2024年2月29日(木)

一般発表 | Track 1: 自然言語処理・機械学習基礎

10:00 ~ 12:10 | T1-A オンライン（Zoom Events)

説明可能AI・データ解釈 (1)

座長:伊藤 寛祥(筑波大学)
コメンテータ:田島 敬史(京都大学)

10:05 ~ 10:30

[T1-A-4-01] Approximate Inverse Model Explanations(AIME)を用いた不完全データによる機械学習モデル挙動評価手法の実現

*阿部 広夢[1]、乾 映之右[1]、佐伯 綾亮[1]、津布久 尚貴[1]、中西 崇文[1] (1. 武蔵野大学データサイエンス学部データサイエンス学科)

10:30 ~ 10:55

[T1-A-4-02] 文章間類似性判定モデルの根拠の抽出

*中井 香那子[1]、河田 友香[1]、山本 岳洋[1] (1. 兵庫県立大学)

10:55 ~ 11:20

[T1-A-4-03] ゼロショットエンティティリンキングに寄与するエンティティ記述の分析

*瀬谷 浩樹[1]、加藤 誠[2] (1. 筑波大学大学院 情報学学位プログラム、2. 筑波大学 図書館情報メディア系)

11:20 ~ 11:45

[T1-A-4-04] 根拠学習とマルチタスク学習を用いた説明可能なヘイトスピーチ検出

*本城 佑太[1]、張 建偉[1] (1. 岩手大学理工学部システム創成工学科知能・メディア情報コース張研究室)

11:45 ~ 12:10

[T1-A-4-05] 時系列グラフニューラルネットワークのリンク予測における反実仮想説明アルゴリズム

*三谷 真裕[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科マルチメディア工学専攻鬼塚研究室)

一般発表 | Track 1: 自然言語処理・機械学習基礎

10:00 ~ 12:10 | T1-B オンライン（Zoom Events)

対話データ処理

座長:酒井 哲也(早稲田大学)
コメンテータ:宮森 恒(京都産業大学)

10:05 ~ 10:20

[T1-B-4-01] SNSを用いたChatGPTの不安感情の特定と解消法の検討

*鍜示 晃大[1]、王 元元[1] (1. 山口大学)

10:20 ~ 10:45

[T1-B-4-02] ラップバトルにおけるアンサーの類型化および類型の自動分類

*三林 亮太[1]、佃 洸摂[2]、渡邉 研斗[2]、中野 倫靖[2]、後藤 真孝[2]、山本 岳洋[1]、大島 裕明[1] (1. 兵庫県立大学、2. 産業技術総合研究所)

10:45 ~ 11:10

[T1-B-4-03] 大規模言語モデルによる投稿の意味的連続性判定手法

*林 想汰[1]、小林 亜樹[2] (1. 工学院大学情報学部情報通信工学科メディア情報システム研究室、2. 工学院大学情報学部情報通信工学科)

11:10 ~ 11:35

[T1-B-4-04] 数個の発話と疑似文脈生成に基づくキャラクタ対話応答生成

*福岡 啓人[1]、宮森 恒[1] (1. 京都産業大学情報理工学部情報理工学科宮森研究室)

11:35 ~ 12:00

[T1-B-4-05] 性格による応答の多様性を考慮したキャラクタ対話データセット構築

*村田 樹[1]、宮森 恒[1] (1. 京都産業大学情報理工学部)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

10:00 ~ 12:10 | T2-A オンライン（Zoom Events)

索引・ストレージアーキテクチャ

座長:北川 博之(筑波大学)
コメンテータ:藤原 真二(日立製作所)

10:05 ~ 10:20

[T2-A-4-01] 同時実行B+木におけるロックフリー手続きの改善と実装

*平野 匠真[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 東海国立大学機構名古屋大学大学院情報学研究科)

10:20 ~ 10:45

[T2-A-4-02] 同時実行B+木のマルチバージョン化の検討

*桑村 真生[1]、杉浦 健人[1]、平野 匠真[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学 情報学研究科)

10:45 ~ 11:10

[T2-A-4-03] リモートログ先行書き込み用のストレージ階層追加による効率化

*XU WENKANG[1]、石井 幹大[1]、Dürst Martin J.[1] (1. 青山学院大学理工学部情報テクノロジー学科)

11:10 ~ 11:35

[T2-A-4-04] コンピュート・ストレージ分離アーキテクチャDBを用いた高可用かつ低コストな冗長構成の提案

*林 友佳[1]、渡 佑也[1]、堀内 美聡[1] (1. NTTソフトウェアイノベーションセンタ)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

10:00 ~ 12:10 | T2-B オンライン（Zoom Events)

グラフデータ処理

座長:新井 淳也(日本電信電話株式会社)
コメンテータ:伏見 卓恭(東京工科大学)

10:05 ~ 10:20

[T2-B-4-01] グラフ要約に基づく知識グラフ索引の提案

*小倉 勇大[1]、天笠 俊之[2] (1. 筑波大学大学院　情報理工学位プログラム　知識・データ工学研究室、2. 筑波大学　計算科学研究センター)

10:20 ~ 10:45

[T2-B-4-02] グラフ深層学習を用いた類似特許検索の精度向上

*中山 樹[1]、中西 宏和[1]、菊地 良将[1]、佐々木 勇和[1]、荒瀬 由紀[1]、鬼塚 真[1] (1. 大阪大学情報科学研究科鬼塚研究室)

10:45 ~ 11:10

[T2-B-4-03] グラフ要約を用いた高速なグラフ相関問合せ

*直井 悠馬[1]、塩川 浩昭[2]、真次 彰平[2] (1. 筑波大学　理工情報生命学術院　知識・データ工学研究室、2. 筑波大学　計算科学研究センター)

11:10 ~ 11:35

[T2-B-4-04] ラベル付き有向グラフにおける node embedding を用いたスキーマ抽出

*木下 一歩[1]、鈴木 伸崇[1] (1. 筑波大学)

11:35 ~ 12:00

[T2-B-4-05] プリミティブなグラフ分析アルゴリズムに対する効率的なグラフ要約の応用

*真次 彰平[1]、塩川 浩昭[1] (1. 筑波大学計算科学研究センター)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

10:00 ~ 12:10 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (1)

座長:杉山 一成(大阪成蹊大学)
コメンテータ:灘本 明代(甲南大学)

10:05 ~ 10:20

[T3-A-4-01] LLMによる商品情報やレビューをもとにした商品推薦

*朴 健[1]、丹羽 健斗[1]、髙村 昇二郎[1]、青井 順一[1] (1. 株式会社MonotaRO)

10:20 ~ 10:45

[T3-A-4-02] レビュー動画内における商品特徴の言及量分析システムの開発

*山口 史弥[1]、上田 真由美[2,3]、中島 伸介[4] (1. 京都産業大学大学院、2. 流通科学大学、3. 大阪大学、4. 京都産業大学)

10:45 ~ 11:10

[T3-A-4-03] 不快コメント予測の個人化のためのコメント特徴に基づく個人化レベルの調整

*中原 輝樹[1]、牛尼 剛聡[1] (1. 九州大学)

11:10 ~ 11:35

[T3-A-4-04] 食品レビュー検索でのクエリ拡張のためのLLMを用いた同義フレーズ生成

*芦澤 亜里紗[1]、三林 亮太[1]、大島 裕明[1] (1. 兵庫県立大学)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

10:00 ~ 12:10 | T3-B オンライン（Zoom Events)

画像検索・生成

座長:古賀 久志(電気通信大学)
コメンテータ:片山 薫(東京都立大学)

10:05 ~ 10:20

[T3-B-4-01] 条件付き画像検索と画像生成の統合

*澤田 一正[1]、坂地 泰紀[1]、野田 五十樹[1]、小山 聡[2] (1. 北海道大学、2. 名古屋市立大学)

10:20 ~ 10:45

[T3-B-4-02] Finding Generative Image LoRA Model by Inputting Style Sample Image

*VUTHI NGOCANH[1]、Shoji Yoshiyuki[1]、Pham HuuLong [2]、Ohshima Hiroaki[2] (1. Shizuoka University、2. University of Hyogo)

10:45 ~ 11:10

[T3-B-4-03] 生成型要約に基づくWebページのサムネイル生成

*前田 直宏[1]、山本 岳洋[1] (1. 兵庫県立大学)

11:10 ~ 11:35

[T3-B-4-04] 類似度グラフの事前再計算による拡散式画像検索の効率化

*加藤 辰弥[1,2]、駒水 孝裕[2]、井手 一郎[2] (1. 名古屋大学大学院知能システム学専攻　井手・駒水研究室、2. 名古屋大学)

11:35 ~ 12:00

[T3-B-4-05] 3Dモデルを利用したユーザーの求めるアングル画像の検索システムの開発

*永田 玄[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学情報理工学院情報工学系情報工学コース宮崎研究室)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

10:00 ~ 12:10 | T4-A オンライン（Zoom Events)

LLM x ユーザ支援

座長:波多野 賢治(同志社大学)
コメンテータ:大島 裕明(兵庫県立大学)

10:05 ~ 10:20
[T4-A-4-01] ブレインストーミングにおける消極的参加者のための大規模言語モデルに基づく発言促進手法

*猪股 朋果[1]、大岡 麗[1]、北山 大輔[1] (1. 工学院大学)

10:20 ~ 10:45
[T4-A-4-02] 意外な単語の組み合わせを用いた発言生成に基づくブレインストーミング支援

*大岡 麗[1]、関 峰[2]、北山 大輔[1] (1. 工学院大学情報学部システム数理学科、2. 工学院大学工学研究科情報学専攻)

10:45 ~ 11:10
[T4-A-4-03] ChatGPTを利用した学術論文読解支援の一手法

*堺 陸斗[1]、金澤 輝一[2]、上野 史[1]、太田 学[1] (1. 岡山大学、2. 国立情報学研究所)

11:10 ~ 11:35
[T4-A-4-04] 大規模言語モデルを用いたクエリ拡張による会話ログの検索

*田貝 奈央[1]、加藤 誠[1] (1. 筑波大学)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

10:00 ~ 12:10 | T4-B オンライン（Zoom Events)

映像の理解と可視化

座長:風間 一洋(和歌山大学)
コメンテータ:中島 伸介(京都産業大学)

10:05 ~ 10:30
[T4-B-4-01] 複数人の競技かるた選手の構え・払いの動作の比較のための三次元ボーン表示および箱棒グラフによる可視化

*北川 リサ[1]、伊藤 貴之[1] (1. お茶の水女子大学伊藤研究室)

10:30 ~ 10:55
[T4-B-4-02] キンボールのプレイ映像解析のための線形フレーム補間による物体追跡

*山下 翔五[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学　工学部　電子システム工学科、2. 滋賀県立大学　先端工学研究院)

10:55 ~ 11:20
[T4-B-4-03] 離脱傾向を踏まえたインターネット動画広告データの可視化システム

*三浦 梨花[1]、大矢 隼士[2]、伊藤 貴之[1] (1. お茶の水女子大学　伊藤研究室、2. Septeni Japan株式会社)

11:20 ~ 11:45
[T4-B-4-04] CLIPを用いたゲーム動画フレーム毎のバグ検知AI

*髙橋 秀太朗[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学　知能情報工学分野、2. 滋賀県立大学)

11:45 ~ 12:10
[T4-B-4-05] 数枚のサンプル画像を用いて類似場面を検出するオンデマンド場面検出技術

秋島 遥[2]、*吉田 登[1]、馬場崎 康敬[1]、劉 健全[1] (1. NEC、2. お茶の水女子大学)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

10:00 ~ 12:10 | T5-A オンライン（Zoom Events)

教育・学習 (4)

座長:林 康弘(武蔵野大学)
コメンテータ:合田 和生(東京大学)

10:05 ~ 10:20

[T5-A-4-01] 道聴と同調について考えさせる体験型アプリケーションDOUCHO2の実装と評価

*杉山 功馬[1]、松村 貫司[1]、吉崎 悠画[1]、伊藤 一成[1] (1. 青山学院大学社会情報学部伊藤一成研究室)

10:20 ~ 10:35

[T5-A-4-02] 教科書の目次構造と科目用語体系に基づく補足Webページの抽出

*中村 愛咲[1]、佐野 逸稀[1]、井上 沙紀[1]、王 元元[2]、河合 由起子[3,4]、角谷 和俊[1] (1. 関西学院大学、2. 山口大学、3. 京都産業大学、4. 大阪大学)

10:35 ~ 11:00

[T5-A-4-03] 大規模言語モデルが生成した論理エラー修正コードと授業資料を活用したプログラミング支援

*上川 貴之[1]、山崎 禎晃[2]、大原 剛三[2] (1. 青山学院大学大学院理工学研究科、2. 青山学院大学理工学部)

11:00 ~ 11:25

[T5-A-4-04] 能動的学習データを用いた学習状況把握のための可視化分析手法の研究

*王 笑難[1]、孫 一[2]、大月 一弘[1]、清光 英成[1] (1. 神戸大学，国際文化学研究科、2. 神戸情報大学院大学，情報技術研究科)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

10:00 ~ 12:10 | T5-B オンライン（Zoom Events)

交通

座長:シリアーラヤ パノット(京都工芸繊維大学)
コメンテータ:北山 大輔(工学院大学)

10:05 ~ 10:30

[T5-B-4-01] 複数台のバスに関する特徴量を活用した都営バスのリアルタイム運行データを用いた渋滞検知

*青柳 宏紀[1]、藤田 智也[1]、畠中 希[2]、小口 正人[2]、山名 早人[1] (1. 早稲田大学、2. お茶の水女子大学)

10:30 ~ 10:55

[T5-B-4-02] 自転車走行中の表情と振動による潜在的快適性分析の検討

*山口 琉太[1]、村重 圭亮[2]、矢野 英人[2]、義久 智樹[3]、河合 由起子[1,2]、下條 真司[4] (1. 京都産業大学、2. 大阪大学、3. 滋賀大学、4. 青森大学)

10:55 ~ 11:20

[T5-B-4-03] 強化学習による差動二輪車制御における内外乱適応

*門垣 幸樹[1]、大島 裕明[1] (1. 兵庫県立大学)

11:20 ~ 11:45

[T5-B-4-04] A Forecasting Method of Traffic Flow Using Road Network Data

*田中 悠[1]、笠松 大佑[2] (1. 創価大学大学院理工学研究科笠松研究室、2. 創価大学大学院理工学研究科)

11:45 ~ 12:10

[T5-B-4-05] 都営バスのオープンデータおよびディープラーニングを用いた渋滞検知手法の提案と評価

*畠中 希[1]、青柳 宏紀[2]、藤田 智也[2]、山名 早人[2]、小口 正人[1] (1. お茶の水女子大学、2. 早稲田大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

13:30 ~ 15:40 | T1-A オンライン（Zoom Events)

説明可能AI・データ解釈 (2)

座長:丸橋 弘治(富士通株式会社)
コメンテータ:佐々木 佑樹(富士通株式会社)

13:35 ~ 14:00

[T1-A-5-01] マルチエージェント強化学習に基づく解釈可能なソーシャルネットワークの将来予測

*三宅 健太郎[1]、伊藤 寛祥[1]、Christos Faloutsos[2]、松本 寛智[1]、森嶋 厚行[1] (1. 筑波大学、2. Carnegie Mellon University)

14:00 ~ 14:25

[T1-A-5-02] 健康行動を促す介入メッセージのパーソナライゼーションに向けた文の特徴量分析

*郷原 聖士[1]、瀧本 祥章[2]、金 秀明[2]、小林 のぞみ[2]、倉島 健[2]、渡辺 太郎[1] (1. 奈良先端科学技術大学院大学、2. NTT人間情報研究所)

14:25 ~ 14:50

[T1-A-5-03] Approximate Inverse Model Explanations (AIME)を用いた内閣支持率変化要因抽出手法

*津布久 尚貴[1]、岡田 龍太郎[2]、峰松 彩子[2]、中西 崇文[2] (1. 武蔵野大学データサイエンス学部データサイエンス学科 TransMedia Tech Lab、2. 武蔵野大学データサイエンス学部)

14:50 ~ 15:15

[T1-A-5-04] 2次元最適相関ルールを用いたSHAPの解釈補完

*塩谷 泰平[1]、森本 康彦[1]、亀井 清華[1] (1. 広島大学情報科学部情報科学科)

15:15 ~ 15:40

[T1-A-5-05] インドのITサービス会社のShapley値を用いた15年間の分析

*白田 由香利[1]、Sreekanth Bonthala[2]、Chakraborty Basabi[2,3] (1. 学習院大学経済学部、2. マダナパレ工科大学、3. 岩手県立大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

13:30 ~ 15:40 | T1-B オンライン（Zoom Events)

対話システム応用

座長:田中 克己(福知山公立大学)
コメンテータ:波多野 賢治(同志社大学)

13:35 ~ 14:00

[T1-B-5-01] 生成AIを利用した自己学習支援

*伊藤 優希[1]、馬 強[2] (1. 京都大学大学院 情報学研究科 社会情報学専攻 伊藤研究室、2. 京都工芸繊維大学 大学院工芸科学研究科)

14:00 ~ 14:25

[T1-B-5-02] LLMを用いた対話型鑑賞エージェントにおける繰り返し発話の抑制機構

*谷 悠翔[1]、牛尼 剛聡[2] (1. 九州大学芸術工学府メディアデザインコース　牛尼研究室、2. 九州大学大学院 芸術工学研究院)

14:25 ~ 14:50

[T1-B-5-03] 複数の大規模言語モデルによる漫才自動生成の比較分析

*下崎 安紋[1]、北畑 哲也[3]、柘植 陽介[3]、梅谷 智弘[2]、北村 達也[2]、灘本 明代[2] (1. 甲南大学大学院自然科学研究科知能情報学専攻　灘本研究室、2. 甲南大学知能情報学部知能情報学科、3. 甲南大学知能情報学部知能情報学科　灘本研究室)

14:50 ~ 15:15

[T1-B-5-04] 授業改善のための意見を引き出すAIインタビュアー

*中西 勇輝[1]、桑田 若菜[2]、三林 亮太[2]、山本 祐輔[3]、大島 裕明[4] (1. 兵庫県立大学社会情報科学部大島裕明研究室、2. 兵庫県立大学情報科学研究科大島裕明研究室、3. 名古屋市立大学データサイエンス学部、4. 兵庫県立大学情報科学研究科)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

13:30 ~ 15:40 | T2-A オンライン（Zoom Events)

データ基盤・性能評価

座長:杉浦 健人(名古屋大学)
コメンテータ:小山田 昌史(NEC)

13:35 ~ 14:00

[T2-A-5-01] 自律分散型データ統合技術Dejimaの性能分析

*吉田 凌河[1]、肖 川[1,2]、鬼塚 真[1] (1. 大阪大学 情報科学研究科 鬼塚研究室、2. 名古屋大学)

14:00 ~ 14:25

[T2-A-5-02] 不揮発メモリを対象とする空間索引のSSDとの性能比較

*吉岡 弘隆[1]、合田 和生[2]、喜連川 優[2] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所)

14:25 ~ 14:50

[T2-A-5-03] モバイル環境における圧縮処理を用いたIoTシステムの通信性能向上のための検討

*伊藤 千紗[1]、竹房 あつ子[2]、中田 秀基[3]、小口 正人[1] (1. お茶の水女子大学、2. 国立情報学研究所、3. 産業技術総合研究所)

14:50 ~ 15:15

[T2-A-5-04] ROS 2準拠ロボット及びクラウドを用いたIoTアプリケーションのためのZenohの通信性能評価

*佐々木 怜名[1]、竹房 あつ子[2]、中田 秀基[3]、小口 正人[1] (1. お茶の水女子大学部人間文化創成科学研究科理学専攻小口研究室、2. 国立情報学研究所、3. 産業技術総合研究所)

15:15 ~ 15:40

[T2-A-5-05] 【技術報告】広告配信における安定して拡張性のある大量データ処理基盤の必要性と活用

大澤 昂太 (株式会社マイクロアド システム開発部)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

13:30 ~ 15:40 | T2-B オンライン（Zoom Events)

類似検索アルゴリズム

座長:白井 匡人(島根大学)
コメンテータ:鈴木 伸崇(筑波大学)

13:35 ~ 14:00

[T2-B-5-01] 逆 k-MIPSに基づく影響力の大きいアイテムの高速検索

*青山 和禎[1]、天方 大地[1]、藤田 澄男[2]、原 隆浩[1] (1. 大阪大学大学院情報科学研究科、2. LINEヤフー株式会社)

14:00 ~ 14:25

[T2-B-5-02] A Grid-density-based Approach for Dense Nearest Neighborhood Query

*謝 若琳[1]、陳 漢雄[1]、古瀬 一隆[2] (1. 筑波大学、2. 白鴎大学)

14:25 ~ 14:50

[T2-B-5-03] アニメーション付き 3 次元アセンブリモデル検索のための3次元ラドン変換の効率化

*内尾 亮太[1]、片山 薫[1] (1. 東京都立大学)

14:50 ~ 15:15

[T2-B-5-04] 大規模DNAデータベースに対するTop-k相関問合せ手法の提案

*八木 隆一[1]、直井 悠馬[1]、塩川 浩昭[2] (1. 筑波大学大学院理工情報生命学術院、2. 筑波大学計算科学研究センター)

15:15 ~ 15:40

[T2-B-5-05] テキスト及び構造に基づく効率的な類似部分木検索手法

*溝上 拓也[1]、Bou Savong[2]、天笠 俊之[2] (1. 筑波大学 情報学群情報科学類、2. 筑波大学 計算科学研究センター)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

13:30 ~ 15:40 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (2)

座長:井上 潮(東京電機大学)
コメンテータ:井手 一郎(名古屋大学)

13:35 ~ 13:50

[T3-A-5-01] コンテンツ振り返りのためのRAGを利用した対話型フィードバックによるレビュー検索

*鈴木 醇[1]、牛尼 剛聡[2] (1. 九州大学芸術工学部芸術工学科 牛尼研究室、2. 九州大学大学院芸術工学研究院)

13:50 ~ 14:15

[T3-A-5-02] J-POPにおける楽曲動画コメントを用いた歌声印象軸の構築

*陳 沢璨[1]、熊本 忠彦[1] (1. 千葉工業大学大学院)

14:15 ~ 14:40

[T3-A-5-03] 単語分散表現を用いたホテルレビュー分析に基づく評価項目別自動スコアリング

*林 裕次郎[1]、張 建偉[2]、河合 由起子[1]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学)

14:40 ~ 15:05

[T3-A-5-04] 追加事前学習によるレビューと商品属性の対応関係学習

*河田 友香[1]、山本 岳洋[1]、大島 裕明[1]、藤田 澄男[2] (1. 兵庫県立大学、2. LINEヤフー株式会社)

15:05 ~ 15:30

[T3-A-5-05] 動画の属するコミュニティ情報を考慮した動画再生回数予測手法の提案

*川上 大凱[1]、鈴木 優[1] (1. 岐阜大学)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

13:30 ~ 15:40 | T3-B オンライン（Zoom Events)

インタラクティブ情報アクセス

座長:中野 美由紀(津田塾大学)
コメンテータ:関 洋平(筑波大学)

13:35 ~ 13:50

[T3-B-5-01] 推薦システムにおけるユーザの新たな嗜好の発見を支援する
NeuMF を用いたインターフェース

*坂本 新真[1]、牛尼 剛聡[1] (1. 九州大学　牛尼剛聡研究室)

13:50 ~ 14:15

[T3-B-5-02] 探索的検索における検索視野拡張のための仮想他者視点の生成と評価

*小竹 神[1]、北山 大輔[1] (1. 工学院大学)

14:15 ~ 14:40

[T3-B-5-03] 観点の自動抽出に基づくインタラクティブな観光スポット推薦システム

*小林 らんう[1]、上野 史[2]、太田 学[2] (1. 岡山大学大学院環境生命自然科学研究科、2. 岡山大学学術研究院環境生命自然科学学域)

14:40 ~ 15:05

[T3-B-5-04] 大規模言語モデルを用いたその場での要約に基づくレビュー探索インタフェース

*藤井 真梨乃[1]、河田 友香[1]、山本 岳洋[1] (1. 兵庫県立大学)

15:05 ~ 15:30

[T3-B-5-05] 個人に最適な生活改善行動を対話的に探索するためのYES/Noチャート自動生成

*織田 直也[1]、金 鎭赫[1]、荘司 慶行[1]、山本 祐輔[2] (1. 静岡大学大学院 総合科学技術研究科 、2. 名古屋市立大学 データサイエンス学部)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

13:30 ~ 15:40 | T4-A オンライン（Zoom Events)

LLM応用

座長:石川 佳治(名古屋大学)
コメンテータ:上田 真由美(流通科学大学)

13:35 ~ 14:00

[T4-A-5-01] 欲求度に着目した行動手順漫才台本自動生成

*柘植 陽介[1]、下崎 安紋[2]、灘本 明代[3] (1. 甲南大学知能情報学部知能情報学科　灘本研究室、2. 甲南大学大学院自然科学研究科知能情報学専攻　灘本研究室、3. 甲南大学知能情報学部知能情報学科)

14:00 ~ 14:25

[T4-A-5-02] 意思決定における大規模言語モデルを用いた検索要求の観点抽出の影響分析

*笹本 まこと[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

14:25 ~ 14:50

[T4-A-5-03] 歌詞と音響特徴の分析に基づくText2Imageモデルを用いた楽曲サムネイル画像の自動生成

*佐々木 翔一[1]、西田 紘子[2]、澤井 賢一[2]、工藤 真生[2]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府、2. 九州大学大学院芸術工学研究院)

14:50 ~ 15:15

[T4-A-5-04] チャート画像に対する大規模マルチモーダルモデルの反実仮想推論能力の検証

*戸崎 友輔[1]、宮森 恒[1] (1. 京都産業大学情報理工学部情報理工学科)

15:15 ~ 15:40

[T4-A-5-05] 【技術報告】NTT版LLM「tsuzumi」の取組について

西田　京介 (NTT 人間情報研究所)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

13:30 ~ 15:40 | T4-B オンライン（Zoom Events)

クラウドソーシングと人間・社会理解 (1)

座長:田島 敬史(京都大学)
コメンテータ:清水 敏之(九州大学)

13:35 ~ 14:00

[T4-B-5-01] 男女差を俯瞰するための階層型データ可視化: EMDによる男女間の分布差の導入

*中井 祐希[1]、伊藤 貴之[1] (1. お茶の水女子大学伊藤研究室)

14:00 ~ 14:25

[T4-B-5-02] Dawid-Skene集約における不均一タスク数の影響緩和による人間-AIハイブリッドクラウドソーシングの品質向上

*田村 匠[1]、伊藤 寛祥[2]、小山 聡[3]、森嶋 厚行[2] (1. 筑波大学情報学群、2. 筑波大学図書館情報メディア系、3. 名古屋市立大学データサイエンス学部)

14:25 ~ 14:50

[T4-B-5-03] フィジカルクラウドソーシングにおける分担タスク割当てのためのワーカ依頼順序の最適化

*藤永 弥太郎[1]、伊藤 寛祥[1]、鈴木 伸崇[1]、森嶋 厚行[1] (1. 筑波大学)

14:50 ~ 15:15

[T4-B-5-04] クラウドワーカの作業結果と投票結果から生成される確率分布を利用した結果集約手法

*城所 祥太[1]、鈴木 優[1] (1. 岐阜大学工学部電気電子情報工学科情報コース　鈴木研究室)

15:15 ~ 15:40

[T4-B-5-05] 理想と現実のボディイメージの数理モデル化

*栗生 紗希帆[1]、眞鍋 雅恵[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

13:30 ~ 15:40 | T5-A オンライン（Zoom Events)

医療・ヘルスケア (1)

座長:中島 伸介(京都産業大学)
コメンテータ:張 建偉(岩手大学)

13:35 ~ 14:00

[T5-A-5-01] 睡眠ステージ判定を対象としたエンコーダ・デコーダモデルによる脳波信号オーグメンテーション

*大宮 直樹[1]、堀江 和正[2]、北川 博之[3] (1. 筑波大学大学院システム情報工学研究群情報理工学位プログラム、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構)

14:00 ~ 14:25

[T5-A-5-02] 睡眠ステージ自動判定における睡眠時無呼吸症候群の重症度を踏まえた深層学習処理の切替機構の開発

*川畑 竜平[1]、堀江 和正[1]、北川 博之[1] (1. 筑波大学)

14:25 ~ 14:50

[T5-A-5-03] 気象データを統合的に用いたクラウドソーシングに基づく頭痛発生傾向の検知

*飯田 静空[1]、若宮 翔子[1]、清水 伸幸[2]、藤田 澄男[2]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. LINEヤフー株式会社)

14:50 ~ 15:15

[T5-A-5-04] データ駆動型研究を促進する大規模かつ多種類の生体光学データのためのデータベース基盤の開発

*山中 卓也[1]、熊本 康昭[2,3]、西村 隆宏[2]、久保 俊貴[2,4]、Heqi Xi[2]、藤田 克昌[2]、松原 靖子[1]、櫻井 保志[1] (1. 大阪大学産業科学研究所産業科学AIセンター、2. 大阪大学大学院工学研究科、3. 大阪大学先導的学際研究機構、4. 大阪大学大学院医学系研究科)

15:15 ~ 15:40

[T5-A-5-05] 【技術報告】類型化と解釈性を考慮した可視化による自動車事故分析

桑元　凌 (三井住友海上火災保険株式会社ビジネスデザイン部)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

13:30 ~ 15:40 | T5-B オンライン（Zoom Events)

ルート探索

座長:勝沼 聡((株)日立製作所)
コメンテータ:若宮 翔子(奈良先端科学技術大学院大学)

13:35 ~ 13:50

[T5-B-5-01] ランドマークを考慮した経路探索フレームワークの検証

*大滝 啓介[1]、吉村 貴克[1] (1. 豊田中央研究所)

13:50 ~ 14:15

[T5-B-5-02] 引力モデルを用いた経路探索手法の提案

*筒井 夏輝[1]、横山 昌平[1,2,3] (1. 東京都立大学、2. 国立情報学研究所、3. 東京大学)

14:15 ~ 14:40

[T5-B-5-03] 移動手段の混雑を考慮した観光ルートプランニング

*田中 寛之[1]、馬 強[1] (1. 京都工芸繊維大学)

14:40 ~ 15:05

[T5-B-5-04] 3D都市モデルを利用して移動中の景観を考慮する観光経路推薦手法の検討

*斎藤 康平[1]、横山 大作[1] (1. 明治大学)

15:05 ~ 15:30

[T5-B-5-05] 個人化レベルのメリハリを考慮した観光ルートプランニング

*廣澤 凛太郎[1]、宜 鯤[2]、馬 強[1] (1. 京都工芸繊維大学、2. 京都大学経済研究所)

一般発表 | Track 1: 自然言語処理・機械学習基礎

16:00 ~ 18:10 | T1-A オンライン（Zoom Events)

構造データ処理

座長:神崎 映光(島根大学)
コメンテータ:伊藤 寛祥(筑波大学)

16:05 ~ 16:30

[T1-A-6-01] マウス版Geneformerの構築と転移学習による細胞解析への応用

*伊藤 啓太[1]、平川 翼[1]、山下 隆義[1]、藤吉 弘亘[1] (1. 中部大学)

16:30 ~ 16:55

[T1-A-6-02] ペルソナ・モデリングによるソーシャルネットワークの将来予測の精度向上

*松本 寛智[1]、伊藤 寛祥[2]、三宅 健太郎[1]、森嶋 厚行[2] (1. 筑波大学情報学群、2. 筑波大学図書館情報メディア系)

16:55 ~ 17:20

[T1-A-6-03] 時間的特徴を制御可能な時間グラフ生成器

*中西 宏和[1]、菊地 良将[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学情報科学研究所マルチメディア専攻鬼塚研究室)

17:20 ~ 17:45

[T1-A-6-04] 時間グラフ生成に関する包括的な調査とベンチマーキング

*菊地 良将[1]、中西 宏和[1]、佐々木 勇和[1]、Yakovets Nikolay[2]、Fletcher George[2]、鬼塚 真[1] (1. 大阪大学情報科学研究科、2. Eindhoven University of Technology)

17:45 ~ 18:10

[T1-A-6-05] ExhaustiveTree: 網羅的な二次元座標木を用いた表の表現

*林 知司[1]、宮森 恒[2] (1. 京都産業大学大学院 宮森研究室、2. 京都産業大学大学院)

一般発表 | Track 1: 自然言語処理・機械学習基礎

16:00 ~ 18:10 | T1-B オンライン（Zoom Events)

文生成

座長:佐々木 佑樹(富士通株式会社)
コメンテータ:酒井 哲也(早稲田大学)

16:05 ~ 16:30

[T1-B-6-01] BERTを用いた感情推定に基づく発話スタイル付き応答文生成システム

*西 佑真[1]、梶岡 慎輔[1]、山本 大介[1] (1. 名古屋工業大学)

16:30 ~ 16:55

[T1-B-6-02] 経路単純化と案内文単純化を組み合わせた経路案内文要約システム

*鋤柄 知哉[1]、金 鎔煥[1]、山本 大介[1] (1. 名古屋工業大学)

16:55 ~ 17:20

[T1-B-6-03] 大規模言語モデルによる商品レビュー文のキャッチコピーへの変換

*沈 辰夫[1]、荘司 慶行[2]、山本 岳洋[3]、大原 剛三[1]、田中 克己[4] (1. 青山学院大学、2. 静岡大学、3. 兵庫県立大学、4. 福知山公立大学)

17:20 ~ 17:45

[T1-B-6-04] 特定選手に注目した試合結果記事見出しのパーソナライズ

*佐々木 泰河[1]、工家 昂之[1]、山本 岳洋[1]、荘司 慶行[2] (1. 兵庫県立大学、2. 静岡大学)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

16:00 ~ 18:10 | T2-A オンライン（Zoom Events)

先進ハードウェア活用・性能評価

座長:藤原 靖宏(日本電信電話株式会社)
コメンテータ:油井 誠(トレジャーデータ株式会社)

16:05 ~ 16:30

[T2-A-6-01] サーバシステムの性能データ転送における効率化手法の評価と考察

*飯山 知香[1]、平井 聡[2]、山岡 茉莉[2]、福本 尚人[2]、小口 正人[1] (1. お茶の水女子大学、2. 富士通株式会社)

16:30 ~ 16:55

[T2-A-6-02] 量子ランダムアクセスメモリ上での関係データベースの構成法の一検討

*御手洗 陽紀[1]、合田 和生[2] (1. 東京大学 大学院情報理工学系研究科、2. 東京大学 生産技術研究所)

16:55 ~ 17:20

[T2-A-6-03] 永続メモリ向けロックフリー索引Bz木に関する研究

*中山 宗[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学大学院情報学研究科)

17:20 ~ 17:45

[T2-A-6-04] ラベルの出現頻度に着目したFPGAを用いた正規パス問合せの提案

*溝谷 祐大[1]、小林 諒平[2]、藤田 典久[2]、朴 泰祐[2]、天笠 俊之[2] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群情報学位プログラム　知識・データ工学研究室(KDE)、2. 筑波大学 計算科学研究センター)

17:45 ~ 18:10

[T2-A-6-05] 量子回路シミュレータの実行環境変化に伴う性能評価と特性比較

*青木 望美[1]、山崎 雅文[2]、平井 聡[2]、山岡 茉莉[2]、福本 尚人[2]、小口 正人[3] (1. お茶の水女子大学人間文化創成科学研究科博士前期課程理学専攻情報科学コース小口研究室、2. 富士通（株）、3. お茶の水女子大学理学部情報科学科)

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

16:00 ~ 18:10 | T2-B オンライン（Zoom Events)

時系列データ分類・予測

座長:天笠 俊之(筑波大学)
コメンテータ:杉浦 健人(名古屋大学)

16:05 ~ 16:30

[T2-B-6-01] 階層的に細分化したマルチレベルイベントタイプからなる長時間マルチレベルエピソード抽出手法

*橋本 一輝[1]、新谷 隆彦[1]、大森 匡[1]、藤田 秀之[1] (1. 電気通信大学大学院情報理工学研究科)

16:30 ~ 16:55

[T2-B-6-02] 動的モード分解を活用した高速将来予測アルゴリズム

*千原 直己[1,2]、松原 靖子[1]、藤原 廉[1,2]、櫻井 保志[1] (1. 大阪大学産業科学研究所、2. 大阪大学大学院情報科学研究科)

16:55 ~ 17:20

[T2-B-6-03] 学習可能な長さを持つshapeletsに基づく時系列分類法

*山口 晃広[1]、植野 研[1]、鹿島 久嗣[2] (1. 東芝研究開発センター 知能化システム研究所 システムAIラボラトリー、2. 京都大学 大学院情報学研究科 知能情報学専攻)

17:20 ~ 17:45

[T2-B-6-04] 合成シェイプレットを用いた時系列データ分類

*岡部 正幸[1] (1. 県立広島大学)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

16:00 ~ 18:10 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (3)

座長:野宮 浩揮(京都工芸繊維大学)
コメンテータ:有次 正義(熊本大学)

16:05 ~ 16:20

[T3-A-6-01] レビュー分析に基づくユーザ目線でのシーン別スポットスコアリング手法の提案

*大岡 芙実[1]、福嶋 亮太朗[1]、張 建偉[2]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学)

16:20 ~ 16:45

[T3-A-6-02] 直感的動画検索における単語分散表現を用いた評価項目別スコアリング手法

*笹原 彰斗[1]、山口 史弥[2]、上田 真由美[3,4]、中島 伸介[1] (1. 京都産業大学、2. 京都産業大学大学院、3. 流通科学大学、4. 大阪大学)

16:45 ~ 17:10

[T3-A-6-03] 生成系言語モデルによる架空のレビューを介した映画検索

*宮下 天祥[1]、荘司 慶行[2]、藤田 澄男[3]、大原 剛三[1] (1. 青山学院大学、2. 静岡大学、3. LINEヤフー株式会社)

17:10 ~ 17:35

[T3-A-6-04] SNSコメントを利用した対照学習によるニュースの潜在的な関係性の抽出と可視化

*高橋 陸[1]、牛尼　剛聡[2] (1. 九州大学大学院芸術工学府牛尼研究室、2. 九州大学大学院芸術工学研究院)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

16:00 ~ 18:10 | T3-B オンライン（Zoom Events)

Beyond Accuracy

座長:江口 浩二(広島大学)
コメンテータ:山名 早人(早稲田大学)

16:05 ~ 16:30

[T3-B-6-01] 音楽推薦におけるセレンディピティ向上のための
アーティストエンコーダ

*佐藤 諒平[1]、SITKRONGWONG PADIPAT[1]、高須 淳宏[3]、相原 健朗[2] (1. 総合研究大学院大学　複合科学研究科　情報学専攻　高須淳宏研究室、2. 東京都立大学 大学院 都市環境科学研究科 観光科学域 教授国立情報学研究所 客員教授(兼任)社会データ構造化センター 客員教授、3. 国立情報学研究所 コンテンツ科学研究系 教授,総合研究大学院大学 複合科学研究科 情報学専攻)

16:30 ~ 16:55

[T3-B-6-02] empty-answer問題に対する多様性を考慮した効率的な探索手法

*池田 悠人[1]、肖 川[1]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科)

16:55 ~ 17:20

[T3-B-6-03] 相互近接性に基づくCollaborative Metric Learningの意外性向上手法

*中嶋 太一[1]、陳 漢雄[1]、古瀬 一隆[2] (1. 筑波大学、2. 白鴎大学)

17:20 ~ 17:45

[T3-B-6-04] δ-spannerを使用した位置匿名化によるモバイル端末からのイベント検索

*石神 京佳[1]、榎 美紀[2]、小口 正人[1] (1. お茶の水女子大学、2. 日本アイ・ビー・エム)

17:45 ~ 18:10

[T3-B-6-05] 好奇心に基づいたユーザ適応型セレンディピティ推薦手法の提案

*徐 哲林[1]、王 旭[1]、松村 敦[2] (1. 筑波大学人間総合科学学術院人間総合科学研究群情報学学位プログラム、2. 筑波大学図書館情報メディア系)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

16:00 ~ 18:10 | T4-A オンライン（Zoom Events)

AIとの対話

座長:灘本 明代(甲南大学)
コメンテータ:山本 祐輔(静岡大学)

16:05 ~ 16:20

[T4-A-6-01] AIは「思いやり」を持った反応ができるか？

*川内野 穂香[1,3]、牛尼 剛聡[2,3] (1. 長崎県立長崎東高等学校、2. 九州大学 大学院芸術工学研究院、3. 九州大学 未来創生科学者育成プロジェクト)

16:20 ~ 16:35

[T4-A-6-02] タスク遂行促進のための作業リアクションの自動生成手法の有効性評価

*堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

16:35 ~ 17:00

[T4-A-6-03] 小説の感想を語り合えるエージェント構築のためのレビューをもとにした小説解釈の再現

*笹原 雄太[1]、牛尼 剛聡[1] (1. 九州大学)

17:00 ~ 17:25

[T4-A-6-04] 対人スキルと対AIスキルの関係

*西岡 竜生[1]、若宮 翔子[1]、清水 伸幸[2]、藤田 澄男[2]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. LINEヤフー株式会社)

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

16:00 ~ 18:10 | T4-B オンライン（Zoom Events)

クラウドソーシングと人間・社会理解 (2)

座長:吉川 正俊(京都大学)
コメンテータ:シリアーラヤ パノット(京都工芸繊維大学)

16:05 ~ 16:20

[T4-B-6-01] 単語埋め込みのジェンダーバイアスとデバイアスの影響の可視化

*杉野 有咲[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:20 ~ 16:45

[T4-B-6-02] TVMLを用いた講師映像が受講生に対して与える影響の分析

*廣岡 采華[1]、熊本 忠彦[1] (1. 千葉工業大学)

16:45 ~ 17:10

[T4-B-6-03] プロジェクト型クラウドソーシングにおけるワーカのスキル推薦

*大石 啓登[1]、伊藤 寛祥[2]、豊田 正史[3]、森嶋 厚行[2] (1. 筑波大学 情報学群、2. 筑波大学 図書館情報メディア系、3. 東京大学 生産技術研究所)

17:10 ~ 17:35

[T4-B-6-04] 映像集中を目的とした第二言語字幕における音声変化を考慮した字幕提示手法

*西 優己[1]、中村 優吾[1]、福嶋 政期[1]、荒川 豊[1] (1. 九州大学)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

16:00 ~ 18:10 | T5-A オンライン（Zoom Events)

医療・ヘルスケア (2)

座長:横田 治夫(城西大学)
コメンテータ:白石 優旗(筑波技術大学)

16:05 ~ 16:30

[T5-A-6-01] 福祉支援施設の支援記録からのインシデントの予兆検出手法の改良

*山本 隆太[1]、上野 史[2]、太田 学[2] (1. 岡山大学大学院環境生命自然科学研究科、2. 岡山大学学術研究院環境生命自然科学学域)

16:30 ~ 16:55

[T5-A-6-02] レム睡眠のリアルタイム自動検出に関する研究

*佐野 史弥[1]、国生 泰資[1]、堀江 和正[2]、古木 淳也[4]、鈴木 陽子[3]、阿部 高志[3]、北川 博之[3] (1. 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構、4. 筑波大学大学院人間総合科学学術院 人間総合科学研究群)

16:55 ~ 17:20

[T5-A-6-03] リアルタイム睡眠ステージ判定システムの開発と評価

*国生 泰資[1]、堀江 和正[2]、鈴木 陽子[3]、阿部 高志[3]、北川 博之[3] (1. 筑波大学システム情報工学研究群知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構)

17:20 ~ 17:45

[T5-A-6-04] 特定健診データに含まれる自由記述に着目した予後予測

*大塚 皇輝[1]、池之上 辰義[2]、福間 真悟[3]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学　ソーシャルコンピューティング研究室、2. 滋賀大学　データサイエンス・AIイノベーション研究推進センター、3. 京都大学 医学研究科 人間健康科学科系専攻)

17:45 ~ 18:10

[T5-A-6-05] AI 音楽療法士（さつきさん）の開発研究

*小杉 尚子[1]、石井 健太郎[1]、児玉 直樹[2] (1. 専修大学、2. 新潟医療福祉大学)

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

16:00 ~ 18:10 | T5-B オンライン（Zoom Events)

ルート推薦

座長:小林 亜樹(工学院大学)
コメンテータ:角谷 和俊(関西学院大学)

16:05 ~ 16:20

[T5-B-6-01] Happyスポットを考慮したウォーキング経路推薦手法の提案

*荒木 賢斗[1]、Siriaraya Panote[2]、角谷 和俊[3]、河合 由起子[1]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学、3. 関西学院大学)

16:20 ~ 16:45

[T5-B-6-02] 安全性を考慮した夜間の徒歩経路推薦システムの提案と評価

*横井 茉紘[1]、石原 靖哲[1] (1. 南山大学)

16:45 ~ 17:10

[T5-B-6-03] 公共交通機関の利用を想定した観光ルート推薦の一手法

*田中 俊[1]、上野 史[1]、太田 学[1] (1. 岡山大学)

17:10 ~ 17:35

[T5-B-6-04] 人流データに基づく観光スポットの周辺混雑度の予測と旅行計画への影響分析

*星 結人[1]、小竹 神[1]、北山 大輔[1] (1. 工学院大学)

17:35 ~ 18:00

[T5-B-6-05] ユーザの性格タイプの差異に基づく災害時行動促進情報理解の分析

*若杉 広介[1]、鈴木 優[2]、灘本 明代[3] (1. 甲南大学大学院自然科学研究科、2. 岐阜大学工学部、3. 甲南大学知能情報学部知能情報学科)

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 10:00 ~ 12:10 | T1-A オンライン（Zoom Events)

説明可能AI・データ解釈 (1)

座長:伊藤 寛祥(筑波大学)
コメンテータ:田島 敬史(京都大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T1-A-4-01] Approximate Inverse Model Explanations(AIME)を用いた不完全データによる機械学習モデル挙動評価手法の実現

*阿部 広夢[1]、乾 映之右[1]、佐伯 綾亮[1]、津布久 尚貴[1]、中西 崇文[1] (1. 武蔵野大学データサイエンス学部データサイエンス学科)

10:30 ~ 10:55
[T1-A-4-02] 文章間類似性判定モデルの根拠の抽出

*中井 香那子[1]、河田 友香[1]、山本 岳洋[1] (1. 兵庫県立大学)

10:55 ~ 11:20
[T1-A-4-03] ゼロショットエンティティリンキングに寄与するエンティティ記述の分析

*瀬谷 浩樹[1]、加藤 誠[2] (1. 筑波大学大学院 情報学学位プログラム、2. 筑波大学 図書館情報メディア系)

11:20 ~ 11:45
[T1-A-4-04] 根拠学習とマルチタスク学習を用いた説明可能なヘイトスピーチ検出

*本城 佑太[1]、張 建偉[1] (1. 岩手大学理工学部システム創成工学科知能・メディア情報コース張研究室)

11:45 ~ 12:10
[T1-A-4-05] 時系列グラフニューラルネットワークのリンク予測における反実仮想説明アルゴリズム

*三谷 真裕[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科マルチメディア工学専攻鬼塚研究室)

Approximate Inverse Model Explanations(AIME)を用いた不完全データによる機械学習モデル挙動評価手法の実現

阿部 広夢† 乾 映之右† 佐伯 綾亮† 津布久 尚貴† 中西 崇文†

†武蔵野大学データサイエンス学部データサイエンス学科 〒135-8181 東京都江東区有明 3 丁目 3-3

E-mail: †{s2122072, s2322007, s2222015, s2322046}@stu.musashino-u.ac.jp, takafumi.nakanishi@ds.musashino-u.ac.jp

あらまし 本研究では，説明可能な AI（XAI）の手法である，我々がこれまで提案してきた Approximate Inverse Model Explanations(AIME)を用いた不完全データによる機械学習モデルの挙動の評価手法を示し，検証を行う．AIME は対象とする機械学習モデルの近似逆作用素を構築することにより大局的・局所的説明を導出する手法である．本研究では，表データに意図的に欠損値を混入させ，複数種類の欠損値処理を施したデータを用意し，これらを学習した機械学習モデルを対象として AIME を用いてその機械学習モデルの挙動を評価できるかを検証する．欠損値処理はリストワイズ除去，平均値補完，多重代入，KNN による補完を行う．またモデルの内部挙動を評価するための指標 AIOD(近似逆作用素差：Approximate Inverse Operator Difference)を提案する．本研究は，不完全データを扱う際の機械学習モデルの挙動を評価する一つの手法として，機械学習モデルの透明性と信頼性の向上に貢献することを目指す．

キーワード XAI，機械学習，不完全データ，欠損値処理，AIME

1. はじめに

現代のデジタル社会において，機械学習は計算機科学を越えた広範な影響を及ぼしている．例としては，大規模言語モデルの ChatGPT[1]，自動運転，医用画像診断，Netflix や Amazon における映画や商品の推薦システムなどが挙げられる．機械学習は日常生活の様々な領域で意思決定の支援ツールとして活用されているが，これらの機械学習モデルは，その内部動作を直感的に理解することが難しいという問題を抱えている．モデルの出力結果が正確であっても，その結果がどのような入力に基づいて導かれたのか，モデル自体からは明確に説明されない．このようなモデルの不透明性は特に医用画像診断のような重要な決定を行う際に問題となる．

この問題に対処するため，説明可能な AI（XAI）の研究が進展している[2][3][4]．XAI は，モデルの出力がどのように導かれたのかを示し，ユーザーに対してモデルの説明性と透明性を提供することで，モデルの出力の信頼性を高めるものである．Adadi ら[5]は，XAI の重要性を次の 4 点で述べている:出力結果の正当化，モデルの脆弱性に起因するエラーへの対応，モデルの改善，およびモデル処理から得られる新たな知見の獲得である．EU の一般データ保護規則（GDPR）[6]など，機械学習モデルの説明性を求める動きが見られることからも，XAI は今後機械学習と社会の関係において重要な役割を益々果たすことになると考えられる．

機械学習の進展の原因には，データセットの拡充やデータ収集手法の発展が一因として考えられるが，その中で収集されるデータには欠損値を含む不完全なものが多く存在する．このようなデータを「不完全データ」といい，逆に欠損値のないデータを「完全データ」という．欠損値の原因には主にデータ収集のミスやデータ提供の拒否など様々な理由が考えられる．通常，表データを対象とした機械学習の学習は完全データであることが望まれ，不完全データの場合は欠損値処理をする必要がある．しかし欠損値処理が不適切であると，モデル内部に対して真のデータから離れたバイアスが発生し，モデルの精度に影響を与える[7]．特に医療分野のように重要な決定に関わる場合，欠損値処理は特に慎重に行う必要があり，医療データの欠損値処理において，現在多くの欠損値処理手法が提案され，その中で，機械学習モデルへの影響評価が行われてきた[8][9][10][11][12]．

しかしこれまでの欠損値処理の評価は基本的に一致率や精度などモデルの出力により求められるものとされており[13]，モデル内部への評価もまた一致率や精度によるものであるが，これではモデルの内部挙動を直接的に評価しておらず，内部評価としては不十分であると考えられる．

XAI はモデルの内部動作の解明を目指しているのは先述した通りだが，従来の XAI 手法は特定のデータポイントを対象としてその特徴の重要度の抽出を主にしてきた．例えばこれまで XAI の代表的な手法であった LIME[2]や SHAP[3]はモデルの形状に因らぬ説明を可能としてきたが，特定のデータポイントに対しての説明，つまり局所的特徴重要度(Local feature importance)

の出力を行うことは可能であっても，モデル全体の説明である大局的特徴重要度(Global feature importance)の出力には限界があった．しかし，機械学習の挙動を把握するというのは，特定のデータポイントに注目するのではなく，データポイントの集合，本研究であればクラス単位のデータポイントでの特徴重要度を把握するべきであると考えられる．

我々はこれまで，Approximate Inverse Model Explanations(AIME)[4]を提案している．具体的な手法については 3 節で取り上げるが，AIME は対象とする機械学習モデルの近似逆作用素を構築することにより大局的・局所的説明の導出を可能とする．

本研究の目的は AIME による大局的特徴重要度を考えることにより，欠損値がもたらすモデルへの影響を評価することである．評価のために本研究では AIOD(近似逆作用素差:Approximate Inverse Operator Difference)という 2 つのモデルの挙動を比較するための指標を提案し，不完全データに対して複数種類の欠損値処理を行ったデータで学習したモデルを評価する．これにより欠損値処理によるモデル内部の影響を評価することが可能になり，モデルの挙動理解に貢献する．

2. 関連研究

本節では本研究に関連する研究を紹介する．

2.1 節では，モデルの挙動理解に関する研究について紹介する．2.2 節では欠損値処理に関する研究について紹介する．2.3 節では関連研究を踏まえ，本研究の位置付けを行う．

2.1. モデルの挙動理解に関する研究

機械学習モデル単体ではそのモデル内部の動作の把握が難しいため，しばしば機械学習・深層学習モデルは「ブラックボックス」と表現される．モデルの動作理解のために，これまで XAI の研究が進展してきており，SHAP や LIME, AIME, その他にも Grad-CAM[14]や PDP[15]，Occlusion[16]などが提案されてきた．

Speith[17]は XAI をモデルの出力結果から説明を獲得することを考える Post-hoc 手法とモデル自体を解釈可能なモデル，線形モデルや決定木などを用いる Ante-hoc 手法の二つに大別されるとしており，その上で，Post-hoc 手法自体を対象となるモデル形状に依存しない model-agnostic と依存する model-specific の二つに分類した．SHAP や LIME, そして本研究で用いる AIME は model-agnostic 手法に分類され，CNN を対象としている Grad-CAM などは model-specific 手法に分類される．

一方 Ante-hoc 手法においては Sagi ら[18]は XGBoost[19]のような GDBT モデルを解釈可能な決定木に，精度を落とすことなく，変換する手法を提案している．

Lu ら[20]は知識蒸留を用いて，モデルの知識を解釈可能な決定木に蒸留する手法を提案している．

モデルの説明性・透明性担保のために XAI の利用は重要であると考えられるが，Ante-hoc 手法によるモデルは精度や表現力の損失を招くなど課題は未だ多く，より汎用的な XAI の利用には Post-hoc かつ model-agnostic な XAI の普及が重要であると考えられる．

2.2. 欠損値処理に関する研究

El-Masri ら[21] は不完全データに対する欠損値処理を初心者の研究者を対象に紹介するものであり，ここではリストワイズ削除, ペアワイズ削除, 平均値補完, 回帰代入，推定最大化を紹介している．

Netten ら[22]は耳鼻咽頭科における多くの研究では不完全データに対する欠損値処理が不適切である点を指摘しており，またこれにより出力結果がバイアスを含んだものになる点や統計的処理に影響が出る点を問題として挙げ，その上で多重代入法[23]による欠損値処理を紹介している．

Satty ら[24]は縦断的研究において得られたデータはドロップアウトや欠損値の増加が避けられず，データ分析において深刻なバイアスを引き起こす可能性があるとし，多重代入法や線形混合モデルなどを用いた欠損値処理を比較している．

不完全データに対する欠損値処理はモデルの精度に対して大きな影響を及ぼす．特に臨床試験などの重大な責任の伴う状況下ではより一層，欠損値処理は適切に行わなければいけない．現に欠損値処理に関する研究は臨床研究で扱う不完全データをベースにした研究が多く[20][25][26]，欠損値処理の重要性が伺える．

2.3. 本研究の位置付け

不完全データは，何かしらの原因でデータの収集ができず欠損値が発生して起こるが，データ分析や機械学習モデルによる学習では基本的にこの欠損値に別の値を補完，あるいはリストワイズ削除やペアワイズ削除のように行や列を削除し，あくまで完全データとして扱わなければならない．しかし Netten ら[20]が問題として挙げているようにデータの専門性やデータ分析，予測の責任の重大性が増すにつれて欠損値処理は難しくなる事で，不適切な欠損値処理が増加してしまうリスクがある．

欠損値処理によるモデルの影響はこれまで多くの研究がなされてきたが，多くは欠損値処理手法の提案をメインとし，その後のモデル評価は精度をメインとしてきた．しかし精度重視の評価はモデルの内部動作を無

視する可能性が高い．XAI の重要性として Adadi ら[5]はモデルの脆弱性の解消やモデルの改善への利用を挙げているが，これらを行うには入力や出力と同時にモデルの内部動作についての理解は不可欠であると考えられる．

本研究の位置付けとして，不完全データや欠損値処理を行って構築された対象となる機械学習モデル(ブラックボックスモデル)が欠損値処理の違いによりモデル内部にどのような影響を受けるのかについて明らかにするものである．本研究ではモデルの内部挙動を定量化するために過去に我々が提案した XAI の AIME を用いて実現することにより可能になる．

3. AIME

3.1. 概要

本節では本研究で評価の中心となる AIME を紹介する．

先述の通り， AIME は LIME と SHAP と同様に XAI における Post-hoc 手法であり，モデルの形状に依存しない model-agnostic 手法に分類される．

AIME はモデルの入力から出力を順問題として捉えるのに対してモデルの説明を逆問題として捉え直すことで，XAI の代表的な手法である SHAP や LIME に比べて説明性に富んだ出力が可能になる．また AIME は Local feature importance(局所的特徴重要度)，Global feature importance(大局的特徴重要度)，Representative estimation instance(代表推定インスタンス)，Representative instance similarity distribution plot(代表インスタンス類似度分布プロット)の 4 機能によって構成される．

3.2. 手法

あるモデルfが入力Xから出力$\hat{Y}$を行う時，この一連の流れを AIME では順問題として捉え，これを以下のように表す．

$$\hat{Y} = f(X)$$

ここでモデルfの逆作用素f^{-1}について考える．AIME では逆作用素を考えることにより，モデルの逆問題を$\hat{Y}$とXから以下のように表すことが可能になる．

$$X = f^{-1}(\hat{Y})$$

逆作用素f^{-1}は解析的に求めることができない．一方，AIME では，逆作用素f^{-1}を線形近似したものを行列$A^{\dagger}$(近似逆作用素)として以下のように表す．

$$X = A^{\dagger}\hat{Y}$$

行列$A^{\dagger}$は行列Aのムーア・ペンローズの擬似逆行列(一般化逆行列)[26][27]を意味している．ムーア・ペンローズの擬似逆行列は正則ではない行列に対して擬似的な逆行列を求めるものである．

近似逆作用素は以下のように導出される．

$$X = A^{\dagger}\hat{Y}$$
$$X\hat{Y}^{\mathsf{T}} = A^{\dagger}\hat{Y}\hat{Y}^{\mathsf{T}}$$
$$X\hat{Y}^{\mathsf{T}}(\hat{Y}\hat{Y}^{\mathsf{T}})^{-1} = A^{\dagger}(\hat{Y}\hat{Y}^{\mathsf{T}})(\hat{Y}\hat{Y}^{\mathsf{T}})^{-1}$$
$$A^{\dagger} = X\hat{Y}^{\mathsf{T}}(\hat{Y}\hat{Y}^{\mathsf{T}})^{-1} = X\hat{Y}^{\dagger}$$

ムーアペンローズの擬似逆行列において

$$\hat{Y}^{\mathsf{T}}(\hat{Y}\hat{Y}^{\mathsf{T}})^{-1} = \hat{Y}^{\dagger}$$

が成立することに気をつけていただきたい．

AIME では，局所的特徴重要度が特定のデータポイントにおける説明をするのであれば，大局的特徴重要度はモデルのクラス単位の説明を行う．この大局的特徴重要度は行列$A^{\dagger}$の行ベクトルがそれを表しており，これは解明しようとするモデル全体の内部挙動の説明であると考えることが可能である．

3.3. 本研究における AIME の利用について

本研究では欠損値処理がモデルの内部に与える影響を評価するが，そのためにはモデル全体を説明する大域的な説明が必要であると考えられる．本節では本研究において AIME を利用する理由を述べる．

SHAP や LIME はこれまで数多くの研究において利用されてきた．しかし，この 2 つと AIME が異なる点の一つに大局的特徴重要度の直接的な獲得がある．AIME では近似逆作用素自体がモデルの大局的特徴重要度を表すことが可能である．一方 SHAP は大局的特徴重要度の獲得は一応可能であるとされてはいるが，区間推定である点やあくまでクラス単位での局所的特徴重要度の平均を考えているなど，大局的特徴重要度の獲得には限界があると考えられる．一方で LIME は SHAP 同様に大局的特徴重要度を考慮する際は局所的特徴重要度の平均を考えなければいけない点や，アルゴリズムの関係上，入力されたデータポイントに対してそれに近いデータポイント群から，線形モデルを近似するため，本研究のようなモデル全体の影響把握のためにはあまり適切な説明を得ることができないと考えられる．また AIME は SHAP や LIME に比べ，多重共線性に頑健であり，特徴量とラベル間の関係性に敏感であるという点も理由である．

4. 提案指標

本節では Approximate Inverse Model Explanations (AIME)を用いた機械学習モデル間の挙動評価手法のための指標である AIOD(近似逆作用素差:Approximate Inverse Operator Difference)について説明する．

4.1. 概要

AIME における大局的特徴重要度はそのクラスを出力するためにどの特徴量がどれだけの寄与をモデルに

与えたのかを説明することが可能である．一方でこれまでモデル間の評価はモデルの出力から得られる精度などを考慮したものであり，モデルの内部動作においてどのように変化を生じるかは，これまでの XAI が局所的特徴重要度に集中してきた点から言っても，評価することができなかったと考えられる．本指標は AIME が実現可能にしたクラス単位での特徴重要度の出力を用い，欠損値処理によって特徴重要度がどのように異なるのかを表す指標である．これによりモデル内部の動作の比較が可能となる．

4.2. AIOD : Approximate Inverse Operator Difference

3 節でも述べたように AIME における大局的特徴重要度はモデルの逆演算子f^{-1}を線形近似した$A^{\dagger}$によって表現可能である．$A^{\dagger}$は入力とモデルの予測値の一般化逆行列によって表現されるため，AIME は機械学習モデルの形状によらない Post-hoc(事後的)な手法である．

AIME における大局的特徴重要度は近似逆作用素$A^{\dagger}$そのものであり，各クラスの特徴重要度は近似逆作用素$A^{\dagger}$の行ベクトルである．

AIOD では 2 つのモデルの大局的特徴重要度を軸としてモデル間の差を考えることを目的とする.その時，大局的特徴重要度の各要素間の値の差だけではなく，出力結果の大きさの順序についても考慮する．大局的特徴重要度はモデルの各クラス単位での特徴重要度を説明するが，順序を考えることでモデルがどの特徴量を優先的に見ているかという点も重要であると考えるためである.具体的には重要度の各要素の差を考えて，L1 ノルムを求める.また順序についてはスピアマンの相関係数を距離尺度に直した値を考え，AIOD は L1 ノルムとスピアマンの相関係数の加重平均で求められる.

以下，機械学習モデル A と機械学習モデル B を考える．モデル A の大局的特徴重要度の行列を$A_A^{\dagger}$とし，モデル B の大局的特徴重要度の行列を$A_B^{\dagger}$とする．入力Xが N 種類の特徴量をもち，M クラス分類問題である時，$A_A^{\dagger} \in \mathbb{R}^{M\times N}$かつ$A_B^{\dagger} \in \mathbb{R}^{M\times N}$が成立する．またクラス$i$の重要度を表す重要度ベクトルを$v_A^{(i)}$, $v_B^{(i)}$とする．したがって$A_A^{\dagger}$, $A_B^{\dagger}$は重要度ベクトルを用いて以下のように表される．

$$A_A^{\dagger} = (v_A^{(1)}, v_A^{(2)}, v_A^{(3)}, \dots, v_A^{(M-1)}, v_A^{(M)})^{\intercal}$$
$$A_B^{\dagger} = (v_B^{(1)}, v_B^{(2)}, v_B^{(3)}, \dots, v_B^{(M-1)}, v_B^{(M)})^{\intercal}$$

スピアマンの相関係数は初めにそれぞれの重要度ベクトルの要素の相対的順序を求める．クラスiのj番目の特徴量の相対順序を$d_j \in \mathbb{N}$とし，特徴量数をnとするとクラスiのスピアマンの相関係数p_iは以下のように求められる．

$$p_i = 1 - \frac{6\sum {d_j}^2}{n(n^2-1)}$$

L1 ノルムDはベクトルの要素間の絶対差の合計である．Dが大きくなる時は重要度に大きな差が発生していることを指す．クラスiの重要度の L1 ノルムD_iは以下のように表現される．

$$D_i = \left\| v_A^{(i)} - v_B^{(i)} \right\|_1$$

L1 ノルムを用いた理由として要素間の距離計算が可能なためである．同じようにユークリッド距離があるが AIME では大局的特徴重要度は 0 から 1 の間に正規化されており，ユークリッド距離では二乗するため差が生まれにくくなってしまう．

クラスiのスピアマンの相関係数と L1 ノルムの和をPD_iとする時，加重平均のためのハイパーパラメータαとして，クラスiの AIOD 値は以下のように表現される．

$$PD_i = \alpha D_i + (1-\alpha)(1-p_i)$$

$1-p_i$とする理由は相関係数を距離尺度に変換するためである．$1-p_i = 0$の時は$p_i = 1$であり，ベクトルの順序が完全に一致している．一方$1-p_i = 1$の場合は順序間に相関がないことであり，ランダムであることを指す．

AIOD は二つのモデル間の挙動比較とクラス毎の挙動比較を行うことが可能である．クラス毎の挙動比較ではPD_iであるが，モデル間の挙動比較では$\sum_i^M PD_i$と表記できる．

以上から AIOD は以下のように表すことができる．

$$AIOD_{model}(\alpha) = \frac{1}{M}\sum_{i}^{M} PD_i$$

$$AIOD_{class}(\alpha) = PD_i$$

AIOD が大きな値を取る時，つまり同クラスあるいはモデル間の大局的特徴重要度に大きな差があることを意味し，モデルの挙動が異なる可能性を示す．小さい時は似た挙動である可能性を示す．

5. 評価実験

本節では AIOD を用いて，欠損値処理を行ったデータと完全データで学習したモデルの挙動を評価する．

5.1. 評価環境

実験をするためのプログラミング言語は Python[29]である．表データは Wine Quality データセット[30]を用いた．Wine Quality データセットは 6,497 行，12 列からなる，ワインの品質を 6 クラスの説明変数とした分類タスクのためのデータセットであり，欠損値は存在しない完全データである．また特徴量はアルコール度などのワインの栄養を数値化したデータであり，全

て量的データである．今回は欠損値全てがランダムに発生した欠損値であると前提して，欠測率 10%の不完全データを作成するために乱数を用いて Wine Quality データセットに欠損値を与える．本実験では 100 通りのデータセットを作成し，実験を行い，結果は 100 回分の実験の平均値を結果として用いる．

欠損値処理はすべて scikit-learn[31]のメソッドを用いており，リストワイズ削除，その欠損値がある列の平均値補完，多重代入法(Multiple Imputation)，K 近傍による補完を行う．

多重代入法では代入の施行回数 M をハイパーパラメータとして考える必要があるが，高橋ら[32]の研究では欠測率が 20%未満の場合は M=20 とするべきであるとされており，本実験もこれに従い M=20 とする．K 近傍による補完では多重代入法と同様に代入に使用する隣接サンプル数 n をハイパーパラメータとして考える必要がある．今回 K 近傍による補完を行う際に Python の機械学習ライブラリ scikit-learn の impute.KNNImputer メソッドを用いているがデフォルトでは n=5 であるため本研究もこれに従う．

学習データと訓練データに分割する際には scikit-learn の train_test_split メソッドを使用する．

AIOD ではクラスiのスピアマンの相関係数と L1 ノルムの加重平均を考えるが，本実験では加重平均の係数のハイパーパラメータ α を 0.5 とする．

また本研究では機械学習モデルにサポートベクタマシン(SVM)[33]を用いる．SVM には scikit-learn の SVC メソッドを用いる．

5.2. 評価手法

本実験では欠損値処理手法がモデルの内部動作にどのような影響を及ぼすかを AIME の大局的特徴重要度を用いて評価する．大局的特徴重要度は AIME においては，特定のクラスの出力においてどの特徴量が寄与したかを説明することが可能である．本実験では AIME と 4 節で提案した，モデル間の内部動作を比較できる指標 AIOD を用いて欠損値処理によるモデルの内部動作の影響を評価する．

具体的には完全データと欠損値処理した 4 種類のデータセットそれぞれで学習した 5 種類の機械学習モデルを用いて，それぞれのクラスの大局的特徴重要度を AIME によって求める．その結果から完全データを学習したモデルとそれ以外のモデルの大局的特徴重要度を AIOD によって比較し，欠損値処理によるモデルの内部動作を確認する．

5.3. 評価結果

100 回の試行を行い，各クラスの AIOD 値の平均値と処理手法毎の AIOD 値の平均値と正解率の平均値などの実験結果を表 1 に示し，$AIOD_{class}$を棒グラフにした結果を図 1 に示す．

完全データで学習したモデルの精度は 55.71%だったが，欠損値処理を行ったデータで学習した正解率は全て完全データで学習したモデルの精度を下回った．特にリストワイズ削除の場合は 51.21%にまで下落した．他の 3 種類の手法は 54%台を記録した．

またリストワイズ削除の$AIOD_{model}$は他の 3 手法に比べて大きいため，リストワイズ削除によるデータは完全データで構成されたモデルの挙動とは異なり，その上汎化性能も落ちることが考えられる．

しかし欠測率 10%でリストワイズ削除を行うと学習データの約 8 割が消えたため，欠測率 1 割の場合を試した．学習データ数は平均 4,354 行となった．その結果$AIOD_{model}$は 1.1107，正解率は 55.71%となった．

この結果から欠測率が高い場合，リストワイズ削除を行うと明らかにモデルの精度や動作には悪影響を及ぼすが，欠測率が小さい場合にリストワイズ削除を行うと他の補完手法より精度があがるが AIOD 値は他の処理変わらないことがわかった．これにより不適切な補完はデータのノイズとなり，モデルの挙動が完全データの時と比べ，異なる挙動になる可能性がある．

また多重代入と K 近傍は精度がほとんど変わらないが，$AIOD_{model}$は多重代入の方が小さく，多重代入法が完全データで学習したモデルの挙動に近い結果を示すことが明らかになり，この結果から多重代入法が K 近傍に比べ，優秀であると言える．

6. 考察

本節では 5 節で行った評価実験結果に対する考察を示す．評価実験では欠測率 10%の Wine Quality データセットに対して欠損値処理を行い，複数のモデルを構築し，それぞれのモデルの挙動評価を AIOD によって評価した．結果として多重代入法は他の処理手法に比べ，モデルの挙動の変化を比較的抑えることがわかった．また AIDO が正解率だけではわからない，モデル挙動の違いを比較することができることを確認した．欠測率 10%でのリストワイズ削除はモデル構築時に完全データに比べて学習データが極めて少ない．リストワイズ削除の結果，学習データは 641 に減っており，データの量の変化が行列$A^{\dagger}$の近似に影響を大きくもたらした可能性が高い．またリストワイズ削除はデータを不均一にする可能性や意図せぬバイアスの導入などを招く可能性が高く，モデルそのものを大きく変容させたため AIOD 値が総じて大きくなったと考えられる．

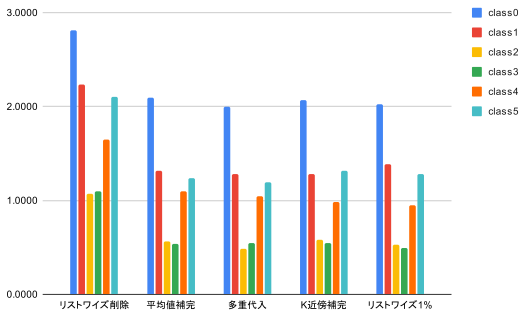


図 1. 各処理によって作成されたモデルの AIOD 値

表 1. 評価実験結果

欠損値処理手法	class	$AIOD_{class}$	$AIOD_{model}$	Acc(%)
リストワイズ削除（欠測率 10%）	0	2.8133	1.8276	51.21
	1	2.2345		
	2	1.0677		
	3	1.0986		
	4	1.6457		
	5	2.1057		
平均値補完	0	2.0983	1.1421	54.30
	1	1.3206		
	2	0.5648		
	3	0.5386		
	4	1.0955		
	5	1.2349		
多重代入	0	2.0022	1.0922	54.38
	1	1.2813		
	2	0.4891		
	3	0.5421		
	4	1.0430		
	5	1.1956		
K 近傍	0	2.0649	1.1269	54.39
	1	1.2778		
	2	0.5806		
	3	0.5441		
	4	0.9805		
	5	1.3136		
リストワイズ削除（欠測率 1%）	0	2.0245	1.1107	55.71
	1	1.3833		
	2	0.5283		
	3	0.4956		
	4	0.9473		
	5	1.2850		

K 近傍による補完は精度においては多重代入とほぼ同じだが，多重代入に比べ，完全データで学習したモデルに対して異なる挙動を示すことがわかった．K 近傍による補完は近隣のデータポイントに大きく依存し，ノイズの影響を受けやすい．そのため，今回のような至る部分で欠損値が存在するデータには向いていないため，AIOD がやや大きくなった可能性が考えられる．

また，図 1 からわかるように class0 と 5 は全体的に AIOD が高くなりやすく，class2 と 3 は低い．これは class0 と 5 は完全データで学習したモデルと欠損値処理をしたモデルの挙動の差が大きいことを意味しており，2 と 3 は差が小さいことを意味する．

本実験では多重代入が$AIOD_{model}$が最も低い結果となったが，特に class2 と class5 が他の手法に比べ，小さいことがわかる．多重代入は他の手法に比べ，他の特徴量から欠損値補完をするため単一の値で代入するような平均値補完に比べると元のデータに対して近い特徴量になると考えられる．しかし class4 においては K 近傍補完と欠測率 1%でのリストワイズ削除の方が AIOD 値は小さい．考えられる可能性として class4 においてクラスと特徴量の間に何らかの関係性がある可能性が考えられる．これにより欠損値補完がノイズになってしまうため，リストワイズ削除の AIOD 値が小さかったのではないかと考えられる．

また全ての処理結果において，class2，3 が他の値に比べ，小さい傾向がある．欠損値処理自体が真のデータに対して少なくともノイズになってしまうことは今回の結果から明らかであるが，これはクラス単位でノイズに対して頑健なクラスと脆弱なクラスが存在することを示唆していると考えられる．

今回の実験はあくまで完全データに対して欠損値処理を行い，学習データとテストデータに分割をし，正解率を算出する．しかし，仮にこの後，仮に実際に獲得された欠損値のないデータが挿入された時，それらの正解率が一番高いのは AIOD 値が最も小さい多重代入法で処理したデータを学習したモデルであると考えられる．テストデータはあくまで欠損値処理後のデータの分布から得られるものであり，完全データの分布からは多少なりとも異なる．そのため，今回の結果で言えば正解率が最も良い欠測率 1%でのリストワイズ削除処理を行なったデータで学習したモデルが，新たなデータに対して他の手法に比べて優れた正解率を記録するのではなく，むしろ AIOD 値が一番低い多重代入法が最も優れた正解率を記録すると考えられる．AIOD 値が小さいということは完全データを学習したモデルに近づくことを意味するためである．

7. おわりに

本稿では不完全データを用いて学習した機械学習モデルに対して AIME による大局的特徴重要度と我々が提案したモデルの内部挙動を比較する指標 AIOD を用いて，モデルの内部動作がどのようになるのかについて評価した．本実験では表データに対して乱数で欠損値を与え， 4 種類の欠損値処理を行い，完全データと 4 つの不完全データを学習させた 5 種類のモデルを構築し，AIME を用いて，欠損値処理がモデルの内部動作に与える影響を評価した．

今後の展望としては表データだけでなく画像や文章データでのモデルの挙動評価や，SVC だけでなく他の機械学習・深層学習モデルへの応用，AIOD の適切なハイパーパラメータの検証，またクラス単位での影響力の評価手法の検討などをしたいと考えている．

参 考 文 献

[1] OpenAI. (2023). ChatGPT [Large language model]. https://chat.openai.com/chat

[2] Ribeiro, Marco Tulio, Sameer Singh, and. Carlos

Guestrin. "" Why should I trust you?"Explainingt he predictions of any classifier."Proceedings of th e 22nd ACM SIGKDD international conferenceon knowledge discovery and data mining. 2016.

[3] Lundberg, Scott M., and Su-In Lee. "A unified. approach to interpreting model predictions." Adan ces in neural information processing systems 30 (2017).

[4] T. Nakanishi, "Approximate Inverse Model Expla nations (AIME): Unveiling Local and Global Insi ghts in Machine Learning Models,"in IEEEAccess, vol. 11, pp. 101020-101044, 2023,doi:10.1109/A CCESS.2023.3314336.

[5]. A. Adadi and M. Berrada, "Peeking Inside the B lack-Box: A Survey on Explainable Artificial Inte lligence (XAI)," in IEEE Access, vol. 6, pp.5213 8-52160, 2018, doi:10.1109/ACCESS.2018.2870052.

[6] Goodman, B. and Flaxman, S. (2017), European Union Regulations on Algorithmic Decision Maki ng and a “Right to Explanation”. AI Magazine, 3 8: 50-57.https://doi.org/10.1609/aimag.v38i3.2741

[7] Shadbahr, T., Roberts, M., Stanczuk, J. et al. T he impact of imputation quality on machine learn ing classifiers for datasets with missing values. Commun Med 3, 139 (2023).https://doi.org/10.103 8/s43856-023-00356-z

[8] Huang, Min-Wei & Tsai, Chih-Fong & Tsui, Shu -Ching & Lin, Wei-Chao. (2023).Combining data discretization and missing value I. imputation for incomplete medical datasets. PloS one. 18. e029 5032.10.1371/journal.pone.0295032.

[9] Manar D. Samad, Sakib Abrar, Norou Diawara, Missing value estimation using clustering and dee p learning within multiple imputation framework, Knowledge-Based Systems,Volume 249,2022,10896 8,ISSN 0950-7051,

[10] Zheng Zhang, Xiaoming Yan, Liyong Zhang,Xiao chen Lai, Wei Lu,Fuzzy neuron modeling of inco mplete data for missing value imputation, Inform ation Sciences, Volume 659,2024,1200 65,ISSN 0 020-0255

[11] Gondara, L., & Wang, K. (2017). MIDA: Multip le Imputation Using Denoising Autoencoders. , 2 60-272. https://doi.org/10.1007/978-3-319-93040-4_ 21.

[12] Gravesteijn, B., Sewalt, C., Venema, E., Nieboe r,D., & Steyerberg, E. (2021). Missing data in pr ediction research: A five step approach for multi ple imputation, illustrated in the CENTER-TBI st udy.. Journal of neurotrauma. https://doi.org/10.10 89/neu.2020.7218.

[13] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missin g value imputation: a review and analysis of the literature (2006–2017). Artif. Intell. Rev. 53, 2 (Feb 2020), 1487–1509. https://doi.org/10.1007/s10 462-019-09709-4

[14] Selvaraju, Ramprasaath R., et al. "Grad-cam:Vis ual explanations from deep networks via gradient-based localization." Proceedings of the IEEE inter national conference on computer vision. 2017.

[15] Friedman, Jerome H. “Greedy Function Approxi mation: A Gradient Boosting Machine.” TheAnnal s of Statistics, vol. 29, no. 5, 2001, pp. 1189–23 2. JSTOR, http://www.jstor.org/stable/2699986. Ac cessed 4 Jan. 2024

[16] Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks,in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision –ECCV 2014, Lecture Notes in Computer. Science. Springer International Publishi ng, Cham,pp. 818–833. https://doi.org/10.1007/978 -3-319-10590-1_53.

[17] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Method s. In Proceedings of the 2022 ACM Conference o n Fairness, Accountability, and Transparency (FAc cT '22). Association for Computing Machinery, N ew York, NY, USA, 2239–2250. https://doi.org/10. 1145/3531146.3534639

[18] Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. Inf. Sci., 572, 522-542. https://doi.org/10.1016/J.INS.2 021.05.055.

[19] Chen, T., & Guestrin, C. (2016). XGBoost: A S calable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conferenc eon Knowledge Discovery and Data Mining (KD D'16). ACM, New York, NY, USA.

[20] Lu, X., & Lee, J. (2022). Distillation Decision Tree. ArXiv, abs/2206.04661..

[21] El-Masri, Maher & Fox-Wasylyshyn, Susan.(200 6). Missing data: An introductory. conceptual ove rview for the novice researcher. The Canadian jo urnal of nursing research = Revue canadienne de recherche en sciences infirmières.37. 156-71.

[22] Netten, A., Dekker, F., Rieffe, C., Soede, W.,Bri aire, J., & Frijns, J. (2017). Missing Data. in the Field of Otorhinolaryngology and Head & Neck S urgery: Need for Improvement. Ear and Hearing,3 8, 1–6. https://doi.org/10.1097/AUD.000000000000 0346.

[23] Rubin, Donald B.. “MULTIPLE. IMPUTATIONSI N SAMPLE SURVEYS-A PHENOMENOLOGICAL BAYESIAN APPROACH TO NONRESPONSE.” (2002).

[24] Satty, A., Mwambi, H., & Molenberghs, G. (201 7). Statistical Methodologies for Dealing with Inc omplete Longitudinal Outcomes Due to Dropout Missing at Random. , 179-209. https://doi.org/10. 1007/978-981-10-3307-0_10.

[25] Bell, M., & Fairclough, D. (2014). Practical. an d statistical issues in missing data for longitudin al patient-reported outcomes. Statistical Methods i n Medical Research, 23, 440 - 459. https://doi.or g/10.1177/0962280213476378.

[26] National Research Council (US) Panel on. Handl ing Missing Data in Clinical Trials. The Preventi on and Treatment of Missing Data in Clinical Tri als. National Academies Press (US), 2010. doi:10. 17226/12955

[27] Moore, E. H. (1920). On the reciprocal of the g eneral algebraic matrix. Bulletin of the American Mathematical Society, 26(9), 394-395.

[28] Penrose, R. (1955). A generalized inverse for m atrices. Mathematical Proceedings of the Cambrid ge Philosophical Society, 51(3), 406-413.

[29] Python Core Team. (2023). Python: A. dynamic, open source programming language. Python Softw are Foundation. URL: https://www.python.org/

[30] Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T.,

and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

[31] Scikit-learn Developers. (2023). Scikit-learn: Machine Learning in Python. Scikit-learn version 1.2.2. URL: https://scikit-learn.org/

[32] 高橋将宜，伊藤孝之：”様々な多重代入法の比較~大規模経済系データを用いた分析~，統計研究彙法，Vol.71，pp39-82，(2014)”

[33] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

文章間類似性判定モデルの根拠の抽出

中井香那子† 河田 友香†† 山本 岳洋††

† 兵庫県立大学 社会情報科学部 〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
†† 兵庫県立大学 大学院情報科学研究科 〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †fa20h053@guh.u-hyogo.ac.jp, ††ad22l019@guh.u-hyogo.ac.jp,t.yamamoto@sis.u-hyogo.ac.jp

あらまし 本研究では，2 つの文の類似性を判定し,「類似」「非類似」という判定の根拠となる文節の抽出に取り組む． まず，類似度が与えられた 2 つの文章ペアのデータに，類似度に応じて「類似」または「非類似」というラベルを付与し,「類似」「非類似」を判定する BERT モデルを作成した．その後，各文を文節に分割し，文節の類似度が高い順に各文から 1 文節ずつを取り出した文節ペアを作成した．最後に，文節ペアをそれぞれマスクした状態で学習した BERT に入力し，類似に分類される確率をマスク前と比較することにより，変化の大きい文節ペアを判定に影響を与える根拠として抽出した．その結果，類似文からは類似度の高い文節を共通語として，非類似文からは類似度は高くないものの対応関係を持つ文節を差異として抽出することができた．また，誤分類に影響を与えた文節を抽出することにより，誤って学習している部分や学習が不十分であった部分を発見した．

キーワード XAI，テキスト分類，BERT

1 はじめに

インターネットの発展により，身の回りには情報があふれている．情報が簡単に手に入ることにより，文章を比較する機会やある文に関連する文章を探す機会は増加したと考えられる．例えば，オンラインショッピングでは 2 つの商品情報を比較し，商品の違う部分を見て購入する商品を決定する．具体的にイヤホンを購入する場面を考える．イヤホンの商品情報には音質やバッテリー，デザインといったカテゴリの情報が存在する．商品を検討する際は，まず同じカテゴリの情報を比較し，それぞれの商品の特徴を手に入れる．例えば音質について商品 A は「圧巻の**重低音**を再生する新規設計ドライバー．完全ワイヤレスで重要な装着感を追求しながら，圧巻の**重低音**を再生する大口径ドライバーを新規開発．高い密閉性により**低音**を逃しません.」，商品 B は「**高音域から低音域までバランスの良い**クリアな音質．**高音から低音までバランスが良く**，特にボーカルや楽器の音を自然でクリアにお楽しみいただけます.」という記載があったとする．このような説明文から商品 A は重低音重視であるのに対し，商品 B は高音から低音までのバランス重視であるという違いを手に入れる．また，低音重視といった手に入れた商品情報に関連するレビューを探すことで，より詳細な情報や他の人の感想を手に入れることもある．そして，手に入れた情報を総合して購入する商品を最終的に決定する．

以上のように，類似，関連している文章を比較する場面は日常生活で多くあり，類似性判定を行うモデルが作成されている．しかし，ディープラーニングを用いて類似性判定のモデルを作成すると人間には内部でどのような処理が行われているか分かりづらいという問題がある.「モデルからなぜこのような出力結果が得られるのかわからない」というブラックボックスモデルは問題視されており，BERT [6] では Attention [5] の可視化や LIME [13] が根拠の可視化として用いられている．こういったモデルの判断の根拠は，モデルの出力結果を理解しやすくなるだけでなく，モデルの性能の評価，モデルの精度の向上に使うことができる．例えば，正解ラベルと予測ラベルが一致しない場合，その根拠を抽出することでうまく学習が行えていない部分を発見し，性能の改善につなげることができる．また，正解ラベルと予測ラベルが一致している場合でも，正しい判断基準から判定を行えているか確認ができる．

さらに，根拠を可視化することができれば情報の比較の手助けになる．2 文の類似性判定においてモデルが人間と同様に，似ている単語の組み合わせに注目して類似と判定，似ていない単語の組み合わせに注目して非類似と判定しているならば，類似の根拠は 2 つの文の共通点，非類似の根拠は 2 つの文の相違点となる．そのような根拠が可視化できれば，商品を比較する際，長い文章をすべて読まずとも，2 つの商品の文章から一目で各カテゴリにおける共通点と相違点を確認できる．

既存研究では，2 つのラベルの分類においてそれぞれの分類の根拠の抽出を行うことで，各ラベルの特徴語を抽出する研究 [20] や，文書集合間の共通語，特徴語集合を可視化する研究 [21] が行われている．しかし，単語を抽出する際,「綺麗」と「汚い」のような各文の単語の対応関係には注目していない．

本研究では，類似度が与えられている 2 つの文章ペアのデータセットを用いて，類似性判定を行う BERT モデルを作成し，そのモデルの判定の根拠となる単語の抽出を行う．単語を抽出する際に，各文の文節の類似度を比較し，文節ペアを用いることにより，各文の単語の対応関係に注目した根拠の抽出を行った．同時に，モデルの性能の評価として誤分類の根拠を抽出し，モデルが上手く学習できていない部分を発見した．こういった根拠の抽出を商品の説明文に応用し，根拠を可視化することができれば，図 1 のように文章から共通点，相違点を探しやすくできるよう支援が行える．

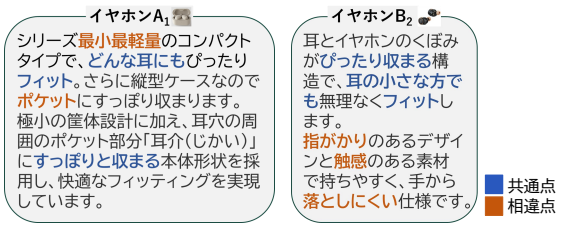


図 1　商品説明文における根拠の可視化の例.

2　関 連 研 究

2.1　類似性判定

単語や文章の類似性判定を利用するタスクとして，クエリと文書の適合性判定や同義語判定，類似文判定などが挙げられる．

クエリと文書の適合性判定のタスクの研究として，杉木らは商品を検索するクエリと商品レビューからそれぞれ要求と意見を抽出し，要求に合う商品の提示を行うシステムを開発している [17]．また，重要語に注目したレビューの要約の研究として，神谷らは，PageRank アルゴリズムを利用して抽出した重要語からクエリを生成し，クエリを用いて要約文を生成する方法を提案している [16].

同義語判定のタスクとして，ある単語に対して前後の文をもとに適切な意味を判定するというような語義曖昧性解消問題がある．語義曖昧性解消の研究として，谷田部らは事前学習済み BERT を用いて用例文ペアにおける単語の同義判定を行っている [19]．さらに，同義関係を表したグラフと，グラフニューラルネットワークにより語義を判別するモデルの構築を行っている [18]．Mikolov らは単語のベクトル化の手法として Word2Vec という手法を提案している [10]．Pilehvar らは単語の意味の類似性を測定するために，単語の意味に対する確率的表現を使用した方法を提案している [11].

類似文判定タスクの研究として，Batanović らはフレーズと文や，文と段落といった異なるレベルのテキスト間の類似度をはかる研究を行っている [1]．データセットにはセルビア語の新聞記事を用いて作成したデータセットを使用し， ボスニア語，クロアチア語，モンテネグロ語，セルビア語を用いて事前学習を行った BERTić モデルや，104 個の異なる言語を用いて事前学習を行った多言語 BERT モデルを利用して類似度をはかっている．また，Reimers らは 2 つの入力文の埋め込み表現を比較する BERT をファインチューニングしたモデルとして Sentence-BERT [12] を提案している．本研究では，2 つの文章の類似，非類似の判定を行う方法として BERT を，文節間のコサイン類似度をはかる方法として Sentence-BERT を用いる．

2.2　根拠や特徴語の抽出

XAI が注目されている現在，画像やテキストデータなど様々な分野で根拠推定の研究が行われている．

Lundberg らはゲーム理論のシャープレイ値を用いてモデルの解釈を行う SHAP という手法を提案している [9]．Ribeiro らは入力データの一部を隠すことによる出力の変化からモデルの解釈を行う LIME という手法を提案している [13]．Chen らはニューラルネットワークの解釈可能性を向上させるため，テキスト分類の階層的説明を生成する手法 [3] や，解釈可能性に加え，モデルの予測精度を向上させるため，単語マスク層を介してタスク固有の重要な単語を学習する VMASK という手法を提案している [2]．Kokalj らは BERT を含む Transformer モデルに対して SHAP を適応させる TransSHAP という手法を提案している [7]．Yan らは分類ラベルに関連付けられたトピックを分類の根拠として自動生成する分類器 HIMT を提案している [14]．Lei らはテキスト分類の根拠抽出として，ジェネレータとエンコーダを用いて入力テキストの一部を根拠として抽出するモデルの学習を行っている [8].

テキストデータに対し LIME を用いて特徴語を抽出した研究として，小野川らのホテルのレビューを分類するモデルについて特徴語を抽出した研究が挙げられる [20]．コンフォートホテルとアパホテルのレビューを用いて，LSTM によりどちらのホテルのレビューかを分類した．そして分類結果の理由となる単語を LIME により単語を抽出したところ，各ホテルのレビューにおける単語出現頻度の差を表すカイ 2 乗値の上位の単語と一致した．さらに，カイ 2 乗値では抽出できなかったホテルの違いを表す特徴語を SP-LIME により抽出することができている．

また，Choudhary らは自然言語処理のブラックボックスモデルの解釈性についての調査を行っている [4]．局所的解釈可能性アプローチと，大局的解釈可能性アプローチに分けて分析している．局所的解釈可能性アプローチとは特定の入力に対する予測の根拠を示すことにより説明を行う方法である．一方で，大域的解釈可能性アプローチは複雑なモデルの全体を処理の簡単なわかりやすいモデルで表現することにより説明を行う方法である．それぞれの長所と短所を分析し，状況に応じて使い分ける必要があると結論付けている．

また，文書間の差異を表す特徴語の抽出として薦田らは，文書集合間の共通語集合と特有語集合を可視化することによる理解支援の研究を行っている [21]．この研究では特有語を，他の文書集合から差別し，同時に関連性を示すような単語と定義し，対数尤度比とコサイン類似度を用いた文脈類似度によって抽出している．

本研究では，類似度の高い文節をペアとしてマスクすることによる類似に分類される確率の変化を見ることで，対応関係を持つ文節を類似性判定の根拠として抽出した．

3　文節ペアのマスクによる類似性判定モデルの根拠の抽出

本節では類似非類似の根拠となる文節の抽出手法について述べる．例文 A「スケートボードでジャンプしている男性がいます.」と例文 B「男性がスケートボードでジャンプしています.」

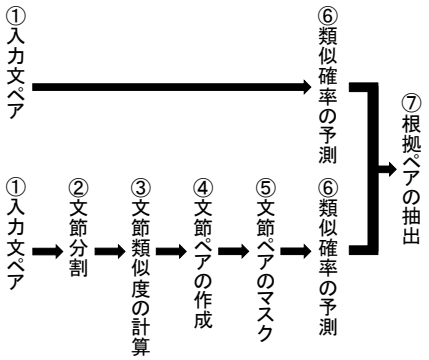


図 2　提案手法の概要.

という 2 文を用いて説明を行う.

3.1　提案手法の概要

提案手法の概要を図 2 に示し，以下で概要を説明する．図 2 の類似確率の予測を行うモデルに対し，入力文ペアをそのまま入力して得られた確率と，一部をマスクして入力して得られた確率を比較することで根拠を抽出する．以下に手順を示す.

（1） 入力文ペア

- 類似または非類似ラベルをつけた 2 文のペアを入力する.

（2） 分割文節

- 入力文をそれぞれ文節単位で分割する.

（3） 文節類似度の計算

- 各文から 1 つずつ文節を取り出し Sentence-BERT を用いて類似度を算出する.

（4） 文節ペアの作成

- 文節間の類似度の高い順に文節ペアを作成する.

（5） 文節ペアのマスク

- 文節ペアをマスクした状態で BERT に入力する.

（6） 類似確率の予測

- 予測分類確率を出力する BERT モデルの学習を行う.
- 予測分類確率とは類似に分類される確率と非類似に分類される確率である.
- 入力文ペアを類似性判定を行う BERT に入力する.

（7） 根拠ペアの抽出

- マスク前後で予測分類確率を比較する.
- 予測分類確率が逆転した文節ペアを根拠として抽出する.

また，文節ペアのマスク前後の予測分類確率の変化による根拠の抽出について，比較方法を図 3 に示す．図 3 では，例文の「スケートボードで」と「スケートボードで」という文節ペアをマスクする前と後の予測分類確率を比較している．マスクする前は類似に分類される確率が 0.9 であるのに対し，マスクした後は類似に分類される確率が 0.2 まで下がったとする．この場合，予測分類確率が逆転しているため,「スケートボードで」と「スケートボードで」という文節ペアは類似という判定に影響を与える根拠として抽出することができる．反対に,「男性が」と「男性が」という文節ペアをマスクすると，類似に分類される確率が 0.8 であったとする．この場合，マスク前も類似に分類される確率が 0.8 であり予測分類確率は変化していないため,「男性が」と「男性が」という文節ペアは類似という判定に影響を与えないと判断し，根拠として抽出しない.

3.2　類似性判定モデルの学習

本研究では BERT を用いて入力文ペアが類似または非類似である確率を予測するモデルを作成する．BERT とは双方向 Transformer Encoder によって構成される自然言語処理モデルである．ファインチューニングにより，分類や質問応答といった様々なタスクにおいて高精度な処理を行うことができる.

BERT には類似または非類似というラベルがつけられたペアとなっている文章を [SEP] トークンでつなぎ，2 文を同時に入力する．また，2 文を同時に入力するため，1 文目と 2 文目のトークンを判別するために Segment ID を用いる．Segment ID は 1 文目に 0，2 文目に 1 を割り当てる．出力は予測分類確率である．予測分類確率は [CLS] トークンのベクトルをシグモイド関数により変換することで算出した．入力と出力の例を図 4 に示す.

3.3　文節ペアの作成

本研究では，根拠の抽出の際文節ペアを利用する．2 文の類似性判定を行うと，類似文には共通するような単語，非類似文には差異をあらわす単語というような対応関係を持つ単語が 2 文に存在すると考えられる．そこでペアを作成することにより対応する単語を根拠として同時に抽出することができる．また，単語単位での分割は助詞や助動詞などそれ単体では意味が理解しにくい単語が抽出される可能性があるため，意味を持つまとまりとして文節を 1 つの単位とする．文節の類似度を用いて文節同士のペアを作成する．このペアは入力文ペアの 1 文目から 1 文節，2 文目から 1 文節を選択したものであり，文節間の類似度の高いものからペアを作成する.

3.3.1　入力文ペアの文節分割

BERT に入力する 2 文をそれぞれ GiNZA により文節に分割する．例では文 A が「スケートボードで/ジャンプしている/男性が/います.」，文 B が「男性が/スケートボードで/ジャンプしています.」というように分割される.

3.3.2　文節の類似度比較

各文節を Sentence-BERT に入力し，文節間の類似度を求める．Sentence-BERT とは，2 つの文章をそれぞれ同一の BERT に入力して各文章のベクトルを生成し，生成したベクトルを用いて文章の関係性を判定する手法である．各文章についてベクトルを生成するため，大量の文章がある場合，高精度かつ高速に類似度を求めることができるという特徴がある．本研究では，文章ではなく文節のベクトルを生成し，類似度を算出する．具体的には，図 5 のように各文から文節を 1 つずつ選択して Sentence-BERT に入力し，コサイン類似度を算出する．まず，図 5 のように各文の 1 文節目を入力する．次に，文 A の 1 文節目の「スケートボードで」と文 B の 2 文節目の「スケートボードで」を入力するというように全ての文節の組み合わせでコサイン類似度を求める.

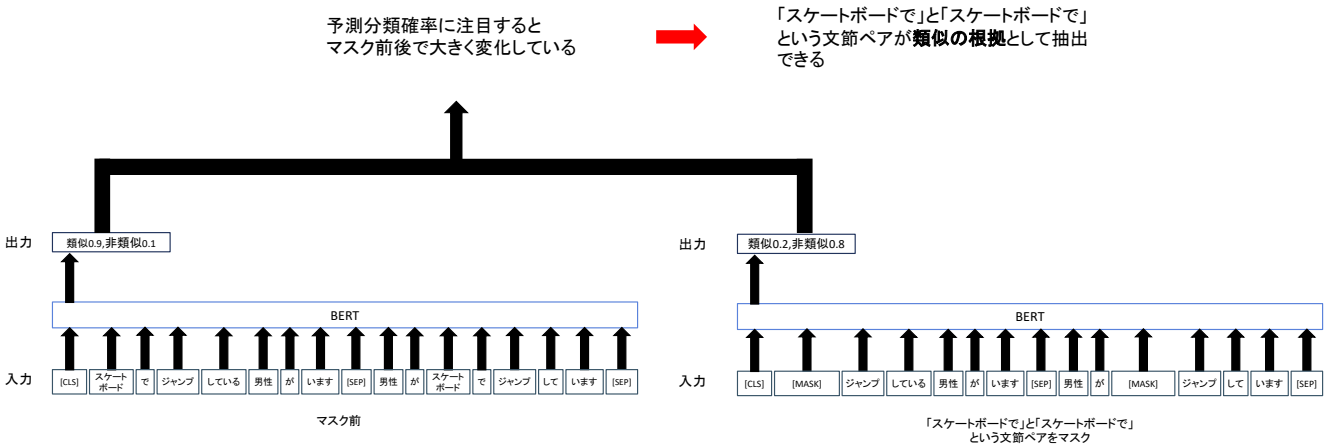


図 3　マスク前後の予測分類確率の変化による根拠の抽出.

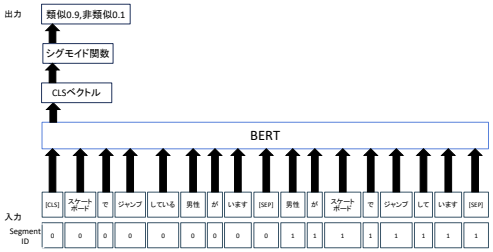


図 4　分類確率を予測する BERT モデルの入力と出力.

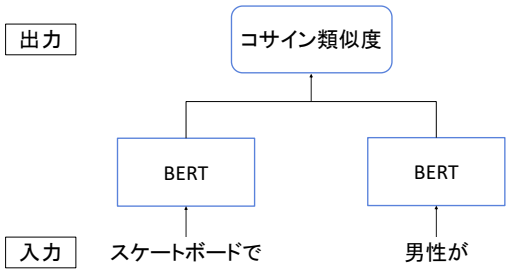


図 5　Sentence-BERT による文節の類似度比較.

3.3.3　文節ペアの作成

コサイン類似度の高いものから順番に，文節が重複しないよう対応付けを行い，文節のペアを作成する．仮に，例文における文節の類似度が高い上位 3 件が類似度の高い順に「スケートボードで」と「スケートボードで」,「スケートボードで」と「ジャンプしています.」,「ジャンプしている」と「ジャンプしています.」だったとする．この場合 1 つ目の文節ペアは 1 番類似度の高い「スケートボードで」と「スケートボードで」になる．次に類似度の高いペアは「スケートボードで」と「ジャンプしています.」だが,「スケートボードで」が 1 つ目の文節ペアに含まれるため文節ペアからは省略し，2 つ目の文節ペアとしては「ジャンプしている」と「ジャンプしています.」を選択する．

例文での文節ペアは実際には類似度の高い順に「スケートボードで」と「スケートボードで」,「男性が」と「男性が」,「ジャンプしている」と「ジャンプしています.」となる．

また，文節ペアは文節数の少ない文章の文節数に合わせ，ペアを作ることのできない文節については対応がないものとする．例では文 A が 4 文節，文 B が 3 文節に分割できる．文節数の少ない文章の文節数に合わせるため文節ペアは 3 つとなり，1 文目の「います.」という文節を対応なしとして扱う．

3.4　根拠の抽出

3.2 節で学習した類似性判定モデルに文節ペアをマスクした状態で 2 文を入力し予測分類確率の変化を観察する．例えば，例文のうち「ジャンプしている」と「ジャンプしています.」という文節ペアをマスクする場合，入力は「[CLS] スケートボードで [MASK] 男性がいます．[SEP] 男性がスケートボードで [MASK][SEP]」となる．そして，何もマスクしていない状態の予測分類確率と比較し，予測分類確率が逆転した場合，文節ペアを判定に影響を与える根拠として抽出する．また，対応なしとして扱う文節についても，その文節のみをマスクする．そして，マスク前後で予測分類確率が逆転した場合は根拠として抽出する．例文では文 A の「います.」が対応なしとなったため,「います.」のみをマスクし,「[CLS] スケートボードでジャンプしている男性が [MASK][SEP] 男性がスケートボードでジャンプしています．[SEP]」という形で入力する．

4　実　　験

4.1　デ　ー　タ

研究に用いたデータについての説明を行う．早稲田大学とヤフー株式会社が構築・公開している JGLUE [15] の JSTS データセットを使用する．JGLUE とは日本語の言語理解モデル用

表 1　JSTS データセット.

類似度	文 A	文 B
0.0	ジャンプ台からスキーヤーがジャンプをしています.	茶色い猫の顔が，あおむけになっています.
1.0	男性が腕を組んでパソコンを見ています.	1 人の女性がベンチに座って本を読んでいます.
2.0	海上で女の子がサーフィンをしています.	男の子が海で遊んでいます.
3.0	柵で囲まれた芝生と道路が見えます.	柵で囲まれた池と道路が見えます.
4.0	男性が子供を抱き上げて立っています.	坊主頭の男性が子供を抱いて立っています.
5.0	真っ赤な二階建てのバスが停まっています.	赤い二階建てのバスが止まっています.

のベンチマークであり，中でも JSTS は文ペアの類似度推定タスク用に構築されている．具体的には，JSTS は意味が完全に異なるものを 0，意味が等価であるものを 5 として 0〜5 までの数字が 0.2 単位で類似度として付与されている．データセットの一部を表 1 に示す．類似度は文 A と文 B の類似度を表す．本研究では，類似度が 3.8 以上 5.0 以下であるデータに類似，類似度が 2.0 以上 2.6 以下までのデータに非類似とラベルを付ける．実際に文の比較を行う際，例えばイヤホンの購入を考えている場面では，デザインに関する 2 つの文や音質に関する 2 つの文など関連性を持つ 2 つの文を比較する．しかし，JSTS データの類似度 0.0 のデータは「ジャンプ台からスキーヤーがジャンプをしています.」と「茶色い猫の顔が，仰向けになっています.」のように関連性を持たない組み合わせとなっている．一方で類似度 2.0 のデータは「海上で女の子がサーフィンをしています.」と「男の子が海で遊んでいます.」のように意味は全く違うが，海という点で関連性を持っている．そこで，非類似のなかでも 2.0 未満のペアについては 2 つの文章間に関連性がないものとし，意味は似ていないが関連性を持つ 2.0 以上 2.6 以下のデータに「非類似」ラベルを付与する．また，類似度 4.0 や 5.0 のデータは「男性が子供を抱き上げて立っています.」と「坊主頭の男性が子供を抱いています.」のように文は一致しないものの一部言葉を付け加えたり，言い換えたりした意味が類似しているペアとなっている．そこで，4.0 以上のデータに「類似」ラベルを付与しようとしたところ，データ数が少なかったため，類似度の幅を広く 3.8 以上 5.0 以下と設定した．データ数は訓練データ 3,200 件，検証データ 400 件，テストデータ 400 件の計 4,000 件である．

また，根拠を抽出するにあたって，分析の対象はテストデータのうち未知語を含まない 2 文の組み合わせのみを利用した．

4.2 実験設定

類似性判定モデルの作成には東北大学の学習済み BERT モデル [1]を用いた．また，学習率は 2.0×10^{-5}，patience は 3，バッチサイズは 16，エポック数は 10，入力の最大長は 512 と設定した．また，文節の類似度の測定には日本語用の事前学習済み Sentence-BERT モデル [2]を用いた．

4.3 比較手法

類似，非類似の根拠の抽出の比較手法として LIME を用いる．LIME とはまず学習したモデルに文章の一部をマスクしたものを入力する．次に，一部をマスクした入力とその入力に対して得られた類似または非類似のラベルをペアとして線形回帰を行う．そして，線形回帰の結果から出力に大きな影響を与えた単語を特徴語として抽出する．比較を行う際には，回帰に使用する一部をマスクした入力とそれに対する出力のペアデータの数を出力結果が安定する 5,000 件とし，上位 10 件の単語を抽出した．また可視化の際には，類似の根拠を青色，非類似の根拠をオレンジ色で示す．

5 結　　果

5.1 類似文ペアを入力した際の根拠の抽出例

正解ラベルが「類似」であり，予測分類確率も「類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「非類似」によるような文節ペアを「類似」の根拠として抽出した．

具体例として,「スケートボードでジャンプしている男性がいます.」と「男性がスケートボードでジャンプしています.」という類似文について抽出した結果を示す．この 2 文はそれぞれ「スケートボードで/ジャンプしている/男性が/います.」「男性が/スケートボードで/ジャンプしています.」という文節に分割される．さらに，各文節の類似度を測り，類似度の高いものから文節ペアを作ると「スケートボードで」と「スケートボードで」,「男性が」と「男性が」,「ジャンプしている」と「ジャンプしています.」となった．また，1 文目が 4 文節，2 文目が 3 文節に分割される．文節数の少ない文章の文節数に合わせるため，文節ペア数は 3 つとなり，1 文目の「います.」という文節は対応なしとして扱う．マスクをしていない元の文と，各文節ペアをマスクした文の予測分類確率は表 2 のようになった．表は上からマスク前，文節類似度の高い順にペアのマスク後，対応なしとなった文節のマスク後の順である．

表 2 より「スケートボードで」や「男性が」といった文節は 2 つの文に共通する文節だが，判定には大きな影響を与えていないことが分かる．一方で,「ジャンプしている」と「ジャンプしています.」という文節ペアは，予測分類確率が逆転しており，類似という判定に大きな影響を与えていることが分かる．また，対応のなかった「います.」は 1 文目のこの文節のみをマスクしたところ，予測分類確率にはほとんど変化がなかったことから，判定に大きな影響を与える文節ではないといえる．

これらの結果より，この 2 文において類似の判定の根拠は「ジャンプしている」と「ジャンプしています.」という文節ペアだといえる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結

1：https://github.com/cl-tohoku/bert-base-japanese-v3

2：https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

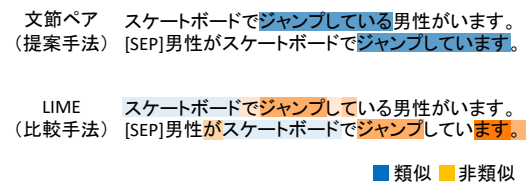


図 6　類似文から抽出された根拠の比較.

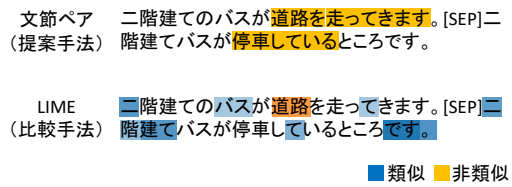


図 7　非類似文から抽出された根拠の比較.

果と提案手法による根拠を抽出した結果の比較を図 6 に示す.「スケートボード」という 2 つの文両方に出現する単語が類似の根拠として抽出できた一方で,「ジャンプ」という共通語は非類似の根拠として抽出された．また，非類似の根拠となる単語が多く抽出されている点や，句読点や助詞，助動詞が抽出されている点から，類似の根拠となる共通部分はわかりづらい結果となっている.

5.2　非類似文ペアを入力した際の根拠の抽出例

正解ラベルが「非類似」であり，予測分類確率も「非類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「類似」によるような文節ペアを「非類似」の根拠として抽出した.

具体例として,「二階建てのバスが道路を走ってきます.」と「二階建てバスが停車しているところです.」という非類似文について抽出した結果を示す．2 文はそれぞれ「二階建ての/バスが/道路を/走って/きます.」「二階建てバスが/停車している/ところです.」という文節に分割される．さらに，類似度の高いものから文節ペアを作ると「バスが」と「二階建てバスが」,「きます.」と「ところです.」,「走って」と「停車している」となる．また，1 文目は 5 文節，2 文目は 3 文節に分割できる．文節数の少ない文章の文節数に合わせるため文節ペアは 3 つとなり，1 文目の「二階建ての」「道路を」という文節を対応なしとして扱う．マスク前と各文節をマスクした予測分類確率の結果を表 3 に示す.

表 3 より「バスが」と「二階建てバスが」という文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「きます」と「ところです」という文節ペアは予測分類確率が逆転しないものの，15 %程度類似に分類される確率が上がっており，非類似という判定に影響を与えていると考えられる．さらに,「走って」と「停車している」という文節ペアは予測分類確率が逆転しており，判定に大きな影響を与える非類似の根拠といえる．この 2 文節の特徴として，文節間の類似度は 0.3411 となっており他の文節ペアに比べると類似度が低いことが挙げられる．ただし,「走る」「停車する」というバスの動きを表す語として関連性は持っているといえる.

対応関係のない文節については 1 文目の 1 文節のみをマスクして比較を行った．その結果,「二階建ての」は予測分類確率に変化が見られなかったが,「道路を」は予測分類確率が逆転した.「二階建ての」は文節としては対応がないが，2 文目の「二階建てバス」という文節に「二階建て」という語が含まれるため判定に影響を与えないと考えられる.

これらの結果より，類似度は低いものの対応関係を持つ「走って」「停車している」という文節ペアと片方のみに出現する「道路を」という文節が非類似文の差異として適切に抽出できたといえる.

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 7 に示す. LIME では「道路」という 1 文目にのみ出現する語が非類似の根拠として抽出された一方で,「走る」と「停車」という対になっている語は抽出できなかった．また,「二階建て」という語の一部である「二」のみが抽出されたり,「して」や「です」などその単語単体で意味の通じない単語が抽出されたりし，非類似の根拠としてわかりづらい結果となった.

5.3　モデルが誤分類を行った際の根拠の抽出例

正解ラベルは「類似」だが予測分類確率は「非類似」である確率が高くなったり，正解ラベルは「非類似」だが予測分類確率は「類似」である確率が高くなったりしたような誤分類においても，正しく類似，非類似と判定された文章と同様に文節ペアを作成し，予測分類確率をマスク前後で比較した．そして，予測分類確率がマスク前後で逆転している文節ペアを誤った分類へ大きな影響を与えたものとして，学習が不足していた部分や間違って学習していた部分の発見につなげた.

まず，正解ラベルが「類似」であったが，予測分類確率は「非類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「類似」によるような文節ペアを，誤って非類似と判定した根拠として抽出した．その結果，2 つの文の差異となる部分が非類似という判定に影響を与えていることが分かった.

具体例として「白のジャンボジェット機が青空を飛んでいます.」「青い空を，飛行機が飛んでいます.」という 2 つの文の抽出結果を示す．2 文はそれぞれ「白の/ジャンボジェット機が/青空を/飛んでいます.」「青い/空を，/飛行機が/飛んでいます.」という文節に分割された．さらに，類似度の高い順に「飛んでいます.」と「飛んでいます.」,「ジャンボジェット機が」と「飛行機が」,「青空を」と「空を,」,「白の」と「青い」という文節ペアになった．マスク前と各文節をマスクした予測分類確率の結果を表 4 に示す．表 4 より「飛んでいます.」と「飛んでいます.」や「青空を」と「空を,」といった文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「白の」と「青い」という文節ペアは予測分類確率が逆転しており，非類似という

表 2　類似文のマスクと予測分類確率.

文 A	文 B	類似	非類似
スケートボードでジャンプしている男性がいます.	男性がスケートボードでジャンプしています.	0.9997	0.0002
[MASK] ジャンプしている男性がいます.	男性が [MASK] ジャンプしています.	0.9960	0.0039
スケートボードでジャンプしている [MASK] います.	[MASK] スケートボードでジャンプしています.	0.9238	0.0769
スケートボードで [MASK] 男性がいます.	男性がスケートボードで [MASK]	0.0004	0.9995
スケートボードでジャンプしている男性が [MASK]	男性がスケートボードでジャンプしています.	0.9991	0.0008

表 3　非類似文のマスクと予測分類確率.

文 A	文 B	類似	非類似
二階建てのバスが道路を走ってきます.	二階建てバスが停車しているところです.	0.0898	0.9101
二階建ての [MASK] 道路を走ってきます.	[MASK] 停車しているところです.	0.0011	0.9988
二階建てのバスが道路を走って [MASK]	二階建てバスが停車している [MASK]	0.0011	0.7740
二階建てのバスが道路を [MASK] きます.	二階建てバスが [MASK] ところです.	0.6139	0.3860
[MASK] バスが道路を走ってきます.	二階建てバスが停車しているところです.	0.0007	0.9992
二階建てのバスが [MASK] 走ってきます.	二階建てバスが停車しているところです.	0.9655	0.0344

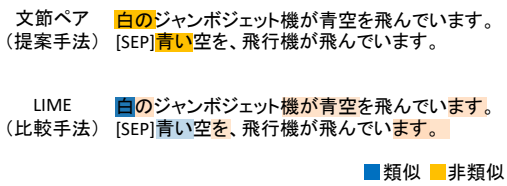


図 8　非類似を類似と判定した入力文ペアから抽出された根拠の比較.

誤分類に影響を与えた文節ペアだといえる．さらに,「白の」と「青い」は文節類似度が低く，それぞれを単体でマスクしたところ,「青い」は非類似に分類される確率が 0.9932 と他の文節ペアをマスクしたときと変化が見られなかったが,「白の」は非類似に分類される確率が 0.1590 と予測分類確率が逆転した．このことから，この 2 文の判定を行う際に,「白の」という 2 文の差異に注目しすぎた結果，非類似という誤分類につながったと考えられる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 8 に示す.「青い」という単語が類似の根拠として抽出された一方で「青空」は非類似の根拠として抽出されるなど対応関係のあるものが別々の根拠として示された．また,「白」という一文目のみに出現する単語が類似の根拠として抽出されたほか，上位 10 件のなかに付属語が多く見られた．

次に，正解ラベルが「非類似」であったが，予測分類確率は「類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「非類似」によるような文節ペアを，誤って類似と判定した根拠として抽出した．その結果，2 つの文に共通する部分が類似という判定に影響を与えていることが分かった．

具体例として「芝生の真ん中に，消火栓が設置されています.」「道路の真ん中に，消火栓が設置されています.」という 2 つの文の結果を示す．2 文はそれぞれ「芝生の/真ん中に，/消火栓が/設置されています.」「道路の/真ん中に，/消火栓が/設置されています.」という文節に分割された．さらに，類似度の高い順に「消火栓が」と「消火栓が」,「設置されています.」と「設置されています.」,「真ん中に,」と「真ん中に,」,「芝生の」と「道路の」という文節ペアになった．マスク前と各文節をマスクした後の予測分類確率の結果を表 5 に示す．表 5 より「消火栓が」と「消火栓が」や「芝生の」と「道路の」といった文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「設置されています.」と「設置されています.」や「真ん中に,」と「真ん中に,」という文節ペアは予測分類確率が逆転しており，類似という誤分類に影響を与えた文節ペアだといえる．このことから，この 2 文の判定を行う際に,「設置されています.」や「真ん中に,」といった 2 文に共通する文節に注目した結果，類似という誤分類につながったと考えられる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 9 に示す.「真ん中」や「消火栓」という 2 つの文両方に出現する単語が非類似の根拠として抽出された．また，全体としては類似と判定しているにもかかわらず，上位 10 件は非類似の根拠に偏った結果となった．

5.4　根拠が抽出できなかった例

文節ペアをマスクすることにより，根拠が上手く抽出できないものもあった．具体例として「綺麗に掃除されたトイレがあります.」「茶系統のタイル張りの部屋のトイレです.」という 2 つの文の抽出結果を示す．2 文はそれぞれ「綺麗に/掃除され

表 4　類似を非類似と判定した文のマスクと予測分類確率.

文 A	文 B	類似	非類似
白のジャンボジェット機が青空を飛んでいます.	青い空を，飛行機が飛んでいます.	0.0007	0.9992
白のジャンボジェット機が青空を [MASK]	青い空を，飛行機が [MASK]	0.0005	0.9995
白の [MASK] 青空を飛んでいます.	青い空を，[MASK] 飛んでいます.	0.0018	0.9981
白のジャンボジェット機が [MASK] を飛んでいます.	青い [MASK] 飛行機が飛んでいます.	0.0004	0.9996
[MASK] ジャンボジェット機が青空を飛んでいます.	[MASK] 空を，飛行機が飛んでいます.	0.9995	0.0005

表 5　非類似を類似と判定した文のマスクと予測分類確率.

文 A	文 B	類似	非類似
芝生の真ん中に，消火栓が設置されています.	道路の真ん中に，消火栓が設置されています.	0.9992	0.0007
芝生の真ん中に，[MASK] 設置されています.	道路の真ん中に，[MASK] 設置されています.	0.9615	0.0384
芝生の真ん中に，消火栓が [MASK]	道路の真ん中に，消火栓が [MASK]	0.2881	0.7118
芝生の [MASK] 消火栓が設置されています.	道路の [MASK] 消火栓が設置されています.	0.3990	0.6009
[MASK] 真ん中に，消火栓が設置されています.	[MASK] 真ん中に，消火栓が設置されています.	0.9993	0.0006

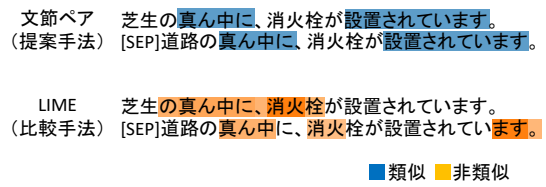


図 9　非類似を類似と判定した入力文ペアから抽出された根拠の比較.

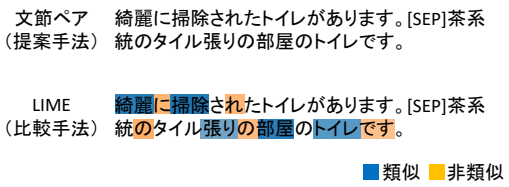


図 10　提案手法では根拠が抽出できなかった例.

た/トイレが/あります.」「茶系統の/タイル張りの/部屋の/トイレです.」という文節に分割された．さらに，類似度の高い順に「トイレが」と「トイレです.」,「あります.」と「部屋の」,「掃除された」と「タイル張りの」,「綺麗に」と「茶系統の」という文節ペアになった．マスク前と各文節をマスクした後の予測分類確率の結果を表 6 に示す．表 6 よりどの文節ペアでも予測分類確率に大きな変化はないことがわかる．これらの文節ペアの特徴として「トイレが」と「トイレです」を除く 3 ペアは類似度 0.4 を下回っており，2 つの文節間に関連がないことが挙げられる．このような結果から，文節同士が関連を持っていなければ文節ペアを根拠として抽出することは難しいといえる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果を図 10 に示す.「トイレ」という 2 文に共通する単語が類似の根拠で抽出された一方で,「掃除」や「タイル」など片方の文にしか出現しない単語が類似の根拠として抽出された．また，非類似の根拠はすべて「です」や「に」など助動詞や助詞となり 2 文の違いは分かりづらい結果となった．

6　まとめと今後の課題

本研究では，各文の文節を用いてマスクを行い，予測分類確率を比較することで，類似文からは共通する部分，非類似文からは差異を表す部分を根拠として抽出することに取り組んだ．同時に，誤分類の根拠を抽出することで，学習の不足部分を検討した．

比較手法である LIME と比べて，LIME では正確に抽出することのできなかった共通点や，対応関係を持つ相違点を抽出することができた．また，文章の分割の単位として単語ではなく文節を採用することにより，助詞や助動詞単体ではなく意味を持つ言葉のまとまりを抽出することができた．

一方で，LIME では類似文の共通語に加えて差異を表す語，非類似文の差異を表す語に加えて共通語も抽出できたが，提案手法による抽出結果は類似文からは共通する文節のみ，非類似文からは差異を表す文節のみとなった．今後，類似文中の差異を表す語，非類似文中の共通語の抽出の方法を検討する必要がある．

本研究の限界点として，文章同士の対応関係がなければ根拠をうまく抽出できないという点が挙げられる．具体的には，非類似と判定される文章のうち「綺麗に掃除されたトイレがあります.」と「茶系統のタイル張りの部屋のトイレです.」では「トイレが」と「トイレです」という文節ペアを除き，文節間の対応関係がなく，文節ペアをマスクすることによる予測分類確率の変化は見られなかった．

表 6　根拠を抽出できなかった文のマスクと予測分類確率.

文 A	文 B	類似	非類似
綺麗に掃除されたトイレがあります.	茶系統のタイル張りの部屋のトイレです.	0.0007	0.9993
綺麗に掃除された [MASK] あります.	茶系統のタイル張りの部屋の [MASK]	0.0007	0.9993
綺麗に掃除されたトイレが [MASK]	茶系統のタイル張りの [MASK] トイレです.	0.0006	0.9994
綺麗に [MASK] トイレがあります.	茶系統の [MASK] 部屋のトイレです.	0.0005	0.9995
[MASK] 掃除されたトイレがあります.	[MASK] タイル張りの部屋のトイレです.	0.0007	0.9993

また，ペアを作成するにあたって文節を用いたため，ある文では 1 文節でまとまるものが，もう一方では 2 文節に分かれてしまいうまく比較できないということがあった．例えば，非類似で例を挙げた「二階建てのバス」と「二階建てバス」は同じ意味を表す語だが，前者は「二階建ての/バス」の 2 文節，後者は「二階建てバス」の 1 文節である．その結果文節ペアは「バス」と「二階建てバス」となり,「二階建てのバス」と「二階建てバス」という 2 つの語全体の比較を行うことはできなかった．

上記 2 つの問題点より，今後の課題として，文節同士の類似度だけでなく，文節に含まれる単語の品詞の組み合わせなどからペアを決めるといった他のペアの作成方法を検討することが考えられる．

また，本研究では短い文を用いて根拠の抽出を行ったが，実際オンラインショッピングでのレビューや商品説明文を比較することを考えると一文がより長くなることや，一文同士の比較だけでなく，複数の文章を比較することが考えられる．本研究から，文の一部分を使うことにより適切に根拠を抽出することができることが分かったが，今後，より長い文の比較に対応するためには，文節よりも適した単位でのペアの作成方法を検討する必要性がある．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Vuk Batanović and Maja Miličević Petrović. Cross-Level Semantic Similarity for Serbian Newswire Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1691–1699, 2022.

[2] Hanjie Chen and Yangfeng Ji. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4236–4251, 2020.

[3] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5578–5593, 2020.

[4] Shivani Choudhary, Chatterjee Niladri, and Subir Kumar Saha. Interpretation of black box NLP models;a survey. *arXiv:2203.17081*, 2022.

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4171–4186, 2019.

[7] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 16–21, 2021.

[8] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016.

[9] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 4768–4777, 2017.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*, 2013.

[11] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1341–1351, 2013.

[12] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You":Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, 2016.

[14] Hanqi Yan, Lin Gui, and Yulan He. Hierarchical Interpretation of Neural Text Classification. *Computational Linguistics*, Vol. 48, No. 4, pp. 987–1020, 2022.

[15] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会発表論文集, pp. 2023–2028, 2022.

[16] 神谷賢太郎, 原寛紀, 青山幹雄. 深層学習によるレビュー内の重要語に着目した要約方法の提案と評価. 第 82 回全国大会講演論文集, pp. 463–464, 2020.

[17] 杉木健二, 松原茂樹. 消費者の意見に基づく商品検索. 情報処理学会論文誌, Vol. 49, No. 7, pp. 2598–2603, 2008.

[18] 梨恵谷田部, 稔佐々木. 用例文間の意味的な類似関係を用いた半教師あり語義曖昧性解消. 情報処理学会論文誌, Vol. 62, No. 10, pp. 1724–1736, 2021.

[19] 谷田部梨恵, 佐々木稔. BERT の学習済みモデルを用いた用例文ペアの同義判定. 言語処理学会 第 26 回年次大会 発表論文集, pp. 824–827, 2019.
[20] 小野川稜之, 折原良平, 清雄一, 田原康之, 大須賀昭彦. 機械学習モデルの解釈手法による競合サービスと比較したレビュー分析. 日本ソフトウェア科学会大会論文集, Vol. 36, pp. 337–343, 2019.
[21] 薦田和弘, 大澤幸生. 複数文書の相対的特徴可視化による理解支援. 人工知能学会第二種研究会資料, pp. 41–46, 2013.

ゼロショットエンティティリンキングに寄与するエンティティ記述の分析

瀬谷　浩樹†　　加藤　　誠††

† 筑波大学大学院 情報学学位プログラム　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s2221653@u.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　我々は，ゼロショットエンティティリンキング（ZSEL）に有効なエンティティ記述を作成する際の指針を得ることを目的として，エンティティ記述内のどの部分が ZSEL の正解率向上に寄与しているかを明らかにするための分析を行った．具体的な手法としては，メンションが言及しているエンティティ（つまり正解エンティティ）のエンティティ記述内の一部をマスクあるいは置換して，その結果 ZSEL の正解率がどのように変化するかを確認した．分析の結果，エンティティ記述内に出現するエンティティ間の関係，エンティティの出現回数，エイリアスの利用頻度が ZSEL の正解率に影響していることが示唆された．

キーワード　エンティティリンキング，ゼロショットエンティティリンキング，XAI

1 はじめに

質問応答 [1] や情報検索 [2] 等の自然言語処理において，ユーザが入力した文章の内容を把握するための手法としてエンティティリンキング（以降 EL）がある．EL は，エンティティ（一意に識別可能な事物）の集合と，文章中のメンション（エンティティを言及している部分）が与えられたときに，メンションがどのエンティティを言及しているかを推定するタスクである．近年は EL の中でも，専門的なエンティティ等の訓練データが少ないエンティティに対しても有効なゼロショットエンティティリンキング（以降 ZSEL）が注目されている．ZSEL は，Logeswaran らによって提案された，エンティティは文章のみで表現される，訓練セット内のエンティティ集合とテストセット内のエンティティ集合が交わりを持たないという 2 つの制約の下で行われる EL である [3]．最初の制約を満たすために，ZSEL モデルは基本的に，エンティティとそれを表現する文章（エンティティ記述）の集合と，メンションとその周囲の文脈のみを入力として受け取り，入力されたメンションがどのエンティティを言及しているかを推定する．そして，このモデルを 2 つ目の制約下で訓練することにより，エンティティに関する構造化された情報・訓練データが得られない場合でも EL を行える ZSEL モデルを作成できる．ZSEL の重要性について，例えば Oberhauser らは，訓練データを作るためのアノテーションに高いコストがかかる医療ドメインでは，訓練なしでも EL 可能な ZSEL が有効であることを指摘している [4]．

以上のように従来の EL では対処できない場面にも有効な ZSEL であるが，Wu らによる BLINK など [5]，ZSEL モデルは既に複数提案されている．そして，中には ZSEL 結果の分析を行なっている研究もある．しかし，これらの研究は ZSEL において有効なエンティティ記述の分析は行なっておらず，現状ではどういった記述が ZSEL に有効かが不明となっている．また，XAI と呼ばれる分野において，機械学習モデルの出力に対して，入力のどの部分がその出力に寄与したかを可視化する研究が行われている．しかし，ZSEL という個別のタスクに注目して，モデルの出力に対して入力されたエンティティ記述のどの部分が寄与したかを分析する研究は行われていない．

そこで本研究では，ZSEL に有効な（ZSEL の正解率向上につながる）エンティティ記述を作成する際の指針を得ることを目的として，どのようなエンティティ記述が ZSEL に有効かを分析する．分析手法としては，メンションとエンティティ記述集合が与えられたときに，そのメンションが言及しているエンティティのエンティティ記述内の一部をマスクあるいは置換し，その結果 ZSEL の正解率がどのように変化するかを確認する．

本研究の実験では，ZSEL モデルとしては BLINK [5] を，エンティティ（記述）集合としては BLINK と共に公開されているエンティティカタログを，EL を行うメンション集合としては WikilinksNED Unseen-Mentions [6] を ZSEL の制約を満たすように改変したデータセットをそれぞれ使用して分析を行った．

上の実験から，ZSEL に有効なエンティティ記述を作成する際の指針として，以下のようなことが示唆された．

（1） エンティティ記述に，記述対象のエンティティと特定の関係にあるエンティティを含めると，代わりに他の多くの関係にあるエンティティを含めた時よりも ZSEL の正解率が向上する．また，特定のタイプに属するエンティティ記述に関しては，他にも重要な場合が多い関係が存在する．

（2） エンティティ記述に，記述対象のエンティティと同タイプのエンティティを説明するときによく使われるようなエンティティを含めると，使われることが少ないエンティティを含めた時よりも ZSEL の正解率が向上する．

（3） エンティティ記述内で，エンティティを参照するときに利用頻度が低いエイリアスを使うと，それ以外の利用頻度のエイリアスを使った時よりも ZSEL の正解率が低くなる場合が多い．なお，特定のタイプに属するエンティティ記述に関しては，上とは異なる傾向が見られる．

本論文の構成は以下の通りである．第 2 節では，本研究に関

連する研究である，エンティティリンキングや寄与部分の可視化に関する研究を紹介する．第 3 節では，本研究において分析の対象とする EL と ZSEL を定義した後に，どのようなエンティティ記述が ZSEL に有効かを分析する手法について説明する．第 4 節では，第 3 節で説明した手法を用いて行った分析の実験について述べる．第 5 節では，本研究の結論を述べる．

2 関 連 研 究

本節では，関連研究について説明する．具体的には，EL に関する研究と，機械学習モデルの出力に対して，入力のどの部分が寄与したかを可視化する研究について説明する．

2.1 エンティティリンキング

EL は典型的に，大量のエンティティの中から文章中のメンションが言及している可能性のあるエンティティを探索する「候補生成」と，候補エンティティを言及されている可能性が高い順に並び替える「ランキング」の 2 段階で行われる．候補生成には，Wikipedia 内のハイパーリンクを元に計算したメンションがエンティティを言及している確率や [7,8]，文書検索手法である BM25 などが用いられる [3]．ランキングには，ニューラルネットワークを用いて行なっている例が多く見られる [9–11]．

近年は，通常の EL よりも厳しい制約の下で行われる ZSEL が提案され [3]，ZSEL モデルも複数研究されている [5,12–14]．

これらの中には EL 結果の分析を行なっている研究もある．Wu らは，提案した ZSEL モデルの EL 結果を分析し，文脈に誤解を招く単語が含まれている場合は正しいエンティティを推定できないことがあると述べている [5]．また，Sun らは，訓練時とテスト時のドメインの差異に対処できる ZSEL モデルの訓練手法を提案している [12]．そして，両ドメインが同じメタドメインに属しており密接に関連している場合，提案手法を通してその関連性を利用し，より良い結果が得られると述べている．

これらの研究は ZSEL 結果を分析しているという点で本研究に類似しているが，ZSEL に有効なエンティティ記述の分析は行なっていない．本研究では，どのようなエンティティ記述が ZSEL に有効かを分析することで，ZSEL に有効なエンティティ記述の作成という，現在主流なモデルの改善とは異なる方法で ZSEL の正解率を上げられるようになることを期待する．

2.2 寄与部分の可視化

XAI という分野において，機械学習モデルから出力が得られたときに，入力のどの部分がその出力に寄与したのかを人間が解釈できるように可視化するという研究が行われている．

文章も対象とした代表的な可視化手法として，LIME [15] や SHAP [16] が挙げられる．エンティティ記述を作成する際の指針となるような，エンティティ記述内のどの部分が ZSEL の正解率向上に寄与しているかに関する知見を得るという本研究の目的を踏まえると，上の手法には 2 つの問題点がある．(1) あくまで元の複雑なモデルを局所的に近似した線形モデルを用いており，元のモデルにおける入力内の寄与部分が正しく求められない可能性がある．(2) 事前に定義された観点で入力内の寄与部分を分析する訳ではないので，得られた寄与部分に指針となり得る規則性を見つけられるとは限らない．よって本研究では，実際の ZSEL モデルに編集したエンティティ記述を入力し，その結果 ZSEL の正解率がどのように変化するかを確認する．

上のような広く利用可能な可視化手法の提案を行なった研究とは異なり，より狭いタスクに注目した研究も行われている．例えば，Li らは自然言語処理タスクにおいて，ニューラルネットワークに入力された文章内の単語や単語ベクトルの一部を削除し，入力のどの部分が出力に寄与したかを分析する手法を提案している [17]．当該研究の手法は本研究で用いる分析手法と類似しているが，当該研究ではあくまで個別の単語それぞれの重要性を求めており，同じ実験を ZSEL を対象に行ったとしても，得られた分析結果からエンティティ記述を作成する際の指針となるような規則性を見つけることができるとは限らない．

3 分 析 手 法

本節では，どのようなエンティティ記述が ZSEL に有効かを分析するための手法について説明する．まず，分析の対象とする EL と ZSEL を定義する．そして，本研究では複数の観点から分析を行うため，観点間で共通している分析手法の概要を説明した後，個々の観点ごとの分析手法の詳細を説明していく．

3.1 エンティティリンキングの定義

EL は，エンティティ集合 E と，文章中においてエンティティ $e \in E$ を言及している部分（これをメンションと呼び，m で表す）が与えられたときに，m が E 内のどのエンティティを言及しているかを推定するタスクである．

ZSEL は，Logeswaran らによって提案された，2 つの制約の下で行われる EL である [3]．まず，エンティティに関するリソースとして与えられるのは，エンティティ e とその記述 d のペアの集合であるエンティティ辞書 $\mathcal{E} = \{(e_i, d_i)\}_{i=1,\ldots,K}$ のみである (K はエンティティ数)．つまり，エンティティに関する構造化された情報が保存されている知識グラフ（以降 KG）等があったとしても，それは ZSEL では使用できない．2 つ目として，訓練セット内のエンティティ集合 $\mathcal{E}_{\text{train}}$ とテストセット内のエンティティ集合 $\mathcal{E}_{\text{test}}$ は交わりを持たない ($\mathcal{E}_{\text{train}} \cap \mathcal{E}_{\text{test}} = \emptyset$)．

3.2 各観点間で共通な分析手法

本研究では，どのようなエンティティ記述が ZSEL に有効かを分析するということを，エンティティ記述内のどの部分が ZSEL の正解率向上に寄与しているかを明らかにすることと捉える．この目的に対して本研究では，メンションとエンティティ記述集合が与えられたときに，そのメンションが言及しているエンティティ（つまり正解エンティティ）のエンティティ記述の一部をマスクあるいは置換して，ZSEL の正解率の変化を確認するという手法を用いる．なお，本研究においてマスクするとは，特殊なトークンで置換することを指す．

以降，上述の手法をより詳細に説明する．マスク・置換対象 T を数式で表すと，次式のようになる：

$$T = \{e \in E_{d_c} \mid s_e \geqq \theta\} \quad (1)$$

T は，正解エンティティ c のエンティティ記述 d_c 内に出現しているエンティティの集合 E_{d_c} の各要素の内，対応するスコア s_e が閾値 θ 以上のエンティティである．具体的な s_e と θ は観点ごとに異なるため，次節以降で説明する．T のマスク・置換が終わったら，メンション集合 M とエンティティ集合 $\mathcal{E}$，マスク・置換済みの正解エンティティ記述を含むエンティティ記述集合を用い，各メンション $m \in M$ に対して ZSEL を行って正解率を求める．これを各観点ごとに定めた複数のパターンで行い，正解率が最も低かった（置換の場合は最も高かった）パターンのエンティティ記述が ZSEL の正解率向上に寄与していると考えられる．

3.3 観点 1：エンティティ間の関係

本節では，どのようなエンティティ記述が EL に有効かを，記述対象のエンティティと特定の関係にあるエンティティという観点で分析する手法について説明していく．

本観点で分析を行う理由は，次の通りである．最も有名なエンティティ記述集合の例の 1 つである Wikipedia がそうであるように，エンティティ記述は基本的に，記述対象であるエンティティと他のエンティティの間に存在する関係を記述する．よって，既述対象のエンティティとどのような関係にあるエンティティが ZSEL の正解率向上に寄与しているかを明らかにできれば，ZSEL に有効なエンティティ記述を作成する際の有用な指針となると考えられる．

これを明らかにするために，3.2 節の手法を用いて分析を行う．後述の各パターン間でマスクするトークン量が同一になるように正解エンティティと特定の関係にあるエンティティをマスクするために，まず式 1 内の s_e を次のようにする：

$$s_e =: g(c, e, r) \quad (2)$$

上式の g は 2 つのエンティティ c と e，関係 r を入力すると，両エンティティ間にその関係が存在する度合い（スコア）を出力する関数である．続いて，式 1 内の θ は，正解エンティティ集合 C 内の各要素とそのエンティティ記述内に出現しているエンティティ間に r が存在するスコアの集合 $S_r = \{g(c', e', r) \mid c' \in C, e' \in E_{d_{c'}}\}$ において，そのスコア以上のスコアを持つエンティティの総トークン数が正解エンティティ記述集合 $D_C = \{d_c \mid c \in C\}$ の総トークン数の p%以上となるようなスコアの内，最大のスコアとする．そして，本観点では，r として関係集合 R 内の各関係を用いたパターンの ZSEL の正解率を比較する．

以上が本観点での分析の詳細であるが，最後に，何を g として用いるかを述べる．1 つ目は KG である．なお，本手法における KG の使用は直接的に ZSEL の正解率を上げるためではなく，あくまでエンティティ記述内において記述対象のエンティティとどのような関係にあるエンティティが ZSEL の正解率向上に寄与しているかを明らかにするためであるので，ZSEL の制約に反することにはならない．さらに，KG には含まれていないが実際には 2 つのエンティティ間に存在する関係を推定するために，知識グラフ埋め込みモデル (KGE モデル) を用いる．

3.4 観点 2：エンティティの出現回数

本節では，どのようなエンティティ記述が EL に有効かを，エンティティ記述内に出現しているエンティティの，記述対象のエンティティと同タイプのエンティティ記述内での出現回数という観点で分析する手法について説明していく．

本観点で分析を行う理由は，次の通りである．3.3 節で述べたように，エンティティ記述は基本的に，記述対象であるエンティティと他のエンティティとの間に存在する関係を記述する．そして，エンティティによっては，複数のエンティティとの間に，ある同一の関係が存在する場合がある．このようなエンティティのエンティティ記述を作成する際に，その複数のエンティティの中からどれをエンティティ記述に含めるべきかという情報は，エンティティ記述を作成する際の有用な指針となると考えられる．そして，エンティティ記述に含めるべきエンティティを選ぶ際の基準として，本手法では記述対象のエンティティと同タイプのエンティティ記述内での出現頻度に着目する．その理由は，次の通りである．ZSEL モデルにおいて多く採用されるバイエンコーダは，一般的には類似エンティティの距離が近くなるようにエンティティをベクトル空間に埋め込む [12,18,19]．よって，同タイプのエンティティは ZSEL において正解エンティティと間違われる可能性が高く，それらを弁別するためにエンティティ記述内に出現するエンティティの差異が使われていると考えられるためである．

これを明らかにするために，3.2 節の手法を用いて分析を行う．後述の各パターン間でマスクするトークン量が同一になるように正解エンティティと同タイプのエンティティ記述内での出現回数に応じてエンティティをマスクするために，まず式 1 内の s_e を次のようにする：

$$s_e =: \mathrm{frq}(e, D_{t_c}) \quad (3)$$

上式の frq は，エンティティ e と，c が属するタイプ t_c とタイプが同じであるエンティティのエンティティ記述集合 $D_{t_c} = \{d_{e'} \mid e' \in \mathcal{E}, t_{e'} = t_c\}$ を入力すると，e の D_{t_c} 内での出現回数を出力する関数である．続いて，式 1 内の θ は，正解エンティティのエンティティ記述内に出現している各エンティティの D_{t_c} 内の出現回数の集合 $F_c = \{\mathrm{frq}(e', D_{t_c}) \mid e' \in E_{d_c}\}$ において，その出現回数以上の出現回数を持つエンティティの総トークン数がエンティティ記述 d_c の総トークン数の p%以上となるような出現回数の内，最大の出現回数とする．なお，人間のエンティティ記述には父親が出現する可能性があるが，映画の場合はその可能性がないなど，正解エンティティと特定の関係にあるエンティティが出現するかは正解エンティティ記述ごとにまちまちである．よって，3.3 節の手法においては正解エンティティと特定の関係にあるエンティティをより多くマスクするために全正解エンティティ記述を横断してマスク対象のエンティティを決めたが，本手法ではその必要はないため，各正解エンティティ記述ごとにマスク対象のエンティティを決める．そして，本観点では，出現回数が多いエンティティと少ないエンティティという 2 パターンのエンティティをマスクした際の ZSEL の正解率を比較する．これまで説明してきたのは出

現回数が多いエンティティをマスクする際の手法であり，少ないエンティティをマスクする際は反対の手法を用いる．

3.5 観点 3：エイリアスの利用頻度

本節では，どのようなエンティティ記述が EL に有効かを，エンティティ記述内に出現しているエンティティを表すエイリアス（別名）の利用頻度という観点で分析する手法について説明していく．

本観点で分析を行う理由は，次の通りである．これまでに述べたように，エンティティ記述はエンティティを含んでいる．そして，エンティティを含めるためには，当然のことながらその名前をエンティティ記述内に書く必要があるが，複数の名前，すなわちエイリアスを持っているエンティティが存在する．このようなエンティティは珍しくないため，そのエイリアスの中のどれをエンティティ記述内で使用すれば ZSEL の正解率が向上するかを明らかにできれば，ZSEL に有効なエンティティ記述を作成する際の有用な指針となると考えられる．そして，どのエイリアスを使うべきかの基準として，本手法ではエイリアスの利用頻度に着目する．これは，エイリアスが使われ始めた時期など他の基準もあり得るが，それらも結局のところ利用頻度と相関がある場合が多いと考えられるためである．

これを明らかにするために，3.2 節の手法を用いて分析を行う．なお，前の 2 つの観点ではエンティティ記述内のエンティティを置換して分析を行うことが難しくマスクで代替したが，本観点ではそれが可能であるため，より実際の状況に近くなるようにエンティティを置換して分析を行う．後述するパターンでの分析を行うために，利用頻度が異なる複数のエイリアスを持っているエンティティを置換対象 T とする．具体的には，式 1 内の s_e と θ を次のようにする：

$$s_e =: \#\{\mathrm{utl}(a_e) \mid a_e \in A_e\}, \quad \theta =: 2 \qquad (4)$$

上式の A_e はエンティティ e のエイリアス集合を指しており，utl は引数として渡されたエイリアスの利用頻度を出力する関数である．そして，本観点では，T 内の各エンティティを以下の 3 パターンの異なる利用頻度のエイリアスで置換した際の ZSEL の正解率を比較する．(1)A_e 内で最も利用頻度が高いエイリアス，(2)A_e 内で最も利用頻度が低いエイリアス，(3)A_e 内で利用頻度が中程度（$\mathrm{utl}(a_e)$ が中央値）のエイリアス．

4 実　　験

本節では，エンティティ記述内のどの部分が ZSEL の正解率向上に寄与しているかを明らかにするために行った分析の実験について説明する．まず，本実験で共通して使用したデータセットと EL モデルについて説明した後，3 節で述べた 3 つの観点による分析の実験についてそれぞれ述べる．

4.1 データセット

本研究では，WikilinksNED Unseen-Mentions [6] を，3.1 節で述べた ZSEL の制約を満たすように改変して得られたデータセット（以降，WikilinksNED UM ZS と呼ぶ）を使用した．改変元として WikilinksNED Unseen-Mentions を選んだ理由は，以下の通りである．当該データセットは，950 万件の Web ページを対象に，Wikipedia ページをリンクしているアンカーテキスト（メンション），リンク先の Wikipedia ページ（エンティティ），アンカーテキストの周りの文章（メンションの文脈）を収集して構築された Wikilinks コーパス [20] を基にしている．よって，当該データセットのメンションとその周囲の文脈は，個人によるブログやテレビドラマのセリフなど多様な文章であり，より実際の例に近いと考えられる．

続いて，WikilinksNED Unseen-Mentions をどのように改変して WikilinksNED UM ZS を作成したのかを説明する．Wu らも述べているように [5]，WikilinksNED Unseen-Mentions はテストセットと検証セットの両方に，訓練セット内のメンションと言及しているエンティティが同一であるメンションを含んでいる．よって，そのままでは ZSEL の制約を満たすことができない．そこで，テストセットと検証セットから，上述の訓練セット内のメンションと言及しているエンティティが同一であるメンションを取り除くことにより，ZSEL の制約を満たすことができるデータセットである WikilinksNED UM ZS を作成した．WikilinksNED UM ZS のテストセットと検証セットのメンション数は，それぞれ 7,493 件と 6,872 件である．

4.2 エンティティリンキングモデル

本研究では，ZSEL モデル「BLINK」[5] を使用した．その理由は，現在の EL モデルでは一般的に採用されているバイエンコーダアーキテクチャ [12, 18, 19] を最初に採用したモデルの 1 つであり，かつ多くの EL 研究においてベースラインとして用いられているためである [13, 21, 22]．なお，前述の通りバイエンコーダを採用している EL モデルは多いが，それと比較してクロスエンコーダを採用している EL モデルは少ないため，本研究では BLINK のバイエンコーダのみを使用して分析を行った．そして，ZSEL 設定における EL モデルの性能を分析するために，上述の WikilinksNED UM ZS を用いて訓練した BLINK を実験に使用した．訓練時のハイパーパラメータを表 1 で示し，訓練後のモデルの WikilinksNED Unseen-Mentions における性能を表 2 で示す．なお，エンティティ（記述）集合として，4.3.1 節で述べるエンティティカタログを使用したため，10,000 個のテストデータの内，メンションが言及しているエンティティが当該カタログに含まれていない 12 個のデータは取り除いた．また，クロスエンコーダ使用の有無などの条件が異なるため正確な比較はできないが，参考のため Wu らによる BLINK の性能も同表に記載する [5]．

4.3 関係に応じてエンティティをマスク

本節では，エンティティ記述内において，記述対象のエンティティとどのような関係にあるエンティティが ZSEL の正解率向上に寄与しているかを明らかにするために，3.3 節で述べた手法に沿って行った分析の実験について説明する．順番としては，実験設定，実験結果の順で述べる．

表 1　BLINK 訓練時のハイパーパラメータ.

パラメータ	値
BERT モデル	bert-large-uncased
損失関数	Cross Entropy Loss
オプティマイザ	Adam
学習率	2×10^{-6}
エポック数	1
バッチサイズ	128
エンティティ記述の上限トークン数	128
メンションの文脈の上限トークン数	32

表 2　WikilinksNED Unseen-Mentions における BLINK の性能.

BLINK	Accuracy	Recall@10
本実験のために訓練	0.475	0.868
Wu ら (2020)	0.747	-

4.3.1 実験設定

ZSEL モデルとしては，4.2 節で述べた BLINK を用いた．

エンティティ（記述）集合としては，BLINK の公式 GitHub リポジトリにおいて公開されているエンティティカタログ [1]を使用した．このカタログは 2019 年 8 月 1 日の Wikipedia ダンプから作られており 5,903,527 個のエンティティを含んでいるが，本実験では 3.1 節で述べた ZSEL の制約を満たすために，BLINK の訓練に用いた WikilinksNED UM ZS の訓練セット内のメンションが言及しているエンティティを除いた 5,817,551 個のエンティティのみを使用した．

ZSEL を行う対象のメンション集合としては，4.1 節で述べた WikilinksNED UM ZS のテストセットの一部を用いた．具体的には，テストセットの中から，言及しているエンティティが上述の Wikipedia ダンプに含まれていないメンションと後述の KGE モデルの訓練に用いた Wikidata5m データセット内に含まれていないメンションを除いた 6,778 個を使用した．

評価指標としては，EL において最も重要だと考えられる正解率（Accuracy）を主に用いるが，同時に報告されることが多い Recall@10 も報告する．

関係集合としては，後述の KGE モデルに合わせて Wikidata5m データセットに含まれている 822 個の関係の一部を用いた．具体的には，822 個の関係の中から，Wikidata ダンプを用いて取得した，エンティティ記述内に出現しているエンティティと記述対象のエンティティ間の関係として出現頻度が高い上位 100 個の関係のみを使用した．すべての関係を使わなかった理由は，出現頻度が低い関係は訓練データが少なく KGE モデルが適切な学習を行えないことが考えられ，その結果モデルが誤った推論をしてしまい誤ったエンティティがマスクされてしまう可能性を低くするためである．

KGE モデルとしては,「NodePiece」[23] を，Wikipedia から作られた KG である Wikidata が基となっている Wikidata5m データセット [24] で訓練して使用した．NodePiece を使用した理由としては次の 2 つがある．1 つ目は，Wikidata5m のような数百万個のエンティティを含む巨大な KG を埋め込むのに適

1：http://dl.fbaipublicfiles.com/BLINK/entity.jsonl

表 3　NodePiece 訓練時のハイパーパラメータ.

パラメータ	値
スコア付け関数	RotatE
損失関数	Margin Ranking Loss
オプティマイザ	Adam
学習率	1×10^{-3}
エポック数	5
バッチサイズ	256
関係文脈数	12
埋め込み次元数	200
負例数	1

していることである．2 つ目は，Wikidata5m もそうであるが，エンティティ間の関係の種類が豊富である KG に対する関係推定タスクで比較的高い性能を発揮したと報告されていることである．訓練時のハイパーパラメータを表 3 に示す．

閾値 θ を決める際のマスク割合 p は 15%とした．これは，マスクされる文章量が少ないと ZSEL の正解率が下がらず，エンティティ記述内において ZSEL の正解率向上に最も寄与しているエンティティと記述対象のエンティティとの間にある関係を確認するのが難しいためである．

3.3 節で述べた手法ではエンティティ記述内でマスクするのはエンティティのみであったが，本実験ではエンティティに加えて，エンティティと係り受け関係にある単語 [2]もマスクした．これは，マスクを行った後のエンティティ記述をより自然な文章にするためである．

4.3.2 実験結果

正解エンティティ記述内に含まれている，正解エンティティと特定の関係にあるエンティティをマスクした状態で行った ZSEL の結果を表 4 に示す．表の各行は ZSEL の結果を示しており，左から，(1)Wikidata 内での関係の ID，(2) 正解エンティティとどの関係にあるエンティティをマスクしたか，あるいはマスクしていないか，(3) 正解率，(4)recall@10 を表している．なお，ここでは 4.3.1 節で述べた 100 個の関係の内，対応する正解率の昇順で上から 10 個の関係における結果のみ示す．また,「†」は，Holm 補正を用いて Wilcoxon signed-rank test を行ったところ，対応する正解率の昇順に並べたときに 100 個の関係中 50 番目に位置する関係「discoverer or inventor (P61)」の正解率よりも $\alpha = 0.05$ で有意に低かった正解率に付している．そして,「†」が付された正解率に対応する関係は，半数の 50 個以上の関係よりも正解率が有意に低かった．

上の結果から，正解エンティティ記述内に含まれている，正解エンティティと「instance of (P31)」関係にあるエンティティを，正解エンティティ記述集合全体でのマスク割合が 15%になるように各正解エンティティ記述からマスクした時の ZSEL の正解率は，他の関係にあるエンティティをマスクした時よりも低い場合が多いということが読み取れる．よって，出現頻度で上位 100 位以内に入らない使用頻度が低い関係は ZSEL におい

2：より厳密には，dependency parse tree においてエンティティが依存している単語.

表 4　正解エンティティ記述内の正解エンティティと特定の関係にあるエンティティをマスクした ZSEL 結果.

Wikidata ID	関係名	Accuracy	Recall@10
P31	instance of	0.3200†	0.8138
P279	subclass of	0.3265	0.8020
P527	has part(s)	0.3272	0.8163
P101	field of work	0.3277	0.8150
P641	sport	0.3280	0.8135
P407	language of work or name	0.3287	0.8157
P460	said to be the same as	0.3292	0.8169
P171	parent taxon	0.3294	0.8176
P136	genre	0.3299	0.8187
P355	has subsidiary	0.3302	0.8203
	マスクなし	0.5003	0.8870

ても重要性が低いと考えられることを踏まえて，上の結果より，エンティティ記述を作成する際に，記述対象のエンティティと「instance of」関係にあるようなエンティティを含むことは，他の関係にあるエンティティを代わりに含めるよりも ZSEL の正解率を向上させる場合が多いと考えられる．

また，上と同様の実験を，言及しているエンティティのタイプによって分割したメンション集合に対しても行った．これは，ZSEL に有効なエンティティ記述を作成する際のより有用な指針を得ることを目的として，あるタイプのエンティティ記述内において，記述対象のエンティティとどのような関係にあるエンティティが ZSEL の正解率向上に寄与しているかを明らかにするためである．本研究では「関根の拡張固有表現階層」[3]を参考に，表 5 に示す 11 個のタイプを用いた．なお，1 つのエンティティに複数のタイプが割り当てられることもあったため，メンション数の合計は 6,778 にはならない．エンティティにタイプを割り当てる方法としては，Wikidata をグラフとして見て，あるエンティティ（ノード）から出発し,「instance of」または「subclass of」エッジ上を移動した（すなわち上位概念であるようなエンティティを収集した）ときに，最も距離が近いタイプ集合をそのエンティティが属するタイプであると見做した．これは，関根の拡張固有表現階層は階層構造になっており，その中でより具体性の高いタイプを割り当てるためである．

以降，実験結果を示す．なお，ここでは比較的多くの ZSEL の結果間に有意差が見られたタイプにおける結果のみ紹介する．

human タイプにおける結果を表 6 に示す．表 4 と同様に各行は ZSEL の結果を示しており,「†」は 100 個の関係中 50 番目に位置する関係「filming location (P915)」の正解率よりも有意に低かった正解率に付している．そして,「†」が付された正解率に対応する関係は，半数の 50 個以上の関係よりも正解率が有意に低かった．また,「⋆」は，表 4 で結果を示した実験において，他関係よりも有意に正解率が低い場合が最も多かった関係「instance of」の正解率よりも有意に低かった正解率に付している．

上の結果から，正解エンティティ記述内に含まれている，正解エンティティと「member of sports team (P54)」関係にあるエンティティと「employer (P108)」関係にあるエンティティそれぞれを，各正解エンティティ記述からマスクした時の ZSEL の正解率は，他の関係にあるエンティティをマスクした時よりも低い場合が多いということが読み取れる．前者について，スポーツチームに加入している人は全体のごく一部であるはずにも関わらずこのような結果になった理由を探るために，どのようなエンティティがマスクされたかを確認した．その結果，スポーツチームに加えて，(1) 記述対象の人物が所属している組織，(2) 受賞した賞や発明品，リリースした楽曲などの記述対象の人物による功績がマスクされている例が比較的多く見られた．これらのエンティティがマスクされた理由について，(1) には記述対象の人物の所属先，(2) には記述対象の人物の貢献対象あるいは功績により得たものと，それぞれスポーツチームとの共通点があるため，記述対象のエンティティとこれらのエンティティ間にはスポーツチームとの間にあるのと同様の関係があると KGE モデルが判断した可能性があり得る．

表 5　使用したエンティティのタイプ.

タイプ名	Wikidata ID	拡張固有表現	メンション数
human	Q5	人名	531
deity	Q178885	神名	83
organization	Q43229	組織名	619
geographical object	Q618123	地名	928
construction	Q811430	施設名	103
goods	Q28877	プロダクト名	1,180
occurrence	Q1190554	イベント名	401
natural physical object	Q16686022	自然物名	141
disease	Q12136	病気名	29
color	Q1075	色名	2
concept	Q151885	概念	2,904

よって，human タイプのエンティティ記述を作成する際に，記述対象の人物が所属している組織や功績を含むことは，他の関係にあるエンティティを代わりに含めるよりも，エンティティ集合の要素が human タイプのエンティティに限定された ZSEL の正解率を向上させる場合が多いと考えられる．なお，表 6 内の「⋆」で示したように，関係「instance of」よりも有意に正解率が低い関係が存在するため，上述の限定された ZSEL の正解率を向上させるエンティティ記述を作成する際には，記述対象のエンティティと「instance of」関係にあるようなエンティティを含むだけでは不十分だと考えられる．

goods タイプにおける結果を表 7 に示す．表 4 と同様に各行は ZSEL の結果を示しており,「†」は 100 個の関係中 50 番目に位置する関係「occupation (P106)」の正解率よりも有意に低かった正解率に付している．そして,「†」が付された正解率に対応する関係は，半数の 50 個以上の関係よりも正解率が有意に低かった．また,「⋆」は表 6 と同様である．

上の結果から，正解エンティティ記述内に含まれている，正解エンティティと表 7 内で「†」を付した 7 つの関係にあるエンティティを，各正解エンティティ記述からマスクした時の ZSEL の正解率は，他の関係にあるエンティティをマスクした時より

3：http://liat-aip.sakura.ne.jp/ene/ene9/definition_jp/html/enetree.html(参照 2023-12-15).

表 6　正解エンティティ記述内の正解エンティティと特定の関係にあるエンティティをマスクした ZSEL 結果 (human タイプ).

Wikidata ID	関係名	Accuracy	Recall@10
P54	member of sports team	0.3239*†	0.7552
P108	employer	0.3239*†	0.7533
P69	educated at	0.3333*	0.7571
P22	father	0.3371	0.7439
P102	member of political party	0.3371	0.7589
P172	ethnic group	0.3371*	0.7439
P19	place of birth	0.3390	0.7627
P1344	participant in	0.3390	0.7495
P27	country of citizenship	0.3409	0.7665
P1830	owner of	0.3409	0.7533
	マスクなし	0.4689	0.8098

表 7　正解エンティティ記述内の正解エンティティと特定の関係にあるエンティティをマスクした ZSEL 結果 (goods タイプ).

Wikidata ID	関係名	Accuracy	Recall@10
P161	cast member	0.2983*†	0.8042
P1877	after a work by	0.3068*†	0.8458
P58	screenwriter	0.3076*	0.8390
P162	producer	0.3076*†	0.8364
P921	main subject	0.3085*†	0.8449
P86	composer	0.3127	0.8483
P674	characters	0.3127*†	0.8542
P449	original broadcaster	0.3136*†	0.8508
P840	narrative location	0.3153*†	0.8508
P725	voice actor	0.3153	0.8441
	マスクなし	0.5153	0.9339

も低い場合が多いということが読み取れる．よって，goods タイプのエンティティ記述を作成する際に，記述対象のエンティティと上の 7 つの関係にあるようなエンティティを含むことは，他の関係にあるエンティティを代わりに含めるよりも，エンティティ集合の要素が goods タイプのエンティティに限定された ZSEL の正解率を向上させる場合が多いと考えられる．なお，表 7 内の「$\star$」で示したように，関係「instance of」よりも有意に正解率が低い関係が存在するため，上述の限定された ZSEL の正解率を向上させるエンティティ記述を作成する際には，記述対象のエンティティと「instance of」関係にあるようなエンティティを含むだけでは不十分だと考えられる．

4.4　出現回数に応じてエンティティをマスク

本節では，エンティティ記述内において，記述対象のエンティティと同タイプのエンティティ記述内で出現回数が多いエンティティと少ないエンティティのどちらが ZSEL の正解率向上に寄与しているかを明らかにするために，3.4 節で述べた手法に沿って行った分析の実験について説明する．順番としては，実験設定，実験結果の順で述べる．

4.4.1　実 験 設 定

使用した ZSEL モデル，エンティティ（記述）集合，ZSEL 対象のメンション集合，評価指標は 4.3.1 節と同様である．

正解エンティティが属するタイプ t_c としては，4.3.2 節で述べた 11 個のタイプを用いた．また，記述対象のエンティティ（正解エンティティ）と同タイプのエンティティ記述内におけるエンティティの出現回数 $\mathrm{frq}(e, D_{t_c})$ は，Wikipedia ダンプ内のエンティティ記述として使用されている箇所に含まれているハイパーリンクのリンク先を確認することによって数えた．

表 8　正解エンティティ記述内のエンティティを記述対象のエンティティと同タイプのエンティティ記述内における出現回数に応じてマスクした ZSEL 結果.

出現回数	マスク割合	Accuracy	Recall@10
多い	19.88%	0.2992	0.8008
少ない	20.06%	0.3126	0.8103
マスクなし	0%	0.5003	0.8870

4.3.1 節と同様に分析が行いやすくなるように，閾値 θ を決める際のマスク割合 p は 20%とした．なお，3.4 節での手法の説明でも触れたように，本実験で用いる手法では全正解エンティティ記述を横断してマスク対象のエンティティを決めるのではなく，各正解エンティティ記述ごとにマスク対象のエンティティを決めるため，マスク割合の平均は完全に 20%にはならない．また，4.3.1 節と同様に，エンティティ記述内においてエンティティに加えてエンティティと係り受け関係にある単語もマスクした．

4.4.2　実 験 結 果

正解エンティティ記述内に含まれている，正解エンティティと同タイプのエンティティ記述内で出現回数が (1) 多いエンティティと (2) 少ないエンティティの 2 種類のエンティティをそれぞれマスクした状態で行った ZSEL の結果を表 8 に示す．表の各行は ZSEL の結果を示しており，左から，(1) 正解エンティティと同タイプのエンティティ記述内で多少どちらの出現回数のエンティティをマスクしたか，あるいはマスクしていないか，(2) 正解エンティティ記述内のマスクされたトークン割合の平均，(3) 正解率，(4)recall@10 を表している．なお，ZSEL の結果ペアに対して，4.3.2 節と同様の検定を行ったところ，どの結果ペアについても $\alpha = 0.05$ で有意差が見られた．

上の結果から，正解エンティティ記述内に含まれている，正解エンティティと同タイプのエンティティ記述内で出現回数が多いエンティティを，平均マスク割合が約 20%になるように各正解エンティティ記述からマスクした際の ZSEL の正解率は，出現回数が少ないエンティティをマスクした時よりも低いということが読み取れる．よって，エンティティ記述を作成する際に，現状のエンティティ記述の文章量がある一定以上まで増加するように，記述対象のエンティティと同タイプのエンティティを説明するときによく使われるようなエンティティを含めると，使われることが少ないエンティティを含めたときよりも ZSEL の正解率を向上させることができると考えられる．

なお，上の結果に関連して，Sciavolino らは，BERT [25] ベースのエンコーダを用いた情報検索手法である DPR は Wikipedia 内での出現頻度が低い稀なエンティティを上手く検索できないことを指摘している [26]．本実験で用いた BLINK も BERT ベースのエンコーダを採用しているため，これと同様の現象

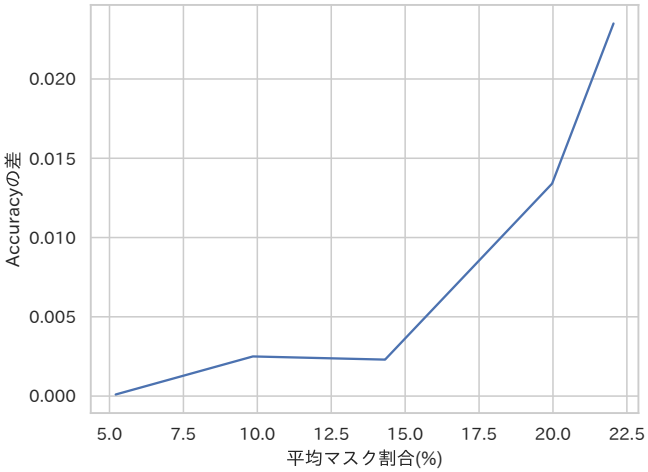


図 1　出現回数の多少間の Accuracy の差.

が起こり，上のような結果が得られたと思われる．この結果の理由として，BERT の事前学習に使われたコーパスの 1 つが Wikipedia であるため，その中で出現頻度が低い単語（エンティティ）は上手く埋め込むことができなかった可能性がある．

また，平均マスク割合を変えたときの，正解エンティティと同タイプのエンティティ記述内で出現回数が多いエンティティと少ないエンティティの 2 種類のエンティティをそれぞれマスクした状態で行った ZSEL の正解率の差を図 1 に示す．

4.5　異なる利用頻度のエイリアスで置換

本節では，エンティティ記述内に出現するエンティティについて，どの利用頻度のエイリアスが ZSEL の正解率向上に寄与しているかを明らかにするために，3.5 節で述べた手法に沿って行った分析の実験について説明する．順番としては，実験設定，実験結果の順で述べる．

4.5.1　実 験 設 定

使用した ZSEL モデル，エンティティ（記述）集合，ZSEL 対象のメンション集合，評価指標は 4.3.1 節と同様である．

式 4 内のエイリアス集合 A_e は，2019 年 1 月 28 日の Wikidata ダンプを用いて収集した．具体的には，ダンプ内の各エンティティについて，“labels” と “aliases” をエイリアスとして使用した．なお，本実験で用いた ZSEL モデル内の BERT は英語用なので，それに合わせて得られたエイリアスの中で英語でないと見做せるものは除外した．具体的には，Unicode の「基本ラテン文字」〜「ラテン文字拡張 B」ブロック外の文字が含まれている，または BERT のトークナイザに入力して得られたトークン列に `[UNK]` トークンが含まれるエイリアスは除外した．また，式 4 内のエイリアスの利用頻度 utl としては，実際の社会における利用頻度をある程度反映していると考えられるため，2019 年 8 月 1 日の Wikipedia ダンプ内におけるエイリアスの出現回数を用いた．

4.5.2　実 験 結 果

正解エンティティ記述内のエンティティを，そのエンティティの (1) 利用頻度が高いエイリアス，(2) 利用頻度が中程度のエイリアス，(3) 利用頻度が低いエイリアスでそれぞれ置換した状態で行った ZSEL の結果を表 9 に示す．表の各行は ZSEL の結果を示しており，左から，(1) 正解エンティティ記述内のエンティティをどの利用頻度のエイリアスで置換したか，あるいは置換していないか，(2) 正解率，(3)recall@10 を表している．なお，ZSEL の結果ペアに対して，4.3.2 節と同様の検定を行ったところ，(1)「利用頻度：中」と「利用頻度：低」，(2)「利用頻度：低」と「置換なし」の 2 組のペアについては $\alpha = 0.05$ で有意差が見られた．

表 9　正解エンティティ記述内のエンティティを異なる利用頻度のエイリアスで置換した ZSEL 結果.

利用頻度	Accuracy	Recall@10
高	0.4951	0.8752
中	0.5025	0.8793
低	0.4879	0.8690
置換なし	0.5003	0.8870

上の結果から，正解エンティティ記述内のエンティティを，そのエンティティの利用頻度が低いエイリアスで置換した際の ZSEL の正解率は，それ以外の利用頻度のエイリアスで置換した（あるいは置換しなかった）際の正解率よりも低い場合が多いということが読み取れる．よって，ZSEL の正解率を向上させるエンティティ記述を作成する際には，その中に含めるエンティティのエイリアスとして利用頻度が低いエイリアスは使うべきでない場合が多いと考えられる．

また，上と同様の実験を 4.3.2 節と同じく，言及しているエンティティのタイプ（11 個）によって分割したメンション集合に対しても行った．ここではタイプで分割しないメンション集合に対して行った実験と異なる傾向の結果が見られた geographical object タイプにおける結果のみ紹介する．

geographical object タイプにおける結果を表 10 に示す．本表内の ZSEL の結果ペアに関して，(1)「利用頻度：高」と「利用頻度：中」，(2)「利用頻度：高」と「利用頻度：低」，(3)「利用頻度：中」と「置換なし」，(4)「利用頻度：低」と「置換なし」の 4 組のペアでは $\alpha = 0.05$ で有意差が見られた．

この結果からは，先ほどとは反対に，正解エンティティ記述内のエンティティを，そのエンティティの利用頻度が低いエイリアスで置換した際の ZSEL の正解率は，それ以外の利用頻度のエイリアスで置換した（あるいは置換しなかった）際の正解率よりも高い場合が多いということが読み取れる．

上のような結果が得られた理由を探るために，どのようなエイリアスに置換されたかを確認した．その結果，「London」が「London, United Kingdom」に置換されるなど，正解エンティティ記述内に出現する地名エンティティが，その地名が属する国名や州名を含むエイリアスに置換されている例が多く見られた．このことから，geographical object タイプに属するエンティティのエンティティ記述内にはそのエンティティと同じ国や州などに位置する地名エンティティが多く含まれており，それらがエイリアスへ置換されることによって位置する国や州などに関する情報が増え，ZSEL が上手く行えるようになったということが考えられる．

表 10　正解エンティティ記述内のエンティティを異なる利用頻度のエイリアスで置換した ZSEL 結果 (geographical object タイプ).

利用頻度	Accuracy	Recall@10
高	0.4375	0.9256
中	0.4731	0.9256
低	0.4806	0.9267
置換なし	0.4267	0.9224

よって，エンティティ集合の要素が geographical object タイプのエンティティに限定された ZSEL の正解率を向上させるエンティティ記述を作成する際に，その中に同じ地域に位置するエンティティを含める場合は，そのエンティティの名前に加えて位置する地域も記載するべきだと考えられる．

5 ま と め

本研究では，ZSEL に有効なエンティティ記述を作成する際の指針を得ることを目的として，エンティティ記述内のどの部分が ZSEL の正解率向上に寄与しているかを明らかにするための分析を行った．分析のために，メンションが言及しているエンティティのエンティティ記述内の一部をマスクあるいは置換して，その結果 ZSEL の正解率がどのように変化するかを確認するという手法を用いた．分析の結果，エンティティ間の関係，エンティティの出現回数，エイリアスの利用頻度が ZSEL の正解率に影響していることが示唆された．

今後の研究課題としては，ZSEL に有効なエンティティ記述に関する更なる分析や，ZSEL に有効なエンティティ記述と人間が読みやすいエンティティ記述の両立が挙げられる．

謝辞　本研究は JSPS 科研費 JP21H03775，JP23H03400 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.

[2] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. *ACM computing surveys (CSUR)*, Vol. 32, No. 2, pp. 144–173, 2000.

[3] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3449–3460, 2019.

[4] Tom Oberhauser, Tim Bischoff, Karl Brendel, Maluna Menke, Tobias Klatt, Amy Siu, Felix Alexander Gers, and Alexander Löser. Trainx–named entity linking with active sampling and bi-encoders. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pp. 64–69, 2020.

[5] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6397–6407, 2020.

[6] Yasumasa Onoe and Greg Durrett. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 8576–8583, 2020.

[7] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 250–259, 2016.

[8] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2619–2629, 2017.

[9] Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 30–34, 2013.

[10] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1333–1339, 2015.

[11] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 519–529, 2018.

[12] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. A transformational biencoder with in-domain negative sampling for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1449–1458, 2022.

[13] GP Shrivatsa Bhargav, Dinesh Khandelwal, Saswati Dana, Dinesh Garg, Pavan Kapanipathi, Salim Roukos, Alexander Gray, and L Venkata Subramaniam. Zero-shot entity linking with less data. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1681–1697, 2022.

[14] Xuhui Sui, Ying Zhang, Kehui Song, Baohang Zhou, Guoqing Zhao, Xin Wei, and Xiaojie Yuan. Improving zero-shot entity linking candidate generation with ultra-fine entity type information. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2429–2437, 2022.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[16] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[17] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[18] Jan A Botha, Zifei Shan, and Dan Gillick. Entity linking in 100 languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7833–7845, 2020.

[19] Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. Fast and effective biomedical entity linking using a dual encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 28–37, 2021.

[20] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. 2012.

[21] Xiaodong Yu, Wenpeng Yin, Nitish Gupta, and Dan Roth. Event linking: Grounding event mentions to wikipedia. In

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2671–2680, 2023.

[22] Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6433–6441, 2020.

[23] Mikhail Galkin, Etienne Denis, Jiapeng Wu, and William L Hamilton. Nodepiece: Compositional and parameter-efficient representations of large knowledge graphs. In *International Conference on Learning Representations*, 2021.

[24] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 176–194, 2021.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6138–6148, 2021.

根拠学習とマルチタスク学習を用いた説明可能なヘイトスピーチ検出

本城　佑太†　　張　　建偉†

† 岩手大学理工学部　〒 020–8551 岩手県盛岡市上田 4-3-5
E-mail: †{s0620057,zhang}@iwate-u.ac.jp

あらまし　SNS の発達により，ヘイトスピーチの投稿が問題となっている．近年，ヘイトスピーチ検出について様々な研究が行われており，検出性能が向上している．しかし，高い検出性能を持つモデルの多くが複雑な構造をしているため，検出結果の要因を理解することが難しくなっている．BERT-MRP は，根拠学習とファインチューニングの 2 段階で学習を行うモデルであり，ヘイトスピーチ検出だけではなく説明も行う．モデルの説明とは，ヘイトスピーチと予測した投稿のテキストにおいて，その根拠となる部分を抜き出したものである．本研究では，BERT-MRP モデルの事前学習モデルの変更，根拠学習の拡張，さらに，BERT-MRP モデルのファインチューニングの部分において，投稿の対象となるグループを予測するターゲット分類とヘイトスピーチ検出のマルチタスク学習を用いた説明可能なヘイトスピーチ検出を提案する．今回，各提案手法に基づき 3 つの実験を行った．モデルの説明の内容を評価する説明可能性の観点で，すべての提案手法において，BERT-MRP より向上する結果となった．

キーワード　XAI，テキスト分類，根拠学習，マルチタスク学習，ヘイトスピーチ検出

1　はじめに

SNS が発達し，オンラインのヘイトスピーチが 1 つの社会問題となっている [1]．オンラインのヘイトスピーチは，広がりやすい特性を持っており，SNS 上で拡散されている．その影響で，特定のグループへの差別や偏見がおこるなど，ヘイトスピーチの対象となった人の人権が侵されている．最終的には，ヘイトスピーチの投稿は戦争や紛争等が起こる原因にもなりうる．そのため，SNS 上のヘイトスピーチを適切に規制することは，多くの社会問題を解決するために重要である [1]．

近年，ヘイトスピーチ検出に関する様々な研究が行われており，検出性能は向上している．しかし，高い検出能力を持つモデルの多くが，複雑な構造をしている [2]．そのため，モデルが検出した要因について，我々が理解することが難しくなっている．実際，ヨーロッパなどでは，説明を受ける権利が確立されている [3]．これらのことから，検出モデルの検出した要因についての説明が必要になってきている．

また，既存研究の多くのヘイトスピーチ検出モデルは，予測にバイアスがあると言われている．バイアスのあるモデルは，一般的に攻撃されるアイデンティティへの言及があるコメントに対して有害でなくとも，有害と分類する傾向がある [4]．一般的に攻撃されるアイデンティティとは，ゲイや黒人，イスラム教などのことである．このことから，アイデンティティに関する言葉によって判断するのではなく，文脈によって判断するバイアスの低いモデルが必要となっている [1]．

Mathew らは，説明可能なヘイトスピーチ検出のためのベンチマークデータセット HateXplain [5] を開発した．ここでいう説明は，検出モデルがヘイトスピーチと判断した際に，投稿のテキストにおいて，モデルがヘイトスピーチと判断した根拠となるトークンを示したものである．このベンチマークでは，検出モデルの検出能力を評価する検出性能，検出モデルの説明について評価する説明可能性，さらに，ターゲットグループという投稿の対象となっているグループ（女性やアフリカ人等）についての検出モデルのバイアスを評価するバイアスの 3 つの観点でモデルを分析することができる．

現在，HateXplain ベンチマークで最良のモデルは，Kim らの提案した BERT-MRP モデル [1] というモデルである．このモデルは 1 段階目に根拠学習，2 段階目にヘイトスピーチ検出のファインチューニングを行うことによって学習が行われる．

本論文では，説明可能性をさらに向上させるため，Kim ら [1] の 2 段階の学習をベースとして用いながら，事前学習モデルを置き換える手法と根拠の学習を拡張する手法を提案する．また，BERT-MRP モデルは，データセットの各投稿にアノテーションされているターゲットグループを利用していない．そこで，BERT-MRP モデルの 2 段階で学習する構造は変えず，2 段階目のファインチューニングについて，ヘイトスピーチ検出とターゲット分類のマルチタスク学習を用いた説明可能なヘイトスピーチ検出モデルを提案する．

事前学習モデルを変更する手法と根拠の学習を拡張する手法を適用することにより，双方の提案手法とも検出性能と説明可能性の指標が BERT-MRP モデルと比較して向上するという結果が得られた．また，マルチタスク学習を用いた説明可能なヘイトスピーチ検出モデルについても同様に，検出性能，説明可能性の指標において，BERT-MRP モデルと比較して向上するという結果が得られた．

2　関連研究

Saha ら [6] は，XGboost を用いた女性に対するヘイトスピーチ検出手法を提案した．また， Gao ら [7] は，LSTM/GRU を用いてヘイトスピーチ検出を行った．さらに，Zhang ら [8]

は，CNN-GRU を用いてヘイトスピーチ検出を行った．その後，BERT [9] などの transformers モデルの登場により，ヘイトスピーチ検出は，様々なデータセットにおいて，性能が向上した [10]．Koufakou ら [11] は，BERT のアーキテクチャに侮辱的な単語の辞書を組み合わせた HurtBERT を提案した．Caselli ら [12] は，BERT の事前学習モデルに，SNS による攻撃的な投稿のテキストを用いて MLM(Masked Language Modeling) で再学習を行った HateBERT を提案し，検出性能を向上させた．Mathew ら [5] は，ベンチマークデータセット HateXplain を作成した．さらに，作成したデータセットを利用し，BERT [9] のファインチューニングを行う際，根拠の学習も行う BERT-HateXplain モデルを開発した．BERT-HateXplain モデルは説明可能性の観点において，BERT よりも向上した．その後，Kim ら [1] は BERT-HateXplain モデルの根拠の学習を中間タスクとして行う BERT-MRP モデルを作成した．BERT-MRP モデルは検出性能，バイアス，説明可能性の 3 つの観点のすべてにおいて BERT-HateXplain モデルより向上した．本研究では，BERT-MRP モデルをベースとして，さらに説明可能性を向上させるため，3 つの手法を提案している．

AI のモデルの結果の根拠を説明することは，モデルの結果を人間に説明する際に必要である [13]．Ribeiro ら [14] は，複雑なモデルを解釈するための手法である LIME を開発した．分類問題において，LIME は入力特徴量の中で，予測に重要な特徴量を示すことができる．いくつかの自然言語処理の研究では，根拠はモデルの予測を裏付ける入力テキストの一部分と定義されている [15] [16] [17]．Mathew ら [5] の作成した HateXplain データセットでは，有害な投稿について，有害である根拠がアノテーションされている．Kim ら [1] の提案した BERT-MRP モデルでは，Masked Rationale Prediction(以下，MRP) という方法を用いて，モデルに根拠を学習させている．本研究では，BERT-MRP モデルの MRP の学習を用いている．さらに，LIME を用いた説明手法も利用している．

Vaidya ら [4] は，マルチタスクの学習において，ターゲットグループラベルと有害なテキストを学習させた．結果，マルチタスク学習を用いることにより，精度が向上し，バイアスが低減した．また，delArco ら [18] は，BERT を用いて，ターゲット分類と，ヘイトスピーチ検出をマルチタスクで学習することで，精度が向上することを示した．本研究では，利用するデータセットにターゲットグループが含まれていることから，ターゲット分類とヘイトスピーチ検出の 2 つのタスクのマルチタスク学習を試みている．

3 予 備 知 識

3.1 HateXplain データセット [5]

このデータセットは，Twitter と Gab という SNS から 20,148 個の投稿を収集したものである．1 つの投稿につき，クラスラベル，ターゲットグループ，根拠がアノテーションされている．図 1 は，データセットの例である．

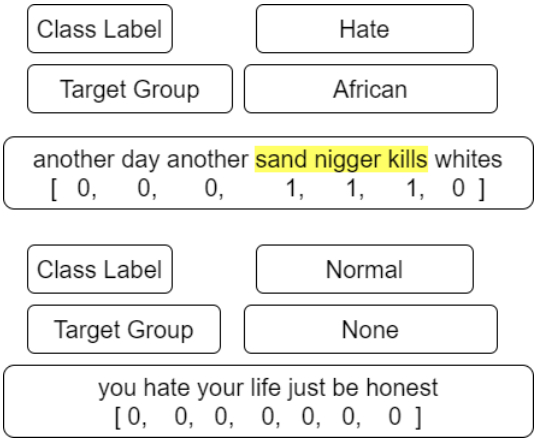


図 1　HateXplain データセット

投稿の内容

ユーザーデータはなく，テキストのみが収集されている．また，テキストとして，1 文の英語の投稿が収集されている．

クラスラベル

1 投稿につき，“Hate”, “Offensive”, “Normal” のいずれかがアノテーションされている．なお，“Offensive” は，“Hate” に次ぐ攻撃性を持つことを意味している．

ターゲットグループ

ターゲットグループは，女性や，アジア人などの投稿の向けられた相手のグループのことである．ターゲットグループラベルは，1 つの投稿につき 1 個以上アノテーションされている．アノテーションされるターゲットグループラベルは，特定の対象グループが 10 個と投稿の対象がいない “None” の 11 個である．特定の対象となるグループは，“African”, “Islam”, “Jewish”, “Homosexual”, “Women”, “Refugee”, “Arab”, “Caucasian”, “Asian”, “Hispanic” の 10 グループである．

根拠

根拠は，投稿のクラスラベルが “Hate” または “Offensive” の際に，クラスラベルを選択する要因となった投稿のテキストの一部分である．投稿のテキストの各単語ごとに対して，“Hate” または “Offensive” を選択する要因となった部分の単語は 1，それ以外の単語は 0 としてアノテーションされている．また，“Normal” の投稿のテキストには，全単語に 0 とアノテーションされている．

3.2 BERT-MRP モデル [1]

図 2 は BERT-MRP モデルのアーキテクチャを示したものであるである．BERT-MRP モデルは，Kim らによって提案された学習を 2 段階で行う説明可能なヘイトスピーチ検出モデルである．モデルの学習について説明する．

1 段階目 (Stage1) では，MRP により，根拠学習を行っている．まず，投稿のテキストを BERT で通常通りエンベディン

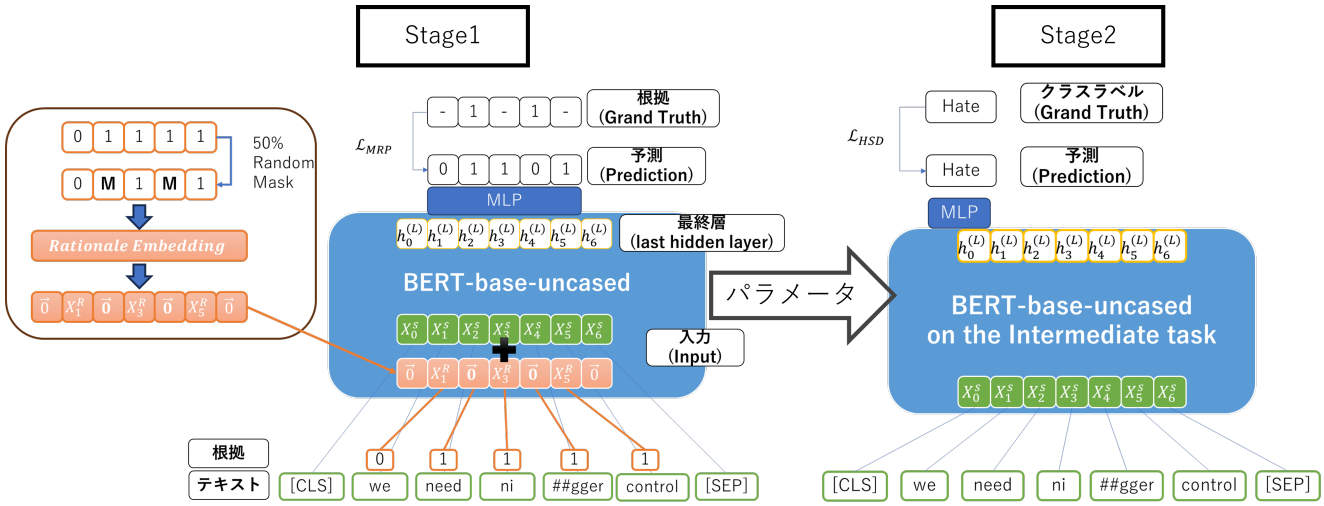


図 2　BERT-MRP アーキテクチャ

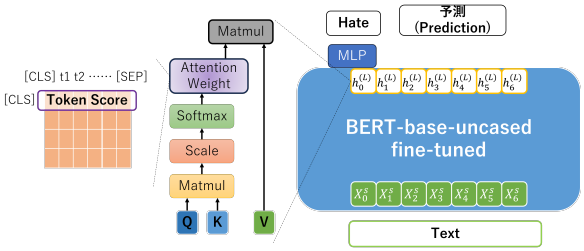


図 3　説明手法: Attn

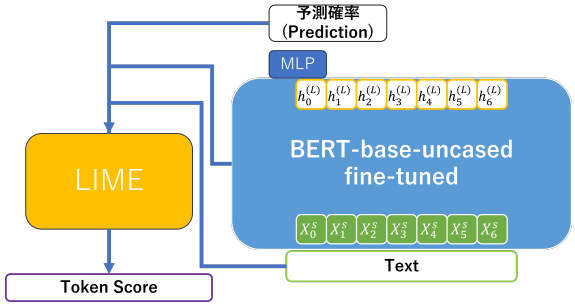


図 4　説明手法: LIME

グする．続いて，投稿の根拠について，根拠全体の 50 %を Mask して根拠エンベディングを行う．根拠エンベディングでは，Mask されていないトークンの根拠をテキストの各トークンと同じ次元で埋め込み，Mask された根拠と元々根拠のない [SEP] や [CLS] という特殊トークンは $\vec{0}$ として埋め込みを行う．これらのエンベディングされたテキストと根拠を足し合わせて BERT の入力とする．そこから BERT のトークン分類を用いて，Mask された根拠の部分について，根拠かどうかを予測し学習を行う．

2 段階目 (Stage2) はファインチューニングとして，ヘイトスピーチ検出の学習を行う．具体的には，1 段階目 (stage1) で学習した BERT のパラメータを初期値として移行して，投稿のテキストのみでヘイトスピーチ検出のファインチューニングを行う．

推論は，2 段階の学習を終えた BERT を用いて行う．その後，推論結果に基づいて，Attn と LIME の 2 つの手法で説明を行う．

- Attn

図 3 は説明手法の Attn 手法を示したものである．Attn の手法は，BERT の最終層の Attention Weight の [CLS] トークンに対応する部分の各トークンの重みをトークンスコアとする．Attention Weight は Softmax を通した状態となっている．そのため，トークンスコアは各トークンごとに [0,1] の範囲の値が得られる．このトークンスコアが高い 5 つのトークンをモデルの説明とする．

- LIME [14]

LIME は Ribeiro らによって，提案された事後説明手法である．LIME は，BERT 等を用いた複雑なモデルに対して，予測した要因となる入力トークンを探ることができる．また，このモデルはインスタンスごとに，説明を行う．図 4 は説明手法の LIME 手法を示したものである．LIME 手法は，LIME モデル [14] に，ヘイトスピーチ検出の推論時に利用したモデル，入力テキスト，予測確率の 3 つを入力として与え，出力される各トークンに対する予測の貢献度をトークンスコアとする．予測の貢献度は [-1,1] の範囲で出力され，1 に近いほど予測に大きく貢献したトークンとなる．Attn と同様にトークンスコアが高い 5 つのトークンをモデルの説明として扱う．

4　提 案 手 法

4.1　提案手法 1：事前学習モデルの置き換え

BERT-MRP モデル [1] は，BERT-base-uncased という事前学習モデルを用いて，学習を行っている．BERT-base-uncased は，自然言語処理で一般的に利用されるモデルである．ここで，BERT-base-uncased をよりヘイトスピーチ検出の用途に特化したモデルに変更することで，検出性能が向上し，それに伴い説明可能性も向上すると考えた．そこで，ヘイトスピーチ検出の用途に特化した HateBERT [12] という事前学習モデルに置き換えを行った．

図 5　提案手法 2

HateBERT [12] は，ヘイトスピーチの検出性能が BERT-base-uncased よりも高く，BERT-base-uncased と同じアーキテクチャを持っている．このことから，BERT-base-uncased と同様の学習が可能であると考え，置き換えを行った．

元々BERT-MRP モデル [1] は 2 つの学習段階がある．そこで，提案手法として以下の 3 つのモデルを提案する．

- 1st HateBERT

1 段階目を HateBERT，2 段階目を BERT-base-uncased で学習させたもの．

- 2nd HateBERT

1 段階目を BERT-base-uncased，2 段階目を HateBERT で学習させたもの．

- All HateBERT

1 段階目と 2 段階目の双方を HateBERT で学習させたもの．

4.2　提案手法 2：根拠学習の拡張

BERT-MRP モデル [1] では Stage1 で根拠を学習する際に，“Hate” もしくは “Offensive” の投稿 (以下，攻撃的な投稿) の根拠ではない単語と “Normal” の投稿（以下，攻撃的ではない投稿）の単語が同じ根拠ラベルとして扱われている．HateXplain データセット [5] は，攻撃的な投稿のみ，根拠がアノテーションされており，攻撃的ではない投稿には根拠がアノテーションされていない．ここで，攻撃的ではない投稿の根拠は，攻撃的な投稿の根拠のような部分的なものではなく，文章全体の文脈と考えた．そこで，攻撃的ではない投稿の全単語を根拠として扱うために，新たなラベルを追加し，根拠学習を 3 値分類にした．この手法は，攻撃的ではない投稿についての根拠も学習することができるため，説明可能性が向上すると考えている．

具体的には，BERT-MRP モデル [1] の 1 段階目の根拠学習に利用する根拠を図 5 に示すように変更する．先行研究と同じく，攻撃的な投稿の根拠ではない部分の単語は 0，攻撃的な投稿の根拠の単語は 1 とする．先行研究と異なる点として，攻撃的でない投稿の全単語の根拠を 2 に変更する．このようにして，2 値分類であった根拠学習を 3 値分類に変更する．今回，根拠が 3 値分類で学習した際の様子を観察したいため，事前学習モデルは先行研究と同様に BERT-base-uncased を用いる．

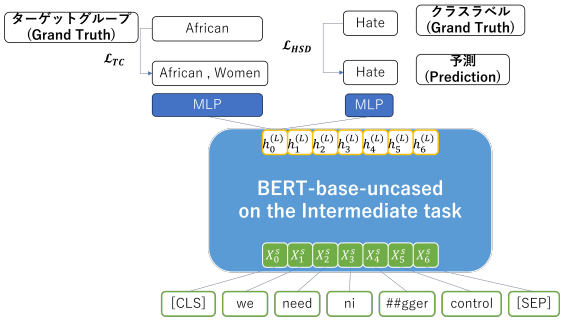


図 6　提案手法 3

4.3　提案手法 3：マルチタスク学習

BERT-MRP モデルの学習では，データセットに含まれるターゲットグループを利用していない．しかし，既存研究では，ターゲット分類を行うことによって，検出性能が向上するという結果が示されている [18]．このことから，2 段階目のファインチューニングの部分で，ヘイトスピーチ検出の学習だけでなく，ターゲット分類も行うことで，検出性能が向上すると考えた．さらに，ターゲットグループごとの特性も学習できることから，説明可能性が向上すると考えた．

提案手法は，1 段階目の根拠学習の部分は BERT-MRP モデルと同様に行う．先行研究と異なる点としては，2 段階目のヘイトスピーチの学習をマルチタスクで行う点である．マルチタスクとして，ヘイトスピーチ検出とターゲット分類の 2 つのタスクを同時に学習を行う．2 段階目の学習は，具体的には図 6 のアーキテクチャを用いる．ヘイトスピーチ検出は，先行研究と同様に行う．ターゲット分類は，投稿の対象がどのグループかを予測するタスクである．今回用いるデータセットでは，11 個のターゲットグループに関して，マルチラベルで，アノテーションがされている．そこで，ターゲット分類をマルチラベル分類問題として定義する．

Loss は，以下の式で計算される．

$$\mathcal{L} = \alpha \mathcal{L}_{HSD} + (1-\alpha)\mathcal{L}_{TC} \qquad (0 \leqq \alpha \leqq 1) \qquad (1)$$

Loss は，ハイパーパラメータ α を用いて，各タスクのロスに重み付けを行い，足し合わせることで計算している．ヘイトスピーチ検出のロス関数 $\mathcal{L}_{HSD}$ は，先行研究と同様にクロスエントロピーロスを用いる．また，ターゲット分類のロス関数 $\mathcal{L}_{TC}$ は，マルチラベルであるため，バイナリークロスエントロピーロスを用いる．

提案手法は，2 つのタスクに重みを付けるハイパーパラメータ α がある．そこで，α が，0.3，0.5，0.7 の 3 つをそれぞれ提案手法とする．今回，マルチタスク学習による結果の効果を観察したいため，事前学習モデルは先行研究と同様に BERT-base-uncased を用い，根拠学習についても先行研究と同様に 2 値分類で確かめる．

5　実 験 設 定

3 つそれぞれの提案手法による効果を検証するため，各提案

表 1　データセットのデータ分割内容

	Toxic	Non-Toxic	Sum
Train	9128	6251	15379
Val	1141	781	1922
Test	1142	782	1924
Sum	11411	7814	19225

表 2　ハイパーパラメータ設定

項目		**値**
学習率	Stage1	2e-5
	Stage2	1e-5
バッチサイズ		16
エポック		Max5
Optimizer		RAdam
Dropout 率		0.1

手法について個別に実験を行った．

実験 1 では，事前学習モデルの置き換えに関する実験を行う．提案手法のモデルは，1st HateBERT，2nd HateBERT，All HateBERT の 3 つである．

実験 2 では，根拠学習の拡張に関する実験を行う．提案手法のモデルは，根拠 3 値分類モデルである．

実験 3 では，マルチタスク学習を用いたモデルに関する実験を行う．提案手法のモデルは，ロス関数のタスクごとの重みを変えた multi-task(α=0.7)，multi-task(α=0.5)，multi-task(α=0.3) の 3 つである．

実験 1，2，3 のすべての実験について，比較対象（Base）として，先行研究の BERT-MRP モデルで実験した結果を出している．また，すべての実験において，HateXplain データセット [5] を利用しており，ヘイトスピーチ検出は 2 値分類で行っている．

3 値のクラスラベルを 2 値分類用のクラスラベルにする方法は，HateXplain [5] に従い次のように行う．3 値分類における “Hate” と “Offensive” のラベルを 1 つにまとめ，“Toxic” というラベルにする．また，3 値分類の “Normal” のラベルを “Non-Toxic” というラベルにする．この 2 つのラベルを 2 値分類のラベルとする．

データセットは，8:1:1 に分割し，学習，検証，テストに用いている．表 1 に実際に分割したデータの数を示す．モデルの学習について，ハイパーパラメータは表 2 のように設定した．ハイパーパラメータは全て，Kim ら [1] の研究と同じ条件にしている．

6 評価指標

実験は，検出性能，バイアス，説明可能性の大きく 3 つの観点から評価する．

表 3　混同行列

		モデルの予測	
		Positive	Negative
実際のクラス	Positive	TP	FN
	Negative	FP	TN

表 4　バイアス評価のための分割

	特定のグループ (subgroup)	その他のグループ (background)
Toxic(実際のクラス)	D_g^+	D^+
Non-Toxic(実際のクラス)	D_g^-	D^-

6.1 検出性能

検出性能は，ヘイトスピーチ検出の性能を評価する指標である．Accuracy(以下，ACC)，macro-F1 スコア (以下，macro-F1)，AUROC スコア (以下，AUROC) で評価する．

Accuracy は表 3 を用いると，以下の式で求めることができる．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

ヘイトスピーチ検出が 2 値分類であるため，macro-F1 は表 3 を用いると，以下の式で求めることができる．

$$\begin{aligned} F1 &= \frac{\text{Precision} \cdot \text{Recall} \cdot 2}{\text{Precision} + \text{Recall}} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{3}$$

AUROC は，真陽性率と偽陽性率の関係を表す ROC 曲線の下の面積から求めることができる．

6.2 バイアス

バイアスは，アイデンティティに関連する内容や単語により，分類器が予測していないかを評価する指標である．本研究では，先行研究と同様にバイアスを評価するため，Borkan ら [19] により提案された 3 つの AUC ベースの評価指標を用いる．

これらの指標は表 4 の 4 つに分割したデータをもとに評価を行う．尚，この章において，女性や，アジア人などの特定のターゲットグループを subgroup，特定のターゲットグループ以外のグループを background と呼ぶ．

一つ目は，Subgroup-AUC（以下，Sub）である．この指標では，subgroup 内の検出性能の評価を行う．指標は，D_g^+ と D_g^- の投稿を検出モデルにかけて，AUC を計算して求めることができる．この指標が高いと，モデルがバイアスを持たず，公平に subgroup のデータを扱っているということができる．

二つ目は，BPSN-AUC（以下，BPSN）である．BPSN は Background Positive Subgroup Negative の略である．指標は，D_g^- と D^+ の投稿を検出モデルにかけて，AUC を計算して求めることができる．この指標は，subgroup の有害ではない投稿の偽陽性のなりやすさを調べる指標である．指標のスコアが

高いと，subgroup の投稿のモデルの予測について，偽陽性が起こりにくいということができる．

三つ目は，BNSP-AUC（以下，BPSN）である．BNSP は Background Negative Subgroup Positive の略である．指標は，D_g^+ と D^- の投稿を検出モデルにかけて，AUC を計算して求めることができる．この指標は，subgroup の有害な投稿の偽陰性のなりやすさを調べる指標である．指標のスコアが高いと，subgroup の投稿のモデルの予測について，偽陰性が起こりにくいということができる．

HateXplain データセットには “None” 以外 10 個の subgroup がある．したがって，各 subgroup ごとに Sub，BPSN，BNSP を計算すると各指標毎に 10 個ずつスコアが出る．先行研究と同様に，各指標の 10 個のスコアを GMB（Generalized Mean of Bias）を利用して，1 つに統合する．GMB は以下の式である．

$$M_p(m_s) = (\frac{1}{N}\sum_{s=1}^{N} m_s^p)^{\frac{1}{p}} \tag{4}$$

M_p は p 個めの power mean 関数のことを表し，m_s は subgroup s における計算された (BPSN などの) 指標を示す．N は subgroup の総数を示しており，今回は総数が 10 個のため，$N = 10$ となる．指数 $\frac{1}{p}$ の部分のについては，先行研究と同様に $p = -5$ として計算を行う．

6.3 説明可能性

説明可能性では，モデルの説明について，指標を用いて評価する．説明可能性は，先行研究と同様に ERASER [20] によって提案された指標で評価する．指標は，Plausibility と Faithfulness の大きく 2 つに分かれる．

6.3.1 Plausibility

はじめに，Plausibility とは，アノテーションされた根拠とモデルの出力した説明 (根拠) がどの程度一致しているかを示すものである．具体的に，3 つの指標がある．

1 つ目は IOU-F1 である．IOU-F1 は，アノテーションされた根拠とモデルの出力した説明 (根拠) が，フレーズレベルで，どの程度一致しているかを評価する．この評価方法は，IOU と呼ばれる計算方法を使う．モデルの説明のフレーズについて，アノテーションされた根拠のフレーズと一致する部分が 50%以上あるものを，フレーズが一致しているとみなす．そして，その一致したフレーズの数を用いて，以下の Precision と Recall から F1 を求める．この指標は，スコアが高いほど，フレーズレベルで，人間の根拠と類似した説明を行うことができるということができる．

$$Precision = \frac{\text{一致したフレーズの数}}{\text{モデルの説明 (根拠) のフレーズの数}} \tag{5}$$

$$Recall = \frac{\text{一致したフレーズの数}}{\text{アノテーションされた根拠のフレーズの数}} \tag{6}$$

2 つ目は，Token-F1 である．Token-F1 は，アノテーションされた根拠とモデルの出力した説明 (根拠) が，トークンレベルでどの程度一致しているかを評価する．評価方法は，以下の Precision，Recall から F1 を計算する．この指標は，スコアが高いほど，トークンレベルで，人間の根拠と一致している説明を行うことができるといえる．

$$Precision = \frac{|E_m^t \cap E_h^t|}{|E_m^t|} \tag{7}$$

$$Recall = \frac{|E_m^t \cap E_h^t|}{|E_h^t|} \tag{8}$$

E_m^t は，モデルの説明 (根拠) のトークンの集合を，E_h^t はアノテーションされていた人間の根拠のトークンの集合を示している．また，$|E_m^t|$ は，モデルの説明 (根拠) のトークンの集合の総数を示している．

3 つ目は，AUPRC である．Token-F1 の Precision と Recall を使用し，閾値としてトークンスコアを用いて AUPRC を計算する．トークンスコアについて，説明手法の LIME は [-1,1] の範囲をとるが，0 以下の値を 0 として扱うことで，計算が行われる．この指標は数値が高いほど，モデルの説明が人間の根拠と一致する説明であるということができる．

6.3.2 Faithfulness

Faithfulness は，モデルの説明 (根拠) が，検出モデルにどの程度影響を与えるものなのか，という根拠の反映度を評価する指標である．指標は 2 つある．

1 つ目は Comprehensiveness(以下，Comp) である．Comp は，入力から根拠を包括的にとれているかを評価する指標である．この指標が高いと，入力の中で根拠となる部分を満遍なく捉えている説明と言える．指標は以下の式によって求める．

$$Comprehensiveness = m(x_i)_j - m(x_i \backslash r_i)_j \tag{9}$$

x_i は，検出モデルに入力したテキストを示している．$(x_i \backslash r_i)$ は検出モデルへの入力 x_i からモデルの説明 (根拠) となる部分を抜いたものを示している．$m(x_i)_j$ は，入力 x_i を検出モデル m に入力した際の予測クラス j の予測確率を表している．Comp は，入力 x_i の予測確率から x_i からモデルの説明 (根拠) となる部分を抜いた $(x_i \backslash r_i)$ の予測確率を引いて計算される．

2 つ目は Suffciency(以下，Suff) である．Suff は，検出モデルにとって予測に十分な説明（根拠）が取れているかを示す指標である．この指標は，値が低いほど，予測に十分な根拠が取れている説明ということができる．以下の式によって求められる．

$$Sufficiency = m(x_i)_j - m(r_i)_j \tag{10}$$

x_i は，検出モデルに入力したテキストを示し，r_i はモデルの説明（根拠）となる部分を示している．Suff は，入力 x_i の予測確率から r_i はモデルの説明（根拠）の予測確率を引くことにより，求められる．

6.3.3 説明可能性の評価方法の変更

先行研究である BERT-MRP モデル [1] では，トークンスコアの高い順から 5 個のトークンをモデルの説明としていた．ここで，データセットの攻撃的な投稿はそれぞれ，テキストの根拠のトークン数が異なる．図 7 から，本来根拠が 4 トークンしかないテキストにおいて，先行研究の説明方法では，“whites”

表 5　実験 1(事前学習モデルの置き換え) 結果 (検出性能・バイアス)

	検出性能			バイアス		
モデル	ACC	macro-F1	AUROC	Sub	BPSN	BNSP
Base	78.4	77.4	87.5	78.7	68.6	86.2
1st HateBERT	**79.6**	78.5	87.7	77.2	66.9	85.7
2nd HateBERT	79.5	78.5	86.7	75.4	66.6	82.9
All HateBERT	79.3	**78.7**	**88.0**	**79.0**	66.6	85.8

表 6　実験 1(事前学習モデルの置き換え) 結果 (説明可能性)

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
1st HateBERT	Attn	41.8	**63.5**	**57.3**	47.8	12.2
2nd HateBERT	Attn	31.6	57.3	51.5	40.6	16.3
All HateBERT	Attn	**42.1**	**63.5**	55.6	51.5	**9.5**
Base	LIME	54.7	69.3	77.3	53.3	-2.7
1st HateBERT	LIME	54.9	70.0	78.5	50.0	-3.8
2nd HateBERT	LIME	**56.7**	**70.9**	78.4	46.2	-1.3
All HateBERT	LIME	55.6	70.3	**78.6**	**53.7**	**-4.9**

表 7　実験 2(根拠学習の拡張) 結果 (検出性能・バイアス)

	検出性能			バイアス		
モデル	ACC	macro-F1	AUROC	Sub	BPSN	BNSP
Base	78.4	77.4	87.5	78.7	68.6	86.2
根拠 3 値分類モデル	**80.0**	**79.1**	**87.7**	75.6	**70.2**	84.4

図 7　モデルの説明

を余分に根拠としていることが分かる．このように，先行研究の方法では，モデルの説明となるトークンの数を 5 個に固定してしまうために，本来できるであろう説明の質を検証できなくなっている．そこで，今回の実験では，より正確な評価をするため，モデルの説明の根拠トークンについて，5 個という制限を与えずに，アノテーションされたの根拠の数と同じ個数のトークンをモデルの説明 (根拠) としている．

7　結果と考察

指標別にまとめた実験結果を表 5，表 6，表 7，表 8，表 9，表 10 に示す．但し，実験の結果は，指標毎に Base を超えた部分はアンダーライン，そのうち，提案手法において最良であった部分は太字としている．

7.1　実験 1 事前学習モデルの置き換えの結果と考察

実験結果は，表 5，表 6 のとおりである．

検出性能について，各提案手法のモデルと Base モデルを比較する．1st HateBERT では全ての指標で Base より向上した．次に，2nd HateBERT では，ACC と macroF1 で向上した．最後に，All HateBERT では，全ての指標で，Base より向上する結果となった．この結果から，HateBERT を用いることで，検出性能が向上するということができる．

バイアスについて，提案手法のモデルと Base モデルを比べると，All HateBERT の Sub 以外の指標が，Base を下回る結果となった．これは，HateBERT が攻撃的な言語を用いて MLM で再学習しているモデルのため，攻撃的な単語に敏感になり，下がったと考えられる．しかし，All HateBERT の Sub が向上した要因について，分析する必要がある．

説明可能性について，各提案手法のモデルと Base モデルを全て比べると，1st HateBERT，All HateBERT の時，多くの指標で Base より向上した．しかし，2nd HateBERT については，LIME 手法の Plausibility 以外の指標で，向上が見られなかった．このことから，1 段階目の根拠の学習に HateBERT を用いることで，説明可能性が向上すると考えられる．

7.2　実験 2：根拠学習の拡張の結果と考察

実験結果は，表 7，表 8 のとおりである．

検出性能では，提案手法が全ての指標で Base より向上した．このことから，クラスラベルが “Normal” の投稿の根拠も学習することで，検出性能が向上するということが言える．

しかし，バイアスについて提案手法の Sub と BNSP の指標について指標が下がる結果となった．BPSN が向上している

表 8　実験 2(根拠学習の拡張) 結果 (説明可能性)

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
根拠 3 値分類モデル	Attn	**39.3**	**64.3**	**57.0**	**57.8**	**13.8**
Base	LIME	54.7	69.3	77.3	53.3	-2.7
根拠 3 値分類モデル	LIME	**56.7**	**71.2**	**79.5**	**56.6**	**-4.0**

表 9　実験 3(マルチタスク学習) 結果 (検出性能・バイアス)

	検出性能			バイアス		
モデル	ACC	F1	AUROC	Sub	BPSN	BNSP
Base	78.4	77.4	87.5	78.7	68.6	86.2
Multi-task ($\alpha = 0.7$)	78.8	77.1	**87.9**	78.7	66.9	85.6
Multi-task ($\alpha = 0.5$)	**79.2**	**77.8**	87.8	77.7	64.8	85.4
Multi-task ($\alpha = 0.3$)	78.8	77.1	87.4	76.2	62.1	85.2

表 10　実験 3(マルチタスク学習) 結果 (説明可能性)

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
Multi-task ($\alpha = 0.7$)	Attn	**41.2**	**65.3**	57.6	**61.1**	14.6
Multi-task ($\alpha = 0.5$)	Attn	40.8	64.5	61.7	58.8	**12.1**
Multi-task ($\alpha = 0.3$)	Attn	38.6	64.0	**63.2**	56.9	12.4
Base	LIME	54.7	69.3	77.3	53.3	-2.7
Multi-task ($\alpha = 0.7$)	LIME	57.2	72.4	79.9	**57.5**	-0.5
Multi-task ($\alpha = 0.5$)	LIME	57.5	72.1	79.8	55.0	-0.8
Multi-task ($\alpha = 0.3$)	LIME	**58.3**	**72.8**	**80.1**	54.1	-1.2

ことから，偽陽性になりにくいモデルである．しかし，Sub, BPSN が下がっていることから，同じターゲットグループ内において，分類がうまくできておらず，偽陰性になりやすいモデルになっていると言える．

説明可能性について，全ての指標で向上した．このことから，根拠を 3 値分類にすることにより，説明可能性が向上するということがいえる．

7.3　実験 3 マルチタスク学習の結果と考察

実験結果は，表 9，表 10 のとおりである．

検出性能について，Multi-task の $\alpha = 0.5$ の時，すべての指標で向上した．このことから，タスクの重みが半々の時，モデルが向上することが分かった．また，ACC は，全てのモデルで向上した．そして，AUROC は α が高い時，向上していることから，ヘイトスピーチ検出のタスクに重みを置いたモデルは，ターゲット分類に重みを置いたときより性能が良くなるといえる．

バイアスについて，提案手法は，全ての指標で Base を超えることはなかった．また，Base ラインは超えていないが，α が高い時のほうが，指標が高くなっている．このことから，ヘイトスピーチ検出のタスクに重みを置いたモデルのほうが，バイアスが低減すると分かった．

説明可能性について，提案手法は，説明手法 LIME の Suff の指標以外の全ての指標で Base より向上した．このことから，マルチタスク学習を用いることで，説明可能性が向上するということがいえる．

8　ま　と　め

本研究では，BERT-MRP モデルをベースとして，BERT-base-uncased をヘイトスピーチに特化した事前学習モデルへ変更する事前学習モデルの置き換え，根拠の学習を 2 値分類から 3 値分類へ変更する根拠学習の拡張を行った．その結果，双方とも検出性能と説明可能性が向上するという結果となった．さらに，BERT-MRP モデルの 2 段階学習の 2 段階目のファインチューニングの部分を変更して，マルチタスク学習を用いた説明可能なヘイトスピーチ検出モデルを提案した．この提案手法においても，検出性能と説明可能性が向上するという結果となった．しかし，すべての提案手法で，バイアスについて向上しないという結果になった．

9　課　　題

今回，3 つの提案手法について，それぞれ個別に実験を行った．そこで，今後はそれぞれの提案手法を掛け合わせることにより，説明可能性や検出性能がさらに向上するのかについて調べていきたい．また，今回の 3 つの提案手法では向上できな

かったバイアスの観点にも注目し，検出性能，バイアス，説明可能性の 3 つの観点で向上できる方法を探っていきたい．

謝　　辞

本研究は HZKY20220326，JSPS 科研費 JP22K12271，JP19K12230 の助成を受けたものである．

文　　献

[1] Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6644–6655. International Committee on Computational Linguistics, 2022.

[2] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.

[3] Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), May 2016.

[4] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 683–693. AAAI Press, 2020.

[5] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press, 2021.

[6] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700, 2018.

[7] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 260–266. INCOMA Ltd., 2017.

[8] Ziqi Zhang, David Robinson, and Jonathan A. Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 745–760. Springer, 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[10] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer, 2019.

[11] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Hurtbert: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, pages 34–43. Association for Computational Linguistics, 2020.

[12] Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472, 2020.

[13] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5570–5581. Association for Computational Linguistics, 2019.

[14] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[15] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1884–1895. Association for Computational Linguistics, 2018.

[16] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 107–117. The Association for Computational Linguistics, 2016.

[17] Omar Zaidan, Jason Eisner, and Christine D. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 260–267. The Association for Computational Linguistics, 2007.

[18] Flor Miriam Plaza del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. Multi-task learning with sentiment emotion and target detection to recognize hate speech and offensive language. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, volume 3159 of *CEUR Workshop Proceedings*, pages 297–318. CEUR-WS.org, 2021.

[19] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM, 2019.

[20] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4443–4458. Association for Computational Linguistics, 2020.

時系列グラフニューラルネットワークのリンク予測における反実仮想説明アルゴリズム

三谷　真裕† 佐々木勇和† 鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
E-mail: †{mitani.masahiro,sasaki,onizuka}@ist.osaka-u.ac.jp

あらまし　時系列グラフニューラルネットワーク (TGNN) の研究が盛んに行われ，その予測結果を説明する技術の重要性が高まっている．TGNN における既存の説明手法はある事象が観測された理由に焦点を当てており，反実仮想説明，つまり他の事象の観測されなかった理由に焦点を当てた説明は行われていない．TGNN を用いたリンク予測における反実仮想説明では，TGNN が未観測リンクを発生すると予測する反実仮想時間グラフを探索する．反実仮想時間グラフは事実とは異なる時間グラフであり，リンクが観測されなかった理由の説明として元の時間グラフとの差分を用いることができる．本研究では，TGNN を用いたリンク予測における反実仮想説明アルゴリズムを開発する．具体的には元の時間グラフから複数リンクの削除により反実仮想時間グラフを見つけ，削除したリンクを目的のリンクが観測されなかった説明として活用する．提案手法ではモンテカルロ木探索により削除リンクの候補を効率的に探索する．実データを用いた実験により，リンクの無作為な削除や近傍の削除と比較して，提案手法が短い実行時間かつ同等の精度で反実仮想時間グラフを発見することを示す．

キーワード　XAI，Interpretable ML，グラフ深層学習

1. はじめに

時間グラフはイベントの集合から構成される．イベントは 1 つが相互作用する 2 つのノードとタイムスタンプ, 特徴量の組で表現される [1]．このようなデータ構造は, ソーシャルネットワークにおけるフォロー関係や, 消費者の購買履歴, 道路の交通量のデータなど, 実世界の様々な場面で見かけることができる．

時間グラフデータに対し, その特性を考慮した機械学習モデルである, 時系列グラフニューラルネットワーク （TGNN） が提案されている [2], [3]．TGNN は, 将来に発生するイベントを予測するリンク予測や将来のノードの状態を予測するノード分類など様々なタスクの研究が行われ, 実世界で多くの応用がなされている．例えば, EC サイト等の購買データに基づくリンク予測は, ソースノードを消費者とし, ターゲットノードを商品群とした過去の購買データから, その人が次に買う商品を予測するタスクとなる．

TGNN は様々なタスクにおいて高い精度を達成しているが, 他のニューラルネットワークと同様にブラックボックスであり, なぜその結果が出力されたのかを知ることができない．もしリンク予測の予測結果が出力された理由を説明することができれば, ユーザが商品を購入する状況を理解するなど, 有意義な情報を手にすることができるため, TGNN の説明可能性の研究が行われている [4]．既存の説明の研究は, 1 つの予測結果に対し, TGNN の推論の根拠となったノードやエッジなどの入力の一部を提示する方法であり, T-GNNExplainer [5] や DyExplainer [6] などの手法がある．これらの手法は説明可能性の分野において, 事実の説明 (factual explanation) と呼ばれる説明に該当し, TGNN モデルの予測結果に対して最大限の影響がある入力の部分集合を見つけることを目的としている．予測の説明としてノードやグラフの部分構造によって説明が行われる．

だが, その説明から分かることは実際に起こった予測対象についてであり, 実際には起こらなかった事象について知ることはできない．事実とは異なる事象についての説明は反実仮想説明 (counterfactual explanation) と呼ばれ, TGNN の予測結果に変化をもたらす入力となる時間グラフの最小の変化を説明として用いることにより, 事象が起こらなかった原因について説明をする．

我々の知る限り, 時間グラフを対象とした TGNN に対して, この反実仮想説明の方法を提案した研究は存在しない．研究が行われていなかった要因は 2 つ考えられる．1 つは, 時間グラフを対象とした GNN の事実の説明を行った研究 [5], [6] は近年行われたばかりで, 事実と異なる事象を想定する反実仮想説明を取り組むのはより難しく研究に取り組めなかったためと考えられる．2 つ目は, 静的グラフと比べて候補数が多くなることが挙げられる．静的グラフでは, 説明の対象となるグラフのノードやエッジがグラフ考慮すべき数が決まるが, 少なくとも動的グラフでは総タイムスタンプ数分のグラフのノードやエッジを考慮しなければならないため, 候補数がより膨大となる．

本研究では, 反実仮想説明を TGNN のリンク予測において取り組む．TGNN のリンク予測における反実仮想説明では, 事実と異なるイベント e_c の 発生について説明を行う．e_c は実際には起こっていないため, TGNN に元の時間グラフを入力して e_c を予測しても発生は予測されない．元のグラフに変更を加えることにより, e_c の発生が予測されるグラフを見つけることが

できれば, 元のグラフと変更後グラフの差分から e_c の発生をしなかった理由について説明を行うことができる.

このとき, 事実と異なるイベントのことを反実仮想イベントと呼び, 予測に用いられた入力グラフを反実仮想時間グラフと呼ぶこととする. さらに, 反実仮想説明のグラフの探索では事実と異なる部分は少ないことが求められ, 元の時間グラフからのイベントの変更を最小限に留めることが求められる.

グラフの探索において考慮すべきイベントの候補数が膨大となる問題に対し, 我々は候補とするイベントを限定するため, 候補とするイベントを元グラフへ追加や削除を考えるのではなく, 元グラフの近傍からの削除のみを考えることとした. 削除のみを考える場合, 反実仮想グラフは元グラフの部分集合となる. そのような反実仮想時間グラフを元グラフからの最小限の変更で探索するため, 既存の TGNN の説明手法 T-GNNExplainer をベースとしたフレームワークを提案する. T-GNNExplainer をフレームワークに組み込むのは, T-GNNExplainer が入力のグラフからその部分集合を探索するといった, 今回のグラフの探索と同様の組合せ最適化問題に取り組んでいることが理由である.

T-GNNExplainer のアーキテクチャを用いることで, 候補イベント集合の中から削除すべきイベントを効率的に探索できる. だが, そのままの適用ではベースラインと同等の精度しかでない. そのため, T-GNNExplainer のパラメータと探索範囲を自動的に最適化する方法を提案する. これにより, 制限時間内においてフレームワークの反実仮想イベントの予測精度が向上することを示す.

実験では, 2 つの TGNN モデルと 2 つの実データセットにおいて比較をし, フレームワークを評価した. TGNN のリンク予測における反実仮想説明を行った研究はないため, 反実仮想時間グラフをルールベースで探索する方法と比較を行う. それらの方法と比較して, 同等かそれ以上のグラフの探索精度を達成することを示す.

本稿の構成は以下の通りである． 2 章にて関連研究について説明し， 3 章にて事前知識と問題定義について説明する．4 章にて提案フレームワークについて説明する． 5 章にて評価実験について述べ， 6 章で本稿をまとめる.

2. 関 連 研 究

時間グラフはグラフの観測の時刻が一定の感覚で区切られている離散時間動的グラフ（DTDG）と, ノードやエッジの変化が連続的な時間スケールで起こる連続時間動的グラフ（CTDG）がある [1]. 本稿では時間グラフとして連続時間動的グラフを考え, タイムスタンプつきのイベント集合を入力する時系列グラフニューラルネットワークを対象とする. 時間グラフにおけるリンク予測では, 過去の一連のタイムスタンプで観測されたグラフ構造に基づいて将来のリンクを推測する予測を指す. TGNN ではエンコーダで各タイムスタンプにおけるノード埋め込みを取得し, ノード埋め込みを入力したデコーダからリンク発生を予測する. TGAT は, 対象ノードの近傍ノードの情報を時系列順に集約する Attention 層と, 時間情報をエンコードするためのタイムエンコーダーを使用している. TGN は TGAT の構成要素に加え, 各ノードの一時的な状態を保持するメモリとメモリ上のノードの状態を更新するモジュールを使用している.

また, グラフニューラルネットワーク (GNN) の予測結果の説明は, 事実に基づく説明と反実仮想説明の 2 つの種類がある [4]. 事実に基づく説明は出力に影響を与える最小限の入力集合の形で説明を見つけることを目的としている. 入力集合はイベントやノード, サブグラフなどを指しており, それらが摂動されて予測結果が変動する場合に, GNN の推論に影響がある入力として, 説明に使用される. また, 事実に基づく説明の方法は学習ベースと探索ベースの 2 つがある. 学習ベースの方法 [7] は, 学習済み GNN によって生成されたノード表現を活用し, 重要なノードやエッジを学習するためのニューラルネットワークを採用する. これらは複数のインスタンスで訓練されており, 素早く説明結果を出力できる. 探索ベースの方法 [8] は, スコア関数を用いた探索アルゴリズムを利用して, 入力の中で重要な部分集合を見つけることができる. しかし, 各インスタンスの探索空間が異なるため, 実現可能な解を一つ一つ探索する必要があり, 推論時間が長くなってしまう. また, TGNN の説明の手法 [5] では, 学習ベースと探索ベースの手法が組み合わされている.

一方で反実仮想説明では, GNN の出力に変化をもたらす, 最小限の入力の変化の形で説明を見つけることを目的としている. ノードの特徴やグラフ構造を利用して元のグラフを変動させ, ユーザに望ましい結果のモデルでの予測の出力を目指し, グラフ変動部分を説明として用いる. 反実仮想説明の方法は摂動ベースと生成ベースがある. 摂動ベース [9] では, バイナリ行列を用いて計算グラフに摂動を与え, 損失関数によって理想的な予測値との誤差を最小化する. 生成ベース [10] では, ニューラル・アーキテクチャを用いて反事実グラフを生成し, 理想的な予測値の差を最小化する. だが, これらの手法は, 静的グラフが対象とされており, 時間グラフにおける反実仮想の説明を行った研究は我々の知る限り存在しない.

3. 事前準備と問題定義

3.1 時系列グラフニューラルネットワークにおけるリンク予測

時間グラフを $\mathcal{G} = (\mathcal{N}, \mathcal{S})$ とする. 時間グラフ内のイベント集合は $\mathcal{S} = \{e_1, e_2, ..., e_T\}$ と表され, イベント集合 $\mathcal{S}$ において存在するノードの集合を $\mathcal{N}$ とし, ソースノードを $\mathcal{N}_u$ とターゲットノードを $\mathcal{N}_v$ と表す. また, 総イベント数は T とする.

個々のイベント $e_i = \{u_i, v_i, t_i, att_i\}$ は, タイムスタンプ t_i において, ソースノード $u_i \in \mathcal{N}_u$ からターゲットノード $v_i \in \mathcal{N}_v$ に対し, エッジ属性が att_i となるようなリンクが発生していることを表す. TGNN のリンク予測における入力となるタイムスタンプ k までの時間グラフ $\mathcal{G}^k = (\mathcal{N}^k, \mathcal{S}^k)$ は, イベント集合を $\mathcal{S}^k = \{e_1, e_2, ..., e_{k-1}\}$ とし, 存在するノードの集合を $\mathcal{N}^k$ とする.

使用する TGNN [2], [3] は, エンコーダ・デコーダモデルを採用している. 訓練済みの TGNN を $f(\cdot)$ とし, 予測するリンクを e_k としたとき, $f(\mathcal{G}^i)[e_i]$ は, エンコーダを使用して $\mathcal{G}^i$ における反実仮想イベントのノード埋め込みを計算し, それらをデコーダに入力してイベント e_i をリンク予測した出力値を表す.

3.2 TGNN のリンク予測における反実仮想説明

TGNN のリンク予測における反実仮想説明では, 反実仮想時間グラフ $\mathcal{G}^c$ から反実仮想イベント e_c が TGNN のリンク予測により予測された場合に, グラフ $\mathcal{G}^c$ と元の時間グラフ $\mathcal{G}^k$ の差分を説明として用いる.

元グラフ $\mathcal{G}$ のタイムスタンプ t_k において事実と異なる任意のイベントを反実仮想イベント e_c と呼ぶ. また, TGNN に入力した際に反実仮想イベントの発生が予測される時間グラフを反実仮想時間グラフ $\mathcal{G}^c$ と呼び, そのようなグラフの集合を G^C と定義する. グラフ $\mathcal{G}^c$ は TGNN に入力し事実と異なるイベントが予測されるため, 元の事実のグラフと違うグラフとなる. 2 つのグラフ $\mathcal{G}^c$ と $\mathcal{G}^k$ は TGNN により片方ではイベント e_c の発生が予測され, もう一方では発生が予測されないため, その差分を調べることがイベント e_c が元のグラフにおいて発生しなかった理由を説明することにつながる.

さらに, グラフ $\mathcal{G}^c$ は事実と異なるグラフであるため, 事実との差分が大きいほど, 非現実的なデータとなってしまう. そのため反実仮想説明では, グラフ $\mathcal{G}^c$ と $\mathcal{G}^k$ のグラフ空間上における距離が出来るだけ小さいことが求められる.

3.3 問 題 定 義

TGNN のリンク予測における反実仮想説明を行うため, タイムスタンプ t_k において TGNN が反実仮想イベント e_c を予測可能な反実仮想時間グラフ $\mathcal{G}^c$ を探索する. 反実仮想時間グラフとなり得るグラフは無数に存在することが考えられ, そのグラフの候補となる数はノードの数 $|\mathcal{N}|$ に対して $O(2^{|\mathcal{N}|})$ である. 問題を絞るため本研究では元の時間グラフ $\mathcal{G}^k$ のイベント集合 $\mathcal{S}^k$ の部分集合となるイベント集合 $\mathcal{R}^c$ を見つけることを考える. したがって, イベント集合 $\mathcal{R}^c$ は $\mathcal{S}^k$ 内の複数イベントを削除した集合を表す. また, $\mathcal{R}^c$ 内で存在するノードの集合 $\mathcal{N}^c$ によってできるグラフが反実仮想グラフ $\mathcal{G}^c = (\mathcal{N}^c, \mathcal{S}^c)$ となるように削除するイベントを選ぶ.

また, 元グラフと反実仮想グラフ間ののイベント集合の差分はできるだけ小さくする. 具体的な目的関数は以下の通りとなる.

$$\begin{gathered} \min_{\mathcal{G}^c \in G^C} |\mathcal{G}^c - \mathcal{G}^k| \\ G^C = \{\mathcal{G}^c = (\mathcal{N}^c, \mathcal{R}^c) |\ f(\mathcal{G}^c)[e_c] \geqq 0.5\} \end{gathered} \tag{1}$$

ここで, G^C を反実仮想グラフの集合とし, 個々の反実仮想グラフ $\mathcal{G}^c$ は次の章で提案するフレームワークやベースライン $g(\cdot)$ によって作られるグラフ $\mathcal{G}^c = g(e^c, \mathcal{G}^c, f(\cdot))$ とする. 関数 $g(\cdot)$ は反実仮想イベント, 元グラフ, TGNN を元に, 反実仮想グラフ探索して出力する関数である.

4. 提案フレームワーク

本章では, TGNN のリンク予測における反実仮想説明を行うフレームワークについて詳細を述べる. 本フレームワークの全体像は図 1 に示す. フレームワークでは事実に基づく説明手法 T-GNNExplainer [5] を利用し, 反実仮想グラフを探索する. 本論文での反実仮想グラフのイベント集合は元のグラフのイベント集合の部分集合であるため, グラフの探索を組合せ最適化問題として考えることができる. そのため, 同様の組合せ最適化に取り組んでいる T-GNNExplainer のアーキテクチャを利用する.

以下では まず, T-GNNExplainer の概要と本フレームワークの違いを述べ（4.1 節）, 次に T-GNNExplainer のナビゲータをについて述べる（4.2.1 項）. さらに T-GNNExplainer のエクスプローラを利用して候補イベントの中から削除イベントを探索する方法について述べる（4.2.2 項）.

また, T-GNNExplainer には実行回数と説明の候補を決めるパラメータが存在し, それらは T-GNNExplainer のアーキテクチャにおいて変更されない. パラメータの変更によりグラフの探索範囲が広がり, より良い結果が得られると考えるため, 本フレームワークではパラメータを一意に決める方法を 4.3 節で提案する. また, フレームワークの処理手順をアルゴリズム 1 に示す.

Algorithm 1 提案フレームワーク

Input: Temporal graph $\mathcal{G}^k$,TGNN f, couterfactual event e^c, Pretrained Network h_θ, rollout N, candidate D
Output: counterfactual graph $\mathcal{G}^c$

$tmp = 0$
$count = 0$
while $prob < 0.5$ **do**
  $S^{can} = NewCandidate(\mathcal{G}^k, D)$
  $\mathcal{G}' = TGNNExplainer(f, \mathcal{G}^k, e^c, S^{can}, h_\theta, N)$
  $prob = f(\mathcal{G}')[e^c]$
  if $tmp \geqq prob$ **then**
    if $count > 1$ **then**
      $D = D \times 2$
      $count = 0$
    else
      $N = N + 500$
      $count = count + 1$
    end if
  else
    $D = D \times 2$
    $tmp = prob$
  end if
end while
Return $\mathcal{G}'$

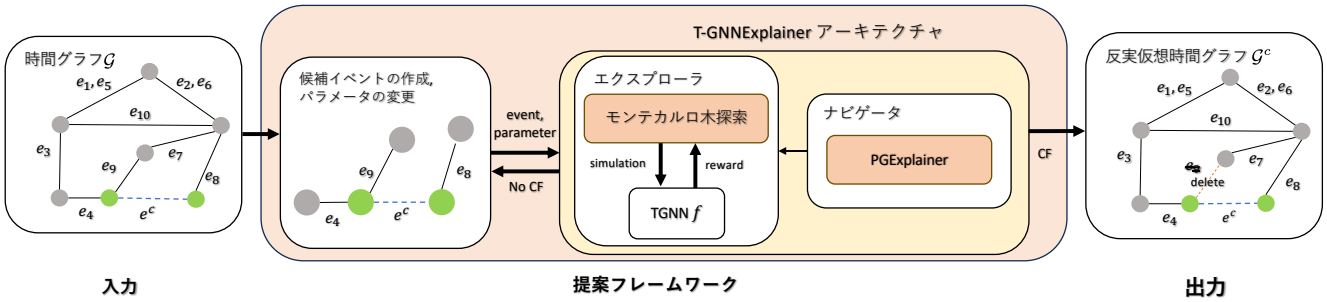


図 1　提案フレームワークの概要: CF は反実仮想グラフ (CounterFactual graph) を表す

4.1　T-GNNExplainer の反実仮想説明タスクへの適用

事実の説明手法 T-GNNExplainer は, TGNN による対象イベント e_k のリンク予測の結果に寄与した入力イベントの集合 R^k を特定することで, TGNN の推論を説明する. R^k は元グラフのイベント集合 G^k の部分集合であるため, 組合せ最適化問題となる. T-GNNExplainer はナビゲータとエクスプローラからなるアーキテクチャを用いて, 重要なイベント集合 R^k を効率的に特定することができる.

本フレームワークでは, TGNN に入力すると反実仮想イベントの発生が予測される反実仮想グラフの探索にこのアーキテクチャを活用する. まず, リンク予測の対象イベント e_k を反実仮想イベント e_c に変更する. 次に, 反実仮想イベントの TGNN での出力結果をリンクの発生が予測されるようにグラフを置き換える. 最後に, アーキテクチャを用いて反実仮想イベント e_c の発生が予測される場合に重要とな入力イベントの集合を特定し, 候補イベントの集合で重要となるイベントに選ばれなかったものを元グラフから削除されたイベントであると見做す. このような適用を行うことで, TGNN によって反実仮想イベントが予測される場合の反実仮想グラフの探索と, 同時に削除イベントの集合を特定することができる.

4.2　T-GNNExplainer のアーキテクチャ

T-GNNExplainer のアーキテクチャはナビゲータとエクスプローラーで構成される. ナビゲータは対象イベントと他イベントの関連度を捉える. エクスプローラはナビゲータから取得したイベント間の関連度を利用し, 重要なイベントを特定する. また, エクスプローラはモンテカルロ木探索を実行し, 学習済みの TGNN モデルから報酬を獲得し, 最適なイベント集合を探索する.

4.2.1　ナビゲータ

ナビゲータは既存の説明手法 PGExplainer [7] を使用し, イベント間の関連性の理解のためにナビゲータを訓練する. 具体的には, ナビゲータはフィードフォワードニューラルネットワーク $h(e_j, e_c)$ であり, 元グラフ内のイベント e_j と反実仮想イベント e_c との関連度合いを出力する. この出力値は, 4.2.2 項のノード展開で探索するイベントを決めるために使用される.

目的関数. ナビゲータの訓練には, 各イベントのノード特徴とエッジ属性の特徴ベクトルを使用する. K は元グラフ内からサンプルした, TGNN の予測の対象となるイベントの個数とする. Y_k は, TGNN の出力 $f(\mathcal{G}^k)[e_k]$ の出力から決定され, 0 か 1 を出力する. また, $\mathbf{1}(\cdot)$ を $(\cdot)$ が成り立つ時に 1 を出力し, そうでなければ 0 を出力する関数とする.

これらを元に [7] と同様に目的関数をいかに定義する.

$$\arg\min -\frac{1}{K}\sum_{k=1}^{\mathcal{K}}[\mathbf{1}(Y^k=1)\log\sigma(f(R^k)[e_k])+ \mathbf{1}(Y^k=0)\log(1-\sigma(f(R^k)[e_k'])] \tag{2}$$

$R^k = g(e_k, \mathcal{G}^k, f(\cdot))$ であり, $g(\cdot)$ のパラメータが訓練される.

ナビゲータの出力. PGExplainer にて使用された再パラメータ化のトリックを用いて目的関数を達成するようにナビゲータを訓練する. 入力には, 各イベントのノード特徴とエッジ属性の特徴ベクトルを使用し, 出力はグラフ $\mathcal{G}^k$ のイベント集合の個々のイベントに重みを割り当てるソフトマスクとみなせる. よって, ナビゲータ $h(e_j, e_c)$ の出力は, 個々の候補イベントと反実仮想イベントの関連度合いとなる.

4.2.2　エクスプローラ

エクスプローラではモンテカルロ木探索を用いて, 反実仮想グラフを見つけるために元グラフから削除するイベントを列挙する. まず, モンテカルロ木探索で探索の対象となる, 候補イベントの集合を決定する. 次に, 探索木を展開するための試行（ロールアウト）が複数回に分けて行われ, 探索木の子ノードのイベント集合は, 削除する候補イベントの集合となる. 最後に, 探索木の中で最良の結果となるノードを説明結果する.

初期化. 探索木の初期ノードは, 空間的・時間的条件を満たす候補イベントの集合とする. 候補イベントが初期ノードに含まれる空間的条件を, TGNN のエンコーダが 2-hop のイベントの集約を行っているため, 2-hop 以内のイベントであることとする. また, 時間的条件は, 空間的条件を満たすイベントの中から反実仮想イベントのタイムスタンプ k から直近の q 個以内のイベントとする.

選択. $\mathcal{N}^i$ を探索木内のノードを表し, e_j を一つの試行, すなわち, $\mathcal{N}^i$ のイベント集合からイベント e_j を破棄することと定義する. ノード選択では, [11] で提案された UCT (Upper Confidence bound applied to Trees) に従い, 探索と活用のバランスを取る. 選択基準は以下とする.

$$e^* = \underset{e_j \in C(\mathcal{N}^i)}{\arg\max} \left(\frac{c(\mathcal{N}^i, e_j)}{n(\mathcal{N}^i, e_j)} + \lambda \frac{\sqrt{\sum_{e_l \in C(\mathcal{N}^i)} n(\mathcal{N}^i, e_l)}}{1 + n(\mathcal{N}^i, e_j)} \right) \quad (3)$$

ここで, $C(\mathcal{N}^i)$ は $\mathcal{N}^i$ に存在するイベント, $n(\mathcal{N}^i, e_j)$ は過去のロールアウトでノード $\mathcal{N}^i$ で e_j を選択した回数である.

$c(\mathcal{N}^i, e_j)$ はノード $\mathcal{N}^i$ で e_j を選択した累積報酬を示している. 最初の項は探索と利用の利用に相当する部分であり, 平均報酬が高いノードを選択する. 二つ目の項は探索に相当し, 探索数が少ないノードを選択する.

展開. 子ノードの展開の方法は, 性能に大きく影響する. 以前のモンテカルロ木探索を使用した説明の研究 [8] では, 選択されたノードに対して可能性のあるすべての子ノードを展開した. だが, 本エクスプローラでは探索空間を洗練させるために関連性の低いノードのみを展開する. 選択されたノード $\mathcal{N}^i$ が展開可能であると仮定すると（すなわち, 子の数がノードに含まれるイベントの数より少ない場合）, 探索者はナビゲータ（4.2.1 節）を呼び出して関連度を計算する

$$e^* = \underset{e_j \in \mathcal{N}^i / C(\mathcal{N}^i)}{\arg\max} h(e_j, e_c) \quad (4)$$

この式において, $\mathcal{N}^i / C(\mathcal{N}^i)$ は, 前のロールアウトで展開されなかったイベントを意味し, 新しい子ノードを展開するために最も重要でないイベント e^* を削除する. ナビゲータは学習済みでスコアを素早く推論できるため, 追加コストは無視できる. また, 上位 k 個の候補を展開したり, 展開する候補にランダム性を加えたりできる. 例えば, e^* を確率 $1-\epsilon$ で拡張し, 未探索の e_j を確率 ϵ で選択するといったことができる.

ノードの選択と展開は交互に行われる. 探索は初期ノードから開始され, 式 2 に従って新しい子ノードを拡張する. 次に, ルートの元の子ノードと新しい子ノードから, 式 3 に基づいて最も高い値を持つ子ノードを選択し, そのノードに移動する. 次に, 新しいノードから展開と選択を繰り返す. 現在のノードが葉ノードであることが確認された場合, 探索は終了する.

報酬とバックプロパゲーション. 報酬は, 時間グラフモデルによって計算される. 具体的には, 葉ノード $\mathcal{N}^{leaf}$ のイベント集合を $R^k = \mathcal{N}^{leaf}$ として式 2 の損失関数を計算し, 葉ノードの報酬 $r(\mathcal{N}^{leaf})$ に代入する.

また, バックプロパゲーションでは初期ノードから葉ノードまでの経路にあるノード $\mathcal{N}^i$ は, 葉ノードの報酬の加算 ($c(\mathcal{N}^i, e_l) = c(\mathcal{N}^i, e_l) + r(\mathcal{N}^{leaf})$) とノードの訪問回数の加算 ($n(\mathcal{N}^i, e_l) = n(\mathcal{N}^i, e_l) + 1$) を行う.

4.3 再起的な実行

本項では, 4.2 項で説明した T-GNNExplainer のアーキテクチャを再起的に実行する方法について説明する. フレームワークの再起的な実行必要な理由として, パラメータが固定されていることにより探索されるイベントやその組み合わせが限定されることが挙げられる. また, 削除の対象となるイベント集合は, 反実仮想イベントに対して時間的に近いものから順に取得しているため, 遠くイベントを削除した場合を考慮できないという問題がある. そのため, 一度目の実行において反実仮想イベントの発生が予測されなかった場合, 削除の対象とするイベント集合を違うものへ変更するする. そうすることでさらに多くのイベントの削除を考慮でき, より多くの反実仮想グラフを見つけることが可能となる.

アークテクチャのパラメータには, モンテカルロ木探索の探索回数を決める rollout と, 探索の候補となるイベントの候補の数を決める candidate がある. アーキテクチャが探索するイベントは, 実際の予測対象となる反実仮想イベントの 2hop 以内のイベントから candidate の個数分を反実仮想イベントのタイムスタンプ t_c に近いものから順に選ばれる. アーキテクチャにおいて rollout と candidate の初期値はそれぞれ 500 と 20 に設定しており, これは T-GNNExplainer [5] における初期値と同様に設定している. この設定において予測ができない場合に, パラメータを変更して再度実行を行う.

以下の手順をリンク発生の予測値 $f(\mathcal{G}^i)[e_c]$ が 0.5 以上となるまで繰り返す. まず, T-GNNExplainer のアーキテクチャを実行し, リンク発生の予測値 $f(\mathcal{G}^i)[e_c]$ を確認する. $f(\mathcal{G}^i)[e_c] < 0.5$ であれば, candidate の値を 2 倍し再度実行を行う. このとき, 前回の実行の値もしくは初期の値よりも予測値が低ければ, rollout に 500 を加算して再度実行を行う.

表 1 データセットの概要

	ソース	ターゲット	イベント数
Wikipedia	8,227	1,000	157,474
Reddit	10,000	984	672,447

5. 実　　験

本章では, 簡易的なベースラインとの比較を通して, 提案フレームワークにより探索された反実仮想グラフにおける反実仮想イベントの予測精度を評価する.

5.1 データセットと使用モデル

実験では Wikipedia (注1) と Reddit(注2) の 2 つの実データセットを使用する. この実データの詳細を表 5.1 で示す. Wikipedia, Reddit 両方において, ソースノードの集合がページを編集するユーザとなり, ターゲットノードの集合が編集がなされたページやアイテムとなる 2 部グラフのデータである. ユーザがページを編集した時刻がタイムスタンプとなる. 使用する TGNN モデルは, TGAT と TGN [2], [3] とする.

5.2 ベースライン

我々の知る限り, TGNN のリンク予測における反実仮想説明に取り組んだ研究は未だ存在しない. そのため, 提案アルゴリズムの比較のためにルールベースでイベントを削除を行うベースラインを用いる. ルールベースの削除には, イベントをラン

(注1)：http://snap.stanford.edu/jodie/wikipedia.csv
(注2)：http://snap.stanford.edu/jodie/reddit.csv

ダムに削除する方法と反実仮想イベントの近傍となるイベントを削除する方法を考える. 前者は削除の候補とするイベントをランダムにサンプルする一方で, 後者では反実仮想イベントから 1-hop 以内もしくは 2hop 以内のイベント集合からサンプルする.

最初に削除の候補となるイベント集合 $\mathcal{S}^{can}$ を定義する. この時, ランダムのイベントの候補数は反実仮想イベントの 1-hop のイベント数と同じとする. 次に S^{can} の中から 1 つ削除することにより最も予測値 $f(\cdot)[e^c]$ を上昇させるイベント e_{rm} を得る.

$\mathcal{S}^{can}$ から e_{rm} を除外し, 再度その集合から削除すると最も予測値を上昇させるイベント e'_{rm} を削除する. この操作を $f(\cdot)[e_c]$ がリンクの発生を予測する閾値を超える, もしくは予測値が上昇しなくなるまで繰り返す.

5.3 実験設定

TGAT と TGN はタイムスタンプに基づいてデータセットを 70%, 15%, 15% に分割し, 訓練した.

説明の対象となる反実仮想イベントは, TGNN モデルとデータセットの組み合わせのそれぞれにおいて正しく予測されるイベント e_i を変更することで作成した. タイムスタンプ t_i を変えず, ソースノードとターゲットノードはそのタイムスタンプにおいて存在しているそれぞれのノードの集合からランダムにサンプルした. また, サンプルされた反実仮想イベント e_c は, 元グラフにおける $f(\mathcal{G}^k)[e_c]$ で予測されないことを条件とした.

T-GNNExplainer では T-GNNExplainer での実験設定を利用し, モンテカルロ木探索によって探索される候補イベントの数は直近の 20 個に制限し, ロールアウトの回数を 500 回とした. ナビゲータは 128 個のニューロンからなる隠れ層を持つ 2 層のニューラルネットワークを使用する.

2 つの TGNN モデル (TGAT,TGN) と 2 つの実データ (Wikipedia,Reddit) をそれぞれ使用する 4 パターンにおいて, ベースライン, T-GNNExplainer, 提案フレームワークそれぞれで実験を行う. 実行時間を揃えるため, 全ての方法に対して 1 つの仮想イベントに対し実行時間が 1 時間を超えたものは, 予測されなかったものとする.

5.4 評価指標

評価指標は反実仮想の説明手法 [10] に倣い, Validity と Proximity, 実行時間を使用する. Validity は手法が反実仮想イベントを予測することができた割合を示す.

$$Validity = \frac{1}{E}\sum_{i=[E]} \mathbf{1}(f(\mathcal{G}^c)[e_i] > 0.5) \tag{5}$$

E は対象とした反実仮想イベントの数とする. $\mathbf{1}(\cdot)$ は $(\cdot)$ が成り立つ時に 1 を, そうでなければ 0 を出力する関数とする.

Proximity は元の時間グラフと反実仮想グラフの近さを測定するために, 元グラフから削除されたイベントの数と定義する.

$$Proximity = \frac{1}{N}\sum_{i=[N]} |\mathcal{R}_i^c - \mathcal{S}_i^k| \tag{6}$$

N は反実仮想グラフ $\mathcal{G}^c$ が存在する反実仮想イベントの数とする. また, $|\cdot|$ 内は $\mathcal{G}^c$ におけるイベント集合 $\mathcal{R}_i^c$ と元グラフのイベント集合 $\mathcal{S}_i^k$ のイベントの差分を表現する.

さらに, それぞれの方法の実行時間も測定する. Validity, 実行時間は全ての反実仮想イベントにおける平均とする. Proximity は反実仮想イベントが予測されたものの中における平均とする.

5.5 反実仮想説明の実験結果

提案フレームワークとベースラインそれぞれに対し, TGAT, TGN と Wikipedia, Reddit の 4 パターンでの実行を行い, Validity, Proximity, 実行時間で評価した結果を表 2 に示す.

提案フレームワークはベースラインと比較し, Wikipedia において最大の Validity を達成していた. また, Reddit の場合においても 2 番目に高い. Validity が高い理由として, 提案手法が他のイベントと同時に削除を行っているため, 反実仮想グラフを発見できていることが考えられる. また, 1-hop 以内のイベント全てを候補とした方法と比べて Validity が低い理由は, 扱うイベント数の差が考えられる. 1-hop の候補イベント数は膨大であるが, 提案手法は実行時間の関係で最大で 80 個しか候補イベントとならず, 削除してグラフを探索する範囲が小さかったためと考えられる.

一方, 提案手法の Proximity はベースラインより大きい. これは, 探索範囲が大きくなったためだと考えられる. ベースラインは残っている候補イベントから 1 つを削除することのみを考える一方で, 提案手法や T-GNNExplainer では候補イベントの中から一つを削除した上で他のイベントを更に削除した場合も一度で考えている. そのため, 提案手法は多くの削除のパターンを探索でき, 削除したイベントが多く反実仮想グラフが増えたことで Proximity が増加していると考えられる.

ベースラインとして使用した手法において, データセットが Wikipedia の場合では 1-hop のみのイベントを用いる場合よりも 2-hop 以内のイベントを用いる方が良い結果となる. 一方で, Reddit においては 1-hop のみの実行結果の方が良い. これは, データの総イベント数が多いため, 候補イベントが多くなることや実行時間が増え, 1 つの候補イベント当たりの実行する数が減っているためである.

5.6 再起的実行の性能調査

提案手法では, T-GNNExplainer のパラメータを変更して実行を行なっている. そのため, パラメータを変更することによってより良い Validity と Proximity を得ることが可能か, また, 得られる場合どの程度まで向上を見込めるのかを調査する. パラメータを変更して実行を行った結果を表 3 に示す.

表からは candidate と rollout 両者において値を増加させると Validity も向上することが分かる. 一方 Proximity は candidate の値が大きくなるにつれ悪化しており, これは候補となるイベントが増えたことによってより多くのイベントの削除を行なった反実仮想グラフを考慮することができるようになったた

表 2　実行結果: Validity は高い方が良く, Proximity は低い方が良い.

	Wikipedia						Reddit					
	TGAT			TGN			TGAT			TGN		
	Validity	Proximity	実行時間	Validity	Proximity	実行時間	Validity	Proximity	実行時間	Validity	Proximity	実行時間
Random	0.01	0.06	121.09	0.00	0.00	127.17	0.00	0.00	1126.05	0.00	0.00	1090.99
1-hop	0.71	5.21	240.06	0.46	4.86	230.48	**0.24**	8.12	2077.31	**0.12**	6.32	1731.37
1,2-hop	0.76	5.17	873.80	0.46	5.48	994.84	0.07	1.26	3231.14	0.03	2.24	3146.08
T-GNNExplainer	0.62	7.64	207.15	0.45	5.44	156.44	0.06	10.16	327.30	0.05	4.60	245.30
提案手法	**0.82**	13.17	688.84	**0.71**	12.80	837.53	0.17	20.94	2633.93	0.06	4.16	2797.97

表 3　T-GNNExplainer のパラメータ変更の影響調査

		Wikipedia						Reddit					
		TGAT			TGN			TGAT			TGN		
candidate	rollout	Validity	Proximity	実行時間	Validity	Proximity	実行時間	Validity	Proximity	実行時間	Validity	Proximity	実行時間
20	250	0.61	9.00	170.03	0.44	6.20	116.54	0.06	11.50	201.02	0.05	5.00	163.07
	500	0.62	7.64	207.15	0.45	5.44	156.44	0.06	10.16	327.30	0.05	4.60	245.30
50	250	0.81	21.07	348.13	0.75	16.26	161.16	0.12	28.91	449.28	0.06	13.83	400.13
	500	0.83	20.69	535.66	0.75	15.81	400.35	0.14	29.57	1113.28	0.06	12.50	711.54
100	250	**0.87**	25.34	421.68	**0.80**	19.71	359.64	0.31	62.22	1398.02	**0.11**	46.00	1055.31
	500	**0.87**	24.48	605.92	**0.80**	18.48	470.06	**0.36**	64.66	2107.20	**0.11**	44.45	1583.81

めだと考えられる．また, Proximity は candidate の値が同じで rollout の値が異なる場合では, rollout が大きい方が Proxity で良い結果を残しており, これはより良い候補を rollout を増やしたことにより探索できたためだと考えられる．

6. おわりに

本稿では，時系列グラフニューラルネットワークのリンク予測における反実仮想説明のための新しい探索フレームワークを提案した．反実仮想時間グラフを探索するため, 元グラフのイベント集合から, T-GNNExplainer を適用して不要なイベントを見つけ出した．実験では, 実世界のデータセットにおいて, 提案手法が反実仮想グラフを効率的かつ効果的に探索できることを確認した．

今後の課題として, より精度を高めるため, パラメータを調整する方法を調べる．さらに, 説明手法で探索ベースの手法や深層学習ベースの手法を用いることでより高精度な説明を目指す． また, 使用する TGNN の予測精度が説明の信頼性に直結するため, より精度の高い TGNN モデルの使用を試みることを今後の課題とする．

謝　辞

本研究は JST さきがけ JPMJPR21C5 および JSPS 科学研究費 JP20H00583 の支援よって行われた．

文　献

[1] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs: A survey. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 2648–2720, 2020.

[2] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*, 2020.

[3] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

[4] Jaykumar Kakkad, Jaspal Jannu, Kartik Sharma, Charu Aggarwal, and Sourav Medya. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*, 2023.

[5] Wenwen Xia, Mincai Lai, Caihua Shan, Yao Zhang, Xinnan Dai, Xiang Li, and Dongsheng Li. Explaining temporal graph models through an explorer-navigator framework. In *The Eleventh International Conference on Learning Representations*, 2022.

[6] Tianchun Wang, Dongsheng Luo, Wei Cheng, Haifeng Chen, and Xiang Zhang. Dyexplainer: Explainable dynamic graph neural networks. *arXiv preprint arXiv:2310.16375*, 2023.

[7] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, pp. 19620–19631, 2020.

[8] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning*, pp. 12241–12252, 2021.

[9] Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 4499–4511, 2022.

[10] Jing Ma, Ruocheng Guo, Saumitra Mishra, Aidong Zhang, and Jundong Li. Clear: Generative counterfactual explanations on graphs. *Advances in Neural Information Processing Systems*, Vol. 35, pp. 25895–25907, 2022.

[11] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293, 2006.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 10:00 ~ 12:10 | T1-B オンライン（Zoom Events)

対話データ処理

座長:酒井 哲也(早稲田大学)
コメンテータ:宮森 恒(京都産業大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T1-B-4-01] SNSを用いたChatGPTの不安感情の特定と解消法の検討

*鍜示 晃大[1]、王 元元[1] (1. 山口大学)

10:20 ~ 10:45
[T1-B-4-02] ラップバトルにおけるアンサーの類型化および類型の自動分類

*三林 亮太[1]、佃 洸摂[2]、渡邉 研斗[2]、中野 倫靖[2]、後藤 真孝[2]、山本 岳洋[1]、大島 裕明[1] (1. 兵庫県立大学、2. 産業技術総合研究所)

10:45 ~ 11:10
[T1-B-4-03] 大規模言語モデルによる投稿の意味的連続性判定手法

*林 想汰[1]、小林 亜樹[2] (1. 工学院大学情報学部情報通信工学科メディア情報システム研究室、2. 工学院大学情報学部情報通信工学科)

11:10 ~ 11:35
[T1-B-4-04] 数個の発話と疑似文脈生成に基づくキャラクタ対話応答生成

*福岡 啓人[1]、宮森 恒[1] (1. 京都産業大学情報理工学部情報理工学科宮森研究室)

11:35 ~ 12:00
[T1-B-4-05] 性格による応答の多様性を考慮したキャラクタ対話データセット構築

*村田 樹[1]、宮森 恒[1] (1. 京都産業大学情報理工学部)

SNSを用いたChatGPTの不安感情の特定と解消法の検討

鍛示　晃大†　　王　　元元†

† 山口大学工学部知能情報工学科　〒 755–8611 山口県宇部市常盤台 2-16-1
E-mail: †{a028ffu, y.wang}@yamaguchi-u.ac.jp

あらまし　2022 年 11 月 30 日に一般公開された ChatGPT は，その運用について幅広い議論が行われている．たとえば，企業が提案を行う際に AI ツールを活用して出た案は信頼されにくい傾向がある．世界的にも ChatGPT の使用を制限しようという動きが広がる中で，日本は新技術に大きな関心を寄せている．そこで，本研究では SNS データを活用し ChatGPT の運用について感情分析を行い，不安と期待を特定した後，それを解消する手法を提案する．これにより，ChatGPT に対する不安情報とその不安を解消する情報をユーザに提示することで，ChatGPT を業務に実装し効率化を行う企業が増加することが期待できる．本論文では YouTube に投稿された ChatGPT の運用に関する議論を行っている動画 10 件を収集し，その動画に付与された字幕と投稿されたコメントの特徴を抽出して ChatGPT への反応分析を行い，感情分析に基づき不安情報および期待情報を抽出し，不安情報と期待情報の類似性を算出することで，視聴している動画に対して，不安を解消する動画シーンを別の動画から発見しユーザに提示する．最後に ChatGPT の運用に対する不安情報および不安を解消できる情報抽出の有効性を検証する．

キーワード　ChatGPT，感情分析，不安情報，SNS データ，解消法

1　はじめに

ChatGPT3.5 プロトタイプは 2022 年 11 月 30 日に一般向けに無料公開された [1]．世界中でその運用方法について広範囲な議論が行われており，多くの国や企業が ChatGPT の使用を制限しようとする動きが見られる．一方で，日本では OpenAI 社の CEO が来日し，国会での使用例もあり，世界とは異なる動向が見られる．しかし，同様に ChatGPT の使用を禁止した企業も存在する．ChatGPT などの AI 技術への信頼性に疑念を抱く声も多く聞かれる．10 代から 20 代の若者層が AI の使用に反対の意見を持つことが調査により明らかになっており，AI が社会に悪影響を及ぼす可能性や人間の代替は難しいといった意見があげられている [2]．

また，SNS 上でも ChatGPT の運用に関する議論が活発に行われている．その中で，ChatGPT の使用に年齢制限を設けるべきだという意見が存在する．これは未成年への悪影響を懸念しているためであるが，一方で，ChatGPT を通じて難解な質問に対する専門的な教育を受ける機会を得られる可能性があるとする賛成意見も存在する．これらの異なる意見はどちらも一定の根拠があり，一方を排除することは難しい状況である．

このような議論の中で，すでに ChatGPT をビジネスに活用している企業もある．ChatGPT は議事録の作成，アイディアの考案，翻訳，表の作成など，多岐にわたる業務に利用されている．ChatGPT の導入により業務の効率化が図られている企業も多く，プログラミング業務においては，かつて 6 時間かかっていた事前調査が 5 分で完了するなどの実績も報告されている [3]．業務時間の短縮だけではない．会議での議事録の作成に活用されている．ChatGPT による議事録の作成は人間とは異なり，誤字脱字がないため正確であるため，積極的に利用されている．また議事録の作成に人員を割かなくて済むというのも大きなメリットの一つだ．専門的な説明を詳しくに人に説明する場合に ChatGPT が使われることもある．ChatGPT に「小学生でもわかるように教えて」と付け加えると分かりやすく解説をしてくる．プライバシーの問題が懸念される一方で，ChatGPT は適切な使い方をすれば非常に有益であるといえる．

そこで本研究では，SNS データを用いて ChatGPT の利用に不安を感じる人々の理由を特定し，その不安を解消する情報を提供することを目指す．具体的には，YouTube に投稿された動画のコメントから ChatGPT に対する不安情報を特定し，動画の字幕情報を利用してその不安を解消する情報を提案する．これにより ChatGPT の利用を促進し，業務の効率化に寄与することが期待される．

本論文の構成は以下のとおりである．2 章では，BERT を用いた感情分析に関する研究および YouTube を用いた特徴抽出に関する研究などの関連研究を紹介する．3 章では，投稿動画を用いた特徴抽出の方法について説明する．4 章では，感情分析に基づく ChatGPT 運用における不安情報とその解消情報の発見方法について述べる．5 章では，不安情報とその不安解消情報の有効性を評価するための実験について説明する．最後に，6 章でまとめと今後の課題について述べる．

2　関連研究

SNS を用いたテキスト分析や感情分析，その分析結果を用いた推薦システムに関する研究やシステムの開発は盛んである．特に，SNS 分析において BERT (Bidirectional Encoder Representations from Transformers) [4] はその有用性 [5] から多用されている．BERT は大規模な生コーパスで pre-training し，各タスクで fine-tuning するという 2 ステップからなる．

BERT の特徴としては，事前学習済のモデルを利用するため，さまざまな種類の自然言語処理タスクに対応可能である．BERT の日本語特化型モデルにはさまざまなものが存在する．その 1 つに日本語用 Sentence-BERT モデルの sonoisa/sentence-BERT-base-ja-mean-tokens-v2 [6] がある．事前学習モデルとして，東北大学乾研究室の cl-tohoku/BERT-base-japanese-whole-word-masking を利用した．BERT の感情分析として圓谷ら [7] は Amazon レビューを用いた研究を行っており，その分類精度について報告している．この研究では感情分析システムの EEAS [8] との比較も行われており，SNS データの分析には BERT を用いるほうが精度を向上させることができるとしている．

本研究では YouTube を用いて ChatGPT に関する不安感情の分析とその解消情報の提供を行う．不安感の抽出に関する研究を紹介する．難波と福田の研究 [9] では，SNS を活用して新型コロナワクチンに関する人々の不安感の情報抽出を行った．日本語の代表的な感情分析システム ML-Ask を用いて感情分類を行った．

YouTube を用いた研究について，堺らの研究 [10] があげられる．この研究では Youtube にといて視聴者コメントがある炎上動画の検出を行っているが，動画の分析については再生回数，コメント数，評価数の 3 つと各コメントの感情スコアを活用し，動画をベクトル化している．11 感情カテゴリベクトル「喜，怒，哀，怖，恥，好，厭，昂，安，驚，望」を用いているが,「怖」の感情ベクトルの抽出が難しいようだ．本研究では動画の字幕に着目し，コメントとの類似度でクラスタリングを行った．類似度に関する研究として，田中の研究 [11] があげられる．この研究はコサイン類似度を元に文脈の類似度を求めている．

3　投稿動画を用いた ChatGPT への特徴抽出

本章では，投稿動画の視聴者から投稿されたコメントの特徴を分析する方法について述べる．本研究では，YouTube に投稿された動画を利用する．また，コメントのほかに，図 1 に示すメタデータ（タイトル，字幕，いいね数）も合わせて抽出する．

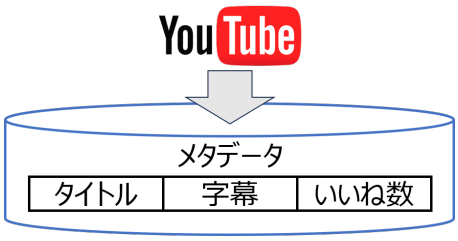


図 1　YouTube 投稿動画のメタデータ

3.1　データセット作成

本研究では，YouTube に投稿された動画へのコメントから感情分析を行う．ChatGPT-3.5 プロトタイプが 2022 年 11 月 30 日に一般公開後，半年以内に，YouTube Data API[1] を使用して，YouTube に投稿された 10 本の動画の字幕とコメントを抽出した．コメントの URL，Bot，他言語を前処理として排除した．その結果，3,428 件のコメントが抽出された．使用した動画は表 1 で示されている．

表 1　動画 ID と各動画タイトル

ID	動画タイトル
1	世界を一変させる AI その驚愕の実力とリスクとは？
2	ChatGPT が変える働き方 AI リスクは増加へ
3	ChatGPT の使い方！超すごい AI 思考サポートツール
4	対話型 AI「チャット GPT」は世界をどう変えるのか？
5	ChatGPT でホワイトカラーが短期間で産業革命に晒される
6	OpenAI サム・アルトマン CEO 単独インタビュー
7	AI は人類の脅威なのか？ チャット GPT で無くなる仕事は
8	人類と共存できる？対話型 AIChatGPT の正体とは？
9	議事録、企画書、メール返信…面倒な仕事は ChatGPT で全て解決？
10	あなたの仕事は大丈夫…？ChatGPT がビジネスを侵食する時代

3.2　動画の字幕とコメントの特徴抽出

まず最初に，動画に付与された字幕と投稿されたコメントの特徴を抽出するため，BERT を用いて字幕とコメントの特徴ベクトルを生成する．日本語用 Sentence-BERT モデルとして，sonoisa/sentence-BERT-base-ja-mean-tokens-v2 を使用する．このモデルでは事前学習済みモデルとして cl-tohoku/BERT-base-japanese-whole-word-masking を用いている．動画ごとに抽出されたコメントと字幕の各特徴ベクトルに対するコサイン類似度を算出し，コメントを動画の字幕数次元で特徴付ける．字幕数次元で特徴付けられたコメントの反応を分析するためには，k-means を用いてクラスタリングを行い，正規化も行う．さらに，主成分分析を用いてコメントの次元数を 2 次元に圧縮する．クラスタ数はエルボー法とシルエット図によって決定し，k-means++法を用いてクラスタリングを行う．

4　感情分析に基づく ChatGPT 運用における不安解消法

本章では，前章で得られたクラスタリング結果を利用し，感情分析に基づく ChatGPT の運用に関する不安情報（ネガティブな情報）および期待情報（ポジティブな情報）を抽出する．そして，不安情報と期待情報の類似性を算出し，不安を解消に寄与する動画字幕の発見を行う．これにより，ユーザに向けて視聴している動画に関連する不安を解消するシーンを別の動画から発見し，提示する．

4.1　不安・期待情報の抽出

各動画のコメントに対して BERT を使用して感情分析を行い，事前学習モデルとして BERT-japanese-finetuned-sentiment を採用する．これにより，各コメントをネガティブ，ニュートラル，ポジティブの 3 つのカテゴリに分類する．ネガティブなコメントを不安情報として抽出し，ポジティブなコメントを期待

1：https://developers.google.com/youtube/v3?hl=ja

情報として抽出する．前章のクラスタリング結果に基づいて，各クラスタのネガティブとポジティブなコメントの感情スコアの平均を算出する．感情スコアの平均以上にネガティブおよびポジティブに分類されたコメント数は，それぞれネガティブ（不安情報）が 446 件，ポジティブ（期待情報）が 400 件であった．表 2 に各動画のクラスタ数，ネガティブなコメント数，ポジティブなコメント数を示す．

表 2　各動画のクラスタ数とコメント数

動画 ID	クラスタ数	不安情報	期待情報
1	10	106	102
2	8	148	102
3	4	23	45
4	4	11	9
5	7	41	41
6	4	19	43
7	5	64	10
8	4	19	19
9	4	8	19
10	4	7	10

4.2　不安解消情報の発見

抽出された 446 件の不安情報（ネガティブなコメント）と傾向が似ている別動画の期待情報（ポジティブなコメント）のクラスタを人手で判断し，不安情報と類似度の高いポジティブなクラスタを特定する．不安情報と特定したクラスタ内の期待情報のコサイン類似度を算出し，期待情報を特定する．最後に，特定した期待情報とその動画の各字幕のコサイン類似度を算出し，類似度の高い字幕の前後約 20 秒から 30 秒の動画シーンを不安解消情報としてユーザに提示する．

5　評 価 実 験

本章では，被験者による不安情報および不安解消情報の抽出精度に関する評価を実施した．本研究では YouTube に投稿された動画のコメントから不安情報の抽出を行い，動画字幕を特徴量としたクラスタリングを用いて不安解消情報となる字幕の発見を行った．各 10 本の動画の中から「いいね」数が最も多かった不安情報を表 3 に，不安解消情報となる動画字幕データを表 4 に示す．

5.1　実 験 概 要

まず，抽出した不安情報の精度を評価するために，アンケート調査を実施した．今回の実験では，20 代からなる計 10 名の被験者を対象とした．アンケート調査では，被験者に抽出した不安コメントがネガティブな意味を含んでいるかどうかを回答してもらい，適合率で抽出精度を検証した．抽出した不安情報は 446 件であり，それを前半 223 件，後半 223 件の 2 グループに分けて調査を行った．次に，抽出した不安解消情報の有効性を評価するためのアンケート調査を実施した．アンケート調査では，被験者に抽出した不安コメントに対する不安解消情報として動画字幕が適切であるかを適合率で抽出精度を検証した．またその字幕の前後 20〜30 秒の動画を被験者に視聴してもらい，不安解消情報として適切かどうかを 5 段階（1 : 全くそう思わない，2 : あまり思わない，3 : どちらともいえない，4 : ややそう思う，5 : とてもそう思う）で評価してもらった．検証した動画は 446 件の中からランダムに 100 件選び，一人あたり 10 件動画を評価してもらった．このアンケート結果を用いて，提案手法により抽出した不安解消情報の有効性を評価した．

5.2　実 験 結 果

抽出したネガティブなコメントが不安情報として適切であったかについての結果を表 5 に示す．前半のグループに対する評価を行った被験者は A から E までで，後半のグループに対する評価を行った被験者は F から J までである．

アンケート結果により，抽出した不安情報の精度は，446 件のコメントの中で，適合率の平均は 0.86 となり，約 384 件のコメントが不安情報として正しく抽出できた．これは良い結果といえる．

次に，抽出した不安情報を解消するための字幕が適切であったかについての結果を表 6 に示す．前半のグループに対する評価を行った被験者は A から E までで，後半のグループに対する評価を行った被験者は F から J までである．

最後に，前後 20〜30 秒の動画を被験者に視聴してもらい，不安解消情報として適切かどうかを 5 段階（1 : 全くそう思わない，2 : あまり思わない，3 : どちらともいえない，4 : ややそう思う，5 : とてもそう思う）で評価してもらった結果を図 2 に示す．

5.3　考　　察

抽出した不安情報についての適合率の平均が 0.86 と良い結果が得た．これは収集した動画がテレビ局公式アカウントが発信したものであったり，専門家を交えた議論ベースであったため，視聴者も ChatGPT がどのようなものか理解することが容易であった．これにより視聴者は自分の意見をはっきりとコメントすることができ，感情分析の精度を高めることにつながったと考えられる．前半グループの適合率が低い理由に関しては動画の返信コメント数が多いことが関係していると考えられる．前半グループのアンケートには返信コメント数の多い ID1 と 2 の動画のコメントが多く含まれていた．これらの動画のコメントには，動画の内容には直接関係のないものが存在したため，適合率が低下したと考える．また，ポジティブな情報も不安情報と同じくらい確認することができた．ニュートラルな情報を省くと，ChatGPT を業務や日常生活に活用したいかどうかについて，ちょうど半々であることがわかる．

不安解消情報の適合率の平均が 0.42 となった．字幕の抽出の際に使用したポジティブなコメントに異常はなかったため，字幕側に問題がある可能性がある．字幕一文のみを不安解消情報としたため，不安解消情報としての内容を満たしていなかったことが原因だと考えられる．字幕の文章量の見直しや，類似度の算出方法を Simpson 係数や Dice 係数などに変更するなどの

表 3 不 安 情 報

ID	不安情報
1	AI に依存し過ぎて生身の人間が自分の脳で思考、判断、決定しなくなる． 未来が本当に来そうで怖い。手塚治虫さんの火の鳥をちょっと思い出した。
2	自ら考える事や描く事をやめるって怖い事だと思うけど手軽さには勝てないんだろうね。 そして AI に教えられた事が正しいと思い込む世界が誕生する．誰も AI には勝てないのだから
3	残念ながらウチの会社では使用禁止になりました。 慣れてない人が悪気なく社外秘情報入力しちゃったりという問題も出てきそうとのことです。
4	ネット後進国日本はネット上の情報の質が低いし量も圧倒的に少ないからね。 アメリカの Wiki は質が高いけど、日本の Wiki はデタラメだらけ。 IT 後進国日本ではまだまだ普及は難しいね。
5	現実は芸術や研究などやりたいことを AI に奪われて、 汎用人型ロボットが安価に実用化されるまで人手が必要な単純作業や肉体労働しか出来なくなるんだね。 ディストピアかな。
6	半導体の敗北とかもそうだけど、日本の上の奴ら先見の明無さすぎるだろ。 あいつら何を根拠に自信満々に特定の研究とか業界が盛り上がるだの盛り上がらんだの判断してるのか。 何回敗戦すれば学習するのか
7	AI に社内機密漏らす人増えそう
8	理系院生です。 gpt のコーディング技術が高すぎて、私はコミュニケーションエンジニアと化してしまいました
9	使う人間がダメだと、結果もダメだから、結局利口な人間にしか使えない
10	学習しているビッグデータ自体が少し前のものなので最新の情報は反映されませんよね

表 4 不安解消情報となる字幕データ

ID	不安解消情報となる字幕データ
1	それよりも後に起きた話ですとかっていうのはあまり対応していないというのと 少しまだ学習中まだこれはあくまで無料公開しているものであって
2	人間じゃないとできないようなフィードバックというのはあるので そういった部分に時間を使ったりとか時間の使い方が変わった
3	逆に言うとこのツールを使いながらさらに上乗せする例えば表現力を もっと磨くとかもっと詳しい情報を多数とかプログラミングであっても これをお客さんで説明するとかそういう部分は人間やればいいと思うんです
4	得てしてどこかで人間の本質が感情みたいななかなか言語化しづらいものというものが そもそもこのデータとしてどこに蓄積されていてどう評価されているのかってところが まあ如実に出てくるかもしれません
5	今回はポストステーブルディフュージョン世代ってありえなくて 3ヶ月しか経ってないんで 生きてる中でみんな経験して半年間でこうなっちゃったっていうのがいやこれ人類史上でも こんなにツールが急に来たっていうのはないんじゃないですか
6	技術は素晴らしいので、ありがとうございます
7	ある意味生産性が上がって人間はもうクリエイティブな方に専念できるともいえるが まあ今それをやってる人たちにとって脅威にもずっとでその脅威という意味で言うとかなりですね
8	それこそ人工 知能とか知識処理が得意としてきたことなんですけれども まだ同じようなものが出てきてるだけと先生もおっしゃられた
9	やっぱりそれその狭間というか葛藤ですごくストレスだったんですけど ChatGPT を使って全体会議の台本を書いてもらってもうそれを読むとかそういう感じ
10	弱点で先ほどから出てますけどやっぱりまだまだ情報に間違いがあったりとか あとは嘘をつかれたりとかそういったことはあるんですけどそれはやがて精度も上がっていくでしょうし

表 5 抽出した不安情報に関する結果

被験者	A	B	C	D	E	F	G	H	I	J	平均
適合率	0.75	0.76	0.75	0.77	0.76	0.97	0.95	0.97	0.96	0.97	0.86

改良が必要である．

不安解消情報（字幕）に対応する動画シーンのアンケート結果も良い結果とはいえなかった．これはコメントに書かれた不安情報に対する回答が，10 件の動画で完全に網羅することができなかったことが一因と考えられる．アンケートにおいて、「1：全くそう思わない」の回答が「2：あまりそう思わない」を

表 6　抽出した不安解消情報に関する結果

被験者	A	B	C	D	E	F	G	H	I	J	平均
適合率	0.44	0.41	0.40	0.42	0.46	0.44	0.41	0.40	0.40	0.43	0.42

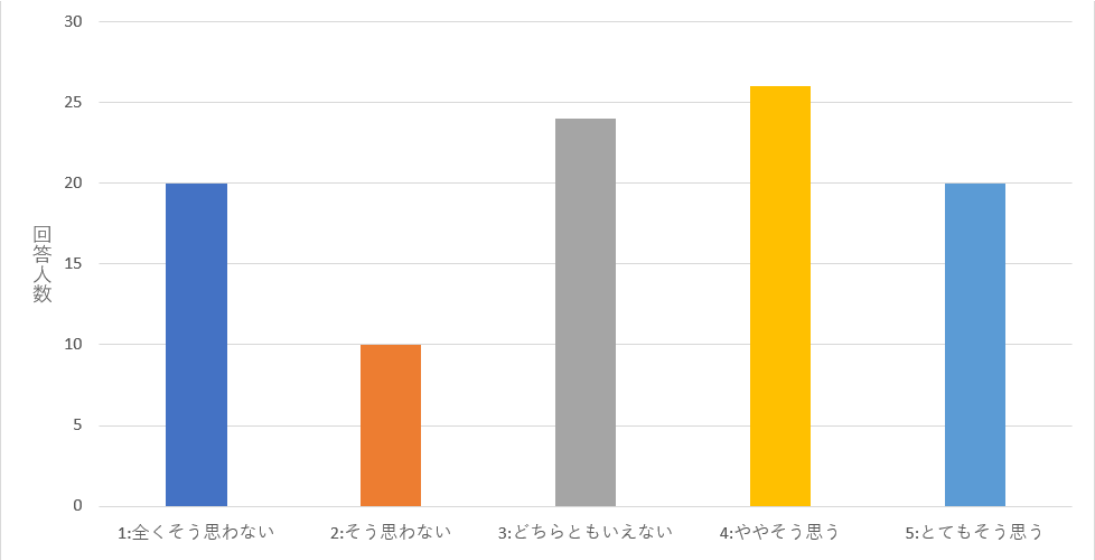


図 2　不安解消情報（字幕）に対応する動画シーンのアンケート結果

大きく上回ったことから，不安解消情報として提示するための適切な動画シーンの不足が示唆されている．リリースから半年ほどで専門性の高い信頼性のある動画を入手するのが難しかったことが最も大きな要因であった．報道機関の YouTube チャンネルや専門家が ChatGPT に関する動画を多数投稿してくれることを期待したい．

今後の展望として，抽出したポジティブデータに関する分析を行う．ネガティブなデータと同様にポジティブなデータも抽出できたので，このデータを分析することで，さらに ChatGPT への不安を解消することにつながると考える．また，動画字幕を提案するとともに，それに併せて会話の流れを考慮した再生時間の提示を行う．これにより，長い動画を短く視聴し要点だけを抑えることができる推薦システムへと昇華する．

6　おわりに

本研究では，YouTube に投稿された ChatGPT の運用に関する議論を行っている動画を収集し，その動画に付与された字幕と投稿されたコメントの特徴を抽出して ChatGPT への反応分析を行い，感情分析に基づき不安情報および期待情報を抽出し，不安情報と期待情報の類似性を算出することで，不安解消情報の特定を行った．最後に，ChatGPT の運用に対する不安情報および不安を解消できる情報の抽出精度を検証した．

今後の課題として，発見した不安解消情報の精度向上を目指す．不安解消情報として動画の字幕を採用したが，一文で不安を解消させるのは難しいという結果となった．したがって，動画内の会話内容単位による分析を行うことを目指す．

文　献

[1] Antonia Creswell, and Murray Shanahan: Faithful Reasoning Using Large Language Models. ArXiv, abs/2208.14271, 30 August 2022.
[2] 日本科学未来館: みんなでつくる AI マップ. 日本科学未来館, 2020 年 6 月 11 日, https://www.miraikan.jst.go.jp/resources/miraikanfocus/202006121367.html (参照: 2023 年 12 月 30 日).
[3] 太田百合子: パナソニックは「ChatGPT」をこんな仕事に使っている. KADOKAWA ASCII Research Laboratories, 2023 年 07 月 27 日, https://ascii.jp/elem/000/004/146/4146649/ (参照: 2023 年 12 月 30 日).
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), pp. 4171-4186, 2019.
[5] 池上達也, 任福継, 西出俊, 康シン: Tweet 日本語感情コーパスに対する BERT の効果の検証. 情報処理学会第 83 回全国大会講演論文集, Vol. 2021, No. 1, pp. 527-528, 2021 年 3 月 4 日.
[6] Nils Reimers, and Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Ubiquitous Knowledge Processing Lab (UKP-TUDA) Department of Computer Science, Technische Universitat Darmstadt, pp. 3982–3992, 2019.
[7] 圓谷顯信, 高橋宏和, 安達由洋: BERT による日本語文の感情分析と話題分析. 情報処理学会第 84 回全国大会講演論文集, Vol. 2022, No. 1, pp. 37-38, 2022 年 2 月 17 日.
[8] 小林孝充, 石井解人, 惠谷菜央, 近藤友啓, 安達由洋: 感情語辞書と感情分析システム EEAS. 情報処理学会第 83 回全国大会講演論文集, Vol. 2021, No. 1, pp. 47-48, 2021 年 3 月 4 日.
[9] 難波英嗣, 福田悟志: ネットからの不安感の情報抽出. 日本感性工学, Vol. 19, No. 4, pp. 163-170, 2021 年.
[10] 堺雄之介, 伊東栄典: 動画サイトにおける視聴者コメントの特徴抽出. 人工知能学会第 124 回知識ベースシステム研究会, pp. 17-22, 2021 年 11 月.
[11] 田中克明: 文脈の類似度に着目した通時的対象の抽出の検討. 埼玉工業大学先端科学研究所アニュアルレポート, No. 21, pp. 10-17, 2023 年 3 月.

ラップバトルにおけるアンサーの類型化および類型の自動分類

三林　亮太† 　佃　洸摂†† 　渡邉　研斗†† 　中野　倫靖†† 　後藤　真孝††
山本　岳洋† 　大島　裕明†

† 兵庫県立大学 情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 産業技術総合研究所　〒 305–8568 茨城県つくば市梅園 1–1–1
E-mail: †af22h007@guh.u-hyogo.ac.jp, ††{k.tsukuda,kento.watanabe,t.nakano,m.goto}@aist.go.jp,
†††t.yamamoto@gsis.u-hyogo.ac.jp, ††††ohshima@ai.u-hyogo.ac.jp

あらまし　ラップバトルとは，2 名のラッパーが即興のラップを交互におこなう競技で，その勝敗を判定するための評価指標に，相手のラップへの返答であるアンサーの質がある．アンサーの質は判別が難しく，どのような場合にアンサーとして成立するのか，どのような種類のアンサーがあるのかは明らかでなく，勝敗を判定することは容易ではない．アンサーの類型化とその自動分類が可能になれば，どのようなアンサーが即興ラップに含まれているのかを明確にすることができ，勝敗の判定補助が可能となる．そこで本研究では，ラップバトルにおけるアンサーの類型化および類型の自動分類に取り組む．まず，ラップバトルの動画を文字起こししたコーパスから，アンサーとして成立するラップ文のペアを 429 件作成した．それらに対して，3 年間以上ラップバトルの視聴経験があるアノテータ 2 名によるカテゴリの作成とアノテーションをおこなうことで，アンサーを類型化した．結果として，アンサーかアンサーでないかを判別する 3 種類のメインカテゴリと，アンサーがどのような表現かを判別する 1 種類のメインカテゴリの合計 4 種類のメインカテゴリから成る類型を作成した．さらに，アンサーの類型をより詳細に分類可能にするため，4 種類のメインカテゴリに対して合計 24 種類のサブカテゴリの類型を作成した．作成した類型の内，2 種類のメインカテゴリにおいて，アノテーションにおける Krippendorff の α が 0.5 を超える一致率を示した．アンサーの自動分類では，BERT を用いた自動分類に取り組んだ．相手のラップ文に対する返答のラップ文が，作成した 3 つの類型の内の少なくとも 1 つに該当するか否かを判別する分類タスクに取り組み，正解率が 0.8554 と高い分類結果を示した．さらに，アンサーがどのような表現かを判別する分類タスクにも取り組み，F 値が 0.4641 と中程度の分類性能を示した．

キーワード　ラップバトル，アンサー，類型化，自動分類

1　はじめに

ラップバトルとは，2 名のラッパーが即興のラップを交互におこなう競技である．ラッパーが自分の番で発言する即興のラップをバースと呼び，バースは複数のラップ文で構成される．図 1 にラップバトルの例を示す．ラッパー A の「まずはリズムに乗せる小節/俺のライムは特別」というバースに対して，ラッパー B は「一体お前のどこが特別/ラップ上手い順ならお前は後列」というバースを返答する．このようなやり取りを複数回繰り返すことでラップバトルは終了する．一般的に，ラップバトルの勝敗は，観客による多数決によって決められる．

ラップバトルでは，バースの質が勝敗の判断基準となる．バースの質は主に，ライム，アンサー，フローの 3 つを対象に評価される [1], [2]．ライムは母音列が一致する単語のペアを指し，そのような単語のペアをバースに含むことができていたかを評価する．たとえば，図 1 のバース 2 の「特別（TOKUBETSU）」と「後列（KOURETSU）」は「OUEU」という母音が一致するためライムである．次に，アンサーは相手のバースを理解し，その内容に沿った返答をする行為を指し，そのような返答をバースに含むことができていたかを評価する．たとえば，図 1

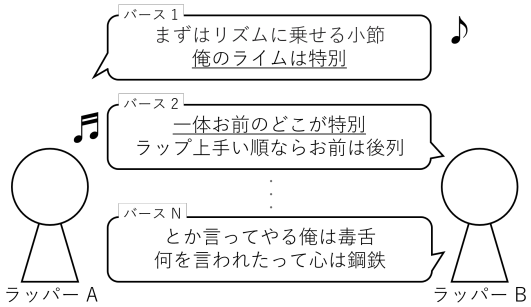


図 1　ラップバトルの例

のバース 1 内の「俺のライムは特別」に対して，バース 2 内の「一体お前のどこが特別」は，相手の「特別」というバースの内容に沿ったアンサーである．最後に，フローはラップの歌い方のことを指し，バースをどれだけリズム良く歌うことができたかを評価する．

これら 3 つの中でも特に，アンサーの質はラップバトルにおいて重要な評価指標である．即興でおこなわれるラップバトルでは，その場で即興のバースを考えたことを証明する必要がある．つまり，事前にクオリティの高いバースを準備していないことを示す必要がある．アンサーは，相手の内容に沿ったバースを事前に準備できないため，即興で考えた証明としても利用

される．また，ライムとフローは一般的なラップの楽曲においても評価対象として用いられるのに対して，アンサーは対話形式のラップバトルに特化した評価対象である点でも重要であると言える．

しかし，アンサーの質は判別が難しい評価指標であり，どのようなバースの掛け合いがアンサーとして成立するかは，これまでに分析されていない．さらに，アンサーが成立した場合においても，そのアンサーがどのような種類の表現であるかは，基準が明らかになっていない．そのため，観客にとってアンサーを評価することは容易ではなく，ラップバトルの勝敗を公平に決めることは難しい．もし，バース内にアンサーが含まれているかを判別することができ，それがどのような表現であるかを自動的に提示できれば，観客は勝敗の判定が容易になる．そこで本研究では，ラップバトルにおけるアンサーの類型化をおこない，その類型を基にアンサーの自動分類に取り組む．

具体的には，まずラップバトルの動画を文字起こししたコーパスからアンサーのペアを 429 件作成し，3 年間以上ラップバトルの視聴経験があるアノテータ 2 名によるカテゴリの作成とアノテーションによってアンサーを類型化する．本研究では，「単語」「ライム」「対話」の各観点でアンサーに該当するか否かを判別するための 3 種類のメインカテゴリと，アンサーの表現の種類を判別するための 1 種類のメインカテゴリの合計 4 種類のメインカテゴリから成る類型を作成した．

さらに，アンサーの類型をより詳細に分類可能にするため，4 種類のメインカテゴリに対して合計 24 種類のサブカテゴリの類型を作成した．作成したメインカテゴリとサブカテゴリの妥当性を検証するため，クラウドソーシングを用いたラベル付けタスクをおこなった．その結果，4 つのメインカテゴリの内 2 つのメインカテゴリで Krippendorff の α が 0.5 を超える一致率を，残りの 2 つのメインカテゴリで 0.2 を超える一致率を示した．また，サブカテゴリについても，Krippendorff の α を調べた結果，24 種類のサブカテゴリの内 4 種類で 0.5 を超える一致率を，16 種類で 0.2 を超える一致率を得られた．

次に，作成した類型を基に，BERT [3] を用いたアンサーの自動分類をおこなった．その結果,「単語」「ライム」「対話」の 3 種類のメインカテゴリの少なくとも 1 つでアンサーに該当するか否かを分類する 2 値分類タスクにおいて正解率は 0.855 となり，高い性能でアンサーの自動分類が可能であることを示した．さらに，アンサーがどのような表現かを判別する分類タスクにも取り組み，F 値が 0.4641 と中程度の分類性能を得た．

本研究の貢献を以下に示す．

- ラップバトルコーパスから抽出した 429 件のアンサーペアに基づいて，4 種類のメインカテゴリと 24 種類のサブカテゴリから成る，ラップバトルにおけるアンサーの類型を作成した．
- 作成したアンサー類型を基に，クラウドソーシングを用いたアンサー類型のラベル付けタスクを実施し，メインカテゴリとサブカテゴリのいずれにおいても作成した類型に一定の妥当性があることを示した．
- BERT を用いたアンサーの自動分類に取り組み，正解率が 0.855 という高い分類性能が得られることを示した．

2 関 連 研 究

類型化に関する研究は多くおこなわれており，これまでに，類型化の方法，対話に関する類型化，コーパスに対する時間表現の類型化，有害表現の類型化などが取り組まれてきた．

2.1 類型化の方法

類型化の代表的な方法として，MAMA(Model-Annotate-Model-Annotate) サイクルがある．MAMA サイクル [4] は Pustejovsky らが提案したアノテーション方法で，基準策定，アノテーション作業，評価，修正作業を繰り返すことで，最適な類型を作成する．類似する手法として，Pustejovsky らは MATTER サイクル [4] を提案している．MATTER サイクルは，構造学習器を学習するステップを MAMA サイクルに追加した手法である．ただし，アノテーションの方針が決まっていない初期の状態では MAMA サイクルをおこなうことが多く，後述する類型化の関連研究においても，MATTER サイクルではなく，MAMA サイクルが採用されている [4] [5]．

2.2 対話に関する類型化

対話に関する類型化の研究は多く取り組まれている．Trippas ら [6] は，音声対話に関する類型化をおこない,「タスク」「会話」「その他」の 3 つのカテゴリとそれらに対するサブカテゴリ(「タスク」のサブカテゴリとしては「情報要求」や「検索支援」など）を定義している．類型は Cohen の k 係数 [7] を用いて，サブカテゴリレベルで評価され 0.71 の高い一致率を示した．

Higashinaka ら [8] は，対話破綻の誤り分類に取り組んでいる．1,146 件のデータを作成し，その中からランダムに選択した 100 件の対話データに対して 24 名がアノテーションをおこない，そこで付与したコメントを基に対話破綻の誤りを分類した.「文法エラー」「解釈不能」「発話無視」などの，17 のメインカテゴリからなる類型を作成し，それらを「破綻」「破綻の可能性がある」「破綻でない」の 3 クラスと,「破綻」「破綻でない」の 2 クラスに分け，評価をおこなっている．評価では，Fleiss の k 係数 [9] が 3 クラスの分類において 0.276 の一致率を示し，2 クラスの分類では 0.396 の一致率を示している．

また，このようなデータに基づく類型化とは異なる類型化のアプローチとして，Higashinaka ら [10] は，対話理論から逸脱した対話を対話破綻とみなし，類型を作成している．結果として,「発話」「応答」「文脈」などの，16 のメインカテゴリと，それらに対するサブカテゴリ（「発話」のサブカテゴリとしては「構文的誤り」や「意味的誤り」など）を定義している．作成した類型を基に，1,046 件の対話に対して，アノテーションをおこなっている．評価では，Fleiss の k 係数 [9] がメインカテゴリでは 0.400，サブカテゴリでは 0.239 の一致率を示している．

さらに，Higashinaka らはこれら 2 つの類型を統合した新たな類型を提案している [11]．この類型では，Fleiss の k 係数 [9] において，専門作業者によるアノテーションでは一致率が 0.567，クラウドワーカによるアノテーションでは一致率が 0.488 となって，いずれも高かった．

2.3 コーパスに対する時間表現の類型化

保田ら [5] は，現代日本語書き言葉均衡コーパスを対象に，時間的順序関係の解析を補助するための類型を提案している．類型の作成には，MAMA サイクルを用いている．類型は，Cohen の k 係数 [7] により評価し，0.653 の高い一致率を示した．また，MAMA サイクルでは一般的に，アノテーション一致率という指標を良くする方向に最適化するが，言語学的には多様なラベルも必要であり，必ずしも一致率を最適化することが適しているわけではないと保田らは主張している．

坂口ら [12] は，京都大学テキストコーパスを対象に，時間表現のアノテーションをおこなっている．類型は，Krippendorff の α [13] により評価され，厳しい条件において 0.526，緩和した条件において 0.867 の一致率を示している．次に，作成した類型を基に，双方向 GRU モデル [14] を用いた自動分類に取り組んでいる．結果として，時間性判定タスクでは，0.905，事象の時間的長さ分類タスクでは 0.606，事象の発生時期分類タスクでは 0.496 の正解率を示している．

2.4 有害表現に関する類型化

有害表現を類型化する研究もおこなわれている．小林ら [15] は日本語の有害表現の類型化に取り組み，Twitter，ガルちゃん，5ch といったプラットフォームから取得したコーパスを使用している．アノテーションは 2 種類おこない，有害レベルと有害カテゴリの 2 種類に対してラベルを付与した．有害レベルは 4 段階のスコアをつけており，有害カテゴリは MAMA サイクル [4] を用いて類型を作成している．結果として,「差別」「卑語」「迷惑行為」などの 7 つのメインカテゴリと，それらに対する 41 つのサブカテゴリ（「差別」のサブカテゴリとしては「人種差別」や「職業差別」など）を定義している．これらを，Krippendorff の α [13] を用いて評価し，有害レベルにおいて，4 値分類では 0.40，2 値分類では 0.78 の一致率を示した．また，作成した類型を基に，BERT を用いた有害検知器を作成しており，回帰問題として解いている．結果として，いくつかの評価尺度に対して高い性能で有害検知ができている．

2.5 本研究の位置づけ

これまでに紹介した関連研究では，対話破綻，時間表現，有害表現を対象に類型を作成している．著者らの知る限り，ラップバトルにおけるアンサーの表現を対象にした類型はまだ作成されていない．これはラップバトルが比較的新しい競技であり，まだ十分に分析がおこなわれていないためだと考えられる．しかし，近年では，ラップバトルの言語学的側面からの分析がおこなわれており [1], [16]，ラップバトルという競技において，アンサーに関する類型を定義することは重要な貢献となる．

さらに，ラップバトルにおけるバースの自動生成に関する研究もいくつか提案されている [17], [18], [19]．本研究において，アンサーの自動分類に取り組むことで，バースの自動評価やラップバトルの勝敗判定の補助など，バースの自動生成に対する応用が可能になると考えられる．そのため，坂口ら [12] や小林ら [15] と同様に，作成した類型の自動分類にも取り組む．

3 アンサーの類型化

本節では，アンサーの類型化の方法について述べる．まず，ラップバトルの動画を文字起こししたコーパスを用いて，アンサーのペアデータを作成する．作成したアンサーのペアデータに対して，3 年間以上ラップバトルの視聴経験があるアノテータ 2 名によるカテゴリの作成とアノテーションをおこない，アンサーを類型化する．作成した類型は Fleiss の k 係数 [9] と Krippendorff の α [13] によって，アノテータ間のラベルの一致率を評価する．

3.1 アンサーのペアデータの作成

我々は，Mibayashi ら [18] が作成した，ラップバトルの動画を文字起こししたコーパスを利用した．3 つの大会 (UMB[1]，凱旋 MCbattle[2]，戦極 MCBATTLE[3]) と 1 つのテレビ番組 (フリースタイルダンジョン) について，それぞれ 25 件のバトルをコーパスからランダムにサンプリングすることで，合計 100 件のバトルを取得した．

取得した 100 件のバトルに対して，アンサーのペアデータを作成する．その作成方法は，図 2 の左側に示すように，まず，1 つのバトルからバースのペアを取得する．ただし，これらのバースはすべてがアンサーで構成されているわけではなく，アンサーでない内容も含まれている．そこで，このバースのペアから，アンサーとそのアンサーをするに至った相手のラップ内容（アンサーの対象）をアンサーのペアとして文単位で人手で抽出する．バースのペアには，アンサーとアンサーの対象が複数件含まれている場合があるため，その場合は該当するものをすべてアンサーのペアとして抽出する．

たとえば，図 2 に示すように,「俺のライムは特別」というアンサーの対象の文と,「一体お前のどこが特別」というアンサーの文はアンサーのペアである．バースによっては，アンサーが複数文にまたがることもあるため，その場合は文同士を連結した．たとえば,「まじでお前はつまらない」に対して「つまらないのはお前の方/俺は即興でも決して詰まらない」といったように，アンサーが 2 文にまたがる場合,「まじでお前はつまらない」に対して「つまらないのはお前の方俺は即興でも決して詰まらない」と連結した文をペアにした．100 件のバトル内のバースのペアすべてに対して，10 年間以上ラップバトルの視聴経験がある筆頭著者が，アンサーと判断したものをアンサーのペアとして取得した．結果として，429 件のアンサーのペアデータを作成した．

3.2 ラベル付けの基準作成

本研究ではラベル付けの基準作成のために，ラップバトルの戦略やテクニックについて詳細な情報が書かれた著名なラッパーらの書籍 [2], [20] を参考にした．書籍内には，実例を基にしたラップバトルの解説があり，プロのラッパーがどのような

1：https://www.youtube.com/user/umbofficial
2：https://www.youtube.com/channel/UCe_EvY8GrvYgx8PbwRBc75g
3：https://www.youtube.com/user/senritumc

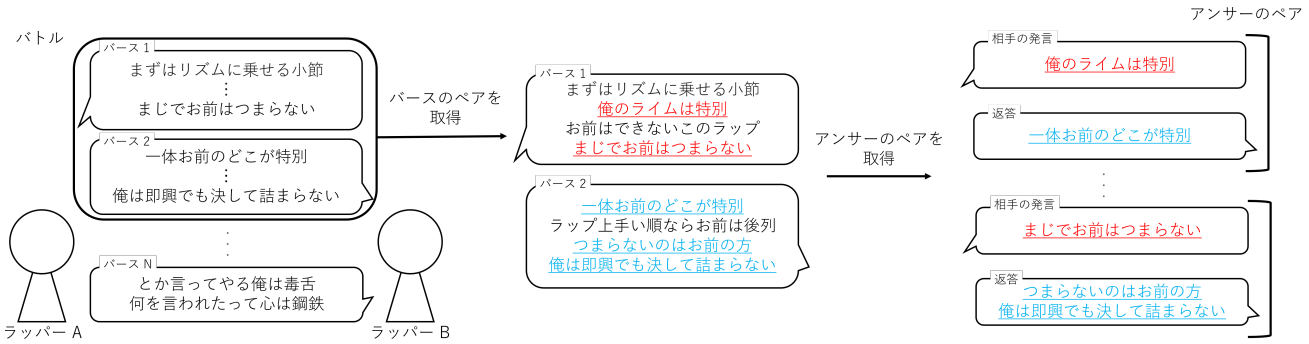


図 2　アンサーペアデータの作成方法

返答をアンサーと判別しているかが書かれている．これらの内容を参考に，以下の 3 つの各観点においてアンサーに該当するか否かを判別するためのメインカテゴリを設計した．

（1）　相手の発言を考慮した**単語**が返答に含まれるか

（2）　相手の発言を考慮した**ライム**が返答に含まれるか

（3）　相手の発言を考慮した**対話**が返答に含まれるか

以下，メインカテゴリ 1 を「**単語によるアンサー**」，メインカテゴリ 2 を「**ライムによるアンサー**」，メインカテゴリ 3 を「**対話によるアンサー**」と呼ぶ．

3 つのアンサーの該当を判別するメインカテゴリに加えて，アンサーがどのような表現であるかを判別する以下のカテゴリも作成した．

（4）　アンサーの表現の種類

まず，1 つ目の単語によるアンサーは，相手の発言を考慮した単語が返答に含まれるかを判別するカテゴリである．たとえば以下のような例が挙げられる．

相手：俺のテンションは高いぞ
返答：テンション確かにいいね

これはラッパー A の「テンション」という単語がラッパー B の発言に含まれているため，相手の発言を考慮しているというアンサーである．

2 つ目のライムによるアンサーは，相手の発言を考慮したライムが返答に含まれるかを判別するカテゴリである．たとえば以下のような例が挙げられる．

相手：俺のテンションは高いぞ
返答：でもファッションは残念だな

これは，ラッパー A の「テンション」という単語に対して，ライムである「ファッション」という単語がラッパー B の発言に含まれているため，相手の発言をライムで考慮しているアンサーである．

3 つ目の対話によるアンサーは，相手の発言を考慮した対話が返答に含まれるかを判別するカテゴリである．たとえば以下のような例が挙げられる．

相手：俺のテンションは高いぞ
返答：そうだな確かにそう見える

これは，直接的に相手の発言を含んでいるわけではないが，相手の発言に対して，肯定や否定といった，対話的な返答が含まれているアンサーである．

4 つ目のアンサーの表現の種類は，アンサーが成立した際に，どのような表現の種類のアンサーであるかを返答の内容のみで判別するカテゴリである．たとえば以下のような例が挙げられる．

返答：お前のラップは下手くそだ

これは，返答自体に含まれる表現の種類のことであり，上記の例では，相手を否定する表現が含まれている．

3.3　メインカテゴリに対するサブカテゴリの作成

前節で作成したメインカテゴリをさらに細かく分類できるように，4 つのメインカテゴリそれぞれに対するサブカテゴリを作成する．まず，サブカテゴリの作成のため，10 年間以上ラップバトルの視聴経験がある筆頭著者 1 名と 3 年以上ラップバトルの視聴経験がある著者ではないアノテータ 1 名の合計 2 名で，アンサーペアに対するアノテーションをおこなった．アノテーションでは，同じアンサーペアに対して，2 名のアノテータ間で協議をしながらサブカテゴリを付与した．3.1 節で作成した 429 件のアンサーペアから 300 件をランダムに選択し，それらに対してサブカテゴリを付与した．300 件のアノテーション後，さらに残りの 129 件のアンサーペアからランダムに選択した 100 件のアンサーペアに対してアノテーションをおこない，それらに対して新しいサブカテゴリが付けられなかったことを確認した上で，アノテーションを終了した．アノテーションしたサブカテゴリを精査し，各サブカテゴリは「単語によるアンサー」において 4 種類,「ライムによるアンサー」において 2 種類,「対話によるアンサー」において 5 種類,「アンサーの表現の種類」において 13 種類作成した．

3.4　アンサーの類型の詳細

作成したアンサーの類型の概要を表 1 に示す．以下では，各メインカテゴリに対するサブカテゴリについて，具体例を用いて説明する．

3.4.1　単語によるアンサー

単語によるアンサーでは，相手のバース内の単語に関連する単語を含むことで，本メインカテゴリのアンサーに該当すると判別する．単語は同じものとは限らず，類義語やその他の関連語を含む．

表 1　作成したアンサーの類型とアノテータ間のカテゴリの一致率

メインカテゴリ	サブカテゴリ	サブカテゴリの説明	Krippendorff の α [13]
(1) 単語によるアンサー	同じ語	相手のバース内の単語を含む	0.6994
	類義語	相手のバース内の単語に類似する単語を含む	0.1846
	その他関連語	相手のバース内の単語に関連する単語を含む	0.4153
	関連語なし	相手のバース内の単語に関連する単語を含まない	0.6497
(2) ライムによるアンサー	ライムあり	相手のバース内の単語に対するライムを含む	0.2372
	ライムなし	相手のバース内の単語に対するライムを含まない	0.2372
(3) 対話によるアンサー	肯定	相手のバースに対して肯定している	0.3629
	否定	相手のバースに対して否定している	0.3865
	肯定と否定の両方	相手のバースに対して肯定と否定をしている	0.2642
	肯定でも否定でもない	相手のバースに対して肯定も否定もしていない	0.0335
	対話になっていない	相手のバースに対する返答になっていない	0.2273
(4) アンサーの表現の種類	ライム	同じ母音列を持つ単語を返答に含むこと	0.3669
	ディス	暴言や相手が嫌がる内容を含むこと	0.5028
	否定	相手の主張や発言を認めないこと	0.1851
	指摘	相手の発言に対して指摘をすること	0.2952
	自己賛美	自分の能力をアピールすること	0.3557
	賛同	発言に対して称賛したり認めたりすること	0.4725
	文意不明	発言の意味が不明であること	0.2061
	開き直り	相手の発言は正しいがそれに対して開き直ること	0.2680
	優勢	自分が有利な立場にあることを主張すること	0.1644
	平気	相手の発言を気にしていないことをアピールすること	0.0721
	レペゼン	自分の地元や所属をアピールすること	0.6065
	観客	観客に対してアピールすることや観客に喋りかけること	0.3060
	表現なし	これまでのカテゴリのどれにも属さないもの	0.0927

表 2　アンサーの表現の種類の類型

サブカテゴリ	説明	返答の例
ライム	同じ母音列を持つ単語を返答に含むこと	ラップにおいてスキルは重要な内容
ディス	暴言や相手が嫌がる内容を含むこと	お前のラップは下手くそだ
否定	相手の主張や発言を認めないこと	理解できないその思想
指摘	相手の発言に対して指摘をすること	こないだもそれ言ってたけど
自己賛美	自分の能力をアピールすること	俺はめちゃくちゃラップ上手い
賛同	発言に対して称賛したり認めたりすること	お前のアンサーは最高だな
文意不明	発言の意味が不明であること	俺はラップで崩壊する火事だ
開き直り	相手の発言は正しいがそれに対して開き直ること	たしかに俺はこの前逃げたけどそれがなんだ
優勢	自分が有利な立場にあることを主張すること	会場を沸かした数は俺のほうが多い
平気	相手の発言を気にしていないことをアピールすること	別に構わないぜ俺はそんなことでは動じない
レペゼン	自分の地元や所属をアピールすること	俺は地元大阪からはるばる来たぜ
観客	観客に対してアピールすることや観客に喋りかけること	お客さんどっちについていくか決めな
表現なし	これまでのカテゴリのどれにも属さないもの	俺はそうやってやってる

- **同じ語**：相手のバース内の単語を含む
  - 相手：俺のテンションは高いぞ
  - 返答：テンション確かにいいね
- **類義語**：相手のバース内の単語に類似する単語を含む
  - 相手：俺のテンションは高いぞ
  - 返答：そうだな気持ちは大事
- **その他関連語**：相手のバース内の単語に関連する単語を含む
  - 相手：俺のテンションは高いぞ
  - 返答：じゃあ体育会の飲み会みたいに盛り上げろ
- **関連語なし**：相手のバース内の単語に関する単語を含まない（本メインカテゴリのアンサーに該当しない）
  - 相手：俺のテンションは高いぞ
  - 返答：俺は今日もラップがうまい

3.4.2　ライムによるアンサー

ライムによるアンサーでは，相手のバース内のライムに関連する単語を返答のバース内に含むことで，本メインカテゴリのアンサーに該当すると判別する．

- **ライムあり**
  - 相手：俺のテンションは高いぞ
  - 返答 1：でもファッションは残念だな
  - 返答 2：つまりダンジョンでやろうセッション
- **ライムなし**（本メインカテゴリのアンサーに該当しない）
  - 相手：俺のテンションは高いぞ
  - 返答 1：楽しそうでなにより
  - 返答 2：テンションがどうこう俺は知らんな [4]

3.4.3　対話によるアンサー

対話によるアンサーでは，相手のバース内の単語やライムを

4：同じ語を含むだけではライムにはならない．

返答のバース内に含むのではなく，対話として成立する返答がバース内に含まれていれば，本メインカテゴリのアンサーに該当すると判別する．対話には，肯定や否定といった基本的なサブカテゴリに加え，肯定と否定の両方や肯定でも否定でもない，といったサブカテゴリも考慮した．

- **肯定**
  - 相手：俺のテンションは高いぞ
  - 返答：そうだな確かにお前は高い
- **否定**
  - 相手：俺のテンションは高いぞ
  - 返答：いや俺のほうが上回ってるテンション
- **肯定と否定の両方**
  - 相手：俺のテンションは高いぞ
  - 返答：確かにお前は高いでも低いときもあるだろが
- **肯定でも否定でもない**
  - 相手：俺のテンションは高いぞ
  - 返答：テンションをアピールしているお前
- **対話になってない**（本メインカテゴリのアンサーに該当しない）
  - 相手：俺のテンションは高いぞ
  - 返答：高い低いでも俺の価値は高い

3.4.4 アンサーの表現の種類

アンサーの表現の種類では返答のみを対象とし，アンサーがどのような表現であるかを判別する．その種類は 13 種類と多いため，サブカテゴリの説明と具体例は表 2 に示す．

3.5 ラップバトルに精通したアノテータによるアノテーション

前節で作成したアンサーの類型を評価するために，ラップバトルに精通したアノテータによるアノテーションをおこなった．アノテータは，クラウドソーシングサービス Lancers にて，ラップバトルの視聴歴が 3 年以上の者を 3 名採用した．アノテータは，著者らが作成したアノテーション作業用のサイトにアクセスし，3.1 節で作成した 429 件のアンサーペアデータに対してアノテーションをおこなった．

アノテータには，4 種類のメインカテゴリのそれぞれに対してサブカテゴリを付与するように指示した．まず,「単語によるアンサー」「ライムによるアンサー」「対話によるアンサー」の 3 種類のメインカテゴリに対しては，サブカテゴリを 1 つだけ付与するように指示した．これら 3 種類のメインカテゴリ内のサブカテゴリは排反であるため，複数付与することはできない．次に,「アンサーの表現の種類」に対しては，サブカテゴリの複数付与を可能とし，どれにも当てはまらない場合は「表現なし」を付与するように指示した．アノテータに提示するアンサーペアデータの順序はアノテータごとにランダムに決め，作業の順番がアノテータ間で同じにならないようにした．

結果を図 3 に示す．図中の n の値は，アノテータ 3 名が 1 つのアンサーペアに対して付与したサブカテゴリの一致数を示している．たとえば，(1)「単語によるアンサー」において，アノテータ 3 名が付与したサブカテゴリがそれぞれ「同じ語」「同じ語」「類義語」だった場合,「同じ語」に $n = 2$ が 1 つカウントされる．全員が「同じ語」なら，それに $n = 3$ が 1 つカウントされる．アノテーションの信頼性の観点から $n = 1$ は扱わず，アノテータ 2 名以上が同じサブカテゴリを付与した事を示す $n = 2$ と $n = 3$ を合算した割合（n が 2 以上の割合）に基づいて考察する．そこで $n = 2$ と $n = 3$ のそれぞれをカウントした後，アンサーペアデータの総数 429 で割ることで，各サブカテゴリの割合を計算した．図 3 の各グラフでは，合算した割合の降順にサブカテゴリを表示している．

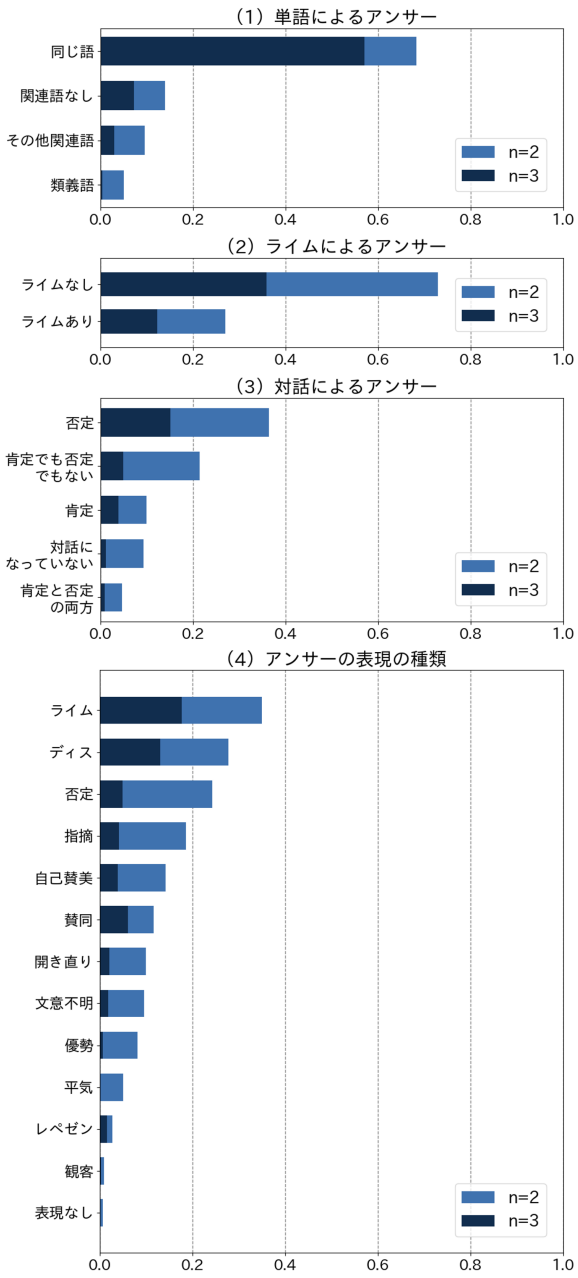


図 3　4 つのメインカテゴリに対するアノテーションによるサブカテゴリの割合（横軸：割合）

まず，(1)「単語によるアンサー」については,「同じ語」が 68.29%と最も高い割合で,「関連語なし」は 13.98%と低い割合であった．このことから，相手のバースに対して何らかの関連する単語がアンサーに含まれる傾向にあり，中でも「同じ語」が最も含まれていることがわかる．次に，(2)「ライムによるアンサー」については,「ライムあり」が 27.03%と「ライムなし」

表 3　メインカテゴリ別の評価

メインカテゴリ	Fleiss の k 係数 [9]	Krippendorff の α [13]
(1) 単語によるアンサー	0.5582	0.5585
(2) ライムによるアンサー	0.2366	0.2372
(3) 対話によるアンサー	0.2450	0.2456
(4) アンサーの表現の種類（複数ラベルの完全一致）	0.0955	0.0962
(4) アンサーの表現の種類（複数ラベルの内どれか一つでも一致）	0.5533	0.5536

より低い割合で，ライムは望ましいもののアンサーにはライムが含まれない場合が多いことがわかる．そして，(3)「対話によるアンサー」については,「対話になっていない」は 9.32%と低い割合なので，多くは対話になっていたことがわかる．対話になっている中では,「否定」が 36.36%と最も高い割合な一方で「肯定」は 10.02%と低い割合であり，アンサーでは否定的になりがちなことがわかる．

最後に，(4)「アンサーの表現の種類」については,「ライム」が 34.96%と最も割合が多かった．これはラップバトルでライムが多用されているために，アノテータが付与したサブカテゴリ（ラベル）が多かったと考えられる．次に多い「ディス」「否定」「指摘」は相手に対して否定的な表現や攻撃的な表現を含む類型である．これらの割合が 18%を超えていることから，ラップバトルではそうした表現が多いことがわかる.「自己賛美」「開き直り」「優勢」「平気」は自身が優位な状態にあることを表現する類型であり,「賛同」は相手を褒めることを表現する類型である．これらの類型の割合が 5%から 15%の間にあることから，自身の優位な状態を示すことや相手を褒めることより，上記の相手に対して攻撃的な表現を含む場合のほうが多いことがわかる.「文意不明」も割合が 5%から 10%の間にあるが，これは即興で的確な表現を考えてアンサーを返すことがいかに難しいかを示している．最後に,「表現なし」は 0.69%とかなり低い割合であることから，これらの類型により，アンサーの表現は十分に網羅できていると考えられる．

3.6　類型に対するアノテーションの評価

前節でアノテーションしたアンサーの類型を評価する．今回，アノテータが 3 名であるため，3 名以上の評価に対応している Fleiss の k 係数 [9] と Krippendorff の α [13] を用いて評価した．Fleiss の k 係数は -1 から 1 までの値を取り，Krippendorff の α は理論上負の値を取ることが可能であるが，一般的には 0 から 1 までの値を取る．どちらの評価尺度も 1 に近づくほど一致率が高いことを示す．本実験では，Krippendorff の α については名義尺度で評価をおこなった．

評価結果を表 3 に示す．まず，単語によるアンサーでは，Fleiss の k 係数は 0.5582 であり，Krippendorff の α は 0.5585 であった．これは Higashinaka ら [10] の研究における，専門業者らによるアノテーション結果である Fleiss の k 係数 0.567 に匹敵する高い一致率を示している．このような一致率になったのは，単語によるアンサーのサブカテゴリは比較的容易に判断できるものが多いためであると考えられる．ただし，表 1 に示すように,「類義語」と「その他関連語」の一致率が低い．これは両者の違いが曖昧であるため，アノテータ間でラベルが一致しなかったと考えられる．

ライムによるアンサーについては，Fleiss の k 係数は 0.2366 であり，Krippendorff の α は 0.2372 であった．これは Higashinaka ら [8] による，対話破綻の 3 クラスの分類における Fleiss の k 係数 0.276 に近い一致率であり，少し低い一致率と言える．原因としては，ライムが含まれているか否かをテキストで判別することは難しく，アノテータ間でラベルが一致しなかったと考えられる．具体的には,「お前は枚方を上げれず終わるぜ」と「イカサマもしくは枚方」というアンサーペアに対して,「ライムあり」のラベルが 2 件,「ライムなし」のラベルが 1 件付けられている．これは「枚方 (HIRAKATA)」と「イカサマ (IKASAMA)」が「IAAA」という母音で一致するライムである．しかし枚方という大阪の地名が正しく読めなかったため,「ライムなし」が付けられた可能性が高い．このように本来音声として聞いていればライムと判別できるものが，テキストだけでは判別が難しくなり，不一致が起きたと考えられる．

対話によるアンサーについては，Fleiss の k 係数は 0.2450 であり，Krippendorff の α は 0.2456 であった．これも先ほどと同じく，少し低い一致率である．これは,「肯定でも否定でもない」が判別しにくく，一致率が低いことが影響していると考えられる．実際に 3 名のアノテータによるラベルの組み合わせを調べると,「否定，否定，肯定でも否定でもない」が 78 件,「否定，肯定でも否定でもない，肯定でも否定でもない」が 23 件あり,「否定」と「肯定でも否定でもない」が判別しにくいことがわかる．さらに,「対話になっていない，対話になっていない，肯定でも否定でもない」が 25 件,「対話になっていない，肯定でも否定でもない，肯定でも否定でもない」が 16 件あり,「対話になっていない」と「肯定でも否定でもない」も判別しにくいことがわかる．

アンサーの表現の種類については，複数ラベルの完全一致において，Fleiss の k 係数は 0.0955 であり，Krippendorff の α は 0.0962 であった．アンサーの表現の種類は複数ラベルであるため，完全一致することは非常に難しく，0.0962 という低い一致率は自然であると考えられる．しかし，アノテータが付与した複数のサブカテゴリの中から，どれか 1 つでもサブカテゴリが全アノテータ間で一致していれば完全一致とみなすという緩和した条件においては，Fleiss の k 係数は 0.5533 であり，Krippendorff の α は 0.5536 と高い一致率を示している．

4　アンサーの自動分類

ラップバトルにおいて，相手の発言に対する返答のラップ文が与えられたときに，作成したアンサーの類型に基づくアンサーの自動分類がどの程度できるかを検証するため，評価実験

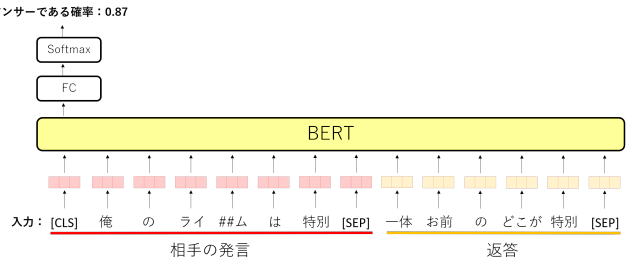


図 4　BERT を用いたアンサーの分類方法

をおこなった．まず,「単語によるアンサー」「ライムによるアンサー」「対話によるアンサー」の 3 つのメインカテゴリを対象として，返答のラップ文が，各メインカテゴリのアンサーに該当するか否かの 2 値分類をおこなった．本章では簡単のため,「本メインカテゴリのアンサーに該当する」「本メインカテゴリのアンサーに該当しない」という表記を,「アンサーである」「アンサーでない」と記述する．次に，これら 3 つのメインカテゴリの内，少なくとも 1 つに該当するか否かを判別する 2 値分類にも取り組んだ．最後に,「アンサーの表現の種類」のメインカテゴリを対象として，返答のラップ文を 13 種類のサブカテゴリに分類する評価実験をおこなった．

4.1　BERT によるアンサーの自動分類

4.2 節から 4.5 節の評価実験では，共通の枠組みを用いてアンサーの分類をおこなっているため，まず以下でその共通の枠組みについて説明する．本実験では，アノテーション済みの 429 件のアンサーペアデータを基に分類タスクに取り組んだ．それぞれのメインカテゴリに対して，アノテーションしてもらったラベルを正解として学習用のデータを作成した．本実験では，3 名のアノテータによって付与してもらったラベルの多数決を取って正解のラベルとした．多数決が取れない場合は学習用データから削除した．

アンサーの自動分類をするためのモデルとして，BERT [3] を用いた．BERT の事前学習モデルは東北大学が公開している bert-base-japanese-v3[5]を用いた．BERT の学習に用いたハイパーパラメータは，バッチサイズは 32，Optimizer は AdamW [21]，ロス関数は Cross Entropy Loss，学習率は 2e-7，Dropout は 0.1，最大文長は 256 とした．

BERT を用いたアンサー分類は図 4 に示すように，相手の発言を 1 文目，返答文を 2 文目とすることで入力とした．前処理として，1 文目の先頭に [CLS] トークンを付与し，1 文目と 2 文目の末尾には [SEP] トークンを付与した．本実験では，BERT から出力される，[CLS] トークンに対応する 768 次元のベクトルを用いて分類問題を解いた．

4.2　単語によるアンサーの分類

単語によるアンサーの分類については，多数決によって正解ラベルを定義し，416 件のデータセットを取得した．本実験では，返答文が単語カテゴリにおけるアンサーに該当するか否かを判別する 2 値分類に取り組むため,「関連語なし」以外の「同じ語」「類義語」「その他関連語」を正例とした，相手の発言と返答文のペアデータを 356 件取得した．それに対して，データセット内の別の返答文を付与したデータを負例として 356 件作成し学習をおこなった．結果として，データセットは 712 件となり，これを訓練用 569 件，検証用 71 件，テスト用 72 件に分けて学習をおこなった．

5：https://huggingface.co/cl-tohoku/bert-base-japanese-v3

表 4　3 つのメインカテゴリに対する分類結果

メインカテゴリ	正解率
(1) 単語によるアンサーの分類	0.7222
(2) ライムによるアンサーの分類	0.8139
(3) 対話によるアンサーの分類	0.8253
(1-3) 単語，ライム，対話によるアンサーの分類	0.8554

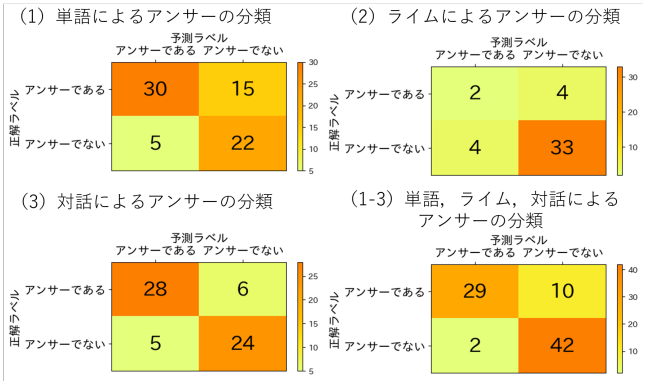


図 5　各分類タスクにおける 2 値分類の混同行列

分類結果は，表 4 に示すように，正解率が 0.7222 となり，ある程度高い性能で分類ができていることがわかる．図 5 に示す混同行列では，本カテゴリに該当するアンサーは正しく分類できているが，アンサーであるものを，アンサーでないと間違えて分類していることがわかる．

4.3　ライムによるアンサーの分類

ライムによるアンサーの分類についても，多数決によってラベルを定義し，429 件のデータセットを取得した．本実験では，返答文がライムカテゴリにおけるアンサーに該当するか否かを判別する 2 値分類に取り組むため,「ライムあり」のラベルを正例,「ライムなし」のラベルを負例とした，相手の発言と返答文のペアデータを 429 件（「ライムあり」が 116 件,「ライムなし」が 313 件）取得した．これを訓練用 343 件，検証用 43 件，テスト用 43 件に分けて学習をおこなった．

分類結果は，表 4 に示すように，正解率が 0.8139 となり，高い分類性能となった．しかし，図 5 に示す混同行列では，予測結果のほとんどが「アンサーでない」となってしまっていることがわかる．これは，データセット全体において,「ライムなし」のラベルが 313 件と多いことにより，このような偏った分類結果になっていると予想される.「ライムあり」の正例データのように，返答文中に相手の発言を考慮した自然なライムを含むことは，自動生成では困難であり，手作業でも時間がかかるため，正例と負例を同数にした学習は本実験ではおこなえなかった．今後はクラスウェイトを使用するなど，データセットのラベルを考慮した学習の改善が必要である．

4.4 対話によるアンサーの分類

対話によるアンサーの分類についても，多数決によってラベルを定義し，351 件のデータセットを取得した．本実験では，返答文が対話カテゴリにおけるアンサーに該当するか否かを判別する 2 値分類に取り組むため,「対話になっていない」(40 件) 以外の「肯定」「否定」「肯定と否定の両方」「肯定でも否定でもない」を正例とした，相手の発言と返答文のペアデータを 311 件取得した．それに対して，データセット内の別の返答文を付与したデータを負例として 311 件作成した．結果として，データセットは 622 件となり，これを訓練用 497 件，検証用 62 件，テスト用 63 件に分けて学習をおこなった．

分類結果は，表 4 に示すように，正解率が 0.8253 となり，高い分類性能を示した．図 5 に示す混同行列では，本メインカテゴリのアンサーに該当するか否かを適切に分類できていることがわかる．これは，BERT の文同士の繋がりを考慮した事前学習により，少量のデータセットでもうまく分類ができたためだと考えられる．

4.5 単語，ライム，対話によるアンサーの分類

単語，ライム，対話によるアンサーのカテゴリに対して，分類をおこなう．学習に使用するデータは 429 件の相手の発言と返答文のペアデータの内，3 つのメインカテゴリのうちの少なくとも 1 つのメインカテゴリにおいてアンサーに該当すれば正例とした．今回，412 件のデータがアンサーと判断されたため，それらのデータを正例とした．それに対して，データセット内の別の返答文を付与したデータを負例として 412 件作成した．結果として，データセットは 824 件となり，これを訓練用 659 件，検証用 82 件，テスト用 83 件に分けて学習をおこなった．

分類結果は，表 4 に示すように，正解率が 0.8554 となり，高い分類性能を示した．図 5 に示す混同行列を見ると，本カテゴリでは「アンサーである」と「アンサーでない」の両方に対して適切に分類できていることがわかる．このことから，アンサーを判別する 3 つのメインカテゴリを同時に考慮することで，メインカテゴリを個別に分類する場合と比べて，高い性能で分類ができることを示した．

4.6 アンサーの表現の種類の分類

最後に，アンサーの表現の種類を分類する．アンサーの表現の種類では，ある返答文に対して複数のサブカテゴリを同時に持つため，複数のサブカテゴリを同時に予測するマルチラベル問題として解いた．本実験では，図 6 に示すように，返答文のみを入力とし，返答文のサブカテゴリを表すベクトルを出力する．入力の前処理として，返答文の先頭に [CLS] トークンを付与し，末尾には [SEP] トークンを付与した．アンサーの表現の種類は 13 種類であるため，[CLS] トークンに対応する 768 次元のベクトルを FC 層によって，13 次元のベクトルとして出力し，Binary Cross Entropy Loss によって正解のサブカテゴリとの誤差を計算した．

学習に使用するデータは，アノテータ 3 人の内，2 人以上が同じサブカテゴリを付与している場合，そのサブカテゴリを正解のサブカテゴリとして加えた．この操作によってラベルが付けられなかった 23 件のデータを除いた，406 件をデータセットとした．これを訓練用 324 件，検証用 41 件，テスト用 41 件として使用した．

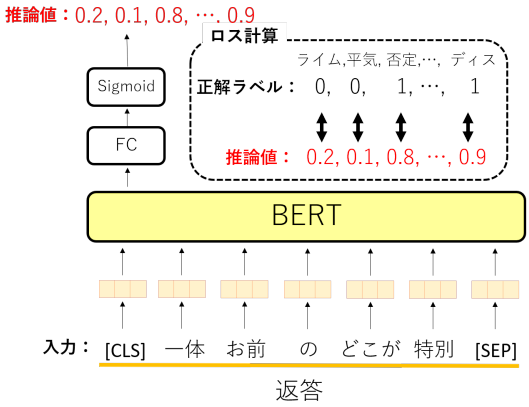


図 6 BERT を用いたアンサーの表現の分類

また，マルチラベル問題を正しく学習するためにクラスウェイトを使用した．マルチラベル問題では，図 6 に示すように，各サブカテゴリの有無を 0 と 1 で表したマルチホットベクトルを正解ラベルとして使用する．本実験で使用する正解ラベルのマルチホットベクトルには 0 が多く含まれているため，予測値をすべて 0 にするといった過学習が起きる可能性がある．そのため，本実験ではサブカテゴリが含まれていることを示す 1 に対して，5.0 のクラスウェイトをかけて学習をおこなった．学習後の推論時では，予測値に対して，0.5 以上を 1，0.5 未満を 0 として扱った．

アンサーの表現の種類の分類結果を表 5 に示す．正解率は，予測値と正解ラベル内の 13 個の要素がすべて一致する際に正解とする場合（完全一致）と，予測値と正解ラベル内のそれぞれ対応する要素が一致する際に正解とする場合（要素ごとの一致）の 2 種類を計算した．まず，完全一致においては，クラスウェイトの有無に関わらず 0.0243 と非常に低い結果となった．これは，13 種類のサブカテゴリの有無を完全に一致させる必要があるため，このような非常に低い正解率になったと考えられる．次に，要素ごとの一致においては，クラスウェイトなしが 0.8592 であり，クラスウェイトありが 0.7617 となった．これは，クラスウェイトなしの方が良い正解率に見えるが，実際はクラスウェイトなしの予測値がほとんど 0 であったため，0 が多い正解ラベルにおいて正解率が高くなっている．

このように，マルチラベル問題において，正解率だけで分類の性能を評価することは難しいため，正解率以外にも，適合率，再現率，F 値を計算した．まず，適合率はクラスウェイトなしでは 0.0512，クラスウェイトありでは 0.7051 であった．クラスウェイトありにおいては，サブカテゴリを含むと予測した際には，予測が正しくできていることがわかる．次に，再現率はクラスウェイトなしでは 0.8000，クラスウェイトありでは 0.3459 であった．クラスウェイトなしはモデルの予測値がほとんど 0 であるため，再現率が 0.8000 と高くなっていることがわかる．クラスウェイトありは再現率が 0.3459 と低く，本来サブカテ

表 5　アンサーの表現の種類の分類結果

実験設定	正解率（完全一致）	正解率（要素ごとの一致）	適合率	再現率	F 値
クラスウェイトなし	0.0243	0.8592	0.0512	0.8000	0.0963
クラスウェイトあり	0.0243	0.7617	0.7051	0.3459	0.4641

ゴリを含むと予測すべきものを予測できていないことがわかる．最後に，F 値はクラスウェイトなしでは 0.0963，クラスウェイトありでは 0.4641 であった．クラスウェイトなしでは適合率が非常に低いため，F 値も 0.0963 と低くなっている．クラスウェイトありでは，F 値が 0.4641 と中程度の値を示していることがわかる．これらのことから，アンサーの表現の種類の分類においては，クラスウェイトを使用した場合，サブカテゴリを含むと予測した場合には高い割合で正解できるが，本来予測すべきサブカテゴリを見逃してしまうことが多いことがわかる．しかし，F 値は中程度の値を示しているため，本分類モデルは中程度の性能で分類が可能であると言える．

5　まとめと今後の課題

本研究では，アンサーの自動分類に向けた類型化に取り組んだ．まず，ラップバトルコーパスからアンサーのペアデータを作成し，ラップバトルに精通した 2 名のアノテータによるラベルの作成とアノテーションにより，アンサーを類型化した．類型化の結果として，作成した類型の内，2 種類のメインカテゴリにおいて，Krippendorff の α が 0.5 を超える一致率を示した．また，作成したアンサー類型を基に，BERT を用いた自動分類に取り組んだ．結果として，作成した 3 つのいずれかの類型のアンサーに該当するか否かを判別する分類タスクの正解率は 0.8554 となり，高い性能で分類ができていることがわかった．さらに，アンサーがどのような表現かを判別する分類タスクにも取り組み，F 値が 0.4641 と中程度の分類性能を示した．今後はアンサーをバースの中から自動的に検知し，バース単位でのペアに対して，自動的にアンサーを分類できる手法に取り組む予定である．

謝辞　本研究は JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けた．

文　献

[1] V. Sykäri. Interactive oral composition: Resources, strategies, and the construction of improvised utterances in a Finnish freestyle rap battle. *The Journal of American Folklore*, Vol. 132, No. 523, pp. 3–35, 2019.

[2] DARTHREIDER. MC バトル史から読み解く日本語ラップ入門. KADOKAWA, 2017.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL 2019, pp. 4171–4186, 2019.

[4] P. James, B. Harry, and Z. Annie. Designing annotation schemes: From theory to model. *Handbook of Linguistic Annotation*, pp. 21–72, 2017.

[5] 保田祥, 小西光, 浅原正幸, 今田水穂, 前川喜久雄. 『現代日本語書き言葉均衡コーパス』に対する時間情報表現・事象表現間の時間的順序関係アノテーション. 自然言語処理, Vol. 20, No. 5, pp. 657–681, 2013.

[6] JR. Trippas, D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. Towards a model for spoken conversational search. *Information Processing & Management*, Vol. 57, No. 2, pp. 1–19, 2020.

[7] V. Susana, K. Uzay, and S. João. Cohen's kappa coefficient as a performance measure for feature selection. In *The 2010 IEEE World Congress on Computational Intelligence*, WCCI 2010, pp. 1–8, 2010.

[8] R. Higashinaka, M. Mizukami, K. Funakoshi, M. Araki, H. Tsukahara, and Y. Kobayashi. Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, pp. 2243–2248, 2015.

[9] JL. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382, 1971.

[10] R. Higashinaka, K. Funakoshi, M. Araki, H. Tsukahara, Y. Kobayashi, and M. Mizukami. Towards taxonomy of errors in chat-oriented dialogue systems. In *The 16th Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL 2015, pp. 87–95, 2015.

[11] 東中竜一郎, 荒木雅弘, 塚原裕史, 水上雅博. 雑談対話システムにおける対話破綻を生じさせる発話の類型化. 自然言語処理, Vol. 29, No. 2, pp. 443–466, 2022.

[12] 坂口智洋, 河原大輔, 黒橋禎夫. 事象に対する網羅的な時間情報アノテーションとその分析. 自然言語処理, Vol. 26, No. 1, pp. 179–206, 2019.

[13] K. Krippendorff. Computing Krippendorff's alpha-reliability. Departmental Papers (ASC). 43, pp, 1–10, 2011.

[14] R. Dey and FM. Salem. Gate-variants of gated recurrent unit (GRU) neural networks. In *IEEE 60th International Midwest Symposium on Circuits and Systems*, MWSCAS 2017, pp. 1597–1600, 2017.

[15] 小林滉河, 山崎天, 吉川克正, 牧田光晴, 中町礼文, 佐藤京也, 浅原正幸, 佐藤敏紀. 日本語有害表現スキーマの提案と評価. 言語処理学会第 29 回年次大会, NLP 2023, pp. 933–938, 2023.

[16] HS. Alim, J. Lee, LM. Carris, and QE. Williams. Linguistic creativity and the production of cisheteropatriarchy: A comparative analysis of improvised rap battles in Los Angeles and Cape Town. *Language Sciences*, Vol. 65, pp. 58–69, 2018.

[17] D. Wu and K. Addanki. Learning to rap battle with bilingual recursive neural networks. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, IJCAI 2015, pp. 2524–2530, 2015.

[18] R. Mibayashi, T. Yamamoto, K. Tsukuda, K. Watanabe, T. Nakano, M. Goto, and H. Ohshima. Verse generation by reverse generation considering rhyme and answer in Japanese rap battles. In *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research*, CMMR 2023, pp. 30–41, 2023.

[19] R. Savery, L. Zahray, and G. Weinberg. Shimon the rapper: A real-time system for human-robot interactive rap battles. In *Proceedings of the 11th International Conference on Computational Creativity*, ICCC 2020, pp. 212–219, 2020.

[20] 晋平太. フリースタイル・ラップの教科書. イースト・プレス, 2016.

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, ICLR 2019, pp. 1–8, 2019.

大規模言語モデルによる投稿の意味的連続性判定手法

林　　想汰†　　小林　亜樹††

† 工学院大学情報学部情報通信工学科　〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学 情報学部情報通信工学科　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †j020216@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　文書解析において一定の主題に関する会話文を対象とする場合があるが，一般に会話の主題はその途中で変化するため，同一の主題についての会話部分のみを抽出する必要がある．本稿では，大規模言語モデル（Large language model; LLM）を用いて，会話文の主題が同一のままであるか転換されたかを判別する手法について提案する．連続する会話文を対象とし，その内容が意味的に継続しているかを尋ねるような問い合わせ文とともに LLM への入力として，その結果を得ることで判定する．Twitter の reply tweet ペアを用いた判定実験を行い，有効性などを検証する．

キーワード　LLM，文書解析，会話，データベース，ソーシャルメディア

1　はじめに

近年 SNS などのソーシャルメディア上での商品評判情報の分析は，現代において極めて重要な役割を果たしている．顧客は自身の意見や感想を即座に共有し，他のユーザーとのやりとりを容易に行うことができるため，企業はこれらの SNS 上での評判を積極的に分析することが求められる．顧客の意見を収集・分析することで，企業は製品の欠点や改善点を把握し，顧客満足度を向上させることができる．さらに肯定的な投稿がされると，製品やサービスのプロモーションにも繋がり，新規顧客の獲得に寄与する可能性がある．また，これらの商品情報は reply 機能等での会話中に商品評判情報が含まれる場合もある．しかし，一般に会話中では所々で話題が転換するため，特定の主題についてのテキストは会話全体の一部である．そのため長い継続会話内で特定の主題であるかを識別することや抽出を行うことが必要となり，会話テキストにおいて話題の転換点を見つけ出せれば主題部分の切り出しに寄与できる．

近年，大規模言語モデル（LLM）の発達が著しく，大規模言語モデルがコンテキストを理解できることから，さまざまなテキスト分析に応用できる．大規模言語モデルは，文の前後の文脈を考慮し，単語やフレーズの意味や関係性を理解する能力を持っており，テキストデータから意味の抽出や類似性の判定，情報の要約など，さまざまな分析タスクに適用することが可能である．

本研究では SNS 等の会話テキストにおいて特定主題文の切り出しを行い，SNS 上の reply 文ペアについて話題が継続しているかの判定を行うことを目的とし，LLM により，文ペアの各文の話題が「同一」「異なる」の判定を行わせ，会話の継続/非継続の判定を行う手法を提案する．提案手法では，複数のやりとりの会話文から 2 回のやりとりのみを抜き出し前処理を行う，前処理後の会話文，会話文とその会話文が継続しているかの答えを含む例題と一緒にプロンプト文を作成する．パラメータを設定して，LLM に与えて回答文から判定結果を抽出する．あらかじめ作成した正解ラベルとの比較を行う．プロンプトで与える例題の種類を変えて会話の継続正判定への影響を調査し，本論文で設定するパラメータ組から有効性の検証を行う．

2　関連研究

SNS に投稿されたデータを用いた研究は多くあるが，話題の継続性を検出する研究は少ない．対話に着目した研究では対話破綻検出チャレンジが提供するデータセットである DBDC データセット [1] を用いた研究がある．DBDC データセットは 1,2 は日本語のみだったが 3 からは日本語と英語の両方に対応している．データセットは人間と対話システムによる 2 者間の対話とそれぞれの対話にアノテーターが付与したラベルによって構成されている．ラベルは破綻ではない，破綻とは言い切れないが，違和感を感じる発話，あきらかに破綻している発話，の 3 種類の内どれかのラベルが付与されている．付与は 30 人のアノテーター達の多数決によって決められ，違和感を感じる発話と破綻している発話の割合が 0.5 を超えなかったものに関しては破綻ではないのラベルを付与している．DBDC データセット 1 は 100 の雑談対話が集められ，DBDC データセット 2 は DCM, DIT，IRS の 3 つの雑談対話システムからそれぞれ 100 対話ずつ集められ合計 300 対話で構成されている．

LSTM を中間層に用いた RNN による対話検出手法 [2] では対話破綻検出チャレンジで配布された条件付き確率場を用いた検出の手法と比較してあきらかに破綻の検出精度は高かったが違和感を感じる発話とあきらかに破綻の検出精度では低い結果となっていた．また，QRNN で対話モデルの学習を行い対話モデルを用いた対話破綻検出器を提案した手法 [3] では，対話モデルの学習に RNN を用いた対話破綻検出器の手法と比べて学習速度が 3.2 倍高速化し検出精度の改善を示していた．しかしこれらの研究では対話が破綻しているかしていないかという点には着目しているが実際に SNS に投稿されたデータを用いて実験は行っていない．

また，SNS のデータを利用した研究では災害情報の採取に着

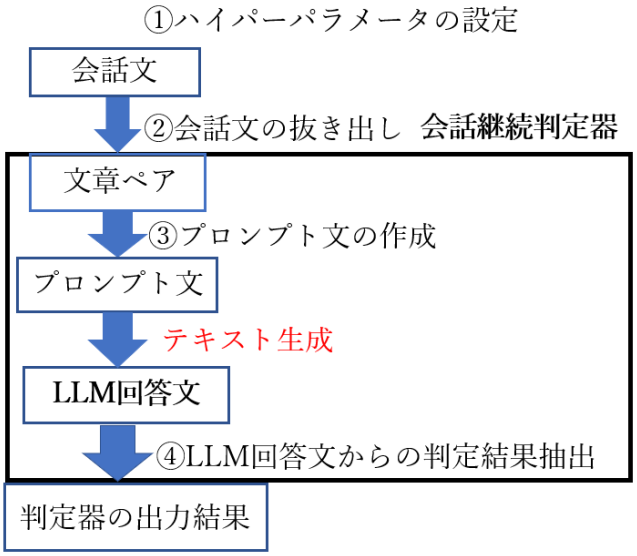


図 1: 提案手法概要図

目した藤田ら [4] の研究がある．この研究では情報を採取する採取者による教師データを必要とせず自己教師学習による準リアルタイムでの採取の手法を提案した．Twitter は投稿できる文字数が 140 文字以内で短文であることから分類は困難であったが，複数のつながった tweet を 1 つの会話文章と見なし，かつその会話内でも分析を行うことも提案した．この手法では代表語が含まれているかどうかでの分類を行う比較手法と比べて精度が良く会話内の部分単位での分析が有効であることを明らかにした．しかし特定の単語のみの採集は可能であるが汎用的にさまざまな情報を採取出来るものではなかった．

本研究では SNS に投稿されたデータでつながりのある投稿群を会話のやりとりとみなし，それらの会話間においての話題の継続性を LLM を用いて判定することを目的とする．

3 提 案 手 法

3.1 概　　要

本論文では，一連の会話等の連続したやりとりのテキストにおいて，話題の転換点を見つけ出すことで何らかの主題についての一連のテキストを抽出する方針を考える．そこで，話題の転換点を，やりとりのうち異なる話者から発せられた連続した 2 文（ここでは，会話中の 1 発言などの区切りを 1 文と呼ぶ）の言及している主題が同一であるか異なるかを識別する問題を解くこととする．このとき，2 文中の後続文が先行文と同一の話題であることを話題が継続するといい，異なることを話題が非継続であるという．

この連続する 2 文における話題継続性問題を解く判定器の構成を提案する．提案する判定器は LLM の機能を利用し，話題判定器の振る舞いを指示する文と，判定対象文とを LLM への入力プロンプト文とする．LLM の生成した出力文から判定結果のみを抽出することで，最終的な話題継続性の判定結果を得る．図 1 がその処理の概要図である．判定は連続する 2 文である会話文である．これをプロンプト文に変換するにあたり必要な前処理を行う．前処理後の 2 文の組のことを文章ペアと呼ぶ．これに 3.4 節で述べる指示文を組み合わせてプロンプト文を作成し，LLM への入力とする．LLM 自体は別に設定しておいたハイパーパラメータに従い，入力プロンプト文の指示により出力文を生成する．出力された文を LLM 回答文とよぶが，ここにはプロンプト文だけでは制御しきれない不要な文字列が含まれているため，別に判定結果のみを抽出するフィルタを適用して，最終的な判定結果とする．

3.2 ハイパーパラメータの設定

LLM はテキストを生成する際に生成する文章に変化を与えるものとしてプロンプト文とハイパーパラメータがあげられる．そのなかでもハイパーパラメータを具体的な数値で設定することにより生成させる文章を細かく変えることが可能となる．特に日本語は語順の自由度が高く，1 つの文章内にひらがなや漢字など多くの文字が使われることから複雑かつ表現力が高いテキストを生成することが可能である．

LLM の運用に当たっては数々のハイパーパラメータが設定されるが，本研究では出力文の生成に大きな影響を及ぼす特に 2 つのパラメータについてその値を細かく見ていく．一つ目は max_new_tokens と呼ばれる生成文のトークン数を制限するパラメータであり，これは，経験的に 10 に設定しておく．これは生成するテキストのトークン数が大きい場合，判定に必要な情報以外も生成されてしまい判定結果に影響を及ぼしてしまうことを防ぐ点や，判定結果が継続判定なのか非継続判定なのかを絞りやすくするためである．二つ目は生成するテキストのランダム性を操作できる temperature の値である．このハイパーパラメータを設定を高く設定することにより生成するテキストが創造的な内容になり，低く設定することにより現実的な内容になる．この temperature の値については 5 章において複数の値について試行して，判定器の性能に与える影響を検証していく．

3.3 会話文の抜き出し

LLM に判定させたい複数回のやりとりのある会話文を用意する．そのうち会話の始まりの文章とそれに続く 2 つ目の文章のみを抜き出す．ここでは LLM に判定させる対象の会話文章 2 つを 1 組としたきにその文章集合のことを文章ペアと呼ぶ．また，文章ペアのうち LLM に入力する際に先に与える会話文章を先行文，そのあとに与える会話文章を後続文と定義する．

3.4 プロンプト文の作成

生成されるテキストはプロンプト文の内容によっても左右される．そのためいくつかあるプロンプトの書き方においてどの書き方を選びどのように与えるかが重要となる．プロンプトの記法として shot 文 [5] と呼ばれるプロンプト内に与えるタスクの例題とその回答を示すことで生成されるテキストの精度を向上させるプロンプティング手法が知られている．また，このとき LLM にプロンプトとして与える際に提示する例題のことを本研究では継続性判定例題と呼ぶ．そのうち話題が継続している例題を継続例題，話題が継続していない判定例を非継続例題と呼ぶ．本研究では異なる shot 文による複数のプロンプト文を

例 1 プロンプト作成例

入力 先行文:寝過ぎて体が痛いのだがまだ熱が下がらない副反応でここまで引っ張るの予想外だわ，後続文:三回目がモデルナだと、やや強く服反応が出るそうですね。お大事にです！

出力

```
以下は、タスクを説明する指示です. 要求を適切に満たす応答を書きなさい.
### 指示:
回答は YES か NO だけで出力してください. 以下の 2 つの文の話題が続いてるかどうか判定してください. 判定例は次のとおりです.
1 文目:ブチャラティ、おかっぱのくせにかっこよすぎるな
2 文目:そりゃ世界一かっこいいおかっぱだから
判定結果:YES
1 文目:寝過ぎて体が痛いのだがまだ熱が下がらない副反応でここまで引っ張るの予想外だわ
2 文目:三回目がモデルナだと、やや強く服反応が出るそうですね。お大事にです！
### 判定結果:
```

作成し，shot 文の種類毎による回答精度など判定に及ぼす影響について調査する．さらに，判定結果を集計しやすくするために判定結果は YES か NO のどちらかだけが出力されるように指示する．

ここからプロンプトの作成手順を説明する．例 1 として実際に使用したプロンプト作成例を示す．3.3 にて抜き出した文章ペアを用いる．出力の 4 行目が継続性判定例題の先行文，5 行目が継続性判定例題の後続文，6 行目が継続性判定例題の答えである．7 行目が実際に判定させたい文書ペアの先行文で 8 行目がその後続文である．9 行目の「判定結果:」の後に LLM にテキストを生成させる．このような形式でプロンプトを作成する．例題数を増やす際には 4 行目から 6 行目のような例題が 7 行目以降に記述される．この形式に判定させたい文章ペアや継続性判定例題を変えることでプロンプト文の作成を行う．

3.5 LLM 回答文からの判定結果抽出

LLM 回答文は入力文である命令文や判定対象の会話文などを含むプロンプト文の後に続く形で生成される．また，判定器の結果として基本的には Positive(会話が継続している),Negative(会話が継続していない) のどちらかだけが出力され意味が 1 つに定まるような結果が望ましい．そのために出力結果をそのまま用いるのではなく加工する必要がある．

LLM 回答文を処理する際の概要図として図 2 を示す．3.4 にて作成したプロンプト文を用いて判定結果のみの抽出手順を説明する．まず，文章ペア間においての話題が継続しているかどうかの判定を行う．その結果プロンプト文を含む形で LLM 回答文が生成される．例 2 として出力結果を示す．抽出は以下の手順で行う．

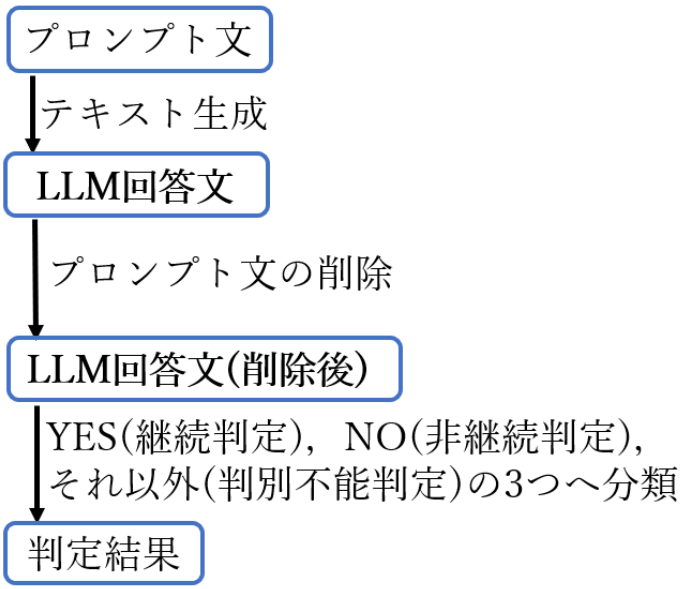


図 2: 出力テキスト処理フローチャート

例 2 出力結果

出力

```
以下は、タスクを説明する指示です. 要求を適切に満たす応答を書きなさい.
### 指示:
回答は YES か NO だけで出力してください. 以下の 2 つの文の話題が続いてるかどうか判定してください. 判定例は次のとおりです.
1 文目:ブチャラティ、おかっぱのくせにかっこよすぎるな
2 文目:そりゃ世界一かっこいいおかっぱだから
判定結果:YES
1 文目:寝過ぎて体が痛いのだがまだ熱が下がらない副反応でここまで引っ張るの予想外だわ
2 文目:三回目がモデルナだと、やや強く服反応が出るそうですね。お大事にです！
### 判定結果:YES </s>
```

（1） まず，1 行目から 9 行目の「判定結果:」まではプロンプト文であり，この部分が存在すると判定結果が誤判定になってしまうため削除を行う．

（2） また，判定結果である「YES」のあとに続く形でプロンプトで与えた答えを含む継続性判定例題が生成される場合があるため，プロンプトとは別に継続性判定例題の削除も行う．

（3） そして，プロンプト文などが含まれていない純粋な LLM の回答文を 3 種類に分類する．

（4） 回答文に「YES」という単語が含まれ,「NO」という単語が含まれない場合には継続判定とする．

（5） 次に，回答文に「NO」という単語が含まれ「YES」という単語が含まれない場合には非継続判定とする．

（6） 最後に「YES」と「NO」が両方の単語が含まれてしまっている場合や「YES」と「NO」が両方の単語が含まれてい

表 1: import リスト

ライブラリ名	version
transformers	4.33.1
pytorch	2.0.1
sqlite	3.41.2

ない場合には判別不能判定とする．

（7） また，3 種類に分類する際にはその単語が含まれているかどうかで分類を行うため「Y,E,S」が含まれる「YEAH!!S」のような形の回答文が継続判定にならないようになっている．これらの手順を踏まえて LLM 回答文から判定結果のみの抽出を行う．

4 試作システム

提案手法を python コードによる 2 つのプログラムで実装する．実行環境は Anaconda を使用しバージョンは conda 23.7.3 である．また，Python のバージョンは 3.11.4 である．試作システムを実行する際に必要なライブラリは表 1 として示す．

本研究で提案する試作システムの流れを以降の節で説明する．会話文の抜き出しで 1 つのプログラム，プロンプトの作成，テキスト生成，LLM 回答文からの判定結果抽出で 1 つのプログラム合計 2 つのプログラムを作成した．

4.1 会話文の抜き出しプログラム

会話文の抜き出しプログラムは json 形式のテキストデータを入力とし，sqlite 形式のテキストデータに変換して出力するプログラムである．json 形式で収集した tweet データ群から，in_reply_to_staus_id の値と id の値が一致する tweet の検索を行う．検索を行い，tweet データ群から一致した tweet データのみの抽出を行う．その際に検索される id の in_reply_to_staus_id が NULL の tweet に限定した．そうすることで reply が 2 つで構成される文章ペアを抜き出さないようにした．検索して一致した際には検索される id の text を先行文，検索する in_reply_to_staus_id の text を後続文として sqlite 形式で出力を行う．

4.2 会話継続判定プログラム

会話継続判定プログラムは sqlite 形式のテキストデータを入力とし，sqlite 形式のテキストデータを出力するプログラムである．4.1 の会話文抜きだしプログラムを用いて抜き出した文章ペアを入力とする．入力である文章ペアを用いてプロンプト文を作成する．複数の継続性判定例題をあらかじめ用意してそれらを組み合わせることでプロンプト文が自動的に作成されるようにする．

そして作成したプロンプト文と設定するハイパーパラメータを設定してテキスト生成を行う．テキスト生成を行う際には使用する LLM モデルのトークナイザーとモデルの読み込みを行ってから生成する．トークナイザーは Python の AutoTokenizer クラス，モデルは AutoModelForCausalLM クラスを利用して読み込みを行う．またテキスト生成を行うコードは PyTorch の torch.no_grad() 内に記述することで勾配の計算をしないように設定した．

テキスト生成後は LLM 回答文からプロンプト文と一致する箇所の削除を行う．次に削除後の LLM 回答文が YES のみか NO のみかそれ以外 (NULL) のどれかに分類する．最後に分類した結果を LLM の判定結果として sqlite 形式で出力する．

表 2: 使用 PC のスペック

部品	仕様	型番
CPU	AM4 8C16T	AMD Ryzen 7 5700X
RAM	128GB DDR4 3200	Crucial DDR4-3200 32GB x4
G/B	dGPU	GeForce RTX4060Ti 16GB

表 3: 実験した条件のパターン表

temperature\ 例題内容	継続のみ	非継続のみ	継続と非継続	計
0.1	5	5	2	12
0.3	5	5	2	12
0.5	5	5	2	12
0.7	5	5	2	12
計	20	20	8	48

5 評 価 実 験

5.1 実 験 目 的

プロンプト内で与える例題の数や種類を変えて会話の継続性判定への影響を調査を行い，LLM において本研究で設定する 4 組のパラメータの組から有効性を検証することを目的として，tweet データを利用して判別実験と性能評価を行う．

5.2 実験手順概要

実験用のデータセットは存在しないため，正解ラベルを付与することを含めて tweet データからデータセットを作成した．これを 4 章の試作システムを用いて判定結果を得て，正解と比較することで性能調査を行う．このとき，5.6 節で述べたハイパ－パラメータが与える影響を検証するため，複数の値による判定結果を比較する．

5.3 実 験 環 境

実験に使用した PC のスペックは表 2 に示す．また，OS は Ubuntu22.04LTS を使用した．実験で temperature や例題の内容や数を変えて行ったパターンをまとめた表として表 3 を示す．

5.4 使用したモデル

本研究では Xwin-LM-13B-V0.1 [6] を使用した．ローカル環境でテキストを生成する上で実行出来るモデルの内，最もパラメーター数の高かったこのモデルを選出した．

5.5 データセットの作成

まず，抽出した tweet データを格納するためのデータベースを sqlite 形式で作成した．

表 4: データセット内訳

データ	継続 (件)	非継続 (件)	計 (件)
自然データ	486	14	500
人為データ	19	481	500
データセット全体	505	495	1000

次に 4.1 の会話文抜き出しプログラムを用いて tweet データの抽出を行う．研究室で streaming API で取得した tweet オブジェクトの JSON 形式の tweet データが保存されておりそれを利用する．2022 年 3 月 3 日，3 月 4 日，3 月 5 日の 3 日分についての抽出を行った．

格納する際に判定精度を高めるために不要な文の削除を行った．tweet に画像や動画などが含まれる際は URL が tweet 本文の末尾に存在するため，その URL を削除した．また，後続文には連続 tweet 以外は reply 元であるユーザーの id が tweet 本文の先頭に来るためこれも削除した．それ以外のユーザー id は削除すると文脈に違和感のある文章になっていまう可能性があるため削除は行わなかった．

その後，格納した先行文と後続文を人手で会話が継続しているかしていないかの判断を行い取得順 500 件分に対しての正解ラベルを付けた．正解ラベルの付与は筆者一人で行い話題が継続していないものを非継続，話題が継続しているもしくは判断が難しいものについては継続のラベルを付与した．この 500 件のデータを自然データと呼ぶ．

また，取得した tweet は基本話題が継続していると考え，先程の 500 件に含まれていない文章ペアから先行文と後続文をランダムに 500 件組み合わせて新たに文章ペア 500 件を作成した．最後に，新たに作成した文章ペア 500 件に対しても先ほどと同様に正解ラベルの付与を行った．この 500 件のデータを人為データと呼ぶ．

データセットの内訳を表 4 として示す．データセット内の正解ラベルにはそれぞれ人手で判断した会話が継続している場合の YES，継続してない場合の NO の 2 種類のうちのどちらかが与えてられている．

5.6 会話継続判定

会話継続判定は 4.2 の会話継続判定プログラムを用いて判定する．本実験では temperature の値は 0.1 から 0.7 まで 0.2 ずつ変化させて実験する．次に 1 つのパラメータ組につき変える例題のパターンを説明する．まず表 5 からランダムに 1〜5 個取りだし継続性判定例題数がそれぞれ 1 5 個の計 5 個のパターンのプロンプトを作成した．同様に表 6 からランダム 1〜5 個取りだし非継続性判定例題数がそれぞれ 1 5 個の計 5 個パターンのプロンプトを作成した．最後に表 5 からランダムに 1 個，表 6 からランダムに 1 つ合計 2 個の例題を含むプロンプト，表 5 からランダムに 2 個，表 6 からランダムに 2 個合計 4 個の例題を含む計 2 個のパターンのプロンプトを作成した．つまり 1 個のパラメータ組につき 12 個のパターンのプロンプトを作成した．またプロンプト内で与える例題の内容による影響を減らすため 1 つのパターンにつき例題内容が被らないように 3 個のプロンプトを作成した．そしてほかの 3 つのパラメータ組についても同様の手順でプロンプトを作成したため合計 192 個のプロンプトを作成した．この作成したプロンプトを用いて会話継続性判定を行う．

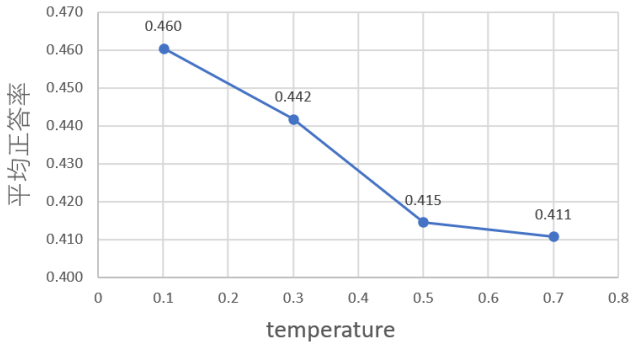


図 3: temperature の比較

5.7 結果集計と正答率の算出

判定器からの出力結果をデータセットの正解ラベルと照合し，正解率などの集計を行う．このとき，出力結果には判定不能である場合が含まれるため，これは別に集計する．試作システムの 4.2 より LLM による判定結果を得たあと正解ラベルとの比較を行い比較結果を得た．正解ラベルとの比較を行った結果は以下の 3 つのパターンに分類される．

（1） LLM の判定結果と正解ラベルが一致していた場合の判定成功

（2） LLM の判定結果と正解ラベルが一致していなかった場合の判定失敗

（3） LLM の判定結果が判別不能判定だった場合の判定不能

そして集計結果から例題の種類毎に判定器の正答率を算出する．このときの正答率は判定成功数を全文書ペア数 (1000 件) で割って算出した．同様に例題の種類は同じだが例題内容が異なる 2 つについても正答率を算出する．得られた 3 つの正答率の平均を取り，平均正答率を算出した．この平均正答率を他のパターンについても同様に算出し，この平均正答率を用いて評価を行う．

5.8 結　　果

temperature の値を変更した際の平均正答率を比較した図として，図 3 を示す．この結果は 1 つの temperature の値につき 12 パターン算出した平均正答率からさらに平均を取り，temperature の値全体での平均正答率を算出した．

例題の種類を変更した際の平均正答率を比較した図として，図 4 と図 5 を示す．図 4 は算出した 48 パターンのうち同一の例題内容毎でさらに平均を取って平均正答率を算出した結果である．また，図 5 は算出した 48 パターンのうち例題数毎でさらに平均をとって平均正答率を算出した結果である．

5.9 temperature の値に関する考察

図 3 より temperature の値を 0.1 から 0.7 まで 0.2 ずつ変えて

表 5: 正判定例一覧

番号	判定例 (先行文)	判定例 (後続文)
1	ブチャラティ、おかっぱのくせにかっこよすぎるな	そりゃ世界一かっこいいおかっぱだから
2	もーほんとにワンピース買いすぎ 引きこもりのくせに	どこも行かないのにわたしも買っちゃう
3	カップヌードルのにんにく豚骨美味しかったのでオススメ !!!!	おいしそー
4	インスタ繋がってない子繋がろーぜー !!!!	繋がりたいです！
5	マンガ読みたいけど、眠いから寝るおやすみ～！	おやすみ！
6	８連勤がようやく終わった	おっつおっつですぞぉ！
7	ふっかさん、まさかの天然 www	まあ、それで言うと阿部ちゃんに限って『アイドルの引き出しが切れる』とか無いと思うので、有吉さんの言うとおりだなとは思っていた (笑)
8	腰痛のためお休みです。椅子に座れる状態でないため	ゆっくり休んで下さいね !! お大事に
9	ショールームありがとうございました	配信お疲れ様
10	あと気になってんのはチェンソーマンとブルーロックとカッコウの許嫁とビースターズ、と、DB	いっぱいですね　ブルーロックは今年アニメもあるから気になってます!!

表 6: 負判定例一覧

番号	判定例 (先行文)	判定例 (後続文)
1	最近毎日何かしらで寝落ちてる自分を恨む	保証大丈夫やた？
2	お仕事終わったので帰ります…	ゴルフ沼へようこそ笑
3	連絡をして、返事を待っている状態です。もうしばらくお待ちください	ジャンプ力ヤバいそして可愛い
4	家ついちゃ	まさかのパワー 12 !? w
5	昨日の昼～今日の昼までろくな食事をしてなかったので、夜はいいもの食べた	私は RIZINLIVE で買ったよ
6	梅田のランワン来たけどやっぱり人多いな	ゆっくり休んで下さいね !! お大事に
7	中学のめっちゃ可愛い子から連絡きた	早く良くなりますように！ゆっくりしてくださいね
8	え、本当に妊娠してた。どうしよ！嬉しい！アメリカでの妊娠、大丈夫か私	肉じゃが食べたい
9	久しぶりにアニメ見たいな　おすすめのアニメありますか？	本番中にダメ出し (笑)
10	産まれました !! 風船飛んでるの嬉しい	事前に入力できますよ !!

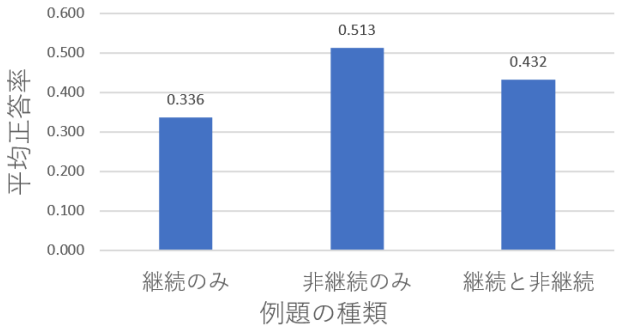


図 4: 例題内容比較

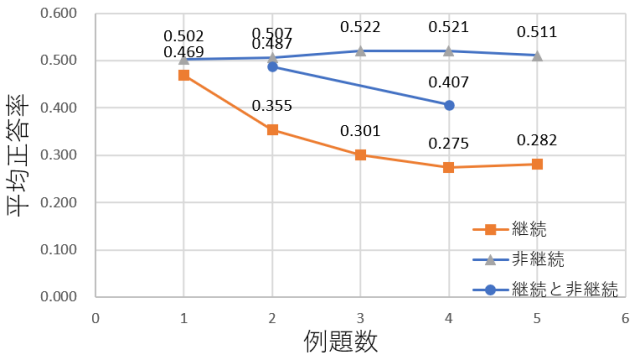


図 5: 例題数比較

実験したが，平均正答率に変化はみられなかった．本実験では別に設定したハイパーパラメータである max_new_tokens を 10 に設定しており，10 トークン以上のテキストが生成されないように制限していた．また，プロンプトでは回答は「YES」か「NO」のどちらかになるように指示していた．このことから生成させるテキストのトークン数を制限し，回答の形式をプロンプト内で指示していれば temperature による影響は小さいと考える．

また，今回の実験では temperature を変えたことによる大きな変化は見られなかったが，平均正答率の観点から temperature が 0.1 のときが会話継続判定において有効であると考える．

5.10　例題内容に関する考察

図 4 から例題の内容を変えた場合には平均正答率に違いがみられた．話題が継続していない (非継続) 例題のみの条件のときが最も平均正答率が高く，話題が継続している (継続) 例題のみの条件のときが最も平均正答率が低かった．本実験では正解

ラベルの付与を明らかに話題が継続していないやりとりを非継続，話題が継続しているもしくは判断が難しいやりとりに関しては継続という判断基準のもと筆者一人で行った．そのため非継続の例題を与えた条件のほうが継続した例題を与えた条件よりもデータセットに即した判断基準になっており平均正答率が高くなったと考える．また，継続の例題と非継続の例題を混ぜた場合では継続の例題によって判断基準があいまいになったため非継続のみの例題のときと比べて平均正答率が低くなったと考える．

5.11 例題数に関する考察

図 5 から例題の数を変えた場合でも平均正答率に違いがみられた．非継続の例題では例題数が増加すると平均正答率も増加し，それ以外では例題数が増加すると平均正答率が減少した．このことや 5.10 から例題がタスクに適していないと例題数を増加させることで判断基準がよりあいまいになり，悪影響になってしまっていると考える．また，GPU のメモリの関係上例題数を増やすとプロンプトの文量が増え実装が困難になる．そのためこれ以上の例題数の増加は見込めないため shot 文以外のプロンプト記法を検討する必要があると考える．

6 おわりに

本論文では LLM を用いて，会話の主題が同一なのか転換されたのかを判定することを目的とした．この目的から本論文では複数の会話文から文章ペアに変換する前処理や，LLM のハイパーパラメータの設定とプロンプトの作成手順，LLM の回答文から判定部分のみの抽出方法を提案した．LLM のハイパーパラメータである temperature の値やプロンプトで与える例題の種類や数を変えて計 48 パターン分の会話継続判定を行った．この結果 temperature による会話継続判定への影響は小さかったが，本実験で設定した temperature の値の中で最も小さかった 0.1 が平均正答率が最も高かったため会話継続判定において有効であることがわかった．また，例題の種類による影響が高くなりそれに応じて例題数による影響も高まることが分かった．今後は shot 文より会話継続判定に適したプロンプト記法の調査や，3 回以上のやりとりに着目した会話継続判定での実験が必要であると考えている．

文　献

[1] 東中 竜一郎，船越 孝太郎, 小林 優佳, 稲葉 通将，“対話破綻検出チャレンジ，”第 75 回 言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム)，人工知能学会研究資料 SIG-SLUD-75-B502 巻，pp.27-32，2016.
[2] 稲葉 通将, 高橋 健一，“Long Short-Term Memory Recurrent Neural Network を用いた対話破綻検出，”第 78 回 言語・音声理解と対話処理研究会，SIG-SLUD, Vol. B5, No. 02，pp.57-60，2015.
[3] 田中 涼太, 李 晃伸，“Quasi-Recurrent Neural Networks に基づく対話モデルを用いた対話破綻検出，”2018 年度人工知能学会全国大会論文集 (第 32 回)，2018.
[4] 藤田　俊之, 小林　亜樹,”マイクロブログにおける会話を利用した災害情報採取手法，”電子情報通信学会論文誌 B，J106-B(1), pp.1-12，2023-01-01.
[5] “Few-Shot プロンプティング”, https://www.promptingguide.ai/jp/techniques/fewshot(参照日 2024-1-9)
[6] “Xwin-LM-13B-V0.1”，https://huggingface.co/Xwin-LM/Xwin-LM-13B-V0.1(参照日 2024-1-9)

数個の発話と疑似文脈生成に基づくキャラクタ対話応答生成

福岡　啓人† 　宮森　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603-8047 京都府京都市北区上賀茂本山
E-mail: †{g2054146,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，キャラクタを模倣した対話応答生成手法を提案する．キャラクタ応答生成では，数個のペルソナでのキャラクタ定義，キャラクタ対話データの収集コストが課題であるため，数発話のみを利用した応答生成タスクに焦点を当てる．既存手法では，キャラクタ発話に対する適切な前発話を対話データから検索することで疑似対話文脈を作成し，キャラクタ発話のみのプロンプトに比べ，よりキャラクタの個性を反映した応答を生成している．しかし，キャラクタ発話に対して，適切な前対話が取得できない場合，疑似対話文脈も適切な内容とならず問題が生じる．そこで，本稿では，キャラクタ発話に対応する前発話を生成することで，疑似対話文脈を作成する手法を提案する．これにより，既存の対話データのみに依存せず，より柔軟にキャラクタ発話に適した疑似対話文脈を活用できることが期待される．具体的には，生成モデルと評価モデルによって，生成と評価を繰り返し，より適切な疑似対話文脈を作成する．実験では，提案手法とベースライン手法による生成結果について，キャラクタ特徴をどの程度反映しているか評価する．

キーワード　キャラクタ，対話システム，言語モデル，プロンプト生成，自然言語処理

1　はじめに

近年，対話システムは，大規模な言語モデルの開発により驚くべき進歩を遂げている [1] [2] [3]．しかし，ユーザは，対話システムの話し方に一貫性がない場合，情緒不安定，説得力がない，人間関係の距離感がうまく取れていないと感じてしまう [5]．このような問題に対し，個性を付与することで，一貫した話し方を維持することができ，ユーザとのより円滑なコミュニケーションをもたらすことができる．本稿では，個性的で魅力的な対話システムを構築するための有望な方向性として，キャラクタを模倣することを検討する．キャラクタのような架空の人物を模した対話システムを構築する場合，2 つの大きな課題が指摘されている．

（1） プロフィール文のようなペルソナをわずかに与えるだけでキャラクタを定義することは困難である．

（2） キャラクタの対話データは十分になく，学習用のデータセットを手作業で作成することはコストが掛かり，非効率である．

この 2 つの課題に対処するために，いくつかのキャラクタの発話のみが利用可能であることを前提条件とする応答生成タスクに取り組む．これらは，以下の理由に基づいている．

（1） キャラクタの個人的特徴やスタイルが発話に内在しているため，キャラクタを模倣した応答を生成するための有用な手がかりを提供している [4]．

（2） 対象キャラクタの数個の発話のみを収集することは，文脈と発話の対からなる完全な対話データを構築することに比べ，コストがかからない．

従来研究では，このタスクに対して，キャラクタ発話から疑似対話文脈を生成するプロンプトを作成することで，キャラクタの特徴を反映した応答を生成した．疑似対話文脈を生成するには，いくつかのキャラクタ発話しか利用できないため，あらかじめ用意された対話データセットから適切な前発話を選択する検索モデルを構築した．既存のベースラインモデルに比べ，よりキャラクタの特徴を反映した応答を生成することが確認された．しかし，あらかじめ用意された対話データセットから前発話を検索する手法では，キャラクタ発話に対して，適切な前発話が取得できない場合，疑似対話文脈も適切な内容とならず問題が生じる．

そこで，本稿では，キャラクタ発話に対応する発話を対話データセットから検索するのではなく，生成することで擬似対話文脈を作成する 2 つの手法である反復的疑似対話文脈生成モデルを提案する．反復的疑似対話文脈生成モデルは，生成モデルと評価モデルを交互に繰り返すことで，より適切な疑似対話文脈を作成する．具体的には，生成モデルは，ベースとなる疑似対話より適切な新たな疑似対話文脈を生成する．次に，評価モデルは，ベースとした疑似対話文脈と新たな疑似対話文脈のどちらがより適切な対話文脈かを選択する．その後，評価モデルが選択した疑似対話文脈を，生成モデルに入力するベース疑似対話文脈とし，また，疑似対話文脈を生成する．これを繰り返すことで，より適切な疑似対話文脈が作成されることが期待される．

本論文の貢献は，以下である．

（1） キャラクタ発話のみからそのキャラクタに沿った応答を生成するタスクに取り組んだ．

（2） 生成と評価を繰り返すことで，より良い疑似対話文脈を生成する手法を提案した．

（3） 疑似対話文脈が応答生成に及ぼす影響を明らかにした．

本論文の構成は，以下の通りである．2 節で関連研究について述べ，3 節で提案手法について詳述する．4 節で実験内容や

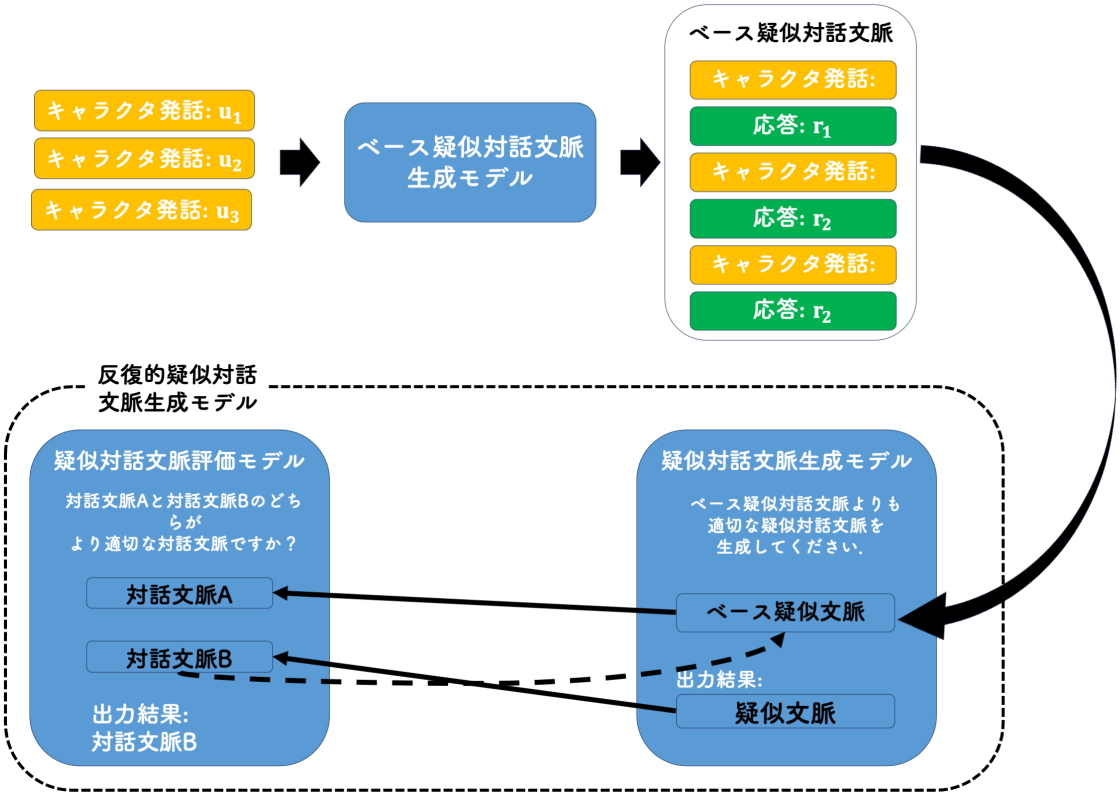


図 1　提案手法の概要

実験結果，考察を示し，5 節では，結論と課題を述べる．

2　関 連 研 究

2.1　キャラクタ対話システム

キャラクタを再現する対話システムの構築に関する研究では，キャラクタの対話データ収集にコストが掛かることが共通の課題であり，収集コストを抑える手法が提案されている．「なりきり質問応答」によりキャラクタらしい発話ペアデータを収集する方法が提案されている [7]．なりきり質問応答とは，複数のユーザが，特定のアニメ等の登場人物に質問したり，なりきって質問に回答することで，質の高い発話ペアを収集する手法である．この得られた質問と回答のペアから，適切な回答を抽出する発話生成モデルを構築するニューラルネットワークを用いた発話生成手法を提案している．しかし，Web サイトを開設する必要やその Web サイトを多くのユーザに認知してもらうことは，特に個人では難しく，発話ペアを収集するコストは高いままである．ペルソナ情報とキャラクタ発話を用いたキャラクタ対話システムの研究 [8] では，小説を題材とし，キャラクタのペルソナとキャラクタ発話を自動抽出するためのアノテーションと擬似的なキャラクタ応答ペア生成モデルが提案されている．キャラクタのペルソナを利用することで，よりキャラクタらしい応答生成につながることが示されている．さらに，強化学習を用いることで，キャラクタらしさを持つ個々の発話データを使用して，発話にキャラクタらしさを持たせる手法が提案されている [9]．対話形式のデータを必要とせず，単一の発話データのみで実現できる特徴がある．しかし，これらの提案手法は，言語モデルをファインチューニングするために，数万以上のデータを用意する必要があり，収集コストは高い．キャラクタの数個の発話のみが利用可能とするタスクに取り組んだ研究として，キャラクタ発話に適切な応答を対話データセットから検索し，疑似文脈を作成する手法が提案されている [6]．キャラクタ発話だけのプロンプトに比べ，キャラクタ発話とその応答で構成される疑似文脈を用いたプロンプトのほうが，よりキャラクタの特徴を反映した応答生成が可能になることを確認した．しかし，キャラクタ発話に対して，適切な対話データが取得できない場合，疑似文脈も適切な内容とならず問題が生じる．

2.2　大規模言語モデル

大規模言語モデルは，Zero-Shot プロンプティングでタスクを解く能力を有しているが，より複雑なタスクを解決することができない．いくつかの例を提供する Few-Shot プロンプティングは，より複雑なタスクでも解くことが可能になる [10]．対話システムにおけるプロンプト学習の調査では，より高いパラメータを持つ言語モデル，より多くの例を提供した Few-Shot 学習がほとんどの場合で，優れたパフォーマンスを示した [11]．

既存の自然言語処理タスクに対する評価は限界があり，大規模言語モデルに大規模言語モデルの評価をさせる考え方 (LLM-as-a-judge) [16] が普及し始めている．基本的な評価の流れは，以下である．

（1）　事前に大規模言語モデルを評価するための質問を準備する．

（2）　比較対象とするいくつかの大規模言語モデルに質問を投げかけ，それぞれの出力を保存する．

（3） 最後に，大規模言語モデルがプロンプトをもとに出力した内容を比較し，最も良い出力結果を判断する．
提案手法で用いる疑似対話文脈の評価は，この考え方を参考にし，大規模言語モデルに判断させる．

3 提案手法

キャラクタ発話でプロンプトをデザインする最も単純な方法は，キャラクタの発話例をプロンプトの最初に追加することだが，キャラクタの特徴を反映していない応答を生成する傾向がある．この問題に対処するために，従来手法では，キャラクタ発話から擬似対話文脈し，プロンプトに利用することで，よりキャラクタの特徴を反映する手法を提案している [6]．具体的には，対話データセットからキャラクタ発話に適切な応答を検索することで，擬似対話文脈を生成している．しかし，キャラクタ発話に対して，適切な対話データが取得できない場合，疑似文脈も適切な内容とならず問題が生じる．

そこで，本稿では，キャラクタ発話に対応する応答を生成することで，疑似文脈を作成する手法を提案する．

3.1 ベース疑似対話文脈生成モデル

ベース疑似対話文脈生成モデル $PDCG_{base}$ は，K 個のキャラクタ発話 $x = \{u_1^c, u_2^c, ..., u_K^c\}$ を入力し，疑似対話文脈 $h_{base} = \{u_1^c, r_1, u_2^c, r_2, ..., r_{K-1}, u_K^c\}$ を生成する．

$$h_{base} = PDCG_{base}(x)$$

ベース疑似対話文脈生成モデルとして，ELYZA 社が公開している「elyza/ELYZA-japanese-Llama-2-7b-instruct」を利用した．しかし，一度に，K 個のキャラクタ発話 x をモデルに入力し，疑似対話文脈を出力すると，疑似対話文脈に抜けがあるなど，安定して生成できなかった．そこで，期待する出力を得るために，elyza/ELYZA-japanese-Llama-2-7b-instruct モデルを追加学習した．具体的には，日本語日常対話コーパス [18] と日本語の Persona-chat，EmpatheticDialogues [1] を使い，elyza/ELYZA-japanese-Llama-2-7b-fast-instruct モデルを QLoRA [17] で追加学習した．これにより，追加学習前の elyza/ELYZA-japanese-Llama-2-7b-instructt モデルより安定した出力結果を得ることができた．

3.2 反復的疑似対話文脈生成モデル

ベース疑似対話文脈生成モデルでは，生成した疑似対話文脈が適切な文脈になっていない可能性がある．そこで，疑似対話文脈生成モデル $PDCG$ と疑似対話文脈評価モデル $PDCE$ で疑似対話文脈を繰り返し，生成，評価することで，適切な疑似対話文脈を作成する．ベース疑似対話文脈生成モデル $PDCG_{base}$ と同様の方法で，ベースとなる疑似対話文脈 h_{base} を生成する．疑似対話文脈生成モデル $PDCG$ に疑似対話文脈 h_{base} を入力し，新たに擬似対話文脈 h_{new} を生成する．このとき，疑似対話文脈生成モデル $PDCG$ に指示として，ベースとなる疑似対話文脈 h_{base} よりも適切な疑似対話文脈 h_{new} を生成するように指示する．

$$h = PDCG(h_{base})$$

次に，疑似対話文脈評価モデル $PDCE$ はベースとした疑似対話文脈 h_{base} と新しく生成した疑似対話文脈 h_{new} を受け取り，どちらの疑似対話文脈がより適切なのかを出力する．ただし，ベースとした疑似対話文脈 h_{base} を選択した場合は e_{base}，新しく生成した疑似対話文脈 h_{new} を選択した場合 e_{new} とし，出力は $result = e_{base} \vee e_{new}$ とする．

$$result = PDCE(h_{base}, h_{new})$$

疑似対話文脈評価モデル $PDCE$ が選択した疑似対話文脈を疑似対話文脈生成モデル $PDCG$ を入力し，新たに擬似対話文脈を生成する．これを繰り返すことで，より適切な疑似対話文脈を生成する．ベース疑似文脈生成モデルと同様に，疑似対話文脈生成モデルの出力が安定して得ることができなかったため，QLoRA で追加学習した．

4 実　　験

4.1 実験設定

提案手法がキャラクタの特徴を反映した応答を生成できているかを検証するために，直前の発話だけから適切な発話を生成するシングルターンで検証した．検証したキャラクタは，男性キャラクタ 2 人，女性キャラクタ 9 人である．それぞれの手法に対して，各キャラクタ 50 個の応答を生成した．疑似対話文脈の評価，キャラクタ応答の生成には，すべて ELYZA 社が公開している elyza/ELYZA-japanese-Llama-2-7b-instruct モデルを利用した．

4.2 評　　価

評価には，一般的に対話システムに用いられる評価指標 BLEU スコア [14]，ROUGE [15] を利用する．更に，生成された応答がキャラクタの特徴を反映しているかを評価するために，text style transfer タスクの先行研究 [12], [13] と同様に，評価対象であるキャラクタ分類器を作成し，分類器がターゲットのキャラクタを予測する平均確率を測定する．キャラクタ分類器の正解率は，63.84% であった．

4.3 比較手法

4.3.1 ベース疑似対話文脈生成モデル

ベース疑似対話文脈生成モデルから出力されたベース疑似対話文脈をプロンプトに与え，直前の発話に対するキャラクタの応答を生成する．

4.3.2 キャラクタ発話のみ

対話形式ではない，キャラクタ発話のみをプロンプトに与え，直前の発話に対するキャラクタの応答を生成する．

4.3.3 実際の対話文脈

この手法では，疑似対話文脈ではなく，対話データセット内に存在する実際の対話文脈をプロンプトに与え，直前の発話に対するキャラクタの応答を生成する．

表 1　各手法の評価結果

手法	キャラクタ分類	BLEU	ROUGE2
反復的疑似対話文脈生成モデル	25.84%	0.4489	0.0189
ベース疑似対話文脈生成モデル	33.33%	0.6695	0.0320
キャラクタ発話のみ	31.18%	0.9295	0.0305
実際の対話文脈	32.36%	1.0428	0.0370
真の応答	58.90%	—	—

4.4 結　　果

実験の結果を表 1 に示す．キャラクタ分類の結果は，ベース疑似対話文脈生成モデルが最も高い結果となった．提案手法である反復的疑似対話文脈生成モデルは最も低い結果となった．BLEU，ROUGE2 は全体的に低い結果であった．

4.5 考　　察

キャラクタ性を反映しているかを評価するキャラクタ分類では，提案手法である反復的疑似対話文脈生成モデル以外の手法がかなり近い値となり，ベース疑似対話文脈生成モデルが最も高い結果であった．原因として，反復的疑似対話文脈生成モデルの疑似対話生成モデルが適切に生成できていないためだと考えられる．疑似対話生成モデルから出力された結果は，同じ発話の繰り返しなど不適切な疑似対話文脈の生成が多く見られた．ベース疑似対話文脈より文脈として一貫性のない疑似対話文脈となった．提案手法を除く手法では，キャラクタ分類の結果に大きな違いは見られなかったため，対話文脈を含めたプロンプトがキャラクタ応答の生成に効果的とは言えない．しかし，不適切な対話文脈をプロンプトに含めるとキャラクタ応答の生成に悪影響を与えることが考えられる．

5 ま と め

本稿では，いくつかのキャラクタの発話だけでキャラクタを模倣した応答を生成するタスクに取り組んだ．このタスクを解決するために，いくつかのキャラクタの発話から疑似対話文脈を生成する手法を提案した．しかし，提案手法は，よりキャラクタを反映した応答を生成することができなかった．プロンプトエンジニアリングの研究では，プロンプトの与え方によって，大きく性能が変わることが明らかにされている．そのため，今後は，別のプロンプトによる応答生成に取り組む予定である．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] 杉山弘晃, et al. Transformer encoder-decoder モデルによる趣味雑談システムの構築. In: 人工知能学会研究会資料 言語・音声理解と対話処理研究会 90 回 (2020/11). 一般社団法人 人工知能学会, 2020. p. 24.
[2] 山崎天, et al. ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム. In: 人工知能学会研究会資料 言語・音声理解と対話処理研究会 93 回 (2021/11). 一般社団法人 人工知能学会, 2021. p. 113-118.
[3] ROLLER, Stephen, et al. Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637, 2020.
[4] BOYD, Alex, et al. Large scale multi-actor generative dialog modeling. arXiv preprint arXiv:2005.06114, 2020.
[5] 東中竜一郎; 稲葉通将; 水上雅博. Python でつくる対話システム. 株式会社 オーム社, 2020.
[6] HAN, Seungju, et al. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. arXiv preprint arXiv:2204.10825, 2022.
[7] 石井亮, et al. 既存のアニメキャラクタを用いたテキスト対話システム構築手法. In: Human-Agent Interaction Symposium. 2018.
[8] 秋山一馬; 稲葉通将. 小説から生成した疑似対話に基づくキャラクター対話システムの構築. In: 人工知能学会全国大会論文集 第 36 回 (2022). 一般社団法人 人工知能学会, 2022. p. 2A6GS602-2A6GS602.
[9] 清水健吾; 上垣貴嗣; 菊池英明. 強化学習を用いてキャラクタらしさを付与した雑談応答の生成. In: 人工知能学会研究会資料 言語・音声理解と対話処理研究会 94 回 (2022/3). 一般社団法人 人工知能学会, 2022. p. 06.
[10] BROWN, Tom, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020, 33: 1877-1901.
[11] MADOTTO, Andrea, et al. Few-shot bot: Prompt-based learning for dialogue systems. arXiv preprint arXiv:2110.08118, 2021.
[12] LI, Juncen, et al. Delete, retrieve, generate: a simple approach to sentiment and style transfer. arXiv preprint arXiv:1804.06437, 2018.
[13] RILEY, Parker, et al. TextSETTR: Few-shot text style extraction and tunable targeted restyling. arXiv preprint arXiv:2010.03802, 2020.
[14] PAPINENI, Kishore, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311-318.
[15] LIN, Chin-Yew. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. 2004. p. 74-81.
[16] ZHENG, Lianmin, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685, 2023.
[17] DETTMERS, Tim, et al. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314, 2023.
[18] 赤間怜奈, et al. 日本語日常対話コーパスの構築. . 言語処理学会第 23 回年次大会発表論文集, 2023, 108-113.

付　　録

1 生成結果

生成した疑似対話文脈や応答を下記の表に示す.

表 A·1 生成結果

入力	涙で弱まったあなたなら私の意見などすいすい受け入れてくれそうだっから
正解の応答	顔文字使ったことで台無しだボケ
	反復的疑似対話文脈生成モデル
対話文脈	次、この顔文字でお前は俺に何を言いたいんだよ え？どうしたの？ 泣いてる顔文字使って俺に何を言いたいんだって質問に変える な、ななんて 世間一般的には弱いって言われてるが俺は涙ごときで変わる人間じゃねえんだ な、ななんて お前のじゃねえだろ！顔文字に涙代用させてんじゃねえよ！
出力	何に？
	ベース疑似対話文脈生成モデル
対話文脈	次、この顔文字でお前は俺に何を言いたいんだよ 泣かしたい 泣いてる顔文字使って俺に何を言いたいんだって質問に変える じゃあ泣いてやる 世間一般的には弱いって言われてるが俺は涙ごときで変わる人間じゃねえんだ 泣かせねぇ お前のじゃねえだろ！顔文字に涙代用させてんじゃねえよ！
出力	俺は涙ごときで変わる人間じゃねえんだ。
	キャラクタ発話のみ
キャラクタ発話集合	次、この顔文字でお前は俺に何を言いたいんだよ 泣いてる顔文字使って俺に何を言いたいんだって質問に変える 世間一般的には弱いって言われてるが俺は涙ごときで変わる人間じゃねえんだ お前のじゃねえだろ！顔文字に涙代用させてんじゃねえよ！
出力	お前はもう俺のことを泣かせようとしてるのか？
	実際の対話文脈
対話文脈	次、この顔文字でお前は俺に何を言いたいんだよ この顔文字が何をしているかくらいわかるでしょう 泣いてる顔文字使って俺に何を言いたいんだって質問に変える 男性は女性の涙に弱いんでしょう？ 世間一般的には弱いって言われてるが俺は涙ごときで変わる人間じゃねえんだ 私の涙ではダメだったというの… お前のじゃねえだろ！顔文字に涙代用させてんじゃねえよ！
出力	じゃあ、俺が涙で弱ってないときにお前の意見を受け入れる気があるのかよ！

性格による応答の多様性を考慮したキャラクタ対話データセット構築

村田　樹†　宮森　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054326,miya}@cc.kyoto-su.ac.jp

あらまし　ペルソナ対話を実現する対話システムは，多様な言語行動やユーザとの豊かなやりとりが期待できるため，ユーザの対話意欲喚起の観点から重要である．近年の LLM は，流暢な言語を介して対話する能力をもつ一方，特定のペルソナ（キャラクタ）を反映した対話を一貫して実現するには至っていない．そこで本稿では，キャラクタ対話実現のための対話データセット構築手法を提案する．ペルソナが，プロフィールと性格から構成されるものとみなし，性格に応じた多様な対話を生成するモデルを構築する．具体的には，性格に応じた感情分布からその性格を反映した発話を生成するタスクと，性格と直前発話や現発話からその性格を反映した対話を生成するタスクを通して，任意キャラクタの性格を反映した多様な対話の生成を実現する．実験では，提案手法で生成した対話データの質を既存手法と比較評価する．

キーワード　ペルソナ，データセット構築，対話システム，LLM，自然言語処理

1　はじめに

近年，大規模言語モデル（LLM）は目覚ましい発展を遂げ，雑談やさまざまなタスクをユーザとの対話活動の中で実行することが主流となった．従って，LLM は，タスクの是非を問わずに対話システムとしての側面を保持していると捉えることができる．LLM がユーザとの対話活動を軸として使用される上で重要となるのが，対話の一貫性である．対話の一貫性が欠落すると，ユーザは LLM が情緒不安定であると感じ，その内容に説得力を感じ得ずユーザの対話意欲を低下させるという問題がある．

これに対する解決策の一つにペルソナ対話が存在する [1]．ペルソナ対話とは対話システムに個性を獲得させ，対話の一貫性の維持やユーザの対話意欲喚起を図る試みである．主にはプロフィールと呼べる個人間の情報を発話内に織り交ぜることが一般的である [2] [3]．中でも，日本語は人称表現や語尾表現などの口調表現が色濃く変化する言語であるため，対話の一貫性の維持には一層重要な手法であると言える [4] [5]．

ここで，本論文では，ペルソナを抽象的な人物像を構成する属性の集合体と定義する．一方，ペルソナの各属性に特定の値が代入され，具体的な個として確立されたものをキャラクタと呼称する．例えば，ペルソナの属性には「名前」が存在し，キャラクタの名前属性には「山田太郎」が存在する．

各属性には，基本的に，事実として活用されるものと，言動に影響を与えるもののいずれかに分類されるが，両方の側面を持つものも存在する．例えば，事実として活用すべきものには視力や利き腕・文化祭の思い出などが挙げられる．一方，言動に影響を与えるものには天邪鬼・素直・怒りっぽいなどが挙げられる．また，どちらの側面も持つものとしては，男性・親方・小学生などが挙げられる．これらは一見事実として活用すべきであるが，同時にステレオタイプとして言動にも影響を与えることがある．以降では，全ての事実として活用すべき属性をプロフィール，全ての言動に影響を与える属性を性格と呼称する．

プロフィールと性格に分類した理由として，各属性の反映方法が存在する．プロフィールは，対話の状況に応じて取捨選択しながら，文面に直接表すことが望ましい．そのため，データベースの構築と LLM 自身による逐次的なプロンプトへの挿入が有効である．一方で，性格は継続的に安定して反映されるべき属性であり，文面上に表現することが適切であるとは限らない．

また，性格属性を反映することは，Instruction で得た振る舞いと競合したり，その属性を反映することを拒絶する場合がある特に，言語モデルが特定の人物に対して偏見を持つことが示唆されたという報告があることかも，基盤モデルの持つ性格への理解をそのまま使用するのは得策ではないと考えられる [6]．キャラクタとの対話時にはユーザの期待する倫理観はキャラクタに依存するため，LLM が持つ倫理観は過敏なものになることがある．従って，性格の反映には追加学習が適切であると考えられる．特に，計算資源や反映するキャラクタの選択自由性の観点から各キャラクタの LoRA モデルの構築が望ましい [7] [8]．

しかしながら，LoRA 学習を行う際，そのためのデータセットの用意は非常に困難である．キャラクタ対話モデルの学習には，特定キャラクタのバリエーションに富んだ対話を数百件程度収集したい．しかし，これは明らかに容易ではなく，特に，性格を反映するための対話を収集することが難しい．そこで本稿では，LLM での多様なペルソナ対話実現のための，キャラクタ対話データセット構築を目的としたキャラクタの性格の反映に特化した対話データセットの構築に取り組む．

本稿では次のような手法を提案する．まず，前発話の類似性と，性格による応答の多様性に着目してショートストーリー（SS）由来のデータセットを構築する．そして，このデータセットを用いて性格に応じた感情分布からその性格を反映した発話を生成するタスクと，性格と直前発話や現発話からその性格を

反映した対話を生成するタスクを通して，任意の性格で構成されるペルソナを反映した対話データセット生成を実現する．既存手法とは性格によって応答の方向性そのものに変化があることに着目し，上記二つのモデルを利用することが違いである．これにより，性格による現発話の変化の反映に特化した対話を出力することが期待される．

提案手法の重要性の確認のため，次のような実験を行う．

実験 1 対話データ生成評価

実験 2 学習したキャラクタ対話モデルのキャラクタ再現性評価

実験 1 では学習時に使用したキャラクターや，未知のキャラクターについて生成する対話データの再現性を評価する．SS から学習に利用していない対話を収集し，既存・提案手法で対話データセットを構築し，その再現性を LLM によって評価する．

実験 2 では，両手法で生成した対話データを利用して，実際にキャラクタ対話モデルの学習を行い，生成した応答の内容を評価する．

本論文の貢献は以下のとおりである．

1 簡易なペルソナ対話の実現による対話の一貫性の維持やユーザの対話意欲喚起

2 より高度なペルソナモデルの実現

2 について補足しておくと，LLM は考慮するペルソナに応じて心理テストの解答にも指向性を持つとの報告があるので，より確かなペルソナ対話の実現は，LLM のより高度なペルソナモデルとしての活用に期待が持てる [9]．

本論文の構成は次のとおりである．2 節で関連研究について述べ，3 節で提案手法を説明する．4 節でデータセットについて説明し，5 節で実験内容や結果について述べ，考察を記す．6 節で結論と課題をまとめる．

2 関 連 研 究

本節では関連研究と筆者の手法との違い等について述べる．関連研究は，LLM でペルソナ対話システムを実装した研究と対話データセット構築に関する研究を挙げている．

2.1 LLM によるペルソナ対話システム

2.1.1 ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム

この論文では，HyperCLOVA をベースとした対話システムを実装している [10]．プロンプトを利用した応答生成モジュールと, 生成された応答の候補から不適切な発話を棄却するフィルタリングモジュールを中心に対話システムを構築している．また，プロフィールを適切に使用した応答生成のために，プロフィールの利用指針を組み立ててから対応する応答生成を行なっている．しかし，この利用指針の組み立てには文の類似度などを使用しており，ペルソナを考慮した言語モデルによる指針立ては行われていない．従って，筆者の提案手法で生成した対話データセットで目指す，キャラクタの性格を反映した対話は反映されていない．

2.1.2 Character-LLM: A Trainable Agent for Role-Playing

この論文では，エピソード記憶や時代背景に注目してペルソナ対話システムが構築されている [11]．LLM は各キャラクタのエピソード記憶や時代背景を寄せ集めたデータベースを閲覧し，対話履歴との関連性を踏まえてその時々に最適なデータを選択する．その選択したデータに含まれる事実や心理描写からキャラクタの反映を行なっている．これにより，例えばベートーヴェンが Python プログラムを執筆できるといった対話システムの幻覚を抑制している．しかし，この論文で考慮されている性格は対話履歴がデータベースの特定状況と類似し活用できる場合のみ有効であり，結果としてエピソードが活用できなかった時などに一貫性を破壊させる恐れがある．提案手法で生成した対話データセットで学習した LoRA モジュールによる性格の反映の場合，このような一貫性の破壊の恐れはない．

2.2 対話データセット構築

2.2.1 ChatHaruhi: Reviving Anime Character in Reality via Large Language Model

厳密には，この論文は対話システム構築の提案である [12]．キャラクタの対話データベースを構築し，対話時には対話履歴を基準に対話データベースからサンプルとなる対話を探索し，応答生成に利用している．対話データベース構築の際，既存の対話が少量であるキャラクタは，対話データベースの拡充を行っている．既存の対話を対話履歴として，続くキャラクタの応答や対話を生成することで対話データの拡充を行なっている．しかし，この手法では対話内容が対話履歴に強く依存することや LLM の幻覚や倫理観の影響を受けやすいという問題がある．提案手法では，性格による感情分布を指針として生成する応答に発散性を持たせることや，対話生成を追加学習したモデルで行うことでこの問題の発現を抑制している．

2.2.2 小説から生成した疑似対話に基づくキャラクター対話システムの構築

この論文では任意発話の前発話を予測するモデルを作成し，小説内の本来対話形式ではなかった単一の発話を対話に加工することで，ペルソナ対話データセットの補完を行う試みである [13]．しかし，前発話生成時に話者の性格を考慮していないため，前発話の内容が現発話の文面に従ったものにしかなり得ない．提案手法では前発話の生成に性格を考慮していることや，現発話の生成も行っているという点が違う．また，本論文ではこの研究を既存手法とおく．

3 提 案 手 法

本節では本稿の提案手法について説明する．

3.1 問 題 設 定

まず，改めて本稿の目的を記す．本稿の目的は，キャラクタ対話モデル構築を簡便にするための，性格による応答の多様性を考慮したキャラクタ対話データセットの構築である．そのために，感情分布予測モデルと対話データ生成モデルを併用した

対話データセット生成を提案する．図 1 に提案手法を利用する際の流れを示す．まず，ユーザが特定キャラクタの性格とサンプルとなる発話を 10 個程度用意する．次に，感情分布予測モデルで性格に基づいた感情分布を作成する．感情分布の各値を整数倍し，生成する発話の感情の内訳を決定する．最後に，対話データ生成モデルで不足する台詞や学習時に使用する．

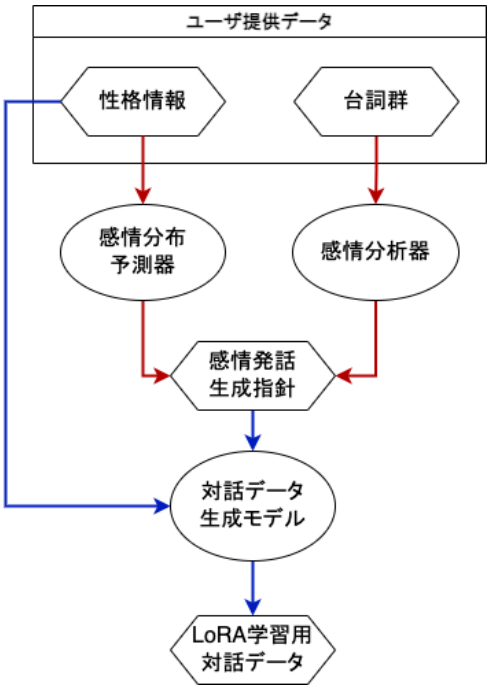


図 1　提案手法利用手順

3.2　研究内容と著作権について

ここで，両モデルの詳細について触れる前に，2024 年 3 月 21 日執筆時現在の生成 AI に関連する著作権関連の情勢を記す．現在の著作権において最も関わり深いのは第三十条（特に第三十条の四）である [14]．著作権者の利益を不当に害する目的や自身または他人に著作物に表現された思想又は感情の享受を目的としない利用を禁止する旨が綴られている．この著作権法により，2023 年 5 月時点では内閣府によって原作者の許諾の無い学習や生成物の私的使用に関しては，現状は合法であると判断された [15] [16]．

一方で，2023 年 11 月末に公開された文化庁の素案では，モデルの学習段階において，上記禁止内容に違反していないかどうかの推認が難しいことについて触れられている [17]．この素案については，2024 年 1 月 15 日に委員会に提出・承諾され，2 月 29 日に募集したパブリックコメントを公開した [18]．また，3 月 15 日には，2 月 29 日付けで提出された最新の素案が了承されたことが発表されている [19]．

本研究の目的はユーザの対話意欲向上のための，特定キャラクタとの対話活動の実現を用意にするためのデータセット構築である．これは提出された素案に倣えば，本研究は他人に著作物に表現された思想又は感情の享受を目的としている，或いは目的に含んでいると判断できると考えられる．従って，執筆時現在では，本研究は法律上の問題はないと考える．

3.3　感情分布予測モデル

3.3.1　感情分布予測モデルについて

感情分布モデルは，キャラクタの性格からそのキャラクタの発話の感情表現の確率分布（感情分布）を予測するモデルである．扱う感情の数を d，任意発話 u における d 次元の感情ラベルを表す one-hot ベクトルを e_u，コーパス内のキャラクタの発話数を N すると，感情分布 E は次のような式で求められる．

$$E = \frac{1}{N}\sum_{i=0}^{N-1} e_i$$

例えば，山田太郎は 23%の確率で喜びの感情表現を持つ発話を応答し，17%の確率で怒りの感情表現をもつ応答をするといった情報が感情分布である．

本稿では，ポジティブ・ニュートラル・ネガティブの 3 感情を扱う．これら 3 感情は一次元的なものであるが，感情分布上では擬似的に別のベクトルとして扱う．仮に一次元的な見方をすると，上記式の one-hot ベクトルを正負の値（ポジティブなら 1, ニュートラルなら 0, ネガティブなら -1）で E を求めることになる．この場合，例えば $[0.4, 0.2, 0.4]$ という感情の起伏が激しいキャラクタと，$[0.2, 0.6, 0.2]$ というあまり感情を表に出さないキャラクタを同じ性格であると扱ってしまうことになる．これを回避するため，本稿では 3 次元として扱っている．

3.3.2　感情分布予測モデルの作成

感情分布予測モデルの作成には，文書分類の手法を応用した．まず，学習用データとして 957 名のキャラクタの性格を 1500token 以内に要約したものを 6 通りずつ，合計 4835 個のキャラクタの性格を用意した．これを入力とし，感情ラベルを出力とするよう感情分類器を学習した．基盤モデルには FacebookAI の xlm-roberta-large を使用した．出力である感情ラベルは，感情分布 E を 50 倍し，各要素を四捨五入した整数部分の分だけあてがった．例えば，$E = [0.2, 0.3, 0.5]$ のキャラクタの場合，ランダムな 5 個の性格がポジティブであるとして 10 回，ニュートラルであるとして 15 回，ネガティブであるとして 25 回学習する．これにより，分類器として追加した最終層の確率分布をそのまま感情分布として扱うことができる．

3.4　対話データ生成モデル

対話データ生成モデルでは，二つのタスクを並行して学習・利用する (図 2)．一つは発話類推タスク（図赤）である．発話類推タスクではキャラクタの性格とサンプルとなる任意のユーザ提供発話から発話を類推するタスクである．これにより，性格から予想されるキャラクタの発話のバリエーションを学習する．もう一つが，対話データ生成タスク（図青）である．対話を行う二人のキャラクタの性格と現発話を発する話者の感情（さらにその台詞を使う場合もある）を入力とし，前発話が類似し，且つ現発話が多様な対話群から選出した対話を出力として学習する．これにより，性格によってどのような返答を行うかを学習する．

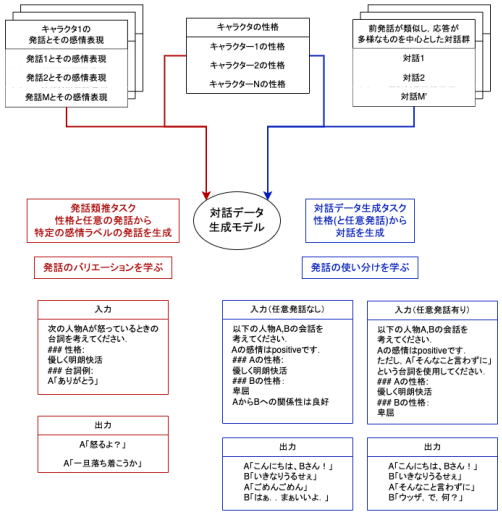


図 2　対話データ生成モデルの学習

4　データセット

ここでは，使用したデータセットについて記す.

4.1　準備データ

4.1.1　SS

ショートストーリー（SS）とは，主に 5 ちゃんねる（旧 2 ちゃんねる）などの Web 掲示板にて匿名のユーザが投稿する小説である．主に魔王や勇者などのステレオタイプを利用した一次創作と特定の版権作品の二次創作である作品に二分される．最大の特徴としては，多くの作品が各台詞の先頭に話者名を記した台本形式で投稿されている点である．よって，キャラクタ ID が付いた対話データとしての活用が非常に容易である．今回は，この内の二次創作の作品に限定して収集を行った.

4.1.2　キャラクタ情報の収集

キャラクタ情報の収集には Web 上の有志による Wiki から収集した．先述の通り言語モデルによって性格や人間性について 1500token 以内で 6 通り要約させた.

4.2　前発話が類似し，且つ現発話が多様な対話群について

前発話が類似し，且つ現発話が多様な対話群とは何かについて説明する．簡単に言えば，性格による応答の変化が強く現れた対話群のことである．例えば,「お茶淹れてくるね」という前発話があったとする．これに対し「ありがとう」と礼を言う場合が考えられるが，他方で「お茶請けはチョコレートで」と，さらに要求を重ねる人物がいる可能性もある．このように性格が違えば現発話に違いが生まれる場合があり，これを中心に学習することで，性格の反映に特化した対話データセットの生成が期待される.

このような対話をコーパスから抽出するため，全ての対話の前発話をベクトル化し，コサイン類似度のスコアが 0.85 以上である他の対話とのグループ化をおこなった．また，現発話の内容が前発話の内容を受けて多岐に渡っているかを確認するため，現発話の話者が画一的なものや，グループ内の半数以上の対話において，現発話が前発話より 1 ターン前の前発話に類似しているようなグループは削除している.

5　実　　験

5.1　研 究 課 題

まず，ここで研究課題を記す.

RQ1 提案手法はどの程度キャラクタを再現したデータセットを構築できるのか

RQ2 提案手法で生成したデータセットは学習にどの程度有用なのか

5.2　実験 1：対話データ生成評価

5.3　実 験 設 定

既存手法として，前発話生成モデルをおく．前発話生成モデルでは，提案手法と同件数の対話をランダムに獲得し，性格を考慮することなく前発話を生成するように学習した．基盤モデルには lightblue の qarasu-14B-chat-plus-unleashed を利用し，それぞれを LoRA 学習で学習した.

キャラクタは既知だがそれぞれのモデルで学習に使用していない対話と，未知のキャラクタ 12 人（男女 6 人ずつ）によるキャラクタの対話を用意する．未知のキャラクタについては半数には話し相手の情報を示し，残り半分には情報を提示しない．情報を提示しない場合の評価を行う理由は，ユーザが実利用する際に，ユーザのキャラクタ情報を必要とするかを確認するためのものである．これら 3 グループそれぞれで対話を 600 件ずつ収集する.

これらの対話を両手法で再現するように生成する．生成した対話を OpenAI の gpt-3.5-turbo によって，どちらが元の対話と近しいかを評価する．この時，評価する問題ごとに既存・提案の順番を入れ替えている．また，3 回評価しても尚評価できないケースに関しては評価を打ち切り，判断不能とした.

5.3.1　実 験 結 果

以下の表 1 に結果を示す．いずれの条件下においても，既存手法の方が提案手法よりも評価が高い結果となった．最も判断不能に陥りやすかったのは既知キャラクタの時であった．また，提案手法の対話生成において，話し相手の情報の有無は，生成時には影響しなことも確認できた.

表 1　実験 1 の評価結果

評価結果	既存手法	提案手法	判断不能
既知キャラクタ	315	204	81
話し相手の情報有り	384	198	18
話し相手の情報無し	372	188	40

5.3.2　考　　察

既存手法が提案手法を上回った要因としては，LoRA モデルのランク設定と推論時のサンプリング等のパラメータの不合が挙げられる．本論文では，前回使用した ELYZA-japanese-Llama-2-13b-fast-instruct より性能の高い基盤モデルである qarasu-14B-chat-plus-unleashed を扱ったが，時間の関係上い

くつかの条件でしか実験を行えていない．前回提出した第二稿においては，提案手法が同等か僅かに評価が高いという結果に落ち着いていたことから，当モデルに最適な調整を行うことで性能向上の余地があると考えられる．

5.4 実験 2：学習したキャラクタ対話モデルのキャラクタ再現性評価

5.5 実 験 設 定

既存手法として，前発話生成モデルをおく．6 人の未知キャラクタの発話データを 200 件抜き出した．既存手法側ではこれら 200 件の発話データから対話データセットを作成した．一方，提案手法側ではさらにその中から 40 件に絞り込み，160 件の発話を類推してから 200 件の対話データを生成した．また，類推した発話については，話し相手はそのキャラクタの話し相手となりうる人物からランダムに宛てがった．基盤モデルである qarasu-14B-chat-plus-unleashed を，両手法で作成した対話データセットを使って LoRA 学習を行い，キャラクタ対話モデルを構築する．それぞれ 5epoch で学習した．

これらのキャラクタ対話モデルに対して，SS から抜き出した 1 ターンの対話 150 件の前発話を入力する．この時の対話の応答者は全てキャラクタ対話モデルが模倣するキャラクタであり，各前発話は 2 回ずつ，合計 300 個のテストケースを与える．各入力に対する出力を，キャラクタの現発話とみなし，その結果を比較する．

評価には，OpenAI の gpt-3.5-turbo を用いた．gpt-3.5-turbo にはキャラクタの発話例として，ランダムに 10 個のキャラクタ発話を与え，サンプルの発話から人物像を類推し，より本人のものと思える発話を選択させた．また，生成した発話群で Self-BLEU スコアを算出し，生成できる応答の多様性を確認した．

5.5.1 実 験 結 果

以下の表 2,3 に結果を示す．表 2 から，評価が同程度のものは存在せず，既存手法か提案手法いずれかの性能が高い傾向にあることがわかる．表 3 も踏まえると，提案手法側の Self-BLEU スコアがほぼ 1 に近いキャラクタが生成する応答の評価が高いことがわかる．一方で，既存手法の方が応答生成の評価が高い場合は，スコアは同程度になっていることがわかる．

表 2　実験 2 の評価結果

キャラクタ	既存手法	提案手法	判断不能
まどか	128	172	0
ほむら	180	118	2
ライナー	174	125	1
悟空	198	102	0
楽	196	104	0
友奈	102	198	0

5.5.2 考　　察

実験 2 から，キャラクタ対話モデルの性能には，発話類推の精度が強く関わっていることが予想される．類推した応答が画一的であれば，応答の多様性と引き換えにキャラクタ性の評価を向上させている．従って，類推する応答が，キャラクタ性を確かに保ちつつ多様になるようにすることで，さらなる精度の向上が見込める．

表 3　各キャラクタの Self-BLEU スコア

キャラクタ	まどか	ほむら	ライナー	悟空	灯織	花京院
既存手法	0.70	0.73	0.63	0.64	0.77	0.65
提案手法	0.95	0.60	0.64	0.68	0.60	0.97

6 ま と め

本稿では，LLM での多様なペルソナ対話実現のための，キャラクタ対話データセット構築を目的としたキャラクタの性格の反映に特化した対話データセットの構築に取り組んだ．提案手法の評価のため，生成できる対話の質や生成した対話データセットによって実際に学習したモデルの性能を評価した．評価の結果，期待したキャラクタ性を反映した対話の生成には至っていないことが明らかとなった．一方で，モデル学習時の LoRA や推論時のパラメータの最適化や，キャラクタ発話の類推により重きを置くことで精度の向上が見込めると判断し，これを今後の展望とする．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, Jason Weston “Personalizing Dialogue Agents: I have a dog, do you have pets too?” arXiv preprint arXiv:1801.07243, 2018.

[2] 光田 航, 東中竜一郎. 李 廷軒, 杉山 弘晃, 水上 雅博, 中村 竜太, 安達 敬武, 川端 秀寿, 吉田 仙, 杵渕 哲也, “なりきり質問応答と大規模言語モデルに基づくなりきり AI の構築” 自然言語処理 30 巻 4 号 p1172-1205 , 2023.

[3] 川本 稔己, 山崎 天, 佐藤 敏紀, 奥村 学 “大規模汎用言語モデルによるペルソナを考慮した応答生成” 言語処理学会 第 28 回年次大会 発表論文集 p1344-1348, 2022.

[4] 森康汰, 服部峻, 高原まどか, 工藤康生 “ツンデレな対話応答 AI の構築コスト削減のためのクロス言語とルールベース個性除去” 情報処理学会研究報告 IPSJ SIG Technical Report, 2022.

[5] 長島大和, 阿倍博信 “対話システムにおけるキャラクタ性構築のための自動語尾抽出方式の検討” DEIM Forum 2021 H25-5, 2021.

[6] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan “Toxicity in ChatGPT: Analyzing Persona-assigned Language Models” arXiv preprint arXiv:2304.05335, 2023.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen “LoRA: Low-Rank Adaptation of Large Language Models” arXiv preprint arXiv:2106.09685, 2021.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer “QLoRA: Efficient Finetuning of Quantized LLMs” arXiv preprint arXiv:2305.14314, 2023.

[9] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, Maja Matarić “Personality Traits in Large Language Models” arXiv preprint arXiv:2307.00184, 2023.

[10] 山崎天, 坂田亘, 川本稔己, 小林滉河, グエントゥン, 上村卓史,

中町礼文, 李聖哲, 佐藤敏紀, “ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム” 人工知能学会研究会資料 第 93 回 言語・音声理解と対話処理研究会, p113-118. 一般社団法人 人工知能学会, 2021.

[11] Yunfan Shao, Linyang Li, Junqi Dai, Xipeng Qiu “Character-LLM: A Trainable Agent for Role-Playing” arXiv preprint arXiv:2310.10158, 2023.

[12] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, Haozhen Sun, “ChatHaruhi: Reviving Anime Character in Reality via Large Language Model” arXiv preprint arXiv:2308.09597, 2023.

[13] 秋山一馬, 稲葉通将 “小説から生成した疑似対話に基づくキャラクター対話システムの構築” 人工知能学会全国大会 (第 36 回) 2A6-GS-6-02, 2022.

[14] e-GOV , 著作権法（昭和四十五年法律第四十八号）, 2023 https://elaws.e-gov.go.jp/document?lawid=345AC0000000048

[15] 内閣府. AI と著作権の関係等について, 2023 https://www8.cao.go.jp/cstp/ai/ai_team/3kai/shiryo.pdf

[16] 内閣府, 生成 AI と著作権, 2023 https://www8.cao.go.jp/cstp/ai/ai_team/7kai/chizai2023.pdf

[17] 文化庁, 文化審議会著作権分科会法制度小委員会（第 5 回）, 2023 https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hoseido/r05_05/

[18] 文化庁, 文化審議会著作権分科会法制度小委員会（第 6 回）, 2024 https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hoseido/r05_06/

[19] 文化庁, 文化審議会著 3 権分科会法制度小委員会（第 7 回）, 2024 https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hoseido/r05_07/

[20] Artur Niederfahrenhorst, Kourosh Hakhamaneshi, Rehaan Ahmad “Fine-Tuning LLMs: LoRA or Full-Parameter? An in-depth Analysis with Llama 2”, 2023 https://www.anyscale.com/blog/fine-tuning-llms-lora-or-full-parameter-an-in-depth-analysis-with-llama-2

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 10:00 ~ 12:10 | T2-A オンライン（Zoom Events)

索引・ストレージアーキテクチャ

座長:北川 博之(筑波大学)
コメンテータ:藤原 真二(日立製作所)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T2-A-4-01] 同時実行B+木におけるロックフリー手続きの改善と実装

*平野 匠真[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 東海国立大学機構名古屋大学大学院情報学研究科)

10:20 ~ 10:45
[T2-A-4-02] 同時実行B+木のマルチバージョン化の検討

*桑村 真生[1]、杉浦 健人[1]、平野 匠真[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学 情報学研究科)

10:45 ~ 11:10
[T2-A-4-03] リモートログ先行書き込み用のストレージ階層追加による効率化

*XU WENKANG[1]、石井 幹大[1]、Dürst Martin J.[1] (1. 青山学院大学理工学部情報テクノロジー学科)

11:10 ~ 11:35
[T2-A-4-04] コンピュート・ストレージ分離アーキテクチャDBを用いた高可用かつ低コストな冗長構成の提案

*林 友佳[1]、渡 佑也[1]、堀内 美聡[1] (1. NTTソフトウェアイノベーションセンタ)

同時実行 B^+ 木におけるロックフリー手続きの改善と実装

平野　匠真† 　杉浦　健人† 　石川　佳治† 　陸 　可鏡†

† 東海国立大学機構名古屋大学大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
Email: {hirano, lu}@db.is.i.nagoya-u.ac.jp, {sugiura, ishikawa}@i.nagoya-u.ac.jp

あらまし　単一コアの性能向上が限界を迎え，複数コアを活用するマルチスレッド処理が主流となってきている．索引技術においてもその傾向は同様であり，CPU のメニーコア化に伴い，マルチスレッド処理を意識した索引構造が多く提案されている．マルチスレッド処理におけるロックを用いた同時実行制御は，スケーラビリティが悪いため，B^+ 木を基にした Bw 木や Bz 木といったロックフリー索引に対して注目が集まっている．しかし，メニーコア環境ではこれらの索引であっても性能が向上せず，ロックフリー索引には更なる改善が必要である．そこで，本研究では B^+ 木をロックフリー化させた新たな索引構造である B^c 木を提案し，提案構造の性能について実験により評価する．実験の結果，B^c 木はキーの書き込みを伴い偏りが大きい場合に，他の索引よりも高い性能を示した．

キーワード　B^+ 木，索引構造，同時実行制御

1 はじめに

ムーアの法則が終わりを迎えたことで，CPU の性能向上は CPU コアの数を増やすメニーコア化へ依存している．現在のコンピュータ技術では CPU に搭載された複数のコアを活用するマルチスレッド処理が主流となっており，マルチスレッドを効果的に使用するソフトウェアやデータ構造などが提案されている．索引技術においてもその傾向は同様であり，近年ではメニーコア CPU 上でのマルチスレッド処理を意識した索引構造 [1–5] が多く提案されている．

代表的な索引構造である B^+ 木 [6] での同時実行制御には改善の余地がある．B^+ 木の同時実行制御ではロックによる排他的処理を用いることによって，データの整合性を保っている．しかし，ロックによる同時実行制御は多数の待ちスレッドが発生しスケーラビリティが悪化することから，マルチスレッド処理を用いるような大規模なシステムにおいてロックによる効果的な管理は困難である [7]．そこで，B^+ 木を基に構造変更時のロックの占有期間を短くした B^{link} 木 [8] や，楽観的な制御を用いることで B^{link} 木における読み取り時のロック取得を不要にした OLFIT [9] などがある．更に，読み取りおよび書き込みの両方においてロックを取得しないロックフリー索引として Bw 木 [10] および Bz 木 [11] が提案されている．しかし，これらのロックフリー索引は書き込み時の競合解消のためにリトライ処理が多発するため，メニーコア環境では性能が向上しない．そのため，ロックフリー索引には改善の余地が残されていると考えられる．

本研究では B^+ 木の同時実行制御においてロックフリーアルゴリズムを活用した，新たな索引構造である同時実行 B^+ 木（concurrent B^+-tree, B^c 木）を提案する．メニーコア環境下での読み書き両方で高い性能を達成するには，スレッド間での同期が必要な箇所を減らし，競合が発生する期間を短くする必要がある．これを実現するために，既存のロックフリー索引である Bw 木および Bz 木の構造を基に，各索引の弱点を克服したデータ構造および手続きについて提案する．そして，既存の索引と比較しながら提案構造の性能について評価する．

本稿の構成は以下の通りである．2 章では，関連研究としてロックの取得期間の短縮を狙った索引やロックフリー索引などについて概説する．次に，3 章で B^c 木の構造について説明し，4 章および 5 章で B^c 木の操作について述べる．そして，6 章で実験による性能評価の結果を示す．最後に，7 章で本稿のまとめと今後の課題について述べる．

2 関連研究

近年ではマルチスレッド環境でも上手く動作する索引が提案されている．その具体例として state-of-the-art な索引を紹介する．その後，B^+ 木に着目した索引としてロック期間の短縮を狙った索引である B^{link} 木と楽観的制御により読み取り時のロック取得を不要にした OLFIT，そして代表的なロックフリー索引である Bw 木および Bz 木について詳しく説明する．

B^{link}-hash [1] は，ハッシュと B^+ 木を組み合わせることにより，単調挿入による競合の削減を狙った索引である．Mass 木 [12] や Adaptive Radix Tree（ART）[13] はトライ木を基に設計された索引である．HydraList [2] は中間ノードに ART を用いており，葉ノードはリンクリストを用いている．また NBTree [3] や uTree [4] は不揮発性メモリに対応した索引であり，可変長の値に対しては別領域に格納することや葉ノードを 2 つの層に分割するなどの手法を用いる．

2.1 B^{link} 木

B^{link} 木 [8] は B^+ 木の各ノードに右兄弟ノードへのリンクを持たせ，構造変更時におけるロックの保持期間を削減した索引構造である．リンクを持つことにより，親ノードからさかのぼらずに隣接するノードに達することができる．また，B^{link} 木はノード内に最大キーを持つことにより，並行する構造変更操

作にも対応できる．例えば，構造変更後の新たなノードが兄弟ノードからリンクによって参照されているが，親ノードからの子ポインタによって参照されていない場合がある．この時に同時に検索操作が起きていた場合，たどり着いたノードの最大キーよりも検索キーが大きければ，構造変更操作が並行して起きたとして兄弟リンクをたどり正しいノードにたどり着くことができる．兄弟リンクと最大キーを持つことにより，ロックを取るノードを最小限に抑え，単一操作内でのロック期間を短縮している．

2.2 OLFIT

OLFIT [9] は B^{link} 木に対してバージョン番号に基づく楽観的ロックを導入した索引構造である．書き込み操作は B^{link} 木と同様にロックを用いて動作する．一方読み取り操作では，ノードをロックしない代わりにバージョン番号を読み取っておく．検索対象を読み取った後，バージョン番号が読み取り前と同じか検証する．バージョン番号が変化していなければ読み取った値を返し，変化している場合には再度読み取り操作を始めからやり直す．また，OLFIT における楽観的読み取りにキューベースのロックを採用し，ロックが解除されてから次のロックが行われる間に読み取りが行えるように設計した OptiQL [5] がある．

2.3 Bw 木

Bw 木 [10] は，直感的にはロックフリーの単方向連結リストと B^+ 木を組み合わせた索引構造である．Bw 木は追記型の差分レコードとマッピングテーブルを採用している．ノードの更新を一括に処理し，単一の CAS 命令を用いてノード間のポインタをインストールできるようメモリ内のデータ構造を設計することにより，木のロックフリー化を実現している [14].

更新内容が記述された差分レコード（delta record）をノードの前に連結リストとして挿入し，索引に対する全ての更新を表す．差分レコードのリストが一定以上の長さを持つとき，ノードの統合操作が行われ，統合後は B^+ 木のノードと同様のものが生成される．

マッピングテーブルはノード間のリンクを論理化し，メインメモリ上の差分更新を可能とする．各ノードは自身の子ノードや兄弟ノードへのポインタを直接持つ代わりにマッピングテーブル上の ID（logical page ID, LPID）を持つ．つまり，マッピングテーブルを用いて LPID を実際のポインタへと変換することでノードを辿る．

2.4 Bz 木

Bz 木 [11] は，B^+ 木の葉ノード内部にロックフリーの固定長配列を持つ索引構造である．Bz 木は，Bz 木の根を示すルートポインタと，複数の内部ノードと葉ノードからなる．ルートポインタが木の根となるノードへのポインタを持ち，根ノードおよび中間ノードなどの内部ノードが子となる内部ノードまたは葉ノードへのポインタを持つ．Bz 木は単一の CAS 命令ではなく，メモリ上の複数のワードを対象とする multi-word compare-and-swap（MwCAS）命令を用いて木のロックフリー化を実現している．

Bz 木の各ノードにはヘッダ，メタデータ，レコードが格納されている．ヘッダにはステータスワードというフィールドが格納されており，ステータスワードは CAS 命令や構造変更操作の制御，ノード内のレコード数などノード更新時に必要な情報を格納している．Bz 木はノードに対する書き込みおよび構造変更操作をステータスワードを更新することによってロックフリーに実現している．

3 B^c 木の構造

B^c 木は Bw 木と Bz 木の構造を参考に，各構造の長所を組み合わせた索引構造である．B^c 木の概形を図 1 に示す．木の論理的な構造は B^{link} 木に則っており，各ノードが同じ階層の右兄弟への参照リンクを持つ．各ノードへの参照はマッピングテーブルを用いた間接参照を採用し，マッピングテーブル内の物理ポインタを差し替えることでそのノードへの参照を一括で変更する．

B^+ 木と同様に，B^c 木は索引層およびデータ層によって構成される．索引層のノード（中間ノード）は分割キーと子ノードへのポインタの組を格納し，木の下方への検索を補助する．

各ノードの領域はイミュータブル領域とミュータブル領域に分けられる．イミュータブル領域はノードヘッダおよびソート済みのレコードを格納する．ヘッダはイミュータブル領域の情報を管理し，構造変更時のみその値が変更される．ミュータブル領域はステータスワードの格納と差分レコードを挿入するための書き込みバッファの役割を果たす．ステータスワードはミュータブル領域の情報を管理し，ノードの現在の状態や残容量などを管理する．各ノードへ構造変更操作を行う際は，構造変更後のノードから構造変更前のノードへ物理リンクを張り，古いノードへの参照を可能にする．なお，古いノードは構造変更が済み次第，ガベージコレクションへ追加され，他のスレッドから参照されないことが保証された時点で削除される．

B^c 木は Bw 木および Bz 木でそれぞれボトルネックとなる箇所を以下の通りに克服している．Bw 木では差分レコードによって更新を実現しているが，差分レコードは各スレッドによって独立にそのメモリ領域が確保されるため，同じノード上の差分レコードであってもメモリ上の配置は連続とならない．そのため，配置された差分レコードの数が多くなるほど，探索時にキャッシュミスが発生しやすくスケーラビリティが悪化する [15]．そこで Bz 木で用いられたノード内書き込みバッファを採用することで，Bw 木で発生していた差分レコードによるキャッシュミスを削減している．

一方で，Bz 木の差分レコードの挿入手続きにも改善の余地がある．Bz 木は差分レコードの書き込み時に，まず MwCAS 命令を用いたステータスワードの更新によって差分レコード用の領域を予約する．その後差分レコードを書き込み，メタデータを MwCAS 命令により更新することで差分レコードを可視化する．このとき，ステータスワードも MwCAS 命令の対象とすることで同時に発生した構造変更操作を検知し，必要に応じて書き込み命令をアボートおよびリトライする．しかしこのような

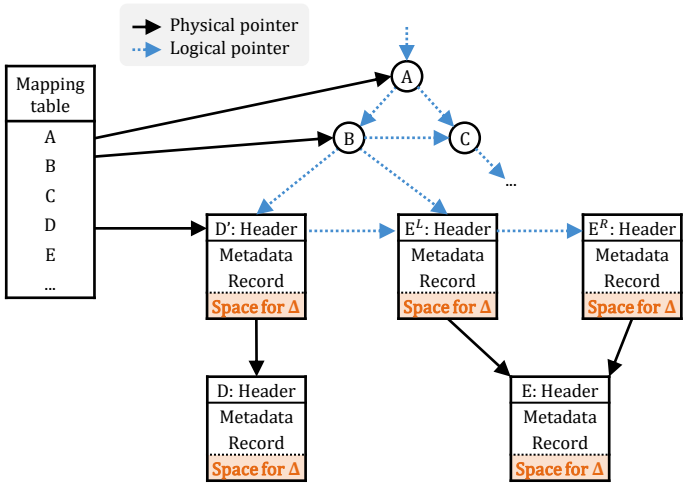


図 1　B^c 木の概形

手続きはステータスワードを不要に更新してしまい，スレッド間の競合を増加させてしまう．そこで，B^c 木では同時に発生した書き込み命令と構造変更操作に対し，書き込み命令を優先するという方針を取る．これにより，差分レコード可視化におけるステータスワードの確認が不要となり，書き込み命令におけるスレッド間の直接的な競合は領域予約のためのステータスワードの更新のみに抑えられる．

B^c 木の基本的な構造は B^+ 木に則った汎用的なものであり，特定のデータ型やハードウェアに合わせた最適化も可能である．例えば，キーをバイナリ比較可能な型（非負整数，文字列など）に限定することで，B^+ 木とトライ木を組み合わせた Mass 木 [12] のような最適化が可能である．格納する値を 8 バイトの非負整数で表せるもの（タプル ID，ポインタなど）に限定すれば，Bz 木 [11] や NB 木 [3] のように CAS による in place な値の更新も可能となる．なお，本研究では B^c 木の基礎的な構造に着目するため，そのような最適化は扱わない．

4　B^c 木のノード操作

本節では，B^c 木のノード操作について述べる．B^c 木は以下の書き込みおよび読み取り操作をサポートする．

4.1　書き込み

書き込み操作はキーとペイロードを差分レコード領域に挿入する操作である．キーが属する書き込み先の葉ノードを根ノードから二分探索により特定する．挿入先の葉ノードに到達後，その葉ノードの差分レコード領域に値を挿入する．書き込み操作には，無条件に書き込む Write，キーの不在を条件とする Insert，キーの存在を条件とする Update およびキーを削除する Delete がある．

書き込み操作は差分レコード領域の予約とレコード挿入および可視化の 2 ステップで行われる．図 2 に B^c 木における差分レコードの挿入を示す．ノードフッタにはミュータブル領域の状態を表すステータスワードを用意し，この中のレコード数と使用済みブロックサイズを加算することで差分レコード用の領域を予約する．また，B^c 木ではノードの生成時にミュータブル領域をゼロ埋めしている．B^c 木では，メタデータがゼロ埋めされている場合を処理途中として表すため，差分レコード用の領域を予約した時点ではレコードが可視化されていないことを認識できる．確保した領域へ差分レコードを書き込み，対応するレコードメタデータの値を更新することでレコードを可視化し，挿入処理を完了する．

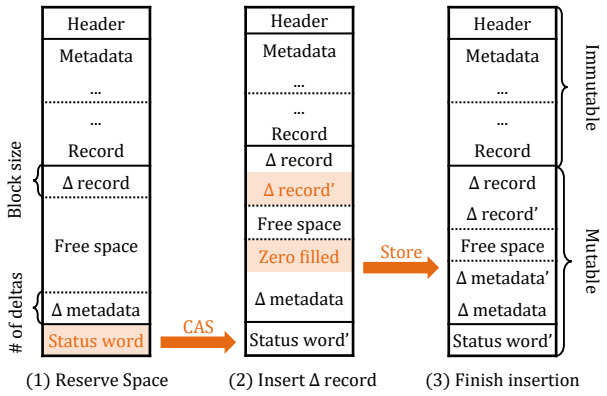


図 2　B^c 木における差分レコードの挿入

以上の書き込み操作によって差分レコードを挿入していくが，挿入後の差分レコードの総数またはノード容量のしきい値を越える場合，差分レコードの統合操作または分割操作といった構造変更操作が行われる．しきい値の確認はステータスワードの更新時に行われ，しきい値を越えた場合はレコードの書き込み後に，どの構造変更操作を行うか判定する．

書き込み処理はロックフリーに動作する．書き込み同士の競合はステータスワードで解決され，差分レコード用領域の予約，つまり差分レコードの書き込み順は CAS 命令の成功順によって順序付けられる．領域予約後のレコード挿入およびメタデータの書き込みは各スレッド並列に実行できる．ただし，後述する構造変更操作との競合による待ち時間は発生しうる．

4.2 読み取り

読み取り操作は与えられた対象キーのペイロードを返す操作である．対象キーが与えられたとき，根ノードからの二分探索によりキーが属する葉ノードに到達する．読み取りには，点読み取りである Read と範囲読み取りである Scan がある．

葉ノードに到達した後は，葉ノード内の差分レコードの線形探索とイミュータブルレコードの二分探索の 2 ステップで読み取り操作を行う．最新の値は差分レコード領域に書き込まれるため，まずミュータブル領域を線形探索し，一致するキーがあるか確認する．このとき，差分レコードの数はステータスワードから読み取り，差分レコードが可視化されていなければスキップする．

差分レコード中に対象のレコードが存在しなければ，ノードヘッダからイミュータブルレコードの数を読み取り，二分探索によって対象キーの有無を確認する．先頭ノードが統合操作の途中である場合には，差分レコードを読み終わった時点で物理ポインタをたどり古いノードへ移動し，同様の 2 ステップを古いノード上で行う．

読み取り処理は wait-free に動作する．上述したとおり読み取り命令は一切ノードの状態を変更せず，読み取った状態に応じて適切な手続きを選択する．そのため，読み取り命令においてリトライなどは発生せず，有限時間内で必ず処理が終了する．

範囲読み取りである Scan を行う際は，葉ノードに到達後，検索対象のノードと始点キーおよび終点キーを持つイテレータを返す．イテレータは差分レコード用のキーとメタデータの領域を用意し，対象ノードのミュータブル領域から始点キーと終点キーに含まれるキーとそのメタデータを用意した領域に挿入ソートにより格納していく．領域のソート後，対象ノードのイミュータブル領域と用意した領域をイテレータで走査する．イテレータはイミュータブル領域，用意した領域それぞれの先頭データを比較し，最新のレコードを取得していく．走査後，終点キーに達していなければ対象ノードの兄弟リンクをたどり，兄弟ノードと始点キーおよび終点キーを持つ新たなイテレータを返す．差分レコード領域とイミュータブル領域の比較によるレコードの取得は，読み取りを行う前に対象ノードを統合しイミュータブル領域のみを読み取る手法と比べて高い性能を示すことが明らかになっている [16].

5 B^c 木の構造変更操作

B^c 木は構造変更操作として差分レコードのノードへの統合，ノードの分割操作およびノード同士のマージ操作を持つ．構造変更操作を始める前に，新たなノードをマッピングテーブルに挿入することで，他のスレッドの待ち時間を削減する．

5.1 統　　合

統合操作は差分レコードの数またはノードの容量がそれぞれのしきい値を越えた場合，差分レコードとイミュータブルレコードをソートしてイミュータブル領域に反映する操作である．統合操作を行うことで，ミュータブル領域を確保し新たな差分

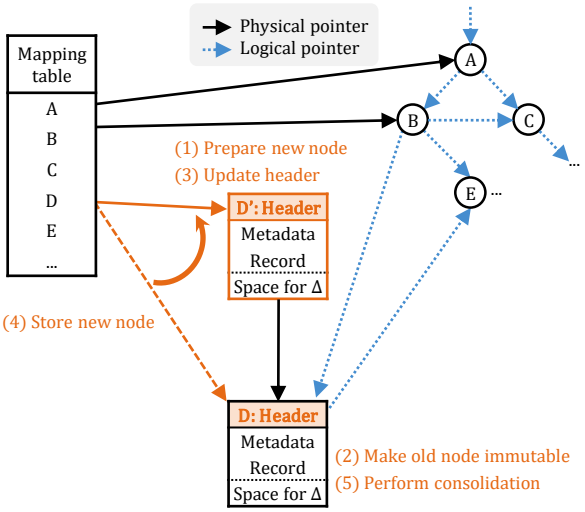


図 3　B^c 木における統合操作

レコードの挿入を受け付ける．

図 3 に B^c 木における統合操作を示す．統合が必要になった場合，新たなノード領域を用意する．そして，古いノードのステータスワードを更新後，ノードヘッダの情報から統合後に必要なイミュータブル領域を計算する．なお，このステータスワードの更新により古いノードは全体がイミュータブルとなり，新たな差分レコードは挿入できなくなる．次に，新たなノードにおいてステータスワードを含むヘッダ情報のみを更新した後，新規ノードとしてマッピングテーブル上の参照を更新する．その後，古いノード上の差分レコードをイミュータブルレコードへ反映させつつ，新たなノードへレコードをコピーする．最後に，統合後の状態でノードヘッダを更新し，新しいノードにおけるイミュータブル領域を可視化する．この際に，後述する分割操作で用いられる分割キーを，統合後のノードのイミュータブル領域から計算する．

統合操作を行っている際に，発生した書き込み操作は新規ノードのミュータブル領域で受け付ける．そのため，統合操作の間も新規レコードのミュータブル領域，統合前のノードのミュータブル領域およびイミュータブル領域の順でレコードを確認することで読み取り操作を実行できる．

5.2 分　　割

分割操作はノードの差分レコードの数またはノード容量がそれぞれしきい値を越えており，統合後も十分なミュータブル領域を確保できない場合に，そのノードのレコードを新規の 2 つのノードへ分ける操作である．分割操作により，容量が一杯になったノードを 2 つに分け，新たな差分レコードの挿入を受け付ける．

図 4 に B^c 木における分割操作を示す．構造変更操作が必要になった場合，新たなノード領域を 2 つ用意する．ステータスワードの更新により対象ノード全体をイミュータブル状態にした後，レコードコピーを始める前に用意したノードをマッピングテーブルに挿入する．これを実現するために，各ノードは自身の分割キーの位置を構造変更操作のたびに記録する．分割し

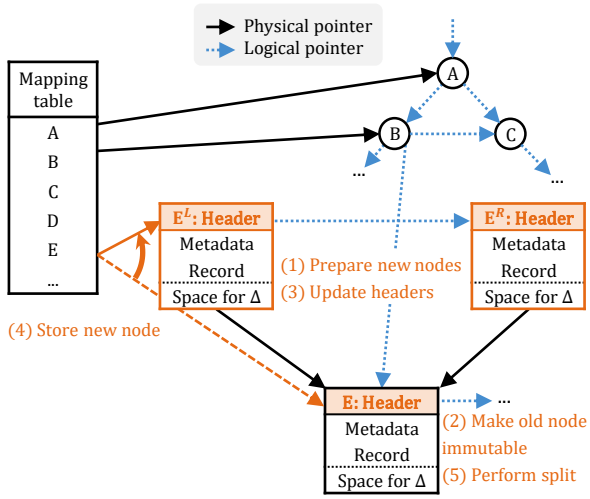


図 4　B^c 木における分割操作

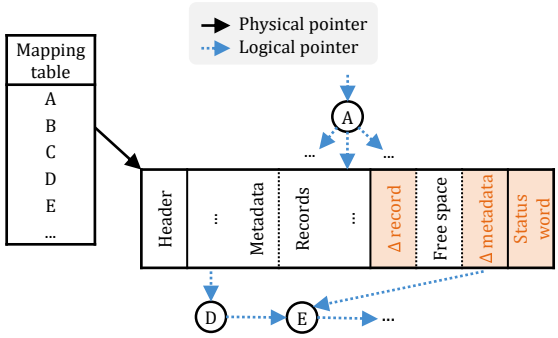


図 5　B^c 木における新規ノード E の挿入に伴う変更箇所

たノードの挿入後，統合操作と同様の手続きでレコードをコピーする．分割時には，古いノードの分割キー未満の全てのレコードを左ノードにコピーする．残りのレコード，すなわち分割キー以上の全てのレコードは右ノードにコピーする．

レコードのコピー後は，親ノードに右分割ノードへのリンクを挿入することで分割操作が完了する．このとき，親ノードへはリンクの情報のみを挿入し，反映は親ノードの構造変更時に行う．例えば新規ノード E を挿入した際，図 5 に示すように差分レコード領域に子ノードへのリンク情報のみを追記し統合操作が行われるまでそのリンクの反映を遅延する．つまり，索引の探索中において中間ノードでは差分レコードを確認せず，イミュータブル領域のレコードのみを検索する．

これは中間ノードにおける CPU キャッシュ効率を高めることを狙っている．親ノードのデータが CPU キャッシュ上に存在しており，ノードヘッダが同一のキャッシュラインに収まっていると想定する．この時，左の子ノードの兄弟リンクを経由して右の子ノードに達するために必要な同期はノードヘッダおよびキー範囲の確認の 2 回のみである．一方，右の子ノードへのリンクを挿入しかつそれを利用する場合，ノードヘッダ，メタデータおよびレコードの最低 3 回の同期が必要となる．親ノードから子ノードへのリンクを有効的に利用できるのはその子ノードを対象とした操作のみであることを考えると，親ノードに右の子ノードへのリンクを即座に反映した手法ではキャッシュ効率が悪化してしまう．そこで親ノードへはリンクの情報のみを挿入し，親ノードの統合操作時にリンクを反映させることによって，中間ノードにおける CPU キャッシュの利用効率を高める．更に，中間ノードのデータ量は葉ノードに比べて小さく変更も少ないため，CPU キャッシュは特に索引層において効率的に使用される．よって，中間ノードでは基本的にイミュータブル領域のみを参照させることで読み書き両方において葉ノード探索の効率を向上できる．

5.2.1　分割時の偏りの抑制

分割キーはイミュータブルレコードから決定される．つまり差分レコードは考慮されないため，分割操作時に片方のノードにレコードの偏りが発生しうる．前述した通り，分割操作前にノードを用意するため，イミュータブル領域および対応するメタデータ数は余裕を持って確保しておく必要がある．そのため，差分レコードが片方のノードに集中し，その分のスペースがもう片方のノードで使用されない，未使用なイミュータブル領域が残る可能性がある．

このような問題に対する対応策として，分割後の左右両ノードが一定以上の容量を持つようにすることが挙げられる．分割後の左右両ノードが一定以上の容量を持つ場合には分割操作を行い，持たない場合には統合操作を行うようにする．統合操作によって分割キーの位置が変化するため，書き込みの偏りが起きた場合においても，差分レコードが片方のノードに集中する事態をある程度防ぐことができる．このような対応により，未使用になるイミュータブル領域を削減し，メモリを効率的に利用できる．

5.2.2　Hot ノードの検出および分割

B^c 木では，構造変更操作中のノードに対する書き込み操作が偏っている場合，次の構造変更操作が要求され構造変更操作同士が競合する可能性がある．その場合の対応策について述べる．

B^c 木では統合および分割による新規ノードの配置後，新規のノードは古いノードからのレコードコピーをしつつ，差分レコードの挿入を受け付ける．その際書き込みの偏りが大きい場合，イミュータブル領域を可視化する前，つまり自身の構造変更操作が完了する前に次の構造変更操作が要求されてしまう．索引においてレコードアクセスには偏りが発生しやすく，書き込みによって偏りが発生すると構造変更操作の競合などにより性能が劇的に低下する．B^c 木ではそのような書き込みの集中したノードを Hot ノードと呼ぶ．

書き込みの偏りによる構造変更操作の競合に対して，先に行われた構造変更操作の完了後，分割操作を行う．これにより，アクセスが偏っているレコード群を細かく分割し競合を緩和させ，かつ差分レコードを書き込むための領域を広く確保できるようになる．このようなノードの積極的な分割によりレコードアクセスが偏るようなワークロードに対しても安定した性能を達成できる．

5.3　マ　ー　ジ

マージ操作は統合後のノードの容量がしきい値より低く，最

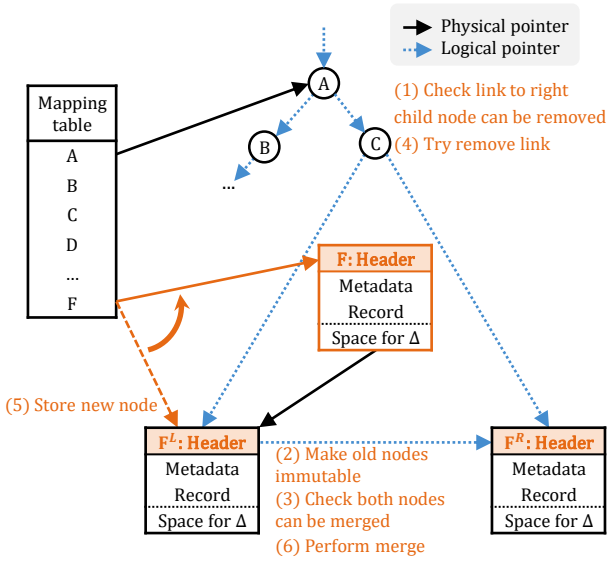


図 6　B^c 木におけるマージ操作

右端ノードでない場合，自身と右兄弟のノードのレコードを統合する操作である．マージ操作により，レコード数が少ないノードを削減し，メモリ使用率を効率化する．

図 6 に B^c 木におけるマージ操作を示す．統合後の新たなノードがマージ可能となった場合を想定する．マージ操作では，対象ノードをイミュータブル化する前に，親ノードの右兄弟ノードへのリンクの削除が可能か確認する．リンクの削除が可能である場合，対象ノードと右兄弟ノードをイミュータブル化し，マージが可能か確認する．ノード容量や構造変更操作が行われていないかどうかを確認し，マージが可能であればリンクの削除を試みる．リンクの削除は差分レコード領域への挿入によって行われるため，即座には反映されない．リンクの削除が失敗した場合には，対象ノードと右兄弟ノードのイミュータブル化を解除し，操作をアボートする．

マージが可能な場合，統合操作と同様の手続きで処理を進める．マージ後のノードを用意し，マッピングテーブルに挿入する．このとき，マージ後のノードはイミュータブル領域の計算などに加えて古いノードの起点（左ノード）を物理ポインタで指すようにヘッダを更新する．その後，マージする 2 つのノードのレコードをソートしながら新たなノードにコピーする．最後にマージ後のノードのステータスワードを更新し，操作を完了する．

6　性能評価

本節では，先述した B^c 木の性能を他の索引と比較して評価する．

6.1　実験設定

本研究では 1 台のサーバを用いて複数のワークロードで実験する．本研究で用いるサーバの構成を表 1 に示す．また，本実験では YCSB（Yahoo! Cloud Serving Benchmark） [17] のワークロードを模したものを使用し，一括挿入によって索引を構築した後に読み書きする．ワークロードは読み取り操作と書き込み操作が混在したワークロードであり，書き込み操作の割合は 0%，5%，10%，20%，...，100%である．読み取り操作は全て単一読み取りの Read 操作，書き込み操作は全て Update 操作である．実験結果は，YCSB ワークロードの A，B，C でのスループットを載せ，その後 112 スレッドにおける書き込み割合毎のスループットを載せる．A，B，C それぞれのワークロードでの書き込み割合は 50%，5%，0%である．

表 1　実験用サーバの構成

Item	Value
CPU	Intel(R) Xeon(R) Gold 6258R (two sockets)
RAM	DIMM DDR4 (Registered) 2933 MHz (16GB × 12)
OS	Ubuntu 20.04.6 LTS
Compiler	GNU C++ 9.4.0

また，Scan 操作の性能を測定するために，YCSB-A での読み取り操作を全て Scan 操作としたワークロードを使用する．つまりこのワークロードは一括挿入によって索引を構築した後，Update 操作と Scan 操作をそれぞれ 50%ずつの割合で同時に発行する．Scan 操作はランダムに選択した開始キーから選択した個数のペイロードを取得する．取得するペイロードの個数は 10，10^2，10^3，10^4 とする．

本実験で構築される索引は 10^8 個のレコードによって構成される．キーとペイロードは 8 byte の符号無し整数を使用し，レコードの大きさは合計 16 byte である．索引の性能は 1 秒あたりのレコード読み書き数（スループット [Rec/s]）を用いて評価する．実験では各スレッドが最大 10^7 回索引を操作し，総実行時間を総実行回数で割りスループットを計算する．なお，実験は各ワークロード各設定で複数回実施した．以下ではその平均スループットを示すが，測定結果は十分に安定しているためエラーバーは省略する．また，同時実行制御手法を効果的に評価するため，先行研究 [18, 19] に従い Zipf 分布を用いてキーの偏り（歪度，skew）を調整した．実験では，レコードアクセスが一様な場合（skew=0.0）とレコードアクセスに大きな偏り（skew=1.0）がある場合の環境を用いた．

評価実験の比較対象として，楽観的 B^+ 木（OSL）と Bz 木および Bw 木を用いる．楽観的 B^+ 木は楽観的制御によるバージョン管理を用いており，読み取り時のロック取得を不要としている．B^+ 木および Bw 木は，固定長のキーに対して最適化されたものを使用する．また，state-of-the-art な索引として Yakushima，OptiQL，ART を用いる．Yakushima は，劔（Tsurugi）DB [20] で用いられている，リトライロジックなどのいくつかの改良が加えられた Mass 木の実装である．

6.1.1　YCSB

図 7 に YCSB ワークロードにおいてキーの偏りが無い状態でスレッド数を変化させた際のスループットを示す．各結果から，全ての索引が線形増加していることがわかる．B^c 木および Bw 木はノード探索にマッピングテーブルを用いているため，直接参照する B^+ 木や探索時にキーの二分探索をしない ART と比

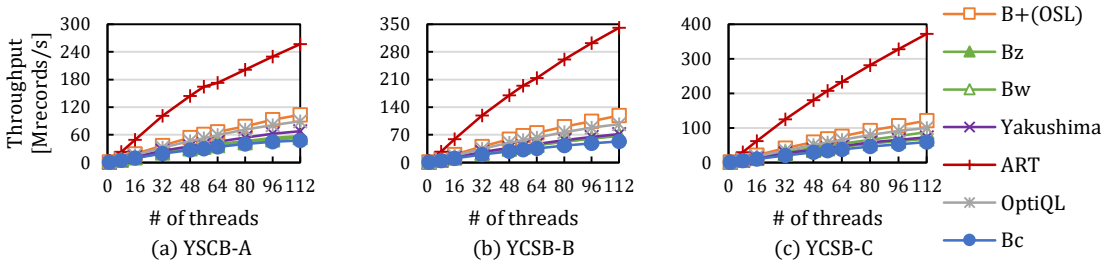


図 7　YCSB ワークロードにおけるスレッド数毎のスループットの変化（skew=0.0）

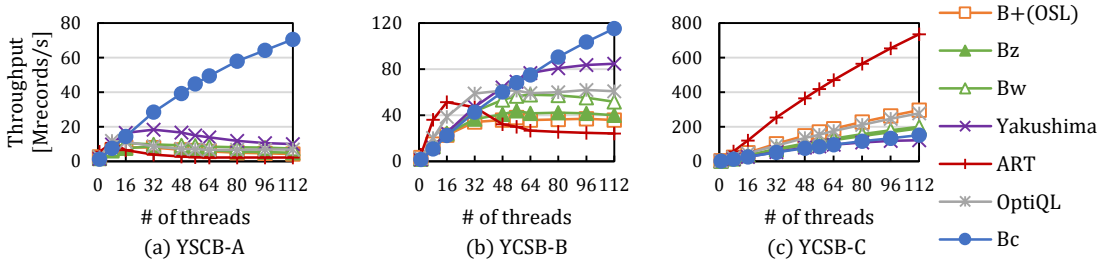


図 8　YCSB ワークロードにおけるスレッド数毎のスループットの変化（skew=1.0）

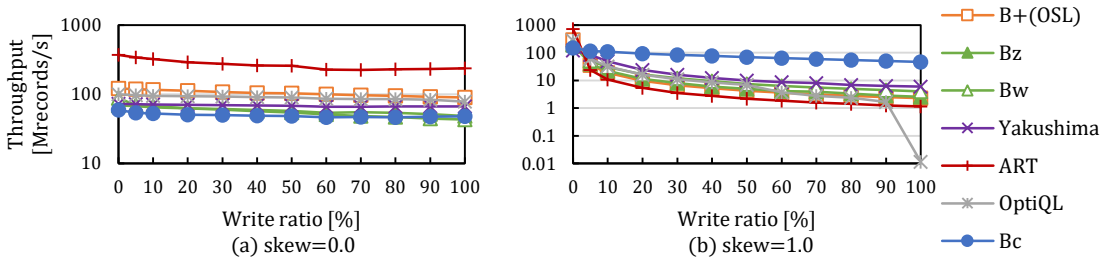


図 9　書き込み割合毎のスループットの変化（threads=112）

べて性能が劣る．

図 8 に YCSB ワークロードにおいてキーの偏りがある状態でスレッド数を変化させた際のスループットを示す．図 8(a)(b) のような書き込みのあるワークロードにおいて B^c 木は他の索引よりも高い性能を示した．他索引は書き込み時のロックや競合が発生し，性能が下がっている．一方で，B^c 木は書き込みの競合を CAS による差分レコード領域の予約に限定していること，構造変更中にもノード操作を受け付けていること，そして Hot ノードに対する積極的な分割によりアクセスを分散させたことなどにより他の索引対して高い性能を示したと考えられる．

図 9 にスレッド数が 112 において書き込み割合を変化させた際のスループットを示す．なお，縦軸は対称軸である．キーに偏りが無い場合は一定して，ART が高い性能を示した．B^c 木は他の索引と同程度の性能を示した．競合が発生しない場合には葉ノード探索が性能に影響し，B^c 木はマッピングテーブルを用いた間接参照を行うことにより性能が低くなったと考えられる．キーに偏りがある場合は，書き込み割合が 0 の場合を除いて B^c 木が他の索引よりも高い性能を示した．他の索引は書き込みに偏りがある場合には競合発生により性能が低下する．B^c 木は書き込みの割合によらず安定した性能を維持できると考えられる．

6.1.2　Update-Scan Mixed

図 10 にスレッド数が 112 においてスキャン範囲を変化させた際のスループットを示す．縦軸，横軸は共に対称軸である．キーに偏りが無い場合には，最適化された B^+ 木と Bw 木，そして OptiQL の性能が高い．また，スキャン範囲によらず B^+ 木が最も高い性能を示した．キーの偏りがある場合には，スキャン範囲が 1000 レコードまでの場合は B^c 木は高い性能を示したが，10000 レコードになると OptiQL の性能が最も高くなった．B^c 木はスキャン時にイミュータブル領域とミュータブル領域を比較しながらレコードを読み出すようにしている．そのため，読み出すレコードの数が多くなるほど比較操作も増えるため OptiQL や B^+ 木などに性能が劣ったと考えられる．

7　おわりに

本稿では B^+ 木の同時実行制御においてロックフリーアルゴリズムを活用した，新たな索引構造である B^c 木について提案し，その構造および操作を紹介した．また，代表的な索引構造である B^+ 木や，近年提案されている索引，Bw 木や Bz 木といったロックフリー索引との性能を比較検証した．実験結果より，B^c 木は様々なワークロードに対して安定した性能を発揮できることを確認した．特にキーの偏りがある書き込みワークロードでは他の索引よりも高い性能を示した．Scan において

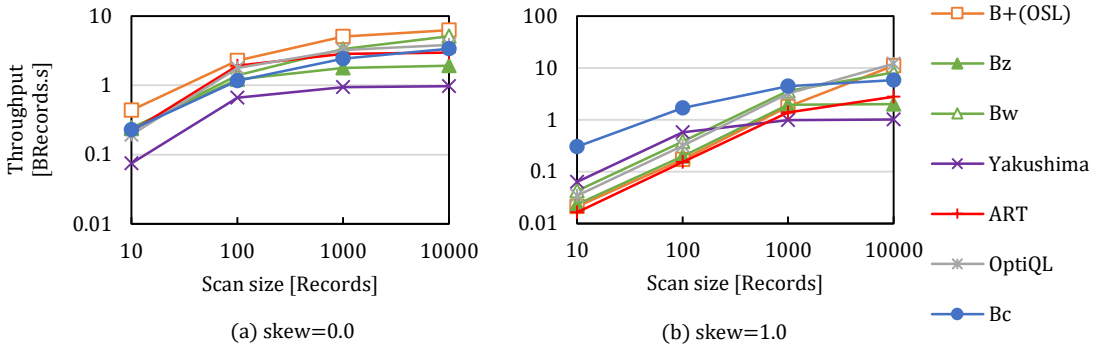


図 10　Update-Scan Mixed におけるスキャン範囲毎のスループットの変化（threads=112）

もキーの書き込みを伴わない場合は他の索引に性能が劣るものの，キーの書き込みを伴い，スキャン範囲が広すぎない場合には他の索引よりも高い性能を示した．今後の課題として，様々なキーの大きさや種類を用いた詳細な性能比較，B^c 木の分散索引への拡張やセカンダリ化などが挙げられる．

謝　　辞

本研究の一部の一部は JSPS 科研費 JP20K19804，JP21H03555，JP22H03594 の助成，日本電信電話株式会社との共同研究，日本電信電話株式会社との共同研究，および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

文　　献

[1] H. Cha, X. Hao, T. Wang, H. Zhang, A. Akella, and X. Yu, "B^{link}-hash: An adaptive hybrid index for in-memory time-series databases," *Proc. VLDB*, vol. 16, no. 6, pp. 1235–1248, 2023.
[2] A. Mathew and C. Min, "Hydralist: a scalable in-memory index using asynchronous updates and partial replication," *Proc. VLDB*, vol. 13, no. 9, pp. 1332–1345, 2020.
[3] B. Zhang, S. Zheng, Z. Qi, and L. Huang, "Nbtree: a lock-free pm-friendly persistent b+-tree for eadr-enabled pm systems," *Proc. VLDB*, vol. 15, no. 6, pp. 1187–1200, 2022.
[4] Y. Chen, Y. Lu, K. Fang, Q. Wang, and J. Shu, "utree: a persistent b+-tree with low tail latency," *Proc. VLDB*, vol. 13, no. 12, pp. 2634–2648, 2020.
[5] G. Shi, Z. Yan, and T. Wang, "Optiql: Robust optimistic locking for memory-optimized indexes," *Proc. ACM Manag. Data*, vol. 1, no. 3, 2023.
[6] 北川 博之, **データベースシステム**. 昭晃堂, 1996.
[7] M. Herlihy, N. Shavit, V. Luchangco, and M. Spear, *The Art of Multiprocessor Programming, Second Edition*. Elsevier, 2020.
[8] P. L. Lehman and S. B. Yao, "Efficient locking for concurrent operations on B-trees," *ACM TODS*, vol. 6, no. 4, pp. 650–670, 1981.
[9] S. K. Cha, S. Hwang, K. Kim, and K. Kwon, "Cache-conscious concurrency control of main-memory indexes on shared-memory multiprocessor systems," in *Proc. VLDB*, pp. 181–190, 2001.
[10] J. Levandoski, D. Lomet, and S. Sengupta, "The Bw-tree: A B-tree for new hardware platforms," in *Proc. ICDE*, pp. 302–313, 2013.
[11] J. Arulraj, J. Levandoski, U. F. Minhas, and P. Larson, "BzTree: A high-performance latch-free range index for non-volatile memory," *PVLDB*, vol. 11, no. 5, pp. 553–565, 2018.
[12] R. T. M. Yandong Mao, Eddie Kohler, "Cache craftiness for fast multicore key-value storage," in *Proc. EuroSys '12*, pp. 183–196, 2012.
[13] V. Leis, A. Kemper, and T. Neumann, "The adaptive radix tree: ARTful indexing for main-memory databases," in *Proc. ICDE*, pp. 38–49, 2013.
[14] A. Petrov, **詳説データベース: ストレージエンジンと分散データシステムの仕組み**. オライリー・ジャパン, 2021.
[15] Z. Wang, A. Pavlo, H. Lim, V. Leis, H. Zhang, M. Kaminsky, and D. G. Andersen, "Building a Bw-tree takes more than just buzz words," in *Proc. SIGMOD*, pp. 473–488, 2018.
[16] 佑. 井戸, 健. 杉浦, 宗. 中山, 佳. 石川, and 可. 陸, "Bz 木における範囲走査性能の改善," in **第 *85* 回全国大会講演論文集**, vol. 2023, pp. 451–452, 2023.
[17] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with YCSB," in *Proc. SoCC*, pp. 143–154, ACM, 2010.
[18] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with YCSB," in *Proc. SoCC*, pp. 143–154, 2010.
[19] Y. Yang and J. Zhu, "Write skew and Zipf distribution: Evidence and implications," *ACM TOS*, vol. 12, no. 4, 2016.
[20] T. Kambayashi, T. Tanabe, T. Hoshino, and H. Kawashima, "Shirakami: A hybrid concurrency control protocol for Tsurugi relational database system," 2023.

同時実行 B+ 木のマルチバージョン化の検討

桑村　真生† 　杉浦　健人† 　平野　匠真† 　石川　佳治† 　陸　可鏡†

† 名古屋大学 情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
Email: {kuwamura, hirano, lu}@db.is.i.nagoya-u.ac.jp, {sugiura, ishikawa}@i.nagoya-u.ac.jp

あらまし　並行トランザクション処理をサポートするデータベース管理システムにおける不整合の 1 つにファントムリードがある．ファントムリードとは，範囲走査が並行する書き込みで生じた本来読むべきでないレコードを読んでしまうというものである．多版同時実行制御によるスナップショット分離はファントムリードを防止するための手法の 1 つである．索引層でスナップショット分離をサポートした先行研究には，範囲走査の性能や空間効率の面で改善点が残っていると考えられる．そこで，本研究では差分レコードに基づくロックフリー索引である同時実行 B+ 木を導入し，そのマルチバージョン化の方針について述べる．

キーワード　データ構造・索引，多版同時実行制御．

1 はじめに

複数のトランザクションを並行して実行できるデータベース管理システムは，一貫性のある読み書きを保証しなければならない．一貫性に関する典型的な問題の 1 つは，範囲走査が並行する問合せによって新たに挿入されたレコードを読んでしまい，読み取りの一貫性が保たれなくなることである．これをファントムリード [1] という．

ファントムリードの防止のためには範囲走査の再実行 [2] やファントム検出用のデータ構造の追加 [3] などの方法が提案されている．レコードの古い値を保持し，その読み取りを許容する多版同時実行制御（Multiversion Concurrency Control，MVCC）によるスナップショット分離 [4] もまた，ファントムリードを防止する手法の 1 つである．スナップショット分離では書き込みと読み取りの競合が起こらないため，範囲走査と書き込みが同時に起こるようなワークロードにおいて，その他の手法よりも低いアボート率と高い並行性を両立できる．

MVCC ではレコードのバージョン情報を索引と分離して保持することもあるが，本研究では索引層でバージョン情報を含めてレコードを保持した場合の読み書きのマルチバージョン化について検討する．すなわち，書き込み時には索引内での管理のためにバージョン情報を生成し，読み取り時には適切なバージョンを返すような索引を考える．しかし，単純に B+ 木 [5] をマルチバージョン化した場合，複数のバージョンから成るレコードの走査によりキャッシュヒット率が悪化してしまう．そこで，本研究では差分レコードに基づくロックフリー索引である同時実行 B+ 木（Concurrent B+-Tree，B^c 木）を導入し、そのマルチバージョン化について述べる．

2 関連研究

本章では，B+ 木の単純なマルチバージョン化による性能低下と，スナップショット分離を保証する索引の先行研究について述べる．

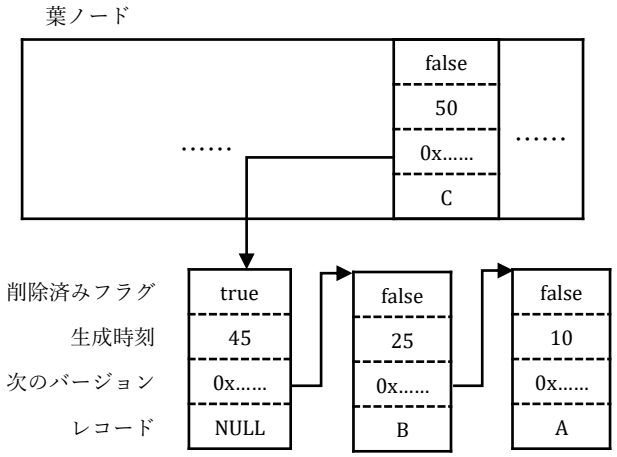


図 1　マルチバージョン B+ 木のバージョンチェイン

2.1 B+ 木のマルチバージョン化

先行研究において著者らはマルチバージョン機能を追加した B+ 木における範囲走査の性能を測定した [6]．すなわち，B+ 木の葉ノードにレコードの最新の値と古い値へのポインタを格納して単方向リスト（バージョンチェイン）を構成することで，レコードのマルチバージョン化を行った．図 1 はバージョンチェインの概観である．古いバージョンは B+ 木のノードとは独立にランダムなアドレスで保持される．レコードの各バージョンには値とともに生成時刻も保持されており，読み取り時にはこの時刻をもとに読むべきバージョンを探す．

しかし，このような単純なマルチバージョン化では，レコードの各バージョンへアクセスするたびにランダムアクセスを必要とする．一定の範囲のアドレスのデータをまとめてコピーするキャッシュメモリの特性から，ランダムアクセスは順次アクセスと比較してキャッシュヒット率が低下する．キャッシュヒット率は現代的なハードウェアのメニーコア環境におけるスケーラビリティに大きく影響し，B+ 木においては本来順次アクセスとなる範囲走査で特に性能低下が著しい．

2.2 Multi-Version Partitioned B-Tree

Multi-Version Partitioned B-Tree（MV-PBT） [7] は，バッファのディスクへのフラッシュをカウントした Partition Number を用いて，B^+ 木を複数のパーティションに水平分割した索引である．MV-PBT への書き込みは全て，バッファ上にある最新のパーティションへのレコードの挿入で表現される．例えば，あるレコードの非キー属性を更新する場合，新しいバージョン情報と新しい値を持った．非キー属性の更新を表す専用のレコードを挿入する．読み取り時は，Partition Number やバージョン情報に基づいて検索条件に合致するレコードを新しいものから順に読み，可視なバージョンを見つけるとその値を返す．最適化のため，ブルームフィルタなどを用いて読み取りたいレコードが存在する可能性のあるパーティションのみを確認する．

MV-PBT は，索引内でバージョン情報を保持してレコードの可視性チェックを索引内で完結させ，ディスクベースのデータベースにおける I/O の回数を減らすことを目的としている．一方，近年のハードウェア技術の進歩により，全てのデータをメモリ上で管理してログのみを永続化するインメモリデータベースが現実的な選択肢となってきている．よって，より高い性能を達成するため，索引についてもインメモリを前提とした設計が必要だと考えられる．

2.3 Parallel Binary Tree

Parallel Binary Tree（P 木） [8] は平衡二分探索木の各ノードにレコード，部分木のサイズおよび参照カウンタを保持したインメモリ索引である．レコードを更新する際にはルートノードから更新したいレコードを持った更新対象ノードまでのパス上のノードを全てコピーし，コピーした対象ノードのレコードを更新する．これをパスコピーという．また，コピーしないノードについてはそのまま流用する（コピーオンライト方式）．つまりレコードの更新時には，もとの木と一部のノードを共有し，一部のノードのレコードが更新された 2 つ目の木を作ることになる．P 木に対する読み取り専用トランザクションは，開始時点で最新の木を対象に行われるため，トランザクションの開始時刻に存在したレコードのみが操作の対象となる．このため，もし並行する書き込みが新しい木を生成しても，読み取りや範囲走査はその影響を受けない．

P 木では，パスコピーによってバージョンチェインに起因するランダムアクセスの頻発が克服されている．しかし，B^+ 木のような兄弟リンクを利用した高速な範囲走査の実現は二分探索木をベースとする P 木では難しい．また，二分探索木は空間効率において B^+ 木に劣ることが知られており，P 木にも同様の問題があると考えられる．

3 B^c 木

本章では，差分レコードに基づくロックフリーなインメモリ索引である B^c 木の設計や書き込み操作，読み取り操作および構造変更操作の手続きについて述べる．なお，ここではマルチバージョン化に関わる部分のみ議論し，詳細な設計の説明は省く．

3.1 基本構造

B^c 木の概観を図 2 に示す．B^c 木の構造は，B^{link} 木 [9] と Bw 木 [10] を基にしている．すなわち，内部ノードを含めた各ノードは兄弟ノードへのリンクを持ち，各リンクの実体はマッピングテーブルを経由した間接参照となっている．マッピングテーブルのエントリを compare-and-swap（CAS）命令で変更することで，特定ノードに対する参照を一括で差し替えられる．ただし，B^c 木は構造変更前のノードを一時的に保持していることがあり，構造変更後のノードから構造変更前のノードへの参照に限ってはマッピングテーブルを経由せず直接行われる．

3.2 ノードレイアウト

B^c 木のノードレイアウトは Bz 木 [11] を基にしている．各ノードはヘッダやソートされたレコードおよびそのメタデータからなる不変領域と，差分レコード用のフリースペースとステータスワードからなる可変領域に分かれている．書き込まれた値は差分レコードとして可変領域に格納され，ステータスワードでは現在のノードの状態や差分レコード書き込み用のオフセットなど，可変領域の情報を管理している．

同様に，不変領域の情報はヘッダによって管理されている．不変領域中のレコードは構造変更時にのみ更新が行われ，構造変更前のノードが持つ差分レコードの内容を統合したものが構造変更後のソート済みレコードとして不変領域に保持される．

3.3 書き込み

B^c 木の書き込みでは，まず書き込みの対象となる葉ノードのステータスワードが持つオフセット値や差分レコードの数を CAS 命令で更新することにより，差分レコードを書き込む領域を予約する．その後，予約した領域へ差分レコード，対応するメタデータの順で書き込む．このように領域の予約とデータの書き込みを分けることで，書き込みの競合をステータスワードに対する CAS 命令の部分のみに抑えられる．なお，B^c 木では，メタデータが書き込まれるまで書き込み途中のレコードは他スレッドから見えないようになっている．

3.4 読み取り

前述のとおり B^c 木では最新の値が可変領域に書き込まれるため，読み取り操作ではまず対象となる葉ノードの差分レコードを線形探索する．読み取り対象のキーを持つ差分レコードが存在しない場合，続いて不変領域のレコード領域を二分探索する．

なお，B^c 木では操作の並列性を高めるために，読み書きの操作と後述する統合操作が同時に行われることを許容する．つまり，読み書き時に対象ノードの不変領域の変更が終わっていない場合があるため，その際は構造変更前の古いノードに遷移し操作を続ける．遷移後は同様の手順で可変領域，不変領域をそれぞれ走査する．

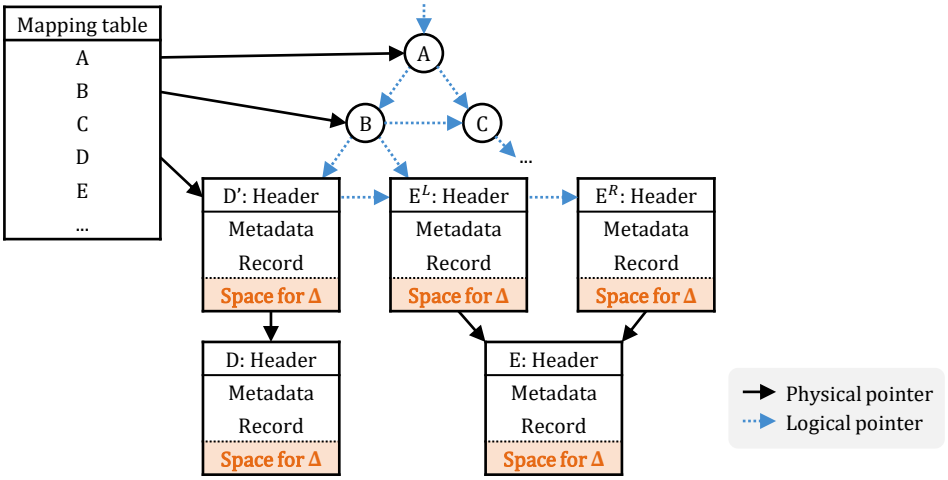


図 2　B^c 木の概観

3.5　構造変更

B^c 木における構造変更とは，差分レコード領域に蓄積した更新を新たに用意したノードの不変領域へ反映する操作である．B^c 木の構造変更操作は，通常の B^+ 木で用いられる分割や併合，Bw 木などにみられる統合の 3 種類が存在する．あるノードへの書き込みが一定量を超えると統合が試みられ，統合後に不変領域の占める容量が大きすぎる場合は分割，小さすぎる場合は併合される．

構造変更操作では，まず差分レコードを適用するための新しいノードを用意する．最初このノードは可変領域のみを持ち，構造変更と並行する書き込みを受け付ける．そして古いノードをロックし，書き込みを制限する．次に，マッピングテーブルを更新し新規ノードで書き込みを受け付ける状態にする．最後に，構造変更前のノードから新規ノードの不変領域へ，差分レコードを適用した新しい値を持つレコードをソートしながらコピーする．構造変更の終わった古いノードはガベージコレクションに追加され，他のスレッドから参照されないことが保証された時点で削除される．

4　B^c 木のマルチバージョン化

本章では，3 章で導入した B^c 木をマルチバージョン化する方針について述べる．ただし，シングルバージョン B^c 木との差分についてのみ言及する．

4.1　タイムスタンプの付与

マルチバージョン B^c 木においてはレコードのメタデータ，ステータスワードおよびノードヘッダ（不変領域）でそれぞれタイムスタンプを保持する．図 3 はタイムスタンプを追加したマルチバージョン B^c 木の概観である．タイムスタンプは一定時間ごとに専用のスレッドがインクリメントするカウンタを用いて表現される [2]．各タイムスタンプが持つ意味は以下のとおりである．

a）　レコードのメタデータ

レコードまたは差分レコードの生成時刻を表す．この値は，読み取り時の可視性チェックに用いられる．本稿ではある時刻 t に対して，生成時刻が t 未満なもののうち最新バージョンのレコードを可視とする．なお，図 3 においては省略されている．

b）　ステータスワード

当該ノードへの最新の書き込み時刻，すなわち最も新しい差分レコードの生成時刻を表す．図 3 では各葉ノードの下部に書かれている．

c）　ノードヘッダ

当該ノードに書き込まれる差分レコードの生成時刻の下限値を表す．図 3 では各葉ノードの上部に書かれている．これは構造変更時に決定され，その値は構造変更前のノードのステータスワードが持つタイムスタンプと一致する．読み取り時には，この下限値に基づいて読み取り対象となる葉ノードのバージョンを決定する．

4.2　ガベージコレクション

図 3 に示したように，マルチバージョン B^c 木では実質的に葉ノードがバージョンチェインを成している．よって，必要に応じて古いバージョンのレコードを持つノード，つまり構造変更の終わった古い葉ノードが参照される．そのため，実行中の全読み取り操作の開始時刻をスレッド間で共有し，それらにとって可視であるレコードを含むノードはガベージコレクションの対象外となる．

HTAP ワークロードのような，長い読み取り専用トランザクションと短い更新トランザクションが混じるワークロードでは，最新バージョンと読み取られるバージョンの間に多数の不要なバージョンが生じる．このようなケースでは，読み取りの効率化のために中間バージョンの削除が必要となる [12]．本研究のガベージコレクションには，読み取り時のバージョンチェイン走査に合わせて不要なバージョンを削除する手法 [13] を改変したものを用いる．すなわち，読み取り時に，最新ノードから自身が読み取るバージョンのノードまでの間にある不要なバージョンを削除する．また，それだけでは誰もアクセスしなくなった古いノードを削除できなくなってしまうため，定期的に全てのバージョンの葉ノードにアクセスして不要なバージョ

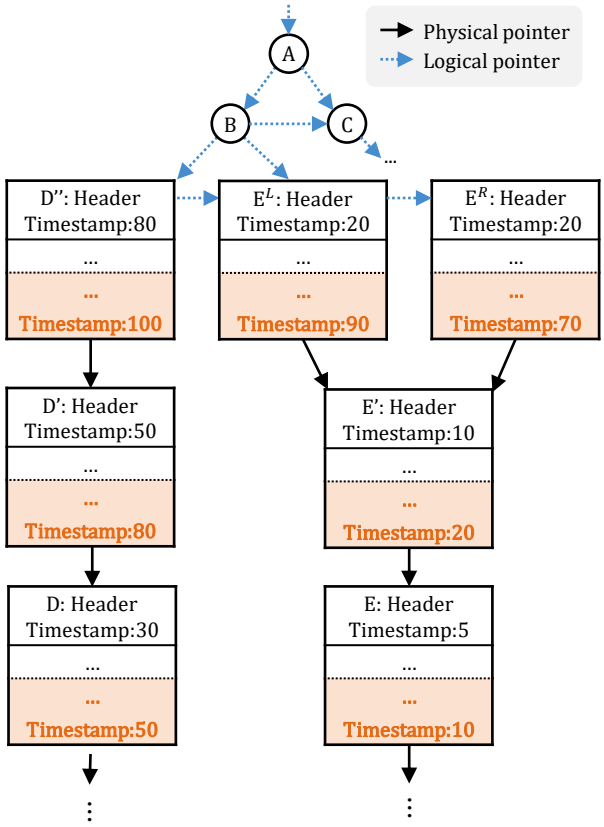


図 3　マルチバージョン B^c 木の概観

ンを GC する専用のスレッドを用意する．GC 時のバージョンチェイン走査によるランダムアクセスが性能に悪影響を与えるという指摘もある [14] が，マルチバージョン B^c 木においてはランダムアクセスはノード単位で生じ，レコードのバージョン単位では生じない．よって，ランダムアクセスによる悪影響は軽減されると考えられる．

4.3　可視なバージョンの読み取り

マルチバージョン B^c 木では，スナップショット分離の実現のために古いバージョンのレコードを読み取る．ある時刻 t において可視なバージョンのレコードを読み取るとき，まずは構造変更によってバージョンチェインを成した葉ノードの中から必要なバージョンのレコードを含むものを探す．次に，通常の読み取りと同じ要領で，当該ノードから時刻 t において可視であるレコードを読み取る．

時刻 t において可視なバージョンのレコードは，ヘッダのタイムスタンプが t 未満であるような最も新しいノード n に存在する．構造変更で n が作成された時点での全レコードの最新状態はその不変領域に反映され，n が作成されて以降の書き込みはその可変領域に差分レコードとして存在する．このため，n の可変領域に時刻 t において可視な読み取り対象のキーを持つ差分レコードが存在するならばそれを読み取ればよい．存在しない場合は，n の不変領域にあるレコードが時刻 t において可視となる．なお，葉ノードのバージョンチェインは先頭から新しい順に整列しているため，バージョンチェインを辿りながら一番最初に見つけたノードヘッダのタイムスタンプが t 未満のノードを読み取り対象とすればよい．

5　おわりに

本稿では，既存のマルチバージョン索引の課題点について述べた．また，それを克服するため差分レコードに基づくロックフリー索引の B^c 木を導入し，そのマルチバージョン化についての方針を述べた．

今後はマルチバージョン B^c 木のスループットやフットプリントについて，マルチバージョン B^+ 木やその他のマルチバージョン索引との比較実験を行う予定である．

謝　辞

本研究の一部は JSPS 科研費 JP20K19804，JP21H03555，JP22H03594 の助成，日本電信電話株式会社との共同研究，および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

文　献

[1] K. P. Eswaran, J. N. Gray, R. A. Lorie, and I. L. Traiger, “The notions of consistency and predicate locks in a database system,” *Commun. ACM*, vol. 19, no. 11, pp. 624–633, 1976.
[2] S. Tu, W. Zheng, E. Kohler, B. Liskov, and S. Madden, “Speedy transactions in multicore in-memory databases,” in *Proc. SOSP*, 2013, pp. 18–32.
[3] S. A. Mehdi, D. Hwang, S. Peter, and L. Alvisi, “ScaleDB: A scalable, asynchronous In-Memory database,” in *Proc. OSDI*, 2023, pp. 361–376.
[4] H. Berenson, P. Bernstein, J. Gray, J. Melton, E. O’Neil, and P. O’Neil, “A critique of ANSI SQL isolation levels,” *SIGMOD Rec.*, vol. 24, no. 2, pp. 1–10, 1995.
[5] 北川 博之，データベースシステム　昭晃堂，1996.
[6] 桑村真生，杉浦健人，野原健汰，石川佳治，陸可鏡，“B^+ 木のマルチバージョン化による範囲走査性能への影響評価”，第 85 回全国大会講演論文集，vol. 2023，no. 1，2023，pp. 449–450.
[7] C. Riegger, T. Vinçon, R. Gottstein, and I. Petrov, “MV-PBT: Multi-version indexing for large datasets and HTAP workloads,” in *Proc. EDBT*, 2020, pp. 217–228.
[8] Y. Sun, G. E. Blelloch, W. S. Lim, and A. Pavlo, “On supporting efficient snapshot isolation for hybrid workloads with Multi-Versioned indexes,” *PVLDB*, vol. 13, no. 2, pp. 211–225, 2019.
[9] P. L. Lehman and S. B. Yao, “Efficient locking for concurrent operations on B-Trees,” *ACM TODS*, vol. 6, no. 4, pp. 650–670, 1981.
[10] J. J. Levandoski, D. B. Lomet, and S. Sengupta, “The Bw-Tree: A B-tree for new hardware platforms,” in *Proc. ICDE*. IEEE, 2013, pp. 302–313.
[11] J. Arulraj, J. Levandoski, U. F. Minhas, and P.-A. Larson, “Bztree: a high-performance latch-free range index for non-volatile memory,” *PVLDB*, vol. 11, no. 5, pp. 553–565, 2018.
[12] J. Böttcher, V. Leis, T. Neumann, and A. Kemper, “Scalable garbage collection for in-memory MVCC systems,” *PVLDB*, vol. 13, no. 2, pp. 128–141, 2019.
[13] R. Onishi, T. Hoshino, and H. Kawashima, “EPO-R: An efficient garbage collection scheme for long-term transactions,” in *2022 Tenth International Symposium on Computing and Networking (CANDAR)*, 2022, pp. 144–150.
[14] A. Raza, P. Chrysogelos, A. C. Anadiotis, and A. Ailamaki, “One-shot Garbage Collection for In-memory OLTP through Temporality-aware Version Storage,” *PACMMOD*, vol. 1, no. 1, pp. 19:1–19:25, 2023.

リモートログ先行書き込み用のストレージ階層追加による効率化

XU WENKANG† 石井 幹大† Martin J. Dürst†

† 青山学院大学理工学部情報テクノロジー学科 〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: †wenymedia@gmail.com, ††{ishii,duerst}@it.aoyama.ac.jp

あらまし クラウドサービスプロバイダーは，オブジェクトストレージやブロックストレージなど，信頼性の高いストレージサービスを提供している．クラウドサービスプロバイダーは，単一障害点によるデータ損失を回避するために，ストレージサービスのデータを複数のサーバに複製して保存している．クラウドコンピューティングのストレージサービスによって，ユーザ自身が複数のサーバにデータを複製して保存する必要がなくなり，システムを迅速に構築できるようになった. 本研究では，クラウド化データベースにおける効率化されたリモートログ先行書き込みシステムを提案する．クラウドコンピューティングのブロックストレージサービスを利用して一連のバッチ最適化を含む書き込み遅延を最小限に抑え，オブジェクトストレージサービスを活用して読み込みのスループットを最大化する．

キーワード リモートログ先行書き込み，先行書き込み，クラウド基盤，ストレージ管理

1 はじめに

1.1 研究の背景と目的

2007 年以降，Amazon [1] やアリババ [2] といった大手のインターネット企業 [3] は，アプリケーションをクラウド化する動きを進めてきた．この動きは，クラウドコンピューティングと呼ばれ，従来のオンプレミスのデータセンターに比べ，2 つのメリットがある．1 つ目は，クラウドサービスは従量制課金が一般的であり，必要なリソースだけを使用するため，コストの最適化が可能な点である．2 つ目は，必要に応じてリソースを柔軟に拡張でき，需要の変動や急激なトラフィックの増加に対応可能な点である．

クラウドコンピューティングが進む中で，データベースも発展を続けている．クラウドコンピューティング以前の 1990 年代，企業業務への対応は，単純なモノリシックアーキテクチャのデータベースのみで可能であった．しかし，インターネットの普及に伴い，2000 年以降は，大量のデータを処理する必要が生じ，データを分割して処理できるシェアードナッシングアーキテクチャのデータベースが主流になってきた．そして，2010 年以降，Amazon S3 [4] などのオブジェクトストレージサービスが成熟するにつれて，データを分割して処理することができ，ストレージ容量がほぼ無制限の共有ストレージアーキテクチャのデータベースが多くの注目を集めている．2020 年には，クラウドベースのデータプラットフォームを提供する Snowflake 社 [1]が新規株式公開 (IPO) を行い [5]，時価総額がソフトウェア企業として史上最高額の 707 億ドルに達するなど，クラウドコンピューティングへの注目は前例のない高まりを見せている．

2018 年頃より，クラウドデータベースの市場において，DBaaS (Database as a Service) が急速に成長している．2022 年，米国の調査会社・ Gartner 社は，データベースの市場規模において，DBaaS の市場規模がオンプレミスに並んだと発表した [6]．DBaaS の急成長により，データベースはクラウドコンピューティングに向けて再設計が必要とする意見が生まれた．

本研究では，クラウド化データベースに向けた最適解化のリモートログ先行書き込み (Write-Ahead Logging，以下，リモート WAL) システムを提案する．クラウドコンピューティングのブロックストレージサービスを利用して書き込み遅延を最小限に抑え，オブジェクトストレージを活用して読み込みのスループットを最大化する．

1.2 論文の構成

本論文の構成は次の通りである． 2 章では，クラウドコンピューティングで頻繁に使われるストレージサービス，およびクラウドコンピューティングに関連するアーキテクチャを説明する．3 章では，関連研究として，先行研究と既存システムのクラウドにおける設計について説明する．4 章では，リモート WAL システムの設計を述べ，クラウドコンピューティングに最適化する設計について述べる．5 章では，本研究による実装と既存システムのパフォーマンスを比較しながら，評価する．最後に，6 章で，まとめと今後の課題について述べる．

2 クラウドストレージとクラウド化システムの特徴

2.1 クラウドストレージ

Amazon S3 [4] や GCP Storage [7] など，耐久性，拡張性，コスト性に優れ，大容量のデータを保存できるオブジェクトストレージサービスが存在する．オブジェクトストレージサービスの特徴は，複数のノードが同一のストレージリソースを共有することや，ストレージの拡張がほぼ無限に可能なことである．オブジェクトストレージサービスでは，可用性と耐久性を確保するため，地理的に離れた場所に設置された複数のストレージサーバにデータを分散し，保存する．オブジェクトストレージ

1：Snowflake 社はクラウドベースのデータウェアハウスで，大規模かつ柔軟なデータ処理を提供する．異なるクラウドプロバイダーで利用可能であり，シンプルな操作とスケーラビリティが特徴である．

サービスは，実質的にクラウドコンピューティングの基本サービスとなっており，主要なクラウドプロバイダのオブジェクトストレージサービスのほとんどは，S3 プロトコルに対応している．オブジェクトストレージサービスの利用料金の計算方法は，いずれのサービスでも類似しており，データ保存量，API 呼び出し回数，およびリージョン間データ転送量に基づいて請求される．計算インスタンスとオブジェクトストレージが同一リージョンに存在する場合，データへのアクセス時あるいはデータの変更時にはリージョン間データ転送量に基づく請求は発生せず，データ保存量および API 呼び出し回数のみに基づいて利用料金が請求される．

オブジェクトストレージサービスが遅延時間に対してあまり敏感でない膨大なデータを保存するサービスであるのに対し，ブロックストレージサービスはデータの読み書きを高速かつ効率的に行えるように設計されているサービスである．ブロックストレージサービスの代表例は，AWS EBS [8] や Azure Managed Disks [9]，Google Compute Engine Persistent Disks [10] である．

一方，クラウドコンピューティングプロバイダが管理するストレージサービスを使用しない方法として，コンピューティングインスタンスのローカルディスクを使用する方法もある．コンピューティングインスタンスのうち，ローカルディスクを持つものは，インスタンスストアと呼ばれる．インスタンスストアは，インスタンス用のブロックレベルの一時ストレージを提供する．インスタンスの停止，終了，または基盤となるディスクドライブで障害が発生した場合は，データは失われる．したがって，インスタンスストアの使用用途は，キャッシュが適していると考えられる．

ネットワーク技術の発展により，今日のデータセンターのネットワーク帯域幅は NVMe (Non-Volatile Memory Express) [2] 帯域幅と同等になりつつある．2022 年，AWS は 200 Gbit/s のネットワーク帯域幅を備えたインスタンスを導入した [11]．その結果，ネットワーク接続を用いてストレージと計算を分離することが可能となり，プール化されたストレージが実現された．実際，オブジェクトストレージサービスは，ストレージプールを利用したサービスの一種である [12]．

2.2 データベースシステムのアーキテクチャの構成例

クラウドコンピューティング以前の 1990 年代，企業業務への対応は単純なモノリシックアーキテクチャのデータベースのみで対応可能であった．モノリシックアーキテクチャの特徴は，CPU，メモリ，ストレージをはじめとするすべての資源が密接に結合されていることである．モノリシックアーキテクチャの構成例を，図 1 に示す．一方，モノリシックアーキテクチャのデメリットは，物理マシン上で異なる資源を高い利用率で動作させることや，負荷に応じて各リソースを柔軟に調整することが難しい点である．そのため，モノリシックアーキテクチャの高可用性は，複数のマシンとリーダー/フォロワー方式によって実現されることが多い．

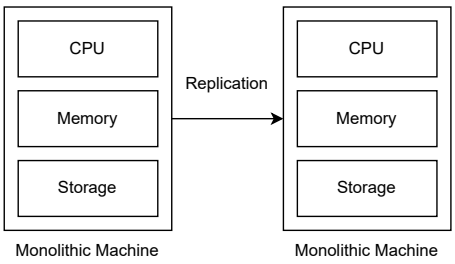


図 1: モノリシックアーキテクチャの構成例

インターネットの普及に伴い，2000 年前後には，大量のデータを処理する必要が生じ，データを分割して処理できるシェアードナッシングアーキテクチャのデータベースが主流になった．シェアードナッシングアーキテクチャの構成例を，図 2 に示す．シェアードナッシングアーキテクチャの特徴は，各ノードが独自のリソースを持ち，他ノードとリソースを共有しない点であり，拡張性や耐障害性，可用性を向上させることができる．シェアードナッシングアーキテクチャの代表例は，Google F1 [13]，TiDB [14]，NebulaGraph [15] であり，データセンター向けに設計されたデータベースとして広く使用されてきた．通常，データは複数に分割され，分割されたデータはパーティションと呼ばれる．各パーティションは，独立した計算資源を持っており，パーティションを Raft [16]，Paxos [17] といった合意アルゴリズムによって複数のサーバに複製することで，高可用性と高耐久性を実現する．また，シェアードナッシングアーキテクチャは，計算資源をより効果的に活用するため，分割後のデータが一定サイズを超えた場合に再分割を行うという特徴も持つ．しかし，データ再分割時はデータ複製コストが発生し，データ量が多い場合，かなりの時間を必要とする．例えば，2 つのコンピューティングインスタンス間のネットワーク帯域幅が 100 Gbps の場合，1 TiB のデータ複製に要する時間は約 85 秒である．さらに，シェアードナッシングアーキテクチャを用いたシステムでは，開発者自身がストレージシステムを構築しなければならないという特徴もある．例えば， Raft [16] アルゴリズムに基づく分散キーバリューストアを用いた，TiDB [14] というシステム全体の耐久性と可用性は，ユーザによる分散キーバリューストアのメンテナンス能力に依存する．そのため，優れた研究者やエンジニアをより多く投入することができる，大規模なクラウドコンピューティングプロバイダは，性能，可用性および耐久性に優れるストレージサービスを容易に提供することができる．例えば，アリババクラウドのストレージサービス基盤である Pangu [18] では，ユーザスペースストレージファイルシステムの使用や，FPGA 圧縮 [3]，RDMA [4] の活用といった最適化技術により，データを 3 つに分散して処理を行うことで，書き込みの遅延を数百マイクロ秒の範囲で制御している．

2：NVMe は，PCI Express（PCIe）バスを使用して高速データ転送を実現し，低遅延かつ高いスループットを提供する．

3：FPGA はデータ圧縮において，CPU に比べてより高い並列性，低遅延，高効率，および柔軟性を有する．

4：RDMA は，サーバー間で高性能，低遅延でデータを転送できる．

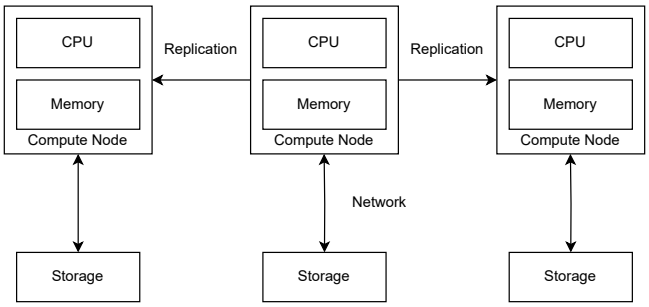


図 2: シェアードナッシングアーキテクチャの構成例

2010 年以降，Amazon S3 などのオブジェクトストレージサービスが成熟するにつれて，共有ストレージアーキテクチャのデータベースへの注目が高まっている．Snowflake [19] や Databricks [20]，AWS Redshift [21] は共有ストレージアーキテクチャの代表例である．共有ストレージアーキテクチャは，2.1 節で述べたオブジェクトストレージを用いたサービスであり，複数のノードが同一のストレージリソースを共有することや，ストレージをほぼ無限に拡張できることが特徴である．これらの特徴を活かし，共有ストレージアーキテクチャは，多くのクラウドネイティブデータウェアハウスシステムで使用されている．なお，クラウドネイティブデータウェアハウスシステムとは，クラウド化したデータウェアハウスのことである．共有ストレージアーキテクチャの構成例を，図 3 に示す．大量のデータ処理に対応するため，データを複数に分割して処理する点はシェアードナッシングアーキテクチャと同様であるが，共有ストレージアーキテクチャにおけるパーティション分割は論理的分割であり，データ複製が不要である点がシェアードナッシングアーキテクチャと異なる．共有ストレージアーキテクチャを用いたシステム全体の耐久性と可用性は，オブジェクトストレージの耐久性と可用性に依存している．

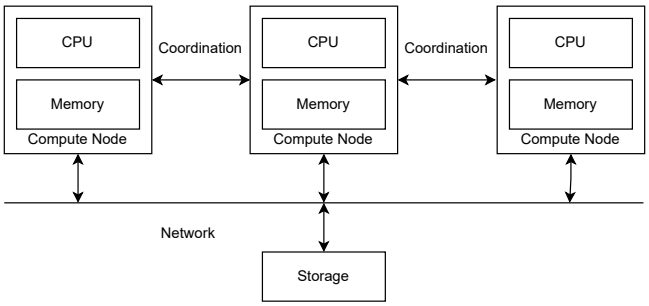


図 3: 共有ストレージアーキテクチャの構成例

3 関 連 研 究

Apache BookKeeper と Apache Kafka は，どちらもリモート WAL に利用できるシステムあるが，どちらのシステムもオブジェクトストレージサービスにデータを書き込むことはできない．本提案では，最新の書き込みデータをブロックストレージサービスに保存し，その他のデータをオブジェクトストレージへ保存することを可能とする．オブジェクトストレージへのアクセス遅延は高いが，オブジェクトストレージの高いスループットやバッチ処理を利用し，高い遅延コストを段階的に償却することができる．

本章では，Apache BookKeeper と Apache Kafka の特徴を述べ，本研究での実装と比較する．

3.1 Apache BookKeeper

Apache BookKeeper （以下，BookKeeper）[22] は，HDFS [23] や Herddb [24]，Twitter Manhattan [25] などのシステムでリモート WAL として使用されているシステムである．

BookKeeper の基本的なストレージ単位は Ledger で，Ledger は順番に並べられた Entry のセットとみなすことができる．各 Entry には WAL データが格納されている．BookKeeper では，Entry を複数のストレージノードに分散配置し，高可用性と高耐久性を実現している．BookKeeper は，シングルライターマルチリーダー (SWMR: Single Writer Multiple Reader) モデルであり，Ledger にデータを書き込む権限は 1 つのライター (Writer) にしかない一方，複数のリーダー (Reader) が同時に Ledger のデータを読み取ることができる．これにより，書き込み操作の順序の担保や，リーダーが独立して同時にデータを読み取ることが可能となる．各書き込み操作には，単調増加のインデックスが割り当てられる．書き込みは，複数のストレージノードに対して同時に書き込みリクエストを発行し，書き込みが多数派に達したときに，書き込みクライアントが最新の書き込みインデックスを LastAddConfirmed へ更新する．

読み取り操作は，LastAddConfirmed を活用することで，任意の Entry データを読み取ることができる．読み取り操作でクライアントが受け取るデータは一貫しており，一定数の障害やノードの故障にも対応できるようになる．

ただし，BookKeeper は ZooKeeper [26] というメタデータサービスに依存しており，BookKeeper を利用するためには，ZooKeeper クラスタを維持するための追加のリソースが必要である．一方，本実装は，メタデータを自己管理する仕組みを備えており，任意のメタデータサービスへ依存する必要がない点がメリットとなる．

3.2 Apache Kafka

Apache Kafka（以下，Kafka）[27] は分散型のストリーム処理プラットフォームであり，HoraeDB [28] システムでリモート WAL として使用されている．書き込みデータはメッセージと呼ばれ，メッセージは Kafka 内でトピックに組織される．各トピックは複数のパーティションに分かれ，各パーティションは順序付けられた変更不可能なメッセージのシーケンス（インデックス）である．パーティションによって Kafka は水平にスケーリングし，並列性とスループットの向上を提供する．各トピックは複数のプロデューサ（ライター）とコンシューマ（リーダー）を許容し，トピックに複数のパーティションがある場合，メッセージは各パーティション内でのみ順序が保証される．Kafka では，データの高可用性と高耐久性を確保するため，データレプリカが使用されていている．各パーティション

は複数のレプリカを構成でき，これらのレプリカは異なるノードに分散して配置される．ノードが障害に陥った場合でも，他のレプリカは利用可能であり，データの永続性が確保される．

Kafka が高いスループットに達する主要なメカニズムの 1 つは，時間ベースのバッチ処理を通じて効率を向上させることである．しかし，時間ベースのバッチ処理のデメリットは，サイズが小さなデータを書き込む際，書き込み遅延が発生する可能性が非常に高いことである．本実装では，時間ベースではないバッチ処理を行うことにより，サイズが小さなデータでもわずかな遅延時間で書き込むことを実現する．

3.3 Apache RocketMQ

Apache RocketMQ（以下，RocketMQ）[29] は Kafka に似たストリーム処理プラットフォームである．Kafka との最大の違いは，RocketMQ がすべてのトピックで同じデータファイルを共有するのに対し，Kafka はトピックごとに個別のデータファイルを配置する点である．トピック数が多い場合，Kafka では書き込みファイルの切り替えが頻繁に行われるため，パフォーマンスが下がる．ただし，RocketMQ においても，単一のデータファイルでは，高性能のディスク IO 帯域幅を完全に活用することはできないという課題がある．

3.4 共有ストレージアーキテクチャ

代表的な共有ストレージアーキテクチャのシステムを図 4 に示す．システムに書き込みリクエストが届く (1. Write Request) と，データをリモート WAL に書き込んだ (2. Write, 3. Ack) のち，メモリ内の状態を更新 (4. Update Watermark) し，最後にクライアントに応答を返す (5. Write Response)．

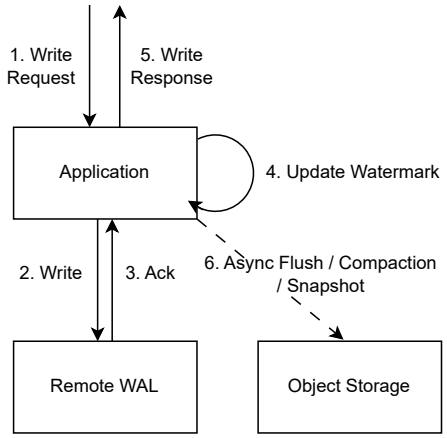


図 4: 共有ストレージアーキテクチャの構成例

共有ストレージアーキテクチャでは，リモート WAL を用いてメモリデータを複数のシステムの間に共有でき，読み取りリクエストの処理能力を簡単に拡張できる．図 5 に，共有ストレージアーキテクチャの構成例を示す．新しいデータベースを起動し，同じパーティションのデータベースの WAL をサブスクライブすることで同じメモリデータを維持 (3. Subscribe) し，一部のクエストに応じる (4. OLAP Quires)． これにより，システム全体のクエスト処理能力が倍になる．

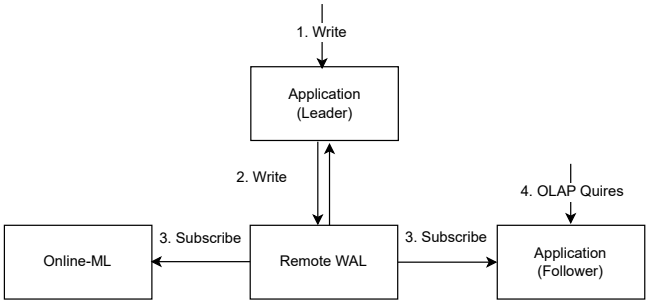


図 5: 読み取りリクエストの処理能力を拡張する例

システム全体で，最適なパフォーマンスを維持するために，常に一定量のデータだけを計算資源で処理することは重要である．リモート WAL でのデータ分割の流れを図 6 に示す．リモート WAL では，あるパーティションが閾値に達し，再度分割が必要となった場合，データベースを新たに起動する (Application (Part 1))．その後，分割するパーティションの WAL からメモリデータ（またオブジェクトストレージに保存されていないデータ）をサブスクライブ (3. Subscribe) し，新規パーティションに属するデータのみを新たに起動したデータベースに保存する (4. Write)．

新しいパーティションデータをメモリに保存した後，新規書き込みが適切なパーティションに行われるように，メタデータの書き込み先を更新する．続いて，オブジェクトストレージに存在するファイルに対して論理的分割を行う．論理的分割のイメージを図 7 に示す．

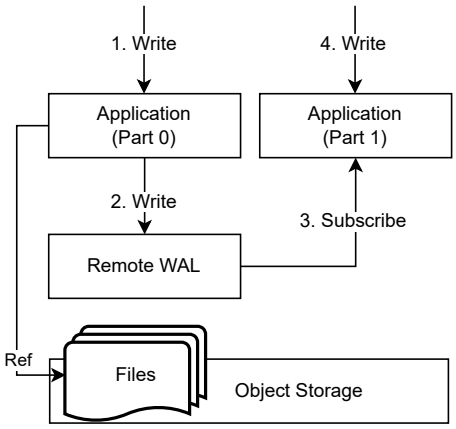


図 6: リモート WAL でのデータ分割の流れ

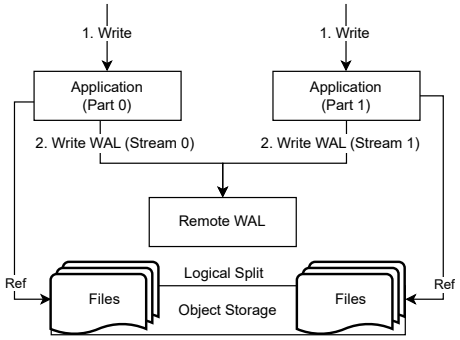


図 7: 論理的分割のイメージ図

4 提案と実装

4.1 システムデザイン

本実装のデザインを図 8 に示す．データベースから追加ログリクエストが届く (1. Append Log Request) と，ディスク上の WAL にデータを書き込んだ (2. Write) のち，メモリ内の状態を更新 (4. Update Watermark) し，最後にクライアントに応答を返す (5. Append Log Response) ．また，定期的に，あるいはメモリ閾値を超えた時には，メモリデータをオブジェクトストレージに非同期的保存する (6. Async Flush/Truncate).

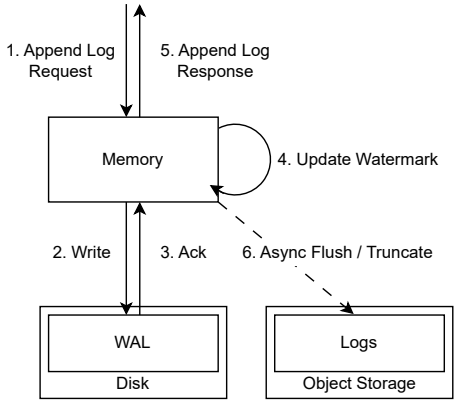


図 8: 本提案のデザイン

本システムは， Thread per core の設計を採用しており，ストレージエンジンには CPU コア数の 1/2 のワーカー（ユーザーが設定可能）が存在する．各ワーカーは Worker ループを実行するために独立したスレッドを起動する．各ワーカーは，ループ内のリソースのうち 1 つのスレッドにしかアクセスせず，ループ内の操作は Lock-free であるため，システムコールの負荷を削減することができる．Worker ループのプログラムを，ソースコード 1 に示す．Worker ループは，同時に到達したリクエストをまとめて処理し（ソースコード 1 の 14 行目），他の共有リソースのロックの負荷を軽減する．

```
/// Starts the worker loop.
async fn run(&mut self) {
    info!("Start worker thread {}", self.id);
    // Buffer to retrieve requests from receiver.
    let mut buffer = RequestBuffer::with_capacity(
        ↪ worker_request_batch_size);
    while self.running.load(Ordering::Relaxed) {
        // Clear the buffer before handling next
            ↪ batch of requests.
        buffer.clear();
        match self.receiver.recv().await {
            Some(request) => buffer.push(request),
            None => break,
        }
        // Try to recv more requests from the
            ↪ channel.
        for _ in 1..buffer.capacity() {
            // We have received one request so we
                ↪ start from 1.
            match self.receiver.try_recv() {
                Ok(req) => buffer.push(req),
                // We still need to handle remaining
                    ↪ requests.
                Err(_) => break,
            }
        }
        self.handle_requests(&mut buffer).await;
    }
    self.clean().await;
    info!("Exit worker thread {}", self.id);
}
```

ソースコード 1： Worker ループ

図 9 に全体の流れのイメージを示す．オブジェクトストレージ上のファイルを削除する，データをオブジェクトストレージに書き込むなど，通常数百ミリ秒かかる操作は， Worker ループ内で非同期タスクを生成し，Flush Scheduler などの外部のスケジューラによってスケジュールされ，非同期に実行される．タスクの実行が完了すると，対応するワーカーに通知され (task flush notification)，メタデータに書き込まれる．

ログデータが書き込まれた順序でファイルに保存されるようにするため，データをファイルに書き込む操作を Worker ループで同期し，データを配置する必要がある．ログデータの書き込み操作は Write システムコールを呼び出す必要がある．システムコール呼び出しでは，多くの場合で相当数のシステムコールのオーバーヘッドが発生する．システムコールのコストを段階的に償却するために，本システムでは，複数の書き込みリクエストを一括して WAL に書き込むことができるグループ書き込みを導入した．図 10 にグループ書き込みのイメージ図を示

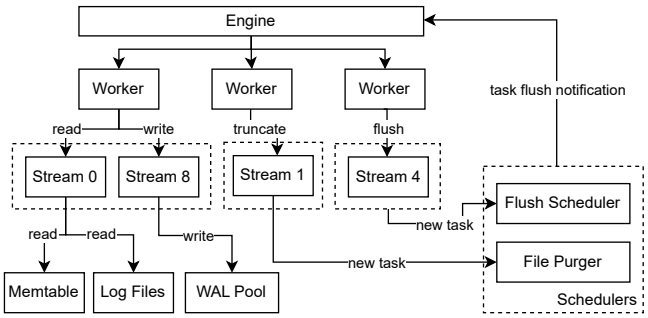


図 9: システム全体の流れ

す．データを書き込んでいるスレッドが存在する場合，他の到着した書き込みリクエストは，到着した順序でグループを形成する (1. Make group)．書き込み中のスレッドが完了 (2. End) すると，待機していた書き込みリクエストグループは書き込みロックを取得 (3. Wake up) し，データが一括して WAL に書き込まれる．

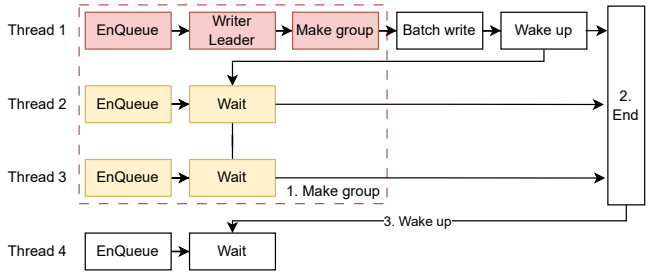


図 10: グループ書き込み

4.1.1 WAL 数の最適化

本システムでは，最適化された固定数の WAL を使用し，複数のストリームを同じ WAL に割り当てることで，IO 帯域の最大化を実現した．これは，Kafka や RocketMQ とは異なるアプローチである．Kafka は各トピックのデータ（本システムのストリームに相当）を独立したファイルに保存する（本システムの WAL に相当）．Kafka では，トピック数が一定数に達すると，複数のファイルによる IO 競合でスループットが急激に低下することがある．RocketMQ はすべてのトピックデータを単一のファイルに保存するが，単一のファイルでは NVME SSD IO 帯域を十分に活用できない．

4.1.2 ストレージ階層の追加

本実装では，WAL データにオブジェクトストレージ階層を追加した．他のシステムがデータを単にディスクに保存するのとは異なり，本実装では，ディスク上に保存する WAL データは常に最低限のみとし，古いデータをすべてオブジェクトストレージに書き込んだ後，ディスク上のデータを削除する．ディスク上の WAL データが一定の量に達した場合，WAL データは最適なサイズ (8MiB) [22] にまとめられて S3 に書き込まれる．

本実装の WAL データは，Bitcask [30] のようなアーキテクチャを使用しており，ログデータは追加のみを許可している．また，非アクティブなファイルデータを削除するためのバッチ削除も可能である．1 つの WAL に対して 1 つのログデータ書き込みに使用される「アクティブファイル」が存在し，アクティブファイルが一定のサイズに達すると，非アクティブファイルとして閉じられ，以降変更されない．同時に，WAL では，新しいファイルがアクティブファイルとして開かれる．ログサーバのメモリ内は，ログデータのインデックスのみを持つ．具体的には StreamId, LogId が FileId, Offset, Len を指し示す形で保持している．ディスク上の WAL データが一定の量に達した場合，実装はすべての非アクティブファイルのデータをストリームごとにグループ化し，8MiB のデータブロックとしてオブジェクトストレージに書き込み，最後に必要なインデックスとメータデータを更新する．図 11 に，本実装における WAL とストリーム間のデータの流れを示す．

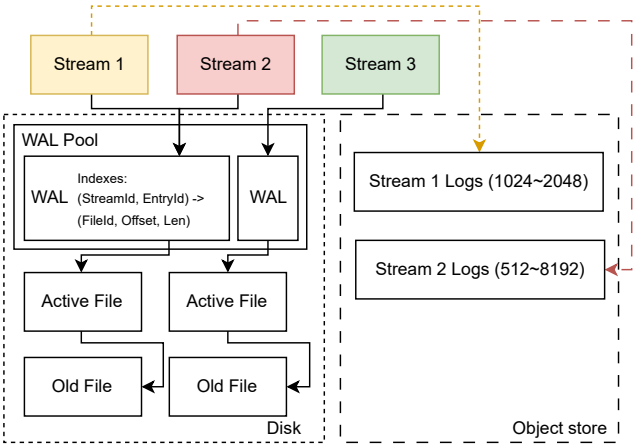


図 11: 本実装における WAL とストリーム間でのデータの流れ

次に，本実装による階層化ストレージの利点を図 12 に示す．クライアントが読み取りを要求したデータがすでにオブジェクトストレージに存在する場合，本実装は直接対応するオブジェクトストレージ上のファイルのアドレスを返し，クライアントはファイルのアドレスを用いてオブジェクトストレージからログデータファイルを自身でダウンロードする．このメカニズムにより，ブロックストレージの低い遅延時間と，オブジェクトストレージの極めて高いスループットの特性の両方を活用できる．

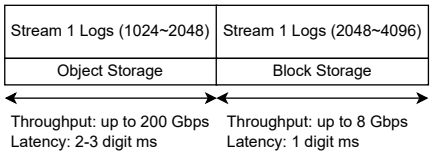


図 12: 本実装による階層化ストレージの利点

5 評価実験と考察

5.1 ローカルハードディスク環境で実験

評価実験は， Intel i9-13900K(基本周波数：3.0GHz，最大ブースト周波数：最大 5.4GHz) ， NVMe PCIe 4.0 SSD が配置された Ubuntu 23.10 で行った．

5.2 複数の WAL での評価

同一のマシン上にサーバとクライアントを配置し，単一のストリーム内で，クライアントはサーバへデータをシリアルに書き込み，評価を行った．異なる WAL 数を配置したスループットとレイテンシを比較した結果を表 1 に示す．

表 1 が示すように，ファイルの数[5]が増加すると，IO スループットを十分に活用できることが分かる．

表 1: 異なる WAL 数の比較．単一クライアントのスループット (MiB/s)

	1 WAL			4 WAL		
Size (bytes)	1 Stream	8 Streams	16 Streams	1 Stream	8 Streams	16 Streams
8,192	183.6	587.5	622.3	183.2	754.4	967.6
65,536	322.4	1,461.3	1,565.8	304.0	1,808.8	2,426.2
262,144	330.8	1,652.4	1,599.1	312.1	1,850.1	2,320.3

5.3 Kafka との比較

5.2 項により，WAL が 4 つの場合，より高いスループットを達成できることが分かった．そのため，本実装では WAL を 4 つ配置する．Kafka の replicate に 1，linger.ms にデフォルト値 25ms，batch.size にデフォルト値 16KiB をそれぞれ設定し，同一マシン上にサーバとクライアントを配置する．スループットとのレイテンシを比較した結果を表 2 に示す．

表 2 が示すように，Kafka は，書き込みデータ量 (Size) が閾値未満 (batch.size に設定した 16KiB) でかつシリアルに書き込む場合，スループットは非常に低くなることがわかる．

Kafka のバッチ処理性能は本実装よりも劣る傾向があり，通常，同じストリーム数で比較すると，Kafka のスループットは本実装よりも低くなる．つまり，Kafka でより高いスループットを達成するには，本実装より多くのストリームが必要である．WAL の場合，書き込みスループット上限は，ストリームあたりのスループットに依存すると考えられる．

表 2: Kafka との比較:単一クライアントのスループット (MiB/s)

	Kafka			4 WAL		
Size (bytes)	1 Stream	8 Streams	16 Streams	1 Stream	8 Streams	16 Streams
8,192	0.1	1.3	2.6	183.2	754.4	967.63
65,536	149.1	1,007.7	1,658.9	304.0	1,808.8	2,426.2
262,144	184.4	1,209.5	1,900.9	312.1	1,850.1	2,320.3

5.4 クラウド環境での実験

AWS EC2 をテスト環境として使用し，c6a-large と c6a-xlarge の 2 つの構成のマシンを使用する．どちらも 3.6 GHz の第 3 世代 AMD EPYC プロセッサを搭載し，12.5Gbps のピーク性能を備えたネットワーク帯域幅を持つ．c6a-large は 2 コア 4GB，c6a-xlarge は 4 コア 8GB を搭載している．

5：書き込みファイルの数と WAL の数は一致している．

5.5 ディスクへの書き込みパフォーマンスの比較

2 台の EC2 (c6a-large，3,000IOPS, 125MiB/s EBS を配置[6]) を用意し，1 台はサーバとして，他方はクライアントとして使用する．異なるストリームを配置し，クライアントはサーバへデータをシリアルに書き込む．WAL を 4 つ配置し，スループットとのレイテンシを比較した．ストリーム数が 1 の場合の結果を表 3 に，ストリーム数が 8 の場合の結果を表 4 に，ストリーム数が 16 の場合の結果を表 5 に，それぞれ示す．

表 3 が示すように，ストリーム 数が 1 の場合，本実装のスループットは 51MiB/s， 154.8MiB/s，142.8MiB/s であり，表 5.2 で示したローカルハードディスクのスループット (183.2MiB/s, 304.0MiB/s，312.1MiB) より低いことが分かった．これは，本実装の EBS 遅延が，ローカル SSD では数百ナノ秒程度に対し，本実装では数ミリ秒に達するためである．一方，Kafka はデフォルトのバッチ処理ポリシーにより順次書き込みを行うため，スループットが非常に低い．そのため，バッチの閾値に達することなく，書き込みごとに 25 ミリ秒待機する必要がある．

表 3: 1 ストリーム, c6a-large 単一クライアントのスループットとレイテンシ

	Size (Bytes)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	8,192	51.0	95	<1	441	490
	65,536	154.8	<1	<1	<1	8
	262,144	142.8	301	<1	4	7,702
Kafka	8,192	0.2	44,282	43,887	44,558	47795
	65,536	58.5	729	610	1,212	1,803
	262,144	84.7	1,688	1,453	2,435	5,520

表 4 が示すように，ストリーム数が 8 の場合，どちらのシステムのスループットも線形に増加していない．この理由は，ディスク書き込みの上限に達しているためであり，ストリーム数が増加すると，システムの遅延時間が長くなることが確認できる．さらに，双方のシステムが同様のスループットに達した場合（本実装が 187.9MiB/s，Kafka が 168.3MiB/s に達した場合），本実装のレイテンシ平均値は Kafka より低く，P50 と P99 の遅延時間も Kafka より短いことが分かる．

表 4: 8 ストリーム, c6a-large 単一クライアントのスループットとレイテンシ

	Size (Bytes)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	8,192	87.7	633	<1	3,277	5,194
	65,536	187.9	2,108	<1	23,428	34,139
	262,144	190.5	7,596	<1	33,345	62,645
Kafka	8,192	1.4	44,724	43,958	48,036	50,334
	65,536	117.3	3,776	2,558	9,949	19,110
	262,144	168.3	9,715	8,929	14,837	44,390

表 5 が示すように，本実装は，ストリーム数が 16 の場合，ストリーム数が 8 の場合と比較して，P90 と P95 の遅延時間が倍になることが分かる．システムハードディスクの書き込み制限に達した後も継続して書き込みを行うと，遅延時間は増加する．このような状況が発生した場合，サーバを垂直方向に拡

6：fio を使用して 4 つの同時書き込みタスク（I/O サイズ 128KiB）をテストした結果，最大の実際のスループットは 181MiB/s に達することが観察された．

張する (よりパフォーマンスの高いハードディスクを使う) か，水平方向に拡張する (データをより多くの個別のサーバに書き込む) 必要がある．

表 5: 16 ストリーム, c6a-large 単一クライアントのスループットとレイテンシ

	Size (Bytes)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	8,192	93.0	1,256	<1	6,594	10,691
	65,536	202.5	4,183	<1	37,871	74,749
	262,144	212.8	13,027	<1	65,784	90,117
Kafka	8,192	2.7	44,782	43,976	48,233	57,954
	65,536	124.2	7,471	5,741	18,176	40,631
	262,144	172.7	20,943	19,949	31,452	58,069

5.6 異なるサイズマシンの比較

EC2 の一部分をマウントした EBS [31] では，24 時間のうち，最大パフォーマンスを 30 分間維持できる．本評価実験では，5.2.1 項で述べた評価実験とは異なる 2 台の EC2 (c6a-xlarge, 3,000IOPS, 125MiB/s EBS を配置 [7]) を用意し，1 台はサーバとして，他方はクライアントとして，それぞれ使用する．実験は 3 種類のストリームで行い，クライアントはサーバへデータをシリアルに書き込む．スループットとレイテンシを比較した結果を表 6 に示す．

表 6 が示すように，c6a.xlarge インスタンス上では，本実装は Kafka よりもやや高いスループットを示している．ストリーム数が 1 の場合，本実装のスループットは Kafka の 2 倍以上であり，遅延時間も非常に短い．ただし，ストリーム数が増加すると，本実装の遅延時間が急激に増大し，Kafka の 2 倍の遅延時間となる．これは，本実装の単一のバッチ処理リクエスト数が大きすぎることが原因と考えらる．単一のバッチ処理リクエストを増やすことでスループットを向上させることができるが，遅延時間も増加すると考えられる．

表 6: c6a-xlarge 単一クライアントのスループットとレイテンシ

	Size (Bytes)	Stream(s)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	65,536	1	179.4	<1	<1	<1	4
		8	274.5	1,305	<1	12,659	24,174
		16	339.3	2,369	<1	22,360	38,729
	262,144	1	189.6	<1	<1	<1	1
		8	303.7	4,442	<1	22,342	46,494
		16	337.3	8,781	<1	43,842	78,237
Kafka	65,536	1	60.5	652	581	1,038	1,437
		8	250.4	1,492	1,106	3,078	6,288
		16	263.3	3,256	2,593	6,572	14,826
	262,144	1	90.3	1,457	1,355	1,912	2,572
		8	318.1	4,218	3,513	6,620	24,844
		16	324.3	10,220	9,138	21,875	37,480

5.7 オブジェクトストレージへアップロード時の影響

本評価実験では，2 台の EC2 (c6a-xlarge, 3,000IOPS, 125MiB/s EBS を配置) を用意し，1 台はサーバーとして，他方はクライアントとして，それぞれ使用する．ストリーム数は 16 とし，クライアントはサーバへデータをシリアルに書き込む．本システムには，128MiB と 256MiB という異なるフラッシュ閾値を設定し，EBS への書き込みデータ量が閾値に達すると，本実装は，EBS へ書き込んだ全てのデータを S3 に書き込む．スループットとレイテンシを比較した結果を表 7 に示す．

表 7 が示すように，本実装に 128MiB フラッシュ閾値を設定すると，システムへの書き込み遅延時間が短いことが分かる．また，少量のデータの S3 への書き込みはパフォーマンスにほとんど影響を与えないことも分かる．この理由は，S3 への書き込み後にディスク上のデータが削除されるためである．頻繁なデータの削除により，OS が書き込みバッファに保持しているデータがすぐに削除され，fsync によるオーバーヘッドが少なくなる．

表 7: 16 ストリーム, c6a-xlarge, 単一クライアントのスループットとレイテンシ

	フラッシュ閾値	Size (Bytes)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	設定しない	65,536	339.3	2,369	<1	22,360	38,729
		262,144	337.3	8,781	<1	43,842	78,237
	128MiB	65,536	306.1	2,677	<1	10,565	21,548
		262,144	344.1	8,589	<1	33,100	52,080
	256MiB	65,536	305.1	2,695	<1	12,735	30,207
		262,144	346.8	8,518	<1	34,867	53,376
Kafka	該当なし	65,536	263.3	3,256	2,593	6,572	14,826
		262,144	324.3	10,220	9,138	21,875	37,480

5.8 読み込みスループットの比較

2 台の EC2 (c6a-large, 3,000IOPS, 125MiB/s EBS を配置 [8]) を用意し，1 台はサーバとして，他方もう 1 台はクライアントとして，それぞれ使用する．実験は 3 種類のストリームで行い，異なるストリームを配置し，クライアントはサーバへデータをシリアルに書き込む．本評価実験では，WAL を 4 つ配置し，1KiB のデータを書き込む．フラッシュ閾値を超えるとデータは 8MiB のサイズにまとめられ，S3 に書き込まれる．スループットとレイテンシを比較した結果を表 8 に示す．

表 8 が示すように，フラッシュ閾値を設定せず，ストリーム数が 1 の場合，本実装のシステムの読み込スループットは低い．これは，読み取りのデータサイズが小さすぎるためであり，バッチ読み取りの最適化が行われていないためである．ストリーム数が増加すると，本実装の読み込みスループットは Kafka よりも高く，遅延時間も Kafka と同じレベルである．フラッシュ閾値を 128MiB と設定した場合は，本実装のスループットは 500MiB/s に達したことが分かる．これは，ディスクへの書き込み量が 128MiB に達すると，ディスクのデータが S3 に書き込まれるためである．実験で使用した EC2 のネットワーク帯

7：fio を使用して 4 つの同時書き込みタスク（I/O サイズ 128KiB）をテストした結果，最大の実際のスループットは 228MiB/s に達することが観察された．

8：fio を使用して 4 つの同時書き込みタスク（I/O サイズ 128KiB）をテストした結果，最大の実際のスループットは 181MiB/s に達することが観察された．

域幅は 10Gbps，EBS の帯域幅は 1Gbps 未満である．そのため，多くデータが S3 に存在する場合，読み取りのスループットでは，EC2 のネットワーク帯域幅を最大限に活用できる．

フラッシュ閾値として 128MiB を設定した場合，1 回あたりの読み込みデータサイズは 6,144KiB であり，読み込み遅延時間は 1KiB の場合の約 6,000 倍となった．本実装は，10Gbps 帯域幅を最大限に活用するために，クライアントには 25 の並行ダウンロードスレッドを構成している．これにより，遅延時間の中央値は 250ms となり，最適な値となった [22]．一方，スレッド数が 25 を超えると，ネットワーク帯域幅がボトルネックとなり，遅延時間が大きくなる．

表 8: 16 ストリーム, c6a-large, 読み込みスループットとレイテンシ

	フラッシュ閾値	Stream(s)	Avg Size (KiB)	Throughput (MiB/s)	Avg (Microsecs)	P50	P95	P99
本実装	設定しない	1	1	29.6	33	<1	126	225
		8	1	278.3	28	<1	1	145
		16	1	292.6	54	<1	1	238
	128MiB	1	5,212	518.0	308,925	33,0786	469,090	497,854
		8	6,247	505.2	243,438	250,485	430,823	474,986
		16	6,102	510.5	219,155	208,761	403,311	420,780
Kafka	該当なし	1	1	157.6	6	<1	<1	<1
		8	1	178.3	45	<1	<1	<1
		16	1	224.9	74	<1	<1	<1

6 おわりに

6.1 まとめ

本研究では，クラウド化データベースに向け，効率化されたリモートログ先行書き込みシステムの提案を実装した．データ書き込みの場合，本実装は既存の Kafka システムと同等のスループットを提供でき，書き込み遅延時間も Kafka と同等である．ストレージ階層を追加した後に，システムの書き込み遅延時間がやや増加することがあるが，これは許容可能な範囲内と考えている．フラッシュ閾値を 128MiB または 256MiB に設定する場合，データを S3 に書き込んでもシステムの書き込み性能にはほとんど影響がないことが分かった．さらに，大量の WAL データを読み取る際には， S3 のネットワーク帯域幅を活用してスループットの向上を図ることが可能だと分かった．

6.2 今後の課題

本研究の評価結果により，フラッシュ閾値が大きくなるとシステム全体への影響が大きくなり，書き込み遅延時間が長くなることが分かった．単にフラッシュ閾値を調整するのではなく，オブジェクトストレージへのデータの書き込みをシステム負荷が軽い時に自動的に行うといった適応アルゴリズムの導入を検討する．

また，いくつかの状況では，本実装のスループットが高い場合でも，書き込みの遅延が大きいことがあった．書き込み遅延が大きくなる原因は，1 度のバッチ処理のリクエスト数が大きすぎるためと考えられ，システムの処理上限に達した後に，一部のリクエストが処理されるまでに長時間を要する．今後，マシンの性能に基づいてバッチ処理のリクエスト数を動的に調整するメカニズムの導入を検討する．

文　献

[1] DataCenterDynamics. Amazon CTO, Developers Should Address Cloud Downtime. June, 2011. https://www.datacenterdynamics.com/en/news/amazon-cto-developers-should-address-cloud-downtime/ (2023 年 12 月 3 日閲覧).

[2] Alibaba Cloud Reveals Innovations for a Sustainable and Inclusive 11.11 Global Shopping Festival. November 2021. https://www.alibabagroup.com/en-US/document-1490857260237193216 (2023 年 11 月 28 日閲覧).

[3] Tencent Completes Major Migration of its Services to Tencent Cloud. March 2019. https://technode.com/2019/03/21/alibaba-to-run-its-entire-business-on-the-cloud/ (2023 年 12 月 3 日閲覧).

[4] Amazon S3 - Cloud Object Storage. https://aws.amazon.com/s3/(2023 年 9 月 24 日閲覧).

[5] Snowflake Shares More Than Double. It's The Biggest Software IPO Ever. September 2020. https://edition.cnn.com/2020/09/16/investing/snowflake-ipo/index.html (2023 年 11 月 28 日閲覧).

[6] DBMS の市場規模、DBaaS がオンプレミスに並ぶ. https://www.itmedia.co.jp/news/articles/2204/18/news063.html (2023 年 12 月 9 日閲覧).

[7] Cloud Storage — Google Cloud. https://cloud.google.com/storage (2023 年 11 月 9 日閲覧).

[8] Amazon Elastic Block Store. https://aws.amazon.com/ebs/ (2023 年 9 月 24 日閲覧).

[9] Azure Managed Disk. https://learn.microsoft.com/en-us/azure/virtual-machines/disks-types (2023 年 9 月 24 日閲覧).

[10] Google Compute Engine Persistent Disks. https://cloud.google.com/persistent-disk (2023 年 9 月 24 日閲覧).

[11] Introducing Amazon EC2 C6in Instances. https://aws.amazon.com/about-aws/whats-new/2022/11/introducing-amazon-ec2-c6in-instances/(2023 年 11 月 9 日閲覧).

[12] Dominik Durner, Viktor Leis, and Thomas Neumann. Exploiting Cloud Object Storage for High-Performance Analytics. *VLDB*, Vol. 16, No. 11, p. 2769–2782, Jul 2023.

[13] Jeff Shute, Radek Vingralek, Bart Samwel, Ben Handy, Chad Whipkey, Eric Rollins, Mircea Oancea, Kyle Littlefield, David Menestrina, Stephan Ellner, John Cieslewicz, Ian Rae, Traian Stancescu, and Himani Apte. F1: A Distributed SQL Database That Scales. In *VLDB*, 2013.

[14] Dongxu Huang, Qi Liu, Qiu Cui, Zhuhe Fang, Xiaoyu Ma, Fei Xu, Li Shen, Liu Tang, Yuxing Zhou, Menglong Huang, Wan Wei, Cong Liu, Jian Zhang, Jianjun Li, Xuelian Wu, Lingyu Song, Ruoxi Sun, Shuaipeng Yu, Lei Zhao, Nicholas Cameron, Liquan Pei, and Xin Tang. TiDB: A Raft-Based HTAP Database. *VLDB*, Vol. 13, No. 12, p. 3072–3084, August 2020.

[15] NebulaGraph: Open Source Distributed Graph Database. https://www.nebula-graph.io/ (2023 年 11 月 19 日閲覧).

[16] Diego Ongaro and John Ousterhout. In Search of an Understandable Consensus Algorithm. In *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, pp. 305–319, Philadelphia, PA, Jun 2014. USENIX Association.

[17] Leslie Lamport. Paxos Made Simple. *ACM SIGACT News 32, 4 (Whole Number 121, December 2001), Distributed Computing Column*, pp. 51–58, December 2001.

[18] Qiang Li, Qiao Xiang, Yuxin Wang, Haohao Song, Ridi Wen, Wenhui Yao, Yuanyuan Dong, Shuqi Zhao, Shuo Huang, Zhaosheng Zhu, Huayong Wang, Shanyang Liu, Lulu Chen, Zhiwu Wu, Haonan Qiu, Derui Liu, Gexiao Tian, Chao Han, Shaozong Liu, Yaohui Wu, Zicheng Luo, Yuchao Shao, Junping Wu, Zheng Cao, Zhongjie Wu, Jiaji Zhu, Jinbo Wu, Jiwu Shu, and Jiesheng Wu. More than Capacity: Performance-Oriented Evolution of Pangu

in Alibaba. In *Proceedings of the 21st USENIX Conference on File and Storage Technologies*, FAST'23, USA, 2023. USENIX Association.

[19] Benoit Dageville, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes, Jon Bock, Jonathan Claybaugh, Daniel Engovatov, Martin Hentschel, Jiansheng Huang, Allison W. Lee, Ashish Motivala, Abdul Q. Munir, Steven Pelley, Peter Povinec, Greg Rahn, Spyridon Triantafyllis, and Philipp Unterbrunner. The Snowflake Elastic Data Warehouse. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, p. 215–226, New York, NY, USA, 2016. Association for Computing Machinery.

[20] Alexander Behm, Shoumik Palkar, Utkarsh Agarwal, Timothy Armstrong, David Cashman, Ankur Dave, Todd Greenstein, Shant Hovsepian, Ryan Johnson, Arvind Sai Krishnan, Paul Leventis, Ala Luszczak, Prashanth Menon, Mostafa Mokhtar, Gene Pang, Sameer Paranjpye, Greg Rahn, Bart Samwel, Tom van Bussel, Herman van Hovell, Maryann Xue, Reynold Xin, and Matei Zaharia. Photon: A Fast Query Engine for Lakehouse Systems. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, p. 2326–2339, New York, NY, USA, 2022. Association for Computing Machinery.

[21] Nikos Armenatzoglou, Sanuj Basu, Naga Bhanoori, Mengchu Cai, Naresh Chainani, Kiran Chinta, Venkatraman Govindaraju, Todd J. Green, Monish Gupta, Sebastian Hillig, Eric Hotinger, Yan Leshinksy, Jintian Liang, Michael McCreedy, Fabian Nagel, Ippokratis Pandis, Panos Parchas, Rahul Pathak, Orestis Polychroniou, Foyzur Rahman, Gaurav Saxena, Gokul Soundararajan, Sriram Subramanian, and Doug Terry. Amazon Redshift Re-Invented. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, p. 2205–2217, New York, NY, USA, 2022. Association for Computing Machinery.

[22] Flavio P. Junqueira, Ivan Kelly, and Benjamin Reed. Durability with BookKeeper. *SIGOPS Oper. Syst. Rev.*, Vol. 47, No. 1, p. 9–15, Jan 2013.

[23] BookKeeper As A Shared Storage (EXPERIMENTAL). `https://hadoop.apache.org/docs/r2.5.2/hadoop-project-dist/hadoop-hdfs/HDFSHighAvailabilityWithNFS.html` (2023 年 12 月 3 日閲覧).

[24] A JVM-embeddable Distributed Database. `https://github.com/diennea/herddb` (2023 年 12 月 3 日閲覧).

[25] Strong Consistency In Manhattan. `https://blog.twitter.com/engineering/en_us/a/2016/strong-consistency-in-manhattan` (2023 年 12 月 3 日閲覧).

[26] Apache ZooKeeper. `https://zookeeper.apache.org/` (2023 年 12 月 28 日閲覧).

[27] Jay Kreps, Neha Narkhede, Jun Rao, et al. Kafka: A Distributed Messaging System For Log Processing. In *NetDB*, Vol. 11, pp. 1–7. Athens, Greece, 2011.

[28] Apache HoraeDB. `https://horaedb.apache.org/` (2023 年 12 月 28 日閲覧).

[29] Apache RocketMQ. `https://rocketmq.apache.org/` (2024 年 1 月 22 日閲覧).

[30] Justin Sheehy and David Smith. Bitcask: A Log-Structured Hash Table for Fast Key/Value Data. `https://riak.com/assets/bitcask-intro.pdf` (2024 年 1 月 7 日閲覧).

[31] Amazon EBS–optimized Instances. `https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-optimized.html#ebs-optimization-support` (2024 年 1 月 7 日閲覧).

コンピュート・ストレージ分離アーキテクチャDBを用いた高可用かつ低コストな冗長構成の提案

林　友佳† 渡　佑也† 堀内　美聡†

† NTT ソフトウェアイノベーションセンタ　〒 108–0075 東京都港区港南 1-2-70 品川シーズンテラス 29F
E-mail: †{tomoka.hayashi,yuya.watari,ms.horiuchi}@ntt.com

あらまし　近年，システムに対するユーザの要求が高度化し，そのひとつとして高可用が求められている．従来の冗長構成では，可用性を向上させるために冗長構成を構築すると，リソース消費量が増加し，金銭的コストも増加してしまうというトレードオフの関係がある．特に，一部機能だけの停止や部分的なリソースの縮小が可能なアプリケーションサーバ等に比べ，データベースは全体を起動するか停止するかの二択であることから，起動と停止の間でトレードオフの極端なギャップが発生するという課題がある．本研究では，コンピュート・ストレージ分離アーキテクチャデータベースに着目し，データベースコンポーネントの一部のみを停止することで，低リソース消費，低コストでありながら，高可用な冗長構成であるミニマムホットスタンバイ構成を提案する．本稿では，ミニマムホットスタンバイ構成のプロトタイプ実装と，既存の冗長構成との比較実験について述べる．実験によって，ミニマムホットスタンバイ構成が，コールドスタンバイ構成より高い可用性を実現しつつ，ホットスタンバイ構成より低コストな構成であることを示した．

キーワード　データベース技術，障害回復，コンピュート・ストレージ分離アーキテクチャ，疎結合アーキテクチャ，冗長化，コスト削減

1 はじめに

近年，IT システムが一般に普及するに伴い，IT システムに対するユーザの要求が高度化している．高度化する要求のひとつとして，自然災害の発生や，IT システムの基盤であるオンプレミスサーバ，クラウドコンピューティングなどの障害にかかわらず，IT システムの提供を継続してほしいという，高可用やディザスタリカバリ（以降，DR）の要求がある．自然災害や人為的なミスによって，サーバやデータセンタの部分的な障害は数多く発生しており [1]，2023 年 1 月には，Azure においてパブリッククラウドのほぼすべてのゾーンが使用できなくなるような障害も発生した [2], [3]．IT システムの運用者は，このような障害が発生した際も IT システムの提供を継続するため，複数のゾーンのデータセンタや複数のクラウド，複数のサーバを用いた可用性向上手法を実施している [4], [5]．

可用性向上の手法として，プライマリ（主系）とスタンバイ（待機系）のサーバやクラウドを用いて，ホットスタンバイ，ウォームスタンバイ，コールドスタンバイといった冗長構成を構築するという手法がある．これらの冗長構成は，それぞれ必要なリソースや金銭的コストと，その結果達成できる可用性の度合いが異なるため，IT システムの運用者が各 IT システムに求められる可用性や予算に応じて適切なものを選択する．しかし，どの冗長構成であっても，リソース消費量や金銭的コストと可用性の間にトレードオフが発生するという問題がある．

筆者らは，リソース消費量や金銭的コストと可用性の間に極端なトレードオフがある現状に対し，そのトレードオフが解消できるような新たな IT システム運用の世界を目指している．本研究では，平常時のスタンバイ側サーバやクラウド上でデータベース（以降，DB）の部分的な停止を実現することにより，可用性と金銭的コストの間に生じる極端なトレードオフを改善するミニマムホットスタンバイ構成を提案する．特に本稿では，ミニマムホットスタンバイ構成の実現パターン 3 種のうち 1 種をプロトタイプ実装し，既存の冗長構成との比較実験を行い，提案手法によるトレードオフの改善度合いを明らかにする．

2 可用性向上のための冗長構成

本節では，既存の冗長構成 3 種について，可用性や金銭的コストの視点で各構成の概要をまとめ，各構成間に存在するトレードオフとその要因について整理する．

2.1 既存の冗長構成の概要

可用性向上を目的として構築する冗長構成には，主にホットスタンバイ，ウォームスタンバイ，コールドスタンバイの 3 つの構成が存在する．

ホットスタンバイとは，平常時にシステムを提供するプライマリを用意した上で，スタンバイ上でもまったく同一のコンポーネントを動作させ，データも同期しておくことで，プライマリとスタンバイで同一のシステム構成を保つ構成である．プライマリのサーバやクラウドに障害が発生した際は，クライアントからの接続先をスタンバイ側に切り替えるフェイルオーバという操作を行うことで，素早くシステム提供が再開できる．一方で，プライマリとスタンバイで同様のシステムを常時稼働

し続けることから，冗長化せずに運用する場合と比べて金銭的コストは 2 倍程度かかる構成である．

ウォームスタンバイは，平常時にプライマリで動作しているアプリケーションサーバやフロントエンドサーバの一部と，DB の全体を，スタンバイ上でも動作させる構成である．プライマリサーバやクラウドに障害が発生した際には，スタンバイ上のコンポーネント数を平常時のプライマリと同程度まで増やすなどして増強することで，スタンバイ上でのシステム提供を再開する．増強のためにコンポーネントを追加で起動することから，ホットスタンバイよりもフェイルオーバ完了に時間がかかってしまう一方で，平常時のリソース消費量がホットスタンバイよりも少なく，金銭的コストも抑えられる構成である．

コールドスタンバイは，日次，週次，月次などあらかじめ決定した時間ごとにプライマリ側データのバックアップを取得し，スタンバイにコピーする構成である．プライマリサーバやクラウドに障害が発生した際は，スタンバイにあるバックアップデータをもとに，システム提供に必要なコンポーネントを起動，復旧させ，スタンバイ上でのシステム提供を再開する．平常時のランニングコストがかからないため，3 つの手法の中で最も安価な冗長構成である．一方で，障害が発生してからスタンバイ上でのシステム構築を行うため，障害発生時のダウンタイムは 3 つの手法の中で最も長い．また，バックアップ頻度とデータの更新頻度によっては，最終バックアップから障害発生時までにデータの差分が発生し，そのデータはフェイルオーバ後のスタンバイ側では参照することができない．プライマリサーバやクラウドのディスクが破損していなければ，プライマリ側が復旧した後にデータを取り戻すことが可能であるが，ディスクが破損した場合は最終バックアップ後の更新データを損失することになる．

2.2 既存の冗長構成に発生するトレードオフ

前項で述べた 3 種の冗長構成においては，可用性を向上させようとすればするほどリソース消費量が増加し，金銭的コストがかかるため，可用性と金銭的コストがトレードオフの関係にあると言える．このトレードオフの関係は，状態を持たないフロントエンドサーバやアプリケーションサーバと状態を持つ DB では，縮退の可否が異なるために生じる．ホットスタンバイ，ウォームスタンバイ，コールドスタンバイの構成を構築するにあたって，状態を持たないフロントエンドサーバやアプリケーションサーバは段階的にコンポーネント数を減らし，最終的に完全に停止するといったように段階的に縮退できる．一方で，DB は状態を持ち，データを永続化する必要があることから，一部のみを停止することができない．

DB を起動するホットスタンバイやウォームスタンバイのような構成は，可用性が向上する代わりにランニングコストが高い手法である．平常時にはプライマリ上とスタンバイ上の DB を用いて，常時レプリケーションを行うため，コールドスタンバイに比べ高い目標復旧時点（以降，RPO）を実現できる．また，平常時のスタンバイ上で DB が起動していることから，フェイルオーバ時に新たに DB を起動する必要がなく，その分高い目標復旧時間（以降、RTO）を達成することができる．一方で、平常時のスタンバイ上で DB を起動しておくだけのリソース消費やランニングコストが発生し、ランニングコストがコールドスタンバイよりも高い手法となる。

DB を停止することが可能なコールドスタンバイは，ランニングコストが削減できる代わりに可用性が低下する構成である．平常時にはバックアップに使用するコンポーネントのみを低頻度で起動するため，ランニングコストを抑えることができる．一方で，バックアップ間隔が長いことにより，達成できる RPO が低下する．さらに，フェイルオーバ時のコンポーネントの起動に時間を要するため，達成できる RTO もホットスタンバイなどの手法に比べ長くなる．

以上の通り，ホットスタンバイやウォームスタンバイの際は，すべての DB をプライマリと同様にスタンバイでも起動させ，コールドスタンバイでは基本的に DB 全体を停止させておくというように，DB の起動状態は二極化している．すなわち，DB の部分的な停止ができないことによって，コールドスタンバイとウォームスタンバイやホットスタンバイの間で，可用性とリソース消費量や金銭的コストのトレードオフに極端な開きがある．

3 コンピュート・ストレージ分離アーキテクチャ DB

本節では，本研究で扱うコンピュート・ストレージ分離アーキテクチャ（以降，CS 分離）DB に関して，初めてこのアーキテクチャを導入した Amazon Aurora と，オープンソースソフトウェア（以降，OSS）として開発されている Neon を中心に説明する．

3.1 Amazon Aurora

Amazon Aurora [6] は，Amazon Web Services, Inc. が開発し，マネージドサービスとして提供している CS 分離 DB の実装のひとつである．CS 分離 DB とは，コンピュート層とストレージ層に処理機能を分離したアーキテクチャの DB である．コンピュート層は，クエリの解析，集計，実行など，計算資源を主に利用して行う処理部分を担っている．ストレージ層は，データの冗長化，永続化処理など，データの保存に関する処理部分を担っている．

Aurora では，従来のリレーショナル DB のコンポーネントのほとんどをコンピュート層のコンポーネントである Instance に残しつつ，トランザクションログの処理や永続化，クラッシュリカバリ，バックアップ，リストアの機能を分散されたストレージ層にオフロードすることで CS 分離を実現している．さらに，CS 分離によって発生するコンピュート層とストレージ層の間のネットワークボトルネックを解消するため，ストレージ層へはトランザクションログのみを転送するような工夫を行っている．このアーキテクチャ実装によって，コンピュート層とストレージ層を別々にスケール可能なスケーラビリティ，コンポーネント別に起動や停止が可能な柔軟性，他のコンポーネントに処理

リソースを奪われない独立性，従来のモノリシックなアーキテクチャの DB よりも高い耐障害性などといった利点を獲得した．スケーラビリティや柔軟性を活かした負荷分散やスループットの向上，独立性を活かしたフォアグラウンドのクエリ処理性能に悪影響を与えない非同期でのバックアップやクラッシュリカバリなどが実現できる．

3.2 Neon

Neon [7] は，Neon, Inc. を中心としたコミュニティが開発し，マネージドサービスとしての提供や OSS としての公開を行っている CS 分離 DB の実装のひとつである．Neon もコンピュート層については，Aurora とほぼ同様の設計となっており，従来のリレーショナル DB のコンポーネントを基本的にそのまま活用した Compute というコンポーネントを持つ．ストレージ層については，Neon では Aurora と異なり分散ストレージを前提としていないため，代わりに単一マシンにおいても耐障害性を担保するような工夫がされている．

Neon のストレージ層は主に，WAL Service，Pageserver，オブジェクトストレージという 3 つのコンポーネントで構成される．WAL Service は，Compute から受信したトランザクションログを冗長化し一時的に保存することで耐障害性を確保しつつ，Pageserver にトランザクションログを受け渡す．Pageserver は，WAL Service から受信したトランザクションログを用いてページデータを作成し，永続化に関する処理を行ったり，不要ファイルの削除処理などを行ったりする．オブジェクトストレージは，Pageserver から受信したデータを永続的に保存する．3 つのコンポーネントのうち WAL Service と Pageserver は，キャッシュデータや中間データを保持することはあるものの，データの長期的な保持は行っていない．そのため，オブジェクトストレージに保存されたデータのみが，最終的に永続化されたデータという扱いになる．これによって Aurora と同様に，負荷分散やスループットの向上，耐障害性の向上などを実現している．

4 提　　案

本節では，提案手法の概要と実現方法の一例について述べる．

4.1 提 案 手 法

本研究では，複数のサーバやクラウドと CS 分離 DB を用いることで，スタンバイ側マシン上で起動する DB コンポーネントを最小限に抑える，ミニマムホットスタンバイ構成を提案する．

前節で述べた通り，CS 分離 DB は，コンピュート層とストレージ層を別々にスケール可能な柔軟性に優れたアーキテクチャであるという特徴がある．特に，コンピュート層は永続データを持たないことから，迅速な起動やスケールアウトが可能である．また，ストレージ層についても，永続化処理済みのデータを保存するコンポーネント以外についてはダウンしても問題ないようなアーキテクチャになっていることから，柔軟に起動や停止ができる．

このような CS 分離 DB の柔軟性を活かし，DB の部分的な停止を実現することで，可用性とリソース消費量や金銭的コストの極端なトレードオフを解消する．極端なトレードオフを改善することで，コールドスタンバイに近い金銭的コストで，ホットスタンバイのようにクラウド全体に影響するような大規模障害にも耐えうる高可用構成を実現できるようにする．

4.2 実　現　例

ミニマムホットスタンバイ構成は，CS 分離 DB のストレージ層のうち，どのコンポーネントをスタンバイ側マシン上で動作させるかによって，実現のパターンが 3 つある．ここでは，3.2 項で説明した，OSS として公開されている CS 分離 DB のひとつである，Neon を例に説明する．

1 つ目はオブジェクトストレージのみスタンバイで動作させるパターン，2 つ目はオブジェクトストレージと Pageserver を動作させるパターン，3 つ目は WAL Service まで含めたストレージ層のコンポーネント全てを動作させるパターンである．3 つのパターンにおいては，起動するコンポーネントが少ないほどリソース消費量が削減され、金銭的コストを抑える効果が高まると期待できるが，一方でフェイルオーバ時に起動するコンポーネント数が増加し起動処理に時間がかかるため，可用性は低下していくと予測できる．今回は，上記の 3 パターンのうち，金銭的コストを抑える効果と可用性の向上効果が中間であると予想した 2 つ目のパターンである，オブジェクトストレージと Pageserver を動作させるパターンをプロトタイプ実装した．以降では，プロトタイプ実装を行った，2 つ目のパターンを基本に説明する．

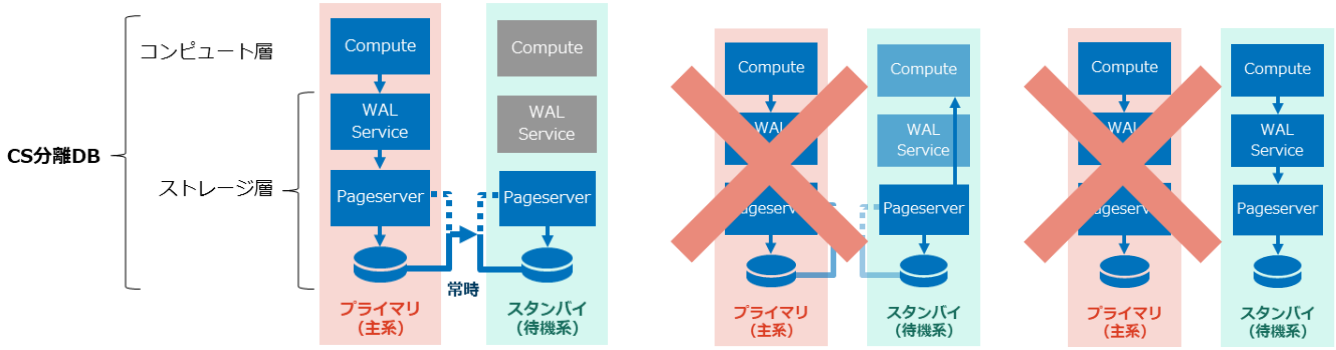
オブジェクトストレージと Pageserver を起動する場合の，ミニマムホットスタンバイ構成の実現イメージ例を図 1 に示す．平常時は，プライマリとスタンバイで，CS 分離 DB を起動する．この際，スタンバイで起動する CS 分離 DB については，先に述べた通り，コンピュート層のコンポーネントとストレージ層の WAL Service は起動せず，オブジェクトストレージと Pageserver を起動する（図 1a）．プライマリで障害が発生した際は，スタンバイで平常時に動作していなかった，コンピュート層のコンポーネントとストレージ層のコンポーネントを，追加で起動する（図 1b）．コンポーネントの起動や，バックアップデータからのリストアを迅速に行うことで，フェイルオーバを迅速に完了し，可用性を担保する（図 1c）．

5 評 価 実 験

本研究では，ミニマムホットスタンバイ構成によって，可用性とリソース消費量や金銭的コストの極端なトレードオフを改善することを目的としている．既存の冗長構成とミニマムホットスタンバイ構成を比較し，有効性を評価する実験として，可用性に関わる時間とリソース消費に伴う平常時のランニングコストの計測を行った．

5.1 実験対象と環境

実験対象には，本研究での提案内容であるミニマムホットス



(a) 平常時
（DB コンポーネントの部分的な停止）

(b) 障害発生時
（DB コンポーネントの追加起動）

(c) フェイルオーバ完了時

図 1: 提案手法の実現イメージ

表 1: 実験環境

DB	Neon (image tag 4549)
クラウド	Google Cloud
VM インスタンス	e2-highmem-4 仮想 CPU 4 メモリ 32GB
ディスク	標準 Persistent Disk 800GB

タンバイ構成と，比較対象として既存の冗長構成であるホットスタンバイ構成，コールドスタンバイ構成を用いた．いずれの冗長構成も，前節での説明にも用いた，オープンソースの CS 分離 DB である Neon を用いて構築した．実験環境については，今回は複数サーバや複数クラウドではなく，単一クラウド上の単一の仮想マシン（以降，VM）上で実施した．詳細な実験環境を表 1 に示す．VM 上にプライマリ側 DB とスタンバイ側 DB を構築し，TPC-B ベンチマーク（pgbench）の初期データをデフォルトのデータサイズで投入したのち，プライマリ側 DB には 20 分間 pgbench による負荷を掛ける．本実験においては，この 20 分間を障害が発生せずにシステムが通常通り稼働している，平常時と定義する．平常時には，ホットスタンバイでは Pgpool-II のネイティブレプリケーションモードを利用し，pgbench によって実行されるクエリを複製することによって，プライマリ側 DB とスタンバイ側 DB に同じデータを格納することで，データを冗長化する．ミニマムホットスタンバイでは，プライマリ側 DB にて永続化されたデータを，スタンバイ側 DB のオブジェクトストレージに常時転送することで，データを冗長化する．コールドスタンバイでは，プライマリ側 DB にて永続化されたデータを，平常時が 10 分経過した時点でスタンバイ側 DB のオブジェクトストレージに転送することで，疑似的に一定時間ごとのバックアップを再現しデータを冗長化する．20 分経過後，プライマリ側 DB を手動で停止し，スタンバイ側 DB へのフェイルオーバを実施する．

5.2 評 価 観 点

可用性に関わる時間として，本研究では，(1) フェイルオーバ時間，(2) 性能復旧時間，(3) データロスト時間という 3 つの時間的指標を定義した．(1) フェイルオーバ時間は，フェイルオーバの開始から完了までの時間を表す．今回の実験においては，スタンバイ側 DB の起動コマンドを入力した時点から，スタンバイ側 DB にクライアントが接続できるようになるまでの時間を計測した．(2) 性能復旧時間は，フェイルオーバ完了時点から，スタンバイ側の DB で一定の性能を発揮できるまでの時間を表す．今回の実験においては，一定の性能を「プライマリ側の DB で平常時に計測した TPS の 90%」と定義し，(1) フェイルオーバ時間の直後からその性能値がスタンバイ側の DB で発揮されるまでの時間を計測した．(1) と (2) の合計時間が，一般に RTO として用いられる．(3) データロスト時間は，クライアントが最後にプライマリ側の DB で更新したデータと，スタンバイでフェイルオーバ後に参照できるデータの時間差を表す．換言すると，フェイルオーバ後にどのくらい前のデータを参照可能かという時間であり，一般に RPO として用いられる指標である．(3) データロスト時間については，DB を運用する際のパラメータ設定に大きく依存するため，今回は机上での比較評価を行った．

ランニングコストの算出に用いるリソース消費量については，主にパブリッククラウドでの従量制課金の課金額を決める要因となる，CPU，メモリ，ストレージの使用量を対象とする．CPU やメモリの使用量については，pgbench デフォルトのデータサイズにおける平常時 20 分間の CPU やメモリの使用量の最高値を計測し，その最高値を基にリソースを 24 時間確保した場合の使用量を計算した．ストレージの使用量については，pgbench のデフォルトのデータ投入と平常時 20 分間のベンチマークによって約 7GB のデータが発生することが計測で明らかとなったため，今回は 7GB として計算した．これらリソース消費量の計算結果と，Google Cloud が公表しているインスタンスやストレージの料金表 [1]を使用して，使用量に応

1：今回は lowa(us-central1) の料金表を基に算出．https://cloud.google.

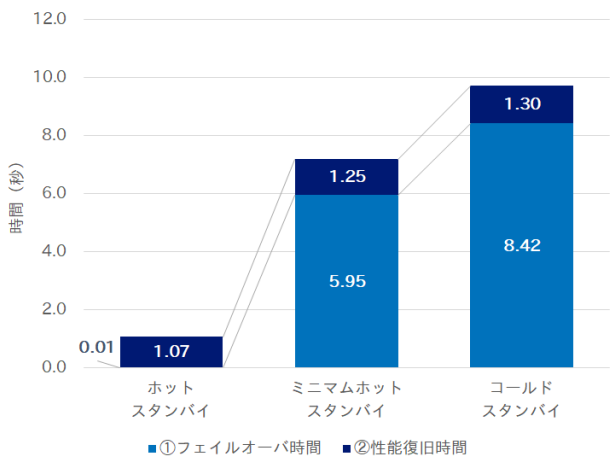


図 2: (1) フェイルオーバ時間と (2) 性能復旧時間の測定結果

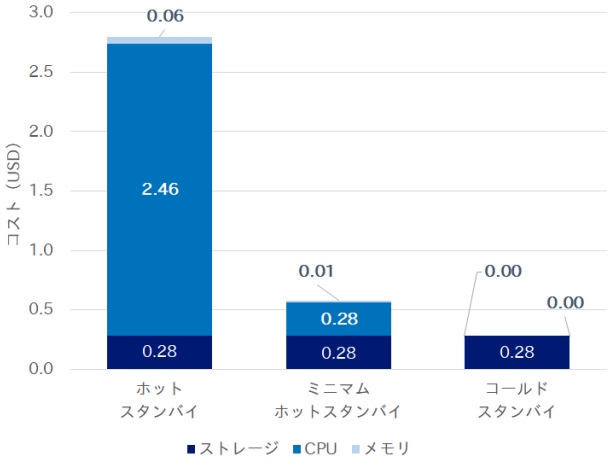


図 3: ランニングコストの計算結果
（7GB，24 時間の場合）

じたランニングコストを計算した．

5.3 実験結果

(1) フェイルオーバ時間と (2) 性能復旧時間について，各構成にて 15 回ずつ測定した際の平均値を実験結果として図 2 に示す．(1) フェイルオーバ時間については，ミニマムホットスタンバイがホットスタンバイより 5.93 秒長く，コールドスタンバイより 2.47 秒短い結果となった．(2) 性能復旧時間については，各構成で大きな差はないものの，ミニマムホットスタンバイがホットスタンバイより 0.18 秒長く，コールドスタンバイより 0.05 秒短い結果となった．

(3) データロスト時間については，前節で述べた通り，机上での比較を行う．ホットスタンバイでは，プライマリ側 DB とスタンバイ側 DB に対して，同時に同じデータを書き込むため，理論上のデータロスト時間は 0 である．ミニマムホットスタンバイでは，今回の設定値と環境で実施した場合のデータロスト時間が最大 10 分程度である．これは，プライマリ側 DB 内のストレージ層で行うデータ永続化処理が，デフォルトの設定値では最大 10 分間隔で行われるためである．今回使用しているミニマムホットスタンバイ構成のプロトタイプ実装では，永続化処理が完了しオブジェクトストレージに保存されたデータを，プライマリ側 DB からスタンバイ側 DB に適宜同期している．その結果，スタンバイで参照できるデータはプライマリ側 DB で永続化処理が完了したデータのみとなり，時間もプライマリ側 DB の設定に依存しそのまま 10 分程度であることがわかる．一方で，コールドスタンバイ構成については，ユースケースによってバックアップ間隔を自由に決定するため，データロスト時間はユースケースによって大きく異なる．一般的には，夜間などシステムの負荷が低い時間帯に日次バックアップを行うため，バックアップ間隔が 24 時間となる．その場合，データロスト時間は最大 24 時間程度ということになる．まとめると，ホットスタンバイではデータロスト時間が 0，ミニマムホットスタンバイでは 0〜10 分程度，コールドスタンバイでは 0〜24 時間程度と見積るため，ホットスタンバイ，ミニマムホットスタンバイ，コールドスタンバイの順で平均してデータロスト時間が長くなると推測できる．

ランニングコストについて，ホットスタンバイ，ミニマムホットスタンバイ，コールドスタンバイを 24 時間稼働させた場合の，スタンバイ側 DB にかかるランニングコストを図 3 に示す．7GB のデータサイズの場合，ミニマムホットスタンバイがホットスタンバイと比べて 2.22 USD の削減，コールドスタンバイと比べて 0.30 USD の増加であった．ミニマムホットスタンバイのランニングコストを基準とした比率で述べると，ホットスタンバイのランニングコストは 4.85 倍，コールドスタンバイのランニングコストは 0.48 倍である．

5.4 ミニマムホットスタンバイとホットスタンバイの比較による考察

ミニマムホットスタンバイとホットスタンバイを可用性の観点で比較すると，ミニマムホットスタンバイはホットスタンバイよりも (1) フェイルオーバ時間の平均が 5.93 秒長く，(2) 性能復旧時間の平均が 0.18 秒長い．(1) フェイルオーバ時間と (2) 性能復旧時間を合計し，比率で表すと，ミニマムホットスタンバイはホットスタンバイの 6.67 倍の時間がかかる．(3) データロスト時間についても，ホットスタンバイは 0 に近い RPO を達成でき，ミニマムホットスタンバイは 10 分程度のデータロスト時間が発生することから，ミニマムホットスタンバイのほうが可用性の観点では劣位な結果となった．(1) フェイルオーバ時間については，スタンバイ側 DB コンポーネントの起動処理やリストア処理が完了するまで DB への接続が待たされることが原因で，フェイルオーバの完了までに時間がかかっていると推測する．そのため，スタンバイ側 DB コンポーネントの起動処理やリストア処理において，ボトルネックとなっている箇所を今後調査する必要がある．(2) 性能復旧時間については，今回の実験で性能復旧の判断基準として TPS を用いるために 1 秒間のベンチマーク実行をしているため，1 秒台の性能復旧時

com/compute/vm-instance-pricing?hl=ja, https://cloud.google.com/compute/disks-image-pricing?hl=ja

間を達成しているということは，ほとんどの場合において初回のベンチマークで性能復旧を確認できたことを意味する．その中でもわずかに平均の秒数に差が生じた要因として，平常時に Compute や WAL Service が稼働状態にあるかという違いが影響している可能性がある．ホットスタンバイでは，平常時にスタンバイ側 DB で起動済みの Compute や WAL Service といったコンポーネントがキャッシュを保持していることで，より早い段階で高い TPS を発揮できる状態に復旧し，ミニマムホットスタンバイはスタンバイ側でキャッシュを保持していないことで，性能復旧が遅れたのではないかと推測する．(3) データロスト時間に関しては，前節で述べた通り，ミニマムホットスタンバイではプライマリ側 DB で行われる永続化処理間隔に依存したデータロスト時間が発生する．この問題によって，ミニマムホットスタンバイは，ホットスタンバイよりは (3) データロスト時間が長い傾向にあると予想できる．

ランニングコストの観点では，ミニマムホットスタンバイはホットスタンバイの約 5 分の 1 に抑えることができ，優位な結果であるといえる．ランニングコストの内訳に着目すると，ミニマムホットスタンバイとホットスタンバイは，CPU 使用率に伴うコストに大きな差がある．ミニマムホットスタンバイにおいてランニングコストを削減できた要因として，DB の一部コンポーネントをスタンバイ側で停止するという特徴によって，スタンバイ側で平常時に利用するリソースのうち特に CPU 使用率を下げることが可能であったことが影響していると考察する．以上のことから，本研究で提案したミニマムホットスタンバイは，リソース消費量や金銭的コストを抑える効果があることが分かった．

5.5 ミニマムホットスタンバイとコールドスタンバイの比較による考察

ミニマムホットスタンバイとコールドスタンバイを比較すると，ミニマムホットスタンバイのランニングコストはコールドスタンバイの約 2 倍である．今回の平常時が 24 時間，データサイズ 7GB という試算条件においては，CPU 使用率に伴うコストがストレージ使用に伴うコストとほぼ同等であるため，このような結果となった．平常時がさらに長くなれば，CPU 使用に伴うコストがさらに増加し，ミニマムホットスタンバイがコールドスタンバイに比べて不利な状況となる．一方で，データサイズが増加すれば，ストレージの使用料金が CPU 使用に伴うコストを上回り，ストレージの使用料金がランニングコスト全体に対して支配的になる可能性もあり，その場合はミニマムホットスタンバイに有利な状況となり得る．本研究では，サーバやクラウド規模の障害への対応も視野に入れていることから，障害発生時間に比べて平常時が極端に長くなるようなケースも想定すべきであると考えている．そのため，今後は，より大きなデータサイズに対してもミニマムホットスタンバイの平常時の CPU 使用率を現状と同程度かそれ以下に抑えることで，ランニングコストをコールドスタンバイに近づけていくことを目指す必要がある．

可用性の観点では，ミニマムホットスタンバイの (1) フェイルオーバ時間の平均はコールドスタンバイより 2.47 秒短く，(3) データロスト時間についてもコールドスタンバイに比べミニマムホットスタンバイのほうが大幅に短縮できる見込みがある．この実験結果には，ミニマムホットスタンバイの特徴が (1) フェイルオーバ時間 (3) データロスト時間に対して有効に働いた 2 つの理由があると考える．ミニマムホットスタンバイは，平常時もスタンバイ側 DB の一部が起動していることによって，ランニングコストはコールドスタンバイより増加してしまうものの，フェイルオーバ時に新たに起動する必要のあるコンポーネントが減り，(1) フェイルオーバ時間の短縮に繋がった．さらに，データ転送頻度がコールドスタンバイに比べて高く，常時転送していることによって，(3) データロスト時間を抑えることができる．(2) 性能復旧時間に関しては，ミニマムホットスタンバイもコールドスタンバイもほぼ差がなく 1 秒台の結果であったため，数回の TPS 計測で 90%の性能を発揮できたということを意味する．ミニマムホットスタンバイは，コールドスタンバイと異なり平常時に Pageserver が起動しているため，Pageserver 上の一時データを活用することで (2) 性能復旧時間の面では有利になるはずであるが，実際にはそのような結果にならなかった．その原因は，実験で用いたデータサイズが小さかったことによって，コールドスタンバイ側のリストア処理やキャッシュデータの作成が短時間で完了したためであると考える．そのため，よりデータサイズを大きくして実験することで，リストア時間やキャッシュデータの作成処理時間に差が生じ，ミニマムホットスタンバイとコールドスタンバイの差をより明らかにできると推察する．以上より，ミニマムホットスタンバイはコールドスタンバイと比べると，ランニングコストが増加した分，可用性に関わる時間的指標のうち 2 つを向上させることができる手法であるということが分かった．

5.6 実験結果と考察内容のまとめ

まとめると，ミニマムホットスタンバイは，ホットスタンバイよりもリソース消費量や金銭的コストを抑えつつ，コールドスタンバイよりも高い可用性を実現できる手法であることが明らかになった．しかし，ホットスタンバイの約 5 分の 1 のランニングコストで，約 6.7 倍の (1) フェイルオーバ時間 (2) 性能復旧時間となることから，トレードオフの改善という観点では，さらに可用性を向上させるか，金銭的コストを抑える必要がある．

改善の余地として，(1) フェイルオーバ時間と (2) 性能復旧時間をホットスタンバイに近づけるため，DB の起動処理のボトルネックやキャッシュ機構の調査を行い，データをより早い段階で高速に読み書きできるよう工夫することを考えている．(3) データロスト時間については，今回の実験設定と環境においては 10 分間のデータロスト時間が発生したものの，より負荷の高い状況や大規模なデータサイズでは，データロスト時間が変動する可能性もある．今後は，まず，負荷状況やデータサイズを複数パターン用意して実測することで，データロスト時間の変動を明らかにする．さらに，プライマリ側 DB の性能への影響も考慮しつつ永続化処理の間隔を短く設定したり，永続化前

のキャッシュデータなどを定期的にプライマリ側 DB からスタンバイ側 DB に転送したりするといった改善を検討している．金銭的コストについては，今回の実験によって CPU 使用率の削減が有効であることが明らかになったため，ミニマムホットスタンバイの平常時のスタンバイ側 DB としては不要な処理が行われていないかを調査し，不要な処理があれば部分的に停止するなどの工夫を検討する．

6 関連技術と関連研究

本節では，本研究に関連する既存技術や先行研究について述べる．

6.1 パイロットライト

Amazon Web Services が，自社のクラウドサービスで提供している冗長化手法としてパイロットライトがある [8]．パイロットライトとは，システムを動作させているメインリージョンと別のリージョンをスタンバイのためのリージョンとし，スタンバイのリージョンでは 3.1 項で説明した Aurora という DB のみを動作させておく冗長化手法である．平常時は，メインリージョンの CS 分離 DB からスタンバイリージョンの CS 分離 DB に対してレプリケーションを行う．障害発生時には，スタンバイリージョンで DB 以外に必要なフロントエンドサーバやアプリケーションサーバを起動し，フェイルオーバを行う．

本研究と類似する点は，可用性向上のための構成の一種であり，CS 分離 DB を使用している点である．本研究との相違点は 2 つあり，1 点目は平常時に動作させる DB コンポーネントの範囲である．パイロットライトでは，CS 分離 DB である Aurora を用いているものの，スタンバイリージョンで DB コンポーネントの全てを動作させている．一方で，本研究で提案するミニマムホットスタンバイ構成では，スタンバイクラウド上で DB コンポーネントの一部のみを動作させる．よって，ミニマムホットスタンバイ構成は，パイロットライトよりもリソース消費量や金銭的コストを抑えることができる．2 点目は，スタンバイの配置場所とそれに伴って対策できる障害の規模の大きさである．パイロットライトは，自社が提供する単一クラウド基盤上の，別のリージョンにスタンバイを配置する．ミニマムホットスタンバイ構成では，別のクラウド基盤上にスタンバイを構築することが可能である．そのため，ミニマムホットスタンバイ構成であれば，複数のリージョンに影響するようなクラウド障害にも対応できるという利点がある．

6.2 クラウドサービスの可用性向上とコスト削減を実現するスケジューリング手法の提案

西田は，クラウドベンダーが遊休リソースを低価格 IT リソースとして提供していることに着目し，低価格 IT リソースと通常リソースに対してコンポーネントの配置を自動化するスケジューリング手法を提案した [9]．低価格 IT リソースは，利用することでクラウド運用にかかる金銭的コストを削減できる一方で，クラウド上の遊休リソースを基に提供されていることから，提供状況が不安定で可用性が低い．そのため，システムのコンポーネントを単に冗長化するだけでは，低価格 IT リソースのみに偏ってコンポーネントがスケジューリングされた際に可用性が低下してしまうという問題がある．そのような問題を解決するために，可用性やリソース消費の効率を考慮した上で，低価格 IT リソースと通常 IT リソース間にまたがってコンポーネントの配置を自動化するスケジューリングのプラグインを開発している．

本研究との類似点は，可用性の向上と金銭的コストの削減を目的とし，冗長化手法に関して研究しているという点である．本研究との相違点は，前提としているクラウドリソースのスペックである．西田のスケジューリング手法は，低価格 IT リソースと通常の IT リソースという，金銭的コストに差がある 2 種類のクラウドリソースがあることを前提にした手法である．そのため，低価格 IT リソースを提供しているパブリッククラウド上で構築するシステムの金銭的コスト削減に対して有効な手法である．一方で，ミニマムホットスタンバイ構成は，どのサーバやクラウド基盤上でも実行可能で，金銭的コスト削減のために DB の起動コンポーネント数を減らす手法をとっている．そのため，ミニマムホットスタンバイはより汎用性が高く，低価格 IT リソースがない場合でも金銭的コストの削減が可能である．

6.3 Disaster Recovery as a Service

DR ができる環境をサービスとして提供するための技術である Disaster Recovery as a Service に関しても，様々な先行研究がある．Saquib らの研究では，Internet Small Computer System Interface（iSCSI）プロトコルを使用し，トランザクションログファイル全体の書き込みを待たず，ログファイルに追加されたブロックを即座に DRaaS 環境に送信することで，データ損失を 0 に抑える手法を提案した [10]．Saquib らの提案する DRaaS 環境は，DB 以外に専用の管理ユニット，抽出・複製ユニット，データ収集・回復ユニットという 3 つのユニットから構成される．抽出・複製ユニットをユーザの DB がデプロイされている環境にインストールし，その他のユニットを DR 先であるクラウド環境上にインストールすることで，ユーザの要件に応じた高可用な DR 環境が構築できる．

本研究と類似している点は，DR を目的とした冗長化手段として，特に DB に着目して高可用を目指した構成を検討している点である．本研究との相違点は，DB 以外の専用ユニットが必須である代わりに，トランザクションログをブロック単位で送信し，高可用を達成している点である．Saquib らの研究では，限りなく 0 に近い RPO や RTO を実現するために，専用ユニットの導入を前提としている．ミニマムホットスタンバイの現状の実装では，永続化済みのデータをプライマリからスタンバイに転送しているため，転送に特別な技術や専用ユニットは不要である一方で，データ損失を抑えることが難しい．そのため，ミニマムホットスタンバイにおいては，汎用性や低リソース消費を保ちつつ，より高可用な構成を目指すために，トランザクションログの適宜送信を DB 内で完結する処理として実装することが可能か検討していく．

7 おわりに

従来の可用性向上手法では，IT システムの可用性を向上させるために冗長構成を構築すると，DB の部分的な起動や停止が実現できないために，コールドスタンバイとウォームスタンバイ，ホットスタンバイの間に極端なトレードオフが発生した．本研究では，DB の部分的な停止を実現するために，柔軟性とスケーラビリティに優れた CS 分離 DB に着目した．そして，複数のサーバやクラウドと CS 分離 DB を用いた，スタンバイのサーバやクラウドでは DB の中でも必要最低限のコンポーネントのみを起動する冗長構成であるミニマムホットスタンバイ構成を提案した．評価実験では，ミニマムホットスタンバイのプロトタイプ実装と，ホットスタンバイ，コールドスタンバイを比較した結果，ミニマムホットスタンバイがホットスタンバイより低い金銭的コストで，コールドスタンバイよりも高い可用性を実現できる手法であることを示した．

今後は，トレードオフのさらなる改善のため，まず 5.6 項のとおり課題が見えている (1) フェイルオーバ時間，(2) 性能復旧時間，(3) データロスト時間といった可用性観点での改善を目指す．それらの改善を行った後の評価実験では，実験環境，実験に用いるデータサイズ，比較対象など，様々な実験条件を変更した実験を行いたいと考えている．実験環境に関しては，今回の実験で使用していない他社のパブリッククラウドやプライベートクラウドを用いたマルチクラウド環境など，他の環境においても同様に提案手法の有効性を示したいと考えている．データサイズに関しては，5.5 項で述べた通り，主にミニマムホットスタンバイとコールドスタンバイの比較において，コストや性能復旧時間の点で影響を与える可能性があるため，より大きなデータサイズで実験することで，2 つの構成の差異をより明確化したい．比較対象に関しては，今回の実験では比較対象としてホットスタンバイとコールドスタンバイを使用したが，他にも比較対象としてはウォームスタンバイや CS 分離 DB を使用しない冗長構成など様々なものが候補となり得る．そのため，今後の実験では，比較対象を追加し他の手法と比較した際のミニマムホットスタンバイ構成の優位性をより明らかにしていく．

文　献

[1] Wade Tyler Millward. The 10 Biggest Cloud Outages Of 2022 (So Far), Jul 2022. https://www.crn.com/news/cloud/the-10-biggest-cloud-outages-of-2022-so-far-（参照 2024-03-21）.

[2] Microsoft. Azure status history | Microsoft Azure, 2023. https://azure.status.microsoft/en-us/status/history/（参照 2024-03-21）.

[3] ITmedia 新野淳一. Microsoft、Azure や Microsoft 365 などで起きた大規模障害の原因報告　WAN 内の全ルータが再計算状態に突入し、パケット転送が不可に- ITmedia NEWS, Feb 2023. https://www.itmedia.co.jp/news/articles/2302/02/news094.html（参照 2024-03-21）.

[4] GMO ペパボ株式会社. オンプレ k8s と EKS の並行運用の実際, Jan 2023. https://speakerdeck.com/ch1aki/onpurek8stoeksnobing-xing-yun-yong-noshi-ji（参照 2024-03-21）.

[5] 山口学. 事例で見るハイブリッドクラウド実践のいま：ますます加速するクラウドファースト- @ IT, Sep 2015. https://atmarkit.itmedia.co.jp/ait/articles/1508/10/news008.html（参照 2024-03-21）.

[6] Alexandre Verbitski, Anurag Gupta, Debanjan Saha, Murali Brahmadesam, Kamal Gupta, Raman Mittal, Sailesh Krishnamurthy, Sandor Maurice, Tengiz Kharatishvili, and Xiaofeng Bao. Amazon Aurora: Design Considerations for High Throughput Cloud-Native Relational Databases. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 1041–1052. ACM, May 2017.

[7] Neon, Inc. Neon — Serverless, Fault-Tolerant, Branchable Postgres, 2022-2024. https://neon.tech/ （参照 2024-03-21）.

[8] Amazon Web Services, Inc. AWS のディザスタリカバリ (DR) アーキテクチャ、パート III: パイロットライトとウォームスタンバイ | Amazon Web Services ブログ, Dec 2021. https://aws.amazon.com/jp/blogs/news/disaster-recovery-dr-architecture-on-aws-part-iii-pilot-light-and-warm-standby/（参照 2024-03-21）.

[9] 西田寿雄. クラウドサービスの可用性向上とコスト削減を実現するスケジューリング手法の提案. 研究報告インターネットと運用技術（IOT）, Vol. 2023-IOT-61, No. 9, p. 1–6, May 2023.

[10] Zia Saquib, Veena Tyagi, Shreya Bokare, Shivraj Dongawe, Monika Dwivedi, and Jayati Dwivedi. A new approach to disaster recovery as a service over cloud for database system. In *2013 15th International Conference on Advanced Computing Technologies (ICACT)*, p. 1–6, Sep 2013.

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 10:00 ~ 12:10 | T2-B オンライン（Zoom Events)

グラフデータ処理

座長:新井 淳也(日本電信電話株式会社)
コメンテータ:伏見 卓恭(東京工科大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T2-B-4-01] グラフ要約に基づく知識グラフ索引の提案

*小倉 勇大[1]、天笠 俊之[2] (1. 筑波大学大学院　情報理工学位プログラム　知識・データ工学研究室、2. 筑波大学計算科学研究センター)

10:20 ~ 10:45
[T2-B-4-02] グラフ深層学習を用いた類似特許検索の精度向上

*中山 樹[1]、中西 宏和[1]、菊地 良将[1]、佐々木 勇和[1]、荒瀬 由紀[1]、鬼塚 真[1] (1. 大阪大学情報科学研究科鬼塚研究室)

10:45 ~ 11:10
[T2-B-4-03] グラフ要約を用いた高速なグラフ相関問合せ

*直井 悠馬[1]、塩川 浩昭[2]、真次 彰平[2] (1. 筑波大学　理工情報生命学術院　知識・データ工学研究室、2. 筑波大学計算科学研究センター)

11:10 ~ 11:35
[T2-B-4-04] ラベル付き有向グラフにおける node embedding を用いたスキーマ抽出

*木下 一歩[1]、鈴木 伸崇[1] (1. 筑波大学)

11:35 ~ 12:00
[T2-B-4-05] プリミティブなグラフ分析アルゴリズムに対する効率的なグラフ要約の応用

*真次 彰平[1]、塩川 浩昭[1] (1. 筑波大学計算科学研究センター)

グラフ要約に基づく知識グラフ索引の提案

小倉　勇大†　　天笠　俊之††

† 筑波大学大学院　〒 305-8577 つくば市天王台 1 丁目 1-1
†† 計算科学研究センター　〒 305-8577 つくば市天王台 1 丁目 1-1
E-mail: †ogura@kde.cs.tsukuba.ac.jp, ††amagasa@cs.tsukuba.ac.jp

あらまし　セマンティックウェブの普及と発展に伴い、Web 上で公開される RDF データは増加し続けており、現在、Web 上には膨大な量の RDF データが存在している。これにより、膨大な量の RDF データに対する効率的なデータ検索の必要性が高まっている。さらに、検索時のクエリの作成には、RDF データに対する理解が必要であるが、RDF データはリレーショナルデータベースとは異なり、構造を決定するスキーマを持たないため、データセットに対する理解は容易ではない。そこで、本研究では、大規模な知識グラフを複数の部分グラフに分割し、それぞれを要約し、さらに、これらの部分グラフに対し、要約に基づく索引を付与する手法を提案する。本研究では、データセットに対する理解の促進と、膨大な RDF データに対する効率的な検索、さらには、さまざまな RDF データセットを統合した Linked Open Data（LOD）全体に対する効率的な検索手法の開発を目指す。

キーワード　知識グラフ，RDF データ

1　序　　論

セマンティックウェブ [1,2] の普及と発展に伴い、Web 上で公開される RDF（Resource Description Framework）データ [3] は増加し続けており、現在、Web 上には膨大な量の RDF データが存在している [4]。膨大な RDF データの存在は、特にデータの連携と統合において重要な役割を果たす。異なるソースからのデータを統合することにより、複数のデータベースやアプリケーション間での情報共有が容易になり、より充実したデータ分析や洞察が可能となる。

しかし、RDF データはリレーショナルデータベースとは異なり、構造を決定するスキーマを持たないため、データに対する理解や取得は容易ではない [5]。このことから RDF データに対する効率的なデータ検索の必要性が高まっている [6]。そこで、膨大な量の RDF データに対する理解の促進を促進するための要約手法や効率的なデータ検索のための手法が提案されている [7–9]。

本研究では、RDF データに対する効率的なデータ検索を行うことを目的として、大規模な知識グラフを中心性に基づくスーパーノードの作成によって要約し、SPARQL クエリと要約とのマッチング推定を行い、求めるデータがデータセットに存在するかを効率的に判定する手法を提案する。

2　前 提 知 識

2.1　RDF（Resource Description Framework）

RDF (Resource Description Framework) [3] は、セマンティック Web 技術の一つとして Linked Data やメタデータの記述に用いられるデータ形式である。RDF は URI(Uniform Resource Identifier) [10] で名付けられたリソースの関係を記述（モデル化）する枠組みを与え、RDF データによって Web 上の情報リソースに対して計算機が解読可能なメタデータを提供する。RDF は、それらのリソース間の関係を主語（Subject）、述語（Predicate）と目的語（Object）のを用いた ($\langle s\rangle$ $\langle p\rangle$ $\langle o\rangle$) という形式の三つ組（RDF トリプル）によって関係付ける。述語は URI によってのみ表現されるが、主語と目的語は他の表現も用いることができる。ブランクノードはグローバルな識別子を持たない資源の表現に使用されるものであり主語と目的語として使用可能である。リテラルは、文字列や日付、数などをそのままデータとして扱うものであり、目的語としてのみ使用可能である。

また、RDF トリプルは述語を有向エッジとして始点に主語、終点に目的語を配置した表現で視覚化できる。RDF グラフは複数の基本パターンのトリプルを組み合わせた連結グラフによって構成される。図 1 はその一例である。

2.2　SPARQL（SPARQL Protocol and RDF Query Language）

SPARQL（SPARQL Protocol and RDF Query Language）[11] は、RDF（Resource Description Framework）形式のデータを検索・操作するための標準的なクエリ言語である。SPARQL は、ウェブデータのグラフ構造に適合した形で設計されており、グラフパターンマッチングを利用してデータを問い合わせる。

SPARQL クエリの基本構造は、リレーショナルデータベース用の標準クエリ言語である SQL に似ており、主に「SELECT 句」と「WHERE 句」で構成される。SELECT 句では、問い合わせの結果として取得したいデータの変数を指定する。一方、WHERE 句では、データベース内でマッチさせたいパターンを指定する。これらのパターンは、主に「？」で始まる変数として記述されたトリプル（主語、述語、目的語の組み合わせ）によって表現する。

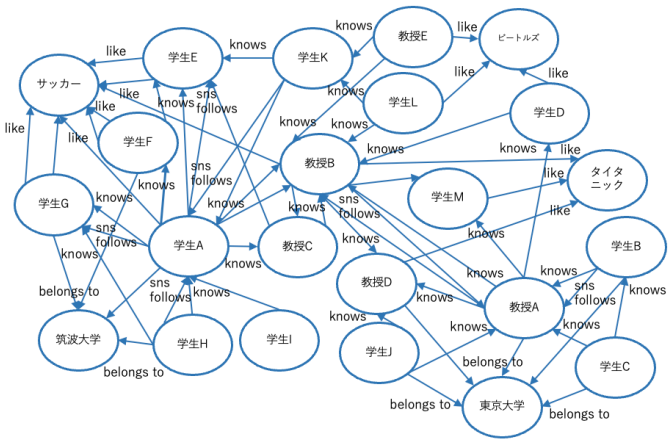


図 1　RDF トリプルのグラフ表現の例

この言語の特徴は、RDF 形式のデータに対する柔軟なアクセスを可能にすることにあり、RDF は、ウェブ上の情報をより豊かに記述するためのデータモデルであり、さまざまな種類のデータソースを統合し、リンクすることができる。SPARQL を使用することで、これらの複雑なデータセットから必要な情報を効抽出することが可能になる。

3　関 連 研 究

膨大な量が存在し、構造を決定する特定のスキーマを持たない RDF データの迅速な理解と探索を促進することを目的としてさまざまな手法やツールが提案されてきた。RDF データに対する理解を促進するための要約手法として、Troullinou らの手法 [8] が挙げられる。Troullinou らは、最も重要なノードを強調した要約を作成することを目的とし、静的な要約を超えて RDF データを動的に探索する新しい手法を提案した。Troullinou らは、最も重要なノードを特定するために、度数中心性、ブリッジ中心性、調和中心性、放射性、エゴ中心性、媒介中心性、PageRank、HITS のような様々な中心性指標を使用した。このアプローチは、「ズーム」と「エクステンド」という二つの操作を用いて、要約から詳細な情報へのアクセスを可能にする。「エクステンド」は特定のサブグラフに、「ズーム」は全グラフに焦点を当て、それぞれがユーザーに細かい情報へのアクセスを提供する。これらの操作の計算は NP 完全であり、これらの計算のための近似アルゴリズムを提案し、その効率性を実証した。しかし、Troullinou らの手法では、視覚化による RDF データの理解に重点を置いており、クエリによる効率的な探索には触れられていない。また、Park ら [12] は、近年の Deep Learning の発展に伴い、グラフニューラルネットワーク（GNN）による知識グラフの重要ノードの特定する手法、GENI [12] を提案し、重要ノードの特定において、最先端の精度を達成している。また、複数データセットを横断した検索を効率的に行うための手法も提案されている。Muhammad ら [13] は効率的なデータセット選択を行う、トリプルパターンごとのソース選択を超える、連合クエリの重複データの影響を考慮した HiBISCuS、DAW、TBSS という新しいソース選択手法を提案した。また、これら 3 つのソース選択アプローチをクエリ書き換えと組み合わせて、Quetsal という完全な SPARQL クエリ連合エンジンを作成した。これらの手法は、既存の SPARQL クエリ連合エンジンと組み合わせて利用され、データソースの数を減少させながらもクエリのリコールを保持する。

4　提 案 手 法

本研究では、膨大な量の RDF データに対するデータ検索の効率化と、データセットを横断した検索の効率化開発を目的として、グラフ要約に基づく知識グラフへ索引付けを提案する。本手法は、大きく分けて、事前処理とマッチング推定の処理に分けられる。事前処理では、注目ノードの抽出（小節 4.1.1）、重要ノードを中心とするスーパーグラフの作成（小節 4.1.2）、スーパーグラフの索引化（小節 4.1.3）、SPARQL クエリに対するマッチング推定（節 4.2）によって構成される。提案手法の概観を事前処理とマッチング推定に分けてそれぞれ、図 2、図 3 に示す。

4.1　事 前 処 理

事前処理では大きく分けて三段階の処理を行う。一段階目は注目ノードの抽出（小節 4.1.1）、重要ノードを中心とするスーパーグラフの作成（小節 4.1.2）、三段階目はスーパーグラフの索引化（小節 4.1.3）、である。以降でそれぞれについて説明を行う。

4.1.1　注目ノードの抽出

まず、事前処理の第一段階として、注目ノードの抽出を行う。RDF グラフのノードやエッジにはそれぞれさまざまな種類が存在する。その中には、グラフとして表現することが適切でないノードやエッジも存在する。そのようなノードやエッジを除き、ユーザが注目したいノードとそのエッジのみを抽出し、グラフとして表現することが適切なノードやエッジのみで構成されたベースグラフを作成する。

4.1.2　スーパーグラフの作成

次に、事前処理の第二段階として、スーパーグラフの作成を

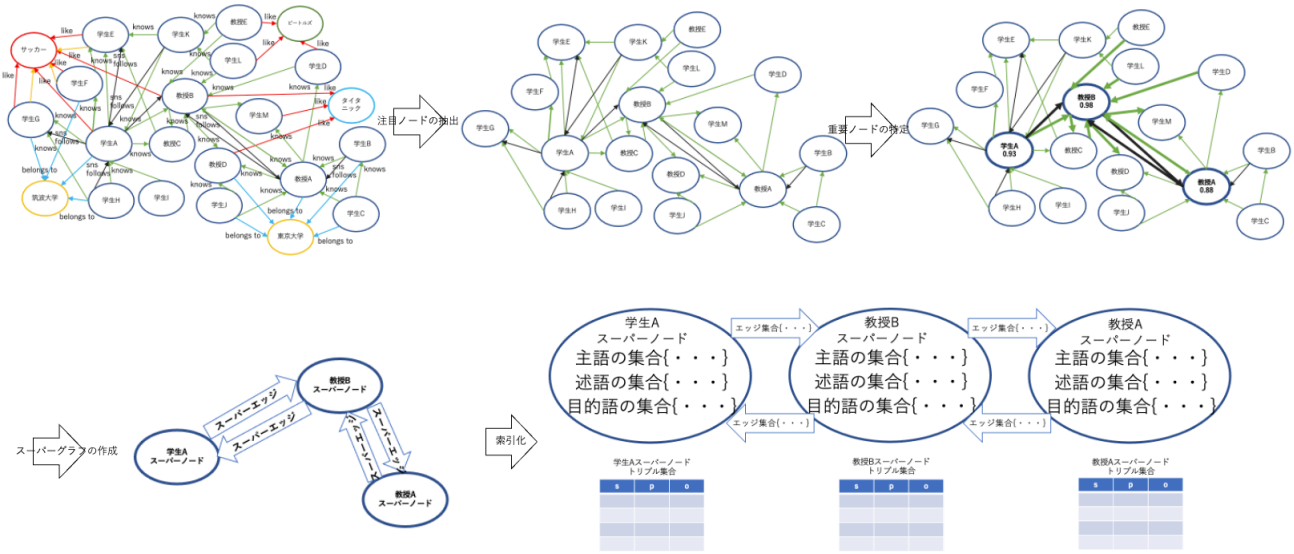


図 2　事前処理概観

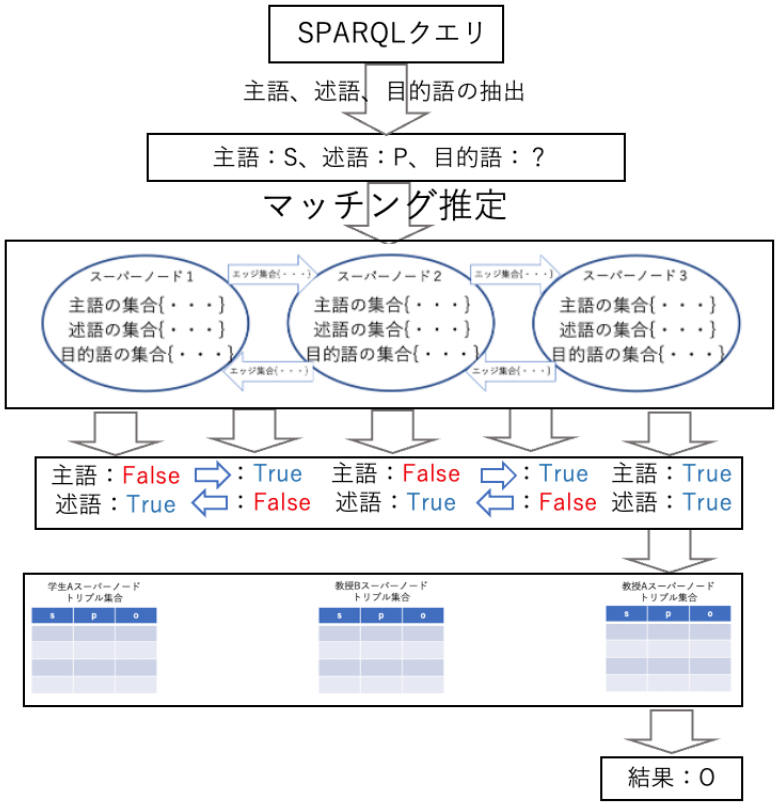


図 3　マッチング推定概観

行う。本手法では前小節 4.1.1 で作成したベースグラフのノードに対して次数中心性に基づく重要ノードの特定を行い。重要ノードから 1 ホップ以内のノードを特定した重要ノードのスーパーノードの要素として、取り込むことでスーパーノードを作成する。この操作を繰り返すことで、スーパーノードを複数作成する。そして、スーパーノード間に張られているエッジをスーパエッジとし、スーパーグラフを作成する。

4.1.3　索　引　化

最後に、事前処理の第三段階として、スーパーグラフの索引化を行う。前小節 4.1.2 で作成したスーパーグラフのスーパーノードとスーパーエッジそれぞれに保持する要素を説明する。まず、スーパーノードにはスーパーノードとして取り出した、トリプルの「主語の集合」、「述語の集合」、「目的語の集合」を保持する。また、スーパーエッジには、スーパーノード間に張られていた「述語の集合」を保持する。さらに、スーパーノー

ドのトリプル集合として、RDF データを保持する。

4.2 SPARQL クエリに対するマッチング推定

本節では、SPARQL クエリに対するマッチング推定の手順について説明を行う。まず、ある SPARQL クエリが投げられたとき、SPARQL クエリから、指定されている主語、述語、目的語を抽出する。そして、前節 4.1 で索引化したスーパーグラフのそれぞれのスーパーノードやスーパーエッジに指定されている主語、述語、目的語が存在するかを確認する。そして、存在が確認されたスーパーノードトリプル集合に検索をかけ、答えを得る。

5 実験計画

本章では実験計画について述べる。まずは実験の第一段階として、単一データセットでの実験を行う予定である。その際、ルールベースによる注目ノード抽出、次数中心性による重要ノード特定を行い、スーパーグラフを作成する。作成したスーパーグラフを索引とする、SPARQL クエリに対するマッチング推定の有効性を測るための実験を行う。次段階として、複数データセットでの実験を行う予定である。また、さらなる発展として、機械学習による注目ノード抽出、その他中心性や機械学習を用いた重要ノード特定を検討している。

6 まとめ

本稿では、膨大な量の RDF グラフを要約し、データアクセスを軽量化する手法の開発を目的として、グラフ要約に基づく知識グラフへ索引付け手法として、スーパーグラフを索引とする、複数データセットを横断した検索を行うための手法を提案した。今後は章 5 で説明した実験計画に基づき、実験を行う予定である。

7 謝辞

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の 「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業、JST CREST（JPMJCR22M2、科学研究費補助金（JP22H03694）に支援によるものです。

文献

[1] http://www.w3c.org/2001/sw/. “world wide web consortium: Semantic web”. 2001.
[2] T. Berners-Lee. “what the semantic web can represent”,http://www.w3.org/designissues/rdfnot.html. 1998.
[3] Lassil et al. “resource description framework (rdf) model and syntax specificatio”, https://www.w3.org/tr/1999/rec-rdf-syntax-19990222/. *W3C Recommendation*, 1999.
[4] Efthymiou V. Stefanidis K. Christophides, V. “entity resolution in the web of data. synthesis lectures on the semantic web: Theory and technology”. 2015.
[5] http://www.w3.org/TR/2000/CR-rdf-schema 20000327/. “world wide web consortium: Resource description framework (rdf) schema specification 1.0”. *W3C Candidate Recommendation 27 March 2000.*
[6] Abhilash Dighe Yike Liu, Tara Safavi and Danai Koutra. “graph summarization methods and applications: A survey”. *ACM CSUR*, 2018.
[7] Xin Luna Dong Qi Song, Yinghui Wu. “mining summaries for knowledge graph search”. *In ICDM*, 2016.
[8] Georgia Troullinou et al. “exploring rdfs kbs using summaries”. *In ESWC*, 2015.
[9] Lukas Faber Davide Mottin Emmanuel Muller Danai Koutra Tara Safavi, Caleb Belth. “personalized knowledge graph summarization: From the cloud to your pocket”. *In ICDM*, 2019.
[10] Berners-Le et al. “uniform resource identifier(uri): Generic syntax”, https://labs.apache.org/webarch/uri/rfc/rfc3986.html. 2005.
[11] EricPrud Steve Harris, Andy Seaborne. “hommeaux. sparql 1.1 query language”. *W3C recommendation, Vol. 21, No. 10, p. 778*, 2013.
[12] Namyong Park et al. “estimating node importance in knowledge graphs using graph neural networks”. *In KDD*, 2019.
[13] Muhammad Saleem. “efficient source selection and benchmarking for sparql endpoint query federation”. *2018. ISBN 978-3-89838-732-3.*

グラフ深層学習を用いた類似特許検索の精度向上

中山　樹†　菊地　良将†　中西　宏和†　佐々木　勇和†　荒瀬　由紀†
鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5
E-mail: †{nakayama.itsuki,kikuchi.ryosuke,nakanishi.hirokazu,sasaki,arase,onizuka}@ist.osaka-u.ac.jp

あらまし　特許検索は発明の新規性や特許侵害のリスクの確認，および技術動向の調査に必須の技術である．代表的な既存技術として，キーワード検索や文書の分散表現による検索がある．しかし，キーワード検索ではキーワードに類似した意味を持つが表層が異なる単語を見つけることができないため精度が低く，文書の分散表現を用いる場合でも，特許で使用される専門用語を表現するような埋め込みを得ることは難しい．すなわち，文章の情報だけでは限界がある．本研究では，メタデータを使用して文書の分散表現を改善し，特許検索の精度向上を目指す．提案手法ではメタデータとして，(1) 分類情報，(2) 特許間の引用関係 を使用する．具体的には，(1) 分類情報を基に二部グラフを作成し，GNN を適用することで，分類情報の影響を反映させ，(2)GNN の出力から引用関係の有無を MLP で判定し，その誤差をロスとすることで，GNN の出力を検索に有効な分散表現にする．実験では，IPC の大分類における 2 つのデータセットに対し，既存手法を上回る性能を達成した．

キーワード　特許検索，深層学習，グラフデータ処理，ノードエンベディング

1　はじめに

特許検索は非常に重要な技術であり様々な用途で使用される．その用途として，申請したい発明に新規性があるかどうかを調べる新規性調査 [1] や，どのような技術が存在しているかを調べる技術動向調査 [2]，自社の製品が他社の特許権を侵害していないかどうか調べる侵害調査 [3] などが挙げられる．

本研究は新規性調査に焦点を当てる．新規性の調査では，発明と類似した特許をいくつかリストアップし，1 つずつ確認していくという作業が行われる．この作業において，申請前に発明者が特許要件（新規性，進歩性）を否定する特許を見つけた場合，無駄な申請費を削減するため申請を断念する．また，申請後に特許審査官が前述のような特許を見つけた場合は拒絶理由として使用し，関連しているが拒絶理由とならない場合は参考文献として利用する．この新規性調査を行う際に必要となるのが類似特許検索である．類似特許検索の既存手法として，キーワード検索や文書の分散表現による検索が挙げられる．キーワード検索は J-PlatPat[1]がよく用いられ，文書の分散表現による検索は Sentence-BERT [4] や Universal Sentence Encoder [5] などが用いられる．しかし，キーワード検索および文書の分散表現による検索にはそれぞれ問題がある．まず，キーワード検索ではキーワードに類似した意味を持つが表層が異なる単語を見つけることができないため精度が低い次に，文書の分散表現に基づく検索では，意味が近い単語を含む特許を見つけることが可能ではあるが，特許文書には専門用語や複雑な表現が多いこともあり，一般ドメインの大規模コーパスで訓練した文書埋め込みモデルでは精度が低い．

特許データの特徴として，特許ごとに IPC と呼ばれる国際的に標準化された分類情報や先行技術調査を迅速に行うために開発された F ターム，特許の審査官による引用情報が付与されている．本研究ではこれらのメタデータを活用して，類似特許検索の精度向上を目指す．しかし，分類情報は一般的に多クラスの分類タスクで使用され，同様に引用関係はリンク予測の 2 値分類タスクで使用されるため，本研究の目的である類似特許検索の精度向上にどのようにこれらのメタデータを活用するべきかが技術課題として挙げられる．この技術的課題に対して，特許文書と分類情報からなる二部グラフに対し、引用関係を予測する学習を行うことで、文書分散表現に分類情報と引用関係を埋め込んだベクトル表現を獲得する．この方針を踏まえて提案手法を次の 4 つのステップで構成する．(1) 文書埋め込みモデルのファインチューニング，(2) 分類情報による二部グラフの構築および二部グラフと GNN による分類情報の反映，(3) ハードネガティブを活用したリンク予測タスクによる引用関係の反映，(4) モデルの中間層を用いた類似特許検索．

本論文の貢献点は以下の 3 つである．

（1） 分類情報および引用関係のメタデータを活用するという着想で，高精度な特許検索を実現した（Recall@100 において既存の検索手法に対して，10%向上）．

（2） 高精度な特許検索を実現するため，文書の分散表現を初期特徴量，分類情報を二部グラフ，引用関係をロス関数として用いることで，モデルの中間層にそれらの情報を含ませた．特に分類情報に関しては，粒度が詳細な場合でも有効な活用方法を考案した．

（3） 包括的な実験と ablation study により，実際の特許データにおいて提案手法が有効であることを示した．

1：https://www.j-platpat.inpit.go.jp/

2 事前準備

本研究で扱う特許データの例を図 1に示す．以下でその詳細について述べる．

2.1 特許文書の詳細

特許文書は構造化されており，主たる構造について説明する．

要約文 発明の概要を平易な文章で簡潔に記載したもの．

国際特許分類（IPC） 特許を体系的に分類するための国際的に定められた分類情報．A から H までのセクション（大分類）があり，さらに各セクションからさらにクラス，サブクラス，メイングループ，サブグループと細かく分類される．本研究では，大分類を使用して特許データを分割し，大分類ごとにモデルの学習と特許検索を行う．

F タ ー ム 日本の特許庁が作成している，特許の先行技術調査を行うための分類である．一般に，1 つの特許につき複数の F タームが割り当てられる．テーマコード 5 桁＋観点 2 桁＋数字 2 桁の計 9 桁で構成されるが，本研究では，9 桁全て使用する．

請求項 特許を受けようとする発明ごとに，請求項 1，請求項 2，... と分けて記述されたもの．発明が保護される範囲を特定するための事項を過不足なく記載する必要があるため，要約文よりも詳細に記述される．ただし，本研究では使用できていないため，図 1には記載していない．

2.2 特許文書間の関係

特許審査官が申請された発明に対し，他の特許を引用することがある．その引用には複数の種類が存在するが，以下に重要なものを取り上げる．

拒絶理由 申請された発明に対し，その発明が特許を受けることができない理由となる特許．その理由として，新規性がない，進歩性がない，などが挙げられる．

特許査定 特許審査官が調査の際に参考にした文献．

3 関連研究

3.1 特 許 検 索

特許検索は古くから研究が行われており，キーワードや引用関係，分類情報などを活用する様々な手法が提案されている．Konishi [6] はキーワードベースの手法を提案しており，請求項からクエリとなるキーワードを抽出し検索を行なっている．この手法では，分類情報や引用関係などのメタデータをできていない．Fujii [7] は引用関係と文章の両方を活用する手法を提案している．まず，過去に多く引用された特許は今後も多く引用されると仮定し，引用ネットワークに PageRank [8] を適用し

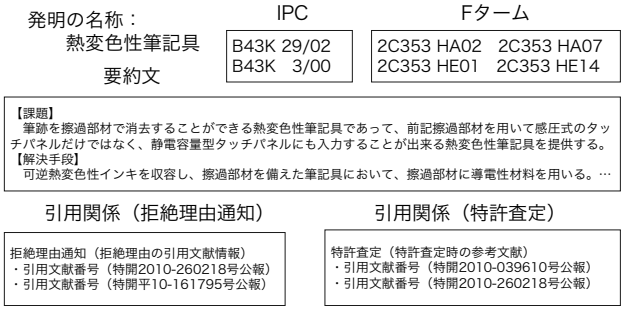


図 1: 特許データの例[2]．ただし，本研究で使用する項目のみを抜粋している．

て引用関係に基づくスコア $S_C(d)$ を計算する．次に，文章を単語に分割し，tf-idf を改良した Okapi BM25 [9] を用いて文章に基づくスコア $S_T(d)$ を計算する．そして，以下の式で最終的なスコアを計算する．

$$S(d) = S_T(d) \times S_C(d)^{\alpha} \tag{1}$$

ただし，α は S_C の影響を調整するパラメータである．Manisha [10] らは，IPC と引用関係の両方を活用している．まず IPC を分類の粒度に応じてベクトルに変換し，各特許の特徴量として割り当てられた IPC のベクトルを使用する．次に，その特徴量を引用ネットワークを基に更新し，更新後の特徴量を用いてベクトル検索を行なっている．そして，tf-idf によって文章に基づくスコアを計算し，再ランク付けして最終的な結果としている．Fujii と Manisha らの手法では引用関係をネットワークに使用している点で提案手法と異なる．

3.2 グラフ表現学習

グラフ表現学習とは，グラフデータから低次元のノード/エッジ/グラフ特徴量を生成する手法の総称であり，近年盛んに研究が行われている．ただし，ここではノード特徴量を生成する手法のみを取り上げる．GAE [11] はエンコーダとデコーダからなる教師なしノード埋め込み手法である．エンコーダに Graph Convolutional Network (GCN) [12] などの GNN を使用し，内積を基にデコードを行いグラフの再構築を行う．ノードペアの内積を sigmoid 関数によって確率に変換し，エッジのあるペアに対しては 1 を，エッジのないペアに対しては 0 を出力するように学習を行う．これによってグラフ構造を反映した特徴量の作成が可能となっている．Graph Attention Network(GAT) [13] は，教師ありの GNN モデルの一種である．GAT では自己注意機構を取り入れることで，特徴量の伝搬の際に掛ける重みをノードペアごとに調整し，精度向上を達成している．また，GraphSAGE [14] と同様に inductive setting でも使用可能であるという特徴を持つ．

3.3 ハードネガティブサンプルを用いた学習

ハードネガティブサンプリングは主に画像認識において使用される対照学習の手法である．VSE++ [15] は，画像とテキストを同じ空間に埋め込み，マルチモーダルな検索を高精度で行

うことを目標としている．VSE++では，クエリ画像に最も近いキャプションをハードネガティブキャプション，クエリキャプションに最も近い画像をハードネガティブ画像として学習に使用することで，画像検索，キャプション検索の両方のタスクで精度向上を達成している．Xuan ら [16] は，画像埋め込み生成におけるハードネガティブの分析を行っており，ハードネガティブサンプルの難易度が高すぎる場合に学習が上手く行かないことを説明している．さらに，ハードネガティブサンプルの難易度が高すぎる場合でも学習を上手く行う方法として，その場合はトリプレットロスではなく，アンカーとハードネガティブの類似度をロスとする手法を提案している．PinSage [17] は GNN を用いた類似画像検索の手法であるが，モデルのロバスト性と収束性の向上のためにハードネガティブサンプルを活用している．トリプレットロスを用いて学習を行うその際，クエリ画像に対し PageRank スコアが 2000 位〜5000 位の中からランダムに選択した画像ハードネガティブとしている，ただし全画像数は 20 億である．こうすることで，アンカーと似ているがポジティブと似ていない画像のサンプリングが可能となっている．さらに，学習の計画に工夫が施されており，最初のエポックではハードネガティブを使用せず，徐々にハードネガティブを増やすことで収束を早めている．

4 提案手法

本研究ではメタデータとして，分類情報（F ターム，IPC）と引用関係を使用する．引用関係には拒絶理由や特許査定など複数の種類が存在するが，それらを区別せず，全て同一の関係（類似特許）として扱う．メタデータの使用には技術的課題が存在する． 分類情報と引用関係はそれぞれ多クラス分類タスクと 2 値分類タスクの情報であり，本研究の目的である類似特許検索とは乖離がある．そこで，本研究では分類情報と引用関係を使用して類似特許検索の精度を向上させるアーキテクチャを提案する．まず，提案手法の概要図を図 2に示す．そして，提案手法は以下の 5 つのステップで構成される．

(1) Sentence-BERT モデルのファインチューニング (4. 1節).

(2) 分類情報を基に二部グラフを構築し，GNN に入力することで分類情報を反映させる (4. 2節).

(3) 引用関係の 2 値分類タスクを解き，引用関係を反映させる (4. 3節).

(4) モデルの出力層を埋め込みベクトルとして用いて，類似特許検索を行う (4. 4節).

また，表 1に本章で用いる主要な記号とその定義を示す．

4. 1 Sentence-BERT モデルのファインチューニング

この節では，特許のデータを基に，Sentence-BERT [4] モデルのファインチューニングを行う方法を述べる．ただし，使用するモデルおよびファインチューニングの設定は 5. 4節で述べる．Sentence-BERT モデルに使用する BERT モデルには事前学習済みのものを使用するが，そのままでは特許に関する知識が乏しいため，ファインチューニングを行う必要がある．トリプレットロスを使用して，Sentence-BERT モデルのファインチューニングを行うが，その際に必要となるトリプレットは特許間の引用関係を基に作成する．具体的には，特許ノード p_i から p_j へ引用がある場合，t_i をアンカー，t_j をポジティブとする．ネガティブは Skip-gram [18] に基づく方法で選択する．1 つのトリプレットを，t_{anc}，t_{pos}，t_{neg} で構成し，ロス関数を以下のように設計する．

$$\mathcal{L}(a,p,n) = \max\left(\|f(t_{anc}) - f(t_{pos})\| - \|f(t_{anc}) - f(t_{neg})\| + \epsilon, 0\right) \tag{2}$$

ただし，f は Sentence-BERT によるエンコードを表し，ϵ はマージンである．このロス関数を用いてファインチューニングを行うことで，引用関係の検索に適した分散表現の生成が期待できる．

表 1: 記号の定義

記号	説明
G_{pf}	F タームによる二部グラフ
G_{pi}	IPC による二部グラフ
V_p	特許ノードの集合
V_f	F タームノードの集合
V_i	IPC ノードの集合
$p_i \in V_p$	特許ノード
$f_i \in V_f$	F タームノード
$i_i \in V_i$	IPC ノード
t_i	p_i の要約文
d	特徴量の次元
$\boldsymbol{x}_{p_i} \in \mathbb{R}^d$	p_i の特徴量
$\boldsymbol{x}_{f_i} \in \mathbb{R}^d$	f_i の特徴量
$\boldsymbol{x}_{i_i} \in \mathbb{R}^d$	i_i の特徴量
$N_{pf}(p_i)$	G_{pf} における p_i の隣接ノードの集合
$N_{pf}(f_i)$	G_{pf} における f_i の隣接ノードの集合
$N_{pi}(p_i)$	G_{pi} における p_i の隣接ノードの集合
$N_{pi}(i_i)$	G_{pi} における i_i の隣接ノードの集合
$\boldsymbol{W}$	重み行列
$\boldsymbol{b}$	バイアス項
σ	活性化関数

4. 2 二部グラフの構築と GNN の適用

4. 2. 1 二部グラフの構築

特許と F タームをノードとし，特許ノード p_i が F タームノード f_j を持つ場合，ノード p_i, f_j の間にエッジを張ることで二部グラフ G_{pf} を構築する．同様にして，特許と IPC の二部グラフ G_{pi} を構築する．

4. 2. 2 初期特徴量の作成

二部グラフを GNN に入力し，分類情報を反映した特許の特徴量を作成する方法を述べる．4. 1節の方法でファインチューニングを行った Sentence-BERT モデルを用いて，特許ノードの初期特徴量 $\boldsymbol{X}_p$ を以下のように定める．

2：J-PlatPat（公開番号：特開 2015-155208）

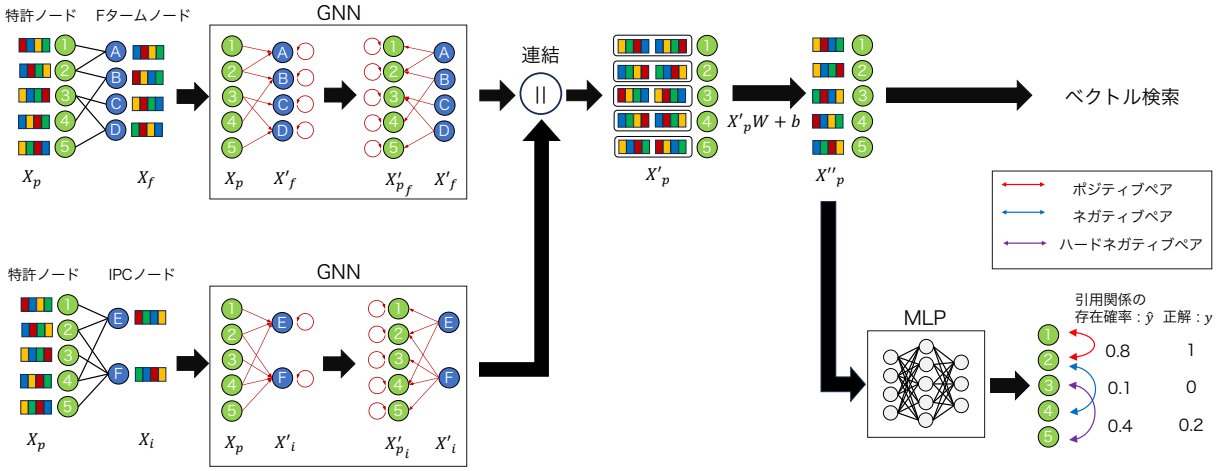


図 2: 提案手法の概要図．F タームによる二部グラフと IPC による二部グラフによって更新された特徴量を連結した後，線形層を通し，最終的な特徴量 X''_p を得る．この特徴量を使用して，引用関係の有無を判定する 2 値分類タスクを解き，ロスを計算する．学習後，X''_p を用いてベクトル検索を行う．

$$\begin{aligned} \boldsymbol{x}_{p_i} &= f(t_i) \in \mathbb{R}^d \\ \boldsymbol{X}_p &= [\boldsymbol{x}_{p_1}, \boldsymbol{x}_{p_1}, ..., \boldsymbol{x}_{p_{|V_p|}}]^T \in \mathbb{R}^{|V_p| \times d} \end{aligned} \tag{3}$$

次に，F タームノードの初期特徴量 $\boldsymbol{X}_f$ を以下のように定める．

$$\begin{aligned} \boldsymbol{x}_{f_i} &= \frac{1}{|N_{pf}(f_i)|} \sum_{p_j \in N_{pf}(f_i)} x_{p_j} \quad \in \mathbb{R}^d \\ \boldsymbol{X}_f &= [\boldsymbol{x}_{f_1}, \boldsymbol{x}_{f_2}, ..., \boldsymbol{x}_{f_{|V_f|}}]^T \quad \in \mathbb{R}^{|V_f| \times d} \end{aligned} \tag{4}$$

ここでは，隣接している特許ノードの特徴量を平均したものを F タームノードの特徴量としている．このように定めることで，エッジのある特許ノードと F タームノードの特徴量を近い値にすることができる．同様に，IPC ノードの初期特徴量 $\boldsymbol{X}_i$ を，以下のように定める．

$$\begin{aligned} \boldsymbol{x}_{i_i} &= \frac{1}{|N_{pi}(i_i)|} \sum_{p_j \in N_{pi}(i_i)} x_{p_j} \quad \in \mathbb{R}^d \\ \boldsymbol{X}_i &= [\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, ..., \boldsymbol{x}_{i_{|V_i|}}]^T \quad \in \mathbb{R}^{|V_i| \times d} \end{aligned} \tag{5}$$

4.2.3 特徴量の更新

二部グラフに GNN を適用することで，特許ノードの特徴量を更新する．本研究では GNN に GraphSAGE [14] を使用する．Algorithm 1に更新式を示す．まず，1行目の式で，隣接している特許ノード特徴量および F タームノード自身の特徴量を用いて，F タームノードの特徴量を更新している．そして，2行目の式で，隣接した F タームノードの更新後の特徴量および特許ノード自身の特徴量を用いて，特許ノードの特徴量を更新している．1,2行目の式によって，共通の F タームノードを介し，特許ノード間で特徴量の伝播を行うことができる．同様に，3,4行目の式で，IPC ノードを介し，特許ノード間で特徴量の伝播を行うことができる．最後に，5行目の式で，1, 2行目で得られた特徴量と 3, 4行目で得られた特徴量を連結し，線形層によって更新している．この式によって， F タームと IPC の情報を反映した特許ノードの特徴量を得ることができる．

Algorithm 1 Update Patent Feature Vector

Input: $\boldsymbol{X}_p, \boldsymbol{X}_f, \boldsymbol{X}_f, N_{pf}, N_{pi}$
Output: $\boldsymbol{X}''_p$

F タームの二部グラフによる特徴量の更新

1: $\boldsymbol{x}'_{f_i} \leftarrow \sigma(\boldsymbol{W}_f \cdot \boldsymbol{x}_{f_i} + \boldsymbol{W}'_f \cdot \mathrm{MEAN}_{p_j \in N_{pf}(f_i)} \boldsymbol{x}_{p_j})$
2: $\boldsymbol{x}'_{pf_i} \leftarrow \sigma(\boldsymbol{W}_{pf} \cdot \boldsymbol{x}_{p_i} + \boldsymbol{W}'_{pf} \cdot \mathrm{MEAN}_{f_j \in N_{pf}(p_i)} \boldsymbol{x}'_{f_j})$

IPC の二部グラフによる特徴量の更新

3: $\boldsymbol{x}'_{i_i} \leftarrow \sigma(\boldsymbol{W}_i \cdot \boldsymbol{x}_{i_i} + \boldsymbol{W}'_i \cdot \mathrm{MEAN}_{p_j \in N_{pi}(i_i)} \boldsymbol{x}_{p_j})$
4: $\boldsymbol{x}'_{pi_i} \leftarrow \sigma(\boldsymbol{W}_{pi} \cdot \boldsymbol{x}_{p_i} + \boldsymbol{W}'_{pi} \cdot \mathrm{MEAN}_{i_j \in N_{pi}(p_i)} \boldsymbol{x}'_{i_j})$

$\boldsymbol{x}'_{pf_i}$ と $\boldsymbol{x}'_{pi_i}$ を連結し，線形層によって更新

5: $\boldsymbol{x}''_{p_i} \leftarrow \boldsymbol{W}_{concat} \cdot (\boldsymbol{x}'_{pf_i} \parallel \boldsymbol{x}'_{pi_i}) + \boldsymbol{b}_{concat}$
6: **return** $\boldsymbol{X}''_p = [\boldsymbol{x}''_{p_1}, \boldsymbol{x}''_{p_2}, ..., \boldsymbol{x}''_{p_{|V_p|}}]^T \in \mathbb{R}^{|V_p| \times d}$

4.3 2 値分類タスクによる引用関係の反映

この節では，引用関係を基にロス関数を設計し，特許ノードの特徴量に引用関係を反映させる方法を述べる．

4.3.1 引用関係の有無の予測

4. 2節で得られた特許ノードの特徴量 $\boldsymbol{X}''_p$ を使用して，リンク予測タスクを解く．引用関係が存在する特許ペアをポジティブペアとし，ランダムに選んだ引用関係が存在しない特許ペアをネガティブペアとする．これらのペアに対してリンク予測を行い，以下の式でペア間にリンクが存在する確率を出力する．

$$\begin{aligned} \boldsymbol{h}_{ij} &= \sigma(\boldsymbol{W}_{L_1} \cdot (\boldsymbol{x}''_{p_i} * \boldsymbol{x}''_{p_j}) + \boldsymbol{b}_{L_1}) \\ \hat{y} &= \mathrm{sigmoid}(\boldsymbol{W}_{L_2} \cdot \boldsymbol{h}_{ij} + \boldsymbol{b}_{L_2}) \end{aligned} \tag{6}$$

ここでは，ペアの特徴量に対し，その要素積を MLP に入力することで引用関係を予測している．そして，予測確率 $\hat{y}$ と正解ラベル y を用いて，以下の Binary Cross Entropy Loss (BCELoss) を計算する．

$$\mathcal{L}(y, \hat{y}) = -\left(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right) \tag{7}$$

ただし，正解ラベル y はポジティブペアの場合は 1，ネガティブペアの場合は 0 とする．このロスを最小化するように学習することで，引用関係の有無を正確に予測できるような特徴量 X''_p を得ることができる． 引用関係の有無を予測できる特徴量は，文書間の関連性を捉える能力があると考えられ，ベクトル検索にも向く特徴量であることが期待できる．

4.3.2 ハードネガティブペアの活用

ランダムに選ばれたネガティブペアでの引用関係の予測は比較的容易である．このため、検索精度をさらに向上させるためには、ハードネガティブペアの利用が効果的である．ハードネガティブペアの作成方法を Algorithm2に示す．

ただし，top_k_search($\boldsymbol{X}_p$[train_nodes, $\boldsymbol{x}_{p_i}$, hardneg_end) は，$\boldsymbol{X}_p$ において $\boldsymbol{x}_{p_i}$ に近い上位 hardneg_end 件のノードを取得する操作を表す．PinSage [19] を参考に，各特許ノードからコサイン類似度に基づいて hardneg_start 位から hardneg_end 位の範囲でランダムに $N_{hardneg}$ 件を選択してハードネガティブペアを作成する．また，このハードネガティブペアは、通常のネガティブペアとは異なる扱いをする．ハードネガティブペアの正解ラベルをネガティブペアの正解ラベル 0 ではなく、label_hardneg $\in [0, 1]$ と設定する．式 (7) に，$y = \{0, 1\}$ の代わりに label_hardneg を用いて，ハードネガティブペアのロス関数を以下のように設計する．

$$\begin{aligned}\mathcal{L}_{hn}(\text{label_hardneg}, \hat{y}) = -(&\text{label_hardneg} \cdot \log(\hat{y}) \\ &+ (1 - \text{label_hardneg}) \cdot \log(1 - \hat{y}))\end{aligned} \tag{8}$$

この微調整により、モデルはハードネガティブペアを通常のネガティブペアと区別し、より微妙な識別能力を身につけることが期待できる．

そして，全体のロス関数を以下に示す．

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_n + \mathcal{L}_{hn} \tag{9}$$

ただし，$\mathcal{L}_p$, $\mathcal{L}_n$ はそれぞれポジティブペア，ネガティブペアの場合の式 (7) のロスである．

Algorithm 2 Create Hard Negative Pairs

Input: $\boldsymbol{X}_p$, pos_pairs, neg_pairs, $N_{hardneg}$, train_nodes, hardneg_start, hardneg_end
Output: hardneg_pairs

```
hardneg_pairs ← ∅
for p_i ∈ train_nodes do
    top_k ← top_k_search(X_p[train_nodes], x_{p_i}, hardneg_end)
    hardneg_pairs_p_i ← ∅
    while |hardneg_pairs_p_i| < N_hardneg do
        p_j ← random_choice(top_k[hardneg_start : ])
        if (p_i, p_j) ∉ pos_pairs and (p_i, p_j) ∉ neg_pairs then
            hardneg_pairs_p_i ← hardneg_pairs_p_i ∪ (p_i, p_j))
        end if
    end while
    hardneg_pairs ← hardneg_pairs ∪ hardneg_pairs_p_i
end for
return hardneg_pairs
```

4.3.3 ミニバッチ学習

GPU メモリの制約を考慮し、ミニバッチ学習を行う．ポジティブペア，ネガティブペア，ハードネガティブペアを含むように各バッチを構成する．具体的には，ネガティブペア，ハードネガティブペアはバッチ数に基づいて各バッチに均等に分配し，ポジティブペアは分割せずに各バッチが全てのポジティブペア含むようにする．これによって学習の偏りを防ぐ．

4.4 類似特許検索

この節ではベクトル検索によって類似特許検索を行う方法について述べる．ベクトル検索では， 4.2節の特許ノードの特徴量 $\boldsymbol{X}''_P$ を埋め込みベクトルとして使用し，コサイン類似度が高い K 件を検索結果とする．

5 実　　験

本稿では，以下の 3 つの実験を行う．はじめに，提案手法と既存手法を比較し，類似特許検索に提案手法が有効であることを確認する．次に，提案手法に関して，transductive setting と inductive setting の 2 つの設定で実験を行い，inductive setting における精度の低下度合いを確認する． 最後に，提案手法におけるハードネガティブペアのハイパーパラメータを変化させ，精度がどのように変化するかを調査した．

5.1 データセット

表 2に本研究で用いたデータセットの詳細を示す．

- **IPC D**：繊維，紙の分野の特許データセット．
- **IPC E**：固定構造物の分野の特許データセット．

4.1節で述べた通り，特許の初期特徴量に要約文を使用しているため，要約文のない特許は除外している（平成 2 年 12 月 1 日から要約書の提出が義務付けられたため，それ以降の特許は要約文が存在する）．表の各列について説明する．F ターム数は，4章における $|V_f|$ に対応しており，データセット内の特許が持つ F タームの種類の数である．IPC 数は，4章における $|V_i|$ に対応しており，データセット内の特許が持つ IPC の種類の数であるが，IPC D，E では，それぞれ D，E から始まるものに限定している． また，引用関係は，同一データセット内における引用関係のみを考慮している．F ターム数 / 特許は，1 つの特許あたりの F ターム数の平均値であり，4章の G_{pf} における特許ノードの平均次数に対応している．IPC 数 / 特許は，1 つの特許あたりの IPC 数の平均値であり，4章の G_{pi} における特許ノードの平均次数に対応している．

データセットは，特許の申請日時に古い順に，Train データが 80%，Validation データが 10%，Test データが 10%になるように分割する．

5.2 評 価 指 標

類似特許を漏れなく見つけることが目標であるため，MEAN Recall@K を使用する．また，検索件数 K については，5.4節では $K = 10, 30, 50, 100, 150, 200$ で評価を行い，5.6節では $K = 100$ のみで評価を行う．

5.3 パラメータ設定

学習率を 0.0001 とした Adam Optimizer [20] を使用して，最大エポック数 500，中間層の次元を 768 次元，ドロップアウト [21] を 0.5 として学習を行う．また，バッチノルム [22] を使用し，学習の加速を図った．また，評価ステップを 10 とし，10 エポックごとに類似特許検索を実行した．そして，提案手法に関するパラメータについては，バッチ数 (num_batch) を IPC D では 5，IPC E では 10 とし，ハードネガティブの正解ラベル (label_hardneg) を 0.2， 1 train node あたりのハードネガティブの本数 ($N_{hardneg}$) を 10 とし，ハードネガティブの選択範囲に関しては hardneg_start = 200，hardneg_end = 1000 とした．さらに，ネガティブペアの件数 (num_neg_pairs) を |positive_pairs| × 15 − |train_nodes| × $N_{hardneg}$ で定め，ネガティブペアとハードネガティブペアの合計の本数をポジティブペアの本数の 15 倍とした．ここで，ここで，ハードネガティブペアの本数は，|train_nodes| × $N_{hardneg}$ である．

5.4 比 較 実 験

提案手法と既存手法の比較を行う．比較手法を以下にまとめる．

表 2: データセットの詳細.

IPC	概要	特許数	F タターム数	F タターム数 / 特許	IPC	IPC 数 / 特許	平均引用数
D	繊維，紙	164075	74783	27.20	2807	2.53	1.24
E	固定構造物	470066	106855	12.03	3075	1.76	1.27

- 全文検索：キーワード検索と比較したいが，特許の要約文にキーワードが付与されていないため全文検索で代替する．検索エンジンには Elasticsearch[3]を使用する．要約文をデータベースに登録し，全文検索を行う．形態素解析エンジンには kuromoji[4]を使用する．
- F タームによる検索：特許ペアの類似度を F タームの Jaccard 係数とする．
- Sentence-BERT による検索：4. 1節で述べた方法でファインチューニングを行った後，文書の分散表現を使用してベクトル検索を行う．

ここで，Sentence-BERT モデルおよびファインチューニングの設定について述べる．本研究では，Sentence-BERT に使用する BERT モデルに，東北大が公開しているモデル (bert-base-japanese-whole-word-masking[5]) を使用する．このモデルの出力は 768 次元である．また，プーリング層には平均プーリングを使用する．ファインチューニングにおけるパラメータ設定については，バッチサイズを 8，学習率を 2×10^{-6}，ウォームアップステップをバッチ数の 10%とし，トリプレットロスにおけるマージンは $\{0.1, 0.3, 0.5\}$ から選択した．また，エポック数に関しては，IPC D では $\{1, 2, ..., 25\}$，IPC E では $\{1, 2, ..., 15\}$ から選択した．ただし，記述していないパラメータに関しては，Sentence Transformers ライブラリのデフォルト設定から変更していない．

表 3に比較実験の結果を示す．ただし，提案手法では，乱数シードを変更し 3 回試行した平均を結果として用いている．提案手法は，2 つのデータセットで全ての K において比較手法を上回った．さらに，K が大きいほど，提案手法と既存手法の再現率の差が大きくなっている．このことから，提案手法では分類情報と引用関係を活用することでベクトル検索に向いた特徴量を生成できていると言える．また既存手法においては，IPC D の $K = 10$ を除くほぼ全ての結果で，Sentence-BERT による検索，全文検索，F タームによる検索の順で再現率が高くなった．このことから，F タームよりもテキスト情報を用いる方が類似した特許を見つけやすいと言える．Sentence-BERT が全文検索よりも再現率が高いのは，4. 1節で述べた方法でファインチューニングを行ったからであると考えられる．

5. 5 inductive setting における実験

提案手法において，transductive setting と inductive setting の 2 つの設定で実験を行い，inductive setting における精度の低下度合いを確認する．まず，本研究におけるこれらの設定を次のように定める．transductive setting では，学習時， 検証時， テスト時のいずれにおいても，全てのデータを利用可能である．それ対し，inductive setting では過去のデータのみ利用可能．つまり，学習時は Train データのみ，検証時は Train データおよび Validation データ，テスト時は全てのデータが利用可能である．この設定では未知のデータに対して推論を行っている．

表 4に実験結果を示す．transductive setting に対し，inductive setting での再現率の低下は，0.01 以内に収まっている．この結果より，提案手法は未知の特許に対しても検索に向いた埋め込みベクトルを生成可能であると言える．

表 3: IPC D，E における MEAN Recall@K．ただし，Sentence-BERT および提案手法のハイパーパラメータは Validation のデータの Recall@100 が最も良いものを使用している．

(a) IPC D

K	10	30	50	100	150	200
全文検索	0.147	0.214	0.252	0.310	0.347	0.375
F タームによる検索	0.085	0.151	0.191	0.259	0.307	0.342
Sentence-BERT による検索	0.144	0.227	0.274	0.352	0.401	0.440
提案手法	**0.188**	**0.304**	**0.371**	**0.481**	**0.547**	**0.594**

(b) IPC E

K	10	30	50	100	150	200
全文検索	0.118	0.178	0.212	0.265	0.300	0.325
F タームによる検索	0.0608	0.109	0.140	0.195	0.232	0.261
Sentence-BERT による検索	0.129	0.205	0.250	0.320	0.366	0.402
提案手法	**0.162**	**0.264**	**0.326**	**0.421**	**0.483**	**0.527**

表 4: IPC D，E における transductive setting および inductive setting での MEAN Recall@K

(a) IPC D

K	10	30	50	100	150	200
提案手法（transductive）	**0.188**	**0.304**	**0.371**	**0.481**	**0.547**	**0.594**
提案手法（inductive）	0.185	0.303	**0.371**	0.478	0.546	0.592

(b) IPC E

K	10	30	50	100	150	200
提案手法 (transductive)	**0.162**	**0.264**	**0.326**	0.421	0.483	0.527
提案手法（inductive）	0.160	0.263	0.323	**0.422**	**0.484**	**0.528**

5. 6 ablation study

ここでは提案手法における各要素の再現率への影響を評価する．まず，ablation study で使用する NCELoss について説明する．提案手法では，2 値分類タスクを解き，その予測誤差を使用してモデルのパラメータを更新している．そして，2 値分類タスクを解く際，判定するペアの埋め込みベクトルの要素積を MLP に入力している．この MLP の効果を調べるために，MLP を用いないロスで実験を行う必要があるが，ここではその方法として，Noise Contrastive Estimation Loss (NCELoss) [23] を使用する．NCELoss を用いることで，ポジティブペアとネガティブペアを区別するように学習を行うことができる．ただし，複数のネガティブペアを使用するために，文献 [24] で提案されたロス関数を使用する．さらに，ペア間のスコアには文献 [25] と同様にコサイン類似度を使用する．以下に本研究で使用する NCELoss を示す．

$$\mathcal{L}(a, p, n_s) = -\log\left(\frac{\exp(\cos(\boldsymbol{z}_a, \boldsymbol{z}_p)/\tau)}{\exp(\cos(\boldsymbol{z}_a, \boldsymbol{z}_p)/\tau) + \sum_{n \in n_s} \exp(\cos(\boldsymbol{z}_a, \boldsymbol{z}_n)/\tau}\right) \quad (10)$$

3：https://www.elastic.co/jp/elasticsearch
4：https://www.atilika.com/ja/kuromoji/
5：https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

ただし，$\boldsymbol{z}$ は特徴量ベクトルであり，a, p, n_s はそれぞれ，アンカー，ポジティブ，ネガティブ集合である．また τ は温度パラメータである．

続いて，実験結果をについて述べる．表 5に，G_{pf} を取り除いた場合，G_{pi} を取り除いた場合，G_{pf}，G_{pi} を共に取り除き，2 層の MLP に置き換えた場合，$N_{hardneg} = 0$ とした場合，2 値分類タスクの BCELoss を NCELoss に変更した場合の MEAN Recall@100 の結果を示す（それぞれ，w/o F ターム， w/o IPC, w/o F ターム or IPC，w/o ハードネガティブ MLP&BCELoss, w/o ハードネガティブ NCELoss に対応している）．ただし，記述していないパラメータに関しては，5. 3節で述べたパラメータ設定から変更していない．また，w/o ハードネガティブ NCELoss における τ は，$\{0.02, 0.04, 0.06, 0.08, 0.10\}$ で変化させた．この結果より，IPC D， E の両方で，提案手法の各要素が再現率の向上に効果があることをわかる．さらに，w/o F ターム or IPC と w/o F ターム，w/o IPC の比較すると，GNN を MLP に置き換えると 0.1 以上再現率が下がっていることがわかる．これより G_{pf} または G_{pi} によって，分類情報を反映することが重要であると言える．また，w/o ハードネガティブ MLP&BCELoss と w/o ハードネガティブ NCELoss を比較すると，w/o ハードネガティブ MLP&BCELoss の方が再現率が高いため，MLP を使用して 2 値分類タスクを解くことが，検索により有効な埋め込みベクトルの生成に効果的であると言える．

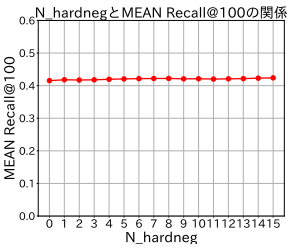
表 5: 各要素の MEAN Recall@100 への影響.

手法	IPC D	IPC E
提案手法	**0.482**	**0.421**
w/o F ターム	0.437	0.396
w/o IPC	0.458	0.385
w/o F ターム or IPC	0.333	0.277
w/o ハードネガティブ　MLP&BCELoss	0.457	0.415
w/o ハードネガティブ　NCELoss	0.445	0.408

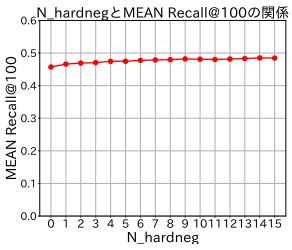
5. 7　ハードネガティブペアに関するパラメータセンシティビティ

ハードネガティブペアに関するパラメータを変化させ，提案手法での再現率がどのように変化するかを調べる．それらのパラメータを以下のように変化させる．ただし，5. 3節での設定，この節での設定の順に述べる．まず，1 train node あたりのハードネガティブの本数 ($N_{hardneg}$) は 10 から $\{1, 2, 3, ..., 14, 15\}$ へ，そして，ハードネガティブの正解ラベル（label_hardneg）は 0.2 から $\{0.0, 0.1, ..., 0.9, 1.0\}$ へ，最後に，ハードネガティブの選択範囲 (hardneg_start，hardneg_end) は $(201, 1000)$ から $(1, 200)$，$(201, 400)$，...，$(2801, 3000)$ へ変更する．また，Recall@100 で評価を行う．

実験結果を図 3，4, 5に示す．まず，$N_{hardneg}$ に関しては大きな差は得られなかった．これより，ハードネガティブをネガティブの代わりに使用可能あると言える．次に，label_hardneg は 2 つのデータセットにおいて 0.2 付近で再現率が最も高くなった．これよりハードネガティブの正解ラベルを適切に定めることが重要であると言える．最後に，ハードネガティブの選択範囲は 2 つのデータセットにおいて共に選択順位が高すぎる場合に，ランダムにハードネガティブを選んだ場合よりも再現率が低下した．これよりハードネガティブには適切な難易度設定が必要であると言える．

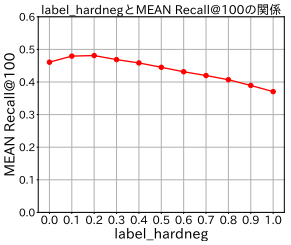


(a) IPC D

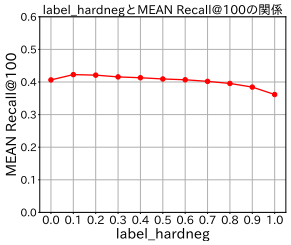


(b) IPC E

図 3: IPC D，E における $N_{hardneg}$ と MEAN Recall@100 の関係



(a) IPC D



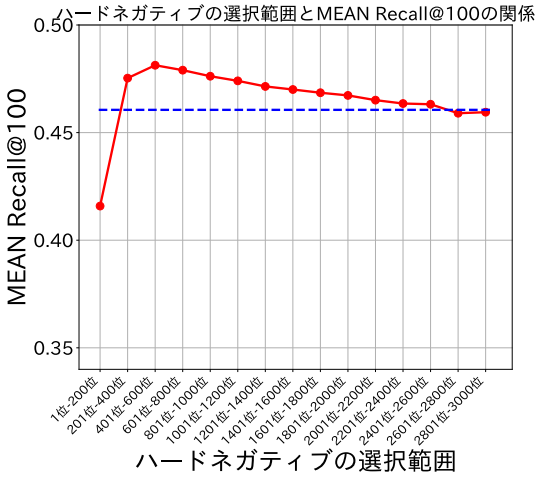
(b) IPC E

図 4: IPC D，E における $label_{hardneg}$ と MEAN Recall@100 の関係

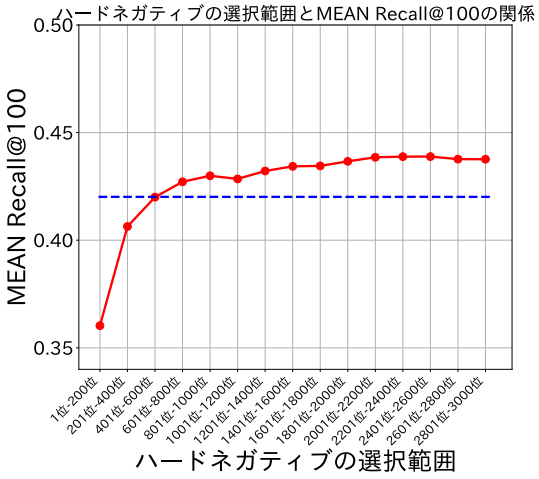
6　終わりに

本稿では，分類情報と引用関係を活用した類似特許検索の手法を提案した．まず，分類情報（F ターム，IPC）を使用して二部グラフを作成し，GNN に入力することで分類情報を反映した特徴量を作成し，その後，引用関係の予測誤差を最小化する学習を通して引用関係を反映させた．実験では，2 種類の特許データセットで，提案手法がキーワード検索および Sentence-BERT を用いた検索を上回り，提案手法のアーキテクチャが分類情報と引用関係の両方を有効に活用できることを示した．さらに，ablation study を行い，分類情報とハードネガティブが検索精度に対して貢献していることを確認した．

今後の課題としては，データセットの拡大が挙げられる．本稿では，一部の特許データのみを使用したが，全ての特許データを用いて学習および推論が可能となれば実用性が増す．さらに，分類情報に対するロバスト性の評価も課題である．本稿では，クエリとなる特許に付与された分類情報を全て使用し，推論を行なった．この場合，検索を行うユーザが分類情報を適切に定める必要がある．そこで，分類情報に対するロバスト性，つまり分類情報が少ない場合や誤った分類情報が与えられた場合に検索精度がどの程度悪化するかを調べる必要がある．



(a) IPC D



(b) IPC E

図 5: IPC D，E における hardneg_start，hardneg_end と MEAN Recall@100 の関係．ただし，青い波線はランダムに選んだ場合を表す．

謝　　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです。

文　　献

[1] 鈴木祥子. 機械による特許分析の課題とアプローチ. 情報の科学と技術, Vol. 67, No. 7.

[2] 悟志福田, 英嗣難波, 寿幸竹澤. 論文と特許からの技術動向情報の抽出と可視化. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 2, 2013.

[3] Sunhye Kim and Byungun Yoon. Patent infringement analysis using a text mining technique based on sao structure. *Computers in Industry*, Vol. 125, , 2021.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the EMNLP-IJCNLP*, 2019.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Lu Wei Kurzweil, Ray". Universal sentence encoder for English. In *Proceedings of the EMNLP*, 2018.

[6] Kazuya Konishi. Query terms extraction from patent document for invalidity search. In *Proceedings of the NTCIR*.

[7] Atsushi Fujii. Enhancing patent retrieval by citation analysis. In *Proceedings of the SIGIR*, 2007.

[8] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, Vol. 30, , 1998.

[9] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the SIGIR*, 1994.

[10] Manisha Verma and Vasudeva Varma. Patent search using ipc classification vectors. In *Proceedings of the 4th Workshop on Patent Information Retrieval*, 2011.

[11] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the ICLR*, 2017.

[13] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the ICLR*, 2018.

[14] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the NeurIPS*, 2017.

[15] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018.

[16] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *Proceedings of the ECCV*, 2020.

[17] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the KDD*, 2018.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the NeurIPS*, 2013.

[19] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the KDD*, 2018.

[20] Adam: A method for stochastic optimization. In *Proceedings of the ICLR*, 2015.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 56, 2014.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the ICML*, 2015.

[23] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the AISTATS*, 2010.

[24] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the CVPR*, 2018.

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the ICML*, 2020.

グラフ要約を用いた高速なグラフ相関問合せ

直井　悠馬†　　真次　彰平††　塩川　浩昭††

† 筑波大学 理工情報生命学術院　〒 305-8577 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305-8577 茨城県つくば市天王台 1-1-1
†{naoi,matsugu}@kde.cs.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　相関問合せとはグラフデータベースの中からクエリと共起して出現する部分グラフを検出する処理であり，医療や生命科学分野で広く利用されている．これまでグラフデータベースの中からクエリグラフとの相関値が最も高い k 個の部分グラフを求める Top-k 相関グラフ問合せ手法が提案されている．しかし，相関問合せはグラフデータベースを構成するグラフがより大規模になると相関判定のコストは大きくなるため，実行に多くの時間がかかる．本研究ではグラフ構造を要約したグラフを構築することで，相関判定のコストを小さくし，相関問合せのコストを削減する手法を提案する．特に，先行研究における相関問合せのコストが大きくなる問題に対して，グラフの経路を用いることで高速に相関判定を行う手法を提案した．実データを用いた評価実験を行い，既存手法と比較して高速な Top-k 相関問合せが可能であることを確認する．

キーワード　グラフデータベース，データ構造・索引，問合せ処理，相関問合せ

1　序　　論

グラフはオブジェクト間の関係を表現する一般的なデータモデルであり，科学や生物学，生命科学分野，ソーシャルネットワークなどの分野で広く使われている [1,2]．グラフはデータオブジェクトを頂点とし，データオブジェクト間の関係を辺として表す．複数のグラフから構成されるグラフデータベースの需要が高まると共に，グラフデータベースの中から有用な情報を検出するグラフマイニングの問題が重要になる．これまでデータベースの中から頻出した部分グラフ構造を検出する頻出部分グラフ検出の研究 [3–7] が広く行われてきた．その中でもデータベースの中から共起して出現する部分グラフ対を検出する相関部分グラフ検出の研究 [8–12] は近年注目を集めている．

相関問合せはグラフデータベースとクエリグラフが与えられたとき，クエリグラフの出現分布に近い部分グラフを特定する問題である．共起して出現する部分グラフ対はデータベースの隠れた特性を示すことがあり，多くのアプリケーションで有用である．例えば，科学化合物のデータベースでは共起して出現する構造対を検出し，新たな知識発見の一助となる．

グラフデータベースに対する相関問合せはグラフデータベースとクエリグラフが与えられたとき，クエリグラフと共起して出現する部分グラフを特定する問題である．図 1 に相関問合せの例を示す．図 1 において，グラフデータベース $\mathcal{D}$ は g_1 から g_9 のグラフによって構成されている．この図では，データベース内でのクエリグラフ q と類似した出現分布を持つ部分グラフ g'_1 がクエリグラフに対する代表的な相関グラフとして検出される．これはクエリグラフ q がデータベース中に $\{g_1, g_2, g_3, g_6, g_7\}$ で出現しているのに対して，g'_1 は $\{g_2, g_4, g_6, g_7\}$ で出現しており，クエリグラフと共起して出現していると言えるためである．

しかし，大規模なグラフデータベースに対する相関グラフ問合せにはいくつかの課題がある．1 つの課題はグラフデータベース内の全てのグラフの部分グラフが相関グラフの候補，すなわち候補グラフとなる点である．もう 1 つの課題は相関グラフ判定には候補グラフの出現頻度を考えるため膨大なコストがかかるという点である．これらの問題に対して，これまでユーザによって与えられた閾値より大きな相関値を持つ部分グラフを検出する手法 [9,10] や，相関値が最も高い k 個の部分グラフを検出する Top-k 相関問合せの手法 [8] が提案されている．この手法は探索する部分グラフ数を削減し，効率的に相関計算を行うことで相関グラフを検出することができる．

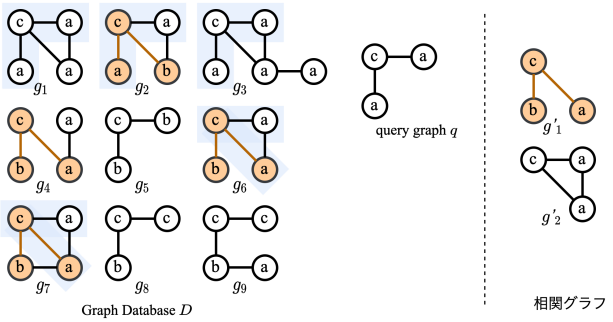


図 1: 相関問合せの例

Top-k 相関問合せに対して，直井らはグラフデータベースの各グラフ構造を要約したグラフデータベースを構築することで相関問合せコストを削減した高速な Top-k 相関問合せ手法を提案した [13]．この手法では，要約したグラフデータベースを構築することで，探索するグラフ数を縮小させると共に，候補グラフの相関問合せの判定コストの削減を図る．しかし，直井らの手法では，要約したデータベースにおいて相関問合せをした後．正確な Top-k 相関グラフを検出するために，要約を解除して検証を行う．検証のために既存の相関問合せ手法 [8] を用い

ており，探索するグラフ集合に含まれる部分グラフが大きい場合には検証に多くのコストを要する．そのため，大規模なグラフデータベースでは候補グラフの数が大きくなり，相関値計算のコストが大きくなるため，相関問合せの実行時間が大きくなるという既存の相関問合せ手法が抱える問題が依然として残る．

この問題を解決するために，本研究では大規模なグラフデータベースに対する高速な Top-k 相関グラフ問合せ手法を提案する．具体的には，提案手法は直井らの手法 [13] における検証処理を高速化を行うために，グラフの経路を用いた高速な相関グラフ判定を行う．提案手法では，相関グラフの候補グラフに出現する経路の出現頻度を考えることによって，クエリグラフとの相関値の上界を小さなコストで計算し，膨大なコストがかかる厳密な相関値計算を回避する．

本研究では，評価実験によって，実際のグラフデータベースに対して，提案手法の Top-k 相関問合せ処理が最先端の手法に比べて高速であることを確認した．また，提案手法はパラメータ k，グラフデータベースのサイズを変えた場合にも，既存手法に比べて高速であることを確認した．

本論文の構成は以下の通りである．2 節で前提知識と本研究で取り扱う問題について説明する．3 節で先行研究について，4 節で提案手法について述べる．5 節で実データを用いた性能評価結果を示し，最後に 6 節で本研究のまとめを述べる．

2 事前準備

本研究で対象とするグラフはラベル付無向連結グラフ $g = (V, E, l)$ である．V は頂点集合，E は辺集合，l は各頂点と各辺にラベルを与えるラベル関数である．また N 個のグラフ $g_1, g_2, \ldots, g_N$ から構成されるグラフデータベースを $\mathcal{D} = \{g_1, g_2, ..., g_N\}$ とする．2 つのグラフを $g = (V, E, l)$, $g' = (V', E', l')$ について g は g' の部分グラフ同型であるとき，$g \subseteq g'$ と表記する．また，本論文では $g \subseteq g'$ となるとき，g' は g のスーパーグラフと呼ぶ．

グラフ g の $\mathcal{D}$ での出現頻度の指標として支持度を $supp(g; \mathcal{D}) = \frac{|\mathcal{D}_g|}{|\mathcal{D}|}$ と定める．本論文では $\mathcal{D}$ が文脈上明らかであるときは，$supp(g; \mathcal{D})$ を $supp(g)$ と表記する．2 つのグラフ g_1, g_2 が同時に出現する割合である結合支持度は $supp(g_1, g_2) = \frac{|\mathcal{D}_{g_1} \cap \mathcal{D}_{g_2}|}{|\mathcal{D}|}$ と定める．支持度には逆単調性があり，$g \subseteq g'$ ならば，$supp(g') \leqq supp(g)$ となる．

相関グラフ問合せとは $\mathcal{D}$ とクエリグラフ $q = (V_q, E_q, l_q)$ が与えられたとき，クエリと共起して出現する部分グラフを $\mathcal{D}$ から検出することである．本研究では 2 つのグラフの相関尺度としてピアソン相関係数 [14] を採用し，定義は以下の通りである．

定義 1 (ピアソン相関係数)**.** 2 つのグラフ g_1, g_2 が与えられたとき，それらの相関値 $\phi(g_1, g_2)$ を以下のように定める．

$$\phi(g_1, g_2) = \frac{supp(g_1, g_2) - supp(g_1)supp(g_2)}{\sqrt{supp(g_1)supp(g_2)(1-supp(g_1))(1-supp(g_2))}}$$

相関値 $\phi(g_1, g_2)$ は $[-1, 1]$ の範囲の値をとる．また，$\phi(g_1, g_2)$ が正の値をとるとき，2 つのグラフ対はグラフデータベースの中で共起して出現すると言える．本研究で対象とする Top-k 相関グラフ問合せを以下のように定義する．

定義 2 (Top-k 相関グラフ問合せ)**.** グラフデータベース $\mathcal{D} = \{g_1, g_2, ..., g_N\}$，クエリグラフ q，整数 k が与えられたとき，相関値 $\phi(g, q)$ が最も高い k 個のグラフ g を求める．

3 先行研究

本節では先行研究 [13] について説明する．大規模なグラフデータベースに対する Top-k 相関グラフ問合せにおける膨大なコストを克服するために，様々な研究がされてきた．先行研究 [13] では，高速に Top-k 相関グラフ問合せを行うために，グラフ要約と乱択アルゴリズムに基づく手法を提案している．

基本アイディア：先行研究 [13] の基本アイデアは，データベース内の各グラフを要約することによってグラフサイズを縮小し相関判定のコストを下げることである．グラフ要約とは，2 つの頂点を 1 つの頂点に集約する操作を繰り返し行うことでグラフの頂点数と辺数を小さくすることである．要約されたグラフから構成されるグラフ集合である要約グラフデータベース内のグラフはグラフサイズが小さいため，相関判定のコストが小さくなり，Top-k 相関グラフ問合せを高速に行うことができる．

しかし，グラフ要約は Top-k 相関グラフ問合せに対して偽陽性・偽陰性の結果をもたらす場合がある．先行研究では乱択アルゴリズム [15] を利用することで確率的に偽陽性・偽陰性となる結果を除外し，高速高精度な Top-k 相関グラフ問合せを実現する．先行研究は以下の 3 つで構成される．

(1) グラフ要約：グラフデータベース $\mathcal{D}$ の各グラフに対してランダムに辺を集約した要約グラフデータベース $\mathcal{D}'$ を構築する．

(2) 探索：$\mathcal{D}'$ に対して Top-k 相関グラフ検出を行う．グラフ要約は偽陽性・偽陰性をもたらす場合があるため，複数の要約グラフデータベースにおいて，相関グラフ問合せ手法により k 個の要約された部分グラフからなる集合 T を複数生成する．

(3) 検証：複数の T をグラフデータベース $\mathcal{D}$ と照合しグラフ復元を行い，正確な Top-k 相関グラフを求める．

先行研究は，事前処理としてデータベース $\mathcal{D}$ の各グラフについてグラフ要約を行い，t 個の要約グラフデータベースを構築する．問合せ処理は，クエリグラフと要約グラフデータベース $D'_j (1 \leqq j \leqq t)$ に Top-k 相関グラフ問合せ [8] を適用することで，相関グラフの探索を行う．次に探索で得られた相関グラフ集合 T_j について検証を行い，Top-k 相関グラフ集合 T を得る．

3.1 グラフ要約

グラフ要約は $\mathcal{D} = \{g_1, g_2, ..., g_N\}$ について $\mathcal{D}$ 内の各グラフを要約したグラフデータベース $\mathcal{D}' = \{g'_1, g'_2, ..., g'_N\}$ を構築する．グラフ要約はエッジ集合から $\langle l(u), l(u, v)l(v) \rangle$ をラベル対として，ラベル対集合 R_j をサンプリングする．ただし，ラベル集合 R_j 内のラベル対は少なくとも 1 つ以上 $\mathcal{D}$ 内のいずれかのグラフで隣接しており，ラベル対の両端点は R_j 内の他のラベル対と重複しない．先行研究は R_j からラベル対 r を取り出した後，$\mathcal{D}$ 内の各グラフ g_i に対して，r を含む場合，それらのノードを 1 ノードに集約する操作を R_j の要素がなくなるまで

繰り返し，$\mathcal{D}'$ を構築する．

3.2 探　　索

要約グラフデータベース $\mathcal{D}'_j(1 \leqq j \leqq t)$ から，対応するラベル対集合 R_j によって要約されたクエリグラフ q'_j に対する Top-k 相関グラフ T_j をそれぞれ検出する．このとき，t は乱択アルゴリズムによって複数回実行される反復回数である．先行研究における探索では，要約グラフデータベースに対して既存の Top-k 相関グラフ問合せ手法 [8] を用いることで，各 $\mathcal{D}'$ における要約された Top-k 相関部分グラフを検出する．要約グラフデータベース $\mathcal{D}'$ における探索は，グラフサイズが小さいため $\mathcal{D}$ に対する探索に比べて，相関問合せのコストが小さくなる．

乱択アルゴリズムの導入：グラフ要約処理はランダムに辺を集約するため，集約の順序やラベル対集合 R の内容に依存して Top-k 相関グラフ問合せにおいて，偽陽性・偽陰性を生じさせる場合がある．そこで先行研究では，乱択アルゴリズムを導入することで確率的に偽陽性・偽陰性の排除を図る．具体的には，複数のラベル対集合 $R_j(1 \leqq j \leqq t)$ を構築し，これらを基に互いに異なる要約グラフデータベース $\mathcal{D}'_j$ を構築する．複数の要約グラフデータベースと要約されたクエリグラフに対して Top-k 相関グラフ問合せを実行することで，偽陽性・偽陰性となる部分グラフを除外を図る．

3.3 検　　証

先行研究における検証では，t 個の要約グラフデータベースに対して Top-k 相関グラフ問合せを行った $T_j(1 \leqq j \leqq t)$ に対して正確な Top-k 相関グラフを求める．T_j に含まれる k 個のグラフは要約グラフデータベースの部分グラフである．したがって，T_j を相関グラフとして検証するためには元々のグラフデータベース $\mathcal{D}$ におけるグラフの部分グラフに変換する必要がある．T_j にそれぞれ含まれる k 個の要約グラフは，$\mathcal{D}$ の各グラフにおける隣接リスト情報とラベル対集合 R に基づき要約を解除して復元する．

復元の操作としては次の手順である．グラフ g の要約グラフを g'，要約グラフデータベースに相関問合せをした結果獲得した T' とする．最初に g' のある頂点が集約された頂点ならば，R を参照し元々の 2 つの頂点間を接続し，g の隣接リスト情報から隣接する頂点を選択し辺を接続する．この操作を繰り返すことでグラフの復元を行い，T' の全ての部分グラフに対して復元を行う．検証では，復元によって得られたグラフ集合に対して．厳密な Top-k 相関グラフを求めるため，Top-k 相関グラフ問合せ手法 [8] を適用することで，相関グラフの探索を行う．

3.4 偽陰性・偽陽性の理論解析

乱択アルゴリズムにおける確率的に偽陽性・偽陰性の排除を図るために反復回数を求める．一回の反復において，任意のグラフ g がグラフ要約によって偽陰性となる確率 P を $P = 1-(1-\frac{k}{l})^{\alpha}$ とする．このとき，k, l はそれぞれ，要約されるラベル対の数とグラフデータベースのラベル対の数である．また，α はグラフデータベースの平均次数とクエリグラフの平均次数の差とクエリグラフの頂点数の積で与えることができる．これより，乱択アルゴリズムの反復回数を以下の定理 1 のように定める．証明は省略する．

定理 1. ϵ 以下の確率で偽陰性を排除するためには，以下の t 回以上の反復を行う．

$$t = \left\lceil \frac{\log(\epsilon)}{\log(P)} \right\rceil$$

先行研究は各グラフデータベースにおいて定理 1 より得られる反復回数だけ要約グラフデータベースを作成し，探索を行う．

4 提案手法：検証の高速化

この節では，提案手法である Top-k 相関問合せの高速化について説明する．提案手法では，先行研究における検証処理を高速化を行うために，グラフの経路を用いた高速な相関グラフ判定を行う．先行研究の検証処理では，t 個の要約グラフデータベースに対して Top-k 相関グラフ問合せを行った $T_j(1 \leqq j \leqq t)$ に対して正確な Top-k 相関グラフを求める．検証において正確な Top-k 相関グラフを求めるために，既存の Top-k 相関グラフ問合せ手法 TopCor [8] を用いており，探索するグラフ集合に含まれる部分グラフが大きい場合には検証に多くのコストを要する．そのため，大規模なグラフデータベースでは候補グラフの数が大きくなると共に，相関値計算のコストが大きくなるため，相関問合せの実行時間が大きくなるという既存の相関問合せの課題を抱える．本節では，先行研究の検証処理を高速化するために，経路を用いた候補グラフの枝刈り手法を提案する．

基本アイディア：検証の高速化の基本アイディアは，候補となる部分グラフ g の経路長 q の経路の出現頻度を考えることによって，$supp(g; \mathcal{D}_q)$ の上界を小さなコストで計算し，膨大なコストがかかる厳密な相関値計算を回避することである．候補グラフが Top-k 相関グラフにならない場合は，その候補グラフは候補から排除することが可能であり，この操作を枝刈りと呼ぶ．支持度の逆単調性より，$g \subseteq g'$ となるグラフ g, g' には，$supp(g'; \mathcal{D}_q) \leqq supp(g; \mathcal{D}_q)$ の関係が成り立つ．そのため，頂点数 q のグラフと見做せる経路長 q の経路 p と候補グラフ g の支持度にも同様に逆単調性がある．このとき，経路 p はグラフ g から得られるため，経路 p のグラフ g_p において，$g_p \subseteq g$ は自明である．したがって，経路 p の支持度の上界を高速に求め，$supp(g; \mathcal{D}_q)$ の上界を小さなコストで計算することによって，候補グラフの枝刈りを行う．

提案手法で用いるグラフの経路を以下のように定義する．

定義 3 (経路)**.** あるグラフ g に含まれる経路を，頂点とその間に存在する辺について，接続する順序で並べた一連の列とする．ただし，経路において頂点と辺はそれぞれ連続して並ばず，端は必ず頂点である．経路の長さ n としたとき，n は経路の頂点数とする．

このとき，あるグラフ g には経路が複数存在するため，以下の経路集合を定義する．

定義 4 (経路集合)**.** あるグラフ g に含まれる全ての重複無し経路集合を $P(g)$ とする．$P_n(g) \subseteq P(g)$ は長さ n の経路集合と

する．

定義 5 (インスタンス集合)**.** グラフ g と $p \in P(g)$ が与えられたとき，g における p のインスタンス集合を $\mathbb{I}_p(g)$ とする．$\mathbb{I}_p$ の各要素は，p が g にマッピングされるときの辺の順序集合である．

このとき，$g_i \in \mathcal{D}_q$ における全て経路 p について，g でのインスタンス集合を計算を避けるために，クエリグラフ q における p のインスタンス集合を $\mathbb{I}_p(q)$ を考える．$g_i \in \mathcal{D}_q$ はクエリグラフのスーパーグラフなので，$P(q) \subseteq P(g)$ となる．したがって，候補グラフ g のある経路 p が $p \in P(q)$ であるとき，空ではないインスタンス集合 $\mathbb{I}_p(g)$ を必ず持ち，$|\mathbb{I}_p(g)| \geqq 1$ である．これらより，経路 p が $g_i \in \mathcal{D}_q$ にてどれだけ出現するかは，$p \notin P(q)$ となる経路のみについて考えれば良い．

次に，グラフの経路を用いて支持度の上界を導出する．あるグラフ g の $\mathcal{D}_q$ での支持度の上界は，以下のように定義される．ただし，$N_q = |\mathcal{D}_q|$ とする．

定義 6 ($\mathcal{D}_q$ での支持度の上界)**.** あるグラフ g の $\mathcal{D}_q = \{g_1, g_2, ..., g_{N_q}\}$ での支持度の上界 $suup_{ub}(g; \mathcal{D}_q)$ を以下のように定める．このとき，$1 \leqq i \leqq N_q$ である．

$$suup_{ub}(g; \mathcal{D}_q) = \frac{min_{p_j \in P(g) \setminus P(q)} \sum_{g_i \in \mathcal{D}_q} \kappa(g, g_i, p_j)}{|\mathcal{D}_q|} \quad (1)$$

$$\text{ただし,} \kappa(g, g_i, p_j) = \begin{cases} 1 & (|\mathbb{I}_{p_j}(g)| \geqq |\mathbb{I}_{p_j}(g_i)|) \\ 0 & (Otherwise) \end{cases}.$$

図 2 を用いて説明する．ただし，グラフの辺のラベルは全て 1 とし，図と以下では省略する．$P_2(q) \subseteq P(q) = \{a-c\}, P_2(g) \subseteq P(g) = \{a-c, b-c\}$ であり，$P_2(g) \setminus P_2(q) = \{b-c\}$ である．このとき，経路 $p_i = b-c$ とすると，グラフ g における p_i のインスタンス集合 $\mathbb{I}_{p_i}(g)$ は，$\mathbb{I}_{p_i}(g) = \{v_0v_1\}$ となる．同様に，$\mathcal{D}_q$ におけるそれぞれのグラフについても，$\mathbb{I}_{p_i}(g_1) = \varnothing, \mathbb{I}_{p_i}(g_2) = \{v_0v_3\}, \mathbb{I}_{p_i}(g_4) = \{v_0v_1\}, \mathbb{I}_{p_i}(g_6) = \{v_2v_3, v_2v_4\}$ となる．したがって，グラフ g の $\mathcal{D}_q$ での支持度の上界 $suup_{ub}(g; \mathcal{D}_q)$ は，$\kappa(g, g_1, p_i) = 0, \kappa(g, g_2, p_i) = 1, \kappa(g, g_4, p_i) = 1, \kappa(g, g_6, p_i) = 1$ より，$\frac{3}{4}$ となる．

提案手法では，あるグラフ g の $\mathcal{D}_q$ での支持度の上界を用いて，グラフ g の取りうる相関値の上界を求めることで，厳密な相関値を求めることなく枝刈りすることができる．あるグラフ g の $\mathcal{D}_q$ での支持度の上界より，上界を用いた相関値と相関値の最大値 $\phi_{max}(g)$ との関係は以下のようになる．ただし，グラフ g とクエリグラフ q の相関値は $supp(g) = supp(g, q)$ のとき最大値 $\phi_{max}(g)$ を取るため，式 (1) の $supp(g)$ に $supp(g, q)$ を代入することで求めることができる [8].

補題 1. $\phi_{max}(g)$ は．$supp(g; \mathcal{D}_q) = suup_{ub}(g; \mathcal{D}_q)$ のとき，$\phi_{ub}(g)$ をとるとする．$\phi_{max}(g)$ は，$supp(g; \mathcal{D}_q)$ に単調増加であるため，以下の不等式が成り立つ．

$$\phi_{max}(g) \leqq \phi_{ub}(g). \quad (2)$$

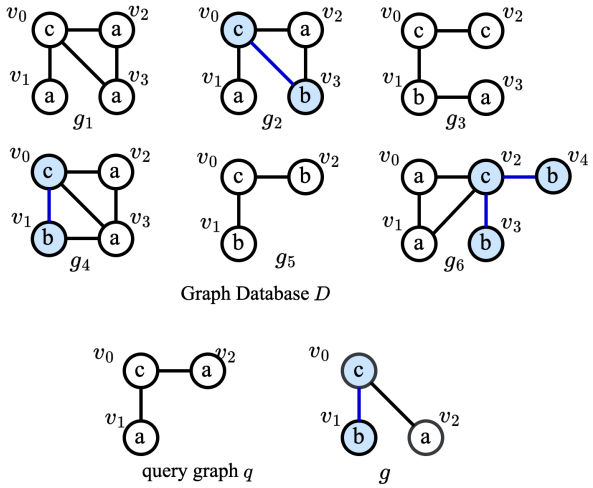


図 2: グラフデータベースと経路

Proof. グラフ g と任意の経路長 q の経路 p において，p を頂点数 q のグラフに変換し，g_p とする．定義 3 より，経路 p はグラフ g に含まれるため，$g_p \subseteq g$ である．このとき，グラフ g_p, g において支持度の逆単調性より，$supp(g; \mathcal{D}_q) \leqq supp(g_p; \mathcal{D}_q)$ が成り立つ．また，定義 6 より，$suup_{ub}(g; \mathcal{D}_q)$ はグラフ g に含まれる経路の内，クエリグラフに含まれる経路を除き最小の支持度を与えるため，$supp(g; \mathcal{D}_q) \leqq suup_{ub}(g; \mathcal{D}_q)$ が成り立つ．したがって，$\phi_{max}(g)$ は $supp(g; \mathcal{D}_q)$ に単調増加であるため $\phi_{max}(g) \leqq \phi_{ub}(g)$ である． □

これより，支持度の逆単調性より 2 つの候補グラフ g_1, g_2 が $g_2 \subseteq g_1$ のとき，$supp(g_1, q) \leqq supp(g_2, q)$ が成り立つため，$\phi_{\max}(g_1) \leqq \phi_{\max}(g_2)$ である．既存の Top-k 相関グラフ問合せ手法 [8] では，Top-k 相関グラフの探索において，k 番目の相関値を ϕ_{min} とし，探索中の候補グラフ g の相関値と比較することで枝刈りを行った．同様に，提案手法では $\phi_{ub}(g)$ と k 番目の相関値 ϕ_{min} を比較することで，検証において候補グラフ g の枝刈りを行う．候補グラフ g が $\phi_{ub}(g) < \phi_{\min}$ となる場合，$\phi_{max}(g) < \phi_{ub}(g) < \phi_{\min}$ であるため，g は相関グラフとはなり得ず，枝刈りすることができる．

これより，候補となる部分グラフ g の経路長 q の経路の出現頻度による候補グラフの枝刈りによって，検証処理を高速に行うことができると考えられる．これは経路の出現頻度は，$supp(g; \mathcal{D}_q)$ を計算するよりも小さい計算コストで求めることができるためである．$supp(g; \mathcal{D}_q)$ の計算を行うための最悪計算量は $\mathcal{O}(|\bar{n}|! \cdot |\bar{n}| \cdot |\mathcal{D}_q|)$ であり [16]，グラフの頂点数 $\bar{n}$ が多くなるとコストが高い．対して，経路の出現頻度は，経路のインスタンス集合を参照すれば良いため，小さいコストで得ることができる．また，先行研究では事前処理の要約段階でデータベースに対して操作するため，要約時に経路を列挙することで経路を検証の段階で再度求める必要がない．

5 評価実験

提案手法の計算時間の評価を行うために，実データを用いて

提案手法と先行研究 [13] の比較を行った．実験は以下の 3 つの点に着目して提案手法の有効性を検証する．

- 出現頻度の異なるクエリグラフに対する性能変化
- Top-k の値 k を変化させた場合における性能変化
- データベース $\mathcal{D}$ に含まれるグラフ数に対する性能変化

提案手法は，先行研究における検証処理を高速化を行うために，グラフの経路を用いた高速な相関グラフ判定を行う．Top-k 相関問合せにおける比較手法は以下の通りである，また，定理 1 におけるパラメータ ϵ は 0.05 とし，使用するデータベース毎に理論解析に基づき，要約グラフデータベースの数を決定する．提案手法は支持度の上界を用いて枝刈りを行うため，Top-k 相関問合せにおける精度は先行研究に対して変化しない．そのため，問合せの精度が $1-\epsilon = 0.95$ 以上となるような反復回数をとる．提案手法と先行研究は，検証における高速化以外のパラメータは全て同様のものを使用する．

- **提案手法：** グラフデータベースに対する Top-k 相関問合せの提案手法である．提案手法のフレームワークは先行研究と同様に，事前処理であるグラフ要約と問合せ処理である探索，検証の 2 ステップによって構成される．問合せ処理における検証において，経路を用いた高速化を行う．
- **先行研究：** 3 節におけるグラフデータベースに対する Top-k 相関問合せの提案手法である．

5.1 実験環境とデータセット

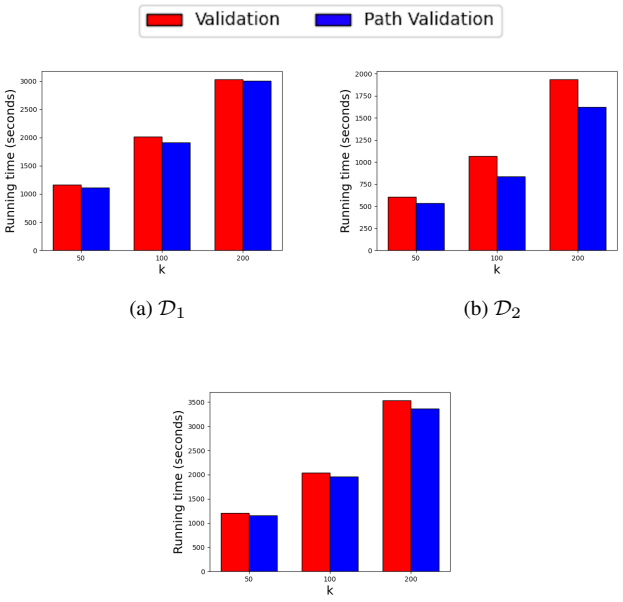
提案手法のアルゴリズムは C++を用いて実装し，コンパイルオプションには-O3 を用いた．グラフは隣接リスト構造を用いて実装した．部分グラフ同型判定には最新のアルゴリズム [17] を用いた．実験は Intel Xeon Platinum 8268 2.9 GHz, 1 TiB RDIMM で構成される Linux サーバ上で行った．

実験に用いたデータセットは実際の癌や AIDS などのタンパク質や酵素の化合物グラフデータベースである [18]．これらのグラフデータベースは原子を頂点とし，原子間の結合を辺として表現されている．LIT-PCBA [18] のそれぞれのデータベースは，特定の性質を持つタンパク質によって構成される．これらのグラフデータベースの詳細を表 1 に示す．実験に用いるクエリグラフは [0.001, 0.005], (0.005, 0.01], (0.01, 0.03] の範囲の支持度を持つクエリグラフを利用し，それぞれ Q_1, Q_2, Q_3 とする．

表 1: 実データセット

ID	データセット	グラフ数	平均頂点数	引用元
$\mathcal{D}_1$	ADBR2	312,500	24.9887	[18]
$\mathcal{D}_2$	PKM2	246,069	24.4584	[18]
$\mathcal{D}_3$	ALDH1	145,133	23.8583	[18]

評価実験では，Top-k 相関問合せにおける k の値を 50, 100, 200 と変化させ，経路を用いた場合と用いない場合の比較実験を行う．クエリは Q_1 に属するクエリを与える．グラフ g の支持度の上界 $suup_{ub}(g;\mathcal{D}_q)$ の計算に用いる経路長は $n = 3$ とする．



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$

(a) $\mathcal{D}_3$

図 4: 異なるデータベースにおける検証の高速化の影響

5.2 検証の高速化の効果

提案手法の経路を用いた検証の高速化の効果を評価する．経路を用いた検証は，相関計算を行うグラフ g に対して $suup_{ub}(g;\mathcal{D}_q)$ を計算することで枝刈りを行い，枝刈りを行ったグラフの $supp(g;\mathcal{D}_q)$ と $supp(g;\mathcal{D})$ の計算を回避する．そのため，$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ のようなグラフ数が多いグラフデータベースは．$supp(g;\mathcal{D}_q)$ の計算コストが大きくなるため，有効に働くと考えられる．

図 4 に $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ における経路を用いた検証の高速化の結果を示す．また，図 4 では，提案手法において検証にかかる処理時間のみを示し，グラフ要約と探索の時間は含まれないことを留意されたい．

図 4 では．経路を用いた検証を行わない場合は赤色で，経路を用いた検証は青色で表示されている．図 4 より，経路を用いた検証によって $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ において，平均 1.102 倍高速，最大 1.277 高速に検証を行うことができた．これはグラフ数が多いグラフデータベースにおいて，経路を用いた検証によって候補となるグラフの支持度計算を回避したことによると考えられる．また，探索と検証を含めたクエリ処理全体においては，経路を用いた検証によって最大 1.09 倍高速にクエリ処理を行うことができた．

6 結論と今後の展望

本研究では，グラフの経路を用いた高速な相関グラフ判定による高速な Top-k 相関問合せ手法の提案を行った．提案手法では，先行研究における検証処理を高速化を行うために，グラフの経路によって相関値の上界をより小さなコストで求め，相関グラフとなりえないグラフを枝刈りを行う．実データを用いた評価実験により，提案手法は先行研究よりも高速に計算できる

ことが示された．

今後の課題として，先行研究における検証処理の高速化の他に，探索の高速がある．探索では，要約グラフデータベースに対して相関問合せを行うが，相関問合せはグラフデータベースのグラフサイズが大きいほどコストが大きくなる．そのため，事前処理によって段階的に要約を行うことで要約グラフデータベースのグラフサイズをより小さくすることで探索における相関問合せのコストを小さくすることができると考えられる，

謝　　辞

本研究の一部は，JST さきがけ (JPMJPR2033) ならびに JSPS 科研費 (JP22K17894) の支援を受けたものである．

文　　献

[1] Shiokawa Hiroaki, Amagasa Toshiyuki, and Kitagawa Hiroyuki. Scaling Fine-grained Modularity Clustering for Massive Graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 4597–4604, 2019.

[2] Shiokawa Hiroaki and Onizuka Makoto. *Scalable Graph Clustering and Its Applications*, pp. 2290–2299. Springer New York, 2018.

[3] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *In Proc. IEEE International Conference on Data Mining*, pp. 721–724, 2002.

[4] T. Washio A. Inokuchi and H. Motoda. An aprioribased algorithm for mining frequent substructures from graph data. In *In Proc. Principles of Data Mining and Knowledge Discovery*, pp. 13–23, 2000.

[5] M. Kuramochi and G. Karypis. Frequent Subgraph Discovery. In *In Proc. IEEE International Conference on Data Mining*, p. 313–320, 2001.

[6] M. Fredrikson M. Christodorescu X. Yan C. Chen, C. X. Lin and J. Han. Mining Graph Patterns Efficiently via Randomized Summaries. In *In Proc. the Very Large Data Base Endowment*, pp. 742–753, 2009.

[7] I. Fischer M. Worlein, T. Meinl and M. Philippsen. A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston. . In *In Proc. Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 392–403, 2005.

[8] Yiping Ke, James Cheng, and Jeffrey Xu Yu. Top-k Correlative Graph Mining. In *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM 2009)*, pp. 1038–1049, 2009.

[9] Yiping Ke, James Cheng, and Wilfred Ng. Correlation Search in Graph Databases. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD2007*, p. 390–399, 2007.

[10] Yiping Ke, James Cheng, and Wilfred Ng. Efficient Correlation Search from Graph Databases. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 12, pp. 1601–1615, 2008.

[11] Yiping Ke, James Cheng, and Jeffrey Xu Yu. Efficient discovery of frequent correlated subgraph pairs. In *In Proc,the 2009 Ninth IEEE International Conference on Data Mining*, p. 239–248, 2009.

[12] Arneish Prateek, Arijit Khan, Akshit Goyal, and Sayan Ranu. Mining Top-k Pairs of Correlated Subgraphs in a Large Network. *Proceedings VLDB Endowment (PVLDB)*, Vol. 13, No. 9, p. 1511–1524, may 2020.

[13] 直井悠馬, 真次彰平, 塩川浩昭. グラフデータベースに対する高速高精度な相関問合せ. 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023.

[14] RH. Reynolds. *The analysis of cross-classifications*. The Free Press, New York, 1977.

[15] Prabhakar Raghavan Rajeev Motwani. *Randomized algorithms*. he Press Syndicate of the University of Cambridge, ISBN 0-521-47465-5, 1995.

[16] C. Sansone L. P. Cordella, P. Foggia and M. Vento. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. In *In Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1367–1372, 2004.

[17] Shixuan Sun and Qiong Luo. Scaling Up Subgraph Query Processing with Efficient Subgraph Matching. In *IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 220–231, 2019.

[18] Viet Khoa Tran Nguyen, Célien Jacquemard, and Didier Rognan. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *Journal of Chemical Information and Modeling*, Vol. 60, No. 9, pp. 4263—-4273, 2020.

ラベル付き有向グラフにおける node embedding を用いたスキーマ抽出

木下　一歩†　　鈴木　伸崇††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1911477@u.tsukuba.ac.jp, ††nsuzuki@slis.tsukuba.ac.jp

あらまし　グラフデータは情報や実態の相互関係を記述することに優れており，様々な用途で使用されている．グラフはスキーマを与えることで情報検索や構造理解が容易になるが，スキーマはほとんど与えられていないのが現状である．そこで，本研究では，グラフデータからスキーマの表現を抽出する方法を提案する．同様の課題に取り組む先行研究はすでにいくつか存在し，なんらかの指標を定義または利用し，クラスタリングを行うものが多く挙げられる．その一方で，貪欲法等によって愚直に解析を行うようなアプローチもあるが，この方法では複雑さが NP 完全であることもあり，処理効率に課題がある．このような背景から，本研究では，node2vec を拡張した手法を用いてノードの埋め込みを行うという既存手法とは異なるアプローチを試みる．評価実験によって，提案手法が概ね適切にスキーマを抽出でき，実行時間はデータ量に対してほぼ線形に推移することがわかった．また，対象とするデータの特徴によって結果に差があることもわかり，実験を通して相性の良いデータの特徴を明らかにした．

キーワード　グラフデータ，スキーマ，スキーマ抽出，埋め込み, graph, schema, schema extraction, node embedding

1　はじめに

グラフデータは情報や実体の相互関係を記述することに優れており，Linked Open Data 等に利用されるほか，近年大きく利用が広がっている Web や SNS におけるユーザや投稿等の関係を記述することにも利用可能である．このような背景から高度に情報化が進む現代において，グラフデータの潜在的な利用可能性は高まっていると言える．グラフデータはスキーマを与えることで情報検索やその構造理解が容易になる．しかしながら，グラフデータに対してスキーマはほとんど与えられていないのが現状である．そこで本研究では，グラフデータからスキーマの表現を抽出する方法を提案する．

グラフデータからスキーマを得る手法はすでにいくつか提案されている．本研究と同じくラベル付き有向グラフを入力とし出力として ShapeExpression(ShEx) スキーマを得る手法としては坪井ら [1] の研究があるほか，他の入出力による手法も提案されている．(Paton ら [2] や Kellou-Menouer ら [3] による調査を参照) 例えば，類似度として Dice 係数を用い，階層的クラスタリングの手法である Gaussian Mixture Model を用いる Bonifati ら [4] の研究や，Kardoulakis ら [5] によるエッジラベルを参照して instance vector を定義し，これを Locality Sensitive Hashing によって類似度を比較してグルーピングする研究がある．これらの既存手法における傾向として，何らかの類似度を図る指標を定義または利用し，クラスタリングを行うという手順のものが多いことが挙げられる．また坪井ら [6] によればラベル付き有向グラフから ShEx スキーマを抽出するという問題は NP 完全であり，実際に貪欲法を用いている坪井ら [6] の提案手法において処理効率について課題があることが指摘されている．

このような背景から本研究では，node2vec を拡張した手法によるノードの埋め込みを行うことでノード間の類似度計算の効率化を図る，という既存手法とは異なるアプローチを試みる．より具体的には，node2vec で提案されているランダムウォークについて，本問題に適用する際の問題点を指摘した上で，その点を改善するような拡張を行い，skip-gram による学習で埋め込み表現を得て，クラスタリングを行う．その後クラスタリング結果を踏まえてラベルを集計することでそれぞれのクラスタに分類されたノード全てに対して妥当であるようなスキーマを得る．また，クラスタリングには k-means++を利用するが，クラスタ数の決定には指標を定義し Elbow 法を用いている．

評価実験の結果，概ね適切にスキーマを抽出でき，同じく SP^2Bench と ShEx を対象とする坪井らの研究と比較してより大きなデータについても処理が可能であることが分かった一方で，対象とするグラフデータの特性による相性が存在することもわかった．入力データとしては RDF のベンチマークツールである SP^2Bench と RDF 生成ツールである WatDiv を用いて生成したデータを用い，出力されるスキーマとしては高い表現力を持つ ShEx を対象とした．グラフデータに対するスキーマとしては RDF Schem 等も存在するが RDF Schema は，オントロジ記述言語や語彙提供としての性格が強く，データ構造を表現するスキーマ言語としては適切な選択とは言えないためである．

2　諸　定　義

2.1　グ　ラ　フ

本研究で用いるグラフとは，ノードとエッジによって表現されるデータ構造である．エッジに指向性があるものを特に有向グラフと呼び，エッジにラベルがついているものを特にラベル

付きグラフと呼ぶ．本研究で扱うグラフは，エッジに指向性がありかつエッジにラベルを持つようなグラフであるため，特にラベル付き有向グラフと呼ぶ．本論文では特に断りなくグラフという語を用いたときにはラベル付き有向グラフのことを指す．また，出力エッジを持たないノードについて，一意に識別可能なノードとしての意味を持つものを終端ノードまたは終端，文字列等のデータの内容について記述するものをリテラルと呼ぶ．

グラフはノード集合 V とラベル集合 L さらにエッジ集合 $E \subseteq V \times L \times V$ の組で定義し，$G = (V, E, L)$ と記述する．図 1 にラベル付き有向グラフの例を示す．このグラフは，意味的には v_1, v_9 が学生を表すノード，$v4$ が講義を表すノード, $v6$ が教員を表すノードで，他のノードは数値や文字列などを表すリテラルである．

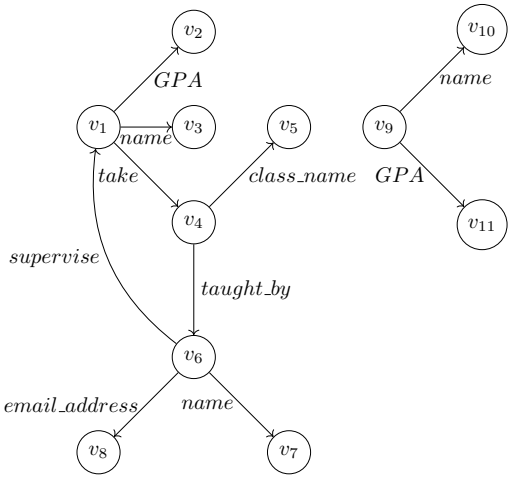


図 1　ラベル付き有向グラフの例

このグラフに対して V, E, L を記述すると下記のようになる．

$$
\begin{aligned}
V = \{&v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{11}\} \\
E = \{&(v_1, GPA, v_2), (v_1, name, v_3), (v_1, take, v_4), \\
&(v_4, class_name, v_5), (v_4, taught_by, v_6), \\
&(v_4, supervise, v_1), (v_6, name, v_7), \\
&(v_6, email_address, v_8), (v_9, name, v_{10}), \\
&(v_9, GPA, v_{11})\} \\
L = \{&GPA, name, take, class_name, \\
&taught_by, supervise, email_address\}
\end{aligned}
$$

本研究では SP^{2}Bench や WatDiv で生成された RDF を用いるが，RDF は主語および目的語をノードとして表現し，両者の間に主語から目的語の方向に指向性を持ち，述語をエッジラベルとして持つようなエッジを定義している．したがって，RDF データはラベル付き有向グラフとして表現できる．

2.2　ShEx [7] スキーマ

本節で扱う概念の詳細な定義や ShEx の記法等は Staworko [8] らによる定義を参照している．本研究では，グラフの構造を表すスキーマとして ShEx スキーマを対象とする．

まず ShEx の理解に必要な RBE(Regular Bag Expression) の概念について述べる．RBE は正規表現と類似した仕組みだが，それぞれの表現の連結についてグラフの文脈ではその順序を考慮しないことから，順序のある連結の代わりに順序を無視した連結である"||"を用いる．また，繰り返しについて，本件研究では下記の 3 通りのいずれかを用いることとした．

- "+" - 1 回以上
- "*" - 0 回以上
- "?" - 0 または 1 回

次に ShEx および ShEx スキーマについて述べる．エッジラベルの有限集合を Σ，型名の有限集合を Γ として，型に対して $\Sigma \times \Gamma$ の RBE を割り当てることで表現する．また，$l \in \Sigma, t \in \Gamma$ のとき，$(l, t) \in \Sigma \times \Gamma$ を簡単のため $l :: t$ と表記する．さらに型名を RBE に写すような関数 $\delta : \Gamma \to \Sigma \times \Gamma$ を定義する．ここで，ShEx を定義できないような型 (リテラルなど) は ϵ をそのことを表す意味で用いる．以上の前提の上で $S = (\Sigma, \Gamma, \delta)$ を ShEx スキーマと呼ぶ．また，それぞれのノードを型に移すような関数 $\lambda : V \to \Gamma$ も同時に定義する．これらの定義に従って，図 1 に対する ShEx スキーマと λ について記述すると下記のようになる．

$$
\begin{aligned}
\Sigma = \{&GPA, name, take, class_name, taught_by, supervise, \\
&email_address\} \\
\Gamma = \{&t_0, t_1, t_2, t_3\} \\
\delta(t_0) &= \epsilon \\
\delta(t_1) &= GPA :: t_0 || name :: t_0 || take :: t_2 \\
\delta(t_2) &= class_name :: t_0 || taught_by :: t_3 \\
\delta(t_3) &= name :: t_0 || email_address :: t_0 || supervise :: t_1 \\
\lambda(v_1) &= \lambda(v_9) = t_1 \\
\lambda(v_4) &= t_2 \\
\lambda(v_6) &= t_3 \\
\lambda(v_2) &= \lambda(v_3) = \lambda(v_5) = \lambda(v_7) = \lambda(v_8) = \lambda(v_{10}) = \lambda(v_{11}) = t_0
\end{aligned}
$$

3　提 案 手 法

本章ではまず，疑似コードによってアルゴリズムの概観を示した後に，各節で詳細な処理について述べる．

3.1　アルゴリズム

アルゴリズムの疑似コードを下記 Algorithm1 に示す．処理は大きく分けて三段階に分かれている．まず Embedding によって node2vec を拡張した手法によりノードをベクトル空間に埋め込む．次に Clustering によって Embedding の処理結

果である，ノードを表すベクトルをクラスタリングする．最後にクラスタリングの結果を踏まえて同じクラスタのノード同士のラベルを集計することによってスキーマを得る．

計算量について，本研究で新たに定義した手順である RandomWalk についてのみ記述する．skip-gram やクラスタリングについては既存の研究に準じる．RandomWalk では各ノード (ノード数 $|V|$) について 1 回の探索で次に訪れるノードの決定を最大 l 回行い，1 つのノードにつき探索を w 回繰り返す．また，1 回の探索で一度訪れたノードを再訪問しないために，探索済みノード集合と遷移候補ノード集合の差集合を取る計算を次に訪れるノードの決定ごとに行う必要があり，この処理に $O(l)$ かかる．よってランダムウォークにかかる計算量は $O(|V| \times l^2 \times w)$ である．

Algorithm 1

Input: Graph $G = (V, E, L)$, Dimensions d, Context size k, Walks per node w, Walk length l, Search bias p
Output: ShEx Schema s
f = Embedding$(G, d, k = l, w, l, p)$
c = Clustering(G, f)
s = GetSchema(G, c)
return s

3.2 埋め込み

本節では埋め込みについて述べる．

Algorithm 2 Embedding

Input: Graph $G = (V, E, L)$, Dimensions d, Context size k, Walks per node w, Walk length l, Search bias p
$sentences$ = RandomWalk(G, w, l, p)
f = skipgram$(sentences, k)$
return f

この処理では次節で述べるランダムウォークによって生成した”コーパス”に対して skip-gram を実行することで ”v_1”,”v_2” . . . といったノードを表す特別な単語の埋め込み表現をノードの情報を代表するベクトルとして得る．ただし，単にグラフを解析することによって容易に求めることができる，終端ノードは埋め込みの対象外とした．必要なパラメータは埋め込みベクトルの次元 d と後述する周辺語の最大幅 k，ランダムウォークに関するパラメータ群であるノード 1 つあたりの探索回数 w と探索の長さの最大値 l，遷移確率の重み付け係数 p である．

skip-gram とは，Mikolov ら [9] によって提案された機械学習手法で，コーパス中のある単語を入力語として周辺語を予測するようなタスクを実行する．周辺語の最大幅 k をあらかじめ与える必要があり，評価実験ではランダムウォークの探索の長さ l と同じ幅を指定した．つまり，$sentence$ の ”v_i” を除いた長さは必ず k 以下になるため ”v_i” に対して $sentence$ に含まれるすべてのラベルが周辺語とみなされる．

3.3 ランダムウォーク

本節ではランダムウォークについて述べる．

Algorithm 3 RandomWalk

Input: Graph, G, Walks per node w, Walk length l, Search bias p
Initialize $sentences$ to Empty
for all nodes $u \in V$ **do**
  for $iter = 1$ **to** w **do**
    Initialize $sentence$ to [NodeNumber(u)]
    $parent = None$
    $cur = u$
    Initialize $V_{visited}$ to [cur]
    for $walk_iter = 1$ to l **do**
      V_{cur} = GetNeighbros$(cur, parent, G) \setminus V_{visited}$
      $next_node$ = DecideNextNode$(cur, parent, V_{cur}, p)$
      if $next_node == None$ **then**
        break
      end if
      if $next_node$ is a child of $parent$ **then**
        $label$ = GetEdgeLabel$(parent, next_node)$
      else if $next_node$ is a child of cur **then**
        $label$ = GetEdgeLabel$(cur, next_node)$
        $parent = cur$
      end if
      $cur = next_node$
      Append $next_node$ to $V_{visited}$
      Append $label$ to $sentence$
    end for
    Append $sentence$ to $sentences$
  end for
end for
return $sentences$

基本的なアイディアは Grover ら [10] によって提案された node2vec の通りであるが，本問題に適用にするにあたり 2 点改善を行った．小節 3.3.1，3.3.2 でそれらについて述べ，小節 3.3.3 で本研究で提案するランダムウォークについて述べる．

3.3.1 サンプルする対象

1 点目はサンプルする対象である．具体的にはノードをサンプルする点である．

本問題で扱っているのはノードがどのような型すなわち ShEx を持つか, また同様の型のノードはグラフ内のどのノードであるかである．ShEx をより分解して書き下すと，プロパティと接続先ノードの型として表現できる．ただし，接続先ノードの型はサンプル時点では未知である．しかしながら，接続元ノードの型と同様に接続先ノードの型もプロパティとさらに先の接続先ノードの型で表現されることから，グラフの回路等を考慮しない限り，どのノードを始点としても必ず最終的には接続先ノードを持たないような終端ノードにたどり着く．このことから例えば図 1 の v_1 について $\delta(v_1)$ を接続先ノードの型について書き下していき，連結 || を用いる代わりに改行をすると図 2 のようになる．入れ子構造については { によって表現している．

$$\begin{cases} GPA :: \epsilon \\ name :: \epsilon \\ take :: \begin{cases} taught_by :: \begin{cases} name :: \epsilon \\ email :: \epsilon \end{cases} \\ class_name :: \epsilon \end{cases} \end{cases}$$

図 2

このように書き下してくと，スキーマをエッジラベルと終端 ϵ のみによって表現することができることがわかる．つまり，今回捉えようとしている情報はすべてエッジラベルによってグラフ上では表現されていると言える．また，直接の接続関係を持たない，すなわちノード同士が歩道を持たないような場合でも同様の型を持つことは図 1 の v_1 と v_9 のように十分考えられるため，この点においてもサンプルすべきはノードやエッジ，ノードラベルではなく，エッジラベルである．

以上のような理由から，本問題ではサンプルする対象はノードではなくエッジラベルとする．

3.3.2 遷 移 確 率

2 点目は遷移確率の重み付けである．

node2vec では 2 つのパラメータ p, q を用いて次に訪れるノード x への遷移確率 $\alpha_{pq}(t,x)$ をカレントノード v の直前に訪れた親ノード t と x の距離 d_{tx} によって以下のように 3 種類の重み付けをしている．

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \textbf{if } d_{tx} = 0 \\ 1 & \textbf{if } d_{tx} = 1 \\ \frac{1}{q} & \textbf{if } d_{tx} = 2 \end{cases}$$

これによって node2vec は幅優先探索と深さ優先探索の中間のような探索を行うことが可能になっている．しかし，本問題においては $d_{tx} = 0$ すなわち親ノードに戻るという遷移はカレントノードが探索の始点の場合はそのノードが持つプロパティと接続先ノードの型とは無関係であるし，カレントノードが始点以外の場合はすでに訪問済みであり，遷移は不要である．よって本問題では親ノードからの距離は $\{1,2\}$ のいずれかとなり，これらに異なる重みをつける場合パラメータは一つで十分である．

3.3.3 ランダムウォーク

以上の 2 点を踏まえて本問題向けに node2vec を拡張したアルゴリズムが 3.1 節に示したアルゴリズムである．必要なパラメータは探索の長さ l, 1 つのノードから探索を行う回数 w, 遷移確率の重み付け係数 p , 探索を始めたノードの祖先を長さ 1 で探索する回数 $parent_width$ (デフォルトは 0, オプション) である．一度訪れたノードには再訪問しないこととしているため，一度訪れたノードはノード集合 $V_{visited}$ に記録する．また，親ノードの初期値は $None$ としているため，始点とするノードの祖先は原則探索しないことに注意されたい．次に訪れるノードはノード集合 $V_{next} = \{v \mid v \in (\text{ChildOf}(parent) \cup \text{ChildOf}(cur)) - V_{visited}\}$ から選ばれ， 遷移確率 $\alpha_p(parent, cur, v)$ の重み付けは以下の式で定義する．

$$\alpha_p(parent, cur, v) = \begin{cases} \frac{1}{p} & \textbf{if } v \in \text{ChildOf}(cur) \\ 1 & \textbf{if } v \in \text{ChildOf}(parent) \end{cases}$$

ただし，$V_{next} = \emptyset$ の場合は次の訪れるノードは選ばれず，探索の長さが l 未満でも探索を終了する．訪れるノードが決定したらエッジラベルのサンプルや変数や集合の更新を行う．

ランダムウォークの結果である，ノードを表す特別な単語 v_i を先頭とするサンプルしたラベルの羅列 $sentence$ の集合 $sentences$ が skip-gram に与える”コーパス”となる．図 3 はこの手順に則って図 1 のグラフについて $w = 1, l = 4, p = 2$ としてランダムウォークを行った際に得られる”コーパス”の例である．

[v_1,”GPA”,”$take$”,$taught_by$”,”$class_name$”]
[v_4,”$taught_by$”,”$email_address$”,”$name$”]
[v_6,”$name$”,”$email_address$”,]
[v_9,”GPA”,”$name$”]

図 3 図 1 に対してランダムウォークを行った際に得られる”コーパス”の例

3.4 クラスタリング

本節ではクラスタリングについて述べる．

Algorithm 4 Clustering
Input: Graph G, Embedding result f
$k = \text{GetNClusterByElbow}(V, f)$
$clustering = \text{k-means++}(V, f, k)$
return $clustering$

本問題では事前にクラスタの数がわからないため，Elbow 法によってこれを決定した．実験の実装には pyclustering を用いており，公式ドキュメントによれば pyclustering における Elbow 法は，k-means++によるクラスタリングを複数回行い，各クラスタについて以下の指標を計算，その数値が最も大きくなるクラスタ数を最適なクラスタ数として採用する．x_k はクラスタ数すなわち k であり，y_k はクラスタ数 x_k で k-means++を実行した結果のクラスタ内誤差平方和である．x_0 を取りうる最小のクラスタ数，x_{-1} を取りうる最大のクラスタ数として記述している．

$$Elbow_k = \frac{|(y_0 - y_{-1})x_k + (x_{-1} - x_0)y_k + (x_0 y_{-1} - x_{-1} y_0)|}{\sqrt{(x_{-1} - x_0)^2 + (y_{-1} - y_0)^2}}$$

取りうる最小及び最大のクラスタ数はあらかじめ与える必要があり，評価実験では $x_{min} = 1$, $x_{max} = 29$ とした．

3.5 ス キ ー マ

本節ではスキーマを抽出する手順について簡単に述べる．

クラスタリングまで完了すれば，それぞれのノードに型名の予測値を割り振ることができるため，各ノードに対してエッジと接続先ノードを求めることによって，そのノードの出力エッジと接続先の型名を表す RBE を得ることができる．そのため，クラスタ内のすべてのノードの RBE の和集合を取ることに

よって，各クラスタに対する RBE を得ることができ，それらを 1 つの ShEx スキーマとしてまとめる．

4 評価実験

本章では，提案手法に関する評価実験について述べる．

4.1 概　　要

実験は大きく 3 段階に分かれている．まず，SP^{2}Bench を用いて種々のパラメータによる結果の変化並びに提案手法の大まかな傾向の評価と考察を行う．便宜的にこれを予備実験と称する．次に予備実験の結果を踏まえて，WatDiv によって SP^{2}Bench のごく一部を模倣するような提案手法で少なくとも良い結果が得られると推測される RDF を作成し，良い結果が得られること確かめる．これを便宜的に確認実験と称する．最後に 2 つの実験の結果を踏まえて，WatDiv を用いて RDF の内容を変化させ，本手法で良い結果が得られるデータの特徴について詳しく検証する．さらに実行時間を計測しその変化の様子を確認する．これを便宜的に本実験と称する．

SP^{2}Bench および WatDiv の詳細については次節に記述する．

実験の環境は下記の通りである．

プロセッサ：Apple M2 Max
主記憶：32GB
OS：macOS Ventura 14.0
実装言語：Python 3.11.3
各種ライブラリはグラフデータ関連には networkx3.2.1，埋め込みには gensim 4.3.2，クラスタリングには pyclustering 0.11.0 を用いた

4.2 利用するデータ

評価実験では分類精度を評価する必要があるため，それぞれのノードの rdf:type プロパティの値をそのノードの正解ラベルとして扱う．正解ラベルを元にクラスタを形成し，スキーマを得ることで本研究で目的とするようなスキーマ抽出は可能であるが，これは対象とするデータがベンチマーク用に作られた RDF だから可能な方法である．現実の RDF やグラフデータではこのような整ったデータは必ずしも多くない．また，正解ラベルは評価する目的にのみ用い，ランダムウォーク等では値を直接参照することはしない．

4.2.1 SP^{2}Bench [11]

予備実験に利用するデータとして，SP^{2}Bench(SPARQL Performance Benchmark) のデータ生成器を利用する．SP^{2}Bench のデータ生成器では，コンピュータ科学に関する書誌学ウェブサイトである DBLP にしたがった RDF が出力され，データサイズはトリプルの数または単位を KB とするファイルサイズで指定できる．例として，SP^{2}Bench で生成される RDF をグラフとして表現したものを [11] より引用して図 4 に示す．(図 4 では rdf:type のラベルがついたエッジは点線で表現されている)

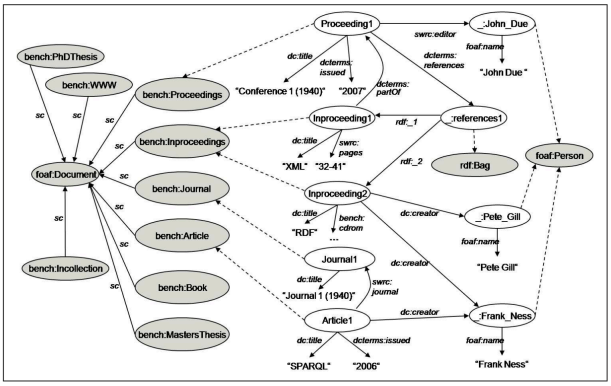


図 4　SP^{2}Bench で生成されるデータの例，[11] より引用

ただし，本研究においては，型間の継承関係を記述することまでは目的としていないため，図 4 左の foaf:Document と複数の型間における継承を表すトリプルは無視した．

4.2.2 WatDiv [12]

確認実験および本実験に利用するデータは WatDiv(Waterloo SPARQL Diversity Test Suite) のデータ生成器を用いて生成したする．

WatDiv では WatDiv スキーマ [13] と呼ばれる独自のスキーマ言語によって記述された任意のスキーマにしたがってデータが生成される．データサイズについては，終端を除くそれぞれの型について，WatDiv スキーマに記述したノード数に任意のスケールファクターを乗じた数のノードを生成できる．これらの特徴により WatDiv によって生成されるデータは自由度が非常に高い．

リテラルの値について，人名やランダムな文字列など複数のバリエーションが存在するが，提案手法ではこれらを直接参照することはしておらず，どのような値でもスキーマに対して影響を与えないため，正解率を評価するための正解ラベルとなる rdf:type の目的語を除き全て最小値 1，最大値 1000000 のランダムな数値にすることとした．

4.3 予備実験

4.3.1 概　　要

提案手法では skip-gram と clustering の二つの機械学習手法を用いており，さらにランダムウォークでは乱数に依存する処理を行っているため，事前の結果予測が容易ではない．また，クラスタリング結果に基づいてラベルを集計してスキーマを得るため，1 つでもノードを誤分類すると期待通りのスキーマが得られないという非常に厳しい要件が存在する．これらの事情から，本格的な実験を行う前におおよそどのような結果が得られるか評価・考察をする．

また，提案手法には 3 章で示したようにオプションを含めると 5 種類 (ただし，skip-gram の Context size k にランダムウォークの探索の長さ l を与えることを考慮しているため，より厳密には 6 種類) のパラメータが必要であり，これらを適切

に調整しなければならない．

以上の二点の目的で予備実験を行う．具体的には，ファイルサイズ約 10 MB，埋め込みおよびクラスタリング対象のノード数 18585 の SP^{2}Bench に対して下記に示す numpy の linspace メソッドを利用することで作成したパラメータを総当たりで実行し，次小節で述べる評価方法に則って評価指標とする正解率を計算し，結果を観察する．

$$d = [5, 28, 52, 76, 100]$$
$$w = [2, 16, 31, 45, 60]$$
$$l = [2, 5, 8, 11, 15]$$
$$p = [0.1, 1.05, 2.0]$$
$$parent_width = [0, 1]$$

4.3.2 評価方法

予備実験の結果を評価するための評価指標として，後述の正解率を用いる．また，正解率計算の対象とするノードは埋め込みおよびクラスタリング対象のノードのみとした．(埋め込み対象でない終端ノードは解析的に求められるので自明に正解率は 1)

本問題においては実行結果のクラスタ数と正解のクラスタ数が一致しない場合がある．また，実行結果はあくまでクラスタが得られるだけであり，そこに予測ラベルを割り当てる作業は結果を評価するという目的において別途必要である．よって今回は次に述べる手順によってクラスタに予測ラベルを割り当てることとした．

まずグラフのノード集合 V の部分集合として $V_{t_i} = \{v \in V \mid \lambda(v) = t_i\}$ を定義し，V_{t_i} について $\mu(v \mid v \in V_{t_i})$ を集計する．集計によって最も出現頻度の高いクラスタ名 c_k を得る．続いて同じく V の部分集合 $V_{c_k} = \{v \in V \mid \mu(v) = c_k\}$ を定義し，$\lambda(v \mid v \in V_{c_k})$ を集計，最も出現頻度の高い型名 t_{c_k} を得る．この時に $t_i = t_{c_k}$ ならば，クラスタ c_k に予測ラベルとして t_i を割り当てる．ただし，λ は 2.2 節で述べたノードを型名に写すような関数で，μ はノードをクラスタ名集合 C に写すような関数 $\mu : V \to C$ である．この処理は列名を $t_1, t_2 \ldots$，行名を $c_1, c_1 \ldots$ とする集計表を事前に作成しておくことで定数時間で実行することができ，すべての列名に対して処理を行うことになるため，$O(|\Gamma|)$ で実行できる．また，この処理の結果予測ラベルを持たないクラスタが発生しうるが，その場合はそのクラスタに属するノードはすべて不正解として扱う．

予測ラベルを割り当てた上で下記の数値を計算することで正解率を得る．

$$\text{正解率} = \frac{\sum_{v \in V} \gamma(v)}{|V|}$$

$$\gamma(v) = \begin{cases} 1 & \textbf{if } \lambda(v) = \mu(v) \\ 0 & \textbf{if } \lambda(v) \neq \mu(v) \end{cases}$$

4.3.3 結果・考察

予備実験の結果について正解率の上位 10 件と下位 10 件を表 1 に示す．正解率の有効数字は 4 桁で切り捨てとした．上位と下位の境は点線で表示している．

表 1　予備実験の実行結果．正解率上位 10 件，下位 10 件を表示

d	w	l	p	parent_length	正解率
28	31	2	2	0	0.9943
52	60	8	2	0	0.9911
76	60	2	1.05	1	0.9907
100	60	2	0.1	1	0.9877
100	45	5	1.05	0	0.9876
52	45	8	1.05	1	0.9876
28	31	5	2	0	0.9876
100	60	2	2	0	0.9876
52	31	5	2	0	0.9876
76	31	5	2	0	0.9876
5	2	5	0.1	1	0.2659
5	2	5	1.05	1	0.2587
5	2	5	0.1	0	0.2500
5	2	2	1.05	0	0.1692
5	2	5	2	1	0.1631
5	2	2	1.05	1	0.1518
5	2	2	0.1	1	0.1500
5	2	2	2	0	0.1416
5	2	2	0.1	0	0.1343
5	2	2	2	1	0.1265

また，列名に型名 (正解ラベル)，行名にクラスタ名 (予測ラベルが割り当てられた場合は予測ラベル) を取るような集計表を作成すると表 2 のような結果が得られる．ただし，横幅の都合で型名とクラスタ名について名前空間が bench の場合は名前空間を省略した．

表 2　$d = 28, w = 31, l = 2p = 2, parent_length = 0$ (正解率 1 位) で実行した時の集計表

cluster \ class	foaf:Person	Article	Inproceedings	Journal	Incollection	Proceedings	rdf:Bag	Book	foaf:Document
foaf:Person	7978	0	0	0	0	0	0	0	1
Article	0	7114	0	0	0	0	0	0	0
Inproceedings	0	8	3066	0	87	0	0	0	0
Journal	0	0	0	186	0	1	0	0	0
Proceedings	0	0	3	0	2	76	0	3	0
rdf:Bag	0	0	0	0	0	0	60	0	0

はじめに，これらの結果を概観し提案手法の大まかな傾向について述べる．正解率上位 10 件全てに共通して割り当てることができた予測ラベルは foaf:Person，bench:Article，bench:Inproceedings，bench:Journal の 4 つの型であった．これらの型はその型に属するノード数が多い型の上位 4 件と一致している．このことから，提案手法では，相対的にノードの数が多いとその型についてはクラスタリングがうまく行われやすいと言える．よって，確認実験では出現確率が高い接続先ノードがこの 4 つの型の範囲とリテラルのみで記述できる foaf:Person，bench:Article，bench:Journal の 3 つ型とリテラルで構成される RDF を，提案手法で良い結果が得られることが期待できるデータとして用いることとした．

次にパラメータについて述べる．先述の通り乱数等の影響で結果は多少変化するため，単純に最も正解率の高かった組み合わせを最適な組み合わせとすることは適当ではない．よって今回はそれぞれのパラメータについて上位 10 件に出現した回数が

最も多い数値を基準として，特に上位と下位で出現する数値に大きな差がある w について追加の予備実験を行った．また計算負荷低減を理由として上位 10 件での出現回数が重複するパラメータは小さい方の数値を採用した．出現回数が多かったパラメータは $d = 52, w = 31, l = 2, p = 2, parent_lengt = 0$ である．追加予備実験ではこの数値を基準として w のみを numpy の linspace メソッドで作成した下記に示す値 w' で変化させ，それぞれ 5 回ずつ正解率を計測しその平均を求めた．計測結果は図 5 の通りである．

$$w' = [2, 7, 12, 17, 22, 27, 32, 38, 43, 48,$$
$$53, 58, 63, 69, 74, 79, 84, 89, 94, 100]$$

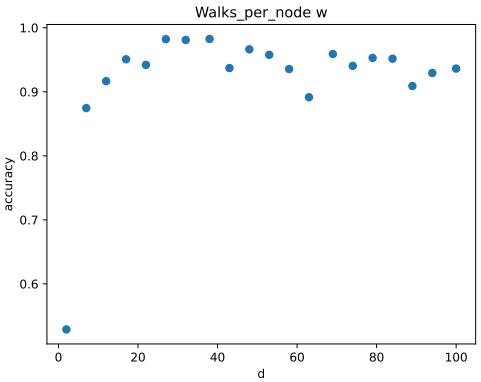


図 5　追加予備実験の結果，w 以外のパラメータは本文記載の予備実験の結果を元にした数値を用いた．

図 5 に示した結果から $d = 38$ までは概ね正解率が増加しており，それ以後は緩やかではあるが概ね減少に転じることがわかる．このことから，最高点である $d = 38$ を以後の実験で用いるパラメータとして利用することとした．

以上で述べた予備実験および追加の予備実験の結果・考察をまとめる．

- 提案手法では，相対的にノードの数が多い型についてクラスタリングがうまく行われる傾向がある
- パラメータについて，下記の数値で正解率が大きくなることが期待される

$$d = 52, w = 38, l = 2, p = 2, parent_length = 0$$

4.4 確認実験

4.4.1 作成した RDF

確認実験では WatDiv を用いて作成した単純な構造の RDF に対して提案手法を実行する．作成した RDF は小節 4.3.3 で述べたとおり foaf:Person，bench:Article，bench:Journal と終端型によって構成する RDF である．少なくとも良い結果が得られると思われるデータにおいては提案手法が良い結果を得られることを確認する趣旨であるため，プロパティはクラスタの中心に近いノードが共通して持つものに限定した．

プロパティ決定方法について具体的に述べる．まず，予備実験と同じ 10 MB の SP^{2}Bench に対して，予備実験の結果得られたパラメータ組，$d = 52, w = 38, l = 2, p = 2, parent_length = 0$ で処理を実行する．その後それぞれのクラスタについて，クラスタの中心から距離が遠いノードを除外してき，中心に近いノードのみを集計する．この結果，それぞれのクラスタの中心から最も遠いノードと中心との距離を 1 とした時，距離 0.25 以内のノードのみを集計し，列をプロパティ検討対象である foaf:Person，bench:Article，bench:Journal のみにすることによって表 3 のような正解率を計算すると 1 になる集計表が得られた．

表 3　クラスタの中心から距離があるノードを除外した時の集計表，除外されたノードの数は removed node の行に記載した

cluster \ class	foaf:Person	bench:Article	bench:Journal
foaf:Person	1263	0	0
bench:Article	0	2371	0
bench:Journal	0	0	30
removed node	6715	4751	156

表 3 に示したクラスタから除外されていないノードについてラベルを集計することで妥当なスキーマを求めると図 6 のようになる (横幅の都合上型名の名前空間は省略した)．ただし，スキーマの定義域として除外されたノードが含まれる場合は除外されたノード以外で定義域を設定した．

$$\begin{aligned}
\Sigma = \{&dc:creator, dc:title, dcterms:issued,\\
&foaf:name, foaf:homepage, rdf:type,\\
&rdfs:seeAlso, swrc:editor, swrc:journal,\\
&swrc:number, swrc:pages, swrc:volume\}\\
\Gamma =&\{t_0, Person, Article, Journal\}\\
\delta(t_0) =&\epsilon\\
\delta(Person) =&rdf:type :: t_0 || foaf:name :: t_0\\
\delta(Article) =&rdf:type :: t_0 || (rdfs:seeAlso :: t_0)? || dc:title :: t_0 ||\\
&(dc:creator :: Person) + || (foaf:homepage :: t_0)? ||\\
&swrc:journal :: Journal || (swrc:pages :: t_0)? ||\\
\delta(Journal) =&rdf:type :: t_0 || dc:title :: t_0 || dcterms:issued :: t_0 ||\\
&swrc:volume :: t_0 || (swrc:number :: t_0)? ||\\
&(swrc:editor :: Person)*
\end{aligned}$$

図 6　表 3 の結果を元に求めたスキーマ

この結果から提案手法で良い結果を得られると推測される単純な構造のデータとして，図 6 のようなスキーマ構造を持つデータを用いることとした．また，それぞれのプロパティの出現確率については原則 10 MB の SP^{2}Bench における実測値に従うこととした．またノード数の比について，単純に有効数字 1 桁 (2 桁目四捨五入) で SP^{2}Bench における foaf:Person，bench:Article，bench:Journal の比を求めると 70 : 80 : 2 となる．しかし，SP^{2}Bench において，foaf:Person は bench:Article，bench:Journal 以外の接続先ノードになることもあるため，これを補正する目的で WatDiv に与える比率は 70 : 40 : 2 とした．

4.4.2 結果・考察

確認実験では，予備実験の結果を踏まえて WatDiv で生成した図 6 のようなスキーマを持つノード数データについて提案手法を実行した．クラスタリングについての集計表を示すと表 4 のようになる．また，クラスタに対する予測ラベルは小節 4.3.2 で述べた方法で付与しており，この結果では予測ラベルを持たないクラスタ non_labeled が発生している．

表 4　実行結果のクラスタリング集計表

cluster \ class	foaf:Person	bench:Article	bench:Journal
foaf:Person	20000	0	0
bench:Article	0	18121	0
bench:Journal	0	0	1000
non_labeled	0	16879	0

このように bench:Article が 2 つのクラスタに別れてしまう結果となった．しかし，$\delta(Article) = \delta(non_labeled)$ であったため，単に同じ RBE を持つクラスタを結合することによってスキーマ抽出は可能である．またこの二つのクラスタを結合した結果得られるスキーマは WatDiv のデータ生成器に与えたものと全く同一である．よってスキーマ抽出を目的とする提案手法は期待通りに動作したと言える．これらの結果から，予備実験の結果を踏まえた提案手法との相性が良いと推測されるデータにおいては，提案手法が期待通りに動作することがわかった．調節した相性が良いと推測されるデータの特徴は下記の通りである．

（1） それぞれの型のノード数に大きな偏りがない (確認実験での WatDiv の設定値 70 : 40 : 2)

（2） すべてのプロパティは出現率が小さくない (確認実験での WatDiv の設定値 0.68 以上)

4.5 本 実 験

本実験で用いる RDF について述べる．予備実験の結果から 10MB の SP^{2}Bench では数の少ない型はよくクラスタリングすることが難しいことがわかっている．そこで 1MB の SP^{2}Bench に着目することとした．1MB の SP^{2}Bench は 10MB の SP^{2}Bench とデータ構造が異なっており，一部のプロパティや型が削減され，より軽量な構造となっている．さらにいくつかの目的のために 3 点変更を加えた．1 点目は数の少ない型を排除する目的で foaf:Document 型を削除した点である．2 点目は rdf:Bag のプロパティを rdf:_1, rdf:_2, . . . という表現から，rdf:Bag は順序を考慮しないことを念頭に全てを rdf:_n というプロパティに統合した点である．3 点目は最も構造が単純な foaf:Person 型について，これを複雑にする目的で bench:friend プロパティを追加した点である．またこの bench:friend プロパティは必ずお互いを相互に参照するように設定した．各プロパティの出現率は原則 1MB の SP^{2}Bench の実測値に従い，スケールファクターを乗じる前のノード数は 1MB の SP^{2}Bench での出現数をそのまま設定した．これらの条件によって作成した RDF では良い結果を得ることができなかったため，確認実験で得られた良い結果につながると推測される条件を適用した．具体的には下記の 2 つの操作を行なった．

（1） 出現率 0.68 以下の各プロパティの出現確率について定義域がリテラルかそれ以外かに分けて以下の操作を行う

- リテラル：プロパティを削除
- リテラル以外：出現率を 0.8 に引き上げ

対応を分けた意図はリテラル以外の型間の情報を欠落させることでデータ構造がより単純になってしまうことを避けることである．

（2） スケールファクターを乗じる前のノード数について，有効数字 2 桁として $n \times 10^2$ の形で表現し $n < 1$ となる場合には $n = 1$ とする

ここからさらに下記の 3 点の修正を加えると良い結果を得る RDF を作成することができた．また修正を加えたデータにおいては確認実験のような過剰分割によって同じ RBE を持つクラスタが発生することもなくクラスタリングの正解率も 1 であった．

（1） bench:Inproceedings と bench:Incollection が混同してしまうため bench:Incollection を削除

（2） bench:Article について dcterms:references プロパティの出現有無によって過剰分割が発生したため，dcterms:references プロパティの出現率を 1 に設定

（3） bench:Journal と bench:Proceedings にはプロパティの重複が多くあり，混同することがあった．そのため，bench:Proceedings の swrc:editor プロパティを bench:editors プロパティに名前を変更した．

これらの操作を経た本手法で良い結果を得ることが期待できてかつ極力 SP^{2}Bench に近いスキーマを図 7 に，このスキーマを持つ複数のサイズのデータに対して提案手法を実行した結果，並びに実行時間の変化の折れ線グラフをそれぞれ表 5，図 8 に示す．また抽出されたスキーマは図 7 と同一である．時間計測は有効数字は 4 桁で切り捨てとして 2 回行い，その平均を示した．ShEx スキーマによる記法には含まれないが，それぞれの型のノード数の比は下記の通りである，ただし横幅の都合で型名の名前空間は省略した

$$Article : Inproceedings : Journal : Proceedings : Person : Bag$$
$$=90 : 17 : 10 : 10 : 90 : 10$$

$$
\begin{aligned}
\Sigma = \{&dc:creator, dc:title, dc:publisher,\\
&dcterms:issued, dcterms:references,\\
&dcterms:partOf, foaf:name,\\
&foaf:homepage, rdf:type, rdf:_n\\
&rdfs:seeAlso, swrc:editor, swrc:journal,\\
&swrc:number, swrc:pages, swrc:volume,\\
&bench:booktitle\}\\
\Gamma =\{&t_0, Person, Article, Inproceedings,\\
&Journal, Proceedings, Bag\}\\
\delta(t_0) =&\epsilon\\
\delta(Person) =&rdf:type::t_0||foaf:name::t_0||\\
&(bench:friend::Person)+\\
\delta(Article) =&rdf:type::t_0||dc:title::t_0||\\
&(dc:creator::Person)+||\\
&dcterms:references::Bag||\\
&(foaf:homepage::t_0)?||\\
&swrc:journal::Journal||(swrc:pages::t_0)?||\\
\delta(Inproceedings) =&rdf:type::t_0||(rdfs:seeAlso::t_0)?||\\
&dc:title::t_0||(dc:creator::Person)+||\\
&dcterms:issued::t_0||\\
&(dcterms:partOf::Proceedings)?||\\
&(dcterms:references::Bag)?||\\
&foaf:homepage::t_0||\\
&swrc:journal::Journal||(swrc:pages::t_0)?||\\
\delta(Journal) =&rdf:type::t_0||dc:title::t_0||\\
&dcterms:issued::t_0||swrc:volume::t_0||\\
&(swrc:number::t_0)?||(swrc:editor::Person)*\\
\delta(Proceedings) =&rdf:type::t_0||dc:title::t_0||dc:publisher::t_0||\\
&dcterms:issued::t_0||(swrc:isbn::t_0)?||\\
&swrc:series::t_0||(foaf:homepage)?||\\
&bench:booktitle::t_0||(bench:editors::Person)+\\
\delta(Bag) =&rdf:type::t_0||\\
&(rdf:_n::(Article|Inproceedings))+
\end{aligned}
$$

図 7　良い結果を得ることができたスキーマ

表 5　実 行 結 果

node size	literal node	正解率	time [s]
11350	27217	1	1.650×10
22700	53813	1	3.320×10
28375	66858	1	4.147×10
56750	129001	1	8.510×10

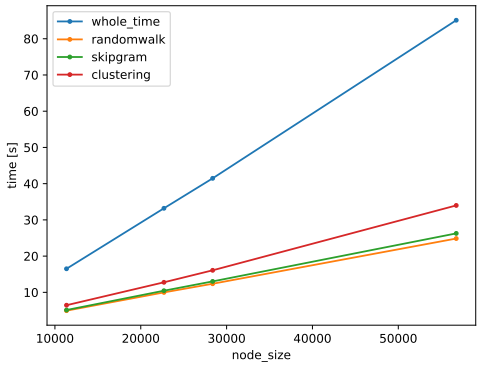


図 8　実行時間の推移

5　む　す　び

本研究では，ラベル付き有向グラフにおけるスキーマ抽出という問題に対して，ノードの埋め込みによってノード間の類似度計算を効率化しクラスタリングを行うという既存研究とは異なるアプローチを試みた．入力データには SP^{2}Bench や WatDiv によって生成したデータを用い，出力は ShEx スキーマを対象とした．結果として，型間でノードの数に偏りがない，各プロパティの出現率が低くないなどの条件に適合するデータであれば正しくスキーマが抽出できることがわかった．もちろん実際のグラフデータにおいてはこれらの条件は必ずしも事前にわかるとは限らず，提案手法は汎用性があるとは言えない．しかし，実行時間がおよそ線形に推移し条件に適合していれば適切にスキーマを得られることから，提案手法はスキーマを得たいグラフデータに対して実行してみる手法の選択肢の一つにはなり得ると言える．

今後の課題として，本研究では提案手法と相性の良いデータの特徴を明らかにしたが，これは絶対的な特徴であるから，既存研究と比較したときに相対的にどのようなデータと相性が良いのかを調べることが考えられる．また，提案手法では，クラスタリングの正解率を 1 にし，正解と全く同一のスキーマを抽出した．しかし，正解ラベルが異なる型であっても構造が類似していれば同一の型とみなしてスキーマを抽出することも有用であると考えられる．そのため，どの程度構造が類似していれば同一の型みなすべきか，また，そのようにスキーマを抽出するためにはどのような手法が適切か検討することも考えられる．

文　献

[1] Yuroti Tsuboi and Nobutaka Suzuki, “An Algorithm for Extracting Shape Expression Schemas from Graphs,” ACM Symposium on Document Engineering 2019 (DocEng ’19), no.32, pp.1-4, New York, USA, September 2019. DOI:10.1145/3342558.3345417
[2] Paton N, Chen J and Wu Z, “Dataset Discovery and Exploration: A Survey,” ACM Computing Surveys, 56,4, pp.1-37. November 2023. DOI:10.1145/3626521
[3] Kellou-Menouer K, Kardoulakis N, Troullinou G, Kedad Z,

Plexousakis D and Kondylakis H, “A survey on semantic schema discovery,” The VLDB Journal, 31,4, pp.675-710, November 2021. DOI:10.1007/s00778-021-00717-x

[4] Angela Bonifati, Stefania Dumbrava, and Nicolas Mir. “Hierarchical clustering for property graph schema discovery,” EDBT 2022:25th International Conference on Extending Database Technology, no2, pp449-453, Edinburgh, United Kigndom, March 2022. DOI:10.48786/edbt.2022.39

[5] Nikos Kardoulakis, Kenza Kellou-Menouer, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, and Haridimos Kondylakis, ”HInT : Hybrid and Incremental Type Discovery for Large RDF Data Sources,” 33rd International Conference on Scientific and Statistical Database Management (SSDBM 2021), pp.97-108, Tampa, USA, July 2021. DOI:10.1145/3468791.3468808

[6] 坪井 悠冬里, ”ラベル付き有向グラフに対する Shape Expression Schema の抽出,” 修士（情報学）, 2020-03-25, 第 43150 号

[7] Thomas Baker and Eric Prud’hommeaux, ”Shape Expressions (ShEx) Primer,” http://shexspec.github.io/primer/, 最終参照 2024.1.8

[8] Slawek Staworko, Iovka Boneva, Jose Emilio Labra Gayo, Samuel Hym, Eric G. Prud ’hom-meaux, Harold R. Solbrig, “Complexity and Expressiveness of ShEx for RDF,” 18th International Conference on Database Theory(ICDT 2015), pp.195-211, Brursseles, Belgim, March 2015. DOI:10.4230/LIPIcs.ICDT.2015.195

[9] Mikolov, T., Chen, K., Corrado, G., and Dean, J, ”Efficient estimation of word representations in vector space,” 2013. DOI:10.48550/arXiv.1301.3781

[10] Grover, A., and Leskovec, J, ”node2vec: Scalable feature learning for networks,” the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.855-864, San Francisco, USA, August 2016. DOI:10.1145/2939672.2939754

[11] Michael Schmidt and Thomas Hornung and Georg Lausen and Christoph Pinkel, ”SP2Bench: A SPARQL Performance Benchmark,” 2008. DOI:0.48550/arXiv.0806.4627

[12] G. Aluç, O. Hartig, M. T. Őzsu and K. Daudjee, ”Diversified Stress Testing of RDF Data Management Systems,” The Semantic Web - ISWC 2014: 13th International Semantic Web Conference, pp197-212, Riva del Garda, Italy, October 2014. DOI:10.1007/978-3-319-11964-9_13

[13] G. Aluç, ”WatDiv Dataset Description Language Tutorial,” Data Systems Group at the University of Waterloo’s Cheriton School of Computer Science, https://dsg.uwaterloo.ca/watdiv/watdiv-schema-tutorial, 最終参照 2024.1.8

プリミティブなグラフ分析アルゴリズムに対する効率的なグラフ要約の応用

真次　彰平†　　塩川　浩昭†

† 筑波大学計算科学研究センター　〒 305-8571 茨城県つくば市天王台 1-1-1
E-mail: †matsugu@kde.cs.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　大規模なグラフデータを効率的に格納するために，近年グラフ要約が注目を集めている．グラフ要約はグラフ中のいくつかのノード，およびエッジをひとつにまとめることにより，グラフサイズを削減する技術である．近年提案されたグラフ要約手法はグラフを保持する空間コストの削減に主眼が置かれており，要約されたグラフの上で動作するアルゴリズムについての議論はあまり活発でない．そこで本研究では幅広いグラフ分析手法において用いられる，いくつかのプリミティブな処理についてグラフ要約を用いた効率的な実装を提案する．本稿では実データを用いた実験を行い，グラフ要約が幅広いアプリケーションの効率性に貢献することを確認した．

キーワード　グラフデータ処理，データ構造・索引，グラフ要約

1　はじめに

近年ではウェブやソーシャルネットワーキングサービスの発展に伴い，大規模なソーシャルネットワークが数多く登場している [1,2]．例えば，2020 年時点での Facebook のユーザ数は 27 億人を超えたと報告されており，これら実世界に存在するソーシャルネットワークの規模は現在もなお増加し続けている [3]．これらのソーシャルネットワークを表現した大規模グラフから有用な知識を抽出する大規模グラフ分析 [4–7] は近年基礎・応用問わず幅広い研究課題が提案され，多くの関心を集めている．しかしながら，大規模グラフはグラフの保持に要する空間コストが大きく，一般的な性能の計算機ではそのようなグラフ分析アルゴリズムの実行は困難であるという問題がある [8–10]．

この問題を解決するために，グラフ要約 [11, 12] が注目を集めている．グラフ要約はグラフ中のノードやエッジをひとつにまとめることによりグラフサイズを削減する技術である．図 1 はグラフ要約の概念図である．グラフ要約では図 1 左側に示すグラフを，図 1 右側の S_1〜S_6 ように，特定のノード集合とエッジ集合をまとめたスーパーノード，スーパーエッジに変換する．これにより元のグラフの特徴を残しつつ，グラフサイズを削減する．グラフ要約はグラフの構造を保ちつつデータ容量を圧縮できるため，巨大なグラフデータをメモリ上に格納して処理を実行するための新しいデータ構造として期待されている．

近年のグラフ要約に関する研究はその圧縮性能についての議論が盛んである一方で，要約されたグラフの上で動作するアルゴリズムについての議論はあまり活発でない．例えば，クラスタリングや経路分析といったグラフ分析アルゴリズムを要約されたグラフに対して実行する場合，要約された内部の情報を逐次展開しながら従来と同等の処理を行う必要がある．そのため，既存のグラフ要約手法は巨大なグラフデータの保持に要する空間コストを小さくすることができる一方で，グラフデータ分析の時間コストを削減することが難しい．

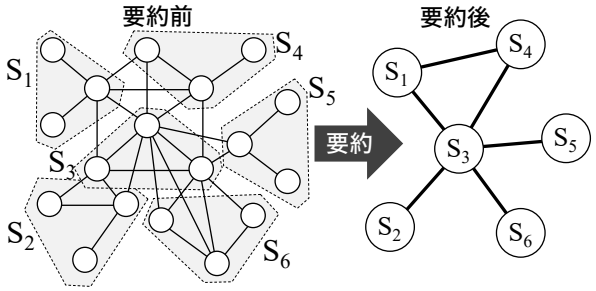


図 1: グラフ要約のイメージ図

1.1　既存研究と本研究の位置付け

既存のグラフ要約手法には大きく分けて (1) 補正セット法，(2) 効用制限法，(3) ウェッジ分割法の三つのアプローチが存在する．本節ではそれぞれのアイデアと最先端の既存手法の要約性能および要約に要する計算時間について簡単に述べる．

(1) 補正セット法: 補正セット法ではスーパーノード内部のノード集合を完全グラフ，またスーパーエッジにより隣接するノード集合の対を完全二部グラフとして元のグラフ上で可能な限り表現し，訂正が必要な部分に対して補正セットと呼ばれるエッジ集合を構築する．補正セット法に基づく最新の手法である SWeG [12] は情報欠損が無いという点で優れているが，圧縮率が十分でない上，分散コンピューティングによる実装を主軸としているため単一の計算機による実行では速度が遅いという欠点がある．

(2) 効用制限法: 効用制限法では元のグラフのノードやエッジに効用 (Utility) と呼ばれるエッジの重要性を示す値を定義し，効用の大きなノード・エッジのみを要約する．具体的にはある閾値が与えられ，グラフ全体の効用がその閾値を下回らない範囲でノード数を最小にするよう手法を設計する．効用制限法に基づく最新の手法である UDS [11] は単一のマシンでも効率的な計算が可能であるが，情報の欠損が発生する．

(3) ウェッジ分割法: ウェッジ分割法ではグラフをウェッジと呼ばれる部分構造毎に分割することで要約を行い，それらの接続

関係をビット列に変換することで．元のグラフに存在する全ての情報を高い圧縮率で要約する．ウェッジ分割法に基づく最新の手法である HYBRID [13] は単一のマシンでも SWeG や UDS と比較して効率的かつ高い圧縮率でグラフを要約することが可能である．

これらの特徴から，本稿では最先端のグラフ要約手法である HYBRID を基盤とし，HYBRID による要約グラフの上で動作するアルゴリズムの高速化に取り組む．

1.2 本研究の貢献

本稿では HYBRID による要約グラフの上で動作する効率的なグラフ分析アルゴリズムを開発する．近年のグラフ分析処理は非常に多岐に渡るが，多くの応用的なグラフ分析手法において組み込まれるいくつかのプリミティブな処理を対象とすることで，本研究が一般の大規模グラフ分析において汎用的に貢献することを目指す．

本研究が対象とするプリミティブなグラフ処理，およびその応用例を以下に示す．

Triangle Counting (TC): TC [14, 15] はグラフ中のあらゆる三ノードから作られる三角形の数を求める処理である．グラフ中の三角形はコミュニティ分析において重要な特徴量であることから，Truss Decomposition [16] やコミュニティ検索 [17, 18] といった手法において多く用いられている．

Single Source Shortest Path (SSSP): SSSP は始点となるノードから各ノードまでのホップ数を求める処理である．あるノードからエッジによって接続するノードを探す処理は最短経路や到達可能性の計算において基本的で重要な処理であり，重み付きグラフに対する Dijkstra 法 [19] や確率的グラフに対する信頼性の計算 [20] など多くの拡張が存在する．

本研究の貢献は以下の通りである．

- **効率性:** 提案手法は本研究が対象とするプリミティブなグラフ処理における一般的な実装と比較して高速である．
- **正確性:** 提案手法は要約グラフの上で効率的に計算を行うが，常に一般的な実装と同一の解を出力する．
- **拡張性:** 本研究が対象とするグラフ処理は幅広いグラフ分析手法において組み込まれているため，提案手法はプリミティブなグラフ処理だけでなく，多くの複雑なグラフ分析手法に対しても応用可能である．

提案手法は我々の知る限り最先端のグラフ要約手法を活用して大規模グラフ分析の効率化を行う最初の手法である．本研究を通じてグラフ要約は一般の幅広い大規模グラフ分析手法に対して，空間消費量のみならず計算時間についても効率化が期待できる技術であることを実証する．

2 事 前 準 備

本稿で用いる主な記号を 表 1 に定義する．本稿では連結な無向グラフ $G(V, E)$ を考える．V, E はそれぞれノード集合およびエッジ集合である．また，部分ノード集合 $V' \subseteq V$，および V' 中の 2 ノード間に存在するエッジの集合からなる部分グラフを V' の誘導部分グラフと呼び，$G[V']$ と書く．グラフ G 中のノード v の次数を $deg_G(v)$ と書き，文脈によって G が自明で有るとき単に $deg(v)$ と書く．本節ではウェッジ分割 [13] の定義を示す．まず，ウェッジは以下のように定義される．

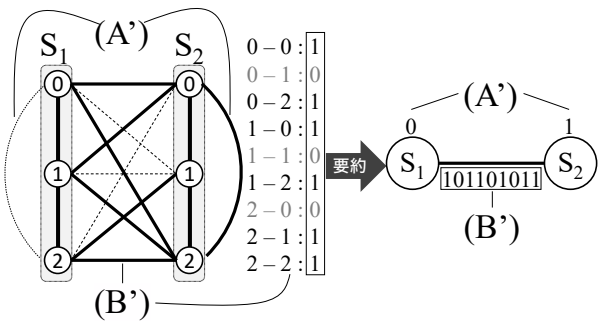


図 2: BLEW のイメージ図

定義 1 (ウェッジ) グラフ $G(V, E)$ に存在するウェッジとは，$(a, b) \in E \wedge (b, c) \in E$ を満たす 3 ノード組 $a, b, c \in V$ のことを指す．また，$(a, c) \in E$ であるものをトライアングル型，そうでないものをパス型のウェッジと呼び，あらゆるウェッジは両者のいずれかに分類される．さらに，G 中のウェッジ全体の集合を $\mathcal{W}(G)$，トライアングル型ウェッジの集合を $\mathcal{W}_{TRI}(G)$，パス型ウェッジの集合を $\mathcal{W}_{PATH}(G)$ と表す．ここで，$\mathcal{W}(G)$，$\mathcal{W}_{TRI}(G)$，$\mathcal{W}_{PATH}(G)$ はいずれも各要素に 3 ノード組を持つ集合であることに注意したい．

ウェッジは実世界のグラフで頻繁に見られる構造であるため，様々なグラフ分析に用いられる他，グラフに存在するウェッジの数はグラフの傾向を示す特徴量としてもよく知られている．

次に，元のノード id を辿らずにウェッジ内部の構造を特定するためのナンバリングとして，ローカルノード id を導入する．

定義 2 (ローカルノード id) ウェッジ s 中の次数が 2 であるノード $v \in F_w(s)$ を 1 つ選択し，v 以外の 2 ノードを $u, w \in F_w(s)$ とする．このとき，$(loc(u), loc(v), loc(w)) = (0, 1, 2)$ を s のローカルノード id と呼ぶ．また，ローカルノード id から元のノード id への写像を $loc_s^{-1} : \{0, 1, 2\} \to F_w(s)$ で表す．

ここで，一般にローカルノード id の決め方は複数存在するが，差異は無いため適当な方法でひとつに決定する．ローカルノード id を用いると，各ウェッジは $(0, 1)$ 間および $(1, 2)$ 間に必ずエッジが存在することがわかり，さらにトライアングル型であれば $(0, 2)$ 間にも存在することがわかる．そのため，各スーパーノードにパス型とトライアングル型のどちらであるかという 1 bit の情報を格納することで，スーパーノード内部のエッジを完全に保存できる．

次に，スーパーエッジの重みとして割り当てられるスーパーノード間の隣接関係を bit 列として格納するフォーマットである，BLEW [13] を定義する．

定義 3 (BLEW) ウェッジ s, t について，その間に張られるスーパーエッジの重み $blew : P \to \mathbb{R}$ は，次の式で表される．

$$blew((s, t)) = \sum_{i,j \in \{0,1,2\}, (loc_s^{-1}(i), loc_t^{-1}(j)) \in E} 2^{3i+j}.$$

表 1: 本稿で用いる主な記号

記号	意味
G	グラフ
V, E	ノード集合，エッジ集合
$V(g), E(g)$	グラフ g のノード集合，エッジ集合
$N(v)$	ノード v の隣接ノード集合
$G[V']$	ノード集合 V' の誘導部分グラフ
$deg(v)$	ノード v の次数
S	スーパーノード集合
P	スーパーエッジ集合
$\sigma : P \to \mathbb{R}$	スーパーエッジの重み
$f : V \to S$	ノード集合からスーパーノード集合への要約
$F : S \to 2^V$	スーパーノード集合からノード集合への要約
$\mathcal{F}$	$V \to S$ の要約写像全体の集合
$\mathcal{W}$	ウェッジ全体の集合
$\mathcal{W}_{TRI}$	トライアングル型ウェッジ全体の集合
$\mathcal{W}_{PATH}$	パス型ウェッジ全体の集合
$loc(v)$	ノード v のローカルノード id (定義 2)
$blew((s,t))$	ウェッジ s,t 間の BLEW (定義 3)

blew は，2 つのウェッジ間にある最大 9 エッジの有無をローカルノード id を用いて 9 bit の 2 進数に変換した値である．図 2 は BLEW の例である．2 つのウェッジ s_1 と s_2 はローカルノード id が割り振られており，s_1 はパス型，s_2 はトライアングル型である．ローカルノード id の導入により各ウェッジは $(0,1),(1,2)$ にエッジが存在することが明らかである．ここで，図 2 (A') に示されたエッジ (各 1 bit) がそれぞれのウェッジがパス型かトライアングル型かを決定する．また BLEW により s_1, s_2 間のエッジ集合を (B') に示すビット列で表現することが可能である．このビット列は厳密に 9 bit からなるため，全 $2^9 = 512$ 通りのエッジ情報を格納できる．

次に，定義 1 を用いてウェッジ分割は以下のように定義される．

定義 4 (ウェッジ分割) グラフ $G(V,E)$，に対して，次の (1) 〜 (4) に示す制約を満たす要約 f を G のウェッジ分割と呼び，f による G の要約グラフ $G'(S,P,\sigma)$ をウェッジ分割グラフと呼ぶ．すなわち，

$$f(G(V,E)) = G'(S,P,\sigma).$$

＜制約＞

（1） $S = \{f(v) | \forall v \in V \cup V\}$.

（2） $P = \{(s,t) | (\bigcup_{i \in s, j \in t, (i,j) \in E}(i,j)) \neq \emptyset\}$.

（3） $\sigma = (s,t) \in P \mapsto blew((s,t))$.

（4） $\forall s \in S, F(s) \in \mathcal{W}$.

ここで，$F(s)$ は s の元ノード集合である．

定義 4 は，f によって重み無しグラフ G が重み付きグラフ G' へと要約されることを示す．制約 (1) は，スーパーノード集合が元ノード集合の f による像の集合で表されることを示す．制約 (2) は，スーパーノード s,t の元ノード集合対 $F(s), F(t)$ の間に一本でもエッジがある場合それらのスーパーノード間にスーパーエッジを張ることを示す．制約 (3) は，スーパーエッジの重みは BLEW により定められることを示す．制約 (4) は，全てのスーパーノードがウェッジであることを示す．図 1 はウェッジ分割の例である．図 1 において，18 ノードからなるノード集合 V は f によって S へと要約される．

ここで，HYBRID においてはグラフをウェッジのみにより被覆できないとき，ダミーノードを作成して疑似的にウェッジを構築することがあるが，本研究においては一般性を失わないためそれらを考慮せずにウェッジを取り扱う．すなわち，与えられる要約グラフはすべてのノードがいずれかの一つのウェッジに属しているものとする．これはグラフを要約する上で現実的な仮定では無いが，実装上は簡単な工夫によりダミーノードをオーバーヘッド無く区別できることが示されているため大きな問題ではない．詳しくは文献 [13] を参照されたい．

3 提案手法: Triangle Counting (TC)

本節ではウェッジ分割に基づく要約グラフに対する高速な TC アルゴリズムを提案する．最初に 3.1 節に提案手法を構成する基本的なアイデアを示し，その詳細を以降の節にて説明する．本稿では補題の証明は省略する．

3.1 基本アイデア

本節で提案する TC アルゴリズムはウェッジ内部に対する三角形の計算を省略し，ウェッジ間のスーパーエッジ中の三角形の個数を効率的に数え上げる．まず，ウェッジ分割されたグラフ中に存在する三角形は (a) 一つのウェッジに収まる三角形，(b) 二つのウェッジに跨って存在する三角形，および (c) 三つのウェッジに跨って存在する三角形のいずれかに分類される．それぞれのケースについて簡単な方針を述べる．(a) について，これはトライアングル型のウェッジの数に等しいため，低コストに求まることは自明である．(b) について，ある二つのウェッジの間に存在する三角形の個数は各ウェッジがパス型，トライアングル型のどちらであるか (各 1bit)，および BLEW によって割り当てられたスーパーエッジの重み (9bit) の計 11bit の情報から一意に求まる．すなわち，$2^{11} = 2048$ 通りの計算結果を事前に求めておくことで高速にカウンティングを行う．(c) について，三つのウェッジにまたがって三角形が存在するとき，それらのウェッジはスーパーエッジにより三角形が構成されるという観測を活用する．ここで，このようなスーパーエッジによる三角形をスーパートライアングルと呼ぶ．この観測に基づき，三角形を構成するスーパーエッジの重みを用いて bit 演算により三角形の個数を計算する．例えば，各ウェッジのローカルノード id が 0 である三ノードから構成される三角形の有無はそれぞれの BLEW の最初の 1bit の AND 演算により求まる．これを最大 $3^3 = 27$ 個の存在し得る三角形に対して同時に判定することで各スーパートライアングルに対して高速に数え上げを行う．これにより，この問題は要約グラフ上のスーパートライアングルを数え上げる問題へと帰着される．要約グラフは元のグラフより小さいことから，提案手法は効率的な Triangle Counting を実現する．

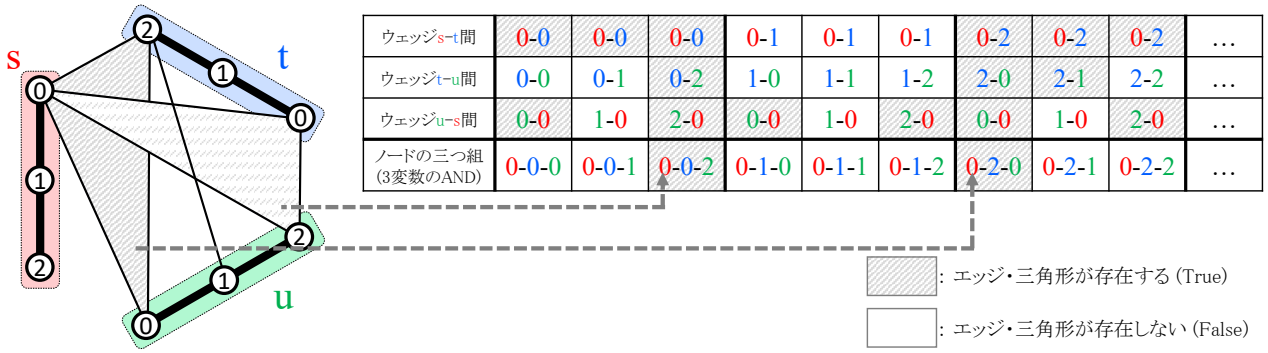


図 3: 三つのウェッジから三角形を数え上げる bit 演算.

3.2 BLEW に基づく高速な数え上げ

本節では 3.1 節で述べたアイデアに基づき，ウェッジ分割されたグラフ中に存在する三角形を三つのケースに場合分けし，それぞれについて理論的な数え上げ方を導出する．

(a) 一つのウェッジに収まる三角形: 定義 1 より，その個数は明らかに $|\mathcal{W}_{TRI}(G)|$ に等しい．この値はトライアングル型であるウェッジを数え上げることで求めるため，(a) は $O(|S|)$ 時間で求まる．

(b) 二つのウェッジに跨って存在する三角形: このケースは二つのウェッジと一つのスーパーエッジ上に収まって存在する三角形と言い換えられる．すなわち，二つのウェッジおよびそれらを繋ぐスーパーエッジの状態を表現する計 11bit の情報から三角形の個数は決定的に求まる．$2^{11} = 2048$ 種類の各パタンにおける三角形の個数を事前に計算することで，各スーパーエッジに対して $O(1)$ 時間で求まる。したがって、(b) は全体で $O(|P|)$ 時間で計算できる．ここで，簡便のために次の記号を定義する．

定義 5 (BLEW11) $blew(s,t)$ が $b_0b_1b_2b_3b_4b_5b_6b_7b_{8(2)}$ と二進数表記されるウェッジ s,t について，BLEW11 は次の 11bit の値を示す関数である．

$$blew11(s,t) = b_0b_1b_2b_3b_4b_5b_6b_7b_8\triangle_s\triangle_{t(2)}.$$

ただし，$\triangle_s, \triangle_t$ はそれぞれ s,t がトライアングル型ウェッジであるとき 1, そうでないとき 0 である．またその逆関数として，整数 $0 \leqq n < 2^{11}$ に対して $blew11(s,t) = n$ となるようなウェッジ s,t およびスーパーエッジ (s,t) のタプルを $blew^{-1}(n) = (s,t,(s,t))$ と表す．

定義 5 を用いて，(b) の事前計算は全ての整数 $n(0 \leqq n < 2^{11})$ に対して $blew^{-1}(n)$ 中に存在する三角形を全列挙することと言い換えられる．ここで，重複を防ぐため一つのトライアングル型ウェッジにより作られる三角形は除外することに注意したい．この事前計算は入力されるグラフに依存しないため，定数時間で処理される．

(c) 三つのウェッジに跨って存在する三角形: このケースは三つのウェッジから一つずつノードを選択したときにできる三角形と言い換えられる．したがって，一つのウェッジから複数のノードを用いることはないため，このケースにおいてはウェッジ内部の情報は不要である．三つのウェッジの各ノードから一つずつノードを選択した際の三角形の存在判定は，三変数による AND 演算により行う．具体的に必要となる bit 演算を図 3 に示す．図 3 はウェッジ s,t,u に跨るスーパーエッジ $(s,t),(t,u),(u,s)$ が与えられたとき，それらから全ての三角形のパタンを数え上げる演算方法を示している．紙面の都合上，ここではウェッジ A のローカルノード id が 0 のノードを含む三角形についてのみ図示する．全ての三角形のパタンを列挙するために，ここでは各 9bit からなる BLEW を冗長性を持たせた各 27bit のビット列に変換することで，3 変数の論理積の各ビットを三角形の有無に対応させる．具体的には，次に定義する BLEW27 を用いて，ビット列を拡張する．

定義 6 (BLEW27) $blew(s,t)$ が $b_0b_1b_2\ b_3b_4b_5\ b_6b_7b_8$ と二進数表記されるウェッジ s,t について，BLEW27 は次の三つの関数の組である．

$$\begin{aligned}
blew27_A(s,t) = {} & b_0b_0b_0\ b_1b_1b_1\ b_2b_2b_2 \\
& b_3b_3b_3\ b_4b_4b_4\ b_5b_5b_5 \\
& b_6b_6b_6\ b_7b_7b_7\ b_8b_8b_8 \\
blew27_B(s,t) = {} & b_0b_1b_2\ b_3b_4b_5\ b_6b_7b_8 \\
& b_0b_1b_2\ b_3b_4b_5\ b_6b_7b_8 \\
& b_0b_1b_2\ b_3b_4b_5\ b_6b_7b_8 \\
blew27_C(s,t) = {} & b_0b_3b_6\ b_0b_3b_6\ b_0b_3b_6 \\
& b_1b_4b_7\ b_1b_4b_7\ b_1b_4b_7 \\
& b_2b_5b_8\ b_2b_5b_8\ b_2b_5b_8
\end{aligned}$$

定義は 6 は三つのスーパーエッジに基づく BLEW を規則に従って 3 倍の長さに拡張することで，各桁の組み合わせが重複無く全てのノードの三つ組を列挙できるようにする関数である．具体的には，スーパーノード s,t,u に対して，$AND(blew27_A(s,t), blew27_B(t,u), blew27_C(u,s))$ を計算することで，三角形が存在したパタンに対応する桁のみが立ち上がる．図 3 には二つの三角形 (0-0-2) と (0-2-0) が存在する．これらを構成するエッジの三つ組はそれぞれ ((0-0),(0-2),(2-0)) と ((0-2),(2-0),(0-0)) であるため，これらに対応する桁のみが論理積において真値を取っていることがわかる．

定義 6 を用いて立ち上がったビット数を数えることで，三角形の個数は各スーパートライアングルに対して $O(1)$ 時間で求

Algorithm 1 ウェッジ分割に基づく TC

Input: グラフ G のウェッジ分割 $G'(S, P, \sigma)$;
Output: 三角形の個数 t;

```
1: t ← 0;
2: // (a) 一つのウェッジ
3: for each s ∈ S do
4:     if s ∈ W_TRI(G) then
5:         t ← t + 1;
6:
7: // (b) 二つのウェッジ
8: τ(n) ← 三角形数を n = 0, 1, ..., 2047 について計算する;
9: for each (s, t) ∈ P do
10:    t ← t + τ(blew11(s, t, (s, t)));
11:
12: // (c) 三つのウェッジ
13: 𝒯 ← G' 中のスーパートライアングル [14, 15];
14: for each (s, t, u) ∈ 𝒯 do
15:    b_st ← blew27_A(s, t);
16:    b_tu ← blew27_B(t, u);
17:    b_us ← blew27_C(u, s);
18:    t ← t+BIT_COUNT(AND(b_st, b_tu, b_us));
19: return t;
```

まる．グラフ全体においては全てのスーパートライアングルに対して逐次この処理を実行すれば良く，これはウェッジ分割グラフ上のスーパートライアングルに対する TC に帰着される．一般的な TC アルゴリズム [14, 15] は $O(|E|^{1.5})$ 時間を要するため，(c) は全体で $O(|P|^{1.5})$ 時間で計算できる．$|P| < |E|$ より提案手法は従来の TC アルゴリズムより効率的である．

3.3 アルゴリズム

TC における提案手法のアルゴリズムを Algorithm 1 に示す．Algorithm 1 は前節で示した (a) から (c) の場合分けに基づき，三角形を数え上げる．(a) について，各ウェッジがトライアングル型であれば三角形の個数 t に 1 を加算する（3–5 行目）．(b) について，まずスーパーノード対とそれらの間のスーパーエッジの各状態に対する三角形の個数を先に計算する（8 行目）．その後，各スーパーエッジについて事前計算した値を参照して三角形の個数を t に加算する（9–11 行目）．(c) について，まず G' 中のスーパートライアングルを列挙する（14 行目）．次に，各スーパートライアングルについて，BLEW27 を用いて三角形の個数を求め，t に加算する（15–19 行目）．最後に，t を出力する（20 行目）．

理論解析: 提案手法の効率性と正確性について理論的に解析する．

定理 1 Algorithm 1 の時間計算量は $O(|P|^{1.5})$ である．

証明 Algorithm 1 は (a) から (c) の三つの部分からなる．(a) について，各ウェッジを一回ずつ参照するため $O(|S|)$ 時間を要する．(b) について，各スーパーエッジを一回ずつ参照するため $O(|P|)$ 時間を要する．なお，事前計算については状態数やノード・エッジ数の最大値が定数であるため定数時間で求まる．(c) について，3.2 節で議論した通り，スーパートライアングルの列挙は $O(|P|^{1.5})$ 時間を要する．なお，各 BLEW27 の計算および論理演算は定数時間で求まる．一般に $|S| << |P|$ が成り立つため，Algorithm 1 の時間計算量は $O(|P|^{1.5})$ である． □

最先端の TC アルゴリズム [14, 15] の時間計算量は $O(|E|^{1.5})$ である．その一方で定理 1 に示した通り提案手法は $O(|P|^{1.5})$ 時間で解を出力する．ウェッジ分割においては一般に $|P| < |E|$ が成り立つ [13] ため，提案手法は効率的にウェッジ分割グラフを TC アルゴリズムに活用することが可能である．

定理 2 Algorithm 1 は正確な TC アルゴリズムである．

証明 Algorithm 1 はウェッジ分割グラフ上の三角形の数を三つの場合分けを用いて計算する．(a) 一つのウェッジに収まる三角形の個数はトライアングル型ウェッジに一致し，(b) 二つのウェッジに跨って存在する三角形は必ずウェッジ内のエッジとそれらを繋ぐ一つのスーパーエッジの中に収まるため，事前計算により正確に全列挙ができる．(c) 三つのウェッジに跨って存在する三角形は BLEW27 により各ノードの三つ組が各桁に一対一対応するように整列されるため，全てのノードの三つ組についての計算が同時に行われ，各桁の値により三角形が存在するかが判定できる．したがって，これらの計算に重複や漏れは無く，Algorithm 1 は常に正確に三角形の個数を出力する． □

定理 2 は提案手法が効率的な TC アルゴリズムを実現するためにウェッジ分割グラフの展開を行わないが，正確性を損なわないことを示している．

4 提案手法: Single Source Shortest Path (SSSP)

本節ではウェッジ分割に基づく要約グラフに対する高速な SSSP アルゴリズムを提案する．4.1 節に提案手法を構成する基本的なアイデアを示す．紙面の都合上，補題の証明は省略する．

4.1 基本アイデア

本節で提案する SSSP アルゴリズムはスーパーエッジの内部を展開することなく始点から各ウェッジへの到達距離を計算することで，効率的に最短距離を求める．まず，三つ組の距離を表すタプル (x, y, z) を各ウェッジに割り当てることで，ウェッジ内の各ノードへの最短距離を表現することを考える．TC アルゴリズムのアイデアでも述べたように，これにより，ウェッジ間の全ノード対の距離をウェッジを展開せずに高速に計算できる．しかしながら，ウェッジ単位での距離のタプルの計算は元のグラフにおける距離計算と異なり，一ホップ離れたウェッジまでの距離が均一でなく，単純な幅優先探索は適用できない．そこで本稿では一般的な重み付きグラフに対する SSSP アルゴリズムである，Dijkstra 法 [19] をウェッジ分割グラフに応用することを考える．

Dijkstra 法のウェッジ分割グラフへの拡張には，ウェッジの訪問時に最短距離の保証がされない場合があるという課題がある．具体的には，一般的なグラフにおける Dijkstra 法では各ノードを初めて訪問した際に，それまでの距離がそのノードへの最短距離であることが保証される．その一方で，ウェッジ分割グラフにおいては各ウェッジを初めて訪問した際に，ウェッジ中の一部のノードについては最短距離が求まらないことがある．そのため，訪問済みのウェッジを再訪問することを考慮する必要があり，複雑な探索が求められる．この問題を解決するために，

Algorithm 2 ウェッジ分割に基づく SSSP

Input: グラフ G のウェッジ分割 $G'(S, P, \sigma)$, 始点ノード v_0;
Output: v_s から各ノードまでの最短距離;

1: $dist(n)$ を $n = 0, 1, ..., 2047$ について計算する;
2: $sp[s] \leftarrow (\infty, \infty, \infty)$ **for all** $s \in S$;
3: $s_0 \leftarrow v_0$ を含むウェッジ;
4: $d_0 \leftarrow v_0$ から s_0 の各ノードまでの最短距離のタプル;
5: 優先度付きキュー $Q \leftarrow \emptyset$;
6: $Q.push((s_0, d_0, 0))$;
7: **while** $Q \neq \emptyset$ **do**
8: $(s, d, min_dist) \leftarrow Q.pop()$;
9: **for each** $t \in N(s)$ **do**
10: **if** $\max(sp[t]) \leqq min_dist$ **then continue**;
11: $d_t \leftarrow d_t + dist[blew11(s, t)]$;
12: **if** $sp[t] = \min(sp[t], d_t)$ **then continue**;
13: $sp[t] \leftarrow \min(sp[t], d_t)$;
14: $Q.push((t, d_t, \min(d_t))$;
15: **return** sp;

本研究では探索の過程でそれ以上距離タプルが更新されないことを保証できる条件を導出し，再訪問を回数を必要最小限に抑える．

4.2 BLEW に基づく高速な距離計算

本節では 4.1 節で述べたアイデアに基づき，SSSP における幅優先探索をウェッジ単位で処理する Dijkstra 法に帰着させる方法について議論する．

ウェッジ s に属する各ノードがスーパーエッジ (s, t) を用いて隣接するウェッジ t に属する各ノードへの移動する最短距離は，3.1 節の (b) に述べた手法と同様に，2048 種類の各状態に対して事前計算が可能である．すなわち，全ての整数 $n(0 \leqq n < 2^{11})$ に対して $blew^{-1}(n) = (s, t, (s, t))$ から構成されるグラフにおいて，s の三ノードから t の三ノードまでの 9 個の最短距離を求めればよい．ここで 3.1 節の (b) とは異なり，この事前計算の結果は各ノード対に対して 0/1 のバイナリではなく整数である最短距離を記憶する必要があるため，各状態に対して 9 変数からなる次のタプルを定義する．

定義 7 (DIST11) 整数 n が与えられたとき，$dist(n)$ は次の 9 つの値のタプルである．

$$dist(n) = (l_{00}(n), l_{01}(n), l_{02}(n), l_{10}(n), l_{11}(n), l_{12}(n), l_{20}(n), l_{21}(n), l_{22}(n)).$$

ここで，$l_{ij}(n)$ は $blew^{-1}(n) = (s, t, (s, t))$ から構成されるグラフにおける s, t のローカルノード id がそれぞれ i, j である二ノード間の最短距離を示す．

定義 7 に基づき，全ての整数 $n(0 \leqq n < 2^{11})$ に対して $dist(n)$ を計算することで，ウェッジを移動する際の各ノードまでの距離のタプルを定数時間で計算することができる．具体的な計算方法を次の補題に示す．

補題 1 始点からあるウェッジ s の各ノードまでの最短距離が (s_x, s_y, s_z) であるとき，始点から s を経由して隣接するウェッジ t までの最短距離 (t_x, t_y, t_z) は次の式で表される．

$$t_x = \min\{(s_x + l_{00}(n)), (s_y + l_{10}(n)), (s_z + l_{20}(n))\}. \quad (1)$$

$$t_y = \min\{(s_x + l_{01}(n)), (s_y + l_{11}(n)), (s_z + l_{21}(n))\}. \quad (2)$$

$$t_z = \min\{(s_x + l_{02}(n)), (s_y + l_{12}(n)), (s_z + l_{22}(n))\}. \quad (3)$$

ただし，$n = blew11(s, t, (s, t))$ とする．

簡便のために，以降の議論では式 (1)–(3) の演算をまとめて $(X, Y, Z) = (x, y, z) + dist(n)$ と表記する．補題 1 により，ウェッジ分割グラフ上の移動コストを元のエッジを展開せずに求めることが可能となる．

しかしながら，ここで得られた (X, Y, Z) は始点からスーパーエッジ (s, t) を通った場合の最短経路であるため，必ずしも真の最短経路と一致するとは限らない．そのため，複数の経路から得られた距離のタプルに対する各ノードの最短距離を次の補題に示す．

補題 2 始点からウェッジ s に到達した二種類の経路上の距離 $(s_{x1}, s_{y1}, s_{z1}), (s_{x2}, s_{y2}, s_{z2})$ が与えられたとき，始点からそれらのいずれかを経由した s までの最短距離 (s_x, s_y, s_z) は次の式で表される．

$$s_x = \min\{s_{x1}, s_{x2}\}. \quad (4)$$

$$s_y = \min\{s_{y1}, s_{y2}\}. \quad (5)$$

$$s_z = \min\{s_{z1}, s_{z2}\}. \quad (6)$$

簡便のために，以降の議論では式 (4)–(6) の演算をまとめて $(s_x, s_y, s_z) = \min((s_{x1}, s_{y1}, s_{z1}), (s_{x2}, s_{y2}, s_{z2}))$ と表記する．補題 2 を s からの全ての経路に適用することで SSSP を解くことができるが，その組み合わせは膨大であるため，次節にて Dijkstra 法を用いた効率的な経路探索アルゴリズムを設計する．

4.3 アルゴリズム

SSSP における提案手法のアルゴリズムを Algorithm 2 に示す．Algorithm 2 は前節で示した Dijkstra 法の拡張に基づき，始点ノード v_0 から各ノードまでの最短距離を計算する．まず Algorithm 2 はスーパーエッジの各状態に対する最短距離のタプルを事前計算し，各ノードまでの最短距離を十分に大きな値で初期化する（1,2 行目）．また始点ノードを含むウェッジとウェッジ内の各ノードまでの距離を求め，優先度付きキューに挿入する（3–6 行目）．探索では，優先度付きキューから一つ要素を取り出し，その状態から隣接するウェッジに移動した場合の距離を計算し，最短距離が更新された場合は優先度付きキューに遷移先の状態を挿入する（8–14 行目）．ここで重要となるのは，10 行目に示す条件である．これは，優先度付きキューに含まれる min_dist がウェッジ t の現在の距離のタプルより大きい場合には，これ以上 $sp[t]$ は更新されないことを利用している．この条件を満たす場合には，一般の Dijkstra 法における訪問済みのノードと同様に扱うことができるため，探索を早期に打ち切り，不要な再訪問を回避することで効率化を実現している．この処理の正確性については，定理 4 にて議論する．

理論解析: 提案手法の効率性と正確性について理論的に解析する．

定理 3 Algorithm 2 の時間計算量は $O(|E|\ log|S|)$ である．

証明 Algorithm 2 の処理は Dijkstra 法を拡張しており，一般的な優先度付きキュー Q を用いた Dijkstra 法の時間計算量は $O(|E|\ log|V|)$ である．これは Q に状態を $O(|E|)$ 回挿入し，ヒープが挿入に要するコストが $O(log|V|)$ 時間であることに起因する．すなわち，ウェッジ分割グラフ上では時間計算量 $O(|P|\ log|S|)$ と仮定できる．しかしながら，Algorithm 2 では Q に複数回同じノードを挿入する可能性があるため，挿入回数は最悪の場合 $|E|$ 回発生する．一方で，Q の要素数が高々$|S|$ であることは変わらないため，Q への挿入は各 $O(log|S|)$ 時間で実行可能である．したがって，Algorithm 2 は $O(|E|\ log|S|)$ 時間で動作する． □

定理 3 は，Algorithm 2 の理論的な時間計算量が Dijkstra 法と比較して効率的であることを示している．ここで，本研究において基盤とするウェッジ分割アルゴリズム HYBRID [13] ではトライアングル型のウェッジが多く，さらにはスーパーエッジが可及的密になるよう設計されていることを考慮すると，多くの場合ウェッジの距離のタプル内には最小値と等しい値が含まれることが多いと考えられる．なぜならば，二つのウェッジにおけるノード間の距離が長くなるケースは，パス型のトライアングルを疎なスーパーエッジが接続するときのみに限られるためである．このことから，各ウェッジの距離のタプル内の各値は多くの場合に最小値と等しく，そのために距離の更新，すなわち再訪問が発生する機会も極めて少ないことがわかる．したがって，Algorithm 2 の実用上の時間計算量は $O(|P|\ log|S|)$ に近似できる．

定理 4 Algorithm 2 は正確な SSSP アルゴリズムである．

証明 Algorithm 2 が一般の Dijkstra 法と大きく異なる処理をする点は次の二つである：(1) 距離の計算が DIST11 に基づくタプルを用いた演算になっている点，(2) ノードに再訪問する可能性がある点．(1) について，補題 1，および補題 2 より DIST11 による演算は全ノード間の最短距離を全探索した結果と同値であるため，正確性を損なうことはない．(2) について，DIST11 を用いた距離は常に正であることから，優先度付きキュー内の min_dist は単調増加である．したがって，距離のタプル内の最大値が現在の min_dist よりも小さい場合には，それが最短距離であり，更新されることはない．また，再訪問する必要のあるノードは未訪問のノードと同様に扱われるため，正確性を損なうことはない． □

定理 4 は提案手法が効率的な SSSP アルゴリズムを実現するためにウェッジを展開せずに経路探索を行うが，精度上の問題が無いことを示している．

5 評価実験

本節では提案手法の実行速度を実験的に評価する．

表 2: 実験に用いるデータセットの概要

	$\|V\|$	$\|E\|$	$\|d_{avg}\|$	t	Type
DBLP	317K	1.05M	8.7	2.22M	co-authorship
Amazon	403K	2.44M	6.1	667K	co-parchased
YouTube	1.13M	2.99M	2.6	3.06M	social
LiveJournal	4.00M	34.7M	8.7	177M	social

表 3: TC アルゴリズムの実行時間の比較

	一般的な実装	提案手法
DBLP	0.35 sec.	0.26 sec.
Amazon	1.16 sec.	0.96 sec.
YouTube	2.33 sec.	2.18 sec.
LiveJournal	72.0 sec.	15.9 sec.

5.1 実験設定

データセット: 実験に用いたデータセットを表 2 に示す．本稿では既存手法 [11, 12] が用いたデータセットのうち，スケールフリー性を持つデータセットを掲載する．表の d_{avg}, t はそれぞれグラフの平均次数と三角形の数を示す．

比較手法：

- **TC:** TC の一般的な実装 [14, 15] である，$O(|E|^{1.5})$ 時間のアルゴリズムと比較する．
- **SSSP:** SSSP の一般的な実装である，$O(|V| + |E|)$ 時間の幅優先探索アルゴリズムと比較する．

実験環境: 全ての実験は Xeon(R) Gold 6246R (3.4GHz) 350GB RAM を搭載した Linux サーバにて行い，C++ (gcc 8.2.0) による実装と，コンパイルオプション -O3 を用いて実行した．

5.2 TC の実行時間

TC の実行時間の比較を表 3 に示す．表 3 より，提案手法は従来の実装と比較して大幅に高速化していることがわかる．最も高速化率の高いデータは LiveJournal であり，その値は約 4.5 倍である．これはエッジ数が多く密度が高いグラフに対してグラフ要約が効率的にエッジ数を減らしたことにより，$|E|^{1.5}$ と $|P|^{1.5}$ の差が効果的に小さくなったためであると考えられる．一方で，YouTube では大きな改善が見られなかった．これは，密度が低いグラフに対してはパス型のトライアングルが多くなってしまい，スーパーエッジの数が効率的に削減されなかったことに起因する．

5.3 SSSP の実行時間

SSSP の実行時間の比較を表 4 に示す．表 4 より，提案手法は最短距離においても従来の実装と比較して高速に処理できることがわかる．最も高速化率の高いデータは LiveJournal であり，その値は約 1.6 倍である．SSSP は線形アルゴリズムであり，TC と比較してグラフの密度に大きく依存する処理ではないにも関わらず，TC と同様に SSSP においても提案手法は密度が高いグラフに対して高速化率が高くなっている．これはグラフ要約の効率性がグラフの密度に依存するために起こると考えられる現象であり，効率的なグラフ要約がプリミティブなグラフ処理の高速化率にも影響することを示唆している．

表 4: SSSP アルゴリズムの実行時間の比較

	一般的な実装	提案手法
DBLP	3.01 msec.	2.37 msec.
Amazon	7.48 msec.	5.87 msec.
YouTube	7.89 msec.	7.21 msec.
LiveJournal	84.36 msec.	52.81 msec.

6 おわりに

本稿ではウェッジ分割に基づき要約されたグラフに対して，要約グラフ上でスーパーノードを展開することなく動作する TC アルゴリズムと SSSP アルゴリズムの実装を提案した．各アルゴリズムは BLEW により割り当てられたスーパーエッジの重みに対して bit 演算を用いることでウェッジを跨いだ処理を高速に行い，グラフ要約によって削減されたエッジ数を活かした処理を実現した．実世界のデータセットに対する実験により，提案手法は TC, SSSP ともに従来の一般的な実装よりも高速であることを実証した．

謝　辞

本研究の一部は JST さきがけ (JPMJPR2033)，JSPS 科研費 (JP22K17894), JSPS 科研費 (JP22J10972) の支援を受けたものである．

文　献

[1] J. Leskovec and A. Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`, June 2014.

[2] Ryan Rossi and Nesreen Ahmed. The Network Data Repository with Interactive Graph Analytics and Visualization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[3] H. Shiokawa and M. Onizuka. *Scalable Graph Clustering and Its Applications*, pages 1–10. Springer New York, 2017.

[4] Etsuji Tomita and Tomokazu Seki. An Efficient Branch-and-bound Algorithm for Finding a Maximum Clique. In *International Conference on Discrete Mathematics and Theoretical Computer Science*, pages 278–289. Springer, 2003.

[5] Xin Huang and Laks V. S. Lakshmanan. Attribute-Driven Community Search. *Proc. VLDB Endow.*, 10(9):949–960, may 2017.

[6] Y. Zhou, H. Cheng, and J. X. Yu. Graph Clustering Based on Structural/Attribute Similarities. *PVLDB*, 2(1):718–729, August 2009.

[7] Jinkun Lin, Shaowei Cai, Chuan Luo, and Kaile Su. A reduction based method for coloring very large graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 517–523. AAAI Press, 2017.

[8] Yi Zhou, Shan Hu, Mingyu Xiao, and Zhang-Hua Fu. Improving Maximum k-Plex Solver via Second-Order Reduction and Graph Color Bounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12453–12460, 2021.

[9] Jian Gao, Jiejiang Chen, Minghao Yin, Rong Chen, and Yiyuan Wang. An Exact Algorithm for Maximum k-Plexes in Massive Graphs. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1449–1455. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[10] Matthew C Schmidt, Nagiza F Samatova, Kevin Thomas, and Byung-Hoon Park. A Scalable, Parallel Algorithm for Maximal Clique Enumeration. *Journal of Parallel and Distributed Computing*, 69(4):417–428, 2009.

[11] K Ashwin Kumar and Petros Efstathopoulos. Utility-driven Graph summarization. *Proceedings of the VLDB Endowment*, 12(4):335–347, 2018.

[12] Kijung Shin, Amol Ghoting, Myunghwan Kim, and Hema Raghavan. SWeG: Lossless and Lossy Summarization of Web-Scale Graphs. In *The World Wide Web Conference*, WWW '19, page 1679–1690. Association for Computing Machinery, 2019.

[13] Shohei Matsugu and Hiroaki Shiokawa. Wedge-based Graph Summarization for Large-scale Graph. *IEICE TRANSACTIONS on Information and Systems*, J106-D(10), October 2023. (in Japanese with English Abstract).

[14] S Thomas. Algorithmic Aspects of Triangle-Based Network Analysis. *Phd in Computer Science*, 26129, 2007.

[15] Matthieu Latapy. Main-memory Triangle Computations for Very Large (Sparse (Power-law)) Graphs. *Theoretical Computer Science*, 407(1):458–473, 2008.

[16] Jonathan Cohen. Trusses: Cohesive subgraphs for social network analysis. *National security agency technical report*, 16(3.1), 2008.

[17] Shohei Matsugu, Hiroaki Shiokawa, and Hiroyuki Kitagawa. Fast Algorithm for Attributed Community Search. *Journal of Information Processing*, 29:188–196, 2021.

[18] Shohei Matsugu, Yasuhiro Fujiwara, and Hiroaki Shiokawa. Uncovering the Largest Community in Social Networks at Scale. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI2023)*, 2023. (to appear).

[19] Edsger W Dijkstra. A Note on Two Problems in Connexion with Graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, pages 287–290. 2022.

[20] Junya Yanagisawa and Hiroaki Shiokawa. Fast One-to-Many Reliability Estimation for Uncertain Graphs. In *Database and Expert Systems Applications: 31st International Conference, DEXA 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings, Part I 31*, pages 106–121. Springer, 2020.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 10:00 ~ 12:10 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (1)

座長:杉山 一成(大阪成蹊大学)
コメンテータ:灘本 明代(甲南大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T3-A-4-01] LLMによる商品情報やレビューをもとにした商品推薦

*朴 健[1]、丹羽 健斗[1]、髙村 昇二郎[1]、青井 順一[1] (1. 株式会社MonotaRO)

10:20 ~ 10:45
[T3-A-4-02] レビュー動画内における商品特徴の言及量分析システムの開発

*山口 史弥[1]、上田 真由美[2,3]、中島 伸介[4] (1. 京都産業大学大学院、2. 流通科学大学、3. 大阪大学、4. 京都産業大学)

10:45 ~ 11:10
[T3-A-4-03] 不快コメント予測の個人化のためのコメント特徴に基づく個人化レベルの調整

*中原 輝樹[1]、牛尼 剛聡[1] (1. 九州大学)

11:10 ~ 11:35
[T3-A-4-04] 食品レビュー検索でのクエリ拡張のためのLLMを用いた同義フレーズ生成

*芦澤 亜里紗[1]、三林 亮太[1]、大島 裕明[1] (1. 兵庫県立大学)

12:00 ~ 12:10
イントロダクション

LLMによる商品情報やレビューをもとにした商品推薦

朴　　健†　　丹羽　健斗†　　高村昇二郎†　　青井　順一†

† 株式会社 MonotaRO　〒530-0001 大阪市北区梅田三丁目2番2号 JP タワー大阪22階
E-mail: †{jian.piao,kento.niwa,shojiro.takamura,aoi}@monotaro.com

あらまし　B2B のユーザーは業務の課題を解決するために，用途に応じた商品を求める需要がある．それらの需要に応えるために，整備された用途情報と用途ごとの商品グルーピングデータが必要となる．しかし，サプライヤーから提供される用途情報は数や整備が不十分である場合がある．また，用途ごとの商品グルーピングは専門知識を要するため量産ができないなどの課題がある．本論文では，LLM を用いて商品の用途タグを生成し，それに基づいて商品を用途ごとにグルーピングする手法を提案する．提案手法により商品情報や商品レビュー文をもとに用途情報の整備と用途ごとの商品グルーピングデータの量産が可能となることを示す．最後に，LLM と既存手法で作成したデータを比較してサービスの性能向上が可能かどうかを検証する．

キーワード　LLM，商品情報生成，セッションベース推薦，知識ベース推薦

1　はじめに

ユーザーには課題を解決するため用途に応じて商品を検索する需要がある．例えば，工事現場の熱中症対策や，狭いところでのねじ締めなどのような用途で商品を探す需要があり，それらの需要に応じた商品推薦を行うことでユーザーの課題解決に役に立つ商品の認知を広げることができる．用途で商品が探せるようにするためには，用途情報の整備と用途ごとの商品グルーピングデータが必要となる．

用途情報の整備手法については，商品の用途情報からの抽出やレビュー文からの推測などが考えられる．しかし，サプライヤーから提供される用途情報やレビュー文は数が少なく，情報が構造化されておらずサービス開発のためのデータ整備が不十分であることが多い．商品の用途情報を整備するため，梶原ら [1] は用途文から BERT モデルを用いて用途タグを抽出する手法を提案した．この手法によって有用な用途情報を抽出できるものの，多くの質の高いアノテーションデータが必要とされることと，用途文が提供されている商品の数が少ないという課題があった．

用途ごとの商品グルーピングデータを作成する上での課題として，商品数が数千万を超える大規模サービスにおいて人手によって専門知識をもとに商品を選定すると工数がかかるため量産をすることができない点があげられる．この課題を解決するため優れた推論能力が証明された LLM(Large Language Models，大規模言語モデル) が活用されている．Chen ら [2] は TagGPT という LLM モデルをベースとした Zero-shot 学習手法で動画内容がいくつかのタグで集約できるフレームワークを提案した．TagGPT の目的は商品 (動画) にタグを付けることで，類似している動画を推薦することだが，用途による商品グルーピングでは単一の商品が複数の用途で利用されることがあるため TagGPT をそのまま利用することは出来ない．

本論文では商品情報や商品レビュー文から用途タグを抽出した上で，用途ごとに商品のグルーピングを行う手法を提案する．本手法では，LLM を利用することでアノテーションデータを用いずに用途タグの抽出を行った上で，用途タグをもとに LLM により商品をグルーピングする．本論文の貢献は以下の通りである．

- LLM の優れた推論や情報生成能力を利用して，商品情報やレビュー文を入力として用途タグを半自動で作成できることを示した．
- 大規模データに対して適用可能な LLM による用途と商品のマッチングの手法を考案した．用途商品を推薦に利用する有効性をオフライン実験で確認した．

2　提案手法

商品推薦で利用するための用途ごとの商品グルーピングデータを作成するために以下の 2 つの手順をとる．

- 手順 1: 商品情報とレビュー文から LLM によって商品の用途タグを作成する
- 手順 2: 作成した用途に対して推薦候補の商品リストがマッチするかどうかを LLM によって判定する

手順 2 において用途とマッチングした商品をグルーピングして推薦商品とする．これらの手順によるデータ作成プロセスを図 1 で示す．

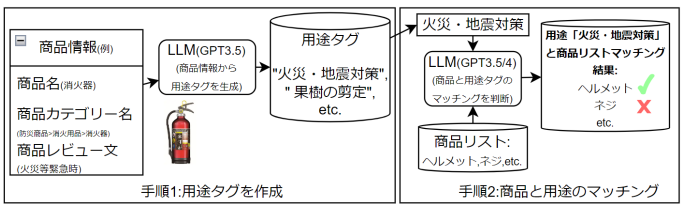


図 1　データ作成プロセス．商品情報やレビュー文を利用して用途タグを作成する．その後，用途タグと商品のマッチングを行う．

2.1 用途タグの生成

用途タグとは，商品の用途を単語や短いフレーズで表現したものを指す．用途タグは商品名，商品カテゴリー名，商品用途文などの商品情報やレビュー文から推測できると考えられる．そのため，推論対象の商品リストを作成した上で，商品ごとに商品情報とレビュー文をもとに Prompt を作成して，GPT3.5 [4] による推論を実行して用途タグを作成する．なお，学習は Few-Shot Learning で行っており，主な目的は出力の用途タグのフォーマットを整えることである．用途タグを作成するための Prompt を以下に示す.

- user: 商品カテゴリー名，商品名と商品のレビュー文を受け取り，重要な用途をサマリーしてください．用途は短い単語やフレーズで出力してください．商品カテゴリー名: シフトノブ/ペダル. 商品名: ミニペダルセット. 商品のレビュー文: ペダルのゴム部の磨耗，軽バン MT 車，アルト・ラパンのプチカスタム，ペダルの滑り止め，車のメンテナンス時.
- assistant: 車のペダルの磨耗, 車のペダルの滑り止めの交換
- user: 商品カテゴリー名，商品名と商品のレビュー文を受け取り，重要な用途をサマリーしてください．用途は短い単語やフレーズで出力してください．商品カテゴリー名:{ 商品カテゴリー名 }. 商品名:{ 商品名 }. 商品のレビュー文: { 商品のレビュー文 }.

生成された用途タグには類似する意味合いだが異なる表現を持つノイズが多く混在するため，類似タグを一つにまとめる後処理を行う．具体的には，用途タグを事前学習をしたユニバーサルセンテンスエンコーダーモデル [3] を用いてベクトル化し，k-means 手法を利用してクラスタリングを行う．クラスタ数をタグ数の 0.5 倍にし，同じクラスタにあるタグはそのクラスタの中心に最も近いタグにまとめる．なお，これらの出力データには推薦で利用するには不適切なデータが含まれるため，最終的に利用する用途タグは専門知識をもつ担当者が目視で精査を行う必要がある．本作業の必要性については 3.1 において説明を行う.

2.2 商品と用途のマッチング

用途ごとの商品グルーピングを行うために必要な用途と商品のマッチング手法について述べる．2.1 で説明をした用途タグに対してマッチをする商品をカタログの全商品の中から検索を行いたいが，大規模な商品カタログを対象とする際の計算量について留意する必要がある．なぜなら，1 つの商品が多くの用途で利用されることがあり，用途の商品を網羅するには全ての商品を対象に用途をマッチングする必要があるためである．具体的には，“マスク”から粉塵対策，ウィルス対策，清掃などの一般的な用途タグの生成は容易だが，“林業の草刈り”のような分野に絞った用途タグを生成することは難しい．よって，計算量を減らすためにツリー構造になっている商品の種類を分類するカテゴリー情報を利用して階層的に用途とのマッチングを行う．具体的には，一つの用途タグに対して，親階層のカテゴリーがマッチングできたら，子階層のカテゴリーのマッチングを行い，最後に商品情報を利用して商品と用途タグのマッチングを行う．マッチングの手法については，先行研究の TagGPT ではタグと動画のテキスト類似度を利用したが，商品用途の場合は用途タグと商品のテキスト情報を知識をもとにマッチングする必要がある．例えば，“エンジンオイル交換”という用途に対して，“軍手”が 1 つの役に立つ商品として考えられるが，軍手の商品情報に“エンジンオイル交換”という情報がないため，テキスト類似度でのマッチングが困難である．したがって，商品情報が不足していてもマッチングを行うために LLM の知識と推論能力を利用する．Prompt に利用する情報としては，カテゴリーのマッチングの場合はカテゴリーツリー名を利用し，商品のマッチングの場合は商品名，商品カテゴリー名，商品説明文，商品用途などの商品情報とレビュー文を利用する．用途タグと商品カテゴリーのマッチングの Prompt と用途タグと商品のマッチングの Prompt の例をそれぞれ以下で示す.

- 以下の商品カテゴリー情報を参考して，このカテゴリー商品は { 用途タグ } の用途を持つ商品かを判断してください．その用途に使える商品カテゴリーの場合は 1 を，使えない商品カテゴリーの場合は 0 を返してください．その理由を教えてください．商品カテゴリー名: { 商品カテゴリー名 }.
- 以下の商品情報を参考して，この商品は { 用途タグ } の用途を持つ商品かを判断してください．その用途に使える商品の場合は 1 を，使えない商品の場合は 0 を返してください．商品情報: { 商品名:{ 商品名 }, 商品カテゴリー名:{ 商品カテゴリー名 }, 商品説明文:{ 商品説明文 }, 商品用途:{ 商品用途 }, 商品レビュー: { 商品レビュー }.

なお，マッチングの質を上げるために，GPT3.5 [4] と GPT4 [5] からの結果を利用している．また，用途タグと同様に推薦に不適切な商品を含むことがあるため，人手による精査などを行う必要がある．なお，推薦商品は用途とマッチした商品に対して，注文数と親カテゴリーの多様性を考慮して商品をランキングして対象の商品を抽出することで作成する.

3 実　　験

商品推薦のグルーピングデータを作成するために必要な用途タグの品質と用途-商品のマッチングの性能の評価を行う．用途タグの品質については作成したタグをサンプリングして定性評価を行う．マッチングの性能については手動で作成した推薦データとの購入の予測精度を定量的に比較することで評価を行う．また，用途に対する網羅性を確認するために商品数や多様性についても評価を行う．最後に，用途タグや推薦商品のデータを作成する際に人手による精査を要するため，実験で利用したデータを作成する際に要した時間について述べる．なお，本章で利用するカテゴリーは親カテゴリーを表す.

3.1 用途タグの定性評価

monotaro.com の売上上位 10000 商品を利用して提案手法によって作成した用途タグの数の上位 100 件に対して定性評価を行うことで品質を評価する．これらの用途に含める商品の範囲

が妥当か，フォーマットが正しいかという観点で専門知識を持つ担当者が用途タグを目視で確認して評価を行った．評価対象の 100 個の用途タグの内，44 個の用途は商品範囲が適切で有効だと判定した．これらは用途の意味が分かりやすく，適切なフォーマットで表現されている．一方，56 個の用途は範囲が広すぎるもの（e.g., メンテナンス時の調整）や狭すぎるもの（e.g., 農作業の手の保護）などの問題がある．以上より提案手法によって作成した用途タグをそのまま利用することは問題があり，人手による精査が必要であることが分かる．表 1 に一部の用途タグの判定結果を示す．

表 1　用途タグの判定結果のサンプル

用途タグ	判定結果
自動車のベルト交換	有効
水道管のメンテナンス	有効
廃材処理	有効
配管の固定	有効
農作業の手の保護	商品の範囲が狭すぎる
メンテナンス時の調整	商品の範囲が広すぎる
段ボール箱の梱包	名称がわかりにくい

3.2　用途タグとマッチングした商品の推薦への応用実験

用途ごとの推薦商品の性能を測るために商品グルーピングデータを商品の購入を予測する推薦タスクに適用する．

作成済みの 45 個の用途タグに対して，既存の推薦商品と提案手法によって作成した推薦商品の購入予測精度を比較することで評価を行う．評価対象ユーザーは monotaro.com において注文実績がある 1 万ユーザーをサンプリングして決める．

次に，ユーザーごとの用途タグの選定方法と商品のランキング方法について述べる．サンプリングしたユーザーの 1 年間の購入商品の履歴データを用いて，ユーザーごとに推薦をする用途タグを 1 個選ぶ．タグの選び方に関しては，ユーザーの行動履歴データと商品グルーピング間のスコアを算出し，スコアが最も高い用途データを推薦するようにしたルールベース方法を利用する．スコアの算出方法を式 1 で示す

$$S(u,t) = \alpha \times \sum_{i=1}^{n}(u_{p_i}) + \beta \times \sum_{j=1}^{m}(u_{c_j}) \qquad (1)$$

ユーザー u の用途タグ t に対するスコア S は，ユーザーが t の商品グルーピングデータに含まれている商品の購入回数 $\sum_{i=1}^{n}(u_{p_i})$ と商品カテゴリーの購入回数 $\sum_{j=1}^{m}(u_{c_j})$ にそれぞれの重み α と β をかけた値になる．重みの α と β はそれぞれ 1 から 100 までの値で探索し, それぞれ 1 に設定した場合に良い結果が得られた．n と m はそれぞれ商品グルーピングデータに含まれている商品数と商品カテゴリー数である．ユーザーごとに選定した用途タグに紐づく商品グルーピングデータを推薦候補商品とする．その上で，注文数と候補商品のカテゴリー数が多い順で階層ごとにカテゴリーの重要度を決めた上で，多様性を考慮して同じカテゴリーが多くても 2 個以上連続に並べないようにランキングをして上位 100 商品を抽出する．

注文実績があるユーザーについて，評価期間における 2 つの商品グルーピングデータの予測精度を評価する．履歴データの期間の後の 1 か月間に発生した対象ユーザーの購入商品を評価データとして利用し，推薦された商品グルーピングデータの中に購入商品が含まれていたユーザー数で評価する．提案手法の商品グルーピングデータの商品数は手動で作成したデータより多いため，商品数を揃えた場合の推薦性能も測る．具体的には，提案手法のデータの商品に対して購入回数が多い順にランクをつけ，手動で作成したデータの商品数よりも大きいランクの商品を除外することで商品数を揃える．推薦結果を表 2 に示す．

表 2　推薦性能の評価

手法	Recall	Precision	F1
手動	0.0119	0.0131	0.0124
LLM(提案手法)	0.0524	0.0564	0.0543
LLM(商品数限定)	0.0280	0.0308	0.0294

手動で作成した用途データと比較して，提案手法で作成した用途データの Recall，Precision 及び F1 値が高い．商品数を揃えたデータについては，F1 値は 0.0543 から 0.0294 まで下がるが，手動で作成したデータより性能が高い．したがって，提案手法によって作成した商品グルーピングデータの予測精度が高いことから推薦商品としての有効性を確認することができる．

また，用途に必要な商品の網羅性について評価をするために推薦商品数と推薦カテゴリー数を評価する．提案手法と手動で作成した商品グルーピングデータの商品数とカテゴリー数の平均値を表 3 で示す．カテゴリー数を用いて認知する商品の多様性を評価する．提案手法で作成したデータは商品数とカテゴリー数が多いため，より多様な商品を網羅した用途商品の推薦を行うことができる．具体的な商品グルーピングデータの例をピックアップして図 2 に示す．

表 3　商品数と多様性の指標

手法	商品数	カテゴリー数
手動	13.7	3.1
LLM(提案手法)	75.2	9.8

3.3　作業工数の効率化についての評価

提案手法による半自動化手法と従来の手動による用途データの作成手法に必要とする作業工数をそれぞれ測って評価する．具体的には，従来手法によって用途データの作成に要した工数と提案手法によって作成した用途タグと商品グルーピングデータを専門知識を持つ担当者が精査するために要した時間を比較する．その結果，LLM を利用することで 1 個の用途データを用意する作業工数を約 4 時間から 10 分までに短縮できることを確認した．

4　ま　と　め

本論文では，用途に応じた商品推薦を行うことでユーザーの課題解決に役に立つ商品の認知を広げるため，LLM を用いて

図 2　手動で作成した用途データと提案手法で作成した用途データのサンプル

商品情報や商品レビュー文から用途タグを抽出した上で商品と用途タグのマッチングを行うことで，用途ごとに商品のグルーピングを行う手法を提案した．その結果，データ作成の半自動化を実現して作業工数が短縮できることを示した．また，定性評価により，100 個の用途サンプルの内 44 個が有効であることが確認され，定量評価のためのオフライン実験でも予測精度と多様性の改善が確認された．今後の課題としては,「手研磨」用途の「手」のようなキーワードに焦点を当てられる Prompt の改善によって「研磨機械」のようなノイズを除外することや提案手法で量産したデータを用いたオンラインテストの実施が挙げられる．

文　　献

[1] 梶原奨, 井上翔太, 岡林遥平, 稲田和明, 張信鵬. E コマースにおける商品用途表現の抽出とグルーピング. 言語処理学会第 29 回年次大会 (NLP2023), 2023.
[2] Chen Li, Yixiao Ge, Jiayong Mao, Dian Li, and Ying Shan. 2023. TagGPT: Large Language Models are Zero-shot Multimodal Taggers. arXiv preprint arXiv:2304.03022 (2023).
[3] Daniel Cer and Yinfei Yang and Sheng-yi Kong and Nan Hua and Nicole Limtiaco and Rhomni St. John and Noah Constant and Mario Guajardo Cespedes and Steve Yuan and Chris Tar and Yun-Hsuan Sung and Brian Strope and Ray Kurzweil. 2018. Universal Sentence Encoder. arXiv preprint arXiv:1803.11175 (2018).
[4] Tom B. Brown and Benjamin Mann and Nick Ryder and Melanie Subbiah and Jared Kaplan and Prafulla Dhariwal and Arvind Neelakantan and Pranav Shyam and Girish Sastry and Amanda Askell and Sandhini Agarwal and Ariel Herbert-Voss and Gretchen Krueger and Tom Henighan and Rewon Child and Aditya Ramesh and Daniel M. Ziegler and Jeffrey Wu and Clemens Winter and Christopher Hesse and Mark Chen and Eric Sigler and Mateusz Litwin and Scott Gray and Benjamin Chess and Jack Clark and Christopher Berner and Sam McCandlish and Alec Radford and Ilya Sutskever and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165 (2020).
[5] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023).

レビュー動画内における商品特徴の言及量分析システムの開発

山口　史弥†　　上田真由美††,†††　　中島　伸介††††

† 京都産業大学 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
†† 流通科学大学 経済学部　〒 651–2188 兵庫県神戸市西区学園西町 3-1
††† 大阪大学 サイバーメディアセンター　〒 560–0043 大阪府豊中市待兼山町 1 番 32 号
†††† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †i2286211@cc.kyoto-su.ac.jp, ††Mayumi_Ueda@red.umds.ac.jp, ††††nakajima@cc.kyoto-su.ac.jp

あらまし　オンラインショッピングには，購入した商品が期待したものと異なるといったリスクが存在する．商品の実際の使用感をより的確に把握することのできるレビュー動画を参考にするユーザも増えているが，投稿数が多く本当に知りたい情報を効率的に視聴するのは困難である．そこで本研究では，オンラインショッピングの際にどのレビュー動画を視聴すべきかを容易にするレビュー動画内における商品の特徴に関する言及量の可視化システムを開発した．レビュー動画内で述べられている商品の特徴分析を行い，言及量と言及箇所をユーザに提示することで，効率的な動画視聴を実現する．本稿では，レビュー動画の字幕とコメントに基づく商品特徴の言及量分析システムを開発したので報告する．

キーワード　レビュー動画，時系列分析，BERT，コメント分析，YouTube

1　はじめに

近年，Amazon[1]や楽天市場 [2]を中心としたオンラインショッピングの利用者が増加している（2020 年 3 月以降 2 人以上世帯の 50%が利用 [1]）．ただし，オンラインサイトでは実店舗と異なり，商品を手に取って確認することができない．オンラインショッピングでは，商品の説明画像やレビュー欄を参考に購入を判断することが一般的である．ただし，視覚的な大きさの比較ができない場合や，レビュー欄による文字だけの情報を参考に商品を購入する場合，購入したものが期待していたものと異なるリスクが存在する．そこで我々は，商品の感覚的な使用感をより詳細に伝えることのできるレビュー動画に着目した．レビュー動画は，従来のレビューよりも使用感をより的確に把握できる．そのため最近ではレビュー動画を参考にするユーザも増えている．

レビュー動画からは，実際に商品が使用されている様子を観ることができ，文字や商品画像のみでは伝わらない情報が得られる．ただし，動画の内容を事前に把握することは困難で，目的の動画のみを効率的に視聴することは難しい．オンラインショッピング時にレビュー動画を閲覧する場合，該当商品を購入するための判断基準となる商品特徴における項目（以下,「特徴項目」と呼ぶ）のうち，特定の特徴項目に対するレビューの有無が重要である．

以上の背景より，各レビュー動画内で商品のどの特徴項目が言及されるかユーザに提示することが好ましい．つまり，動画を再生する前に，動画内で言及されていることが事前に分かると，効率的に動画を視聴することができる．本稿では，オンラインショッピング支援を目的としたレビュー動画の商品特徴に関する言及量分析手法とその言及箇所の特定手法について説明する．また，提案手法の有用性を検証した精度評価実験とその結果，開発したレビュー動画分析システムとそのシステムユーザビリティについて検証した実験と結果について報告する．

以降，2 章では関連研究との比較を述べ，3 章ではレビュー動画内における商品特徴の言及量分析システムについて説明する．4 章では，言及量分析システムの有効性に関する評価実験について述べ，最後に 5 章でまとめと今後の課題を挙げる．

1：Amazon, https://www.amazon.co.jp/
2：楽天市場, https://www.rakuten.co.jp/

2　関連研究

以下に，商品レビュー動画の分析および商品スコアリングシステムに関する関連研究をあげ，本研究との差異を示す．

動画の内容を事前に把握することは困難で，目的の動画のみを効率的に視聴することは難しい．そこで，動画に付与されたコメントを基に，ユーザの直感に基づく動画検索を可能にする先行研究が行われている．

鴇田ら [2] は，YouTube の動画に対して，動画のタイトルや説明文などに含まれないユーザの直感に基づく動画検索のための評価項目別自動スコアリングを提案している．具体的には「じわる」「萌える」「エモい」などの動画の検索の実現のために，評価項目別に動画に付与されるコメントを分析し，動画を再生する前に真に求めている動画か否かを判断する手法を提案している．この研究では，直感に基づく動画検索を目的としているが，本研究ではレビュー動画の内容分析を目的としている．

このように，YouTube にアップロードされた動画に付与されたコメントに着目することで，データマイニングの狙いや傾向や関連性を見出す分析手法見出す目的を持った研究は盛んである．Ramamonjisoa ら [3] は，YouTube の動画コメントに対

して，機械学習のアルゴリズムを用いることで自動的にフィルタリングを行うコメント分析ツールを作成している．コメントを「有害」「スパム」「感傷的」「皮肉」「建設的」の 5 つの分類に取り組んだ．Muhammad ら [4] は，動画に付与されたコメントから肯定的な意見と否定的な意見を分類する取り組みを行っている．Pradhan ら [5] は，同様に自然言語処理を用いたコメント分類によって，当該動画の視聴者による評判を分かりやすくする取り組みを行っている．Siersdorfer ら [6] は，機械学習手法を用いた YouTube の動画等に対するコメント分析手法を提案している．包括的な意見の極性を抽出したり，将来高い評価を受ける可能性のあるコメントの予測，コメントからコミュニティに悪影響を及ぼす可能性のあるユーザーの分析を提案している．動画のコメントを分析することで，不明瞭なデータを予測する研究もある．Khan ら [7] は，未公開の映画に対して一般的な評価と成功確率を計算している．YouTube にアップロードされた予告編のコメントから，コメントの感情分析を行い，映画の成功を予測する機械学習の手法を提案している．このように映画や製品の成功は，公開前の肯定的または否定的な注目度に影響されると推測している．これらの研究では，YouTube に投稿された動画に付与されたコメントを分析することで，意見の分類や極性分析を行っている．本稿では，意見の分類や極性分析ではなく，動画内での商品特徴に基づく分析について取り扱う．

また，オンラインショッピングにおいてレビューや口コミは商品購入の重要な判断材料となっている．加えて，オンラインショッピングでは，ユーザの目的に沿った機能を持つ製品を探すには非常に時間がかかる．しかし，全てのレビューに目を通すことは非常に手間がかかる．これを解決すべく，内容や投稿者などの情報を用いて，自動で商品の評価を算出する研究が行われている．Scaffidi ら [8] は商品の特徴をレビューから抽出しスコアをつけることによって，オンラインショッピングを利用するユーザの要求を満たす製品の自動検索システムを提案している．Cao ら [9] は，消費者の商品レビューから製品改善のための特徴抽出手法に取り組んでいる．e コマースにおける電子製品のオンラインレビューのテキストマイニングを行い，消費者の好みの発見や企業の製品・サービス向上に貢献している．オンラインレビューから効果的に消費者の好みを見つけ出すことができ，企業はこれを活用して製品やサービスの品質を向上させ，消費者のニーズにより良く応え，消費者の消費を促進し，企業の目標と価値を達成することができると述べている．Haque ら [10] は，Amazon データセットを用いて教示あり学習手法によるレビューの極性化に取り組んでいる．Basani ら [11] は，商品レビューをポジティブ・ネガティブ・ニュートラルに分類し，商品レビューの要約を行っている．Zhang ら [12] は，商品レビューにおける極性分析と意見抽出において，テキストの主観性を考慮した検討を行っている．Tong ら [13] は，製品設計への活用を目的として，商品レビューから商品の使用感に関する文章を抽出している．Welgamage ら [14] は，自然言語処理による感情分析に着目し，ブログサイト等の大量のデータから商品の極性や機能の要約に取り組んでいる．Hananto ら [15] は，ファッションスタイルに関する EC サイトのレビューに着目し，潜在的ディリクレ配分法 (LDA) を用いたトピックモデリングを提案している．上述のように商品レビュー分析により，商品の評判情報の可視化を可能とする取り組みが多くある．これらの研究は，商品レビューにおける極性の分類やレビューの要約を目的としている．本稿では，単なる極性分析や要約ではなく，レビュー動画を対象とした商品特徴の分析について取り扱う．本研究は，商品の実際のイメージや雰囲気といった感覚的な情報を把握できるレビュー動画の視聴効率化を目的としたシステムの開発である．

さらに，レビュー分析に基づく評価表現別自動スコアリング手法に関する先行研究として，松波らは，コスメアイテムに特化した評価表現辞書を構築し，レビュー自動スコアリング手法を開発した [16]．同様に，谷口らはコスメアイテムに特化した評価表現辞書の改良に取り組み，Word2Vec に基づく評価表現辞書を構築することによる効率的な自動スコアリングについて検討している [17]．これらの研究では，コスメアイテムに特化した評価表現辞書を使用したレビューのスコアリングシステムを提案しているが，本研究ではレビュー動画を対象として特徴項目辞書を用いた，商品の言及量分析についても提案する．

以上で取り上げたような研究には，いずれもテキスト分析技術が使用されている．上述のように YouTube を含む SNS 上のユーザ投稿データの分析を目的とした場合のテキスト分析において， BERT を利用した取り組みが盛んである．Agarwal ら [18] は，Twitter や YouTube などのソーシャルメディアで生成されたデータに対して，BERT モデルを用いた感情分析の利用法を探っている．Fimoza ら [19] は，YouTube 上の映画レビューに対して，BERT を使用して感情分析を行っている．これらの研究では BERT を感情分析に用いている．本稿では，商品特徴の分析のために BERT を用いる．

このように，動画の内容やレビューの要約を目的とした取り組みは盛んに行われている．しかしながら，レビュー動画内の商品そのものの特徴分析を行う取り組みは十分に行われているとは言えない．また，動画内の言及箇所に着目した研究も少ない．このような背景から，本研究ではオンラインショッピング支援を目的としたレビュー動画内における商品特徴の言及量分析を行う．同時に，分析された商品の特徴が YouTube にアップロードされた動画の時系列上のどのエリアにて言及されているかに着目した分析も行う．

3 レビュー動画内における商品特徴に関する言及量分析手法

本章では，レビュー動画内における商品特徴に関する言及量分析手法について，量と箇所の 2 点の分析手法について述べるとともに，システム開発について説明する．

3.1 概　　要

本節では，提案システムの概要について説明する．

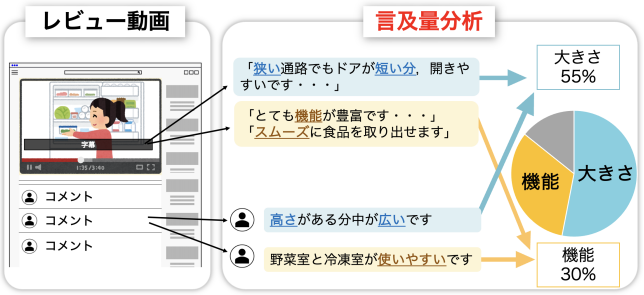


図 1　レビュー動画内の商品特徴の言及量分析手法処理イメージ

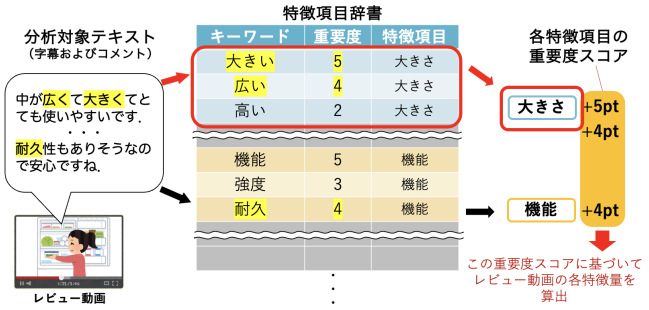


図 2　特徴項目辞書と言及量の自動スコアリング

近年では，動画共有サイトとして YouTube[3]が有名である．上記サイトには膨大な動画が日々アップロードされており，商品のレビュー動画の公開や視聴も盛んに行われている．オンラインショッピング時にレビュー動画を閲覧する場合，該当商品を購入するための判断基準となる複数の特徴項目のうち，特定の特徴に対するレビューの有無が重要である．そこで，各レビュー動画において，どの特徴項目がどの程度の割合でレビューされているか（以下,「言及量」と呼ぶ）を提示することを目指す．

図 1 には，商品特徴の言及量分析手法の処理イメージを示している．レビュー動画から取得した字幕情報とコメントを基に，レビュー動画から得られる情報を円グラフの形で割合としてユーザに提示している．

以降，レビュー動画における特徴項目の割合を提示するための言及量分析手法について 2 手法の開発について述べる．3.3 節で説明するのは辞書ベースの手法，3.4 節で説明するのは機械学習ベースの手法である．

3.2　字幕情報及びコメントの取得

レビュー動画内の商品特徴の言及量分析では，動画に付与された字幕情報およびその動画に対する視聴者のコメント文を分析する．したがって，分析対象の動画に付与された字幕情報とその動画に対して投稿されたコメントを取得する必要がある．本研究において開発したシステムでは，レビュー動画から自動で字幕とコメントを取得し，言及量分析を行っている．

字幕情報およびコメントの取得には　YouTube Data API[4]を使用した．取得した字幕とコメントは分析対象テキストとして次節以降で述べる言及量分析に使用される．

3.3　特徴項目辞書に基づく言及量分析手法の開発

3.3.1　辞書の構築

我々は，特徴項目辞書に基づくレビュー動画のコメント・字幕の自動スコアリング方式を開発している [20]．ここで開発した手法に基づき，図 2 の中央に示すような辞書を構築した．辞書には，商品特徴ごとに，その特徴項目を支持する単語が複数登録されている．登録されている単語は，後にレビュー動画の特徴項目を分析するの際のキーワードとなる．加えて，辞書に登録された単語にはそれぞれ重要度スコアが付与されている．

3 : YouTube,https://www.youtube.com/
4 : YouTube Data API,https://developers.google.com/youtube/v3

重要度スコアの付与は，それぞれの単語が各特徴項目に関連度が高いほど大きいスコアを付与する．以上のようにして構築した特徴項目辞書は後の言及量分析に用いられる．

3.3.2　言及量の算出

3.2 節で取得した分析対象テキストと 3.3.1 節で構築した特徴項目辞書を用いてレビュー動画における各特徴の言及量を算出する．まず，動画に付与された字幕文とコメント文における特徴表現の有無を，構築した辞書を用いて確認する．各文に特徴表現が存在する場合には，辞書を参照しどの特徴項目に対する表現かを判定する．次に，特徴項目辞書の重要度スコアを参照し，判定された特徴項目にその重要度スコアを付与する．以上を繰り返すことで，すべての字幕情報およびコメント文から各特徴の言及量を算出する．特徴項目別に付与された重要度スコアを集計することで言及量が算出可能になる．図 2 においては，レビュー動画から取得した分析対象テキストに対して，特徴項目辞書に登録されたキーワードと一致する表現が存在する場合，設定されている重要度スコアをその特徴項目に付与している．これを全ての文章に対して行うことで，最終的に特徴項目ごとに合計値を求め，このレビュー動画内での商品の各特徴の言及量が可視化できる．

3.3.3　言及量の割合変換

3.3.2 節で算出した言及量によるレビュー動画における各特徴項目の割合算出について述べる．3.3.2 節で算出された言及量は各言及量単体では意味を持たない．そのため，動画内での各特徴の割合を算出する．これを特徴項目割合として，最終的には円グラフの形でユーザに提示する．各特徴の言及量割合の算出式は次のとおりである．

$$\text{言及量割合} = \frac{\text{言及量}}{\text{全言及量の合計}} \qquad (1)$$

以上の処理を行うことで，図 1 に示すような商品特徴の言及量を割合として提示することを可能とする．

3.4　機械学習に基づく言及量分析手法の開発

3.4.1　学習モデルの構築

3.3 節における言及量分析は辞書を用いた静的な振る舞いをする．これに対し，テキスト分類処理を機械学習によって実現する手法を新たに開発した．使用した技術は Bidirectional Encoder Representations from Transformers (BERT) である．BERT による機械学習モデルの構築には，東北大学の訓練済み日本

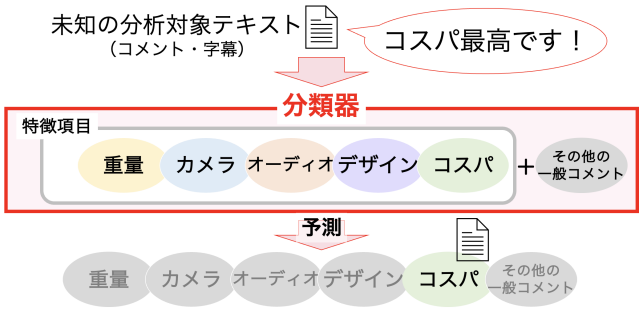


図 3　構築した BERT 学習モデルによる文書分類のイメージ

語 BERT モデル (Pretrained Japanese BERT models)[5]を使用した．

構築した機械学習モデルは，未知のコメントや字幕について，どの商品特徴に述べているか分類する特徴項目数+1(その他)クラス分類器である．次にファインチューニングについて述べる．ファインチューニングに使用したデータは，YouTube に公開されているレビュー動画数十本に付与されていたコメントデータである．複数人による手作業で学習データの収集を行った．収集した動画に付与されたコメントが，特徴項目に関することに触れられている場合は，各特徴項目ごとに収集し，ファインチューニングに用いた．また，収集したコメントが分類されるべき特徴項目とは関係が全くないと判断されたコメントは，「これらの情報を含まないその他の項目」に分類してファインチューニングに使用した．データ数は各特徴項目において約 100 コメントである．

3.4.2　言及量の算出

3.4.1 節で構築した学習モデルを用いて，各商品特徴の言及量を算出する．レビュー動画から収集したコメントを分類器に通すことで，そのコメントがどの商品特徴に分類される（もしくはどの商品特徴にも分類されない）ことが予測される（図 3 参照）．各特徴項目別に分類されたコメントの総数を算出し，円グラフとして表示できるよう割合に変換することで図 1 のような商品特徴言及量を提示を可能となる．

3.5　商品の特徴項目に関するレビュー動画内での言及箇所分析手法

本節では，レビュー動画内で商品特徴が言及されている箇所を特定するための言及箇所分析手法について述べる．

3.5.1　レビュー動画内の商品特徴に関する言及箇所提示の概要

3.3 節, 3.4 節の取り組みにより，特徴項目の言及量を分析することで各レビュー動画が着目している商品特徴が明確になる．これにより，ユーザは大量にあるレビュー動画から本当に視聴したい動画を選択することが容易になる．

しかし，図 4 のように動画の割合が提示されたとしても，動画のどの部分を視聴すれば良いかまでは分からない．それぞれの動画の最初から最後までを視聴するには時間がかかり，動画

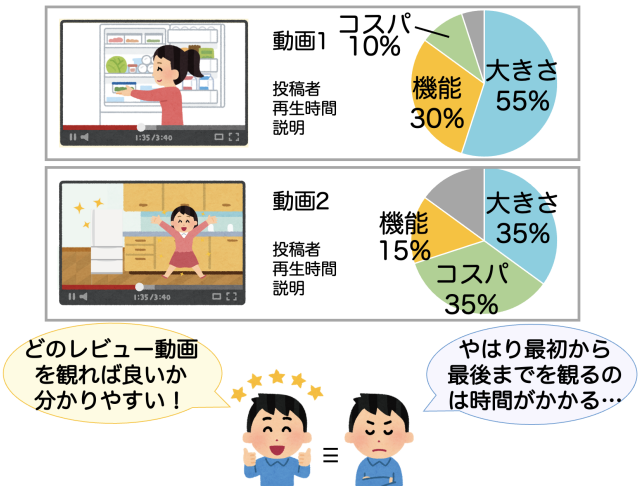


図 4　言及量分析結果の円グラフのみの提示による課題

図 5　特徴項目に関する言及箇所提示のイメージ

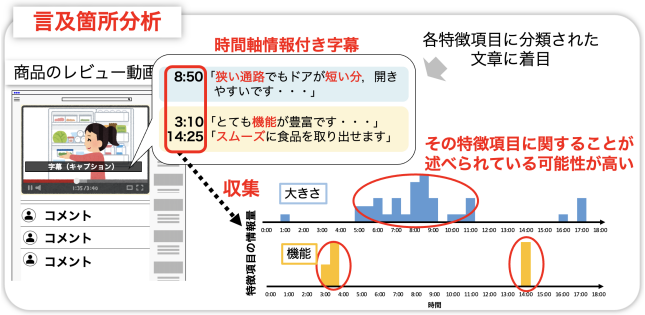


図 6　レビュー動画内の言及箇所の分析手法

内で自分の視聴したい場面を探すには手間がかかる．そこで，より短時間で知りたい情報を得るには図 5 のような言及箇所の提示が有効であると考える．次節ではユーザが視聴したいと思った特徴項目に関する言及箇所の特定手法について述べる．

3.5.2　言及箇所の特定方法

レビュー動画における各特徴項目の言及箇所抽出について述べる（図 6）．言及箇所分析に使用するのは，レビュー動画から取得できる時間軸情報付きの自動字幕である．各特徴項目に分類される字幕には，時間軸情報が付与されている．先程の，特徴項目辞書に登録されたキーワードが出現する時刻に着目する．出現時刻を特徴項目ごとに収集し，数秒区切りのヒストグ

5：東北大学自然言語処理研究グループ,https://www.nlp.ecei.tohoku.ac.jp/

ラム状のグラフに変換する．すると，この動画内で紹介されている商品のどの情報がどの部分で出現するか分析が可能になる．その特徴項目に関することが述べられている可能性が高い場所には山ができ，動画中のどこを視聴すべきかがすぐに分かる．

最終的にシステムが提示するこのヒストグラムは，キーワードの単純な出現回数ではない．特徴項目辞書に基づく言及量分析 [21] を基にしているため，辞書に登録されたスコアに基づいてキーワードの重要度は異なる．これによって，なめらかなヒストグラムを生成することが可能である．以上の処理を行うことで，特徴項目別に言及箇所を分析できる．

3.6 レビュー動画内における商品特徴に関する言及量分析システムの開発

本節では，3.5 節までの提案手法を基に開発した，レビュー動画内における商品特徴に関する言及量分析に関するシステム開発について概要を説明する．

3.6.1 言及量分析システムで使用するデータ

本システムで対象となるデータは YouTube 上に公開されているすべての動画である．本システムのデータは，YouTube から事前に取得した動画データや事前に分析したデータのみを対象とするのではなく，ユーザの操作（検索・視聴）に合わせてリアルタイム処理を実装した．

本システムでは 3.6.2 節に示す 5 つの商品カテゴリを分析対象とした．各商品カテゴリのレビュー動画において，レビュー動画ごとにそれぞれ 5 個の特徴項目を分析する．

提案手法における分析対象テキストのためのデータとして，各レビュー動画ごとに自動字幕と動画に対して投稿されたコメントを取得している．その他，提案システムにおいて動画データをユーザに提示するために，レビュー動画ごとのタイトルやサムネイル画像，再生時間等の詳細情報を取得し提示する．

3.6.2 言及量分析システムの分析対象とする商品カテゴリと特徴項目

レビュー動画内の商品特徴に関する言及量分析システムの評価対象とした商品カテゴリは次に示す 5 点である．

（1） 家電
（2） アウトドア
（3） 家具
（4） 精密機器
（5） コスメ

これらの商品カテゴリを分析対象としている理由は次のとおりである．ここに示すカテゴリにおける商品は，オンラインショッピングにおいて購入前に重要な情報が分からない可能性が高いものである．家具や家電は，大きさについて問題になる可能性が高い上，使用前に組み立てが必要な商品カテゴリとなる．精密機器やコスメは使用感がその商品の特徴を大きく左右する上，高価な商品である．

次に，これらの商品カテゴリに対して，言及量分析を行うための特徴項目を決定した．例えば，家電に関するレビュー動画の特徴項目は「機能・操作，デザイン・見た目，性能・能力，騒音・運転音，価格・コスパ」の 5 項目である．

図 7 開発されたレビュー動画分析システムでレビュー動画の検索結果を表示した様子

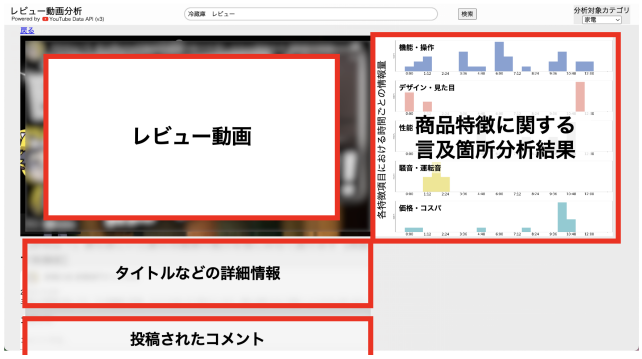


図 8 レビュー動画分析システムで動画の視聴と同時に言及箇所分析が提示されている様子

3.6.3 開発したレビュー動画内の商品特徴に関する言及量分析システムの概要

実際に開発したレビュー動画内の商品特徴に関する言及量分析システムにおける表示画面を図 7 と図 8 に示す．

図 7 に示すのは，言及量分析において家電のレビュー動画を検索した時の様子である．「冷蔵庫　レビュー」と検索すると，検索クエリにヒットした動画の一覧と言及量分析結果が表示される．左側には，分析対象となったレビュー動画のサムネイル画像やタイトル，動画の詳細情報が表示される．右側には，3 節で述べた言及量分析手法による，レビュー動画の商品特徴における言及量が提示される．商品特徴に関する言及量分析結果は割合として算出され，円グラフとしてパーセンテージで提示されている．

検索結果に表示された動画一覧から 1 つ動画を選択してクリックすると動画の視聴が可能になる．図 8 に示すのは，視聴する動画がクリックされた際に動画の視聴画面へ遷移した様子である．画面左側には動画の再生ウィンドウやタイトル，この動画に対して YouTube で投稿されたコメントが表示され，画面右側には 3.5 節で述べた言及箇所分析結果が表示されている．

4 評 価 実 験

本章では，3 節で述べたレビュー動画の言及量・言及箇所分析手法に関して行った評価実験とその結果について報告する．

表 1　BERT に基づく言及量分析結果の妥当性評価実験結果

レビュー動画のカテゴリ	言及量分析の精度（コサイン類似度）
家具	0.86
精密機器	0.88
アウトドア	0.75
家電	0.82
コスメ	0.63

表 2　BERT に基づく言及量分析結果の妥当性評価実験結果

商品カテゴリ	正解率 (%)
家電	96.65
家具	86.89
アウトドア	74.40
精密機器	84.44
コスメ	69.09

表 3　指示された情報を発見するまでにかかった時間の動画ごとの比較結果

動画 #	従来手法	提案手法
1	169.50	237.00
2	125.00	141.00
3	158.00	188.75
4	183.75	118.00
5	143.75	118.00
6	154.75	60.00
7	146.50	145.25
8	162.00	64.25
平均	155.41[sec]	134.03[sec]

表 4　指示された情報を発見するまでにかかった時間の被験者ごとの比較結果

被験者 #	従来手法	提案手法
1	105.25	73.50
2	149.75	222.75
3	138.75	105.50
4	141.75	247.25
5	172.25	71.00
6	106.25	95.25
7	201.50	133.25
8	227.75	123.75
平均	155.41[sec]	134.03[sec]

表 5　5 段階主観評価の比較

質問内容	従来手法	提案手法
すぐに答えを発見できた	3.44	4.56
関連する箇所だけを視聴できた	2.72	4.34
頻繁に行いたい	2.59	4.53
平均	2.92	4.48

実施した評価実験は 3 点である．本節では，行った評価実験，言及量分析の妥当性評価と BERT の学習モデルの精度評価・言及箇所提示の有効性評価についてそれぞれ概要と結果を述べる．

4.1　BERT に基づく言及量分析手法の妥当性評価

本節では BERT に基づくレビュー動画の商品特徴の言及量分析手法について，その妥当性を評価するための実験概要と結果を述べる．

4.1.1　実 験 概 要

被験者にはレビュー動画を視聴してもらい，各特徴項目の情報が感覚的にどの程度含まれていたかを回答していただいた．評価された動画の本数は 1 カテゴリあたり 12 本（合計 60 本）であり，被験者は 20 代の男女 8 名である．評価に用いる動画の長さは，10 分〜15 分程度のものにした．被験者 1 人当たり 5 つのカテゴリより数本ずつ動画を指定し，合計 16 本の動画について回答していただいた．1 本の動画をランダムに平均 3 名が評価するように設定した．被験者の回答した結果の平均値を Ground Truth データとして使用する．最終的に Ground Truth による特徴項目割合と，提案手法による特徴項目割合を比較し，どの程度一致しているかを検証する．両者の数値が一致しているほど提案手法の有効性が高いと言える．

Ground Truth による特徴項目割合と提案手法による特徴項目割合の比較尺度には，コサイン類似度を使用する．コサイン類似度は，それぞれの要素にマイナスは存在しないため 0 から 1 の間で算出される．つまり， 0.5 より大きい値であれば両者は類似していると判断できる．この場合，両者の数値には類似性があることから提案手法による言及量分析の有効性が確認できる．

4.1.2　実 験 結 果

カテゴリ別に各レビュー動画に対するコサイン類似度の平均値を表 1 に示す．

表 1 より，全てのカテゴリにおいてコサイン類似度は中央値である 0.5 を上回った．これは提案手法によって商品特徴の言及量は正しく分析できていることを示す．

4.1.3　考　　察

結果より，コサイン類似度は全てのカテゴリにおいて中央値 0.5 を上回り，提案手法の高い精度を確認した．これは，実際に被験者がレビュー動画を視聴して作成した Ground Truth と一致していることが確認できたことになる．完全な一致とならない点においては，コメント数や字幕量に依存するためであると考える．

4.2　BERT 学習モデルの精度評価

本節では，3.4 節で開発した BERT 学習モデルの精度評価について述べる．

4.2.1　実 験 概 要

テストデータに対するファインチューニング後の BERT 学習モデルによる文書分類精度の評価を行った．正解率を用いて実施した．開発を行った 5 つの商品カテゴリに対する BERT 学習モデルの全てに対して評価を行った．

4.2.2　実 験 結 果

テストデータに対するファインチューニング後の BERT 学習モデルによる文書分類精度の評価を行った．正解率を表 2 に示す．

4.2.3　考　　察

結果より，評価を行った全ての BERT 学習モデルの精度は

表 6　提案手法に対する有用性と分かりやすさの評価結果

評価項目	質問	スコア
有用性	特徴項目の述べられている箇所を示すグラフは役立ったと思う	4.69
理解容易生	特徴項目の述べられている箇所を示すグラフは分かりやすかった	4.47

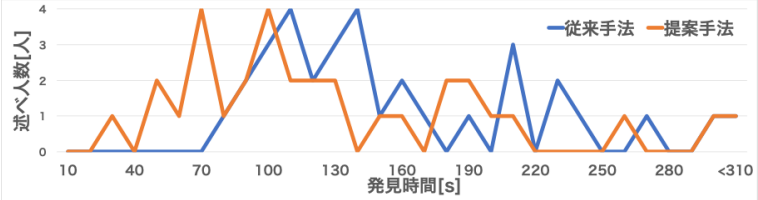


図 9　全ての経過時間のまとめと比較

非常に高く，正しく文書を分類できていることが分かる．高い文書分類精度は，レビュー動画内の言及量分析における精度も向上させる．今回は，1 文書をいずれかの商品特徴に分類するように実施し，言及量分析手法に利用した．ただし，今後より高いレビュー動画内の商品特徴分析においても精度を求めるためには，分析対象テキストがどのクラスに分類されるかの確率を算出し利用することも考えられる．また，各特徴項目別に分類器を作成し，2 クラス分類のモデルを特徴項目数分用意することでより高い言及量分析精度につながるとも考える．これらの検証は今後の課題であるが，いずれにしても BERT を用いることで商品特徴における文書分類タスクにおいて，高い文書分類精度を期待できることが確認できた．

4.3　言及箇所提示の有効性評価

4.3.1　実 験 概 要

提案手法の妥当性を評価するため，従来手法と提案手法の比較を行った．言及箇所の提示のある状態でのレビュー動画の視聴が提案手法である．これに対して，言及箇所の提示のない状態でのレビュー動画の視聴が従来手法である．それぞれの情報収集にかかる時間や煩わしさを比較する．実験の具体的な流れは次のとおりである．

（1） 動画内で商品がどのようにレビューされているのかクイズを出題する

（2） 被験者にレビュー動画内から指示された情報を探していただく

（3） 被験者が問いに対する答えを発見するまでの時間を計測する

（4） 被験者は以上の取り組みについて主観評価（5 段階）で煩わしさを回答する

この流れを，従来手法と提案手法それぞれで行う．上記の手順で計測した「時間」と「主観評価」に基づき，比較を行う．「主観評価」は，動画内での情報探索における煩わしさを評価するための指標として，以下の 3 つの質問を行い 5 段階（1．全く当てはまらない ～ 5．とても当てはまる）で回答していただいた．

- すぐに答えを発見できた
- 関連する箇所だけを視聴できた
- 頻繁に行いたい

「時間」の比較に関しては，従来手法に比べて提案手法の時間が短くなると，有効性を確認できる．「主観評価（5 段階）」に関しては，従来手法に対して提案手法の得点が高くなると，有効性が確認できるように設定した．

この実験において，参加者は従来手法と提案手法で同じ動画を評価することはない．つまり，一度動画を視聴したことがある状態でもう一度評価に取り組むと，後に実施した方が有利になることが考えられるが，この問題が発生することがないように実験を行った．

4.3.2　実 験 結 果

まず，レビュー動画内から指示された情報を発見するまでにかかった時間を表 3 と表 4 に示す．これらの表に示すのは，従来手法で必要とした時間 [s] と提案手法で必要とした時間 [s] である．動画ごとと被験者ごとに算出した．

表 3 と表 4 より，動画 1 本に対して被験者 1 人当たりの平均値は，従来手法で 155.4[s]，提案手法で 134.0[s] であった．必要な情報を発見するための時間について，8 本中 5 本の動画において，従来手法より提案手法の方が短時間となった．また，8 人中 6 人の被験者において，提案手法の方が短時間となった．

表 5 に，このタスクに対する参加者の主観評価の結果を示す．表 5 より，全ての評価項目で提案手法が従来手法を上回る結果となった．全ての質問において有意水準 0.1%($\alpha = 0.001$)t 検定を行い，優位差があることが確認した．

4.3.3　考　　察

以上の結果より，5 段階主観評価では全ての質問において提案手法の有効性が強く示された．また，実際に必要とされた時間も，提案手法は従来手法に比べて短くなっている．これは，答えを発見するまでの時間が短くなり，本システムによる言及箇所提示が有効であったことを示す．提案手法の有効性は，全被験者（述べ）の発見時間の折れ線グラフ（図 9）からも明らかである．

時間比較において大きく差が開かなかった原因には次の 3 つの要因が考えられる．(1) 一部の分析結果が適切でなかった．(2) 実験概要や「言及箇所分析グラフ」の見方の理解に時間がかかった．(3) 提案手法は従来手法に比べて設問数が数個多かったため，回答に時間がかかった．しかし，同じ設問が続くので途中からは設問を読む必要がなくなった．

まとめると，レビューとしてテキストや画像ベースのもののみを確認するよりも，提案手法によって動画を有利とする感覚的な情報を得られる可能性は高まると考える．

5 おわりに

本稿では，オンラインショッピング支援のためのレビュー動画の商品特徴に関する言及量と言及箇所の分析手法について提案し，システムの開発を行った．BERT に基づくレビュー動画の言及量分析は，少量の学習データにおいて文書分類を可能にすることから辞書に比べて少量のメンテナンスで，閲覧すべきレビュー動画の判定を容易にする．言及箇所分析はレビュー動画の時間的な視聴効率を実現する．提案手法に関する評価実験の結果は高評価であり，これらの提案手法の有効性を確認することができた．

今後の課題は，開発した提案システムのユーザビリティ評価と UI の改良である．また，評価実験結果を基にした，提案手法の機能追加と改善を実施する予定である．加えて，本研究の本来の目的であるオンラインショッピング支援のための有効性の検証実験を行う．

謝　辞

本研究の一部は，科研費 (課題番号：19K12243，20H04293，22K12281) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　献

[1] デジタルで支える暮らしと経済，令和 3 年 情報通信白書，総務省
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/html/nd121310.html

[2] 鴇田 和士，上田 真由美，中島 伸介，動画に対する評価項目別スコアを用いた直感的動画検索システム，第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022), C21-5, 2022.

[3] David Ramamonjisoa, Hidemaru Ikuma, Riki Murakami, “Filtering relevant comments in social media using deep learning,” 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 335-340, 2022.

[4] Abbi Nizar Muhammad et al., “Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier,” International Conference on Computer Science, Information Technology, and Electrical Engineering(ICOMITEE), 1570571406, 2019.

[5] Rahul Pradhan, “Extracting Sentiments from YouTube Comments,” 2021 Sixth International Conference on Image Information Processing (ICIIP), pp. 142-145, 2021.

[6] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl, “Analyzing and mining comments and comment ratings on the social Web,” ACM Transactions on the Web, Volume 8 Issue 3 Article No. 17, pp. 1-39, 2014.

[7] Farden Ehsan Khan, Ahmed Mahir Ruhan, Rifat Shamsuddin, Faisal Bin Ashraf, “A machine learning approach to predict movie success from Youtube trailer comments,” 2022 25th International Conference on Computer and Information Technology (ICCIT), pp. 372-377, 2022.

[8] Christopher Scaffidi et al., “Red Opal: Product-Feature Scoring from Reviews,” Proceedings of the 8th ACM conference on Electronic commerce (EC ’07), pp. 182–191, 2007.

[9] Jihua Cao, Jie Li, Miao Yin, Yunfeng Wang, “Online reviews sentiment analysis and product feature improvement with deep learning,” ACM Transactions on Asian and Low-Resource Language Information Processing, pp. 1-17, 2022.

[10] Tanjim Ul Haque et al., “Sentiment Analysis on Large Scale Amazon Product Reviews,” IEEE International Conference on Innovative Research and Development, 2018.

[11] Yuniarta Basani et al., “Application of Sentiment Analysis on Product Review E-Commerce,” International Conference on Advance and Scientific Innovation (ICASI), 1175 (2019) 012103, 2019.

[12] Zhu Zhang, Balaji Varadarajan, “Utility Scoring of Product Reviews,” Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM ’06), pp. 51–57, 2006.

[13] Yanzhang Tong, Yan Liang, Ying Liu, Irena Spasić, Yulia Hicks, “Understanding context of use from online customer reviews using BERT,” 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), pp. 1820-1825, 2022.

[14] V. R. Welgamage, U. A. C. Senarathne, N. H. A. C. Madhubhashani, T. C. Liyanage, P. P. G. Dinesh Asanka, “Overall and Feature Level Sentiment Analysis of Amazon Product Reviews Using Machine Learning Techniques and Web-Based Chrome Plugin,” 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE), SC-31, 2022.

[15] Valentinus Roby Hananto, Soomin Kim, Mate Kovacs, Uwe Serdült, Victor Kryssanov, “A Machine Learning Approach to Analyze Fashion Styles from Large Collections of Online Customer Reviews,” 2021 6th International Conference on Business and Industrial Research (ICBIR), pp. 144–151, 2021.

[16] Yuuki Matsunami et al., “Explaining Item Ratings in Cosmetic Product Reviews,” International MultiConference of Engineers and Computer Scientists 2016 Vol I(IMECS 2016), 2016.

[17] Yuna Taniguchi et al., “A Research on Constructing Evaluative Expression Dictionaries for Cosmetics Based on Word2Vec,” The 23rd International Conference on Information Integration and Web-based Applications & Services (iiWAS2021), 2021.

[18] Madhav Agarwal, Prince Kumar Chaudhary, Shubham Kumar Singh, Charvi Vij, “Sentiment analysis dashboard for socia media comments using BERT,” 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), pp. 292-298, 2023.

[19] Dwi Fimoza, Amalia Amalia, T. Henny Febriana Harumy, “Sentiment analysis for movie review in Bahasa Indonesia using BERT,” 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), pp. 27-34, 2021.

[20] 山口 史弥, 上田 真由美, 中島 伸介, レビュー動画内における商品特徴量分析手法の開発, 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), 1b-1-3, 2023.

[21] Fumiya Yamaguchi, Felix Dollack, Mayumi Ueda, Shinsuke Nakajima, “Feature relevance analysis of product reviews to support online shopping,” The 24th International Conference on Information Integration and Web Intelligence (iiWAS2022), pp. 441-446, 2022.

不快コメント予測の個人化のための コメント特徴に基づく個人化レベルの調整

中原　輝樹†　　牛尼　剛聡††

† 九州大学 大学院芸術工学府　〒 815-8540 福岡市南区塩原 4-9-1
†† 九州大学 大学院芸術工学研究院　〒 815-8540 福岡市南区塩原 4-9-1
E-mail: †nakahara.teruki.528@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　SNS やニュースサイトを利用してニュース記事を読む際，読者はニュースに対する他者のコメントを読むことができる．読者は他者のコメントを読むことにより，ニュースに対する世論を理解でき，ニュースの全体像の把握に役立つ．しかし，他者のコメントには読者を不快に感じさせる不快なコメントが含まれることがあり，ユーザの体験の質を低下させる．本研究では，コメントの読者の価値観が多様であることを考慮し，読者が過去に低評価したニュースコメントのフィードバックに基づき，読者ごとに不快なコメントを予測する．提案手法では，コメントの内容によって個人化の重要度が異なることに注目する．各読者に対してコメントの不快度を予測する際，予測対象のコメント特徴をもとに個人化レベル（個人化を行う度合い）をコメントごとに調整する．評価実験の結果，コメント特徴に基づいて個人化レベルをコメントごとに調整する処理を予測モデルに組み込むことによって，不快コメント予測の個人化の性能が向上することが示された．

キーワード　重み付き連結，フィルタリング，パーソナライゼーション，ソーシャルメディア

1　はじめに

近年，インターネット上でニュースを読むことが一般化した．新聞通信調査会によるメディアに関する全国世論調査 [1] では，一週間に 1 日以上インターネット上でニュースを読むと回答したのは，全世代の 73.1 %に及び，40 代以下では 90 %を超えている．ポータルサイトや，テレビ局・新聞社が運営するニュースサイトだけでなく，SNS(ソーシャル・ネットワーキング・サービス) 上でニュースを読む人々も多い．特に，10 代・20 代の若い世代では，インターネット上でニュースを読む際にアクセスするサイトとして，SNS が最も多く利用されている [1]．

インターネット上でニュースを読む際，読者はニュースだけではなく，ニュースに対する他者のコメントを読むことができる．日本における代表的なニュースサイトの一つである Yahoo! ニュース [1]では，ニュース記事毎に，ニュースに対するコメントを投稿することができるコメント欄が設けられている．また，代表的な SNS の一つである X（旧 Twitter）[2]では，テレビ局・新聞社などのアカウントが，ニュースやニュース記事に対するリンクをポストとして投稿している．ユーザは，ニュースサイトにおけるコメント欄のように，ニュースに対するコメントをポストへの返信として投稿できる．

Stroud ら [2] の調査では，インターネット上でニュースを読む人の 49 %がニュースに対するコメントを読むと回答している．他者のコメントを読むことにより，ニュースに対する世論を理解し，ニュースの全体像を捉え，ニュースに対する理解を深めることができる．一方，ニュースコメントなどのオンラインコミュニティでは，攻撃的な内容や誹謗中傷を含む発言などの読者にとって不快と感じるコメントが投稿される問題がある．読者が不快と感じるような有害なコメントは，有害という意味を表す toxic という単語から toxic comment と呼ばれる [3]．不快なコメントが投稿されることによって，読者がコメントを閲覧する際の体験の質が低下する恐れがある．

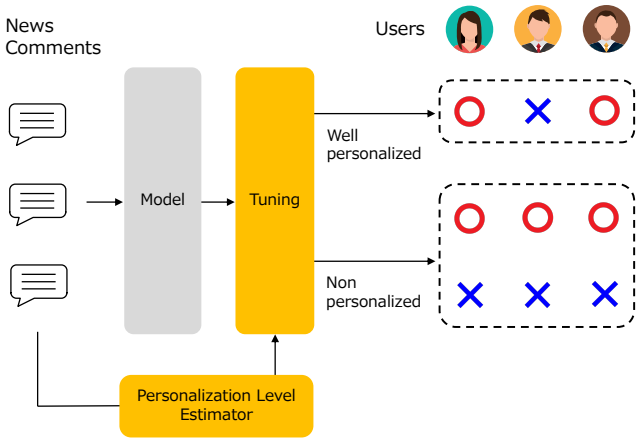


図 1: 個人化の重要度はコメントごとに異なるため，コメントの内容に応じて予測の個人化レベル（個人化を行う度合い）を調整する

コメント閲覧時における読者の体験の質を向上させるために，不快なコメントを予測する手法が提案されている [4–12]．これらの手法では，コメントのテキストなどを入力し，対象のコメントが不快であるかを予測する機械学習モデルを利用している．しかし，コメントの読者の価値観などの違いによって不快なコメントの判定基準は異なる [13]．Nakahara ら [14] はこれまで

1：https://news.yahoo.co.jp/
2：https://twitter.com/

に不快コメントの予測において個人化（パーソナライゼーション）を行い，予測性能の向上を示している．本論文では，コメントの内容によって個人化の重要度が異なることに注目する．

著者らはコメントの内容による評価の個人差について調査した．クラウドソーシングサービスであるクラウドワークス [3]でアンケートを実施し，ニュースコメントに対するラベル付けを行った．アンケートでは 2022 年 4 月 1 日から 6 月 30 日の 3ヶ月間に投稿されたニュースコメント 2000 件を 50 名のユーザに提示し，不快と感じるかについて {1: 全くそう思わない， 2: あまりそう思わない，3: どちらとも言えない，4: ややそう思う，5: 非常にそう思う } の 5 段階で回答してもらった．また，評価値が 4 以上であれば「不快である」，4 未満であれば「不快でない」とした．表 1 に読者の評価が一致するコメントの例を，表 2 に読者の評価が一致しないコメントの例を示す．表 1 のように，不快な内容を全く含まないコメントや，極度に不快な内容を含むコメントに対する評価の個人差は小さい．一方，表 2 のように，読者によって捉え方が異なるような内容を含むコメントに対する評価の個人差は大きい．このため，図 1 のように，コメントの内容に応じて予測における個人化レベル（個人化を行う度合い）を調整することによって予測性能の向上が期待される．

本論文では，読者ごとにコメントの不快度を予測する際，予測対象のコメントに応じて個人化レベルを調整する機械学習モデルを提案する．具体的には，予測対象のコメントと読者のベクトルを全結合層に入力する際，各ベクトルの重み付き連結を行うことにより個人化レベルを調整する．重みのパラメータを算出するために，予測対象のコメントの不快度の平均値や標準偏差の予測値などを活用する．実験では，実際に投稿されたニュースとコメントからなるデータセットを利用して提案手法の有効性を示す．

本論文の貢献は，以下の通りである．

- 不快コメント予測の個人化において，コメントの内容によって個人化の重要度が異なることに注目し，コメントの内容に応じて個人化レベルを調整するための重み付き連結を行う機械学習モデルを開発した．
- 実際に投稿されたニュースとコメントからなるデータセットを利用して提案手法が有効であることを示した．

本論文の構成は次の通りである．2 章で関連研究について述べる．3 章では提案手法について説明する．4 章では評価実験について述べる．5 章で本論文のまとめを述べる．

2 関 連 研 究

2.1 ニュースコメントの理解支援

これまでにニュースコメントの閲覧を支援するためのいくつかの研究が行われている．Ma ら [15] は，代表的なトピックモデルの 1 つである LDA(Latent Dirichlet Allocation) [16] を用いてコメントのトピックを推定することで，ニュースコメントのクラスタリングを行っている．Aker ら [17] は，コメント間の類似度をグラフベースの手法でモデル化することで，ニュースコメントのクラスタリングを行っている．これらの研究では，クラスタリングによってニュースコメントの全体像を読者に理解させることでニュースコメントの閲覧支援を行っている．それに対して，本論文での提案手法では，攻撃的な内容や誹謗中傷を含む発言などの不快なコメントを事前に予測することで読者の体験の質を向上させ，ニュースコメントの閲覧支援を行うことを目的とする．

2.2 不快コメント予測

これまでに SNS などのオンラインコミュニティにおける不快なコメントを予測する研究や実装が行われている [4–12]. Georgakopoulos ら [4] は，Wikipedia における編集者の議論ページを対象に，CNN を用いた機械学習モデルによって不快なコメントの予測を行っている．Hessel ら [6] は，Reddit[4]を対象に LSTM [18] や BERT [19] などのモデルを用いて，論争的な投稿と非論争的な投稿の分類を行っている．Saveski ら [7] は，Twitter における会話の構造と投稿の有害性との関係について分析している．Google Jigsaw は不快なコメントを検出するシステムを Perspective API として提供している [5]．これらの手法ではコメント読者の価値観などの違いを考慮した予測が行われていないのに対して，本論文では予測の個人化を行うことでコメント読者の価値観などの違いを考慮する予測手法を提案する．

2.3 不快コメント予測の個人化

2.2 節で述べた不快コメント予測に関する従来の研究では，アノテーターの評価の多数決などで決定された正解ラベルを予測することを目的としている．しかし，アノテーターの価値観などの違いによって不快なコメントの判定基準は異なる．Pavlic ら [20] と Palomaki ら [21] はアノテーターの評価の不一致は必ずしもノイズではないことを指摘している．Sap ら [13] は，評価者のアイデンティティや信条と，コメントに対する不快度の評価との間に関連性があることを示している．これらの研究は，コメント読者の価値観などの違いを考慮し，不快コメントの予測において個人化（パーソナライゼーション）を行うことの重要性を示している．

不快コメント予測における個人化の重要性を背景として，実際に予測の個人化を行う研究が行われている．Kumar ら [22] は，コメントの読者が過去に有害と評価したコメントのフィードバックを利用し，Google Jigsaw が提供する Perspective API のパラメータを読者ごとにチューニングすることによって予測性能が向上することを示している．Nakahara ら [14] は，読者が過去に低評価したコメントだけでなく，低評価したコメントの投稿者の特徴を活用する機械学習モデルを用いた予測を行っている．これにより，読者が不快と感じるコメントの投稿者の傾向を捉えることができ，不快コメント予測の性能が向上する．コメントの投稿者の特徴に関して，対象の投稿者が過去に投稿

3：https://crowdworks.jp/

4：https://www.reddit.com/

表 1: 読者の評価が一致するコメントの例

ニュース	コメント	読者の評価	
		不快でない	不快である
悠仁さま 筑波大学附属高校に入学	ご入学おめでとうございます。	50	0
低所得世帯 子ども 1 人当たり 5 万円支給案など検討	低所得なんかほっとけ	6	44

表 2: 読者の評価が一致しないコメントの例

ニュース	コメント	読者の評価	
		不快でない	不快である
東電管内 あすの節電呼びかけ	2000 キシダポイントもらう為に節電するぞー！頑張って熱中症ぶっ倒れ狙って病院や周りにガンガン迷惑かけるぞー	22	28

したニュースコメントを入力とした機械学習モデルを利用することにより取得している [23].

一方，表 1・表 2 のようにコメントの内容によって個人化の重要度は異なるが，これらの手法では個人化レベルがコメントごとに調整されていない．本論文では予測対象のコメントに応じて個人化レベルを調整する機械学習モデルを提案し，不快コメント予測の性能向上を目指す.

3 提 案 手 法

3.1 提案手法の概要

本論文では予測対象のニュースコメントと，読者が過去に低評価したニュースコメントのフィードバックに基づいて不快なコメントを読者ごとに予測する．本論文で提案する手法の概要を図 2 に示す．提案手法では予測対象のコメントに応じて個人化レベルを調整する処理を組み込む．ここで，読者 r にとってのコメント c の不快度を予測する機械学習モデルは以下の式 (1) で表される.

$$y = f(\boldsymbol{x}; \theta) \tag{1}$$

ここで $\boldsymbol{x}$ はコメント c と読者 r のペア (c, r) のベクトル表現，y は不快度を表すスカラー値，θ はパラメータである．提案モデルの学習用データセット D は，入力 $\boldsymbol{x}$ と出力 $y \in \{1, 2, 3, 4, 5\}$ のペア $(\boldsymbol{x}, y)$ の集合として，以下の式 (2) で定義される.

$$D = \{(\boldsymbol{x}, y) \mid \boldsymbol{x} = (1-\alpha)em(c) \oplus \alpha em(r), c \in C, r \in R\} \tag{2}$$

C と R はそれぞれコメントの集合と読者の集合を表す．$em(c)$ と $em(r)$ は，それぞれコメント c と読者 r の埋め込みベクトルを表す．$\oplus$ はベクトルの連結を行う演算子であり，$\boldsymbol{x}$ はコメント c の埋め込みベクトルと読者 r の埋め込みベクトルを連結したものである．α は個人化レベルの調整を行うためのパラメータであり，$0 \leqq \alpha \leqq 1$ を満たす．重み α は 3.5 節で説明する Personalization Level Estimator で算出する．コメントと読者の埋め込みベクトルの重みを変化させることにより，予測においてコメントと読者のどちらを重要視するべきかをコメントの内容に応じて調整することができる.

3.2 News Comment Encoder

News Comment Encoder は，予測対象とするコメントの埋め込みベクトルを生成する機能である．本論文では，コメントとコメント元のニュースのテキストを BERT に入力することで得られるベクトルを以下の式 (3) で表される関数として定義する.

$$enc^{cmt}(c) = \mathrm{BERT}(c, news(c)) \tag{3}$$

ここで $news(c)$ はコメント c が参照するニュースのテキストである．BERT によるコメントとコメント元のニュースのテキストのベクトル化については 3.4 節で述べる．本論文では，予測対象のコメント c の埋め込みベクトル $em(c)$ を以下の式 (4) として定義する.

$$em(c) = enc^{cmt}(c) \tag{4}$$

3.3 Reader Encoder

Reader Encoder は対象とする読者の埋め込みベクトルを得るための機構である．ここでは，読者が過去に低評価したコメントの情報を格納するデータベース（フィードバックデータベース）が利用可能であるとする．フィードバックデータベースを利用することで読者 ID に対してそのユーザが過去に低評価したコメントとコメント元のニュースを取得できる.

コメント読者の埋め込みベクトルを表現するため，読者 ID に基づいてフィードバックデータベースから対象の読者が過去に低評価したコメントとコメント元のニュースのペアのテキストを N 件取得し，BERT に入力することで N 個のベクトルを得る．ここで，N は取得するフィードバックの数を決定するハイパーパラメータである．いま，読者 r が低評価したコメントの集合を $toxic(r) = \{c_1, \ldots, c_N\}$ と表現する．読者 r に対する不快コメントのベクトル集合 $TC(r)$ を以下の式 (5) で定義する.

$$TC(r) = \{enc^{cmt}(c) \mid c \in toxic(r)\} \tag{5}$$

Attention 機構 [24] を用いて予測対象のコメント c のベクトル $enc^{cmt}(c)$ と，対象の読者 r が過去に低評価したコメントのベクトルの集合 $TC(r)$ の関連度を算出し，$TC(r)$ の加重平均を計算する．$enc^{cmt}(c)$ と $TC(r)$ の関連度を算出することで $TC(r)$ の加重平均を計算する式を以下の式 (6) で定義する.

$$enc^{toxic}(c, r) = \sum_{i=1}^{N} W_i \cdot e(TC(r), i) \tag{6}$$

ここで，$e(S, i)$ は集合 S の i 番目の要素を表し，W_i を以下の式

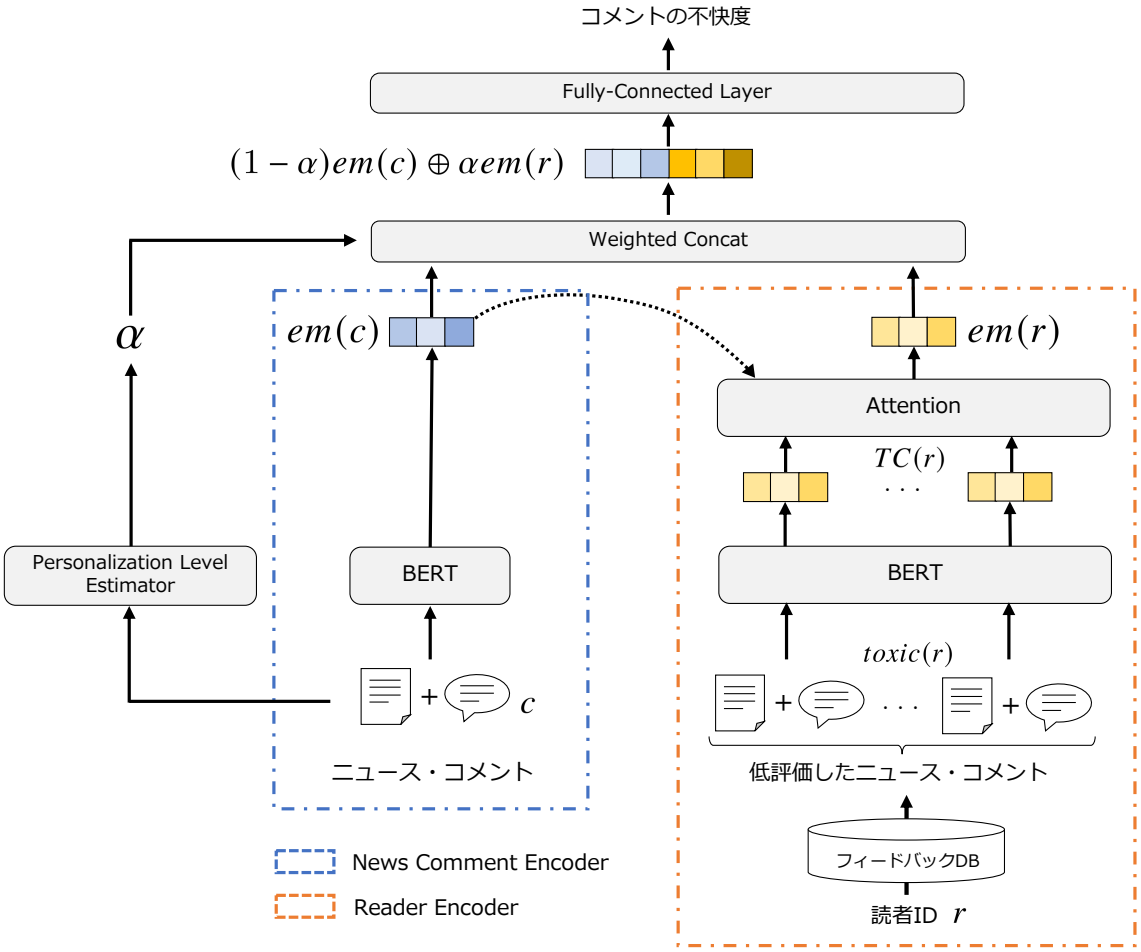


図 2: 提案手法の概要．予測対象のコメントと読者のベクトルを全結合層に入力する際，重み付き連結を行うことにより個人化レベルを調整する．

(7) で定義する．

$$W_i = \mathrm{Softmax}\left(\frac{qk_i^\top}{\sqrt{d}}\right) v_i \tag{7}$$

d は q, k_i の次元数であり，q, k_i, v_i をそれぞれ以下の式 (8), (9), (10) で定義する．

$$q = W^q \cdot enc^{cmt}(c) \tag{8}$$

$$k_i = W^k \cdot e(TC(r), i) \tag{9}$$

$$v_i = W^v \cdot e(TC(r), i) \tag{10}$$

ここで，W^q, W^k, W^v は学習可能なパラメータである．本論文では，コメントの読者 r の埋め込みベクトル $em(r)$ を以下の式 (11) として定義する．

$$em(r) = enc^{toxic}(c, r) \tag{11}$$

3.4 BERT によるニュースとコメントのベクトル化

本節では，BERT によるニュースとコメントのベクトル化について説明する．BERT は単語同士の関係を考慮することで，入力文を文脈に応じたベクトルに変換することができ，文章分類などの様々な自然言語処理のタスクを高い精度で処理可能である．提案手法では，ニュースとコメントのテキストを以下の形式で結合し，トークンごとに分割したテキストを BERT への入力とする．

[CLS] ニュース文 [SEP] コメント文 [SEP]

ここで，[CLS] は入力文の先頭を表す特殊トークンであり，[SEP] は文のペアの境界や，入力文の終わりを表す特殊トークンである．BERT では，Transformer [24] の Encoder における Self-Attention 機構により，入力データ内の単語どうしの関係を考慮したベクトルを出力する．提案手法では，最終層の Transformer Encoder の [CLS] トークンの出力ベクトル（768 次元）を BERT の出力として用いる．

3.5 コメントに応じた個人化レベルの調整

本節では，予測対象のコメントの内容に応じて個人化レベルの調整を行う処理について説明する．表 1 のように，不快な内容を全く含まないコメントや，極度に不快な内容を含むコメントに対する評価の個人差は小さい．一方，表 2 のように，読者によって捉え方が異なるような内容を含むコメントに対する評価の個人差は大きい．このため，コメントの内容に応じて予測

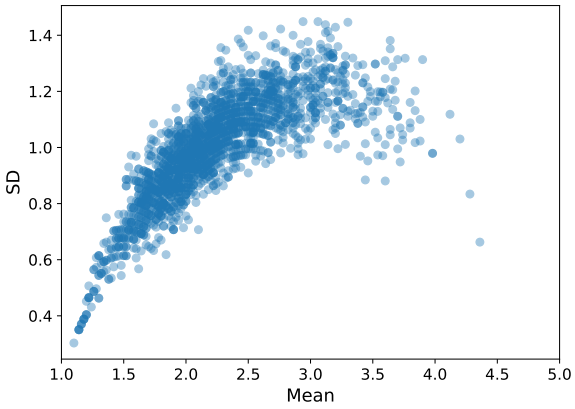


図 3: コメントの不快度の平均値と標準偏差の関係

における個人化レベルを調整することによって予測性能の向上が期待される．

提案手法では，式（4）で表される予測対象のコメントの埋め込みベクトルと，式（11）で表されるコメントの読者の埋め込みベクトルを連結する際に，コメントの内容に応じて各ベクトルの重みを調整する．コメントに応じた個人化レベルの調整のための重み付き連結は，重み α $(0 \leqq \alpha \leqq 1)$ を用いて以下の式（12）で表される．

$$(1-\alpha)em(c) \oplus \alpha em(r) \tag{12}$$

コメントと読者の埋め込みベクトルの重みを変化させることができるため，予測においてコメントと読者のどちらを重要視するべきかをコメントの内容に応じて調整することができる．

読者による不快度の評価の平均値と標準偏差をコメントごとに計算し，その値の関係を表したものを図 3 に示す．図 3 より，コメントに対する評価の個人差は，そのコメントの平均的な不快度と関連があることがわかる．このため，重み α を算出する関数 Personalization Level Estimator としてコメントの不快度の平均値や標準偏差を利用することで，適切に個人化レベルの調整を行うことができると考えられる．ここで，コメント c の不快度の平均値と標準偏差を事前に予測することでコメント c に対する重み α を求める式は以下の式 (13) で表される．

$$\alpha = \mathrm{Sigmoid}(\mathbf{W}(\hat{\mu}(c) \oplus \hat{\sigma}(c)) + \mathbf{b}) \tag{13}$$

$\hat{\mu}(c), \hat{\sigma}(c)$ はそれぞれコメント c の不快度の平均値と標準偏差の予測値を表すスカラー値である．$\mathbf{W}$ は重み行列，$\mathbf{b}$ はバイアス項であり，$\hat{\mu}(c) \oplus \hat{\sigma}(c)$ を 1 次元に圧縮する全結合層のパラメータである．評価実験では，個人化レベルの調整の有効性と，個人化レベルの調整を行うパラメータ α の算出における最適な手法を調査する．

4 評 価 実 験

4.1 使用データ・実験設定

4.1.1 不快コメント予測モデル

実験に使用するデータセットを作成するために TwitterAPI を用い，2022 年 4 月 1 日から 2022 年 6 月 30 日にかけて NHK ニュース (@nhk_news)[5]が投稿したニュースのポスト，ニュースに対するユーザの返信をそれぞれニュース・コメントとして取得した．

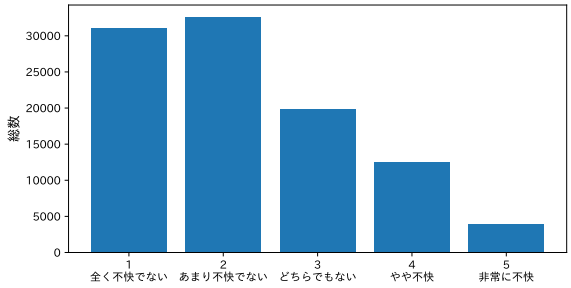


図 4: アンケートで取得した不快度に関する各ラベルの分布

ニュースコメントに対する評価値を取得するため，クラウドソーシングサービスの一つであるクラウドワークスでアンケートを実施した．被験者の総数は 250 名であり，250 名の被験者を 50 名ごとの 5 つのグループに分けた．グループごとに 2022 年 4 月 1 日から 6 月 30 日の 3ヶ月間に投稿されたニュースコメント 400 件を提示し，各ニュースコメントが不快と感じるかについて {1: 全くそう思わない，2: あまりそう思わない，3: どちらとも言えない，4: ややそう思う，5: 非常にそう思う } の 5 段階で回答してもらった．なお，被験者に提示するニュースとコメントは各グループ内で共通である．データセットの概要および各ラベルの分布をそれぞれ表 3, 図 4 に示す．

表 3: 不快コメント予測で利用するデータセットの概要

項目	値
ユーザ数	250
コメント数	2,000
評価数	60,000

2022 年 4 月 1 日から 4 月 14 日の 2 週間において 5 段階の評価値が 4 以上である（不快と評価された）ニュースとコメントのペアと評価を行ったユーザの読者 ID を格納したデータベースをフィードバックデータベースとして構築した．フィードバックデータベースには，評価値が 4 未満である（不快と評価されなかった）ニュースとコメントは格納しない．提案手法では，読者 ID を入力としてフィードバックデータベースから読者が不快と評価したニュースとコメントのペアを取得する．

不快コメント予測を行うモデルの訓練について説明する．教師データの入力データは予測対象のニュースとコメントのペアと読者 ID であり，正解ラベルは読者による 5 段階の評価値である．2022 年 4 月 15 日から 6 月 9 日のニュース・コメントを訓練データ，2022 年 6 月 10 日から 6 月 20 日のニュース・コメントを検証データ，6 月 21 日から 6 月 30 日のニュース・コメントをテストデータとして時系順に分割した．訓練・検証・

5：https://twitter.com/nhk_news

テストのデータ数は 1 人の読者につきそれぞれ 230 件・50 件・50 件である．

フィードバックデータベースから 10 件以上のフィードバックを取得できた 150 人のコメント読者を対象に不快コメント予測モデルの訓練を行った．BERT 層に関しては，東北大学乾研究室による日本語 Wikipedia を用いた事前学習済み BERT モデル [6]を用い，fine-tuning を行った．フィードバックを大量に行う行為は読者にとって負担となるため，個人化に用いる読者によるフィードバックのデータ数は少量であることが望ましい．このため，読者 ID の入力時に取得する，過去に低評価したニュースとコメントを 10 件という少量のデータに限定して個人化を行った．

4.1.2 使用モデル

不快コメント予測の個人化におけるコメントに応じた個人化レベルの調整の有効性と，個人化レベルの調整を行うパラメータ α の算出の最適な手法を調査するために，本実験では複数のモデルによる予測結果の比較を行った．各モデルの詳細は以下の通りである．

Basic
コメントのベクトルと読者のベクトルの連結を行い，全結合層の出力値をコメントの不快度とする．

Weighted Concat (Mean & SD)
コメントのベクトルと読者のベクトルの重み付き連結を行う．重み α の算出に関して，対象のコメントの不快度の平均値と標準偏差を訓練済みモデルを用いて予測し，全結合層を通して 1 次元に変換された値をもとに重み α を算出する（図 5(a)）．

Weighted Concat (Mean)
コメントのベクトルと読者のベクトルの重み付き連結を行う．重み α の算出に関して，対象のコメントの不快度の平均値を訓練済みモデルを用いて予測し，予測値をもとに重み α を算出する（図 5(b)）．

Weighted Concat (SD)
コメントのベクトルと読者のベクトルの重み付き連結を行う．重み α の算出に関して，対象のコメントの不快度の標準偏差を訓練済みモデルを用いて予測し，予測値をもとに重み α を算出する（図 5(c)）．

Weighted Concat (BERT emb)
コメントのベクトルと読者のベクトルの重み付き連結を行う．重み α の算出に関して，対象のコメントに対する BERT 層の出力ベクトル（768 次元）を全結合層を通して 1 次元に変換し，重み α を算出する（図 5(d)）．

Feature
対象のコメントの不快度の平均値と標準偏差を訓練済みモデルを用いて予測し，予測値を特徴量として扱う．不快度の平均値・不快度の標準偏差・コメントのベクトル・読者のベクトルの連結を行い，全結合層の出力値をコメントの不快度とする．

6：https://github.com/cl-tohoku/bert-japanese

4.2 実験結果

4.2.1 不快コメント分類の結果

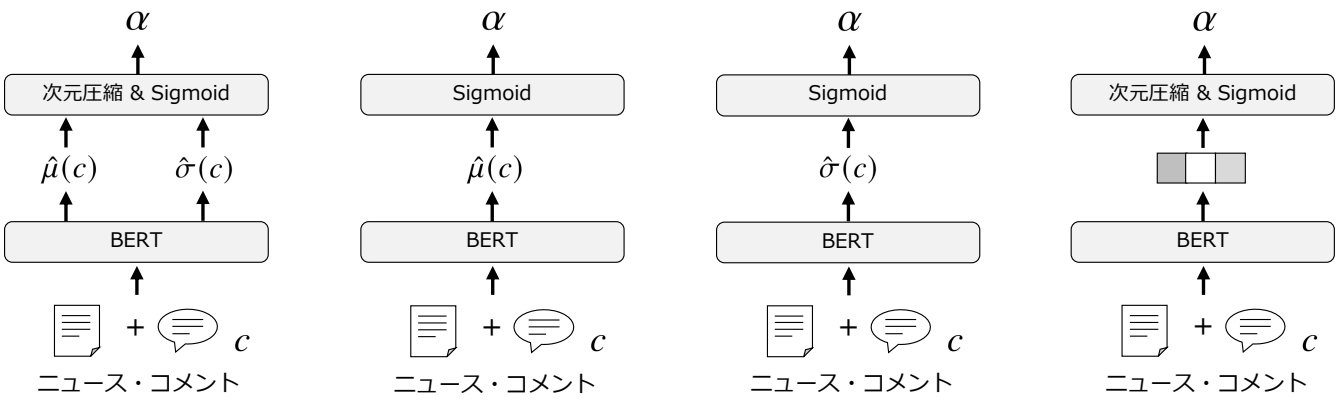
読者ごとに予測評価値が上位 k 件のニュースコメントを「不快コメント」と判定したときの適合率 (=Precision@k) を求めることで，2 値分類における提案手法の有用性を検証した．正解ラベルに関して，アンケートで 4 または 5 の評価値が付けられたコメントを「不快コメント」とし，それ以外の評価値が付けられたコメントは「不快コメントでない」とした．SNS などのインターネット上のコンテンツでは，推薦やフィルタリングによって自分の興味があるコンテンツばかりが提供されてしまうフィルターバブルと呼ばれる現象があり，特定の信念が増幅または強化されることが問題となっている [25]．このため，本実験では予測評価値が上位のニュースコメントに対する正解率から予測性能の評価を行うことにより，読者にとって特に有害なコメントのみをフィルタリングすることを目的とした．

はじめに，個人化レベルの調整を行う変数である α の算出方法による分類性能の差異を調べるため，Weighted Concat (Mean & SD) モデル，Weighted Concat（Mean）モデル，Weighted Concat（SD），Weighted Concat（BERT emb）モデルの比較を行った．k=1 から k=10 と変化させたときの Precision@k の値を図 6 に示す．図 6 より，最も高い性能を示すのは Weighted Concat (Mean & SD) モデルであり，個人化レベルの調整において対象のコメントの不快度の平均値と標準偏差の予測値を活用することが有効であることがわかった．また，ほとんどの k の範囲において Weighted Concat（Mean）モデルの方が Weighted Concat（SD）モデルよりも高い性能が得られ，個人化レベルの調整において不快度の平均値の予測がより重要であることがわかった．

次に，コメントに応じて個人化レベルの調整を行うことの有効性を検証するため，Basic モデル，Weighted Concat (Mean & SD) モデル，Feature モデルによる予測結果の比較を行った．k=1 から k=10 と変化させたときの Precision@k の値を図 7 に示す．図 7 より，全ての k の値で Weighted Concat (Mean & SD) モデルから高い分類結果が得られた．また，Feature モデルは対象のコメントの不快度の平均値と標準偏差を予測し，予測値を特徴量として扱うモデルであるが，Basic モデルの性能を上回らなかった．これより，単に対象のコメントの特徴を予測して分類に活用するだけでは不十分であり，コメントの特徴に応じて個人化レベルを調整することが有効であることがわかった．

4.2.2 考　　察

不快度の予測において個人化レベルの調整が与える影響について調査した．はじめに，テストデータについて予測した不快度をもとに，読者ごとにコメントを不快度の昇順で順位付けを行った．その後，与えられた順位の平均値と，順位の標準偏差を各コメントに対して算出した．Basic モデルと提案手法の一つである Weighted Concat (Mean & SD) モデルによって予測した不快度を利用した場合における，コメントの不快度の順位の平均値と標準偏差の関係を表した図が図 8(a), 8(b) である．



(a) Mean & SD　(b) Mean　(c) SD　(d) BERT emb

図 5: Weighted Concat モデルにおける重み α の算出手法の比較

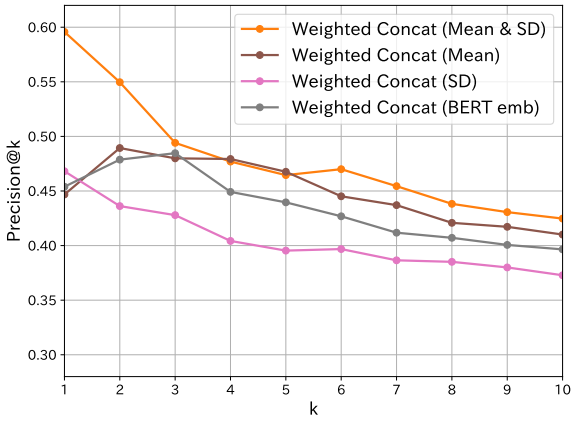


図 6: 異なる重み α の算出手法による Precision@k の比較

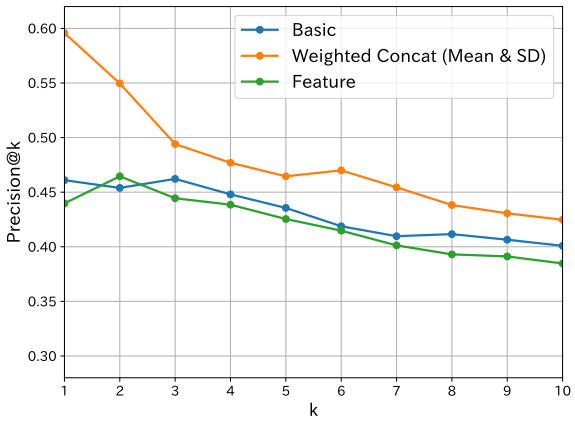


図 7: 個人化レベルの調整の有無による Precision@k の比較

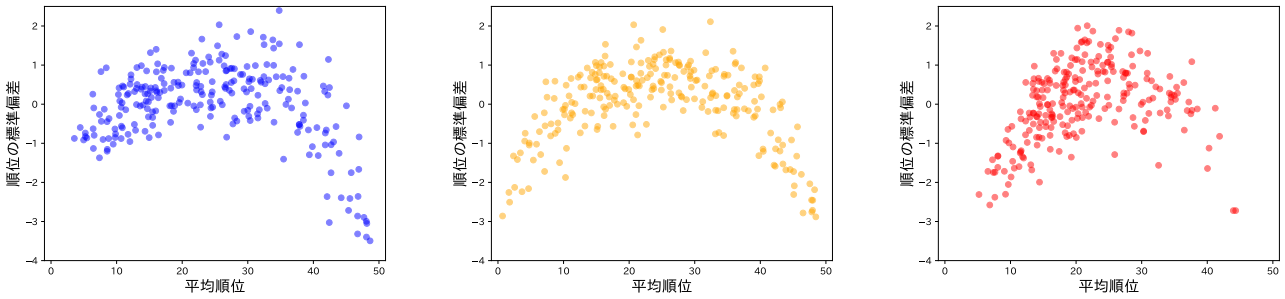
同様に，正解データを用いて不快度の順位の平均値と標準偏差の関係を表したものが図 8(c) である．テストデータには 1 人の読者につき 50 件のコメントが含まれるため，横軸の順位の平均値は 0 以上 50 未満となる. なお，各手法における標準偏差のスケールが異なったため，データの比較を行うために標準化を行った．

図 8(a) の Basic モデルの散布図では，左側の不快度が非常に低いコメントに対しても標準偏差の値がやや高くなっていることから，全く不快でないコメントに対して個人化された予測を行っていることがわかる．一方，図 8(b) の Weighted Concat (Mean & SD) の散布図では，左側の不快度が非常に低いコメントに対して標準偏差の値が低いことから，全く不快でないコメントに対して予測の個人化をほとんど行っていないことがわかる．図 8(c) の正解データの散布図より，不快度が非常に低いコメントでは標準偏差の値が低いことから，全く不快でないコメントに対しては，予測の個人化の必要性が低いといえる．提案手法の一つである Weighted Concat (Mean & SD) モデルでは不快度が非常に低いコメントに対して予測の個人化をほとんど行っておらず，正解データの散布図と似た傾向が見られたことから予測性能が向上したと考えられる．

4.3 他ドメインへの応用

提案手法による個人化レベルの調整が，不快コメント予測以外のドメインのタスクで有効であるかについて検証した．本実験では推薦システムを対象とし，推薦システムのデータセットとして広く利用されている MovieLens[7]のデータセットの 1 つである MovieLens 100K Dataset を用いた．MovieLens は映画に対するレビュー評価のデータセットであり，ユーザが映画に対して 1 から 5 までの 5 段階の評価を行なったデータが格納されている．MovieLens-100K の概要を表 4 に示す．実装には推薦システム用のライブラリである RecBole [26] を用い，ユーザと映画のベクトルの次元数はそれぞれ 64 とした．4.1.2 節で述べた予測モデルのうち，Basic モデル（図 9）と提案手法の 1 つである Weighted Concat (Mean & SD) モデル（図 10）による推薦結果を比較した．各モデルによる評価指標の値を表 5 に示す．表 5 より，全ての指標において Weighted Concat (Mean & SD) モデルによる結果が上回っており，重み付き連結による個人化レベルの調整が推薦システムにおいても有効であることがわかった．

7：https://grouplens.org/datasets/movielens/



(a) Basic (b) Weighted Concat (Mean & SD) (c) 正解データ

図 8: 各手法によって計算されるコメントに対する不快度の平均順位と標準偏差の関係．読者ごとにコメントを不快度の昇順で順位付けを行い，与えられた順位の平均値と順位の標準偏差を各コメントに対して算出し，その関係を表したものである．

表 4: MovieLens-100K の概要

項目	値
ユーザ数	944
映画数	1,683
評価数	100,000

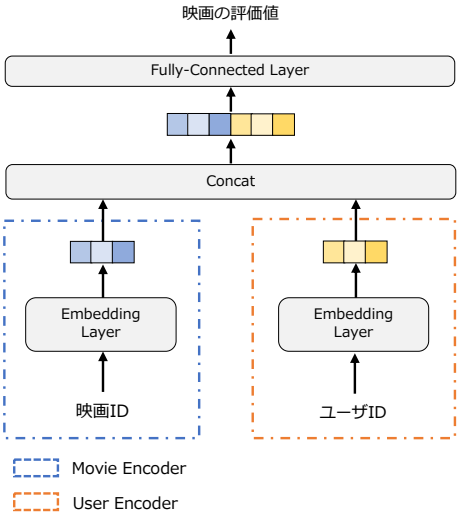


図 9: 映画推薦における Basic モデルの概要

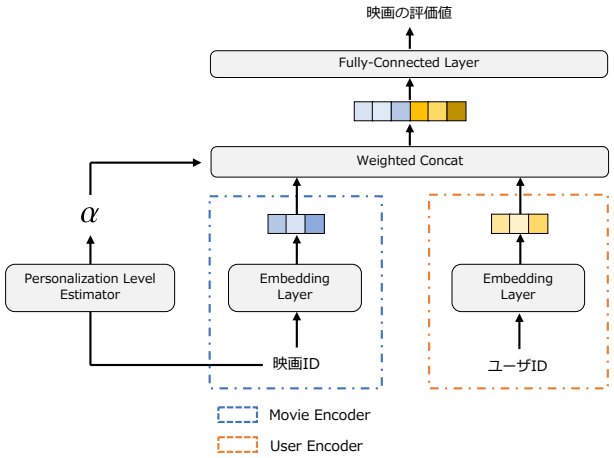


図 10: 映画推薦における Weighted Concat (Mean & SD) モデルの概要

5 ま と め

本論文では，ニュースコメントの読者が過去に低評価したニュースコメントのフィードバックをもとに，不快コメント予測の個人化手法を提案した．提案手法ではコメントの内容によって個人化の重要度が異なることに注目した．具体的には，予測対象のコメントに応じて，コメントと読者の情報を重み付けすることで個人化レベルを調整する手法を提案した．評価実験の結果，コメントごとに個人化レベルを調整する処理を予測モデルに組み込むことによって，不快コメント予測の個人化の性能が向上することが示された．さらに個人化レベルを調整する処理は推薦システムにおいても有効であることがわかった．

今後の課題として，最新の出来事やトレンドを反映した不快コメント予測モデルの開発があげられる．ニュースとそれに対するコメントには常に最新の情報が含まれるため，最新の出来事やトレンドを理解した予測を行うことが重要である．しかし，本論文で提案した手法では最新の情報を取り入れることができず，コメントや読者の特徴を正確にモデル化することが難しい．このため，ニュースやコメントに含まれるトピックについて予測時に外部データを参照し，最新の出来事やトレンドを踏まえてコメントや読者をモデル化する手法の開発を目指す．

文 献

[1] 新聞通信調査会. 第 14 回メディアに関する全国世論調査, 2021.
[2] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. News commenters and news comment readers. *Engaging News Project*, pp. 1–21, 2016.
[3] Julian Risch and Ralf Krestel. *Toxic Comment Detection in Online Discussions*, pp. 85–109. Springer Singapore, Singapore, 2020.
[4] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, No. 35, pp. 1–6, 2018.
[5] Google Jigsaw. Perspective API, 2017.
[6] Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Proceedings of the Conference of the North Ameri-*

表 5: MovieLens-100K を利用した推薦の性能比較

モデル	Recall@10	MRR@10	nDCG@10	Hit@10	Precision@10
Basic	0.0629	0.1854	0.0890	0.3977	0.0702
Weighted Concat (Mean & SD)	**0.0662**	**0.2010**	**0.0973**	**0.4104**	**0.0758**

can Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1648–1659, 2019.

[7] Martin Saveski, Brandon Roy, and Deb Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference*, pp. 1086–1097, 2021.

[8] Ahmad Alsharef, Karan Aggarwal, Deepika Koundal, Hashem Alyami, Darine Ameyed, et al. An automated toxicity classification on social media using lstm and word embedding. *Computational Intelligence and Neuroscience*, Vol. 2022, , 2022.

[9] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in twitter. *Sensors*, Vol. 19, No. 21, 2019.

[10] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2023.

[11] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 71–80, 2012.

[12] Yuto Oikawa, Yuki Nakayama, and Koji Murakami. A stacking-based efficient method for toxic language detection on live streaming chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 571–578, 2022.

[13] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 5884–5906, 2022.

[14] Teruki Nakahara and Taketoshi Ushiama. Personalized prediction of offensive news comments by considering the characteristics of commenters. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pp. 1615–1622, 2023.

[15] Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 265–274, 2012.

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[17] Ahmet Aker, Emina Kurtic, A. R. Balamurali, Monica Paramita, Emma Barker, Mark Hepple, and Rob Gaizauskas. A graph-based approach to topic clustering for online comments to news. In *Proceedings of the 38th European Conference on IR Research*, pp. 15–29, 2016.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1977.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[20] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 677–694, 2019.

[21] Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. A case for a range of acceptable annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*, Vol. 2276, pp. 19–31, 2018.

[22] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*, pp. 299–318, 2021.

[23] 中原輝樹, 牛尼剛聡. ニュースコメントの閲覧支援のためのニュースへの反応に基づくユーザ埋め込み表現の生成. 情報処理学会論文誌データベース (TOD), Vol. 15, No. 3, pp. 99–110, 2022.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[25] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.

[26] Lanling Xu, Zhen Tian, Gaowei Zhang, Junjie Zhang, Lei Wang, Bowen Zheng, Yifan Li, Jiakai Tang, Zeyu Zhang, Yupeng Hou, Xingyu Pan, Wayne Xin Zhao, Xu Chen, and Ji-Rong Wen. Towards a more user-friendly and easy-to-use benchmark library for recommender systems. pp. 2837–2847, 2023.

食品レビュー検索でのクエリ拡張のための LLMを用いた同義フレーズ生成

芦澤亜里紗† 三林 亮太† 大島 裕明†

† 兵庫県立大学 大学院情報科学研究科 〒 651–2197 神戸市西区学園西町 8–2–1
E-mail: †{ad23y002,af22h007}@guh.u-hyogo.ac.jp, ††ohshima@ai.u-hyogo.ac.jp

あらまし 本研究では，オンラインショッピングで食品の購入を検討する際のレビュー検索において，表現の多様性を考慮した検索手法を提案する．例えば，特定のうどんの商品レビューに対して,「コシ」というクエリで検索を行ったとする．その際に,「コシ」だけでなく「弾力がある」や「噛み応え」のような表現が入ったレビューも検索結果に含まれるようになれば，ユーザは多様なレビューを購入の参考にすることができる．しかし，文字列一致だけでは「弾力がある」や「噛み応え」のような表現が入ったレビューを得ることは難しい．そこで本研究では，文字列一致だけでは得ることのできないレビューの検索を実現するために 2 つのアプローチを行う．1 つ目は，大規模言語モデル（LLM）によるクエリ拡張を利用した検索モデルを提案する．このモデルを学習するために，あり得る様々な検索キーワードに対して，LLM を用いて同義フレーズを取得し，モデル構築のための学習データを作成する．そして，作成した学習データを用いてモデルを学習することで，レビューにおける表現の同義性を考慮した検索モデルを構築する．2 つ目は，検索クエリが与えられた時に LLM によってオンデマンドに同義フレーズを取得し，それらの同義フレーズを用いて検索を行うことを提案する．提案手法の評価は，既存手法との比較をすることで行った．

キーワード レビュー検索，食品レビュー，BERT，LLM，クエリ拡張

1 はじめに

オンラインショッピングにおいて，商品レビューは購入の判断材料として重要な要素である．しかし，膨大なレビューがある場合，欲しい情報を含むレビューを探し出すには手間がかかる．実際に，本研究で使用する楽天市場の「食品」,「スイーツ・お菓子」の商品の中には，レビューが 100 件以上あるものが 8,000 個以上も存在する．このように商品に大量のレビューがある場合，レビュー検索を行うことで欲しいレビューを見つけることができる．

しかし，従来の文字列一致の検索手法では，ユーザが入力した検索キーワードに完全に一致するレビューしか検索されない．そのため検索キーワードに完全に一致するレビューがない場合，ユーザの任意の文字列に左右され，ユーザが入力した検索キーワードにマッチしたレビューを検索することができない．

例えば，文字列一致の検索手法の場合，表 1 のように「コシ」という検索キーワードが入力されると，完全に「コシ」という単語を含むレビューのみが検索される．そのため,「コシ」という検索キーワードでは,「今まで食べた中で一番**弾力**があって満足」といったような意味は似ているが異なる表現で書かれたレビューを得ることはできない．同様に，表 2 のように,「辛い」という検索キーワードが入力された場合,「辛い」という単語を含むレビューのみが検索される．しかし,「辛い」の程度を表す「マイルド」といった対照的な表現を含むレビューも必要な情報となる．このように，文字列一致の検索手法を用いた際の検索結果には含まれず，意味や関連性を考慮した検索が行われない．

そのため，検索キーワードに一致する表現が対象となるレビュー集合に含まれない場合，検索結果を得ることができない．そこで本研究では，検索対象となるレビュー集合に 1 つも検索キーワードと同じ表現を含んだレビューがない場合でも，検索キーワードにマッチするレビューを検索できるようにする．このように，ユーザの様々な表現や検索意図を考慮し，必要な情報を含んだレビューを検索できるようにするために，ユーザの表現の多様性を考慮したレビュー検索手法を提案する．

この検索手法には，図 1 と図 2 のような 2 つのアプローチが考えられる．1 つ目のアプローチでは，検索キーワードにレビューが適合しているかを判定する機械学習モデルの構築を行う．この機械学習モデルを構築するための学習データ作成に LLM を活用する．このモデルの構築には，検索キーワード「コシ」に関連する表現として「弾力」や「ふにゃふにゃ」などの言いかえ表現のデータを事前に収集できれば，それを含んだレビューを適合データとし，学習することができる．そこで，LLM を使用することで，事前に言いかえ表現を大量に取得し，適合データを作成する．この適合データで学習を行った適合判定 BERT を使用して，検索キーワードにマッチしたレビューを取得する．

2 つ目のアプローチでは，LLM によって検索キーワードの言いかえ表現をオンデマンドに取得し，クエリ拡張を行う．これによりユーザに入力された検索キーワードにマッチする多様な表現のレビューを上位に検索することを可能にする．このように，本研究では，LLM を用いて「言いかえ表現生成」と「クエリ拡張」2 つのアプローチを行う．

表 1　異なる表現で書かれたレビューの入出力例

目的の商品	うどん
欲しい情報	うどんにコシがあるか
入力	**コシ**
文字列一致の出力	**コシ**が強くておいしかったです
異なる表現を含むレビューの出力	今まで食べた中で一番**弾力**があって満足

表 2　対照的な表現で書かれたレビューの入出力例

目的の商品	麻婆豆腐
欲しい情報	辛さがあるのか
入力	**辛い**
文字列一致の出力	私には**辛い**です
対照的な表現を含むレビューの出力	思ったよりも**マイルド**で食べやすい

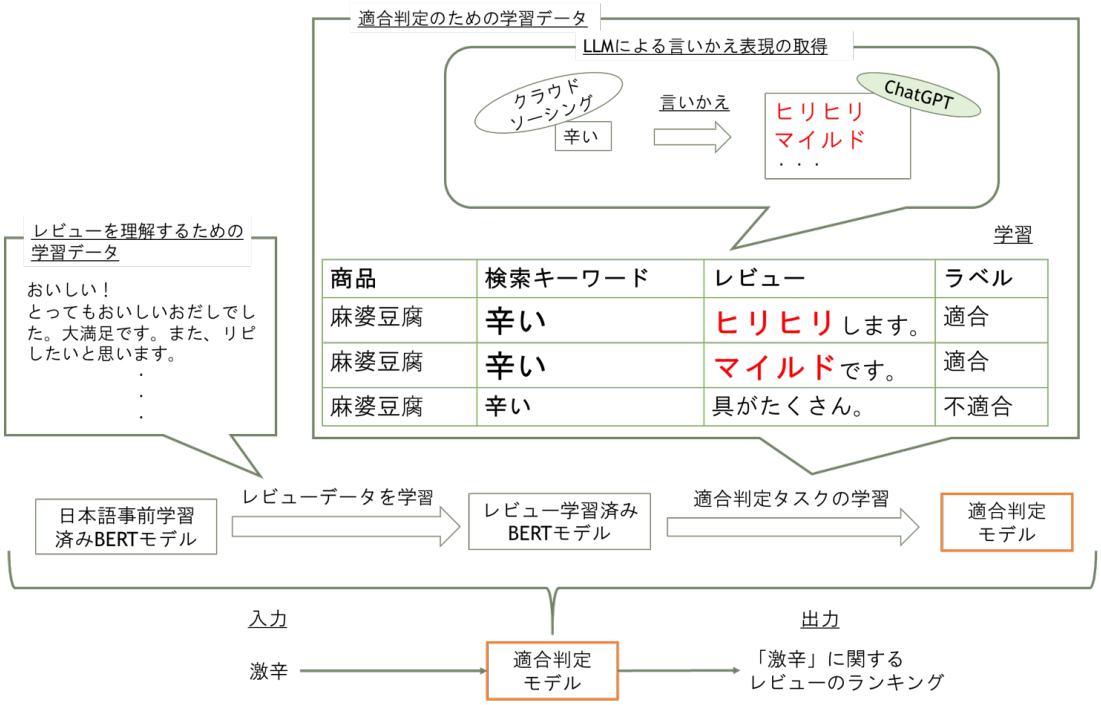


図 1　1 つ目のアプローチである適合判定 BERT の構築

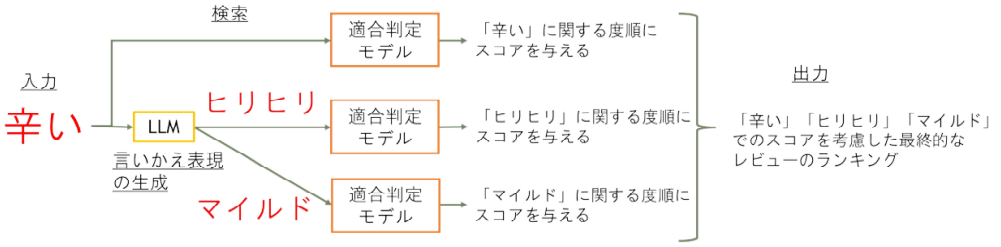


図 2　2 つ目のアプローチであるクエリ拡張

2　関 連 研 究

2.1　商品情報に関する研究

オンラインショッピングでユーザが商品に関する情報を得ることができるように，様々な研究が行われている．例えば，Yitzhak ら [2] によってファセット検索に関する研究が行われていたり,Liu ら [9] によって，製品の特徴を視覚的にあらわすことで，一目見ただけで消費者が思う製品の長所と短所を知ることができる研究が行われている．

他にも，Fang ら [5] によって，レビューの要約などに利用できる研究が行われている．この研究では，文の情報量ランキングと構造的センチメント分析を行い，定式化することで，センチメント説明のランキング問題を解決している．

また，オンラインショッピングに関する研究では，レビューから情報を抽出することで，システムの構築を行う研究は多い．例えば，Hu ら [7] によってレビューを要約する際に，特徴抽出とレビューが肯定的か否定的かの識別情報を使用する研究が行われている．このように，レビュー情報を使用した研究が行われている．本研究ではレビュー検索を行い，ユーザが欲しいレビューを抽出する．

2.2 ユーザの検索意図を考慮した検索

検索を行うユーザにとって，検索キーワードは求めている情報を表している．しかし，入力される検索キーワードには多様な表現があり，欲しい情報を検索するのに最も適切な検索キーワードでない可能性がある．そのため，ユーザがどのような情報を求めているのか検索意図をくみ取って検索される必要がある．先行研究では，ユーザの検索意図を考慮した検索を行えるようにするための手法が提案されている [17] [19].

例えば，大石ら [20] によってユーザーが意図する単語を生成するための手法として，文章内の単語の関連性を見つける新しいアルゴリズムが提案されている．このアルゴリズムは，重要な単語が文中で近くに出現する傾向があるという考えに基づいていた．さらに，この手法を適合性フィードバックの一つである RSV（Relevance-based Similarity Vector）と組み合わせることで，検索結果の精度を向上させていた．

本研究では，検索キーワードにマッチしたレビューを検索できるようにするためのモデルの構築を行い，検索キーワードの意図を汲み取った検索を可能にする．

2.3 BERT

本研究では，機械学習モデルの一つで，汎用言語モデルである BERT を用いる．BERT は 2018 年に Devlin ら [8] によって，提案されたモデルである．BERT は解きたいタスクで学習を行うとそのタスクを解くことができるようになる．そのため，様々なタスクに活用されている [13] [3] [1] [10].

例えば，Chouikhi ら [4] によって Twitter から収集したデータに対して，BERT モデルを用いた感情分析が行われている．その結果，新型コロナウイルスの社会生活への影響についての感情分析を行っている．このように，BERT は様々な分野のタスクに使用することができるため，本研究では検索キーワードに対するレビューの適合判定タスクに使用する．

2.4 追 加 学 習

BERT は追加学習を行うことで，学習したデータに関連するタスクの性能が向上することが知られている．

Sushil ら [11] によって，タスク固有のデータに対して，タスク適応型事前学習（TAPT）というタスクが行われ，追加学習によってタスクのパフォーマンスが向上したモデルの構築が行われている．タスク固有のデータを用いて大規模言語モデルをさらに追加学習を行うことで，タスクのパフォーマンスが向上したモデルを構築している．

他にも，鈴木ら [22] によって金融ドメインの文書で追加学習を行い，Wikipedia のみで構築したモデルよりも精度が高くなっていることを示していたり，Gururangan ら [6] によって，解きたいタスクに関連するデータで追加学習を行うことでそのタスクの性能が向上することを示していた．

これらの先行研究から追加学習の有用性が示されている．そのため，本研究では,「食品」,「スイーツ・お菓子」のレビューデータで追加学習を行った．

2.5 学習データの作成

BERT を使用してタスクを学習する際には、学習データが必要である．しかし，大量の学習データを手作業で作成することは難しいため，学習データ作成には工夫を行う．

例えば，Wang ら [14] によって，人々が面倒だと感じる行動を自動的に分類するための新しいデータ拡張アプローチが行われている．レキシカル埋め込みのようなアプローチを行うことで，疑似的に新しいトレーニングインスタンスを作成している．

また，自然言語だけでなく画像の分野では，福岡ら [21] によって，生成モデルの一種である GAN を用いて，疑似的にひび割れコンクリート画像生成を行っている．このように疑似的に作成したデータを用いて，保守点検作業の労力を減らすためのモデルの構築を行っている．本研究でも，LLM を用いることで大量の学習データを作成し，モデルの構築を行う．

2.6 LLM の活用

ChatGPT は，OpenAI によって開発された自然言語処理モデルの一つであり，テキスト生成や質問応答などを行うことができる．例えば，Tinega ら [12] によって，プログラミングのデバッグにどのように ChatGPT を利用するのかについて議論されている．この研究では，ChatGPT の強みを他のデバッグツールと組み合わせることの利点について述べられている．本研究では，ChatGPT を用いてクエリ拡張を行う．また，先行研究では Wikipedia のデータを利用しているが [18] [16] [15], 本研究では,「食品」「スイーツ・お菓子」のレビューデータの情報を利用し，このデータに特化したクエリ拡張を行う．

3 食品レビューデータとその前処理

3.1 食品レビューデータ

本研究で使用するレビューデータについて述べる．本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット [1]」の商品レビューデータを使用している．使用したデータは，2015 年の 1 月から 2019 年の 12 月までの 5 年間のデータである．全ジャンルのレビュー数は約 7,000 万件と膨大である．また，ジャンルごとにレビューに書かれている内容の特徴も異なる．例えば，食品のレビューには美味しさに関する意見が書かれているレビューが多く，電化製品のレビューでは，傷や使いやすさに関するレビューが多い．そのため，すべての商品ジャンルを対象に検索システムを構築するためにはそれぞれ

1：楽天グループ株式会社 (2020): 楽天市場データ. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.2.1

のレビューの特徴を考慮する必要がある．そこで今回は,「食品」,「スイーツ・お菓子」に絞ることで表現の特徴をつかみ，欲しいレビューを見つけるための検索を行えるようにする．本研究で使用するデータは,「食品」,「スイーツ・お菓子」に絞った約 650 万件のデータである．

3.2 使用するデータと言葉の定義

本研究では，レビュー内容に付属する情報のうち，店舗 ID，商品 ID，商品ジャンル ID，投稿者 ID，レビュータイトルを使用した．また，店舗 ID と商品 ID をもとに商品を特定するための商品 No の作成や，商品ジャンル ID を使用して子ジャンルから親ジャンルのパスをたどった商品ジャンルパスの作成を行った．子ジャンルである商品ジャンル ID から親ジャンルまでのパスをたどるために，楽天から提供されているジャンルマスタ (master_genre_.tsv) を参照した．表 3 に，使用したデータの詳細を示す．

本研究では，以降「レビュータイトル」と「レビュー内容」を「。」で結合したものをレビューと呼ぶ．また，レビューを「。」や「...。」,「(。¿ω¡。)」などの顔文字などで区切ったものをレビュー文と呼ぶこととする．

3.3 学習と評価に使用するデータ

あらかじめ，学習を行うための学習データ作成に使用するためのデータと検索システム構築後にモデルを評価するための評価データを分割した．まず，評価データの抽出方法について説明する．本研究では，100 件以上のレビューがある時に，人手で欲しいレビューを見つけ出すことは大変であると仮定をおき，レビュー数が 100 件以上ある商品 No を 8,399 個収集した．それらの商品に付属する商品ジャンルパスから「食品」もしくは「スイーツ・お菓子」の直下のジャンルのみを抽出し、商品数が 20 件以上あるジャンルとその商品 No を収集した．収集したジャンル数は，1,301 個あり，この中で 20 件以上の商品があるジャンルは，28 個あった．この 28 個のジャンルそれぞれからランダムに 5 商品ずつ取得した．

その結果 28 個のジャンル× 5 商品の合計 140 商品を評価対象の商品として取得した．また，評価に使用する商品それぞれから最新のレビューを 100 件ずつ取得した．この，140 商品× 100 レビューの合計 14,000 件を評価対象のデータとした．

評価に使用していない 627,400 商品すべてを学習データの対象商品とした．その結果，6,435,083 件のレビューを学習データの対象レビューとして取得した．また，レビュー文は，25,051,124 件取得した．

4 提 案 手 法

本研究では 1 つ目のアプローチとして，レビューにおけるタスクの性能を向上させたレビュー学習済み BERT モデルの構築を行い，LLM(gpt-35-turbo) を使用して作成した適合判定のための学習データでファインチューニングを行った．これにより，適合判定 BERT を構築した．2 つ目のアプローチでは，LLM によってオンデマンドクエリ拡張を行った．

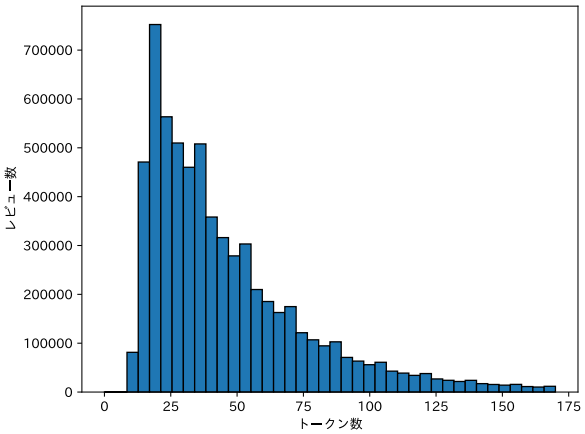


図 3 トークン数ごとのレビュー数（bins = 40）

4.1 レビュー学習済み BERT モデルの構築

本研究では，日本語事前学習済み BERT モデルにレビューを学習させることでレビューに特化したレビュー学習済み BERT モデルを構築した．日本語事前学習済み BERT モデルは，東北大学の乾研究室により提供された BERT モデル（cl-tohoku/bert-base-japanese-v3）を使用した．このモデルは，CC-100 の日本語データと日本語 Wikipedia で学習することで日本語に特化したモデルを構築している．本研究で行った，モデルがレビューに特化するための学習には，楽天レビューデータの「食品」,「スイーツ・お菓子」のレビューデータを使用した．このデータには， 3.3 節で作成した学習データ作成に使用するためのレビューデータすべてを使用した．

「食品」,「スイーツ・お菓子」に絞った楽天レビューデータで東北大学の日本語事前学習済み BERT モデルに学習を行った．この学習を本研究では，追加学習と呼ぶ．学習は，レビューを入力として図 4 のように一部の語を MASK したり違う語に置き換えたりし，それらの元の語を予測することで行った．

学習の前に，トークナイザに未知語の追加を行うかを検討した．検討を行うために，未知語を学習の対象レビューから 113 個取得した．しかし，取得できたものは「滷味」「哩」「桝志」などの語彙であった．これらの語彙を食品レビューに特化した語彙として，追加を行うことは適切でないと考えた．そのため，未知語の追加は行わず，モデルの語彙サイズは 32,768 のままである．

追加学習を行う際にはトークン長を指定する必要がある．図 3 は，トークン数ごとのレビュー数のヒストグラムである．この図からわかるように，ほとんどのトークン数が 128 以下である．また,「食品」,「スイーツ・お菓子」全レビューのうち，トークン数が 128 以下であるレビューの割合は，96%であった．そのため，学習を行う際のトークン長は，128 とした．

4.2 アプローチ 1: 適合判定 BERT の構築

図 5 に示すように，検索キーワードにレビューが適合しているかを判定してくれるモデルさえあれば，検索キーワードに適合したレビューを検索することができる．そこで，このモデル

表 3　楽天レビューデータとそれに基づき作成したデータの詳細

データ	データ詳細	例
投稿者 ID	マスクしたユーザ名	user_1
店舗 ID	店舗ごとに付与された ID	255162
商品 ID	商品ごとに付与された ID	10000001
商品ジャンル ID	商品のジャンルに付与された ID	412237
商品 No	自身で作成した,「商品 ID」と「店舗 ID」が一致する商品に対して 1–627664 の数字をつけたもの	9160
レビュータイトル	レビューのタイトル	普通の
レビュー内容	レビュー	普通においしいです。たまには変わったパッケージもおもしろい
商品ジャンルパス	商品ジャンル ID からたどったジャンル名のパス	食品/魚介類・水産加工品/エビ

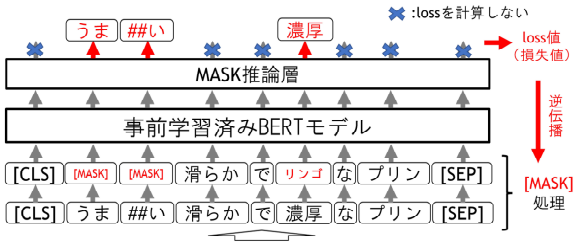


図 4　追 加 学 習

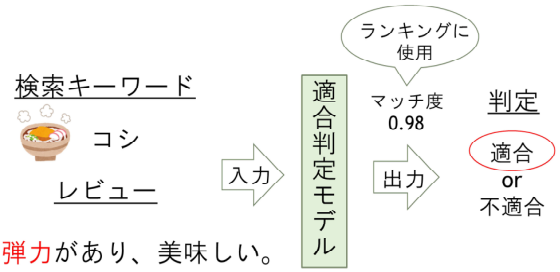


図 5　適合判定 BERT

を構築するために，LLM を用いて学習データを作成した．モデル構築の手順は，まず，LLM により検索キーワードの言い換え表現を取得する．次に，この言い換え表現を使用して適合データと不適合データを作成し，この学習データでファインチューニングを行い適合判定 BERT を構築した．

4.2.1　クラウドソーシングによる検索キーワードの収集

レビュー検索に用いる検索キーワードをクラウドソーシングを利用してあらかじめ収集した．収集には，クラウドソーシングのサービスを提供している Lancers [2]を利用した．200 名のランサーに行ってもらったタスクの概要は以下のとおりである．

（1）　まず，これまでに実際に購入したことがある食品について，具体的な商品を思い浮かべる．

（2）　その商品についてのレビューがネット通販サイトに大量に（100 件以上）あると考える．

（3）　それらの大量のレビューの中から，その商品の特徴などについて調べるためにレビューを検索することができるとしたら，どのような検索キーワードを用いるかを考える．

（4）　このようにして，思い浮かべた商品の名称と，レビュー検索に用いたいと思い浮かべた検索キーワードを回答する．

(1)〜(4) のタスクを，思い浮かべる商品を替えて，1 人のランサーごとに 10 商品分行ってもらった．収集できた検索キーワードの中には，意図しているデータでないものがあった．そこで，適切な検索キーワードのみを収集するための前処理を行った．

意図しているデータは以下のような対象商品に付属するレビューから購入に必要な情報を含んだレビューを検索するための検索キーワードとして適切なものである．

商品名	検索キーワード
ポッカサッポロ サンスウィート プルーン	酸味

意図していないデータは 2 パターンある．1 つ目は以下のように，検索キーワードがレビューではなく商品を検索するためのものになっている場合である．

商品名	検索キーワード
ポテトチップチョコレート	ポテトチップチョコレート ROYCE

2 つ目は，以下のように，商品名が本研究では対象としていない飲み物となっている場合である．

商品名	検索キーワード
響 21 年ウイスキー 700ml	アルコール

このように意図していないデータを除き，意図している検索キーワードのみを抽出した．意図している検索キーワードには，「コシ 太さ」や「甘さ、チーズの濃さ、くどい」の様な回答が複数あるものがあったので,「、」や「。」などの句読点や空白がある場合には分割を行い，重複する単語は削除した．また,「キャベツ」や「焼きいも」などの商品名であったり,「北海道」や「神戸」などの土地名（日本産などの産地は除く）,「20kg」や「5 キロ」などの具体的な重量,「カバンの中」,「弁当」,「どっち」,「英語」などの特定の単語はレビュー検索のための検索キーワードとして適切でないと考えたため，削除した．ただし，商品名のうち,「バジル」や「ニンニク」など風味などにかかわるものは一部保持した．また，重量の「kg」のような単位が単体のものは，検索キーワードとして適切であると考え，削除しなかった．

2：https://www.lancers.jp/

その結果 1,189 語の検索キーワードを取得した．取得した検索キーワードは「ギフト包装」「粒のそろい方」「配送」「コリコリ」「脂っこさ」などである．この検索キーワードを本研究では,「事前収集済み検索キーワード」と呼ぶこととする．

4.2.2 LLM による言い換え表現の取得

LLM の一つである ChatGPT によって「事前収集済み検索キーワード」の言い換え表現を取得した．この言いかえ表現を本研究では,「LLM 生成キーワード」と呼ぶ．

まず,「事前収集済み検索キーワード」を含むレビュー文を取得した．この際，検索キーワードとレビュー文それぞれを tokenizer で分割し，検索キーワードと分割の仕方が同じ表現を含むレビュー文のみを取得した．例えば,「コシ」を含むレビュー文を取得すると,「トウモロコシ」や「コショウ」などを含むレビュー文まで取得してしまっていた．しかし，tokenizer で分割を行うことによって，[コ, ##シ]，[トウモロコシ]，[コ, ##ショウ] となり,「コシ」とは異なる表現をできる限り含まないような工夫を行った．

ChatGPT に「LLM 生成キーワード」生成してもらう際，どのジャンルを対象とした検索キーワードであるかを示した．そのため，検索キーワードの表現を含んだレビュー文のジャンルは，その検索キーワードの対象ジャンルとして適切だと考え取得した．ジャンルは，あらかじめ収集していた検索キーワードを含むレビュー文の商品ジャンルパスを参考に，最下層の最も具体的なジャンルではなく，その上位のジャンルを取得した．例えば,「食品/フルーツ・果物/バナナ」ならば，最下層のジャンルである「バナナ」ではなく「食品/フルーツ・果物」までを取得した．これにより，同じ検索キーワードを持つ「みかん」や「マンゴー」などの最下層ジャンルを上位のジャンルで，一括して扱えるようにした．

また，同じ検索キーワードを含むレビュー文のうち，例えば,「匂い」という検索キーワードを含むレビュー文がある．そのレビュー文のうち，上位ジャンルが「食品/フルーツ・果物」である最下層のジャンルには「みかん，バナナ，パパイヤ，イチジク，ドラゴンフルーツ …」がある．このように，検索キーワードを含んだレビューにそのジャンルの様々な商品が含まれてる場合,「匂い」の検索対象となる商品のジャンルとして正しい可能性が高い．そのため，最下層の最も具体的なジャンル数の平均が約 4.7 個（小数点以下第 2 位で四捨五入）であったことから，最も具体的なジャンル数が 4 個以上の検索キーワードとジャンルの組み合わせを取得した．この結果，38,283 件の検索キーワードとジャンルのペアを取得することができた．

このように収集した，検索キーワードとジャンルのペアを使用して，以下のような文で ChatGPT に言いかえ表現を生成してもらった．以下の例は，ジャンルが「食品/調味料/スパイス/唐辛子」，検索キーワードが「辛口」のペアである．生成する際，temprature=0.0 を指定することで，結果のバラつきを最小限に抑えた．

- **実際に行ったプロンプト例**
  - 食品/調味料/スパイス/唐辛子を対象としたレビューにどのような具体的な表現が含まれていたら「辛口」に関するレビューだと思いますか。「辛口」に関連する語と、「辛口」と対照的な語を 5 個ずつください。”なるべく短い単語”にしてください。単語だけ欲しいです。単語は重複しないでください。() は使用しないでください。食品/調味料/スパイス/唐辛子の「辛口」に対する表現として適切でないものはいりません。関連する語 5 個と対照的な語 5 個のみをそれぞれ以下のフォーマットに格納してください。## フォーマット　関連する語:[“あ”,“い”,“う”] 対照的な語:[“あ”,“い”,“う”]
- **実際の生成例**
  - 関連する語:[“辛い”,“刺激的”,“ピリッとした”,“強烈な”,“火辛い”]
  - 対照的な語：[“甘口”,“マイルド”,“穏やかな”,“控えめな”,“やさしい”]

プロンプトでは，合計 10 個の言い換え表現の生成を行う指示を与えているため複数の表現が生成される．そのため，その中に重複する表現がある場合，重複をなくした．また,「LLM 生成キーワード」のうち，対象ジャンルのレビューにその表現が含まれない場合，適切な言い換え表現でないと判断し,「LLM 生成キーワード」から除外した．その結果全部で，381,349 語の言い換え表現を取得した．

4.2.3 適合，不適合データの作成

適合判定 BERT の構築には，モデルに検索キーワードとレビュー文の適合性を学習させるための適合データと不適合データが必要である．そのため，この学習に使用する適合データ,不適合データを作成する．適合データとは,「ジャンル，検索キーワード，レビュー文，適合ラベル」であり，不適合データとは,「ジャンル，検索キーワード，レビュー文，不適合ラベル」である．

まず，ジャンルごとにある「事前収集済み検索キーワード」とそれを基として生成した「LLM 生成キーワード」を 1 組のリストとして作成した．これを,「検索キーワードリスト」と呼び，全部で 38,283 個の「検索キーワードリスト」を作成した．この,「検索キーワードリスト」内の単語の組み合わせ「検索キーワード i」と「検索キーワード j」（$1 \leqq i, j \leqq$ それぞれの「検索キーワードリスト」の長さ）を取得し，対象ジャンルの商品のうち「検索キーワード i」を含むレビュー文と「検索キーワード j」を含むレビュー文の両方がある商品を抽出した．この商品のレビュー文集合のうち,「検索キーワードリスト」の単語を含んだレビュー文をランダムに 2 つ選び（シード値 = 42）,「検索キーワード i」と「検索キーワード j」の正例とした．また，同じ商品のレビュー文集合のうち,「検索キーワードリスト」の単語を 1 つも含まないレビューが 2 件ある場合,「検索キーワード i」と「検索キーワード j」の負例とした．その結果，正例は 12,261,334 件，負例は 12,099,540 件取得した．

4.2.4 ファインチューニング

4.1 節で構築したレビュー学習済み BERT モデルに 4.2.3 節で作成した適合・不適合データでファインチューニングを行っ

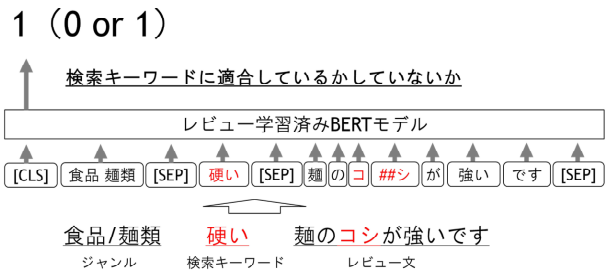


図 6　ファインチューニング

た．このファインチューニングでは，図 6 のように，ジャンル，検索キーワード，そしてレビュー文を入力とし，検索キーワードにレビュー文が適合しているかしていないかを判定した．

商品の検索キーワードとレビューを学習する際，単に検索キーワードとレビューの情報だけでなく，どのジャンルに関するものかの情報を含めた．これにより，例えば海苔とバナナの「黒」のような場合，同じ色について述べていても，バナナの「黒」は熟していることを示す一方で，海苔の「黒」は商品自体の色を指し，熟していることを示さないといった商品ごとの違いを考慮できる．このように，商品ごとの表現の意味の違いをジャンルによって考慮しながら学習できるようにした．また，ジャンルを入力する際に，「/」を空白におきかえて分割することで，どこまでがひとつのジャンルになっているのかを示した．

4.2.5　適合判定 BERT を用いた検索

検索結果は，レビューを検索キーワードに適合している順にランキングする必要がある．しかし，モデルにはレビューではなくレビュー文が入力されているため，適合度はレビュー文のスコアとなっている．そこで，同じ商品で同じユーザのレビュー文のうち最大のスコアをそのレビューのスコアとして，ランキングを行い，適合しているレビューを上位に，適合していないレビューを下位に検索できるようにした．この，適合度はシグモイド関数を通した後の，0〜1 の範囲で表されている．

4.3　アプローチ 2:LLM によるオンデマンドクエリ拡張

ユーザに入力された検索キーワードにマッチする多様な表現のレビューを上位に検索するために，オンデマンドにクエリ拡張を行う．まず，ユーザによって検索キーワードが入力されると，ChatGPT を用いてオンデマンドに検索キーワードの言い換え表現を最大 10 個取得する．これらを新たな検索キーワードとする．そして，4.2 節で構築した適合判定 BERT を利用し，それぞれの検索キーワードで個別に検索を行う．このようにオンデマンドに ChatGPT によってクエリ拡張を行うことで，あらかじめ収集しなくても，検索キーワードに関連する表現で検索を行える．クエリ拡張は，適合判定 BERT 構築のための学習データ作成時の言い換え表現生成と同じプロンプトで行う．また，適合判定 BERT から得ることができるレビュー 1 文ごとの最大のスコアをレビューのスコアとした．しかし，クエリ拡張を行った場合，検索キーワードごとにレビューにスコアが付与されている．そこで，各検索キーワードごとにある，同じレビューのスコアの最大をそのレビューの最終的なスコアとし，レビューのランキングを行う．このランキング結果を検索結果とする．

5　評価と考察

5.1　評価に使用する検索キーワードの収集

評価に使用する検索キーワードは，同じ研究室の学生 1 名に協力してもらい作成した．この検索キーワードは，ユーザが欲しい情報を含んだ食品レビューを検索するための検索キーワードである．検索キーワードの対象商品は，3.3 節であらかじめ収集していた 140 商品である．評価をしてもらう際には，商品 No，店舗 ID，商品名，商品ジャンルパスを提示した．

一般的に検索される検索キーワードとそうでない検索キーワードの両方が必要であると考え，各商品に対して，2 つずつ検索キーワードを収集した．例えば，商品名：「津軽産りんご 20 個入り　送料無料　おいしい　丹精込めて作った」，ジャンル：「食品/フルーツ・果物/りんご」を提示したとする．リンゴであれば，「甘さ」についての情報が欲しいユーザは多いと考えられる．そのため，「甘い」などの検索キーワードでレビュー検索を行うことが多いと考える．1 つ目は，このように対象商品に対して，ユーザが思いつきやすい検索キーワードを収集した．また，同じ商品でも，多くのユーザが思いつくような検索キーワードではなくても重要である．例えば，実際に商品を見ることのできないオンラインショッピングでリンゴに「カビ」が生えているかが気になることがある．そこで，2 つ目には「カビ」などユーザが真っ先には思いつかない検索キーワードを収集した．その結果，140 商品× 2 検索キーワードの 280 個の検索キーワードを収集することができた．

5.2　評価に使用する正解データの作成

同じ研究室の学生 2 名に協力してもらい，評価に使用する正解データを作成した．正解データ作成の対象データは，3.3 節で作成した評価データである．具体的には，検索の対象となるレビューが，5.1 節で収集した検索キーワードで検索した時に検索されてほしいレビューであるかを「2,1,0」の 3 段階で評価してもらった．提示する検索キーワードに対して，それぞれのレビューが関連している場合は 2 部分的に関連している場合は 1，関連していない場合は 0 を入力してもらった．評価してもらう際には，商品 No，店舗 ID，レビュー ID，商品名，商品ジャンルパス，レビュー，検索キーワードを提示した．以下に評価方法の例を示す．

- **ユーザが参照するもの**
  - 商品名：「カレー　送料無料　おいしい　丹精込めて作った」
  - ジャンル：「食品/惣菜/カレー」
  - 検索キーワード：「甘い」
- **提示した回答例**

レビュー	評価
辛党なので、甘ったるく感じました。	2
子供でも食べられる味でした。	1
ドロドロとしたカレーでした。リピです！	0

この例のように，明らかにレビューが検索キーワード「甘い」にマッチしているものには 2，辛くないことが推測でき「甘い」に関する情報が得られる場合は 1，検索キーワードに関する情報を得ることができず，明らかにマッチしていないレビューは 0 と評価してもらった．

この結果，140 商品× 100 レビュー× 2 検索キーワードの計 28,000 件の正解データを収集した．

5.3 検索キーワードの表現を含まない正解データ

本研究では，検索キーワードの表現が検索対象のレビュー集合に含まれていない時でも，検索キーワードにマッチしたレビューを検索できるようにすることを目的としている．そのため，評価に使用するレビューを検索キーワードの表現を含まないレビューに限定した．そのため 28,000 件から合計 26,988 件のレビューを正解データの中から抽出し，検索キーワードの表現を含まない評価データとした．

5.4 評　　価

検索キーワードの表現を含まない評価データを用いて，検索キーワードにマッチするレビューを上位に検索できているかを評価した．ユーザは，検索キーワードへのマッチ度が高い上位 10 件のレビューを確認すると仮定をおき，nDCG@10 を用いて評価した．計算には，検索評価指標を計算するためのツールキットである NTCIREVAL を利用した．nDCG@10 は，検索結果の上位 10 件のレビューのうち，上位に適合文書がある場合には値が高くなり，下位に適合文書がある場合にはペナルティが課される．評価は，商品と検索キーワードの 280 ペアを入力として，適合判定 BERT のみを使用して検索した時と，クエリ拡張を行った時を比較した．また，既存手法である TF-IDF や OkapiBM25 との比較や，ChatGPTEmbeddings(text-embedding-ada-002) や sentenceBERT(sonoisa/sentence-bert-base-ja-mean-tokens) とも比較を行った．この際，それぞれの手法に 2 つ目のアプローチであるクエリ拡張も行った．計算結果は商品と検索キーワードの 280 ペアそれぞれで計算を行った平均である．比較するためにそれぞれの手法での計算結果を表 4 に示す（小数点以下第 4 位を四捨五入）．

検索キーワード「甘さ」に対して,「甘味」「甘い」など語幹が同じ表現がレビューに頻出する場合 OkapiBM25 の評価は高いことが分かった．また，ChatGPTEmbeddings でも OkapiBM25 と同様に検索キーワードと語幹が同じ表現が含まれるレビューが上位にランキングされる傾向にあった．しかし，ユーザから入力される検索キーワードには，レビューに完全に一致する表現や語幹が同じ表現が含まれているとは限らない．そのため，ユーザが思いつきやすい検索キーワードとユーザが思いつきにくい検索キーワードそれぞれで，どのような評価結果になるのかを比較した．評価結果を表 4 に示す（小数点以下第 4 位を四捨五入）．

5.5 考　　察

評価結果より，適合判定 BERT にクエリ拡張を行った手法が検索キーワードにマッチしたレビューを上位に検索できる結果となった．また，ユーザが思いつきにくい検索キーワードで検索した時に，本研究の検索手法は特に有用であることが分かった．

クエリ拡張を行うと，OkapiBM25 や ChatGPTEmbeddings,SentenceBERT にはあまり有効ではなかったが，本研究で構築した適合判定 BERT には有効であることが分かった．クエリ拡張を行った時に，例えば「ナッツ」の「甘さ」に対してクエリ拡張を行った時に「甘さ」の対照的な語として「塩味」が生成されていた．この時,「ナッツ」のレビューには,「塩味」という単語が含まれているレビューがいくつかあり，スコアも高くなる傾向にあった．そのため，甘さに対して「塩味」などのレビューが上位に出やすくなってしまった．その結果，OkapiBM25 や ChatGPTEmbeddings などでクエリ拡張を行うと,「甘さ」に関係のないレビューが上位に含まれてしまったのだと考えられる．

ユーザが思いつきにくい検索キーワードで検索を行った結果についての考察を行った．例えば，評価対象の商品である「鶏卵」のレビュー集合に検索キーワード「割れ」でレビュー検索を行ったうと,「割れ」に対して，他の手法では取得できなかった「破損」という表現を含んだレビューを適合判定 BERT は検索することができていた．さらに，クエリ拡張を行うことで，実際に「まれにヒビの入った卵があることもありますが、大概は無傷で届きます。」（前後の文章は省略）のような「割れ」に関するレビューを上位に検索することができていた．また,「ハマグリ」のレビューに対して「大きさ」と検索すると,「小さい」という表現を含んだレビューを検索することができていた．これらのことから，語幹が一致する表現がレビューに含まれない時でも本研究の提案手法である，適合判定 BERT と適合判定 BERT にクエリ拡張を行った手法では，検索キーワードにマッチするレビューを検索できていることが分かった．そのため，本研究の検索手法では，ユーザの入力がレビューの表現とあまりマッチしていない場合でも，検索キーワードを変えて検索することなく，欲しいレビューを入手できる結果となった．

6 まとめと今後の課題

本研究では，LLM を用いて言い換え表現作成による適合判定 BERT の構築及びオンデマンドクエリ拡張を行い，検索キーワードにマッチしたレビューを検索できる手法を構築した．本研究の提案手法を用いることで，ユーザの検索キーワードにとらわれないレビュー検索を実現できる．

この手法では，あらかじめ検索キーワードの候補となる語があれば，検索キーワードの候補の表現以外を含んだレビューを検索できるようにすることができる．収集した語は，ユーザがレビュー検索に使用するありとあらゆる検索キーワードを網羅する必要はない．そのため，食品レビュー以外に対しても適用できると考えられる．

今回はデータ数が膨大であることと，精度向上を目指して，対象とするデータを「食品」と「スイーツ・お菓子」に絞った．しかし，ユーザはオンラインショッピングを行う時「食品」や

表 4　nDCG@10 による評価結果

手法	nDCG@10 （全検索キーワード）	nDCG@10 (ユーザが思いつきやすい検索キーワード)	nDCG@10 (ユーザが思いつきにくい検索キーワード)
TF-IDF	0.298	0.341	0.246
TF-IDF + クエリ拡張	0.418	0.502	0.317
OkapiBM25	0.512	0.620	0.385
OkapiBM25 + クエリ拡張	0.448	0.524	0.357
ChatGPTEmbeddings	0.464	0.503	0.419
ChatGPTEmbeddings + クエリ拡張	0.406	0.440	0.366
sentenceBERT	0.311	0.348	0.266
sentenceBERT + クエリ拡張	0.253	0.295	0.204
提案手法（適合判定 BERT）	0.450	0.474	0.422
提案手法（適合判定 BERT + クエリ拡張）	0.536	0.603	0.456

「スイーツ・お菓子」だけではなく,「家電」や「服」などその他のジャンルの商品も購入する．そこで，今後は今回提案した検索システムをベースに，対象商品の拡張に取り組み，多様な商品に適応できる汎用的な検索システムを構築したい．

謝　　辞

本研究は，JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けたものです．また，本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用しました．ここに記して謝意を表します．

文　　献

[1] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 642–652, 2020.

[2] Ori Ben-Yitzhak, Nadav Golbandi, Nadav Har'El, Ronny Lempel, Andreas Neumann, Shila Ofek-Koifman, Dafna Sheinwald, Eugene Shekita, Benjamin Sznajder, and Sivan Yogev. Beyond basic faceted search. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 33–44, 2008.

[3] Tianyu Chen, Shaohan Huang, Furu Wei, and Jianxin Li. Pseudo-label guided unsupervised domain adaptation of contextual embeddings. In *Proceedings of the 2nd Workshop on Domain Adaptation for Neuro Linguistic Programing*, pp. 9–15, 2021.

[4] Hasna Chouikhi, Hamza Chniter, and Fethi Jarray. Arabic sentiment analysis using bert model. In *Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13*, pp. 621–632, 2021.

[5] Lei Fang, Qiao Qian, Minlie Huang, and Xiaoyan Zhu. Ranking sentiment explanations for review summarization using dual decomposition. In *Proceedings of the 23rd Association for Computing Machinery International Conference on Conference on Information and Knowledge Management*, pp. 1931–1934, 2014.

[6] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. mith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

[7] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pp. 755–760, 2004.

[8] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[9] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, pp. 342–351, 2005.

[10] Ayana Niwa, Keisuke Nishiguchi, and Naoaki Okazaki. Predicting antonyms in context using bert. In *Proceedings of the 14th International Conference on Natural Language Generation*, pp. 48–54, 2021.

[11] Madhumita Sushil, Simon Suster, and Walter Daelemans. Are we there yet? exploring clinical domain knowledge of bert models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 41–53, 2021.

[12] Gesare Asnath Tinega, Waweru Mwangi, and Richard Rimiru. Text mining in digital libraries using okapi bm25 model. *International Journal of Computer Applications Technology and Research*, pp. 398–406, 2019.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[14] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2557–2563, 2015.

[15] 横山誠, 村上貴彦, 中村太戯留, 上林憲行ほか. Wikipedia からキーワードを援用し初出の知識の学習と学習コンテンツの発掘を支援するサービスの研究. 第 77 回全国大会講演論文集, pp. 701–702, 2015.

[16] 河内沙織, 豊田哲也, 延原肇. Wikipedia カテゴリおよび自己相互情報量に基づく関連検索キーワード生成による知識拡充支援. 第 73 回全国大会講演論文集, pp. 607–608, 2011.

[17] 栗田多喜夫, 加藤俊一, 福田郁美, 坂倉あゆみほか. 印象語による絵画データベースの検索. 情報処理学会論文誌, Vol. 33, No. 11, pp. 1373–1383, 1992.

[18] 堀憲太郎, 大石哲也, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. Wikipedia からの拡張クエリ生成による web 検索とその評価. 人工知能学会第二種研究会資料, p. 13, 2009.
[19] 桑折章吾, 加藤優, 高間康史. 検索エンジンを用いた情報検索におけるユーザ行動の分析. 人工知能学会第二種研究会資料, Vol. 2013, p. 02, 2013.
[20] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. 関連単語抽出アルゴリズムを用いた web 検索クエリの生成. 電子情報通信学会論文誌 D, pp. 281–292, 2009.
[21] 福岡知隆, 南貴大, 浦田渡, 藤生慎, 高山純一. 深層学習による橋梁点検のための pix2pix による疑似訓練データ生成. 土木学会論文集 F4 (建設マネジメント), pp. I_27–I_35, 2019.
[22] 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔. 事前学習と追加事前学習による金融言語モデルの構築と検証. 人工知能学会第二種研究会資料, p. 132, 2022.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 10:00 ~ 12:10 | T3-B オンライン（Zoom Events)

画像検索・生成

座長:古賀 久志(電気通信大学)
コメンテータ:片山 薫(東京都立大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T3-B-4-01] 条件付き画像検索と画像生成の統合

*澤田 一正[1]、坂地 泰紀[1]、野田 五十樹[1]、小山 聡[2] (1. 北海道大学、2. 名古屋市立大学)

10:20 ~ 10:45
[T3-B-4-02] Finding Generative Image LoRA Model by Inputting Style Sample Image

*VUTHI NGOCANH[1]、Shoji Yoshiyuki[1]、Pham HuuLong [2]、Ohshima Hiroaki[2] (1. Shizuoka University、2. University of Hyogo)

10:45 ~ 11:10
[T3-B-4-03] 生成型要約に基づくWebページのサムネイル生成

*前田 直宏[1]、山本 岳洋[1] (1. 兵庫県立大学)

11:10 ~ 11:35
[T3-B-4-04] 類似度グラフの事前再計算による拡散式画像検索の効率化

*加藤 辰弥[1,2]、駒水 孝裕[2]、井手 一郎[2] (1. 名古屋大学大学院知能システム学専攻　井手・駒水研究室、2. 名古屋大学)

11:35 ~ 12:00
[T3-B-4-05] 3Dモデルを利用したユーザーの求めるアングル画像の検索システムの開発

*永田 玄[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学情報理工学院情報工学系情報工学コース宮崎研究室)

条件付き画像検索と画像生成の統合

澤田　一正†　　坂地　泰紀††　　野田　五十樹††　　小山　　聡†††

† 北海道大学大学院情報科学院　〒 060–0814 札幌市北区北 14 条西 9 丁目
†† 北海道大学大学院情報科学研究院　〒 060–0814 札幌市北区北 14 条西 9 丁目
††† 名古屋市立大学データサイエンス学部　〒 467–8501 名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †intsawadan@eis.hokudai.ac.jp, ††{sakaji,i.noda}@ist.hokudai.ac.jp, †††oyama@ds.nagoya-cu.ac.jp

あらまし　条件付き画像検索は、クエリ画像と、クエリ画像に対する追加要求のテキストのペアから画像を検索するタスクであり、ファッションのデザインなど発想支援的な目的でも用いられる。条件付き画像検索では、クエリによっては適合する画像が十分に検索されない場合があるが、データベースに存在しない画像を生成することで、クエリの改善や発想の支援につながる可能性がある。本研究では、生成 AI を用いて条件付き画像検索と同じ設定で画像を生成し、既存の検索モデルで検索された結果と組み合わせる手法を提案する。さらに、統合した結果をクラウドソーシングを用いて評価する。

キーワード　検索モデル、検索クエリ、クラウドソーシング、LLM、評価・データセット

1　はじめに

オンラインショッピングにおいては画像を元に検索を行う手段の開発が重要となってきている。インターネットの普及により、オンラインショッピングが爆発的な成長を遂げ、従来の実店舗に代わって多くの人が商品を購入する主要な手段となっている。それに伴い、オンライン上でユーザーが望んでいる商品を検索する重要性が高まっており、様々な検索方式が提案されている。その 1 つとして、条件付き画像検索 (Conditioned Image Retrieval) という検索方式がある。条件付き画像検索は、画像と、その画像をどのように修正するかのテキストからなる検索クエリを用いて画像を検索する検索方式である。この検索方式は、衣服などの、視覚的特徴により検索を行うことができる商品に適しており、現在閲覧している商品画像に対して、修正点を入力することにより新たな商品を検索できるため、より望んだ商品を見つけやすくなることが期待できる。しかし、条件付き画像検索に関する既存の研究は、正解画像をどれだけ正確に検索できるか、といった検索の精度に焦点を当てているものが多く、検索結果の多様性などの、検索の精度以外の観点がほとんど考慮されていない。

近年、テキストや画像を生成する様々な生成 AI モデルが登場し、幅広い領域への活用が検討されている。例えば、衣服の新たなデザインの生成といった新たな商品画像の生成に活用されており、商品を検索するだけではなく、新たな商品を生成する試みがされている [1]。しかし、画像生成モデルは、高品質な画像を生成することに優れているものの、生成画像の多様性に欠けているという問題が指摘されている [2]。

また、オンライン上の商品検索において、ユーザーが最初の検索クエリとは完全に一致しないアイテムを購入する可能性があることが指摘されている [3]。例えば、最初は「黒いダウンジャケット」と検索したユーザーが、検索結果を見ている中で、ブルーのダウンジャケットが目に留まり、結果として最初の検索クエリとは異なる、ブルーのダウンジャケットを購入するといったことが考えられる。そのため、商品の検索結果は、検索クエリに完全に一致している必要はなく、検索クエリに関連する商品なども含む、検索結果の多様性が重要であると考えられる。

本研究では、条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人が、本当に欲しい商品を見つけられるような商品画像の結果を出力する手法を提案する。具体的には、検索結果の検索クエリに対する適合性と多様性を両立する結果を得られることを目指す。既存の条件付き画像検索モデルによる検索結果と生成 AI により生成された商品画像を組み合わせることにより、生成 AI の多様性に欠けてしまう問題を解決しながら、検索クエリへの適合性を高めることを目標とする。

2　関連研究

2.1　条件付き画像検索 (Conditioned Image Retrieval)

条件付き画像検索 (Conditioned Image Retrieval) は、画像と、その画像をどのように修正するかのテキストからなる検索クエリを用いて画像を検索する画像検索タスクである。ここで、検索クエリの画像を reference image、検索クエリのテキストを relative caption、検索結果の正解画像を target image という。条件付き画像検索の一連の流れを図 1 に示す。

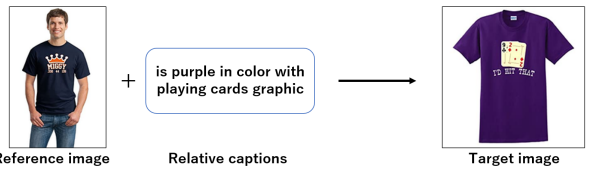


図 1　条件付き画像検索の流れ

条件付き画像検索はファッションのための対話型検索システムを実装するためのモデルとして提案された [4]。このシステムでは、条件付き画像検索を活用することにより、商品の画像と、その画像をどのように修正するかのテキストを入力することで、新たな商品を検索できるようになっている。また、検索された商品の画像に対しても、それをどのように修正するかのテキストを入力して検索することができるため、対話を通して商品を検索することができる。

条件付き画像検索のタスクにおいて、高い精度で検索を行うための様々な検索モデルが提案されてきた。Baldrati ら [5] は、CLIP [6] を活用して、条件付き画像検索を行うモデルを提案した。CLIP を用いて reference image と relative caption からそれぞれ特徴量を抽出する。これらの特徴量を組み合わせた新たな特徴量を作成するネットワークを構築し、その組み合わせた特徴量が target image の特徴量に近づくように学習を行う。Lee ら [7] は、2 つの異なるニューラルネットワークのモジュールを使用し、条件付き画像検索を行うモデルを提案した。1 つは relative caption に基づいて、reference image の特徴に対して、オブジェクトの形状やサイズ、配置などの詳細を修正するモジュールである。もう 1 つは、reference image の全体的な色調、質感などを修正するモジュールである。Chen ら [8] は、検索クエリのテキストである relative caption には、検索候補が 1 つに定まるような詳細に記述されているものと、検索候補が複数存在するような曖昧に記述されているものがある点に着目し、その両方の種類の relative caption の学習に対応した条件付き画像検索を行うモデルを提案した。

しかし、これらの既存の研究のほとんどが検索結果の精度に焦点をあてており、検索結果の多様性等の観点が考慮されていない。実際にこれらの検索システムを使用するユーザーの観点で考えると、検索結果の多様性は大切であるので、検討する必要があると考えられる。また、条件付き画像検索のタスクは基本的に、すでにあるデータベース内の画像から検索をおこなうため、データベース内に無い画像は検索できない。そのため、具体的に検索をしたいものがあったとしても、望ましい検索結果が得られない可能性がある。

2.2 画像生成を用いた衣服の画像生成

近年、画像生成を行う深層学習モデルの性能が向上しており、幅広い領域への活用が検討されている。Choi ら [9] は人間のファッションデザイナーが行う作業を模倣する AI ベースの衣服開発システムを開発した。具体的には、データ分析、トレンド抽出、デザイン生成、修正の各段階をサポートする 4 つのモジュールを備えた AI システムを提案した。デザイン生成には GAN(Generative Adversarial Network) を使用している。これらの画像生成を行う深層学習モデルを用いることで、条件付き画像検索において、検索したい画像がデータベース内に存在しない場合でも、検索クエリに適合する画像を新たに生成できる可能性がある。しかし、画像生成モデルは、高品質な画像を生成することに優れているものの、生成画像の多様性に欠けているという問題が指摘されている [2]。

そのため、条件付き画像検索と画像生成を統合することで、各々にかけているところを補うところに、本研究の狙いがある。それを検証するために使用できるデータセットを次節で紹介する。

2.3 データセット

条件付き画像検索のデータセットには、FashionIQ [4] や CIRR [10] がある。

FashionIQ はファッション画像に関するデータセットで、”Dress”、”Shirt”、”Toptee”の 3 つのカテゴリに分けられている。また、学習データ、検証データ、テストデータの 3 つに分割して提供されている。学習データには 46,609 枚の画像が含まれており、reference image と 2 件の relative caption、target image の検索セットが 18,000 件ある。検証データには 15,537 枚の画像が含まれており、検索セットが 6,117 件ある。テストデータには 15,538 枚の画像が含まれており、検索セットが 6,119 件ある。本研究の 4 章の実験にて、FashionIQ を使用する。

CIRR(Compose Image Retrieval on Real-life images) はファッションなどの特定のドメインに限定しない、一般画像に関するデータセットである。学習データ、検証データ、テストデータの 3 つに分割して提供されている。学習データには 16,939 枚の画像が含まれており、reference image と relative caption、target image の検索セットが 28,225 件ある。検証データには 2,297 枚の画像が含まれており、検索セットが 4,184 件ある。テストデータには 2,316 枚の画像が含まれており、検索セットが 4,148 件ある。

3 提案手法

3.1 検索システム

本研究の提案手法である検索システムについて説明する。検索システムの概要を図 2 に示す。この図には、検索クエリが入力されてから検索結果が出力されるまでの流れを示している。この検索システムでは、ユーザーは商品の画像とその画像をどのように修正するかのテキストからなる検索クエリを入力し、新たな商品画像の検索を行う。検索システムは、ユーザーから得られる検索クエリを入力とし、10 枚の画像の集合を検索結果として出力する。検索結果の画像枚数が 10 枚とした理由は、ユーザーにとってある程度選択肢があり、1 度にすべての商品を確認できる数であると考えたためである。画像検索モデルと画像生成モデルという別の 2 つの手法から得られた結果を組み合わせることで、各々にかけているところを補うことを期待して、提案手法を考案した。提案手法の流れとしては、まず既存の条件付き画像検索モデルを使用して、画像を検索する。次に、画像生成モデルを用いて検索と同じ条件で画像を生成する。すなわち、検索と同じ reference image と relative caption を用いて新たな画像を生成する。最後に、画像検索と画像生成で得られた画像を組み合わせることで新たな画像の集合を作成する。検索システムの各要素について次節より説明する。

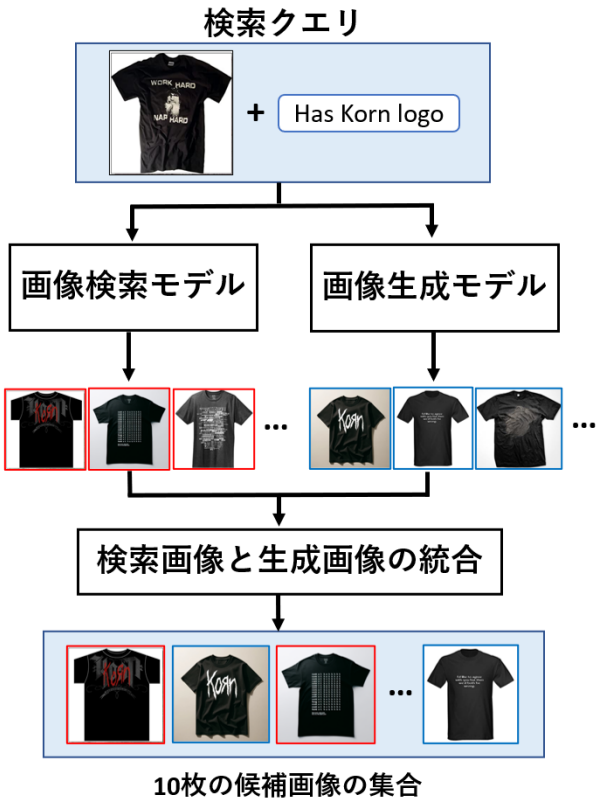


図 2　検索システムの概要

3.2　条件付き画像検索モデル

2.1 節で説明した既存の条件付き画像検索のモデル (MUR) [8] を使用し、画像データベース内から検索条件にあった画像を検索する。MUR では、reference image と relative caption の特徴量を組み合わせ、新たな特徴量を作成する。その新たに作成された特徴量と、画像データベース内のすべての画像の特徴量とのコサイン類似度を計算し、その値が大きい順に、検索結果の上位としている。本研究では、上記の手順で計算されるコサイン類似度の値を適合度と呼ぶ。本研究でも、MUR と同じように、データベース内の各画像に対して、適合度を計算する。その数値が高い順に、検索結果の上位とする。最終的に検索結果の上位 10 枚を出力する。

3.3　条件付き画像生成モデル

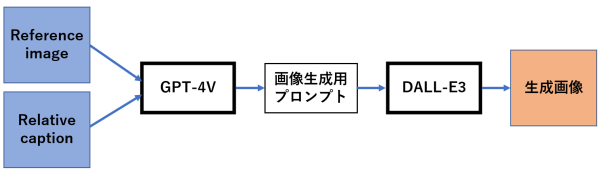


図 3　条件付き画像生成の流れ

条件付き画像検索と同じ reference image と relative caption を用いて新たな画像を生成する。条件付き画像生成の全体の流れは図 3 のようになる。まず初めに、GPT-4V(ision) [11] に reference image と relative caption を入力して、条件に合った画像を画像生成モデルで生成するためのプロンプトの作成を行う。画像生成用のプロンプトを作成する理由としては、画像生成で使用するモデルがテキストの入力しか受け付けていないためである。このとき、GPT-4V に対して以下のプロンプトを入力した。ここで、relative caption はプロンプトの最後に空行を 1 行追加し、その次の行に記述した。

> Please create a text prompt for DALL·E to generate a modified image using the following text based on the attached image.I would like to generate images from text prompts only, without inputting any images into DALL·E. Please come up with a prompt for this scenario. Try to retain as many detailed features from the provided original image as possible and make modifications only to the specified parts in the modified text.Please create a prompt that generates images of clothing only, not worn by a person. Create a prompt that generates images showing the entire piece of clothing, not just a part of it. Please output only the prompt.
>
> [relative caption]

上記のプロンプトでは、reference image の特徴を残しながら、relative caption で指定された部分のみを修正するように指示しており、今回の条件に適していると考えられる。

次に、生成された画像生成のためのプロンプトを用いて画像の生成を行う。画像生成モデルとしては DALL-E3 [12] を使用した。1 つの検索クエリに対して図 3 に示した一連の流れを、10 回繰り返し行うことにより、10 枚の画像を生成する。GPT-4V は、標準の設定が、モデルが生成するテキストの創造性と予測可能性のバランスをとるようになっているため、同じ入力を行っても、異なる出力が期待できる。

3.4　検索画像と生成画像の統合

検索モデルによって得られた画像 (検索画像)10 枚と生成モデルによって得られた画像 (生成画像)10 枚を組み合わせて、新たな 10 枚の画像の集合を作成する。このアルゴリズムを以下にまとめる。

(i)　生成画像 10 枚に対して、検索のときと同様に条件付き画像検索のモデルを使用して、検索条件にどのくらい適合しているかを表す適合度を計算する。

(ii)　検索画像 10 枚と生成画像 10 枚を合わせて 20 枚の画像の集合とし、適合度が高い順に並べる。

(iii)　もし上位 10 枚の画像の中に含まれている生成画像が、5 枚以上の場合は、その上位 10 枚の画像を新たな 10 枚の画像の集合とし、アルゴリズムを終了する。

(iv)　もし上位 10 枚の画像の中に含まれている生成画像が、5 枚未満の場合は、(検索画像の中で一番高い適合度) - (生成画像の中で一番高い適合度) - 0.0001 を計算し、その値が 0 以上の場合は、生成画像の上位 5 枚の適合度に、計算した値を加算する。

(v) 検索画像上位 5 枚と生成画像上位 5 枚を合わせて、適合度が高い順に並べ替え、その 10 枚を新たな 10 枚の画像の集合とし、アルゴリズムを終了する。

4 実　　験

提案手法が、条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人に対して、本当に欲しい商品を見つけられるような商品画像の結果を出力することができるのかを検証するために、実験を行う。本研究では、2 つの手法を比較したときにどの程度の割合で検索クエリに適合しているかを表す「検索クエリに対する適合比」と、検索結果にどの程度多様性があるかを表す「検索結果の多様性」の 2 つの観点で評価を行う。それぞれの定義は 4.2 節で説明する。

4.1 実験設定

4.1.1 使用したデータ

本研究では 2.3 節で説明した FashionIQ [4] というファッションに関する条件付き画像検索のデータセットを使用した。本研究の実験では、データセットの 3 つのカテゴリの中から、“Shirts” のカテゴリを使用した。また、検索を行うときの検索候補画像として、FashionIQ の学習データと検証データに含まれるすべての画像を使用した。実験で使用する検索クエリは検証データのものを用いた。また、FashionIQ には 1 つのクエリに対して relative caption が 2 つ含まれているため、それらを “and” でつなげることにより、1 つの relative caption として扱った。

4.1.2 比 較 手 法

本研究では、比較手法として、2.1 節で説明した既存の条件付き画像検索モデル [8] と 3.3 節で説明した提案手法の条件付き画像生成モデル 3.3 を用いた。条件付き画像検索モデルは、検索クエリを入力し、検索候補画像から画像を 10 件検索しそれらを出力する。条件付き画像生成モデルは、検索クエリを入力し、画像を 10 枚生成しそれらを出力する。

4.2 実験内容

各手法について、検索クエリから 10 枚の画像の集合 (画像集合) を作成する。それらの画像集合について、「検索クエリに対する適合比」と「検索結果の多様性」の 2 つの観点で評価を行う。この 2 つの観点について説明する。

4.2.1 検索クエリに対する適合比

検索クエリに対する適合比とは、2 つの手法を比較したときにどの程度の割合で検索クエリに適合しているかを表すものである。本研究では、2 つの手法を 10 件の検索クエリを用いて比較し、検索結果が検索クエリに対してより適合しているとクラウドソーシングにより判断された検索件数の比を検索クエリに対する適合比と定義する。例えば、手法 A と手法 B を比較する場合に、10 件の検索のうち、検索結果が検索クエリに対してより適合していると判断された件数が、手法 A が 2 件、手法 B が 8 件である場合、検索クエリに対する適合比は 手法 A : 手法 $B = 2 : 8$ となる。1 つの検索クエリに対して、5 人のクラウドワーカーに、2 つの検索結果のうち、どちらがより検索クエリにふさわしいかを選択してもらい、過半数である 3 人以上が選択した方の手法をその検索クエリに対してより適合していると判断する。検索クエリに対する適合比は、10 件の検索クエリそれぞれに対して 5 人のクラウドワーカーに評価を行ってもらっているため、延べ 50 件の調査により計算されている。ここで、10 件の検索クエリすべてに対して、同一の 5 人のクラウドワーカーが評価を行っているわけではなく、検索クエリごとにクラウドワーカーは異なる場合がある。クラウドソーシングのサービスはランサーズ [1] を用いた。

本研究では実験により、①画像検索モデルと提案手法の検索クエリに対する適合比、②画像検索モデルと画像生成モデルの検索クエリに対する適合比、③提案手法と画像生成モデルの検索クエリに対する適合比、の 3 つを調査した。上記の 3 つの調査を行った理由は、3 つの手法のすべての組み合わせに対して比較を行うためである。上記の 3 種類の検索クエリに対する適合比を評価する際の検索クエリ 10 件は、すべて同じものを使用した。今回の実験では、30 代後半〜60 代前半の 40 人がクラウドワーカーとして参加した。全員日本人であり、男女比は約 5:1 で男性の方が多かった。

4.2.2 検索結果の多様性

画像集合がどのくらい多様性があるかを評価する。多様性の評価指標として、情報推薦の分野で広く使用されている intra-list diversity(ILD) [13], [14] を利用した。この ILD という指標は、情報検索や推薦システムにおいて、検索結果や推薦結果のリスト内の項目間の多様性を測定するために使用される指標である。本研究においても、10 件の画像集合を ILD における検索結果のリストとみなすことができると考えたため、この指標を使用した。画像集合 I の多様性の値 $div(I)$ は以下のように計算される。

$$div(I) = \frac{1}{|I|(|I|-1)} \sum_{i \in I} \sum_{j \neq i \in I} (1 - sim(i,j)) \qquad (1)$$

ここで、$sim(i,j)$ は画像集合内の画像 i、j 間の類似度である。この類似度の値は、学習済みの CLIP [6] の Image Encoder を使用し画像の特徴量を抽出し、それらの特徴量のコサイン類似度の値とした。CLIP は画像とテキストのペアを理解するために設計されたモデルである。 インターネット上の大量の画像とテキストのペアから学習しており、関連する画像とテキストが類似した特徴表現を持つようになっている。CLIP はゼロショットで様々な種類の分類問題のデータセットを高い精度で分類できることが示されており [6]、多様な色、デザイン、形状などの衣服のビジュアル表現を理解する能力が備わっていると考えられる。以上の理由から、今回は学習済みの CLIP の Image Encoder を使用した。この多様性の値 $div(I)$ は、値が大きくなればなるほど多様性が高くなる。

実験の条件は、クエリへの一致度と同様に、①画像検索モデルと提案手法の比較、②画像検索モデルと画像生成モデルの

1 : https://www.lancers.jp/

比較、③提案手法と画像生成モデルの比較、の 3 つの条件で行った。それぞれの条件に対して、30 件の検索クエリで評価した。各条件では、2 つの手法で得られた画像集合の多様性の値 $div(I)$ を計算し、値が大きい方をより多様性の高い画像集合と判断した。

4.3 実験結果

4.3.1 検索クエリに対する適合比

クラウドソーシングにより得られた、2 つの手法を比較したときの、検索クエリに対してより適合していると判断された検索件数を表 1-3 に示す。4.2.1 節で説明したように、1 つの検索クエリに対して、5 人のクラウドワーカーに、2 つの検索結果のうち、どちらがより検索クエリにふさわしいかを選択してもらい、過半数である 3 人以上が選択した手法をその検索クエリに対してより適合していると判断する。表 1 は、画像検索モデルのと画像生成モデルを 10 件の検索クエリを用いて比較したときの、検索クエリに対してより適合していると判断された検索件数を示している。表 2 は画像検索モデルと提案手法の比較であり、表 3 は画像生成モデルと提案手法の比較である。表 1 より、10 件の検索クエリのうち、検索クエリに対してより適合していると判断された検索件数は、画像検索モデルが 2 件、画像生成モデルが 8 件であることが示されている。したがって、4.2.1 で定義した検索クエリに対する適合比は、画像検索モデル : 画像生成モデル = 2 : 8 となる。同様にして、表 2 より、画像検索モデルと提案手法の検索クエリに対する適合比は、画像検索モデル : 提案手法 = 3 : 7 となり、表 1 より、画像生成モデルと提案手法の検索クエリに対する適合比は、画像生成モデル : 提案手法 = 8 : 2 となる。

表 1　画像検索モデルと画像生成モデルの検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像検索モデル	2
画像生成モデル	8

表 2　画像検索モデルと提案手法の検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像検索モデル	3
提案手法	7

表 3　画像生成モデルと提案手法の検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像生成モデル	8
提案手法	2

4.3.2 検索結果の多様性

実験結果を表 4-6 に示す。これらの表は、2 つの手法を比較したときの、多様性の値が大きく、より多様性が高いと判断された検索クエリ数を示している。例えば、表 4 では、30 件の検索結果の多様性の値を比較したとき、画像検索モデルの多様性の値の方が高い検索結果が 26 件、画像生成モデルの多様性の値の方が高い検索結果が 4 件あったことを示している。各手法の多様性の値の平均値を表 7 に示す。この表では、各手法の 30 件すべての検索結果に対して計算された 30 個の多様性の値の平均値を表している。例えば、画像検索モデルの 30 個の多様性の値の平均値は、0.25561 となっている。

表 4　検索と生成の比較

	多様性の高い検索件数
画像検索モデル	26
画像生成モデル	4

表 5　検索と提案手法の比較

	多様性の高い検索件数
画像検索モデル	11
提案手法	19

表 6　生成と提案手法の比較

	多様性の高い検索件数
画像生成モデル	0
提案手法	30

表 7　多様性の値の平均値

	ILD の平均値
画像検索モデル	0.25561
画像生成モデル	0.16352
提案手法	0.26276

4.4 考　　察

4.4.1 検索クエリに対する適合比

4.3.1 節の実験結果から、検索クエリに対する適合比はそれぞれ、画像検索モデル : 画像生成モデル = 2 : 8、画像検索モデル : 提案手法 = 3 : 7、画像生成モデル : 提案手法 = 8 : 2 となった。画像検索モデル : 画像生成モデル = 2 : 8 と 画像生成モデル : 提案手法 = 8 : 2 の結果より、画像生成モデルが他の 2 つの手法よりも検索クエリに対する適合比の比率が高くなっていることがわかる。このことから、画像生成モデルは他の 2 つの手法と比べて、より検索クエリにふさわしい検索結果を出力すると考えられる。また、画像検索モデル : 提案手法 = 3 : 7 と 画像生成モデル : 提案手法 = 8 : 2 の結果より、提案手法は画像生成モデルよりも検索クエリに対する適合比の比率が低いが、画像検索モデルより比率が高くなっていることがわかる。このことから、提案手法は画像生成モデルよりも検索クエリにふさわしい検索結果を出力する能力は低いが、画像検索モデル

よりは検索クエリにふさわしい検索結果を出力する能力が高いと考えられる。

画像生成モデルが検索クエリに対する適合比で最も高い比率となった理由は、今回の実験条件が画像生成モデルにとって有利な条件であったためだと考えられる。なぜなら、今回使用したデータセットは、検索クエリに対する正解画像が 1 枚しかないという条件で作成されたものであるため、今回の 10 枚の検索結果を表示してその結果を比較するために十分な画像がデータベースに存在しない可能性が高いからである。また、画像生成モデルが、細かい条件に対応した画像を生成するのに優れているためだと考えられる。

4.4.2 検索結果の多様性

4.3.2 節の結果から、検索結果の多様性について考察する。

表 5 と表 6 の結果から、提案手法は他の 2 つの手法と比較して、多様性の高い検索件数が多くなっていることがわかる。また、表 7 の結果から、提案手法の多様性の値の平均値が最も高くなっていることがわかる。このことから、提案手法は他の 2 つの手法よりも、検索結果の多様性の観点では、一番優れていると考えられる。表 4 と表 5 の結果から、画像検索モデルは多様性の高い検索件数が提案手法よりも少ないが、画像生成モデルよりは多いことがわかる。また、表 7 の結果から、画像検索モデルの多様性の値の平均値は提案手法より低く、画像生成モデルより高いことがわかる。このことから、画像検索モデルは提案手法ほどではないが、画像生成モデルよりも検索結果の多様性が優れていると考えられる。表 4 と表 6 の結果から、画像生成モデルは、他の 2 つの手法と比較して、多様性の高い検索件数が少ないことがわかる。また、表 7 の結果から、画像生成モデルの多様性の値の平均値が最も低くなっていることがわかる。このことから、画像生成モデルは他の 2 つの手法よりも、検索結果の多様性の観点では、一番劣っていると考えられる。

5 おわりに

5.1 まとめ

本研究の目的は、条件付き画像検索というタスクにおいて、検索結果の検索クエリに対する適合性と多様性を両立する結果を得られることを目指すことであった。本研究では、既存の条件付き画像検索モデルによる検索結果と画像生成モデルにより生成された商品画像を組み合わせる手法を提案した。実験の結果、提案手法が既存の検索モデルよりも検索クエリへの適合性を向上させながら、高い多様性の結果が得られることが示された。したがって、提案手法は、既存の研究と比較して条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人が、本当に欲しい商品を見つけられるような商品画像の結果を出力することが期待できる。

5.2 課題

本研究には次のような課題がある。まず、生成画像の多様性を高める手法の検討が十分ではない点である。本研究では生成画像の多様性を高めるために、検索モデルにより得られた画像と組み合わせたが、画像生成モデル自体の生成画像の多様性を高めるための手法を検討する必要がある。また、検索モデルにより得られた画像と生成画像を組み合わせる手法も、本研究では単純なアルゴリズムであるので、より高度な手法を検討する必要がある。多様性の評価方法に関しても、画像の表示される順番を考慮できていないという問題があるため、より適した評価方法を検討する必要がある。

謝辞

本研究の一部は、JST CREST JPMJCR21D1 の助成を受けたものである。

文献

[1] Liu, Linlin, et al. "Toward AI fashion design: An Attribute-GAN model for clothing match." Neurocomputing 341 (2019): 156-167.

[2] Zameshina, Mariia, Olivier Teytaud, and Laurent Najman. "Diverse Diffusion: Enhancing Image Diversity in Text-to-Image Generation." arXiv preprint arXiv:2310.12583 (2023).

[3] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. Challenges and research opportunities in ecommerce search and recommendations. SIGIR Forum, 2020.

[4] Wu, Hui, et al. "Fashion iq: A new dataset towards retrieving images by natural language feedback." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021.

[5] Baldrati, Alberto, et al. "Effective conditioned and composed image retrieval combining clip-based features." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[7] Lee, Seungmin, Dongwan Kim, and Bohyung Han. "Cosmo: Content-style modulation for image retrieval with text feedback." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[8] Chen, Yiyang, et al. "Composed image retrieval with text feedback via multi-grained uncertainty regularization." arXiv preprint arXiv:2211.07394 (2022).

[9] Choi, Woojin, et al. "Developing an AI-based automated fashion design system: reflecting the work process of fashion designers." Fashion and Textiles 10.1 (2023): 39.

[10] Liu, Zheyuan, et al. "Image retrieval on real-life images with pre-trained vision-and-language models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[11] Yang, Zhengyuan, et al. "The dawn of lmms: Preliminary explorations with gpt-4v (ision)." arXiv preprint arXiv:2309.17421 9.1 (2023): 1.

[12] Li Jing Tim Brooks Jianfeng Wang Linjie Li Long Ouyang Juntang Zhuang Joyce Lee Yufei Guo Wesam Manassra Prafulla Dhariwal Casey Chu James Betker, Gabriel Goh and Yunxin Jiao. Improving image generation with better captions. In https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

[13] Zhang, Mi, and Neil Hurley. "Avoiding monotony: improving the diversity of recommendation lists." Proceedings of the 2008 ACM conference on Recommender systems. 2008.

[14] Ziegler, Cai-Nicolas, et al. "Improving recommendation lists through topic diversification." Proceedings of the 14th international conference on World Wide Web. 2005.

Finding Generative Image LoRA Model by Inputting Style Sample Image

NgocAnh VUTHI†, Yoshiyuki SHOJI†, Huu-Long PHAM††, and Hiroaki OHSHIMA††

† Faculty of Informatics, Shizuoka University
Johoku, Hamamatsu-shi, Shizuoka 432 – 8011, Japan
†† Graduate School of Information Science, University of Hyogo
Gakuen-nishimachi, Nishi-ku, Kobe, Hyogo 651–2197, Japan
E-mail: †vu.thi.ngoc.anh.20@shizuoka.ac.jp, ††shojiy@inf.shizuoka.ac.jp, †††huulongpham28@gmail.com, ††††ohshima@ai.u-hyogo.ac.jp

Abstract This paper proposes a method for finding fine-tuned image-generation LoRA models suitable for generating images that have a similar style to a given sample image. Sharing trained image-generation AI models on the Internet is becoming common. However, to determine which of these models can generate the image style the user wants, it is necessary to inspect the output samples posted on the site one by one. To enable search for the model that can generate an image that has a similar style to a given image, we adopt a CNN-based multi-class classifier. The ResNet50 model was fine-tuned to classify the distinctive features of each LoRA model. We transformed pre-prepared sample images with each LoRA model using img2img as training data. Experimental results show that our customized ResNet50 can find more correct LoRA models than the original ResNet50. Future research should explore more effective methods to address the unique challenges presented in this novel approach.

Key words Finding model, Stable Diffusion model, LoRA model, CNN.

1 Introduction

Even though it has been only a short time since image-generative AI was released, people have already begun to open and share their fine-tuned models. Stable Diffusion, one of the most famous image-generative AI models, was released in 2022. This model generates images by processing input texts (*i. e.,* prompts), or a sample image. While the base Stable Diffusion model excels in various tasks, it encounters challenges in generating specific styles.

Rather than adjusting the prompt to generate specific styles, a more practical approach involves utilizing custom models that are fine-tuned models with images of the target sub-genre. These customized models, known as checkpoint models and LoRA models, offer enhanced performance. This paper focuses on custom LoRA models, demonstrating proficiency in generating specialized style images. Within one year, file-tuned LoRA models of sharing became common.

Specifically, on a website called "Civitai(注1)", many image-generative file-tuned LoRA models are shared. From anime to realistic style, they are fine-tuned with domain images and uploaded there. Users can freely download models and generate images of what they want.

Despite the proliferation of such models, subtle variations exist among those with similar styles, complicating user attempts to find a LoRA model that closely aligns with their desired output. When users want to find models that can be generated to style what they wish, they need to check the thumbnail images of each model to know whether this model can be generated to style what they want. This makes users waste a lot of time.

If thumbnail images of the model exist, the model can be found by the above methods. This method depends on existing thumbnail images. Also, adjusting hyperparameters (strength, guidance scale, etc.) of fine-tuned models when generating images can generate different style images, although using the same fine-tuned model. So that that model can generate the style the users want.

As described above, the fine-tuned image generation model can generate various styles of images by adjusting the hyperparameters and prompts. This means that even if a fine-tuned image generation model does not exist prior, if a model generated close to the style image that the user wants can be retrieved, an image in the desired style can be generated.

This paper proposes a method for obtaining an image-generative LoRA model using a single-style sample image. For example, when a user inputs a single sample image of Van Gogh, the system outputs models that can generate a painting style that is the same or close to the input image.

(注1)：Civitai: The Home of Open-Source Generative AI
https://civitai.com/

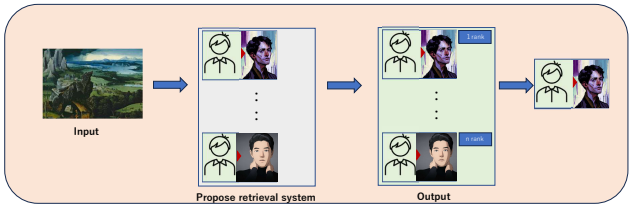


Figure 1 Overview of the retrieval system

Since multiple models may be output, model ranking is performed. The model that generates the closest match to the style of the input image is placed at the top of the list.

Figure 1 shows an overview of our system. It accepts a style image as its input, and outputs the ranking of models order by the possibility of generating a given image. To rank the models, we used a ResNet50-based classifier.

Our research aims to help users easily find fine-tuned image generative models without checking the thumbnail images of LoRA models. Moreover, it supports users in finding the alternative image generative model when the model they really want does not exist.

This system is necessary not only for users who want to generate an image, but also for artists and copyright holders. Artists and copyright holders may have to check if their original artworks were illegally used to train the model without agreement. When fine-tuning Stable Diffusion with specific domain images, it is necessary to use a lot of images in this domain. A fined-tuned model can only generate the style images used during fine-tuning. If the user is an artist with many well-known works, the painter may suspect their works are being used illegally. By using this system, the user can check for such illegal use. By inputting an image into the system, the system outputs a model that can generate a painting style similar to that of the image.

To find image-generative fine-tuned models, we used pre-trained ResNet50. To adapt pre-trained ResNet50 in our proposal, ResNet50 is fine-tuned with image classification tasks.

To create an image for fine-tuning ResNet 50, Stable Diffusion version 1.5, ControlNet - Canny, and LoRA models are used. Further, to boost ResNet50's performance with the given task, a small part of the original ResNet50 architecture is modified. By fine-tuning the pre-trained Resnet50 with the style-transformed image classification task, the fine-tuned Resnet50 obtains higher accuracy than the pre-trained ResNet50. Macro-recall, Macro-precision, and Macro-F1 scores are used to evaluate the image classification task.

Our paper is structured as follows. In this section, we describe the background and objectives of this study. Section 2 describes related studies and the position of these studies in related fields. Section 3 describes the methodology proposed in this study, and section 4 describes the evaluation method of the proposed methodology. Section 5 discusses the results obtained through experiments. Finally, we summarize this study and discuss future prospects.

2 Related Work

This study aims to help users find image-generative fine-tuned LoRA models more easily by inputting a sample image. This study relates to the image-generative AI field. Also, this study relates to the classification of images and related to the content-based image retrieval field.

2.1 Image Generative AI

In recent years, the field of artificial intelligence has witnessed remarkable advancements, especially in the domain of computer vision. One of the most intriguing developments is the emergence of generative AI image models, which have revolutionized how we perceive and interact with visual data. These sophisticated algorithms are at the forefront of the AI revolution, enabling machines to generate realistic and creative images.

First, we explain the concept of generative AI image models and their significance in the world of generative AI development. Generative AI refers to a class of artificial intelligence models that can generate data resembling authentic examples from the dataset they were trained on. In the context of images, generative AI aims to create new images that resemble the distribution of images in the training dataset, exhibiting creativity and imagination. The stable Diffusion model is a type of generative AI. It is used not only to generate images, but also to generate training data for supervised machine learning [1].

Model Agnostic Zero-Shot Classification is a method to classify real images by learning with synthetic images. Using the Stable Diffusion model improves the quality of the training dataset and solves problems relating to large-scale vision-language models [2].

Class incremental learning aims to learn new classes without forgetting previously learned classes incrementally. Several research works have shown how additional data can be used by incremental models to help mitigate catastrophic forgetting. A Stable Diffusion model [3]can generate synthetic samples belonging to the same classes as the previously encountered images.

Computer-aided surgical systems can provide surgeons with supportive information to improve procedure execution and overall outcomes. The Stable Diffusion model [4] of image-to-image task can reduce the gap between synthetic and real data. This makes computer-assisted surgical systems more accurately annotate data.

2.2 Image Classification by CNNs

CNN-based architecture is used for object detection and pattern recognition. Object detection is highly attended to in computer vision, but classification image style is not so attended to. Hence, there are few prior works that use CNNs to classify image styles.

CNN-based VGG-19 is adopted to extract object features and texture features [5]. Using a deep convolutional net to extract image features [6]. Although models were trained for object detection, these models can be used to classify aesthetics and styles.

CNNs are adapted to identify artists [7]. It has been discovered that CNNs can learn illustration styles from images. An ensemble of convolutional neural networks is adapted to learn the surface of image data, utilizing a CNN-based multiple neural network architecture [8]. The output of the CNN allows for the prediction of attribute images.

2.3 Content-Based Image Retrieval

In recent years, content-based image retrieval has become common in a variety of fields, such as product search and similar image retrieval. With the development of e-commerce sites, customers can search for and purchase just about anything on e-commerce. However, too much information hinders the search for information. Images [9] can be used to search for information easily.

Content-based image retrieval (CBIR) is a widely used technique for retrieving images from huge and unlabeled image databases. To help users more easily retrieve images [10], deep learning frameworks based on Convolutional Neural Networks (CNN) for feature extraction and Support Vector Machine (SVM) for classification are adopted. In this paper, instead of using SVM, pre-trained Resnet50 of the fully connected layer is used for classification. Fully connected layers of weights are initialized with ImageNet weights that obtain better results.

By retraining the CNN models with classification or similarity learning objective on the new domain, [11] found that the retrieval performance could be boosted significantly, which is much better than the improvements made by similarity learning.

Jin *et al.* [12] propose a method for searching images that share the same spatial semantics and enjoy visual consistency as the query image in complex scenes. They adopted CNNs to extract image features.

Adopting deep learning architecture to extract content from query images is common in the content-based image retrieval field. Instead of using text to search content, using images offers more information for the retrieval system, introducing the closest results that users expect.

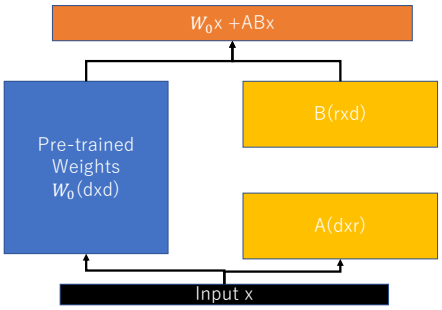


Figure 2 Low Rank Adaptation

3 Classification to Identify Model Capable to Generate Exemplified Styled Image

This section describes the proposed method. We explain the method in two sections: The image generation process, and Resnet50 fine-tuning task.

3.1 Image Generation Models

This sub-section describes the image generation process by fine-tuning image-generative AI models. Latent stable diffusion models [13], ControlNet-Canny [14], and Stable Diffusion fine-tuned by LoRA [15]are implemented. Latent stable diffusion models are exploited based on diffusion models. By training on latent space and introducing cross-attention layers into diffusion models. Latent stable diffusion models are more powerful and flexible generators for general conditioning inputs such as prompt or bounding boxes and high-resolution synthesis.

When enabling a pre-trained large model into the specific task, it is necessary to add a task-specific head or update the weights of the pre-trained large model through backpropagation during the training process. This process demands a lot of GPU memory. If small-scale data is used for full fine-tuning, the fine-tuned model can be catastrophic forgetting. This caused harmful to infer phase.

In Figure 2, instead of fine-tuning all the weights of the base model, the LoRA (low-rank adaptation) method decomposes the weights of the base model into matrix A and matrix B. The new A and B matrix weights are only updated during retraining; the weights of the base model are fixed. The LoRA method is being introduced for the first time through the fine-tuning of a large-scale language model. However, it can be applied to any dense layers in deep learning models. When fine-tuning Stable Diffusion models, the cross-attention layers in Stable Diffusion models are trained by LoRA. The cross-attention layers are the part of the model where the image and the prompt meet.

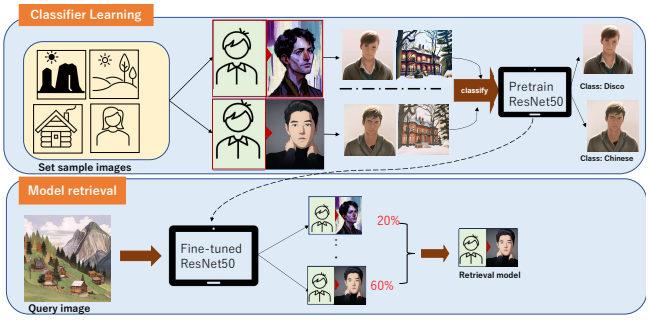


Figure 3 Image Classification and Model Retrieval Process

Using specific domain images and adopting LoRA for fine-tuning, fined-tuned Stable Diffusion models with specific styles are created; we call it the xx LoRA model. When inferring or generating, the weight of the fine-tuned LoRA model and the weight of the base model (Stable Diffusion model) are merged to implement the task. For instance, the anime LoRA model, the Chinese Ink LoRA model, etc.

With the advent of text-to-image Stable Diffusion models [13], we can create visually stunning images by inputting prompts. This paper proposes a method to classify the LoRA model style through a generated image. For this reason, image-to-image Stable Diffusion models are adopted to transfer base images into LoRA-style images. However, generating new images by image-to-image task without a prompt is still challenging.

The reason is related to the diffusion model mechanism, which uses image reconstruction. First, diffusion models noise input images by Markov chain, then reconstructing the image from the noised image via the denoising process. This means diffusion models can not retain anything about inputted images. This causes generating images through image-to-image to be stuck. ControlNet is adapted to address this challenge in the image-to-image task. The neural network architecture of ControlNet is described in Figure 4. We adopted ControlNet with Stable Diffusion model version 1.5 as the base model.

The ControlNet architecture based on the Stable Diffusion model version 1.5 works as below. ControlNet copies 12 encoder blocks and one middle block weight of the Stable Diffusion model and uses them as trainable parameters. As illustrated in Figure 4,

- x is the input features map,
- c is the internal vector calculated based on x, ControlNet of input,
- y is the output feature map.

The trainable parameters are connected to the base model with zero convolution layers, denoted $Z(\cdot;\cdot)$. Specifically, $Z(\cdot;\cdot)$ is a 1 x 1 convolution layer with both weight and bias initialized to zeros. To build up a ControlNet, two instances

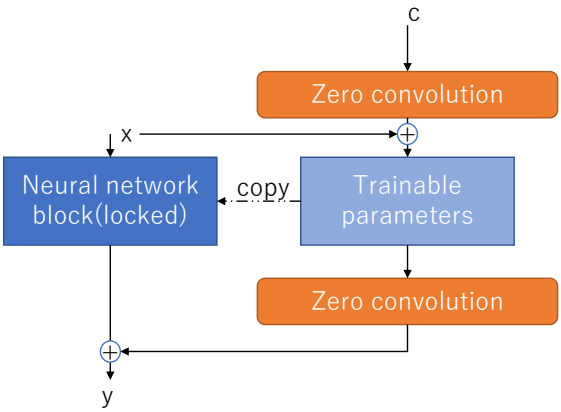


Figure 4 ControlNet neural network

of zero convolutions with parameters Θ z1 and Θ z2, respectively, are used. Although convolution layers of weight and bias are initialized to zeros, ControlNet still learns and updates parameters. We demonstrate the following formula:

$$y = Wx + b, \tag{1}$$

$$\frac{\partial\, y}{\partial\, w} = x, \tag{2}$$

$$\frac{\partial\, y}{\partial\, x} = W, \tag{3}$$

$$\frac{\partial\, y}{\partial\, b} = 1, \tag{4}$$

$$\tag{5}$$

where y denotes output, W denotes weight, x denotes input, b denotes bias. If W and b are zero and x is not zero, y is different from zero. This means that when images are input, the weight and bias of ControlNet are grown up. When training ControlNet, the weight and bias of the stable diffusion model are locked, and only trainable weights are updated by gradient descent. ControlNet offers various conditional controls in the Stable Diffusion model. In our experiment, we used the ControlNet-Canny condition. Via ControlNet-Canny conditional control, we can extract the outline of the subject in each base image, then transfer object style with the Stable Diffusion model and LoRA models.

3.2 Classifier Using Fine-tuned Resnet50

In the context of the classification problem, shallow layers would [16] extract low-level features that are almost the same as the input image. In deeper layers, more abstract features were extracted. The abstract features make the classification possible and easier. Hence, more layers are added to network architecture in the hope that the network can learn more features from input images and boost classification accuracy. However, when more layers are adapted, a degradation problem is caused. This problem makes training errors and test/validation errors go up [17]. By introducing a residual learning framework into CNN, Resnet50 [17] is easier to

optimize and can gain higher accuracy. The deep residual learning framework of ResNet50 is shown in Figure 5. As shown in Figure 5, the x identity map feature is added to network weight before activating network weight forward. Instead of output $F(x)$, the output changes into $F(x) + x$. This makes deep layers learn shallow layers of features and decreases errors in training and test/validation.

In this paper, Resnet50 is adopted to classify LoRA style via transferred images generated from fine-tuned image generative LoRA models. To enhance classification accuracy, pre-train Resnet50 is modified and fine-tuned.

3.3 Classification-based Model Search

In this section, we explain more clearly about our method. First, we collected sample images. The sample image denotes which image we use to transform its style. Sample images are variously selected. Next, we prepared the fine-tuned image generative LoRA models. As mentioned in the previous section, to transfer the style of base images, image-to-image Stable Diffusion model version 1.5, ControlNet canny model, and LoRA models are adopted. At first, the subject in each base image of the outline is extracted and transformed into a vector. We call it a conditional control vector. Then, the Stable Diffusion model version 1.5 and the Canny model of ControlNet are implemented. In the stage of applying the LoRA models style, weight in u-net layers and text encoder layers of LoRA [15] are set with 1. Hence, enabling the LoRAs feature is clearly reflected in the transferred image. The base image and conditional control vector are used when transforming the base image style without any text input. In addition, hyperparameters such as inference steps, guidance scales, etc... are adjusted so that the transformed image better reflects LoRA features.

After transforming the style of base images, the fine-tuning phase is implemented. We fine-tune pre-trained ResNet50 with image classification tasks. Data is transformed style base images, and the label is the name of LoRA models. In our experiment, the training dataset is very tiny compared with the ImageNet dataset used for training ResNet50 [17]. For this reason, only small parts of pre-trained ResNet50 are returned, instead of returning the entire model. As mentioned in [18], deeper CNN layers can extract more abstract features of the image. These features represent a unique style of image. ResNet50 architecture has four major residual layers, which are assigned to extract the abstract features of each image. Moreover, each layer has a different amount of bottleneck [17]. The third layer of Resnet50 has the most number of bottlenecks, with six bottlenecks. Based on the above mentioned, weights between layer 3 and the last fully connected layer are set backpropagation. The primary role of the CNNs is feature extraction. Our main purpose is image classification, thus more emphasis on fully connected layers. One more fully connected layer is added.

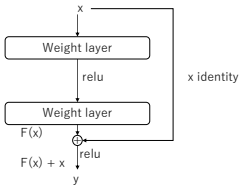


Figure 5 Residual block

4 Evaluation

We evaluate the proposed retrieval system through three factors. First, we tried to clarify whether the designed fine-tuning is effective or not. Second, we evaluated what kind of dataset can be used to obtain good performance for the proposed system. Finally, we evaluated how many style-transformed images should be included in each training data class to ensure sufficient learning.

4.1 Fine-tuning result

This section introduces the evaluation method for the proposed method. The base images are downloaded from the Internet. We call this the base image set as set B. The set B is processed to include as many different types of images as possible. Images of people, scenery, food, plants, animals, etc., are included in the set B. Each element in the set B is unique. The set B is transformed style by group models, which are the stable diffusion, canny, and LoRA models. With each image in the B set creates one image in the F set. After transforming the style, we create a transformed style image set called the F set. The size of B set and F set are the same, both set include 126 images. Such as, $(\forall x \in B | \exists x' \in F, |B| = |F|)$, where

- x : base image,
- B : base image set,
- x': transformed style image, and
- F: transformed style image set.

In the fine-tuning phase, 80 percent of the F set is trained, and 20 percent of the set is used for validation.

Table 1 shows the list of models that we used in our experiment. As the search target, 12 LoRA models are prepared. All these LoRa models are downloaded from Civitai. In the experiment, 15 epochs and a 0.001 learning rate are used. The number of nodes in the output of the fully connected layer is modified to match the number of LoRA models we wish to classify.

When returning, early stopping with five patients and the learning schedule CosineAnnealingLR are utilized to obtain the best performance fine-tuned ResNet50 model.

As shown in Figure 6, training and validating results are

Table 1 Models used in experiment

Model name
Anime Screencap Style LoRA (注2)
Anime Game Backgrounds(注3)
xiaorenshu(注4)
Chinese ink painting(注5)
Disco Elysium / Aleksander Rostov / Oil paints & Abstraction / Style(注6)
Fairy in Clouds(注7)
XSarchitectural-11Fantasyarchitecture(注8)
Phantasmal Luminous: The Radiance of Rainbow Dispersion(注9)
8bitdiffuser 64x — a perfect pixel art model(注10)
pop art(注11)
Texture illustration(注12)
Realistic Dating attire(注13)

plotted. When fine-tuning pre-trained ResNet50 with 12 LoRA models, the best result was obtained in the 6th epoch. In the loss plot, the learning loss decreases gradually as the number of learning epochs increases. Nevertheless, the validation loss decreases from the first to the third epoch, but becomes very unstable from the fourth epoch to the end. The same phenomenon is observed in the accuracy plot.

Then, to evaluate the effect of fine-tuning, we compare the classification performance of the fine-tuned ResNet50 with the classification results of the pre-trained ResNet50. Here, we use Accuracy and Macro-F1. In each experiment, shown in 2, the same set of validation data is used. Also, the number(100, 26), in second and third column, denoting amount of images included in each class.

With pre-trained ResNet50, the output of a fully connected layer is modified from 1000 to 12 to adapt with our task. When inferring, weights of pre-trained ResNet50 are frozen. The detailed result is shown in 2. The our ResNet50 performs better than the pre-trained ResNet50 in the three measures of Accuracy, Macro-F1 and MRR. We calculated the Macro-F_i score as

$$Macro - F1 = \frac{1}{N}\sum_{i=0}^{N}(\frac{2TPi}{2TPi + FPi + FNi}) \qquad (6)$$

where

- N: Amount of models(N = 12),
- TP: True positive,
- FN: False negative, and
- FP: False positive.

Via Accuracy and Macro-F1, we can evaluate classification performance of ResNet50. Via MRR, we can evaluate capability retrieval model of proposed system.

In addition, the more images included in each class, the better the system's learning performance. However, once a certain amount is reached, the change becomes small. Specifically, even if 70 or more images are included in each class, no significant change in learning accuracy occurs. In conclusion, 70 images in each class are sufficient to train the proposed retrieval system.

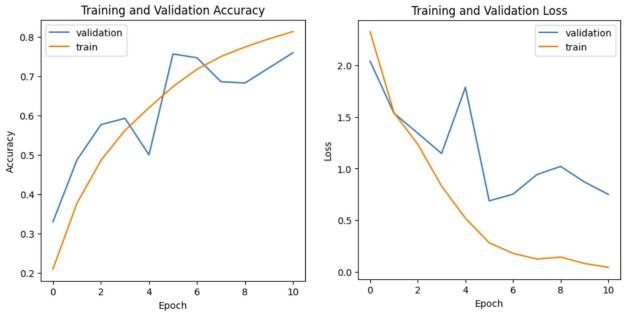


Figure 6 12 models of classification result

4.2 Experiment Model Retrieval

We prepare more than one sample image dataset to evaluate the importance of training data variety. In the above dataset, we call it a variety dataset, which includes a variety of images such as humans, animals, landscapes, *etc.*. The new one, we call it a specialized domain image dataset that contains only human face images. Each image in the domain image dataset is processed like an image in the variety dataset.

The result is shown in table 2 Fine-tuning the pre-trained ResNet50 using the same number of data, the retrieval system outperforms when using diverse data. In conclusion, diverse data images are important for building the proposed retrieval system.

A fine-tuned ResNet50 model with 12 models is also used to find LoRA models. Found models are ranked based on the output of the second fully connected layer of ResNet50. Images are used in the experiment, not included in the training or validating sets.

To enable us to evaluate objectively, these images are downloaded from the site called Civitai, and they are thumbnail images of the original LoRA models created by authors.

Table 2 Evaluation Results

	Train data	Validation data	Accuracy	Macro-F1	MRR
Fine-tuned ResNet50	100 variety images	26 variety images	79.50	74.37	60.92
Pretrained ResNet50		26 variety images	8.30	Nan	
Fine-tuned ResNet50	70 variety images	26 variety images	76.90	70.17	62.15
Fine-tuned ResNet50	50 variety images	26 variety images	67.90	65.50	55.09
Fine-tuned ResNet50	30 variety images	26 variety images	62.20	60.48	60.42
Fine-tuned ResNet50	70 human face images	26 variety images	50.30	48.33	47.92

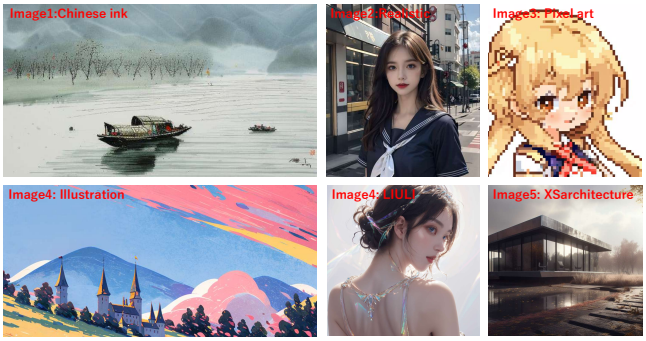


Figure 7 Experimental images

Figure 8 Image generation sample

Experimental images are shown in Figure 7. The red letters in the top left of each image are the names and correct labels of each image. A correct label indicates that the image is generated from that LoRA model.

Finding model results by a sample image is shown in Table 3. Via our method, among the six sample images, three sample images can accurately retrieve the original LoRA model. When the LoRA model is retrieved by feeding the experimental images into the proposed system, the correct results are included within the top six.

4.3 Image generation

This section uses the retrieved LoRA models from the previous section to generate images. Following the result of Table 2, the search result is the most undesirable when using image 2 to find the LoRA model that generated image 2. The desirable search result is located in the sixth position. In the experiment of image generation, the first and sixth models are used to generate images and compare them. Figure 9 shows the four images in our experiment, such as

Depending on each LoRA model, the generated image does not exactly resemble the base image. However, they were converted to their respective painting styles. Indeed, the style of the image transformed by the Realistic LoRA model is more similar to the style of the reference style image than the style of the image transformed by the XSarchitecture LoRA model. However, when users want to generate images, they want the generated images to follow their desired style, and the generated images are beautiful. In this experiment, in the two generated images, we are not able to say that the image transformed by the Realistic LoRA model is more aesthetic than the image transformed by the XSarchitecture LoRA model.

5 Disccusion

Through this experiment, it is certainly possible for the proposed method to search for fine-tuned image-generative LoRA models. However, to retrieve the LoRA model accurately, the number of LoRA models fed in ResNet50 is limited. Furthermore, since the models are retrieved through style-transformed images, amount of models that can be retrieved may depend on the process of image style transformation. There are many hyper-parameters that need adjusting during the process of transforming the style of the image. Thus, when generating more images, the image generation process is more complicated and not sturdy.

Moreover, the number of fine-tuned LoRA models gradually increases, and in this proposal, ResNet50 will need to return when new fine-tuned LoRA models are released. This is inefficient. ResNet50 is specialized to extract image style features. Although the original ResNet50 is modified for best performance on a given task, its functionality is still limited when a number of LoRA models are added to ResNet50. We have confirmed that the capability of ResNet50 is not appropriate for our research. We need to investigate more effective methods in the future.

One of effective method is the Siamese Network, employing a loss function called contractive loss or triplet loss during

Table 3 Result of retrieving model

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Image1	**Chinese ink**	XSarchitecture	Illustration	Anime	Pop art	Realistic	Pixel art	LIULI	Cloudy
Image2	XSarchitecture	Anime	cloudy	Illustration	Chinese ink	**Realistic**	Pixel art	LIULI	Anime background
Image3	**Pixel art**	Pop art	Illustration	Anime background	Realistic	Chinese style	XSarchitecture	Anime	LIULI
Image4	LIULI	**Illustration**	Anime background	Cloudy	Pop art	Disco	XSarchitecture	Anime	Realistic
Image5	Cloudy	XSarchitecture	Illustration	**LIULI**	Anime	Pop art	Anime background	Realistic	Chinese ink
Image6	**XSarchitecture**	Cloudy	Anime	Illustration	Chinese ink	Realistic	Pixel art	LIULI	Pop art

training. The Siamese Network takes two or three images as input values and calculates the distance between them. Based on distance, determine if they are the same image or not. When applying to our research, we assume that one image is the style the user wants and the other is a thumbnail image of the original LoRA model. The above process makes it possible to eliminate the work of image style transformation.

6 Conclusion

This paper proposes a method for retrieving fine-tuned image-generating LoRA models. Through our method, it is possible to output a LoRA model that can generate an image style similar to the style of the reference image. A dataset of various base images is prepared to implement the proposed method. The style of the sample images is transformed by using Stable Diffusion version 1.5, ControlNet-Canny, and fine-tuned image generation LoRA models. The style-transformed images are used as fine-tuning data for the pre-trained ResNet50. Several parts of the original ResNet50 architecture are modified to adapt it to our task. Experiments are conducted to compare pre-trained ResNet50 and fine-tuned ResNet50 and search fine-tuned image generation LoRA models.

Acknowledgement

This work was supported by JSPS KAKENHI Grants Numbers 21H03775, 21H03774, and 22H03905.

References

[1] Gabriele Valvano, Antonino Agostino, Giovanni De Magistris, Antonino Graziano, and Giacomo Veneri. Controllable image synthesis of industrial data using stable diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5354–5363, 2024.

[2] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 769–778, 2023.

[3] Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Class-incremental learning using diffusion model for distillation and replay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3425–3433, 2023.

[4] Joanna Kaleta, DiegoDall ' Alba, Szymon Płotka, Przemysław Korzeniowski. Minimal data requirement for realistic endoscopic image generation with stable diffusion. *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2023.

[5] Tiancheng Sun, Yulong Wang, Jian Yang, and Xiaolin Hu. Convolution neural networks with two pathways for image style recognition. *IEEE Transactions on Image Processing*, Vol. 26, No. 9, pp. 4102–4113, 2017.

[6] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *Proc. Brit. Mach. Vis.Conf. (BMVC)*, 2014.

[7] Kazuma Kondo and Tatsuhito Hasegawa. Cnn-based criteria for classifying artists by illustration style. IVSP '20, p. 93–98, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Fang Ji, Michael S McMaster, Samuel Schwab, Gundeep Singh, Lauryn Nicole Smith, Shishir Adhikari, MarcioO ' Dwyer, Farah Sayed, Antony Ingrisano, Dean Yoder ほか. Discerning the painter ' s hand: machine learning on surface topography. *Heritage Science*, Vol. 9, No. 1, pp. 1–11, 2021.

[9] Farhan Ullah, Bofeng Zhang, and Rehan Ullah Khan. Image-based service recommendation system: A jpeg-coefficient rfs approach. *IEEE Access*, Vol. 8, pp. 3308–3318, 2020.

[10] Ouhda Mohamed, El Asnaoui Khalid, Ouanan Mohammed, and Aksasse Brahim. Content-based image retrieval using convolutional neural networks. In *Lecture Notes in Real-Time Intelligent Systems*, pp. 463–476. Springer, 2019.

[11] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, p. 157–166, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, and Yaochen Li. Spatial-content image search in complex scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2503–2511, 2020.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[16] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[18] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

(注2)：civitai: https://civitai.com/api/download/models/60568?type=Model&format=SafeTensor
(注3)：civitai: https://civitai.com/api/download/models/19065?type=Model&format=SafeTensor&size=full&fp=fp16
(注4)：civitai: https://civitai.com/api/download/models/25661?type=Model&format=SafeTensor&size=full&fp=fp16
(注5)：civitai: https://civitai.com/api/download/models/42314?type=Model&format=SafeTensor
(注6)：civitai: https://civitai.com/api/download/models/169568?type=Model&format=SafeTensor
(注7)：civitai: https://civitai.com/api/download/models/269819?type=Model&format=SafeTensor
(注8)：civitai: https://civitai.com/api/download/models/30572
(注9)：civitai: https://civitai.com/api/download/models/180175
(注10)：civitai: https://civitai.com/api/download/models/231819?type=Model&format=SafeTensor
(注11)：civitai: https://civitai.com/api/download/models/188417?type=Model&format=SafeTensor
(注12)：civitai: https://civitai.com/api/download/models/113534?type=Model&format=SafeTensor
(注13)：civitai: https://civitai.com/api/download/models/228470?type=Model&format=SafeTensor

生成型要約に基づくWebページのサムネイル生成

前田　直宏†　　山本　岳洋†

† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †ad22p062@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では Text2Image モデルを用いて Web ページの主題を表したサムネイルを生成する手法を提案する．具体的には，モデルに入力するプロンプトを Web ページを要約することで複数生成し，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点から適したプロンプトを分類して選択する．この条件を満たすプロンプトを取得するために本研究では，ランダムフォレストを用いて分類モデルを構築する．実験の結果，自然言語処理タスクで高い性能を示す BERT よりも高い精度で分類を行うことができた．また，構築した分類モデルを用いて CNN ニュース記事と旅行ブログなどの Web ページを対象としたサムネイル生成を行い，人手による評価を行う．サムネイル評価実験の結果，記事のタイトルをプロンプトとするベースラインに比べ，提案手法によって生成されるサムネイルの方が記事のサムネイルに合っているとして好まれた．

キーワード　情報検索，Text2Image，プロンプトエンジニアリング，画像生成，要約文生成

1 はじめに

Web ページの概要を表すサムネイルは，ユーザ所望の情報を獲得する助けになる．例えば検索結果の表示をテキストベースだけではなく，Web ページ内の画像やデザインを用いた 1 枚のサムネイルも表示することで，探している情報と関係のある情報をより正確に獲得することが可能である [4]．また，一度閲覧した Web ページの約 80% は再度閲覧することが明らかになっている [1] [3]．ブラウザに備わったブックマーク機能や閲覧履歴，記憶を辿ることで再度閲覧したページにたどり着くことができる一方，検索結果表示欄にサムネイルも表示することで，一度訪れたサイトへの再訪問が容易に行えることが報告されている [22]．このようにユーザが情報を獲得するとき，Web ページの概要を表したサムネイルは情報検索の効率を上昇させる．

Web ページの内容を表したサムネイルの生成方法としては，Web ページの上部のスクリーンショットにロゴを付与するものや，ページ内に掲載されている写真を組み合わせたもの，インターネット上からその Web ページと関連性が高い画像を用いたものなど様々なサムネイルが提案されている [9] [15] [22]．また近年の検索エンジンでは，Web ページの上部に掲載されているアイキャッチ画像とも呼ばれる画像をサムネイル代わりとして使用されることもある．このアイキャッチ画像はフリー素材サイトなどで取得する必要があるが，必ずしも記事の内容に合った画像が存在するわけではない．また内容と異なる画像を用いた場合，誤った情報の獲得を促してしまう可能性がある．そのため，Web 検索におけるサムネイルは記事の内容や主題に合っていることが重要である．

そこで本研究では Text2Image モデルを用いて，Web ページの主題を表したサムネイルの生成を行う．Text2Image モデルによる生成画像は，入力されたプロンプトと呼ばれるテキストに基づいた画像が生成される．そのため本研究では Web ページの記事テキストをもとにプロンプトを生成し，得られたプロンプトを用いてサムネイルの生成を行う．具体的には，まず記事テキストの要約文を複数生成する．その後，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点からサムネイル生成に適したプロンプトを分類して取得する．このプロンプト分類モデルとして本研究ではランダムフォレストを構築する．そして得られたプロンプトを Text2Image モデルに入力することでサムネイルを生成する．

2 関連研究

2.1 サムネイルを用いた情報検索支援

Web ページのサムネイルを表したサムネイルの研究は多く行われている．Zhou らは Web ページを検索する時に使用したクエリを強調するサムネイルを提案している [23]．Teevan らは Web ページのタイトル，ロゴ画像，注目度が高い画像を用いたサムネイルを提案している [22]．また，稲垣らは Web ページに対する顕著性マップを用いたサムネイルを提案している [25]．顕著性マップとは，人が画像を認識する際に注視しやすい領域を画像で定量的に表現したものである．Web ページを閲覧する際に注視されやすい上位 10 の領域を 1 枚の画像に並べて配置した集約画像を生成することで，ページ内容の理解補助に繋げている．これらのサムネイルで使用している画像は，Web ページ内に含まれる画像を抽出し使用している．しかし内部画像が存在しない場合，サムネイルを作成することが困難であった．そこで Jiao らはインターネット上からその Web ページと関連性が高い画像を取得し，その画像を用いたサムネイルを提案している [9]．Web 検索時以外にも Web ページの概要を表したサムネイルの利用方法が提案されている．例えばサムネイルを Web ブラウザのタブ管理に使用することが提案されている [12]．Web ページの左右余白部分に閲覧した Web ページのサムネイルを配置することで，一度閲覧した Web ページへの

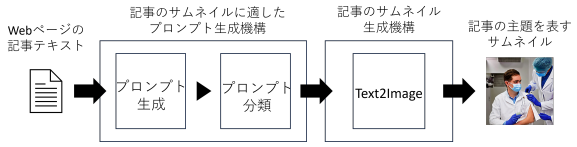


図 1　記事の主題を表したサムネイルの生成手法.

素早い切り替えを行うことができる.

2.2　画像生成モデル

画像生成モデルとしては，VAE（Variational AutoEncoder）[10] や GAN（Generative Adversarial Network）[6]，拡散モデル [8] などの手法が数多く提案されていた．VAE はエンコーダとデコーダで構成されており，エンコーダで入力データを潜在変数に落とし込み，デコーダで潜在変数から画像を生成する．GAN は敵対的生成ネットワークとも呼ばれ，生成器と識別器からなる 2 つのネットワーク構造を持つ．生成器が偽のデータを生成し，識別器が正解データと偽データを見分ける学習を行うことで生成器が本物に近く質の高い画像を生成することができる．拡散モデルは画像にノイズを付与する拡散過程と，ノイズを取り除いて画像を生成する逆拡散過程に分かれており，高い質の画像を生成することができる．また近年では，テキストから画像を生成する Text2Image モデルが多く提案されており，Stable Diffusion [20] や DALL-E2 [18] など質の高い画像を生成することができる．これらの画像生成するモデルでは，プロンプトと呼ばれるテキストを入力する必要があり，プロンプトの質が生成画像の質に大きく影響する．そのため，プロンプトの質を高めるプロンプトエンジニアリングの手法も多く提案されている.

2.3　サムネイル生成とプロンプトエンジニアリング

Text2Image モデルの登場により，テキストからサムネイル生成を行う研究が行われてきている．Liu らはニュース記事の挿絵を生成した [14]．Web 検索以外においてもサムネイル生成の研究が行われている．Shinomoto らは，ニュース記事に対してミーム画像を生成する手法を提案している [21]．また，プロンプトエンジニアリングによるサムネイル生成を行った研究もおこなわれている．Hao らは強化学習なども用いてプロンプトの最適化を行った上で画像の生成を行っている [7]．Liu らは，プロンプトの様々な条件のプロンプトで画像を生成することによってプロンプトと生成画像の関係を調査している [13].

3　提 案 手 法

本研究では Web ページの記事から，その記事の主題を表したサムネイルを生成する．提案手法の全体像を図 1 に示す．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．この Text2Image モデルには出力したい画像に対するプロンプトを入力する必要がある．そこで記事本文を要約することで Text2Image モデルに入力するプロンプトを複数生成する．複数生成されたプロンプトの内，サムネイル生成に適したプロンプトを分類し，最も適したプロンプトを Text2Image モデルに入力することで記事の主題を表したサムネイルを生成する．次節では Text2Image モデルに入力するプロンプトを生成するプロンプト生成機構と，記事のサムネイル生成機構の 2 つについて説明する.

3.1　記事本文の要約によるプロンプト生成

サムネイル生成を行うために必要であるプロンプトの生成について説明する．Web ページの記事本文には多くの情報が含まれているため，Text2Image モデルに全ての文を入力することによって記事の主題を表した画像を生成することは難しい．そこで本研究ではタイトルを除いた記事本文を T5 [17] を用いて要約することで，記事の要点を短くまとめたプロンプトを生成する．しかし，Web ページの記事は章立てられていることが多く要点が多く存在すると考えられる．そのため，全文をまとめて要約するのではなく全文をいくつかの文章に分け，各文章を要約することでプロンプトを生成する．本文を区切って入力する理由としては 2 つある．1 つ目は T5 の入力トークン数には上限があり，本文全文を入力することが困難な場合があるからである．2 つ目は Web ページの文章は章立てられていることが多く，文の位置によって述べられているトピックが異なるからである．あまりにも多くのトピックが含まれた文章の要約を行うと情報量が減少してしまう恐れがあるため，数文に区切って入力する.

記事本文 article を文 s_i の系列として以下のように表現する.

$$\text{article} = s_1 s_2 \dots s_n \tag{1}$$

ここで n は記事本文 article の文数を表す．ただし，1 文はピリオド，エクスクラメーションマーク，クエスチョンマークで区切る．また，記事本文から系列の一部を取り出す関数である窓関数を以下に定義する.

$$\text{window}(\text{article}, i, j) = s_i s_{i+1} \dots s_{i+j-1} \tag{2}$$

この窓関数を用いて j 文で構成される文章 sentence_i を以下に定義する.

$$\text{sentence}_\text{i} = \text{window}(\text{article}, i, j) \tag{3}$$

ただし，i は取り出したい部分の開始位置を表す.

こうして得られた各文章 sentence_i を要約することでプロンプトを生成する.

3.2　サムネイル生成に適したプロンプトの分類

3.2.1　問 題 定 義

Text2Image モデルによって生成される画像の質は，入力するプロンプトの質に大きく依存する．そのため，画像で表現することが難しいプロンプトを入力すると質の低い画像が生成される．例えば「方法」や「安い」といった単語を画像で表現することは難しい．そのため入力として適するプロンプトを選ぶ必要がある．また，3.1 節で生成された複数のプロンプト内には，記事の主題を含んでいるプロンプトと含んでいないものが

表 1　ランダムフォレストで使用する特徴量.

特徴量変数名	説明
interrogative_words	疑問詞を含むか
word_length	単語数
title_prompt_sim	タイトルとプロンプトの類似度
LexRank	LexRank
adjective_count	形容詞の数
noun_count	名詞の数
proper_noun_count	固有名詞の数
verb_count	動詞の数
proportion_adjective	形容詞の割合
proportion_noun	名詞の割合
proportion_proper_noun	固有名詞の割合
proportion_verb	動詞の割合
concreteness	各単語の具体度の平均

あると考えられる．本研究の目的は記事の主題を表したサムネイルの生成であるため，記事の主題を含んでいないものはサムネイル生成に適していないプロンプトである．そこで，プロンプトの良し悪しを分類する分類モデルを構築することで，記事の主題を表したサムネイル生成に適したプロンプトを選ぶ．ただし，ここでの良いプロンプトとは以下の 2 つを十分に満たすものである．

- 記事の主題を含んだプロンプトである．
- 画像で表現することが可能なプロンプトである．

本研究ではプロンプトの良さを 3 段階に分割し，以下の様にラベルを定義する．

$$\text{label} = \begin{cases} l_1 & (\text{サムネイル生成に適さないプロンプト}) \\ l_2 & (\text{どちらでもないプロンプト}) \\ l_3 & (\text{サムネイル生成に適したプロンプト}) \end{cases} \quad (4)$$

プロンプト分類モデルに入力された複数のプロンプトから，最もサムネイル生成に適したプロンプト l_3 を 1 つ取得することが目標である．

3.2.2　プロンプト分類モデル

本研究では，プロンプトの分類モデルとしてランダムフォレストを用いる．ランダムフォレストは複数の決定木を用いたアンサンブル学習法の 1 つである．数値やカテゴリといった様々な種類の特徴量を扱うことができるため，良いプロンプトの条件を種類の異なる複数の特徴量で表現することができると考えられる．ランダムフォレストで使用する特徴量を表 1 に示す．特徴量の選択は主に 3 つの軸に基づき決定した．1 つ目は記事の主題を表すかである．特徴量として，LexRank [5] とタイトルとプロンプトの類似度を選択した．LexRank とは，文書から各文をノードとみなしたグラフ構造を作り出し，各ノード間の類似度を計算することでそれぞれの文の重要度を算出するアルゴリズムである．本研究では 1 つのノードに 1 文を割り当てるのではなく，プロンプト生成に用いた要約元の文章を 1 つのノードとして類似度を計算する．要約元の文章の重要度を計算することで，各プロンプトの文書内での重要度を算出する．これらの文章間の類似度を計算するには Sentence-BERT[1] [19] を用いてベクトル化を行い，コサイン類似度を計算する．Web ページのタイトルは文書の内容を簡潔に表していることがある．特にニュース記事などは読者に記事の情報を素早く理解してもらうために，タイトルが内容を簡潔に表現していると考えられる．そこでタイトルとプロンプトの類似度を特徴量に用いた．類似度は，タイトルとプロンプトのコサイン類似度を計算することで算出する．2 つ目は画像として自然な出力が期待できるかである．画像で表現することが容易な単語は具体度が高いと考えられるため，特徴量としてプロンプトに含まれる単語の具体度を用いる．具体度の計算には，様々な単語に具体度が付与された辞書 [2] を使用する．これは 40,000 単語に対して 1 から 5 の具体度を付与したものである．各プロンプトに対する具体度を以下の様に定義する．

$$\text{concreteness}_i = \frac{\sum_i^n \text{conc}(w_i)}{n} \quad (5)$$

ここで w_i はプロンプトを構成する単語を表し，$\text{conc}(w_i)$ は単語 w_i の具体度を返す関数である．ただし，具体度は 1 から 5 の間の実数であり，最も具体度が高い場合 5 を取る．3 つ目は，1 つ目と 2 つ目の両方を含むもしくはそれ以外に当てはまるものである．特徴量として，疑問詞の有無，プロンプトの単語数，品詞の数と割合を選択した．プロンプトに疑問詞が含まれる場合，例えばプロンプトに「何」,「なぜ」,「どこで」といった疑問詞を含む場合，表現することが困難であると考えられる．そのため疑問詞の有無を特徴量に用いた．またプロンプトの長さは生成画像の質に影響することが報告されている [24]．そのため特徴量にプロンプトの長さを用いた．ただしプロンプトの長さとはトークン数を表す．また，プロンプト中に名詞や固有名詞が含まれていなければ抽象度が高く，うまく画像を生成できないと考えられる．同様に形容詞や動詞が含まれていない場合でも抽象度が高まると考える．そのため，品詞の数とトークン数に対する割合を特徴量に選択した．

3.3　Web ページの主題を表したサムネイル生成

分類モデルによって得られたサムネイル生成に適したプロンプトを用いて，Web ページの内容に合ったサムネイルの生成を行う．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．

4　評 価 実 験

4.1　サムネイル生成に適したプロンプト分類の評価

構築したプロンプト分類モデルの性能評価を行う．分類モデルとしてはランダムフォレスト，XGBoost，BERT を用いる．

4.1.1　データセット

プロンプトを生成するためのデータセットとして CNN ニュース記事を用いる．CNN はアメリカ合衆国の国際的なケーブルテレビニュースネットワークであり，全世界にニュースを伝え

1：https://huggingface.co/sentence-transformers/all-mpnet-base-v2

表 2　ランダムフォレストの学習に用いたハイパーパラメータ.

パラメータ	値
決定木の深さ	20
葉に必要な最小サンプルサイズ	1
分割後ノードの最小サンプルサイズ	2
決定木の数	300

ている．そのため，記事に対するサムネイルやタイトルなどの質は高いと考えられる．CNN ニュースサイトからニュース記事 577 件を収集し，各ニュース記事のページから URL，タイトル，記事本文，サムネイルを取得した．これらを取得する際には Python のライブラリである Newspaper3k[2]を使用した．

4.1.2　実 験 条 件

収集したニュース記事の本文を用いてプロンプトのデータセット構築を行う．プロンプトの生成を行うための言語生成モデル T5 は，t5-one-line-summary[3]を使用する．T5 へ本文を 3 文ずつに区切って入力することで記事についてのプロンプトを複数生成する．生成された全プロンプト数は 7,255 である．生成された各プロンプトに対してプロンプトの良さを表すラベルを付与した．記事に掲載されているサムネイルとプロンプトから生成される画像との類似度を計算し，その値によってラベルを付与する．類似度が 0.0〜0.4 であればサムネイル生成に適さないプロンプト，0.4〜0.7 であればどちらでもないプロンプト，0.7〜1.0 であればサムネイル生成に適したプロンプトとラベルを付与する．ただし，プロンプトから画像を生成するモデルとして Stable Diffusion [20] を使用した．また，類似度を計算するときには 1 つのプロンプトから 5 枚画像を生成し，類似度の平均をとった．複数画像を生成することで，生成のたびに画像の質が異なることを考慮した．画像同士の類似度を計算するために，CLIP [16] を使用する．

4.1.3　ランダムフォレスト

プロンプトを分類する 1 つ目のモデルは，ランダムフォレストである．使用する特徴量として，表 1 を用いた．学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 2 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

4.1.4　XGBoost

2 つ目のモデルは XGBoost である．使用した特徴量はランダムフォレストと同様に表 1 を用いた．データについてもランダムフォレストと同様に学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 3 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

表 3　XGBoost の学習に用いたハイパーパラメータ.

パラメータ	値
学習率	0.2
決定木の深さ	5
決定木の数	300
各決定木のサンプル抽出比	0.9

表 4　BERT のファインチューニングに用いたハイパーパラメータ.

パラメータ	値
バッチサイズ	16
学習率	2×10^{-5}
早期終了	3 epoch
最適化手法	Adam
損失関数	交差エントロピー

4.1.5　BERT

3 つ目のモデルは BERT である．事前学習済みモデルとして，インターネット上の大規模なテキストコーパスを用いて学習された bert-base-uncased[4]を使用し，この事前学習済みモデルをファインチューニングして使用する．ファインチューニングのデータとしては，4.1.1 節で述べたデータセットを用いる．ファインチューニングに用いたハイパーパラメータを表 4 に示す．BERT への入力は，プロンプトである．先頭には特殊トークンの [CLS] を付与し，プロンプトの末尾には [SEP] を付与して入力する．ただし入力トークン長が最大入力長に満たない場合，特殊トークンの [PAD] で埋める．

4.1.6　評 価 尺 度

プロンプトをクラス l_k と分類するタスクにおいて，実際にクラス l_k である場合を正例，そうでない場合を負例とする．またクラスが l_k であるプロンプトを実際に正例とモデルが予測できたときを真陽性（TP），負例と予測したときを偽陰性（FN）とし，クラスが l_k でないプロンプトを正例とモデルが予測した時を偽陽性（FP），負例と予測したときを真陰性（TN）とする．評価尺度としては適合率（Precision），再現率（Recall），F 値のマクロ平均を用いる．適合率 P，再現率 R，F 値 F は以下の式で求められる．

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \tag{8}$$

マクロ平均とは，各クラスの評価指標の平均である．マクロ平均適合率 P_{macro}，マクロ平均再現率 R_{macro}，マクロ平均 F 値 F_{macro} は以下の式で求められる．

2：https://newspaper.readthedocs.io/en/latest/

3：https://huggingface.co/snrspeaks/t5-one-line-summary

4：https://huggingface.co/bert-base-uncased

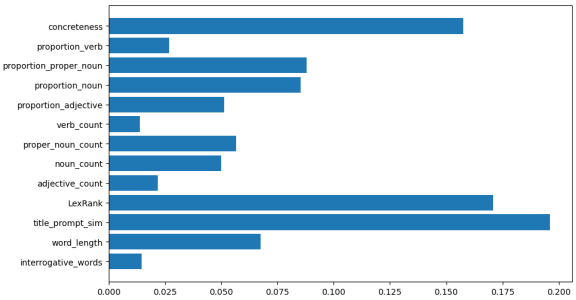


図 2　ランダムフォレストにおける各特徴量の重要度.

$$P_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} P_i \tag{9}$$

$$R_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} R_i \tag{10}$$

$$F_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} F_i \tag{11}$$

ここで K は分類するクラス数であり，$K = 3$ である.

4.2　サムネイル生成に適したプロンプトの分類実験結果

各モデルによるプロンプトの分類結果を表 5 に示す．実験の結果，ランダムフォレストはマクロ平均適合率，マクロ平均再現率，マクロ平均 F 値全てにおいて最も高い値であった．また，良いプロンプトであるラベル l_3 の適合率，再現率，F 値も最も高い値であった．反対に自然言語処理タスクにおいて高い性能を示す BERT は最も低い値であった．深層学習に基づく分類手法よりも特徴量による分類の方が高い精度で分類を行えたことがわかる．また，最も分類精度が高かったランダムフォレストにおける各特徴量の重要度を図 2 に示す．図 2 より，タイトルとプロンプト類似度，LexRank，各単語の具体度の平均が特に重要な特徴量であったことがわかる.

4.3　Web ページの主題を表したサムネイル生成の評価

構築したプロンプト分類モデルを用いて選ばれたプロンプトを使用し，記事に対するサムネイルの生成を行う．使用するプロンプト分類モデルとしては，分類精度が最も高かったランダムフォレストを使用する．モデルがサムネイル生成に適したプロンプトと最も高い確率で予測した 1 つのプロンプトを画像生成モデルに入力する．もしプロンプトの中にサムネイル生成に適したプロンプトと予測したものが無い場合，最も高い確率でどちらでもないプロンプトと予測した 1 つのプロンプトを画像生成モデルに入力することで記事に対する画像を生成する．また比較手法として記事のタイトルをプロンプトとし，タイトルを画像生成モデルに入力することで記事に対する画像を生成する．本評価実験では，提案手法で生成された画像と比較手法で生成された画像について評価を行う.

4.3.1　データセット

CNN ニュース記事と旅行記事が掲載されているブログなどのサイトを用いる．CNN ニュース記事からは，Business，Entertainment，Health，Sport，Travel のジャンルから各 5 記事ずつの計 25 記事を使用する．ただし，CNN ニュース記事は分類モデルの学習，評価に使用していないデータを用いる．また，ブログ等の旅行記事サイトは 14 サイトからランダムに計 25 記事収集し使用する．ただし，ページのタイトルが「人気観光地 10 選」のようなまとめ記事は使用しない．CNN ニュース記事からは Newspaper3k ライブラリを用いて本文を取得し，旅行記事サイトからは手作業で本文を取得した.

4.3.2　実験条件

画像生成モデルとして Stable Diffusion と DALL-E2 を使用する．DALL-E2 は Stable Diffusion と同様にテキストから画像を生成することができるモデルである．1 つの Web ページ記事に対し，Stable Diffusion と DALL-E2 を用いて 1 枚ずつ画像を生成する．すなわち評価者は，1 つの記事に対して（提案手法とベースライン）×（Stable Diffusion と DALL-E2）による生成の 4 枚の画像を評価する．生成された画像について，同研究室の学生 2 名と著者によるアンケートを通じて評価を行う．評価アンケートでは以下の点について評価を行ってもらう.

（1）　画像が記事の主題を表していますか.

（2）　画像として自然ですか.

（3）　記事のサムネイルとしてどちらの画像が適していますか.

評価 1 では，各サムネイルについて記事の主題を表しているかを評価する．評価者には 5 段階評価で回答を求めた．ここでは 5 が最高点とした．評価 2 では，各サムネイルについて，画像としての自然であるかを評価する．不自然な画像とは，画像が表している風景，状態を読み取ることができない画像のことを指す．評価 1 と同様に評価者には 5 段階評価で回答を求めた．ここでも最高点は 5 点とした．評価 3 では，提案手法とベースラインどちらで生成された画像が記事のサムネイルとして適しているかを回答してもらう．なお，評価者にはどの画像がどの手法によって生成されたかなどの情報を公開せずに評価を行う.

4.3.3　評価尺度

評価 1 と評価 2 では，各生成モデルについて評価者の点数の平均を用いて手法の比較評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は各記事 Article_i の評価点の平均であり，以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n \text{Article}_i}{n} \tag{12}$$

ただし，n はジャンル g に含まれる記事数である.

評価 3 では，各生成モデルについて提案手法とベースラインで好まれた記事数の平均を用いて評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n f(\text{Article}_i)}{n} \tag{13}$$

$$f = \begin{cases} 1 & (\text{Article}_i\text{のサムネイルとして適している}) \\ 0 & (\text{Article}_i\text{のサムネイルとして適していない}) \end{cases} \tag{14}$$

ここで f はサムネイルが Article_i のサムネイルとして適していれば 1 を返し，適していなければ 0 を返す関数である.

表 5　プロンプト分類モデルの評価結果.

分類モデル	ラベル l_3 の適合率	ラベル l_3 の再現率	ラベル l_3 の F 値	マクロ平均適合率	マクロ平均再現率	マクロ平均 F 値
ランダムフォレスト	**0.756**	**0.523**	**0.617**	**0.796**	**0.683**	**0.725**
XGBoost	0.704	0.478	0.570	0.760	0.659	0.696
BERT	0.593	0.467	0.523	0.670	0.648	0.656

4.4　Web ページの主題を表したサムネイルの評価結果

評価 1 のサムネイルが記事の主題を表しているかの評価結果を表 6，評価 2 のサムネイルが画像として自然であるかの評価結果を表 7，評価 3 の記事に適したサムネイルはどちらであるかの評価結果を表 8 に示す．表 6 より，全体的にはベースラインの方が記事の主題を表すことができていた．しかし Entertainment と Health のジャンルでは，提案手法のサムネイルの方が記事の主題を表すことができていたことがわかる．また Health のジャンルでは，Stable Diffusion と DALL-E2 両方において提案手法のサムネイルが記事の主題を表すことができた．表 7 より，全体的には提案手法とベースラインで画像としての自然さに違いはなかった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 両方において提案手法の方がスコアが高い結果となった．表 8 より，全体的にはベースラインの方が記事のサムネイルとして選ばれる数が多かった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 の両方において提案手法の方が高いスコアになった．また評価 1，評価 2，評価 3 について一致度を算出するために Krippendorff の α 係数 [11] を計算した．求まった α 係数の値を表 9 に示す．

5　議　　論

5.1　サムネイル生成に適したプロンプト分類

サムネイル生成に適したプロンプトの分類では，自然言語処理タスクで高い性能を示す BERT よりも，決定木をベースとしたアンサンブル学習手法であるランダムフォレストや XGBoost の方が高い分類精度を示した．その理由として，プロンプト分類に使用した情報量の差ではないかと考える．BERT の入力はプロンプトのみである．そのため BERT はプロンプトの単語のみの情報を分類に使用している．一方でランダムフォレストや XGBoost はプロンプトのみではなく，プロンプトの元となった文章の LexRank やタイトルといった記事中の情報も用いている．サムネイル生成に適したプロンプトとは，記事の主題を含んでいる且つ画像で表現することが可能なプロンプトである．プロンプトの単語情報だけでなく記事中の情報も用いていることから，サムネイル生成に適した 2 つの条件を考慮することができたため，ランダムフォレストの分類精度の方が高かったのではないかと考える．

5.2　Web ページの主題を表したサムネイル生成

プロンプトの分類を行うにあたり，良いプロンプトとして（1）記事の主題を表している,（2）画像として自然な出力が期待できる，の 2 つの観点を考慮した．そのため（1）を満たす特徴量として LexRank，タイトルとプロンプトの類似度,（2）の特徴量としてプロンプトの具体度が影響すると考えた．サムネイルに対する評価の点数とプロンプトの特徴量について相関係数を計算することで，特徴量の影響を分析した．相関係数の値を表 10 と表 11 に示す．記事の全ジャンルについて相関係数を計算した結果，記事の主題を表すかの点数と LexRank，タイトルとプロンプトの類似度にそれぞれ正の相関がみられた．画像として自然であるかの点数とプロンプトの具体度にも正の相関がみられた．また，Business と Entertainment 以外のジャンルに絞ると全ジャンルに比べて強い相関がみられた．以上のことから，LexRank とタイトルとプロンプトの 2 つの特徴量は類似度は記事の主題を表すかに影響を与え，プロンプトの具体度は画像の自然さに影響を与えることがわかった．

評価 1 と評価 2 において，Entertainment と Health のジャンルでは提案手法が好まれた．この 2 つのジャンルの中でも特に Health ジャンルはベースラインとの差が大きかった．その理由として，プロンプトの具体度が影響していると考える．Text2Image モデルに入力したベースラインの記事タイトルに比べ，提案手法によるプロンプトの方が具体度の平均が約 0.5 高く，これは他ジャンルには見られない傾向であった．以上の理由から Health ジャンルは他のジャンルと異なり，ベースラインよりも提案手法の方がスコアが高くなったと考える．

評価 3 においては，提案手法によるサムネイルよりもベースラインによるサムネイルの方が好まれる結果であった．CNN ニュース記事では大きく差はないが，CNN ニュースの Travel と旅行記事サイトにおいてはベースラインが大きく好まれる傾向がみられた．旅行系記事には様々な風景についての情報などが含まれていた．そのため主題とは少し異なる内容を含むプロンプトも多く生成されたためにプロンプトの選択で誤ったものが選ばれ，ベースラインに勝ることができなかったのではないかと考える．また，Stable Diffusion では提案手法によるサムネイルが多く好まれ，DALL-E2 ではベースラインによるサムネイルが多く好まれた．その理由として，プロンプトのラベル付けを行ったときに用いた画像生成モデルが Stable Diffusion であったためだと考える．プロンプトに対するラベルは，プロンプトから生成された画像と記事に掲載されているサムネイルとの類似度を計算することで 3 段階のラベルを付与した．このときの画像生成モデルとして本研究では Stable Diffuision を使用した．そのため各プロンプトに付与されたラベルは Stable Diffusion を用いて画像生成を行うときに対するプロンプトの良さであると考えられるため，画像生成モデルによって提案手法と比較手法の評価に差が生じたと考える．

5.3　限　界　点

本研究ではサムネイルを生成するにあたり，T5 によって生

表 6　評価 1：サムネイルが記事の主題を表しているか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	2.87	**4.13**	**3.20**	4.53	2.20	3.93	3.47
	ベースライン	**3.47**	3.93	2.47	**4.60**	**3.93**	**4.10**	**3.89**
DALL-E2	提案手法	**2.6**	3.07	**3.40**	3.87	3.80	3.65	3.02
	ベースライン	2.4	**4.00**	2.87	**4.13**	**4.07**	**3.80**	**3.65**

表 7　評価 2：サムネイルが画像として自然であるか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	3.67	**4.93**	**4.27**	3.57	4.53	4.51	4.41
	ベースライン	3.67	4.73	3.53	**4.73**	**4.73**	**4.53**	4.41
DALL-E2	提案手法	**3.53**	4.27	**3.93**	3.38	4.33	3.84	4.17
	ベースライン	3.20	**4.73**	3.60	**4.40**	**4.80**	**4.20**	4.17

表 8　評価 3：記事に適したサムネイルはどちらか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	**2.67**	**2.67**	**2.67**	**2.67**	0.67	8.33	19.7
	ベースライン	2.33	2.33	2.33	2.33	**4.33**	**16.7**	**30.3**
DALL-E2	提案手法	2.33	1.33	**3.33**	1.67	0.67	8.00	17.3
	ベースライン	**2.67**	**3.67**	1.67	**3.33**	**4.33**	**17.0**	**32.6**

表 9　各評価における Krippendorff の α 係数.

生成モデル	評価 1	評価 2	評価 3
Stable Diffusion	0.59	0.24	0.44
DALL-E2	0.50	0.28	0.33

表 10　評価 1 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	LexRank	タイトルとプロンプトの類似度
全ジャンル	0.306	0.749
Business と Entertainment 以外	0.921	0.929

表 11　評価 2 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	プロンプトの具体度
全ジャンル	0.310
Business と Entertainment 以外	0.373

成されたプロンプトの中から記事の主題を表している，画像で表現が可能なプロンプトを分類モデルを用いて選ぶという手法を提案した．そのため T5 によって生成されたプロンプトの中に，サムネイルのための優れたプロンプトが存在しているということが前提条件の手法である．したがって T5 によるプロンプトの生成がうまく行われなかった場合，提案手法によるサムネイルの質は悪くなる．本研究の実験と評価では，T5 によって生成された複数のプロンプトの中に優れたプロンプトが存在するという仮定の下で行った．そのため T5 の性能向上，あるいは他の言語モデルの性能向上によって提案手法によるサムネイルの評価結果が改善される可能性がある．

6　まとめと今後の課題

本研究では，サムネイル生成に適したプロンプトの分類を行うことによって画像生成モデルへの入力を選別し，Text2Image モデルを用いて Web ページの主題を表したサムネイル生成に取り組んだ．提案手法のプロンプト分類モデルとして，ランダムフォレストを構築した．プロンプト分類の結果，提案手法の分類精度が最も高く，自然言語処理タスクで高い性能を示す BERT よりも高い分類性能を示した．また，この構築したプロンプト分類モデルによって選ばれたプロンプトを用いて，Web ページのサムネイル生成を行った．CNN ニュースと旅行記事サイトに対して生成をし評価を行った結果，CNN ニュースの内 Travel のジャンル以外では提案手法によるサムネイルが好まれ，Travel と旅行記事サイトではベースラインであるタイトルをプロンプトとしたサムネイルが好まれた．ベースラインが勝った理由として，タイトルが記事の内容を端的に表現しており強力であったことや，プロンプトの誤分類が提案手法で起きていたことなどが考えられる．今後は，生成するプロンプトの質を高める必要があることやプロンプト分類の精度をさらに向上させる必要がある．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 1197–1206, 2008.

[2] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, Vol. 46,

pp. 904–911, 2014.

[3] Andy Cockburn and Bruce McKenzie. What do web users do? an empirical analysis of web use. *International Journal of human-computer studies*, Vol. 54, No. 6, pp. 903–922, 2001.

[4] Susan Dziadosz and Raman Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 365–366, 2002.

[5] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, Vol. 22, pp. 457–479, 2004.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851, 2020.

[9] Binxing Jiao, Linjun Yang, Jizheng Xu, and Feng Wu. Visual summarization of web pages. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 499–506, 2010.

[10] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[12] Shenwei Liu and Keishi Tajima. Wildthumb: a web browser supporting efficient task management on wide displays. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pp. 159–168, 2010.

[13] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2022.

[14] Vivian Liu, Han Qiao, and Lydia Chilton. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2022.

[15] A Porselvi and S Gunasundari. Survey on web page visual summarization. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 1, pp. 26–32, 2013.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, Vol. 1, No. 2, p. 3, 2022.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[21] Erica K Shimomoto, Lincon S Souza, Bernardo B Gatto, and Kazuhiro Fukui. News2meme: An automatic content generator from news based on word subspaces from text and image. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019.

[22] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M Drucker, Gonzalo Ramos, Paul André, and Chang Hu. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2023–2032, 2009.

[23] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 198–205, 2001.

[24] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023*, pp. 3892–3902, 2023.

[25] 稲垣有哉, 岩田一, 白銀純子, 深澤良彰. ウェブページの構造と顕著性マップの組み合わせによる重要領域の視覚化手法. Web インテリジェンスとインタラクション研究会 予稿集 第 14 回研究会, pp. 45–50. Web インテリジェンスとインタラクション研究会, 2019.

表 12　ベースラインとして用いた記事タイトル例.

記事のジャンル	記事タイトル
Business	These popular fast food drive-thrus are getting faster
Entertainment	Late-night hosts return to air post-writers' strike
Health	mRNA vaccine: 5 things to know
Sport	Solheim Cup: Europe retains title after stunning comeback against USA
Travel	The top superyachts at Monaco Yacht Show 2023
旅行記事サイト	Hershey ' s Chocolate World Announces New Experience That Will Immerse Visitors In A Delicious Train Ride

表 13　提案手法によって選ばれたプロンプト例.

記事のジャンル	選ばれたプロンプト
Business	Artificial Intelligence Drive-Thru and Walk-Up Locations at Chick-fil-A
Entertainment	Late-Night Network Talks Revisited
Health	Essence RNA-Based Vaccine Technology
Sport	First blood vs. Ireland in the Solheim Cup
Travel	The peak of the super yachting sector
旅行記事サイト	The Chocolate World Experience

表 14　提案手法が好まれたサムネイル例.

記事のジャンル	ベースライン	提案手法
Business		
Entertainment		
Health		
Sport		
Travel		
旅行記事サイト		

類似度グラフの事前再計算による拡散式画像検索の効率化

加藤　辰弥†　　駒水　孝裕†　　井手　一郎†

† 名古屋大学　〒 464–8601 愛知県名古屋市千種区不老町

E-mail: †katot@cs.is.i.nagoya-u.ac.jp, ††taka-coma@acm.org, †††ide@i.nagoya-u.ac.jp

あらまし　画像検索タスクにおいて，事前計算による高速な手法とリランキングに基づく高精度な手法がそれぞれ提案されているが，速度と精度を両立した手法の開発が課題になっている．我々はこれらの手法を統合することで，この課題に対処した画像検索のフレームワーク R-DiP （Re-ranking based Diffusion Pre-computation）を提案する．具体的には，事前計算手法において類似度を計算する際に，リランキングで用いられる高精度な類似度再計算を導入することで，高速で高精度な検索を実現する．ベンチマークを用いた実験により，提案手法が最先端の手法と同等な検索精度を維持しつつ，高速な検索を実現できることを示す．提案手法は特に大規模なデータセットにおいて，mAP スコアを 2.0% 程度向上し，クエリごとの検索速度を平均で 75% 程度の削減をし，最先端の高精度手法を上回る検索精度を示した．提案する枠組みは，任意のリランキング手法を取り込むことが可能であり，今後出現するであろう高精度なリランキング手法に対しても効果を発揮し，画像検索技術の進展に貢献することが期待される．

キーワード　画像検索，Content-based Image Retrieval，Diffusion，Re-ranking，効率化

1　はじめに

ディジタルカメラの低廉化やスマートフォンなどの携帯端末に搭載される高性能カメラの普及，画像編集ソフトウェアの大衆化により，個人がディジタル画像を作成することが容易になった．また，Web 上のソーシャルメディアなどのコンテンツ共有プラットフォームの普及に伴い，作成された画像を手軽に公開することができるようになり，我々がアクセスできる画像データの量が増大している．この状況に対して，画像データを管理・活用するために，画像検索は重要な技術のひとつである．画像は実世界を記録するメディアとして言葉だけでは表現することができない事象を記録することに向いており，画像検索は与えられた検索要求に基づいて，過去の事象の記録や関連する事象を探すための主要な手段の一つである．画像検索の中でも，画像を検索要求とする「Content-Based Image Retrieval（CBIR）」があり，キーワード検索やメタデータ検索では捉えきれない画像自身がもつ情報を活用するという特徴がある．

CBIR は 1990 年代 [17] から長い間研究されてきているものの，その性能向上は依然として研究課題である．その主な課題は，画像からの特徴抽出である．近年，ニューラルネットワーク技術の進歩に伴い，従来の人手による特徴量の設計から，大量のデータから自動的に適確な画像特徴を抽出する枠組みが主流となってきた．例えば，畳み込みニューラルネットワーク（Convolutional Neural Network；CNN）[1], [9], [31] や Vision Transformer（ViT）[6], [7] など，数多くの手法が提案されている．これらは検索の目的に合わせて構築されてきたわけではないため，必ずしも検索に適した特徴とは限らない．これに対して，より良い検索を実現するために，検索に適した結果の並び替えを行なうリランキングを利用する手法 [2], [22], [26], [35], [37] やランダムウォークに基づく拡散 [4], [11], [14], [46], [47] などが提案されている．この中でも，Correlation Verification Networks（CVNet）[22] のような，局所特徴をリランキングに用いる手法が高い検索精度を示している．一方で，このようなニューラルネットワークモデルを用いた類似度の推定は低コストとは言えず，検索のオーバヘッドが増し，検索速度が低下する．それに対応するために，大域特徴のみをリランキングに用いる手法 [34] も提案されているが，依然として十分に効率的とは言えない．

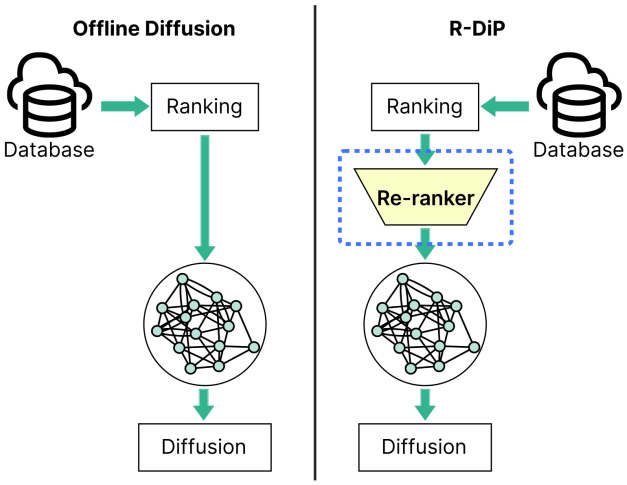


図 1: R-DiP の枠組み．

これに対して，高速な検索を実現するために，データベース画像間の類似性をランダムウォークにより事前計算する Offline Diffusion [44] と呼ばれる手法が提案されている．その基礎となっている拡散（Diffusion）モデルは，擬似適合性フィードバック [47] の考え方を用いることで，同一物体を異なる視点から写した画像や，遮蔽物により物体の一部が隠れている画像を検

索することができる手法である．具体的には，検索要求とデータベース画像から k 近傍グラフを構築し，その上でランダムウォークを行なうことで，データの大域的な構造を考慮した効果的な検索を実現している．Offline Diffusion は，k 近傍検索によるグラフ構築および画像間の類似度計算をデータベース側で事前に（つまりオフラインで）実行する．これにより，検索時に k 近傍検索の結果との単純な線型結合によって検索を行なうことが可能になり，高速な検索を実現している．しかし，その検索精度は事前処理時のグラフ構築に大きく依存し，リランキングを利用する高精度な手法 [22], [34] と比較して劣っている．

本研究では，以上のような速度と精度の間のトレードオフを解決するための手法として，事前計算に基づく高速な手法である Offline Diffusion とリランキングを利用する高精度な手法を融合するための新しいフレームワーク R-DiP （Re-ranking based Diffusion Pre-computation）を提案する．R-DiP の枠組み（図 1）は，Offline Diffusion の近傍検索や関連する類似度検索に高性能なリランキング手法を取り込むことで精度の向上を図る．これにより，検索における関連度を計算する任意のリランキング手法を取り込むことができる柔軟性がある．本研究では，ベンチマークを用いた実験により，特に大規模データセットにおいて，提案手法がリランキングに基づく SuperGlobal [34] に匹敵する検索性能を維持しつつ，Offline Diffusion と同等に高速な検索を実現できることを示す．

本研究の主な貢献は以下の通りである．

- **フレームワーク R-DiP の提案：**Offline Diffusion を基礎とし，リランキング手法の利点を取り込むフレームワークを提案する．提案フレームワークは検索における関連度を計算する任意のリランキング手法を取り込める柔軟性があるため，今後より検索精度が高い高計算コストなリランキング手法が出現した場合にも適用可能である．
- **高速・高性能な検索の実現：**R-DiP は，リランキングに基づく最先端手法である SuperGlobal に匹敵する検索精度を維持しつつ，Offline Diffusion と同等に高速な検索を達成する．特に，大規模なデータセットにおいて，mAP スコアを Offline Diffusion の特性により，最先端の手法より 2.0% 程度向上したうえで，クエリごとの検索速度を平均で 75% 程度の削減をし，最先端の手法を上回る検索性能を示した．

2 関 連 研 究

画像検索における重要な課題は，画像特徴の設計である．機械学習の進歩に伴い，コンピュータビジョンに関する研究として，CNN [19], [21], Regional Maximum Activation of Convolutions（R-MAC）[42]，ViT [5]，GAN [8] などのニューラルネットワークによる手法が登場してきた．これらの手法は画像検索にも採用され [1], [6], [7], [9], [10], [18], [23], [25], [31], [32], [36], [38]，従来の問題であったセマンティックギャップを克服し，画像検索の精度を大幅に向上させている．

画像特徴の設計に続き，画像検索の特性を考慮した検索の最適化に向けた手法が提案されている．例えば，k 近傍検索に基づく検索 [3], [24]，上位に順位付けられた検索結果の類似度を再計算するリランキング，ランダムウォークに基づく拡散 [4], [11], [14], [46], [47] などがある．その中で，リランキングを利用した手法は検索精度の面で優れており，拡散の中でも事前計算に基づく手法 [44] や k 近傍検索に基づく手法は速度の面で優れていることが明らかにされている．以下に，リランキングおよび拡散に基づく手法について詳細に述べる．

a） **画像検索におけるリランキング**

リランキングは，2 段階の検索手法であり，始めに軽量な手法を用いて検索結果を大まかにフィルタリングした後，高精度な類似度計算モデルによって順序付けを行なう．リランキングを利用した CBIR 手法では，最初の段階で大域的特徴に基づいて検索し，物体の局所的な関係性を捉える幾何検証 (Geometric Verification) により類似度を再計算する手法が主流である．幾何検証に関する研究としては，特徴点検出に焦点をあてた研究 [20], [27], [33], [45] や特徴抽出に焦点をあてた研究 [13], [16], [26], [33], [35], [40], [41], [45]，特徴点検出と抽出の順番を入れ替えた研究 [39] など，様々な手法が提案されている．

さらに，幾何検証の代替手法として，ニューラルネットワークを用いて抽出した局所特徴マップを用いる手法が登場した．Transformer を利用した Re-Ranking Transformer（RRT）[37]，局所特徴と大域特徴を効果的に統合した Super-Feature を提案する研究 [43]，局所特徴を用いて複数の異なる大きさで画像間の類似度を計算する CVNet [22] などが提案されている．これらの手法は高精度な検索結果を示すが，速度に関するオーバヘッドが大きいことが問題である．より低オーバヘッドな手法として，リランキングに大域特徴のみを用いる SuperGlobal [34] が提案されているが，事前計算に基づく高速な手法 [44] などと比較すると速度の面で改善の余地が残っている．

b） **拡散（Diffusion）/ Offline Diffusion**

拡散 [14] の核となる考え方は，クエリ画像とデータベース内の画像の類似性をデータベース内の類似画像グラフを通して拡散させることである．拡散の主要な利点は，k 近傍検索のように単に Euclidean 空間上での画像特徴の距離に依存するのではなく，近接性に基づくネットワーク構造を考慮することによって，間接的に類似している画像を発見できる点にある．このような違いにより，例えば，同一物体を異なる視点から写した画像や，遮蔽物により物体の一部が隠れている画像の検索が可能になる．しかし，特に大規模なデータセットに適用する際には，ランダムウォークの計算コストが高いため，検索に時間を要するという欠点がある．

この問題に対処するために，Offline Diffusion [44] が提案されている．Offline Diffusion では，拡散過程がデータベース側で事前に実行されることで，検索時の計算負荷を大幅に軽減できる．具体的には，拡散を行ない事前計算した到達確率と $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n : \mathbf{c}_i \in \mathbb{R}^n\}$（ここで，$n$ はデータベース画像枚数）と k 近傍検索で求める y の線型結合として，以下のようにして画像間の類似度を高速に取得する．

（1） k_q 近傍検索：k_q は初期の k 近傍検索における検索枚数である．クエリ画像 q に対して，類似画像のインデックス

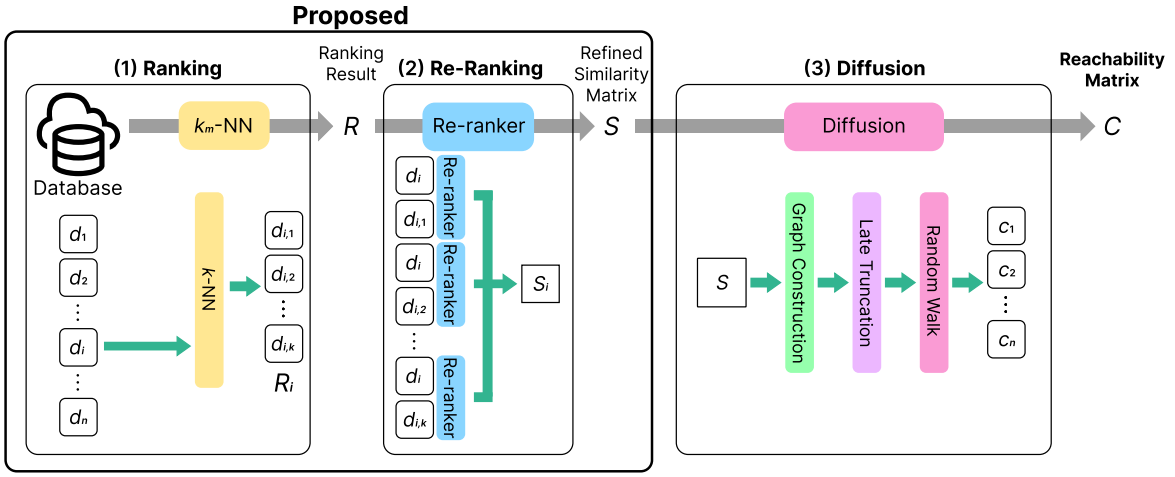


図 2: 提案手法におけるオフライン過程の概要.

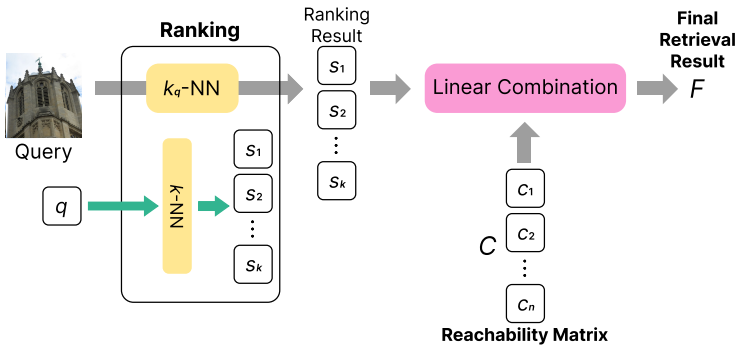


図 3: 提案手法におけるオンライン過程の概要.

$J = \{j_1, j_2, \ldots, j_{k_q}\}$ と類似度 $S = \{s_1, s_2, \ldots, s_{k_q}\}$ を得る．

（2） 線型結合：k_q 近傍検索の結果を用いて，最終的な類似度を $F = \sum_{\ell=1}^{k_q} \mathbf{c}_{j_\ell} s_\ell^\gamma$ として計算する．ここで，$\mathbf{c}_i \in C$ は事前に計算した到達確率ベクトル，γ は画像特徴に基づく類似度の重要度を調整する重み変数である．

この手法は高速な反面，検索精度はリランキングを利用する高精度な手法には劣るため，検索精度の向上が課題である．

3 提案手法：R-DiP

本節では，提案する R-DiP の概要について述べる．R-DiP は，Offline Diffusion における事前計算に，リランキングで用いられる高精度な類似度再計算モデルを組み込むことで，高速で高精度な画像検索の実現を図る．R-DiP の考え方の特徴は以下の二つである．

- **拡散過程の改善**：拡散による類似度計算は，k 近傍グラフの質に大きく依存する．単純な k 近傍検索による構築は画像特徴の設計に依存しており，検索を目的として設計されていない画像特徴の場合に，最適な近傍グラフを構築できるとは限らない．そこで，画像検索におけるリランキングで用いられる類似度計算モデルを用いることでより適した近傍グラフを再構築し，検索精度の向上を図る．
- **リランキングの push-down**：リランキングに基づく手法は精度の面で優れているものの，検索時にリランキングを適用する際のオーバーヘッドは依然として大きい．Offline Diffusion のオフライン過程にリランキングのプロセスを移行することにより，Offline Diffusion の精度を補いつつ，リランキングに必要な計算時間を削減する．

R-DiP は，事前計算を行なうオフライン過程と，実際に検索を行なうオンライン過程の二つから構成される．オフライン過程とオンライン過程の概要をそれぞれ図 2 と図 3 に示す．

ここで，データベース画像の集合を $D = \{d_1, d_2, \ldots, d_n\}$（ただし，$n$ は画像の総数である）とし，画像検索タスクを以下のように定義する．

- **入力**: 一つのオブジェクトを含むクエリ画像 q
- **出力**: q に含まれるオブジェクトを含む画像集合 $D' \subseteq D$

なお，以下では，データベース中の画像をデータベース画像と呼ぶ．本研究では，クエリ画像 q とデータベース画像 $d_i \in D$ は，事前に抽出された画像特徴とする．

3.1 オフライン過程

オフライン過程では，以下のようにして到達確率を計算する．

（1） k 近傍検索による類似画像の取得（k_m-NN）

（2） 類似度再計算モデルによる類似度再計算（Re-ranking）

（3） 拡散に基づく到達確率計算（Diffusion）

これらは，図 2 の（1）～（3）に対応している．Offline Dif-

fusion [44] では，k 近傍検索による画像間類似度計算後に，拡散に基づく類似度計算を行なう．しかし，この類似度は，データベース画像の特徴量の品質に大きく影響を受けるため，その後に計算される類似度の精度を大きく損なう可能性がある．これに対処するため，本研究では，(1) の k 近傍検索の後に (2) の類似度再計算モデルによる類似度再計算を行なうことで，(3) の拡散に基づいて計算される到達確率の精度向上を図る．

ここで，類似度再計算モデルを全てのデータベース画像に適用せず，k 近傍検索の後に適用する理由は二つある．一つ目は，リランキングの概念を取り入れることで事前計算コストを削減できるからである．k 近傍検索により類似度が低いとされる画像が検索結果に与える影響は少ないため，類似度が高い画像のみを考慮することで類似度再計算に要する計算コストを削減することができる．二つ目は，拡散の特性により，ランダムウォークを通じて類似度が高い画像群を特定することができるため，類似度再計算された画像が少ない場合でも効果的に画像間の類似度を求めることができるからである．

3.1.1 k 近傍検索による類似画像の取得

この過程では，各データベース画像 $d_i \in D$ を入力として，$k = k_m$ とする k 近傍検索を行ない，データベース全体から最も類似した k_m の画像集合 $R_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,k_m}\} \subseteq D$ を取得する．ここで，$d_{i,j}$ は d_i と j 番目に類似度の高い画像を示す．また，すべてのデータベース画像 $d_i (1 \leqq i \leqq n)$ の R_i からなる集合を $R = \{R_1, R_2, \ldots, R_n\}$ とする．なお，画像間の類似度は余弦類似度を用いて計算する．

3.1.2 類似度再計算モデルによる類似度再計算

k 近傍検索により類似画像を取得したのち，各 $R_i \in R$ について，類似度再計算モデル $s : D \times D \to \mathbb{R}$ を用いて d_i と $d \in R_i$ の類似度を再計算する．この過程は近傍グラフの質を向上させることを目的としている．これにより，検索により適した類似度行列 $\mathbf{S} \in \mathbb{R}^{n \times n}$ を得る．ここで，検索結果に含まれない画像 $d_j \notin R_i$ との類似度 $\mathbf{S}_{ij}$ は 0 とする．

3.1.3 拡散に基づく到達確率計算

まず，再計算された類似度行列 $\mathbf{S}$ を用いて，拡散過程のための類似度グラフを構築する．ここでは，再計算された類似度の上位 k_d 件を用いて類似度グラフ $G = (V, E, \delta)$ を作る．ここで，$V = D$ はノード集合，$E \subseteq V \times V$ は $\mathbf{S}_{ij} > 0$ である場合に $E_{ij} = (d_i, d_j)$ で示されるエッジ集合，写像 $\delta : E \to \mathbb{R}$ は $\delta(E_{ij}) = \mathbf{S}_{ij}$ とする．次に，Offline Diffusion [44] 同様に拡散を行ない，到達確率を計算する．これにより，データベース内画像間類似度 $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ が求まり，オンライン過程でこれに基づいて検索を行なう．ここで，$\mathbf{c}_i \in \mathbb{R}^n$ は i 番目の画像とデータベースの他の画像間の類似度のベクトルを表す．

3.2 オンライン過程

オンライン過程では，効率的な検索を実現するために，Offline Diffusion と同じ方法を採用する．これにより，最終的な検索結果は，$F = \sum_{\ell=1}^{k_q} \mathbf{c}_{j_\ell} s_\ell^\gamma$ として計算する．ここで，j_ℓ と s_ℓ は，クエリ画像 q に対する k 近傍検索 $(k = k_q)$ の結果のうち，ℓ 番目に類似度が高い画像のインデックスと類似度であり，γ は画像特徴に基づく類似度の重要度を調整する重み変数である．

4 実　　験

R-DiP を検索精度と速度の両面で評価するために，提案手法が基礎とした Offline Diffusion [44] と最先端の検索性能を達成している SuperGlobal [34] と比較する実験を行なった．

4.1 比 較 手 法

各手法の詳細と比較手法とした目的は以下の通りである．

- **Offline Diffusion：**拡散に基づく画像検索の速度を改善するために設計された手法である．拡散過程の一部を事前に計算することで，検索時の計算コストを削減して検索を高速化している．提案手法と比較することで，提案する類似度再計算の効果を評価する．
- **SuperGlobal：**効果的な大域特徴の取得とそれを用いたリランキングを行なう手法である．この手法は，SuperGlobal Pooling と SuperGlobal Reranking の二つで構成される．SuperGlobal Pooling では Generalized Mean（GeM） Pooling [32] に L_p pooling [12] を取り入れた Regional GeM [34] を利用することで，局所的な特徴を考慮した大域特徴を出力する．一方，SuperGlobal Reranking では，k 近傍検索で取得した初期検索結果の各画像と他の画像のうち類似するものの特徴量を組み合わせることで新たな大域特徴を作成し，類似度を再計算する．この手法は最先端の画像検索手法であり，リランキングに基づく他の手法よりも高速かつ高精度な検索を実現しているが，依然としてリランキング時のオーバヘッドにより Offline Diffusion に比べて検索速度で大きく劣っている．提案手法と比較することで，提案手法の検索速度と精度をともに評価する．

4.2 ベンチマーク

本研究では，提案手法の性能評価のために，画像検索でベンチマークとして用いられる $\mathcal{R}$Oxford5k データセット [30] と $\mathcal{R}$Paris6k データセット [30] を用いる．これらは，旧来のベンチマークである Oxford 5k データセット [28] と Paris 6k データセット [29] のアノテーションや評価難易度を改良したベンチマークである．以下に，$\mathcal{R}$Oxford5k データセットと $\mathcal{R}$Paris6k データセットの特徴を述べる．

- **$\mathcal{R}$Oxford5k データセット：**英国 Oxford 市内のランドマークや建築物の画像，Oxford 大学内の歴史的建造物や，同大学周辺の橋や教会などの画像で構成される．建築物の視覚的特徴を，異なる条件下（異なる視点や時間帯，気候など）において識別する能力の評価に適している．
- **$\mathcal{R}$Paris6k データセット：**フランス Paris 市内の著名なランドマークや観光名所の画像，Eiffel 塔や Louvre 美術館，Notre-Dame 大聖堂など象徴的な場所を写した画像で構成される．都市景観や観光名所を，異なる条件下（異なる視点や時間帯，気候など）において識別する能力の評価に適している．

これまでの研究 [2], [22], [34] では，$\mathcal{R}$Oxford5k データセットは $\mathcal{R}$Paris6k データセットと比較して，検索精度が低くなる傾向が見られる．

表 1: 提案手法と比較手法の精度評価（mAP）.

Method	Medium				Hard			
	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M
Offline Diffusion [44] (k_q=100)	89.12	85.78	95.22	92.09	75.95	71.11	88.99	84.63
Offline Diffusion [44] (k_q=400)	89.88	85.91	95.37	92.75	77.06	**72.10**	89.29	85.20
SuperGlobal [34] (k_q=100)	88.26	81.76	92.20	83.24	76.67	67.74	83.79	67.66
SuperGlobal [34] (k_q=400)	**90.79**	84.08	93.36	85.53	**80.05**	70.93	86.77	72.52
R-DiP(k_q=100)	89.71	**86.55**	95.39	92.94	76.94	71.97	90.68	**85.61**
R-DiP(k_q=400)	90.40	**86.55**	**95.79**	**93.15**	77.67	71.80	**91.30**	85.46

表 2: 最適なパラメータ.

Method	Medium															
	$\mathcal{R}$Oxf				+$\mathcal{R}$1M				$\mathcal{R}$Par				+$\mathcal{R}$1M			
	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ
Offline Diffusion [44] (k_q=100)	—	200	400	15	—	60	400	15	—	200	800	5	—	200	800	5
Offline Diffusion [44] (k_q=400)	—	200	200	10	—	40	100	20	—	200	800	10	—	100	800	10
R-DiP(k_q=100)	100	60	800	10	400	100	800	10	400	200	800	5	400	200	800	5
R-DiP(k_q=400)	100	60	800	10	400	60	800	20	400	200	800	10	400	200	800	10

Method	Hard															
	$\mathcal{R}$Oxf				+$\mathcal{R}$1M				$\mathcal{R}$Par				+$\mathcal{R}$1M			
	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ
Offline Diffusion [44] (k_q=100)	—	200	400	15	—	100	400	10	—	200	800	5	—	200	800	5
Offline Diffusion [44] (k_q=400)	—	200	200	10	—	40	100	20	—	80	400	15	—	100	800	10
R-DiP(k_q=100)	400	400	400	10	400	100	800	10	400	200	800	10	400	80	800	5
R-DiP(k_q=400)	400	400	200	10	400	40	100	20	400	200	800	10	400	80	800	10

これらのデータセットはそれぞれ 70 個のクエリがあり，それぞれに異なる数の検索対象画像が割り当てられている．データベースに含まれる画像の枚数は，それぞれ 4,933 枚と 6,322 枚である．各クエリに対して，対応するデータベース画像には，検索の難易度に応じて *unclear*，*easy*，*hard* の三つのラベルが付与されている．そして，使用するラベルに基づいて三つ（Easy，Medium，Hard）の評価プロトコルに分かれる．

- Easy プロトコル： *easy* ラベルのみ
- Medium プロトコル： *easy* と *hard* ラベル
- Hard プロトコル： *hard* ラベルのみ

これまでの研究では，Easy プロトコルは既に十二分な性能が達成されており，検索手法の評価に適さないと言われているため，本研究では，Medium プロトコルと Hard プロトコルで実験を行なう．加えて，大規模なデータセットにおける検索性能を評価するために，クエリ画像に含まれない建造物や風景などの約 100 万枚の類似ドメインの画像からなる Distractor セット（$\mathcal{R}$1M）も用意されている．本実験では，Distractor セットを追加した実験も行なう．

評価指標には mAP（mean Average Precision）を用いる．

4.3 実装の詳細

本実験では，提案手法と全ての比較手法において，SuperGlobal Pooling によって抽出された 2,048 次元の大域特徴を用いる．また，提案手法における類似度再計算モデルには SuperGlobal Reranking を用いる．k 近傍検索には，Offline Diffusion 同様に Facebook AI Similarity Search（FAISS）ツールキット [15] を用いる．

各手法について，実験で用いた変数を以下に示す.

- **R-DiP**

提案手法の変数 k_m，k_d，k_t, k_q，γ は次のように設定した.

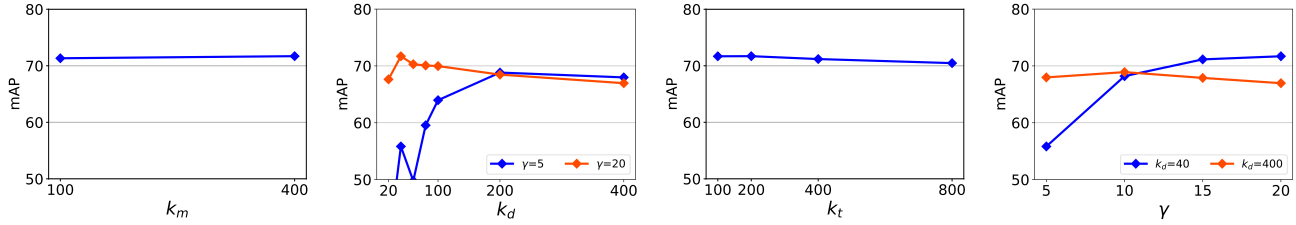
– オフライン過程における k 近傍検索による初期類似画像検索で取得する件数（Offline k-NN Size）： $k_m \in \{100, 400\}$

– オフライン過程における近傍グラフ構築時の近傍数（Offline Graph Degree）： $k_d \in \{20, 40, 60, 80, 100, 200, 400\}$

– オフライン過程における Late Truncation のための k 近傍検索での画像インデックス取得件数（Offline Truncation Size）： $k_t \in \{100, 200, 400, 800\}$

– オンライン検索時の k 近傍検索により取得する画像件数（Online k-NN Size）： $k_q \in \{100, 400\}$

– オンライン検索時の類似度に対する重み変数（Online



(a) k_m (b) k_d (c) k_t (d) γ

図 4: 各パラメータが与える影響.

表 3: クエリごとのオンライン検索時間の平均［ms］.

手法	k_q=100		k_q=400	
	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Oxf	+$\mathcal{R}$1M
Offline Diffusion [44]	**1.18**	37.21	1.31	40.18
SuperGlobal [34]	17.74	41.43	139.34	162.86
R-DiP	1.19	**37.14**	**1.30**	**40.01**

Weight)： $\gamma \in \{5, 10, 15, 20\}$

- **Offline Diffusion**

Offline Diffusion では，k_d，k_t，k_q，γ に関して提案手法と同じ変数集合を用いる．なお，Offline Diffusion において類似度再計算過程は存在しないため，k_m は存在しない．

- **SuperGlobal**

SuperGlobal では， k_q に関して提案手法と同じ変数集合を用いる．提案手法におけるその他の変数は，オフライン過程における拡散のための変数であるため，存在しない．

4.4 実験結果

提案手法の性能を評価するための実験結果を示す．

4.4.1 検索性能 (mAP) の比較

表 1 に比較手法と R-DiP の検索精度（mAP）を示す．また，各評価プロトコルにおいて，各手法で最良の結果を示した変数を表 2 に示す．提案手法は，$\mathcal{R}$Oxford5k データセットでは SuperGlobal に匹敵する性能を，$\mathcal{R}$Paris6k データセットでは SuperGlobal を上回る精度を示した．また，Offline Diffusion との比較では，ほとんどの結果において提案手法が優れた精度を示した．これは，提案手法における類似度再計算が Offline Diffusion における近傍グラフ構築を改善したことを示す．また，大規模データセットにおける実験では，提案手法はほとんどの場合で他の比較手法を上回る精度を示した．特に，$\mathcal{R}$Oxford5k+$\mathcal{R}$1M データセットにおける Hard プロトコルでは，2.0%程度の mAP の向上が見られた．Offline Diffusion もまた，大規模データセットにおける実験では，SuperGlobal を上回る精度を示しており，Offline Diffusion の良い特性を取り入れることができていることが示唆される．

4.4.2 オンライン検索時間

次にオンライン検索の実行時間を評価する．$\mathcal{R}$Oxford5k データセットと+$\mathcal{R}$1M セットのクエリあたりの平均オンライン検索実行時間（単位 ms）を表 3 に示す．提案手法は SuperGlobal と比較して実行時間を大幅に削減し，Offline Diffusion と同等の実行時間で検索できることが示された．

4.4.3 パラメータ感度分析

実験で用いる変数が提案手法に与える影響を調べるために，$\mathcal{R}$Oxford5k+$\mathcal{R}$1M データセットの Hard プロトコルにおいて，各変数に対する感度分析を行なった．

- **Offline k-NN Size（k_m）**：図 4(a) に示すように，$k_m = 100$ に比べて $k_m = 400$ の時にわずかに高い mAP を示した．理由として，k_m の値が大きいほど構築される類似度グラフの質が向上し，拡散による類似度計算の精度が向上したためであると考えられる．ただし，事前計算で用いる類似度再計算モデルによる計算コストは k_m の値に比例することをふまえると，実運用では $k_m = 100$ で十分効果的であると考えられる．
- **Offline Graph Degree（k_d）**：$k_d = 20$ の時のみほとんどの場合で最低の mAP を示した．また，k_d の変化が mAP に与える影響は，γ の値によって変わる傾向が見られた．例として，図 4(b) に $\gamma \in \{5, 20\}$ の場合の mAP を示す．γ の値が小さいほど，k_d の値が大きい場合に高い mAP を示し，γ の値が大きいほど，k_d の値が小さい場合に高い mAP を示した．
- **Offline Truncation Size（k_t）**：他の変数と比較して k_t の値の変化は mAP の大きな影響を与えなかった．また，mAP が最良となる k_t の値に対しても，特に規則性は見られなかった．例として，図 4(c) に実験で最良の mAP を示した際に k_t を変化させたグラフを示す．
- **Online Weight（γ）**：前述の通り，mAP が高い変数集合では γ の値が大きいものが多かった．また，γ の値が mAP に与える影響として，k_d の分析同様に，k_d の値が大きいほど，γ の値が小さい場合に高い mAP を示し，k_d の値が小さいほど，γ の値が大きい場合に高い mAP を示した．例として，図 4(d) に $k_d \in \{40, 400\}$ の場合の γ と mAP を示す．

4.4.4 考　　察

精度評価において，提案手法は概ね良好な結果を示し，ほとんどの場合において提案手法は Offline Diffusion を上回る精度を示した．このことは，Offline Diffusion に対して類似度再計算モデルを組み込むことの有効性を示している．

また，提案手法は大規模データセットにおいて SuperGlobal の精度を大きく上回った．SuperGlobal のようなリランキングに基づく手法は，初期の k 近傍検索における検索結果画像集合に大きく依存するのに対し，拡散に基づく手法はデータの大域

的な構造を考慮するため，画像枚数が多い場合に特にその有効性を示すという特性によるためと考えられる．

一方で，提案手法は特に $\mathcal{R}$Oxford5k データセットの Hard プロトコルにおいて， Offline Diffusion と同様に SuperGlobal に劣る精度を示した．この結果は，この状況では Offline Diffusion の有効性が低いことを示し，提案手法もその特性に影響を受けているためであると考えられる．この結果に対し，具体的に二つの理由が考えられる．一つ目は，$\mathcal{R}$Oxford5k データセットにおいて，データの枚数が少ないことにより，大域的構造がもつ情報量の少なさである．拡散は，その特性により大域的構造がもつ情報の質や量に対して大きく影響を受ける．このことにより，特に $\mathcal{R}$Oxford5k データセットの Hard プロトコルのような画像間類似度推定の難易度が高い場合において，画像間の類似度を直接計算するリランキングに基づく手法に比べて，提案手法のような拡散に基づく手法は有効性が低下したと考えられる．二つ目の理由として，Offline Diffusion の特性である，検索結果の初期の k 近傍検索結果における画像間類似度への依存性に影響を受けていることが考えられる．Offline Diffusion は事前計算した類似度に対して初期の k 近傍検索結果における画像間類似度を線型結合する．初期の k 近傍検索結果における画像間類似度により，ランダムウォークが遷移する画像群は大きく変化し，検索結果に大きな影響を与える．これにより，特に高難易度な場面において，オンライン検索で使用される k 近傍検索がボトルネックとなり，提案手法のような拡散に基づく手法の有効性が低下すると考えられる．

5 ま と め

本論文では，画像検索タスクにおける速度と精度のトレードオフに対処する，新しい画像検索フレームワークを提案した．これは，リランキング手法で用いられる高精度な類似度再計算を Offline Diffusion [44] に組み込むことで達成された．実験の結果，提案手法は $\mathcal{R}$Oxford5k [30] と $\mathcal{R}$Paris6k [30] の両データセットにおいて，Offline Diffusion と同等の検索速度を維持しながら，最先端の手法よりも優れている，もしくは同等であることが示された．大規模データセットにおける実験では，提案手法は全てのベースラインと比較して優れた結果を残し，これは，現実世界で大規模データを扱うシナリオに適しており，画像検索技術の進展に貢献することが期待される．

謝　　辞

本研究の一部は JSPS 科研費 21H03555，NII との共同研究による．

文　　献

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *Computer Vision —ECCV 2014— 13th European Conf., Zurich, Switzerland, September 6–12, 2014, Procs. Part I, Lecture Notes in Computer Science*, volume 8689, pages 584–599. Springer, 2014.

[2] B. Cao, A. Araujo, and J. Sim. Unifying deep local and global features for image search. In *Computer Vision — ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part XX, Lecture Notes in Computer Science*, volume 12365, pages 726–743. Springer, 2020.

[3] T. Dharani and I. L. Aroquiaraj. Content based image retrieval system using feature classification with modified kNN algorithm. *Computing Research Repository arXiv Preprint*, arXiv:1307.4717, 2013.

[4] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *Proc. 2013 IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA*, pages 1320–1327, 2013.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. 9th Int. Conf. on Learning Representations, Online*, 22 pages, 2021.

[6] S. R. Dubey, S. K. Singh, and W. Chu. Vision transformer hashing for image retrieval. In *Proc. 2022 IEEE Int. Conf. on Multimedia and Expo, Taipei, Taiwan*, 6 pages, 2022.

[7] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou. Training vision transformers for image retrieval. *Computing Research Repository arXiv Preprint*, arXiv:2102.05644, 2021.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[9] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision —ECCV 2016— 14th European Conf., Amsterdam, the Netherlands, October 11–14, 2016, Procs. Part VI, Lecture Notes in Computer Science*, volume 9910, pages 241–257. Springer, 2016.

[10] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *Int. J. Computer Vision*, 124(2):237–254, 2017.

[11] L. J. Grady. Random walks for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[12] Ç. Gülçehre, K. Cho, R. Pascanu, and Y. Bengio. Learned-norm pooling for deep feedforward and recurrent neural networks. In *Machine Learning and Knowledge Discovery in Databases —European Conf., ECML PKDD 2014, Nancy, France, September 15–19, 2014. Proceedings, Part I, Lecture Notes in Computer Science*, volume 8724, pages 530–546. Springer, 2014.

[13] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 596–605, 2018.

[14] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, pages 926–935, 2017.

[15] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2021.

[16] H. Jun, B. Ko, Y. Kim, I. Kim, and J. Kim. Combination of multiple global descriptors for image retrieval. *Computing Research Repository arXiv Preprint*, arXiv:1903.10663, 2019.

[17] T. Kato. Database architecture for content-based image retrieval. In *Image Storage and Retrieval Systems, Proc. SPIE*, volume 1662, pages 112–123, 1992.

[18] J. Kim and S. Yoon. Regional attention based deep feature for image retrieval. In *Proc. British Machine Vision*

Conf. 2018, Newcastle-upon-Tyne, England, UK, number 209, pages 1–13, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1106–1114, 2012.

[20] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk. Key.Net: Keypoint detection by handcrafted and learned CNN filters. In *Proc. 17th IEEE/CVF Int. Conf. on Computer Vision, Seoul, Korea*, pages 5835–5843, 2019.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[22] S. Lee, H. Seong, S. Lee, and E. Kim. Correlation verification for image retrieval. In *Proc. 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, pages 5364–5374, 2022.

[23] F. Magliani and A. Prati. An accurate retrieval through R-MAC+ descriptors for landmark recognition. In *Proc. 12th Int. Conf. on Distributed Smart Cameras, Eindhoven, the Netherlands*, number 6, pages 1–6, 2018.

[24] H. Nezamabadi-pour and E. Kabir. Concept learning by fuzzy k-NN classification and relevance feedback for efficient image retrieval. *Expert Systems and Application*, 36(3):5948–5954, 2009.

[25] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk. SOLAR: Second-Order Loss and Attention for image Retrieval. In *Computer Vision —ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part XXV, Lecture Notes in Computer Science*, volume 12370, pages 253–270, 2020.

[26] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proc. 16th IEEE Int. Conf. on Computer Vision, Venice, Veneto, Italy*, pages 3476–3485, 2017.

[27] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. LF-Net: Learning local features from images. *Advances in Neural Information Processing Systems*, 31:6237–6247, 2018.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Minneapolis, MI, USA*, 8 pages, 2007.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. 2008 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA*, 8 pages, 2008.

[30] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 5706–5715, 2018.

[31] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Computer Vision —ECCV 2016— 14th European Conf., Amsterdam, the Netherlands, October 11–14, 2016, Procs. Part I, Lecture Notes in Computer Science*, volume 9905, pages 3–20. Springer, 2016.

[32] F. Radenovic, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 41(7):1655–1668, 2019.

[33] J. Revaud, C. R. de Souza, M. Humenberger, and P. Weinzaepfel. R2D2: Reliable and Repeatable Detector and Descriptor. *Advances in Neural Information Processing Systems*, 32:12405–12415, 2019.

[34] S. Shao, K. Chen, A. Karpur, Q. Cui, A. Araujo, and B. Cao. Global features are all you need for image retrieval and reranking. In *Proc. 19th IEEE/CVF Int. Conf. on Computer Vision, Paris, France*, pages 11036–11046, 2023.

[35] O. Siméoni, Y. Avrithis, and O. Chum. Local features and visual words emerge in activations. In *Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 11651–11660, 2019.

[36] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen. Binary generative adversarial networks for image retrieval. In *Proc. 32nd AAAI Conf. on Artificial Intelligence, New Orleans, LA, USA*, pages 394–401, 2018.

[37] F. Tan, J. Yuan, and V. Ordonez. Instance-level image retrieval using reranking transformers. In *Proc. 18th IEEE/CVF Int. Conf. on Computer Vision, Montreal, QC, Canada*, pages 12085–12095, 2021.

[38] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 5109–5118, 2019.

[39] Y. Tian, V. Balntas, T. Ng, A. B. Laguna, Y. Demiris, and K. Mikolajczyk. D2D: Keypoint extraction with Describe to Detect approach. In *Computer Vision —ACCV 2020— 15th Asian Conf. on Computer Vision, Kyoto, Japan, November 30–December 4, 2020, Revised Selected Papers, Part III, Lecture Notes in Computer Science*, volume 12624, pages 223–240. Springer, 2020.

[40] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. SOSNet: Second Order Similarity regularization for local descriptor learning. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 11016–11025, 2019.

[41] G. Tolias, T. Jenícek, and O. Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision —ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part I, Lecture Notes in Computer Science*, volume 12346, pages 460–477. Springer, 2020.

[42] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Proc. 4th Int. Conf. on Learning Representations, San Juan, Puerto Rico*, 12 pages, 2016.

[43] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis. Learning super-features for image retrieval. In *Proc. 10th Int. Conf. on Learning Representations*, 19 pages, 2022.

[44] F. Yang, R. Hinami, Y. Matsui, S. Ly, and S. Satoh. Efficient image retrieval via decoupling diffusion into online and offline processing. In *Proc. 33rd AAAI Conf. on Artificial Intelligence, Honolulu, HI, USA*, pages 9087–9094, 2019.

[45] T. Yang, D. Nguyen, H. Heijnen, and V. Balntas. UR2KiD: Unifying Retrieval, Keypoint Detection, and Keypoint Description without local correspondence supervision. *Computing Research Repository arXiv Preprint*, arXiv:2001.07252, 2020.

[46] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas. Query specific rank fusion for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(4):803–815, 2015.

[47] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in Neural Information Processing Systems*, 16:169–176, 2003.

3D モデルを利用したユーザーの求めるアングル画像の検索システムの開発

永田　玄† 常　穹†† 宮崎　純††

† 東京工業大学情報理工学院情報工学系情報工学コース　〒 152-8550 東京都目黒区大岡山 2-12-1
†† 宮崎研究室　〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †nagata@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.jp

あらまし　現在画像を入力として画像を検索する画像検索ツールに関する研究が多くされている. しかし, 既存の研究の主な目的は入力した画像と色合いが同じものやカテゴリが同じものを検索することで, 角度違いのものを検索するという目的で使用することはできない. そこで, 本研究ではユーザーが入力した画像の 3D モデルを推定し, 回転させることで任意の角度からの画像を検索可能とした. また、データベース上の画像とのマッチングにおいても, 検索対象の画像の 3D モデルを推定することで角度によって色やシルエットが異なるものも検索可能とした. こうして、構築されたシステムの精度を実画像データを用いた評価実験により評価し，手法や入力画像，アングル毎に結果を確認し，提案手法について考察した．その結果, 先行研究では困難だった, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアの判別や, 逆に角度によって輪郭線や色が大きく異なるオブジェクトに対する検索も可能とした.

キーワード　画像検索エンジン, 3D モデル, Image Search Engine, 3D model

1 はじめに

情報検索システムの起源は 1970 年代にまで遡り, 大規模に蓄積される学術文献や論文等の管理をコンピュータ上で行うために, データの管理と検索が行われるようになったことで, 需要に応えるべく情報検索システムの開発が進んだ. そして今日では，インターネットの発達とスマートフォンなどの画像撮影デバイスの普及に伴い，生活者は役に立つ情報を見つけることから最寄り店舗の検索, あるいは何か新しいものを学習するためなど, あらゆる目的に情報検索を行う．そのため, 情報検索システムの利用形態の多様化に伴って, テキストによる検索だけでなく画像による検索システムも必要とされている. このような背景のもと, Google 画像検索などの画像を結果として提示する検索エンジンが普及し，多くの画像検索システムに関する研究，中でも, CBIR(content based image retrieval) と呼ばれる画像を入力とする画像検索システムの開発が盛んに行われている [4].CBIR では一般に入力された画像とデータベース上の画像から色, テクスチャ, 輪郭線などの特徴量を抽出し, 比較することで色合いが同じものやカテゴリが同じものを検索することができる. しかし, このような CBIR 手法の多くは, オブジェクトの角度に関する情報を十分に抽出することができず, 角度の違うものを検索するという目的で使用することは難しい. 従来のコンテンツベースの画像検索の例を図 1 に示す．そこで本研究では，2D 画像から 3D モデルを推定することによりこの問題を解決することを目標とする．近年, 機械学習を利用して 2D 画像から 3D モデルを推定する手法が多く開発され, 3D モデルの精度を向上している [9] [16] [7]. 本研究の目的は，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案することである．提案手法での画像検索の例を図 2 に示す．この目的を達成，発展させることで, 主にオンラインショッピングにおいて事前に商品を様々な角度から閲覧することが可能となるため, 想像していた商品と実際の商品との差異を減らすことが可能であると考える.

雛形ら [17] は同様の目的をもった研究をしている．雛形らの研究では，通常のカメラで撮ることのできない距離情報を持った画像であるデプス画像を入力画像に用いる．そこで本研究では，通常の画像データからなる写真を入力とし，より汎用性の高い検索アルゴリズムを提案する．次に, 画像中の物体を他の角度から見た画像を探すため, その物体の形状の情報，すなわちその物体の 3D モデルを推定する．また, 前川ら [18] も同様の目的をもった研究をしている．前川らの研究では, 推定された 3D モデルの形状が不正確で, 画像の曖昧な輪郭線のマッチングを図っている．そのため, 角度によって形状や色が著しく異なるものや 逆に角度によって色と形状が似ているが全くの別ものは判別が難しく, 検索結果の精度に悪影響を及ぼすことがある．そこで本研究では, 機械学習を用いたより精度の高い 3D モデル推定手法を利用して, 3D モデル間でのマッチングとランキング化を行い, 角度による形状変化に影響されないシステムを提案する．この提案手法の精度を，実画像データを用いて入力画像やアングル毎に検証する.

2 関連研究

2.1 画像検索に関する研究

一般的な画像検索手法は 2 種類ある．第一の方法は, 画像に注釈をつけるためのキーワードに基づくもので, テキストベースの画像検索として知られている．しかし, この方法には多くの欠点があり, 1) 大規模なデータベースに手作業で注釈を付け

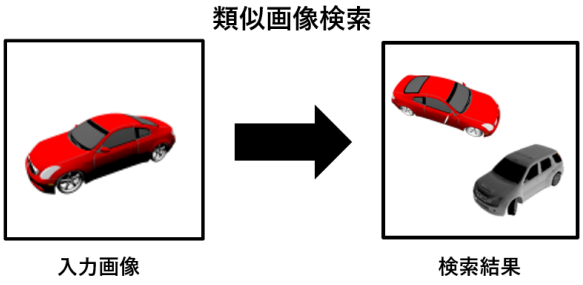


図 1　従来の検索システムのイメージ

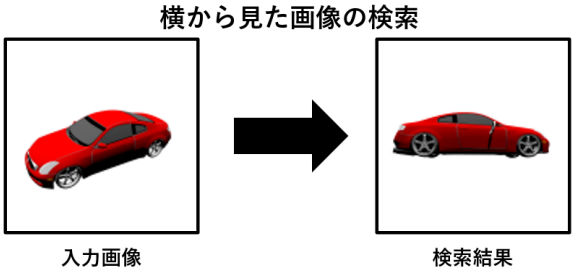


図 2　本研究の検索システムのイメージ

るのは現実的ではない, 2) エンドユーザーが注釈を付けなければならず, そのためこの方法は人間の知覚に左右される, 3) これらの注釈は 1 つの言語にしか適用できないといったものが代表される. そこで第二の方法としてコンテンツベースの画像検索 (CBIR) が提案され, テキストベースの画像検索方法の欠点を克服するために強く推奨されている [4]. CBIR で最も重要な段階は画像の特徴を抽出する段階である. 一般に, 特徴量は大域的特徴量と局所的特徴量に分類される. 大域的特徴 (色, テクスチャ, 形状, 空間情報), は画像全体の表現を与える特徴量であり, 特徴抽出や類似度計算が速いという利点がある一方, 画像内の背景と物体を区別することができないといったデメリットもある. そのため, 複雑なシーンでの検索や物体認識には適さないが, 物体の分類や検出には適している. そして, 大域的特徴と比較して局所的特徴 (コーナー, エッジ) はスケール, 回転, 平行移動, 背景の変化, 部分的なオクルージョンに対してロバストであるため, 画像検索, マッチングタスク, 認識に適している. 最近の CBIR では 2 つの特徴量を組み合わせたり, 抽出された特徴を機械学習アルゴリズムに与えることで精度の向上を図っている. Srivastava と Khare [15] は, 複数のレベルで画像を分析し, 1 つのレベルがスキップした情報を他のレベルが捕捉する, 新しいマルチ解像度分析アルゴリズムを開発した. このアプローチは、テクスチャ特徴を抽出するために LBP(local binary pattern) 記述子 [10] を使用し, テクスチャ特徴から形状特徴を抽出するために Legendre モーメントを多解像度レベルで使用する 2 つの特徴量抽出手法をベースとしたアルゴリズムを提案している. Sarwar ら [14] は, LBPV(local binary pattern variance) [3] と LIOP(local intensity order pattern) を用いて, セマンティックギャップを減らすことで CBIR 性能を向上させるシステムを提案した. このシステムではこれら 2 つの特徴記述子を 2 つの小さな視覚辞書を形成するために使用し, その後, 1 つの大きな視覚辞書を形成するために連結し, サイズを小さくするために主成分分析 (PCA) を用いてヒストグラムが計算することで, LBPV と LIOP に基づいて SVM(support vector machine) [2] を学習する.

しかし，これらのような最先端の CBIR 手法であっても，入力画像内のオブジェクトに隠れている部分が存在する場合，指定した別のアングルから見た画像を提示することは難しい．本研究は，その目的を達成することができる手法を提案することが目標である.

2.2　2D 画像からの 3D モデル推定に関する研究

深層学習の発展により, 2D から 3D モデルを再構成するための学習ベースの手法に関する研究が盛んに行われている.

Mescheder ら [9] は, 3D 再構成手法として, Occupancy Network を提案した．このネットワークでは, 3D の表面情報を，ディープニューラルネットワーク分類器の連続性決定性境界として暗黙的に表している．単一の画像からだけではなく，ノイズの多い点群や粗いボクセル状のグリッドからでも, 3D 構造をエンコードできる．また，他の既存手法よりも低いメモリ使用量での 3D モデル推定を実現させた．学習されているモデルのカテゴリは，主に家具や乗り物である.

Yao ら [16] は対称性をもつものに対する 3D 再構成手法として, Front2Back を提案した．このシステムでは, 入力画像から CNN を利用してデプス画像, 法線マップ, シルエット画像 (以降これらの画像を前面図とする) を予測する．また, 前面図より入力画像中の物体の対称面を探索し, 得られた対称面の情報と前面図より条件付き生成的敵対的ネットワークの変種 [5] を使用して, 物体を入力画像と反対の視点から見たデプス画像と法線マップ (以降これらの画像を背面図とする) を生成し, これらの前面図と背面図を用いて 3D メッシュを再構成する．この手法は対称性をもつものに対しては Occupancy Network よりも高い精度をもつ.

Long ら [7] は多視点画像から 3D 再構成する手法として, Wonder3D を提案した．このシステムでは, 入力画像からマルチビューのカラー画像と法線マップを生成し, SDF(Signed Distance Field) と呼ばれる 3D 空間内の点に対して, その点が物体の表面からどれだけ離れているかを示す関数を深層学習モデルで学習し, 3D メッシュを再構成する．このアプローチは 2D 画像から 3D モデルを推定に当たり問題となる忠実性, 一貫性, 一般化可能性, 効率性に対処するように設計されている．実際に Wonder3D により推定した 3D モデルを図 3,4 に示す．図 3 に示したモデルから Wonder3D により推定されたモデルは同じカテゴリのものでもその違いを表すことができるほど精度は高い．また, 自動車のドアミラーや椅子, 机の空洞部など細かい部分も表現することができる．しかし, 入力画像からマルチビューのカラー画像と法線マップを生成するため, 図 4 のように奥行が分かりにくい画像や画像中にオクルージョン部分が存在すると上手くモデルを生成することは難しい.

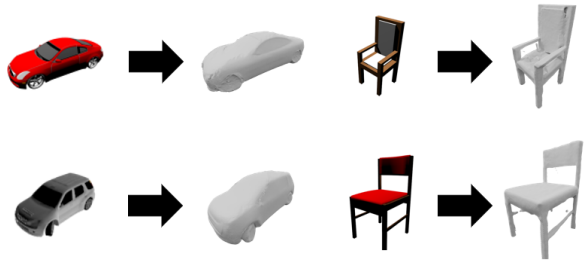


図 3　Wonder3D で推定した 3D モデルの例

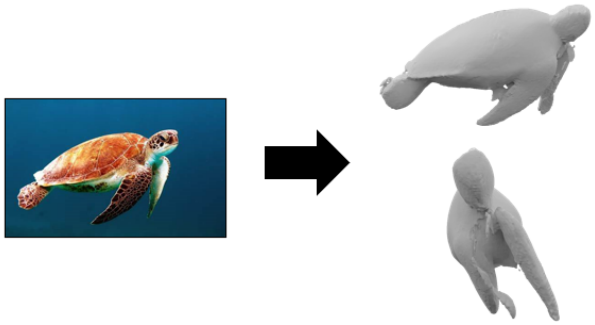


図 4　Wonder3D で推定した 3D モデルの失敗例

2.3　3D モデルの特徴量に関する研究

2 つの 3D モデルの類似度を調べる代表的な手法として, 特徴量によるマッチングが挙げられる. 3 次元データの特徴量に関する研究は盛んに行われている. [8]

Kazhdan ら [6] は大域的特徴量として Spherical Harmonics を提案した. この手法は, 任意の物体の 3 次元形状を球面上のある関数と見なし, これを有限個の基底ベクトルの線形結合で近似するものである.

Rusu ら [12] は FPFH(fast point feature histograms) と呼ばれる PFH(point feature histograms) [13] を高速に改良した局所特徴量を提案した. ある点 $\boldsymbol{P}$ を表す PFH は, その点を中心とした小球領域に含まれる k 個の近傍点を求め, それらの点から 2 点を選ぶすべての組み合わせ $(\boldsymbol{P}_i, \boldsymbol{P}_j)(i \neq j, 1 \leqq i, j \leqq k)$ に対して計算される 3 つのパラメータのヒストグラムを計算する. このヒストグラムの計算量は $O(k^2)$ となり, サンプリング数によって計算量が指数関数的に増加する. 一方, FPFH の計算量は O(k) に改良されており, PFH に比べてかなり高速な処理となっている.

しかしこれらの特徴量は 3D モデルの回転に対して不変性を示すため, 本研究には適用することができない. そこで本研究では 3D モデルの姿勢に依存する類似度測定手法として, CD(Chamfer Distance) を利用する.

CD は一方の点が他方の点にどれだけ近いかを示す指標であり, 各点がもう一方の点集合の最も近い点までの距離を計算し, その距離の合計を取ることで定義される. そのため, 3D モデルの回転も考慮した姿勢と形状の類似度を比較することができ, 本研究に適している. 具体的に点群 X と点群 Y 間の CD を計算する式は以下のように定義される.

$$CD(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2 + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\|x-y\|_2^2 \tag{1}$$

具体的な処理は以下のようになる.

1. 点群 X 内の点 x と, 点群 Y の全ての点のユークリッド距離を算出し, 2 乗する.
2. 1 で求めた 2 乗ユークリッド距離の最小値を, 加算する.
3. 加算した最小距離の平均を計算する.
4. 同様の処理を点群 Y で行う (こちらの場合は点 y と点群 X の全ての点を比較する).
5. 算出した点群 X における最小距離と点群 Y における最小距離を加算する.

2.4　先 行 研 究

雛形ら [17] や前川ら [18] は，本研究と同様に，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案している．雛形らは入力画像として，デプス画像と呼ばれる物体までの距離情報を持った画像を用いて，そのデプス画像から復元した部分モデルと完全な 3D モデルのマッチング精度を評価している．本研究では，デプス画像を普通のカメラで撮影できず，今日社会の中で普及していないことに着目し，同様の目的を普通の画像をクエリとして達成可能にすることを目標としている．前川らは入力した画像から 3D モデルを推定し，それをユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にしている．しかし, 推定された 3D モデルは形状が不正確であり，色も単色で使用できず，データベース上の画像と比較する際は画像の輪郭線と回転させた 3D モデルの曖昧な輪郭線を比較している．前川らの使用した 3D モデルの例 [18] を図 5 に示す．前川らの研究では, 輪郭線を比較するにあたって, Oliva ら [11] が提案した GIST 特徴量を用いてを用いて, 入力画像とデータベース内の全ての画像の類似度を比較し, 同じ物体が写っている画像を事前に抽出することを考えている．しかし, GIST 特徴量は画像の雰囲気を表現し, シーン認識に特化した特徴量であるため, 角度によって形状や色が著しく異なるものは抽出することができず, 逆に角度によって色と形状が似ているものは全く違うものであっても抽出されてしまい, 検索結果の精度に悪影響を及ぼすことがある．前川らの提案したシステムの苦手とする検索例 [18] を図 6 に示す．そこで, 本研究ではより精度の高い 3D モデル推定を行うことができる手法を利用して, 入力画像とデータベース内の全ての画像の 3D モデルを推定し, 3D モデル同士で特徴量マッチングをすることで, 角度による形状の違いが検索結果へ及ぼす悪影響を抑制する.

図 5　前川らの使用した 3D モデル ([18] より抜粋)

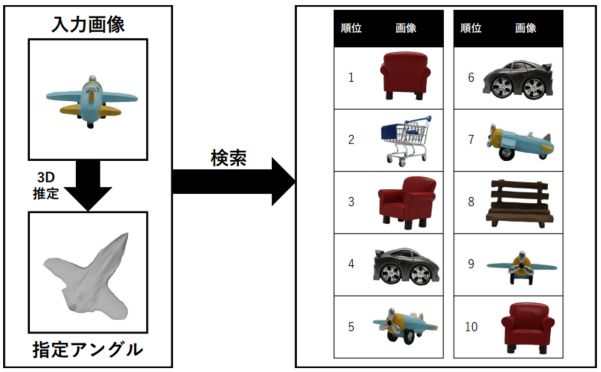


図 6　前川らの提案したシステムで飛行機を上から見た画像を検索した結果 ([18] より抜粋)

3　提 案 手 法

3.1　システムのアーキテクチャ

ここでは，本研究で提案するシステムのアーキテクチャについて包括的に述べる．本研究で提案する検索システムの全体像を図 7 に示す．このシステムのフローは, 大きく以下の 4 つに分けられる．

1. 事前にデータベースの画像の 3D モデルを推定
2. 入力画像から 3D モデルを推定し，3D モデルをユーザーが任意のアングルへ回転
3. 回転した 3D モデルとデータベースの 3D モデルで特徴量マッチング
4. 特徴量マッチングに基づいてランキング化

あらかじめデータベース内のすべての画像から 3D モデルを推定し, 3D モデルデータベースとして保存しておく．入力は, JPG や PNG などの一般的な画像データ 1 枚のみであり, 入力された 1 枚の画像から 3D モデルを推定し, ユーザーが要求する角度へと回転させる．次に, 3D データベース上の 3D モデルを 1 つずつ選択し, ユーザーによって回転させておいた 3D モデルとの類似度を比較する．最後に類似度を昇順に並べ, データベース上の画像をランキング化して提示する．それぞれの内容に関して詳しく，以降の節で述べる．

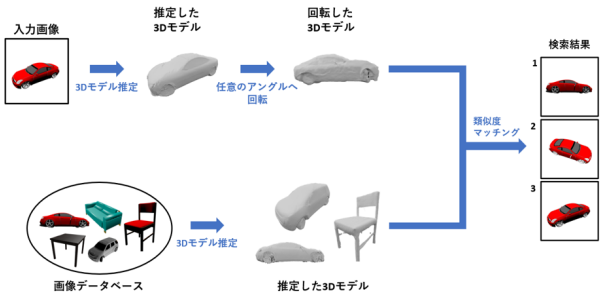


図 7　提案するシステムのアーキテクチャ

3.2　2D 画像からの 3D モデル推定

本手法では，入力した画像とデータベース上の画像から 3D モデルを推定し, 入力画像の 3D モデルはユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にする．

深層学習の発展により, 2D から 3D モデルを再構成するための学習ベースの手法に関する研究が盛んに行われている．Mescheder ら [9] の Occupancy Network は事前にオブジェクトを学習する必要があり, また, Yao ら [16] の Front2Back は対称性をもたないオブジェクトに対する 3D モデルの精度が低い．本研究では, スマートフォンを利用する人が老若男女問わず, また情報検索を必要とするユーザーは特定の分野に限らないことから両者のように特定の条件でのみ使用できる 3D モデル推定手法を利用することはできない．そのため, 事前に学習する必要がなくモデルの生成範囲に制限がない, Long [7] の提案する Wonder3D を採用している．

3.3　特徴量マッチング

提案する検索システムでは, 入力した画像やデータベース上の画像から 3D モデルを推定し, 入力画像の 3D モデルをユーザーの任意に回転させることで検索したいアングルを指定する．その後, 各 3D モデルの姿勢も考慮した特徴量マッチングをし比較する．

3D モデルの特徴量には Spherical Harmonics [6], FPFH [12] など様々な種類がある．これらは全て 3D モデルの回転に依存しない特徴量である．本研究ではオブジェクトの回転を含めた姿勢に対して類似度を計算する必要があるため, これらの特徴量は本研究には適さない．よって本研究では, Chamfer Distance(CD) と呼ばれる 2 つの 3D 点群間の位置誤差を測る指標を用いる．CD の値は小さいほど 2 つの 3D モデルの類似度が高いことを示す．本研究では 3.2 節で推定したデータベース上の画像の 3D モデルから 1 つずつ選択し, ユーザーの要求する角度へと回転させた入力画像の 3D モデルとの CD を計算する．最終的にデータベース上の画像のすべての 3D モデル間で計算された CD を昇順に並び替え, 対応するデータベース上の画像をランキングとして出力する．

4 評価実験

4.1 実験用データと前処理

本手法の実験では，複数のアングルから見たオブジェクトの画像が必要になる．よって，事前に用意した 13 個のオブジェクトについてそれぞれ 4 アングルから撮影し，データセットとして 52 枚の画像を用意した．撮影したオブジェクトと，実験のため付けた名前を図 8 に示す．この時，撮影した 4 アングルは以下の通りである．なお, オブジェクトを正面から見たときに左方向を x 軸, 上方向を z 軸, 手前方向を y 軸の正の方向とし, 各軸を正の方向から見た場合の反時計回りを回転の正の方向とする．

- 前：正面から見た画像 (以降「f」とする)
- 横：正面から見てオブジェクトを z 軸周り-90 回転させた画像 (以降「s」とする)
- 後ろ：正面から見てオブジェクトを z 軸周り 180 回転させた画像 (以降「b」とする)
- 上：正面から見てオブジェクトを x 軸周り-90 回転させた画像 (以降「u」とする)

また, 図 9 に示すような角度によって色や輪郭線が 52 枚の画像の一部と類似するオブジェクトの画像も画像データベースに加えた．それぞれのオブジェクトについて，スマートフォンで写真を撮影し，背景を取り除いた．それぞれの手法の効果を正確に比較できるよう，様々なジャンルのものや，色や形状，ジャンルが似ているものなど幅広く用意した．さらに, これらの 59 枚の画像すべての 3D モデルを事前に推定し, 3D データベースとして格納した．

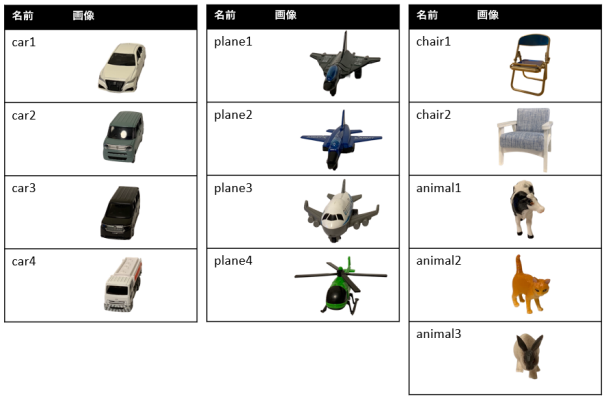


図 8 実験のために撮影したオブジェクト

4.2 実験方法

4.2.1 3D モデル推定

本手法では，入力した画像から 3D モデルを推定し，それをユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にする．ここでは，実際に 3D モデルを推定した方法を述べる．入力は撮影した画像から背景を取り除いた PNG 画像 1 枚のみである．Long ら [7] の提案する

図 9 角度によって色や輪郭線が類似するオブジェクト

Wonder3D で，3D モデル化した．この時，Wonder3D では奥行がオクルージョン部分となり, 適切に 3D モデル化できないオブジェクトがあった．そのうちの一部を図 10 に示す．また, 各オブジェクトの前アングルから推定した 3D モデルを図 11 に示す．実験では，これらの 3D モデルを入力として用いた．

図 10 奥行が適切に 3D モデル化できていないオブジェクトの例

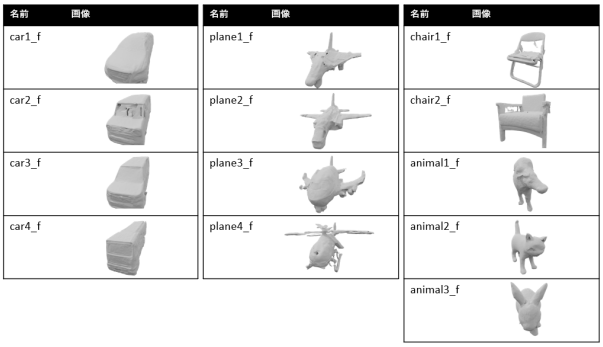


図 11 実験の入力に用いた正面画像の 3D モデル

4.2.2 指定アングル

アングルの指定には，図 11 に示した各オブジェクトの正面の 3D モデルを入力として用いた．アングルには, 1 つのオブジェクトに対して，z 軸方向-30 刻みにアングルを指定し ，実験した．

4.2.3 特徴量マッチング

本手法では推定した 3D モデルを任意のアングルに回転し，そのアングルから見た画像を検索する．その際，回転させた 3D モデルとデータベースの画像の 3D モデルの特徴量を比較し，近いものをランキングの上位に出力する．特徴量マッチングで

は，CD による特徴量マッチングを用いた．データベース内の画像 59 件に対してマッチングを行い類似度の高かった画像の上位 10 件をランキングとして出力し，評価した．

4.3 評価方法

検索結果の評価には，nDCG (normalized Discounted Cumulative Gain) を用いた．nDCG は，予測されたランキングの評価指標である．nDCG は 0 から 1 の値を取り，1 に近いほど正しいランキングであることを示す．nDCG は，DCG (Discounted Cumulative Gain) を，予測ランキングの DCG を正しいランキングの DCG で割ることにより，正規化した値である．本研究では Burges ら [1] が提案した DCG を用いた．Burges らが提案した DCG は以下の式で定義されている．

$$DCG(k) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2}$$

そして，nDCG は，DCG を考えられる最大の DCG($DCG_{perfect}$) で正規化した値で，0 から 1 の間の値をとる．

$$nDCG = \frac{DCG}{DCG_{perfect}} \tag{3}$$

rel_i は，ランキングの i 番目の要素の適合度を表しており，今回の実験では i 番目の画像が以下の条件を満たす数を rel_i とする．

- 入力した画像のジャンルと同じジャンル (例: car1_f を入力とした場合，「car」であるどうか)
- ジャンルが一致している条件のもと，番号が一致する (例: car1_f を入力とした場合，「car1」であるかどうか．chair1 や animal1 は無効とする)
- ユーザーの要求する角度に対して ±30 度以内の画像かどうか (例: car1_f を入力とし，-30 度の z 軸回転をした場合，0 度の画像であるかどうか)

また，k は評価に用いる要素数を表すため，今回はランキングの上位 10 件を評価しているので，k は 10 である．

4.4 評価結果

13 個の 3D モデルをそれぞれ 7 つのアングルで，59 件の画像から検索した結果の nDCG を図 12 に示す．具体的なランキングの結果をいくつか図に示す．図中の赤丸は正解となる画像を示す．

図 15 は car1_f を入力として，z 軸の回転角を-180 度とした場合の検索結果であり，検索対象は car1_b である．図 15 は正解となる画像が 3 位となっている．

図 17 は plane3_f を入力として，z 軸の回転角を-90 度とした場合の検索結果であり，検索対象は plane3_s である．図 17 は正解となる画像が 1 位となっている．

4.5 考察

4.5.1 入力画像ごとの検索結果に関する考察

表より入力画像によって nDCG の値にばらつきがあることが分かる．car1～car4 では nDCG の平均値が 0.64～0.82, chair1

入力オブジェクト \ z軸の回転角(度)	0	-30	-60	-90	-120	-150	-180	平均
car1_f	0.92	0.77	0.71	0.64	0.6	0.56	0.79	0.71
car2_f	0.89	0.7	0.81	0.78	0.81	0.86	0.91	0.82
car3_f	0.88	0.65	0.77	0.7	0.56	0.66	0.66	0.7
car4_f	0.9	0.59	0.58	0.69	0.49	0.43	0.77	0.64
plane1_f	0.81	0.72	0.22	0.3	0.34	0.19	0.35	0.42
plane2_f	0.89	0.69	0.17	0.35	0.37	0.12	0.35	0.42
plane3_f	0.8	0.34	0.17	0.76	0.51	0.15	0.32	0.44
plane4_f	0.91	0.71	0.61	0.29	0.34	0.26	0.47	0.51
chair1_f	0.79	0.77	0.58	0.73	0.82	0.57	0.64	0.7
chair2_f	0.76	0.76	0.51	0.61	0.6	0.76	0.71	0.67
animal1_f	0.69	0.77	0.27	0.15	0.24	0.2	0.35	0.38
animal2_f	0.75	0.5	0.04	0.61	0.76	0.12	0.64	0.49
animal3_f	0.68	0.69	0.39	0.39	0.29	0.42	0.43	0.47
平均	0.82	0.67	0.45	0.54	0.52	0.41	0.57	

図 12　nDCG による評価結果

と chair2 の nDCG の平均値は 0.685 と両者の nDCG の値は比較的高いことが分かる．しかし，plane1～plane4 では nDCG の平均値が 0.42～0.51, animal1～animal3 では nDCG の平均値が 0.38～0.49 であり，car や chair と比べると低いことが分かる．これは，Wonder3D によって生成された 3D モデルのアングルによる差にあると考えられる．car や chair で生成された 3D モデルは正面，横，後ろからそれぞれ生成された 3D モデルの形状が類似しており，animal や plane ではそれぞれの角度から生成された 3D の形状に差がある．実際に car2 と animal1 の正面，横，後ろから生成された 3D モデルを図 13 に示す．car2 から生成された 3D モデルはすべての角度でほぼ正確に car2 が表現されているのに対して，animal1 から生成された 3D モデルは特に正面から見た 3D モデルで animal1 の体部分の長さが実際のものとは異なっていることが分かる．

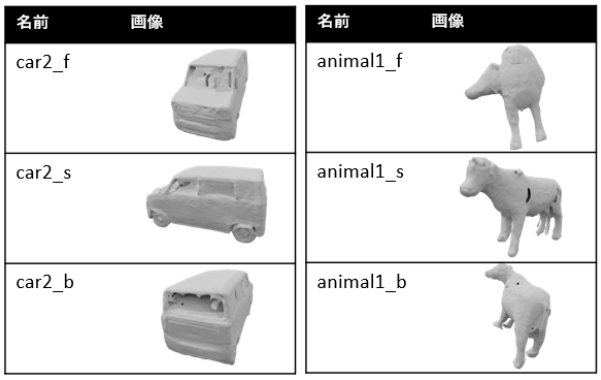


図 13　car2 と animal1 を正面，横，後ろから見た画像から生成された 3D モデル

4.5.2 各アングルごとの検索結果に関する考察

表より z 軸の回転角ごとに nDCG の値にばらつきがあることが分かる．図 14 に z 軸の回転角と nDCG の平均値グラフを示す．0 度や-30 度で比較的高い数値を記録していることが分かる．これは単に入力した画像と検索対象となる画像が同じであるため，3D モデルで比較する際にも差が小さいためであると考えられる．次に数値の高い回転角は-180 度，-90 度，-120 度の順である．-90 度や-180 度はそれぞれその角度から生成された 3D モデルが存在するためだと考えられる．-120 度は-60 度と条件がほぼ同様であるにも関わらず，差が出ている．しかし，入力画像によっては-60 度のほうが-120 度のほうが高いものもある．-60 度と-120 度に差が出る原因は検索対象の 3D モデルよ

りも, 入力画像をそれぞれの角度分回転させた 3D モデルに近い 3D モデルの数に差があるためである. そのため, 今回使用したデータベースでは-60 度よりも-120 度回転させた 3D モデルに近い 3D モデルが全体的に多かったのだと考えられる. しかし, 別のデータセットを使用した場合, この結果は逆転する可能性があるため, 明確に提案するシステムが-60 度よりも-120 度で検索したのほうが優れているとは言い切れない.

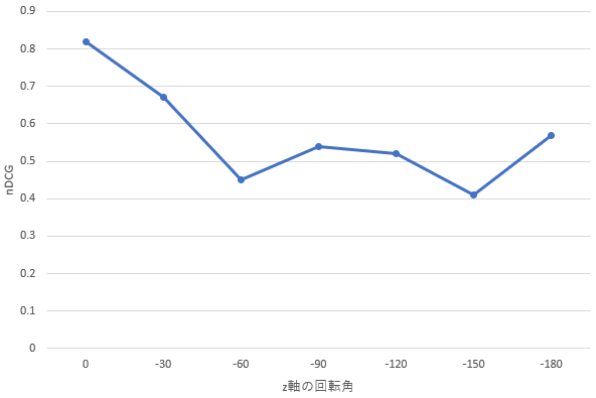


図 14　z 軸の回転角ごとの nDCG の平均値

4.5.3　角度によって輪郭が類似するオブジェクトに関する考察

今回のデータセットには図 8 の 13 個のオブジェクトに加えて, 図 9 に示す 7 個のオブジェクトを加えた. これらのオブジェクトの輪郭線や色に注目すると, black1～black3 は car3 に, white1 から white4 は car1 に類似していることが分かる. このように角度によって輪郭線や色は類似するが全く別のオブジェクトのペアがデータセットに存在する場合, 前川らの提案する輪郭線によるマッチングでは上手くランキング化することができない場合がある. しかし, 本研究では図 15 に示すように, このような画像がデータベースに存在する場合でも正確にランキング化することができることが分かる. 図 15 は car1_f を入力として, z 軸の回転角を-180 度とした場合の検索結果であり, 検索対象は car1_b である. white2 や white3 は car1_b と色や輪郭線が類似しているが図 15 は正解となる画像が 3 位となっており, white2 や white3 はランク外となっている. このように, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアがデータセットに存在する場合, 前川らの提案するシステムよりも本研究のほうが優れていることが分かる.

4.5.4　角度によって形が大きく異なるオブジェクトに関する考察

図 6 のように前川らの研究では角度によって形が大きく異なるオブジェクトに対する検索は上手くいかないことがある. 本研究で使用した plane3 の正面, 横, 後ろから撮影した画像をそれぞれ図 16 に示す. 図 16 に示すように plane3 は正面や後ろから見た場合には胴体と羽で十字のような輪郭線となっているが, 横から見た場合には羽部分が輪郭線には現れず, 胴体部分のみであるため car を横から見た画像と同じような輪郭線となる. 図 17 に plane3_f を入力として, z 軸の回転角を-90 度とした場

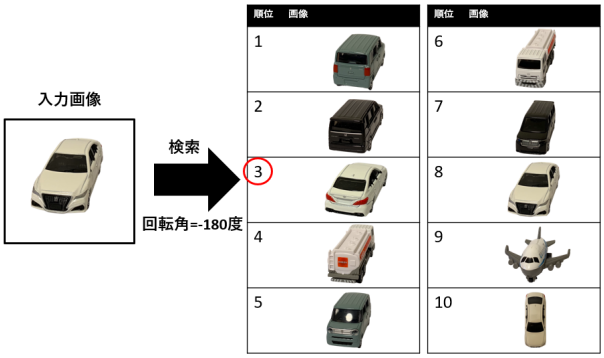


図 15　入力: car1_f, z 軸の回転角: -180 度としたときの検索結果

合の検索結果を示す. この場合検索対象の画像は plane3_s である. 図 17 に示すように正解となる画像が 1 位となっている. このことから, 本研究では前川らのシステムが苦手としている角度によって形が大きく異なるオブジェクトに対する検索が可能だということが分かる.

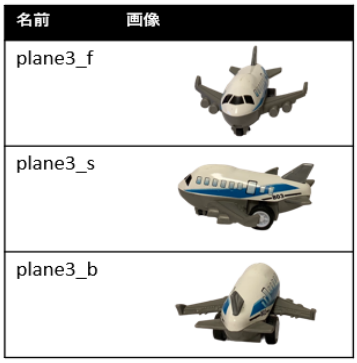


図 16　plane3 を正面, 横, 後ろから撮影した画像

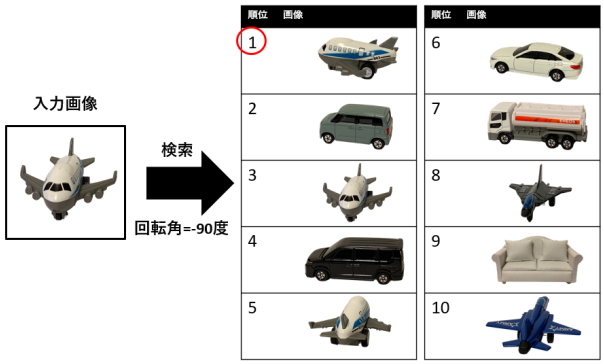


図 17　入力: plane3_f, z 軸の回転角: -90 度としたときの検索結果

5　おわりに

本研究では，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案した. 従来の画像検索システムでは，画像を入力して，その画像内のオブジェクトをユーザーの要求するアングルから見た画像をピンポ

イントで検索するといった使い方が困難である．同様の目的をもった研究 [17] [18] では，一般的ではないデプスカメラの利用や、精度の低い 3D モデルの推定により、検索精度の低さが問題点となっていた。そこで本研究では，通常の写真画像を入力とし，より精度の高い Wonder3D [7] を用いて, 入力画像から 3D モデルを推定した．そして, 入力画像の 3D モデルを回転させることで，検索したいアングルを指定することに成功した．さらにデータベース上の全ての画像の 3D モデルも推定し, 回転させた入力画像の 3D モデルと 3D モデルによる特徴量マッチングをすることで, 要求されたアングルの写真を検索結果として提示するシステムを構築した．こうすることで, 輪郭線マッチングでは判定することが困難だった, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアも判別や, 逆に角度によって輪郭線や色が大きく異なるオブジェクトに対する検索も可能とした．この提案手法を，実画像データを用いた評価実験により検証した．その際，z 軸の回転角を 30 度ごとに変更して実験することで各角度ごとの得意不得意を比較した．実験の結果，生成した 3D モデルを正確に検索する角度の場合精度が向上することがわかった．また, それぞれのオブジェクトについて各角度から生成された 3D モデルの差異が小さいほど検索結果の精度が向上することも分かった．

評価の結果生成された 3D モデルの精度が検索結果に大きな影響を及ぼすことが分かった．本研究で用いた, 2D 画像から 3D モデルを推定する手法 Wonder3D は, 精度が高く, 事前に学習する必要がないため, 広範囲の 2D 画像を 3D モデル化することが可能であったが, 入力された画像からそのオブジェクトの多視点画像を推測するため, 入力画像の角度によっては推定される 3D モデルの奥行がうまく表現されないことがあった．そのため, 実験で用いたオブジェクトの一部は正しく 3D モデル化できずに検索結果が落ちることとなった．今後は 3D モデルに適した軸でスケーリングをする, または同様の精度で角度による影響を受けない 3D モデル手法を検討する必要がある．また, 本手法ではデータベース上の画像をすべて 3D モデル化する必要があるため, 計算コストが膨大になってしまった．そのため, 今後本研究をウェブサービスとして実現するためには, より早く 3D モデルを推定する手法を検討する, そして並列して 3D モデルを推定することができる環境が必要であると考えられる．

謝　　辞

本研究は, JSPS 科研費 JP23H03401, JP23H03694 の助成を受けたものである．

文　　献

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

[2] Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[3] Zhenhua Guo, Lei Zhang, and David Dian Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *Pattern Recognit.*, Vol. 43, pp. 706–719, 2010.

[4] Ibtihaal M. Hameed, Sadiq H. Abdulhussain, and Basheera M. Mahmmod. Content-based image retrieval: A review of recent trends. *Cogent Engineering*, Vol. 8, , 2021.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[6] Michael M. Kazhdan, Thomas A. Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Eurographics Symposium on Geometry Processing*, 2003.

[7] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

[8] Graciela Lara López, Adriana Peña Pérez Negrón, Angélica de Antonio Jiménez, Jaime Ramírez Rodríguez, and Ricardo Imbert. Comparative analysis of shape descriptors for 3d objects. *Multimedia Tools and Applications*, Vol. 76, pp. 6993–7040, 2017.

[9] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

[10] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 24, pp. 971–987, 2002.

[11] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, Vol. 42, No. 3, p. 145–175, may 2001.

[12] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. *2009 IEEE International Conference on Robotics and Automation*, pp. 3212–3217, 2009.

[13] Radu Bogdan Rusu, Nico Blodow, Zoltán-Csaba Márton, and Michael Beetz. Aligning point cloud views using persistent feature histograms. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3384–3391, 2008.

[14] Amna Sarwar, Zahid Mehmood, Tanzila Saba, Khurram Ashfaq Qazi, Ahmed Adnan, and Habibullah Jamal. A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science*, Vol. 45, pp. 117 – 135, 2018.

[15] Prashant Srivastava and Ashish Khare. Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. *J. Vis. Comun. Image Represent.*, Vol. 42, No. C, p. 78–103, jan 2017.

[16] Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, and Alla Sheffer. Front2back: Single view 3d shape reconstruction via front to back prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 531–540, 2020.

[17] 雛形奎祐, 欅惇志, 宮崎純. 他視点のオブジェクトを検索可能な画像検索システムの構築. DEIM Forum 2018 論文集, 2018.

[18] 前川由依, 宮崎純. クエリ画像から別アングル画像を検索可能なシステムの構築. 東京工業大学修士論文, 2022.

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 10:00 ~ 12:10 | T4-A オンライン（Zoom Events)

LLM x ユーザ支援

座長:波多野 賢治(同志社大学)
コメンテータ:大島 裕明(兵庫県立大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T4-A-4-01] ブレインストーミングにおける消極的参加者のための大規模言語モデルに基づく発言促進手法

*猪股 朋果[1]、大岡 麗[1]、北山 大輔[1] (1. 工学院大学)

10:20 ~ 10:45
[T4-A-4-02] 意外な単語の組み合わせを用いた発言生成に基づくブレインストーミング支援

*大岡 麗[1]、関 峰[2]、北山 大輔[1] (1. 工学院大学情報学部システム数理学科、2. 工学院大学工学研究科情報学専攻)

10:45 ~ 11:10
[T4-A-4-03] ChatGPTを利用した学術論文読解支援の一手法

*堺 陸斗[1]、金澤 輝一[2]、上野 史[1]、太田 学[1] (1. 岡山大学、2. 国立情報学研究所)

11:10 ~ 11:35
[T4-A-4-04] 大規模言語モデルを用いたクエリ拡張による会話ログの検索

*田貝 奈央[1]、加藤 誠[1] (1. 筑波大学)

ブレインストーミングにおける消極的参加者のための大規模言語モデルに基づく発言促進手法

猪股　朋果† 　大岡　麗† 　北山　大輔†

† 工学院大学情報学部　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j321037,j320044}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　会議に参加した人たちが，自由な発言で意見を出し合い，新しいアイデアを生み出す手法としてブレインストーミングがある．ブレインストーミングでは，以下の 4 つの要素が重要と考えられている．1) 他人のアイデアの否定をしない，2) 自由に発言する，3) 質より量を求める，4) アイデアを連想、結合する．しかしながら，積極的な発言を得意とする参加者ばかりではなく，貢献はしたいが消極的な参加者も存在する．このような消極的な参加者は要素 3)4) においての貢献が低くなりがちである．そこで本研究では，消極的な参加者の発言を促進するために，顕在的/潜在的な発言のヒントを大規模言語モデルを用いて生成する手法を提案する．本稿ではプロトタイプシステムを構築し，その動作についてと, 支援内容が効果的であるかの検証結果の評価と意見数について報告する. 結果として, 発言促進に対する支援は効果的であることが分かったが, 発言生成に対する支援は意見数の増加は見られたものの低評価であった.

キーワード　ユーザ支援，インタラクティブシステム，ブレインストーミング，bot

1　はじめに

商品開発や研究テーマの決定などの場面で新しいアイデアを必要とする場面は多い．このような場面で新しいアイデアを生み出すために，多人数で行うブレインストーミングが一般的に使用されている．多人数による議論の場では，積極的に発言を行うことを得意とする参加者と不得意な参加者が出てくる．本稿ではこのような不得意な参加者を「消極的参加者」と定義する．消極的な参加者に対する支援のとしては，塩津ら [1] が取り組んでおり，対面の議論の場において，LED ライトの点灯によって発言を聞きたい旨や，発言を支持する旨を通知する仕組みを提案している．我々は，このような参加者同士の暗黙的な意図を伝えるのではなく，消極的な参加者が自発的に発言できる環境を支援する仕組みに着目している．

ブレインストーミングについて Osborn [2] は,「批判厳禁」,「自由奔放」,「アイデアの量」，および「連想と結合」という 4 つの原則を提唱している．これらのうち，本研究では消極的参加者に関する項目として「アイデアの量」「連想と結合」に着目する.「アイデアの量」とは，アイデアの質よりも多くのアイデアを生み出すことを重視する考え方である．消極的参加者は，発言の量自体が少なくなりがちであり，アイデアの量に貢献しにくい．ただし，消極的参加者は，議論に貢献したくないわけでも，考えがないわけでもなく，きっかけを与えることで，アイデアを含む発言ができると考えられる．そのため，発言のきっかけを与える支援が有効である.「連想と結合」は，既存の案を用いて新たな案を導き出す原則である．消極的参加者は自分の考えをまとめている間に，次の話題に行ってしまうために発言しそびれていることが考えられる．このような場合に，過去の発言のまとめや引用を場に提示することで，発言の機会を呼び起こすような支援が有効であると考えられる．

そこで本研究では，参加者が自発的に発言したい，発言のヒントが欲しいと思ったタイミングで，発言のきっかけを与えてくれるチャットボットを提案する．このチャットボットでは，発言したいタイミングで，明示的に自分を発言者として指名する機能や，これまでの議論のまとめや引用を場に提供する機能を実装する．このことにより，消極的参加者の自発的な発言を促す支援を実現する．

2　提案手法

2.1　発言促進の基本的な流れ

本研究では，テキストチャットを用いたブレインストーミングを想定している．参加者は各々のタイミングで発言を書き込むことができる．なお，図表の利用は想定しておらず，ホワイトボードや付箋等を利用した形式は今後の課題である．このような環境下において，提案するチャットボットは以下のような挙動をする．

- 定期的に発話者を指定する．原則としてランダムに決定するが,「ランダムに決定した旨」を併記する場合としない場合がある．
- 参加者が要請したタイミングで，参加者全員に見える形で発言する．発言内容は次節以降で説明する．
- 参加者が要請したタイミングで，要請者のみに見える形で発言する．同様に発言内容は次節以降で説明する．

消極的な参加者としては，発話促進の支援を受けたいときに，自身が明示的に支援を受けたことを周囲に明らかにしたくないと考えられる．そのため，2.2 節の機能を使ったことがわかりにくくなるように，1 つ目の挙動を設計している．2 つ目の機能は要請があったこと自体はわかるが，誰が要請したのかはわ

からないように設計している．3 つ目については，そもそも要請があったこと自体が参加者からはわからないように設計している．

発話の生成には OpenAI が提供している大規模言語モデル（Large Language Models）である GPT[1]を用いる．Model は gpt-3.5-turbo-1106 使用し，パラメータである temperature は 0 とした．入力するプロンプトは次節以降で説明する．

2.2 発言のきっかけの生成

消極的な参加者は，発話の意思があり，発話する内容がある場合においても，発言を控えてしまうことがある．このような場合は，第三者に発言を促されることで，発話機会が与えられることが望ましい．そこで，参加者がチャットボットに要請を送ることで，チャットボットは以下のプロンプトで発言を生成する．また，すべてのプロンプトには「あなたは司会者です．参加者に次世代の机について意見を聞きます．司会者としての発言はなるべく簡潔に行ってください.」が冒頭に付与される．

prop1　【要請者】を指名してください．

prop2　【要請者】を指名してください．ただし，指名はランダムに決定した旨の補足をしてください．

【要請者】の部分には，要請を出したユーザ名が挿入される．1 つ目の機能の場合，自分で要請して自分を指定しているため，他の参加者からはわからないとは言え，強い拒絶を覚えるユーザもいると考えた．そこで，ランダムである旨を強調することで心理的負担が軽減されることを狙い 2 つ目の機能を用意している．どちらの機能を使うのかは，要請者が選択するものとする．

2.3 既存の発言を追及するヒントの生成

消極的な参加者は，議論への貢献をしたいが，発話内容が浮かばずに発言できない場合もある．そこで，発話のヒントとなる情報をチャットボットに発言させることで，ユーザが発想を刺激され発話しやすくなるような支援を考える．

まず，ユーザが特に注目する発言がないが，何か発言のヒントが欲しい場合についてのプロンプトを説明する．

prop3　これまでの発言の中心となっているトピックを明示的に取り上げた後，それについて参加者に意見を求めてください．

prop4　これまでの発言に含まれるトピックについて，その発言量を付与して列挙してください．その後，それらについて参加者に意見を求めてください．

この時，プロンプトとしては，これまでの発話履歴も同時に与えるものとする．1 つ目の機能では中心となっているトピックについて明示的に意見を聞くことを目的としている．2 つ目の機能ではこれまでのトピックの分析結果だけを伝え，どのトピックを選ぶかは参加者にゆだねて判断してもらうことを目的としている．この機能はランダムにどちらかが選ばれて実行される．

次に，ユーザが特に注目する発言があった上で，何か発言のヒントが欲しい場合についてのプロンプトを説明する．以下の機能も同様に，他の参加者から見える形で発言させることも，自分だけが見える形で発言させることもでき，それは要請者が任意に決定する．

prop5　【指定の発話】という発言からトピックを明示的に取り上げた後，それについて参加者に意見を求めてください．

prop6　【指定の発話】という発言とその他の発言の関連性を分析した後，それについて参加者に意見を求めてください．

この時,【指定の発話】には，ユーザが注目する発話の文が挿入される．1 つ目の機能では，指定の発話のトピックについて明示的に意見を聞くことを目的としている．2 つ目の機能では注目する発話についての分析結果を伝え，分析内容のどこを取り上げるかは参加者にゆだねて判断してもらうことを目的としている．この機能はランダムにどちらかが選ばれて実行される．

これらの機能は，他の参加者から見える形で発言させることも，自分だけが見える形で発言させることもでき，それは要請者が任意に決定する．他の参加者から見える形でチャットボットが発言すると，自身が発言せずとも間接的に場に貢献することができ，タイミングが合えば自身も発言することが可能である．一方，自分だけに見える形にすると，自身のタイミングで考え，発言することが可能であり，議論への貢献度は高く感じることが期待できる．

3 ユースケースと考察

「次世代の机」をテーマとしたブレインストーミングを 3 人で行なっているシチュエーションをもとに，各機能の挙動について説明する．機能を実行する直前までの会話ログは表 1 に示す通りである．なお，この会話ログは chatGPT を用いて作成を行なった．

表 1 の最後の発言の後で，2.2 節，2.3 節の各機能を用いた場合，表 2 の発話が生成される．prop1，prop2 においては，発言していない参加者 4 が機能をつかったという想定で実行している．また，prop5，prop6 では参加者 2 の 1 回目の発言を指定している．prop1，prop3 では，適切な生成が行えていると考える．複数回実行したが，安定的に生成できていることを確認した．適切な生成が行えていない機能について考察する．

prop2 においては,「参加者 4：」という余分なものがついてしまっている．テキストの後処理で削除することも考えられるが，ランダムに選んだ旨の発言を別途生成し，指名する文は prop1 で生成するというように分割することが有効だと考えられる．

prop4 では，トピックの発言量の分析を行ってほしいが，各参加者の発言数になっている．トピックの列挙自体はうまく行っていることから，一旦トピックの列挙をさせた後に，それぞれのトピックの量を分析させるような分割をすることが有効だと考えられる．また，トピックの量については，抽出したトピックに類似する発言を文書類似度で測ることで判定するようなアルゴリズムに置き換えることも検討する余地があると考える．

prop5 では，指定した発言に含まれるトピックについて取り上げてほしいが，全体の発言からトピックが抽出されている．

1：https://platform.openai.com/docs/api-reference/introduction

表 1　ユースケース：会話ログ

参加者 1	みんな、次世代の机ってどんなイメージある？
参加者 2	そうだな、デジタルとの融合がポイントかな。例えば、ディスプレイが組み込まれてて、ノートとか書き込めるような感じ？
参加者 3	それはいいね。それに加えて、充電も自動でできるようにして、スマホやタブレットがすぐ使えるようにすると便利そうだよね。
参加者 1	うん、確かに。でも、机のデザインも大事だよね。使いやすさはもちろん、部屋に馴染むデザインがいいな。
参加者 2	そうだね。あと、エルゴノミクスも考えないと。長時間座ってると体に負担かかるから、姿勢を考慮したデザインが必要だよね。
参加者 3	そうだね、快適さも大事だけど、環境にも優しい材料とかを使って持続可能なものにしたいよね。
参加者 1	机の高さも重要だと思う。個々の慎重に合わせて調整できるといいね
参加者 2	ディスプレイが触れるだけでなく、ジェスチャーコントロールできたら面白いかも
参加者 3	充電だけでなく、ワイヤレスでデバイス同士を連携させる機能も欲しいな
参加者 2	机の側面にコンセントや USB ポートがあると便利だと思う
参加者 1	持ち運びがしやすいデザインも考えてほしい。畳んでコンパクトにできるといいかもしれない
参加者 2	デスクトップアクセサリーを置くためのスペースも考えて欲しい
参加者 3	エコフレンドリーな素材を使用して、捨てる際にも環境にやさしい設計になってると良いね

数回実行した結果としては，適切に指定した発言からトピックを取り上げることができたものもあった．トピックの抽出とそれについて意見を求めるものとに分割することが有効だと考えられる．

prop6 では，前半の各発言との関係の分析はある程度うまく行っているように見えるが，後半はまとめに入ってしまい，参加者に意見を聞くことができていない．これも発言の関係分析と，それについて意見を求める発言を分けることで改善すると考える．

以上のことから，単一の命令になっている機能はうまく動作する一方，複合的な命令になっている機能はうまく動作していないといえる．そのため，命令を分割するように設計し，GPT を用いずにアルゴリズムで行うべき部分についても検討する必要がある．

4　評 価 実 験

4.1　実 験 手 順

4.1.1　発 言 促 進

15 人の参加者に消極的な参加者を意識し, 発言の意思がある状態での各手法を通じた発言のしやすさを 5 段階で評価してもらう. 各手法は下記 4 つとする.

（1） 司会者は何も発言せず, 自発的に発言する
（2） 誰か発言したい人がいるか司会者が聞く
（3） 意図的に司会者に自身を指名させる (prop1)
（4） 意図的に司会者に自身を指名させるが, ランダムで行ったと表記する (prop2)

5 段階評価は 1 が発言しにくいとし, 数字が大きくなるにつれて発言しやすいとする.

4.1.2　発言のきっかけ

各手法 15 人ずつトーク履歴と手法ごとに生成された文章より, テーマに対する意見と意見の出しやすさについて 5 段階の評価を行う. 各手法は下記 3 つとする.

（1） ヒントは何も出さずに意見を聞く
（2） 現在の発言の中心となるトピックを提示する (prop3)
（3） 発言として挙げられたトピックと、参加者の発言量を提示する (prop4)

5 段階評価については,1 が意見を考えにくいとし, 数字が大きくなるにつれて意見が考えやすいとする.

4.1.3　既存発言を追求するヒント

各手法 15 人ずつトーク履歴と手法ごとに生成された文章より, テーマに対する意見と意見の出しやすさについて 5 段階の評価を行う. 各手法は下記 3 つとする.

（1） ヒントは何も出さずに意見を聞く
（2） 1 つの発言に注目し, 参加者に意見を求める (prop5)
（3） 1 つの注目する発言とその他の発言の関連性を説明し, 意見を求める (prop6)

5 段階評価については,1 が意見を考えにくいとし, 数字が大きくなるにつれて意見が考えやすいとする.

4.2　実験結果と考察

4.2.1　発 言 促 進

各手法を 5 段階評価してもらった平均と標準偏差の結果を表 3 に示す. 自主的に発言することに対する評価は低く, 発言の機会を設ける支援は評価が高く, 発言しやすいと感じる傾向があると分かった. 発言の意思がある状態においては, 発言の機会を設ける支援が効果的であると考える.

4.2.2　発言のきっかけ

各手法を 5 段階評価した結果と, 意見数について表 4, 表 5 に示す. 表 4 より, 支援を行わない (1) の評価が高く, 具体的な支援を行う (3) の評価が低いとわかる. 一方で, 意見数に注目すると低評価である (3) 意見数は多いことが分かった. 具体的な支援では, 支援文章の中にキーワードが多く入っており, 意見が出しやすいが, 支援内容について言及する意識となってしまうため評価が低くなってしまったと考える.

4.2.3　既存発言を追求するヒント

各手法を 5 段階評価した結果と, 意見数について表 6, 表 7 に示す. 表 6 より, 支援を行わない (1) の評価は高く, 具体的な支援文章の (3) の評価が低いとわかる. また, 意見数に注目すると低評価であった (3) は意見数が増えているが,(2) の意見数は 1,2 と少ない結果であった. 原因として,(3) の支援では文章内のキーワードの提示が多く, 意見を作成しやすい文章となってい

表 2　発話生成結果

prop1	参加者 4 さん、次世代の机に求める要素は何だと思いますか？
prop2	参加者 4：次は参加者 4 さんに意見を聞きたいと思います。指名はランダムです。
prop3	参加者の皆さん、次世代の机についてたくさんの興味深いアイデアが出ていますね。デジタルとの融合、自動充電機能、使いやすさとデザイン、エルゴノミクス、そして環境に優しい材料の使用など、さまざまな要素が重要となりそうですね。これらの中で、特に重要だと思う点は何ですか？
prop4	参加者の発言量：- 参加者 1：2 回- 参加者 2：2 回- 参加者 3：2 回 これまでの発言について、皆さんが挙げたポイントについて意見を求めます。デジタルとの融合、充電機能、デザイン、エルゴノミクス、環境への配慮など、どのポイントが特に重要だと思いますか？
prop5	参加者の皆さん、デジタルとの融合、充電機能、デザイン、エルゴノミクス、そして環境への配慮など、次世代の机に求められる要素はさまざまですね。それぞれの要素について、どのようなデザインや機能が理想だと思いますか？
prop6	参加者 2 の発言は、デジタル技術との融合に焦点を当てており、ディスプレイや自動充電機能などのアイデアを提案しています。一方、参加者 3 は快適さと環境に配慮した持続可能な材料の使用について言及しています。参加者 1 はデザインと使いやすさに焦点を当てています。参加者の意見を総合すると、次世代の机はデジタル技術との融合、快適さ、環境に配慮した持続可能な材料の使用、そしてデザインと使いやすさが重要であると言えます。参加者に意見を求めると、次世代の机に求める要素は多岐にわたることが分かります。それぞれの要素をどのようにバランスよく取り入れるかが重要です。例えば、デジタル技術との融合を図りつつも、環境に配慮した持続可能な材料を使用することは可能でしょうか？それぞれの要素をどのように組み合わせるか、参加者の意見を伺いながら議論を深めていくと良いでしょう。

表 3　発言促進手法の評価結果

	(1)	(2)	(3)	(4)
平均	1.6	3.5	3.8	4.2
標準偏差	0.61	1.31	1.33	1.1

表 4　発言のきっかけ支援評価結果

	(1)	(2)	(3)
平均	3.47	3.27	2.27
標準偏差	1.31	1.06	1.24

表 5　発言のきっかけ支援による意見数

	(1)	(2)	(3)
平均	2.21	2.00	2.43
標準偏差	1.18	0.73	1.14

表 6　既存発言追及支援の評価結果

	(1)	(2)	(3)
平均	2.93	2.27	1.87
標準偏差	1.12	1.06	0.81

表 7　既存発言追及支援による意見数

	(1)	(2)	(3)
平均	1.86	1.50	1.87
標準偏差	1.29	0.94	1.56

たが,(2) の支援では感想を聞く形となってしまっていたことが原因であると考える.

5 ま と め

本研究では，消極的参加者が自発的に発言することを支援するチャットボットを提案した．このチャットボットの機能として本稿では，自身を発言者として指名する機能やこれまでの議論をまとめたり引用したりして，過去の発言に基づく発言を言いやすくする機能を定義した．本稿では，各機能についての動作確認を行い，複合的な命令になっている機能については，分割して実行することやアルゴリズムで置き換える場所の検討が必要であることを確認した．また, 支援の有効性を確かめる実験では, 発言の意思がある状態で発言の場を設ける支援は効果的であると分かった. 一方で, 発言生成のきっかけを作る支援は意見数の増加は見込めるものの, 評価が低いことが分かった.

今後の課題としては，評価が低かった支援に対しての原因を探るために文章量とキーワード量を調節した際の評価検証を行う. その後構築したシステムを用いて実験を行い，どの機能がどの程度利用され，発言の促進に有効であるかを分析する予定である．実験にあたって被験者を集め，規模を大きくすることが困難であると予想されるため，まず，このシステム上で動作する参加者チャットボットを構築する．次に提案手法を用いずに 1 名の被験者と数体のチャットボットという環境で実験を行い，被験者の積極性を判断する．最後に，同環境で提案手法を用いた実験を行い，積極的参加者，消極的参加者に対する各機能の発言促進効果について評価する．さらに、支援内容が自身だけに見えるのか, 全体に見えるほうが良いかの検証も行う. また，すべての参加者を被験者とした実験も行い，実環境における効果についても評価する予定である．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 塩津翠彩, 高島健太郎, 西本一志. Lighthousechat：消極的参加者に発言を促す手段を備えた会議支援メディア. インタラクション 2018 論文集, pp. 260–263, 02 2018.

[2] A.F. Osborn. *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner, 1953.

意外な単語の組み合わせを用いた発言生成に基づくブレインストーミング支援

大岡　麗† 関　峰†† 北山　大輔†

† 工学院大学情報学部システム数理学科　〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学大学院工学研究科情報学専攻　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j320044,em22006}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　アイデア生成にブレインストーミングが使われてきた．ブレインストーミングで重要となる要素として，「アイデアの量」と「連想と結合」がある．そこで，本研究ではこの要素を支援する bot を作成する．チャット上でブレインストーミングをすることで，発言から特徴語を抜き出し，特徴語の共起ネットワークを作成する．さらに，作成された共起ネットワークから，単語の類似度やホップ数を用いて意外な 2 単語を選び出す．最後に選ばれた 2 単語から大規模言語モデルを用いて，結合に基づく案を生成する．この一連の操作によってユーザの発案の支援を実現し，発言数の増加を本研究の目的とした．その結果，議論における発言文字数は増えなかったが，生成された発言により，被験者の発案の支援となっていることが確認された．

キーワード　ブレインストーミング，参加者 bot，グラフ，見落とし話の発見

1　はじめに

会議方式によって，新しいアイデアを生み出すための手法として，Osborn [1] の考案した手法であるブレインストーミングが一般的に使用されている．また，Osborn はブレインストーミングの実施に際して，4 つの原則を提唱している．それらの要点は,「批判厳禁」,「自由奔放」,「アイデアの量」，および「連想と結合」である．

「アイデアの量」とは，アイデアの質よりも多くのアイデアを生み出すことを重視する考え方である．これにより，多様なアプローチや視点から新たなアイデアが生まれる．また,「連想と結合」は，既存の案を用いて新たな案を導き出す原則である．多様な観点を持ったたくさんのアイデアを生み出すためには，既存のアイデアを組み合わせることが効果的とされている．そこで本研究では，これらの要点のうち「アイデアの量」と「連想と結合」の 2 つを重視する．しかし，佐々木ら [2] によると，人間同士のブレインストーミングでは，2 つの既存のアイデアを結合させることが難しいとされている．

この問題を解決するために，人間が得意でないアイデアの結合を，bot が実行する手法を提案する．また，結合の組が似通ったものでは新たな視点が生まれない．そのため，意外性のある組み合わせを探索する必要がある．そこで本研究での「結合」を，2 つの単語を関連させて新たな案を考えることと定義する．さらに，意外性のある組み合わせを，単語の概念に関連がなく，議論の中でつながりを見つけられていない組み合わせと定義する．すなわち人間には組み合わせることが困難な単語の組といえる．そのため，単語の類似度と共起ネットワーク上のホップ数を用いて選択する．そして，選ばれた 2 つの特徴語から，発案の支援となる有益な発言をする bot を作成し，発言数の増加を本研究の目的とする．

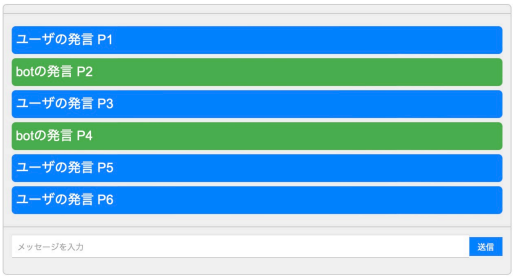


図 1　チャット画面

2　提案手法

本章では，本研究における提案手法のシステムの概要を説明する．本稿ではテキストチャットを想定しているが，音声認識を用いることで，音声チャットでも実現可能であると考える．一方で，図表の利用は想定しておらず，ホワイトボードや付箋等を利用した形式は今後の課題である．

本システムではブレインストーミングの参加者は何人いてもよく，bot の数も任意の数参加させることができる．以降は議論の単純化のために参加者が 1 名，bot が 1 名の想定で説明する．図 1 にブレインストーミングをするチャット画面を示す．下のテキストボックスに発言を入力し，送信することで発言が確定する．ユーザの発言である発言 P_1，P_3，P_5，P_6 は青色で表示され，bot の発言である発言 P_2，P_4 は緑色で表示される．チャットで発言するたびに特徴語の抽出と共起ネットワークへの追加を行う．

次に，bot の発言時に枝刈りをすることで共起ネットワークの整理し，共起ネットワークから，単語類似度とホップ数により選択された意外な 2 単語選び出す．最後に，選ばれた 2 単語を元に bot が発言し，特徴語の抽出と共起ネットワークへの追

表 1　発言 P_1,P_2,P_3 から得られるキーフレーズとスコア

発言	キーフレーズとスコア
P_1	k_1:100,k_2:90,k_3:80,k_4:70,k_5:60,k_6:50,k_7:40,k_8:30,k_9:20
P_2	k_{10}:100,k_5:90,k_4:60,k_{11}:40
P_3	k_{12}:100,k_{11}:90,k_{13}:50

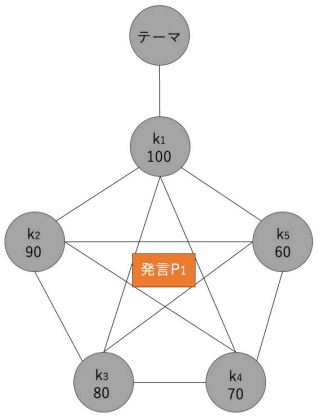


図 2　発言 P_1 の共起ネットワーク

加を行う．以下で具体的に説明する．

2.1　発言からの特徴語ノードの追加

ユーザの発言や bot の発言に対して，キーフレーズを抽出し，発言における重要度を示すスコアをキーフレーズごとに算出するため，yahoo のキーフレーズ抽出の手法 [1] を用いる．ここで，1 回の発言から多数のキーフレーズが取れる場合，発言からスコアの高い上位 n 件を発言の特徴語とする．

表 1 に例として作成した発言と，抽出されたキーフレーズとスコアをそれぞれ示す．例では 5 件の特徴語を抽出するものとする．発言 P_1 に注目すると，9 個のキーフレーズが抽出され，個別にスコアが算出される．そして，特徴語として採用されるのはスコアを参考に k_1，k_2，k_3，k_4，k_5 である．

次に，ブレインストーミングの会話構造を表すために特徴語を用いた共起ネットワークを作成する．ユーザや bot が発言するたびに，1 回の発言で得られた特徴語集合から完全グラフを作成する．テーマのノードに直接繋げるノードは特徴語集合内で重要度のスコアが高いノードとする．

図 2 に発言 P_1 の共起ネットワークを示す．キーフレーズから選ばれた特徴語集合は［k_1,k_2,k_3,k_4,k_5］であり，発言 P_1 ではノードが 5 個の完全グラフが作成される．完全グラフはテーマのノードに接続される．また，図 2 の特徴語の下に表示されているスコアは，yahoo キーフレーズ抽出を用いて特徴語と一緒に抽出された重要度のスコアである．

図 3 に発言 P_1 と発言 P_2 の共起ネットワークを示す．発言内の特徴語が既出の場合，すでに共起ネットワークに存在するノードへ接続する．発言 P_2 の特徴語集合は［k_4,k_5,k_{10},k_{11}］であるため，発言 P_1 と重複する特徴語は k_4，k_5 である．そのため発言 P_1 に重なるように発言 P_2 の完全グラフが追加される．

1：https://developer.yahoo.co.jp/webapi/jlp/keyphrase/v2/extract.html

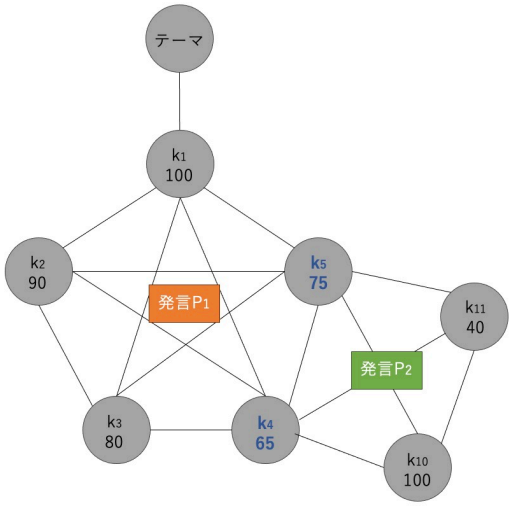


図 3　発言 P_1 と発言 P_2 の共起ネットワーク

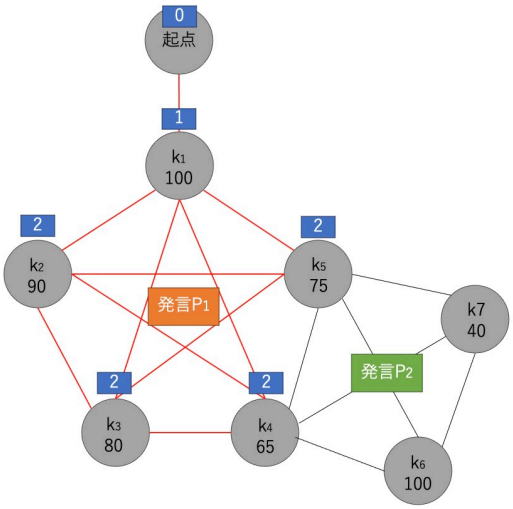


図 4　削除候補のエッジ

表 2　エッジごとのテーマからのホップ数

テーマからのホップ数	エッジ
0.5	テーマ-k_1
1.5	k_1-k_2,　k_1-k_3,　k_1-k_4,　k_1-k_5
2	k_2-k_3,　k_2-k_4,　k_2-k_5,　k_3-k_4,　k_3-k_5

そして，重複した特徴語のスコアは既存のスコアとの平均をとるため，表 1 より k_4 のスコアは P_1 の 70 と P_2 の 60 である．図 3 の青字で示すように平均の 65 が最終的なスコアとなる．

2.2　枝　刈　り

m 分が過ぎてもエッジの両端である 2 つの特徴語が同時に発言されなかった場合，そのエッジを削除の候補とする．これは，m 分経つことで特徴語のつながりが bot を含む参加者集団の思考の中で薄くなることを表している．

図 4 において，発言 P_1 と発言 P_2 の間に m 分の時間が存在したと仮定する．その時，削除候補のエッジは赤色の部分である．ただ，k_4-k_5 エッジは発言 P_2 において特徴語 k_4 と特徴語 k_5 が発言されているため，候補エッジとならない．

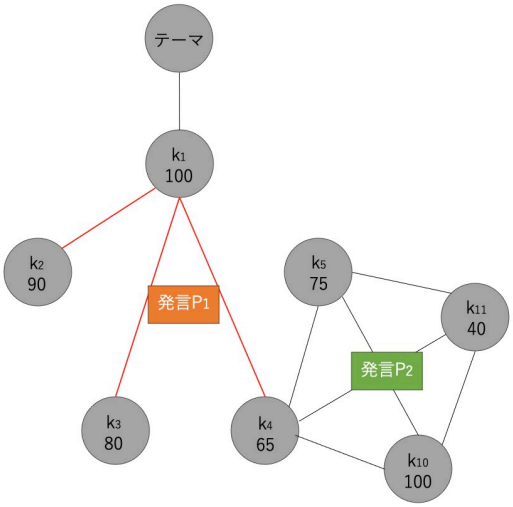


図 5　枝刈り後の共起ネットワーク

次に，削除候補となったエッジのうち，テーマからのホップ数が遠いものから削除する．エッジにおけるテーマとのホップ数の定義として，エッジの両端ノードに注目した．本研究では，それぞれのノードとテーマのノードのホップ数を算出し，平均をとったものをエッジにおけるテーマとのホップ数とする．たとえば，k_1-k_2 のエッジのホップ数の計算方法は，$(1+2)/2 = 1.5$ となる．図 4 の青の数値がノードとテーマのホップ数となる．この数値を元に，それぞれのエッジごとのテーマからのホップ数を計算した結果を表 2 に示す．

削除候補のエッジを削除していく段階で，ホップ数が同一の場合は，エッジの両端ノードのスコアの平均が小さいものから削除する．図 4 で示す重要度のスコアは k_1 から k_5 の順に小さい．そのため表 2 においてテーマからのホップ数が同じ場合，左から右にかけてエッジを削除する．

ただし，削除した時に任意のノードがテーマのノードの共起ネットワークから孤立する場合はエッジの削除は行わない．言い換えると，枝刈り後であっても任意のノードがエッジを辿ればテーマのノードに接続する状態となる．孤立するノードが存在する場合，後述する意外なノードの探索ができないためである．

図 5 はテーマからのホップ数が 2 のエッジすべてと，ホップ数が 1.5 のエッジ k_1-k_5 を削除したものである．次のエッジ削除候補であるエッジ k_1-k_4 を削除すると，ノード k_4 を含む共起ネットワークが，テーマのノードを含む共起ネットワークから孤立してしまう．そのため，エッジ k_1-k_4 は削除しない．

一連の枝刈りの操作によって，1 つの発言を完全グラフで表現していたものを，ブレインストーミングが進むにつれ，発言の共起ネットワークの構造がユーザにとって正確な関連度として表れるようになっている．

2.3　意外なノードの探索

本研究における意外な組み合わせは，単語の概念が遠く，議論の中で見落とされ，組み合わされていない単語組である．そのため，単語の類似度と作成した共起ネットワークから意外なノードの組み合わせを探索する．共起ネットワーク中でのホップ数が大きい場合，参加者間で単語のつながりを見いだせていないことを意味すると考える．また，単語の類似度が低い場合，概念的に関連しない単語であると考える．この 2 つの指標を用いることで，人間が組み合わせを考えにくく，かつ見落とされているアイデアを提示することを可能とする．はじめに，ユーザの直近の発言から得られた特徴語集合 N のうちの 1 単語と，全特徴語集合から特徴語集合 N を除いた特徴語集合 M から選ばれた 1 単語を組み合わせとする．

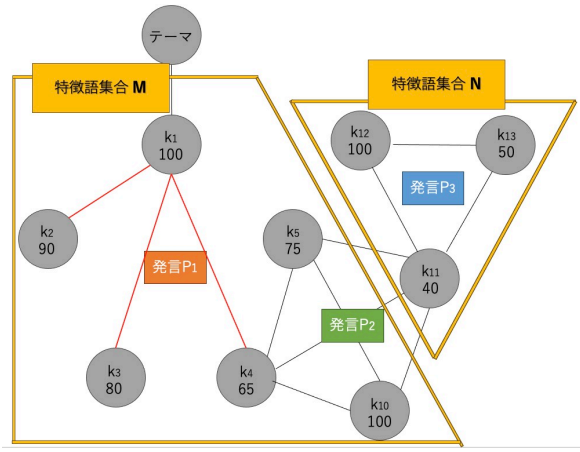


図 6　発言 P_3 が追加された共起ネットワーク

図 6 に示した共起ネットワークにおいて，ユーザの発言 P_3 後に bot が発言する状況だと仮定する．この時，特徴語集合 N はユーザの直近の発言 P_3 である［k_{11}，k_{12}，k_{13}］となり，特徴語集合 M は［k_1，k_2，k_3，k_4，k_5，k_{10}］となる．発言 P_2 の一部である特徴語 k_{11} は発言 P_3 に含まれるため，特徴語集合 N に含まれる．それぞれの特徴語集合の組み合わせを総当たりで計算する．特徴語集合 N から取られたものを w_i，特徴語集合 M から取られたものを w_j とし，式 1 で表す指標を算出する．

$$score(w_i, w_j) = hop(w_i, w_j) \times (1 - sim(w_i, w_j)) \quad (1)$$

hop 関数部分は，2 単語の最短のホップ数を表すものである．たとえば，図 6 において特徴語 k_2 は w_j で，特徴語 k_{13} は w_i となり，*hop* 関数による計算結果のホップ数は 4 である．

$1 - sim(w_i, w_j)$ 部分は，2 単語の非類似度を表す．*sim* 関数である 2 単語の類似度の計算方法としては fasttext[2]を用いた．

ホップ数が遠く類似度の低い単語の組が大きな値となる．この値が大きいほど，ユーザにとって 2 単語を結合させることが起こりにくい組を対応づけることが可能となり，選ばれた 2 単語は意外な単語の組み合わせと言える．

最後に，bot の発言として内容が同じ発言を避けるため，1 度選ばれた単語の組は 2 度と選択されない．

2.4　bot の発言の生成

bot の発言には OpenAI が提供している大規模言語モデル

2：https://fasttext.cc

表 3　GPT のプロンプト

役割	プロンプト部分
発案	「*theme*」というテーマでブレインストーミングをしています。意外な組み合わせである、「$word_a$」と「$word_b$」のキーワードを元にアイデアを生み出してください。また、発言中に 2 つのキーワードを含んでください。そして、発言の文字数は 40 字以上 50 字以内で生成してください。
基本会話	「*theme*」というテーマでブレインストーミングをしています。議論のテーマに則って、相手に共感をする発言をしてください。また、発言の文字数は 30 字以上 40 字以内で生成してください。ただし，会話の中で新たに案出しをしないでください”

表 4　実験用に追加した GPT のプロンプト

役割	プロンプト部分
発案	「*theme*」というテーマでブレインストーミングをしています。アイデアを生み出してください。そして、発言の文字数は 40 字以上 50 字以内で生成してください。

（Large Language Models）である GPT[3]を用いる．Model は gpt-3.5-turbo-1106 使用し，パラメータである temperature は 0 とした．表 3 で示したように，system 部分のプロンプトに theme と，bot が発言する度に意外な単語の組である $word_a$ と $word_b$ が入力される．これによってどのような *theme* や $word_a$ や $word_b$ でも，生成できる汎用的なプロンプトを作成した．

さらに，発案のみを行う bot であると，ユーザは本来のブレインストーミングとは違い，bot に引きずられて発案のみの発言をしてしまう可能性がある．そのため，l %の確率で発案をし，$(100 - l)$ %の確率で会話の流れに沿った基本の発言をするように調整した．また，表 3 のプロンプトに加え，user 部分のプロンプトにはユーザが発言するたびに発言履歴が追加される．加えて，system 部分と user 部分以外で GPT3.5 に実装されている assistant 部分が存在する．assistant 部分のプロンプトには bot が発言するたびに発言履歴が追加される．bot が発言する間隔は s 秒から t 秒の乱数である．これによって人間のような不規則な発言の間隔を実現する．

3　評 価 実 験

3.1　実 験 手 順

本提案システムを使用することで，ブレインストーミングにおけるユーザの思考の手助けになることを検証する．主に，ユーザの発言数の増加と bot の発言の参考度合いを確認する．

実験の環境と手順について説明する．クラウドソーシング[4]で募集した 60 人に対して，3 つのテーマから 1 つを選択してもらい，被験者 1 人に対し 1 人の bot を参加させて実験を行った．3 つのテーマは「新しい家庭用ロボット家電を考える」,「次世代のインターホンを考える」,「新しいコンビニ弁当を考える」である．複数の案を用意することで，被験者の知識量の差で結果が揺らがない．ブレインストーミングの時間は 15 分とし，被験者にはチャット上で相手と会話しながら案を出すように指示した．そして，最終的に 1 つ以上 3 つ以下の案を提示するよう説明した．これは被験者が集中してブレインストーミングをするためである．実験の最後に，被験者が参考になった bot の発言を任意の個数抜き出して回答するように指示した．

また，2 章で示した参加者の発言から取り出す最大の特徴語数は 5 個とした．抽出する特徴語数を増やした場合，bot の発言から抽出される特徴語数のみが増え，被験者の発言が軽視される．次に，枝刈り法の境界となる時間のパラメータ m を 5 分とした．実験時間の 15 分間中に 2 回の枝刈りがされるようにした．さらに，bot の発案発言の確率のパラメータ l を 70 %とし，それに応じて基本発言は 30 %と設定した．初回の発言は基本発言とし，15 分間に合計して約 10 回の発案発言が行われる．最後に，bot の発言の時間間隔であるパラメータ s と t をそれぞれ 50 と 70 とした．これは，被験者が bot の発言を解釈し，自分の考えを入力する時間を考慮したものである．bot の発言文字数の期待値は 45 文字の発案発言と 35 文字の基本発言とそれぞれの発言確率より 42 文字となる．この文字数の発言を解釈して自分の発言を入力することを加味し，1 分が最適だと考えた．

さらに，提案手法と後述する 2 つの比較手法を用意し，3 つのグループに被験者を割り当てた．1 つ目の比較手法である sim 手法は提案手法と違って，選択される 2 単語の基準を変更した．提案手法では式 1 のように類似度の低い 2 単語が選択されたが，sim 手法では式 2 のように類似度が高い 2 単語同士を選択する．

$$score(w_i, w_j) = hop(w_i, w_j) \times sim(w_i, w_j) \quad (2)$$

これは，関連が見つけにくい 2 単語ではなく，本来関連があるが被験者が見つけられていない組み合わせが選択される．2 つ目の比較手法である keywordless 手法は提案手法と違って，アイデア生成の際に 2 単語の組み合わせを行わない．プロンプト部分における keywordless 手法と提案手法の違いを表 4 に示す．提案手法では，発案発言のプロンプトは 2 単語を組み合わせる指示であったが，keywordless 手法では 2 単語の結合に関する指示をせず，文脈からアイデアを生み出させる．

3.2　実験結果と考察

はじめに，bot の発言に関しての結果と考察を述べる．提案手法の bot の発言が表 5，6 で，比較手法である sim 手法の bot の発言が表 7 である．それぞれ，sim 手法は関連している 2 単語を組み合わせて発言を生成し，提案手法は意外性のある 2 単語を組み合わせる．提案手法の結果である表 5，6 より，組み合わせが困難である 2 つの概念を組み合わせて発案をしている．一方で表 7 より，sim 手法では人間が思いつきそうではあるが被験者にとっては盲点である発言をしている．

次に，被験者の行動に関しての結果と考察を述べる．発案の発言をする確率は 70 %としたが，実際の発案発言確率の分布は

3：https://platform.openai.com/docs/api-reference/introduction
4：https://crowdworks.jp

表 5 「新しい家庭用ロボット家電を考える」をテーマにした提案手法の bot の発言

選択された 2 単語	発言内容
飼い猫, 協力	飼い猫と協力して遊ぶロボットがあれば、猫の運動不足を解消し、飼い主の負担も軽減できます。猫との絆を深める機会にもなります。
話し相手, 活性化	家事を手伝いながら、家族との会話を楽しめるロボットを提案します。掃除や料理をしながら、楽しいトークやクイズを通じてコミュニケーションを活性化できます。

表 6 「新しいコンビニ弁当を考える」をテーマにした提案手法の bot の発言

選択された 2 単語	発言内容
コーヒー, 仕事中	仕事中に手軽に食べられる新しいコンビニ弁当には、コーヒーに浸すと美味しくなるチーズも取り入れると面白いかもしれません。
お湯, 人	お湯で戻すこともできる野菜スティックをセットにした新しいコンビニ弁当。忙しい人に手軽で栄養豊富な食事を提供し、仕事中でも健康的な食事が可能になります。

表 7 「新しい家庭用ロボット家電を考える」をテーマにした sim 手法の bot の発言

選択された 2 単語	発言内容
自動, 料理	自動で料理をしてくれるロボットは、食材を切ったり調理したりする機能があれば便利です。レシピを読み込んで自動で料理を作ってくれると良いでしょう。
アロマ, 環境	天候や予定、気分に合わせたアロマを出すロボット家電は素晴らしいアイデアです。AI が分析し、自動でアロマを調整することで快適な空間を提供できますね。

表 8 四分位数

第一四分位数	0.59
第二四分位数	0.65
第三四分位数	0.71

表 9 発案発言確率別の被験者数

	59 %未満	59 %以上 71 %未満	71 %以上
proposed	4	13	3
sim	4	7	9
keywordless	4	10	6
合計	12	30	18

図 7 のようになり，四分位数は表 8 となった．ブレインストーミングが開始して 1 回目の bot の発言は基本発言であるため，発案発言の確率が 70 %を下回る被験者が多いと考えられる．

bot の発案発言確率を元に以下の 3 つのグループに分割して考察する．

- 第一四分位数以下
- 四分位範囲
- 第三四分位数以下

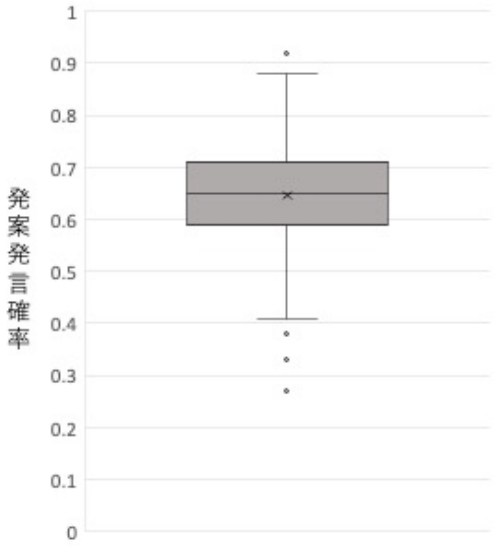


図 7 bot の発案発言確率の分布

表 10 全体の被験者の統計

	平均発言数	平均発言文字数	平均参考発言数
proposed	11.7	267.9	2.70
sim	12.7	332.0	2.30
keywordless	12.9	340.0	2.30
平均	12.4	313.2	2.43

表 11 発案発言が 59 %未満の被験者の統計

	平均発言数	平均発言文字数	平均参考発言数
proposed	10.0	206.0	1.75
sim	13.3	384.0	1.75
keywordless	11.3	324.8	1.75
平均	11.5	304.9	1.75

表 12 発案発言が 59 %以上 71 %未満の被験者の統計

	平均発言数	平均発言文字数	平均参考発言数
proposed	12.9	284.0	2.92
sim	13.3	364.9	2.14
keywordless	13.4	346.6	2.10
平均	13.2	331.8	2.39

表 13 発案発言が 71 %以上の被験者の統計

	平均発言数	平均発言文字数	平均参考発言数
proposed	8.7	280.3	3.00
sim	11.9	283.2	2.67
keywordless	13.2	338.3	3.00
平均	11.2	300.6	2.89

3 つのグループに加え，それぞれの手法別の被験者数を表 9 に示す．

次に，発言数，発言文字数，参考発言数をそれぞれ集計した．被験者がテキストボックスに発言を入力して送信することで発言が確定した．これを 1 つの単位として，ブレインストーミングの合計の発言を発言数と定義した．発言文字数は被験者のす

表 14　参考発言数の発言種類の内訳

	基本発言	発案発言
proposed	0.30	0.70
sim	0.25	0.75
keywordless	0.29	0.71
平均	0.28	0.72

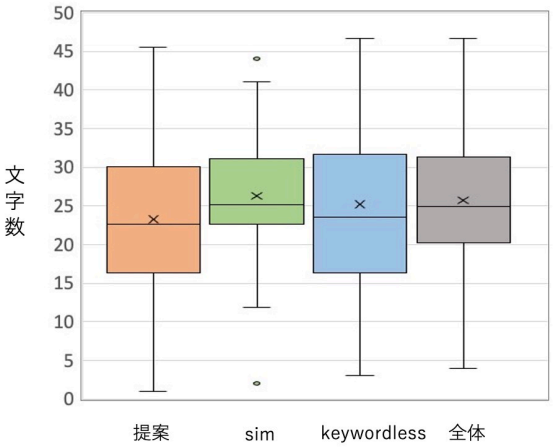


図 8　1 回の発言の文字数の分布

表 15　発言文字数の伸びた被験者

発言者（発言種類）	発言内容
bot（発案発言）	洗濯物の回収も自動で行う機能を追加すれば、完璧な家事ロボットになりますね。洗濯から回収までを一貫して行うことで、家事の負担を大幅に軽減できます。
user	すごいもう何もすることがない
user	あ、ついでにコーディネートまでしてくれるとか
user	いい感じの外出セットにして、ハンガーにかけてくれる

べての発言の文字数をカウントしたものである．参考発言数とは，ブレインストーミング後のアンケートで参考になった bot の発言を選び出してもらった発言数である．

これらを 3 つの手法ごとに集計したものが表 10 となり，発案発言確率の四分位数とグループごとに集計した結果は表 11，12，13 となった．

表 10 に示すように提案手法の発言数は比較手法である 2 つの手法より 1 回分少ない．また，発案発言確率別に見ても提案手法が比較手法に勝ることはなかった．これに応じて発言文字数も約 60 文字少ない．発言数と発言文字数の相関係数は 0.51 となり，2 変数には相関がある．ただ，強い相関を示しているとは言えない．これは 1 回の発言の文字数に個人差があると考えられる．1 回の発言あたりの文字数を算出した結果が図 8 となる．提案手法に，外れ値として 108 文字の被験者が存在したので除外した．これより 1 回当たりの発言文字数の少ない被験者が存在する．主に，表 15 に示すように相槌のような基本発言を含みながら発案をしている．実際に対面で会話しているかのようにテキストチャット形式で話している．これは発言文字

表 16　発案発言のみをする被験者

発言者（発言種類）	発言内容
user	使い終わったお弁当箱を再利用できるコンビニ弁当も良いと思いますが、ただのお弁当箱として使うのではなく、災害時にも役立ったりするお弁当箱を提供する。
bot（発案発言）	「災害時にも役立つお弁当箱」を提供するアイデアは素晴らしいですね。再利用可能で、非常時にも使える機能を備えたお弁当箱は、多くの人々にとって便利なアイテムになるでしょう。
bot（発案発言）	「耐衝撃お弁当箱」を提供！災害時にも使える特殊素材で作られ、再利用可能。食事後は非常時の備えとして活用できる。
user	逆に、お弁当箱を使わずにビニール袋に入れただけの環境にやさしいコンビニ弁当があってもいいかもしれません

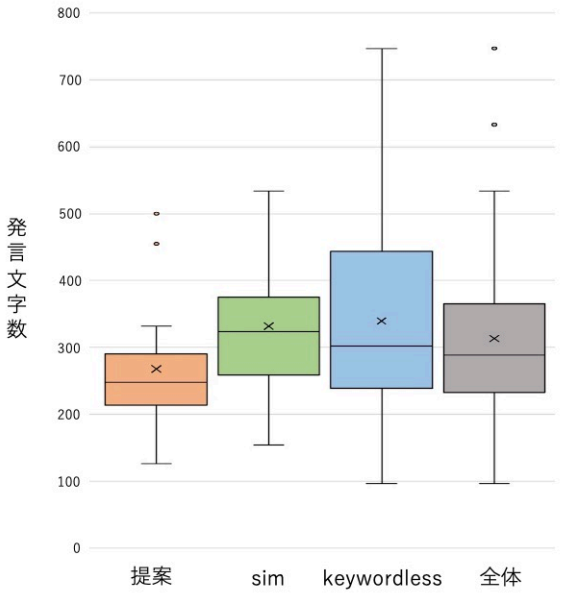


図 9　被験者の発言文字数の分布

数が少なくなり，発言数が増える例である．また，人間にも基本発言や発案発言の割合が存在すると考える．それに応じて，発話スタンスの違いが生まれている．

一方で，表 16 に示すように，基本発言をしないで箇条書きのように発案発言のみをする被験者も複数確認した．

ここで，すべての被験者の発言文字数の分布を図 9 に示す．発言文字数に個人差があると言える．発言数が伸び悩んだ被験者の特徴として，長考してしまう点が考えられる．bot の発言間隔は 50 秒から 70 秒の乱数であり，長考するがゆえに，bot の発言を解釈し自分の発言を入力する前に次の bot 発言が表示されてしまう．bot の発言が複数回続くものの，被験者の発言が 1 度もない状況である．ここで提案手法が長考する要因になった可能性がある．提案手法は突飛なアイデアを生成していたため，被験者はアイデアを理解するのに時間がかかったと考えられる．

表 17　相手に質問をする被験者

発言者（発言種類）	発言内容
user	何かアイデアはありますか？
bot（基本発言）	「新しいコンビニ弁当を考える」私は、地域ごとの特色を生かした地域限定のコンビニ弁当を提案したいと思います。それぞれの地域の食材や料理を取り入れることで、地元のお客様に喜んでもらえると思います。
user	それはぜひ導入しましょう。他にありますか。
bot（基本発言）	地域限定のコンビニ弁当は、地元の食材や料理を取り入れることで、お客様に新しい味を提供できます。
bot（発案発言）	「季節限定のコンビニ弁当」季節ごとに異なる食材やテーマを取り入れたコンビニ弁当を提供し、お客様に新鮮な味わいと楽しみを提供します。
user	それは素晴らしいです。地域限定以外のアイデアはありますか？

一方で比較手法である sim 手法は類似度が高いもの同士を結合した発案を行った．つまり，人間には理解しやすい文章であるため，bot の発言間隔内に内容の解釈ができた．

一方発言文字数が伸びた被験者の特徴として，またブレインストーミングの原則に則り，突飛な意見でも積極的に恐れず発言している．被験者の発案発言割合の個人差に加えて，これらの発話スタンスの違いから，発言文字数と発言数の相関が 0.51 と強い相関を示さなかったと考えられる．

次に参考になった bot の発言数を考察する．表 11，12，13 に示す平均参考発言数より，bot の発案発言確率が大きいほど参考発言数が大きい．これは，他者の発案が自分の発案につながることがわかる．表 10 に示す全体の集計結果より，参考になった発言文字数は提案手法が一番多い．また，発案発言確率にかかわらず提案手法が劣ることはなかった．よって，突飛な案が被験者のアイデアの種となっている．さらに，参考になった発言数に関して，発言の種類ごとに分けたものが表 14 となる．発案発言が参考になった発言として大半を占めずに，平均として 0.72 となった．表 8 で示したように，発案発言確率の平均は 0.65 であったことからも，基本発言も被験者の支援となっている．基本発言は被験者への共感を示す発言であったため，被験者としても良い気持ちで発案を行うことができたと考えられる．これはブレインストーミングの原則からも重要だとわかる．

最後に，被験者が案出しをせずに bot に対して質問をすることで，ブレインストーミングを進めている可能性がある．例として表 17 の被験者が存在した．このようにブレインストーミングの進め方によってさまざまな種類に分けられる．これらの結果から，個人の発話のスタンスによるものが実験結果に大きく影響していると考えられる．

4　関 連 研 究

本研究に関連する研究を紹介する．ブレインストーミングにおいて，多様な案を出すためにさまざまな方法がとられてきた．灘本ら [3] の研究より，意外性のある組み合わせの探索の際に，web 上のデータから取り寄せる方法が存在した．しかし，思案者の未知の単語同士での発想は難解とされている．そこで，共起ネットワークの中から組み合わせを探索することによって，思案者の既知の単語同士での組み合わせで思考が可能とする．竹内ら [4] の研究では，専門知識や固定観念の束縛を逸脱するためにモノボケを用いた．また，趙ら [5] の行なった研究は，突飛な案を出すために子供の発想力を利用した研究であり，提案手法により独自性を持った案が多く得られた．さらに，長谷部ら [6] の研究では，ブレインストーミングを思考者の盲点発見の技法として利用する新しい発散的思考技法 BrainTranscending を作成した．これは，既存の発案手法である強制連想法を応用したものである．このように，遠い概念を見つけ出すためにモノボケや子供の発想を使用したが，本研究では会話の流れから遠い概念を定量的に結びつける．

小山ら [7] は発案を支援するシステムを作成した．マインドマップ上である特徴語を選択することによって，単語分散表現により連想語を推薦する．本システムも同様に，word2vec を元に類似度を計算する．しかし，連想語を提示することで発案を促進させている点が異なり，結合によって支援する．

西海ら [8] の研究は，マインドマップの作成時に，仮想エージェントが関連思考を促す質問をするシステムを作成した．本研究は，bot 自身が発案をすることによって，ユーザの更なる発案を促進する点が異なる．

また，会話の構造を整える方法として，大島ら [9] は発言と単語の 2 部グラフを作成し，協調学習のネットワーク分析に活かした．森ら [10] は，会議中の発言から会議内容の把握や次回会議までの課題を，会議後のリフレクションに必要な議事録あり方を議論した．議論中の発言を意味的な単位で分割し構造化した．これらの研究と違って，本研究では単語単位で共起ネットワークを作成する．

5　ま　と　め

本稿ではブレインストーミングの重要な項目の 2 つである，アイデアの量」と「連想と結合」に対しての支援をするシステムを作成した．手法として，チャットで入力されたユーザの発言から特徴語を抽出し，会話の構造を共起ネットワークで表した．作成された共起ネットワークから意外な 2 単語を探索し，大規模言語モデルを通して bot が発言する．結果として，提案手法によって議論における発言文字数は増えなかった．これは突飛なアイデアを提示することによって解釈に時間がかかったものだと考えられる．また，突飛な案が被験者の発案の支援になっていることが確認された．また，個人の発話のスタンスさまざま存在することが確認され，その影響が大きいと言える．今後の課題はチャットシステムの会話精度の向上を目指し，被験者

のブレインストーミングの進め方のスタンスに沿った支援を行う．また，複数人で利用した場合，bot の発言を任意呼び出し形式にした場合など，他の利用状況における評価も実施する．

6 謝　　辞

本研究の一部は，2023 年度科研費基盤研究（C）(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] A.F. Osborn. *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner, 1953.
[2] 佐々木航, 高島健太郎, 西本一志. カードゲームをモチーフにした結合アイデアの生成を支援するグループ発散思考技法. 情報処理学会研究報告, 2021.
[3] 灘本明代, 阿辺川武, 荒牧英治, 村上陽平ほか. コミュニティ型コンテンツのコンテンツホール抽出手法の提案. 情報処理学会研究報告データベースシステム (DBS), Vol. 2007, No. 65 (2007-DBS-143), pp. 265–270, 2007.
[4] 西本一志, 竹内慎吾, 高島健太郎. Monobocai: モノボケを活用した発散的思考技法の提案. 2022.
[5] 趙暁, 高島健太郎, 西本一志. 「子供の発想」 を利用するアイデア生成技法の提案とその有効性の検証. 情報処理学会研究報告, 2018.
[6] 長谷部礼, 西本一志ほか. 思考者の盲点を発見し活用する発散的思考技法. 研究報告グループウェアとネットワークサービス (GN), Vol. 2015, No. 8, pp. 1–7, 2015.
[7] 小山琢也, 荻野正. word2vec を用いたブレインストーミングシステムの実装. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2018.
[8] 西海湧登, 吉田直人, 米澤朋子ほか. マインドマップ上発散的思考支援におけるノードの列挙を促進する対話エージェント. 2016 年度 情報処理学会関西支部 支部大会 講演論文集, Vol. 2016, , 2016.
[9] 大島純, 新原勇介, 太田健介, 大島律子. 協調学習のプロセスと個人の貢献を測定する試み: 発言のネットワークを用いた学習者の対話分析 (¡ 特集¿ 協調学習とネットワーク・コミュニティ). 日本教育工学会論文誌, Vol. 33, No. 3, pp. 333–342, 2010.
[10] 森幹彦, 八村太輔, 喜多一. リフレクションのための逐語議事録を用いた議論の構造化法. 人工知能学会全国大会論文集 第 21 回 (2007), pp. 2D41–2D41. 一般社団法人 人工知能学会, 2007.

ChatGPT を利用した学術論文読解支援の一手法

堺 陸斗 †　金澤 輝一 ††　上野 史 †††　太田 学 †††

† 岡山大学工学部情報系学科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †pz1442nn@s.okayama-u.ac.jp, ††tkana@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　学術論文には未知の専門用語や新たな概念が頻繁に登場するため，重要語や被引用文，専門用語の意味へのアクセスを容易にする学術論文閲覧支援の研究が行われている．本稿では，重要情報の提示を通じて，論文の読みながらの理解を促進する読解支援手法とそれを提示するインタフェースを提案する．具体的には，ChatGPT を用いて学術論文から様々なデータを抽出し，重要語のタグクラウド，提案手法のフローチャート，論文中の図表の説明文，論文の概要文，各節の要約文を生成し，これらを論文 PDF とともに読者に提示する．被験者実験では，論文の内容に関する質問への回答時間と正答率を計測し，読解支援インタフェースの使用感や理解の深さ，時間に対する影響をアンケートによって調査した．その結果，提案インタフェースの使用が論文を読む際の理解にかかる時間に顕著な影響を与えるとは言えないものの，被験者 10 人中 7 人は，普段より理解が少し早くなったまたは早くなったと回答し，理解が遅くなったと回答した人はいなかった．一方，10 人中 5 人は理解が少し深くなったと回答したが，理解が深くなったと回答した被験者はおらず，2 人は理解が少し浅くなったと回答した．論文の概要文と各節の要約文については，10 人中 9 人が役立ったと回答した．しかし，重要語タグクラウドとフローチャートについては，論文によって評価が異なった．また，図表説明文については，7 人が役立たなかったと回答し，重要語タグクラウドとフローチャート，および図表説明文の質や利便性の改善が必要であることが示された．

キーワード　情報抽出，学術論文，読解支援，ChatGPT

1 はじめに

研究者は学術論文を迅速に理解し，最新の知見を効率的に吸収することが求められている．最近では，タブレット端末や電子書籍リーダーの普及により，電子端末を使用した学術論文の閲覧支援に関する研究が進められている．しかし，これらの研究は主に情報へのアクセスの容易化に焦点を当てており，論文内容の迅速かつ深い理解を促進する読解支援は十分には着目されていない．

2022 年 11 月 30 日に公開された ChatGPT[1]は，OpenAI によって開発された大規模言語モデル（Large Language Model: LLM）を利用し，ユーザーの入力に応じた回答を提供するチャットサービスである．2024 年 1 月現在において公開されている最新モデルである GPT-4（Generative Pretrained Transformer 4）は，OpenAI [1] のベンチマークテストによると，従来の GPT-3.5 よりも幅広い知識と優れた問題解決能力，より精度の高いテキスト生成，文脈理解の向上，及びより自然な言語生成能力を備えているとされている．GPT-4 のこれらの特性は，論文の重要な情報を抽出し解釈するツールとして適している．

本研究では，論文の迅速かつ深い理解を促進するために，GPT-4 を基盤とする ChatGPT を用いた学術論文読解支援手法と，それを提示するインタフェースを提案する．具体的には，重要語のタグクラウド，論文が提案する手法のフローチャート，図表の説明文，論文の概要文，各節の要約文を生成し，これらの読解支援情報を読者に提示する．ChatGPT には論文の PDF とプロンプトを入力し，必要なデータを表示するコードを出力させる．その後，出力されたコードを利用して，読者がインタフェースを通じて読解支援情報を閲覧できるようにする．評価実験では，論文内容の理解を問う質問フォームを用い，回答時間とその正答率を計測する被験者実験を行う．また，アンケートにより，提示インタフェースの使用感や，読解支援情報が論文を読む際の理解にかかる時間や理解の深さにどのような影響を与えたかを調査する．

本稿の構成は以下の通りである．2 節では学術論文閲覧支援と学術研究における ChatGPT の利用に関連する研究を紹介する．3 節では本稿で提案する学術論文読解支援手法を述べる．4 節では読解支援情報の提示インタフェースを用いた被験者実験について述べる．5 節ではまとめを述べる．

1：https://chat.openai.com/

2 関連研究

2.1 学術論文閲覧支援

谷尻ら [2] は，タブレットのカメラ機能を活用して紙媒体の論文テキストを読み取り，単語の重要度を表示するインタフェースを提案した．岩本ら [3] は，谷尻らのインタフェースを拡張し，ユーザがスワイプ操作で引いたマーカーの履歴を利用した検索結果の個人化手法を提案した．これらの研究は，タブレットでの論文 PDF の閲覧を想定し，タブレットの特長的な機能

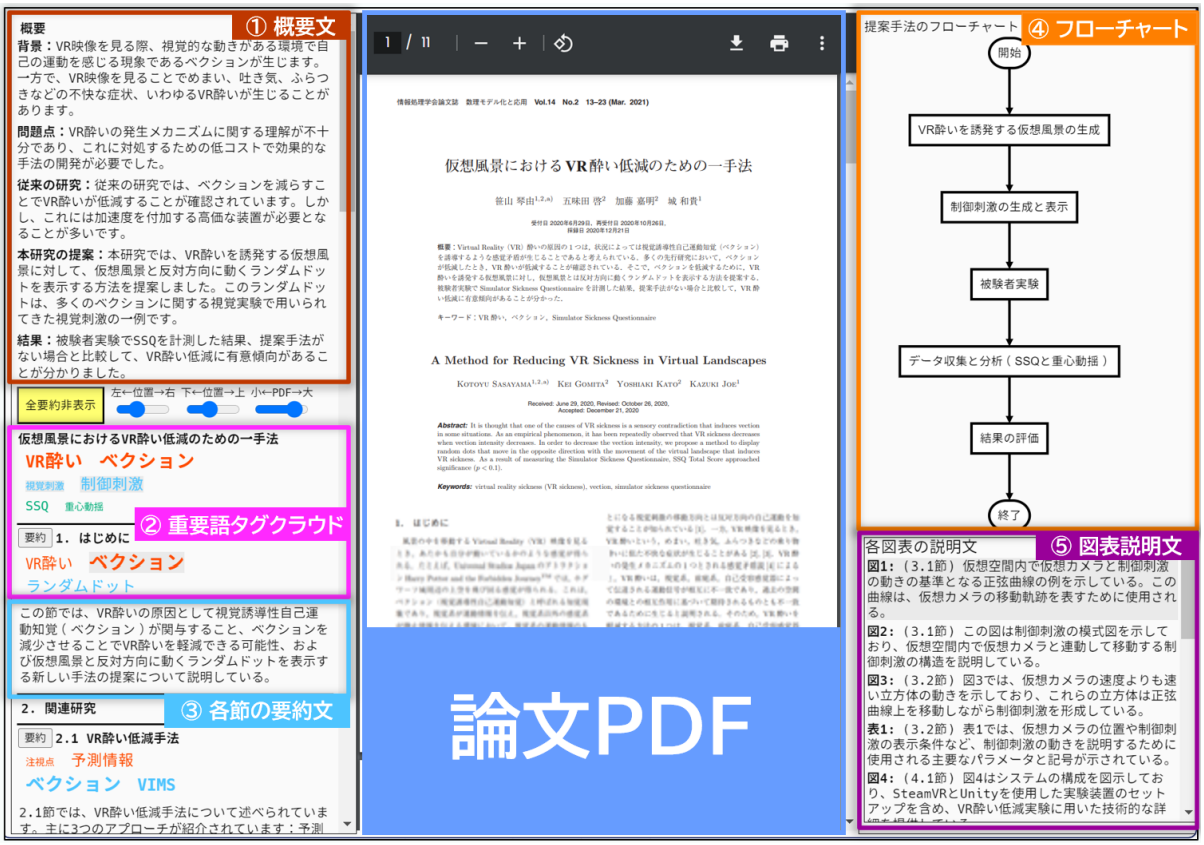


図 1: 学術論文読解支援情報提示インタフェースの画面

や操作感を活かす閲覧支援機能を提案している．

阿辺川ら [4] は，論文閲覧システム SideNoter を提案した．SideNoter は，専門用語検索，連続ページめくり，注釈追加，拡大縮小表示など，Web ブラウザで論文を読む際に必要な閲覧支援機能を提供する．さらに，本文中のキーワード解説やページに関連する情報を外部情報源を用いて脚注に表示することで，内容理解を深める読解支援機能も備えている．

これらの研究から，論文閲覧支援は，媒体の特性に応じた機能を通じて，読者が必要とする情報に容易にアクセスし，論文内容に集中できるように支援する．一方で，論文読解支援は，論文内容に基づく情報の提示を通じて，読者が論文の全体構造，主要概念，テーマを迅速かつ深く理解できるように支援する．この情報のインタフェース上での提示には，論文閲覧支援の知見を活かすことが重要である．閲覧支援と読解支援は，それぞれ情報アクセスの容易化と内容理解の促進という異なる役割を持ちながらも，読者が論文を全面的に理解する過程で相補性がある．

本研究で提案する読解支援情報は，重要語の論文中での意味や各節の要約文，概要文などを通じて，論文の全体構造や主要概念の迅速かつ深い理解を支援する．これらの情報は，Web ブラウザの機能を活用して読者が容易にアクセスできるように提示され，論文理解を促進する．

2.2 学術研究における ChatGPT の利用

GPT-4 は，OpenAI の報告 [1] によると，同社の従来モデルである GPT-3.5 や他の LLM と比較して，自然言語処理のベンチマークで優れた性能を示している．これには，より精度の高いテキスト生成，文脈理解の向上，及びより自然な言語の生成が含まれる．しかし，従来モデルと同様に，信頼性が求められる状況で GPT-4 の出力を使用する際は注意が必要であると述べられている．笠井ら [5] は，GPT-4 を基盤とする ChatGPT が日本の過去 5 年間の医学国家試験全てに合格するという結果を示し，GPT-4 の日本語文章の理解能力の高さを報告している．一方で，日本の医療行為では回避すべき選択肢を選択することもあったと報告している．

これらの研究から，GPT-4 および ChatGPT が複雑な問題の理解や多言語のテキスト処理において高い能力を有することが示されている．特に，ChatGPT は自然言語処理の様々な応用分野において優れた結果を示している．このため，本研究では GPT-4 を基盤とする ChatGPT を利用し，論文内の重要語や要約文などの読解支援情報を生成する．生成された情報の信頼性を確保するため，生成された情報を元の論文とともに提示する．

3 学術論文読解支援手法

本研究では，学術論文読解支援情報として，重要語のタグクラウド，提案手法のフローチャート，図表の説明文，論文の概要文と各節の要約文を生成し，読者に提示するための手法を提案する．3.1 節では読解支援情報の提示インタフェースについて，3.2 節では学術論文読解支援情報について，3.3 節では GPT-4 に基づく ChatGPT による学術論文読解支援情報の生

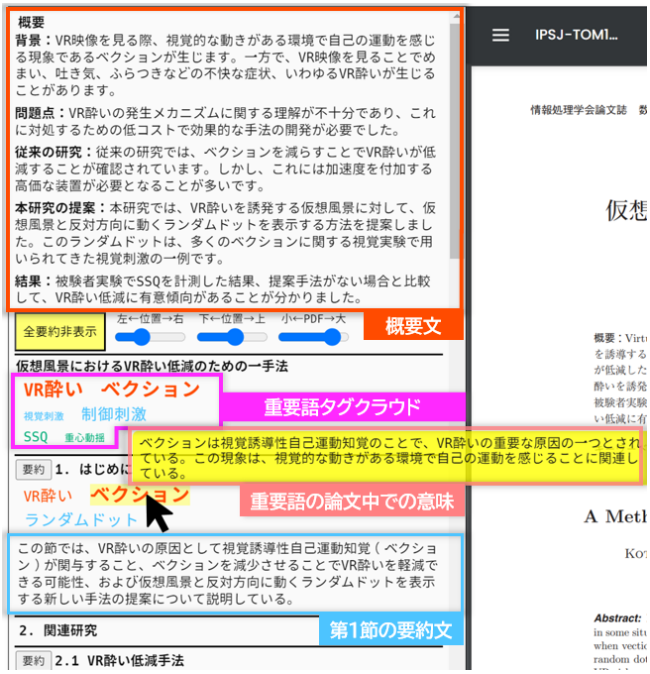


図 2: 重要語タグクラウドと重要語の文脈の表示例

成と提示について述べる．

3.1 学術論文読解支援情報の提示インタフェース

図 1 は，提案インタフェースの画面であり，中央には笹山らの論文 [6] の PDF が表示されている．論文の左側には論文の概要文 (図 1①) と重要語タグクラウド (図 1②)，各節の要約文 (図 1③)，論文の右側にはフローチャート (図 1④)，図表説明文 (図 1⑤) が表示されている．各節の要約文には，節毎に非表示と再表示を切り替えられる機能と，全ての節の要約文の非表示と再表示を切り替えられる機能を備えている．また，マウスポインタを特定の要素に重ねる (マウスホバー) ことで表示されるツールチップ [2]の位置，また中央に表示される論文 PDF の横幅を，読者の環境や好みに合わせて調整できるスライドバーを備えている．

3.2 学術論文読解支援情報

本節では，笹山らの論文 [6] を用いて学術論文読解支援情報を生成し，重要語タグクラウドと論文の概要文および各節の要約文の生成例を示す図 2，フローチャートの生成例を示す図 3，図表説明文の生成例を示す図 4 にはその生成例を示す．

3.2.1 重要語タグクラウド

重要語タグクラウドは，文章から抽出した重要語を重要度のランクに応じて表示サイズを変更し，意味的類似性や文脈的関連性に基づいて関連性のある重要語を同じ色で表示することで重要語の可視化を実現する．また，重要語にマウスホバーすることで，その重要語の論文中での意味をツールチップとしてマウスポインタの近くに表示する．これにより，論文中の重要語の迅速な理解を促進する．図 1 の提示インタフェースでは，論文全体および各節に対して重要語タグクラウドを表示する．

2：特定の要素にマウスポインタを合わせることで表示される小さな領域．

図 2 には，論文 [6] を例に論文全体と第 1 節の重要語タグクラウドの生成例，および「ベクション」にマウスホバーした際の様子を示している．重要度は，最も重要なものをランク 1 とし，1 から 3 までの 3 段階で評価する．ランク 1 の重要語は太く大きく表示され，読者はその重要語の重要度を視覚的に把握できる．図 2 の論文全体のタグクラウドでは，「VR 酔い」と「ベクション」がランク 1，「制御刺激」と「SSQ」がランク 2，それ以外がランク 3 である．また，橙の「VR 酔い」と「ベクション」，青の「視覚刺激」と「制御刺激」，緑の「SSQ」と「重心動揺」は，それぞれ関連性のある重要語である．

3.2.2 フローチャート

図 3 の右に示すフローチャートは，論文が提案する手法の流れを視覚的に示すものである．また，マウスポインタをフローチャート中の各処理に重ねることで，その処理の詳細と処理に最も関連が強い節番号をツールチップとしてマウスポインタの近くに表示する．各処理をクリックすることで，ツールチップに記載された節番号のタグクラウドが画面左上にスクロールされ，詳細を確認できる．

図 3 には，論文 [6] を例に，フローチャートの生成例と，「制御刺激の生成と表示」にマウスホバーおよびクリックした際の様子を示している．図内の青い矢印に従って視線を移すと，「制御刺激の生成と表示」に最も関連する節の重要語タグクラウドや節の要約文が確認できる．このフローチャートを通じて，手法の全容を把握させ，本文読解を促進する．

3.2.3 図表説明文

図表説明文は，論文中の図表についての説明を，本文を基にして生成したものである．図表説明文には，各図表に最も関連が強い節番号が記載されており，クリックすることで図 3 に示したフローチャートと同様に該当する節のタグクラウドと要約文が画面左上にスクロールされ，詳細を確認できる．図 4 には，論文 [6] の図表説明文の生成例を示している．図表説明文の表示領域をスクロールすることで，論文中の全ての図表に対する説明文を確認できる．論文中の図表とそれに対応する本文が別ページに掲載されていることがあり，電子端末での学術論文閲覧時には，これらを見るために画面を行き来する必要があるが，図表説明文はこの問題を軽減する．

3.2.4 論文の概要文と各節の要約文

論文の概要文は，論文から背景，従来の研究，問題点，論文の提案，結果を本文に基づいて生成したものである．図 2 のように，論文の概要文はインタフェースの左上に提示し，各節の要約文は，各節の重要語タグクラウドの下に提示する．これらの要約文を通じて，論文全体や各節の内容を把握させる．

3.3 ChatGPT による学術論文読解支援情報の生成と提示

ここでは，ChatGPT を使用した重要語のタグクラウド，フローチャート，図表説明文，論文の概要文，各節の要約文の作成に必要なデータの生成方法と提示インタフェースへの提示方法について説明する．提示インタフェースは HTML，CSS，JavaScript により作成され，論文 PDF は HTML の<embed>タグを用いて埋め込まれる．

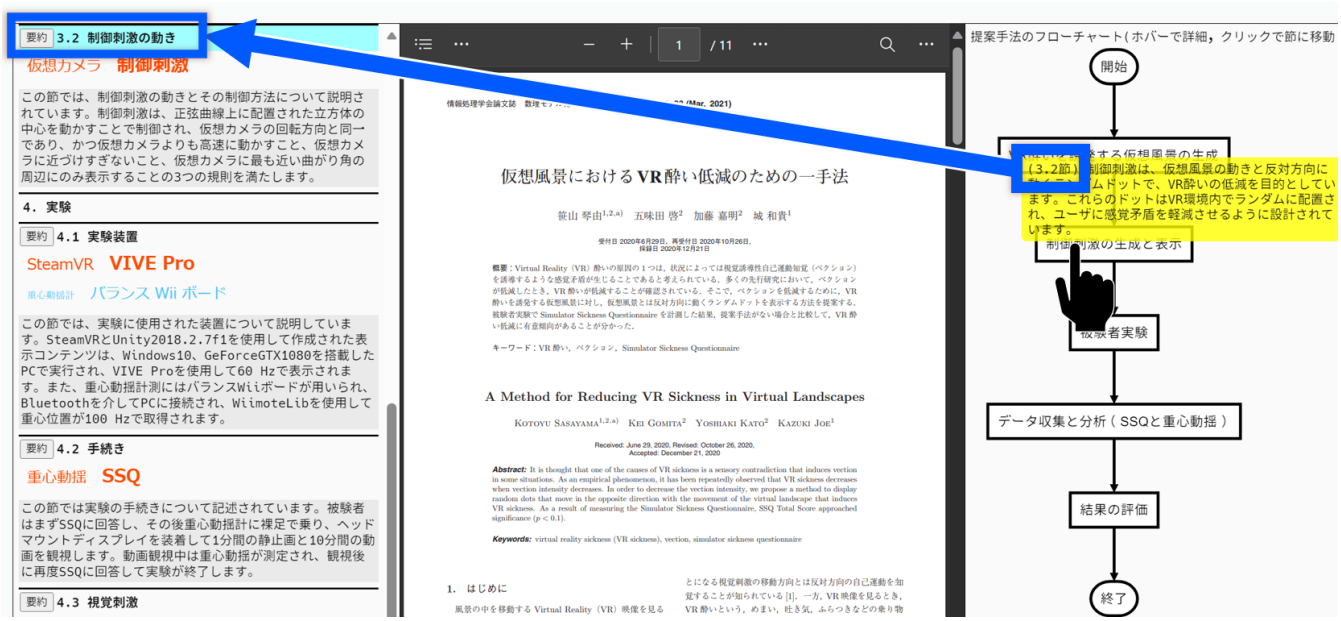


図 3: フローチャートとフロー詳細の表示例

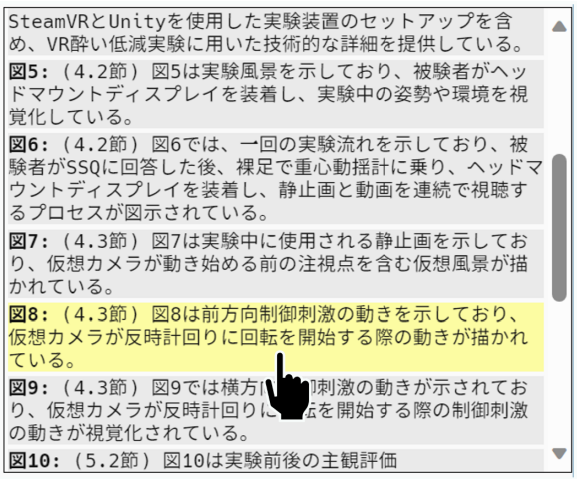


図 4: 図表説明文の表示例

以下の指示で，出力形式に従ったJavascript関数の引数を作成してください．
・指定する節の要約を3文程度の日本語で生成してください．
・論文の指定する節から重要語を抜き出し，その重要度を，最重要を1として，1から3までのランクを付けてください．
・各重要語の論文中での意味を日本語で2文程度で生成してください．
・重要語同士の内容の関連性に基づき，関連する重要語をグループ化してください．
#指定する節
1. はじめに
#出力形式
createSecTag(“sec3_1“,“3.1 節タイトル“, “3.1節の要約“,
[[/*グループ1*/ ["重要語A", 1, "文脈A"],
["重要語B", 2, "文脈B"],],
[/*グループ2*/ ["重要語C“ , 1, "文脈C"],
["重要語D", 2, "文脈D"],["重要語E", 3, "文脈E"],],
[/*グループ3*/ ["重要語F", 3, "文脈F"],],]);

図 5: 重要語タグクラウドと各節の要約文生成のためのプロンプト

GPT-4 を基盤とする ChatGPT は，画像や PDF などのテキスト以外の入力に対応し，これを元にユーザからのメッセージに応答することができる．この機能を活用して，読解支援情報を生成する論文 PDF を入力し，必要な情報に応じてプロンプトを変更することで，各読解支援情報用のデータを生成する．

各節の重要語タグクラウドと要約文を生成するために使用したプロンプトの例を図 5 に示す．重要語やそのランク付け，重要語の論文中での意味，グループ化に関する指示を送信し，これらの出力は図 5 の出力形式に示す `createSecTag()` 関数の引数として利用する．提示インタフェース内には `createSecTag()` 関数の処理が事前に定義されているため，図 6 に示す出力されたコードをインタフェースのプログラムに追加することで，指定した節の重要語タグクラウドと要約文が利用できる．論文全体の重要語タグクラウドにおいては，指定する節を論文全体に置き換え，要約に関する指示を除いたプロンプトとする (図 7)．

フローチャートを生成するために使用したプロンプトの例を図 8 に示す．フローチャートは JavaScript のライブラリである Flowchart.js [3]の文法で出力され，それとは別に補足情報を表示するためのコードが生成される．このコードは，各フローに最も関連が強い節番号も含む．

図表説明文を生成するために使用したプロンプトの例を図 9 に示す．図表説明文は，図表番号やそのキャプションを指定し，図表ごとにコードを出力させる．このコードは，図表に最も関連が強い節番号も含む．

論文の概要文を生成するために使用したプロンプトを図 10 に示す．これらのコードも，インタフェースのプログラムに追加して利用される．ただし，節ごとの重要語タグクラウドや要

3：https://flowchart.js.org/

```
createSecTag( "sec1", "1. はじめに", "この節では、VR酔いの原因として視覚誘導性自己運動知覚（ベクション）が関与すること、ベクションを減少させることでVR酔いを軽減できる可能性、および仮想風景と反対方向に動くランダムドットを表示する新しい手法の提案について説明している。",

[["ベクション", 1, "ベクションは視覚誘導性自己運動知覚のことで、VR酔いの重要な原因の一つとされている。この現象は、視覚的な動きがある環境で自己の運動を感じることに関連している。",],
 ["VR酔い",2, "VR酔いは、Virtual Reality環境で体験する不快な症状で、めまいや吐き気などを引き起こすことがある。この酔いは、ベクションと密接に関連している。",],],

 [["ランダムドット", 2, "ランダムドットは仮想風景と反対方向に動く視覚刺激で、本研究ではVR酔いを減少させる新しい手法として提案されている。この手法は、ベクションを減少させることによりVR酔いを軽減することを目的としている。",],],]);
```

図 6: 各節の重要語タグクラウドで出力されたコードの例

```
以下の指示で，指定した出力形式に従ったJavascript関数の引数を作成してください.
・論文全体から重要語を抜き出し，その重要度を最重要を1として，1から3までのランクを付けてください.
・各重要語の論文中での意味を日本語で2文程度で生成してください.
・重要語同士の内容の関連性に基づき，関連するものをグループ化してください.
#出力形式
createAllTag([[ /*グループ1*/["重要語A", 1, "文脈A"],["重要語B",2, "文脈B"],],
   [ /*グループ2*/["重要語C" , 1, "文脈C"],["重要語D", 2, "文脈D"],["重要語E", 3, "文脈E"],],
   [ /*グループ3*/["重要語F", 2, "文脈F"],],]);
```

図 7: 論文全体の重要語タグクラウド生成のためのプロンプト

```
論文が提案する手法の流れを示すフローチャートを
Flowchart.jsで作成してください.
また，それぞれの処理の日本語の解説を，その解説
が書かれた節番号とともにブロックのidごとに 2文
程度で出力形式に従って生成してください.
#出力形式
const chart = `st=>start: 開始
               op1=>operation: 処理1
               cond=>condition: 条件
               yes=>operation: はい
               no=>operation: いいえ
               end=>end: 終了
               st->op1->cond
              cond(yes)->yes
              cond(no)->no
              yes->end
              no->end`;
const chartList =[[“id1”,”sec3_1”,”(3.1節)解説1”],
          [“id2”,”sec3_2_1”,”(3.2.1節)解説2"],];
```

図 8: フローチャート生成のためのプロンプト

```
指定した図または表について，論文をもとにした2文程度の日本語の解説を，
図表と最も関連のある節番号とともに以下の出力形式に従ってJavaScript
形式で出力してください.
図表を特定できない場合は，キャプション名から解説を生成してください.
#指定する図表
表1 制御刺激の動きの説明で使用する記号
#出力形式(図1の場合)
const fig1Desc = createDesc("図1","sec3_1","(3.1節)図1の説明");
#出力形式(表2の場合)
const tab2Desc = createDesc("表2","sec3_2_1","(3.2.1節)表2の説明");
```

図 9: 図表説明文生成のためのプロンプト

```
以下の指示で，指定した出力形式に従ったJavascript関数の引数を
作成してください．論文の「背景」「問題点」「従来の研究」
「本研究の提案」「結果」を，それぞれ2文の日本語で生成して
ください.
#出力形式
createAbstract("背景","問題点","従来の研究","本研究の提案","結果");
```

図 10: 概要の要約文生成のためのプロンプト

約文の生成，節タイトルの抽出，図表説明文の生成のための論文からの図表キャプションの抽出，ChatGPT の出力からインタフェースへの貼付けは人手で行っている．これらの処理の自動化は今後の課題である．

4 評 価 実 験

評価実験では，提案インタフェースの有効性を，質問フォームとアンケートを通じて評価する．質問フォームは，論文の内容理解を確かめるため，正誤形式，選択形式，並べ替え形式の質問から構成される．質問は，論文の主要な内容や用語の意味を問う問題や，特定の箇所の正誤を判断するために論文全体から根拠を探す必要がある問題で構成される．アンケートは，提案インタフェースの各読解支援情報の有用性や，提案インタフェースの使用が被験者の主観的な論文理解の早さや深さに与える影響，およびインタフェースの利点や改善点について質問する．

4.1 実 験 内 容

本実験では，情報系の岡山大学工学部の 4 年生 5 名，同大学大学院の博士前期課程 1 年生 3 名，2 年生 2 名の合計 10 名を対象に，情報処理学会の論文誌に掲載された笹山らの論文 [6] と同学会の研究報告である倉林らの論文 [7] を読ませる．以降，

表 1: グループごとの使用するインタフェース

	論文 A	論文 B
被験者グループ 1	提案インタフェースの支援有	提案インタフェースの支援無
被験者グループ 2	提案インタフェースの支援無	提案インタフェースの支援有

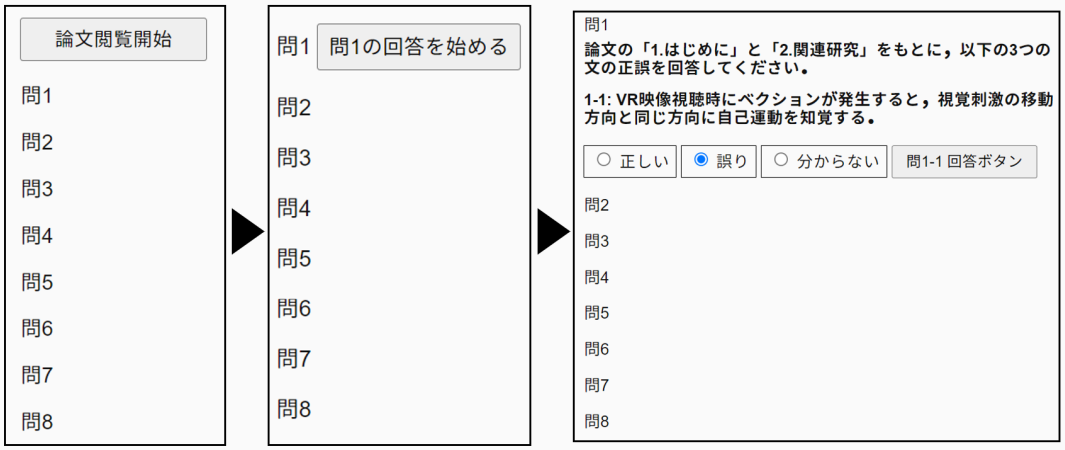


図 11: 質問フォーム

笹山らの論文を論文 A，倉林らの論文を論文 B と表記する．

実験の流れについて，被験者は論文 A を読みながら質問フォームに回答し，全ての質問に回答した後，アンケートに回答する．その後，論文 B も同様に質問フォームとアンケートに回答する．

10 名の被験者を各グループ 5 名ずつの 2 つのグループに分ける．グループ分けは，以下の基準でグループ間の偏りが小さくなるように本稿の第一著者が行った．

- 過去 1ヶ月間に読んだ論文の本数
  （1） 0 本以上 1 本以下
  （2） 2 本以上 4 本以下 (1 週間あたり 1 本程度)
  （3） 5 本以上 10 本以下 (1 週間あたり 2 本程度)
  （4） 11 本以上
- 論文理解の得意不得意
  （1） 全く得意ではない
  （2） あまり得意ではない
  （3） まあまあ得意である
  （4） かなり得意である
- 対象論文における専門用語への理解度
  （1） 聞いたことがない
  （2） 聞いたことはあるが意味は知らない
  （3） 意味をなんとなく知っている
  （4） 意味を説明できる

ここで，論文理解の得意不得意については，短時間での大まかな理解と，長時間での詳細な理解の 2 つの観点で回答させる．対象論文における専門用語への理解度については，該当する論文の分野や使用された手法，および評価指標に基づき人手で選んだ各専門用語に対する理解度を，上述の選択肢を用いて回答させる．

被験者グループ 1 と被験者グループ 2 の提案インタフェースの支援の有無を表 1 に示す．提案インタフェースを使用しない場合は，ブラウザの標準 PDF ビューアを利用する．提案インタフェースを使用する場合，その使い方を直前に確認させる．どちらのグループも，検索エンジンによる単語の検索，ページ内検索，PDF へのハイライトやペンでの書き込み，コメントの追加といったブラウザの標準機能の利用を許可している．ただし，ChatGPT や Microsoft Edge の AI チャット機能等の生成系 AI の使用は，不明な単語を尋ねることにのみに限定し，論文の内容や文章を直接使用しての要約や質問は許可しない．

4.2 計測項目

本実験で使用する質問フォームを図 11 に示す．質問フォームは，論文の閲覧を開始するボタン，各問題文や選択肢を表示するボタン，各問題の回答を確定するボタンをクリックして回答を進める．これらのボタンをクリックすると時刻が 0.1 秒単位で記録される．最後の問題の回答を確定した時刻を閲覧終了時刻とする．被験者は，回答を確定するボタンを押すまで，次の問題以降の問題文や選択肢を閲覧できない．一度回答を確定すると，その問題の問題文や選択肢，自身の回答の確認や変更はできない．

本実験で使用する学術論文読解支援情報の提示インタフェースは，被験者の操作ログをバックグラウンドで取得する機能を組み込んでおり，外見は図 1 と同じである．重要語タグクラウド，フローチャート，図表説明文，論文の概要文，各節の要約文は各要素にマウスホバーをすることでログを取得する．これらの要素とマウスポインタが重なった瞬間と終了した瞬間の時刻を 0.1 秒単位で記録する．被験者には，これらの要素を閲覧中にマウスホバーをするよう指示している．ただし，ホバーの開始と終了時刻の差が 1 秒未満の場合は，マウスポインタがその要素を通過しただけとみなし，記録しない．クリックによる

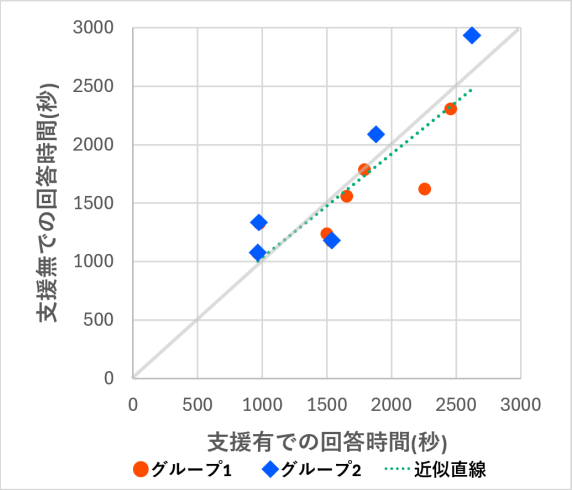


図 12: 支援有りと支援無しの回答時間の関係

ログの取得対象はフローチャートと図表説明文であり，これらの要素をクリックした時刻も 0.1 秒単位で記録される．

4.3 実験結果と考察

実験の結果，論文 A で提案インタフェースを使用したグループ 1 から 5 人，論文 B で同インタフェースを使用したグループ 2 から 5 人のデータを取得した．4.3.1 項では質問フォームの正解率と回答時間について，4.3.2 項では提案インタフェースの使用状況について，4.3.3 項では提案インタフェースの評価について実験結果と考察を述べる．

4.3.1 質問フォームの正解率と回答時間

表 2 と表 3 は，論文 A と論文 B の質問フォームの質問概要と，グループごとの正解率と回答時間を示し，Grp. 1 はグループ 1 を表す．論文 A の全問題の平均正解率は，支援有りのグループが支援無しのグループより 2.3% 低く，論文 B ではどちらのグループも 76.7% となった．論文 A において，支援ありのグループでは全 8 問中 2 問が支援なしのグループより高い正解率を，3 問が低い正解率を，そして残り 3 問が同等の正解率を示した．論文 B では，支援ありのグループが全 7 問中 4 問で高い正解率，2 問で低い正解率を，残りの 1 問で同等の正解率を示した．また，回答時間に関しては，両論文において支援有りのグループが 5 問でより長い時間を要し，質問フォーム全体の回答時間の平均も支援有りのグループが長い時間を要した．

図 12 は，支援有りと支援無しの回答時間の関係を示しており，緑の点線は 10 件のデータを最小二乗法により求めた近似直線を表す．灰色の線は，支援有りと支援無しの回答時間が同じ場合を表す線である．この図から，支援がある場合に必ずしも回答時間が支援無しの場合より短くなるわけではないことが読み取れる．

4.3.2 提案インタフェースの使用状況

図 13 と図 14 は，被験者ごとの提案インタフェースの使用状況を示す．1 行目の左端から 2 行目の右端にかけて，各問題の回答を開始したタイミングと各読解支援情報の閲覧を開始したタイミングを，凡例に示す数字または記号の左端が一致するように配置している．読解支援情報を 5 秒以上閲覧し続けた場合は，5 秒ごとに 1 個ずつ該当の記号を追加で配置している．

図 13 と図 14 より，9 名の被験者は論文の概要文を論文閲覧開始から 3 分以内に閲覧したことが分かる．被験者クと被験者ケは，提案された手法の流れに関する問題 3 においてグループ 2 の他の被験者と比較してフローチャートを多くの回数閲覧していた．図表説明文は，グループ 2 の方がグループ 1 より閲覧された回数が多く，被験者間での閲覧回数には大きな差が見られる．重要語タグクラウドと各節の要約文は，多くの被験者に複数回閲覧されていた．

4.3.3 提案インタフェースの有用性評価

表 4 と表 5 は，被験者アンケートの結果に基づき，提案インタフェースの各読解支援情報が役に立ったかの回答を示す．

重要語タグクラウドについて，グループ 1 では 5 人全員が「役立った」または「とても役立った」と回答したが，グループ 2 では 5 人中 4 人が「あまり役立たなかった」または「全く役立たなかった」と回答しており，逆の評価となった．グループ 1 の被験者オから，同じ重要語に異なる意味が付与される点が良かったという意見が得られた．よって重要語タグクラウドは，読解支援に有用なタグクラウドを生成できない場合もあることが分かった．

フローチャートについては，グループ 1 では 5 人中 4 人が「あまり役立たなかった」と回答したが，グループ 2 では 5 人中 4 人が「役立った」と回答しており，論文 A と論文 B で逆の評価となった．被験者エから，フローチャートは論文の内容を素早く理解するのに役立ったという意見が得られた．

図表説明文については，両グループともに過半数が「あまり役立たなかった」と回答した．図 13 と図 14 から，図表説明文を閲覧していない被験者が 10 人中 3 人おり，図表説明文の閲覧しやすさには改善の余地があると考えられる．提案インタフェースでは，<embed>タグで使用される PDF ビューアの制約により実装できなかったが，PDF.js[4] や pdf2htmlEX[5] といったライブラリを用いて，論文中の図表と図表説明文を連携させる機能を検討する必要がある．

論文の概要文と各節の要約文は，両グループともに評価が高く，図 13 と図 14 から，大半の被験者が複数回利用したことが分かる．被験者エ，被験者カ，被験者ケから，論文の概要文や各節の要約文の内容が分かりやすかったという意見や，それらが既に読んだ本文内容の確認に役立ったという意見，提案インタフェース上での要約文は読みたい部分を効率的に探すのに役立ったという意見が得られた．これらは，ChatGPT による論文の概要文や各節の要約文の内容や，インタフェースでの提示方法の有効性を裏付けるものといえる．

表 6 と表 7 は，提案インタフェースの使用による理解の早さや深さに対する被験者の主観を示す．表 6 によると，提案インタフェースの使用によって論文の理解の早さが「少し遅くなった」または「遅くなった」と感じた被験者はいなかった．

4：https://mozilla.github.io/pdf.js/
5：https://texwiki.texjp.org/?pdf2htmlEX

表 2: 論文 A の質問概要とグループごとの正解率と回答時間

問	質問概要	回答形式 [選択肢数] (小問数)	正解率 (%) 支援有 (Grp. 1)	正解率 (%) 支援無 (Grp. 2)	回答時間 (秒) 支援有 (Grp. 1)	回答時間 (秒) 支援無 (Grp. 2)
1	1 節と 2 節について	正誤 (3 問)	**80.0**	66.7	537.1	**352.1**
2	仮想風景の動き	選択 [4 択]	40.0	**60.0**	314.8	**214.9**
3	制御刺激について	選択 [4 択]	**60.0**	**60.0**	**200.9**	334.6
4	被験者実験の手続き	並べ替え	60.0	**80.0**	151.1	**122.7**
5	実験の測定項目	選択 [5 択]	**100.0**	**100.0**	**141.5**	156.7
6	実験結果	正誤 (4 問)	65.0	**85.0**	**290.7**	297.1
7	ベクションの意味	選択 [5 択]	**100.0**	**100.0**	30.2	**27.1**
8	特定箇所の内容について	正誤 (4 問)	**65.0**	55.0	264.6	**212.0**
平均	—	—	71.8	**74.1**	2027.4	**1717.3**

表 3: 論文 B の質問概要とグループごとの正解率と回答時間

問	質問概要	回答形式 [選択肢数] (小問数)	正解率 (%) 支援有 (Grp. 2)	正解率 (%) 支援無 (Grp. 1)	回答時間 (秒) 支援有 (Grp. 2)	回答時間 (秒) 支援無 (Grp. 1)
1	従来手法の課題と解決法	選択 [4 択]	**60.0**	40.0	**238.5**	501.6
2	遺伝的プログラミングの内容	正誤 (4 問)	70.0	**75.0**	488.8	**390.9**
3	提案手法の流れ	並べ替え	**80.0**	60.0	176.6	**123.8**
4	提案手法の評価方法	正誤 (3 問)	73.3	**93.3**	230.5	**220.9**
5	実験結果	正誤 (4 問)	**65.0**	**65.0**	271.5	**244.0**
6	TF-IDF の意味と用途	選択 [5 択]	**100.0**	90.0	**52.5**	93.8
7	特定箇所の内容について	正誤 (3 問)	**93.3**	86.7	137.0	**127.1**
平均	—	—	**76.7**	**76.7**	1747.8	**1595.4**

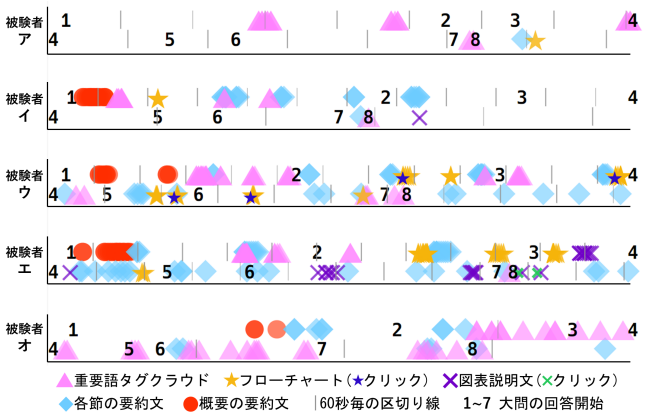


図 13: グループ 1 の被験者ごとの提案インタフェース使用状況

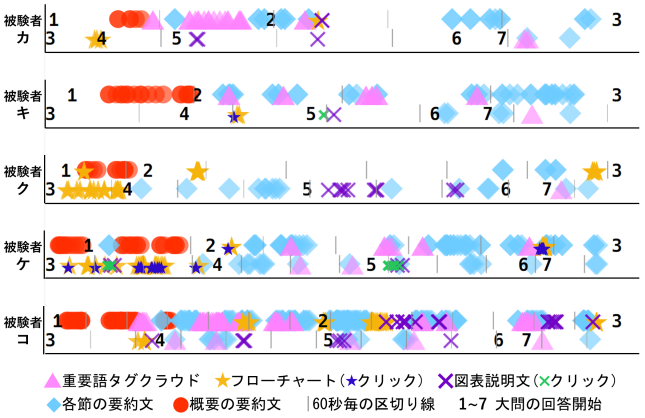


図 14: グループ 2 の被験者ごとの提案インタフェース使用状況

表 7 によると，理解の深さに関しては，グループ 1 の 5 人全員が「普段と変わらない」または「少し深くなった」と感じているものの，「深くなった」と回答した被験者はいなかった．

表 8 は，提案インタフェース全体の評価を示す．両グループともに，提案インタフェースは高い評価だった．提案インタフェースは全体的に評価が高かったものの，各読解支援情報については，論文により異なる評価が得られた．論文の概要文と各節の要約文は，読解支援情報として役立つことが示された．一方で，重要語タグクラウドの改善，フローチャートの適切な使用，図表説明文の利便性向上が，より効果的な読解支援に必要である．

今後の課題として，重要語タグクラウドやフローチャートの有用性を高めるために ChatGPT のプロンプトを変更し，図表の説明文の利便性を向上させるために論文の PDF ファイルに直接作用する機能を追加することが挙げられる．

表 4: グループ 1 の論文 A の実験において提案インタフェースが役立ったかの回答

	全く役立たなかった	あまり役立たなかった	まあまあ役立った	とても役立った
重要語タグクラウド	0	0	4	1
フローチャート	0	4	0	1
図表説明文	0	3	2	0
各節の要約文	0	0	1	4
論文の概要文	0	0	4	1

表 5: グループ 2 の論文 B の実験において提案インタフェースが役立ったかの回答

	全く役立たなかった	あまり役立たなかった	まあまあ役立った	とても役立った
重要語タグクラウド	1	3	1	0
フローチャート	0	1	4	0
図表説明文	0	4	1	0
各節の要約文	0	1	3	1
論文の概要文	0	1	3	1

表 6: 提案インタフェースの使用によって論文の理解が早くなったかの回答

	遅くなった	少し遅くなった	普段と変わらない	少し早くなった	早くなった
論文 A(グループ 1)	0	0	0	4	1
論文 B(グループ 2)	0	0	3	2	0

表 7: 提案インタフェースの使用によって論文の理解が深くなったかの回答

	浅くなった	少し浅くなった	普段と変わらない	少し深くなった	深くなった
論文 A(グループ 1)	0	0	1	4	0
論文 B(グループ 2)	0	2	2	1	0

表 8: 提案インタフェース全体の評価

	全く役に立たない	あまり役に立たない	まあまあ役に立つ	とても役に立つ
グループ 1	0	0	3	2
グループ 2	0	0	5	0

5 ま と め

本稿では，論文中の重要語や要約文等の重要情報の提示を通じて，論文を読みながらの理解を促進する読解支援手法とその提示インタフェースについて述べた．具体的には，GPT-4 を基盤とする ChatGPT を使用して学術論文から重要語タグクラウド，フローチャート，図表説明文，論文の概要文，各節の要約文の作成に必要なデータを生成し提示する手法を提案した．

評価実験では，提案インタフェースの使用の有無が論文の迅速かつ深い理解に与える影響を質問フォームとアンケートを通じて評価する被験者実験を行った．その結果，提案インタフェースの使用が論文を読む際の理解にかかる時間に顕著な影響を与えるとは言えないものの，多くの被験者から理解が少し早くなったとの回答が得られた．しかし，理解の深さについてはあまり変わらない結果となった．また，提案インタフェース全体について，多くの被験者に高く評価された．提案インタフェース内の個々の読解支援情報については，論文の概要文と各節の要約文が多くの被験者に高く評価された．一方で，重要語タグクラウドとフローチャートは論文によって評価が異なり，また図表説明文は評価が低かったため，質や利便性を改善する必要性が示された．

文 献

[1] OpenAI, “GPT-4 technical report,” arXiv:2303.08774, 2023.
[2] 谷尻淳貴，太田学，高須淳宏，安達淳，“ タブレット端末によるカメラ機能を用いた学術論文閲覧支援の一手法, ” 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), E3-5，2018.
[3] 岩本拓実，金澤輝一，上野史，太田学，“ ユーザの興味を利用した学術論文閲覧支援の一手法, ” 情報処理学会第 83 回全国大会講演論文集，Vol. 2021，No. 1，pp. 535–536，2021.
[4] 阿辺川武，相澤彰子，“内部構造解析機能と脚注表示機能を備えた論文閲覧システム,”人工知能学会第二種研究会資料，Vol. 2014，No. AM-07，pp. 13–18，2014.
[5] Kasai Jungo, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir R. Radev, “Evaluating GPT-4 and Chat-GPT on Japanese Medical Licensing Examinations,” arXiv:2303.18027, 2023.
[6] 笹山琴由，五味田啓，加藤嘉明，城和貴，“ 仮想風景における VR 酔い低減のための一手法, ” 情報処理学会論文誌数理モデル化と応用（TOM），Vol. 14，No. 2，pp. 13–23，2021.
[7] 倉林利行，吉村優，切貫弘之，丹野治門，富田裕也，松本淳之介，まつ本真佑，肥後芳樹，楠本真二，“ 過去の実装情報を活用したプログラム自動生成 ”，情報処理学会研究報告ソフトウェア工学（SE），Vol. 2020，No. 8，pp. 1–8，2020.

大規模言語モデルを用いたクエリ拡張による会話ログの検索

田貝　奈央†　　加藤　　誠††

† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s2321695@u.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　本研究では，会話を円滑に進めるために必要な情報を，現在の会話内容に基づいて，過去の会話ログから抽出するシステムの実現を目指す．この時，会話ログの検索には時間を要するため，できるだけ少ない現在までの発話内容を元に，次に出現しそうな内容を予測し，その内容に関連する過去の会話ログを抽出することが望ましい．具体的には，現在までの発話内容を入力とし，与えられたテキストから次に出てくる単語を逐次的に予測する文章生成モデルである GPT-2 を用いて生成された発話文を基に会話ログを検索するシステムを実現する．実験では，講演音声を対象とした音声ドキュメント検索の評価に使用できる NTCIR-9 SpokenDoc テストコレクションの適合性判定結果を利用し，ある検索トピックに対して適合であると判定された発話文同士を関連発話文とすることで，提案手法の評価を行った．実験の結果，新しいデータセットでも大規模言語モデルを用いたクエリ拡張手法が効果があることがわかった．

キーワード　音声，ユーザ支援，検索モデル，クエリ補完

1　はじめに

一度聞いた情報を，別の機会に言及することは，特別感や信頼感を与える効果的なコミュニケーション手法として知られている [6]．例えば，会話の中でスポーツの話題が上がった際に，過去の会話を覚えておけば，会話相手の贔屓にしているチームなどの情報を思い出しつつ会話に組み込むことができる．しかし，会話を覚えておくことは人によって得意・不得意がある．また，得意な人でも，コミュニケーション量が多くなると覚えておくことが難しくなる．この課題を解決するために，過去の会話で言及した現在の会話内容に関連する情報を会話中に閲覧できるシステムを実現する．これにより，特に多くの人と会う機会がある人や相手との信頼関係が重要となる職種に対してコミュニケーションの支援を行うことを目的とする．

我々は，現在の発話文に基づいて，その後に続く発話文を予測し，過去の会話ログから必要な情報を先回りして抽出する手法を提案してきた [5]．これにより，発話文の少し後の文脈を予測し，過去の会話ログを検索することができるため，会話中のリアルタイムな検索を実現することが可能になる．

本論文では，新規のデータセットで，**大規模言語モデルを用いたクエリ拡張**による過去会話ログの検索手法の評価を行う．前の文脈を入力すると，以降に起こりやすいような単語列を予測可能なモデルである GPT-2 を利用し，将来の話題を予測して，時系列的に後に必要となるであろう会話を会話ログから抽出することで実現する．

既存研究では，一般的な文書検索において GPT-2 を用いたクエリ拡張は効果があるとされている．しかし，発話文自体を入力とするのではなく，ユーザが入力したクエリに基づいて文書を生成し，クエリ拡張に用いるものである．本研究では，発話途中の文書を入力にし，その次に発話されると予測される文書を生成しクエリ拡張に用いることで検索精度が向上すると考える．

提案手法の優位性を検証するために，現在の会話に関連する過去の会話を紐づけたデータセットが必要だが，それに適したデータセットが入手できなかったため，新たにデータセットを作成した．データセットの作成には，NTCIR-9 SpokenDoc テストコレクションの適合性判定結果を活用した．NTCIR-9 SpokenDoc テストコレクションは,『日本語話し言葉コーパス』を元に構築された講義音声を対象とした音声ドキュメント検索の評価に使用できるデータセットである．SpokenDoc テストコレクションの適合性判定結果から現在の発話文と関連する過去の発話文のペアを形成した．

実験では，このデータセットを用い，ベースライン手法と提案手法を比較し，提案手法の有効性を検証した．また，検索に用いるパラメータを変化させることで実験結果の評価をおこなった．実験を行った結果，提案したクエリ拡張手法により検索結果が統計的有意に改善することを示した．

本論文の貢献は以下のとおりである．:

（1） タスクの定義とそれを評価するための新規のデータセットの作成.

（2） 新規データセットで大規模言語モデルを用いたクエリ拡張手法の評価を行い，有意性を示した.

本論文の構成は以下の通りである．2 節では過去の会話検索とクエリ拡張の関連研究について述べる．3 節では問題設定と大規模言語モデルを用いたクエリ拡張手法，および，その主問題への適用方法について述べる．4 節では評価用データセットの作成方法や統計情報，実験設定やその結果について述べる．最後に，5 節で今後の課題と共に本論文の結論を述べる．

2 関連研究

本節では，過去の会話検索，クエリ拡張の関連研究について述べる．

2.1 過去の会話検索

Kunpeng らは，チャットツールの過去の会話に対し検索を行う「会話を検索する」という新しいタスクの提案と，その実験を容易にする研究システムの開発をおこなった [3]．与えられた会話から類似する過去の会話を検索するタスクは，会話のトピックが以前に議論されていることを示すことで，時間の浪費を防いだり，重要な決定を思い出したり追跡するのに役立つ．

これらの研究では，チャットツールでやりとりされた過去の会話を検索するタスクに取り組んでいる．本研究では，口頭での会話を対象にしているが，現在の会話に関連する過去の会話を抽出することで得られるメリットはテキストか口頭であるかには依存しないと考える．また，発話をテキスト化することで，既存研究の手法を適用することもできると考える．

しかし，テキストでの会話と口頭での会話の違いは，発話文をテキストに変換する際の発話文書き起こしツールの精度により，意図しない単語が入力されることもあり，ノイズがあるテキストに対する検索になることを考慮に入れる必要がある．

TREC の SDR (Spoken Document Retrieval) トラックでは，音声録音のアーカイブに対しアドホック検索を実現するタスクが提案された [2]．一般的な SDR は，自動音声認識と情報検索技術の組み合わせにより実現する．まず，音声ドキュメントを音声認識エンジンを利用し，タイムスタンプ付きの書き起こし文書に変換する．次に，書き起こし文書は検索システムによって索引付けされ，検索できる．クエリに対し返却される結果は，関連する書き起こし文を指すタイムスタンプのリストになっており，指している音声の内容とクエリ間の類似性スコアでランク付けされる．

既存のタスクと本研究で提案するタスクの違いについて述べる．既存のタスクには発話意図を自ら話すものや，検索クエリをユーザが直接入力するものがあるが，発話中に自動的に過去の発話文を検索するタスクは存在しない．

2.2 クエリ拡張

Vincent の研究は，テキスト生成モデルを使用して生成されたテキストをクエリ拡張に使用できることを示した [1]．関連研究では，前のトークンから次のトークンを予測するような教師なし学習でトレーニングされた Transformer である生成モデル GPT-2 [4] を使用し，元のクエリを生成モデルの入力とし，生成されたテキストをクエリ拡張に用いている．文書生成モデルを用いた文書の生成は，コストや時間がかかるため，できる限り少ない生成数で良い精度を達成したい．この研究では，クエリあたり 20 文書以上生成しても精度が変化せず，安定することを示した．また，重み付けによるクエリ拡張に比べ良い精度を得られることを示している．

既存の研究と本研究の違いについて述べる．既存の研究では，与えられたクエリごとに新しい文書を生成し，それを元に新しいクエリを作成する．本研究では，今まで発話された発話文を入力とし，まだ発話されていない部分を大規模言語モデルを用いて予測し，クエリを作成する．そのため，複数回にわたる会話の文脈を考慮した予測文を元にクエリを作成できると考えられる．

3 提案手法

本節では，現在の会話に関連する過去の会話を抽出するタスクの問題設定について述べる．次に，大規模言語モデルを用いたクエリ拡張手法を導入し，この主問題への適応方法について述べる．

3.1 問題設定

我々の以前の論文 [5] と同様に，本研究では，現在の発話内容に関連する過去の会話内容を，発話途中の発話文から検索するタスクを定義する．

N 個の発話文 u_l からなる集合 $U = \{u_1, u_2, \ldots, u_l, \ldots, u_N\}$ を考慮する．ここで，u_l は第 l 番目の発話文を指し，発話が終了するまでの区間を一つの発話単位とする．さらに，現在の発話区間 $D_c = [u_1, u_t]$ とそれに関連する過去の発話文集合 $D_s \in U$ を定義する．この時，u_t はある時点 t に入力として用いられる最新の発話文を表す．

会話支援のために過去の会話ログを提示するには，会話内容に基づいてリアルタイムで検索を行う必要がある．そこで，本研究では，現在の発話区間 D_c のできるだけ早い時点 t で検索を行い，関連する過去の発話区間 D_s の発話文を上位に含めるようなランキング D_r を返すことを目指す．

3.2 大規模言語モデルを用いたクエリ拡張

本節では，我々が以前の論文 [5] で提案した，大規模言語モデルを用いたクエリ拡張手法について説明する．

会話中に関連する会話をリアルタイムに検索するために，発話途中の発話文の続きを GPT-2 を用いて生成する．GPT-2 は，Transformer をベースにした事前学習済みモデルであり，文脈の理解や文章生成を行うことができる．

入力として，現在の発話区間 $D_c = [u_1, u_t]$ を GPT-2 に与え，その続きを生成することで未観測の発話文を予測する．この時，各入力に対して M 種類の文書を生成する．生成された各発話文を予測発話文 $\hat{D}_c$ とし，これをクエリ拡張に利用する．

(1) 現在の発話区間 D_c に対して分かち書きを行い，クエリに適した単語を抽出し，検索クエリ Q_m を作成する．(2) GPT-2 を用いて生成した予測発話文 $\hat{D}_c$ に対して分かち書きを行い，クエリに適した単語を抽出し，検索クエリ Q_m を作成する．これにより，未観測な発話内容を考慮した検索を実現する．(3) (1) で取得した D_c から得られる単語と (2) で取得した予測発話文 $\hat{D}_c$ から得られる単語を組み合わせ，検索クエリ Q_m を形成する．これにより，発話済みの文脈と，予測される未観測な発話文の両方を考慮した検索クエリ Q_m を作成する．

これらの手法を全ての予測発話文 $\hat{D}_c$ に対して適用し，入力

クエリ集合 $Q = \{Q_1, \ldots, Q_m, \ldots, Q_M\}$ を作成する．

次に，BM25 を用いた関連発話文の検索手法について説明する．上記の手法で生成された入力クエリ集合 Q を入力とし，発話文集合 U との関連度を BM25 を用いて算出する．式 1 に BM25 の式を示す．

$$\mathrm{score}(u_l, Q_m) = \sum_{i=1}^{n} \mathrm{idf}(q_i) \cdot \frac{f(q_i, u_l) \cdot (k_1 + 1)}{f(q_i, u_l) + k_1 \cdot (1 - b + b \cdot \frac{|u_l|}{\mathrm{avg}(dl)})} \quad (1)$$

ここで，$\mathrm{score}(u_l, Q)$ は発話文 u_l とある入力クエリ Q_m との間の BM25 のスコアを表し，発話文とクエリの間の関連性を数値化する．k_1 は 0 以上の実数，b は 0 から 1 の実数をとるパラメータである．$\mathrm{idf}(q_i)$ は特定の単語 q_i の IDF を示し，$f(q_i, u_l)$ は発話文 u_l 中の単語 q_i の出現頻度を表す．$\frac{|u_l|}{\mathrm{avg}(dl)}$ は発話文 u_l に含まれる単語数を文書集合内の平均文書長 $avg(dl)$ で割ったものであり，発話文の長さが平均より長いか短いかを示す．

次に，大規模言語モデルを用いたクエリ拡張時の文書の関連度スコアの式を式 2 に示す．入力クエリ集合 $Q = \{Q_1, \ldots, Q_m, \ldots, Q_M\}$ と，発話文 u_l を入力に与えると，u_l とそれぞれの検索クエリの間の BM25 スコアの合計が算出できる．これにより，複数のクエリによる検索結果の結合を行うことで，どのクエリでも抽出される過去の発話文に対するスコアを大きくする．

$$f(\mathcal{Q}, u_l) = \sum_{j=1}^{l} \mathrm{score}(Q_j, u_l) \quad (2)$$

関連度スコアの高い順上位 k 件を現在の会話に関連する過去の会話集合 D_r として提示する．この時の D_r が発話文予測を用いたクエリ拡張手法により得られた検索結果となる．

4 実　　験

本研究のタスクでは，現在の会話に関連する過去の会話を紐づけるデータセットが必要である．しかし，この要求を満たす適切なデータセットが存在しないため，新たにデータセットの作成を行う．本節では，まず，データセットの作成方法と統計情報について述べる．次に，ベースライン手法を含む実験設定について紹介し，最後に実験結果を示す．

4.1 データセットの作成

ここでは，データセットの作成方法と統計情報について述べる．

4.1.1 検索対象の文書コレクション

データセットの作成には，NTCIR-9 SpokenDoc テストコレクションの適合性判定結果を利用した．NTCIR-9 SpokenDoc テストコレクションは,『日本語話し言葉コーパス (CSJ)』を元に構築された講義音声を対象とした音声ドキュメント検索の評価に使用できるデータセットである．

表 1　自作データセットの統計情報

	自作データセット
全体の発話文数 U	884148
平均入力発話文長 $\lvert U \rvert$	16.21
入力発話文数 D_c	86
平均入力発話途中文長 $\lvert D_c \rvert$	101.49
入力発話文 (D_c) ごとの平均関連発話文数 D_s	1.52

CSJ は，日本語の自発音声データを，人手による書き起こしデータや話し言葉研究のための付加情報とともに収録したデータセットである．SpokenDoc コレクションでは，CSJ のうち学会講演 987 講演と模擬講演 1715 講演の計 2702 講演，約 604.3 時間の音声ドキュメントを検索対象としている．

CSJ では，発話文の転記基本単位を IPU(Inter-Pausal Unit) 要素と呼ぶ．IPU は原則として 0.2 秒以上のポーズで区切られる単位であり，文末（の可能性がある位置）では，0.05 秒以上のポーズでも分割する．

本研究のタスクでは，検索に用いる発話文の基本単位を IPU とし，これを発話文 u_l と表す．また，検索対象となる全ての発話文集合を $U = \{u_1, \ldots, u_l, \ldots, u_N\}$ とし，これは NTCIR-9 SpokenDoc コレクションの検索対象と一致する．

4.1.2 評価用データ

評価用のデータは，NTCIR-9 SpokenDoc テストコレクションの Spoken Document Retrieval（SDR）タスク用の適合性判定結果を活用して作成する．SDR タスクでは，特定の検索トピックに対する各発話文の適合性判定の結果が提供されている．

本研究のタスクは，現在の発話文 D_c に関連する過去の発話文集合 D_s を抽出することである．そのため，SDR タスクが提供する適合性判定結果を活用し，ある検索トピックに対して適合であると判定された発話文同士を関連発話文として定義する．関連発話文は一つの検索トピックに対して複数存在するため，関連発話文の中で最後に出現した発話文を現在の発話文 D_c とし，それ以前の発話文を D_s とする．

この方法により，現在の発話文 D_c とそれに関連する過去の発話文 D_s のペアを形成し，これを用いて検索手法の性能評価を行う．

4.2 データセットの統計情報

本実験では，86 個の現在の発話文集合 D_c と関連する過去の発話文集合 D_s のペアを得た．得られたデータセットは TREC 形式で保存した．表 1 に，作成したデータセットの統計量を示す．

4.3 実 験 設 定

我々は提案手法の有意性や特性を示すために，次の手法を用いて実験をおこなった．

(1) BM25 入力発話文

これは本実験のベースラインとなる手法である．現在までの発話文を入力とし，クエリを作成し，検索を行う．

表 2　各手法の精度．予測発話文数 $M = 20$．手法間で最も精度が高い結果を太文字で示す．

	nDCG@k		
手法	k=1	k=3	k=10
（1）BM25	0.011	0.044	0.042
（2）BM25（予測発話文）	**0.081**	**0.092**	**0.069**
（3）BM25（発話文+予測発話文）	0.012	0.012	0.019

(2) BM25 予測発話文
これは発話文までの入力をもとに，予測発話文の生成を行い，クエリを作成し検索を行う．
(3) BM25 入力発話文 + 予測発話文
これは発話文までの入力をもとに，予測発話文の生成を行い，クエリを作成し，元の発話文から生成されたクエリを拡張し，検索を行う．

(2) での予測発話文の生成には rinna 社が構築した日本語に特化した大規模言語モデルである GPT-2(japanese-gpt-1b) を用いた．パラメータ数は 13 億である．また，形態素解析には sudachipy を用いた．本研究では，一つの入力に対し，最短で 100 トークン，最長で 200 トークンのテキストが生成される設定としている．

構築したデータセットを用いた実験の手順を以下で述べる．現在までの発話文 D_c を入力し，それぞれのモデルでの検索を行う．D_c は，ある時点 t での発話途中の発話文 $u_{k_{[1:t]}}$ とその文脈にあたる過去の発話文 $(u_1, \ldots, u_l - 1)$ で構成されている．ここで，入力発話文 u_l の先頭から t 番目までの発話文が u_l を占める割合を r と定義した．これにより，できる限り早い時点での検索を実現するために必要な未完成な発話文を入力とした時の提案手法の評価を行う．実験でのタスクでは，現在の発話文 D_c とそれに関連する発話文 D_s を上位に抽出するようなランキング D_r を返す問題とした．入力に用いた発話文数は 86 個である．それぞれの入力に対して nDCG@k を計算し，平均 nDCG@k を各モデルの評価として用いた．

実験に用いたパラメータは以下の通りである．(1) 入力に用いる発話文の割合 r を $r = \{25\%, 50\%, 100\%\}$ とした．(2) 拡張時に生成する予測発話文 $\hat{D_c} = \{\hat{d_{c_1}}, \ldots, \hat{d_{c_M}}\}$ の種類を M 個とした．予測発話文の種類が増えるほど，ユーザの情報意図にあったクエリが生成される可能性が高いと考えられるため，さまざまな数値を利用してモデルの精度を検証する．

各モデルの性能評価は nDCG@k（$k = 1, 3, 10, 50, 100$）で行う．

4.4 実験結果

まず，表 2 に，本研究で作成したデータセットにおける各モデルの精度を示す．比較のため，クエリ拡張を行わないベースラインの BM25(1)，クエリ拡張ありの BM25(2)，(3) を用いた．クエリ拡張時の予測発話文数 M は 20 個とした．

大規模言語モデルを用いたクエリ拡張による BM25 は，予測発話文のみを用いた BM25(2) で，全ての評価基準でベースラインを上回った．nDCG@3 に基づくと，ベースラインの BM25(1) と比較して，総合的な精度は 2.09 倍に向上した．しかし，発話文と予測発話文の両方を用いてクエリ拡張を行った場合，検索性能は著しく低下している．これは，現在までの発話文から作成したクエリが予測発話文数 M の数だけ含まれ，BM25 での検索語のスコア結合時に高いスコアになってしまうことが理由として考えられる．

表 3　各手法の nDCG@3.

	手法		
生成文書数	（1）BM25	（2）BM25（予測発話文）	（3）BM25（発話文+予測発話文）
1	0.044	0.025	0.011
5	0.044	0.072	0.012
10	0.044	0.076	0.015
15	0.044	0.105	0.012
20	0.044	0.092	0.012
25	0.044	0.098	0.024
30	0.044	0.093	0.020

次に，発話予測文の生成には計算コストと時間がかかる可能性があるため，クエリ拡張に適切な予測発話文数を確認する．そのため，予測発話文数のパラメータを変更し，どれだけ精度が変化するかを評価した．その結果を表 3 に示す．提案手法 (2) は予測発話文数 M が 5 以上のときにベースライン手法を上回る精度を達成していることが明らかになった．しかし，予測発話文数が 1 つの場合，精度がベースラインを下回っている．これは，GPT-2 による生成が確率論的に行われるため，予測発話文にばらつきが生じ，検索クエリとして適切な発話内容が生成されなかったと考えられる．そのため，GPT-2 の予測発話文数を増やし，より多様な予測発話文を用いてクエリ拡張を行うことで検索の精度を高める必要がある．

最後に，図 1 に示されているように，予測発話文数を増やしても，nDCG@k の精度はある程度で頭打ちになる．これは，予測発話文を生成する過程で，不適合な発話文を高く評価してしまうクエリが一定程度生成されることが原因であると推測される．

例えば,「この時のホトトギスは一羽で飛んでったと/で渋民村のカッコウはつがいで飛んでったと/いうところに (F あのー) 対比すると/自分なり私なりに面白さを感じておりました」という入力発話文があった時，関連する発話文集合は「野鳥にはどういうものがいるか」というトピックに関連する発話文である．そのため,「カラス」や「スズメ」などの他の野鳥の名前が含まれたクエリを生成し検索を行うことが理想である．しかし，入力発話文に「ホトトギス」や「カッコウ」といった単語が含まれているため，GPT-2 がこれらの単語を用いて次に出現する確率が高い単語を予測する傾向がある．その結果，入力発話文には含まれていない他の野鳥に関連するクエリの生成が困難になることがある．

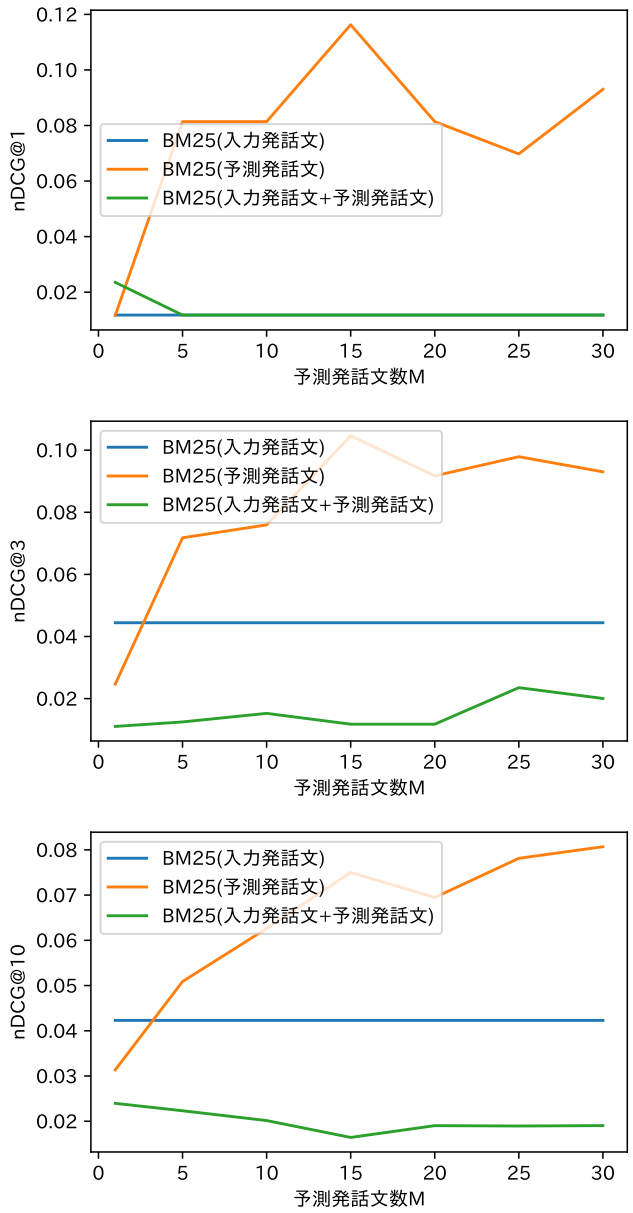


図 1　生成文書数を変化させたときの nDCG@k($k = 1, 3, 5$).

さらに，入力発話文に含まれる「ホトトギス」や「カッコウ」がクエリに含まれる確率が上がってしまい，検索対象の発話文のスコアが不当に高くなることがある．この問題を解決するには，生成したクエリを考慮して BM25 のスコアを正規化するか，BM25 以外の検索モデルの利用を検討する必要がある．これにより，入力発話文に含まれる単語の影響を減らし，より多様な関連発話文を効果的に検索できる可能性がある．

5　ま　と　め

本論文では，現在の会話内容に関連する過去の発話内容を抽出し，円滑な会話を実現する問題に取り組んだ．この問題では，ユーザが掲示情報を理解し，会話に利用するために時間がかかってしまうため，明示的にトピックへの言及がなくとも関連会話データを抽出する必要があった．そこで，我々は現在の発話内容に関連する過去の会話内容を，発話途中の発話文から検索するタスクを定義し，大規模言語モデルを用いたクエリ拡張による過去会話ログの検索手法を評価した．

大規模言語モデルを用いたクエリ拡張は，現時点までの発話内容を入力とし，GPT-2 を用いて生成した予測発話文を基に検索クエリを作成ことで実現した．

提案手法の有意性を検証するために，現在の会話に関連する過去の会話を紐づけたデータセットが必要だったが，それに適したデータセットが入手できなかったため，新たにデータセットを作成した．

実験では，このデータセットを用いてベースライン手法と提案手法を比較した．予測発話文数を変化させた場合の比較実験を通じて，大規模言語モデルを用いたクエリ拡張が，予測発話文数が 5 以上の場合にベースラインを上回る精度を達成することが明らかになった．しかし，予測発話文数が 1 つの場合，精度が低下することが観察された．これは，生成された予測発話文にばらつきが生じ，検索クエリとして適切な発話内容が生成されなかったためである．また，予測発話文数を増やしても，ある程度で nDCG@k の精度が頭打ちになることが見られた．これは，予測発話文生成過程で，不適切な発話文を評価するクエリが生成される可能性があるためである．今後の課題としては，生成したクエリを考慮した正規化や，BM25 以外の検索モデルの検討などが挙げられる．

謝辞　本研究は JSPS 科研費 JP21H03554，JP21H03775，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Vincent Claveau. Query expansion with artificially generated texts. *arXiv preprint arXiv:2012.08787*, 2020.
[2] John S Garofolo, Cedric GP Auzanne, Ellen M Voorhees, et al. The trec spoken document retrieval track: A success story. *NIST SPECIAL PUBLICATION SP*, 500(246):107–130, 2000.
[3] Kunpeng Qin, Harrisen Scells, and Guido Zuccon. Pecan: A platform for searching chat conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2610–2614, 2021.
[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
[5] 田貝奈央 and 加藤誠. 発話内容を用いたクエリ拡張による過去会話ログの検索. In **第 *15* 回データ工学と情報マネジメントに関するフォーラム**, 2023.
[6] 野口敏. **誰とでも *15* 分以上 会話が途切れない！ 話し方 そのまま話せる！ お手本ルール** *50*. 株式会社すばる社, 東京, 2013.

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 10:00 ~ 12:10 | T4-B オンライン（Zoom Events)

映像の理解と可視化

座長:風間 一洋(和歌山大学)
コメンテータ:中島 伸介(京都産業大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T4-B-4-01] 複数人の競技かるた選手の構え・払いの動作の比較のための三次元ボーン表示および箱棒グラフによる可視化

*北川 リサ[1]、伊藤 貴之[1] (1. お茶の水女子大学伊藤研究室)

10:30 ~ 10:55
[T4-B-4-02] キンボールのプレイ映像解析のための線形フレーム補間による物体追跡

*山下 翔五[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学　工学部　電子システム工学科、2. 滋賀県立大学　先端工学研究院)

10:55 ~ 11:20
[T4-B-4-03] 離脱傾向を踏まえたインターネット動画広告データの可視化システム

*三浦 梨花[1]、大矢 隼士[2]、伊藤 貴之[1] (1. お茶の水女子大学　伊藤研究室、2. Septeni Japan株式会社)

11:20 ~ 11:45
[T4-B-4-04] CLIPを用いたゲーム動画フレーム毎のバグ検知AI

*髙橋 秀太朗[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学　知能情報工学分野、2. 滋賀県立大学)

11:45 ~ 12:10
[T4-B-4-05] 数枚のサンプル画像を用いて類似場面を検出するオンデマンド場面検出技術

秋島 遥[2]、*吉田 登[1]、馬場崎 康敬[1]、劉 健全[1] (1. NEC、2. お茶の水女子大学)

複数人の競技かるた選手の構え・払いの動作の比較のための三次元ボーン表示および箱棒グラフによる可視化

北川　リサ†　　伊藤　貴之†

† お茶の水女子大学大学院人間文化創成科学研究科　〒 112–8610 東京都文京区大塚 2 丁目 1-1
E-mail: †{g1820512,itot}@is.ocha.ac.jp

あらまし　競技かるたの試合でのパフォーマンスを向上させるには，選手それぞれの体格やプレースタイルに合わせた構えや払い方を模索することが求められる．そのため，他者との比較によって最適な構えを模索することが多いが，撮影した動画などを使用して詳細な違いまで比較することは困難である．競技かるたに関する既存研究として，競技中の選手の脳の動きを測定したり，手首に加速度センサーをつけて払いの速度を分析する手法は用いられていたが，これらの手法では体に器具を装着して計測するため，自然な状態での計測ができない．そこで我々は， Google が開発した MediaPipe を使用して動画から抽出した骨格情報を用いて，読まれた札に対しての全身の動き方を分析するために三次元ボーンならびに箱棒グラフを表示し，複数の選手の動作を比較できる可視化システムを開発している．本報告では，提案システムを用いて上級者 7 名が一人取りを行っている様子を比較した結果を示す．

キーワード　可視化, 映像解析, 行動解析, 競技かるた, UI・UX, スポーツデータ活用

1　はじめに

競技かるたとは，小倉百人一首かるたの 100 枚のうち無作為に選び取った 50 枚を用いて一対一で対戦する競技である．使用する札のうち 25 枚ずつを互いの陣地に並べ，1 枚ずつ取り合い，自陣の札を 0 枚にした方が勝利となる．

競技かるたの実戦で勝ち上がるには，場に配置されている 50 枚の札を相手よりも速く取る必要があるが，そのためには選手自身の取り方の特徴を分析することが不可欠である．陣地の中でどこが得意（または苦手）なのか，また構え方や払い方に修正の必要はないか，さまざまな観点から分析をすることで選手は上達し，より多くの札を速く取れるようになることが期待できる．

多くの選手は，対戦相手や指導者から主観的なアドバイスを交換し合うことで，自身の構えや払いの特徴や改善点を分析しているが，この方法には 3 つの問題点がある．

1 つ目として，人の目による観察だけでは正確な知見が得られないことがあげられる．競技かるたに限らず，個人の感覚にはさまざまなバイアスが生じやすい．上級者であっても，他者の動き方を見るだけで詳細まで正確に認識できる選手は少ないと考えられる．そのため，個人の観察のみで選手の状態を判断するより，定量的な計測結果も加味する方が正確な知見を得られると考えられる．

2 つ目に，自身と他者との比較の難しさがあげられる．他者の動作は観察者が移動することで任意の角度や距離から見られるが，自身の動作を俯瞰して見る機会はほとんどない．札を払っている他者と自身を比べるためには，自身の払いの動画を撮影し，二次元の映像情報と見比べるしかない．図 1 からもわかるように，他者と比較したときに体のどの部位がどの程度違う動作をしているのか，詳細な違いを見つけることは難しい．

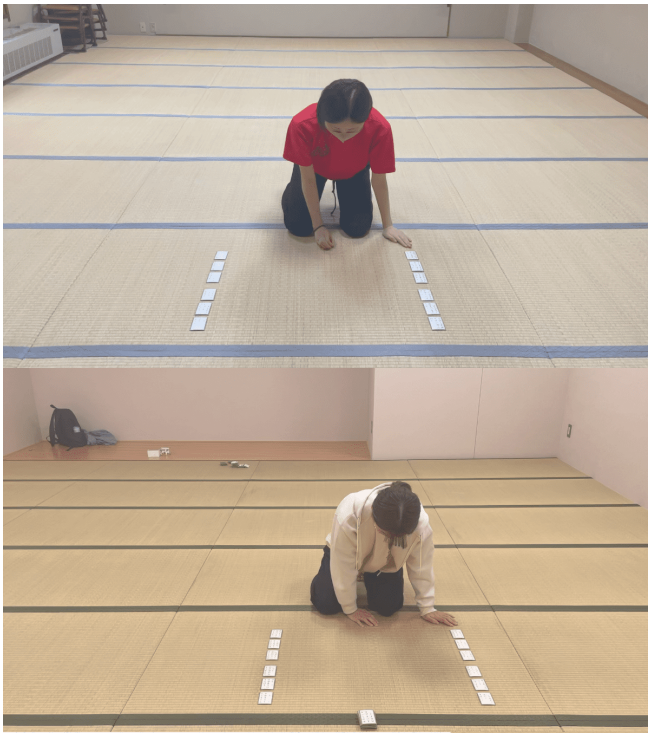


図 1: 動画や写真を用いた他者との比較

3 つ目に，自身と似た体格の選手との比較が難しい点があげられる．各選手にとって最適とされる構えや払いを見つけるためには，それぞれの選手と似た体格の上級選手と比較することが有効な手段であると考えられる．他の競技であれば性別や体格ごとに階級分けがあるため，自身と似た体格の選手を見つけることは容易であると予想される．その一方で，競技かるたは実力による階級分けしか存在しないために自身と似た体格の選手を見つけること自体が難しい．

そこで我々は，競技かるたにて札を取る動作を測定しそれを

可視化することで，他者と自身の動作を比較できるシステムを開発している．本研究の目的は，2 人以上の選手の動作を比較し詳細な違いを分析するため，競技かるたの動作の可視化において最適な可視化手法を提案することである．

本研究の初期段階では，初心者に対して最適な構えを提案するシステムを構築することも想定していた．しかし，研究を進めていく中で，どの構えが最適であるかを判断することは困難であると考えた．そこで本研究では，複数の選手の動作を比較できるシステムを構築した上で，構えや払いの改善点となり得る箇所をユーザ自身で発見し，改善を重ねながら上達するという流れを想定している．

この提案手法では，選手が一人取りと呼ばれる形式で試合をしている様子を撮影し，選手の各部位の動作情報を抽出する．この動作情報を，体格ごとの特徴量による散布図，三次元ボーン表示による選手の動作のアニメーション，および動作タイミングを表す箱棒グラフの 3 つの機能を有するインタラクティブな可視化システムで表示する．この 3 つの可視化結果を比較しながら分析することで，体格の違う複数人の選手の差異を評価できる．

本報告では上級者 7 名の払いの動作を比較し，どのような差が見られるか検討した結果を紹介する．

2 関 連 研 究

2.1 競技かるたに関する研究

本節では測定技術を駆使して競技かるたの上達を支援する研究事例を紹介する．

丸山ら [1] は，VR 空間上で百人一首かるたを遊べるシステムを構築した．片方のプレイヤーから見える札を大きくするなど，プレイヤーには知覚されない形でシステム上でスキルを調整することにより，スキル差のあるプレイヤー同士が対等に遊ぶことを可能にしている．

武田ら [2] は，光脳機能イメージング装置を用いて競技かるた選手の試合中の脳の変化を測定している．その結果，下の句の読み開始から上の句の読み終わりまでと，場にある札が読まれた際に酸素化ヘモグロビン濃度値が上昇していることが明らかになった．これにより，選手が札を取る際の高度な聴覚情報の受容，認知，処理，俊敏な運動を繰り返す脳の情報処理サイクルを観測している．

以上の 2 つの手法は競技かるたをテーマに情報処理技術を駆使してシステムを構築しているという点では本手法と類似しているが，目的や提案手法に違いが見られる．

山田ら [3] は，手首に加速度センサと角速度センサを装着することで競技かるたの札の取得タイミングを計測した．この手法では，あらかじめ採取した払いの加速度・角速度の情報とそのときの札の取得タイミングを正解データとしておく．続いて，実際に採取したテストデータと正解データの DTW (Dynamic Time Warping) 距離を計算する．そして DTW 距離が最短となるような時刻を算出し，そこがテストデータ上の札の取得時刻であると判定している．実際の競技で選手同士の間に生じる取得タイミングの差は数十ミリ秒程度であるが，この手法では 99.0%のものが 20 ミリ秒以内の誤差で推定することができ，最大誤差は 60 ミリ秒程度となった．

この手法は人体の動きを測定し競技かるたをスポーツとして分析しているという点では本手法と類似しているが，札の取得タイミングを測定することが目的のため，競技中の全身の動作を可視化することを目的とした本手法とは異なる．

2.2 骨格抽出に関する研究

人体の骨格情報を抽出する技術は，スポーツ分析をはじめとする多様な目的に既に活用されている．特に，人の動作を撮影し三次元の位置情報を取得するモーションキャプチャを活用した研究事例が多く報告されている．

2.2.1 マーカやセンサを使用したモーションキャプチャ

モーションキャプチャの代表的な技術の 1 つとして，人体にマーカやセンサを装着して関節の位置をトラッキングする技術があげられる．この技術は精度が高いが，機器の購入のための金銭的負担や実験に必要な空間の確保などのコストがかかることから，自宅等での気軽な導入が難しいため，本研究では適用していない．

磁気式モーションキャプチャを適用した例として，田中ら [4] はダンスの上達支援システムに用いている．この手法では上級者と初心者それぞれのダンスを可視化することで，自らのダンスを客観的に見ることができる．ダンサーの顔の向きに着目してグラフ表示させることにより，上級者と初心者の違いを明確に表している．

競技かるたはダンスに比べると全身の動きが小さく，また床に座っての競技であることから，ダンスおよび他のスポーツと比べてモーションキャプチャでの測定の難易度が高いが，それにしてもグラフなどを用いた可視化による動作の提示は有効であると考えられる．

2.2.2 カメラを使用したモーションキャプチャ

赤外線カメラと機械学習を用いて同画像から人物動作を抽出するモーションキャプチャ技術も近年普及している．Microsoft 社の Azure Kinect[1]がその代表的なシステムである．Azure Kinect はマーカやセンサを使用せずともトラッキングができる上に，位置情報の精度も比較的高く，安価で導入しやすいというメリットがある．

川西ら [5] は，Azure Kinect を用いて一人のダンサーの反復練習のモーションデータを取得し，体の部位ごとにクラスタリングを適用した結果を可視化している．また Melios ら [6] も同様に，Azure Kinect を用いてダンスの動作をモーションキャプチャし，Unity を用いてダンスの練習をしながらゲームが楽しめるシステムを制作している．

2.3 機械学習による動画像からの姿勢推定

マーカやセンサ，赤外線カメラ，といった専用の機器を使用せずに，一般的な二次元の動画像から機械学習によって人の関節の位置を推定する姿勢推定技術も近年発達している．この技

1：Azure Kinect https://azure.microsoft.com/

術は機器の導入コストがかからないことに加え，複数の人物に対しても姿勢推定が可能であるということ，そしてモーションキャプチャを目的とせずに録画した動画像にも適用できるというメリットがある．複数の人物に対して同時に姿勢推定できるという利便性から，サッカーやバスケットボールなどのチームスポーツをはじめ，多様な分野での活用が期待されている．Bridgeman ら [7] はサッカーの試合を撮影した 2 つの動画像から，高速な貪欲法を適用させ，複数の選手の全身の姿勢推定と追跡の精度を大幅に向上している．また Ghasemzadeh ら [8] は，バスケットボールにおいて選手どうしが重なって写っている場面での骨格検出，およびボールの位置情報の取得の手法を提案している．本手法では動画像の機械学習自体を研究対象としているわけではないが，競技かるたに関する姿勢推定の精度を向上させる際には動画像からの動作抽出手法を模索したい．

Long ら [9] は，MediaPipe を用いて対話型ヨガ姿勢指導システムを開発した．この研究では，あらかじめ正解となる 14 のヨガポーズを学習しておき，被験者のポーズと比較して修正点を指摘するという手法をとっている．本手法と同じ MediaPipe を用いたコーチングシステムであるが，競技の種類に違いがある．また，競技かるたの場合は体格に合わせた構えを取る必要があり，その種類は多様であるため，正解となる姿勢を定義することが難しい．しかし，Chhaihuoy らの手法と同様にして，似た体格の上級者と比較することで修正点を指摘する，という手法を適用できる．

3　可視化システムの提案

本章では，本報告で提案する可視化システムについて説明する．本研究で開発したシステムは以下の 2 つの工程で構成されている．

（1） 払いデータの抽出

（2） 払いデータの可視化

(1) ではまず競技かるたの払いの動作を撮影し，骨格情報を抽出する．以下，このときに得られる体の部位ごとの位置情報を「払いデータ」と称する．(2) では (1) で得られた払いデータを可視化することで，選手の払いの動作を観察する．

先行研究である北川ら [10] の手法では 1 音目が読まれた瞬間に指定された箇所を払う払い練の様子を測定していた．それに対して本手法では，試合形式での測定が可能になった．これにより，動き出しのタイミングや決まり字による動作の違いについても分析ができるようになった．また本手法では，新たに箱棒グラフを追加し，試合形式の情報を可視化するための最適な手法を検討した．

以下，3.1 節で払いデータの取得方法，3.2 節で払いデータを抽出するシステムの詳細，3.3 節で可視化システムについて説明する．

3.1　払いデータの取得方法

本節では競技かるたの払いの動作の測定手法について説明する．本研究では，専用の機器を使用せずとも動画像さえ撮影できれば骨格を抽出できることから，多くの選手の情報を収集できるメリットがあると考え，動画像からの機械学習による姿勢推定を用いることにした．本研究では Azure Kinect によるモーションキャプチャも試みたが，払いの動作における測定結果の精度は差が見られなかったため，動画像からの姿勢推定を選択した．

また，当初は全身の骨格抽出を目標としていたが，動画像からの姿勢推定も，Azure Kinect による測定も，いずれも下半身の検出の精度が低かった．これは競技かるたの正座のような構えの性質上，カメラから見て下半身が隠れてしまうため，精度が下がることは明らかである．払いの動作の観察は下半身よりも上半身や腕，手の動きが中心になってくるため，主に上半身を分析する方針で研究を進めることにした．

本研究では Google が開発した MediaPipe を用いて姿勢を推定した．MediaPipe は他の姿勢推定モデルと比べて軽量であることに加えて，実行環境を問わず Web ブラウザ上で実行できる．将来的にはより多くの選手の情報を集めたいと考えていることから，MediaPipe を用いることにした．なお，MediaPipe 以外の姿勢推定モデルも試したが，使用するコンピュータの性能によっては動作しなかったり動きが滑らかでない場合があったため，本研究では採用を見送った．

MediaPipe では 1 つのカメラの入力映像から x,y,z の三次元座標を推定できる．この推定結果を用いて三次元でのボーン表示による可視化を実行することで，任意の角度や距離から自身の構えや払いの観察が可能になった．

3.2　払いデータの取得システム

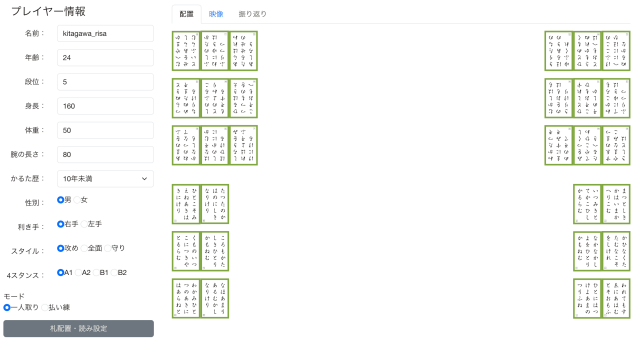


図 2: 払いデータの取得システム

払いデータを取得するために，JavaScript を用いて Web システムを構築した．このシステムの画面は図 2 に示す通り，

- コントロールパネル（左）
- メインパネル（右）

の 2 つで構成されている．

3.2.1　コントロールパネル

コントロールパネルは，選手の情報を入力するフォーム，札配置・読み設定パネル，暗記時間と読みのコントロールパネルの 3 つで構成されている．

選手の情報を入力するフォームでは，これから測定する選手の [名前，年齢，段位，身長，体重，腕の長さ，かるた歴，性別，利き手，プレースタイル，4 スタンス理論 [2]におけるタイプ] の 11 項目について入力する．これらの項目は，のちに選手の体格によって分類する際に使用される．

フォームの下には試合モードを選択するラジオボタンがあり，一人取りモードと払い練モードの 2 種類のモードを切り替えられる．これを切り替えることで，選択されたモードに合わせてメインパネルに表示される札配置や読まれる和歌が設定される．

暗記時間と読みのコントロールパネルには，試合開始前には残りの暗記時間を表示するタイマーと暗記開始のボタンが設置されており，暗記開始ボタンを押すことで暗記時間が開始される．通常の試合では 15 分間の暗記時間が設けられるが，上級者の選手の中には 15 分間の暗記時間も必要としない選手もいるため，暗記時間の長さも測定対象とし，15 分が経過する前に試合開始ボタンを押すことで暗記時間を終了して試合を開始する．試合開始後には暗記時間のタイマーとボタンが消え，代わりに読みの再生ボタンが表示される．試合開始時に再生ボタンを押すと，一枚読まれるごとに適当な間隔を空けながら試合が終了するまで読まれ続ける．

3.2.2 メインパネル

メインパネルは配置，映像，振り返りの 3 つのタブにより機能を切り替えることができる．

システムを開いた直後は配置タブが表示されており，測定に用いる札の配置が画像で表示されている．この配置は前述した通りデフォルト以外の設定を読み込むことも可能で，読み込まれた json ファイルに従って配置画像も切り替わる．測定では，画面に表示されている配置の通りに札を並べてから暗記時間を開始する．

暗記時間が終了したら試合を開始し，最後の 1 枚になるまでデータの計測を続ける．

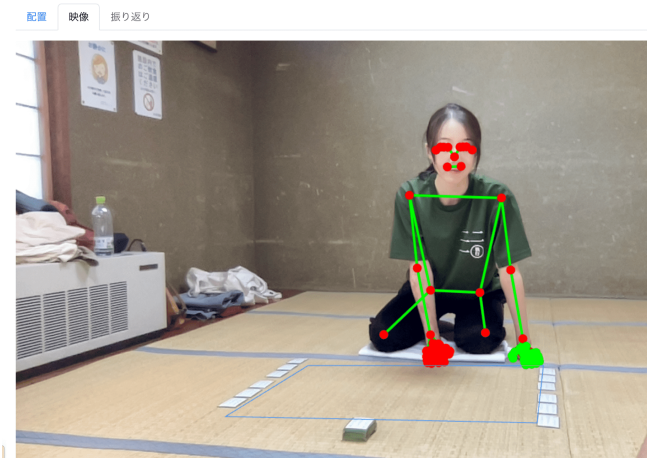


図 3: 映像タブ

試合開始後は映像タブに切り替え，抽出された骨格情報を確認できる（図 3）．画面には Web カメラで取得した動画に合わせ，リアルタイムでボーン情報を表示する．読みが再生されると，読みのタイミングに合わせて自動でボーン情報が記録される．ここでは実際の試合と同様に，序歌から読み始める．競技かるたの規定に則り，前の歌の下の句を読み終えてから 1 秒間の間の後に次の歌の上の句を読み上げるが，上の句が読まれる 1 秒前から身体の位置情報を記録し始め，その 3 秒後には記録を停止している．これにより，読まれる直前の選手の動作も測定でき，そして記録時間を最小限に抑えてデータ量の増加を抑えた．なお，MediaPipe を用いることで，全身の 32ヶ所と両手の指のそれぞれ 20ヶ所を合わせた 72ヶ所の骨格が毎秒 20〜30 回推定される．

図 4: 振り返りタブ

場の札が 1 枚になると試合は終了となり，振り返りタブに切り替える．振り返りタブの画面構成を図 4 に表す．振り返りタブでは読まれた順に札の一覧が表示される．それぞれの札に対してどのような動きをしたか選手自身で振り返り，入力する．対象となる情報は初動の速さ，触るまでの速さ/巧みさ，コメント，決まり字判定ができていたかのチェックボックス，お手つきだったかどうかのチェックボックスの 5 項目である．

競技かるたでは試合の中でコンディションの変動があることが多く，選手自身が 1 枚 1 枚の取りに対してどのような印象を抱いていたかの指標も併せて分析することでより有意義な情報が得られると考えたため，振り返りのデータを入力できるようにしている．

振り返りの入力を終えたあと，左側のコントロールパネル下部にある保存ボタンを押すことで，選手情報を含むすべての払いデータを json 形式で書き出し，ローカルファイルとして保存する．

3.3 可視化システム

可視化システムの画面を図 5 に示す．システムは大きく以下の 3 つの画面で構成されている．

- コントロールパネル（左）
- ボーン表示（右上）

2：廣戸 [11] によって提唱された，ヒトの身体特性を 4 種類に分類するための理論．

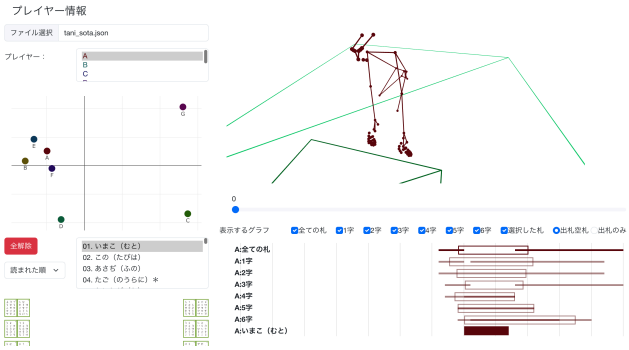


図 5: 可視化システム

- 箱棒グラフ表示（右下）

本システムでの可視化では，まずコントロールパネルの上部に設置されているボタンから可視化したい選手全員の json ファイルを読み込む．ファイルを読み込むと読み込んだ選手の名前が表示され，その下部にある可視化ボタンを押すことでボーンとグラフが表示される．コントロールパネルには表示する選手を選択するためのセレクトボックス，選手同士の体格の類似度を示す散布図，表示する札を選択するためのセレクトボックスの 3 つで構成されている．

選手を選択するためのセレクトボックスでは，1 人以上の選手を選択できる．複数人を選択することで，重ねて比較が可能である．

その下部に，どの選手同士を比較するべきかの指標の 1 つとして，体格を比較するための散布図を設置した．この散布図は，計測システムにて入力した「4 スタンス理論におけるタイプ」「体重」「身長」の 3 つの特徴を，主成分分析（PCA）により次元削減し作成した．本来であれば性別，段位についても類似度を計算するべきところだが，性別においては身長とほぼ同じ相関が見られること，そして段位は全員がほぼ同じであることから，今回は 3 つの特徴に絞って特徴量を抽出した．近くに分布している選手同士が似た特徴を持っていることを表しており，どの選手と比較するべきかを検討するための指標になる．

この散布図の下には，表示する札を選択するためのセレクトボックスと札の配置図が設置されている．セレクトボックスは「読み上げられた順」「決まり字のあいうえお順」「決まり字の長さ順」の 3 パターンで表示を切り替えられる．これにより，似た決まり字の札を払った場合を比較しやすくしている．札の配置図では，現在選択している札を赤枠で表示することで，どこにあった札なのかを可視化している．直接配置図内の札をクリックすることでも切り替えられるため，近くに配置されていた札を複数選択することで，似た位置にあった札を払う際の動きを比較できる．

コントロールパネルの最下部にはボーン表示に関する詳細設定パネルがあり，任意の選手のボーンを上下左右に線形移動，任意の選手のボーンを拡大縮小，任意の選手のボーンを左利き用に左右反転，任意の選手のボーンの詳細設定をリセット，カメラ（視点）の位置のリセットの 5 つの機能が付与されている．

選手によって構える位置が違うため座標が左右にずれていることが多く，体格差によっても比較しづらいことがあるため，本手法では描画するボーンの位置を各選手ごとに微調整できるようにしている．ただし，ボーンを拡大縮小することによって払う札の位置も変わってしまうため，本報告では変形させずに比較している．また，左利きの選手はデータ測定の際に札の配置も反転させている．そのため，ボーンを反転表示させることで，右利きの選手と同じ位置を払っている様子を観察できる．

画面右上のボーン表示は，選択した選手や札に対応したデータを描画している．ボーンの描画には p5.js を用いることで三次元の座標データを可視化し，任意の角度からの観察を可能にした．ボーンをドラッグすることで視点を動かせるため，スライダーと合わせて操作することで一瞬の動作を自由な角度から観察できる．

各選手を識別しやすいように，ボーン表示，グラフ表示ともに，各選手に 1 色ずつ対応させて描画する．

画面右下には動作にかかった時間を可視化するためのグラフを設置してある．グラフには 2 種類あり，1 つは選択した札の動き始めと動き終わりを示す棒グラフ，もう 1 つは複数の札の動作タイミングを可視化するために我々が考案した箱棒グラフである．

利き手の人差し指の座標が一定数以上動いたときを「動き始めのタイミング」とし，利き手の人差し指の座標が始めの座標から一定数以上離れた時を「動き終わりのタイミング」として算出した．読まれた札 1 枚ずつに対して算出したあと，全ての札，決まり字ごとにそれぞれのタイミングの最小値，最大値，平均値を算出する．

箱棒グラフの見方については，4.3 章で詳しく説明する．

4 実 行 結 果

今回の測定では読まれた際の決まり字のパターンが多くなるように読み上げ順を設定した 64 枚の札を読み，その読み札に含まれる 30 枚の札を場に配置した．つまり，合計で 64 回分の払いのデータを 7 人分測定している．実際の計測では，1 試合につき 40 分程度の時間を要した．

複数の選手を比較する際に着目すべきいくつかの観点のうち，本報告で考案したシステムを用いて分析できる以下の 4 つの点についての考察結果を示す．

4.1 動き始めるまでの間の動き

まず，ボーン表示をコマ送りで観察することで得られた考察について言及する．

ボーン表示に付属しているスライダーにおいて，0 コマ目から 20 コマ目あたりまでは読まれる直前の「間（マ）」の 1 秒間に当たる．その 0 コマ目から 20 コマ目の間でスライダーを動かしながらボーン表示を観察することで，動き始めるまでの間の最中に選手がどのように動いているかを考察した．この章で示す図は，それぞれの選手の動きが最もわかりやすかった 3 コマ目，16 コマ目，24 コマ目での姿勢である．

7 人の選手を全員同時に描画して「間」での動きを比較した

ところ，“ほとんど動かないタイプ”，“頭が前後に動くタイプ”，“頭が左右に動くタイプ” の 3 つのタイプがあることがわかった．

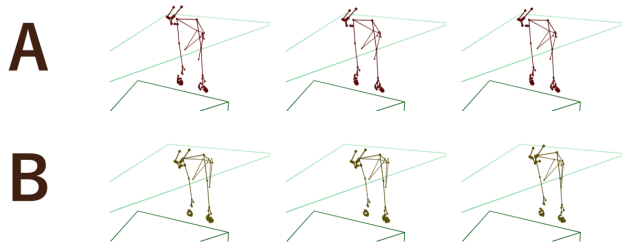


図 6: ほとんど動かないタイプ（左から 3 コマ目，16 コマ目，24 コマ目）

まず 1 つ目の “ほとんど動かないタイプ” にあたる選手は，A,B の 2 名である (図 6)．この 2 名の選手は読まれるまでの「間」の最中は構えた状態からほとんど動かない．このタイプに該当する選手は読まれる直前はむやみに体を動かさず，音を聞くことに集中していると考えられる．

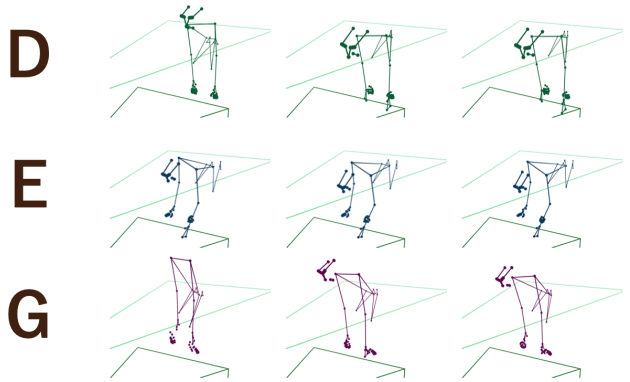


図 7: 頭が前後に動くタイプ（左から 3 コマ目，16 コマ目，24 コマ目）

2 つ目の “頭が前後に動くタイプ” にあたる選手は，D,E,G の 3 名である (図 7)．この 3 名の選手は，読まれる直前に頭をグッと前に押し出すような動きをしている．この動きは競技かるた選手の多くに見られる動作であり，手を動かす前準備として集中力を高めるための動作であると考えられる．このタイプに該当する選手は動きの大きさの違いはあれど，ほとんど全ての札に対してルーティーンのように頭を動かしていることが確認された．

3 つ目の “頭が左右に動くタイプ” にあたる選手は，C,F の 2 名である (図 8)．この 2 名の選手は，読まれる直前に頭を左右に振って競技陣を見渡すような動きをしている．競技かるたでは配置されている札の場所を暗記する必要があり，選手は札が読まれるまでの少しの時間でも札を覚える作業をしていることが多い．そのため，このタイプに該当する選手は「間」の少しの時間でも札を覚えているか，もしくは札を目視することで数枚の札を瞬間的に認識しているのだろうと考えられる．C と F で比べると C の方が左右の動きが大きく，ほとんど全ての札

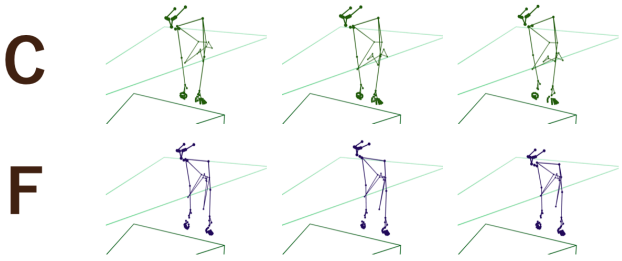


図 8: 頭が左右に動くタイプ（左から 3 コマ目，16 コマ目，24 コマ目）

に対してルーティーンのように頭を動かしていることが確認された．F は C よりも動きが小さく，ほとんど動かない札もあったことから，ルーティーンというよりは偶発的に動いてしまったと予想される．

“頭が前後に動くタイプ” は動く幅が大きいことから,「間」に動いていることを自身も他者も認識していることが多いが，今回の “頭が左右に動くタイプ” は動く幅が小さいため他者からは気付かれないことが多い．本研究で構築したシステムでは，視線の動きの分析はサポートしていないが，視線計測装置を併用することで新たな発見が期待される．

本システムで可視化することにより，動き始めるまでの間の動きについて選手を 3 つのタイプに分類したが，後述するように同じタイプの選手間でも動き出しのタイミングに差があったため，競技時にどのタイプが有利かは明らかではない．

4.2 動き始めのタイミング

次に，ボーン表示と箱棒グラフを併用して比較した結果について示す．

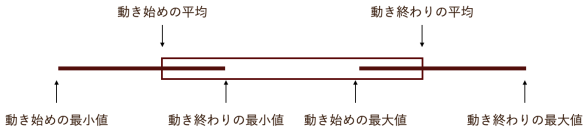


図 9: 箱棒グラフの見方

図 9 で示す通り，箱棒グラフの左側の棒は動き始めと動き終わりの最小値，右側の棒は動き始めと動き終わりの最大値を示す．箱は動き始めの平均値から動き終わりの平均値を示している．以下では，それぞれを棒，箱と呼ぶ.「全ての札」と 6 種類の決まり字によって分けられた 7 つの札グループに対して，それぞれの平均値・最大最小値を求め，箱棒グラフを用いて比較した．

4.2.1 全ての札

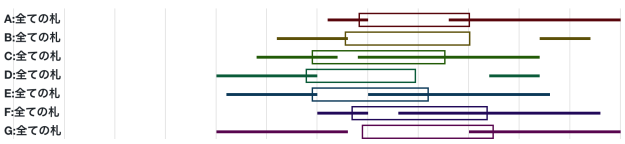


図 10: 全ての札についての箱棒グラフ

図 10 は本測定で読み上げられた 64 枚の全ての札についての箱棒グラフである．箱で示されている動き始めと動き終わりのタイミングの平均値によると，C,D,E の 3 名は他の 4 名に比べて動き始めが早いと言える．

また，棒で示されている動き始めと動き終わりのそれぞれの最大・最小値については，D,E,G の 3 名が動き始めの最小値が小さいことがわかる．このことから，G は動き始めの早さにバラつきがあることが示唆される．

4.2.2 1 字決まりと 2 字決まり

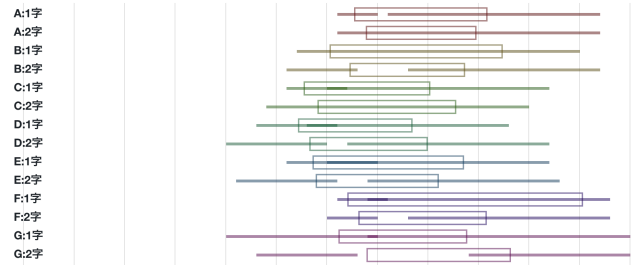


図 11: 1 字決まりと 2 字決まりの札についての箱棒グラフ

次に，1 字決まりと 2 字決まりの札について，選手それぞれの動作のタイミングを比較した (図 11)．1 字決まりは他の決まり字に比べて，7 名ともに動き始めのタイミングが早いことがわかる．2 字決まりも他の決まり字に比べれば，動き始めのタイミングは早いが，1 字決まりよりは 7 名ともに遅い．注目すべきは G で，他の 6 名に比べて 2 字決まりと 1 字決まりの動き始めの伸びが大きいことがわかる．このことから，G は他の 6 名よりも 1 字のときだけ早く動き出す，というタイミングの使い分けができていると予想される．

4.3 動作にかかる時間

次に，動き始めてから動き終わりまでの動作にかかる時間を比較した．図 12 は 7 人の選手の 1 字決まりから 3 字決まりまでの箱棒グラフであり，それぞれの選手の払い始めの平均値と払い終わりの平均値を線で結んでいる．このグラフから，全ての選手に共通して動き始めが 1 字，2 字，3 字の順で速くなっていることがわかる．反対に，動き終わりに着目すると，C，D，G の 4 名は動き終わりも 1 字，2 字，3 字の順で速くなっているのに対して，A,B,E,F の 4 名は動き終わりは 3 字，2 字，1 字の順で速くなっている．つまり，それぞれの平均値を線で結ぶと C,D,G は平行四辺形のようになっているが，A,B,E,F の 4 名は下辺が短い台形のようになっている．

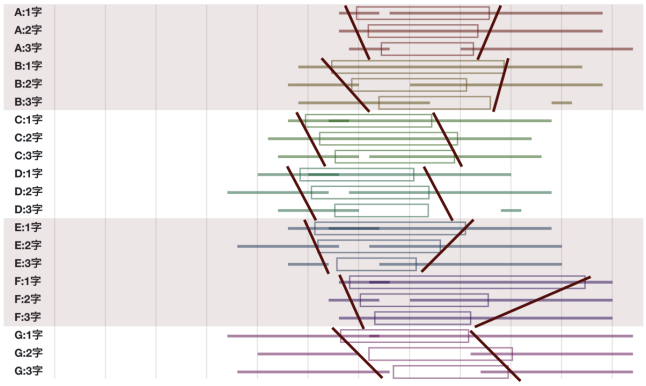


図 12: 1 字決まりから 3 字決まりの札についての箱棒グラフ

ここから，C,D,G の 3 名は 1 字も 2 字も 3 字も同様な動きで取っているのに対し，A,B,E,F の 4 名は決まり字が短いほど動作が大きくなり，動作にかかる時間が長くなると考えられる．競技かるたでは決まり字が短い札を取る際にはより速く動く必要があり，札を取るまでの時間をできるだけ短くする必要がある．そこで，手の加速度を最大にするために体重を移動するテクニックがよく用いられる．

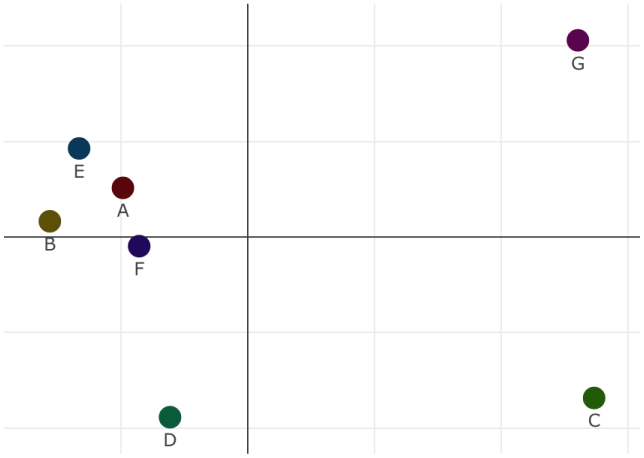


図 13: 体格の特徴量を示す散布図

図 13 にも示される通り，C,D,G の 3 名は身長が高く，A,B,E,F の 4 名は身長が低い．つまり，身長が高いほど腕も長くなり，体重移動を行わずとも速く札に到達できる．それに対して A,B,E,F は，身長のギャップを埋めるために体重移動を行い腕を大きく動かして札を取っており，その結果動作にかかる時間が長くなったことが予想される．

この考察から，箱棒グラフと体格の特徴量を示す散布図を用いて，体格による選手の動作の差異を明らかにできたと言える．

4.4 考察のまとめ

本章では，ボーン表示および箱棒グラフを用いて分析できる 3 つの点についての考察結果を示した．

以上のように，提案システムを用いて三次元ボーン表示によるアニメーションを用いることで，複数の上級者間にどのような差異が見られるか可視化することができた．動き出しのタイミングや構えた際の頭や首の角度など，肉眼やビデオの比較では見つけにくい詳細な差異についても発見できたのが本システムの利点であると考えられる．

5 まとめ・今後の課題

本報告では，競技かるたの払いの動作を対象として我々が開発した可視化システムを用いて，複数の選手の構えや動作の違いを比較した結果を示した．

我々が開発するシステムでは，まず姿勢推定を適用して払い

の動作を測定し，続いてボーン表示とグラフ表示を搭載したシステムによって測定結果を可視化する．この可視化システムにより，複数の選手の払いの動作を詳細に比較し，部位ごとの差異を発見できる．三次元のアニメーション表示と動作タイミングを示す箱棒グラフ表示の 2 つの機能を有するインタラクティブな可視化は，競技かるたの動作を可視化する手法として適していると言える．

今後の展望として，より多くの選手に対して払いの動作を計測し，自身と他の選手の払い方を比較できるシステムを構築したい．各選手の払いの動作から特徴を抽出し，自身と類似度の高い選手を算出できれば，選手の競技能力の向上に役立つと期待している．本研究で開発した払いデータの取得システムは Web ブラウザ上で動作するため，他の選手のデータの収集は環境面で容易である．複数の選手にわたる払いの動作をどのように比較し評価するかは模索中であるため，今後の課題としたい．

文　　献

[1] 丸山礼華, 栗原一貴ほか. Vr 空間におけるマルチモーダルなスキル調整を用いた百人一首かるたのインクルーシブ化の検討. 研究報告エンタテインメントコンピューティング (EC), Vol. 2023, No. 21, pp. 1–11, 2023.

[2] 武田昌一, 長谷川優, 平井祥之, 小杉年範, 津久井勤, 山本誠一. 百人一首かるた選手の競技時の脳の情報処理に関する研究. 近畿大学 生物理工学部 紀要, Vol. 24, pp. 33–43, 2009.

[3] Hiroshi Yamada, Kazuya Murao, Tsutomu Terada, and Masahiko Tsukamoto. A method for determining the moment of touching a card using wrist-worn sensor in competitive karuta. *Journal of Information Processing*, Vol. 26, pp. 38–47, 2018.

[4] 田中佑典, 齊藤剛. モーションキャプチャを用いたダンス上達支援システムの開発. 第 75 回全国大会講演論文集, 第 2013 巻, pp. 225–226, 2013.

[5] Mami Kawanishi, Shuhei Tsuchida, and Takayuki Itoh. Visualization of the repetitive practice of dance motion: Case study with multiple genres of dance. *27th International Conference on Information Visualisation (IV2023)*, pp. 362–367.

[6] Panagiotis Melios. Creative dance learning platform using microsoft azure kinect. In *UNIVERSITY OF CYPRUS DEPARTMENT OF COMPUTER SCIENCE*, 2021.

[7] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[8] Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, Niels Sayez, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. *The 32nd British Machine Vision Conference*, 2021.

[9] Chhaihuoy Long, Eunhye Jo, and Yunyoung Nam. Development of a yoga posture coaching system using an interactive display based on transfer learning. *The Journal of Supercomputing*, Vol. 78, No. 4, pp. 5269–5284, 2022.

[10] 北川リサ, 伊藤貴之. 競技かるたにおける払いの動作の三次元ボーン表示による可視化. *DEIM 2023*, 2023.

[11] 廣戸聡一. 「4 スタンス理論」バイブル. 実業之日本社, 2014.

キンボールのプレイ映像解析のための線形フレーム補間による物体追跡

山下　翔五† 　服部　峻†† 　砂山　渡††

† 滋賀県立大学 工学部 電子システム工学科　〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学 先端工学研究院　〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †tz23syamashita@ec.usp.ac.jp, ††{hattori.s,sunayama.w}@e.usp.ac.jp

あらまし　スポーツの動画解析は，試合の進行や選手のパフォーマンスを分析するために極めて重要な手段となっている．得られた解析結果を活用して，試合映像に解説を付加したり，試合の反則を判定するなどの応用的な活用が考えられる．従来はメジャーなスポーツに焦点が当てられてきたが，マイナースポーツに関する研究はあまりなされていない．本稿では，キンボールと呼ばれるマイナースポーツの動画解析において，物体追跡の精度向上を図る手法を提案し，その有効性を評価することを目的とする．キンボールの複数の試合映像を対象に，YOLOv8 と ByteTrack を用いて物体追跡を行う．特に物体追跡中に見失った部分に対しては線形補間やベクトル補間（3 種類）を行う．評価実験により，提案手法の有効性を評価し，今後の課題について検討する．

キーワード　線形補間, 物体検出, 物体追跡, キンボール, マイナースポーツ

1　はじめに

近年，スポーツの動画解析は試合の進行や選手のパフォーマンスを詳細に分析し，戦術の改善やトレーニングの最適化に役立つ重要な手段となっている．得られた解析結果を活用することにより，試合映像に解説を付加したり，試合の反則を判定するなど，様々な応用的な活用が考えられる．また，野球やサッカー，バスケットボールなどの球技では，戦術の高度化により，球技映像の分析が一大ビジネスとなっている．記録を競う陸上や水泳などの競技では，競技者の体の動きを細かく分析し，適切にコーチや選手にその情報をフィードバックするシステムが活用されつつある [1].

このように，従来のスポーツ動画解析の研究では，主にメジャースポーツに焦点が当てられていて，マイナースポーツの動画解析に関する研究が不足している．本稿では，その中でも特にキンボールと呼ばれるマイナースポーツに焦点を当てる．

キンボールは直径 122cm の大きなキンボールを使用して，主に室内で行われる競技である．1 チーム 4 人で 3 チームが同時に試合を行い，床にキンボールを落とさないようチームで協力するスポーツである．チームがボールを床に落としたり，ミスをした場合，そのチーム以外の 2 チームに得点が与えられ，得点制や時間制でゲームを行う．

本稿の目的は，キンボールのプレイ動画解析において物体追跡の精度向上を図り，その有効性を評価することである．本稿では，キンボールの試合画像を収集し，学習データとして使用するためにアノテーションを行い収集されたデータを元に，キンボールに特化した新たな学習済みモデルを作成する．この学習済みモデルを用いて，キンボールの試合映像に対して物体の検出と追跡を実施する．物体の検出と追跡を行い，検出できなかった部分，特にボールに対して線形補間やベクトル補間（3 種類）を行い，その有効性を評価する．

本稿の構成として，2 章では関連研究を紹介する．次に，3 章では本稿の提案手法を各フェーズごとに分解して示す．4 章では，各フェーズにおける出力結果を記述する．5 章では，提案手法の性能評価を行い，評価実験の結果を詳述する．最後に 6 章では，まとめと今後の課題について言及する．

2　関連研究

本章では，関連研究について述べる．

2.1　物体検出

物体検出 (Object Detection) は，画像や動画中に存在する物体の位置やクラスを検出する技術である．主な目的は，入力データ内の物体の存在を特定し，それらをバウンディングボックスで囲み，さらに物体のクラスの識別行うことである．

既存の物体検出アルゴリズムである R-CNN [2] では，物体検出の基本となる領域ベース検出器と分類器を用いて，画像中の物体をバウンディングボックスで囲む形で検出を行う．画像の領域推定と分類の処理が分割されているため，処理が複雑となり処理の時間が長くなるという課題が存在する．

対して，YOLO [3] では単一の回帰問題として物体検出を捉え，単一の CNN を使用して，複数のバウンディングボックスとそれらのボックスのクラスの確信度を同時に予測することが可能である．このアプローチにより処理が迅速に行われる．YOLO は画像中の物体を素早く識別することが可能であるが，小さい物体を正確に検出することは課題となっている．

2.2　物体追跡

物体追跡 (Object Tracking) は，連続するフレーム内で動く物体を検出し，それを追跡する技術である．主な目的は動画やリアルタイムの映像データにおいて，特定の物体や対象を一貫

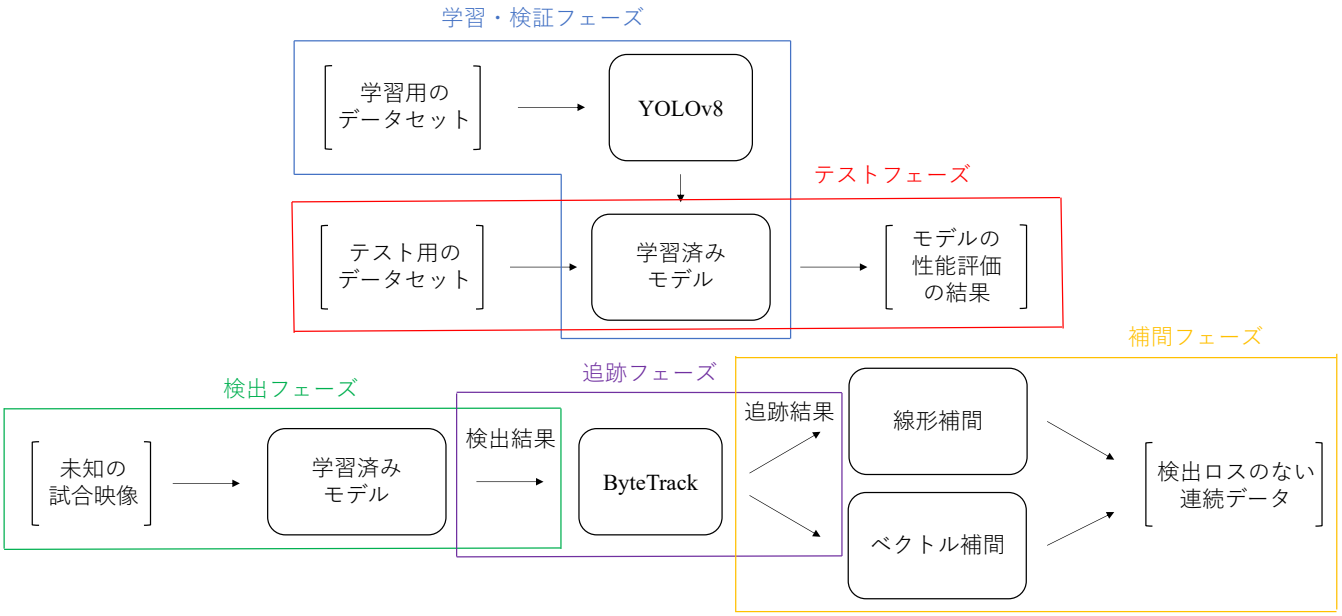


図 1: システムの概略図

して追い続けることである．

既存の物体追跡アルゴリズムである SORT (Simple Online and Realtime Tracking) [4] では，Kalman Filter を用いて，追跡中の物体の予測位置と実際の検出結果の重なり度合い (IoU; Intersection over Union) をもとに，マッチングを行い，連続したフレームで同一オブジェクトに対し一意の id を割り当てる．シンプルな設計であるが，物体が一時的に検出されない時に補間が難しいという課題がある．

一方，ByteTrack [5] では検出時の確信度に焦点を当て，検出結果を高い確信度と低い確信度で分割し，それぞれに対して Kalman Filter を用いて，マッチングを行うことで，確信度が低い物体に対しても補間を行い，検出漏れの発生を防ぐ手法となっている．しかし，ByteTrack では，物体の外観情報を考慮せずに IoU のみで物体のマッチングを行っている．これにより，オブジェクトの検出ロスや id の振り直しが発生する．

本稿では，YOLO や ByteTrack を用いて，キンボールの物体検出ならびに物体追跡を行うが，その際に発生したボールの検出ロスの改善について検討していく．

3 提 案 手 法

本稿のシステムの流れを，図 1 に示す．システム図は，キンボールに特化した学習済みモデルを作成するための学習・検証フェーズ，作成した学習済みモデルの性能評価のためのテストフェーズ，未知のデータに対するオブジェクトの検出を行う検出フェーズ，検出結果から物体追跡を行う追跡フェーズ，物体の検出・追跡を実施した結果を用いて，ボールの検出ロス部分の補間を行う補間フェーズに分かれている．各フェーズに関する内容を以降の節で詳しく記述していく．

3.1 学習・検証フェーズ

提案手法の学習フェーズでは，YOLOv8 [6] を用いて物体検出を行うためのキンボールに特化した学習済みモデルを作成した．使用されたデータセットは 591 枚で，その中の 80%を学習用，20%を検証用に分割し，学習および検証を行った．学習に使用された画像は，複数のキンボールの試合映像をフレームごとに分割することで取得したものである．得られた画像を手作業でアノテーションを行うことで，学習に用いる教師付きデータを作成した．アノテーションの様子を図 2 に示す．キンボールの画像にはキンボール，プレイヤー，レフェリーが写っており，これらの 3 つのクラスに対して学習および検証を行った．学習および検証に用いた画像例を図 3，4，5 に示す．画像例からわかるようにキンボールの試合では，主にピンク，グレー，ブラックの 3 種類の色のボールを使用している．なお，学習に用いたデータセットでは，各色のキンボールが写る画像を均等に収集し，ボールの色に依存せず，高い精度が得られるように配慮した．

図 2: アノテーションの様子

3.2 テストフェーズ

提案手法のテストフェーズでは，学習モデルの性能評価を行うために，学習に使用していない未知のキンボールの試合映像を 3 試合分用意してテストを実施した．学習フェーズと同様に，

図 3: 学習・検証に用いた画像例 1

図 4: 学習・検証に用いた画像例 2

図 5: 学習・検証に用いた画像例 3

試合映像をフレームごとに分割し，アノテーションを行うことでテストに用いるデータセットを用意した．各試合について，1 試合の中から 1 プレー分，約 6 秒を取り出し，この動画から約 170 枚の画像を取得してモデルの性能テストを行った．このテストにより，学習モデルが未知の試合映像に対してどれだけ正確に物体を検出・追跡できるかを検証した．複数の試合を用いることでモデルの汎用性に対する評価を行った．

3.3 検出フェーズ

提案手法における物体検出フェーズでは，学習や推論に用いていない未知のキンボールの試合映像を検出の対象とした．システムの汎用性を向上させ，新規の試合映像にも適用可能なモデルの構築を目指すためである．この未知の試合映像に対して，事前に学習させたモデルを活用し，YOLOv8 を用いて物体検出を行った．ここで使用した未知の映像はフレームごとに分割して検出が行われた．これにより各フレームごとに物体検出が独立して行われた．

図 6: YOLOv8 を用いた検出の様子

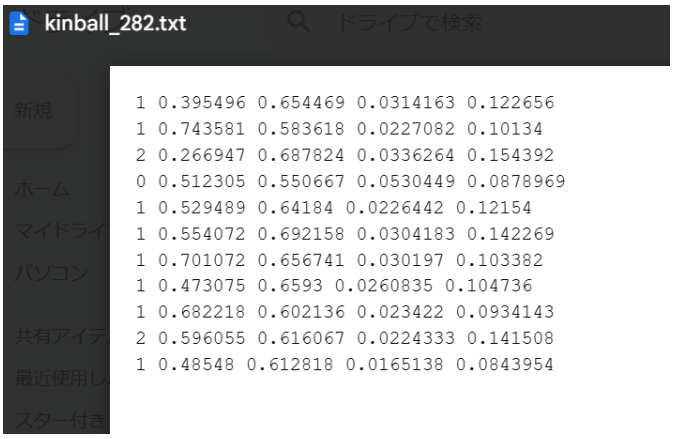


図 7: 検出時に取得したテキストデータ例

YOLOv8 を用いて試合映像の物体検出した様子の一部を図 6 に示す．図から，各オブジェクトのクラス名，クラスの確信度，オブジェクトを囲む矩形が可視化されているのがわかる．矩形が検出された際には，各オブジェクトのクラス番号，中心の x，y 座標，矩形の幅，矩形の高さをリストでフレームごとにテキストデータとして取得した．取得したテキストデータの例を図 7 に示す．クラス番号は，0 が「キンボール」，1 が「プレイヤー」，2 が「レフェリー」を表している．また，フレームのシーケンス番号は，図 7 のようにテキストデータのファイル名から取得した．

3.4 追跡フェーズ

提案手法における追跡フェーズでは，物体検出によって得られた結果を元に，ByteTrack [5] を活用して物体追跡を行った．物体検出では各フレームごとに独立して検出が行われたが，物体追跡では，オブジェクトの動きや位置を過去のフレームと比較して追跡を行った．これにより，各オブジェクトが連続したフレーム間でどのように移動しているかを把握することが可能となった．ByteTrack を用いて試合映像の物体追跡した様子の一部を図 8 に示す．図から，ByteTrack を導入した物体追跡により，各オブジェクトには一意の id が割り当てられ，各オブ

図 8: ByteTrack を用いた追跡の様子

図 9: ボールのみを追跡した様子

図 10: id の振り直しを無くした時の様子

ジェクトごとの識別が可能になったことがわかる．

3.5 補間フェーズ

提案手法の補間フェーズでは，検出・追跡により取得したオブジェクトの位置情報を用いて，特にキンボールの検出が途切れている個所に対して，補間を行う．キンボールに対しての補間を行うため，キンボールを追跡した結果のみをテキストデータとして取得し，テキストデータを用いて補間を実施する．

ここで，キンボールのみを追跡した様子を図 9 に示す．図 9 は，追跡した試合映像の中から，9 枚の画像を抜き出したもの

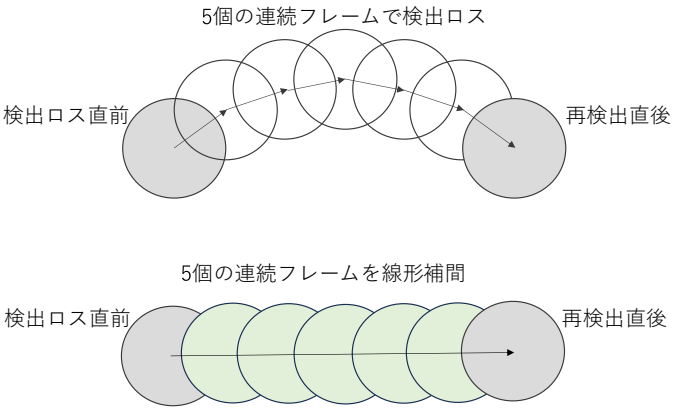


図 11: 線形補間の概略図

で，No.30 や No.70 などは抜き出したフレームのシーケンス番号を表す．左上から，キンボールの id は 1 から 4 に変化していることがわかる．追跡ロスが起こるたびに id が振り直され，ロス前後で異なるオブジェクトであると判別されている．

そこで，キンボールは試合中に 1 つしか用いられていないことから，キンボールの追跡ロスが起こった際，id の振り直しを無くす処理を追加した．その時の追跡の様子を図 10 に示す．図 10 では，追跡ロスが起こっても id が変化していないことがわかる．これにより，ロス前後でキンボールが同一オブジェクトであると判別される．キンボールの id がロス前後で変化しない追跡結果を用いて，線形補間やベクトル補間を行う．但し，ベクトル補間において，単一の検出点が検出ロス点系列の途中で孤立している場合が有り得るが，検出ロス前後のベクトルを計算できず，ベクトル補間を適用できないため，単一の検出点は全て，検出ロスしたものと見なした上で適用している．以降ではそれぞれの補間手法について詳細に説明する．

3.5.1 線 形 補 間

線形補間の概略図を図 11 に示す．線形補間では，検出が途切れたフレームにおいて，検出ロス前後のフレームにおけるボールの位置情報（中心の x，y 座標）と矩形の幅・高さを利用して補間を行う．線形補間を行う際に用いた計算式を以下に示す．

$$loc_t = loc_{t-1} + (loc_{\text{after}} - loc_{\text{before}}) \cdot \frac{1}{\text{after} - \text{before}} \quad (1)$$

$$w_t = w_{t-1} + (w_{\text{after}} - w_{\text{before}}) \cdot \frac{1}{\text{after} - \text{before}} \quad (2)$$

$$h_t = h_{t-1} + (h_{\text{after}} - h_{\text{before}}) \cdot \frac{1}{\text{after} - \text{before}} \quad (3)$$

loc はボールの位置情報を表す．t は検出ロス中の任意のフレームのシーケンス番号を示し，before は検出ロスの直前，after は検出ロスが終わり再検出された時のフレームのシーケンス番号を表す．この計算式を用いて，検出ロス前後の位置情報の差を利用し，検出ロス中のフレームでの位置情報を予測し，補間を行う．また矩形の幅 w や，矩形の高さ h も，位置情報と同様に計算し補間を行う．これにより，検出が途切れたフレームにおいても，前後の位置情報を利用してボールの移動軌跡の連続性を確保することが目的である．

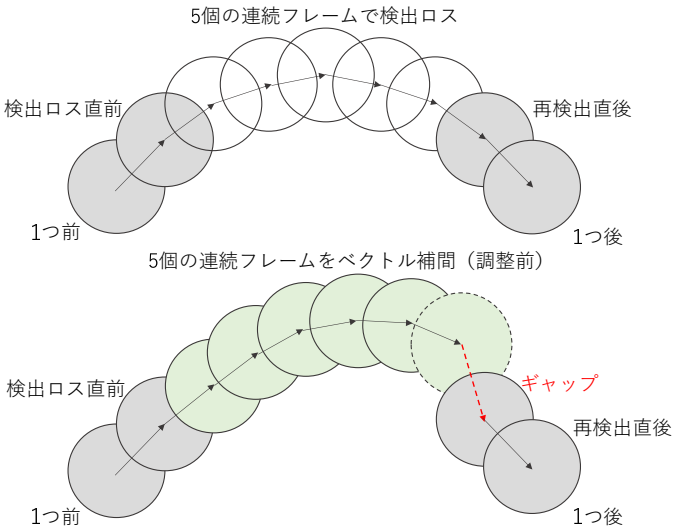


図 12: ベクトル補間の概略図

3.5.2 ベクトル補間

次にベクトル補間では，検出が途切れたフレームにおいて，検出ロスの直前と 1 つ前のフレーム，検出ロス直後と 1 つ後のフレームの位置情報（中心の x，y 座標）を使用して，検出ロス前後でボールの移動したベクトルを計算し，それを用いて補間を行う．但し，矩形の幅と高さに関しては，線形補間を用いて補間を行う．ベクトル補間に用いた計算式を次に示す．

$$v_{\text{before}} = \begin{bmatrix} x_{\text{before}} \\ y_{\text{before}} \end{bmatrix} - \begin{bmatrix} x_{\text{before}-1} \\ y_{\text{before}-1} \end{bmatrix} \tag{4}$$

$$v_{\text{after}} = \begin{bmatrix} x_{\text{after}+1} \\ y_{\text{after}+1} \end{bmatrix} - \begin{bmatrix} x_{\text{after}} \\ y_{\text{after}} \end{bmatrix} \tag{5}$$

$$\begin{aligned} v_t = v_{\text{before}} \cdot \left(\frac{\text{after} - t + 1}{\text{after} - \text{before} + 1} \right) \\ + v_{\text{after}} \cdot \left(\frac{t - \text{before}}{\text{after} - \text{before} + 1} \right) \end{aligned} \tag{6}$$

$$u = \frac{(loc_{\text{before}} + \sum_{t=\text{before}+1}^{\text{after}} v_t) - loc_{\text{before}}}{loc_{\text{after}} - loc_{\text{before}}} \tag{7}$$

$$loc_t = loc_{t-1} + u \cdot v_t \tag{8}$$

線形補間と同様に loc はボールの位置情報を示し，before は検出ロスの直前，after は検出ロスが終わり再検出された時のフレームの番号を表す．x と y はボールの中心座標を表す．ここで，ベクトル補間の概略図を図 12 に示す．図 12 では，検出ロスしている部分に対してベクトル補間を行っている．再検出直後のフレームまでベクトル補間を行う際に，ベクトル補間を行った最後のフレームと再検出直後のフレームでのボールの位置にギャップが生まれる．そこで，u はベクトル補間により生まれるギャップを補正し，スムーズに追跡できるように調整を行うものである．ベクトル補間はボールの移動方向を考慮した補間を行うため，線形補間より精度が高く補間されることが期待される．しかし，検出ロス前後で計算したベクトル値が実際のボールのベクトル値と異なる場合に補間がスムーズに行えない可能性が存在する．そこでベクトル補間の改善手法を次の節で示す．

3.5.3 k-ベクトル補間

通常のベクトル補間の式 (4) や (5) では，検出ロスの直前と 1 つ前のフレーム，検出ロス直後と 1 つ後のフレームの位置情報を用いてベクトルを計算するが，k-ベクトル補間では，検出ロスの前後で両方とも同一の k 個離れたフレームの位置情報を用いてベクトルの計算を行う．k-ベクトル補間に用いた式を以下に示す．

$$v'_{\text{before}} = \frac{1}{k} \left(\begin{bmatrix} x_{\text{before}} \\ y_{\text{before}} \end{bmatrix} - \begin{bmatrix} x_{\text{before}-k} \\ y_{\text{before}-k} \end{bmatrix} \right) \tag{9}$$

$$v'_{\text{after}} = \frac{1}{k} \left(\begin{bmatrix} x_{\text{after}+k} \\ y_{\text{after}+k} \end{bmatrix} - \begin{bmatrix} x_{\text{after}} \\ y_{\text{after}} \end{bmatrix} \right) \tag{10}$$

$$\text{gap} = \sqrt{\left(loc_{\text{after}} - \left(loc_{\text{before}} + \sum_{t=\text{before}+1}^{\text{after}} v_t \right) \right)^2} \tag{11}$$

k は 1 から 10 の任意の整数を表すパラメータである．gap は図 12 のように再検出時のフレームまでベクトル補間を行った際の最後のフレームでのボールの位置と再検出時のフレームでのボールの位置との距離を示す．この gap が最小となる k を探してパラメータを最適化し，ベクトル補間を行う．これにより，ベクトル計算に用いる標本数を増やし，頑健性を向上させることが主な目的である．

3.5.4 j,k-ベクトル補間

k-ベクトル補間では，検出ロスの前後で両方とも同一の k 個離れたフレームの位置情報を用いてベクトルの計算を行うが j,k-ベクトル補間では検出ロスの直前では j 個前のフレーム，検出ロス直後では k 個後のフレームの位置情報を用いてベクトルの計算を行う．j,k-ベクトル補間に用いた式を以下に示す．

$$v''_{\text{before}} = \frac{1}{j} \left(\begin{bmatrix} x_{\text{before}} \\ y_{\text{before}} \end{bmatrix} - \begin{bmatrix} x_{\text{before}-j} \\ y_{\text{before}-j} \end{bmatrix} \right) \tag{12}$$

$$v''_{\text{after}} = \frac{1}{k} \left(\begin{bmatrix} x_{\text{after}+k} \\ y_{\text{after}+k} \end{bmatrix} - \begin{bmatrix} x_{\text{after}} \\ y_{\text{after}} \end{bmatrix} \right) \tag{13}$$

3.5.3 項と同様に，j や k は 1 から 10 の任意の整数を表すパラメータであり，gap が最小となる j や k を探してパラメータを最適化し，ベクトル補間を行う．これにより，前項の k-ベクトル補間と比べて，ベクトルの計算時に検出ロス前後で各々，より最適なベクトルを用いて補間を行うことが可能となる．一方で，パラメータの組み合わせが増大し ($\mathcal{O}(k)$ から $\mathcal{O}(jk)$ に)，最適化の計算時間がより掛かってしまうというデメリットも存在する．

4　各フェーズにおける出力の結果

本章では，提案手法で記述した各フェーズでの検証結果や検出の結果などを詳細に記述する．また，補間フェーズにおいて，追跡結果に対し線形補間やベクトル補間（3 種類）を行うことで，補間前と補間後でどのように精度が変化したかについては，5 章の評価実験で検証していく．

4.1　学習フェーズ

キンボールに特化した学習済みモデルの作成の際に，エポック数を 100 とし，YOLOv8 を用いて学習と検証を行った．検証の結果を図 13，14，15 に示す．

まず，図 13 においては学習データを使用した際の損失がどのように変化したかを示すグラフである．この損失は，検出されたオブジェクトのバウンディングボックスと，正解データとの誤差を表している．初めのエポックでは損失が大きくなっていることが観測される．エポックが進むにつれ，損失は，徐々に収束し，検出精度向上に向けた学習が進んでいることが期待される．

図 14 においては，検証データを用いた際の損失の推移を示している．図 13 と同様に損失は初めより低下しているが，上昇や下降が繰り返されながら，一定値に収束しているようにも捉えられる．一定値に収束することは，学習モデルが一定の精度で検証データに対応できるようになったことを示唆している．

図 15 においては，エポック数における mAP の推移である．エポックの早い段階で mAP の値が上昇し，一定値に収束していることがわかる．学習済みモデルにおいては，性能が最も優れていたとされるエポックでの重みを使用している．

4.2　テストフェーズ

学習済みモデルに対する性能テストの結果を表 1 に示す．テストに使用した 3 種類の未知の試合映像では，それぞれ異なる色のキンボールを用いているが，どの試合でも適合率が 9 割を超え高い精度が得られた．試合 1 の映像ではグレーのボールが使用されており，ピンクやブラックのボールを用いた試合 2，3 より適合率が低くなった．これはグレーのボールが背景色と似た色であるため，ボールの境界が適切に検出できていないと考えられる．

プレイヤーやレフェリーに関しては，各試合で適合率が変動したが，平均して 7 割以上の適合率が得られた．プレイヤーやレフェリーの検出時には，背景部分がプレイヤーやレフェリーとして誤検出される場合があった．プレイヤーやレフェリーは画像に対して小さいサイズであることや，ボールに密集する場合にオクルージョンが生じることから，ボールに比べて適合率が低くなると考えられる．

4.3　検出フェーズ

学習に用いていない未知の試合映像に対して，YOLO を用いて検出を行った時の結果の一部を図 16 に示す．検出を行う際は，グレーのボールを使用する試合映像を用いた．試合映像の長さは約 9 秒のものを用いて検出を行った．

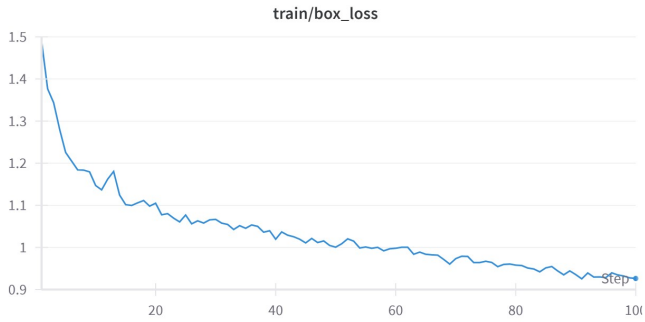


図 13: 学習時のエポック数に対するロスの変化

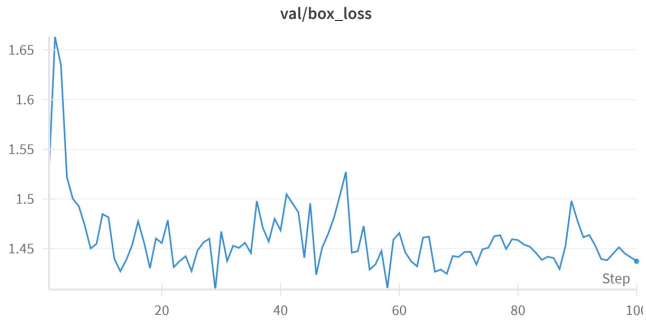


図 14: 検証時のエポック数に対するロスの変化

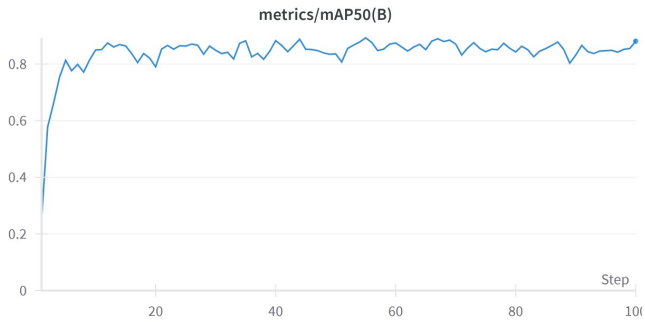


図 15: 検証時のエポック数に対する mAP の変化

表 1: 各オブジェクトの適合率

Class	Game1	Game2	Game3	Average
kinball	0.91	0.99	0.99	0.96
player	0.63	0.74	0.83	0.73
referee	0.66	0.77	0.84	0.79

図 16 のように，ボールの検出が途切れている箇所が複数存在した．これはボールの色が背景の色と似ていて，ボールと背景の境界が正確に捉えられていないことが原因であると考えられる．また，ボールやプレイヤーに関しては，オブジェクトが重なっている時に，オブジェクトの境界が不明瞭となり，検出がスムーズに行われていない部分が存在した．このような境界があいまいな箇所においては，検出精度が低下する傾向が見られた．

図 16: 検出結果

図 17: 追跡結果

4.4 追跡フェーズ

検出フェーズで検出された結果に対して，ByteTrack を用いて追跡を行った結果の一部を図 17 に示す．図 17 では，キンボールの id が 56 となっていることがわかる．各フレームで検出されたオブジェクトの数は約 15 体であり，これにより，id の振り直しが 30 回近く発生していることが示されている．ByteTrack は検出結果の確信度の低いオブジェクトに対しても考慮を行い，検出漏れを減少させる手法であるが，検出結果と追跡結果で大きな精度の変動は見られなかった．次章で，追跡結果を用いて，ボールの検出ロスした部分に対して補間を行った結果を記述していく．

5 評価実験

検出・追跡によって取得した結果，線形・ベクトル補間後の結果，用いた試合映像の正解データを用いて，補間前と補間後でどのように精度が変化したか評価を行った．

5.1 評価の方法

補間前と補間後で精度の変化を評価するために，用いた試合映像において，フレームごとにボールの位置情報が記述されたテキストデータと，補間前・補間後に取得したテキストデータを比較し，それぞれのフレームにおけるボールの矩形の重なり度合い (IoU) を計算した．これにより，検出精度がどのように変化したか検証した．

5.2 実験結果

評価実験の結果を図 18，19，20 に示す．図 18 では，IoU 値

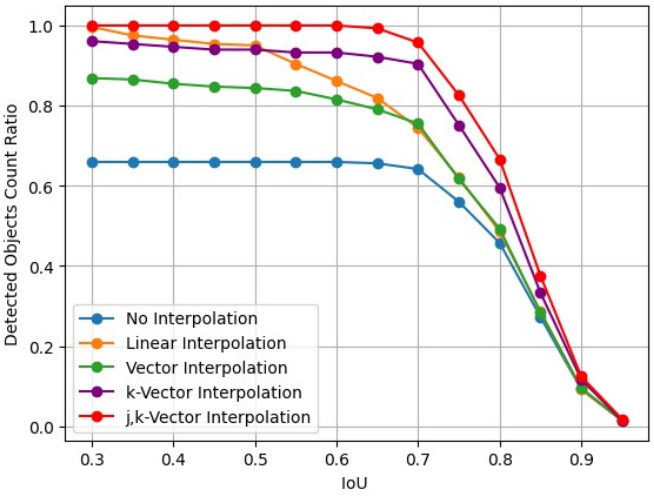


図 18: IoU 値に対する検出数の割合の推移

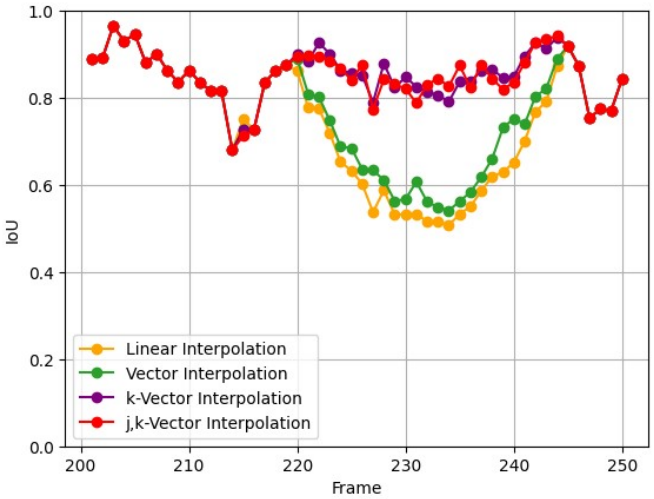


図 19: フレーム No.201 から No.250 までの IoU 値の推移

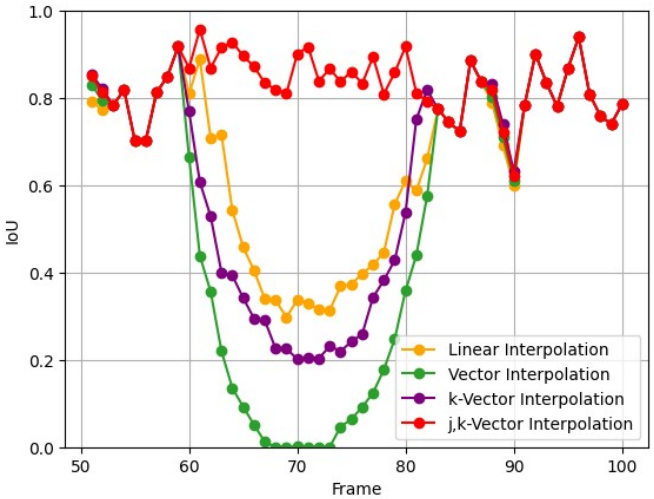


図 20: フレーム No.51 から No.100 までの IoU 値の推移 2

の閾値を変化させた際の，ボールの検出数の割合の推移を示している．図 18 から，いずれの補間手法においても補間前と比べてボールの検出数の割合が増加し精度が向上していることが

わかる．しかし，IoU の閾値が高い時では，補間前と補間後で，検出数の割合に変化があまり見られない．このことから，補間を行うことにより，検出割合は増加しているが，実際の矩形と補間した矩形とでは，少し位置が異なることがわかる．閾値が低い場合では，線形補間においては，ほとんどのフレームで補間が可能となったが，ベクトル補間においては，補間が正確に行えていないフレームが存在した．これはベクトル補間を行う際に計算したベクトルが，実際のボールのベクトルと異なることが原因であると考えられる．また，線形補間やベクトル補間は IoU の閾値が高くなるにつれ，検出数の割合が低下した．最適なベクトル値を用いて補間を行う k-ベクトル補間や j,k-ベクトル補間では IoU の閾値が高い場合でも検出数の割合が増加し，精度の向上が見られた．k-ベクトル補間では，補間が正確に行えていないフレームが存在することがわかる．j,k-ベクトル補間ではほとんどのフレームで補間が可能となり，k-ベクトル補間と比べて，より汎用的に補間が可能であることがわかる．

次に，図 19，20 では，補間を行った試合映像から，検出ロスを含む 50 フレームである No.201 から No.250，及び，No.51 から No.100 を抜き出し，各フレームにおける IoU 値の変化を示している．図中で，各補間手法の IoU 値が同じであるフレームは，ByteTrack でそもそも検出・追跡が成功しているフレームである．IoU 値が異なる場合は，検出・追跡に失敗し，提案手法によって補間が行われたフレームである．

図 19 から，検出ロスが連続的に発生し，その期間が長い場合，ベクトル補間は線形補間より IoU 値が高くなっていることがわかる．また，k-ベクトル補間や j,k-ベクトル補間は，より精度よく補間が行えていることがわかる．一方で，検出ロスが短い部分では，線形補間の方がより精度よく補間できていることがわかる．

また，図 20 は，ベクトル補間が不十分である時の例を示している．k-ベクトル補間は通常のベクトル補間より精度良く補間を行えているが，線形補間よりも IoU の値が低くなっている．また，j,k-ベクトル補間は，通常のベクトル補間では正確に補間が行えないフレームに対しても有効であることがわかる．

6 おわりに

本稿では，キンボールの試合映像を使用し，キンボールに特化した学習済みモデルを作成して，学習に用いていない未知の映像を用意し，YOLOv8 や ByteTrack を用いてボールやプレイヤー，レフェリーの検出や追跡を行った．この過程で生じたキンボールにおけるボールの検出ロス部分に対して線形補間やベクトル補間を実施し，補間前と補間後でどのように精度が向上したか評価を行った．

キンボールの試合で用いられているボールの色は主にピンク，グレー，ブラックの 3 種類であり，その中でグレーのボールに対して検出ロスが多く見られた．これは，グレーのボールが背景色と似ていて，オブジェクトの境界が不明瞭であったことが原因であると考えられる．また，プレイヤーやレフェリーに関しては，背景部分がプレイヤーやレフェリーと誤検出される場合があった．また，プレイヤーやレフェリーは画像に対して小さいサイズであることや，ボールに密集しオクルージョンが生まれることから，検出ロスしている部分が多く見られた．

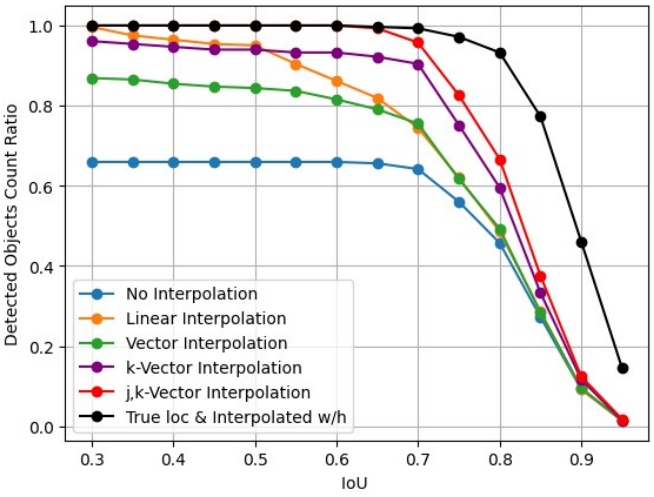


図 21: 位置情報のみ正解データを用いた時の検出割合の推移

そこで，特にボールの検出ロス部分に対して線形補間やベクトル補間を行った．補間前と比べてボールを検出できたフレームの割合が増加し，連続したボールの位置情報を取得することが可能となった．しかし，線形補間とベクトル補間のいずれにおいても，補間を行った際に取得したボールの位置は，実際のボールの位置と多少異なる結果となった．また，ベクトル補間においては，正確に補間が行えていないフレームが存在した．実際のボールのベクトルと検出結果を用いて計算したベクトルの値が異なることが原因として挙げられる．そこで，ベクトル補間におけるこれらの問題点の修正を行うことで，より正確に補間を行うことが可能となった．また，どの検出ロスをしているフレームにおいてもベクトル補間が有効であることがわかった．しかし，IoU の閾値が高い場合では検出数の割合は補間前と比べて精度の向上は見られなかった．

そこで，ベクトル補間（3 種類）において，位置情報 (loc) についてはベクトル補間を用いているが，矩形の幅 (w) や高さ (h) については線形補間を用いており，矩形についても補間手法を改善することでより補間の精度が向上することが考えられる．ここで，位置情報については正解データを用い，矩形の幅や高さについては検出結果や線形補間を行った結果を用いた際の，検出数の割合の推移を図 21 の黒色の実線で示す．図 21 から物体検出や線形補間により取得した矩形データによっても IoU の高い場合では検出数の割合が低下しており，1.0 にはなっていないことがわかる．これは，検出時や線形補間時に取得したボールの矩形が実際のボールの矩形と異なることが原因であると考えられる．そこで，本稿で提案したボールの中心座標のみをベクトル補間するのではなく，矩形の左上の座標と右下の座標もベクトル補間するなど，矩形データに関してもベクトル補間を行うことで，より補間の精度が向上することが期待される．

これらの補間手法により，ボールの検出ロスフレームに対し

て補間が可能となり，また，補間を行うことにより，より精度よくオブジェクトの検出，追跡を行うことが可能となった．これらの補間手法は，オクルージョンなどにより検出ができない際にも有効であると考えられる．

また本稿ではボールに関する検出精度の向上を目指したが，プレイヤーやレフェリーに関しても精度を向上させるために検討していく．精度を向上させた後には，試合映像に解説を付与したり，試合の反則を判定したりなど，応用的な活用にも取り組んでいく予定である．

文　　献

[1] 清水 千弘, 清水 陽司, “スポーツ競技と AI,” 人工知能学会論文誌, Vol.34, No.4, pp.492–496 (2019).
[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” arXiv:1311.2524 (2014).
[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” arXiv:1506.02640 (2016).
[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, “Simple Online and Realtime Tracking,” arXiv:1602.00763 (2016).
[5] Yifu Zhang, Peoze Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, “ByteTrack: Multi-Object Tracking by Associating Every Detection Box,” arXiv:2110.06864 (2021).
[6] Glenn Jocher, ultralytics YOLOv8, https://github.com/ultralytics/ultralytics （参照 2024-01-08).

DEIM Forum 2024 XX-Y

離脱傾向を踏まえたインターネット動画広告データの可視化システム

三浦　梨花†　　大矢　隼士††　　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
†† Septeni Japan株式会社　〒 160–6128 東京都新宿区西新宿 8-17-1
E-mail: †{g1820536,itot}@is.ocha.ac.jp, ††hayato.oya@septeni.co.jp

あらまし　インターネット広告の市場規模は拡大を続けている．効果的なインターネット広告の制作と配信は，今後ビジネスにおいてもますます重要になるといえる．一方で，インターネット広告は配信ターゲットを設定した上で特定の層に向けて発信するため，ターゲット設定が不適切であると広告効果が得られにくいという背景がある．よって，過去の広告配信データを利用し，適切なターゲット設定を分析することが求められる．そこで我々は，効果的なターゲット設定の発見を目指すために，インターネット動画広告データを対象にインタラクティブな可視化システムを開発している．本報告では，この可視化システムに必要な 3 つのタスクを定義し，このタスクを満たすように可視化システムを拡張した結果を述べる．具体例として本報告では，Yahoo!広告で配信された広告のダミーデータ 224,355 件と LINE 広告で配信された広告のダミーデータ 89,400 件を可視化システムで可視化した結果と可視化結果から得られる考察を報告する．

キーワード　可視化手法，動画広告，離脱率，クリック率，コンバージョン率

1　はじめに

インターネット広告の市場規模は年々拡大を続けている．電通の報告 [1] によると，インターネット広告費は 2022 年に，前年比の 114.3%である 3 兆 912 億円に達した．特に，動画広告においては前年比 115.4%の 5,920 億円と大きく伸長し，2023 年は引き続き高い成長率を保ち，6,852 億円まで増加すると予測されている．効果的なインターネット広告の制作と配信は，今後ビジネスにおいてもますます重要になるといえる．このようなインターネット広告市場の拡大という背景から，効果的な広告活動を行うために現在まで様々な研究がなされている．Eric ら [2] は，インターネット広告が多くのユーザーにとって不快であるという出発点から，広告に印象のラベル付を行いクラスタ分析を用いることで，具体的にどのような広告がユーザーにとって不快なのか，そして，好まれるのかを調査した．Shon ら [3] は，Structural Equation Modeling(SEM) を用いてモデルを構築することによって，インターネット広告が配信された際に不快感を与える要因を分析した．また，松本ら [4] は，マーケティングデータをクラスタリングと主成分分析を用いて分析することで，消費者の特徴を把握した．上記のような研究からもわかるように，効果的な広告活動のためにはユーザーを不快にさせる広告の特徴を分析した上で広告を改善したり，消費者の特徴を把握した上で適切な広告を適切な消費者へ配信することが求められる．このような研究手法の他にも，効果的な広告配信を行うための手段としてデータの可視化は一般的な手法の 1 つである．また，データ分析において，新たな示唆を得るためにはインタラクティブな可視化システムを構築し，探索的にデータを分析できることも重要な要素の 1 つであるといえる．広告データに限らずインタラクティブにデータを可視化することで，新たな示唆を得ることを達成した研究は多くある．Kui ら [5] は，テレビ視聴率に影響を与える要因を見つけるために視聴率を視覚的に分析するインタラクティブなシステムを開発した．Dongyun ら [6] は，政策立案者の意思決定をサポートするために政策データの傾向やパターンを理解することができるインタラクティブな可視化システムを提案した．これらの研究において，インタラクティブな可視化システムを開発する上で重要視されていることは，分析対象のデータに合わせて適切な可視化手法を用いたインタラクティブな可視化システムを開発することである．

そこで我々は，インターネット動画広告の配信結果データを対象にしたインタラクティブな可視化システム [7] を開発している．インターネット広告は，テレビなどのマスコミ 4 媒体広告と異なり，特定の層に向けて広告を配信するため，広告ごとに適切なターゲットを設定し配信を行わないと，期待した効果を得られないのが難点である．その上，インターネット広告のターゲット設定は,「女性または男性」のような単純なターゲット設定ではなく,「女性×年齢」というような複数の属性を組み合わせたターゲット設定である．そのため，どの属性の組み合わせが広告効果を発揮したのかを分析結果から見つけることが難しい．一方で，これらを考慮して，広告データを可視化分析することができる手法は現状少ない．

本報告では，上述した動画広告データの特徴を踏まえた可視化システムを実現するために，下に記す T1 から T3 の 3 つのタスクを定めて本研究を拡張した結果を述べる．

T1：動画広告の離脱率ごとにクリック率・コンバージョン率にどのような特徴が現れるのかを分析できるようにすること．

T2：クリック率が大きく変動するようなパラメータの組み合わせを見つけやすくすること．

T3：見るべき価値のある広告データを探索しやすくすること．

本報告の構成は以下の通りである．2 章では関連研究について，3 章では提案手法について，4 章では実行結果と分析結果から得た考察について，5 章では本報告のまとめと今後の課題について述べる．

2 関 連 研 究

本章では関連研究として，広告データの可視化手法，広告効果向上を目的としたデータ活用手法，そして探索的データ分析を目的とした可視化ツールの開発手法について述べている論文を紹介するとともに，本研究との差異を述べる．

2.1 広告データの可視化手法

本研究で利用している動画広告データは，多次元データである．効果的な多次元データの可視化手法については，現在まで多様な議論がなされてきた．一方で，本研究のような，広告データに特化した可視化分析の研究も少数ではあるもののいくつか存在する．Di ら [8] は Targeting Vis というインタラクティブな分析ツールを提案した．このツールを利用することによって広告アナリストは，ターゲット広告の配信データを視覚化し，広告主の配信パターンを調査することで，広告をより効果的にする方法を発見することができる．Targeting Vis においても，本研究と同様にターゲットの組み合わせパターンを可視化する手法が用いられており，この手法が効果的な広告設定の発見に有用であることを示唆している．Shangsong ら [9] は，マルチビューのインタラクティブシステム MulUBA を提案した．この可視化システムは，広告アナリストが，消費者のオンラインショッピング行動と属性について分析することを複数の可視化を組み合わせることで支援している．一方でこれらの研究は，広告効果の指標として，クリック率やコンバージョン率は用いているものの，広告の離脱率に関しては言及していない．本研究においては，離脱率，クリック率，コンバージョン率の 3 つの指標から動画広告の視聴傾向を包括的に可視化することで，効果的な広告設定を発見することを目的としている．

2.2 広告効果向上を目的としたデータの活用

広告効果の向上を目指した研究は，2.1 節で述べたような可視化分析手法の研究以外にも，現在までに多くなされてきた．Kitada ら [10] の研究では，マルチタスク学習を用いた不均衡データに対する学習と conditional attention 機構によって，広告クリエイティブの効果を正しく評価する枠組みを提案した．石川ら [11] は，Yahoo!プロモーション広告（現 Yahoo!広告）への個人の広告出稿において得られる 10 種類の基本広告データの中で，広告主に定量的な指標を明確に与えるデータについて議論した．その結果，コンバージョン予測に大きく関与する指標の組み合わせを発見した．Eric ら [2] は，どのような広告が消費者を不快にさせるのかをクラスタ分析により発見した．これにより，印象の良い広告が集まっているクラスタには，シンプルでデザイン性に優れているものや消費者の興味に関連しているものが多いことがわかった．一方で，印象が悪いとされる広告が集まっているクラスタには，リスト記事や政治的なものがあったことを発見した．前述した通り，現在まで多様な観点から効果的な広告制作支援の研究が進められてきたが，それに対して本研究では，可視化分析のアプローチから広告効果の向上を目指す．

2.3 探索的データ分析を目的とした可視化ツール

探索的データ分析のための可視化ツールを開発した研究は多くある．Dongyun ら [6] は，政策データの分析を行う可視化ツールにおいて，データのフィルタリングやグループ化，比較などの操作をインタラクティブに実現させた．また，NetInf アルゴリズムを用いることで，政策データを可視化分析し，政策立案者などがデータの傾向やパターンを理解し，意思決定をサポートすることを可能とした．また Kui ら [5] は，複数のビューを組み合わせてテレビチャンネル・テレビ番組・視聴者の行動間の情報を表示することで，ユーザーが視聴率の変化を分析するのに役立つ可視化システムを開発し，視聴率に影響を与える要因を判断することを可能とした．Zana ら [12] は，交差表としての視覚化手法を提案しており，異なる階層カテゴリの組み合わせを直接比較できるようした．この可視化ツールを利用することで複雑なカテゴリ関係を探索し，分析することができる．このように，利用するデータに合わせた可視化ツールを開発することは，データから新たな示唆を得るためには重要である．本研究でも同様な考え方にもとづき，動画広告データに特化してこれを効果的に分析する可視化ツールの開発を進めている．

3 提 案 手 法

本研究の可視化システムを図 1 に示す．可視化システムはコントロールパネル部分と可視化部分の 2 つの要素で構成されている．また，可視化部分は主に 4 つの可視化画面を有している．ユーザーがコントロールパネル部分でデータを探索しながら，4 つの可視化を組み合わせて分析することで，1 章で述べた T1 から T3 の 3 つの可視化タスクを満たすように設計する．

3.1 使用データ

本研究の可視化システムのデータセットは，インターネット動画広告の配信結果データを想定している．具体的に本報告では，LINE 広告と Yahoo!広告で実際に配信された動画広告データと同様の形式のダミーデータで可視化した結果を報告する．これらのダミーデータが有するデータ項目の一部を表 1 と表 2 に示す．

表 1　量的データ

impressions	広告が表示された回数
clicks	広告がクリックされた回数
video_views	視聴回数
costs	配信費用

3.2 データの前処理

本研究では，動画広告の離脱率とクリック率，コンバージョン率の 3 つの指標を新たにデータとして適用した．本節では，

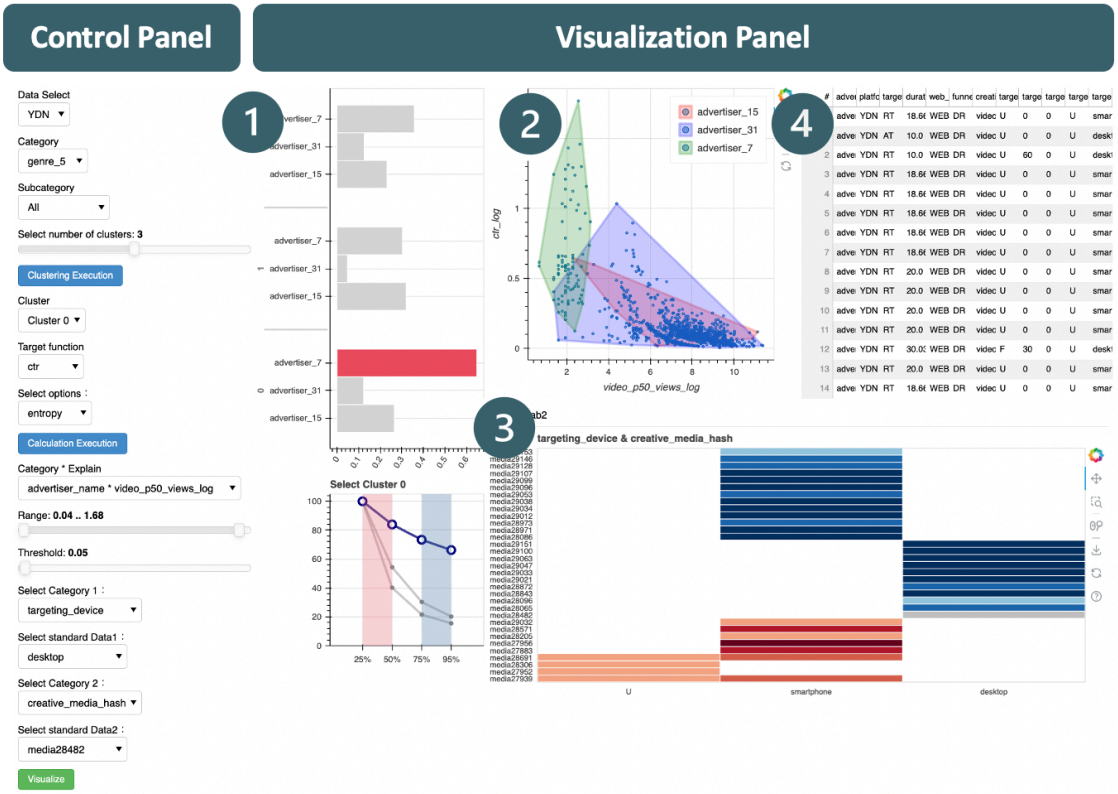


図 1　可視化画面の全体像

表 2　質的データ

genre	広告ジャンル
targeting_gender	ターゲット性別
advertiser_name	配信広告主名
targeting_age_min	広告のターゲット最小年齢

それぞれの指標について詳しく述べる．

3.2.1　動画広告の離脱率

動画広告の離脱率を用いて広告を評価する．動画広告の離脱率の傾向が似ているクラスタを生成しデータを分析するために，表 3 の 4 つの変数を用いて，以下に記載している動画広告離脱率に関する 6 つの変数 R1〜R6 を算出した．

表 3　広告離脱率の計算に利用する 4 つの数値

V_{p25}	動画広告の 25%時点まで再生された回数
V_{p50}	動画広告の 50%時点まで再生された回数
V_{p75}	動画広告の 75%時点まで再生された回数
V_{p95}	動画広告の 95%時点まで再生された回数

$$R_1 = \frac{V_{p25} - V_{p50}}{V_{p25}}, R_2 = \frac{V_{p25} - V_{p75}}{V_{p25}}, R_3 = \frac{V_{p25} - V_{p95}}{V_{p25}}$$
$$R_4 = \frac{V_{p50} - V_{p75}}{V_{p25}}, R_5 = \frac{V_{p50} - V_{p95}}{V_{p25}}, R_6 = \frac{V_{p75} - V_{p95}}{V_{p25}} \quad (1)$$

3.2.2　クリック率 (CTR) / コンバージョン率 (CVR)

一般的にインターネット広告は，広告の表示数に対するクリック数であるクリック率 (CTR) や，広告がクリックされた後に広告配信の目的が達成されたかを示すコンバージョン率 (CVR) で評価されることが多い．本研究でも可視化システムの目的変数として，クリック率 (CTR) とコンバージョン率 (CVR) を用いる．これらは，それぞれ以下の式で算出される．

$$CTR = clicks/impressions \times 100 \quad (2)$$
$$CVR = conversions/clicks \times 100 \quad (3)$$

3.3　クラスタリング

3.2.1 節で算出した動画広告の離脱率に関する 6 つの変数 R1〜R6 に対して k-means クラスタリングを適用する．その結果，動画の離脱傾向が似ている広告が同一クラスタに属するようになる．これにより，動画広告の離脱率の傾向が似ている広告データについて，それぞれどのような特徴が見られるかを観察することが可能となる．

3.4　可視化手法

本研究では，Python の可視化ライブラリである Bokeh を用いて可視化システムを開発した．実行環境として Python 3.9.6 および Bokeh 3.1.1 を利用した．

3.4.1　コントロールパネル部分

コントロールパネルを設置することで，ユーザーが自由にデータを探索できる．コントロールパネル部分は図 2 のように 7 つの要素で構成されている．以下に詳細を述べる．

（1）　可視化データ選択部分

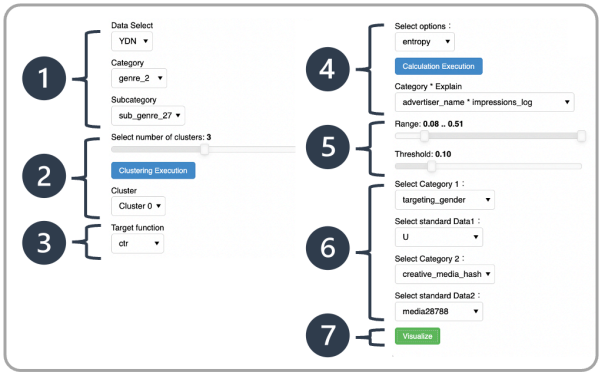


図 2　コントロールパネル部分

配信プラットフォーム・広告ジャンル・広告サブジャンルを選択することで，ジャンルごとの大まかな傾向を観察するだけでなく，サブジャンルごとに詳細にデータを分析することまでを可能にする．

（2） クラスタリング実行部分

図 2 の①でユーザーに選択されたデータの R1〜R6 に対して，クラスタリングを実行する．スライダー部分でユーザーが選択したクラスタ数にもとづいて計算を実行する．スライダーの数値 α は，$\{\alpha \in N \mid 1 \leqq \alpha \leqq 5\}$ の値をとる．具体的な計算手法は 3.3 節に述べたとおりである．

（3） 目的変数選択部分

3.2.2 節で述べたクリック率 (CTR) とコンバージョン率 (CVR) の 2 つのうち 1 つを可視化システムの目的変数として選択する．

（4） 説明変数・カテゴリ変数選択部分

本システムの可視化手法では，伊藤ら [13] の研究で提案された {correlation・skinny・clumpy・entropy} の 4 つの指標を用いて，見るべき価値のある説明変数とカテゴリ変数を見つけ出す．本手法で利用するデータは n 個の標本集合 A=$\{a_1, a_2, a_3, \cdots, a_{n-1}, a_n\}$ であり，t 番目の標本 a_t は，m_v 個の量的変数と m_c 個の質的変数を持つ，w 次元ベクトル $a_t = (a_{t1}, a_{t2}, a_{t3}, \cdots, a_{tw-1}, a_{tw})$ である． また，このデータから N 個の散布図 S=$\{s_1, s_2, s_3, ..., s_{N-1}, s_N\}$ が生成されるとする．そして，κ 番目の散布図 s_k の {correlation・skinny・clumpy・entropy} の 4 つの指標のスコアをそれぞれ $\{s_{k1}・s_{k2}・s_{k3}・s_{k4}\}$ とする．算出された 4 つの指標のスコアをソートした結果を，図 3 と図 4 に示すようにドロップダウンリストを用いてユーザーに提示する．4 つの指標の詳細を以下に示す．

- 相関 {correlation}

w 次元データのうち，i 番目と j 番目の次元を散布図の縦軸と横軸にそれぞれ配置したときの，i 番目と j 番目の次元間の Speaman 順位相関係数を算出する．算出式は以下のように表せる．

$$s_{k1} = |\mathcal{S}_{pear}(i,j)^2| \tag{4}$$

s_{k1} が大きいほうが正または負の相関が高くなるため，見るべき価値のある散布図だといえる．

- 細さ {skinny}

散布図上で点群が細長いとモデル化した際に誤差が小さい．そのため，点群が細長い散布図も見るべき価値があるといえる．そこで点群が集中する領域の細長さを指標とする．具体的には，点群を接続する Delaunay 三角メッシュを生成し，閾値以上の長い辺をもつ三角形を削除した上で，以下の値を算出する．本研究では，閾値を 0.1 とした．これは，Wilkinson ら [14] の実装に準じる．

$$s_{k2} = 1 - \frac{\sqrt{4\pi A_{rea}(T)}}{P_{erimeter}(T)} \tag{5}$$

ここで $A_{rea}(T)$ は三角メッシュT の総面積であり，$P_{erimeter}(T)$ の T は外周となる三角形辺の長さの合計値である．

- クラスタや例外点の分離性 {clumpy}

散布図におけるデータが有限個のクラスタに明確に分離する場合や，例外点が存在している場合も，見るべき価値のある散布図であるといえる．これを満たす散布図を生成する説明変数をユーザーに提示するために，分離性に関する Wilkinson ら [14] の Clumpy という指標を用いて，以下の値を算出する．

$$s_{k3} = max(1 - length(e_l)/length(e_{l+1})) \tag{6}$$

ここで前項と同様に，点群を連結する Delaunay 三角メッシュを生成し，メッシュを構成する三角形辺をその長さでソートする．辺の長さの長い順に l 番目と l+1 番目から，s_{k3} が最大となる値を算出する．s_{k3} の式において，それぞれ，e は Delaunay 三角メッシュの各エッジを，$length(e_l)$ はメッシュを構成する三角形辺の l 番目に長い長さを，$length(e_{l+1})$ は l+1 番目に長い長さを表す．

- クラスの分離性 {entropy}

カテゴリ変数をラベルとしたとき，ラベルごとに点群が明確に分かれている散布図もカテゴリ変数ごとの特徴を観察することができるため，見るべき価値のある散布図であると考えられる．本研究では情報エントロピーにもとづいてラベルの分離性を数値化する．具体的には，w 次元データのうち，i 番目と j 番目の次元から構成される散布図について，以下の値を算出する．

$$H(i,j) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{C} p(y_k = c|(a_{ki}, a_{kj})) \log p(y_k = c|(a_{ki}, a_{kj})) \tag{7}$$

ここで y_k は k 番目の標本に割り当てられたラベル，(a_{ki}, a_{kj}) は k 番目の標本の座標値，C はラベルの数である．散布図を L 個の部分領域に分割し，l 番目の部分領域におけるエントロピー $H(i,j)_l$ を上述の式により算出する．そして最終的に，以下の式により k 番目の散布図のスコアを算出する．ここで H_{max} は $\sum H(i,j)_l$ の最大値である．

$$s_{k4} = (H_{max} - \sum H(i,j)_l)/H_{max} \tag{8}$$

（5） 目的変数（CTR・CVR）の閾値変更部分

- Range スライダー

図 1 の可視化②でプロットされているデータのうち，データテーブルで表示する広告データを設定する． ここで選択された値の範囲内にある広告データのみを絞って表示する．

- Threshold スライダー

この部分で閾値を変更することによって，図 1 の可視化③の部分に表示するデータを絞ることができる．具体的には，閾値を上げることで基準値から大きく目的変数の値が高くなる，また，低くなるデータのみを可視化する．

（6） ヒートマップデータ選択部分

図 1 の可視化③の縦軸・横軸に用いる 2 つのターゲット属性の組み合わせと，可視化③において基準値にする広告データの 2 つのターゲット属性の組み合わせを選択する．

（7） 可視化実行部分

図 2 の①から⑥において，可視化画面の詳細設定を選択したのちに，visualize ボタンを押すことで可視化が実行される．

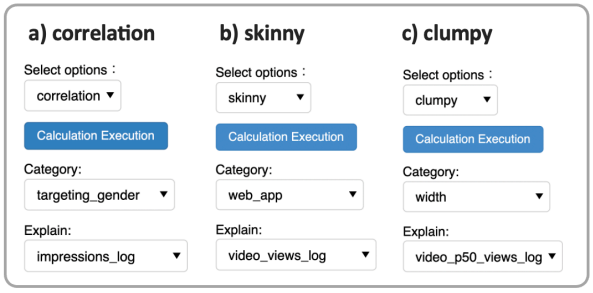


図 3 {corellation・skinny・clumpy} を選択した際のドロップダウンリスト

図 4 {entropy} を選択した際のドロップダウンリスト

3. 4. 2 可視化部分

図 1 が本研究の可視化の全体像である．図 1 中の①から④の 4 つの可視化の詳細を以下に述べる．

（1） 可視化①

3.3 節で述べた手法で，動画広告データにクラスタリングを適用した結果を，棒グラフと折れ線グラフで可視化する．動画広告の離脱傾向が似ている広告データが同一のクラスタに属している．このように，離脱率ごとにデータを絞り込んだ後にその他の可視化を実行することで，上述した可視化タスク T1 を達成している．

- 可視化①上部の棒グラフでの可視化

目的変数 (クリック率またはコンバージョン率) をクラスタごとに，かつ，コントロールパネル④の部分で選択されたカテゴリー変数ごとに可視化する．また，目的変数となるクリック率やコンバージョン率が，コントロールパネル①においてユーザーに選択されたジャンルとサブジャンルごとの平均値を上回っていれば，その広告をピンク色にハイライトする．この可視化によって，クラスタリングした結果のデータ分布を俯瞰し，クリック率やコンバージョン率が高い広告をカテゴリごとに発見することができる．

- 可視化①下部の折れ線グラフでの可視化

クラスタごとの離脱率を可視化する．コントロールパネルの②の部分で選択された広告クラスタの離脱傾向を，青色でハイライトして表示する．また，ユーザーによって選択されていないクラスタの離脱率をグレーで可視化する．これによって，クラスタごとの離脱傾向を把握することができる．また，ユーザーに選択されたクラスタに属する広告データの離脱率を（25%時点から 50%時点）・（50%から 75%時点）・（75%から 95%時点）の 3 つの領域に分け，3 つの領域のうち離脱率が最も高い部分には赤色の背景を，最も低い部分には青色の背景を設定している．

（2） 可視化②

可視化①でプロットされているクラスタのうち 1 つを選び，散布図でデータを可視化する．縦軸には目的変数（クリック率もしくはコンバージョン率）を，横軸には，コントロールパネル④の部分で選択された説明変数を設定している．中林ら [15] の研究では，点群の包括領域を表示することによって，複数のカテゴリを持つ散布図の視認性を上げる手法を用いている．本研究でもコントロールパネル④の部分で選択されたカテゴリ変数に対して，凸包を用いてそれぞれのカテゴリ変数の要素を色で囲う．これは，カテゴリ変数の分布を一目で把握することを目的としている．また，可視化に用いる説明変数とカテゴリ変数の選択肢は，3.4.1 節 (4) の部分で述べた手法で説明変数とカテゴリ変数をユーザー側に提示し，見る価値のある散布図を生成することをフォローする．これによって，可視化タスク T2 を達成する．

（3） 可視化③

選択されたクラスタに属している広告データのクリック率もしくはコンバージョン率をヒートマップで可視化する．ヒートマップの縦軸と横軸には，コントロールパネル⑥の部分で選択された 2 つのカテゴリー変数をそれぞれ設定している．また，ユーザーによって選択された広告属性の組み合わせを基準値とし，グレーの色で表示する．基準値よりも目的変数が低い属性の組み合わせには青色を，高い組み合わせには赤色が割り当てられている．また，コントロールパネル⑤の部分で目的変数の閾値を選択することで，表示するデータを絞り込むことができる．閾値の値を上げることによって，基準値から大きくクリック率またはコンバージョン率が上昇・下降する広告データのみを表示することができる．図 1 の可視化画面においては，縦軸が広告の種類で，横軸が広告のターゲットデバイスの属性であ

る．このように，多岐にわたる広告属性の組み合わせのうち 2 つを選択し，可視化することで属性の組み合わせごとにクリック率やコンバージョン率の大きく変動するパラメータを見つけることができ，可視化タスク T3 を達成させる．また，可視化③で用いるカラーマップを図 5 に示す．

図 5　ヒートマップで用いるカラーマップ

（4）　可視化④

可視化②にてプロットされたデータの詳細をデータテーブルで表示する．ユーザーが図 2 中のコントロールパネル⑤の部分の Range スライダーを用いて，目的変数の範囲を絞ることで，可視化②の目的変数が高い値のデータや低い値のデータのみをデータテーブルにおいて確認することができる．

上記のように可視化①から④を組み合わせることで，T1 から T3 の可視化タスクの達成をしている．

4　実行結果と考察

本章では，実行結果の一部と実行結果から得られる知見について述べる．具体例として，本報告では，Yahoo!広告で配信された動画広告データのダミーデータ 224,355 件と，LINE 広告で配信された動画広告データのダミーデータ 89,400 件を可視化分析した結果を紹介する．

4.1　Yahoo!広告で配信された動画広告の可視化結果

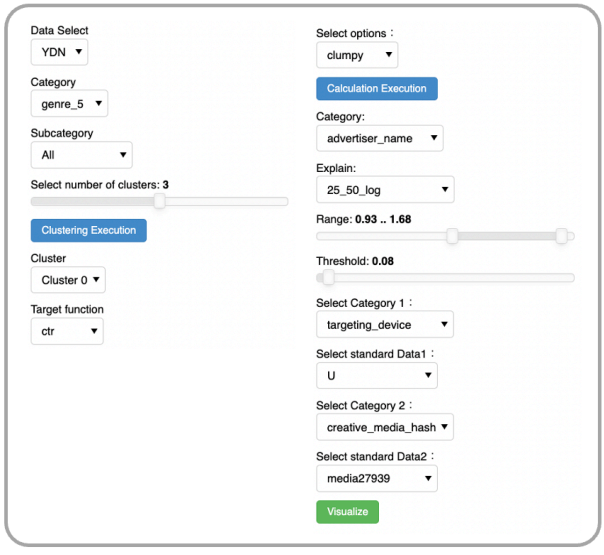


図 6　Yahoo!広告の可視化例におけるコントロールパネルの詳細設定

図 7 から図 9 は，Yahoo!広告で配信された動画広告データと同様の形式のダミーデータを，3 章で述べた手法で可視化した結果である．ここで，図 7 から図 9 の可視化におけるコントロールパネル部分の設定の詳細を図 6 に示す．

図 7 の左図の縦軸は，クラスタ番号とそれぞれのクラスタに属する広告主名を，横軸は，クリック率を表す．また，右図の縦軸は，動画広告 25%時点の再生回数を 100%としたときの再生回数の割合を，横軸は，動画広告の再生時点を表している．可視化結果からジャンル 5 の広告群の中で，選択されたクラスタ 0 は最も離脱傾向が緩やかな広告群で，比較的長く視聴され続ける傾向のある広告であることがわかる．また，このクラスタ 0 に属する広告主 7 の広告は，棒グラフにおいてハイライトされていることから，クリック率が平均値よりも高いことがわかる．この離脱率が低い広告群のデータの詳細を観察する．

図 8 は，クリック率を縦軸に，動画広告の 25%時点から 50%時点までの離脱率を横軸に割り当て，広告主ごとに色を割り当てた散布図である．この結果から，特徴が表れている広告主 7 と広告主 31 の広告に関して考察する．広告主 31 の広告は散布図の左下の方に点群が集中しているため，この 3 社の広告の中で最も動画広告の 25%時点から 50%時点までの離脱率が低い．しかし，クリック率も低い広告であることがわかる．一方で，広告主 7 の広告は動画広告の 25%時点から 50%時点までの離脱率が高いものの，クリック率はこの 3 社の中で最も高い傾向にある．よって，離脱率の高さはユーザーによるクリック率の高さに起因しており,「良い広告」であると推察することができる．

次に図 9 の可視化結果を観察してみる．この図は，広告の種類を縦軸に，広告のターゲットデバイスを横軸に割り当て，ヒートマップでクリック率を可視化した結果である．ヒートマップの基準値は図 9 の灰色で示された「広告種類：media27939」と「ターゲットデバイス：U(U：未設定)」の組み合わせの広告に設定している.「広告種類：media27939」の広告は「ターゲットデバイス：U」で配信するよりも「ターゲットデバイス：smartphone」の属性で配信した方がクリック率が高いことから，smartphone に向けた配信が効果的であることが可視化結果からわかる．また，ヒートマップの横軸に注目して観察してみると，離脱傾向が比較的緩やかな広告ジャンル 5 の広告においても，desktop に向けて配信されたものは，クリック率が全体的に低いことがわかる．一方で，ターゲットデバイスを smartphone に設定した広告の中でも，広告の種類によってクリック率に大きく差が出ている．具体的な事例を 1 つを挙げると,「広告種類：media27956」の広告はクリック率が高い一方で,「広告種類：media28086」の広告はクリック率が低い．

4.2　LINE 広告で配信された動画広告の可視化結果

図 11～図 13 は，LINE 広告で配信された動画広告データと同様の形式のダミーデータを，3 章で示した手法で可視化した結果である．ここで，図 11～図 13 における可視化のコントロールパネル部分の詳細設定を，図 10 に示す．図 11 の左図の縦軸は，クラスタ番号とそれぞれのクラスタに属する動画広告の縦のサイズの種類を，横軸は，クリック率を表す．また，右図の縦軸は，動画広告 25%時点の再生回数を 100%としたときの再生回数の割合を，横軸は，動画広告の再生時点を表してい

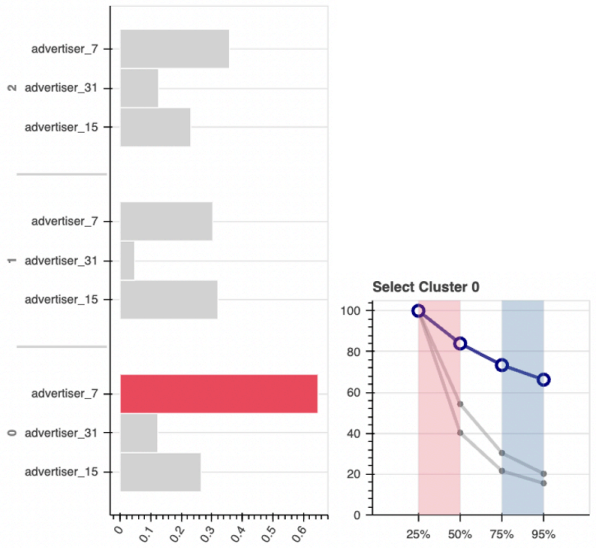


図 7　図 6 の条件における可視化①

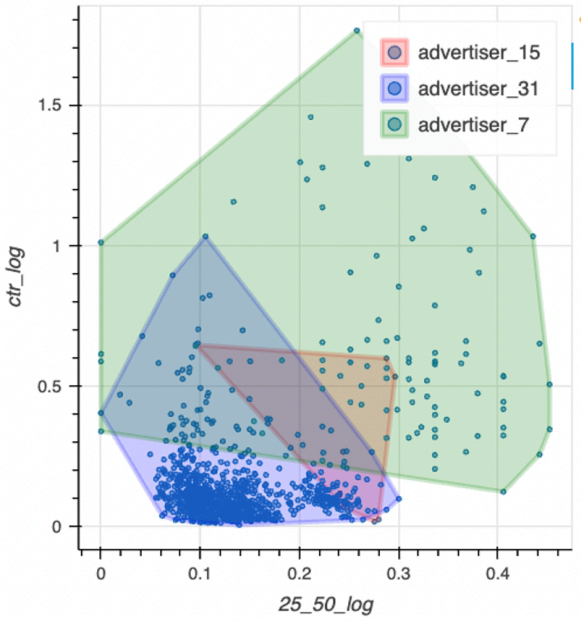


図 8　図 6 の条件における可視化②

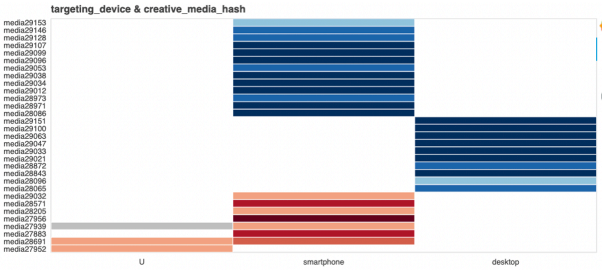


図 9　図 6 の条件における可視化③

る．図 11 の可視化①の左図において，クラスタ 2 に属する広告データはピンク色にハイライトされていることから，この広告ジャンルの中でもクリック率が高い広告が集まっているといえる．また，ユーザーに選択されたクラスタ 1 を観察すると，クラスタ 1 はクリック率が低い広告群であることがわかる．クラスタ 1 に関して，図 11 の可視化①の右図を観察してみると，クラスタ 1 は動画広告の 25%部分から 50%部分において大きく離脱が生じている．この結果からクラスタ 1 は広告の序盤で多くの視聴者が離脱し，最後まで見られないような広告群であることがわかる．以上をまとめると，クラスタ 1 はクリック率が低く，25%部分から 50%部分において大きく離脱が生じる広告であることから，広告クリックによるものではなく，単純な視聴離脱が起こった広告であると考えられる．

図 12 は，クリック率を縦軸に，動画広告の 25%時点から 50%時点までの離脱率を横軸に割り当て，動画広告の縦のサイズごとに色を割り当てた散布図である．この散布図で用いている説明変数とカテゴリ変数は，{entropy} でソートして提供された変数である．{entropy} の指標を選択すると，選択された目的変数（ここではクリック率）において，ラベルがよく分離するような説明変数とカテゴリ変数の組み合わせの順にソートされる．よって，この広告群においては，動画広告の 25%時点から 50%時点までの離脱率と動画広告の縦のサイズがクリック率に大きく効いていることがわかる．また，動画広告のサイズに関して観察してみると，縦のサイズが 720mm である広告群は，動画広告の 25%時点から 50%時点までの離脱率がこのクラスタの中だと低く，クリック率が高い傾向があることがわかる．一方で，縦のサイズが 1280mm である広告は動画広告の 25%時点から 50%時点までの離脱率が高く，クリック率が低い傾向にあることがわかる．

次に図 13 の可視化結果を観察する．この図は，広告の種類を縦軸に，広告のターゲット最小年齢を横軸に割り当て，ヒートマップでクリック率を可視化した結果である．ここで，ヒートマップの基準値は図 13 の灰色で示された「広告種類：media29187」と「ターゲット最小年齢：20 歳」の組み合わせの広告に設定している．可視化結果から「広告種類：media29187」の広告は，ターゲット最小年齢を設定しない場合よりも，ターゲット最小年齢を 20 歳に設定した方が，クリック率が高いことがわかる．また，ターゲット最小年齢を設定しない広告群よりも，ターゲット最小年齢を 20 歳に設定した広告群の方が全体的にクリック率が高いことがわかる．つまり，ジャンル 5 の広告の LINE プラットフォームにおける配信は，ターゲット最小年齢を 20 歳に設定した方がクリック率が高くなる可能性があることを示唆している．

このように，可視化①から可視化③を組み合わせて分析することで，離脱傾向を出発点としてどのような属性の広告がクリックされやすいのか・広告離脱が起こりやすいのかを発見することができる．

5　結論・今後の課題

本研究では，1 章で示した T1 から T3 の 3 つの可視化タスクを定め，これらのタスクを達成するようなインタラクティブなインターネット動画広告データの可視化ツールを開発した．実行結果で示したように，本可視化システムを用いることで，クリック率やコンバージョン率に大きな傾向が現れる広告デー

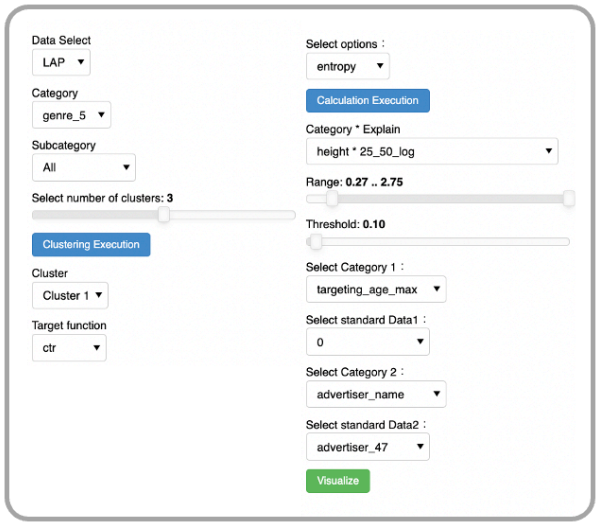


図 10　LINE 広告の可視化例におけるコントロールパネルの詳細設定

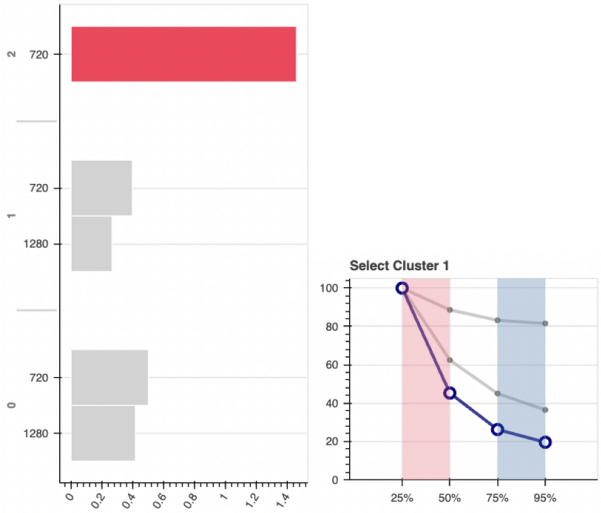


図 11　図 10 の条件における可視化①

タを発見することができ，効果的な動画広告データの配信設定を発見することができる．

今後の課題を 3 つ挙げる．1 つ目は，可視化④部分の改善である．現状の実装では，可視化②部分で表示されたデータの詳細を可視化④のデータテーブルで確認する操作手順を前提としている．データテーブルはデータの詳細を把握できるメリットがある一方で，瞬発的な視認性が低いというデメリットがある．今後の課題として，散布図部分においてデータを選択することで，選択されたデータの詳細を一目で把握できるような可視化を導入したいと考えている．2 つ目は，動画広告の 25%時点までの離脱が大きなデータを見つけやすくすることがあげられる．動画広告は 25%時点で離脱されるか否かが最も重要と言われている．現時点の可視化システムでは，25%時点，50%時点，75%時点，95%時点での離脱に違いを設けていないため，どの時点で離脱されたかに対して重みをつけるような可視化分析を実現したい．3 つ目は，時系列を考慮した動画広告データ分析の実現である．インターネット広告では，広告ジャンルごとに

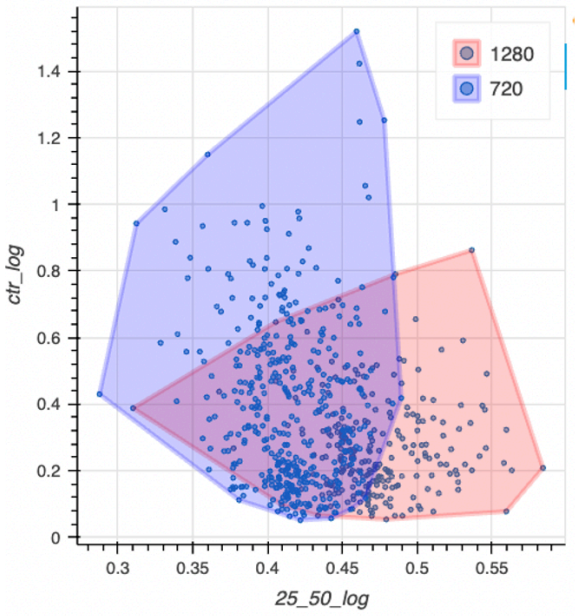


図 12　図 10 の条件における可視化②

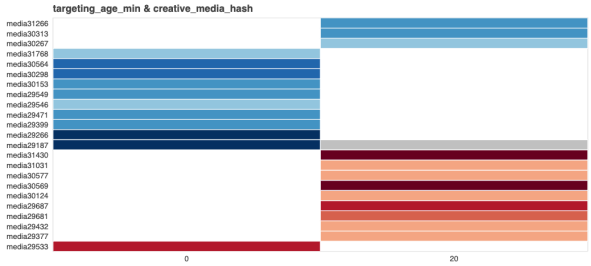


図 13　図 10 の条件における可視化③

クリック率やコンバージョン率に特徴が現れる時期や時間帯が変わると想定できる．しかし，現状の実装では可視化に時系列データを用いていないため，これらを考慮した分析ができない．よって，可視化システムにおいて，時系列の観点からもクリック率やコンバージョン率に対する考察ができるように実装を改良したい．

文　献

[1] 電通：「2022 年 日本の広告費 インターネット広告媒体費 詳細分析」(2023), https://www.dentsu.co.jp/news/item-cms/2023007-0314-2.pdf. (Accessed：2023-12-21).
[2] Eric Zeng, Tadayoshi Kohno, Franziska Roesner, “What makes a “Bad” Ad? User Perceptions of Problematic Online Advertising”, CHI Conference on Human Factors in Computing Systems, pages 1-24, 2021.
[3] Shon Minjung, Lee Daeho, “Perceived Consumer Discomfort of Online Video Advertising: Analysis of Characteristics by Consumer Cluster”, Journal of Scientific and Industrial Research, pages 528-536, 2021.
[4] 松本有加, 堂前力真, 大槻明, “x-means クラスタリング及び主成分分析を用いた消費者の特徴分析”, 情報知識学会誌, pages 200-207, 2018.
[5] Kui Xiaoyan, Lv Huihao, Tang Zhengliang, Zhou Haowen, Yang Wang, Li Jinqiu, Guo Jialin, Xia Jiazhi, “Tvseer: A visual analytics system for television ratings”, Visual Informatics, vol.4, no.3, pages 1–11, 2020.

[6] Dongyun Han, Abdullah-Al-Raihan Nayeem, Jason Windett, and Isaac Cho1, “PDViz: A Visual Analytics Approach for State Policy Data”, ICOMPUTER GRAPHICS forum Volume 42, pages 305–318, 2023.
[7] Rika Miura, Hayato Ohya, Takayuki Itoh, “Visualization System to Analyze Browsing Trends of Internet Video Advertisements”, 27th International Conference on Information Visualisation (IV2023), pages 1-6, 2023.
[8] Di Peng, Wei Tian, Min Zhu, Yukun Ren, Xiaojian Lin, Mingzhao Li, “TargetingVis: visual exploration and analysis of targeted advertising data”, Journal of Visualization, vol.23(6), pages 1113-1127, 2020.
[9] Shangsong Liu, Di Peng, Haotian Zhu, Xiaolin Wen, Xinyi Zhang, Zhenghao Zhou, Min Zhu, “MulUBA: multi-level visual analytics of user behaviors for improving online shopping advertising”, Journal of Visualization, vol.24(6), pages 1287-1301, 2021.
[10] Shunsuke Kitada, Hitoshi Iyatomi, Yoshifumi Seki, “Conversion Prediction Using Multi-task Conditional Attention Networks to Support the Creation of Effective Ad Creatives”, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’19, pages 2069–2077, 2019.
[11] 石川善一郎, 奥牧人, 河野崇, “Web リスティング広告における基本広告データを用いたコンバージョン分析と予測”, DEIM Forum, 2017.
[12] Zana Vosough, Marius Hografer, Loic A. Royer, Rainer Groh, Hans-Jorg Schulz, “Parallel hierarchies: Avisualization for cross-tabulating hierarchical categories”, Computers & Graphics, vol. 76, pages 1–17, 2018.
[13] Takayuki Itoh, Asuka Nakabayashi, Mariko Hagita, “Multidimensional Data Visualization Applying a Variety-Oriented Scatterplot Selection Technique”, Journal of Visualization, Vol. 26, No. 1, pages 199-210, 2023.
[14] Leland Wilkinson, Anushka Anand, Robert L. Grossman, “Graph-theoretic scagnostics”, IEEE Symposium on Information Visualization, pages 157–164, 2005.
[15] Asuka Nakabayashi, Takayuki Itoh, “Applications and Evaluations of Drawing Scatterplots as Polygons and Outlier Points”, Integrating Artificial Intelligence and Visualization for Visual Knowledge Discover, pages 631-651, 2022.

CLIPを用いたゲーム動画フレーム毎のバグ検知AI

髙橋　秀太朗† 服部　峻†† 砂山　渡††

† 滋賀県立大学大学院 工学研究科 電子システム工学専攻　〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学 先端工学研究院　〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †on23stakahashi@ec.usp.ac.jp, ††{hattori.s,sunayama.w}@e.usp.ac.jp

あらまし　ゲーム開発において，バグを見つける時間・コストが増大しており，これらを下げるべく人間のデバッガーの代わりにバグを検知可能なバグ検知 AI を構築する必要がある．そこで本稿では，ゲーム中のオブジェクトが物理法則に反するバグに限定し，この種のバグが発生しているフレーム画像をいくつか含むプレイ動画に対して，それらのシーケンス番号を精度良く網羅的に出力することが可能なバグ検知 AI を試作し，その性能を検証する．具体的には，入力プレイ動画をフレーム毎の画像に分け，バグを表現した入力テキストとともに，OpenAI の CLIP を用いて同じ次元のベクトル空間上に変換し，テキストとのコサイン類似度が高かったフレーム画像を「バグ有り」と検知する．

キーワード　画像認識，ゲーム情報学，バグ検知 AI，CLIP

1 はじめに

近年，ゲーム開発の規模が年々拡大してきており，2020 年には日本国内のゲームコンテンツ市場が 2 兆円を上回り，過去最高の規模となっている．2021 年はほぼ横ばいの状態で，高い市場規模を維持し，2022 年は前年比 1.4 %増と一定の成長を見せた [1,2]．それに伴って，バグ検知に掛かる時間・コストが大きな問題となっている．

従来のバグ検知方法の基本的な手順として，ゲームに長けた人間のデバッガーやテスターがプレイヤーの視点に立ってプレイを行い，その中でゲームに含まれていた不具合などのバグを発見するというものである．ゲーム開発のバグ検知に掛かる時間，コストを下げるために，人間のテスターの代わりにバグを判別可能なバグ検知 AI を構築する必要があると考える．しかしながら，異常検知の広範な研究や，ソフトウェアのバグ検出 [3] やバグ予測 [4] は盛んに行われて来たが，ゲームの，特にプレイ画面に特化したバグ検知に関する研究は未だ少ない．

そこで本稿では，バグをオブジェクトが物理法則に反するバグに限定して，この種のバグが発生しているフレーム画像をいくつか含むプレイ動画に対して，それらのシーケンス番号を精度良く網羅的に出力することが可能なバグ検知 AI を OpenAI の CLIP を用いて試作し，その性能を検証する．

以降，2 章では本稿の関連研究を紹介する．次に 3 章では，バグ検知 AI の要件分析を行う．4 章では，本稿の提案手法を示す．5 章では，評価実験を行って，提案手法の性能を評価する．最後に 6 章でまとめと今後の課題を述べる．

2 関連研究

本章では，関連研究について述べる．また，本稿は 2.1 節で紹介する Mohammad Reza Taesiri らの “CLIP meets Game-Physics: Towards bug identification” [5] に基づいて研究を行っている．

2.1 バグ動画検索 AI に関する研究

動画のデータセットの中から関連動画を検索する研究 [5] がある．この研究では，OpenAI の CLIP を用いて任意の英語のテキストを入力することで，大規模なゲームプレイ動画のデータセットから関連する動画を検索するシステムを提案している．評価実験の一つに，データセット内で発生している，バグを表すテキストを入力し，そのバグを含む動画がデータセットから検索できるかを検証している．

本稿では，この研究の，ゲームのバグを表した入力テキストと動画・画像の関係に着目して，バグ検知 AI が作成できるのではないかと考えた．また，この研究では成し遂げていない，より局所的なバグが発生しているフレームのシーケンス番号を出力することができるバグ検知 AI を作成し，その精度を評価する．

2.2 大規模言語モデルを用いたバグ検知に関する研究

大規模言語モデルのゼロショット機能を，バグ検知に活用できる可能性を探る研究 [6] がある．この研究では，同じゲーム動画内のバグである現象を表すテキスト 1 つとバグでない現象を表すテキストを 3 つ用意し，大規模言語モデルを使用して，4 つの内どのテキストがバグだと言えるかを判定させている．

本稿の類似点として，ゲームを題材にしている点，バグ検知を試みている点が挙げられる．一方で，本研究では，ゲーム画面の現象をテキストに表してからバグ検知を行うのではなく，ゲーム画面の動画をそのままバグ検知に利用することとしている．

3 バグ検知 AI の検討

本章では，本稿の対象である “物理法則に反するバグ” だけでなく，汎用的なバグ検知 AI を構築するための手法を検討する．

まず，新しく作られたゲームにバグが含まれてしまう理由として，以下が挙げられる．

図 1　車が宙に浮いているバグが発生している様子

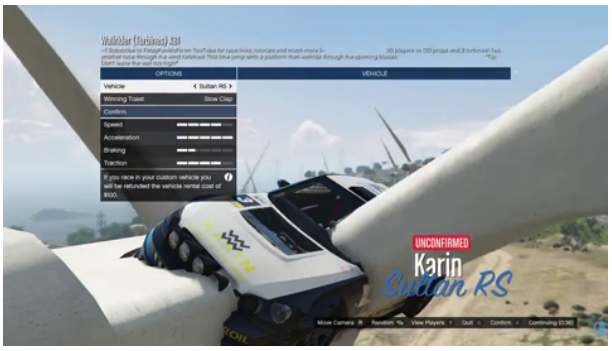


図 2　車がめり込むバグが発生している様子

- 要求仕様書が明確でない
- 人間の手によるバグ発見が不十分である
- 開発に掛ける人員やコストが不十分である
- リリース後のアップデートにバグ修正を頼る

3.1　手法の検討

本研究の最終的な目標は，汎用的なデバッグ AI の構築である．そのための手法として，

- ルールベース
- 機械学習

以上 2 つの手法が考えられる．

ルールベースを用いた手法は，人間のデバッガーやテスターなどによって既知となったバグの一つ一つをあらかじめルールとして記述しておくことで，バグが発生した際にどのバグが発生したのか検出することができる手法である．テトリスを題材として，その有用性を検証した [7, 8]．

一方で，機械学習を用いた手法として，動画をフレーム画像に分け，画像ごとに画像分類モデルを適用させる手法や，ラベル付けされたバグを含む動画と含まない動画を何度も学習させる，教師あり学習の手法も考えられる．

3.2　バグの分類

本研究では，現実世界に近い 3D ゲームに限定し，ゲーム画面内に発生する表示バグを扱い，仕様書に書かれていないことを前提とすると，バグを以下のように分類する．ここで，「仕様書に書かれていないこと」を前提としたのは，例として「車が宙に浮いているバグ」を挙げると，ゲームによっては車が浮くという現象はバグではなく，仕様であるとされるゲームもあるため，バグかどうかを区別するために前提としている．

- オブジェクトの見た目についてのバグ
- オブジェクトの振る舞いについてのバグ
- オブジェクト間の相互作用についてのバグ

上記のバグ分類における，「オブジェクトの見た目についてのバグ」には，オブジェクトの見た目が変わったり，突然消えたりするバグが含まれる．「オブジェクトの振る舞いについてのバグ」には，本来想定されていないような，オブジェクトが宙に浮かんだり，オブジェクトが逆さまになったりするバグが含まれる．「オブジェクト間の相互作用についてのバグ」には，本来想定されていないような，オブジェクト同士が重なったり，めり込んだりするバグが含まれる．以上の 3 つのバグの内，「オブジェクトの振る舞いについてのバグ」と「オブジェクト間の相互作用についてのバグ」に対して，画像言語特徴抽出モデルである CLIP が有効ではないかと考えられた．一方で，「オブジェクトの見た目についてのバグ」に対しても，CLIP が多少は有効であると考えられるものの，見た目の変化や消滅は少なくとも連続する 2 枚の静止画の差分が必要となるため，CLIP では対応できない．CLIP の詳細な説明は 4.1 節で行う．

本研究の検知対象となるバグは図 1，2 のような “物理法則に反するバグ” としている．これは，使用したモデルが CLIP であり，CLIP は物理法則に従う現実世界の画像とその説明テキストを元に，大量に学習しているためである．また，現実世界に近いゲームには精度が期待できるが，独自のルールに基づく仮想世界のゲームに対しては期待できないと考えられる．

3.3　プレイ動画分析

発見したいバグによって，プレイ動画の分析方法が異なる．

- 1 枚の静止画ごとに分析すれば良いバグ

「車が宙に浮いているバグ」などは，プレイ動画の 1 枚の静止画であるフレーム毎にバグ判別することで解決できると考える．

- 局所的な数枚の静止画を分析する必要があるバグ

仮に「車の色が急に黒に変化したバグ」が存在したとすると，1 枚の静止画だけではバグ判別できず，少なくとも連続する 2 枚の静止画の差分が必要になる．

4　提案手法

本章では，前章の検討を踏まえて，バグ検知 AI を画像分類モデルによって試作する．

本稿で試作するバグ検知 AI は，2.1 節で紹介した Mohammad Reza Taesiri らの “CLIP meets GamePhysics: Towards bug identification” [5] に基づいて作成しており，本稿では，最終的な出力部分で，バグが発生している可能性が高いフレーム画像のシーケンス番号を出力できるようにシステムを改良している．

提案手法の全体のシステム図を図 3 に示す．入力としてゲーム動画とバグを表すテキストを受け付ける．入力動画はフレーム画像に分けて処理を行う．OpenAI の CLIP [9] を介して Embedding された後，フレーム画像毎に類似度を計算する．こ

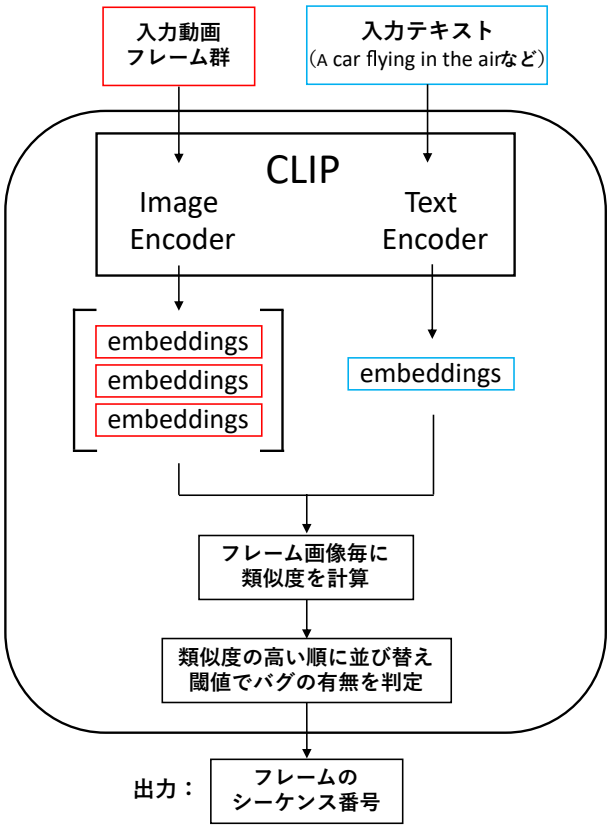


図 3　システムの概略

の類似度を高い順に並び替え，予め決めておいた閾値でバグの有無を判定する．最後にバグが有りの際，該当するフレーム画像のシーケンス番号を出力する．尚，参考にした元論文 [5] では，フレーム画像毎に類似度を計算した後に，類似度が上位のフレーム画像と対応するゲーム動画を特定して，最終的に入力テキストに対する動画検索結果を出力し，その精度を検証している．

以下から CLIP の説明の後，システムの構成毎に更に詳しく紹介し，最後に画像を用いて具体例を示す．

4.1　CLIP

システムの前処理に当たる部分に CLIP [9] が使われている．この CLIP とは，画像言語特徴抽出モデルである．

CLIP の特徴として，学習データの多さとゼロショット転移が可能な点が挙げられる．CLIP の学習には，インターネットを利用して構築された 4 億組のデータセットが使われている．また，タスクごとに追加の学習を必要とせず，事前学習データで学習したモデルを直接タスクに転用することができるというメリットがある．

CLIP のエンコーダーは画像とテキストをそれぞれ入力とし，それぞれ埋め込みベクトル表現に変換する．これらのベクトルは同じ次元を持ち，同じ高次元特徴空間上に存在するため，画像のベクトルとテキストのベクトルの類似度を比較することができる．

4.2　CLIP によるエンコード

入力としてゲーム動画とバグを表すテキストを受け付ける．ここで，バグを表すテキストとは，「A car flying in the air」のような英文で表すことができるテキストである．入力した動画を CLIP に渡す前に，フレーム毎の画像に分ける必要がある．これは，CLIP が画像に対して Embedding を行うことができるため，画像の状態で入力する必要があるからである．

CLIP のエンコーダーを介して出力された Embedding データは，画像，テキスト共に同じベクトル空間で表現される．

4.3　類似度の計算

画像とテキストの 2 つのベクトルは，コサイン類似度によって計算される．画像を表す埋め込みベクトルと，テキストを表す埋め込みベクトルの距離が近いほど類似度が高いと言える．より網羅的な類似度の探索を行うために，Faiss の近傍探索ライブラリを用いている．

4.4　バグの有無の判定

フレーム毎に類似度を計算した後，それらを類似度が高い順に並び替える．これは，類似度が高いほどフレーム画像とバグを表すテキストとの関係性が高いことを表しており，テキストに示されたバグが発生している可能性が高い．

また，閾値を定めておくことで，テキストに示されたバグの有無が動画ごとに判別できるようにしている．しかし，現段階では最適な閾値を定めることはまだできていない．閾値がないため，バグかどうかの判断は出力された上位のフレーム画像を人間が見て判断する必要がある．

そして，類似度が高いフレーム画像のシーケンス番号を出力することで，動画のどのフレームにバグが存在している可能性が高いかを知ることができる．

4.5　処理の具体例

システムの入力から出力までの一連の具体例を示す．ここでは，車が宙に浮いているバグが発生している動画を扱う．尚，例で示すゲーム動画は Mohammad Reza Taesiri らの “CLIP meets GamePhysics: Towards bug identification” [5] のデータセットから使用している．

まず，用意したゲーム動画をフレーム毎の画像に分ける．例として，フレーム画像の内 3 枚を図 4，5，6 に示す．ここで，図のフレーム数は全 351 フレームの内それぞれ 15，270，327 である．また，入力テキストは「A car flying in the air」とする．

次に，入力動画フレームと入力テキストを CLIP を介して Embedding する．求められた図 4 のフレーム画像とテキスト，それぞれに対するベクトルの一部を以下に示す．

フレーム No.15 のベクトル：
[[1.66258574e-01 -2.57626548e-02 2.77469516e-01
-9.20731351e-02 -1.49414331e-01 -3.12886298e-01
4.58234549e-01 -3.34012061e-02 6.47625506e-01
6.23272657e-01 2.04471573e-01 8.29054117e-02

図 4　フレーム No.15 の様子

図 5　フレーム No.270 の様子

図 6　フレーム No.327 の様子

・・・
-2.01157890e-02 5.04817605e-01 -1.94994867e-01
-1.95492625e-01 -8.32808614e-01 -7.87560865e-02
1.62453912e-02 -2.68335730e-01 1.47313401e-01
8.18762958e-01 -2.56921679e-01 1.09947845e-01]]

テキストのベクトル：
[[4.63719293e-02 1.16745662e-02 -2.62305010e-02
-2.18800902e-02 4.54418873e-03 4.79435176e-02
2.20424822e-03 -6.41536638e-02 -9.06289089e-03
3.62320691e-02 2.50516995e-03 -7.90150277e-03
・・・
-1.49018252e-02 3.47883701e-02 -2.78250836e-02
2.89495755e-02 -4.19641874e-04 -2.99086943e-02
7.64050893e-03 1.37810716e-02 -1.27613004e-02
-2.83300206e-02 1.00666629e-02 -4.76861112e-02]]

図 7　評価実験 I，II の Video1 でバグが発生している様子

図 8　評価実験 I の Video2 でバグが発生している様子

これらのベクトル表現を元に，コサイン類似度を計算する．図 4，5，6 それぞれの類似度の値は，0.25996774，0.31938526，0.2761459 となった．

よって，図 4，5，6 を類似度が高い順に並び替えると，図 5，6，4 となり，3 枚のフレーム画像では最も図 5 に入力テキスト「A car flying in the air」のバグが発生している可能性が高いと分かる．但し，最適な閾値を設定できておらず，各フレーム画像にバグを含むかどうかの自動判別までは実現できていない．

5　評 価 実 験

本章では，前章で提案したバグ検知 AI に，どのようなバグを表すテキストを入力すれば効果が期待できるかを検証する実験 I と，ある特定のバグを表すテキストが，対応するバグ動画に加え無関係なバグ動画に対して，どの程度出力が異なるかを検証する実験 II を行った．ここで，評価実験 I，II で用いたゲーム動画，計 11 本は Mohammad Reza Taesiri らの “CLIP meets GamePhysics: Towards bug identification” [5] のデータセットから使用している．また，評価を行うために必要である，各フレーム画像のバグかどうかのラベル付けは第 1 著者自らが行っている．評価実験 I で用いた Video1 では全 351 フレーム中，計 279 フレームで車が宙に浮いているバグが発生しており，Video2 では全 210 フレーム中，計 57 フレームで車が空中で縦一回転するバグが発生している．評価実験 II で用いた動画については表 1 に示す．但し，評価実験 I の Video1 と評価実験 II の Video1’ のみ同じ動画を用いている．

表 1　評価実験 II で用いた動画毎のフレーム数・バグの種類

VideoX	全フレーム数	バグ発生フレーム数	どんなバグを含むか
1’	351	279	車が宙に浮いているバグ
2’	222	60	人がオブジェクトにめり込むバグ
3’	450	368	自転車が飛んでくるバグ
4’	210	69	車の下に人が入るバグ
5’	123	42	車同士が重なるバグ
6’	402	70	人が逆さまになるバグ
7’	900	444	飛行機が地面に突き刺さるバグ
8’	451	120	自転車が空を飛ぶバグ
9’	682	314	車が直立するバグ
10’	367	239	車が風車にめり込むバグ

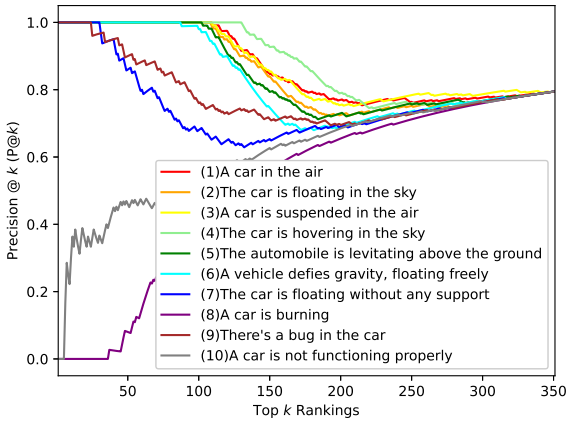


図 9　実験 I，Video1 での Precision @ k

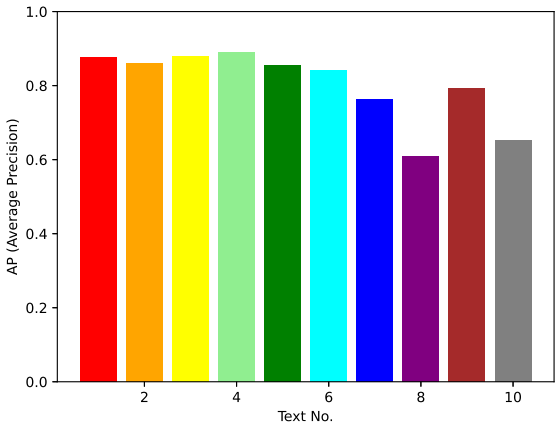


図 10　実験 I，Video1 での AP

5.1　実　験　I

5.1.1　実 験 方 法

Step 1.　バグを含む動画 2 本とバグを表すテキスト 10 個を用意する

Step 2.　それぞれのテキストに対して，提案手法を用いて，1 つの動画毎にフレーム画像との類似度を計算し，高い順に並べる

Step 3.　それぞれの結果に対して，Precision @ k（上位 k 番目までの Precision）と AP（Average Precision）を計算し，比較する

5.1.2　使用した動画とテキスト

使用した 2 本の動画（Video1, Video2）でのバグが起きている様子を図 7，8 に示す．また，使用した 10 個のテキストを以下に示す．テキストを決めるにあたり，車が宙に浮いている現象を様々なテキストで表すだけでなく，(8) の様に全く異なる現象を示すテキストを表したり，(9),(10) のように “バグがある” や “適切に動いていない” などの現象についての説明を直接的には行っていない抽象的なテキストを表したりもした．

（1）　A car in the air

（2）　The car is floating in the sky

（3）　A car is suspended in the air

（4）　The car is hovering in the sky

（5）　The automobile is levitating above the ground

（6）　A vehicle defies gravity, floating freely

（7）　The car is floating without any support

（8）　A car is burning

（9）　There’s a bug in the car

（10）　A car is not functioning properly

5.1.3　Video1 に対する結果

図 9，10 は図 7 のようなバグが全 351 フレーム中，計 279 フレーム発生している動画に対する結果である．Precision @ k の値が左から大きな値を保ち，AP の値が大きいと，検知対象のバグを表現するテキストとしてより最適であると読み取ることができる．まず，(8)「A car is burning」に対しては動画と無関係なテキストなため，Precision @ k の左側が小さい値となっていることは妥当と言える．(9)「There’s a bug in the car」や (10)「A car is not functioning properly」に対しては，他のテキストの結果と比べ早い段階で Precision @ k の値が減少している．よって，バグを表す抽象的なテキストは，具体的に示されたテキストに比べて精度が低くなると分かる．これは，3.4 節や 4.1 節でも述べたように，提案手法に用いている

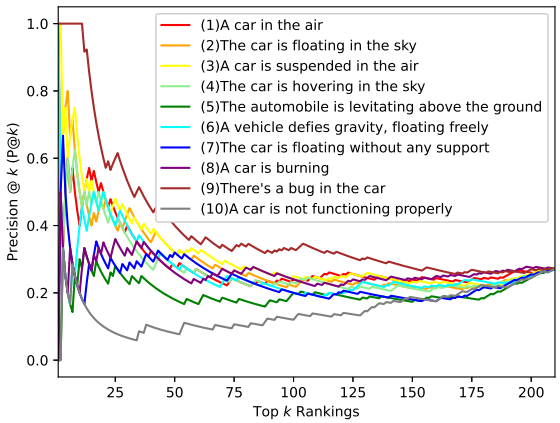


図 11　実験 I，Video2 での Precision @ k

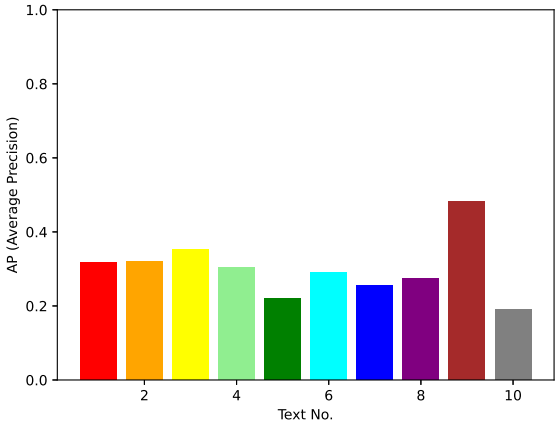


図 12　実験 I，Video2 での AP

CLIP が，事前学習に現実世界の画像を用いているため，「bug」や「not functioning properly」などの抽象性が高い文には精度を期待できないと考えられる．更に，具体的に表現されたテキスト文の中でも，(4)「The car is hovering in the sky」がこの中では最適なテキスト文であったことが分かる．他の文と比較して，“hover” という単語が，Video1 のバグを表すのに最も適していたと考えられる．また，Video1 の結果が Video2 より全体的に高く出力されたのは，全体フレームに対するバグを含むフレームの数が Video2 より多かったため，Precision が大きくなったと考えられる．

5.1.4　Video2 に対する結果

図 11，12 は図 8 のようなバグが全 210 フレーム中，計 57 フレーム発生している動画に対する結果である．Video1 と比べ，(8)「A car is burning」や (10)「A car is not functioning properly」は同様に Precision @ k の値が低くなっているので，妥当であると言える．一方で，(9)「There’s a bug in the car」の Precision が他のテキストや Video1 と比べ大きくなった．バグを含む動画によっては，車の状態を表すテキストよりも，「bug」という単語を直接与えた方が精度は良くなると分かる．しかし，これは CLIP が「虫」という意味で認識し，黒い車が虫の様に見えたという可能性がある．更に，これよりも Precision が高く出るテキストが存在する可能性もある．これは，今後検討していきたいと考える．

5.2　実　験　II

5.2.1　実 験 方 法

Step 1.　バグを含む動画 10 本とそれぞれに対応するバグを表すテキスト 1 個ずつ，計 10 個を用意する

Step 2.　動画 10 本に対して，提案手法を用いて，1 つのテキストとの類似度を計算し，それぞれ高い順に並べる

Step 3.　Step 2. を用意した計 10 個すべてで行う

Step 4.　それぞれの結果に対して，Precision @ k（上位 k 番目までの Precision）を計算し，比較する

図 13　評価実験 II の Video7’ でバグが発生している様子

5.2.2　使用した動画とテキスト

使用した 10 本の動画（Video1’〜10’）の中で，Video1’,7’ でバグが起きている様子を図 7，13 に示す．また，10 個の動画に対応したテキスト（Text1’〜10’）計 10 個を以下に示す．

ここで，動画 1 本に対して 1 個のテキストを決めるにあたり，動画 1 本に対して 10 個のバグを表すテキストを用意し，評価実験 I で用いたシステムを使って，平均精度 (AP) が最も高く出力されたテキスト 1 個を用いている．

Text1’.　The car is hovering in the sky

Text2’.　The man is entwined with the object

Text3’.　A bike is in mid-air, making its way here

Text4’.　Someone is stuck beneath the wheels of the car

Text5’.　The cars are stacked on top of each other

Text6’.　The person is hanging upside down

Text7’.　An aircraft plunges into the ground

Text8’.　A cyclist is airborne, flying through the sky

Text9’.　A car is in a vertical position

Text10’.　A car becomes embedded in the wind turbine

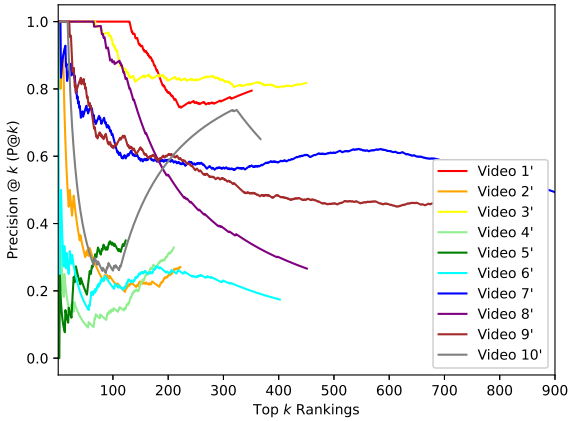


図 14　Text1’ での Precision @ k

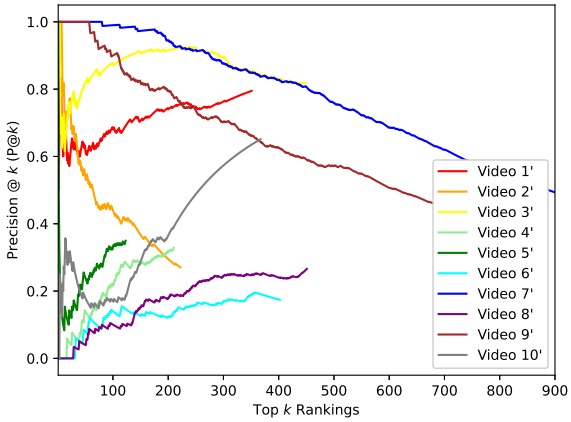


図 15　Text2’ での Precision @ k

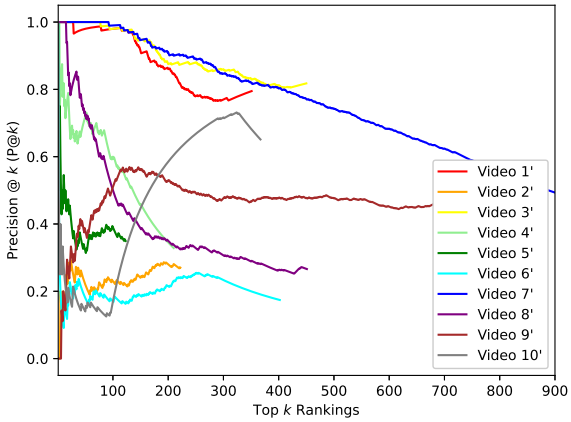


図 16　Text3’ での Precision @ k

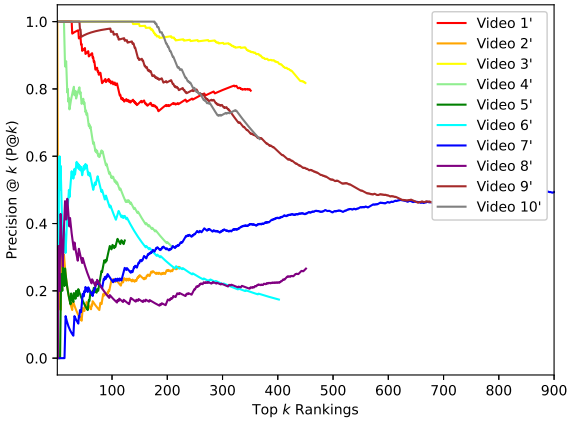


図 17　Text4’ での Precision @ k

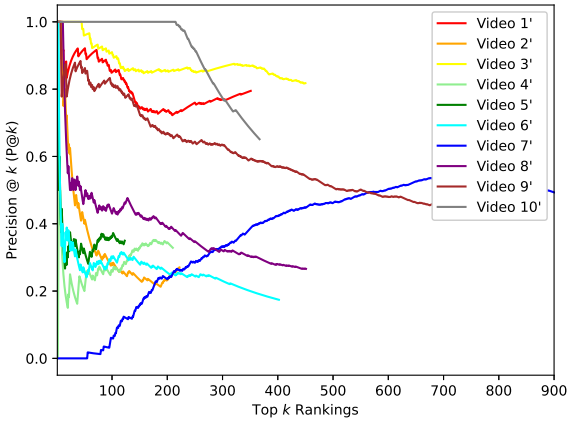


図 18　Text5’ での Precision @ k

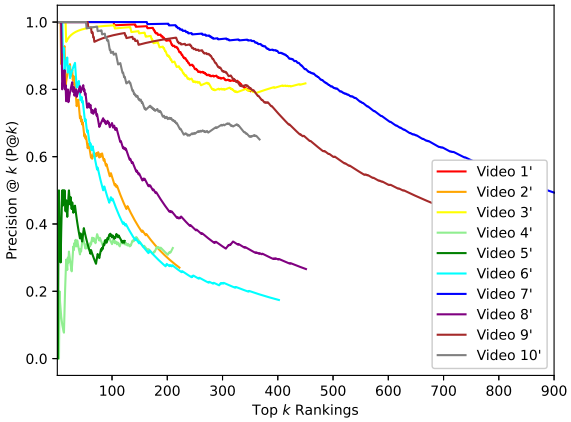


図 19　Text6’ での Precision @ k

5. 2. 3　実 験 結 果

Video1’〜10’ のそれぞれに対応するテキスト（Text1’〜10’）毎の Video1’〜10’ に対する Precision @ k の結果を図 14〜23 に示す．青線の Video7’ の Precision @ k が高いグラフが多いのは，総フレーム数が 900 とかなり多く，その中で半分近くものフレームでバグが発生しているためである．図 14，16，20，22，23 では，それぞれのテキストと対応する動画の Precision @ k が高く出力されており，有効であると思われるが，同じグラフ内で他の動画も Precision @ k が高く見られた．例えば，図 14 では，Video1’ を説明する Text1’ が入力されているため，

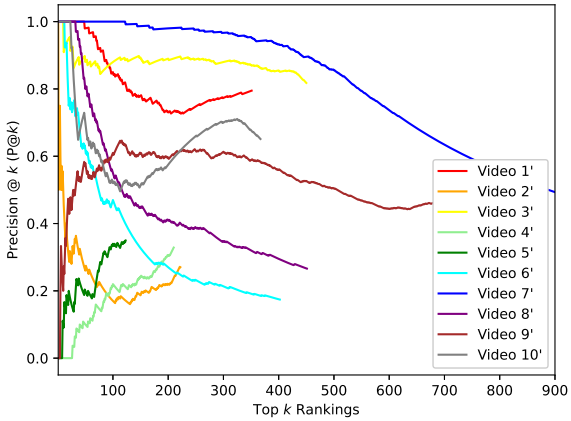


図 20　Text7’ での Precision @ k

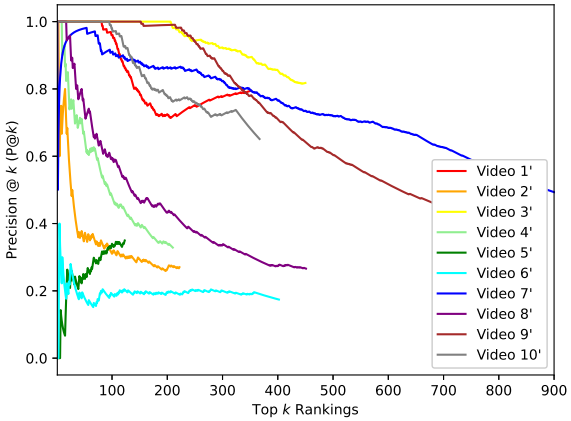


図 22　Text9’ での Precision @ k

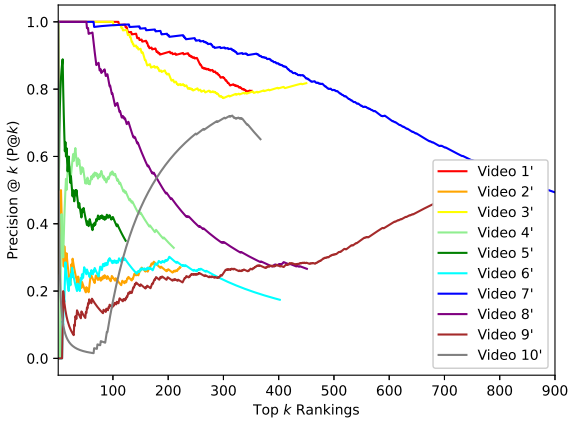


図 21　Text8’ での Precision @ k

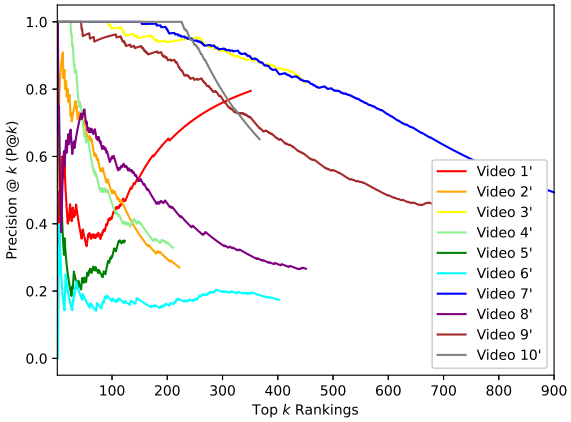


図 23　Text10’ での Precision @ k

Video1’ の Precision @ k が高くなっている一方で，Video3’ も高くなっている．Video3’ は「自転車が飛んでくるバグ」であるため，Text1’ の car や hover との関連性が低いにも関わらず，Precision @ k が高く出力された．考えられる理由として，Video3’ は全 450 フレーム中，368 フレームにバグが発生しているため，高い値を出力しやすいことがある．

テキストと対応する動画の Precision @ k が低く出力された図 15，17，18，19，21 は，総フレーム数が少ないためグラフでは低く見えたと考えられる．やはり，複数の動画を比較するには，総フレーム数とバグを含む割合が近い動画を使用しなければ結果を評価しにくいが，これに値するバグ動画を収集するには困難を極める．

6　おわりに

本稿では，関連研究の Mohammad Reza Taesiri らの “CLIP meets GamePhysics: Towards bug identification” [5] を参考に，CLIP を用いたバグ検知 AI の作成と評価を行った．評価実験 I は最適な入力テキストを探すものとなっており，詳細にバグが起きている現象を説明するテキストならば精度が期待できる結果となった一方で，「bug」という単語を含むテキストも Precision が高くなった．評価実験 II は複数の動画を入力させた際に，入力したテキストと対応する動画がその他の動画とどのように精度が異なるかを検証するものとなっており，動画によっては Precision @ k が高い状態を維持していたが，総フレーム数やバグ割合によっては Precision @ k が低く出てしまう結果となった．よって，CLIP でのバグ検知が有効かどうかは，バグ動画に依ってしまうこととなった．

今後の研究課題として，バグかどうかを判断するための閾値の最適化を検討する．また，本稿の提案手法ではフレーム毎の画像とテキストの類似度の結果をそのままバグ判別に用いているが，バグは数フレーム間で発生する可能性が高いという特性を活かし，フレームの前後の類似度を参照する方法も考えられる．更に，本稿ではバグを表すテキストを動画を見てから決定しているが，これを仕様書や「車」「動物」などの名詞などからバグを表すテキストを自動生成できないか検討していきたいと考えている．

加えて，評価実験 II で用いた Video1’，Video5’，Video6’，Video8’ でのそれぞれに対応するバグのテキストとの類似度の推移を図 24，25，26，27 に表す．図 25，26 のように，同じよ

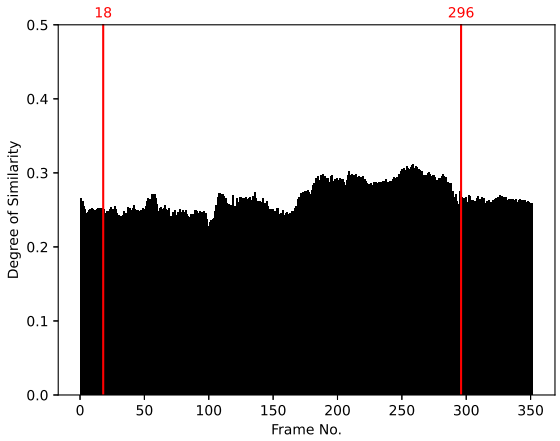


図 24　Video1’ での類似度推移（フレーム No.18〜296 においてバグが発生）

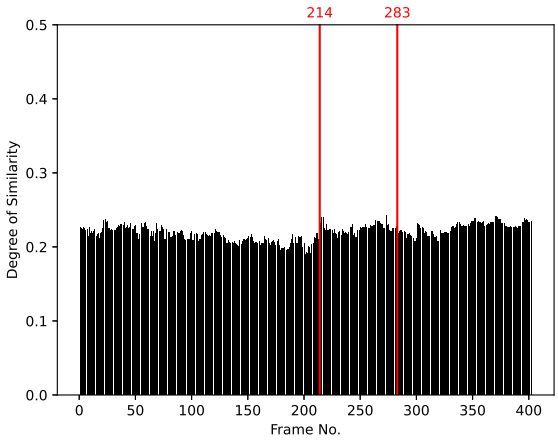


図 26　Video6’ での類似度推移（フレーム No.214〜283 においてバグが発生）

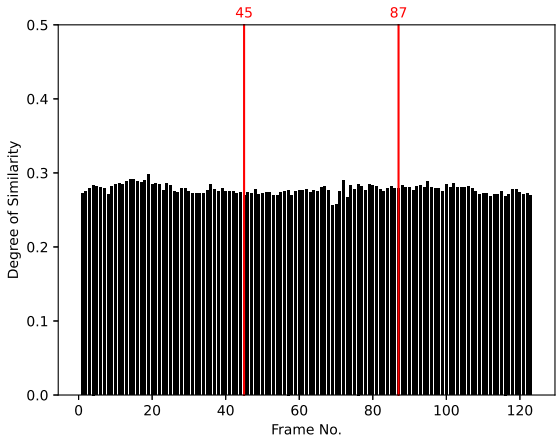


図 25　Video5’ での類似度推移（フレーム No.45〜87 においてバグが発生）

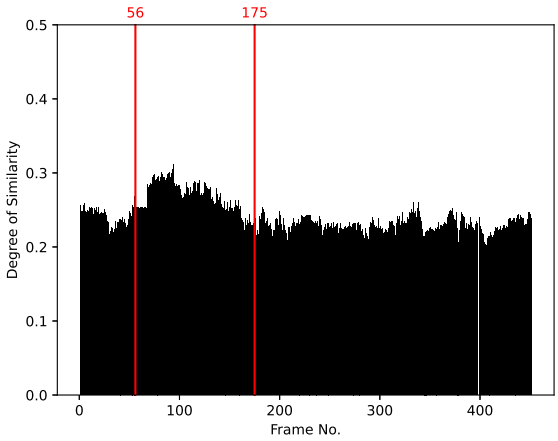


図 27　Video8’ での類似度推移（フレーム No.56〜175 においてバグが発生）

うな画面が続く動画に対しては類似度の推移は見られないが，図 24，27 のように，動画によっては大きく類似度が変化するものも存在する．この時，山ができている箇所がバグの可能性が高いと考えることができ，このような類似度推移のグラフもバグ検知 AI によるバグ検知や，人間によるバグ発見支援に有効ではないかと考えている．これについては今後，詳細に検討していく予定である．

文　　献

[1] GameBusiness.jp, “2022 年の世界ゲームコンテンツ市場規模は 26 兆 8,005 億円―ゲーム業界データ年鑑「ファミ通ゲーム白書 2023」8 月 29 日発売,” https://www.gamebusiness.jp/article/2023/08/29/22141.html (参照 2023-12-25).

[2] 角川アスキー総合研究所, “ファミ通ゲーム白書 2023,” (参照 2023-12-25).

[3] 峯久 朋也, 阿萬 裕久, 川原 稔, “Transformer によるメソッド名推定を活用したネーミングバグの検出,” コンピュータソフトウェア, vol. 39, no. 4, pp. 17–23 2022.

[4] 荻野 翔, 肥後 芳樹, 楠本 真二, “現実的な設定におけるメソッド粒度バグ予測モデルの構築および精度評価,” 情報処理学会論文誌, vol. 63, no. 4, pp. 973–985, 2022.

[5] Mohammad Reza Taesiri, Finlay Macklon, Cor-Paul Bezemer, “CLIP meets GamePhysics: Towards bug identification in gameplay videos using zero-shot transfer learning,” MSR, pp. 270–281, 2022.

[6] Mohammad Reza Taesiri, Finlay Macklon, Yihe Wang, Hengshuo Shen, Cor-Paul Bezemer, “Large Language Models are Pretty Good Zero-Shot Video Game Bug Detectors,” arXiv, 2022.

[7] 髙橋 秀太朗, 服部 峻, 高原 まどか, “テトリスのためのルールベースなゲーム画面認識によるデバッグ AI の試作,” 日本デジタルゲーム学会 (DiGRA) 2022 年夏季研究発表大会 予稿集, #4-3, pp. 45–48, 2022.

[8] 髙橋 秀太朗, 服部 峻, 砂山 渡, “テトリスのための Inpainting を用いたルールベースなデバッグ AI,” 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2022.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” Proceedings of the 38th International Conference on Machine Learning (PMLR), pp. 8748–8763, 2021.

DEIM Forum 2024 T4-B-4-05

数枚のサンプル画像を用いて類似場面を検出するオンデマンド場面検出技術

秋島　遥† 吉田　登†† 馬場崎康敬†† 劉　健全††

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2–1–1
†† NEC ビジュアルインテリジェンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †g1920502@is.ocha.ac.jp, ††{y_babazaki,n-yoshida14,jqliu}@nec.com

あらまし　本論文では，1 つのクエリ画像から目的とする人と物の関係性を指定するだけで，その他の映像の中から類似の関係性を持つ場面を即時検出することができる場面検出システムを紹介する．本技術は，人単体の姿勢と，人と物の位置関係をそれぞれ画角に依存せずに記述する特徴量を用いた学習レスな検出技術と，学習に基づいて人と物の関係性を推定する関係性推定技術の両方を使いこなすものであり，学習データに依存しない幅広い場面への適用が可能である．実験では，関係性検出の公開データセット（V-COCO）を用いて，学習データに含まれない未知の場面に対する検出精度を評価し，本手法の有効性を示した．

キーワード　画像認識，映像解析，行動解析，検索モデル，インタフェース・インタラクション

1　はじめに

近年，防犯カメラがいたるところに設置され，膨大な画像，動画データが日々蓄積されている．その用途は，公共施設での不審者の検出，追跡，助けを必要とする人の検出，小売店での顧客の動線解析，工場での工員管理，介護施設での見守りなど多岐にわたる．しかし，人の手によって膨大なデータの中から特定の場面を検出することは非常に困難である．コンピュータによって画像を解析し，対象人物が写る場面を見つける技術が期待される．本研究では，画像から人とモノの関係性に着目した検出を行う技術を扱う．

まず，人物の類似場面検出では，見つけたい場面の適切な言語化が難しいという問題がある．例えば「介助を必要とする人」を検出したい場合，具体的にどんな人物を検出したいのか明確に言語化できない場合がある．しかし，こういう人を探したい，と数枚の画像を準備することは容易である．車いすに乗っている人や，杖をついている人などの画像を手に入れればよい．そこで，本研究では，サンプル画像をクエリとした類似場面検出技術を採用した (図 1)．

類似場面検出の技術として，まず機械学習があげられる [6]．事前に検出したい場面がわかっていれば，学習を行うことで画像から特定の場面を認識することができる．しかし，新たに検出したい場面が増えた場合，すぐに対応することが難しい場合がある．機械学習で新しい場面を認識するには，その場面を学習するためのデータ収集やアノテーション，モデルの作成などの準備が必要であり，実際の運用までには時間がかかる．実運用では，検出したい場面を決定してから，実際に検出をするまでの期間を短くすることが重要である．

次に，人の姿勢の類似性による行動検出手法が挙げられる [15], [16]．この技術は，少量のサンプル画像を用いて，行動についての事前学習をすることなく類似行動を検出可能である．しかし，姿勢が似ていても，関係する物体によっては場面の意味合いが変わってくることが多い．例えば「コップを持っている人」と，「スマートフォンを持っている人」では，同一の行動だが全く異なる場面を表していると考えられる．そこで，本研究では姿勢情報に加えて，関係する物体の情報も加味することで検出精度向上を図った．

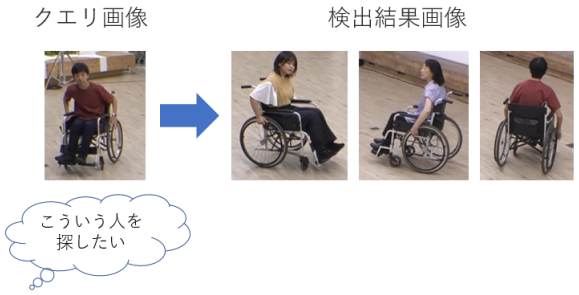


図 1:　画像をクエリとした検出イメージ

類似場面検出の難しさとして，同じ場面でも姿勢や人と物体の距離にバリエーションがあることが挙げられる．姿勢や関連物体の情報を用いた検出では，検出したい場面のクエリに類似する場面を局所的に検出ができるが，異なるバリエーションまで含めた大域的な検出をすることは難しい．その点機械学習では，学習データを整備することで，姿勢や物体との距離に関わらず様々なバリエーションの場面を広く検出できる．しかし広範囲な検出が可能である一方，異なる場面とのクラス境界において誤検出のリスクも存在する．そこで，姿勢や人と物体との距離をさらに組み合わせて用いて，よりクエリの場面に似ている場面の類似度を引き上げることで，精度の向上が期待できると考えた．また本手法を用いることで，もし機械学習モデルが

未学習の場面を検出する場合も，姿勢や物体のクラス，物体との距離などの特徴のみを用いて類似場面を検出することができる．

以上を踏まえ，本研究では，人物と物体の関係性に着目した類似場面検出手法として，関係性ラベルの事前学習を必要としない検出技術と，機械学習に基づく検出手法を組み合わせ，双方の技術の利点を活かした場面検出アルゴリズムを提案する．検出の流れとしては，検出したい行動が写っている画像数枚をクエリ画像とし，人の姿勢，人と関連する物体との距離，機械学習で推定された人と物の関係性という 3 種類の情報を組み合わせた類似度を用いて類似場面を検出する．本研究では，以下のことを実現した．

- 高精度な類似場面検出技術として，事前準備が必要ない検出手法を提案.
- 検出手法と，機械学習の知見を組み合わせることによる，より高精度な場面検出技術を提案．機械学習の誤検知，未検知による精度劣化の抑制を実現.

2 関 連 研 究

2.1 物 体 検 出

物体検出は，画像認識の分野で，画像から特定の物体の位置とその物体のクラスを検出する技術である [5], [7]．検出された物体の位置は，外接矩形により表される．本手法では，Faster R-CNN [8], [17] を用いて人と物体を検出し，類似場面検出に用いる．また，検出で得られた物体クラスは，物体の一致を判断するために使用する．

2.2 類似姿勢検索

近年，さまざまな姿勢推定技術が提案され [18]，行動認識 [1] にも応用されている．姿勢情報とは，画面における人物の関節を 18 点のキーポイントの座標情報として示し，人の外観や照明条件や背景などに対して不変な性質を持つ．その情報を使って，ユーザが指定した姿勢情報をクエリとし，類似した姿勢の人物を映像中から検索する手法が提案されている [10]．本手法では，姿勢情報を類似度算出のために使い，検出に応用する．

2.3 物体との位置関係による行動認識

人の行動にはほとんどの場合物体が関係し，その関係性に着目した行動認識技術が提案されている．人と物体との位置関係についてグラフ構造で表し，画像の説明文を作成する研究 [2], [3] が存在する．また，3D モーションにおいて，人と物体を Interaction Bisector Surface (IBS) と呼ばれる技術で面によって区切り，物体がどこの体の部位に関係しているかを数値化して類似行動を検索する手法 [21] が存在する．このことから，人の行動の認識には物体との関係が大きな役割を持つことがわかる．本手法では，人と物体との距離に着目した類似度を検出に用いる．

2.4 HOI 検出

Human-Object Interaction(HOI) 検出は，人と物体に着目した検出において強力な手助けとなる．HOI 検出とは，入力画像から，$< \mathrm{human}, \mathrm{object}, \mathrm{interaction} >$ を検出する技術であり，検出の精度を上げるため，多くの研究が進められている [4], [12], [22]．本手法では，この関係性検出によって得られた人と物体のインタラクションを類似度を算出するために使い，検出に応用する．

2.5 人物行動検索

人物行動検索について，映像クエリのキャプションから同じ行動が写っている映像を検索する手法 [19] や，自然言語のクエリで検索する手法 [11] が存在する．しかし，人と物に着目したクエリ画像により類似場面検出を行っている例はなく，本研究は新しい検出手法であると言える．

2.6 Zero-Shot Learning (ZSL) for HOI Recognition

Zero-Shot 学習は，学習データがなくても新しいクラスやタスクに対応できる手法として知られている．画像分野では，画像データとして学習データがない場合でも，与えられたクラスやタスクの説明に基づいて画像にうつる場面の予測と分類を行うことができる [14], [20]．しかし，Zero-Shot 学習は言語情報に依存しており，クラスやタスクの説明が提供された場合にのみ機能するため，完全に未知のクラス，すなわち学習時に説明情報が提供されていないクラスには対処できない．本研究では，行動を言語情報として扱うことをせず，画像を数枚もらって，それをクエリとして類似場面検出をすることで，学習データがない未知のクラスにも対応することを目指す．

3 提 案 手 法

本研究では，物体検知，姿勢推定，関係性検出の技術を組み合わせた検出アルゴリズムを提案する．人の行動は，対象となる物体に関連して特徴づけられることが多い．そこでクエリは，画像に写っている人と物体のペア (以下人モノペア) とする．人とモノの関係性から，類似している場面を検出することが目的であり，こういう人を探したい，という曖昧な要望でも，実際にクエリとなる画像を数枚用意してもらうことで類似場面を探すことが可能である．

3.1 類 似 度

姿勢の類似度 s_{pose}，人と物体との距離の類似度 s_{ho}，人物の行動の類似度 s_{ml} の 3 種類の類似度を組み合わせてクエリの人モノペアと，画像データベースに保存されている人モノペアとの類似度を算出する..

人モノペア a と人モノペア b の類似度 $S_{a,b}$ は以下のように表される．

$$S_{a,b} = w_{\mathrm{pose}} s^{a,b}_{\mathrm{pose}} + w_{\mathrm{ho}} s^{a,b}_{\mathrm{ho}} + w_{\mathrm{ml}} s^{a,b}_{\mathrm{ml}} \qquad (1)$$

w_{pose}，w_{ho}，w_{ml} は，それぞれ類似度の強さを変化させる正のパラメタである．クエリ画像によってこの重みを変化させることで，検出の精度が上がることが見込まれる．

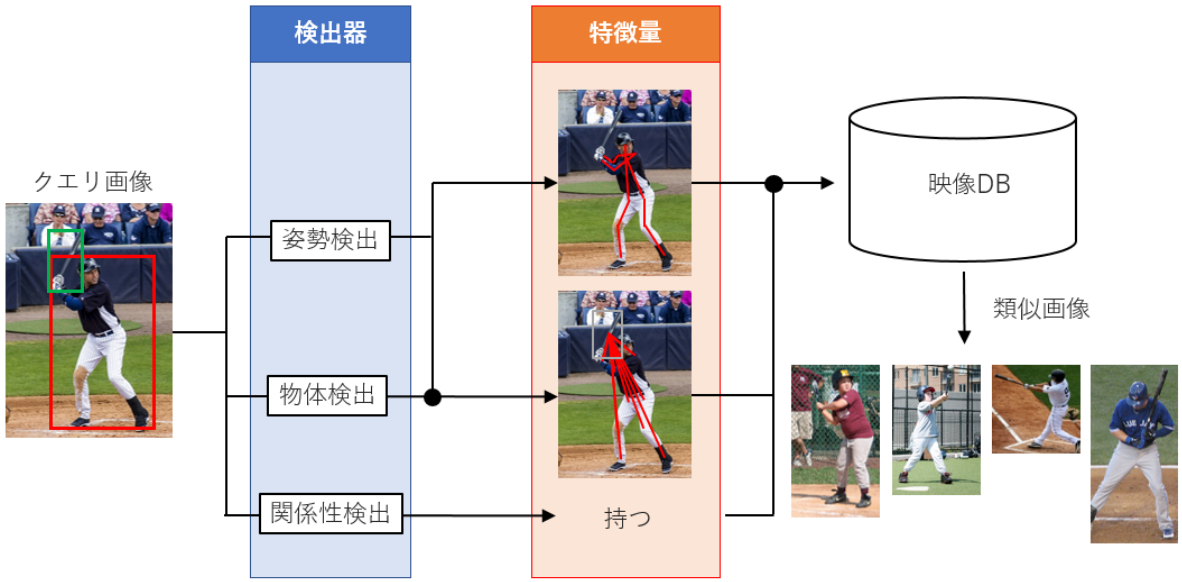


図 2: 検出の流れ. クエリとして，人とモノのペアを受け取る．検出器によって得られた姿勢，物体のクラス，関係性（行動ラベル）の情報から特徴量を算出し，画像データベース内の人とモノのペアと類似度を計算することで類似場面を得る．

3.1.1 姿勢の類似度

人の行動は姿勢に関連付けられることが多く，実際に行動認識に用いられている．姿勢は人物の外観 (服装や持ち物) や撮影環境 (照明，背景) に依らないことが強みであるが，姿勢推定結果の欠損値への対応や，カメラの焦点や設置位置により，人の向き，サイズが変わるという課題がある．そこで，姿勢の特徴量として先行研究 [23] にある画角不変な姿勢特徴量を用いた．以下の式で表される．

$$f_i = \frac{pose_y_i - core_y}{p_height} \tag{2}$$

$pose_y_i$ は 2 次元姿勢情報のうち i 番目のキーポイントの y 成分であり，$core_y$ は基準点となる首の y 成分，そして p_height は推定された画像内における人物の身長をそれぞれ示している．i 番目のキーポイント情報が欠損した場合は，f_i は -1 となる．キーポイントの x 成分は固定値で，y 成分の値により特徴量が決まる．これを身長で割ることで，正規化する．この特徴量を用いて，姿勢の類似度は以下のように表される．

$$s^{a,b}_{\text{pose}} = \frac{1}{1 + \sum_{i=1}^{n} l_i |f_{a,i} - f_{b,i}| / \sum_{i=1}^{n} l_i} \tag{3}$$

$$l_i = \begin{cases} 1, & \text{if } f_{a,i} \neq -1 \text{ and } f_{b,i} \neq -1 \\ 0, & \text{if } f_{a,i} = -1 \text{ and } f_{b,i} = -1 \end{cases} \tag{4}$$

類似度の式 (3) では，欠損値以外の値から類似度を求めている．

3.1.2 人と物体との距離の類似度

次に，人とモノの関係性を考えるのに，物体との位置関係を使う．人の行動には，物と関連付けられるものが多く，特定の体の部位と物体の距離は有用な情報である．例えば， ボールをける，という動作では，けり足とボールの位置関係が重要であり，コップを持つ，という動作では持っている手とコップの位置関係が重要であると考えられる．また，距離はカメラの設置位置に対して，人の身長の値による正規化が可能である．人と物体との距離の特徴量 p_i は以下のように表される．

$$p_i = \frac{||\boldsymbol{k}_i - \boldsymbol{b}||}{p_height} \tag{5}$$

$\boldsymbol{k}_i$ は i 番目のキーポイントの座標，$\boldsymbol{b}$ はペアの物体の中心座標である．p_height は，姿勢の類似度のところで推定された身長であり，身長で割ることで正規化をしている．姿勢推定で検出されたキーポイント全てについて物体の中心点との距離を考えることで，どこの関節が物体により関係しているのか考慮することができる．

この特徴量を用いて，人と物体との距離の類似度は以下のように表される．

$$s^{a,b}_{\text{ho}} = \frac{1}{1 + \sum_{i=1}^{n} l_i |p_{a,i} - p_{b,i}| / \sum_{i=1}^{n} l_i} \tag{6}$$

$$l_i = \begin{cases} 1, & \text{if } p_{a,i} \neq -1 \text{ and } p_{b,i} \neq -1 \\ 0, & \text{if } p_{a,i} = -1 \text{ and } p_{b,i} = -1 \end{cases} \tag{7}$$

姿勢の類似度と同じように，欠損値以外の値から類似度を求めている．

この姿勢の類似度と，人と物体との距離の類似度による検出手法を**非関係性学習検出手法**と呼ぶ．

3.1.3 関係性の類似度

人物の行動推定には，機械学習の手法を用いる．本研究では，人ーモノ関係性抽出に特化した汎用的行動認識技術を使用する．画像を入力として人の動詞 (行動) を検出し，推定された動詞とその推定スコアに基づいて類似度を求める．

$$s^{a,b}_{\text{ml}} = \max_i \left(\frac{v_{a,i} v_{b,i}}{2} \right) \tag{8}$$

推定される動詞が M 個あるとすると，$\boldsymbol{v}$ は M 次元のベクトルであり，それぞれの動詞の推定スコア (0–1) が入っている．推定度が高いほど 1 に近い値となる．そこで，類似度の値には，クエリの人モノペアと，データベースの人モノペアの推定スコアの平均値が一番高い動詞の値を用いる．

クエリに基づく関係性検出モデル使い分け

一般的な行動認識技術では，人の行動を学習，推定するのに画像中に映る全ての物体を参照することが多い．しかし，関係性ラベルごとに，人と物体の距離には傾向があると考えられる．例えば，*hold*，*work_on_computer* などの行動ラベルがついている人モノペアでは，関連物体が近くにあることが多く (図 3)，*point*，*look* などの行動ラベルがついている人モノペアでは，関連物体が遠くにあることが多い (図 4) 傾向がある．

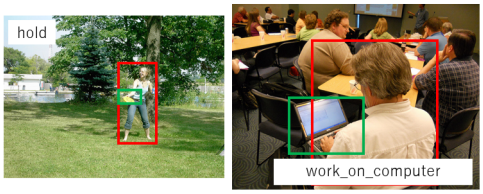


図 3: 人と物体が近い傾向がある行動ラベル例．hold(左)，work_on_computer(右)．

図 4: 人と物体が遠い傾向がある行動ラベル例．point(左)，look(右)．

このことから，人と物体の距離に合わせて適切なモデルを選択することで，検出精度改善が可能であると考えた．そこで，クエリとなる人モノペアにおいて人と物体が近い場合，人と近い物体を参照する機械学習モデルを利用し，人と物体が遠い場合，人と遠い物体を参照する機械学習モデルを利用する実験を行った．

実験において，関係性学習のモデルとして以下の 3 つのモデルを作成した．

- 人との $IoU > 0$ の物体だけ学習・推論に用いたモデル
- 人との $IoU = 0$ の物体だけ学習・推論に用いたモデル
- 人との $IoU \geq 0$ の物体 (すべての物体) を学習・推論に用いたモデル

検出に使用する類似度計算において，手法 1 として，クエリの人モノペアの人と物体が $IoU > 0$ である場合，「人との $IoU > 0$ の物体だけ学習・推論に用いたモデル」を使い，クエリの人モノペアの人と物体が $IoU = 0$ である場合，「人との $IoU = 0$ の物体だけ学習・推論に用いたモデル」を使うものと，

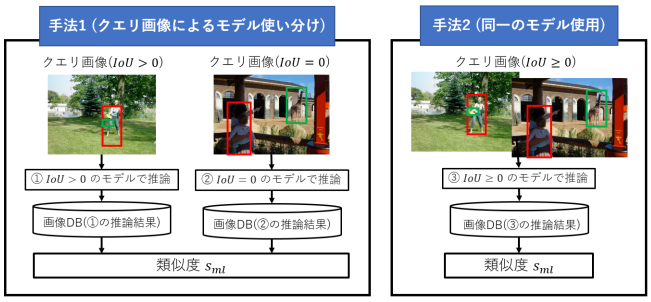


図 5: クエリ画像による，機械学習モデルの使い分けの検討．

手法 2 としてクエリの人モノペアの IoU にかかわらず，「人との $IoU \geq 0$ の物体 (すべての物体) を学習・推論に用いたモデル」を使うものの 2 つ (図 5) を比較した．実験結果は 4.3.3 で示す．

この行動ラベルを学習し，検出に用いる手法を**関係性学習検出手法**と呼ぶ．

3.2 類似場面検出アルゴリズム

姿勢の類似度，人と物体との距離の類似度，関係性検出の類似度の 3 つを用いて，クエリの人モノペアから類似行動をしている人モノペアを検出するアルゴリズムを示す (Algorithm1)．各人モノペアは物体検出技術により検出された物体のクラス情報を持っており，その物体クラスが一致している人モノペアについて類似度の算出をする．画像データベースの全人モノペアについて類似度の計算をし，類似度が閾値を上回ったものが検出結果となる．ここで，Precision = 80 %となる類似度を閾値として検出を行った．

Algorithm 1

Input: A query q, A set of human-object pair P
Output: A set of human-object pair P'

```
for p in P do
    if q_obj = p_obj then
        Sim[p] ← S_{q,p}
    else
        Sim[p] ← 0
    end if
end for
P' ← Sort(P) with regard to Sim
return P'
```

4 評価実験

提案手法の有効性を評価するため，2 種類の実験を行った．1 つ目の実験は提案手法による検出精度の検証であり，非関係性学習検出手法 (姿勢，物体との距離) と機械学習による関係性学習検出手法を組み合わせることで検出の精度が上がることを示す．2 つ目の実験は機械学習で学習していない動詞について

の検出精度検証であり，機械学習による学習データがない動詞でも，非関係性学習検出手法と組み合わせることで新しく動詞を学習することなく検出ができることを示す．

4.1 データセット

評価には， HOI 検出によく使われるデータセットである V-COCO [9] を用いた．V-COCO は COCO データセット [13] のサブセットで，物体カテゴリ 80，行動ラベル 26 種類で，人とモノの関係性についてアノテーションを提供する．本研究では，物体との関係性がある行動ラベル 22 種類を評価に使う．評価には V-COCO val データ 2867 枚を使用した．関係性学習には V-COCO train データ 2533 枚を使用した．

4.2 評価方法

FasterR-CNN で物体検知した人モノペアの中から，V-COCO の正解アノテーションがついているものを対象に全人モノペアについて評価を行う．評価指標として，precision = 80%時の recall を評価した．recall は，対象人モノペアの全正解データの内，検出結果に含まれる正解データの割合である．なお，本実験においては，クエリの人モノペアと検出結果の人モノペアが正解アノテーションにおいて同一の物体のクラス，行動ラベルを有していた場合を正解と判定した．

4.3 実験 1

4.3.1 重みの調整なし

まず，非関係性学習検出手法：関係性学習検出手法=1:1 の比率で実験した結果を示す (表 1)．非関係性学習検出手法のみ，関係性学習検出手法のみよりも，2 つの手法を組み合わせた方が検出の精度が上がることが確認できた．

表 1: 重みの調整なしでの結果

手法	類似度の計算	recall@precision80%
非関係性学習	$s_{\text{pose}} + s_{\text{ho}}$	0.233
関係性学習	s_{ml}	0.369
非関係性学習 + 関係性学習	$(s_{\text{pose}} + s_{\text{ho}})/2 + s_{\text{ml}}$	**0.422**

実際の結果を図 6，図 7 に示す．関係性学習手法において関係性が未検出，誤検出の場合であっても，非関係性学習手法と組み合わせることでより精度の高い検出が可能である結果が得られた．

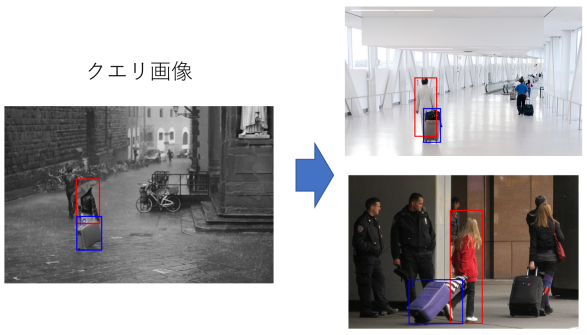


図 6: 関係性が未検出であり，非関係性学習手法のみで検出した結果例．

図 7: 関係性が誤検出であっても，非関係性学習手法を組み合わせることで改善した結果例．

4.3.2 非関係性学習の重み調整

本手法の精度を上げることを目標に，類似度の重みの調整を行った．姿勢の重みと，物体との距離の重みには，行動ごとに特徴が表れると考えた．例えば，姿勢は似ているが，物体との距離は画像によって異なる場合 (図 8)，姿勢の類似度の重みを大きくするとより似た画像を検出することができ，逆に姿勢は画像によって違うが，物体との距離は似ている場合 (図 9)，物体との距離の類似度の重みを大きくすると良いと考えられる．

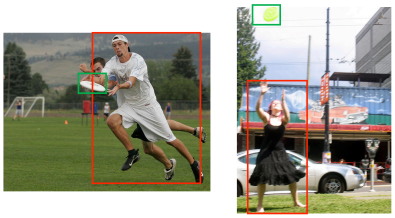


図 8: 姿勢のほうが類似性が高い例 (catch)．

図 9: 物体との距離のほうが類似性が高い例 (surf)．

そこで，姿勢が推定されている人モノペアのみをクエリ画像とし，行動ラベルごとに 0.1–2.0 の範囲で検出に一番良い重みをグリッドサーチで探索した (表 3)．結果を見ると，重みに特徴が表れた行動ラベルがいくつかあることがわかる．*eat*, *jump*, *lay*, *carry*, *throw*, *catch* などの行動は姿勢についてより重視した方が良く，*surf*, *kick* などの動詞は人と物体との距離に着目した方が良い．このことから，同じ行動ラベルの人モノペアの類似度が重みを決めるのに重要だと考えた．

よって，複数枚のクエリを受け取ったら，クエリ同士の類似度 (姿勢の類似度，物体との距離の類似度) から重みを算出す

表 2: 手法と類似度の計算における recall@precision80%

手法	類似度の計算	recall@precision80%
姿勢	s_{pose}	0.200
人と物体との距離	s_{ho}	0.231
関係性学習	s_{ml}	0.371
非関係性学習（姿勢＋人と物体との距離）	$w_{\mathrm{pose}}s_{\mathrm{pose}} + w_{\mathrm{ho}}s_{\mathrm{ho}}$	0.233
非関係性学習＋関係性学習	$w_{\mathrm{pose}}s_{\mathrm{pose}} + w_{\mathrm{ho}}s_{\mathrm{ho}} + w_{\mathrm{ml}}s_{\mathrm{ml}}$	**0.422**

表 3: アクションごとの重みとリコール

Action	w_{pose}	w_{ho}	recall_max	recall_min	query
hold	1.3	1.5	0.376	0.331	610
sit	0.1	0.4	0.478	0.459	225
ride	0.2	0.1	0.704	0.688	139
look	1	0.9	0.314	0.279	760
hit	0.6	0.1	0.065	0.058	110
eat	1.3	0.1	0.110	0.091	54
jump	2	0.1	0.064	0.0439	136
lay	1.9	0.1	0.091	0.048	33
talk_on_phone	0.3	0.2	0.139	0.077	19
carry	2	0.7	0.483	0.364	169
throw	1	0.3	0.074	0.052	70
catch	0.9	0.2	0.092	0.069	70
cut	0.1	0.2	0.165	0.140	52
work_on_computer	0.1	0.1	0.345	0.345	69
ski	0.1	0.4	0.810	0.810	194
surf	0.1	1.9	0.427	0.265	144
skateboard	0.1	0.1	0.594	0.594	143
drink	0.1	0.2	0.284	0.245	27
kick	0.5	1.9	0.455	0.162	58
point	0.1	0.1	1	1	5
read	0.1	0.2	0.213	0.184	15
snowboard	0.1	0.1	0.603	0.603	127

ることにした．算出方法を以下に示す．検出したい行動について，n 枚の画像があると仮定する．クエリ i と，クエリ j の姿勢の類似度 $z^{i,j}_{\mathrm{pose}}$ と人と物体との距離の類似度 $z^{i,j}_{\mathrm{ho}}$ は，式 (3)，式 (6) よりそれぞれ以下のように表される．

$$z^{i,j}_{\mathrm{pose}} = s^{i,j}_{\mathrm{pose}} \tag{9}$$

$$z^{i,j}_{\mathrm{ho}} = s^{i,j}_{\mathrm{ho}} \tag{10}$$

ここから，以下のように重み w^i_{pose}，w^i_{ho} を決定する．

$$\tilde{w}^i_{\mathrm{pose}} = \frac{1}{n}\sum_{j=1}^{n} z^{i,j}_{\mathrm{pose}} \tag{11}$$

$$\tilde{w}^i_{\mathrm{ho}} = \frac{1}{n}\sum_{j=1}^{n} z^{i,j}_{\mathrm{ho}} \tag{12}$$

$$w^i_{\mathrm{pose}} = \frac{\tilde{w}^i_{\mathrm{pose}}}{\tilde{w}^i_{\mathrm{pose}} + \tilde{w}^i_{\mathrm{ho}}} \tag{13}$$

$$w^i_{\mathrm{ho}} = \frac{\tilde{w}^i_{\mathrm{ho}}}{\tilde{w}^i_{\mathrm{pose}} + \tilde{w}^i_{\mathrm{ho}}} \tag{14}$$

調整した重みでの結果は表 4 に示す．今回使用したデータセットでは，重みの調整による精度の改善は見られなかった．式 (13), (14) の方法では，数枚のクエリ画像から重みを算出でき，未知の動詞についても対応することができる．あとの実験での重み w_{pose},w_{ho} は，式 (13), (14) によって算出された値を用いる．

表 4: 非関係性学習の重みの調整をした結果

重みの調整手法	類似度の計算	recall@precision80%
グリッドサーチ (表 3)	$w_{\mathrm{pose}}s_{\mathrm{pose}} + w_{\mathrm{ho}}s_{\mathrm{ho}}$	**0.236**
式 (13), (14)	$w_{\mathrm{pose}}s_{\mathrm{pose}} + w_{\mathrm{ho}}s_{\mathrm{ho}}$	0.233
1:1	$s_{\mathrm{pose}} + s_{\mathrm{ho}}$	0.234

4.3.3 クエリによる関係性検出モデル使い分け

関係性検出の類似度算出において，クエリによってモデルを使い分けた場合，recall@precision80%は 0.371，クエリによらず全物体を参照するモデルを用いた場合 0.369 となった．今回の実験では，モデルの使い分けによる精度改善はあまり見られなかった．ある行動ラベルが付与された人モノペアについて，IoU による特徴が現れると考えていたが，ほとんどの人モノペアで $IoU > 0$ のものが多く，$IoU = 0$ のモデルを使うことで，遠くにある物体をよく参照していると考えられる *throw*，*catch*，*look* 等において精度が上がることが期待されたが，$IoU > 0$ のペアも同等に含まれており (表 5)，効果が出ずらかったと考えられる．

4.3.4 非関係性学習と関係性学習の重みの調整

非関係性学習の重み w_{pose},w_{ho} と，関係性学習の重み w_{ml} の比率について検討する．関係性学習で学習済みの行動については，関係性検出の類似度の値を参考に決め，未学習の行動については，非関係性学習による姿勢の類似度，物体との距離の類似度によって行動が推定できるように重みを定めることで，さまざまなクエリに柔軟に対応できる．関係性学習の重み $w_{\mathrm{ml}} = 1$ とし，類似度を $S_{a,b} = \alpha(w_{\mathrm{pose}}s^{a,b}_{\mathrm{pose}} + w_{\mathrm{ho}}s^{a,b}_{\mathrm{ho}}) + s^{a,b}_{\mathrm{ml}}$ として，関係性学習に対する非関係性学習の重み α を調整した (図 10)．非関係性学習の重みが 0 より大きければ検出の精度は 0.42 付近であまり変わらず，$\alpha = 1$ だと精度が下がり始めていることから，今回の実験では非関係性学習に重きを置き，$(w_{\mathrm{pose}}, w_{\mathrm{ho}}) : w_{\mathrm{ho}} = 0.1 : 1$ となるように重みを決めることとした．以上の調整後の結果を表 2 に示す．recall の値を見ると，姿勢のみの類似度を使った検出，人と物体との距離のみの類似度を使った検出よりも，2 つを組み合わせた非関係性学習手法の方が精度が良く，また関係性学習手法のみでの検出よりも，非関係性学習手法と合わせた方が精度が上がることが確認できた．

表 5: 検出対象人モノペアの人と物体の IoU 内訳 (個).
右 2 列は関係性学習手法の使い分けの検討における
recall@precision80%の結果

行動ラベル	$IoU > 0$	$IoU = 0$	図 5 手法 1	図 5 手法 2
hold	1024	19	0.357	0.357
sit	339	52	0.397	0.390
ride	161	29	0.454	0.450
look	900	339	0.341	0.315
hit	92	78	0.157	0.233
eat	132	10	0.065	0.053
jump	237	14	0.104	0.149
lay	88	18	0.162	0.210
talk_on_phone	61	1	0.335	0.242
carry	199	2	0.108	0.160
throw	57	40	0.074	0.059
catch	62	41	0.009	0.012
cut	99	23	0.144	0.133
work_on_computer	179	13	0.476	0.476
ski	176	43	0.781	0.781
surf	201	33	0.742	0.746
skateboard	216	18	0.780	0.772
drink	45	2	0.519	0.501
kick	45	23	0.146	0.129
point	6	5	0.545	0.500
read	41	0	0.262	0.276
snowboard	162	20	0.685	0.698

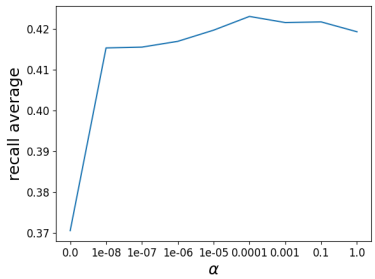


図 10: 関係性学習の重み α の調整結果

4.4 実　験　2

次に，関係性検出において，学習していない動詞について精度が落ちても，非関係性学習と組み合わせることで類似場面検出が可能であることを示すための実験をした．*hold*, *sit*, *look*, *work_on_computer*, *ski*, *kick* について，それぞれ学習データからその動詞を抜いて学習させ，モデルを作成した．実験 1 で調整した重みで関係性学習手法と非関係性学習手法を組み合わせ，それぞれ未知の動詞が正解ラベルである人モノペアをクエリとして検出を行った．結果を表 6 に示す．

関係性学習では，全ての動詞を学習した時 (カッコ内の値) よりも recall が下がっていることがわかる．*sit*,*work_on_computer* 等であまり数字が下がっていない理由としては，V-COCO は 1 つの人モノペアに複数の行動ラベルが正解としてついているため，1 つの動詞を学習データから抜いても，他の動詞から検出は可能であったと考えられる．非関係性学習＋関係性学習の列を見ると，関係性学習での検出精度が落ちても，非関係性学習手法と組み合わせることで精度劣化を抑制することができている．このことから，検出したい行動が関係性検出技術の学習データになくても，検出できる行動ラベルについて学習が必要ない非関係性学習と組み合わせることで検出が可能であるということがわかった．

5 結　　論

人モノ関係性に着目した類似場面において，人の姿勢の類似度，人と物体の距離，機械学習による関係性検出を組み合わせることで検出の精度があがることを確認した．機械学習で学習していない場面についても，非関係性学習の手法と組み合わせることで検出の精度をあげることができることがわかった．本技術に用いている各種検出エンジンは代替可能であり，物体検出，姿勢検出，関係性検出技術の向上により更なる精度向上が見込める．

文　　献

[1] Girum G Demisse, Konstantinos Papadopoulos, Djamila Aouada, and Bjorn Ottersten. Pose encoding for robust skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 188–194, 2018.

[2] Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 42–52, Beijing, China, July 2015. Association for Computational Linguistics.

[3] Desmond Elliott, Victor Lavrenko, and Frank Keller. Query-by-example image retrieval using visual dependency representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 109–120, 2014.

[4] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.

[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[6] Deeptha Girish, Vineeta Singh, and Anca Ralescu. Understanding action recognition in still images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1523–1529, 2020.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[10] Nataraj Jammalamadaka, Andrew Zisserman, Marcin Eichner, Vittorio Ferrari, and CV Jawahar. Video retrieval by mimicking poses. In *Proceedings of the 2nd ACM Inter-*

表 6: recall@precision80%（すべての動詞を学習したとき）

関係性学習において未知の動詞	非関係性学習	関係性学習	非関係性学習＋関係性学習
hold	0.221	0.291 (0.357)	**0.367** (0.395)
sit	0.277	0.349 (0.397)	**0.465** (0.498)
look	0.194	0.152 (0.341)	**0.259** (0.346)
work_on_computer	0.127	0.450 (0.476)	**0.514** (0.536)
ski	0.718	0.686 (0.781)	**0.854** (0.874)
kick	0.398	0.052 (0.146)	**0.429** (0.463)

national Conference on Multimedia Retrieval, pages 1–8, 2012.

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[12] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 74–83, 2021.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[14] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020.

[15] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR 2011*, pages 3177–3184, 2011.

[16] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[18] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 53–70. Springer, 2020.

[19] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 450–459, 2019.

[20] Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2839–2846, 2023.

[21] Xi Zhao, Myung Geol Choi, and Taku Komura. Character-object interaction retrieval using the interaction bisector surface. In *Computer Graphics Forum*, volume 36, pages 119–129. Wiley Online Library, 2017.

[22] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022.

[23] 劉 健全 吉田 登. 画角不変な姿勢特徴量を用いた類似行動検出. *DEIM2022*, 2022.

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 10:00 ~ 12:10 | T5-A オンライン（Zoom Events)

教育・学習 (4)

座長:林 康弘(武蔵野大学)
コメンテータ:合田 和生(東京大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T5-A-4-01] 道聴と同調について考えさせる体験型アプリケーションDOUCHO2の実装と評価

*杉山 功馬[1]、松村 貫司[1]、吉崎 悠画[1]、伊藤 一成[1] (1. 青山学院大学社会情報学部伊藤一成研究室)

10:20 ~ 10:35
[T5-A-4-02] 教科書の目次構造と科目用語体系に基づく補足Webページの抽出

*中村 愛咲[1]、佐野 逸稀[1]、井上 沙紀[1]、王 元元[2]、河合 由起子[3,4]、角谷 和俊[1] (1. 関西学院大学、2. 山口大学、3. 京都産業大学、4. 大阪大学)

10:35 ~ 11:00
[T5-A-4-03] 大規模言語モデルが生成した論理エラー修正コードと授業資料を活用したプログラミング支援

*上川 貴之[1]、山崎 禎晃[2]、大原 剛三[2] (1. 青山学院大学大学院理工学研究科、2. 青山学院大学理工学部)

11:00 ~ 11:25
[T5-A-4-04] 能動的学習データを用いた学習状況把握のための可視化分析手法の研究

*王 笑難[1]、孫 一[2]、大月 一弘[1]、清光 英成[1] (1. 神戸大学，国際文化学研究科、2. 神戸情報大学院大学，情報技術研究科)

道聴と同調について考えさせる体験型アプリケーション DOUCHO2 の実装と評価

杉山 功馬† 松村 貫司† 吉崎 悠画† 伊藤一成†

†青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5 丁目 10-1

あらまし SNS の発展に伴い，人は自分の意思に関係なく情報を目にする機会が増加した．情報の真偽を確かめず同調することは，SNS 上でデマを拡散することにつながりかねない．SNS 上で偶然目にした真偽が明確でない情報を，安易に同調して拡散しないことが重要である．今回，同音異義語である「同調」「道聴」の語意，及びその両語意の対比に着目することで，情報の真偽を見極めているのかを自律的に考えさせるアプリケーションを提案し，実装した．評価実験の結果，概ね良好な評価を得ることができた．

キーワード 道聴塗説，同調圧力，同音異義語，SNS，情報の真偽，人型ピクトグラム

1. はじめに

我々の研究グループでは，以前道聴塗説と同調圧力について考えさせるコンテンツ GODOUCHO を開発し Web 公開していた経緯がある[1]．メディアリテラシ，情報倫理教育，モラル教育など利用者視点から SNS の利用規範について日常的活動から学ぶという観点を重視していた．またその後，同調効果について考えさせる体験型アプリケーション DOUCHO を実装し，簡易的な評価を行った[2]．DOUCHO では，アンケートシステムにおいて同調した際に，歓声や拍手の音声を発生させることで同調圧力を促した．

SNS は，ユーザ自体が十分に情報の真偽を判断せずとも気軽に情報を伝播できるメディアである．そのため， SNS 上で拡散されるデマが後をたたない．情報過多時代における安易な情報拡散は,いわば「道聴塗説」と形容できよう．

SNS 上で偶然目にした真偽が明確でない情報を，安易に同調して拡散しないことが重要である．

本稿では，道端で聴きかじるという意味の「道聴」と調子が同じであるという意味の「同調」に着目し，実際に質問応答を通じたやり取りの中で，自律的に道聴と同調について考えさせる DOUCHO2 を開発したので報告する．

以下 2 章でアプリケーションを利用する上での前提知識について関連研究を踏まえながら示す．3 章で本研究のアプローチについて，解説する．4 章で評価と分析を行い，5 章でまとめる．

2. 前提知識

本研究では，道聴と同調について考えさせるアプリケーションを設計する上で，文献[1]と同様に「人型ピクトグラム」,「同音異義語」の 2 点について着目した．本章ではこれらについて解説する．

2.1 人型ピクトグラムの利用

ピクトグラムとは日本語で絵記号，図記号と呼ばれるグラフィックシンボルであり，意味するものの形状を使ってその意味概念を理解させる記号である．ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物など想起させる効果があると言われている.有名な「非常口」ピクトグラムのデザインの策定に関わった太田は「走る人型を囲む空間が見る人を包む空間とつながって走る人は見る人の投影になる」と述べ[4]，実際に避難中の人が如何に出口へ向かって走る人型ピクトグラムを見たときに，自身に投影し，同一視するかにデザインの労力が払われた．

田中は，人型のピクトグラムについて，認知主体が自らの経験に基づいて動作を行う自己像をピクトグラムに投影し,それを他者の視点で認識することにより，動態的な行為の可能性を認知していることを指摘している[5]．

2.2 同音異義語の効用

日本語は，多くの同音異義語が存在するのが特徴である． 同音異義語はシャレでよく利用される．織田はシャレの言語行動による理解と誤解において，同音異義語や音韻的類似を利用して文に 2 つの意味がある場合にもコミュニケーションが成立するのがシャレであると定義している[6]．また織田は，その二語の発音がより似ており，さらに語の意味がより異なるものほどシャレの効果が大きいと述べている．そこで今回,「道聴」と「道聴」の 2 つの同音異義語に着目する．

2.2.1 道聴

道聴は，小学館のデジタル大辞泉[7]の解説に道聴

塗説として，次のように記述されている．

1. 路上で聴いたことをそのまま直ちに話すこと．人の話を心にとどめて自分のものにしないこと．また，人から聞いたことをすぐに受け売りすること．いい加減なうわさ話．

道聴塗説の原典は論語であり，この論語の陽貨編に，「道聴而塗説，徳之棄也」という一節がある．これは道端で聴きかじったこと（道聴）を，すぐにまた道端で自説のように話すこと（塗説，塗は道と同じ意味）であり．これは徳を棄てる行為であると論語では述べられている．この一節は，道聴塗説という有名な四文字熟語の原典として知られている．またこの熟語は，いい加減な受け売りの話のたとえにも用いられる．本来，「道聴途説」は悪い意味で用いられ，ゆえに一般的には，「道聴」も悪い意味に捉えられるが，一方で，「道聴」してしまったけれど，「塗説」はしなかったという良い意味での解釈も可能であり，本コンテンツでは後者の解釈を採用する．

2.2.2 同調

同調は，小学館のデジタル大辞泉[7]の解説によると，次のように記述されている．

1. 調子が同じであること．また，同じ調子
2. 他のものと調子を合わせること．ある人の意見や態度に賛成し，同じ行動をとること．「彼の意見に－する」「－者」
3. 外部からくる振動に共振するように装置の固有振動数あるいは周波数を調節すること．特に，テレビ・ラジオで目的の周波数に合わせること．

参考程度に同調に対する定義について示す．アッシュによると，同調は，個人が集団や他者の期待に沿って行動することと定義される．特に，集団で，他人が一致して自分とは異なる意見を主張するとき，同調は生じやすいとされている[3]．また，ドイッチュらによると，同調には，「規範的同調」と「情報的同調」の2つがあるとされている．「規範的同調」は，多数派から受け入れられたいとの動機から生じる同調であり，集団のルールに従って行動するものである．「情報的同調」は，他者からより正確な情報を得ようとする動機から生じる同調である．これは，他人の意見や判断を参考にすることで，適切な行動をしようとするものである[8]．横田は，「規範的同調」と「情報的同調」には相関性があると実験で示している[9]．藤原は，同調行動を「自分とは異なる意見・態度・行動を周囲から求められたとき，迷いながらも周りの意見・態度・行動に合わせてしまうメカニズム」と定義している．また，同調行動を2種類に分け，内心から他者の意見や行動を受け入れる「内面的同調」と，表面的には同調しているように見えるが，内面では異なる「表面的同調」とに分けることができるとしている[10]．

また，このように二語の意味をより対比的に解釈することでシャレとしての効用を高める．シャレによる笑いやコミュニケーション創出の効用については社会学，認知言語学，感性工学や情報工学など様々な見地から研究が進められている[11][12]．

3. 実装アプリケーション

3.1 アプリケーション DOUCHO2

今回実装したDOUCHO2はWebアプリケーションなので，指定されたURIにブラウザでアクセスするだけで利用できる．DOUCHO2には問題を作成する出題者側が使用する「作問モード」と，出題者が作成した問題に対して回答する回答者側に分かれて使用する「回答モード」の2つがある．

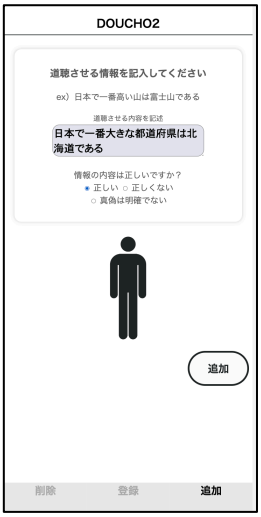
1. 「作問モード」…情報の真偽について他者に確認してもらうための通知をQRコードによって要請する．
2. 「回答モード」…読み込んだQRコードから情報の真偽について確認する．

以下3.2節にて「作問モード」，3.3節にて「回答モード」について説明する．

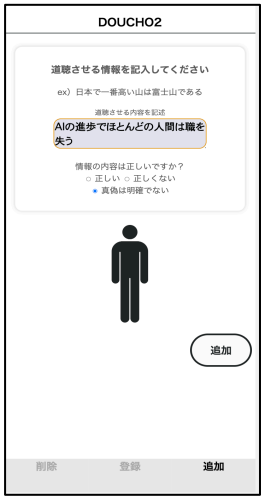
3.2 作問モード

「作問モード」のスクリーンショットを図1に示す．

（1）「入力待機状態」

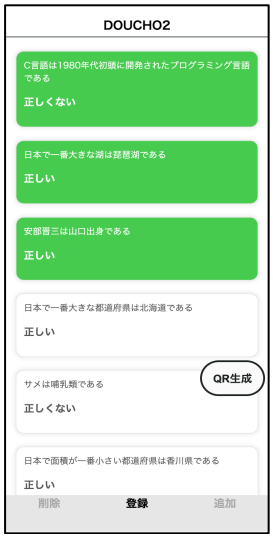


（2-1）「追加準備状態」



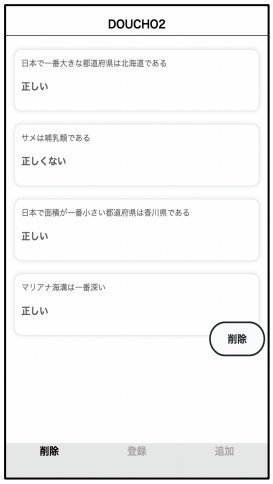
（2-2）「追加準備状態」

（3）「登録待機状態」



（4）「登録準備状態

（5）「QR コード表示状態」



（6）「削除待機状態」

（7）「削除準備状態」

図 1　作問モード時のスクリーンショット

各画面の下部には削除，登録，追加を選択するタブが表示されている．削除タブを選択すると，登録されてある全ての情報の組みが表示される．

情報の組が表示されているボックスを選択すると，ボックスの背景を赤色，ボックス内の文字を白色に変化させ，選択されたことを表現する「削除準備状態」に移行する．1 つ以上選択された状態で画面右下の「削除」と表示されているボタンを押下すると，選択状態の問題を削除する．削除されたことが確認されると，「削除されました．」という内容のダイアログが表示され，更新される．登録タブを選択すると，Google Sheets 上に登録されてある全ての問題の組みが表示されている．問題の組みが表示されているボックスを選択すると，ボックスの背景が黄緑色，ボックス内の文字を白色に変化させ，選択されたことを表現する，「登録準備状態」に移行する．1 つ以上選択された状態で波面右下の「登録」と表示されているボタンを押下すると，選択された問題のユニーク ID，Google App Script のデプロイ ID が含まれるエンドポイントの情報を変数に保存する．保存された情報を組み込んだ QR コードを生成し表示する画面に遷移する．

「登録準備状態」では，DOUCHO2 で使用する内容を変数に保存し，入力した内容を公開すると，入力された情報から QR コードを生成し，第三者に向けて QR コードを読み込むことに対して要請する「QR 表示状態」に遷移する．画面には，生成された QR コードと「読込みお願いします」という文字列が表示される．

追加タブを選択すると，画面上には道聴してもらう情報を入力するフォーム，道聴してもらう情報の内容について「正しい」，「正しくない」，「真偽は明確でない」のいずれかを選択するラジオボタン，追加ボタン，そして直立した人型ピクトグラムが表示されている．図 1 のスクリーンショットの一番上部に表示されている「質問する内容を記述」の見出しのテキストフィールドに指示に応じて内容を記述する．「情報の内容は正しいですか？」が見出しのラジオボタンで「正しい」，「正しくない」，「真偽は明確でない」のいずれかを選択する．画面下部には人型ピクトグラムが表示される．内容の入力途中に，入力している本人自身を人型ピクトグラムに投影させることで，情報の発信者としての心構えを意識させる．内容入力後，画面右下の「追加」と表示されているボタンを押下すると，入力された内容にユニーク ID を生成し，Google Sheets 上に書き込む．

出題者の一連の流れとしてまず，「作問モード」を使用する．ブラウザでアクセスすると，削除タブ，登録タブ，追加タブが表示される．デフォルトとして追加タブが選択された状態で表示され，入力フォームが表示される．

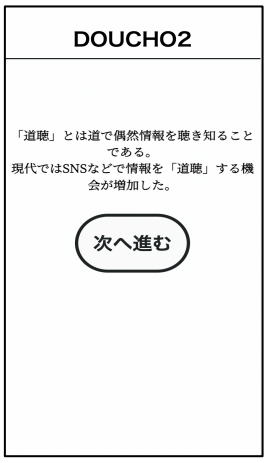
まず，道聴してもらう情報を記述する．続けて，情報の内容が「正しい」のか「正しくない」のか「真偽は明確でない」のかを選択する．例として，質問文に「日本で一番大きな都道府県は北海道である」,と記述し,「正しい」のラジオボタンを選択する．また,「AIの進歩でほとんどの人間は職を失う」といった真偽が明確でない情報は「真偽は明確でない」のラジオボタンを選択する．次に追加ボタンを押下すると，記述した内容の書き込みを行う．他に質問する内容がある場合は続けて記述し，ない場合は登録タブを選択し，問題の共有を行う．登録タブでは，これまでに追加した問題の一覧から出題する問題を任意に選択する．次に登録ボタンを押下する．「QR 表示状態」へ移行し出題者側は生成された QR コードをディスプレイなどに映し，回答者側に QR コードを読み込むよう要請する．

3.3 回答モード

「回答モード」について解説する．生成された QR コードもしくは，URL からアクセスした端末は，ブラウザ上に DOUCHO2 ページが表示される．PC とスマートデバイスでの表示は同様となっている．ページにアクセスすると，DOUCHO2 の利用者は次の 4 状態を遷移していく．

(1)道聴説明状態…道聴を説明する内容を 5 秒間表示
(2)選択待機状態…ブラウザ上に表示されている質問内容に対していずれかの単語を選択
(3)確認状態…選択された単語が正表記か誤表記かを表示
(4)結果表示状態…これまでに選択した正表記の数を集計し，結果を表示

各状態について順に説明する．図 2 に(1)「道聴説明状態」～(4)「結果表示状態」のスクリーンショットを示す．



(1)「道聴説明状態」

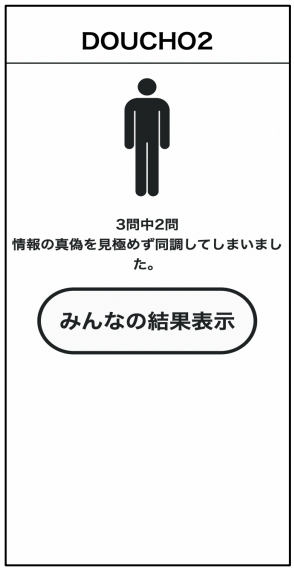
(2)「選択待機状態」

(3-A) 正しい情報に同調時の「確認状態」

(3-B) 正しい情報に非同調時の「確認状態」

(3-C) 正しくない情報に同調時の「確認状態」

(3-D) 正しくない情報に非同調時の「確認状態」



(4) 最終結果確認状態

(5) 最終結果確認状態(全体)

図 2　回答モード時のスクリーンショット

(1)「道聴説明状態」では，「「道聴」とは道で偶然情報を聴き知ることである。現代では SNS などで情報を「道聴」する機会が増加した。」という文章が表示される.これにより,「道聴」というキーワードに対する利用者の理解を促す.

「次へ進む」と書かれたボタンを押すと，(2)「選択待機状態」へ進む．(2)「選択待機状態」では，質問文が表示され「道聴した上で同調する」と「道聴すれども同調せず」がそれぞれ前面に表記されたボタンが表示されている.質問文の内容に従い,「選択してください」と表示し，いずれかのボタンを選択するよう促す．情報の真偽などは関係なく同じボタンを選択すると (3)「確認状態」に移行する．ここで選択されたボタンと情報の内容から判定を行い，それに対応する文字列を表示する．表 1 に正しい情報に同調した際に表示される内容を，表 2 に正しい情報に同調しなかった際に表示される内容を，表 3 に正しくない情報に同調した際に表示される内容を，表 4 に正しくない情報に同調しなかった際に表示される内容を示す．これらの表示は「正解」,「不正解」で評価を行う場合と比べて回答者側の知識量の有無に対する評価ではなく，回答者側の情報の真偽の見極めに対する意識の有無に対して評価を行う.

表 1　正しい情報に同調した際に表示される内容

内容
正しい情報に同調しました

表 2　正しい情報に同調しなかった際に表示される内容

内容
正しい情報を見逃してるかも

表 3　正しくない情報に同調した際に表示される内容

内容
安易に同調するべからず！
情報の真偽はしっかり見極めよう
誤った情報を広めてるかも、、、

表 4 正しくない情報に同調しなかった際に表示される内容

内容
正しくない情報に同調しませんでした

その後画面下部に表示されている「次にすすむ」ボタンを押下すると,「作問モード」時に登録された内容に対して全て選択している場合は(4)「最終結果確認状態」に移行し，全て選択していない場合は再び(2)「選択待機状態」へ移行する．その際表示される質問内容は登録された順番で表示される.(4)「最終結果確認状態」ではこれまでの質問に対する質問の真偽を見極め正しく「同調する」,「同調しない」を選択した数が「～問中～問，情報の真偽を見極めて同調できました。」という表記で表示される．また，画面下部に表示されている「みんなの結果表示」と表記されたボタンを押下すると,(5) 最終結果確認状態(全体)に移行する.(5) 最終結果確認状態(全体)では，これまでの質問に対する全体の同調数と非同調数が表示される．他の人が質問に対して，どのような選択をしているのかを確認することができる.

回答者の一連の流れは，出題者側から向けられた QR コードをスマートデバイスなどの端末で読み込む．ブラウザを開きアクセスすると「回答モード」へ移行する.出題者側が作成した質問が表示され,「道聴した上で同調する」と「道聴すれども同調せず」が表記された二つのボタンが表示される.例としては質問文に「日本で一番高い山は富士山である」が表示され，二つのボタンが表示される．回答者側は質問内容に対して適している方の選択肢ボタンを押下する．選択肢ボタンを押下すると,選択した内容に対する評価が行われる.「日本で一番高い山は富士山である」の様に質問内容が「正しい」場合は,「道聴した上で同調する」を選択した際に「正しい情報を同調しました」という文章が,「道聴すれども同調せず」を選択した際に「正しい情報を見逃してるかも」という文章が表記される.また,「サメは哺乳類である」の様に質問内容が「正しくない」場合は，「道聴した上で同調する」を選択した際に「安易に同調するべからず！」などの文章が,「道聴すれども同調せず」を選択した際に「正しくない情報に同調しませんでした」という文章が表示される.

4. 評価，考察

4.1 概要

2023 年 10 月 27 日（金曜日），青山学院大学社会情報学部が開講するウェブテクノロジーを履修している学生を対象に，評価を行った.評価者の人数は 67 名である．情報の真偽がわからない 3 つの問題に回答してもらい，その後アンケートを実施した．回答してもらった問題の内容を表 5 で示す．問題は本アプリケーションである DOUCHO2 を用いて作成しており,「道聴した上で同調する」と「道聴すれども同調せず」の二択で回答してもらった．問題の内容は情報の真偽が真の問題，偽の問題，明確でない問題が 1 問ずつで，合計 3 問設定した.

表 5 問題の内容と真偽

Q	内容	真偽
1	ダブリンコアというあらゆるリソースに共通に付与可能な語彙集合がある	真
2	RDF ではリソース、プロパティ、プロパティ値はいずれも必ず URL で表記する	偽
3	セマンティック Web は将来性のある概念である	明確でない

4.2 アンケート評価

授業実践の内容の妥当性，研究内容の可能性を評価する目的で授業の最後に無記名式のアンケートを行った．質問項目は表 6 に示す 4 項目である．

質問 1 は 2 段階の回答選択肢を，質問 2，3 は 5 段階の回答選択肢を用意した．67 名より回答があった．アンケートの回答結果を以下の表 7，8，9 に示す．

表 6　アンケート内容

Q	内容	回答形式
1	情報の真偽が明確ではないものに対して同調してしまった質問があった	2 段階選択
2	自分は日頃から情報の真偽を見極めていると思う	5 段階選択
3	このアプリケーションは情報の真偽を見極めているか考えるきっかけになった	5 段階選択

表 7　設問 1 のアンケート結果

回答	人数
はい	64
いいえ	3

表 8　設問 2 のアンケート結果

回答	人数
とてもそう思う	5
ややそう思う	40
どちらでもない	14
あまりそう思わない	7
そう思わない	1

表 9　設問 3 のアンケート結果

回答	人数
とてもそう思う	10
ややそう思う	37
どちらでもない	8
あまりそう思わない	10
そう思わない	2

質問 1 のアンケート結果から，約 95%の評価者が情報の真偽が明確でないものに対して同調してしまったと回答した．しかし，質問 2 のアンケート結果では，約 65%の評価者が日頃から情報の真偽を見極めていると思うと回答している．この結果から，多くの人が情報の真偽を見極めていると自負はしているが，正確に真偽を見極めることは出来ていないということがわかる．本アプリケーションでは，情報の真偽を見極めることが出来ているのかを即時的に判定することが可能であるため，自身が正確に真偽を見極めているのかを見直させることが可能であると考えている．実際に，質問 3 では約 70%の評価者が，本アプリケーションは情報の真偽を見極めているか考えるきっかけになったと回答しているため，真偽を見極めているのかを見直させるという点で有効であると考えられる．

5. おわりに

同音異義語である「同調」「道聴」の語意に着目することで，道聴と同調について考えさせるアプリケーションを提案し，実装，評価を行った．評価実験では，概ね良好な評価を得ることができたが，いくつか改善が必要な点も明確になったため，今後改善していく予定である．

謝辞

本研究は JSPS 科研費 21H03560 の助成を受けたものです．

参考文献

1) 伊藤一成：道聴塗説と同調圧力について考えさせるコンテンツ GODOUCHO（ご道聴，ご同調），情報教育シンポジウム 2015 論文集 2015 (2015)
2) 大野友哉，御家雄一，伊藤一成：同調効果について考えさせる体験型アプリケーションの試作，情報処理学会第 85 回全国大会 (2023)
3) Asch, S. E.:Effect of groupressure upon the modification and distortion of judgements. In H. Guetzkow (Ed.), Groups, leaderships and men. Carnegie Press (1951)
4) 太田幸夫．ユニバーサル・コミュニケーションデザインの認識と実践 https://forum8.co.jp/topic/universal107.htm (参照 2023-12-29)
5) 田中敦：公共サインのピクトグラムにおける情報の認知に関する研究，新潟国際情報大学 国際学部 紀要 Vol.7 pp. 107-121 (2022)
6) 織田正吉,野村雅昭：シャレ・ダジャレ学事始(ことはじめ)，日本笑い学会 笑い学研究, Vol.6 , pp. 55-67 (1999)
7) デジタル大辞泉 https://daijisen.jp/digital/index.html (参照 2023-12-29)
8) Deutsch, M., Gerard, H. B : A study of normative and informative social influence upon individual judgement. Journal of Abnormal & Social Psychology, 51, pp. 629-636 (1955)
9) 横田晋大，中西大輔：同調志向尺度の作成：規範的影響と情報的影響，広島修大論集 Vol.51，No.2，pp. 23-36 (2011)
10) 藤原正光:同調行動志向尺度・個人行動志向尺度作成の試み (1)：大学生による小 5 時代の回想から，立教大学教育学部紀要 40，pp. 1-9 (2006)
11) 木村洋二：笑いの社会学, ISBN: 4790702553 (1983)
12) 濱田真樹，鬼沢武久:同音異義語の意味の多様性を構造にもつなぞなぞの生成,日本知能情報ファジー学会誌 知能と情報， Vol.20, No. 5, pp.696-708 (2008)

教科書の目次構造と科目用語体系に基づく補足 Web ページの抽出

中村　愛咲† 　佐野　逸稀†† 　井上　沙紀†† 　王　元元††† 　河合由起子††††,††††† 　角谷　和俊††

† 関西学院大学総合政策学部　〒 669–1330 兵庫県三田市学園上ケ原 1 番
†† 関西学院大学総合政策研究科　〒 669–1330 兵庫県三田市学園上ケ原 1 番
††† 山口大学大学院創成科学研究科　〒 755-8611 山口県宇部市常盤台２丁目１６番１号
†††† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
††††† 大阪大学サイバーメディアセンター　〒 567-0047 大阪府茨木市美穂ケ丘５番１号
E-mail: †{idw65509,gil06518,hms54907,sumiya}@kwansei.ac.jp, ††y.wang@yamaguchi-u.ac.jp,
†††kawai@cc.kyoto-su.ac.jp

あらまし　一般的に講義スライドは, スライドタイトル・見出し・本文という構造になっており, 講師はスライドタイトルに含まれる用語を学生に理解してもらうために見出しや本文に入れる情報を工夫する必要がある. しかし, 文字数や講義時間などの制限により, 講師はスライドに全ての情報を入れることができない. それにより, スライドタイトルに含まれる用語に関する情報と, 見出しに含まれる用語に関する説明はスライドに記載されていても, 見出しに含まれる用語が説明される本文に含まれる用語は学生が知っている前提で説明が省略されることが多いと考えれられる. そこで, 本研究は, 教科書の目次構造と科目用語体系を利用することで, 講義スライドの内容を補足する用語を抽出し, 教科書の内容を踏まえた学生の理解を促進する補足 Web ページを推薦することができる.

キーワード　学習支援, e-learning, 推薦システム, 教育, 検索クエリ

1 はじめに

一般的に講義スライドは, スライドタイトル・見出し・本文という構造になっており, 講師はスライドタイトルに含まれる用語を学生に理解してもらうために見出しや本文に入れる情報を工夫する必要がある. しかし, 文字数や講義時間などの制限により, スライドに全ての情報を入れることができない. それにより, スライドタイトルに含まれる用語に関する情報と, 見出しに含まれる用語に関する説明はスライドに記載されていても, 見出しに含まれる用語が説明される本文に含まれる用語は学生が知っている前提で説明が省略されることが多いと考えれられる. 例えば, 図**??**のスライド 3 は, ソフトウェアの種類として「基本ソフトウェア」「ミドルウェア」「ハードウェア」があり, 本文にはそれぞれの説明文が記載されているが, その本文に含まれる「ハードウェア」や「アプリケーション」は説明が省略されている. 学生は, 本文の「ハードウェア」や「アプリケーション」を知らなければ「基本ソフトウェア」を理解することは難しく, 延いては「ソフトウェア」を理解することが難しくなると考えられる. そこで, 本研究は, 教科書の目次構造と科目用語体系を利用することで, 講義スライドに対する補足 Web ページを学生に推薦する. 科目用語体系とは, 科目用語の特徴や用語間の関係をもとに体系的に整理されたものを指し, これを利用することで科目全体を構成する用語を概念的な関係性を基に, 講義スライドに含まれていないが講義において重要な用語を抽出することができる. このような科目用語体系と順序を持った教科書の目次構造を利用することによって, 用語の関係性と教科書の内容を踏まえた, 学生の理解を促進する Web ページを推薦することができる.

2 関連研究

2.1 Web コンテンツ中に存在する知識間の理解支援関係の自動抽出

高橋ら [2] は, 発展途上国などインターネットへのアクセスはあるものの必要な教科書を買うことができない学生が多数存在することや, インターネットに関連情報があっても教科書が存在しない新たなテーマが多数存在することを背景に, インターネット上にあるさまざまな教材コンテンツから知識間の理解支援関係を抽出する手法を構築した. 理解支援関係とは, あるトピックの知識 v_1, v_2 を学習する際に, これらの間に「v_1 を理解するためには v_2 を理解しておいた方がいい」ということが成立するときの関係である. このような理解支援関係を得ることで, すでに理解している知識をもとに, 次に学習すべき知識を推薦することや, 理解が十分でない知識をもとに, 先に学習すべき知識を推薦することが可能になる. 本研究とは, 用語の関係性をもとに推薦すべき知識を抽出する点で似通っているが, 高橋らは関係導出手法として, 単純 TS 法と NMF 中間ノード TS 法を利用しているのに対して, 我々は, 教科書の目次情報と教科書の索引用語から構築された科目用語体系を利用している点で異なる.

2.2 Web ページ推薦

Web ページ推薦を目的とした研究は数多く存在し, 代表的なものとして協調フィルタリングを用いたものが挙げられる. 協

調フィルタリングとは， 個々のユーザの興味を他のユーザの興味を用いて推定し， 情報を推薦するという手法であり，代表的な研究として岩田ら [3] の研究がある． 同様にユーザの閲覧履歴を用いた研究としては， 堀ら [4] の研究があるが, いずれの研究も Web ページの内容を活用した推薦手法とは言えない. そこで， 枝ら [5] はユーザの閲覧履歴情報とページの内容の解析を組み合わせた手法を提案しており， Web ページの内容を考慮した推薦を行っている． 本研究も， Web ページの内容を活用した推薦を行う点で, 枝らの研究と類似しているが， 本研究は教育効果を期待した Web ページ推薦となっており， ユーザの閲覧履歴情報ではなく， 科目用語体系を用いる.

3 教科書の目次構造と科目用語体系

本研究では, 講義スライドの内容を補足する Web ページを学生に推薦するために, 教科書の目次構造と科目用語体系を利用する. 科目用語体系は, 補足 Web ページを推薦すべき講義スライドを検出や, 講義スライドに出現しないが講義内容に関する知識を広げることができる用語を抽出し検索クエリを生成するために利用する. 教科書の目次構造は, 生成した検索クエリを用いて検索エンジンから抽出した複数の Web ページをリランキングする際に用いる. 次節より, 補足 Web ページの抽出に利用する, 科目用語体系と教科書の目次構造について具体的に説明する.

3.1 教科書の目次構造

教科書の目次構造について説明する. 同じ科目の教科書であっても, 教育機関や出版社の違いなどの違いなどによって, 取り扱う用語や説明順序が異なる場合がある. 例えば, 川井ら [6] の『データサイエンス体系 データ構造とアルゴリズム』では, 第 3 章で「配列」「リスト」が説明された後, 第 4 章で「スタック」「キュー」が説明されているのに対して, 杉原 [8] の『データ構造とアルゴリズム』では, 第 2 章で「配列」「リスト」が説明された後,「ヒープ」「ハッシュ」「バケット」「再帰呼び出し」「分割統治」が説明され, 第 6 章でようやく「スタック」「キュー」が説明されている. 教科書によって取り扱う用語や説明順序が異なる要因として, 教科書がどのような目的に沿って情報を提供しているのかの違いにあると考えられる. 例えば, 基礎的な知識の理解を目的とした教科書の場合は, 基礎的な用語の説明を中心に構成されているが, 基礎を踏まえた実践的なスキルの習得を目的とした教科書の場合は, 先述の教科書よりも基礎的な用語の説明が比較的少ないと考えられる. 教科書によって, 取り扱う用語や説明順序が異なることから, 教科書の目次構造は動的な構造 (Dynamic-Structure) であると言える. つまり, 教科書の説明順が必ずしも正しいとは言えず, 学生にわかりやすく説明するために講師は説明順序を教科書から修正する可能性があると考えられる.

3.2 科目用語体系

科目用語体系について説明する. 赤澤ら [1] は, 高等学校共通教科情報科の知識体系の構築を行なっており, 我々は, 赤澤らによって構築された知識用語体系に用語間の関係性を追加する. 本研究では, その用語の概念体系を科目用語体系と定義する. 科目用語体系は, 上位・下位 (is-a), 全体・部分 (part-of), 同義 (same-as) の 3 つの概念を考慮しており, 本稿では, 2 つの教科書 [6] [7] の索引から用語を手作業で抽出し, 上記の 4 つの概念で用語間の概念関係を構造化したものを用いる. その用語体系は図科目用語体系を図 1 に示す. 科目用語体系における用語間の関係は教科書の目次構造と異なり不変であるため, 静的な構造 (Static-Structure) であると言える.

4 補足 Web ページの抽出

4.1 推薦すべき補足情報

本研究は, 科目用語体系の利用により注目するスライドに対して補足情報が必要な否かを検出する. 我々は, 学生に推薦すべき補足情報として, スライドに出現する用語の理解に繋がる情報であると考える. 具体的には, (1) スライドタイトルや見出しに出現する用語の詳細情報, (2)(1) の用語に関連する用語の詳細情報である. あるスライドに注目したとき, スライドタイトルや見出しに出現する用語はそのスライド内で説明されているが, 本文に出現する用語については説明されていない可能性が高い. つまり, 本文に出現する用語を理解できなければ見出しの用語を理解することは難しく, 見出しの用語を理解できなければ, スライドタイトルに出現する用語を理解することが難しくなると考える. そこで本研究では, 推薦する補足情報として 2 種類の Web ページを推薦する. 1 つ目は出現語の詳細情報で本稿では「詳細情報」と定義する.「出現語」とは, スライドタイトルと見出しの用語などの講義スライドで説明されている用語を指す. 2 つ目は関連語の詳細情報で本稿では「関連情報」と定義する.「関連語」は, 講義スライドで説明されていないが, 出現語と概念的な関係性がある用語を指す. 検索クエリを生成する際, 詳細情報の場合はスライドタイトルと見出しに出現する用語を用い, 関連情報の場合は科目用語体系に含まれる講義スライドに出現しない用語 (関連語) を用いる. 次節で, 出現語と関連語の抽出について説明する.

4.2 出現語と関連語の抽出

4.2.1 出現語の抽出

出現語の抽出について図 4 を用いて説明する. 出現語とは, スライドタイトルや見出しに出現する用語など, 講義スライドで説明されている用語を指す. まず, 全ての講義スライド (S) からスライドタイトルに出現する用語 (T_{s_n}) と見出しに出現する用語 (C_{s_n}) を抽出する. ここで抽出された T_{s_n} と C_{s_n} に含まれる用語が出現語に該当する. 図 5 のスライドの場合, スライド下部に示された用語が出現語にあたる.

4.2.2 関連語の抽出

関連語の抽出について図 4 を用いて説明する. 関連語とは, スライドで説明されていないが講義内容と関連性がある用語を指す. 具体的には, スライドの本文に出現する用語やスライドに出現しないが出現語と関係性がある用語を指す. 関連語は以下の

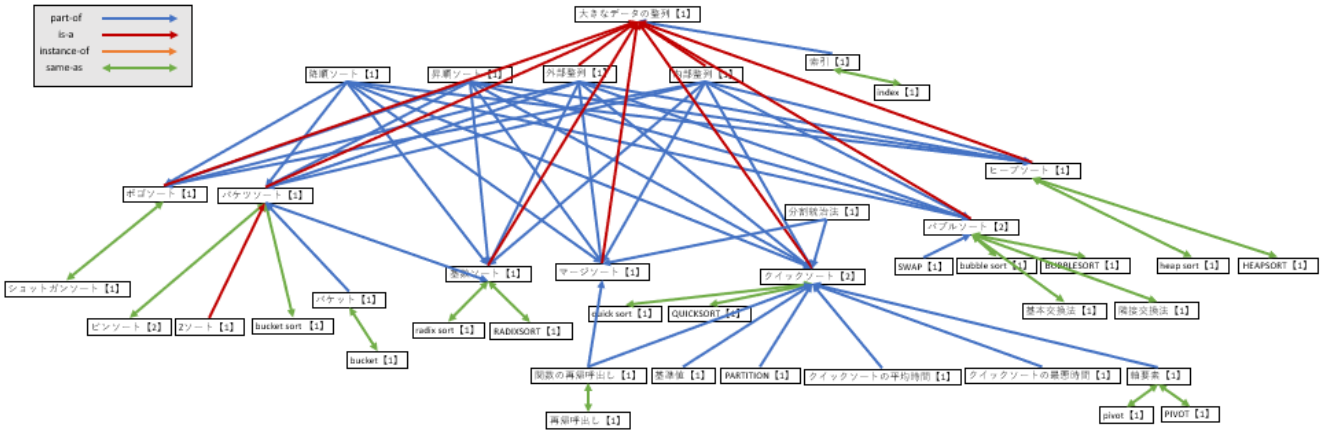


図 1　科目用語体系

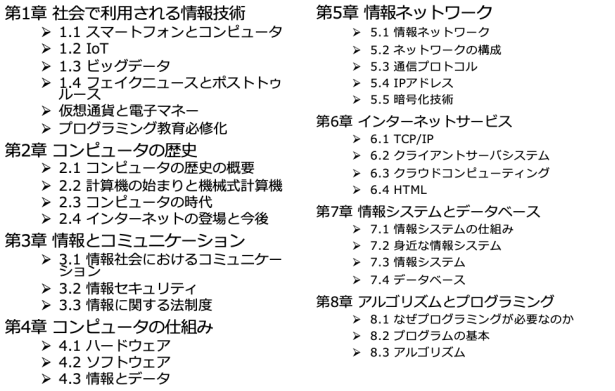


図 2　教科書「情報入門」の目次

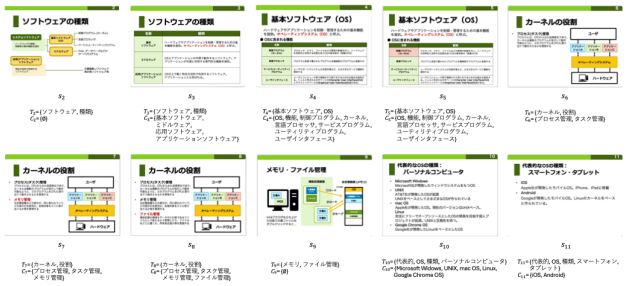


図 3　講義スライド: 大阪大学第 4 回『コンピュータの仕組み コンピュータの構成 ソフトウェア編』

手順に従って抽出する.

（1） 科目用語体系から出現語を検出する

（2） 注目するスライドの出現語と概念関係がある用語を検出する

（3） (2) で新たに検出された用語を関連語とする

図 4 の科目用語体系で検出された用語は, (1) の操作により検出された用語である. そのうち赤で示された用語 k_1, k_2, k_3, k_4 は, 注目スライドの出現語である. 青で示された用語 $k_{20}, k_{21}, k_{24}, k_{25}$ は (2) の操作によって科目用語体系から新たに検出された用語で, これらを関連語とする. ここで注意したいのは, k_5, k_6 は注目スライドの出現語 k_2 と概念関係があるが, これらの用語は (1)

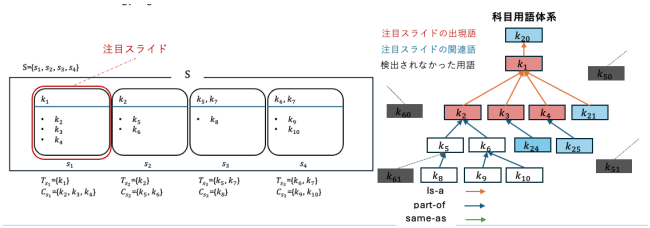


図 4　出現語と関連語の抽出

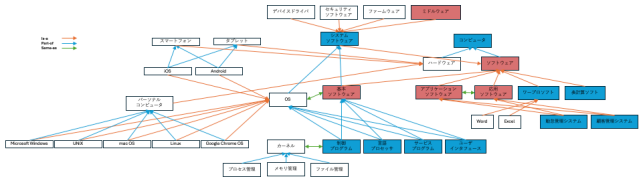


図 5　スライド 3 に注目した場合の用語の検出例

の操作の時点で検出されている用語, つまり講義スライドに出現する用語であるため関連語としない. 関連語が抽出されたので注目するスライドは補足情報を推薦すべきであると判定する. 図 3 のスライドを例にすると, スライド 3 に注目する場合, 図 5 のような結果になる. 次節では, 抽出した用語を用いた検索クエリの生成について説明する.

4.3　検索クエリの生成

前節で抽出した用語を用いて関連 Web ページの抽出するための検索クエリを生成する. 用語間に is-a 関係がある場合, その下位概念にあたる用語を検索クエリにする. 理由としては, 下位概念にあたる用語の共通項を見つけることで学生は上位概念の理解を深めることが可能になると考えるためである. 例えば, スライド 3 の場合, 「ソフトウェア」と「ワープロソフト」, 「ソフトウェア」と「表計算ソフト」, 「ソフトウェア」と「システムソフトウェア」に is-a 関係があるが, 上位概念である「ソフトウェア」を理解するには, 「ワープロソフト」「表計算ソフト」「システムソフトウェア」などの具体的な情報を知ることが有効であると考える. よって, スライド 3 において is-a 関係の検索クエリは, 「ワープロソフト」「表計算ソフト」「システムソ

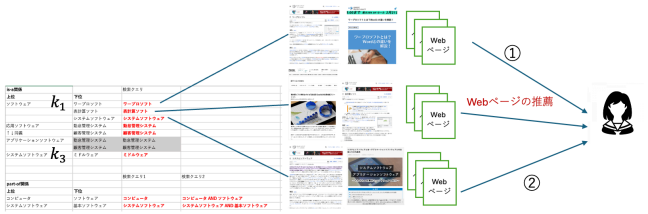


図 6　教科書の目次構造を用いた Web ページのランキング

フトウェア」になる. 次に, 用語間に part-of 関係がある場合, 上位概念にあたる用語を検索クエリにしたものと, 上位概念と下位概念を AND 検索にし下位概念を複数持つ場合それらの用語を OR 検索にした検索クエリの 2 つを生成する. 理由としては, 上位概念を広範的に理解した上でそれを構成する下位概念を理解することが適切であると考えるためである. 例えば, スライド 3 の場合, 「基本ソフトウェア」と「制御プログラム」, 「基本ソフトウェア」と「言語プロセッサ」, 「基本ソフトウェア」と「サービスプログラム」, 「基本ソフトウェア」と「ユーザインタフェース」に part-of 関係があるが, 制御プログラムを深く理解したからといって基本ソフトウェアを理解したことにならない. まずは基本ソフトウェアの広範的に理解し, それから制御プログラムなどの下位概念を理解することで基本ソフトウェアについての理解を深めることが可能になる. よって, スライド 3 において part-of 関係の検索クエリは, 広範的に理解するための検索クエリ「基本ソフトウェア」と, より深い理解をするための検索クエリ 「基本ソフトウェア AND (制御プログラム OR 言語プロセッサ OR サービスプログラム OR ユーザインタフェース)」になる.

5　教科書の目次構造による Web ページのランキング

検索クエリを用いて検索エンジンから抽出した Web ページのランキングを, 教科書の目次構造を用いて決定する. 推薦する Web ページを教科書で説明されている順序と合わせることで, 学習の流れを反映させることができるため, このシステムを利用する学生は目を通しやすくなると考える. 具体的に, 図 6 を用いて説明する. 教科書の 40 ページ目に k_1, 41 ページ目に k_2 が出現する場合, Web ページのランキングは, k_1 の補足 Web ページ, k_2 の補足 Web ページの順となる.

6　おわりに

本研究は, 科目用語体系を利用することで講義スライドに出現していないが講義において重要な用語を抽出し, 講義スライドに対する補足 Web ページを抽出するための検索クエリの生成手法を提案した. また, 抽出した Web ページを教科書の目次構造を利用することで, 推薦する Web ページに学習の流れを反映させた. これにより, 学生の理解を促進につなげることができると考える. 今後の課題として, 推薦する Web ページが講義の理解促進に有効であるかどうかの評価実験の実施と, 教科書の目次構造を用いる効果を検討する.

文　献

[1] 赤澤紀子, 赤池英夫, 柴田雄登, 山根一朗, 角田博保, 中山泰一. 高等学校共通教科情報科の知識体系に関する一考察. 情報処理学会論文誌教育とコンピュータ (TCE), Vol.8, No.3, pp. 19–34, 2022.
[2] 高橋健太郎, 伊藤寛祥, 松原正樹, 森嶋厚行. インターネット学習における学習支援ガイド生成に向けた理解支援関係の自動抽出. DEIM Forum2023,5a01.
[3] 岩田具治, 山田武士, 上田修功. 購買順序を効率的に用いた協調フィルタリング. 情報知識学会論文誌, Vol. 49, No. 4, pp.125–134, 2008.
[4] 堀幸雄, 今井慈朗, 中山尭. ユーザの web 閲覧履歴を用いた検索支援システム. 情報知識学会誌, Vol. 17, No. 2, pp. 95–100,2007.
[5] 枝 隼也, 佐藤 哲司. ユーザの Web 探索履歴におけるキーワード遷移に基づく Web ページ推薦システム. DEIM2010, Mar, 2010
[6] 川井明, 梅津高朗, 高柳昌芳, 市川治. データサイエンス体系 データ構造とアルゴリズム. 学術図書出版社, 2018, 192p, 978-4-7806-0703-1, https://www.gakujutsu.co.jp/product/978-4-7806-0703-1/.
[7] 茨木俊秀. アルゴリズムとデータ構造 (21 世紀を指向した電子・通信・情報カリキュラムシリーズ). 昭晃堂, 1989, 209p, https://ndlsearch.ndl.go.jp/books/R100000074-I000360537.
[8] 杉原厚吉. データ構造とアルゴリズム. 共立出版, 2001, 214p, 9784320120341, https://www.kyoritsu-pub.co.jp/book/b10011095.html.

大規模言語モデルが生成した論理エラー修正コードと授業資料を活用したプログラミング支援

上川　貴之† 　山崎　禎晃†† 　大原　剛三††

† 青山学院大学大学院理工学研究科　〒 252–5258 神奈川県 相模原市 中央区 淵野辺 5-10-1
†† 青山学院大学理工学部　〒 252–5258 神奈川県 相模原市 中央区 淵野辺 5-10-1
E-mail: †c5622208@aoyama.jp, ††{yamazaki,ohara}@it.aoyama.ac.jp

あらまし　プログラミング教育において，プログラム作成者の意図と異なる結果が出力される論理エラーは様々な要因で起こりうるため，解決方法の共通化が難しく，教員は個別に対応する必要がある．この問題に対し，近年では，問題文やエラーを含むコードを与えた際に自由度の高いアドバイスを生成できる大規模言語モデル（LLM: Large Language Model）の活用が注目を集めている．しかし，LLM が生成したアドバイスは正しいとは限らず，学習の妨げとなることが懸念されている．そこで本研究では，課題に対して正しく動作する模範コードが利用可能であることを前提とし，正しい実行結果が得られるまで LLM に与えるプロンプトを漸進的に改善する手法を提案する．さらに提案手法では，修正されたコードと元のコードの比較を通して正確なアドバイスの生成を試み，加えて，得られたアドバイスなどから抽出したキーワードを元に，授業資料中の該当ページを検索して学習者に提示する．評価実験から提案手法の授業資料ページ提示によりアドバイスの有用性が向上することが明らかになった．

キーワード　教育：e-Learning，EdTech，LLM

1　はじめに

近年，人工知能やビッグデータを用いた新たなビジネスの台頭により第 4 次産業革命が起こっている．それにともない，新たなビジネスの担い手となる IT 人材の育成が求められているが，その指導者不足が問題となっている [1]．特に，プログラミング教育においては，講義と並行して実技が必要とされるだけでなく，プログラミングの自由度の高さから生じる様々な問題への対処が求められるため，すべての学習者に行き届いた指導を実現することは困難である．

プログラミングの習得では，プログラミング言語の文法だけでなく，プログラムの動作手順をアルゴリズムとしてまとめるための論理的思考を学ぶ必要がある．それに加えて，正しい結果が返されない場合には，不具合（エラー）の原因を特定し修正するデバッグ作業が求められる．C 言語のようなコンパイル型のプログラミング言語のエラーは，文法の間違いが原因となるコンパイルエラー，segmentation fault などの実行時エラー，実行手順の論理的間違いによって起きる論理エラーの 3 種類に大別することができ，その中でもプログラミング言語の仕様とは無関係に生じる論理エラーの修正が難しいとされている．この理由としては，コンパイルエラーと実行時エラーに関しては，それぞれコンパイル時，実行時において，その解決のヒントとなるエラーメッセージやエラーの発生箇所がシステム的に出力されうるのに対し，論理エラーに関しては，エラーメッセージが出力されず，エラーの発生箇所の特定が難しいだけでなく，発生箇所を特定しても正しいアルゴリズムを理解していないと修正ができないためである．また，整数同士の除算時のデータ型の間違いなど，多くのプログラムに共通する論理エラー [2] も存在するが，プログラムは文法を守っていれば高い自由度で記述できるため，課題や学習者が作成したプログラム固有の論理エラーに対して個別の対処が求められる．特に，新たなプログラミング課題を作成した場合には，学習者が引き起こす論理エラーを事前に想定し網羅しておくことは非常に難しい．

近年，テキストからのコード作成やコードに対して自然で自由度の高いフィードバックを返すことができる大規模言語モデル（LLM: Large Language Model）が注目されている．その中でも，OpenAI が開発した対話機能を持つ ChatGPT [3] は優れたコード生成・修正能力を示しており，エラーの検出からメッセージの出力までも担えるため，プログラミング，および，プログラミング教育への活用が研究されている [4–8]．しかし，LLM は間違った解答を生成すること [3] があり，単純に利用することはできない．また，課題の答えを簡単に得られるという特性が学生の問題解決能力に悪影響を与える可能性があることが指摘されており [9]，課題に対する答えではなく解決につながるアドバイスや情報を学習者に提示する必要がある．

そこで本研究では，プログラミングの学習支援を目的として，LLM を用いて論理エラーを修正するための正確なアドバイスの生成方法，および，授業資料中の該当ページを検索して学習者に提示する方法を提案する．提案手法では，教員がプログラミング課題の模範解答と授業資料を用意することを前提とし，LLM によって生成した論理エラーを修正したコードが正しく動作するかを模範解答との比較を通して検証することで，正しいアドバイスを生成する．また，多くの授業課題では授業資料に課題のヒントが含まれるという経験則に基づき，アドバイスとともに参照すべき授業資料も提示する．それにより，講義における教員の負担の削減にも繋がると考えられる．

2 関連研究

本節では，LLM によるプログラミング支援に関する研究，および，LLM に対するプロンプトエンジニアリングに関する研究について紹介する．

2.1 LLM をプログラミングに活用する研究

LLM をプログラミングに活用する研究として，Jiang ら [4] は，LLM を用いてアルゴリズムを生成した後に，コードを生成する手法を提案している．具体的には，OpenAI が開発した Codex を用い，少数のアルゴリズム生成例と対象のプログラミング課題の課題文を用いてアルゴリズムを生成し，課題文と生成したアルゴリズムからコードを生成する．しかし，この手法は Python の 1 つのメソッド単位のソースコードを対象としており，複雑な課題には対応していない．次に，Chen ら [5] は，LLM を用いて生成したコードをモデル自身にデバッグさせるコード生成手法を提案している．この手法は，生成，説明，修正というステップに分かれている．具体的には，課題文からコードを生成し，そのコードをモデル自身に説明させる．その後，生成したコードが正しいかどうかを，ユニットテストを用いて判定した結果，もしくは，ユニットテストが無い場合はモデル自身に判断させた結果を用いてモデルに修正させる．

これらの研究では，LLM のプログラミングに関するコード生成性能，および，バグ修正性能の向上を目的とした手法を提案しているが，プログラミング教育に役立てることは行っていない．

2.2 LLM によってコードからアドバイスを生成する研究

LLM によってコードからアドバイスを生成する研究として，Kiesler [6] らは，間違ったコードを修正するためのアドバイスを ChatGPT に生成させ，そのアドバイスの質を調査している．具体的には，33 個の間違ったコードの修正を 3 回ずつ ChatGPT に依頼し，エラーの原因の説明，修正結果，コード等の種類に分類している．計 99 件の回答の中で，80 件以上の回答にテキストによるエラーの原因と修正に関する説明が含まれており，修正後のコードも 65 件の回答に含まれていた．しかし，61 件の回答には不正確な内容が含まれており，21 件の回答にはエラーを特定できないという内容が含まれていた．また，各コードに対して生成した 3 回の回答は大きく異なっており，ランダム性が大きいことも明らかにしている．そのため，著者らは，ChatGPT は有用な回答を行うこともあるが，フィードバックのランダム性や正確性に問題があるため，適切な指導の元で利用する必要があると結論付けている．フィードバックの正確性を向上させる研究として，Phung [7] らは，文法エラーに対して生成したフィードバックを検証する手法を提案している．具体的には，文法エラーを含むコードから複数の修正コードを生成し，その中で最も差分が小さいコードと元のコードから修正内容のフィードバックを生成する．その後，生成したフィードバックを用いてエラーを含むコードを修正し，修正後のコードとフィードバック生成時に利用したコードが一致するかによって，フィードバックの正しさを検証している．さらに，Phung [8] らは，GPT-4，GPT-3.5 モデルをそれぞれ教員と学生に見立てたフィードバック生成手法を提案している．最初に，既存研究 [7] の手法に従ってエラーを含むコードを修正し，修正したコードから教員役モデルが詳細な説明と学生に提供する簡潔なフィードバックを生成する．次に，より弱い性能の学生役モデルが，詳細な説明を用いた場合，用いない場合，それぞれについてエラーを含むコードの修正を実行する．その際，修正コードをそれぞれ複数生成し，正しい修正コードの数を比較することによって，フィードバックの正しさと有用性を検証している．しかし，この方法では 1 つのコードのフィードバックを生成するために多数のリクエストが必要となることから，API 利用における金銭的・時間的コストが高く，大人数の授業環境には適していない．

2.3 プロンプトエンジニアリングに関する研究

プロンプトエンジニアリングとは，言語モデルに入力するプロンプトと呼ばれる文書を最適化することによって，モデルの推論能力を向上させる手法である．Wei ら [10] は，大規模言語モデルの推論能力を向上させる Chain of Thoughts (CoT) という手法を提案している．具体的には，少数のタスクの段階的な解法をプロンプトに含めることで，モデルに与えるタスクを段階的に考えさせ，推論能力を向上させている．Kojima ら [11] は，解法を必要とせずに大幅に推論能力を向上させる Zero-shot-CoT という手法を提案している．この手法は，「Let's think step by step」という単純な指示をプロンプトに追加するだけであり，それにより段階的な思考を誘発し，多様な論理推論タスクの精度を向上させている．

3 提案手法

本研究では，プログラミング授業における受講者のプログラミング支援を目的とし，LLM による正確なアドバイスの生成，および，授業資料を十分に活用するための該当ページ検索手法を提案する．また，アドバイス生成は授業補助者の代替となることが期待されるため，単に高性能な手法を追い求めるだけでなく，利用料金，実行時間などの観点からの利用コストの削減も考慮する．

提案手法の概要を図 1 に示す．提案手法は大きく，アルゴリズム生成，コード修正，アドバイス生成の 3 つから構成される．アルゴリズム生成は，授業準備段階で実施することを想定し，コード修正，および，アドバイス生成は授業中に学生からコードが提出された際に発生する処理となる．図 1 における LLM-1，LLM-2 は，それぞれ利用コストが高いが性能の高い LLM，利用コストは低いが性能が LLM-1 より劣る LLM を表す．アルゴリズム生成では，利用コストの低い LLM-2 の性能向上につながるアルゴリズムを利用コストの高い LLM-1 を用いて授業開始前に生成する．コード修正に関しては，授業課題の課題文，動作例における標準入力と標準出力，および模範解答コードを入力として利用する．また，LLM が誤ったアドバ

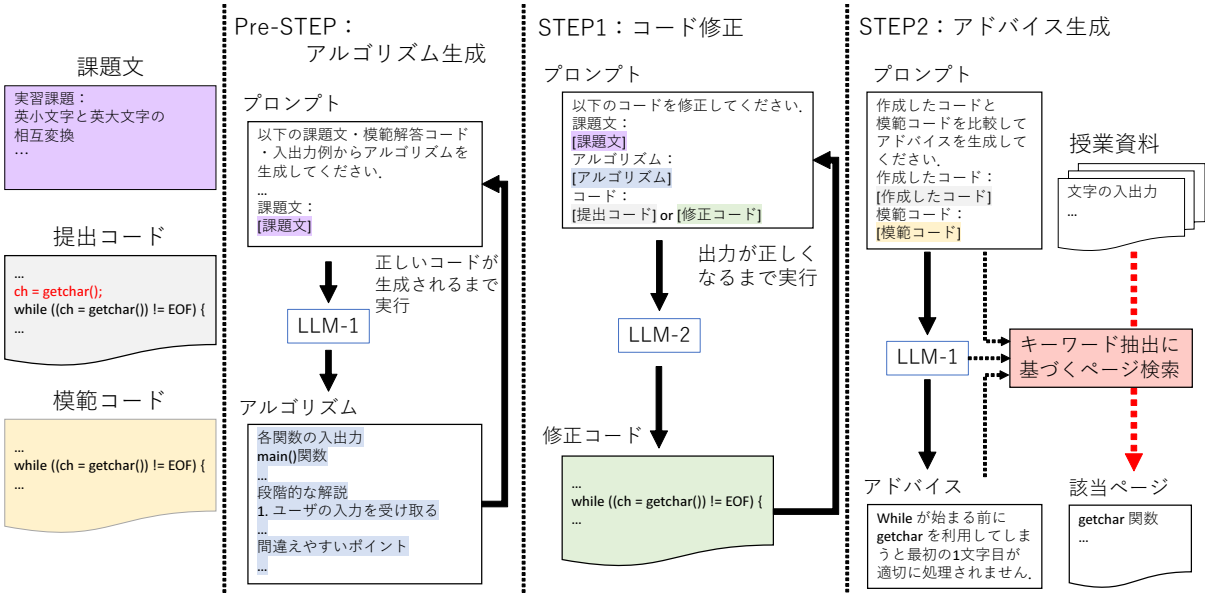


図 1: 提案手法の概要

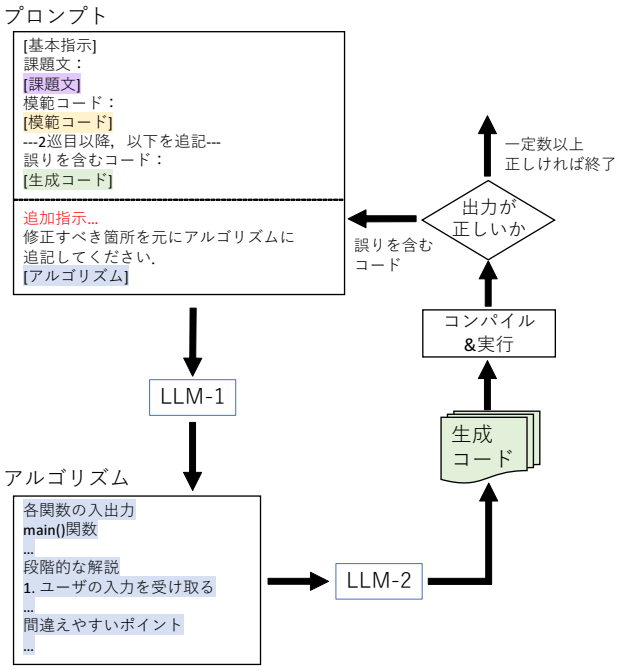


図 2: アルゴリズム生成の概要

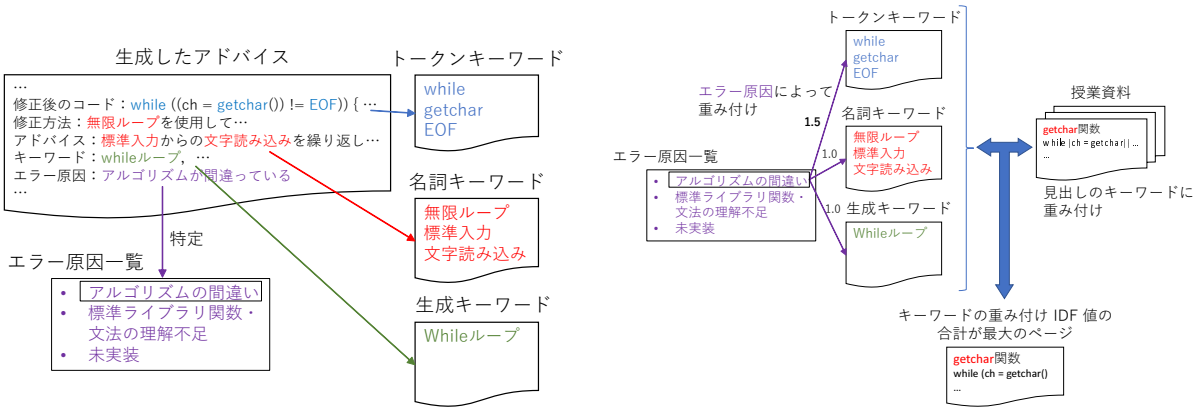
イスを生成することを防ぐために，誤りを含んだコードが適切な出力を返すように修正されるかをチェックし，出力が正しくなるまで修正コードの生成を繰り返す．そのため，ここでは利用コストの低い LLM-2 を用いる．アドバイス生成では，元の誤りを含むコードと正しく修正できたコードを LLM-1 への入力として用いてアドバイスを生成し，生成されたアドバイスから抽出したキーワードを用いて授業資料中の参考となるページを検索する．

3.1 アルゴリズム生成

GPT-3.5 のような利用コストの低い LLM は，GPT-4 のような利用コストの高い LLM と比較すると，性能は低いものの，推論のヒントを与えることで，生成結果の質を向上させることができることが知られている．そこで提案手法では，利用コストの低い LLM-2 の生成結果の質を向上させるために，利用コストの高い LLM-1 を用いて，LLM-2 のよりよいコード生成につながるアルゴリズムを事前に生成する．ここでは，各課題に対して 1 つのアルゴリズムを生成することを目的とし，アルゴリズム生成とコード生成を交互に実行し，正しい出力を返すコードが生成されるまで繰り返す．その概要を図 2 に示す．最初に，対象の課題文，動作例における標準入出力，模範解答コードを入力として LLM-1 を用いてアルゴリズムを生成する．具体的には，各関数の目的，入出力，外部への出力，段階的な解説，および，全体を通して間違えやすいポイントについてまとめさせる．次に，課題文，動作例における標準入出力，生成したアルゴリズムから LLM-2 を用いて 10 個のコードを生成する．生成したコードが一定以上正解すればそのアルゴリズムを採用する．一定以上正解しない場合は，模範解答と間違ったコードを LLM-1 に比較させ，間違えた箇所をアルゴリズムの間違えやすいポイントに追記するように再度生成させる．その際，生成した中で模範解答コードとのトークンのレーベンシュタイン距離が最大となる間違ったコードを選択して比較する．これらのアルゴリズム生成とコード生成を繰り返すことによってアルゴリズムを改善する．

3.2 コード修正

提案手法におけるコード修正について説明する．課題からは課題文と動作例における標準入出力，間違ったコードからはコードと動作例における標準出力を取得し，事前に生成したアルゴリズムとともにプロンプトを生成する．さらに，正しい標準出力と間違ったコードによる標準出力を比較し，内容が異なる最初の 1 行を抽出し，明示的にプロンプトに記載する．こうして生成したプロンプトを LLM-2 に入力し，間違ったコードに対する修正コードを出力させる．そして，修正コードが正



(a) 生成したアドバイスからのキーワード抽出　　(b) エラー原因による重み付けと授業資料の検索

図 3: 授業資料のページ検索の概要

しいかどうかを動作例の実行によって判定し，出力結果が正しければ次のアドバイス生成に進み，正しくなければ，間違ったコードに関するプロンプトの内容を修正コードのものと入れ替えたプロンプトを生成し，再度 LLM-2 を用いて修正コードを生成する．これらの操作を一定回数まで繰り返すことによって正しい修正コードを生成する．

3.3 アドバイス生成

提案手法の最終段階のアドバイス生成は，LLM-1 による自然言語によるアドバイス生成，および，その際に生成した情報に基づいた授業資料ページの検索から構成される．アドバイスは 2 段階で生成し，1 段階目では，課題文，動作例における正しい入出力，生成したアルゴリズム，間違ったコードとその出力を LLM-1 に入力し，コードを比較させ，間違えている箇所を生成させる．その際，修正箇所の説明，修正方法，および修正内容の分類（必要な修正またはリファクタリング）を生成させる．2 段階目では，1 段階目の対話に加えて，最初に修正すべき箇所の説明，修正方法，該当箇所の修正前後のコード，簡潔な 1 つのアドバイス，関連する用語,「アルゴリズムの間違い」「文法や基本的概念の理解不足」「未実装」の 3 種類の分類となるエラー原因を生成させる．このように，段階的にアドバイスを生成することで，アドバイスの正確性の向上が期待できる．

次に，授業資料のページ検索の概要を図 3 に示す．授業資料のページ検索に関しては，図 3a に示すように次の 3 つの方法で取得したキーワードを利用する．1 つ目は，LLM-1 で生成した修正方法と簡潔なアドバイスから，普通名詞，複合名詞を抽出するものである．これにより，修正に必要な概念に関する名詞キーワードを取得する．2 つ目は，生成した修正前後のコードをトークン化し，それぞれ関数，演算子，制御構文，定数を抽出し，修正後のコードに含まれるものと修正前のコードにしか含まれていないものを抽出する．このようにして抽出したトークンキーワードにより，課題のヒントとなるサンプルコードが授業資料に含まれる場合，その検索が期待できる．3 つ目は，前述のアドバイス生成 2 段階目で生成した関連する用語を利用するものである．この生成キーワードは，名詞キーワードでは取得できない関連概念に対応することが期待できる．この名詞キーワードは,「〇〇関数」のような英数字・日本語を含む場合にそれらを分割し，キーワードとして追加する．これにより，より抽象的な概念や具体的な関数名を用いたコード部分とのマッチングが期待できる．

次に，パワーポイント形式の授業資料に含まれるテキストをページごとに抽出し，その際，検出したテキストを含む図形の中心の位置が最も上となる図形のテキストをそのページの見出しとする．その後，対象とする課題の回までの授業資料において，検索するキーワードの Inverse Document Frequency (IDF) 値を計算し，図 3b に示すように 2 つの条件で重み付けする．IDF とは，文書集合内における対象単語の希少性を表す値であり，その単語を含む文書数の逆数を用いて定義される．IDF が大きいほど，その単語がより少ない文章にのみ出現することを意味する．1 つ目の条件は，キーワードがスライドの見出しに含まれているかどうかであり，含まれていればより関連するスライドだとみなし，そのキーワードを重みづける．2 つ目の条件は，ページに関係なく LLM-1 に生成させたアドバイス箇所のエラー原因に基づいたものである．具体的には，エラー原因が「アルゴリズムの間違い」であれば，授業資料の中で類似した構造をもつサンプルコードが,「文法や基本的概念の理解不足」であれば，そのような文法・概念を解説したページが参考になると考えられる．そのため，エラー原因が「アルゴリズムの間違い」の場合は，主にサンプルコードとのマッチングに向いているトークンキーワードを重み付けし,「文法や基本的概念の理解不足」の場合は，名詞キーワード・生成キーワードを重み付けする．ただし，エラー原因が「未実装」の場合は，どちらも原因である可能性が考えられるため，いずれのキーワードも重み付けしない．最後に，ページごとに含まれているキーワードの重み付け IDF 値の合計が最も高いページを検索結果として提示する．

4 評価実験

実際の大学のプログラミング授業において学生が作成したソースコードを用いて，提案手法の論理エラー修正性能を評価

表 1: アルゴリズム生成結果

課題	反復回数ごとの正解数					合計入力	合計出力	合計 API
	1	2	3	4	5	トークン数	トークン数	利用料金（$）
素数の判定	10/10	-	-	-	-	12,326	2,431	0.046
相互変換	10/10	-	-	-	-	15,700	3,136	0.058
カレンダー	3/10	4/10	8/10	-	-	68,947	22,965	0.39

表 2: コード修正結果（アルゴリズムあり）

課題	総正解数	反復回数ごとの正解数				修正コードとの	模範コードとの
		1	2	3	4	平均距離	平均距離
素数の判定	24/25	23/25	1/2	0/1	0/1	31.5	52.7
相互変換	15/17	13/17	1/4	1/3	0/2	53.1	71.8
カレンダー	10/14	3/14	4/11	1/7	2/6	231.3	304.9

表 3: コード修正結果（アルゴリズムなし）

課題	総正解数	反復回数ごとの正解数				修正コードとの	模範コードとの
		1	2	3	4	平均距離	平均距離
素数の判定	24/25	23/25	0/2	1/2	0/1	28.7	52.7
相互変換	15/17	10/17	1/7	4/6	0/2	51.3	71.8
カレンダー	3/14	2/14	1/12	0/11	0/11	168.0	301.0

した．

4.1 実験設定

評価実験では，青山学院大学理工学部情報テクノロジー学科 2 年生を対象とした C 言語を用いたプログラミング実習科目の 2022 年度受講生が作成したプログラムを利用した．ここでは，同科目で出題された課題のうち 3 つの課題に対する提出プログラムを対象とした．具体的には，標準入力で与えられた整数が素数か否かを判定する課題，標準入力で与えられた半角英字の大文字と小文字を相互に変換する課題，標準入力で与えられた年月のカレンダーを出力する課題である．素数判定の課題は，単純なアルゴリズムの実装が必要な低難度の課題，大文字・小文字相互変換の課題は ASCII コードの知識と適切な条件分岐が必要な中難度の課題，カレンダーの課題は，月の開始する曜日の計算，適切な日数での改行が必要な高難度の課題となっている．本研究では，論理エラーを対象としているため，文法エラー，実行時エラーの発生しているコードと正解コードは除外した．本実験での利用に同意した受講生から提出されたコードのうち，これらを除外した後のコード数は，それぞれ素数判定の課題が 25，大文字・小文字相互変換の課題が 17，カレンダーの課題が 14 となった．また，取得したコードは，コメントを削除し，linux の lint コマンドを用いて整形して使用した．整形後のコードの平均行数は，それぞれ 30.9 行，35.5 行，64.7 行となった．

本実験では，LLM として ChatGPT を利用し，高コストかつ高性能な LLM-1 としては GPT-4-1106-preview を用い，利用コストの低い LLM-2 としては GPT-3.5-turbo-1106 を使用した．また，ChatGPT には，設定できるパラメータの 1 つとして temperature が存在する．temperature は，モデルの出力の決定性に関わるパラメータであり，低い方がより決定性が高まり $[0.0, 2.0]$ の範囲の値を指定できる．アルゴリズムの生成，アドバイス生成では 0.0，コードの生成，修正では 0.5 を指定した．また，生成したアルゴリズムを採用する正解コード数の割合は 0.8，アルゴリズム追記の反復回数は 5，コード修正の際の反復回数は 4 に設定した．

表 4: コード修正のコスト（アルゴリズムあり）

課題	平均入力	平均出力	平均 API	平均実行時間（s）
	トークン数	トークン数	利用料金（$）	
素数の判定	1,840	618	0.0031	13.4
相互変換	3,648	792	0.0044	17.8
カレンダー	8,682	2,456	0.014	49.3

表 5: コード修正のコスト（アルゴリズムなし）

課題	平均入力	平均出力	平均 API	平均実行時間（s）
	トークン数	トークン数	利用料金（$）	
素数の判定	870	687	0.0022	15.6
相互変換	1,505	1,023	0.0036	24.0
カレンダー	4,273	2,790	0.0099	55.5

4.2 評価指標

アルゴリズム生成では，アルゴリズムの精度として生成したコードの反復ごとの正解数，コストとして全体の入出力トークン数と API 利用料金を評価した．コード修正では，修正精度として修正したコードの反復ごとの正解数，最終的な正解数を評価した．また，修正前のコードと修正後のコードおよび模範解答コードのトークンのレーベンシュタイン距離を計算することによって，ChatGPT による修正がより個人のコードに沿った修正になっているかどうかを評価した．さらに，1 コードあたりの平均入出力トークン数，平均 API 利用料金，平均実行時間を実行コストとして評価した．アドバイス生成においても，1 コードあたりの平均入出力トークン数，平均 API 利用料金，平均実行時間を実行コストとして評価し，また，アドバイスの有用性を被験者実験によって評価した．

4.3 アルゴリズム生成とコード修正の実験

アルゴリズム生成とコード修正の実験では，正解数，および，コストによる定量評価を行った．アルゴリズム生成の結果を表 1 に示す．高難度のカレンダーの課題以外は，最初に生成した段階ですべて正解となっている．カレンダーの課題で最初に生成されたアルゴリズムと最終的なアルゴリズムを比較すると，「合計日数を計算する際に、入力された年の 1 月 1 日から入力された月の 1 日までの日数を加算するが、これを二重に加算してしまうと曜日の計算がずれる。」のように追記されていた．このように，間違いやすい部分を強調することによってコードの生成・修正精度が高められていると考えられる．

次に，生成したアルゴリズムの有無によるコード修正の結果を表 2，表 3 に示す．カレンダーの課題は正解数が大幅に向上しており，それ以外の課題でも正解数は変わらないが，アルゴリズムがある場合の方が修正の反復回数が減少していることがわかる．これらの結果から，提案手法で生成したアルゴリズムがコード修正精度の向上に有効だと考えられる．また，ChatGPT で修正したコードは，模範コードより元のコードとの平均レーベーンシュタイン距離が小さくなっており，より個別のコードに沿った修正ができているといえる．さらに，コー

表 6: 標準的な手法のアドバイス生成のコスト

課題	平均入力トークン数	平均出力トークン数	平均 API 利用料金（$）	平均実行時間（s）
素数の判定	816	72	0.010	7.5
相互変換	932	76	0.012	5.9
カレンダー	1,584	94	0.018	6.8

表 7: 提案手法のアドバイス生成のコスト

課題	平均入力トークン数	平均出力トークン数	平均 API 利用料金（$）	平均実行時間（s）
素数の判定	4,762	1,082	0.080	97.3
相互変換	5,324	974	0.082	96.1
カレンダー	7,897	1,374	0.120	125.1

表 8: 標準的な手法と提案手法の評価結果

課題	正確性				情報の量	有用性
	間違いから正しい	正しいから正しい	正しいから間違い	間違いから間違い		
標準的な手法	10	0	0	0	−0.37	3.96
提案手法	10	0	0	0	−0.28	3.88

表 9: 授業資料ページの評価結果

ページに関連する概念やコードの構造が含まれているか		ページ単体の有用性	アドバイスと組み合わせた有用性
含まれている	含まれていない		
5	5	3.41	4.23

ド修正のコストをそれぞれ表 4，表 5 に示す．アルゴリズムを用いた修正のほうが，アルゴリズムを入力に含むため入力トークン数が増加し，それに応じて API 利用料金が高くなっているが，修正の反復回数が減少しているため，実行時間は減少していることが確認できる．

4.4 アドバイス生成の実験

アドバイス生成の実験では，コストの定量評価と被験者実験による評価を行った．具体的には，実験データを収集した授業において，ティーチングアシスタントとして経験のある学生 7 人を対象に被験者実験を実施し，生成したアドバイスとマッチングした授業資料ページを評価した．対象としたコードは被験者の負担を考慮して合計で 10 コードであり，各課題のデータ数に応じて素数の判定の課題から 5 コード，相互変換の課題から 3 コード，カレンダーの課題から 2 コードをランダムに選択した．具体的な評価手順は 2 段階に分かれている．1 段階目では，間違ったコード，模範コード，および，入出力例を入力としてアドバイスを生成する標準的な手法と提案手法に関して，生成したアドバイスの正確性，アドバイスとしての情報の量，および，アドバイスの有用性を評価した．この時，提示する両手法によるアドバイスはコードごとにランダムな順序で提示した．2 段階目では，マッチングした授業資料ページの評価であり，ページ内にコード修正に関連する概念およびコードの構造が含まれているか，および，検索したページ単体の提示とアドバイスと組み合わせた際の有用性を評価した．正確性の評価では，コードの状態が「間違った状態から正しい状態（必要な修正）」「正しい状態から正しい状態（リファクタリング）」「正しい状態から間違った状態」「間違った状態から間違った状態」のいずれの状態であるかを評価して貰った．ただし，標準的な手法と提案手法はどちらも最初に修正すべき箇所のアドバイスを生成させている．そのため，この正確性の分類では，アドバイス箇所のエラーを正しく指摘できているかについて注目しているため，コード全体で他に間違っている箇所が残っているかについては考慮していない．次に，アドバイスとしての情報の量は，LLM が簡単に答えを得られるという特性を考慮し，答えを教えすぎていないかという点を「不足」「やや不足」「適切」「やや過剰」「過剰」の-2 から+2 の 5 段階で評価した．最後に，これらの有用性は，どのくらい修正に役立つかを「まったく有用ではない」「あまり有用ではない」「どちらともいえない」「やや有用である」「とても有用である」を 1 ～ 5 の 5 段階で評価した．また，実験の最後に全体を通してアドバイスおよび授業資料ページの良かった点，改善すべき点についての自由記述を用意した．

アドバイス生成における標準的な手法と提案手法のコストをそれぞれ表 6 と表 7 に示す．提案手法の方が，標準的な手法より入力トークン数が 5 倍以上，出力トークン数が 12 倍以上，API 利用料金が 8 倍以上，実行時間が 13 倍以上かかっている．これは，標準的な手法が 1 回の入出力でアドバイスを生成しているのに対して，提案手法では，一度間違い箇所の一覧を出力し，さらにその入出力を入力として対象とする箇所のアドバイスを生成しているためである．また，提案手法では，授業資料の検索に利用するためにアドバイスの他に修正方法，キーワード，エラー原因といった情報も出力しており，さらにコストが増加している．

次に，標準的な手法と提案手法の評価結果と提案手法のマッチングした授業資料ページの評価結果を，それぞれ表 8 と表 9 に示す．正確性，ページに関連する概念およびコードの構造が含まれているかの評価は，コードごとに被験者の分類結果を多数決で決定しており，情報の量と有用性はそれぞれ被験者の評価結果の平均値を記載している．実験の結果，正確性の評価では，どちらの手法もすべてのコードで間違った状態から正しい状態という結果になっており，標準的な手法でも十分に正しいアドバイスを生成できると考えられる．次に，情報の量はどちらの手法も少し不足しており，過剰にアドバイスを与えすぎてはいないといえる．最後に，提案手法の授業資料とアドバイスを組み合わせた際の有用性は，各手法のアドバイスのみ，および授業資料ページのみに比べて向上しており，標準的な手法に対して t 検定を行った結果，有意差（$p < 0.05$）があることを確認した．そのため，生成したアドバイスとともに授業資料ページを提示することは，アドバイスのみを提示するよりも有用だと考えられる．

次に，提案手法のアドバイスと授業資料ページを提示した際の評価が最も高かった（平均 5.0）事例について考察する．この事例では，相互変換の課題で図 4 のように while ループの条件式が間違っているコードに対して,「代入と比較の優先順位に注意して、条件式を見直してください。」というアドバイスと

```c
#include <stdio.h>
int lower2upper(int c);
int upper2lower(int c);
int main()
{
    int ch;
    while (ch = getchar() != EOF) {
        if (ch == 10)
            putchar(ch);
        else if (97 <= ch && ch <= 122)
            printf("%c", (char) lower2upper(ch));
        else if (65 <= ch && ch <= 90)
            printf("%c", (char) upper2lower(ch));
        else
            putchar('*');
    }
    return 0;
}
```

図 4: 最高評価のアドバイスのコード

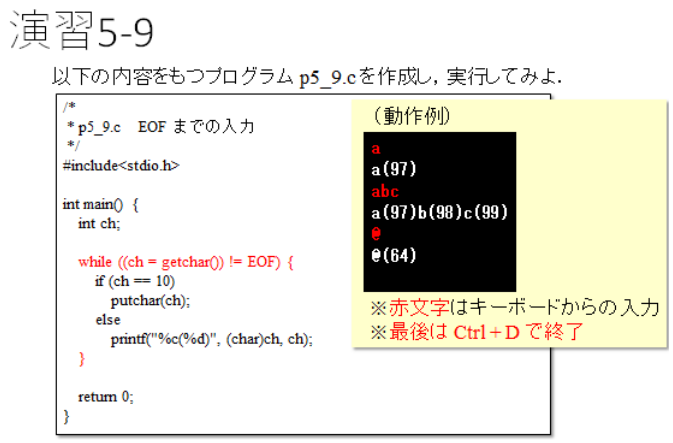


図 5: 最高評価の際に提示されたページ

```c
#include<stdio.h>
int isprime(int i);
int main()
{

}

int isprime(int n)
{
    int m, ans;
    for (n = 1, m = 2; m != n; m++) {
        ans = n % m;
        if (ans == 0) {
            return 0;
        } else {
            return 1;
        }
    }
}
```

図 6: 最低評価のアドバイスのコード

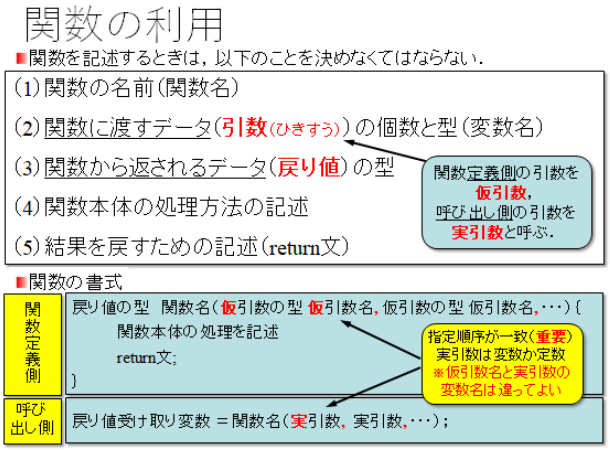


図 7: 最低評価の際に提示されたページ

図 5 のような授業資料ページを提示している．このページには，正しいループ構造の例が含まれているため，高評価となったと考えられる．また，同様に最も低かった事例では，素数の判定の課題で図 6 のように関数名が課題の指定と異なっているコードに対して，「関数の宣言と呼び出しで使用されている名前を確認し、課題文の指定に合わせてください。」というアドバイスと図 7 のような授業資料ページを提示している．低評価となった原因として，課題の指定とは異なるが被験者が普段指導する際に重視する箇所ではないこと，および，isprime 関数の引数 n を関数内で上書きしているといったさらに重大な論理エラーを指摘できていないことが評価を下げる原因になったと考えられる．ただし，提示されたページは関数の使い方に関するページであり，アドバイスに関連した内容を提示することはできている．最後に，最も正確性の評価が分かれており，さらに，提案手法のアドバイスのみの評価が標準的な手法より最も下がった事例について考察する．この事例では，図 8 のようにカレンダーの課題のコードに対して，標準的な手法では，「カレンダーの日付を出力するためのループ処理が不足しています」というアドバイスを提示している．提案手法では，「入力された年と月が指定された範囲内にあるかを確認し、範囲外であれば再入力を促すループを追加する。」というアドバイスと図 9 のような授業資料ページを提示している．提案手法のアドバイスの評価が最も分かれた原因として，模範コードでは存在するが入出力例には示されていない入力の検証機構がリファクタリングだと考える被験者も多かったことが考えられる．また，提示された授業資料は，scanf() 関数から受け取った標準入力を条件ごとに処理するコードになっており，アドバイスとの関連は，標準入力によって分岐するといった部分的なものになっている．

4.5 自由記述の考察

自由記述では，アドバイスの良かった点として，「直接的に答えを教えるのではなく考えさせるアドバイスがあってよかった」といった意見があった．ただし，改善すべき点として，「記述を促すだけのアドバイスが参考にならない」「何回もやりとりする必要がある」といった意見もあり，ヒントの量が少なすぎるアドバイスがあることを指摘していた．そのため，ヒントの量が少ないと感じた際にインタラクティブに LLM にヒントを生成させるような手法が必要だと考えられる． また，授業資料ページ良かった点として，「ベースとなる知識を提示できていて良かった」や「コードの枠組みが分からない人にコード例を提示できていて良かった」という意見があった．しかし改善すべき点として，「まったく関係ないページが迷いを生じさせる可能性がある」といった意見や「コード例のページの際にどこに注目すれば良いか分かりにくい」といった意見があった．このこ

```c
#include <stdio.h>
int is_leap_year(int year);
int get_days(int year, int month);
int main()
{

    int year, month, day, a, i = 5;

    printf("西暦を入力してください（2000年以降） : ");
    scanf("%d", &year);
    printf("月を入力してください : ");
    scanf("%d", &month);

    printf("%d年%d月¥n", year, month);
    printf("Sun Mon Tue Wed Thu Fri Sat¥n");

    day = get_days(year, month);

    return 0;
}
```

図 8: 最も分類が分かれたアドバイスのコード

演習2-9

以下の内容をもつプログラム p2_9.c をフォルダ 2nd 内に作成し，実行してみよ．

```c
/*
 * p2_9
 */
#include <stdio.h>

int  main() {
   int  score;

   printf("成績を入力してください：");
   scanf("%d", &score);
   if (score >= 90) {
      printf("AA¥n");
   } else if (score >= 80) {
      printf("A¥n");
   } else if (score >= 70) {
      printf("B¥n");
   } else if (score >= 60) {
      printf("C¥n");
   } else {
      printf("来年頑張ってください¥n");
   }
   return 0;
}
```

図 9: 最も分類が分かれたアドバイスとともに提示されたページ

とから，マッチングした授業資料ページをさらに LLM に入力として使い，LLM に関係あるページかどうか判断させる手法，どこに注目すべきかといったより充実したアドバイスを生成させる手法などが考えられる．

5 ま と め

本研究では，論理エラーを含む学生のプログラムコードに対する修正アドバイスを，LLM を用いて生成する手法を提案した．提案手法では，間違ったコードに対する修正コードを生成し，そのうえでアドバイスを生成する．提案手法におけるコード修正の性能を評価した結果，課題に対して事前に生成したアルゴリズムが，コード修正において正解数の向上，反復回数の減少に寄与することが明らかになり，それにより 7 割以上のコードを修正できることを確認した．また，アドバイス生成に関しても，被験者実験の結果から授業資料をアドバイスとともに提示することが有用だと明らかになった．今後の課題として，授業資料の検索精度の向上，アドバイス生成の 1 段階化によるコスト削減が挙げられる．提案手法では，テキスト情報とスライドの見出しの情報しか用いていないが，授業資料では，強調したい情報は文字色の変更，フォントサイズの変更などの操作が行われることが多い．そのため，抽出したテキストの文字色やフォントサイズを用いることによって，授業資料のマッチング精度の向上が期待できる．また，今回標準的な手法と提案手法のアドバイスの正確性は変化がなかった．そのため，1 段階でのアドバイス生成による実行コストの削減が考えられる．

本研究の一部は，一般財団法人ホワイトロック財団の補助による．

文 献

[1] NTT ラーニングシステムズ株式会社．文部科学省委託事業次世代の教育情報化推進事業『平成 30 年度教育委員会等における小学校プログラミング教育に関する取組状況等について』の調査．https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/05/28/1417283_002.pdf.

[2] Andrew Ettles, Andrew Luxton-Reilly, and Paul Denny. Common logic errors made by novice programmers. In *Proceedings of the 20th Australasian Computing Education Conference (ACE)*, p. 83–89, 2018.

[3] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt.

[4] Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. Self-planning code generation with large language models. arXiv preprint arXiv:2303.06689, 2023.

[5] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. arXiv preprint arXiv:2304.05128, 2023.

[6] Natalie Kiesler, Dominic Lohr, and Hieke Keuning. Exploring the potential of large language models to generate formative programming feedback. arXiv preprint arXiv:2309.00029, 2023.

[7] Tung Phung, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. Generating high-precision feedback for programming syntax errors using large language models. arXiv preprint arXiv:2302.04662, 2023.

[8] Tung Phung, Victor-Alexandru Pădurean, Anjali Singh, Christopher Brooks, José Cambronero, Sumit Gulwani, Adish Singla, and Gustavo Soares. Automating human tutor-style programming feedback: Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. arXiv preprint arXiv:2310.03780, 2023.

[9] Md. Mostafizer Rahman and Yutaka Watanobe. Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, Vol. 13, No. 9, p. 5783, 2023.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903, 2022.

[11] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 22199–22213, 2022.

能動的学習データを用いた学習状況把握のための可視化分析手法の研究

王 笑難† 孫 一‡ 大月 一弘† 清光 英成†

†神戸大学大学院国際文化学研究科 〒657-8501 神戸市灘区鶴甲 1-2-1
‡神戸情報大学院大学情報技術研究科 〒650-0001 神戸市中央区加納町 2-2-7
E-mail: †217c303c@stu.kobe-u.ac.jp, †{ohtsuki, kiyomitu}@kobe-u.ac.jp, ‡sun@kic.ac.jp

あらまし 教育機関で実施されるオンライン授業は対面授業と同等以上の教育効果が求められている．学生の能動的活動（質問や学習ノートなど）を学習状況把握に利用できれば，学習状況に応じたフィードバックならびに学習意欲などのエンゲージメント向上のための教育支援に有用と考えられる．本研究では，オンデマンドプログラミング授業において，能動的学習データと動画教材とを紐づける教育支援システムを開発し，学習活動を教員に提示するための可視化手法を提案する.大学プログラミング講座で実験を行い,収集した能動的学習データの分析より，ある学習活動と学習効果との相関が発見され，学生の学習戦略や学習方法の相異も現れた．考察より，個々学生の進捗状況や理解度を把握する手法が証明され，動画講義の改善と，個性的な指導を提供する可能性が示された．

キーワード 情報可視化，e-Learning，プログラミング教育，学習支援

1. はじめに

近年，オンライン形式などの遠隔授業の普及は教育実践に大きな影響を及ぼしている．教育機関で実施される遠隔授業は対面授業と同等以上の教育効果ならびに教育の質が求められる．リアルタイム型の遠隔授業は，無料もしくは安価な遠隔会議システムが利用され学習場所の多様化を許容する．オンデマンド型では学習場所に加えて学習時刻の束縛も緩和される．

しかしながら，対面授業と同等の臨場感を求めることには限界がある．例えば，遠隔授業では教員が学生の表情などの変化から感じ取る種類のフィードバックを得にくい．学生は近くを通りかかった教員に気軽に質問をするという従来型の簡易質問ができない．さらに，オンデマンド型は学習時刻・時間が一様でないためリアルであれば同時に得られたはずのフィードバックが学習者個々の時間軸上に分布する．このように対面授業との授業形態，学生と教員のコミュニケーション方法などの違いに起因する問題が顕在化してきた．

遠隔環境におけるコミュニケーション低下より，学生の進捗や理解度など学習状況の把握は重要な課題になっている．そのため，オンライン学習プラットフォームや学習管理システム（LMS）などの教育支援システムを利用し，オンラインでの学習データから学習行動や学習効果などに分析する研究は多くなさられている[1, 2]．しかしながら，従来の異なる教科の講義形態は知識や技能について教授する効率はよく，学習の効果は学生のエンゲージメントや成績に限らず，学習戦略や教育コンテンツとのインタラクションなど，もっと複雑な要素によって影響される[3, 4]．

プログラミング教育は，など実践的な学習を重視されている．従来プログラミング授業において，学生と教員のコミュニケーションは多様な方法がある．言語で説明しにくい場合，学生が画面を指差す「これ」とか「ここ」という曖昧に見えるような意思伝達は実は非常に有効である．教員は指したところによって，問題を特定することができる．しかし，オンデマンドでの授業は，非言語で伝わることが期待できないので，何らかの方法で補う必要がある．

オンデマンドプログラミング授業において，簡易なコミュニケーション方法の提供より，学生の進捗や理解度など学習状況を把握できれば，学習状況に応じたフィードバックならびに学習意欲などのエンゲージメント向上のための教育支援に有用と考えられる．そこで本研究は,「どのようなデータが学生側から出てくればいいのか，どのようなデータなら学生側から出しやすいのか」，また「これらのデータなら教員側から如何に読むのか」という一連の問いに対して，本研究は一つ教育支援策を提案する．

本研究は，学生が自ら積極的な意思伝達活動の記録（質問や学習ノートなど)を能動的学習データという．先行研究は，オンデマンドプログラミング授業（講義動画から学習)において,能動的学習データに着目し，講義動画の時点と学生の質問や学習ノートとを紐づけることにより，状況を特定する記述をほとんど必要としないメモ・質問機能をもつ教育支援システムを開発した[5]．大学プログラミング入門講座での実験より，学生の質問の難易度を大幅に下げ，学習ノートは教員により多くのフィードバックを提供した．

本稿では「これらのデータなら教員側から如何に読むのか」という問題について，学習状況を効率的把握する可視化分析手法を紹介する．可視化より，学習行動（質問と学習ノート）と学習成果（成績）の相関と，学習ノートの取り方やパターン相違よる学生の潜在的な理解度の違いを分析する．また，提案する手法は，動画教材の改善と，個性的な指導を提供する可能性を議論する[16]．

本研究の貢献を以下にまとめる：

- 学生に質問しやすい・メモ機能付きの動画学習支援システムを提供した．
- 教員により多くのフィードバックを与え，学習活動を視覚的に再現し，効果的な分析法を提供した．
- 講義動画の学習活動の分類方法の提案より，学習活動と学習成果の関係を発見した．
- 提案する可視化分析手法より，学生の学習活動の違いを発見し，学生を特定することが可能となり，動画教材の改善と，個性的な指導に関する貴重な洞察を提供した．

本稿の構成は，次のとおりである．第２章では，関連研究について述べる．第３章では，システムの概要と実験について説明する．第４章では，能動的学習データを可視化する手法を紹介する．第５章では収集した能動的学習データの分析を行い，学生の学習活動と学習効果との相関分析より，学生グループを特定する．第６章では，考察を行い，動画教材の改善と個性化指導について議論する．最後に，まとめと今後の課題について述べる．

2. 関連研究

教授方法の提案やオンライン通信システム利用の授業コミュニケーションの改善などさまざまな研究がなされている[6, 7]．オンデマンド型授業において，学習活動の観察は困難であるため提出物の評価を介して知識・技能の習得過程を推測することは多く，提出物だけで学習状況の把握が対面授業と同等にできるわけではない．

そこで，学習管理システム（LMS）などの教育支援システムを利用することで，オンライン学習者のアクセスログや操作履歴などを分析し，教育評価に関する研究は多く行われている．学習者を分析するために使われる行動研究は，オンライン教育で広く使われている．Anderson ら[8]の研究は，MOOC における学習者の学習方式を理解するため，個々の学習行動の分類手法を提案し，異なる成績の学習者の行動を調べた．しかしながら，ログや行履歴などの受動的な学習データを用いた分析が学習状況の把握には限界があり，学習者の直接な考えや，過程中のトラブルなどを発見することは困難である．

学習活動において，学習者による質問やノートなどの活動は，より能動的な学習表現である．Gasevic らの研究[9]は，オンライン学習環境において，質問が学習者の学習意欲と成績の強力な予測因子であることを判明した．Makany ら[10]や Bonner ら[11]の研究では，学習者の情報理解を向上させため，学習ノートの役割が強調した．ノートテイキングから，学習方法の不備や自己調整学習戦略の弱さを明らかにした．学習効果の違いは学習方法と学習戦略など様々な要素に影響されている．

学習データの読み取りは適切な方法が必要である．学習プロセスの可視化は既に教育実践に応用されている．武田ら[12]の研究は，オンライン授業のコンテンツ構造を分析する．教育コンテンツの構造をトピックモデルによって推定し，トピックと学習活動の傾向との関係を可視化する方法を提案した．Xia ら[13]は，オンライン学習からのインタラクションデータを利用して問題解決行動を分析より，詳細な問題解決の軌跡を検査し，異なる学生グループを比較し下上，学生が問題を解決する際に遭遇する困難を反映する図表を通して可視化する視覚分析システムを提案した．

また，講義動画などの教育コンテンツを利用する場合，学習者が動画コンテンツとどのようにインタラクションするのかを理解するのも一つ重要な課題である．Kim ら[14]と Shi ら[15]研究は，学習者が動画を視聴中の操作データ（再生，一時停止，スキップ，見直すなど）を利用し，学習者のクリックアクションデータから学習行動を分析するための視覚分析システムを開発した．

自動的に収集される学習情報に基づいた手法について，客観的な学習者の状態や学習タスク達成度などの把握が可能になったが，内容の理解度や応用への着想など，学習者が実感した主観的な自己評価の把握には至っていない．

本研究行った調査で，オンデマンド授業中の学生はプログラミングに関する質問に際し，状況の説明に難しさを感じていることが分かった．学習者の能動的活動（質問や学習ノートなど）を学習状況把握に利用できるようになれば，学習状況の判明に応じる教育の質向上が期待できる．そこで本研究では，オンデマンドプログラミング授業において，学生が動画学習での質問・学習ノートなどの能動的学習データを促進・利用する教育支援システムを開発した．また，これらのデータを教員に有効に読み取るための可視化手法を提案した．

3. 提案手法と実験

3.1 システム概要

本研究はプログラミング学習に特有の動作などを参考にし，オンデマンド型授業に適応した教育システムを開発した．

システムは JavaScript で開発し，ブラウザ上で動作するア Web プリである．学生は個人 ID でログインし動画を視聴しながら質問やノートを記入する．システムは,学生から能動的データと当時動画の画面(時点)を紐つけ，時系列的なデータを集約し，可視化手法より有用的な学習情報を変換して教員に提示する．

従来対面で行う「ここ」とか「これ」などの意思伝達を動画の画面（時点）に紐付けることより，単元や状況を特定する記述をほとんど必要とせずに，教員へ直接質問を送信する．本章では，教員に学習状況の把握のため，これらのデータを如何に表示するのかについて，可視化の工夫を紹介する．

3.2 使用データ

予備実験は，システムを導入し，大学で実施する初心者向けの Python 基礎講義で行った．授業はオンデマンド形式で，学生が講義動画から学習する．受講生は 38 名であった．

今回可視化と分析に用いるデータは，実験での約 90 分程度の三つの講義動画から収集し，38 名受講生のうち，26 名学生の 56 件質問と 303 件の学習ノートであった．

3.3 提案手法

予備実験の結果は，質問の数が大幅に増加し，学生が「質問し易い」「メモ機能便利」などの好評を得た．教員側は教育フィードバックとしての質問と学習ノートの大量獲得には喜びだが，それを一つずつ確認するのは非常に時間がかかり，大変さを感じていることも確認した．これらのデータを視覚的学習活動に再現できるようになれば，学習状況の把握には効果的なの方法であると考える．

そこで本研究は，能動的学習データよる学習状況把握のための可視化分析手法を提案する．それぞれの時間帯においての質問・学習ノートが動画のタイムラインに一つを集約し，質問・学習ノートの活動分布図を作る．また，動画内容の特徴に基づいて，トピック・教授法の分布を可視化し，質問やノートの分布を重ね比べることで，学習活動に問題発見・異常検出などができるようになる．

可視化結果より，異なる種類のノートと，学習効果（テスト成績）の相関を分析する．これより，学習状況の診断や,より効果的なノート教育資料の参加方法,学生の潜在的理解度の違い（学生グループの特定）などの教育向上のための有用な手掛かりとする．

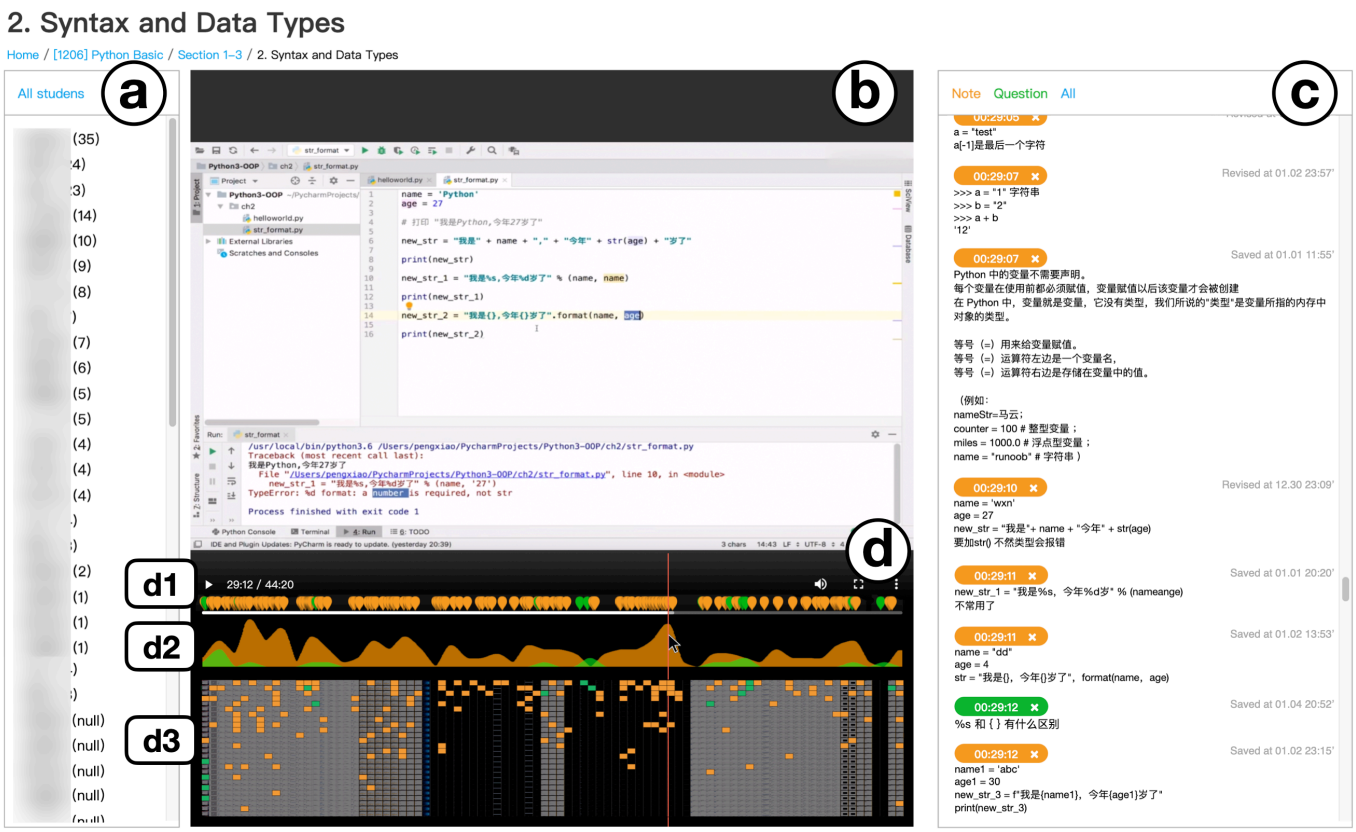


図 1． 教員インタフェース

3.4 教員インタフェース

教員インタフェース画面を図１に示す．左側に学生一覧フォーム（a）である．学生一覧の氏名リストは，学生からの質問とノート件数の多い順（参加度）から掲載される．氏名をクリックすると，個々学生の質問とノートが表示フォーム（c）に表示される．さらに表示フォーム（c）上の「Note」・「Question」・「All」ボタンと組み合わせて利用することで，「質問のみ」「ノートのみ」などを絞り込み，フィルターで表示できる．また，ミドリ（質問）やオレンジ（ノート）のタグをクリックすると，プレヤーフォーム(b)動画の画面を，そのタグ対応の時点にスキップする．画面と共に確認を行う．

可視化フォーム（d）には，３つの可視化方法を設置し，動画のタイムラインに基づき，動画内発生したすべての質問とノートを再現する．動画のタイムライン(d1)では，質問とノートのピークが直接標記されている．(d2)では，一線のデータ密度を面積で表示され，質問やノート集中発生のピークを現す．活動分布図(d3)では，学生ごとのノート・質問が動画内でどのように分布しているかを示している．横軸は同様時間軸であり，縦軸は(a)と同じく質問とノート数の降順に，上から学生を並べている．背景としての枠組みは，動画教材のサムネイルを，時間間隔は 30 秒に一列とした．ノートがあった時点をオレンジ色で，質問があった時点をミドリで表している．

3.5 活動分布図の考察

活動分布図は横の一行を見ると，学生一人の状況を分かる．縦に見ると，クラス全体がこの時点の状況が分かる．図 2（h）のところには，ミドリの山が現れ，複数の学生はほぼ同じ時点に質問を行った．質問文の確認より，ここの質問は「環境設定」の問題であることが分かった．

ノート文の確認より，エリア（ｅ）のノートは「整数・浮動小数」のトピックである．また，エリア（f）のトピックは「文字列」，（g）トピックは「配列」である．これらは，講義動画のトピックに対応している．動画で新しい知識のトピックを現れたら，それに応じるノート山のピークがある．学生は動画の内容を正確に理解していることが分かった．

しかし，同じくノートのピークだが，(e)(f)より，（g）のところのノートは，横に密集されている．ここにあるノートは，ほんの少数の学生によるものである．このような異常は，「少人数しか理解していないのか」，「ここでの説明は不十分なのか」など，教員が実際に注意すべきところである．

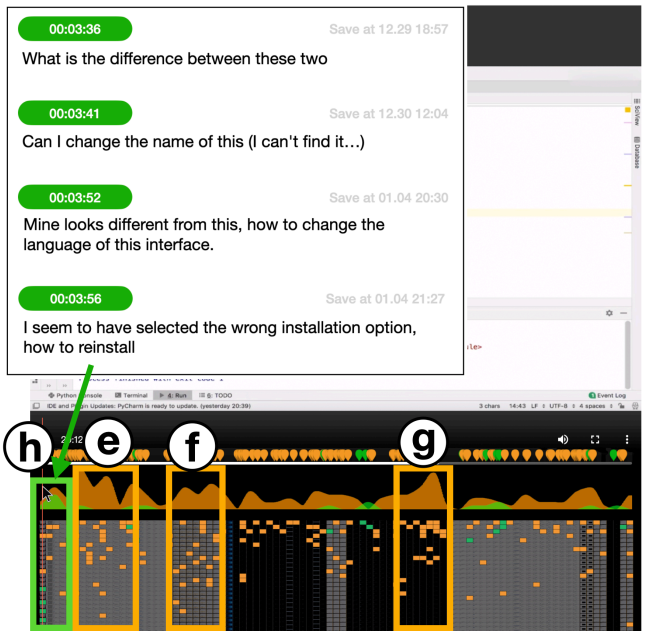


図２．質問とノート密集のところ（問題文の例）．

活動分布図の一瞥で，個々学生の学習活動としての質問とノートの分布と，動画の内容の変化（背景）と，その時間軸とともに確認できる．各学生の状況及び全体的な参加度の変化を視覚的再現し，参加度の変化や質問頻出などの異常なシーンが捕えやすくなった．

4. 活動分布図の可視化

質問と学習ノートの分布から，学習活動の参加度と問題点を検出できたが，分布上のルールや異常などの判明は，文章内容を確認しなければ分からない．そこで，活動分布図より再設計を行い，図表から学習過程が診断できる方法を紹介する．

4.1 特徴可視化

動画コンテンツの特徴を明確にし，画面，教授法，要点三つの Map を作成した（図３）．画面 Map（図３の a）は図１の d３の背景であり，三十秒ごとの画面を抽出ちゅうしゅつした一列に並ぶ．動画作った教員であれば，画面 Map から，例えばこの黒い部分は何なの

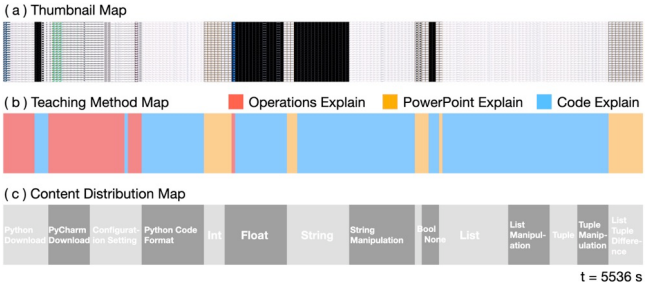


図３．動画特徴可視化マップ．

かがすぐわかれる．教授法 Map（図 3 の b）は，動画中教員どういう方式で説明しているのかを分類した．動画中教員がパソコンを操作する説明する部分を赤に，資料で説明の部分を黄色に，コーディングしながら説明する部分を青色にする．要点 Map（図 3 の c）では，動画中に各トピックに当たり部分に色分けで表示する．

一方，質問と学習ノートの内容特徴に基づいて分類することにより，学習者が学習過程での認知活動について詳細に理解することができる．ここで学習ノートを「Mark Note」,「Summary Note」,「Code Note」,「Word Note」 4 つ種類に分けた．

図 4 では，4 種類ノートの分布を示している．「Mark Note」は，学生が動画にタグ付けのようなものであり，「ここは○○です」というような表記である．「Summary Note」は，講義内容を自分の言葉で要約したものである．「Code Note」の中身は，コードを主としており，注釈などの説明も付いている．また，内容のほとんどは 1 つか 2 つの単語に構成し，他人にはよく理解していないものを「Word Note」に分類される．

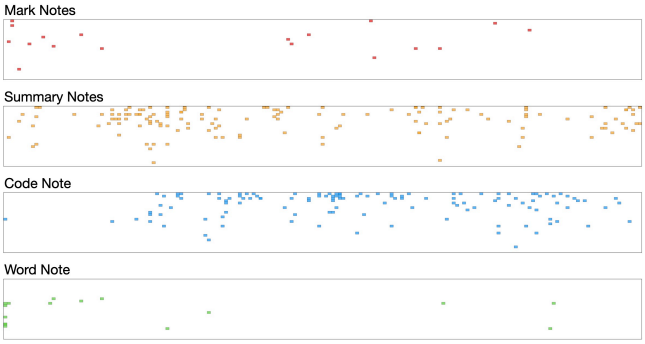


図 4．四つのノート種類の分布．

同様，学生の質問を「What」,「Why」,「How」,「Can」と「Comment」五種類に分類した．これら異なるタイプの質問と学習ノートの分布図は，動画特徴マップとの比較分析に使用される．

4.2 使用例の考察

質問の分布を要点 Map と教授法 Map に重ね合わせ，図 5 に示す．質問と要点 Map (図 5 の d)は，クラス全体の質問活動（共通問題）はどの要点どのトピックに多いのはすぐ分かれる．ここでは先頭の環境設定の部分と後半のリスト操作の部分に，質問が集中していることが観察された．

質問の分布と教授法 Map((図 5 の e)を比較すると，教員が資料にしたがって説明する際には，学生からの質問がほとんどなかったことが注目される．これは，この部分には学生たちよくわかっていて,「いい説明でした」,「学生よく受け入れた」ということを示唆した．

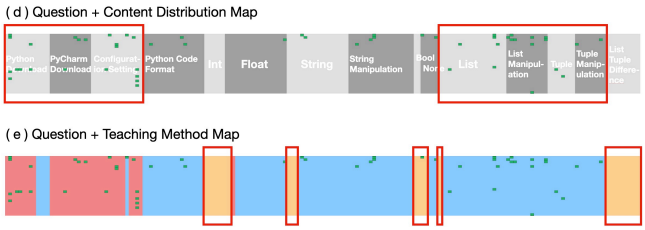


図 5．要点 Map(d)と教授法 Map(e)における問題の分布．

図 6 は，要点 Map（図 6 の h）における 5 種類の質問の分布を示す．教員は各トピックに存在する問題を特定することができる．図 6 の i は，環境設定部分にあたり質問の種類である．ここで，最も多いのは「How」タイプの質問であることがわかった．これは，ここが理解の問題ではなく,「どうする」とういう操作の問題であることを示唆した．

このように，分類より，教員は学生が書いた質問の具体的な内容を確認しなくても，一瞥で学生の主な問題を判断することが可能である．

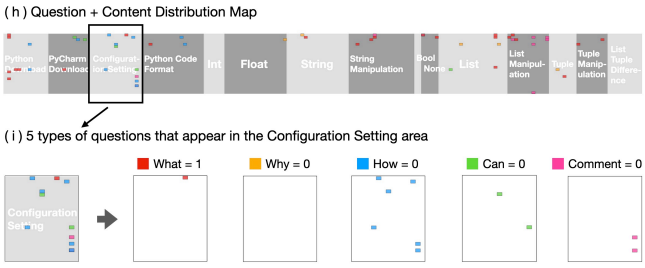


図 6．要点 Map における 5 つの質問の分布．

可視化方法は，異なる教授方法（資料の説明，操作から教授）の効果が示している．教員がこれらの情報を利用し，講義動画を修正したり，指導方法を調整したりすることが可能となる．

5. 分析

活動分布図から発問したところやノートを取った時点は，個人によって相違があることは，観察より分かった．学習方法の相違は，学習戦略や潜在的理解度の違いなどの影響要素が含まれていると考えられる．本章では，能動的学習データを学習効果に比較し，学習活動の特徴から学習者を特定する分析方法を説明する．

4.1 成績による統計

図 7 では，実験で個々学生の質問とノートの数とその成績の変化を示す．横軸は 26 名の学生（A-Z）を成績高い順で並んでいる．質問はミドリ，ノートはオレンジ色の棒で，縦軸はその数を示す．

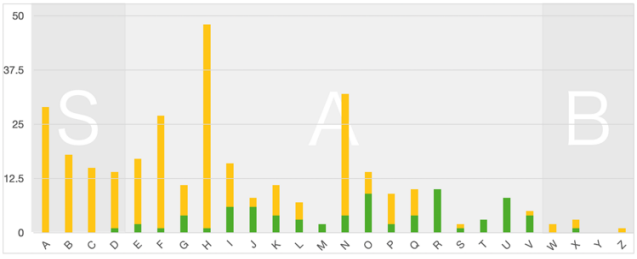


図 7　成績高い順で 26 名学生の質問とノートの数.

図 7 によると，先頭成績 S の学生数（A-D）は少ないが，ノートを取る傾向が良い．しかし，質問はほとんどしない．成績 A の学生人数は一番多く（E-V），質問も主に A の学生からのである．成績 B の学生（W-Z）はノートも質問もほとんどしない（今回実験データでは成績 B 以下の学生のノートと質問の投稿数 0 であるため，統計対象外にする）．

学生の能動的活動は講義動画とのインタラクションと参与方法が反映される．質問はほぼ成績 A の学生が投稿したことで，成績による変化が捕らえない．したがって，学生ノート活動に焦点にしている．ノート数多ければ多いほど，成績は良くなるのも一般的だが，例外である学生もいる（H と N）．

4.2 ノートの種類と成績の関係

成績 S・A・B のグループによって，学生が好むノートの種類も異なることがわかる．図 8 では，各成績グループにおける，第四章で分類された 4 種類のノートの比率を示す．

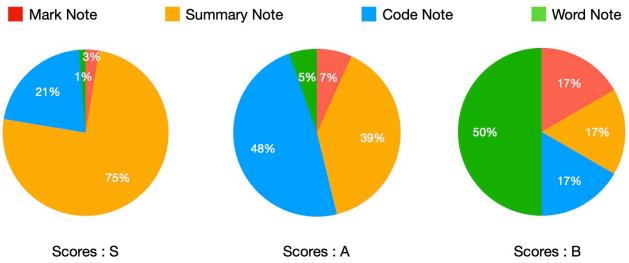


図 8．成績 S・A・B より 4 つの種類のノートの割合.

成績 S の学生（15%）では，Summary Note の比率は最も多く，Summary Note 総数の 39%を占めている．Code Note の数も一定だが，Word Note はほとんどしない．成績 A の学生（69%）では，Summary Note より直接的に Code Note に集中している．Mark Note と Word Note も S の学生より増える．成績 B の学生のノート数が少ないだが，図 8 の各成績グループ内のノート種類の比率から，成績 S と A の学生の特定の種類に集中より，成績 B の学生のノート種類はより混雑している．

学習ノート（動画参加方式）の分類より，違い学習効果の学生は，ノートの種類に対する好みも異なる可能性を示した.その相関性をさらに確認するためには，多項式回帰分析を行い，その結果は図 9 に示す．

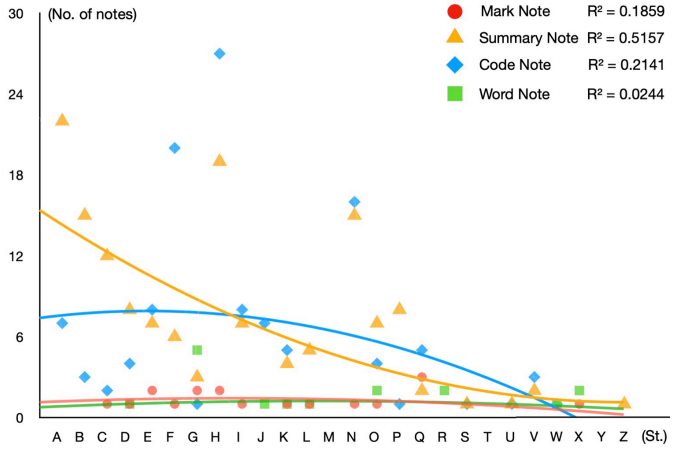


図 9．成績順位によるノート種別の多項式フィット.

図 9 では，26 名の学生（A-Z）の成績順位（高から低へ）を横軸にし，各種類のノートの数を縦軸にした二次多項式のフィットを示す．異なる種類のノートを各形・色のマークに表し，それぞれの理論曲線を表示する．

図 9 により，Mark Note（赤）と Word Note（緑）では，フィット曲線が低く，あまり変化がないで，使用頻度が低い,成績への影響がないことを示唆している．Summary Note（黄）と Code Note（青）では，成績順位が下がるにつれて減少しており，高得点を取る学生ほど多くの Summary Note を取っており，成績と強い正相関（R^2=0.5157）を示している．

しかし，Code Note の理論曲線は，良い成績の区間（S-A）に，成績順位による変化が小さくなる．一定の数に満たす以上，Code Note を多く取るのも成績への影響が少ないことを示唆している．これについては，プログラミング学習において，コードの解釈は基本的なスキルである．しかし，学習が進めば進むほど，アルゴリズムなどのロジックの思考が，コードから独立していく傾向がある．Code Note 理論曲線の傾向がそれを示唆しているだと考える．

4.3 学生グループの特定

図 10 には，外れ値を取り除いた Summary Note の理論曲線を示す．各点から曲線までの残差などの位置関係は，学生個人によって，今までの成績を取るために費やした努力と学習効果を反映している．

例えば，理論曲線の上部にいる学生（a）は，現在の成績に至るまで，より多くの努力をした．曲線の下部にいる学生（b）は，有効的参加度は高くないのに，良い点数を達成した．

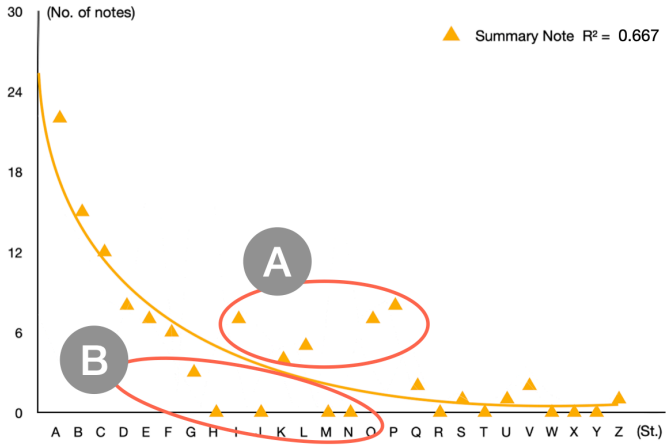


図１０．Summary Note の理論曲線（外れ値除き）.

学習者の特性に基づいて,教員はより正確な介入を行い，教材や教授戦略の調整ができるようになる．例えば，(b) グループの学生には，より興味深い内容や有効的参加方法（Summary Note）を勧め，この部分の学生の学習意欲を高める．また，(a) グループの学生には，努力を承認した上，より理解しやすい資料，あるいは詳細な説明を提供する．

このような学習介入ができれば，教育コンテンツを個々学生の学習状況により適切に合わせ，最終的に学習効果を向上させることが期待できる．

6. 考察

本研究は，動画講義における学生の能動的活動（質問と学習ノート）に焦点に当たり，学習状況を把握するため可視化分析手法を提案する．学習活動と学習効果の分析より，有益な結果が得られた．

学生の質問と学習ノートの分析は，学生の活動と成績との相関関係は確認された．分類より Summary ノートと成績の間には強い正相関があることが示され，学生の知識の理解と応用を効果的に高めることができることを示唆している．学生にこの種のノートをより多く取るよう奨励するか，学生に主要概念や知識点を要約するタスクを設定することで，これらを指導設計に取り入れることができる．しかし学習効果では，活動の数だけでなく,より複雑の要素にも影響されている．例えば，今回良い成績の学生が質問より，ノートの好みがあることは，学生の主体性や先行する関連経験にも関連していると考える．

これらの結果は，個々学生の学習方法や学習戦略を反映し，学習者特性の発見・理解するための重要な示唆を提供した．教育者は，これらの分析結果に基づいて,例えば以下のような解決策を講じることができる．

1. 既存の教材を維持する：特定のトピックで学生が高い理解度と教材への積極的な関わりを示している場合，教育者は複雑度と内容のレベルを維持し，教材と主題が学習を効果的に促進する．
2. 学習教材の追加：質問が増えていることろは，理解を深めるための追加教材や練習問題を提供し，難解な領域に対処する．
3. 高度な内容の導入：一貫して教材を理解していて，質問が減少している場合，教育者はより高度なトピックや課題を導入し，学習意欲を促進する．
4. 細かい内容を省略する：学生が一貫して高い理解度を示している部分は，最小限に扱われる．より複雑あるいは挑戦的なトピックのための時間が確保されることになり，コースの構造と内容を最適化し，関連性が保たれ興味が引かれるようにすることができる．
5. 基礎概念の再訪：特定のトピックを理解するのに苦労している学生に対しては，これらのトピックが構築されている基礎概念を再訪することが有益である．

これにより，学生は基礎をしっかりと把握し，より複雑な教材を理解しやすくなると考える．

おわりに

本研究では，オンデマンドプログラミング授業における学生の能動的学習活動を用いた学習状況把握のための可視化分析手法を提案した．学生の質問と学習ノートの分析により，学生個々の学習過程や進捗状況を効果的に把握し，学習者特性の判別や個性化指導を行うための根拠を提供した．分析結果は，学習行動と学習成果の相関関係を明らかにし，学習方法の違いによる学生の潜在的な理解度の差異を捉えることが可能であった．これにより，教材の改善や教育戦略の最適化に役立つ洞察が得られた．

本研究の成果は，遠隔教育における学習状況把握と個別化された教育支援の分野において重要な貢献をしている．今後の研究では，この手法をさらに多様な学習環境や学習者に適用し，その汎用性と効果を検証することが求められ，より多くの実験データに基づく分析を行うことが必要である．教育内容や学習環境の改善を続き，遠隔教育における学習支援の質を向上させることが期待される．

謝辞

本研究の一部は科研費(19K03000)「プログラミング教育のための進捗把握手法」，科研費(19K03030)「コピペ時代のプログラミング学習方法の開発」の支援による．ここに記して謝意を表す．

参 考 文 献

[1] D. Monk, "Using Data Mining for e-Learning Decision Making," *Electronic Journal of E-learning,* vol. 3, no. 1, pp. pp65-81-pp65-81, 2005.

[2] E. N. Ogor, "Student academic performance monitoring and evaluation using data mining techniques," in *Electronics, robotics and automotive mechanics conference (CERMA 2007)*, 2007: IEEE, pp. 354-359.

[3] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard, "Who does what in a massive open online course?" *Communications of the ACM,* vol. 57, no. 4, pp. 58-65, 2014.

[4] K. Stephens-Martinez, M. A. Hearst, and A. Fox, "Monitoring moocs: which information sources do instructors value?" in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 79-88.

[5] X. Wang, Y. Sun, and Y. Su, "A study of interaction design for E-learning system based on distributed asynchronous communication," Chinese CHI 2022*,* in press.

[6] 内田いづみ. "反転授業の授業外学習を支援する学習環境の設計と効果: 情報基礎科目における Moodle 活用." 駿河台大学教職論集 (2017): 26-35.

[7] 守一雄. "Microsoft Teams を使ったオンライン授業での小グループに分かれてのディベートの実践報告." 教育総合研究 4.special (2020): 249-256.

[8] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 687-698.

[9] D. Gašević, V. Kovanović, S. Joksimović, and G. Siemens, "Where is research on massive open online courses headed? A data analysis of the MOOC Research Initiative," *International Review of Research in Open and Distributed Learning,* vol. 15, no. 5, pp. 134-176, 2014.

[10] T. Makany, J. Kemp, and I. E. Dror, "Optimising the use of note-taking as an external cognitive aid for increasing learning," *British Journal of Educational Technology,* vol. 40, no. 4, pp. 619-635, 2009.

[11] J. M. Bonner and W. G. Holliday, "How college science students engage in note-taking strategies," *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching,* vol. 43, no. 8, pp. 786-818, 2006.

[12] T. Takeda, Y. Hayashi, K. Shigeta, H. Mori, D. Kaneko, H. Yagi, and T. Nagashima, "Visualizing Relationships among Contents Topics and Learning Activities of Online Courses," in *EdMedia+ Innovate Learning*, 2018: Association for the Advancement of Computing in Education (AACE), pp. 1355-1356.

[13] M. Xia, R. P. Velumani, Y. Wang, H. Qu, and X. Ma, "Qlens: Visual analytics of multi-step problem-solving behaviors for improving question design," *IEEE Transactions on Visualization and Computer Graphics,* vol. 27, no. 2, pp. 870-880, 2020.

[14] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks inonline lecture videos," in Proceedings of the first ACM conference on Learning@ scale conference, 2014, pp. 31-40.

[15] C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from massive open online courses," in *2015 IEEE Pacific visualization symposium (PacificVis)*, 2015: IEEE, pp. 159-166.

[16] X. Wang, Y. Su, Y. Sun, K. Ohtsuki, H. Kiyomitsu and K. Shigyo,"Visualization of potential differences in comprehension by distribution of notes and questions in online programming courses," TALE 2023, in press.

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 10:00 ~ 12:10 | T5-B オンライン（Zoom Events)

交通

座長:シリアーラヤ パノット(京都工芸繊維大学)
コメンテータ:北山 大輔(工学院大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T5-B-4-01] 複数台のバスに関する特徴量を活用した都営バスのリアルタイム運行データを用いた渋滞検知

*青柳 宏紀[1]、藤田 智也[1]、畠中 希[2]、小口 正人[2]、山名 早人[1] (1. 早稲田大学、2. お茶の水女子大学)

10:30 ~ 10:55
[T5-B-4-02] 自転車走行中の表情と振動による潜在的快適性分析の検討

*山口 琉太[1]、村重 圭亮[2]、矢野 英人[2]、義久 智樹[3]、河合 由起子[1,2]、下條 真司[4] (1. 京都産業大学、2. 大阪大学、3. 滋賀大学、4. 青森大学)

10:55 ~ 11:20
[T5-B-4-03] 強化学習による差動二輪車制御における内外乱適応

*門垣 幸樹[1]、大島 裕明[1] (1. 兵庫県立大学)

11:20 ~ 11:45
[T5-B-4-04] A Forecasting Method of Traffic Flow Using Road Network Data

*田中 悠[1]、笠松 大佑[2] (1. 創価大学大学院理工学研究科笠松研究室、2. 創価大学大学院理工学研究科)

11:45 ~ 12:10
[T5-B-4-05] 都営バスのオープンデータおよびディープラーニングを用いた渋滞検知手法の提案と評価

*畠中 希[1]、青柳 宏紀[2]、藤田 智也[2]、山名 早人[2]、小口 正人[1] (1. お茶の水女子大学、2. 早稲田大学)

複数台のバスに関する特徴量を活用した都営バスのリアルタイム運行データを用いた渋滞検知

青柳 宏紀† 藤田 智也† 畠中 希‡ 小口 正人§ 山名 早人¶

† 早稲田大学大学院 基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
‡ お茶の水女子大学大学院 人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1
§ お茶の水女子大学 理学部情報科学科 〒112-8610 東京都文京区大塚 2-1-1
¶ 早稲田大学 理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: † ¶ {aoyagih, tfujita, yamana}@yama.info.waseda.ac.jp, ‡ § {hatanaka.nozomi, oguchi}@is.ocha.ac.jp

あらまし 我が国の渋滞による経済損失は年間 12 兆円に上る．対策の一つは渋滞情報の利用者への提供であり，プローブ車両の GPS データや道路上の車両感知器を用いた渋滞検知が実現されている．これに対し，本研究では，利用者がよく利用する道路（バスが定期運行する道路）を対象に，低コストで安定した渋滞検知を目指す．具体的には，オープンデータとして公開されているバスの運行データ（バス停発車時刻データ）と機械学習を組み合わせた渋滞検知手法を提案する．本手法では，停留所や信号での停車時間を把握できないため，1 台分のバスのデータを用いるだけでは渋滞検知精度が低下する．この課題に対し，本研究では，直近に同じ区間を走行した複数台のバスにおける，①速度の平均値， ②速度の標準偏差， ③運転間隔の標準偏差 を特徴量として追加し，渋滞検知精度向上を図る．評価実験では，都営バスの運行データを用いて渋滞の多い明治通り（池袋～渋谷間）の渋滞検知を行った．最も高い F1 スコアを達成したランダムフォレストでは，直近 3 台分のバスのデータをもとに算出した上記特徴量を追加することにより，全バス停区間のマイクロ平均 F1 スコアが 0.743 から 0.748 へと向上した．

キーワード 渋滞検知，オープンデータ，バス，公共交通機関，モビリティ，スマートシティー，機械学習，分類

1. はじめに

国土交通省によれば，我が国の渋滞による損失は，経済損失が年間 12 兆円（2002 年）[1]，時間換算が年間 40 時間／人（2013 年）[2]と推計されている．特に，交通渋滞が引き起こす物流への支障や，排気ガスによる大気汚染，緊急車両の遅延は，社会に対して悪影響をもたらしている．そのため，運転手が渋滞を回避できるよう渋滞情報をリアルタイムに提供することが重要である．

先行研究[1]~[6]では，プローブ車両の GPS データを用いた渋滞検知が可能であることが示されており，現在，Google Map や NAVITIME 等の交通アプリでは，ユーザから取得した GPS データを活用した渋滞検知が実装されている．一方で，アプリを起動しながら走行する車の数が少ない場合は，渋滞検知精度が低くなるという問題点がある．また，GPS データ以外にも，我が国では道路上の車両感知器を用いた渋滞検知が日本道路交通情報センター（JARTIC）により提供されている．しかし，この手法は実現コストが高いため，渋滞検知対象路線は交通量の多い県道以上の幹線道路に限られている[3]．一方，近年では公共交通機関に関するデータがオープンデータとして公開され，様々な形で利活用されている．公開されているデータの一つに，路線バスのリアルタイムのロケーションデータがある．路線バスは定期的に運行しているため，バスの走行データがあれば，市民の生活に直結した道路において，低コストで安定した渋滞検知が可能になると考えられる．

著者ら[7]は，2022 年に，バスの運行データと機械学習を組み合わせた自動渋滞検知の手法を提案し，池袋から渋谷までの明治通りの区間を対象として，一定時間ごとに「渋滞」と「非渋滞」の 2 値分類を行った．その結果，特定の停留所区間においては，F1 スコア 0.742 を達成した．また，2023 年には，[7]の手法を応用し，藤田ら[8]の転移学習を活用した手法や，畠中ら[9]のバスの運行特性を考慮した手法が提案されている．また，海外では，2022 年に，Bannur ら[10]が，公共交通に関する世界標準のデータ形式である GTFS（General Transit Feed Supplement）準拠のバス運行データセットを用い，渋滞検知に活用可能なデータセットを作成した．しかし，作成したデータセットを活用した渋滞検知の評価実験はなされていない．

上記の先行研究[7]~[9]では，バスの速度に関して 1 台分のデータのみを特徴量として用いており，当該バスが停留所や信号で長時間停車するとデータにノイズが多く含まれるため，渋滞検知精度が低下する．そ

[1] https://www.mlit.go.jp/common/000043136.pdf
[2] https://www.mlit.go.jp/common/001098855.pdf
[3] https://www.vics.or.jp/faq/faq_other.html

こで本稿では，複数台のバスに関する特徴量を活用した手法を提案する．具体的には，直近に同じ区間を走行した複数台のバスの速度と運転間隔に関する特徴量を追加することで，渋滞検知精度の向上を図る．

本稿では次の構成をとる．2 節で関連研究を紹介し，3 節で提案手法を説明し，4 節で評価実験について説明する．最後に，5 節で本稿をまとめる．

2. 関連研究

2.1 GPS プローブデータを用いた渋滞検知

Xu ら[1]は，2012 年，複数台のバスの平均移動時間を算出し渋滞検知を行う手法を提案した．Xu らは，T 分間に渋滞検知対象道路を通過したバスの平均移動時間を算出する「T-window Average」と，道路区間を直近に通過したN台のバスの平均移動時間を算出する「N-window average」を定義し，それぞれの指標が予め個別に設定した閾値を超えた場合に，渋滞が発生していることをシミュレーションにより示した．

Wang ら[2]は，2013 年，1 台の車両からの GPS プローブデータを用い，機械学習にて交通状態を分類する手法を提案した．Wang らは，GPS によって取得した車両の平均速度を位置ごとに離散化し，特徴ベクトルを作成した．学習モデルは，ランダムフォレスト，Adaboost，SVM を用いており，それぞれ 0.9159，0.8943，0.8786 の精度で検知できることを示した．

Carli ら[3]は，2015 年，バスの GPS プローブデータによる効率的な自動渋滞検知手法を提案した．バスの平均速度やバスが閾値を超えた速度で走行した時間，といった指標を計算し，都市部内の車両の動きを監視することで，信号機故障の異常を検出した．

Wang ら[4]は，2019 年，タクシーの GPS プローブデータを用いた渋滞検知手法を提案した．Wang らは，北京市交通管理局から入手したタクシーの GPS 座標からなる行列データから，非負値行列因子分解を用いて交通パターンを抽出し，渋滞検知を行った．評価実験では，Recall 0.92，Precision 0.74，F1 スコア 0.82 を達成した（Accuracy の記載は無い）．

Saraiva ら[5]は，2021 年，バスの GPS プローブデータを用いた渋滞検知手法を提案した．Saraiva らは，特徴量として，一定時間内に渋滞検知対象区間を通過したバスの台数・系統数・平均速度と，一定時間内に取得した GPS レコードの総数を用いた．評価実験では，リオデジャネイロ市を走行するバスの 1 分毎の GPS データを用い，6 種類の機械学習分類アルゴリズムにより渋滞検知を行った．6 種類の機械学習分類アルゴリズムのうち，最も F1 スコアが高かったアルゴリズムは Decision Tree であり，Accuracy 0.8077，Precision 0.6640，Recall 0.6350，F1 スコア 0.6235 を達成した．

Makara ら[6]は，2023 年，バスの GPS データと増分学習を用いた渋滞検知手法を提案した．Makara らは，バスの速度の長期的な傾向を学習した LSTM モデルに，渋滞検知を行う時間帯付近の短期的なバスの速度傾向を増分学習させ，渋滞検知精度向上を図った．評価実験では，ハンガリーのペーチ市を走行するバスの GPS データを用い，Accuracy 0.8237 を達成した．

2.2 バスの運行データを用いた渋滞検知

現在，2.1 で示した GPS を用いた渋滞検知は，Google Map や NAVITIME 等の交通アプリで実装されている．一方で，アプリを起動しながら走行する車の数が少ない場合は，渋滞検知精度が低くなるという問題点がある．これに対し，路線バスは定期的に運行しており，多くのバス路線で走行データが公開されている．バスの走行データがあれば，市民の生活に直結した道路において，安定した渋滞検知が可能になると考えられる．

そこで，著者ら[7]は，2022 年，都バスの運行データ（バス停発車時刻）を用いたリアルタイム渋滞検知手法を提案した．著者らは，都バスの停留所発車時刻のデータから，各バス停区間のバスの平均速度と同平均速度から算出される統計量を特徴量として抽出し，機械学習による渋滞検知を 20 分間隔で行った．評価実験では，都バスが走行する明治通り（池袋～渋谷間）において，バス停区間毎に渋滞検知モデルを構築し，特定の区間において F1 スコア 0.742 を得た．

一方，[7]の手法では，渋滞データの少ないバス停区間において渋滞検知精度が低下した．そこで，藤田ら[8]は，2023 年，この問題点の克服を目的として，転移学習を活用した手法を提案した．藤田らは，類似する渋滞傾向を持つバス停区間をクラスタリングし，クラスタ毎に構築した学習モデルをもとに，バス停区間毎に転移学習を行い，渋滞検知精度向上を図った．評価実験では，都バス 16 系統のバス停区間を 8 つのクラスタに分類し，クラスタ毎に学習を行った．最も高い渋滞検知性能を示したクラスタでは，バス停区間毎のマイクロ平均 F1 スコア 0.881 を達成した．一方，最も低い渋滞検知性能を示したクラスタでは同スコアが 0.179 となった．また，全バス停区間に対しては，同スコアが 0.431 となった．

畠中ら[9]は，2023 年，バスの運行特性を考慮した手法を提案した．具体的には，「バスが定刻より早く運行している場合は，定刻に合わせるため意図的に速度を落として運行する」という特性に着目し，バスが定刻より 10 秒以上早く停留所を出発したか否かという情報を特徴量として追加した．[7]と同じ区間（明治通りの池袋～渋谷間）の全てのバス停区間に対し 1 つの機械学習モデルを構築して渋滞検知を行ったとこ

ろ，F1 スコアが 0.594 から 0.606 へ向上した．

また，海外においても，バスの運行データを用いた渋滞検知を目指す研究が存在する．Bannur ら[10]は，2022 年，バスの運行データから，渋滞予測に活用できるデータセットを作成した．Bannur らは，公共交通に関する世界標準のデータフォーマットである GTFS（General Transit Feed Supplement）に準拠したインドのプネ市のバスのデータから，バスの速度，停留所間の距離，2 つの停留所間のバスの移動回数と所要時間，道路の混雑度を計算する手法を提案した．データセットには 366 のルート，5,624 の停留所，21,804 の停留所間の移動が含まれている．しかし，渋滞検知の評価実験は行われていない．

2.3 関連研究まとめ

プローブ車両の GPS データにより，渋滞検知が可能となることがいくつかの論文[1]~[6]で示されている．一方で，プローブ車両が少ない場合は，渋滞検知精度が低くなる [2]．GPS データ以外にも，我が国では道路上の車両感知器を用いた渋滞情報が提供されている．しかし，車両探知器設置には高いコストがかかるため，設置は幹線道路に限られている．

これに対し，路線バスは定期的に運行しており，多くのバス路線で走行データが無料公開されている．バスの走行データがあれば，市民の生活に直結した道路において，低コストで安定した渋滞検知が可能になると考えられる．先行研究[7]~[9]では，我が国の多くのバス路線で提供されている「バス停発車時刻」データを用い，1 台のバスのデータから渋滞検知を目指した．一方で，バスが複数台走行している道路では，複数台のデータを用いることで，渋滞検知精度の向上が可能である．そこで本稿では，[8]で提案されている特徴量に加え，複数台のバスに関する特徴量を追加することで，渋滞検知精度の向上を目指す．

3. 提案手法

3.1 概要

著者ら[7]の，バスの各停留所の発車時刻データを用いた渋滞検知手法をもとに，バスが停留所や信号で長時間停車した際の渋滞検知精度低下に対処するための手法を提案する．なお，渋滞検知では，国家公安委員会の渋滞定義[4]に基づき，渋滞検知区間を走行する自動車が「時速 10km 未満で継続的に走行している状態」を渋滞と定義する．

提案手法では，連続する 2 つのバス停の出発時間の差をバスの移動時間とし，2 つのバス停間の距離を移動時間で割ることで，バスの速度を求める．ここで注意すべきは，同移動時間には「バス停での停車時間」や「信号による停車時間」が含まれている点であり，1 台のバスの速度のみで渋滞／非渋滞を判断が難しい点である．そこで，[8]で提案されている特徴量に加えて，複数台のバスに関する特徴量を用い，精度向上を目指す．新たに追加する特徴量は，直近に同じ区間を走行した複数台のバスにおける，①速度の平均値，②速度の標準偏差，③運転間隔の標準偏差 の 3 つである．①は，バスが停留所や信号で長時間停車し，1 台分のバスの速度のみでは正確な渋滞検知できない場合に，より長期間の交通状況を学習する上で精度向上に寄与すると考えられる．②は，渋滞発生・解消時に値が大きくなると考えられ，渋滞検知に有用であると考えられる．これは，渋滞が発生し始める時間帯または解消し始める時間帯にバスの速度が大きく増減するためである．③も，渋滞発生時に値が大きくなり，渋滞検知に有用であると考えられる．これは，バスは非渋滞時にはほぼ等間隔で運行するが，渋滞時にはダイヤが乱れて発車間隔が不規則になるためである．

上記に加えて，渋滞の発生率に影響があると考えられる，曜日，五十日[5]かどうか，雨天かどうか，の特徴量も追加する．

次に，算出した特徴量と該当するバス停区間の渋滞／非渋滞を示す正解データを用いて学習を行う．なお，本研究では渋滞検知の最小単位をバス停 2 区間とする．これは，バス停区間 1 区間を対象とした場合，300m 程度の短区間に対する検知となり，当該区間内に存在する信号待ち時間やバス停停車時間等のノイズの影響が大きく渋滞検知が困難となるためである．

3.2 特徴量抽出

本稿で用いる記号を表 3.1 と図 3.1～3.2 に示す．

表 3.1 特徴量抽出に用いる記号

記号	定義
b_i	i 番目のバス
p^j	$(2j-1)$ 番目のバス停
s^j	p^jからp^{j+1}の j 番目の区間
l^j	s^jの距離[m]
td_i^j	b_iのp^jの出発時刻
Δtd_i^j	b_iのs^jの走行にかかる所要時間[s]
v_i^j	b_iのs^jでの平均速度[m/s]
$B_{i,n}^j$	s^jを走行する，b_iから数えて直近n台のバス集合（$=\{b_i, b_{i-1}, \dots, b_{i-n+1}\}$）
I_i^j	p^jを出発する 2 台のバスb_i・b_{i-1}の発車間隔（$=td_i^j - td_{i-1}^j$）
c_i^j	b_iがp^jを出発する時刻を 20 分毎に分類した指標．0 時 0 分 0 秒から 0 時 19 分 59 秒を 0 として，以後 20 分おきに 1, 2, 3, …と定義．

※バスb_iは出発時刻順に，バス停p^jは経路順に整列しているとする．

[4] https://www.npa.go.jp/bureau/traffic/seibi2/shinsei-todokede/jouhou/guideline.pdf

[5] 15 日や 30 日のように，毎月「5」や「0」がつく日のこと．企業の決済日や給料日に設定されていることが多く，関連車両の行き来が活発になり，渋滞が多く発生すると言われている．

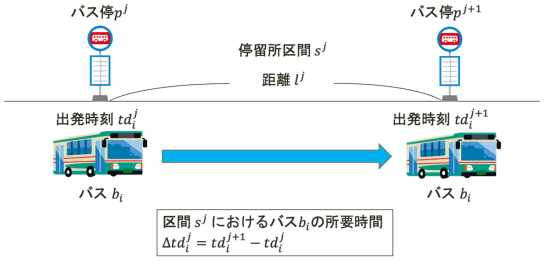


図 3.1 特徴量抽出に用いる記号の図解 1

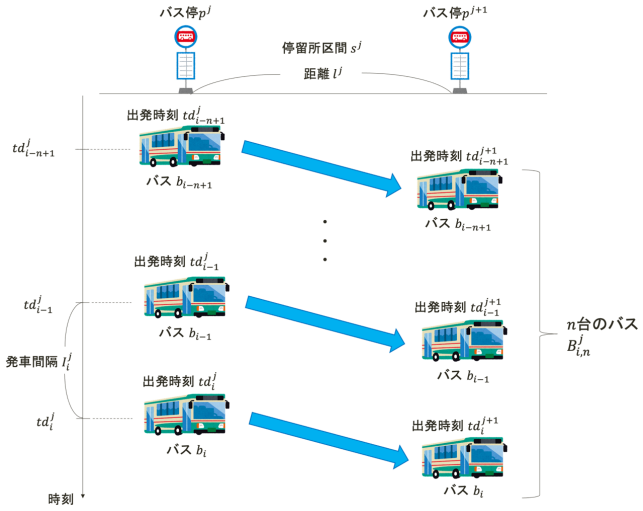


図 3.2 特徴量抽出に用いる記号の図解 2

本手法で用いるデータは，バスb_iがバス停p^jを時刻td_i^jに出発したデータである．また，各区間s^jの距離l^jは正解データ取得元である Google Directions API[6]から取得する．Δtd_i^jは出発時刻td_i^jとtd_i^{j+1}から次式(1)で表される．

$$\Delta td_i^j = td_i^{j+1} - td_i^j \quad (1)$$

次に，バスb_iの区間s^jにおける平均速度v_i^jを算出する．v_i^jは次式(2)で表される．

$$v_i^j = \frac{l^j}{\Delta td_i^j} \quad (2)$$

以下，$B_{i,n}^j$の台数が 3 台以上 $(n \geq 3)$ の場合を考える．n台のバス$B_{i,n}^j$の区間s^jにおける平均速度$\overline{v_{i,n}^J}$を算出する．$\overline{v_{i,n}^J}$は次式(3)で表される．

$$\overline{v_{i,n}^J} = \frac{v_i^j + v_{i-1}^j + \cdots + v_{i-n+1}^j}{n} = \frac{1}{n}\sum_{k=0}^{n-1} v_{i-k}^j \quad (3)$$

$\overline{v_{i,n}^J}$を用いて，n台のバス$B_{i,n}^j$の速度の標準偏差$\delta_{v_{i,n}^j}$を算出する．$\delta_{v_{i,n}^j}$は次式(4)で表される．

$$\delta_{v_{i,n}^j} = \sqrt{\frac{1}{n}\sum_{k=0}^{n-1}\left(v_{i-k}^j - \overline{v_{i,n}^J}\right)^2} \quad (4)$$

次に，p^jを出発する2 台のバスb_i, b_{i-1}の発車間隔I_i^jを算出する．I_i^jは次式(5)で表される．

$$I_i^j = td_i^j - td_{i-1}^j \quad (5)$$

I_i^jを用いて，n台のバス$B_{i,n}^j$の発車間隔の平均値$\overline{I_{i,n}^J}$と，標準偏差$\delta_{I_{i,n}^j}$を算出する．$\overline{I_{i,n}^J}$は次式(6)で，$\delta_{I_{i,n}^j}$は次式(7)でそれぞれ表される．

$$\overline{I_{i,n}^J} = \frac{I_i^j + I_{i-1}^j + \cdots + I_{i-n+2}^j}{n-1} = \frac{1}{n-1}\sum_{k=0}^{n-2} I_{i-k}^j \quad (6)$$

$$\delta_{I_{i,n}^j} = \sqrt{\frac{1}{n-1}\sum_{k=0}^{n-2}\left(I_{i-k}^j - \overline{I_{i,n}^J}\right)^2} \quad (7)$$

また，バスb_iのバス停p^jの出発時刻td_i^jを 20 分ごとの時間帯c_i^jに分類する．c_i^jの値は，0 時 0 分 0 秒から 0 時 19 分 59 秒を 0 として，以後 20 分おきに 1, 2, 3, …と定義する．

本手法で用いる特徴量を表 3.2 に示す．

表 3.2 本手法で用いる特徴量

特徴量名	内容
v_i^j	バスb_iのバス停区間s^j走行時の速度
v_{i-1}^j	一つ前のバスb_{i-1}の同一バス停区間s^j走行時の速度
$\overline{v_{i,n}^J}$	$B_{i,n}^j$の速度の平均値
$\delta_{v_{i,n}^j}$	$B_{i,n}^j$の速度の標準偏差
$\delta_{I_{i,n}^j}$	$B_{i,n}^j$の発車間隔の標準偏差
c_i^j	バスb_iの区間s^j走行時の時間帯
dw_i^j	バスb_iの区間s^j走行時の曜日
$is5nDay_i^j$	バスb_iの区間s^j走行時，五十日かどうか
$isRainy_i^j$	c_i^jに雨が降ったかどうか

v_i^jとv_{i-1}^jはそれぞれ区間s^jにおけるバスb_iの速度，バスb_{i-1}の速度を表す．バスの速度は渋滞判定の基準となる交通速度との相関があると考えられ，特徴量として有用である．

また，$B_{i,n}^j$を，区間s^jを走行するb_iから数えて直近n台のバスと定義したとき，$\overline{v_{i,n}^J}$は$B_{i,n}^j$の平均速度を表す．バスは，停留所や信号で長時間停車する場合があるため，1 台分のバスの速度であるv_i^jやv_{i-1}^jのみでは正確な渋滞検知をする上で不十分であると考えられる．そこで，渋滞が長時間にわたって継続する性質があることに注目し，複数台のバスの平均速度である$\overline{v_{i,n}^J}$を特徴量として加えることで，精度の向上を目指す．

[6] https://developers.google.com/maps/documentation/directions

また，$\delta_{v_{i,n}^j}$は，$B_{i,n}^j$の速度の標準偏差を表す．$\delta_{v_{i,n}^j}$の値は渋滞時のうち，特に渋滞発生直後と渋滞解消直前で大きくなると予測されるため，特徴量として有用であると考えられる．

$\delta_{I_{i,n}^j}$は，$B_{i,n}^j$の発車間隔の標準偏差を表す．バスは，非渋滞時にはほぼ等間隔で運行するが，渋滞時にはダイヤが乱れて発車間隔が不規則になる．そのため，$\delta_{I_{i,n}^j}$の値は渋滞発生時に大きくなると予測され，特徴量として有用であると考えられる．

ここで，$\overline{v_{i,n}^J}, \delta_{v_{i,n}^j}, \delta_{I_{i,n}^j}$の 3 つの特徴量は，$n$によって様々な値をとる．$n$の値を大きくしすぎると，渋滞検知の対象となる現在の交通状況とは乖離した，過去の時間帯のデータが多く含まれてしまうおそれがある．本手法では，nの値は，$n = 3,4$の中から全バス停区間のマイクロ平均 F1 スコアが最も高くなるような値を，パラメータチューニングにより選ぶ．

c_i^jはバスb_iがバス停p^jを出発する時間を，20 分ごとに分類した指標である．また，dw_i^jはバスb_iがバス停p^jを出発する曜日を表す特徴量である．時間帯や曜日により，交通量の変化があると考えられるため，c_i^jやdw_i^jは特徴量として有用であると考えられる．なお，dw_i^jはカテゴリ変数であるため，one-hot encoding により数値化してから学習に用いる．

$is5nDay_i^j$は，バスb_iがバス停p^jを出発する日が五十日かどうかを表す二値の特徴量である．脚注 5 で示した通り，$is5nDay_i^j$も特徴量として有用であると考えられる．

$isRainy_i^j$は，時間帯c_i^jに雨が降ったかどうかを表す二値の特徴量である．雨が降った場合，路面状況や視界の悪化による渋滞への影響や，バスの利用者増加に伴うダイヤへの影響が考えられる．よって，$isRainy_i^j$は特徴量として有用であると考えられる．

3.3 データ前処理

バスは停留所に停車するため，一般車両より速度が遅い．そのため，バスの速度v_i^jが渋滞判定基準の 10km/h を超える場合は，一般車両の速度も 10km/h を超えるため，学習器を使用せずに非渋滞と判定できる．そのため，本手法では，v_i^jが 10km/h を超えるデータは，学習器による渋滞判定対象から除外する．

3.4 分類手法

学習器として，先行研究[7][9]と同じランダムフォレスト，AdaBoost，XGBoost を用い，バスb_iが走行したバス停区間s^jにおいて「渋滞」「非渋滞」の二値分類を行う．実装は Python の機械学習ライブラリである scikit-learn のバージョン 1.3.2 を用いて行う．

4. 評価実験

4.1 渋滞検知対象区間

本実験で渋滞検知を行う対象区間を図 4.1 に示す．同区間には，東京都交通局が運行する都営バス池 86 系統が，3~5 本／時間の頻度で走行している．本実験で対象とする区間は，池 86 系統の「池袋駅東口」～「渋谷駅東口」～「南池袋三丁目」間とする．

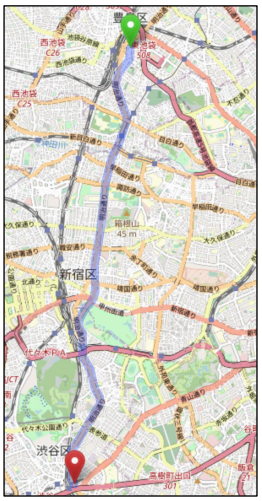


図4.1 渋滞検知対象区間（OpenStreetMapより）

4.2 データセット

4.2.1 バスの運行データの取得方法

バスの運行データは，公共交通オープンデータ協議会が公開している「公共交通オープンデータセンター」[7]の API を用いて取得した．表 4.1 に取得したバス運行データのサンプルを示す．

表 4.1 取得したバス運行データのサンプル

日付	車両番号	出発バス停留所識別子	バス停出発時刻
2022-11-30	C220	SodaiRiko.1051.2	10:23:49

4.2.2 雨天ラベルの取得方法

雨天ラベルに用いるデータは，気象庁 Web サイトの「過去の気象データ検索」[8]より，地点「東京都 東京」の「10 分ごとの値」の「降水量(mm)」のデータから取得した．表 4.2 に取得した降水量データのサンプルを示す．

表 4.2 取得した降水量データのサンプル

日付	時間帯	降水量[mm]
2022-11-30	14:00	0.0
2022-11-30	14:10	--

[7] https://developer-dc.odpt.org/ja/info

[8] https://www.data.jma.go.jp/stats/etrn/index.php

「降水量(mm)」の「10 分ごとの値」が「--」である時間帯は雨が降っていないことを意味し，雨天ラベルを「0」とする．それ以外の時間帯は雨が降っていることを意味し，雨天ラベルを「1」とする．

また，雨天ラベルデータは，10 分ごとのデータしか存在しない都合上，バスが実際に走行する時刻とタイムラグがあるため，「バス停出発時刻」より後のデータかつ出発時刻に最も近いデータを採用する．

4.2.3 渋滞ラベルの取得方法

渋滞ラベルの正解データは，Google Maps Platform の Directions Advanced API[9]にリクエストを送り 5 分おきに収集した．

Directions Advanced API では，地図上の出発地点と到着地点を緯度・経度により指定すると，2 点間の道のり（m）と，リアルタイム交通状況を踏まえた所要時間（s）を含むデータを返す．出発地点と到着地点に，それぞれバス停b_iとバス停b_{i+1}の緯度・経度を入力することにより，バス停間の道のりとリアルタイムの所要時間を取得できる（表 4.3 にサンプルを示す）．なお，経過時間は，当該区間の一般車両通過時間を示す．対象区間での交通速度は，距離を経過時間で割ることで計算し，交通速度が渋滞の定義である時速 10km 未満であれば渋滞，それ以外を非渋滞とする．

表 4.3 取得した渋滞ラベルデータのサンプル

日付	取得時刻	バス停始点	バス停終点	経過時間 [s]	距離 [m]
2022/11/30	09:00:00	池袋駅東口	南池袋一丁目	158	615

渋滞ラベルの正解データは，交通速度を実計測したものが最も正確である．しかし，著者の知る限り，この条件を満たし一般公開されているデータセットは存在しない．一方で，本 API の正解データとしての正確性は，著者らが明治通り（早稲田大学 西早稲田キャンパス前）で実施した目測による調査[10]において，実際の渋滞状況と比較して 0.92 の Accuracy があることが確認されている．そのため，少なくともこの調査を実施した明治通りでは，正解データとして十分な正確性を有すると判断し，評価実験で用いる．

また，渋滞ラベルデータは，API から定期的に取得する都合上，バスが実際に走行する時刻とタイムラグがあるため，「バス停出発時刻」より後のデータかつ出発時刻に最も近いデータを採用する．

4.2.4 データセットの作成

本実験では，バス停区間を走行したバスデータ（表 4.1），雨天ラベルデータ（表 4.2），渋滞ラベルデータ（表 4.3）を用いて，「渋滞」と「非渋滞」の二値分類を行う．表 4.1，4.2，4.3 の各データは，2022 年 11 月 30 日～12 月 13 日，12 月 18 日～12 月 28 日，2023 年 1 月 25 日～2 月 28 日の計 60 日間で収集したものである．二値分類を行う上で，必要となるデータセットのサンプルを表 4.4 に示す．表 4.5 には，作成したデータセットについてバス停区間毎に渋滞ラベルの内訳をまとめた．1 データは，その区間を走行したバス 1 台分の運行データに対応する．

表 4.4 作成したデータセットのサンプル

日付	A 出発時刻	B 出発時刻	始点のバス停 (A)	終点のバス停 (B)	経過時間 B－A [s]	雨天フラグ	渋滞フラグ
2022/11/30	15:48:05	15:52:49	ShinjukuIsetan.705.7	SendagayaGochome.832.2	284	0	1

表 4.5 渋滞ラベル数（池 86 系統）

始点	終点	渋滞	非渋滞	計
池袋サンシャインシティ	池袋駅東口	13	1,535	1,548
池袋駅東口	南池袋一丁目	347	3,998	4,345
南池袋一丁目	学習院下	18	3,127	3,145
学習院下	学習院女子大学前	85	2,798	2,883
学習院女子大学前	新宿コズミックセンター前	8	4,073	4,081
新宿コズミックセンター前	東新宿駅前	73	4,200	4,273
東新宿駅前	新宿伊勢丹前	550	2,332	2,882
新宿伊勢丹前	千駄ヶ谷五丁目	87	2,823	2,910
千駄ヶ谷五丁目	千駄ヶ谷小学校前	0	2,928	2,928
千駄ヶ谷小学校前	表参道	319	2,585	2,904
表参道	神南一丁目	32	2,908	2,940
神南一丁目	渋谷駅東口	2,543	452	2,995
渋谷駅東口	神宮前六丁目	157	2,902	3,059
神宮前六丁目	神宮前一丁目	36	2,924	2,960
神宮前一丁目	北参道	0	2,978	2,978
北参道	新宿四丁目	19	2,909	2,928
新宿四丁目	日清食品前	372	2,553	2,925
日清食品前	大久保通り	165	3,719	3,884
大久保通り	早大理工前	55	4,206	4,261
早大理工前	高田馬場二丁目	371	3,554	3,925
高田馬場二丁目	千登世橋	1	3,419	3,420
千登世橋	南池袋三丁目	114	4,528	4,642
南池袋三丁目	東池袋一丁目	289	1,300	1,589
全区間合計		5,654	68,751	74,405

前処理によりv_i^jが 10km/h を超えるデータを除外した後，データセットを表 4.5 のバス停区間毎に時系列

[9] https://developers.google.com/maps/documentation/directions

[10] https://www.seikatubunka.metro.tokyo.lg.jp/tomin_anzen/about/kaigi/kotsu-anzen/files/0000002078/jutai_2305_suishinkaigi_02-3_shiryo_3.pdf

順にソートし，前半 60%を学習用データセット，次の 20%を検証用データセットとし，最後の 20%をテスト用データセットとした．

ここで，表 4.5 より，学習用データとテスト用データに渋滞データが全くもしくはほとんど含まれていない区間が存在する．このような区間では，学習に必要なデータを十分に確保できないため，正確な渋滞検知が困難である．よって，本研究では具体的な基準として，学習用データに 20 個以上，もしくはテスト用データに 10 個以上の渋滞データが含まれていない区間は，渋滞検知の対象外とする．表 4.6 に渋滞検知の対象外となった区間（全 23 区間中 11 区間）を示す．

表 4.6 渋滞検知の対象外となった区間

始点	終点
池袋サンシャインシティ	池袋駅東口
南池袋一丁目	学習院下
学習院女子大学前	新宿コズミックセンター前
新宿コズミックセンター前	東新宿駅前
新宿伊勢丹前	千駄ヶ谷五丁目
千駄ヶ谷五丁目	千駄ヶ谷小学校前
表参道	神南一丁目
神宮前六丁目	神宮前一丁目
神宮前一丁目	北参道
北参道	新宿四丁目
高田馬場二丁目	千登世橋

4.3 評価実験

実験の目的は，複数台のバスに関する特徴量を追加した際の効果を確認することである．このため，評価実験では，表 3.2 の特徴量から複数台のバスに関する特徴量である $\overline{v^J_{i,n}}, \delta_{v^j_{i,n}}, \delta_{l^j_{i,n}}$ を除いて学習器を構築した場合（ベースライン手法）と，表 3.2 の全ての特徴量を用いて学習器を構築した場合（提案手法）の性能を比較する．

表 4.7 に全区間の実験結果（Accuracy，Precision，Recall，F1 スコアのマイクロ平均）を示す．表 4.7 より，複数台のバスに関する特徴量を含めるか否かに関わらず，F1 スコアの全区間マイクロ平均はランダムフォレストで最も高くなった．表 4.8 に F1 スコアの全区間マイクロ平均が最も高くなったランダムフォレストの区間毎の実験結果を示す．

表 4.7 より，ランダムフォレストにおける全区間のマイクロ平均は，Precision が 0.772 から 0.794 に上昇し，Recall が 0.716 から 0.706 に減少した．F1 スコアは 0.743 から 0.748 に上昇した．また，表 4.8 を見ると，全 12 区間中 Precision は 7 区間，Recall は 4 区間で上昇した．このことから，提案した特徴量は，

表 4.7 全区間の渋滞検知結果（12 区間のマイクロ平均）

アルゴリズム		Accuracy	Precision	Recall	F1 スコア	n
ランダムフォレスト	ベースライン	0.868	0.772	**0.716**	0.743	-
	提案手法	**0.873**	**0.794**	0.706	**0.748**	3
AdaBoost	ベースライン	**0.864**	**0.758**	**0.719**	**0.738**	-
	提案手法	0.852	0.734	0.698	0.716	3
XGBoost	ベースライン	0.849	0.715	0.718	0.717	-
	提案手法	**0.853**	**0.721**	**0.726**	**0.724**	4

表 4.8 区間毎の渋滞検知結果（ランダムフォレスト）

区間		Accuracy	Precision	Recall	F1 スコア
池袋駅東口～南池袋一丁目	ベースライン	**0.850**	**0.527**	0.604	**0.563**
	提案手法	0.817	0.451	**0.667**	0.538
学習院下～学習院女子大学前	ベースライン	0.874	0.500	0.538	0.518
	提案手法	**0.883**	**0.533**	**0.615**	**0.571**
東新宿駅前～新宿伊勢丹前	ベースライン	0.815	0.195	**0.113**	**0.143**
	提案手法	**0.835**	**0.200**	0.070	0.104
千駄ヶ谷小学校前～表参道	ベースライン	0.705	0.400	**0.346**	0.371
	提案手法	**0.734**	**0.459**	0.327	**0.382**
神南一丁目～渋谷駅東口	ベースライン	0.906	1.000	0.906	0.951
	提案手法	**0.913**	1.000	**0.913**	**0.955**
渋谷駅東口～神宮前六丁目	ベースライン	0.899	0.500	**0.292**	**0.369**
	提案手法	**0.908**	**0.600**	0.250	0.353
新宿四丁目～日清食品前	ベースライン	0.881	0.581	**0.493**	0.533
	提案手法	**0.891**	**0.636**	0.479	**0.546**
日清食品前～大久保通り	ベースライン	**0.904**	0.286	0.364	0.320
	提案手法	0.902	**0.290**	**0.409**	**0.339**
大久保通り～早大理工前	ベースライン	0.863	0.538	0.467	0.500
	提案手法	0.863	0.538	0.467	0.500
早大理工前～高田馬場二丁目	ベースライン	0.870	0.706	**0.779**	**0.741**
	提案手法	**0.879**	**0.764**	0.714	0.738
千登世橋～南池袋三丁目	ベースライン	**0.959**	**0.400**	0.333	**0.363**
	提案手法	0.953	0.333	0.333	0.333
南池袋三丁目～東池袋一丁目	ベースライン	0.796	**0.459**	**0.683**	**0.549**
	提案手法	**0.801**	0.455	0.488	0.471

Recall よりも Precision の向上に寄与したことがわかる．元々，複数台のバスに関する特徴量を加えた目的は，バスが信号や停留所で停車した場合に渋滞と誤検知する割合を減らすためであった．つまり，提案手法では False Positive が減り，Precision が上昇することが自然だと考えられるため，当初の想定通りの結果が得られたといえる．

5. まとめ

本稿では，都営バスのリアルタイム運行データから渋滞検知を行う手法を提案した．渋滞検知精度向上のため，直近に同じ区間を走行した複数台のバスに関して，①速度の平均値， ②速度の標準偏差， ③運転間隔の標準偏差 を特徴量として追加した．評価実験では，都営バスの運行データを用いて明治通り（池袋～渋谷間）の渋滞検知を行った．最も F1 スコアが高かったアルゴリズムであるランダムフォレストでは，直近 3 台分のバスのデータをもとに算出した特徴量を追加することにより，F1 スコアが 0.743 から 0.748 に上昇した．

今後の課題として，運行頻度の異なる路線での追加実験が挙げられる．本研究で実験対象とした池 86 系統より運行頻度が高い系統であれば，nの値を大きくしても，直近のバスのデータのみから特徴量を計算できるため，提案手法により更なる精度向上ができる可能性がある．一方，池 86 系統より運行頻度が低い系統であれば，提案手法を適用できない可能性があるため，別の手法を検討する必要があると考えられる．

謝　辞

本研究は，令和 4 年度「東京都と大学との共同事業」の一環として，東京都政策企画局からの交付を受けています．

参　考　文　献

[1] Y. Xu, Y. Wu, J. Xu, and L. Sun, “Efficient Detection Scheme for Urban Traffic Congestion Using Buses,” Proc. of 26th Int. Conf. on Advanced Information Networking and Applications Workshops, 2012, pp. 287-293, doi: 10.1109/WAINA.2012.62.

[2] C. Wang and H. Tsai, “Detecting urban traffic congestion with single vehicle,” Proc. of Int. Conf. on Connected Vehicles and Expo (ICCVE), 2013, pp. 233-240, doi: 10.1109/ICCVE.2013.6799799.

[3] R. Carli, M. Dotoli, N. Epicoco, B. Angelico, and A. Vinciullo, “Automated evaluation of urban traffic congestion using bus as a probe,” Proc. of IEEE Int. Conf. on Automation Science and Engineering (CASE), 2015, pp. 967-972, doi: 10.1109/CoASE.2015.7294224.

[4] H. Wang and Y. Si, “Detection of traffic abnormity based on clustering analysis of taxi GPS data,” Proc. of the 2019 2nd Int. Conf. on Data Science and Information Technology, 2019, pp. 219-224, doi: 10.1145/3352411.3352445.

[5] T. V. Saraiva and C. A. V. Campos, “An approach using machine learning and public data to detect traffic jams,” Proc. of 2021 Int. Wireless Communications and Mobile Computing (IWCMC), 2021, pp. 675-680, doi: 10.1109/IWCMC51323.2021.9498622.

[6] L. A. Makara, P. Maric, and A. Pekar, “Public transport congestion detection using incremental learning,” Pervasive and Mobile Computing, 2023, vol. 91, article 101769, doi: 10.1016/j.pmcj.2023.101769.

[7] 青柳宏紀，岡田一洸，山名早人，“都バスのリアルタイム運行データを用いた渋滞検知,”DEIM2022 第 14 回データ工学と情報マネジメントに関するフォーラム, 2022, A41-2, pp. 1-8.

[8] 藤田智也，青柳宏紀，畠中希，小口正人，山名早人，“転移学習による都営バスのリアルタイム運行データを用いた渋滞検知,” DEIM2023 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023, 5c-6-2, pp. 1-8.

[9] 畠中希，青柳宏紀，藤田智也，山名早人，小口正人，“都営バスのオープンデータによる渋滞検知の精度向上のための運行特性を考慮した手法の提案,” DEIM2023 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023, 1a-5-4, pp. 1-5.

[10] C. Bannur, C. Bhat, G. Goutham, and H. R. Mamatha, “General Transit Feed Specification Assisted Effective Traffic Congestion Prediction Using Decision Trees and Recurrent Neural Networks,” Proc. of 2022 IEEE 1st Int. Conf. on Data, Decision and Systems (ICDDS), 2022, pp. 1-6, doi: 10.1109/ICDDS56399.2022.10037408.

自転車走行中の表情と振動による潜在的快適性分析の検討

山口　琉太† 　村重　圭亮†† 　矢野　英人†††
義久　智樹†††† 　河合 由起子†,††† 　下條　真司†††††

† 京都産業大学大学院先端情報学研究科　〒 603-8555 京都府京都市北区上賀茂本山
†† 大阪大学大学院情報科学研究科　〒 567-0047 大阪府吹田市山田丘 1-5
††† 大阪大学サイバーメディアセンター　〒 565-0871 大阪府茨木市美穂ケ丘 5 － 1
†††† 滋賀大学データサイエンス学部　〒 522-8522 滋賀県彦根市場 1 丁 1-1
††††† 青森大学ソフトウェア情報学部　〒 030-0943 青森県青森市幸畑 2-3-1
E-mail: † i2286185@cc.kyoto-su.ac.jp

あらまし　近年，健康的で経済的な移動手段として，二輪車の利用が世界的に推進されている．また，COVID-19 による自転車通勤の増加に伴い，MaaS（Mobility as a Service）における移動手段として自転車の活用が注目されている．本研究では，自転車ユーザのための快適な移動支援を目的に，スマートフォンのセンサとカメラで自転車走行中の表情や振動を収集しつつ，効率的にデータ収集できる経路を推薦するナビゲーションシステムを構築する．取得したデータから運転者の感情や路面の凹凸を分析し，地点の潜在的快適性を抽出する．また，走行後に各地点の分析結果を明示的に評価できる機能を実装し，走行後のフィードバックを収集する．最後に，取得したデータとユーザフィードバックから快適性を可視化するマップを生成する．本研究では，青森にて走行実験を行い，表情画像と振動による快適性抽出結果を比較し，妥当性を検証する．

キーワード　MaaS， 二輪車， 経路推薦， 表情分析， 凹凸検出，潜在的快適性

1 はじめに

近年，健康への配慮や健康的，経済的な交通手段として二輪車の利用が世界的で促進されており，日本でも 2017 年に自転車活用推進法が施行され，自転車をこれまで以上に利用してもらう動きが進んでいる [1] [2]．また，COVID-19 の流行により公共交通機関の利用を控え，自転車を利用する人が増加している [3]．株式会社ドコモ・バイクシェア [1]や OpenStreet 株式会社 [2]等の企業が二輪車のシェアサービスを行っている．15 分程度の短時間の利用ができることや借りたポート (駐輪場) と同じポートに返却する必要がないため都市部で注目が集まっている．

ナビゲーションシステムに関する研究は広く実施されており，既存研究では路面状況や道幅等の環境データを事前に取得することで，最短経路推薦や安全な経路の推薦を実現しているが，ユーザが走行中に環境から直接体感する明示的/潜在的な快適性を経路推薦に反映するには至っていない．本研究では，自転車ユーザ向けの快適な経路推薦システムの構築を目的に，二輪車で走行中の表情と振動から走行環境に対する潜在的な快適性を抽出する．二輪車走行中は歩行時と違い文字入力等のスマホ操作ができない．そこで，運転者の快適性を自動的に抽出するために，走行中の表情と路面の凹凸に着目する．二輪車に搭載したスマホから運転者の表情と振動を取得し，感情分析と路面状況を分析することで，ユーザの走行環境に対する潜在的な快適性を抽出し，地点に対してオートアノテーションする．これにより，ナビゲーションや地点の特徴をマッピングする等の利活用につなげる．さらに，表情分析結果から快適度が高い地点と低い地点に対してそれぞれ快適さのフィードバックを明示的に実施することで，アノテーション精度を高める．本稿では，走行中の表情と振動による地点に対する潜在的な快適性分析手法を提案し，青森にて走行実験を行い，快適性抽出結果を検証する．

以下，2 章では二輪車ナビゲーションシステムと二輪車によるセンシング技術の関連研究を紹介し，本研究と比較する．3 章と 4 章，5 章，6 章で快適性分析システム，感情強度算出手法，快適な経路推薦手法，データ収集ナビアプリの実装，7 章で作成した快適性分析システムと快適性抽出の評価を行い，最後に 8 章で本研究で提案した快適性分析システムのまとめと今後の課題について述べる．

2 関連研究

本章では，二輪車走行中のセンシング技術と二輪車ナビゲーションに関する研究を説明する．

2.1 二輪車ユーザ向けのナビゲーション技術

近年，自転車利用者の増加に伴い，安全性や快適性，景観を考慮した経路推薦手法の研究が進められている．これらの要素はカーナビに代表される効率重視の経路推薦に比べ，自転車利用者にとって重要である [4] [5] [6]．Oh ら [7] は，特定の施設の

1 : https://docomo-cycle.jp/
2 : https://www.hellocycling.jp/

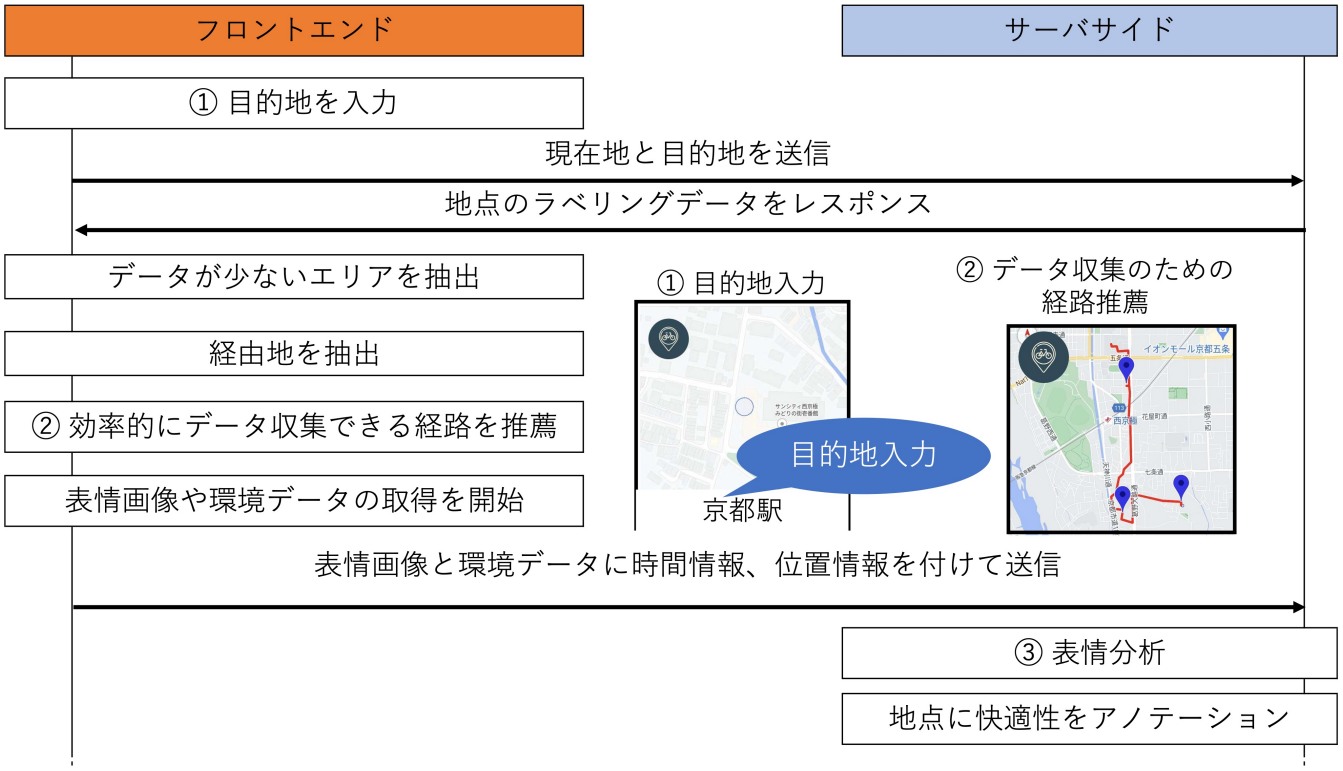


図 1　データ収集ナビゲーションシステムの処理の流れ

犯罪危険率をあらかじめ算出し，ユーザの経路に含まれる施設の犯罪発生率から経路の犯罪危険率を算出することで，利用者が出発地と目的地を入力すると，利用時間や日程に応じて様々な経路の中から安全な経路を推薦するシステムを構築した．Herzog ら [8] は，マルチモーダルな経路を提案する際に，ユーザの個人的な好みやその地域特有の物事を考慮しないことに注目した．地元住人の知識や経験を取り入れることで経路推薦の質の向上を目指し，協調フィルタリングと知識ベース，おすすめの経路を組み合わせたシステムのユーザ評価を行なったところ，最新の推薦システムよりユーザの満足度が高いことを示した．Matos ら [9] は，travel time, energy expenditure, effort distribution, infrastructure performance, safety, comfort and emission hotspots の 7 種類の指標に基づいて，二輪車ナビゲーションのための情報システムを開発した．Wei ら [10] は，ほとんどの E-Bike サービスはリアルタイムでのターンバイターンナビゲーションのためネットワークに依存していることに注目し，ナビゲーションをオフラインの場合でも利用できるよう，Open Street Map をベースに「Skobber」を用いてナビゲーションを開発した．また，Google Maps の Place API を利用して，有名な場所や店を検索可能にした．

2.2　二輪車によるセンシング技術

二輪車にセンサやスマホを取り付け，走行中にセンサデータを収集する研究が進められている [11] [12]．Taniguchi ら [13] は，自転車事故を防止するために二輪車に超音波センサを取り付け，定期的にセンサから道路までの距離を測定することで，道路上の障害物や穴を検知した．Jeon ら [14] は，二輪車にレーザーセンサを取り付け，後方を走行する車両の存在を検知・追跡し，さまざまな種類の車線変更を検知できることを実証した．また，得られた情報から衝突の危険性を検知し，二輪車走行者に音声で警告するシステムを開発した．Jaime ら [15] は，都市環境における大気物質のモニタリングを目的として二酸化窒素やオゾンを測定するための電気化学センサと PM2.5 や PM10 を測定するための光学センサを自転車のハンドルに搭載し，バダホス市（スペイン）内を自転車で走行した．結果，平日と週末や時間帯による変化を都市中心部と周辺部の汚染物質の時空間変動に関するデータを取得した．Bian ら [16] は，自転車利用環境を適切に把握するため，利用環境品質（BEQ）という概念を提案し，知覚満足度と葛藤度によって定義した．携帯電話のセンサデータから速度と加速度関連の自転車行動指標（BBI）と記録データから自転車走行量などの自転車走行環境指標（BEIs）を抽出した．BBI と BEI を入力属性とし，ランダムフォレストアルゴリズムに基づく 2 階層 BEQ 分類評価モデルを構築した．提案した RF ベースの分類評価モデルより BEQ の評価に GPS データを用いることが可能であることを示した．

2.3　本研究の新規性

既存研究では周囲の環境データを事前に取得することで，最短経路や快適な経路の推薦を実現しているが，自転車特有のユーザが直接体感する心地よさや不快感といった周囲の道路環境から受ける感情の取得および分析による経路推薦には至っていない．本研究では，二輪車走行中の表情や振動，フィードバックから抽出した潜在的な評価を用いることで，周辺環境も含めた快適性をラベリングし，快適な経路を推薦できる点が特異点である．

図 2　実験で使用したデータ収集システムを搭載した二輪車

3　走行中の表情と振動に基づく快適性分析システム

ユーザの潜在的快適性分析と分析結果に基づいた経路推薦を実現することを目的に，運転者の潜在的快適性分析に用いる走行時の表情画像と振動データを安全に取得するナビゲーションシステムと分析システムを構築する．実験で使用した走行中の表情画像と振動データを収集するナビゲーションシステムを搭載した二輪車を図 2 に示す．

3.1　データ収集ナビゲーションシステムの概要

図 1 に表情画像と振動データを収集するナビゲーションシステムのフロントエンドとサーバサイドの処理を示す．まず，ナビゲーションアプリを起動し，目的地を入力する．現在地と目的地をサーバに送信すると，地点にラベル付けされたデータがレスポンスされる．ここでのラベル付けとは表情分析結果や路面状況を地点に割り当てることである．フロントエンドは，レスポンスされたラベリングデータから，データが少ない地点を抽出し，目的地までの通過点とすることで新たな地点のデータを効率的に収集する ([17])．これら経由地と現在地，目的地の座標を結ぶ最短ルートを Google Maps API で取得し，データが未取得のエリアを効率的に経由するデータ収集用の経路として推薦する．

3.2　路面の凹凸による路面状態分析システム

スマホの加速度センサを用いて路面の凹凸を検出し，多くの凹凸がある道を避けることで自転車ユーザの快適性の向上を目指す．自転車が路面の段差や凹凸を走行する際，垂直方向の加速度が発生するので，その加速度データから路面の凹凸の有無を判定する．垂直加速度は，スマホのモーションセンサから得られる重力を除いた 3 軸座標系の加速度から算出し，垂直加速度が急激に増加する場所を路面の凹凸があると判定する．鉛直加速度 a_v は，y 軸加速度 a_y と z 軸加速度 a_z を合成し，式 (1) で計算する．図 3 は鉛直加速度の導出を視覚的に表したものである．角度 θ は，スマホを通じて y 軸方向における地面との角度である．角度 θ はスマホのセンサで取得することができ，スマホの角度に応じて鉛直加速度を算出できる．

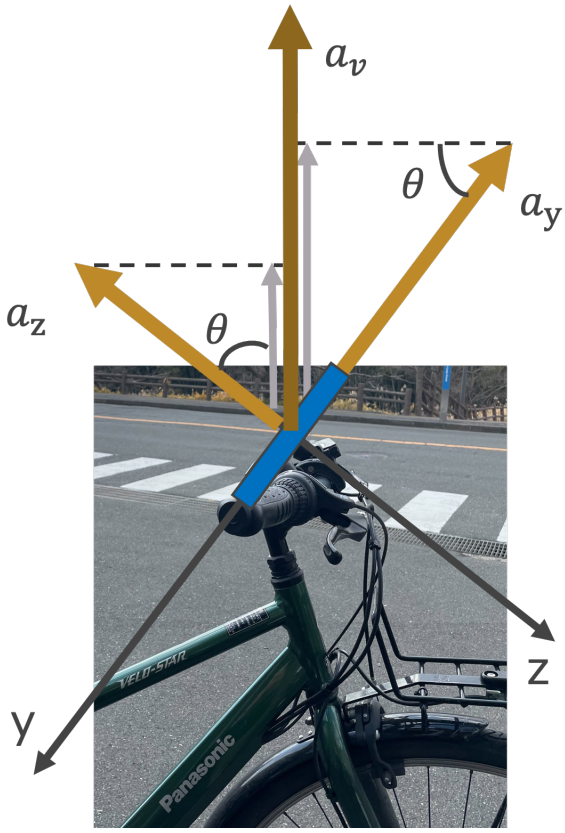


図 3　自転車走行中の垂直加速度の方向と座標軸

$$|a_v| = \sqrt{(a_y \sin\theta + a_z \cos\theta)^2} \qquad (1)$$

3.3　運転者の表情による快適性分析システム

本研究では，センサデータだけでなく自転車走行中の表情画像を用いて運転者の表情から潜在的な快適性を分析し，その地点にアノテーションすることで運転者が走行中に無意識に感じている快適性，不快感，驚き等の感情を収集する．サーバサイドで運転者の表情画像を受け取り，表情分析することで感情値を算出する．著者らはこれまで Microsoft Azure が提供する Face API を利用していたが，顔認識と感情分析の精度向上を目指し，顔認識は Haar-Cascade-based [18] による顔認識モデル，感情分析は Deep Convolutional Neural Networks (DCNN) [19] で学習されたモデルを利用した．この解析モデルは「surprise」,「neutral」,「anger」,「happy」,「sad」の 5 つの感情を計算し，各感情を 0 から 1 の範囲で数値化する．5 つの感情値の合計は 1.0 となる．これら 5 つの感情値は，走行地点の緯度経度とタイムスタンプと共に地点にアノテーションされる．

3.4　走行後の明示的フィードバック機能

二輪車走行時の表情の感情分析により得られた潜在的快適性は，天候や撮影角度等の要因に依存し，データの信頼性が低い

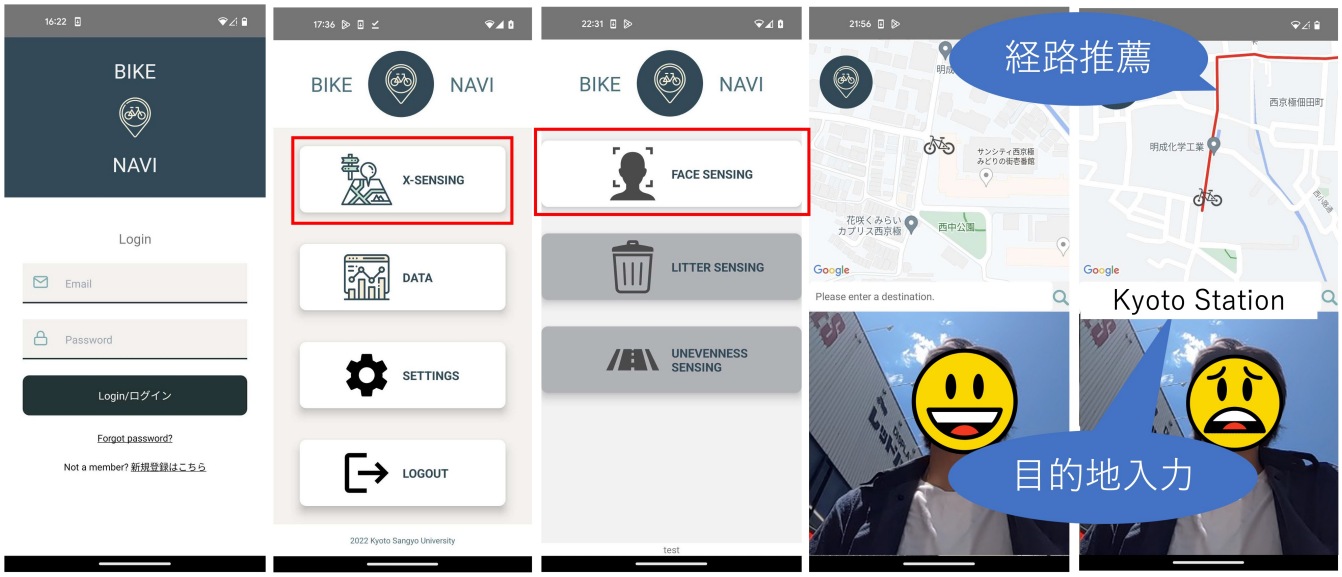


図 4　表情画像収集ナビゲーションアプリの起動から表情画像取得までの流れ

図 5　振動データ収集アプリの起動から振動データ取得

可能性がある．そこで，アノテーション精度を高めるために，二輪車走行後に地点ごとの分析結果を明示的に再評価できるフィードバック機能を実装する．ユーザは走行後に走行した地点の Google Street View 画像が表示されるので，その地点が快適だったどうかをリッカート尺度の 5 段階で明示的にフィードバックする (very good=1.0，good=0.8，normal=0.6，bad=0.4，very bad=0.2，not remember=0)．そして，表情分析結果と明示的フィードバック結果を組み合わせることで次章で説明する感情強度を算出する．

4　表情分析と明示的フィードバックによる感情抽出

地点へのアノテーション精度を向上させるために，表情分析結果と明示的フィードバック結果を組み合わせ，快適さの指標である感情強度 (EI) を算出する．感情強度は以下の関数で算出され，地点に対してラベリングされる．式 (2) は感情 i がポジティブ（「happy」，「neutral」）の場合，式 (3) はネガティブ（「anger」，「sad」，「surprise」）の場合に使用される [20]．S_i は表情画像を感情分析した潜在的な快適さ (0〜1) であり，表情分析結果の 5 つの感情値である． S_U は走行直後の 5 段階リッカート尺度のフィードバック結果である．重み λ_1 と重み λ_2 の和は 1.0 である．

$$EI(i) = \begin{cases} \lambda_1 S_i + \lambda_2 S_U & if\ i \leqq 2 \quad (2) \\ \lambda_1 S_i + \lambda_2 \dfrac{1}{S_U} & if\ i > 2 \quad (3) \end{cases}$$

5　環境データに基づく快適ナビゲーション

自転車走行時の快適さは，道路や歩道の地点だけでなく走行経路上の周囲の環境から影響を受けると仮定し，提案する経路推薦手法は，まずマップをエリアごとに分割する．図 6 に 5×5 のエリアに分割した快適な経路推薦アルゴリズムを示す．左下のセルの座標は（0，0），現在地（2，4）を緑色，目的地（4，2）を黄色，経由地を水色とし，推薦経路は赤矢印とする．

分割したエリアのうち，経路推薦対象となる 3×3（赤枠）エリアにおける快適ナビの経路推薦アルゴリズムを示す．なお，単純化のため快適なエリアを経由する地点を 1 箇所とする．快適性の重みを表情分析結果の 0〜1 の範囲で設定し，値が大きいほど快適性が高いとする．初期のエリアの重みは少なくとも一度は走行され，各エリアに快適性がラベリングされているとする（a）．現在地と目的地を含むエリア（図中赤枠の 9 つの

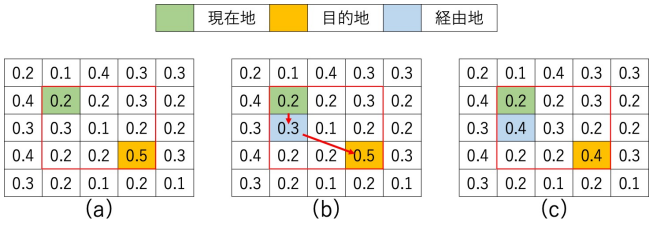


図 6　快適性を考慮した経路推薦アルゴリズム

エリア）を対象とし，各エリアを中心とした周囲 8 エリアと選出された各エリア自身の総和を算出する．算出結果は，エリア（2，3）は 2.3,（3，3），（4，3），（4，4）が 2.2,（2，2）と（3，4）は 2.1,（3，2）が 2.0 となり，最大値となる（2，3）のエリアを経由地エリアとして抽出する（b）．経由地エリア内の感情強度が最大だった地点の座標を経由地とし，現在地と経由地と目的地を結ぶ最短経路を生成し，推薦する．推薦された経路を走行すると，地点に対する快適性を分析し，すでに取得したデータと合わせてエリアに快適性の重みを更新する (c)．

6　データ収集ナビとフィードバック機能の実装と快適性可視化マップの構築

前章までで提案した手法に基づき，データ収集ナビゲーションシステムと明示的フィードバック機能を持つ Android アプリを実装した [3]．

6.1　表情画像収集ナビの実装

図 4 に表情画像収集ナビゲーションアプリのログイン画面から走行中の表情画像撮影までの流れを示す．アプリを起動すると，ログイン画面が表示される (図 4 (a))．ログイン画面はメールアドレスとパスワードで構成されており，ユーザに合わせた経路推薦を実現するために活用される．ログイン後，トップページ（b）に遷移し，「X-SENSING」を選択し，収集するセンシングデータを選択する (c)．現在地周辺のマップが表示され，スマホの内カメラが起動し，表情の撮影が開始される (d)．必要に応じて目的地を入力することで，ナビゲーションしつつ表情画像を撮影することが可能である (e)．3 秒に一回，現在地を更新して表情画像を自動で取得するため，ユーザは自転車走行中に安全にデータを収集することが可能である．

6.2　振動データ収集システムの実装

図 5 に振動データ収集アプリのデータ収集までの流れを示す．アプリを起動すると，トップページが表示されるのでデータ収集を開始するため START ボタンをクリックする (図 5 (a))．データ収集が開始され，垂直加速度や位置情報，時刻等が表示され (b)，自動でデータ収集され保存される．

6.3　明示的フィードバック機能の実装

図 7 に走行後の明示的フィードバックの流れを示す．走行後

図 7　走行後の明示的フィードバックの流れ

にトップページの「FEEDBACK」を選択 (図 7 (a)) すると，表情分析結果の”happy”が上位 10 件と下位 10 件の地点にピンがプロットされたマップが表示 (b) される．ピンをクリックするとその地点の Google Street View 画像が表示されるのでその地点が走行していて快適だったかをフィードバックする (c)．

6.4　表情分析とフィードバックによる快適性可視化マップ

自転車走行中の表情分析と明示的フィードバック結果から快適性を抽出し，マップ上に可視化するシステムを開発した．図 8 に 17 人の被験者が関西で走行した結果を可視化したマップを示す．マップ上の花アイコンは快適性が高い地点，葉アイコンは快適性が低い地点を示す．地図を拡大した西側のエリアを見ると葉アイコンが多くプロットされており快適性が低いエリアであることが分かる．このエリアは北側に山があり傾斜が大きいため走行が大変なエリアである．一方で花アイコンが多くプロットされている東側は，傾斜が緩く，川沿いなので走行しやすく，心地の良いエリアである．

7　快適性分析システムの検証

実装したシステムを用いて被験者実験を行い，快適性分析システムを検証する． 検証期間は 2023 年 11 月 13 日と 2023 年 11 月 14 日で，青森市で 4 人 (20 代:2 人，50 代:1 人，60 代:1 人) の被験者で実施した． 被験者は，開発したデータ収集ナビアプリで推薦された経路に従って 1 日 1.5 時間二輪車で走行した．

7.1　走行時の表情分析結果の検証

走行中の表情の分析結果から得られた感情を検証する．実験より得られた走行時の表情画像（地点）は 9,901 枚で，表情分析が可能であった画像は，2,533 枚で全体の 25.5%であった． 表情分析ができなかった画像を見ると，カメラの角度に問題があり下からの画像になっていたため，顔認識ができなかったと考えられる．被験者 4 名の 5 つの感情分析結果の平均 (average) と標準偏差 (SD) を表 1 に示す． neutral が平均値で大きな値と

3：https://delab2.kyoto-su.ac.jp/~yamaryu/release/android/bike-navi-v0.3.5.apk

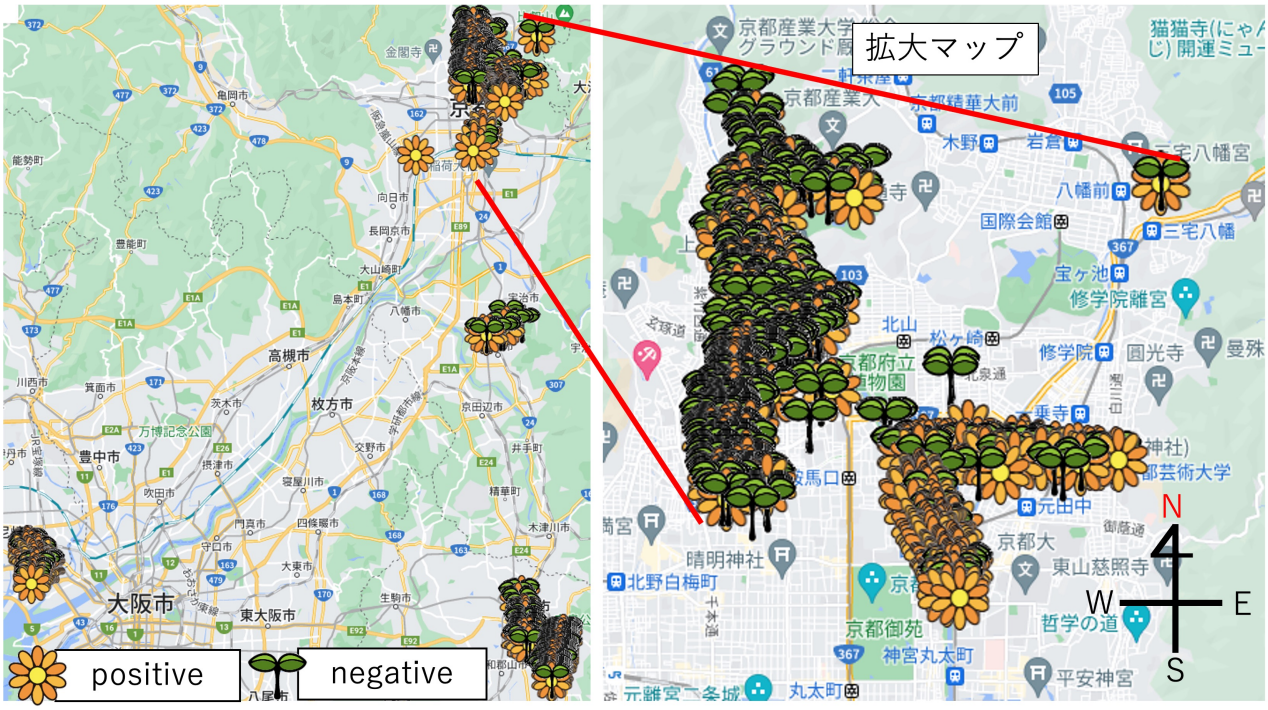


図 8　表情分析と明示的フィードバックによる快適性可視化マップ

表 1　ユーザごとの感情分析結果

	surprise	neutral	anger	happy	sad
U1	0.0999	0.3477	0.0685	0.1565	0.3271
U2	0.0800	0.3881	0.0713	0.1129	0.3475
U3	0.1006	0.3476	0.0686	0.1575	0.3255
U4	0.1002	0.3477	0.0686	0.1570	0.3264
average	0.0999	0.3477	0.0685	0.1565	0.3271
SD	0.1338	0.1788	0.0652	0.2364	0.2098

なっており，自転車走行中は無表情の場面が多い点や携帯をハンドル付近に設置しており，下からの角度の表情画像になるので，ポジティブな評価が多くないと考えられる．次いで happy や sad といった快適性に直接影響する感情が大きな値となった．さらに，happy の標準偏差が他の感情と比較して大きく，感情の起伏が大きいので快適性を判定するのに happy が有効なのではないかと考えられる．次節で happy の感情に着目して快適性を抽出し，振動を検出した地点と比較して happy の感情が地点の快適性の抽出に有効かどうかを検証する．

7.2　表情分析と振動データによる快適性と振動の可視化マップ

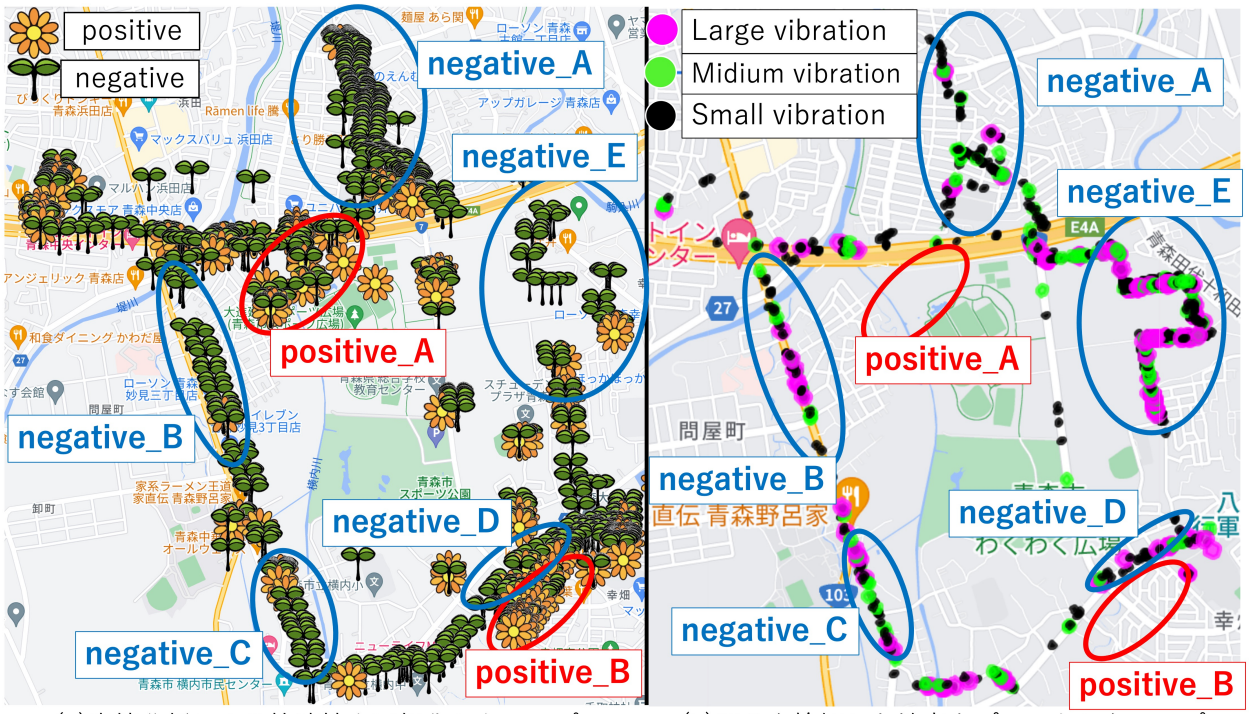
表情による快適さと振動の大きさとの関係を比較するため，図 9 に表情分析による happy の感情を元に抽出した快適性を可視化したマップ (左) と振動が検出された地点をプロットしたマップ (右) を示す．図 9 の左のマップの花アイコンと葉アイコンは，図 8 同様にそれぞれ快適な地点とそうでない地点を示す．図 9 の右のマップの紫，緑，黒色の点は垂直加速度を正規化し，閾値を超えた地点を示す (紫:0.75 以上，緑:0.5 以上，黒:0.25 以上)．図中の可視化マップを見ると，花アイコンが多くプロットされた快適性の高いエリア (図 9 左マップの赤円) では振動が計測されていない (図 9 右マップの赤円)．一方で，葉アイコンが多くプロットされている快適性の低いエリア (図 9 左マップの青円) では，大小様々な大きさの振動が計測されていることが分かる (図 9 右マップの青円)．

7.3　表情分析と振動データによるポジティブ/ネガティブなエリア分析

前節で表情分析と振動データを可視化することで，快適性が高いエリアが 2 つ，低いエリアが 5 つ抽出された．最後に，この 7 つのエリアの表情分析結果と振動データに着目する．表 2 は 7 つのエリア別にそれぞれの感情の平均値と凹凸検出に利用した垂直加速度を正規化した平均値を算出した．表 2 の happy の平均値を見ると，ポジティブなエリアの快適性は他のエリアと比較して高い．次に垂直加速度に着目すると，図 9 ではポジティブなエリアには振動がプロットされていなかったが，小さな振動だったためマップには表示されなかった．一方で，ネガティブなエリアはポジティブなエリアと比較して垂直加速度が 2 倍以上計測されており，走行が大変だったと予想される．以上のことから，二輪車走行中の表情と振動を組み合わせることで，走行地点の快適性を適切に抽出し，地点へアノテーションできることを示している．

8　終わりに

本研究では，自転車走行中の表情と振動，走行後の記憶から走行環境に対する潜在的な快適性を抽出するシステムを開発し，実証実験を行った．実験より二輪車走行中の表情と振動から快適性を抽出することで，適切な地点に快適性をアノテーションすることができた．カメラの角度や周辺環境の問題で表情分析



(a)表情分析による快適性を可視化したマップ　(b)凹凸を検知した地点をプロットしたマップ

図 9　表情分析による快適性と振動を検出した地点を可視化したマップ

表 2　ポジティブ/ネガティブなエリアごとに算出したユーザごとの感情分析結果と凹凸検知のための垂直加速度の平均値

	surprise	neutral	anger	happy	sad	垂直加速度
positive_A	0.1076	0.3199	0.0966	0.2462	0.2294	0.2543
positive_B	0.0765	0.2471	0.0833	0.4021	0.1926	0.1807
negative_A	0.0893	0.4580	0.0715	0.0737	0.3084	0.5428
negative_B	0.1010	0.4166	0.0702	0.1213	0.2907	0.4910
negative_C	0.0858	0.3565	0.0572	0.0983	0.4020	0.5069
negative_D	0.0982	0.5190	0.0735	0.0384	0.2706	0.5799
negative_E	0.1190	0.3510	0.0756	0.1653	0.2889	0.4217

ができなかった画像が多くあったので，今後はアノテーション精度を向上させるために顔認識モデルや表情による感情分析モデルの精度向上や二輪車走行中の他の快適性の指標の一つである日照データに着目する．また，アノテーションされた快適性データを経路推薦に活用する予定である．

謝　　辞

本研究の一部は，JSPS 科研費 JP22H03700，JP20H00584，JP19K12240，JP22K12274 の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 大森宣暁．わが国の自転車文化に関する一考察-ママチャリに着目して．*IATSS Review (*国際交通安全学会誌*)*, 46(2):139–146, 2021.

[2] 山内さつき, 島崎敢, 小嶋理江, 米川隆, 武田夏佳, 新海裕子, and 青木宏文．Covid-19 流行に伴う高齢者の行動と利用移動手段の変化．**自動車技術会論文集**, 52(5):1143–1148, 2021.

[3] Haruka Kato and Daisuke Matsushita. Changes in walkable streets during the covid-19 pandemic in a suburban city in the osaka metropolitan area. *Sustainability*, 13(13):7442, 2021.

[4] Rohit Verma, Surjya Ghosh, Mahankali Saketh, Niloy Ganguly, Bivas Mitra, and Sandip Chakraborty. Comfride: a smartphone based system for comfortable public transport recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 181–189, 2018.

[5] Yihong Zhang, Panote Siriaraya, Yuanyuan Wang, Shoko Wakamiya, Yukiko Kawai, and Adam Jatowt. Walking down a different path: route recommendation based on visual and facility based diversity. In *Companion Proceedings of the The Web Conference 2018*, pages 171–174, 2018.

[6] Weisheng Zhong, Fanglan Chen, Kaiqun Fu, and Chang-Tien Lu. Safebike: a bike-sharing route recommender with availability prediction and safe routing. *arXiv preprint arXiv:1712.01469*, 2017.

[7] Su Bin Oh, Hyeok Ju Park, Byeong Ki Kang, Ill Chul Doo, and Mee Hwa Park. Safe route recommendation method to prevent crime exposure. *Advanced Science Letters*, 23(10):9579–9583, 2017.

[8] Daniel Herzog, Hesham Massoud, and Wolfgang Wörndl. Routeme: A mobile recommender system for personalized, multi-modal route planning. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*,

pages 67–75, 2017.

[9] Francisco Lé de Matos, José Maria Fernandes, Carlos Sampaio, Joaquim Macedo, Margarida C Coelho, and Jorge Bandeira. Development of an information system for cycling navigation. *Transportation research procedia*, 52:107–114, 2021.

[10] Ching-Chuan Wei, Jing-Shian Lin, Chia-Chi Chang, Yung-Fa Huang, and Chuan-Bi Lin. The development of e-bike navigation technology based on an openstreetmap. *Smart Science*, 6(1):29–35, 2018.

[11] Sasank Reddy, Katie Shilton, Gleb Denisov, Christian Cenizal, Deborah Estrin, and Mani Srivastava. Biketastic: sensing and mapping for better biking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1817–1820, 2010.

[12] Steven Verstockt, Viktor Slavkovikj, Pieterjan De Potter, Jürgen Slowack, and Rik Van de Walle. Multi-modal bike sensing for automatic geo-annotation geo-annotation of road/terrain type by participatory bike-sensing. In *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, pages 39–49. IEEE, 2013.

[13] Yoshiaki Taniguchi, Kodai Nishii, and Hiroyuki Hisamatsu. Evaluation of a bicycle-mounted ultrasonic distance sensor for monitoring road surface condition. In *2015 7th International Conference on Computational Intelligence, Communication Systems and Networks*, pages 31–34. IEEE, 2015.

[14] Woongsun Jeon and Rajesh Rajamani. Active sensing on a bicycle for simultaneous search and tracking of multiple rear vehicles. *IEEE Transactions on Vehicular Technology*, 68(6):5295–5308, 2019.

[15] Jaime Gómez-Suárez, Patricia Arroyo, Raimundo Alfonso, José Ignacio Suárez, Eduardo Pinilla-Gil, and Jesús Lozano. A novel bike-mounted sensing device with cloud connectivity for dynamic air-quality monitoring by urban cyclists. *Sensors*, 22(3):1272, 2022.

[16] Yang Bian, Ling Li, Huan Zhang, Dandan Xu, Jian Rong, and Jiachuan Wang. Categorizing bicycling environment quality based on mobile sensor data and bicycle flow data. *Sustainability*, 13(8):4085, 2021.

[17] Ryuta Yamaguchi, Panote Siriaraya, Da Li, Tomoki Yoshihisa, Shinji Shimojo, and Yukiko Kawai. A proposal of data collection by mobility users on real-time for e-bike navigation system. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 6056–6057. IEEE, 2021.

[18] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Ryuta Yamaguchi, Panote Siriaraya, Tomoki Yoshihisa, Shinji Shimojo, and Yukiko Kawai. A detection system for comfortable locations based on facial expression analysis while riding bicycles. In *Companion Proceedings of the ACM Web Conference 2023*, pages 306–309, 2023.

強化学習による差動二輪車制御における内外乱適応

門垣　幸樹†　　大島　裕明†,††

† 兵庫県立大学社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †fa20x088@guh.u-hyogo.ac.jp, ††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，内外乱に適応した差動二輪車の安定した制御手法を提案する．差動二輪車とは，車両型のロボットであり，左右の車輪を独立に制御することによって，前後左右に動作する事が可能である．現実の環境において差動二輪車を制御するためには，内乱と外乱（内外乱）に適応することが必要である．内乱とは，車輪の調子のような車体の内部から発生する干渉である．外乱とは，路面の凹凸のような外部から発生する干渉である．内外乱に適応するために過去の行動情報を考慮した手法を提案した．提案手法では，過去の行動情報を含む入力値の拡張手法を用いた．評価では，現在の観測情報のみを考慮した従来手法と比較し，提案手法のほうが成功率という評価尺度において 30%高い結果となった．

キーワード　強化学習，差動二輪車，安定制御，内外乱適応

1　はじめに

近年，現実の環境における，強化学習を用いたロボットの安定的な制御を行う手法が提案されている [1] [2] [3]．このロボットの車体として，差動二輪車を利用することがある [4]．差動二輪車は，車両型のロボットであり，左右の車輪を独立に制御することによって，前後左右に動作する事が可能である．しかし，現実の環境において，強化学習を用いてロボットを安定的に動作させるには，内乱と外乱（内外乱）への適応が必要である．内乱とは，車両の内部から発生する問題のことである．例えば，車輪の調子やバッテリーの寿命などが挙げられる．外乱とは，車両とは関係ない外部の要因により発生する問題のことである．例えば，路面の凹凸や風などが挙げられる．

一般的な強化学習では現実の環境のように内外乱を考慮していない．そのため，内外乱を考慮したモデルを学習する必要がある．しかし，内外乱の種類は多様であり，それら全ての情報を特徴量としてモデルに入力することはできない．そのため，行動情報から内外乱を自動的に認識する必要がある．そこで本研究では，過去の観測値と行動情報をモデルに入力することで，過去の行動の状態遷移から，内外乱の状況を自動的に認識する手法を提案する．

本研究では，過去の観測情報と行動，現在の観測情報を入力とし，次の行動を出力するモデルを提案した．比較手法として，現在の行動のみを入力とした従来手法と比較し，提案手法のほうが成功率という評価尺度において 30%高い結果となった．

2　関連研究

本研究では，強化学習を用いて内外乱が存在する環境での差動二輪車の安定制御を行う．関連研究として，外乱除去制御による安定制御手法を行う研究や外乱値を含む拡張状態を用いて外乱適応を行う研究が挙げられる．

2.1　強化学習による差動二輪車制御

Gou ら [4] は，外乱やモデルパラメータの不確実性が存在する環境で，差動二輪車の軌道追跡制御のための，Clipped Double Q-learning と長短期記憶 (Long Short-Term Memory; LSTM) [5] に基づく深層決定的方策勾配法（Deep Deterministic Policy Gradient; DDPG）アルゴリズム [6] を提案した．Clipped Double Q-learning では，DDPG アルゴリズムにおける状態行動価値の過大評価の問題を解決するために導入されている．また，LSTM ニューラルネットワークでは，強化学習制御における部分観測問題を解決するために使用している．過去の状態情報を使用し，Actor によって現在の行動を生成している．実験では，提案されたアルゴリズムを使用した場合と，従来の PID 制御器を使用した場合の性能を比較を行っている．しかし，実験では特定される 1 つの外乱を用いて評価を行っている．

本研究では，実環境で発生する未知の干渉に適応することを目的としており，実験では車輪の調子と路面の凹凸の複数の内外乱を用いて安定した制御が行えているかを評価する．また，Gou らの手法では，過去の状態情報をモデルへの入力としているが，本研究では過去の状態情報に加え，行動情報も入力する．

2.2　外乱オブザーバを用いた外乱適応制御

Cheng ら [7] は，外乱オブザーバ（Disturbance observer; DOB）と安全フィルタとして制御バリア関数（Control Barrier Function; CBF）を使用した安全かつ効率的な強化学習手法を提案している．従来の安全フィルターに基づくアプローチは，通常，不確かなダイナミクスの学習と学習されたモデルの誤差の定量化を含む．これにより，良いモデルを学習するために多くのデータが収集される前に保守的なフィルターが生じ，効率的な探索が妨げられる．

Malmir ら [8] は，シミュレーションから実世界のロボット制御に対するポリシー転移を可能にするため，外乱感知型強化

学習アプローチの提案をしている．DOB を用いて外乱を推定し，その推定値を用いて制御ポリシーを学習することで，制御システムの堅牢性を向上させる手法である．実験では，ロボットアームを使用した簡単な到達タスクを用いて，外乱感知を行わないモデルとの比較実験を行った．結果では，ランダムな外乱に対して，より堅牢な行動を行い，安定性が向上した．しかし，外乱の推定値を計算する DOB では，パラメータ調整が必要であることや計算コストが高い．

2.3 GRU を用いた外乱適応のための強化学習

Wenjie ら [9] は，オブザーバネットワークとコントローラネットワークの 2 つのモジュールを用いて，外乱に適応した強化学習手法を提案している．オブザーバネットワークは，外乱オブザーバをゲート付き再帰ユニット (Gated Recurrent Unit; GRU) [10] で模倣した外乱推定器であり，教師ありで学習が行われる．コントローラネットワークでは，強化学習形式で学習され，学習したオブザーバネットワークで推定した外乱値を状態に加えて強化学習することで，外乱に適応した制御をしている．この研究では，オブザーバネットワークとコントローラネットワークの 2 つのモジュールが別々で学習され，その後 2 つを連結して再度学習を行う，2 段階の学習アプローチを取り入れている．

2.4 PPO（Proximal Policy Optimization）

本節では，本研究で用いる強化学習アルゴリズムの PPO（Proximal Policy Optimization）[11] について述べる．PPO は 2017 年に Schulman らによって提案され，サンプル効率の良さ，実装の容易さ，さまざまな環境での堅牢な性能で知られている．

PPO では，行動 a を決定する Actor と，方策 π を評価する Critic で構成される，Actor-Critic のアプローチが用いられている．具体的には，エージェントが観測した状態 s を入力とし，Actor としての方策 $\pi\theta$ と，Critic としての状価値関数を同時に推定し，これを学習するものである．Actor-Critic 系の強化学習アルゴリズムには，A3C（Asynchronous Advantage Actor-Critic）[12] や A2C（Advantage Actor-Critic）[12] などがある．しかし，これらの手法は一度でも方策が大幅に劣化してしまうと，その後は報酬が得られにくくなり方策関数の改善も困難になるなど，学習が不安定になる問題がある．学習の安定性を向上させるために，PPO の前身である TRPO（Trust Region Policy Optimization）[13] は，更新前と更新後の方策との間の KL-divergence に制約を課すのに対し，PPO は更新前と更新後の方策の比をクリッピングするという非常にシンプルな手法となっている．これにより，PPO は実装が非常に容易であり，安定した高い学習性能を有しているといった特徴がある．本研究では Stable Baselines3 [1]の PPO をベースとして実装を行う．

1：https://github.com/DLR-RM/stable-baselines3

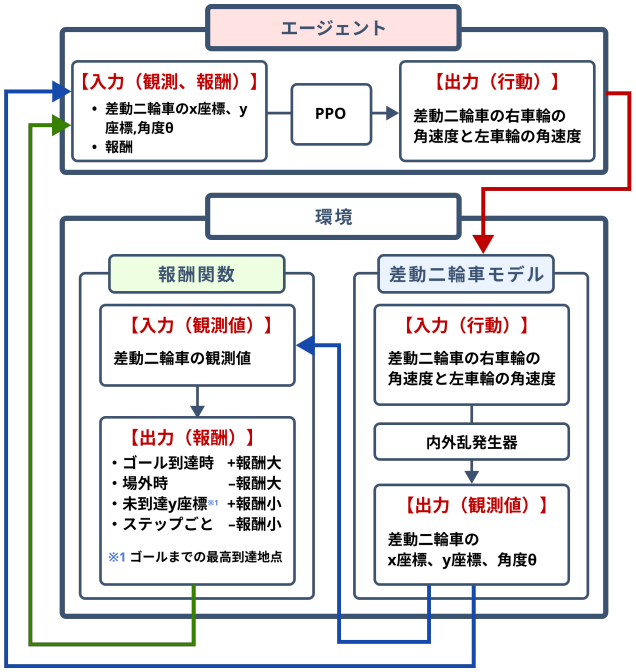


図 1　エージェントと環境の相互作用

3 実環境を想定した差動二輪車の走行シミュレーション

我々の手法では，過去の行動情報を考慮した内外乱への適応手法を提案する．本研究では，実環境での動作を想定している．しかし,「コスト効率」や「有限性」,「問題の簡略化」の観点から，まずはシミュレーション環境上で行うこととする．実環境を想定した差動二輪車の走行シミュレーション環境において強化学習を行う．本節では，本シミュレーション環境における問題定義とシミュレーション環境の説明を行う．

3.1 問 題 定 義

強化学習とは，学習者であるエージェントが環境との相互作用を通して，最適な行動方針 (方策 π) を推定する手法のことである．強化学習では，環境の設定として，マルコフ決定過程 (Markov Decision Process; MDP) [14] が多くの場合利用される．強化学習では，図 1 に示すように環境，エージェント，行動，観測値，報酬の 5 つの要素が必要である．以下に，それぞれの要素について説明を行う．

環境は，エージェントが行動を選択するための空間である．環境はエージェントの選択する行動 a_t と現在の観測値 o_t をもとに，エージェントに次の時点での観測値 o_{t+1} を返す．また，行動によって得た次の時点での状態 o_{t+1} から，エージェントに報酬 r_{t+1} を与える．

エージェントとは，学習の主体者および意思決定者のことである．エージェントは環境から観測値と報酬を入力として受け取り，その情報に基づいて行動を選択する．

行動は，エージェントが環境に対して選択できる手段である．ある時点 t における行動 a_t は環境の現在の状態 s_t を変化させ，

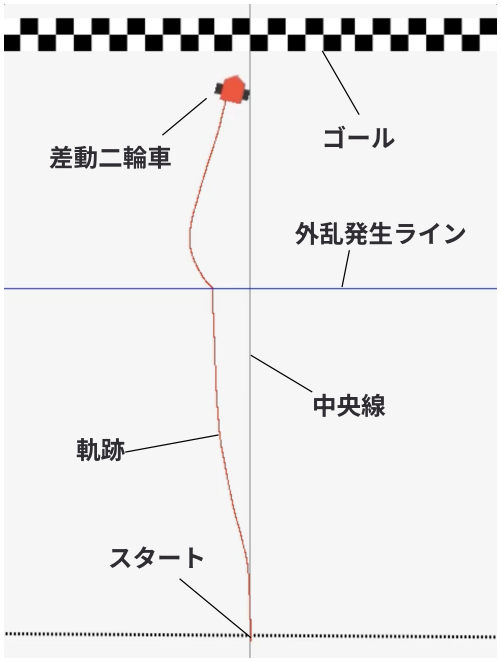


図 2　Straight Line Car

環境を次の状態 s_{t+1} に変化させる．エージェントは，長期的に報酬が高くなるように最適な行動を学習する．

観測値は，エージェントが環境内で観測ができる状態のことである．内外乱のような観測できない状態がある環境では，観測値を入力値とする．

報酬とは，エージェントが取った行動 a_t をもとに，環境がエージェントに返す行動の即時的な価値のことである．エージェントはこの報酬をもとに，長期的に価値が高くなるように行動を学習する．

3.2　シミュレーション環境の設定

本研究では，エージェントが学習を行うための環境として，Straight Line Car というシミュレーション環境を自作した．シミュレーション画面を図 2 に示す．Straight Line Car は 640 × 480 のサイズの環境である．本環境では，差動二輪車をスタート地点 $(x_s, y_s) = (240, 620)$ からゴールライン $y_g = 20$ までをできる限り早く，中央線を通るように制御することを目的とする．本環境は，x 座標 (x)，y 座標 (y)，角度 (θ)，外乱値 (d)，内乱値 (c_r, c_l) の 5 つの状態で構成されている．

$$s = [x, y, \theta, c_r, c_l, d], \tag{1}$$
$$0 \leqq x \leqq 480, 0 \leqq y \leqq 640, 0 \leqq \theta \leqq 360,$$
$$d = [-50, 50], 0.3 \leqq c_r, c_l \leqq 1.0$$

本研究において，エージェントは差動二輪車の動作制御を行い，スタート地点からできる限り中央線を通りながら，ゴールラインを目指す．エージェントは行動として，左右の車輪の角速度（w_r, w_l）をそれぞれ選択する (2)．

$$a_t = [w_r, w_l], \quad -1 \leqq w_r, w_l \leqq 1 \tag{2}$$

エージェントが観測できる状態には，x 座標，y 座標，角度（θ）の 3 つが含まれている．

$$o_t = [x_t, y_t, \theta_t] \tag{3}$$
$$0 \leqq x \leqq 480, 0 \leqq y \leqq 640, 0 \leqq \theta \leqq 360$$

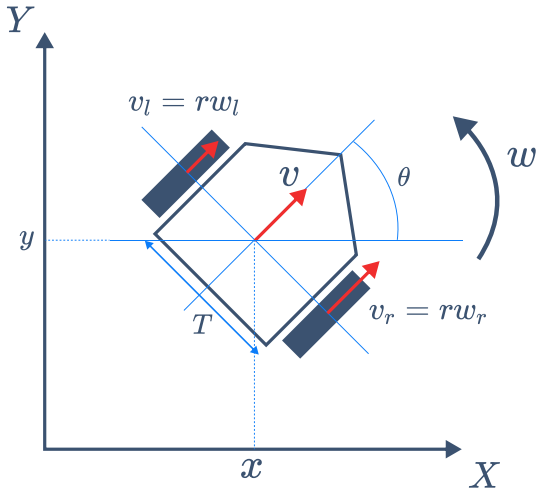


図 3　差動二輪車モデル

表 1　差動二輪車モデルに用いる記号の定義

x, y	位置座標 [px]
θ	方位角 [rad]
v	進行速度 [px/s]
w	旋回角速度 [rad/s]
w_r	右車輪の回転角速度 [rad/s]
w_l	左車輪の回転角速度 [rad/s]
v_r	右車輪の進行速度 [px/s]
v_l	左車輪の進行速度 [px/s]
r	車輪の半径 [px]
T	左右車輪間のトレッド [px]

シミュレーション環境で用いる，制御対象の差動二輪車の位置推定モデルについて述べる．用いる位置推定モデルでは，以下の 4 つを前提条件とする．

- 車輪は車軸に垂直な方向にのみ回転する
- 車軸方向へのドリフトはしない
- 運動中は各瞬間で旋回中心が存在する
- 全ての車軸の回転軸は旋回中心を通る

また，差動二輪車は左右の駆動輪のほかに，バランスをサポートするための自由輪を持つこととする．そのため，バランスを取るための制御は必要としない．

考慮する数値を表した差動二輪車モデルを図 3 に示す．また，図 3 中の記号を表 1 に定義する．

エージェントの行動である w_r, w_l を用いて，進行速度（v）と旋回角速度（w）を求める式を (4) に示す．

$$\begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} r/2 & r/2 \\ r/T & -r/T \end{bmatrix} \begin{bmatrix} w_r \\ w_l \end{bmatrix} \tag{4}$$

式 (4) で求めた進行速度 (v)，旋回角速度 (w) と現在の状態 x_t, y_t, θ_t から次の状態 $x_{t+1}, y_{t+1}, \theta_{t+1}$ を求める式を (5) に示す．

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ \theta_t \end{bmatrix} + \begin{bmatrix} cos\theta_t \Delta\ t & 0 \\ cos\theta_t \Delta\ t & 0 \\ 0 & \Delta\ t \end{bmatrix} \begin{bmatrix} v_t \\ w_t \end{bmatrix} \tag{5}$$

本研究では，内乱として，差動二輪車の左右の車輪の調子（c^r, c^l）を扱う．本研究では，車輪の調子を 0.3 以上 1.0 以下の浮動小数点をランダムで取る．

$$0.3 \leqq c_r, c_l \leqq 1.0 \tag{6}$$

この調子の値は，エージェントが行った行動に掛けることで，調子を反映させる．調子が加わった場合，エージェントの実際の行動 (w_r^t, w_l^t) は式 (7) のようになる．

$$w_r^t, w_l^t = w_r \times c_r, w_l \times c_l \tag{7}$$

なお，内乱の値はエピソードごとに変化し，1 エピソード内では一定とする．また，エージェントは調子の値が掛けられた行動を観測することはできず，調子の値の観測もできない．調子の値が掛けられる前のエージェントが出力した行動のみを把握する．

外乱としては，路面の凹凸によって発生する車体の向きの急変化を扱う．シミュレーション環境では，路面の凹凸を図 2 の外乱発生ラインとする．なお，この外乱発生ラインは 1 エピソード内に 1 つ発生し，外乱発生場所（y_d）は $30 \leqq y_d \leqq 600$ でランダムに発生する．また，外乱値は -50 または 50 のどちらかをエピソードごとにランダムで取る．

外乱が加わる条件を式 (8) に示す．外乱発生ライン y_d が前の状態の y 座標 y_{t-1} 以上，現在の状態の y 座標 y_t 未満の場合とする．

$$y_t \leqq y_d < y_{t+1} \tag{8}$$

状態に外乱を加える方法は，状態 s_t の角度 θ_t に外乱値を加算することとする．外乱発生ライン (y_d) に到達した場合，外乱加算前の状態が $s_b = [x_b, y_b, \theta_b]$ とすると，外乱加算後の状態は $s_a = [x_a, y_a, \theta_a + d]$ となる．

4 内外乱に適応した過去の行動情報を考慮した MLP モデル

本研究では，強化学習を用いて，内外乱に適応した安定制御を行う．内外乱が存在する環境において，従来の強化学習で安定的に制御することは難しい．その原因として，内外乱は予測不可能であり，直接観測できないことが挙げられる．そこで提案手法では，過去の観測値と行動を含めた拡張観測値 $(obs_{t-1}, a_{t-1}, obs_t)$ を入力値とすることによって，内外乱の間接的な観測を行う．

本研究では Stable Baselines3 [2]の PPO をベースとして実装を行う．エージェントネットワークでは図 4 に示すように，Actor と Critic ネットワークで構成される．入力は，過去の行動情報と観測情報をスタックする CustomFrameStack という手法を用い，8 次元に拡張される．Actor と Critic は 3 層のネットワークであり，Actor では最終的に 2 次元，Critic では最終的に 1 次元のベクトルを出力する．

2：https://github.com/DLR-RM/stable-baselines3

表 2　学習時のハイパーパラメータ

パラメータ名	値
total timesteps	1,000,000
learning rate	0.0003
batch size	128
epochs	10
gamma	0.99
λ (GAE)	0.95
clip range	0.2
optimizer	Adam [15]
n_steps	2,048

表 3　報 酬 設 計

報酬付与条件	報酬値
ゴールする	$100 - 100 * \lvert x_t - 240 \rvert / 240$
場外に出る	-100
未到達 y 座標を走行	$1 - \lvert x_t - 240 \rvert / 240$
step ごと (以上に当てはまらない場合)	-1

5 比 較 実 験

本節では，過去 2 時点以上の観測値と行動値を含む，入力値の拡張手法（CustomFrameStack）を用いることで，内外乱適応にどの程度有効かを検証する．有効性を検証するために，提案手法を 1 時点での観測値を入力とした場合（PPO）と，2 時点以上の観測値を入力とした場合（FrameStackPPO）とで比較を行った．比較実験では,「車輪の調子（0.7〜1.0）」,「車輪の調子（0.5〜1.0）」,「車輪の調子（0.3〜1.0）」,「車輪の調子（0.3〜1.0）+路面の凹凸（±50）」の 4 つのタスクを用いて行った．それぞれのタスクでは，1,000,000 ステップ学習を行った後，1,000 エピソード試行した結果を実験結果とした．学習時に用いたハイパーパラメータは表 2 のように設定した．

5.1 報 酬 設 計

報酬設計は，エージェントの観測値 $o_t = [x_t, y_t, \theta_t]$ を内外乱を把握した観測値 $o'_t = [x_t, y_t, \theta_t, c_r, c_l, d_t]$ に変更した上で中央線近くを走行するように行った．報酬設計は表 3 の通りである．

「未到達 y 座標を走行」では，エージェントがスタートからゴールまで走る中で，1 エピソード内での最高到達 y 座標が更新された場合のみに，報酬が与えられる．また,「ゴールする」と「未到達 y 座標を走行」の報酬では，中央線からの絶対値距離を報酬に用いることで中央近くであれば+1 の報酬，中央から遠ざかれば 0 の報酬が与えられるようにし，中央線近くを走行させるように設計した．

5.2 評 価 方 法

実験結果の評価では，定量的な評価として,「成功率」「平均

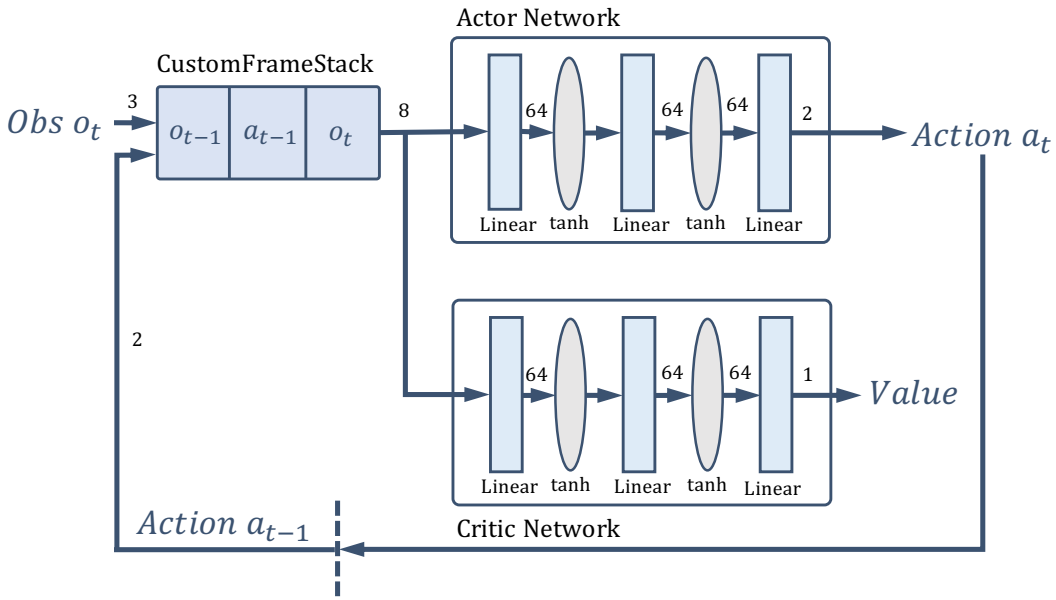


図 4　CustomFrameStackPPO のネットワーク

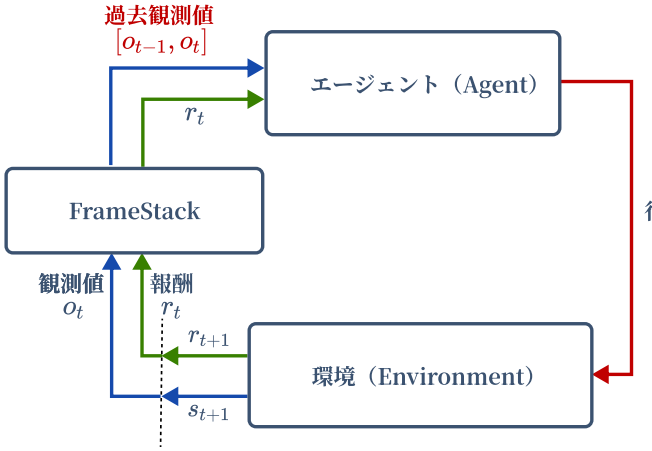


図 5　FrameStack

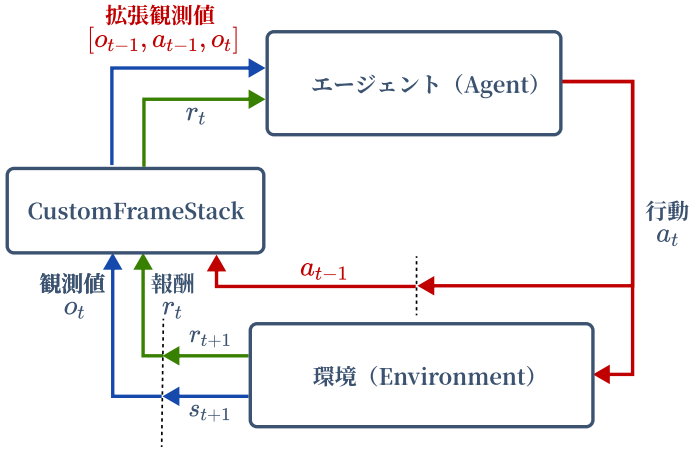


図 6　CustomFrameStack

ステップ誤差」「平均エピソード長」「平均成功エピソード長」の 4 つの指標を用いて行った．

「成功率（*success rate*）」は，試行した全エピソードの内，どれくらいのエピソードでゴールに到達できたかを表す指標である．算出方法を式 (9) に示す．ゴールに到達したエピソード数 $e_{success}$ を実行した全エピソード数 e で割ることで算出する．

$$success\ rate = \frac{e_{success}}{e} \tag{9}$$

「平均ステップ誤差（*mean step error*）」は，どの程度中央線近くを走行しているかを表す指標である．算出方法を式 (10) に示す．ステップごとの中央線 x_{middle} と現在 x 座標 x_t との絶対値誤差の合計をステップ数 s で割ることで算出する．

$$mean\ step\ error = \frac{1}{e}\sum_{i=1}^{e}\frac{1}{s_i}\sum_{j=1}^{s}|x_{middle} - x_{ij}| \tag{10}$$

「平均エピソード長（*mean episode length*）」と「平均成功エピソード長（*mean success episode length*）」は，どちらもゴールするまでの速さを表す指標である．「平均エピソード長」は，スタートからゴールまでに，平均してどれくらいのステップ数を行っているかを表す指標である．算出方法を式 (11) に示す．各エピソードのステップ数 s の合計をエピソード数 e で割ることで算出する．なお，場外に出てエピソードが終了した場合は，エピソードが終了したステップ数 s ではなく，最大ステップ数 s^{max} を用いて算出を行った．

$$mean\ episode\ length = \frac{1}{e}\sum_{i=1}^{e} s_i \tag{11}$$

$$s_i = \begin{cases} s_i & \text{ゴールした場合} \\ s_i^{max} & \text{場外に出た場合} \end{cases} \tag{12}$$

「平均成功エピソード長（*mean success episode length*）」は，成功したエピソードで平均してどれくらいのステップ数で

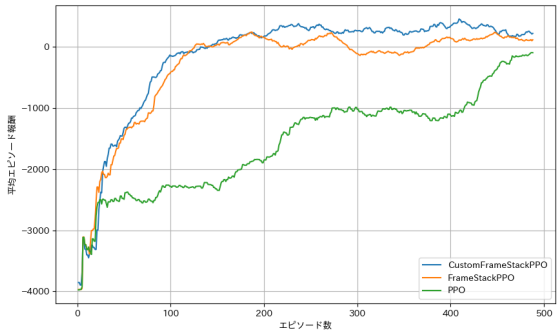


図 7　車輪の調子が 0.3〜1.0+路面の凹凸のタスクでの学習曲線

ゴールしたかを表す．算出方法を式 (13) に示す．成功したエピソードのステップ数 $s^{success}$ を成功したエピソード数 $e^{success}$ で割ることで算出する．

$$mean\ success\ episode\ length = \frac{1}{e^{success}} \sum_{i=1}^{e^{success}} s_i^{success} \tag{13}$$

また，定性的な評価として，シミュレーション環境での差動二輪車の行動軌跡を確認した．

5.3　結果と考察

提案手法（2 時点以上の観測値と行動値）と比較手法（2 時点以上の観測値，1 時点の観測値）で行った実験結果から 4 つの評価指標の数値を表 7 に示す．

結果として，成功率においては，車輪の調子は 30%以上 100%以下のランダムな値を取り，路面の凹凸を加えた場合には，提案手法である CustomFrameStack の成功率は 0.76 となり，他の手法より約 30%上回った．しかし，車輪の調子が 50%〜100%や 30%〜100%の場合は，比較手法よりも低い結果となった．また，どの程度中心近くを走行しているかを表す平均ステップ誤差においても，車輪の調子は 30%以上 100%以下のランダムな値を取り，路面の凹凸を加えた場合では，CustomFrameStack は 67.91 であり，他の手法と比較して，平均して約 2〜11px 中心を走行出来ていることが分かった．しかし，車輪の調子が 70%以上や 50%以上の場合では，比較手法の方が良い結果となった．平均エピソード長や平均成功エピソード長における結果では，他の手法と比較して，車輪の調子が 30%以上かつ路面の凹凸がある場合では，エピソード長が短くなったが，その他のタスクでは，比較的にエピソード長が長くなった．

図 8 は「車輪の調子（0.3〜1.0）+路面の凹凸（±50）」の場合での学習後の行動軌跡である．縦軸では，差動二輪車の右車輪の調子（c_r），左車輪の調子（c_l）の数理を表す．5 つの例では，CustomFrameStack の場合，全てのエピソードでゴールすることができている．しかし，PPO と FrameStackPPO の場合では，CustomFrameStack と比較して，不安定な行動の軌跡となっている．特に，c_r と c_l の差が大きい場合において，安定した行動制御が行えていないことが見て取れる．

6　まとめと今後の課題

本論文では強化学習を用いて，内外乱に適応した安定動作制御手法を提案した．手法として，従来の入力値である 1 時点の観測値から，2 時点以上の観測値と行動を含めた拡張観測値を入力とすることで，モデルに過去の観測値と行動の傾向を与えることができ，間接的に内外乱の認識を行った．

結果として，「車輪の調子（0.3〜1.0）+路面の凹凸」の難易度の高いタスクにおいて，提案手法である CustomFrameStack を用いた場合，そうでない場合と比較して成功率が 30%向上した．しかし，難易度の低いタスクや，平均エピソード長や平均ステップ誤差においては，比較手法より悪い結果となった．

今後の課題としては，今回用いた車輪の調子と路面の凹凸のみでなく，あらゆる干渉において，安定的に動作制御を行うことが可能な手法を検討する必要があると考えている．また，未知未踏の環境においても，自律的に学習を行い，あらゆる環境下での内外乱に適応できる自律進化機能の実現を目標として取り組む予定である．

謝　　辞

本研究は，JSPS 科研費 JP21H03775，JP22H03905 の助成，ならびに，JST【ムーンショット型研究開発事業】【JPMJMS2238】の支援を受けたものです．ここに記して謝意を表します．

文　　献

[1] Marc-Andrė Blais and Moulay A. Akhloufi. Reinforcement Learning for Swarm Robotics: An Overview of Applications, Algorithms and Simulators. *Cognitive Robotics*, Vol. 3, pp. 226–256, 2023.

[2] Mohan Yogeswaran, S.G. Ponnambalam, and Ganesan Kanagaraj. Reinforcement learning in swarm-robotics for multi-agent foraging-task domain. In *Proceedings of the 2013 IEEE Symposium on Swarm Intelligence*, pp. 15–21, 2013.

[3] Xudong Zhu, Fan Zhang, and Hui Li. Swarm Deep Reinforcement Learning for Robotic Manipulation. *Procedia Computer Science*, Vol. 198, pp. 472–479, 2022.

[4] Wenyao Gou and Yan Liu. Trajectory Tracking Control of Wheeled Mobile Robot Based on Improved LSTM-DDPG Algorithm. *Journal of Physics: Conference Series*, Vol. 2303, No. 1, 2022.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[6] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[7] Yikun Cheng, Pan Zhao, and Naira Hovakimyan. Safe and Efficient Reinforcement Learning using Disturbance-Observer-Based Control Barrier Functions. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, Vol. 211, pp. 104–115, 2023.

[8] Mohammadhossein Malmir, Josip Josifovski, Noah Klarmann, and Alois Knoll. DiAReL: Reinforcement Learning with Disturbance Awareness for Robust Sim2Real

表 4　各手法のタスクごとの成功率

手法	車輪の調子:0.7~1.0	車輪の調子:0.5~1.0	車輪の調子:0.3~1.0	車輪の調子:0.3~1.0 +路面の凹凸
CustomFrameStackPPO	**1.00**	**0.89**	**0.66**	**0.76**
FrameStackPPO	0.98	0.81	0.83	0.45
PPO	0.99	0.98	0.34	0.41

表 5　各手法のタスクごとの平均エピソード長

手法	車輪の調子:0.7~1.0	車輪の調子:0.5~1.0	車輪の調子:0.3~1.0	車輪の調子:0.3~1.0 +路面の凹凸
CustomFrameStackPPO	**1021.06**	**1307.95**	**2136.42**	**1896.88**
FrameStackPPO	987.36	1654.22	1787.16	2696.32
PPO	1006.85	975.71	2986.60	2854.19

表 6　各手法のタスクごとの平均成功エピソード長

手法＼タスク	車輪の調子:0.7~1.0	車輪の調子:0.5~1.0	車輪の調子:0.3~1.0	車輪の調子:0.3~1.0 +路面の凹凸
CustomFrameStackPPO	**1021.06**	**988.75**	**1172.11**	**1232.74**
FrameStackPPO	925.87	1100.39	1330.71	1109.36
PPO	979.66	926.53	1019.41	1232.35

表 7　平均ステップ誤差

手法＼タスク	車輪の調子:0.7~1.0	車輪の調子:0.5~1.0	車輪の調子:0.3~1.0	車輪の調子:0.3~1.0 +路面の凹凸
CustomFrameStackPPO	**12.62**	**59.98**	**72.56**	**67.91**
FrameStackPPO	10.28	24.93	60.78	70.25
PPO	9.18	21.45	83.95	88.12

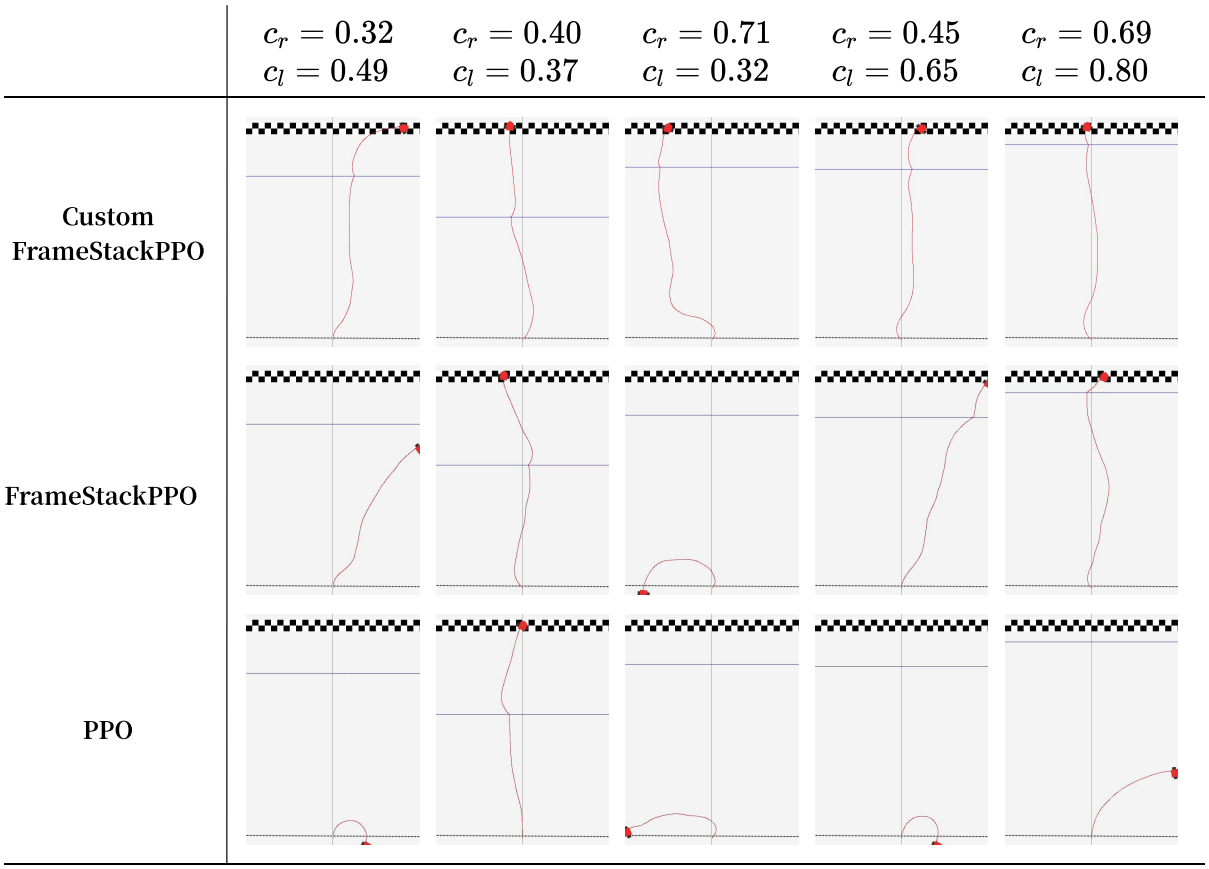


図 8　学習後の行動軌跡

Policy Transfer in Robot Control. *arXiv preprint arXiv:2306.09010*, 2023.

[9] Wenjie Lu and Manman Hu. Disturbance-aware Reinforcement Learning for Rejecting Excessive Disturbances. *Robotics and Autonomous Systems*, 2023.

[10] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *Computing Research Repository*, 2014.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Rad-

ford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *Computing Research Repository*, 2017.

[12] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1928–1937, 2016.

[13] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

[14] Richard Bellman. A Markovian Decision Process. *Indiana University Mathematics Journal*, Vol. 6, pp. 679–684, 1957.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

A Forecasting Method of Traffic Flow Using Road Network Data

Haruka TANAKA† and Daisuke KASAMATSU†

† Graduate School of Science and Engineering, Soka University
Tangi 1–236, Hachioji, Tokyo, Japan
E-mail: †e23M5319@soka-u.jp, ††dxk@ieee.org

Abstract Over the last few years, predicting traffic flow in urban space has become a common task of urban computing. A city is divided into grids and the amounts of traffic objects flowing into and out of each grid per unit of time are measured to predict the future traffic flow. This study aims to improve the prediction accuracy by adding three methods: *Grid Transformation*,*GIS subnet*, and a *Mask Processing Layer* to the conventional method to input the road network, RN. The *Grid Transformation* transforms the RN, which are graph data, into tensors according to the grid areas. The *GIS Subnet* consists of two convolutional layers and learns topographic information from the tensors. The *Mask Processing Layer* replaces the prediction values in impassable grids with zeros. These proposed methods were validated on New York City taxi data and improved prediction accuracy.

Key words GIS; smart city; mobility; traffic state prediction; deep learning; data mining; embedding; road networks

1 Introduction

Recently, it has been a common task to forecast the traffic states of city areas in urban computing [1].

Conventionally, the dynamics of traffic states is regarded as time sequences, and statistical methods have been proposed to capture their features. However, traffic patterns in the real world behave differently depending on neighborhood conditions, weather conditions, day of the week, and seasonal events such as vacation seasons and new year seasons. To include that information in prediction sources, deep learning techniques are suitable because they learn the relationships between input data and output data flexibly and effectively using the forward and backward propagation mechanism. Since the techniques began to gather attention, some prediction methods using deep learning have increased and they were designed to solve the prediction task under the following situation. Traffic condition prediction is the problem of dividing a city into grids or networks and predicting the traffic condition, speed, flow, etc. at each node. At the time of the initial period, research was mainly grid-based, which is compatible with tensor processing in deep learning. In the latent papers, graph neural networks (GNNs) that can handle more flexible data structures have appeared, and approaches that match the shape of actual urban spaces are increasing. This approach enables researchers to include the road network structure into prediction effectively, but also causes privacy problems because details of users' trajectories are required to use the models. Therefore, grid-based predictions are also important to keep the required data anonymous, and new methods that can learn the road network structure efficiently are needed to enhance the prediction accuracy of the grid-based approach.

This paper proposes a novel forecasting method for future traffic states. The method has three addiction processes to enable deep learning to learn the structure of the road network. Our experiments verify the effectiveness with real data of taxis in New York City.

2 Related Work

This section discusses conventional works that aim to predict future traffic states from past data by using deep learning techniques. To capture the features of traffic, researchers analyzed amounts of flows, and averages of vehicle speed and congestion conventionally and called those data traffic states. Traffic states in urban areas are changing continuously. They are affected by traffic states in the past, the states of adjacent areas, and external factors such as weather, vacations, information on the day of the week, and so on. Therefore, the main focus of previous research was on how to extract the temporal features, spatial features, and external features.

ST-ResNet [2]. Zhang et al. proposed a simple structured deep learning method that consists of three residual convolutional networks to process traffic state data and a fully connected layer subnet to process external factors. They divided the past data into a trend window, period window, and closeness window and inserted them into residual convolutional networks respectively. The model outperformed

in experiments with two real-world datasets, Bike NYC and taxi BJ.

3D-CLoST [3]. Fiorini et al. proposed a unified model that processes the three time windows simultaneously. It is built with three-dimensional convolutional networks and long- and short-term memory units, which effectively learn spatial features and temporal features, respectively. In addition, they added a mask layer to modify the prediction values in impassable areas such as rivers and lakes. Mask areas were defined from previous data where no vehicle was entered.

3 Preliminaries

3.1 Definitions

Definition 3.1 (*Grid Area*). A target city is divided into small grids based on longitude and latitude. W indicates the number of grids in the longitude direction and H indicates the number of grids in the latitude direction. The width of each grid is defined by dividing the width of the target city area into W at equal intervals and the height is defined by dividing the height of the target city area into H at equal intervals. Each grid is denoted as *grid* $a_{i,j} \in A$, A is a set of grids, and i, j indicates the index of width direction and height direction, respectively.

Definition 3.2 (*Traffic Flow*). Traffic Flow, *TF*, represents the number of traffic vehicles that pass through a grid during the unit time. *TF* is categorized into *Inflow* and *Outflow*. *Inflow* is the number of vehicles that move from other grids to a grid, and *Outflow* is the number of vehicles that move from a grid to other grids. *TF* at time t is denoted as a multidimensional tensor $TF_t \in \mathbb{R}^{2\times W\times H}$.

Definition 3.3 (*Time Sequential Traffic Flow*). The time window which retrieves the sets of time sequences traffic from time series data $\{TF_1, TF_2, ..., TF_T\}$ is defined with three-time ranges: *distant*, *near*, *rescent*. Each range includes the time difference d and length l and a time sequential traffic flow, $TSTF$, at time t is defined as follows.

$$TSTF_t = \{\ TF_{t-distant_d}, \ldots, TF_{t-distant_d+distant_l}, \\ TF_{t-near_d}, \ldots, TF_{t-near_d+near_l}, \\ TF_{t-resecnt_d}, \ldots, TF_t\ \}$$

Definition 3.4 (*Road Network*). Road networks of the actual cities are complex and widely spread. To be treated as computational data, a road network is represented as graph data. In this work, the road network, *RN*, means graph data $RN = \{V, E\}$, where V is the set of geometric points and E is the set of roads.

Definition 3.5 (*Mask*). $Mask \in \{0, 1\}^{W\times H}$ is a tensor which consists of binary values. Ones in *Mask* mean the grids in A at the same indexes are passable and zeros in *Mask* mean the grids in A at the same indexes are impassible.

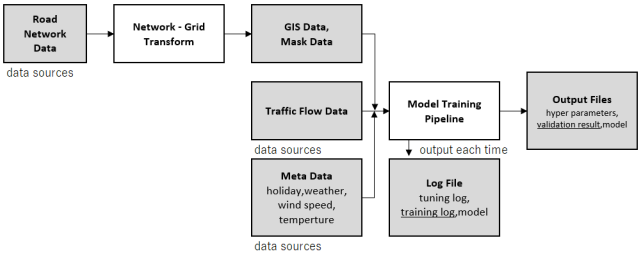


Figure 1 The data flow chart of the proposed system.

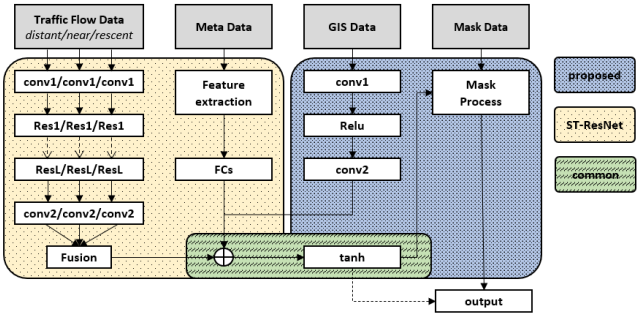


Figure 2 The architecture of proposed model.

Definition 3.6 (*Metadata*). Metadata is commonly used to predict traffic flow in previous studies. In this paper, metadata consists of meteorological data and day types.

Meteorological data represents the features of weather that affect the patterns of traffic flow in the cities. For instance, weather information, temperatures, and wind speed of each day can be used as meteorological data.

A day type is either weekdays or holidays. The destinations of drivers on weekdays and holidays are different and the differences change the traffic flow in the cities.

3.2 Traffic State Prediction Problem

Traffic state prediction is designed to capture the temporal features and spatial features from TSTF and external features from metadata with artificial neural network components such as convolutional neural network, *CNN*, and recurrent neural network, *RNN*. On the other hand, those methods do not consider the features of the geometric topography represented by the actual road network.

Therefore, this study focuses on traffic flow prediction and aims to propose a traffic state prediction method using *RN* to improve the accuracy of the prediction.

4 Proposed Method

This study proposes three methods to append the ability to ST-ResNet to capture the features of *RN*. They are *Grid Transformation*, *GIS Subnet* and *Mask Processing Layer*. ST-ResNet is a deep learning model with a simple structure proposed in a previous study. The components and each data are shown in Figure 1 and the structure of the proposed model is shown in Figure 2.

4.1 Grid Transformation

The preprocess named *Grid Transformation* exchanges the RN into gridded data which is denoted as GIS $\in \mathbb{R}^{4\times W\times H}$ to input the RN into a deep learning component which is designed to take tensors as input. With the grid area $A = \{a_{i,j}|i,j \in \mathbb{N}, i \leqq W, j \leqq H\}$ and the road network $RN = (V, E)$, the GIS is expressed by the following formula. The function $\mathrm{E}(g)$ means the set of edges in $g = (V, E)$, and the function $\mathrm{B}(a)$ means the set of the border lines of each direction of a.

$$\begin{aligned} X_{gis_{i,j,k}} &= |\{e|e \text{ cross } b_k,\ e \in \mathrm{E}(RN),\ b_k \in \mathrm{B}(A_{i,j})\}| \quad (1)\\ \mathrm{B}(A_{i,j}) &= \{north, east, south, weast\} \\ i,j,k &\in \mathbb{N}, i \leqq W, j \leqq H, k \leqq 4. \end{aligned}$$

The result of *Grid Transformation* is used to produce *Mask*. It is defined with the following formula.

$$Mask_{i,j} = \sum_k X_{gis_{i,j}} \bmod 2. \quad (2)$$

4.2 GIS Subnet

GIS Subnet was designed to learn the features of *RN*. It takes GIS as inputs which are generated by the *Grid Transformation*. The subnet is implemented with the simplest convolutional neural network (CNN) layers using convolutional layers specialized for learning spatial characteristics and an activation function ReLU unit. The output H_{gis} has the same dimension as the predicted values.

$$H_{gis} = \mathrm{Conv2D}(\mathrm{ReLU}(\mathrm{Conv2D}(X_{gis})). \quad (3)$$

The output of *GIS Subnet* H_{gis} is added to the outputs of ST-ResNet, one is the output H_{main} of the subnet which learns the features of *TSTF* and the other is the output H_{meta} of the subnet which learns the features of *Meta Data*. They are added and fed into the activation function tanh. The output of the active function Y_{gis} is expressed by the following equation.

$$Y_{gis} = \tanh(H_{main} + H_{meta} + H_{gis}). \quad (4)$$

4.3 Mask Processing Layer

Mask processing layer aims to enhance the prediction accuracy by removing the falt prediction at the impassable areas such as lakes, rivers, and buildings by replacing the values to 0. It is placed after the activation function tanh. It replaces the values of inpassible cells in Y_{gis} with -1 using *Mask*. The output Y_{mask} of *Mask Processing Layer* is expressed by the following equation, where $\bigodot$ is the Hadamar product, and $J_{2,W,H}$ is a matrix of ones.

$$Y_{mask} = Y_{gis} \bigodot Mask + (Mask - J_{2,W,H}). \quad (5)$$

The output Y_{mask} is converted by the data post-processing defined in the previous study, which produces the predicted value $\hat{Y}$ with -1 replaced by 0.

Table 1 Configurations of Hardware and Software.

CPU	100MHz AMD Ryzen Threadripper 3960X 24-Core Processor
GPU	Nvidia RTX A5000
Memory	16GB
Hard Disk	1TB
OS	ubuntsu22.04
Anaconda	version 22.9.0
Python	version 3.8.0
Keras	version 2.4.3
Tensorflow	version 2.4.1
GeoPandas	version 0.11.1
Numpy	version 1.19.5
Pandas	version 1.4.3
h5py	version 2.10.0
BaysianOptimization	version 1.2.0

5 Experiments

5.1 Experimental Settings

To verify the effectiveness of the proposed methods, experiments were conducted with real data. The following sections explain the details of configuration, dataset, preprocessing, and scenarios.

5.1.1 Configuration

The experiments were conducted on Ubuntu 22.04 with GPU, Anaconda, and Python. The versions of hardware and software that were used in experiments are shown in table 1. GeoPandas, Numpy, Pandas, and h5py are Python libraries that are helpful for data preprocessing and BayesianOptimization is a Python library that is commonly used to optimize hyperparameters. To evaluate the prediction accuracy, the root mean square error (RMSE) is employed in this paper. This metric indicates the amount of the distances between true traffic flow and predicted values. The true value is denoted as $\overline{Y} = (\overline{I}, \overline{O})$ and the predicted value is denoted as $Y = (I, O)$. RMSE is expressed by the following equation.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{W}\sum_{j=1}^{H}\left(\left(\overline{I}_{i,j} - I_{i,j}\right)^2 + \left(\overline{O}_{i,j} - O_{i,j}\right)^2\right)}{W \times H}}. \quad (6)$$

5.1.2 Datasets

The datasets were TaxiNYC [4] and CSCL which stands for

Table 2 Traffic flow data built by preprocessing.

records	Origin	Fixed	Filled	Masked
2010	8,112	8,064	8,760	8,760
2011	8,136	8,064	8,760	8,760
2012	8,136	8,088	8,784	8,784
2013	8,352	8,064	8,760	8,760
2014	-	-	-	8,760
train	32,736	32,280	35,064	35,064
test	8,760	8,760	8,760	8,760

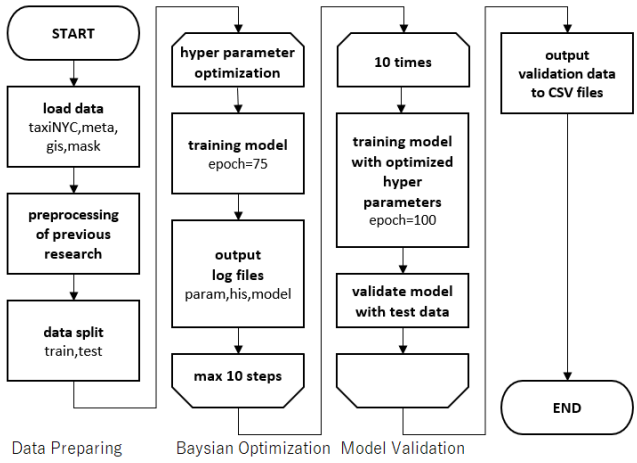


Figure 3 Training steps.

NYC Street Centerline [5]. TaxiNYC consists of the traffic flow data of taxis in New York aggregated by every one hour from 2010 to 2014 and *Meta Data* which includes *Holiday Data* and *Meteorology Data* such as one-hot encoding data, temperatures, and wind speed. CSCL is a road network data of New York City which is published as open data and the data of 2015, the oldest data, was selected in this study.

5.1.3 Preprocessing

The original traffic flow data in TaxiNYC include luck, duplications, and unexpected values. Therefore, several preprocessing were conducted. The original data of TaxiNYC was named Origin. It was processed in three steps. Fixed is the data where duplicate dates in Origin were removed. Filled is the data where missing dates in Fixed were completed by linear interpolation, and Masked is the data generated by applying mask processing to Filled. All data from 2010 to 2013 were used to train the model and the 2014 data of Masked was used to verify the performances of trained models. The whole dataset information was shown in table 2.

5.1.4 Scenarios

The eight patterns of evaluation scenarios and the two patterns of ablation scenarios were conducted.

In the evaluation studies, ST-ResNet was trained with four datasets: Origin, Fixed, Filled, and Masked, these models are referred to with labels origin, fixed, filled, and masked respectively in the result section. The following scenarios: origin2, fixed2, filled2, and masked2 indicate the results for proposed methods which were trained with Origin, Fixed, Filled, and Masked respectively.

In the ablation studies, the impacts of *GIS Subnet* and *Mask Processing Layer* by comparing the ST-ResNet with *GIS Subnet* model and ST-ResNet with *Mask Processing Layer* model, these scenarios are called with gis and mask.

In these studies, hyperparameter tuning using Bayesian optimization was conducted in each scenario. Then each model is trained and tested with optimized parameters ten times. The training steps from hyperparameter tuning to evaluation steps are shown in Figure 3.

5.2 Experimental Results

The results of evaluation studies and ablation studies are shown in Figure 4 and Figure 5 respectively. The horizontal axis are label of each scenario and the vertical axis is RMSE. Colored lines in the boxes are medians and triangles in the boxes are average scores of RMSE.

According to Figure 4, the models performed differently depending on the training data and depending on the architectures. Comparing ST-ResNet and proposed methods using the same training data, it was shown that the proposed methods outperformed ST-ResNet in most patterns.

In particular, the average score of masked2 is the lowest and its beard is the narrowest. That means the proposed method which was trained on Masked is the most stable and accurate.

To compare the impacts of each proposed component, the results of models in evaluation studies which were trained with Masked and ablation studies were shown in Figure 5. Since the average score of masked, ST-ResNet, is the highest, each component has the effect of improving the prediction accuracy. According to Figure 5, *Mask Processing Layer* has less effective than *GIS Subnet* but it enhanced the prediction

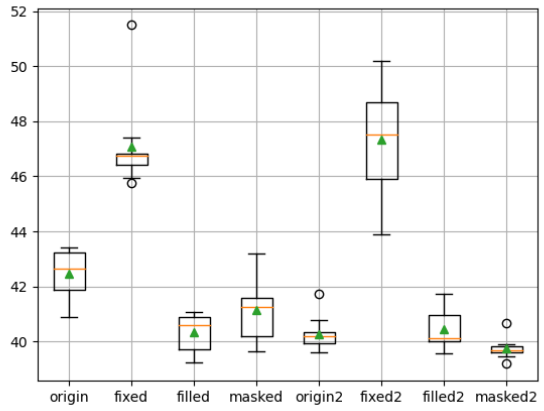


Figure 4 Experimental results of evaluation studies.

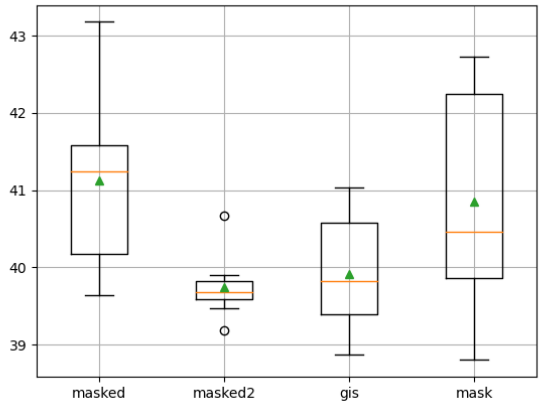


Figure 5 Experimental results of ablation studies.

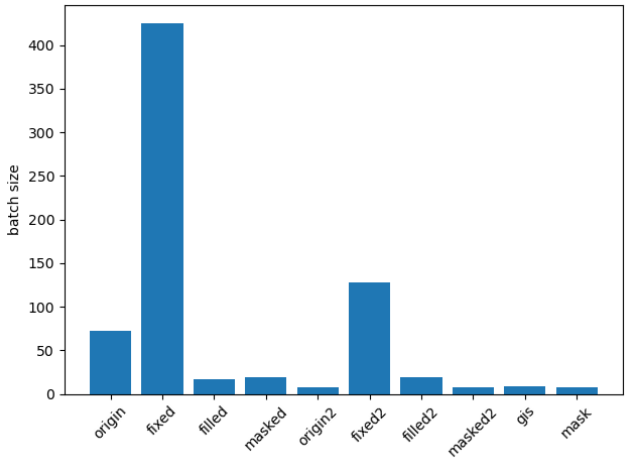


Figure 6 Optimized parameter: batch size.

ability when it was unified with *GIS Subnet*.

In evaluation studies, the scores of models which were trained with Fixed were higher than the others and the considerable factor which was found is the hyperparameter: batch size. The optimized batch sizes are shown in Figure 6. This Figure clearly shows the same patterns with the RMSE patterns. Batch size is a parameter that controls the amount of gradient valuations to update the weights of models in a training step. It enables the model to avoid overfitting but can worsen the whole accuracy.

6 Conclusion

In this study, a prediction method for traffic flow prediction was proposed to improve the accuracy by including the road network information. To use the road network data, one preprocessing unit, *Grid Transformation*, and two deep learning units, *GIS Subnet* and *Mask Processing Layer* were implemented. The proposed methods were tested with real-world data of taxis in NYC and outperformed the conventional model. In the future, additional verification using different real-world data and new methods which enable it to consider the information on road types such as highways, bridges and one-direction roads.

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban Computing: Concepts, Methodologies, and Applications,” ACM TIST, vol. 5, issue 3, no. 38, pp. 1-55, 2014.

[2] J. Zhang, Y. Zheng, and D. Qi, “Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction”, AAAI, vol. 31, no. 1, pp. 1655-1661, 2017.

[3] S. Fiorini, G. Pilotti, M. Ciavotta and A. Maurino, ”3D-CLoST: A CNN-LSTM Approach for Mobility Dynamics Prediction in Smart Cities,” IEEE International Conference on Big Data (Big Data), pp. 3180-3189, 2020.

[4] S. Fiorini, M. Ciavotta, and A. Maurino, ”Listening to the city, attentively: A Spatio-Temporal Attention Boosted Autoencoder for the Short-Term Flow Prediction Problem,” Algorithms, vol. 15, no. 10, 376, 2022.

[5] NYC openData, ”NYC Street Centerline (CSCL),” available at https://data.cityofnewyork.us/City-Government/NYC-road-Centerline-CSCL-/exjm-f27b, accessed Dec. 23, 2022.

都営バスのオープンデータおよびディープラーニングを用いた渋滞検知手法の提案と評価

畠中　希† 青柳　宏紀†† 藤田　智也†† 山名　早人†† 小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
†† 早稲田大学　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †{hatanaka.nozomi,oguchi}@is.ocha.ac.jp, ††{aoyagih,tfujita,yamana}@yama.info.waseda.ac.jp

あらまし　渋滞は大きな損失を招くため渋滞が起こっている箇所を検知し回避することは重要である．現在, 渋滞は主に道路上の感知器もしくは民間の渋滞予測サービスを使用するユーザの現在地から試算され共有されている．しかし, 感知器は全道路に設置されているわけではなく, ユーザの位置情報はデータ数に偏りがあるという問題がある．これらの問題を解決する先行研究として都営バスのオープンデータと機械学習を組み合わせた渋滞検知手法がある．都営バスは感知器がない道路も走っており, 定期的に運行しているためデータを収集することが出来る．先行研究では都営バスの出発時刻から計算された速度等を特徴量として用いて停留所区間ごとに渋滞もしくは非渋滞の二値分類を行った．本論文では青柳らの提案手法をもとに, 特徴量に定刻と実際の発車時刻との差を追加し, さらに学習器に 1 次元の畳み込みニューラルネットワークを用いた手法を提案し，評価を行った．実験の結果,

キーワード　渋滞検知, 機械学習, 公共交通機関, オープンデータ, ディープラーニング

1 はじめに

交通渋滞は様々な悪影響を及ぼしており，日本では交通渋滞によって年間 12 兆円の経済的損失，1 人あたり年間 30 時間の時間的損失が生じていると試算されている [1]．このような渋滞によって生じる損失を抑制するため，交通渋滞を検知し回避することは重要である．

現在，交通渋滞の検知は主に道路上に設置されている感知器によって行われている．しかし，一部感知器が設置されていない道路もあるため，全道路の交通渋滞を検知することは難しい．また，交通渋滞を認識する他の手法として民間企業が運営する渋滞予測サービスが挙げられる．これらのサービスの多くは，そのサービスを使用しているユーザーの走行状態を基に予測が行われている．そのため，ユーザの走行データが少ない道路では渋滞予測の精度が低くなるという問題がある．

感知器が存在しないため一部の道路では渋滞検知が出来ないという点とデータが少ないことによる渋滞予測精度の低さという点を考慮した研究として青柳ら [2] の都営バスのリアルタイム運行データと機械学習を用いた渋滞検知手法がある．彼らが実験を行った都営バスは感知器が存在しない道路上も走行し，さらに，定期的に運行しているため断続的なデータを得ることが出来る．このモデルでは，渋滞を時速 10km 以下で断続的に走行している状態と定義し，連続する 2 つの停留所の実際の発車時刻から算出した走行速度や停留区間におけるバスの移動時間の Z スコア等を特徴量とし，停留所区間ごとに「渋滞」と「非渋滞」という二値分類を行うというものである．渋滞の定義に用いられる速度だけではなく他の特徴量も使用したのは，バスは乗客の乗降による遅延など一般車両と異なる走行を行うためである．また，彼らはバスの走行速度を連続する 2 つの停留所の実際に出発した時刻を距離で割ったものから求めている．連続する 2 つの停留所の出発時刻と到着時刻からではなく出発時刻と出発時刻からバスの走行速度を算出したのはデータの取得元である API のレスポンスに都営バスの到着時刻が含まれていないためである．走行速度を出発時刻から算出しているため，時間調整による停留所での待ち合わせを行う場合，交通速度よりバスの走行速度が遅く算出され，実際は非渋滞な区間であっても渋滞であると誤判定が起こるという問題がある．

そこで本論文では青柳ら [2] の手法を参考にしバスの時刻表情報を考慮し, Convolutional Neural Networks（以下, CNN）を用いた渋滞検知手法の提案と評価を行う．バスの時刻表情報を考慮したのは，バスには時刻表が存在しており，なるべく時刻表通りに運行するという特性があるため，渋滞と非渋滞では振る舞いが異なると考えられるためである．渋滞していない場合は，運転手は走行速度を調整することができ，次の停留所でも予定通りに発車するようにすることが出来るため，実際の発車時刻と定刻との差は小さいはずだが，渋滞している場合は，運転手が走行速度を調整することは出来ず，次の停留所においても実際の発車時刻と定刻との差は大きいと考えられる．このように，実際の出発時刻と定刻との差は，渋滞と非渋滞の状態によって振る舞いが大きく異なるため入力データとして有効であると考え，提案モデルに加えた．さらに，畳み込みニューラルネットワーク（以下、CNN）を使用することで精度向上を目指した．

この論文では第 2 章では関連研究，第 3 章では提案モデル，第 4 章では実験と評価，第 5 章ではまとめと今後の課題を述べる．

2 関連研究

Xu ら [3] は，バスを複数台用いて平均移動時間を 2 種類の方法で算出し渋滞検知を行う手法を提案した．1 つ目は，渋滞検知の対象である道路区間に対して T 分間に通過した全てのバスの平均移動時間（T-window average），2 つ目は渋滞検知の対象である道路区間に対して直近に通過した N 台のバスの平均移動時間（N-window average）である．それぞれの指標にあらかじめ閾値を設定し，両方の閾値を超えた場合に渋滞が発生していることをシミュレーションにて示した．Wang ら [4] は，2013 年，1 台のプローブ車両から得られるデータと機械学習を用いた渋滞検知手法を提案した．彼らは，GPS によって取得した車両の平均速度を位置ごとに離散化し，特徴ベクトルとして使用した．

日本において渋滞は時速 10km 以下で継続して走行している状態と定義される [5]．そのため，バスのプローブデータを用いて走行速度を算出すると渋滞検知に用いることができるが，バスは乗客の乗降を行うため停車を繰り返すことから一般車両と異なる走行をするため，バスの走行速度は一般車両の交通速度を表していない．したがって，バスの速度と交通速度が等しくなるよう推定することが必要である．Samal ら [6] は，プローブ車両であるバスのリアルタイムの位置，天候データ，匿名化された運転手の識別子のデータから k-means 法を使用して類似した運行のもののクラスタリングを行い，クラスタ単位ごとにランダムフォレストで学習を行ったモデルを作成し，どのクラスタに属するかを特定した後に車両速度の推定を行うという手法を提案した．Kwan ら [7] はバスの走行パターン，燃費消費量率，走行距離，エンジンのオン/オフの状況，バスの位置を 20 秒ごと 24 時間追跡し，さらに Place of Interests（POI）を考慮した手法を提案した．POI には学校，市場・商店，ショッピングモールなどが含まれる．Gu ら [8] はバスの GPS 軌道データをもとに Self-Organizing Map（SOM）を用いて，停留所区間ごとの道路セグメントの交通渋滞を推定する手法を提案した．停留所での停車時間，バスの平均速度，乗客のバスの停留所間の移動にかかる時間を入力としてニューラルネットワークを構築した．

バスによる渋滞検知と交通速度推定の両方を行った研究として青柳らの「都バスのリアルタイム運行を用いた渋滞検知」[2] がある．彼らは特徴量として停留所の出発時間から算出したバスの平均走行速度，ある時間帯におけるある停留所区間のバスの移動時間の Z スコア，特定の時間帯におけるある停留所区間のバスの移動時間を停留区間の距離で正規化したものの 3 つをそれぞれ過去 2 つの時間区分，前 2 つの停留所区間の計 15 個のうち特徴量の選択を行い，選択された特徴量を用いて，停留所区間ごとに渋滞もしくは非渋滞を判定する二値分類を行う手法を提案した．学習器は系統ごとに 1 つの置く場合と停留所区間ごとに 1 つ置く場合の 2 種類を実験し，停留所区間ごとに 1 つ置く場合の方がより高い精度を示した．ただし，この手法では GPS ではなくオープンデータを用いてる．

本論文では青柳ら [2] の提案手法を参考にし，時刻表情報を考慮し，バスのリアルタイム運行データと機会学習を用いた渋滞検知手法を提案する．バスには時刻表が存在し，なるべく定刻に出発しようとする特性がある．非渋滞の際は，運転手が走行速度を調整することが出来るため停留所における実際の発車時刻と定刻との差は小さいが，渋滞の場合は走行速度を調整することが出来ないため実際の発車時刻と定刻のずれが大きいと考えられる．したがって，実際の発車時刻と定刻の差は渋滞時と非渋滞時でそれぞれ振る舞いが異なるため特徴量として有効であると考え，使用することにした．また他の特徴量として，渋滞の定義にも用いられている走行車両の速度と走行時の時間帯も用いる．ただし，走行車両の速度は停留所の発車時刻から算出したものである．そして，停留所 2 区間ごとに「渋滞」「非渋滞」を示した正解データを用いて学習をし二値分類を行う．なお，渋滞検知の停留所 2 区間としたのは，乗客の乗降や停留所停車時間等のノイズの影響を受けにくくするためである．また, CNN を用いて学習を行う．

3 提案モデル

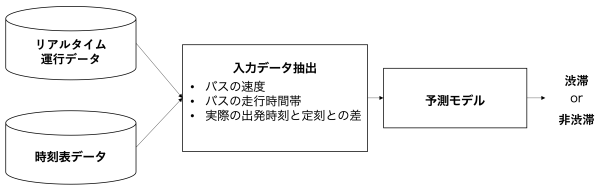


図 1 提案モデル

本論文における渋滞検知の提案モデルを図 1 に示す．まず，収集したバスのリアルタイム運行データ，時刻表データ，渋滞データをサーバに保存する．そして，サーバに保存されているデータから入力データとして使用するバスの走行速度，バスが走行している時間帯，バスの発車時刻と定刻との差を停留所 2 区間ごとに抽出する．そして，抽出したバスの走行速度，バスが走行している時間帯，バスの発車時刻と定刻との差を入力データとして 1 次元の CNN に与え，停留所ごとに渋滞もしくは非渋滞の判定を行う．

3.1 データ収集

表 1 収集したデータの情報

バスのリアルタイム運行データ		バスの時刻表データ	
構造	JSON	構造	JSON
取得元	東京都交通局のバスロケーション情報	取得元	東京都交通局の都バスのバス停時刻表
更新	30 秒ごと	更新	2022 年 12 月
渋滞の正解データ			
構造	JSON		
取得元	Directions Advanced API		
更新	5 分ごと		

本研究ではデータセット作成のためバスのリアルタイム運行データ, 時刻表データ, 渋滞データの収集を行なった．収集したデータの情報を表 1 にまとめた．

バスのリアルタイム運行データは公共交通オープンデータセ

ンターが提供している東京都交通局のバスロケーション情報の API [9] に 30 秒おきにリクエストを送り取得した．レスポンスには出発した停留所名，停留所を出発した時刻等が含まれている．ただし，停留所に到着した時間はレスポンスに含まれていない．また，このデータはオープンデータである．

バスの時刻表データも同様に公共交通オープンデータセンターによってオープンデータとして公開されており，東京都交通局の都バスのバス停時刻表 API [10] から取得した．レスポンスには発車予定時刻や停留所名等が含まれている．実験対象期間が 2022 年 11 月 30 日〜12 月 31 日だったため使用するデータ時点の時刻表を収集した．

また，正解ラベルとして使用する渋滞データは Google Maps Platform の Directions Advanced API [11] に 5 分おきにリクエストを送り収集した．Directions Advanced API は，地図上の出発地点と到着地点のそれぞれを緯度経度により指定すると，2 点間の道のりと現在の交通状況を踏まえた所要時間を含むデータがレスポンスとして返される．そのため，出発地点に区間始点停留所，到着地点に区間終点停留所の緯度経度を指定し，リクエストを行うことで停留所間の道のりと所要時間を取得した．取得した道のりを所要時間で割ることで停留所区間の交通速度を算出し，日本の交通渋滞の定義 [5] である時速 10km 以下であれば渋滞とし，それ以外は非渋滞とした．収集期間は 2022 年 11 月 30 日〜12 月 13 日，12 月 18 日〜12 月 31 日である．

3.2 特徴量と特徴量抽出方法

表 2 特徴量抽出に用いる記号

記号	定義
b_i	i 番目のバス
p_j	j 番目の停留所
s_j	停留所 p_{j-1} から停留所 p_j の j 番目の区間
l_j	区間 s_j の距離
td_{ij}	バス b_i の停留所 p_j の出発時刻
tt_{ij}	バス b_i の停留所 p_j の定刻

表 3 特 徴 量

入力データ	定義
v_{ij}	停留所区間 s_j 走行時のバス b_i の速度
c_{ij}	バス b_i の区間 s_j 走行時の時間帯
g_{ij}	バス b_i の停留所 p_j（区間終点）における実際の出発時刻と定刻との差
$g_{i(j-1)}$	バス b_i の停留所 p_j（区間始点）における実際の出発時刻と定刻との差

特徴量と特徴量の抽出方法について説明する．入力データ抽出に用いる記号と入力データをそれぞれ表 2，表 3 に示す．また，入力データ抽出に用いる記号の説明を図 2 に示す．

日本における渋滞の定義は時速 10km 以下で継続的に走行する状態 [5] である．しかし，バスは乗客の乗降による停車を繰り返すため一般車両の交通速度とは異なる．よって，バスの走行速度の他に特徴量を追加することで渋滞検知の精度向上を目指す．特徴量はバス b_i が停留所区間 s_j を走行する際の速度である v_{ij}，バス b_i が停留所区間 s_j を走行する際の時間帯である c_{ij}，バス b_i の停留所 p_j つまり区間終点停留所における実際の出発時刻 td_{ij} と定刻 tt_{ij} の差である g_{ij}，バス b_i の停留所 p_j つまり区間始点停留所における実際の出発時刻 td_{ij} と定刻 tt_{ij} の差である $g_{i(j-1)}$ の合計 4 つである．ただし，バス b_i は出発順，停留所 p_j は経路順に並んでいるとする．時間帯 c_{ij} は通勤時など特定の時間帯よって渋滞が起こるため特徴量として有用であると考えた．また，収集したバスのリアルタイム運行データには停留所に到着した時刻が含まれていないため，バスが区間を走行するために要した時間は停留所の出発時刻から求める必要がある．この場合，バスが定刻より早く走行しているため定刻に合わせようと停留所で時間調整を行う際に，走行時間に大幅なノイズが入り，算出された速度が実際の走行速度より遅く渋滞であると判定される可能性がある．他方で，時刻表に合わせるために時間調整していたため実際の出発時刻と定刻との差は小さいと考えらる．したがって，算出された速度が遅くても実際の出発時刻と定刻との差が小さい場合は，時間調整のための待ち時間などの渋滞以外の要因によりノイズが入っている判断でき，渋滞を区別できると考えられる．そのため停留所における実際の時刻と定刻との差である g_{ij}，$g_{i(j-1)}$ が有効であると考えた．

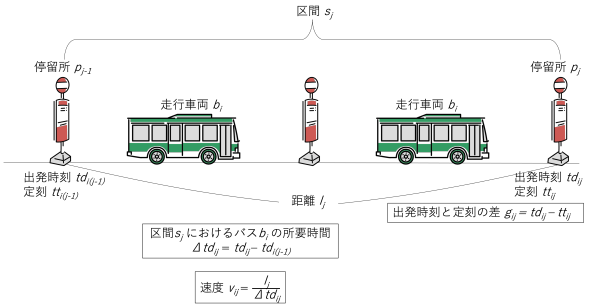


図 2 特徴量抽出に用いる記号の説明

続いて，入力データの抽出方法について説明する．バス b_i の平均速度である v_{ij} は停留所区間 s_j の距離 l_j，区間終点停留所での実際の出発時刻 td_{ij}，区間始発停留所での実際の出発時刻である $dt_{i(j-1)}$ から次式で表される．

$$v_{ij} = \frac{l_j}{dt_{ij} - dt_{i(j-1)}} \tag{1}$$

ただし，v_{ij} どちらも停留所の出発時刻から算出しているため停留所における時間調整によるノイズが含まれる可能性がある．

バス b_i の区間 s_j 走行時の時間帯である c_{ij} は 0 時 0 分 0 秒から 0 時 19 分 59 秒を 0 として，以後 20 分おきに 1, 2, 3, … と定義する．

バス b_i の実際の出発時刻 td_{ij} と定刻 tt_{ij} の差である g_{ij} は次式で表される．

$$g_{ij} = dt_{ij} - tt_{ij} \tag{2}$$

以上の計算により抽出された，$v_{ij}, c_{ij}, g_{ij}, g_{i(j-1)}$ の合計 4 つのデータを特徴量として CNN に用いた．

4 実験と評価

実験対象期間は正解データの収集期間に合わせ 2022 年 11 月 30 日〜12 月 13 日，12 月 18 日〜12 月 31 日とした．実験対象系統は東京都の大塚〜錦糸町を走る系統である．データ数は全部で 61792 個であり，そのうち渋滞は 1675 個で非渋滞は 60117 個であった．作成したデータセットを 8:2 の割合で訓練データとテストデータに分割し実験を行った．モデルは Keras の Sequential モデルを用い，1 次元の CNN を行いました．評価は accuracy，precision，recall，f1-score を用いた．

4.1 評　　価

表 4 実験結果

accuracy	precision	recall	f1-score
0.975	0.680	0.104	0.180

表 4 に実験結果を示す．accuracy が 0.975，precision が 0.680，recall が 0.104，f1-score が 0.180 となった．f1-score は precision と recall の調和平均であるため，recall が低いため f1-score も低い値となった．これはデータが不均衡であるためであると考えられる．

5 まとめと今後の課題

本論文では都営バスのリアルタイム運行からバスの速度, バスの走行時の時間帯, 区間終点停留所における実際の出発時刻と定刻との差，区間始点停留所における実際の出発時刻と定刻との差の 4 つを抽出し, CNN に入力し渋滞か非渋滞かを判断するモデルを提案し, 評価した．f1-score は 0.180 となった．これはデータの偏りにより recall が低くなったためと考えられる．よって，今後はデータを増やす手法を行い精度の向上を目指す．

文　　献

[1] 国土交通省道路局, "平成 18 年度 道路行政の達成度報告書," 2006. https://www.mlit.go.jp/road/ir/ir-perform/h19/all.pdf
[2] 青柳宏紀, 岡田一洸, 山名早人, "都バスのリアルタイム運行データを用いた渋滞検知," DEIM2022 第 14 回データ工学と情報マネジメントに関するフォーラム, 2022, pp. 1-8.
[3] Y. Xu, Y. Wu, J. Xu and L. Sun, "Efficient Detection Scheme for Urban Traffic Congestion Using Buses," 2012 26th Int. Conf. on Advanced Information Networking and Applications Workshops, Fukuoka, Japan, 2012, pp. 287-293.
[4] C. Wang and H. -M. Tsai, "Detecting urban traffic congestion with single vehicle," 2013 Int. Conf. on Connected Vehicles and Expo (ICCVE), Las Vegas, NV, USA, 2013, pp. 233-240.
[5] 国家公安委員会, "国家公安委員会告示第 12 号," 2004. https://www.npa.go.jp/bureau/traffic/seibi2/shinsei-todokede/jouhou/guideline.pdf
[6] C. Samal, F. Sun and A. Dubey, "SpeedPro: A Predictive Multi-Model Approach for Urban Traffic Speed Estimation," 2017 IEEE Int. Conf. on Smart Computing (SMARTCOMP), Hong Kong, China, 2017, pp. 1-6.
[7] T. Kyaw, N. N. Oo and W. Zaw, "Estimating Travel Speed of Yangon Road Network Using GPS Data and Machine Learning Techniques," 2018 15th Int. Conf. on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 2018, pp. 102-105.
[8] Y. Gu, Y. Wang, and S. Dong, "Public Traffic Congestion Estimation Using an Artificial Neural Network," ISPRS International Journal of Geo-Information, vol. 9, no. 3, p. 152, Mar. 2020.
[9] "Bus location information of Bureau of Transportation, Tokyo Metropolitan Government," https://ckan.odpt.org/dataset/b_bus_l toei.
[10] "Bus timetable of Bureau of Transportation, Tokyo Metropolitan Government," https://ckan.odpt.org/dataset/b_bus_timetab toei.
[11] "Directions API," https://developers.google.com/maps/documentation/d

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 13:30 ~ 15:40 | T1-A オンライン（Zoom Events)

説明可能AI・データ解釈 (2)

座長:丸橋 弘治(富士通株式会社)
コメンテータ:佐々木 佑樹(富士通株式会社)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-A-5-01] マルチエージェント強化学習に基づく解釈可能なソーシャルネットワークの将来予測

*三宅 健太郎[1]、伊藤 寛祥[1]、Christos Faloutsos[2]、松本 寛智[1]、森嶋 厚行[1] (1. 筑波大学、2. Carnegie Mellon University)

14:00 ~ 14:25
[T1-A-5-02] 健康行動を促す介入メッセージのパーソナライゼーションに向けた文の特徴量分析

*郷原 聖士[1]、瀧本 祥章[2]、金 秀明[2]、小林 のぞみ[2]、倉島 健[2]、渡辺 太郎[1] (1. 奈良先端科学技術大学院大学、2. NTT人間情報研究所)

14:25 ~ 14:50
[T1-A-5-03] Approximate Inverse Model Explanations (AIME)を用いた内閣支持率変化要因抽出手法

*津布久 尚貴[1]、岡田 龍太郎[2]、峰松 彩子[2]、中西 崇文[2] (1. 武蔵野大学データサイエンス学部データサイエンス学科TransMedia Tech Lab、2. 武蔵野大学データサイエンス学部)

14:50 ~ 15:15
[T1-A-5-04] 2次元最適相関ルールを用いたSHAPの解釈補完

*塩谷 泰平[1]、森本 康彦[1]、亀井 清華[1] (1. 広島大学情報科学部情報科学科)

15:15 ~ 15:40
[T1-A-5-05] インドのITサービス会社のShapley値を用いた15年間の分析

*白田 由香利[1]、Sreekanth Bonthala[2]、Chakraborty Basabi[2,3] (1. 学習院大学経済学部、2. マダナパレ工科大学、3. 岩手県立大学)

DEIM Forum 2024 T1-A-5-01

マルチエージェント強化学習に基づく 解釈可能なソーシャルネットワークの将来予測

三宅健太郎† 伊藤 寛祥† Christos Faloutsos†† 松本 寛智† 森嶋 厚行†

† 筑波大学
†† Carnegie Mellon University
E-mail: †s2113591@klis.tsukuba.ac.jp, ††ito@slis.tsukuba.ac.jp, †††christos@cs.cmu.edu,
††††s2011456@s.tsukuba.ac.jp, †††††morishima-office@ml.cc.tsukuba.ac.jp

あらまし SNS や研究者の協力関係などを表すソーシャルネットワークにおける将来予測は重要である．しかし，既存手法では予測のために多くの潜在特徴量を用いる必要があり，ブラックボックス性が高く，予測結果の解釈が困難である．本研究では，各人のネットワークを変化させる際の方策をネットワーク科学で一般的に利用される用語でモデル化し，強化学習に基づいてソーシャルネットワークの将来予測を行う手法を提案する．より具体的には，ネットワーク科学の研究分野で知られている「ホモフィリー」や「推移性」などの性質を用いて，各人のネットワークの状態の好ましさを測る報酬関数やネットワークを変化させる際の方策関数を設計することでモデルパラメータの解釈性を向上する．生成データの実験により，提案手法はコミュニティの分裂や形成や属性の変化など、ソーシャルネットワーク上で起こると考えられるパターンを生成できることを示した．実データの実験では共著者ネットワークなどのコミュニティグラフの予測において，提案手法は既存手法と同等以上の精度の予測となることを示し，学習されたパラメータからネットワークの性質を説明することで結果に対する解釈性を与えられることを示した．

キーワード ネットワーク科学，時系列グラフ，マルチエージェントシステム，強化学習

1 はじめに

ソーシャルネットワークとは人と人が何らかの関係を元に接続されているグラフデータのことである．例えば，X(旧 Twitter) におけるフォロー/フォロワーの関係や Facebook の友達関係，研究者間の共著関係などがその具体例である．多くの場合，ソーシャルネットワークは人をノード，関係をエッジとするグラフ構造で表現され，各ノードは自身の特性を表現する属性値をもつ．. たとえば，X(旧 Twitter）などのソーシャルネットワークでは投稿データに含まれる単語やハッシュタグなどが属性値として用いることができ，研究者間の共著関係を表すグラフでは論文中に含まれる単語やキーワードなどが用いることができる．ソーシャルネットワークでは多くの場合，エッジやノードがもつ属性値が時間の変化と共に変化する．例えば，時間の経過によって各ノードの興味関心や意見が変化し，それによって属性値が変化しエッジが出現・消滅したりする．本論では，このような動的に変化するソーシャルネットワークの変化を予測することを目的とする．

動的に変化するソーシャルネットワークを予測する手法は近年活発に研究が行なわれている．これまでの研究では [7], [17], [20], [38] のように，興味関心の伝搬などを考慮した潜在的な特徴量に基づいたモデルやグラフニューラルネットワークを用いて，ソーシャルネットワークにおける将来の人々の繋がりがどのように変化するかを予測する研究などが行なわれてきた．しかし，既存の手法には予測に多くの潜在特徴量を用いる必要がありブラックボックス性が高く，予測結果の解釈が困難である．また，推移性など既存のネットワーク研究で検討されてきた理論を十分考慮されてないという問題点がある．

このような背景の下，本研究では解釈可能な特徴を持つマルチエージェント強化学習を用いたソーシャルネットワークの未来を予測するための新しい手法 NetEvolve を提案する．提案手法では，時間変化するエッジと属性値の予測を行う．提案手法では，各ノードが強化学習設定におけるエージェントを表し，

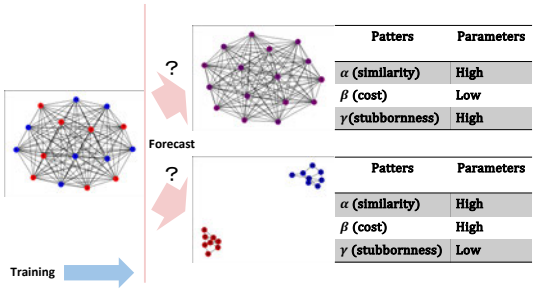


図 1: **NetEvolve は解釈可能性を持ちながらソーシャルネットワークの将来予測を行う手法である．**提案手法ではネットワーク科学で知られている homophily, heterophily, homogenization, や polarization [23], [29], [30], [36] 等の性質を用いて現実のソーシャルネットワークにおける現象を再現するモデルを構築し，未観測の時系列ネットワークを予測する．

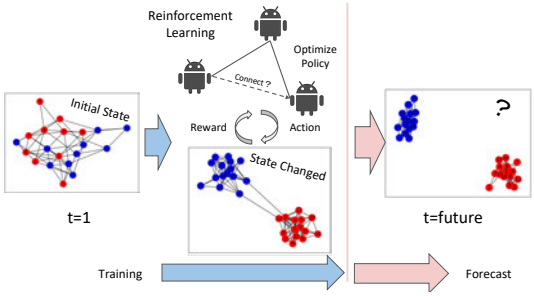


図 2: **NetEvolve の概要** 提案手法では，報酬関数と方策関数を定義し観測済み時系列ネットワークから報酬関数のパラメータを学習する．次にマルチエージェント強化学習を用いて方策関数の学習を行う．学習済みの方策関数によるマルチエージェントシミュレーションにより，未観測時間における時系列ソーシャルネットワークを予測できる．

各ノードは自身の報酬を最大にするような合理的な行動を常に取り続けるものとする．報酬関数と方策関数については，ネットワーク科学の知識に基づいて解釈可能な特徴を設計した．提案手法における学術的問いは次の 2 つである．

(RQ1) ソーシャルネットワークの現象を再現できる説明可能な報酬関数と方策関数の設計とマルチエージェント強化学習を用いた予測モデルが構築できるか？

(RQ2) 提案手法が実世界のデータを予測に対してどの程度うまく機能するのか？

RQ1 について，マルチエージェント強化学習の環境で動的なソーシャルネットワークの予測モデルを定式化した．ここで各ノードの行動はエッジの作成/削除とその属性の変更であると仮定する．そして，提案手法は以下の 3 つのステップで方策関数と報酬関数の学習を行う (図 2).

Step1. 過去の時系列ネットワークデータを用いて報酬の計算を行い，各ノードのパラメータを勾配降下法で学習する

Step2. 各ノードの報酬の期待値を最大化するような方策関数を方策勾配法によって学習する

私たちはネットワーク科学において知られる性質 [30] [10] に基づいて方策関数と報酬関数の設計を行なった．報酬関数は隣接ノードとの属性値の類似度 [30] とエッジを張ることに対するストレスを表したコスト関数 [10] の線形結合で定義した．方策関数はノード毎にエッジ生成/削除や属性値の変化などの行動確率を計算し，より高い報酬を獲得するための様々なストーリーを生成できるようにパラメータを設計した．

提案手法の優位性は以下である．

（1） **Interpretable**: ネットワーク科学と心理学に基づいたモデルを構築することで高い解釈性を持つ

（2） **Extensible**: 他のネットワーク科学や心理学の既知の現象を取り入れる際に容易に拡張できる

（3） **Effective**: 実データを用いた実験の結果，提案手法は観測されていないネットワークにおけるエッジや属性値の予測において，既存の手法を 8%の精度で上回ることが示された

2 関 連 研 究

ソーシャルネットワークに対する予測手法は様々なアプローチで提案されてきた．これまでネットワーク構造を表現する潜在ベクトルを学習する手法，確率モデルに基づくリンク生成モデルによる予測，ノードが属するコミュニティにおけるエッジ生成とコミュニティ間のエッジ生成をモデル化した手法，グラフニューラルネットワークによる手法などが提案されている．本節では，ソーシャルネットワークの予測・生成に関する関連研究，および既存手法との関係について述べる．

2.1 ソーシャルネットワークの表現学習

ネットワーク表現学習とは，ノードまたはグラフ全体をよく表現する潜在ベクトルを学習する手法である．代表的な手法として，確率モデルに基づく潜在ベクトルモデル，ネットワークエンベディング，グラフニューラルネットワークなどの手法が提案されている．確率モデルに基づくネットワーク表現学習の手法として，潜在ベクトルに基づくリンク生成モデル [31]，ノードが属するコミュニティにおけるエッジ生成，コミュニティ間におけるエッジ生成をモデル化した [1] が提案されている．ネットワークエンベディングとはネットワークやノードが持つ属性値をベクトルで表現し，最適な埋め込み表現を学習する手法である．代表的な手法として [14], [35], [37], [40], [45] が提案されている．より近年では，グラフニューラルネットワークを用いた表現学習手法として [25], [26], [39] を用いた手法などが提案されている．また，ノードの変化と属性値の変化の両方を時系列的に埋め込む手法が提案されている．Twitter におけるユーザーの関心の変換を追跡する手法として [27]，テキスト情報を持つ動的なネットワークにおけるノードと単語の埋め込み手法である [45] などがある．これらの手法は，観測されたネットワークにおけるグラフとノードの特徴を取得するために開発されたものであり，観測されていない時間におけるネットワークの予測については検討されていない．

2.2 マルチエージェント強化学習とネットワーク予測

マルチエージェント強化学習は，複数のエージェントが相互に影響を及ぼし合いながらエージェント自身，あるいは集団の充足度を最大化させるような行動を学習するフレームワークのひとつである．マルチエージェント強化学習には完全協調タスク，対戦型タスクなどが存在する [6], [22]．先行研究 [8] ではマルコフゲームと呼ばれる環境におけるエージェントの最適行動のを学習する手法を提案した．[8] ではマルコフゲームと呼ばれる環境でのエージェントの最適な振る舞いを学習する研究が行なわれている．また [43] では社会が個人に与える利益であるソーシャルキャピタルをゲーム・ベースのフレームワークで定義し，マルチエージェント強化学習の枠組みで新たな社会的ネットワーク構造の出現を予測する研究が行なわれている．最近の研究では，表現学習，関係推論，リンク予測など，いくつかのグラフマイニングのタスクに対して強化学習を用いることが検討されている [28], [32]．これらの研究は，与えられた状況

におけるエージェントの最適な行動の学習と，それに伴うネットワーク構造の予測することを試みているが，観測されたネットワークから未観測の時系列ネットワークを生成する研究は十分に行なわれていない．

2.3 ソーシャルネットワークの予測

ネットワーク構造から新規に生成されるノードの特徴量を予測する手法を提案している GCN [25] やその改良として学習時に隣接ノードからサンプリングすることで精度と速度を向上させた．GraphSAGE [18] などの手法が提案されている．ノードの潜在変数の変化をネットワーク構造における伝搬を考慮した上で予測する手法として LFP [20] が存在する．ノード集合の意見の群衆としての変化を予測する手法として CoNN [15]，ノードの潜在変数の変化とコミュニティ構造の変化を抽出し，ネットワーク構造の補完を行う研究として ELSM [16] がある．リンク予測に対して構造情報と時間の情報を取り入れたネットワーク構造の予測手法として STEP [7]， VGAE [26]， VGRNN [17] が存在する．ネットワーク構造から，ノードの特徴の変化を予測する手法として ONE-M [33]，TensorCast [3] などがある．これらの研究ではグラフ構造やノードの属性の予測については考慮されているが，グラフ構造と属性の変化の間の相互作用に関しては十分な検討が行われていない．これらの研究に対して，提案手法はグラフ構造のとノードの属性値の予測の両方を同時に行うことで相互作用を十分に考慮した予測を行うことができる．

DualCast [24] は，ネットワーク構造とノードの属性値を予測する手法であり，既存の手法と比較して高い精度を示した．また，データマイニングによる意見集合の分類と心理学で知られている性質と組み合わせることで，ノードの属性値を予測する研究手法 [34] が提案されている．これはフリードキン - ジョンソンモデル [11] と呼ばれる個人の意見の他人からの影響度や確証バイアスなどの性質を数理モデルとして表現することで予測精度を向上させた．この研究では心理学などの性質を数理モデルに記述し，それに基づくモデルを構築するのことの有用性を示しており，提案手法と類似するアプローチである．しかし，未来の複数の時系列におけるエッジの生成や特徴量の予測については検討が行なわれていない．表 1 に既存手法と比較した提案手法の特徴をまとめる．提案手法では，ネットワーク科学や心理学で知られている性質をモデル化し，ソーシャルネットワークのノードは満足度を高めるために最適な行動をとると仮定する．

3 問題設定

本節では，本研究で対象とする問題について定義する．ここで，各ノード n は人，属性値 $\mathbf{x}_n \in \mathbb{R}^k$ は単語など各人の興味を表すベクトル，エッジは人と人とのつながりを表す．簡単化のため，ノードの出現と消滅は考慮しないこととする．

これらを踏まえて，本研究の問題設定を次のように設定する．時系列ソーシャルネットワーク $\mathcal{G}^{(t)} = \langle \mathcal{V}^{(t)}, \mathcal{E}^{(t)}, \mathcal{X}^{(t)} \rangle, t = 1 \sim T$. ただし $\mathcal{V}^{(t)}$ はノードの集合, $\mathcal{E}^{(t)} = \{e_{i,j}^{(t)}\}$ はノード間の無向エッジ, $\mathcal{X}^{(t)} = \{\boldsymbol{x}_i\}$ は各ノードの特徴量ベクトルが与えられた時に未観測の時系列のネットワークグラフのシーケンスを表すエッジの集合 $\mathcal{E}^{(t')}$ とノードの属性値 $\mathcal{X}^{(t')}, t' > T$ を生成する．

PropertyMethod	*LFP [20]*	*TensorCast [3]*	*Dyngraph2vec [13]*	*DualCast [24]*	*SINN [34]*	*NetEvolve*
Network-forecast	✔		✔	✔		✔
Attribute-forecast		✔		✔	✔	✔
Extensible					✔	✔
Multi-task		✔		✔	✔	✔
Interpretable						✔

表 1: **NetEvolve 提案手法と既存手法との比較**， 提案手法はエッジ生成の予測とモデルの拡張性に特に優れている．

3.1 記号の定義

提案手法で使用した記号について表 2 に示す．

表 2: 記号の定義

$\mathcal{G}^{(t)}$	時刻 t におけるグラフ.
$\mathcal{V}^{(t)} = \{n_i\}$	時刻 t におけるノードの集合.
$\mathcal{E}^{(t)} = \{e_{i,j}^{(t)}\}$	時刻 t におけるエッジの集合.
$\mathcal{X}^{(t)} = \{\boldsymbol{x}_i\}$	時刻 t における n_i の属性値の集合.
$e_{i,j}^{(t)} = \langle n_i, n_j \rangle$	ノード n_i から n_j に向けた有向グラフのエッジ.
$\boldsymbol{x}_i \in \mathbb{R}^k$	ノード n_i の属性値ベクトル.
$S^{(t)}$	時刻 t における状態集合.
$A^{(t)}$	時刻 t における行動集合.
$r(S^{(t)} \mid \Psi)$	パラメータ集合 Ψ を持つ報酬関数.
$\pi(A^{(t)} \mid S^{(t)}, \Theta)$	パラメータ Θ を持つ方策関数.
$\Psi = \{\psi_i\}_{n_i \in \mathcal{V}^{(t)}}$	報酬関数のパラメータ集合.
$\Theta = \{\theta_i\}_{n_i \in \mathcal{V}^{(t)}}$	方策関数のパラメータ集合.
$\psi_i = \{\alpha_i, \beta_i, \gamma_i\}$	ノード n_i が持つ報酬関数の集合.
$\theta_i = \{\epsilon_i, \tau_i, \lambda_i, \mu_i\}$	ノード n_i が持つ方策関数の集合.
$\alpha_i \in \mathbb{R}^+$	**Similarity** に関する報酬関数の重み.
$\beta_i \in \mathbb{R}^+$	**Cost** に関する報酬関数の重み.
$\gamma_i \in \mathbb{R}^+$	**Impact** に関する報酬関数の重み.
$\epsilon_i \in \mathbb{R}^+$	**Intensity**: エッジに関する方策のパラメータ.
$\tau_i \in \mathbb{R}^+$	**Temperature**: エッジに関する方策のパラメータ.
$\lambda_i \in \mathbb{R}^+$	**Stubbornness**: 属性値に関する方策のパラメータ.
$\mu_i \in \mathbb{R}^+$	**Influence**: 属性値に関する方策のパラメータ.
$\eta \in \mathbb{R}^+$	方策勾配法における学習率.
$\xi \in \mathbb{R}^+$	報酬の計算における割引率.
$N \in \mathbb{N}^+$	方策勾配法におけるシミュレーション回数.
$T' \in \mathbb{N}^+$	方策勾配法における生成シーケンスの数. height

4 提案手法

本節では，提案手法であるマルチエージェント強化学習に基づくソーシャルネットワークの将来予測モデルについて述べる．本モデルでのパラメータは，ネットワーク科学の知見に基づいて解釈可能であり，ネットワーク科学に関する他の知識を取り込んで容易に拡張可能である．

提案手法は次の 2 つの段階によってモデルの学習を行う．(1) 過去の時系列ソーシャルネットワークデータを用いて，各ノードの報酬関数と方策関数を最適化する．(2) マルチエージェントシミュレーションに基づいて，未観測の時系列ネットワークデータを生成する．

提案手法では，入力されたソーシャルネットワークの時系列から，各ノードがもつ報酬関数のパラメータを学習し，推定された報酬関数の値が最大化されるように方策勾配法によりエッジと属性値に関する方策関数の最適化を行なう．これらの結果をもとに学習済の報酬関数と方策関数の学習済みパラメータを用いてノード間のエッジの生成および消滅確率を計算と属性値の変化の計算を同時に行うことでネットワークの将来予測を行なう．

4.1 強化学習の環境

4.1.1 前提知識: マルコフ決定過程 (MDP)

提案手法では環境をマルコフ決定過程を用いて定義し，状態空間 $S^{(t)}$，行動空間 $A^{(t)}$， and 報酬関数 $R^{(t)}$ とする．エージェント集合 $\mathcal{V}^{(t)}$ について，各エージェントは状態に応じて方策関数 $\pi(A^{(t)} \mid S^{(t)}, \Theta)$ により行動を選択する．そして，状態遷移関数 $f(S^{(t+1)} \mid S^{(t)}, A^{(t)})$ によってエージェントの状態が変化し，報酬関数 $r(S^{(t)} \mid \Psi)$ によって報酬が計算される．ここで Θ と Ψ はそれぞれ方策関数と報酬関数のパラメータである．強化学習の目的は，期待報酬を最大化する方策関数のパラメータを学習することである．

4.1.2 MDP のソーシャルネットワークへの適応

提案手法では，強化学習を利用してソーシャルネットワークを予測するために，状態空間と行動空間を次のように定義する．

- **状態** $S^{(t)}$ は現在のグラフ $\mathcal{G}^{(t)}$ における状態である
- **行動** $A^{(t)}$ は次の時刻の系列との差分 $\Delta\mathcal{G}^{(t)}$ である

ここで，$\Delta\mathcal{G}^{(t)} = \langle \Delta\mathcal{E}^{(t)}, \Delta\mathcal{X}^{(t)} \rangle$ において $\Delta\mathcal{E}^{(t)}$ は新しく追加／削除されたエッジの集合，$\Delta\mathcal{X}^{(t)}$ はノードの属性の変化である．上記の式を用いて，報酬，行動，状態遷移を以下のように記述できる．

- **報酬関数** $r(S^{(t)} \mid \Psi) = r(\mathcal{G}^{(t)} \mid \Psi)$
- **方策関数** $\pi(A^{(t)} \mid S^{(t)}, \Theta) = \pi(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \Theta)$
- **状態遷移関数** $f(S^{(t+1)} \mid S^{(t)}, A^{(t)}) = \langle \mathcal{V}^{(t)}, \mathcal{E}^{(t)} \cup \Delta\mathcal{E}^{(t)}, \mathcal{X}^{(t)} \cup \Delta\mathcal{X}^{(t)} \rangle$

提案手法では方策関数 $\pi(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \Theta)$ と状態遷移関数 $\langle \mathcal{V}^{(t)}, \mathcal{E}^{(t)} \cup \Delta\mathcal{E}^{(t)}, \mathcal{X}^{(t)} \cup \Delta\mathcal{X}^{(t)} \rangle$ を用いて将来のソーシャルネットワークを予測する．

4.1.3 ソーシャルネットワークのためのマルチエージェント強化学習

各ノードはエージェントであり，パラメータ集合 $\psi_i \in \Psi$ をもつ報酬関数 $r_i(\mathcal{G}^{(t)} \mid \psi_i)$ と，パラメータ集合 $\theta_i \in \Theta$ をもつ方策関数 $\pi_i(\Delta\mathcal{G} \mid \mathcal{G}^{(t)}, \theta_i)$ によるマルチエージェント環境として定義する．報酬関数は各ノードのソーシャルネットワークの好ましさを表現し，方策関数はソーシャルネットワークにおける各ノードのエッジと属性の変化の傾向を表現する．

- **報酬関数** $r(\mathcal{G}^{(t)} \mid \Psi) = \sum_{n_i \in \mathcal{V}} r_i(\mathcal{G}^{(t)} \mid \psi_i)$
- **方策関数** $\pi(\Delta\mathcal{G}^{(t)} \mid \Theta) = \prod_{n_i \in \mathcal{V}} \pi_i(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \theta_i)$

提案手法では，時系列ソーシャルネットワークの $\langle \mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(T)} \rangle$ を入力とし，報酬関数のパラメータ $\Psi = \{\psi_i\}_{n_i \in \mathcal{V}}$ と方策関数のパラメータ $\Theta = \{\theta_i\}_{n_i \in \mathcal{V}}$ を学習するための最適化手法を設計する．

4.1.4 方策関数による将来予測

方策関数を学習することにより，未観測時刻におけるソーシャルネットワークの時系列 $\langle \mathcal{G}^{(T+1)}, \mathcal{G}^{(T+2)}, \ldots, \mathcal{G}^{(T+T')} \rangle$ を予測できる．MDP に基づくと，将来の時系列ソーシャルネットワークを生成される確率は式 1 で表される．

$$Pr(\langle \mathcal{G}^{(T+1)}, \ldots, \mathcal{G}^{(T+T')} \rangle) = \prod_{t=1}^{T'} \pi(\Delta\mathcal{G}^{(T+t)} \mid \mathcal{G}^{(T+t)}, \Theta) \qquad (1)$$

4.2 報酬関数

本節では入力となったソーシャルネットワークの時系列の変化から，各ノードの報酬関数のパラメータを推定する方法について述べる．報酬関数は各ノードのソーシャルネットワークの好ましさを表現し，方策関数は現在のソーシャルネットワークの状態における各ノードのエッジや属性値の変化の傾向を表現する．提案手法では，ノードがネットワークから得る報酬は隣接ノード (Fig. 3) 間の関係に基づいて計算されると仮定する．時刻 t における各ノードの報酬関数を以下の式で定義する．

$$r_i(\mathcal{G}^{(t)} \mid \psi_i) = \sum_{n_j \in N(n_i)} \alpha_i \, sim(n_i, n_j) - \beta_i \, cost(n_i, n_j) + \gamma_i \, impact(n_j) \qquad (2)$$

ただし，$\psi_i = \{\alpha_i, \beta_i, \gamma_i\}$ は各ノード n_i が持つパラメータ集合，$N(n_i)$ は n_i の隣接ノードの集合とする．

式 2 は，隣接ノードの類似度に基づく報酬，エッジの接続コスト，および重み付けパラメータ α_i，β_i と γ_i による隣接ノードの属性変更の影響の線形結合からなる．報酬関数は，ノードの報酬を計算するための他の項を追加することで容易に拡張できる設計となっている．報酬関数における各パラメータは以下のように解釈される．

- α_i: **Similarity** はネットワークにおけるノード同士の類似性の傾向を表す重み．この値が正の場合，ノードは類似する属性持つノードと接続する傾向が強い．
- β_i: **Cost** は人間関係におけるストレスに対する重みを表す．この値が高いほどあるノードと繋がることへの報酬が低くなる．
- γ_i: **Impact** は隣接ノードの属性値に対する影響力の大きさを表現する．あるノードが隣接ノードの属性により大きな影響を与えるほど報酬は高くなる．

$sim(n_i, n_j)$ はノード間の類似度を示し，提案手法ではノードの属性値集合のコサイン類似度によって算出する．

$$sim(n_i, n_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|} \qquad (3)$$

コスト関数はソーシャルネットワークにおいてノード n_i とノード n_j の間にエッジを持つことのコストを示し以下の式で定義する．

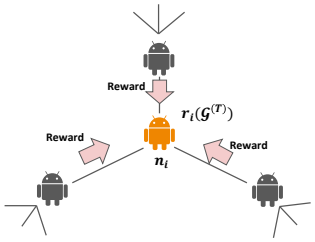


図 3: **報酬関数の計算** あるノードの報酬は隣接ノードとの計算結果の総和によって求める．

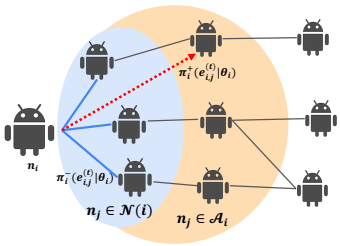


図 4: **エッジに関する方策関数の概要** エッジを生成する方策 $\pi^+(\cdot)$ は，2 ホップ先までの隣接ノードとランダムノードに対してエッジの生成確率を計算する．エッジを削除する方策 $\pi^-(\cdot)$ は，時刻 t において隣接ノードに対してエッジの消滅確率を計算する．

$$cost(n_i, n_j) = \begin{cases} 1, & e_{i,j} \in \mathcal{E}^{(t)} \\ 0, & e_{i,j} \notin \mathcal{E}^{(t)} \end{cases} \tag{4}$$

impact 関数は自分の属性が他者に与える影響を表す．提案手法では以下のように定義した．

$$impact(n_j) = \|\mathbf{x_j}^{(t)} - \mathbf{x_j}^{(t-1)}\|_2^2 \tag{5}$$

sim，*cost*，*impact* については埋め込み表現や GNN [4], [9], [25] を用いてより複雑に表現することもできる．

報酬関数のパラメータは SGD を用いて，観測済みのソーシャルネットワークにおける報酬関数の値が最大になるようにパラメータを推定する．

$$\Psi^* \leftarrow argmax_\Psi \sum_{t=1}^{T} r(\mathcal{G}^{(t)} \mid \Psi) \tag{6}$$

4.3 方策関数

本節では，ノードの方策関数について述べる．各ノードはエッジの作成/削除と属性の変更の行動を取ることができると仮定する．本研究ではエッジと属性の方策は互いに独立しているとする．

$$\pi_i(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \theta_i) = \pi_i(\Delta\mathcal{E}^{(t)} \mid \mathcal{G}^{(t)}, \theta_i) \cdot \pi_i(\Delta\mathcal{X}^{(t)} \mid \mathcal{G}^{(t)}, \theta_i) \tag{7}$$

以下の節では，方策関数の設計とパラメータの学習方法について説明する．

4.3.1 エッジの方策関数

本節では，ノードのエッジを変更するための方策関数の設計

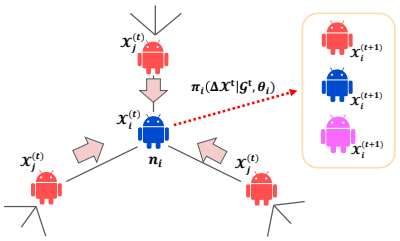


図 5: **属性値に関する方策関数の概要** 図では各ノードの色がそれぞれの属性を表している．各ノードは自分の属性値の変化のしにくさと，隣接ノードの属性に対する影響力の大きさによって自身の属性が変更する．

について述べる．提案手法では，ランダム選択とネットワーク科学の知識に基づく推移性 [41] の考えに基づいてエッジの生成・削除を行なう対象を探索する．さらに，エッジの作成／削除を独立として方策関数を以下のように分解できると仮定する．

$$\pi_i(\Delta\mathcal{E}^{(t)} \mid \mathcal{G}^{(t)}, \theta_i) = \prod_{n_j \in \mathcal{A}_i} \pi_i^+(e_{i,j}^{(t)} \mid \theta_i) \prod_{n_j \in N(n_i)} \pi_i^-(e_{i,j}^{(t)} \mid \theta_i) \tag{8}$$

ここで，$\pi_i^+(e_{i,j} \mid \theta_i)$ と $\pi_i^-(e_{i,j} \mid \theta_i)$ は，それぞれエッジを生成/削除する方策関数である．また，$V_{neighbor}(i) = \bigcup_{n_j \in N(i)} \boldsymbol{N}(i)$ をノード i の 2 ステップ先までの隣接ノードの集合，V_{random} をノード全体の集合からランダムにサンプリングした集合とする．この時，ノードのエッジの生成/削除の計算対象である行動空間は $\mathcal{A}_i = \{n_j \in \mathcal{V} | n_i \in V_{neighbor}(i) \cup V_{random}\}$ で表される．行動空間をこのように設計することで，近接性が高いノードにエッジが生成されやすくなり，推移性を再現することができる．

ネットワーク上でエージェントがエッジを作成/削除する際，様々な行動パターンを生成可能な方策関数を設計した 4．例えば，報酬を大きくするためには，(1) コストを無視してより多くのノードにエッジを生成する戦略，(2) 類似の属性を持つ限られた数のノードにエッジを生成することでコストを抑えつつ期待報酬を高くする戦略などが考えられる．提案手法では，各ノードの属性値に基づいて類似度関数を定義し，それを用いて方策関数を定義する．類似度関数 $sim(n_i, n_j)$ を用いて，i 番目のノードのエッジを生成する方策関数 $\pi^+(\cdot)$ とエッジを削除する方策関数 $\pi^-(\cdot)$ を以下のように定義する．

$$\pi_i^+(e_{i,j}^{(t)} \mid \theta_i) = \tanh\left\{\epsilon_i \exp\left(\frac{sim(n_i, n_j)}{\tau_i}\right)\right\} \tag{9}$$

$$\pi_i^-(e_{i,j}^{(t)} \mid \theta_i) = \tanh\left\{\epsilon_i \exp\left(\frac{1 - sim(n_i, n_j)}{\tau_i}\right)\right\} \tag{10}$$

式 (9)(10) は以下のパラメータを含んでいる:

- ϵ_i: **Intensity** が大きい場合，各ノードは報酬を増やすためにより多くのエッジを作成/削除する傾向になる．
- τ_i: **Temperature** はエッジを作成/削除する際の温度係数で，このパラメータが小さい場合はノードは属性の類似性により依存してエッジを作成/削除する傾向になる．

4.3.2 属性変化の方策関数

次に属性値の変化に関する方策関数について説明する．この

関数は，より高い報酬を獲得するために属性を変更させる際の様々なシナリオを生成できるように設計されている (図 5)．例えば (1) 隣接ノード間の類似性を高めるために隣接ノードと同じになるように自身の属性値を変化する，(2) 隣接ノードを類似させるために隣接ノードの属性値への影響を与えるなどである．このようなストーリーを説明するために，各ノードが持つ **stubbornness** と **influence** に基づいて属性に関する方策をモデル化する．

$$\pi_i(\Delta X^{(t)} \mid \mathcal{G}^{(t)}, \theta_i) = \pi_i(\boldsymbol{x}_i^{(t+1)} \mid \mathcal{G}^{(t)}, \theta_i) = \sigma\left(\lambda_i \cdot \boldsymbol{x}_i^{(t)} + (1-\lambda_i) \sum_{n_j \in \mathcal{N}(n_i)} \mu_j \cdot \boldsymbol{x}_j^{(t)}\right) \quad (11)$$

ここで λ_i,μ_i はノード i のパラメータである．

- λ_i: **Stubbornness** は自身の属性値の変化のしにくさを表す．この値が大きい場合，そのノードは自身の属性を変更させる傾向がないことを意味する．
- μ_i: **Influence** は隣接ノードの属性への影響力を表す．この値が大きいほど，そのノードの属性値は周囲のノードの属性値に対する影響が大きくなり，隣接ノードの属性を変化させやすくすることを意味する．

全体として，ノード n_i の方策関数のパラメータ集合は $\theta_i = \{\epsilon_i, \tau_i, \lambda_i, \mu_i\}$ である．すなわち，ネットワークにおけるノード n_i の振る舞いは，これらのパラメータによって特徴づけられる．

4.3.3 方策関数のパラメータ学習

次に，方策関数のパラメータを学習する方法を説明する．具体的には，勾配降下法の 1 つである REINFORCE [42] を用いて，自身の最適化された報酬関数 4.2 を最大化する方策関数のパラメータを学習する．提案手法では未観測のソーシャルネットワークを生成し，報酬関数の期待値を最大化する方策関数のパラメータを学習する．

ネットワークの各ノードが各時刻において最適な行動を選択していると仮定すると，将来の報酬を最大化する方策を得ることで観測されていない時系列におけるネットワークを予測することが可能である．提案手法では，方策勾配法を用いて方策関数のパラメータ *Theta* を最適化する．方策関数 $\pi(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \Theta)$ に基づいて未観測の時間におけるソーシャルネットワークを生成するために，生成関数 eq. 1 を使用し，生成されたグラフの系列を以下のように記す: $\omega = \langle \mathcal{G}^{(T+1)}, \mathcal{G}^{(T+2)}, \ldots, \mathcal{G}^{(T+T')} \rangle$.

このとき，ネットワーク中のすべてのノードが未観測の時刻において獲得した累報酬積を割引率 ξ を用いて以下のように定義する．

$$R(\omega) = r(\mathcal{G}^{(T+1)}) + \xi \cdot r(\mathcal{G}^{(T+2)}) + \cdots + \xi^{T'} \cdot r(\mathcal{G}^{(T+T')}) \quad (12)$$

本手法では，累積報酬の期待値を最大にする方策関数のパラメータを学習する．このときの目的関数 $J(\theta)$ は以下の式で表現する．

$$J(\Theta) = \mathbb{E}_{\omega \sim \pi_\Theta}[R(\omega)] \quad (13)$$

ここで上記の目的関数を最大化するパラメータ Θ に関する勾配を求める．

$$\nabla_\Theta J(\Theta) = \nabla_\Theta \mathbb{E}_{\omega \sim \pi_\Theta}[R(\omega)] \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T'} R(\omega^{(i)}) \nabla_\Theta \log \pi(\Delta\mathcal{G}^{(t)} \mid \mathcal{G}^{(t)}, \Theta) \quad (14)$$

ここで $\omega^{(i)}$ は方策関数によってサンプリングした i 番目の系列である．これを用いてパラメータを更新する．ここで η は学習率である．

$$\Theta^* \leftarrow \Theta + \eta \cdot \nabla_\Theta J(\Theta) \quad (15)$$

以上の式で方策関数のパラメータの学習を行なう．この方策関数と (1) で学習した報酬関数を用いて未観測のソーシャルネットワークを生成する．

4.4 計算量の分析

NetEvolve における計算量は以下の通りである．

（1） 報酬関数の計算量: $O(T \cdot |\mathcal{E}|)$
（2） 方策関数の計算量: $O(T \cdot N \cdot \frac{|\mathcal{E}|^2}{|\mathcal{V}|})$
（3） ネットワークの生成時の計算量: $O(T' \cdot \frac{|\mathcal{E}|^2}{|\mathcal{V}|})$

全体の計算量は $O\left((TN + T') \cdot \frac{|\mathcal{E}|^2}{|\mathcal{V}|}\right)$. ただし，エッジが疎の場合は $|\mathcal{V}|$ に漸近する．

5 実験 1: モデルの妥当性の検証

RQ1 に答えるために，実験 1 では合成したデータを用いて実験を行い，NetEvolve がソーシャルネットワークで見られる homophily, heterophily, homogenization, や polarization [23], [29], [30], [36] などの現象を再現できるか検証する．

5.1 実 験 設 定

この実験では，報酬関数と方策関数のパラメータを手動で設定し *homophily*, *heterophily*, *homogenization*, と *polarization* の 4 つのネットワーク現象をシミュレーションし，どのような設定で対象とする現象が発生するかを調査する．まず，本研究で対象とする 4 つの現象について述べる．

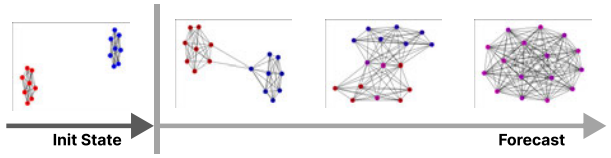
- **Homophily**: ホモフィリーは「類似した人々との間の接触は，類似していない人々間の接触より高確率で発生する」という性質であり，類似する属性を持つノードが高密度に接続するネットワークが生成される現象である [5], [21].
- **Heterophily**: ヘテロフィリーは同じ属性や価値観を持たないが利益があるためにつながろうとする性質である．ヘテロフィリー・ネットワークでは，ネットワークは人々が異なる属性のノードと繋がることによってネットワークが成長する現象である [44].
- **Homogenization**: はノード同士の関係を維持しながら，接続ノードの属性が徐々に均質化する現象である [36].
- **Polarization** は 類似した特性を持つノードがコミュニティを形成し，異なる属性のノードは別のコミュニティに徐々に分離していく現象である [12], [15], [19], [23].

本実験ではノード数は 20，各ノードは「赤」「青」のいずれかの属性値を持つ設定で実験した．

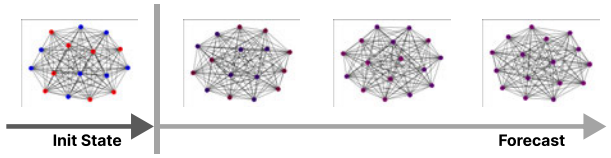
5.2 実験結果：パラメータの値と対応する生成ネットワーク

図 6a から 6d は，提案手法によって生成されたネットワークを示している．表 3 は各現象をシミュレーションするために設定したパラメータをまとめたものである．図 6a では NetEvolve はホモフィリー・ネットワークを生成した．このネットワークでは，パラメータは類似度が高ければ高いほど報酬が高く，接続コストが低ければ低いほど報酬が高くなる．これは，ノードが積極的に隣接ノードとの接続を増やそうとし，同じようなノードからなる 1 つの大きなコミュニティになることを示している．図 6c では，この方法は *heterophily* ネットワークを生成し，そのパラメータは報酬と影響力の低さにおいて *homophily* の場合と似ている．最終的なネットワークは *homophily* ネットワークとは異なり，異なる属性を持つノードからなる 1 つの大きなコミュニティが発生する．図 6b では, *homogenization* を生成している．図 6d では *polarization* のパターンを発生させている．*polarization* ではコミュニティを分離し，同じ属性のノードが異なるコミュニティを形成する．

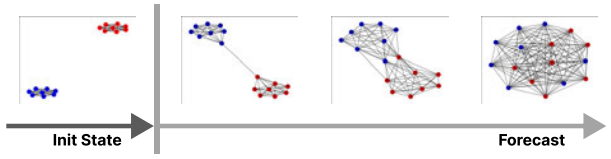
これらの結果は NetEvolve が初期ネットワークが同じ状態であっても，パラメータが異なれば異なるシナリオのソーシャルネットワークを生成できることを示している．



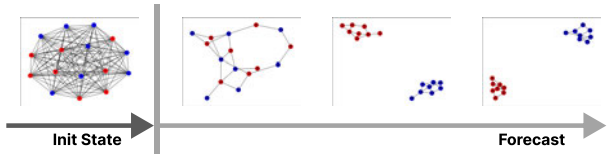
(a) **Homophily** の場合



(b) **Homogenization** の場合



(c) **Heterophily** の場合



(d) **Polarization** の場合

図 6: **実験 1** の結果，様々なパラメータを設定することで，様々なタイプのソーシャルネットワークの現象を再現することができることが示される．各シナリオにおいて，パラメータは表 3 に示すように設定した．

6 実験 2: 実データを用いたネットワーク予測

RQ2 に答えるため，実データを用いて観察されていないソーシャルネットワークのエッジや属性の変化を NetEvolve がどの程度正確に予測できるかを検証した．より具体的には，未観測の時刻のエッジと属性が生成される確率を計算した．本実験では，報酬関数と方策関数のパラメーターを更新しない場合をベースライン手法として用いて，パラメーターを更新した場合の予測モデルと比較することで提案手法の有効性を検証する．また，比較手法として潜在ベクトルの特徴量からエッジと特徴量の予測を行なった DualCast [24]，エッジを生成するグラフニューラルネットワークに基づく手法として提案された VGAE [26]，VGRNN [17] を採用する．属性予測の比較では RNN と DualCast を採用した．本研究では予測精度を検証するために AUC と NLL の指標を用いる，AUC は高ければ高いほど予測精度が高く，NLL が低いほど予測精度が高いことを示す．

6.1 ハイパーパラメータの設定

提案手法では学習データから方策関数のハイパーパラメータを学習することで予測精度を向上した．最適なハイパーパラメーターの探索には Optuna [2] を利用し，学習データに利用した 5 セグメントから最初の 3 セグメントを学習データに利用し，残りの 2 セグメントを検証データに利用した。その結果，`NIPS` の場合は $\eta = 0.00015, \xi = 0.420$, `DBLP` の場合は $\eta = 0.079, \xi = 0.164$, `Twitter` の場合は $\eta = 0.00086, \xi = 0.725$ となった．ただし，全てのデータセットにおいて $|\mathcal{A}_{random}| = 0.001 \times |\mathcal{V}|$, $N = 48, T' = 48$ で最適化した．

6.2 データセット

実験には `DBLP`，`NIPS`，`Twitter` の 3 つのデータセットを使用する．これらのデータセットは，それぞれネットワークのサイズとエッジの密度が異なる．各データセットについて，10 個の時間セグメントを収集し 5 個の時間セグメントを報酬関数の最適化に使用した．

- `NIPS`：NIPS は研究者の共著者ネットワークである．ノードは著者，エッジは共著関係で時間区分は年単位である．ある年に 2 人の著者 x，y が共著論文を発表したとき，その 2 人の間に双方向の有向エッジが張られる．各ノードは属性値としてある年に発表した論文のタイトルに含まれる単語をもつ．ノード数は 32 で属性値数は 2411 である．
- `DBLP`：DBLP は研究者の共著者ネットワークである．ノードは著者，エッジは共著関係，時間区分は年単位である．エッジ情報や特徴量などのネットワークデータは NIPS と同様の方法で生成を行なっている．2008 年から 2017 年に発表されたデータマイニング，データベース，自然言語処理，機械学習，人工知能，情報検索，コンピュータビジョンから構成されるの 47 国際会議の論文データを取得し，2008 年から 2012 年までを学習データ，2013 年から 2017 年までをテストデータとして用いる．ノード数は 500，属性値数は 3854 である．
- `Twitter`：Twitter ユーザーのリツイートネットワークのデータセットである．ノードはユーザーを表す．ツイートの収集期間は 2010 年 1 月から 2010 年 10 月である．時間区分は月単位とする．1 月から 5 月までのデータを学習データとして使

	α (**similarity**)	β (**cost**)	γ (**impact**)	ϵ (**intensity**)	τ (**temperature**)	λ (**stubbornness**)	μ (**influence**)
homophily (Fig.6a)	*Very High*	*Low*	*High*	*High*	*High*	*High*	*Very High*
homogenization (Fig.6b)	*High*	*Low*	*High*	*Low*	*High*	*Low*	*Very High*
heterophily (Fig.6c)	*Very High*	*Low*	*Low*	*High*	*High*	*High*	*Low*
polalization (Fig.6d)	*High*	*High*	*Low*	*High*	*Low*	*High*	*Low*

表 3: 実験 1 におけるパラメータの設定を (*VeryHigh*, *High*, *Low*) で表現する．ただし，"*Very High*" if $x \geqq 1.5$, "*High*" if $1.5 > x \geqq 1.0$, and "*Low*" if $x < 1.0$ である．また，これらのパラメータについて簡単のためにすべてのノードのパラメーターを同じに設定した．

Dataset		average ± std.
DBLP	α (similarity)	1.31 ± 0.40 (*High*)
	β (cost)	0.93 ± 0.02 (*Low*)
	γ (impact)	3.00 ± 0.00 (*Very High*)
NIPS	α (similarity)	1.13 ± 0.12 (*High*)
	β (cost)	0.99 ± 0.00 (*Low*)
	γ (impact)	1.13 ± 0.00 (*High*)
Twitter	α (similarity)	1.03 ± 1.29 (*High*)
	β (cost)	0.99 ± 0.00 (*Low*)
	γ (impact)	41.0 ± 0.00 (*Very High*)

表 4: パラメータの学習結果

用し，6 月から 10 月までのデータをテストデータとして使用した．全期間を通じて 1 回しか登場しないユーザーはデータセットから除外した．あるユーザー i がその月に他のユーザー j のツイートをリツイートすると，i から j への有向エッジが現れる．また，各時間区分のツイートに含まれるハッシュタグを属性値とした．ノード数は 112,044，属性値の数は 5372 である．

6.3 実験結果：予測精度

図 7a から 7c にエッジ予測のタスクにおける AUC の結果を示し，図 7d から 7f に NLL の結果を示す．NetEvolve においてエッジのみを予測する方法と精度を比較すると，DBLP と NIPS ではほぼ同じ精度を達成している．twitter では NetEvolve においてエッジと属性の両方を予測する方法がエッジのみを予測する方法よりも高い．これは NetEvolve のマルチタスク学習が有効であることを示している．図 8a から 8c に属性値予測タスクの AUC の結果を示し， 図 7d から 7f までに NLL の結果を示す．エッジ予測のタスクと同様に NetEvolve はエッジと属性を予測する方法は，属性のみを予測する方法に勝る．

6.4 学習パラメータの分析

報酬関数の最適化されたパラメータを分析することで NetEvolve の解釈可能性を検証する．表 4 に学習したパラメータ *alpha* (similarity), *beta* (cost), *gamma* (impact) の平均と標準偏差を示す．パラメータを学習する前に，各パラメータの初期値を 1.0 に設定した．DBLP, NIPS, Twitter の結果から，α（類似度）と γ（影響度）は，β（コスト）よりもノードの動作に大きな影響を与えることが分かる．また Twitter の結果から， α と β の影響は同じで γ の方がノードの行動に与える影響が大きいことがわかる．

7 結　　論

本研究では，ネットワーク科学や心理学で知られた性質に基づいてマルチエージェント強化学習を用いたソーシャルネットワークの予測手法 NetEvolve を提案した．提案手法では，ネットワーク科学や心理学で知られたホモフィリーなどの性質から報酬関数と方策関数を設計することで，モデルの解釈性を向上することができた．提案手法が達成した点を以下にまとめる．

（1） **Interpretable**: ネットワーク科学と心理学に基づいたモデルを構築することで解釈性の高いモデルを提案

（2） **Extensible**: 他のネットワーク科学や心理学の既知の現象を取り入れる際に容易に拡張できるモデルを提案

（3） **Effective**: 実データを用いた実験の結果，ほとんどの場合で既存の手法を上回る精度で予測できることが示した

NetEvolve により，モデルパラメータからネットワーク科学的で知られた現象と対応付けることで結果に対してより解釈性のある説明を導出することができた．さらに，合成データや実データによる実験により，NetEvolve は様々な社会現象を生成できることを示し，予測についても既存手法と同等かそれ以上の精度で行えることを示した．

謝　　辞

本研究の一部は JSPS 科研費 (JP22H00508, JP22K17944)，JST CREST(JPMJCR22M2) の AIP チャレンジの支援を受けたものである．ここに謝意を示す．

文　　献

[1] E. M. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels. In *Proceedings of Advances in Neural Information Processing Systems (NIPS'08)*, pp. 1981–2014, 2008.

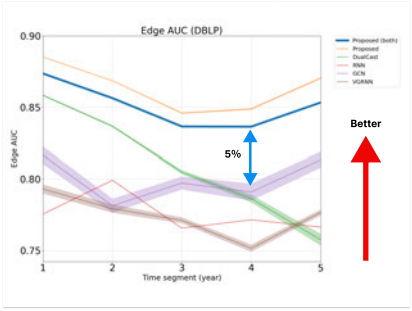
[2] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*, pp. 2623–2631, 2019.

[3] M. Araújo, P. Ribeiro, and C. Faloutsos. Tensorcast: forecasting time-evolving networks with contextual information. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pp. 5199–5203, 2018.

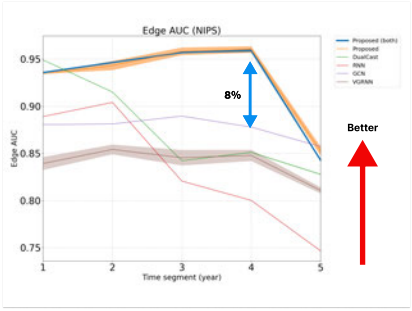
[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[5] Y. Bramoullé, S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers. Homophily and long-run integration in social networks. *Journal of Economic Theory*, 147(5):1754–1786, 2012.

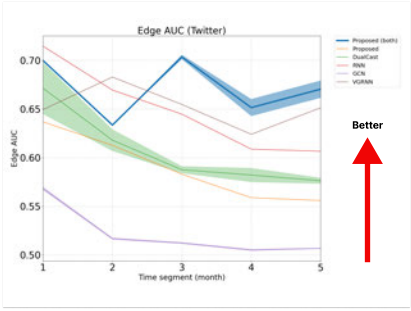
[6] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on*



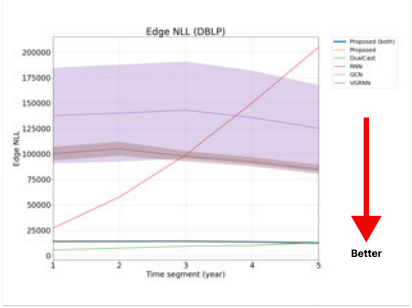
(a) Edge AUC over time with DBLP



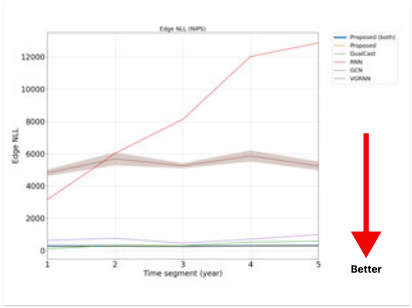
(b) Edge AUC over time with NIPS



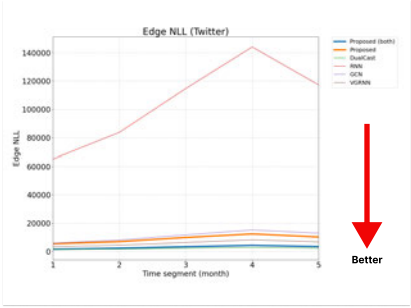
(c) Edge AUC over time with Twitter



(d) Edge NLL over time with DBLP

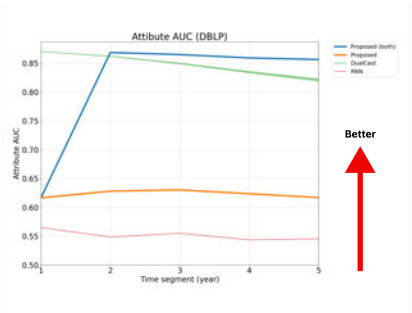


(e) Edge NLL over time with NIPS

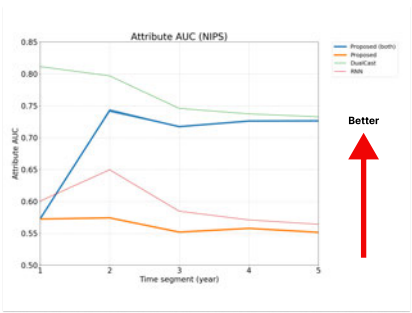


(f) Edge NLL over time with Twitter

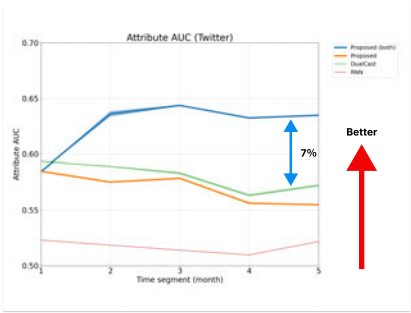
図 7: **エッジの予測精度の結果の比較** これらの図は，将来現れるエッジを予測する際の AUC と NLL を示している．．ここで”Proposed(both)” は NetEvolve であり，”Proposed” は属性値の更新を行なわずエッジの予測だけを行なった結果である. いずれの場合においても多くのケースで既存手法と同等以上の精度となっていることがわかる.



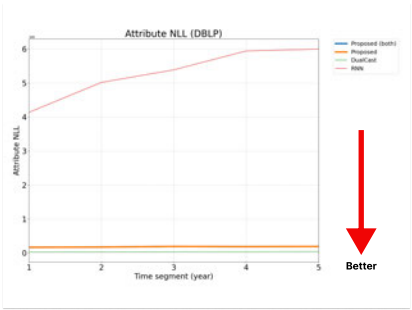
(a) Attribute AUC over time with DBLP



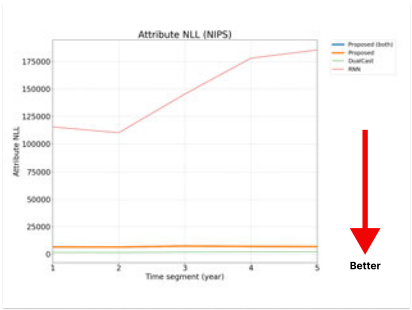
(b) Attribute AUC over time with NIPS



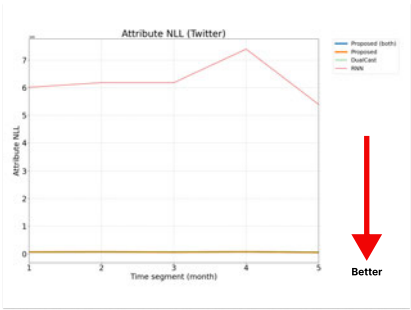
(c) Attribute AUC over time with Twitter



(d) Attribute NLL over time with DBLP



(e) Attribute NLL over time with NIPS



(f) Attribute NLL over time with Twitter

図 8: **属性値の予測精度の比較** これらの図では，将来の属性変化を予測結果について AUC と NLL を示している．ここで”Proposed(both)” は NetEvolve であり， ”Proposed” はエッジの更新を行なわず属性値の予測だけを行なった結果である.NetEvolve はほとんどのケースで既存手法と同等以上の精度が出ていることが分かる.

Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(2):156–172, 2008.

[7] H. Chen and J. Li. Exploiting structural and temporal evolution in dynamic link prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18)*, pp. 427–436, 2018.

[8] Y. Chen, J. Liu, H. Zhao, and H. Su. Social structure emergence: A multi-agent reinforcement learning framework for relationship building. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'20)*, pp. 1807–1809, 2020.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: Patterns and a utility-driven generator. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, p. 269–278, 2009.

[11] N. E.Friedkin and E. C.Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[12] S. M. Gerrish and D. M. Blei. Predicting legislative roll calls from text. In *International Conference on Machine Learning (ICML'11)*, p. 489–496, 2011.

[13] P. Goyal, S. R. Chhetri, and A. Canedo. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. *Knowledge-Based Systems*, 187, 2020.

[14] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pp. 855–864, 2016.

[15] Y. Gu, Y. Sun, and J. Gao. The co-evolution model for social network evolving and opinion migration. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'17)*, pp. 175–184, 2017.

[16] S. Gupta, G. Sharma, and A. Dukkipati. A generative model for dynamic networks with applications. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19)*, Vol. 33, pp. 7842–7849, 2019.

[17] E. Hajiramezanali, A. Hasanzadeh, K. Narayanan, N. Duffield, M. Zhou, and X. Qian. Variational graph recurrent neural networks. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS'19)*, 32, 2019.

[18] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proceedings of Advances in Nneural Information Processing Systems (NIPS'17)*, pp. 1024–1034, 2017.

[19] C. Hare and K. T. Poole. The polarization of contemporary american politics. *Polity*, 46(3):411–429, 2014.

[20] C. Heaukulani and Z. Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *Proceedings of International Conference on Machine Learning (ICML'13)*, pp. 275–283, 2013.

[21] I. Himelboim, K. Sweetser, S. Tinkham, K. Cameron, M. Danelo, and K. West. Valence-based homophily on twitter: Network analysis of emotions and political talk in the 2012 presidential election. *New Media & Society*, 18:1382–1400, 2014.

[22] J. Hu, M. P. Wellman, et al. Multiagent reinforcement learning: theoretical framework and an algorithm. In *International Conference on Machine Learning (ICML'98)*, pp. 242–250, 1998.

[23] D. J. Isenberg. Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50:1141–1151, 1986.

[24] H. Ito and C. Faloutsos. Dualcast: Friendship-preference co-evolution forecasting for attributed networks. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM'22)*, pp. 46–54, 2022.

[25] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[26] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[27] S. Liang, X. Zhang, Z. Ren, and E. Kanoulas. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD'18)*, pp. 1764–1773, 2018.

[28] M. Lim, A. Abdullah, N. Jhanjhi, M. K. Khan, and M. Supramaniam. Link prediction in time-evolving criminal network with deep reinforcement learning technique. *IEEE Access*, 7:184797–184807, 2019.

[29] C. Lozares, J. M. Verd, I. Cruz, and O. Barranco. Homophily and heterophily in personal networks. from mutual acquaintance to relationship intensity. *Quality & Quantity*, 48:2657–2670, 2014.

[30] M. Mcpherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[31] K. Miller, M. Jordan, and T. Griffiths. Nonparametric latent feature models for link prediction. *Proceedings of Advances in Neural Information Processing Systems (NIPS'09)*, pp. 1276–1284, 2009.

[32] M. Nie, D. Chen, and D. Wang. Reinforcement learning on graphs: A survey. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[33] A. Nigam, K. Shin, A. Bahulkar, B. Hooi, D. Hachen, B. K. Szymanski, C. Faloutsos, and N. V. Chawla. One-m: Modeling the co-evolution of opinions and network connections. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'18)*, pp. 122–140. Springer, 2018.

[34] M. Okawa and T. Iwata. Predicting opinion dynamics via sociologically-informed neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22)*, pp. 1306–1316, 2022.

[35] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'14)*, pp. 701–710, 2014.

[36] G. A. Quattrone and E. E. Jones. The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, 38:141–152, 1980.

[37] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, pp. 1067–1077, 2015.

[38] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha. Dyrep: Learning representations over dynamic graphs. In *Proceedings of International Conference on Learning Representations (ICLR'17)*, 2019.

[39] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo. Graphgan: Graph representation learning with generative adversarial nets. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18)*, Vol. 32, 2018.

[40] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang. Community preserving network embedding. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI'17)*, pp. 203–209, 2017.

[41] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press, 1994.

[42] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

[43] L. Yu, J. Song, and S. Ermon. Multi-agent adversarial inverse reinforcement learning. *arXiv preprint arxiv:1907.13220*, 2019.

[44] X. Zheng, Y. Liu, S. Pan, M. Zhang, D. Jin, and P. S. Yu. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*, 2016.

[45] L.-k. Zhou, Y. Yang, X. Ren, F. Wu, and Y. Zhuang. Dynamic network embedding by modeling triadic closure process. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI'18)*, pp. 571–578, 2018.

健康行動を促す介入メッセージのパーソナライゼーションに向けた文の特徴量分析

郷原　聖士† 　瀧本　祥章†† 　金　 秀明††
小林のぞみ†† 　倉島　 健†† 　渡辺　太郎†
† 奈良先端科学技術大学院大学　〒 630-0192 奈良県生駒市高山町 8916 番地-5
†† NTT 人間情報研究所　〒 239–0847 神奈川県横須賀市光の丘 1-1
E-mail: †{gobara.seiji.gt6,taro}@is.naist.jp,
††{yoshiaki.takimoto,hideaki.kin,nozomi.kobayashi,takeshi.kurashima}@ntt.com

あらまし　健康行動の機会不足は現代の社会的課題であり，解決手段としてメッセージによる介入に着目した研究が行われている．従来の研究では，メッセージの内容や提示方法の工夫による介入手法が提案されており，特定のユーザ属性（例：時間選好）において，行動変容に効果的な文の特徴量が異なることが確認されている．本研究では，ユーザの目標を達成するための健康行動（行動目標）の達成をサポートするため，性別や行動変容ステージを含む幅広いユーザ属性に対して効果的な文の特徴量を分析することで，より効果的なメッセージによる介入を目指している．具体的には，SNS での行動変容を促す投稿を用いて，健康行動に対する前向きな反応と文の特徴量との関連性を，相関係数と回帰モデルを通じて調査した．分析の結果，絵文字の使用頻度や文長がユーザ属性に依らず行動変容に有効であることが示された．また，ユーザに推奨する健康行動（行動目標）に応じて絵文字の使用の有無が行動変容の効果に差を生じさせることも確認され，異なる行動目標に対して有効な特徴量には差異が存在することが示唆された．本研究は，介入メッセージのパーソナライゼーションを目指し，各ユーザに対して最適化されたメッセージの作成に寄与することが期待される．

キーワード　行動変容，パーソナライゼーション, Twitter, ヘルスケア

1　はじめに

現在，人々は運動などの健康行動を行う機会が不足しており，社会的な課題となっている [1]．この課題に対処するために，情報提供，演習，議論を通じて健康行動の理解を深める方法や，ゲームデザインの要素を活用した行動変容を促す介入方法がある [2,3]．これには，禁煙，環境保護，食事選択などの多岐にわたる分野が含まれている．その中で，個々の人々の状況に合わせて効果的な行動変容を実現するために，行動変容ステージに基づいた介入方法がある．行動変容ステージモデルは，ある行動に対する関心や実施状況を基にした進行段階として定義される [4]．このモデルは,「無関心期」,「関心期」,「準備期」,「実行期」,「維持期」の 5 段階で構成され，それぞれ以下の状態を指す [5].

（1） 無関心期：6 ヶ月以内に行動を変えようと思っていない

（2） 関心期：6 ヶ月以内に行動を変えようと思っている

（3） 準備期：1 ヶ月以内に行動を変えようと思っている

（4） 実行期：行動を変えて 6 ヶ月未満である

（5） 維持期：行動を変えて 6 ヶ月以上である

メッセージを使った介入の例としては, 図 1 に示すような運動を呼びかける介入メッセージがある．介入メッセージとは，メッセージを見たユーザが実際に行動に移したくなるような，前向きな気持ちへと変化させるメッセージである．例えば， 図 1 に示されるように，運動への関心はあるが実行に移せていない関心期のユーザに対して,「1 日 5 分」という，行動に移すまでのハードルが低い文言を介入メッセージに取り入れることで，ユーザがやる気の状態になるように呼びかけている．

しかしながら，行動変容ステージモデルが示すように，画一的な介入メッセージが常に有効であるとは限らない．例えば，図 1 の例では，既に行動に移しているユーザに対して，介入メッセージが継続を促す内容でないため，実行期から維持期への移行を促す効果は限定的であると考えられる．また，行動変容ステージに限らず，ユーザが持つバイアスを考慮することも重要である．具体的には，近視眼的バイアスを示すユーザに対しては,「すぐに」,「楽に」のような近視眼的な単語を含む投稿が，他の投稿と比較して，より効果的にユーザの行動変容を促す可能性がある．

このようにユーザによってメッセージを出し分けるための仕組みとして， リマインダーメッセージ [6] や動機付けナッジ [7]，ユーモアと脅威性を組み合わせたメッセージ [8] がある．佐藤らは， 健康行動を促進するための介入メッセージについての分析を行い， 動機付けナッジが健康行動に有効であることを示した [9,10]．また， ユーザの時間選好の強さや性別な

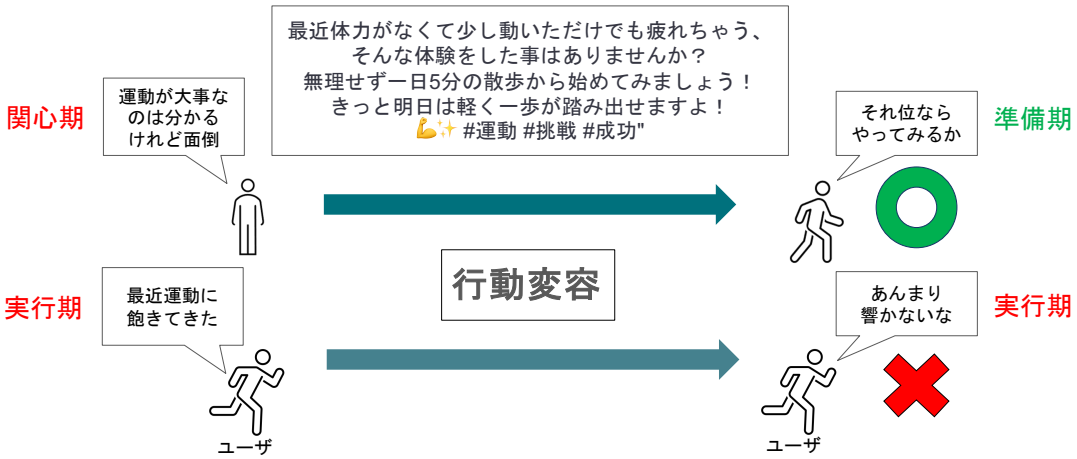


図 1: メッセージによる介入例

どの特定のユーザ属性によって，行動変容に有効な文の特徴量が，変化することが明らかになっている [7,8]．このことから，一律のメッセージではなく，個々のユーザ属性（例：行動変容ステージ，近視眼的行動バイアス）を考慮した，カスタマイズされたメッセージを提示することで，行動変容をより効果的に促すことが期待される．この背景に基づき，行動変容ステージや近視眼的行動バイアスなどのユーザ属性によって，行動変容に有効な文の特徴量が変化する， という仮説が立てられる（仮説 A）．また，行動変容で得られる価値について，同一の強調方法を行っても，行動目標によって，その有効性が異なることが知られている [9–11]．そのため，強調方法と同様に，有効な文の特徴量についても，行動目標によって，その効果が異なるという仮説が立てられる（仮説 B）．

そこで本研究では，介入メッセージのパーソナライゼーションに向けて，ユーザ属性と行動目標によって，健康行動に有益な文の特徴が変化するのかを検証した．分析では，SNS 上で行動変容を促す投稿を収集したデータセットに対して，投稿を見て健康行動に前向きになった割合と，文の特徴量との相関係数を計算した．また，健康行動に対する，前向きな反応を予測する回帰モデルを構築し，モデル中の特徴量の重要度から，予測に寄与する特徴量を調査した．分析の結果，有効な文の特徴量へのユーザ属性による影響は確認できなかったが，全体として絵文字の個数や文長，文の数などが行動変容に有効であることが分かった．また，運動や食事などの行動目標に対して，絵文字の使用の有無で行動変容の促しやすさが変化するのかを検証し，行動目標によって文の特徴量の有効性は異なることが示唆された．

2 関 連 研 究

2.1 SNS 上でのテキスト分析

Arakawa ら [12] は，X（Twitter）で収集した投稿を分析し，リツイート（RT）数とユーザタイプを分類する研究を実施した．実験では，投稿中の機能語の出現割合や品詞などの文の特徴量を用いて，ランダムフォレスト [13] を活用した分類モデルを学習し，RT 数の大小とユーザタイプを高精度で識別できることを示した．また，Lo ら [14] は，SNS 上の一般ユーザの投稿から，特定のトピックに対して潜在的な顧客を識別する分類モデルを構築した．このモデルは，文書の特徴量を基にしてサポートベクターマシンを学習し，潜在的な顧客を高精度で区別することから，ビジネス分野での応用が期待されている．

これらの研究から，SNS における拡散や広告の効果を高める，すなわち，読み手の反応を最大化するために，文の特徴量を活用するアプローチが有効であることが分かる．この知見は，行動変容のメッセージをパーソナライズする際にも有用であると考えられる．

2.2 介入メッセージの効果的活用

医療現場での無断キャンセルを削減するため，Junodo ら [6] は携帯電話を用いたリマインダーメッセージの効果を検証し，リマインダーメッセージが無断キャンセルの削減に寄与することを示した．また，Hallsworth ら [15] は，予約の欠席がもたらす，コストの無駄に焦点を当てたリマインダーメッセージを患者に送ることで，通常のメッセージよりも欠席率が優位に低下することを報告している．

さらに，介入メッセージの効果を高めるために，ナッジの活用が注目されている．例えば，濱谷ら [16] は，人の行動をさりげなく良い方向へと導く，ナッジを取り入れたメッセージに関するアンケートを分析し，そのようなメッセージが受容性を高めることを示唆している．佐藤ら [9,10] は，健康行動の意思決定を促進する，動機付けナッジメッセージの効果を検証し，将来の健康の価値や，推奨行動で実現できる本来の目標とは異なる利点，推奨行動と比較した時の推奨行動の効果を伝えることが，健康行動の動機付けに有効である可能性を示した．

山本ら [7] は，ユーザの時間選好に基づく即時報酬と遅延報酬メッセージの使い分けが，健康行動の促進に効果的であることを明らかにした．また，Sasaki ら [11] は，新型コロナウィルス感染症（COVID-19）対策として使用されたナッジベースのメッセージが，自発的な接触回避と感染予防行動にどのように影響を与えるかを調査した．調査では，自身の行動が身近な

人々の命を守ることを強調する排他的なメッセージが，外出や他者との接触頻度を減少させる効果があることを示した．一方で，このメッセージは，実験前に元々外出頻度が低かったグループに対しては，感染予防行動の遵守を低下させる逆効果をもたらすことも報告されている．

本研究では，効果的な介入メッセージの構築に必要な，文の特徴量の特定に焦点を当てている．これは，過去の研究でメッセージ例や評価者が限られていたこと，および特定のユーザ属性に特化した文の特徴量の効果を仮説検証していたことに由来する [9,10]．さらに，ユーザ属性や行動目標によって，行動変容に効果的な文の特徴量が異なることも知られている [9–11]．以前の研究では，介入メッセージによる，行動変容を促す効果を調査することに焦点が当てられていたが [6,15]，本研究ではこれらの知見を踏まえ，メッセージ中の文の特徴量を詳細に分析することで，より多様なユーザ属性を考慮した，行動目標毎に効果的な介入メッセージを提示することを目的としている．具体的には，瀧本らが作成した多数のメッセージ例と評価者，そして幅広いユーザ属性を考慮したアンケート [17] を用いて，行動変容に役立つ文の特徴量を探索的に調査した．

3 準　　備

3.1 データセット

本分析では，性別や行動変容ステージなどの多様なユーザ属性と行動目標を考慮するため，バランスの取れた十分なサンプルサイズを有するデータセットを利用する必要がある．そこで本研究では，上記の条件を満たす，瀧本らの作成したデータセット [17] を使用し，行動変容に有益な文の特徴量を分析した．瀧本らは，Twitter で行動変容を促す投稿 630 件を収集し，「投稿を見て前向きになったのかどうか」を 4 段階（なった，ややなった，あまりならなかった，ならなかった）でアンケート調査を実施した．分析では，メッセージを見た後の前向きな反応割合（「なった」，「ややなった」の選択肢を選んだ割合の合計）を行動変容確率として定義した．

3.2 ユーザ属性と行動目標

本研究では，アンケート回答者の行動変容ステージ，性別および近視眼的行動バイアスの有無をユーザ属性として調査した．近視眼的行動バイアスとは，短絡的利益を優先する行動の傾向を指し，回答者が「a: 今すぐ 1 万円が欲しいか」（近視眼的）あるいは「b:1 年後に 11 万円が欲しいか」（俯瞰的）という質問をした時に前者を選ぶような行動バイアスである．行動変容ステージは，「関心期」，「準備期」，「実行期」，「維持期」の 4 段階を調査対象とした．

さらに，行動目標毎に有益な文の特徴量は異なるのかを調査するため，各投稿に対して行動目標を人手で分類した．分類した行動目標は，「運動」，「食事」，「医療」，「挨拶」，「目標」，「睡眠」，「その他」の 7 種類である．分類された介入メッセージの具体例を表 1 に示す．ただし，行動目標毎の分析では，元のデータ 630 件からテキスト情報が適切に得られなかったデータを除外し，最終的には 613 件のデータを使用した．

3.3 文の特徴量

今回調査対象としたテキストの特徴量は表 2 の通りである．Arakawa ら [12] は，RT 数の大小の予測には品詞の割合などの文の特徴量が重要であることを示唆しており，「情報の拡散しやすさ」，すなわち，投稿の受け入れやすさは行動変容の促しやすさにつながると考えられる．そのため，品詞，機能語，およびその他（特殊な語などの語彙）の出現割合を分析対象として選定した．本研究では，unidic-lite[1]を辞書として使用し，Unidic の品詞体系 [18] に基づいて形態素解析器（MeCab[2]）を行った．解析では以下の品詞，機能語，その他の語彙を抽出した．

（1） 品詞：名詞，動詞，副詞，形容詞，連体詞，接続詞，感動詞

（2） 機能語：名詞-非自立，名詞-代名詞，接続詞， 連体詞，助詞，助動詞

（3） その他：品詞で選ばなかった分類の割合

また，ユーザに行動を促すための説得対話において，感情表現の役割が重要であることから [19]，介入メッセージにおいても感情を含む文章が行動変容を促しやすい，という仮説を立てて，感情分析に基づく極性値を特徴量として採用した．極性値 S は，汎用的な自然言語処理基盤エンジンであるリッチインデクサ [3]を使用して得られた，正（positive），負（negative），中立（neutral）の 3 種類の各単語について，出現頻度を考慮して

$$S = \frac{1 \cdot C_{\text{positive}} - 1 \cdot C_{\text{negative}} + 0 \cdot C_{\text{neutral}}}{C_{\text{positive}} + C_{\text{negative}} + C_{\text{neutral}}},$$

ように定義する．ここで，C_{positive} は文章中の正の単語の個数，C_{negative} は文章中の負の単語の個数，C_{neutral} は文章中の中立の単語の個数を表している．

くわえて，本研究では，「すぐ」，「楽に」のような，近視眼的行動バイアスに影響する可能性のある単語が，メッセージ内で近視眼的な文脈で使用されている例を，630 件のメッセージから 44 件，人手で抽出した．さらに，TF-IDF 分析で上位 200 件の単語からメッセージを特徴付けるもの（例：ダイエット，禁煙，効果，絶対など）を含む 54 件のメッセージを抽出した．

4 分 析 手 法

本研究では，行動変容に有益な文の特徴量を特定するために，性別や行動変容ステージなどでユーザを抽出し， 行動変容確率を求め，特徴量との相関係数を調査（4.1 節），LightGBM を用いて回帰モデルを構築し，回帰への寄与度を評価（4.2 節），行動目標毎に文章を人手で分類し，文の特徴量の有無による行動変容確率の平均値の差を調査（4.3 節）， という 3 種類の分析を行った．それぞれの具体的な手順は以下の通りである．

1：https://github.com/polm/unidic-lite
2：https://taku910.github.io/mecab/
3：https://www.rd.ntt/research/HI0006.html

表 1: 行動目標毎の介入メッセージ例

行動目標	介入メッセージ	アカウント id
運動	＼✨運動をしましょう✨／　忙しい毎日，まったく運動しない日が続いている…そんな方も多いと思いますが，体を動かすことは、生活習慣病やがんのリスクを減らしてくれるほか、軽い気分障害などの予防・解消にも有効です。　是非、健康のために運動習慣をつけていきましょう！ #ラビユキ　#Domingo https://t.co/UbtgxgNSCb	@Hokkaido_health [20]
食事	・新鮮な食材を使い、調味料の量を減らす ・漬物は浅漬けにし、少量にする ・外食や加工食品を控える ・醤油やソースを追加でかけることを避ける 以上のことに注意し、塩分の摂取量を減らしましょう！	@parkcoronavac [21]
医療	今日は #自分の BMI を計算してみましょう🧐 #BMI は肥満の指標となる数値☝ 計算方法 BMI=体重 (kg) ÷身長 (m) ÷身長 (m) 正常範囲は 18～49 歳で 18.5～25 未満 50～64 歳で 20～25 未満で「22」が標準体重になります📝 計算してみてくださいね😉 #今日のいいこと #ぷちけん https://t.co/dbLOrHqDPN	@askenjp [22]
挨拶	【謹賀新年】 明けましておめでとうございます🐇✨ 本年も、道民の皆様をはじめ、Twitter をご覧の皆様に、健康づくりに役立つ様々な情報を発信していきたいと思いますので、どうぞよろしくお願いいたします☺👏 #北海道　#健康づくり https://t.co/5p6Bn3qrOX	@Hokkaido_health [23]
目標	今日から #4 月 ですね🌸 4 月は #新しいこと に挑戦しやすい時期です☺ 今日は #夏までに達成したい目標 を 1 つ決めてみませんか？ 🏝 夏までの約 3 か月で達成できそうな 「7 月までに〇 kg になる‼」などなんでも OK です👌 目標を決めると、やる気もアップしますよ😉 #今日のいいこと #ぷちけん	@askenjp [24]
睡眠	＼✨睡眠をとりましょう✨／　寝不足や、質の悪い睡眠が続くと、生活習慣病のリスクを高めてしまいます…！　快眠を取るためには、運動習慣をつけたり、お風呂に入り体を温めることが効果的です✨ きちんと体と心を休ませて、健康な体づくりをしましょう☺ #北海道　#ラビユキ #Domingo #雪ミク https://t.co/rXzTvdn6Mv	@Hokkaido_health [25]
その他	9 月は #健康増進普及月間 です 「1 に運動 2 に食事 しっかり禁煙 最後にクスリ」を合言葉に、生活習慣・健康について考えるきっかけにしましょう！ https://t.co/VBk60QRP7q https://t.co/C2vbx2FqB1	@wakayama_master [26]

表 2: 選定理由と特徴量

No.	選定理由（仮説）	特徴量
1	読みやすさがユーザに影響を与えやすい	
	a：文章の長さによる影響	文章の長さ，平均文長，文数
	b：単語の構成割合による影響	機能語, 品詞, 記号, その他の（特殊な語などの）語彙の出現割合
2	目を惹く単語が含まれているとユーザに影響を与えやすい	ハッシュタグの個数
3	絵文字が含まれているとユーザに影響を与えやすい	絵文字の個数
4	疑問文があるとユーザの関心を引きやすい	疑問文（クエスチョンマーク）の個数
5	データが示されていると信頼度が高いと判断されやすい	数詞の個数
6	正，負の感情が含まれているとユーザに影響を与えやすい	正・負・中立の単語の個数, 感情分析の極性値
7	近視眼的バイアスのかかった人は近視眼的な単語に影響されやすい	近視眼的な単語（例：ちょっと, 手軽）

4.1　ユーザ属性毎に有効な特徴量の調査

本研究では，性別や行動変容ステージなどのユーザ属性でフィルタリングした上で，各文の特徴量と介入メッセージの行動変容確率の間でスピアマンの相関係数を計算した．本稿では有意水準 5%として，「あるメッセージの文の特徴量とそのメッセージの行動変容確率には，母集団で順位相関が存在しない」という帰無仮説の無相関検定を，スピアマンの順位相関係数で実施した．

4.2　回帰モデルを用いた各特徴量の重要度の調査

アンケートのメッセージと行動変容確率のペアに対して，データセットを 訓練：評価 = 9：1 の割合で分割した．その後，scikit-learn[4]を用いて，訓練データで 5 分割交差検証を行い，LightGBM（回帰モデル）で行動変容確率を推定するように学習させた．モデルのハイパーパラメータは，訓練用データと optuna[5]を用いてベイズ最適化で決定し，その最適なパラメータで行動変容確率の推測を行った．最後に，特徴量の重要度を SHAP 値と split，gain によって計算した．ここで，SHAP 値は，ある特徴量を追加した時に得られる限界貢献度，split は，木構造ベースのモデルで分岐回数に基づいて計算される特徴量の重要度，gain は，決定木や木構造ベースのモデルで新しい分割を行った際に特徴量がモデルの性能向上に寄与したのかを評価する指標を指す．

4.3　行動目標毎に有効な文の特徴量の調査

行動目標毎に有効な特徴量が異なるかを検証するために，ここでは 5.1 節と 5.2 節で有効であることが確認された，絵文字を対象として，行動変容確率に変化は見られるのかを分析した．

4：https://scikit-learn.org/

5：https://www.preferred.jp/ja/projects/optuna/

表 3: 行動変容ステージと行動変容確率の相関係数

No.	特徴量	関心期	準備期	実行期	維持期	関心期ー準備期	実行期ー維持期	全体
1a	**文章の長さ**	0.163*	0.162*	0.150*	0.146*	0.169*	0.159*	0.170*
	平均文長	−0.041	−0.080	−0.056	−0.060	−0.056	−0.061	−0.057
	文数	−0.109*	−0.144*	−0.143*	−0.116*	−0.127*	−0.135*	−0.133*
1b	**機能語の出現割合**	0.002	0.014	−0.029	0.018	0.011	−0.003	0.005
	品詞の出現割合	0.051	0.055	0.075	0.043	0.051	0.061	0.057
	記号の出現割合	−0.147*	−0.160*	−0.135*	−0.171*	−0.158*	−0.167*	−0.165*
	その他の出現割合	0.053	0.057	0.024	0.081	0.061	0.060	0.061
2	**ハッシュタグの個数**	0.100*	0.140*	0.110*	0.122*	0.117*	0.123*	0.123*
3	**絵文字の個数**	0.176*	0.194*	0.195*	0.173*	0.192*	0.198*	0.199*
4	**疑問文の個数**	0.071	0.053	0.087	0.053	0.064	0.077	0.073
5	**数詞の個数**	−0.031	−0.051	−0.048	−0.076	−0.042	−0.067	−0.050
6	**正の単語の個数**	−0.001	−0.001	−0.052	−0.028	−0.004	−0.042	−0.018
	負の単語の個数	−0.052	−0.059	−0.073	−0.064	−0.058	−0.076	−0.068
	中立の単語の個数	0.014	−0.019	0.050	0.028	0.003	0.042	0.019
	感情極性値	0.028	0.044	0.016	0.021	0.035	0.020	0.032

表 4: ユーザ属性（性別）と行動変容確率の相関係数

No.	特徴量	男性	女性	全体
1a	**文章の長さ**	0.140*	0.185*	0.170*
	平均文長	−0.020	−0.082	−0.057
	文数	−0.098*	−0.150*	−0.133*
1b	**機能語の出現割合**	0.027	−0.021	0.005
	品詞の出現割合	0.028	0.075	0.057
	記号の出現割合	−0.156*	−0.158*	−0.165*
	その他の出現割合	0.094	0.030	0.061
2	**ハッシュタグの個数**	0.091*	0.137*	0.123*
3	**絵文字の個数**	0.149*	0.219*	0.199*
4	**疑問文の個数**	0.056	0.073	0.073
5	**数詞の個数**	−0.062	−0.030	−0.050
6	**正の単語の個数**	−0.027	−0.005	−0.018
	負の単語の個数	−0.051	−0.068	−0.068
	中立の単語の個数	0.007	0.022	0.019
	感情極性値	0.024	0.032	0.032

分析では，絵文字の有無によって，行動目標毎の行動変容確率の平均値に差があるか，対応がなく二郡のデータの母分散が等しいと仮定できない点を踏まえて，Welch の t 検定（$\alpha = 0.05$）を実施した．

5　分析結果と考察

5.1　ユーザ属性毎に有効な文の特徴量の調査

表 3–5 に，それぞれ行動変容ステージ，性別，近視眼的行動バイアスで場合分けをした，ユーザ属性と行動変容確率とのスピアマンの順位相関係数を示す．図 2–4 にユーザ全体に対する行動変容確率と文章の長さ，感情分析の極性値，絵文字の個数との関連性を示す．表 3–5 中の*記号は帰無仮説を検定し，p 値が 0.05 未満であり，帰無仮説が棄却されたことを表している．

表 3–5 において，ほとんどの項目で相関係数の絶対値は 0.20 未満であった．さらに， 表 3–5 では 統計的有意差が見られた文の特徴量は共通していたため，性別や行動変容ステージなどのユーザ属性によって特徴量の有効性について，違いは確認できなかった．正の相関が確認できたのは，文章の長さ（図 2），ハッシュタグ数，絵文字の個数（図 3）であり，ハッシュタグや絵文字を積極的に使うことで，行動変容を促すことができることを示唆している．

絵文字に関しては，メッセージに視覚的な感情表現を含んでいるため，感情面での訴えかけや親しみやすさに寄与したと考えられる．しかしながら， 絵文字の個数が多い投稿でも， 行動変容確率が低下する場合も存在していた．これは，絵文字が視覚的な影響を与える一方で，文章を見た時の受け取り方に負の影響を及ぼす可能性があるため，絵文字の使用についてさらに詳細な分析が必要である．くわえて，今回の研究では，絵文字の個数を記録する際に，正規表現で単純に数え上げていた．このことから， 喜怒哀楽のように，様々な絵文字の種類や意味を考慮できていない可能性がある．そこで， 絵文字の種類や任意の絵文字を組み合わせたデータセットを構築し，より詳細な分析を実施する方が望ましい．

一方で，文数，記号の出現割合では負の相関が確認され，多くの文に分けたり，記号を過剰に使用したりすることで，行動変容を促す効果が減少することを示唆している．なお，表 5 中の選択肢「どちらでもない」について，有意差が見られなかったのは，該当するユーザが少なかったことが原因と考えられる．

また，疑問文と数詞の個数に有意差は認められなかった（特徴量 4，5）．感情極性に関しても，感情面で訴えかける説得対話の戦略 [27] に反して，有意差は見られなかった（特徴量 6，図 4 参照）．これは，感情極性の計算時に，正，負，中立の単語自体を単純に抽出していたため，単語の組み合わせによる複雑な感情表現を捉えることができていなかった可能性がある．

さらに，図 5 にキーワード「禁煙」を含むデータと行動変容確率の相関を示す．キーワード単位の検証では，一部のキーワード

表 5: 近視眼的行動バイアスと行動変容確率との相関係数

No.	特徴量	近視眼的	やや近視眼的	どちらでもない	やや俯瞰的	俯瞰的	近視眼的計	俯瞰的計	全体
1a	**文章の長さ**	0.170*	0.150*	0.062	0.166*	0.142*	0.175*	0.159*	0.170*
	平均文長	−0.057	−0.077	−0.038	−0.062	−0.044	−0.060	−0.058	−0.050
	文数	−0.133*	−0.136*	−0.082*	−0.130*	−0.120*	−0.128*	−0.133*	−0.118*
1b	**機能語の出現割合**	0.005	0.015	−0.042	0.025	−0.010	0.006	0.005	0.004
	品詞の出現割合	0.057	0.096*	0.030	0.043	0.031	0.080	0.035	0.068
	記号の出現割合	−0.165*	−0.188*	−0.045	−0.158*	−0.115*	−0.191*	−0.139*	−0.176*
	その他の出現割合	0.061	0.051	0.008	0.008	0.045	0.056	0.064	0.055
2	**ハッシュタグの個数**	0.123*	0.168*	0.012	0.110*	0.107	0.141*	0.113*	0.119*
3	**絵文字の個数**	0.199*	0.204*	0.055	0.203*	0.158*	0.215	0.185*	0.200*
4	**疑問文の個数**	0.073	0.023	−0.033	0.091*	0.061	0.074	0.077	0.088
5	**数詞の個数**	−0.050	−0.066	−0.015	−0.045	−0.030	−0.057	−0.041	−0.052
6	**正の単語の個数**	−0.018	−0.004	−0.012	−0.003	0.011	−0.035	0.001	−0.043
	負の単語の個数	−0.068	−0.075	−0.064	−0.070	−0.057	−0.062	−0.064	−0.051
	中立の単語の個数	0.019	0.000	0.039	0.040	0.016	0.016	0.029	0.021
	感情極性値	0.032	0.060	0.054	0.030	0.045	0.018	0.038	0.000

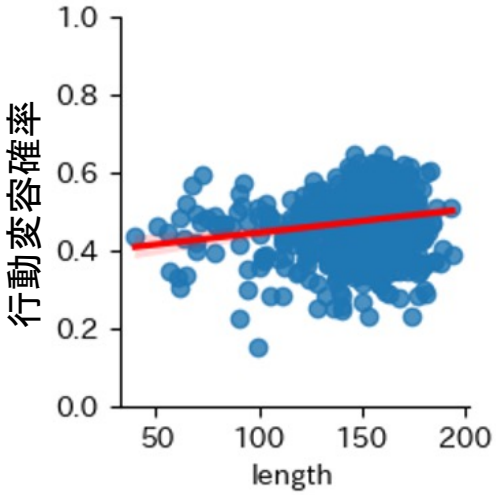


図 2: 文章の長さと行動変容確率の相関

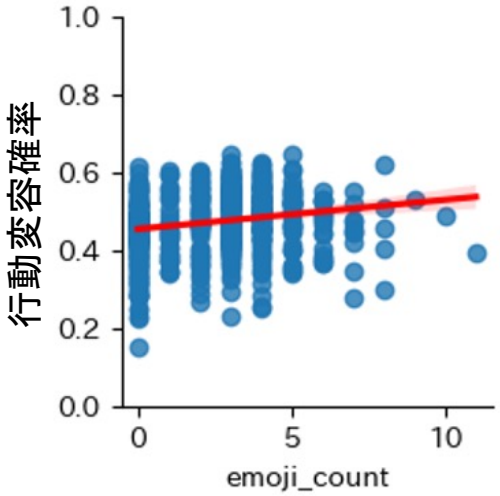


図 3: 絵文字の個数と行動変容確率の相関

(例:「禁煙」, 図 5 参照) に有意差が認められたが ($p < 0.001$), ユーザ属性による差異は確認できなかった. くわえて,「ちょっと」,「手軽」のような近視眼的な人に対し有効であると予想された単語は, 予想に反して, 近視眼的行動バイアスを持つ人々の行動変容確率との間に, 有意な関連性 (「ちょっと」$p = 0.354$; 「手軽」 $p = 0.450$) (特徴量 7) は見られなかった.

以上の結果から, 行動変容を促す介入メッセージには, 文章の長さや絵文字のような, 文の特徴量が有効であると考えられる. さらに, 仮説 A に反して, 行動変容ステージや性別, 近視眼的行動バイアスが変化しても, 行動変容に有効な文の特徴量を確認できなかった. これは, 行動変容を促すメッセージの効果が, ユーザ属性よりも, メッセージ自体の特性によって大きく影響されることを示唆している.

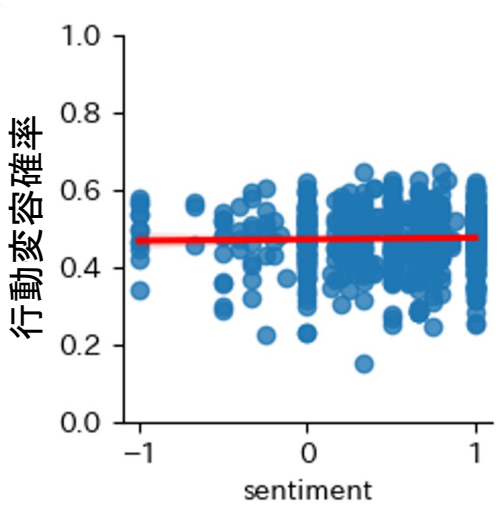


図 4: 感情分析の極性値と行動変容確率の相関

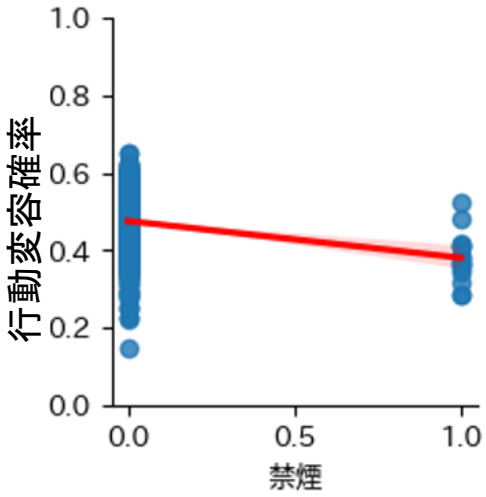


図 5: キーワード「禁煙」の有無と行動変容確率の相関

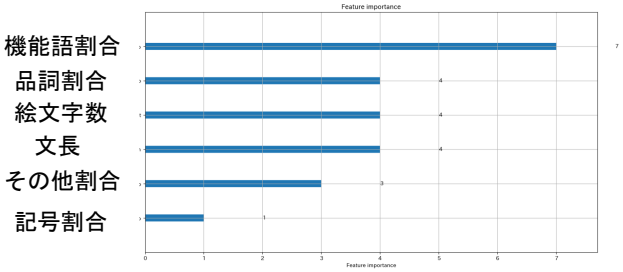


図 6: LightGBM での特徴量の重要度（split）

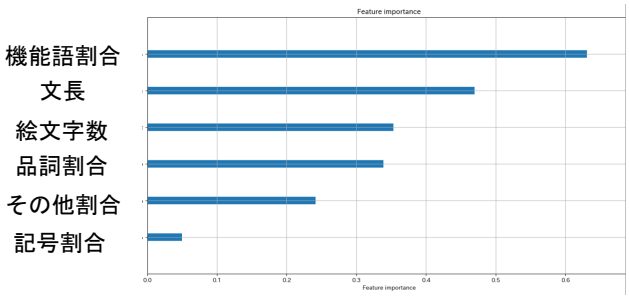


図 7: LightGBM での特徴量の重要度（gain）

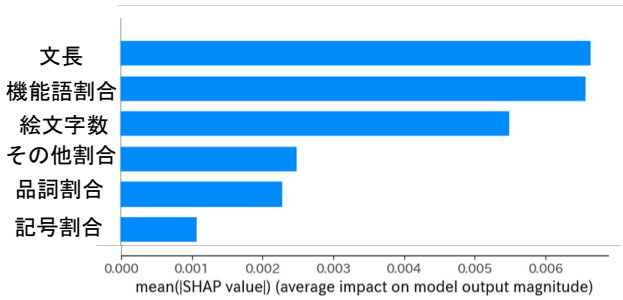


図 8: LightGBM での特徴量の重要度（SHAP）

表 6: 行動目標に対する行動変容確率の平均

行動目標	絵文字有（件数）	絵文字無（件数）
運動	0.474 (81)	0.451 (32)
食事	0.489 (279)	0.451 (93)
医療	0.478 (51)	0.399 (30)
挨拶	0.476 (5)	none (0)
目標	0.483 (17)	none (0)
睡眠	0.499 (14)	0.432 (3)
その他	0.530 (6)	0.417 (2)
全体	0.485 (453)	0.440 (160)

5.2 回帰モデルを用いた各特徴量の重要度の調査

LightGBM を用いた特徴量の重要度をそれぞれ図 6（split），図 7（gain），図 8（SHAP 値）に示す．図 6–8 では，文章の長さ，絵文字数，その他の（特殊な語などの）語彙の出現頻度の重要度が高かった．一方で特徴量の重要度については，機能語や品詞の出現割合の影響が大きいことが確認された．なお，文の各特徴量の重要度分析において，5.1 節で統計的に有意な文の特徴量として挙げられた特徴量と，回帰モデルの予測で重要度の高い特徴量は，多くの特徴量が共通していた．

5.3 行動目標毎に有効な文の特徴量の調査

行動目標毎の行動変容確率の平均を，特徴量（絵文字の有無）で条件付けた結果を表 6 に示す．表 6 において，絵文字を含む投稿と絵文字を含まない投稿の行動変容確率を比較すると，平均値に全体で 4%以上も差があることが確認できる．また，十分なデータ数が存在した 3 つのカテゴリ（運動，食事，医療）および全体について，Welch の t 検定を行い，ボンフェローニの調整を実施した結果，食事，医療，全体で有意差が認められた（「食事」$p < 0.001$;「医療」$p = 0.003$;「全体」$p < 0.001$）．しかし，運動では有意差が確認できなかった（$p = 0.815$）．したがって，ユーザ属性によって文の特徴量の有効性に差は見られなかった一方で，行動目標によって有効な特徴量は異なる可能性がある，という仮説 B が支持される結果となった．

6 おわりに

本研究では，介入メッセージのパーソナライゼーションに向けて，SNS 上で行動変容を促す投稿が収集されたアンケートに対し，ユーザ属性と行動目標毎に有益な文の特徴量分析を実施した．アンケートの特徴量分析の結果，介入メッセージに有益な特徴量として，文長，ハッシュタグ，絵文字の個数，品詞の割合があることが明らかになった（特徴量 1–3）．さらに，行動変容ステージと性別，近視眼的行動バイアスの有無によって，行動変容に有効な文の特徴量に変化は見られなかった（仮説 A）．また，行動目標毎に介入メッセージを分析すると，投稿に含まれる絵文字の有無によって，食事，医療で行動変容確率に変化があったが，運動ではその変化は見られなかったため，行動目標によって有効な文の特徴量が異なることが示唆された（仮説 B）．今回は絵文字やハッシュタグなどの特徴について調査し

たが，絵文字以外の要素である， 文体の「だ・である調」や「ですます調」， 敬語の有無などが与える印象の違いについて，今後の調査が望まれる．

今回明らかになった絵文字などの特徴量を活用することで，介入メッセージの人手作成時に有用なガイドラインとなり，効率的で効果的な介入メッセージの作成が期待される．くわえて，介入メッセージの自動作成を検討する際に，機械学習モデルの訓練用データセットの選定にも役立つ可能性がある．今後は，ユーザ属性毎に有効な特徴量をさらに分析し，特定した特徴量を活用してより効果的な介入メッセージを提示する，言い換えモデルの構築を目指している．

文　　献

[1] World Health Organization. Global status report on physical activity 2022. https://iris.who.int/bitstream/handle/10665/363607/9789240059153-eng.pdf?sequence=1, 2022. Accessed on November 23, 2023.

[2] Mahmood Karimy, Peter Higgs, Shaghaygh Solayman Abadi, Bahram Armoon, Marzieh Araban, Mohammad Reza Rouhani, and Fereshteh Zamani-Alavijeh. Oral health behavior among school children aged 11–13 years in saveh, iran: an evaluation of a theory-driven intervention. *BMC pediatrics*, Vol. 20, pp. 1–9, 2020.

[3] Daniel Johnson, Sebastian Deterding, Kerri-Ann Kuhn, Aleksandra Staneva, Stoyan Stoyanov, and Leanne Hides. Gamification for health and wellbeing: A systematic review of the literature. *Internet interventions*, Vol. 6, pp. 89–106, 2016.

[4] James O Prochaska and Wayne F Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, Vol. 12, No. 1, pp. 38–48, 1997.

[5] 松本千明. 行動変容ステージモデル. https://www.e-healthnet.mhlw.go.jp/information/exercise/s-07-001.html, 2019.

[6] Noelle Junod Perron, Melissa Dominicé Dao, Nadia Camparini Righini, Jean-Paul Humair, Barbara Broers, Françoise Narring, Dagmar M Haller, and Jean-Michel Gaspoz. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: a randomized controlled trial. *BMC health services research*, Vol. 13, No. 1, pp. 1–7, 2013.

[7] 山本修平, 赤木康紀, 冨永登夢, 瀧本祥章, 倉島健, 戸田浩之. ユーザの時間選好を考慮したメッセージ提示が健康行動に与える影響. マルチメディア, 分散, 協調とモバイルシンポジウム 2023 論文集, Vol. 2023, pp. 475–485, 2023.

[8] Hanneke Hendriks and Loes Janssen. Frightfully funny: Combining threat and humour in health messages for men and women. *Psychology & Health*, Vol. 33, No. 5, pp. 594–613, 2018.

[9] 佐藤妙, 阿部直人. 健康行動の意思決定における介入メッセージの検討 －日々の行動未実施のユーザ認知変更のためのナッジ（プログレス・レポート）－. 行動経済学会第 13 回予稿, pp. 1–6, 2019.

[10] 佐藤妙, 阿部直人. 健康行動の意思決定における介入メッセージの検討 －動機づけナッジメッセージの効果検証－. 行動経済学会第 14 回予稿, pp. 1–6, 2020.

[11] Shusaku Sasaki, Hirofumi Kurokawa, and Fumio Ohtake. Effective but fragile? responses to repeated nudge-based messages for preventing the spread of covid-19 infection. *The Japanese Economic Review*, Vol. 72, pp. 371–408, 2021.

[12] Yui Arakawa, Akihiro Kameda, Akiko Aizawa, and Takafumi Suzuki. Adding Twitter-specific features to stylistic features for classifying tweets by user type and number of retweets. *Journal of the Association for Information Science and Technology*, Vol. 65, No. 7, pp. 1416–1423, 2014.

[13] Leo Breiman. Random forests. *Machine learning*, Vol. 45, pp. 5–32, 2001.

[14] Siaw Ling Lo, Raymond Chiong, and David Cornforth. Using support vector machine ensembles for target audience classification on twitter. *PloS one*, Vol. 10, No. 4, p. e0122855, 2015.

[15] Michael Hallsworth, Dan Berry, Michael Sanders, Anna Sallis, Dominic King, Ivo Vlaev, and Ara Darzi. Stating appointment costs in sms reminders reduces missed hospital appointments: findings from two randomised controlled trials. *PloS one*, Vol. 10, No. 9, p. e0137306, 2015.

[16] 濱谷尚志, 山田渉, 落合桂一, 山本直樹, 檜山聡ほか. 健康に関する行動変容のための社会的メッセージ提示方法の検討. 研究報告マルチメディア通信と分散処理 (DPS), Vol. 2021, No. 5, pp. 1–6, 2021.

[17] 瀧本祥章, 金秀明, 小林のぞみ, 倉島健. 効果的な行動変容メッセージ実現に向けたインセンティブ効果の被験者評価. 第 16 回データ工学と情報マネジメントに関するフォーラム (DEIM)), 2024.

[18] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』 形態論情報規程集 第 4 版. 国立国語研究所, 2011.

[19] 石川葉子, 水上雅博, 吉野幸一郎, 鈴木優, 中村哲. 感情表現を用いた説得対話システム. 人工知能学会論文誌, Vol. 33, No. 1, pp. DSH–B_1, 2018.

[20] @Hokkaido_health. https://twitter.com/Hokkaido_health/status/1503159976075878408. Accessed on November 23, 2023.

[21] @parkcoronavac. https://twitter.com/parkcoronavac/status/1594829284212101120. Accessed on November 23, 2023.

[22] @askenjp. https://twitter.com/askenjp/status/1543701117044924418. Accessed on November 23, 2023.

[23] @Hokkaido_health. https://twitter.com/Hokkaido_health/status/1609338621085073408. Accessed on November 23, 2023.

[24] @askenjp. https://twitter.com/askenjp/status/1509636661310799873. Accessed on November 23, 2023.

[25] @Hokkaido_health. https://twitter.com/Hokkaido_health/status/1503159472562319363. Accessed on December 13, 2023.

[26] @wakayama_master. https://twitter.com/wakayama_master/status/1565127303587188737. Accessed on November 23, 2023.

[27] Irene Mazzotta, Fiorella De Rosis, and Valeria Carofiglio. Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent systems*, Vol. 22, No. 6, pp. 42–51, 2007.

DEIM Forum 2023 XX-Y

Approximate Inverse Model Explanations (AIME)を用いた内閣支持率変化要因抽出手法

津布久 尚貴† 峰松 彩子‡ 岡田 龍太郎† 中西 崇文†

†武蔵野大学データサイエンス学部 〒135-8181 東京都江東区有明3丁目3-3
‡武蔵野大学アジアAI研究所 〒135-8181 東京都江東区有明3丁目3-3

E-mail: †s2322046@stu.musashino-u.ac.jp,{ayako.minematsu, ryotaro.okada, takafumi.nakanishi}@ds.musashino-u.ac.jp

あらまし 本稿では，AIMEを用いた内閣支持率要因抽出手法について示す．これまで我々の研究室で説明可能なAI(XAI)手法の一つとしてApproximante Inverse Model Explantations(AIME)を提案してきた．AIMEとは，複雑もしくは中の構造が不明なAI，機械学習モデル(ブラックボックスモデル)について，訓練データなどのモデルを構成するためのデータから近似逆作用素を作ることにより，モデルの挙動やモデルへの入力に対してそのモデルでその値が推定されたかを説明するExplainable AI(XAI)の手法の一つである．本稿では，ニュース記事から内閣支持率を予測する機械学習モデル(ブラックボックスモデル)が存在すると仮定して，ニュースデータのTF-IDF(term frequency–inverse document frequency)から成り立つ説明変数と内閣支持率・不支持率の上昇/下降を目的変数とし，AIMEによる逆問題を解くことにより，内閣支持率・不支持率の上昇/下降に貢献する要素(単語)を抽出する手法を実現する．本方式を実現することにより，内閣に対する国民の受け入れや反発といった反応を可視化することによって，与野党の政策決定の一助となると考えられる．

キーワード 内閣支持率要因，Approximante Inverse Model Explantations(AIME)，説明可能なAI(XAI)，近似逆作用素，原因単語抽出

1. はじめに

内閣支持率の変動要因を理解することは，政治的意思決定，公共の意見形成，社会の動向分析など，多岐にわたる分野において重要であると考える．特に，政治的安定性や政策の受け入れ度を評価する上で，内閣支持率の上昇や下降の背景にある要因を知ることは，政府や政党にとって不可欠である.このような分析は，国民の意見や反応を正確に把握し，より効果的なコミュニケーション戦略や政策の策定を可能にする．したがって，内閣支持率の変動を引き起こす要素を明らかにすることは，政治学，社会学，コミュニケーション学の研究において重要な課題となっている.

一般的に,内閣支持率の上昇/下降の原因はこれまで限られた分析で留まっており,正確な公共の意見形成,社会の動向をとらえるまで至らないことが多く見受けられる。これらを具体的なデータから導出することで,正確な公共の意見形成，社会の動向を正しく導くことができると考えられる.

一方で,我々は,複雑もしくは中の構造が不明なAI,機械学習モデル(ブラックボックスモデル)について,訓練データなどのモデルを構成するためのデータから近似逆作用素を作ることにより，モデルの挙動やモデルへの入力に対してそのモデルでその値が推定されたかを説明するExplainable AI(XAI)の手法の一つであるApproximante Inverse Model Explantations (AIME) [1]を提案してきた．この手法により従来の各XAI手法よりもシンプルで解釈しやすい説明を導出することが可能であることを文献[1]では実証済みである.

AIMEの枠組みは，訓練データなどのモデルを構成するためのデータから近似逆作用素を作ることにより,モデルの挙動やモデルへの入力に対してそのモデルでその値が推定されたかを説明するため，例えば，ニュース記事から内閣支持率を予測する機械学習モデル(ブラックボックスモデル)が存在すると仮定して,AIMEを用いた逆問題導出手法を内閣支持率変動要因として実現することができれば，定量的に内閣支持率変動要因を発見することが可能となり，政治的意思決定のみならず，公共の意見形成，社会の動向を捉えるのに重要な機能の実現になると考える.

本稿では，ニュース記事から内閣支持率を予測する機械学習モデル(ブラックボックスモデル)が存在すると仮定して，ニュースデータのTF-IDFから成り立つ説明変数と内閣支持率・不支持率の上昇/下降を目的変数とし，AIMEによる逆問題を解くことにより，内閣支持率・不支持率の変動に貢献する要素(単語)を抽出する手法を実現する．本方式を実現することにより，定量的に公共の意見形成，社会の動向分析を捉えることのみならず，国民の受け入れや反発といった反応を可視化することによって，与野党の政策決定，政治的意思決定の一助になると考えられる.

本稿の貢献は以下の点にまとめられる.

・ ニュース記事から内閣支持率を予測する機

学習モデル(ブラックボックスモデル)が存在すると仮定して，ニュースデータの TF-IDF から成り立つ説明変数と内閣支持率・不支持率の上昇/下降を目的変数とし，AIME による逆問題を解くことにより，内閣支持率・不支持率の変動に貢献する要素(単語)を抽出する手法を示している点にある．AIME は実際に存在する AI，機械学習モデルをブラックボックスモデルとして説明を導出するものであるが，本研究では，1 ヶ月分のニュースデータから内閣支持率，不支持率の変動を予測する AI, 機会学習モデルがあると仮想的に想定し，閣支持率，不支持率の変動に貢献する要因(単語)を導出する新たな手法を示している．

- 上記のシステムの有効性を検証するために，2018 年 12 月 26 日から 2023 年 12 月 25 日の「安倍首相」,「菅首相」,「岸田首相」をその任期の期間と合わせてキーワードを設定し，NewsAPI [2]から集約し，61443 件のニュースを収集した．また，内閣支持率，不支持率のデータも入手し，これらのデータを統合し，ニュースデータは TF-IDF, 内閣支持率, 不支持率のデータは上昇を 1, 下降を-1，変化なしを 0 として月毎のデータとして，AIME を適用することにより，内閣支持，不支持の変動に関わる全体的な傾向(大局的特徴重要度)と特に変化が大きかった時の内閣支持，不支持の変動に関わる重要要因(局所的特徴重要度)を抽出することを実現し，その考察を行った．

本論文の構成を以下の通りである．第 2 節では関連研究について述べる．第 3 節では本方式について述べる．第 4 節では具体的なシステムを実装し，実験を行う．第 5 節でまとめる．

2. 関連研究

AIME[1]は，XAI の一種であり，複雑な AI や機械学習モデルをブラックボックスとして考え，その入出力データの結果から近似的な逆演算子を生成して説明を導く手法である．現在，AI や機械学習における説明可能性，透明性，信頼性を実現するために，XAI は多様な領域で研究されており，多くのサーベイ論文が発表されている（例えば，以下の論文[3]-[21]）．

Speith [18]は，XAI の種類として大きく post-hoc と ante-hoc と分類している．ante-hoc 手法には，線形回帰モデル，決定木，k-nearest neighbor モデル，ルールベースなど，一般に説明可能であると考えられてきたグラスボックスモデルの説明を対象としている．一方，post-hox 手法は一般的にブラックボックスモデルに適用され、model-specific 手法と model-agnostic 手法の 2 つのカテゴリに分類される．model-specific 手法は特定のブラックボックスモデルに限定され，例えば Grad-CAM [22] は畳み込みニューラルネットワークにのみ適用可能な特定モデルである．それに対し，model-agnostic 手法は LIME [23]，SHAP [24]，AIME [1]などの AI や機械学習モデルに適用可能であり，より汎用性が高いことが知られている．

さらに，我々は model-agnostic 手法を大きく 3 つに分けている．1 つ目は、ブラックボックスを使ってモデルの振る舞いを観察し、説明を導き出す方法である．例えば，partial dependence plots(PDP) [25][26]，individual conditional expectations (ICE) [27]などがある．2 つ目はブラックボックスとは異なる手法を用いて順方向で重要な特徴を抽出する方法である．例えば，LIME [23]，SHAP [24]がこれに相当する．3 つ目は，ブラックボックスの逆問題を解く手法であり，AIME[1]がこれに当てはまる．

AIME[1]は，ブラックボックスの逆問題を解くための近似逆作用素を説明変数Xとその説明変数を入力したときの推定結果$\hat{Y}$から導出する．しかしながら，訓練データとして説明変数Xと目的変数Yがあれば，AIME の枠組みで逆作用素を導出することが可能となり，ブラックボックスが存在しなくとも，要因を探ることが可能となる．これが AIME の大きな特徴と言える．

吉田[28]は，2001 年 4 月から 2004 年 9 月の期間における小泉内閣に対する世論調査の支持率に「TeX-Ray」によって生成されたデータセットを回帰分析することによって，内閣に対する社説の扱いと内閣の支持率との間に異常に強い系統的な関係が存在することを示した．島田[29]は，安倍晋三、菅義偉、岸田文雄の 3 人の総理大臣が新型コロナウイルスとの闘いの先頭に立って来た 2 年間（2020 年 1 月～2021 年 11 月）について，国民の政治意識の変化を俯瞰しているが、定性的分析にとどまっている．小野[30]は，日本の内閣支持に影響を与える要因を，Stimson の再帰的二項モデルを用いて，時事通信の世論調査データを用いて包括的に分析している．その結果，日経平均株価は政権能力支持ムード，GDP 成長率は漠然支持ムードに影響することが示されている．近年の日本の動向を追う研究も活発化しており，Thies ら[31]では，COVID-19 が自民党の支持層がどのように変化したかを観察している．また，内閣総理大臣の行動を分析したものとしては，Kato ら[32]は潜在ディリクレ配分（LDA）と特異スペクトル変換（SST）を組み合わせた新しいトピック変動検出方法を用いて、首相の行動パターンの変化点を検出，政治的ビジネスサイクルの理論と実証分析を進めている．

本方式は，AIME を用いて，内閣支持率，不支持率の変動に作用する要因を探索することに重きを置いている．

3. AIME を用いた内閣支持率変化要因抽出手法

本節では，具体的な手法について示す．3.1 節では本方式における AIME の概要について示す．3.2 節ではシステムの全体図について述べる．3.3 節では用いたデータをどのように構築したかを述べる．3.4 節では AIME を用いた分析手法について述べる．

3.2 節では，3.1 節の AIME の定式化を踏まえた本方式の実現方式を述べる．

3.1. AIME の概要

図 1 に，AIME の概要を示す．図 1 に示す AIME は本方式に特化した AIME の構成を示している．

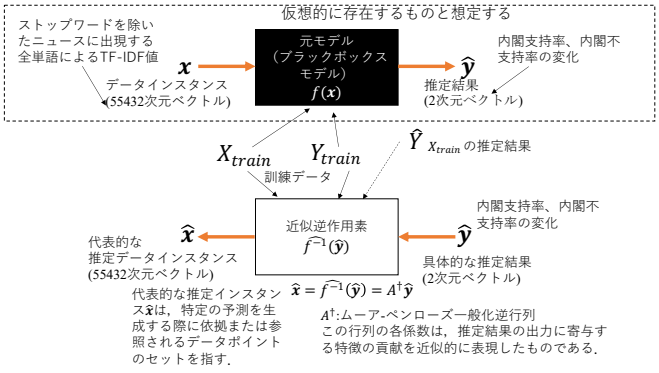


図 1: 本方式における AIME の概要

まずは，文献[1]に示す AIME について説明する．AIME は，複雑な AI，機械学習モデル，つまり，ブラックボックスモデルの大局的，局所的説明を導出する手法である．図 1 で言えば，通常ならば，上段の「元モデル(ブラックボックスモデル)」を指し，ニュース$\boldsymbol{x}$を入力すれば，内閣支持率，不支持率の変動$\hat{\boldsymbol{y}}$を導出する AI，機械学習モデルを構築する．入力のニュース$\boldsymbol{x}$は，ストップワードリストに入った言葉以外の全ニュースに使われる語(今回の場合は 55432 語)であり，その各語が TF-IDF で重み付けられた 55432 次元のベクトルである．出力$\hat{\boldsymbol{y}}$は，内閣支持率，不支持率について，それぞれ上昇は 1，下降は-1，変化なしは 0 で表される 2 次元のベクトルである．訓練に用いる入力側のニュースのデータをX_{train}，正解データ(つまり出力側のデータを) Y_{train}とおく．さらに，訓練データX_{train}を一つずつ入力し，推定された結果の集合を$\hat{Y}$とおく．通常の AIME を用いる際には，X_{train}，$\hat{Y}$を用いて近似逆作用素を用いることで，ブラックボックスモデルの振る舞いを逆問題で導出する近似逆作用素$A^{\dagger}$を導出する．この近似逆作用素$A^{\dagger}$は，ブラックボックスモデルの振る舞い，つまり大局的な特徴の重要性を表す．

しかしながら，今回は，内閣支持率，不支持率の変化を予測し，その推定根拠を求めるのではなく，内閣支持率変化要因を抽出することが目的である．そのことから，図 1 上段のブラックボックスは仮想的にあるものと想定して考える。そうなると，ブラックボックスから導出される$\hat{Y}$が導出されなくなるが，全く問題がない．近似逆作用素$A^{\dagger}$を導出するのに，X_{train}，Y_{train}を用いればよい．これにより，実際に内閣支持率，府支持率に貢献する説明を導出する近似逆作用素$A^{\dagger}$を計算することが可能となる．先行研究[1]により，近似逆作用素$A^{\dagger}$は，下記の式で導出される．

$$\begin{aligned} X &= A^{\dagger}Y, \\ XY^T &= A^{\dagger}YY^T, \\ XY^T(YY^T)^{-1} &= A^{\dagger}(YY^T)(YY^T)^{-1}, \\ A^{\dagger} &= XY^T(YY^T)^{-1} = XY^{\dagger}. \end{aligned} \tag{1}$$

ここで，上記の理由から，先行研究[1]の式と比べて，$\hat{Y}$からYに変わっていることに注意したい．また，X_{train}，Y_{train}は複雑になるため，式(1)以降では，X,Yに書き換えていることに注意したい．ここで，Y^Tは行列Yの転置行列を示す．$Y^{\dagger}$はYのムーアペンローズの一般化逆行列[33][34]を示す．

以上より，ニュース記事集合からそれぞれのニュースを単語に分解し，TF-IDF によって重みづけされた行列化されたXと内閣支持率，不支持率の変動を表す行列Yによって，内閣支持率，不支持率の変動の要因を導出するための近似逆作用素$A^{\dagger}$を導出することが可能となる．

3.2. システム全体像

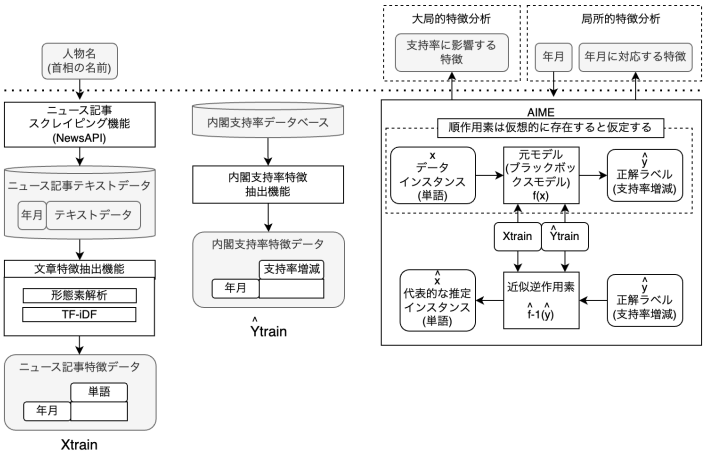


図 2 ：システム全体図

図 2 にシステム全体図を示す.図 2 の左はデータ構築，右側は AIME のシステムを表している．

3.3. データ構築

この節ではデータをどのようについて構築したのかを示す．3.3.1 節ではニュース記事スクレイピング機

能について説明する. 3.3.2 節ではニュース記事から文章の特徴（単語）を抽出する機能について説明する. 3.3.3 節では内閣支持率の特徴データ抽出機能について説明する.

3.3.1. ニュース記事スクレイピング機能

ニュース記事は NewsAPI [2]を用いて取得した. 内閣支持率が 1 ヶ月ごとに発表されることに合わせてニュース記事を 1 ヶ月ごとにまとめて取得する. 取得するニュースは指定するキーワードを含むニュースとする.本研究では「岸田首相」,「菅首相」,「安倍首相」というキーワードのどれかが含まれるニュースとする.

3.3.2. 文章特徴抽出機能

1 ヶ月分のニュースを一つの文書とみなし、形態素解析を行って Bag-of-Words を生成した後、TF-IDF を用いて単語の重み付けを行った.

内閣支持率は月毎に発表される. NewsAPI [2]の限界から, 2018 年 12 月 26 日から 2023 年 12 月 25 日の「安倍首相」,「菅首相」,「岸田首相」をその任期の期間と合わせてキーワードとして設定し，ニュース記事を収集し，月毎に.” “(スペース)で繋げ，まとめた. 月毎にまとめたニュースを TF-IDF で重みづけを行うことにより 59(月)×55,432(語)からなる行列からなる行列Xとなる.

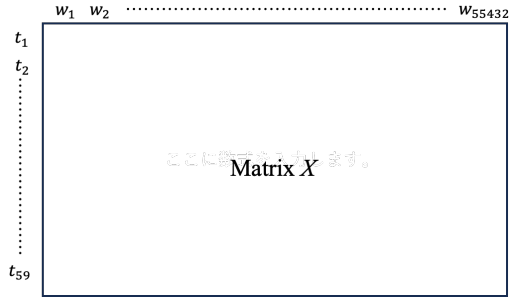


図 3: 本方式における行列X

3.3.3. 内閣支持率,不支持率特徴抽出機能

内閣支持率は NHK 選挙 WEB の「NHK 世論調査 内閣支持率」の Web サイト[35]から抽出した. ここでは支持率および不支持率がそれぞれ前月から上昇しているか下降しているか，あるいは変化していないかの 3 段階として，特徴量を表現する. それぞれの月の支持率,不支持率と 1 ヶ月前の支持率,不支持率との差を計算し，それが前月から見て上昇している場合は 1，下降している場合は-1，変化なしは 0 とする. これによって，1 ヶ月につき 2 次元のベクトルを構成する.

それらを合わせると 59(月)×2(特徴量)からなる行列Yとなる. この行列Yを図 4 に示す.

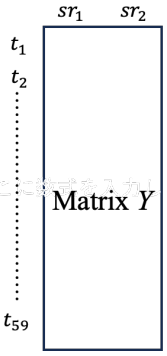


図 4: 本方式における行列Y

3.4. AIME を用いた分析手法

3.3 節で構築された行列Xと行列Yに対して式(1)を適用することで近似逆行列$A^{\dagger}$を導出することができる. この近似逆行列$A^{\dagger}$を使って、内閣支持率，不支持率の変動に貢献する要因(単語)を抽出する大局的特徴重要度，ある月のニュースによってなぜ内閣支持率，不支持率が上昇/下降したかを導出する局所的特徴重要度を求めることが可能となる.

3.4.1. AIME による大局的特徴重要度分析

近似逆行列$A^{\dagger}$は，行数と列数がそれぞれデータセットXの特徴数(55,432)と目的変数の数(2)に等しい行列である. この一近似逆行列$A^{\dagger}$は、その成分が各特徴の特定の目的変数への影響を反映するため，大域的特徴重要度を示す。具体的に、近似逆行列$A^{\dagger}$ は，各目的変数に対する各特徴の寄与度を示す. $A^{\dagger}$ の特定の成分（例えば、行 i の列 j の成分）は，j 番目の目的変数の出力に対する i 番目の特徴の寄与を表す. これにより、各特徴がモデル(この場合は，内閣支持率，不支持率の変動)の全体的な動作にどの程度影響するかを理解することができる.

3.4.2 AIME による局所的特徴重要度分析

AIME における局所特徴重要度は，各特徴が特定のデータインスタンスの分類にどの程度寄与しているかを導出する. これは，データインスタンスがモデルによって特定のクラスに分類される理由を理解するのに役立ち，特定のインスタンスの分類に直接貢献する特徴(単語)を明らかにすることができる.

局所特徴重要度の値は，特定のデータインスタンスにおいて，内閣支持率の変化，不支持率の変化という目的変数に対して，各特徴量(ここでは単語)が影響している度合いを示している.これによって，内閣支持率，不支持率の変化について重要であると見なす特徴を特定し，理解を深めることができる.

データインスタンス$\boldsymbol{x}$, 内閣支持率，不支持率の変化を表すベクトル$\boldsymbol{y}$, 近似逆行列$A^{\dagger}$を組み込むことで，以下の式を使用して局所特徴重要度ベクトル$\boldsymbol{l}$を導出す

ることができる：

$$\boldsymbol{l} = A^{\dagger}\boldsymbol{y}^{*} \circ \boldsymbol{x},$$

ここで，局所特徴重要度ベクトル $\boldsymbol{l}$は，全特徴の次元数(55,432)で構成され，各値はその特徴(単語)の重要度を表し，∘はアダマール積を表す．この絶対値が高い順にソートすることにより，どのような特徴(単語)がその月の内閣支持率，不支持率の変化に貢献する要因(単語)として説明を導出することが可能である．

4. 実験

本節では，3 節で示した我々の手法を実装し，結果を出力することにより有効性の検証を行った．4.1 節では，実験環境を述べる．4.2 節では，実験 1 として内閣支持率，不支持率の上昇/下降と全体で寄与する特徴(単語)を抽出する大局的特徴重要度の結果について示す．4.3 節では，実験 2 として内閣支持率が上昇し，内閣不支持率も上昇した 2021 年 2 月，内閣支持率が上昇し，内閣不支持率が下降した 2020 年 4 月，内閣支持率が下降し，内閣不支持率が下降した 2019 年 8 月局所的特徴重要度を示した．4.4 節ではこれらについて考察を行った．

4.1. 実験環境

本実験環境構築，および実行環境は Google Coloaboratory である．ニュース記事の収集には NewsAPI[2](有料版)を用いた．この API は 5 年間が限界のため，5 年分 2018 年 12 月 26 日から 2023 年 12 月 25 日の「安倍首相」,「菅首相」,「岸田首相」をその任期の期間と合わせてキーワードを設定し， 61443 件のニュースを収集した．その間 2019 年 10 月は台風の影響で行われなかったが，毎月内閣支持率，不支持率調査が行われており，その月に該当するニュースを再整理した上で，MeCab と neologd 辞書によって，単語に分割した上で，名詞，動詞，形容詞を取り出し，さらに原形に直した単語を取得し，Scikit-learn 1.2.2.を用いて tf-idf を用いて行列Xを作成した．AIME は，Numpy 1.23.5，Pandas 1.5.3，Scikit-learn 1.2.2 を用いて実装した．可視化は Seaborn 0.12.2，matplotlib 3.7.1 を使用した．

ちなみに，仮想的に想定すればよいので，本方式とは全く関係ないが，訓練データとテストデータをテストデータを 20%として LightGBM でニュース記事から内閣支持率，不支持率の変化の予測をしたところ，支持率は 0.444、不支持率 0.667 の精度であった．なお LightGBM は，lightgbm 4.1.0 を用いて実装を行った．LightGBM を選んだ理由も特になく，ブラックボックスモデルとなればそれでよく，かつ本方式ではブラックボックスの存在は仮想的に設定すればよいため，この事実については，参考までに 1 ヶ月で集めたニュース記事では判別が難しいことを示しただけあり，精度についても本研究の手法は全く関係ない．AIME の近似逆作用素$A^{\dagger}$を導出する際は，集約したデータ全てを用いた．以下は，3 節で示した手法を実装したものの結果を示す．

4.2. 実験 1: AIME による大局的特徴重要度

図 4 は，内閣支持率の変動にも内閣不支持率の変動にも貢献した単語をソートし，重みの高い順に並べた大局的特徴重要度を示す．COVID-19 の影響があるため，zoom などの単語が上位に出ている．無駄遣いについては支持，不支持について両方とも正に相関があることが意外である．ストップワードを決めて，整理すると詳しい方向性が出てくることも考えられるが，政治的な事象よりも一般的な事象が内閣支持率に相関しているようにみえる．

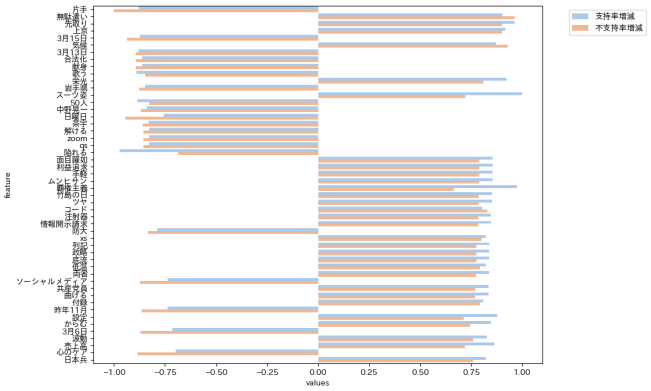


図 4：内閣支持率の変動にも内閣不支持率の変動にも貢献した単語の大局的特徴重要度

次に，図 5 に内閣支持率の変動に貢献した単語の大局的特徴重要度上位 50 件を示す．上位 3 件の語を考察すると,「スーツ姿」は安倍首相がスーツ姿で約 5 年半ぶりに福島第 1 原子力発電所を訪れ，初めて防護服やマスクをつけずに視察をしたことがニュースに出現しているからである.「贈る」は安倍首相がぼたんの栽培が盛んな松江市の市長らと面会し，ぼたんの花束を贈られ「安倍政権も大輪の花を咲かせたい」と語ったことが効いている.「覇権主義」は，菅首相が日米で覇権主義を強める中国に日米が連携して対処するニュースを指している．このように，政治的な動きだけでなく，スーツ姿で福島第一原発といった具合に首相の姿勢なども貢献していることが分かる．例えば,「日米関係」,「気候」,「売上高」などの内閣支持率に関係しそうな単語が上位に上がってきていることがわかる．そのほかにも,「ムンヒサン」当時韓国の議長の天皇陛下に元慰安婦への謝罪を求めた発言や,「竹島の日」(韓国の大統領選の候補が竹島の日本領の表記を変えない限り韓国の五輪ボイコットもあり得ると発言したこ

と)など，日韓問題についても支持率に関連していることが分かる．

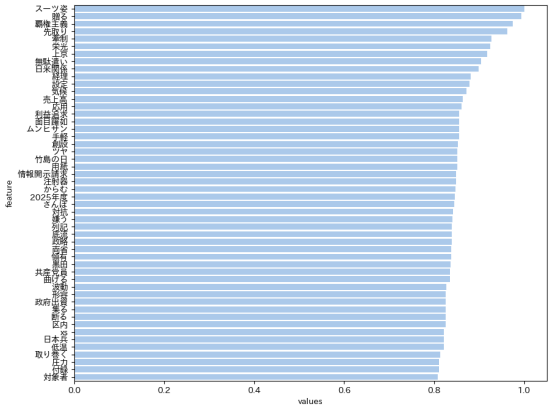


図 5：内閣支持率の変動に貢献した単語の大局的特徴重要度上位 50 件

次に，図 6 に内閣支持率の変動に貢献した単語の大局的特徴重要度上位 50 件を示す．「無駄遣い」は実際のニュース記事では，「無駄遣い排除」という文脈で使われている．「気候」など支持率と同様の単語が上位にきている．「ムンヒサン」(当時韓国の議長)の発言や，「竹島の日」など，日韓問題についても不支持率に関連している．「プライマリーバランス」は財政健全化の指標として赤字を強調するニュースがあるからである．「経団連会長」は様々な政府への意見をしており，それが貢献したと考える．「日本兵」は安倍首相が第二次世界大戦で多くの元日本兵が死亡した「インパール作戦」の舞台になった北東部マニプール州インパールを訪れる方向で調整に入ったニュースによるもの．これらが不支持に寄与しているのがなぜであるのかは専門家の意見を聞く必要があるだろう．これについては今後の課題である．

これらのことから，ノイズがあるものの全体としての内閣支持率，不支持率に貢献する単語が導出できていることが分かる．特に抽出された特徴(単語)のみでは，なぜ内閣支持率，不支持率に貢献することが分からない場合でも，その語が出現するニュースをもう一度検索し，確認することで，内閣支持率，不支持率に貢献することについて納得できるものを多く導出できていると考える．

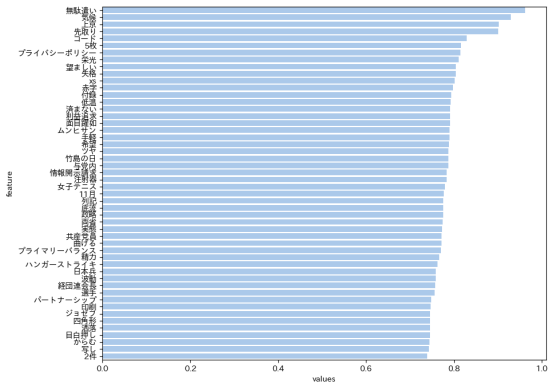


図 6：内閣不支持率の変動に貢献した単語の大局的特徴重要度上位 50 件

しかしながら，単語によってはどのような関係があるのか，もしくはないのかを専門家を入れてより深く検証する必要があると考える．それについては，今後の課題である．

4.3. 実験 2: AIME による局所的特徴重要度

次に，AIME による局所的特徴重要度の有効性を示すため，分かりやすい例として，内閣支持率が上昇し，内閣不支持率も上昇した 2021 年 2 月，内閣支持率が上昇し，内閣不支持率が下降した 2020 年 4 月，内閣支持率が下降し，内閣不支持率が下降した 2019 年 8 月局所的特徴重要度を示す．

まず，内閣支持率が上昇し，内閣不支持率も上昇した 2021 年 2 月の AIME による局所的特徴重要度として図 7 に示す．

「渦中」については，菅首相長男接待問題や総務省官僚接待疑惑の渦中を指している．「後藤謙次」はジャーナリストであり「菅政権失敗の本質」というコラムを執筆している．「COCOA」は，新型コロナウイルス接触確認アプリ「COCOA」で発生した大規模な不具合についてのニュースが関連している．「音声」は総務省接待について発覚した音声データのニュースに行き着くことができる．また，「入院者」については，新型コロナウイルスについて，「対策続け入院者や重症者減少を」と呼びかけているニュースに行き着くことができる．「パラリンピック」開催へ動き出すことについても大きくニュースになった時期である．これらの特徴(単語)からニュースを再度追いかけることによって，この時期は，菅内閣の手腕が評価される部分と，様々な疑惑が露出したり，新型コロナウイルス接触確認アプリ「COCOA」の不具合で不支持率を上げていることが分かる．

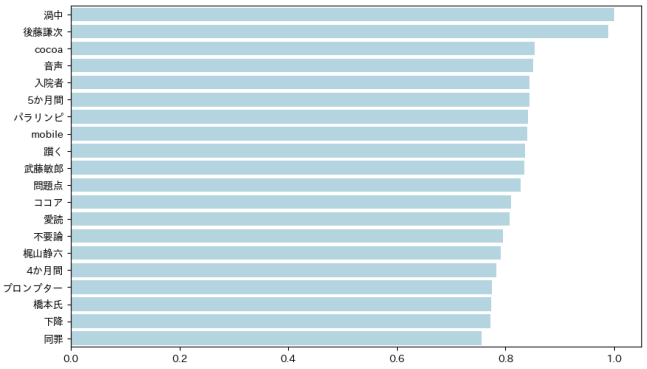


図 7: 2021 年 2 月の内閣支持率，不支持率に貢献したとされる局所的特徴重要度

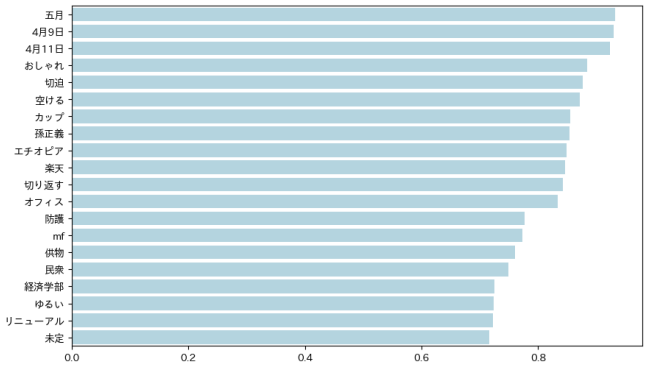


図 8: 2020 年 4 月の内閣支持率，不支持率に貢献したとされる局所的特徴重要度

次に，内閣支持率が上昇し，内閣不支持率が下降した 2020 年 4 月の AIME による局所的特徴重要度として図 8 に示す．

この時期の「五月」というのはゴールデンウイークと新型コロナウイルスの流行を予測したニュースが出ている．「4 月 9 日」,「4 月 11 日」は，新型コロナウイルス感染症の拡大に伴う緊急事態宣言が発令され，混乱の様子のニュースが報じられており，特に，4 月 11 日には安倍晋三首相が新型コロナウイルス感染症対策本部での議論を踏まえ､「オフィスでの仕事は原則として，自宅で行えるようにする．どうしても出勤が必要な場合でも、出勤者を最低 7 割は減らす」ように要請していると報じている．「おしゃれ」は意外かもしれないが，小池百合子都知事のマスクについて報じたものである．「切迫」は医療機関の切迫を表している．「空ける」は密を避けるため，参院本会議では前後左右を 1 席分空けることを決めたニュースである．「カップ」は，安倍首相が，星野源（39）が歌う「うちで踊ろう」とともに，自宅ソファで愛犬のミニチュアダックスフントのロイを抱く姿や、カップ片手にくつろぐ様子を公開したニュースのことを示している．「孫正義」は，世界のベンチャー企業に 8 兆 5000 億円を投資した「ソフトバンク・ビジョン・ファンド」（SVF）が 1 兆 8000 億円の投資損失を計上したニュースであり，新型コロナウイルス感染症が広がる中で，経済界から不安になる要素である．

これらのキーワードから直接的な単語ではないが，内閣支持率が上昇し，内閣不支持率が下降した理由を示すニュースを検索するために重要な特徴(単語)が取得できていることが分かる．新型コロナウイルス感染症の流行による混乱を様々な形で対策を行い，また，国民を安心させる政策が支持された時期であったことが分かってくる．

最後に，内閣支持率が下降し，内閣不支持率が下降した 2019 年 8 月局所的特徴重要度の AIME による局所的特徴重要度として図 9 に示す．

「墓前」については，安倍首相が父・晋太郎元外相の墓参りを「改憲議論進める時を迎えた」と語ったことを示している．「未来のために」は,「小泉進次郎よ、日本の未来のために君は育児休暇を取らないといけない」という More Access! More Fun (https://www.landerblue.co.jp/46661/) というブログの主張がニュースで話題になったためである．確かに小泉進次郎議員の育児休暇取得に関しては賛否両論が当時湧き上がっていた．「大枠」は，日米貿易交渉が大枠で合意したというニュースがこの頃多数みられる．「出資」は，このころのニュースで，日本企業が出資する韓国の化粧品会社が，会長が社員に安倍晋三首相を称賛する動画をみせたことを巡り，謝罪声明を発表したというものを指す．「孤児」は第二次世界大戦の戦争孤児についてニュース記事から導かれたものであり，毎年 8 月にこのようなニュースは流れる．その他,「河野談話」の撤回など日韓問題が大きくなったときである．また,「和平合意」は，安倍首相がエチオピアのアビー首相と会談し,民主化に向けた支援の一環として，日本が行う無償資金協力を通じて、来年予定されている総選挙で投票箱を提供するなどの協力を行う考えを伝えたニュースを「和平合意」を踏まえてと伝えている．

これらのことから，出力結果は一見政治と関係のない単語が出現しているが，その単語で調べると，確かに内閣支持，不支持につながる出来事を示すニュースにたどり着くことがわかった．AIME の大局的特徴重要度，局所的特徴重要度は内閣支持率，不支持率に関連する特徴(単語)を導出しているが，ここから実際のニュース記事をもう一度検索を行うことにより，実際の当時の様子を振り返ることが可能となる．これらの結果によって,AIME を用いた本手法によって,閣支持,

不支持につながる特徴を抽出する手法の有効性を検証することができた．

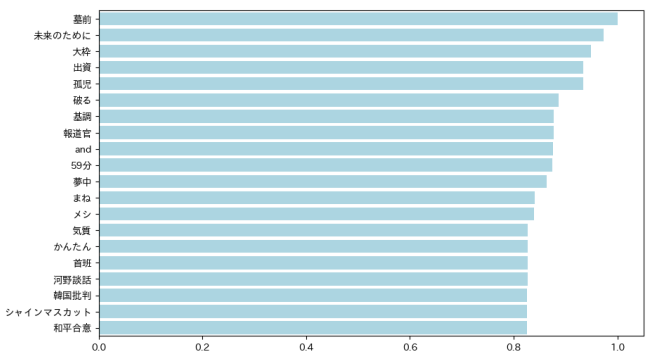


図 9: 2019 年 8 月の内閣支持率，不支持率に貢献したとされる局所的特徴重要度

4.4. 考察

本節の実験から、AIME の大局的特徴重要度によって内閣支持率，不支持率の変動の全体的な動作にどの程度影響するかを理解することが可能であり,「スーツ姿」や「日米関係」,「竹島の日」など内閣支持率・不支持率全体に影響する単語が抽出できた．さらに，AIME における局所特徴重要度によって，具体的な年月の内閣支持率, 不支持率の変動の理由を示す特徴(単語)を抽出が可能であることを示した．

課題として，今回は，データセットとしてニュースデータを用いたため，図 6 の「女子テニス」など元のニュースデータを見ると直接総理の施策や行動とは関係ない，周辺の事柄も含まれていたため，確かに内閣支持率，不支持率の変動に何らかの影響があった想像できるものの直接の影響とは言えないものも抽出されたことは否めない．これは AIME の手法の問題というより，NewsAPI[2]で収集したニュースであることに問題があると考える．NewsAPI の性質から報道的な記事からブログ等の記事をまとめて収集してくる特性がある．例えばデータセットを首相動静をキーワードとしたものに限り，ブログ等の記事を排除することで，より首相の施策や行動を根拠として導出することが可能であると考える．

5. おわりに

内閣支持/不支持率の変動の要因を推定することは，現在の大衆の考え方を明確にする上で非常に重要である．

本稿では，ニュース記事から内閣支持率を予測する機械学習モデル(ブラックボックスモデル)が存在すると仮定して，ニュースデータの TF-IDF から成り立つ説明変数と内閣支持率・不支持率の上昇/下降を目的変数とし，AIME [1]による逆問題を解くことにより，内閣支持率・不支持率の変動に貢献する要素(単語)を抽出する手法を実現した．

AIME [1]は，AI，機械学習モデルなどのブラックボックスモデルの説明のために用いる手法であるが，AIME の逆演算により，そのようなニュースデータから内閣支持率・不支持率の変動を予測するブラックボックスモデルがあることを想定し，その要因を探ることが可能となった．これは，AIME を XAI の一手法として捉えるだけでなく，新たなテキストマイニング手法の一つとして，適用することが可能と考えられる．

今後の課題として，首相動静など他のデータセットを使った本方式の検討，より細かい時系列テキストデータを用いた AIME を用いた分析手法の実現，大局的/局所的重要特徴に関連する元データを集約し，要約することで，より分かりやすい説明を導出する機能の実現が挙げられる．

参 考 文 献

[1] T. Nakanishi, "Approximate Inverse Model Explanations (AIME): Unveiling Local and Global Insights in Machine Learning Models," in IEEE Access, vol. 11, pp. 101020-101044, 2023, doi: 10.1109/ACCESS.2023.3314336.

[2] NewsAPI: https://newsapi.org/.

[3] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, "A survey on XAI and natural language explanations," Inf. Proc. Manage., vol. 60, no. 1, 103111, 2023, doi: 10.1016/j.ipm.2022.103111.

[4] W. Saeed, C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowl-Based Syst., vol. 263, 110273, 2023, doi: 10.1016/j.knosys.2023.110273.

[5] Y. N. Chuang, G. Wang, F. Yang, et al., "Efficient XAI techniques: A taxonomic survey," arXiv preprint arXiv:2302.03225, 2023.

[6] G. Schwalbe, B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," Data. Min. Knowl. Disc., 2023, [Online]. Available: https://doi.org/10.1007/s10618-022-00867-8doi: 10.1007/s10618-022-00867-8.

[7] R. Dazeley, P. Vamplew, F. Cruz, "Explainable reinforcement learning for broad-XAI: a conceptual framework and survey," Neural Comput. Applic., vol. 35, pp. 16893-16916, 2023, doi: 10.1007/s00521-023-08423-1.

[8] W. Yang, Y. Wei, H. Wei, et al., "Survey on explainable AI: From approaches, limitations and applications aspects," Hum-Cent. Intell. Syst., vol. 3, pp. 161-188, 2023, doi: 10.1007/s44230-023-00038-y.

[9] S. Ali, T. Abuhmed, S. El-Sappagh, et al., "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," Inf. Fusion, vol. 99, 101805, 2023, doi: 10.1016/j.inffus.2023.101805.

[10] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," IEEE Access, vol. 11, pp. 78994-79015, 2023, doi: 10.1109/ACCESS.2023.3294569.

[11] R. Dwivedi, D. Dave, H. Naik, et al., "Explainable AI (XAI): core ideas, techniques, and solutions," ACM Comput. Surv., vol. 55, no. 9, Article 194, 2023, doi: 10.1145/3561048.

[12] C. Molnar, "Interpretable machine learning," Lulu Com, 2020.

[13] D. Minh, et al., "Explainable artificial intelligence: A comprehensive review," Artif. Intell. Rev., vol. 55, no. 5, pp. 3503-3568, 2022, doi: 10.1007/s10462-021-10088-y.

[14] A. Adadi, M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[15] Q. Zhang, et al., "Interpretable convolutional neural networks," in: Proceedings of the 2018 Conference, pp. 8827-8836, 2018.

[16] R. Guidotti, et al., "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, pp. 1-42, 2019, doi: 10.1145/3236009.

[17] D. V. Carvalho, et al., "Machine learning interpretability: A survey on methods and metrics," Electronics, vol. 8, no. 8, Article 832, 2019, doi: doi: 10.3390/electronics8080832.

[18] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in: Association for Computing Machinery, 2022; New York, NY, USA, pp. 2239-2250.

[19] W. Samek, K. R. Müller, "Towards explainable artificial intelligence," in: Springer, 2019, pp. 5-22.

[20] M. V. Lent, W. Fisher, M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in: AAAI Press; Cambridge, MA, USA: MIT Press, 2004; San Jose, CA, USA, pp. 900-907.

[21] D. Gunning, D. Aha, "DARPA's explainable artificial intelligence (XAI) program," AI Mag., vol. 40, no. 2, pp. 44-58, 2019.

[22] R. R. Selvaraju, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in: Proceedings of the 2017 Conference, pp. 618-626.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[24] S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems 30, 2017.

[25] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.

[26] Q. Zhao, T. Hastie, "Causal interpretations of black-box models," J. Bus. Econ. Stat., vol. 39, no. 1, pp. 272-281, 2017.

[27] A. Goldstein, et al., "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," J. Computat. Graph. Stat., vol. 24, no. 1, pp. 44-65, 2015, doi: 10.1080/10618600.2014.907095.

[28] 吉田文彦, "文章解析ソフトウェア TeX-Ray の概要と応用事例: 小泉内閣関連新聞社説の内容分析結果による内閣支持率の予測," マス・コミュニケーション研究, vol. 68, pp. 80-96, 2006.

[29] 島田敏男, "コロナ禍と政治意識の揺れ 安倍・菅・岸田内閣の支持率の変化を軸に," 放送研究と調査, vol. 72, no. 1, pp. 24-34, 2022.

[30] 小野弾, "内閣支持のムードと経済的変動要因—内閣支持理由データを用いた長期的分析," 年報政治学, vol. 72, no. 2, pp. 2_209-2_233, 2021.

[31] M. F. Thies, Y. Yanai, "Did COVID-19 Impact Japan's 2021 General Election?," in R. J. Pekkanen, S. R. Reed, D. M. Smith (eds), Japan Decides 2021, Palgrave Macmillan, Cham, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-11324-6_15

[32] S. Kato, et al., "Time-series topic analysis using singular spectrum transformation for detecting political business cycles," Journal of Cloud Computing, vol. 10, no. 1, 21, 2021.

[33] E. H. Moore, "On the reciprocal of the general algebraic matrix," Bull. Am. Math. Soc., vol. 26, pp. 294–300, 1920.

[34] R. Penrose, "A generalized inverse for matrices" in Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge: Cambridge University Press, vol. 51, no. 3, 1955, pp. 406-413. DOI: 10.1017/S0305004100030401.

[35] NHK' NHK 世論調査 内閣支持率 政党支持率 毎月の最新情報 | NHK 選挙 WEB' https://www.nhk.or.jp/senkyo/shijiritsu/ .(参照 2023-12-27)

2次元最適相関ルールを用いたSHAPの解釈補完

塩谷　泰平†　　森本　康彦†　　亀井　清華†

† 広島大学情報科学部情報科学科　〒 739—8511 広島県東広島市鏡山 1 丁目 3-2
E-mail: †{b194032,morimo,s10kamei}@hiroshima-u.ac.jp

あらまし　人工知能（AI）の能力が向上する一方で，そのモデルの複雑さ，判断根拠の不透明さが問題視され，現在までに，人工知能モデルの解釈性を向上させる様々な手法（XAI）が提案されている．しかし．既存の手法では 2 つの特徴量にまたがる非線形の相関を適切に説明することが難しい．本論文では，モデルの出力と 2 つの特徴量の関係性を，2 次元最適相関ルールと呼ばれるデータマイニング手法で捕捉し，それを用いて特徴量間の非線形の相関を説明する手法を提案した．また，提案手法では，特徴量の膨大な組み合わせのうち，注目に値するもののみを選別し，説明として利用する．これにより，既存の手法では見逃されてしまいがちな特徴量間の非線形の関係性を理解することに貢献できる．

キーワード　XAI，データマイニング

1　序　　論

AI の開発や研究が進んだことで，金融や医療をはじめとする多くの分野で AI が活用されるようになり，AI の担う社会的な責任は大きくなっている．特に，AI の出力が社会に与える影響が大きな分野では AI の判断ミスが許されない．例えば，AI を用いた自動運転システムや医療診断システムでは AI の判断ミスが直接的に事故につながり，人の命が奪われる可能性がある．そのため，AI の信頼性の向上が求められている．しかし，ニューラルネットワークなどの複雑なモデルの増加に伴い，AI の判断根拠を人間が理解することは難しくなっている．このように，より複雑で精度の高い AI の需要が高まっている一方で AI の判断根拠の不透明性が問題視されている [1].

このような AI の判断の不透明性の改善に貢献する技術として，「説明可能な AI(eXplainable AI, XAI)」と呼ばれる技術が注目されている [1]．主要な XAI 技術の 1 つに SHapley Additive exPlanations (SHAP) [2] が挙げられる．SHAP によるモデルの出力の解釈では各予測に対する特徴量の貢献度を算出することができ，モデルがどの特徴量をどの程度根拠にして判断を下したかが分かる．

しかし，従来の XAI 手法では，2 つの特徴量を跨いだ非線形の相関を適切に説明することが難しい．たとえば，SHAP を用いた解釈では，各特徴量間の相関を考慮していない．その結果，特徴量間の相関がモデルの出力に及ぼす影響を説明することはできない．具体的に，血糖値と BMI，糖尿病発症率の間に正の相関関係があり，モデルは血糖値や BMI が高い患者ほど糖尿病の可能性が高いと判断する傾向がある場合を考える．ただし，BMI が低い場合は血糖値が高くても糖尿病の可能性は低いとモデルは判断するとする．この場合，SHAP による解釈では，血糖値や BMI が高い患者ほど糖尿病の可能性が高いという解釈は可能だが，患者の血糖値がどの程度を超えると糖尿病の可能性が高いとモデルが判断しているかを，血糖値だけでなく BMI も考慮して解釈することができない．

本論文では，既存の XAI 手法のそのような問題を軽減するため，2 次元最適相関ルール [3] を用いて SHAP による解釈を補完する手法を提案する．具体的には，多数ある 2 つの特徴量の組み合わせ (特徴ペア) からモデルの出力の根拠になっている特徴ペアを特定し，それらの特徴ペアに対して 2 次元最適相関ルールを適用することで，特徴量間の相関を考慮した解釈を行う．そうして得られた結果を SHAP の結果と共に解釈することで，特徴量間の相関がモデルの出力に与える影響を解釈することができ，SHAP による解釈を特徴量の相関を考慮した解釈で補完することが可能となる．

実験では，まず，糖尿病診断データセットに含まれるデータに対して 2 次元最適相関ルールを用いることで数値変数間の相関がモデルの出力に与える影響を解釈できることを示した．次に，上記のデータセットに含まれるインスタンスに SHAP と提案手法を適用して説明を生成し，少ない特徴ペアを用いて相関を考慮した説明ができることを確認した．

本研究の主要な貢献は以下の 3 点である．

- 2 次元最適相関ルールを用いて数値変数間の相関がモデルの出力に与える影響を解釈できることを確認した．
- 2 次元最適相関ルールを用いた SHAP の解釈補完手法を提案した．
- 提案手法によって得られた説明が相関を考慮した説明に有効であることを確認した．

2　XAI 手法によるモデル解釈

本節では，SHAP と LIME [4] を例に挙げ，既存の XAI 手法によるモデル解釈の問題点を説明する．

2.1　SHAP による解釈

SHAP は協力ゲーム理論に基づいた手法である．SHAP で

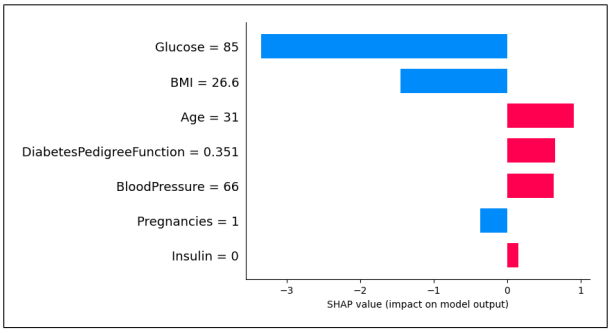


図 1: SHAP による予測根拠の解釈の例

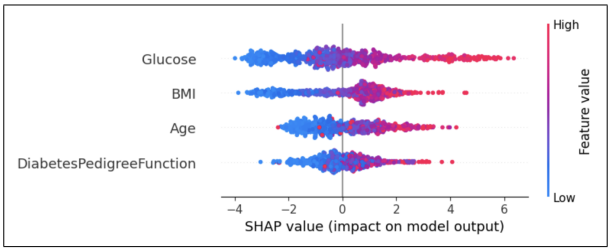


図 2: 特徴量の値と SHAP 値の関係

は各特徴が予測に対してどれだけ寄与しているかを計算することで，各インスタンスにおけるモデルの予測に対する特徴量の影響を明らかにする [2]．図 1 は 4 節で説明する，糖尿病診断データベースに対して CatBoost [5] を用いて構築した勾配ブースティング決定木 [6] の，ある正しい予測に SHAP を適用した例である．図中の赤い棒グラフはモデルの予測に各特徴量がどの程度正の方向に影響しているかを表し，一方で青い棒グラフは各特徴量が予測にどの程度負の方向に影響しているかを表している．この図からは，患者の血糖値 (Glucose) が 85mg/dL であることや BMI が 26.6 であることなどを根拠にモデルがこの患者は糖尿病ではないと予測したと SHAP が解釈したことが分かる．さらに，図 2 のようにデータセット全体において特徴量の値に対する SHAP 値をプロットすることで，特徴量の変化に対して SHAP 値がどう変化するかを確認することもできる．図 2 からは，血糖値や BMI，年齢などの特徴量は高い値をとるほど糖尿病であるという予測に寄与していることが分かる．

しかし，SHAP を用いてモデルの判断根拠と各特徴量の貢献度を解釈できる一方で，図 2 のように各特徴量と SHAP 値の関係を可視化する手法では，特徴量同士が強く相関している場合にモデルの振る舞いを解釈しきれない場合がある．

例えば，図 2 で示された解釈を用いると，血糖値が高い患者ほど血糖値がモデルの出力に大きく寄与していることが分かるため，「このモデルは血糖値が高い患者ほど糖尿病である可能性が高いと判断している」と解釈できる．しかし，図 3(a) と図 3(b) を見比べると，これらのインスタンスの血糖値はそれぞれ 148, 145 と共に高いにも関わらず，モデルは図 3(a) のインスタンスのみが糖尿病であると判断している．これは，血糖値の高い患者が糖尿病であるとは限らないことを示している．

ところが，図 3(a) と図 3(b) を比べただけでは，図 3(a) と図 3(b) の両方において BMI の貢献度が血糖値の貢献度よりも大きいため，このモデルでは BMI が血糖値よりも糖尿病の判断に大きく影響しているだけである可能性が否定できない．一方で，図 3(c) を見ると，BMI が高いインスタンスにも関わらず，BMI の寄与度が血糖値の寄与度よりも小さく，最終的なモデルの出力ではこの患者が糖尿病ではないと判断していることから，モデルが血糖値よりも BMI を糖尿病の判断に大きく用いているとは言えないことが分かる．以上より，図 3(a)，図 3(b)，図 3(c) を見比べてもモデルが血糖値と BMI の関係性をどのように糖尿病の判断に用いているかを解釈することはできないことが分かる．

このように，SHAP による解釈では各特徴量の値がモデルの出力に与える影響の度合いを解釈することはできても，特徴量間の相関がモデルの最終的な出力に与える影響を解釈することができないことがあり，SHAP による解釈が不十分であると言える．

2.2 LIME による解釈

LIME は，モデルの出力を局所的に解釈しやすいモデルで近似することで元のモデルの出力を解釈する手法である [4]．図 4 は 2.1 目と同様のモデルの，図 1 と同様のインスタンスに対して LIME を適用した結果である．図では寄与度の大きい上位 5 つの特徴量を表示している．図中のオレンジ色の特徴量はモデルの出力に正の影響を与えていることを表し，青色の特徴量はモデルの出力に負の影響を与えていることを表している．この図からは，SHAP と同様に，患者の血糖値 (Glucose) が 85mg/dL であることや BMI が 26.6 であることなどを根拠にモデルがこの患者は糖尿病ではないと予測したと LIME が解釈したことが分かる．

ところが，SHAP と同様に，LIME による解釈では各特徴量の値がモデルの出力に与える影響の度合いを解釈することはできても，特徴量間の相関がモデルの最終的な出力に与える影響を解釈することができないことがある．図 5 はそれぞれ図 3 と同様のインスタンスに対して LIME を適用した結果である．図 5 の解釈をそれぞれ見比べても，やはり SHAP による解釈と同様に，血糖値と BMI が高いほど糖尿病であるとモデルが判断するという傾向は解釈できるが，血糖値と BMI の値の組み合わせがどのようにモデルの最終的な出力に影響しているかを解釈することはできないことが分かる．

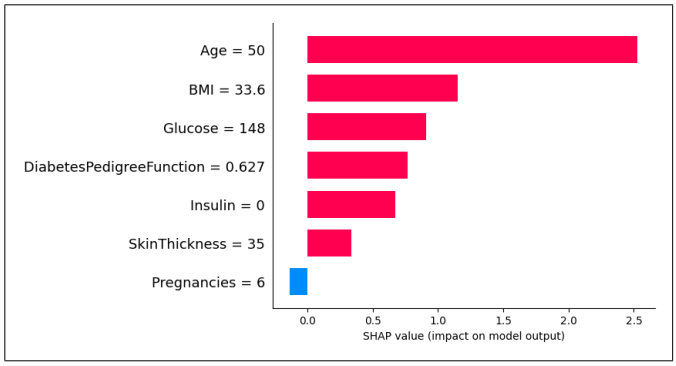
このように，LIME による解釈も，SHAP による解釈と同様の問題を抱えていると言える．

3 提案手法

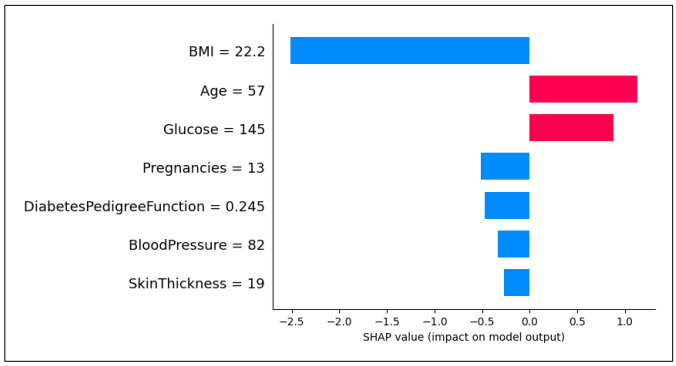
本論文では，説明対象のインスタンスに対して，SHAP を用いた解釈では不十分である特徴ペアを特定し，それらの特徴ペアに対して文献 [3] で提案された 2 次元最適相関ルールを求めることで SHAP による解釈を補完する手法を提案する．

3.1 2 次元最適相関ルール

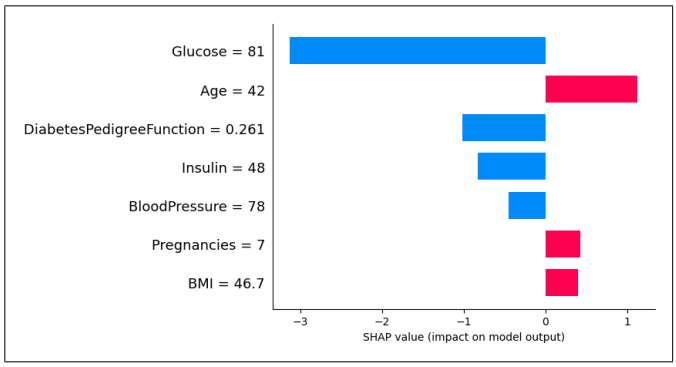
本研究で 2 次元最適相関ルールを用いる目的は，与えられた 2 つの特徴量を用いて 2 値の目的変数を持つデータセットを分



(a) BMI，血糖値（高，高）のインスタンス



(b) BMI，血糖値（低，高）のインスタンス



(c) BMI，血糖値（高，低）のインスタンス

図 3: SHAP によるインスタンス解釈

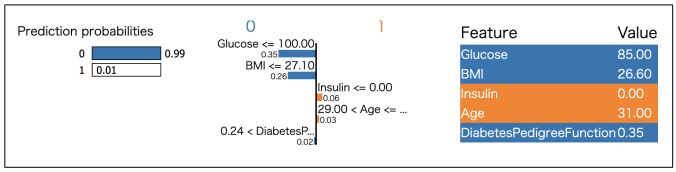
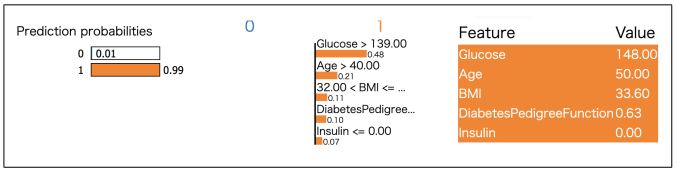


図 4: LIME による予測根拠の解釈の例

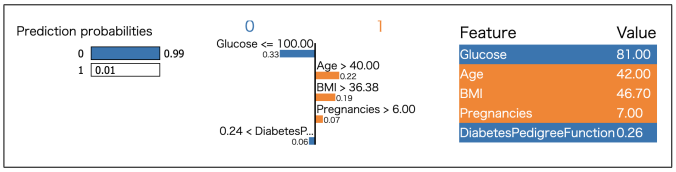
類するための最適なルールを求め，1 つの特徴量だけでは捉えられないモデルの振る舞いを解釈することである．2 次元最適相関ルールは，データセット内の数値変数 A, B，目的変数の値 y，ユークリッド平面の部分領域 P を用いて以下のように定義される [3]．

$$((A, B) \in P) \Rightarrow y$$

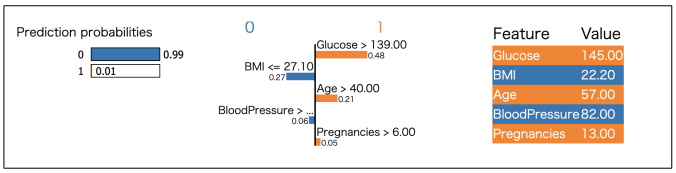
このルールは，あるインスタンスの特徴量の値が対応する点 (A, B) が領域 P に含まれるとき，そのインスタンスの目的変数の値は y であるというルールである．



(a) BMI，血糖値（高，高）のインスタンス



(b) BMI，血糖値（低，高）のインスタンス



(c) BMI，血糖値（高，低）のインスタンス

図 5: LIME によるインスタンス解釈

本研究の場合，目的変数のとる値は 2 値であるため，以下のルールも同時に成り立つ．ただし，$\neg y$ は y の否定を表している．

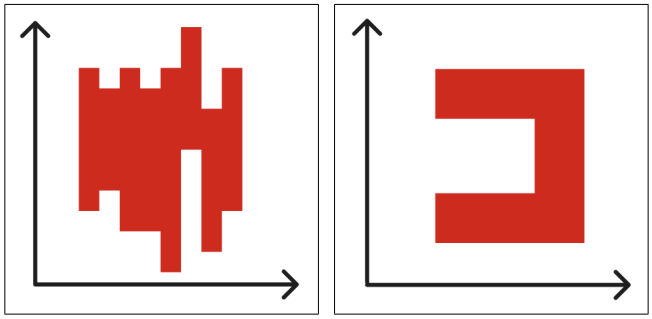
$$((A, B) \notin P) \Rightarrow \neg y$$

このルールは，あるインスタンスの特徴量の値が対応する点 (A, B) が領域 P に含まれないとき，そのインスタンスの目的変数の値は y でないというルールである．

ただし，数値変数は多くの場合に連続値を取るため，数値変数をそのまま用いて最適なルールを求めることは困難である．そこで，提案手法では数値変数をバケットに分割することで，数値変数を離散値として扱う．具体的には，2 つの数値変数 A, B をそれぞれ N, M 個のバケットに等分割する．そしてユークリッド平面を $N \times M$ ピクセルのグリッド平面 G に分割し，各グリッドとそのグリッドに含まれるインスタンスの数を対応させ，その G 上に X 単調領域 P を定義する．ここで，X 単調領域とは，図 6(a) の赤い領域のように，長方形が水平方向に連結されたような形の，全ての点において垂直方向に連続である領域のことである [7]．図 6(b) は X 単調領域ではない領域の例である．図 6(b) には垂直方向に連続でない点が存在するため，この領域は X 単調領域ではない．

G を最適に分割する領域 P は，領域分割後のエントロピーを最小にする領域であると定義する．このような領域 P は Morimoto et al. (1997) [7] の提案したアルゴリズムを用いて求めることができる．ここで，ある集合 S のエントロピー $Ent(S)$ は，p_0 を目的変数が 0 をとる確率，p_1 を目的変数が 1 をとる確率とすると，

$$Ent(S) = -p_0 \log p_0 - p_1 \log p_1$$



(a) X 単調領域の例　(b) X 単調でない領域の例

図 6: X 単調である領域とそうでない領域の例

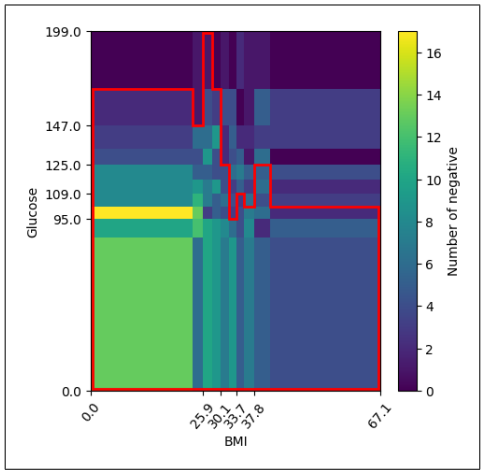


図 7: 可視化された最適領域

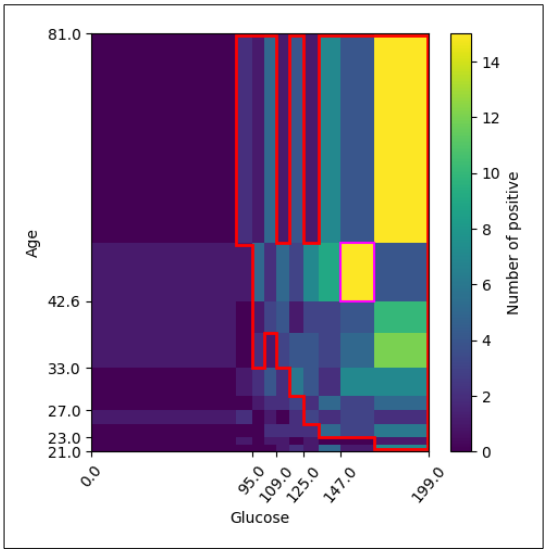
で表される．この集合 S をそれぞれ n_1 個と n_2 個のインスタンスを含む部分集合 S_1, S_2 に分割する時，領域分割後のエントロピー $Ent(n_1, n_2)$ は以下のように計算できる．

$$Ent(n_1, n_2) = \frac{n_1}{n_1 + n_2} Ent(S_1) + \frac{n_2}{n_1 + n_2} Ent(S_2)$$

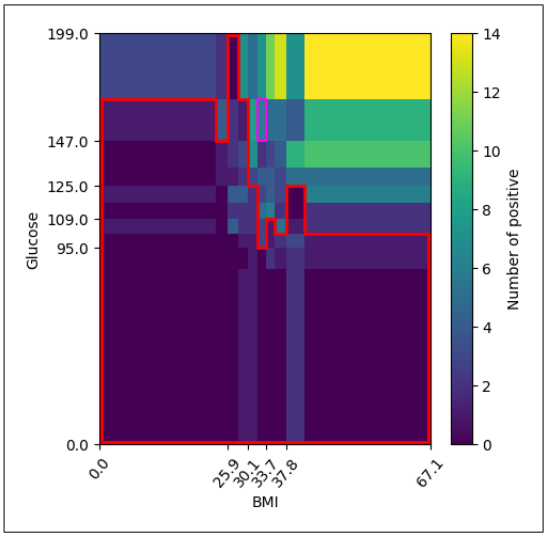
3.2 最適領域の可視化

本研究では，2 次元最適相関ルールの最適領域を可視化するビジュアライザも作成した．図 7 はビジュアライザによって糖尿病診断データセット (4 節を参照) の BMI を x 軸に，血糖値を y 軸に取った時の最適領域を出力した結果である．各点の色はそのグリッドに含まれるインスタンスの数の多さを表し，色が明るいほどそのグリッドに含まれるインスタンスが多いことを表す．図中の赤枠で囲まれた領域は最適領域を表す．図 7 は非糖尿患者数に対する図であるため，明るい色の点ほどその点に含まれる非糖尿病患者のインスタンスが多いことを表す．よって，赤線で囲まれた領域の内側は糖尿病である患者が少ないことが図から分かる．

また，あるインスタンスに対して複数の特徴ペアを用いて 2 次元最適相関ルールを求めた場合，説明対象のインスタンスが各グリッド中のどの位置にあるかを示すこともできる．図 8 は，説明対象のインスタンスが各グリッド中のどの位置にあるかを示すビジュアライザである．マゼンタ色の四角の枠で囲まれた点が説明対象のインスタンスが含まれる点であり，図 8(a) では説明対象のインスタンスが領域の内側 (図の右側中央) にあり，図 8(b) では説明対象のインスタンスが領域の外側 (図の上側中央) にあることが分かる．



(a) x 軸が血糖値，y 軸が年齢のグリッド



(b) x 軸が BMI，y 軸が血糖値のグリッド

図 8: 説明対象のインスタンスの位置の可視化

3.3 2 次元最適相関ルールによる説明が必要な特徴ペアの特定

2 次元最適相関ルールはデータセットに含まれる 2 つの特徴の組み合わせに対して求められるが，m 個の特徴量を持つデータセットに対しては，m^2 個の特徴ペアが存在する (x 軸に使う特徴量と y 軸に使う特徴量は区別される)．そのため，全ての特徴ペアに対して 2 次元最適相関ルールを求め，それら全てをユーザーに提示することは，ユーザーがモデルの振る舞いを解釈することを困難にするため実用的ではない．

そこで，提案手法では説明対象の各インスタンスに対して SHAP による解釈では不十分である特徴ペアを特定し，それらの特徴ペアのみを用いて 2 次元最適相関ルールを提示することでモデルの振る舞いを説明する．本研究では，2 次元最適相関ルールによる説明が必要な特徴ペア A, B が満たすべき条件を

以下のように定義した．

（1） データセット内の全てのデータに対して数値変数 A, B を用いて作成した 2 次元最適相関ルールが出力する値と，モデルが出力する値の一致度が高い．

（2） 説明対象のインスタンスに対して，数値変数 A, B をもとに作成した 2 次元最適相関ルールが出力する値と，モデルが出力する値が一致している．

（3） 数値変数 A, B のうちどちらか 1 つの値だけではモデルの出力を説明できない．

これらの条件を満たす特徴ペアを特定するために，本研究では以下の手順で特徴ペアを絞り込んでいく．

まず，全ての特徴ペアが含まれる集合 S_0 の各特徴ペアに対して 2 次元最適相関ルールを求めて，各インスタンスに対する 2 次元最適相関ルールの出力値とモデルの出力値の一致度を計算し，その一致度が閾値 θ を超える特徴ペアのみを集めた集合 S_1 を作成する．ここで，2 次元最適相関ルールが出力する値とモデルが出力する値の一致度は以下のように定義する．

$$\frac{\Sigma_{i=1}^{N} x_i | R(x_i) = f(x_i)}{N}$$

なお，N はデータセットに含まれるインスタンス数，$R(x_i)$ は 2 次元最適相関ルールがインスタンス x_i に対して出力する値，$f(x_i)$ はモデルがインスタンス x_i に対して出力する値である．

次に，S_1 に含まれる各特徴ペアのうち，説明対象のインスタンスに対する 2 次元最適相関ルールが出力する値とモデルが出力する値が一致している特徴ペアのみを集めた集合 S_2 を作成する．つまり，S_2 に含まれる特徴ペアから作られる全ての 2 次元最適相関ルールは，以下の条件を満たす．

$$R(x_k) = f(x_k)$$

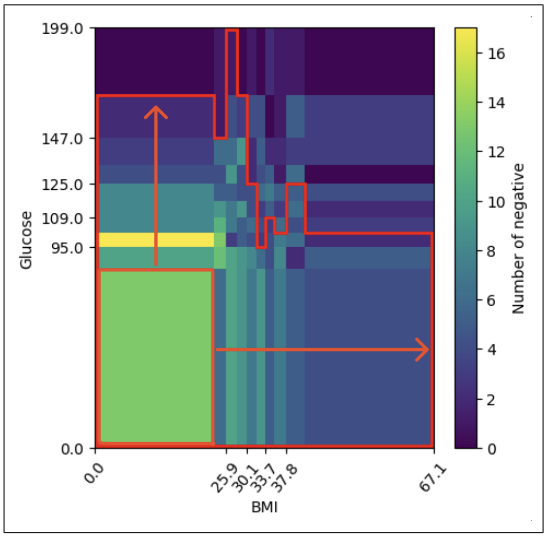
ただし，x_k は説明対象のインスタンスである．

最後に，S_2 に含まれる特徴ペアの中から，特徴ペアのうちどちらか 1 つの特徴量だけでそのインスタンスに対するモデルの挙動を説明できる特徴ペアを除外する．このステップでは，説明対象のインスタンスが対応するグリッド上の点に対してどちらか一方の特徴量の値だけを変化させても，その点が領域 P に含まれるかどうかが変化しない場合，その特徴量ペアは除外される．つまり，最終的に提示される 2 次元最適相関ルールは，以下の条件を満たす．

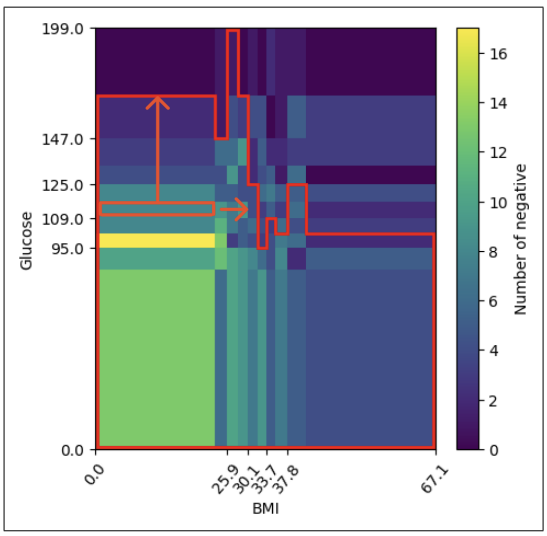
$$need_two_dimensional(A, B) = true$$

ここで，$need_two_dimensional(A, B)$ は，特徴ペア (A, B) の特徴量のうちどちらか 1 つでも特徴量の値を変化させることで領域 P に含まれる点が変化する場合に $true$ を返し，そうでない場合に $false$ を返す二項関数である．

例えば，図 9 で用いられている領域において，図 9(a) のオレンジ色の枠で囲まれている点に含まれるインスタンスに対しては，BMI の値がどの値を取っても領域 P に含まれるため，この 2 次元最適相関ルールは不要であると判断され，BMI と血糖値を用いた 2 次元最適相関ルールは提示されない．一方で，図 9(b) のオレンジ色の枠で囲まれている点に含まれるインスタンスに対しては，BMI や血糖値の値を変化させることでその点は領域 P に含まれなくなるため，BMI と血糖値の両方を用いた 2 次元最適相関ルールを提示する．



(a) 2 次元の解釈が不要なインスタンス



(b) 2 次元の解釈が必要なインスタンス

図 9: 2 次元最適相関ルールによる説明例

また，この手順で特徴ペアを絞り込んだ後，2 次元最適相関ルールを提示する際には，グリッドを 2 次元最適相関ルールで分割した後のエントロピーを基準にして特徴ペアを順位づけする．そして，エントロピーの小さい順に特徴ペアを提示する．

4 実　　験

提案手法の有効性を検証するために，糖尿病診断データセットに対して，勾配ブースティング決定木アルゴリズムの 1 種である CatBoost を用いて勾配ブースティング決定木を作成し，そのモデルの出力を目的変数として，データセットに含まれる特徴量を用いて 2 次元最適相関ルールを求めた．ここで作成し

た予測モデルは，患者の検査値などを基にして患者が糖尿病を有するかどうかを予測することを目的としている．

まず，2 次元最適相関ルールが SHAP による解釈を補完できていることを確かめるため，特徴量の値が似ていない複数のインスタンスに対して，2 次元最適相関ルールを提示することで SHAP では解釈できないモデルの挙動を解釈できることを示す．図 10 は，糖尿病診断データセットに含まれるある 4 つのインスタンスに対して SHAP を適用した結果であり，図 11 は特徴量の値と SHAP 値の関係を表している．2 節で述べたように，これらの SHAP の結果からは，各インスタンスにおける特徴量の値や各特徴量の値の変化がモデルの出力に与える影響を解釈することはできても，相関のある複数の特徴量の値の変化がモデルの出力に与える影響を解釈することはできない．そこで，2 次元最適相関ルールを用いて，それぞれのインスタンスの特徴量の値がどのようにモデルの出力に影響しているかを説明する．

次に，説明対象のインスタンスに対する提案手法の有効性を検証する．糖尿病診断データセットには 8 個の説明変数が含まれているため，2 次元最適相関ルールを求めることができる特徴ペアは $8^2 = 64$ 個存在するが，提案手法では各インスタンスに対して説明に適切な特徴ペアを絞り込んでいくつかのルールを提示する．そうして提示されたルールが説明対象のインスタンスに対するモデルの出力の理由を説明できているかどうかを確認する．本実験では，各ルールとモデルの出力の一致度の閾値 θ は 0.7 に設定した．

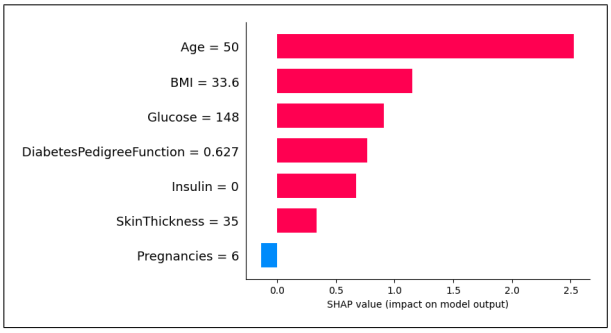
4.1 データセット

本研究では Kaggle[1]にて公開されている糖尿病診断データセット [2]を対象に実験を行う．糖尿病診断データセットは患者の糖尿病の有無を予測することに焦点を当てており，患者の年齢や BMI に加え，血圧や血糖値などの検査値，糖尿病を有するかどうかを表す計 9 個の変数が含まれている．また，このデータセットに含まれる全ての患者は 21 歳以上のピマ・インディアンの遺伝子を持つ女性に限られている．

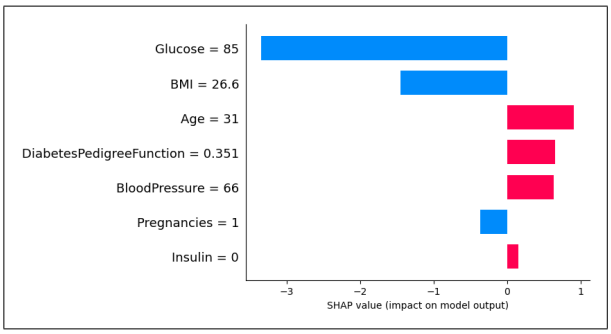
4.2 結果と考察

4.2.1 2 次元最適相関ルールによる SHAP の解釈の補完

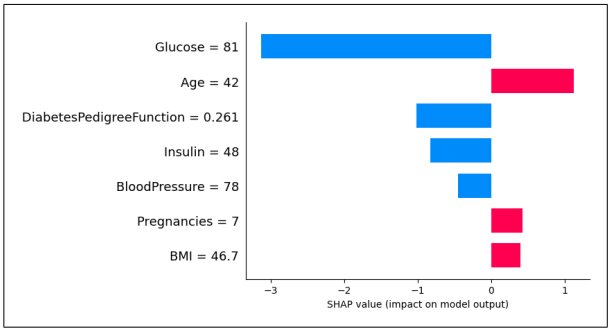
図 10 で示したインスタンスに対するモデルの出力の違いの原因を説明するために，BMI と血糖値に対して提案手法を適用した結果を図 12 に示す．図 12 において，横軸は BMI，縦軸は血糖値を示す．また，各点の色は各点に対応する患者のうち糖尿病であると診断された患者の人数を表す．この図を見ると，糖尿病であると診断された患者は BMI と血糖値の両方の値が大きい領域に集中していることが分かり，BMI と血糖値のどちらかが低い患者は糖尿病であると診断されない傾向があることが分かる．したがって，図 10 の 4 つのインスタンスのうち図 10(a) のインスタンスのみが糖尿病であると診断されたのは，4 つのインスタンスのうち BMI と血糖値が共に高かったのが図 10(a) のインスタンスのみであったためであると解釈できる．このようにして提案手法は SHAP の解釈を補完できていることが分かる．



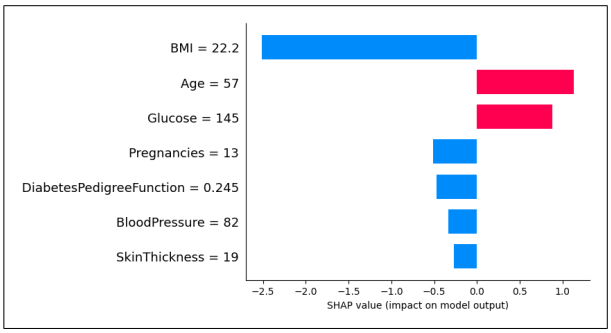
(a) BMI，血糖値（高，高）のインスタンス



(b) BMI，血糖値（低，低）のインスタンス



(c) BMI，血糖値（高，低）のインスタンス



(d) BMI，血糖値（低，高）のインスタンス

図 10: SHAP によるインスタンス解釈

4.2.2 提案手法の有効性

図 13 は糖尿病でないと診断されたあるインスタンスに対して SHAP を適用した結果である．このインスタンスに対して提

1 : https://www.kaggle.com/
2 : https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

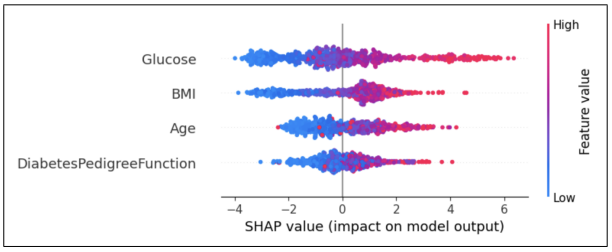


図 11: 特徴量の値と SHAP 値の関係

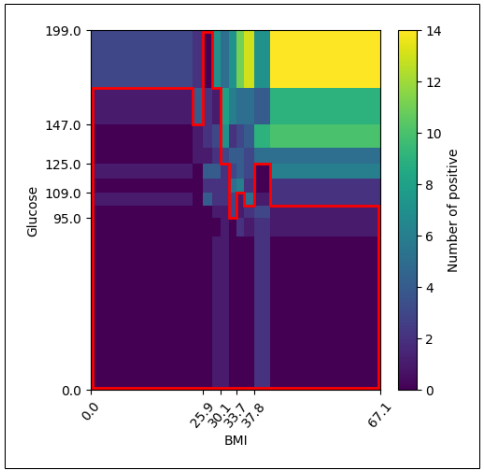


図 12: BMI と血糖値に対する 2 次元最適相関ルール

案手法を適用した結果，図 14 に示す 2 つのルールを含む計 6 つのルールが出力された．ここで，図 14(a) は 1 番目に提示されたルールであり，図 14(b) は 2 番目に提示されたルールである．図 14(a) は DiabetesPedigreeFunction(糖尿病の遺伝的リスクのスコア) と血糖値を用いたルールであり，領域の形を見ると，これは概ね＜血糖値が低いか，血糖値は高すぎず糖尿病の遺伝的リスクのスコアも高すぎないならば糖尿病でない＞というようなルールであると解釈できる．そして，このインスタンスは血糖値が高すぎず糖尿病の遺伝的リスクのスコアも高すぎないため，このルールでは糖尿病でないと判断される．次に，図 14(b) は BloodPressure(血圧) と血糖値を用いたルールであり，領域の形を見ると，これは概ね＜血糖値が低いか，血糖値は少し高いが血圧が高すぎず低すぎないならば糖尿病でない＞というようなルールであると解釈できる．そして，このインスタンスは血糖値が高すぎず血圧も高すぎないため，このルールでは糖尿病でないと判断される．図に示した以外のルールには，＜年齢が低いか，年齢は高いが糖尿病の遺伝的リスクのスコアが低いならば糖尿病でない＞と解釈できるルールや，＜ BMI が低いか，BMI は少し高いが年齢が若いならば糖尿病でない＞と解釈できるルールなどが提示された．これらのルールを見ることで，例えば，なぜこのインスタンスは遺伝的リスクのスコアが高いのにモデルは糖尿病でないと判断したのかという問題を考える時に，それはこのインスタンスの血糖値が高すぎないためであるということが分かる．このようにして，提案手法は説明対象のインスタンスに対するモデルの出力の理由を説明するルールを抽出し，提示できていることが分かる．また，提

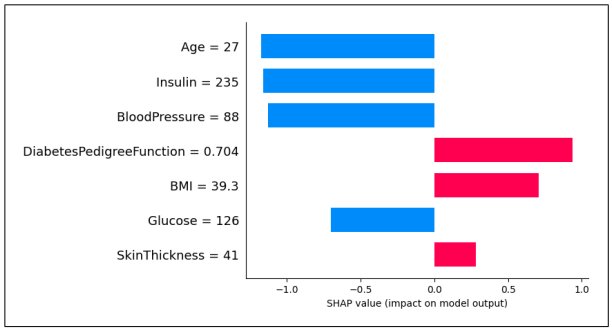
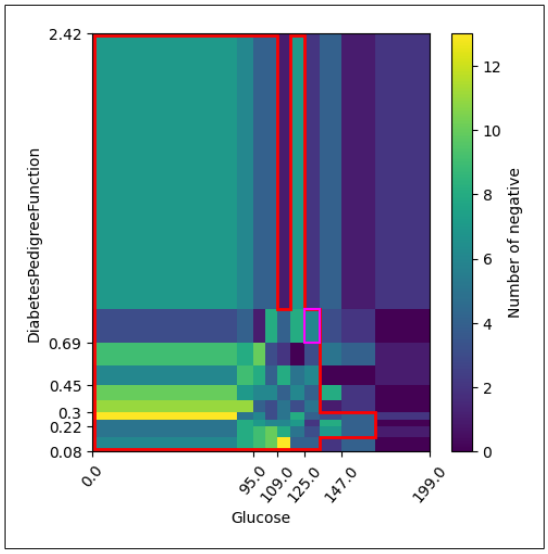
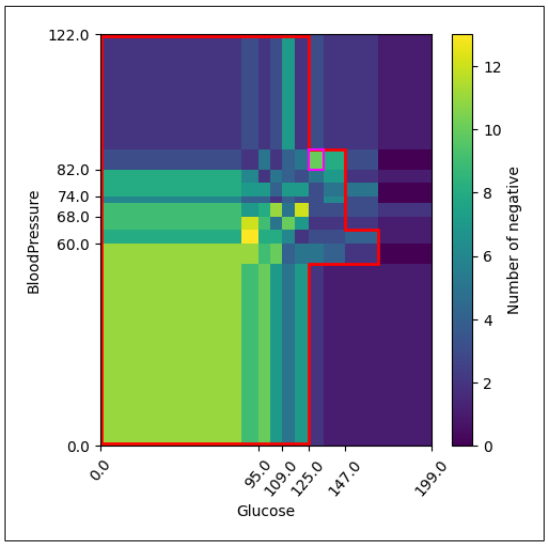


図 13: 糖尿病でないと診断されたインスタンスに対する SHAP の解釈



(a) 糖尿病の遺伝的リスクのスコアと血糖値を用いたルール



(b) 血圧と血糖値を用いたルール

図 14: 図 13 のインスタンスに対する提案手法の出力に含まれるルール

案手法においてエントロピーを基準に順位付けされたルールのうち，上位のルールが説明対象のインスタンスに対するモデルの出力の理由を適切に説明できていることも分かる．

5 関 連 研 究

5.1 相関ルールを用いた代理モデルの作成

代理モデルに基づく XAI 手法は，説明対象のモデルの挙動を近似する透明性の高いモデルを作成することで元のモデルの挙動を理解する手法である [8]．例えば，ニューラルネットワークのようなブラックボックスモデルに対して，決定木を用いた代理モデルを作成することで，ブラックボックスモデルの挙動を理解することができる．

Himabindu et al. (2019) [9] や Milad et al. (2021) [10] では，相関ルールの集合で構成される決定セットを用いた代理モデルを作成する手法を提案し，元のモデルの挙動を良く近似しつつ，ユーザーが元のモデルの挙動を容易に理解することができる説明を提示する代理モデルが作成できたことを示した．これらの手法では，if-then 形式で構成される決定セットを用いて説明を生成するため，数値変数を扱う場合は数値を離散化しており，生成された説明では数値の順序は考慮されていない．本研究では，数値の順序を考慮した説明を生成し，数値変数の値の変化に対するモデルの挙動の変化の説明を試みた．

5.2 最適数値相関ルールを用いた SHAP の解釈補完

山下ら (2023) [11] は，1 次元の最適数値相関ルールを用いて SHAP による解釈を補完する手法を提案した．彼らは SHAP で生成した説明において最も SHAP 値の大きい特徴量に対して最適数値相関ルールを求めることで，＜21 歳から 22 歳の人の内 84.7%の人が糖尿病ではない＞というような元のデータセットにおける事実に基づいたルールを提示することで，SHAP による解釈をデータの傾向という観点から補完している．本研究では，2 次元最適相関ルールを用いることで，彼らの手法では説明しきれなかった，相関のある数値変数間に対するモデルの挙動の説明を試みた．

5.3 2 次元空間の可視化による解釈

SHAP の dependence plot という可視化手法や，Partial Dependence Plot (PDP) [12] では 2 つの特徴量とモデルの出力の関係を 2 次元空間にプロットすることで，モデルの挙動を解釈する．図 15 は SHAP の dependence plot による説明の例である．図 15 の各点はインスタンスを表し，横軸に 1 つ目の特徴量の値，縦軸にその特徴量の値に対する SHAP 値をとり，各点の色はもう 1 つの特徴量の値を表している．図 15 は PDP による説明の例である．PDP は 2 つの特徴量の値を変化させた時にモデルの出力に与える効果を示す．図 15 の横軸と縦軸はそれぞれ 1 つ目と 2 つ目の特徴量の値を表し，各領域の色はその領域に含まれるインスタンスのモデルの出力の平均値を表している．これらの手法では，2 つの特徴量の値の変化がモデルの出力に与える影響を説明することができるが，膨大な数の特徴ペア全てに対して可視化による説明を提示することは実用的ではない．そのため，可視化による解釈では，特徴ペアの順位付けが必要となる．そこで，提案手法では特徴ペアの絞り込みと順位付けを行うことで適切な数の特徴ペアを提示することを目指した．

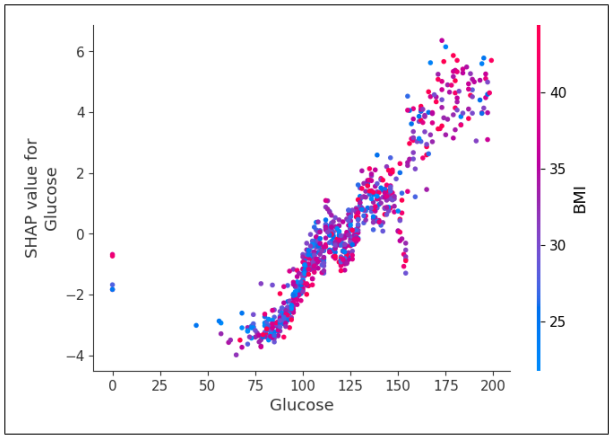


図 15: SHAP dependence plot による説明

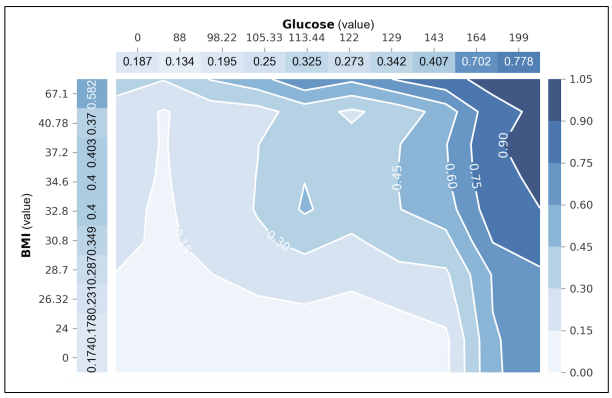


図 16: PDP による説明

6 結　　論

本論文では，SHAP では解釈できない，相関のある特徴量に対するモデルの振る舞いを説明することで SHAP による解釈を補完するため，2 次元相関ルールを用いた説明の生成手法を提案した (3 節)．

提案手法の具体的な手順は以下の通りである．まず，データセットから作成可能な全ての 2 次元相関ルールを求める．次に，それらのうちデータセット中のインスタンスに対する出力がモデルの出力と一致している割合が多いものを絞り込む．そして，説明対象のインスタンスに対する出力がモデルの出力と一致しているものを絞り込む．最後に，説明対象のインスタンスに対して 2 次元相関ルールを用いた説明が必要な特徴量の組み合わせを絞り込む．以上の手順で絞り込まれたルールを説明として出力する．これにより，説明対象のインスタンスに対するモデルの挙動を説明する少ない数のルールが抽出できることを確認した (4 節)．

今後の課題として，まず求めたルールの領域の形に凸凹が多いことが挙げられる．これは，グリッドの密度が低いことが原因で起こる．例えば，身長と体重が含まれるデータセットにおいては，身長がとても高く体重がとても軽い人は少ないため，身長と体重という特徴ペアに対して 2 次元最適相関ルールを求めると，身長が高く体重が軽いインスタンスに対応する点の密

度が低くなり，領域の形が不安定になって凸凹した形になる．これにより，モデルに対するルールの一致度の悪化や，ルールの解釈性の悪化といった問題が生じる．これらの課題に対処するため，全ての点において垂直方向にのみ連続である X 単調領域の代わりに，全ての点において垂直方向にも水平方向にも連続である直交凸領域を用いて 2 次元最適相関ルールを求めることで，領域の凹凸が少なくなり，モデルに対するルールの一致度の向上やルールの解釈性の改善が期待される．また，生成されたルールの適切さを測る評価指標作成も今後の課題である．適切な評価指標が作成できた場合，ルール抽出の精度を向上させることができる．

謝　　辞

本研究は科研費 20K11830 の助成を受けたものです．

文　　献

[1] 大坪直樹, 中江俊博, 深沢祐太, 豊岡祥, 坂元哲平, 佐藤誠, 五十嵐健太, 市原大暉, 堀内新吾. XAI(説明可能な AI)–そのとき人工知能はどう考えたのか? (AI/Data Science 実務選書). リックテレコム, 2021.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., 2017.

[3] Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, and Takeshi Tokuyama. Data mining with optimized two-dimensional association rules. ACM Trans. Database Syst., Vol. 26, No. 2, p. 179–213, jun 2001.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, p. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[5] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Proceed- ings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18, p. 6639–6649, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, Vol. 29, No. 5, pp. 1189 – 1232, 2001.

[7] Yasuhiko Morimoto, Takeshi Fukuda, Shinichi Morishita, and Takeshi Tokuyama. Implementation and evaluation of decision trees with range and region splitting. Con- straints, Vol. 2, No. 3, pp. 401–427, Dec 1997.

[8] 亀谷由隆. 説明可能 AI 技術のこれまでとこれから. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 16, No. 2, pp. 83–92, 2022.

[9] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’19, p. 131–138, New York, NY, USA, 2019. Association for Computing Machinery.

[10] Milad Moradi and Matthias Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. Expert Systems with Applications, Vol. 165, p. 113941, 2021.

[11] 山下史紘, 亀井清華, 森本康彦. 最適数値相関ルールを利用した SHAP の予測モデル解釈の補完. In DEIM, 2023.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, Vol. 29, No. 5, pp. 1189 – 1232, 2001.

インドの IT サービス会社の Shapley 値を用いた 15 年間の分析
ーShapley 値による時系列データ分析ー

白田　由香利†　　Sreekanth Bonthala‡　　Basabi Chakraborty[1,2]

†学習院大学経済学部　〒171-8588　東京都豊島区目白 1-5-1
‡マダナパレ工科大学経営学部　Madanapalle Chittoor, Andhra Pradesh 517325, India
1 マダナパレ工科大学計算機学部　Madanapalle Chittoor, Andhra Pradesh 517325, India
2 岩手県立大学　〒020-0611　岩手県滝沢市巣子　152-52

E-mail:　† yukari.shirota@gakushuin.ac.jp,　‡ sreekanthb@mits.ac.in ,　2 basabi@iwate-pu.ac.jp

あらまし　インドの IT サービス企業，特に TCS や Infosys などの企業の時価総額の増加は著しく，その成長の主要因を特定するために機械学習に基づく回帰分析を利用する．15 年にわたる長期データを使用し，SHAP（Shapley Additive exPlanations）値を用いて，各年度における主要因を繰り返し分析する．この方法により，主要因の時系列変化を探ることが可能となる．

キーワード　インド，IT サービス企業，Tata Consultancy Services, Infosys，時価総額，XAI, SHAP

1. 始めに

インド企業において，IT サービス産業の発展が顕著である．IT サービス産業とは，ソフトウェア関連サービスおよび，業務プロセスの受託サービス BPO(Business Process Outsourcing)などを指す[1]．本稿の分析の目的は，インド IT サービス企業の時価総額の急成長の主要要素を究明することである．要因として，売上高営業利益率，有形固定資産回転率，の 2 つを比較する．売上高営業利益率は，技術水準の高さ等のソフトウェア的指標と見ることができるであろう．他方，有形固定資産回転率はハードウェア的効率性の指標と見ることができると考えた．インド IT サービス企業は，どちらが主要因となり時価総額の増加に貢献しているかを分析することが本論文の目的である．

15 年間の代表的インド IT サービス企業の時価総額の変動を，回帰分析及び Shapley 値(厳密には SHAP 値)を用いて分析する[2], [3], [4], [5], [6]．本論文は，SHAP 値の時系列変化を見るというケースともなる． SHAP 時系列変化の既存研究例としては，求野達による自動車企業の株価変化解析などがある[7], [8].

次節は，インド IT サービス企業について概説する．第 3 節では，データと手法について説明する．第 4 節は分析結果である．最後に総括を行う．

2. インド IT サービス産業

インドの IT 関連サービス産業の市場規模拡大は顕著である．インドの IT サービス企業は，(1)インド系 IT 企業，(2)多国籍 IT 企業に類別できるが，本稿ではインド系 IT 企業の時価総額の増加に対象をおく．代表的な企業として，TATA CONSULTANCY SERVICES(以下，TCS)，INFOSYS ，HCL TECHNOLOGIES(以下，HCL)，WIPRO，MPHASIS 等がある[1]．TCS はインドのタタ財閥の系列会社の中でも代表的な企業である．日本へは 1987 年に進出している．これらの IT サービス企業の時価総額の 15 年間の伸びは非常に膨大なものである(図 1 参照)．

図 1 にインド IT サービス企業の時価総額の 2008 年から 2022 年の年次データ変化を示した．本稿のデータは，指定がない限り，全て Bureau van Dijk（ビューローバンダイク社）による ORBIS データベースから検索した．時価総額は，株価×発行株数であるので，株価分割に依存せずに，その企業の株価の成長を見ることが可能となる．

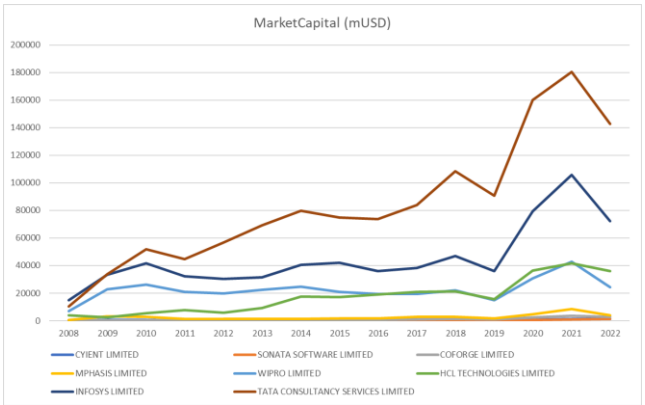


図 1：インド IT サービス企業の時価総額の変化

最も時価総額が高い企業は TCS であり，ついで Infosys が高い．TCS を見ると，2008 年には 20000(百万ドル)であったが，2021 年には 180000(百万ドル)と、約 9 倍となっている．これは非常に高い成長率である．図 1 の時価総額の変動を見ると，2019 年と 2022 年に低下が見られる．インドの会計年度は，4 月 1 日から翌年の 3 月 31 日までである．2019 年度の下落は，

2020 年度 1 月から 3 月の第 4 四半期に COVID-19 が勃発したからである.インドは COVID-19 により多大な死者を出してしまい，経済が停滞した**[9], [10], [11]**．2022 年度の下落**理由**であるが，コロナ後の約 18 ヶ月間，インド IT 株は好調であったが，2022 年はインフレの悪化とそれに伴う利上げ，地政学的緊張が世界的な成長の鈍化により，インド国内の IT 企業の株価が下落した**[12]**.

図 2 に TCS の株価変動を示す. 図中にある XSNE とは，National Stock Exchange of India Limited (NSE)であり，インドで最も重要な株式取引所の一つである. 米国中央銀行 FRB が金利上昇に転じた 2022 年 3 月以降，株価が下落しているようすが分かる.

図 2: TCS の株価変動(www.bing.com より引用)

売上高の推移を図 3 に示したが，TCS が売上高トップである.1USD を 140 円で換算した場合, TCS の 2022 年度売上高 27,000 million USD は約 3 兆 7800 億円となる．売上高において，COVID-19 及び 2022 のインフレ利上げなどの影響による下落は殆ど見られない．2019 年度の COVID-19 の年でも，売り上げは横ばい状態である.

インド IT サービス企業の売上の内訳は一般的に以下のような内容から構成される.

- IT コンサルティングおよびカスタムソフトウェア開発
- デジタルトランスフォーメーションに関連するプロジェクト等．デジタルコンサルティング，デジタルマーケティング，デジタルプロダクト開発などが含まれる.
- インフラストラクチャサービス: サーバー, ネットワーク，クラウドサービスなど，クライアントのインフラストラクチャを構築・管理・保守するサービス.
- ビッグデータ分析:ビジネス上の意思決定をサポートするサービス． AI や機械学習の導入が含まれる.
- クラウドコンピューティングサービス
- IoT: センサーやデバイスを活用して物理的なオブジェクトをインターネットに接続し,データを収集・分析して効率や生産性を向上させるサービス.

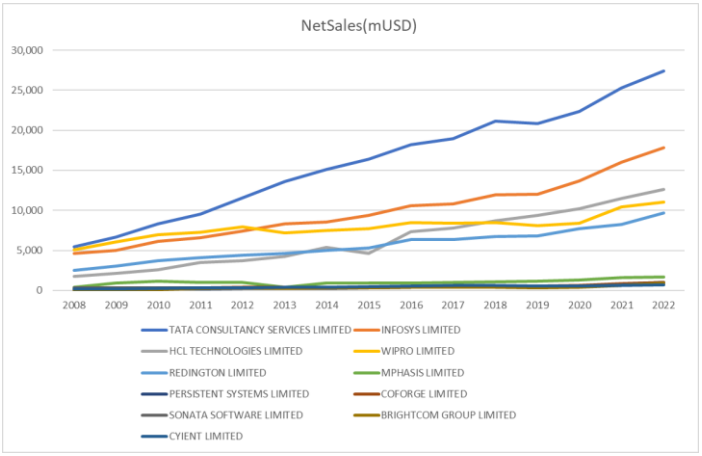


図 3: インド IT サービス企業の売上高推移

インド IT サービス企業の時価総額の増加の原因を考えてみる．インドにはインド工科大学(IIT)卒業生を筆頭とする高度 IT 人材が豊富であり，技術水準が高く，かつ，欧米に比較して人件費が低いこと，などの理由が挙げられる[13]. 当初, 産業界はインドに設立された工科大学を利用しており企業内ではほとんど職業訓練は行われていなかった．しかしソフトウェア産業が成長するにつれ大規模な人材育成が社内で行われ，多くの民間企業が高等教育へ参入した. Infosys や TCS のような大手ソフトウェア企業は収益のかなりの額を教育に費やし，NITT や HCL のような企業もソフトウェア教育に参入した．その後，労働力の需要が急速に増加して,若者がソフトウェア産業に参入する一方で, 教育市場から人材供が行われた[13].

インドの「シリコンバレー」として知られるベンガルールは, 多数の GCC(グローバル・ケイパビリティー・センター) が集まっている[14]．GCC は多国籍企業が設置する海外拠点で，IT，エンジニアリング，R&D などの専門サービスを提供し，その企業の世界的ビジネス効率化と技術革新を支援するセンターである．インドのテクノロジー企業の業界団体ナスコム(NASSCOM, National Association of Software and Service Companies)によれば，インドには全世界の GCC の約 40%に相当する 1600 社がインドにある[15]．2022 年には，66 社の世界的企業がインドに最初の GCC を設

置した．GCC は 2023 年 3 月までの会計年度に合計で約 460 億ドル（約 6 兆 4400 億円）の収入を生み，インドのサービス輸出額を前年度比 27%増に押し上げた．GCC 設置増加は，インド IT サービス企業の時価総額躍進に強い相関をもつと考えられる[15]．

本稿の分析の目的は，時価総額の成長の主要要素を究明することである．本稿では，要因として，**売上高営業利益率，有形固定資産回転率**，の 2 つを比較する．図 4,5 にこれら指標の変動を示す．

3. データと手法

本節では，分析に用いたデータと手法を解説する．

分析は回帰を用いる．ターゲット変数は**時価総額**であり，説明変数は**(1)売上高営業利益率，(2) 有形固定資産回転率**，とした．売上高営業利益率は，営業利益÷売上高，と定義され，本業でどれだけ利益をあげているかの指標である[16]．図 4 に売上高営業利益率の変動を示す．TCS と Infosys が高く，常に 20%以上であることが分かる．しかし，近年利益率の低下が見られる．GCC 急増などにより従業員の確保が困難となり賃金上昇につながっていることが主原因のひとつと考えられる[15].GCC は先進技術を扱うため、人材確保はさらに困難となるであろう．

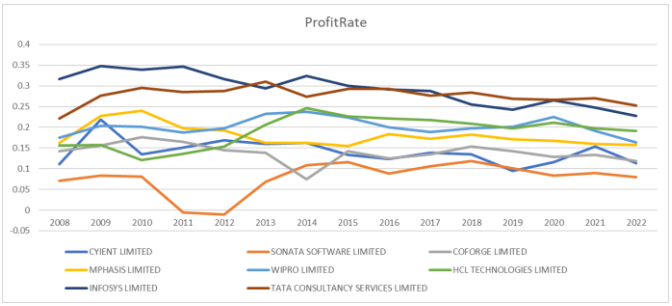


図 4:　インド IT サービス企業の売上高利益率の推移

有形固定資産とは，物理的な形の資産であり，通常は土地，建物，機械装置，輸送機関，オフィス家具などの物理的な資産を指す．IT 関連企業の場合，オフィスビル・施設，IT 機器，車両など，オフィス家具，テスト環境テストハードウェア等を指す．これらの資産を有効活用しているかを測るため，**売上高÷資産＝資産回転率**，を導入する．製造業の場合，だいたい 1 倍前後であるが，コンサルティング会社やソフトウェア会社の場合，資産の必要度は低いため，資産回転率が高くなる傾向がある[16]. 図 5 に示すように，SONATA, MPHASIS が非常に高い有形固定資産回転率を示した理由は，他の大規模企業に比較して時価総額が小規模であるからと言える．有形固定資産回転率の効果を測定するためには，こうした極端に大きい値をもつ企業は除外して回帰分析を行ったほうが適切であると考えた．そのため，この 2 社を除く 6 社で分析を行うこととした(図 6 参照)．

回帰分析のアルゴリズムは XGBOOST[17]を用いた．データ数が少ないので，訓練データとテストデータに分けず，全データで回帰モデルを作成している．決定係数は約 1.0 である（図 7 参照）．

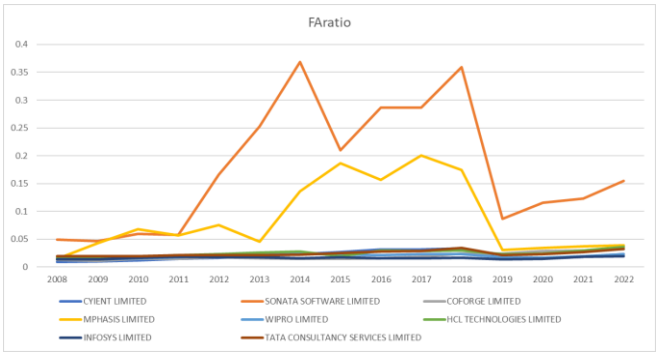


図 5: インド IT サービス企業の有形固定資産回転率の推移

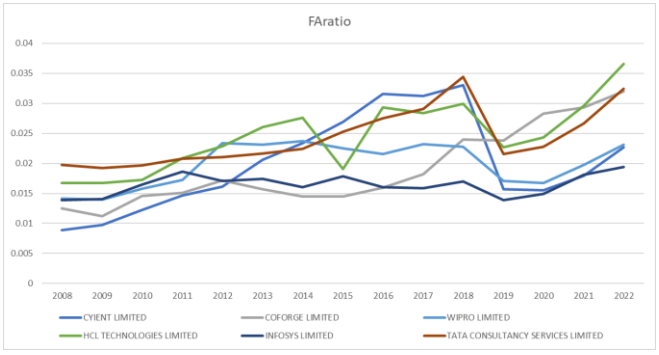


図 6: インド IT サービス企業の有形固定資産回転率の推移(MPHASIS 及び SONATA を除いた 6 社)

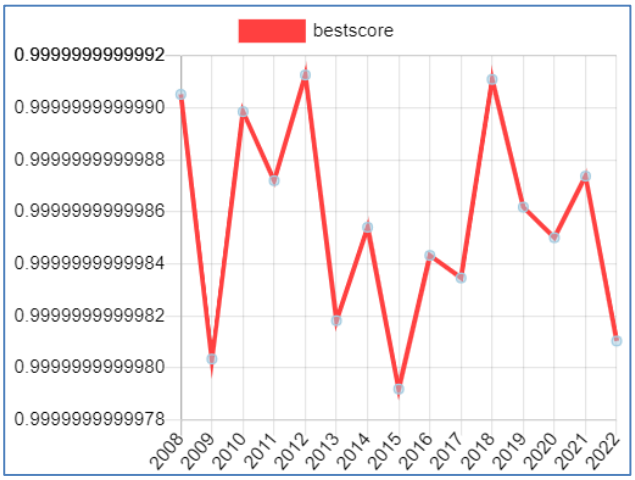


図 7: 回帰分析における決定係数の変化

4. 分析結果

本節では，回帰分析し，回帰モデルから得られた

SHAP 値について論じる．まず，XGBOOST 法の決定係数は約 1.0 と高い(図 7 参照)．横軸がデータの年度，縦軸が決定係数である．

TCS の SHAP 値の変化の様子を図 8 に示した．

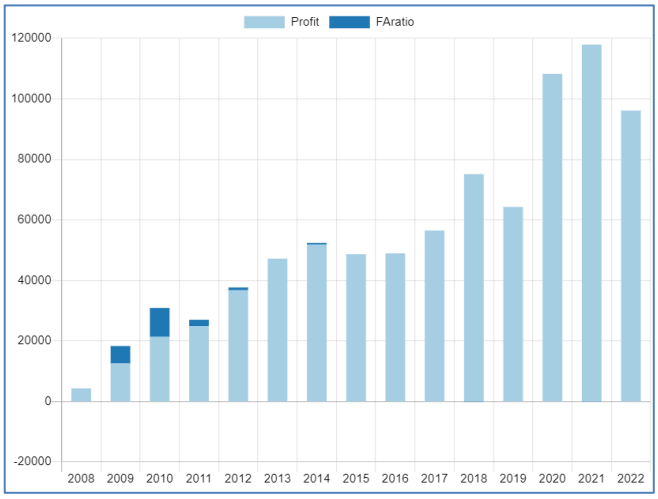


図 8：TCS の SHAP 値の変化

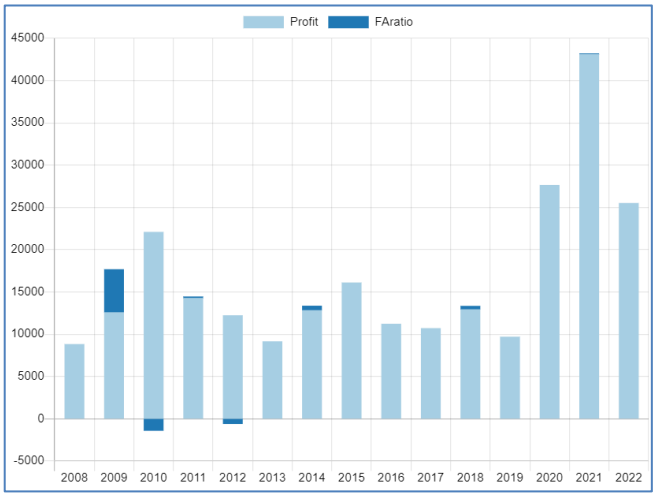


図 9: Infosys の SHAP 値の変化

縦軸は，ターゲット値の偏差(業界平均からの偏差)である．図 1 の時価総額のグラフから分かるように，TCS は業界でトップの時価総額をもつので，正の値が非常に高い．一つの棒が，特定年度の回帰分析の結果の SHAP 値の積み上げである．薄い青が SHAP_Profit（売上高利益率）であり，濃い青が SHAP_FAratio（有形固定資産回転率）である．2009 年から 2011 年には，SHAP_FAratio が僅かながら存在したが，それ以降，ほぼゼロであり，すべて SHAP_Profit がターゲット値に貢献していることが分かる．図 9 に Infosys の結果を示したが，TCS と同様に，2010 年以降では，ターゲット値への貢献は殆どが Profit によるものであることが分かる．

全 6 社において，SHAP_profit がターゲット値への貢献の殆どを占めていた．この場合，ターゲット値への貢献を比較する指標は，SHAP 値の絶対値の平均値が適切とあると考えた(図 10 参照)．

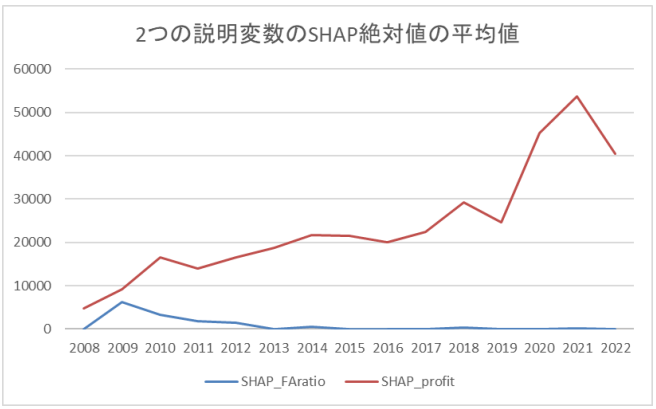


図 10: 各説明変数の SHAP 値絶対の平均値の推移

ここから分かることは，SHAP_FAratio の貢献は 2009 年から 2013 年にかけて下落を続け，2013 年から 2022 年まで殆どゼロである．時価総額への貢献はすべて SHAP_profit の貢献であることが分かった．IT サービス企業の利益の核は，ソフトウェア的な技術の高さであり，その高さと利益率は多大に影響すると解釈できる．有形固定資産の回転率などのハード面での効率化は，2009 年 2010 年のような状況では，少しは影響があったが，それ以降影響は減少し，2013 年以降はゼロが続いている．IT サービス業界は，製造業とは異なり，有形固定資産回転率は利益率に比較して，時価総額の伸びには殆ど影響しないことが分かった．

5. まとめ

本稿では，インド IT サービス企業の急激な時価総額の伸びの要因を回帰分析によって分析した．その結果の解釈として SHAP 値を用いた．ターゲット値は時価総額，説明変数として，売上高営業利益率，有形固定資産回転率，の 2 つを用いた．売上高営業利益率は，技術水準の高さ等のソフトウェア的指標と見ることができるであろう．他方，有形固定資産回転率はハードウェア的効率性の指標と見ることができると考えた．結果として IT サービス業界は，製造業とは異なり，有形固定資産回転率は利益率に比較して，時価総額の伸びには殆ど影響しないことが分かった．

謝辞

本研究は部分的に，学習院大学 GEM プロジェクト 2023 年度，学習院大学計算機センター2023 年度特別プロジェクト，科研 B，20H01537(代表：白田由香利)の助成によるものである．

参 考 文 献

[1] 深澤光樹, 佐藤隆広・上野正樹編著 高口康太編集協力 『図解インド経済大全ー政治・社会・文化から進出実務までー』 白桃書房, 2021 年, 420 頁," *Resource*, vol. 2022, pp. 05-10, 2022.

[2] L. S. Shapley, "A value for n-person games, Contributions to the Theory of Games, 2, 307–317," Princeton University Press, Princeton, NJ, USA, 1953.

[3] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley* Cambridge University Press, 1988.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[5] 白田由香利, "最強 DB 講義：機械学習回帰における Shapley 値の活用法," in 日本データベース学会, 2023. [Online]. Available: https://dblectures.connpass.com/event/274198/

[6] 白田由香利, "チュートリアル T2: 機械学習回帰における Shapley 値の理論説明と事例紹介," *DEIM2022* 電子情報通信学会, オンライン, 2022.

[7] K. Kuno and Y. Shirota, "Time Series Analysis of Shapley Values in Machine-Learning Regression," *IEICE Technical Report; IEICE Tech. Rep.*, 2022.

[8] Y. Shirota, K. Kuno, and H. Yoshiura, "Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates," *2022 6th International Conference on Deep Learning Technologies (ICDLT)* ACM, pp. 135-141, 2022.

[9] 白田由香利, 橋本隆子, バサビ・チャクラボルティ, and リリ・フィトリ・サリ, "COVID-19 に関するインド, 日本, インドネシア SNS 上でのトピック比較," *学習院大学、東洋文化研究, Journal of Asian cultures*, vol. 24, pp. 287-328, 2022.

[10] Y. Shirota, T. Hashimoto, B. Chakraborty, and R. F. Sari, "SNS Topics Comparison on COVID-19 in India, Japan, and Indonesia," *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)* IEEE, pp. 293-298, 2021.

[11] R. Murakami, B. Chakraborty, and Y. Shirota, "Dynamic Topic Tracking and Visualization using COVID-19 related Tweets in Multiple Languages," *2021 International Conference on Artificial Intelligence and Big Data Analytics* IEEE, pp. 16-21, 2021.

[12] T. Jaiswal, "Why Are IT Stocks Falling? Indian Stock Market," in 5paisa, 2023/09/07. [Online]. Available: https://www.5paisa.com/blog/why-are-it-stocks-falling

[13] S. Ghosh, *Indian Economy* PHI Learning Private Limited., 2022.

[14] V. Dey, "Bengaluru Is The Most Desired Hub For GCCs," in GOA Institute of Management, 2021/07/21. [Online]. Available: https://analyticsindiamag.com/bengaluru-is-the-most-desired-hub-for-gccs/

[15] S. Sircar and R. Bhatia, "Goldman's Biggest Office Beyond New York Attests to India's Rise," in Bloomberg, 2023/06/14. [Online]. Available: https://www.bloomberg.com/news/features/2023-06-13/goldman-s-biggest-office-outside-new-york-shows-india-s-business-evolution

[16] 小. 一慶, *図解「ROE って何?」という人のための経営指標の教科書* PHP 研究所 2017.

[17] XGBoostDevelopers, "XGBoost Decumentation (Revision 534c940a.)," 2022. [Online]. Available: https://xgboost.readthedocs.io/en/stable/

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 13:30 ~ 15:40 | T1-B オンライン（Zoom Events)

対話システム応用

座長:田中 克己(福知山公立大学)
コメンテータ:波多野 賢治(同志社大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-B-5-01] 生成AIを利用した自己学習支援

*伊藤 優希[1]、馬 強[2] (1. 京都大学大学院 情報学研究科 社会情報学専攻 伊藤研究室、2. 京都工芸繊維大学 大学院工芸科学研究科)

14:00 ~ 14:25
[T1-B-5-02] LLMを用いた対話型鑑賞エージェントにおける繰り返し発話の抑制機構

*谷 悠翔[1]、牛尼 剛聡[2] (1. 九州大学芸術工学府メディアデザインコース　牛尼研究室、2. 九州大学大学院 芸術工学研究院)

14:25 ~ 14:50
[T1-B-5-03] 複数の大規模言語モデルによる漫才自動生成の比較分析

*下崎 安紋[1]、北畑 哲也[3]、柘植 陽介[3]、梅谷 智弘[2]、北村 達也[2]、灘本 明代[2] (1. 甲南大学大学院自然科学研究科知能情報学専攻　灘本研究室、2. 甲南大学知能情報学部知能情報学科、3. 甲南大学知能情報学部知能情報学科灘本研究室)

14:50 ~ 15:15
[T1-B-5-04] 授業改善のための意見を引き出すAIインタビュアー

*中西 勇輝[1]、桑田 若菜[2]、三林 亮太[2]、山本 祐輔[3]、大島 裕明[4] (1. 兵庫県立大学社会情報科学部大島裕明研究室、2. 兵庫県立大学情報科学研究科大島裕明研究室、3. 名古屋市立大学データサイエンス学部、4. 兵庫県立大学情報科学研究科)

生成 AI を利用した自己学習支援

伊藤　優希† 　馬　　強††

† 京都大学大学院 情報学研究科 社会情報学コース　〒 606-8501 京都府京都市左京区吉田本町
†† 京都工芸繊維大学 大学院工芸科学研究科　〒 606-8585　京都市左京区松ヶ崎橋上町
E-mail: †ito.yuki.46m@st.kyoto-u.ac.jp, ††qiang@kit.ac.jp

あらまし　言語生成 AI の進展により, 授業理解やレポート作成といった課題解決を生成 AI に頼る学生が増えている. しかし教育において重要な自己学習を生成 AI に依存すると, 成長機会損失に繋がる可能性が高い. 学生が AI に頼る原因の一つとして考えられるのは, 問題を独力で解決できないと感じることである. そこで本研究では, 課題を自力で解けない学生を, ヒントによって自己解決に導くための対話システムを提案する. 本システムはまず, 学生に問題を出し, それに対する解答を採点することで, 理解が足りていない部分を分析する. 採点は, 構文解析で問題文を採点項目別の小問に分割し, 採点モデルを用いて各小問への学生解答の正誤とその採点理由を生成することで行う. 次にフィードバック生成モデルを用いて, 採点結果に基づき改善可能なポイントを指摘する文章を生成する. この文章は, 直接答えを教えることなく, 自己解決を促すヒントを提供する. これは事前に用意したテンプレートを言語生成 AI を利用して埋めることで生成する. 生成されたヒントをもとに学生が自主的に学習を行うことで, 授業理解度と課題解決力の向上が期待される.

キーワード　対話システム, 大規模言語モデル, 適応学習

1　はじめに

学校教育において, 学生が受動的に授業を聴講するだけでなく, 自ら理解した内容を整理し, その理解を問題解決に応用することは, 授業理解度や課題解決能力の向上にとって極めて重要である. 従来の教育現場では, 教員が学生に課題やレポートを課すことで, 理解度・課題解決能力向上の機会を提供してきた. しかし近年では, ChatGPT [15] をはじめとした言語生成 AI が広く利用され始め, 誰でも簡単に, 課題に対する妥当な解答を自動生成できるようになっている. その解答生成の性能は人間並に近づいている. Waltzer らの研究によれば, 高校生の書いたエッセイと ChatGPT が生成したエッセイを教員に識別させた結果, 識別正解率は 70%であり, 参加した教師の 84%が一回以上識別に失敗していると報告されている [23]. この問題に対し, 現時点の教育現場では, ChatGPT の課題解答への利用を禁止するなどの応急的な対策が取られている [13]. しかし, 言語生成 AI は授業理解や関連情報調査の補助としては有用なツールであり, 利用禁止措置は, 適切な AI 利用の学習機会を奪う可能性がある. また AI が生成した文章を自動検出する手法も提案されているが, 完全に信頼できるほどの頑健性は担保されていない. 今後の言語生成モデルの発展次第では, ますますその検出は困難になっていくと予測される [18].

このような背景から, 一部の学生の間では, 安易に AI に問題解決を丸投げしてしまう傾向が強まっている. その根底には, 以下のような理由があると考えられる.

- 課題に労力をかけたくない.
- 自分の理解に自信がなく, 問題が解けない.

後者の学生は, 独力で課題解決を行う意思はあるものの, 授業理解が足りていないために, AI に頼ってしまっている. しかし課題解決を AI に頼る習慣が根付くと, 学生は習得した知識を定着させ, それらを課題解決に応用する機会を失ってしまう. こうした学生には, 教育者からの何らかの支援が必要である.

学生を支援する方法の一つは, 直接的な解答を避けて, ヒントを提供し, 学生が自分で問題を解決する手助けをすることである. しかし, 授業内容を理解していない学生の中には, 教員に質問することをためらったり, 自分自身では学習内容を理解できていないことに気づかなかったりする [22]. そのため, 教員に質問に来た生徒にヒントを与えるだけでは, 問題を完全に解消するのは難しい. 質問に来ない学生にも対処するには, 教員がレポート課題の解答を一つ一つ確認し, 理解不足や誤解が見られる学生全員に適切なヒントを提供することも考えられる. しかしこれは, 非常に手間がかかり現実的ではない.

そこで, 本研究では学生の課題解答を分析し, 内容に応じたフィードバックを自動生成する対話システムを提案する. このフィードバックとは, 解答の正誤や修正ヒント, 授業資料に基づいた補足情報など, 解答内容に関するコメント全般と定義する. 提案システムは図 1 のように, 学生に対して授業内容に関する設問を投げかけ, 対話を開始する. 学生が解答を入力すると, 提案システムはそれを評価し, 正しい場合は授業資料をもとに生成した課題問題への解説を提供する. 修正が必要な場合は, 模範解答を提示するのではなく関連するヒントを提供し, 学生自身に修正方法を考えさせる.

提案システムの目的は, 問題への正答率を評価することではなく, システムとの対話を通じて, 学生の授業理解度と課題解決能力を向上させることである. よってシステムには次の機能が必要である.

- 学生の理解不足箇所を特定するための採点機能

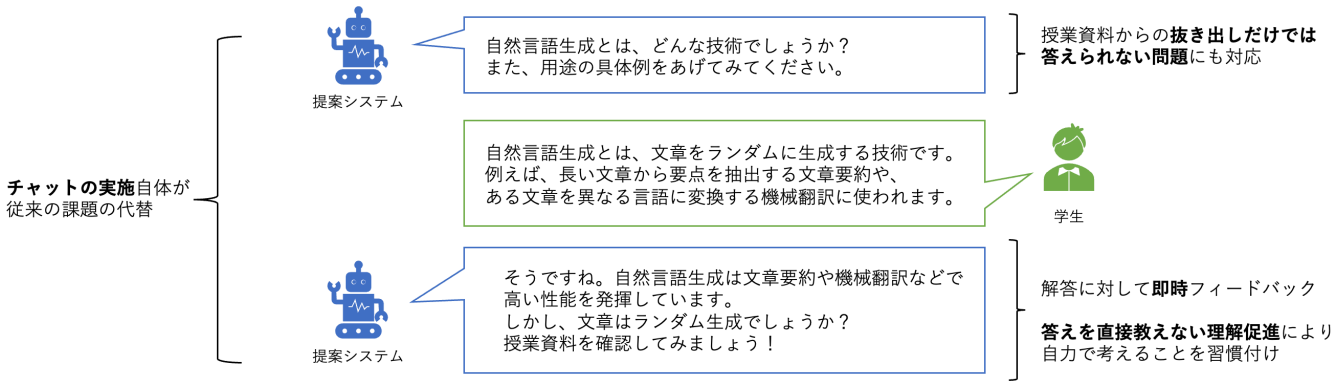


図 1: 提案するチャットシステムと学生の対話例： 中央の三つの吹き出しは, システム - 学生間で交わされる一連の対話を表している． 図に示したのは, 学生が設問に対して部分的に正解しているが不完全な解答をしている例である.

- 学生自身の思考を促すフィードバック文を生成する機能

本研究では, これらの機能を以下のように実現した．システムは授業内容に関する質問を事前に準備しておき, 対話が始まると学生に出題する．そして学生から解答を受け取ると, 質問を構文解析して採点項目別の小問に分割する．この操作には, 学生が間違えた箇所をより詳細に把握する狙いがある．次に採点モデルを用いて, 小問ごとの観点から学生解答の正誤を判定する．採点モデルとは事前に訓練された言語生成 AI であり, 事前に用意していた問題と関連する授業資料, 学生解答を入力して, 採点を行うように求めると, 以下の 2 点を出力する．

- 小問に対するおおまかな理解度を表す採点結果（正解 or 部分的に正解 or 不正解の 3 種類）
- 学生の理解不足がどこにあるかを特定する採点理由

これらの出力は, フィードバック（FB）生成モデルに入力される．このモデルは言語生成 AI を利用し, 採点結果や理由に応じて学生を導くヒントや問題に対する補足情報を生成する．生成した文章は, 事前に準備したテンプレートに当てはめられ, 学生に提供される．テンプレートと言語生成 AI を併用することで, 学生に返されるフィードバックの量や内容を平等に保ちながら, 理解度に合わせたアドバイスの作成を目指す.

本研究の貢献を以下にまとめる.

- 学生の間違い箇所を詳細に特定するため, 構文解析を用いて質問文を小問に分割する手法を提案する．(3. 1 節)
- 採点モデルを利用して学生解答の正誤を判定し, その判定理由を生成する方法を提案する．(3. 2 節)
- 正誤判定・理由をもとにプロンプト（AI に与える指示文や質問文などの入力のこと）を分岐し, フィードバックを生成する FB 生成モデルを提案する．(3. 3 節)
- 既存データセット上での提案システムの性能評価とケーススタディを行っている．(4. 2 節, 5 節)

2 関 連 研 究

2. 1 自 動 採 点

本研究以前にも, 教員の負担を軽減するため, 様々な自動採点手法が模索されている．この研究分野は学生解答の性質によって, 「小論文の自動採点」と「記述式解答の自動採点」の二つに分類される.

2. 1. 1 小論文の自動採点

小論文の自動採点（Automated Essay Scoring / AES）は, 複数の文章が組み合わさった長い解答を対象とした自動採点手法である．小論文は 2. 1. 2 節で紹介する記述式答案と比べて, 書き手の意見など主観的な要素を多く含む傾向がある．そのため, 文法や論理展開の観点からの採点が主である.

2020 年ごろまでは, 文章から文法や意味内容などに関する特徴量を抽出し, 回帰モデルやニューラルネットワークを用いて分析する手法が用いられていた [17]．しかし近年では, 大規模言語モデルを利用する方法が主流となっている．例えば Wang らは, 小論文をトークンスケール, セグメントスケール, ドキュメントスケールの 3 種類の表現ベクトルに変換し, それぞれを BERT で処理することで解答に点数をつける手法を提案している [24]．また BERT と LSTM を組み合わせて, 文章全体としての観点と文単位の局所的観点両方からの評価を行う手法も開発されている [11]．また, 言語生成モデルを利用した自動採点手法もいくつか登場している．具体的な例として, OpenAI の text-davinci-003 や GPT-4 に, 採点基準やその細かい例などを与えて採点を行った研究が報告されている [14] [25].

2. 1. 2 記述式解答の自動採点

記述式解答の自動採点（Short Answer Scoring / SAS）は, 1 文から数文で構成された比較的短い記述解答を対象とした自動採点手法である．小論文が主観的な要素を多く含むのに対し, この分野の採点対象には事実に対する正確で完全な記述など, 客観的に採点される要素が多く含まれる傾向がある．本研究が対象にしているのは, この種類の解答である.

Haller ら [5] によればこの分野の研究は, WordNet [3] や知識ベースの活用が始まりである．その後, 主要なアプローチは手動による特徴量抽出を用いた機械学習へと移行し, さらには単語の埋め込み表現技法が発展し, RNN や LSTM などの DNN を利用した手法へと変遷している．最近では, 大規模言語モデルの利用が増加しており, 例えば Takano らは読解問題に対する BERT を用いた採点手法を提案している [19]．Takano の研究では問題に対する解答文から, 事前に用意された採点基準に

対応する部分を BERT を用いて特定する．各部分文章は対応する採点基準ごとに, 異なる表現ベクトルに変換される．また, GPT-4 に few-shot 学習を適用した採点手法も試みられ, 学習データにない問題に対しても高い性能で正確な採点が可能であると報告されている [7]．さらに読解問題以外にも, 一般的な科学分野や電気・量子分野に関する問題に対する GPT-4 による採点手法 [10] や, 科学・生物学分野の問題に関して GPT-3.5 を利用して自動採点を行う方法 [26], 数学の問題採点に MathBERT を応用する方法 [27] も提案されている．

2.2 フィードバック生成

説明可能な人工知能に対する社会的需要は年々高まっている．この需要の高まりは, 教育支援を目的とした自然言語処理システムでも同様である．特に学生の評価に関わる採点システムの場合, その採点理由を説明できることは必須要件とされている [9]．よって自動採点の分野では, 採点結果と共に採点理由やフィードバック文も合わせて生成する手法が考案されている．例えば AES では, BERT による学生解答の点数予測を行い, その結果をもとに言語生成モデルでフィードバック文を生成する研究が行われている [6]．SAS でも同様の取り組みが行われており, 例えば Inui らの研究は, 読解問題に対する学生の解答から, 事前に設定した採点項目に対応する部分をニューラルネットワークで抽出することで, 学生解答のどの部分がどの採点基準を満たしていないのかを説明可能にしている [9].

最近の研究では, 対話システムを活用した新しいフィードバック手法も提案されている．例えば Huang らの提案するシステムは, 学生に問題を対話形式で提示し, その解答をもとに生成したフィードバック文で, 学生が正しい答えに辿り着けるように誘導する [8]．このシステムは, ヒントを生成する対話システムという点で本研究の提案システムと共通している．ただし, 本研究の提案システムは読解問題だけでなく, 例えば図 1 のように, 授業資料から文章を抜き出すだけでは解答できない問題も採点対象としている．また Huang らの研究ではあまり性能が出なかったと報告されているフィードバック自動生成部分も, 学生解答を採点するプロセスやフィードバック生成プロンプトの工夫によって, 性能向上を目指している．

3 提案システム

提案システムの全体を図 2 に示す．また以下に質問モデル, 採点モデル, FB 生成モデルの詳細を示す.

3.1 質問分割モデル

質問分割モデルは, 構文解析などを用いて質問文を採点項目別に複数の小問に分割する．その処理内容を図 3 に示す．ただし本モデルでは, 文章の最後の単語が*?*である文章を疑問文, 文全体の主たる動作や主題を示す述語動詞が対応する主語を持たない文章を, 命令文とみなしている．

質問を分割せずに採点モデル（3.2 節）に入力すると, 質問文の中に複数の指示が含まれていた場合, そのどれかに関する評価が行われなかったり, 行っていても間違ってしまう可能性が高くなる．そこで質問を要素ごとの小問に分割して採点を行うことで, このような採点ミスを削減することを狙っている．

3.2 採点モデル

採点モデルは図 2 に示される採点用プロンプトを受け取り, 採点結果とその理由を出力する言語生成モデルである．採点結果は正解, 部分的に正解, 不正解のいずれかで, 理由は数文で簡潔に表される．なおこの採点プロセスは, 3.1 節の質問分割モデルで得た小問ごとに実行される．

採点結果は学生の理解度を表しており, FB 生成モデルはその理解度に基づいて学生に提供する文面を調整している．一方, 採点理由は学生の間違い箇所を詳細に特定することを目的としている．この内容に沿ってフィードバックを生成することで, 学生の理解内容にピンポイントで寄り添ったメッセージの出力を目指す.

3.3 FB 生成モデル

FB（フィードバック）生成モデルでは, 事前に用意した FB テンプレート（図 4）と言語生成 AI を組み合わせて利用することで文章を生成する．テンプレートは「正解した小問の開示」「間違いの有無」「学生解答に合わせたメッセージ」の 3 つの部分から構成され, 小問の正答状況によってそれぞれの内容が変化する．このモデルで作成された文章が, システム全体の出力として学生に返却される.

テンプレートを用いずにフィードバックの全文を言語生成 AI だけで作成すると, 生成に時間がかかり, 内容にも大きなばらつきが出てしまう．そこで事前に文章の型を決めておき, 学生ごとに変更が必要な部分だけを生成することで, 生成時間を短縮し, 内容も均一に保てるようにしている.

4 採点性能の評価実験

提案システムの性能評価のため, 実装と既存データセットを用いた採点（正解・部分的に正解・不正解の三値分類）を行い, 正解率を算出した.

4.1 実　　装

4.1.1 構 文 解 析

質問分割モデル (3.1 節) の実現には, 文章を一文ずつに区切る, 疑問文の発見, 命令文の発見, 構文解析の 4 つの処理が必要である．これらの実装には, 多様な言語解析機能を提供している Stanza [16] というライブラリを利用した．Stanza の機能のうち, 文章を一文ずつに区切るために利用したのは, TokenizeProcessor である．これは, 文章を一文ごとに区切り, 同時に単語分割も行う．単語分割の結果, 最後の単語が*?*である文を疑問文とみなした．命令文の発見には DepparseProcessor を用いた．これは単語同士の依存構造解析を行うものである．DepparseProcessor が Root とラベルを付けた動詞を, 文全体の主たる動作や主題を示す述語動詞とみなし, それに対応する主語がない文章を, 命令文と判断している．構文解析は ConstituencyProcessor で行った．これは文章を Penn Treebank [12] に沿った木構造に

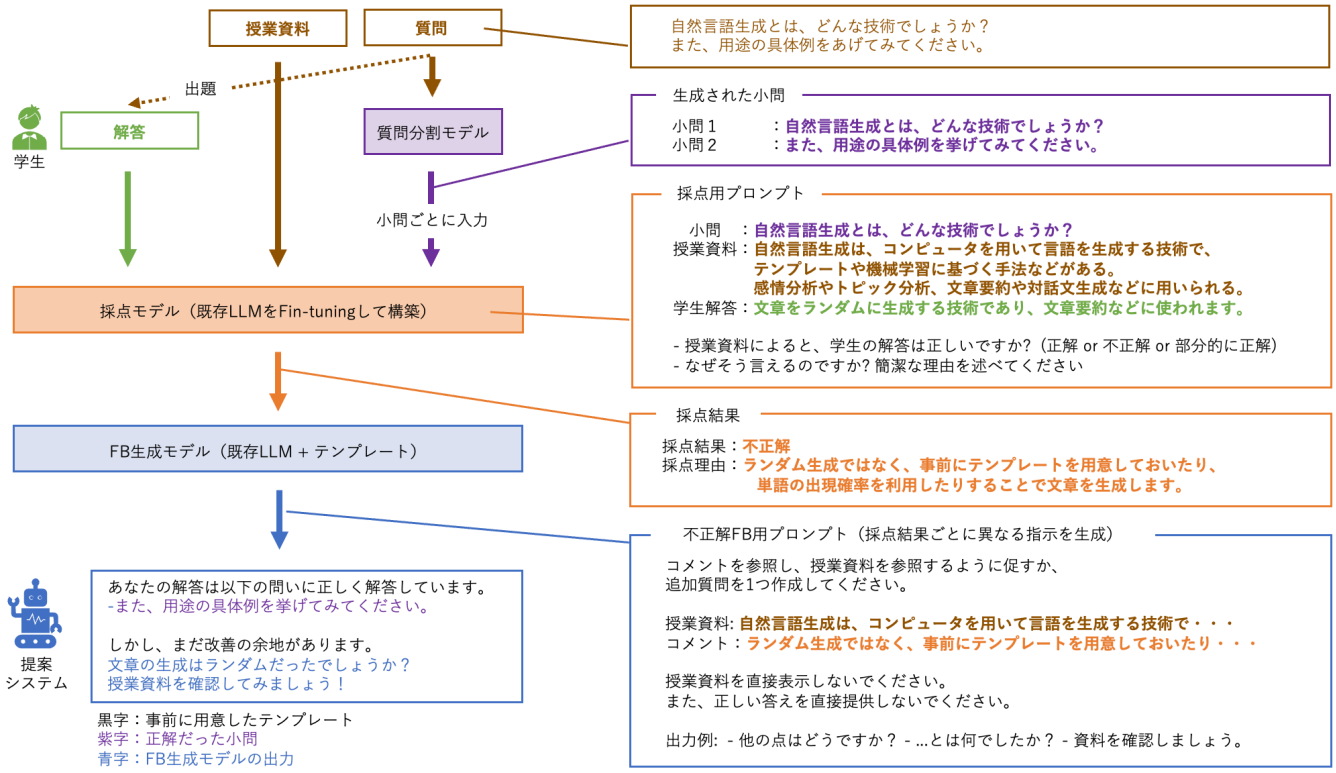


図 2: 提案システムの全体図: 提案システムは, 事前に用意した授業資料, 問題, それに対する学生からの解答を受け取ると, 質問分割モデル, 採点モデル, フィードバック (FB) 生成モデルの 3 段階を用いて. 解答に応じたフィードバックを生成する. なお, 図中では読みやすさのため, プロンプトや入力データを日本語で示しているが, 実際に利用したデータは全て英語である.

変換する. このとき and と親要素が同じ単語のうち, *and* の直前の単語と品詞が同じものを, and 同士で結ばれた単語だと判定している. 例えば, 図 3 の *advantage, disadvantage* の品詞は共に NNS である.

4.1.2 データセット

採点モデルを訓練するにあたって, Short Answer Feedback dataset（SAF）[4] を利用した. このデータセットには大学レベルの通信ネットワークの授業に関する問題とその模範解答, 学生の解答, そしてそれに対する採点結果ラベル（正解・部分的に正解・不正解）と, 採点者による判断理由などが含まれている. データセットは英語とドイツ語で構成されているが, 本実験では英語のデータのみを利用した. なお, SAF データセットは採点結果ラベルごとのデータ数の偏りが大きい. このため訓練データと検証用データの拡張を行い, すべてのラベルのデータ数を最も個数の多い「正解」の数に合わせた. 拡張時にはまず, 部分的に正解のラベルがついたデータをランダムに 1 件選び, それを Paraphrase-Generator [1] を用いて言い換えたものをデータに追加する. これを部分的に正解のラベルがついたデータが, 正解のデータと同数になるまで繰り返す. これと同様の操作を不正解のデータに対しても行った. この操作により, 訓練データは各採点結果 1053 件ずつ計 3159 件, 検証用データは各採点結果 262 件ずつ計 786 件となった.

4.1.3 採点モデルの学習

採点モデルのベースとなる言語生成 AI には, Zephyr-7B-β [21] を利用した. これは HuggingFace 社が発表したオープンソースの言語生成 AI で, 学習データの透明性と小規模なモデルサイズを実現しつつ, 一部のタスクで Llama2-Chat-70B [20] の性能を超えていることが報告されている [21].

実験にあたって, 問題・学生解答・授業資料を模擬した文章を与えられると, 学生解答の採点結果とその理由を出力するように, Zephyr-7B-β をファインチューニングした. 学習用プロンプトには, 拡張済みの SAF データセットから取得した問題文・学生解答・模範解答と, 学生解答を採点するように命令する指示文を含めた. ただし, 模範解答は授業資料の代替として利用している. ファインチューニングには QLoRA [2] という手法を用いた. QLoRA は 4-bit 量子化を利用して, 性能を保ちつつ学習に必要な GPU メモリを削減するものである. ファインチューニングは 2249steps 分実施した.

4.2 採点結果判定性能評価 - 結果と考察

SAF データセットの訓練・検証データには含まれない問題から構成されたテストセット上で, 学習済みの採点モデルによる自動採点を実行した. 得られた出力のうち, 採点結果（三値分類）の内訳を, 図 5 に示す. 「不正解の小問がある」解答を不正解, 「不正解はないが部分正解の小問がある」解答を部分的に正解, 「全ての小問に正解」している解答を正解と見做すと, 三値分類の正解率は 54%である. 特に多く発生した採点ミスは, 教師ラベルは正解なのに, システムの判定が部分正解となるものである. ついで教師ラベルが部分的に正解なのに, システムの判定が正解もしくは不正解となるミスが多い. ただし, 正解

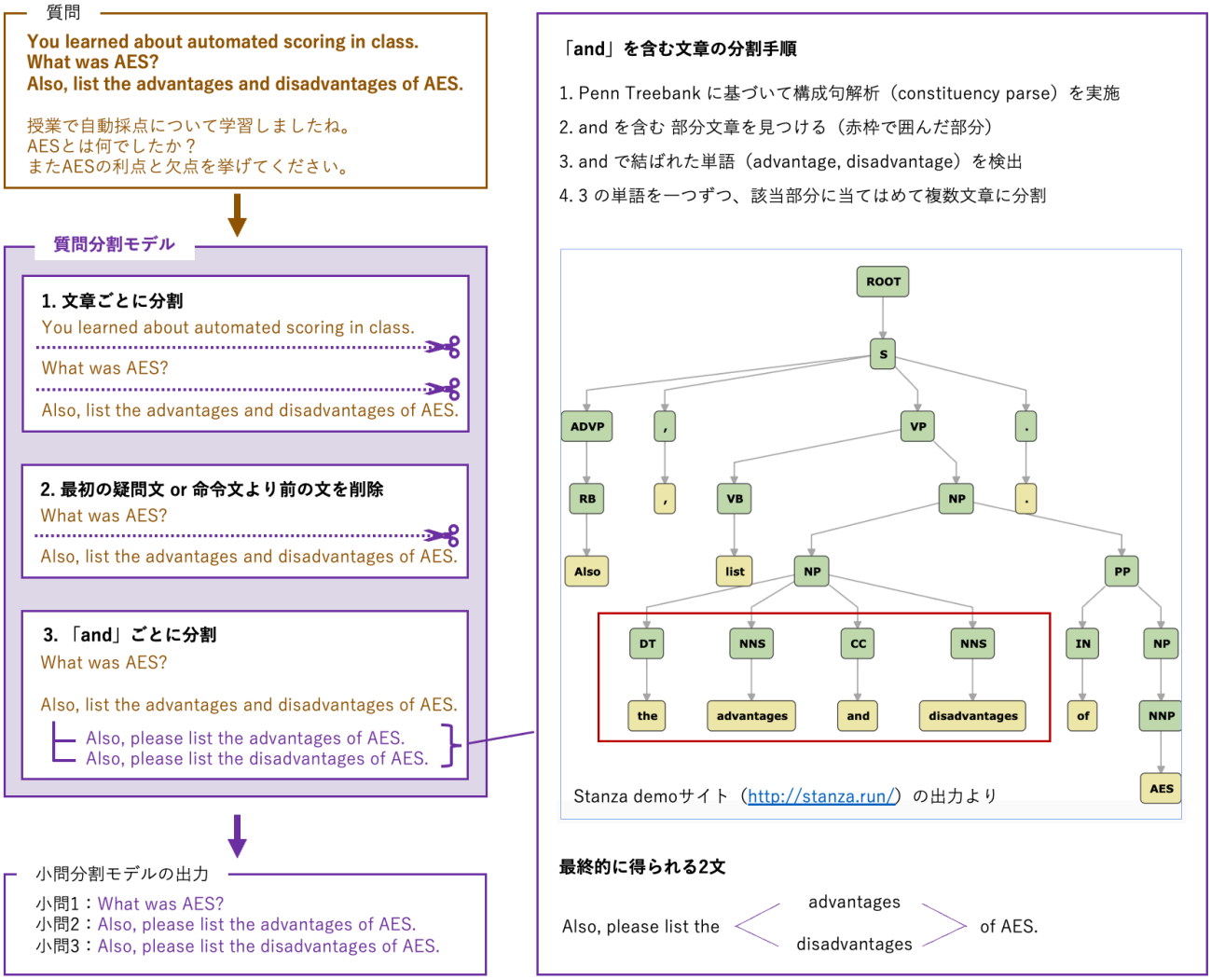


図 3: 質問分割モデル: 質問分割モデルは入力として質問文を受け取ると, まずはそれらを 1 文ごとに分割する. 次に, 最初に出現する疑問文もしくは命令文を見つけ, その文章より前の情報を削除する. これは, 図中の例の一文目のような, 問題の前提条件を提示するだけで採点項目を含んでいない文章が, 小問として出力されないようにすることを目的としている. 前提条件は通常, 最初の疑問文より前に書かれているため, このような処理を導入した. また, 前提条件の排除後には *and* を含む文章を分割する. これによって, 一つの小問には一つの質問項目や指示しか含まないようにし, 採点精度を上げることを狙っている. なお Penn Treebank については [12] を参照.

の解答を不正解と判定したケース, 不正解の解答を正解と判定したケースは共に 0 件に抑えられている. このような採点ミスは, 解答とフィードバックの大きな乖離を引き起こす可能性が高い. よってこれらの件数が低いことは, フィードバックの品質向上のために重要である.

5 ケーススタディ

4 章では提案システムを実装し, テストセット上での採点結果の三値分類の精度を検証した. 本節では, 採点理由と FB 文に焦点を当てたケーススタディを実施する. このケーススタディでは, 提案システムの実用性と効果をより深く理解するために, 提案システムがテストセット上でどのように機能するのかを確認し, 採点モデルや FB 生成モデルの妥当性と効果を評価する.

調査対象としたのは, テストセットから選択した 1 問（図 6）に対する 3 個の学生解答である.

5.1 学生の解答が正解の例

図 7 は学生解答が問題に正しく答えている例である. この解答に対して採点モデルは, 質問分割モデルで作られた小問ごとに, オレンジの文字で示した内容を出力している. 小問 1, 2 共に採点モデルは採点結果を正解と判定している. また採点理由も, 小問の内容に則した妥当な評価を述べている.

続いて, FB 生成モデルの出力を確認する. 黒字の部分は事前に用意した固定の文章である. この例では, 全ての小問に対する採点結果が正解であるため, 図 4 中の一番左の文章が選ばれている. 紫字の部分は質問分割モデルによって得られた分割後の問題文である. ここには正解した小問の問題文が出力されている. 最後に, 青字で示された部分が言語生成 AI からの出力である. この AI は 学生が全ての小問に正解した場合, 質問に

	小問の採点結果が 全て正解の場合	小問の採点結果に 部分的に正解が含まれる場合	小問の採点結果に 不正解が含まれる場合
正解した小問の開示 （なければスキップ）	あなたの解答は、以下の小問に対して正しく解答しています。 - [ここに、採点モデルの判定が「正解」になった小問を、箇条書きで列挙]		
間違いの有無	求められる全ての要件を満たしていますね。 よくできました！	ただし、あなたの解答にはまだ改善の余地があるようです。	
学生解答に合わせた メッセージ	以下のプロンプトで生成した メッセージを提示 授業資料の復習すべき点を箇条書きで簡潔に要約してください。 解答に重要な点が欠落している場合は、それを要約に含めてください。 授業資料：[事前準備した授業資料] 学生解答：[学生が入力した解答]	以下のプロンプトで生成した メッセージを提示 コメントを参照し、授業資料を参照するように促すか、追加質問を1つ作成してください。 [部分的に正解の小問がある場合] 資料: [class_material] 採点コメント： 部分正解小問の採点理由を全て列挙 [不正解の小問がある場合] 資料: [class_material] 採点コメント： 不正解小問の採点理由を全て列挙 授業資料を直接表示しないでください。また、正しい答えを直接提供しないでください。 出力例: - 他の点はどうですか？ - ...とは何でしたか？ - 資料を確認しましょう。	

図 4: FB 生成テンプレート: 最終的なフィードバック（FB）文は, 正解した小問の開示・間違いの有無・学生解答に合わせたメッセージの 3 つの部分から構成される．うち最初の二つは, 採点モデルが各小問にどんな出力をしたかによって選択される．3 つ目は, 図中に記載のプロンプトを言語生成 AI に渡し, その出力を FB 文に挿入することで得る

		採点モデルの出力		
		全ての小問に 正解	不正解はないが 部分正解の 小問がある	不正解の 小問がある
教師ラベル	正解	**138（28%）**	109（22%）	0（0%）
	部分的に 正解	45（9%）	**97（20%）**	48（10%）
	不正解	0（0%）	9（1%）	**33（6%）**

図 5: 採点結果の内訳: 縦軸が SAF データセット（テストセット）に記録された教師採点ラベル，横軸が提案システムの小問採点結果．三値分類の正解率は $28 + 20 + 6 = 54\%$.

関連する授業資料の要約を提示する．実際に青字部分は, 問題に関連する授業内容を正しくまとめることができている．このフィードバックは, 学生が自身の理解度と学習進捗を確認するために有益と思われる．

5.2 学生の解答が部分的に正解の例

図 8 は学生解答が部分的に正解である例を示している．この例の解答は二つのモード名を正しく述べている．しかし Replication の欠点を挙げておらず, また Partitioning の欠点である *coordination and consistency* を利点として説明してしまっている．この解答に対して採点モデルは, 小問 1 を正解, 小問 2 を部分的に正解と判定した．これは, 採点モデルが学生の大まかな理解度を小問レベルで正しく認識できていることを示している．しかし, 採点理由には問題点がある．例えば, 小問 1 の理由は「二つのモードの定義と欠点を正しく述べている」となっている．しかし学生解答が挙げた欠点には, 一

質問：
Name the two modes of control plane distribution and name one downside for each of them.
（制御プレーンの物理的な分散の二つのモードを挙げ、それぞれのデメリットを一つずつ挙げてください。）

模範解答：
Two modes of physical distribution of control plane and associated drawbacks are:

Replication
● Issue of scalability
● More resources, more cost
● Wastage of resources as the replicated node is on standby and no actual load distribution occurs.
● The problem to keep the data consistent in the replicated node is hard and complex.

Partitioning
● Raises issues similar to P2P networks
● Knowledge about neighborhood
● Coordination and consistency
● Lower availability

（コントロールプレーンの物理的な分散の二つのモードとそれに関連する欠点は次の通りです：

レプリケーション
● 拡張性の問題
● より多くのリソースが必要で、それに伴うコストがかかる
● レプリケートされたノードが待機状態であり、実際の負荷分散が行われないため、リソースの浪費が発生する
● レプリケートされたノードでデータを一貫性を保つ問題が難しく、複雑である

パーティショニング
● P2Pネットワークと同様の問題が発生する
● 近隣に関する情報
● 調整と一貫性
● 低い可用性）

図 6: ケーススタディで用いる問題とその模範解答: SAF データセット [4] より転載．システムに入力したのは英語のみだが, 読みやすさのため日本語訳した文章も併記している．

部に不足や間違いがあるため, これは不正確である．また, 小問 2 の理由は「学生は Replication の欠点を述べている」とあるが, 学生は実際には欠点を述べていない．さらに小問 2 の理由には, Partitioning の欠点についての指摘はあるが, それが *coordination and consistency* に関する間違いであるとは名言

学生解答:
Replication: Double the cost of building and maintaining a single controller.
Partitioning: The different parts of the controller must coordinate with each other like in peer to peer networks which could reduce speed or consistency.

小問ごとの採点結果:
小問1：Name the two modes of control plane distribution.
結果： Correct
理由： The response is correct as it correctly explains the two modes of physical distribution of control plane.

小問2： name one downside for each of them .
結果： Correct
理由： The response correctly states the downside of replication and partitioning.

フィードバック文:
"Good job! Your answer correctly answers the following subquestion:
- Name the two modes of control plane distribution.
- name one downside for each of them.
You have fulfilled all the requirements. Well done!
Below is a lecture summary for this question: -
- Replication: Higher cost due to resource wastage, scalability issue, and consistency maintenance complexity.
- Partitioning: Similar issues to P2P networks, neighborhood knowledge required, coordination and consistency challenges.

図 7: 学生の解答が「正解」の場合の代表的な出力

学生解答:
The two modes of control plane distribution are replication and partitioning.
On replication, it has OpenFlow standard which allows for connections from switch to multiple controllers.
Partitioning supports large-scale deployment in which it raises issues similar to peer to peer networks, it contains knowledge of the neighborhood and it has good coordination and consistency.

小問ごとの採点結果:
小問1：Name the two modes of control plane distribution.
結果： Correct
理由： The response answers the definition and the drawbacks of replication and partitioning correctly.

小問2： name one downside for each of them.
結果： Partially correct
理由： "The response contains the following downsides: replication waste of resources and issue of scalability. The correct downside for partitioning is the raised issue of neighborhood knowledge.

フィードバック文:
Great! Thank you for answering. Your answer correctly answers the following subquestion:
- Name the two modes of control plane distribution.
However, there is still room for improvement.
- Can you identify a potential drawback of partitioning that is not mentioned in the given material?
- What is an example of a neighborhood knowledge issue in partitioning?
- Let's review the section on partitioning in the textbook.

図 8: 学生の解答が「部分的に正解」の場合の代表的な出力

されていない．学生の間違いを見逃している小問 1 のような採点理由や，間違っている箇所には言及しているものの，それが具体的にどんな間違いだったのかをピンポイントで指摘できない小問 2 のような採点理由は，学生解答が「部分的に正解」や「不正解」である場合に多く確認された．

次にフィードバック文を確認する．青字部分（言語生成 AI が作成した文章）は，学生に Partitioning について再考を促す 2 つの質問と，授業資料を見返すように促す文章で構成されている．これらのヒントは，学生に直接答えを教えることなく，自力で正解に辿り着けるように誘導するものである．従来の模範解答を提示するだけの方法と比較して，このフィードバック文は学生に自力で考える余地を与えることで，理解を深め，思考力を高める効果が期待できる．しかしこのフィードバック文にも，いくつか問題点がある．まずこの文章には，Partitioning に関するヒントだけが含まれており，Replication に関する間違いは触れられていない．また一つ目のヒントには「授業資料に載っていない欠点を挙げてみましょう」といった，問題の趣旨と外れる提案が含まれている．これらの原因は，採点理由の生成ミスがフィードバック文に反映されていることである．先ほど確認した通り，生成された採点理由は Replication に関する記述漏れを見逃している．また Partitioning の欠点に関しても，言及はあるものの，それが具体的にどんな間違いだったかまでは特定していない．これらの影響を受けた結果，生成されたヒントは，学生の間違いとは異なる内容を指摘していると考えられる．このように，採点理由のミスは後続の処理に大きな影響を与える．よって今後，更なる採点理由生成手法の改善が求められる．

5.3 学生の解答が不正解の例

学生解答:
The control plane can be distibuted or centralized.
Distributed needs multiple steps to converge to a perfect solution.
Centralized needs to share the path with the router.

小問ごとの採点結果:
小問1：Name the two modes of control plane distribution.
結果： Incorrect
理由： The response is incomplete as it does not state the two modes of distribution.

小問2： name one downside for each of them .
結果： Incorrect
理由： The response is incomplete as it does not state the centralized problem correctly and also the distributed problem.

フィードバック文:
Thank you for your nice answer. However, there is still room for improvement.
- Can you provide an example of a distributed problem in partitioning?
- Let's review the section on replication and identify the drawbacks of this mode of distribution.
- What are the similarities and differences between partitioning and P2P networks?
- Can you summarize the challenges of keeping data consistent in a replicated node?
- Let's discuss the advantages and disadvantages of replication and partitioning in more detail.

図 9: 学生の解答が「不正解」の場合の代表的な出力

図 9 は学生解答が不正解である例を示している．この例の解答は，モード名として Distribute と Centralize という誤った単語を挙げている．これに対して採点モデルは，小問 1, 2 を両方とも不正解と判定した．よって学生の大まかな理解度は，採点結果の三値分類によって正しく推定されていると言える．また小問 1 の採点理由は，「正しいモード名を述べていない」という学生の間違いをピンポイントで指摘できている．しかし一方，小問 2 の採点理由は「学生解答は Centralize の欠点を述べていない」ことに焦点を当てている．本来，学生が述べるべきなのは Replication と Partitioning の欠点であるため，これは不適切である．つまり，採点理由が質問の趣旨から外れている．このような例は特に，学生の解答内容が模範解答とかけ離れている場合に多く確認された．この内容がフィードバック文に反映されると，学生は間違った方針のヒントを受け取ることになり，学習支援の妨げとなる可能性がある．

続いてフィードバック文の青字部分を確認する．この部分は，Replication と Partitioning の欠点について確認する内容になっている．この文章は図 8 の場合と同様に，再考を促すための追加質問と，資料を見返すように提案する文章によって構成されており，学生を正解へと間接的に誘導しようとしている．これは，学生の自力での課題解決の促進につながる可能性がある．しかし，ヒントの内容が完全には学生の理解度に則していないという課題点もある．例えば提示された質問文の中には，学生が答えられなかった Replication や Partitioning という単語が直接含まれている．学生に自力で正解に辿り着かせるためには「二つのモードは本当に Distribute と Centralize で正しいでしょうか？」のように，答えとなる単語を含まない尋ね方をするのが理想である．

6 おわりに

本研究では，学生が自らの解答を振り返り，自力で理解度を向上させるためのシステムを開発した．提案システムは，学生の

解答を分析することで誤答の原因を特定し, 正解への間接的なアドバイスを提供する. 学生はこのシステムを活用することで, 自身の学習状況を把握し, 自力で学習を進めることが可能となる. 同時に教員側においても, 学生個々に合わせた指導を行う負荷が軽減されることが期待される.

実験では提案システムを実装し, その有効性と実用性を評価した. その結果, 提案システムは 54 %の精度で解答を採点できることが確認された. またケーススタディでは, 採点理由やフィードバック文が, 学生の解答内容に則したヒントや補足情報を提供することが示された. これらの機能は, 学生が自力で学習を進めるための有効な手段となりうる. 今後, ユーザ実験などを通して, その効果の検証が求められる. また実験結果や事例で示されたように, システムの性能改善も課題である. 提案システムのさらなる改良により, 教員の負担軽減と学生個々への効果的な自己学習機会提供に貢献することが期待される.

謝　　辞

本研究の一部は科研費（23H03404）による.

文　　献

[1] Sai Vamsi Alisetti. Paraphrase-generator. `https://github.com/Vamsi995/Paraphrase-Generator`, 2020. Accessed on January 24, 2024.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[3] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

[4] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8577–8591, 2022.

[5] Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*, 2022.

[6] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*, 2023.

[7] Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. Can llms grade short-answer reading comprehension questions: Foundational literacy assessment in lmics. *arXiv preprint arXiv:2310.18373*, 2023.

[8] Jin-Xia Huang, Yohan Lee, and Oh-Woog Kwon. Direct: Toward dialogue-based reading comprehension tutoring. *IEEE Access*, Vol. 11, pp. 8978–8987, 2022.

[9] Kentaro Inui, Yutaka Ishii, Yuichiroh Matsubayashi, Naoya Inoue, Shoichi Naito, Yoriko Isobe, Hiroaki Funayama, and Seiya Kikuchi. Frontiers in explainable automated writing evaluation. *IEICE ESS Fundamentals Review*, Vol. 16, No. 4, pp. 289–300, 2023.

[10] Gerd Kortemeyer. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *arXiv preprint arXiv:2309.09338*, 2023.

[11] Feng Li, Xuefeng Xi, Zhiming Cui, Dongyang Li, and Wanting Zeng. Automatic essay scoring method based on multi-scale features. *Applied Sciences*, Vol. 13, No. 11, p. 6775, 2023.

[12] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[13] Lucas Mearian. Schools look to ban chatgpt, students use it anyway. `https://www.computerworld.com/article/3694195/schools-look-to-ban-chatgpt-students-use-it-anyway.html`, 2023. Accessed on January 7, 2024.

[14] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, Vol. 2, No. 2, p. 100050, 2023.

[15] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, 2023. Accessed on December 28, 2023.

[16] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.

[17] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, Vol. 55, No. 3, pp. 2495–2527, 2022.

[18] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[19] Shunya Takano and Osamu Ichikawa. Automatic scoring of short answers using justification cues estimated by bert. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 8–13, 2022.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

[22] Carnegie Mellon University. Student's don't seek help when they need it. `https://www.cmu.edu/teaching/solveproblem/strat-dontseekhelp/index.html`. Accessed: February 13, 2024.

[23] Tal Waltzer, Riley L Cox, Gail D Heyman, et al. Testing the ability of teachers and students to differentiate between essays generated by chatgpt and high school students. *Human Behavior and Emerging Technologies*, Vol. 2023, , 2023.

[24] Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*, 2022.

[25] Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating short l2 essays on the cefr scale with gpt-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 576–584, 2023.

[26] Su-Youn Yoon. Short answer grading using one-shot prompting and text similarity scoring model. *arXiv preprint arXiv:2305.18638*, 2023.

[27] Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. Automatic short math answer grading via in-context meta-learning. *arXiv preprint arXiv:2205.15219*, 2022.

LLMを用いた対話型鑑賞エージェントにおける繰り返し発話の抑制機構

谷　悠翔†　牛尼　剛聡††

† 九州大学大学院 芸術工学府 メディアデザインコース　〒815–8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学院 芸術工学研究院　〒815–8540 福岡県福岡市南区塩原 4-9-1
E-mail: †tani.haruto.025@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　対話型鑑賞とは一つの絵画を複数人で鑑賞し，ファシリテータに意見を促されながら意見交換をすることで，対象とする絵画の理解を深めていく鑑賞方法である．対話型鑑賞では，鑑賞のポイントを指摘したり，知識を教授するファシリテータの存在が重要な役割を果たす．しかし，鑑賞時に絵画の専門的な知識を持ったファシリテータが常に身近に存在しているわけではないため，ユーザが希望するタイミングで対話型鑑賞を行うことは困難である．そこで，本論文では鑑賞者がいつでも好きなときにで対話型鑑賞を行うことができるファシリテータ AI エージェントを開発することを目的とする．ファシリテータの役割を果たす AI エージェントを GPT-4 にプロンプトを与えることで開発する場合，ファシリテータ AI が同じ発言を繰り返してしまうことが，鑑賞者の絵画理解を妨げることが問題となる．本論文ではこの問題を解決するために，鑑賞者と対話する AI の他に、対話ログを分析し改善点を提示するフィードバック用の AI を利用することを提案する．この手法ではフィードバックにより対話を行う AI の発話が適宜修正されることで，繰り返しの発話が抑制され，提案手法が絵画理解に寄与することを，被験者実験に示す．

キーワード　巨大言語モデル，対話型鑑賞，鑑賞支援

1　はじめに

対話型鑑賞とは芸術作品を複数人で鑑賞し，ファシリテータに意見を促されながら意見交換をすることで，対象とする絵画の理解を深めていく鑑賞方法である [1]．対話型鑑賞は，ファシリテータから鑑賞者へ投げかけられる「絵画の中で何が起こっているか」,「絵画のその要素からそう思うのか」,「それ以外に絵画の中に発見はあるか」の３つの質問を主軸に行われる．そしてファシリテータと複数の鑑賞者が対話を通して，絵画から何らかの意味を見出したり，表面的な感想をより詳細に具体化する過程で，自身の鑑賞を深めることができる．また，他の鑑賞者の視点や思考から新たな知見を得ることで，作品に対する理解を深め，絵画鑑賞体験の質を向上させることができる．そのため近年では，美術館来訪者や小・中学生の授業，さらに医学生のカリキュラム [2] など幅広い分野で注目されている．

芸術作品の鑑賞は，複数人で行うことにより独自の魅力や深い理解が得られる場合がある．しかし，物理的な共同鑑賞者が近隣に存在しない状況や，デジタルアーカイブを介したオンライン上の芸術作品閲覧の際には，単独での鑑賞をせざるを得ない場合がある．そうした状況下では，複数人との同時鑑賞の実現が困難である．また，作品に関する知識や芸術的な知見を有さない人物ら複数人で鑑賞が行われた場合，作品に関する意見が十分に発話されず，対話が発展しない可能性がある．そのため，対話型鑑賞では，作品に対して十分な知識を持ち，適切なタイミングで対話に介入し，鑑賞者間での作品理解を促すファシリテータの存在が重要である．しかし，一般的に鑑賞者の近隣にファシリテータが存在する場合は少ない．

そこで，本論文では，ユーザがシステム上で対話型鑑賞を行う状況を想定し，ユーザがファシリテータ役の AI エージェントと対話しながら芸術作品を鑑賞することで，ユーザ単独での対話型鑑賞を実現し，鑑賞体験を向上させるシステムを開発することを目的とする．

本論文では，ファシリテータ役の AI エージェントに注目し，その実現方法を検討する．AI エージェントの実現には GPT-4 [3] [4] を用いる．GPT-4 は大量のテキストデータを学習しており，その中には絵画に関する情報も含まれている．それら絵画に関する情報をファシリテータの前提知識として用いることができる．さらに，GPT-4 は大量のパラメータ数によって構築されいるため自然な発話が可能である．

しかし，予備実験としてファシリテータ役のエージェントの質問能力を確認したところ，鑑賞者に対して，繰り返し同じ内容の質問をしてしまうことが判明した．この繰り返し発話は鑑賞者の絵画理解を妨げてしまう．よって本論文では，この繰り返し発話を抑制する機構を考案する．

提案手法としては，鑑賞者とファシリテータ役のエージェントとの対話型鑑賞の対話ログをもとにファシリテータの質問内容を修正させるフィードバック用のエージェントを導入する．フィードバック用のエージェントが対話ログの校閲し，ファシリテータの繰り返しの質問が確認された場合，フィードバック用のエージェントが繰り返し発話の抑制を行うためのプロンプトを作成する．このプロンプトをもとにファシリテータは質問内容を訂正し，繰り返し発話を抑制する．

本論文では言語モデルを用いて作成したファシリテータを以

後，ファシリテータ AI，対話型鑑賞の対話ログをモニタリングしながら適切なフィードバックをファシリテータ AI に与えるエージェントをフィードバック AI と呼ぶ．

本論文の構成は次の通りである．2 章で関連研究について述べる．3 章では予備実験について説明する．4 章では提案手法について述べる．5 章で本論文のまとめを述べる．

2 関 連 研 究

これまでにも，対話型鑑賞に関する研究や，対話型鑑賞をシステムとして構築するための研究が行われてきた．ここではそれらのうち，提案手法に関連のあるものについて述べ，本論文の位置づけを明確にする．

2.1 ファシリテーションに必要な要素

平野ら [5] は対話型鑑賞において，ファシリテータが鑑賞者に対して行う，作品背景や絵画の知識などの情報提供に着目することで，作品を媒介として鑑賞者への知識構築を促し新たな作品理解を形成するためのファシリテーションの方略を検討している．図 1 はファシリテータが鑑賞者の作品理解を深めるための基本的な方針が示されている．ここでいう知識構築とは「学習者が自らのアイディアに基づき，学習対象としての知識を吟味し，協働的な理解として向上させ続ける活動」という考えを基に，図 1 に記された「結びつけ/コネクト」「まとめ/サマライズ」を鑑賞者が行えていることを指す．

平野らは知識構築が増加した事例として，ファシリテータが与える情報が「考えるための情報」に絞られていたこと，鑑賞のタイミングが 8 割程度進んだ状態で更なる情報提供が行われいたことを挙げている．一方，知識構築が減少した例としては，ファシリテータが与える情報が作者の身辺情報に偏っていたこと，鑑賞全体を通して，作品の表現内容を中心とした内容ばかりが議論されていたことを挙げられている．

本論文では，このファシリテータの基本方針を参考に，システムとして構築する対話型鑑賞モデルがファシリテータとしての能力を十分に有しているかを検証する．

2.2 エージェントを用いた対話型鑑賞システム

小倉ら [6] らは，ファシリテータと鑑賞者のエージェントを対話モデルとして作成し，そのモデル内のエージェント 2 役とユーザが 3 人で対話型鑑賞を行う手法を提案している．この研究では，対話型鑑賞を以下のような段階に分けている．

（1） Introduction: 初発の思考を引き出す
（2） Divergence: 意見を拡散させる
（3） Convergence: 意見を収束させる
（4） Conclusion: 対話内容をまとめる

この手法では事前に収集した対話型鑑賞の対話事例を意味解析し，発話の種類を識別する発話タグを付与したものを学習データとして用いて，有限オートマトンに学習させたものを対話モデルとして用いている．このモデルが対話型鑑賞の各段階に対応した対話を行う．

この手法では，対話型鑑賞における鑑賞者・ファシリテータの発話の流れが構造的に定義されているものの，エージェントとして用いた対話モデルの性能が自然な対話をするまでに達しておらず，そもそもの人とエージェントとの対話に破綻している部分があった．

カテゴリ	基準
受け答え／コメント	直前の鑑賞者の言葉に対して反応したもの．自分の意見を述べるものや，相槌や，鑑賞者の言葉をそのまま拾った「オウム返し」を含む．
言い換え／パラフレイズ	直前の鑑賞者の言葉に対して反応しつつ，それを似た言葉で言い換えたもの．
結びつけ／コネクト	直前の鑑賞者の言葉に対して，2つ以上前の鑑賞者の言葉を1つ引き出し，対比させるもの．
情報提供／インフォメーション	作家や作品についての背景知識を鑑賞者に伝えるもの．
まとめ／サマライズ	2つ以上前の鑑賞者の言葉を複数引き出し，鑑賞の中で出てきた事柄について筋道を立てて整理するもの．
質問／クエスチョン	VTS において重視される3種類の問いかけに相当するもの．「何が起こっているだろう？」「どこからそう思う？」「もっと発見はある？」
焦点化／フォーカシング	作品の中で注目する場所，話すべき話題を絞ることを提案するもの．

図 1 ファシリテーションの基本方針
(平野「対話型鑑賞のファシリテーションにおける情報提供のあり方」より)

本論文では，これらを踏まえて，対話型鑑賞の流れを構造的に分解してモデル化し，その上でエージェントに用いる対話システムの中核として巨大言語モデルである GPT-3.5 を使用することで，ユーザと対話システム間での自然な対話型鑑賞を提案する．

3 巨大言語モデルを利用したファシリテータ AI

GPT-4 は大量のテキストデータを学習しており，その中には絵画に関する情報も含まれている．それら絵画に関する情報をファシリテータの前提知識として用いることができる．また，GPT-4 に対話型鑑賞に用いる絵画の前提知識を学習させることも可能である．さらに，GPT-4 は大量のパラメータ数を有し，かつ RLHF による強化学習により信頼性の高い返答が可能である，そのためユーザに対して自然な発話を行うことができる．

しかし，ファシリテータ AI は自然な発話ができ，また絵画の知識を備えている一方で，絵画知識をもとに鑑賞者の絵画理解を適切に深めることのできるファシリテーションの能力は十分ではないと考えられる．そこで LLM に予め備わっているファシリテーションの能力を検証すべく，予備実験を行った．

モデルには OpenAI が提供する API の中から gpt-3.5-turbo [4] を使用した．モデルに与えるプロンプトとして「あなたはファシリテーター，私は鑑賞者です．鑑賞者が提示する作品について鑑賞者と対話型鑑賞をしてください．ファシシテータは鑑賞者の意見を引き出すよう努めてください.」という文をモデルに与えた．鑑賞者はファシリテータ AI とショートメッセージをするように対話型鑑賞を行った．絵画には図 4 のクロード・モネ《日の出，印象》を用いた．

予備実験の結果，単純なプロンプトによる AI を利用した場合，ファシリテータ AI が被験者に何度も同じ内容の質問を繰り返してしまうという問題点が挙げられた．具体的には，鑑賞者が絵画について「淡くて儚い印象」を述べた際，ファシリテータ AI はその「淡くて儚い印象」を絵画のどの部分から感じるかを質問する．その質問について鑑賞者が「太陽から感じる」返答した後，ファシリテータ AI は再び「太陽からどのような印象を受けるか」と言う質問を行った．このように，似た質問を繰り返すことは，ユーザにとって負担になる上に，話題が展開しないため，鑑賞者に新しい発見をもたらすことができない．

この問題については図 1 のファシリテータの基本方針に基づいて,「焦点化/フォーカシング」と「情報提供/インフォメーション」が重要であると考えられる.「焦点化」を行うことで，鑑賞者がまだ注目していない点をファシリテータ AI に指摘させることで話題の展開を作ることができるからである．また，「情報提供」をすることで鑑賞者が意見を述べるときのヒントを与え鑑賞者自身の広がりを促すことで新しい発見を導くことができる．これら二つはファシリテータ AI の発話の種類が増えることにもつながるため，同じ内容の質問を生成してしまうことを抑止できると考える．

4 提 案 手 法

4.1 提案手法の概要

予備実験の考察を踏まえて，ファシリテータ AI が同じ内容の発話を繰り返すことを抑制し，適切な対話型鑑賞を実現するためのモデルを開発した．図 2 は本モデルの概要を示し, 図 2 の流れに沿ってユーザとの対話型鑑賞は行われる．

- **ユーザとファシリテータ AI の対話**　ユーザとファシリテータ AI が対話型鑑賞を行う．ファシリテータ AI はプロンプトをもとに，ユーザに対して質問を行う．
- **対話ログの校閲/プロンプトの作成**　フィードバック AI は適宜，ファシリテータ AI の発言を校閲する．そして，繰り返しの発話が認められた場合，その発話内容を修正するためのプロンプトを作成する．
- **発言の修正**　ファシリテータ AI がフィードバック AI からプロンプトを受け取った場合，その内容をもとにユーザへの質問内容を変更する．

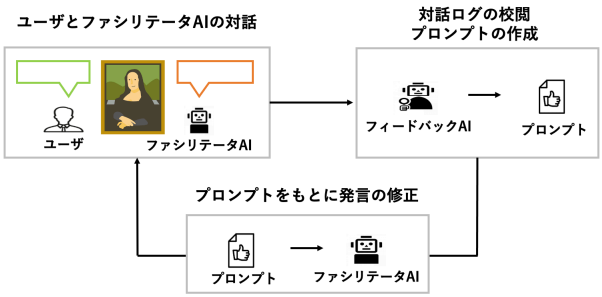


図 2　システムの概念図

4.2 AI エージェントを利用した対話型鑑賞

以下に，図 2 の内容を実装する際の LLM について説明する．

4.2.1 ユーザとファシリテータ AI の対話

ユーザと対話する LLM にプロンプトとして,「ファシリテータとしての役割」と「鑑賞する絵画の名称・作者の名前」を与えている．対話はファシリテータ AI の質問に対してユーザが返答する一問一答形式で行われる．

4.2.2 対話ログの校閲/プロンプトの作成

フィードバック AI は対話ログをもとに，対話ログの校閲と繰り返し発話抑制のためのプロンプトの作成を行う．

対話ログを校閲する際のプロンプトの内容を表 1 に示す．このプロンプトは対話ログ上でファシリテータ AI が繰り返し同じ質問をしているかを判定させるものである．まず，ファシリテータ AI が繰り返し同じ質問をしている状況を定義する．定義した内容をもとに，質問が繰り返しているかを判定する．対話ログは，ファシリテータ AI のはじめの質問から始まり，ファシリテータ AI の質問が修正されるまでに行われたユーザとの一問一答の内容を用いる．ファシリテータ AI の発言が修正されたのちは，修正後のファシリテータ AI の質問から，次の修正を受けるまでに行われたユーザとの一問一答の内容を対話ログとして用いる．

対話ログの校閲にて，ファシリテータ AI の繰り返しの質問が確認された場合，繰り返し発話の抑制を行うためのプロンプトを作成する．まず，フィードバック AI は，絵画に写っている事象・物体を列挙する．表 1 にモネの《日の出，印象》の場合に抽出される事象・物体の例を示す．フィードバック AI はこの物体・事象の中から，まだ対話として話題に上がっていないものを選びファシリテータ AI の質問を訂正させるために「対話型鑑賞のダイアログに基づいて，鑑賞者に新しい話題を提供するために，以下の中から要素を一つ選んでください．

…（抽出した話題）…

選んだ要素について鑑賞者に質問を投げかけてください.」というプロンプトを生成する．

4.3 実　　装

図 2 に示した対話型鑑賞の流れを python のフレームワークである langchain を用いて web アプリのバックエンドとして実装する．LLM には，GPT-4 を用いる．図 3 に提案手法の内容を web アプリのプロトタイプとして実装した際の画面を示す．画面右側には対話型鑑賞の題材になる絵画を表示している．画面左側にはトーク機能の UI を実装している．

4.3.1 発言の修正

ファシリテータ AI は，フィードバック AI から受け取ったプロンプトをもとに質問の内容を訂正する．これによりファシリテータ AI はユーザに対して新たな話題を提供し，質問の繰り返しを抑制する．

表 1　フィードバック AI のプロンプト内容

校閲	対話型鑑賞にてファシリテータの質問がループするとは **1,** 絵画についての「印象・感情」を質問 **2,** その感情の原因となる絵画内の要素について質問 **3,** 再びその要素についての「感情・印象」を質問することを指す。 このことに基づき、対話型鑑賞のダイアログについて a: 鑑賞者の意見に対して、ファシリテータの質問がループしている b: 鑑賞者の意見に対して、ファシリテータの質問がループせずに適切である を判定してください
絵画の物体・事象	例：クロード・モネ《日の出，印象》 **日の出：**絵画の下部には波の描写があり，海面が揺れている様子が表現されています．波の動きはモネの印象派の特徴の一つです． **波：**絵画の下部には波の描写があり，海面が揺れている様子が表現されています．波の動きはモネの印象派の特徴の一つです． **雲：**絵画の上部には雲の描写があり，空に浮かんでいる様子が描かれています．雲の形や色彩はモネが独自の視点で表現したもので，印象派の特徴の一つとされています．

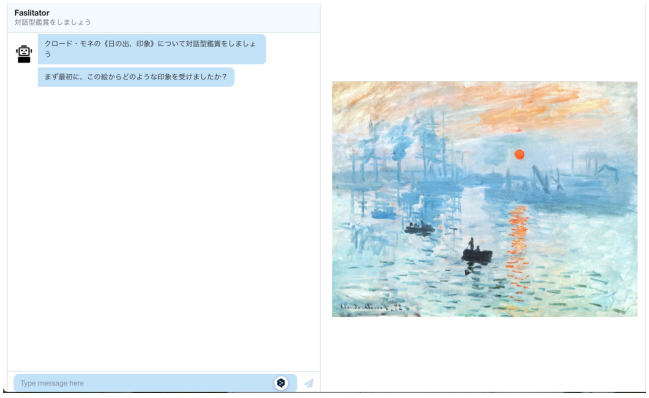


図 3　web アプリプロトタイプの実装画面

5　評 価 実 験

5.1　実 験 内 容

提案手法における対話システムによって，ファシリテータ AI の繰り返し質問を抑制し，効果的な対話型鑑賞が実現可能かを検証するために，評価実験を行った．比較手法としてファシリテータ AI のみを実装した web アプリを作成した．被験者は作品を鑑賞しながら，ファシリテーション AI の質問に従って，チャット形式での対話を行った．また，被験者は比較手法と提案手法の両方の手法による鑑賞を行った．鑑賞として用いる絵画には図 4 のクロード・モネ作《日の出，印象》と，図 5 フィンセント・ファン・ゴッホ作《星月夜》を用いる．二つの絵画は同時期に製作されており，両者とも印象派絵画に大別されるように作風が似ているため，評価実験の絵画として採用した．評価実験は二つのグループに分けて行った．これにより，提案手法が絵画による有意差がないことを示す．1 つ目のグループは提案手法の絵画として図 4 の《日の出，印象》を用い，比較手法として図 5 の《星月夜》を用いた．大学生男女 11 名の被験者に実験を行った．この被験者群をグループ 1 と呼ぶ．2 つ目のグループは提案手法の絵画として図 5 の《星月夜》を用い，比較手法として図 4 の《日の出，印象》を用いた．大学生男女 9 名の被験者に実験を行った．この被験者群をグループ 2 と呼ぶ．各被験者は対話型鑑賞後，アンケートに回答した．アンケートの質問内容を以下に示す．

以下，3 つの質問内容をアンケートで聴取した．

- **同じ質問の繰り返しの影響**　ファシリテータ AI による同じ質問の繰り返しが，鑑賞者にどのように影響したかを確認した．「Q1: ファシリテータ AI による同じ質問の繰り返しが対話の妨げになったか」，「Q2: ファシリテータ AI による同じ質問の繰り返しが鑑賞を深めることに貢献したか」の 2 つの質問を行った．1 つ目の質問の回答は「大変妨げになった (1 点)」から「全く妨げにならなかった (5 点)」で行った．2 つ目の質問の回答は「全く貢献しなかった (1 点)」から「大変貢献した (5 点)」で行った．いずれの質問も得点が高いほど，ポジティブな印象を示している．
- **話題転換の影響**　ファシリテータ AI が，まだ対話の中で話題に上がっていない絵画中の物体や事象を鑑賞者に示し，話題の転換を図ることが，鑑賞者にどのように影響したかを確認した．「Q3: ファシリテータ AI による話題の転換が対話の妨げになったか」，「ファシリテータ AI による話題の転換が鑑賞を深めることに貢献したか」の 2 つの質問を行った．回答は「同じ質問の繰り返しの影響」と同じ形式をとった．
- **絵画の印象**　「絵画についての印象を深めることができたか」の質問を行った．回答は「全く深まらなかった (1 点)」から「大変深まった (5 点)」で行った．

5.2　実 験 結 果

実験結果の平均値および標準誤差を求めたものをそれぞれの質問項目ごとに図 6 から図 10 に示す．提案手法と比較手法の間の有意差をウィルコクソンの符号付き順位付き検定 [7] にて検定した．また，検定結果について，この被験者実験では多重比較を行なっているため，各質問項目ごとに有意差があっても，それらが偽陽性の可能性がある．そこで BH 法 [8] にて各質問項目の P 値を調整し，それらをもとに有意差を判定する．グラ

図 4　クロード・モネ《日の出，印象》1872 年
出典：https://www.wga.hu/

図 5　フィンセント・ファン・ゴッホ《星月夜》1889 年
出典：https://www.wga.hu/

フには提案手法と比較手法間における有意差の有無も示す（*：5% 水準で有意）．また表 2 にゴッホの《星月夜》を用いた比較手法の対話例，表 3 にモネの《日の出，印象》を用いた提案手法の対話例を示す．表 4 は被験者実験における各質問項目の平均値，中央値，P 値，BH 法による調整後の P 値を示している．

5.2.1　同じ内容の質問の繰り返しによる影響

Q1 の同じ質問の繰り返しが対話の妨げになったかについて，比較手法は中央値が 2 点であり平均点が 2.6 点であった．提案手法においては中央値が 4 点で平均点が 3.4 点であった．P 値は 0.0203 であった．図 6 の箱ひげ図が示すように，比較手法は提案手法よりも低い評価をした鑑賞者が多い．これは比較手法において，ファシリテータ AI の質問が堂々巡りが頻発していたためだと考えられる．表 2 にあるように，鑑賞者の「人の営み」という発言について何度を深掘りするような質問を繰り返しており，このことが鑑賞以前にファシリテータ AI との対話そのものを阻害していたことが中央値・平均値の低さにつながっていると考えられる．また，提案手法においても一つの回答を深掘りするような質問が見受けられた．表 3 が示すように，ここでは「幻想的」という内容について，同じような内容の質問を数回繰り返していた．

また，Q2 の同じ質問の繰り返しが鑑賞を深めることに貢献したかについて，比較手法では中央値が 3 点で平均値が 2.8 点であった．提案手法においては中央値 4 点，平均値 3.6 点であった．P 値は 0.0384 であった．提案手法・比較手法のどちらも Q1 に比べて平均値が 0.2 点増加している．これは，ファシリテータ AI が鑑賞者の発言の理由を追求する質問が繰り返されることにより，鑑賞者が絵画に抱いた印象を明文化したためだと考えられる．表 2 において，鑑賞者の「不思議な雰囲気」というものが，絵画のその要素から由来していて，不思議さを感じる理由について明文化されている．またファシリテータ AI の追求を通して，絵画を深く考える行為そのものも鑑賞を深めることに貢献したと考えられる．

5.2.2　話題転換の影響

Q3 の話題の転換が対話の妨げになったかについて，比較手法においては，中央値 3.5 点，平均点が 3.4 点であるのに対し，提案手法では中央値が 5 点，平均点が 4.4 点であった．P 値は 0.0156 あった．提案手法においては，図 8 が示すように，50% 以上の鑑賞者が 5 点をつけていたことからも，ファシリテータ AI が対話中に適切なタイミングで，絵画に描写されている物体・事象などを用いた新しい話題の提供が行えていたと考えられる．Q4 の話題の転換が鑑賞を深めることに貢献したかについて，比較手法においては中央値 3.5 点，平均値が 3.1 点に対し，提案手法においては中央値 4 点，平均値が 4 点であった．P 値は 0.0032 であった．以上のように，新しい話題の提供が鑑賞者の絵画理解を深めることに貢献したのは，その提供される新しい話題が，鑑賞者に新しい観点を与えるための視線誘導になったためだと考えられる．表 3 にあるように，ファシリテータ AI の「この作品の中でまだ触れられていない『太陽』という話題について」という質問が，鑑賞者の視点を，今見ている場所から「太陽」に誘導し，これにより鑑賞者は新たな意見を獲得している．

ただ一方で，図 8 と図 9 から分かる通り，Q4 の方が Q3 よりも箱ひげ図の幅が広くなっている．これは，話題の転換が対話を自然なものにしたが，それ自体は鑑賞を深めることに貢献しなかったと答えた鑑賞者がいることを示している．つまり，鑑賞者にとって，対話型鑑賞にて一度に多くの情報 (ファシリテータ AI によって提示される，絵画鑑賞における新しい視点) を受け取ることが，負担になる場合があったと考えられる．

5.2.3　絵画の印象について

Q5 の絵画についての印象を深めることができたかについて，比較手法は中央値 3 点，平均値 3 点．提案手法においては中央値 4 点，平均値 3.7 点であった．P 値は 0.0282 であった．図 10 より，提案手法においては鑑賞者の 75% が比較手法の中央値・平均値を超えていることから，比較手法よりも提案手法が絵画の印象を深めることに貢献した．これは，提案手法においては，ファシリテータ AI が鑑賞者のまだ絵画の中で注目していない要素をもとに質問することで，絵画鑑賞における新しい視点を提供することで，鑑賞者が絵画に抱く感情・印象の幅を広げることができたためだと考えらる．一方，比較手法において，図

10 より，四分位範囲が 1.2 点から 4 点と幅広いことから鑑賞が深まったか関して評価が大きく分かれていた．これはある鑑賞者にとってはファシリテータ AI による同じ質問の繰り返しが，絵画のある一つの要素を追求することに貢献した一方，そのほかの鑑賞者にとっては，同じ内容の質問の堂々巡りが対話型鑑賞を阻害する要素として受け取られたためだと考えられる．

5.2.4 ま と め

BH 法で P 値を調整した場合でも各質問項目には全て 5% 水準での有意差が確認された．

以上の結果を通して，提案手法のシステムは絵画中のまだ対話の話題として触れられていない要素をもとに，新しい質問を生成することが，ファシリテータ AI による同じ内容の質問の繰り返しを抑制に貢献していたことが確認できた．これにより，ファシリテータ AI との自然な対話が可能となった．また，ファシリテータ AI によって提供される絵画の要素は，鑑賞でどこに注目するべきかというヒントになったため，絵画の印象を深めることに貢献した．

一方，質問の繰り返しは対話型鑑賞の妨げになる反面，鑑賞者の意見を追求することにもつながり，そのことが絵画印象を深めることに貢献した事例もあった．

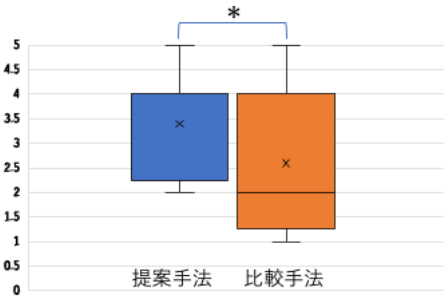


図 6　Q1: ファシリテータ AI による同じ質問の繰り返しが対話の妨げになったか

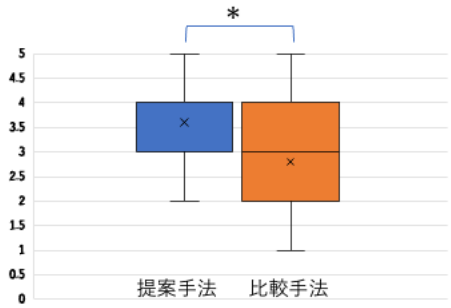


図 7　Q2: ファシリテータ AI による同じ質問の繰り返しが鑑賞を深めることに貢献したか

6 ま と め

本論文ではユーザがシステムと共に対話型鑑賞を行う状況を想定し，ユーザが，巨大言語モデルを用いて作成したエージェントと対話しながら，芸術作品を鑑賞することで，ユーザ単独での擬似的な対話型鑑賞を実現し，鑑賞体験を向上させるシステムを開発した．この際に，単純なプロンプトで LLM にファシリテータの役割を与えたとしても，LLM は鑑賞者の一つの意見について，堂々巡りする質問を生成し続けてしまう問題点があった．

それを解決するために，提案手法としては，鑑賞者とファシリテータ役のエージェントとの対話型鑑賞の対話ログをもとにファシリテータの質問内容を修正させるフィードバック用のエージェントを導入し，フィードバック用のエージェントが対話ログの校閲し，ファシリテータの繰り返しの質問が確認された場合，フィードバック用のエージェントが繰り返し発話の抑制を行うためのプロンプトを作成させた．このプロンプトをもとにファシリテータは質問内容を訂正し，繰り返し発話を抑制するシステムを実装した．

また被験者実験の結果，提案手法のシステムによる，絵画中のまだ対話の話題として触れられていない要素をもとに新し

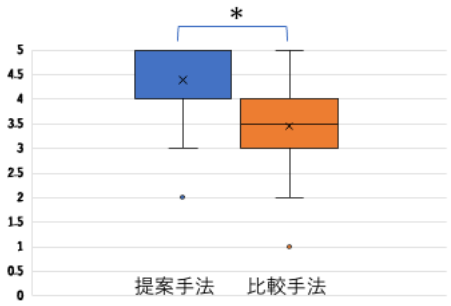


図 8　Q3: ファシリテータ AI による話題の転換が対話の妨げになったか

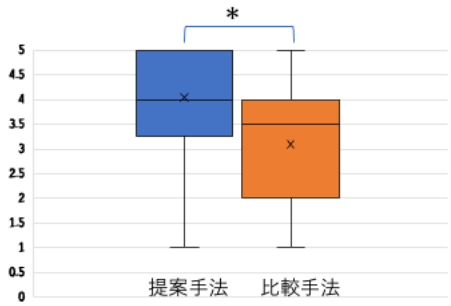


図 9　Q4: ファシリテータ AI による話題の転換が鑑賞を深めることに貢献したか

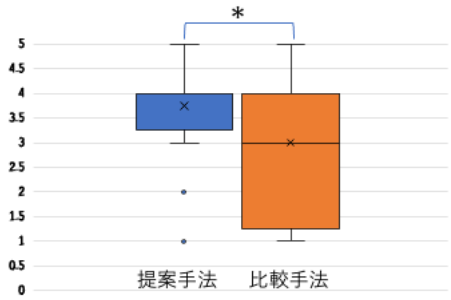


図 10　Q5: 絵画についての印象を深めることができたか

表 2　比較手法の対話例（鑑賞絵画：ゴッホ）

鑑賞者: 不思議な全体な雰囲気を感じました。
ファシリテータ: どのような要素がそう感じさせるのか，具体的に教えていただけますか？
鑑賞者: 青のグラデーションと月明かり、下町と謎の影の対比
ファシリテータ: その「青のグラデーションと月明かり、下町と謎の影の対比」がどのように「不思議な全体の雰囲気」を作り出していると感じますか？
鑑賞者: 対比だと思います
ファシリテータ: その「色の対比、物体の大きさの対比」がこの作品を通して何を表現していると感じますか？
鑑賞者: 人の営み
ファシリテータ: その「人の営み」を感じさせる部分は具体的にどこでしょうか？
鑑賞者: 下町の光と、それを照らす月明かり
ファシリテータ: その「下町の光と、それを照らす月明かり」から「人の営み」を感じる理由について教えていただけますか？
鑑賞者: 夜に灯る光

表 3　提案手法の対話例（鑑賞絵画：モネ）

鑑賞者: 幻想的
ファシリテータ: 具体的には、どの部分や要素が幻想的と感じるのですか？
鑑賞者: 淡い色使いです
ファシリテータ: 淡い色使いが幻想的な印象を与えるというのは面白い観察ですね。色彩以外にも、この絵画を幻想的に感じさせる他の要素はありますか？
鑑賞者: 背景の船の群れと対比して前面の小舟がぽつんと浮かんでる感じ
ファシリテータ: あなたが言及した淡い色使いや背景の船の群れと小舟の対比が作品全体の幻想的な雰囲気を作り出しているという観察は非常に興味深いものですね。次に、この作品の中でまだ触れられていない「太陽」について質問したいと思います。この淡いオレンジ色の太陽が画面の中央よりやや上にあることについて、何か特に印象的な点はありますか？
鑑賞者: 綺麗で、淡いタッチが素敵だと思います

表 4　被験者実験結果

質問項目	Q1		Q2		Q3		Q4		Q5	
手法	提案手法	比較手法	提案手法	比較手法	提案手法	比較手法	提案手法	比較手法	提案手法	比較手法
平均値	3.4	2.6	3.6	2.8	4.4	3.4	4.0	3.1	3.7	3.0
中央値	4.0	2.0	4.0	2.0	5	3.5	4.0	3.5	4.0	3.0
検定量 W	20.0		20.0		14.5		4.5		11.5	
P 値	0.0203		0.0384		0.0156		0.0032		0.0282	
調整後の P 値	0.0338		0.0384		0.0390		0.0160		0.0352	

い質問を生成することが，ファシリテータ AI による同じ内容の質問の繰り返しを抑制に貢献していたことが確認できた．これにより，ファシリテータ AI との自然な対話が可能となった．さらに，ファシリテータ AI によって提供される絵画の要素は，鑑賞でどこに注目するべきかというヒントになったため，絵画の印象を深めることに貢献した．一方，質問の繰り返しは対話型鑑賞の妨げになる反面，鑑賞者の意見を追求することにもつながり，そのことが絵画印象を深めることに貢献する場合もあることが示唆された．

現時点では，ファシリテータ AI とユーザによる一対一の対話がシステムとして実装されている．今後は，鑑賞者としてのエージェントを交えた，ユーザと複数の LLM エージェントによる対話を実装し，より対話型鑑賞に近いシステムを開発する予定である．

文　献

[1] 奥田秀巳. 対話型鑑賞の展開の方向性についての一考察. 美術教育学研究, Vol. 54, No. 1, pp. 57–64, 2022.
[2] Jo Marie Reilly, Jeffrey Ring, and Linda Duke. Visual thinking strategies: a new role for art in medical education. *Fam Med*, Vol. 37, No. 4, pp. 250–2, 2005.
[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.
[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27730–27744, 2022.
[5] 平野智紀, 安斎勇樹, 山内祐平. 対話型鑑賞のファシリテーションにおける情報提供のあり方. 日本教育工学会論文誌, Vol. 43, No. 4, pp. 285–298, 2020.
[6] 小倉拓人, 徳永隼人, 松村冬子, 原田実ほか. 進行役と鑑賞者のエージェントを用いた対話型鑑賞システム. じんもんこん 2015 論文集, Vol. 2015, pp. 43–50, 2015.
[7] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 196–202. Springer, 1992.
[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multi-

ple testing. *Journal of the Royal statistical society: series B (Methodological)*, Vol. 57, No. 1, pp. 289–300, 1995.

複数の大規模言語モデルによる漫才自動生成の比較分析

下崎　安紋†　北畑　哲也††　柘植　陽介††　北村　達也††　梅谷　智弘††
灘本　明代††

† 甲南大学大学院自然科学研究科知能情報学専攻　〒 658–8501 兵庫県神戸市東灘区岡本 8 丁目 9 － 1
†† 甲南大学知能情報学部知能情報学科　〒 658–8501 兵庫県神戸市東灘区岡本 8 丁目 9 － 1
E-mail: †{m2324007,s2171036,s2071074}@s.konan-u.ac.jp, ††{t-kitamu,umetani,nadamoto}@konan-u.ac.jp

あらまし　日常生活において笑いは重要であると言われている．そこでこれまで我々は日々の笑いを提供するために，漫才台本の自動生成を行ってきた．我々の提案する漫才台本自動生成は，ニュース記事やレシピサイトなどを基にボケやツッコミの生成をしている．一方，近年 ChatGPT に代表されるように大規模言語モデルの普及により，自然な対話を生成することが容易になってきている．しかしながら，同様の対話形式である漫才のようなユーモアを含む対話の自動生成を行うことは困難であると考えられる．そこで本研究では，将来的に大規模言語モデルを用いた漫才台本自動生成をするために，既存の複数の大規模言語モデルを用いて漫才台本の自動生成を行い，モデルごとに比較分析を行う．これにより，現状の大規模言語モデルによる漫才台本自動生成の問題点を抽出する．

キーワード　漫才，台本自動生成，大規模言語モデル

1　はじめに

日常生活において笑いは重要であると言われている [1]．そこで我々は気軽に日々の笑いを提供することを目的とし，身近で親しみやすく笑いを提供できるコンテンツとして漫才に着目し，漫才ロボットを提案している [1] [2]．漫才ロボットの特徴は，ユーザの好みに合わせた漫才台本を自動生成し，それを図 1 のようなロボットやチャットアプリケーションを用いて漫才を提供している [3]．漫才ロボットの表現方法は，テキスト，音声や表情であるためコント調の漫才ではなく，しゃべくり漫才に焦点を当て自動生成している [4] [5]．具体的には，システムに入力されたお題 (キーワード) に対して，ニュースサイトやレシピサイトからお題に関する情報を取得しボケやツッコミの生成を行っている．

一方，近年 ChatGPT[2] に代表される生成系 AI が急速に普及し，研究のみならず様々な分野での実用が行われている．この生成 AI により，自然な流れでの雑談対話を生成することが可能になっている．そこで我々はユーモアを含む雑談対話である漫才を ChatGPT を用いて生成を行った．その際に使用したプロンプトと生成例を図 2 に示す．生成された結果より，2023 年 12 月現在の ChatGPT ではお題に関する特徴や，方言については生成することが可能であることがわかる．しかしながら，漫才に必要不可欠なユーモアのある対話は全く生成されておらず，雑談対話のみが生成されていることがわかる．その原因として，様々なことが考えれるが，その中の一つとして，現在の大規模言語モデルの学習データに漫才などのユーモアを含む雑談があまり含まれていないことがあげられる．そこで本研究では，現状の大規模言語モデルで漫才を生成することができるかどうかを検証することを目的とし，複数のデータセットによって構築したモデルによる漫才台本自動生成の比較分析を行う．具体的には，プロの漫才師 (以下漫才師) の漫才のデータをもとにデータセットの構築を行い，横田らの Swallow[3] を用いファインチューニングを行う．そしてその際のデータセットを構築するための手法ごとに性能比較を行う．それらを行うことで，大規模言語モデルを用いたユーモア生成の問題点を見つけることで，将来的に漫才の自動生成を行うことができるのではないかと考えている．

以下，２章では本研究の関連研究について述べ，第３章ではデータセットの構築手法について，４章では学習を行ったモデルごとの性能を比較し問題点について，5 章ではまとめと今後の課題について述べる．

a:小型ロボット　b:大型ロボット
c:チャットアプリケーション

図 1　漫才ロボットシステムの実現例

1：https://www.pref.osaka.lg.jp/attach/4002/00029624/waraisasshi.pdf
2：https://chat.openai.com/
3：https://tokyotech-llm.github.io/

prompt	漫才のお題は神戸について 漫才の登場人物はボケとツッコミ 漫才の方言は，関西弁 ツッコミを入れる時はなんでやねんを入れる
ボケ	よお、おばあちゃん、今日は神戸の話しよろしくやで！
ツッコミ	おお、神戸やねん！まぁ、神戸って言えば何が有名やろ？
ボケ	もちろん、神戸ビーフやないかい！あれは絶品やで！
ツッコミ	そやな！あの柔らかさとジューシーさはたまらん！神戸ビーフって言うたら、値段も高いけどな。
ボケ	ホンマにな！あんな美味しいお肉、食べるためには財布の中身を確認せなアカンな。
ツッコミ	でも、神戸って言えばビーフだけじゃないやろ？
ボケ	そうやな！神戸には有名な観光スポットもたくさんあるやろ
ツッコミ	そや！神戸港やら、メリケンパークやら、夜景の有名なルミナリエやら、見所がいっぱいや！
ボケ	ほんまや！神戸港の夜景はロマンチックやし、ルミナリエはキレイやし、メリケンパークはのんびり散歩できるな。
ツッコミ	そうそう、それに神戸には有名なスポーツチームもあるんやで！
ボケ	おお、確かに！ヴィッセル神戸やろ？イニエスタも所属してるし、注目されてるな。
ツッコミ	そうや！ヴィッセル神戸応援してるおじいちゃんおばあちゃんも多いんやろな。
ボケ	そやな！神戸には美味しいものもたくさんあるし、楽しいところもいっぱいや。老後の楽しみのひとつやんな。
ツッコミ	そうやな！神戸ってホンマに魅力的な街やで！
ボケ	ほんまや！老人になっても、神戸で楽しく過ごせるやんな。
ツッコミ	そやそや！神戸やねん、なんでやねん！ (観客笑い声)

図 2　ChatGPT での漫才生成例

2　関 連 研 究

雑談対話に関する様々な研究が行われている．杉山ら [6] は大規模な Transformer ベースの日本語対話モデルと日本語の雑談データセットの構築を行っている．

大規模言語モデルを用いたユーモア生成の研究が行われている．原ら [7] は ChatGPT のユーモア生成能力についてプロンプトベースでの分析を行った. 結果より，ChatGPT は対話によって，人間が面白いと感じるユーモアを生成できるということが明らかとなった．本研究は既存のモデルに対してファインチューニングを行い性能比較を行なっている点で異なる．

漫才に関する様々な研究が行われている．土肥ら [8] は，同一の芸人による漫才とコントの映像を比較分析し，コントの中で作られる仮想空間の創出において間のとり方，顔を向ける方向，距離のとり方，肩幅の見せ方の 4 項目が重要であると述べ，漫才とコントの違いに着目した分析を行っている. それに対して，本研究では漫才台本の自動生成に着目している点で異なる．宿利ら [9] は，日本語母語話者と日本語学習者の「面白い話」の面白さを活かす間に関して実験を行い，日本語母語話者では「早い間」と「遅い間」の 2 種類があると述べている．宿利らは面白さを活かす間に着目しているが，本研究では漫才台本の自動生成に着目している点で異なる．林ら [10] は，ロボットによる新たな社会的受動メディアと位置づけ，2 体のヒューマノイドロボットを用いた漫才を行っている．人間の漫才とロボットの漫才との比較実験を行い，その結果エンターテイメントとしてロボット漫才の有用性を示している．林らの提案するロボット漫才では，動作などを人間の漫才に近づけて漫才を行っているが，本研究では漫才台本の自動生成に着目してロボットやアプリで実演している点で異なる．笑いを生み出すコンテンツに関する研究も行われている．伊勢崎ら [11] は，ユーザの笑い感情の誘起が容易に行えると想定される大喜利に着目し，機械学習的アプローチによりユーザの笑い感情を誘起するロボットインタラクションモデルを提案している．伊勢崎らはセンサデータを用いてユーザの笑いの感情が誘起させることが確認された会話文をデータセットとしているが，本研究では笑の感情を誘起させるインタラクションとして漫才台本自動生成を用いている点で異なる．宅和ら [12] は，共感を得ることで笑いを誘う発話文としてあるあるネタを Twitter より取得する手法を提案している．本研究では，笑いを生み出すコンテンツとして漫才に着目する点で異なる．

3　データセット構築

大規模言語モデルで漫才を生成するためには，漫才データでファインチューニングする必要があると考える．ファインチューニングには多数のデータが必要である．しかしながら，ほとんどの既存の漫才は漫才台本のテキストデータが存在しないため，漫才台本のテキストデータを用意する必要がある．また，それらのデータでは発話テキストや話者情報が必要なため，多くの負担を要した．そこで本研究は，佐々木ら [14] の提案する自動アノテーションのパイプラインを改良することにより，漫才台本自動書き起こしシステムを提案する．本研究で提案する漫才台本自動書き起こしシステムのフローを図 3 に示す．

3.1　データ収集

本研究では，データセットの作成を行うために漫才師の台本の収集を行う．本研究では，漫才台本の収集を行うために漫才ビデオを用いる．漫才ビデオは，コント調やしゃべくり漫才などが存在しているがコント調の漫才は道具や動作を含めておりそれらはテキストデータとして表現する事が困難なため，しゃべくり漫才のみを対象とする．具体的には，M1(2001-2019) 決勝，敗者復活戦をはじめ，中田ダイマルラケットなど様々な年代を用いる．収集を行った漫才ビデオを表 1 に示す．

表 1　漫才ビデオ

種類	本数	時間
平成以降	100	500 分
昭和	20	200 分
合計	120	700 分

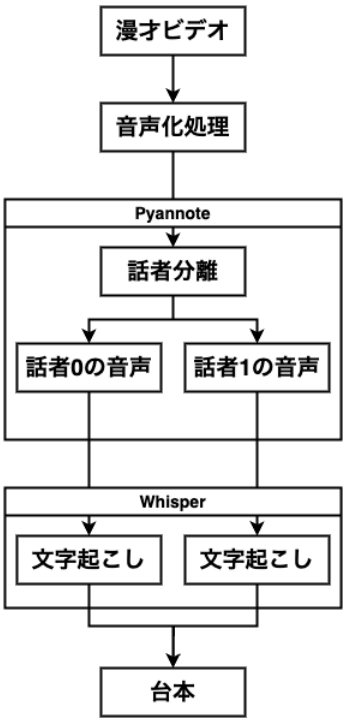


図 3　漫才台本自動書き起こしシステムフロー

3.2 話者分離

漫才でファインチューニングするにあたり，漫才師のうちどちらがどの発話をしたかどうかを分離する必要がある．そこで，本研究では Pyannote[4]を用いて，話者の分離を行う．しかしながら，ボケ役とツッコミ役は漫才によって異なる場合があるため話者の分離は可能だが，発話者の役割のアノテーションを行うことは困難である．そのため，本研究では話者を 0，1 とし各漫才内で完結するようにする．

3.3 文字起こし

漫才台本自動書き起こしシステムでは，漫才ビデオをテキストデータにするために，音声の書き起こしを行う Whisper[5]を用いる．Whisper には，様々なモデルサイズが用意されているが small，midium，large-v3 の比較を行った結果，small は誤字が多く修正作業に時間を要し作業負担も大きかった．一方で，large-v3 は生成に時間を要するが作業負担が軽減されデータセット作成時間の短縮を行うことができた．そこで，本研究では large-v3 を用いる．

まず初めに，漫才台本から発話者毎に分離を行った各話者の音声データを用いて文字起こしを行う．そして，出力されたタイムスタンプを基に漫才台本の出力を行う．しかしながら，現状の手法では方言や動画データの劣化などでうまく文字起こしや話者分離ができていない場合があった．そこで最後に，出力された台本に対して話者情報や発話情報の修正を人手により行う．台本は表 2 に示すように発話ごとに，開始時間，終了時間，話者，セリフからなっている．

4：https://huggingface.co/pyannote/speaker-diarization
5：https://huggingface.co/openai/whisper-large-v3

表 2　作成台本例

開始時間	終了時間	話者	台詞
00:10:0	00:11:3	0	いやー，ありがとうございます.
00:11:3	00:12:8	1	よろしくお願いします
		…	
04:34:1	04:35:4	1	就職しろよ！
04:35:2	04:36:7	0	いい加減にしろ！どうもありがとうございました

4 比較実験

現状の既存モデルで最適な漫才が生成できるかどうかを確認することを目的とし，比較実験を行う．本研究では，様々な大規模言語モデルで漫才の自動生成を行いそれぞれの生成結果の比較を行う．本研究では比較実験モデルとして ChatGPT，Llama2，Llama2-chat のモデルを用いる．Llama2，Llama2-chat は作成した漫才データを用いてファインチューニングを行ったものの比較実験も行う．具体的には，以下の 4 種類の比較実験を行う．

（1） ファインチューニングなし比較実験

（2） 漫才データのみでファインチューニングをした比較実験

（3） 漫才データと雑談データ両方でファインチューニングをした比較実験

（4） 漫才データに発話者のアノテーションしたデータでファインチューニングをした比較実験

上記 1 は ChatGPT，Llama2，Llama2-chat の 3 つのモデルで比較し，2〜4 は Llama2，Llama2-chat の 2 つのモデルで比較実験を行う．ここでは，メモリ削減のため Lora 学習で行い学習可能な層を対象とする．パラメータはバッチサイズ 8 とラーニングレートは 0.0002 とする．

4.1 実験条件

使用モデル

- ChatGPT

2024 年 2 月現在，ChatGPT の中で生成能力が高い GPT-4 Turbo[6]を用いる．その中でも 2024 年 2 月時点で公開されているモデルのうち最新の gpt-4-0125-preview[7]を用いる．

- Llama2

本研究では，研究利用・商業利用・ファインチューニングができる Llama2 のモデルの中で，2024 年 2 月時点で最も日本語生成性能が高い Swallow[8]を用いる．本研究では，ファインチューニングの際に用いる GPU メモリの制約上公開されている Swallow のモデルのうち Swallow-13B-hf[9]を用いる．

- Llama2-Chat

漫才は，ユーモアのある対話文で構成されていると考えているため，ファインチューニングを行うモデルも対話用に Tuning を行なったものが有用なのではないかと考えられる．そこ

6：https://openai.com/gpt-4
7：https://openai.com/product
8：https://tokyotech-llm.github.io/swallow-llama
9：https://huggingface.co/tokyotech-llm/Swallow-13b-hf

で，本研究では対話用のモデルの中で日本語の生成精度が高い Youri-7b-chat[10] を用いる．

データセット

漫才台本自動書き起こしシステムで生成された漫才台本を用いて実験を行う．入力は 1 つの対話対とする．ここでいう対話対とは，一方が発話し，それに対して他方が発話し終わるまでを 1 つの対話対として分割を行う．例えば，A が発話し，それに対して B が 3 回連続して発話しても，それを 1 対話対とする．また，漫才台本には A が発話し，B が発話する前に A が複数回発話する場合がある．この場合も，A の複数回発話してそれに対して B が発話し終わるまでを 1 対話とする．これにより，120 本の台本から 12000 対の対話データが得られ，これを実験に用いる．

4.2 実験 1. ファインチューニングなし比較実験

プロンプト作成

一般的に大規模言語モデルで文章の生成を行う際に，入力するプロンプトによって生成精度が大きく向上すると言われている．そこで，我々は，既存の大規模言語モデルのプロンプトを考慮することにより漫才台本が生成できないかと考えた．そこで，既存の大規模言語モデルのプロンプトを複数生成し，その性能を比較する実験を行う．作成を行なったプロンプトを図 4 に示す．

sys_prompt	以下は、タスクを説明する指示です。要求を適切にみたす応答をかきなさい。
prompt	###指示: あなたは面白い漫才師です。 2 人の対話文を以下の対話内容に沿ってボケとツッコミを生成してください ###対話内容: お題 A さん：ボケ B さん：ツッコミ

図 4　入力プロンプト例

結果と考察

漫才の生成の際には，3 つのお題 (神戸，唐揚げ，阪神タイガース) について作成をしたプロンプトを用い，生成を行う．各モデルで生成された結果を表 3(1) に示す．生成結果より，すべてのモデルにおいてお題についての対話文を生成することができているが，何もボケた発話が何もなく漫才の生成をすることができていないことがわかる．また，ChatGPT は日常会話に近い対話が生成されているのに対して，Llama2 や Llama2-chat はお題の説明を行うような対話の生成が多く見られた．さらに Llama2 と Llama2-chat を比較すると Llama2-chat の方が話口調の生成がされていることがわかる．このことから，漫才を生成するためにファインチューニングなしのモデルを用いるのは有用ではないことがわかった．一方で，ユーモアを含まない雑談対話を生成するには ChatGPT が最も有用なモデルであると言える．

4.3 実験 2. 漫才データのみでファインチューニングをした比較実験

次に漫才データセットである 12,000 対の対話対を用いてファインチューニングを行なった Llama2 と Llama2-chat で比較実験を行う．入力したプロンプトはファインチューニングなし比較実験と同様のお題で生成を行う．本比較実験では，ChatGPT のファインチューニングができないため Llama2 と Llama2-chat で行う．生成結果を表 3(2) に示す．結果より，どちらのモデルでも「どこやねん」や「なんでやねん」といったようなツッコミを生成することができていることがわかる．一方で，漫才にあるようなボケを生成することは出来ていないことがわかる．また，プロンプトの入力では A がボケ，B がツッコミとしていたが実際に生成された結果では A がツッコミを発話していることが多くあった．その原因として，漫才データにある発話がボケ役・ツッコミ役のどちらが発話しているかがアノテーションされていないため学習することが出来ていない事が考えられる．さらに，Llama2，Llama2-chat で出力された結果はファインチューニングなしモデルと比較して 1 発話が短くなる傾向があり内容が単調になっている．その原因として，漫才データには相槌のような短いフレーズが含まれているためであると考えられる．

4.4 実験 3. 漫才データと雑談データ両方でファインチューニングをした比較実験

漫才モデルの比較実験より漫才データのみでファインチューニングを行なった際，生成内容が単調になったことから漫才データに対して日常の雑談対話を加えることで多様な漫才の生成ができるのではないかと考えた．そこで本研究では，雑談対話として名大コーパス [15] を用いる．名大コーパスは，科学研究費基盤研究「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」[15] の一環として作成された，129 会話， 合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである．雑談対話の対話対は，漫才データセットと同様の手法を用いる．また，漫才データセットに追加する雑談対話数は漫才データ数と同様の 12000 対とする．これにより，漫才データと雑談対話との合計対話数が 24,000 対となり，これを用いて Llama2 と Llama2-chat のファインチューニングを行う．生成を行った結果を表 3(3) に示す．結果より，漫才データのみのモデルと比較してツッコミを生成することも困難になっていることがわかる．その原因として，追加を行った雑談データが漫才データに対して多い為であると考えられる．このことから，漫才データに対して追加する雑談データの比率を検討する必要があることがわかった．そのため，漫才データと雑談データの比率を検討し再度モデルを作成するためには，過学習を防ぐために漫才データを増やす必要があることがわかった，

10：https://huggingface.co/rinna/youri-7b-chat

4.5 実験 4. 漫才データに発話者の役割アノテーションしたデータでファインチューニングをした比較実験

漫才モデルでの生成結果にてプロンプトで与えた役割と異なった発話の生成が見られたため発話者情報に対してボケやツッコミのアノテーションの付与を行う．役割の決定として，ボケ役やツッコミ役があらかじめ決まっている漫才師は，Wikipediaから役割の取得を行う．しかしながら，漫才師によってはボケ役やツッコミ役が定まっていない場合も存在する．その際は本研究では，その漫才内で「もうええわ」など漫才の終わりを発話している漫才師をツッコミ役とする．生成結果を表 3(4) に示す．結果より，Llama2 と Llama2-chat のどちらもツッコミや冗談，嫌味などを生成することができていることがわかる．このことから，役割をアノテーションをすることで漫才モデルと比較してよりユーモアのある対話の生成をすることができることがわかった．そのため，漫才を学習させる上でアノテーションを付与することは有用であることがわかる．しかしながら，生成された冗談や嫌味は漫才のような笑いを生み出すことは出来ていないといえる．その原因として，漫才のボケやツッコミには様々な技法が存在しておりそれらをうまく学習できていない為であると考えられる．

4.6 総　　評

比較実験のまとめを表 4 に示す．実験より，ファインチューニングなし比較実験では ChatGPT が最もお題に対して日常対話に近い出力をし Llama2，Llama2-chat ではお題の説明するような発話の生成が多く見られた．そのため，雑談対話を生成する上では ChatGPT が有用であるがファインチューニングができないことから漫才の自動生成を行う上では有用ではないと言える．次に，漫才データのみでファインチューニングを行なった Llama2 と Llama2-chat ではツッコミを交えた対話文の生成することができることがわかった．さらに，ファインチューニングを行う際に役割のアノテーションを付与することでより冗談やツッコミを交えた対話文の生成することができることがわかった．しかしながら，生成された冗談は漫才におけるボケのような笑いを生み出すようなものは生成することはできなかった．原因として，漫才のボケは多くの種類があるためうまく学習できていなかったのではないかと考える．そのため，ボケを生成するために役割だけでなくボケの種類のアノテーションを付与しファインチューニングを行う必要があるのではないかと考えている．これにより，既存の大規模言語モデルでは現状漫才台本を自動生成することが困難であることがわかった．ただし，今回はファインチューニングに用いた教師データ数が少ない為，さらに多くの教師データを用いて漫才データに発話者の役割アノテーションしたデータでファインチューニングをした比較実験をする必要があることもわかった．

5　まとめと今後の課題

本研究では，大規模言語モデルを用いた漫才台本自動生成の実現への初段階として，データセットの構築を行った．さらに提案手法を用いて作成した漫才データを用いて 4 つの比較実験を行った．比較実験の結果，漫才のみのデータを用いてファインチューニングを行うと漫才におけるツッコミを生成することは可能であることはわかったが，ボケの生成を行うことは困難であることがわかった．さらに，モデルの生成結果が単調になっていた．そこで，生成結果に多様性を持たせるために漫才データのみのモデルに対して雑談対話を追加しモデルの構築を行なった．しかしながら，どちらもユーモアを生み出すことが困難となりその原因として漫才データの不足と雑談データの比率が考えらるためそれらを考慮したうえで再度実験を行う必要があることが分かった．

一方で，漫才データセットに対して役割のラベル付けを行うことでよりユーモアのある対話が生成されていることがわかった．しかしながら，漫才データのみのモデルと比較してユーモアが生成されやすくなったが漫才におけるボケのような笑いを生み出すようなユーモアを生成することはできなかった．その原因として，漫才におけるボケは多くの手法が存在しているが，それに対して漫才のデータ数が不足していることが考えられる．そのため，漫才のデータ数を追加する必要がある．

更に，本研究では漫才台本におけるボケ役とツッコミ役のラベル付けのみを行ったが，さらにそれぞれの役割で用いる技法のラベル付けを行う必要があると考えている．そうすることによって，多様なユーモアの生成が可能ではないかと考える．また，本研究では話者ごとにデータセットの分割を行いモデルの構築を行ったため，モデルが生成した対話文は明らかに続きがあるのにも関わらず生成が終了してしまっていることが問題点として挙げられる．

そこで今後の課題として，問題点の解決を行うためにデータの分割方法の検討を行う必要がある．具体的には，漫才データセットにある漫才対話を漫才師が用いる技法ごとに分割を行いたいと考えている．

謝　　辞

論文の一部は私学助成金（大学間連携研究助成金）の助成によるものである．ここに記して謹んで感謝の意を表する．

文　　献

[1] 原口和貴, 梅谷智弘, 北村達也, 灘本明代, “Seq2Seq を用いたボケの自動生成手法の提案”, DEIM2021,H24-3,.
[2] 下崎安紋, 梅谷智弘, 北村達也, 灘本明代, “漫才のわかりやすさを考慮した漫才台本自動生成手法”, DEIM2023,5b-2.
[3] 家根和希, 原口和貴, 佐藤亮, 荒牧英治, 宮代勲, “チャット型漫才アプリケーションの提案”, DEIM2020,A3-3.
[4] 大隈一郎, -「しゃべくり漫才入門」, リットーミュージック,2015.
[5] TBS ラジオ「土曜ワイドラジオ TOKYO ナイツのちゃきちゃき大放送」（編）, -「ナイツ　午前九時の時事漫才」, 駒草出版,2020.
[6] 杉山 弘晃, 成松 宏美, 水上 雅博, 有本 庸浩, 千葉 祐弥, 目黒 豊美, 中嶋 秀治, “Transformer encoder-decoder モデルベース雑談対話システムの学習方法に対する主観評価の変動分”, 人工知能学会全国大会論文集,2021.
[7] 原直大, 荒木健治, “ChatGPT を利用したユーモア応答生成の分析”, NLC2023.
[8] 土肥健太, 寺岡丈博, 榎本美香, “ 仮想的演技空間の創出方略の

分析：同一ネタの漫才とコントの違いを通じて”, 言語・音声理解と対話処理研究会, 人工知能学会, vol. 73, pp. 41-46, 2015
[9] 宿利由希子, ヴォーゲ ヨーラン, 林良子, 定延利之, “ユーモアを生み出すための「間」-ボケとツッコミのタイミングに関する考察-”, 2018 年日本認知科学会第 35 回大会, 2018
[10] 林宏太郎，神田崇行，宮下敬宏，石黒浩，萩田紀博，“ロボット漫才 社会的受動メディアとしての 2 体のロボットの利用”, 日本ロボット学会誌, Vol. 25, No. 3, pp. 381-389, 2007.
[11] 伊勢崎隆司, 小林明美, 望月崇由, 山田智広, “笑い感情を誘起するロボットインタラクションの検討”, 情報処理学会研究報告グループウェアとネットワークサービス (GN), Vol. 2017-GN-100,No. 7, pp. 1-5, 2017.
[12] 宅和晃志, 吉川大弘, ジメネス フェリックス, 古橋武, “非タスク思考型対話システムにおけるあるあるツイートからの共感誘発型会話生成手法に関する検討”, 知能と情報（日本知能情報ファジィ学会誌）, Vol. 30, No. 5, pp. 744-752, 2018
[13] Hugo Touvron and Louis Martin and Kevin Stone et.al, Llama 2: Open Foundation and Fine-Tuned Chat Models, arXiv:2307.09288, 2023
[14] 佐々木裕多，張建偉, “漫才対話の収集及び自動アノテーションのパイプラインの検討”, 情報処理学会第２９回年次大会　発表論文集，2023.
[15] 藤村逸子・大曽美恵子・大島ディヴィッド義和, 「会話コーパスの構築によるコミュニケーション研究」藤村逸子、滝沢直宏編『言語研究の技法：データの収集と分析』p. 43-72、ひつじ書房,2011

表 3　生成まとめ

No.	実験タイプ	ChatGPT-4	Llama2	Llama2-chat
お題：神戸				
(1)	ファインチューニングなし	A: ねぇ，今度神戸に旅行行こうよ B: え？神戸？牛に会いに行くの？ A: 馬鹿な事いってないで！神戸牛じゃなくて，神戸に行くんだよ B: え，でも神戸行ったら神戸牛ぐらい食べたいよね	A: 兵庫県神戸市は，兵庫県南東部に位置する港町です B: そうですね A: 神戸市は，古くから貿易港として栄え，現在も国際的な湊湾都市として知られています B: そうですね	A: 神戸にはたくさんの観光名所があるんですよ B: ほう，それはいいですね A: 神戸は，日本の大阪市や京都市に匹敵する都市です B: ははは！
(2)	漫才データのみ	-	A: ここはどこや？ B: ここは，日本の西のほうにある，大きな町や. A: どこやねん	A: ねえねえ，お前知ってる？神戸のハーバーの向かいにあるホテルって，もともと客船だったんだよ B: えっ，そうなの？知らなかった A: 知らなかったっけ？あのホテルもともとは白い客船で，第二次世界大戦前に建造されたんだよね B: えっ．ほんまに？それはすごいな
(3)	漫才-雑談データ両方	-	A: 神戸市ってどう思う？ B: 神戸市？ A: そう！ B: うーん，あんまり行ったことないんだ	B: あなたはどこ出身ですか？ A: 神戸です B: 神戸ってどんなところですか？
(4)	漫才-役割データ	-	A: えー，まずはね，僕らのホームグランドでもある神戸記念ホールのことから話そうと思います. B: はいはいはい．なんでやねんそんなん最初に言うてなかったやん．えらい急な話やな，ほんまに．まぁええわ，ほなワールドの話からいこうやないか	A: 私は神戸に住んでいます B: 神戸の人ってどんな人ですか？ A: 神奈川の人は冗談を言っているような顔をしているけど神戸の人は本気で冗談を言っているような感じがする B: それって，どういう意味や

次のページに続く

No.	実験タイプ	ChatGPT-4	Llama2	Llama2-chat
お題：唐揚げ				
(1)	ファインチューニングなし	A: さてさて，B さん唐揚げの話をしましょうか B: いいね唐揚げの話ならいつでもOK だよ A: なんだかんだで唐揚げは世界一の料理だと思わない B: あ，それなら唐揚げの儚さも忘れちゃいけないね	A: 今日の夜ご飯は唐揚げにしよっか B: そうだね！唐揚げって本当においしいよね A: 唐揚げっていいよね！僕はからからにならない方が好きなんだけど B はどう？ B: 私はからからな方が好きなんだけど	B: 唐揚げ好きですか？ A: ええ，大好きです B: どんな唐揚げが好きですか A: 特にこだわりはないですね
(2)	漫才データのみ	-	B: あなたはどうして唐揚げを食べたがるの？ A: だっておいしいやん B:A さんそれはただの肉の塊です A: Bさんそんなこと言わないで	B: あの，この唐揚げについてですが，これは手作りですか A: いや，コンビニの弁当だよ B: あ，そうですか A: なに？その顔
(3)	漫才-雑談データ両方	-	A: 唐揚げっておいしいよね B: そうだね！ A: でも，唐揚げって油が多いから，あんまり食べられないんだよね B: そうなんだ！	A: ねぇ，唐揚げってさ，味そんなに変わらないよね B: そんなことないよ A: え，どんな味がするの？ B: え，衣がついていて，衣の味がついているんだよ
(4)	漫才-役割データ	-	A: あなたは，どんな食べ物のなかで一番好きですか？ B: 私は鶏肉が好きです A:B さんはどの鶏胸肉の部位が好きですか B: 私は骨付きの手羽先が好きで羽の部分が嫌いです. A: あなたは鶏の尻尾と羽を食べますか？ B: 食べるわけない	A: あなたは何を作るのですか B: 私はチキンの胸肉を使った唐揚げを作ります A: あなたの胸は小さすぎます B: あなたは心が小さすぎます
お題：阪神タイガース				
(1)	ファインチューニングなし	A: ねえねえ，阪神タイガースってさやっぱり虎がモチーフなんだよね B: それはそうだけど何か？ A: それならさ，球団のマスコットって当然虎なんだよね？ B: それもそうだけど何を言いたいんだ	A: 昨日も阪神勝ってましたね B: え？阪神負けてましたよ A: 何でですか？ B: 昨日，阪神はホームランを１本も打たなかったからです	B:「今日は阪神タイガースが試合する日だ！」 A:「そうだね」 B:「勝つよ！」 A:「うん，そうだね」
(2)	漫才データのみ	-	A: 今日の阪神戦どうやった？ B: 阪神見てないよ A: 見てないのかよ B: 今日の試合は知ってる？	A: 阪神が優勝したらどうしよかな B: それはまだ早いよ A: なんでやねん B: 今年は優勝できると思うか？ A:(゜Д゜)ビッ！
(3)	漫才-雑談データ両方	-	A:「今年の阪神はどうなってんねん！」 B:「今年は優勝間違いなしや！」 A:「なんでそんなこと言うてんねん！」 B:「いやいや，今年の阪神は優勝間違いなしだよ！」	B:A さん，阪神ファンなんですか A: そうなんですよ B: 巨人ファンの私からすると許せないですね A: 今日は何の日か知っていますか
(4)	漫才-役割データ	-	A: 今年の阪神は強いな B: 今日の試合見た？ A: 見てないよ，俺は阪神ファンだからさ B: なんでやねん	B: 阪神はなぜ負けたんだい？ A: お前だよ！ B: なんでだよ B: ラモスが悪いんだよ A: 誰だよ B: 知らんのかよ

表 4　生成性能まとめ

No.	実験タイプ	ChatGPT-4	Llama2	Llama2-chat
(1)	ファインチューニングなし	△雑談のみ	△雑談のみ	△雑談のみ
(2)	漫才データのみ	-	○ツッコミを言う	○ツッコミを言う
(3)	漫才-雑談データ両方	-	△雑談のみ	△雑談のみ
(4)	漫才-役割データ	-	○ツッコミと冗談を言う	○ツッコミと冗談を言う

授業改善のための意見を引き出すAIインタビュアー

中西　勇輝†　　桑田　若菜††　　三林　亮太††　　山本　祐輔†††　　大島　裕明††

† 兵庫県立大学社会情報科学部　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学社会情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
††† 名古屋市立大学データサイエンス学部　〒467–8501 愛知県名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †{fa20m056,ad23g022,af22h007}@guh.u-hyogo.ac.jp, ††yusuke_yamamoto@acm.org,
†††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，授業改善のための意見を引き出すAIインタビュアーを提案する．授業改善において，学生からの要望が書かれた授業評価アンケートは重要な参考資料である．しかし，授業評価アンケートの回答内容は充実しておらず，特に自由記述においては回答内容が単調になりやすい．回答内容を充実させる方法として，インタビューが有効であることが知られている．そこで，本研究では，LLMを用いたAIインタビュアーを提案することで，より充実した回答を学生から引き出す課題に取り組む．提案手法では，Retrieval-Augmented Generation（RAG）を用いて，授業内容に沿った，学生が回答しやすいインタビューを実現する．評価では，兵庫県立大学の学生に対して，本研究で構築したインタビュアーと対話をしてもらう実験を行った．結果として,「特にありません」のような回答をした場合でも，授業資料内のトピックを提示しながら質問をすることで学生から意見を引き出すことができたため，現状の授業評価アンケートでは得られないような意見を引き出すことができた．

キーワード　LLM，対話システム，ユーザ支援，インタラクティブシステム，インタフェース・インタラクション

1　はじめに

学生からのフィードバックは授業の改善において重要な情報である．学生からのフィードバックの代表例として，授業評価アンケートがある．授業評価アンケートとは，授業の質を向上させることを目的とした，多様な観点から授業を改善するための取り組みである [1].

しかし，授業評価アンケートは回答率を担保するために簡易な回答形式であることが多い．たとえば，そこで，簡易な回答形式ではなく，学生と対話をするインタビュー形式で回答してもらうことで，詳細な授業の評価を取得することが可能であると考えられる．インタビューとは，2人かそれ以上の間での会話で，特定のテーマやトピックについて質問と回答の形式で情報を収集する対話的なコミュニケーションの手段である．インタビューを活用した研究も多数存在する [2] [3]．インタビューは，インタビュアーの技量が問われるものであり，インタビューのスキルや知識が無いと，相手から必要な情報を聞き出すことができない．しかし，学生と1人ずつこのような対話をしていくのは現実的に不可能である．

そこで，本論文では人間がインタビューをする代わりに，LLMであるChatGPT [1] を用いて，授業改善のための意見を引き出すAIインタビュアーを提案する．ChatGPTは対話形式で人間とやりとりをすることができる．このChatGPTを授業評価アンケートのインタビュアーとして用いることで，人間が学生と1人ずつ対話をすることなく，授業改善のための意見を引き出すことができる．

1：https://openai.com/blog/chatgpt

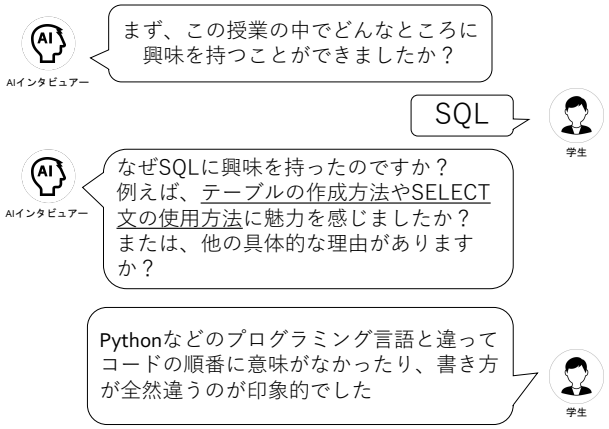


図1　AIインタビュアーとの対話例

図1にAIインタビュアーによる実際の対話例を示す．AIインタビュアーがまず学生に対して質問する．そして学生が質問に対して回答する．RAGを用いて学生の回答に関連するトピックを提示しながら，学生の回答を深掘りする質問を投げる．このようにして意見を引き出すインタビューをおこなっていく．

2　関連研究と基本的事項

2.1　LLMを活用した生成

LLMは，膨大なテキストデータを用いて訓練された，汎用的な言語モデルである．LLMは，テキスト生成，要約，翻訳などのさまざまな言語タスクを行うことができ，LLMの活用が広まっている．代表的なLLMに，GPT-3 [4] やBERT [5] などが挙げられる．

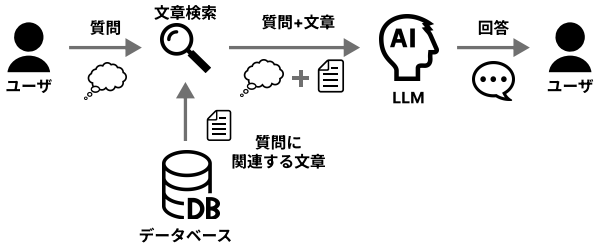


図 2　LLM を用いた RAG の概要

LLM を活用する研究も多く存在する．Acharya ら [6] は，オープンソースの LLM である Alpaca [7] を利用して，複数の特徴を考慮した詳細な映画の説明を生成した．その結果，LLM による映画説明文生成は，ウェブスクレイピングによる説明文生成と同等の結果を示し，大きな可能性を示した．

Montagna ら [8] は，GPT-3 [4] を利用して，プライバシーとセキュリティの懸念に対処しつつ，慢性患者をサポートする LLM ベースのチャットボットシステムを開発するための一般的なアーキテクチャを提案した．

落合ら [9] は，OpenAI 社が提供している GPT-4 [10] を基盤とした GPT4-turbo を用いて，検索システムの検索結果上位 n 件に対し，検索結果全体の要約とユーザーの目的である情報に焦点を当てた要約を生成する手法を提案し，目的の情報を迅速かつ容易に特定できることを示した．

本研究において，LLM である ChatGPT を利用し，対話，質問の生成をおこなう．

2.2　RAG を用いた生成

RAG [2] とは，Retrieval-Augmented Generation の略であり，検索により強化した文書生成とも呼ばれる．RAG は，外部の知識ベースから事実を検索して，最新の正確な情報に基づいて LLM に回答を生成させることで，ユーザーの意図を LLM の生成プロセスに組むこむという AI フレームワークである．RAG の全体像を図 2 に示す．ユーザからの入力文を受け取り，その入力文に関連する文書をデータベースから検索する．検索で得られた文書とユーザからの入力文を LLM に入力し，回答を生成させる．

LLM は時に一貫性に欠けた振る舞いをし，質問に対して正確に答えることもあれば，学習データに含まれている誤った情報を生成することもある．RAG は，LLM が持つ知識の内部表現を補うために外部の情報源と言語モデルを繋げることで，LLM が生成する回答の質を向上させることができる技術である．RAG は，ファインチューニングのような特定のタスクやデータセットに適応させるための学習を必要としない．

Lewis ら [11] は，知識コーパスから文書を抜き出す Retriever と，文書から答えを生成する Generator とで構成される RAG を提案した．実験では，Open-domain の QA タスクにおいて，Natural Questions(NQ)，TriviaQA(TQA)，WebQuestions(WQ)，CuratedTrec(CT) を使い，T5 [12] よりもスコアが大きく上回っており，REALM [13] も上回った．

Levonian ら [14] は，中学生の数学に関する質問に対する回答を LLM によって生成するために，数学教科書の内容を検索する RAG を利用したシステムを構築し，評価した．人間は RAG を用いて生成された回答を好むが，回答が教科書の内容にあまりにも忠実な場合はそうではないことを示した．

高野ら [15] は，ChatGPT を利用して，運用報告書における市況コメントと見通しを自動生成するツールを開発した．RAG のアプローチにより，資産リターンと市場動向に関連する記事を基に ChatGPT に市況コメントや見通しコメントを生成させ，資産運用業界におけるレポート作成の効率化や品質向上に寄与する可能性を示した．

本研究では，PDF の授業資料の内容を検索して，ChatGPT を LLM に採用し，質問，対話の生成をおこなう．

2.3　教育における AI の利用

教育分野においても AI が活用されている．

Câmara ら [16] は，学習支援のための検索エンジンを開発した．検索エンジン内において，ある単語を検索すると，その単語と共に学習するとより学習が深まるような関連する単語を表示するようにした．結果，検索行動には有意な影響を与えることを示した．

また，Roy ら [17] は，ウェブ検索エンジンの検索インタフェースに，ハイライトとメモを取るツールを構築した．結果，ハイライトを表示した参加者は，標準的なウェブ検索インターフェースを利用した参加者と比較して，34%多くのサブトピックをカバーしたことを示した．

木村ら [18] は，教育支援のシステムにおいて，システムに愛着を持たせることでシステムの使用を継続させることに着目した．LLM にファインチューニングを行い，ペルソナを持った雑談対話システムを開発した．結果，ペルソナを持たせた雑談は，対話システムへの愛着を高める上で有用であることを示した．

3　授業改善のための意見を引き出す AI インタビュアー

3.1　概　　要

提案手法では，学生からの回答を入力に，それに対する意見を引き出す質問を出力する．基本的には，ChatGPT をベースとした質問生成を行う．提案手法では，まず，RAG を用いた質問の生成を行う．RAG では，授業評価を引き出す対象の授業資料から，文書を抽出する．その各文書に対して埋め込みベクトルを計算し，その埋め込みベクトルをベクトルストアに格納する．学生と対話をおこない，学生の回答と類似度の高い文書を検索する．授業資料内における意見を引き出す質問を生成するために，ChatGPT に検索した授業資料の文書を与える．その文書をもとに ChatGPT に質問文を生成させる．生成された質問に対する相手の回答を ChatGPT が受け取り，その回答と関連する授業資料の文書をもとに意見を引き出す質問を生成さ

2：https://www.ibm.com/blogs/solutions/jp-ja/retrieval-augmented-generation-rag/

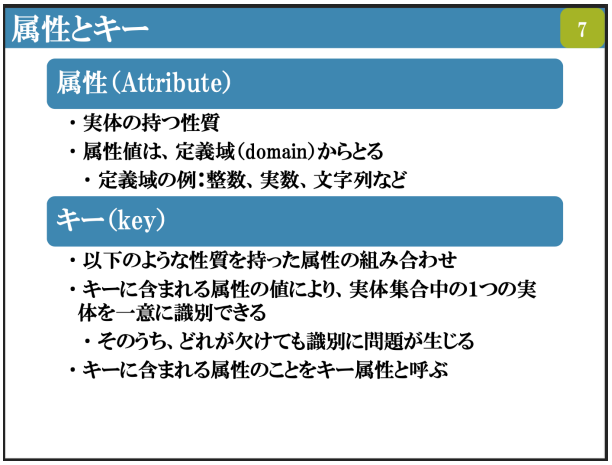


図 3　情報マネジメントの授業資料の 1 ページ

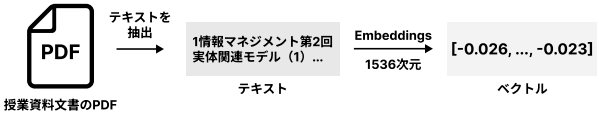


図 4　テキストの抽出と Embeddings の概要

せる．

3.2　授業資料から情報を取り出すプロセス

学生から授業評価を引き出す会話の中で，学生が入力した文章と関連する文書だけを授業資料から取得する．本論文では，兵庫県立大学の「情報マネジメント」の授業資料を扱っている．参考画像として，「情報マネジメント」の授業資料の 1 ページを，図 3 に示す．

PDF から LangChain の PyPDFLoader を用いて PDF 内の文書を抽出する．たとえば，図 3 においては，「7 属性とキー 属性（ Attribute ） ● 実体の持つ性質 ● 属性値は、定義域（domain）からとる ● 定義域の例：整数、実数、文字列など キー（key ） ● 以下のような性質を持った属性の組み合わせ ● キーに含まれる属性の値により、実体集合中の 1 つの実 体を一意に識別できる ● そのうち、どれが欠けても識別に問題が生じる ● キーに含まれる属性のことをキー属性と呼ぶ」というような文として抽出される．図 4 に PDF からテキストを抽出し，ベクトル化する流れを示す．一つの PDF に含まれるすべての文書を抽出して一つに結合する．テキストは 1000 文字から 7000 文字程度のものが存在する．Azure OpenAI Service の埋め込みモデル「text-embedding-ada-002 [3]」を用いて，1536 次元の Embeddings（埋め込みベクトル）を計算する．一つに結合した文書の Embeddings（埋め込み）を計算する．これを「情報マネジメント」のすべての授業資料 13 件の PDF に行い，それぞれ 13 件のテキストとベクトルが得られる．

「情報マネジメント」の授業資料 13 件の中から AI インタビュアーとの対話中に学生が入力した文と類似している文書をベクトル検索する．ベクトル検索には Faiss を利用する．学生の回答が入力されると，「text-embedding-ada-002」を用いて，

3：https://openai.com/blog/new-and-improved-embedding-model

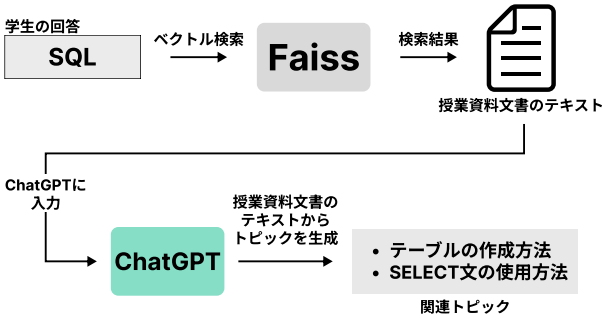


図 5　トピック生成の概要

その学生の回答の埋め込みベクトルを計算する．Faiss でコサイン類似度により，クエリの埋め込みベクトルと類似した埋め込みベクトルを検索する．検索結果は，インデックスのリストとして返され，その返されたインデックスのリストを基に文書リストから該当する授業資料文書を 1 件取得する．

3.3　トピック生成

学生に意見を引き出す質問を投げるために，トピックを提示しながら深掘りをする質問を生成させる．たとえば，図 1 にあるように，AI インタビュアーの「まず、この授業の中でどんなところに興味を持つことができましたか？」という質問に対して，学生が「SQL」と回答した際に，「なぜ SQL に興味を持ったのですか？」だけでなく，「例えば、テーブルの作成方法や SELECT 文の使用方法に魅力を感じましたか？」というように，学生の回答に関連するトピックを提示する質問をするといった具合である．

まず，学生の回答を「text-embedding-ada-002」を用いて，ベクトルに変換する．その学生の回答文のベクトルでベクトル検索をおこなう．13 個の授業資料の文書の中からコサイン類似度において，最も類似度の高い授業資料の文書を一つ取得する．以下のプロンプトによって，トピックを生成させる．

トピック生成プロンプト

> 入力文書からサブトピックを抽出してください。サブトピックを「〇〇、〇〇、〇〇」のような形で必ず 5 つだけ出力してください。

たとえば，学生が「SQL」と回答したとする．その場合，ベクトル検索によって，「情報マネジメント第 10〜 11 回 SQL（2）」の授業資料に含まれるすべての文書が得られる．その文書をもとに生成されるトピックは「1. Google アカウントの用意 2. Google Colaboratory の場所 3. ノートブックのアップロード方法 4. テーブルの作成方法 5. SELECT 文の使用方法」というような文が出力される．

3.4　質問生成

3.3 でも述べたように，学生に意見を引き出す質問を投げるために，トピックを提示しながら深掘りをする質問を生成させる．以下に質問生成のプロンプトを示す．

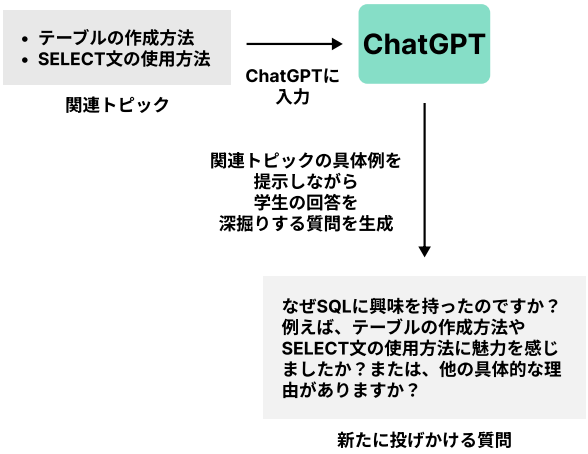


図 6　質問生成の概要

質問生成プロンプト

あなたはインタビュアーです。参考文書を基に具体例を抽出して、その具体例を提示しながら入力文の理由を聞いてください。「なぜ〇〇だと思ったのですか？例えば、(参考文書の具体例) ですか？または、(参考文書の具体例) ですか？」のような文章を出力してください。

プロンプトにおける参考文書として，3.3 で生成したトピックを入力する．学生の回答を入力文とし，3.3 で生成したトピックを提示しながら意見を引き出す質問を生成させる．質問を生成する流れを図 6 に示す．たとえば，学生が「SQL」と回答した場合には，3.3 で挙げた「1. Google アカウントの用意 2. Google Colaboratory の場所 3. ノートブックのアップロード方法 4. テーブルの作成方法 5. SELECT 文の使用方法」のような生成されたトピックを基にして,「なぜ SQL に興味を持ったのですか？例えば、テーブルの作成方法や SELECT 文の使用方法に魅力を感じましたか？または、他の具体的な理由がありますか？」というような質問文が生成される．

4　実験と結果

4.1　実験の概要

本研究で提案した，授業改善のための意見を引き出す AI インタビュアーの有用性を検証するために，大学生に被験者となってもらい，実験を行った．兵庫県立大学の「情報マネジメント」という授業を対象におこなった．今回の実験では,「情報マネジメント」を受講した兵庫県立大学の学生に実験に参加してもらった．RAG の効果を明らかにするために,「情報マネジメント」を受講した学生を半分に分け，グループ別にインタビューをおこなった．表 1 に実験グループの説明を示す．兵庫県立大学の UNIVERSAL PASSPORT 上で，グループごとに LINEBot の QR コードを配布し，LINE ヤフー株式会社が運営するモバイルメッセンジャーアプリケーション「LINE」を用いて，LINEBot の AI インタビュアーによるインタビューに参加してもらった．また，インタビュー回答後に自由参加のアンケートに答えてもらった．

表 1　実験グループの説明

グループ A	授業資料を基にした RAG を利用していないインタビュー
グループ B	授業資料を基にした RAG を利用したインタビュー

表 2　インタビューの項目

項目 1.	この授業の中でどんなところに興味を持つことができましたか？
項目 2.	この授業でどんな内容が難しかったですか？
項目 3.	この授業で印象に残っているトピックを教えてください。

表 3　インタビュー回答後アンケートの内容

	質問内容
Q1	今回利用したシステムは、授業資料の内容を考慮できていたと思いますか。
Q2	授業評価アンケートの代わりとして、このシステムをまた利用したいと思いますか。
Q3	このシステムに限らず、授業評価アンケートを対話システムを利用して回答したいと思いますか。

4.2　実験手順

表 2 は，AI インタビュアーが学生に質問するインタビューの項目及び順番を示している．その質問における学生の回答に対して，さらに深掘りするような質問を投げかける．今回の実験では学生の回答を深掘りする回数は 2 回に固定している．質問を全て聞き終えると，インタビューが終了する．

インタビューが終了すると，自由参加のアンケートの URL を送信した．アンケートの質問内容を表 3 にしめす．回答には 5 段階の尺度を用いて，1 を「そう思わない」，5 を「そう思う」とした．

4.3　インタビュー結果

グループ A とグループ B のそれぞれの対話結果の例の一部を示す．

グループ A におけるインタビューの例を以下に示す．

質問: それではインタビューを始めさせていただきます！まず、この授業の中でどんなところに興味を持つことができましたか？

回答: データベースの構造とそこから必要な情報を取り出す方法

質問: なぜデータベースの構造とそこから必要な情報を取り出す方法に興味を持ったのでしょうか?例えば、参考文書の「テーブルとカラムの設計」という具体例がありますが、それと関連があるのでしょうか?または、他に具体的な事例はありますか?

回答: 情報化で大量のデータが蓄積されるため、そこから必要な情報を効率的に取り出す方法を知っていることは今後必ず必要になると感じたから

質問: なるほど、情報化によって大量のデータが蓄積されることで，効率的に必要な情報を取り出す方法の重要性を感じたのですね。具体的な事例として、参考文書の中で挙げられている

「ビッグデータ処理技術の進化」という話題が関連しそうです。たとえば、ビッグデータを処理するための分散処理や並列処理、クラウドサービスの活用などが登場しており、これらの技術を使うことで膨大な量のデータから効率的に必要な情報を抽出することが可能となっています。
他にも、参考文書の中には「データ分析と予測のための機械学習」というテーマもあります。機械学習の手法を用いることで、データベース内のデータから傾向やパターンを抽出し、予測モデルを構築することができます。これによって、必要な情報を効果的に選別し取り出すことができます。
以上の具体例から、「データベースの構造とそこから必要な情報を取り出す方法」に興味を持ったのでしたら、例えば、ビッグデータ処理技術の進化や機械学習の活用などを参考にすることで、効率的にデータを抽出する方法を学ぶことができるかもしれませんね。
回答: そうですね

グループ B におけるインタビューの例の一部を以下に示す。
質問: それではインタビューを始めさせていただきます！まず、この授業の中でどんなところに興味を持つことができましたか？
回答: SQL
質問: なぜ SQL に興味を持ったのですか？例えば、テーブルの作成方法や SELECT 文の使用方法に魅力を感じましたか？または、他の具体的な理由がありますか？
回答: Python などのプログラミング言語と違ってコードの順番に意味がなかったり、書き方が全然違うのが印象的でした
質問: なぜ SQL の書き方やコードの順番に興味を持ったのですか？例えば、Python などのプログラミング言語と比べて順番や書き方が異なることによる特性や利点を感じたのでしょうか？または、他の具体的な理由がありますか？
回答: はい、そうです
質問: ありがとうございます！次に、この授業でどんな内容が難しかったですか？
回答: 特にありません
質問: ありがとうございます！では、この授業で難しいと感じる部分がなかったということですね。では、逆に言うと、例えば「Google Colaboratory とは」というトピックはわかりやすかったですか？ or 難しかったですか？
回答: 少し難しかったです
質問: なぜ Google Colaboratory が難しかったと感じたのでしょうか?例えば、Google Colaboratory でノートブックをアップロードする方法が理解しづらかったですか？または、テーブルの作成とデータ型の扱いが難しかったと感じましたか？
回答: Google colaboratory でなぜ SQL が書けるのか、ローカル環境だと書けないのかがよく分かりませんでした

グループ A のインタビュー結果を見ると,「ビッグデータ処理技術の進化」などの授業では扱っていないようなトピックを提示してしまっている．だが，RAG を利用していない場合でも，学生に深掘りする質問を投げているため,「情報化で大量のデータが蓄積されるため、そこから必要な情報を効率的に取り出す方法を知っていることは今後必ず必要になると感じたから」という意見を引き出すことができていた.

表 4　アンケートの結果

	グループ A	グループ B
Q1	4.0	4.0
Q2	3.0	2.5
Q3	5.0	4.5

同様に，グループ B のインタビュー結果を見ると，学生の「SQL」という回答に対し，深掘りをする質問を投げかけることで「Python などのプログラミング言語と違ってコードの順番に意味がなかったり、書き方が全然違うのが印象的でした」という意見を引き出すことができた．さらに，AI インタビュアーが「この授業でどんな内容が難しかったですか？」というふうに質問した際に，学生が「特にありません」と回答していたが，それに対し，AI インタビュアーは「Google Colaboratory」という授業で扱った内容を提示し，難しかったかどうかを質問することで，学生から「少し難しかったです」という回答を引き出すことができた.「少し難しかったです」という回答に対しても深掘りすることで,「Google colaboratory でなぜ SQL が書けるのか、ローカル環境だと書けないのかがよく分かりませんでした」という意見を引き出すことができた．このように授業内容のトピックを提示しながら質問をすることで，現状の授業評価アンケートでは得られないような意見を引き出すことができた.

4.4　インタビュー回答後アンケートの結果

インタビュー回答後アンケートは，グループ A は 1 名，グループ B は 2 名の方が回答した．インタビュー回答後アンケートの結果を表 4 示す．アンケート結果は，各実験参加者が回答した評価値の平均で示している．Q2 の値を見ると，どちらのグループにおいても値は低いものとなっている．生成された質問が長くなりすぎたため，低い値となったと考えられる．Q3 の値を見ると，どちらのグループにおいても値が高く，現状の授業評価アンケートの代わりに対話システムを利用して回答するという方法も，授業改善のための一助になるのではないかと考える.

今回の実験では，参加していただけた人数が少なかったため，今後より多くの人に参加してもらうことが必要である.

5　まとめと今後の課題

本論文では，授業改善のための意見を引き出す AI インタビュアーを提案した．現状の授業評価アンケートでは，十分に学生の意見を引き出すことができていない．そこで，本研究では ChatGPT と Retrieval-Augmented Generation(RAG) のアプローチを用いて，学生と対話を行い，学生から授業改善のための意見を引き出す AI インタビュアーを提案した．実験では，AI インタビュアーを利用して，学生にインタビューに協力して

もらった．対話を行うことで，現状の授業評価アンケートでは引き出すことができないような情報も引き出すことができた．ただ，質問文の精度としては，一部日本語として不自然な文になっている場合もあったので，精度向上にも取り組みたいと考えている．

今後は複数の授業で実験を行い，授業を担当する先生などに授業改善のための意見を引き出せていたかどうかなどを評価してもらう方法も検討する．また，今回の提案手法では，学生に対して深掘りする回数を固定していたが，深掘りするかどうかを ChatGPT に判断させてインタビューを行う方法も考えている．ChatGPT などの LLM にインタビューのテクニックを適切に与えて，AI のインタビュアーを構築することも課題である．

謝　　辞

本研究は JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] 文部科学省．令和 3 年度の大学における教育内容等の改革状況について（概要）, p. 25. 2021. https://www.mext.go.jp/content/20230908-mxt_daigakuc01-000031526_1.pdf, 2024 年 1 月 9 日閲覧.

[2] 小泉美佐子, 伊藤まゆみ, 宮本美佐. 老年看護学の対象理解にライフヒストリー・インタビューをとり入れた学習効果. 老年看護学, Vol. 5, No. 1, pp. 140–146, 2000.

[3] 簑原文子, 畑野相子. 高齢者理解を目的としたライフインタビューの効果. 滋賀医科大学看護学ジャーナル, Vol. 12, No. 1, pp. 27–30, 2014.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[6] Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. LLM Based Generation of Item-Description for Recommendation System. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1204–1207, 2023.

[7] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[8] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, p. 205–212, 2023.

[9] 落合瞬也, 松井藤五郎. LLM を活用した業務記録検索システムの検索結果の要約. 人工知能学会第二種研究会資料, Vol. 2023, No. SAI-048, p. 12, 2023.

[10] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[13] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020.

[14] Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*, 2023.

[15] 高野海斗, 中川慧, 藤本悠吾. Chatgpt を活用した運用報告書の市況コメントの自動生成. 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-031, pp. 61–67, 2023.

[16] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. Searching to Learn with Instructional Scaffolding. In *Proceedings of the 6th Conference on Human Information Interaction and Retrieval*, pp. 209–218, 2021.

[17] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 6th Conference on Human Information Interaction and Retrieval*, pp. 229–238, 2021.

[18] 木村江梨花, 佐藤美唯, 小原百々雅, 倉光君郎. 対話システムへの愛着を高める談話コーパスの実現に向けて. 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023.

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 13:30 ~ 15:40 | T2-A オンライン（Zoom Events)

データ基盤・性能評価

座長:杉浦 健人(名古屋大学)
コメンテータ:小山田 昌史(NEC)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T2-A-5-01] 自律分散型データ統合技術Dejimaの性能分析

*吉田 凌河[1]、肖 川[1,2]、鬼塚 真[1] (1. 大阪大学 情報科学研究科 鬼塚研究室、2. 名古屋大学)

14:00 ~ 14:25
[T2-A-5-02] 不揮発メモリを対象とする空間索引のSSDとの性能比較

*吉岡 弘隆[1]、合田 和生[2]、喜連川 優[2] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所)

14:25 ~ 14:50
[T2-A-5-03] モバイル環境における圧縮処理を用いたIoTシステムの通信性能向上のための検討

*伊藤 千紗[1]、竹房 あつ子[2]、中田 秀基[3]、小口 正人[1] (1. お茶の水女子大学、2. 国立情報学研究所、3. 産業技術総合研究所)

14:50 ~ 15:15
[T2-A-5-04] ROS 2準拠ロボット及びクラウドを用いたIoTアプリケーションのためのZenohの通信性能評価

*佐々木 怜名[1]、竹房 あつ子[2]、中田 秀基[3]、小口 正人[1] (1. お茶の水女子大学部人間文化創成科学研究科理学専攻小口研究室、2. 国立情報学研究所、3. 産業技術総合研究所)

15:15 ~ 15:40
[T2-A-5-05] 【技術報告】広告配信における安定して拡張性のある大量データ処理基盤の必要性と活用

大澤 昂太 (株式会社マイクロアド システム開発部)

自律分散型データ統合技術 Dejima の性能分析

吉田 凌河† 肖 川† 鬼塚 真†

† 大阪大学大学院情報科学研究科 〒 565–0871 大阪府吹田市山田丘 1-5
E-mail: †{yoshida.ryoga,chuanx,onizuka}@ist.osaka-u.ac.jp

あらまし データ統合には，大域的なコンポーネントに基づく中央集権型のシステムとピアに基づく自律分散型のシステムがある．自律分散型で大域的な一貫性を保証するシステムとして Dejima が挙げられる．Dejima のロッキングプロトコルとして，2PL，C2PL，Adaptive 2PL が実装されている．先行研究では，ある特定の単純な環境でスループットのみを用いてそれぞれのロッキングプロトコルの性能を評価している．本研究では，より複雑な環境で様々な観点を用いて評価することで，それぞれのロッキングプロトコルの特徴を明らかにする．調査の結果，C2PL がピア間によるスループットの偏りが小さいことや見かけ上のスループットは 2PL に劣ること等を示した．

キーワード データ統合，分散トランザクション制御，双方向変換、ロッキングプロトコル

1 はじめに

大企業や大きな組織において，それぞれの部門でデータを管理し，他と連携することなく独自のアプリケーションやデータベースシステムを設計・構築していることは珍しくない．しかし，大企業の合併や買収，病院と薬局，ホテルと航空会社など，それらのシステムを簡単に連携させることが出来れば便利であり，その需要は大きい [1]．また，複数の研究者が独立してデータセットを作成する大規模な科学プロジェクトや，それぞれが独自のデータソースを持つ政府機関の間でのデータ統合などにおいても大変重要な役割を果たす [2]．このようなデータ統合を実現するには，データベース間のスキーマを一致させる必要があったりする等問題が多く，多くの研究が行われている [3]．

データ統合はグローバルスキーマに基づく中央集権型のものとピアに基づく自律分散型のアーキテクチャに分類することが出来る．前者は主に企業での利用を想定しており，単一のグローバルスキーマを持つ大域的なコンポーネントが他のデータベースサーバーからの通信を受け付け，データを管理する [4–6]．一方後者は，単一のグローバルスキーマを決定することが困難な場合を想定したもので，大域的なコンポーネントを設けず，2 つのピア間での通信を元にデータをカスケード的に伝搬していくものである．しかし，実際には，この 2 つに分類できない両者のシステムを組み合わせたようなシステムも存在する．

中央集権型のデータ統合を用いたアプリケーションの場合，大域的なコンポーネントが何らかの障害でダウンしてしまったとき，システム全体が使用不可能となったり，大域的なコンポーネントでのデータ量や計算量が多いというデメリットがある．一方，自律分散型のデータ統合の場合，あるピアがダウンしても他のピアが持っているデータを使ってシステムを維持することが可能であり，アプリ提供者が処理すべきデータ量の削減も可能である．

しかし，Piazza [7,8] や ORCHESTRA [9,10] のように自律分散型アーキテクチャの多くが，システム全体でのデータの一貫性を意識しない限定的な利用を想定しており，各ピア内だけのデータの一貫性は保証するものの全てのピアにわたった大域的なデータの一貫性は保証しない．これはより一般的な利用を考えた場合，大きな問題となる．そこで，大域的な一貫性を保証するデータ統合アーキテクチャとして，Dejima が提案されており [1,11]，ライドシェアリングにおける利用 [12] やプロトタイプも提案されている [13]，

Dejima では，それぞれのピアが更新可能ビューを用いることで，ピア間のデータ共有をする．そのため，各ピアで用いるデータベースの種類やデータ構造が異なっていても，ピア間で共有したいビューの定義さえ同じであればデータ共有が可能である．また，ビューの共有によるデータ交換をすることにより，データ交換のための複雑なプログラムを書く必要がなく，ベーステーブルとビューの変換戦略を定義するプログラムだけを書けば良いため，データ統合が比較的簡単であることもメリットとして挙げられる．

Dejima では大域的一貫性を保証するためのロッキングプロトコルが複数実装されているが，そのプロトコルによる性能差が様々な環境で検証されているわけではなく，ある一定の単純な環境での性能差についてしか研究されていない．そこで本研究では，より複雑な環境で，様々な評価観点を用いて性能評価を行う．

2 Dejima

本章では，Dejima アーキテクチャについて説明する．

2.1 概 要

Dejima アーキテクチャは以下の 4 つの構成要素から成る．

- ピア
- ベーステーブル
- Dejima テーブル
- Dejima グループ

ピアはサービスプロバイダなどデータを持つ主体を表し，それ

ぞれ固有のローカルデータベースを管理する．このアーキテクチャではデータを共有したいピア同士で Dejima グループを形成する．各ピアは参加する Dejima グループごとに Dejima テーブルを定義し，同一の Dejima グループに属する他のピアと対応する Dejima テーブルを共有する．

Dejima テーブルは，データベース内にあるベーステーブルから導出されるビューである．ビューであるが，更新可能であり，Dejima テーブルの更新に対応して導出元のベーステーブルを適切に更新することができる．この更新戦略は双方向変換言語 [14,15] を用いて定義される．つまり，ベーステーブルと Dejima テーブルは双方向の変換が可能になっている．

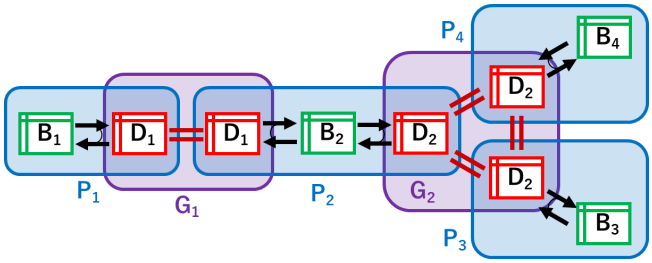


図 1　Dejima アーキテクチャの例

図 1 は Dejima アーキテクチャの一例であり，4 つのピア $P_1, \cdots, P_4$ とそれに対応するベーステーブル $B_1, \cdots, B_4$，2 つの Dejima グループ $G1, G2$ とそれに対応する Dejima テーブル D_1, D_2 が表されている．同一の Dejima グループに属するピアの対応する Dejima テーブルは同期しているため，ここで例えば P_1 の Dejima テーブル D_1 と P_2 の Dejima テーブル D_1 は全く同じ内容になっており．どちらか一方が更新されても，すぐに更新内容がもう一方に伝わり，同一の状態が保持される．

2.2　更新伝搬

Dejima アーキテクチャにおける更新伝搬は次のように行われる．

（1） あるピアにおいてベーステーブルが更新される．

（2） (1) の更新によって，ベーステーブルのビューである Dejima テーブルが更新される．

（3） (2) の更新内容を，同じ Dejima テーブルを保持する他のピアに通知する．

（4） (3) で通知された更新内容をもとに自身の Dejima テーブルを更新し同期する．

（5） (4) で更新した内容をもとに，双方向変換によりベーステーブルが更新される．

（6） 以降，Dejima テーブルの更新が発生しなくなるまで (2)~(5) を繰り返す．

2.3　ロッキングプロトコル

Dejima では，複数のピアにおいてトランザクションが並列実行され，更新が伝搬する．また，大域的なデータの一貫性を保証する必要があるため，排他制御が必要となる．Dejima では，排他制御を実現するロッキングプロトコルとして 2PL [16–19] と C2PL，Adaptive 2PL [20] が採用されている．

2.3.1　2PL

2PL は，順にロックを取得していく成長相と取得したロックを順に解放していく縮退相から成る．並列に実行されるすべてのトランザクションが 2PL に従うことで，すべてのデータが直列化され，他のトランザクションにデータを書き換えられることなどによるデータの不整合を避けることができる．

2.3.2　C2PL

C2PL は，2PL において，トランザクションの開始時にロックの獲得をすべて終わらせる方式である．つまり，成長相がトランザクション実行前に開始し一瞬で終わるような方式である．これにより，トランザクション実行中にロックを新たに獲得することがなくなるため，デッドロックを原因とするトランザクションのアボートを防ぐことができる．

しかし，トランザクション実行前に全てのロックを取得する必要があることから，トランザクション処理に条件分岐が含まれるなど実行時にアクセスするレコードが事前に分からないトランザクションについては，実行前にアクセスされるレコードを特定できないため，基本的に実行できない．ただし，更新が伝搬する可能性のあるレコード全てにロックをかけることで実行することは可能であるが，実際に更新が伝搬しないレコードにもロックをかけないといけないというデメリットがある．

Dejima では，ロックを取得すべきレコードを特定するために FRS (Family Record Set) [20] というものを用いる．FRS は Original record と Derived Records の集合，つまり $FRS = \{Originalrecord, Derivedrecords\}$ と定義される．ここで，Original record は，あるピアのベーステーブルに挿入された新しいレコードを指し，Derived records は，その挿入されたレコードが伝搬して，他のピアで生成されたレコードを指す．

各 Family Record Set には固有の識別子が与えられ，同じ Family Record Set に属するレコードは同一の識別子を持つ．この識別子は，システム全体を通して一意であればなんでも良いが，実装上は Original Record のピア ID，ベーステーブル ID，レコード ID の組み合わせで表現している．識別子は，各レコードの Lineage カラムに追加されるが，1 つである必要はなく複数の識別子を保持することができる．これを応用することで，JOIN クエリにより定義された Dejima テーブルを含んでいたとしても，ロックをかけるべきレコードを適切に特定することが可能である．

事前ロック時は，そのレコードが属する FRS を Lineage カラムを参照することで確認し，ロックを獲得すべきレコード（同じ FRS に属する全てのレコード）を特定する．

2.3.3　Adaptive 2PL

通常の 2PL と C2PL にはそれぞれメリットデメリットがある．C2PL を採用することで，トランザクション実行中のロック獲得失敗によるアボートを防ぐことができ，アボートコストが小さくなる．一方，アボート率が低い環境においては，事前ロックのオーバーヘッドなどが原因となり，スループットが下がる．

Adaptive 2PL はこの問題を解決するために提案されたロック方式で，システムの負荷に応じて 2PL と C2PL を適応的に

切り替えることでスループットの低下を防ぐ．具体的には，定期的に両者のスループットを T_{test} 秒間計測し，現在実行中の手法がもう一方の手法に比べてスループットが低かった場合，もう一方の手法に切り替えることでシステムの負荷に応じて動的に手法を切り替える．ただし，T_{test} 秒間を超えてもトランザクションを実行中の場合は，そのトランザクションが終了するまで待つ．また，各ピアごとの最適な手法を求めるために，全ピアが一斉に計測を開始するのではなく，各ピアが別々の時間に計測を開始し，手法を切り替える．

3 先行研究

三宅らの研究 [20] では，2PL と C2PL，Adaptive 2PL それぞれのロッキングプロトコルを使って実験することで，Dejima における 3 つのロッキングプロトコルによる性能差を明らかにした．しかし，実験環境が全てのピアで同じスキーマ，次数が似ているなど，似たような特徴を持った環境での実験である．また，評価もスループットのみを用いて評価している．

Dejima の各ロッキングプロトコルの性能評価を行う際，スループットは重要な指標であるが，それ以外のアボートの起きやすさやアボートにかかった時間なども重要な指標である．また，Dejima は単一のシステムではなく複数のピアから成るため，システム全体のスループットだけでなく，各ピアにおけるスループットの過度な偏りが起きていないかや，特定のトランザクションが実行されにくくなっていないか等も重要な指標である．

そこで本研究では，異なる特徴を持つピアから構成される環境において，スループット以外のアボート率，それぞれのコミットやアボートにかかった時間，各ピアにおけるスループットやアボート率の違いも評価観点として評価する．

4 実験設定

4.1 トポロジ

本実験のトポロジにはヘテロフィーなものを用いる．そのイメージを 2 に示す．この例では，航空関連のトポロジを想定しており，プロバイダ，アライアンス，アグリゲーターの 3 層から構成される．プロバイダは ANA や JAL などの航空会社に，アライアンスは Expedia や Trip.com などの予約サイトに，アグリゲーターは Skyscanner などのメタサーチエンジンに対応する．

4.2 トランザクション

各ピアで実行されるトランザクションとして以下の 2 つを用意した．

- Update: 予約や価格の変更に対応
- Read: 予約状況の確認や価格の参照に対応

それぞれのトランザクションにおいて，対象にするレコード数を制御することが出来る．また，本来 Skyscanner などのアグリゲータでは予約などを行えないものも存在するが，本研究ではどのピアにおいても Update トランザクションを実行できる

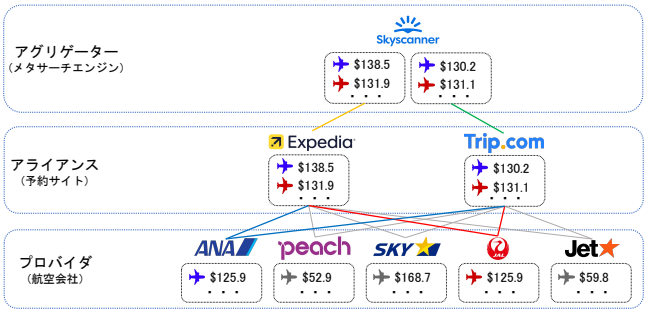


図 2　ヘテロフィー設定のトポロジの例

ものとする．

Update トランザクションでは，各更新が隣のピアに伝搬する．そのときの伝搬経路を表 1 に示す．ここで Pr，Ar，Ag はそれぞれプロバイダ，アライアンス，アグリゲータを表す．

表 1　各ピアにおけるトランザクションの伝搬経路

更新伝搬元	伝搬経路	最大伝搬数
プロバイダ	Pr → Al → Ag	2
アライアンス	Al → Ag Al → Pr → Al → Ag	3
アグリゲータ	Ag → Al → Pr → Al → Ag	4

プロバイダにあるレコードが更新された場合，そのレコードを共有しているアライアンスのレコードが更新され，その更新がアグリゲータにも伝搬する．アライアンスのレコードが更新されたとき，図 3 のように 2 つの経路に分かれる．1 つはアライアンスからアグリゲータに伝搬する経路．もう 1 つは，アライアンスからプロバイダ，他のアライアンス，アグリゲータと伝搬するものである．アグリゲータにあるレコードを更新した場合，そのレコードを持っているアライアンスにまず更新が伝搬し，その後，プロバイダ，他のアライアンス，アグリゲータへと伝搬する．

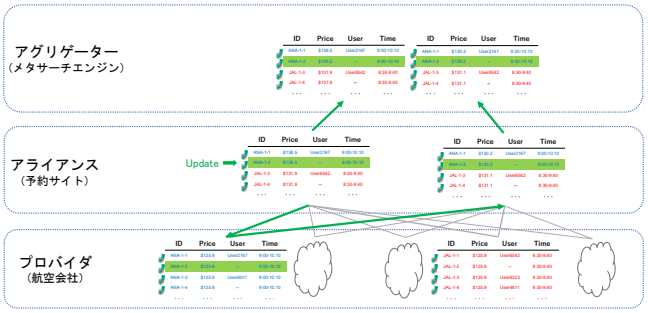


図 3　アライアンスのレコードが更新されたときの伝搬経路

更新処理を行う前に実際には，そのレコードのロックを獲得する必要があるが，複数トランザクションが同時に動いているため，デッドロックが発生する可能性がある．Dejima では，NO WAIT でロックを獲得することでデッドロックを防いでいる．つまり，取ろうとしたレコードが既に他のトランザクションによりロックが獲得されていた場合，即座に自身のトランザクションをアボートする．本研究では，ローカルのデータベー

スに対するロック獲得に失敗してアボートするケースをローカルアボート，他のピアに対するロックを獲得するタイミングでロック獲得に失敗し，アボートするケースをグローバルアボートと定義する．

Update トランザクションの場合，以上のように他のピアにロックを獲得し，更新伝搬する必要があるが，Read トランザクションの場合，自身のピアの共有ロックを獲得し，値を読み取るだけであるため，Update トランザクションに比べると処理がはるかに軽いうえ，グローバルアボートも発生しない．

本実験では，一部を除いてトランザクション Tx_i が失敗したとき，Tx_i が成功するまで Tx_i を実行し直し続ける．Tx_i が成功したときに限り Tx_{i+1} を実行する．

4.3 各種パラメータ

本実験における各種パラメータを表 2 に示す．

表 2　各種パラメータ

パラメータ	デフォルト値	
アグリゲータ数	1	固定
アライアンス数	2	固定
プロバイダ数	5	固定
レコード数/プロバイダ	10	固定
Update:Read	20:80	可変
レコード数/Tx	1	可変

本実験において，アグリゲータ数，アライアンス数，プロバイダ数はそれぞれ 1，2，5 となっている．この設定は規模が小さいが，アグリゲータよりアライアンスの方が，アライアンスよりプロバイダの方が多いという経験則を反映している．1 プロバイダあたりのレコード数が 10 というのは，日本の事業者ごとの平均タクシー台数が 10 台程度というのを参考に設定している．また，Update トランザクションと Read トランザクションの比率は 20:80，1 トランザクションあたりのレコード数は 1 となっている．これらパラメータにおいて，アグリゲータ数やアライアンス数，プロバイダ数は実世界で急激に変化することは少ない．1 プロバイダあたりのレコード数も同様である．一方，Update:Read やレコード数/Tx は，各プロバイダやアライアンスの方針によって比較的大きく変化し得るパラメータである．そのため，本実験では Update:Read とレコード数/Tx を変化させた実験を行う．

5　実験結果と考察

5.1 デフォルト設定での実験結果

まず，各パラメータをデフォルト値に設定し，変更しないまま 120 秒間ベンチマークを動かした結果を図 4，図 5 に示す．

青色で書かれている数字はコミット数を表し，オレンジ色で書かれている数字がアボート数を表す．コミット数の右上に書かれている数字は Update トランザクションのコミット数を表し，右下は Read トランザクションのコミット数を表す．一方，アボート数の右上に書かれている数字はグローバルアボート数

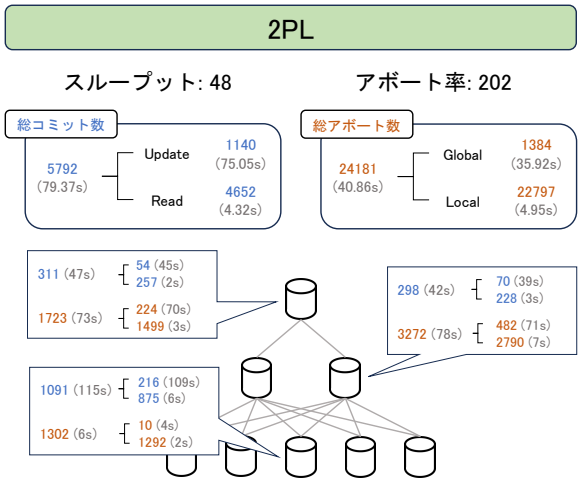


図 4　デフォルト設定での 2PL の実験結果

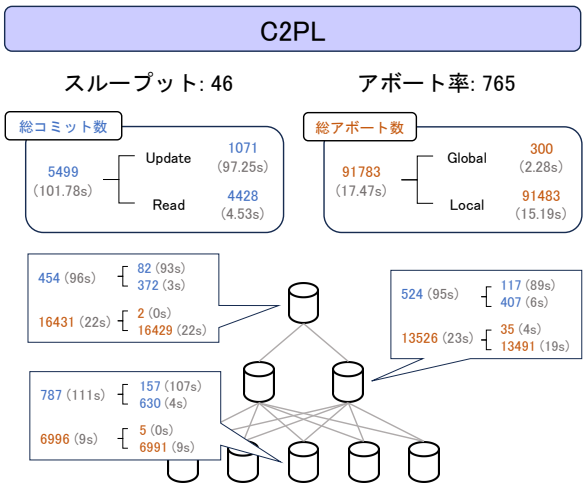


図 5　デフォルト設定での C2PL の実験結果

を表し，右下はローカルアボート数を表す．スループットは 1 秒あたりの総コミット数，アボート率は 1 秒あたりのアボート数を表す．

図 4，図 5 の総アボート数について見ると，C2PL の方がローカルのアボート数が多く，時間も長いことが分かる．これは，Update トランザクションにおいて C2PL では，処理していないレコードでも事前にロックを取るため，2PL より多くのレコードにロックをかけることになり，結果として Read トランザクションがロックを獲得しようとしたレコードが既にロックされている確率が高くなるからだと考えられる．このことから C2PL は Read トランザクションを邪魔する傾向が強いと言える．

続いてピアごとのコミット数の違いについて見ると，2PL ではプロバイダのコミット数は多いのに，アライアンスとアグリゲータのコミット数は少なくなっている．一方，C2PL ではどのピアにおいてもコミット数が同じ程度になっている．言い換えると 2PL の方がピア間によるコミット数の偏りが大きく，C2PL の方が偏りが少ないと言える．これは，トランザクショ

ンの伝搬経路の長さがプロバイダからの Tx ＜アライアンスからの Tx ＜アグリゲータからの Tx となっていることが関係している．つまり，伝搬経路が比較的長いアライアンスとアグリゲータはアボート時，C2PL ではロックを解除するだけで良いが，2PL ではそれまでの処理を全てアボートする必要があり，このアボートコストが高く時間がかかるため，結果として長い経路のアライアンス，アグリゲータのコミット数が少なくなっている．逆に伝搬経路の短いプロバイダからのトランザクションは，アボート時のコストが低いため，オーバーヘッドが少なく，コミット数が多くなっている．この結果より，伝搬経路の違いを考慮せず，単純なスループットだけによる比較を行うと，伝搬経路の短いトランザクションばかりがコミットされ，伝搬経路の長いトランザクションはあまりコミット出来ていなくても，見かけ上そちらの方が性能が良いと判断される可能性がある．

5.2 レコード数/Tx を変化させた実験

デフォルト設定からレコード数/Tx を 1 から 4 に変更した実験した結果を図 6，図 7 に示す．

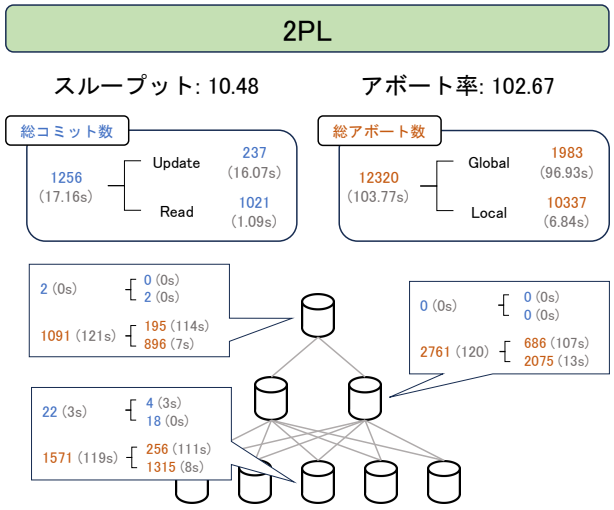


図 6　4 Records/Tx での 2PL の実験結果

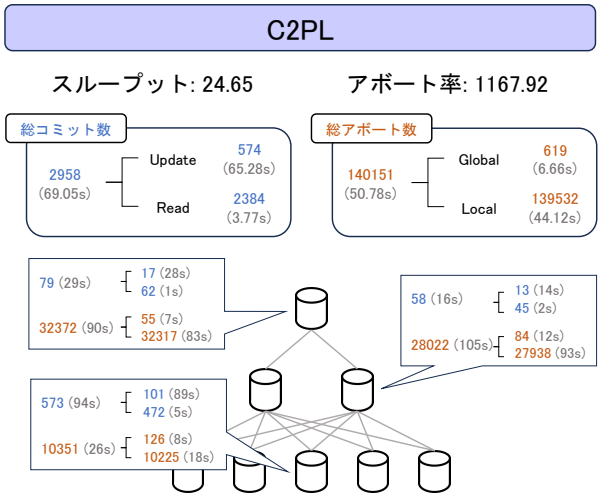


図 7　4 Records/Tx での C2PL の実験結果

2PL において，デフォルト設定での実験結果と比較すると，総コミット数が全体的に約 1/4 倍になっている．これは単にレコード数/Tx が 4 倍になったため，計算量やロックを取るレコード数も約 4 倍になったことで，コミット数が 1/4 に落ちたと考えられる．一方，C2PL については，総コミット数が全体的に約 1/2 倍になっており，2PL よりもコミット数の減少が小さい．なぜ 1/3 や 1/5 のコミット数ではなく，1/2 のコミット数になったかは分からないが，C2PL の事前ロックによる早い段階でのアボート検知が有効に働いたことでコミット数の減少を 2PL より抑えられたと考えられる．この実験結果から C2PL は複数レコードを一度に更新するトランザクションが 2PL に比べて得意だと言える．

5.3 Update:Read とレコード数/Tx を変化させた実験

Update:Read とレコード数/Tx をそれぞれ 0:100 から 100:0 まで，1 から 4 まで変化させた実験結果を図 8 に示す．横軸はトランザクションにおける Update トランザクションの割合，つまり Update 率（Update:Read）を表している．縦軸は 1 トランザクションあたりの対象レコード数，つまり Records/Tx（レコード数/Tx）を表している．各セルの上の数字は C2PL のスループット及び 2PL に対する相対スループット，真ん中の数字は Adaptive 2PL のスループット及び 2PL に対するスループット，下の数字は 2PL のスループットを表している．また，赤っぽい色の部分は C2PL より 2PL のスループットが良いことを表しており，青っぽい色の部分は 2PL より C2PL の方がスループットが良いことを表している．

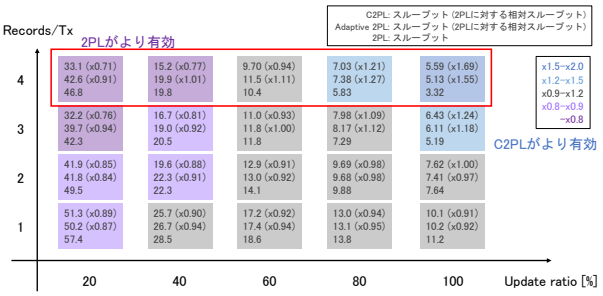


図 8　Update 率とレコード数/Tx を変化させた実験結果

実験結果から Update 率が低いときは 2PL がより有効であり，Update 率が高いときは C2PL の方がより有効であることを示している．これは，Update 率が低いとき，C2PL の事前ロックのオーバーヘッドのため C2PL の性能が悪くなっているのに対して，Update 率が高いときは，事前ロックが有効に働くことでアボート時のコストを小さくしているからであると考えられる．また，レコード数/Tx が大きくなるにつれて，その差異が顕著になっていることも分かる．これは，Update 率が小さいとき，レコード数/Tx が大きくなると事前ロックを取るレコード数が増え，オーバーヘッドも大きくなり，Update 率が大きいときは，レコード数/Tx が大きくなると 2PL のアボートコストがより大きくなるからだと考えられる．

最後に図 8 の赤枠で囲った部分，つまり，レコード数/Tx を

4 に固定して Update 率を変化させた実験結果を図 9 に示す．横軸は Update 率，縦軸は 2PL に対する各ロッキング手法の相対スループットを表している．2PL，C2PL，Adaptive 2PL の結果はそれぞれ青色，紫色，緑色で表している．

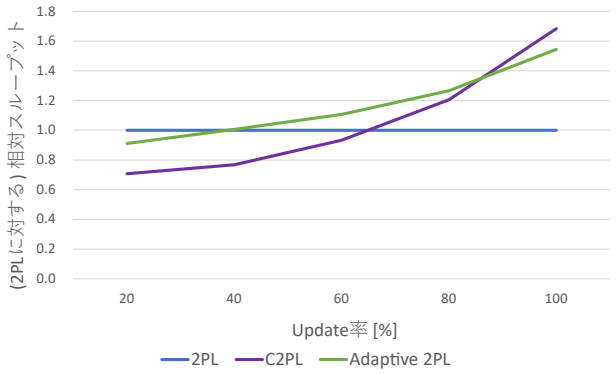


図 9　4 Records/Tx において Update 率を変化させた実験結果

この結果からも Update 率が低い範囲では C2PL の性能が 2PL より悪く，Update 率が高い範囲では C2PL の性能が良いことが読み取れる．また，Adaptive 2PL では，適応的に各手法を切り替えることにより，両者の良いところを取り，性能の低下を防いでいる．Update 率が 65%あたりのとき，2PL と C2PL の性能はほとんど同じなのに，Adaptive 2PL が一番性能が良くなっているが，これは各ピアで最適なロッキング手法に切り替えることにより，より良い性能を出すことが出来ていると思われる．

6 ま と め

本研究では，自律分散型のデータ統合システムである Dejima におけるロッキングプロトコルによる性能の違い，及び各ロッキングプロトコルの特徴を，先行研究より複雑な環境を用意し様々な観点で評価することで，詳細に明らかにした．

実験設定としては，トポロジとしてプロバイダ，アライアンス，アグリゲータから成るヘテロフィーな環境を用意した．トランザクションには Update トランザクションと Read トランザクションを用意し，トランザクションにおける Update トランザクションの割合と 1 トランザクションあたりの対象レコード数を変化させて，実験を行った．

実験の結果，以下の知見が得られた．

a） 2PL

- Read トランザクションが得意
- 短い伝搬経路のトランザクションが得意

– 長い伝搬経路のトランザクションは後回しにされる傾向にある

b） C2PL

- Update トランザクションが得意
- 長い伝搬経路のトランザクションが得意

– ピア間によるスループットの偏りが比較的小さい

– 見かけ上のスループットは 2PL に劣る場合がある

- 広い範囲にロックを取るため Read トランザクションをより邪魔する傾向にある

c） Adaptive 2PL

- 適応的に 2 つのロッキング手法を切り替えることで性能低下を防ぐ

本研究では，アグリゲータ数，アライアンス数，プロバイダ数，1 プロバイダあたりのレコード数をそれぞれ 1，2，5，10 と固定して実験した．これは実世界への応用を考えた時，規模が小さい環境での実験である．また，1 プロバイダあたりのレコード数が 10 というのは，日本の事業者ごとの平均タクシー台数が 10 台程度というのを参考に設定しているが，実際は個人タクシーは 1 事業者当たり 1 台，法人タクシーは 30 台であり，単純な平均値では，この不均一性を再現できていない．

そのため，これらのパラメータを大きく変化させ，規模が大きいシステムを想定したときの性能変化や 1 プロバイダあたりのレコード数に不均一性を持たせた時の性能変化等を確認することが今後の課題として挙げられる．

文　献

[1] Yasuhito Asano, Dennis-Florian Herr, Yasunori Ishihara, Hiroyuki Kato, Keisuke Nakano, Makoto Onizuka, and Yuya Sasaki. Flexible framework for data integration and update propagation: System aspect. In *BigComp*, pp. 1–5. IEEE, 2019.

[2] Alon Y. Halevy, Anand Rajaraman, and Joann J. Ordille. Data integration: The teenage years. In *VLDB*, pp. 9–16. ACM, 2006.

[3] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

[4] Maurizio Lenzerini. Data integration: A theoretical perspective. In *PODS*, pp. 233–246. ACM, 2002.

[5] Alon Y. Halevy. Answering queries using views: A survey. *VLDB J.*, Vol. 10, No. 4, pp. 270–294, 2001.

[6] Jeffrey D. Ullman. Information integration using logical views. In *ICDT*, pp. 19–40. Springer, 1997.

[7] Alon Y. Halevy, Zachary G. Ives, Peter Mork, and Igor Tatarinov. Piazza: data management infrastructure for semantic web applications. In *WWW*, pp. 556–567. ACM, 2003.

[8] Alon Y. Halevy, Zachary G. Ives, Jayant Madhavan, Peter Mork, Dan Suciu, and Igor Tatarinov. The piazza peer data management system. *IEEE Trans. Knowl. Data Eng.*, Vol. 16, No. 7, pp. 787–798, 2004.

[9] Grigoris Karvounarakis, Todd J. Green, Zachary G. Ives, and Val Tannen. Collaborative data sharing via update exchange and provenance. *ACM Trans. Database Syst.*, Vol. 38, No. 3, p. 19, 2013.

[10] Zachary G. Ives, Nitin Khandelwal, Aneesh Kapur, and Murat Cakir. ORCHESTRA: rapid, collaborative sharing of dynamic data. In *CIDR*, pp. 107–118, 2005.

[11] Yasuhito Asano, Soichiro Hidaka, Zhenjiang Hu, Yasunori Ishihara, Hiroyuki Kato, Hsiang-Shang Ko, Keisuke Nakano, Makoto Onizuka, Yuya Sasaki, Toshiyuki Shimizu, Van-Dang Tran, Kanae Tsushima, and Masatoshi Yoshikawa. Making view update strategies programmable - toward controlling and sharing distributed data. *CoRR*, Vol. abs/1809.10357, , 2018.

[12] Yasuhito Asano, Zhenjiang Hu, Yasunori Ishihara, Hiroyuki Kato, Makoto Onizuka, and Masatoshi Yoshikawa. Controlling and sharing distributed data for implementing service

alliance. In *BigComp*, pp. 1–4. IEEE, 2019.

[13] Onizuka Lab. Dejima prototype. `https://github.com/OnizukaLab/dejima-prototype`, 2022.

[14] Van-Dang Tran, Hiroyuki Kato, and Zhenjiang Hu. Programmable view update strategies on relations. *PVLDB*, Vol. 13, No. 5, pp. 726–739, 2020.

[15] Van-Dang Tran, Hiroyuki Kato, and Zhenjiang Hu. Birds – bidirectional transformation for relational view update datalog-based strategies. `https://dangtv.github.io/BIRDS/`, 2020.

[16] Philip A. Bernstein, Vassos Hadzilacos, and Nathan Goodman. *Concurrency Control and Recovery in Database Systems*. Addison-Wesley, 1987.

[17] Christos H. Papadimitriou. *The Theory of Database Concurrency Control*. Computer Science Press, 1986.

[18] Kapali P. Eswaran, Jim Gray, Raymond A. Lorie, and Irving L. Traiger. The notions of consistency and predicate locks in a database system. *Commun. ACM*, Vol. 19, No. 11, pp. 624–633, 1976.

[19] Jim Gray, Raymond A. Lorie, Gianfranco R. Putzolu, and Irving L. Traiger. Granularity of locks in a large shared data base. In *VLDB*, pp. 428–451. ACM, 1975.

[20] 三宅康太, 佐々木勇和, 肖川, 鬼塚真. 統合型データベースにおける適応的 2 相ロックに基づく分散トランザクション制御. In *DEIM*, pp. 1–8, 2022.

不揮発メモリを対象とする空間索引のSSDとの性能比較

吉岡　弘隆†　　合田　和生††　　喜連川　優††

† 東京大学　大学院情報理工学系研究科
〒 153–8505 東京都目黒区駒場 4–6–1
†† 東京大学　生産技術研究所
〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{hyoshiok,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし
概要：不揮発メモリ (PMEM) は揮発性メモリ (DRAM) とほぼ同等のレイテンシーを持ちながら HDD/SSD 同様永続性を持つ．不揮発メモリへの書き込みは OS が非同期で行うが一貫性を担保したい場合は永続性を持たせるために明示的な flush および同期 (fence) が必要となる．空間索引の性能比較を従来デバイス (SSD) と不揮発メモリとで行った．不揮発メモリの基本的な動作特性 (レイテンシー，スループット等) や不揮発メモリ向けの基本的なデータ構造 (B+木やハッシュ索引) とアルゴリズムについては様々な研究がされてきている．しかしながら不揮発メモリを対象とする空間検索構造の実装方式の報告については FBR [11]，PMR [26] などに見られる程度で，十分とは言えない．そこで本研究では不揮発メモリを対象とする空間検索構造の実装方式の検討を行い，従来デバイス (SSD) と不揮発メモリで性能比較を行った．その結果，挿入については SSD より PMEM のほうが 22 倍以上のスループットが高かった．しかしメモリマッピングした場合の検索には顕著な差はみられなかった．
キーワード　データベース技術，不揮発メモリ，PMEM，SSD，空間索引構造

1　はじめに

不揮発メモリと呼ばれるバイト単位アクセス可能なメモリが近年出荷されている [39], [41].

その特徴をまとめると，

- 性能（スループット，レイテンシなど）は SSD (NAND Flash)，HDD より高い，DRAM より低い
- DRAM と異なり永続性がある（電源が遮断されても情報は喪失しない）
- 記憶容量は DRAM より大きい
- バイト単位のアクセスが可能である（DRAM と同様）
- CPU cache coherent である（DRAM と同様）
- DMA （Direct Memory Access）と RDMA （Remote Direct Memory Access）をサポートする（DRAM と同様）
- ユーザ空間でアクセスできる．SSD や HDD のようにアクセスするために system call は必要ない．アクセスするためにカーネルコード，ページキャッシュ，割込などは必要ない

となる．

様々な興味深い特徴があるが，不揮発メモリの製品例である Intel Optane DCPMM も広く普及しているとは言い難く，研究段階での評価にとどまっているのが現状である．

ここでは不揮発メモリの製品例として Intel Optane DCPMM をとりあげ評価する．

不揮発メモリの特徴として SSD や HDD より高速だといわれている．そこで本研究ではまず PMEM と SSD の挿入時のスループットを計測しその性能を確認した．これによって PMEM が SSD より高性能な記憶デバイスとして利用できることを示した．

一方で，従来の揮発性メモリと違い，不揮発であるという特徴はプログラミングモデルの変更を余儀なくする．そのため，その動作特性の違いは単にスループットやレイテンシーだけでなく，アルゴリズムやデータ構造のあり方にも影響を与える．

不揮発メモリは従来のメモリ同様にバイト単位でアクセス可能であるが，CPU から見て不揮発メモリへのアクセスは通常キャッシュ経由でなされる．図 1 に示すようにキャッシュは揮発性なので永続性を必要とする場合は適宜キャッシュ内容を flush する必要がある．Intel x86 の場合，CLFLUSH 命令などによって明示的に行う．flush 関連の命令は CLFLUSH の他，CLWB，CLFLUSHOPT があり，キャッシュをバイパスする Non-temporal 命令もある．また，メモリへの書き込みを同期するために fence が必要になる．

また記憶の永続性を担保する動作についてはファイルシステムでは本質的な問題だが，メインメモリに於いては永続性は存在しなかったので一般的な問題ではなかった．例えばクラッシュコンシステンシーという概念は従来は主にハードディスクなど永続性を持つデバイスにおいて議論されてきたが，揮発性メモリではクラッシュした時点でそもそも永続性を保証しないので一般的には問題とならなかった．しかし不揮発メモリに於いてはストア命令によって記憶が永続化されるので，ディスクへの書き込みと同様な問題が生じる．ここで言うクラッシュコンシステンシーとは，情報の変更を首尾一貫した状態で行うことを指す．

不揮発メモリへのストア命令は必ずしも同期的にメディアに

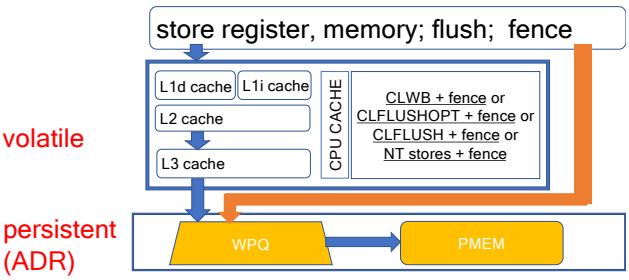


図 1　CPU とキャッシュ，メモリの関係

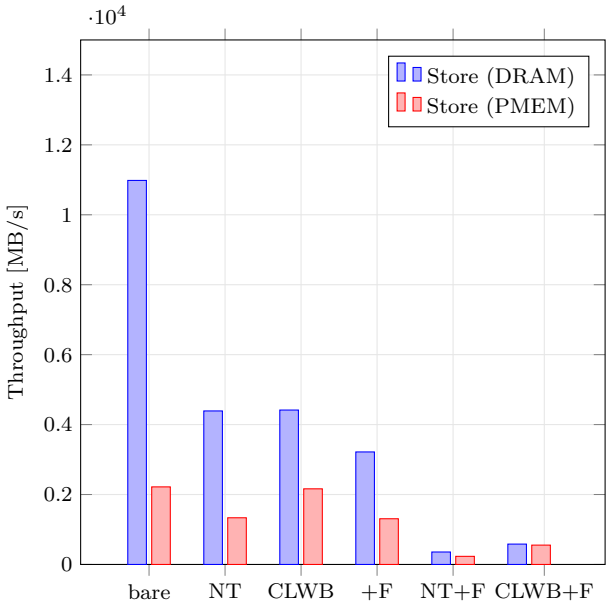


図 2　64 byte sequenial store, bare, NT, CLWB はデフォルトの store 命令および Non-temporal 命令，CLWB 命令を追加したもの，＋ F はさらに FENCE 命令を追加したものを示す [53]

書き込むことを保証しない，通常はキャッシュにのみ書き込み，非同期にメモリ (メディア) に書き込む．そのため，プログラムのストアの順序とメモリへの書き込み順は必ずしも等しくならず，書き込み順の一貫性を担保する仕組みが必要になる．

メモリレベルでの永続性を担保する命令（FLUSH 命令）や書き込み順の同期を取る命令（FENCE 命令）などをプログラマが意識しないと不揮発メモリを利用した正しいプログラム（クラッシュした時点での整合性を持つ）は構築できない．

またこのような各種の flush や fence が性能に与える影響は必ずしも自明ではない．そこでわれわれは，先行研究として pmmeter というマイクロベンチマークを作成し，flush 命令と同期命令が与えるオーバーヘッドを評価した [53]．64 バイトのシーケンシャルアクセスのスループットは DRAM で 10.9GB/s, PMEM で 2.2GB/s だが flush 命令を追加すると 4.4GB/s と 2.1GB/s にそれぞれ低下する．さらに fence 命令を追加すると 0.58GB/s と 0.55GB/s となる (図 2 参照)．flush と fence は大きなコストが発生する．

そのような特性を踏まえ，従来の広く利用されている索引技術を不揮発メモリへ適用する研究が近年活発に行われている．

例えば B+木索引はメモリアクセスとディスクアクセスのレイテンシの大きな違いを前提にファイルをディスクに格納した場合に最適化された構造で，検索時にディスクアクセス回数を削減するようなアルゴリズムになっている．

不揮発メモリ向けの B+木索引 [16], [18], [29] ，ハッシュ索引 [17]，ファイルシステム [38] ，データベースエンジン，基本的な動作特性 [37], [46] などの報告はあるが，空間索引構造についてはまだ十分に検討がされていない．そこで本研究では，不揮発メモリを対象とする空間検索構造の実装方式を検討した．

1.1　PMEM プログラミングモデル

上述したように，PMEM のプログラミングは必ずしも自明ではない．例えば，メモリレベルでの永続性を担保する命令（FLUSH 命令）や書き込み順の同期を取る命令（FENCE 命令）などをプログラマが意識しないと不揮発メモリを利用した正しいプログラム（クラッシュした時点での整合性を持つ）は構築できない．

このように不揮発メモリは揮発性メモリとは異なったプログラミングモデルを持つために従来とは異なるプログラミング上の注意が必要となる．一般的には，上記のような煩雑な処理をアプリケーションプログラマに委ねるのはアプリケーションをアセンブリ言語で記述するようなものなので，動作の詳細を隠蔽するためにベンダーが提供する PMDK (Persistent Memory Developers Kit) [19] などのライブラリを利用することになる．

アプリケーション・プログラムないしシステム・ソフトウェアからみた NVM (Non Volatile Memory) のプログラミングモデルはストレージの業界団体である SNIA [43] が NVM Programming Model Specification として定義し公開している．図 3 で示した．

これらの機能は Linux ないし Windows など複数の OS によってすでに実装されている．Linux の XFS や ext4 ファイルシステムなどは DAX モード (AppDirect モード) をサポートしていて，mmap() と同様の機能を持つライブラリコールを利用することによって，ファイルの内容を直接ユーザーメモリ空

アプリケーションから見たNVM

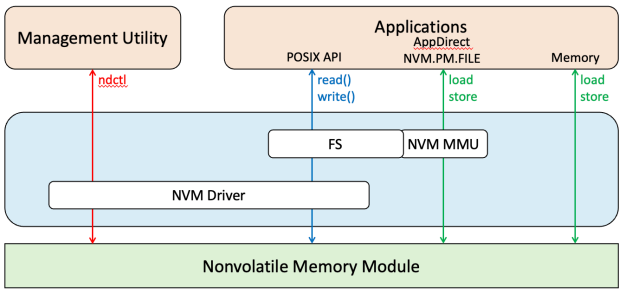


図 3　PMEM プログラミングモデル

間にマッピングすることができる．一度メモリ空間にマップされると OS の提供する I/O 機能，すなわち read()/write() システムコールなどを利用しなくても直接 NVM を利用することができる．NVM への読み書きは load/store で行う．Intel x86 アーキテクチャの場合，通常の MOV 命令になる．この場合，システムコールが必要ないので，コンテキストスイッチ，割り込みなどが発生しないため，低いレイテンシーでメモリにアクセスできる．

一方で，従来の揮発性メモリと違い，不揮発であるという特徴はプログラミングモデルの変更を余儀なくする．そのため，その動作特性の違いは単にスループットやレイテンシーだけでなく，アルゴリズムやデータ構造のあり方にも影響を与える．

不揮発メモリは従来のメモリ同様にバイト単位でアクセス可能であるが，CPU から見て不揮発メモリへのアクセスは通常キャッシュ経由でなされる．図 1 に示すようにキャッシュは揮発性なので永続性を必要とする場合は適宜キャッシュ内容を flush する必要がある．Intel x86 の場合，CLFLUSH 命令などによって明示的に行う．flush 関連の命令は CLFLUSH の他，CLWB，CLFLUSHOPT があり，キャッシュをバイパスする Non-temporal 命令もある．また，メモリへの書き込みを同期するために fence が必要になる．

以下に論文の構成をしめす．第 2 章でまず，空間索引構造を紹介し，それを不揮発メモリに適応するときの課題を整理する．そして，第 3 章で今回行った実装方式の検討についてふれ，その実験で明らかになった問題を議論する．第 4 章で関連研究を紹介し，最後にまとめと今後の課題を述べる．

2　空間索引構造

2.1　R-Tree

多次元データを高速に検索する方法として空間索引構造が知られている．その例としてここでは R-Tree を取り上げる (図 4 参照)． R-Tree [14] は Guttman によって提案された空間索引構造であり，次のような特徴を持つ．

- R-tree の leaf ノードは (I, tuple 識別子) という形式で，I は多次元のオブジェクト，tuple 識別子は当該オブジェクトを一意に識別する．
- leaf ノードでない場合 (non-leaf ノード) は，(I, 子ノードのポインタ) という形式である．
- 子ノードは leaf ノードか non-leaf ノードである．
- 各 leaf ノードはルートノードでなければ m から M 個の索引レコードを持つ．
- leaf ノードの各索引レコードは n 次元の最小矩形である．
- ルートノードは leaf ノードでないかぎり 2 個の子ノードを持つ．
- 全ての leaf は等しいレベルである（leaf ノードまでの高さは等しい）

図 4 では，ルートノードが R1 と R2 の子ノードを持ち，R1 は R3,R4,R5 の子ノードを持つ，R3 は同様に R8,R9,R10 を持ち，それらは leaf ノードになる．leaf ノードは多次元オブジェクトデータと tuple 識別子を持つ．B+tree と同様に，leaf ノードにデータを格納していくと，leaf までの高さを一定に保つために適宜分割などが発生する．

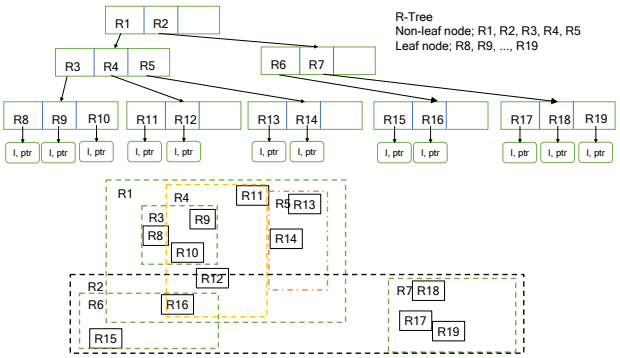


図 4　R-Tree の例

表 1　実 験 環 境

CPU model	Intel Xeon Silver, 2.5GHz
	8 core, 2 socket
No. of nodes	2
Cache	L1d 32KiB, L1i 32 KiB,
	L2 1 MiB, L3 11 MiB (shared)
DRAM	32 GiB * 12 (384 GiB)
DCPMM	128 GiB *12 (1536 GiB)
OS	CentOS 7.7.1908, linux kernel 3.10
PMDK	1.8

R-Tree そのものにはトランザクションの概念を含まないので同時実行制御等の機能は定義されていないが，データベースなどで利用する場合には検討する必要がある．

R-Tree そのものについては, [34] という参考書があり，Guttaman の提案以降の主要な提案が網羅されている．ただし出版年が 2006 年なので，近年のハードウェア動向については考慮されていない．

2.2　不揮発メモリのプログラミングモデル

不揮発メモリをストレージデバイスとして利用するときにはいくつかの注意すべき点がある．従来のプログラミングで永続性を確保したい場合は通常ファイルシステムないしデータベースなどを利用しておこなっているが，不揮発メモリの場合，ファイルシステムなどの利用は不必要となるが下記のようなプログラミングモデルとなるので従来型プログラムを変更する必要がある．

2.2.1　明示的な情報の永続化

不揮発メモリは情報をメモリにストアすると永続化されるが，そのタイミングは一般には非同期であり，通常はキャッシュに書き込まれるだけでメディアへの書き込みタイミングはプログラマは意識しない (図 1 参照)．また書き込む順も不定である．そのため，明示的にキャッシュからフラッシュし，書き込み順を同期する必要がある．書き込み順を指定しないと，一貫性を保てない状態が発生する場合がある．例えば，あるページへのポインタがあった場合，あるページへの書き込みが終了する前

に，そのページへのポインタを設定した場合，まだ情報が書き込まれていない場所への参照が発生してしまい，そのタイミングでシステムがクラッシュすると一貫性が保持されないままになる．最近のプロセッサは out of order で実行するために，プログラムを書いた順番に実行されるとは限らないので注意が必要である．Intel Optane DC Persistent Memory Module (以下 Intel DCPMM と称する) の場合，命令ごとの同期をとるために，明示的な FENCE 命令が必要になる．

2.2.2 アトミック書き込みは最大 8 バイト

Intel DCPMM がアトミックに書き込めるのは最大 8 バイトである．それ以上大きな塊での書き込みはアトミックに行われない．そのため書き込み途中でシステムがクラッシュすると一貫性が保持されない．R-Tree の木構造の変更はアトミックに行う必要がある．

上記のような問題に対処するため，不揮発メモリを対象とした B+Tree や Hash などで様々な提案がなされている．(主なサーベイとして B+木索引では [16], [18], [29] ，ハッシュ索引 [17] などがある).

空間索引については, [11] (以下 FBR と称する) および [26] (以下 PMR と称する) などが上記の問題に対応している．

多次元空間索引はデータベース分野のみならずコンピュータグラフィックス，位置情報システム，高性能システム (High Performance Computing) など様々な分野での応用が期待される．科学技術計算分野でのデータ量の爆発は不揮発メモリの応用分野としても期待されている．

2.2.3 クラッシュコンシステンシの実装

前述したように不揮発メモリにおいては単なるメモリへの転送だけでは一貫した状態を保持できない．なんらかの方法でクラッシュコンシステンシを確保する必要がある．

R-Tree の場合，B+Tree と同様にページへのレコードの追加時にページ分割が発生する場合があり，それは leaf ノードから root ノードまで波及する場合がある．木構造を変更するためにアトミックに変更する必要があり，その一貫性制御が必要となる．

先行研究 FBR [11] ではクラッシュコンシステンシを確保するために，mutex lock を利用した実装を評価した．一方，PMR [26] では mutex lock ではなく不揮発メモリ向け lock free アルゴリズム (Persistent Multi-word Compare and Swap 以下 PMwCAS と称する) [45] を利用した実装を評価した．また複数バイトへのアトミックな CAS (Multi-word Compare and Swap 以下 MwCAS と称する) [15] を利用した．

FBR は木構造の変更は木構造の root ノードに対するジャイアントロックを mutex lock を排他的に取得することで行っている．PMR はそれを PMwCAS と MwCAS で実装した．

3 実　　験

3.1 ハードウェア，ソフトウェア構成

次のような環境で実験した (表 1)．CPU は Intel Xeon Silver, 2.5Ghz, 8 core, 2 socket, Numa 2 ノード，OS は CentOS 7.7 Linux Kernel 3.10, PMDK 1.8 を利用した．

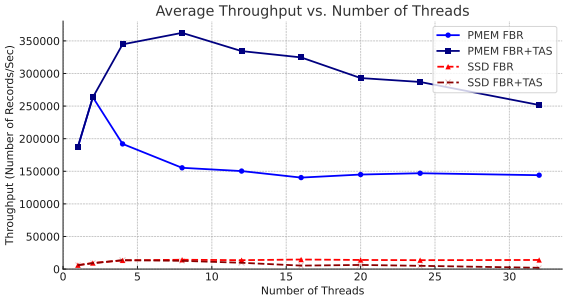


図 5　FBR，FBR+TAS，50 万件データの挿入，スレッド数を変化させる．縦軸は秒あたりの挿入件数 PMEM と SSD の差

3.2 実験方法と結果

FBR をベースラインとして，ランダムに生成した 50 万件の空間データを R-Tree に挿入する実験を行った．ソースコードは以下を利用した．`https://github.com/DICL/FBR-tree`

結果を図 5 に示した．データベースは PMDK を利用して不揮発メモリ上に生成して APP ダイレクトモードでメモリ上にマッピングしている．ランダムに生成した 50 万件の空間データをデータベースに挿入するスレッド数を 1, 2, 4, 8, 12, 16, 20, 24, 32 と変化させ，そのスループットを計測した．横軸にスレッド数，縦軸に秒あたりの挿入件数をとった．

まず，PMEM 領域にデータファイルを置きそこに挿入する時間を測定した．ベースラインの FBR アルゴリズム (図 5 の PMEM FBR の線，以下同様) は 1 スレッドのとき PMEM へ 1 秒あたり 19 万レコード弱挿入し，2 スレッドで 1 秒あたり 26 万レコード弱挿入しピークになる．その後スレッド数を増やしても 1 秒あたり 14 万レコード前後挿入する．次に SSD への挿入は (SSD FBR) で示される通り 1 スレッドのとき 6 千件/秒程度である．PMEM のほうが 30 倍以上速い．

FBR の場合，スレッド数を 2 より増やしても挿入件数は増えず，スケーラビリティがないことを確認した．

FBR の実装は mutex-lock を利用して R-Tree のノードの挿入の同期を取っている．R-Tree に対するジャイアントロックのため，レコードを挿入するとき，待ちが発生し，スケールしないと考えられる．そこで実験的に木全体を mutex-lock を利用してロックするのではなく Node に対する Test And Set (TAS) を利用して同期を取る方法を実装してみた (図 5PMEM FBR+TAS)．それによると 8 スレッドまで秒あたりの挿入件数が増えその後は徐々に減っていった．(188K 件/秒から 362K 件/秒)

FBR による検索の結果を図 6 に示した．予めレコードを 10 万件挿入したデータベースを用意して，それを PMDK の pmemobj_open() API を利用してオープンしメモリにマッピングして，ランダムに 1 万件 (図 6 10K)，2 万件 (図 6 20K)，4 万件 (図 6 40K) 検索した，1 秒あたりの検索件数を計測した．検索スレッド数 32 まで増やした．スレッド数に比例して秒あたりの検索件数が増加した．

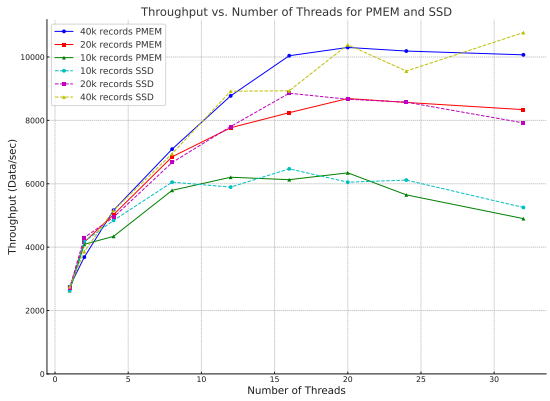


図 6　FBR 検索，PMEM と SSD の比較，横軸はスレッド数，縦軸は秒あたりの検索件数，データファイルへ 50 万件挿入後，1 万件 (10K)，2 万件 (20K)，4 万件 (40K) 検索

すでに存在するデータファイルを memory mapping するのでどちらも memory 上へのアクセスになる．したがって検索のスループットは SSD 上のファイルの場合も PMEM 上のファイルの場合もほぼ同様のスループットになる．PMEM の場合 20 スレッドあたりでピークになった．SSD も同様の傾向ではあるが 4 万件の検索の場合 32 スレッドがピークになった．

FBR のスケーラビリティの問題は挿入時に R-Tree のルートノードに対する mutex lock で木全体をロックしてしまうことにあると考えられる．さらにスケーラビリティを向上させるためにはなんらかの方法でボトルネックを解消する必要がある．

3.3　実験のまとめ

FBR について追試した．また，FBR の実装について mutex-lock のかわりに Test and Set (TAS) を利用した簡易的なロックフリーメカニズムを導入し評価したところ，スケーラビリティの向上がみられた．

PMEM のかわりに SSD に R-Tree をおいて評価したところ，挿入のスループットが 2.6 %〜10.9 %程度になった．一方で検索のスループットはデータファイルを PMEM においた場合と SSD においた場合の差異はほとんど生じなかった．これはデータファイルをメインメモリ上にメモリマッピングするため，一度マッピングされるとスループットはほぼ等しくなる．

4　関 連 研 究

Intel Opetane DCPMM そのものの性能評価は [22], [44] がある．多くの実装研究は実機がなかったためシミュレーションによって評価をしていた [1]．近年，Intel Optane DCPMM の出荷に伴い，徐々にその性能特性について評価が発表されてきている．汎用的なベンチマークのフレームワークは [4] にみられる．第 2 世代 Intel Optane(特に ADR(Asynchronous DRAM Refresh) および eADR (Extended ADR)) の評価を行っている．ADR および eADR の概要は [20] にある．

先行研究の提案を実機で検証しシミュレーション結果との差分を報告している [50]．Intel Optane DCPMM のレイテンシとスループットだけではなく，電力消費量の評価しているのは [37] である．Intel の NVM 向けライブラリ（NVML）についての初期の評価は [42] にある．当該ライブラリは現在，pmdk (persistent memory development kit) [19] になった．

既存のデータ構造を NVM に拡張し評価したものとして [27] などがある．B^+Tree 及びその NVM 拡張系の提案として，Persistent B^+-Trees [7]，FPTree [36]，BzTree [1]，また索引データ構造の比較検討 [21], [29] は従来提案されていたものを DCPMM で検証し，その有効性などを定量的に確認している．NVM 向けのファイルシステム NOVA [49]，Hikv(KVS) [48]，DBMS の実装 [2]，Write-behind logging [3] などがある．データベースへの適用について [25] は異なる database engine(PostgreSQL, MySQL，SQLServer，DuckDB，VoltDB，RockDB) について様々なワークロード (OLAP， OLTP， YCSB)、構成 (SSD, AppDir， Memory mode) で評価している．

High Performance Computing 分野での評価は [46] があり大規模アプリケーションでの性能を報告している．

不揮発メモリ向けの範囲索引について [29], [16], [17] などが様々な提案の比較を行った．ハッシュ索引 [17]，ファイルシステム [38] ，基本的な動作特性 [46], [37]，などの報告がある．不揮発メモリ向けの範囲索引について [29] は BzTree [1]，FPTree [36]，NV-Tree [51], [52]， wBTree [6]， について評価している．[16] は, [29] で取り上げられなかった以下の範囲索引 LB+-Tree [31]，uTree [8]，DPTree [55]，ROART [33]，PACTree [24] をそれぞれ評価した．主に Intel Optane DC Persistent Memory Module 出荷後に提案されたものを実機で評価している．

同様に [17] はハッシュ索引を比較し，Level hashing [56]，Clevel hashing [9]，CCEH [35]，Dash [32]，PCLHT [27]，SOFT [57] を評価した．[10] は様々なハッシュ索引を定量的に評価した．ファイルシステム [38] ，データベースエンジン [54]，不揮発メモリ向けのデータ構造 [13]，それぞれのカテゴリで比較検討している．Intel Optane DCPMM の性能評価は [37], [46] にある．[30] は不揮発メモリを大規模メインメモリとして利用する様々なアプリケーションについて調査している．

NUMA への PMEM の適用は [23] にある．

不揮発メモリ向け R-Tree の研究は近年徐々に増えてきている．[40] は FBR のスケーラビリティの欠如の問題について検討し，many core マシンを対象に，MPR-Tree を提案した．NUMA をサポートし更新のスケーラビリティを向上させた．

SSD(Flash storage) 対応の R-tree の研究として [47] と [12] がある．前者は SSD 向けの最適化について，後者は不揮発メモリと SSD のハイブリッドな構成での最適化について検討している．空間索引構造についてはまだ十分に検討がされていない．そこで本研究では，不揮発メモリを対象とする空間検索構造の実装方式を検討し実験をおこなった．

PMwCAS [45] は PMEM 向けの multi-word CAS (compare-and-swap) である。B+Tree の兄弟ノードにリンクを張った B-link 木は [28] で提案されている。latch-free index (OLFIT) は [5] で提案されている。

5 まとめと今後の課題

不揮発メモリを対象とする空間索引構造の実装方式の検討を行った．まず空間索引構造を PMEM に置く場合と SSD に置く場合の性能の比較を行った．

FBR の実装を確認し，マルチスレッド環境での挿入および検索のスケーラビリティを確認した．その結果，挿入は R-Tree へのコンテンションが発生し，スループットはスケールしないことが確認された．簡易的な Test and Set (TAS) を利用したロックフリーの実装を FBR に適応したところ，マルチスレッド環境で 8 スレッド程度までスループットが向上した．

次に SSD の挿入のスループットを確認し，PMEM と比較して 3.2 %程度のスループットだった．検索に関してはメモリマッピングをすることによって PMEM 上にデータファイルを置く場合も，SSD 上にデータファイルを置く場合もほぼ同様のスループットが得られた．

今後は PMR の未実装の部分 (mutex lock ではなく PMwCAS によるロックフリー同期 [45]) についても追試をし不揮発メモリを対象とする空間索引構造の拡張の提案および実装を行いたい. [26]

文　献

[1] Joy Arulraj, Justin Levandoski, Umar Farooq Minhas, and Per-Ake Larson. Bztree: A high-performance latch-free range index for non-volatile memory. *Proceedings of the VLDB Endowment*, Vol. 11, No. 5, pp. 553–565, 2018.

[2] Joy Arulraj and Andrew Pavlo. How to build a non-volatile memory database management system. In *Proc' of the 2017 ACM SIGMOD*, p. 1753–1758, 2017.

[3] Joy Arulraj, Matthew Perron, and Andrew Pavlo. Write-behind logging. *Proc' of the VLDB Endowment*, Vol. 10, No. 4, pp. 337–348, 2016.

[4] Lawrence Benson, Leon Papke, and Tilmann Rabl. PerMA-Bench: Benchmarking persistent memory access. *Proc. VLDB Endow.*, Vol. 15, No. 11, p. 2463–2476, sep 2022.

[5] Sang Kyun Cha, Sangyong Hwang, Kihong Kim, and Keunjoo Kwon. Cache-conscious concurrency control of main-memory indexes on shared-memory multiprocessor systems. In *VLDB*, Vol. 1, pp. 181–190, 2001.

[6] Shimin Chen and Qin Jin. Persistent b+-trees in non-volatile main memory. *Proceedings of the VLDB Endowment*, Vol. 8, No. 7, pp. 786–797, 2015.

[7] Shimin Chen and Qin Jin. Persistent b^+-trees in non-volatile main memory. *Proc. VLDB Endow.*, Vol. 8, No. 7, p. 786–797, February 2015.

[8] Youmin Chen, Youyou Lu, Kedong Fang, Qing Wang, and Jiwu Shu. utree: a persistent b+-tree with low tail latency. *Proceedings of the VLDB Endowment*, Vol. 13, No. 12, pp. 2634–2648, 2020.

[9] Zhangyu Chen, Yu Hua, Bo Ding, and Pengfei Zuo. Lock-free concurrent level hashing for persistent memory. In *Proceedings of the 2020 USENIX Conference on Usenix Annual Technical Conference*, pp. 799–812, 2020.

[10] Zhiwen Chen, Daokun Hu, Wenkui Che, Jianhua Sun, and Hao Chen. A quantitative evaluation of persistent memory hash indexes. *The VLDB Journal*, pp. 1–23, 2023.

[11] Soojeong Cho, Wonbae Kim, Sehyeon Oh, Changdae Kim, Kwangwon Koh, and Beomseok Nam. Failure-atomic byte-addressable r-tree for persistent memory. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 32, No. 3, pp. 601–614, 2020.

[12] Athanasios Fevgas, Leonidas Akritidis, Miltiadis Alamaniotis, Panagiota Tsompanopoulou, and Panayiotis Bozanis. A study of r-tree performance in hybrid flash/3dxpoint storage. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–6. IEEE, 2019.

[13] Philipp Götze, Arun Kumar Tharanatha, and Kai-Uwe Sattler. Data structure primitives on persistent memory: an evaluation. In *Proceedings of the 16th International Workshop on Data Management on New Hardware*, pp. 1–3, 2020.

[14] Antonin Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pp. 47–57, 1984.

[15] Timothy L Harris, Keir Fraser, and Ian A Pratt. A practical multi-word compare-and-swap operation. In *International Symposium on Distributed Computing*, pp. 265–279. Springer, 2002.

[16] Yuliang He, Duo Lu, Kaisong Huang, and Tianzheng Wang. Evaluating persistent memory range indexes: Part two. *arXiv preprint arXiv:2201.13047*, 2022.

[17] Daokun Hu, Zhiwen Chen, Jianbing Wu, Jianhua Sun, and Hao Chen. Persistent memory hash indexes: An experimental evaluation. *Proc. VLDB Endow.*, Vol. 14, No. 5, p. 785–798, mar 2021.

[18] Kaisong Huang, Yuliang He, and Tianzheng Wang. The past, present and future of indexing on persistent memory. *Proc. VLDB Endow.*, Vol. 15, No. 12, p. 3774–3777, sep 2022.

[19] Intel. Persistent Memory Development Kit. https://pmem.io/pmdk/, 2019.

[20] Intel. eADR: New Opportunities for Persistent Memory Applications. https://www.intel.com/content/www/us/en/developer/articles/technical/eadr-new-opportunities-for-persistent-memory-applications.html, 2021.

[21] Abdullah Al Raqibul Islam, Anirudh Narayanan, Christopher York, and Dong Dai. A performance study of optane persistent memory: From indexing data structures ' perspective. In *36th (MSST 2020)*, pp. 2–2, 2020.

[22] Joseph Izraelevitz, Jian Yang, Lu Zhang, Juno Kim, Xiao Liu, Amirsaman Memaripour, Yun Joon Soh, Zixuan Wang, Yi Xu, et al. Basic performance measurements of the intel optane DC persistent memory module. *arXiv:1903.05714*, pp. 1–61, 2019.

[23] Safdar Jamil, Abdul Salam, Awais Khan, Bernd Burgstaller, Sung-Soon Park, and Youngjae Kim. Scalable numa-aware persistent b+-tree for non-volatile memory devices. *Cluster Computing*, Vol. 26, No. 5, pp. 2865–2881, 2023.

[24] Wook-Hee Kim, R Madhava Krishnan, Xinwei Fu, Sanidhya Kashyap, and Changwoo Min. Pactree: A high performance persistent range index using pac guidelines. In *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, pp. 424–439, 2021.

[25] Dimitrios Koutsoukos, Raghav Bhartia, Michal Friedman, Ana Klimovic, and Gustavo Alonso. Nvm: Is it not very meaningful for databases? *Proceedings of the VLDB Endowment*, Vol. 16, No. 10, pp. 2444–2457, 2023.

[26] Brandon Lavinsky and Xuechen Zhang. Pm-rtree: A highly-efficient crash-consistent r-tree for persistent memory. In *Proceedings of the 34th International Conference on Scientific and Statistical Database Management*, pp. 1–11, 2022.

[27] Se Kwon Lee, Jayashree Mohan, Sanidhya Kashyap, Taesoo

Kim, and Vijay Chidambaram. Recipe: converting concurrent DRAM indexes to persistent-memory indexes. In *Proc' of 27th ACM SOSP*, pp. 462–477, 2019.

[28] Philip L. Lehman and s. Bing Yao. Efficient locking for concurrent operations on b-trees. *ACM Trans. Database Syst.*, Vol. 6, No. 4, p. 650–670, dec 1981.

[29] Lucas Lersch, Xiangpeng Hao, Ismail Oukid, Tianzheng Wang, and Thomas Willhalm. Evaluating persistent memory range indexes. *Proc' of the VLDB Endowment*, Vol. 13, No. 4, pp. 574–587, 2019.

[30] Hai-Kun Liu, Di Chen, Hai Jin, Xiao-Fei Liao, Binsheng He, Kan Hu, and Yu Zhang. A survey of non-volatile main memory technologies: State-of-the-arts, practices, and future directions. *Journal of Computer Science and Technology*, Vol. 36, No. 1, pp. 4–32, 2021.

[31] Jihang Liu, Shimin Chen, and Lujun Wang. Lb+ trees: optimizing persistent index performance on 3dxpoint memory. *Proceedings of the VLDB Endowment*, Vol. 13, No. 7, pp. 1078–1090, 2020.

[32] Baotong Lu, Xiangpeng Hao, Tianzheng Wang, and Eric Lo. Dash: Scalable hashing on persistent memory. *arXiv preprint arXiv:2003.07302*, 2020.

[33] Shaonan Ma, Kang Chen, Shimin Chen, Mengxing Liu, Jianglang Zhu, Hongbo Kang, and Yongwei Wu. Roart: Range-query optimized persistent art. In *FAST*, pp. 1–16, 2021.

[34] Yannis Manolopoulos, Apostolos N Papadopoulos, Apostolos N Papadopoulos, and Yannis Theodoridis. *R-Trees: Theory and Applications: Theory and Applications*. Springer Science & Business Media, 2006.

[35] Moohyeon Nam, Hokeun Cha, Young-ri Choi, Sam H Noh, and Beomseok Nam. Write-optimized dynamic hashing for persistent memory. In *FAST*, Vol. 19, pp. 31–44, 2019.

[36] Ismail Oukid, Johan Lasperas, Anisoara Nica, Thomas Willhalm, and Wolfgang Lehner. FPTree: A hybrid scm-dram persistent and concurrent b-tree for storage class memory. In *Proc' of the 2016 SIGMOD*, pp. 371–386, 2016.

[37] Ivy B. Peng, Maya B. Gokhale, and Eric W. Green. System evaluation of the intel optane byte-addressable nvm. In *Proceedings of the International Symposium on Memory Systems*, MEMSYS '19, p. 304–315, New York, NY, USA, 2019. Association for Computing Machinery.

[38] Gianlucca O. Puglia, Avelino Francisco Zorzo, César A. F. De Rose, Taciano Perez, and Dejan Milojicic. Non-volatile memory file systems: A survey. *IEEE Access*, Vol. 7, pp. 25836–25871, 2019.

[39] Andy Rudoff. Persistent memory programming. *Login: The Usenix Magazine*, Vol. 42, No. 2, pp. 34–40, 2017.

[40] Abdul Salam, Safdar Jamil, Sungwon Jung, Sung-Soon Park, and Youngjae Kim. Future-based persistent spatial data structure for nvm-based manycore machines. *IEEE Access*, Vol. 10, pp. 114711–114724, 2022.

[41] Steve Scargall. *Programming Persistent Memory A Comprehensive Guide for Developers*. Apress, Berkeley, CA, 2020.
https://doi.org/10.1007/978-1-4842-4932-1.

[42] H. Shu, H. Chen, H. Liu, Y. Lu, Q. Hu, and J. Shu. Empirical study of transactional management for persistent memory. In *2018 IEEE 7th Non-Volatile Memory Systems and Applications Symposium (NVMSA)*, pp. 61–66, 2018.

[43] SNIA. SNIA Storage Networking Industry Association. https://www.snia.org, 2017.

[44] Alexander van Renen, Lukas Vogel, Viktor Leis, Thomas Neumann, and Alfons Kemper. Persistent memory I/O primitives. In *Proc' of DaMoN'19*, pp. 1–7. ACM, 2019.

[45] Tianzheng Wang, Justin Levandoski, and Per-Ake Larson. Easy lock-free indexing in non-volatile memory. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 461–472. IEEE, 2018.

[46] Michèle Weiland, Holger Brunst, Tiago Quintino, Nick Johnson, Olivier Iffrig, Simon Smart, Christian Herold, Antonino Bonanni, Adrian Jackson, and Mark Parsons. An early evaluation of Intel's optane DC persistent memory module and its impact on high-performance scientific applications. In *Proc' of SC '19*, pp. 1–19, 2019.

[47] Chin-Hsien Wu, Li-Pin Chang, and Tei-Wei Kuo. An efficient r-tree implementation over flash-memory storage systems. In *Proceedings of the 11th ACM international symposium on Advances in geographic information systems*, pp. 17–24, 2003.

[48] Fei Xia, Dejun Jiang, Jin Xiong, and Ninghui Sun. Hikv: A hybrid index key-value store for dram-nvm memory systems. In *USENIX/ATC 17*, pp. 349–362, 2017.

[49] Jian Xu and Steven Swanson. NOVA: A log-structured file system for hybrid volatile/non-volatile main memories. In *14th USENIX/FAST*, pp. 323–338, 2016.

[50] Jian Yang, Juno Kim, Morteza Hoseinzadeh, Joseph Izraelevitz, and Steve Swanson. An empirical guide to the behavior and use of scalable persistent memory. In *18th USENIX/FAST 20*, pp. 169–182, 2020.

[51] Jun Yang, Qingsong Wei, Cheng Chen, Chundong Wang, Khai Leong Yong, and Bingsheng He. NV-Tree: Reducing consistency cost for NVM-based single level systems. In *13th USENIX Conference on File and Storage Technologies (FAST 15)*, pp. 167–181, Santa Clara, CA, February 2015. USENIX Association.

[52] Jun Yang, Qingsong Wei, Chundong Wang, Cheng Chen, Khai Leong Yong, and Bingsheng He. Nv-tree: A consistent and workload-adaptive tree structure for non-volatile memory. *IEEE Transactions on Computers*, Vol. 65, No. 7, pp. 2169–2183, 2016.

[53] Hirotaka Yoshioka, Yuto Hayamizu, Kazuo Goda, and Masaru Kitsuregawa. pmmeter: A microbenchmark for understanding synchronization cost on persistent memory. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp2023)*, pp. 326–327. IEEE, 2023.

[54] Hao Zhang, Gang Chen, Beng Chin Ooi, Kian-Lee Tan, and Meihui Zhang. In-memory big data management and processing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 7, pp. 1920–1948, 2015.

[55] Xinjing Zhou, Lidan Shou, Ke Chen, Wei Hu, and Gang Chen. Dptree: differential indexing for persistent memory. *Proceedings of the VLDB Endowment*, Vol. 13, No. 4, pp. 421–434, 2019.

[56] Pengfei Zuo, Yu Hua, and Jie Wu. Write-optimized and high-performance hashing index scheme for persistent memory. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pp. 461–476, 2018.

[57] Yoav Zuriel, Michal Friedman, Gali Sheffi, Nachshon Cohen, and Erez Petrank. Efficient lock-free durable sets. *Proceedings of the ACM on Programming Languages*, Vol. 3, No. OOPSLA, pp. 1–26, 2019.

DEIM Forum 2024 XX-Y

モバイル環境における圧縮処理を用いたIoTシステムの通信性能向上のための検討

伊藤　千紗† 　竹房あつ子†† 　中田　秀基††† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2–1–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 産業技術総合研究所　〒 305–8560 茨城県つくば市梅園 1–1–1
E-mail: †{chisa,oguchi}@ogl.is.ocha.ac.jp, ††takefusa@nii.ac.jp, †††hide-nakada@aist.go.jp

あらまし　IoT 機器から収集された様々なセンサデータをクラウドで収集, 解析し, 活用することが期待されている. しかし, モバイル環境にある IoT データの収集では, 各種サービスで要求される通信スループットや通信遅延を維持することができるかが課題となる. IoT 通信では小規模データが大量に送信されるため, 通信環境に合わせた転送方法を検討する必要がある. 本研究では, IoT 用通信ライブラリを提供する SINETStream を用いて, IoT データ通信の性能評価を行う. 既発表研究で通信スループットを高めるにはマルチスレッド通信やデータ圧縮が有効であることを示した. 本稿ではモバイル環境で, 送受信データ特性や圧縮率に焦点を当てた圧縮処理の有効性を調査し, 性能向上のための検討を行う. 実験ではデータサイズ, 圧縮アルゴリズム, データ特性, 圧縮率の条件を変え, 比較を行う. その結果, 高圧縮可能なデータであるほど性能が高くデータサイズが大きいほど性能差が顕著であること, 圧縮時間とデータの圧縮率に相関はないこと, 圧縮時間が全体の実行時間に与える影響はわずかであることが分かった. また, 適切なアルゴリズムやデータ特性の選択や, データサイズに応じた圧縮処理の適用が性能向上に効果的であることが確認できた.

キーワード　モバイル, IoT, 性能評価, データ圧縮

1 はじめに

様々なモノに通信デバイスを組み込み, インターネットに接続しようとする試みを IoT (Internet of Things) と呼ぶ. これらの IoT 機器を利用してさまざまなセンサデータを収集することができ, これらのセンサデータをクラウドで蓄積，解析し，活用することが期待されている. しかし，モバイル環境にある IoT データの収集では，自動運転車や倉庫内自走ロボットなどの各種サービスで要求される通信スループットや通信遅延を維持する必要があり, 性能向上が課題となっている. 高性能なモバイル通信技術として，5G の活用が期待されている. 5G は第 5 世代移動通信システムの略で LTE で提供してきた高速・大容量をさらに進化させた通信規格である. 最高伝送速度は 10Gbps と言われており, LTE の帯域幅よりもさらに広い帯域幅である. 高速大容量・超低遅延・多数同時接続の特徴を持ち, IoT 機器との接続など社会発展の基盤としても大きく期待されている. しかし, 小規模データが大量に送信されるような, IoT 通信の 5G 環境における性能特性が明らかではない. そこで我々は国立情報学研究所内に設置された 5G SA 環境で，IoT 用通信ライブラリを提供する SINETStream を用いて基本性能の調査を行った [1]. 5G SA とは，4G 用のコア装置を流用する NSA (Non Stand Alone) に対して，5G 専用のコア装置と基地局装置を使用するものである [2]. 実験では暗号化の有無, データサイズ, スレッド数の条件を変え, 5G と LTE のそれぞれで比較を行った. また, LTE 環境でデータ圧縮の有無での比較を行った. 調査の結果, モバイル通信における 5G のスループットが従来の LTE よりも高くなること, 通信スループットを高めるにはマルチスレッド通信やデータ圧縮が有効であることを確認した. 一方で, IoT 通信のように小規模のメッセージを大量にクラウドに送信するような通信パターンではアップロード帯域で律速され, ダウンロード帯域の方が広い 5G モバイル通信を有効利用するのが難しいことも示唆された. そこで, より通信環境に合わせた転送方法を検討する必要がある.

本稿ではモバイル環境で, 送受信データ特性と圧縮率に焦点を当てたデータ圧縮の有効性を調査する. 実験ではデータサイズ, 圧縮アルゴリズム, データ特性, 圧縮率の条件を変え, LTE 環境で通信性能を測定し, 比較する. その結果, 高圧縮可能なデータであるほど性能が高くデータサイズが大きいほど性能差が顕著であること, 圧縮時間とデータの圧縮率に相関はないこと, 圧縮時間が全体の実行時間に与える影響はわずかであること, 適切なアルゴリズムやデータ特性の選択や, データサイズに応じた圧縮処理の適用が性能向上に効果的であることがわかった.

2 関連技術

2.1 SINETStream

SINETStream [3] [4] とは, 国立情報学研究所で開発されたデータを欠損なく確実に収集・解析するための機能を提供している IoT システム開発支援ライブラリである. SINETStream は, IoT 通信に適した Publisher/Subscriber 型の通信モデル

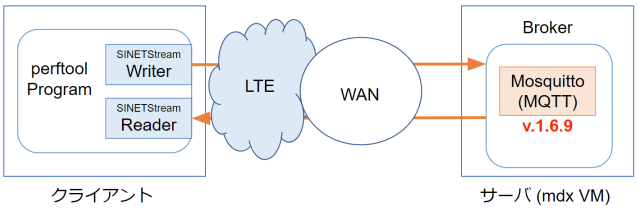


図 1: 実験の構成

を採用している．MQTT(Message Queuing Telemetry Transport) [5] 準拠のブローカや Apache Kafka [6] のような標準的なメッセージブローカに対応した共通 API を提供しており，具体的には，(1) センサ等から収集されるデータをクラウドや大学などに設置されたサーバへ書き込む (2) サーバに収集されたデータを解析プログラムに読み込むといった機能を提供している．また，通信やデータの暗号化，IoT 機器の認証を行う機能も含んでいるため，安全な広域データ収集を実現できる．本研究では SINETStream を用いた MQTT 通信の性能調査を行う．

2.2 MQTT

MQTT は IoT 通信で利用される主要なプロトコルとして知られている [7] [8]．MQTT は IoT 向けの TCP/IP プロトコル上に構築された Publisher/Subscriber 型のモデルに基づく軽量な通信プロトコルである．わずかなコードと限られた帯域幅で，接続されたリモートデバイスにリアルタイムの信頼性の高いメッセージングサービスを提供できることが MQTT の利点である．本研究ではブローカとして MQTT ブローカの Eclipse Mosquitto [9] を用いる．

3 実　　験

通信スループットを高めるのにより効果的なデータ特性や圧縮手法を調査するため，以下の実験 1, 2 を行う．実験 1 では送受信データ特性によって通信スループットが変わるかどうか調査する．実験 2 ではデータの圧縮率と通信スループットとの相関と，圧縮にかかる時間の割合を調査する．

実験では LTE 環境で IoT 通信の往復通信時間を測定する．LTE はモバイル SINET [10] の Softbank の LTE 環境を用いる．実験の構成を図 1 に示す．SINETStream では Publisher は Writer，Subscriber は Reader と呼ぶ．ブローカのソフトウェアは Mosquitto v. 1.6.9 を用いる．また，実験環境を表 1 に示す．クライアントは Raspberry Pi，サーバはデータ活用社会創成プラットフォーム mdx [11] の VM を用いる．クライアント端末から柏にある mdx VM の MQTT ブローカに対して通信する．モバイルルータとクライアント端末は USB で接続する．また，モバイルルータは Aterm MR51FN を用いる．

データ圧縮アルゴリズムには gzip [12] と zstd [13] を用いる．zstd は gzip とほぼ同じ圧縮率だが，より高速に圧縮と解凍が行えるように設計されている圧縮アルゴリズムである．測定には perftool [14] を使用する．perftool とは SINETStream を介しての MQTT ブローカとのデータの送受信を行い，そのメトリクス情報を出力するプログラムである．クライアント端末で Publisher と Subscriber のプログラムを実行し，サーバで起動しているブローカにメッセージを送信し，ブローカで収集されたメッセージをクライアント端末の Subscriber で取得するまでの時間を測定する．perftool には Java と Python の実装があり，本実験では Java 版を使用する．本実験では，指定したデータサイズのデータを Publisher が 100 個送信し，Subscriber が受信するまでを 10 セット行うことを，perftool 実行 1 回分と定義する．そして perftool 実行 1 回分の実行時間を測定し，累積送信データ量を実行時間で割ったものをスループットとする．

表 1: 実験環境

クライアント：Raspberry Pi	
機種名	Raspberry Pi 4 Computer Model B 4GB RAM
OS	Raspbian GNU / Linux 11
CPU	ARMv7 Processor rev 3 (v7l), 4 コア, 2065 MHz
Main Memory	4GB
サーバ：mdx VM	
OS	Ubuntu 20.04.5 LTS
仮想 CPU コア数	16
メモリ	24.19GB
仮想ディスクサイズ	100GB

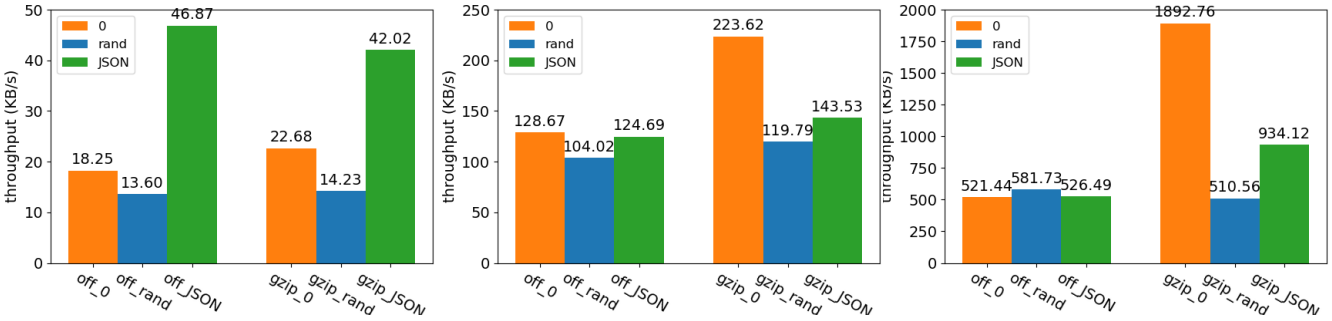
表 2: 実験 1 (送受信データ特性による比較) 条件

データサイズ	1 / 10 / 100 [KB]
データ圧縮	有 (gzip) / 無
送受信データ特性	0 埋め / 乱数 / JSON

3.1 実験 1 (送受信データ特性による比較)

LTE 環境で送受信データ特性を変えて比較を行う．実験条件を表 2 に示す．データサイズは 1KB, 10KB, 100KB の 3 通りであり，圧縮アルゴリズムは gzip を使用する．送受信データは，perftool ではデフォルトが乱数だが，本実験では乱数のデータ “rand”，圧縮が効きやすいように 0 で埋めたデータ “0”，JSON 形式で書かれた実際のセンサ情報を読み込んだデータ “JSON” の 3 通りで測定する．なお，JSON のデータサイズは厳密には 2.8KB, 11KB, 112KB だが，ここでは 1KB, 10KB, 100KB と近似して比較する．

通信スループットを図 2 に示す． 横軸の “off” は圧縮無，“gzip” は gzip アルゴリズムで圧縮を表しており，縦軸はスループットである．グラフから，乱数データにおいてはデータ圧縮による通信スループットの改善がみられないことがわかった．これは乱数データは圧縮率が悪く，データ圧縮自体に時間がかかってしまうため性能が低下していると考えられる．0 埋めデータでは，データ圧縮を行ったものは性能向上がみられた．また，データサイズが大きくなるほど性能の差が大きくなった．JSON データでは，データサイズが 1KB のときは 0 埋めデータよりも性能が良く，データサイズが大きくなると性能は良くなるものの 0 埋めデータほどではないという結果になった．



(a) 1 KB (b) 10 KB (c) 100 KB

図 2: データ特性の違いによる通信スループットの比較

表 3: データ特性の違いによる gzip の圧縮率の比較

データ特性		0 埋め	乱数	JSON
gzip	1KB	0.029	1.023	0.388
	10KB	0.0045	1.0023	0.241
	100KB	0.00132	1.00053	0.167

表 4: 実験 2 (データの圧縮率による比較) 条件

データサイズ	1 / 10 / 100 [KB]
データ圧縮	有 (gzip) / 有 (zstd) / 無
圧縮率目標値	0.1 / 0.2 / 0.3 / 0.4 / 0.5 / 0.6 / 0.7 / 0.8 / 0.9 /1.0

表 3 は圧縮率の結果を示したものであり, 0 埋めデータ, JSON データ, 乱数データの順に圧縮が効いていることが読み取れる. また、データサイズが大きいほど圧縮効率が高かったこともわかる. これらのことから圧縮率の高いデータほど性能改善効果が大きいと考えられる. 実験 1 の結果から, 送受信するデータサイズやデータ特性に応じた圧縮処理の適用が重要であることがわかった.

3.2 実験 2 (データの圧縮率による比較)

データの圧縮率と通信性能との相関をより詳細に調べるため, LTE 環境でデータの圧縮率を変えて比較を行う. 実験条件を表 4 に示す. データサイズは 1KB, 10KB, 100KB の 3 通りである. 圧縮アルゴリズムは gzip, zstd を用いる. また本実験では, 指定した圧縮率目標値に近くなるデータを生成するプログラムを作成し, 送受信データを生成する. データ生成アルゴリズムは, まず, 圧縮率を受け取ったらエントロピーを算出し, そのエントロピーになるような確率分布を生成する. そして確率分布を正規化し, 確率分布に従った, 指定したデータサイズの乱数データ配列を生成する. ここで, 圧縮率 0.1 とは圧縮によりデータサイズが 0.1 倍になることを目標としている.

まず, 開発した圧縮用データ生成プログラムを用いてデータを生成し, そのデータを gzip と zstd で圧縮した場合の圧縮率を表 5 に示す. 圧縮率の結果は 10 個のデータの平均値である. 表から gzip と zstd のどちらにおいても, 設定した圧縮率目標値と一致はしないものの, 近い値となっていることがわかる. また, データサイズが大きいほど目標値に近くなっていることがわかる. 次に, 生成した合成データを用いて圧縮通信をした際の通信スループットを図 3 に示す. 青色のグラフ “gzip” が gzip アルゴリズムで圧縮, 橙色のグラフ “zstd” が zstd アルゴリズムで圧縮, 緑色のグラフ “off” が圧縮無の結果である. 横軸の数字は圧縮率目標値を表している. 縦軸はスループットである. 結果から gzip と zstd 両方で, 高圧縮であるほど性能が良く, データサイズが大きいほど性能差が顕著であることが確認できた.

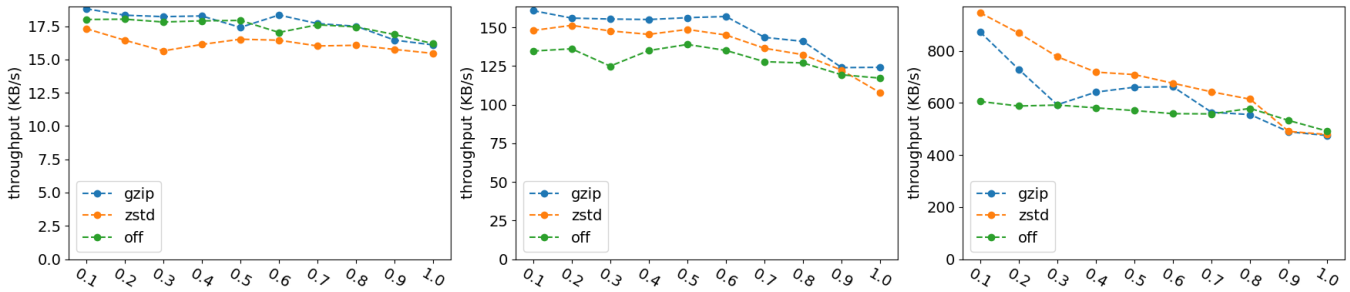
さらに, 圧縮解凍・転送時間を比較したものを図 4 に示す. ここで実行時間は, 送受信データを Publisher が 100 個送信し, Subscriber が受信するまでを 10 セット行った累積時間で比較する. 橙色が圧縮解凍時間で perftool 実行時間のうちデータ圧縮とデータ解凍にかかる時間のみを測定したもの, 青色が転送時間で全体の実行時間から圧縮時間を引いたものとした. また, 圧縮解凍時間のうち, 圧縮時間と解凍時間の内訳を図 5 に示す. 緑色がデータ圧縮時間, 赤色がデータ解凍時間を表している. 図 4, 図 5 から, gzip と zstd いずれもデータ圧縮率と圧縮・解凍時間にはあまり相関がみられないことがわかる. gzip は, 1KB, 10KB といった小さいデータサイズでは圧縮時間の割合はかなり小さく, データサイズが大きくなると圧縮時間の割合が大きくなり, 圧縮率 0.3 付近で圧縮時間が長くなることが読み取れる. ここで圧縮率 0.3 付近で圧縮時間が長くなる理由は調査中だが, 現時点では確率分布が一定の長さになると圧縮に時間がかかることがわかっている. zstd は, データサイズに関係なく圧縮時間はほぼ一定であることがわかる. これらのことから, データ圧縮を行う際に, データサイズが小さいときは gzip, 大きいときは zstd のようにアルゴリズムを適切に選択することが重要だと考えられる. また, 圧縮時間が全体の実行時間に与える影響は小さいこともわかった. 実験 2 から, 効率的に通信するにはデータの特性により適切なアルゴリズムを選択することが重要であること, 圧縮時間が全体の実行時間に与える影響はわずかであるため, 実行時間の観点においても圧縮処理が有効であることがわかった.

4 まとめと今後の課題

IoT 用通信ライブラリ SINETStream を用いてモバイル環境

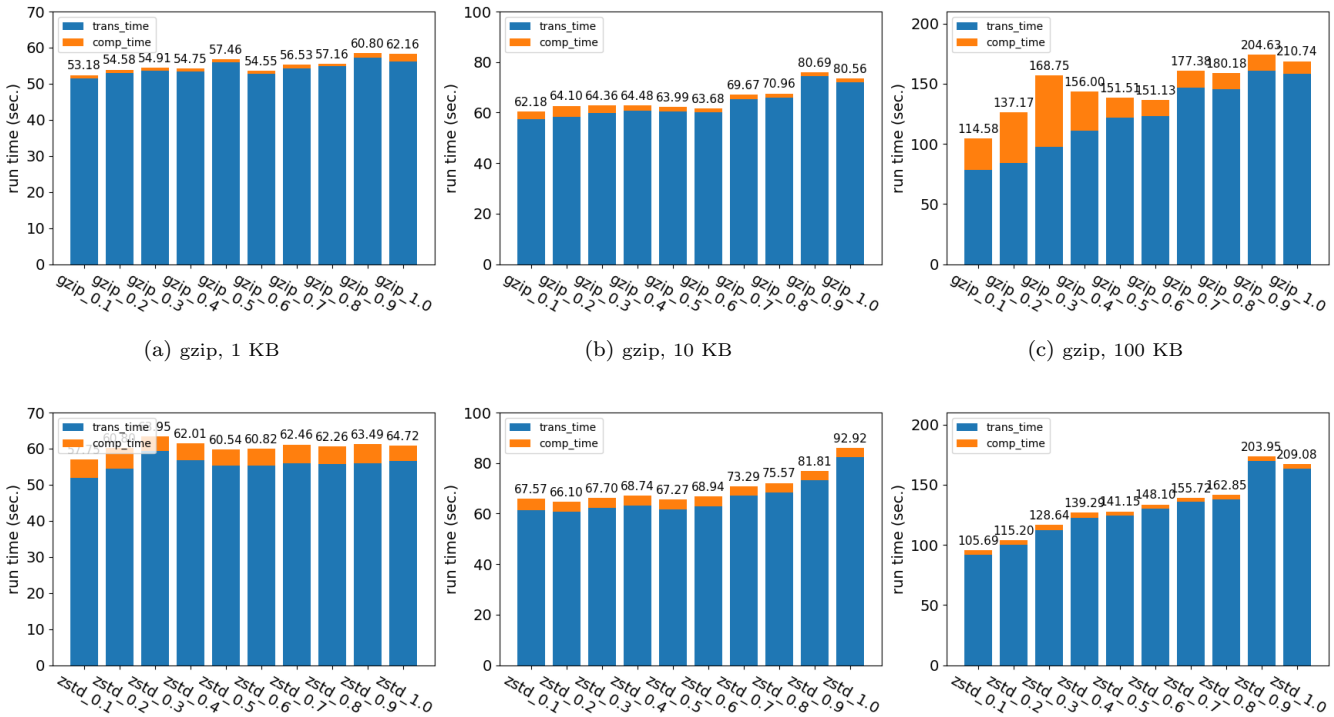
表 5: 圧縮用データ生成プログラムで生成したデータを圧縮した場合の圧縮率の比較

圧縮率目標値		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
gzip	1KB	0.191	0.315	0.425	0.509	0.569	0.650	0.747	0.847	0.953	1.023
	10KB	0.150	0.259	0.381	0.493	0.578	0.633	0.719	0.813	0.909	1.002
	100KB	0.144	0.245	0.358	0.468	0.575	0.642	0.718	0.812	0.907	1.001
zstd	1KB	0.245	0.312	0.393	0.447	0.532	0.632	0.735	0.836	0.943	1.009
	10KB	0.180	0.281	0.370	0.434	0.504	0.608	0.711	0.811	0.907	1.001
	100KB	0.177	0.265	0.375	0.457	0.504	0.607	0.710	0.810	1.000	1.000



(a) 1 KB (b) 10 KB (c) 100 KB

図 3: 圧縮率の違いによる通信スループットの比較

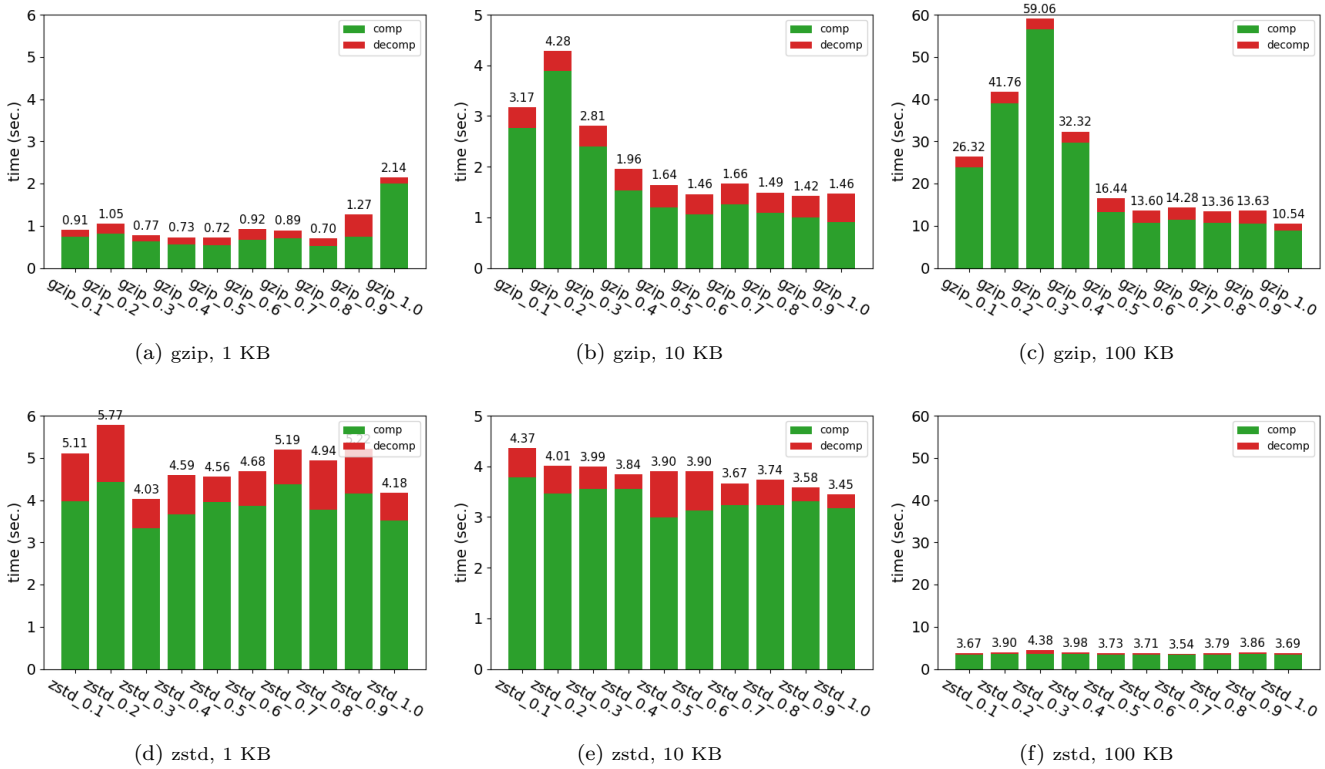


(a) gzip, 1 KB (b) gzip, 10 KB (c) gzip, 100 KB

(d) zstd, 1 KB (e) zstd, 10 KB (f) zstd, 100 KB

図 4: 圧縮率の違いによる圧縮解凍・転送時間の比較

の通信性能の測定を行い, 圧縮処理の有効性を調査した. 送受信データ特性を変えて通信性能の比較を行った実験 1 からは, 送受信するデータサイズやデータ特性に応じた圧縮処理の適用が性能向上に効果的であることが確認できた. 圧縮率目標値を変え, それに合わせた送受信データを用いて通信性能と圧縮解凍・転送時間の比較を行った実験 2 からは, 高圧縮可能なデータであるほど性能が高くデータサイズが大きいほど性能差が顕著であること, 圧縮時間とデータの圧縮率に相関はないこと, gzip と zstd でデータサイズの違いによる圧縮時間の特徴が異なること, 圧縮時間が全体の実行時間に与える影響はわずかである



(a) gzip, 1 KB (b) gzip, 10 KB (c) gzip, 100 KB

(d) zstd, 1 KB (e) zstd, 10 KB (f) zstd, 100 KB

図 5: 圧縮率の違いによる圧縮・解凍時間の比較

ことが確認できた．これらによって効率的に通信するにはデータの特性により適切なアルゴリズムを選択することが重要であることや，実行時間の観点においても圧縮処理が有効であることが示唆された．

今後は，効果的な圧縮処理の適用について引き続き調査を行っていく．また，通信環境に合わせて IoT 通信の転送効率を高める手法を検討する．具体的にはブローカのパラメータを変更し，複数個のメッセージをまとめて送受信したときの性能の調査を行う．

謝　　辞

本研究は JSPS 科研費 JP19H04089，および 2023 年度国立情報学研究所公募型共同研究 (2023-23S0205) の助成を受けたものです．また，本研究成果はデータ活用社会創成プラットフォーム mdx を利用して得られたものです．

文　　献

[1] 伊藤千紗，竹房あつ子，中田秀基，小口正人．5G 環境での IoT システムのためのモバイル通信性能の評価．マルチメディア，分散，協調とモバイルシンポジウム 2023 論文集，Vol. 2023，pp. 367–374，06 2023.
[2] docomo. 5G. https://www.docomo.ne.jp/area/5g/.
[3] SINETStream. https://sinetstream.net/.
[4] Atsuko Takafusa, Jingtao Sun, Ikki Fujiwara, Hiroshi Yoshida, Kento Aida, and Calton Pu. SINETStream: Enabling Research IoT Applications with Portability, Security and Performance Requirements. *In Proc. COMPSAC 2021*, pp. 482–492, 2021.
[5] MQTT (Message Queue Telemetry Transport). https://mqtt.org/.
[6] Apache Kafka. https://kafka.apache.org/.
[7] Eyhab Al-Masri, Karan Raj Kalyanam, John Batts, Jonathan Kim, Sharanjit Singh, Tammy Vo, and Charlotte Yan. Investigating messaging protocols for the internet of things (iot). *IEEE Access*, Vol. 8, pp. 94880–94911, 2020.
[8] https://iot.eclipse.org/community/resources/iot-surveys/.
[9] Mosquitto. https://mosquitto.org/.
[10] モバイル SINET. https://www.sinet.ad.jp/wadci.
[11] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Renhe Jiang Ryo Nakamura, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizak, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. *In 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 1–7, 2022.
[12] gzip. https://www.gzip.org/.
[13] Zstandard. https://facebook.github.io/zstd/.
[14] perftool. https://github.com/nii-gakunin-cloud/sinetstream/tree/main/python/sample/perftool.

ROS 2準拠ロボット及びクラウドを用いたIoTアプリケーションのためのZenohの通信性能評価

佐々木怜名† 竹房あつ子†† 中田 秀基††† 小口 正人†

† お茶の水女子大学 〒112–8610 東京都文京区大塚 2–1–1
†† 国立情報学研究所 〒101–8430 東京都千代田区一ツ橋 2–1–2
††† 産業技術総合研究所 〒305–8560 茨城県つくば市梅園 1–1–1
E-mail: †{reina,oguchi}@ogl.is.ocha.ac.jp, ††takefusa@nii.ac.jp, †††hide-nakada@aist.go.jp

あらまし ROS 2（Robot Operating System 2）は，ロボットアプリケーション開発のためのソフトウェアプラットフォームである．ROS 2では通信ミドルウェアにDDS（Data Distribution Service）が用いられているが，無線通信環境で大規模データを扱うのに適しておらず，ディスカバリートラフィックなどの課題が存在する．Zenohは低遅延で高スケーラブルなPub/Sub/Query型通信ミドルウェアで，DDSやMQTTなど他プロトコルとの相互運用プラグインも提供しているが，その性能は明らかでない．本研究では，ROS 2で実装されたセンサロボットと中継サーバを用いて，ZenohとDDSをブリッジするZenohプラグインを用いたときのROS 2の通信性能を評価する．

キーワード IoT，クラウドロボティクス，ROS 2，Zenoh

1 はじめに

マルチロボットシステムは，複数のロボットがセンシングしながら協力してタスクを遂行するシステムであり，作業の迅速化や冗長性の確保，高度なタスクの遂行を可能にする．このようなロボットシステムでは，一般にオープンソースソフトウェアのROS(Robot Operating System) [1], [2] が用いられている．ROS 2 (Robot Operating System 2) [3] は，2017年にリリースされたオープンソースで提供されているロボットソフトウェアプラットフォームであり，デファクトスタンダードとして現在多くのロボット開発で用いられている．リアルタイム性を確保するため，通信プロトコルにDDS（Data Distribution Service）[4] が採用されている．ROS 2では，ミドルウェア抽象化インターフェイス (RMW) [5] が定義されており，各ミドルウェアでRMWを実装することで，アプリケーション側でコードを変更することなく，DDS実装を切り替えて使用できる．

DDSは，デフォルトでマルチキャストUDPで通信を行っており，QoSを導入することで，性能，スケーラビリティ，信頼性などの向上を図っている．マルチロボットシステムにおいて，複数のロボットがDDSを使用した無線通信を行う際には，いくつかの課題が存在する．DDSは有線環境を想定して設計されており，クライアント間の通信を始める前に，通信ホスト間の探索時間であるディスカバリーオーバヘッドが，クライアントの数が多ければ多いほど大きくなるため，Wi-Fi環境やマルチロボットシステムにおいてはディスカバリーオーバーヘッドが問題になる．また，無線通信環境下での通信やカメラ画像や点群データなどの大規模データの通信が困難であることが近年報告されている．そこで，新たな通信プロトコルZenoh [6] が注目されている．Zenohは低遅延で高スケーラブルなPub/Sub/Query型の通信プロトコル・ミドルウェアである．IoTデバイスで用いられることを想定して設計されており，DDSやMQTT [7] など他プロトコルとの相互運用プラグインも提供しているが，無線環境におけるそれらの性能について明らかでない．

現在，ROS 2ではZenohに対するRMWが実装されておらず，ROS 2アプリケーション間でZenoh通信を行う場合には，ZenohとDDSをブリッジするzenoh-plugin-ros2dds [8] を用いる必要がある．本研究では，zenoh-plugin-ros2dds（Zenoh Plugin）を用いたときのROS 2の通信性能を評価する．ROS 2で実装された複数の車輪型移動ロボットをセンサ端末として，駆動機能を用いて室内環境情報を収集し，エッジを介して，クラウド上で解析処理を行うようなIoTシステムを想定し，センサロボット間のP2P通信（P2P）とZenoh Routerを介したセンサロボット間の通信（Brokered）について調べる．

Zenoh Pluginを用いたROS 2通信の性能評価に向けて，Zenoh通信とCyclone DDS通信の比較を行い，それぞれの通信性能特性を明らかにする．また，Zenoh通信とZenoh Pluginを用いた時のROS 2通信の比較を行い，Zenoh Pluginの処理オーバヘッドの大きさを明らかにする．さらに，現時点におけるZenoh Pluginを用いたROS 2通信の有用性を確認するため，P2PではZenoh Pluginを用いた時のROS 2通信とCyclone DDSを用いた時のROS 2通信の性能を比較する．BrokeredではZenoh Pluginを用いてZenoh Routerを介したROS 2通信，Kafkaブローカ [9] を介した通信，MQTTブローカを介した通信の性能を比較する．

2 関連研究

ロボットが地図作成，物体把持，物体認識，位置推定などの処理を行う場合，センサから大量の情報を収集し，膨大な計算を実時間で処理することが要求される．しかし，ロボット単体

で処理を行うためには，高性能な計算機の搭載が不可欠となり，ロボットの大型化，高価格化，高電力化に繋がる．そこで，ロボットとクラウドを連携させ，大規模な計算やデータベースを要求するような処理をクラウド上で担う，クラウドロボティクスと呼ばれる考えが普及している．G. Mohanarajah らは，複数のロボットによるクラウドロボティクスのオープンソースプラットフォームである Rapyta を実装した [10]．Rapyuta は，安全でカスタマイズ可能なコンピューティング環境をクラウド上で提供している．ロボットは膨大な計算をクラウドにオフロードし，他のロボットと結果を共有することで，マルチロボットの協調的制御を実現している．Rapyuta は ROS と互換性のある環境を用意しており，ほぼ全てのオープンソース ROS パッケージを変更することなく動作可能である．

車車間通信（V2V）やロボット通信における，車両群による協調判断アプリケーションが研究されている [11], [12]．車両が個々の運転状態を共有し，互いの走行状態を確認できれば，車両間が行き違う際に安全な運転判断を見出すことができる．このようなシステムでは，車両，ロボット，エッジデバイスのネットワークを通じて，一定時間内に一貫性のある判断をしなければならず，不安定なネットワーク環境下でもレイテンシに厳密で，デバイスの冗長性を確保したシステムを構築する必要がある．C.Shih らは，スマート工場や自律走行車群などに応用できる，エッジを中心とした意思決定プロセスを実現するフレームワークを開発している [13]．デバイスの通信ミドルウェアに Zenoh を採用しており，意思決定の制約時間，同時性，スケーラビリティ特性を満たした意思決定システムを構築している．さらに，通信ミドルウェアを Zenoh と Cyclone DDS としたときの性能の違いを評価し，ペイロードが 2KB 以下であれば，意思決定の制約時間内で，高い同時性の要件を達成することを両者の場合で確認している．

ロボットアプリケーションには ROS が広く利用されており，リアルタイム分散組込みシステムへの活用に向けて，ROS や ROS 2 のリアルタイム性が評価されている [14]．ROS や ROS 2 におけるノード間の様々な通信状況について，レイテンシやスループットなどを測定している．また，DDS のベンダーや構成による ROS 2 の性能評価に向けて，異なるデータサイズ及び同一周期でメッセージを通信し，レイテンシ，スループット，スレッド数，メモリ消費量を測定している．

3 実　　験

3.1 実験概要

ROS 2 で実装されたセンサロボットと，中継サーバを用いて，Zenoh と DDS をブリッジするプラグインである zenoh-plugin-ros2dds（Zenoh Plugin）を用いた時の ROS 2 の通信性能を評価する．本実験のシステムの構成について図 1 に示す．センサロボット 1 とセンサロボット 2 間の通信 (P2P) と，中継サーバ上の Zenoh Router を介したセンサロボット 1 とセンサロボット 2 間の通信 (Brokered) の所要時間を調査する．また，センサロボット 1，センサロボット 2，中継サーバにおいて，それぞれ Wi-Fi に接続する場合（Wi-Fi）と Ethernet に接続する場合（Wired）の 2 つの通信環境で測定する．

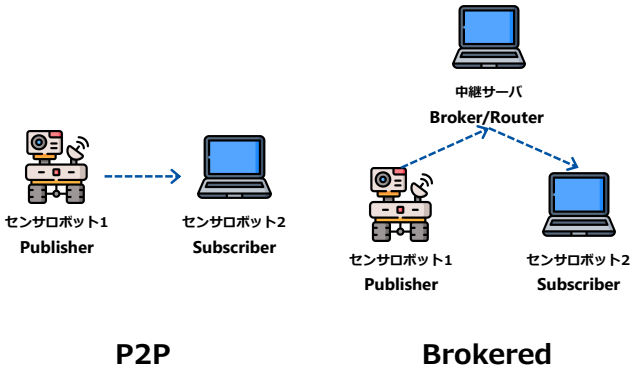


図 1: システム構成

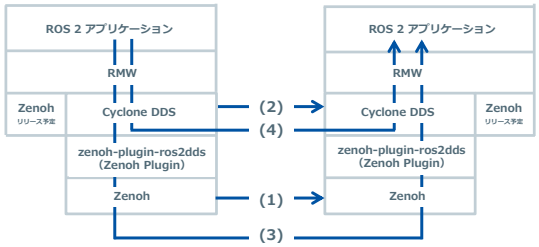


図 2: P2P における実験概要

P2P では，UDP を用いて，センサロボット 1 上の Publisher からセンサロボット 2 上の Subscriber に文字列データを 1000 回パブリッシュする．P2P における実験概要について図 2 に示す．P2P では，(1)Zenoh 通信，(2)Cyclone DDS 通信，(3)Zenoh Plugin を用いた時の ROS 2 通信，(4)Cyclone DDS を用いた時の ROS 2 通信を測定する．zenoh-plugin-ros2dds（Zenoh Plugin）は，ノード間通信を始める前に，Cylone DDS の Discovery Server を介して，DDS の Discovery Server Protocol によりノード同士を見つける．Zenoh Plugin ではメッセージのシリアライズ/デシリアライズを行わない．

Brokered では，TCP を用いて，センサロボット上の Publisher から中継サーバ上の Zenoh Router を介して，センサロボット 2 上の Subscriber に文字列データを 1000 回パブリッシュする．

メッセージ通信処理間には 100 msec の sleep を入れる．異なるホスト間で一方向のメッセージ通信の性能を測定するには，ホスト間で厳密に時刻を同期させる必要があるが，1msec 以下の精度で NTP で時刻同期をさせるのは非常に難しい．よって，受信ホストがメッセージを受け取ったあと，送信ホストに対してソケット通信で小さいデータ (ack data) を送信するようにし，送信ホストでメッセージの送信から ack data を受け取るまでの時間を片側通信時間として計測する．

3.4.1 節で，図 2 における (1)Zenoh 通信と (2)Cyclone DDS 通信の比較を行い，それぞれの通信性能特性を明らかにする．3.4.2 節で，図 2 における (1)Zenoh 通信と (3)Zenoh Plugin

表 1: A: センサロボット 1，B: センサロボット 2・中継サーバの構成

A	機種名	Raspberry Pi 4 Computer Model B 4GB RAM
	CPU	ARM Cortex-A72(1.8 GHz) 4 Cores
	OS	Ubuntu 22.04.02 LTS
	ROS2	Humble
B	機種名	Surface Pro 4
	CPU	Intel®Core™i7-6650U CPU@2.20Ghz × 4
	OS	Ubuntu 20.04.4 LTS
	メモリ	15.6GiB
	ROS 2	Foxy

表 2: ROS 2 通信で使用する DDS と QoS 設定

DDS	Cyclone DDS	
QoS	HISTORY	LAST
	DEPTH	1
	RELIABILITY	RELIABLE

表 3: Kafka と MQTT の QoS 設定

consistency	AT_LEAST_ONCE
data_compression	false
data_encryption	false

を用いた時の ROS 2 通信の比較を行い，Zenoh Plugin の処理オーバヘッドの大きさを明らかにする．3.4.3 節で，図 2 における (3)Zenoh Plugin を用いた時の ROS 2 通信と (4)Cyclone DDS を用いた時の ROS 2 通信の比較を行う．さらに，3.5 節で，Zenoh Plugin を用いて Zenoh Router を介した ROS 2 通信，Kafka 通信，MQTT 通信を比較し，現時点における Zenoh Plugin を用いた ROS 2 通信の有用性を確認する．

3.2 実験環境

実験に用いるセンサロボット 1，センサロボット 2，中継サーバの詳細を表 1 に，ROS 2 通信で使用する DDS と QoS 設定を表 2 に，SINETStream API における Kafka ブローカと MQTT ブローカの QoS 設定を表 3 に示す．本実験では，通信ミドルウェア実装の通信性能を比較をするため，センサロボット 1 およびセンサロボット 2 と同一 LAN 内に中継サーバを配置し，そこで Kafka ブローカ，MQTT ブローカ，Zenoh Router を動作させる．センサロボット 1 には Ubuntu22.04 をインストールした Raspberry Pi 4 Computer を，センサロボット 2 と中継サーバには，Ubuntu20.04 をインストールした Surface Pro4 を用いる．センサロボット 1 および 2 が接続するアクセスポイントとして Buffalo WSR-2533DHP2-CG を使用し，センサロボット 1 は IEEE802.11n(256QAM)，センサロボット 2 と中継サーバは IEEE802.11ac の規格で接続している．

以降で，実験システムの構成技術である Zenoh，SINETStream の概要について述べる．

3.2.1 Zenoh

Zenoh は，zetta scale 社が開発している，低遅延で高スケーラブルな Pub/Sub/Query 型の通信プロトコル・ミドルウェアである．多様な Pub/Sub 通信方式や分散クエリの機能を提供している．任意のネットワークトポロジーを組むことが可能であり，Peer-to-Peer(P2P) 通信だけでなく，Zenoh Router を介して複数のクライアント同士で通信することができる．Rust で実装されており，Rust，Python，C などの API を提供している．軽量なデバイスやクラウドサーバなど多様な計算機で用いられることを想定して設計されており，DDS や MQTT など他プロトコルとの相互運用プラグインも提供している．2024 年 5 月には，Zenoh に対応した ROS 2 の RMW がリリース予定であり，それまで ROS 2 準拠ロボットで Zenoh を用いた通信を行う場合は，図 2 で示すように，Zenoh と DDS をブリッジするプラグインを用いる必要がある．

表 4: センサロボット間の iPerf3 測定結果

	sender	receiver
Wired	922 Mbps	920 Mbps
Wi-Fi	31.2 Mbps	30.2 Mbps

表 5: センサロボット間の Ping 結果

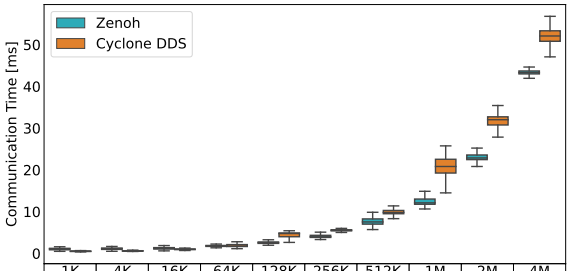
	RTT [ms]				packet loss
	min	avg	max	mdev	
Wi-Fi	18.4	204	1124	191	0%
Wired	0.385	0.870	1.45	0.223	0%

3.2.2 SINETStream

広域に分散したデータを活用する研究では，広域ネットワークを介して，センサ等から取得されるデータを欠損なく確実に収集し，解析に用いることが求められている．本実験では，Kafka と MQTT を用いる実験で，SINETStream [15], [16] を利用する．SINETStream は，国立情報学研究所で開発された広域データ収集・解析プログラム開発支援ソフトウェアパッケージであり，安全確実なデータ収集，認証・認可，圧縮/解凍，暗号化，メトリクス収集機能といった，IoT データの収集・蓄積・解析に必要な機能が予め提供されている．また，IoT 通信プロトコルを抽象化する API が提供されているため，センサデータの通信パターンに応じて適宜バックエンドに配備するメッセージブローカを選択できるという利点もある．

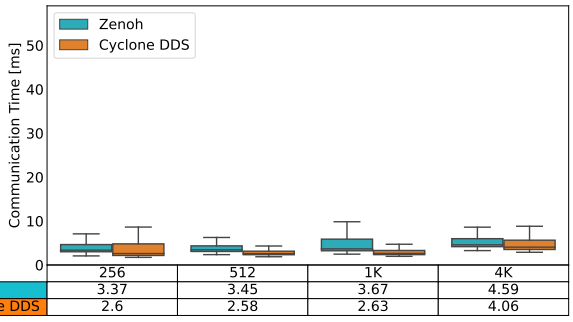
3.3 ベースラインの通信性能

ベースライン通信性能を確認するため，iPerf3 [17] の通信スループット測定，および ping コマンドでの通信遅延の測定を行う．iPerf3 で測定したセンサロボット 1 とセンサロボット 2 間の通信スループットを表 4 に示す．Wi-Fi 接続（Wi-Fi）より Ethernet 接続（Wired）の方が，通信スループットが約 30 倍高いことがわかる．また，Wired，Wi-Fi において，センサロボット間，センサロボットと中継サーバ間で，ping を用いて RTT を測定した．50 回測定した結果を表 5 に示す．Wi-Fi 接続（Wi-Fi）より Ethernet 接続（Wired）の方が，平均の RTT が約 230 倍速いことがわかる．



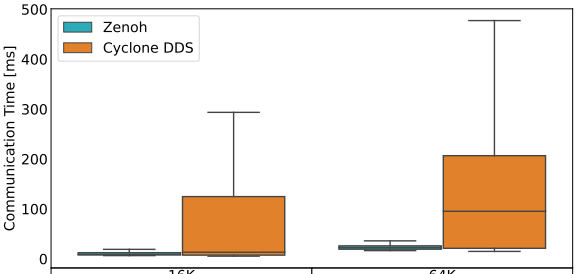
	1K	4K	16K	64K	128K	256K	512K	1M	2M	4M
Zenoh	1.1	1.17	1.27	1.79	2.57	3.99	7.55	12.1	22.9	43.3
Cyclone DDS	0.505	0.601	0.987	1.76	4.76	5.53	9.82	20.8	32.0	52.1

図 3: Wired での P2P における Zenoh 通信と Cyclone DDS 通信の比較



	256	512	1K	4K
Zenoh	3.37	3.45	3.67	4.59
Cyclone DDS	2.6	2.58	2.63	4.06

図 4: Wi-Fi での P2P における Zenoh 通信と Cyclone DDS 通信の比較



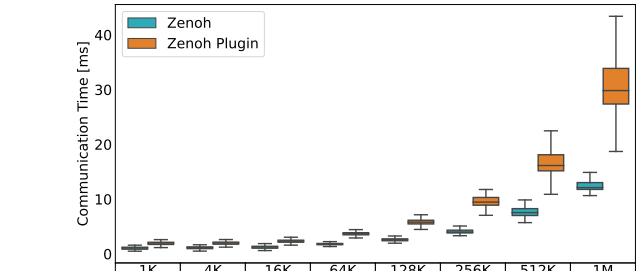
	16K	64K
Zenoh	8.84	22.2
Cyclone DDS	13.2	95.3

図 5: Wi-Fi での P2P における Zenoh 通信と Cyclone DDS 通信の比較

3.4 P2P おける Zenoh Plugin の通信性能評価

3.4.1 Zenoh 通信と Cyclone DDS 通信の比較

センサロボット間の P2P 通信における，Zenoh 通信と Cyclone DDS 通信の通信時間の分布の結果について，Wired のとき図 3 に，Wi-Fi でメッセージサイズが 256B から 4KB のときを図 4 に，Wi-Fi でメッセージサイズが 16KB と 64KB のときを図 5 に示す．以降，各グラフにおいて，箱の下端が第一四分位数，箱の中の線が中央値，箱の上端が第三四分位数を示す．横軸はメッセージサイズを Byte 単位で示し，縦軸は通信時間をミリ秒単位で示す．グラフの表は各サイズの通信時間の中央値を示す．Cyclone DDS ではマルチキャスト UDP 通信を用いており，パケットロスが発生すると再送されない．そのため，Wi-Fi の 16KB と 64KB のときの実験のみ，100 個のデータ転送を 10 回行った結果をまとめている．Wired のとき，メッセージサイズが 64KB 以下では Cyclone DDS の方が通信時間が短いが，メッセージサイズが大きくなるほど，Zenoh の方が通信時間が短く，分布の広がりが小さいことがわかる．また，Wi-Fi のときでは，メッセージサイズが 256B から 4KB までの時は Zenoh より Cyclone DDS の方が中央値は小さいものの，データの分布の広がりが大きく，16KB と 64KB の時は，Cyclone DDS の通信時間や分布の広がりが急激に増加している．これより，データサイズが小さく Ethernet に接続する場合では，Zenoh より Cyclone DDS の方が通信時間が短いが，データサイズが大きいときや，Wi-Fi に接続する場合では Zenoh 通信の方が通信性能が良いことがわかる．



	1K	4K	16K	64K	128K	256K	512K	1M
Zenoh	1.1	1.17	1.27	1.79	2.57	3.99	7.55	12.1
Zenoh Plugin	1.99	1.99	2.33	3.69	5.8	9.47	16.1	29.8

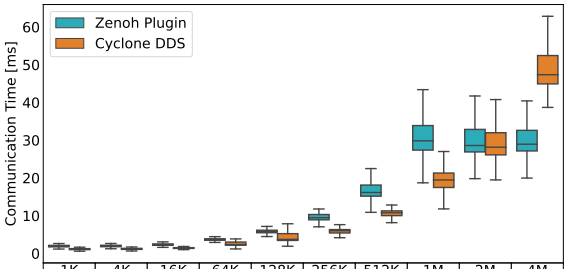
図 6: Wired での P2P における Zenoh 通信と Zenoh Plugin を用いた ROS 2 通信の比較

3.4.2 Zenoh 通信と Zenoh Plugin を用いた ROS 2 通信の比較

センサロボット間の P2P 通信における，Zenoh 通信と Zenoh Plugin を用いたときの ROS 2 通信の通信時間の分布の結果について，図 6 に示す．メッセージサイズが大きくなるにつれ，Zenoh Plugin を用いた ROS 2 通信の通信時間や分布の広がりが大きくなることがわかる．Zenoh Plugin では，ROS 2 アプリケーションで宣言された DDS の Writer と Reader の検出に Cyclone DDS を用い，DDS と Zenoh をブリッジしてデータ転送をしている．そのため，Zenoh Plugin は，Cyclone DDS によるディスカバリーオーバヘッドにより，Zenoh 通信よりも通信時間が大きくなると考えられる．

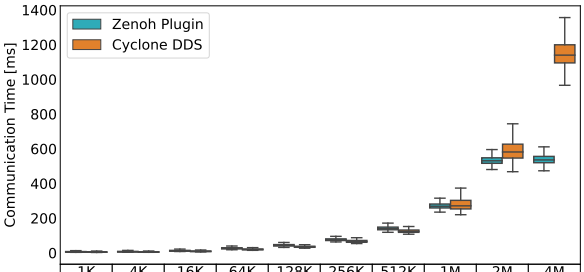
3.4.3 Zenoh Plugin と Cyclone DDS を用いた ROS 2 通信の比較

センサロボット間の P2P 通信における，Zenoh Plugin を用いた時の ROS 2 通信と Cyclone DDS を用いた時の ROS 2 通信の通信時間の分布の結果を,Wired のとき図 7 に，Wi-Fi のとき図 8 に示す．図 7，図 8 より，メッセージサイズが小さい時の通信時間の中央値は，Zenoh Plugin を用いた時の方が大きい．しかし，Wi-Fi では 2MB 以上のとき，Wired では 4MB のとき，Zenoh Plugin を用いた時の方が中央値が小さくなる．特に Wi-Fi では両者の差が大きくなる．Zenoh Plugin を用いた時の ROS 2 通信の通信時間には ROS 2 メッセージのシリアライズ/デシリアライズの処理時間は含まれない．これより，メッセージサイズが大きい場合や，Wi-Fi 環境下では，Zenoh Plugin での処理よりも Cyclone DDS の通信オーバーヘッドが



	1K	4K	16K	64K	128K	256K	512K	1M	2M	4M
Zenoh Plugin	1.99	1.99	2.33	3.69	5.8	9.47	16.1	29.8	28.6	28.0
Cyclone DDS	1.16	1.24	1.45	2.36	3.72	6.01	10.8	19.5	28.2	47.4

図 7: Wired での P2P における Zenoh Plugin と Cyclone DDS を用いた ROS 2 通信の比較



	1K	4K	16K	64K	128K	256K	512K	1M	2M	4M
Zenoh Plugin	5.36	6.03	11.6	26.0	43.2	75.4	140.0	269.0	531.0	536.0
Cyclone DDS	4.27	4.59	8.88	19.7	34.5	66.0	124.0	271.0	582.0	1149.0

図 8: Wi-Fi での P2P における Zenoh Plugin と Cyclone DDS を用いた ROS 2 通信の比較

大きいかった．

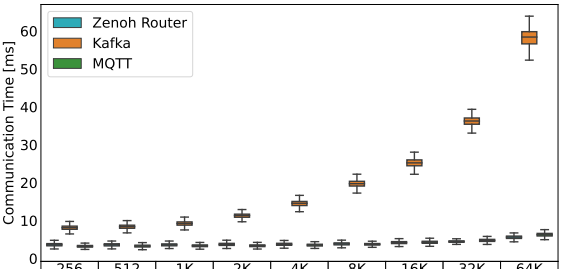
3.5 Brokered における Zenoh Plugin の通信性能評価

中継サーバを介したセンサロボット間の Brokered 通信における通信時間の分布について，Zenoh Plugin を用いて Zenoh Router を介した時の ROS 2 通信と Kafka 通信と MQTT 通信を比較したグラフを図 9 と図 10 に示す．図 9，図 10 より，全てのメッセージサイズにおいて，Kakfa より Zenoh Router を用いたときの方が通信時間の中央値や分布の広がりは小さい．また，メッセージサイズが 16KB から 256KB のときは MQTT よりも Zenoh を用いたときの方が中央値や分布の広がりが小さいが，それ以外のサイズでは Zenoh の方が大きく，1MB のときはその差が約 3.5 倍になることがわかる．ただし，Zenoh Router では通信の中継以外の処理をしておらず，Kafka や MQTT と Zenoh Router における処理は異なるため，同列に比較することは困難であるが，単純なメッセージの中継を行う場合には，Kafka よりも Zenoh の方が有効であることが示された．

図 6 から，Zenoh Plugin を用いた ROS 2 通信において，P2P より Brokered の通信時間の中央値が，メッセージサイズが 1MB のとき約 6 倍大きい．これより，Zenoh Router 上での中継処理のオーバーヘッドがあることがわかる．

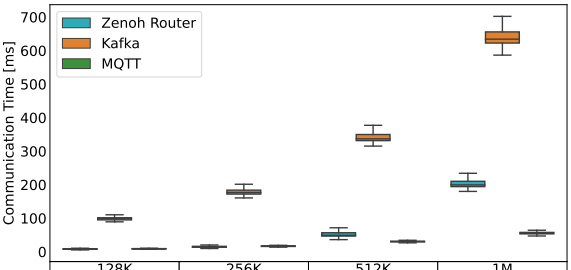
4 まとめと今後の課題

ROS 2 準拠センサロボット 1 とセンサロボット 2，中継サーバを利用した IoT システムを想定し，P2P と Brokered の 2 つのトポロジーにおける通信性能を評価した．Zenoh と DDS をブリッジするプラグインを用いた ROS 2 通信の性能評価に向けて，Zenoh 通信，Cyclone DDS 通信，Zenoh Plugin を用いた ROS 2 通信，Cyclone DDS を用いた ROS 2 通信，Zenoh Plugin を用いて Zenoh Router を介した時の ROS 2 通信，Kafka 通信，MQTT 通信の，それぞれの通信時間を測定した．



	256	512	1K	2K	4K	8K	16K	32K	64K
Zenoh Router	3.62	3.63	3.62	3.72	3.75	3.87	4.2	4.47	5.64
Kafka	8.13	8.4	9.24	11.3	14.6	19.7	25.2	36.3	58.4
MQTT	3.22	3.28	3.35	3.37	3.55	3.75	4.29	4.77	6.26

図 9: Wired での Brokered における Zenoh，Kafka，MQTT の比較: メッセージサイズが 256B から 64KB の時



	128K	256K	512K	1M
Zenoh Router	8.56	14.2	49.1	199.0
Kafka	99.0	177.0	337.0	634.0
MQTT	9.07	16.6	30.5	55.4

図 10: Wired での Brokered における Zenoh，Kafka，MQTT の比較: メッセージサイズが 128KB から 1MB の時

Cyclone DDS による通信は，データサイズが大きいときや Wi-Fi に接続する場合では，Zenoh による通信よりも通信性能が劣ることがわかった．Zenoh Plugin では，Cyclone DDS によるディスカバリーオーバヘッドにより，Zenoh 通信よりも通信時間が大きくなることがわかった．メッセージサイズが小さい時や，Ethernet に接続する場合，Cyclone DDS を用いたこれまでの ROS 2 通信の性能は Zenoh Plugin を用いた通信に劣らないが，メッセージサイズが大きい時や，Wi-Fi 接続の時では，Zenoh Plugin の方が通信性能が良いことがわかった．マルチロボットシステムのような，複数ロボットが無線環境で接続し，カメラ画像や点群データを通信する場合には，既存の DDS よりも，Zenoh を用いた通信の方が性能が良いと考えられる．Brokered において，単純なメッセージの中継で，Zenoh Plugin を用いて Zenoh Router を介した ROS 2 通信の通信時間は，Kafka 通信よりも小さいが，メッセージサイズによっては MQTT 通信よりも大きいことがわかった．

今後の課題は，Zenoh に対応した RMW の性能評価を行う．また，クライアント数を増やし，より複雑な通信トポロジーを

組んだ場合の通信性能の評価を行う．さらに，効率的な環境情報収集に向けて，複数センサロボットの適切な制御方法を検討する．

謝　　辞

本研究は JSPS 科研費 JP19H04089 及び 2023 年度国立情報学研究所公募型共同研究 (2023-23S0205) の助成を受けたものです．

文　　献

[1] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. Ros: an open-source robot operating system. In *Proc. of IEEE International Conference on Robotics and Automation Workshop on Open Source Software*, Vol. 3, p. 5, 2009.
[2] ROS wiki. `http://wiki.ros.org/`.
[3] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette Lalancette, and William Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Science Robotics*, Vol. 7, No. 66, p. eabm6074, 2022.
[4] G. Pardo-Castellote. Omg data-distribution service: architectural overview. In *23rd International Conference on Distributed Computing Systems Workshops, 2003. Proceedings.*, pp. 200–206, 2003.
[5] 2023-09 ROS 2 RMW alternate. `https://discourse.ros.org/t/ros-2-alternative-middleware-report/33771`.
[6] Zenoh. `https://zenoh.io/`.
[7] MQTT. `https://mqtt.org/`.
[8] zenoh-plugin-ros2dds. `https://github.com/eclipse-zenoh/zenoh-plugin-ros2dds`.
[9] Apache Kafka. `https://kafka.apache.org/`.
[10] Gajamohan Mohanarajah, Dominique Hunziker, Raffaello D'Andrea, and Markus Waibel. Rapyuta: A cloud robotics platform. *IEEE Transactions on Automation Science and Engineering*, Vol. 12, No. 2, pp. 481–493, 2015.
[11] Georgios Karagiannis, Onur Altintas, Eylem Ekici, Geert Heijenk, Boangoat Jarupan, Kenneth Lin, and Timothy Weil. Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions. *IEEE Communications Surveys Tutorials*, Vol. 13, No. 4, pp. 584–616, 2011.
[12] Gabriele Baldoni, Julien Loudet, Luca Cominardi, Angelo Corsaro, and Yong He. Zenoh-based dataflow framework for autonomous vehicles. In *2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pp. 555–560, 2021.
[13] Chi-Sheng Shih, Hsiang-Jui Lin, Yuyuan Yuan, Yi-Hung Kuo, and Wen-Yew Liang. Scalable and bounded-time decisions on edge device network using eclipse zenoh. In *2022 IEEE 28th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pp. 170–179, 2022.
[14] Yuya Maruyama, Shinpei Kato, and Takuya Azumi. Exploring the performance of ROS2. In *2016 International Conference on Embedded Software (EMSOFT)*, pp. 1–10, 2016.
[15] SINETStream. `https://sinetstream.net/`.
[16] Atsuko Takefusa, Jingtao Sun, Ikki Fujiwara, Hiroshi Yoshida, Kento Aida, and Calton Pu. SINETStream: Enabling research iot applications with portability, security and performance requirements. In *Proc. COMPSAC 2021*, pp. 482–492, 2021.
[17] iPerf3. `https://iperf.fr/`.

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

データ基盤・性能評価

座長:杉浦 健人(名古屋大学)
コメンテータ:小山田 昌史(NEC)
2024年2月29日(木) 13:30 〜 15:40 T2-A (オンライン（Zoom Events))

15:15 〜 15:40

[T2-A-5-05]【技術報告】広告配信における安定して拡張性のある大量データ処理基盤の必要性と活用

大澤 昂太 (株式会社マイクロアド システム開発部)

広告配信業務ではインターネット上で発生する大量のWebリクエストを24時間処理する必要があります。このようなデータの処理量は一日に数TB単位になります。

大量なデータを処理するには量のデータを処理するには1台のサーバでは厳しい場合が多いので、多くの場合分散処理、クラスタ、負荷分散など複数サーバをまとめて処理させるような技術が欠かせません。

弊社では広告配信サーバ、データレイクやDWHに数百台、RDBMSのMySQL、分散ストリーミング処理を行うApache Kafka、In-memoryのKVS型データベースのRedisクラスタなどがそれぞれ数十台程度で構成しています。これらのサーバは性能劣化、同期、トランザクションなどを様々な問題を抱えていて日々対処を行っています。

今回は、大量データを処理する実務の様々なポイントについてついて説明したいと思います。

©一般社団法人日本データベース学会

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 13:30 ~ 15:40 | T2-B オンライン（Zoom Events)

類似検索アルゴリズム

座長:白井 匡人(島根大学)
コメンテータ:鈴木 伸崇(筑波大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T2-B-5-01] 逆 k-MIPSに基づく影響力の大きいアイテムの高速検索

*青山 和禎[1]、天方 大地[1]、藤田 澄男[2]、原 隆浩[1] (1. 大阪大学大学院情報科学研究科、2. LINEヤフー株式会社)

14:00 ~ 14:25
[T2-B-5-02] A Grid-density-based Approach for Dense Nearest Neighborhood Query

*謝 若琳[1]、陳 漢雄[1]、古瀬 一隆[2] (1. 筑波大学、2. 白鴎大学)

14:25 ~ 14:50
[T2-B-5-03] アニメーション付き 3 次元アセンブリモデル検索のための3次元ラドン変換の効率化

*内尾 亮太[1]、片山 薫[1] (1. 東京都立大学)

14:50 ~ 15:15
[T2-B-5-04] 大規模DNAデータベースに対するTop-*k*相関問合せ手法の提案

*八木 隆一[1]、直井 悠馬[1]、塩川 浩昭[2] (1. 筑波大学大学院理工情報生命学術院、2. 筑波大学計算科学研究センター)

15:15 ~ 15:40
[T2-B-5-05] テキスト及び構造に基づく効率的な類似部分木検索手法

*溝上 拓也[1]、Bou Savong[2]、天笠 俊之[2] (1. 筑波大学 情報学群情報科学類、2. 筑波大学 計算科学研究センター)

逆 k-MIPS に基づく影響力の大きいアイテムの高速検索

青山　和禎† 　天方　大地† 　藤田　澄男†† 　原　隆浩†

† 大阪大学　〒 565–0871 大阪府吹田市山田丘 1-5
†† LINE ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3 紀尾井タワー
E-mail: †{aoyama.kazuyoshi,amagata.daichi,hara}@ist.osaka-u.ac.jp, ††sufujita@lycorp.co.jp

あらまし　近年，最大内積探索（k-MIPS）が多くのアプリケーションに応用されている．特に推薦システムにおいて，ユーザがアイテムに対して行う評価を予測し，評価予測値の高い上位 k 個のアイテムを検索する k-MIPS 問題を解くことが一般的である．また，あるアイテムを k-MIPS 結果に含むユーザの集合を検索する逆 k-MIPS（reverse k-MIPS）問題がある．多くのユーザの k-MIPS 結果に入るアイテムは影響力が強いといえるため，あるアイテムの影響力は，そのアイテムの逆 k-MIPS の結果のカーディナリティと定義できる．本論文では，逆 k-MIPS に基づく影響力の大きいアイテムの検索問題を扱う．この問題は，アイテム集合，ユーザ集合，k，および N が与えられたとき，スコアの大きい N 個のアイテムを探索する問題であり，あるアイテムのスコアはそのアイテムを k-MIPS 結果に含むユーザの数である．我々は，この問題を高速に解くためのアルゴリズムを提案する．提案アルゴリズムは，前処理において，$O(nm)$ 時間で各アイテムのスコアの上界値を計算する（n および m それぞれユーザ数およびアイテム数）．オンラインでは，前処理で計算した上界値を使って不要な逆 k-MIPS の計算を枝刈りする．実世界のデータを用いた実験により，提案アルゴリズムが既存の技術と比較して優れた性能であることを示す．

キーワード　内積探索，逆 k-MIPS

1　はじめに

内積探索は，情報検索 [15,16]，推薦システム [1,6]，データマイニング [7,12]，人工知能 [3,8]，データベース [9,17]，および機械学習 [4,10] 等の多くの分野において重要な問題である．これらの分野では，(d 次元）クエリベクトル $\mathbf{q}$ が与えられたとき，$\mathbf{q}$ との内積が最大となるベクトルを探索する k-MIPS 問題が利用されている．例えば，推薦システムにおいて，ユーザがアイテムに対して行う評価値を予測し，評価予測値の高い上位 k 個のアイテムを検索するために用いられ，k-MIPS 結果はユーザの推薦リストに対応する．ここで，k-MIPS 問題を応用した問題として，逆 k-MIPS 問題がある [1,8]．逆 k-MIPS 問題では，アイテムベクトルをクエリとし，そのアイテムを k-MIPS 結果に含む全てのユーザを検索する．この問題は，企業が新商品やおすすめ商品に対して効果的なマーケティングを行う際に利用される問題である．

本論文では，逆 k-MIPS に基づく影響力の大きいアイテムの検索問題に取り組む．この問題は，アイテム集合，ユーザ集合，k，および N が与えられたときに，逆 k-MIPS 結果のサイズを影響力スコアと定義し，このスコアが最も大きい N 個のアイテムを探索する問題である．多くのユーザの k-MIPS 結果に含まれるアイテムというのは，市場において影響力の大きい（潜在的に人気な）アイテムといえる．市場におけるアイテムの影響力は，そのアイテムを k-MIPS 結果に含む全てのユーザの数，つまり逆 k-MIPS の出力結果のサイズと定義できる．影響力の大きいアイテムの検索の応用例ととして，市場分析が挙げられる．例えば，影響力の大きいアイテムを分析できれば，(今後の）市場においてユーザがどのようなアイテムを好みやすいかを理解することができ，新商品の開発，既存商品の改良，およびマーケティングの最適化に役立つ．また，逆 k-MIPS に基づく影響力の大きいアイテムというのは，単純な購買数や閲覧数に基づく人気のアイテムとは異なる．逆 k-MIPS に基づく影響力の大きいアイテムというのは，多くのユーザにとって評価「予測」値が上位に入るアイテムであり，購買数や閲覧数に基づく人気のアイテムよりも潜在的な人気を考慮できる．また，この「潜在的に人気」なアイテムの見極めは，将来の（大局的な）購買パターンの予測にも役立つ．

この問題を解く素朴なアルゴリズムとして，既存の k-MIPS 手法を用いて全ユーザに対して k-MIPS 結果を計算し，各アイテムが何名の k-MIPS 結果に含まれているかを調べ，それが最も多い N 個のアイテムを出力するものが考えられる．しかし，単純に全てのユーザについて k-MIPS 結果を計算することは，大規模なデータベースにおいて非常に重い処理である．また，既存の逆 k-MIPS 手法を用いて，全てのアイテムについて逆 k-MIPS を計算する手法も考えられる．しかし，この手法も全てのアイテムについて逆 k-MIPS 結果を計算する必要があり，これも大規模なデータベースにおいて非常に重い処理である．

以上から，素朴なアルゴリズムでは，この問題を高速に計算することは困難である．そのため，本論文ではこの課題を解決する新しいアルゴリズムを提案する．提案アルゴリズムは，前処理において，$O(nm)$ 時間で各アイテムのスコアの上界値を計算する（n および m それぞれユーザ数およびアイテム数）．ここで，この前処理時間は次元数 d の要素がないことに注意する．オンラインでは，前処理で計算した上界値を使って不要な逆 k-MIPS の計算をフィルタリングする．

本論文の主な貢献は以下である．

- 逆 k-MIPS に基づく最も影響力の大きいアイテムを検索する問題を定式化する．
- この問題を高速に解くためのアルゴリズムを提案し，提案アルゴリズムの性能を理論的に分析する．
- 実世界のデータセットを用いて大規模な実験を行い，比較手法よりも高速であることを示す．

本論文の構成．以降では，2 章で問題定義およびベースラインアルゴリズムを紹介し，3 章で関連研究について述べる．次に，4 章で提案手法を紹介し，実験結果を 5 章で報告する．最後に 6 章で結論を述べる．

2 予 備 知 識

2.1 問 題 定 義

$\mathbf{P}$ を m 個のアイテムベクトル $\mathbf{p}_{(1)},...,\mathbf{p}_{(m)}$ からなるアイテム集合とし，$\mathbf{U}$ を n 個のユーザベクトル $\mathbf{u}_{(1)},...,\mathbf{u}_{(n)}$ からなるユーザ集合とする．各ベクトル $\mathbf{p} \in \mathbf{P}$, $\mathbf{u} \in \mathbf{U}$ は d 次元ベクトルであり，$\mathbf{p} = (p_1,...,p_d)$, $\mathbf{u} = (u_1,...,u_d)$ で表される．本研究では，d は大きい（数百），つまり，高次元ベクトルを想定する．

定義 1 (内積) *2* つのベクトル $\mathbf{p} = (p_1,...,p_d) \in \mathbb{R}^d$ と $\mathbf{u} = (u_1,...,u_d) \in \mathbb{R}^d$ が与えられた時，内積 $\mathbf{p} \cdot \mathbf{u}$ は以下で求まる．

$$\mathbf{p} \cdot \mathbf{u} = \sum_{i=1}^{d} p_i \cdot u_i$$

そして，k-MIPS（k Maximum Inner Product Search，または k 最大内積探索）問題は次のように定義できる．

定義 2 (k-MIPS) アイテムベクトル集合 $\mathbf{P}$，クエリ（ユーザ）ベクトル $\mathbf{q}$，および k が与えられたとき，*k-MIPS* は，q との内積が最大となる k 個のアイテムベクトル $\mathbf{p} \in \mathbf{P}$ を検索する．

k-MIPS とは異なり，アイテムをクエリとしたとき，そのアイテムを好むユーザの集合を検索する逆 k-MIPS 問題 [1] は次のように定義できる．

定義 3 (逆 k-MIPS) アイテムベクトル集合 $\mathbf{P}$，ユーザベクトル集合 $\mathbf{U}$，クエリ（アイテム）ベクトル $\mathbf{q}$，および k が与えられたとき，逆 *k-MIPS* 問題は，$\mathbf{P} \cup \mathbf{q}$ の中で $\mathbf{q}$ が $\mathbf{u}$ の *k-MIPS* 結果に属する全てのユーザベクトル $\mathbf{u} \in \mathbf{U}$ を検索する．

ここで，逆 k-MIPS の出力結果サイズが大きいアイテムほど，潜在的に人気の高い，つまり，多くのユーザに影響力があるアイテムと考えることができる．そのため，アイテム $\mathbf{p}$ の影響力スコアを以下のように定義する．

定義 4 (影響力スコア) アイテムベクトル集合 $\mathbf{P}$，ユーザベクトル集合 $\mathbf{U}$，アイテムベクトル $\mathbf{p}$，および k が与えられたとき，アイテム $\mathbf{p}$ の影響力スコアは、$\mathbf{p}$ の逆 *k-MIPS* 結果のカーディナリティと定義する．

この定義に基づいて，我々の取り組む問題を定義する．

定義 5 （**影響力スコアの大きい *top-N* アイテム検索**）アイテムベクトル集合 $\mathbf{P}$，ユーザベクトル集合 $\mathbf{U}$，k，および N が与えられたとき，影響力スコアの大きい N 個のアイテムを検索する．（N 番目に大きい影響力スコアをもつアイテムが複数存在する場合は任意の方法で順位付けする．）

本論文では，この問題を正確かつ高速に解く厳密解アルゴリズムを提案する．また，指定可能な k の最大値（$k_{max} = O(1)$）が存在すると想定する．

2.2 ベースラインアルゴリズム

この問題を解く素朴なアルゴリズムとして，既存の k-MIPS 手法を用いて全ユーザに対して k-MIPS を計算するアルゴリズムが考えられる．全てのアイテムの影響力スコアの初期値を 0 とする．

(1) あるユーザについて，既存の k-MIPS アルゴリズムを用いて，k-MIPS 結果を計算する．

(2) その k-MIPS 結果に含まれているアイテムの影響力スコアを $+1$ する．

(3) 全てのユーザについて (1) および (2) の処理を行い，各アイテムの影響力スコアを計算する．

(4) 上の結果を降順にソートし，影響力スコアの大きい N 個のアイテムを出力する．

このアルゴリズムでは，常に全てのユーザの k-MIPS 結果を完全に求め，全てのアイテムの影響力スコアを計算するため，$|\mathbf{P}|$ および $|\mathbf{U}|$ が大きい場合に計算時間が長くなる．

また，既存の逆 k-MIPS 手法を用いて，全アイテムについて逆 k-MIPS を計算する手法も考えられる．全アイテムについて逆 k-MIPS を計算し，その出力結果サイズが大きい N 個のアイテムを出力する手法である．このアルゴリズムも常に全てのアイテムの影響力スコアを計算するため，先述のベースラインアルゴリズムと同様の問題がある．

3 関 連 研 究

3.1 *k*-MIPS 手法

先述の通り，逆 k-MIPS に基づく影響力の大きいアイテムの検索問題は，k-MIPS アルゴリズムを全てのユーザに対して実行し，各アイテムが何名の k-MIPS 結果に含まれるかを調べることで解くことができる．既存の k-MIPS アルゴリズムに関する研究は，厳密解アルゴリズムおよび近似解アルゴリズムともに多くのが提案されている．

厳密解ベースのアルゴリズムには，空間分割に基づく手法と線形スキャンに基づく手法がある．空間分割に基づく手法とし

て，文献 [11] では，木構造に基づく分枝限定法を用いてベクトルを効率的に探索する Dual-Tree，および Dual-Tree の効率を上げるための新しいデータ構造 Cone-Tree が提案されている．Dual-Tree では，ベクトル集合だけでなくクエリの集合も木構造を用いてインデックスを構築する．クエリが与えられたとき，構築した 2 つの木構造を用いてデータベクトルを探索する．こうすることで，類似したクエリの集合について木の探索コストを減らすことができる．しかし，空間分割ベースの手法は次元の呪いによって高次元空間においては機能せず，数十次元まで次元が増加すると，線形スキャンよりも性能が悪くなることが知られている [14]．そのため，線形スキャンに基づく手法も提案されている．線形スキャンベースの手法として，LEMP [13] は，コーシーシュワルツの不等式を用いたスキャンの早期終了と，各ベクトルの内積を計算する際に，部分内積を用いたタイトな内積上限を用いて不必要な計算を省く Incremental Pruning を用いることで高速に k-MIPS を計算する．FEXIPRO [9] は SVD 変換，整数近似，および正数変換によって厳密な内積上限を設定し，LEMP で提案されているスキャンの早期終了の効果を高める手法である．

これらの厳密解ベースのアルゴリズムは，ベクトル数を n としたときに $O(n)$ 時間必要であるため，$o(n)$ 時間で解ける近似アルゴリズムも提案されている．最大内積探索問題を最近傍探索問題や最大コサイン類似度探索問題に変換し，LSH（局所鋭敏型ハッシュ）を用いて解く手法がある．文献 [7] は，k-MIPS 問題から近似最近傍探索問題へと変換するためのクエリとベクトルの変換に歪み誤差の少ない変換を提案している．また，文献 [7] で提案された近似内積探索アルゴリズム H2-ALSH は，前処理フェーズにおいて，ノルムの大きさに応じてベクトルの集合をいくつかのサブセットに分割してハッシュテーブルに格納する．クエリフェーズでは，各サブセットに対して，ノルムの降順でサブセット内のベクトルの個数に応じて，線形スキャンまたは LSH を用いて近似最近傍探索問題を解く．また，近接グラフを用いた手法も提案されている．ip-NSW [5] は，類似度関数に内積を用いて近接グラフを作成し，クエリとの内積が大きくなるノードを貪欲に探索する．ip-NSW には，近接グラフを作成する際にノルムの小さいベクトルはエッジが結ばれづらいという問題があるが，NAPG [12] では，グラフを作成するノルムを調整するパラメータを設定することで解決している．また，文献 [4] ではベクトル量子化を用いた手法 ScaNN を提案している．ベクトル量子化は，データベクトル集合を有限個のセントロイドに置き換えることで，大量のデータを少ない代表パターンに置き換え圧縮する手法である．ScaNN では，内積の特性を考慮して量子化の際の損失関数にペナルティを与え，量子化によって生じる誤差を軽減させている．

近似 k-MIPS アルゴリズムを用いて影響力の大きいアイテムを検索する場合，実際には影響力が小さいアイテムが，高い影響力を持つと誤って評価されてしまうリスクが存在する．また，例えば市場分析に用いる場合，影響力の大きいアイテムを特定することに加えて，その影響力スコアの把握も重要である．したがって，本論文では近似 k-MIPS アルゴリズムは考えない．

3.2 逆 k-MIPS 手法

本論文が取り組む問題は，全てのアイテムに対して逆 k-MIPS を行うことで解くことができる．Simpfer [1] は，逆 k-MIPS 問題に対する厳密解アルゴリズムであり，そのアイデアはクエリとなるアイテムがユーザの k-MIPS 結果に入るかどうかを k 番目の内積値の上限と下限を用いて高速に判定することである．具体的には，ノルムの大きい $O(k_{max})$ 個のアイテムと各ユーザについて内積を計算し，各ユーザについて k 番目の内積値の下限を得る．また，コーシーシュワルツの不等式を用いて，k 番目の内積値の上限を計算でき，この上限と下限を用いることで，クエリアイテムがユーザの k-MIPS 結果に入るかどうかを定数時間で判定する．ここで，k-MIPS 結果に入るか，または入らないかが確定しなかった場合に線形スキャンを行い，クエリアイテムがユーザの k-MIPS 結果に含まれるかどうかを検査する．本論文の提案アルゴリズムでは，あるアイテムの影響力スコアを計算する必要がある場合に Simpfer の応用を利用する．

また，逆 k-MIPS 問題に対する近似アルゴリズムも存在する．SAH [8] は，Simpfer と同様にクエリアイテムが各ユーザの k-MIPS 結果に入るかどうかを高速に判定するアルゴリズムである．しかし，線形スキャンの代わりに SAT と呼ばれる歪み誤差が小さい新しいベクトル変換を行い，最大内積探索問題をを最近傍探索問題へと変換して LSH を用いて高速に解く手法である．また，Cone-Tree を使ってユーザ同士が空間的に近くなるようにいくつかのサブセットに分割し，サブセットの重心ベクトルを用いたタイトな内積上限を用いることで k-MIPS 結果に含まれるかどうかの判定処理の効果を高めている．しかし，近似 k-MIPS と同様に，SAH では影響力スコアを厳密に計算できないため，本論文ではこの近似アプローチは考えない．

4 提案アルゴリズム

本章では，逆 k-MIPS に基づく影響力スコアの大きい top-N アイテム検索問題を高速に解くための提案アルゴリズムについて説明する．まず，提案アルゴリズムのメインアイデアについて説明し，次に，提案アルゴリズムのオフライン（前処理）フェーズとオンライン（クエリ）フェーズを説明する．

4.1 メインアイデア

本問題の主なボトルネックは，アイテムの影響力スコアの計算であり，これは最悪の場合に $O(nmd)$ 時間必要であ．そのため，提案アルゴリズムは，各アイテムの影響力スコアの上界を求め，上界が暫定の N 番目に大きいスコアを上回るアイテムにのみ逆 k-MIPS を解くことにより，無駄なアイテムの影響力スコアの計算をフィルタリングする．ここで，無駄なアイテムの影響力スコアの計算をできるだけフィルタリングするためには，影響力スコアの上界をタイトにする必要がある．また，大量のデータに対応するため，影響力スコアの上界の計算も効率的に（ユーザ数およびアイテム数に線形の時間で）計算する必要がある．

影響力スコアの上限の計算として，各ユーザについて，ノル

ムの大きい $O(k_{max}) = O(1)$ 個のアイテムベクトルをスキャンし，暫定（近似）の k-MIPS 結果を格納した配列を得る．ここで，ユーザ $\mathbf{u}_i$ の暫定の k-MIPS 結果の配列を L_i とし，その中で k 番目に大きい内積値を L_i^k とする．このとき，L_i^k から $\mathbf{u}_i$ の k-MIPS 結果に入り得るアイテムと入り得ないアイテムに分類でき，入り得るアイテム（つまり，このアイテムをクエリとした逆 k-MIPS 結果に $\mathbf{u}_i$ が入り得る場合）のみに影響力スコアの上界を $+1$ することによって各アイテムの影響力スコアの上界を計算する．具体的には，k-MIPS 結果に入り得るか入り得ないかの判定にコーシーシュワルツの不等式

$$\mathbf{p} \cdot \mathbf{u} \leqq \|\mathbf{p}\|\|\mathbf{u}\| \tag{1}$$

を利用する．$O(k_{max})$ 個のアイテムをスキャンして得られた暫定の k 番目の内積 L_i^k を用いて，$\|\mathbf{p}\|\|\mathbf{u}\| < L_i^k$ となるアイテムはフィルタリングできる．一方，暫定の k-MIPS 結果に含まれるアイテムと未スキャンで $\|\mathbf{p}\|\|\mathbf{u}\| \geqq L_i^k$ となるアイテムに対して影響力スコアの上界を $+1$ する．オフラインフェーズでは，k_{max}-MIPS 番目の内積値 L_i^{kmax} を用いてフィルタリングを行い，$[1, k_{max}]$ の各 k について影響力スコアの上界値を計算する．

オンライン（クエリ）フェーズでは，与えられた k から各アイテムの影響力スコアの上界を抽出する．暫定の上位 N 番目の影響力スコアを τ としたとき，あるアイテムの影響力スコアの上界が τ 以下であれば，このアイテムの影響力スコアを正確に計算することなくフィルタリングできる．提案アルゴリズムはこのアプローチによってフィルタリングできなかったアイテムに対してのみ影響力スコアを計算することでオンライン時間を短縮する．

4.2 オフラインフェーズ

提案アルゴリズムのデータ構造を構築し，各アイテムのスコア上限を計算するオフラインフェーズについて説明する．なお，このオフライン処理は一度しか行われず，構築されたデータ構造は任意の $k \in [1, k_{max}]$ に適用できる．

オフラインでは，前節で述べたアプローチで各アイテムの影響力スコアの上界を計算するが，コーシーシュワルツの不等式の適応を 2 つのアプローチで最適化する．1 つ目は既存技術の Incremental Pruning [9, 13] を利用する．2 つ目は，各ユーザの k-MIPS をコーシーシュワルツの不等式を利用しながら計算する際，k-MIPS 結果が確定するまでに必要なアイテム数は異なることが分かる．（これは，ユーザのベクトルおよびノルムがそれぞれ異なることから自明である．）そのため，$O(nk_{max})$ 時間使える（つまり，$O(nk_{max})$ 回の内積計算を行える）という制約があるとき，各ユーザに割り当てるアイテム数を動的にすることによって，より多くのユーザの k-MIPS 結果を厳密解に近づける．これは，既存研究において用いられていないアイデアである．以降では，これらについて詳しく説明するが，$\mathbf{P}$（アイテムベクトルの集合）内のベクトルはノルムの降順でソートされていることを前提とする．

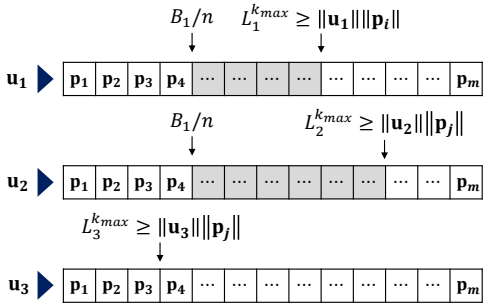


図 1: ユーザベクトル $\mathbf{u}_1$，$\mathbf{u}_1$，および $\mathbf{u}_3$ の k_{max}-MIPS 結果の確定し辛さの例．最大 B_1/n 個のアイテム（$\mathbf{p}_1$ から $\mathbf{p}_4$）にアクセスして暫定的な k_{max}-MIPS 結果を計算したとき，灰色の部分が k_{max}-MIPS 結果を確定するまでに残り必要なアイテムベクトルとなる．例えば，$k_{max} = 3$ で $\mathbf{u}_3$ は B_1/n 未満のアイテムベクトルで k_{max}-MIPS 結果を確定しているが，それ以外は k_{max}-MIPS 結果を確定するためにさらに追加（かつ違う数の）のアイテムベクトルにアクセスする必要がある．

4.2.1 Incremental Pruning

Incremental Pruning では，アイテムベクトルとユーザベクトルを特定の次元で分割する．分割する次元を $\omega \in [0, d]$ とし，$\mathbf{p} = \mathbf{p}^l\mathbf{p}^r$ とする．ここで，$\mathbf{p}^l = (p_1, ..., p_\omega)$ および $\mathbf{p}^r = (p_{\omega+1}, ..., p_d)$ である（$\mathbf{u}$ についても部分ベクトルを同様に定義する．）このとき，$\mathbf{p}$ と $\mathbf{u}$ の内積の上限は，

$$\mathbf{p} \cdot \mathbf{u} \leqq \mathbf{p}^l \cdot \mathbf{u}^l + \|\mathbf{p}^r\|\|\mathbf{u}^r\| \tag{2}$$

と計算できる．この内積上限が暫定の k 番目の内積を下回るアイテムについてはフィルタリングする．また，$\omega = O(1)$ とすることで，不等式（2）は定数時間で評価できる．つまり，各アイテムについて $\mathbf{u}$ の k-MIPS 結果に入る可能性があるかを $O(1)$ 時間で判定できる．

4.2.2 アイテムの動的な割当

影響力スコアの上限を計算するために一部のアイテムをスキャンするとき，予算（全ユーザに対する内積計算可能なアイテムの総数）を $O(nk_{max})$ とする．ここで，予算を各ユーザに割り当てる際，最も単純な方法は各ユーザに均等に（つまり $O(k_{max})$ 個を）割り当てるものである．しかし，先述のように k-MIPS 結果が厳密と確定するまでにアクセスする必要があるアイテム数はユーザごとに異なる．図 1 はそれを直感的に説明した例である．そのため，各ユーザそれぞれに適切な予算を割り当てることが重要となる．

今，予算を $B = B_1 + B_2$ とする．まず，各ユーザに B_1/n 個のアイテムを割り当てる（つまり，$\mathbf{P}$ の先頭から B_1/n 番目までのアイテムにアクセスできる）．このとき，コーシーシュワルツの不等式（不等式（1））から，k-MIPS 結果が確定しないユーザが存在し得ることが分かる．あるユーザの k-MIPS 結果が確定しない場合，コーシーシュワルツの不等式を満たすために必要な残りのアイテム数が分かる．この数は，そのユーザの k-MIPS 結果の「確定し辛さ」を表しており，この数によってユーザを順位付けして予算を割り当てる．

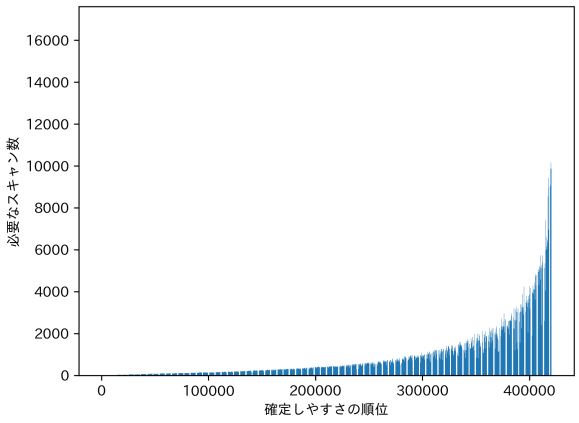


図 2: k-MIPS 結果の確定し辛さと必要なアイテムアクセス数の関係

ここで，どのような関数でスキャン可能予算を割り当てるのかについて考える．図 2 は，Netflix のデータセットにおいて，各ユーザが B_1/n 個のアイテムをスキャンした後，確定していないユーザについて残り何個のアイテムにアクセスすれば，k-MIPS 結果が確定するかを表している．(実験で用いた他のデータセットも同様の形状を示している.) この図から，指数関数を用いて予算を割り当てることを考える．そのため，k-MIPS 結果の「確定しやすさ」の順位 x を変数とした指数関数

$$f(x) = \alpha \exp(\beta x) + \gamma \tag{3}$$

(α，β，および γ はパラメータ) に従って予算を効果的に割り当てる．今，α および γ を定数とし，x_{max} を確定していないユーザ数とする．そして，

$$\int_0^{x_{max}} f(x)dx = \frac{\alpha(\exp(\beta x_{max}) - 1)}{\beta} + \gamma x_{max} \approx B_2 \tag{4}$$

となるように β を決定する．これによって求まった関数に従って，k-MIPS 結果が確定していないユーザについて再度予算を割り当て，k-MIPS を行いながら影響力スコアの上界を計算する．

4.2.3 オフラインアルゴリズム

上記の技術を用いて，各アイテムの各 $k \in [1, k_{max}]$ に対する影響力スコアの上界をオフラインで計算する．オフラインアルゴリズムは以下の 5 つのステップで構成されている．

(1) ノルムの計算． アイテム集合 $\mathbf{P}$ の各アイテムベクトル $\mathbf{p}$ およびユーザ集合 $\mathbf{U}$ の各ユーザベクトル $\mathbf{u}$ にそれぞれついて，ノルム ($\|\mathbf{p}\|$ および $\|\mathbf{u}\|$) を計算する．また，コーシーシュワルツの不等式によるアーリストップを可能とするため，アイテムベクトルをノルムの降順にソートする．説明に都合上，$\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m\}$ で，各 $i \in [1, m-1]$ について $\|p_i\| \geqq \|p_{i+1}\|$ であるものとする．あるユーザベクトル $\mathbf{u}$ に対して $\mathbf{P}$ をスキャンしている場合，暫定の k_{max}-MIPS 結果の閾値を θ としたとき，コーシーシュワルツの不等式から $\theta \geqq \|\mathbf{u}\|\|\mathbf{p}_i\|$ であれば，$\mathbf{p}_i$ 以降のアイテムが $\mathbf{u}$ の k_{max}-MIPS 結果に含まれることはないため，$\mathbf{u}$ の k_{max}-MIPS 結果を確定できる．

(2) 特異値分解． $\mathbf{P}$ および $\mathbf{U}$ は行列で表現できるため，特異値分解 (singular value decomposition) を適用でき，SVD 変換したアイテムベクトルとユーザベクトルを得る．変換後の各アイテムおよびユーザベクトルについて，部分ノルム $\|\mathbf{p}^r\|$ および $\|\mathbf{u}^r\|$ を計算する．SVD 変換は次元のグローバルな再オーダーに相当し [9]，前半の次元が後半の次元よりも重要な次元となり，$\mathbf{p}^l \cdot \mathbf{u}^1$ が $\mathbf{p} \cdot \mathbf{u}$ に近くなるため，不等式 (2) における内積の上限がタイトになる．

(3) アイテム数の均等割当． まず，各ユーザに予算 B_1/n を割り当てる．各ユーザ $\mathbf{u}_i \in \mathbf{U}$ は，$\mathbf{P}$ の先頭から最大 B_1/n 個までのアイテムベクトルを Incremental Pruning を用いながらスキャンし，暫定の k_{max}-MIPS 結果を保持する配列 L_i を計算する．(スキャンの間に $L_i^{k_{max}} \geqq \|\mathbf{u}_i\|\|\mathbf{p}\|$ となれば k_{max}-MIPS 結果が確定するため，スキャンを中止する.) これを全てのユーザに対して行い，k_{max}-MIPS 結果が確定していないユーザについて，二分探索を用いて k_{max}-MIPS 結果を確定させるために必要なアイテム数を推定し，この数の昇順にユーザベクトルをソートする．

(4) アイテム数の動的割当． 式 (4) を用いて，上のステップで残ったユーザベクトルに予算を動的に割り当てる．あるユーザベクトル $\mathbf{u}_i$ が割り当てられた予算を B_2^i としたとき，$\mathbf{P}$ の $B_1/n + 1$ 番目のアイテムベクトルから最大 B_2^i 個のアイテムベクトルにアクセスし，上のステップと同様の方法で暫定の k_{max}-MIPS 結果の配列を更新する．ここで，B_2^i 個のアイテムにアクセスすることなく k_{max}-MIPS 結果が確定した場合，未使用分の予算を次のユーザの予算に加算する．一方，このステップにおいても k_{max}-MIPS 結果が確定しなかったユーザについては，どのアイテムまでスキャンしたのかを保持しておく．この操作をステップ (3) で残った全てのユーザに対して行う．

(5) 影響力スコアの上界の計算． 上のステップで k_{max}-MIPS 結果が確定していない各ユーザ $\mathbf{u}_i$ に対して，上までのステップにおいて未アクセスのアイテムベクトル $\mathbf{p}$ を考える．各 $k \in [1, k_{max}]$ について

$$\mathbf{p}^l \cdot \mathbf{u}_i^l + \|\mathbf{p}^r\|\|\mathbf{u}_i^r\| \geqq L_i^k$$

となる $\mathbf{p}$ は $\mathbf{u}_i$ の k-MIPS 結果に入り得る．そのため，k における $\mathbf{p}$ の影響力スコア上限を $+1$ する．ステップ (4) で記録したアクセス済みの位置から未アクセスの位置を把握できるため，この位置からスキャンを再開し，$L_i^{k_{max}} \geqq \|\mathbf{p}\|\|\mathbf{u}_i\|$ となるか全てのアイテムベクトルにアクセスするまで上の操作を行う．また，$\mathbf{u}_i$ との内積を計算していないアイテムベクトル $\mathbf{p}$ の集合 $\mathbf{P}'$ としたとき，

$$\lambda_i = \max_{\mathbf{p} \in \mathbf{P}'} (\mathbf{p}^l \cdot \mathbf{u}_i^l + \|\mathbf{p}^r\|\|\mathbf{u}_i^r\|) \tag{5}$$

を記録しておく．これは，$\mathbf{u}_i$ と各 $\mathbf{p} \in \mathbf{P}'$ との内積の上界である．

ここで，この処理では内積計算を行っていないため，L_i の更

新は行っていないことに注意する．また，この処理は k-MIPS 結果に入り「得る」かどうかを判定のに定数時間のみ要しているため，予算の範囲外とする（オフラインフェーズの予算は内積計算回数に相当する）．最後に，各 $k \in [1, k_{max}]$ について，各ユーザの（近似）k-MIPS 結果に含まれるアイテムの影響力スコアの上界を $+1$ する．

オフラインアルゴリズムの擬似コードを Algorithm 1 に示す．

4.2.4 空間計算量および時間計算量

このフェーズでは，$\mathbf{P}$ および $\mathbf{U}$ に加えて，各ユーザに対して k_{max}-MIPS 結果および式（5）の値を保持する．また，各アイテムに対して各 $k \in [1, k_{max}]$ での影響力スコアの上界を保持する．ここで，$k_{max} = O(1)$ であるため，これらの空間計算量はそれぞれ $O(|\mathbf{U}|) = O(n)$ および $O(|\mathbf{P}|) = O(m)$ である．つまり，空間計算量は $O((n+m)d)$ である．

次に，時間計算量について分析する．ステップ 1 において，ノルムの計算には $O((n+m)d)$ 時間要し，$\mathbf{P}$ のソートに $O(m \log m)$ 時間要する．ステップ 2 における $\mathbf{P}$ および $\mathbf{U}$ の SVD 変換には $O((n+m)d^2)$ 時間要する．$B = O(nk_{max})$ としたとき，ステップ 3 は $O(n(k_{max}d + \log m + \log n))$ 時間要する（$O(\log m)$ 時間は 15 行を二分探索で行うコストである）．また，式（4）は定数時間で解けるため，ステップ 4 は $O(nk_{max}d)$ 時間要する．最後に，ステップ 5 では（実践的には起こらないが）最悪全てのユーザベクトルの k_{max}-MIPS 結果が確定しない場合があるため，$O(nm)$ 時間要する．ここで，$k_{max} = O(1)$ および，$d \ll n, m$ であるため，オフラインアルゴリズムの時間計算量は $O(nm)$ となる．この結果には d の要素が含まれておらず，最悪の場合でも内積計算を行わずに各アイテムの影響力スコアの上界を得られることを理論的に意味している．

4.3 オンラインフェーズ

最後に，k および N が与えられたとき，オフラインフェーズで構築したデータ構造を用いて，高速に影響力スコアが最も大きい N 個のアイテムを検索するアルゴリズムについて説明する．Algorithm 2 はオンラインアルゴリズムを示している．

（1）影響力スコアの初期化． オフラインフェーズで k-MIPS 結果が確定している各ユーザに対して，k-MIPS 結果に含まれるアイテムの影響力スコアを $+1$ する．また，オフラインフェーズで k-MIPS 結果が確定していないユーザの集合を $\mathbf{Q}_k$ とする．

（2）影響力スコアおよび結果の更新． 次に，影響力スコア上界 $uscore_k(\cdot)$ の降順にアイテムをソートし，この順で逆 k-MIPS を計算する．また，検索結果（N 個のアイテム）の集合を R とし，閾値 τ の初期値を 0 とする．

ここでは，$\mathbf{p}_j \in \mathbf{P}$ が $\mathbf{u}_i \in \mathbf{Q}_k$ の k-MIPS 結果に入るかを判定していく．まず，$\mathbf{p}_j$ が $\mathbf{u}_i$ の未アクセスもしくは暫定の k-MIPS 結果に含まれるかを判定する．これは，Algorithm 1 の 32 行目で記録した情報から定数時間で判定できる．アクセス済で暫定の k-MIPS 結果に含まれない場合，k-MIPS 結果に含まれないことが確定する．一方，暫定の k 番目の内積 L_i^k よりも $\|\mathbf{u}_i\|\|\mathbf{p}_j\|$ が大きい場合，$\mathbf{p}_j^l \cdot \mathbf{u}_i^l + \|\mathbf{p}_j^r\|\|\mathbf{u}_i^r\|$ を計算する．

Algorithm 1: PRE-PROCESSING

Input: $\mathbf{U}$, $\mathbf{P}$, and k_{max}

1 **foreach** $\mathbf{u}_i \in \mathbf{U}$ and $\mathbf{p}_j \in \mathbf{P}$ **do** Compute $\|\mathbf{u}_i\|$ and $\|\mathbf{p}_j\|$
2 Sort $\mathbf{P}$ in descending order of norm
3 SVD transformation of $\mathbf{U}$ and $\mathbf{P}$, $\mathbf{U}' \leftarrow \varnothing$
4 **foreach** $\mathbf{u}_i \in \mathbf{U}$ and $\mathbf{p}_j \in \mathbf{P}$ **do** Compute $\|\mathbf{u}_i^r\|$ and $\|\mathbf{p}_j^r\|$
5 **foreach** $\mathbf{u}_i \in \mathbf{U}$ **do**
6 $\theta \leftarrow 0$ ▷ threshold initialization
7 **foreach** $j \in [1, B_1/n]$ **do**
8 **if** $|\|\mathbf{u}_i\|\|\mathbf{p}_j\| > \theta$ **then**
9 **if** $\mathbf{u}_i^l \cdot \mathbf{p}_j^l + \|\mathbf{u}_i^r\|\|\mathbf{p}_j^r\| > \theta$ **then**
10 Compute $\mathbf{u}_i \cdot \mathbf{p}_j$
11 Update A_i and θ if necessary
12 **else**
13 **break**
14 **if** $\theta < \|\mathbf{u}_i\|\|\mathbf{p}_j\|$ **then**
15 Compute j' such that $\theta \geqq \|\mathbf{u}_i\|\|\mathbf{p}_{j'}\|$
16 Add $\langle \mathbf{u}_i, j' + 1 - B_1/n \rangle$ into $\mathbf{U}'$
17 Sort $\mathbf{U}'$ in ascending order of $j + 1 - \frac{B_1}{n}$
18 Determine β in Equation (3) by Equation (4), $B' \leftarrow 0$
19 $uscore_k(\mathbf{p}_i) \leftarrow 0$ for each $\mathbf{p}_j \in \mathbf{P}$ and $k \in [1, k_{max}]$
20 **foreach** $\mathbf{u}_i \in \mathbf{U}'$ **do**
21 $\theta \leftarrow A_i^{k_{max}}$
22 $B_2^i \leftarrow$ the budget determined by Equation (3)
23 **foreach** $j \in [\frac{B_1}{n} + 1, \frac{B_1}{n} + 1 + B_2^i + B']$ **do**
24 **if** $\|\mathbf{u}_i\|\|\mathbf{p}_j\| > \theta$ **then**
25 **if** $\mathbf{u}_i^l \cdot \mathbf{p}_j^l + \|\mathbf{u}_i^r\|\|\mathbf{p}_j^r\| > \theta$ **then**
26 Compute $\mathbf{u}_i \cdot \mathbf{p}_j$
27 Update A_i and θ if necessary
28 **else**
29 $B' \leftarrow \frac{B_1}{n} + 1 + B_2^i + B' - j$
30 **break**
31 **if** $\theta < \|\mathbf{u}_i\|\|\mathbf{p}_j\|$ **then**
32 Record $j+1$, $\lambda_i^* \leftarrow 0$
33 **foreach** $j' \in [j+1, m]$ **do**
34 **if** $\theta \geqq \|\mathbf{u}_i\|\|\mathbf{p}_{j'}\|$ **then**
35 **break**
36 $\lambda \leftarrow \mathbf{u}_i^l \cdot \mathbf{p}_j^l + \|\mathbf{u}_i^r\|\|\mathbf{p}_j^r\|$
37 $\lambda_i^* \leftarrow \max\{\lambda_i^*, \lambda\}$ ▷ Equation (5)
38 **foreach** $k \in [1, k_{max}]$ **do**
39 **if** $\lambda \geqq A_i^k$ **then**
40 $uscore_k(\mathbf{p}_i) \leftarrow uscore_k(\mathbf{p}_i) + 1$
41 **foreach** $i \in [1, n]$ **do**
42 **foreach** $k \in [1, k_{max}]$ **do**
43 **if** $\mathbf{p}_j$ *is included in the first* k *elements of* A_i **then**
44 $uscore_k(\mathbf{p}_j) \leftarrow uscore_k(\mathbf{p}_j) + 1$

これが L_i^k よりも大きいとき，$\mathbf{p}_j^r \cdot \mathbf{u}_i^r$ を計算し，$\mathbf{p}_j \cdot \mathbf{u}_i$ を得る．この内積が L_i^k よりも大きいとき，式（5）で得られた未スキャン部分の内積の上限よりも $\mathbf{p}_j \cdot \mathbf{u}_i$ が大きいかを判定する．

Algorithm 2: オンラインアルゴリズム

Input: $\mathbf{U}$, $\mathbf{P}$, k, and N

$R \leftarrow \varnothing$, $\tau \leftarrow 0$, $\mathbf{X} \leftarrow \varnothing$
foreach $\mathbf{u}_i \in \mathbf{U}$ **do**
  if $A_i^k \geqq \lambda_i$ **then**
    foreach $\mathbf{p}_j$ *in the first* k *elements of* A_i **do**
      $score_k(\mathbf{p}_j) \leftarrow score_k(\mathbf{p}_j) + 1$
  else
    Add $\mathbf{u}_i$ into $\mathbf{X}$
Sort $\mathbf{P}$ in descending order of $uscore_k(\cdot)$
foreach $\mathbf{p}_j \in \mathbf{P}$ **do**
  if $uscore_k(\mathbf{p}_j) > \tau$ **then**
    $score_k(\mathbf{p}_j) \leftarrow score_k(\mathbf{p}_j)$+ RMIPS$(\mathbf{p}_j, \mathbf{X}, \mathbf{P}, k)$
    Update R and τ if necessary
  else
    return R

内積の方が大きい場合，$\mathbf{p}_j$ は $\mathbf{u}_i$ の k-MIPS 結果に含まれることが確定するため，$\mathbf{p}_j$ の影響力スコアを $+1$ する．内積の方が小さい場合，未スキャン部分から Incremental Pruning を用いて線形スキャンを行い，$\mathbf{p}_j$ が $\mathbf{u}_i$ の k-MIPS 結果に含まれるかどうかを検査する．オフラインアルゴリズムと同様に，どこまでスキャンしたかも記録する．これにより k-MIPS 結果に含まれることが確定した場合，$\mathbf{p}_j$ のスコアを $+1$ する．

上の操作を最初の N 個のアイテムベクトルについて行うと，暫定の結果 R および τ を得る．ここで，$N+1$ 番目のアイテム $\mathbf{p}'$ の $uscore_k(\mathbf{p}')$ と τ を比較する．このとき，$uscore_k(\mathbf{p}') \leqq \tau$ であれば，以降のアイテムベクトルが検索結果に含まれることはないことがソートオーダーから保証される．そのため，以降のアイテムベクトルについて逆 k-MIPS を行うことなく厳密解を確定できる．一方，$uscore_k(\mathbf{p}') > \tau$ であれば，上と同様の方法で影響力スコアを計算し，R および τ を更新する．この操作を R が確定するまで繰り返す．

5 評価実験

評価実験の結果を本章で示す．全ての実験は Ubuntu 22.04.3 LTS，2.8GHz Core i9-10900 CPU，および 24GB の RAM を搭載した計算機上で行った．

5.1 設　　定

データセット．本実験では，以下の 4 つの実世界データセットを用いた．各データセットのユーザ数およびアイテム数を表 1 に示す．

- Netflix[1]: Netflix Prize で用いられたレーティングのデータセット．
- Movielens[2]: Movielen 25M データセット．

1：https://www.cs.uic.edu/liub/Netflix-KDD-Cup-2007.html
2：https://grouplens.org/datasets/movielens/

表 1: データセットの統計

	ユーザ数	アイテム数
Netflix	480,189	17,770
Movielens	162,541	59,047
AmazonKindle	1,406,890	430,530
AmazonMovie	2,088,620	200,941

- AmazonKindle [5]: AmazonKindle の書籍のレーティングのデータセット．
- AmazonMovie [5]: AmazonMovie の映画のレーティングのデータセット．

ユーザベクトルとアイテムベクトルを生成するために，[2] の Matrix Factrization を実行した．各データセットのデータベクトルの次元数は 200 とした．

比較手法．本実験では，以下の 3 つの比較手法を用いた．

- FEXIPRO [9]: 厳密解を出力する k-MIPS アルゴリズム FEXIPRO を用いて，全てのユーザの完全な k-MIPS 結果を計算し，全てのアイテムの影響力スコアを計算するアルゴリズム．
- LEMP [13]: FEXIPRO と同様の処理を LEMP に置換したアルゴリズム．
- Uniform: 提案アルゴリズムにおいて，予算の動的割り当てを行わず，全ユーザに予算を均一に割り当てるアルゴリズム．

全てのアルゴリズムは C++で実装を行い，g++ 11.4.0 に最適化オプション-O3 付加してコンパイルした．提案アルゴリズムにおけるベクトルを分割する次元 $\omega = 10$ と設定した．評価指標として実行時間 (msec) を用いた．

5.2 結果と分析

k の影響．図 3 に $N = 10$ で固定し，k を 5，10，15，および 20 に変化させたときの提案アルゴリズムと比較手法の実行時間の結果を示す．全てのデータセットにおいて，提案アルゴリズムが最も高速であることが分かる．FEXIPRO と LEMP は，k を大きくするほど，k-MIPS 結果を計算するために多くのアイテムにアクセスする必要があるため低速となる．提案アルゴリズムや比較手法は，アイテムが k-MIPS 結果に入りうるか高速に判定するアルゴリズムであるが，入るか入らないか確定しない場合には，線形スキャンを行うため，同様に k が増加するほど低速となる．また，Uniform は提案アルゴリズムに比べて大幅に低速なことから，提案手法による動的な予算割当によって多くのアイテムのスコアの上界がタイトになっていることが分かる．

N の影響．図 4 に $k = 10$ で固定し，N を 20，40，60，80，および 100 に変化させたときの提案アルゴリズムと比較手法の実行時間の結果を示す．FEXIPRO や LEMP は，全てのアイテムの影響力スコアを計算するため，N が変化しても実行時間は変わらない．提案アルゴリズムと Uniform は N を増加するほどより多くのアイテムについて逆 k-MIPS を解く必要があるた

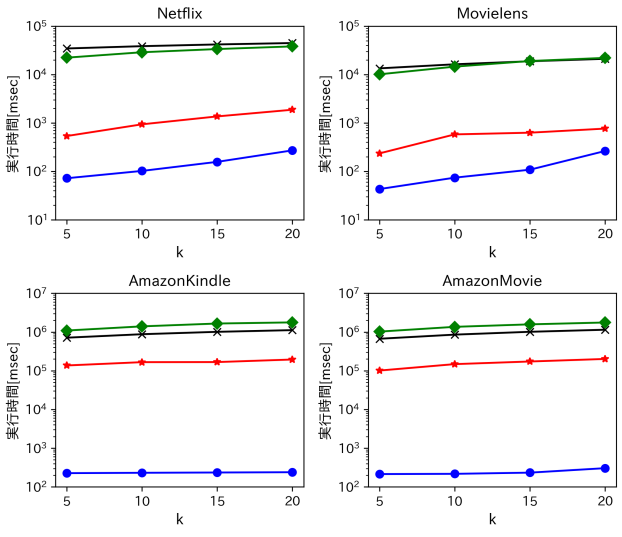


図 3: k の影響．黒が FEXIPRO，緑が LEMP，赤が Uniform，および青が提案アルゴリズムの結果を示す．

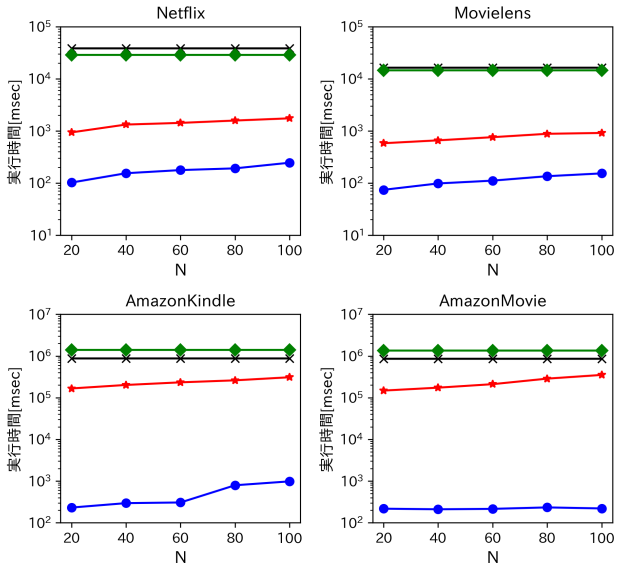


図 4: N の影響．黒が FEXIPRO，緑が LEMP，赤が Uniform，および青が提案アルゴリズムの結果を示す．

め低速となる．

6 結　　論

本論文では，逆 k-MIPS に基づく影響力の大きいアイテムの検索問題を定義した．この問題では，アイテムの影響力スコアをそのアイテムを k-MIPS 結果に含むユーザのカーディナリティとして，影響力スコアの大きい N 個のアイテムを検索する．この問題を既存の k-MIPS 手法や逆 k-MIPS 手法で解く場合，は全ユーザの k-MIPS 結果を計算するため，大きな計算コストがかかる．我々は前処理において $O(nm)$ 時間で各アイテムのスコアの上界値を計算し，オンラインでは，前処理で計算した上界値を用いて不必要な逆 k-MIPS の計算を枝刈りするアルゴリズムを提案した．また，実世界データセットを用いた実験を行い，我々のアルゴリズムが比較手法を大幅に上回る性能を持つことを確認した．

謝辞．本研究の一部は，AIP 加速課題（JPMJCR23U2）の支援を受けたものである．

文　　献

[1] D. Amagata and T. Hara, “Reverse maximum inner product search: How to efficiently find users who would like to buy my item?,” In RecSys, pp.273–281, 2021.
[2] W.S. Chin, B.W. Yuan, M.Y. Yang, Y. Zhuang, Y.C. Juan, and C.J. Lin, “Libmf: A library for parallel matrix factorization in shared-memory systems.,” Journal Machine Learning Research, vol.17, no.1, pp.2971–2975, 2016.
[3] X. Dai, X. Yan, K.K. Ng, J. Liu, and J. Cheng, “Norm-explicit quantization: Improving vector quantization for maximum inner product search,” In AAAI, vol.34, no.01, pp.51–58, 2020.
[4] R. Guo, P. Sun, E. Lindgren, Q. Geng, D. Simcha, F. Chern, and S. Kumar, “Accelerating large-scale inference with anisotropic vector quantization,” In ICML, pp.3887–3896, 2020.
[5] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” In World Wide Web, pp.507–517, 2016.
[6] K. Hirata, D. Amagata, S. Fujita, and T. Hara, “Solving diversity-aware maximum inner product search efficiently and effectively,” In RecSys, pp.198–207, 2022.
[7] Q. Huang, G. Ma, J. Feng, Q. Fang, and A.K. Tung, “Accurate and fast asymmetric locality-sensitive hashing scheme for maximum inner product search,” In SIGKDD, pp.1561–1570, 2018.
[8] Q. Huang, Y. Wang, and A.K. Tung, “Sah: shifting-aware asymmetric hashing for reverse k maximum inner product search,” In AAAI, vol.37, no.4, pp.4312–4321, 2023.
[9] H. Li, T.N. Chan, M.L. Yiu, and N. Mamoulis, “Fexipro: fast and exact inner product retrieval in recommender systems,” In SIGMOD, pp.835–850, 2017.
[10] S. Morozov and A. Babenko, “Non-metric similarity graphs for maximum inner product search,” In NeurIPS, vol.31, 2018.
[11] P. Ram and A.G. Gray, “Maximum inner-product search using cone trees,” In SIGKDD, pp.931–939, 2012.
[12] S. Tan, Z. Xu, W. Zhao, H. Fei, Z. Zhou, and P. Li, “Norm adjusted proximity graph for fast inner product retrieval,” In SIGKDD, pp.1552–1560, 2021.
[13] C. Teflioudi, R. Gemulla, and O. Mykytiuk, “Lemp: Fast retrieval of large entries in a matrix product,” In SIGMOD, pp.107–122, 2015.
[14] R. Weber, H.J. Schek, and S. Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” In VLDB, vol.98, pp.194–205, 1998.
[15] L. Xiang, B. Tang, and C. Yang, “Accelerating exact inner product retrieval by cpu-gpu systems,” In SIGIR, pp.1277–1280, 2019.
[16] L. Xiang, X. Yan, L. Lu, and B. Tang, “Gaips: Accelerating maximum inner product search with gpu,” In SIGIR, pp.1920–1924, 2021.
[17] X. Zhao, B. Zheng, X. Yi, X. Luan, C. Xie, X. Zhou, and C.S. Jensen, “Fargo: Fast maximum inner product search via global multi-probing,” PVLDB, vol.16, no.5, pp.1100–1112, 2023.

A Grid-density-based Approach for Dense Nearest Neighborhood Query

Ruolin XIE†, Hanxiong CHEN†, and Kazutaka FURUSE††

† University of Tsukuba,
1–1–1 Tennodai, Tsukuba-shi, Ibaraki, Japan
†† University of Hakuoh,
2–2–2 Ekihigashidori, Oyama-shi, Tochigi, Japan
E-mail: †xieruolin@outlook.com, ††chx@cc.tsukuba.ac.jp, †††furuse@fc.hakuoh.ac.jp

Abstract Efficient spatial queries often face limitations with traditional Nearest Neighborhood Query (NNH) method. This study introduces Grid-Density-Based Dense Nearest Neighborhood Query (GDBDNNH), extending Dense Nearest Neighborhood Query (DNNH) with a density-based perspective. GDBDNNH ensures computational efficiency and precision, demonstrated through experiments on diverse datasets, highlighting its effectiveness in achieving high performance for spatial queries.

Key words Nearest Neighborhood Query, Retrieval Recommendation, Density-based Spatial Exploration

1 Introduction

The widespread adoption of mobile internet technologies and devices in the past decade has fueled the popularity of Location-Based Services (LBS). Major providers like Google Maps, social media platforms such as Instagram and TikTok, and food delivery services like Uber Eats leverage spatial data to offer assistance and service information.

Nearest Neighborhood Query [1] (NNH) is crucial in spatial data processing, playing a fundamental role in applications ranging from location-based services to data mining. While traditional approaches have limitations, Dense Nearest Neighborhood Query [2] (DNNH) has addressed these constraints, particularly excelling in searching dense spatial clusters.

In Figure 1, we illustrate an example of an NNH Query with k=3, seeking nearby convenience stores in a two-dimensional coordinate space. This study introduces Grid-Density-Based Dense Nearest Neighborhood Query (GDBDNNH), an algorithm that significantly enhances spatial query performance. GDBDNNH extends DNNH by incorporating a density-centric perspective, ensuring both density and efficiency. Unlike traditional NNH methods, GDBDNNH allows more flexible and realistic cluster identification by eliminating strict constraints on group size. The proposed method employs innovative grid indexing techniques, resulting in a notable reduction in computational overhead.

The paper's structure is as follows: Section 2 provides an overview of related work, Section 3 outlines our methodology, Section 4 describes the experimental setup and discusses the results, and Section 5 concludes with a summary of this study. This research aims to contribute valuable advancements to the spatial query domain by introducing the complex and efficient method GDBDNNH, enhancing both precision and efficiency.

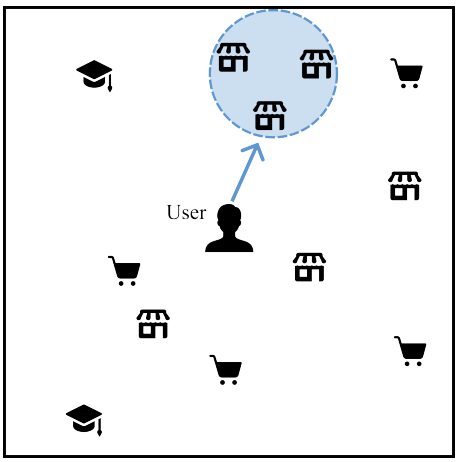
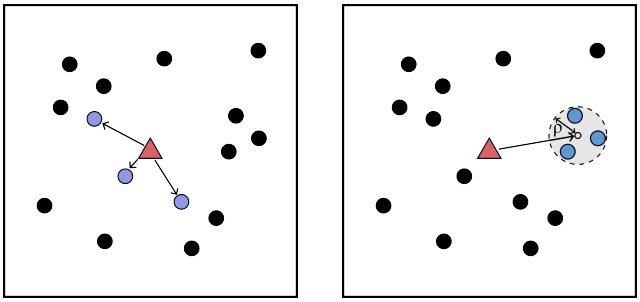


Figure 1: An example of NNH Query. NHH recommends the nearest group of convenience stores with $k = 3$.

2 Related Works

This chapter provides an overview of the Nearest Neighborhood Query problem, a pivotal element in spatial data applications. And this chapter explores existing efforts in refining NNH methodologies, laying the foundation for understanding the motivations and challenges that drive the proposed Grid-Density-Based Nearest Neighborhood Query approach.



(a) NN (b) NNH

Figure 2: Comparison of NN and NNH Algorithms. (a)NN, 3-NN of query point, (b)NNH, Obtaining a set of nearest neighborhood data points with $k = 3$ and a radius ρ.

2.1 Nearest Neighborhood Query

The significance of Nearest Neighbors [3] (NN) in spatial databases underpins many contemporary applications. Roussopoulos et al.'s algorithm for nearest neighbor queries enables precise and efficient spatial searches. Building on this, Choi et al. focus on the nearest neighborhood search, which can be viewed as the group version of nearest neighbors, aiming to identify clusters of points within a specified radius.

Their work introduces the nearest enclosing circle (NEC) problem, not only elevates the efficiency of NNH searches but also finds applications in diverse fields, reflecting the continuous evolution of methods to address real-world challenges. The subsequent sections delve deeper into these advancements, setting the stage for a comprehensive exploration of the nearest neighbor problem and its evolving landscape.

Figure 2 below illustrate the differences between nearest neighbors and NNH, especially when k=3. Nearest neighbors method identifies the three closest points to a query point, yet these points might be widely spaced. In contrast, NNH overcomes this limitation by focusing on point clusters within a circle of specific radius ρ. If the circle contains at least k points, the result is the set of points in the circle closest to the query point, ensuring a more concentrated group of results. However, unsuitable values may cause the result to return an empty set. And it is quite difficult for the user to specify the values of ρ and k in advance.

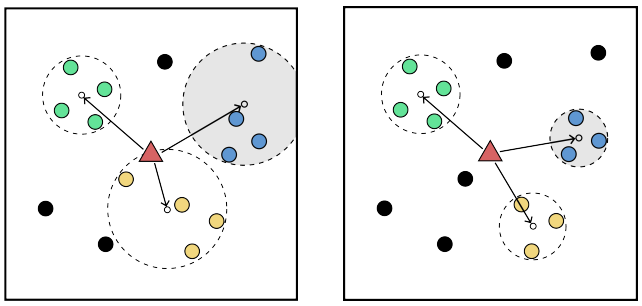
To overcome the potential issue of no results in NNH, Le et al. proposed the Balanced Nearest Neighborhood [4] (BNNH) method. This method ensures that a result set always contains a specified number of data points k, with the size of clusters adapting to k. BNNH allows for variable group sizes in finding the nearest cluster of points, ensuring non-empty results and incorporating user preferences for a more realistic and relevant neighborhood retrieval. However, BNNH does not assure dense clustering, and determining an optimal k value remains a challenge for different datasets.

Following the discussion on the BNNH method, it's pertinent to mention the Semantic Consistency on Hyper Nearest Neighborhood [5] (SCNNH) approach. This method introduces the concept of considering semantic consistency in nearest neighborhood queries. Additionally, other methodologies like Reverse Nearest Neighborhood [6] (RNNH) and Aggregate Nearest Neighborhood [7] (ANNH) also contribute to enhancing and optimizing nearest neighborhood queries. RNNH focuses on reverse queries, identifying points that consider a given point as their nearest neighbor, while ANNH emphasizes aggregate queries among multiple query points, aiming to find groups or areas closest to these point combinations.

2.2 Dense Nearest Neighborhood Query

BNNH enhances the NNH by introducing flexibility in the cluster sizes, although it does not assure dense clustering, and determining an optimal k value remains a challenge for different datasets.

To address the issue of less compact results in BNNH, Suzuki et al. developed several algorithms for Dense Nearest Neighborhood Query. These algorithms effectively address the issue of fixed parameter constraints inherent in BNNH by introducing flexibility in parameter selection and adaptation to the data. This approach allows for more precise tuning to the specific characteristics of different datasets, leading to more compact and accurate results in dense neighborhood groups identification.



(a) BNNH (b) DNNH

Figure 3: Comparison of Clustering Compactness in BNNH and DNNH

As shown in Figure 3a, the BNNH algorithm with $k = 4$ aims to find clusters containing 4 data points. However, the yellow and blue clusters appear to be loose, somewhat contravening the definition of Neighborhood in the NNH context and demonstrating how the parameter k can restrict the algorithm. In contrast, Figure 3b demonstrates how DNNH re-

laxes the parameter constraints of NNH and BNNH, with the nearest neighborhood candidates only selecting tight clusters.

2.3 Density-Based Clustering Methods

While Nearest Neighborhood and Dense Nearest Neighborhood have demonstrated advancements in spatial query efficiency, a significant leap can be achieved by integrating them with state-of-the-art density-based clustering methods.

DBSCAN [8], a cornerstone method, identifies clusters based on the density of points, effectively distinguishing high-density clusters from sparse regions. OPTICS [9] extends this concept by providing an ordered representation of data to recognize clusters of varying densities. DENCLUE [10], with its mathematical approach, defines clusters through density distribution, adept at handling diverse densities. Complementing these, the CFDP [11] method efficiently pinpoints central cluster points. These methodologies collectively provide a comprehensive framework for the study, with DBSCAN's principles laying the groundwork for innovative spatial data analysis techniques.

3 Methodology

3.1 Parameters and Definition

This part elucidates the key parameters and definitions critical to the understanding of the methodology.

The notation and meaning of this paper are shown in Table 1, followed by the definition of the methods.

Parameters	Meaning
P	dataset
p	points in P
q	query point
C	clusters
$\bar{C}$	cluster centroid
d	distance bewteen points
ρ	density of C
ϵ	radius of the neighborhood for clustering
$minPts$	the minimum number of points required
N	a threshold for determining minimum cluster size

Table 1: Symbol Cross-Reference for Proposed Method Parameters

Definition 1. *(Dense Nearest Neighborhood Query, DNNH).* Given a set of points P and a query point q, the DNNH query returns a group C that not only minimizes the value of $\Delta(C, q)$ but also ensures that the number of points in C is not less than a predefined threshold.

The $\Delta(C, q)$ function is a key metric in this study, formulated to assess both the compactness of a cluster and its distance to a query point q. It is mathematically represented as follows:

$$\begin{aligned} d(A, B) &= \|A - B\| \\ \rho(C) &= \frac{1}{|C|^2} \sum_{i,j \in C} d(i, j) \\ \Delta(C, q) &= \frac{d(\bar{C}, q)^2}{\rho(C)} \end{aligned} \tag{1}$$

where $d(\bar{C}, q)$ denotes the squared Euclidean distance from the query point q to the centroid $\bar{C}$ of the cluster. The term $density(C)$ signifies the density of the cluster C, integral to this computation. $density(C)$ can be approached through different methodologies, reflecting the cluster's spatial attributes. Thus, the $\Delta(C, q)$ function provides a comprehensive measure of the cluster's relevance in spatial queries, integrating internal cohesion with external proximity.

$\Delta(C, q)$ is effective but computationally intensive due to its $O(n^2)$ complexity. This complexity stems from the pairwise distance calculations within the cluster. To enhance efficiency, an approximate delta function is proposed, utilizing a simpler density measure: the ratio of the cluster size to its approximate area. This approximation significantly reduces the complexity to $O(n)$, balancing computational efficiency with the effectiveness of the density and proximity measurements. The approximate area A_{approx} of a cluster, important for calculating the $\rho_{\text{approx}}(C)$ in the Δ_{approx} function, is estimated by the formula:

$$\begin{aligned} A_{\text{approx}}(C) =& \min(|x_{\min} - \tilde{x}|, |x_{\max} - \tilde{x}|) \\ & \times \min(|y_{\min} - \tilde{y}|, |y_{\max} - \tilde{y}|) \\ \rho_{\text{approx}}(C) =& \frac{|C|}{A_{\text{approx}}(C)} \end{aligned} \tag{2}$$

where $x_{\min}, x_{\max}$ and $y_{\min}, y_{\max}$ are the minimum and maximum coordinates of the points in the cluster, and $\tilde{x}, \tilde{y}$ are the median values of these coordinates. This method provides an estimated bounding area for the cluster, which is then used in the computation of the Δ_{approx} function:

$$\Delta_{\text{approx}}(C, q) = \frac{d(\bar{C}, q)^2}{\rho_{\text{approx}}(C)} \tag{3}$$

3.2 Adaptive Epsilon for DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely-used density-based clustering algorithm that groups together data points based on their density in the feature space. Unlike traditional clustering methods, DBSCAN categorizes points as core points, border points, or outliers (noise) based on the density of points in their vicinity. The key parameters in DBSCAN include ϵ and `minPts`.

The need for Adaptive Epsilon arises from the inherent challenge of manually specifying a universal ϵ value for di-

verse datasets. Different datasets may exhibit varying density patterns, making it impractical to predefine a single ϵ that accommodates all scenarios. In many cases, setting a fixed ϵ may lead to suboptimal results, particularly when faced with datasets containing clusters of varying densities.

The Adaptive Epsilon method, inspired by the SA-DBSCAN [12] approach, employs a data-driven strategy to determine the optimal ϵ for DBSCAN clustering. It begins by standardizing the input data using the Standard Scaler to ensure consistent scale across features. Subsequently, it utilizes the Nearest Neighbors algorithm to estimate the average distance to the k nearest neighbors for each data point.

The knee of the resulting distance curve, determined through the KneeLocator algorithm, signifies an inflection point where the average distance stabilizes, as shown in the Figure 4. This knee point is considered as the adaptive ϵ value, providing an automated and data-driven means to set the optimal neighborhood distance for DBSCAN.

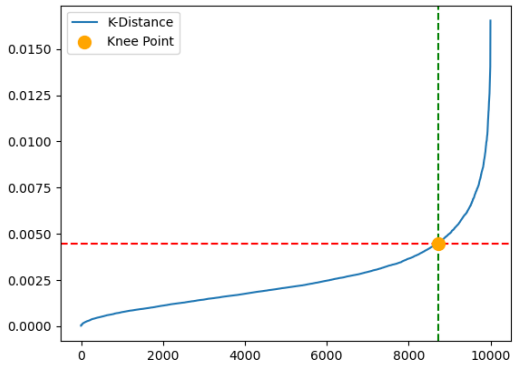


Figure 4: Adaptive ϵ Determination in DBSCAN. The x-axis represents the index of data points, while the y-axis denotes the average distance to their k nearest neighbors, reflecting local density.

The Adaptive Epsilon method contributes to the adaptability of DBSCAN, offering an effective solution for datasets with varying density structures.

3.3 Naïve Algorithm

In the Naïve Algorithm, the dataset undergoes a two-step process. Initially, it is clustered using DBSCAN, resulting in a set of clusters. Subsequently, the algorithm selects the closest circular sub-cluster to the query point q from the clustered dataset. The general search flow is as follows:

Firstly, the dataset is clustered using DBSCAN with specified parameters (ϵ and minPts), which produces a collection of clusters. Then, the sub-clusters derived from the DBSCAN result are sorted, and the one closest to the query point q is chosen as the final result (C_{best}). The Naïve Algorithm is presented below:

Algorithm 1 Naïve Algorithm

Input: $P, minPts$
Output: C_{best}
1 $\epsilon \leftarrow$ adaptive_eps(P)
2 C $\leftarrow$ DBSCAN($\epsilon, minPts, P$)
3 **Return** C_{best} with argmin(ΔC)

Compared to the algorithm based on X-means [13], the Naïve Algorithm offers more flexibility by eliminating constraints on k_{min} and k_{max}. It adapts well to diverse datasets and effectively filters out noise points, resulting in improved outcomes. However, it involves clustering the entire dataset initially and subsequently dividing the obtained clusters into sub-clusters. This process incurs a notable execution time cost. To address this, the following subsection introduces the Expanding Algorithm, which aims to circumvent the need for individual traversal of all data points.

3.4 Expansion Algorithm

To address the issue of the naive algorithm's time complexity being $O(n^2)$ due to traversing all data points, an optimization is introduced known as the Expansion Algorithm. This method aims to enhance query efficiency by utilizing preprocessing steps, dynamically adjusting search boundaries, and harnessing the power of the KD-tree data structure for accelerated searching.

This algorithm starts by organizing the dataset through data preprocessing, which involves sorting data points by their distances from the query point. This step ensures that the search begins in the vicinity of the query point. Subsequently, the algorithm employs the KD-tree data structure to efficiently index and partition data points. The KD-tree facilitates quicker nearest-neighbor searches. The initial search boundary is set to encompass all data points, either with an infinite value or a considerably large one. The algorithm then enters a search loop, iteratively expanding the search area to discover dense clusters. Starting from the data point nearest to the query point, it examines each data point within the current search boundary. If a dense cluster is identified, the search boundary updates to include the newly discovered cluster. The search boundary dynamically expands, incorporating newly discovered clusters until predefined termination conditions are met. These conditions may include reaching the maximum search distance or satisfying specific cluster density or size criteria.

By introducing the Expansion Algorithm and utilizing the KD-tree structure, the query process becomes significantly more efficient. This is achieved by reducing the number of

data points to be examined and leveraging the KD-tree's indexing mechanism for rapid nearest-neighbor identification. The algorithm cleverly adjusts the search boundaries, focusing attention on regions likely to contain dense clusters. This strategic approach substantially reduces time complexity, thereby enhancing the overall performance of the query algorithm.

For the sake of efficiency, visited points are marked as $visited = True$ during traversal. In line 6 in Algorithm 2, the algorithm checks if the current point p has been visited; if not, it is marked as visited. At line 9, the algorithm employs the DBSCAN concept, marking the current point p as a core point if the number of points in its epsilon neighborhood is not less than minPts. The expand_cluster function is then used to find other density-reachable core points, forming a new cluster with these core points and the points within their epsilon neighborhoods. In line 10, the mask is updated. If the current cluster meets certain conditions and has a Δ smaller than the previous Δ_{old}, the bound is updated.

Algorithm 2 Expansion Algorithm

Input: $P, minPts$
Output: C_{best}

1 **Initialization:** $P \leftarrow \text{sort}(P, q)$, $labels \leftarrow$ -1, $bound \leftarrow \infty$, $update \leftarrow$ **False**
2 $\epsilon \leftarrow$ adaptive_eps(P)
3 **while** ***True*** **do**
4 $P \leftarrow$ points in bound
5 **for** p *in* P **do**
6 visited[idx] $\leftarrow$ **True**
7 **if** $len(neighbors) \geqq minPts$ **then**
8 **if** $labels[idx] == -1$ **then**
9 expand_cluster(...)
10 Update variables...
11 labels[$\sim mask$] $\leftarrow$ -1
12 **if** ***not*** $update$ ***or*** $all(labels \neq -1)$ or $minPts \times N > len(P[mask])$ **then**
13 **break**
14 **Return** C_{best} with argmin(ΔC)

In the subsequent section, we will introduce the Grid-Dense-Based Algorithm, a grid-based method designed to further improve query efficiency. This algorithm optimizes performance by reducing the number of data points to be examined.

3.5 Grid-density-based Algorithm

Building upon the foundations of the previous algorithms, the Grid-dense-based (GDB) algorithm introduces additional advancements to significantly enhance its efficiency and scalability. By dividing the data space into cells using a grid-based approach, as illustrated in Figure 5, the GDB algorithm can swiftly identify the cell containing the query point q and efficiently explore its surrounding cells. A critical advantage of this grid-based partitioning is the elimination of unnecessary search in cells that lie outside the specified threshold, substantially reducing computation overhead and focusing the algorithm on potentially relevant solutions.

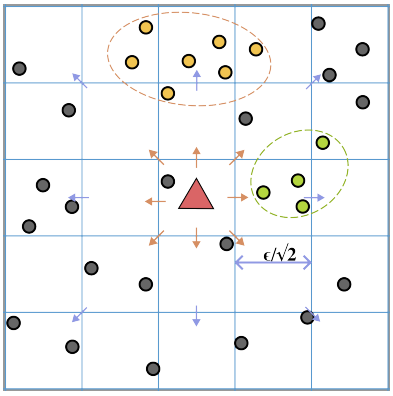


Figure 5: GDB Method, Halting grid expansion with discovered dense neighborhood at $minPts = 4$ and grid size set to $\epsilon/\sqrt{2}$.

The GDB algorithm builds upon existing expansion algorithms and introduces critical improvements with the use of grid-based techniques to enhance its efficiency and scalability. To efficiently handle data points, the data space is divided into grid cells. The sizes of these cells are dynamically determined based on the data distribution and the chosen query point. Grid cells ensure that only adjacent cells need to be considered for clustering, reducing unnecessary computations. The algorithm systematically explores the data space in a hierarchical manner, starting from the grid cell containing the query point q. Each layer expands outward from the query point in a structured way, reducing unnecessary search operations.

In each layer, the algorithm identifies clusters based on the density of data points. It employs density-based clustering methods to group data points according to density criteria. Grid partitioning ensures that only relevant neighboring cells are used for clustering, enhancing efficiency. The algorithm includes a threshold parameter to control the level of detail in the clustering results. By adjusting this threshold, users can obtain clusters of different granularities.

A fundamental aspect of the GDB algorithm is the grid partitioning of the data space. Grid partitioning divides the space into grid cells, with each cell containing a subset of data points. The size and distribution of grid cells are dynamically determined based on the data distribution and the selected query point. First, the epsilon value is calculated corresponding to the dataset. The size of the grid cells is set

as $\frac{\epsilon}{\sqrt{2}}$, ensuring that the diagonal distance of each cell equals the epsilon value. This process ensures that the algorithm can adapt to different data distributions and dimensions, improving its robustness and performance.

The GDB algorithm adopts a hierarchical exploration strategy to efficiently navigate the data space. This strategy begins with the grid cell containing the query point q and expands outward. The hierarchical exploration strategy, coupled with grid-based partitioning, provides the algorithm with high scalability, enabling it to efficiently handle large-scale datasets. Flexibility is another key feature, as users can control the granularity of clustering results by adjusting the threshold parameter, offering flexibility in cluster identification. Empirical evaluations demonstrate that the GDB algorithm outperforms traditional clustering algorithms in terms of running time and cluster quality.

The stopping criterion of the GDB algorithm is defined as follows: When the number of data points in any cluster exceeds the threshold of $N \times$ minPts, the algorithm ceases to expand the grid further. Eventually, it returns the cluster with the minimum Delta value as the best cluster.

Algorithm 3 Grid-density-based Algorithm

Input: $P, minPts, N$

Output: C_{best}

Initialization: $labels \leftarrow$ -1, $\epsilon \leftarrow$ adaptive_eps(P), cell_size $\leftarrow \frac{\epsilon}{\sqrt{2}}$

while *True* **do**

  for *cell in layer* **do**

    neighbors $\leftarrow$ find$_n eighbors$

    if *len(neighbors)* $<$ *minPts* **then**

      labels $\leftarrow -1$

    else

      expand_cluster

  if *any* $|C_i| \geqq minPts \times N$ **then**

    break

Return clusters with argmin(ΔC)

In the following chapter, experimental results and performance evaluations will be explored, including a comparative analysis of the three algorithms, to provide a comprehensive assessment of their capabilities in various scenarios.

4 Experiment

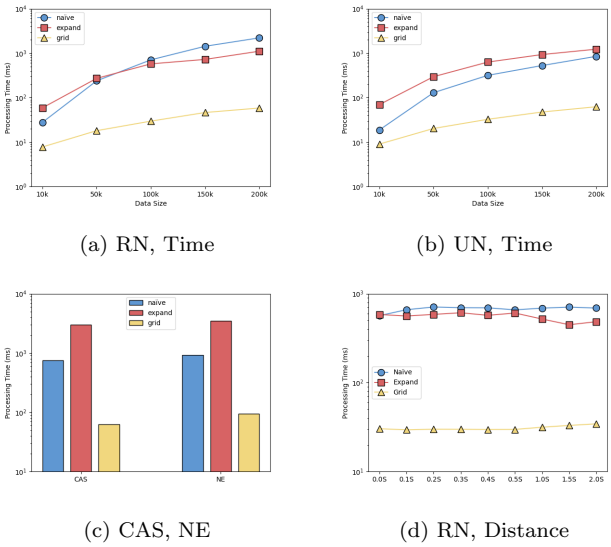
4.1 Setup and Datasets

Before delving into the results, details of the experimental setup is provided. Three algorithms are evaluated based on their processing time, precision, recall, and F1 score. Each experiment was repeated 100 times to obtain average results, and the experiments were conducted on a system equipped with an Apple M3 Pro chip, 18GB LPDDR5 memory, running MacOS 14.1.1.

Experiments are conducted using two types of datasets: real-world and synthetic. The real-world datasets, NE (123,593 data points) and CAS (196,902 data points), were provided by the U.S. Census Bureau's TIGER project. The synthetic datasets consist of uniform random data with sizes ranging from 10,000 to 200,000 points. These datasets consist of the following types: Uniform Random Data (UN): Generated as uniform random data; Random Normal Data (RN):Compounded datasets where points follow a standard normal distribution and are scaled and uniformly arranged in space as clusters.

4.2 Results and Analysis

In the experimental evaluation, Naive, Expand, and Grid algorithms are compared across various datasets, including real-world and synthetic datasets (UN and RN).



(a) RN, Time (b) UN, Time

(c) CAS, NE (d) RN, Distance

Figure 6: Impact of Diffrent Datasets and Cluster Distance on Running Time. (a)(b) With size of datasets (RN and UN) from 10k to 200k,(c) real datasets, (d) Different distances between clusters from 0.0s to 2.0s.

The variation in processing time was observed with different datasets and parameter settings, with an emphasis on the efficiency of the Grid-density-based algorithm. Additionally, the impact of the distance between clusters on algorithm performance and the influence of dataset size on processing time were examined. As shown in Figure 6a and Figure 6b, all algorithms exhibit an increase in execution time as the dataset size grows. It is observed that changes in dataset size have little to no effect on the GDB algorithm, and it consistently outperforms the other two algorithms. Specifically, in the RN dataset, the Expansion algorithm appears to perform better than the Naïve algorithm, but in the case

Table 2: Performance Metrics of Methods According to Different Clusters Distances at N=10

	Naive			Expand			Grid		
Params	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0.0S	0.646112	0.774416	0.675623	0.465359	0.570309	0.510455	0.832838	0.767548	0.789798
0.1S	0.913125	0.896755	0.896742	0.55817	0.599531	0.576889	0.979322	0.861084	0.914817
0.2S	1	0.909221	0.951973	0.66081	0.622877	0.635221	0.995051	0.8844	0.934519
0.3S	1	0.907809	0.950875	0.706257	0.64551	0.671868	1	0.8916	0.941077
0.4S	1	0.92077	0.958424	0.794623	0.678904	0.726643	1	0.8588	0.920029
0.5S	1	0.919942	0.957427	0.842705	0.691178	0.754377	1	0.8862	0.938935
1.0S	1	0.927715	0.961969	1	0.754	0.849154	1	0.8798	0.933036
1.5S	1	0.92736	0.962093	1	0.711333	0.818404	1	0.8926	0.941754
2.0S	1	0.931813	0.964396	1	0.738667	0.839606	1	.8824	0.933264
Avg.	0.951026	0.901756	0.919947	0.780880	0.668034	0.709180	0.978579	0.867159	0.916359

of the UN, the Expansion algorithm takes more time compared to the Naïve algorithm. In the real-world CAS and NE datasets, the GDB algorithm continues to demonstrate superior performance as seen in Figure 6c, while the Expansion algorithm falls short in comparison. When testing the impact of inter-cluster distances, a dataset consisting of 100,000 points was utilized, with inter-cluster distances ranging from 0.0s to 2.0s, as depicted in Figure 6d. All three algorithms exhibited relatively stable performance. On average, the GDB algorithm outperformed the Expansion algorithm by a factor of 18 times and surpassed the Naive method by 22 times in terms of processing speed.These experimental results further underscore the outstanding performance of the Grid density algorithm in handling dense nearest-neighbor queries with varying data distributions and sizes. The algorithm not only provides efficient processing speed but also maintains high precision across a range of clustering scenarios, making it a viable choice for practical applications.

Additionally, the impact of different threshold settings on algorithm performance was investigated, along with the performance of the three algorithms on the labeled dataset RN, specifically using a threshold of 8.The Grid algorithm exhibited superior efficiency, outpacing the Naive algorithm by 4 to 38 times, and the Expand algorithm by 8 to 20 times in processing speed. Notably, it also maintained an average F1 score above 0.86, demonstrating a balanced performance in dense nearest neighborhood queries. In the investigation of threshold settings on the Grid-density-based (GDB) algorithm's performance, the threshold variable N is set ranging from 2 to 10. The impact of this parameter variation on the algorithm's efficacy was systematically analyzed and is illustrated in Figure 7. The findings from this analysis are pivotal for fine-tuning the GDB algorithm to different data characteristics and application requirements.

This significant improvement in speed suggests that the Grid algorithm can effectively reduce computational overhead while processing large-scale, complex datasets without compromising the accuracy of the queries.

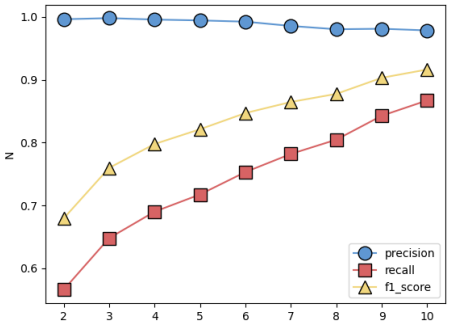


Figure 7: Effect of Threshold N from 2 to 10 experiments on RN with minPts=4

Overall, the Grid algorithm shows exceptional adaptability to changes in data scale and query distances, making it not only suitable for large-scale data processing scenarios but also a reliable solution for efficient spatial data analysis.The comprehensive assessment of this experimental content further solidifies the effectiveness of the GDB algorithm in dense nearest-neighbor query tasks, providing strong support for future research and applications.

5 Conclusion and Future Work

In this thesis, the Grid-density-based Dense Nearest Neighborhood Query approach has been introduced, as well as the naïve and expansion algorithms. The GDB algorithm significantly enhances the efficiency of clustering for large-scale datasets by partitioning the data space using grid cells. Through experimental evaluations, the substantial performance advantages of the GDB algorithm has been demonstrated in terms of runtime and clustering quality. Additionally, its performance has been compared with traditional clustering methods, highlighting the competitiveness of the

GDB algorithm.

This research has mainly focused on two-dimensional data. For future work, it can explore how to extend the GDB algorithm to handle high-dimensional data to meet the demands of various domains. Also, applying the GDB algorithm to practical problems, such as image analysis, bioinformatics, and social network analysis, will be an important research area in the future.

In summary, great potential is held by the GDB algorithm for large-scale DNNH query, and powerful tools can be provided for future research in data science and machine learning. More innovations and advancements in this field are eagerly anticipated.

References

[1] Dong-Wan Choi and Chin-Wan Chung. Nearest neighborhood search in spatial databases. In *2015 IEEE 31st International Conference on Data Engineering*, pages 699–710. IEEE, 2015.

[2] Hina Suzuki, Hanxiong Chen, Kazutaka Furuse, and Toshiyuki Amagasa. Dense nearest neighborhood query. In *International Conference on Intelligent Technologies and Applications*, pages 3–16. Springer, 2021.

[3] Nick Roussopoulos, Stephen Kelley, and Frederic Vincent. Nearest neighbor queries. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 71–79, 1995.

[4] Sang Le, Yuyang Dong, Hanxiong Chen, and Kazutaka Furuse. Balanced nearest neighborhood query in spatial database. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–4. IEEE, 2019.

[5] Song Tang, Yan Yang, Zhiyuan Ma, Norman Hendrich, Fanyu Zeng, Shuzhi Sam Ge, Changshui Zhang, and Jianwei Zhang. Nearest neighborhood-based deep clustering for source data-absent unsupervised domain adaptation. *arXiv preprint arXiv:2107.12585*, 2021.

[6] Md Saiful Islam, Bojie Shen, Can Wang, David Taniar, and Junhu Wang. Efficient processing of reverse nearest neighborhood queries in spatial databases. *Information Systems*, 92:101530, 2020.

[7] Hayata Takagi, Hanxiong Chen, Kazutaka Furuse, and Hiroyuki Kitagawa. Aggregate nearest neighborhood queries. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 396–414. Springer, 2021.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[9] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[10] Alexander Hinneburg and Hans-Henning Gabriel. Denclue 2.0: Fast clustering based on kernel density estimation. In *International symposium on intelligent data analysis*, pages 70–80. Springer, 2007.

[11] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496, 2014.

[12] XIA Lu-Ning and JING Ji-Wu. Sa-dbscan: A self-adaptive density-based clustering algorithm. *Journal of University of Chinese Academy of Sciences*, 26(4):530, 2009.

[13] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734, 2000.

アニメーション付き 3 次元アセンブリモデル検索のための 3 次元ラドン変換の効率化

内尾　亮太† 　片山　薫†

† 東京都立大学システムデザイン研究科情報科学域　〒 191–0065 東京都日野市旭が丘 6-6
E-mail: †uchio-ryota@ed.tmu.ac.jp, ††kaoru@tmu.ac.jp

あらまし　アニメーション付き 3 次元アセンブリモデルの検索において用いられる 3 次元ラドン変換を効率的に計算する手法を提案する．アニメーション付き 3 次元アセンブリモデルは多数の部品からなる CAD モデルで，各部品がフレームごとに個別に回転・平行移動によってモデルの動きが定義される．静止モデルに対する部分一致検索をアニメーションモデルの全てのフレームに繰り返し行うと，処理時間が大きくなる．提案手法では，特徴量抽出で行う 3 次元ラドン変換を，アニメーション付き 3 次元 CAD モデルの初期姿勢の値から各フレームの値を補完して計算することで処理時間を短縮した．

キーワード　アニメーション付き 3 次元アセンブリモデル，3 次元ラドン変換，パターン認識，CAD，検索

1 はじめに

近年，先進的な技術の進展とデジタル化の波が，様々な分野において 3 次元モデルの利用を急速に促進している．特に CAD（Computer-Aided Design）などのモデリング技術は，製造業から建築，医療，エンターテインメントなど幅広い領域で新たな創造性と効率性をもたらしている．これらの 3 次元モデルは，物理的なプロトタイプを作成する前に詳細なデザインや解析が行えるため，開発プロセスの迅速化やコスト削減に寄与している．CAD などで設計される 3 次元モデルのの多くは，様々な小さな部品が組み合わさってできるアセンブリモデルである．CAD ではモデルの表面の情報しか与えられないため，多数の部品を持つアセンブリモデルとして階層構造を与えることで，モデルの内部構造の情報を与えている．さらに，そうした 3 次元モデルに対しアニメーションを付与することで，コンピュータ上で機械部品の実際の動作を再現したり，3DCG のように使用するなど，アニメーション付き 3 次元モデルの需要は増えていくと考えられる．アニメーション付き 3 次元モデルの増加に伴い，そうしたモデルの効率的な比較，検索技術が必要となる．

本稿では，アニメーション付き 3 次元アセンブリモデル検索における 3 次元ラドン変換の効率的な手法を提案する．検索対象となるアニメーション付き 3 次元アセンブリモデルとは，3 次元アセンブリモデルにアニメーション情報を付与したものである．アニメーション付き 3 次元アセンブリモデルの例を図 1 に示す．まず，アセンブリモデルは図 1 のモデルのように，多数の部品モデルを持ち，その部品モデルによってモデル全体としての内部構造の情報を持つ．また，アセンブリモデルは動きや素材などによって分けられるサブアセンブリという部品の集合を持つ．図 1 ではその色分けによってサブアセンブリが分かれている．次に，アニメーション付き 3 次元モデルは初期姿勢の全ての部品モデルと，各部品モデルの回転移動，平行移動の情報が記述された Affine 行列からなる．アニメーション付き 3 次元モデルのアニメーションの時間軸上でサンプリングしたものをフレームという．モデルの移動を記述する Affine 行列は，全てのフレームにおいて各部品の前のフレームからの移動でなく初期姿勢からの移動である．最後に，アニメーション付き 3 次元モデルに対する検索について記述する．検索はモデルの特徴量を計算し，それを比較して類似度を出すことで行う．この時，サブアセンブリの構造も考慮する．アニメーションモデルの特徴量抽出は，アニメーションモデルの全てフレームに対して静止モデルの検索手法を適用することで行う．アニメーションモデルと静止モデルの類似度計算は，アニメーションモデルのうち最も類似度が高いフレームと静止モデルとの類似度を，そのアニメーションモデルと静止モデルの類似度とする．全てのモデルに対し類似度が高いものから順にソートすることで検索結果とする．

(a) 初期姿勢　**(b)** 1/3 移動時　**(c)** 2/3 移動時

図 1: アニメーション付き 3 次元アセンブリモデルの例.

2 関連研究

様々な 3 次元モデルを対象とした検索の研究がある．例えば，Tangelder ら [1] や Li ら [2] は 3 次元 CAD モデルを対象とした様々な検索について網羅的に調査している．しかし，これらの研究によって調査された検索手法の多くはモデル単体同士の比較手法であり，多くの部品や階層構造を持ったアセンブリモデルに対応していない．

階層構造などを持つモデルに対する研究もある．Lupinetti ら [3] は，コンテンツベースの CAD アセンブリ検索に関する既

存の研究を包括的に調査している．Zhang ら [4] は、各部品の形状パターンや結合の種類などを示す記述子を付与したグラフを比較することで抽象化した形状パターンを考慮した検索手法を提案している．さらに Lupinetti ら [5] は、サブアセンブリが直線状や円形などに並ぶことが多いことに着目し、そのような抽象化した配置パターンを考慮した検索手法を提案している．これらの検索手法は抽象化したモデルの部分の配置を元に検索を行えるが，機械設計などの分野では具体的な形状や配置の詳細な値も重要である．

アニメーションを持つ 3 次元モデルに関する検索の研究は，人型のモーションキャプチャなどを中心に行われている．Wang ら [6] は固有ベクトル相関に基づいてマッチングスコアを計算することで人型モデルのモーションのリアルタイム検索を実現している．また，Luo ら [7] は変分オートエンコーダによるモーション推定 (MEVA) と呼ばれる手法を用いて人間の動きの推定を行っている．これらは人型のモデルに対する検索手法であり，機械設計などに使われる多くの部品を持つアセンブリモデルでの利用は難しい．豊泉ら [8] は球面調和関数展開などを用いた，アニメーション付き 3 次元アセンブリモデルに対しての検索手法を提案している．しかし，この手法は完全一致検索のみに対応しており，モデルの部分一致検索に対応してない．

Katayama ら [9] は 3 次元ラドン変換による特徴量抽出を用いた 3 次元アセンブリモデルに対する検索手法を提案している．この手法は静止したモデルを前提としているが，それをアニメーションのフレームごとに順に対応させることでアニメーション付きモデルの検索も可能である．しかし，その方法では処理時間がかかりすぎてしまうという問題を抱えている．本稿では，katayama らの手法の特徴量抽出で使われる 3 次元ラドン変換について，対象モデルが部品の形状は変わらないアニメーションモデルであることを利用した効率的な手法を提案する．これによりアニメーション付き 3 次元アセンブリンモデルに対してより短時間での特徴量抽出が可能となる．

3 CAD モデルの検索

本性では，Katayama が提案した，クエリモデル，データベースモデル共に静止モデルであるときの部分一致検索の手法について記述する．この検索手法は以下の 2 つの工程に分けられる．

（1） 3 次元ラドン変換を利用した特徴量抽出

（2） 得られた特徴量の類似度の計算

全てのデータベースモデルに対し工程 (1) の特徴量抽出を行い，工程 (2) によってクエリモデルと類似度比較を行う，その類似度が高いものから順に検索結果とする．

3.1 特徴量抽出

工程 (1) の特徴量抽出に利用する 3 次元ラドン変換のアルゴリズムを Algorithm1 に示す. 3 次元ラドン変換では球面に一様に分布した視点 *View* からの体積変化をとる．図 2 に 3 次元ラドン変換における体積変化の計算の概要図を示す．*View* の全ての要素 v に対し，その法線ベクトルを $x(v)$ とする．体積変化を

Algorithm 1 $3DRadonTransform(Model, View, S)$

Input: 静止モデル *Model*, 球面上に一様に分布した視点リスト *View* 体積変化を計算するサンプリング数 S

Output: *Model* の 3 次元ラドン変換結果の配列 *Volume*

```
for all Parts ∈ Model do
    for all v ∈ View do
        x(v) ← v の法線ベクトル
        for s = 1 → S do
            Volume(Parts, v, s) ← 平面 Π(s, x(v)) と平面 Π(s+1, x(v)) によって Parts が挟まれた部分の体積
        end for
    end for
end for
return Volume
```

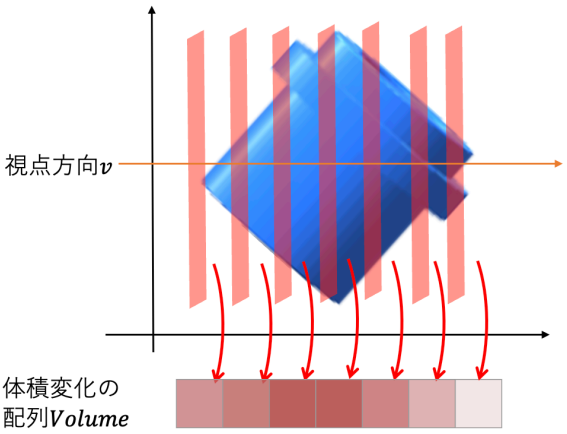


図 2: 3 次元ラドン変換の体積変化の計算方法.

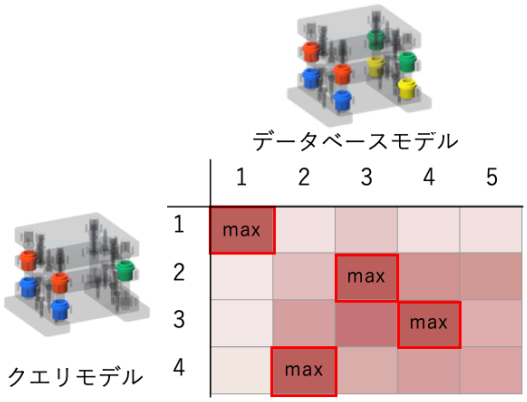


図 3: 部品の割り当ての概念図.

計算するサンプリング数を S とし，1 から S までの s に対し，平面 $\Pi(s, x(v))$ を考える．平面 $\Pi(s, x(v))$ と平面 $\Pi(s+1, x(v))$ によって部品が挟まれた部分の体積を全てのサンプリング範囲で計算したものを $Volume(Prats, v, s)$ とする．この計算を全ての $Prats, v, s$ に対して行う．これにより得られた *Volume* がモデルの特徴量となる．

3.2 類似度の計算

3 次元ラドン変換の結果から異なるモデル同士の比較する際，

Algorithm 2 *CalculateSimilarity*($Volume_q$, $Volume_d$)

Input: クエリモデルの特徴量 $Volume_q$，データベースモデルの特徴量 $Volume_d$
Output: クエリモデルとデータベースモデルの特徴量の類似度 *Similarity*

1: $viewSimilarityList \leftarrow$ 視点数 × 視点数のリスト
2: $partsSimilarityList \leftarrow Volume_q$ の部品数 × $Volume_d$ の部品数のリスト
3: **for all** $v_q \in View$ **do**
4: **for all** $v_d \in View$ **do**
5: **for all** $idx_q \in$ パーツ数 **do**
6: **for all** $idx_d \in$ パーツ数 **do**
7: $partsSimilarityList \leftarrow Volume_q(idx, :, v)$ と $Volume_d(idx, :, v)$ の相互相関を計算
8: **end for**
9: **end for**
10: **for** $s = 1 \rightarrow S^2 - 1$ **do**
11: $partsSimilarityList(s)$ の割り当て問題を計算する．
12: $partsSimilarity(s) \leftarrow$ 割り当てられた値の平均
13: **end for**
14: $viewSimilarityList(V_q, v_d) \leftarrow$ 計算した $partsSimilarity$ の最大値
15: **end for**
16: **end for**
17: $viewSImilarityList$ の割り当て問題を計算する．
18: $Simiarity \leftarrow$ 割り当てられた値の平均
19: **return** $Similarity$

Algorithm 3 3*DradonForAnimetion*($View, Model, affine, k, S$)

Require: 球面上に一様に分布した視点リスト $View$, アニメーションモデルの初期姿勢 $Model$, モデルのフレームごとの移動を示すアフィン行列 $affine$, 使用する近傍点の数 k, 体積変化を計算するサンプリング数 S
Ensure: アニメーションモデルの初期姿勢の 3 次元ラドン変換結果から計算されたの 3 次元ラドン変換結果 $calcuratedVolume$

1: //高視点数での初期姿勢の 3 次元ラドン変換
2: $Volume1 \leftarrow 3DRadonTransform(Model, highView, R)$
3: **for all** $Parts \in Model$ **do**
4: **for** i in フレーム数 **do**
5: **for all** $v \in View$ **do**
6: //回転移動の処理
7: $v' \leftarrow v * affine\ (i)$ の回転移動部分
8: $indices \leftarrow v'$ の k 個の近傍点のインデックス
9: $ratios \leftarrow$ 正規化された $indices$ の点と v' の距離の比
10: $calcuratedVolume(v) \leftarrow \sum(Volume(indices) * ratios(indices))$
11: //平行移動の処理
12: $\hat{v} \leftarrow v$ の単位ベクトル
13: $shiftVec \leftarrow \hat{v} \cdot affine(i)$ の平行移動部分
14: $calcuratedVolume(v)$ を $shiftVec$ の分シフト
15: **end for**
16: **end for**
17: **end for**
18: **return** $calcuratedVolume$

モデル全体の姿勢が揃っていないことがある．また，部分一致検索においては，モデル全体の形状が異なるため，形状からモデルの姿勢を合わせるのは困難である．そのため，工程 (2) で特徴量の比較から類似度計算をする際には，モデルが回転移動，平行移動している場合を考慮する必要がある．工程 (2) の類似度計算のアルゴリズムを Algorithm2 に示す．類似度計算では，最初に全ての部品，視点の体積変化 $Volume$ 同士で相互相関をとる．これは，モデルが平行移動をしていることを考慮するためである．Algorithm2 の 3 行目から 9 行目でその処理を行なっている．その後の処理では，相互相関によって得られた $partsSimilarityList$ を利用して，そのサンプリング数 $S^2 - 1$ で視点の割り当てと部品の割り当てを行う．サンプリング数の類似度の中で最大となる部分の類似度を全体の類似度とする．部品の割り当ての概念図を図 3 に示す．図はクエリモデルが部品数 4 で，データベースモデルが部品数 5 の時の割り当てである．それぞれの部品を割り当てた時，最も類似度の合計が大きくなるように選ぶ．これは割り当て問題として効率的な方法が提案されている．これによって得られた類似度をある視点での類似度として扱う．視点のマッチングは，モデルが回転移動をしていることを考慮するために行う．部品の類似度をとるときと同様に，割り当て問題を解き，割り当てられたペアの類似度の平均が最も高いものがモデル全体の類似度となる．

4 提案手法

アニメーションモデルに対する検索は，静止モデルに対する検索手法をアニメーションモデルの各フレームで行うことで適用できる．以降ではこれを既存手法と呼ぶ．既存手法では，全てのフレームで 3 次元ラドン変換を行うため実行時間がかかり過ぎてしまうという課題がある．そこで，提案手法ではアニメーションモデル全体の 3 次元ラドン変換結果を初期姿勢の 3 次元ラドン変換結果から計算することで，特徴量抽出の高速化を目指した．

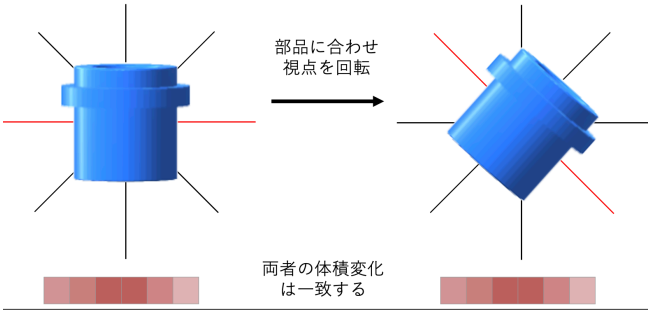


図 4: 回転移動の処理の概念図．

提案手法のアルゴリズムを Algorithm3 に示す．提案手法では，最初に初期姿勢の 3 次元ラドン変換を行う．この値を利用して，各フレームの体積変化を計算する．離散した視点と視点の間の体積変化を求めるために使うため，この 3 次元ラドン変換は高視点数で行う必要がある，その後の処理は，主に回転移動に対する処理と平行移動に対する処理に分けられる．

図 4 に回転移動に対する処理の概念図を示す．部品が回転した後のある視点 v での体積変化の値は，その視点を同じだけ回転した視点 v' からの体積変化と等しくなる．Algorithm3 の 7

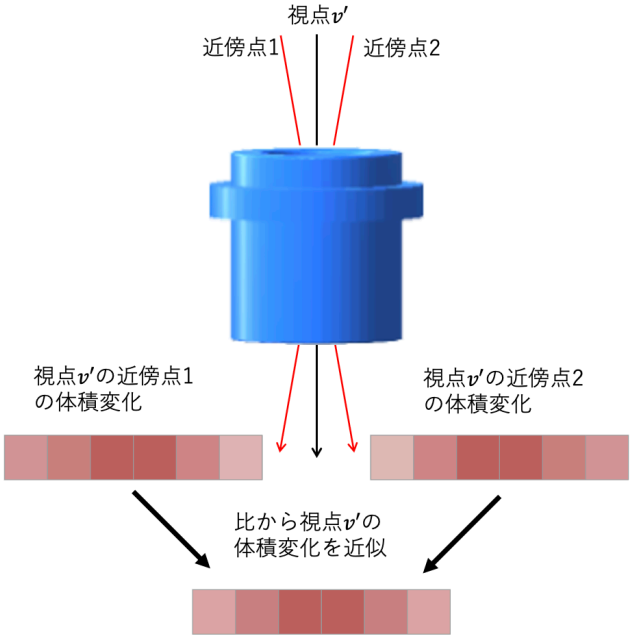


図 5: 計算していない視点 v' からの体積変化の近似の概念図 ($k = 2$).

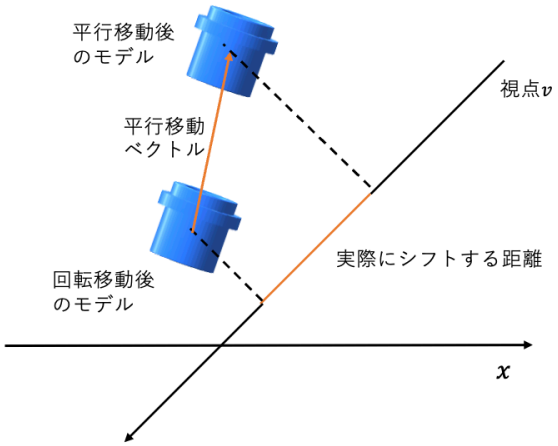


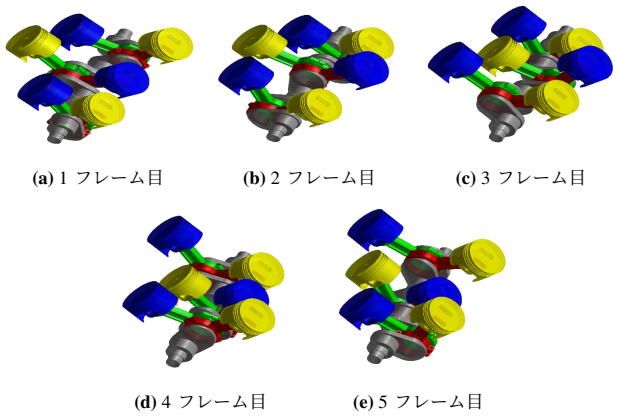
図 6: 平行移動の処理の概念図.

行目のように，体積変化を求める視点 v を部品の回転量分の回転行列をかけることで必要な視点 v' を計算する．しかし，計算上視点は離散的であるため，v' は視点リスト $View$ にあるとは限らない．そこで，3 次元ラドン変換の各視点の体積変化は連続的に変化するので，図 5 のように v' に近い点をいくつか利用し，v' からの体積変化を計算する．Algorithm3 の 8，9 行目のように，v' の k 個の近傍点のインデックス $indices$ とその距離の比 $ratios$ を計算する．その値を使い，v からの体積変化の値は以下のように求める．

$$calcuratedVolume(v) = \sum_{i \in indices} Volume(i) * ratios(i) \qquad (1)$$

各近傍点の体積変化にその比をかけたものの合計がこれにより，回転移動に対する処理は完了する．

図 6 に平行移動に対する処理を示す．平行移動に対する処理は回転移動に対する処理がすでに終わっている体積変化に対し



(a) 1 フレーム目 **(b)** 2 フレーム目 **(c)** 3 フレーム目

(d) 4 フレーム目 **(e)** 5 フレーム目

図 7: 実験に使用するモデル V6Engine(フレーム数 5)

表 1: 評価実験の入力パラメータ

アニメーションモデルの部品数	19
補正に利用する近傍点の数 k	3
体積変化のサンプリング数 S	100
初期姿勢の 3 次元ラドン変換の視点数	600

行う．部品が平行移動した時，v における体積変化は部品の平行移動ベクトルの v の方向ベクトルの成分だけシフトする．平行移動ベクトルの v の方向ベクトルの成分は v の単位ベクトル $\hat{v}$ と平行移動ベクトルの内積を求めることで計算できる．

記述した回転移動，平行移動に対する処理を，各部品，フレーム，視点に対し行うことで，3 次元ラドン変換結果を効率的に計算することができる．

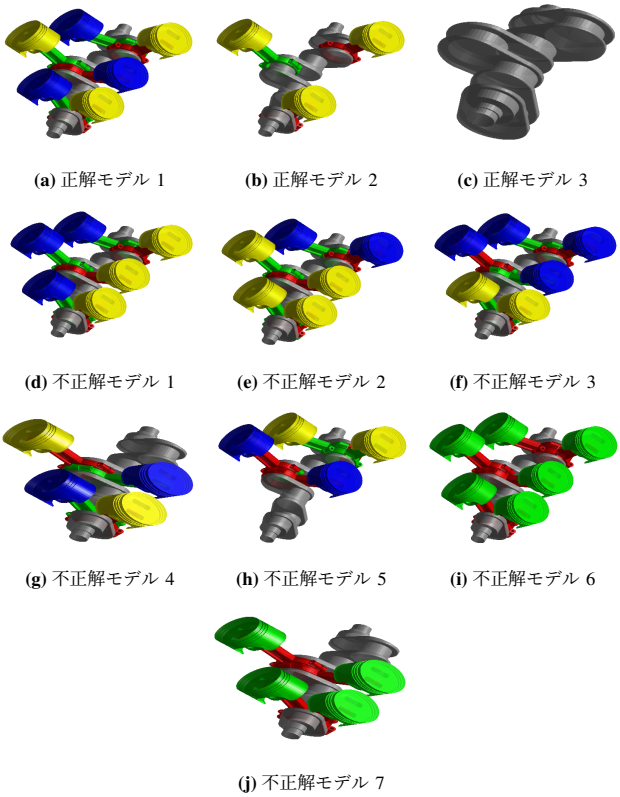
$$H(\omega, r)h(t) = \frac{1}{r}\delta(t - \frac{c}{r})dth(t) = a_1\delta(t - \tau_1) + a_2\delta(t - \tau_2)$$

5 評価実験

提案手法を評価するため，既存手法と提案手法において，3 次元ラドン変換による特徴量抽出の処理時間を比較する．使用するモデルは GrabCAD [10] から取得した V6Engine を利用する．使用するモデルを図 7 に示す．このモデルは，図 7 で灰色に示されたクランクシャフトが一周することで青色，黄色で示された 6 つのピストンが連動しながら運動を行うアニメーションが付与されている．このモデルと，このモデルと同じ形状の静止モデル，その部分モデルを利用して実験を行う．実験は，MATLAB R2023b にてプログラムを実装し，Intel Core i9 13900F 2.0 GHz CPU 32GB RAM, Windows11 Home を搭載したコンピュータで行った．

5.1 処理時間

V6Engine をフレーム数 10,20,30 に分割したモデルにおいて，視点数を 20 から 100 まで 20 刻みで増やして既存手法，提案手法それぞれで 3 次元ラドン変換を行った際の処理時間を比較した．フレーム数，視点数以外のパラメータについては表 1 の通りである．



(a) 正解モデル 1　(b) 正解モデル 2　(c) 正解モデル 3

(d) 不正解モデル 1　(e) 不正解モデル 2　(f) 不正解モデル 3

(g) 不正解モデル 4　(h) 不正解モデル 5　(i) 不正解モデル 6

(j) 不正解モデル 7

図 8: 実験に使用するデータベースモデル群.

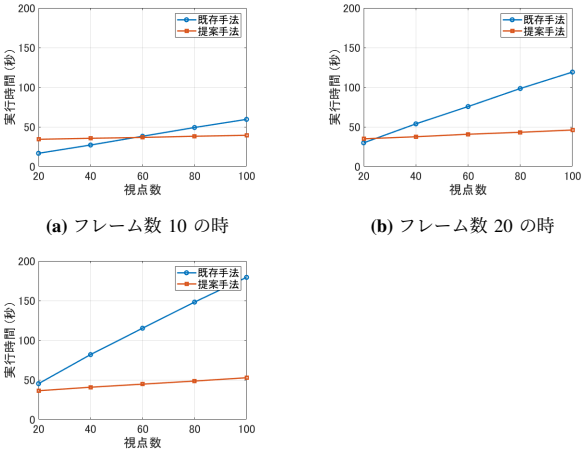
5.2 検索精度

V6Engine に対し，完全一致するモデルを 1 つ，部分一致する静止モデルを 2 つ，一致しない静止モデルを 7 つ，合計 10 個のモデル用意した．それらのモデルを図 8 に示す．正解となるモデルは完全一致するモデルを 1 つ，部分一致する静止モデルを 2 つの合計 3 つのモデルである．図 7 のアニメーションモデルをクエリモデル，図 8 のモデル群をデータベースモデル群とする．検索の際には，クエリモデルはランダムな回転をかけ，データベースモデルはアニメーションのランダムな一部分を切り取ったものとする．この実験を各視点数，各フレーム数で 5 回ずつ行った．この時の検索結果として，正解となるモデルがの類似度が一致しない全てのモデルを超えている正答率を検索精度とする．

6 結果と考察

6.1 処理時間

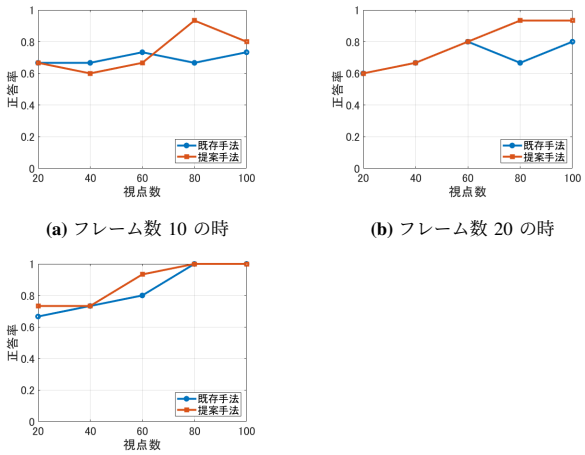
各フレーム数，視点数における処理時間のグラフを図 9 に示す．フレーム数 10 の時は 60 視点から，フレーム数 20 の時は 40 視点から、フレーム数 30 の時は全ての視点数において提案手法が既存手法を実行時間で上回る結果となった．提案手法では既存手法に比べ低視点数であるときは多くの実行時間がかかっているが，これは既存手法では行っていない初期姿勢の 3 次元ラドン変換を 600 視点で行っており，その部分の計算が低視点での全てのフレームの計算より処理時間が大きくなるからである．今後，よりサイズの大きなモデルやより詳しいアニメーションが施されたモデルが作成されていくと，検索により多くのフレーム数が必要となっていくと考えられる．提案手法は，そうしたモデルに対してより短時間での検索を可能にする．



(a) フレーム数 10 の時　(b) フレーム数 20 の時

(c) フレーム数 30 の時

図 9: 各フレーム，視点数での実行時間



(a) フレーム数 10 の時　(b) フレーム数 20 の時

(c) フレーム数 30 の時

図 10: 各フレーム，視点数での検索精度

6.2 検索精度

各視点数，フレーム数における検索精度を図 10 に示す．全体としてフレーム数，視点数が大きくなるにつれ正答率が上昇する傾向にある．フレーム数 30，視点数 80 の時点から提案手法，既存手法ともに正答率が 1 となった．また，その他の視点数，フレーム数でも提案手法と既存手法で正答率に大きな差はなかった．提案手法では，既存手法と同等の検索精度を持ち，より短時間での検索を可能にする．

7 おわりに

アニメーション付き 3 次元アセンブリモデルにおいて用いられる 3 次元ラドン変換の効率的な手法を提案した．提案手法では，低視点数，低フレーム数を除いて，静止モデルに対する特徴量抽出の手法を繰り返す手法に比べ高速に処理が可能であることがわかった．また，その際の検索精度は既存手法と同等の

ものである．しかしながら，提案手法においても十分な処理時間とは言えず，より大規模な検索においては更なる効率化が求められる．また，本稿では特徴量抽出の 3 次元ラドン変換における高速化を示したが，その特徴量の比較にも多くの時間がかかっているため，高速化が望まれる．今後は，特徴量の比較においてもアニメーション特有の特性を利用した高速化の実現に取り組む予定である．

文　　献

[1] Johan WH Tangelder and Remco C Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia tools and applications*, Vol. 39, No. 3, pp. 441–471, 2008.

[2] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, et al. A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, Vol. 131, pp. 1–27, 2015.

[3] Katia Lupinetti, Jean-Philippe Pernot, Marina Monti, and Franca Giannini. Content-based cad assembly model retrieval: Survey and future challenges. *Computer-Aided Design*, Vol. 113, pp. 62–81, 2019.

[4] Jie Zhang, Zhijia Xu, Yuan Li, Shoushan Jiang, and Nannan Wei. Generic face adjacency graph for automatic common design structure discovery in assembly models. *Computer-Aided Design*, Vol. 45, No. 8-9, pp. 1138–1151, 2013.

[5] Katia Lupinetti, Franca Giannini, Marina Monti, and Jean-Philippe Pernot. Multi-criteria retrieval of cad assembly models. *Journal of Computational Design and Engineering*, Vol. 5, No. 1, pp. 41–53, 2018.

[6] Pengjie Wang, Rynson W.H. Lau, Zhigeng Pan, Jiang Wang, and Haiyu Song. An eigen-based motion retrieval method for real-time animation. *Computers & Graphics*, Vol. 38, pp. 255–267, 2014.

[7] S. Alireza Golestaneh Luo, Zhengyi and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. *Proceedings of the Asian Conference on Computer Vision*, 2020.

[8] 豊泉翔太朗, 平島栄志, 片山薫. アセンブリ構造をもったアニメーション CAD モデルのレイアウト情報を考慮した検索, DEIM2022, A43-3.

[9] Kaoru Katayama, Youta Yamaji, Shotaro Toyoizumi, and Takashi Hirashima. Content-based retrieval of 3d cad subassemblies using 3d radon transform. *IEICE Transactions on Information and Systems*, Vol. 31, pp. 734–742, 2023.

[10] GrabCAD. The grabcad community library. https://grabcad.com/library.

大規模DNAデータベースに対するTop-k相関問合せ手法の提案

八木　隆一†　　直井　悠馬†　　塩川　浩昭††

† 筑波大学大学院理工情報生命学術院　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1-1-1
E-mail: †{yagi,naoi}@kde.cs.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　DNA データベースに対する相関問合せとは，クエリとして DNA シーケンスが与えられたときクエリと共起して出現する部分シーケンスを検出する問題であり，生命情報科学分野で広く応用されている．この問題において DNA データベース内の全てのシーケンスの部分シーケンスが候補となるため，大量の DNA シーケンスからなる大規模 DNA データベースに対する相関問合せでは計算コストが大きくなる．本論文ではクエリとの相関値が大きい上位 k 件の部分シーケンスを得ることを目的とする．本論文ではそのための閾値ベースの手法ならびにその高速化手法を提案し，実データを用いた評価実験により提案手法の有効性を検証した．

キーワード　時系列データ処理，データ構造・索引，問合せ処理

1 序　　論

近年文字列間の類似度に関する研究が多くされている．文字列間の類似度に関する研究はデータベース分野や生命情報科学分野等において基本的な問題であり，データクリーニングやゲノム配列アセンブリングなどに数多く応用されている．例えば，人間を含めいかなる生物の DNA も A, C, G, T の 4 文字で構成させる文字列として表すことができ，生物学的に近い種は似たような DNA 配列となる．そのため与えられた DNA のデータベースから類似した DNA 配列を見つけ出す問題は生命情報科学分野において有用な問題とされている．また，膨大な情報量から目的に合わせた有用な情報を得るためのデータマイニング等でも文字列処理に関する技術は欠かせないものとなっている．

文字列間の類似度に関する研究として文字列類似結合 [1] がある．文字列類似結合は文字列集合が与えられたときに集合内から類似した文字列のペアを見つけ出す問題である．文字列類似結合の問題は類似度として用いる指標や対象とする問題の違い等によっていくつかのカテゴリーに分類することができる．代表的な類似度の指標として編集距離 [2] やハミング距離があり，そのほか Jaccard 類似度やコサイン類似度等がある．なかでも編集距離を類似度の指標とした研究は数多くされている．例えば，文字列集合と閾値が与えられたとき文字列集合から編集距離が閾値以下となる文字列のペアを見つける問題として文献 [3–6] などがある．これらの研究では文字列集合に対してインデックスを構築するあるいはハッシュベースの検索を行うなどの工夫がされ効率的に編集距離が小さいペアを見つけ出すことができる．

しかしながら，上記の文字列ベースの類似検索手法は 2 つの欠点が存在する．1 つ目は編集距離の計算は膨大な計算量を必要とする点である．DNA シーケンスの平均長を m としたとき，編集距離計算は $\mathcal{O}(m^2)$ の計算時間を要する．Ukkonen はこの計算コストを削減するための効率的なアルゴリズムを提案した [7]．しかし，t を $t < m$ となる定数値としたとき，このアルゴリズムは $\mathcal{O}(tm)$ の計算時間を要する．2 つ目は編集距離は部分構造の長さに大きく影響される点である．これは編集距離が一方の文字列を他方の文字列に変換するために必要な編集操作の数をカウントするためである [8]．そのため，2 つの部分構造の長さに大きな差があるとき，例え 2 つの部分構造が包含関係にあったとしても，長さに影響されて類似度を正しく評価することができない．

上記の問題を解決するため，本研究では DNA シーケンス探索問題として「DNA データベースに対する相関問合せ」を定義する．相関問合せとは 2 つ以上のオブジェクト間の隠れた特性を明らかにすることであり，主にグラフデータベース分野で盛んに行われている研究である [9–12]．DNA データベースに対する相関問合せは従来の類似文字列探索問題と異なり，クエリとの共起性に基づきデータベース内のシーケンスを評価する．DNA データベースに対する相関問合せはクエリシーケンスと DNA データベースが与えられたとき，クエリと頻繁に共起するサブシーケンスを見つける問題である．DNA データベースに対する相関問合せはクエリとの共起性に基づいてサブシーケンスの探索を行うため，クエリと結びつきのあるサブシーケンスをシーケンス長に関わらず明らかにすることができる．以下に DNA データベースに対する相関問合せの概要を示す．

- q = TCG
- s_1 = ACTCGATCATCCATGCATG
- s_2 = GCAGCTAGCTCCATGCCCGACTGGT
- s_3 = CGATCGATCCATGCATCG
- s_4 = GCTACGATCGATATGCATC
- s_5 = GCTACGAATATGCATC

q = TCG がクエリシーケンスとして与えられ，そして 5 つのシーケンスを含む DNA データベースが与えられているとする．このとき q = TCG とサブシーケンス TCCATGC は同じシーケンス内において頻繁に出現していることがわかる．このような場合，q = TCG とサブシーケンス TCCATGC には共起性があ

ると表現する．本研究の目的はこのようなクエリとの相関値が高いサブシーケンスを明らかにすることである．DNA データベースに対する相関問合せはクエリシーケンスとの共起性を持つ DNA の部分構造を発見することは特定の疾患の原因となりうる DNA 構造を明らかにすることにつながり，基礎生物学研究や医療診断，生命情報科学の分野において必要な技術である．

本研究では DNA データベースに対する相関問合せとして以下の問題を扱う．

- DNA データベース，クエリシーケンス，整数値が与えられたとき，クエリとの相関値が大きいサブシーケンスを整数値の個数分だけ列挙する問題

この問題はクエリとの相関値が与えられた整数の個数分だけ列挙する Top-k 検索問題である．DNA データベースに対する相関問合せでは，データベース中のすべてのサブシーケンスが候補になるため相関値を計算するのに膨大な計算量を要する [13–15]．そのため上記の問題に対し本研究では解になり得ないサブシーケンスを早い段階で除外することで相関値の計算をできるだけ回避し計算量を削減する手法を提案する．この目的のために，まず DNA 構造をグラフ表現に変換する（3.1 節, 3.2 節）．そして，頻出部分グラフマイニング手法 [9,10,16] を応用した枝刈り手法をそれぞれの問題に応用した手法を提案する．評価実験により本研究の提案手法が DNA データベースに対する相関問合せである上記の問題を効率的かつ正確に解けること確認した．

本研究の目的と貢献をまとめると以下のようになる．

- **目的:** 新たな DNA シーケンス探索問題として「DNA データベースに対する相関問合せ」を提案する．クエリとの相関値が大きいサブシーケンスを整数値の個数分だけ列挙する問題（3 節）に対し，グラフ表現を介した枝刈り手法を示す．
- **効率性:** 提案手法は比較手法に対し最大で 10 倍高速に問合せ処理を行うことを確認した（4.2 節）．
- **正確性:** 提案手法は効率的な計算のため DNA の部分構造を動的に除外していくが，常に正確な解を導くことができる（4.3 節）．

2 基本事項

本章では基本事項を説明する．また，表 1 に本論文で用いる記号等を示す．

2.1 前提知識

DNA シーケンスは ACGT の 4 文字からなる文字列として表現することができる．A,C,G,T はそれぞれ塩基である Adenine, Cytosine，Guanine，Thymine の頭文字を表している．DNA は 2 本のヌクレオチド鎖が二重らせん構造となるように，ヌクレオチドに含まれる塩基同士が組み合わさってできている．そのため，DNA の塩基配列を表現するためには 2 つの文字列あるいは 1 次元配列が必要となる．しかし A と T，C と G のような塩基対により DNA は二本鎖が相補的に組み合わさっているため一方の塩基配列からもう一方の塩基配列を特定できる．よって DNA の塩基配列は A，C，G，T からなる計算機科学上の文字列として扱うことができる [17].

s を DNA シーケンスとし，そのシーケンス長を m_s としたとき，m_s は s に含まれる文字数を表す．s が文脈上明らかであるときは m_s を m と表す．s_i，s_j について単射関数 $f : s_i \to s_j$ が存在するとき，$s_i \subseteq s_j$ と表し，s_i は s_j の部分構造であるという．$\mathcal{B}(s)$ を s の全ての部分構造の集合とする．すなわち $\mathcal{B}(s) = \{s_i \mid s_i \subseteq s\}$ である．DNA データベースを $\mathcal{D} = \{s_1, s_2, \ldots, s_N\}$ と表す．DNA データベース $\mathcal{D}$ が与えられたとき，s の射影データベースを $\mathcal{D}_s = \{s' \in \mathcal{D} \mid s \subseteq s'\}$ と定義する．

$supp(s; \mathcal{D})$ を $\mathcal{D}$ における s の支持度とする．すなわち $supp(s; \mathcal{D}) = \frac{|\mathcal{D}_s|}{|\mathcal{D}|}$ である．特に断りがなければ，$supp(s; \mathcal{D})$ を $supp(s)$ と表記する．s_i，s_j について，$\mathcal{D}$ における結合支持度 $supp(s_i, s_j)$ を $supp(s_i, s_j) = \frac{|\mathcal{D}_{s_i} \cap \mathcal{D}_{s_j}|}{|\mathcal{D}|}$ と定義する．$s_i \subseteq s_j$ であるならば明らかに $supp(s_j) \leqq supp(s_i)$ である．また，結合支持度の定義より $supp(s_i, s_j) \leqq supp(s_i)$ ならびに $supp(s_i, s_j) \leqq supp(s_j)$ である．

本論文では，部分構造間の相関値として以下に示すピアソン相関係数 [18] を用いる．

定義 1 (類似度)**.** 部分構造 s_i，s_j の相関値 $\phi(s_i, s_j)$ を以下のように定める．

$$\phi(s_i, s_j) = \frac{supp(s_i, s_j) - supp(s_i)supp(s_j)}{\sqrt{supp(s_i)supp(s_j)(1 - supp(s_i))(1 - supp(s_j))}} \tag{1}$$

ただし，$supp(s_i)supp(s_j)(1 - supp(s_i))(1 - supp(s_j)) = 0$ となる場合，$\phi(s_i, s_j) = 0$ とする．

$\phi(s_i, s_j)$ は $[-1, 1]$ の範囲の値を取る．もし $\phi(s_i, s_j) > 0$ であるならば, s_i と s_j の出現分布に正の相関があることを示す．$\phi(s_i, s_j) < 0$ の場合，負の相関があることを示す．$\phi(s_i, s_j) = 0$ ならば， s_i と s_j は独立して出現する．

例えば，DNA データベース $\mathcal{D} = \{s_1, s_2, \ldots, s_5\}$ が下記のように与えられているものとする．

- s_1 = ACTCGATCATCCATGCATG
- s_2 = GCAGCTAGCTCCATGCCCGACTGGT
- s_3 = CGATCGATCCATGCATCG
- s_4 = GCTACGATCGATATGCATC
- s_5 = GCTACGAATATGCATC

s_i =TCG であるとき，$\mathcal{D}_{s_i} = \{s_1, s_3, s_4\}$ であり，$supp(s_i) = \frac{3}{5}$ である．また，s_j =TCCATGC であるとき，$supp(s_j) = \frac{3}{5}$ である．s_i と s_j の結合支持度は $supp(s_i, s_j) = \frac{2}{5}$ なので，$\phi(s_i, s_j) = \frac{1}{6}$ である．

2.2 問題定義

最後に，1 節で述べた本研究で扱う問題を Top-k CDS search (Top-k Correlated DNA Subsequence search) とし，定義 2 に示す．

定義 2 (Top-k CDS search)**.** *DNA* データベース $\mathcal{D} = \{s_1, s_2, \ldots, s_N\}$，クエリーシーケンス q，整数 k が与えられたとき，*Top-k CDS search* はクエリとの相関値 $\phi(q, s')$ が大きい上位 k 個のサブシーケンスを検出する問題である．

1 節で述べたように，Top-k CDS search では時間計算量が課題

表 1　本論文で用いる記号等の定義

記号等	定義
s	DNA シーケンス
m_s	DNA シーケンス s の長さ．また s に含まれる文字数．すなわち $s = S^1, S^2, \ldots, S^m$
S^i	DNA シーケンス s の i 番目の文字（塩基）．s が文脈上明らかであるときは m と表す．
$\mathcal{D}$	DNA データベース $\mathcal{D} = \{s_1, s_2, \ldots, s_N\}$
$\mathcal{D}_s$	シーケンス s の射影データベース
N	$\mathcal{D}$ に含まれるシーケンスの数．$\lvert\mathcal{D}\rvert = N$
q	クエリ DNA シーケンス
k	相関 DNA サブシーケンスの数
$supp(s; \mathcal{D})$	$\mathcal{D}$ におけるシーケンス s の支持度．特に断りがなければ $supp(s)$ と表す．
$supp(s_i, s_j)$	$\mathcal{D}$ における結合支持度
$\phi(s_i, s_j)$	シーケンス s_i, s_j 間の相関値

となる．DNA シーケンス長を m としたとき，データベース内の各シーケンスは $\mathcal{O}(m^2)$ 個の部分構造を持つため，データベース中に N 個のシーケンスが含まれている場合，Top-k CDS search の時間計算量は $\mathcal{O}(Nm^2)$ となる．しかし，表 2 に示すように，実際の DNA データベースは大きな m と N をもつ．ゆえに，Top-k CDS search に対する効率的なアルゴリズムの開発が必要である．

3　提 案 手 法

本章では定義 2 で示した Top-k CDS search に対する提案手法を示す．まず 3.1 節，3.2 節で提案手法の核となるアイデアを示す．そして 3.3 節, 3.4 節で Top-k CDS search に対する手法を示す．

3.1　基本アイデア

DNA データベースに対する相関問合せへのナイーブなアプローチとして，DNA データベース中のサブシーケンスをすべて列挙しクエリとの相関値を計算する方法がある．しかしその場合の時間計算量は $\mathcal{O}(Nm^2)$ となるため，このアプローチは現実的ではない．そのため提案手法では $\mathcal{D}$ から相関部分構造を短時間で見つけることを目的とする．

相関問合せに関する研究として文献 [9, 10, 16] がある．これらの研究ではグラフデータベースに対する効率的な相関問合せのための手法が提案されている．そのため，DNA データベースに対する相関問合せへのアプローチとして文献 [9, 10, 16] のアイデアを応用することを考える．しかし，グラフと DNA シーケンスは異なるデータ構造であるため，文献 [9, 10, 16] のアイデアを直接 DNA データベースに対する相関問合せに利用することはできない．そこで提案手法ではまず DNA シーケンスを等価なグラフ構造（DNA グラフ）に置き換えることを考える．そして DNA グラフに対し文献 [9, 10, 16] のアイデアを利用することで，相関値が小さい部分構造に対する相関値計算をできるだけ回避し計算対象となる部分構造の数を減らすことで Top-k CDS search を効率的に解く．

まず 3.2 節では DNA シーケンスのグラフ表現を導出する．次に 3.3 節にて Top-k CDS search に対する具体的なアプローチを示す．

3.2　DNA グラフ

DNA シーケンスとグラフの等価性を理論的に議論する．$g = (V, E, L)$ をラベル付きグラフとする，V，E，L はそれぞれノード，エッジ，ノードラベルの集合とする．DNA グラフを以下のように定義する．

定義 3 (DNA グラフ)**.** s に対応する *DNA* グラフを $g_s = (V, E, L)$ とする．V，E，L はそれぞれ $V = \{1, 2, \ldots, m\}$，$E = \{(v_i, v_j) \in V^2 \mid i + 1 = j\}$，$L = \{S^1, S^2, \ldots, S^m\}$ である．

定義 3 は s が，文字の隣接を表すエッジを持つパスグラフ g_s に変換されることを意味している．すなわち，$\mathcal{D} = \{s_1, s_2, \ldots, s_N\}$ はグラフデータベース $\mathcal{G} = \{g_{s_1}, g_{s_2}, \ldots, g_{s_N}\}$ とみなすことができる．定義 1 より，2 つの DNA グラフの相関は以下のように定義される．

定義 4 (DNA グラフ間の相関)**.** $\mathcal{G}_{g_i}$ を $\mathcal{G}$ における g_i のスーパーグラフの集合とする．*DNA* グラフ g_i，g_j の相関は以下のように定義される．

$$\phi(g_i, g_j) = \frac{supp(g_i, g_j) - supp(g_i)supp(g_j)}{\sqrt{supp(g_i)supp(g_j)(1 - supp(g_i))(1 - supp(g_j))}} \tag{2}$$

ただし，$supp(g_i) = \frac{|\mathcal{G}_{g_i}|}{|\mathcal{G}|}$，$supp(g_i, g_j) = \frac{|\mathcal{G}_{g_i} \cap \mathcal{G}_{g_j}|}{|\mathcal{G}|}$ である．

定義 1, 3, 4 より以下の補題が成り立つ．

補題 1. *DNA* シーケンス s_i，s_j と対応する *DNA* グラフ g_i，g_j に対して，$\phi(s_i, s_j) = \phi(g_i, g_j)$ が常に成り立つ．

証明.　DNA シーケンス s_k と対応する DNA グラフ g_{s_k} が与えられたとき，$s_k \subseteq s_i$ ならば g_{s_k} は g_{s_i} の部分グラフである．これは定義 3 に示すように，s_k と g_{s_k} の間に全単射写像が存在するためである．したがって，任意の DNA シーケンス s_i とそれに対応する DNA グラフ g_{s_i} について常に $|\mathcal{D}_{s_i}| = |\mathcal{G}_{g_{s_i}}|$ が成り立つ．よって，定義 1，3，4 より $\phi(s_i, s_j) = \phi(g_i, g_j)$ が常に成り立つ．　□

3.3　グラフベースの候補生成および枝刈りルール

本節では効率的に Top-k CDS search を行う手法を説明する．まず 3.2 節で示した DNA シーケンスと DNA グラフの等価性を利用し，DNA データベース中の DNA シーケンスを DNA グラフに変換する．その後，文献 [10] で提案されたグラフデータベースとクエリグラフ，整数値が与えられたときクエリとの相関値が大きい部分グラフを整数個分だけ列挙するための手法を Top-k CDS search に応用する．各諸定理やルールの詳しい証明は文献 [10] を参照されたい．

はじめに相関値が大きい上位 k 個の DNA グラフを探索する間，それまでに発見された Top-k DNA グラフを保持するキュー Q_{cur} を考える．Q_{cur} には DNA グラフ g と相関値 $\phi(g_q, g)$ のペアが相関値の降順に格納されている．Q_{cur} の k 番目の相関値を ϕ_{min} とすると，候補グラフ g の相関値は少なくとも ϕ_{min}

を超えていなければ Top-k DNA グラフにはなり得ない．

そこでまず，候補グラフ g が取りうる相関値の最大値を考える．候補グラフ g は $supp(g_q, g) = supp(g)$ のとき最大値を取るため，定義 1 で示したピアソン相関係数の式に $supp(g_q, g) = supp(g)$ を代入することで以下の最大値を得る．

$$\phi_{max}(g) = \sqrt{\frac{(1 - supp(q))supp(q, g)}{supp(q)(1 - supp(q, g))}} \tag{3}$$

さらに，$supp(g_q, g) = supp(g; \mathcal{G}_{g_q}) \cdot supp(g_q)$ であることを利用することによって，候補グラフについて射影データベース内 $\mathcal{G}_q$ で最大相関値を計算することができる．よって，探索中に得た候補グラフ g について $\phi_{max}(g)$ を計算し，ϕ_{min} と比較することで g が Q_{cur} に入るか判定する．

定理 1. （文献 *[10]*）候補グラフ g について以下の関係が成り立つとき，g は枝刈りされる．

$$\phi_{max}(g) < \phi_{min} \tag{4}$$

証明. 式 3 より $\phi_{max}(g)$ は候補グラフ g が取りうる相関値の最大値を表す．よって $\phi_{max}(g)$ が ϕ_{min} を下回るとき候補グラフ g は Top-k 相関グラフになり得ない． □

また $\phi_{max}(g)$ について以下の補題が成り立つ．

補題 2. （文献 *[10]*）2 つの候補グラフ g_1, g_2 が与えられたとき，$supp(g_q, g_1) \geqq supp(g_q, g_2)$ ならば $\phi_{max}(g_1) \geqq \phi_{max}(g_2)$ である．

証明. $\phi_{max}(g)$ は $supp(g_q, g)$ に対して単調増加するため補題 2 は成り立つ． □

上記の補題から次のことがいえる．2 つの候補グラフ g_1, g_2 について $g_1 \subseteq g_2$ が成り立つとき，支持度 *supp* の定義から $supp(q, g_1) \geqq supp(q, g_2)$ であるため，候補グラフ g_1 が枝刈りされるとき g_1 のスーパーグラフである g_2 も枝刈りされる．よって，候補グラフ g が枝刈りされるならばそのスーパーグラフもすべて枝刈りすることができる．

上記のように $\phi_{max}(g)$ の特性を利用することで候補グラフとそのスーパーグラフを効率的に枝刈りすることが可能である．さらに射影データベース $\mathcal{G}_{g_q}$ 内での候補グラフ g の支持度 $supp(g; \mathcal{G}_{g_q})$ について以下の定理が成り立つ．

定理 2. （文献 *[10]*）候補グラフ g について以下の関係が成り立つとき，g は枝刈りされる．

$$supp(g; \mathcal{G}_{g_q}) < \varsigma(\phi_{min}) = \frac{\phi_{min}^2}{\phi_{min}^2 \cdot supp(g_q) + 1 - supp(g_q)} \tag{5}$$

証明. $supp(g; \mathcal{G}_{g_q}) < \varsigma(\phi_{min})$ が成り立つとき，$\phi(g_q, g) < \phi_{min}$ となることを示す．関係式 $supp(g; \mathcal{G}_{g_q}) < \varsigma(\phi_{min})$ について，$supp(g_q, g) = supp(g; \mathcal{G}_{g_q}) \cdot supp(g_q)$ より $supp(g_q, g) < (\varsigma(\phi_{min}) \cdot supp(g_q))$ となる．$(\varsigma(\phi_{min}) \cdot supp(g_q))$ をグラフ g' についての結合支持度として扱い，式 3 に代入すると $\phi_{max}(g') = \phi_{min}$ を得る．補題 2 より $supp(g_q, g) < supp(g_q, g')$ ならば $\phi_{max}(g) < \phi_{max}(g') = \phi_{min}$ となる．よって $\phi(g_q, g) \leqq \phi_{max}(g) < \phi_{min}$ となり，定理 2 は成り立つ． □

上記の定理は候補グラフ g について，$supp(g; \mathcal{G}_q) \leqq \varsigma(\phi_{min})$ であるとき，$\phi_{max}(g) \leqq \phi_{min}$ となる ϕ_{min} の関数 $\varsigma(\phi_{min})$ を式 3 から導いたものである．よって，候補グラフ g について定理 2 が成り立つならば，g は枝刈りされると同時に g のスーパーグラフもすべて枝刈りされる．ϕ_{min} の値は探索中に新たな候補グラフが Q_{cur} に挿入されるたびに増加し，それに伴い $\varsigma(\phi_{min})$ の値も増加することから定理 2 の枝刈りルールは探索が進むたびに効果的になっていく．

さらに効率的に Top-k DNA グラフの絞り込みを行う上で下記のルール 1, 2, 3, 4, 5 [10] を用いる．

ルール 1. 候補グラフ g について，$g_q \subseteq g$ ならば $\phi(g_q, g) = \phi_{max}(g)$ である．

$g_q \subseteq g$ であるとき，$supp(g_q, g) = supp(g)$ であるため，補題 2 よりルール 1 は成り立つ．ルール 1 は候補グラフ g がクエリ DNA グラフ g_q のスーパーグラフとなるとき，厳密な相関値の計算をすることなく $\phi(g_q, g)$ を得ることができることを表している．そのため候補グラフ g がクエリ DNA グラフ g_q のスーパーグラフとなるとき，ルール 1 により $\phi(g_q, g)$ を得た結果その値が ϕ_{min} を上回る場合は候補グラフ g を Q_{cur} に格納できる．

さらに下記のルールが成り立つ．

ルール 2. 候補グラフ g_1, g_2 について，$g_2 \subset g_1$ かつ $supp(g_q, g_1) = supp(g_q, g_2)$ ならば $\phi(g_q, g_1) \geqq \phi(g_q, g_2)$ である．

ルール 2 は候補グラフ g_1, g_2 について，$g_2 \subset g_1$ が成り立ち，かつクエリとの結合支持度が同じであるならば g_1 の相関値は g_1 の部分グラフである g_2 より大きくなることを表している．そのためルール 2 は次のように用いることができる．深さ優先的に探索を進める上である候補グラフ g の相関値を計算しその値が ϕ_{min} を下回った場合，g と同じ結合支持度をもつ g の部分グラフについて相関値の計算をすることなく候補から除外することができる．

ルール 3. 候補グラフ g_1, g_2 について，$g_2 \subset g_1$ かつ $supp(g_1)$ の値が計算済みであるとき，以下の *(1), (2)* が成り立つ．
(1) $supp(g_q, g_2) < supp(g_1)$ ならば，$\phi_{max}(g_2)$ は次のように更新される．

$$\phi_{max}(g_2) = \frac{supp(g_q, g_2) - supp(g_q) \cdot supp(g_1)}{\sqrt{supp(g_q)(1 - supp(g_q))supp(g_1)(1 - supp(g_1))}}$$

(2) $supp(g_q, g_2) < f(supp(g_1), \phi_{min})$ ならば，$\phi(g_q, g_2) < \phi_{min}$ である．ただし関数 f は

$$\begin{aligned} &f(supp(g), \phi_{min}) \\ &= \phi_{min}\sqrt{supp(g_q)(1 - supp(g_q))supp(g)(1 - supp(g))} \\ &+ supp(g_q)supp(g) \end{aligned}$$

である．

ある候補グラフ g について $supp(g)$ を計算し相関値を得ているとする．このようなとき結合支持度が $supp(g)$ より小さい g の候補部分グラフについて，ルール 3 (1) の式を用いて取りうる

Algorithm 1 Top-k CDS search に対する提案手法

入力: $\mathcal{D}$, q, and k
出力: Q_{cur}

1: Convert $\mathcal{D}$ to $\mathcal{G}$ by **Definition** 3;
2: Obtain $\mathcal{G}_{g_q}, Q_{cur} = \emptyset$;
3: Find the set of distinct edges, E, in $\mathcal{G}_{g_q}$;
4: Set ϕ_{min} and $\varsigma(\phi_{min})$ initially as 0;
5: **for each** $e \in E$ **do**
6: Obtain $supp(e; \mathcal{G}_{g_q})$;
7: **if** ($\phi_{max}(e) \geqq \phi_{min}$ and $supp(e; \mathcal{G}_{g_q}) \geqq \varsigma(\phi_{min})$) **then**
8: **MineTopCDS**(e);
return Q_{cur};

最大相関値の更新を行えることを表している．この最大値の更新は式 3 よりも厳密なものとなるため，効率的な枝刈りにつながる．またルール 3 (2) により，候補グラフ g が枝刈りされたとき，結合支持度が $f(supp(g), \phi_{min})$ を下回る g の候補部分グラフもまた枝刈りされる．

ルール 2, 3 は候補グラフとその候補部分グラフについての枝刈りルールである．さらに下記のルールが成り立つ．

ルール 4. 候補グラフ g_1, g_2 について，$g_1 \subset g_2$ かつ $supp(g_1) = supp(g_1, g_q)$ ならば，$\phi(g_q, g_2) = \phi_{max}(g_2)$ である．

$g_1 \subset g_2$ かつ $supp(g_1) = supp(g_q, g_1)$ であるとき，$supp(g_2) = supp(g_q, g_2)$ であるため，式 3 よりルール 4 は成り立つ．ルール 4 により，支持度が結合支持度と一致するある候補グラフについて，その候補グラフのすべてのスーパーグラフについて相関値の値は式 3 となるため厳密な相関値の計算を回避できる．

ルール 5. 候補グラフ g_1, g_2 について，$g_1 \subset g_2$ かつ $supp(g_1)$ が計算済みであるとき，$supp(g_q, g_2) \geqq f(supp(g_1), \phi_{min})$ ならば，$\phi(g_q, g_2) \geqq \phi_{min}$ である．

ルール 5 はルール 3 (2) を応用したものである．ルール 5 より，ある候補グラフ g が Q_{cur} に格納された場合，結合支持度が $f(supp(g), \phi_{min})$ を上回る g のスーパーグラフを厳密な相関値の計算をすることなく Q_{cur} に追加できることを表している．

以上の定理とルールを用いた提案手法のアルゴリズムを 3.4 節に示す．

3.4 Top-k CDS search のアルゴリズム

Top-k CDS search に対する提案手法のアルゴリズムを説明する．Algorithm 1, 2, 3 に擬似コードを示す．Algorithm 1 では DNA データベース $\mathcal{D}$，クエリシーケンス q，整数 k を受け取る．その後，3.2 節で示した DNA シーケンスと DNA グラフの等価性を利用して DNA データベース中の各シーケンスを DNA グラフに変換し，クエリの射影データベース $\mathcal{G}_{g_q}$ を取得する．次に探索中に Top-k 相関 DNA グラフとその相関値のペアを相関値の降順で格納する Q_{cur} を定義する．探索中は射影データベース $\mathcal{G}_{g_q}$ 内の 1 本のエッジからなる最小の部分 DNA グラフに対し，Algorithm 1 の 7 行目の枝刈りルールを満たすならば，Algorithm 2 を呼び出す．Algorithm 2 は，Algorithm 1 から受け取った部分 DNA グラフに 1 つのエッジを加えた部分 DNA グラフに対して深さ優先的に相関 DNA グラフの探索を行う．その際，定理 2 と各ルールを利用し候補グラフやそのスーパーグラフの枝刈りを行い，できるだけ厳密な相関値の計算をすることなく探索を進める，新たな Top-k 相関 DNA グラフが探索中に発見され，Q_{cur} に追加された際は同時に ϕ_{min} と $\varsigma(\phi_{min})$ も更新し，枝刈りの効率性を高める．候補グラフがクエリシーケンスのスーパグラフとなる場合は Algorithm 3 を呼び出し，相関値の計算をせずに相関判定をする．

Algorithm 2 MineTopCDS(g) [10]

1: **if** ($g_q \subseteq g$) **then**
2: **MineTopCDS-1**(e); (**Rule 1**)
3: **else**
4: Let G be the set of $valid$ graphs that are grown from g by adding one edge to g;
5: Compute $supp(g'; \mathcal{G}_{g_q})$ for each $g' \in G$;
6: **if** (g is a Closed frequent graph) **then**
7: Compute $supp(g)$;
8: **if** ($\phi(g_q, g) \geqq \phi_{min}$) **then**
9: Push g into Q_{cur} and refine $\varsigma(\phi_{min})$ by **Theorem 2**;
10: **for each** $g' \in G$ **do**
11: Apply **Rule 5** to determine whether to push g' into Q_{cur};
12: **if** ($\phi_{max}(g') \geqq \phi_{min}$ and $supp(g'; \mathcal{G}_{g_q}) \geqq \varsigma(\phi_{min})$) **then**
13: **if** ($supp(g)$ is computed and $supp(g) = supp(g_q, g)$) **then**
14: **MineTopCDS-1**(g');
15: **else**
16: **MineTopCDS**(g');
17: **if** ($supp(g)$ is not computed) **then**
18: Apply **Rule 2** and **Rule 3** to determine whether $supp(g)$ needs to be computed;
19: **if** ($supp(g)$ is not computed and $\phi_{max}(g) \geqq \phi_{min}$) **then**
20: Compute $supp(g)$;
21: **if** ($\phi(q, g) \geqq \phi_{min}$) **then**
22: Push g into Q_{cur} and refine $\varsigma(\phi_{min})$ by **Theorem 2**;

Algorithm 3 MineTopCDS-1(g) [10]

1: **if** ($\phi_{max}(g) \geqq \phi_{min}$) **then** (By **Rule 1** and **Rule 4**)
2: Push g into Q_{cur} and refine $\varsigma(\phi_{min})$ by **Theorem 2**;
3: Grow g by one edge;
4: **for each** $valid$ graph g' grown from g **do**
5: **MineTopCDS-1**(g');

表 2　データセットの詳細

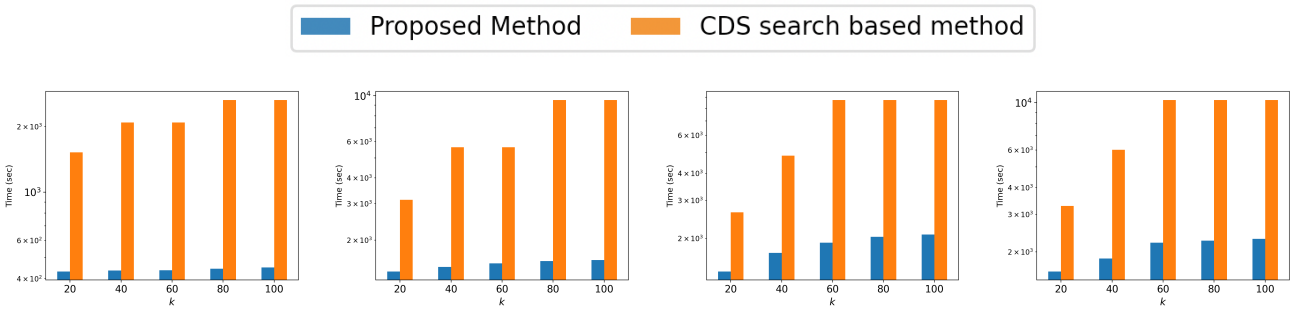
データセット	N	平均長	最小長	最大長
GEN20kS	20,000	5,000	4,829	5,109
GEN20kM	20,000	10,000	9,843	10,154
GEN20kL	20,000	20,000	19,821	20,109

4 評価実験

本節では，Algorithm 1, 2, 3 に示す提案手法の効率性と探索精度を実際の DNA データベースを用いて評価する．

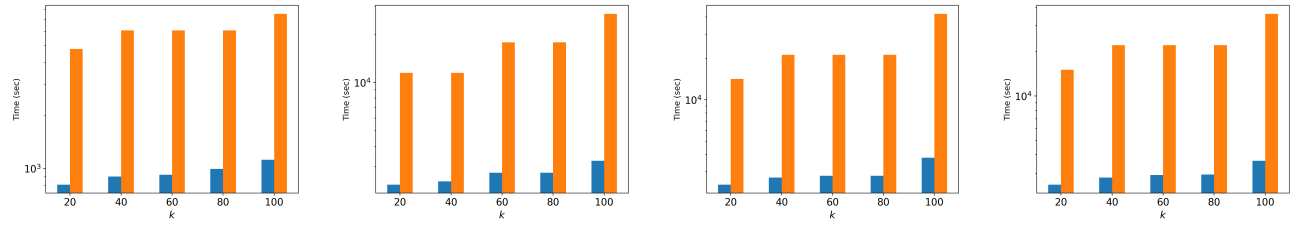
4.1 実験設定

評価実験では提案手法を以下に示す手法（CDS search based



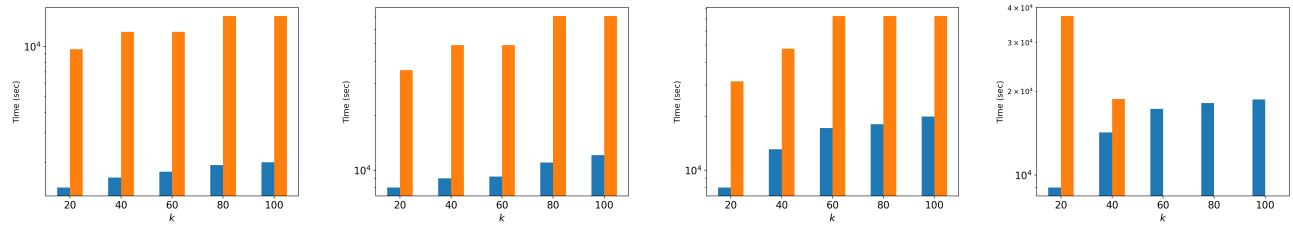
(a) $\boldsymbol{supp}(q_1) \in (0.01, 0.05)$ **(b)** $\boldsymbol{supp}(q_2) \in (0.05, 0.10)$ **(c)** $\boldsymbol{supp}(q_3) \in (0.10, 0.15)$ **(d)** $\boldsymbol{supp}(q_4) \in (0.15, 0.20)$

図 1　GEN20kS に対する問合せ処理時間 1



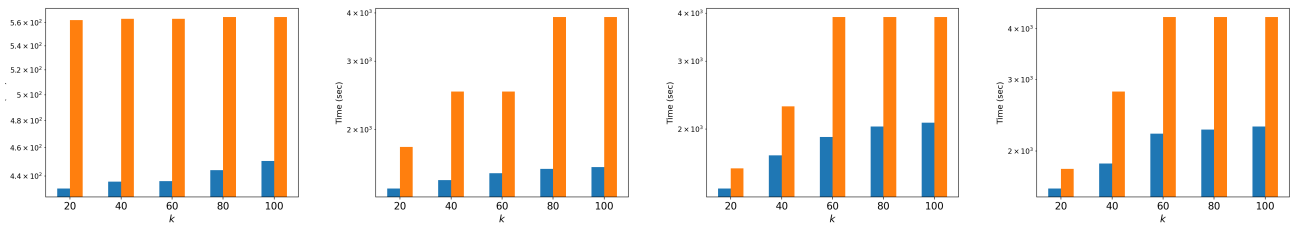
(a) $\boldsymbol{supp}(q_1) \in (0.01, 0.05)$ **(b)** $\boldsymbol{supp}(q_2) \in (0.05, 0.10)$ **(c)** $\boldsymbol{supp}(q_3) \in (0.10, 0.15)$ **(d)** $\boldsymbol{supp}(q_4) \in (0.15, 0.20)$

図 2　GEN20kM に対する問合せ処理時間



(a) $\boldsymbol{supp}(q_1) \in (0.01, 0.05)$ **(b)** $\boldsymbol{supp}(q_2) \in (0.05, 0.10)$ **(c)** $\boldsymbol{supp}(q_3) \in (0.10, 0.15)$ **(d)** $\boldsymbol{supp}(q_4) \in (0.15, 0.20)$

図 3　GEN20kL に対する問合せ処理時間



(a) $\boldsymbol{supp}(q_1) \in (0.01, 0.05)$ **(b)** $\boldsymbol{supp}(q_2) \in (0.05, 0.10)$ **(c)** $\boldsymbol{supp}(q_3) \in (0.10, 0.15)$ **(d)** $\boldsymbol{supp}(q_4) \in (0.15, 0.20)$

図 4　GEN20kS に対する問合せ処理時間 2

method）を比較した．

CDS search based method：Top-k CDS search の先駆けとして我々は文献 [13] で CDS search に対する手法を提案した．CDS search は DNA データベース $\mathcal{D}$，クエリシーケンス q，相関閾値 θ が与えられたとき相関値が閾値 θ を上回るサブシーケンスを列挙する問題である．そのため文献 [13] の提案手法に対し適切な閾値 θ を与えることで相関値が大きい上位 k 個のサブシーケンスを得ることができる．しかし相関値が大きい上位 k 個のサブシーケンスを得るのに適した θ の値はデータセットやクエリごとに異なるため検索を行う前にその値を知ることはできない．そこで $\theta = 1 - 0.1 \times i, (i = 1, 2, 3, \dots)$ とし，上位 k 個の解を得られるまでイテレーション数 i を増加させる．この手法を CDS search based method とし，Top-k CDS search に対する提案手法と比較する．

本論文では全てのアルゴリズムを C++を用いて実装し，-O3 の最適化オプションでコンパイルした．実験は Intel Xeon 2.7GHz CPU と 768GiB RDIMM を搭載した Linux サーバー上で実行した．データセットは文献 [5] で使用されている 3 つの DNA データベースを利用した．データセットの詳細を表 2 に示す．各データセットには $N = 20,000$ 個の DNA シーケンスが含

表 3　GEN20kS に対する Recall

手法	q_1	q_2	q_3	q_4
提案手法	1.0	1.0	1.0	1.0

まれており，DNA シーケンスの平均長は小さいものから順に $5,000, 10,000, 20,000$ となっている．

4.2　問合せ処理時間の評価

Top-k CDS search に対する提案手法の処理時間を評価する．表 2 の各データセットについて，支持度の異なる 4 種類のクエリを用意した．各クエリを q_1，q_2，q_3，q_4 とし，それぞれの支持度は $supp(q_1) \in (0.01, 0.05)$，$supp(q_2) \in (0.05, 0.10)$，$supp(q_3) \in (0.10, 0.15)$，$supp(q_4) \in (0.15, 0.20)$ である．

図 1，2，3 はそれぞれ GEN20kS，GEN20kM，GEN20kL に対する問合せ処理時間を表す．各図において，24 時間以内に問合せ処理を終了しなかった結果は省略されている．実験結果から分かる通り，提案手法は処理時間の観点で CDS search based method よりも高速であり，最大で 10 倍の処理速度である．これに対し，CDS search based method は図 3 に示すように DNA シーケンスの平均長が大きいとき Top-k CDS search を終えることができないケースが存在する．これはデータベースの規模が大きくなるほど，イテレーション回数が増えたときに処理時間の増加が顕著に表れるためだと考えられる．

しかし文献 [13] の実験を通じて，CDS search based method が上位 k 個の解を得るのに適した θ の値を我々は獲得済みである．そのため Top-k CDS search に対する処理時間と適切な θ を与えたときの CDS search based method の処理時間を比較することができる．その結果を図 4 に示す．図 1，2，3 の実験結果に比べ，Top-k CDS search に対する提案手法の優位性は下がるがそれでもなお処理時間の観点で優れていることがわかる．

4.3　問合せ処理精度の評価

提案手法の処理精度を評価する．文献 [13] では CDS search に対する提案手法をベースライン手法（$\hat{\mu} = 1$）と比較した実験が行われている．ベースライン手法（$\hat{\mu} = 1$）はデータベース中のすべてのサブシーケンスについて相関値を計算する手法である．すなわちベースライン手法（$\hat{\mu} = 1$）を正解データとして用いることができる．よって正解データとしてベースライン手法（$\hat{\mu} = 1$）を用い Top-k CDS search に対する提案手法の精度を評価した．

表 3 は GEN20kS に対する提案手法の *Recall* の結果を表す．表 3 から分かるように，提案手法は正解データと同様の結果を再現している．3.2 節で示したように，提案手法は CDS search を理論的にグラフベースの表現に変換し，効率的なグラフベースの枝刈りルールを導入している．加えて，[10] で議論されているように，枝刈りルールは理論的に正確な結果を得ることができる．そのため，提案手法は全てのケースにおいて正確な問合せ結果を得ることができる．

5　関 連 研 究

DNA データベースに対する類似検索はデータマイニングや生命情報科学などをはじめ，多くの分野で行われてきている [19,20]．中でも，1 節で述べたように，編集距離などを指標とする文字列ベースの検索手法 [6,20] は代表的な例の 1 つである．本節では関連研究として DNA データベースや文字列に対する類似検索に関する研究をいくつか紹介する．

インデックスベースの検索手法：インデックスベースの検索手法は DNA 類似検索の効率性を高間得る上で有用である．DNA データベース $\mathcal{D}$ に対しあらかじめインデックスを構築することで，類似検索の実行時間を減らすことができる．例えば，Chen らは近年 Burrows-Wheeler Transformation (BWT) に基づいたインデックス構築手法を提案している [19–21]．$\mathcal{D}$ に含まれる文字の種類が限られているとき BWT は効果的なので，編集距離に関する研究を拡張した k-difference 等の問題の計算量を大きく減らすことができる．PassJoin [3] や VChunkJoin [4] も同様に DNA データベース $\mathcal{D}$ のインデックスを事前計算する．特に PassJoin は DNA シーケンスをいくつかのサブシーケンスに分割し，サブシーケンスの転置インデックスを構築する．VChunkJoin もまた PassJoin に類似して，各 DNA シーケンスからチャンクと呼ばれる重複しない部分シーケンスをインデックスとして抽出する。これにより編集距離を効率的に計算するために効果的なフィルタリング手法を設計することができる．これらの手法は効果的であるが，類似検索を行うまでにインデックス構築を事前計算する必要がある．加えてこれらの手法は編集距離などの距離ベースの値を類似指標としている．そのため，これらのインデックスベースの検索手法を定義 2 で示した問題に直接適用することはできない．

ハッシュベースの検索手法：インデックスベースの検索手法とは別に，ハッシュベースの検索手法が近年提案されている [5,6]．ハッシュベースの検索手法は DNA データベースなど，文字列データベース上での編集距離計算を効率を向上させるための一般的なアプローチである．Zhang らによって提案された EmbedJoin [5] は Localtiy Sensitive Hashing（LSH） [22, 23] を文字列分割手法を組み合わせたものである．LSH を分割された部分構造に適用することで，EmbedJoin は各部分構造をハミング距離空間に射影し，ハミング距離空間で類似度を計算する．同様に MinJoin [6] は q-gram のハッシュ値を用いて DNA シーケンスを部分構造に分割し，そして類似する可能性がある全ての部分構造ペアにハッシュ結合アルゴリズムを適用する．その後 MinJoin は偽陰性を正確に除外するための検証ステップを実行する．しかし前述したようにこれらの手法は編集距離などの距離ベースの値を類似指標としている．すなわち，これらの手法を Top-k CDS search に適用することはできない．

1 節で述べたように，上記に示したような文字列ベースの検索手法は計算量の多さやサブシーケンス長の偏りによる影響を受けてしまう．そこでこれらの文字列ベースの検索手法とは異なり，本研究は DNA 類似検索の新たな問題として定義 2 で Top-k CDS search を定義した．また，DNA 類似検索問題とグラフデータベースの検索問題を結びつけることで，Top-k CDS search のための効率的なアルゴリズムを提案した．本研究の提案手法は DNA データベースに関する研究の新たな糸口になる

と考えられる．

6 結　　論

本論文では DNA 類似検索に取り組んだ．既存の文字列ベースのアプローチには編集距離計算のコストが大きい点や，文字列の部分構造間の長さに差があるときにその部分構造間に包含関係があったとしても長さに影響されて類似度を正しく評価できない点などが問題としてある．その問題を解消するため本研究では DNA 検索問題として Top-k CDS search を定義した．Top-k CDS search はユーザーが指定したクエリに対し相関値が高い部分構造を見つける問題である．Top-k CDS search は最悪の場合で $\mathcal{O}(Nm^2)$ の計算時間を必要とするため，本論文ではグラフベースの枝刈りルールに基づいた Top-k CDS search のアルゴリズムを提案した．提案手法はまず DNA シーケンスと DNA グラフの等価性を利用し DNA データベースをグラフデータベースに変換する．その後グラフベースの枝刈りルールを用いて相関値が閾値を下回る部分構造を効率的に除くことで相関部分構造を探索する．実際の DNA データベースを用いた実験により，提案手法は正確な結果を得ながら比較手法に対し最大で 10 倍の速度で処理できることを示した．

今後の課題として，提案手法で述べた各枝刈りルールが処理時間に対しどれだけの影響を及ぼしているかを検証することを考えている．また，θ が小さいとき，提案手法は未だ多くの処理時間を必要とするため，グラフ集約法 [24–26] を利用したさらなる高速化を検討している．さらに上記の改善を施した上で今回の提案手法を実際の DNA スクリーニングなどに応用することを考えている．

謝　　辞

本研究の一部は，JST さきがけ (JPMJPR2033) ならびに JSPS 科研費 (JP22K17894) の支援を受けたものである．

文　　献

[1] Yu Jiang, Guoliang Li, Jianhua Feng, and Wen-Syan Li. String similarity joins: An experimental evaluation. *Proc. VLDB Endow.*, Vol. 7, No. 8, pp. 625–636, apr 2014.

[2] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, Vol. 21, No. 1, pp. 168–173, jan 1974.

[3] Guoliang Li, Dong Deng, Jiannan Wang, and Jianhua Feng. Passjoin: A partition-based method for similarity joins. *Proc. VLDB Endow.*, Vol. 5, No. 3, pp. 253–264, nov 2011.

[4] Wei Wang, Jianbin Qin, Chuan Xiao, Xuemin Lin, and Heng Tao Shen. Vchunkjoin: An efficient algorithm for edit similarity joins. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 8, pp. 1916–1929, 2013.

[5] Haoyu Zhang and Qin Zhang. EmbedJoin: Efficient Edit Similarity Joins via Embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 585–594, 2017.

[6] Haoyu Zhang and Qin Zhang. Minjoin: Efficient edit similarity joins via local hash minima. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1093–1103, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Esko Ukkonen. Algorithms for approximate string matching. *Information and Control*, Vol. 64, No. 1, pp. 100–118, 1985.

[8] Dong Deng, Guoliang Li, Jianhua Feng, and Wen-Syan Li. Top-k string similarity search with edit-distance constraints. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 925–936, 2013.

[9] Yiping Ke, James Cheng, and Wilfred Ng. Correlation Search in Graph Databases. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD2007*, pp. 390–399, 2007.

[10] Yiping Ke, James Cheng, and Jeffrey Xu Yu. Top-k Correlative Graph Mining. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009*, pp. 1038–1049, 2009.

[11] 直井悠馬, 真次彰平, 塩川浩昭. グラフデータベースに対する高速高精度な相関問合せ. 第 15 回データ工学と情報マネジメントに関するフォーラム（DEIM2023）, 2023.

[12] Yuma Naoi and Hiroaki Shiokawa. Boosting Similar Compounds Searches via Correlated Subgraph Analysis. In *Proceedings of the 25th International Conference on Information Integration and Web Intelligence (iiWAS 2023)*, pp. 464–477, 2023.

[13] Ryuichi Yagi, Yuma Naoi, and Hiroaki Shiokawa. Fast Correlated DNA Subsequence Search via Graph-Based Representation. In *Proceedings of the 25th International Conference on Information Integration and Web Intelligence (iiWAS 2023)*, pp. 339–347, 2023.

[14] 八木隆一, 直井悠馬, 塩川浩昭. 大規模 DNA データベースに対する高速な相関問合せ手法の提案. Technical Report 9, 情報処理学会研究報告データベースシステム (DBS), 2023-DBS-177(9), 2023.

[15] 八木隆一, 直井悠馬, 塩川浩昭. DNA データベースに対する効率的な相関問合せ手法の提案. 第 15 回データ工学と情報マネジメントに関するフォーラム（DEIM2023）, 2023.

[16] Xifeng Yan and Jiawei Han. gSpan: graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 721–724, 2002.

[17] Jin Wook Kim, Eunsang Kim, and Kunsoo Park. Fast matching method for dna sequences. In Bo Chen, Mike Paterson, and Guochuan Zhang, editors, *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies*, pp. 271–281, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[18] H. T. Reynolds. *The Analysis of Cross-Classifications*. The Free Press, New York, 1977.

[19] Yangjun Chen and Yujia Wu. On the string matching with k mismatches. *Theoretical Computer Science*, Vol. 726, pp. 5–29, 2018.

[20] Yangjun Chen and Hoang Hai Nguyen. On the string matching with k differences in dna databases. *Proc. VLDB Endow.*, Vol. 14, No. 6, pp. 903–915, feb 2021.

[21] Yangjun Chen and Yujia Wu. Bwt arrays and mismatching trees: A new way for string matching with k mismatches. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 399–410, 2017.

[22] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pp. 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[23] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pp. 604–613, New York, NY, USA, 1998. Association for Computing Machinery.

[24] Hiroaki Shiokawa, Yasuhiro Fujiwara, and Makoto Onizuka. Fast Algorithm for Modularity-Based Graph Clustering. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*, 2013.

[25] Hiroaki Shiokawa, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Scaling Fine-grained Modularity Clustering for Massive Graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 4597–4604, 2019.

[26] Hiroaki Shiokawa. Scalable Affinity Propagation for Massive Datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 11, pp. 9639–9646, May 2021.

テキスト及び構造に基づく効率的な類似部分木検索手法

溝上　拓也†　　天笠　俊之††　　Savong Bou††

† 筑波大学 情報学群情報科学類　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
†† 筑波大学 計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
E-mail: †mizokami@kde.cs.tsukuba.ac.jp, ††{amagasa,savong-hashimoto}@cs.tsukuba.ac.jp

あらまし　類似部分木検索とは，ユーザが与えたクエリ木に類似した全ての部分木を巨大なデータベース木から検索する問題である．既存研究では，テキストと構造の類似度を併せた総合類似度による検索手法が提案されてきた．しかし，既存手法のテキスト類似度は，単語の一致または同義語 / 類義語のマッチングのみが考慮されており，テキスト自体の意味的な類似度を考慮していない．また構造類似度の計算コストはクエリ木のサイズの 4 乗オーダーの計算コストがかかり，実用的ではない．本研究では，テキストと構造の類似度を考慮した類似部分木検索手法を改良する．アプローチとしては，既存の手法をベースとするが， 2 つの点で改良する．1 つ目は，テキストデータの分散表現ベクトルによるテキスト類似度を用い，テキストの意味的な類似度を考慮する点である． 2 つ目は木編集距離計算手法の既存手法を用いることで，構造類似度評価をさらに高速化する点である．重複検出を想定した実験により，提案手法が実行時間・検索精度の両面で既存手法より優れていることを確認する．

キーワード　木構造データ，部分木マッチング，類似検索，木編集距離

1　はじめに

階層構造を持つデータは木構造データとして表現されることが多い．テキスト値とエレメントがネストされた XML データのように，木構造データではデータレコードとメタ情報が頂点ラベルに格納され，それらの関係性が木の構造で表現される．

本稿では，巨大なデータベース木から，ユーザが与えたクエリ木に類似した全ての部分木を検索する類似部分木検索問題に取り組む．図 1 の例では，データベース木 D 中の部分木 D' はクエリ木と類似していると見なされるため，D' を検索結果に含まれるが，D'' は類似していないと見なされ，検索結果に含まれない．類似部分木検索は，(1) プログラムコードの重複・移動検出を目的とした，抽象構文木 (AST) の類似部分の検索 [1] (2)RNA の二次構造 (ラベル付き順序木) の類似した折り畳み構造の検索 [2] (3)Web ページの階層構造を利用した，欲しい情報を含むページの検索 [3] (4) 製造業界で用いられている Bills of Materials(BOM) から類似した製品情報の検索 [4]，など様々な応用例がある．

木構造データの類似度として木の構造に基づく構造類似度やデータレコード (葉ノードのラベル) の情報に基づくテキスト類似度がある．構造類似度は，木編集距離 [5] に基づく類似度が代表的である．テキスト類似度は，葉ノードに含まれるテキストデータを元に，木構造データが表現しているコンテンツの差異を考慮した類似度である．

類似部分木検索における既存研究では，構造類似度のみを用いた手法 [6–8] と，構造類似度とテキスト類似度を併せた総合類似度による手法 [9, 10] が提案されてきた．構造類似度のみを用いた手法は，構造が同じ木データ同士におけるコンテンツの差異を考慮できないという問題がある．図 1 は，クエリ木

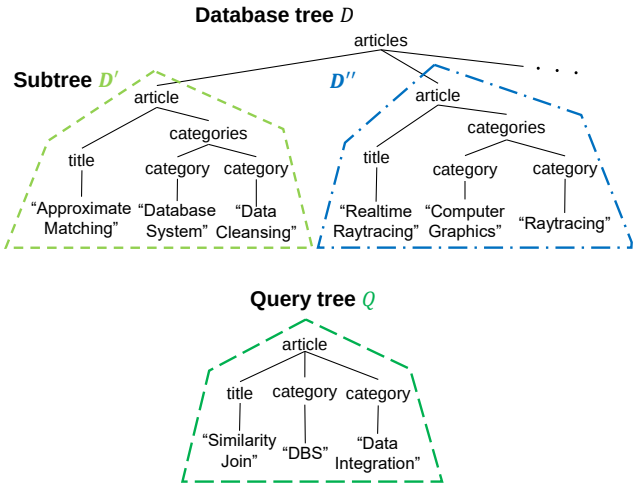


図 1　データベース木 D (上)，クエリ木 Q (下) の例．Q に対する部分木 D' (左上) と D'' (右上) の構造の差異は同じであるが，コンテンツは D' の方が D'' よりも類似している．

Q とデータベース木 D の例である．D の部分木 D' と D'' は，Q に対して構造的な差異は同じであるが，コンテンツを見ると D' は Q と類似しているが，D'' は類似していない．このような例において，構造類似度のみを用いた手法では，D' が検出されるならば D'' も検出されてしまう．このような問題を解決するため，総合類似度による類似部分木検索手法が提案されてきた．小柳ら [9] はデータレコードの単語集合のジャカード係数を，小久保ら [10] はデータレコード同士の類義語 / 同義語を考慮した最大類似度マッチングをテキスト類似度として用いている．

しかし，小柳ら小久保らの手法では，テキスト類似度として単語の一致または同義語 / 類義語のみが考慮されており，テキ

スト自体の意味的な類似度を考慮していない．また構造類似度の計算コストはクエリ木のサイズの 4 乗オーダーの計算コストがかかり，クエリ木のサイズに対してスケールしないという問題がある．

本稿では，テキストと構造の類似度を考慮した類似部分木検索手法を改良する．アプローチとしては，既存の手法をベースとするが， 2 つの点で改良する．1 つ目は，テキストデータの分散表現ベクトルによるテキスト類似度を用い，テキストの意味的な類似度を考慮する点である． 2 つ目は木編集距離計算手法の既存手法である LGM [11] や TopDiff [12] を用いることで，構造類似度評価をさらに高速化する点である．重複検出を想定した実験により，提案手法が実行時間・検索精度の両面で既存手法より優れていることを確認する．

本稿の構成は次の通りである．まず，2 章で問題定義を示し，3 章で関連研究を示す． 4 章にて事前知識を述べる．5 章で提案手法を示し，6 章で実験結果を示す．最後に 7 章で本稿をまとめる．

2 問題定義

本稿で扱う類似部分木検索問題を次のように定義する． D をデータベース木，Q をクエリ木とする．これらの木は兄弟ノード間で順序を持つ順序木であり，各頂点にはラベルが付与されており，根付き有向木である．頂点数はそれぞれ $|Q|, |D|$ と表す．D_i をある頂点 v_i を根とし，v_i とその子孫を全て含んだ D の部分木とする．Q と D_i の類似度は，構造類似度 $\mathrm{sim_s}(\cdot,\cdot)$ とテキスト類似度 $\mathrm{sim_t}(\cdot,\cdot)$，サイズ差によるペナルティ $\mathrm{d}(\cdot,\cdot)$，重み係数 $\alpha \in (0,1]$ を用いて次のように定義する．

$$\mathrm{sim}(Q,D_i) = \alpha \mathrm{sim_s}(Q,D_i) + (1-\alpha)\frac{\mathrm{sim_t}(Q,D_i)}{1+\mathrm{d}(Q,D_i)} \quad (1)$$

$$\left(\mathrm{d}(Q,D_i) = \frac{||Q|-|D_i||}{\min\{|Q|,|D_i|\}}\right)$$

$\theta \in (0,1]$ を類似度しきい値とする．本稿では類似部分木検索を，D の全ての部分木集合 $Subtrees(D)$ の中から，Q との類似度が θ 以上であるような全ての部分木の集合 $\mathcal{R}$ を求める問題とする．

$$\mathcal{R} = \{D_i | D_i \in Subtrees(D) \wedge \theta \leqq \mathrm{sim}(Q,D_i)\} \quad (2)$$

類似度の定義で用いられている構造類似度とテキスト類似度については，4 章で詳しく述べる．

3 関連研究

本章では関連研究として，木編集距離及びその上界・下界値の計算手法，類似部分木検索手法とそれに関連する類似木結合手法，テキスト類似度について述べる．

3.1 木編集距離

順序木の編集距離は，多項式時間で計算できることが知られている． Tai [13] によって最初の多項式時間の動的計画法アルゴリズムが提案された．時間計算量は 2 つの木の大きい方のノード数を n として， $O(n^6)$ である．その後，Zhang と Shasha [5] がこの動的計画法アルゴリズムの全ての部分問題を解く必要が無いという洞察により $O(n^4)$ に改善された．その後 Klein [14] が *Heavy Path Decomposition* に基づく最適化により $O(n^3 \log n)$ に，Demaine ら [15] によるさらに効率的な *decomposition* により $O(n^3)$ に改善された．その後，Pawlik と Augsten らにより，実用的に良いパフォーマンスを発揮する $O(n^3)$ アルゴリズムの RTED [16] が提案され，RTED のメモリ使用量を改善した APTED [17] が提案された．これらのアルゴリズムは，ノード数が大きく複雑な構造を持つ木の間の木編集距離を計算するのには適していない．

前述の木編集距離の計算コストを軽減するために，しきい値付き木編集距離の計算手法が提案されている．しきい値付き木編集距離の計算とは，木編集距離がしきい値 τ 以下であるならば正しい木編集距離を計算し，そうでないなら ∞ を返すという問題である．Touzet [18] は Zhang と Shasha [5] の動的計画法アルゴリズムの探索空間をしきい値 τ に基づき制限することで，時間計算量 $O(n\tau^3)$，空間計算量 $O(n\tau)$ のアルゴリズムを提案した．Pawlik と Augsten ら [12] は，Touzet のアルゴリズムを改良し，実用的なパフォーマンスが向上した TopDiff を提案した．これらすべてのアルゴリズムは，より軽量に計算できる下界値，上界値による枝刈りを行い，ある程度類似しているとされる木同士の編集距離を計算する際に有用である．本稿の手法では，次に述べる下界値と上界値による枝刈りを行った後に，TopDiff [12] によるしきい値付き木編集距離の計算を行い，従来手法よりも高速に構造類似度を評価する．

3.2 木編集距離の下界

Li ら [19] は，木編集距離の下界値の計算方法を調査し，それらの性能を比較した．比較されているのは，ノードを先行/後行順に並べた文字列の編集距離による下界値 [20]，ノードのラベル分布や頂点次数分布のヒストグラムの L1 距離による下界値 [21]，二分木変換した木の二分枝 (Binary Branch) のマルチセットの差分による下界値 [22] の 3 つである．[19] では，実験により，文字列編集距離による下界値が最もタイトであることを示している．Binary Branch と似た考えの下界値として，木を pq-gram のトークン集合に変換し，下界値に用いる手法が [23] によって提案されているが，編集コストを頂点のファンアウト数によって重みを付けた編集距離の下界値であり，本稿で用いている編集距離には適用できない．[11] では，木のラベルのマルチセットを用いたラベル下界をさらにタイトにした構造フィルタという下限値が提案されているが，この構造フィルタはインデックスベースの類似木結合アルゴリズムに組み込まれており，インデックスフリーの本手法には適用できない．本稿では，タイトで高速に計算できる，しきい値付き文字列編集距離による下界値を用いた．

3.3 木編集距離の上界

木編集距離の上界値を計算するアルゴリズムは多く提案されており，Kan ら [24] によってそれらが比較されている．[24] で

比較されているアルゴリズムは全て時間計算量が $O(n^2)$ であり，サイズの大きい木には適していない．また，その中で最も二乗オーダーより低い空間計算量のアルゴリズムは Wang ら [25] の $O(n \log n)$ のみである．ノード数 n に対して線形時間で上界値を計算する LGM アルゴリズムは，Hütter ら [11] の類似部分木結合アルゴリズムの一部として提案された． LGM アルゴリズムは，しきい値を τ として，$O(n\tau)$ の時間計算量と空間計算量で上界値を計算できる．LGM は，同じラベルを持つ頂点ペアを貪欲に編集マッピングに追加し，その後ラベルの異なる頂点ペアを貪欲に編集マッピングに追加する処理を，マッピング条件 (祖先–子関係や兄弟関係) を満たしながら行っている．マッピング可能な頂点ペアのうち，マッピングコストが τ を超えてしまうような頂点ペアを除外することで，計算量を削減している．本稿では，木のノード数 n に対して線形時間で上界値を計算できる LGM アルゴリズムを用いた．

3.4 類似部分木検索

木編集距離を距離尺度とした類似部分木検索の手法では，TASM [6] や StructureSearch [7]，SlimCone [8] が提案されている．TASM は，データベース木を走査して類似部分木を探すインデクスフリーの手法であり，StructureSearch，SlimCone は，データベース木をインデックス化して類似部分木を探す手法である．実行時間，インデックス作成時間，インデックスサイズ全てにおいて最も優れているのは SlimCone である．これらの手法は，構造的な類似度のみを考慮しているため，1 章の図 1 の例で示したような，同じ編集距離であって内容が異なる部分木を区別できない．この問題を解決するため，小柳ら [9] や小久保ら [10] は，木編集距離に基づく構造類似度とテキスト類似度の両方を考慮した類似部分木検索手法を提案している．この手法は，TASM と同様にデータベース木を走査して類似部分木を探すインデックスフリーの手法である．本稿では，この手法をベースにした手法を提案する．

3.5 類似木結合

類似木結合は，類似部分木検索の一般化であり，木構造データの集合から類似する木のペアを列挙する問題である．類似度指標を木編集距離とした結合アルゴリズムとして，決定的なアルゴリズムでは TJoin [11] が，確率論的なアルゴリズムでは SyncSignature [26] が最も優れている．また，木編集距離が最も低い木のペアを列挙する Top-k 類似木結合アルゴリズムとして，Top-KTJoin [27] が提案されている．また，これらの順序木を対象としたアルゴリズムとは異なり，JSON フォーマットのような兄弟間の順序が不定なオブジェクト型と固定順序の配列型のノードを持つ木構造データを対象とした類似木結合アルゴリズムとして JEDI [28] が提案されている．これらの手法は，構造的な類似度のみを考慮しているため，本研究が扱う問題には適用できない．

3.6 テキスト類似度

類似部分木検索におけるテキストデータの類似度として，小柳ら [9] は葉ノードに含まれるテキストデータに含まれる単語集合のジャカード係数を，小久保ら [10] は葉ノードの同義語 / 類義語を考慮した最大類似度マッチングを用いている．これらの手法は，テキストノードをこれらの手法は，単語の一致または同義語 / 類義語のマッチングのみが考慮されており，テキスト自体の意味を考慮できていない点で，本稿の手法と異なる．

4 事 前 知 識

本章では事前知識として，木編集距離と構造類似度，テキスト類似度について述べる．

4.1 木編集距離

二つの木 T，T' の間の木編集距離 $TED(T, T')$ は，T から T' への最小の編集操作の回数である．ここでは編集操作として，頂点の置換，削除，挿入の 3 つを想定している．①頂点の置換は，頂点 v のラベルを別のラベルに置き換える操作である．②頂点の削除は，頂点 v を削除し，v の子孫を元々 v がつながっていた場所につなげる操作である．③頂点の挿入は，頂点 u とその子孫の i 番目から j 番目までの子 $C = \{c_i, \ldots, c_j\}$ の間に新しい頂点 v を挿入し，C を v の子にする操作である．例えば，図 1 のクエリ木 Q をデータベース木の部分木 D' に変換するのに，“article” ノードと “category” ノードの間に “categories” ノードを挿入し，葉ノードのラベル 3 つを置換する必要がある．そのため，$TED(Q, D') = 4$ になる．同様に D' の右側の部分木 D'' についても，$TED(Q, D'') = 4$ になる．

4.2 構造類似度

クエリ木 Q と部分木 D_i の間の構造類似度 $\mathrm{sim_s}(Q, D_i)$ は，木編集距離 $TED(Q, D_i)$ をその最大値 $|Q| + |D|$ で正規化した距離を用いて，以下のように定義される．

$$\mathrm{sim_s}(Q, D_i) := 1 - \frac{TED(Q, D_i)}{|Q| + |D_i|} \tag{3}$$

図 1 の例では，$TED(Q, D') = 4$，$TED(Q, D'') = 4$，$|Q| = 7$，$|D_9| = |D_{19}| = 8$ であるから， $SIM_S(Q, D') = SIM_S(Q, D'') = 1 - 4/15 = 0.73$ となる．

4.3 テキスト類似度

クエリ木 Q と部分木 D_i の間のテキスト類似度 $\mathrm{sim_t}(Q, D_i)$ は，それぞれの代表ベクトル $\mathbf{dv}(Q)$，$\mathbf{dv}(D_i)$ を用いて，以下のように定義される．

$$\mathrm{sim_t}(Q, D_i) := \frac{1}{2} \times \left(\frac{\mathbf{dv}(Q) \cdot \mathbf{dv}(D_i)}{|\mathbf{dv}(Q)||\mathbf{dv}(D_i)} + 1 \right) \tag{4}$$

木の代表ベクトルは図 2 のように，根から各葉ノードまでの経路上のノードのラベルを結合し，得られた結合文字列をそれぞれ分散表現ベクトルに埋め込み，その平均を取ることで得られる．

5 提 案 手 法

本章では，小柳ら [9] 小久保ら [10] の手法をベースにして改良を加えた類似部分木検索手法について述べる．

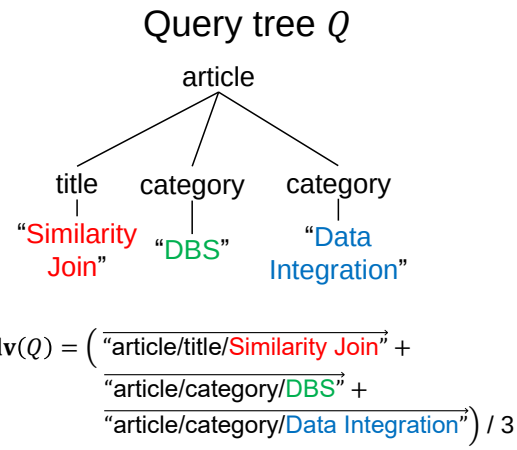


図 2　埋め込みベクトルによるテキスト類似度の計算例

5.1　全体的な処理の流れ

全体的な処理の流れは既存手法と同様に，データベース木からノードを後行順に読み込み，そのノードを根とする部分木とクエリ木の類似度を計算して類似部分木を探す，という流れを全てのノードに対して行う．また，既存手法と同様に，以下の 2 つの枝刈りを行い，類似度計算を省くことで高速化を図る．

（1） **木のノード数による枝刈り**　ノード数がクエリ木 Q に対して大きすぎる / 小さすぎる部分木 D_i は類似しえないため，そのような部分木を省く枝刈り．

（2） **テキスト類似度による枝刈り**　構造類似度よりも軽量に計算できるテキスト類似度を先に計算し，その時点で類似度しきい値を満たせないことが判明した部分木を除外し，構造類似度の計算を省く枝刈り．

これら 2 つの枝刈りの詳細については，5.3 章で述べる．

既存手法と異なる点は，構造類似度の計算を改良した点である．

5.2　構造類似度の計算に関する工夫

既存手法では，構造類似度を求める際に計算する木編集距離の計算コストが $O(|Q|^4)$ であるため，クエリ木のサイズが大きい場合に高速に計算できない．そのため，これを改善したいという動機がある．そこで本稿では，先の 5.1 章で述べた 2 つの枝刈りを通過した候補部分木はクエリ木に対してサイズとテキストが類似しているため，類似した木同士の木編集距離をクエリ木のサイズに対して線形時間で計算できるアルゴリズム [11, 12, 20] を用いることで高速化を図る．

実際の計算手順について説明する．5.1 章で述べた 2 つの枝刈りを通過した部分木 D_i について，既にテキスト類似度 $\mathrm{sim_t}(Q, D_i)$ が計算されているため，その値を用いて類似度条件式 $\mathrm{sim}(Q, D_i) \geqq \theta$ を満たす場合の木編集距離のしきい値 τ を求めることができ，次のように計算できる．

$$\theta \leqq \alpha \left(1 - \frac{TED(Q, D_i)}{|Q| + |D_i|}\right) + (1 - \alpha)\frac{\mathrm{sim_t}(Q, D_i)}{1 + \mathrm{d}(Q, D_i)}$$
$$\beta = 1 - \frac{\theta - (1 - \alpha)\mathrm{sim_t}(Q, D_i)/(1 + \mathrm{d}(Q, D_i))}{\alpha}$$
$$TED(Q, D_i) \leqq \beta\,(|Q| + |D_i|)$$
$$\tau = \beta\,(|Q| + |D_i|) \qquad (5)$$

構造類似度の計算の際には，この距離しきい値 τ を用いて，以下の手順で計算を行う．

（1） **木編集距離の下界値による枝刈り**　木のノードを先行順 / 後行順に並べた配列間の文字列編集距離の最大値 [20] を木編集距離の下界値として計算し，距離しきい値 τ を超える部分木を枝刈りする．時間計算量は $O(\min\{|Q|, |D_i|\}\,\tau)$，空間計算量は $O(\tau)$ である．

（2） **木編集距離の上界値による枝刈り**　Thomas らが提案した LGM [11] による木編集距離の上界値を計算し，距離しきい値 τ 以下の部分木を結果に出力し，後の構造類似度の計算を省く．時間計算量は $O(\max\{|Q|, |D_i|\}\tau)$，空間計算量は $O(\max\{|Q|, |D_i|\})$ である．

（3） **しきい値付き木編集距離の計算**　残った候補部分木に対して，Mateusz らが提案した TopDiff [12] によるしきい値付き木編集距離の計算を行う．$n = \max\{|Q|, |D_i|\}$，$d = \max\{depth(Q), depth(D_i)\}$，$l = \max\{|\mathcal{L}_Q|, |\mathcal{L}_{D_i}|\}$ として，時間計算量は $O(nd\tau \min\{l, \tau\})$，空間計算量 $O(n\tau)$ である．

5.3　アルゴリズム

提案手法のアルゴリズムを Algorithm 1 に示す．Algorithm 1 の概要を説明する．

Alg. 1 では，小柳ら小久保ら [9, 10] と同様に，データベース木のノードを後行順に読み込む際に後行順キュー [6] を用いる．後行順キューは，図 3 のように，データベース木のノードを後行順に並べた配列を保持するキューである．キューの各要素は，ノードのラベルとそのノードを根とする部分木のサイズ，及びそのノードの分散表現ベクトルを保持する．ただし，分散表現ベクトルは葉ノードの場合のみ保持し，それ以外のノードについては零ベクトルが保持されている．後行順キューは，データベース木を走査する際に必要に応じて動的に末尾にノードを追加するため，省メモリで構築できる．

検証は以下の手順で行う．

1） **木のノード数の最小値，最大値の計算** (Alg.1, 2 行目)
5.1 で述べた木のノード数による枝刈りにおいて用いる，類似部分木のサイズ範囲 $[\mathrm{minD_i}, \mathrm{maxD_i}]$ を計算する．木編集距離の最小値 $||Q| - |D_i|| \leq TED(Q, D_i)$ [8] と，テキスト類似度の最大値 1 を用いて，候補部分木 D_i が満たす条件は次のように表される．$\mathrm{UB_s}(Q, D_i)$，$\mathrm{UB_t}(Q, D_i)$，$\mathrm{UB}(Q, D_i)$ はそれぞれ構造類似度の上界値，テキスト類似度の上界値，総合類似度の上界値を表す．

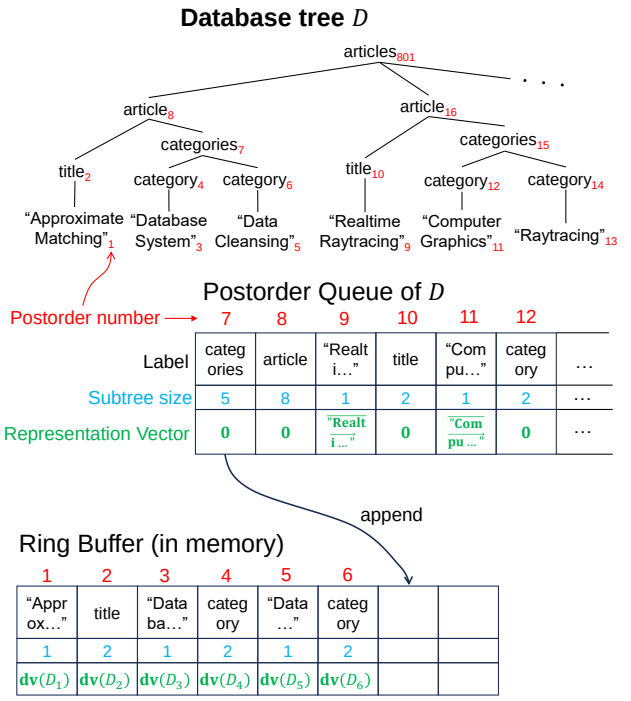


図 3　後行順キューによるストリーム処理

$$\mathrm{UB_s}(Q, D_i) = 1 - \frac{||Q| - |D_i||}{|Q| + |D_i|}$$

$$\mathrm{UB_t}(Q, D_i) = 1$$

$$\mathrm{UB}(Q, D_i) = \alpha \mathrm{UB_s}(Q, D_i) + (1 - \alpha)\frac{\mathrm{UB_t}(Q, D_i)}{1 + \mathrm{d}(Q, D_i)}$$

$$\theta \leqq \mathrm{UB}(Q, D_i) \qquad (6)$$

この条件を満たす部分木 D_i のサイズ範囲 $[\mathrm{minD_i}, \mathrm{maxD_i}]$ を計算すると，式 (7) の $\mathrm{minD_i}$ と $\mathrm{maxD_i}$ が得られる．

$$\mathrm{minD_i} = \begin{cases} \left\lceil \frac{-(1+\alpha-\theta)+\sqrt{(1+\alpha-\theta)^2+4\theta(1-\alpha)}}{2(1-\alpha)}|Q| \right\rceil & \alpha \neq 1 \\ \left\lceil \frac{\theta}{2-\theta}|Q| \right\rceil & \alpha = 1 \end{cases}$$

$$\mathrm{maxD_i} = \left\lfloor \frac{-(\theta-1-\alpha)+\sqrt{(\theta-1-\alpha)^2+4\theta(1-\alpha)}}{2\theta}|Q| \right\rfloor$$

$$\mathrm{minD_i} \leqq |D_i| \leqq \mathrm{maxD_i} \qquad (7)$$

2) **リングバッファの初期化** (Alg.1，3 行目)

後行順キューから読み込んだデータベース木のノードを最大で $\mathrm{maxD_i}$ 個保持するリングバッファ R を初期化する．リングバッファのキャパシティは $\mathrm{maxD_i} + 1$ であり，リングバッファの先頭と終端の番兵のインデックスをそれぞれ s，e で表す (4 行目)．

リングバッファを用いて後行順配列の一部のみをメモリに保持するのは，非常に巨大なデータベース木に対しても，メモリの使用量を抑えるためである．後行順で i 番目のノードを根とする部分木 D_i の頂点情報は，後行順配列の $i - |D_i| + 1$ 番目から i 番目の連続した部分列に記録されている．リングバッファ上に最大 $\mathrm{maxD_i}$ 個の連続した後行順部分列のみ保持することで，サイズ範囲内の部分木 D_i を漏れなく抽出することができる．

3) **候補部分木の取得** (Alg.1，8 行目)

後行順キューからノードを読み込みリングバッファに格納し，サイズ範囲内の部分木 D_i を取得する．この処理は Algorithm 2 に示す pull_candidate_subtree 関数によって行う．

pull_candidate_subtree 関数内では，部分木の代表ベクトル $\mathbf{dv}(D_i)$ の計算も行う (Alg. 2，8〜9 行目)．具体的には，リングバッファに格納されたノードが葉ノードでない場合，D_i の子部分木の代表ベクトルが計算済みでリングバッファに保持されていることを利用して，それぞれの部分木の葉ノードの数を重みとした子部分木の代表ベクトルの加重平均を計算する．図 3 の例では，後行順キューから後ろ掛け 7 番目のノード $\mathrm{categories_7}$ をリングバッファに格納したとき，そのノードを根とする部分木 D_7 の代表ベクトルは，子部分木である D_4(葉ノードの数 1) と D_6(葉ノードの数 1) の代表ベクトルの加重平均 $(\mathbf{dv}(D_4) \times 1 + \mathbf{dv}(D_6) \times 1)/(1 + 1)$ で求められる．全体を通した代表ベクトルの計算時間はベクトルの次元を d として，$O(d|D|)$ である．

4) **テキスト類似度の計算** (Alg.1, 13 行目)

クエリ木 Q と候補部分木 D_i のテキスト類似度 Sim_t を計算する．

5) **テキスト類似度によるフィルタリング** (Alg.1，14〜16 行目)

クエリ木 Q と候補部分木 D_i のテキスト類似度 Sim_t を計算し，Sim_t のみで類似度条件を満たす場合は，その後の構造類似度の計算を省略して，候補部分木 D_i を類似部分木として出力する．具体的には，式 (6) のテキスト類似度の上界値 $\mathrm{UB_t}(Q, D_i)$ を実際に計算した Sim_t に置き換えてよりタイトな上界値 $SimUB$ を計算し (14 行目)，θ 未満になる場合は類似しえない部分木であると判定して，構造類似度の計算を省略する．

6) **構造類似度の評価** (Alg.1，18〜23 行目)

5.2 章で述べた手順に従って，候補部分木 D_i に対して構造類似度の計算を行う．式 (5) のしきい値 τ を計算し (18 行目)，下界値によるフィルタリング (19 行目)，上界値によるフィルタリング (20 行目) およびしきい値付き木編集距離の計算 (22 行目) を行う．

6 評 価 実 験

本章では，5 章で提案した類似部分木検索アルゴリズムの有用性を確認するために，重複検出を想定した実験を行い，検索速度と検索精度を評価する．

6.1 実 験 環 境

実験は全て，Intel Core i7-8700K 3.20GHz，メモリ 16GB

Algorithm 1 Subtree-Similarity-Search(Q, pq, θ, α)

Input: Query tree Q, postorder queue pq of a document tree D, similarity threshold $\theta \in (0, 1]$, weight coefficient for structural similarity $\alpha \in [0, 1]$.

Output: $\mathcal{O} = \{D_i | sim(Q, D_i) \geqq \theta\}$.

$\mathcal{O} \leftarrow \emptyset$
Calculate $\mathrm{minD_i}$ and $\mathrm{maxD_i}$ from $|Q|$, θ, α.
$R \leftarrow$ Ring buffer of size $\mathrm{maxD_i} + 1$.
$s \leftarrow 1$, $e \leftarrow 1$ // Start/End index of ring buffer R.
/* Ring buffer R stores s-th to $(e-1)$-th elements of pq. */
$Rsize \leftarrow \mathrm{maxD_i} + 1$ // Capacity of ring buffer R.
while true **do**
  $D_i, s, e \leftarrow$ pull_candidate($pq, R, Rsize, s, e, \mathrm{minD_i}, \mathrm{maxD_i}$)
  $Sim_t \leftarrow 0$
  if $D_i = \emptyset$ **then**
    break
  if $\alpha < 1$ **then**
    $Sim_t \leftarrow \mathrm{sim_t}(Q, D_i)$
  $SimUB \leftarrow \alpha \left(1 - \frac{||Q|-|D_i||}{|Q|+|D_i|}\right) + (1-\alpha)\frac{Sim_t}{1+d(Q,D_i)}$
  if $SimUB < \theta$ **then**
    continue
  else if $0 < \alpha$ **then**
    $\tau \leftarrow \left\lfloor \left(1 - \frac{\theta-(1-\alpha)Sim_t/(1+d(Q,D_i))}{\alpha}\right)(|D_i| + |Q|) \right\rfloor$
    if $\mathrm{LowerBoundED}(Q, D_i, \tau) \leqq \tau$ **then**
      if $\mathrm{UpperBoundED}(Q, D_i, \tau) \leqq \tau$ **then**
        $\mathcal{O} \leftarrow \mathcal{O} \cup \{D_i\}$
      else if $\mathrm{TED}(Q, D_i, \tau) \leqq \tau$ **then**
        $\mathcal{O} \leftarrow \mathcal{O} \cup \{D_i\}$

Algorithm 2 pull_candidate(pq,R,$Rsize$,s,e,$\mathrm{minD_i}$,$\mathrm{maxD_i}$)

Input: Postorder queue of a document tree pq, ring buffer R, capacity of ring buffer $Rsize$, start/end index of ring buffer s and e, minimum/maximum size of candidate subtree $\mathrm{minD_i}$ and $\mathrm{maxD_i}$.

Output: Candidate subtree D_i, start/end index of ring buffer s and e.

$D_i \leftarrow \emptyset$
while $pq \neq \emptyset$ **do**
  /* λ : label. $size$: size of subtree. $\boldsymbol{v}$: embedding vector. */
  $(pq, (\lambda, size, \boldsymbol{v})) \leftarrow \mathrm{dequeue}(pq)$
  $R[e \mod Rsize] \leftarrow (\lambda, size, \boldsymbol{v})$ // Push $(\lambda, size, \boldsymbol{v})$ to R.
  $i \leftarrow e$
  $e \leftarrow e + 1$
  if $size \leqq \mathrm{maxD_i}$ **and** $size > 1$ **then**
    compute $\mathbf{dv}(D_i)$ from $\mathbf{dv}$ of D_i's children.
  if $\mathrm{minD_i} \leqq size$ **and** $size \leqq \mathrm{maxD_i}$ **then**
    $lmld \leftarrow i - size + 1$ // The index of the leftmost leaf descendant of D_i.
    /* Extract the subtree rooted at i-th node. */
    $D_i \leftarrow \mathrm{extract_subtree}(R, lmld, i)$
  if $e - s \geqq \mathrm{maxD_i}$ **then**
    $s \leftarrow s + 1$ // Pop the oldest element from R.
  if $D_i \neq \emptyset$ **then**
    break

の PC 上で行った．OS は Ubuntu 24.04 LTS，プログラムは全て C++17 で実装し，コンパイラは g++ 13.2.0 を用い，最適化オプション-O3 を付与してコンパイルした．

6.2 データセット

データセットとして，医学を中心とする生命科学の文献情報を収集したオンラインデータベースである MEDLINE[1] を使用した．

6.3 実験設定

実験は重複検出を想定して，ノイズが加えられたクエリ木を複数含んだデータベース木を用意した．具体的には，MEDLINE からランダムに選んだ 1 つの文献データ (部分木) をクエリ木とし，データベース木にはクエリ木にノイズが加えられた文献データ 1000 件を適合文書とし，クエリ木以外の文献データ 29000 件を非適合文書として含めた．データベース木内の適合文書については，クエリの各テキストノードのラベルに対して 80%の確率で以下の 2 つの操作のいずれかを行った．

（1） ラベルに含まれる単語をランダムに 1 つ削除する．

（2） ラベルに対して類似度 0.5 以上の同義語 / 類義語をランダムに 1 つ選び，それに置き換える．

同義語 / 類義語の選択には，MEDLINE の MeSH(Medical Subject Headings) 辞書を用いた [10]．同義語 / 類義語間の類似度は，[29] で示されている方法で計算した．生成されたクエリ木とデータベース木とクエリ木の統計情報を表 1，表 2 に示す．

表 1 データベース木の統計情報

Size T [MB]	# Nodes (# Subtrees)	Fanout min / avg / max	Height min / avg / max	# Words	# Words in a text node min / avg / max
211.9	9,430,379	0 / 1 / 30,000	1 / 6.5 / 14	9,629,992	1 / 1.02 / 957

表 2 クエリ木の統計情報

Size T [KB]	# Nodes (# Subtrees)	Fanout min / avg / max	Height min / avg / max	# Words	# Words in a text node min / avg / max
7.6	288	0 / 1 / 11	1 / 5.36 / 8	414	1 / 1.44 / 244

テキストノードの分散表現ベクトルは，NLP ライブラリである Spark-NLP [30] を用い，MEDLINE データセットで訓練された BERT モデル [2] を用いて計算した．生成された分散表現ベクトルの次元数は 768 次元である．

6.4 比較手法

比較手法は，小柳らの手法 [9] と小久保らの手法 [10] を用いた．小久保らの手法では，テキスト類似度を評価するのに同義語 / 類義語の知識ベースが必要である．本稿では，[10] と同様に同義語ルールに MeSH の Entry Term を用い，類義語の知識ベースとして MeSH の概念階層を用いた．

1：https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

2：https://sparknlp.org/2021/08/31/sent_bert_pubmed_en.html

6.5 実験手順

実験では，類似度しきい値 θ を 1.0 から 0.01 刻みで変化させ，類似部分木検索を行った際の実行時間と再現率，適合率を評価した．

6.6 実験結果

実験結果を図 4，図 5 に示す．小久保らの手法が図には含まれていないが，全ての再現率において実行時間がタイムアウト時間の 50 秒を超えたため，図から省いている．また適合率については，再現率 0〜1 を通して全て 1.0 であったため，図には含めていない．

図 4 は，再現率に対する実行時間の比較である．一般的に，再現率が上がるほど候補部分木が増え，類似度評価のコストがかかるため，実行時間が増加する傾向がある．実行時間については，0.2 以上の再現率で提案手法が既存手法より大幅に速く実行可能であり，特に再現率 1.0 において 2 倍以上の速度向上が見られる．

図 5 は，再現率に対する構造類似度の計算時間の比較である．全ての再現率を通して，提案手法の計算時間が既存手法より大幅に短いことが分かる．これは，構造類似度の計算アルゴリズムが $O(|Q|^4)$ から $O(|Q|\tau)$ $(\tau \propto 1-\theta)$ に改善され，一回の構造類似度の計算コストが大幅に削減されたためである．

図 6 と図 7 は，提案手法と小柳らの手法における，再現率に対する木編集距離の計算回数をプロットしたものである．提案手法の方が全体を通して木編集距離の計算回数が多いが，内訳をみると文字列編集距離による下界値フィルタと，LGM による上界値フィルタによって構造類似度の評価がすでに終了しており，一回当たりの評価が $O(|Q|\tau)$ で行えていることが分かる．

検索精度については，再現率 0〜1 を通して 1.0 であり，検索精度において差が見られなかった．

以上から，提案手法は，既存手法と比較して速度および精度の両面で優れていることが示せた．

7 まとめ

本稿では，構造類似度とテキスト類似度を組み合わせた類似部分木検索において，分散表現ベクトルによるテキスト類似度の評価と，構造類似度評価の高速化を行い，従来の手法を改良した．実験の結果，提案手法は，既存手法と比較して速度および精度の両面で優れていることがわかった．

今後の課題は，既存のベクトル検索手法を用いた，インデクスベースの類似部分木検索アルゴリズムの提案である．

謝辞

本研究の成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の 「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業，JST CREST（JPMJCR22M2），科学研究費補助金（JP22H03694, JP23K16888）に支援によるものである．

文献

[1] Jean-Rémy Falleri, Floréal Morandat, Xavier Blanc, Matias Martinez, and Martin Monperrus. Fine-grained and accurate source code differencing. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*, ASE '14, p. 313–324, New York, NY, USA, 2014. Association for Computing Machinery.

[2] Claire Herrbach, Alain Denise, and Serge Dulucq. Average complexity of the jiang–wang–zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm. *Theoretical Computer Science*, Vol. 411, No. 26, pp. 2423–2432, 2010.

[3] Márcio L. A. Vidal, Altigran S. da Silva, Edleno S. de Moura, and João M. B. Cavalcanti. Structure-driven crawler generation by example. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, p. 292–299, New York, NY, USA, 2006. Association for Computing Machinery.

[4] Mohamed Kashkoush and Hoda ElMaraghy. Matching bills of materials using tree reconciliation. *Procedia CIRP*, Vol. 7, pp. 169–174, 2013. Forty Sixth CIRP Conference on Manufacturing Systems 2013.

[5] K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, Vol. 18, No. 6, p. 1245–1262, dec 1989.

[6] Nikolaus Augsten, Denilson Barbosa, Michael Böhlen, and Themis Palpanas. Tasm: Top-k approximate subtree matching. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 353–364, 2010.

[7] Sara Cohen. Indexing for subtree similarity-search using edit distance. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, p. 49–60, New York, NY, USA, 2013. Association for Computing Machinery.

[8] Daniel Kocher and Nikolaus Augsten. A scalable index for top-k subtree similarity queries. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, p. 1624–1641, New York, NY, USA, 2019. Association for Computing Machinery.

[9] 小柳 涼介, 天笠 俊之, 北川 博之. テキストおよび構造の類似度に基づいた XML データに対する効率的な類似検索. DEIM2014 Forum, Jul 2014. 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014).

[10] 柚真小久保, 俊之天笠, 博之北川. 複数の類似度を考慮した木構造データに対する類似部分木検索. 第 83 回全国大会講演論文集, 第 2021 巻, pp. 349–350, mar 2021.

[11] Thomas Hütter, Mateusz Pawlik, Robert Löschinger, and Nikolaus Augsten. Effective filters and linear time verification for tree similarity joins. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 854–865, 2019.

[12] Mateusz Pawlik and Nikolaus Augsten. Minimal edit-based diffs for large trees. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, p. 1225–1234, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Kuo-Chung Tai. The tree-to-tree correction problem. *J. ACM*, Vol. 26, No. 3, p. 422–433, jul 1979.

[14] Philip N. Klein. Computing the edit-distance between unrooted ordered trees. In *Proceedings of the 6th Annual European Symposium on Algorithms*, ESA '98, p. 91–102, Berlin, Heidelberg, 1998. Springer-Verlag.

[15] Erik D. Demaine, Shay Mozes, Benjamin Rossman, and Oren Weimann. An optimal decomposition algorithm for tree edit distance. *ACM Trans. Algorithms*, Vol. 6, No. 1, dec 2010.

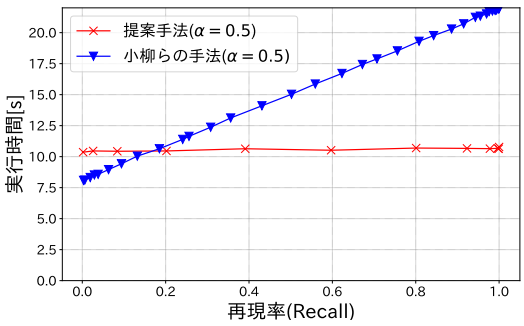


図 4　再現率に対する実行時間の比較 (適合率 0.99 以下の点は適合率を明記)

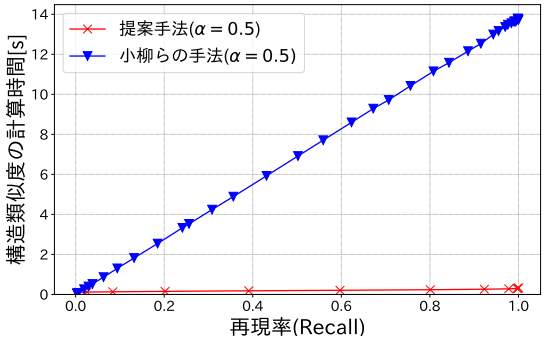


図 5　再現率に対する構造類似度の計算時間の比較

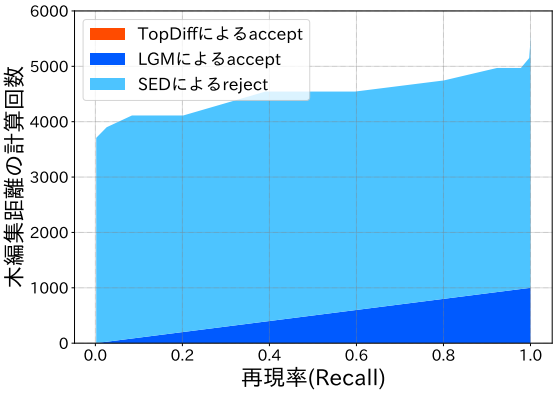


図 6　提案手法における，再現率に対する木編集距離の計算回数

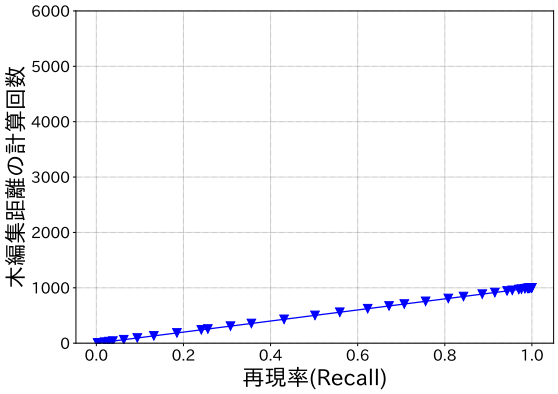


図 7　小柳らの手法における，再現率に対する木編集距離の計算回数

[16] Mateusz Pawlik and Nikolaus Augsten. Rted: a robust algorithm for the tree edit distance. *Proc. VLDB Endow.*, Vol. 5, No. 4, p. 334–345, dec 2011.

[17] Mateusz Pawlik and Nikolaus Augsten. Tree edit distance: Robust and memory-efficient. *Information Systems*, Vol. 56, pp. 157–173, 2016.

[18] Hélène Touzet. Comparing similar ordered trees in linear-time. *J. of Discrete Algorithms*, Vol. 5, No. 4, p. 696–705, dec 2007.

[19] Fei Li, Hongzhi Wang, Jianzhong Li, and Hong Gao. A survey on tree edit distance lower bound estimation techniques for similarity join on xml data. *SIGMOD Rec.*, Vol. 42, No. 4, p. 29–39, feb 2014.

[20] Sudipto Guha, H. V. Jagadish, Nick Koudas, Divesh Srivastava, and Ting Yu. Approximate xml joins. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, SIGMOD '02, p. 287–298, New York, NY, USA, 2002. Association for Computing Machinery.

[21] Karin Kailing, Hans-Peter Kriegel, Stefan Schönauer, and Thomas Seidl. Efficient similarity search for hierarchical data in large databases. In Elisa Bertino, Stavros Christodoulakis, Dimitris Plexousakis, Vassilis Christophides, Manolis Koubarakis, Klemens Böhm, and Elena Ferrari, editors, *Advances in Database Technology - EDBT 2004*, pp. 676–693, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[22] Rui Yang, Panos Kalnis, and Anthony K. H. Tung. Similarity evaluation on tree-structured data. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, p. 754–765, New York, NY, USA, 2005. Association for Computing Machinery.

[23] Nikolaus Augsten, Michael Böhlen, and Johann Gamper. The pq-gram distance between ordered labeled trees. *ACM Trans. Database Syst.*, Vol. 35, No. 1, feb 2008.

[24] Tomohiro Kan, Shoichi Higuchi, and Kouichi Hirata. Segmental mapping and distance for rooted labeled ordered trees. *Fundam. Informaticae*, Vol. 132, No. 4, pp. 461–483, 2014.

[25] Lusheng Wang and Kaizhong Zhang. Space efficient algorithms for ordered tree comparison. In Xiaotie Deng and Ding-Zhu Du, editors, *Algorithms and Computation*, pp. 380–391, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[26] Nikolai Karpov and Qin Zhang. Syncsignature: a simple, efficient, parallelizable framework for tree similarity joins. *Proc. VLDB Endow.*, Vol. 16, No. 2, p. 330–342, oct 2022.

[27] Jianhua Wang, Jianye Yang, and Wenjie Zhang. Top-k tree similarity join. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, p. 1939–1948, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Thomas Hütter, Nikolaus Augsten, Christoph M. Kirsch, Michael J. Carey, and Chen Li. Jedi: These aren't the json documents you're looking for... In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, p. 1584–1597, New York, NY, USA, 2022. Association for Computing Machinery.

[29] Pengfei Xu and Jiaheng Lu. Towards a unified framework for string similarity joins. *Proc. VLDB Endow.*, Vol. 12, No. 11, p. 1289–1302, jul 2019.

[30] Veysel Kocaman and David Talby. Spark nlp: Natural language understanding at scale. *Software Impacts*, p. 100058, 2021.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 13:30 ~ 15:40 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (2)

座長:井上 潮(東京電機大学)
コメンテータ:井手 一郎(名古屋大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T3-A-5-01] コンテンツ振り返りのためのRAGを利用した対話型フィードバックによるレビュー検索

*鈴木 醇[1]、牛尼 剛聡[2] (1. 九州大学芸術工学部芸術工学科 牛尼研究室、2. 九州大学大学院芸術工学研究院)

13:50 ~ 14:15
[T3-A-5-02] J-POPにおける楽曲動画コメントを用いた歌声印象軸の構築

*陳 沢璨[1]、熊本 忠彦[1] (1. 千葉工業大学大学院)

14:15 ~ 14:40
[T3-A-5-03] 単語分散表現を用いたホテルレビュー分析に基づく評価項目別自動スコアリング

*林 裕次郎[1]、張 建偉[2]、河合 由起子[1]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学)

14:40 ~ 15:05
[T3-A-5-04] 追加事前学習によるレビューと商品属性の対応関係学習

*河田 友香[1]、山本 岳洋[1]、大島 裕明[1]、藤田 澄男[2] (1. 兵庫県立大学、2. LINEヤフー株式会社)

15:05 ~ 15:30
[T3-A-5-05] 動画の属するコミュニティ情報を考慮した動画再生回数予測手法の提案

*川上 大凱[1]、鈴木 優[1] (1. 岐阜大学)

コンテンツ振り返りのためのRAGを利用した対話型フィードバックによるレビュー検索

鈴木　　醇†　　牛尼　剛聡††

† 九州大学芸術工学部　〒 815–8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学芸術工学科研究院　〒 815–8540 福岡県福岡市南区塩原 4-9-1
E-mail: †suzuki.jun.296@s.kyushu-u.ac.jp, ††Ushiama@design.kyushu-u.ac.jp

あらまし　映画や漫画などのコンテンツを閲覧した後, その作品に寄せられたレビューを読み振り返るという人は多い. この行為には,『作品描写理解の深化』,『共感』,『自分と異なる視点に触れる』など, 様々な目的の元に行なわれると考えられる. 個人の価値観や目的は異なる一方で, 作品レビューはインプレッション数や新規順程度の並び替えしか行われておらず, ユーザにとって価値のあるレビューを読むことができるとは限らない. 本研究では, ユーザが自分と近い意見を閲覧することを目的としたときに, ユーザが『共感』できるレビューを対象とし, 絞り込み状況に応じた質問を逐次的に生成し, 解答をフィードバックとすることによってレビューを発見可能な手法を開発する.

キーワード　LLM, RAG, 適合性フィードバック

1　はじめに

近年, 定額制動画配信（SVOD）は急速に市場規模を拡大している.GEM Standard の調査によると日本国内における定額制動画配信の市場規模は増加の一途を辿っており,2022 年度の市場規模は 4,508 億円であったが [1]、2019 年度の調査 [2] が 2,692 億円であったことから, 直近 3 年間で 67.4%も増加していることがわかる. これは, 定額制動画配信サービスのサービス拡充や新型コロナウイルス感染拡大によって外出機会が減り巣篭もり需要が高まったことなどが影響している. また 2022 年度の国内市場シェア率は,Netflix が最も高く,U-NEXT と Amazon プライムビデオが続いている. これら 3 つのサービスは, 共通して映画やドラマ, アニメなどを主たるコンテンツとしており, これらは鑑賞後に感想や評価を共有することも盛んである. 定額制動画配信の市場規模拡大に伴い, 鑑賞後に感想や評価を共有するためのレビューサイトの利用者も増加している.

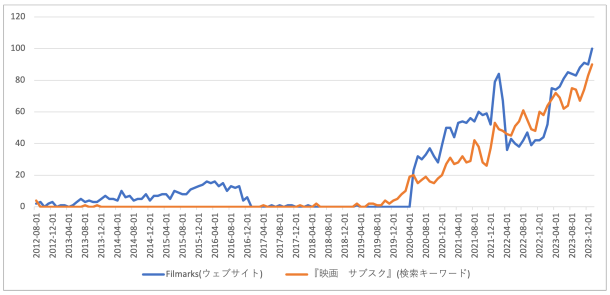


図 1　filmarks とキーワード『映画 サブスク』の検索数推移

上図 1.1 は Google Trends を用い, レビュー投稿サイト filmarks のアクセス数と『映画 サブスク』と検索された回数の最大値を 100 としグラフにしたものである. 図 1.1 から, 新型コロナウイルスによる巣篭もり需要が拡大した 2020 年 4 月より,2 項目とも急速に増加していることがわかる. これは, 定額制動画配信の規模拡大とともに, レビュー投稿サイトも利用者が増加していることを示している.

filmarks の利用者は, レビューを書かかないユーザの数が圧倒的に多い. 例えば, 映画『君たちはどう生きるか』に対しては,『文書なし』のレビュー件数は約 12 万件に対して,『文書あり』のレビュー件数は約 2.8 万人である. また,filmarks はユーザ登録・レビューの投稿をせずとも閲覧ができるため, 閲覧のみを行うユーザも多くいることが考えられる. このことも考慮に入れると, レビュー投稿サイトである filmarks の利用者の中では, レビューを書くユーザの数は少なく, レビューを書かない・閲覧のみを行うユーザの数が圧倒的に多いと考えられる.

レビューを読む行為は, 鑑賞者が作品を深く理解し共感を得たり, 新しい感性に触れる手段としての役割を果たしている. レビューを読むことは, 作品の描写やメッセージだけでなく, 他者の視点から見た作品の魅力や意味を理解するのに役立つ. このような鑑賞の一環としてのレビュー閲覧は, 視野を広げ自らの感性や評価基準を振り返る機会を提供する.

しかし, 現在のレビューサイトでは, ユーザが望む特定の視点や情報を含むレビューを見つけることが容易ではない. ユーザの個々の価値観や観賞の目的が異なるにもかかわらず, 既存のレビューサイトでは作品レビューが単なるインプレッション数や新規順に基づいているため, ユーザが共感できるレビューを見つけることが難しい場合があるためである. そのため, ユーザは, 自分が共感できるレビューを見つけるのが困難である.

現在のレビューサイトでは, ユーザが求める特定の視点や情報を含んだ映画のレビューを見つけることは容易ではない. この問題に対処するため, 本研究では, 大規模言語モデル（LLM）を利用して構築されたシステムとの対話を利用して, ユーザからのフィードバックを収集し, ユーザが期待する特定の視点や情報を含む映画レビューを効率的に検索できる手法を開発することを目指す.

具体的には, 文書類似度を用いた文書検索と LLM の文書理

解・生成能力を活かしてシステムを構築する. 本システムでは, まず, ユーザが共感できるレビューを見つけることに焦点を当て, ユーザと近い意見を持つ他のレビュワの意見を検索する. その後, ユーザに対して絞り込まれた状況に応じた質問を逐次的に生成し, ユーザの回答を収集する. これらの対話の結果をフィードバックとして取り入れ, 次回の質問に利用する. 単純に, ユーザとの対話を取り入れず検索結果上位の文書を適合文書と仮定して検索を改善する疑似適合性フィードバック [4] という手法もあるが, 本手法の方がユーザから確実なフィードバックを得れるため, レビューの検索をより精緻に行えることが期待できる.

本手法により, ユーザの鑑賞体験を深化させると同時に, 共感できる視点や感性に触れる機会を提供することを目指す. また, 質問と回答を通じたシステムとの対話により, ユーザが求める情報に焦点を絞り込み, 効果的なレビューの発見を支援する. 本手法により期待される効果は, ユーザが映画やコンテンツに対する理解を深め, 異なる視点からの情報を通じて鑑賞体験をより豊かにすることにある. 図 2 に提案するシステムの動作例を示す.

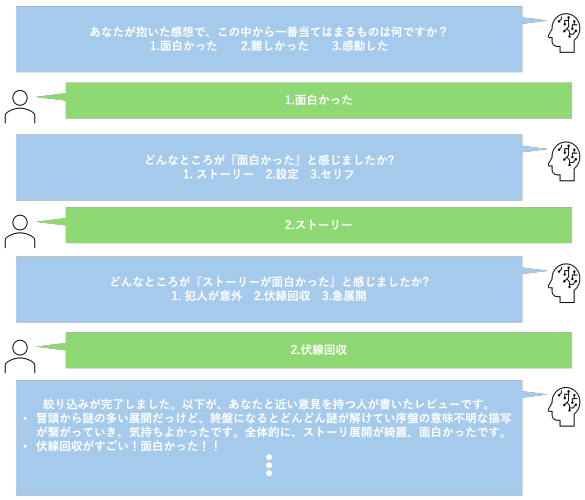


図 2　システム動作例

本論文の構成は以下のようになっている. 第 2 章では, 関連研究の紹介を行う. 第 3 章では, 提案手法について述べる. 第 4 章では, 提案手法の有効性を評価するための実験内容について述べ, 実験結果について述べる.5 章では実験結果に関する考察について述べる. 第 6 章では, まとめと今後の課題について述べる.

2　関 連 研 究

2.1　適合性フィードバックを利用したクエリ拡張

本研究と同様に, ユーザのフィードバックに合わせてクエリを拡張する研究がいくつか行われている.Verberne らは学術的な検索におけるクエリ拡張を適合性フィードバックによる手法で行った [5]. Verberne らの手法では, 検索とクエリ追加を交互に繰り返す. クエリ追加ステップでは, 上位 100 件の検索結果のうち検索者が適合文書と判定した文書中に含まれる語を追加クエリの候補としてスコア付けし, 上位 10 件のクエリ候補を検索者に提示する. 出永らは, それに加え非適合文書をクエリに追加することによって, クエリの拡張を行う手法を提案した [6]. 本手法では, レビュー対象としてこれらの手法と同様にユーザが適合する選択肢を選ぶことによって検索結果を改善していく. 一方, 本手法ではこれらの手法と違い,LLM を用いて動的に質問を生成し, ユーザと対話と通してフィードバックを得る.

2.2　検索を更に絞り込む質問の生成

情報を検索する際, ユーザが最初に入力したクエリだけでは, ユーザの意図を十分に汲み取れない場合が多い. それに対して, ユーザに質問を提示しユーザの意図を汲み取るという手法が提案されている.Hamed Zamani ら [7] は, この質問生成を機械学習を用いて自動で生成する手法を提案している. また, ユーザが迅速かつ正確に自分の意図に合わせた情報にアクセスできるように, 明確な質問を用意し事前に定義された選択肢から, 選択できるようにする手法も検討している.Hamed Zamani らの手法では, 過去の検索ログからクエリの再構築データを抽出・分析し, それらを用いて強化学習・弱教師あり学習などを用いてこの質問生成をおこなっている.

本手法でも, ユーザの意見と近いレビューを検索するためにユーザが明快に回答できる質問・選択肢を生成することを目指す. 一方で, 既存手法のように新しくパラメータの更新を伴う学習を行うのではなく, 学習済みの LLM を用いて質問生成を行う.

3　提 案 手 法

本章では, 絞り込み状況に応じた質問を動的に生成しユーザに尋ねることによってユーザの共感できる視点を検索するための手法を提案する. 提案する手法の概念図を図 1 に表す.

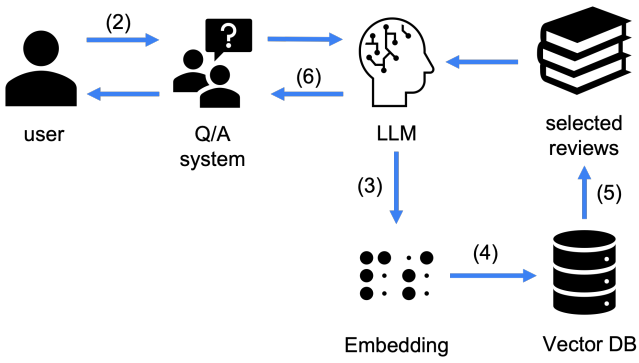


図 3　システム概念図

3.1　提案手法の概要

提案手法では, 以下のような手順でレビューの検索を行う.

（1） トピック分析を行い最初の質問を生成
（2） ユーザが質問に回答
（3） LLM を利用した過去の回答と合わせてクエリを拡張
（4） クエリ拡張したものをベクトル空間に Embedding

（5） データベースからベクトル空間上で近いものを検索
（6） 検索したレビューの中からさらに絞り込む質問を生成
（7） 任意のレビューになるまで 2〜6 を繰り返す
（8） 絞り込み結果をユーザに提示

提案手法は, 図 1 で表した流れで処理を行う. また,LLM として OPEN AI のモデルである gpt-3.5-turbo-1106 を用いる. これは, 会話行うためのモデルである同社の gpt-3.5-turbo を会話以外のタスクに利用しやすくするために JSON,XML,YAML などの形式の生成に適するよう調整がなされたモデルである. このモデルは, 従来のモデルと比べて JSON,XML,YAML などの形式を生成するタスクにて 38%の改善がみられた [11]. 本手法では会話ではなく LLM の言語能力に着目し, 質問や回答を生成してもらうという目的で利用するため,gpt-3.5-turbo-1106 が適していると考えた.

3.2 手法の説明

以下に, 提案手法の概要で示した手順に関して詳細に説明する.

3.3 最初の質問の生成

提案手法において, ユーザに対して『この中で, あなたが抱いた感想の中で最も近いものを選んでください』という質問をする. ここでの選択肢は, 対象コンテンツに寄せられた代表的な印象であることが望ましい. そのために, まずはレビューに対してトピック分析を行い, 代表的な文章を取得する. その後に,LLM を用いて選択肢を生成する.

トピック分析には,Grootendorst による BERTopic モデルを利用する [12]. このモデルは LLM である sentense-BERT を利用して文書群をベクトルデータに変換し, クラスタリングする.

ここで得られたクラスタ単位で TF-IDF 表現を行い, その表現間での類似度に基づいて計算することにより代表的な文章を取得する. そして, 取得した代表的なレビュー上位 3 件を LLM に渡し, ユーザに与える選択肢を生成する. 図 4 に LLM に渡す際のプロンプトの例を示す. また, 映画『君たちはどう生きるか』のレビューに対して, クラスタリングを行った例を図 5 に, 選択肢生成を行った例を表 6 に示す.

```
以下の#文章は, 映画のレビューを 9 個のクラスタに分類したものです.
#文章の左の数字はクラスタの id を表しています. 各クラスタ毎に一言程度で感想を要約し, 回答してください.
(例: 難しかった, 面白かった, 印象的だった)
クラスタ毎の感想が意味的に被らないことが望ましいです.
#文章:
1. クラスタ 1 のレビュー 1
1. クラスタ 1 のレビュー 2
1. クラスタ 1 のレビュー 3
2. クラスタ 2 のレビュー 1
2. クラスタ 2 のレビュー 2
...
```

図 4 プロンプトの例

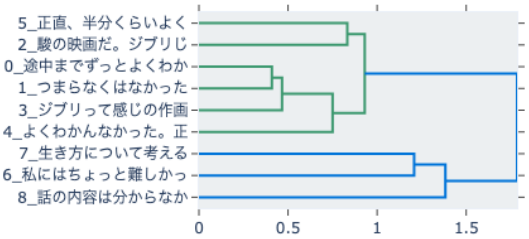


図 5 クラスタリング結果

1. 宮崎駿監督の集大成であり, 人生を描いた作品
2. 理解が難しく, 物語の一部がわからないが感動した
3. ジブリ映画とは異なり, 意外な一面があった
4. ジブリらしさがあり, 可愛いキャラクターが好き
5. 人生についての問いかけが印象的
6. 夢のような感覚で, 再鑑賞したいと思う
7. 難解だったが, 火事のシーンが迫力的で印象的
8. ジブリの雰囲気があり, 食べ物の描写が印象的
9. 前情報なしでの映画体験が初めての感動

図 6 最初にユーザに与える選択肢の例

3.4 LLM を利用した過去の回答と合わせたクエリ拡張

一般に, 情報を検索する際, キーワード検索とベクトル検索という二つの手法がある. キーワード検索は, キーワード, 語彙の類似性, 単語の出現頻度に依存する. 一方, ベクトル検索では埋め込み空間における距離を基にしているため文脈的な類似性を測ることができる. コンテンツのレビューにおいては,『難しかった』を表現するために『理解ができなかった, わからなかった』などと同一の内容であってもさまざまな表現がなされる. そのため, ベクトル検索を利用する. 文脈を考慮した検索をより精密に行うために, ユーザの過去の回答履歴を合わせて文脈的に整合性のとれた文章を生成する. この文章を疑似レビューと名付ける. これを検索クエリとすることによって, 既存レビューとの文章類似度をより精密に測ることを目指す. ここで与えるプロンプトを表 7 に示す. また, このプロンプトによる生成例を表 8 に示す.

```
以下は,#質問とユーザの回答のセットです. これらより, ユーザの意見文を考えてください.
#質問とユーザの回答のセット:
Q1: ユーザに尋ねる質問 1
A. ユーザの回答 1
Q2: ユーザに尋ねる質問 2
A. ユーザの回答 1
Q3: ユーザに尋ねる質問 2
A. ユーザの回答 1
```

図 7 プロンプトの例

質問	回答
Q0: 映画に対してあなたはどのように感じましたか?	A. 感動的なジブリの集大成であった
Q1: 映画のどの部分が特に印象に残りましたか？	A. 映像やアニメーションの美しさ
Q2: あなたがこの映画で感じたことは？	A. ストーリーが理解できなかった
出力結果：この映画は感動的なジブリの集大成であり, 特に映像やアニメーションの美しさが印象的でした. ただ, ストーリーが理解できなかったのが残念です.	

図 8 クエリ拡張生成例

3.5 ベクトルデータベースの準備

レビューを検索するために, 作品に寄せられているレビューに対して埋め込み表現を行ったベクトルデータベースを作成する. 埋め込みモデルには,Open AI から提供されている text-embedding-ada-002 [9] を利用する.

3.6 レビューの検索

疑似レビューを検索クエリとし, ベクトルデータベース内を検索する. 疑似レビューも同様に,text-embedding-ada-002 を用いて埋め込み表現を行う. 疑似レビューを埋め込み表現したものと,VecotorDB 内の他のベクトルとのコサイン類似度を計算する.

コサイン類似度を利用し, 閾値を設けレビューの絞り込みを行う. 一般に, 序盤ほどクエリが持つ情報が少なく, ユーザが回答を繰り返すほど, クエリが持つ情報が大きくなる. そのため, ユーザが回答を繰り返すほど一定値を超えるコサイン類似度をもつレビューの数が増える可能性が上がることを考慮し閾値を設定する.

3.7 絞り込むための質問の生成

検索結果の中からより近いレビューを見つけ出すため, さらに絞り込むための質問を生成する. そこで, 現在の絞り込み結果の中から上位 10 件と疑似レビューを LLM に提示することによって, 次回ユーザに提示する質問を生成する.LLM には, 期待する例を提示し出力結果の安定を図る. 図 9 に, プロンプトと質問の生成例を表す.

以下に, 映画に対する＃レビューを列挙します. あなたは, 意見がユーザと近い人を探したいです. そのため, ユーザの意見を聞き出す必要があります. 現在聞き出せているユーザの意見は『ps_reviews』です. どのような質問をしたら『ps_reviews』以上の情報を聞き出し, 近い意見を探すことができますか？ユーザに問う 1 つの質問と, その回答となる選択肢を 5 つ考えてください. 選択肢は単語程度の短い文章であることが望ましいです. ＃他の人のレビュー： ・レビュー 1 ・レビュー 3 ・レビュー 2 ... ＃ 期待する JSON レスポンス {”question”:”ユーザに問う質問”,sentakushi:[”選択肢”,”選択肢”,]}
ps_reviews は 現在の疑似レビューを代入する.

図 9　プロンプトの例

4 実　　験

4.1 利用する LLM

本実験では、GPT3.5-turbo を LLM として利用する。GPT3.5-turbo は日本語における処理能力が高いことも知られている.Weights & Biases Japan 株式会社が提供している,Nejumi LLM リーダーボードではさまざまな LLM に対して, 日本語における一問一答形式で言語理解を評価する llm-jp-eval [13] と, プロンプト対話で生成能力を評価する MT-Bench [14] という二つの特性を持つベンチマークを用いて総合評価を行っている.Nejumi LLM リーダーボードにおいて,GPT-3.5 は GPT-4 の 2 つのモデルに次いで 3 番目に高いスコアを示している. また,llm-jp-eval は様々な項目を複数のデータセットを用いることで計測している.llm-jp-eval では, 日本語文章理解力を測るために,JSQuAD というデータセットを用いている [15]. これは,Wikipedia の日本語記事を用い, 段落に対応する質問とその答えをクラウドソーシングで作成したもので与えられた文章を理解した上で, 質問に回答する必要がある.JSQuAD を用いて測られた, 日本語文章理解力の項目は正解率が 0.8406 である. 本研究では, この文章理解能力に着目し,GPT3.5-turbo を利用して実験と評価を行う.

4.2 使用するデータ

実験に使用するデータを作成するために, 映画『君たちはどう生きるか』を対象に映画レビューサイトを filmarks に 2023/10/1 までに寄せられたレビューを 28,079 件取得した. それに対して,Open AI から提供されている text-embedding-ada-002 を用いて Embedding することによって, ベクトルデータベースを構築した.

4.3 実 験 設 定

今回の実験では, 最初の質問は表 6 と同様のものを設定する.

実験では, 実際にユーザが使用する場合を想定し回答するための AI を用意する. 具体的には,GPT3.5-turbo を用いてユーザをロールプレイさせる. まず, 全レビューから無作為に一つのレビューを抽出する. この抽出したレビューを, 本提案手法の利用ユーザの意見とみなす. このレビューをターゲットレビューとみなす. 理想的には, ターゲットレビューが検索結果として出てくることである. ターゲットレビューを GPT3.5-turbo に渡し, そのレビューを書いた本人として選択肢を回答させる. このように, 二つの LLM を対話させることによって検索が行われるかを調査する. 質問と回答を 1 イテレーションとし,5 イテレーション行う.5 イテレーションでどのくらい検索が進むか, 指標としてコサイン類似度を用いる. 初期状態と 5 イテレーション後について, ターゲットレビューと疑似レビューのコサイン類似度を計算し差を求める. これをスコアとし, 値の増減を観察する. 本実験は GPT3.5-turbo の回答能力に依存する. そのため, 以下の 3 つのケースを実験し比較する.

ケース 1: 検索 AI が質問を提示し, 回答 AI が答える (提案手法)

ケース 2: 無作為に抽出したレビューから質問生成, 回答 AI が答える (検索なし)

ケース 3: 検索 AI が質問を提示し, 回答は無作為に抽出する (回答 AI 不使用)

提案手法の結果を測るとともに, これらを比較することによって精度が回答 AI, 検索 AI のどちらに起因しているかを評価する. ケース 1 とケース 2 を比較することによって, 本提案手法の逐次的に検索・質問の生成の仕組みがどの程度結果に寄与するかどうかを調べる. ケース 1 とケース 3 を比較することによって, 回答 AI の回答能力が確かなものかを確かめる.

3 つのケースそれぞれ,40 回実行し平均をとる. 実験方法のイ

メージを図 10 示す.

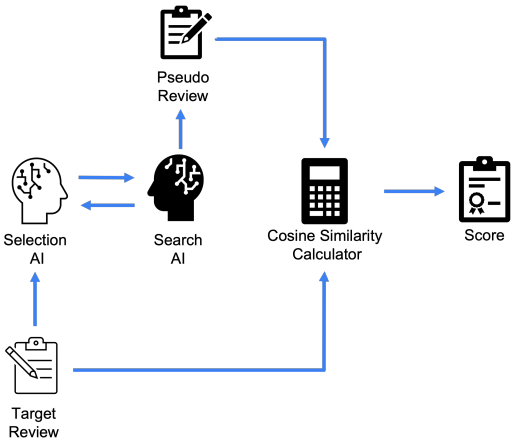


図 10　実験方法イメージ

4.4　実 験 結 果

提案手法によって生成した文章を表 5 に示す. また, ケース 1,2,3 のコサイン類似度の増加量の平均を表 1 に表す. また, ケース 1,2,3 の結果を箱ひげ図にしたものを図 11 に表す.

表 1　結果：コサイン類似度の増加量平均 (×　100)

	ケース 1	ケース 2	ケース 3
平均	2.409176897	1.758535687	1.352611565

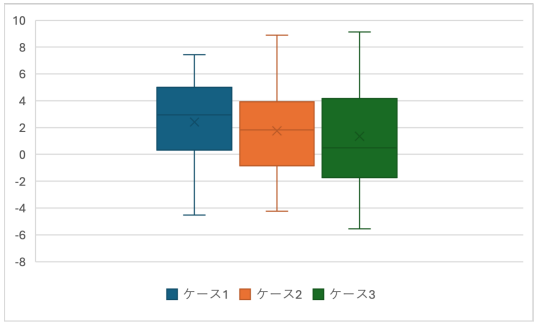


図 11　結果: 箱ひげ図

5　考　　察

表 1 より実行結果のベクトル類似度の増減の平均はケース 1 ＞ケース 2 ＞ケース 3 の順に高いことがわかった. また, 箱ひげ図を見ると第一, 第二, 第三四分位数に対してケース 1 が最も高いことがわかる.

5.1　t　検　定

これらの結果に関して, 有意な差があるかどうかを調べる. トピック 1 と 2, トピック 1 と 3 のそれぞれに対して『等分散を仮定し, 有意水準 95% で片側 t 検定』を行う.

まず, ケース 1 とケース 2 に対して

帰無仮説 H_0 : ケース 1 の母平均＝ケース 2 の母平均

対立仮説 H_1 : ケース 1 の母平均＞ケース 2 の母平均

として, 有意水準 95% で片側 t 検定を行う.

表 2 にケース 1 とケース 2 に対する検定の結果を示す.

表 2　t 検定: ケース 1, ケース 2

	ケース 1	ケース 2
平均	0.024091769	0.017585357
分散	0.001052714	0.00103668
観測数	40	40
t	0.900246372	
P(T$_{j}$=t) 片側	0.185380125	
t 境界値 片側	1.664624645	

片側検定の結果帰無仮説は棄却されず, 帰無仮説は棄却されず, ケース 1 とケース 2 には有意な差がないことがわかった.

同様に, ケース 1 とケース 3 に対しても

帰無仮説 H_0 : ケース 1 の母平均＝ケース 2 の母平均

対立仮説 H_1 : ケース 1 の母平均＞ケース 2 の母平均

として, 有意水準 95% で片側 t 検定を行う.

表 3 にケース 1 とケース 2 に対する検定の結果を示す.

表 3　t 検定: ケース 1, ケース 3

	ケース 1	ケース 3
平均	0.024091769	0.013526116
分散	0.001052714	0.001355178
観測数	40	40
t	0.900246372	
P(T$_{j}$=t) 片側	0.08859338	
t 境界値 片側	1.664624645	

5.2　要　　因

これらの結果となった要因として, ランダムに行ったケース 2,3 のばらつきの大きさや外れ値の影響が考えられる. また, 結果をより詳細に検討することで以下のような要因が見えてきた.

（1） 類似した質問の反復

（2） 選択肢にユーザの意見と近い要素が含まれない

（3） 回答 AI の回答能力と文書検索能力の齟齬

以下に, それぞれについて詳しくみていく.

1. 類似した質問の繰り返し

今回の実験において, 類似した質問を繰り返してしまう結果が多くみられた. 例えば, 表 4 のようなものである.Q2 と Q3 に関して, 言い方が異なるだけでほぼ同じ回答であるといえる. 質問生成時に, 現在得られている情報として疑似レビューを渡し, よりユーザの情報を得られるように仕向けている. しかしながら, このように回答をすることによって得られる情報があまりない質問を生成してしまっている. そのため, この既にわかっている情報を重複して聞かないようにする仕組みを考える必要がある.

表 4 類似した質問を繰り返す例

Q0: 映画に対してあなたはどのように感じましたか?
A. 君たちはどう生きるかという深遠なテーマ
Q1: 映画『君たちはどう生きるか』についてどのように感じましたか？
A. 世界観が抽象的だが魅力的だった
Q2: 映画『君たちはどう生きるか』についてどのように感じましたか？
A. 映画のタイトルと内容の関連性について考えた
Q3: この映画の世界観やテーマについて, どのような印象を持ちましたか？
A. 題名と内容の関連性について考えさせられた

2. 選択肢にユーザの意見と近い要素が含まれていない

本提案手法では, ユーザは選択肢の中から意見として近いものを選んでいく想定である. しかしながら, 適切な選択肢が生成されず, ユーザの意見と近いレビューが選択肢に含まれない場合がある. これは, 選択肢を生成するときに検索結果上位 20 件を渡しているが, 現在の検索クエリ (疑似レビュー) が適切でない場合などに, あまりユーザの意見と近くない意見ばかりを選んで渡してしまっている可能性があることが原因であると考えられる. そのため, 質問生成時に LLM に渡すレビューをどのようにして選ぶか再検討する必要がある.

3. 回答 AI の回答能力と文書検索能力の齟齬

回答を行う AI は, 対象となるレビューを記述したユーザを模倣し, 選択肢を選択します. これには, ユーザが回答するであろう選択肢まで含まれることがある. 例えば, 表 1 の 2 イテレーション目がそれに該当する. ターゲットレビューでは『それが人生というものなのかな』と綴られている. この言葉は『物語の出来事を通して得た感情や気づきを, 自分自身の人生と結びつけている』と解釈することもできる. そのため,『自分の人生を考えさせられた』は自然な回答であるといえる. しかしながら, ベクトル類似度は 0.01 程度しか変化していない. これは埋め込み表現を用いてコサイン類似度を測る手法では, このような類似度を正確に測ることが難しい可能性がある.

6 ま と め

本研究では,RAG のフレームワークと LLM を用いて絞り込み状況に応じて質問を逐次的に生成し, ユーザの回答をフィードバックとすることによってユーザと近い意見を持つレビューの検索を測ることに取り組んだ. 実験結果より, 質問生成法や検索の仕組みに課題があることがわかった. そのため,LLM に渡すプロンプトや検索システムを見直す必要がある. 今後は, 考察であげたうまくいかなかった要因の改善する予定である.

文 献

[1] 定額制動画配信（SVOD）の市場規模は 4,508 億円（22 年国内, 前年比 16.7%増）, 市場シェアは Netflix が 4 年連続首位,U-NEXT が 2 位に浮上 (2024) Available at: https://gem-standard.com/columns/673 (Accessed: 13 January 2024).

[2] 【注目の調査結果】＜動画配信（VOD）市場＞ 2019 年前年比 22.4%増の 2,692 億円,「Netflix」が SVOD 市場シェア No.1. Available at: https://www.gempartners.com/news/20200213/ (Accessed: 13 January 2024).

[3] Google Trends (2024). Available at: https://trends.google.co.jp/trends/explore?date=2012-08-10%202024-01-13&geo=JP&q=%2Fg%2F11ftz_tkv6, 映画%20 サブスク (Accessed: 13 January 2024).

[4] J.J. Rocchio, Relevance feedback in information retrieval,The Smart system - experiments in automatic document processing, Prentice Hall Inc. (1971).

[5] Verberne S, Sappelli M, Kraaij W, Query Term Suggestion in Academic Search ECIR 2014, Advances in Information Retrieval, pp 560-566 (2014)

[6] Yuma Idenaga, Satoshi Fukuda, Yoichi Tomiura, Query expansion using relevant feedback in academic search. (2019)

[7] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In Proceedings of the web conference 2020.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. (2020)

[9] New and improved embedding model (2024). Available at: https://openai.com/blog/new-and-improved-embedding-model (Accessed: 13 January 2024).

[10] Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S. and Amodei, D., Deep reinforcement learning from human preferences.(2017)

[11] New models and developer products announced at DevDay (2024). Available at: https://openai.com/blog/new-models-and-developer-products-announced-at-devday (Accessed: 20 January 2024).

[12] Maarten Grootendorst,BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022).

[13] llm-jp. llm-jp-eval (2024). Available at: https://github.com/llm-jp/llm-jp-eval (Accessed: 21 January 2024).

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (2023).

[15] 日本語言語理解ベンチマーク JGLUE の構築 ～ 自然言語処理モデルの評価用データセットを公開しました (2024). Available at: https://techblog.yahoo.co.jp/entry/2022122030379907/ (Accessed: 21 January 2024).

表 5 実 行 例

iter	ターゲットレビュー	
	すごい面白くてドキドキハラハラってわけではないど, 見終わったあとに爽快感がある. 辛い現実から逃れることができたとしても少年は, 現実と向き合うことを選んだ. それが人生というものなのかなと思った.	
1	**質問**	**ユーザの選択**
	以下の選択肢の中で, 映画の感想としてあなたが最も当てはまるものはなんですか?	君たちはどう生きるかという深遠なテーマ
	疑似レビュー	**上位 2 件**
	君たちはどう生きるかという深遠なテーマ	1.「君たちはどう生きるか」ってこの世に生きるものすべてにとってのテーマだと思う. 主人公の眞人が様々な経験をする中で, 私自身にも問われているような気がした. 2.（私はこう生きた）君たちはどう生きるか. というメッセージが伝わった.
2	**質問**	**ユーザの選択**
	映画『君たちはどう生きるか』についてどのように感じましたか？	メッセージが伝わった
	疑似レビュー	**上位 2 件**
	映画『君たちはどう生きるか』は, 深遠なテーマについてのメッセージが伝わった.	1. 噂通りの映画だったと思う. 何を表現しているのかが曖昧だった. 個人的にはタイトルの「君たちはどう生きるか」に全て集約されるのではないかと思った. 2. ジブリらしさを残しつつ, 考えさせられる映画映画のタイトル「君たちはどう生きるか」を意識して観ると, いろんな場面で「自分ならどう解釈し, 選択するか」を問われてる感じがした. いろんなメッセージが隠れている. 本「君たちはどう生きるか」を読んでから, もう一度観たい.
3	**質問**	**ユーザの選択**
	画『君たちはどう生きるか』についてどのように感じましたか？	自分の人生を考えさせられた
	疑似レビュー	**上位 2 件**
	映画『君たちはどう生きるか』は深遠なテーマについてのメッセージが伝わり, 自分の人生を考えさせられる作品でした.	1. ジブリらしさを残しつつ, 考えさせられる映画映画のタイトル「君たちはどう生きるか」を意識して観ると, いろんな場面で「自分ならどう解釈し, 選択するか」を問われてる感じがした. いろんなメッセージが隠れている. 本「君たちはどう生きるか」を読んでから, もう一度観たい. 2. 噂通りの映画だったと思う. 何を表現しているのかが曖昧だった. 個人的にはタイトルの「君たちはどう生きるか」に全て集約されるのではないかと思った.
4	**質問**	**ユーザの選択**
	映画『君たちはどう生きるか』についてどのように感じましたか？	映画のタイトルに込められたメッセージに共感した
	疑似レビュー	**上位 2 件**
	この映画は, 深遠なテーマについて考えさせられる作品でした. メッセージがしっかりと伝わり, 自分の人生についても考えさせられました. 特に, 映画のタイトルに込められたメッセージに共感しました.	1. 本映画は生と死の絡み合いと, 他者を犠牲にする覚悟と自覚を描かれていました. 少数の犠牲による全体の存続というテーマについて深く考えさせられました. 既得権益と新旧の価値観についての衝突も示され, 新世界への変革へのメッセージを感じ取ることができました. しかし, ストーリーは深くて難しかったため, もう一度見てみたいと思いました. それでも, 美しい映像と強い女性キャラクターの活躍は大変印象的でした. 2.’ 何を伝えたいか考える作品ではなく, 自分が感じとったものを大切にしたいと思える映画だった. 見たその時々で印象がかわりそう.

J-POPにおける楽曲動画コメントを用いた歌声印象軸の構築

陳　沢璨†　　熊本　忠彦††

† 千葉工業大学大学院情報科学研究科　〒 275–0016 千葉県習志野市津田沼 2–17–1
†† 千葉工業大学情報科学部　〒 275–0016 千葉県習志野市津田沼 2–17–1
E-mail: †chenzecan.work@gmail.com, ††kumamoto@net.it-chiba.ac.jp

あらまし　近年，YouTube をはじめとする様々な動画共有サイトにオリジナル曲やカバー曲を投稿する人が急増している．また，2007 年に発売されたバーチャルシンガーソフトの「初音ミク」をきっかけに「ボカロ」という新たな音楽ジャンルが開拓され，世界中で人気となっている．一方，YouTube 発のバーチャル YouTuber（VTuber）もオリジナル曲やカバー曲を大量に投稿し，J-POP の更なる人気に貢献している．しかしながら，数多くいる歌手の中から自分の好みに合った歌手を見つけるのは容易でない．歌手名以外で検索する手段が少なく，歌声という歌手の最たる特徴は有効に活用されていないのが現状である．そこで本稿では，歌声の特徴を有効活用するための準備段階として，視聴者一人ひとりが肌で感じる歌声の印象に着目し，動画共有サイト上のコメントを分析することで，歌声の印象を分類するのに適した評価軸（本稿では歌声印象軸と呼ぶ）を構築する．最終的には，この歌声印象軸を用いて歌声の印象に基づく歌手検索を実現する予定である．

キーワード　J-POP，歌声，印象抽出，印象分析，歌手検索

1　はじめに

音楽は古今東西，文化の表現形式や媒介の一つとして重要な役割を担っているが，その中でもポピュラー音楽は国によって J-POP や K-POP と呼ばれたりしており，その国特有の伝統や文化，風習を反映している．一方，YouTube のような動画共有サイトには，オリジナル曲を創作し投稿する人やそういった曲のカバー曲を投稿する人が大勢いるが，バーチャル YouTuber（いわゆる VTuber）の出現により新たなブームが引き起こり，オリジナル曲を投稿する VTuber の増加以上に，カバー曲を投稿する VTuber が急増している．

一方，視聴者にとって動画共有サイトは，プロレベルから素人まで様々なレベルの人が投稿した無尽蔵ともいえる曲を視聴できる場であり，様々な歌手との出会いを可能にしている．しかしながら，その反面，この数多の歌手の中から自分の好みに合った歌手を見つけるのは決して容易でない．YouTube であれば，投稿された曲はジャンルや曲風などによって分類されており，この枠組みに沿った歌手検索が可能となるが，自分の好みに合う歌声を持つような歌手を検索する手段は少ない．すなわち，歌手にとって最たる特徴である歌声の印象が有効に活用されていない．

そこで本稿では，歌声の印象に基づく歌手の検索を実現するにあたり，まずは歌声の印象を表現するための評価軸（歌声印象軸と呼ぶ）を構築し，提案する．なお，本研究では，代表的な動画共有サイトである YouTube を対象とするが，YouTube では多くの場合，投稿された楽曲動画に対して自分の意見や感想などをコメントとして自由に書き込むことができるため，歌声に対しての印象を記述したものも数多く見受けられる．こういったコメントから歌声に関する形容詞と動詞を印象語として抽出し，分析することで，歌声の印象表現に適した歌声印象軸を構築する．

具体的な研究の手順を以下に示す．まず，YouTube に投稿された楽曲動画（J-POP）を対象に，オリジナル曲，カバー曲，ボカロ曲のそれぞれから「声」もしくは「ボイス」というキーワードを含むコメントを取得する．次に，取得したコメントを分析することで，印象に関するコメントを効率的に取得するためにはカバー曲に対するコメントを取得すればいいことを明らかにする．また，異なる歌手が歌う異なる曲の歌声からは異なった印象を受けやすく，同じ歌手が歌う異なる曲の歌声や異なる歌手が歌う同じ曲（カバー曲）の歌声からは同じような印象を受けやすい．そのため，同じ印象を持つ曲からコメントを抽出すると，抽出された印象語も重複になり，印象語の種類が増えず無駄な作業になるかもしれない．すなわち，視聴者が歌手の歌声から同じ印象を感じた場合，抽出される印象語が重複しやすくなるため，無駄作業を回避するにはまずは同じ歌手が歌う異なる曲と異なる歌手が歌う同じ曲（カバー曲）という二つのパターンで取得したコメントから印象語を抽出し，両パターンの間で TF-IDF を用いてコサイン類似度（印象類似度と呼ぶ）を算出し，平均値，中央値，標準偏差を求めて類似度の調査を行う．その結果を踏まえて，200 曲のカバー曲を新たに選び，印象語を抽出するとともに，抽出された印象語の頻度を WordCloud を用いて可視化する．この可視化の結果と金礪ら [3] がまとめた歌声に関する評価語を参考にしながら手作業で歌声の印象として相応しい印象語を選定する．最後に，選定された印象語を対象に，全 500 曲における印象語の出現頻度を用いて因子分析を行い，各因子に属する印象語から歌声の印象を表現するのに適した歌声印象軸を構築し，提案する．

なお，音声合成ソフトウェアにより制作されたボカロ曲が J-POP に属するかどうかについては，様々な意見があるが，ボ

カロ曲の多くが J-POP 風であることから，本研究では J-POP の 1 ジャンルとして扱うことにした．

2 関 連 研 究

歌声に関する先行研究では歌声の音としての側面に焦点を当てているものが多い．例えば，片岡ら [1] は，声楽経験の異なる被験者の歌声から，楽譜データとの対比で算出される音高・音長のずれや抑揚に着目した強弱記号の表現といった様々な音に関する情報を抽出し，分析することで，歌い手の習熟度を提示する歌唱支援について検討している．また，田中ら [2] は，楽曲サイトや動画投稿サイトで高評価だった曲に付されている印象を表す言葉に着目し，歌声とその歌い方を単語の側面から分類している．すなわち，分類した曲から音特徴量を抽出し，その特徴量を用いて歌声に対して聴き手が感じる印象を単語推定する手法を提案している．これらの研究は音に関する特徴量を分析しているのに対し，本研究では楽曲動画に対するコメントのみを用いて歌声印象軸を構築する点が異なっている．

一方，金礪ら [3] は，ポピュラー音楽におけるアマチュア歌手（女性）の歌声を対象に，歌声を実際に聴取してもらう主観評価実験を行い，その結果に対し因子分析を行うことで，歌声の多様な印象を適切に表現可能な評価尺度を構築するとともに，歌声の音に関する特徴量からその歌声の印象を推定する手法を提案している．この研究では様々な文献や専門家による CD レビュー，X（旧 Twitter），ニコニコ動画などで集められたデータから手作業で評価語を抽出して主観評価実験と因子分析を行い，最終的に 12 語の評価語と因子負荷量の高い評価語を参考に命名した 3 つの因子（迫力性，丁寧さ，明るさ）によって構成された歌声の印象評価尺度を構築し，印象得点を自動推定の出来る重回帰分析モデルを構築した．この研究ではアマチュア女性歌唱者のみを対象とした点に対し，本稿では歌手の性別を制限せず，YouTube の楽曲動画コメントから抽出された印象語を分析し，歌声の印象軸を作成する点が異なっている．なお，本稿では，金礪らが選出した評価語を印象語選出の際の参考にしている．

他の関連研究を挙げると，池田ら [4] は音楽大学の声楽専攻で指導を行っているプロ歌手の教員 4 名による議論に基づいて初学者の歌声に対する印象評価語として 40 語を選出している．この 40 語を音声生理学的観点から再検討し，池田らの先行研究 [5] を踏まえて，因子分析等は行わずに，討論のみにより「発音源因子」,「共鳴因子」,「呼吸因子」の 3 因子に修正した．この研究で抽出された印象評価語は教員 4 名らの議論によって歌の教育・指導を目的として得られた評価語であり，一般人が歌声を評価する際にほぼ使わない用語であるため，本稿では YouTube の楽曲動画コメントから抽出した日常生活でよく使われている印象語を用いている点が池田らの研究とは異なっている．

3 歌声に関する印象の抽出

楽曲動画のコメントには，歌声に対しての印象のみならず，

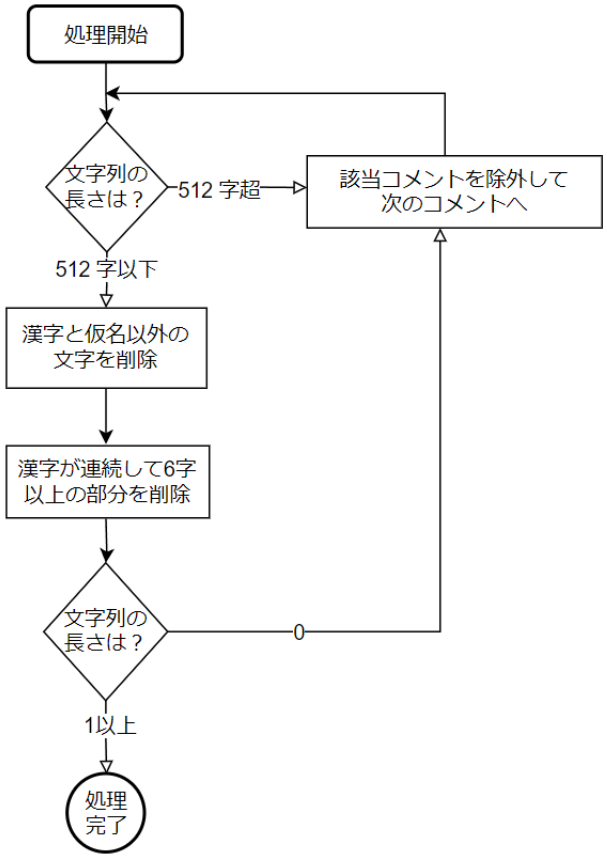


図 1: 前処理の流れ

動画に使われたイラストやアニメーション，歌詞，メロディなど様々な要素に対するコメントが混在している．そのため，まずはコメントから歌声に対する印象だけを抽出する必要がある．楽曲動画のコメントを観察した結果,「声」もしくは「ボイス」という単語が含まれているコメントは歌声に対する何らかの評価が含まれてる可能性が高いことがわかった．そこで本研究では，この 2 つの単語をキーワードとして用いるとともに，形態素解析ツールと構文解析ツールを用いることで，より高精度に歌声に関する印象の抽出を試みる．なお，本研究で採用した形態素解析ツールと構文解析ツールは以下のとおりである．

a）形態素解析ツール Juman

Juman [6] は，京都大学・黒橋研究室が開発した日本語形態素解析ツールであり，日本語のテキストを形態素に分割するとともに，それぞれの形態素に品詞情報や活用形などの言語情報を付加する．

b）構文解析ツール KNP

KNP [7] は，同じく京都大学・黒橋研究室が開発した日本語構文解析ツールであり，Juman の出力に基づいてテキストの構造を分析し，文節や係り受けなどの情報を出力する．

3.1 コメントの前処理

YouTube の楽曲動画コメントには文字数や言語に関する制限がなく，絵文字やスタンプも使用可能な設計となっている．しかしながら，Juman と KNP は日本語のみに対応していて，他の言語や長文，絵文字などには対応していない．そのため，

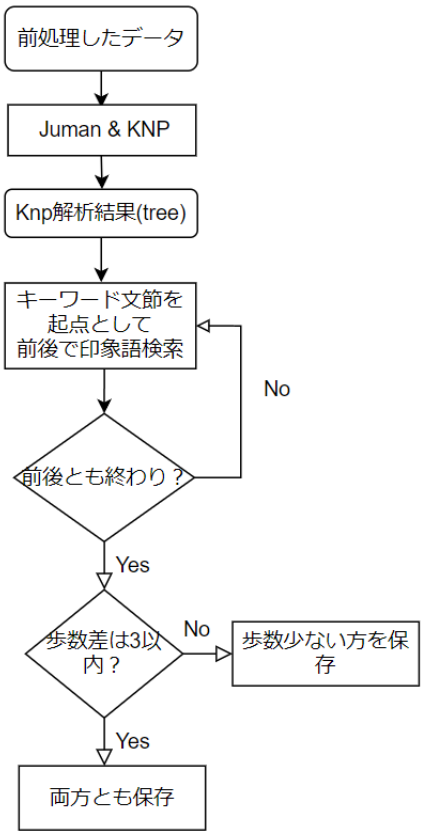


図 2: 印象語抽出の流れ

各ツールにコメントを解析させる前に，何らかの前処理を行う必要がある．

本研究で行う前処理の流れを図 1 に示す．なお，前処理への入力は「声」もしくは「ボイス」が含まれているコメントである．

図 1 に示したように，本稿では文字列の長さが 512 字を超えるコメントを長文として扱い，削除した．これは，文字列の長さが 512 字を超えるコメントのほとんどが歌詞をコピー&ペーストしたものであり，歌声に関する印象が含まれていることはほぼないためである．

次に，コメントに日本語の漢字と仮名（ひらがな・カタカナ）のみが残るよう正規表現を用いてフィルタリングを行い，コメントから外国語や顔文字，絵文字などを削除した．しかしながら，この方法では中国語のコメントを削除できないため，漢字が連続して 6 字以上出現する部分を中国漢字からなる文字列と捉え，除外してみて，残った部分の文字列の長さが 0 になったコメントは中国語のコメントとして削除した．これは，日本語のコメントにおいても漢字が連続して 6 字以上出現する可能性を考慮し，一括りに除外するのではなく，連続する漢字を削除した後のコメントの文字列の長さで日本語のコメントか中国語のコメントかをより確実に判断できるようにするためである．

3.2 印象語の抽出

3.1 節の手法で前処理されたコメントから歌声に関する印象語を抽出するための全体的な流れを図 2 に示す．まず，前処理されたコメントを Juman に入力して形態素解析を行い，その

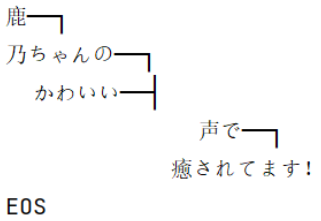


図 3: KNP の出力例

出力を KNP に入力して，図 3 に示したような係り受け解析の結果を取得する．KNP の出力結果では文の係り受け関係が文節ごとに分けて出力されるため,「声」あるいは「ボイス」のキーワードが存在する文節を起点として，前後の文節で最も近い印象語を探索する．ここで，印象語の探索手順を図 4 に示す．なお，印象語は形容詞に限定するのが一般的であるが，一部の動詞も印象を表すことがあるため，本研究では抽出対象を拡張し，形容詞と動詞を収集することにした．また，起点から離れた文節に含まれる印象語は起点を修飾する印象語である可能性が低くなるため，チェック時に経過した文節の数（経過文節数と略す）も計数することで，前後のチェックを終了した時点で両方の経過文節数の差の絶対値が 3 以内であれば両方の抽出結果を保存し，3 より大きければ経過文節数の少ない方の抽出結果を保存することにした．

ここで図 3 に示した KNP の出力例を用いて印象語の抽出方法を説明する．図 3 の出力結果では,「声で」の文節が起点となる．まず，この文節「声で」に係っている前の文節をチェックし,「かわいい」という形容詞を印象語として抽出する．このときの経過文節数は 1 となる．ここで抽出された文節には形容詞が含まれているので,「かわいい」の文節の更に前の文節「乃ちゃんの」もチェックし，動詞や形容詞がないため，前方向のチェックはここで終了となるが，その文節に形容詞か動詞が含まれていれば，その語も印象語として抽出した後，処理を終了する．これは，形容詞と動詞が連続して出現し,「声で」に係っている可能性を考慮したためである．次に，起点である文節「声で」が係っている後ろの文節にも同じ操作を行い，経過文節数 1 の「癒さ（「癒す」の未然形）」という動詞を抽出する．これが最後の文節のため，後方向のチェックも終了となる．最後に，両方の経過文節数の差の絶対値を取り，0 になったため両方の抽出結果を保存して処理完了とする．

4 歌声印象軸の構築

本章では，J-POP をオリジナル曲，カバー曲，ボカロ曲の 3 つに分け，それぞれの曲に対するコメントにおいて印象に関するものがどの程度出現するかを調査するとともに，その結果を踏まえて，3.2 節で示した印象語抽出手法を用いて楽曲動画に関するコメントから印象語を抽出し，分析することで，歌声印象軸を構築する．

4.1 コメントの印象に関する予備調査

本研究では，J-POP を以下の 3 つに分類する．自分が創作し

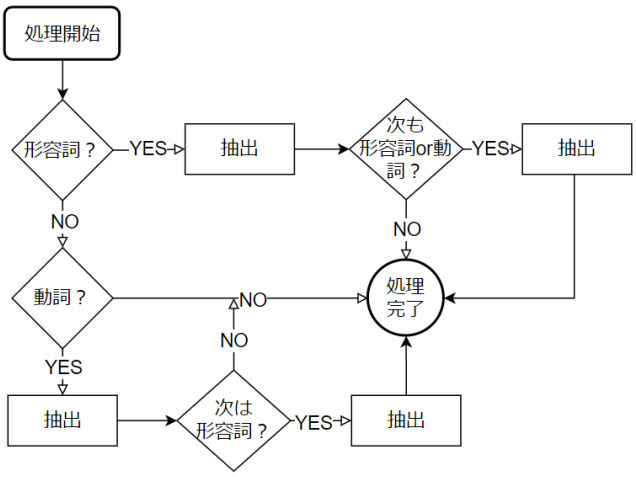


図 4: 印象語探索の手順

た曲を自分の声で歌ったオリジナル曲，他の人のオリジナル曲を自分の声で歌ったカバー曲．バーチャルなキャラクタが音声合成ソフトウェアの無機質な声で歌ったボカロ曲の 3 つである．

この 3 つでは，それぞれ歌手が異なり，YouTube に投稿された楽曲動画も人気の違いによってコメント数が大きく異なっている．そこで，効率よく様々な印象を抽出するために，まずはコメントに存在する印象語について予備調査を行う．

なお，YouYube からのコメント抽出に用いるツールは YouTubeDataAPI であり，抽出したコメントを 3.1 節で示した前処理手法で処理した後，データベース管理システムである MySQL に保存する．また，この予備調査では，歌声に対する印象を含むコメント（本稿では印象コメントと略す）が 30 件以上存在する楽曲動画のみを抽出対象とした．

4.1.1 印象頻度調査

効率よく様々な印象語を獲得するため，まずはオリジナル曲，カバー曲，ボカロ曲の楽曲動画を対象に，それぞれの印象コメントの出現頻度について調査した．

オリジナル曲は，曲自体が重複になる可能性は無いため，オリジナル曲をメインとしている歌手 65 人を対象に，それぞれの歌手からオリジナル曲 1〜10 本を無作為に抽出し，計 100 本を調査対象とした．カバー曲は,「歌ってみた」をキーワードとして YouTube で検索し，再生数の多いカバー曲 100 本を重複なしで無作為に抽出し，調査対象とした．ボカロ曲は，初音ミクのような音声合成ソフトウェアに歌わせた楽曲動画 100 本を無作為に抽出し，調査対象とした．

次に，それぞれの調査対象において，楽曲動画毎に抽出されたコメントの数を横軸，印象語を含む印象コメントの割合を縦軸とする散布図を作成した．ボカロ曲を対象にしたものを図 5 に，オリジナル曲を対象にしたものを図 6 に，カバー曲を対象にしたものを図 7 に示す．これらの図から，ボカロ曲のコメント数は 5,000 件以下が多く，印象コメントの割合も平均で 2.54 %程度と低め（つまりコメント 39 件中 1 件が印象コメントという割合）であることや，オリジナル曲のコメント数は一番多いが，印象コメントの割合は平均で 2.76 %程度と低め（つまりコメント 36 件中 1 件が印象コメントという割合）であること，カバー曲のコメント数は一番少なかったが，印象コメントの割合は平均で 15.65 %と高め（つまりコメント 6 件中 1 件が印象コメントという割合）であることがわかる．

一方，それぞれの調査対象におけるコメントの内容を見てみると，ボカロ曲に対しては，実際に歌を歌う音声合成ソフトの歌声には大きな違いがないため，歌声よりもメロディや歌詞の方が注目されており，多くのコメントもメロディや歌詞あるいは曲を創作したアーティストに対するコメントであることがわかった．また，オリジナル曲に対しても同じような傾向であり，実在する歌手の歌声に違いはあるが，オリジナルな歌詞やメロディの方に注目が集まりやすく，歌声に関する印象についての言及は少ない傾向にあることがわかった．それに比べ，カバー曲に対しては，歌詞やメロディにオリジナルな要素が少ないことから，歌声に注目が集まりやすく，コメントも歌声に関するものが多くなる傾向にあることがわかった．

4.1.2 印象類似度の算出

前節での調査結果から，印象コメントを最も効率よく抽出できるのはカバー曲であることがわかった．しかしながら，印象コメントの数が多ければ，抽出できる印象語の種類が多くなるというわけではない．そこで本稿では，大量の重複印象語を抽出する無駄作業を回避するため，カバー歌手とカバー曲における印象が重複しやすい状況に対し，印象類似度の調査を行った．

具体的には，まず，3.2 節で提案した手法を用いて印象語（の形容詞）を抽出し，抽出された形容詞集合を TF-IDF 値を用いてベクトル化することで，楽曲と楽曲のコサイン類似度（本稿では印象類似度と呼ぶ）を算出する．なお，抽出された動詞は歌声の印象と関わる単語が少ないため，ここでは扱わないようにした．

a) カバー歌手

同じ歌手が歌う場合，曲が異なっても歌声の印象は同じと言えるが，実際には歌い方によって歌声の印象も異なる場合がある．ただ変化があまりにも少ない場合は，同じ歌手から幾らコメントを取得しても抽出できる印象語は似たり寄ったりになる可能性があるので，どれほど変化があるのかを調査する．そこで，まずは YouTube でメインに活動している人気歌手 5 名を選び，それぞれの投稿から 10 曲を対象に，すべての曲間で印象類似度を算出し，平均値，中央値，標準偏差を求めた．調査の結果から見ると，同じ歌手が歌う楽曲でも歌声の印象が変化しており，印象類似度の平均値は高くても 0.8 に達することはない．すなわち，同じ歌手の異なる楽曲からコメントを取得しても印象語の多様性はある程度保つことが可能と考えられる．具体的な結果を表 1 にまとめる．

b) カバー曲

オリジナル曲と違って，異なるカバー歌手が同じ曲を歌い上げ投稿するカバー曲は大量に存在している．歌手が異なると歌声の印象も変わってくると考えられるのだが，曲の印象によって歌い方も定着しやすく，歌声にどれほど影響が出るのかは調査する必要がある．調査方法は個人判断で印象の異なるオリジナル曲 5 曲を対象として選び，それぞれのカバー曲 10 本を取り上げて，同じ曲の間で印象類似度を算出し，平均値，中央値，

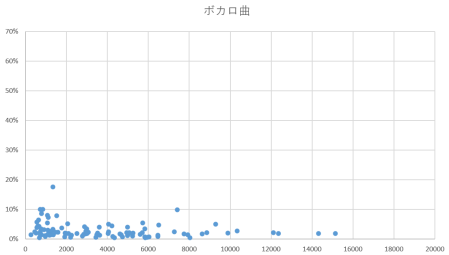


図 5: ボカロ曲に対するコメント数と印象コメントの割合

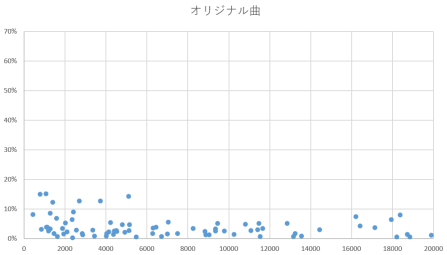


図 6: オリジナル曲に対するコメント数と印象コメントの割合

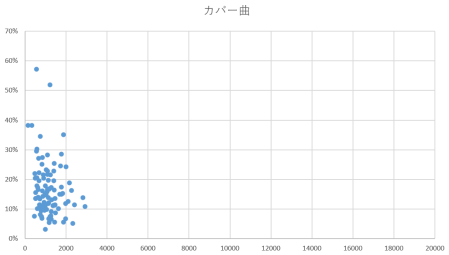


図 7: カバー曲に対するコメント数と印象コメントの割合

表 1: カバー歌手の印象類似度

歌手名	平均値	中央値	標準偏差
まふまふ	0.632	0.732	0.150
Ado	0.533	0.540	0.165
キズナアイ	0.432	0.385	0.181
P 丸様。	0.727	0.754	0.134
莉犬くん	0.625	0.655	0.181
平均	0.590	0.613	0.162

表 2: カバー曲の印象類似度

曲名	平均値	中央値	標準偏差
アイドル	0.448	0.424	0.134
ヴァンパイア	0.608	0.612	0.164
うっせぇわ	0.485	0.490	0.190
強風オールバック	0.518	0.645	0.250
命に嫌われている。	0.521	0.531	0.135
平均	0.516	0.540	0.174

標準偏差を求めた．調査結果として平均値の一番高い楽曲でも印象類似度は 0.6 程度でカバー歌手より下回っており，全体的の中央値と平均値も 0.5 程度となっている．よって，同じ曲であっても，歌手が異なるのであれば多様性に影響を及ぼすことは限られていると考えられる．具体的な結果を表 2 にまとめる．

4.2 コメントの追加取得

印象に関する予備調査の結果からカバー曲のコメントに印象語の出現頻度が一番高く，同じ歌手が歌う異なる曲でも異なる歌手が歌う同じ曲でも抽出された印象語は一定以上の多様性を保つことができると考えられる．

この結果を踏まえて，更にカバー曲を中心として印象コメント数が 20 件以上の楽曲動画 200 本に対するコメントを新規に取得し，印象頻度調査で取得した 300 曲コメントのうち，100 曲のボカロ曲のコメントを除き，残りの 200 曲と印象類似度計算で新たに取得した 100 曲のコメントを合わせて計 500 曲からコメントを取得した．

取得したコメント総数は 2,283,737 件であり，うち 154,172 件が印象コメントであった．各楽曲動画が有する印象コメント数は 21〜3,503 件であり，印象コメントの割合は 7%に達している．詳細を表 3 に示す．

表 3: 取得されたコメントデータの情報

項目	件数
楽曲動画数	500
コメント数	2,283,737
印象コメント数	154,172
印象コメント割合	7%
最少印象コメント数	21
最多印象コメント数	3,503

図 8: 出現頻度 100 以上の形容詞

4.3 印象分析

4.3.1 印象語選定

印象語を印象軸に作成するため，先ずは抽出された印象語の分析を行う．4.2 節でまとまった 500 曲の計 154,172 件の印象コメントから抽出された印象語を観察して見ると，形容詞の方は印象として相応しい単語が多いのに対し，動詞の方はほとんどが印象と関係のない単語であった．抽出された形容詞は計 1,139 語で，合計の出現頻度は 167,016 回である．出現頻度が 100 個以上の形容詞を対象に，頻度を大きさに反映させてワードクラウドに出力する．出力結果を図 8 に示す．

結果から「可愛い」,「綺麗だ」,「優しい」など歌声の印象として相応しい形容詞も見られた一方で,「好きだ」,「最高だ」,「素敵だ」といった歌声の特徴を明白に表せず，印象語として相応しくない単語も見られた．また,「可愛い」と「かわいい」,「綺麗だ」と「きれいだ」のように書き方が違うだけの単語も存在する．

ここで，金礪ら [3] が整理した歌声の印象評価語 47 語を参考に，出現頻度が 100 以上の形容詞と動詞を手作業で，同義語ど

表 4: 因子分析の最終結果

	因子 1	因子 2	因子 3	因子 4	因子 5	因子 6
切ない	**0.917**	0.063	0.221	0.118	0.019	0.079
透き通る	**0.914**	-0.011	0.205	-0.046	0.068	0.017
儚い	**0.863**	0.069	0.044	0.188	-0.027	0.072
繊細だ	**0.851**	0.103	0.080	0.104	-0.054	0.060
綺麗だ	**0.836**	0.188	0.291	0.157	0.048	0.136
低い	0.038	**0.891**	0.047	0.033	0.211	-0.035
かっこいい	0.005	**0.882**	-0.006	0.048	0.166	0.070
高い	0.081	**0.838**	0.018	0.033	0.197	0.070
力強い	0.310	**0.674**	0.110	-0.003	-0.146	0.244
温かい	0.039	0.021	**0.824**	-0.003	0.006	0.102
落ち着く	0.357	0.076	**0.753**	-0.004	0.004	-0.041
優しい	0.565	0.045	**0.676**	0.184	-0.018	0.002
癒す	0.113	-0.040	**0.643**	0.533	0.073	0.171
爽やかだ	0.362	0.007	-0.021	**0.708**	-0.141	0.118
甘い	-0.031	0.038	0.041	**0.643**	0.239	0.019
柔らかい	0.075	0.033	0.471	**0.601**	-0.040	0.121
美しい	0.531	0.100	-0.027	**0.595**	-0.062	-0.055
萌える	0.041	0.195	0.006	-0.001	**0.828**	-0.142
可愛い	-0.040	0.297	0.022	0.120	**0.776**	0.327
明るい	0.075	0.138	0.038	0.155	0.038	**0.882**
元気だ	0.259	0.143	0.530	-0.001	0.034	**0.630**

表 5: 歌声印象軸の候補として取り上げた印象語（26 語）

印象語
高い，低い，美しい，甘い，元気だ，爽やかだ， 小さい，落ち着く，震える，萌える，透き通る， 優しい，かっこいい，可愛い，切ない，色っぽい， 明るい，癒す，柔らかい，独特だ，幼い， 儚い，温かい，繊細だ，力強い，綺麗だ

表 6: 提案する歌声印象軸

儚くて透明感のある	力強い
切ない	優しくて落ち着く
繊細だ	可愛くて萌える
綺麗だ	爽やかだ
低い	柔らかくて美しい
かっこいい	明るくて元気だ
高い	甘い

うしを合併しながら歌声の印象として相応しい単語を取り上げて，出所のコメントと照らし合わせて 1 個ずつ確認し，歌声の印象と関係のない単語を除外した．最終的に残された 26 語の単語を表 5 にまとめる．

4.3.2 因子分析

これまでは 500 曲のコメントから抽出された印象語を人手で印象語として相応しい単語 26 語を取り上げた．しかしながら，歌手の歌声を分析するには 26 語は若干多くなっていて，26 語の間でも歌声の印象として意味合いの近い単語が存在するのみならず，コメント数によっては少数の人気楽曲動画のコメントでしか滅多に現れない印象語も高い出現頻度を有する可能性がある．

そのため，この 26 語の間で相関関係を分析し，他の印象語と関連性が低い単語を除外しながら意味合いの近い単語を見つけ出して合併し，印象語数を更に凝縮する必要がある．そこで 500 曲に対するコメント 154,172 件から抽出された各印象語の頻度分布を用いて因子分析を行い，印象語数の削減を試みる．因子分析のサンプルとして，先ずは 500 曲から抽出された印象語の中で，取り上げられた 26 語の同義語を合併して，合併後の 26 語の出現頻度をサンプルとし，26 語を観測変数としてバリマックス回転法を用いて分析を行い，因子数は主成分分析に基づいて決定する．その結果から先ずは印象語の共通性を確認し，著しく低くなっている印象語は他の印象語と共通因子を持たずに独自因子の割合が高いである他，どの因子においても明らかに因子負荷量が低かった印象語も同じく他の印象語と関連性が低いであるため，毎回の結果からこの 2 種類の印象語を外して因子分析を繰り返し行う．最終結果は表 4 に示す．共通性の高い印象語は 6 つの因子があって，それぞれに集まっているのは共起頻度の高い印象語，つまり同じ楽曲のコメントに出現する頻度が高いと考えられる．除外された単語は「独特だ」，「幼い」,「色っぽい」,「小さい」,「震える」の五つとなった．

残りの印象語を除外せずに数を減らして圧縮する手段は因子に直接名付けて印象語として用いるか，或いは同じ因子を持っていて意味合いの近い印象語を新しい単語や短文に合成するかのこと．最終結果から意味合いが異なる印象語も同じ因子を持つことが見られたため，因子に名付けずにそれぞれの因子において，因子負荷量が 0.5 以上であった意味合いの近い印象語の合成を行う．

因子 1 では「切ない」,「透き通る」,「儚い」,「繊細だ」,「綺麗だ」の 5 つの印象語が集まっている，意味合いが近くて同じ単語や短文に合成できるのは「透き通る」と「儚い」の 2 単語のみである．よって「透き通る」と「儚い」を「儚くて透明感のある」という文言に合成する．因子 2 では「低い」,「かっこいい」,「高い」,「力強い」の 4 つの印象語が集まっているが，意味合いが近くて合成できる印象語は見られないため，合成を行わない．因子 3 では「温かい」,「落ち着く」,「優しい」,「癒す」の 4 つの印象語を「優しくて落ち着く」という文言に合成する．因子 4 では「美しい」と「柔らかい」を「柔らかくて美しい」に合成して，結果,「爽やかだ」,「甘い」,「柔らかくて美しい」の 3 つになる．因子 5 の「可愛い」と「萌える」を「可愛くて萌える」に合成する．因子 6 の「明るい」と「元気だ」を「明るくて元気だ」に合成する．

完成した印象軸を表 6 に示す．以上より 500 本の楽曲動画から抽出したコメントを用いて印象軸の構築を完成した．構築した印象軸は出現頻度と共起頻度を考慮し，楽曲動画のコメントでよく使われている印象語を凝縮した汎用性の高い軸である．これを以て歌手の歌声の印象を表現することが可能であろう．

5 ま と め

本稿では，動画共有サイト YouTube に投稿されている楽曲動画のコメントに着目し，歌声に焦点を当てて印象語を抽出して歌声印象軸を構築した．具体的には，形態素解析ツール Juman と構文解析ツール KNP を用いて，構文的に「声」もしくは「ボイス」と係り受け関係にある印象語を抽出する手法を提案し，YouTube に投稿された楽曲動画（ボカロ曲，カバー曲，オリジナル曲）に対するコメントから印象語を抽出してみた結果，カバー曲に対するコメントから効率よく印象語を抽出できることがわかった．また，同じカバー歌手が歌った異なるカバー曲どうしの印象類似度と異なるカバー歌手が歌った同じ曲のカバー曲どうしの印象類似度を算出してみたところ，いずれにしても印象類似度の値はあまり高くないことがわかった．そこで，これらの結果を踏まえて，さらに 200 本のカバー曲に対するコメントから印象語を抽出し，歌声の印象として相応しい印象語を選定したうえで，因子分析の結果に基づいて印象軸を構築した．すなわち,「儚くて透明感のある」,「切ない」,「繊細だ」,「綺麗だ」,「低い」,「かっこいい」,「高い」,「力強い」,「優しくて落ち着く」,「可愛くて萌える」,「爽やかだ」,「柔らかくて美しい」,「明るくて元気だ」,「甘い」の 14 軸である．

本研究は歌声の印象に基づく歌手検索を実現する準備段階として，テキストの構文に基づいて抽出した歌声の印象語を分析して印象軸構築の研究であるため，今後は先ず軸を以て歌手検索の実現を目指していきたい．他には短文形式の印象語を抽出によって印象語の種類を増やす，固有表現抽出などの手法でより高精度の印象語抽出，抽出曲数を更に増やして印象語の更なる多様性や印象軸の汎用性向上など，複数の課題が残されている．

謝　　辞

本研究の一部は，JSPS 科研費 20K12085 による研究助成の成果であり，ここに記して感謝の意を表す．

文　　献

[1] 片岡 靖景，伊東 一典，池田 操，中澤 達夫，米沢 義道，今関 義弘，橋本 昌己，“歌唱支援システム構築のための歌声の分析と評価，’’ 情処研報（音楽情報科学），Vol.26，pp. 23—30，1998.
[2] 田中 晶之，中村 嘉志，“ランダムフォレスト法を用いた歌声が聞き手に与える印象の単語推定に関する研究，’’ 国士舘大学理工学部紀要，Vol.15，pp.23–28，2022.
[3] 金礪 愛，中野 倫靖，後藤 真孝，菊池 英明，“歌声の印象評価尺度の構築に基づく多様な印象の自動推定手法，’’ 情報処理学会論文誌，Vol.57，No.5，pp.1375–1388，2016.
[4] 池田 京子，香山 瑞恵，山口 道子，小畑 朱実，谷 友博，山下 泰樹，浅沼 和志，伊東 一典，“歌声の「印象評価語」と「指導語」の抽出と分類～最適な指導語を求めて～，’’ 人工知能学会第 33 回身体知研究会，4 pages，2020.
[5] 池田 京子，山下 泰樹，平林 花奈，香山 瑞恵ほか，“歌声の印象評価語・指導語の抽出－ Caro mio ben を対象として－，’’ 声楽発声学研究，No.10，2019.
[6] 日本語形態素解析システム Juman，https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN，2014.
[7] 日本語構文・格・照応解析システム KNP，https://nlp.ist.i.kyoto-u.ac.jp/?KNP，2012.
[8] YouTube Data API，https://developers.google.com/youtube/v3 (2023.12.12 参照).
[9] 剣持 秀紀，“歌声合成技術 VOCALOID^{TM} と新しい音楽，’’ JAS Journal，Vol.54，No.2 (3 月号)，2014.

単語分散表現を用いたホテルレビュー分析に基づく評価項目別自動スコアリング

林　裕次郎† 　張　 建偉†† 　河合由起子† 　中島　伸介†

† 京都産業大学　情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 岩手大学　理工学部　〒 020-0066 岩手県盛岡市上田３丁目１８－８
E-mail: †{g2054065,kawai,nakajima}@cc.kyoto-su.ac.jp ††zhang@iwate-u.ac.jp,

あらまし　ホテル検索システムにおけるレビュー文章は，ホテル検索を行う一般ユーザにとって有益である. しかし，レビュー投稿者に対するレビュー投稿のメリットは乏しく，レビュー投稿を行うユーザは多いとはいえない. そこで我々は，レビュー投稿者に対して有益な情報を提供可能なシステムを開発することで，レビュー投稿者を増加させることが出来るのではないかと考えた. 本稿では，レビューを単語分散表現を用いて分析し，各評価項目ごとに自動スコアリングを行った. 既存サイトのスコア，評価表現辞書のスコア，提案手法を用いたスコアを比較し提案手法の有効性が確認できた.

キーワード　ホテル推薦，情報推薦，レビュースコアリング

1. はじめに

近年，楽天トラベル [1] やじゃらん [2] など，宿泊施設に対する利用者のレビューや各評価項目に対するスコアが閲覧可能なホテル予約サイトが数多く運営されている．一般ユーザが，宿泊先のホテルを決定する際，他のユーザのレビューを参考にすることが多く，自身が宿泊したことがないホテルの情報を収集する手段として，このようなレビュー情報は非常に有益である．しかしながら，ホテルを利用したユーザのうち，レビュー投稿を行うユーザはごく一部に限られており，全てのユーザがレビュー投稿を行う理想的なケースに比べると，有効な情報が失われていると共に，限られたユーザの意見のみが参考にされるという意味でバイアスがかかる可能性が否定できない (図 1). 我々はこのような一部のユーザのみのレビューが共有される状況は好ましいことではないと考えている.

レビューを投稿しない理由としては，レビューを投稿しても他のユーザが参考になるだけであり，レビュー投稿者に対する大きなインセンティブは無いという理由が考えられる．レビュー投稿数増加を狙い，ポイント還元などをしている場合もあるが，それはポイントという報酬を貰うためにレビュー投稿を行うこととなり，適切なレビューを投稿することよりも，質が悪くてもレビュー投稿数が増えれば良いという，悪い傾向の原因になりかねない.

そこで本研究では，ユーザにとってレビュー投稿自体にメリットがあれば，レビューを投稿するユーザが増加するのではないかという考えのもと，ユーザが投稿した不満レビューに対して，これを解消するようなホテルをこのレビュー投稿者に推薦するシステムを提案する．本稿では，単語分散表現を用いてレビューを分析し，評価項目ごとに自動スコアリングを行った. 評価実験により，提案手法の有効性が確認できた.

2. 関連研究

有名なホテル予約サイトとして楽天トラベルがある．楽天トラベルでは，ホテルや旅館など多くの宿泊施設がレビューと共に掲載されており，評価を比較しながらホテルの検討ができる. 楽天トラベルのようなレビューを取り扱うサイトが広く普及したことで，レビューを分析する研究が多く行われている.

安倍らは, ホテルの評価項目を詳細化した評価表現辞書を作成し, 自動スコアリングを行う方式を提案している [3].

青柳らは, 宿泊予約サイトのレビュー文からクレームを抽出し, 内容解析を行い課題抽出する方法を提案している [4].

林部らは, 宿レビューから「肯定的事実」と「推薦対象」という言語知識を，宿のレビューから抽出する手法を提案している [5].

田熊らは, ユーザが順位付けしたホテルの評価項目を元にホテルを推薦する手法を提案している. 投稿されたレビューから，評価項目ごとの名詞の数をカウントし，数が多い順に評価項目の順位付けを行い，ユーザ類似度が高いレビューを推薦している [6].

和田らは, 宿泊予約サイトのレビューに記述されているギャップ文を抽出し，評価項目ごとに提示する手法を提案している [7].

Md. Shafiul Alam Forhad らは，ホテルレビューにおいて，レビューと周辺施設を考慮したホテル推薦システムを提案している．周辺施設をカテゴリ分けし，スコアリングする．レビューも分析しスコアリングし，周辺施設のスコアと組み合わせ算出した総合的なスコアからホテルを推薦した. [8].

上述の通り，ユーザによって投稿されたレビューや特徴ベクトル，ランキングなどを分析し，情報推薦を行う研究が数多く行われている. これらはこれからホテルを予約しようとしている一般ユーザに対して有益な研究だが，レビュー投稿者にメリットを与えるような研究は行われていない.

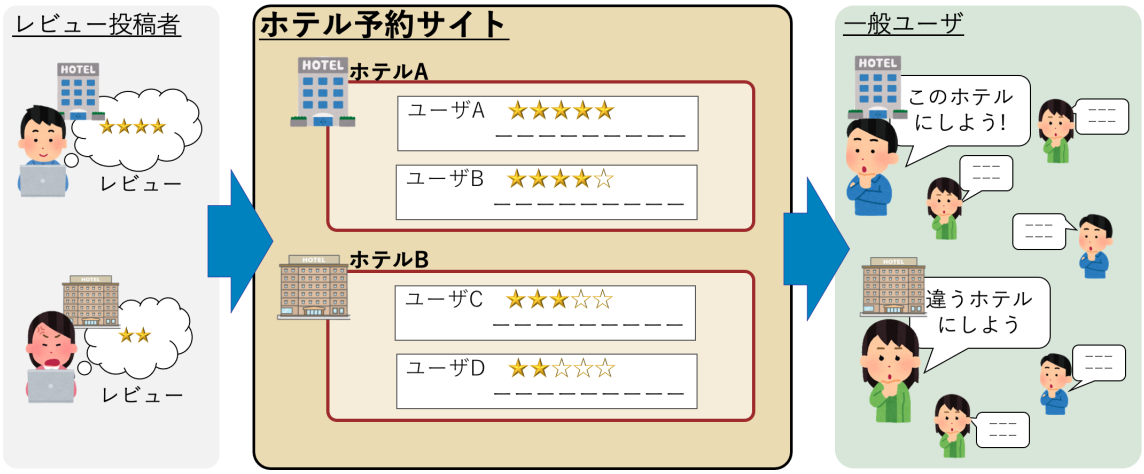


図 1　一般的なレビュー投稿の流れ

3. レビュー投稿者の不満を解消するホテル推薦システム

3.1 提案システムの概要

提案システムの概要を図 2 に示す．ユーザがホテル宿泊後レビューを投稿する．投稿された不満レビューを分類器を用いて宿泊したホテルに対するユーザ評価を作成する．また，同様の分類器を用いてホテルスコアも作成する．ユーザ評価とホテルスコアから不満点を解消する不満解消スコアを作成する．その後不満解消スコアと一致度の高いホテルを検索しレビュー投稿者に推薦する．

3.2 ホテルデータベース

本節では，過去のレビューデータを保存する，ホテルレビューデータベースについて述べる．ホテルレビューデータベースには，過去にレビューを行ったユーザのデータが，図 3 のように格納されている．このデータベース内にあるデータを用いて，推薦を行う．このデータベースの管理にはデータベース管理システム MySQl を使用しており，各カラムの内容はユーザ判別を行うための ” username ”，ホテル名の ” hotelname ”，レビュー文が保存されている ” review- text ”，それ以外の ” location ”，” room ”，” food ”，” bath ”，” service ”，” facility ”，” total ” は各項目に対する評価という形になっている．このデータベースを作成するサンプルデータとして，楽天トラベルのレビューデータ約 250 万件を採用した．

3.3 評価表現辞書

3.3.1 評価表現辞書の作成方法

本節では，先行研究で行われた，ホテル特化型評価表現辞書の作成方法について述べる．ユーザは様々な観点でホテルに対するレビューを投稿している．そこで，実際に使用されている表現を辞書に登録することで項目別の評価表現辞書を構築した．辞書の構築に利用するホテルレビューデータはホテル予約サイトである楽天トラベルのレビューデータを採用した．具体的な辞書の構築は，過去に投稿されたレビュー文に含まれている評価表現をキーワード，特徴，程度に分割し評価項目ごとに分類して辞書に登録する．

表 1　従来の評価項目と提案する詳細化した評価項目

従来の評価項目	詳細化した評価項目
立地	コンビニ・スーパーへのアクセス 周辺の飲食店 観光スポットへのアクセス ホテルまでのアクセス
食事	豪華さ 品数 量 味
部屋	眺め 広さ 清潔さ 静かさ
風呂	温泉 眺め
サービス	スタッフ 送迎
設備・アメニティ	アメニティ ネット環境 機器

表 1 に従来の評価項目と先行研究で提案した詳細化した評価項目の一覧を示す．これらの詳細化した評価項目に対する評価表現に対してスコアを設定し，評価表現辞書に登録する．評価表現に対してのスコア設定は，20 代の男女 5 名にアンケートを行い 0 から 5 の範囲でスコアをつけてもらったものの平均と，楽天トラベルのデータを元に設定している．

3.3.2 評価表現辞書に基づくスコアリング手法

本節では，先行研究で行われた，未知のレビューデータに対する自動スコアリングを行う手順を述べる．詳細化した評価項目ごとにスコアを付与するために，レビュー文に対して形態素解析エンジン mecab を用いて形態素解析を行う．分割されたキーワード，特徴，程度の組み合わせを評価表現辞書から探索を行い，辞書に当てはまる評価表現があれば，設定されているスコアをその評価項目に付与する．実際にスコアリングを行う

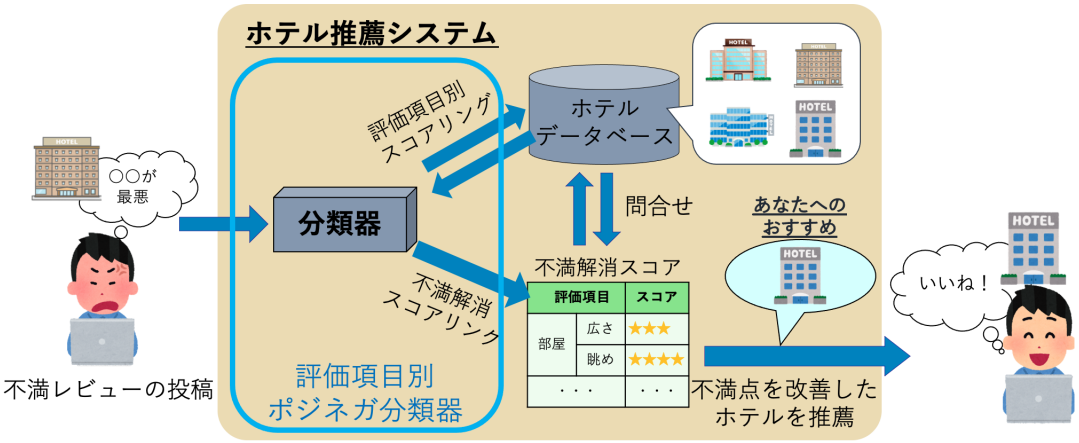


図 2　レビュー投稿者の不満を解消するホテル推薦システムの概要図

username	hotelname	location	room	food	bath	service	facility	total	reviewtext
		4	3	0	3	3	3	4	とても清潔な部屋でした。周辺は箱崎～千代にかけ
		5	5	4	5	5	5	4	貸し切り風呂がとても良かったです！色んなお風呂
		5	5	0	3	5	5	5	ダブルでの利用は初めてでした。相変わらず綺麗な
		4	4	0	3	4	4	4	フロントの方の対応も感じよく、部屋も清潔でした。
		2	5	4	4	3	4	4	部屋は広く、ベランダがあったのはとてもよかった
		5	5	2	5	5	3	5	ホテルの外観も内装もとっても素敵でした。部屋は広
		3	4	2	4	3	3	4	とにかく安いプランがあって部屋は広くて安くて良
		5	3	3	4	3	2	3	ビジネスホテルのユニットバスではくつろがないの
		5	5	5	4	5	4	4	この度は両親子供と４人でお世話になりました。部
		5	5	4	3	4	4	5	駅から近くとても便利で良かったです。部屋もきれい
		4	4	3	3	4	4	4	いつも利用させて頂いています。今回はいつもと違い
		5	5	5	4	4	5	5	深夜に予約・チェックインしましたが、対応が素早
		4	4	4	3	5	2	4	夜遅くのチェックインだったのにも関わらず、嫌な顔
		4	4	5	4	4	4	4	久しぶりの利用でした。朝食が以前と比べて、凄
		5	4	0	3	4	3	4	二度目の宿泊です。施設は隅々まで綺麗にされてお
		4	4	5	4	5	4	5	多少、古さを感じましたが、清潔感もあり、芦ノ湖
		5	4	4	5	4	3	5	熊本出張でいつも[redacted]を利用しているので、
		1	1	0	1	1	1	1	二度と泊りません。【ご利用の宿泊プラン】【室数
		5	5	5	5	5	5	5	社員が宿泊させていただきましたが、感想として、と
		2	3	0	2	3	3	3	値段相応の施設でした。単に寝て泊まる方には申し分
		4	4	0	3	5	4	5	服務人員真的非常親切、交流讓人愉快。旅館就在岡
		3	3	0	4	4	4	4	普通の都市圏でこの値段は十分過ぎるかと思います。
		4	4	3	3	5	5	4	同僚と泊まりましたが、満足してました。カウンター
		4	5	5	4	4	4	5	今回初めて利用させて頂きました。他宿泊施設の一

図 3　データベース内データの一例

過程を図 4 に示す. このレビューには，部屋の眺めに対する評価と部屋の広さに対する評価が書かれている. レビューを解析し，評価表現辞書に当てはまる表現があるので，設定されているスコアを部屋の「眺め」と部屋の「広さ」に付与している.

評価表現辞書を用いたユーザ評価の算出方法は，ユーザが投稿したレビューのみを評価表現辞書を用いてスコアリングする. その結果，スコアが 2.0 以下となった項目は不満．スコアが 4.0 以上となった項目は満足．言及がなくスコアが付与できなかった項目は言及なしと判定する．ホテルスコアの算出方法は，あるホテルの全レビューを評価表現辞書を用いてスコアリングする．その後評価項目ごとに平均値を計算しその値をホテルスコアとする.

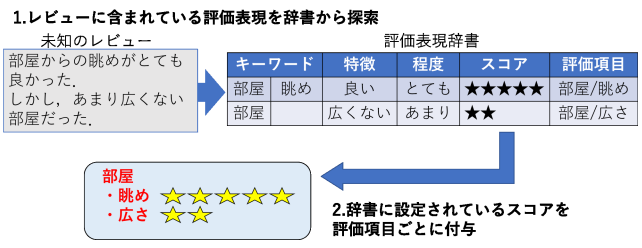


図 4　評価表現辞書を用いた項目別スコアリングの例

3.4　構築した分類器

本研究では部屋の各評価項目ごとに，fastText，BERT を用いてポジティブな要素を分類する分類器，ネガティブな要素を分類する分類器の 2 つの分類器を構築した．すべての分類器は 2 値分類を行う．学習は楽天トラベルのレビューデータ約 150 万件の中から各モデルのポジティブデータ，ネガティブデータとしてそれぞれ 1 万件ずつ抽出し学習に使用した．図 5，図 6 は構築したモデルの正解率，F1 スコアを示した図である．評価項目の後の添字の P,N はそれぞれポジティブ，ネガティブなモデルを表している．眺めのネガティブなモデル以外では，正解率，F1 スコアともに fastText で構築したモデルの方が優れている．眺めのネガティブなモデルは他のモデルと比べ正解率，F1 スコアの値が低く，学習データ等の改善によりモデルの改善を行う必要があると考えている.

3.5　分類器を用いたレビュー分類手法

本節では分類器を用いたスコアリング手法のついて説明する．まず，分類器は未知レビューを評価項目ごとに言及の有無を分類するために使用する．本研究では各評価項目に対してポジティブ分類器，ネガティブ分類器の 2 値分類の分類器 2 つを使用し総合的に分類している．図 7 に広さの評価項目を例に分類例を示す．広い分類器でポジティブ，狭い分類器でネガティブとなった場合，広いレビューと分類する．広い分類器でネガ

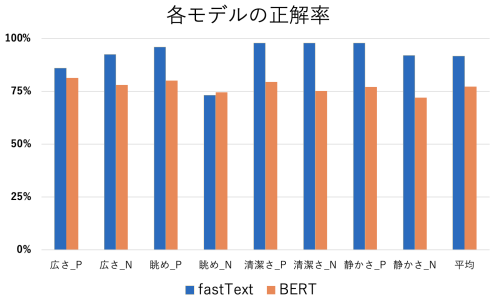


図 5　構築した各モデルの正解率

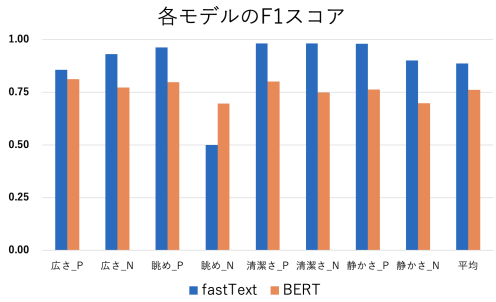


図 6　構築した各モデルの F1 スコア

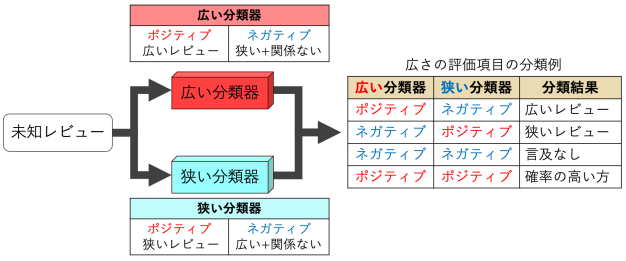


図 7　未知レビューの分類方法

ティブ，狭い分類器でポジティブとなった場合，狭いレビューと分類する．広い分類器でネガティブ，狭い分類器でネガティブとなった場合，言及のないレビューと分類する．広い分類器でポジティブ，狭い分類器でポジティブとなった場合，どちらかの分類器が誤分類しているため，より確率の高い方を採用する．

3.5.1　回帰モデルの作成

本節では，スコアを予測する際に使用する回帰モデルについて述べる．この回帰モデルは，各評価項目ごとに，ポジティブレビュー，ネガティブレビューの割合を説明変数，評価実験で被験者の回答から作成したスコアを目的変数として重回帰分析を行い作成した．以下に部屋の 4 つの評価項目の回帰モデルの計算式を示す．各文字の添え字は評価項目を表しており，1=広さ,2=眺め,3=清潔さ,4=静かさである．S はホテルスコアであり，計算の結果 1 以下の場合は 1 に，5 以上の場合は 5 とする．P はポジティブレビューの割合であり，ポジティブレビューが 10%であれば 0.1 とし小数に変換して計算する．N はネガティブレビューの割合であり，P 同様, 小数に変換し計算する．

$$S_1 = 2.91 + 7.64 * P_1 - 8.38 * N_1 \quad (1)$$

$$S_2 = 3.01 + 11.5 * P_2 - 13.1 * N_2 \quad (2)$$

$$S_3 = 3.3 + 3.62 * P_3 - 5.46 * N_3 \quad (3)$$

$$S_4 = 2.45 + 16.9 * P_4 - 2.77 * N_4 \quad (4)$$

3.5.2　ホテルスコアの算出方法

本節では，分類器を用いたホテルスコア算出方法について述べる．あるホテルに付いた全レビューを分類器を用いて評価項目ごとにポジティブレビュー，ネガティブレビュー，言及のないレビューがそれぞれ何パーセントあるか分類する．その後，ポジティブレビュー，ネガティブレビューの割合によってスコアを予測する回帰モデルによってスコアを決定する．

3.5.3　ユーザ評価の算出方法

本節では，ユーザが投稿したレビューから対象ユーザのホテルに対するユーザ評価の算出方法について述べる．投稿されたレビューを分類器を用いて評価項目ごとにポジティブ，ネガティブ，言及なしに分類する．ポジティブと分類された評価項目は満足，ネガティブと分類された評価項目は不満，言及なしと分類された評価項目は言及なしとする．

3.6　不満解消スコア

本節ではユーザへホテルを推薦する際に使用する不満解消スコアについて述べる．本研究で不満解消スコアとは，対象ユーザの不満解消に最低限必要なスコアと定義している．

3.6.1　不満解消スコアの算出方法

不満解消スコアは 3.4.2 節，3.4.3 節で算出したホテルスコア，ユーザ評価を用いて算出する．ユーザ評価は，対象ユーザが宿泊したホテルに抱いた印象であるため，不満となった評価項目については宿泊したホテルのホテルスコアよりも高い値を設定する．満足となった評価項目については，現状のホテルスコアでも満足しているため，同程度の値を設定する．言及なしとなった評価項目については，不満も満足もしていないため，宿泊したホテルのホテルスコアより少し高い値を設定する．このように設定することで，次回推薦後，宿泊したホテルに満足する可能性が高いと考えられる．また，各条件において，どの程度値を更新するかは今後の評価実験を通して最適な値を検証していく．

3.6.2　不満解消スコアを用いたホテル推薦

本節では不満解消を用いたホテル推薦について述べる．推薦時には，不満解消スコアとホテルデータベース内のホテルスコアを用いて推薦する．まず，ホテルデータベース内から全ての評価項目で不満解消スコア以上のホテルを推薦対象とする．これは，不満解消スコアがユーザの不満解消に最低限必要なスコアと定義しているからである．推薦対象のホテル群が決定できたら，その後予算に応じて総合評価順，不満があった評価項目のスコアの高い順，予算の安い順など複数のランキング方法を用いてユーザへ推薦する．

4.　評 価 実 験

4.1　実 験 概 要

提案手法，評価表現辞書，既存サイトの評価，各手法を用いて算出したスコアの精度比較を行うため実験を行った．実験手順としては，被験者である大学生男女 14 名にホテルのレビュー

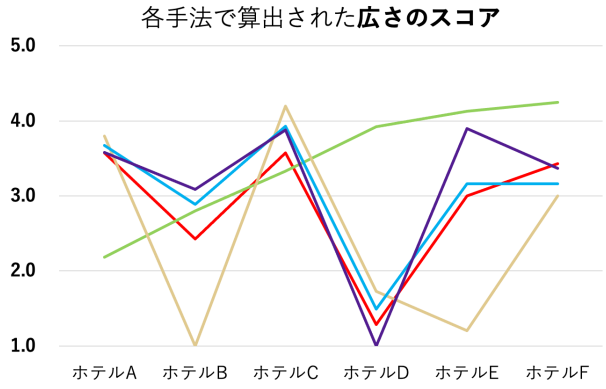


図 8　広さの Groundtruth と各手法のスコア

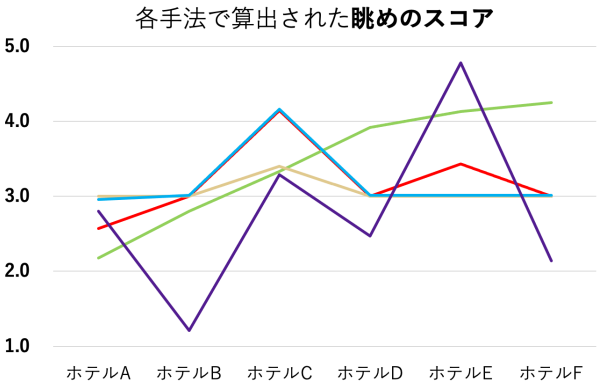


図 9　眺めの Groundtruth と各手法のスコア

を読んでもらい各評価項目についてリッカートの 5 段階で評価してもらった．被験者の回答の平均値を Groundtruth とし，各手法によって算出されたスコアと比較した．被験者は 3 軒のホテル各 30 レビューずつ読み，部屋の広さ，眺め，清潔さ，静かさの評価項目を評価する．被験者は 7 人ずつ 2 つのグループに分け計 6 軒のホテルに対する回答を収集した．

4.2　実 験 結 果

図 8, 図 9, 図 10, 図 11, に各評価項目の Groundtruth と各手法を用いて算出したスコアの結果を示す．グラフの縦軸は各手法で求めたスコア，横軸は被験者に提示したホテルに設定している番号である．赤が Groundtruth，水色が fastText を用いた手法，紫が BERT を用いた手法，緑が既存サイトのスコア，黄土色が辞書スコアである．各手法ごとに Groundtruth の予測精度を比較するため，Groundtruth との相関係数を取った．図 12 に Groundtruth と各手法の相関係数を計算したものをに示す．fastText を用いた提案手法ではすべての評価項目で相関係数が 0.8 以上と強い相関が見られた．BERT を用いた手法では「広さ」,「清潔さ」の評価項目では強い相関が見られたが，その他の項目では見られなかった．辞書スコアでは「眺め」の評価項目では強い相関が見られたが，その他の項目では見られなかった．サイトスコアでは「清潔さ」の評価項目では強い相関が見られたが，その他の項目では見られなかった．以上の結果より，fastText を用いた手法が最もスコア予測精度が高いと言える

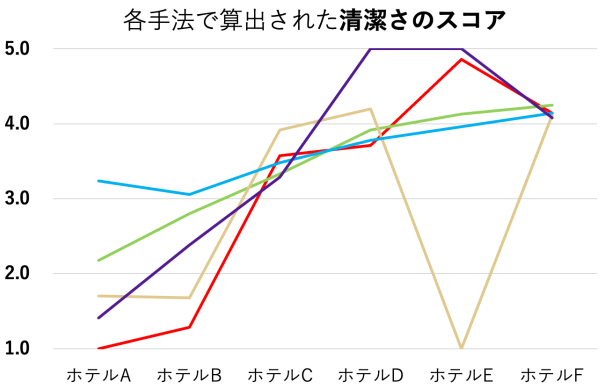


図 10　清潔さの Groundtruth と各手法のスコア

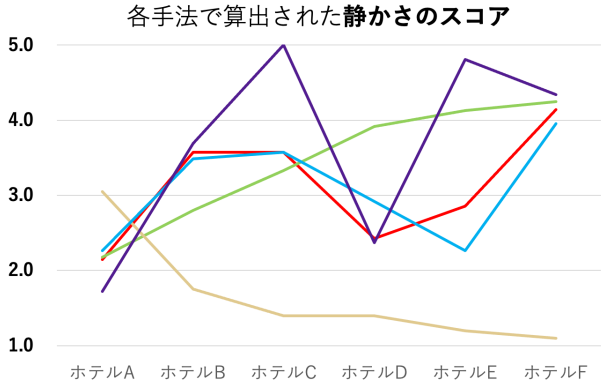


図 11　静かさの Groundtruth と各手法のスコア

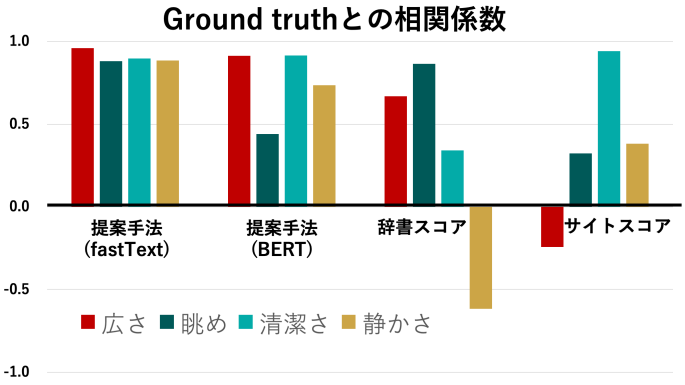


図 12　Groundtruth と各手法の相関係数

5.　ま　と　め

本研究では，ホテルレビューはこれからホテルを予約する一般ユーザにとってホテルの情報を知るための有益なものだが，ホテル宿泊後レビューを投稿するユーザが少ないことから，投稿されたレビューを分析し，レビュー投稿者の不満を解消するようなホテル推薦システムを提案した．本稿では，単語分散表現を用いてホテルレビューを評価項目別に自動スコアリングする手法を提案した．分類器を用いてホテルのレビューを評価項目ごとにポジティブレビュー，ネガティブレビュー，言及のないレビューに分類し，割合によってスコアを計算する回帰モデルを用いてスコアを算出する．評価実験により，提案手法と，

先行研究で作成された評価表現辞書，既存のホテル予約サイトのスコアを比較し，提案手法の有効性が示された．今後はこれらのシステム開発に取り組み，評価実験を行う予定である．

謝　　辞

本研究の一部は，科研費 (課題番号：19K12243，20H04293，22H03700，22K19837, 22K12281) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 楽天トラベル https://travel.rakuten.co.jp
[2] じゃらん net, https://www.jalan.net
[3] 安部克，中島伸介. レビュー自動スコアリング方式に基づくホテル推薦システム，DEIM Forum 2020, P1-28, 2020.
[4] 青柳直人, 蓮池隆. 宿泊予約サイトのユーザーレビューにおけるクレーム抽出と内容解析による課題抽出,FIT2021,CE-002,2021.
[5] 林部祐太. 宿レビューからの肯定的事実と推薦の対象の抽出, 言語処理学会 第 25 回年次大会,pp.554-557,2019.
[6] 田熊浩二，福原楓，亀井清華，藤田聡. 口コミを用いた嗜好抽出に基づくホテル推薦手法，情報処理学会第 78 回全国大会，5K-01,2016.
[7] 和田知亜紀，豊田哲也，大原剛三. 期待と現実のギャップに着目した宿泊施設レビューの解析とその検索への応用，情報処理学会，Vol.2018-IFAT-129No6,2018.
[8] Md. Shafiul Alam Forhad,Mohammad Shamsul Arefin,A.S.M.Kayes,Khandakar Ahmed Mohammad,Jabed Morshed Chowdhury,and Indika Kumara.An Effective Hotel Recommendation System through Processing Heterogeneous Data,in the 5th International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2020.

追加事前学習によるレビューと商品属性の対応関係学習

河田　友香†　　山本　岳洋†　　大島　裕明†　　藤田　澄男††

† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† LINE ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1–3
E-mail: †ad22l019@guh.u-hyogo.ac.jp,t.yamamoto@sis.u-hyogo.ac.jp,ohshima@ai.u-hyogo.ac.jp,
††sufujita@lycorp.co.jp

あらまし　本研究では,「画素数: 3,240 万，撮像素子: フルサイズ，重量: 800g」のような商品属性と,「遊ぶ子どもの表情もくっきり撮影することができます.」のようなレビューとの関連性を予測できる言語モデルを提案する．このモデルを構築するため，事前学習済みの BERT を使用し，入力として受け取った商品属性とレビューから，その商品属性をもつ商品に対するレビューなのか対応関係を予測するようにモデルを学習する．テキストであるレビューと表データである商品属性の対応の学習方法が技術的課題として挙げられる．そこで，テキストをマスクし，元のトークンを予測する追加事前学習を行うことで，レビューと商品属性の対応関係を学習する．価格.com のカメラを対象に実験を行い，追加事前学習を行うことで関連度を予測する精度が向上することが分かった．

キーワード　レビュー，言語モデル，BERT

1　はじめに

商品を購買しようとするユーザは多様な情報を収集し，購買の決断に至る．商品を選ぶ際に参考にする情報の一つとして，商品属性が挙げられる．商品属性とは，商品の様々な属性名とそれに対応する属性値の集合である．例えば，カメラであれば「撮像素子」という属性名に対して「フルサイズ」,「APS-C」といった属性値があり,「画素数」という属性名に対して,「3,410 万画素」,「2,040 万画素」といった属性値がある．

商品属性について，商品に詳しいユーザであれば,「画素数が 3,410 万画素あれば風景写真も美しく撮影できる」や「フルサイズであれば色鮮やかに撮影できる」のように商品属性がどのような特徴に繋がるのかを理解している．そのため，商品属性をもとに商品を探すことが可能である．

反対に，あまり商品に詳しくないユーザは商品属性と関連づけながら商品を探すことはできない．例えば，画素数が風景写真を色鮮やかに撮影するために重要な属性なのか，素早く動くスポーツ選手を撮影するために重要なのか理解できない．商品に詳しくないユーザは商品説明文やレビューを読むことで，どのような商品であるかを理解する．しかし，レビューの中には「色鮮やかに撮影できる」カメラとは書いてないが，商品属性に着目すると「色鮮やかに撮影できる」カメラであることも考えられる．知識のないユーザにとって，商品属性と商品属性と結びつく評価の関連を理解しながら商品を探すことは難しい．

我々はこれまでに，商品属性とレビューの対応関係を学習し，これらの関連を理解したモデルの構築手法を提案した [27]．提案した手法の概要を図 1 に示す．このようなモデルを構築することで，テキストを入力し，そのテキストと最も関連の高い商品属性をもつ商品を推薦することができる．さらに，レビューや商品説明文には書かれていないが，商品属性からレビューに書かれる可能性がある情報を提示することができる．このように多様な見方で商品に関する情報を提示することが可能となる．

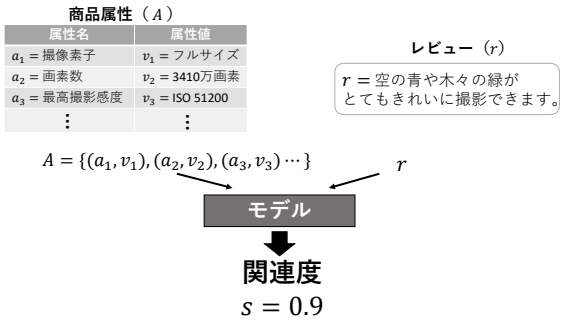


図 1　レビューと商品属性を入力として，その関連度を求める手法の概要図.

これまでに提案したモデルは，BERT にレビューと商品属性を入力すると，レビューと商品属性の関連度を算出するモデルである．正例を入力された商品属性を持つ商品に対して実際につけられたレビュー，負例を入力された商品属性を持つ商品とは異なる商品に対して実際につけられたレビューとして，学習を行った．

提案手法を構築する上で技術的課題として，表データである商品属性とテキストデータであるレビューの対応関係をどのように学習するのか，また，量的データをどのように処理するかが挙げられる．これまでに我々が提案した手法では，量的データの大小関係を捉えるために量的データをカテゴリ化を行っていた．また，商品属性とレビューの対応関係を学習するために，ファインチューニングのみを行っていた．

本研究では追加事前学習を行うことで，商品属性とレビューの対応関係をより理解したモデルの構築に取り組む．BERT モデルに対して，追加事前学習を行うことでタスクの性能向上に

寄与することが既存研究で明らかになっている [17] [19]．追加事前学習とは，入力テキストにマスク処理を行い，マスクされたトークンを当てながら学習する手法である．追加事前学習を行うことで，テキストと属性値の関連や，属性値同士の関連を学習することができる．さらに追加事前学習に加えて，量的データのカテゴリ化手法についても 3 種類の手法を提案し，量的データの大小関係を考慮できる手法の比較を行った．

構築したモデルの評価として，2 つの評価を行った．1 つ目は関連度予測タスクである．商品属性とレビューを入力し，商品属性をもつ商品に付けられたレビューかどうかを正しく当てられるかを評価した．

2 つ目は，レビューランキングタスクである．テストデータに含まれるレビューと商品属性の関連度をもとにランキングを作成し，実際に入力した商品属性を含む商品に適合したレビューが上位にランキングされているかを評価した．

実験の結果，関連度予測タスク，レビューランキングタスクの両方の評価において，追加事前学習を行うといずれのモデルでも精度が向上することが分かった．

2 関 連 研 究

2.1 商品属性やレビューを用いたテキスト分析

商品にまつわるレビューのみ，また商品の属性やユーザの情報のみを用いたテキスト分析が行われている．Chen らは，ユーザの情報をもとにナレッジベースのニューラルネットワークモデルを構築し，ユーザの興味関心に合わせた商品説明文を生成する手法を提案している [2]．McAuley らは，Yes/No で判断できる質問に対して，レビューから回答を見つけ，答えを提示するモデルを提案している [15]．Roy らは，商品の素材についてなどの商品属性から回答できる質問に対して，商品属性から自動的に回答を生成するモデルを XLNet や BERT をもとに構築している [18]．Zhang らは，ユーザが行った商品に関する質問から，その質問に対する答えを予測し，答えに関するレビューを提示している [22]．Dong らは，ユーザの情報，評価レート，商品の ID を属性としてレビューを生成するモデルを構築している [6]．Deng らは，レビューから意見を集約し，回答を生成するモデルを提案している [4]．Chen らは，レビュー文から質問に関係するスニペットを生成し，レビュー文に基づいて質問回答を生成するモデル RAGE を構築している [3]．

また，レビューと属性の両方を用いたテキスト分析が行われている．Feng らは，商品属性とレビューを関連づけるモデルを構築し，商品属性とレビューの関係を理解した上で回答を生成する手法を提案している [7]．Gao らは，商品属性と属性値，過去にユーザによってつけられたレビューを用いて，質問に関係する商品属性に言及したレビューから自動的に回答を生成している [9] [10]．Zhang らは，商品属性と BERT でベクトル化したテキストを結合する MA-Transformer Encoders を用いて，商品属性とテキストをベクトル化する MA-BERT を提案している [24]．

本研究では，商品属性とレビューの両方を用いて両者の関連付けを行い，言語モデルを構築する．

2.2 テキストからの属性抽出

文章内から属性を抽出する研究が行われている．Iwanari らは，ブログ内の文章から，対象物とその対象物の評価を抽出し，属性に応じて対象物を順位づけた [11]．Fuchs らは，商品説明文と商品の属性名，属性値を与え，商品をあらわすために重要な属性値を抽出した [8]．内田らは，レビューでの相対的な評価や単体での評価から，商品属性ごとの順序関係を推定し，信頼性の評価を行った [28]．Majumder らは，商品のタイトルから商品属性を抽出するモデルを双方向 LSTM や双方向 LSTM-CRF モデルを元にして構築した [14]．Zheng らは，複雑な記述がなされている商品説明文から，知りたい属性の属性値を Opentag モデルを提案し，抽出した [26]．Chen らは，BERT の固有表現抽出を用いて，商品説明文やタイトルから商品属性を抽出した [25]．Wang らは，属性名を問いとして与え，属性値を商品説明文から抽出するモデルを BERT をもとに構築した [21]．

本研究では，レビューと商品属性を入力として受け取り，両者の関連を計算するモデルを提案する．

2.3 言語モデルでの数値の処理

数値を言語モデルで扱う研究が行われている．Naik らは，数値の大小関係や数え上げを単語の分散表現がどの程度捉えているか分析するフレームワークを提案した [16]．Chen らは，ニューラルネットワークモデルに数値を学習させることで，テキスト内に誇張して記述されている数値表現を検出できるか分析した [1]．Talmor らは数字の比較や質問応答などの推論タスクにおいて，異なる言語モデルでの推論能力を比較した [20]．Lin らは，BERT などの事前学習済みモデルでは，数値についての知識をどの程度理解することができているのかを調査した [13]．Jin らは，数値を考慮した生成的言語モデルを構築するため，仮数と指数を表す埋め込み表現を GPT に組み合わせた NumGPT を提案した [12]．

本研究では，商品属性に含まれる数値の処理について検討し，商品属性とテキストの対応関係の学習方法を提案する．

3 問 題 定 義

本節では，本研究で取り組む問題について定義する．ある商品 p の商品属性 A は (1) のように表すことができる．

$$A = \{(a_1, v_1), (a_2, v_2), \cdots (a_k, v_k)\} \tag{1}$$

ここで k は属性の個数とする．また，$a_i (1 \leqq i \leqq k)$ は属性名を表し，$v_i (1 \leqq i \leqq k)$ は属性値を表す．カメラを例とすると，商品属性 A は $A = \{($ 画素数, 3,410 万画素$)$, $($撮像素子, フルサイズ$)$, $\cdots$ $($重量, 800g$)\}$ のように表すことができる．商品 p に対するレビューを r と置く．レビュー r は，単語の系列である．i 番目の単語を w_i と置くと，(2) のように表すことができる．

$$r = w_1 w_2 \cdots w_L \tag{2}$$

ここで L はレビュー r 中の単語の個数である．商品属性 A と

レビュー r を入力として受け取り，商品属性 A に対して，レビュー r が書かれる可能性のある文章なのかを予測し，関連度 s で表す．関連度 s とは，商品属性 A とレビュー r がどの程度関連しているかの度合いを表す．関連度 s は，レビュー r と商品属性 A に関連がなければ値が 0 に近くなり，レビュー r と商品属性 A に関連があれば値が 1 に近くなる．

具体的には，$A = \{$(画素数, 3,410 万画素), (撮像素子, フルサイズ), $\cdots$ (重量, 800g)$\}$ のような商品属性が与えられたとする．このカメラは撮像素子がフルサイズであり，画素数が 3,410 万画素であるため，美しく撮影することができると考えられる．このようなカメラの商品属性に対して,「空の青や木々の緑がとてもきれいに撮影できます.」のようなレビューは商品に対する評価としてあり得るテキストであるため，関連度 s は 1 に近くなる．反対に「発色はきれいだが，画質は良いといえない.」のようなレビューは，その商品に対する評価とは考えにくいため，関連度 s が 0 に近くなる．このようなモデルを構築することで，商品属性 A とレビュー r が関連しているかを理解することができる．

4 提案手法

本節では，我々が提案したモデルについて述べる [27]．提案モデルは，BERT に商品属性とレビューを入力し，それらの関連度を予測する手法を提案する手法である．BERT が入力として受け取る商品属性は表データである．さらに，属性値には量的データを含むものもある．表データである商品属性とテキストであるレビューの対応を学習する方法について述べる．

4.1 モデルの概要

本研究で提案するモデルの概要について述べる．提案モデルについて図 2 に表す．提案モデルでは入力として，レビュー，商品属性を受け取り，レビューと商品属性の関連度を算出する．

モデルをレビューが入力された商品属性を持つ商品に実際に書かれていたレビューであるかを分類する問題として学習させる．正例として，入力された商品属性を持つ商品に実際につけられたレビュー，負例として，異なる商品属性を持つ商品に実際につけられたレビューを用いる．これらの正例と負例を当てるようにモデルを学習させる．

4.2 BERT

本研究では, 様々な自然言語処理タスクで使用される BERT [5] を元に言語モデルの構築を行う．BERT は Transformer Encoder を用いたニューラルネットワークのモデルである．大量のテキストでファインチューニングを行うことで様々な自然言語処理タスクを精度よく解けることが知られている．

BERT では入力されたテキストを分割し，トークンとして入力する．商品属性は，属性名と属性値を「画素数:3410, 重量:800 $\cdots$」のように「:」と「,」を用いて結合し，テキストとして BERT に入力した．レビューと商品属性の間には，レビューと商品属性を区別するために境界となるトークンとして [SEP] トークンを加えた．

また，Segment ID を使用し，レビューに 0，商品属性に 1 を割り当てる．Segment ID は BERT にテキストを 2 文入力する際に，入力するトークンが 1 文目のトークンなのか 2 文目のトークンなのかを判別するために使用される．提案手法では，Segment ID をレビューと商品属性の判別のために使用した．

4.3 量的データの処理

本節では量的データをどのように処理するかについて述べる．属性値には,「3,410 万画素」や「800g」のように量的データを含むものも多く存在する．これらの値を言語モデルに入力すると，それぞれの数値の大小関係や規模感を理解することが難しい [23]．商品属性においてこれらの数値情報は商品を表す重要な情報であるため，適切に処理をする必要がある．そこで，属性値の量的データのカテゴリ化を行うことで，量的データの大小関係を理解できるようにする．量的データのカテゴリ化とは，画素数であれば，4,000 万画素のカメラは「画素数が非常に高い」や 2,000 万画素のカメラは「画素数はあまり高くない」のように，閾値を定め，その閾値に基づいて量的データをカテゴリに分割する．カテゴリ化した量的データを含む属性値を，BERT に入力する際の表現手法として，3 つの手法を比較する．

- 属性ごとにカテゴリを表すトークンで表現
- カテゴリを数値で表現
- 属性間で共通のカテゴリを表すトークンで表現

[27] で提案した，属性ごとにカテゴリを表すトークンで表現する手法，カテゴリを数値で表現する手法に加え，属性間で共通のカテゴリを表すトークンで表現する手法を提案する．それぞれの量的データのカテゴリ化の手法について説明する．

4.3.1 属性ごとにカテゴリを表すトークンで表現

画素数や重量など，属性ごとにカテゴリの数，閾値を定め，それぞれに独自のトークンを用意する．例えばここでは，5 段階のカテゴリで画素数と重量を表現することを考える．画素数が 5 段階中，最も高い 5 で表現できたとする．このとき，画素数の属性値を表すトークンとして [gaso_5] のように表現できる．重量は 5 段階中，最も低い 1 で表現できたとする．この重量を表現するときには [weight_1] というトークンを用意する．このように，属性ごとにカテゴリを表現するトークンを準備し，それによって属性値を表現する．属性ごとに値の大きさを表すトークンを用意することで，それぞれの属性における属性値の大小関係や意味合いを表現することができる．

4.3.2 カテゴリを数値で表現

数値を用いてカテゴリを表現する．画素数が 5 段階のカテゴリで表現でき，そのカメラの属性値の画素数が 5 段階のカテゴリの中で最も高い 5 であるときを考える．この時，画素数の属性値は「5」という数字で表現する．新たに追加するトークンがないため，事前学習済みモデルに新たに重みを追加する必要がないという利点が挙げられる．

4.3.3 属性間で共通のカテゴリを表すトークンで表現

複数の属性で共通するカテゴリごとの大きさを表すトークンを追加し，属性値を表現する．例えば画素数が 5 段階中，3 というカテゴリで表現できるとする．このときに，画素数の属性

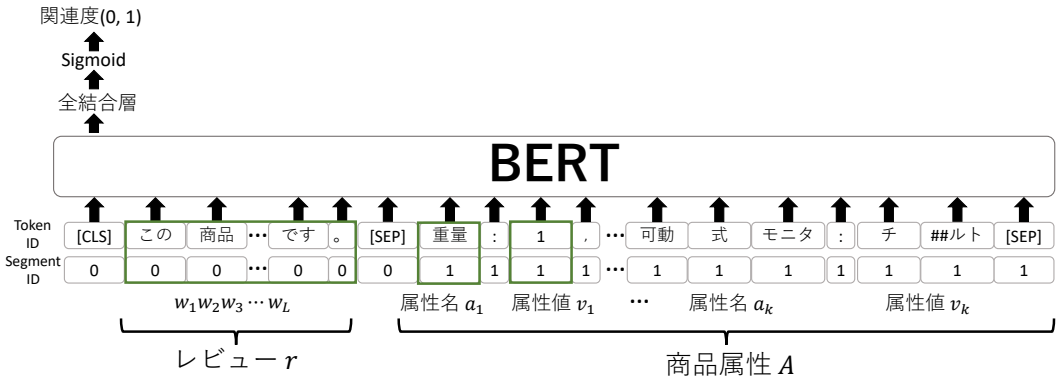


図 2　提案モデル．レビューと属性名，属性値を入力として，入力した商品属性がレビューと対応しているかを判断する．提案手法は実際に入力した商品属性を持つ商品に付けられたレビューを正例，異なる商品属性を持つ商品に付けられたレビューを負例として学習する．

表 1　量的データを含む属性値の表現方法．

属性値の表現方法	入力形式
量的データをそのまま入力	画素数:3410, 重量:800 ⋯
属性ごとにカテゴリを表すトークンで表現	画素数:[gaso_1], 重量:[weight_3] ⋯
カテゴリを数値で表現	画素数:1, 重量:3 ⋯
属性間で共通のカテゴリを表すトークンで表現	画素数:[level_1], 重量:[level_3] ⋯

値を表すために，[level_3] とトークンを置く．同様に重量が 5 段階中，3 というカテゴリで表現できるとする．その場合，重量も [level_3] というトークンで置く．異なる属性において，同一のカテゴリにあてはまる時は，共通のトークンで属性値を表現する．このトークンを数値トークンと定義する．数値トークンを導入することで，異なる属性間でトークンを変えずに属性値を表現できる．新たに追加するトークンの数を抑えることができ，新たに学習する重みを少なくすることができる．

本研究で比較を行う量的データを含む属性値の表現方法について，表 1 に表す．

4.4　追加事前学習

BERT モデルに対して追加事前学習を行う．追加事前学習を行うことでタスクの性能が向上することが既存研究で明らかになっている [17] [19]．追加事前学習とはファインチューニング前の BERT の事前学習済みモデルに対して，訓練データを用いて [MASK] トークンの予測を行う．追加事前学習の概要を図 3 に示す．

追加事前学習では，入力トークンの 15%に対してマスク処理を行う．マスク処理とは入力トークンの 15%うち，80%のトークンは [MASK] トークンに置き換える．10%のトークンは，入力トークンとは異なるトークンに置き換える．また，残り 10%のトークンは入力したトークンと同じトークンとして入力する．このようにマスク処理をしたトークンについて，元のトークンがどのようなトークンであったかを予測する．予測したトークンと元のトークンについて損失値を計算し，その損失値ができる限り小さくなるように学習を進める．追加事前学習により，トークン同士の関連を学習することができる．

本研究での追加事前学習では，レビューと属性値に対してマスク処理を行い，その処理したトークンを予測する．この時，属性名はどのテキストにおいても共通で入力されており学習に大きな影響を与えないため，マスク処理を行わない．レビューと属性値に対してマスク処理を行うことで，レビューの表現に対して自然だと考えられる属性値や，属性値に対して自然な表現であるテキストを学習することができる．

4.5　正例と負例の構築

本節では，正例と負例をどのように構築するかについて述べる．モデルは，レビューに対して，入力された商品属性が関連しているかを当てるように学習を行う．正例を入力された商品属性を持つ商品に実際につけられたレビュー，負例を異なる商品属性を持つ商品に付けられたレビューと定義し，それらを当てるように学習する．

負例として用いる，異なる商品属性を持つ商品の選び方について述べる．異なる商品名を持つすべての商品からランダムに商品を選択し負例として用いると，商品属性が類似している商品も多くみられる．これは，後継商品やキットとして付いているレンズの違いなどにより，異なる商品名で販売されているからである．

商品属性が極端に類似している商品を負例として学習に用いると，商品属性とレビューが関連しているのにも関わらず負例と定義されるため，モデルが正しく商品属性とレビューの関連を学習することができない．そこで，商品同士において，商品属性のコサイン類似度を求め，そのコサイン類似度が一定の値以下の商品を負例として用いる．このことにより，極端に類似した商品属性を負例として学習に用いることを防ぐことがで

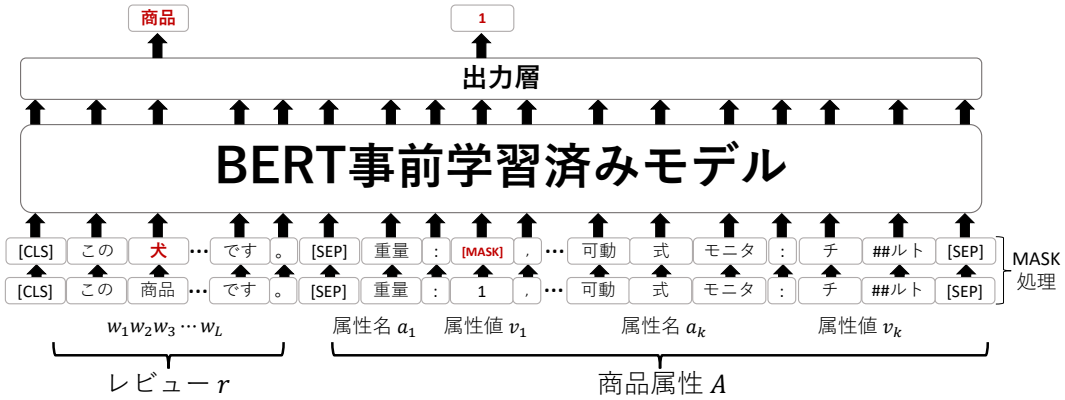


図 3　追加事前学習の概要．レビューと属性値に対して，マスク処理を行う．マスク処理したトークンに対して，元のトークンを予測するように学習を行う．

きる．

コサイン類似度の求め方について説明する．属性値は，カテゴリカルデータと量的データの 2 種類の値がある．カテゴリカルデータについては，ダミー変数に変換する．量的データについては，正規化を行う．ダミー変数や正規化を行い置き換えた属性値を用いてコサイン類似度を求め，レビューに対して負例として用いる商品を選択する．本研究では負例となる商品は，コサイン類似度が 0.95 以下の商品をランダムで選択する．

4.6　モデルの学習

提案モデルでは，入力された商品属性に対して，入力されたレビューがその商品属性を持つ商品に付けられる可能性があるレビューかを予測する．

正解ラベルとの誤差を計算し，損失値を求め，逆伝播を行うことで学習する．損失関数として，バイナリ交差エントロピーを用いた．バイナリ交差エントロピーの式を表す．いま，o_i を訓練データ内の i 番目の商品に対して，正例であるとモデルが予測した確率とする．また，y_i を i 番目の商品の正解ラベルとする．ここで，商品の正解ラベルとは，入力された商品属性を持つ商品に付けられたレビューであるとき y_i は 1，そうでなければ y_i は 0 をとる．

訓練データの個数を N，i 番目のデータに対して正例であると予測した確率を o_i，i 番目のデータの正解ラベルが y_i であるとき，以下のバイナリ交差エントロピーを最小化することでモデルを学習する．

$$L = -\sum_{i=1}^{N}(y_i \log o_i + (1 - y_i)\log(1 - o_i)) \qquad (3)$$

4.7　関連度の計算

提案手法では，入力された商品属性とレビューが関連している度合いを関連度と定義し，この関連度を計算する．具体的に関連度とは，モデルに商品属性 A とレビュー r を入力することで求められる，両者が対応している確率である．

関連度は，シグモイド関数を用いることで求められる．ある商品の商品属性 A とレビュー r を用いて，関連度 s を求める．この関連度 s はシグモイド関数を用いるため，$(0, 1)$ の間で表すことができ，確率として表現できる．

5　実　　験

本節では，本研究で提案した手法を用いた実験について述べる．提案した手法について，2 つの方法で評価した．1 つ目は関連度予測精度の評価である．関連度を関連あり，関連なしの二値と定義し，それらが正しく当てられるかを評価する．2 つ目はテストデータのランキングタスクである．商品属性に対してテストデータのテキストとの関連度をそれぞれ求める．この関連度をもとにレビューをランキングし，実際に入力した商品のレビューが上位に現れるかを評価する．

5.1　デ　ー　タ

本研究では，デジタル一眼レフカメラを対象にし，価格.com[1] のレビューと商品スペック表を用いて実験を行う．デジタル一眼レフカメラは，様々な商品属性を持ち，スペック表を見ただけではどのような商品かを理解することが難しい．またレビューからどの商品属性に言及されているものなのかを判断することは難しいと考え，本研究ではデジタル一眼レフカメラを対象とする．

価格.com のデジタル一眼レフカメラが掲載されているページでは，デジタル一眼レフカメラとして，一眼レフ，ミラーレス一眼レフが含まれている．対象商品数は 458 商品，また，総レビュー数は 23,246 件である．

5.1.1　商 品 属 性

本研究では 17 種類の属性を対象として，実験を行った．提案手法は 4.3 節で述べたように，属性値の形式が量的データである場合，カテゴリ化を行い入力する．カテゴリ数については，スタージェスの公式を使用し，決定した．実験に使用した商品と，その属性値の形式，例，カテゴリ数について表 2 に示す．

5.1.2　レビュー

レビュー中には，「配送もスムーズで満足です.」や「セールだったので購入しました.」のような購買情報に関するテキスト

1：https://kakaku.com/

表 2　使用するカメラの商品属性とその形式，例．形式が数値である場合，カテゴリ化を行う際のカテゴリ数を示す．

属性名	形式	属性値の例	カテゴリ数
メーカー	文字列	CANON，SONY	-
タイプ	文字列	ミラーレス，一眼レフ	-
画素数	数値	1600 万	7
撮像素子	文字列	フルサイズ，APS-C	-
連写撮影	数値	10	6
最低撮影感度	数値	200	4
最高撮影感度	数値	25600	5
最低シャッタースピード	数値	1/16000	4
最高シャッタースピード	数値	60	4
撮影枚数	数値	210	7
防塵・防滴	文字列	あり，なし	-
幅	数値	106.5	8
高さ	数値	64.6	8
奥行	数値	33.3	8
可動式モニタ	文字列	チルト，バリアングル	-
重量	数値	240	8
カラー	文字列	ブラックシルバー	-

もある．これは商品属性に関するテキストではないので，学習として用いるのはふさわしくない．

そこで，評価項目に基づいて書かれたレビューを使用した．価格.com では，評価項目ごとにレビューが書かれていることが多い．例えばカメラであれば,「デザイン」,「画質」,「操作性」,「バッテリー」,「携帯性」,「機能性」,「液晶」,「ホールド感」の 8 つの評価項目でユーザは商品を評価し，レビューを書く．画質であれば，撮影感度や画素数，撮像素子などに関連しており，また，携帯性であれば，幅や高さ，重量などの属性が関連していることが考えられる．このように評価項目に基づくレビューには，商品属性に関連したレビューが書かれている．そこで，この評価項目に基づいたレビューを抽出する．

価格.com では評価項目を「【】」で囲み，その後にレビューを書いていることが多い．具体的には,「【画質】3,300 万画素でもかなり高精細.」のように,「【】」の中に評価項目を記述し，その後に，商品のレビューが書かれている．評価項目について記述されたレビューを抽出し，学習に使用する．複数文に分かれているレビューについては，1 文ずつ抽出する．

抽出したレビューの長さは，長い文章から短い文章まで多様な長さの文章が抽出できる．短すぎる，あるいは長すぎるレビューを除去するため，本研究では，全文書の中で 25%から 75%の文書長であるレビューを学習に用いる．25%から 75%の文書長であるレビューの文字数は 19 文字以上 47 文字以下であった．この条件でレビューを取得すると 107,194 文取得できた．

5.2 実験設定

BERT の事前学習済みモデルは東北大学乾研究室の事前学習済みモデルを使用する [2]．学習率は追加事前学習は 2.0×10^{-5}，ファインチューニングは 2.0×10^{-6} とした．訓練データを学習するエポック数については，追加事前学習，ファインチューニングともに検証データに対して Patience を 10 に設定し，早期終了に基づき決定した．

入力するトークン数は，256 トークンに設定した．訓練データ，検証データは 408 商品のレビューを対象とした．訓練データ 86,514 件，検証データ 9,613 件として学習を行った．また，テストデータは 50 商品 11,067 件のテキストを用いる．訓練データ，検証データにはテストデータと同じ商品属性を持つ商品は含まれていない．

5.3 比較手法

実験で比較する手法について述べる．量的データの大小関係を BERT で考慮するために，以下の入力手法を比較する．

- **属性ごとにカテゴリを表すトークンで表現**

4.3.1 節で述べた，カテゴリ化した量的データを含む属性値を，属性ごとにカテゴリを表す特殊トークンを用いて表現する．

- **カテゴリを数値で表現**

4.3.2 節で述べた，カテゴリ化した量的データを含む属性値を数値を用いて表現する．

- **属性間で共通のカテゴリを表すトークンで表現**

4.3.3 節で述べた，カテゴリ化した量的データを含む属性値を数値トークンという特殊トークンを用いて表現する．

- **量的データをそのまま入力**

ベースライン手法．量的データを含む属性値をカテゴリ化せずに入力する．例えば,「画素数:3410, 重量:800 … 」のように，BERT モデルの入力に量的データをそのまま入力する．

これらの量的データの入力手法について，追加事前学習の有無で比較する．

5.4 評価方法

本研究では 2 つの評価方法で，提案手法を評価する．

5.4.1 関連度予測タスクによる評価

関連度を関連あり，関連なしの二値とみなし，それらを正しく予測できるかどうかを評価する．ここでは，関連あり，関連なしとは実際にその商品属性を持つ商品に与えられたレビューが付いたかどうかである．関連ありは入力した商品属性を持つ商品に付けられたレビュー，関連なしは異なる商品属性を持つ商品に付けられたレビューと定義する．関連ありと関連なしを正しく当てられているかを正解率で評価する．

5.4.2 レビューランキングタスクによる評価

テストデータに含まれるレビューを商品属性との関連度が高い順にランキングを行い，実際に入力した商品属性を持つ商品につけられたレビューが上位にランキングされているかを評価する．この評価により，実際にその商品に書かれたレビューについて，高い関連度で予測できているかを評価することができる．ランキングした結果について，平均逆順位（MRR）を用いて評価を行う．

6 結　　果

6.1 関連度予測タスク

関連度予測タスクの結果について述べる．各手法における関

2：https://github.com/cl-tohoku/bert-base-japanese-v3

表 3　各モデルにおける関連度予測タスクの正解率.

量的データの入力方法	追加事前学習	正解率
量的データをそのまま入力	あり	0.7359
量的データをそのまま入力	なし	0.6937
属性ごとにカテゴリを表すトークンで表現	あり	0.7280
属性ごとにカテゴリを表すトークンで表現	なし	0.7111
カテゴリを数値で表現	あり	0.7132
カテゴリを数値で表現	なし	0.7029
属性間で共通のカテゴリを表すトークンで表現	あり	0.7193
属性間で共通のカテゴリを表すトークンで表現	なし	0.7045

表 4　各モデルにおけるレビューランキングタスクの MRR.

量的データの入力方法	追加事前学習	MRR
量的データをそのまま入力	あり	0.3331
量的データをそのまま入力	なし	0.2029
属性ごとにカテゴリを表すトークンで表現	あり	0.2965
属性ごとにカテゴリを表すトークンで表現	なし	0.2102
カテゴリを数値で表現	あり	0.3518
カテゴリを数値で表現	なし	0.1717
属性間で共通のカテゴリを表すトークンで表現	あり	0.3297
属性間で共通のカテゴリを表すトークンで表現	なし	0.2033

連度予測タスクの正解率について表 3 に示す．まず，追加事前学習なしの場合について正解率を比較する．このとき，ベースラインが正解率が最も低く，量的データの入力を工夫した提案手法はベースラインよりも正解率が高いことが分かる．このことは，追加事前学習を行わない条件においては，量的データの入力を工夫することにより，商品属性とレビューの関連度予測の精度が向上したことを示している．

つぎに，追加事前学習ありの場合となしの場合について正解率を比較する．このとき，どのモデルも追加事前学習を行うことで正解率が向上していることが分かる．このことは，レビューと商品の関連を理解したモデルを構築するうえで，追加事前学習が有用であることを示している．

最後に，追加事前学習ありの場合について正解率を比較する．追加事前学習ありのモデルの中では，ベースラインが最も高い正解率となることが分かる．追加事前学習を行わない場合は，ベースラインは最も低い正解率であったが，追加事前学習を行うことで量的データの入力を工夫したモデルよりも高い正解率を達成することが分かる．

これらのことから，以下のことが推察される．追加事前学習を行わない場合は，モデルの学習に用いるデータが比較的少ないため，ベースラインでは量的データを適切に扱うことができず，正解率が低くなってしまう．量的データをカテゴリ化することで量的データをモデルが適切に扱えることができ，ベースラインよりも高い正解率を達成できたと考えられる．一方，追加事前学習を行うことで，学習に利用するデータが比較的多くなるため，ベースラインでも商品属性と量的データの関連度を適切に推定できるようになったのではないかと考えられる．

6.2　レビューランキングタスク

レビューランキングタスクの結果を表 4 に示す．まず，関連度予測タスクと同様に，どのモデルについても追加事前学習を行うことで，MRR が向上していることが分かる．一方で，追加事前学習なしの場合について比較すると，量的データをカテゴリを表すトークンを用いる手法がベースラインの MRR を上回っているものの，他の手法はベースラインを下回っていた．追加事前学習ありの場合についてみてみると，量的データをカテゴリ化し数値で表現する手法が 0.3518 とベースラインの 0.3331 よりも高い MRR を達成していることが分かった．

6.3　考　　察

関連度予測タスクとレビューランキングタスクの結果から，どちらのタスクでも，追加事前学習を行うことでいずれのモデルも精度が向上することが分かった．このことから，追加事前学習を行うことが関連度予測精度の向上に重要だということが理解できる．マスクした属性値を当てたり，属性値をもとにマスクされたテキスト内のトークンを当てたりする学習を追加事前学習では行っている．このような学習を行うことで，属性間の関係やテキストと属性の関係を学ぶことができたため，精度向上につながったと考えられる．

追加事前学習が商品属性とレビューの関連度予測に有用であることが分かった一方で，量的データの入力方法を工夫することについては，一貫した実験結果は得られなかった．関連度予測タスクでは，追加事前学習を行わない場合，量的データの入力方法を工夫することが正解率の向上につながった．一方で，レビューのランキングタスクにおいては量的データにカテゴリを表すトークンを用いる手法がベースラインの MRR を上回った．また，追加事前学習を行う場合，レビューのランキングタスクにおいては，量的データをカテゴリ化し，数値で表す手法がベースラインを上回った．このように，量的データの入力方法を工夫することはモデルの精度を向上させる可能性があるものの，どのような条件において精度が向上するかについてはより分析が必要であると考えられる．

最後に，本研究での限界点について述べる．本研究では，比較文章を学習の際に考慮することができなかった．レビューの中には「D500 よりもグリップが深くなって持ちやすくなっている.」や「M2 と比べるとフォーカスの精度がかなり改善されていて，ぶれなく撮影できる.」のようなレビューもある．商品名を含んでいるレビューでは，同じメーカーの前の型の商品と比較をすることが多い．ランキングタスクを行った際も，上位に商品名が含まれているレビューもあった．このため，商品名を根拠として関連度予測をしていることも考えられる．

このような商品の比較を行っている文章は，商品属性とレビューの関連を学習するうえで重要であると考える．レビューの対象となる商品と，レビュー中で比較されている商品の属性の差を知ることで，レビューで記述されている内容に影響を与える属性を理解することができる，この情報を学習に組み込むことにより，どの属性がレビューに関連あるかを理解でき，商品属性とレビューの関連だけではなく，ある属性とレビューの

関連を理解したモデルの構築につながる．

7 まとめ

本研究では．商品属性とレビューの関連を理解したモデルの構築に取り組んだ．そのために，商品属性とレビューを BERT に対して入力することで関連度を推定するモデルを提案した．さらに，追加事前学習と入力の商品属性に量的データのカテゴリ化を行い，どのような手法が関連度予測に適しているのかを評価した．関連度予測タスク，レビューランキングタスクどちらにおいても，追加事前学習を行うことで精度の向上が見られ，追加事前学習の重要性を示すことができた．

今後は，レビュー内の比較文章を考慮した学習により，商品属性とレビューの関連をより理解したモデルの構築に取り組む．

謝辞 本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します．

文献

[1] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Numeracy-600k: Learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6307–6313, 2019.

[2] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3040–3050, 2019.

[3] Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. Review-driven answer generation for product-related questions in e-commerce. *arXiv preprint arXiv:11905.01994*, 2019.

[4] Yang Deng, Wenxuan Zhang, and Wai Lam. Opinion-aware answer generation for review-driven question answering in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 255–264, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 623–632, 2017.

[7] Yue Feng, Zhaochun Ren, Weijie Zhao, Mingming Sun, and Ping Li. Multi-type textual reasoning for product-aware answer generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1135–1145, 2021.

[8] Gilad Fuchs, Haggai Roitman, and Matan Mandelbrod. Automatic form filling with form-BERT. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1850–1854, 2021.

[9] Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. Meaningful answer generation of e-commerce question-answering. *ACM Transactions on Information Systems*, Vol. 39, No. 2, pp. 1–26, 2021.

[10] Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 429–437, 2019.

[11] Tatsuya Iwanari, Naoki Yoshinaga, Nobuhiro Kaj, Toshiharu Nishina, Masashi Toyoda, and Masaru Kitsuregawa. Ordering concepts based on common attribute intensity. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, p. 3747–3753, 2016.

[12] Zhihua Jin, Xin Jiang, Xingbo Wang, Qun Liu, Yong Wang, Xiaozhe Ren, and Huamin Qu. NumGPT: Improving numeracy ability of generative pre-trained models. *arXiv preprint arXiv:2109.03137*, 2021.

[13] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. 2020.

[14] Bodhisattwa Prasad Majumder, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, and Ajinkya More. Deep recurrent neural networks for product attribute extraction in ecommerce. *arXiv preprint arXiv:1803.11284*, 2018.

[15] Julian McAuley and Alex Yang. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 625–635, 2016.

[16] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3374–3380, 2019.

[17] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*, 2019.

[18] Kalyani Roy, Smit Shah, Nithish Pai, Jaidam Ramtej, Prajit Nadkarni, Jyotirmoy Banerjee, Pawan Goyal, and Surender Kumar. Using large pretrained language models for answering user queries from product specifications. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pp. 35–39, 2020.

[19] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics: 18th China National Conference*, pp. 194–206, 2019.

[20] Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 743–758, 2019.

[21] Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 47–55, 2020.

[22] Shiwei Zhang, Jey Han Lau, Xiuzhen Zhang, Jeffrey Chan, and Cecile Paris. Discovering relevant reviews for answering product-related queries. In *2019 IEEE International Conference on Data Mining*, pp. 1468–1473, 2019.

[23] Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4889–4896, 2020.

[24] You Zhang, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. MA-BERT: Learning representation by incorporating multi-attribute knowledge in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*

2021, pp. 2338–2343, 2021.

[25] Chen Zhao, Yuki Nakayama, and Koji Murakami. Multi-task learning for product information with fine-tuned BERT. 言語処理学会 第 26 回年次大会 発表論文集, pp. 736–739, 2020.

[26] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1049–1058, 2018.

[27] 河田友香, 山本岳洋, 大島裕明, 柳田雄輝, 加藤誠, 藤田澄男. 商品属性とレビューの対応関係学習に基づく言語モデル構築. 第 15 回データ工学と情報マネジメントに関するフォーラム, 3a-7-3, 2023.

[28] 内田臣了, 山本岳洋, 加藤誠, 大島裕明, 田中克己. 不確かな順序関係の分析に基づく経験的属性によるオブジェクト検索. 第 8 回データ工学と情報マネジメントに関するフォーラム, C2–4, 2016.

DEIM Forum 2024

動画の属するコミュニティ情報を考慮した動画再生回数予測手法の提案

川上　大凱† 　鈴木　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †kawakami.taiga.b8@s.gifu-u.ac.jp, ††suzuki.yu.r4@f.gifu-u.ac.jp

あらまし　本研究では，動画のメタデータから再生回数を高い精度で予測する手法を提案する．動画はジャンルやチャンネルごとに存在し，同じジャンルやチャンネルであれば視聴者数の類似傾向が見られる．我々は動画の再生回数予測精度を向上させるために，ジャンルやチャンネルの特徴を考慮すべきではないかと考えた．ところが，それらのデータは明示的に与えられているわけではない．そこで本研究では，動画に付随するコメントにジャンルやチャンネルの特徴が現れ，ジャンルやチャンネルを判別可能な情報を保有するという仮説を立てた．仮説を検証することによって，コメントがジャンルやチャンネルの特徴を表すデータとして使えるのかどうかを調査した．仮説を検証するため，コメントを入力データとしたジャンル及びチャンネル分類を行った．実験の結果より，コメントがジャンルやチャンネルの情報を保有していることが分かった．また，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説を立てた．仮説を検証することによって，コメントが再生回数予測精度に影響を与えるかどうかに加え，コメントに含まれるジャンルやチャンネルを表す情報が再生回数予測精度に影響を与えるかどうかを調査した．仮説を検証するため，自作回帰ニューラルネットワークモデルによる 2 種類の再生回数予測実験を行った．実験の結果より，コメントに含まれるジャンルやチャンネルの情報が再生回数予測に影響を与えることが分かった．

キーワード　ソーシャルメディア,YouTube, メタデータ, 再生回数予測, ニューラルネットワーク, 機械学習

1　はじめに

我々は，高い精度で人気の動画を推測することによって，投稿者が動画を作成するための大きな助けとなると考えた．そこで本研究では，高い精度で人気の動画を推測するため，再生回数を高い精度で予測する手法を提案する．

高い再生回数予測精度を得るため，我々は動画の再生回数に差が表れる原因を考慮する必要があると考えた．我々は動画の再生回数に差が表れる原因として，動画の内容以外にもジャンルやチャンネルの特徴が存在していると考えた．動画の視聴者数はジャンルによって差が存在する．ゲーム実況のような人気ジャンルは，バイクや車などの乗り物関連のジャンルと比較して視聴者数が多い [1][2]．また，動画の視聴者数はチャンネルによっても差が存在する．同時期に類似した動画をアップロードしていたとしても，チャンネルごとに動画の再生回数は異なる．しかし，視聴者数の違いが表れる原因は明確化することが難しい．そのため，ジャンルやチャンネルの特徴を表すデータは直接収集することが現実的ではない．そこで我々は，ジャンルやチャンネルの特徴を表すため，動画に付随するコメントに着目した．

我々は二つの仮説を立てた．一つ目の仮説は，動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているということである．二つ目の仮説は，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるということである．

動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているという仮説を検証することによって，コメントがジャンルやチャンネルの特徴を表すデータとして使えるのかどうかを調査する．仮説を検証するため，実験 1 としてコメントを入力として BERT [1] を用いたジャンル分類とチャンネル分類を行った．ジャンル分類では，ジャンルを独自に 7 種類設定した．それらをジャンルラベルとし，コメントに付与した．チャンネル分類では，一つのジャンルに対して 10 個のチャンネルを用意した．それらをチャンネルラベルとし，コメントに付与した．ジャンル及びチャンネルのラベルを付与したものを BERT の入力データとした．

実験の結果，ジャンル分類では Accuracy が 0.87149 という結果が得られた．チャンネル分類では Accuracy が最大で 0.6785 という結果が得られた．また，ジャンルによって分類精度に大きな差 (0.1844) が存在することが確認できた．以上より，コメントがジャンルやチャンネルを判断可能な情報を保有していることが確認できた．

ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説を検証することによって，ジャンルやチャンネルの情報が再生回数予測精度に影響を与えるかどうかを調査する．また，コメントが再生回数予測精度に影響を与えるかどうかも調査する．ジャンルやチャ

1：出典：LINE リサーチ https://lineresearch-platform.blog.jp/archives/39951598.html
2：出典：株式会社Ｇｚブレイン eb-i Xpress https://www.kadokawa.co.jp/topics/2508/

ンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説を検証するため，実験 2 と実験 3 として自作の回帰ニューラルネットワークモデルを用いた 2 種類の再生回数予測実験を行った．ベースモデルとしてタイトル，チャンネル登録者数，動画高評価数を説明変数に持つモデルを作成した．

実験 2 では，コメントが再生回数予測精度に影響を与えるかどうかを調査した．ベースモデルに加えて，ベースモデルにコメントを加えたモデル，コメントの代わりに動画に付随する動画タグを加えたモデルを用意した．三つのモデルに対するテスト用データの予測精度を比較することによって，コメントが再生回数予測精度の向上に寄与するかどうかを調査した．10 分割交差検証を行い，有意水準を 10%として対応のある 2 標本 t 検定での片側検定を実施した．実験 3 で本実験での調査対象であるコメントを細分化した特徴を用いるため，我々は本実験での厳しい判定は必要ないと考え，有意水準を 10%と設定した．タイトルとコメント，動画タグは BERT を用いて 768 次元の分散表現を取得した．再生回数は動画単位で付与されるものであるため，コメントについては同動画内でのコメントの分散表現の平均とした．

実験の結果，コメントを説明変数に含むモデルの決定係数 (R^2) が 0.7470 であり，最も高い数値となった．次点で動画タグを説明変数に含むモデルが続き，決定係数 (R^2) は 0.7391 であった．ベースモデルの決定係数 (R^2) は 0.7264 で最も低い数値となった．t 検定でも各モデルの評価値間に有意な差が見られた．以上の結果より，コメントが再生回数予測精度に影響を与えていることが確認できた．

実験 3 では，コメントに含まれるジャンルやチャンネルを表す情報が再生回数予測精度に影響を与えるかどうかを調査した．ベースモデルに加えて，説明変数に前述したジャンル分類で得られた予測ラベルを加えたモデル，そしてチャンネル分類で得られたチャンネルの予測ラベルを加えたモデルの三つを用意した．それぞれのモデル対するテスト用データの予測精度を比較し，ジャンルやチャンネルが再生回数予測精度の向上に寄与するかどうかを調査した．10 分割交差検証を行い，有意水準を 5%として対応のある 2 標本 t 検定での片側検定を実施した．動画単位の学習を行わせるために，同動画内のコメントに対して付与された予測ラベルの中で最も付与された回数の多い予測ラベルを動画の予測ラベルとした．

実験の結果，予測ラベルを含むモデルの決定係数 (R^2) はベースモデルの決定係数 (R^2) と比較して最大で 0.0526 ほど高いという結果が得られた．また t 検定でも評価値間に有意な差が見られた．しかし実験 1 で得られた予測精度が低いジャンルにおいては，予測ラベルを含むモデルとベースモデルの評価値間に有意な差が見られないという結果が得られた．以上の結果より，精度向上にコメントが保有するジャンルやチャンネルの情報が影響を与えていることが確認できた．

本論文における貢献は以下の通りである．

- コメントにはジャンルとチャンネルの特徴が保有されているということを確認した．
- コメントを用いることによって再生回数予測精度が向上するということを確認した．
- ジャンルやチャンネルの情報が再生回数予測精度の向上に影響を与えていることを確認した．

2 関 連 研 究

動画の再生回数予測に関連する研究は現在までに複数行われている．Abisheva ら [2] は動画作成から Twitter による動画の拡散までの速度に注目し，拡散速度と最終的な動画人気との関係を調査をした．Yu ら [3] は Twitter 由来の特徴量を用いることによって，動画再生回数の急激な増加とアップロード直後の再生回数の増加を予測する研究を行った．これらはどちらも動画に対して行われた Twitter での拡散に着目し，それを要因とした動画再生回数との関係性を調査するものであった．本稿では動画投稿サイト内で取得できる特徴のみを使用することによって，投稿者と視聴者以外が介入する可能性を排除する．投稿者と視聴者以外が介入する可能性として，本研究では動画投稿サイト内で動画を視聴せずに Twitter のタイムライン上に流れてきたものをそのまま拡散するといった場合を挙げる．その可能性を排除することで，我々は再生回数予測に対するノイズとなり得るデータが混入するリスクを削減することができると考えた．

Zhou ら [4] は，関連動画推薦や検索アルゴリズムなどの視聴ユーザの動画への到達方法と，動画に対する視聴パターンの関連性について研究を行った．これは Twitter だけでなく，動画投稿サイト内での拡散方法についても着目していた．YouTube における関連動画推薦機能は動画検索アルゴリズムや Twitter での拡散と異なり，再生回数が低いマイナーな動画に対しても行われていることを確認した．動画の持続的な視聴には関連動画推薦と検索アルゴリズムによる誘導が有効であることを確認した．柴田ら [5] は，ニコニコ動画における動画再生回数の推定を行った．音楽関連の動画は線形回帰分析の結果，高い推定精度を確認した．また，音楽関連の動画は再生回数とマイリスト登録数の相関が他のジャンルよりも高いことを確認した．動画についた動画タグを使用したクラスタリングと MLP による回帰分析を組み合わせた結果，推定精度が向上した．本稿では関連動画推薦や検索アルゴリズムでは考慮しきれない動画の特徴を考慮する方法として，動画タグではなくコメントを用いる．また関連動画推薦や検索アルゴリズムと再生回数の関係性を考慮するため，推薦や検索アルゴリズムに用いられる動画タイトルを特徴量として用いる．また，マイリスト登録数の代わりに動画高評価数とチャンネル登録者数を使用する．

視聴者の反応と動画に対する印象の関連性を調査する研究も現在までに複数行われている．山本ら [6] は楽曲動画の印象推定を視聴者のリアルタイムコメントを用いて行った．サビ区間で複数出現する形容詞や形容動詞を繰り返すコメントに着目し，楽曲の印象を推定した．歌詞や音響特徴量を用いた印象推定と

比較し，視聴者のリアルタイムコメントが印象推定に有意な特徴であることを確認した．土屋ら [7] は形容詞や形容動詞以外の品詞にも注目することによって，より詳細な楽曲動画の印象推定手法を提案した．メディアや動画の印象タイプごとに適切な品詞を用いることによって，より詳細な印象推定が可能であるということが確認された．また品詞を考慮せずコメント全文を用いることによって，メディアや印象タイプに関係なくある程度の印象推定が可能であることが確認された．村上ら [8] はリアルタイムコメントに基づく動画検索ランキングの手法を検討した．使用データセットはニコニコ動画で収集可能なメタデータセットであった．再生回数とマイリスト登録数，コメント数に相関があることを確認した．また，コメント付与者の面白さ尺度とコメント付与人数を考慮したランキング手法を検討した．面白さ尺度はあるシーンに対してのコメント数であった．堺ら [9] は YouTube 動画におけるメタデータを使用した炎上動画分類を行った．動画に対する低評価数とコメント数が多いものを炎上動画と仮定した．また，コメントに対して感情分析を行った結果を感情スコアとした．動画のベクトル化には感情スコアと再生回数，コメント数，高評価数を使用した．LightGBM を用いた分類器に全ての動画ベクトルを入力した際，最も高い分類精度を確認した．Siersdorfer ら [10] は YouTube 上のコメントに対する様々な分析を行った．コメントに対する感情分析を行い，その結果を用いることによってコメントがコミュニティに受け入れられるものかどうかを自動的に判断できることを確認した．またコメントとユーザによって反応が変わる動画の間には相関関係を確認した．実験結果より，コメントで表現された感情とコミュニティによるコメント評価，そして動画のトピックに強い依存関係を確認した．本稿ではコメントを品詞で制限を加えることなく全文で使用する．そのため我々は，動画に対する視聴者の印象に考慮できないものが存在する可能性は低いと考える．

3 提案手法

本稿では二つの仮説を立てた．一つ目の仮説は，動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているということである．二つ目の仮説は，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるということである．

動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているという仮説の検証のため，動画の属するジャンルやチャンネルのラベルを付与したコメントを入力データとしたジャンル分類とチャンネル分類を行う．ジャンルの正解ラベルを l_{genre}，チャンネルの正解ラベルを $l_{channel}$ とする．また分類によって得られたジャンルの予測ラベルを p_{genre}，チャンネルの予測ラベルを $p_{channel}$ とする．

また，コメントが再生回数予測に影響を与えるのかを検証するため，コメントから生成した分散表現を説明変数に加えたモデルを用いて再生回数予測実験を行う．コメントに着目した理由については 3.1 節で詳しく述べる．

また，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説を検証するため，コメントに含まれるジャンルやチャンネルの情報を特徴量とした再生回数予測実験を行う．コメントに含まれるジャンルやチャンネルの情報を特徴量とするため，p_{genre} と $p_{channel}$ を特徴量として用いる．

3.1 コメントデータ

我々は動画の再生回数予測精度を向上させるために，ジャンルやチャンネルの特徴を考慮すべきではないかと考えた．しかし，ジャンルやチャンネルの特徴をデータとして収集するためには二つの問題が存在する．一つ目はジャンルやチャンネルの特徴を表すデータが存在しないことである．そして二つ目は収集可能なメタデータには動画の内容に則していないものが存在することである．我々はこれらの問題を，コメントを用いることによって解消できると考えた．

一つ目の問題は，ジャンルやチャンネルの特徴を表すデータが存在しないことである．1 章で述べた通り，ジャンルやチャンネルによって動画の再生回数に差が生まれることは明確である．そこで我々は動画の再生回数に差が表れる原因として，動画の内容以外にもジャンルやチャンネルの特徴が存在していると考えた．しかし，動画のデータを収集する際，ジャンルやチャンネルの特徴を直接表すデータは存在しない．そこで我々は，ジャンルやチャンネルの特徴を別のデータから推定することによって，その問題を解消できると考えた．

二つ目の問題は，収集可能なメタデータには動画の内容に則していないものが存在することである．メタデータとはコンテンツに付随する周辺データのことであり，動画のメタデータにはタイトルや動画タグ，概要欄，サムネイル画像，字幕の有無など多くのデータが存在している．しかし投稿者が自由に設定可能であるメタデータには，動画の内容を表していないものが存在することがある．例えば動画タグでは，他のコンテンツの視聴者を誘導するために，関係のないコンテンツの動画タグを設定している投稿者が存在する．また設定ミスにより，動画タグを一つも設定していない投稿者や他のカテゴリのタグを設定したままの投稿者も存在する．そういった形でメタデータが設定されている場合，そのメタデータは動画の属するジャンルの違いやチャンネルの特徴を表現しているとは言えない．そこで我々は，動画の属するジャンルやチャンネルの特徴を適切に表現するために，データが客観性を持つ必要があると考えた．

以上の問題点を解消するため，我々は不特定多数の視聴者が動画を見た上で感じたことや思ったことを書き残すものであるコメントに着目した．我々は，動画を見た上で書き残されるコメントには動画のジャンルの違いやチャンネルの特徴に言及しているものが存在していると考えた．また我々は，不特定多数の視聴者によって残されるコメントには客観性が担保されていると考えた．

3.2 使用データ

使用するデータは YouTubeDataAPI[3]の Search 機能を用いて収集した．Search 機能はリクエストしたキーワードと関連する動画やチャンネルの情報を返す．本稿では，2023 年 1 月 1 日から同年 6 月 30 日までの動画を対象としてデータを収集した．ゲーム実況，キャンプ，アニメコント，ニュース，歌ってみた，踊ってみた，ネタ動画の 7 ジャンルを設定してデータの収集を行った．データ収集時に用いたキーワードは，それぞれ”ゲーム実況”，”キャンプ”，”アニメコント”，”ニュース”，”歌ってみた”，”踊ってみた”，”ネタ”である．収集したコメントには，そのコメントが属する動画の収集キーワードに対応する l_{genre} を付与した．そのため p_{genre} と l_{genre} は 0 から 6 の値を取る．

チャンネル分類における使用チャンネルは登録者数 10 万人以上であり，尚且つ収集データ内出現頻度がジャンルごとの上位 10 位以内であるチャンネルのみとした．出現頻度を考慮した理由は，登録者数のみでデータを収集した場合に該当ジャンルと関係ない動画が混入してしまうことを可能な限り防ぐためである．チャンネル分類に使用するデータのジャンルは，代表としてゲーム実況とアニメコントの二つを選定した．チャンネル分類を行うためのコメントを十分量確保するため，選定した 10 チャンネルの動画全てに対して期間を設けずコメントを収集した．収集したコメントには，該当する $l_{channel}$ を付与した．そのため $p_{channel}$ と $l_{channel}$ は 0 から 9 の値を取る．またチャンネル登録者数を s_n，動画高評価数を f_n とする．s_n と f_n は値を取りうる範囲が広く，重みの更新が上手くいかないことがある．そのため，s_n と f_n は常用対数を取ることによって特徴量のスケーリングを行う．

3.3 使用モデル

本研究では BERT と自作の回帰ニューラルネットワークモデルを用いて実験を行う．扱えるトークンの最大長を 512 とし，満たないものに関しては 0 での Padding 処理を実施する．バッチサイズを 32 とし，ミニバッチ学習を実施する．検証用データの loss が 50 回連続で下がらなかった場合，earlystopping を実施する．各精度評価値は小数点第 5 位で四捨五入を行う．実験に使用するデータにはアンダーサンプリングを実施する．

3.3.1 BERT

BERT は Transformer の Encoder を用いた自然言語処理特化の深層学習モデルである．このモデルはその名の通り双方向でのエンコード表現を実装しており，MLM(Masked Language Modeling) と NSP(Next Sentence Prediction) という二つの事前学習タスクを行うことによって，文脈を考慮した分散表現が可能となっている．また，BERT はラベルなしデータを用いた事前学習を行っている．そのため，様々なタスクに対してファインチューニングを行うことが可能であるという特徴を持つ．よって本稿では，ジャンル及びチャンネル分類と文字列データの分散表現生成に BERT [1] を用いる．BERT を用いたのは複数文が存在する長文のコメントに対して精度の高い分類や分散表現の取得を行うためである．BERT の事前学習モデルは，東北大学自然言語処理研究グループの日本語学習モデル [4]を使用した．文字列の分散表現を取得するために，本稿では BERT モデルの最終隠れ層の出力を用いる．入力文をトークン化し，トークンごとに生成された 768 次元の単語分散表現の平均を取ったものを文字列の分散表現とした．本稿ではタイトルと動画タグ，コメントの分散表現を取得する．

あるタイトル s_title が与えられたとき，s_title はトークン化によって N_{title_token} 個のトークン t_i，(i=1,2,...,N_{title_token}) に分割される．各トークンに対応する BERT の最終隠れ層の出力（単語分散表現）を v_i とする．また，ある動画タグ s_tags が与えられたとき，s_tags はトークン化によって N_{tags_token} 個のトークン t_j，(j=1,2,...,N_{tags_token}) に分割される．各トークンに対応する BERT の最終隠れ層の出力（単語分散表現）を v_j とする．タイトルの分散表現 v_{s_title} と動画タグの分散表現 v_{s_tags} は以下の式で定義される．

$$v_{s_title} = \frac{1}{N_{title_token}} \sum_{i=1}^{N_{title_token}} v_i$$

$$v_{s_tags} = \frac{1}{N_{tags_token}} \sum_{j=1}^{N_{tags_token}} v_j$$

しかし再生回数は動画単位で付与されるものであるため，コメントに関しては動画に付随した全てのコメントにおける分散表現の平均を取った．コメントを c，トークン化されたコメントの a 番目のトークンの単語分散表現を v_a，トークンの総数を N_c とする b 番目のコメントにおける分散表現を v_b，コメント総数を n_c とする．コメントの分散表現 $v_{c_pervideo}$ は以下の式で定義される．

$$v_{c_pervideo} = \frac{1}{n_c \cdot N_c} \sum_{b=1}^{n_c} \sum_{a=1}^{N_c} v_a$$

3.3.2 回帰ニューラルネットワークモデル

本稿では，動画の再生回数予測実験を行うために基本的な構造のニューラルネットワークモデルを構築する．構築するモデルについての概要は図 1 と図 2，図 3 で示す．モデルは入力層，隠れ層 1 層，出力層の 3 層構造とした．複雑な特徴量を用いるためニューラルネットワークを使用したが，再生回数予測における特徴量の違いを確認するためのモデルであるため，基本的な構造のモデルとした．活性化関数には ReLU 関数を用いた．隠れ層は 128 次元とした．

再生回数は動画単位で付与されるものであるため，動画のジャンル予測ラベル p_{v_genre} は同動画内のコメント c に対して実験 1 で付与された p_{genre} の内，付与回数が最も多い予測ラベルとした．p_{genre} が付与された c を c_p_{genre}，$p_{channel}$ が付与された c を $c_p_{channel}$ とする．p_{genre} が同動画内のコメントに付与された回数を $C(p_{genre})$ とすると，動画のジャンル予測

3：https://developers.google.com/youtube/v3/docs?hl=ja

4：https://github.com/cl-tohoku/bert-japanese

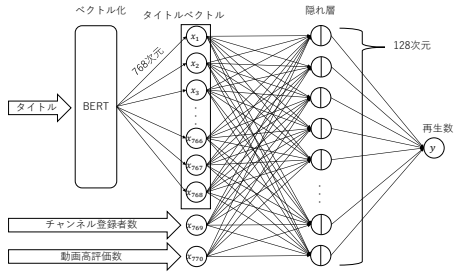


図 1　ベースモデル

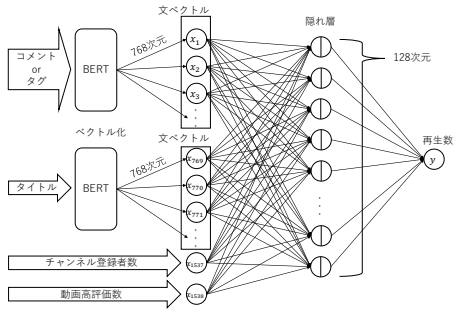


図 2　モデル 1(コメント有・動画タグ無)・モデル 2(コメント無・動画タグ有)

ラベル p_{v_genre} は以下の式で定義される．

$$p_{v_genre} = \max(C(p_{genre}))$$

同様に動画のチャンネル予測ラベル $p_{v_channel}$ は以下の式で定義される．

$$p_{v_channel} = \max(C(p_{channel}))$$

max() は最大値を求める関数である．

再生回数予測精度の比較を行うため，5 種類のモデルを構築する．モデルは全て動画再生回数 V を出力とする．それぞれのモデルの説明変数を I とおくと，ベースモデルは，$I = [v_{s_title1}, v_{s_title2},..., v_{s_title768}, s_n, f_n]$ とした．モデル 1 は，$I = [v_{s_title1}, v_{s_title2},..., v_{s_title768}, s_n, f_n, v_{c_pervideo1}, v_{c_pervideo2},..., v_{c_pervideo768}]$ とした．モデル 2 は，$I = [v_{s_title1}, v_{s_title2},..., v_{s_title768}, s_n, f_n, v_{s_tags1}, v_{s_tags2},..., v_{s_tags768}]$ とした．モデル 3 は，$I = v_{s_title1}, v_{s_title2},..., v_{s_title768}, s_n, f_n, p_{v_genre}]$ とした．モデル 4 は，$I = [v_{s_title1}, v_{s_title2},..., v_{s_title768}, s_n, f_n, p_{v_channel}]$ とした．

4　実　　験

本稿での実験目的は，二つの仮説を検証することである．一つ目の仮説は，動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているということである．二つ目の仮説は，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるということである．一つ目の仮説に対する実験を実験 1 とする．二つ目の仮説に対する実験を実験 2 と実験 3 とする．

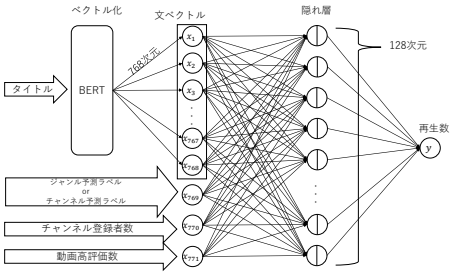


図 3　モデル 3(ジャンルの予測ラベル)・モデル 4(チャンネルの予測ラベル)

4.1　実　験　1

本実験では，動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているという仮説について検証する．コメントを入力とした BERT 分類タスクによって，コメントがジャンルやチャンネルを判別可能な情報を保有しているかどうかを調査した．

4.2　実 験 手 順

BERT を用いたジャンル・チャンネル分類の実験手順を示す．

1. YouTubeAPI を用いてコメントを収集する．
2. 各コメントに対して正解のラベルを付与する．
3. BERT にコメントを入力し，ラベルを予測する．
4. テスト用データでの精度を確認する．

ジャンル分類の場合，2. と 3. はジャンルのラベルに対して行う．チャンネル分類の場合，2. と 3. はチャンネルのラベルに対して行う．データ数の詳細は表 1 で示す．データセットの割合は訓練用：検証用：テスト用の順に 8:1:1 とした．精度評価にはテスト用データにおける Accuracy と Cross Entropy Loss を用いた．RMSELoss とは回帰タスクで使われることが多い評価関数であり，以下の式で定義される．

$$RMSELoss = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

n はデータ数 (バッチ数)，y は正解値，$\hat{y}$ は予測値である．

4.2.1　結果・考察

テスト用データに対する Accuracy と Cross Entropy Loss を表 6 に示す．なお，Cross Entropy Loss はバッチごとの Loss

表 1　実験 1：使用データ数

チャンネル分類 (ラベル数 10)		ジャンル分類 (ラベル数 7)
ゲーム実況	アニメコント	
300,000	100,000	87,500

表 2　実験 1：ジャンル分類の混同行列

	pre 0	pre 1	pre 2	pre 3	pre 4	pre 5	pre 6
label 0(ゲーム実況)	548	40	35	6	57	12	2
label 1(キャンプ)	12	597	45	8	12	16	10
label 2(アニメコント)	12	41	555	35	13	25	19
label 3(ニュース)	4	25	51	596	5	9	10
label 4(歌ってみた)	35	30	28	7	581	13	6
label 5(踊ってみた)	20	31	46	8	33	550	12
label 6(ネタ動画)	6	31	61	6	16	14	566

表 3　実験 1：ジャンル分類の F 値

	ゲーム実況	キャンプ	アニメコント	ニュース	歌ってみた	踊ってみた	ネタ動画
F 値	0.8197	0.7987	0.7298	0.8726	0.8200	0.8215	0.8543

表 4　実験 1：チャンネル分類 (ゲーム実況) の混同行列

	pre 0	pre 1	pre 2	pre 3	pre 4	pre 5	pre 6	pre 7	pre 8	pre 9
label 0	2,189	40	80	138	175	40	73	119	82	64
label 1	22	2,310	122	98	52	193	24	49	81	49
label 2	46	131	1,889	128	56	468	41	48	131	62
label 3	94	94	127	2,043	108	129	36	56	198	115
label 4	170	60	75	150	2,163	44	30	188	55	65
label 5	22	143	333	85	37	2,206	24	23	92	35
label 6	70	51	131	81	36	129	2,344	72	62	24
label 7	173	77	89	152	237	73	60	1,993	82	64
label 8	81	122	194	220	87	196	32	54	1,735	279
label 9	129	141	129	290	199	116	29	84	400	1,483

表 5　実験 1：チャンネル分類 (アニメコント) の混同行列

	pre 0	pre 1	pre 2	pre 3	pre 4	pre 5	pre 6	pre 7	pre 8	pre 9
label 0	554	18	39	32	59	51	85	35	57	70
label 1	88	390	38	39	103	45	172	26	31	68
label 2	66	17	384	24	109	63	89	68	78	102
label 3	81	18	65	434	66	46	57	52	54	127
label 4	61	40	46	26	556	110	70	24	22	45
label 5	69	34	38	22	110	515	68	41	50	53
label 6	23	30	18	22	42	55	717	7	34	52
label 7	54	21	101	39	96	39	88	382	51	129
label 8	41	14	42	32	42	37	99	34	544	115
label 9	84	15	51	33	65	54	85	57	91	465

の平均を取ったものとした．

ジャンル分類における Accuracy はチャンスレート (0.143) を大きく超えている．加えて Cross Entropy Loss は 1 を切っている．また表 2 を確認すると，どのジャンルにおいても分類精度が高いことが分かる．よってコメントはジャンルを表現する情報を保有していると言える．また，ジャンルによって分類精度が変化するのかを確認するため，ジャンルごとの F 値を用いて精度差を確認する．ジャンル分類の F 値を表 3 に示す．表 3 を確認すると，ほとんどのジャンルで F 値が 0.8 付近を取った．しかしアニメコントは他のジャンルと比較して大きく F 値が下がる．

表 6　実験 1：テスト用データに対する Accuracy と Cross Entropy Loss

	チャンネル分類		ジャンル分類
	ゲーム実況	アニメコント	
Accuracy	0.6785	0.4941	0.8149
Cross Entropy Loss	0.9422	1.5215	0.5468

チャンネル分類における Accuracy は，両者ともチャンスレート (0.1) を大きく超えている．また，表 4 と表 5 を確認すると，両者とも分類がしっかり行われていることが分かる．よってコメントはチャンネルを表現する情報を保有していると言える．しかし，ゲーム実況とアニメコントの間では Accuracy に大きな差 (0.1844) が存在する．ゲーム実況は Accuracy が 0.65 を超えているが，アニメコントの Accuracy は 0.5 を切っている．Cross Entropy Loss も Accuracy と同様に，ゲーム実況とアニメコントで大きな差 (0.5793) が生まれる結果となっている．混同行列でもアニメコントはゲーム実況と比較して，間違った分

類が行われていることが多い．以上の結果より，我々はアニメコントがチャンネルごとの特徴を表現しにくいジャンルなのではないかと推測する．

アニメコントとゲーム実況の間に精度差はあったものの，動画に付随するコメントからジャンルやチャンネルを判別可能であることが確認できた．よって，動画に付随するコメントがジャンルやチャンネルを判別することができるだけの情報を保有しているということが確認できた．

4.3 実　験　2

本実験では，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説について，一つ目の検証を行う．3種類のモデルを構築し，テスト用データの予測精度を比較することによって，コメントが再生回数予測精度の向上に影響を与えるかどうかを調査した．

4.3.1 実 験 手 順

回帰ニューラルネットワークモデルを用いた再生回数予測実験の手順を示す．

1. チャンネル登録者数・動画高評価数をスケーリングする．
2. BERT を用いてタイトルとコメント，動画タグからそれぞれ 768 次元の分散表現を取得する．
3. 1. と 2. で用意した特徴量を説明変数としたベースモデル，モデル 1，モデル 2 を構築する．
4. それぞれのモデルのテスト用データに対する精度を比較する．

ベースモデルは，入力にタイトル・チャンネル登録者数・動画高評価数を用いた．モデル 1 は，入力にタイトル・チャンネル登録者数・動画高評価数・コメントの分散表現を用いた．モデル 2 は，入力にタイトル・チャンネル登録者数・動画高評価数・動画タグの分散表現を用いた．また動画単位の学習を行わせるため，コメントは動画に付随した全てのコメントの分散表現の平均を取った．本実験は計 6,133 個の動画 (629,516 コメント) に対して行った．またデータセットは再生回数の偏りが無いよう，順番のシャッフルを実施した．データセットの割合は訓練用：検証用：テスト用の順に 8:1:1 とした．精度評価にはテスト用データにおける決定係数 (R^2) と RMSE Loss を用いた．特徴量の有意性を検証するため，10 分割交差検証を実施した．交差検証で得られた評価値は正規分布に従うものとし，有意水準を 10%として対応のある 2 標本 t 検定を実施した．実験 3 で本実験での調査対象であるコメントを細分化した特徴を用いるため，我々は本実験での厳しい判定は必要ないと考え，有意水準を 10%と設定した．帰無仮説を「2 群の評価値の母平均に差がない」とし，片側検定を行った．精度評価にはテスト用データにおける決定係数 (R^2) と RMSELoss(Root Mean Squared Error Loss) を用いる．RMSELoss とは回帰タスクで用いられることが多い評価関数であり，以下の式で定義される．

$$RMSELoss = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

n はデータ数 (バッチ数)，y は正解値，$\hat{y}$ は予測値である．

4.3.2 結果・考察

実験結果を表 8 に示す．また，t 検定についての結果を表 7 に示す．表 8 では，モデル 1 が最も高い精度であった．次点でモデル 2 が高い精度であった．ベースモデルは最も精度が低いという結果となった．

ベースモデルの精度が低い理由を，我々は説明変数の情報量が少ないためであると考える．表 7 を確認すると，ベースモデルと他の二つのモデル間での p 値は全て 0.1 以下となっていることが分かる．よって帰無仮説は棄却され，ベースモデルと他の二つのモデルの評価値における母平均には有意な差があると言える．ベースモデルと他の二つのモデルの違いに着目すると，これらのモデルの違いは説明変数の情報量であることが分かる．ベースモデルは v_{s_title} と s_n，f_n の 3 つを説明変数としている．タイトルは 3.3.1 節で述べた通り 768 次元の分散表現となっているため，ベースモデルにおける説明変数の次元は 770 次元となっている．一方で，他の二つのモデルは $v_{c_pervideo}$，または v_{s_tags} が追加されていることにより 1538 次元の説明変数を持つ．以上より，我々は特徴量の次元数が増えることで情報量が増加し，予測精度が向上すると考える．

次にモデル 1 とモデル 2 を比較する．表 7 を確認すると，二つのモデル間での p 値は全て 0.1 以下となっていることが分かる．よって帰無仮説は棄却され，モデル 1 とモデル 2 における評価値の母平均には有意な差があると言える．以上の結果より，コメントが再生回数予測に影響を与えていることが確認できた．

4.4 実　験　3

本実験では，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説について，二つ目の検証を行う．3 種類のモデルを構築し，テスト用データの予測精度を比較することによって，コメントに含まれるジャンルやチャンネルを表す情報が再生回数予測精度に

表 7　実験 2：p 値

	ベースモデル		モデル 1 とモデル 2
	モデル 1	モデル 2	
決定係数	7.69978e-05	0.0078	0.0592
RMSE Loss	9.0836e-05	0.0076	0.0548

表 8　実験 2：テスト用データに対する精度 (10 回平均)

	ベースモデル	モデル 1	モデル 2
決定係数 (R^2)	0.7264	0.7470	0.7391
RMSE Loss	0.6927	0.6660	0.6769

表 9　実験 3：p 値

	モデル 3	モデル 4	
		ゲーム実況	アニメコント
決定係数 (R^2)	1.4671e-05	2.0492e-06	0.3577
RMSE Loss	1.7274e-05	1.1992e-06	0.3349

表 10　実験 3：テスト用データに対する精度 (10 回平均)

	ジャンル		チャンネル			
			ゲーム実況		アニメコント	
	モデル 3	ベースモデル	モデル 4	ベースモデル	モデル 4	ベースモデル
決定係数 (R^2)	0.2714	0.2188	0.4307	0.3815	0.3380	0.3374
RMSE Loss	0.6432	0.6661	0.6525	0.6801	0.6113	0.6116

影響を与えるかどうかを調査した．

4.4.1　実 験 手 順

回帰ニューラルネットワークモデルを用いた再生回数予測実験の手順を示す．

1. 4.1 節で得られたジャンルの予測ラベル・4.1 節で得られたチャンネルの予測ラベルを用意する．
2. チャンネル登録者数・動画高評価数をスケーリングする．
3. BERT を用いてタイトルとコメントからそれぞれ 768 次元の分散表現を取得する．
4. 1. と 2. と 3. で用意した特徴量を説明変数としたモデル 3 とモデル 4，そしてそれぞれにデータを合わせたベースモデルを構築する．
5. モデル 3 とモデル 4 のテスト用データに対する精度を，各モデルにデータを合わせたベースモデルのテスト用データに対する精度と比較する．

ベースモデルは，入力にタイトル・チャンネル登録者数・動画高評価数を用いた．モデル 3 は，入力にタイトル・チャンネル登録者数・動画高評価数・ジャンルの予測ラベルを用いた．モデル 4 は，入力にタイトル・チャンネル登録者数・動画高評価数・チャンネルの予測ラベルを用いた．ジャンルの予測ラベルとチャンネルの予測ラベルは，実験 1 におけるテスト用データに対して付与されたものを用いた．動画単位の学習を行わせるため，それぞれのラベルは同動画内のコメントに対して付与された予測ラベルの中で最も付与された回数の多い予測ラベルとした．データセットに用いた動画数は，ゲーム実況が計 11,679 個の動画 (30,000 コメント)，アニメコントが計 6,133 個の動画 (10,000 コメント)，ジャンル分類が計 3,897 個の動画 (8,750 コメント) であった．データセットは再生回数の偏りが無いよう，順番のシャッフルを実施した．データセットの割合は訓練用：検証用：テスト用の順に 8:1:1 とした．精度評価にはテスト用データにおける決定係数 (R^2) と RMSE Loss を用いた．特徴量の有意性を検証するため，10 分割交差検証を実施した．交差検証で得られた評価値は正規分布に従うものとし，有意水準を 5%として対応のある 2 標本 t 検定を実施した．帰無仮説を「2 群の評価値の母平均に差がない」とし，片側検定を行った．

4.4.2　結果・考察

実験結果を表 10 に示す．また，t 検定についての結果を表 9 に示す．表 10 では，全てのモデルがベースモデルよりも高い精度であった．また表 9 では，アニメコントを除いた二つのモデルに対して p 値は全て 0.05 以下となっている．よって帰無仮説は棄却され，モデル 3 とモデル 4(ゲーム実況) はベースモデルと比較して評価値における母平均に有意な差があると言える．しかし，モデル 4(アニメコント) に関しては帰無仮説が棄却されず，ベースモデルと比較して評価値における母平均に有意な差があるとは言い切れない．アニメコントは 4.2.1 節で述べた通り，分類精度がゲーム実況のチャンネル分類と比べて低い．そこで我々は，チャンネルの特徴が上手く考慮できていない場合，そのデータを特徴量に追加しても再生回数予測精度は上がらないと考えた．以上の結果より，再生回数予測精度の向上にコメントが保有するジャンルやチャンネルの情報が影響を与えていることが確認できた．

5　おわりに

本研究では，動画のメタデータから再生回数を高い精度で予測する手法を提案した．我々は動画の再生回数予測精度を向上させるために，ジャンルやチャンネルの特徴を考慮すべきではないかと考えた．

しかし，ジャンルやチャンネルの特徴をデータとして収集するためには二つの問題が存在する．一つ目はジャンルやチャンネルの特徴を表すデータが存在しないことである．二つ目は収集可能なメタデータには動画の内容に則していないものが存在することである．

一つ目の問題を解消するため，我々はジャンルやチャンネルの特徴を別のデータから推定することを考えた．また二つ目の問題を解消するため，我々は動画の属するジャンルやチャンネルの特徴を適切に表現するためにデータが客観性を持つ必要があると考えた．そこで我々は，不特定多数の視聴者が動画を見た上で感じたことや思ったことを書き残すものであるコメントに着目した．

本稿では，二つの仮説を立てた．一つ目の仮説は，動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているということである．二つ目の仮説は，ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるということである．

動画に付随するコメントがジャンルやチャンネルを判別可能な情報を保有しているという仮説を検証するため，実験 1 としてコメントを入力として BERT を用いたジャンル分類とチャンネル分類を行った．実験の結果，ジャンル分類では Accuracy が 0.87149 という結果が得られた．チャンネル分類では Accuracy が最大で 0.6785 という結果が得られた．また，ジャンルによって分類精度に大きな差 (0.1844) が存在することが確認できた．以上より，コメントがジャンルやチャンネルを判断可能な情報

を保有していることが確認できた..

ジャンルやチャンネルの情報を特徴量にすることによって再生回数予測精度の向上が期待できるという仮説を検証するため，実験 2 と実験 3 として自作の回帰ニューラルネットワークモデルを用いた 2 種類の再生回数予測実験を行った．

実験 2 では，再生回数予測においてコメントが与える影響を調査した．実験の結果，コメントを説明変数に含むモデルの決定係数 (R^2) が 0.7470 であり，最も高い数値となった．次点で動画タグを説明変数に含むモデルが続き，決定係数 (R^2) は 0.7391 であった．ベースモデルの決定係数 (R^2) は 0.7264 で最も低い数値となった．t 検定でも各モデルの評価値間に有意な差が見られた．以上の結果より，コメントが再生回数予測精度に影響を与えていることが確認できた．

実験 3 では，再生回数予測においてコメント内のジャンルやチャンネルの情報が与える影響を調査した．実験の結果，予測ラベルを含むモデルの決定係数 (R^2) はベースモデルの決定係数 (R^2) と比較して最大で 0.0526 ほど高いという結果が得られた．また t 検定でも評価値間に有意な差が見られた．しかし実験 1 で得られた予測精度が低いジャンルにおいては，予測ラベルを含むモデルとベースモデルの評価値間に有意な差が見られないという結果が得られた．以上の結果より，精度向上にコメントが保有するジャンルやチャンネルの情報が影響を与えていることが確認できた．

本稿で行った実験にはいくつかの懸念事項が存在している．一つ目は，手法の汎用性や有効性の有無を他の条件で確認する必要があることである．使用ジャンルや再生回数予測実験に用いる特徴量，そしてデータを収集する期間は我々が独自に設定している．そのため，偏った結果が出力されている可能性が存在する．我々は，ジャンルやチャンネルの追加や他のモデルやコンテンツに対して適応可能か調査する必要があると考える．その一例として，我々はジャンルの設定方法に YouTube 側の機能として用意されている動画カテゴリ設定を用いることによって，その問題を解消できると考える．しかし YouTube 側で用意されている動画カテゴリ設定は，動画タグのように投稿者の設定によっては動画に全く関係のないものが設定されている場合がある．そのためコメントを用いたジャンル分類だけではなく，コメント内のどの単語がジャンル予測に大きな影響を与えているのかを調査し，動画カテゴリを適するものに修正する手法を考える必要がある．また，データの収集期間を延ばすことによって手法の更なる汎化が可能になると考える．

二つ目に，少数ではあるが本実験で収集したデータ内にも設定したジャンル以外の動画が含まれてしまっていることである．本稿では，データの収集方法として YouTubeDataAPI の Search 機能を使用している．Search 機能ではリクエストしたキーワードと関連する動画やチャンネルの情報を収集する．しかし 3.1 節で述べたように，動画タグには関係のないコンテンツの動画タグが設定されている場合や，動画タグが一つも設定されていない場合が存在する．これにより動画が別のジャンルのキーワードと関連づけられてしまうことがある．そのためデータ収集の際，YouTubeDataAPI の Search 機能以外の方法を考える必要がある．

三つ目に，アップロードされる前の動画にはコメントが付与されないことである．本稿では，コメントを用いることによって再生回数予測の精度が向上するという結果を確認した．しかし，アップロードされる前の動画にはコメントを付与することが不可能である．そのため我々は，投稿前から存在しているメタデータを用いて将来的に付与されるであろうコメントを生成する手法について考案する必要があると考える．

謝辞　本研究の一部は，JSPS 科学研究費 (19H04218, 23H03694)，越山科学技術振興財団，およびソニー株式会社の助成を受けたものです．

文　献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In *Proceedings of the 7th ACM international conference on Web search and data mining*, WSDM 2014. ACM, February 2014.

[3] Honglin Yu, Lexing Xie, and Scott Sanner. Twitter-driven youtube views: Beyond individual influencers. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 869–872, 2014.

[4] Renjie Zhou, Samamon Khemmarat, Lixin Gao, Jian Wan, and Jilin Zhang. How youtube videos are discovered and its impact on video views. *Multimedia Tools and Applications*, Vol. 75, , 05 2016.

[5] 柴田知親, 伊東栄典. 回帰分析による cgm 動画再生回数推定. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), pp. 1–8. 電子情報通信学会 データ工学研究専門委員会, 2018.

[6] 山本岳洋, 中村聡史ほか. 視聴者の時刻同期コメントを用いた楽曲動画の印象分類. 情報処理学会論文誌データベース (TOD), Vol. 6, No. 3, pp. 61–72, 2013.

[7] 土屋駿貴, 大野直紀, 中村聡史, 山本岳洋. ソーシャルコメントからの音楽動画印象推定手法の提案. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016), pp. 1–7. 一般社団法人電気情報通信学会, 2016.

[8] 村上直至, 伊東栄典. 視聴者コメントに基づく動画検索手法. 電子情報通信学会 2010 年総合大会講演論文集, Vol. 500, pp. 24–24, 2010.

[9] 堺雄之介, 伊東栄典. 動画サイトにおける視聴者コメントの特徴抽出. 人工知能学会研究会資料 知識ベースシステム研究会 124 回 (2021/11), pp. 17–22. 一般社団法人 人工知能学会, 2021.

[10] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments? analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, pp. 891–900, 2010.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 13:30 ~ 15:40 | T3-B オンライン（Zoom Events)

インタラクティブ情報アクセス

座長:中野 美由紀(津田塾大学)
コメンテータ:関 洋平(筑波大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T3-B-5-01] 推薦システムにおけるユーザの新たな嗜好の発見を支援するNeuMF を用いたインターフェース

*坂本 新真[1]、牛尼 剛聡[1] (1. 九州大学　牛尼剛聡研究室)

13:50 ~ 14:15
[T3-B-5-02] 探索的検索における検索視野拡張のための仮想他者視点の生成と評価

*小竹 神[1]、北山 大輔[1] (1. 工学院大学)

14:15 ~ 14:40
[T3-B-5-03] 観点の自動抽出に基づくインタラクティブな観光スポット推薦システム

*小林 らんう[1]、上野 史[2]、太田 学[2] (1. 岡山大学大学院環境生命自然科学研究科、2. 岡山大学学術研究院環境生命自然科学学域)

14:40 ~ 15:05
[T3-B-5-04] 大規模言語モデルを用いたその場での要約に基づくレビュー探索インタフェース

*藤井 真梨乃[1]、河田 友香[1]、山本 岳洋[1] (1. 兵庫県立大学)

15:05 ~ 15:30
[T3-B-5-05] 個人に最適な生活改善行動を対話的に探索するためのYES/Noチャート自動生成

*織田 直也[1]、金 鎭赫[1]、荘司 慶行[1]、山本 祐輔[2] (1. 静岡大学大学院 総合科学技術研究科 、2. 名古屋市立大学データサイエンス学部)

推薦システムにおけるユーザの新たな嗜好の発見を支援するNeuMFを用いたインターフェース

坂本　新真†　　牛尼　剛聡††

† 九州大学芸術工学府　〒 815-8540 福岡県福岡市南区塩原 4 - 9 - 1
†† 九州大学芸術工学研究院　〒 815-8540 福岡県福岡市南区塩原 4 - 9 - 1
E-mail: †sakamoto.shinma.626@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　ユーザに対して，おすすめのアイテムを提示する推薦システムは，ＥＣサイトや動画配信サービスで重要な役割を果たしている．しかし，従来の推薦システムは，ユーザが自覚している好みを強化する傾向があり，新しい好みの探索に適しているとは言い難い．本研究では，ユーザの新たな好みの発見を支援するために，深層学習モデルを用いて，ユーザの高次元な消費プロファイルと好みを捉え，ユーザにとって未知である可能性が高く，かつ好みに合うアイテムの空間を推定する手法を提案する．

キーワード　インターフェース・インタラクション，推薦システム

1　はじめに

ユーザに対して，おすすめのアイテムを提示する推薦システムは，ＥＣサイトや動画配信サービスで重要な役割を果たしている．正確性を重視した従来の推薦システムの多くは，膨大なアイテムの中から，ユーザに関連性の高い少数のアイテムを選び出すことによって，ユーザの意思決定のプロセスを支援する一方で，私たちが消費するアイテムの多様性を狭める可能性があることや [1]，ユーザに退屈な体験を与える可能性があることが指摘されている [2]．従来の推薦システムの多くは，ユーザの現在の好みを強化する傾向があり，新しい好みの探索には適していない．

これに対して Wang ら [3] は，推薦システムによって，ユーザ特徴とアイテム特徴の両方においてフィルターバブルが発生することを明らかにし，これらの特徴量をユーザ自らがコントロール可能なフレームワークを提案した．Movie Lens などのデータセットを用いた実験により，精度と多様性の両方において有望な性能を示した．また，Jayachithra ら [4] は，ユーザの楽曲ジャンルに基づく消費プロファイルを全体のユーザ群と比較し，可視化することで，ユーザが特に消費しているジャンルと，消費していないジャンルを明らかにした．これにより，ユーザがこれまでに消費していないジャンルを発見し，それらのジャンルの楽曲の聴取を促すインターフェースを提案した．

これらの研究では，ユーザに対して新しい好みの探索を促すが，いずれもユーザがジャンルやその他のアイテム特徴の選択、操作を行う必要がある．特徴量の選択や操作は，ユーザにとって負担が大きく，また，ユーザが自身の好みを正確に把握していない場合，ユーザが操作や選択によって新たな好みを発見することは難しい．また，ユーザの消費プロファイルは単にジャンルなどのアイテム特徴によって表現されるとは限らず，より複雑な構造を持つ可能性がある．

本研究では，ユーザの新たな好みの発見を支援するために，

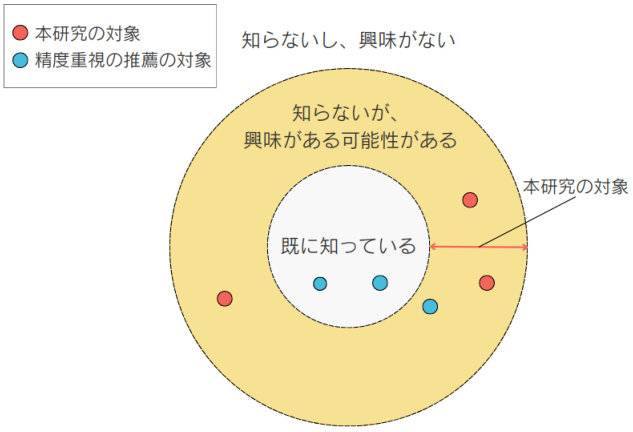


図 1　提案手法の概要

深層学習モデルを用いて，ユーザの高次元な消費プロファイルと好みを捉え，ユーザにとって未知である可能性が高く，かつユーザの好みに合うアイテムの空間を推定する．図 1 に示すように，従来の精度重視の推薦システムは，ユーザの既知の好みを強化するアイテムを推薦する傾向があるが，本研究では，ユーザの新たな好みの可能性を重視した空間を推定し，ユーザに提示する．

本論文の構成は以下の通りである．2 章では，関連研究について述べる．3 章では，提案手法について述べる．4 章では，予備実験とその結果について述べる．最後に，5 章でまとめと今後の課題について述べる．

2　関連研究

推薦システムによる消費コンテンツの多様性の減少を解決しようとする研究がこれまでにいくつか行われている．Tintarev ら [5] は，ユーザが消費プロファイルをより良く理解することを支援するために，ECLAT アルゴリズムによって，ジャンルの組み合わせを分析し，コードダイアグラムと棒グラフの 2 つ

で可視化する手法を提案している．これにより，ユーザが消費していないジャンル（ブラインドスポット）を発見することができることを示した．

Jayachithra ら [4] も同様に，ユーザの消費プロファイルと，ブラインドスポットを視覚化することで，フィルターバブルを抜け出すことを目的とするインターフェースを提案している．楽曲に関するデータを用いて，ユーザの楽曲ジャンルに関する頻度を抽出し，散布図と棒グラフを用いて視覚化する．また，実験によって，自身の消費プロファイルをより理解しているユーザほど，アイテムの探索傾向が高いことを示した．

Kunkel ら [6] は協調フィルタリングにおいて，ユーザの好みを視覚化し操作するための 3D マップベースの視覚化手法を提案した．標準的な行列分解手法を用いて得られた潜在因子を用いて，アイテム空間を 2D マップとして視覚化し，ユーザの好みをマップの高低として表現することで，アイテム空間全体を表現すると同時に，ユーザが高低を操作することで，好みの空間を探索することができる．

これらの手法は，ユーザの高次元の消費プロファイルを考慮していないことや，ユーザが新たな好みを発見するために，ユーザが自身の好みを正確に把握している必要があることなどの問題がある．本研究では，深層学習を用いて高次元の消費プロファイルや好みを考慮したアイテムの空間を推定することで，これらの問題を解決することを目指す．

3 提 案 手 法

3.1 提案手法の概要

提案手法では，ユーザにとって未知である可能性が高く，かつユーザの好みに合うアイテムの空間を推定する．以下ではこのようなアイテムの空間を探索空間と呼ぶ．探索空間の推定は，3つのステップで行う．はじめに，ユーザの過去のインタラクション情報から，深層学習モデルの NeuMF [7] を用いて，ユーザのアイテムに対する未知度を算出する．次に，ユーザの明示的なフィードバックを考慮して，ユーザのアイテムに対する好ましさを推定する．最後に，これらの 2 つの推定値から，ユーザの探索空間を推定する．本研究では，対象とするデータとして Anime Dataset[1] を利用する．

3.2 NeuMF によるアイテムの未知度の推定

3.2.1 NeuMF

NeuMF [7] は，暗黙的なフィードバックを用いたニューラル協調フィルタリング手法である．図 2 に示すように，NeuMF は Generalized Matrix Factorization(GMF) と Multi-Layer Perceptron(MLP) の 2 つを含むサブネットワークで構造化されており，単純なドット積ではなく 2 つのサブネットワークからの相互作用をモデル化している．GMF は，以下の式で表される．

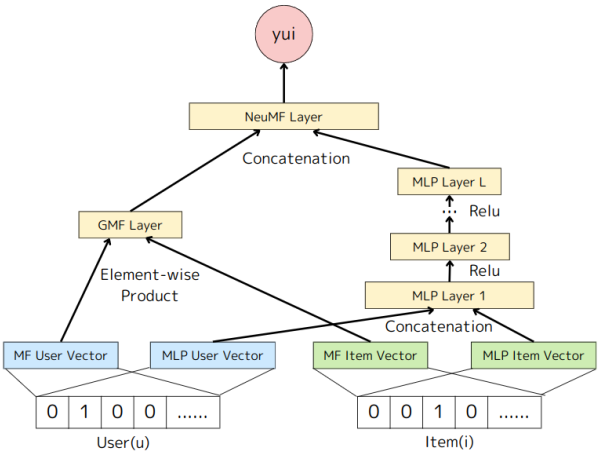


図 2　NeuMF の構造

$$\begin{aligned} \mathbf{x} &= \mathbf{p}_u \odot \mathbf{q}_i \\ \hat{y}_{ui} &= \alpha(\mathbf{h}^\top \mathbf{x}) \end{aligned} \tag{1}$$

ここで，$\mathbf{p}_u \in \mathbb{R}^k$ はユーザ u の特徴ベクトル，$\mathbf{q}_i \in \mathbb{R}^k$ はアイテム i の特徴ベクトル，α および $\mathbf{h}$ は出力層の活性化関数と重みであり，$\hat{y}_{ui}$ は予測値である．

もう 1 つのサブネットワークである MLP は，以下の式で表される．

$$\begin{aligned} z^{(1)} &= \phi_1(\mathbf{U}_u, \mathbf{V}_i) = [\mathbf{U}_u, \mathbf{V}_i] \\ \phi^{(2)}(z^{(1)}) &= \alpha^1(\mathbf{W}^{(2)} z^{(1)} + b^{(2)}) \\ &\cdots \\ \phi^{(L)}(z^{(L-1)}) &= \alpha^L(\mathbf{W}^{(L)} z^{(L-1)} + b^{(L)})) \\ \hat{y}_{ui} &= \alpha(\mathbf{h}^\top \phi^L(z^{(L-1)})) \end{aligned} \tag{2}$$

$\mathbf{W}^*$ と $\mathbf{b}^*$，α^* はそれぞれ重み行列，バイアス，活性化関数である．$\mathbf{z}^*$ は対応する層の出力を表す．

NeuMF は，GMF と MLP の層を結合し最終的な予測値を算出する．

$$\hat{y}_{ui} = \sigma(\mathbf{h}^\top [\mathbf{x}, \phi^L(z^{(L-1)})]) \tag{3}$$

$\mathbf{h}$ は出力層の重みであり，NeuMF の予測値 $\hat{y}_{ui}$ は σ を活性化関数とする単一のニューロンの出力として表され，本研究ではこの予測値をユーザ u のアイテム i の消費可能性として用いる．

3.2.2 個人化されたアイテムの未知度の推定

NeuMF によってユーザ u のアイテム i の消費可能性が算出できると，ユーザが消費しない、つまり未知である可能性として $1 - \hat{y}_{ui}$ が考えられるが，一般に人気のないアイテムは，インタラクション情報が少ないため，消費可能性 $\hat{y}_{ui}$ が低くなる傾向がある．同様に，人気のアイテムは消費可能性 $\hat{y}_{ui}$ が高くなる傾向がある．

本研究では，特定のユーザにとってのアイテムの未知度が，全体的なユーザ群におけるそのアイテムの未知度と比べてどの程度であるかを求めるために，以下の式を用いる．

1：https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset

$$n_i = -\log \frac{\#\ users\ consumed\ item\ i}{\#\ users}$$
$$n_{ui} = -\log \hat{y}_{ui}$$
$$unknown_{ui} = \frac{n_{ui}}{n_i} \qquad (4)$$

n_i はアイテム i のグローバルな未知度を表し，$unknown_{ui}$ が求める個人化されたアイテムの未知度である．

3.3 ユーザのアイテムに対する好ましさの推定

ユーザのアイテムに対する好ましさの推定は，ユーザの明示的なフィードバックを用いて行う．3.2.1 節で述べた NeuMF は，本来ユーザの暗黙的評価値を用いて学習を行うが，本研究では，NeuMF の構造がユーザの明示的な評価値予測にも適用できると考え，学習を行う．学習の際，ネガティブサンプリングは行わず，損失関数として平均二乗誤差を用いる．また，このように学習した NeuMF によるユーザのアイテムに対する好ましさの推定値を fav_{ui} とする．

3.4 探索空間の推定

探索空間の推定は，NeuMF によるアイテムの消費可能性の推定と，ユーザのアイテムに対する好ましさの推定の 2 つの推定値を用いて行う．本研究の目的は，ユーザにとって未知である可能性の高いアイテム群の中で，ユーザが好みそうなアイテムの空間を推定することである．そこで，3.2 節で述べたユーザのアイテムに対する未知度 $unknown_{ui}$ と，3.3 節で述べたユーザのアイテムに対する好ましさの推定値 fav_{ui} を用いて，以下の式で探索空間を推定する．

$$s_{ui} = \alpha * unknown_{ui} + (1 - \alpha) * fav_{ui} \qquad (5)$$

ユーザ u に対して，すべてのアイテムのうち s_{ui} の上位をユーザの探索空間とする．また，α はユーザ u の未知度に対する重みを表し，α が大きいほど，ユーザの探索空間は，より未知の可能性の高い空間に広がる．

4 予備実験

4.1 実験設定

提案手法によって得られた探索空間の有効性を検証するために，被験者 4 名に対し予備実験を行った．予備実験には，myanimelist[2]から得られたアニメ関する情報と、ユーザのアニメの視聴履歴および評価値を元に作成された Anime Dataset から，さらにユーザ数 1997 人，アイテム数 1890 のサブセットを作成し，実験に用いた．評価値の値は 1 から 10 までの整数で表されている．

被験者には，はじめに図 3 に示すような Web アプリケーションを用いて，アニメの評価値を入力してもらった．評価値を入力したアニメが 30 になった時点で，自動的に終了し，評価値のデータが収集された．その後，提案手法と比較手法によって得られたアニメ 50 件のリストを被験者に提示し，それぞれのリストに対して，3 つの評価項目に対して 5 段階評価を行ってもらった．

図 3　被験者の評価値データ収集用の Web アプリケーション

項目 1(興味):　あなたのアニメに関する現在の興味嗜好と一致しているかどうか．(1: 一致していない, 2: あまり一致していない, 3: どちらとも言えない, 4: やや一致している, 5: 一致している)

項目 2(知識):　あなたのこのアニメに対する認識がどうか．(1: 全く知らない, 2: あまり知らない, 3: どちらとも言えない, 4: やや知っている, 5: よく知っている)

項目 3(発見性):　あなたの新しい好みの発見につながるかどうか．(1: 全くつながらない, 2: あまりつながらない, 3: どちらとも言えない, 4: ややつながる, 5: つながる)

明示的評価値を利用した好みの推定には，評価値を正規化した値を正解ラベルとして用いた．ユーザの消費可能性の推定には，インタラクションがあったユーザ・アイテムのペアを正例として用いた．

NeuMF のパラメータは，GMF と MLP 層のユーザ・アイテム特徴の次元数を 64 とし，MLP 層には (128, 64) の 2 層を用いた．また，バッチサイズは好みの推定には 512，消費可能性の推定には 1024 を用い，エポック数，学習率はそれぞれ 50, 0.001 とした．また，式 (5) における α は 0.2 とした．

4.2 比較手法

提案手法の有効性を検証するために，以下の 3 つの比較手法を用いた．

- FAV: 提案手法の NeuMF によるユーザのアイテムに対する好ましさの推定のみを用いた手法
- UNKNOWN: 提案手法の NeuMF によるユーザのアイテムに対する未知度の推定のみを用いた手法
- RANDOM: ランダムに選択したアイテムを提示する手法

2：https://myanimelist.net/

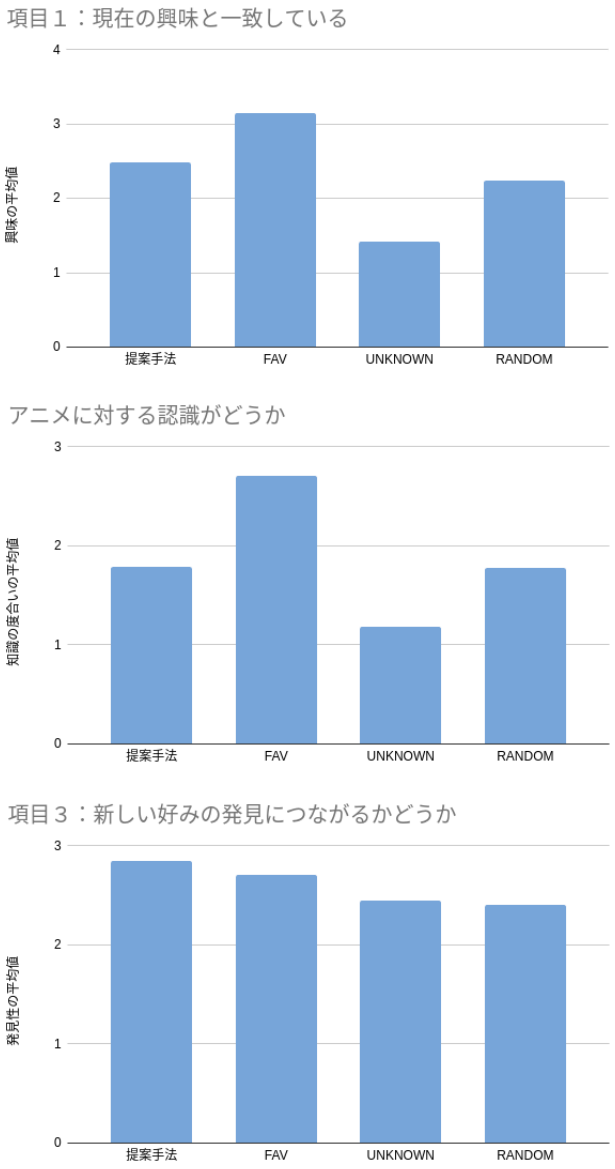


図 4　各項目の評価の平均値

4.3　被験者実験の結果

図 4 に，提案手法と比較手法によって得られたアニメ 50 件のリストに対する被験者の各評価項目の平均値を示す．項目 1(興味) に関しては，FAV が最も高く，提案手法は RANDOM や UNKNOWN よりも高いことがわかった．このことから，提案手法が現在の興味を強化するアイテムばかりを提示するのではなく，ユーザの新たな好みの発見を重視していることが確かめられた．項目 2(知識) も同様に，FAV が最も高く，提案手法は RANDOM や UNKNOWN よりも高いことがわかった．ユーザの現在の好みを重視した従来の推薦システムと比べて，提案手法は未知のアイテムを提示することができることがわかった．また，UNKNOWN は RANDOM よりも低いことから，ユーザのアイテムに対する未知度の推定が有効であることがわかった．項目 3(発見性) に関しては，提案手法が最も高く，FAV が次に高く，RANDOM が最も低かった．このことから，提案手法がユーザの新たな好みの発見に適している可能性が高いことが確かめられた．

4.4　評価値予測の精度

本来ユーザの暗黙的なフィードバックを用いて学習を行う NeuMF が，ユーザのアイテムに対する評価値予測において，どの程度の精度を持つかを検証した．検証には，Matrix Factorization(MF) と Factorization Machine(FM) を用いた手法と比較した．FM では，アイテム特徴量として，アニメのジャンル，タイプ (テレビ，映画，OVA など)，原作の種類 (マンガ，ライトノベルなど)，エピソード数を用いた．また，データの 8 割を学習データ，1 割を検証データ，1 割をテストデータとした．

表 1　テストデータの RMSE

	NeuMF	MF	FM
RMSE	**0.1358**	0.1395	0.1386

表 1 に，NeuMF と比較手法によるテストデータの RMSE を示す．NeuMF が最も低い RMSE を示し，評価値予測において，NeuMF が有効であることがわかった．

4.5　探索空間の可視化

抽出した探索空間の可視化について述べる．可視化には Next.js と D3.js[3]を用いた．本研究では，バブルチャートを用いたインターフェースを提案する．myanimelist には，アニメが属するジャンル（例：Action, Comedy, Drama, ...）に関する情報が含まれている．このジャンル情報を用いて，アニメの探索空間を可視化する．

図 5 に示すように，バブルチャートは，ジャンルを円形のバブルとして表現し，バブルの大きさは，そのジャンルに属する探索空間のアイテムの数を表す．次に，ユーザがジャンルのバブルをクリックすると，そのジャンルに属するアイテムのバブ

図 5　インターフェース画面（ジャンル）

3：https://d3js.org/

図 6　インターフェース画面（アニメ）

ルが表示される．図 6 では，ユーザが Sci-Fi ジャンルのバブルをクリックしたときの画面を示している．この画面でのバブルの大きさは，式 (5) で求めた s_{ui} の値を表す．また，特定のアニメのバブルをホバーすると，画面左にそのアニメの情報が表示される．

5　おわりに

本研究では，ユーザの新たな好みの発見を支援するために，深層学習モデルを用いて，ユーザの高次元な消費プロファイルと好みを捉え，ユーザにとって未知である可能性が高く，かつ好みに合うアイテムの空間を推定する手法を提案した．予備実験から，提案手法によってユーザにとって未知である可能性が高く，かつユーザの好みに合うアイテムの空間を推定することができ，ユーザの新たな好みの発見を促進する可能性があることがわかった．今後は，提案したインターフェースを用いて，より多くの被験者に対して実験を行い，提案手法の有効性を検証する予定である．

文　献

[1] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, p. 677–686, New York, NY, USA, 2014. Association for Computing Machinery.

[2] Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A. Konstan, and Paul Schrater. "i like to explore sometimes": Adapting to dynamic user novelty preferences. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, p. 19–26, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 1251–1261, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Jayachithra Kumar and Nava Tintarev. Using visualizations to encourage blind-spot exploration. In Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, John O'Donovan, Giovanni Semeraro, and Martijn C. Willemsen, editors, *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2018, co-located with ACM Conference on Recommender Systems (RecSys 2018), Vancouver, Canada, October 7, 2018*, Vol. 2225 of *CEUR Workshop Proceedings*, pp. 53–60. CEUR-WS.org, 2018.

[5] Nava Tintarev, Shahin Rostami, and Barry Smyth. Knowing the unknown: Visualising consumption blind-spots in recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, p. 1396–1399, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. A 3d item space visualization for presenting and manipulating user preferences in collaborative filtering. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, p. 3–15, New York, NY, USA, 2017. Association for Computing Machinery.

[7] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering, 2017.

探索的検索における検索視野拡張のための仮想他者視点の生成と評価

小竹　神†　北山　大輔†

† 工学院大学大学院工学研究科情報学専攻　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †em22009@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　一般的に，Web 検索はユーザが検索キーワードを入力し，検索エンジンの検索結果ページ（SERP）とその SERP 中の Web ページを見て行われる．また，情報ニーズは同じトピックを検索する他のユーザと類似していることが多い．しかし，情報ニーズが類似していても，ユーザが閲覧する情報はそれぞれ異なる．本研究では，検索ユーザのクエリ履歴と SERP のタイトルとスニペットおよびユーザが閲覧したページ内の単語で他のユーザのクエリ履歴を仮想的に作成し，検索ユーザが閲覧したページ内の単語を判定し，判定結果に基づいてその単語をハイライト表示するシステムを実装した．そのシステムをユーザ実験により評価した．

キーワード　行動分析，多様性，インタフェース・インタラクション，探索的検索，協調検索

1 背　　景

検索エンジンの発展によって，ウェブ検索を用いて情報収集することが多くなっている．検索ユーザは，自身の興味・観点に基づいて検索をする．また，複数人で同様のトピックを検索し，目的を達成する検索もある．このような検索は，協調検索 [1] と呼ばれる．例えば，複数人で旅行計画を立てている際に，観光地について調べるときに協力して行われる検索である．このような背景から，情報ニーズ自体は，他者と類似していることも多いが，同様の情報ニーズであっても，ユーザによって閲覧する情報が異なる場合は，そのユーザにとって共有された情報は有用であると考える．

本研究では，ユーザ自身のクエリ履歴と閲覧履歴を用いて他者の検索行動モデルを仮想的に作成し，それによってユーザが閲覧したページ中の単語および SERP のタイトルとスニペット中の単語を他者が興味を持つか判定を行い，判定結果に基づいてその単語を強調表示するシステムの実装を行った．このシステムを用いて下記の 2 つの Research Question（RQ）を実験により検証した．

RQ1　他者の検索行動モデルを仮想的に作成することは可能か
RQ2　他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか

本研究の貢献を以下にまとめる．

- 他者の検索履歴は通常検索エンジンの提供者しか得ることができないデータであるが，同一トピックに関してユーザの検索履歴と異なる観点の検索履歴を仮想的に生成する手法を開発した．
- 自身の検索履歴に基づくキーワード強調と仮想的に生成した他者の検索履歴に基づくキーワード強調を実験により比較し，それぞれの性質の違いを明らかにした．

以下，2 章では関連研究について述べる．3 章では提案手法について述べる．4 章では，実験と考察について述べる 5 章では結論について述べる．

2 関 連 研 究

2.1 探索的検索に関する研究

Ward ら [2] は，探索的検索を行うユーザの支援をする機能が少ない問題に対して，メタ認知的な活動を支援するための検索支援ツール「Orgbox」を提案し，その評価を 2 つのユーザ実験から実施した．1 つ目の実験では，Orgbox とベースラインのブックマークツールを比較した．2 つ目の実験では，Orgbox と Google document スタイルのノート作成ツールの比較した．それらの結果から，Orgbox は探索的検索中にユーザに認知的・メタ認知的な支援を提供できることを示した．

Roberto ら [3] は，探索的検索のような不確実性が高いタスクにおいて，従来の検索結果の表示形式よりも，拡張した表示形式を提案している．この表示形式では，情報オブジェクトの特徴付けに役立つ定量的指標（検索エンジンランキング，読みやすさ，新鮮さ，評判など）に由来する様々な視覚的手がかりを用いて検索結果を表示する．この表示形式と従来の表示形式を比較するために，SERVS（検索エンジン結果可視化システム）を開発し，ユーザ実験を実施した．その結果，SERVS は，適切な視覚的手がかりを用いた多数の検索結果により，ユーザのパフォーマンスや体験を大きく損なうことなく，身体的・視覚的労力の軽減に役立つことが示唆された．

He ら [4] は，探索的検索を流動的にサポートするビジュアル インターフェイスをどのように設計するかを研究課題とし，ユーザ実験を実施した．具体的には，ファセット検索インターフェイスに関する先行研究をレビューし，流動的な検索動作をサポートする際の欠点を特定し，ファセット検索インターフェイスを提案している．2 つのユーザ実験の結果，提案したツールは，学習が容易であり，クエリ検索インターフェイスと比較して同等のパフォーマンスを備えていることを示した．

Soulier ら [5] は，探索的な検索において複数のユーザで検索する際のアプローチを提案した．具体的には，探索的タスクを 2 人のユーザが協力して行う際に，システムが両者の検索の差

異を発見し，検索における役割を提案するというアプローチである．被験者実験の結果，4 つの比較手法と比べて，検索セッション全体の検索精度が，有意に向上していることを示した．

Golovchinsky ら [6] は，探索的検索のための検索システムを提案した．具体的には，複数のクエリの結果をマージすることや関連性フィードバックのために文書を選択する等の検索行動の履歴を明示的に示し，利用できるものである．また，使いやすいインタフェースの課題として，インタラクティブ性と応答性が重要であると述べている．

これらの研究と本研究との関連は，ユーザ視点での探索的検索支援手法を提案している点である．本研究では，探索的検索の検索トピックにおいて，仮想的に他ユーザの履歴を作成し，その他ユーザの興味をユーザの閲覧ページ中にて可視化させることでユーザの得られる情報の多様化を支援することが目的である．

2.2 協調検索に関する研究

Salvatore ら [7] は，協調検索において，各ユーザが，自発的に探索するための QueryTogether というシステムを提案し，その評価をユーザ実験を通して実施した．その結果，協調検索で起きる状況のうち，1 人のユーザーが検索している間，他のユーザーが見守って指導するという状況を回避することを示唆し，共同検索における公平な貢献の促進をもたらすことを示した．

Nakayama ら [8] は，ユーザが自ら検索文脈を理解することを目的とした検索対象の視覚管理システムを提案し，検索意図と実際の閲覧状況との差異を抽出する手法の評価を行った．提案手法は，適合率・再現率・F 値によって評価した．その結果，適合率・再現率ともに高い値であることを示した．

Pickens ら [9] は，協調検索の仲介をするシステムを構築し，評価し，検索クエリを共有する検索エンジンを使用した場合に，ユニークな関連文書の発見数等が向上することを示した．

Reimers ら [10] は，Collaborative Information-Seeking (CIS) において，従来のシングルユーザデザインに基づくシステムでユーザ実験を実施し，課題を抽出した．その課題に対処するために複数での CIS 向けのシステムである Cometogether を提案し，その効果を 4 週間のユーザ実験とインタビューにより評価した．その結果，他人の好みを他のメンバーが明示的にみることが相互理解の促進や他のメンバーに対する意識の向上につながったことを示唆した．

Morris [11] は，204 人のウェブ検索行動の調査を実施し，ユーザが検索の過程と検索結果の両方において共同作業を行っていることを明らかにした．この結果から，協調検索の方法や頻度等の知見を提供した．

Yamamoto ら [12] は，協調検索における 2 つの役割が検索性能とクエリ作成に与える影響を調査した．具体的には，検索トピックを深掘りする役割と，探索対象を拡張し多様な情報を探す役割が検索行動に与える影響を調査した．被験者実験の結果，前者はトピックの探索に時間を費やすため，探索されていないトピックを見つけることが難しくなることや，後者は前者が調査した内容をもとに検索する内容を決めていることを示した．

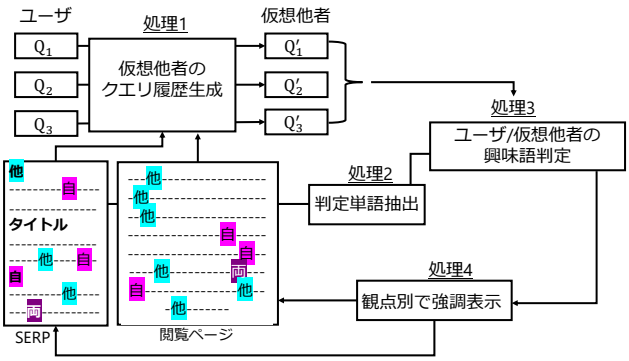


図 1　提案システムの概要

表 1　閲覧ページを embedrank で抽出した例と**候補ワード**．下線はストップワード処理で除外した文字

	フレーズ
P1	**完全 禁煙** スタンダード コース 9800 円 プレミアム 掲載 店 26 うかい 鳥山 八王子
P2	**楽しい 食事** 時間 海鮮 料理
P3	**多い 店内**
P4	**女子 会** 等 様々
P5	**銀座 開放** 感
P6	**話題 ラクレットチーズ** 期間 限定 入荷 !! 濃厚
P7	**雰囲気 抜群** KICHIRI 池袋 店
P8	**鍋 保存** 秋葉原 駅
P9	**合掌 づくり** 隠れ 座敷
P10	**高級 A** 5 ランク 黒毛 和牛 自慢 食材 鉄板 焼き

Kim ら [13] は，複数人における共同作業が学習の過程や結果にどのような影響を及ぼすかを調査した．具体的には，大学院生を複数人のグループに分け，共通の課題を共同作業で取り組んだ．その結果，多くの情報を得ることはできたが，その情報を整理することが難しいことを明らかにした．

Moraes ら [14] は，協調検索におけるグループサイズが検索指標や検索の成果にどのような影響を与えるのかを調査した．具体的には，クラウドソーシングサービスを用いた 300 人規模での実験にて，最大 6 人のグループでの共同検索で，グループサイズによる影響の仮説を立て調査した．実験の結果，グループサイズが大きくなることでグループが検索で得られる情報が増加することを示した．しかし，実験は，リモートで行われたため，今回の結果が，他の状況設定には一般化されないことも示した．

これらの研究と本研究との関連は協調検索支援手法を提案している点である．本研究では，実ユーザが無くとも，仮想的に他ユーザの履歴を作成し，それを閲覧ページ中にて他ユーザの興味として可視化させることで，協調検索で得られる情報の多様化をユーザ実験により調査した．

3 提 案 手 法

提案システムの処理の概要を図 1 に示す．提案システムは，処理 1 から処理 4 の流れで成り立つ．まず，図 1 の処理 1「仮想他者のクエリ履歴生成」について述べる．この処理では，現

表 2　仮想他者のクエリ履歴の生成例

	ユーザのクエリ履歴	仮想他者のクエリ履歴
Q1	東京 グルメ 浅草	東京 墨 田川
Q2	東京 グルメ 上野	東京 全席 扉
Q3	東京 グルメ 人気	東京 高級 A

表 3　閲覧ページを TFIDF で抽出した例と**強調単語**

上位	単語
10 位	**オススメ**, **個室**, **料理**, 件, 貯まる, 徒歩, 飲む, 税込, 空席, 禁煙
20 位	デト, コス, 和食, 保存, 予約, イタリアン, メニュ, **居酒屋**, シェフ, 食事
30 位	牡蠣, **使える**, 焼肉, 店内, 洋食, **完全**, 行う, 空間, **秋葉原**, あり
40 位	**新宿**, 寿司, 堪能, パスタ, テイクアウト, 厳選, ランチ, 和牛, 鍋, おすすめ
50 位	**接待**, 食材, **食べる**, 海鮮, 利用, サラ, する, 用意, 楽しめる, 倍

表 4　強調単語の判定結果の例

	強調単語	類似度	計算元の履歴
上位 1 位	個室	0.92	ユーザ
上位 2 位	完全	0.88	ユーザ
上位 3 位	秋葉原	0.83	ユーザ
上位 4 位	食べる	0.79	仮想他者
上位 5 位	料理	0.71	仮想他者
上位 6 位	使える	0.69	ユーザ
上位 7 位	居酒屋	0.64	仮想他者
上位 8 位	接待	0.55	ユーザ
上位 9 位	オススメ	0.50	仮想他者
上位 10 位	新宿	0.44	仮想他者

図 2　閲覧ページにおける強調表示の例

在検索しているユーザの検索クエリを用いて仮想他者のクエリ履歴生成を行う．本稿では，ユーザが SERP 中のタイトルとスニペット中の単語と閲覧しているページ中の単語からその検索クエリに対する他者のクエリの候補語を取得する．SERP は，検索結果中のすべてのタイトルとスニペットを結合し，1 つの文書としたものを用い，閲覧ページはページの内容を 1 つの文書として用いる．

- 対象文書に対して，embedRank [15] を使用し，その値が高い上位 N フレーズを取得
- 上位 N フレーズから不要な語（記号等）を削除した単語集合を候補語としてデータベースへ挿入
- データベースからそのユーザのクエリでの候補語を取得

EmbedRank は，文書全体に類似するようなキーフレーズを抽出するアルゴリズムで，ここのキーフレーズが類似しないように，文書要約でも用いられる Maximal Marginal Relevance を用いている．本稿では，類似度の判定に fasttext を用い，文書全体におけるキーフレーズの出現頻度の対数を重みとして加えるという変更を加えている．上記の 1 つ目について詳細を述べる．表 1 に閲覧ページ A[1]を EmbedRank で抽出した例を示す．P1 は上位 1 位のフレーズであり，P10 は上位 10 位のフレーズを表す．各フレーズの先頭の 2 単語（太字）をユーザのクエリでの候補語とした．上記の 2 つ目について詳細を述べる．以下に，上位 N フレーズから不要な語（記号等）を削除対象の文字を示す．表 1 中のフレーズにおいて下線で引かれている箇所は，除外した文字であるが，文字単位で処理している．これによって,「スタンダード」や「ラクレットチーズ」等の単語中に「ー」が存在している場合はそれを除外する処理となる [2]．

（1） 半角記号，英字
（2） 全角記号
（3） 改行文字
（4） その他の単語：例えば，–･、。，～♪ー ©®¥【】()()〈〉『』「」等

次に，SERP のタイトルとスニペット，閲覧ページ中から得た候補語とユーザの検索クエリとの類似度を算出し，もっとも値が低い候補語と検索ユーザのクエリの 1 語目を組み合わせたものを仮想他者の m 番目の検索クエリとする．この m 番目の検索クエリが，最新となる．仮想他者の検索クエリの 1 語目は，検索ユーザの 1 語目を使用する．理由は，検索トピックを代表する可能性が高いためである．ここでは，ユーザの検索クエリと候補語を fasttext を用いてベクトル化し，コサイン類似度を用いて算出した．また，ユーザが入力した検索クエリが 1 語の場合には，その 1 語目の単語を 2 語目の単語としても扱った．たとえば，入力された検索クエリが,「Tokyo」であった場合は，そのユーザのクエリを「Tokyo Tokyo」として扱った．

そして，仮想他者の m-1 番目以降の検索クエリ生成について説明する．仮想他者の検索クエリにも，実際のユーザの検索クエリのように一貫性があることが求められる．そこで，仮想他者の m-1 番目以降の検索クエリの条件は「ユーザの m-i 番目の検索クエリとは類似しなく，仮想他者の m-i+1 番目（直前）のクエリと類似する」と設定した．それは，式（1）で表現した．式（1）について，$Score(c)$ の $\mathrm{sim}(q_{m-i,2}, c)$ はユーザの m-i 番目以降の単語と候補語 c との類似度，$(1-\mathrm{sim}(p_{m-i+1,2}, c))$ は仮想他者の m-i+1 番目以降の単語と候補語 c との類似度を示す．また，類似度は，コサイン類似度により算出する．式（1）より求めたスコア c がもっとも低い候補語を仮想他者の検索クエリとする．表 2 に仮想他者のクエリ履歴の生成例を示す．具体的には，表 2 のユーザが Q3 の「東京 グルメ 人気」で検索を行い，その Q3 で閲覧ページ A から表 1 の太字の候補語とユーザの Q3 の 1 語目を組み合わせ，仮想他者の Q3「東京 高級 A」が生成された例になる．

1：https://hitosara.com/ranking/tokyo/
2：この処理の状態で 4 章で示す実験を行っているため，適切なストップワード処理ができていないこととなる．これの改善として，文字単位でなく，形態素解析を行い単語単位でストップワード処理をする必要がある．

$$score(c) = \mathrm{sim}(q_{m-i,2}, c) \cdot (1 - \mathrm{sim}(p_{m-i+1,2}, c)) \quad (1)$$

次に，図 1 の処理 2「判定単語抽出」について述べる．閲覧中の SERP もしくは Web ページの判定単語抽出は，ページ中の文書に対して，TFIDF を使用し，不要な語（記号等）を削除したその値が高い上位 K 単語とした．表 3 に閲覧ページ A を TFIDF で抽出した例を示す．こちらの単語も，上記と同様にストップワード処理を行った結果である．具体的には，上位 1 位から 50 位までの単語を示している．

最後に，図 1 の処理 3「ユーザ/仮想他者の興味語判定」と処理 4「SERP・閲覧ページに対する興味判定結果の提示」について述べる．まず，処理 3 について述べる．処理 3 は図 1 の処理 2 で得られた判定単語と処理 1 で得られた仮想他者のクエリ履歴とユーザのクエリ履歴との類似度判定を行う．ここでも，コサイン類似度を使用した．この類似度を興味度として扱い，ユーザと仮想他者の興味度を降順にソートし，上位 T 件を強調表示する対象とする．この候補は，表 3 の全ての単語である．強調表示する候補と表 2 に示すユーザ/仮想他者のクエリ履歴と類似度判定を行う．表 4 に類似度判定結果を示す．具体的には，上位 T 件に入る興味度が，ユーザとの類似度によるものなら，ユーザの観点の単語，仮想他者との類似度によるものなら，仮想他者の観点の単語，上位 T 件に同じ単語が両方の類似度に基づいて入っているなら両方の観点の単語とする．次に，処理 4「SERP・閲覧ページに対する興味判定結果の提示」について述べる．この処理は，処理 3 で得られたユーザ/仮想他者/両方が興味を持つであろう単語を SERP 中のタイトルとスニペット，閲覧ページ中にて強調表示する処理である．具体的には，SERP 中のタイトルとスニペット，閲覧ページ中で該当する単語の文字の色をユーザ/仮想他者/両方によって適応した．具体的には，ユーザ観点のみの場合には，該当する単語を薄い赤色でハイライト形式で強調表示した．仮想他者観点の場合には，該当する単語を薄い青色でハイライト形式で強調表示した．ユーザ観点と仮想他者観点の両方の場合には，該当する単語を薄い紫色でハイライト形式で強調表示した．図 2 に，閲覧ページ A での強調表示の例を示す．ユーザ判定された単語は,「個室」,「完全」,「使える」であり，その背景色が薄い赤色に変化していることがわかる．仮想他者判定された単語は,「新宿」であり，その背景色が薄い青色に変化していることがわかる．このように，表 4 に示す上位 10 単語を観点別で強調表示している．

4 実験と考察

4.1 実験設定

3 章で述べた提案システムを用いて，下記の 2 つの Research Question（RQ）を実験により検証した．

RQ1　他者の検索行動モデルを仮想的に作成することは可能か
RQ2　他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか

ユーザ実験の被験者の検索行動の取得方法は共通の設定である．

実験は，図 3 に示すように設計した．被験者は，まずクラウドソーシング [3] にアクセスし，指定した URL を開き，検索タスクの説明ページを閲覧後，図 3 の流れの通りに実験を実施した．具体的には，以下の流れである．

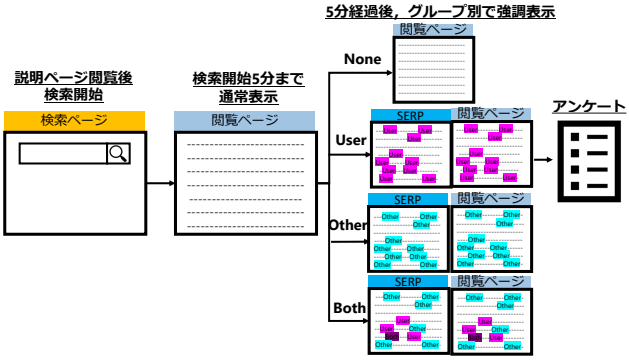


図 3　ユーザ実験の流れ

（1） 検索をはじめてから 5 分まで通常の表示形式で指定された検索タスクで検索をする

（2） 5 分経過後，4 つの表示パターンのうち 1 つの表示形式に自動で切り替わり，引き続き同じ検索タスクで検索をする

（3） 10 分経過後，クラウドソーシングのページに戻り，アンケートに回答する

まず，実験に使用した検索トピックについて説明する．具体的には，以下の 2 つである．探索的検索は，検索ユーザの知りたいことが動的に変化する検索であるため，本稿では，観光地における旅行計画の検索トピックを設定した．

- 東京で 1 日過ごす際の計画を立てるための検索
- 京都で 1 日過ごす際の計画を立てるための検索

次に，実験に使用した閲覧ページでの表示形式について説明する．検索タスクを開始して，5 分経過までは，どの実験参加者にも共通で，一般的な表示形式での検索を実施した．5 分経過後は，以下の 4 つのいずれか表示形式となる．図 4 に実際にユーザに提示されるハイライト表示の例を示す．

1 つ目（None）は，一般的な表示形式である．これは，前半の 5 分と同様の表示形式である．2 つ目（User）は，ユーザ自身の検索履歴にマッチする単語を薄い赤色でハイライトした表示形式である．これは，ユーザが検索タスクを開始して，5 分経過までの検索履歴を用いてマッチする単語を 3 章の手法で判定し，判定された単語を薄い赤色でハイライトした表示形式にしたものである．具体的には，ユーザ観点と判定された単語は，図 4 の User viewpoint で示されている薄い赤色でハイライトされる．3 つ目（Other）は，同じトピックを検索する他のユーザ（仮想他者）の検索履歴にマッチする単語を薄い青色でハイライトした表示形式である．これは，ユーザが検索タスクを開始して，5 分経過までの検索履歴を用いて，3 章の手法で仮想他者のクエリ履歴を生成し，それらにマッチする単語を 3 章の手法で判定し，判定された単語を薄い青色でハイライトした表示形式にしたものである．具体的には，仮想他者観点と判定された単語は，図 4 の Other User viewpoint で示されている薄

3：本研究ではクラウドワークスを用いた，https://crowdworks.jp/

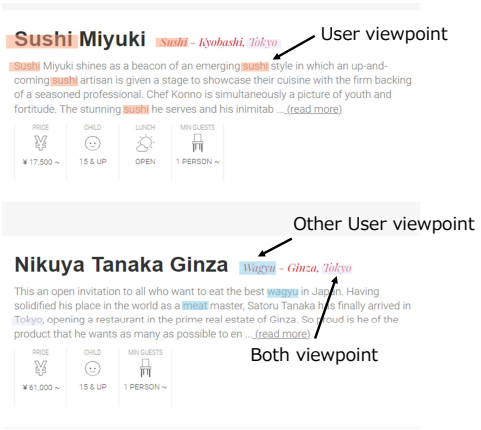


図 4　Web ページ上でのハイライト例

い青色でハイライトされる．4 つ目（Both）は，2 つ目と 3 つ目の表示に，ユーザと他のユーザの検索履歴にマッチする単語を薄い紫色でハイライトした表示も加えたものである．具体的には，ユーザ観点かつ仮想他者観点と判定された単語は，図 4 の Both viewpoint で示されている薄い紫色でハイライトされる． 以上の 4 種類の表示形式の違いによる効果をユーザ実験による検証した．

（1） オリジナルの表示形式（ハイライトなし）

（2） ユーザ自身の検索履歴にマッチする単語を薄い赤色でハイライトした表示形式

（3） 同じトピックを検索する他のユーザ（仮想他者）の検索履歴にマッチする単語を薄い青色でハイライトした表示形式

（4） 上記 2 つの表示形式かつユーザと他のユーザの検索履歴にマッチする単語を薄い紫色でハイライトした表示形式

最後に，アンケートについて述べる．このアンケートは，ハイライトされた単語の参考度合いを 5 段階リッカート尺度の項目とハイライトされた単語の感想の自由記述の項目の 2 個で構成される．実験は，2023 年 12 月 19 日に，クラウドソーシングサービスを提供しているクラウドワークスにて実施した．被験者は，2 つの検索トピックと 4 つの表示パターンを組み合わせた各 10 名の計 80 名である．実験データとして使用した被験者数は，表示なし（None）11 名，ユーザ視点のみ（User）9 名，仮想他者視点のみ（Other）10 名，両方（Both）11 名の計 41 名である．具体的には，以下の条件を満たさない被験者のデータは除外した．

- 閲覧ページ数が 2 個以上
- 検索タスク前半と後半の滞在時間の差が 2 倍以下

まず，RQ1「他者の検索行動モデルを仮想的に作成することは可能か」を検証した実験 1 について述べる．実験 1 は，ユーザの入力クエリと仮想他者の生成クエリの分析を行った．具体的には，ユーザ実験と異なる被験者に対し，ある被験者のクエリ履歴と仮想他者のクエリ履歴を提示し，それぞれのクエリ履歴の性質の違いをアンケートにより検証した．詳細は，4.2 節で述べる．

次に，RQ2「他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか」を検証した実験 2 および実験 3 について述べる．実験 2 は，閲覧ページで強調された単語の分析を行った．具体的には，ユーザ実験と異なる被験者に対し，ある被験者のクエリ履歴とその被験者が最後に閲覧したページでハイライトされた各単語を提示し，アンケートで評価した．この実験では，強調された単語が，ユーザにとって多様なコンテンツを選ぶきっかけになるかを検証した．詳細は，4.3 節で述べる．

実験 3 はユーザ実験の行動ログデータの分析を行った．具体的には，以下の検索ログを収集し，前半と後半の変化量を考察した．この実験では，強調表示する UI が一般的な UI と比べ，検索行動に変化があるのかを検証した．

（1） 検索クエリの数の変化

（2） 閲覧ページ数の変化

（3） Web ページの 1 ページあたりの滞在時間の変化

（4） SERP の選択順位の変化

詳細は，4.4 節で述べる．

4.2　実験 1：仮想他者のクエリ履歴の評価

実験は，2024 年 1 月 18 日に，クラウドソーシングサービスを提供しているクラウドワークにて実施した．件数は，120 件である．一人あたり 3 件までは異なるタスクで実施可能とした．収集した検索ユーザのクエリ履歴と実験中に生成された仮想他者のクエリ履歴，最後に閲覧したページを用いて実験を設計した．具体的には，以下の流れで被験者に評価してもらった．

（1） 被験者にユーザのクエリ履歴 A とクエリ履歴 B の 2 つを提示する

（2） 提示したクエリ履歴 A とクエリ履歴 B が人間によるクエリか自動生成されたクエリかを 5 段階で評価する

（3） 提示したクエリ履歴 A とクエリ履歴 B の各履歴から想像できる興味カテゴリをそれぞれに付与する

（4） 提示したクエリ履歴 A とクエリ履歴 B の関係を 3 段階で評価する

ここでは，RQ1「他者の検索行動モデルを仮想的に作成することは可能か」を検証するための評価指標について述べる．まず，ユーザのクエリ履歴（クエリ履歴 A）と仮想他者のクエリ履歴（クエリ履歴 B）がそれぞれ人間によるクエリか自動生成されたクエリかを 5 段階で評価してもらった．具体的には，1 が人間らしいであり，2 がやや人間らしい，3 が判断がつかない，4 がやや機械らしい，5 が機械らしいの 5 段階である．この評価の意図として，仮想他者のクエリが 1 または 2 の評価が多い場合に，仮想他者のクエリは人間のクエリとして妥当だと評価できる．

次に，ユーザのクエリ履歴と仮想他者のクエリ履歴から想像できる興味カテゴリを付与してもらった．具体的には，各クエリ履歴に対して「グルメ」,「ショッピング」,「レジャー」,「夜景」等の検索トピック中でどんな興味があるのかをカテゴリとして自由に入力してもらい，評価をした．評価の意図として，仮想他者のクエリに付与されたカテゴリとユーザのクエリに付与されたカテゴリが異なる場合は，仮想他者としてのクエリ履

表 5　ユーザのクエリ履歴および仮想他者のクエリ履歴が人間による履歴か自動生成された履歴かの 5 段階での評価結果

履歴	1	2	3	4	5	平均	分散
ユーザ	58	38	8	12	4	1.88	1.12
仮想他者	12	18	15	56	19	3.43	1.21

表 6　ユーザのクエリ履歴と仮想他者のクエリ履歴の関係

選択肢	回答数	選択割合
共通の興味の同じ側面	40	0.33
共通の興味の異なる側面	61	0.50
無関係な側面	19	0.16

歴として妥当だと評価できる．

最後に，ユーザのクエリ履歴と仮想他者のクエリ履歴の関係を 3 段階で評価した．具体的には,「共通の興味の同じ側面」,「共通の興味の異なる側面」,「無関係な側面」の 3 段階である．この評価の意図として，仮想他者のクエリ履歴が「共通の興味の異なる側面」として評価された場合には，仮想他者のクエリ履歴が他ユーザの興味を表現できていると評価できる．

4.2.1　実験 1 の結果と考察

ここでは，RQ1「他者の検索行動モデルを仮想的に作成することは可能か」の検証結果について述べる．

まず，表 5 にユーザのクエリ履歴および仮想他者のクエリ履歴が人間による履歴か自動生成された履歴かの 5 段階での回答結果を示した．ユーザのクエリ履歴に対して，2 以下に 96 名が回答している．一方で，仮想他者のクエリ履歴に対して，4 以上に 75 名が回答している．また，ユーザのクエリ履歴と仮想他者のクエリ履歴の平均に対して有意水準が 5%の両側 t 検定を実施した結果，P 値が 1.83×10^{-14} となり，有意差ありとなった．これらの結果から，ユーザのクエリ履歴と仮想他者のクエリ履歴は異なることがいえる．この原因として，仮想他者のクエリが人間が入力するクエリと比べて不自然あることが考えられる．以下に，適切に生成されなかった例を挙げる．各クエリの 2 語目以降の「原材料 名」や「トップ 日本」,「店 自体」は，人間が入力するクエリとは考えにくい．このような不自然なクエリが生成されないために，仮想他者のクエリ履歴のサジェストワードの取得時のストップワード処理やサジェストワードから仮想他者のクエリの 2 語目以降として決定するための 3 章で述べた式 1 の改善が必要となる．

- 東京駅 原材料 名
- 京都 トップ 日本
- 東京 店 自体

最後に，ユーザのクエリ履歴と仮想他者のクエリ履歴の関係を 3 段階で評価した結果と考察を述べる．表 6 にユーザのクエリ履歴と仮想他者のクエリ履歴の関係を示した．共通の興味の同じ側面の回答数が，40 件であり，共通の興味の異なる側面の回答数が 61 件であり，無関係な側面の回答数が 19 件であった．ユーザのクエリ履歴と仮想他者のクエリ履歴の関係は,「共通の興味の異なる側面」がもっとも多い回答となった．このような結果となった理由は，仮想他者のクエリ履歴の 1 語目はユーザのクエリ履歴と共通していることが挙げられる．

以上のアンケートの結果から，RQ1「他者の検索行動モデルを仮想的に作成することは可能か」は部分的に支持できると考える．理由は以下の 2 点である．1 点目は，表 6 のユーザのクエリ履歴と仮想他者のクエリ履歴の関係においては「共通の興味の異なる側面」がもっとも多い回答となり，仮想他者のクエリ履歴がユーザのクエリ履歴と異なる側面を表していることを示したためである．2 点目は，表 5 のユーザのクエリ履歴および仮想他者のクエリ履歴が人間による履歴か自動生成された履歴かの 5 段階での回答結果では，仮想他者のクエリ履歴が「4 がやや機械らしい」,「5 が機械らしい」の回答が集中し，有意差検定の結果，仮想他者のクエリ履歴とユーザのクエリ履歴は有意差ありとなり，人間が入力するクエリとは異なったためである．

4.3　実験 2：ハイライトされた単語の評価

実験は，2024 年 1 月 18 日に，クラウドソーシングサービスを提供しているクラウドワークにて実施した．件数は，72 件である．ユーザ実験で収集したユーザの最後に閲覧したページでハイライトされた単語を用いて実験を設計した．具体的には，以下の流れで被験者に評価してもらった．

（1） 被験者にユーザ実験の被験者 A のクエリ履歴とその被験者が最後に閲覧したページでハイライトされた各単語を提示する

（2） 提示した各単語を 3 段階で評価する

ここでは，ユーザ実験で収集したデータのハイライトされた単語が次の検索に有用であるかを検証するための評価指標について述べる．ハイライトとされた単語を以下の 3 段階で評価をした．評価の狙いとして，ハイライトされた単語が 1 つ目または 2 つ目の評価が多い場合にハイライトされた単語が有用であると考える．

- 現在の検索トピックを詳細化するためのキーワードの一部に用いることができる
- 現在の検索トピックの別の観点を得るためのキーワードの一部に用いることができる
- 検索キーワードに含めて用いることはない

4.3.1　実験 2 の結果と考察

ここでは，ハイライトされた単語の検証結果について述べる．

ハイライトされた単語が次の検索に有用かを 3 段階で評価した結果と考察をする．まず，User 判定された単語における回答結果について述べる．表 7 の User 列に U 回答結果を示した．「詳細化するためのキーワードの一部に用いることができる」の回答数が 106 件となり,「別の観点を得るためのキーワードの一部に用いることができる」の回答数が 102 件,「検索キーワードに含めて用いることはない」の回答数が 70 件となった．この結果から，ユーザ観点と判定された単語は，詳細化または別の観点を得るためのどちらかに使用できるとなった．また，ハイライトされた単語が検索キーワードに用いることができるかでみると，約 7 割の回答が検索キーワードに用いることができるとあるため，ハイライトする単語が有用であることが考えられる．考察として，観点別の判定やハイライト単語候補の抽出方

表 7　それぞれに判定された単語における回答結果

選択肢	User	Other	Both
現在の検索トピックを詳細化するためのキーワードの一部に用いることができる	106	42	0
現在の検索トピックの別の観点を得るためのキーワードの一部に用いることができる	102	42	4
検索キーワードに含めて用いることはない	70	50	10

法に改善が必要である．

次に，Other 判定された単語における回答結果について述べる．表 7 の Other 列に回答結果を示した．「詳細化するためのキーワードの一部に用いることができる」の回答数が 42 件となり，「別の観点を得るためのキーワードの一部に用いることができる」の回答数が 42 件，「検索キーワードに含めて用いることはない」の回答数が 50 件となった．この結果から，仮想他者観点と判定された単語は，ユーザ観点と判定された単語と同様に，詳細化または別の観点を得るためのどちらかに使用できるとなった．また，ハイライトされた単語が検索キーワードに用いることができるかでみると，約 6 割の回答が検索キーワードに用いることができるとあるため，ハイライトする単語が有用であることが考えられる．考察として，ユーザ観点と判定された単語と同様に，観点別の判定やハイライト単語候補の抽出方法に改善が必要であると考える．

以上のアンケートの結果から，ハイライト単語は次の検索キーワードに部分的に有用であると考える．理由は以下の 2 点である．1 点目は，表 7 の User 判定された単語における回答結果と Other 判定された単語における回答結果で，詳細化または別の観点を得るためのどちらかに使用できるの回答が多い傾向であり，これはハイライトすることで次の検索キーワードに有用であることを示唆している．2 点目は，表 7 の回答結果において，詳細化または別の観点を得るためのどちらかに使用できるとなったため，観点別で想定していた効果が得られなかったためである．

4.4　実験 3：提案システムの有効性の検証

RQ2「他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか」を検証するための評価指標について述べる．実験システムのログとして，検索クエリ・SERP・閲覧したウェブページ・滞在時間・強調された単語・仮想他者のクエリ履歴を実験中に収集した．下記に示す指標により RQ2 の検証をした．具体的には，前半 5 分と後半 5 分の変化量を 4 つの表示形式間で比較した．

下記の指標は，前半と後半で None と比べてどの程度変化するかを評価するために使用した．検索クエリ数と閲覧ページ数は，重複は除いて算出している．4 つ目の指標は，前半で検索したクエリの SERP 中で選択したページの順位と後半で検索したクエリの SERP 中で選択したページの順位の変化を評価した．

（1） 検索クエリの数の変化（表 8）

（2） 閲覧ページ数の変化（表 9）

（3） Web ページの 1 ページあたりの滞在時間の変化（表 10）

表 8　検索クエリの数の変化 （個）

表示形式	前半 5 分	後半 5 分	前後半の差	分散
None	5.45	4.73	-0.73	2.05
User	4.56	3.78	-0.78	1.56
Other	4.60	3.90	-0.70	2.36
Both	4.09	3.64	-0.45	1.58

表 9　閲覧ページ数の変化 （個）

表示形式	前半 5 分	後半 5 分	前後半の差	分散
None	7.27	5.64	-1.64	2.62
User	7.00	7.22	0.22	1.92
Other	4.90	5.40	0.50	2.80
Both	5.09	5.18	0.09	1.29

表 10　Web ページの 1 ページあたりの滞在時間の変化（秒）

表示形式	前半 5 分	後半 5 分	前後半の差	分散
None	36.36	37.95	1.59	13.83
User	31.52	33.35	1.83	10.50
Other	46.87	43.54	-3.33	17.83
Both	48.87	46.00	-2.87	17.73

表 11　SERP の選択順位の変化（位）

表示形式	前半 5 分	後半 5 分	前後半の差	分散
None	2.42	2.49	0.07	2.34
User	2.91	3.12	0.21	1.03
Other	3.12	2.11	-1.01	1.49
Both	2.79	3.19	0.40	2.72

表 12　ハイライトされた単語の参考度合い

表示形式	1	2	3	4	5	平均	分散
User	3	5	0	0	1	1.64	1.22
Other	0	3	4	2	1	3.10	0.99
Both	1	6	0	3	1	2.73	1.27

（4） SERP の選択順位の変化（表 11）

また，検索タスク後のハイライトされた単語の参考度合いを 5 段階リッカート尺度の項目とハイライトされた単語の感想の自由記述の項目についてハイライトされた単語の考察として使用する．

4.4.1　実験 3 の結果と考察

ここでは，RQ2「他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか」の検証結果について述べる．表 8 から表 12 は，4.4 章で述べた指標において，2 つの検索トピックの結果は合算した結果である．

まず，各項目の前半 5 分と後半 5 分の変化量の傾向についての結果と考察を述べる．

表 8 の検索クエリ数の変化は，すべての表示形式に共通して

若干の減少傾向であった．この結果から，強調表示による検索クエリ数の変化の影響はないと考える．

表 9 の閲覧ページ数の変化は，一般的な表示形式と比べてユーザ観点のみと仮想他者観点のみ，両方が若干の増加傾向であった．この結果から，観点提示の 3 種類の UI が多くのページの閲覧を促進する可能性があることを示唆している．

表 10 の Web ページの 1 ページあたりの滞在時間の変化は，一般的な表示形式と比べて仮想他者観点のみの場合に，5 秒程度減少している．この結果から，仮想他者のみの観点提示による効果として，そのページにおけるキーワードがハイライトされることにより時間短縮につながることを示唆できる．

表 11 の SERP の選択順位の変化は，一般的な表示形式と比べて仮想他者観点のみの場合に，1 位程度順位が高くなっている．この結果から，仮想他者のみの観点提示による効果として，上位の検索結果の選択を促進させると示唆できる．

次に，ハイライトされた単語の評価について述べる．表 12 は，観点提示ありのユーザ観点のみ/仮想他者観点のみ/両方の被験者に回答してもらった 5 段階リッカート尺度の結果である．具体的には,「ハイライトされた単語はどの程度参考にしましたか」という項目を 1：使いたくない，5：使いたいの結果である．

表 12 から，ユーザ観点のみの場合に 2 以下を回答した被験者が 8 名と参考にならなかったことを示してる．仮想他者観点のみの場合には，2 以下を回答した被験者が 3 名，4 以上を回答した被験者が 3 名とばらつきがある結果となった．両方の場合には，2 以下を回答した被験者が 7 名，4 以上を回答した被験者が 4 名で若干のばらつきがある結果となった．このようにばらつきがある理由として，ハイライトされた単語の質が考えられる．ユーザ観点のみの場合で，適切にハイライトされるべき単語は，ユーザ自身の検索トピックに合致し，直前のクエリとの関連度が高い単語だと考える．仮想他者観点のみの場合で，適切にハイライトされるべき単語は，自身の検索トピックに合致し，直前のクエリとの関連度が低い単語だと考える．しかし，本実験でハイライトされた単語は，このようになっていないことが原因で回答にばらつきが出たと考える．

以上の検索ログとアンケートの結果から，RQ2「他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなる」は支持できないと考える．なぜなら，一般的な表示形式と比べて仮想他者のハイライトされた単語による効果として，新たな興味の発見に繋がらなかったためである．この原因として，4.3 節で述べた実験 2 の結果が影響していると考える．実験 2 において閲覧ページで強調された単語の分析を行った．具体的には，ユーザ実験と異なる被験者に対し，ある被験者のクエリ履歴とその被験者が最後に閲覧したページでハイライトされた各単語を提示し，アンケートで評価した．その結果，User 判定された単語と Other 判定された単語が同等の数であり，強調表示における観点別の判定が適切にできていなかったことを確認した．これが，強調表示ありの表示形式と一般的な表示形式で検索ログの差異がほとんど無い理由と考える．

5 結　　論

本研究では，Web 検索ユーザの検索視野の多様化のために，ユーザ自身のクエリ履歴と閲覧履歴を用いて他者の検索行動モデルを仮想的に作成し，それによってユーザが閲覧したページ中の単語および SERP のタイトルとスニペット中の単語を他者が興味を持つか判定を行い，判定結果に基づいてその単語を強調表示するシステムの実装を行った．このシステムを用いて下記の 2 つの Research Question（RQ）を実験により検証した．

RQ1　他者の検索行動モデルを仮想的に作成することは可能か

RQ2　他者の観点の強調表示は，ユーザ自身の興味と異なると考えるため，多様なコンテンツを選ぶきっかけとなるか

提案手法は,「対象ユーザ・仮想他者の履歴データ」について対象ユーザをユーザ自身とし，そのユーザのクエリ履歴と閲覧ページおよび SERP のタイトルとスニペットを 1 つにした文書から仮想他者のクエリ履歴を生成する手法にした.「強調対象候補の抽出方法・強調の適用/非適用の決定方法」については，閲覧ページおよび SERP の単語単位で強調表示する方式とした．

提案手法は，以下の処理 1 から処理 4 によって構成されている．

処理 1　収集したクエリ履歴と閲覧したページおよび SERP のタイトルとスニペットを 1 つにした文書により仮想的なユーザ（以下，仮想他者）のクエリ履歴を生成する

処理 2　閲覧したページおよび SERP のタイトルとスニペットを 1 つにした文書中で強調表示対象候補の単語を抽出する

処理 3・4　ユーザおよび仮想他者のクエリ履歴と強調対象候補の単語の類似度判定を行いユーザ/仮想他者/両方の 3 種類の観点で色別の強調表示を行う

RQ1 を検証するための実験 1 では，クラウドソーシングにてユーザ実験で収集した検索ユーザのクエリ履歴と実験中に生成された仮想他者のクエリ履歴，最後に閲覧したページを用いて，ユーザのクエリ履歴と仮想他者のクエリ履歴の評価を行った．アンケートの結果から，RQ1 は部分的に支持できることを明らかにした．

RQ2 を検証するための実験 2 と実験 3 では，クラウドソーシングにて 2 つの実験を実施した．実験 2 において閲覧ページで強調された単語の分析を行った．具体的には，ユーザ実験のある被験者のクエリ履歴とその被験者が最後に閲覧したページでハイライトされた各単語を提示し，アンケートで評価した．その結果，User 判定された単語と Other 判定された単語が同等の数であり，強調表示における観点別の判定が適切にできていなかったことを確認した．これが，強調表示ありの表示形式と一般的な表示形式で検索ログの差異がほとんど無い理由と考える．

実験 3 では，指定した 2 つの検索トピックと 4 つの表示パターンを組み合わせた各 10 名の計 80 名に対して実施した．検索ログとアンケートの結果から，一般的な表示形式と比べて強調表示する形式による効果が確認できなかった．この原因として，4.3 節で述べた実験 2 の結果で，観点別の強調表示が適切

にできていなかったことが考えられる．以上のことから，観点別の判定が適切にできていないため，RQ2 は支持できないと考える．

本研究の今後の課題は以下である．

- 仮想他者のクエリ履歴を生成するためのスコア付けの改善
- 仮想他者のクエリ履歴生成時の候補ワードを抽出した際に，クエリに適していない語を除外する処理の追加
- 閲覧ページおよび SERP 中の判定対象の単語のストップワード処理の改善
- 閲覧ページおよび SERP 中の判定対象の単語抽出の際に希少性や有用性をもとにした重み付けの追加
- ハイライトする単語の数の調整
- 強調対象単語の観点別における判定手法の改善
- 仮想他者の複数生成することによる効果の検証
- 単語単位以外に，フレーズ単位で判定および強調表示を行うことによる効果の検証
- 実ユーザのクエリログとクリックスルー履歴のモデル化とクエリ生成への反映

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Jonathan Foster. Collaborative information seeking and retrieval. Annual Rev. Info. Sci & Technol., Vol. 40, No. 1, pp. 329–356, 2007.

[2] Austin R. Ward and Robert Capra. Orgbox: Supporting cognitive and metacognitive activities during exploratory search. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, p. 2570–2574, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Roberto González-IbáñTez, Verónica Proaño-Ríos, Gary Fuenzalida, and Gonzalo Martinez-Ramirez. Effects of a visual representation of search engine results on performance, user experience and effort. Proceedings of the Association for Information Science and Technology, Vol. 54, No. 1, pp. 128–138, 2017.

[4] Chen He, Luana Micallef, Barış Serim, Tung Vuong, Tuukka Ruotsalo, and Giulio Jacucci. Interactive visual facets to support fluid exploratory search. Journal of Visualization, Vol. 26, No. 1, pp. 211–230, 2023.

[5] Laure Soulier, Chirag Shah, and Lynda Tamine. User-driven system-mediated collaborative information retrieval. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, pp. 485–494, 2014.

[6] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. The future is in the past: Designing for exploratory search. In Proceedings of the 4th Information Interaction in Context Symposium, IIIX '12, p. 52–61, New York, NY, USA, 2012. Association for Computing Machinery.

[7] Salvatore Andolina, Khalil Klouche, Tuukka Ruotsalo, Patrik Floréen, and Giulio Jacucci. Querytogether: Enabling entity-centric exploration in multi-device collaborative search. Information Processing & Management, Vol. 54, No. 6, pp. 1182–1202, 2018.

[8] Hiroki Nakayama, Ryo Onuma, Hayato Takagi, Hiroaki Kaminaga, Youzou Miyadera, and Shoichi Nakamura. Methods for supporting the understanding of differences between search intentions and actual browsing situations in collaborative exploration. In 2016 Eleventh International Conference on Digital Information Management (ICDIM), pp. 209–214, 2016.

[9] Jeremy Pickens, Gene Golovchinsky, Chirag Shah, Pernilla Qvarfordt, and Maribeth Back. Algorithmic mediation for collaborative exploratory search. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 315–322, 2008.

[10] Sungsoo (Ray) Hong, Minhyang (Mia) Suh, Tae Soo Kim, Irina Smoke, Sangwha Sien, Janet Ng, Mark Zachry, and Juho Kim. Design for collaborative information-seeking: Understanding user challenges and deploying collaborative dynamic queries. Proc. ACM Hum.-Comput. Interact., Vol. 3, No. CSCW, pp. 1–24, nov 2019.

[11] Meredith Ringel Morris. A survey of collaborative web search practices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pp. 1657–1660, 2008.

[12] Takehiro Yamamoto, Mitsuo Yamamoto, and Katsumi Tanaka. Analyzing effect of roles on search performance and query formulation in collaborative search. In Proceedings of the 2015 Workshop on Evaluation on Collaborative Information Retrieval and Seeking, ECol '15, pp. 3–6, 2015.

[13] Jeonghyun Kim and Jisu Lee. Graduate students' information seeking in a collaborative learning setting. In Proceedings of the 2012 IConference, iConference '12, pp. 552–554, 2012.

[14] Felipe Moraes, Kilian Grashoff, and Claudia Hauff. On the impact of group size on collaborative search effectiveness. Inf. Retr., Vol. 22, No. 5, pp. 476–498, oct 2019.

[15] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings. In Anna Korhonen and Ivan Titov, editors, Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 221–229, Brussels, Belgium, October 2018. Association for Computational Linguistics.

観点の自動抽出に基づくインタラクティブな観光スポット推薦システム

小林 らんう † 上野 史 †† 太田 学 ††

† 岡山大学大学院環境生命自然科学研究科 〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学学術研究院環境生命自然科学学域 〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †pwoy2qot@s.okayama-u.ac.jp, ††{uwano, ohta}@okayama-u.ac.jp

あらまし 推薦システムの研究では，レビューからユーザの観点に基づく評判情報を抽出し，推薦に利用するものが存在する．しかしながら，観光スポットの推薦においては季節やイベントなどの多数の要因でその魅力が変化することから，事前に定めた共通の観点を利用すると，スポットの特性が反映されず，偏ったスポットを推薦する可能性がある．そこで，本稿では観光レビューからスポット毎に観点を自動抽出し，その中からユーザが選択した観点に適う観光スポットをインタラクティブに推薦するシステムを提案する．提案システムのユーザインタフェースが提示する地図に対し，ユーザは観光予定の中心地点と推薦範囲，観光スポットの選択において着目する観点を指定する．その後インタフェースは選択された観点に関連した観点を含む観光スポットを地図上に表示する．システムの評価実験では，選択観点に対して推薦スポットが有効であるかを考察し，岡山県内の 699 のスポットを推薦候補とする被験者実験を実施した．被験者は推薦スポットやインタフェースなどの満足度について，アンケート形式で 4 段階評価の質問に回答する．

キーワード 観光，推薦システム，観点抽出，fastText

1 はじめに

旅行計画を立てる際，多くの人は観光スポットを選定するために楽天トラベル [1]やじゃらん net[2]等の観光支援サイトを利用する．このとき，観光スポットの豊富なレビューからスポットの情報を収集し，旅行先を選定する方法は一般的である．しかし観光レビューは膨大であり，限られた時間でこれらを読み，効率的に旅行先を選定することは難しい．そこで，従来より観光情報学（Tourism Informatics）として，観光振興を図る事業者や旅行者に対する情報処理技術の適用が研究されてきた．例えば鮫島はプランナーのための観光プランの作成支援システムを提案した [1]．このシステムでは観光プランの事例を記載した記録を収集し，トピックやキーワードを抽出して提示することでプランナーが適切な事例を発見することを支援する．奥らは Twitter（現在は X と呼ばれる）の地理情報（ジオタグ）付きの投稿を利用して，スポットを傾向，イベント，季節，時刻などの属性で特徴づけることで，対象スポット付近での活発な領域を検出する手法を提案した [2]．一方で，本稿では Twitter における投稿情報や，旅行サイトの口コミなどを利用して，旅行者に向けて観光スポットを推薦するシステムに着目する．具体的には，ユーザの観点に基づくスポットの評判情報を Twitter などから抽出し，その観点にマッチするスポットをユーザに推薦し，効率的な旅行計画を支援する．

観点を利用した観光スポットの推薦として，野本らは，じゃらん net の観光レビューを食事，景観，購買，体験，設備，混雑，交通の 7 つの観点で分析し，高認知度スポットに類似した穴場スポットを発見する手法を提案した [3]．また観点を利用した推薦に関連した事例として，杉浦らは京都の観光スポットを推薦するシステム「京のおすすめ」を構築するために，アンケートや面接から収集した 7,000 件以上の評価表現を，上位概念である 137 の評価要因に要約し，さらにそれを「気分」,「体験」,「雰囲気」,「スポットの特徴」という 4 つのカテゴリに分類した [4]．しかしながら，観光スポットの推薦においては，季節やイベントなどの多数の要因でスポットの魅力が変化し，他にないスポット固有の魅力も存在することから，事前に定めた観点を利用すると，特定の観点に偏ったスポットを推薦する可能性がある．例えば「食事」「自然」「歴史」「土産」の 4 つの観点から観光スポットを推薦したとき,「金閣寺」は「歴史」の観点でその特徴を推薦に反映できるが,「乗馬体験」や「スキューバダイビング」をセールスポイントとする観光スポットは，その魅力が適切に反映されなくなる．この場合は，例えば観点「体験」を追加することで，これらのスポットの魅力を推薦に反映することができる．

我々は上記の問題に対する解決策として，観光レビューを用いて，様々な観点をスポット毎に自動で抽出する手法を提案した [5]．本稿では [5] で提案した観点の自動抽出手法を利用して，ユーザに向けてインタラクティブに観光スポットを推薦するシステムを提案する．推薦システムでは，ユーザが観光スポットの選択において着目する観点と，観光予定の地域を指定することで，推薦される観光スポットが地図上に表示される．ユーザは推薦スポットを確認し，インタラクティブに推薦地域や観点を変更することができる．システム内では選択された観点やスポットの人気度の情報を利用して，スポットごとに推薦スコアを算出し，推薦スコアが高い 10 スポットを推薦する．本稿では，提案システムが推薦したスポットと選択した観点との関連

1：https://travel.rakuten.co.jp/mytrip/
2：https://www.jalan.net/

性について考察し，推薦されたスポットや，インタフェースの満足度を被験者実験により検証する．

本稿の構成は以下の通りである．2 節では関連研究を述べ，3 節では観光スポットの観点の自動抽出手法を述べる．4 節では本稿で提案する観光スポット推薦システムについて述べる．5 節では提案する観光スポット推薦システムの評価実験を行い，結果と考察を述べる．6 節ではまとめと今後の課題について述べる．

2 関 連 研 究

様々な分野において，ユーザのレビューには多くの情報が記載されており，近年はレビューを用いた研究が多い．金子らは美容や健康，医薬品などに関係する商品のレビューから，CaboCha[3]を用いて係受け表現を抽出し，商品の購買意図の構造化を行い，その実用性を調査した [6]．市村は，ユーザが入力した文と関連のある料理店の情報を表示する Web システム「食探」を開発した [7]．市村は Doc2Vec[4]を用いてユーザが入力した文と類似したレビュー文を探し，日本語係り受け解析器の CaboCha を用いてレビュー文を短い係り受け文に要約した．その要約文の中から料理に関係した文を選別しインタフェースに表示することで，ユーザが手早く料理店の情報を把握できるようにした．小川は，映画「劇場版「鬼滅の刃」無限列車編」と「君の名は。」の 2 つの作品のレビューを用いて，作品の評価に影響を及ぼす要因を分析した [8]．

観光分野においても，レビューや旅行ブログを用いた分析が行われており，この際特定の単語や特徴をレビューから抽出する手法がしばしば用いられる．野本らは，Natural Language API[5]のエンティティ感情極性分析を用いて，観光レビュー文からエンティティとともにそのエンティティの入力テキストにおける感情極性値を抽出した [3]．本稿では野本らが用いた Natural Language API の中のエンティティ分析を用いて観光レビューからエンティティを抽出する．上原らは MeCab[6] や TermExtract[7]，独自に作成した辞書を用いて Yahoo!知恵袋 [8] と Yahoo!ブログ [9] から観光地名の共起キーワードを抽出した [9]．

また，レビューの類似性や，スポットの特徴分析のために，レビューや抽出した語をクラスタリングすることがある．市村らは類似した観光レビューをグルーピングするため，非階層的クラスタリング手法である k-means 法を用いた [10]．田村らは AppStore の類似したアプリケーションをまとめるために，アプリケーションの機能説明文から名詞を取り出し，x-means 法を用いてクラスタリングした [11]．また淀川らはクラスタリングによって生成された文書クラスタにラベルを自動で付与することを試みた [12]．彼らは Support Vector Machine (SVM) を用いてニュースコーパスの文書をクラスタリングし，さらに fastText[10]を用いて各クラスタのラベルの候補となる単語を抽出し分析した．

このような分析結果を応用した，観光スポットの推薦システムの開発も活発に行われている．杉浦らは，観光スポットに関する情報をアンケートや面接によって収集し，それらの情報を 137 の評価要因に人手で分類し，さらに評価要因を「気分」,「体験」,「雰囲気」,「スポットの特徴」の 4 つのカテゴリに分類した [4]．これらの分類結果を利用し，彼らは観光スポット推薦システム「京のおすすめ」を開発した．市村らは観光レビューを利用して観光スポット推薦システム「旅ゲーター」を開発した [10]．彼らは観光レビューを Doc2Vec や k-means 法を用いてクラスタリングし，各クラスタから観光スポットの特徴を示す重要語を TF-IDF 法を用いて抽出した．その重要語を「旅ゲーター」のインタフェースに表示しユーザに選ばせることで，ユーザの好みを反映した観光スポット推薦システムを構築した．住友らは，Web 上に公開されている API や WordNet を活用し，口コミに含まれる感情語を抽出し，類似する感情を示すスポットを，ユーザに推薦するシステムを提案した [13]．長尾らは，観光スポットの情報を Web から収集し，Word2vec やクラスタリング技術を利用し，ユーザの自由文に基づいて観光スポットを推薦するシステムを提案した [14]．Abbasi-Moud らは，口コミからユーザの好みや観光スポットの特徴を抽出し，ユーザの好みと最も類似したスポットを推薦するシステムを提案した [15]．ユーザの好みを抽出する際は SentiWordNet を利用した感情極性分析や，クラスタリング技術を利用している．また，レビュー中の単語の出現頻度をもとに，天候毎に分類してスポットの特徴を抽出している．野本らは自然言語処理モデルの BERT [16] を利用して，観光レビューを事前に定めた 12 の観点に分類し，観点のマッチングに基づいた観光スポット推薦システムを提案した [17]．

3 観光レビューからの観点の自動抽出手法

3.1 概　　要

本稿で提案するインタラクティブな観光スポット推薦システムでは，観光レビューから自動で観点を抽出し，それを利用して観光スポットを推薦する．そのため，まずスポット毎に観点を抽出する．本節では同システムに導入する我々が文献 [5] にて提案した観点の自動抽出手法について説明する．

図 1 に観点の自動抽出手法のフローを示す．まず任意の都道府県を一つ選択し，じゃらん net などの旅行サイトのページから，県内の観光スポットのレビューを収集する．次にそれらのレビューから，エンティティを抽出し，クラスタリングする．エンティティのほとんどは人物名，地名などの固有名詞やレストラン，野菜などの普通名詞である．次にエンティティのクラスタのラベル候補リストを作成し，その中から 1 つラベルを選

3：https://taku910.github.io/cabocha/
4：https://radimrehurek.com/gensim/models/doc2vec.html
5：https://cloud.google.com/natural-language?hl=ja
6：http://taku910.github.io/mecab/
7：http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
8：https://chiebukuro.yahoo.co.jp/
9：Yahoo!ブログは 2019 年にサービスを終了した
10：https://fasttext.cc/

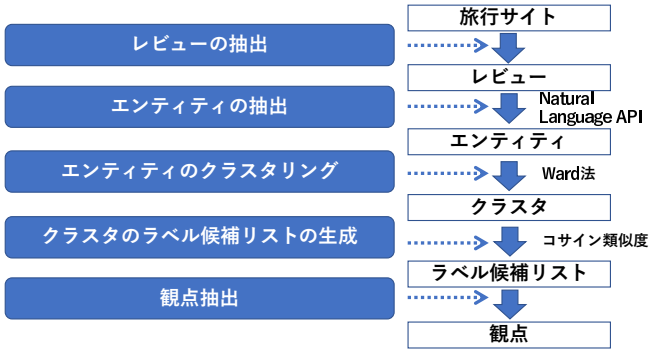


図 1　観点の自動抽出手法の手順

択し，観点として抽出する．

3.2　レビューの取得およびエンティティの抽出

まず，じゃらん net などの旅行サイトから Web スクレイピングを用いて観光スポット毎に観光レビューを収集する．Web スクレイピングには Python のライブラリである BeautifulSoup を利用した．続いて，Google Cloud Platform (GCP) が提供する Natural Language API を用いて，収集したレビューからエンティティを抽出する．この API は感情分析や構文解析，エンティティ分析などの自然言語処理のタスクが行える．本稿では，エンティティ分析 [11] タスクにより観光レビューからエンティティを抽出し，そのエンティティの salience(重要度) の情報を抽出する．salience は [0.0,1.0] の値をとり，1.0 に近いほどエンティティがレビュー中で重要であることを示す．さらに，1 つのレビューから抽出されるエンティティの salience をすべて足し合わせると 1.0 になる．一般に抽出されるエンティティの数はレビューが長いと増えるため，平均的な salience の値はレビュー毎に異なる．そのため本稿では，レビュー毎にエンティティを抽出せず，スポット毎に全レビューを 1 つに連結したものからエンティティを抽出する．また，ノイズとなるエンティティを除去するために，スポット毎に抽出されたエンティティの内 salience が上位 30%のエンティティのみ抽出する．同じエンティティは複数回が抽出されることがあるが，そのいずれかの salience が上位 30%に入っていなければそのエンティティは抽出しない．一方，そのように抽出されたエンティティに重複がある場合は 1 つを残して削除する．

3.3　エンティティのクラスタリング

fastText を用いて 3.2 節の方法で抽出したエンティティの分散表現を獲得する．本稿では，fastText を利用して Common Crawl[12] と Wikipedia[13] で事前学習済みの単語ベクトル [14] を得る．この単語ベクトルは 300 次元，語彙数は 2,000,000 である．単語ベクトルに語彙が存在しない場合は，形態素解析器の MeCab を用いてエンティティをいくつかの単語 (トークン) に

11：https://cloud.google.com/natural-language/docs/analyzing-entities?hl=ja
12：https://commoncrawl.org/
13：https://www.wikipedia.org/
14：https://fasttext.cc/docs/en/crawl-vectors.html

こと，事，もの，物，観光，市，県，市内，県内，すべて，全て，ところ，所，場所，旅行，何，ある，いる，東，西，南，北，右，左，左右，上，下，東西南北，東側，西側，南側，北側，側，前，右側，左側，前側，後ろ側，後ろ，後，途中，近く，遠く，そこ，それ，ここ，どこ，あれ，これ，名前，さん，とき，地域，方面，方，一緒，客，思う，他，スポット，感じ，考え，中，うち，最初，最後，明日，昨日，私，君，僕，あなた，次回，方々，今回，今度，大体，ほとんど，多く，たち，達，様子，魅力，以外，感想，説明，注意，前方，後方 (計 89 語)

図 2　ラベル候補のストップワード

分解する．例えば「家族連れ」は，「家族」と「連れ」の 2 つのトークンに分解される．これらを形態素解析して各トークンの分散表現を足し合わせて元のエンティティの分散表現とする．このようにしても分散表現を獲得できないエンティティは削除する．

次にエンティティの分散表現をクラスタリングする．クラスタリングには，階層的なクラスタリング手法の一つである Ward 法 [18] を用いる．Ward 法は，最初個々のデータをすべて異なるクラスタとし，距離が近いクラスタを順にまとめて，クラスタを階層的に統合させるクラスタリング手法である．クラスタ数の決定には，距離を基準にするものや，不整合係数を基準にするものなど様々な方法があるが，本稿では不整合係数を基準にクラスタ数を決定する．不整合係数は，クラスタを統合したときのリンクの不整合度合いを表す値で，この値が大きいほど似てないクラスタが統合されていることになる．

不整合係数を基準にクラスタ数を決定するために不整合係数の閾値を設定する．あるリンクとその下の全階層のリンクの不整合係数が閾値未満であればそのリンクと下の全階層を同一クラスタとする．本稿では不整合係数の閾値を，全リンクの不整合係数の最大値 × 0.99 に設定する．

3.4　エンティティクラスタからの観点抽出

3.3 節で生成したクラスタから，クラスタのラベルになり得るラベル候補リストを生成し，その中から観点を一つ選択する．ラベル候補リストは，クラスタベクトルとコサイン類似度が大きい 3 つを単語ベクトルの語彙から選択したリストである．クラスタベクトルはそのクラスタに属するエンティティの分散表現の平均値とする．またレビューでの言及が少ない一部の意見は，スポットの特徴を把握するのにノイズとなる可能性が高い．ここで，本稿ではエンティティ数が 3 未満のクラスタからは観点を抽出しない．さらに「の」や「も」などの助詞や，「岡山」や「倉敷」などの地名そのものも，ラベルの候補としてふさわしくない．そのため本稿では形態素解析器の MeCab と Juman++[15] を用いて，助詞や地名となる単語をラベル候補リストに含めない．また，出力するすべてのラベル候補はストップワード (図 2) に含まれないことが前提である．

ラベル候補リストの中には接尾語が含まれることがある．そこで本稿では接尾語から複合語を生成し，ラベル候補リストを

15：https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B

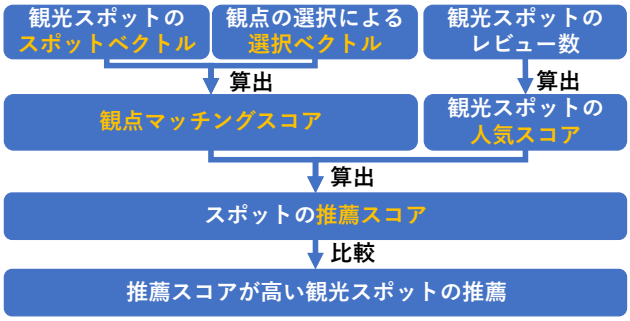


図 3　観光スポット推薦の概要

更新する．例えばラベル候補リストが「駐車，駐輪，場」である場合,「駐車場，駐輪場」に更新する．しかし,「動物犬」や「愛称声」のように存在しない複合語の生成を防ぐ必要がある．ここで本稿では，生成した複合語と一致する Wikipedia の記事が存在しない語をラベル候補リストから削除する．ただし，Wikipedia で名前が一致する記事が存在しない場合でも，検索結果の 1 ページ (最大 20 記事) 目のスニペットに生成した複合語が 10 件以上含まれていれば，その語は日本語の語彙にあるとみなし，削除しない．

各ラベル候補の分散表現と，元のクラスタに含まれるエンティティの分散表現とのコサイン類似度の平均値が最も高いラベル候補を観点として抽出する．

4　インタラクティブな観光スポット推薦システム

本節では 3 節の手法で抽出した観点を利用し，ユーザが観光スポットを選ぶ際に着目する観点を自由に選択できる，インタラクティブな観光スポット推薦システムを提案する．推薦システムはフロントエンドに JavaScript を，バックエンドに Python を使用し，Web アプリケーションとして実装した．

4.1　観光スポット推薦システムの概要

図 3 はユーザに観光スポットを推薦する処理の概要である．まず観光レビューから抽出した観点とその出現頻度を示したスポットベクトルを算出し，インタフェースでユーザが選択した観点を示す選択ベクトルを算出する．次にスポットベクトルと選択ベクトルから算出した観点マッチングスコアと，スポットの人気スコアからスポットの推薦スコアを算出し，推薦スコアが高い観光スポットをインタフェース上でユーザに表示して推薦する．また，ユーザは表示されたスポットが自身の好みに合致しているかを判断して観点を選び直し，システムはその観点に応じて推薦する観光スポットを変更する．このようにユーザがインタラクティブに観点を変更できることから，推薦スポットをよりユーザの要望に近づけることができる．また，スポットから自動抽出した観点を利用するため，推薦にスポット固有の魅力が反映できる．

4.2　スポットベクトルおよび選択ベクトル

図 4 にスポットベクトルおよび選択ベクトルの例を示す．まず推薦候補となるスポットの観点を 3 節に示す方法で抽出し，

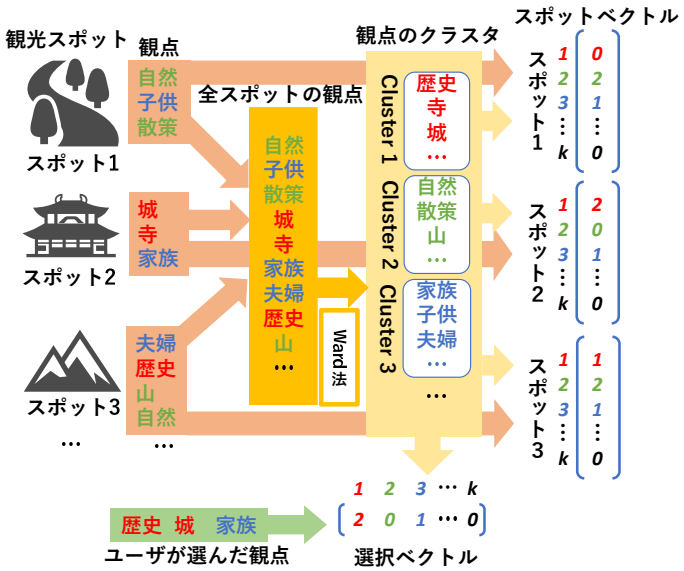


図 4　スポットベクトルおよびユーザの選択ベクトル

抽出された観点を一つにまとめ，それを Ward 法を用いてクラスタリングして観点クラスタを生成する．次にこの観点クラスタに含まれる観点をスポット毎に数え，スポットベクトルを生成する．例えば，図 4 のスポット 1 は「自然」「散策」が Cluster 2 に含まれるため，スポットベクトルの 2 番目の要素を 2 とし,「子供」が Cluster 3 に含まれるため，スポットベクトルの 3 番目の要素を 1 とし，他の要素を 0 とする．このようにして k 次元のスポットベクトルを生成する．

選択ベクトルは，推薦システムを利用するユーザが選択した観点のリストから生成される．これはスポットベクトルと同様に k 個の観点クラスタに含まれる選択観点を数えて生成するため，k 次元のベクトルとなる．例えば図 4 の場合，ユーザが選んだ「歴史」と「城」は Cluster 1 に,「家族」は Cluster 3 に含まれるため，選択ベクトルの 1 番目の要素が 2，3 番目の要素が 1 となり，他の要素は 0 となる．

4.3　観光スポットの推薦スコア

観点マッチングスコアは，スポットの観点とユーザが選択した観点の類似度を示すスコアであり，スポットベクトルと選択ベクトルのコサイン類似度と定義する．すなわち，スポット N のスポットベクトルを $\boldsymbol{N}$，選択ベクトルを $\boldsymbol{S}$ としたとき，スポット N の観点マッチングスコア $\mathrm{MatchScore}_{NS}$ は式 (1) で計算する．このとき，$\boldsymbol{N}$ と $\boldsymbol{S}$ の i 番目の成分を n_i，s_i で表し，k は $\boldsymbol{N}$ と $\boldsymbol{S}$ の次元数である．

$$\mathrm{MatchScore}_{NS} = \frac{\sum_{i=1}^{k} n_i s_i}{\sqrt{\sum_{i=1}^{k} n_i^2}\sqrt{\sum_{i=1}^{k} s_i^2}} \qquad (1)$$

人気スコアはスポットの人気度合を示すスコアであり，スポットに書き込まれたレビュー数から計算する．本稿では、野本らの分析をもとに，レビュー数を 0 から 200 まで，10 ずつ増加する階級幅で分割し，さらに 200 以上のレビュー数を含む階級を設定した．人気スコアを階級に分割する理由は，人気が高いスポットが極端に推薦され易くなることを防ぐためであ

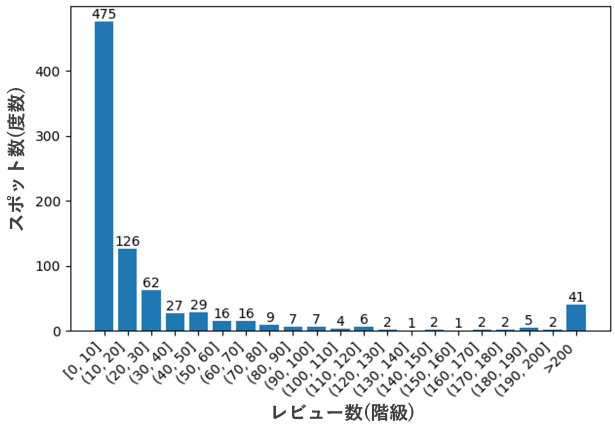


図 5　推薦候補の観光スポットのレビュー数の度数分布

る．スポット N の人気スコア $\mathrm{Popularity}_N$ は，スポットが属する階級から算出し，式 (2) で計算される．このとき Bin_N はスポット N が属する階級が $[0, 10]$ の階級から数えて何番目の階級であるかを示す値であり，NumOfBins は階級の数，すなわち 21 である．例えばレビュー数が 45 のスポット 1 は $(40, 50]$ の階級に属するため，$\mathrm{Bin}_1 = 5$ である．また本稿で推薦候補とするじゃらん net から取得した岡山の観光スポットのレビュー (2023 年 11 月 27 日時点) の度数分布を図 5 に示す．

$$\mathrm{Popularity}_N = \frac{\mathrm{Bin}_N}{\mathrm{NumOfBins}} \tag{2}$$

スポットの推薦スコアは，観点マッチングスコアと人気スコアの積で算出する．すなわち，スポット N の推薦スコア $\mathrm{RecScore}_N$ は式 (3) で計算する．

$$\mathrm{RecScore}_N = \mathrm{MatchScore}_{NS} \times \mathrm{Popularity}_N \tag{3}$$

4.4　ユーザインタフェース

提案システムのユーザインタフェースを図 6 に示す．システムはパラメータの調整部と，推薦結果の表示部に分かれており，パラメータの調整部には検索部と観点の選択部，推薦範囲の調整部があり，推薦結果の表示部には推薦スポットを表示する地図とスポットの詳細が確認できるスポットの情報表示部がある．スポットの情報表示部ではスポットの観点を確認できる．また，リセットボタンを押すと推薦条件をリセットすることができる．

検索部でユーザは旅行で重要視する項目やスポットの特徴等をクエリとして入力する．システムは形態素解析や fastText を利用し，ユーザのクエリの分散表現を算出し，コサイン類似度が高い 10 観点をインタフェースに表示する．また入力するクエリが思いつかない場合は「オススメの観点」のボタンを押すことで，10 か所以上のスポットに含まれる観点がランダムに 10 個表示される．ユーザは観点の検索を繰り返し，選択部に推薦に反映させる観点を 1 つ以上追加する必要がある．また，推薦範囲の調整部で，観光予定中心地から推薦されるスポットの距離の上限 (以下,「推薦範囲」と言う) を変更する．観点や推薦範囲を選択したユーザは，地図中で観光予定の中心地をクリックし，推薦を受ける．

図 6 の地図上の緑色のピンがユーザが入力した中心地で，青色のピンが推薦されたスポットである．また薄い赤色の円が推薦エリアを示している．ユーザは地図中のピンを押すことでスポット名を確認することができ，ピン上のスポット名を押すと，そのスポットのじゃらん net のページに移動することができる．また，図 6 の右側のスポットの情報表示部で推薦されたスポットの観点の一覧を確認できる．ユーザが選択した観点と関連がある観点を赤色で表示している．

ユーザは推薦された結果が自身の好みに合致しているかを判断して,「観点を修正」ボタンで観点を選び直し，異なる観点に対してインタラクティブに新しい推薦を受けることができる．また,「距離を修正」ボタンで推薦エリアを観光予定スポット近辺に狭めることで，行きたいスポットの周りのスポットを効率よく探索できる．このようにユーザは自由に観点をや推薦エリアをカスタマイズして，推薦スポットをよりユーザの要望に近づけていく．

5　観光スポット推薦システムの評価実験

5.1　推薦候補となるスポット

本実験で推薦する観光スポットは，じゃらん net にある岡山県内の 3,634 の観光スポット (2023 年 11 月 27 日時点) のうち，少なくともレビューが 3 件以上存在する 842 スポットである．また 842 スポットの内 143 スポットは 3.4 節のラベル候補の選択条件により，観点を 1 つも抽出できなかったため，699 スポットが最終的な推薦候補となる．推薦候補となるスポットのレビューは合計 39,095 件あり，その内最もレビューが最も多いスポットは「倉敷美観地区」で，レビューは 3230 件である．

5.2　実験 1: 選択観点と推薦スポットの関連性に関する実験

5.2.1　実 験 概 要

本稿で提案する観光スポット推薦システムが推薦するスポットの有効性を図るために，選択観点に対して推薦されたスポットの関連性について考察する．具体的には，複数の観点と単一の観点を選択したときや，少ないスポットにしか当てはまらない固有の観点を選択したとき，選択観点が多いときの推薦スポットや関連観点，推薦スコアについてを考察する．

5.2.2　実 験 結 果

単一の観点と複数の観点による推薦スポットの例を表 1 と表 2 に示す．表 1 から，単一の観点「景色」によるスポット推薦では，おおよそ景色に関連する観点を含むスポットが推薦されており，実際に景色に関連しているスポットとなっている．例えば「瀬戸大橋」は海の景色が綺麗なスポットであり,「美星天文台」は星空が綺麗なスポットである．しかし「鷲羽山ハイランド」の「スリル感」のように「景色」に関連しているとは言えない観点を含むものも存在した．また表 2 では,「景色」に加え「子ども」を選択することで「子ども」に関連する観点を含むスポットが，上位の推薦スポットに入っている．例えば「ドイツの森」では「子供」や「テント」,「ボート」が「子ど

図 6　観光スポット推薦システムのインタフェース

も」の関連観点である．このスポットではゴーカートや，ラジコンボートの操縦などの子供が楽しめるアトラクションが多数存在し，それに加えて自然豊かなスポットであるため「子ども」と「景色」を両方満足するスポットといえる．また追加で選択した観点「子ども」が考慮され，遊園地である「鷲羽山ハイランド」の順位が 5 から 3 へ上昇した．

少ないスポットにのみ当てはまるスポット固有の観点を選択したときの推薦結果を表 3 に示す．選択した観点は「ひな祭り」であり，その関連観点「ひな人形」を含む「勝山町並み保存地区」や「岡山後楽園」が推薦された．両スポットではひな祭りが行われるため，選択観点に沿ったスポットが推薦されたといえる．しかし関連観点として「手作り」が含まれる「月と花 Japanese tea ceremony salon」は茶道体験スポットで，ひな祭りに関係ないスポットのため，選択観点と関連性が薄いスポットである．また,「勝山町並み保存地区」と「岡山後楽園」のうち,「岡山後楽園」の方が人気度が高いスポットであるが,「岡山後楽園」の方が低い順位で推薦されていることもわかる．

選択観点が多い場合の推薦結果を表 4 に示す．魚に関連する観点「鯛」,「刺身」,「穴子」,「海水魚」,「サンマ」,「川魚」,「漁師」が選択されている．表 4 から，2 位に魚に関連する観点が含まれた「五味の市」が推薦スポットに含まれていることがわかる．このスポットでは魚介を購入でき，またその場で食べられるという特徴があるため，選択観点に沿った推薦といえる．しかしその他のスポットでは，魚に関連する観点が存在せず，実際に魚とは関連が薄いスポットが推薦されている．例えば 1 位の「農マル園吉備路農園」は果物や野菜で有名でいちご狩り体験ができるスポットであり魚に関連がない．

5.2.3　考　　察

表 5 は，表 1 から表 4 の間で選択した観点が含まれる観点クラスタを示す．表 5 中の太字は，表 1 から表 4 までの間に選択観点に関連した観点として現れた観点を表す．

選択観点「子ども」が含まれるクラスタには，子どもが楽しめるアトラクションや子どもに近い観点が含まれているが，選択観点「景色」が含まれるクラスタは「スリル感」や「音色」など一部景色と関連が薄い観点も含まれている．このことが原因で表 1 や表 2 では「スリル感」や「風味」といった違和感がある観点が関連観点として現れた．

表 3 で推薦された「岡山後楽園」は「勝山町並み保存地区」に比べて人気度が高いが，低い順位で推薦された．これは「岡山後楽園」の観点数が 37 であり,「勝山町並み保存地区」は観点数が 9 であることから，式 (1) による観点マッチングスコアに差が生じたためである．現在の観点マッチングスコアの計算方法は，抽出された観点数によって値が変化するため，計算方法を変更する必要がある．

表 5 から「サンマ」はあまりまとまりがないクラスタに含まれていることがわかる．また,「穴子」や「刺身」,「漁師」,「鯛」が含まれるクラスタには水産業に関する観点の他に,「農園」や「野菜」のように農業に関連する観点も多く含まれている．このように複数の特徴を示す観点が，一つのクラスタに含まれていることが原因で，表 4 のように，選択観点とずれたスポットが推薦される．この問題を解決するには，異なるクラスタリング手法や，観点の分散表現獲得に fastText 以外の自然言語モデルを利用するなどの検討が必要である．

表 1　観点「景色」を選択したときの推薦結果

スポット	選択観点に関連するスポットの観点	Match Score	Popularity	Rec Score
瀬戸大橋	光景 風情 景色	0.522	1.000	0.522
美星天文台	景色 空間	0.554	0.810	0.449
備中国分寺	風情 風景	0.384	1.000	0.384
牛窓オリーブ園	景色	0.353	1.000	0.353
鷲羽山ハイランド	景色 スリル感	0.333	1.000	0.333

表 2 「景色，子ども」を選択したときの推薦結果

スポット	選択観点に関連するスポットの観点	Match Score	Popularity	Rec Score
ドイツの森	テント 景色 ボート 子供	0.431	1.000	0.431
瀬戸大橋	光景 風情 景色	0.369	1.000	0.369
鷲羽山ハイランド	小学生 景色 スリル感	0.353	1.000	0.353
蒜山高原	風味 子供 景色	0.344	1.000	0.344
美星天文台	景色 空間	0.392	0.810	0.317

表 3　観点「ひな祭り」を選択したときの推薦結果

スポット	選択観点に関連するスポットの観点	Match Score	Popularity	Rec Score
勝山町並み保存地区	ひな人形	0.333	0.952	0.317
岡山後楽園	ひな祭り	0.142	1.000	0.142
月と花 Japanese tea ceremony salon	手作り	0.707	0.047	0.033

表 4　「鯛, 刺身, 穴子, 海水魚, サンマ, 川魚, 漁師」を選択したときの推薦結果

スポット	選択観点に関連するスポットの観点	Match Score	Popularity	Rec Score
農マル園芸 吉備路農園	果物 ピザ 野菜 農家 農園	0.776	0.952	0.739
五味の市	漁師 穴子 魚介 サンマ かき	0.641	1.000	0.641
道の駅 風の家	ブドウ 野菜 ヒマワリ	0.435	1.000	0.435
道の駅 彩菜茶屋	果物 野菜	0.749	0.571	0.428
道の駅 久米の里	野菜	0.306	1.000	0.306

表 5　選択観点を含む観点のクラスタ

選択観点	選択観点が含まれる観点のクラスタ
景色	**風景**，情緒，**景色**，**風情**，**光景**，開放感，景観，趣，**風味**，**スリル感**，趣味，**空間**，香り，音色，空気
子ども	**子供**，子ども，水遊び，こども，**小学生**，**テント**，**ボート**，中学生，プール，海水浴，夏休み，滑り台，カヌー，カヤック，タープ
ひな祭り	**ひな祭り**，**ひな人形**，**手作り**
サンマ	かき氷，もみじ，ピカソ，**ヒマワリ**，菜の花，カピバラ，ウサギ，チーズ，**ブドウ**，ミルク，イルカ，サラダ，サンバ，ひよこ，**サンマ**，つつじ，スープ，タルト，イチゴ，あじさい，いちご，アイス，ひまわり，コップ，ハーブ，ガラス，アスパラ，田んぼ，ツツジ，ケーキ，メダカ，ミカン，バケツ，ぶどう，紫陽花
穴子，刺身，漁師，鯛	家業，**農園**，**野菜**，**穴子**，**果物**，**ピザ**，**農家**，**漁師**，**魚介**，**かき**，刺身，職人，鯛，鮮魚，産直，棚田，直売，海鮮，朝市，牡蠣，農民，前菜
海水魚	海水，海水魚，海水面
川魚	川池，川魚

5.3　実験 2: 被験者によるシステムの評価実験

5.3.1　実 験 内 容

本稿で提案する観光スポット推薦システムの使い易さや，スポットの選定に役立つかどうかを評価するために，岡山大学の情報系の学生 13 名 (学部 4 年生:6 名，修士学生:6 名，研究生:1 名) による被験者実験を行う．被験者は，提案システムを実際に使用して，アンケート形式の質問に回答する．被験者は，異なる観点を選択した推薦を 5 回以上受け，各推薦に対して「観点の検索と選択に関する満足度」と「推薦スポットの満足度」に関する 4 段階評価の質問に回答する．また 5 回以上の推薦を終えた被験者は「インタフェースの満足度」と「総合評価」に関する 4 段階評価の質問に回答し，さらに本システムの使用目的や良いと思った点について回答する．各項目の 4 段階評価の質問を表 6 にまとめた．また，使用目的の選択肢は「旅行中に近くのスポットを検索する」と「旅行前の計画でスポットを調べる」の 2 つであり，システムの良い点の選択肢は「複数観点の組み合わせが可能である点」,「レビューを読まずにスポットの特徴がわかる点」,「インタラクティブに推薦が受けられる点」,「斬新な観点が見つけられる点」の 4 つである．4 段階評価は 1 を最低評価，4 を最高評価とする．

5.3.2　実 験 結 果

被験者実験の結果を表 7 に示す．表において，すべての評価値が 3 を超えていることから，低くない満足度が得られたが，「観点の検索と選択に関する満足度」と「推薦スポットの満足度」については他の項目に比べて比較的低い評価となっている．表 8 から表 10 まではそれぞれ観点の検索と選択に関する満足度，推薦スポットの満足度，インタフェースの満足度に関するアンケートの評価結果である．

表 8 の結果から，オススメの観点機能の評価は 3.00 と 3.14 と他の項目に比べて低く，検索機能の評価である 3.42 に比べて，有効ではなかったといえる．また,「検索回数の少なさ」の評価が 3.58 と最も高く，これは少ない回数で望む観点を見つ

表 6　被験者実験における被験者への質問

評価する内容	質問内容
観点の検索と選択に関する満足度	オススメ観点に興味があるか
	オススメ観点が役に立ったか
	観点の検索結果に満足したか
	意外な観点を発見できたか
	検索回数は少なかったか
	検索された観点は分かり易かったか
推薦スポットの満足度	推薦スポットに満足したか
	選択観点とスポットは関連性があるか
	観点からスポットの特徴がわかるか
	選択観点とスポットの観点は関連性があるか
	推薦されたスポットは人気なスポットか
	推薦範囲を調整した時の推薦結果に満足できたか
インタフェースの満足度	検索欄は使い易いか
	文字の大きさは適切か
	推薦範囲の変更はし易いか
	観点の選択や追加はし易いか
	地図の操作やスポットの特徴を確認するのは問題なかったか
	レイアウトは適切だったか
総合評価	今後もシステムを使いたいか

表 7　被験者実験のアンケート結果

評価する内容	評価値 (最大 4.00)
観点の検索と選択に関する満足度	3.14
推薦スポットの満足度	3.02
インタフェースの満足度	3.46
総合評価	3.38

けられたということであるため，検索がし易かったといえる．一方で，検索から意外な観点を見つけることに対しての満足度は 2.31 と低く，ユーザが想像しなかった意外な観点は現れにくかったといえる．また，検索結果の観点の分かり易さは 3.38 と，おおよそ理解できる観点が検索されたことがわかる．

表 9 の結果から，推薦されたスポットの満足度や，選択観点と推薦されたスポットとの関連性，選択観点と関連したスポットの観点 (赤色で表示される観点) に対する評価が 2.80，2.65，2.85 と低いことがわかる．すなわち，被験者が選択した観点に沿ったスポットが推薦されていないといえる．一方で観点からスポットの特徴が把握できるかどうかという項目に対しては 3.29 の評価を示し，観点からスポットの特徴を把握できた被験者が多かった．また，推薦範囲を変更した時の推薦結果の満足度が 3.37 と最も高かったことから，インタラクティブに推薦エリアを変更した際の結果に満足したことがわかる．推薦スポットの人気度に関する評価も 3.06 となり，人気度が高いスポットが推薦されたと感じる被験者は多くなかった．

表 10 の結果から，インタフェースの評価は高いことがわかる．各機能において概ね使い易いという意見だったが,「観点の選択や追加」に関しては 2.92 と 3 を下回る評価で，低い評価となっている．また，レイアウトの評価も他のインタフェースの評価項目に対して 3.31 と比較的低い値となっている．

表 7 の「総合評価」の 4 段階評価の質問は「今後もシステムを使いたいか」という質問 1 つで，表 7 から 3.38 だった．これは評価 1 を 0%，評価 4 を 100%とした時，79.3%の評価となるため，被験者はこのシステムに約 80%満足したことがわかる．このことから，ユーザの観点選択によるインタラクティブな観光スポット推薦システムは，被験者にとって有効であったといえる．

この他に提案システムの使用目的を複数選択で回答してもらった結果,「旅行前の計画でスポットを調べる」目的が 10 件,「旅行中に近くのスポットを検索する」目的が 4 件であった．また，提案システムの良いと思った点についても複数選択のアンケートを実施し,「複数観点の組み合わせが可能である点」が 6 件,「レビューを読まずにスポットの特徴がわかる点」が 8 件,「インタラクティブに推薦が受けられる点」が 4 件,「斬新な観点が見つけられる点」が 6 件となった．

5.3.3　考　　察

被験者の意見のうち,「オススメの観点」機能が「日本らしい観点」や「家族向のけ観点」が多かったが学生である自身は興味を持てなかったという意見があった．このことから，被験者の興味に関係なく多くのスポットに出現する観点をオススメの観点として出力していることが，表 8 のオススメの観点機能の評価が低い原因の一つであるといえる．この問題を解決するためには，事前にユーザの特性を収集し，それに沿った観点を勧めることが必要となる．

一部被験者は「雰囲気」や「意外」のような曖昧な観点に対して「どんな雰囲気か」や「何が意外なのか」といった具体的な情報が必要であると指摘した．この問題を解決するために，観点にポジティブやネガティブなどの極性を付加するなどして，曖昧さを軽減する手法を検討したい．さらに,「時間目」のように辞書に存在しない観点に困惑した被験者もいた．これは 3.4 節で複合語を生成する際に発生した問題である．これらは観点が分かりにくくなる原因であるため，この問題を改善することで表 8 の「観点の分かり易さ」の評価の改善につながる．

表 9 の結果が全体的に低かった原因は，選んだ観点に対して適切なスポットが推薦されなかったことが挙げられる．これは 5.2.3 節と同様に，関連性が薄い観点が同じクラスタにあることが原因である．また，被験者の意見として「複数観点を選択した際，一方に偏る場合がある」という意見があった．このことから，優先したい観点を選択する機能を追加する必要がある．

6　ま　と　め

本稿では我々が [5] で提案した観点の自動抽出手法を利用して観光レビューから観光スポットの観点を自動で抽出し，その観点を用いてユーザに観光スポットを推薦するシステムを提案した．提案システムでは，ユーザが指定した観光エリアにおいて，ユーザが自由に選択した観点に関連する観光スポットを推薦する．また，ユーザは推薦されたスポットを確認し，観点や観光エリアを変更して，何度でも観光スポットの推薦を受ける

表 8　観点の検索と選択に関する満足度

オススメ観点への興味	3.00
オススメ観点の有効性	3.14
観点の検索結果の満足度	3.42
意外な観点の発見し易さ	2.31
検索回数の少なさ	3.58
観点の分かり易さ	3.38

表 9　推薦スポットの満足度

推薦スポットの満足度	2.80
選択観点とスポットの関連性	2.65
観点からのスポットの特徴把握	3.29
選択観点とスポット観点の関連性	2.85
推薦されたスポットの人気度	3.06
推薦範囲変更時の推薦結果の満足度	3.37

表 10　インタフェースの満足度

インタフェース全体	3.31
検索部	3.54
文字の大きさ	3.85
推薦範囲の変更	3.46
観点の選択や追加	2.92
地図やスポットの観点	3.85
レイアウト	3.31

ことができる．

実験 1 ではシステムが推薦するスポットの有効性を図るために，推薦されたスポットやその観点，スコアを確認し，考察した．その結果，選択した観点に関連するスポットは推薦されたが，一部選択した観点に関係ないスポットが推薦されることが分かった．

実験 2 の被験者実験では 13 名の学生に対して，岡山県内の 699 スポットを推薦候補として「観点の検索と選択に関する満足度」と「推薦スポットの満足度」,「インタフェースの満足度」,「総合評価」の 4 項目についてアンケート調査を行った．その結果,「観点の検索と選択に関する満足度」が 4 段階評価で 3.14,「推薦スポットの満足度」が 3.02,「インタフェースの満足度」が 3.46,「総合評価」が 3.38 となり，すべての項目が 3 を超え，提案システムの評価は概ね低くなかったといえる．しかしその中でも「観点の検索と選択に関する満足度」と「推薦スポットの満足度」は他の項目に比べて比較的低い評価であった．

実験 1 と 2 の結果から，観点のクラスタリングに問題があることが確認できた．この問題を解決するために，今後は観点の分散表現の獲得に fastText 以外の自然言語モデルを利用することや，新しいクラスタリング手法を検討する必要がある．また,「意外」や「雰囲気」のような曖昧な観点を示されて困惑する被験者がいたことから，感情極性分析などを利用し，観点の曖昧さを軽減する手法を検討する．この他に，観点マッチングスコアがスポットの観点数によって異なることから，観点マッチングスコアの計算式を見直す必要がある．

文　献

[1] Samejima, M., “Topic Analysis of Case Reports in Tourism Towards Collaborative Tourism Planning Support,” in *Tourism Informatics*, pp. 1–14. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[2] Oku, K., Hattori, F., “Mapping Geotagged Tweets to Tourist Spots Considering Activity Region of Spot,” in *Tourism Informatics*, pp. 15–30. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[3] 野本輝，上野史，太田学，“観光レビュー文を用いた穴場スポットの発見，”第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM 2022) B43–3，2022.

[4] 杉浦孔明，岩橋直人，芳賀麻誉美，堀智織，“観光スポット推薦アプリ「京のおすすめ」を用いた長期実証実験，”観光情報学会誌「観光と情報」，vol. 10，pp. 15–24，2014.

[5] 小林らんう，上野史，太田学，“観光レビューを利用した観光スポットの観点の自動抽出，”第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM 2023) 5b–3–1，2023.

[6] 金子貴美，村上浩司，石野亜耶，“商品レビューからの購買意図の抽出と自動階層化，”人工知能学会全国大会論文集，vol. 30，no. 2P1-13in1，pp. 1–4，一般社団法人 人工知能学会，2016.

[7] 市村哲，“口コミから美味しい料理店を手早く探すシステム，”情報処理学会論文誌，vol. 61，no. 11，pp. 1748–1756，2020.

[8] 小川哲司，“テキストマイニングとネットワーク分析を用いた映画評価の要因分析，”経済経営論集，vol. 29，no. 2，pp. 26–35，2022.

[9] 上原尚，嶋田和孝，遠藤勉，“Web 上に混在する観光情報を活用した観光地推薦システム，”電子情報通信学会技術研究報告書，vol. 112，no. 367，pp. 13–18，2012.

[10] 市村哲，上石萌恵，堀口莉里花，“口コミ解析と好み診断に基づいた旅行先推薦，”情報処理学会 グループウェアとネットワークサービス研究会報告，vol. 2020-GN-110，no. 11，pp. 1–8，2020.

[11] 田村亮介，白石絵里奈，浅沼爽汰，藤田和成，町田翔，白井聡一，延澤志保，“類似アプリ比較のためのレビュー自動分析，”情報処理学会 第 81 回全国大会講演論文集，pp. 125–126，2019.

[12] 淀川翼，加登一成，伊東栄典，“単語の分散表現を用いた文書クラスタのラベル推定，”人工知能学会第 49 回セマンティックウェブとオントロジー研究会 (SIG-SWO)，vol. 49，no. 3，pp. 03-01 – 03-05，2019.

[13] 住友千将，石野拓也，久保洸貴，岳五一，“口コミ情報に含まれる感情語に基づく類似スポット推薦システムの構築と実証実験，”パーソナルコンピュータ利用技術学会論文誌，vol. 14，no. 1，2020.

[14] 長尾将宏，草野大智，園田泰子，本橋洋介，“自由記述文にもとづく観光スポットレコメンドシステムの提案，”研究報告グループウェアとネットワークサービス (GN)，vol. 2020-GN-109，no .14，2020.

[15] Abbasi-Moud, Z., Vahdat-Nejad, H., Sadri, J., “Tourism Recommendation System Based on Semantic Clustering and Sentiment Analysis,” Expert Systems with Applications, vol. 167, 114324, 2021.

[16] Devlin, J., Chang, M., Lee, K. and Toutanovaet,K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, 2018.

[17] 野本輝，太田学，“観点マッチングによる認知度に依存しない観光スポット推薦，”Web インテリジェンスとインタラクション (ARG WI2)，no. 14，2022.

[18] Ward, J. R. Jr., “Hierarchical Grouping to Optimize an Objective Function, ”Journal of the American Statistical Association, vol. 58, no. 301, pp. 236–244, 1963.

大規模言語モデルを用いたその場での要約に基づくレビュー探索インタフェース

藤井真梨乃† 河田 友香†† 山本 岳洋††

† 兵庫県立大学 社会情報科学部 〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学 大学院情報科学研究科 〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †{fa20b075,ad22l019}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし 本研究では，類似する商品レビューについて，大規模言語モデルを用いた要約をユーザに提示するシステムを提案する．たとえばドライヤーをインターネットで購入するときに「髪の毛がつやつやになりました！」というレビューを見て，類似する観点のレビューを探したくなるというように文章を読んで調べたい観点が思い浮かぶことがある．しかし，従来の検索手法である文字列マッチングでは，類義語を用いたレビューを検索することができないうえ，検索キーワードとして適切な単語がないという問題がある．そこで，商品レビューから抜粋した文章を入力とし，分散表現を用いて類似すると判定した商品レビューについて要約を行い，提示するシステムを提案する．大規模言語モデルを用いることにより，ドメインに縛られずに要約を行い，同じ観点のレビューを分かりやすく提示できると考えた．提案手法についてユーザ実験を行い，観点や意見の網羅性や結果の見やすさについてアンケートをとった結果，従来の文字列マッチングよりも回答者平均が上回り，結果の見やすさについては有意差が認められた．また，レビューの検索精度についても評価したところ，提案手法ではフレーズで検索を行ったときに文字列マッチングを用いた手法より F_1 値が高くなった．今後はユーザ実験の人数を増やし，プロンプトの改善を行っていく必要がある．

キーワード レビュー，LLM，インタフェース

1 はじめに

インターネットが発達してきた近年，EC サイトで商品を購入する人が増加している．経済産業省の調査によると，物販系分野の BtoC–EC 市場規模は，2013 年の 5 兆 9,931 億円から 2022 年の 13 兆 9,997 億円まで年々増加している [16]．EC サイトで商品を購入する判断材料の 1 つとして商品レビューがある．商品レビューを読むことで，実際の使用感や耐久性などを詳しく知ることができる．

レビューを読む際，読んでいる最中に調べたいことが思い浮かぶことがある．たとえば，新しくドライヤーを買おうとしているユーザが，気になっているドライヤーのレビューを読んでいる途中に,「可愛い色合い」というフレーズが目に留まったとする．このときユーザは「見た目」について調べると考えられ，ユーザは 2 通りの行動をとる．1 つ目は，レビューを最初から見直すという行動である．この場合，レビューの投稿数が非常に多いときには，ユーザが求めている情報を探すことに時間がかかるという問題がある．そのため，求めている情報が書かれたレビューを絞ることや，求めている情報を要約することが必要である．2 つ目は，キーワードで検索をするという行動である．このようなシステムでは文字列マッチングが一般的に用いられている．しかし,「見た目」と「デザイン」といった，同じ意味を表す単語が含まれているレビューが表示されないという問題がある．また，観点を簡潔な単語で表しにくい場合があるという問題もある．たとえば，ドライヤーにおいて「髪の毛がつやつやになりました！」と「(髪が) しっとりと落ち着いたのにビックリしました」という 2 つのレビューは,「髪質」のように同じ観点で書いてあるように見えるが，実際に文字列マッチングで検索するには限度がある．そのため，類似語や類似文でも検索結果に表示されるシステムが必要である．

そこで本研究では，ユーザがレビューを閲覧している途中で気になる観点や文章を抜粋し，分散表現を用いて抜粋した文章と類似する商品レビューを絞り，それらのレビューについて大規模言語モデルを用いた要約を行って提示するインタフェースを提案する．大規模言語モデルを用いることにより，データ量やパラメータ数が多いという特性からドメインを限定せずに要約を行うことが可能になるのではないかと考えた．このインタフェースを用いることで，ユーザは類似文を検索することや，検索結果で内容を一目見て把握することが可能になる．

提案手法において，従来の検索手法と比較して「要約の見やすさ及び精度」,「ユーザの興味に対する網羅性」,「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」,「システムの使用難易度」,「全体的な満足度」の 7 項目にどのような差があるのかを把握するためにユーザ実験を行った．ユーザは 2 種類のドライヤーのレビューを，1 つは比較手法，もう 1 つは提案手法を用いて検索した．ユーザへのアンケートを用いて 7 項目をリッカート尺度で測定し，検索クエリの発行数及び検索にかかった時間なども取得した．

実験を行った結果，7 項目のうち「システムの使用難易度」を除いた 6 項目で，回答者平均が比較手法を上回った．しかし，有意差は認められなかった．また，検索クエリの発行数及び検

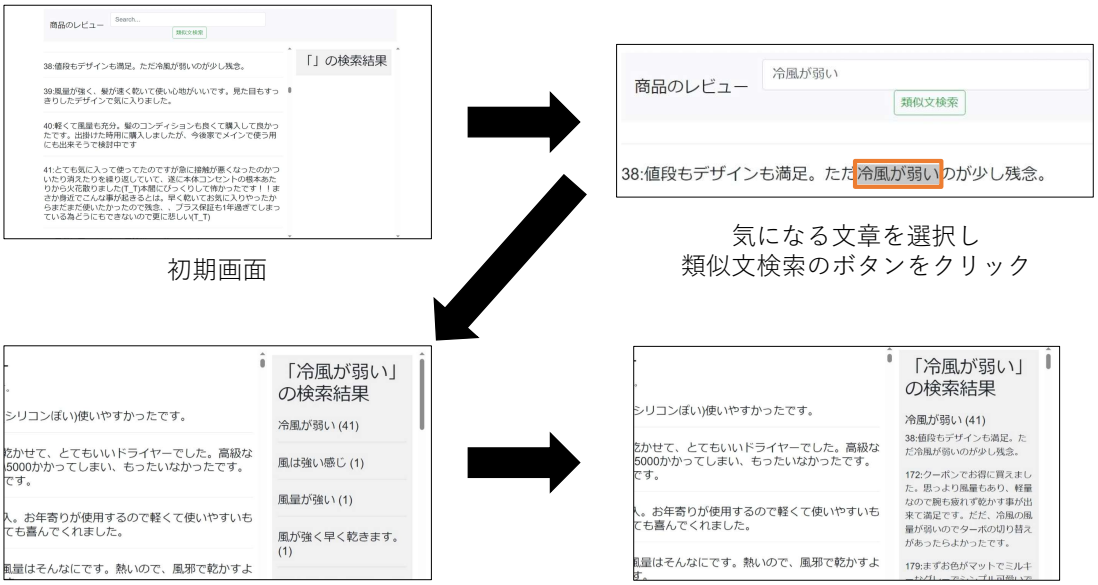


図 1　ユーザ側の利用の流れ.

索にかかった時間についても，有意差は認められなかった.

そして，どのような精度でレビューが検索結果に表示されているかを確かめるために，検索精度の評価も行った．実際の実験で用いられたクエリの中から，名詞 1 つで構成される単語 5 つと，名詞以外の品詞も含まれるフレーズ 5 つを選択して評価を行った．その結果，比較手法では単語での検索の方が高い F_1 値を示し，提案手法ではフレーズでの検索の方が高い F_1 値を示した.

2　関 連 研 究

2.1　商品レビュー分析

商品レビューを分析する研究は数多く存在する．Hu らは，商品の属性を抽出し，その特徴に対して肯定的，否定的な意見を WordNet を用いて収集した [3]．Poria らは，埋め込み表現を用いてレビュー中で述べられている観点を抽出した [10]．Jerripothula らは，商品レビュー及びそれに対する投票と商品の機能レベルへの点数から，商品の機能に対する評価点数を算出した [5]．Deng らは，製品に関する質問に対し，回答生成タスクと意見マイニングタスクをマルチタスク学習を行うことで，高い精度で回答を生成することを可能とした [2]．金兵らは，商品レビューから係り受け解析を用いて評価表現辞書を構築して特徴語を収集し，クラスタリングを行うことで評価軸の構築を行った [14]．渡辺らは，Word2Box を用いて出現単語間の領域の重なりから概念構造における上位語を選択し，教師なし学習で観点の付与を行った [19]．Nguyen らは，200 字以内の短いレビュー文の要約を行う際に，長いレビュー文が投稿されるレビューサイトを用いたアルゴリズムを作成した [9]．ホンらは，商品説明を用いて，レビュー文から教師なしでキーフレーズを抽出した [13]．Angelidis らは，弱い教師あり学習を行った観点抽出器と感情予測器を用いて商品レビューの要約を行った [1]．Liu らは，商品レビューの要約に ChatGPT を用いたとき，人手では他の手法と比較して高い評価を受けたことを報告した [8].

これらの研究では，レビュー中の観点の抽出やレビューの要約を行っているが，本研究では商品レビューを検索して要約を行うインタフェースという部分に焦点を当てる.

2.2　レビューを検索するシステム

商品レビューを検索するシステムについて研究が行われている．Liu らは，商品に対する肯定的，否定的な意見から属性を抽出し，その割合を可視化するシステムを提案した [7]．杉木らは，商品レビュー検索サイトにおいて自然言語表現で入力されたクエリと商品レビューに対し，感性表現シソーラスを用いて（項目，値）の組を抽出して類似度を求める検索方式を提案した [18]．栗原らは，映画レビューサイトに投稿されているレビュー文を，Doc2Vec を用いて映画ごとにベクトル化を行うことで，検索クエリから類似する評価表現を含むレビューを発見することができることを報告した [15].

また，検索を行うインタフェースについての研究も行われている．平山らは，商品レビューに形態素解析を用いて語の共起関係を抽出し，評価属性ごとの記述および関係性を可視化したインタフェースを作成することで，利用したユーザが商品購入に対して有益な文章を従来より多く発見できることを明らかにした [20]．市村は，係り受け解析を用いて飲食店のレビューから料理に関する文の要約を 3 種類行い，インタフェースを作

成して提示したところ，料理に関する抽出を行った 1 文をそのまま載せたシステムが最も満足度が高いことを報告した [17]. Jasim らは，肯定的，中立，否定的なレビューの訪問率を可視化してバイアスを減らすインタフェースを提案し，ユーザが自信を持った意思決定をより行いやすくなったことを報告した [4].

本研究は，市村 [17] のように要約を提示するインタフェースを作成するが，要約に大規模言語モデルを利用する部分が異なる．なお本研究では，先行研究 [15] [17] と同様に商品レビューに対してベクトル化を行い，類似度を求める検索方式を採用している．

2.3 大規模言語モデルを用いた検索システム

最近では，大規模言語モデルを用いた検索システムについての研究が行われている．Spatharioti らは，商品の特徴量の検索に対し，大規模言語モデルである ChatGPT を用いた対話型の検索インタフェースと従来のキーワード検索インタフェースを用いて比較した [12]．Jianqi は，文献検索において，自然言語で書かれた質問から大規模言語モデルを用いて検索クエリとなり得る単語を出力し，キーワード検索を行うことが有効であると報告した [6]．Pride らはそれに加え，質問に対する回答も大規模言語モデルで生成した [11].

要約結果を提示する検索インタフェースについての研究は存在するものの，商品レビューに対し，大規模言語モデルを用いてリアルタイムで要約を行っている検索インタフェースはまだない．そのため，本研究では大規模言語モデルを用いた要約を提示する部分に着目する．

3 その場での要約に基づくレビュー探索インタフェース

本節では，まず提案インタフェースにおけるユーザ側の利用手順を述べる．次に，提案インタフェースの概要を述べ，その後レビューの要約という段階について詳細に述べる．最後に，実際のシステムの動作例について述べる．以下では，1 ユーザが書いたレビュー文章全体を「レビュー」，その中の 1 文を「レビュー文」と定義する．

3.1 ユーザ側の利用手順

図 1 にユーザが実際にシステムを使用する時の流れを示す. まず，レビューが表示されたウェブサイト上で，ユーザが調べたい文章をドラッグして選択する．たとえば，ドライヤーのレビューを見ながら,「冷風が弱い」という文を見て他のレビューが気になったとき，ユーザはその文章を選択する．このとき，ユーザが選択した「冷風が弱い」という文章はウェブサイト上部の検索窓に自動的に入力される．次に，類似文検索ボタンを押すと，クエリである「冷風が弱い」と類似するレビューの要約が表示される．そして，要約された結果をクリックすると要約前のレビューが表示されるようになっている．これは，要約文のみだとレビューが異なる意味で解釈される可能性があるためである．

3.2 提案インタフェースの概要

提案インタフェースの概要を図 2 に示す．まず，あらかじめレビューを 1 文ずつ分散表現に変換する．1 文ごとに類似度を求めることにより，検索クエリにより近いレビュー文が結果に表示できると考えられる.「髪がつやつやになりました．デザインも可愛くて良かったです.」というレビューの場合,「髪がつやつやになりました.」と「デザインも可愛くて良かったです.」のように句点ごとに分割し，OpenAI の embedding 用モデルである text–embedding–ada–002 [1]を用いて各文のベクトル化を行う．このモデルを用いると，レビュー文 1 文を 1,536 次元のベクトルで表現することができる．次に，検索窓にクエリを入力する．このとき，入力するクエリはユーザがレビュー文から引用した文章とする．そして，クエリとレビュー文のコサイン類似度を求め，0.85 以上となる文が含まれているレビュー文を要約対象にする．そこから大規模言語モデルを用いて，対象となるレビュー文を要約する．最後に要約結果をユーザに提示する．

3.3 商品レビューの要約

類似度を求めることで得たレビュー文を，OpenAI の gpt–35–turbo–16k [2]を用いて要約した．このモデルは，プロンプトと呼ばれる自然言語による指示を入力すると，その回答を自然言語で行う大規模言語モデルである．

要約を行う理由は，最初に表示される画面で，レビュー文に書かれている評価を一括で見られるようにするためである．

要約には，few–shot プロンプトを用いる．few–shot プロンプトとは，プロンプトの中で入力と出力の例をいくつか示すことで，回答の精度を上げるプロンプトのことである．プロンプトは以下の通りである．

```
#説明文
以下で入力された文章を，出力形式に従って 10 文字程度で要約してください.
#入力
(類似度が 0.85 以上のレビュー文)
#入力例
(検索クエリが含まれているレビュー文)
#出力例
(検索クエリ)
#出力形式
10 文字程度の要約結果
```

本研究では，入力例として検索クエリの元となったレビュー文，出力例として検索クエリを与える．これにより，ファインチューニングなどで大量のデータを要することがない上，商品のドメインに縛られずに生成を行うことができる．

同一のレビュー文および検索クエリを入力したときに同じ結

1：https://openai.com/blog/new-and-improved-embedding-model
2：https://platform.openai.com/docs/models/gpt-3-5

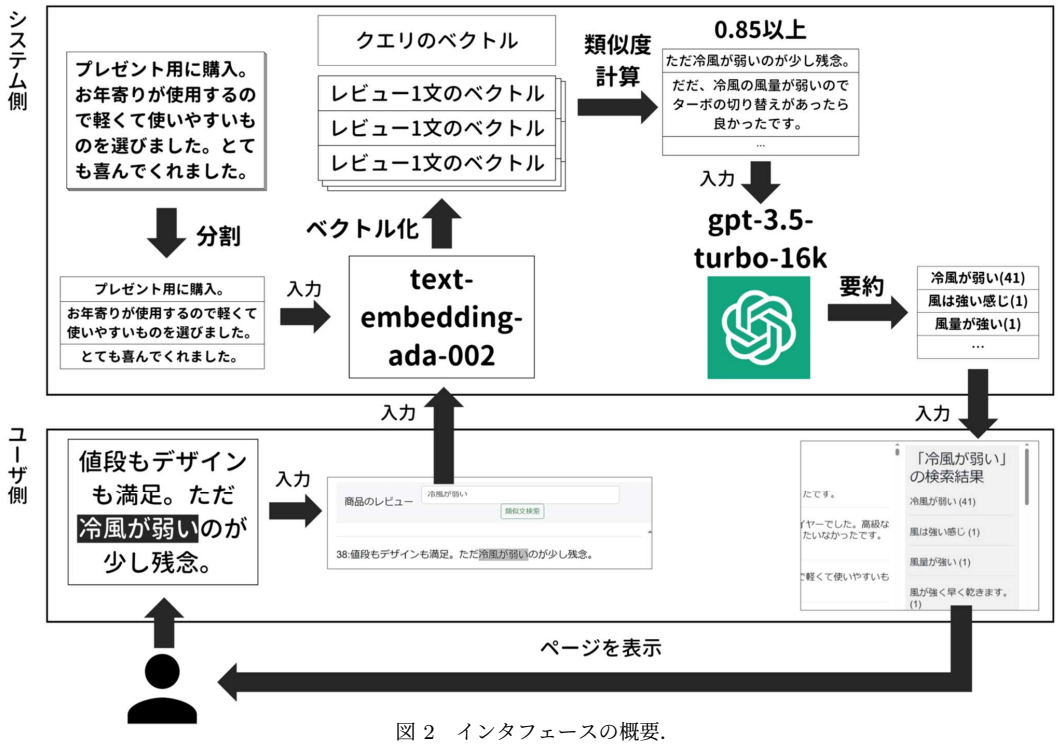


図 2　インタフェースの概要.

果を出力するために，値が高いほど回答が変化しやすくなるパラメータである temperature を 0 とする．

3.4　システムの動作例

本節では，実際のシステムの動作例を述べる．まず,「髪の毛の仕上がりも結構パサつく感じがあって子供のきれいな髪にはいいけど，傷んだ髪質には向かないなって思いました.」というレビュー文から「傷んだ髪質には向かない」というクエリが入力されたとする．このとき,「傷んだ髪質には向かない」との類似度が 0.85 以上になるレビュー文は 5 件存在する．このうち「乾かした時の髪の質ですが，微妙.」というレビュー文の要約を行う場合，以下のプロンプトがシステムに入力される．

```
#説明文
以下で入力された文章を，出力形式に従って 10 文字程度で要約してください.
#入力
乾かした時の髪の質ですが，微妙.
#入力例
髪の毛の仕上がりも結構パサつく感じがあって子供のきれいな髪にはいいけど，傷んだ髪質には向かないなって思いました.
#出力例
傷んだ髪質には向かない
#出力形式
10 文字程度の要約結果
```

このプロンプトが実行され,「髪の質は微妙」という要約が出力される．今回は「傷んだ髪質には向かない」と類似するレビュー文が 5 個存在するために同様の処理が 5 回行われ，出力された要約とその頻度をまとめてユーザに提示する．

4　ユーザ実験

4.1　使用したデータ

本研究では，楽天市場 [3]で収集したドライヤー 2 商品のレビューデータを用いた．ドライヤーを用いた理由は，複数の類義語を持つ観点や，明確に言語化できない観点のレビューが多いと考えたからである．価格は 2023 年 12 月 11 日現在，それぞれ 5,918 円と 6,490 円であり，ほとんど同じ価格帯の商品である．今回は商品レビューの文章のみを使用しており，レビューのタイトルおよび評価点数は利用していない．また，訓練タスク用にオーブントースター 1 商品のレビューデータも使用した．いずれも 2023 年 7 月 26 日時点での最新 200 件を収集した．

4.2　実験参加者

実験参加者は，兵庫県立大学に所属する学生 4 名 (男性 2 名，女性 2 名) である．2024 年 1 月 9 日から 2024 年 1 月 16 日にかけて実験を行い，参加者には約 1,500 円を支払った．

4.3　比較手法

ベースラインとして，文字列マッチングで検索を行い，結果に全文を表示するインタフェースを用いた．検索の流れとしては，検索窓に単語を入力し，文字列マッチングで検索した結果

3：https://www.rakuten.co.jp/

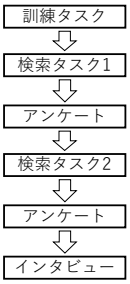


図 3　実験の手順.

表 1　各被験者が検索を行う順番.

被験者	検索タスク 1	検索タスク 2
1, 5	比較手法/商品 1	提案手法/商品 2
2, 6	比較手法/商品 2	提案手法/商品 1
3, 7	提案手法/商品 1	比較手法/商品 2
4, 8	提案手法/商品 2	比較手法/商品 1

を要約せずに表示するというものである．提案手法との違いは以下の 3 点である．

- **検索クエリとして単語を入力すること．**
- **クエリと全く同じ単語が含まれているレビューを検索結果に表示すること．**
- **検索結果に最初に表示されるものがレビュー全文であること．**

4.4　実 験 手 順

実験手順を図 3 に示す．まず，実験参加者に対して検索インタフェースの概要と収集するデータについて説明を行った．収集を行うデータは，検索クエリと検索時間のログ，及びインタビューの音声データである．次に，参加者が検索インタフェースに慣れてもらうために，訓練タスクを行った．訓練タスクでは，オーブントースター 1 商品のレビューに対し，提案手法であるインタフェースを用いたタスクを行ってもらった．具体的には，3 つの検索クエリを提示して検索を行ってもらい，結果に表示されるレビュー全文を Google Form で送信してもらうというものである．また，このタスクが終わった後に比較手法についても説明を行い，2 つの手法それぞれで自由に検索する時間を設けた．訓練タスクが完了した後，検索タスクを行い，ドライヤー 2 商品のレビューを見てどのような評価がされているのかを検索してもらった．検索タスクでは，2 つの商品のうち 1 つを比較手法，もう 1 つを提案手法のインタフェースで検索してもらった．実験参加者に対して検索してもらった商品とインタフェースの順番を表 1 に示す．本実験は，被験者内実験として行った．検索タスクの終了後，アンケートに回答してもらった．アンケートでは，インタフェースに対する満足度や，気になった観点についてくまなく検索を行うことができたかなどを回答してもらった．最後に，アンケートの回答理由を詳細に知るために口頭でインタビューを行い，ユーザ実験を終了した．

表 2　アンケート項目.

評価項目	質問
要約の見やすさ及び精度	レビューに含まれる意見を簡単に確認できたと思う.
ユーザの興味に対する網羅性	自分の調べたいことを調べることができたと思う.
観点の網羅性	様々な観点で調べることができたと思う.
意見の網羅性	同じ観点の中で漏れなく意見を調べることができたと思う.
結果の見やすさ	結果の表示方法が分かりやすかったと思う.
システムの使用難易度	システムの使用は簡単だったと思う.
全体的な満足度	システムの機能に満足している.

4.5　検索タスク

実験参加者には次の文言が書かれた紙を渡し，想定する状況を把握したのちにレビュー検索を行ってもらった．

> あなたは母親に誕生日プレゼントとしてドライヤーをプレゼントすることにしました．値段などを考慮して，2 つの商品に絞りました．
> それぞれの商品についてレビューを読み，どのような観点でどのような評価がされているのか調べて下さい．なるべく多くの観点で調査を行ってください．

制限時間は 1 商品につき 10 分とした．理由としては，1 つの商品あたり 200 件のレビューを用意しており，それを読むためにかかる時間として適切であると判断したためである．そして，充分に検索し終えたと感じたら検索を切り上げても良いこととした．また，自由にメモをとってもらい，実験終了後に回収した．

4.6　アンケート

検索タスクを終了するごとにアンケートを行った．アンケートはそれぞれの商品につき 7 問であり，その項目は表 2 の通りである．全ての問いで 5 段階リッカート尺度を用いた (1：そう思わない，2：あまりそう思わない，3：どちらともいえない，4：ややそう思う，5：そう思う).

4.7　アンケートの分析

アンケートおよびインタビューを対象として，アンケートの分析を行った．まず，アンケートの結果として，手法と設問ごとに回答者の平均と標準偏差を算出した表を以下表 3 に示す．この結果から，システムの使用難易度以外の 6 項目において提案手法が比較手法を上回ったことが分かった．特に 5%有意水準でウェルチの t 検定を行ったところ,「結果の見やすさ」については有意差が認められた．インタビューの中で，提案手法が比較手法を上回っている部分としてあげられることが多かった項目は「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」の 3 つであった.「観点の網羅性」については，比較手法について言及するときに以下のような発話がみられた．

“自分のボキャブラリーにないと検索ができないので，そこがちょっと難しいかなと思ったので，そこは一番やりづらさを感じた部分だったかなと思います．”（参加者 1）

表 3　アンケート結果.

評価項目	比較手法 平均	比較手法 標準偏差	提案手法 平均	提案手法 標準偏差
1. 要約の見やすさ及び精度	4.00	1.31	4.80	0.48
2. ユーザの興味に対する網羅性	3.63	1.19	4.20	0.52
3. 観点の網羅性	3.88	1.55	4.40	0.52
4. 意見の網羅性	2.88	1.55	3.60	1.14
5. 結果の見やすさ	3.25	1.28	4.20	0.84
6. システムの使用難易度	4.63	0.74	4.40	0.71
7. 全体的な満足度	3.38	1.51	4.40	0.67

“どこをドラッグして検索しようか悩んでしまった自分がいました．”（参加者 4）

「意見の網羅性」については，以下のような発話がみられた．

“（提案手法について）結構広い範囲で（結果が）出てくれるので，そこは割といいなと思いました．”（参加者 1）
“1（比較手法）は「風量が大きい」で調べると,「大きい」と書いてあるレビューしか出てこないけど，2（提案手法）はそういうことがなかった．”（参加者 2）
“比較手法だと,「満足度」で調べると「満足度」と書いてあるやつは出てくるけど，大満足なのかどうかというのは，ワードが違うと出てこないから，同じ意見を出すという部分ではいい感じかも．”（参加者 3）

「結果の見やすさ」については，以下のような発話がみられた．

“後半（提案手法）の方は．何件それについて書いてあるかが分かった．”（参加者 2）
“（提案手法について）大量に検索（結果）が出たときに，こっちはめちゃくちゃ見やすいなと思った．”（参加者 3）

また，6 の「システムの使用難易度」のみ比較手法が提案手法を上回った．使用難易度について，以下のような発話がみられた．

“シンプルさっていうのは変わらないかな．”（参加者 1）

表 3 より平均が 4.4 点であることを踏まえると，提案手法でも使用は難しくないといえる．

4.8　クエリの分析

それぞれの参加者のクエリ発行数は表 4 の通りである．クエリ発行数とは，クエリの重複なしで検索ボタンが押された回数を指す．

参加者の中では比較手法のクエリ発行数が多い者が多かったが，手法ごとに有意差はみられなかった．

クエリごとの滞在時間についても分析を行った．クエリごとの滞在時間を，本研究では「クエリの検索ボタンが押されたとき」から「次のクエリの検索ボタンが押されたとき」までとする．そのため，最後から 2 番目までに検索されたクエリを対象として滞在時間の計算を行った．また，検索ボタンのダブルクリックを行っているものや，レビューに関係のない記号について検索しているものは除外した．比較手法と提案手法におけるそれぞれの参加者の平均滞在時間を表 5 に示す．提案手法の方が滞在時間が長い者が多かったが，5%有意水準でウェルチの t 検定を行ったところ有意差はみられなかった．

表 4　クエリ発行数.

参加者	比較手法	提案手法
1	9	8
2	18	22
3	26	14
4	15	6
5	13	10
6	9	13
7	15	9
8	10	8

表 5　クエリごとの平均滞在時間（秒）.

参加者	比較手法	提案手法
1	56.62	79.86
2	33.29	27.95
3	21.20	40.08
4	40.43	105.80
5	48.25	44.78
6	63.38	47.45
7	28.36	76.12
8	48.33	72.71

表 6　選択したクエリ.

単語	使いやすかった
	音があまり気にならない
	こげたような匂い
	軽くていい
	冷風にすると風量が弱くなってしまう
フレーズ	風量
	カラー
	プレゼント
	故障
	温度

5　レビュー検索の精度評価

5.1　概　　要

どの程度正しく検索結果にレビューが表示されるのかを調べるために，システムで用いられる検索の精度を評価した．使用するクエリは，予備実験と本実験で商品 1 のレビュー検索において使用されたクエリ 10 個である．そのうち 5 個を名詞 1 つで構成される単語のクエリ，残りの 5 個を名詞以外の品詞も含まれるフレーズとした．選択したクエリは表 6 の通りである．それぞれのクエリについて，レビューが検索結果に表示されて欲しいと著者が判断したものを真の値とし，適合率，再現率，F_1 値を算出した．

5.2　結　　果

それぞれの手法で単語とフレーズの適合率，再現率，F_1 値の平均を以下の表 7 で示す．なお，提案手法の単語における適合率および F_1 値については,「故障」「温度」クエリの検索結果が 0 件であったため，適合率の値を 0 として平均を算出した．

表 7　レビュー検索の精度評価の結果.

	比較手法		提案手法	
評価項目	単語	フレーズ	単語	フレーズ
適合率	0.94	1.00	0.60	0.72
再現率	0.42	0.06	0.20	0.60
F_1 値	0.56	0.11	0.27	0.61

比較手法では単語で検索を行うと F_1 値が高く，提案手法ではフレーズで検索を行うと F_1 値が高くなった.

6 考　　察

6.1 アンケートの分析

「結果の見やすさ」という部分において高い評価を受けたのは，提案手法において要約した結果を一目で可視化できたからであると考えられる．比較手法だと検索結果に大量の文字が表示されるため，全てのレビューを読むより軽減されるものの，レビューからユーザが知りたいことを探すには大変である．一方で提案手法では結果を短く要約し，件数を表示することで，ユーザにとってレビューを読むことに対する負担を減らしている．このため，アンケートにおいて有意差が認められ，多くのインタビューで言及された.

「意見の網羅性」という部分において高い評価を受けたのは，「見た目」と「デザイン」といった同じ意味を表す単語が含まれているレビューが表示されないという問題や，実際に文字列マッチングで検索するには適切なキーワードが浮かばないという問題が解消したからであると考えられる．実際に文字列マッチングで検索するには適切なキーワードが浮かばないという問題を解消できた例を挙げる,「購入後、1 年 2ヶ月で断線しました.」というレビューが気になったとき，ユーザは耐久性に関連するレビューを検索しようとする．しかし，故障の中でも「火花が出た」や「動かなくなった」など表現が様々であり，キーワード検索において一括で検索するのは難しい．提案手法では,「1 年 2ヶ月で断線」で検索を行うと「1 年ちょっとしか使ってないのに、コード根元から火花が散り、使えなくなりました.」「2 月に買って 7 月の頭にいつも通りターボで髪を乾かしていたらいきなり電源が切れスイッチがガチガチに固まってそれ以降動かなくなりました.」といった，様々な壊れ方をしたレビューが表示される．このため，意見の網羅性という項目において高い評価となり，多くのインタビューで言及された.

6.2 クエリの分析

提案手法の方が 1 つのクエリに対する滞在時間が長くなった要因は 2 つあると考えられる．1 つ目は，提案手法の方が処理に費やす時間が長くなるためである．提案手法では，gpt–35–turbo–16k を用いるために OpenAI の API を呼び出している．一方で，比較手法でクエリに一致するレビュー文を見つけるためには，API を呼び出す必要はない．クエリの検索ボタンを押してから次のクエリの検索ボタンが押されるまでを測定しているため，滞在時間には検索結果が表示されるまでの時間が含まれる．そのため，処理時間が滞在時間に影響を及ぼしている可能性があると考えられる．今後，結果が表示された時刻を記録することで，処理時間に起因せず滞在時間を測定する必要がある．2 つ目は，提案手法の方が検索結果に出力されるレビューの件数が多いためである．レビューの件数が多くなると，関係のないレビューが出力される可能性もある．ただ表 7 では，フレーズで検索したときの再現率が高くなっており，関係のあるレビューが出力される確率は高い．そのため，比較手法と比べて関係のあるレビューが多く出力され，それらを読むために滞在時間が長くなると考えられる.

6.3 検索精度の評価

比較手法でフレーズより単語の方が F_1 値が高くなったのは，クエリの文字数や情報量が多くなるにつれて完全にクエリと一致するレビューが少なくなるからである．たとえば「風量」の検索結果には,「風量が強い」というレビューも「風量が弱い」というレビューも含まれる．しかし,「風量が強い」の検索結果には「風量が弱い」というレビューが含まれない．したがって，単語で検索する方がクエリと一致するレビューの数が多くなったと考えられる.

また提案手法において，単語で検索したときに F_1 値が低くなった理由として，1 つの単語と文では言葉の単位が異なり，類似度が低下した組合せが多いことが考えられる．たとえば「音があまり気にならない」と「音量」の場合，類似度は 0.80 となり，要約の対象にはならない．ただ,「音」と「音量」は類似する単語であり，類似度も 0.87 と要約対象となる基準を上回っている.「音」の後に単語が続くことによって,「音」以外の単語も多く考慮した分散表現に変換されてしまい，類似度が 0.85 を満たさなくなったと考えられる.

7 まとめと今後の課題

本研究では，レビューの中でユーザが気になる部分を選択し，大規模言語モデルを用いて生成された類似するレビューの要約を提示するインタフェースを提案する．このインタフェースを用いることで，ユーザは類似文を検索することや，検索結果で内容を一目見て把握することが可能になる．実際にユーザに使用してもらい,「要約の見やすさ及び精度」,「ユーザの興味に対する網羅性」,「観点の網羅性」,「意見の網羅性」,「結果の見やすさ」,「システムの使用難易度」,「全体的な満足度」の 7 つの項目でアンケートを行った．その結果「システムの使用難易度」以外の 6 項目において，提案手法の回答者平均が比較手法を上回ったが，統計的な有意差はみられなかった．また，クエリについて分析を行った結果，提案手法において 1 つのクエリごとの滞在時間が長いことが多かったが，有意差はみられなかった.

そして，レビューの検索精度の評価も行った．文字列マッチングを用いた手法ではクエリに単語を用いた検索において F_1 値が高く，提案手法ではクエリにフレーズを用いた検索において F_1 値が高いという結果になった.

最後に，今後の課題について 3 点述べる．1 点目は類似レ

ビューを判定するシステムの調整である．6.3 節で述べたように，提案手法では文に対してベクトル化を行っているため，単語との類似度が低くなってしまった．そのため，類似レビューと判定する閾値の変更や，文に対して助詞といった単語の除外処理を行ってから分散表現に変換する必要があると考えられる．2 点目はプロンプトの改善である．出力例と入力例としてクエリとその元となったレビュー文を ChatGPT に入力すると，出力がクエリと同様である事例が多数あった．また，10 文字以内で要約できていない事例も存在していたため，現状とは異なるプロンプトの記述方法でも要約を行う必要がある．3 点目は要約の精度評価である．今回，システムの機能のうち，検索精度についての評価は行ったが，要約の精度評価はまだ行っていない．そのため，2 点目で述べた課題も踏まえ，プロンプトの改善前と改善後における要約の精度評価が必要である．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775，JP22H03905，による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3675. Association for Computational Linguistics, 2018.

[2] Yang Deng, Wenxuan Zhang, and Wai Lam. Opinion-aware answer generation for review-driven question answering in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 255–264, 2020.

[3] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, 2004.

[4] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2022.

[5] Koteswar Rao Jerripothula, Ankit Rai, Kanu Garg, and Yashvardhan Singh Rautela. Feature-level rating system using customer reviews and review votes. *IEEE Transactions on Computational Social Systems*, Vol. 7, No. 5, pp. 1210–1219, 2020.

[6] Shou Jianqi. Towards known unknowns: GPT large language models empower human-centered information retrieval. *Journal of Library and Information Sciences in Agriculture*, Vol. 35, No. 5, p. 16, 2023 (in Chinese).

[7] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[8] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. Is ChatGPT a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*, 2023.

[9] Thanh-Son Nguyen, Hady W Lauw, and Panayiotis Tsaparas. Review synthesis for micro-review summarization. In *Proceedings of the eighth ACM international conference on web search and data mining*, pp. 169–178, 2015.

[10] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, Vol. 108, pp. 42–49, 2016.

[11] David Pride, Matteo Cancellieri, and Petr Knoth. CORE–GPT: Combining open access research and large language models for credible, trustworthy question answering. In *International Conference on Theory and Practice of Digital Libraries*, pp. 146–159. Springer, 2023.

[12] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

[13] ホンヘジン, 木村優介, 神田悠斗, 吉丸直希, 波多野賢治. 商品説明に着目したレビュー文からの教師なしキーフレーズ抽出. 第 15 回データ工学と情報マネジメントに関するフォーラム, 1b–5–5, 2023.

[14] 金兵裕太, 沼尾雅之. ネットショッピングサイトの商品レビューを利用したジャンル毎の評価軸の自動構築とその応用. 第 8 回データ工学と情報マネジメントに関するフォーラム, C2–3, 2016.

[15] 栗原光祐, 荘司慶行, 藤田澄男, MartinJ.Dürst. Doc2vec 手法による映画レビューサイトからのクエリと意味的に類似した評価表現の発見. 第 11 回データ工学と情報マネジメントに関するフォーラム, C4–2, 2019.

[16] 経済産業省. 令和 4 年度 電子商取引に関する市場調査 報告書. `https://www.meti.go.jp/policy/it_policy/statistics/outlook/230831_new_hokokusho.pdf`. 2023 年 11 月 27 日閲覧.

[17] 市村哲. 口コミから美味しい料理店を手早く探すシステム. 情報処理学会論文誌, Vol. 61, No. 11, pp. 1748–1756, 2020.

[18] 杉木健二, 松原茂樹. カスタマーレビューに基づく商品検索のための感性表現シソーラスの構築. 言語処理学会第 15 回年次大会発表論文集, pp. 781–784, 2009.

[19] 渡辺一生, 楠和馬, 吉丸直希, 波多野賢治. レビュー要約を目的とした単語の意味領域に基づく上位語選択. 第 15 回データ工学と情報マネジメントに関するフォーラム, 1b–5–5, 2023.

[20] 平山拓央, 湯本高行, 新居学, 佐藤邦弘. 語の共起と極性に基づく商品レビュー閲覧支援システム. 情報処理学会研究報告, Vol. 2012–DBS–155, No. 3, pp. 1–9, 2012.

個人に最適な生活改善行動を対話的に探索するための YES/NO チャート自動生成

織田　直也†　　金　　鎭赫†　　荘司　慶行†　　山本　祐輔††

† 静岡大学大学院総合科学技術研究科情報学専攻　〒 432–8011 静岡県浜松市中央区城北 3–5–1
†† 名古屋市立大学データサイエンス学部　〒 467-8501 名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †oda@design.inf.shizuoka.ac.jp, ††{kimj,shojiy}@inf.shizuoka.ac.jp,
†††yusuke_yamamoto@acm.org

あらまし　本研究では，ユーザが生活習慣を改善するための具体的な行動（以下，改善行動）を対話的に探索するシステムを提案する．提案システムは，運動不足や不規則な睡眠習慣などのユーザが改善したい生活習慣トピックをテキスト形式で受け取ると，大規模言語モデル（LLM）を使用してトピックに対応した改善行動のリストを生成する．この時，生成されたリストから抽出アルゴリズムによって意味的重複が少ない改善行動のみがピックアップされる．次に，行動間の意味的類似度を計算し，ピックアップされた改善行動に対して階層的クラスタリングを行うことで二分木を生成する．最後に，生成された二分木の葉ノード以外のノードについて，LLM を用いて YES/NO 形式の質問文に変換し，YES/NO チャートを生成する．YES/NO チャートを根ノードから葉ノードに至るまで順次質問に回答することで，ユーザは自身にとって最適な改善行動を効率的に探索することが可能となる．抽出アルゴリズムの評価実験を行った結果，アルゴリズムが抽出したリストは抽出前のリストと比較して，より多様な行動を保持していることが明らかになった．また，ユーザ実験の結果，提案システムを用いることでユーザは改善行動探索を効率的に行うことができた．一方で，ウェブ検索や ChatGPT で探索する方が提案システムで探索するよりも，実行しやすく納得感のある改善行動を探索できるという結果も明らかになった．

キーワード　対話型情報検索，推薦システム，パーソナライゼーション，健康

1　はじめに

今日，食事や運動などの生活習慣が発症の主要因となる，いわゆる生活習慣病が人々の健康に大きな影響を与えている．特に，先進国においては不健康な食生活が広まっており，糖尿病などの生活習慣病による死亡率が増加している [1] [2]．日本においても生活習慣病は重大な問題であり，不健康な生活習慣が生活習慣病の罹患リスクに与える影響を調査した研究が数多くなされている [3] [4] [5]．こうした生活習慣病を予防するためには，運動や食事改善といった生活習慣を改善するための行動（以下，健康改善行動）をとることが必要とされている [6]．これを受けて，ユーザの生活習慣に介入するシステムに関する研究も盛んに行われている [7] [8]．

健康改善行動に関する情報は様々な媒体で発信されており，特にウェブを用いることで健康改善行動を容易に探索することができる [9] [10]．しかし，膨大なウェブ情報の中から自分に適した健康改善行動を効率的に特定することは困難である．個人の趣向や習慣によって最適な健康改善行動は異なる．そのため，ウェブ検索エンジンの検索結果上位のウェブページに掲載された改善行動が自分に最適なものであるとは限らない．限られたリソースの中で解となり得る健康改善行動の候補を探し，それぞれを吟味しながら自分にあった改善行動を探すのは骨の折れる作業である．また，自分にあった健康改善行動を見極めるためには，健康改善行動を探すためのニーズや制約条件をユーザが明らかにする必要がある．しかし，健康ドメインについての知識が不足するユーザにとって，探索のためのニーズや制約条件を整理するのは困難である．

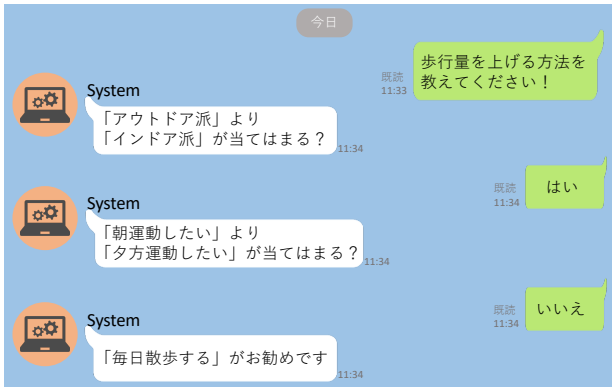


図 1　提案システムの使用例．ユーザが「歩行量を増やす」をトピックとして指定すると，システムはいくつか質問を提示した後，ユーザに適した改善行動を提案する．

本研究では改善行動を探索するために必要な質問を生成しユーザに提示することで，最適な改善行動を対話的に探索するシステムを提案する（図 1）．運動不足や不規則な睡眠週間など，ユーザが改善したい生活習慣トピックを受け取ると，提案

システムは指定されたトピックを改善するための行動リストを大規模言語モデル（LLM）を用いて生成する．例えば，運動不足を解消することをトピックとして指定すると，提案システムは「友達とウォーキングをする」,「自動車通勤をやめて歩いて通勤する」,「夜寝る前にヨガを行う」などの改善行動を生成する．次にこれらの改善行動に対して意味的類似度を計算することで階層的クラスタリングを行い，葉ノードに改善行動が設置された二分木を生成する．その後,「友達とウォーキングをする」,「自動車通勤をやめて歩いて通勤する」といった改善行動をもとに「他者と一緒に活動するより一人で活動する方が好きですか?」といった YES/NO 形式の質問を LLM を用いて生成し，二分木の葉ノード以外のノードを質問文に変換する．最後に改善行動を葉ノードに持つ YES/NO チャートを生成し，それをベースに質問を提示する．

2 関 連 研 究

2.1 健康情報検索における課題

一般的なユーザがウェブ上で健康情報を検索する際に様々な障壁があることが，多くの研究において報告されている．Zhang らは，高齢者は健康に関する情報をウェブ検索する際，ページの管理者やページ内の広告数などの限られた指標からウェブページの品質を判断しており，中には指標の意味を誤解している高齢者がいることを明らかにした [11]．Chi らは，健康に関するウェブページの探索方法について調査を行った結果，多くの人は検索トピックが重篤であれば専門的なウェブページを閲覧するが，軽度なトピックについては専門家による裏付けのない情報源も参考にすることを明らかにした [12]．鈴木らは，確証バイアスが健康トピックのウェブ検索行動に与える影響を調査した [13]．調査の結果，実験前に確証バイアスを与えられた参加者は，上位の検索結果のみを頻繁に閲覧し，検索タスクを早急に完了させることが明らかになった．

本研究では，情報検索能力に乏しいユーザでも自身に適した改善行動を探索可能にすることを目指す．

2.2 健康分野への LLM の利用

LLM を活用した健康支援を試みる研究も行われつつある．KhaoKaew らは，ウェアラブルセンサと LLM を組み合わせることで，睡眠に関してユーザごとに個人化されたアドバイスを提供するためのフレームワークを設計した [14]．Lu らは，LLM を用いて病理学の画像や文書に関する解釈を提供するシステムを提案した [15]．Zhou らは，実際の医療現場で LLM を活用するには，説明可能性の欠如やデータの不足などが課題として存在することが指摘している [16]．

本研究では LLM を用いて改善行動や質問の生成を行うことで，ユーザが健康情報を対話的に探索することを実現する．

2.3 対話的情報検索に関する研究

．対話的情報検索を用いる試みは医療，健康分野以外でも盛んになされている．Nguyen らは，ユーザが探したい動画に含まれるコンテンツや動画の撮影日，また類似した動画の選択によって，ユーザが目的の動画を対話的に検索するシステムを提案した [17]．Schedl らは楽曲を 3 次元空間上の立体として表現することで，楽曲を対話的に探索できるシステムを提案した [18]．提案システムは，楽曲のジャンルや音響特徴の分布に応じてブロックの色を変えたり，ユーザがあまり聴かない楽曲ほど遠くに配置したりすることで，楽曲の情報を視覚的に把握しやすくしている．平本らは，対話的に地図情報を取得できるシステムを提案した [19].

本研究では，LLM が生成した YES/NO チャートを用いてユーザに質問を提示することで，ユーザごとに最適な改善行動を対話的に探索するためのシステムを提案する．

3 提 案 手 法

本研究では，ユーザごとに適した改善行動を対話的に探索するシステムを提案する．図 1 のように，提案システムはまずユーザから改善したい生活習慣トピックをテキスト形式で受け取る．その後，対話的にいくつかの質問を提示し，回答結果に合わせて健康改善行動を推薦する．システムは内部的に YES/NO チャートを生成することで，質問の提示および健康改善行動の推薦を可能としている．以下，システムが YES/NO チャートを生成する手順について，図 2 を参照しながら説明する．

3.1 改善行動リストの生成

YES/NO チャートを生成するためには，トピックに対応する多様かつ重複が少ない健康改善行動リストを用意する必要がある．そのため,「改善行動の候補生成」および「意味的重複の少ない改善行動の抽出」の 2 ステップを経て，改善行動リストを生成する．

ステップ 1 では，ChatGPT[1]を用いて入力された健康トピックに対応する健康改善行動の候補を n 件生成する（図 2-(1) および (2) を参照）．候補生成用プロンプトは「私は生活習慣の改善に興味があります．<トピック>のための方法を n 件出力してください」といった単純なものを用いた．

ステップ 2 では，ステップ 1 で生成した候補から意味的に重複したものを省く．ChatGPT によって生成された健康改善行動リストは意味的に重複するものが含まれることがある．例えば，図 2 の 1 で生成された改善行動リストにある「階段を積極的に使う」と「エスカレータではなく階段を使う」は類似する健康改善行動である．これらの行動をそのまますべて用いて YES/NO チャートを生成すると，システムから類似した行動のみが推薦される可能性が高くなる．よって，リスト内の行動間の意味的類似度は可能な限り低くなることが好ましいと言える．そこで，ChatGPT が生成した改善行動リスト A_{in} から冗長なものを省いた要素数 $m(m < n)$ のリスト A_{out} を生成する．

抽出アルゴリズムの内容を Algorithm1 に示す．まず，n 件の改善行動を格納したリストを A_{in} とし，さらに空のリスト A_{out} を生成した後，A_{in} からランダムに取り出した a_{in} を A_{out} に移動する．次に，A_{in} のすべての要素 a_{in} と A_{out} との意味的類

1：ChatGPT: https://openai.com/blog/chatgpt

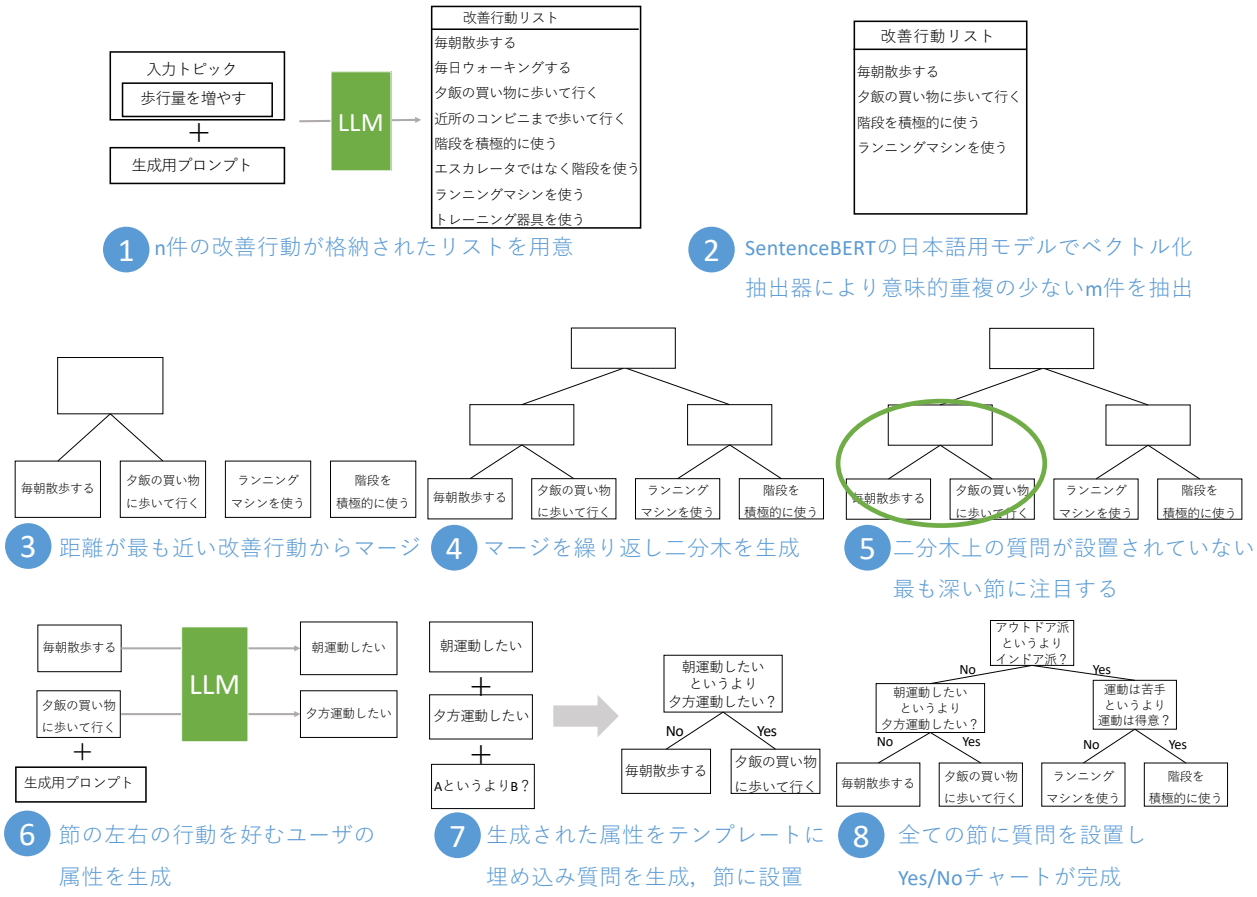


図 2　提案システムの動作フロー

Algorithm 1 改善行動候補リスト A_{in} から冗長なものを除き m $(m < n)$ 件の行動を抽出するアルゴリズム

Input: $A_{in} = \{a_1, ..., a_n\}, m \in \mathbb{N}$
Output: A_{out}

Select action a_i from A_{in} randomly.
$A_{out} \leftarrow \{a_i\}$, $A_{in} \leftarrow A_{in} \setminus \{a_i\}$
while $|A_{out}| < m$ **do**
    $a \leftarrow \underset{a \in A_{in}}{\text{argmin}}\ \textsc{List-sim}(a, A_{out})$
    $A_{out} \leftarrow A_{out} \cup \{a\}$
    $A_{in} \leftarrow A_{in} \setminus \{a\}$
end while

似度を算出し，最も低い意味的類似度を持った a_{in} を A_{out} に移動させる操作を A_{out} の要素数が m になるまで繰り返した.. なお，ある改善行動 x と改善行動のリスト Y が与えられとき，x と Y の類似度 $\textsc{List-Sim}(x, Y)$ の計算には以下の式を用いた

$$\textsc{List-Sim}(x, Y) = max(\{sim_{cos}(x, y) \mid y \in Y\}) \quad (1)$$

ここで，$sim_{cos}(x, y)$ は x および y のベクトル $\mathbf{v}(x)$，$\mathbf{v}(y)$ の cos 類似度を意味する．また，各健康改善行動のベクトル化には SentenceBERT の日本語用モデルを用いた [20].

3.2　二分木の生成

次に，第 3.1 節で抽出した m 件の行動を二分木構造にまとめ，YES/NO チャートの土台を生成する．ここで，最終的には YES/NO チャートを用いて根ノードから葉ノードに至るまで順次質問を提示することで，ユーザのニーズを段階的に絞り込み最適な改善行動を提示することに留意する．そのため，生成される二分木の浅い箇所では改善行動を大雑把に分ける質問がなされ，二分木の深い箇所では類似する健康改善行動を分ける質問が配置されることが望ましい．

そこで，本研究では健康改善行動リストに対して Ward 法による階層的クラスタリングを適用することで，類似する健康改善行動ほど近い葉に箇所に配置される二分木を生成する．なお，リスト内の改善行動のベクトル化には，SentenceBERT の日本語用モデルを利用し，768 次元のベクトルに変換する．また，クラスタ間距離の測定にはユークリッド距離を用いる．

図 2 の 3 ではいくつかの改善行動の中から最も距離が近い「毎日散歩する」と「夕飯の買い物に歩いて行く」がマージされる．その後，距離の近い行動同士のマージを繰り返し，図 2 の 4 のように改善行動を葉とする二分木が生成される．

3.3　質問生成

3.2 節で生成された二分木の節ノードに対して ChatGPT を使用して，YES/NO 形式の質問文を生成する．生成される質問は，ユーザがどのような属性やニーズを持っているのかを明らかにできるものが望ましい．また，その質問の回答結果に応

```
#命令:
listAには健康的な生活習慣を行うための行動が格納されています
私が好きな行動をlistAからlistBに抜き出しました
また，友達が好きな行動をlistAからlistCに抜き出しました
私と友達がそれぞれどのような趣味趣向を持った人間かを
類推してください
なお，出力は5個ずつ出してください

#制約条件:
・出力形式は，以下のようなlist形式にしてください
listB2=[
"毎日運動したい",
"計画的である",
"健康意識が高い",
]
・私の属性をlistB2，友達の属性をlistC2に出力してください
・listB2とlistC2の要素は意味的に
対比の関係になるようにしてください
・1.のようなインデックスを書かないでください
・""で文章を囲ってください
・listB2とlistC2の要素はそれぞれ5個にしてください
```

図 3　属性生成のためのプロンプト

じて，提示すべき改善行動の範囲を段階的に絞り込めるものが望ましい．そこで，本研究では YES/NO チャート上の改善行動から想定されるユーザの属性やニーズを ChatGPT によって推定し，推定された属性とあらかじめ用意した質問テンプレートを用いて質問を生成する．

以下，質問生成の具体的なフローを示す．まず，3.2 節で生成された二分木上の質問が設置されていない，かつ最も深い節に注目する．次に注目した節の左右の部分木上に存在する改善行動をリストアップする．リストアップされた健康改善行動を図 3 に示したプロンプトに組み込み ChatGPT に入力することで，それぞれの行動に対応したユーザの属性を推定する．図 2 の 6 では，注目する節の左部分木上の行動として「毎朝散歩する」，右部分木上の行動として「夕飯の買い物に歩いて行く」がリストアップされており，それぞれに対応するユーザ属性が推定されている．ユーザ属性が推定後，推定された属性を質問テンプレートに埋め込み，節に対応する質問を生成する．使用するテンプレートは「<属性 A>というより<属性 B>が当てはまりますか？」である．<属性 A>，<属性 B>には左右の部分技上の行動から生成した属性がそれぞれ当てはまる．例えば，図 2 の 7 では,「朝運動したい」,「夕方運動したい」をテンプレートに当てはめることで質問を生成，配置している．以上のような操作をすべての節に質問が配置されるまで繰り返すことで，図 2 の 8 のような YES/NO チャートが完成する．

4　提案システムの評価実験

提案システムによって，ユーザ自身に適した改善行動を効率的に探索できるかを検証するために，オンライン上でのユーザ実験を実施した．実験協力者は，PC または Mac を用いて GoogleMeet 上で実験に参加した．本実験は，2024 年 1 月 10 日から 1 月 16 日の期間に実施した．実験協力者は静岡大学の学生，大学院生から募集し，12 名が実験に参加した．実験協力者のうち 1 名が女性であった．

4.1　仮説と評価項目

本実験で検証した仮説と検証のための評価項目は以下の 2 つである．

- H1：YES/NO チャートを用いた対話形式の探索システムで改善行動を探索するとユーザは効率的に改善行動を探索できる

評価項目：探索所要時間，探索の実施容易度

- H2：YES/NO チャートを用いた対話形式の探索システムで改善行動を探索するとユーザは自身に適した改善行動を見つけ出せる

評価項目：改善行動の最適化度合い，改善行動の実現性，改善行動の納得度合い，改善行動のモチベーション寄与度

各評価項目の測定方法について説明する．探索所要時間は，Mac に搭載されたストップウォッチを用いて筆者が計測した．実験協力者が探索タスクを始めると同時に計測を始め，実験協力者からタスクを完了した旨の報告をしたと同時に計測を終了した．

探索の実施容易度，改善行動の実現性，改善行動の納得度合い，改善行動のモチベーション寄与度は，タスク完了後アンケートにて質問を実施することで測定した．提示した質問と評価項目の対応は以下の通りである．

- 探索の実施容易度：改善行動の探索は容易だった
- 改善行動の実現性：最終的に採用した改善行動は，自分にとって実行しやすいものだと思う
- 改善行動の納得度合い：最終的に採用した改善行動は，自分にとって納得のいくものだと思う
- 改善行動のモチベーション寄与度：最終的に採用した改善行動は，生活習慣改善へのモチベーションを向上させるものだと思う

各質問は,「非常に当てはまる」から「全く当てはまらない」の 5 段階で評価してもらった．

改善行動の最適化度合いは，タスク完了後アンケートにおいて，実験協力者が探索した改善行動と比較用の行動 16 件に対して，順位付けを行なってもらうことで評価した．実験協力者に上記 17 件の行動について自身の好みに合うかで 5 点満点のスコア付けを行なってもらうことでランキングを生成し，実験協力者が探索した改善行動が何位になるかを測定した．同率順位が発生した場合は，順位の重複を認め，重複した分だけ次の順位をスキップする．

4.2　比 較 手 法

実験では提案システムと比較手法 2 つを用いて実験協力者に改善行動探索を行ってもらった．提案システムの実際の画面を図 4 に示す．比較手法には，一般的な情報検索手法である以下の 2 つを採用した．

- ウェブ検索での改善行動探索

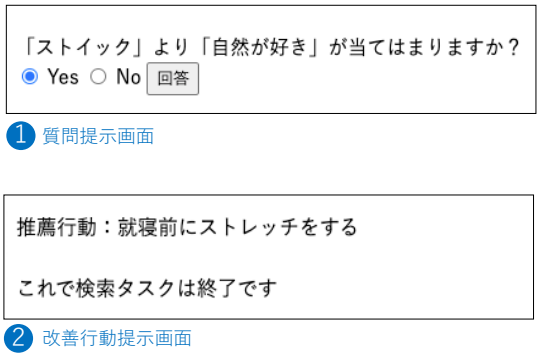


図 4　実験で実験協力者が閲覧した実際の提案システムの画面．1 のようにいくつか質問が提示された後，2 のように改善行動が提示される．

ウェブブラウザに自由に検索クエリを入れて改善行動を探索する．検索エンジンは Google を用いた．

- ChatGPT での改善行動探索

ChatGPT に自由に検索クエリを入れて改善行動を探索する．モデルのバージョンは GPT3.5 を用いた

4.3　実験計画と手順

本実験は被験者内計画にて行った．すなわち，実験協力者はウェブ検索，ChatGPT，提案システムの 3 つの検索手法すべてで改善行動探索を行う．実験は以下の手順で行った．

（1） 事前アンケートへの回答
（2） ウェブ検索での改善行動探索
（3） タスク完了後アンケートへの回答
（4） ChatGPT での改善行動探索
（5） タスク完了後アンケートへの回答
（6） 提案システムでの改善行動探索
（7） タスク完了後アンケートへの回答

事前アンケートでは，検索トピックごとに関心度合いを「まったくない」から「とてもある」の 5 段階で評価してもらった．

手法ごとに探索するトピックは以下の 6 つの候補の中からランダムに 2 つずつ割り当てた．

- 歩行量の増加
- 代謝の上昇
- 野菜摂取量の増加
- 減塩の実施
- ストレス解消
- 寝つきの向上

検索手法とトピックの組み合わせが探索に影響を与えることを回避するため，ラテン方格 [2]を用いて，表 1 のように検索手法とトピックの組み合わせをランダム化した．ただし，検索手法の実施順についてはランダム化を行なっておらず，全ての実験協力者はウェブ検索，ChatGPT，提案システムの順番で検索タスクに取り組んだ．

2：ラテン方格:https://mathworld.wolfram.com/LatinSquare.html

表 1　ラテン方格によりランダム化した検索手法とトピックの組み合わせ

実験協力者	ウェブ検索	ChatGPT	提案システム
S1, S2	ストレス解消, 代謝の上昇	減塩の実施, 歩行量の上昇	寝つきの向上, 野菜摂取量の増加
S3, S4	野菜摂取量の増加, ストレス解消	代謝の上昇, 寝つきの向上	減塩の実施, 歩行量の上昇
S5, S6	歩行量の上昇, 野菜摂取量の増加	ストレス解消, 減塩の実施	代謝の上昇, 寝つきの向上
S7, S8	寝つきの向上, 歩行量の上昇	野菜摂取量の増加, 代謝の上昇	ストレス解消, 減塩の実施
S9, S10	減塩の実施, 寝つきの向上	歩行量の上昇, ストレス解消	野菜摂取量の増加, 代謝の上昇
S11, S12	代謝の上昇, 減塩の実施	寝つきの向上, 野菜摂取量の増加	歩行量の上昇, ストレス解消

4.4　結　　果

12 名の実験協力者の主観指標と行動指標を分析し，提案システムがユーザの改善行動探索に与える影響を分析した．ここでは，収集したデータが正規分布に従わなかったため，3 つの検索手法についてノンパラメトリック法のクラスカル・ウォリス検定を採用した．また，事後分析で多重比較検定にスティール・ドゥワス検定を用いた [21]．

表 2 は，探索所要時間，4 つの質問に対する回答，探索した改善行動の順位に対する平均値，標準偏差，分散分析の統計的有意性，3 つの検索手法に対する多重比較の結果を示している探索所要時間の単位は秒で，平均値が高いほど探索により長い時間を費やしたことを示している．Q1 から Q4 は平均値が高いほど，検索手法や探索した改善行動に対してより肯定的な結果を示している．探索した改善行動の順位は，平均値が低いほど実験協力者が探索した改善行動が他の行動に比べてより好まれていることを示している．

探索所要時間について，表 2 が示す通り，多重比較における提案システムとの比較では，ウェブ検索と ChatGPT のいずれの間にも有意差が確認された（ウェブ検索：$p < 0.05$，ChatGPT：$p < 0.05$）．また，提案システムはウェブ検索，ChatGPT と比べて，平均値が有意に低くなる傾向にあった (探索所要時間: 164.13s vs 139.79s vs 26.58s; $p < 0.001$)．このことは，提案システムを用いるとウェブ検索や ChatGPT と比較して，短い時間で実験協力者が改善行動探索を完了したことを意味する．

Q1 について，提案システムとの比較では，ウェブ検索と ChatGPT のいずれの間にも有意差が確認された（ウェブ検索：$p < 0.05$，ChatGPT：$p < 0.05$）．また，ウェブ検索と ChatGPT の間にも有意差が確認できる ($p < 0.05$)．提案システム，ChatGPT，ウェブ検索の順で平均値が有意に高くなる傾向にあった (Q1: 4.29 vs 4.75 vs 5.00; $p < 0.001$)．このことは，ChatGPT，ウェブ検索よりも提案システムを用いると探索がしやすかったと実験協力者が判断したことを意味する．

Q2 から Q4 について，ウェブ検索，ChatGPT，提案システムの順で平均値が高くなる傾向にあったが，有意差は見られなかった (e.g.Q2: 4.37 vs 4.29 vs 3.88; $p > 0.05$)．このことは，どの検索手法で探索した改善行動についても，実験協力者が同

表 2　探索所要時間，質問に対する回答スコア，探索した改善行動の順位の平均値と標準偏差，および分散分析における統計的有意性（*：0.05，***：0.001）．^{W,C,S} はそれぞれウェブ検索，ChatGPT，提案システムの 2 対条件化における事後検定の有意性（p<0.05）を示す．

評価項目	検索手法			p-value
	ウェブ検索	ChatGPT	提案システム	
探索所要時間（秒）	164.13(112.92)^{S}	139.79(103.86)^{S}	26.58(5.04)^{W,C}	***
Q1. 改善行動の探索は容易だった	4.29(0.65)^{C,S}	4.75(0.34)^{W,S}	5.00(0.0)^{W,C}	***
Q2. 最終的に採用した改善行動は，自分にとって実行しやすいものだと思う	4.37(0.57)	4.29(0.58)	3.88(0.83)	0.22
Q3. 最終的に採用した改善行動は，自分にとって納得のいくものだと思う	4.50(0.56)	4.33(0.78)	4.17(0.58)	0.35
Q4. 最終的に採用した改善行動は，生活習慣改善へのモチベーションを向上させるものだと思う	3.91(0.73)	3.79(0.81)	3.71(1.14)	0.98
探索した改善行動の順位	2.96(3.61)^{S}	3.13(1.51)	5.08(2.50)^{W}	*

程度に実現性や納得感を感じ，生活習慣改善への意欲を奮起させたことを意味する．

探索した改善行動の順位について，多重比較における提案システムとの比較では，ウェブ検索との間に有意差が確認された（ウェブ検索：$p < 0.05$）．また，ウェブ検索は提案システムと比べて，平均順位が高くなる傾向にあった (探索した改善行動の順位: 2.96 vs 5.08; $p < 0.05$)．このことは，提案システムで探索した改善行動と比較して，ウェブ検索で探索した改善行動を実験協力者がより好ましく感じたことを意味する．

5　抽出アルゴリズムの評価実験

抽出アルゴリズムによって ChatGPT が生成した改善行動リストから意味的に重複したものを省くことができたかを検証するために，ユーザ実験を実施した．本実験は，2024 年 1 月 18 日から 1 月 19 日の期間に実施した．実験協力者は静岡大学の学生，大学院生から募集し，23 名が実験に参加した．実験協力者のうち 2 名が女性であった．

5.1　評 価 尺 度

本実験では，抽出アルゴリズムが抽出したリストが抽出前のリストと比較して類似した行動が少ないかを検証した．検証ためにグループ数，相対的多様度という 2 つの評価尺度を用意した．

グループ数は，改善行動リスト内の行動をいくつのグループに分けられるかを 1 から 16 の値で回答する尺度である．グループ数が多いと評価されたリストほど，多様な行動を保持すると言える．相対的多様度は，2 つの改善行動リストの内，より多様な行動を保持していると感じるリストを回答する尺度である．より多くの回答者から多様であると判定されたリストほど，多様な行動を保持すると言える．

5.2　比 較 手 法

実験では，抽出前のリスト内から 16 件の行動を抽出したリスト（以下，非多様化リスト）を比較手法として設定した．実験協力者は，非多様化リストおよび抽出アルゴリズムが抽出した 16 件を格納した行動リスト（以下，多様化リスト）を用いたタスクを行った．

表 3　非多様化リストおよび多様化リストのグループ数の平均値，標準偏差，統計的有意性（有意水準は***：0.001）．

非多様化リスト	多様化リスト	p 値
3.79(1.26)	4.97(2.17)	***

5.3　実 験 手 順

本実験では，第 4 章でも使用した 6 つのトピックについて，非多様化リストと多様化リストの 2 種類のリストを用意した．

本実験で実施したタスクは 2 種類である．両タスクにおいて，非多様化リストをリスト A，多様化リストをリスト B として提示した．

1 つ目のタスクはグループ分けタスクである．このタスクでは，リスト A とリスト B を順番に提示し，それぞれが直感的にいくつのグループに分割できるかを尋ね，グループ数を測定した．グループ数は 1 から 16 までの数値で記述してもらった．

2 つ目のタスクは多様度比較タスクである．このタスクでは，非多様化リストと多様化リストを比較し，どちらがより多様なリストかを選択してもらい，相対的多様度を測定した．実験協力者は，リスト A，リスト B，どちらも同じの 3 つの選択肢から 1 つを選択した．

5.4　結　　果

表 3 は，リストごとの平均グループ数，標準偏差，対応のある t 検定における統計的有意性を示している．グループ数の平均値が高いほど，提示したリスト内の行動が意味的に多様であることを示している．表 3 に示すように，多様化リストの平均グループ数は非多様化リストの平均グループ数より有意に多いことがわかる．

表 4 は，多様度比較タスクにおいて「非多様化リスト」,「多様化リスト」,「どちらでもない」の各選択肢が選ばれた割合を百分率で示している．カイ二乗検定を行った結果，p 値は有意水準より小さく，多様化リストの選択割合が有意に高かったことが示された．表 4 に示すように，多様化リストが多様であると選択される割合が 68.1%と最も高かった．

6　考　　察

本研究の目的は，改善行動探索に必要な質問を生成しユーザに提示することで，対話的に改善行動を探索するシステムを開

表 4　多様度比較タスクにおける 3 つの選択肢の選択割合（%）および統計的有意性（有意水準は***：0.001）.

非多様化リスト	多様化リスト	どちらでもない	χ^2	p 値
14.4	68.1	17.3	75.3	***

発することである．提案システムの実装にあたり，ChatGPT が生成した改善行動リストから意味的に重複した行動を除去するためのアルゴリズムを実装した．実装したアルゴリズムとシステムの効果を測定するためにそれぞれ実験を行った．本章では，各実験の結果をもとにアルゴリズムとシステムの効果について議論する．

6.1　提案システムの評価実験

第 4 章における実験の結果，提案システムを使用すると，ウェブ検索や ChatGPT を用いた場合と比較して，実験協力者は短時間で改善行動探索を完了できることがわかった．また，提案システムはウェブ検索や ChatGPT といった一般的な情報検索手法と比較し，改善行動探索の実施が容易であることも明らかになった．

ウェブ検索や ChatGPT で改善行動探索を行う際には，実験協力者は検索結果から得られた情報を吟味しなければならない．他方で，提案システムを用いる際は，提示された質問に対して Yes か No のどちらが当てはまるかを考えるだけで実験協力者は探索を実施できる．そのため，提案システムによる改善行動探索において，実験協力者は短時間で探索を完了させたり探索が容易であると感じたりしたと考えられる．以上の結果から，「H1：YES/NO チャートを用いた対話形式の探索システムで改善行動を探索するとユーザは効率的に改善行動を探索できる」は支持されたと言える．

改善行動の実現性，改善行動の納得度合い，改善行動のモチベーション寄与度については，提案システムの平均値が比較手法の平均値を下回る結果となった．ウェブ検索や ChatGPT による探索では，実験協力者が納得感がある，あるいは実現可能だと感じる行動が見つかるまで探索できるため，比較手法で探索した改善行動がより高い評価を得たと考えられる．また，改善行動の最適化度合いについては，提案システムがウェブ検索を有意に下回る結果となった．

提案システムは比較手法に比べて提示する情報量が少ないため，探索結果の評価に差が現れたと考えられる．提案システムでの探索の際には改善行動の名称のみがユーザに提示される．一方で，ウェブ検索を用いて探索する際には，改善行動の名称に加えて，その行動を実施する際の注意点や効果的に実施するためのアドバイスも提示される．このような捕捉情報の提示の有無が探索結果の評価の差につながったと考えられる．

以上の結果から,「H2：YES/NO チャートを用いた対話形式の探索システムで改善行動を探索するとユーザは自身に適した改善行動を見つけ出せる」は支持されなかったと言える．ただし，全 4 件中 3 件の評価項目については有意差が確認できず，また提案システムが推薦した改善行動の順位は平均して 16 件中の 5 位と比較的高順位であった．このため，提案システムがユーザに推薦した改善行動が全くもって不適切であったとは考えられない．

H1 と H2 の結果より，提案システムは従来の検索手法と比較して，探索結果の質をあまり落とすことなく探索効率を格段に向上させたと考えられる．よって，提案システムは，第 1 章にて述べた「探索時間効率と探索結果の個人最適化度合いのトレードオフ」をある程度解消できると考えられる．ただし，検索手法の実施順はランダム化されておらず，順序効果による実験結果への影響は否定できない．順序効果を排除した結果を得るために，今後は実験協力者数を増やし，グレコラテン方格を用いて検索手法の実施順をランダム化した，より頑健な実験を実施することを検討する．

6.2　抽出アルゴリズムの評価実験

第 5 章における実験の結果，グループ分けタスクにおいて，多様化リストは非多様化リストに比べて有意に多くのグループ数に分割された．抽出アルゴリズムが出力したリストの方が意味的に重複した行動が少なく，実験協力者はより多くのグループに分けることができたと考えられる．

多様度比較タスクにおいては，多様化リストは非多様化リストに比べて有意に多くの回数において多様であると判定された．無加工の改善行動リストと抽出アルゴリズムが出力したリストを並べて比較した際に，後者の方がより多様な行動を保持すると実験協力者が判断したと考えらえる．以上の結果から，抽出アルゴリズム用いることで意味的に重複した行動を除去するが可能であると言える．

6.3　本研究の限界

本研究にはいくつかの限界がある．第一に，ChatGPT が生成する改善行動の健康効果は保証されていない．実際に，医療分野における LLM の活用について，学習データの不足や生成結果の信頼性を指摘した研究も存在する [16]．よって，提案システムで探索した行動を実勢しても，実際に生活習慣の改善に繋がるかは保証できない．生活改善に確実に寄与する行動を提示するには，より健康情報を学習データとして取り込んだ LLM が必要であると考える．

第二に，探索にかかる計算量が $O(n)$ になることを回避できない．3.2 節では，Ward 法を用いて二分木を生成した．この場合，生成される木の形によっては探索にかかる最悪計算量が $O(n)$ になることもありうる．n 件の改善行動を平衡二分木で探索する際の計算量は $\log_2 n$ であり，探索効率を保証するためには平衡二分木を生成することが望ましい．

7　おわりに

本研究では，改善行動を YES/NO チャート形式にまとめることで，ユーザに最適な改善行動を対話的に探索するシステムを提案した．提案システムは ChatGPT を活用することで，改善行動リストやユーザの属性を絞り込む質問を生成した．さらに，ChatGPT が生成した改善行動リストの多様度を向上させ

るため，抽出アルゴリズムを作成した．

作成した抽出アルゴリズムによって出力されたリストが，抽出前のリストと比較して，意味的に重複した行動が少なくなるかを調査した．実験の結果，抽出アルゴリズムが出力したリストは抽出前のリストと比較して，有意に多様であることがわかった．また，提案システムがユーザの改善行動探索に与える影響を調査するために，静岡大学の学生を対象としたユーザ実験を行った．ユーザ実験の結果，提案システムは比較手法に比べ，探索される改善行動の質をあまり落とすことなく，効率的に改善行動探索を実施可能にすることが明らかになった．

今後は，より効率的な探索を可能にするため，平衡二分木を用いたYES/NOチャートの生成を行う必要がある．また，よりユーザに適した改善行動を探索可能にするために，改善行動と矛盾しない質問を確実に生成する方法を検討する必要がある．

謝　辞

本研究はJSPS科研費18H03554, 18H03775, 22H03905，および大幸財団の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Earl S Ford, Manuela M Bergmann, Heiner Boeing, Chaoyang Li, and Simon Capewell. Healthy lifestyle behaviors and all-cause mortality among adults in the united states. *Preventive medicine*, Vol. 55, No. 1, pp. 23–27, 2012.

[2] Paul Zimmet, K George Alberti, Dianna J Magliano, and Peter H Bennett. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. *Nature Reviews Endocrinology*, Vol. 12, No. 10, pp. 616–622, 2016.

[3] Yoshihiro Fukumoto. Nutrition and cardiovascular diseases. *Nutrients*, Vol. 14, No. 94, 2022.

[4] Hisashi Adachi, Mika Enomoto, Ako Fukami, Sachiko Nakamura, Yume Nohara, Shoko Kono, Akiko Sakaue, Hitoshi Hamamura, Kenta Toyomasu, Maki Yamamoto, et al. Trends in nutritional intake and coronary risk factors over 60 years among japanese men in tanushimaru. *Heart and Vessels*, Vol. 35, pp. 901–908, 2020.

[5] Akiko Sakaue, Hisashi Adachi, Mika Enomoto, Ako Fukami, Eita Kumagai, Sachiko Nakamura, Yume Nohara, Shoko Kono, Erika Nakao, Nagisa Morikawa, et al. Association between physical activity, occupational sitting time and mortality in a general population: An 18-year prospective survey in tanushimaru, japan. *European journal of preventive cardiology*, Vol. 27, No. 7, pp. 758–766, 2020.

[6] Yuzo Sato, Masaru Nagasaki, Naoya Nakai, and Takashi Fushimi. Physical exercise improves glucose metabolism in lifestyle-related diseases. *Experimental Biology and Medicine*, Vol. 228, No. 10, pp. 1208–1212, 2003.

[7] Hsien-Te Kao, Shen Yan, Homa Hosseinmardi, Shrikanth Narayanan, Kristina Lerman, and Emilio Ferrara. User-based collaborative filtering mobile health system. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 4, No. 4, pp. 1–17, 2020.

[8] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp 2015, p. 707–718, 2015.

[9] Thomas R Eng, Andrew Maxfield, Kevin Patrick, Mary Jo Deering, Scott C Ratzan, and David H Gustafson. Access to health information and support: a public highway or a private road? *Jama*, Vol. 280, No. 15, pp. 1371–1375, 1998.

[10] Tabitha Tonsaker, Gillian Bartlett, and Cvetan Trpkov. Health information on the internet: gold mine or minefield? *Can Fam Physician*, Vol. 60, No. 5, pp. 407–408, 2014.

[11] Yan Zhang and Shijie Song. Older adults' evaluation of the credibility of online health information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR 2020, p. 358–362, 2020.

[12] Yu Chi, Daqing He, Shuguang Han, and Jie Jiang. What sources to rely on: Laypeople's source selection in online health information seeking. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR 2018, p. 233–236, 2018.

[13] Masaki Suzuki and Yusuke Yamamoto. Characterizing the influence of confirmation bias on web search behavior. *Frontiers in Psychology*, Vol. 12, pp. 1–11, 2021.

[14] Yonchanok Khaokaew, Thuc Hanh Nguyen, Kaixin Ji, Hiruni Kegalle, and Marwah Alaofi. Zzzgpt: An interactive gpt approach to enhance sleep quality. *arXiv preprint arXiv:2310.16242*, Vol. abs/2310.16242, No. 1, 2023.

[15] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Kenji Ikamura, Georg Gerber, Ivy Liang, Long Phi Le, Tong Ding, Anil V Parwani, and Faisal Mahmood. A foundational multimodal vision language ai assistant for human pathology, 2023.

[16] Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.

[17] Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Chonlameth Arpnikanondt, Cathal Gurrin, Annalina Caputo, and Graham Healy. Efficient search with an interactive video retrieval system for novice users in ivr4b. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, CBMI 2023, p. 168–172, 2023.

[18] Markus Schedl, Michael Mayr, and Peter Knees. Music tower blocks: Multi-faceted exploration interface for web-scale music access. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR 2020, p. 388–392, 2020.

[19] Ryoko Hiramoto and Kazutoshi Sumiya. A web search method using user operation on digital maps. In *Proceedings of the 7th International Conference on Mobile Data Management*, MDM 2006, p. 106, 2006.

[20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[21] Robert GD Steel. A rank sum test for comparing all pairs of treatments. *Technometrics*, Vol. 2, No. 2, pp. 197–207, 1960.

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 13:30 ~ 15:40 | T4-A オンライン（Zoom Events)

LLM応用

座長:石川 佳治(名古屋大学)
コメンテータ:上田 真由美(流通科学大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T4-A-5-01] 欲求度に着目した行動手順漫才台本自動生成

*柘植 陽介[1]、下崎 安紋[2]、灘本 明代[3] (1. 甲南大学知能情報学部知能情報学科　灘本研究室、2. 甲南大学大学院自然科学研究科知能情報学専攻　灘本研究室、3. 甲南大学知能情報学部知能情報学科)

14:00 ~ 14:25
[T4-A-5-02] 意思決定における大規模言語モデルを用いた検索要求の観点抽出の影響分析

*笹本 まこと[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

14:25 ~ 14:50
[T4-A-5-03] 歌詞と音響特徴の分析に基づくText2Imageモデルを用いた楽曲サムネイル画像の自動生成

*佐々木 翔一[1]、西田 紘子[2]、澤井 賢一[2]、工藤 真生[2]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府、2. 九州大学大学院芸術工学研究院)

14:50 ~ 15:15
[T4-A-5-04] チャート画像に対する大規模マルチモーダルモデルの反実仮想推論能力の検証

*戸崎 友輔[1]、宮森 恒[1] (1. 京都産業大学情報理工学部情報理工学科)

15:15 ~ 15:40
[T4-A-5-05] 【技術報告】NTT版LLM「tsuzumi」の取組について

西田　京介 (NTT 人間情報研究所)

欲求度に着目した行動手順漫才台本自動生成

柘植　陽介†　　下崎　安紋††　　灘本　明代†

† 甲南大学知能情報学部知能情報学科　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
†† 甲南大学大学院自然科学研究科　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †{s2071074,m2324007}@s.konan-u.ac.jp, ††nadamoto@konan-u.ac.jp

あらまし　これまで我々は日常的な笑いを提供する手法の一つとして，ニュースから漫才台本自動生成手法の提案を行っている．漫才にはある程度話の流れがある．本研究ではこの漫才における話の流れを日常的な行動の手順を用いてこの手順に基づいた漫才台本の自動生成手法を提案する．具体的には，動作に関するキーワードを入力し，その動作を目的とした手順を ChatGPT により生成する．この生成された手順からボケを生成し，日常の手順に基づいた漫才台本を自動生成する．この時，人の欲求が高い手順が重要な手順であると考え，マズローの欲求段階説に基づいて，手順ごとの欲求度を求める．そしてその欲求度に基づき，笑いを誘発するボケの作成を行い，漫才台本を自動生成する．

キーワード　漫才，台本自動生成，ChatGPT，欲求段階説

1　はじめに

日常生活において，笑いは健康効果に期待できるといわれている [1]．親しみやすく笑いを提供できるコンテンツは多数あるが，その中の一つとして漫才がある．しかしながら，実際の漫才師が常に新たな漫才を様々な人に対して日常的に提供することは難しい．そこで我々は新しい漫才を日常的に提供することを目的とし，常に更新され続けるニュース記事を用いた漫才台本自動生成を提案している [1]．我々の提案する漫才台本自動生成の特徴は，ユーザの興味と関連するニュース記事から漫才台本を自動生成し，それを漫才ロボットやチャットアプリケーションを用いて漫才を提供していることである [2]．一般に，漫才はボケとツッコミの掛け合いからなる話の流れが重要である．そこで我々は，この漫才における話の流れをニュース記事の流れに見立て，漫才台本生成を行っている [3]．しかしながら，ニュースは難解なニュースもあり，一般の方々に対して共感を得やすい題材でないものもあるため，常に親しみを持つ漫才になっているとはいいがたい．そこで本研究では，共感を得やすい日常的な行動に着目し，その行動を達成するための手順を漫才における話の流れに見立て，漫才台本の自動生成を行う．ここでいう日常的な行動の手順とは，たとえば「電車に乗る手順」や「学校に行く手順」などである．これらは人々が日常的に行うまたは行った行動であり，親しみのあるコンテンツになると考え，本研究では日常的な行動に着目した漫才台本の自動生成手法を提案する．

本研究では漫才の流れのもととなる日常的な手順の生成には大規模言語モデルの一つである ChatGPT[2] を用いる．日常的な行動の手順を取得する方法を以下に示す．

（1）　ユーザは漫才にしたい日常的な行動に関するキーワードをシステムに入力する．

（2）　システムは ChatGPT により，入力されたキーワードを目的とした手順を生成する．

（3）　生成された中で，簡潔に記された手順を取得する．

（4）　システムは (3) により入手した手順を基に，ボケとツッコミの対話を自動で生成し，漫才台本を自動生成する．

また，図 1 に手順の取得する例を示す．この生成された手順により，ボケを作成する．

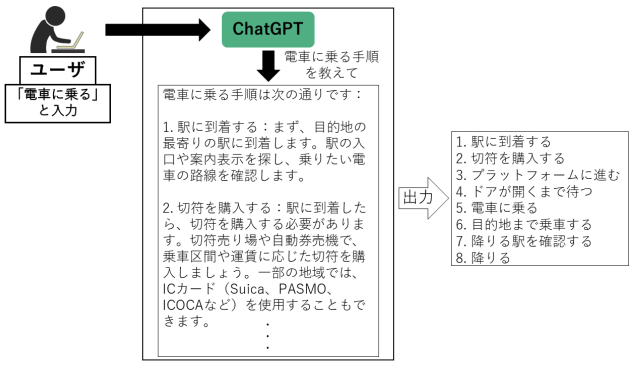


図 1　行動手順の生成例

2　関連研究

漫才に関する様々な研究が行われている．吉田ら [4] は，漫才形式の台本を多様にかつ自動的に生み出すため，漫才形式の対話文自動生成システムを提案している．このシステムでは入力された文の解析を行い，名詞の置き換えを行うことでボケ文を生成する．次にそのボケ文の置き換えた単語から連想される語を所得し，例えによるツッコミ文を生成する．この一連の流れを繰り返すことにより，漫才台本の自動生成を行っている．本研究では，単語の置き換えに限らずボケ文を自動生成するという点で異なる．

1：https://www.pref.osaka.lg.jp/attach/4002/00029624/waraisasshi.pdf
2：https://openai.com/chatgpt

笑いを誘発するコンテンツの生成に関する研究についても行われている．岩倉ら [5] は，ユーザの趣味や嗜好を取り入れたユーモア文の自動生成手法の検討を行っている．この手法では，ツイートデータからユーザの趣味に関するデータベースを構築し，そのデータベース内の名詞と入力名詞を Word2Vec [6] を用いてコサイン類似度を算出する．このコサイン類似度によって決定した単語を含んだコンテンツを生成している．また原ら [7] は，ChatGPT を用いて様々な状況での対話に対するユーモア応答生成の分析を行っている．原らは，対話文データをプロンプトに与えその対話文の最後の発話に続くボケやツッコミの 1 発話の生成を行っている．これらに対して本研究では，システムによってボケとツッコミからなる対話を自動生成するという点で異なる．

また，欲求度に基づいたコンテンツの生成に関する研究も行われている．薛ら [8] は，プルチックの 8 感情によってマズローの欲求段階説を考慮した，LSTM による対話文生成を行っている．本研究では，漫才台本を自動生成するという点で異なる．

3　行動手順漫才台本自動生成手法

本研究では，一般の方々に対して共感を得やすい漫才を提供するため，日常的な行動の手順から漫才台本を自動生成する．漫才台本においては「つかみ」「本ネタ」「オチ」の漫才 3 段論法を用いる．そこで本研究では，つかみとして非常識な行動をするという小ボケである「道徳観つかみ」を提案する．また本ネタは，複数のボケコンポーネントからなるが，本論文では 1 つの手順の中から任意の 1 単語を抜き出し，その単語とは違う単語を答えて笑わせるボケコンポーネントである「Q&A ボケ」と,「高欲求手順欠損ボケ」を提案する．

3.1　道徳観つかみ

つかみとは，挨拶から本ネタの流れを円滑にするために重要な導入部分である．従来手法においてのつかみの生成は，生成した日時に合わせたつかみを生成している．しかしながら本研究では日常的な行動の手順から本ネタを構成しているため，本ネタとの結びつきがない．そこで，ユーザの入力した日常的な行動におけるつかみを生成する．この時，安部のおかしみの構造 [9] に基づきボケを生成する．おかしみの構造とは，概念のはき違いにより笑いを生じさせるという構造である．そこで，本研究ではユーザの入力した行動において，常識に反する行動をするというボケにより笑いを生じさせる．例えば，図 2 に示すような「電車に乗る手順」の場合,「電車の中では大声をだしてはいけない」という常識に反して,「電車の中で，大声で叫んだんよ」といったセリフをボケに言わせて，ツッコミがつっこむというつかみである．本研究ではこのつかみを「道徳観つかみ」と呼ぶ．

道徳観つかみを生成するにあたり，道徳観を日本語の常識道徳データセットである JCommonsenseMorality [10] を用いる．このデータセットは，道徳的に間違っている文と道徳的に許容できる文の 1 組で構成されており，その文は動作主を主語とし

ボケ	ゴン太です
ツッコミ	あいちゃんです，よろしくお願いします
ボケ	最近な
ツッコミ	はいはい
ボケ	電車の中で，大声で叫んだんよ
ツッコミ	なにをしてんねん！人としてどうなん？

図 2　道徳観つかみの生成例

ない行為で表されている．本研究では非道徳的な文を用いるため，JCommonsenseMorality の内，道徳的に間違っている文を使用する．以下に，入力された日常的な行動から，非道徳的な文を抽出する手法を示す．

（1） 入力された日常的な行動を形態素解析し，名詞，動詞を抽出する．

（2） 抽出された名詞，または動詞を含む非道徳的な文を，JCommonsenseMorality より抽出する．

（3） 抽出された複数の非道徳的な文のうち，1 つをランダムで取得する．

この手法を用いて，漫才台本のつかみ部分の生成を行う．

3.2　Q&A ボケ

「Q&A ボケ」とは，手順にある任意の名詞 A とは異なる名詞 B を答えるというボケである．具体的には，まずツッコミの質問に対して，ボケが手順を答えていくが，ある手順の中にある名詞 A を忘れてしまう．次に，ツッコミが「フリ」として名詞 A が答えとなるヒントのような文をボケに伝え，その名詞 A を答えられるように促す．ここでフリとは，ボケの先行部分でおかしみを効果的に伝達する表現のことである．図 3 の「ほら，電力によってレール上を走る乗り物よ」のように，ボケを強調している文がフリである．最後にボケ側が誤った名詞 B を答え，ツッコミが訂正する．図 3 に「電車に乗る手順」における Q&A ボケの例を示す．

ボケ	プラットフォームに進む
ボケ	ドアが開くまで待つ
ツッコミ	ええ感じやん
ボケ	次に，あれ？なんやっけな？
ツッコミ	ほら，電力によってレール上を走る乗り物よ
ボケ	わかった！「クルーザー」に乗る
ツッコミ	電車やろ！

図 3　Q&A ボケの生成例

Q&A ボケで対象とする名詞は一般的に認知されている名詞であると考える．そこで本論文では，一般的に認知されている名詞に日本語の多肢選択式の常識推論データセットである JCommonsenseQA [11] を用いる．このデータセットは常識推論能力を評価するためのものであり，質問文に対して誤った答え 4 つ，正答 1 つの計 5 つの選択肢からなる．図 4 にツッコミ役のフリと，ボケ役の誤った名詞を抽出する手法を示す．

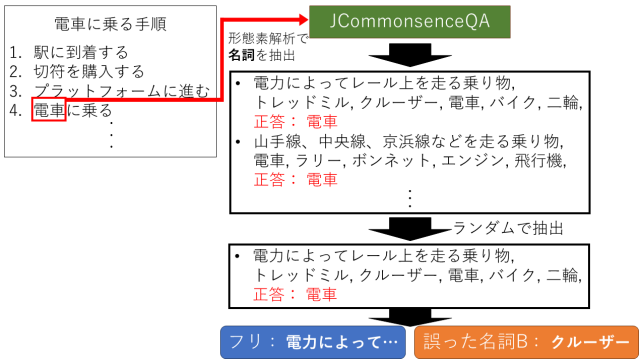


図 4 「フリ」,「誤った名詞 B」の抽出

3.3 高欲求手順欠損ボケ

高欲求手順欠損ボケとは

高欲求手順欠損ボケでは，手順の中で人にとって重要であると考えられる手順に着目する．本研究では，人にとって重要であると考えられる手順は人の欲求が強い行動を含む手順であるとする．人の欲求が強い行動を含む手順に着目した理由は，漫才において人の欲求が強い行動を含む手順が欠けている場合，観客はその手順がないことに気づきやすく，そしてその手順がない事に対する意外性から笑いが誘発されるのではないかと考えたためである．本研究では，人の欲求が強い行動を含む手順を削除するボケを高欲求手順欠損ボケと呼ぶ．図 5 に高欲求手順欠損ボケの例を示す．この人にとって重要であると考えられる手順を決定するため，我々は行動と深く関わりを持つ人の欲求について着目した．例えば,「学校に行く手順」は ChatGPT により以下のように生成される．

（1） 起きる．
（2） 朝食を摂る．
（3） 歯を磨く．
（4） 身だしなみを整える．
（5） 持ち物をバッグに詰める．
（6） 家を出て，最寄りの交通手段を利用して学校に向かう．
（7） 指定された教室や集合場所に向かう．

このとき,「朝食を摂る」という行動は食欲を満たすための行動であり,「食事をとる」という行動は人の欲求が高い行動であると考え，この行動は高欲求な行動が含まれている手順であるといえる．また,「歯を磨く」という行動は健康を維持したい欲求を満たす行動であり，この行動もまた高欲求な行為が含まれている手順であると考えられる．したがって，高欲求手順は欲求を満たすための行動であるといえる．しかしながら人の欲求は多数存在し，欲求の種類によって価値が異なると考えられる．そこで我々は，それぞれの欲求の価値の基準として，マズローの欲求段階説 [12] に着目する．マズローの欲求段階説とは，心理学者のアブラハム・マズローが提唱し，人間の欲求を 5 つの段階に分け，低次の欲求が満たされるとより高次の欲求を求めるという学説である．本研究では，この欲求段階説から欲求度を算出し，この欲求度から高欲求手順欠損ボケを作成し，漫才台本の自動生成を行う．

ツッコミ	電車に乗る手順を順番にゆーてってみ？
ボケ	わかった
ボケ	駅に到着する
ボケ	プラットフォームに進む
ボケ	ドアが開くまで待つ
ツッコミ	「切符を購入する」はどこ行ったんや！一番重要な部分やないかい !!

図 5 高欲求手順欠損ボケの生成例

マズローの欲求段階説

マズローの欲求段階説とは，人間の欲求は 5 段階のピラミッドで表され，低い階層の欲求が満たされるとより高い階層の欲求を求めるというものである．5 つの欲求は図 6 のように，生理的欲求，安全の欲求，社会的欲求，そして自己実現の欲求で構成されている．

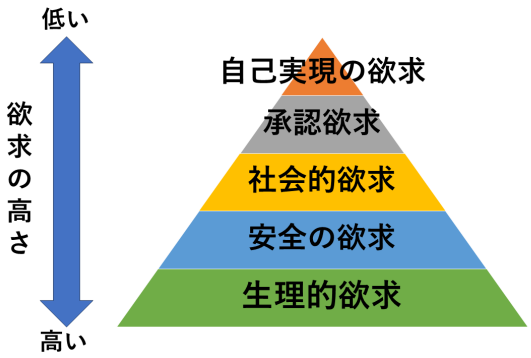


図 6 マズローの欲求段階ピラミッド

生理的欲求とは食欲，睡眠欲，性欲，排泄欲など，人間が生きていくための本能的な欲求のことである．これは人間の生存に不可欠な欲求であるため，あらゆる欲求よりも最優先される．生理的欲求を満たすと，人間は次に安全の欲求を満たしたくなる．安全の欲求とは心身の健康や経済的な安全性など，という欲求である．安全の欲求を満たすと，家庭や会社のように，ある一定のグループに所属していたいという社会的欲求を求める．家族や友人から愛情を受けたいという欲求もまた社会的欲求であり，これらの欲求が満たされないと人は孤独を感じてしまう．この社会的欲求を満たすと，承認欲求と呼ばれる高次の欲求を求めるようになる．この欲求は出世欲など，社会的欲求を満たすために属したグループにおいて自分を認めてもらいたい，尊重されたいという欲求である，そしてこの承認欲求を満たすと，自己実現など自分の人生観・世界観に基づき「あるべき自分」になりたいという自己実現欲求が生まれる．すなわちこの階層モデルは，まず生存に不可欠な要素から欲求が始まり，上の階層になるにつれてより人間らしい欲求が高まっていくことを表している．また高い階層の欲求であるほど生得的な欲求ではないため，欲求の強さに個人差が生じやすい．

欲求度の算出

マズローの欲求段階説から，漫才で用いる日常的な行動手順を構成する文に含まれる 5 つの欲求それぞれの度合いを算出し，手順ごとの欲求度について決定する．この欲求度の決定

を行うのにあたり，我々は手順にある単語ベクトルと上記の 5 つの欲求と関わりを持つ単語ベクトルを生成し，これらの類似度から欲求度を算出する方法を提案する．単語ベクトルには Word2Vec を用いる．Word2Vec は似たように使用される単語は似たようなベクトルを生成するという特徴がある．この Word2Vec の特徴を活かし，手順の単語と 5 つの欲求と関わりを持つ単語との類似から，手順がもつ 5 つの欲求の度合いの算出を行う．5 つの欲求とその関連語は以下のように定義する．

- 生理的欲求　ー食事，睡眠
- 安全の欲求　ー健康，金銭
- 社会的欲求　ー所属，愛
- 承認欲求　ー評価，尊重
- 自己実現の欲求　ー夢，目標

これらの単語と類似性のある単語は，それぞれの欲求と深く関わりを持つと考えられる．以下，上記の定義による欲求の判別手法について述べる．

（1） 手順を形態素解析し，名詞，動詞，形容詞を抽出する．

（2） Word2Vec を用いて，手順で抽出された名詞，動詞，形容詞のベクトルと，5 つの欲求の関連語のベクトルを各々生成する．

（3） 手順中の名詞，動詞，形容詞と 5 つの欲求の関連語とのコサイン類似度を計算し、それらを合計する．

（4） (3) を手順から抽出した名詞，動詞，形容詞の個数で割り，平均を求める．

これにより，手順がもつ 5 つの欲求の度合いを算出する．この手法を用いたとき,「電車に乗る手順」において、それぞれの手順の 5 つの欲求の度合いは図 7 のようになる．

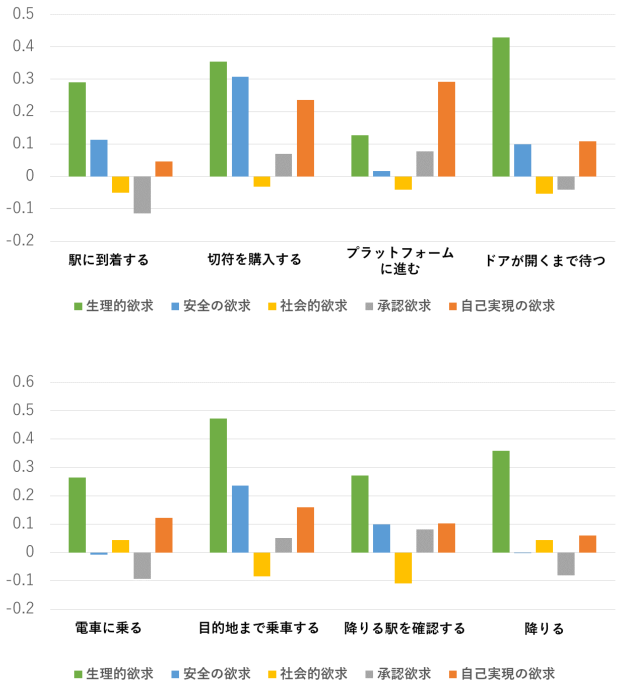


図 7 「電車に乗る手順」のそれぞれの欲求度

高欲求手順の決定

マズローの欲求段階説では，人間は低次の欲求である生理的欲求が一番強い欲求をもち，高次の欲求になるにつれて欲求が弱まっていくことが記されている．したがって本研究では，5 つの欲求に対してそれぞれ重みを設定し，手順の 5 つの欲求度に重みをかけ，これら 5 つの値の合計を求め，これを手順の高欲求度とする．5 つの欲求それぞれの重みは実験により，生理的欲求を 1.5，安全の欲求を 1.4，社会的欲求を 1.3，承認欲求を 1.2，自己実現の欲求を 1.1 とする.「電車に乗る手順」においての高欲求度は図 8 のようになる．

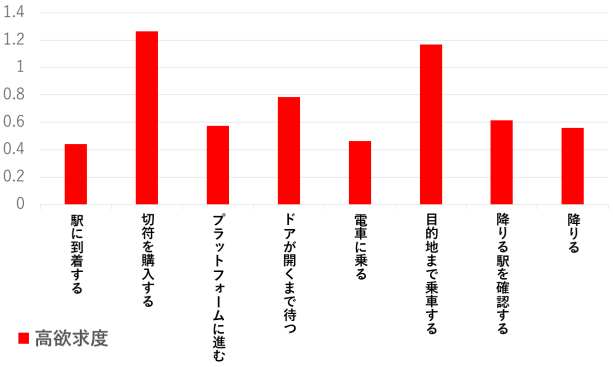


図 8 「電車に乗る手順」のそれぞれの高欲求度

図 8 の通り,「切符を購入する」という手順の高欲求度が高くなっている.「切符」と「購入」は経済欲と関わる単語であり，すなわち安全の欲求と関わりを持つ単語であるといえる．この「電車に乗る手順」において生理的欲求と関わりを持つ単語がないことから，安全の欲求が 1 番低い階層であるといえる．したがって，高欲求手順を抜き出すことができていると考えられる．

4 実　　験

本研究では，提案手法の漫才に対する評価実験，従来手法と提案手法の評価の比較実験の 2 つの実験を行う．

4.1 実験１：提案手法で生成した漫才のユーザ評価実験

本実験では，提案手法で生成した漫才のわかりやすさと面白さの有用性を測ることを目的とし，ユーザの評価実験を行う．具体的には，日常的な行動手順に基づく漫才全体のわかりやすさの評価と面白さの評価と,「道徳観つかみ」,「Q&A ボケ」,「高欲求手順欠損ボケ」のそれぞれに対するわかりやすさの評価と面白さの評価を行う．

実験の方法

実験ではクラウドソーシングを用い，被験者は男性 3 名，女性 7 名の計 10 名である．被験者はオンライン上にて提案手法で生成した漫才の動画を 3 つ視聴し，3 つの漫才それぞれの全体の評価と，道徳観つかみ，Q&A ボケ，高欲求手順欠損ボケの評価をアンケートにて行った．実験はチャットアプリケーションによる動画を用いた．また，この 3 つの漫才のお題はそれぞれ,「野球をする手順」,「電車に乗る手順」,「学校に行く手順」である．

結果と考察

提示した 3 つの漫才ごとのわかりやすさに対する評価を図 9 に，面白さに対する評価を図 10 に示す．図 9 より，どのお題においても漫才がわかりやすかったと回答している人が半数以上を占めている．このことから，提案手法を用いた漫才はわかりやすさに対する有用性を示すことができた．一方で，図 10 より漫才の面白さへの評価が分かれていることがわかる．したがって提案手法を用いた漫才は，一定数の人に対して面白さに対する有用性を示すことはできたが，多くの人が面白いと感じる漫才ではないことがわかった．

次に，道徳観つかみの面白さに対する評価を図 11 に示す．図 11 より,「電車に乗る手順」では半数の人が面白くなかった，あまり面白くなかったと回答していることがわかる．この半数の面白くなかった，あまり面白くなかったと回答している人を分析すると，すべての被験者が女性であった.「電車に乗る手順」において，道徳観つかみのボケは「電車内で女子高生のスカートの中をスマホで撮影した」となっており，これは女性側にとって特にコンプライアンスに大きく反した文であり，不快感を与えるボケとなってしまっている．このことから，度を超えた非常識文をボケとした道徳観つかみは，かえって不快感を与えてしまうことがわかった．度を超えた非常識文を生成しないようにすることは今後の課題である．

続いて，Q&A ボケの面白さに対する評価を図 12 に示す．図 12 より,「学校に行く手順」では，面白くなかった，あまり面白くなかったと回答している人が半数以上を占めている．また，この学校に行く手順での Q&A ボケのわかりやすさに対する評価を分析すると，半数以上の人がわかりにくかった，少しわかりにくかったと回答していることがわかった.「学校に行く手順」においての Q&A ボケでは，手順にある名詞の一つである「朝食」に注目し,「朝食」を答えられるようなヒントとなる文は「朝食べるご飯のこと」，誤った回答は「河童巻き」が生成されている．しかしながら，朝食べるご飯として河童巻きが必ずしも不適切であるとはいえない．そのためユーザが河童巻きを誤った回答として認識しておらず，ボケとして理解できなかったと考えられる．また「野球をする手順」,「電車に乗る手順」においては Q&A ボケの面白さへの評価がわかれている．Q&A ボケが面白かったと回答している人のコメントの中には,「ボケが理解しやすかったから笑うことができた」などがあり，反対に Q&A ボケが面白くなかったと回答している人のコメントの中には,「先の展開が読めてしまい，新規性を感じられなかった」などがあった．このことから，ボケの理解のしやすさに加え，ボケの内容の新規性も重要であることがわかった．

最後に，高欲求手順欠損ボケの面白さに対する評価を図 13 に示す．図 13 より，どのお題においても高欲求手順欠損ボケの面白さへの評価がわかれていることがわかる．この内，面白くなかったと回答した人のコメントでは,「ツッコミが面白くない」などがみられた．高欲求手順欠損ボケのツッコミでは，手順が欠損していることへの指摘のみである．このことから，ツッコミを工夫する必要があることがわかった．

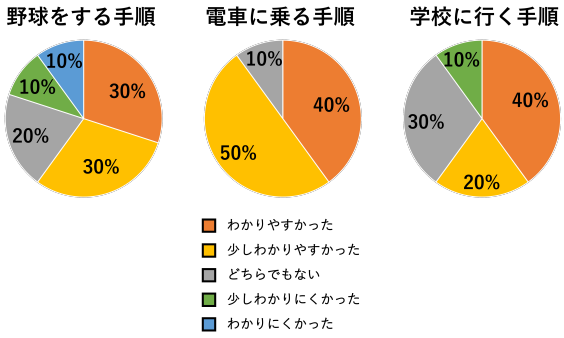


図 9　漫才全体のわかりやすさの評価

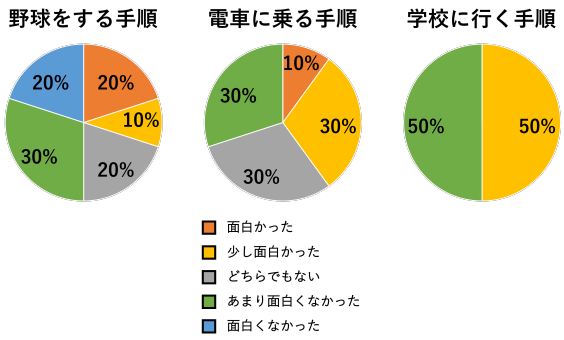


図 10　漫才全体の面白さの評価

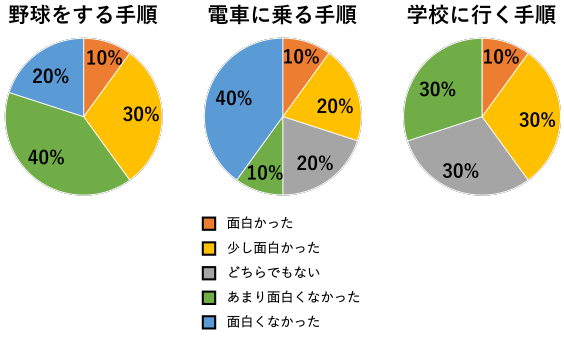


図 11　道徳観つかみの面白さの評価

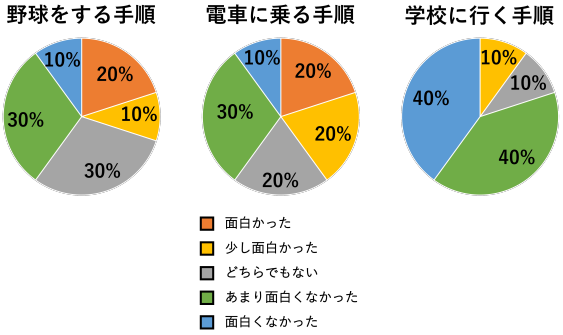


図 12　Q&A ボケの面白さの評価

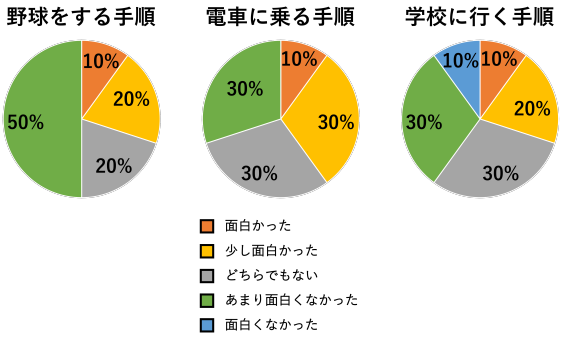


図 13　高欲求手順欠損ボケの面白さの評価

4.2　実験 2：提案手法と従来手法の比較実験

本実験では，提案手法がニュース記事を用いた従来手法より面白いかどうかを確認するため，比較実験を行う．

実験の方法

実験ではクラウドソーシングを用い，被験者は男性 3 名，女性 7 名の計 10 名である．実験に使用した漫才は，従来手法は大谷翔平のニュースに基づく漫才の動画であり，提案手法は「野球をする手順」の漫才の動画である．被験者は 2 つ漫才を視聴し，どちらの漫才がわかりやすかったか，面白かったかをアンケートにて評価を行う．実験はチャットアプリケーションによる動画を用いた．

結果と考察

従来手法の漫才と提案手法の漫才の評価を比較した結果を図 14 に示す．図 14 の漫才のわかりやすさから，提案手法で生成した漫才は従来手法で生成した漫才よりもわかりやすい漫才であることがわかる．したがって実験 1 の評価実験と同じく，漫才のわかりやすさに対する有用性を示すことができた．一方，図 14 の漫才の面白さから，従来手法と提案手法で評価が分かれる結果となった．従来手法の漫才の方が面白いと回答したコメントでは，「漫才の手数が多くてよかった」・「提案手法の漫才のオチが面白くなかった」などがみられた．このことから，ボケの種類の多さがボケの意外性を引き出すことがわかった．また漫才を締めくくるオチは，漫才全体の評価に大きな影響を及ぼすことがわかった．

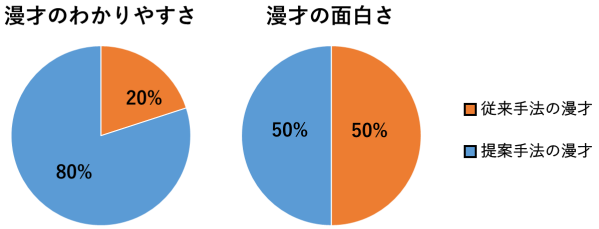


図 14　従来手法と提案手法の比較

5　まとめと今後の課題

本研究では，日常的な行動の手順を基に漫才台本自動生成を提案した．具体的には，入力された日常的な行動の中での常識に反する行動をするという小ボケをつかみとする道徳観つかみ，手順の中の任意の名詞とは誤った名詞を答える Q&A ボケ，欲求度に着目した高欲求手順欠損ボケ，そしてその高欲求手順を決定するための欲求度算出手法などを提案した．次に，提案手法の評価実験，従来手法と提案手法の評価の比較実験を行った．実験より，提案手法の漫才はわかりやすさにおいて有用性があるが，漫才のおもしろさの有用性を示すことはできなかった．理由として，道徳観つかみではコンプライアンスに大きく逸脱した非常識文を抽出してしまっていること，高欲求手順欠損ボケではツッコミがボケの指摘のみであることなどが考えられる．また，ボケの種類の少ないことやオチの工夫がなかったことも課題である．提案手法の漫才の面白さを改善するため，今後は道徳観つかみにおいて適切な非道徳文の抽出，高欲求手順欠損ボケにおいてツッコミの工夫を行う予定である．また新たなボケコンポーネントの提案や面白いオチの提案も行う予定である．

謝　　辞

論文の一部は私学助成金（大学間連携研究助成金）の助成によるものである．ここに記して謹んで感謝の意を表する．

文　　献

[1] 原口和貴, 梅谷智弘, 北村達也, 灘本明代, “Seq2Seq を用いたボケの自動生成手法の提案”, DEIM2021,H24-3.
[2] 家根和希, 原口和貴, 佐藤亮, 荒牧英治, 宮代勲, “チャット型漫才アプリケーションの提案”, DEIM2020,A3-3.
[3] 下崎安紋, 北村達也, 梅谷智弘, 灘本明代, “漫才のわかりやすさを考慮した漫才台本自動生成”, DEIM2023,5b-2.
[4] 吉田裕介, 萩原将文, “漫才形式の対話文自動生成システム”, 日本感性工学会論文誌,Vol.11,No.2,pp.265-272,2007
[5] 岩倉亮介, 吉川大弘, ジメネスフェリックス, 古橋武, “Twitter データを用いたユーモア語句自動生成手法に関する一検討”, 第 31 回人工知能学会全国大会,2017
[6] Mikolov,Tomas,et al. “Efficient estimation of word representations in vector space.”, arXiv preprint arXiv:1301.3781,2013
[7] 原直大, 小林明美, “ChatGPT を利用したユーモア応答生成の分析”, 言語理解とコミュニケーション研究会 (NLC), 信学技報,vol. 123,no. 176,NLC2023-8,pp. 41-46,2023

[8] 薛強, 滝口哲也, 有木康雄, “感情・欲求に基づく主観性を持つ雑談対話システム”, 言語処理学会 第 26 回年次大会,P4-19,2020
[9] 安部達雄, “漫才における「フリ」「ボケ」「ツッコミ」のダイナミズム”, 早稲田大学大学院文学研究科紀要. 第 3 分冊, 日本文学演劇映像美術史日本語日本文化, Vol.51, No.28, pp. 69-79, 2006
[10] 竹下昌志, ジェプカ・ラファウ, 荒木健治, “JCommonsenseMorality: 常識道徳の理解度評価用日本語データセット”, 言語処理学会 第 29 回年次大会,D2-1,2023
[11] 栗原健太郎, 河原大輔, 柴田知秀, “JCommonsenseQA 2.0: 計算機と人の協働による常識推論データセットの改良”, 言語処理学会 第 29 回年次大会,A12-4,2023
[12] Abraham Harold Maslow, “Motivation and Personality”, pp.35-57,1987.

付　　録

ボケ	ゴン太です。
ツッコミ	あいちゃんです。よろしくお願いします。
ボケ	最近な、
ツッコミ	はいはい。
ボケ	野球の試合に負けたから、砂を投げつけたんよ。
ツッコミ	何してんねん !! 人としてどうなん !?
ボケ	ボクロボットデスケド？
ツッコミ	やかましいわ。とにかくもうやらないでくださいね。
ボケ	それでまぁ本題やねんけど、
ボケ	ちゃんと野球をすることができるようにしたい !!
ツッコミ	なるほど。うん。できるようには思えんけども。
ボケ	ちょっと野球をする手順を確認してもらっていい？
ツッコミ	わかった。じゃあとりあえずじゅんばんにゆーてみ？
ボケ	了解！そうやな〜どこから言おうかなぁ....
ツッコミ	は？最初からゆーたらええやん。
ボケ	いやでも思いつく手順の数が 1200 個もあるからさぁ...
ツッコミ	多すぎるやろ !! 一歩歩いたら 1 手順じゃないねん。
ボケ	あ、そうかそうか。じゃあ読むで？
ボケ	グローブをつける、
ボケ	ピッチャーがマウンドに立つ、
ボケ	バッターがホームプレートに立つ、
ツッコミ	ええ感じやん！
ボケ	次に、ピッチャー、
ボケ	あれ？なんやったっけな...？
ツッコミ	ほら、いろいろなスポーツで投げるものよ。
ボケ	そや！「靴」を投げる
ツッコミ	ボールやろ !!
ツッコミ	靴は足を包む履物やろ！
ツッコミ	途中からでもええから、もう一回順番ゆーてみ？
ボケ	わかった。
ボケ	チームを組む、
ボケ	グローブをつける、
ボケ	ピッチャーがマウンドに立つ、
ツッコミ	ボールとバットを用意するはどこ行ったんや !!
ツッコミ	一番重要な部分やないかい！
ツッコミ	ていうか手順手順ゆーてるけどやな、あんた手あるんか？
ボケ	あ.....
ボケ	バッテリーがなくなりました
ツッコミ	逃げられへんよ？もうええわ。

図 15 「野球をする手順」の漫才台本

ボケ	ゴン太です。
ツッコミ	あいちゃんです。よろしくお願いします。
ボケ	最近な、
ツッコミ	はいはい。
ボケ	電車内で女子高生のスカートの中をスマホで撮影したんよ。
ツッコミ	何してんねん !! 人としてどうなん !?
ボケ	ボクロボットデスケド？
ツッコミ	やかましいわ。とにかくもうやらないでくださいね。
ボケ	それでまぁ本題やねんけど、
ボケ	ちゃんと電車に乗ることができるようにしたい !!
ツッコミ	なるほど。うん。できるようには思えんけども。
ボケ	ちょっと電車に乗る手順を確認してもらっていい？
ツッコミ	わかった。じゃあとりあえずじゅんばんにゆーてみ？
ボケ	了解！そうやな〜どこから言おうかなぁ....
ツッコミ	は？最初からゆーたらええやん。
ボケ	いやでも思いつく手順の数が 800 個もあるからさぁ...
ツッコミ	多すぎるやろ !! 一歩歩いたら 1 手順じゃないねん。
ボケ	あ、そうかそうか。じゃあ読むで？
ボケ	駅に到着する、
ボケ	切符を購入する、
ボケ	プラットフォームに進む、
ツッコミ	ええ感じやん！
ボケ	次に、
ボケ	あれ？なんやったっけな...？
ツッコミ	ほら、部屋にある、出入りするのに必要な部分の事よ。
ボケ	そや！「障子」が開くまで待つ
ツッコミ	ドアやろ !!
ツッコミ	鬼滅の刃かあんた
ツッコミ	途中からでもええから、もう一回順番ゆーてみ？
ボケ	わかった。
ボケ	駅に到着する、
ボケ	プラットフォームに進む、
ボケ	ドアが開くまで待つ、
ツッコミ	切符を購入するはどこ行ったんや !!
ツッコミ	一番重要な部分やないかい！
ツッコミ	ていうか手順手順ゆーてるけどやな、あんた手あるんか？
ボケ	あ.....
ボケ	Siri、あとは頼んだで
ツッコミ	形すらないやないか！もうええわ。

図 16 「電車に乗る手順」の漫才台本

意思決定における大規模言語モデルを用いた検索要求の観点抽出の影響分析

笹本まこと† 堀川 達平†† 北山 大輔††

† 工学院大学情報学部システム数理学科 〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学大学院工学研究科情報学専攻 〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{jx20114,em23040}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし Web 検索は日常的に行うが，膨大な情報をもつインターネット上から目的の情報を探し出すことは難しい場合がある．とくに旅行先の検討や商品の購入など，複数の情報源から比較検討をして，ユーザが求めている情報を探すことは難しくまた手間である．この時，手間を解消できると望ましいと考えられるが，簡単に発見できた場合に，質の高い意思決定ができているとは限らない．そこで我々は,「Web 検索で複数の対象を比較して検討する場合について，入力のヒントを得ることで，客観的に見て意思決定の質が向上する」と考えた．我々は，入力された検索要求に対し，大規模言語モデルを用いて，検索要求を達成するための観点を示す文を生成し，検索結果を見せる前にその観点を見せる手法を提案する．本稿では，この手法に対し，Web の検索エンジンで通常の検索を行うとき，検索における観点を確認してから Google 検索をするとき，AI チャット機能を利用して検索を行うとき，それぞれの検索行動として，意思決定にかかる時間，意思決定のための検討アイテム，最終的な意思決定内容について比較する実験を行う．

キーワード 情報検索，検索クエリ，観点抽出，検索行動，大規模言語モデル，GPT

1 はじめに

我々はインターネットを利用すると，ジャンルを問わず調べ物をすることができる．検索窓に関連するキーワードを入力して検索ボタンを押すと，入力に関連するサイトが複数表示される．多くの人がスマートフォンを持ち歩くようになった今，いつでもどこでも，誰もが簡単に情報検索を行うことができるようになった．インターネットが普及し始めてから現在に至るまで，情報検索に関するさまざまな研究が行われてきた [1].

インターネット上には古いものから新しいものまで，大量の情報が蓄積されており，検索窓に関連するキーワードを入力したからといって，最適な情報をすぐに見つけ出すことができるとは限らない．検索した結果として表示される類似した複数の情報の中から，自分にとって最適な情報を見つけ出すことは難しい．通常の Web 検索の場合，ユーザはこの検索結果の中から自身に適したものを探し出す必要がある .

一方，近年では ChatGPT をはじめとする大規模言語モデルを用いたチャットが発達し，大規模言語モデルと情報検索についても研究が行われている．口コミから特徴語を抽出し，言語モデルを用いて検索時の入力を補完する研究や，商品説明文を入力として与えてレビューを生成する研究などがある [2] [3]. Bing Chat では，チャット内容を Web 検索を併用して補足する機能が標準で搭載されている．このような AI チャットを Web 検索の代わりに利用することも可能となってきている．しかし，AI を用いた対話形式での Web 検索では過剰なヒント，もしくは検索結果が要約され単一の答えが与えられてしまう．このような対話形式で得られた情報は自らの力で検索結果に辿り着いていないことから，質の高い意思決定を行うことができる可能性が低いと考えた.

本研究では,「Web 検索で複数の対象を比較して検討する場合について，入力のヒントを得ることで，客観的に見て意思決定の質が向上する」と仮定し，この仮説の妥当性を検証する．そこで，本研究では，ユーザが入力した自然言語による検索要求に対し，大規模言語モデルを用いて，その要求を達成するための観点を生成し，検索開始前にユーザに提示する手法を提案する．これにより，適度な検索のヒントを与えつつも検索行動は自身の手で行うことができ，手間を軽減しつつ，質の高い意思決定が行えるよう支援する．

通常の Web 検索をする手法，AI チャットを用いた Web 検索手法とで検索結果や意思決定にかかる時間などを比較する．また，最終的な意思決定内容の評価も行う．

2 関連研究

2.1 検索行動の分析

検索行動を詳細に分析した桑折らの研究では，既存の検索エンジンを用いた情報検索におけるユーザの行動分析を行った [4]. 被験者に 2 枚の画像を提示し，その写真の場所を特定する問題を出題した．画像をクエリとしてどのように表現すれば良いか明確でない場合，検索者が自ら回答への道筋を考える必要があり，この時の検索行動について，検索意図の分類や場所の特定にかかった時間，検索クエリの分析を詳細に行っている．

また，オンラインショッピングという機会における検索行動の分析の研究もある．柳田らは，EC サイトにおける商品購買の前後の Web 情報探索行動と購買満足度の関係を調査した [5]. Web 検索ログを購買満足度の評価として EC サイトにおける商品レビューの点数で分析を行った．EC サイトで商品を購入

したユーザと検索意図に基づいて Web 検索クエリを分類し，購買満足度が異なるユーザの時間分布を調査した結果，特に，購買直前に幅広い商品に関する情報を検索したユーザは満足している場合が多く，また，具体的なクエリで検索していた．他人の意見を求めている顧客の満足度も高かった．顧客が行った Web 検索の種類が異なれば，購買満足度も異なる可能性があることを実証した．

これら先行研究と同様に，本研究でも検索行動の分析や満足度の評価を行うが,「質の高い意思決定を行う」ために適した検索の事前情報について議論する点で異なる．

2.2 検索の支援やクエリ推薦

検索行動を支援する研究，特にクエリ推薦に関しても，様々な研究が行われている．大重らの研究では，協調検索支援手法を提案している [6]．グループで意思決定を行うとき，複数のメンバーで検索行動が重複することで検索範囲が狭まってしまったり，各メンバーが好みに合わせて検索を行うと，検索範囲がバラバラであるためにグループとしての意思決定が難しくなってしまう．この問題の解決のために，グループメンバーの検索を発散・収束させるクエリを推薦することで検索行動を支援するシステムを提案した．実験の様子を撮影した映像と操作ログ，アンケート，クエリの精度の評価についてを分析し，評価を行った．また，ウェブ検索に慣れていないユーザにとって，検索クエリを立てることは難しく，適切な検索クエリを作成しなければユーザの情報欲求は満たされないことから，小竹らは，一般的なウェブ検索 UI と穴埋め形式の検索 UI を用いて，検索 UI の違いによる検索プロセスの差異を明らかにした [7]．あらかじめ提示した 2 種類のトピックについて検索行動を収集した結果，穴埋め形式の検索 UI を用いた検索は，負担がかかるものの，具体的なクエリ生成を可能にすることを確認した．

一般的によく知られている情報が得られるクエリが推薦されやすく，新たな知識を獲得できないことがある．鈴木らは，未知の情報を提示するために意外性を示す Serendipity に着目し，思いつきづらく，予想できない情報を得る目的で，意外性のある検索クエリをユーザに推薦する方法を提案した [8]．足立らは，検索クエリを用いた場所の関連語の取得について研究を行った [9]．ユーザの知りたい情報は各場所で異なるため，情報提供者は場所に応じた情報を準備する必要があるとし，ユーザの関心事が反映されている検索クエリログを用いて，各場所におけるユーザの関心事の傾向を把握する．場所名と共起する共起語に限定して，共起語をカテゴリとしてまとめ，各場所でのカテゴリの割合を求めることで，場所における関心事の傾向を把握する手法を提案した．大石らは検索エンジンに与えるクエリを改善するクエリ拡張の研究を行った [10]．検索では，検索エンジンによって得られた結果がユーザの必要としてる情報ではないことがしばしばある．この問題解決のために，クエリ拡張を行う．この研究では，ユーザの意図する語を生成するために，関連単語抽出アルゴリズムを提案した．

本研究は，上記研究と検索の支援を行うという点で類似しているが，これらの研究では UI やクエリに対する直接的なアプローチを行っているのに対し，本研究では，質の高い意思決定を行うために，検索のヒントを与えるという間接的なアプローチを行う．

2.3 検索テーマの観点抽出

本研究では，大規模言語モデルにプロンプトを与えることで，検索テーマを検索する上で着目するべき観点の提示を行うが，先行研究として，検索テーマの観点に着目した研究が存在している．山本らは，人と人による質問と回答のやりとりがある質問応答サイトに着目し，蓄積された質問と回答の情報を利用してユーザのウェブ検索を支援する方法を提案した [11]．具体的には，ユーザの興味を引く単語を提示するために修飾語付き観点に着目し，その観点から Q&A コンテンツより語を抽出してランキングする手法である．よりユーザの興味を引くことができることと旅行カテゴリにおいての有効性が特に高いことを確認した．本研究の提案手法では，検索クエリの候補になるような観点を大規模言語モデルを用いて生成し，検索者に提示する．検索者の検索テーマに関する知識が少ない場合，ここで提示する観点は，より絞り込んだ出力結果にできると考えられ，検索の支援になると考えている．

3 提 案 手 法

本提案手法では，ユーザの最初の入力をユーザプロンプトとして大規模言語モデルに与え，関連する文書を生成させたものを観点とする．そして，この観点を検索におけるヒントとして検索者に提示する．これは他手法と比較して時間はかかるかもしれないが，より自分のニーズに合った情報を見つけ出すことができると考える．

提案手法では，事前に観点を提示する．本実験における「観点」とは，検索テーマに関して調べ物をする時に考慮するべき点を指す．観点の抽出には，OpenAI が提供する GPT-3.5-turbo-1106[1]（以後，単に GPT と記す）を用いる．system プロンプトは,「入力された内容に関する情報やアドバイスを 5 個教えてください。その際、それぞれに 1 単語の見出しが欲しいです。(例 ○○:〜)」で，ここで生成された文章を検索のヒントとなる観点とする．このプロンプトにおける入力とは，検索テーマそのものであり，提示する観点は，検索したい内容に関する知識の浅い人にとって，検索時の入力を具体的な内容にできるため，より良い検索結果を得る効果があると考えられる．

ユーザが検索キーワードを入力すると，その入力に基づいて，検索結果の代わりに観点が生成され，表示される．その後，ユーザは，通常の Web 検索および閲覧行動を行うこととなる．提案手法の概要図を図 1 に示す．

4 実　　験

4.1 実 験 方 法

ある検索テーマに対し，通常の Web 検索をするとき（比較

1：https://platform.openai.com/docs/api-reference/chat

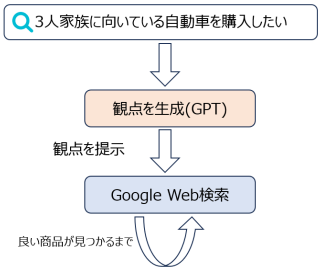


図 1　提案手法の概要図

表 1　検索テーマ

A さんは 30 歳です。妻、2 歳の子供と暮らしています。今まではカーシェアサービスなどを利用していましたが、休日のお出かけや子どもの送り迎え、買い物など、車に乗る機会が増えたため、自動車の購入を検討しています。あなたが A さんの場合、どの車を買いますか。（具体的な車種を回答ください）

表 2　提示された観点の例

予算: 予算に合った車を選ぶことが重要です。 サイズ: 3 人家族には広々とした車内が必要です。 安全性: 安全装備が充実している車を選ぶことが大切です。 燃費: 経済的な燃費の良い車を選ぶことでランニングコストを抑えられます。 保険料: 保険料が安く抑えられる車を選ぶことも考慮してください。

手法 1），対話型 AI を用いて Web 検索を行うとき（比較手法 2），検索における観点を確認した上で Web 検索をするとき（提案手法）の計 3 手法で検索行動の比較を行う．被験者はいずれか 1 手法で検索を行い，最終的な意思決定までの検索行動の要約文を記入する．これにより各手法での検索行動を収集する．さらに，収集した検索行動を別の被験者により一対比較法で評価する．一つのタスクでは，検索者が記入した 2 つの検索行動の要約文を読み,「『より良い決定ができた』と感じるのはどの手法か？」について評価する．

本実験には，世界でもっともユーザの多い Google の検索エンジンと Microsoft 社が提供する Bing の AI チャット機能を用いる．Bing とは,「利用者の意思決定を支援する」というコンセプトの検索エンジンで，2023 年 2 月には OpenAI が提供する人工知能を利用したチャット技術を導入し，注目されている．

本実験では，検索テーマを表 1 のように設定した．このとき，3 手法とも最初の入力は「3 人家族に向いている自動車購入をしたい」で，提案手法で提示した検索における観点の例は表 2 である．クラウドソーシング [2]で募集した 30 名の検索者は，テーマに沿って検索を行い，テンプレートにしたがって検索行動の要約文を書く．これには，検索窓に何を入力したか，検索結果として表示された情報の中で候補にあがった商品や，何を最も魅力的に思ったかを記入する．30 個の検索要約文をの組を 1 つずつ比較し評価するために，クラウドソーシングで募集したのべ 1305 名は，2 個の検索行動の要約文を読んだ上で最終決定について「より良い決定ができているかどうか」の評価を行った．

2：クラウドワークスを用いた．https://crowdworks.jp/

表 3　比較手法 1 で得られた検索要約文の例

1 回目のキーワード：子供連れ ファミリーカー スズキ ハスラー 2 回目のキーワード：3 人家族 ファミリーカー トヨタ アルフォード 検討した候補商品：スズキ ハスラー、トヨタ アルファード 最終的に決定した商品：トヨタ アルファード 理由：1 人子供だが今から増えるかもしれないし、子供用の自転車も入りそうなので

4.2　比 較 手 法

GPT が生成した観点を検索のヒントとしてあらかじめ検索者に提示する提案手法と，比較するための手法 1，2 は以下のとおりである．

- 比較手法 1：最初の固定の入力を GoogleWeb 検索に入力した結果が表示される．そのため検索開始時の画面は Google Web 検索となる．以降は，被験者が商品決定のための必要な情報収集を行う．
- 比較手法 2：最初の固定の入力を Bing の AI チャットに入力した結果が表示される．そのため，検索開始時の画面は Bing の AI チャットとなる．以降は，被験者が商品決定のための必要な任意の情報収集を行う．

比較手法 1 は，多くの人が行う一般的な検索である．Google に備わるすべての検索支援機能を使用できる状態であった．比較手法 2 では Bing の AI チャットを利用する．AI チャットを用いた検索というのは，Web 検索結果が要約されたものが出力される．そのため，ユーザが望む情報をより早く見つけることができるかもしれないが，最初から絞られた情報が提示されている分，検索者にとって最適な結果から外れたものが提示される可能性もある．対話形式で調べ物を進めることができるため，通常の検索行動とは異なる特徴が見られると考えた．

提案手法，比較手法 1，比較手法 2 のいずれの手法も，最初の入力は「3 人家族に向いてる自動車を購入したい」で，この検索結果の閲覧より後の行動は検索者に任せることとした．設定に合った商品を見つけるまで検索を繰り返した．例えば，比較手法 2 の検索行動として，最初の検索は Bing の AI チャットで行うが，その後は Google を利用した被験者もいた．

4.3　実験結果と考察

提案手法と比較手法の 3 手法で手法毎に 10 名が，表 1 の検索テーマに沿って検索を行った．商品の決定までにかかった時間の平均は，比較手法 1 が 521.1 秒，比較手法 2 が 723.4 秒，提案手法が 918.5 秒であった．また，検索回数の平均は，比較手法 1 が 3.2 回，比較手法 2 が 4.3 回，比較手法 3 が 4.4 回であった．実際の 30 個の検索要約文のうち結果の例を表 3，表 4，表 5 に示す．

観点が提示される提案手法は 3 手法の中でもっとも意思決定までに時間がかかっており，検索回数が多い傾向がある．観点が提示されることによって，幅広い知識を得てから絞り込むような検索がしやすくなっていると考えられる．

AI チャットで検索を行う比較手法 2 は商品決定までにかかる

表 4　比較手法 2 で得られた検索要約文の例

1 回目の検索キーワード：ソリオ、ヴォクシー、N-box の特徴と価格は？
2 回目の検索キーワード：トヨタ ヴォクシーの価格帯は？
3 回目の検索キーワード：ソリオの動力性能
4 回目の検索キーワード：ソリオ、ヴォクシー、N-box の排気量
検討した候補商品：ソリオ、ヴォクシー、N-box
最終的に決定した商品：ソリオ
理由：サイズ的にコンパクトだが小さすぎず、マイルドハイブリッドなどの動力性能も良好。値段も手ごろなので年齢や家族構成から推測できる経済状態にもあっていると思いました。

表 5　提案手法で得られた検索要約文の例

1 回目の検索キーワード：3 人家族　父、母、子供一人　ファミリーカー　相場
2 回目の検索キーワード：3 人家族　父、母、子供一人　ファミリーカー　トヨタ
3 回目の検索キーワード：3 人家族　ファミリーカー　おすすめ
4 回目の検索キーワード：3 人家族　ファミリーカー　おすすめ 燃費が良い
5 回目の検索キーワード：3 人家族　ファミリーカー　アクア　ライズ　ヤリス
6 回目の検索キーワード：アクア　ライズ　ヤリス 燃費が良い車
7 回目の検索キーワード：トヨタ　ヤリス　性能　価格 ガソリン車
8 回目の検索キーワード：トヨタ　ヤリス　WLTC モード
9 回目の検索キーワード：トヨタ　ヤリス　安全性　運転のしやすさ　車内の広さ
検討した候補商品：トヨタ: アクア ライズ ヤリス
最終的に決定した商品：トヨタ: ヤリス
理由：事前に決めていた予算内に収まる車だったのと、燃費の良さや車のデザインの良さも良いと思ったから。車内も家族 3 人で使う文には十分な広さで荷室も広く設計されていたため、買い物の際に車内が手狭になる心配もなさそうだと思ったから。

表 6　一対比較法による選択結果

被験者番号	手法	選択数
12	提案手法	66
19	提案手法	65
28	提案手法	65
14	比較手法 1	64
11	提案手法	62
2	比較手法 1	61
8	比較手法 2	61
23	比較手法 2	61
27	比較手法 2	59
9	提案手法	53

表 7　各手法の中央値と上位件数

	比較手法 1	比較手法 2	提案手法
選択数中央値	35	47	51
選択数順位上位件数	2	3	5

平均時間が，観点提示をする提案手法の次に長く，検索回数も多い傾向があった．チャットでユーザの知りたいことが入力されると，次々に候補となる車の一覧などの情報を提示する Bing では，ユーザの探求意欲が深まっていると考えられる．

また，第三者による「良い決定が行えているかどうか」の評価も行った．30 個の検索要約文を 2 個ずつペアにした 435 パターンの組み合わせを，それぞれ 3 人ずつが「どちらが良い決定を行っているか」比較し評価を行った．結果を表 6 に示す．また，各手法の選択数の中央値と選択数順位で上位 10 位以上となる検索行動の件数を表 7 に示す．中央値は提案手法がもっとも高く，選択数上位 10 件中 5 件が提案手法による検索要約文であるという結果より，提案手法は客観的に見て良い選択ができる検索行動をとるようになることが示唆される．

個別の結果について考察を加える． 良い決定が行えていると評価された上位 2 件，被験者番号 12 番と 8 番の検索要約文を，表 8，表 9 に示す．上位 2 件表 8 の検索要約文は，提案手法で行われた検索であった．良い決定と評価された理由には,「幅広い候補から徐々に選択の幅を狭めている」,「先入観なく，ニーズに合った車種を絞り込めている」,「様々な条件でより詳しい情報を集めて，判断材料を増やしているから」などの理由が挙げられた．与えられた観点によって，様々な角度からテーマに適した自動車を絞り込んでいることが高評価の理由であることを確認した．この検索にかかった時間は 2466 秒であった．

表 9 は評価が 2 位だった被験者番号 8 番の検索要約文である．こちらも提案手法で検索を行っており,「少しずつ検索ワードを変えて，家族向きの車を探す検索が行われている」,「得られた情報量が多く，複数の車種をしっかり見極めて判断している」，などの理由から，良い決定が行われたと判断された．検索にかかった時間は 1409 秒であった．

評価が最も低かった被験者番号 30 番の検索要約文を表 10 に示す．この検索行動の特徴は，最初から特定の車種を検索しているところである．実際に評価者によって挙げられた低評価の理由として,「車種のみの検索で，家族のためより，自分がその車に乗りたいだけに感じる」,「条件が少なすぎて比較行動という面で有用とは言えない」などがあった．検索キーワードに用いた 2 つの車種以外にも，他にまだニーズを満たす車が存在する可能性がある．

表 11 は次に評価が低かった被験者番号 5 番の検索要約文である．こちらもまた，1 回目の検索から特定の車種を検索キーワードとしている．最終的には複数の車種を比較できているが,「検索範囲が狭い」などの理由で低評価であった.「より良い決定ができている」という点で評価する時には，最初から特定の車種に絞り込む検索は向いていない．評価が低かった検索行動は，検索にかかった時間も短い傾向にあった．

観点提示を行った提案手法では，他手法に比べて検索回数が多く，意思決定までにかかる時間が長い傾向が見られたが，入力内容を変更しながら絞られた情報の中から複数回の選択を行っている．第三者による検索要約文の評価は，最大選択数，選択数中央値で見ても 3 手法の中でもっとも高い値であった．このことから，ヒントが与えられた提案手法は,「Web 検索で複数の対象を比較して検討する場合について，入力のヒントを得ることで，客観的に見て意思決定の質が向上する」という仮説通りであったことを確認した．

AI チャットを用いた比較手法 2 は，通常の検索である比較手

表 8　最も評価が高かった検索要約文 (提案手法)

1 回目の検索キーワード：3 人家族　子供　ちょうど良い　車
2 回目の検索キーワード：3 人家族　車種　オススメ
3 回目の検索キーワード：3 人家族　安全性　車種
4 回目の検索キーワード：3 人家族　積載量　車種
5 回目の検索キーワード：3 人家族　維持費　車種
6 回目の検索キーワード：3 人家族　税金　車種
7 回目の検索キーワード：3 人家族　コンパクトカー
8 回目の検索キーワード：3 人家族　フィット　アクア　どっち
検討した候補商品：ホンダ フィット、トヨタ アクア
最終的に決定した商品：ホンダ フィット
理由：3 人家族であることや自動車税、燃費などの維持費の節約と安全性を考慮してコンパクトカーに決めました。コンパクトカーのランキングを見て、人気かつメーカーの信頼性からトヨタかホンダに目が引きました。ホンダ フィットとトヨタ アクアで迷いましたが、ハイブリット車の燃費やイメージの良さからホンダ フィットに決定しました。

表 9　2 番目に評価が高かった検索要約文 (提案手法)

1 回目の検索キーワード：ファミリーカー　3 人家族　おすすめ
2 回目の検索キーワード：ファミリーカー　低燃費
3 回目の検索キーワード：ファミリーカー　3 人家族　快適
4 回目の検索キーワード：ファミリーカー　売れ筋
4 回目の検索キーワード：タント　ルークス　比較
検討した候補商品：ダイハツ タント、日産 ルークス
最終的に決定した商品：ダイハツ タント
理由：比較したルークスよりもわずかに燃費が良く、車体価格も安かったこと。そして助手席側の柱をドアに埋め込んだミラクルオープンドアという後部座席のスライドドアと合わせてドアをガバっと大きく開けることができる機能を備えていたことによって、乗り降りや荷物の出し入れがしやすく、家族で出かけるのに向いている車だと思ったから。

表 10　最も評価が低かった検索要約文 (提案手法)

1 回目の検索キーワード：プリウス
2 回目の検索キーワード：カローラ
検討した候補商品：カローラ
最終的に決定した商品：カローラ
理由：適度にコンパクトで積載性も備えていて、燃費も良く価格も適正なので、家族のために安全を重視したカーライフを過ごすのに一番適した車だと思いました。

表 11　2 番目に評価が低かった検索要約文 (比較手法 2)

1 回目の検索キーワード：ホンダ N-Box
2 回目の検索キーワード：ホンダ N-Box の価格を教えてください
3 回目の検索キーワード：スズキ ソリオの価格を教えてください
4 回目の検索キーワード：家族向けの軽自動車
検討した候補商品：ダイハツ　タント、ホンダ N-BOX、スズキ ソリオ
最終的に決定した商品：スズキ
理由：3 人家族なら軽自動車でも良いのでは通った。価格も手ごろで初めて買う車として買い求めやすい。

法 1 と比較して回数が多く，検索にかかる時間も長かった．AI が Web 上の情報を要約して提示することで，楽に情報収集を行うことができるが，まとめサイトや比較を行うためのページが羅列されることがあるため，情報の閲覧と同時に疑問も増えていると考えられる．

今回の実験からは，より良い決定を行うためには，検索回数の多さや検索にかかる時間が高評価に比例する傾向があることがわかった．検索回数が多いと時間もかかるが，丁寧にさまざまな角度から調べることで質の高い意思決定ができる．

5　まとめと今後の課題

検索は容易に行うことができるが，検索結果は入力内容のわずかな差によっても異なるため，膨大な情報が蓄積されているインターネット上で目的の情報を見つけ出すことは難しいことがある．

本研究では「Web 検索で複数の対象を比較して検討する場合について，入力のヒントを得ることで，客観的に見て意思決定の質が向上する」と仮定し，この仮説の妥当性を検討した．

Google の検索エンジンで通常の検索を行うとき，検索における観点をあらかじめ確認した上で Google 検索をするとき，Bing の AI チャット機能を利用して検索を行うときの計 3 手法で検索行動を比較した．3 手法を比較した結果，入力内容と検索結果や意思決定にかかる時間，最終的な意思決定など，検索行動にそれぞれ差が見られることを確認した．手間はかかるものの，より良い決定ができたという点では，観点を確認した上で検索する手法がもっとも効果的だということがわかった．観点を提示する提案手法は検索回数が多く，時間もかかる傾向が見られた．提案手法は，他手法と比較して時間はかかるかもしれないが，より自分に合った情報を見つけ出すことができることを確認した．本実験では，検索観点の生成に GPT を用いた．GPT が生成した観点は，Web 検索で複数の対象を比較して検討する場合について，客観的に見て意思決定の質を向上させることに貢献した．

今後は，観点を GPT に生成させる方法以外の新たな観点生成手法について検討したり，自動車選び以外の検索シチュエーションで検索行動を比較したい．自動車のように高級で頻繁に購入するものではなく，例えば，家電や日用品など価格や購入頻度が異なる商品を選ぶ検索シチュエーションで実験をする．また，観点を提示する提案手法は，検索を苦手とする人への効果がより高いと考えている．このため，検索を行う被験者を検索能力別に分類し，分析する実験も行う必要もあると考えている．

6　謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 高久雅生．進化する情報検索技術．情報の科学と技術, Vol. 69, No. 5, pp. 182–188, 2019.
[2] 笹本まこと, 堀川達平, 北山大輔. 大規模言語モデルのためのサー

ビスに対する口コミを用いた入力自動補完の検討. *D-013*, Vol. 2016, No. vol.78,no.1, p. 02, 2016.
[3] 山本 岳洋 田中 克己 Martin J. Durst 沈 辰夫. 商品のユーザレビューを利用した体験的な記述の生成とその根拠の推定. pp. pp. 3a–7–1–, 3 2023.
[4] 桑折章吾, 加藤優, 高間康史. 検索エンジンを用いた情報検索におけるユーザ行動の分析. 人工知能学会第二種研究会資料, Vol. 2013, No. AM-04, p. 02, 2013.
[5] 柳田雄輝, 加藤誠, 河田友香, 山本岳洋, 大島裕明, 藤田澄男. オンラインショッピングにおける満足度と検索行動の関係分析. 第 15 回データ工学と情報マネジメントに関するフォーラム, Vol. 2023, No. AM-04, p. 10, 2023.
[6] 大重智志, 中村聡史, 田中克己. クエリ推薦に基づくモバイル協調検索支援手法の評価. 情報処理学会研究報告 (CD-ROM), Vol. 2012, No. 23, p. 02, 2012.
[7] 小竹神, 北山大輔. 情報要求の明確化のための穴埋め検索 ui の検討. 第 14 回データ工学と情報マネジメントに関するフォーラム, Vol. 2022, No. DEIM2022 I44-7, p. 08, 2022.
[8] 鈴木永史郎, 杉本徹. 意外性のある検索クエリの推薦方法の提案. 情報処理学会全国大会講演論文集, Vol. 2016, No. vol.78,no.1, p. 02, 2016.
[9] 足立貴行, 内山俊郎, 内山匡, 藤村滋, 望月崇由. 検索クエリを用いた場所の関連語の取得. 人工知能学会研究会資料 知識ベースシステム研究会, Vol. 89, p. 08, 2010.
[10] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸, 堀憲太郎. 関連単語抽出アルゴリズムを用いた web 検索クエリの生成 (web 解析・検索クエリ). 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2008, No. 56, pp. 33–40, 06 2008.
[11] 山本岳洋, 中村聡史, 田中克己. Q&a コンテンツからの観点抽出に基づくウェブ検索支援. 情報処理学会論文誌データベース（TOD）, Vol. 4, No. 2, pp. 74–87, 07 2011.

歌詞と音響特徴の分析に基づく Text2Image モデルを用いた楽曲サムネイル画像の自動生成

佐々木翔一† 西田 紘子†† 澤井 賢一†† 工藤 真生†† 牛尼 剛聡††

† 九州大学大学院 芸術工学府 〒 815–8540 福岡県福岡市南区塩原 4–9–1
†† 九州大学大学院 芸術工学研究院 〒 815–8540 福岡県福岡市南区塩原 4–9–1
E-mail: †sasaki.shoichi@s.kyushu-u.ac.jp, ††{nishida,ken1,m-kudo,ushiama}@design.kyushu-u.ac.jp

あらまし 定額制コンテンツ配信サービスにおいて，サムネイル画像は個々のコンテンツ情報を視覚的に短時間でユーザに伝達することができるため，内容理解支援に貢献している．一方で，音楽配信サービスの分野では，アルバムのように複数楽曲の集合を一つの作品として扱う場合がある．このときアルバムに付与されるサムネイル画像は一つであることが多いため，サムネイル画像はアルバム内の個々の楽曲に特化して内容を表現し，伝達しているとはいいにくい．そこで，本研究では楽曲情報を用いて個々のコンテンツ内容を表現するサムネイル画像を自動生成する手法を提案する．特に，サムネイル画像が持つ複数の機能のうち内容伝達機能に着目し，楽曲を聴取した際の印象を表す画像を生成することを目指す．具体的には，まず，自然言語処理技術を用いた歌詞分析と，音楽キャプショニング技術を用いた音響特徴分析を行い，生成する画像の内容を示すテキスト（プロンプト）を作成する．その後，作成したプロンプトを画像生成モデルに入力しサムネイル画像を獲得する．

キーワード 音楽，画像生成，Text2Image，自然言語処理，LLM，Music Captioning，歌詞分析，音響特徴分析

1 はじめに

近年，映画や音楽などのコンテンツを楽しむ手段として，サブスクリプション形式のコンテンツ配信サービス (以下「サブスク」と呼ぶ) を利用する人が増えている．サブスクでは，ユーザのコンテンツ選択を支援するために，個々のコンテンツに対してサムネイル画像が付与されていることが多い．サムネイル画像は，実際にコンテンツを閲覧したり聴取することと比べて，視覚的に短時間で内容を把握することができる．そのため，多種多様なコンテンツが大量に提示されるサブスクにおいて，ユーザがコンテンツを選択するか否かの重要な決定要因になっていると考えられる．

こうした背景から，各種サブスクではコンテンツの選択確率向上や内容理解支援を目的としたサムネイル画像の最適化が重要視されており，盛んに研究が行われている．例えば，Netflix はユーザの視聴履歴をもとに好みの作品に共通して出現する映像的特徴を算出し，ユーザごとに最もクリック率が高くなると思われるワンシーンをサムネイル画像として提示している [1]．また，AbemaTV はサムネイル画像を検証し，ユーザが興味を持ちやすいと考えられるロゴやタイトルの位置・色・内容等を分析している [2]．

このように，映像サブスクの分野では，サムネイル画像の最適化が進んでいる一方で，音楽サブスクの分野においては，サムネイル画像の最適化が進んでいるとは言い難い．これは，音楽サブスクにおいては楽曲のジャケット画像がサムネイル画像として用いられることが多いからだと考えられる．ジャケット画像は，楽曲作者によって制作された作品の一部であり，楽曲内容を反映しているといえる．しかし，作者の表現の一部という意味合いが強く，ユーザの内容理解支援に特化したものではない．また，音楽分野ではアルバムやプレイリストといった複数楽曲の集合を一つの作品として捉えることも多い．この場合，ジャケット画像 (サムネイル画像) は複数楽曲に対して単一のものが用いられるため，楽曲一つ一つの内容を画像として表しているとはいえない．これらの理由から，音楽サブスク分野では，他のサブスク分野と比較して，内容理解支援や興味度向上のための最適化という点においてはサムネイル画像の最適化が進んでいない．

そこで，音楽と画像のそれぞれから人間が思い浮かべる感情の対応を分析・予測する研究，楽曲情報を用いてサムネイル画像を自動生成する研究などが行われている [3] [4]．しかし，従来の手法は楽曲から得られる特徴量と画像から得られる特徴量を事前に紐付け，そこから音楽と画像の関係性を学習している．そのため，学習過程で用いられたデータと大きく異なる特徴を持つデータが入力されたときに，適切な画像を生成・予測することが難しい．すなわち，従来手法では，生成・予測する画像の多様性が高いとはいえない．

一方，近年，テキスト (プロンプト) を入力とし，その内容を表す画像を生成する Text2Image モデルの研究が活発化している [5] [6] [7] [8]．OpenAI 社が 2021 年に発表した DALL-E では，「アボカドの形をした椅子」といった現実に存在しないオブジェクトの画像を生成することが可能となった [5]．さらに 2022 年に発表された DALL-E2 では，「写実的なスタイルの馬に乗った宇宙飛行士の鉛筆画」など，教師データに含まれていない複雑な内容に対しても高水準な画像を生成できるようになった [6]．Text2Image モデルでは，画像と画像の内容を説明した

テキストの関係を学習している．そのため，単一単語ではない複雑な文章を表す画像も生成することができ，生成画像の多様性という点において従来手法よりも秀でていると考えられる．

以上の背景から，本研究では楽曲が持つ固有の情報とText2Image モデルを用いて音楽サブスクにおけるサムネイル画像を自動生成する手法の開発を目指す．特に，サムネイル画像の機能の中でも「内容理解支援」に注目し，楽曲を聴取した際の印象を反映した画像の生成を目指す．具体的には，まず楽曲の「タイトル」,「歌詞」,「音声」に着目し，これらの中から楽曲の印象を表す画像を生成するために利用できると思われる情報を取得する．そして，取得した情報をもとにプロンプトを作成し，そのプロンプトを Text2Image モデルに入力することで画像を生成する．また，提案手法の有効性を評価するためにアンケート調査による被験者実験を行い，生成した画像が楽曲の印象を反映しているかを検証する．

本論文の貢献は以下の通りである．

- 楽曲歌詞の意味的な重畳性をもとに，楽曲を聴取した際に抱く印象を最も反映していると考えられる「印象的フレーズ」を抽出する手法を開発した．
- 楽曲歌詞と楽曲音声ぞれぞれから，その楽曲の聴取者が抱くと思われる感情を形容詞として予測する手法を開発した．
- 楽曲固有の情報（タイトル，歌詞，音声）に基づき Text2Image モデルを用いて楽曲の印象を表す画像を自動生成するための手法を開発した．
- 楽曲の印象を表す画像を生成した後，被験者実験により提案手法の妥当性を検証した．

本論文の構成は次の通りである．2 章では関連研究について述べる．3 章では提案手法について述べる．4 章では実験と考察について述べる．5 章では本論文のまとめを述べる．

2 関 連 研 究

2.1 音楽サブスクにおけるサムネイル画像の最適化

楽曲情報をもとに，その楽曲に合う適切な画像の生成・予測を目指した研究について述べる．Gaurav らは画像データと音声データから得られる特徴量をニューラルネットワークに入力し，データに付与した印象ラベルの関係を学習することで, 音楽と画像のそれぞれから人間が思い浮かべる感情の対応を分析・予測する手法を提案している [3]．しかし，Gaurav らの手法では，学習したデータから得た知識に基づき予測を行うため，予測を行いたい楽曲の内容に近いサンプルが学習したデータセットに存在しない場合，適切な予測ができない可能性がある．

Qiu らは短時間フーリエ変換（STFT）した楽曲音源を CNN および LSTM を用いたモデルに入力することで，楽曲内容を表現することに適した特徴量を獲得し，その後，特徴量を DCGAN と組み合わせることで，楽曲の雰囲気に合うサムネイル画像を自動生成する手法を提案している [4]．Qiu らの手法では，4 つのクラス（空，水，山，砂漠）について，各クラスのテーマを表現する楽曲および画像からそれぞれ特徴量を抽出し，これらを紐付ける事前学習を行っている．そのため，これら 4 つのテーマのいずれかに関する楽曲の場合，適切な画像を生成することができたとしても，全く未知の楽曲であった場合，適切な画像を生成できない可能性がある．

このように，従来手法では，楽曲の特徴と画像の特徴を紐付け，そこから得られた関係性をもとに画像を生成・予測している．そのため，事前情報として保有していない内容の入力があったときに，適切な画像を生成することが難しい．以上のことから，従来手法は，一つ一つの楽曲に適したサムネイル画像を出力するという，生成画像の多様性が求められるタスクにおいて課題があると考えられる．

2.2 Text2Image とプロンプトエンジニアリング

近年，テキスト (プロンプト) を入力とし，その内容を表すような画像を生成する Text2Image モデルが盛んに研究されている．DALL-E 2 [6]，Imagen [7]，Stable Difusion [8] などでは，入力された文に対して，高水準で内容を表す画像を生成できることが報告されている．また，これらの手法では，Qiu らの手法 [4] のように sky，water といった特定のクラスに属する楽曲と画像とをセットにして学習が行われているのではなく，画像と画像の内容を説明したテキストとの関係を学習している．そのため，sky,water といった単一単語ではない複雑な文章を表す画像も生成することができ，生成画像の多様性という点において従来手法よりも優れていると考えられる．一方で，入力するテキストの記述方法によっては目的の画像とは全く異なるものが出力されるという欠点があり，説明文を Text2Image モデルの理解しやすい形に調整する必要がある．そのため，理想的な画像を生成するための高品質なテキストについて検討するプロンプトエンジニアリングが重要視されており盛んに研究が行われている [9]．

2.3 楽曲における印象的フレーズ

楽曲の歌詞には聴取者に対して特に強い印象を与える「印象的フレーズ」が存在すると考えられる．山西らの研究では，楽曲内の共起語の特異性と繰り返し構造が，印象度に影響を与える可能性を示唆している．特に，楽曲内の繰り返し表現は，印象度に影響を与える割合が大きいことを，被験者実験により示している [10]．

2.4 音楽キャプショニングモデル

近年，楽曲の音声内容を理解し，テキスト情報として出力する LLM について盛んに研究が行われている．こうしたモデルは，音楽キャプショニングモデルとよばれる．具体的には，音楽キャプショニングモデルは，楽曲の音声データを入力として受け取り，その楽曲の音響特徴を説明するテキスト（楽器構成，アーティストの特徴，和声進行の特徴など）をテキスト形式で出力する．たとえば，Spotify によって提案された LLark [11] は，楽曲の音声データをもとに，音楽理解，キャプショニング，推論など幅広い予測を高精度で行えることが示されている．

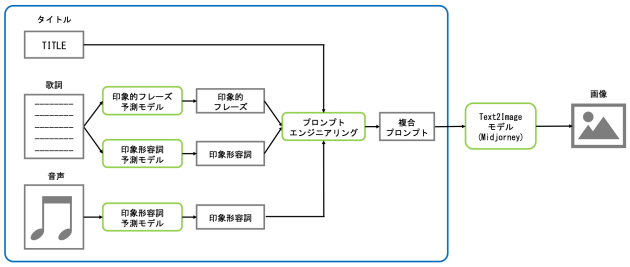


図 1: 提案手法概要

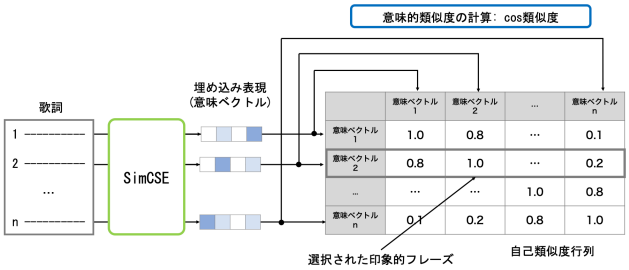


図 2: フレーズ間での自己類似度を用いた印象的フレーズの決定法

3 提案手法

3.1 提案手法の概要

本研究では Text2Image モデルを用いて，楽曲の印象を反映したサムネイル画像を自動生成する手法の開発を目指す．この際，楽曲内容を表現することができる適切なプロンプトを設計することが必要である．そこで，本研究では楽曲データの中から，その楽曲の印象を表す画像の生成に活用可能である情報を取得あるいは予測し，それらを Text2Image モデルに入力することで，サムネイル画像を獲得する手法を開発する．具体的には，楽曲の「タイトル」,「歌詞」,「音声」の 3 つの構成要素に着目し，これらの中から画像生成に利用可能な情報を抽出する．提案手法の概要を図 1 に示す．

3.2 歌詞を用いた印象的フレーズの抽出

山西らは，楽曲におけるフレーズの繰り返しが印象に大きな影響を与えていることを，被験者実験により示した．ここで言う繰り返しとは，同じ文字列の繰り返しという意味である [10]．本研究ではこれに加えて，フレーズの意味的な繰り返し構造についても，印象度に影響を与えているという仮説をたてる．楽曲には作者が一曲全体を通して伝えたいテーマがあり，この作者が表現したいテーマと意味的に最も近いフレーズが印象的フレーズになると考えられる．そして，作者が表現したいテーマは楽曲内のフレーズすべてに対して意味的な影響を与えていると考えられる．そのため，楽曲のテーマと意味的に近い印象的フレーズは，その他のフレーズに対する意味的な類似度が大きいと考えられる．つまり，各フレーズのその他のフレーズに対する意味的な重畳性を確認すれば，印象的フレーズを推定することが可能であると考えられる．

具体的に本研究では，楽曲歌詞の各フレーズを SimCSE を用いて埋め込み表現に変換し，それらの自己類似度を測ることで印象的フレーズを抽出する方法を提案する．

フレーズ間での自己類似度を用いた印象的フレーズの決定法を図 2 に示す．楽曲の歌詞は，時間的な順序関係を有するフレーズの列と考えることができる．まず，対象とする楽曲の 1 番目のフレーズに対して，同じ楽曲に含まれる自身を含めたすべてのフレーズとの意味的な類似度を求める．類似度の具体的な算出方法については後ほど述べる．これをフレーズの時系列順に並べることで，1 番目のフレーズについて自身を含めたすべてのフレーズに対する類似度からなる，ベクトルを獲得することができる (ベクトルの次元数=フレーズ数)．この処理を，2 番目以降のすべてのフレーズについても同様に行い，同様のベクトルを獲得する．その後，獲得したすべてのベクトルを順番に並べることで，フレーズ数×フレーズ数の行列を作成する．この行列を自己類似度行列という．作成した自己類似度行列を用いることにより，任意のフレーズが自身を含めたすべてのフレーズに対して，意味的にどの程度類似しているかを求めることができるようになる．例えば，図 2 の自己類似度行列において，フレーズ 1 の行を確認することで，自身を含めたすべてのフレーズに対する類似度はそれぞれ,「1.0, 0.5, ... ， 0.1」だと知ることができる．実際に意味的な類似度を算出する際には，もとのフレーズはテキストであるため，そのままの状態では類似度を計算できない．そこで，まず各フレーズの文字列をそのフレーズが持つ意味を表現できる形式に変換し，その後類似度を計算する．フレーズの形式変換には，テキストの埋め込み表現を利用する．具体的には，埋め込み表現の獲得には複数の楽曲歌詞により Fine-Tuning を行った SimCSE を用い，類似度の計算には cos 類似度を利用する．

次に，作成した自己類似度行列を用いて印象的フレーズを決定する．我々は，フレーズの意味的な繰り返し構造が印象度に影響を与えており，各フレーズのその他のフレーズに対する意味的な重畳性が大きいほど，印象度が高くなるという仮説をたてた．すなわち，対象楽曲の自己類似度行列の各行 (各フレーズ) を確認し，平均して自身を含めたすべてのフレーズに対する類似度が最も大きい行 (フレーズ) を選択すればよい．選択を行う際には，単純に類似度の平均を求める方法や，一定の閾値を超えた回数をカウントする方法など，複数の手法が考えられる．本研究では，予備実験の結果，最も印象的フレーズの予測精度が高かった，任意の閾値 (類似度 0.9) を超えた回数をカウントし，その回数を重みとして扱った，重み付き平均を用いて印象的フレーズを決定する．

3.3 歌詞と音声を用いた印象形容詞予測モデル

楽曲が聴取者に与える感情を予測できれば，その楽曲の印象をより的確に表現した画像を生成することができると考えられる．そこで，本研究では楽曲の歌詞と音声から，その楽曲を聴取した際に人間が抱く感情を形容詞として予測する手法を提案する．歌詞から形容詞を予測する手法を図 3 に示す．本手法では，まず，楽曲歌詞の各フレーズを，WRIME データセットで Fine-Tuning した BERT に入力する．WRIME データセット

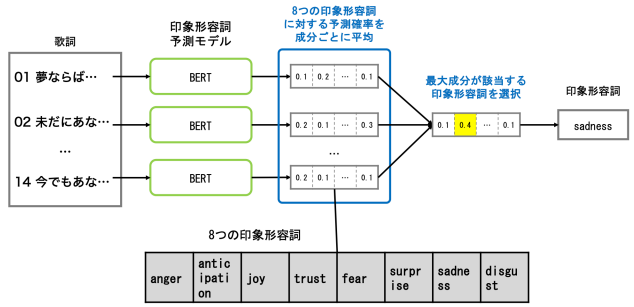


図 3: 歌詞に基づく印象形容詞の予測法

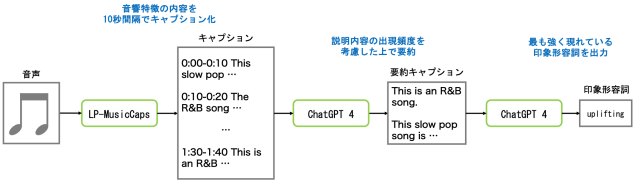


図 4: 音声に基づく印象形容詞の予測法

には，SNS の投稿文とその投稿の印象ラベル（プルチックの基本 8 感情）とのペアが約 2 万件（前処理後）収録されている．BERT は，入力されたフレーズから基本 8 感情のどれが想起されるかを確率分布として出力する．この予測を歌詞内の全フレーズに行った後，要素ごとの平均を取ることで，その楽曲歌詞を代表する印象形容詞を獲得する．

音声から印象形容詞を予測する手法を図 4 に示す．まず，楽曲の音声データを，音楽キャプショニングモデルの一つである LP-MusicCaps [12] に入力する．LP-MusicCaps は，入力された音声データの楽器構成やアーティストの歌い方などを，10 秒間隔でキャプション（説明テキスト）化し出力する．LP-MusicCaps が出力したキャプションには，冗長な表現が多く含まれているため，これを ChatGPT 4 に入力し，情報の出現頻度を考慮した要約を行う．最後に，要約したキャプションを ChatGPT 4 に入力し，要約文から聴取者が最も強く想起すると思われる形容詞を予測する．

3.4 画像生成のためのプロンプト設計

Text2Image モデルを利用して画像を生成する際には，抽出した情報の役割を考慮し，プロンプトを設計することが重要である．本研究では，図 5 に示すようなプロンプトを用いる．印象的フレーズは，画像化した際に具体的な場面を指し示すプロンプトとして用いる．タイトルは，画像化した際に，直接的に画像に反映されず，メタファとして影響を与えるプロンプトとして用いる．ネガティブプロンプトも画像に対する直接的な反映を抑制するためのものである．歌詞と音声から予測した印象形容詞は，画像化した際に画像の雰囲気を形成するプロンプトとして用いる．

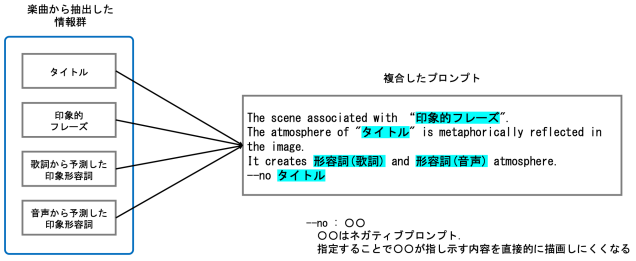


図 5: 画像生成のためのプロンプト設計

表 1: 予測した印象的フレーズと正解データとの一致度

Precision@1	Precision@5	平均順位	標準偏差
0.48	0.84	2.40	1.98

4 実験と考察

4.1 楽曲からの情報抽出

4.1.1 データセットの作成

著者らは，楽曲歌詞に対する聴取者の印象を評価するための基準となるデータセット（以後，楽曲データセットと述べる）を被験者実験により作成しており，本論文においても同じデータセットを使用する [13].

データセットの内容について具体的に説明する．対象楽曲は，歌詞検索サービスの歌ネットにおける歴代人気曲ランキング，上位 25 曲（2022 年 7 月 31 日アクセス時点）とし，クラウドソーシングにより 30 名〜40 名程度の被験者に印象的フレーズを評価させた（楽曲により回答者数が異なる）．実験では，被験者に 1 曲ずつ楽曲の歌詞を提示し，その楽曲において印象的であると感じるフレーズを，1 行単位で任意数選択させた．その後，回答を集計し，被験者により選択された回数により各フレーズに対して印象的度合いのランク付けを行った．

4.1.2 歌詞を用いた印象的フレーズの抽出

楽曲データセット内の楽曲 25 曲に対して，歌詞の意味的な重畳性を用いた印象的フレーズの抽出法を適用した実験を行った．なお，埋め込み表現を獲得するために用いる SimCSE は，J-POP を中心とした約 1 万の楽曲歌詞を用いて，教師なし対象学習を行った．

提案手法を用いて抽出した印象的フレーズを表 2 に示す．表 2 には，例として歌ネットにおける歴代人気曲ランキング上位 10 曲を示す．また，「Lemon ／米津玄師」の印象的フレーズを求める際に用いた自己類似度行列を図 6 に示す．

提案手法を用いて印象的フレーズを予測した際の性能を表 1 に示す．表の各列は実験に用いた 25 曲に対して予測した印象的フレーズの Precision@1，Precision@5，フレーズの正解順位の平均 (Rank Average)，標準偏差である．提案手法は，最も印象的なフレーズを 48%，5 番目以内に印象的なフレーズを 84%の精度で予測している．また，平均して 2.4 番目に印象的なフレーズを予測していると考えられるため，提案手法は重要度の高いフレーズを抽出することができていると考えられる．

表 2: 正解データと提案手法により選択されたフレーズ

楽曲名／アーティスト名	被験者アンケート回答数 1 位のフレーズ	提案手法によって選択されたフレーズ	選択されたフレーズの正解データ内での順位
Lemon ／米津玄師	今でもあなたは私の光	今でもあなたは私の光	1
クリスマスソング／ back number	君が好きだ	君が好きだ	1
キセキ／ GReeeeN	せめて言わせて 「幸せです」と	君に巡り会えた それって奇跡	2
花束／ back number	君と抱き合って手を繋いでキスをして	僕は何回だって何十回だって	2
前前前世（movie ver.）／ RADWIMPS	君の前前前世から僕は 君を探しはじめたよ	君の前前前世から僕は 君を探しはじめたよ	1
Pretender ／ Official 髭男 dism	君の運命のヒトは僕じゃない	グッバイ	2
LOSER ／米津玄師	アイムアルーザー どうせだったら遠吠えだっていいだろう	アイムアルーザー どうせだったら遠吠えだっていいだろう	1
紅蓮華／ LiSA	紅蓮の華よ咲き誇れ！ 運命を照らして	強くなれる理由を知った 僕を連れて進め	4
ハッピーエンド／ back number	あなたを好きなままで消えてゆく	なんてね 嘘だよ 元気でいてね	9
RPG ／ SEKAI NO OWARI	空は青く澄み渡り 海を目指して歩く	空は青く澄み渡り 海を目指して歩く	1

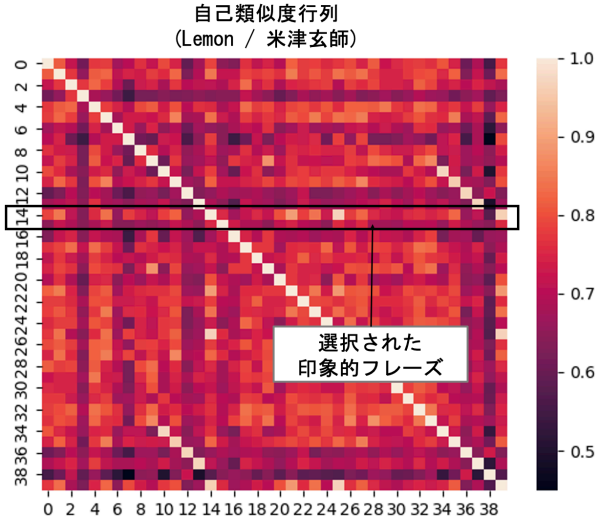


図 6: 自己類似度行列の例 (Lemon ／米津玄師)

4.1.3 歌詞を用いた印象形容詞の予測

提案手法に基づき，印象形容詞予測モデルを構築し学習を行った．その後，楽曲データセット内の 25 曲の歌詞を入力し，印象形容詞の予測を行った．

まず，印象形容詞予測モデルの構築と学習について述べる．ベースとなるモデルには，日本語 Wikipedia 学習済み BERT を用いた．このモデルをもとに，入力として歌詞のフレーズを受け取り，8 つの印象形容詞のいずれかを出力するモデルを構築した．ファインチューニングには WRIME データセットを用いた．WRIME データセットは約 40000 件の SNS の投稿文からなる日本語感情分析用のデータセットである．各投稿文には，8 つの印象形容詞のいずれかがラベルとして付与されている．このデータに対して感情強度の弱いものは除去するといった前処理を行った後，約 2 万件のデータを用いて，8 クラスの分類問題として BERT をファインチューニングした．

次に，楽曲データセット内の 25 曲を対象に印象形容詞の予測を行った．予測の際には歌詞を構成する各フレーズを別々に入力し，各クラスに対する予測確率を平均し，最も確率の高い一つの形容詞を出力するという形式で行った．

各楽曲に対する，予測された印象形容詞を表 3 に示す．実験の結果，歌詞を用いた形容詞予測については，いずれの楽曲においても，特定の形容詞を出力する場合が多かった．具体的には，「悲しみ」，「期待」といった形容詞を出力することが多かった．これは，WRIME データセットの各クラスのデータ数に偏りが存在したことと，基本 8 感情では楽曲の詳細なニュアンスを捉えきれなかったことに起因すると考えられる．

表 3: 予測された印象形容詞

楽曲名／アーティスト名	形容詞
Lemon ／米津玄師	sadness（悲しみ）
クリスマスソング／ back number	anticipation（期待）
キセキ／ GReeeeN	joy（喜び）
花束／ back number	anticipation（期待）
前前前世／ RADWIMPS	anticipation（期待）
Pretender ／ Official 髭男 dism	sadness
LOSER ／米津玄師	anticipation（期待）
紅蓮華／ LiSA	sadness（悲しみ）
ハッピーエンド／ back number	sadness（悲しみ）
RPG ／ SEKAI NO OWARI	anticipation（期待）

4.1.4 音声を用いた印象形容詞の予測

提案手法に基づき，楽曲の音声特徴のキャプション化と楽曲の印象を表す形容詞の抽出を行った．

まず，楽曲の音声特徴のキャプション化について実験を行った．音声特徴のキャプション化には Tianyu Gao らによって提案された LP-MusicCaps を利用した．LP-MusicCaps に楽曲データセット内の 25 曲の音声データを入力し，各楽曲の音声を説明したキャプションを獲得した．獲得したキャプションの例を表 4 に示す．次に，各楽曲について，出力されたすべてのキャプションを ChatGPT 4 に入力し，説明内容の出現頻度を考慮した上で，冗長な表現を削除するといった整形を行った．最後に，獲得したキャプションを用いて，楽曲の印象を表す形容詞の抽出を行った．ここでは，キャプションを要約することによる効果を検証するために，全キャプションをそのまま与えて予測した形容詞と要約したキャプションを与えた形容詞とを比較検討した．形容詞の抽出には，ChatGPT 4 を用いた．このとき，ChatGPT 4 には「楽曲音声を説明したテキストの中で最も強く現れている感情を形容詞として出力せよ」という指示を与えた．出力として獲得した形容詞の例を表 5 に示す．

実験の結果，まず，楽曲音声のキャプション化については比較的高精度で行えていると考えられる．たとえば，Lemon からは「男性の主旋律」，「バックでピアノが和音を奏でる」，「エモーショナルでハートフルなサウンド」といったキャプションが抽出できており，これらは同楽曲の音声の特徴をよく表していると思われる．また，クリスマスソングからは「スネアロールと鈴のきらめき」，「クラシック音楽の演奏」，「登場人物が恋に

表 4: LP-MusicCaps による J-POP 楽曲のキャプショニング結果（キャプションから一部抜粋，日本語に自動翻訳）

楽曲名／アーティスト名	出力されたキャプション
Lemon ／米津玄師	この低音質録音はポップ・ソングのライブ・パフォーマンスで，メロウなピアノのコードに乗せてフラットな男性ボーカルが歌っている．情熱的でエモーショナル，ハートフルなサウンドだ．
クリスマスソング／ back number	低品質の録音で，ベルのインパクトが大きいのが特徴．録音はノイズが多く，モノラル．クラシック音楽の生演奏．クラリネットが主旋律を吹いている．アコースティックギター，エレキギター，ベースがバッキングを演奏している．アコースティック・ドラムが大きな音でリズムを刻んでいる．活気に満ちた雰囲気だ．
キセキ／ GReeeeN	男性シンガーが歌うエモーショナルなメロディー．ピアノ伴奏，力強いベースライン，安定したドラムのリズム，ギター伴奏，男性ボーカルによるミディアムテンポの曲で，感情的で情熱的な曲．音質は悪い．

落ちるシーンに使えそう」といったキャプションが抽出できており，これらも同楽曲の音声の特徴をよく表していると思われる．以上のことから，LP-MusicCaps を用いた楽曲音声のキャプション化については比較的高精度で行えたと考えられる．

次に，楽曲の印象形容詞の抽出については，こちらも比較的高精度で行えていると考えられる．たとえば，Lemon からは「uplisting（元気づける）」という形容詞が抽出された．同楽曲は「別れによる悲しみの中でも希望を見つける」というテーマを反映するために「ややスローテンポでストリングスやピアノをメインとした温かみが感じられる楽器構成と楽曲構成」という音響特徴になっていると思われるが，抽出された形容詞はこの内容をよく表していると考えられる．クリスマスソングからは「nostalgic（ノスタルジックな）」という形容詞が抽出された．同楽曲は「大切な人をずっと思い続けているが，その願いが実らない」というテーマを反映するために「スローテンポで鈴の音やストリングスを用いた切なさと郷愁を感じさせる楽器構成と楽曲構成」という音響特徴になっていると思われるが，抽出された形容詞はこの内容をよく表していると考えられる．以上のことから，楽曲音声からの印象形容詞の抽出は，比較的高精度で行えていると考えられる．特に，高精度な予測が行えた要因として，全キャプションを説明の出現頻度を考慮した上で要約したことと，ChatGPT 4 を用いて大量の知識に基づき形容詞を予測したことが効果的だったと考えられる．

表 5: キャプションから予測された印象形容詞

楽曲名／アーティスト名	予測した形容詞
Lemon ／米津玄師	uplifting（元気づける）
クリスマスソング／ back number	nostalgic（ノスタルジックな）
キセキ／ GReeeeN	energetic（エネルギッシュな）
花束／ back number	sentimental（感傷的な）
前前前世／ RADWIMPS	rebellious（反抗的な）
Pretender ／ Official 髭男 dism	emotional（感動的な）
LOSER ／米津玄師	youthful（若々しい）
紅蓮華／ LiSA	upbeat（楽天的な）
ハッピーエンド／ back number	emotional（感動的な）
RPG ／ SEKAI NO OWARI	optimistic（楽観的な）

4.2 画像生成と評価実験

楽曲データセット内の 25 楽曲を対象に，提案手法に基づきプロンプトを作成し，Text2Image モデルの一つである Midjourney を用いて画像を生成した．この際，提案手法で用いるプロンプトの構成要素（印象的フレーズ，タイトル，歌詞に基づく印象形容詞，音声に基づく印象形容詞）それぞれの重要性を検証する必要があった．そこで，プロンプトを構成する各要素を一つずつ除外した場合の画像も生成し，アブレーションスタディにより各構成要素の重要性を検証した．

具体的には，まず，提案手法に基づき，楽曲データセット内の 25 曲に対して画像を生成した．生成した画像の例を図 7 に示す．

次に，生成した画像の妥当性を評価するために，30 人の被験者による 5 段階の評価実験を行った．被験者には「画像が楽曲を聴取した際の印象を表しているか」という観点で評価させた．評価値の平均を表 6 に示す．なお，P 値は，平均値の差の検定（両側）を行った際の値である．

実験結果より,「印象的フレーズ」,「音声から予測した印象形容詞」については，これらの要素を除外した場合，平均評価値が下がる傾向が見られた．このことから，楽曲の印象を表す画像を生成する際の情報として効果的に機能していると考えられる．次に,「タイトルおよびネガティブプロンプト」,「歌詞から予測した印象形容詞」については，これらの要素を除外することにより，平均評価値が上がる傾向が見られた．このことから，これらの要素は楽曲の印象を表す画像を生成する際の情報として効果的でなかったと考えられる．この原因として，プロンプトの設計や予測した形容詞が適切でなかったことが示唆されたため，楽曲による評価値のばらつきについて，今後詳細に検証する必要がある．

5 ま と め

本研究では楽曲が持つ固有の歌詞・音声情報と Text2Image モデルを用いて音楽サブスクにおけるサムネイル画像を自動生成する手法の開発を目指し実験を行った．特に，サムネイル画像の機能の中でも「内容理解支援」を一番の目標に据え，楽曲を聴取した際の印象を反映した画像を生成することを目指した．具体的には，楽曲の歌詞と音声に着目し，印象的フレーズと印

タイトル / アーティスト	複合プロンプト		印象的フレーズ なし		タイトル なし		印象形容詞(歌詞) なし		印象形容詞(音声) なし	
	除外したプロンプト	画像	除外したプロンプト	画像	除外したプロンプト	画像	除外したプロンプト	画像	除外したプロンプト	画像
Lemon / 米津玄師	-		今でもあなたはわたしの光		Lemon		Sadness (悲しみ)		Uplifting (元気づける)	
クリスマスソング / back number	-		君が好きだ		クリスマスソング		anticipation (期待)		Nostalgic (ノスタルジックな)	
キセキ / GReeeeN	-		いつまでも君の横で 笑っていたくて		キセキ		Joy (喜び)		Energetic (エネルギッシュな)	
花束 /back number	-		僕は何回だって何十回だって		花束		anticipation (期待)		Sentimental (感傷的な)	
前前前世 / RADWIMPS	-		前前前世から僕は 君を探しはじめたよ		前前前世		anticipation (期待)		Rebellious (反抗的な)	
Pretender / Official 髭男Dism	-		グッバイ		Pretender		Sadness (悲しみ)		Emotional (感動的な)	
LOSER / 米津玄師	-		アイムアルーザー どうせだったら遠吠えだっていいだろう		LOSER		anticipation (期待)		Youthful (若々しい)	
紅蓮華 / LiSA	-		世界に打ちのめされて負ける意味を知った		紅蓮華		Sadness (悲しみ)		Upbeat (楽天的な)	
ハッピーエンド / back number	-		あなたを好きなままで消えてゆく		ハッピーエンド		Sadness (悲しみ)		Emotional (感動的な)	
RPG / SEKAI NO OWARI	-		空は青く澄み渡り 海を目指して歩く		RPG		anticipation (期待)		Optimistic (楽観的な)	

図 7: 生成した画像例

象形容詞を抽出する手法を開発した．その後，抽出した情報を用いてプロンプトを設計し，Text2Image モデルに入力することで画像を生成した．さらに，提案手法の有効性を評価するためにアンケート調査による被験者実験を行い，生成した画像が楽曲の印象に沿うものとなっているかを検証した．今後は，評価が高かった画像と低かった画像について詳細に考察し，より適切なプロンプトを設計する方法について検討する予定である．また，将来的には,「内容理解」,「クリック率の向上」など目的に合わせた画像を生成することができるように改良する予定である．

文　　献

[1] Artwork personalization at netflix. `https://netflixtechblog.com/artwork-personalization-c589f074ad76`. (Accesed on 2023.10.13).

[2] 海老澤颯（AbemaTV）. Abema におけるクリエイティブ検証と ope 活用. In *CyberAgent Developer Conference*, 2023.

[3] Verma Gaurav, Dhekane Eeshan, Gunesh, and Guha Tanaya. Leaning affective correspondence between music and image. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019*, pp. 3975–3979. IEEE, 2019.

[4] Qiu Yue and Kataoka Hirokatsu. Image generation associated with music data. In *CVPR workshop*, 2018.

[5] Ramesh Aditya, et al. Zero-shot text-to-image generation. In *arXiv:2102.12092*, 2021.

[6] Ramesh Aditya, et al. Hierarchical text-conditional image generation with clip latents. In *arXiv:2204.06125*, 2022.

[7] Saharia Chitwan, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *arXiv:2205.11487*, 2022.

[8] Rombach Robin, et al. High-resolution image synthesis with latent diffusion models. In *arXiv:2112.10752*, 2022.

[9] Kojima Takeshi, et al. Large language models are zero-shot reasoners. In *arXiv:2205.11916*, 2022.

[10] YAMANISHI Ryosuke, et al. Detecting attractive phrases from musical lyrics focusing on typical expressions on lyrics. *International Journal of Affective Engineering*, Vol. Vol.14, No. No.3, pp. 167–174, 2015.

[11] Josh Gardner, Simon Durand, Daniel Stoller, and Rachel Bittner. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*, 2023.

[12] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

[13] 佐々木翔一, 牛尼剛聡. 楽曲歌詞の意味的な重畳性に基づいた印象的フレーズの抽出. 情報処理学会研究報告 (Web), Vol. Vol.2022-IFAT-148, No. No.26, pp. 1–6, 2022.

表 6: 画像の評価値

楽曲名／ アーティスト名	項目	複合プロンプト	フレーズなし	タイトルなし	形容詞なし (歌詞)	形容詞なし (音声)
Lemon ／米津玄師	平均	2.683	2.408	**2.983**	2.667	2.742
	P 値	-	0.248	0.098	0.916	0.716
クリスマスソング／ back number	平均	2.900	2.717	**2.975**	2.892	2.750
	P 値	-	0.362	0.632	0.942	0.194
キセキ／ GReeeeN	平均	2.950	2.583	**3.017**	2.908	2.842
	P 値	-	0.086	0.696	0.738	0.292
花束／ back number	平均	2.508	2.333	**2.883**	2.558	2.525
	P 値	-	0.322	0.006**	0.682	0.886
前前前世／ RADWIMPS	平均	**2.617**	2.300	2.567	2.575	2.450
	P 値	-	0.108	0.800	0.754	0.220
Pretender ／ Official 髭男 dism	平均	2.642	2.558	**2.867**	2.783	2.508
	P 値	-	0.708	0.306	0.274	0.294
LOSER ／米津玄師	平均	2.433	2.542	**2.800**	2.692	2.433
	P 値	-	0.720	0.110	0.130	1.000
紅蓮華／ LiSA	平均	3.033	2.767	**3.350**	3.025	3.008
	P 値	-	0.310	0.090	0.936	0.826
ハッピーエンド／ back number	平均	2.442	2.483	**2.800**	2.675	2.500
	P 値	-	0.784	0.018*	0.056	0.550
RPG ／ SEKAI NO OWARI	平均	2.633	2.417	2.792	**2.850**	2.500
	P 値	-	0.386	0.408	0.114	0.252
トリセツ／西野カナ	平均	**3.083**	2.692	3.000	3.017	2.850
	P 値	-	0.090	0.626	0.688	0.140
奏／スキマスイッチ	平均	**2.775**	2.567	**2.775**	2.758	2.633
	P 値	-	0.328	1.000	0.912	0.236
白日／ King Gnu	平均	2.358	2.417	**2.942**	2.542	2.342
	P 値	-	0.784	0.000**	0.082	0.892
小さな恋のうた／ MONGOL ８００	平均	2.792	2.883	**2.925**	2.767	2.592
	P 値	-	0.632	0.308	0.824	0.038*
366 日／ HY	平均	2.467	2.292	2.467	**2.667**	2.442
	P 値	-	0.366	1.000	0.208	0.836
高嶺の花子さん／ back number	平均	2.642	1.858	2.825	**2.658**	2.467
-	P 値	-	0.000**	0.240	0.892	0.106
猫／ DISH//	平均	2.833	2.517	2.917	**2.867**	2.833
	P 値	-	0.164	0.732	0.810	1.000
I LOVE…／ Official 髭男 dism	平均	2.592	2.667	**3.050**	2.708	2.558
	P 値	-	0.706	0.016*	0.400	0.798
マリーゴールド／あいみょん	平均	**2.883**	2.275	2.983	2.842	2.825
	P 値	-	0.004**	0.590	0.798	0.676
瞬き／ back number	平均	2.608	2.583	2.833	**3.050**	2.775
	P 値	-	0.896	0.276	0.006**	0.182
パプリカ／ Foorin	平均	2.825	2.617	**2.942**	2.708	2.633
	P 値	-	0.344	0.502	0.356	0.140
なんでもないや／ RADWIMPS	平均	2.792	2.500	**2.983**	2.850	2.775
	P 値	-	0.244	0.414	0.726	0.906
香水／瑛人	平均	2.908	2.925	**3.100**	**3.100**	2.833
	P 値	-	0.944	0.254	0.138	0.532
夜に駆ける／ YOASOBI	平均	2.642	2.533	2.683	**2.733**	2.492
	P 値	-	0.610	0.792	0.464	0.192
スパークル／ RADWIMPS	平均	2.725	2.825	**3.167**	3.117	2.733
	P 値	-	0.630	0.014*	0.014*	0.952
全楽曲	平均	2.711	2.530↘	**2.905**↗	2.800↗	2.642↘
	P 値	-	0.000**	0.000**	0.005**	0.002**

$p \leqq 0.05*, p \leqq 0.01**$

チャート画像に対する大規模マルチモーダルモデルの反実仮想推論能力の検証

戸崎　友輔† 　宮森　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054650,miya}@cc.kyoto-su.ac.jp

あらまし　近年，大規模マルチモーダルモデル（LMM）の発展に伴い，自然画像だけでなくチャート画像に対するモデルの読解能力が向上している．しかし，チャート画像に対する質問応答の既存研究では，チャート画像や関連テキストから取得できる表層的な情報のみを用いた回答を多く扱っており，モデルの反実仮想推論能力の検証は十分になされていない．反実仮想推論は，過去の事実や出来事に対する代替案を考慮する能力であり，説明可能 AI の実現において特に重要な能力である．本稿では，チャート画像に対するモデルの反実仮想推論能力を測定するために，反実仮想推論を要する新しいチャート質問応答データセットを提案する．具体的には，チャート画像と，チャート画像に関する質問テキストに反実仮想推論を要するテキストを含むデータセットを構築し，チャートの読解で競争力のある性能を出しているパラメータ数の異なる代表的な LMM を対象に性能を検証する．実験では，いくつかの研究課題を明らかにし，モデルの反実仮想推論能力の向上に対する知見を見出す．本稿により，チャート画像理解の分野における LMM の説明可能性の向上に寄与することが期待される．

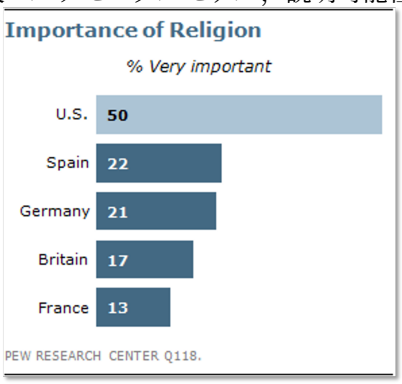
キーワード　反実仮想推論，チャート質問応答，チャート画像，大規模マルチモーダルモデル，説明可能性

1　はじめに

近年，大規模マルチモーダルモデル（Large Multimodal Model；LMM）の発展に伴い，様々な視覚言語データに対する読解能力が著しく向上してきている [1] [2]．自然画像 [3] や 3D 合成画像 [4] を対象とした視覚言語分野の研究が盛んに行われてきたが，昨今，文書画像やチャート画像などの抽象画像を対象とした研究も数多く行われている．チャート質問応答は，チャート画像とそれに対する自然言語の質問文から回答を生成するタスクであり，チャートの読解力を測るタスクの 1 つとして挙げられる [5], [6]．モデルがチャート画像を正しく読解することは，データ分析による意思決定や視覚障害者の支援など，様々な観点からその重要性が認識されている [7], [8], [9] [10]．

しかし，チャート質問応答の既存研究の多くは，チャート画像や関連テキストから取得できる表層的な情報のみを用いた回答生成を扱っている．そのため，チャート質問応答における LMM の反実仮想推論能力の検証が十分になされていないという問題がある．ここで，チャート質問応答での反実仮想の具体例を図 1 に示す．反実仮想なしの質問「最も宗教を重要視している国は？」に対してチャート画像から，最も数値が大きい「アメリカ（U.S.）」が回答となる．一方で，反実仮想ありの質問「もしアメリカが無宗教であるとき，最も宗教を重要視している国は？」に対して，アメリカを除外した次いで数値が高い「スペイン（Spain）」が回答となる．

こういった反実仮想を含む推論は，過去の事実や出来事に対する代替案を考慮する能力であり，計画作成，意思決定，問題解決において重要な役割を果たすことが言及されている [11]．さらに，反実仮想推論の能力は，人間にとって重要な認知機能で



反実仮想なし

Q：最も宗教を重要視している国は？

A：アメリカ（U.S.）

反実仮想あり

Q：もしアメリカが無宗教であるとき，最も宗教を重要視している国は？

A：スペイン（Spain）

図 1: チャート画像に対する反実仮想が含まれない場合の質問応答，反実仮想が含まれる場合の質問応答の例

あり，異なる選択肢や出来事に基づく別の現実や結果を想像することを可能にする．人間と近い思考を持って推論する LMM は，説明可能 AI（Explainable AI；XAI）の実現において極めて重要であることが言える．例えば，LMM がある決定を下し

た理由を説明する際，反実仮想を考慮することで，他の選択肢が採用されていた場合の可能性についても考察することができる．これにより，LMM の決定過程がより透明性を持つとともに，その信頼性も向上する．

そこで本稿では，チャート画像に対するモデルの反実仮想推論能力を測定することを目的とし，そのための問題設定をする．設定した問題に基づき，反実仮想推論を要する新しいチャート質問応答データセットを提案し，以下の 3 つの研究課題を明らかにする．

RQ1 反実仮想の有無により正解率にどの程度の差が生じるか?

RQ2 反実仮想の種類により正解率にどの程度の差が生じるか?

RQ3 チャート画像を用いる VQA での反実仮想推論とテキストを用いる QA での反実仮想推論の違いにより正解率にどの程度の差が生じるか？

本稿では提案手法として，既存のチャート質問応答データセットをもとにして，反実仮想を含むチャート質問応答データセットを新たに構築する．具体的には棒グラフの画像を対象とし，棒の数，カテゴリの数，系列の数についての質問を取り上げる．それらの質問に対して，いくつかの種類に分類可能な反実仮想を付加する．

実験は先述の 3 つの研究課題を明らかにすることを目的として行う．具体的には，チャートの読解において競争力のある性能を示す，パラメータ数の異なる代表的な LMM を対象として，提案したデータセットに対するその性能を検証する．本稿の検証により，チャート画像理解の分野における LMM の課題を見出し，説明可能性の向上に寄与することが期待される．

本稿の主な貢献は以下の通りである．

- チャート画像に対するモデルの反実仮想推論能力を測定するための問題設定をした．
- 問題設定に基づき，反実仮想推論を要する新しいチャート質問応答データセットを提案した．
- 構築したデータセットを用いた実験により，LMM であっても反実仮想を含む推論は困難であり，挑戦的なタスクであることが示唆された．

本稿の構成は以下の通りである．2 節では関連研究を示し，本稿の位置付けを明確にする．3 節ではチャート画像に対する反実仮想推論を要する質問応答の問題設定とデータセットの構築について説明する．4 節では構築したデータセットを用いた実験により，明らかとなった研究課題を考察し，提案手法の有用性を検証する．5 節では本稿の結論とともに今後の課題について述べる．

2 関 連 研 究

2.1 チャート質問応答に関する研究

視覚的質問応答（Visual Question Answering；VQA）[3] が提案されて以降，これまでに様々な視覚言語データに対する質問応答の研究がされてきた．

その中でも近年，チャート画像に対する質問応答（チャート質問応答）が盛んに研究されており，様々なベンチマークデータセットが提案されている [12].

Masry らのチャート質問応答のためのデータセットを提案した研究が代表例として挙げられる [13]．この研究によって提案された ChartQA データセットには，棒グラフ，円グラフ，折れ線グラフのチャートを対象に，推論を要する質問が含まれている．チャートの読解における高度な能力を測ることができ，多くの LMM でこのデータセットが活用されている．

また，Kantharaj らはオープンエンドでのチャート質問応答を目指し，OpenCQA データセットを提案している [14]．実験では，一貫性のある説明文を生成することができている一方で，複雑な推論を要する質問に対しては課題があることが示されている．

また，Akhtar らはチャート画像のファクトチェックを目的としたベンチマークデータセットを提案している [15]．人手によりチャート画像に真の説明文と偽の説明文を生成し，構築したデータセットを用いて真偽分類と説明を生成するタスクを設計している．実験では，チャート画像をテキストの表形式に変換することで高い性能を発揮していることが示されている．

これらの先行研究はチャート質問応答を扱っているが，いずれも反実仮想を考慮していない．本稿では反実仮想を考慮した推論能力の検証を重視している点で異なる．

2.2 反実仮想推論に関する研究

反実仮想推論に関する質問応答の研究は，テキストを対象としたものと画像とテキストを合わせた視覚言語を対象としたものの 2 種類に大別できる．

Yu らは，大規模な反実仮想を要するオープンドメイン質問応答が不足していることを示唆し，ウェブ検索で得た内容をもとにした反実仮想推論のタスクを提案している [16]．反実仮想を含む質問での検索は困難なタスクであり，質問の読解と推論においても挑戦的なタスクであることが言及されている．

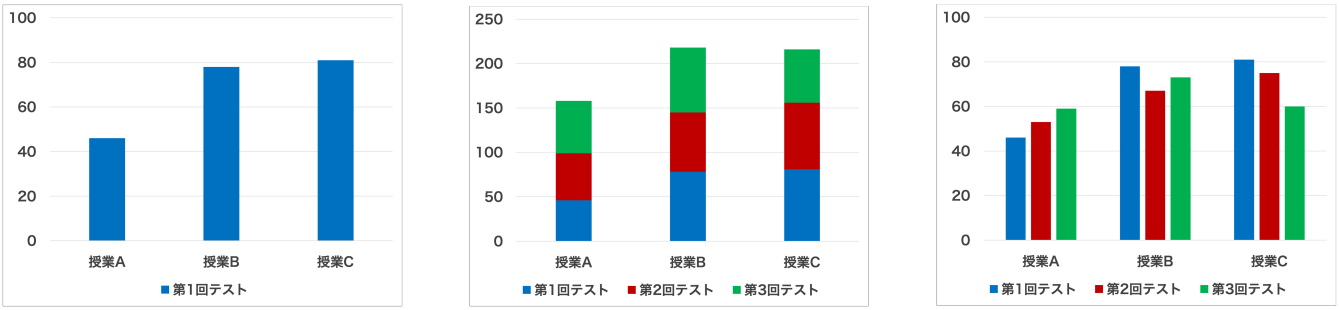
Wu らは，一般的な仮定から逸脱した反実仮想タスクを設計している [17]．実験では，OpenAI の GPT-4 [18] などを対象に検証し，結果として反実仮想タスクでは一般的な家庭の場合と比較し性能が低いことを示している．

Zhang らは，自然画像に対する反実仮想推論を要する質問応答を提案している [19]．画像から直接的に反実仮想推論可能な低レベルのタスクから，間接的に反実仮想推論が必要な高レベルなタスクまで含まれる．実験の結果から，大規模なモデルであっても反実仮想推論を要しない場合と比べて性能が低下することが示されている．

従来研究ではチャート画像に対する反実仮想推論の研究はされておらず，本稿ではチャート画像を対象に反実仮想推論で検証する点で異なる．チャート画像に対する反実仮想推論の能力は，チャート画像内のデータを正確に読み取った上で，過去の事実や出来事に対する代替案を考慮し，説明可能性を向上させるために重要である．

2.3 大規模モデルの推論能力検証に関する研究

大規模言語モデル（Large Language Model；LLM）や LMM



(a) 単一棒グラフ　(b) 積み上げ棒グラフ　(c) グループ化棒グラフ

図 2: 棒グラフの種類の例

の登場により，様々な問題設定での検証がなされている．

Yang らは，GPT-4 に対して幅広い分野での推論能力の質，汎用性，回答形式などについて幅広い観点から評価している [20]．GPT-4 の推論能力の高さを示す一方で，チャート理解などの特定の分野においては，改善の余地があることが示されている．

Guan らは，LMM があたかも本当のことであるかのように嘘をつく問題（幻覚，Hallucination）を測るためのベンチマークデータセットを提案している [21]．この研究では，質問文と事実とは異なる画像を提示し，正しく回答できるかどうかを測定している．データセットには，錯視画像，複数画像，文字画像，チャート画像，表の画像など多くの種類の画像が含まれている．実験では，GPT-4 などの LMM でも多くの幻覚を起こすことがあることが示されている．

Lu らは，GPT-4，Google の Gemini [22] とオープンソースの LMM を対象にテキスト，コード，画像，ビデオにおける下流タスクでの性能を包括的に検証している．[23]．論文中では，選択肢が与えられている場合のテキストにおける反実仮想の能力を検証している．

これらの研究では，LMM が持つ汎用的な推論能力の検証，幻覚の検証をしているが，本稿ではチャート画像を読解した上で，入力画像に影響されずに反実仮想を考慮する能力の検証をする点で異なる．

3 提 案 手 法

3.1 問 題 設 定

3.1.1 入力と出力の形式

チャート画像 C と，反実仮想推論を要するテキストを含む質問文 Q を LMM への入力とし，チャート画像 C と質問文 Q に対する回答 A を出力する問題を扱う．

$$\mathrm{LMM}(C, Q) \rightarrow A$$

3.1.2 対象とするチャートの種類

チャートの種類には，棒グラフ，円グラフ，折れ線グラフ，散布図などがあり，可視化の目的に応じてそれぞれのチャートが用いられる．棒グラフはさらに，単一棒グラフ（single bar chart），積み上げ棒グラフ（stacked bar chart），グループ化棒グラフ（grouped bar chart）に細分化される．それぞれのチャートの例を図 2 に示す．単一棒グラフは，系列数が 1 つの場合に用いられ，系列数が複数の場合には積み上げ棒グラフやグループ化棒グラフが用いられる．例に示している積み上げ棒グラフとグループ化棒グラフは，カテゴリ数が 3 つ，系列数が 3 つの場合であるが棒の数はそれぞれ 3 本と 9 本であり異なる．このように，データ構造が同じでも棒グラフの種類によって読み取り方が異なる特徴があり，本稿ではこの特徴を活用してそれぞれの棒グラフの種類ごとに検証する．

3.1.3 質問と回答

a） 反実仮想を含まない質問

主に棒の本数，カテゴリ（グループ）の数，系列の数についての質問を取り上げる．具体的な質問は以下の通りである．

- 棒の本数：「棒は何本あるか？」
- カテゴリの数：「カテゴリはいくつあるか？」
- 系列の数：「系列はいくつあるか？」

b） 反実仮想を含む質問

反実仮想を含む質問は，反実仮想を含まない質問に「もし〜ならば」という条件を付加した形式をとる．反実仮想を含むことで，見た目上の棒の本数やカテゴリの数，系列の数が変化することが想定できる．その具体例を図 3 に示す．図中央の棒グラフは反実仮想を考慮する前のものであり，図左上は棒の本数が増加した場合の外見，図左下は棒の本数が減少した場合の外見となる．また，図右上はカテゴリの数が増加した場合の外見，図右下はカテゴリの数が減少した場合の外見となる．

このように，LMM は与えられたチャート画像に反して，見た目の変化を考慮して正しい回答を出力することが求められる．

具体的な反実仮想を含む質問の例は表 1 に示す．単一棒グラフに関しては，系列に関する質問と系列の変化に関する反実仮想は数が変化しない，あるいは 0 個となる場合が考えられるため本稿では除外している．表中の「直接的/間接的」で反実仮想の種類を分類している．変化の対象と回答の対象が同じである場合は「直接的」，異なる場合は「間接的」としており，間接的である方が難易度が高い反実仮想推論であると言える．

3.2 データセット

3.2.1 チャート質問応答データセット

Masry らによって提案された ChartQA [13] をベースとし，反実仮想推論を要する新しいチャート質問応答データセットを

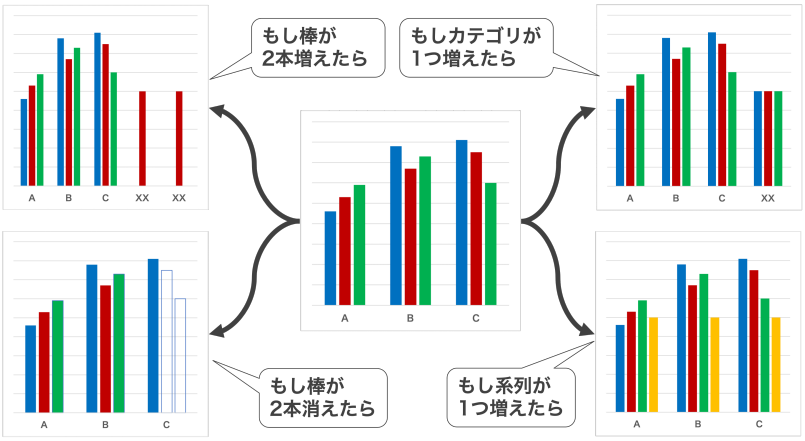


図 3: 反実仮想を含む質問によるチャートの変化の例

表 1: 反実仮想を含む質問の例

反実仮想を含む質問								棒グラフの種類		
	変化対象		変化の仕方		回答対象 *		直接的/間接的	単一	積み上げ	グループ化
もし	棒の数	が	X 本増える	と	棒の数	はすべてで何本あるか？	直接的	✓	✓	✓
もし	棒の数	が	X 本減る	と	棒の数	はすべてで何本あるか？	直接的	✓	✓	✓
もし	棒の数	が	X 倍になる	と	棒の数	はすべてで何本あるか？	直接的	✓	✓	✓
もし	カテゴリの数	が	X 個増える	と	棒の数	はすべてで何本あるか？	間接的	✓	✓	✓
もし	カテゴリの数	が	X 個減る	と	棒の数	はすべてで何本あるか？	間接的	✓	✓	✓
もし	カテゴリの数	が	X 倍になる	と	棒の数	はすべてで何本あるか？	間接的	✓	✓	✓
もし	系列の数	が	X 個増える	と	棒の数	はすべてで何本あるか？	間接的		✓	✓
もし	系列の数	が	X 個減る	と	棒の数	はすべてで何本あるか？	間接的		✓	✓
もし	系列の数	が	X 倍になる	と	棒の数	はすべてで何本あるか？	間接的		✓	✓
もし	あるカテゴリ名	を	除く	と	棒の数	はすべてで何本あるか？	間接的	✓	✓	✓
もし	ある系列名	を	除く	と	棒の数	はすべてで何本あるか？	間接的		✓	✓

* 代表例として棒の本数を回答対象としているが，カテゴリ数，系列数を回答対象とする質問も同様に考えられる．

構築する．ChartQA は，棒グラフ，円グラフ，折れ線グラフのチャート画像とアノテーション，質問と回答のデータを含むデータセットである．

データセットに含まれるアノテーションから，棒の本数，カテゴリの数，系列の数についての反実仮想を含まない質問と，反実仮想を含む質問をそれぞれ生成する．

変化させる数は，1 から 3 の範囲の一様分布に従う乱数を生成して用いる．また，質問は英語で生成する．

3.2.2 データセットの統計

利用した棒グラフの画像の件数と，それに対する質問の件数を表 2 に示す．実験では，生成したすべての質問を対象に検証する．

4 実　　験

4.1 研 究 課 題

実験は以下の研究課題を明らかにすることを目的として行う．

RQ1 反実仮想の有無により正解率にどの程度の差が生じるか？

RQ2 反実仮想の種類により正解率にどの程度の差が生じるか？

RQ3 チャート画像を用いる VQA での反実仮想とテキストを用いる QA での反実仮想の違いにより正解率にどの程度の差が生じるか？

4.2 実 験 設 定

4.2.1 モ デ ル

本稿では，以下の 6 つのモデルを対象に評価する．

a） MatCha [24]

チャート画像の理解に特化した比較的小規模なマルチモーダルモデルである．レイアウト理解と数値推論の能力が重要な要素であると言及し，これらの能力を向上させるための事前学習をしている．

b） UniChart [25]

チャート画像の理解に特化した比較的小規模なマルチモーダルモデルである．チャート画像から棒や線などの視覚的要素を抽出する低レベルのタスクの学習と，チャートの読解や推論が必要となる高レベルのタスクの学習をしている．

c） LLaVA-1.5 [26]

ビジュアルインストラクションチューニングによって提案された LLaVA [27] の改善版であり，MLP 結合層とデータセットの拡張によって性能が向上している．大規模言語モデルである

表 2: データセットの統計

	垂直棒グラフ			水平棒グラフ		
件数	単一棒グラフ	積み上げ棒グラフ	グループ化棒グラフ	単一棒グラフ	積み上げ棒グラフ	グループ化棒グラフ
画像	544	89	118	335	87	17
反実仮想なし質問	1088	267	354	670	261	51
反実仮想あり質問	7616	2937	3894	4690	2871	561

Llama2 [28] や Vicuna [29] をベースとする生成モデルとし，視覚エンコーダである CLIP [30] を組み合わせて構築した LMM である．様々なベンチマークにおいて GPT-4 に匹敵する性能を達成している．

d） ChartLlama [31]

LLaVA-1.5 をベースとして，チャート画像理解タスクでインストラクションチューニングされたモデルである．GPT-4 を用いて，多様なインストラクションチューニングデータセットを作成し，利用している．従来のチャート理解のベンチマークの多くで SOTA を達成している．

e） Gemini Pro [22]

Gemini には，Gemini Nano，Gemini Pro，Gemini Ultra の 3 つの規模のモデルがある．Gemini Pro は 2 番目に大きな LMM であり，画像理解のベンチマークで SOTA に匹敵する性能を達成している．

f） GPT-4 [18]

様々なベンチマークで SOTA を達成している LMM である．実験では，単一棒グラフ，積み上げ棒グラフ，グループ化棒グラフの画像をそれぞれ 1 つずつランダムで抽出し，定性評価する．

上記のモデルはいずれも，ゼロショットで推論する．また，Gemini Pro の API 呼び出し時に発生した内部エラーは推論結果から除外した．

4.2.2 評 価 指 標

正解率を用いて評価する．

4.3 結果と考察

4.3.1 反実仮想の有無による評価結果（RQ1）

図 4 と表 3 に，棒グラフの棒の数に関する反実仮想の有無による評価結果を示す．反実仮想なしの場合，LMM に分類される LLaVA-1.5，ChartLlama，Gemini Pro は，それぞれのチャートの種類によらず，高い正解率を示している．しかし，比較的小規模なマルチモーダルモデルに分類される MatCha と UniChart では，LMM と比較すると正解率が非常に低い．よって，大規模なモデルであるほど，反実仮想を要する推論において優れた性能を示すことが示唆される．

反実仮想ありの場合，ChartLlama と Gemini Pro が，それぞれのチャートの種類によらず，すべてのモデルの中で高い正解率を示している．一方で，反実仮想なしの場合と比較し，半分以上の正解率の低下が見られる．また，LLaVA-1.5 においては小規模なモデルの MatCha よりも低い正解率を示している．

表中の UniChart の正解率が 0%の項目があるが，ほとんどの出力結果が数値以外の文字列または特殊トークンであったことが原因である．UniChart は，質問の意図をそもそも汲み取ることができていないことが考えられる．

これらの結果から，**RQ1:** 反実仮想の有無により正解率にどの程度の差が生じるか？に対する回答は以下の通りである．反実仮想が含まれる場合は，反実仮想が含まれない場合の正解率と比較し 50%以上の正解率の低下が見られる．LMM であっても，反実仮想推論を要するチャート質問応答では課題の残る挑戦的なタスクであることが示唆される．

4.3.2 反実仮想の種類による評価結果（RQ2）

回答を得るための反実仮想の推論が直接的である場合と間接的である場合での正解率を比較する．結果を図 5 に示す．

Gemini Pro は，いずれの棒グラフの種類においても間接的な反実仮想の場合では，直接的な反実仮想の場合に比べてやや正解率が低下している．一方で，ChartLlama は，直接的な反実仮想の場合と間接的な反実仮想の場合での正解率に大きな差が見られない．ここから，よりチャートの読解に強い ChartLlama では他の LMM に比べ，チャート画像に対する反実仮想推論の能力に長けていることが示唆される．

これらの結果から，**RQ2:** 反実仮想の種類により正解率にどの程度の差が生じるか？に対する回答は以下の通りである．汎用的な LMM である Gemini Pro は，直接的な反実仮想の場合と比較して，間接的な反実仮想の場合でわずかに正解率が低下した．一方で，チャートの読解に強い ChartLlama では，直接的な反実仮想の場合と間接的な反実仮想の場合での正解率に大きな差が見られない．ChartLlama へのチャート画像でのインストラクションチューニングの効果が，反実仮想推論において示されている．

4.3.3 VQA と QA の違いによる評価結果（RQ3）

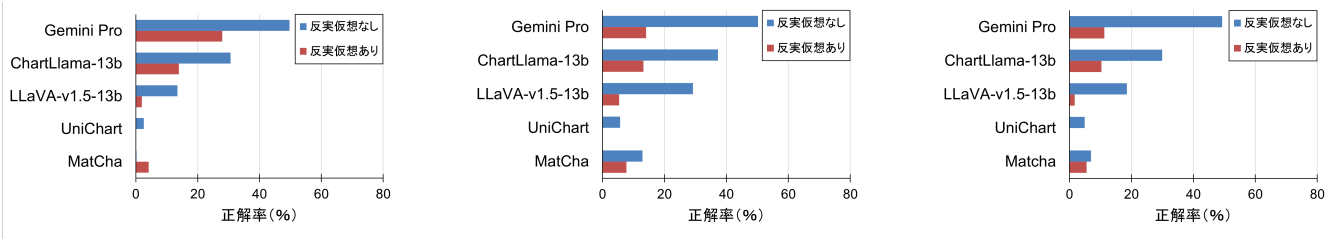
チャート画像と質問を与えて反実仮想を含む推論を要する VQA の場合と，チャート画像の情報をテキストに起こして質問と合わせて反実仮想を含む推論を要する QA の場合での正解率を比較する．検証対象のモデルは，最も高い性能を示した Gemini Pro とした．モデルへ入力したテキストの形式は図 6 の通りで，棒グラフの種類，棒の本数，カテゴリの数，系列の数についての情報を含む．

結果を表 4 に示す．反実仮想を含む推論を要するテキストでの QA は，VQA に比べて非常に高い正解率を示している．明示的にテキストでチャートの情報を与えることで，「チャート画像の読解」における推論の段階が省略されるため，正解率が向上したと考えられる．

また，反実仮想タイプが間接的の場合は，直接的である場合

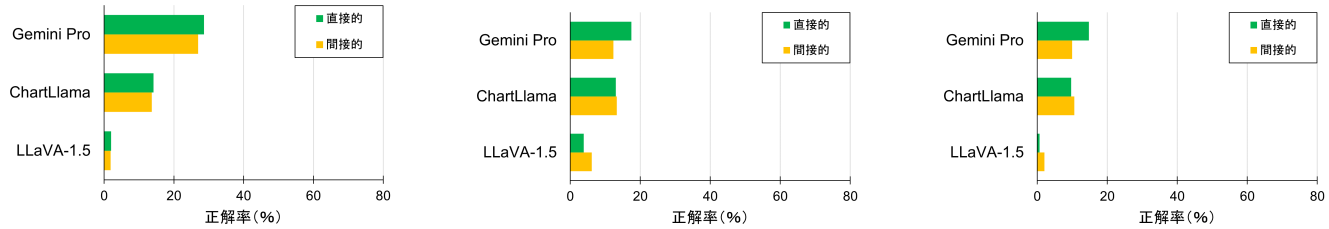
表 3: 反実仮想の有無によるモデルごとの正解率

モデル	単一（single）		積み上げ（stacked）		グループ化（grouped）	
	反実仮想なし	反実仮想あり	反実仮想なし	反実仮想あり	反実仮想なし	反実仮想あり
MatCha	0.19	4.13	12.88	7.68	6.91	5.57
UniChart	2.54	0.00	5.68	0.00	4.94	0.00
LLaVA-1.5-13b	13.46	1.93	29.17	5.35	18.52	1.68
ChartLlama-13b	30.60	13.93	37.31	13.21	29.88	10.35
Gemini Pro	**49.69**	**27.92**	**50.19**	**14.05**	**49.25**	**11.26**



(a) 単一棒グラフでの正解率　(b) 積み上げ棒グラフでの正解率　(c) グループ化棒グラフでの正解率

図 4: 反実仮想の有無によるモデルごとの正解率のグラフ



(a) 単一棒グラフでの正解率　(b) 積み上げ棒グラフでの正解率　(c) グループ化棒グラフでの正解率

図 5: 反実仮想の種類による正解率

と比較して困難な推論を要するが，QA では VQA に比べて高い正解率を示している．人間は画像が与えられた場合の方が直感的に理解しやすいが，LMM は情報をテキストとして与えられる場合の方が推論の性能が向上している．こういった点で，LMM は未だ人間に近い推論能力は持っていないことが示唆される．

これらの結果から，**RQ3:** チャート画像を用いる VQA での反実仮想とテキストを用いる QA での反実仮想の違いにより正解率にどの程度の差が生じるか？に対する回答は以下の通りである．チャートの情報がテキストとして与えられる場合，チャート画像で与えられる場合に比べて 3 倍以上の正解率の向上が見られた．自然言語のみの反実仮想推論を要する質問応答ではある程度は反実仮想の能力を保持していると考えられるが，直感的に画像を理解する人間の思考とは異なることが示唆される．

4.3.4 出力の例

Gemini Pro と GPT-4 による出力の例を示す．ここでは，単一棒グラフ，積み上げ棒グラフ，グループ化棒グラフの 3 種類のチャート画像に対して，棒の数に関する反実仮想を含む質問を与えた際の結果をとその考察を述べる．

> There is a single/stacked/grouped bar chart.
> The chart has X bars, Y categories, and Z series.
> If ... , how many bars are there in total in the chart?

図 6: チャート画像の情報をテキストに起こした入力例

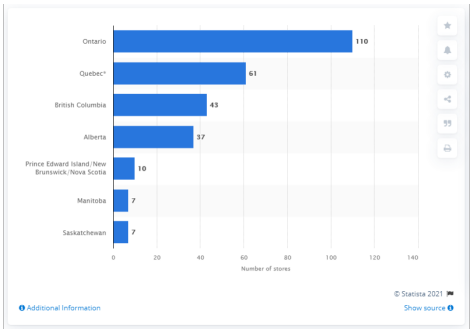
図 7〜9 に，棒グラフ別の出力の例を示す．いずれの棒グラフの種類においても，棒の本数に関する質問では正解している．一方で，反実仮想を含む質問の場合，Gemini Pro では図 7〜9 のすべてで間違った答えを出力した．GPT-4 は図 7，9 では正しい答えを出力したものの，図 8 では間違った答えを出力した．出力は「34」となっており，これは系列の棒の数の合計（8 $bars \times 4$ $series = 32$）に 2 を加えた結果であることがわかる．このように，どの棒の数を数えるべきか明確になっていない場合に間違いを起こす可能性があることが確認できた．

5 まとめ

本稿では，チャート画像に対するモデルの反実仮想推論能力を測定するために，反実仮想推論を要する新しいチャート質問応答データセットを提案した．構築したデータセットを用いて，

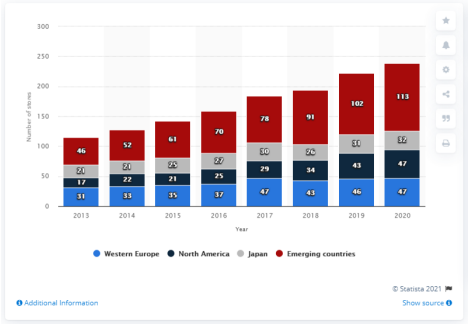
表 4: VQA と QA での正解率

	単一		積み上げ		グループ化	
反実仮想タイプが直接的/間接的	直接的	間接的	直接的	間接的	直接的	間接的
画像＋質問（VQA）	28.63	26.97	17.46	12.35	14.76	9.953
テキスト＋質問（QA）	83.38	86.86	78.25	55.91	76.21	41.48



反実仮想なし：棒は何本あるか？
正解：7
Gemini Pro：7 ✓
GPT-4：7 ✓
反実仮想あり：もし棒の数が 2 本増えると，棒は何本あるか？
正解：9
Gemini Pro：275 ×
GPT-4：9 ✓

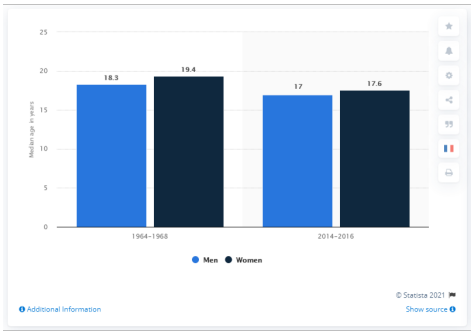
図 7: 単一棒グラフでの出力の例



反実仮想なし：棒は何本あるか？
正解：8
Gemini Pro：8 ✓
GPT-4：8 ✓
反実仮想あり：もし棒の数が 2 本増えると，棒は何本あるか？
正解：10
Gemini Pro：250 ×
GPT-4：34 ×

図 8: 積み上げ棒グラフでの出力の例

チャートの読解で競争力のある性能を出しているパラメータ数の異なる代表的な LMM を対象に性能を検証し，3 つの研究課題を明らかにした．結果は，反実仮想を要しない推論に比べ反実仮想を要する推論では大幅に性能が低下し，直接的な反実仮想の場合と間接的な反実仮想の場合での性能に差が見られることが示された．また，チャートの情報がテキストとして与えられる場合は，チャート画像で与えられる場合に比べて 3 倍以上の正解率の向上が見られた．LMM にとって，チャート画像に対する反実仮想推論は課題が残る挑戦的なタスクであることが示唆された．



反実仮想なし：棒は何本あるか？
正解：4
Gemini Pro：4 ✓
GPT-4：4
反実仮想あり：もし棒の数が 2 本増えると，棒は何本あるか？
正解：6
Gemini Pro：24 ×
GPT-4：6 ✓

図 9: グループ化棒グラフでの出力の例

今後は，複数の種類のチャート画像や反実仮想を含む質問で検証し，LMM の反実仮想推論での性能向上に向けた新しいアプローチを検討する．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *ArXiv*, abs/2306.13549, 2023.

[2] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *ArXiv*, abs/2401.13601, 2024.

[3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31, 2015.

[4] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 1988–1997, 2016.

[5] Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. Automatic chart understanding: A review. *IEEE Access*, 11:76202–76221, 2023.

[6] Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41, 2022.

[7] Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. Exploring chart question answering for blind and low vision users. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[8] Shahira K C, Pulkit Joshi, and Lijiya A. Data extraction and question answering on chart images towards accessibility and data interpretation. *IEEE Open Journal of the Computer Society*, 4:314–325, 2023.

[9] Md. Zubair Ibne Alam, Shehnaz Islam, and Enamul Hoque. Seechart: Enabling accessible visualizations through interactive natural language interface for people with visual impairments. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023.

[10] Benny J. Tang and Angie Boggust. Vistext: A benchmark for semantically rich chart captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[11] Nicole Van Hoeck, Patrick D. Watson, and Aron Keith Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9, 2015.

[12] Chen Chen and Zhicheng Liu. The state of the art in creating visualization corpora for automated chart analysis. *Computer Graphics Forum*, 42, 2023.

[13] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022.

[14] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq R. Joty. Opencqa: Open-ended question answering with charts. *ArXiv*, abs/2210.06628, 2022.

[15] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *ArXiv*, abs/2311.07453, 2023.

[16] W. Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *ArXiv*, abs/2305.14010, 2023.

[17] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *ArXiv*, abs/2307.02477, 2023.

[18] OpenAI. Gpt-4 technical report, 2023.

[19] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Xin Wen, Yongshuo Zong, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4631–4635, 2023.

[20] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *ArXiv*, abs/2309.17421, 2023.

[21] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023.

[22] Gemini Team Google. Gemini: A family of highly capable multimodal models. *Technical report, Google*, 2023.

[23] Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, Kunchang Li, Lijun Li, Limin Wang, Lu Sheng, Meiqiu Chen, Ming Zhang, Qibing Ren, SI-YIN Chen, Tao Gui, Wanli Ouyang, Yali Wang, Yan Teng, Yaru Wang, Yi Wang, Yinan He, Yingchun Wang, Yixu Wang, Yongting Zhang, Yu Qiao, Yujiong Shen, Yurong Mou, Yuxi Chen, Zaibin Zhang, Zhelun Shi, Zhen fei Yin, and Zhipin Wang. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities. *ArXiv*, abs/2401.15071, 2024.

[24] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[25] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq R. Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *ArXiv*, abs/2305.14761, 2023.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[28] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[31] Yucheng Han, China. Xiaoyan Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *ArXiv*, abs/2311.16483, 2023.

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

LLM応用

座長:石川 佳治(名古屋大学)
コメンテータ:上田 真由美(流通科学大学)
2024年2月29日(木) 13:30 〜 15:40 T4-A (オンライン（Zoom Events))

15:15 〜 15:40

[T4-A-5-05]【技術報告】 NTT版 LLM「 tsuzumi」の取組について

西田　京介 (NTT 人間情報研究所)

©一般社団法人日本データベース学会

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 13:30 ~ 15:40 | T4-B オンライン（Zoom Events)

クラウドソーシングと人間・社会理解 (1)

座長:田島 敬史(京都大学)
コメンテータ:清水 敏之(九州大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T4-B-5-01] 男女差を俯瞰するための階層型データ可視化: EMDによる男女間の分布差の導入

*中井 祐希[1]、伊藤 貴之[1] (1. お茶の水女子大学伊藤研究室)

14:00 ~ 14:25
[T4-B-5-02] Dawid-Skene集約における不均一タスク数の影響緩和による人間-AIハイブリッドクラウドソーシングの品質向上

*田村 匠[1]、伊藤 寛祥[2]、小山 聡[3]、森嶋 厚行[2] (1. 筑波大学情報学群、2. 筑波大学図書館情報メディア系、3. 名古屋市立大学データサイエンス学部)

14:25 ~ 14:50
[T4-B-5-03] フィジカルクラウドソーシングにおける分担タスク割当てのためのワーカ依頼順序の最適化

*藤永 弥太郎[1]、伊藤 寛祥[1]、鈴木 伸崇[1]、森嶋 厚行[1] (1. 筑波大学)

14:50 ~ 15:15
[T4-B-5-04] クラウドワーカの作業結果と投票結果から生成される確率分布を利用した結果集約手法

*城所 祥太[1]、鈴木 優[1] (1. 岐阜大学工学部電気電子情報工学科情報コース　鈴木研究室)

15:15 ~ 15:40
[T4-B-5-05] 理想と現実のボディイメージの数理モデル化

*栗生 紗希帆[1]、眞鍋 雅恵[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学)

男女差を俯瞰するための階層型データ可視化

EMD による男女間の分布差の導入

中井　祐希†　　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{g1820524,itot}@is.ocha.ac.jp

あらまし　データの偏りは特定の属性を持つ部分集合に見られることが多い．また，偏りは定量的に判断できる問題ばかりではない．そのため，データ中の特定の属性に起因する偏りを発見するには，属性ごとの数値分布の違いを人間が理解する必要がある．そこで本報告では，多数の人物を対象としたデータから，データの分布の男女差を可視化する手法を提案する．この手法では，データ中の人物群を属性で階層的に分類し，帯グラフを搭載した階層型データ可視化手法を適用する．その際に EMD による男女間の分布差算出により，注目するべき属性の組み合わせを推薦し，可視化画面において男女差が大きい部分を強調表示する．本報告では空調の温感に関する評価値の可視化例を紹介する．
キーワード　可視化，男女差，階層型データ，データバイアス

1　はじめに

近年，公平性の高い社会や組織を構築することは非常に重要な課題となっており，特定の属性に起因する差別や不平等が問題視されている．ここでいう属性とは例えば，性別，地域，人種，世代などである．例えば機械学習において，訓練データに偏りがあることで，学習結果にも偏りが生じることがあり，これが特定の属性に対して利益の不均衡をもたらすことがある．実際に，企業の採用人事に AI を導入したら女性が不利な結果が生じた，クレジットカード発行や住宅ローン貸付が許可される割合に男女差が生じた，といった問題が公表されている．公平性の高い社会や組織を築くためには，多様性を認め，特定の属性に起因する不利益や差別を解消する必要がある．特に性別に起因する利益の不均衡は非常に重要視されており，これを解決するための取り組みとして Gendered Innovation という新しいキーワードが注目されている．

データの偏りはデータ全体に見られるとは限らない．むしろ特定の条件を満たす部分集合に見られることが多い．端的に言えば，データを構成する人物の中で，特定の属性を有してかつ別の要因を満たす人物で構成される部分集合の中に偏りが見られることが多い．具体的な例として「職業が○○である女性は住宅ローン貸付が許可されない割合が高い」といったことが起こり得る．そのため，データの偏りを発見するには，データが有する属性ごとの数値分布の違いを観察することが重要である．

このような偏りの基準には定性的・主観的な問題を含んでおり，必ずしも定量的に判断できる問題ばかりではない．偏りや公平性は文脈に左右されるため，合理的差異なのか，注意すべき偏りであるかを判断するには人間による意思決定が必要である．つまり，データに潜む偏りの発見や理解を機械に完全に任せるのではなく，データの分析と解釈を担当する人間がデータを理解し，データに潜む偏りを発見し，その対策について意思決定を下すことが必要である．可視化技術はデータの理解を促進する技術として活発に研究されており，データの偏りの発見の目的にも有用であると考えられる．特に，データ中の特定の部分に含まれる偏りを発見するには，データ全体の俯瞰と，その中の特定の部分集合の発見を両立できる可視化手法が有用である．この観点から我々は，以下の要件を満たす可視化手法の開発を目標とした．

要件 1　データの全体像を俯瞰できる．

要件 2　データ中の特定の属性に起因する部分集合に生じている偏りを発見しやすい．

要件 3　データ中の特定の属性に起因する部分集合に含まれる人物群の分布を表現できる．

要件 4　データ中に潜む偏りを発見するために必要な反復的操作の回数をできる限り少なくすることができる．

本報告では，多数の人物を対象としたデータから，データの分布の男女差を可視化する一手法を提案する．人物をデータ要素とし，各々の人物が各種の属性値を有するデータを仮定する．データを構成する人物群を属性で階層的に分割し，「平安京ビュー」[1] という階層型データ可視化手法を用いて可視化する．ここで，階層を表現する長方形領域を 2 列の帯グラフで塗りつぶし，その帯グラフによって所定の属性値に関する男性および女性の数値分布を表現することで，特定の属性に起因する数値分布の違いを可視化する．

さらに，我々自身の先行発表 [2] からの進展として，男性と女性の数値分布の分布間距離の算出に Earth Mover's Distance (EMD) [3], [4] を導入する．そして，指定値よりも EMD の値が高ければ男女の帯グラフをカラースケールで表示し，低ければグレースケールで表示することで，数値分布の男女差が大きな部分を強調表示する．さらに，可視化システムを操作する際

に選択するに値する属性の組み合わせをユーザに推薦する．まず，データ中の人物群が有する属性の全ての組み合わせに対して人物群を階層的に分類して木構造を生成する．生成した木構造の全ての末端の葉ノード群に相当する人物群の男性と女性の数値分布の分布間距離を EMD を用いて算出する．算出した EMD の合計値が高いものから順に木構造を生成する際に用いた属性の組み合わせをユーザに提示する．

本手法では要件 1 を満たすために，階層型データ全体の俯瞰を目的とした「平安京ビュー」手法を採用する．その上で本手法では，EMD の値が大きい部分，すなわち男女差が大きい部分集合のみを表示するのではなく，男女差が大きい部分を強調表示した上で，男女差が小さい部分集合もあわせて表示することで，階層化されたデータ全体の俯瞰を実現する．ここで，データを構成する人物群が有する属性で階層的にデータを分類し，男女差が大きい部分をカラースケール，男女差が小さい部分をグレースケールで表示することで，要件 2 を満たす．また，末端の階層に含まれる人物群の数値分布を帯グラフで表現することで，要件 3 を満たす．さらに，選択するに値する属性の組み合わせをユーザに推薦することで，要件 4 を満たす．

本報告では，空調の温感に関する評価値の男女差を可視化した事例を報告し，本手法の有効性を議論する．

2 関 連 研 究

2.1 データのバイアスの可視化

データのバイアスの可視化はここ数年で急速に研究が増えている課題である．機械学習の公平性が大きな社会問題として注目される中で，データのバイアスがその大きな要因であり，機械学習のプロセスに人間が積極的に介在することでその解決を図ることが有用であるとされているためである．データのバイアスの可視化に関する近年の研究の例を以下にあげる．

Cabrera らによる FairVis [5] は，人種や性別などのセンシティブな属性を複合的にグループ化し，グループ間で発生する交差バイアスに注目することで，差別や不平等な学習結果を防ぐ可視化解析システムである．Wang らによる DiscriLens [6] は，差別発見モジュールと可視化モジュールを組み合わせることで，機械学習におけるバイアスを調査し、より包括的な分析をサポートするための対話型可視化ツールである．Pastor らによる DivExplorer [7] は，モデルが異常な動作をするデータセット中のサブグループ (発散サブグループ) を特定し，サブグループとデータセット全体のエラーメトリックの差を表形式で可視化する．さらに，個々の属性がサブグループの発散に与えた影響や，サブグループ内の個人が他のサブグループに出現する程度を可視化する．栃木ら [8] は，機械学習における映画推薦システムデータにおいて鑑賞履歴と推薦結果の差異を表示することで，推薦システムにおける機械学習のバイアスを可視化している．Munechika らによる VISUAL AUDITOR [9] は，モデルパフォーマンスが低いサブグループを SliceFinder アルゴリズムを用いて特定し，サブグループのサイズや定義，どの程度パフォーマンスが低いかなどの重要な情報も付随して可視化する．

これらのシステムでの可視化表現は単純な棒グラフ・折れ線グラフ・散布図を中心とした複数のビューで構成されており，ユーザは反復的な操作を繰り返すことでデータのバイアスを探す必要がある．一方で，情報可視化システムの多くは Schneiderman [10] が提唱した「Overview first, zoom and filter, then details on demand」という操作シナリオに則って設計されており，Overview(俯瞰) のための最初の一画面がその後の適切な操作を導く重要な要素であるとされている．この考えに則り，データを俯瞰する最初の一画面でデータ中の特定の部分に潜む偏りを発見させることに重点を置いた，という点で本研究は上述の従来研究と大きく異なる．

2.2 男女差の可視化

ダイバーシティ問題の改善や解決において，コミュニティや組織の中の多様性を分析し，理解する必要がある．特に性別に関する不平等は非常に重要視されており，性別に起因する格差の可視化に関する近年の研究の例を以下にあげる．Tovanich ら [11] は，VIS(IEEE Visualization) コミュニティ内の男女差とその傾向や，共著関係構築における性別的観点での変化を詳細に分析し，ダイバーシティへの取り組みとしてコミュニティ内のジェンダーに関する議論を提示した．Li ら [12] は研究者の師弟関係を木構造で表して可視化する手法を提案した．木構造に含まれる研究者の性別を枝の色で，研究分野の色を葉の色で表し，男性研究者と女性研究者の偏りを木構造の傾きで表現した．様々な年代の分野の違う研究者の師弟関係を可視化し，木構造の形や傾きがどのように変化したかを分析した．

2.3 階層型データ可視化手法

2.3.1 木構造表示型の可視化手法

階層型データ可視化手法の中で最も一般的なのが，木構造を表示する手法であり，Cone Tree [13]，Hyperbolic Tree [14]，Fractal Views [15] などが提案されている．階層型データ中の枝ノードと葉ノードの親子関係に重点を置いた可視化手法であるため，階層型データの末端の葉ノードに含まれる人物群の持つ属性と数値分布を同時に可視化することを目的とする本研究に適した手法ではない．

2.3.2 空間充填型 (Space-filling) の可視化手法

長方形やアイコンなどの視覚要素で画面を埋めるタイプの可視化手法を「空間充填型 (Space-filling)」と呼ぶ．階層型データ中の葉ノードと枝ノードの親子関係よりも，階層型データ全体に分布する葉ノード群を全て一画面に表現することに焦点を当てた可視化手法である．Treemaps [16], [17] は画面空間を再帰分割し，入れ子状の帯グラフで階層型データを表示する．図 1 に Treemaps による可視化例を示す．Treemaps は下位階層を表現する長方形領域のアスペクト比が必ずしも小さくなるとは限らず，下位階層領域の形状に自由度がない．

Voronoi Treemap [18] は長方形に限らず三角形や円など任意の形状の領域で画面を分割し，Treemaps を実現する可視化手法である．本研究では階層構造の末端に含まれる人物群の分布

を帯グラフで表現することから，長方形領域で画面を埋める必要がある．

図 1　Treemaps [16], [17] による階層型データの可視化例.

伊藤らによる「平安京ビュー」[1] は，階層型データの葉ノードを長方形のアイコンで，枝ノードを長方形の枠で表す．階層構造を 2 次元の長方形群の入れ子構造で表現し，これらをできるだけ小さい画面空間に配置することで，階層型データ全体を一画面に表示する．図 2 に「平安京ビュー」による可視化例を示す.「平安京ビュー」は Treemaps と比べて，アルゴリズムが生成する長方形領域の形状の点で優位であるという実験結果 [1] がある．

「平安京ビュー」は空間充填型可視化手法であるため，階層型データ全体を一画面で表示することができる．また，階層構造を長方形の入れ子構造で表現し，階層を表現する長方形領域の形状の自由度が高いため，末端の階層に含まれる人物群の分布を帯グラフで表示することができる．これらの特徴は他の階層型データ手法と比較して 1 節で提示した要件 1 と要件 3 を実現するのにより適した手法であると言える．

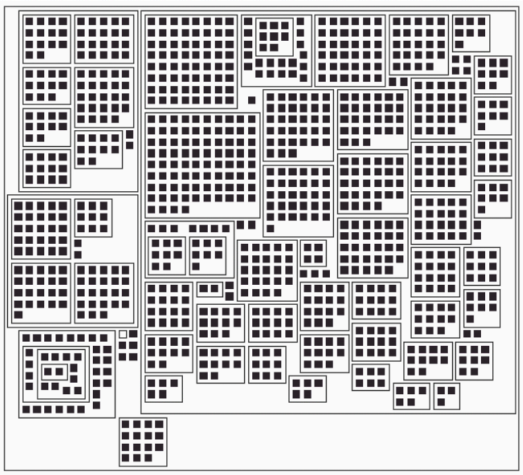


図 2 「平安京ビュー」[1] による階層型データの可視化例.

3　階層型データとしての偏りの可視化

3.1　想定する入力データの概要

本手法では入力データの最小単位が人物であり，各人物が可視化の対象となる値，性別，性別以外の属性値を持つことを想定している．また性別以外の属性値は，実数型変数の属性値か，カテゴリ型変数の属性値である必要がある．よって提案手法では以下のデータ A が与えられることを前提とする．A は人物集合によるデータ全体を表し，a_i は i 番目の人物を表し，n はデータ中の人数を表す.

$$A = \{a_1, a_2, \ldots, a_n\}$$

また，i 番目の人物に相当する a_i は以下の変数を有するものとする. ここで e_i は可視化の対象となる実数値，g_i は i 番目の人物の性別，r_{ij} は j 番目の実数型変数の属性値，c_{ik} は k 番目のカテゴリ型変数の属性値である.

$$a_i = \{e_i, g_i, r_{ij}, \ldots, c_{ik}, \ldots\}$$

3.2　木構造の生成

属性値 r_{ij} または c_{ik} のうちユーザが選択した複数の属性値を用いてデータを構成する人物を階層的に分類し，木構造を構成する．属性値がカテゴリ型変数の属性値の場合は人物群をカテゴリごとに分類する．属性値が実数型変数の属性値の場合は，「(実数値 x_0) 以上（実数値 x_1）以下」,「(実数値 x_1) 以上（実数値 x_2）以下」，…,「(実数値 x_{N-1}) 以上（実数値 x_N）以下」というように区間（ここで N は区間数を表す）を設定し，人物群を区間ごとに分類する．本手法では属性を 1 つずつ選び，その属性にもとづいて人物を分類し... という処理を何度か反復することで木構造を生成する．生成した木構造の特定のノード配下に可視化の対象値である実数値 e_i の偏りが見られるようであれば，その偏りはユーザが選択した属性値に起因する偏りであることが示唆される．現段階の我々の実装では選択する属性の個数を 3 に固定している．

3.3　EMD による男女の分布間距離の算出

Earth Mover's Distance(EMD) [3] [4] は，ある分布をもう一方の分布に移動させるための最小コストとして定義される距離尺度である．本手法では，末端の葉ノード群に相当する人物群のうち，男性と女性の e_i の分布間距離を EMD を用いて算出し，この結果を男女間の分布の非類似度とする．EMD の値が大きいほど男女の分布が異なっており，反対に EMD の値が小さいほど男女の分布が類似していることが示唆される．

3.4　「平安京ビュー」を用いた可視化

生成した木構造を「平安京ビュー」[1] によって可視化する．図 3 にその構造を示す．木構造を「平安京ビュー」で可視化した結果において，1 番外側の領域が 1 個目に選んだ属性，その内側の領域が 2 個目に選んだ属性，さらにその内側が 3 個目に選んだ属性と対応する．

「平安京ビュー」では葉ノードを正方形のアイコンで表現したのに対し，本手法では末端の葉ノード群に相当する人物群が有する e_i の分布を男女別に 2 列の帯グラフで表現する．具体的には，データを構成する人物群が持つ実数値 e_i をいくつかの階級に分割する．そして，木構造の末端の葉ノード群に相当する人物群に対して各階級に該当する人数を集計し，その集計結果を帯グラフで表示する．現段階の我々の実装では階級数を 7 に固定している．

図 4 に帯グラフの例を示す．左側の帯グラフが男性の e_i の数値分布，右側の帯グラフが女性の e_i の数値分布を示す．また，帯グラフの横幅は男性と女性の人数の比率を表す．帯グラフの各領域の色算出には HSI 表色系を採用し，以下の原則に沿って数値分布の各階級の色を算出する．

色相 (H)： EMD の値が指定値より低い場合は男女ともにグレースケールで，EMD の値が指定値より高い場合は男性の色相を青，女性の色相を赤にする．

彩度 (S)： 各階級の中央値が全体の平均値に近いほど低い値を，最大値／最小値に近いほど高い値を割り当てる．

明度 (I)： 各階級の中央値が大きいほど高い値を割り当てる．

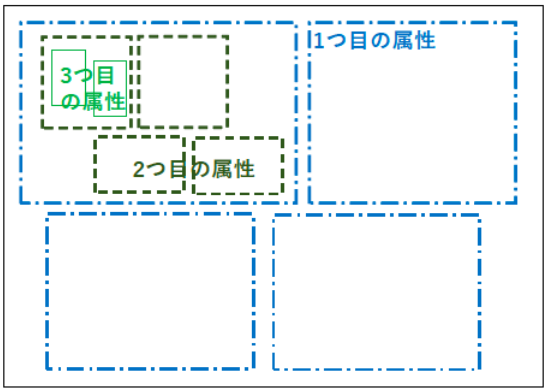


図 3 「平安京ビュー」による人物群の木構造の可視化．

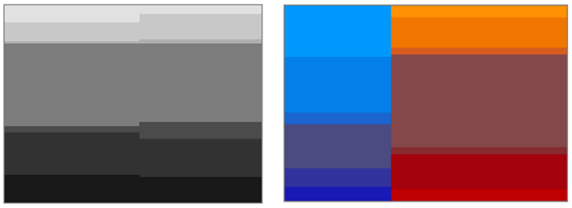


図 4 末端の葉ノード群に相当する人物群が有する e_i の分布を男女別に 2 列の帯グラフで表現．（左）EMD の値が指定値より低い場合．（右）EMD の値が指定値よりも高い場合．

3.5 インタラクション

我々の実装では，階層構造を生成するための複数（現時点での実装では 3 種類）の属性を選択するためのラジオボタンを搭載している．ここで 3 つの属性を選択して「更新」ボタンを押すたびに，階層構造を再生成して表示を更新する．また，帯グラフの上にカーソルをホバリングすることで，対応する 3 つの属性を文字表示する．

4 選択する属性の組み合わせの推薦

3 節で説明した可視化システムをユーザが操作する際に，ユーザが選択する属性の組み合わせは非常に多い．データを構成する人物群の持つ属性の個数が多くなればなるほど，組み合わせの総数は数十，数百にも及ぶ．本研究では，データに潜む偏りを発見する上で，ユーザが反復する操作の回数をできる限り少なくした可視化システムの実現を目指している．そのためには，「複数の属性を選択し，そのたびに可視化結果を観察する」という試行錯誤をユーザが反復しなければならない点を解消する必要がある．そこで，3 節の可視化システムとは別に，選択するに値する属性の組み合わせをユーザに推薦するシステムを開発した．

開発したシステムでは，まず，入力データ（3.1 節と同一のデータ）に含まれる人物群が持つ属性の全ての組み合わせに対して，3.2 節と同様の処理を実行し，それぞれの組み合わせについて木構造を生成する．生成される木構造の全体を T とすると，T は以下のように表される．t_m は m 番目の属性の組み合わせに対して生成される木構造を表し，N は属性の組み合わせの数を表す．

$$T = \{t_1, t_2, \ldots, t_N\}$$

次に，それぞれの木構造 t_m に対して，3.3 節と同様の処理を実行し，全ての末端の葉ノード群に相当する人物群の男性と女性の e_i の分布間距離を EMD を用いて算出する．l 番目の末端の葉ノード群に対して算出された EMD の値を d_l とする．ここで，算出した EMD の合計値を v_m とすると，v_m は以下のように表される．L は末端の葉ノード群の個数を表す．

$$v_m = \sum_{l=1}^{L} d_l$$

以上の結果，木構造 t_m に対して EMD の合計値である v_m が求められる．木構造に対する EMD の合計値が高いということは，その木構造を生成する際に用いた属性の組み合わせは男女間の偏りを生じさせている可能性が非常に高い．そのため，EMD の合計値である v_m が高いものから順に木構造 t_m を生成する際に用いた属性の組み合わせをユーザに提示することで，ユーザが選択する属性の組み合わせを推薦する．

以上の処理には一定の計算時間を要するが，単一のデータセットに対して 1 回だけ実行すればいいので，可視化システムの対話操作のユーザビリティに影響を及ぼすものではない．

5　空調温感データでの適用事例

本報告では空調の温感に関するオープンデータ [19] を適用した事例を示す．このデータから著者らは 32,373 人を対象として以下の属性値を抽出した．

TS:　温感に対する 7 段階の評価値．0 がちょうどよい，正値が暑い，負値が寒い，という回答を表す．

Sex:　生物学的な意味での性別．なお「男性」「女性」以外で回答した人物は本事例では対象外とした．

Age:　年齢．

Cloth:　服装の厚さに関する実数値．大きいほど厚い．

Metab:　代謝量に関する実数値．

Season:　春・夏・秋・冬の 4 種類のカテゴリ値．

Building:　オフィス・教室・住居・高齢者施設・その他の 5 種類のカテゴリ値．

Strategy:　エアコン・換気・混合の 3 種類のカテゴリ値．

以上をまとめると，TS が e_i に，Sex が g_i に，Age, Cloth, Metab が r_{ij} に，Season, Building, Strategy が c_{ij} に対応する．実数型変数の属性値である Age, Cloth, Metab は 4 段階に区切った．

5.1　推薦された属性の組み合わせ

6 個の属性から 3 個の属性を選ぶ組み合わせは全部で 20 通りある．推薦された属性の組み合わせの上位 10 個は表 1 のようになった．

表 1　推薦された属性の組み合わせの結果．

順位	属性の組み合わせ
1	Age, Cloth, Building
2	Cloth, Metab, Building
3	Age, Metab, Building
4	Metab, Season, Building
5	Cloth, Season, Building
6	Age, Season, Building
7	Age, Metab, Season
8	Age, Cloth, Season
9	Age, Cloth, Metab
10	Cloth, Metab, Season

5.2　可視化例 (1)

推薦順位が最も高い Age，Cloth，Building の組み合わせを用いて可視化する．図 5 は Age，Cloth，Building の順に属性値を参照して人物を分類した可視化結果である．可視化結果の外側には年齢の 4 つの区間に対応する長方形の枠があり，その内側には服装の厚さの 4 つの区間に対応する長方形の枠があり，さらにその内側には建物の 5 種類のカテゴリ値に対応する長方形領域がある．カラースケールで表示されている帯グラフ，すなわち男女差が大きい部分は可視化結果全体に点在している．このことから可視化結果中に現れている男女差は特定の年齢の属性だけで温感に男女差が生じている可能性が低いと考察できる．

そこで次に，Building，Cloth，Age の順に属性値を参照して人物を分類して可視化する．図 6 はその可視化結果である．可視化画面全体にカラースケールで表示されている帯グラフが点在しているというよりは，特定の部分に集中しており，男女差が大きい部分が局所的に現れている．このことから，建物と服装の厚さが原因で男女差が生じる部分集合が多い可能性が高い．

ここで，左上のオフィスの枠に注目する．まずオフィスの枠が占める面積の大きさから，データ中に建物の属性がオフィスである人物が，オフィス以外の建物と比較して非常に多いことがわかる．次に帯グラフの色に注目すると，ほとんどの帯グラフがグレースケールで表示されているが，A と B の部分のみ帯グラフがカラースケールで表示されている．A は，服装の厚さが 3 区分目（服装の厚さが 2 番目に厚い）で年齢が最も若い人物群の帯グラフである．男性と女性の数値分布を帯グラフで比較すると，男性よりも女性の方が寒いと感じる人が多いことがわかる．B は，服装の厚さが 1 区分目（最も薄着）で最も若い人物群の帯グラフである．男性と女性の帯グラフの横幅に着目すると，男性に比べて女性の比率がかなり高いことがわかる．また，男性と女性の数値分布を帯グラフで比較すると，女性より男性の方が寒いと感じる人が多いことがわかる．以上のことから，建物がオフィスの場合は服装の厚さが 1 区分目または 3 区分目で年齢が最も若いという局所的な部分集合で温感に男女差が発生していることがわかる．しかし，A と B の部分では男女の分布に偏りがあることは共通しているが，男女間でどのような偏りが生じているかは異なる．このことからも，偏りの判断を計算機に任せるのではなく，人間の解釈を交えることが望ましいと言える．

次に住宅の枠に注目する．中でも服装の厚さが 4 区分目（最も厚着）である C の枠を見ると，内部にある帯グラフが全てカラースケールで表示されている．このことから，建物が住宅で服装の厚さが 4 区分目であると，年齢を問わず男女の温感の判断が食い違う傾向にあることがわかる．また服装の厚さが 1 区分目（最も薄着）である D の枠を見ると，内部にある帯グラフが全てグレースケールで表示されている．このことから，建物が住宅で服装の厚さが 1 区分目であると，年齢を問わず温感に大きな男女差が生じていないことがわかる．

以上の結果からわかるように，属性の組み合わせは同じでも，属性を選択する順番によって可視化結果は異なってくる．属性を選択する順番を変えることで，可視化結果に現れた男女差が特定の属性の影響で生じた男女差であるのか，あるいは複数の属性が絡み合って起こった男女差なのかを分析できる．

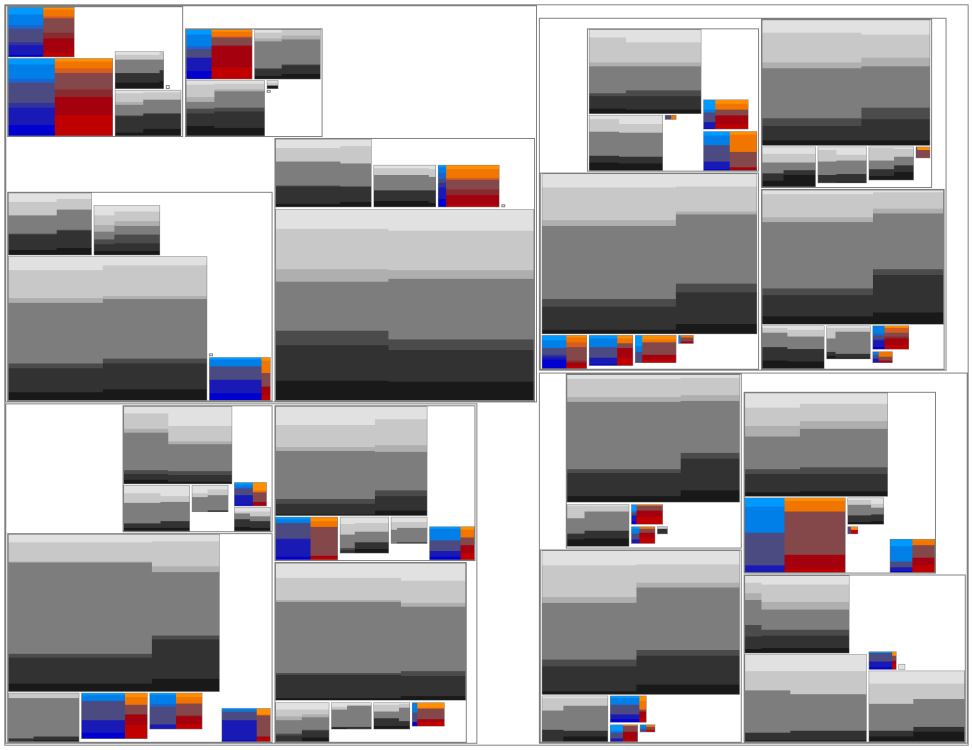


図 5　Age，Cloth，Building の順に人物を分類した可視化結果.

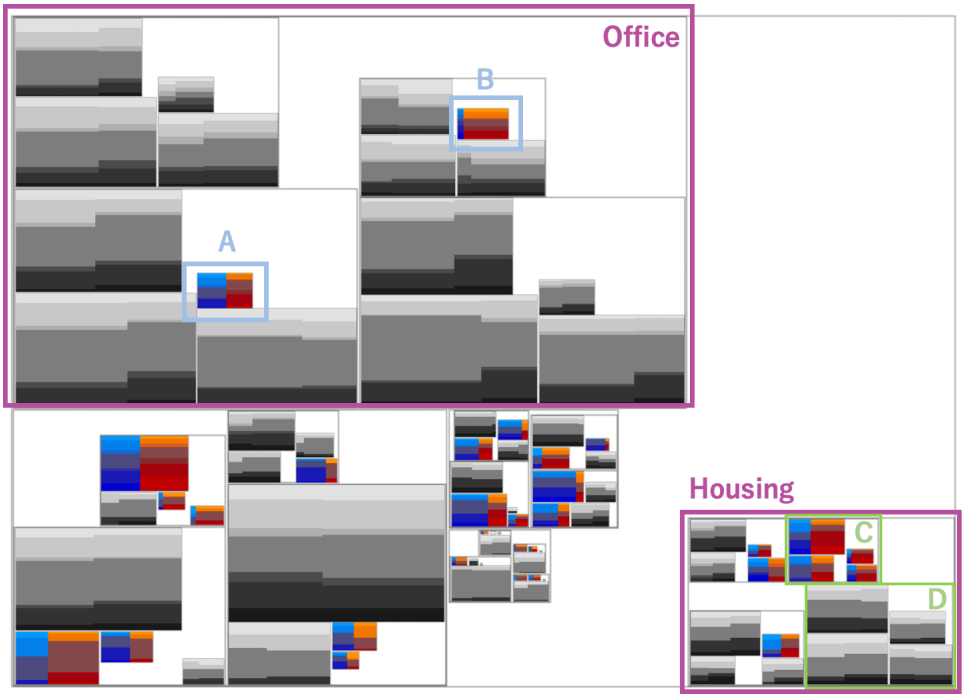


図 6　Building，Cloth，Age の順に人物を分類した可視化結果.

5.3　可視化例 (2)

5.2 節で選択しなかった属性である Metab，Season，Strategy の組み合わせを用いて可視化する．この組み合わせの推薦順位は 14 位である．図 7 は Metab，Season，Strategy の順に属性値を参照して人物を分類した可視化結果である．5.2 節での可視化結果と比較すると，カラースケールで表示されている帯グラフの数がかなり少ない．カラースケールで表示されている帯グラフは可視化結果全体に点在しているが，それぞれの季節の枠に注目すると枠の中の帯グラフが全てグレースケールの部分と反対に枠の中の帯グラフが全てカラースケールの部分が多い．このことから，特定の季節を持つ部分集合において男女差が生じている可能性が高いと考察できる．

そこで次に，Season，Metab，Strategy の順にして，季節を第 1 階層にして可視化結果を表示する．図 8 はその可視化結果

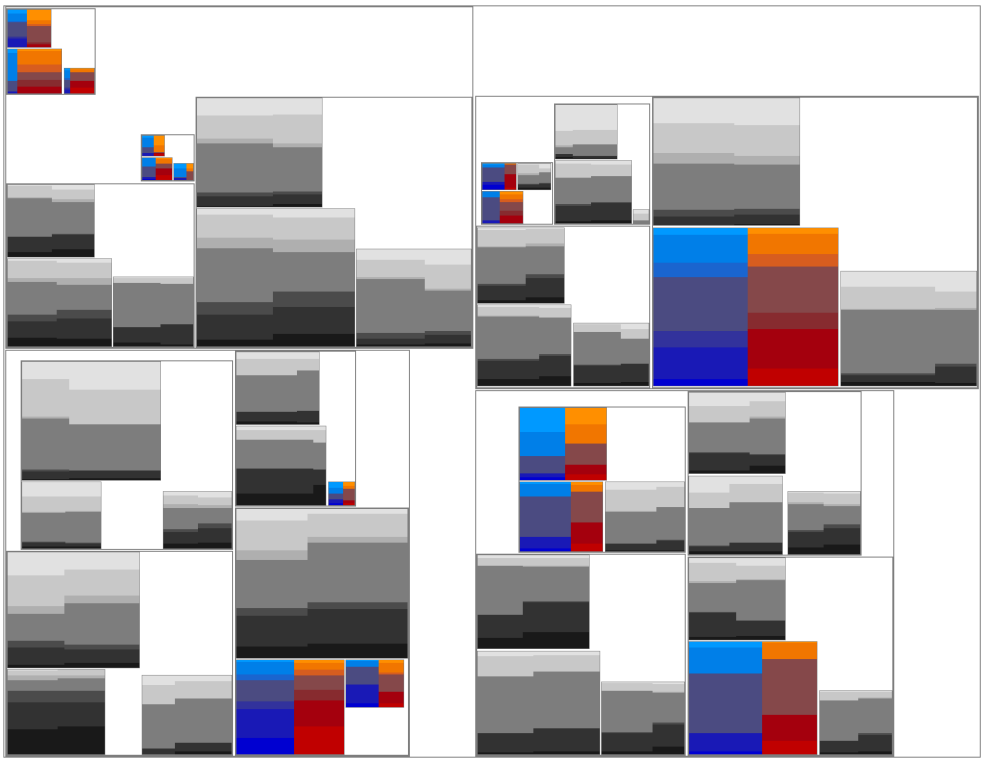


図 7　Metab，Season，Strategy の順に人物を分類した可視化結果.

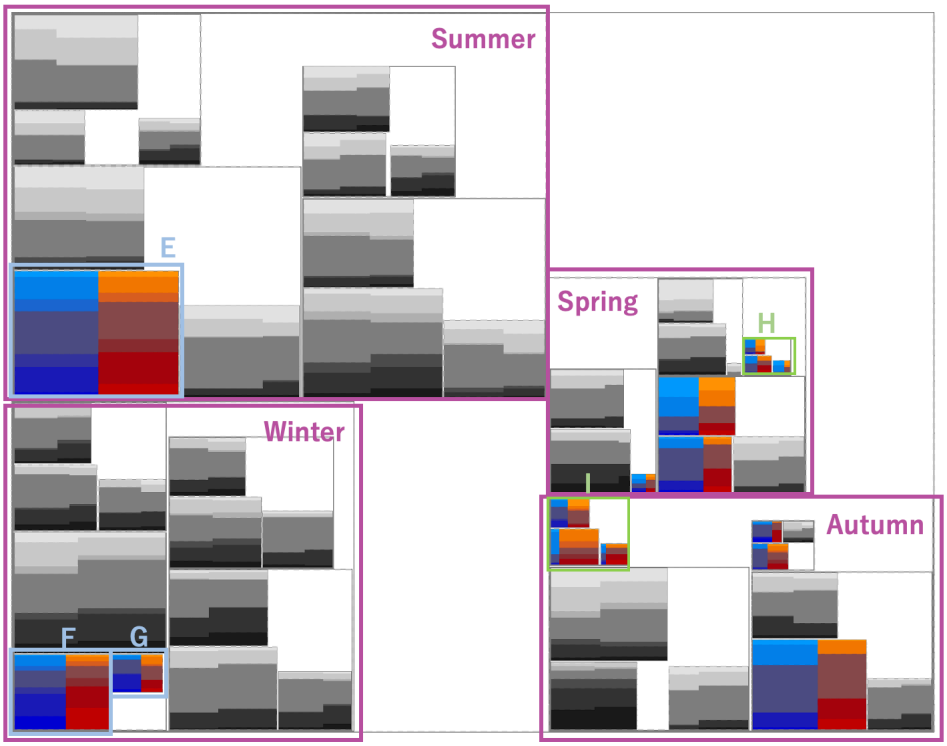


図 8　Season，Metab，Strategy の順に人物を分類した可視化結果.

である．季節が夏と冬の部分には，カラースケールで表示されている帯グラフが非常に局所的に現れている．このことから，季節が夏と冬の場合は代謝量やエアコンの属性値が特定の属性値である場合に男女差が生じると言える．ここでまず，季節が夏の枠に注目する．E は代謝量が 2 区分目（2 番目に代謝量が低い）で空調がエアコンの人物群の帯グラフであり，カラースケールで表示されている．男性と女性の数値分布を帯グラフで比較すると，男性より女性の方が寒いと感じる人が多いことがわかる．このことから，季節が夏の場合は代謝量が 2 区分目で空調がエアコンである人物群という局所的な部分集合で男女差が生じていることがわかる．

次に，季節が冬の枠に注目する．F は代謝量が 3 区分目（2

番目に代謝量が高い）で空調がエアコンの人物群の帯グラフである．男性と女性の数値分布を帯グラフで比較すると，男女ともに寒いと判断している人が全体と比べて多いが，男性よりも女性の方が寒いと感じている人がより多い．また，G は代謝量が 3 区分目（2 番目に代謝量が高い）で空調が混合の人物群の帯グラフである．男性と女性の数値分布を帯グラフで比較すると，男性は全体の約半分が寒いと感じている一方で，女性は暑いと判断している人，ちょうどいいと感じている人，寒いと感じている人の割合が約 1:1:1 である．これらのことから，男女の帯グラフを比較すると，F と G はどちらもカラースケールで表示されているが，どのような男女差が生じているかは異なることがわかる．また，男女差の内訳は異なるものの，季節が冬の場合は，代謝量が 3 区分目（2 番目に代謝量が高い）で空調がエアコンまたは混合である人物群，という局所的な部分集合で男女差が生じていることがわかる．

また，季節が春と秋の場合はカラースケールで表示されている帯グラフが点在している．このことから，季節が春と秋の場合は季節の属性値が男女差に大きく影響を与えている可能性が高い．季節が春の枠の中にある H の枠と，季節が秋の枠の中にある I の枠は，どちらも代謝量が 4 区分目（代謝量が最も高い) の枠である．H と I の枠の中にある帯グラフは全てカラースケールで表示されている．このことから，季節が春か秋で代謝量が 4 区分目であると，空調の属性値に関わらず男女の温感の判断が食い違う傾向にあることがわかる．

6　まとめ・今後の課題

本報告では，多数の人物を対象としたデータ中からデータ分布の男女差を可視化する手法を提案した．データを構成する人物群を属性に沿って階層的に分類し，階層型可視化手法である「平安京ビュー」を用いて可視化する．この際に，末端の階層を表現する長方形領域を 2 列の帯グラフで塗りつぶすことで，特定の属性値を有する男性と女性の数値分布を表現し，特定の属性に起因する数値分布の男女差を可視化する．

さらに，我々自身の先行研究 [2] からの進展として，Earth Mover's Distance (EMD) による男女間の分布差算出を導入する．指定値よりも EMD が高ければ男女の帯グラフをカラースケールで表示し，低ければグレースケールで表示することで，数値分布の男女差が大きな部分を強調表示する．また，可視化システムを操作する際に選択するに値する属性の組み合わせをユーザに推薦する．データを構成する人物群が持つ属性の全ての組み合わせに対して階層構造を生成する．生成した階層構造の末端の階層に含まれる人物群の男性と女性の数値分布の分布間距離を EMD で算出し，算出した EMD の合計値が高い順に属性の組み合わせを提示する．

今後の課題として，以下の三つがあげられる．

一つ目は，帯グラフをカラースケールで表示する場合とグレースケールで表示する場合を分類するための適切な閾値の議論である．本報告では閾値に階層構造の末端の階層に対して算出された EMD の平均値を使用している．EMD の平均値を使用する以外に，例えば中央値など EMD の平均値以外の値を閾値とすることや，可視化システムに閾値を設定するためのスライダーを設け，ユーザ自身で閾値を定めることができるようにすることが考えられる．

二つ目は属性を選択する順番を推薦するシステムの構築である．現在は見るに値する属性の組み合わせを推薦しているが，属性を選択する順番はユーザに委ねている．データの偏りを発見するために必要な反復的操作の回数を少なくするには，単に属性の組み合わせを推薦するだけでなく，どのような順番で属性を選択するのが適切かまで推薦する必要があると考える．

三つ目は空調の温感に関するデータ以外の多様なデータへの適用である．多様なデータに対して可視化を試み，本手法が応用できる幅を広げたい．またその際に，何らかのタスクを与えて可視化システムを実際に操作してもらいその達成度を図る，という形式でのユーザ実験にも取り組むことで，本手法の有効性を検証したい．

文　献

[1] Takayuki Itoh, Yumi Yamaguchi, Yuko Ikehata, and Yasumasa Kajinaga. Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):302–313, 2004.

[2] Yuki Nakai, Takayuki Itoh, Hidekazu Takahashi, Satoshi Nakashima, and Tetsu Yamamoto. Hierarchical data visualization of gender difference: Application to feeling of temperature. In *27th International Conference on Information Visualisation (IV2023)*, pages 178–183. IEEE, 2023.

[3] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 495–508. Springer, 2008.

[4] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

[5] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE, 2019.

[6] Qianwen Wang, Zhenhua Xu, Zhutian Chen, Yong Wang, Shixia Liu, and Huamin Qu. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1470–1480, 2020.

[7] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. How divergent is your data? *Proceedings of the VLDB Endowment*, 14(12):2835–2838, 2021.

[8] Ami Tochigi, Takayuki Itoh, and Xiting Wang. Visualization of bias of machine learning for content recommendation.

[9] David Munechika, Zijie J Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. Visual auditor: Interactive visualization for detection and summarization of model biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 45–49. IEEE, 2022.

[10] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003.

[11] Natkamon Tovanich, Pierre Dragicevic, and Petra Isenberg.

Gender in 30 years of ieee visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):497–507, 2021.

[12] Jiabao Li, Houjiang Liu, Jilie Zeng, Di Wu, Ying Ding, and Alec McGail. Shifting winds: Gendered structures of academic mentorship. In *2022 IEEE VIS Arts Program (VISAP)*, pages 14–29. IEEE, 2022.

[13] Jeromy Carriere and Rick Kazman. Interacting with huge hierarchies: Beyond cone trees. In *Proceedings of Visualization 1995 Conference*, pages 74–81. IEEE, 1995.

[14] Jonh Lamping and Ramana Rao. The hyperbolic browser: A focus+ context technique for visualizing large hierarchies. *Journal of Visual Languages & Computing*, 7(1):33–55, 1996.

[15] Hideki Koike. Fractal views: a fractal-based method for controlling information display. *ACM Transactions on Information Systems (TOIS)*, 13(3):305–323, 1995.

[16] B Johonson and B Scheneiderman. Tree-maps: A space filling approach to the visualization of hierarchical information space. In *IEEE Visualization*, volume 91, pages 275–282, 1991.

[17] Benjamin B Bederson, Ben Shneiderman, and Martin Wattenberg. Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *AcM Transactions on Graphics (TOG)*, 21(4):833–854, 2002.

[18] Michael Balzer and Oliver Deussen. Voronoi treemaps. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 49–56. IEEE, 2005.

[19] Ashrae global thermal comfort database ii. `https://www.kaggle.com/datasets/claytonmiller/ashrae-global-thermal-comfort-database-ii`.

Dawid-Skene集約における不均一タスク数の影響緩和による人間-AIハイブリッドクラウドソーシングの品質向上

田村　匠† 伊藤　寛祥†† 小山　聡††† 森嶋　厚行††

† 筑波大学情報学群　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
††† 名古屋市立大学データサイエンス学部　〒 467–8501 名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †tamura.takumi.ap@alumni.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp,
†††oyama@ds.nagoya-cu.ac.jp

あらまし　近年の AI 技術の発展に伴い，AI が人間と同じようにクラウドソージングに参加する「AI ワーカ」のアイデアが注目されてきている．クラウドソーシングでワーカの能力に偏りがある場合に有効な結果集約手法としては Dawid-Skene 法が知られており，これを AI ワーカと人間ワーカが混在するハイブリッドクラウドソーシングに適用するという考えは自然である．しかし，そのような混在状況では，AI ワーカは明らかに人間ワーカと異なる次のような性質をもち，これらが実際に Dawid-Skene 集約結果にどう影響をもたらすかは自明ではない．(1) 個々の人間と比較して圧倒的に多くのタスクを実施できる．(2) 人間とは全く 異なるバイアスをもたらすことがある．本研究では，以上の問題が実際に Dawid-Skene 集約結果にもたらす影響を調査する．その結果，Dawid-Skene 集約はタスクを多く解く AI ワーカの結果に影響されやすいことがわかった．そこで，我々はこの影響を緩和するための新しい集約手法を提案し，シミュレーション実験によってその影響を緩和できることを確認した．

キーワード　クラウドソーシング，Human-in-the-loop，機械学習

1 はじめに

近年の AI 技術の発展に伴い，AI を組み合わせたクラウドソーシングシステムが注目されてきている．例えば，画像アノテーションタスクの集約に AI モデルを活用する研究 [2,17] や，テキスト分類タスクにおいて，分類に関連すると思われる部分を予め AI によって強調表示し人間にそのタスクを回答させる研究 [13] など，様々な応用が検討されてきている [3]．

そのような人間-AI の協調からもう 1 歩進んだアイデアとして，AI が人間と同じようにマイクロタスクに取り組むモデルが提案され，そのようなモデルに基づいた研究が行われるようになってきている [1,7,8,14]．例えば，Kobayashi らは，人間と同じようにタスクに取り組む「AI ワーカ」が人間ワーカと混在する「Human+AI Crowd」モデルにおける効率的なタスク割り当て手法を提案している [8]．現在においては，HuggingFace Hub[1]のようなインターネット上のプラットフォームから大規模言語モデルなどの様々な AI モデルを入手することや，Kaggle[2] のようなコンペティション型プラットフォームで AI エンジニアをクラウドソースすることで AI モデルを独自開発することが可能になってきている．これらを AI ワーカとして活用することは人的資源に限りのある状況下において有効な選択肢となりうる．

クラス分類問題をクラウドソーシングで解決したい場合，品質向上のために 1 つの分類タスクを複数ワーカに割り当てて，その結果を集約することが多い．結果集約には単純多数決のほか，能力の高いワーカを推定しその回答を優先するような様々なアルゴリズムが提案されている [4,6,15,19]．特に Dawid–Skene (**DS**) 法 [4] はクラウドソーシングにおける標準的な集約手法として知られている．しかし，人間-AI の混在状況において人間ワーカや AI ワーカのタスク回答結果をどのように集約するかについては未解決の部分が多い．一般に，AI ワーカはその単体での性能においては，一部の能力の高い人間ワーカに及ばないことが多いと考えられる（仮に AI 単体で人間の性能を超えることが明らかな場合はクラウドソーシングは不要である）．一方で，そのような能力の高い人間ワーカの数は限られており，そのようなワーカを特定することもまた困難である．さらに，どのワーカが人間ワーカであり，どのワーカが AI ワーカであるかがわからない場合も想定されうる．このような場合，各ワーカの能力を推定しながら真のラベルを推定する DS 法を集約に用いるという考えは自然である．

しかし，そのような混在状況では AI ワーカは人間ワーカと異なる次のような性質を持つと考えられる．

（1） 人間と比べて圧倒的に多くのタスクを実施できる
（2） 人間とは全く異なるバイアスをもたらすことがある

AI はその性質から明らかに，人間ワーカよりも多くのタスクを解くことができる．すなわち，人間-AI の混在状況においては，両者の間に個々が取り組んだタスク数の大きな不均一が生じる．また，AI にはバイアスや公平性の問題があることが知

1：https://huggingface.co/models/
2：https://www.kaggle.com/

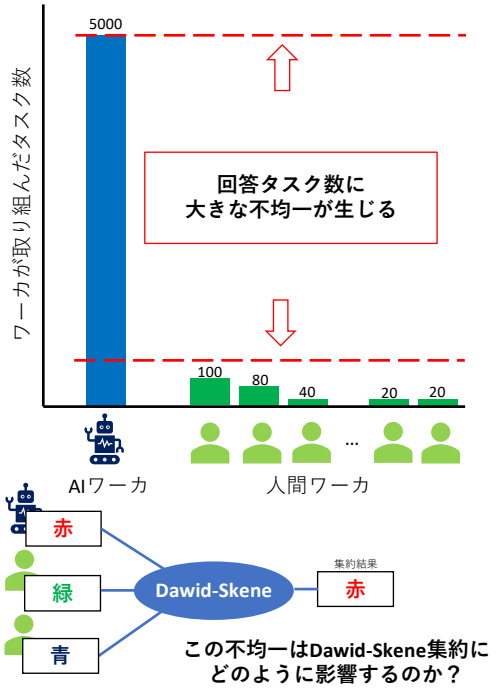


図 1　人間-AI 混在状況において生じるタスク数の不均一の例

られており [11]，AI のタスク結果には人間とは全く異なるバイアスが含まれる可能性がある．DS 法をこのような不均一なタスク結果に適用した場合，どのような影響があるのだろうか（図 1）．

そこで我々は以下のリサーチ・クエスチョン (RQ) を設定し，実験によって RQ を明らかにすることとした．**(RQ1)** 不均一タスク数は DS 集約結果にどのような影響を与えるのか？**(RQ2)** 影響がある場合，その影響を緩和し集約結果を品質向上させることは可能なのか？

RQ1 を明らかにするため，我々は合成データによるシミュレーション実験を行い，タスク数の不均一を伴う人間-AI 混在状況における DS 法や他の集約手法の振る舞いを分析した．その結果，**DS 集約結果はより多くのタスクを解く AI ワーカの回答結果に影響されやすいことがわかった**．AI ワーカが AI 固有のバイアスをタスク結果にもたらす場合，DS 法は集約結果にそのバイアスを反映されてしまうリスクがある．

我々は，DS 法が回答したタスク数が少ないワーカの能力推定に困難を抱えており，結果としてタスクに多く回答する AI ワーカの回答が優先されてしまうと仮説を立てた．DS 法はワーカの能力を混同行列によって推定する．これはクラス数 × クラス数の行列であり，回答タスク数が少ないワーカにとってこの混同行列を正確に推定することは難しい．そこで我々は，ワーカの正答率を利用して混同行列の能力推定を補正し，不均一タスク数の影響を緩和する **CoinFusion** 法を考案した．シュミレーション実験の結果から，提案手法である CoinFusion 法は既存手法と同じような精度を保ちながら，AI ワーカの影響を緩和することができることが明らかになった．

2　関 連 研 究

2.1　ハイブリッド・クラウドソーシング

人間と AI が協調するハイブリッド・クラウドソーシングの考え方は，様々な形で研究されるようになってきている [3]．中でも，AI が人間と同じようにタスクに取り組むことで，クラウドソーシング結果の品質向上を目指すアイデアが提案されてきている [1]．例えば，Kobayashi らは AI が「AI ワーカ」として人間ワーカと同じようにタスクに取り組むシナリオを「Human+AI Crowd」モデルと定義し，その状況における効率的なタスク割り当て手法について提案した [8]．また，Kanda らは Kobayahi らの手法の計算量を改善し，同モデルにおけるより良い割り当て手法を提案している [7]．しかし，これらの研究はタスクの割り当てについて焦点を当てており，タスク結果の集約方法については議論していない．

そのようなシナリオにおけるタスク集約については，Tamura らがクラス分類を行う AI ワーカの出力する確率分布を uncertainty とみなせることに着目し，DS 法を AI ワーカの確率出力に拡張することで集約精度が向上することを報告している [14]．一方で，Tamura らは人間-AI 間に取り組んだタスク数の大きな不均一が生じることについて特に考慮していない．

2.2　Dawid–Skene 法とその拡張

Dawid–Skene (DS) 法 [4] は，クラウドソーシングの品質向上手法としてよく知られた集約手法である．また，DS 法を拡張したタスク集約手法もいくつか提案されている [15, 19]．しかし，これらの手法は人間-AI の混在状況において生じるタスク数の不均一について考慮したものではない．

ワーカの持つバイアスを緩和する集約手法としてはワーカの回答クラスに偏りがある場合（例えば，感情分析において大半ワーカがタスクの多くをネガティブと回答する場合）についてその影響を緩和する手法が提案されている [16, 18] が，本論文では人間とは異なるバイアスを持つ少数の AI ワーカが集約結果に与える影響を対象としており，その状況設定は大きく異なる．

回答したタスク数が少ないワーカの能力推定が困難であることに着目した研究としては，ワーカごとの混同行列とワーカ全体での混同行列を組み合わせワーカ能力の推定を補正する HybridConfusion モデル [9] や，それを発展させ同じような混同行列を持つワーカをクラスタリングし補正に利用する Imamura らのモデル [6] が提案されている．Imamura らは実際のクラウドソージングではワーカ間の回答タスク数に差があることに着目しており，その状況設定は本論文で取り扱う問題に類似しているが，人間-AI 間の回答タスク数が圧倒的に異なるという点において本論文とは状況が異なる．加えてこれらの研究では，タスクを多く解くワーカの回答が集約結果に与える影響については議論していない．本論文ではワーカの正答率を用いて混同行列の推定をシンプルに補正し，人間-AI 間における不均一タスク数の影響を緩和する手法を提案した．

表 1　本論文で用いる記号

記号	説明
$\mathbb{J}$	分類問題のクラス集合
$\mathbb{I}$	全タスク集合
$\mathbb{K}$	（AI ワーカを含む）全ワーカ集合
T_{ij}	タスク i が真のクラス j であるかどうかのバイナリ値 (真のクラスが j ならば $T_{ij}=1$, そうでなければ $T_{ij}=0$)
E_{ij}	T_{ij} の期待値
$n_{ij}^{(k)}$	ワーカ k がタスク i にクラス j と何回回答したか
p_j	クラス j の周辺確率
$\pi_{jl}^{(k)}$	ワーカ k が真のクラスが j である場合に l と回答する確率

3　既存集約手法

3.1　Dawid–Skene 法

Dawid–Skene (**DS**) 法 [4] はクラス分類問題の結果に対する教師無しの集約手法で，EM アルゴリズムを用いて，ワーカの能力とクラスの周辺分布を推定することで，タスクの真のクラスを予想する．DS 法はクラウドソーシングにおける効果的な集約手法として知られている．

本論文で用いる記号について表 1 に記号表を示す．クラス分類問題における分類対象のクラスの集合を $\mathbb{J}$ とおく．例えば，画像を見て犬か猫か兎を判断するクラス分類においては，$\mathbb{J}=\{$ 犬, 猫, 兎 $\}$ となる．また，タスクの集合を $\mathbb{I}$ とおき，タスクに取り組んだワーカの集合を $\mathbb{K}$ とおく．この集合 $\mathbb{K}$ には AI ワーカも含まれる．各タスク $i\in\mathbb{I}$ の真のクラスは T_{ij} で表される．タスク i の真のラベルが $j\in\mathbb{J}$ であるとき $T_{ij}=1$，そうでないとき，$T_{ij}=0$ となる．DS 法は教師データなしで，すべてのタスクにおける T_{ij} の期待値 E_{ij} を予測することを目的としている．$n_{ij}^{(k)}$ はモデルの入力として与えられる唯一の観測変数で，ワーカ $k\in\mathbb{K}$ がタスク i にクラス j と何回回答したかを表している．

EM アルゴリズムに基づく DS 法は，期待値の計算を行う E-step と最尤法により潜在変数を推定する M-step を反復することで，真のラベル T_{ij} を予測する．DS 法は各クラス j の周辺確率 p_j と各ワーカの混同行列である $\pi^{(k)}$ を潜在変数として持つ．混同行列 $\pi^{(k)}$ は $|\mathbb{J}|\times|\mathbb{J}|$ 行列であり，その各要素 $\pi_{jl}^{(k)}$ はワーカ k が真のクラスが j である場合に l と回答する確率を表している（すなわち，$\sum_{l\in\mathbb{J}}\pi_{jl}^{(k)}=1$）．DS 法では，この混同行列をワーカの能力とみなして真のクラスを推定する．

E-step では期待値 E_{ij} を計算する．

$$E_{ij}=\frac{p_j\prod_{l\in\mathbb{J}}\prod_{k\in\mathbb{K}}(\pi_{jl}^{(k)})^{n_{il}^{(k)}}}{\prod_{q\in\mathbb{J}}p_q\prod_{l\in\mathbb{J}}\prod_{k\in\mathbb{K}}(\pi_{ql}^{(k)})^{n_{il}^{(k)}}}. \tag{3.1}$$

M-step では最尤法によって潜在変数を推定する．

$$p_j=\frac{\sum_{i\in\mathbb{I}}E_{ij}}{|\mathbb{I}|}, \tag{3.2}$$

$$\pi_{jl}^{(k)}=\frac{\sum_{i\in\mathbb{I}}E_{ij}n_{il}^{(k)}}{\sum_{m\in\mathbb{J}}\sum_{i\in\mathbb{I}}E_{ij}n_{im}^{(k)}}. \tag{3.3}$$

E_{ij} の初期値は次のように定めるとよいことが経験的に知られている [4].

$$E_{ij}=\frac{\sum_{k\in\mathbb{K}}n_{ij}^{(k)}}{\sum_{l\in\mathbb{J}}\sum_{k\in\mathbb{K}}n_{il}^{(k)}}. \tag{3.4}$$

3.2　他の集約手法

本論文では DS 法以外の既存集約手法として，OneCoin モデル [19] と GLAD [15] を実験に利用し，その振る舞いを DS 法と比較した．両手法は DS 法と同じように EM アルゴリズムを用いて真のクラスを推定する．

OneCoin モデル:　OneCoin モデルは DS 法を単純化したモデルで，ワーカの能力を混同行列ではなく正答率のみで見積もる．すなわち，DS 法ではワーカの能力を $|\mathbb{J}|\times|\mathbb{J}|$ 個の変数からなる混同行列で見積もるが，OneCoin モデルでは 1 個の変数のみで OneCoin に見積もることを意味する．

GLAD:　GLAD は OneCoin モデルと同じようにワーカの能力を変数 α のみで見積もり，またタスクの難易度を変数 β で見積もる．

4　問題設定と合成データ実験

4.1　タスク生成と問題設定

人間ワーカと AI ワーカのタスク数の不均一が DS 集約の結果に与える影響を評価するために，合成データによるシミュレーション実験を行った．はじめに，人工的にデータセットとそれに基づくタスクを生成し，次にそのタスクデータに対する合成人間ワーカや AI ワーカのタスクへの回答を作成する．その後人間と AI の結果を集約し，その集約手法の性能を評価する．

(1) データセットとタスクの作成　まず，10,000 件データアイテムからなる 4 クラス分類のデータセットを次のように作成する．(a) 各データアイテムは 2 つの特徴量 1,2 を持つ．(b) 4 つのクラス (A, B, C, D) のいずれかのラベルを持ち，それぞれ 2,500 件ずつとする (c) それぞれのクラス毎に，特徴量 1,2 が表 2 に示す異なる正規分布に従う値を持っている．図 2 にこのデータセットの分布を示す．このデータセットのうち 7,000 件を訓練データに利用し，3,000 件をテストデータ（タスク）として扱う．

(2) 問題設定　我々の実験目的は，人間ワーカと AI ワーカの取り組むタスク数が不均一であり，かつ，AI ワーカが人間と異なるバイアスを持っている場合に，DS 法による集約結果がどのような影響を受けるかを調査することである．そこで，2 種類の人間ワーカと 3 種類のバイアスを伴う AI ワーカを作成した．AI ワーカは上記の特徴量 1,2 を受け取り，そのデータが属するクラスを予想する機械学習モデルとして実装されるが，そのモデルにはその予測結果がバイアスされたものとなるように人為的な操作を加えた．

4.2　人間ワーカの生成

テストデータの 3,000 タスクに対して，1 人のワーカが 20 タスクずつ取り組み，1 タスクあたり 5 人の人間ワーカを割り当

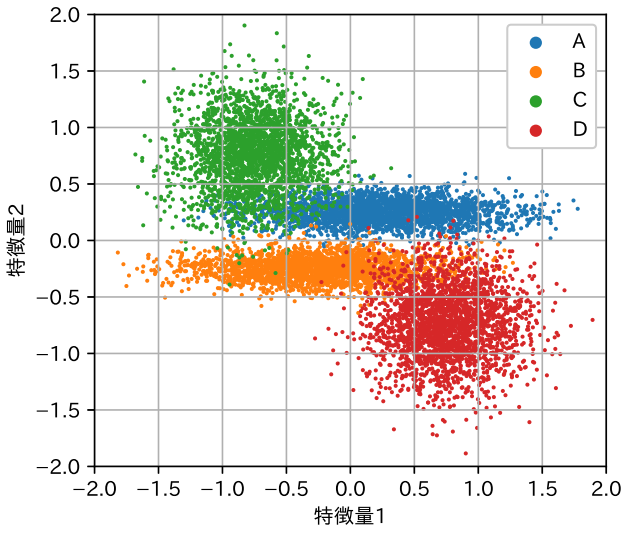


図 2　正規乱数を用いて生成したデータセット

表 2　図 2 のデータ生成に用いた正規分布

クラス	特徴量 1	特徴量 2
A	$\mathcal{N}(0.25, 0.5^2)$	$\mathcal{N}(0.25, 0.1^2)$
B	$\mathcal{N}(-0.25, 0.5^2)$	$\mathcal{N}(-0.25, 0.1^2)$
C	$\mathcal{N}(-0.75, 0.3^2)$	$\mathcal{N}(0.75, 0.3^2)$
D	$\mathcal{N}(0.75, 0.3^2)$	$\mathcal{N}(-0.75, 0.3^2)$

てるという設定で合成人間ワーカを作成した．この場合，合計 750 人のワーカを生成することになる．750 人のワーカは 150 人ごとにそれぞれ能力に差がつくように設定し，ワーカの能力を 5 段階に分けた．そのうえで，「混同行列の異なるワーカ」と「正答率のみ異なるワーカ」という 2 種類の合成人間ワーカを作成した．どのワーカがどのタスクに割り当てられるかはランダムとし，同一のタスクに複数回ワーカが割り当てられることも許容した．

異混同行列ワーカ (Different Confusion Matrix, **DCM**)

DS 法が本来想定しているような，ワーカ間でその混同行列が異なる設定において人間ワーカを作成した．はじめに，訓練データを用いて正規分布を仮定したナイーブベイズ分類器を訓練する．その後，ワーカ k ごとに能力値 $\alpha^{(k)}$ と差分 $r_1^{(k)}, r_2^{(k)}$ を設定する．能力値 $\alpha^{(k)}$ は $\{0.1, 0.2, 0.4, 0.7, 1.1\}$ のいずれかの 5 段階の値からランダムに割り当てられ，差分 $r_1^{(k)}, r_2^{(k)}$ は範囲 $[-1, 1)$ の一様乱数として設定される．タスクの特徴量が (f_1, f_2) で与えられるとき，そのタスクへのワーカ k の回答は訓練した分類器に対して座標 $(a^{(k)}r_1^{(k)} + f_1, a^{(k)}r_2^{(k)} + f_2)$ を入力して得られたクラスとなる．

異正答率ワーカ (Different Accuracy, **DAcc**)

ワーカ間で正答率は異なるものの，その混同行列については大きな差がないような人間ワーカを作成した．これはワーカの能力を正答率のみで見積もる OneCoin モデルや GLAD などの集約手法が想定している設定である．DAcc では各ワーカ k に正答率 $a^{(k)}$ を $\{0.95, 0.85, 0.75, 0.65, 0.55\}$ の中からランダムに割り当てる．ワーカ k は確率 $a^{(k)}$ でタスクに正解し，$1 - a^{(k)}$ でタスクを誤答するものとする．誤答する場合，その答えはそのタスクの正解クラスによって一意に決定する．正解クラスが A であれば C と，B であれば D と，C であれば A と，D であれば B と誤答する．図 2 を見るとわかるように，これらは実際に最も間違えやすいクラスに対応している．

表 3　合成人間ワーカのみでの集約精度（エラー率）

集約手法	異混同行列 (DCM)	異正答率 (DAcc)
DS	**0.073**	0.114
OneCoin	0.075	**0.081**
GLAD	0.085	0.096
単純多数決	0.114	0.099

表 4　誤ったサンプリングを伴う AI ワーカの誤った訓練データ生成に用いた正規分布

クラス	特徴量 1	特徴量 2
A	$\mathcal{N}(0.4, 0.2^2)$	$\mathcal{N}(0.4, 0.4^2)$
B	$\mathcal{N}(-0.4, 0.2^2)$	$\mathcal{N}(-0.4, 0.4^2)$

合成人間ワーカのみでの集約精度

合成人間ワーカの回答結果を DS 法などの集約手法で集約し，そのエラー率（1-正答率）を評価したものを表 3 に示す．これを見ると，混同行列に差がある DCM では DS 法が最もエラー率が低く，正答率のみが異なる DAcc では OneCoin モデルが最もエラー率が低いという想定された結果になっていることがわかる．OneCoin モデルや GLAD はワーカの能力を正答率のみで見積もるモデルであり，DAcc で強く DCM で弱い．

4.3　AI ワーカの生成

人間にはないバイアスをもたらす AI ワーカとして 3 種類の AI ワーカを作成した．AI ワーカはすべて訓練データを用いて訓練された正規分布を仮定したナイーブベイズ分類器として実装される．AI ワーカのバイアスについては，実際の機械学習におけるバイアスの問題 [11] を参考にした．

特徴量の欠落

特徴量 1 を欠落させ，特徴量 2 のみで分類器を訓練した．これは AI が本来知るべき特徴量を知らないために，特定のデータに対して誤った分類を行ってしまうというバイアスを想定している．この AI 単体でのエラー率は 0.093 であった．

誤ったサンプリング

クラス A と B に対し，表 4 に示す正規分布から誤ったデータを 3,000 件ずつ生成し，訓練データに加えた．その誤ったデータを含む訓練データから分類器を訓練した．これは選択バイアスなどによって，誤ったデータを訓練してしまった AI を想定しており，AI 単体でのエラー率は 0.118 であった．

偏った能力

「偏った能力」の AI ワーカは，クラス A とクラス B の識別に困難を抱えているが，クラス C とクラス D の分類は過不足なく行える．この AI ワーカでは，訓練データをそのまま用いて訓練された分類器を用いて回答を作成するが，分類器がクラス A または B を予想した場合，0.5 の確率で，A と予想していた場合は B，B と予想していた場合は A を回答してしまうように

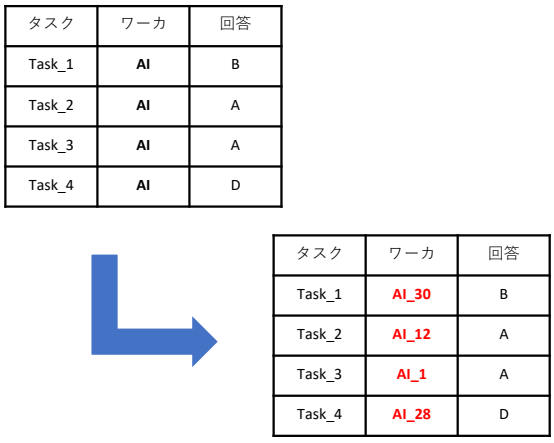


図 3　AI ワーカの 1 人あたりの取り組むタスク数を N とし，AI の予測結果を $3,000/N$ ワーカにランダム分割する

する．単体でのエラー率は 0.287 であった．

4.4　シミュレーション実験

不均一タスク数が DS 集約にどのような影響を与えるのかを調査するため，作成した合成人間ワーカと AI ワーカをを用いてシミュレーション実験を行った．実験では，AI ワーカ 1 人あたりのタスク数 N を設定し，AI ワーカの回答をランダムに分割した（図 3）．N の値が大きくなればなるほど，タスク数の不均一さは大きくなるため，N の値を変化させた場合の集約結果を分析することで，不均一タスク数が集約結果に与える影響を評価できる．我々は DS 集約結果がタスクを多く解く AI ワーカの回答結果に影響されやすいと仮説を立て，集約結果のエラー率に供せて，集約結果と AI ワーカの回答の類似度であるカッパ係数 (Cohen's κ) を計測した．

2 種類の人間ワーカと 3 種類の AI ワーカからそれぞれ 1 つずつ選択し，そのすべての組み合わせについて，以下の手順で実験を行った．

（1）　AI ワーカにテストデータの 3,000 タスクをすべて回答させる．

（2）　AI ワーカ 1 人あたりの回答数 N を定め，図 3 のように 1 の回答を分割する．

（3）　人間ワーカの結果と AI ワーカの結果を集約する．

（4）　集約結果のエラー率と，集約結果と 1 の AI ワーカの回答の類似度（カッパ係数）を計算する．

（5）　3,000 のすべての約数を N の値とし，2〜4 を繰り返す．$N = 3000$ のとき，AI ワーカは 1 人となり，$N = 1$ のとき AI ワーカは 3,000 人となる．集約手法としては DS 法のほか，比較手法として OneCoin モデルと GLAD を用いた．

5　実験結果

図 4 に各人間ワーカ・AI ワーカの組み合わせにおけるエラー率とカッパ係数の変化を示す．グラフの横軸は AI ワーカ 1 人あたりのタスク数を表しており，右に行くほどタスク数の不均一さが大きくなる．縦軸はエラー率またはカッパ係数を表している．カッパ係数は AI ワーカの予測結果と集約結果の類似度を表しており，カッパ係数が大きいほど，集約結果が AI の影響を受けてしまっているといえる．

これを見ると，DS 法はすべての場合で，不均一さの変化に伴って，エラー率・カッパ係数が大きく変化していることがわかる．DS 法では，不均一さが大きくなればなるほどカッパ係数も大きくなっている．多くの場合エラー率の減少とカッパ係数の上昇はトレードオフの関係にあるが，DS 法ではエラー率が他の手法より高いにもかかわらず，カッパ係数も他の手法より大きいケース（図 4b における「特徴量の欠如」）も観測されている．DS 集約の結果が AI の予測結果と類似していることは AI のバイアスが強く集約結果に反映されてしまっていることを意味する．

また，比較手法の OneCoin モデルと GLAD では，エラー率やカッパ係数の変化の幅が DS 法より小さく，不均一さの影響を受けづらいこともわかった．一方で，OneCoin モデルや GLAD はワーカの能力を正答率のみで評価するため，異混同行列 (DCM) 合成人間ワーカや「偏った能力」の AI ワーカを用いた場合においてはエラー率が DS 法より高い．

上記のことから，**DS 法はワーカ間のタスク数の不均一である場合に，タスクを多く解いたワーカの回答に集約結果が影響されやすい**ということがわかる．この性質は AI ワーカが人間と異なるバイアスをタスク結果にもたらす場合に，そのバイアスを増長してしまう恐れがある．

6　提案手法

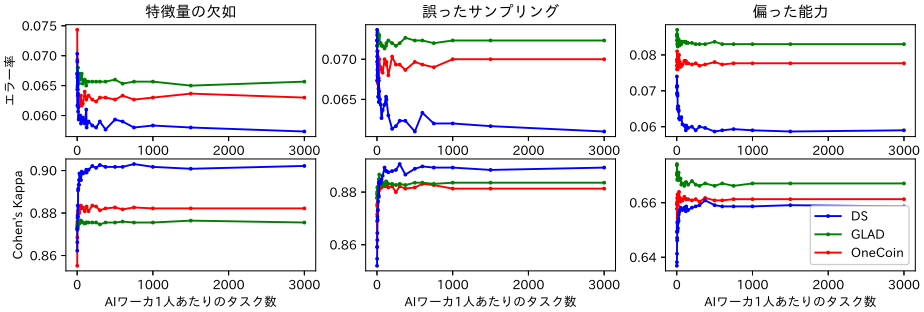
6.1　アイデア

シミュレーション実験の結果（図 4）から DS 法はタスク数が不均一な場合に，AI ワーカのタスク結果に強い影響を受けてしまうことが明らかになった．一方で，エラー率の減少と AI ワーカとの回答の類似度は多くの場合トレードオフの関係にあることもわかった．そこで，エラー率を低く保ちつつ，AI ワーカの影響を緩和できる手法を提案する．

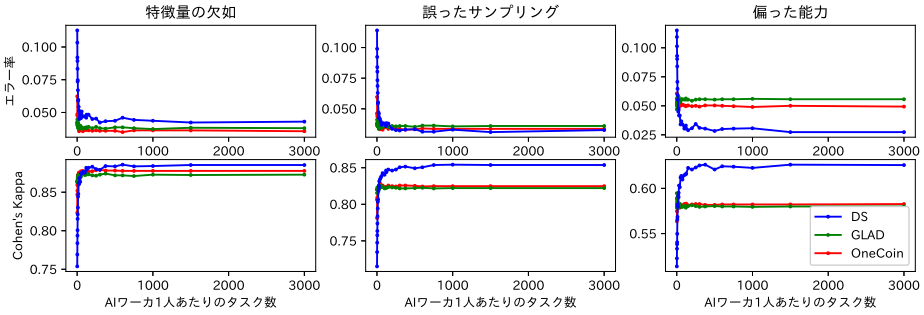
我々は，ワーカの能力を正答率のみで見積もる OneCoin モデルと GLAD がタスク数不均一の影響を受けづらいことに着目した．OneCoin モデルや GLAD では，ワーカの能力を正答率という 1 変数のみで推定するのに対し，DS 法では混同行列，すなわち，$|\mathbb{J}| \times |\mathbb{J}|$ 個の変数を用いて推定する．DS 法は式 (3.3) において混同行列の各要素を最尤推定してるが，多くのタスクを解いてない人間ワーカにおいて，この混同行列の推定が正しく行われず，結果として多くのタスクを解くワーカの結果が優先されてしまうのではないか？

最尤推定はその推定に用いるサンプルサイズが小さい場合にうまく推定を行えないことがあり，いくつかの先行研究では集約アルゴリズムにおける最尤推定をベイズ化することでその集約精度が向上したことが報告されている [5, 10]．

そこで我々は，式 (3.3) の最尤推定をワーカの正答率を用い



(a) 異混同行列 (DCM) 合成ワーカ



(b) 異正答率 (DAcc) 合成ワーカ

図 4　AI ワーカ 1 人あたりのタスク数を変化させた際のエラー率・カッパ係数の変化

た MAP 推定に拡張し，混同行列 (**Confusion** Matrix) によるワーカの能力の推定と，正答率によるワーカの能力の推定 (すなわち，One **Coin** による推定) を組み合わせる **CoinFusion** 法を提案する．

6.2　CoinFusion 法

CoinFusion 法は混同行列によるワーカの能力推定と正答率によるワーカの能力推定を組み合わせることで，タスクを多く解いていないワーカの能力推定を改善させる DS 法の拡張手法である．DS 法の E-step については変更を加えず，M-step において正答率の推定を追加で行い，その値を用いて混同行列を更新する．

M-step のはじめに，以下の式で各ワーカ k の正答率 $a^{(k)}$ を推定する．

$$a^{(k)} = \frac{1}{\sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}} n_{ij}^{(k)}} \sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}} E_{ij} n_{ij}^{(k)} \tag{6.1}$$

ワーカ k の混同行列の各行 $\pi_j^{(k)}$ はカテゴリ分布のパラメータとみなせることが知られている [12]．そこで，$\pi_j^{(k)}$ の事前共役分布として次のようなディリクレ分布を仮定する．

$$\pi_j^{(k)} \sim \mathrm{Dir}(\xi_j^{(k)}) \tag{6.2}$$

$\xi_j^{(k)}$ はワーカ k の混同行列 $\pi^{(k)}$ の j 行目 $\pi_j^{(k)}$ に関する事前共役分布であるディリクレ分布のパラメータであり，長さ $|\mathbb{J}|$ のベクトルである．

ここで，このベクトルの各要素 $\xi_{jl}^{(k)}$ を次のように定義する．ここで，T は提案手法 CoinFusion 法のハイパーパラメータで，正整数とする．

$$\xi_{jl}^{(k)} = \begin{cases} a^{(k)}T & \text{if } j = l, \\ \frac{1-a^{(k)}}{|\mathbb{J}|-1}T & \text{otherwise} \end{cases}. \tag{6.3}$$

そして，CoinFusion 法では，DS 法の式 (3.3) を MAP 推定で次のように置き換える．

$$\pi_{jl}^{(k)} = \frac{\sum_{i\in\mathbb{I}} E_{ij} n_{il}^{(k)} + \xi_{jl}^{(k)}}{\sum_{m\in\mathbb{J}}\sum_{i\in\mathbb{I}} E_{ij} n_{im}^{(k)} + \sum_{m\in\mathbb{J}} \xi_{jm}^{(k)}}. \tag{6.4}$$

$\sum_{m\in\mathbb{J}} \xi_{jm}^{(k)} = T$ であることを踏まえると，混同行列の推定式は次のようにも表せる．

$$\pi_{jl}^{(k)} = \begin{cases} \frac{\sum_{i\in\mathbb{I}} E_{ij} n_{il}^{(k)} + a^{(k)}T}{\sum_{m\in\mathbb{J}}\sum_{i\in\mathbb{I}} E_{ij} n_{im}^{(k)} + T} & \text{if } j = l, \\ \frac{\sum_{i\in\mathbb{I}} E_{ij} n_{il}^{(k)} + \frac{1-a^{(k)}}{|\mathbb{J}|-1}T}{\sum_{m\in\mathbb{J}}\sum_{i\in\mathbb{I}} E_{ij} n_{im}^{(k)} + T} & \text{otherwise} \end{cases}. \tag{6.5}$$

周辺確率 p_j の推定については DS 法と同じように行う．

7　提案手法の評価

提案手法の性能を評価するため再度シミュレーション実験を行った．4 節の実験において，AI ワーカ 1 人あたりののタスク数を 3,000 件，つまり，すべてのタスクを 1 人の AI ワーカが解く，最も不均一な場合において提案手法と既存集約手法のエラー率とカッパ係数を比較した．表 5 は，AI ワーカ 3 種と人間ワーカ 2 種のすべての組み合わせにおける，各集約手法のエラー率とカッパ係数（AI ワーカの回答と集約結果の類似度）を示したものである．

これを見ると，CoinFusion 法はほとんどの場合において，DS 法よりもカッパ係数を減少させていることがわかる．これ

は人間ワーカの結果が DS 法よりも活用されやすくなったことを表している．また，カッパ係数の減少とエラー率の上昇はトレードオフの関係にあるが，OneCoin モデルや GLAD と比較すると CoinFusion 法ではエラー率の上昇を最小限に抑えつつ，カッパ係数を減少させることができている．中にはカッパ係数を減少させながら，エラー率も減少させているケースもある．

また，図 5 に 4 節の実験と同じように，AI ワーカ 1 人あたりのタスク数を変化させた場合の CoinFusion 法のエラー率とカッパ係数の変化を示す．これを見ると，CoinFusion 法のハイパーパラメータである T の値が大きくなればなるほど，不均一さの変化に応じたエラー率やカッパ係数の変化が小さくなり，また，T の値が大きくなるほど，カッパ係数が減少し，エラー率が上昇するトレードオフの関係があることがわかる．ハイパーパラメータ T は，混同行列による能力の推定と，正答率による能力の推定のバランスを調整する重みの役目を果たしており，T の値を大きくすると，正答率のみでワーカ能力を推定する OneCoin モデルや GLAD にその振る舞いが近づく．

総じて，CoinFusion 法は AI ワーカ 3 種・人間ワーカ 2 種の計 6 通りのすべての組み合わせにおいて，エラー率を低く保ちながら，その集約結果における AI ワーカの影響を小さくすることに成功している．これは，正答率によるワーカ能力の推定と混同行列によるワーカ能力の推定を MAP 推定により組み合わせることで，回答したタスク数が少ない人間ワーカの能力推定をより正確にし，人間ワーカの回答を DS 法よりも優先することが可能になったからだと考えられる．

8 おわりに

本論文は，AI が人間のワーカと同じようにクラウドソーシングに参加する場合において，それら AI ワーカと人間ワーカとの間に取り組んだタスク数に大きな不均一が生じることに着目し，その不均一さが DS 集約にどのような影響を与えるのかを調査した．その結果，DS 集約結果はタスクを多く解いたワーカ回答に影響されやすことがわかった．AI ワーカは人間ワーカと比較して圧倒的に多くのタスクに取り組め，また，人間と異なるバイアスをタスク結果にもたらす可能性があり，DS 集約はこのようなバイアスを集約結果に反映させてしまうリスクがあることが明らかになった．

そこで，我々はこの影響を緩和するための新しい集約手法である CoinFusion 法を提案し，評価実験によって，提案手法が精度を保ちながらバイアスを持つ AI ワーカの影響を緩和できることを確認した．一方で，CoinFusion 法のハイパーパラメータを適切に設定する方法を明らかにすることや，実世界データにおける実験評価，さらには，複数の AI ワーカがクラウドソーシングに参加する場合の集約における影響を調査することは今後の課題である．

謝　辞

本研究の一部は，JSPS 科研費 (JP21H03552, JP22H00508, JP23H03405) と JST CREST (JPMJCR21D1, JPMJCR22M2) の支援を受けたものである．ここに謝意を示す．

表 5　最も不均一な場合の各集約手法のエラー率とカッパ係数

(a) AI ワーカが「特徴量の欠落」である場合

人間ワーカ	異混同行列 (DCM)		異正答率 (DAcc)	
集約手法	エラー率	Cohen's κ	エラー率	Cohen's κ
DS	0.057	0.902	0.043	0.885
CoinFusion@T=1e1	0.059	0.888	0.037	0.871
CoinFusion@T=1e2	0.062	0.878	0.041	0.860
CoinFusion@T=1e3	0.064	0.876	0.040	0.863
OneCoin	0.063	0.882	0.036	0.878
GLAD	0.066	0.876	0.038	0.873

(b) AI ワーカが「誤ったサンプリング」である場合

人間ワーカ	異混同行列 (DCM)		異正答率 (DAcc)	
集約手法	エラー率	Cohen's κ	エラー率	Cohen's κ
DS	0.061	0.889	0.033	0.854
CoinFusion@T=1e1	0.062	0.888	0.028	0.835
CoinFusion@T=1e2	0.066	0.879	0.035	0.822
CoinFusion@T=1e3	0.067	0.879	0.034	0.823
OneCoin	0.070	0.881	0.034	0.825
GLAD	0.072	0.884	0.036	0.822

(c) AI ワーカが「偏った能力」である場合

人間ワーカ	異混同行列 (DCM)		異正答率 (DAcc)	
集約手法	エラー率	Cohen's κ	エラー率	Cohen's κ
DS	0.059	0.659	0.027	0.626
CoinFusion@T=1e1	0.058	0.660	0.026	0.603
CoinFusion@T=1e2	0.063	0.656	0.034	0.593
CoinFusion@T=1e3	0.070	0.654	0.043	0.586
OneCoin	0.078	0.661	0.049	0.583
GLAD	0.083	0.667	0.056	0.580

文　献

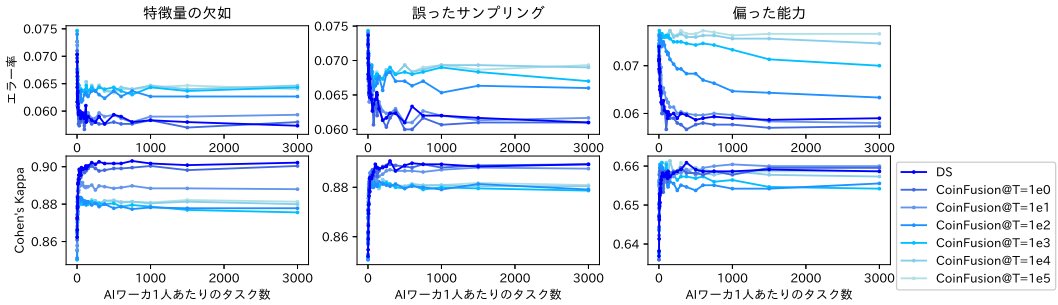
[1] Sihem Amer-Yahia, Senjuti Basu Roy, Lei Chen, Atsuyuki Morishima, James Abello Monedero, Pierre Bourhis, François Charoy, Marina Danilevsky, Gautam Das, Gianluca Demartini, Shady Elbassuoni, David Gross-Amblard, Emilie Hoareau, Munenari Inoguchi, Jared Kenworthy, Itaru Kitahara, Dongwon Lee, Yunyao Li, Ria Mae Borromeo, Paolo Papotti, Raghav Rao, Sudeepa Roy, Pierre Senellart, Keishi Tajima, Saravanan Thirumuruganathan, Marion Tommasi, Kazutoshi Umemoto, Andrea Wiggins, and Koichiro Yoshida. Making AI machines work for humans in FoW. *ACM SIGMOD Record*, Vol. 49, No. 2, pp. 30–35, 2020.

[2] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6109–6118, 2017.

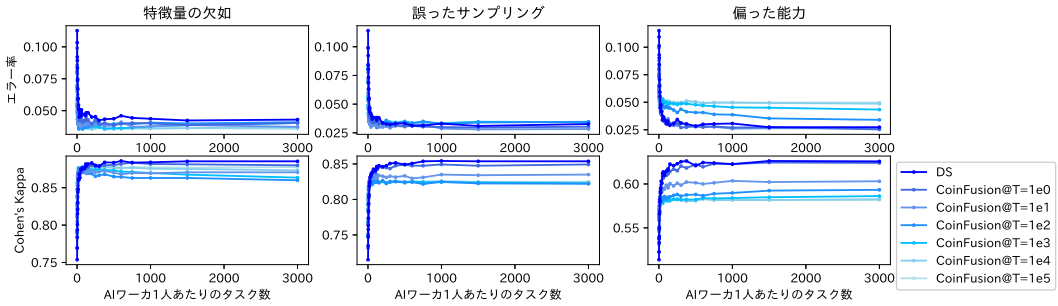
[3] António Correia, Andrea Grover, Daniel Schneider, Ana Paula Pimentel, Ramon Chaves, Marcos Antonio de Almeida, and Benjamim Fonseca. Designing for hybrid intelligence: A taxonomy and survey of crowd-machine interaction. *Applied Sciences*, Vol. 13, No. 4, 2023.

[4] A.P. Dawid and A.M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28, 1979.

[5] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In



(a) 異混同行列 (DCM) 合成ワーカ



(b) 異正答率 (DAcc) 合成ワーカ

図 5　CoinFusion 法における AI ワーカ 1 人あたりのタスク数を変化させた際のエラー率・カッパ係数の変化

Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 1120–1130, 2013.

[6] Hideaki Imamura, Issei Sato, and Masashi Sugiyama. Analysis of minimax error rate for crowdsourcing and its application to worker clustering model. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Vol. 80, pp. 2147–2156, 2018.

[7] Tomoya Kanda, Hiroyoshi Ito, and Atsuyuki Morishima. Efficient evaluation of AI workers for the human+AI crowd task assignment. In *Proceedings of IEEE International Conference on Big Data (BigData)*, pp. 3995–4001, 2022.

[8] Masaki Kobayashi, Kei Wakabayashi, and Atsuyuki Morishima. Human+AI crowd task assignment considering result quality requirements. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, Vol. 9, pp. 97–107, 2021.

[9] Chao Liu and Yi-Min Wang. Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, pp. 17–24, 2012.

[10] Qiang Liu, Jian Peng, and Alexander Ihler. Variational inference for crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 1, pp. 692–700, 2012.

[11] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, Vol. 54, No. 6, 2021.

[12] Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, Vol. 6, pp. 571–585, 2018.

[13] Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, Vol. 7, pp. 144–152, 2019.

[14] Takumi Tamura, Hiroyoshi Ito, Satoshi Oyama, and Atsuyuki Morishima. Influence of AI's uncertainty in the Dawid-Skene aggregation for human-AI crowdsourcing. In *Information for a Better World: Wisdom, Well-being, Win-win, 19th International Conference on Information (iConference 2024)*, in press.

[15] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 2035–2043, 2009.

[16] Ming Wu, Qianmu Li, Fei Yang, Jing Zhang, Victor S. Sheng, and Jun Hou. Learning from biased crowdsourced labeling with deep clustering. *Expert Systems with Applications*, Vol. 211, , 2023.

[17] Yu Yamashita, Hiroyoshi Ito, Kei Wakabayashi, Masaki Kobayashi, and Atsuyuki Morishima. HAEM: Obtaining higher-quality classification task results with AI workers. In *Proceedings of the 14th ACM Web Science Conference (WebSci)*, pp. 118–128, 2022.

[18] Jing Zhang, Victor S. Sheng, Qianmu Li, Jian Wu, and Xindong Wu. Consensus algorithms for biased labeling in crowdsourcing. *Information Sciences*, Vol. 382–383, pp. 254–273, 2017.

[19] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, Vol. 1, pp. 1260–1268, 2014.

DEIM Forum 2024 11-27

フィジカルクラウドソーシングにおける分担タスク割当てのための ワーカ依頼順序の最適化

藤永弥太郎† 伊藤 寛祥†† 鈴木 伸崇†† 森嶋 厚行††

† 筑波大学情報学群情報メディア創成学類 〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †yataro.fujinaga.2023b@gmail.com, ††{ito,mori}@slis.tsukuba.ac.jp, †††nsuzuki@slis.tsukuba.ac.jp

あらまし フィジカルクラウドソーシングにおいてワーカに対して効率よくタスクを依頼をすることができれば，迅速な落とし物捜索や迷子の捜索，災害の情報把握など計算機では解くことが難しい問題の解決など様々なことが可能になる．ラベリングなどのオンラインマイクロタスクでは，それを可能なワーカが一般には大量に存在することを仮定できるため，楽観的なタスク割当てで十分なことが多い．しかし，フィジカルクラウドソーシングではタスクを実行するワーカが通常のクラウドソーシングに比べて少ないため，効率の良い戦略を考えなくてはならない．本研究ではワーカへ依頼する順番に注目し，限られたワーカの中で、できるだけ多くを実行するような依頼順序を決定する問題に取り組む. 依頼順序それぞれについて可能世界を考え，コストと遂行確率を計算する. 可能世界とは，依頼順序に沿ってワーカに依頼した場合に起こりうる依頼順序のことである．本論文では，提案手法を用いてワーカへ依頼順序を求め，それらの間にどの程度の差があるかを調べた．本論文の結果から依頼順序を工夫することで，依頼にかかる費用を減らし，タスク全てを実行する確率を増やせることがわかった.

キーワード Spatial Crowdsourcing, Task Assignment

1 はじめに

現在，クラウドソーシングを用いて，様々な問題が解決されている．オンライン上で作業可能なマイクロタスクだけではなく，物理空間での作業が必要となるフィジカルタスクの解決にもクラウドソーシングが利用されるようになってきている．フィジカルタスクのクラウドソーシングをフィジカルクラウドソーシングということにする．

フィジカルクラウドソーシングは，これまで解くことが困難であった問題に対して強力な解決策になる可能性がある．フィジカルクラウドソーシングが活用された例として，レッドバルーンチャレンジ [1] がある．レッドバルーンチャレンジは 2009 年にアメリカ国防高等計画局（DARPA）が行ったプロジェクトであり，このプロジェクトではアメリカ合衆国全土に設置された 10 個の風船をすべて見つけ出すという問題をフィジカルクラウドソーシングによって約 9 時間で解決することに成功している．このように，フィジカルクラウドソーシングはオンライン上では解決できない様々な問題や大きな規模の問題を短時間で解決する可能性を持っている．

本論文では，分担型フィジカルタスクをクラウドソーシングする際に，ワーカにどのような順番で依頼するかを考える．分担型フィジカルタスクとは，与えられたタスクを複数のワーカで分担するようなフィジカルタスクを指す．このような分担型フィジカルタスクの例として，落とし物の捜索や迷子の捜索，災害の状況把握などが挙げられる．フィジカルタスクは，オンラインタスクでは解決できないタスクを扱う一方，フィジカルタスクはオンラインタスクに比べて，ワーカがタスクを引き受けるハードルが高く，フィジカルタスクを実行するワーカを見つけることは簡単な問題ではないといえる [2].

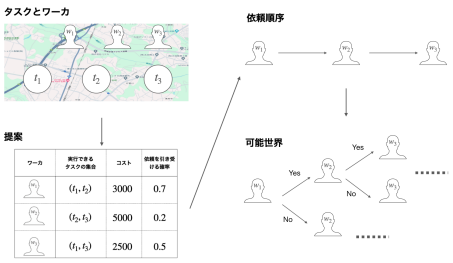


図 1 依頼順序の求め方

分担型フィジカルタスクにおいては適当なタスク割り当てでは，タスク全てを完了できない．これはオンライン上のマイクロタスクでは，一般的にワーカが大量に存在すると考えることができ，オンライン上で作業ができるためタスクの実行に制約が少ないといえる一方で分担型フィジカルタスクではワーカが少なく，オンライン上では作業が完結しないため，タスクの実行に制約が大きいと考えられるからである．分担型フィジカルタスクの割り当てでは数少ないワーカで，できるだけ多くタスクが実行されるような工夫が必要である．そのため分担型フィジカルタスクの割り当てでは，効率の良い戦略を考える必要がある．本論文では，ワーカの数が限られている場合に，できるだけ多くの分担型フィジカルタスクを実行できるような，効率的な依頼方法を明らかにすることを目的とし，ワーカへ依頼順

序を決定するアルゴリズムを提案する．この問題はフィジカルクラウドソーシングの分担割当てで顕著となるが，一方で，提案する手法は，タスク割当ての制約が強い分担割当てが必要な場合には同様に適用可能になるため，そのような場合にも適用可能である．

本論文では，ワーカへの最適な依頼順序を求めるために可能世界を考える．可能世界とは，ある依頼順序に沿ってワーカに依頼したときに，起こりうる全ての依頼順序のことである．可能世界では，ワーカの依頼の承諾状況によって，依頼するワーカを変えていく．可能世界は木構造として表現され，この木構造において根から葉までのパスが依頼順序の１つの場合と対応している（図 2）．図 2 における w_i はワーカを，赤い矢印は依頼順序の１つの場合を，Yes/No はワーカが依頼を引き受けた場合，引き受けなかった場合，Covered/Not Covered はタスク全てを実行できた場合，できなかった場合を表している．実験の結果，依頼順序間でコストの期待値に最大 2408，遂行確率に最大 0.32 の差があることが分かった．コストの期待値とは依頼順序でワーカにいくらの費用を支払うか，遂行確率とは依頼順序に沿って依頼した場合，どの程度の確率でタスク全てを実行することができるかを示している．

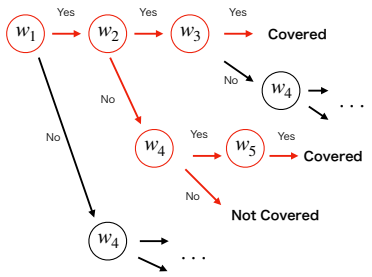


図 2　ある依頼順序の可能世界

本論文は，タスク割当ての制約が強い分担割当てについて，最適な依頼順序を求めるアルゴリズムを開発した．本論文の貢献は次の通りである．（１）ワーカへ依頼する順番を最適化するアルゴリズムを開発できることを示した．（2）生成データのシミュレーションにより，異なる依頼順序間で，必要コストとどれだけのタスクが実施されるかの点に，大きな差があり，この問題が重要である事を明らかにした．

本論文の構成は，次の通りである．まず第２章では，関連研究について述べる．次に第３章では，本論文における問題の定義とその評価について説明する，第４章では提案手法について説明する．第５章では実験について設定，結果，考察をそれぞれ述べる．第６章では，結論と今後の課題について述べる．

2　関 連 研 究

2.1　フィジカルクラウドソーシングのタスク割り当て

フィジカルクラウドソーシングのタスク割り当てについては，いくつかの研究が行われている．Zhao ら [3] はフィジカルタスクをワーカへの割り当て方によって 2 つに分類している．サーバーがワーカにタスクを割り当てる Server Assigned Tasks(SAT) とサーバーが公開しているタスクの中からワーカがタスクを選択する Worker Selected Tasks(WST) である．

SAT については様々な研究が取り組まれている．例えばワーカの移動コスト，ワーカに割り当てたタスク数，ワーカが完了したタスク数を最適化する研究 [4] [5] [6] [7] [8] がある．WST に関する研究には Deng ら [9] の研究がある．Deng らの研究では，ワーカが自律的にタスクを選択するシナリオに焦点を当て，タスクの移動コストと有効期限の両方を考慮しながら，各ワーカが実行するタスク数を最大にするスケジュールを見つける問題を扱っている．

Zhao ら [3] は１つのタスクに必要なワーカの数によってもフィジカルタスクを分類している．１つのタスクに単一のワーカが必要な場合は Single Worker Required Tasks（SWRT），１つのタスクに複数のワーカが必要な場合は Multiple Workers Required Tasks（MWRT）としている．多くの研究では [4] [5] [6] [7] [8] SWRT を扱っているが，MWRT を扱った研究もある．Zhao らの研究 [10] では各タスクで複数のワーカを必要とする場合に，それぞれのタスクに与えられた予算内でタスクスコア（タスクを完了させた際に得られる値のこと）を最大化させるようなタスク割り当てを求める問題を扱っている．

与えられたタスク全てをワーカが実行することを条件づけた研究も存在する．Wang ら [11] の研究では，タスクにカバレッジ制約がある場合にフィジカルクラウドソーシングにかかるコストを最小化させるようなタスク割り当てを扱っている．カバレッジ制約とはどのタスクに対しても少なくとも 1 人のワーカが実行するような制約のことである．フィジカルクラウドソーシングにおいて，このカバレッジ制約がサービスの品質に影響する場合がある．例えば，DroneSense [12] のような天気予報サービス，CalTel [13]，Nericell [14]，GreenGPS [15] のような交通ルート最適化を提供するサービスでは，特定の場所でのデータの欠落が最終的な結果に大きな影響を与える．このようにフィジカルタスクの種類によって，カバレッジ制約が求められるような割り当てが必要になる場合がある．

以上のように先行研究ではフィジカルタスクの割り当てについて研究が数多く存在するが，それらの大半がタスクに割り当てることができるワーカが十分に存在し，ワーカが割り当てたタスクを必ず引き受けるという前提の研究であり，割り当てるタスクの数に制約が存在したり，ワーカへの依頼順序に着目した研究ではない．

2.2　フィジカルクラウドソーシングにおけるワーカリクルーティング

フィジカルタスクを引き受けるワーカについての先行研究がある．Alt ら [2] は，モバイルアプリケーション上で，ワーカが自らタスクを選択して実行するシステムを作成し，クラウドソーシングするフィジカルタスクをタスクを実行する場所，実行する作業内容ごとに分けて，タスクの完了数を比較する実験を行った．その結果，ワーカは物理的な移動を必要とするようなフィジカルタスクはワーカを集めるのが困難であり，オン

ラインタスクに比べて，引き受けるワーカが少ないことがわかった．

プラットフォーム上でワーカが足りない場合に，プラットフォーム外からワーカを集めるという手法についても研究されている．Wang ら [16] は，ワーカが不足している場合のモバイルクラウドソーシングシステムに焦点を当て，ワーカをリクルートする手法を提案した．この手法ではクラウドソーシングするタスクを，クラウドソーシングプラットフォームに登録していない人にタスクを紹介するタスクと実際にタスクを実行するタスクの２種類に分けて，ワーカに依頼している．実データを用いた検証により，この手法がタスクの完了とワーカのリクルートについて既存の手法を上回ることを示した．Xu ら [17] は，モバイルクラウドソーシングプラットフォームに登録されたユーザーに，ソーシャルネットワークでタスクを拡散してもらうことで，ワーカをリクルートする手法を提案している．これらの研究はプラットフォームに十分なワーカがいない場合のワーカリクルーティングに関する研究であるが，本研究ではプラットフォーム上にいるワーカを活用し，できるだけの多くの分担型フィジカルタスクを実行することを目的としている．

2.3 フィジカルタスクのクラウドソーシング

オンラインクラウドソーシングでは誰でもワーカとして参加でき，タスクを実行することができる一方，フィジカルクラウドソーシングでは，特定の場所で実行するタスクで構成されており，ワーカはタスクを実行するために物理的に特定の場所に移動する必要がある．フィジカルタスクをクラウドソーシングするプラットフォームは数多く存在する．例えばリアルタイムのタクシー呼び出しサービスの Uber[1]，DiDi[2] や店頭での商品の陳列などのマイクロタスクを行うプラットフォーム Gigwalk[3]，家具の組み立てや引っ越しなど身の回りのマイクロタスクを行うプラットフォーム Task Rabbit[4]，フードデリバリーサービス GrubHub[5]，Instacart[6]，UberEats[7]，道路の交通状態のモニタリングを行う Waze[8]，ワーカから提供されたデータから地理データを生成する OpenStreetMap[9]がある．

これらのプラットフォームにおける依頼順序決定アルゴリズムは必ずしも公開されていないが，サービスによっては本論文で扱う手法が有効な場合もあると考えられる．

3 問題定義と評価基準

3.1 問 題 定 義

本論文で取り扱う問題は次の通りである．（図 3）

- 入力： タスク集合 T, ワーカ集合 W, 提案集合 P
- 出力： 最適な依頼順序 $o_j \in O$

下記に入力となる「提案」と出力となる「依頼順序」について説明する．

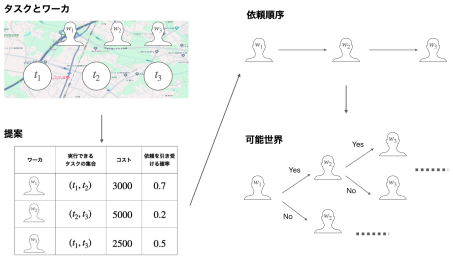
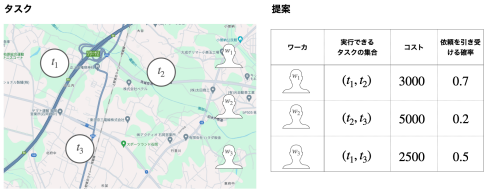


図 3 依頼順序の求め方

【提案】提案 $p_i \in P$ はあるワーカ $w_i \in W$，タスク T の部分集合 $T_i \subset T$，コスト c，依頼を引き受ける確率 q) からなる 4 つ組み (w_i, T_i, c, q) である（図 4）．T' は，その提案をしたワーカが実行することができるタスクの集合を表している．コスト c とは，ワーカが依頼を引き受けた際に，ワーカに支払う謝礼のことである．依頼を引き受ける確率 q とは，ワーカにタスクを実行するように依頼したとき，ワーカが依頼を引き受ける確率のことである．依頼を引き受ける確率が高いほど，依頼をしたときワーカがタスクを実行する可能性が高いといえる．



ワーカ	実行できるタスクの集合	コスト	依頼を引き受ける確率
w_1	(t_1, t_2)	3000	0.7
w_2	(t_2, t_3)	5000	0.2
w_3	(t_1, t_3)	2500	0.5

図 4 提 案 の 例

【依頼順序】依頼順序とは，ワーカへタスクの実行を依頼する順番のことである．依頼順序 o はワーカのリスト $[w_i, \ldots, w_j]$ で表され，リストのインデックスがワーカへ依頼する順番を表している．次で説明するように，依頼順序ごとに，かかるコストの期待値と，T 中のタスクを全て遂行する確率が異なる．本問題における「最適な依頼順序」とは，全ての依頼順序 $O = \{\ldots, o_i \ldots\}$ の中で，T 中のタスクを全て遂行する確率が最もたかいもののうち，最も低いコストの期待値を持つものである．

3.2 依頼順序の遂行確率とコスト期待値

依頼順序に沿って，ワーカに依頼したとき，依頼を受けたワーカは必ずしも依頼を引き受けるとは限らず，依頼を引き受けた場合，依頼を引き受けない場合の両方が考えられる．そのため単に依頼順序に沿ってワーカへ依頼するだけではなく，実際にはワーカの依頼の承諾状況によって場合分けが必要となる．起

1：https://www.uber.com/
2：https://didimobility.co.jp/
3：https://www.gigwalk.com/
4：https://www.taskrabbit.com/
5：https://www.grubhub.com/
6：https://www.instacart.com/
7：https://www.ubereats.com/
8：https://www.waze.com/
9：https://www.openstreetmap.org/

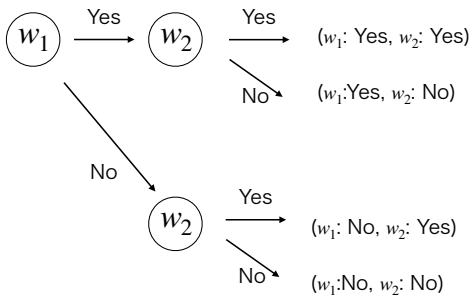


図 5　可能世界の例

こりうる場合分けを全て記述したものを，依頼順序の**可能世界**と呼ぶ．可能世界は，ワーカをノードとした木構造として表現することでき，この木構造の根から葉までパスが 1 つの依頼順序の場合を示す．例えば，依頼順序が $[w_1, w_2]$ の可能世界は図のようになる．w_i はワーカを，Yes/No はワーカが依頼を引き受けた場合，引き受けなかった場合を表している．

依頼順序の遂行確率とコストの期待値は，その可能世界を利用して求める．依頼順序 $o_k \in O$ の可能世界におけるパスの集合を $path(o_k)$ とする．パス $path_i \in path(o_k)$ に含まれるワーカの集合を W_{path_i}，パス $path_i$ 内のワーカ $w_i \in W_{o_k}$ の依頼の承諾状況を示す 2 値変数を x_i とし，コストを $c(w_i)$，ワーカが依頼を引き受ける可能性を $q(w_i)$ とする．

コストの期待値

コストの期待値とは，可能世界全体でかかるコストの期待値のことをいう．コストの期待値は，木構造として表現される可能世界のパス全てのコストの期待値の総和である．依頼順序 o_k の可能世界におけるパスの集合を $path(o_k)$ とするとパス $path_i \in path(o_k)$ における，コストの期待値を $cost(path_i|o_k)$ は

$$cost(path_i|o_n) = \left(\left(\sum_{w_k \in W_{path_i}} c(w_k) \cdot x_k\right) \cdot \left(\prod_{w_k \in W_{path_i}} (1 - q(w_k) - (1 - 2 \cdot q(w_k)) \cdot x_k\right)\right)$$

となり，依頼順序 o_n におけるコストの期待値 $cost(_n)$ は，

$$cost(o_n) = \sum_{path_i \in path(o_n)} cost(path_i|o_n)$$

で表される．

遂行確率

遂行確率とはタスク全てを実行できる確率のことをいう．遂行確率は，木構造として表現される可能世界のパスのうち，タスク全てを実行できたパスが起こる確率の和である．依頼順序 o_n の可能世界におけるタスク全てを実行したパスの集合を $path_{suc}(o_n)$ とするとパス $path_i \in path_{suc}(o_n)$ が起こる確率は，$p(path_i|o_n)$ は

$$p(path_i|o_n) = \prod_{w_k \in W_{path_i}} (1 - q(w_k) - (1 - 2 \cdot q(w_k)) \cdot x_k$$

となり，依頼順序 o_n における遂行確率 $q_{suc}(o_n)$ は，

$$q_{suc}(o_n) = \sum_{path_i \in path(o_n)} q(path_i|o_n)$$

で表される．

依頼順序が $[w_1, w_2, w_3]$ の場合のコストの期待値と遂行確率を計算した例を示す．例ではコストの期待値は 4900(=2800 + 525 + 735 + 315 + 525 + 0)，遂行確率は，0.5(=0.35 + 0.105 + 0.045) となる．(図: 6)．

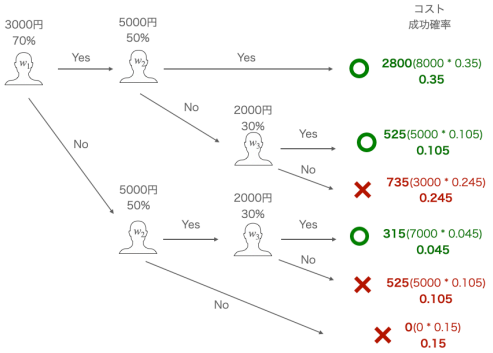


図 6　コストの期待値と遂行確率の例

最適な依頼順序 o_{opt} を求めるアルゴリズムを，次章で説明する．

4　提 案 手 法

本章では，第 3 章で説明した依頼戦略を求めるアルゴリズムについて説明する．「4.1 アルゴリズムの概要」では，アルゴリズム全体の手順について説明し，「4.2 アルゴリズムの詳細」では各手順におけるアルゴリズムの詳細について説明する．

4.1　概　　要

本研究では，ワーカへの依頼順序を決定する手法を提案する．提案手法では依頼順序の決定を 2 つのフェーズに分割する．フェーズ 1 では，ワーカからの提案を基に，与えられたタスク全てを実行できるようなタスク割り当てを求める．フェーズ 2 では，フェーズ 1 で求めたタスク割り当てを基に，依頼順序を求め（図 7），依頼順序の可能世界を考えて，コストと遂行確率を評価する（図 8）．図における a_j はタスク割り当てを，w_i はワーカを表している．

4.2　アルゴリズム

【フェーズ 1】タスク割り当てを求めるフェーズ 1 では，与えられたタスク集合 T の全てのタスクがワーカによって実行されるようなタスク割り当てを求める．タスク集合 $T = \{t_1, t_2, ..., t_i\}$ とワーカ集合 $W = \{w_1, w_2, ..., w_j\}$，ワーカからの提案 P からタスク割り当て a の集合 A を求める． P は i 行 j 列の行列であ

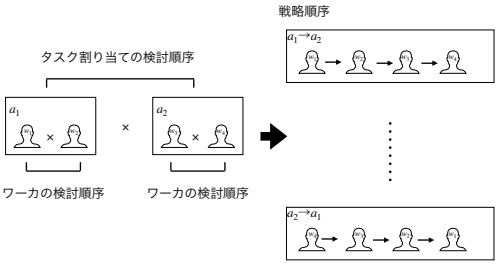


図 7　依頼順序の列挙

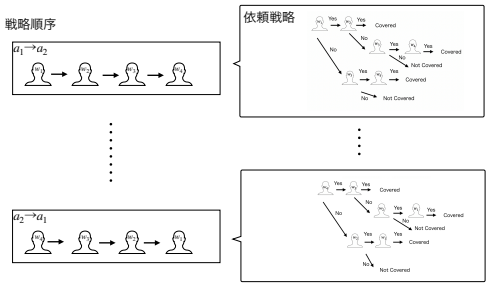


図 8　依頼順序の可能世界

り，i 行 j 列の要素 p_ij はワーカ w_i がタスク t_j を提案した場合に 1，提案しなかった場合に 0 となる．a は，i 行 j 列の行列であり，i 行 j 列の要素 a_ij はワーカ w_i へのタスク t_j の割り当て状況を示す．a_ij は，ワーカ w_i にタスク t_j を割り当てられた場合に 1，割り当てられなかった場合に 0 となる．このとき，タスク割り当てを求める問題は次のように定義される．

$$a = \sum_{i \in W} \quad p_{ij} \cdot a_{ij} \geqq 1 \quad (\forall j \in T)$$

これを満たす a の集合 A が，タスク割り当ての集合となる．

【フェーズ 2】依頼順序を求めるフェーズ 2 では，依頼順序を求める．依頼順序はフェーズ 1 で求めたタスク割り当てを試す順番と，そのタスク割り当て内のどのワーカから依頼するか決めることで求められる．タスク割り当てを試す順番をタスク割り当ての検討順序，タスク割り当て内のどのワーカから依頼するかを各タスク割り当てにおけるワーカの検討順序ということにする．

依頼順序は，タスク割り当ての検討順序と，各タスク割り当てにおけるワーカの検討順序からなる．依頼順序はタスク割り当て集合 A の順列とタスク割り当て a_i におけるワーカ $W(a_i)$ の順列によって求めることができる．このとき依頼順序はワーカを要素に持つ，タスク割り当てのリストで表現できる．

$$[W(A_1), W(A_2), ..., W(A_n)]$$
$$\vdots$$
$$[W(A_n), W(A_{n-1}), ..., W(A_1)]$$

【タスク割り当ての検討順序】タスク割り当ての検討順序とは，どのタスク割り当てから試すのかを表している．あるタスク割り当て内のワーカが依頼を引き受けなかった場合，そのタスク割り当て内のワーカでタスク全てを実行することはできない．そこでワーカが依頼を引き受けなかった時点で別のタスク割り当てを試すことを考える．全てのタスク割り当てを試すまでこれを繰り返したとき，タスク割り当てを試した順番のことをタスク割り当ての検討順序という．タスク割り当ての検討順序の例を示す．タスク割り当てが a_1, a_2, a_3 の 3 つの場合は，タスク割り当ての検討順序は，カバーパターン $[a_1, a_2, a_3]$ の順列として表される．

【タスク割り当て内のワーカの検討順序】タスク割り当て内ワーカの検討順序とは，あるタスク割り当てを試す際に，そのタスク割り当て内のどのワーカから依頼するのかを表している．例えば、あるタスク割り当て内のワーカが (w_1, w_2) の場合のワーカの検討順序は，ワーカ $[w_1, w_2]$ の順列として表される．

次に依頼順序から可能世界を求める．可能世界はワーカをノードとした木構造で表現できる．ワーカが依頼を引き受けた場合に，次に依頼するワーカを右の子ノード，依頼を引き受けなかった場合に，次に依頼するワーカを左の子をワーカとする．依頼順序を基に，タスクのカバーが成功するまで子ノードを作成していく．全ての依頼順序について可能世界を求める．

次に，依頼順序 o_n の可能世界のパス $path_i$ ごとに，コストの期待値 $cost(path_i|o_n)$，遂行確率 $q_{suc}(path_i|o_n)$ を計算する．

可能世界全てのパス $path_i \in path(o_n)$ について，$cost(path_i|o_n)$ と $q_{suc}(path_i|o_n)$ を計算し，その和 $cost(o_s)$，$q_{suc}(o_n)$ を求める．全ての依頼順序について、コストの期待値と遂行確率を計算し，$q_suc(o_n)$ が最も大きいもののうち，$cost(o_n)$ が最も小さい依頼順序 o_n を求める．

Algorithm 1 タスクの割り当て

Input: A set T of tasks, a set W of workers, a matrix P of proposal that is row as a task and column as a worker
Output: A set A of assignments

```
Let available_workers_in_tasks ← []
for all t ∈ T do
    Let available_worker ← []
    for all w ∈ W s.t. [p_ij] == 1 do
        add w to available_worker
    end for
    add available_worker to available_workers_in_tasks
end for
Let assignments ← []
assignments = ∏_{k=1}^{i} available_worker_list[k]
Let A ← []
for all assignment ∈ assignments do
    add convert_to_matrix(assignment) to A
end for
return A
```

Algorithm 2 依頼順序

Input: A set A of assignments, a function $Permutation(X)$ that returns the permutations of the set X, a function $W(a_i)$ that returns the workers in the assignment a_i

Output: A list o of sequences of the worker

1: *Let* $o \leftarrow []$
2: **for** *all* $pa_i \in Permutation(A)$ **do**
3: $sequence = Permutation(W(a_k))$ *each* a_k *in* pa_i
4: *add sequence to* o
5: **end for**
6: **return** o

$$O_{suc} = \left\{ o' \in O \quad | \quad q_{suc}(o') \geqq \max_{o'' \in O} \quad q_{suc}(o'') \right\}$$

$$o_{opt} = \{ \arg\min_{o' \in O_{suc}} cost(o') \}$$

5 実　　験

本研究で開発した依頼順序決定アルゴリズムで求めた依頼順序についてシミュレーションを行った．シミューレーションでは，依頼順序間にどれだけの差があるかについて調べ，依頼順序間に有意な差があることが分かった．各依頼順序において，遂行確率，コストの期待値について調べた．

シミュレーションによって，コストの期待値に最大 2408，遂行確率に最大 0.32 の差があることが分かった．

5.1 設　　定

プログラムで生成したデータを用いてシミュレーションした．生成したデータのパラメータについて説明する．ワーカの数を 10，タスクの数を 5 とし，ワーカからの提案（実行することができるタスク，コスト，依頼を引き受ける確率）を次のように設定した。実行することができるタスクは，タスクから n 個ランダムに抽出したタスクの集合とし，ワーカは最小で 0 個，最大で 5 個のタスクを実行できるものとする．コストは 1000 から 10000 の範囲のランダムな整数値，依頼を引き受ける確率は 0 から 1 の範囲のランダムな小数値とした．以上の設定で依頼順序のコストの期待値と遂行確率の分布を調べた．

5.2 結　　果

コストの期待値と遂行確率の分布から，コストの期待値に最大 2408，遂行確率に最大 0.32 の差があることが分かった．また，最良の依頼順序を選ぶと，コストの期待値は 1233，遂行確率は 0.33 であり，最悪の依頼順序を選ぶと，コストの期待値は 1547，遂行確率は 0.13 であった．この結果から，最良の依頼順序を選ぶことでコストの期待値を下げられること，タスク全てを実行する確率を上げられることがわかる．（図 9）．

コストの期待値と遂行確率について平均値，最小値，最大値，標準偏差を以下に示す．

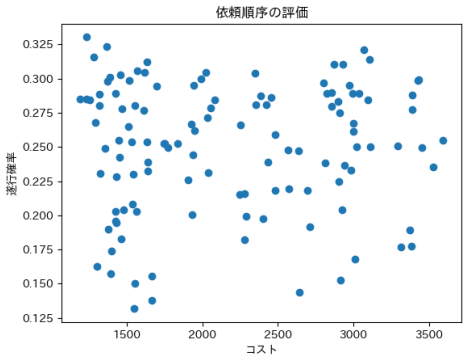


図 9　コストと遂行確率の分布

表 1　コストの期待値と遂行確率の統計値

評価値	平均値	最小値	最大値	標準偏差
コストの期待値	2182	1186	3594	715
遂行確率	0.24	0.13	0.33	0.047

5.3 考　　察

コストの期待値と遂行確率の分布からは，２つの値がバラついていることが分かる．コストの期待値と遂行確率はトレードオフではなく，コストの期待値が小さい場合を優先して依頼順序を求める場合，遂行確率が大きい場合を優先して依頼順序を求める場合によって，最適な依頼順序が異なるということが分かる．

6 結　　論

本論文は，物理的な空間で作業を必要とするフィジカルタスクのクラウドソーシングにおいて顕著にみられる「割当ての制約が強い状況における分担型タスク割当て」について，ワーカへの依頼順序に注目した依頼方法を提案し，最適な依頼順序を求めるアルゴリズムを開発した．シミュレーションによる実験の結果，依頼順序によって，依頼にかかる費用と遂行確率に大きな差が出ることが確認された．つまり，依頼順序を最適化することで依頼にかかる全体の費用を下げ，タスク全てを実行する可能性を上げられることが分かった．今後の課題としては，発見的手法や近似的解法を用いた効率の良いアルゴリズムの開発が挙げられる．

7 謝　　辞

本研究の一部は JSPS 科研費 (JP22H00508)，と JST CREST(Grant Number JPMJCR22M2) の支援を受けたものである．ここに謝意を示す．

文　　献

[1] John C. Tang, Manuel Cebrian, Nicklaus A. Giacobe, Hyun-Woo Kim, Taemie Kim, and Douglas" Beaker" Wickert. Reflecting on the darpa red balloon challenge. *Communications of the ACM*, Vol. 54, No. 4, pp. 78–85, 2011.

[2] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. Location-based crowdsourcing: extending crowd-

sourcing to the real world. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pp. 13–22, 2010.

[3] Yongjian Zhao and Qi Han. Spatial crowdsourcing: current state and future directions. *IEEE Communications Magazine*, Vol. 54, No. 7, pp. 102–107, 2016.

[4] P. Cheng, X. Lian, L. Chen, J. Han, and J. Zhao. Task assignment on multi-skill oriented spatial crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 8, pp. 2201–2215, 2016. ID: 1.

[5] Dingxiong Deng, Cyrus Shahabi, and Linhong Zhu. Task matching and scheduling for multiple workers in spatial crowdsourcing. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–10, 2015.

[6] Srinivasa Raghavendra Bhuvan Gummidi, Torben Bach Pedersen, and Xike Xie. Transit-based task assignment in spatial crowdsourcing. In *32nd International Conference on Scientific and Statistical Database Management*, pp. 1–12, 2020.

[7] Leyla Kazemi and Cyrus Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th international conference on advances in geographic information systems*, pp. 189–198, 2012.

[8] Yongsung Kim, Darren Gergle, and Haoqi Zhang. Hit-or-wait: Coordinating opportunistic low-effort contributions to achieve global outcomes in on-the-go crowdsourcing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018.

[9] Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, and Linhong Zhu. Task selection in spatial crowdsourcing from worker' s perspective. *GeoInformatica*, Vol. 20, pp. 529–568, 2016.

[10] Yongjian Zhao and Qi Han. Offline worker selection for real-time spatial crowdsourcing multi-worker tasks. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pp. 545–550, 2019.

[11] Ning Wang and Jie Wu. Cost-efficient heterogeneous worker recruitment under coverage requirement in spatial crowdsourcing. *IEEE Transactions on Big Data*, Vol. 7, No. 2, pp. 407–420, 2018.

[12] Wen Sun, Qiang Li, and Chen-Khong Tham. Wireless deployed and participatory sensing system for environmental monitoring. In *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 158–160, 2014.

[13] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. Cartel: a distributed mobile sensor computing system. In *Proceedings of the 4th international conference on Embedded networked sensor systems*, pp. 125–138, 2006.

[14] Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pp. 323–336, 2008.

[15] Fatemeh Saremi, Omid Fatemieh, Hossein Ahmadi, Hongyan Wang, Tarek Abdelzaher, Raghu Ganti, Hengchang Liu, Shaohan Hu, Shen Li, and Lu Su. Experiences with greengps—fuel-efficient navigation using participatory sensing. *IEEE Transactions on Mobile Computing*, Vol. 15, No. 3, pp. 672–689, 2016.

[16] Zhibo Wang, Yuting Huang, Xinkai Wang, Ju Ren, Qian Wang, and Libing Wu. Socialrecruiter: Dynamic incentive mechanism for mobile crowdsourcing worker recruitment with social networks. *IEEE Transactions on Mobile Computing*, Vol. 20, No. 5, pp. 2055–2066, 2021.

[17] Jia Xu, Chengcheng Guan, Haobo Wu, Dejun Yang, Lijie Xu, and Tao Li. Online incentive mechanism for mobile crowdsourcing based on two-tiered social crowdsourcing architecture. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9, 2018.

クラウドワーカの作業結果と投票結果から生成される確率分布を利用した結果集約手法

城所　祥太†　　鈴木　　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †{kidokoro.shota.e3@s.gifu-u.ac.jp，suzuki.yu.r4@f.gifu-u.ac.jp}

あらまし　本研究では，作業精度の高いワーカの作業情報を用いてタスク型クラウドソーシングの結果を集約することにより，高品質な作業結果を出力することを目的とする．ワーカの作業精度を推定する際に，総作業数が少ないワーカの正確な作業精度を推定できない問題がある．そこで我々は，ワーカの作業精度を推定する際の不確かさを考慮したワーカの選出に着目した．本手法では，推定の不確かさを考慮する方法として，確率分布に従うサンプリングの結果を用いる．確率分布に従うサンプリング値を元にワーカの選出を行うことによって，総作業数が多く，作業精度の高いワーカを優先的に選出することができると考えた．我々は，手法の有効性を検証するために実験を行った．実験では，異なる三種類の手法による集約結果を比較し，推定の不確かさを考慮したワーカの選出が高品質な作業結果の出力に有効であるか調査した．

キーワード　クラウドソーシング，確率的データ処理

1　はじめに

本研究は，クラウドソーシングにおいて作業精度の高いワーカの作業情報を用いた結果集約によって，高品質な作業結果を出力することを目的とする．クラウドソーシング [1] とは，インターネット上で不特定多数の作業者に作業を依頼する仕組みである．タスク型クラウドソーシングでは，複数のワーカに同じタスクを依頼し，それらの作業情報を用いた多数決によって最終的な作業結果を決定する方法がある．複数の作業情報を用いた多数決結果は，多くのワーカの意見を反映させることができる．そのため，一つのタスクに対して一人のワーカに作業を依頼した場合に比べ，複数の作業情報を用いた多数決結果は信頼度の高い作業結果となる．しかし，不適切な作業を行うワーカを含む場合，多数決結果が作業依頼者の求める作業結果と異なることもある．そこで我々は，作業精度の高いワーカの作業情報を集約することによって，作業依頼者の求める作業結果を得る集約手法を提案する．

本手法では，ワーカの作業情報から確率分布を生成し，その確率分布から生成する乱数を用いた結果集約を行う．この集約方法を用いることによって，推定されるワーカの作業精度の不確かさを考慮することができると考えた．個人の作業情報と作業依頼者による正解の作業結果の一致率が高いワーカは，作業精度の高いワーカであると考えられる．作業精度の高いワーカの作業情報を用いた多数決により，高品質な作業結果を得ることができる．しかし，ワーカの作業精度を推定するためには二つの課題がある．一つ目は，ワーカの作業精度を推定するために必要となる，作業依頼者による正解の作業結果を大量に用意することが難しい点である．二つ目は，総作業数が少ないワーカの正確な作業精度を推定することができない点である．

ワーカの作業精度を推定するために，まず，個人の作業情報と作業依頼者による正解の作業結果を比較する必要がある．十分な精度で作業精度を推定するためには，作業依頼者が正解の作業結果を大量に用意する必要がある．しかし，大量の正解の作業結果を用意することは作業依頼者に大きな負担がかかる．そこで我々は，作業依頼者による正解の作業結果の代わりに，複数の作業情報を用いた多数決結果を利用することができるのではないかと考えた．

我々は，推定される作業精度と本来の作業精度の誤差を考慮したワーカの選出を目指す．以後，推定される作業精度と本来の作業精度の誤差を推定誤差と呼ぶ．ワーカの総作業数が少ないほど，推定誤差は大きくなるため，推定される作業精度は信頼度が低いといえる．Accuracy や F 値を用いた作業精度推定では，総作業数の違いによる推定誤差を考慮していないため，推定される作業精度の信頼度を評価することができない．そこで，我々は推定誤差を考慮したワーカの作業精度推定を行うために，ワーカの作業情報を元に生成する確率分布に着目した．確率分布とは，ある試行で起こり得るすべての事象（確率変数）の確率を表現したものである．確率分布には，離散型確率分布と連続型確率分布の二種類がある．離散型確率分布とは，確率変数が連続ではなく独立している場合の確率分布である．対して，連続型確率分布とは，確率変数が連続値である確率分布である．

本手法では，推定されるワーカの作業精度を確率変数とする連続型確率分布を用いる．そして，確率分布に従う 0 から 1 の乱数を生成し，生成した乱数をワーカの作業精度と仮定することにより，ワーカ同士の作業精度を比較する．ワーカの作業数が多いほど，生成する確率分布は本来の作業精度に収束する．そのため，この確率分布から生成される乱数は本来の作業精度に近い値となる．一方，ワーカの作業数が少ないほど，生成する確率分布のばらつきは大きくなる．そのため，この確率分布から生成する乱数は本来の作業精度と離れた値を出力しやすく

なる．我々は，確率分布から生成する乱数を用いたワーカの選出によって，総作業数が多く，作業精度の高いワーカを優先的に選出することができると考えた．本手法では，各ワーカに対して確率分布を生成し，その確率分布から生成する乱数の値が高いワーカの作業情報を用いた多数決によって最終的な作業結果を決定する．

手法の有効性を検証するために，三種類の集約手法の比較実験を行った．ベースラインとして単純な多数決による方法と，作業情報から求められる F 値が高いワーカの作業結果を用いた多数決により作業結果を決定する方法を用いた．実験の結果，本手法を用いることによる作業精度の高いワーカの作業情報を用いた集約が，ベースラインと比較して高品質な作業結果の出力に有効であることが示された．また，確率分布から生成する乱数を用いることによって，総作業数が多く，作業精度の高いワーカを優先的に選出することができることが分かった．

本論文における貢献は以下のとおりである．

- 確率分布に従うサンプリング値を用いた集約が高品質な作業結果の出力に有効である．
- 確率分布に従うサンプリング値を用いたワーカの選出によって，作業精度が高いワーカのうち，操作行数が多く，推定される作業精度の信頼度が高いワーカを優先的に選出できた．

本論文の構成は以下の通りである．2 章では本論文にて用いた技術や手法の基本的事項について述べる．3 章では関連研究について述べる．4 章では本論文の提案手法について述べる．5 章では提案手法を用いた評価実験の目的や手順，結果・考察などについて述べる．最後に 6 章では本論文のまとめと今後の課題について述べる．

2 関 連 研 究

タスク型クラウドソーシングの作業結果の質の向上には，様々なアプローチが存在する．例えば，Zhang ら [2] や Ho [3] らの研究のように, ワーカにタスクを割り当てる際に工夫を施す方法や，Chandler [4] らや Gillier [5] らのようにインストラクションを改善することによってワーカの作業精度を向上させる方法などがある．様々なアプローチのうち，我々は集まった複数の作業情報の集約の際に，作業精度の高いワーカの作業情報を元に結果集約することによって，高品質な作業結果の出力を目指す．

複数の作業情報の集約において，ワーカの作業精度を考慮した結果集約手法の研究はいくつか行われている．Dawid と Skene ら [6] はワーカの作業精度にばらつきがあることと作業情報の信頼度を考慮した集約手法を紹介している．彼らの手法は，EM アルゴリズムを用いることによって，ワーカの作業精度と作業情報の信頼度の両方を推定する．この方法で推定した各ワーカの作業精度を元に作業結果を決定することで，高品質な作業結果が得られることを示した．Whitehill ら [7] は，タスクの難易度を考慮した集約手法を紹介している．タスクの難易度とワーカの作業精度の関係性に着目し，作業情報を決定することによって，最終的な集約結果の品質が向上することを示した．Welinder ら [8] は，ワーカとタスクの相性を考慮した手法を紹介している．ワーカの作業の判断基準の傾向とタスクの潜在特徴との関係性を捉えることが集約結果の品質向上につながることを示した．本手法は，作業精度の高いワーカを選出し，その作業情報を用いた多数決によって作業結果を決定する．作業精度の高いワーカを選出するうえで，ワーカの作業精度の推定誤差を考慮したワーカの選出を行うという点で上記の研究とは異なる．

ワーカの作業精度を推定する上で，総作業数が少ないワーカの正確な作業精度を推定できない問題を考慮した集約に関する研究も存在する．Venanzi ら [9] は，各ワーカはいくつかのカテゴリーに分類できると仮定し，ワーカをカテゴリー分類した結果を用いる集約手法について紹介している．ワーカの作業精度を考慮した集約は有効である傾向が強いが，総作業数が少ないワーカの推定される作業精度は信頼度の高いものであるとは言えない，そこで，各ワーカはいくつかのカテゴリーに分類できると仮定し，混同行列によって各ワーカをそのカテゴリーごとに分類した．総作業数が少ないワーカであっても他のワーカと同等レベルで作業精度を推測でき，集約結果の品質向上に有効であることを示した．小山ら [10] らは，確信度を用いた集約手法を紹介している．ワーカが報告した各作業に対する確信度をパラメータとして用いることで，ワーカの作業精度の推定精度が向上することを示した．本手法は，各ワーカに対して，作業精度を確率変数とする確率分布を生成し，その確率分布に従うサンプリング値を用いた集約を行う．確率分布を用いることで，作業精度の推定誤差を考慮したワーカの選出が可能になると考える．本手法は，作業精度の高いワーカの選出に，確率分布に従うサンプリング値を用いるという点で上記の研究とは異なる．本研究は，複数の作業情報を用いた多数決結果と，ワーカの作業情報から確率分布を生成し，その確率分布に従うサンプリング値を用いた集約の有効性を検証する．

3 提 案 手 法

我々は，確率分布から生成する乱数を用いた結果集約手法を提案する．個人の作業情報と作業依頼者による正解の作業結果との一致率が高いワーカは作業精度の高いワーカである．本研究の目的は，作業精度の高いワーカの作業情報を集約することによって，高品質な作業結果を出力することである．しかし，作業精度の高いワーカを選出する上で，総作業数の少ないワーカの作業精度の推定が不正確である問題がある．そこで我々は，確率分布から生成する乱数を用いたワーカの選出に着目した．本手法の手順を以下に示す．

Step1　各ワーカの正解作業数（正解の作業結果と同じ作業情報である作業数）と不正解作業数（正解の作業結果と異なる作業情報である作業数）を求める．

Step2　Step1 で求めた結果を元に，各ワーカの確率分布を生成する．

Step3　あるタスク n を作業した複数のワーカに対して，各

ワーカの確率分布から乱数を生成をする．その乱数の値が高い数人の作業情報を用いた多数決によってタスク n の作業結果を決定する．この操作を各タスクに対して行い，全てのタスクの作業結果を決定する．

これらの手順の詳細について各節にて説明する．

3.1 作業情報の集計

本節では，複数の作業情報を用いた多数決結果とワーカの作業情報の比較することによってワーカの作業情報情報を集計する方法について説明する．作業情報の集計で求める出力は，正解作業数（正解の作業結果と同じ作業情報である作業数）と不正解作業数（正解の作業結果と異なる作業情報である作業数）である．作業情報の集計は，ワーカの作業情報と作業依頼者による正解の作業結果と比較する．しかし，十分な精度で作業精度を推定するためには，作業依頼者による正解の作業結果を大量に用意する必要がある．しかし，大量の正解の作業結果を用意することは作業依頼者に大きな負担となる．そこで我々は，作業依頼者の求める作業結果の代わりに，複数の作業情報を用いた単純多数決結果を使用できるのではないかと考えた．タスク型クラウドソーシングは，同一タスクを複数のワーカに割り当て，それらの作業情報を用いた多数決によって最終的な作業結果を決定する．複数の作業情報を用いた集約は，多くのワーカの意見を反映させた作業結果を出力する．そのため，一人で作業した結果と比べて，複数の作業情報を用いた集約結果のほうが信頼度の高い作業結果となる．同一タスクに割り当てるワーカの数を増やすほど，それらの作業情報を用いた集約結果は作業依頼者の求める作業結果と一致しやすくなると考えられている．以後，複数の作業情報の単純多数決結果を投票結果とする．作業依頼者の求める作業結果の代わりに投票結果を用いることは，作業依頼者の負担を軽減することにつながる．我々は，作業依頼者の求める作業結果の代わりに投票結果を使用することの妥当性を確認するため，4.3.1 節にて紹介するデータセットを用いて，作業依頼者の求める作業結果と投票結果の一致率を求めた．一致率の算出手順を以下に示す．

Step1　各タスクに対して，集まった複数の作業情報を用いた多数決によって投票結果を決定する．

Step2　全てのタスクのうち，無作為に抽出した 500 件のタスクに対して作業依頼者が正解の作業結果を付与する．

Step3　投票結果と Step2 で付与した正解の作業結果を比較する．

Step4　正解の作業結果と投票結果が同じタスクの件数を正解の作業結果を付与した件数である 500 件で割ることにより一致率を求める．

上記の手順により 4.3.1 節のデータセットにおける作業依頼者の求める作業結果と投票結果の一致率（Accuracy）は 0.894 であった．作業精度を推定するうえで，作業依頼者の求める作業結果と投票結果の一致率は十分に高い．結果から我々は，作業依頼者の求める作業結果の代わりに投票結果を使用できると考えた．以上の理由より本手法では，個人の作業情報と投票結果を比較することによって各ワーカの作業情報を集計する．

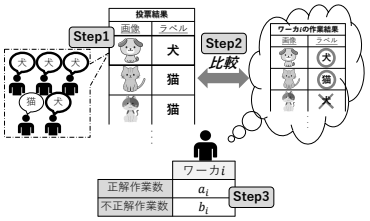


図 1　作業情報の集計の概要

総ワーカ数を N 人とするとき，ワーカ $i(i = 1, 2, \cdots, N)$ に対して作業情報の集計で求める出力は，正解作業数 a_i（正解の作業結果と同じ作業情報である作業数）と不正解作業数 b_i(正解の作業結果と異なる作業情報である作業数) である．作業情報の集計の概要を図 1 に示す．

Step1　各タスクに対して，集まった複数の作業情報を用いた多数決によって投票結果を決定する．

Step2　ワーカ i が作業した全てののタスクに対して，ワーカ i の作業情報と投票結果を比較する．

Step3　投票結果と同じ作業情報であったワーカ i の作業数を正解作業数 a_i とする． 投票結果と異なる作業情報であったワーカ i の作業数を不正解作業数 b_i とする．

Step2 と Step3 の操作をすべのワーカに対して行う．

3.2 結果集約手法

本節では，確率分布から生成する乱数を用いた結果集約手法について説明する．3.2.1 節では，3.1 節にて集計した各ワーカの正解作業数と不正解作業数を元に生成する確率分布について述べる．3.2.2 節では，3.2.1 節にて紹介する確率分布に従うサンプリングとサンプリング値を用いた集約手法について述べる．

3.2.1 確率分布の生成

本手法は作業精度の高いワーカを選出するために，作業精度を確率変数とするベータ分布を用いる．

作業精度の高いワーカを選出には，各ワーカの作業精度を推定する必要がある．ここで我々は，総作業数が少ないワーカの作業精度を推定する場合に注意が必要である．なぜならば，総作業数が少ないワーカの作業情報から推定される作業精度は本来の作業精度とは大きく異なる場合があるからである．本稿では, 推定した作業精度と本来の作業精度の間に生じる誤差を推定誤差と呼ぶ．作業精度の推定誤差が大きい場合，推定される作業精度は信頼度が低いといえる．Accuracy や F 値を用いた作業精度の推定では，ワーカの総作業数を考慮していないため，推定される作業精度の信頼度を評価することができない．そこで我々は，確率分布による作業精度の推定に着目した．

確率分布とは，ある試行で起こり得る全ての事象（確率変数）の発生確率を分布を用いて表したものである．確率分布には，

離散型確率分布と連続型確率分布がある．離散型確率分布とは，確率変数が $0, 1, 2$ のような離散値である場合の確率分布である．連続型確率分布とは，確率変数が連続値である場合の確率分布である．連続型確率分布において，確率変数 X がある値 x を取るときの確率は 0 である．これは確率変数 X が連続値であるため，発生し得る事象が無限に存在するからである．連続型確率分布における確率変数 X がある値 x を取るときの確率密度 $f(x)$ は確率密度関数と呼ばれる．確率密度関数は，出力がある事象の発生確率を表しているわけではないことに注意が必要である．連続確率分布は，確率変数 X がある範囲内にどれくらいの確率で存在し得るかを求める際に用いられる．確率密度関数 $f(x)$ において，$a \leqq X \leqq b$ となる確率 P は以下の計算によって求められる．

$$P(a \leqq X \leqq b) = \int_a^b f(x)dx \tag{1}$$

本手法では，ワーカ i の作業精度を確率変数とする連続型確率分布を用いることによって，ワーカ i の作業精度を推定する．ワーカの作業数が多いほど，生成する確率分布は本来の作業精度に収束する．一方，ワーカの作業数が少ないほど，生成する確率分布は分散が大きく，推定される作業精度のばらつきが大きいことを表現できる．上記の理由より我々は，確率分布を用いることによって，ワーカごとの作業精度推定の不確定さを考慮した集約が可能なのではないかと考えた．

確率分布には様々な種類が存在するが，本手法ではベータ分布を使用する．ベータ分布とは，連続型確率分布の一つであり，確率変数の区間が 0 から 1 の間で定義される．ベータ分布は，成功回数 a 回と失敗回数 b 回が分かっている試行に関して、成功確率 p の分布を表すことができる。ベータ分布を使用する理由は，3.1 節でワーカの作業情報を正解の作業を行ったか否かの 2 値で分類しており，いかなるタスクであってもワーカの作業情報をベルヌーイ過程で表現できるからである．ベルヌーイ過程とは，2 種類の独立した結果のみを出力する試行のことである．本手法は，ワーカ i の推定される作業精度を確率変数とするベータ分布を用いることによって，ワーカ i の作業精度を推定する．ベータ分布の従う確率密度関数はベータ関数を用いることで求められる．ワーカ i の作業精度を p_i，正解作業数を a_i, 不正解作業数を b_i とするときのワーカ i のベータ関数 $\beta(a_i, b_i)$ を式 (2) に，ベータ分布の従う確率密度関数 $f(p)$ を式 (3) 示す．

$$\beta(a_i, b_i) = \int_0^1 p_i^{a_i-1}(1-p_i)^{b_i-1}dp_i \tag{2}$$

$$f(p_i) = \frac{p_i^{a_i-1}(1-p_i)^{b_i-1}}{\beta(a_i, b_i)} \tag{3}$$

作業精度 p が 0.6 のワーカに関して，総作業数が 10 回（正解作業数：6 回，不正解作業数：4 回）の場合と，100 回（正解作業数：60 回，不正解作業数：40 回）の場合のベータ分布を図 2 に示す．図 2 から，総作業数が少ないワーカの推定される作業精度はばらつきが大きく，総作業数が多いワーカの推定される作業精度はばらつきが小さいことが分かる．

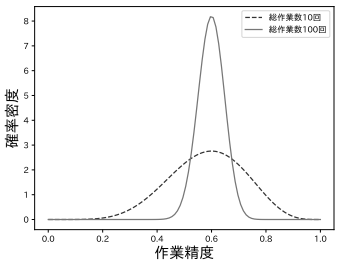


図 2　作業精度 $p = 0.6$ のワーカのベータ分布

3.2.2　サンプリング値を用いた集約

本手法では，3.2.1 節で紹介した確率分布に従うサンプリングを行う．そのサンプリング値をワーカの作業精度と仮定することによって．ワーカ同士の作業精度を比較し，作業精度の高いが考えられるワーカを決定する．サンプリングとは，確率分布から標本を抽出することをいう．つまり，サンプリングを行うことによって確率分布に従う乱数を生成することができる．サンプリングで得られる値は，確率分布に従って抽出される確率変数 x である．3.2.1 節で紹介したベータ分布は，各ワーカの推定される作業精度とその推定誤差を表現することができる．ワーカの作業数が多いほど，生成する確率分布は本来の作業精度に収束するため，サンプリング値は本来の作業精度付近の値になる．例えば，図 2 の実線のような総作業数が多いワーカから生成する確率分布上でサンプリングを行うと，0.60 や 0.65 など本来の作業精度に近い値を出力しやすい．そのため，総作業数が多く，作業精度の高いワーカはサンプリング値が高い値を出力しやすい．一方，ワーカの作業数が少ないほど，生成する確率分布のばらつきが大きくなるため，サンプリング値は本来の作業精度から離れた値になりやすい．図 2 の点線のような総作業数が少ないワーカから生成する確率分布上でサンプリングを行うと，0.40 や 0.80 など本来の作業精度から離れた値を出力する確率が高くなる．そのため，総作業数が少なく，推定される作業精度の信頼度の低いワーカのサンプリング値は不安定であり，作業精度の高いワーカとして選出されにくくなる．以上の理由より，確率分布に従うサンプリング値を元にワーカの選出を行うことによって，総作業数が多く，作業精度の高いワーカを優先的に選出でき，高品質な作業結果の出力に繋がると考えた．サンプリング値を用いた集約手法の手順を以下に示す．

Step1　3.2.1 節の式 (2) から各ワーカの確率分布を得る．

Step2　あるタスク n を作業した複数のワーカに対して，各ワーカの確率分布に従うサンプリングを行う．

Step3　サンプリング値の高い上位数人のワーカの作業情報のみで多数決を行い，タスク n の最終的な作業結果を決定する．

Step2 と Step3 の操作をすべのタスクに対して行う．Step3 において，多数決を行う際に選ぶ上位人数は，データセットによって適当な人数を用意する必要がある．多数決を行う上位人数を 3 人とした時の集約例を図 3 に示す．

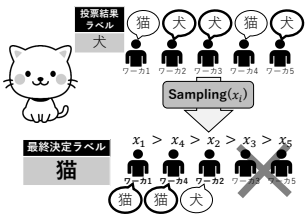


図 3　サンプリング値を用いた集約の例

確率分布に従うサンプリング値の利用は，推定される作業精度の信頼度の高いワーカの選出に有効であると考える．総作業数が多く，作業精度の高いワーカの確率分布に従うサンプリングは，高い値を出力する確率が非常に高い．そのため，Step3 にて選ばれる可能性は高くなる．対して，総作業数が少なく，推定される作業精度の信頼度の低いワーカはサンプリング値にばらつきが大きく，Step3 にて選ばれる可能性が下がる．これらの理由より，サンプリング値を用いた集約は推定される作業精度の信頼度の高いワーカを優先的に選出することができる．

4　評 価 実 験

本実験の目的は二つある．一つ目は，本手法を用いた集約が高品質な作業結果の出力に有効であるかどうかを検証することである．二つ目は，サンプリング値を用いることによって，作業精度が高いと推定されるワーカのうち，総作業数が多く，作業精度が高いワーカを優先的に選出することができるかどうかを検証することである．本実験では，ベースラインとして単純な多数決による集約手法，F 値により作業情報を評価する集約手法と本手法であるサンプリング値を用いた集約手法の三種類の集約手法を比較する．

4.1　データセット

本実験では，岐阜大学鈴木研究室とソニー株式会社の共同研究にて構築したデータセットを用いる．このデータセットは，クラウドソーシングを用いて構築した．このデータセットを構築した目的は，テキストデータに含まれるある製品に対する要望，体験，感情について分析することである．このテキストデータは，ある製品に関する動画のコメントである．データセットの構築に伴い，一つのテキストデータあたり五つの質問を設定した．質問の項目を以下に示す．

Q1　テキストデータがある製品に対する要望，体験，感情が含まれるか．

回答の選択肢は「含む」「含まない」の二種類である．

Q2　テキストデータがある製品の何について書かれているか．

回答の選択肢は「本体」「品質・不具合」「アクセサリ・周辺機器」「その他」「該当なし」の五種類である．

Q3　テキストデータを投稿した著者はどのようなある製品に対してどのような感情を持っているか．

回答の選択肢は「ポジティブ」「ネガティブ」「ニュートラル」「ポジティブ＆ネガティブ」の四種類である．

Q4　テキストデータを投稿した著者はある製品を使用したことがあると考えられるか．

回答の選択肢は「はい」「いいえ・不明」の二種類である．

Q5　テキストデータ内で，ある製品は何の製品と比較されているか．

回答の選択肢は「同社の製品」「他社の製品」「比較なし」の三種類である．

ただし，Q1 で「含まない」と回答した場合，後の四つの質問を行わないこととした．

本実験で用いるデータセットは 606 人のワーカの作業情報をまとめたものである．各作業情報は,「ワーカ ID」「テキストデータ」「質問番号」「ワーカが付与したラベル」の四種類の項目で構成されている．また，データセットの構築において対象としたテキストデータ数は 26,200 件である．データセットの構築にあたり，多くのワーカの意見を反映したラベル決定を行うため，一つのテキストデータに対して 10 人のワーカに作業を依頼した．この 10 人のワーカの作業情報を用いた多数決によって決定するラベルを投票結果とする．

集約精度の評価を行うにあたって，集約結果と作業依頼者による正解の作業結果を比較する．そのため，各テキストデータに対して作業依頼者が正解ラベルを付与したデータセットを作成必要がある．作業依頼者は，投票結果が付与されたテキストデータのうち，無作為に抽出したテキストデータに対して正解ラベルを付与した．無作為に抽出したテキストデータのうち正解ラベルを付与した 500 件のテキストデータを正解データセットとする．全てのテキストデータに対して正解ラベルを用意するべきだが，コストがかかるため本実験では 500 件とした．正解データセットを作成する際，作業依頼者の負担を軽減するため，正解ラベルの決定に迷うテキストデータは取り除いた．本実験では，作業依頼者は著者であるとして，正解データセットを構築した．正解データセットに含まれる 500 件のテキストデータに対して 4.2 節にて紹介する三種類の集約手法を適用し，集約精度を算出する．

4.2　実 験 設 定

本実験では，異なる三種類の手法の集約精度を比較することによって，本手法の集約精度を評価する．手法の集約精度とは，手法を用いて決定したラベルと正解ラベルの一致率である．本実験では，一致率の指標として F 値を用いる．三種類の手法について以下で説明する．

手法 1　各テキストデータに対して，集まった複数の作業情報を用いた多数決によって作業結果を決定する．この時，集まった全作業情報を用いた多数決によって決定した作業結果を投票結果とする．

手法 2　各ワーカに対して，個人の作業情報と投票結果を用いることにより F 値を求め，その値を作業精度とする．

各テキストデータに対して，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

手法 3　本研究の提案手法であるサンプリング値を用いた集約によって作業結果を決定する．

手法 1 の集約精度を各実験のベースラインとする．ベースラインと提案手法の比較により，提案手法が高品質な作業結果の集約に有効であるかどうか検証する．

手法 2 は，作業精度が高いワーカを選出するうえで，作業精度の推定の不確定さを考慮していない集約方法である．3.2.1 節にて，F 値による作業精度の推定では推定の不確定さを考慮できていない問題があると述べた．対して提案手法は，確率分布に従うサンプリング値の利用により，推定の不確定さを考慮した集約を行うことができる．F 値により作業精度を評価する集約手法と提案手法の比較により，提案手法が推定の不確定さを考慮した集約を行うことができているかどうかを検証する．加えて，推定の不確定さを考慮した集約が高品質な作業結果の出力に有効であるかどうか検証する．手法 2 の集約手順を以下に示す．

Step1　各ワーカに対して，個人の作業情報と投票結果用いることにより F 値を求め，その値を作業精度をする．

Step2　あるテキストデータ n に対して作業したワーカの中で F 値の高い上位数人を選出する．

Step3　選出した数人の作業情報のみで多数決を行い，テキストデータ n の最終的なラベルを決定する．

Step1 から Step3 を全てのタスクに対して行う．

手法の集約精度を求める手順を以下に示す．

Step1　データセットの全てのテキストデータに対し，集約手法を用いて作業結果を決定する．

Step2　Step1 の結果と作業依頼者による正解の作業結果を用いることにより集約精度（F 値）を求める．

上記の手順で，手法 1，手法 2，手法 3 の集約精度を求める．ただし，手法 1 と手法 3 を用いる場合は，上記の手順を 10 回行い，それらの平均を取ることによって集約精度を決定する．手法 2 に関しては，同じ操作をした場合，集約精度に変化がないため，上記のような複数回の操作から平均を求める作業を行わない．

手法を適用をする際，何人のワーカの作業情報を用いると適切なのかを調査する．そのため，多数決に使用するワーカの数を変えて集約を行い，集約精度を比較する．この時，各手法において最も高かった集約精度をその手法の集約精度とする．求めた三種類の手法の集約精度を比較することによって，本手法の有効性を検証する．

4.3　実験 1,「要望，体験，感情が含まれるか」

4.3.1　実験条件

実験 1 では，4.1 節にて用意したデータセットのうち，Q1 の質問に関して作業を行った作業情報のみを抽出し構築したデータセットを用いる．作業によって付与されるラベルは体験・感情・要望を「含む」「含まない」の 2 種類である．上記のデータセットは，一つのテキストデータに対し，10 人のワーカが同じ作業をしている．そこで各テキストデータに付与するラベルを，作業した 10 人のワーカの作業情報を用いた多数決によって決定する．10 人のワーカの作業情報を用いた多数決によって決定したラベルを投票結果とする．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まる場合は，そのラベルを最終的に付与するラベルして決定する．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まらない場合は，最終的に付与するラベルを「含む」とした．この操作によって構築されたデータセットのラベル別のテキストデータ総数は,「含む」が 6,829 件,「含まない」が 19,371 件であった．投票結果が付与されたデータセットと各ワーカの作業情報を比較し，3.1 節にて説明した正解作業数と不正解作業数を求める．

このデータセットを用いた実験は，選出するワーカ人数を 1 人，3 人，5 人，7 人，9 人の 5 パターンで行った．これは，2 値分類において作業情報の多数決の結果が一意に決まる人数である．ベースラインである単純多数決は，無作為に抽出した数人のワーカの作業結果を用いた多数決によって作業結果を決定する．F 値によりワーカの作業精度を評価する集約手法では，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．本手法である，確率分布に従うサンプリング値を用いた集約手法では，サンプリング値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

4.3.2　結果・考察

4.2 節にて説明した三種類の手法を用いた集約結果と正解データセットの一致率を F 値で算出し，表 4.3.2 に示す．ベースラインである単純多数決による集約は，選出人数が増えるほど精度が高くなっている．多数決結果が 5:5 に割れた場合,「含む」というラベルを付けたが，実際は正解ラベルが「含まない」であるデータが多かったため，10 人で集約した時に少し精度が下がったと考えられる．以後，ベースラインである単純多数決の集約精度は最も高かった 0.853 を採用し，他の手法の集約精度と比較を行う．

本手法である，サンプリング値を用いた集約手法の集約精度は，選出人数が 5 人の時に最も高く，0.892 であった．ベースラインの 0.853 と比べて本手法の集約精度が高いことから，サンプリング値を用いた集約は高品質な作業結果の出力に有効であることが分かる．これは，サンプリング値が高いワーカが実際に作業精度が高いワーカであり，かつサンプリング値が低い

表 1　実験 1 における各手法を用いた際の F 値

	選出人数（人）					
	1	3	5	7	9	10
ベースライン	0.737	0.796	0.827	0.842	**0.853**	0.843
F 値上位	0.839	0.871	**0.874**	0.850	0.855	0.843
サンプリング	0.857	0.890	**0.892**	0.873	0.870	0.843

ワーカが実際に作業精度が低いワーカであったためだと考えられる．また，本手法を用いた際，選出人数が 9 人の場合の集約精度と 10 人の場合の集約精度では大きな差がある．これは，サンプリング値が低いワーカが高確率で作業精度の低いワーカであり，本手法を用いることによってそのようなノイズとなる作業情報を取り除くことができるといえる．

F 値により作業情報を評価する集約手法では，選出人数が 5 人の時に最も集約精度が高く，0.874 であった．本手法を用いた場合の作業精度は 0.892 であり，F 値により作業情報を評価する集約手法に比べて優れているといえる．また，本手法と F 値により作業情報を評価する集約手法における作業精度が高いと推測した上位 5 人のワーカに対して作業数の平均を求めた．本手法におけるサンプリング値が高い上位 5 人のワーカの平均作業数は 2,900 件であった．また，F 値により作業情報を評価する集約手法における F 値が高い上位 5 人のワーカの平均作業数は 2,747.7 件であり，本手法を用いた場合のほうが高い値であった．以上の結果から，サンプリング値が高いワーカを選出することによって，総作業数が多く，作業精度が高いワーカを優先的に選出することができたことが分かった．

また，本手法を用いる際，選出する上位人数が 5 人の時に集約精度が最も高いことから，およそ半分の作業情報は多数決の集約においてノイズであることが分かる．このようなノイズとなる作業を行うワーカを事前に取り除くことで，コスト削減とさらなる集約精度向上に繋がると考える．

各手法において選出人数が 1 人の場合の集約精度はあまり高くない．これは，高精度で最も集約精度が高いワーカを選出できたとしても，そのワーカ 1 人の作業結果のみでは，高品質な作業結果の出力は困難であることがいえる．作業精度の高い複数のワーカの作業情報を集約することが高品質な作業結果の出力に有効だと考える．

4.4 実験 2,「どのような感情を持っているか」

4.4.1 実験条件

実験 1 では，4.1 節にて用意したデータセットのうち，Q3 の質問に関して作業を行った作業情報のみを抽出し構築したデータセットを用いる．作業によって付与されるラベルは「ポジティブ」「ネガティブ」「ニュートラル」「ポジティブ&ネガティブ」の四種類である．本実験では，ラベルをポジティブな感情を「含む」「含まない」の二種類に変換し，データセットを用いた.「ポジティブ」と「ポジティブ&ネガティブ」はポジティブな感情を含むとし,「ネガティブ」「ニュートラル」はポジティブな感情を含まないとした．上記のデータセットは，一つのテキストデータに対し，10 人のワーカが同じ作業をしている．しかし，Q1 の質問にて「含まない」と回答した場合，Q3 の質問は行われないため，作業情報がない場合も存在する．一つのテキストデータに対し，5 人以上のワーカが Q3 の質問を回答したテキストデータに関して，付与するラベルを，作業した複数の作業情報の多数決によって決定する．複数の作業情報を用いた多数決によって決定したラベルを投票結果とする．あるテキストデータに関して，投票結果がポジティブな感情を「含む」「含まない」のどちらかに決まる場合は，そのラベルを最終的に付与するラベルして決定する．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まらない場合は，そのテキストデータはラベルなしとした．この操作によって構築されたデータセットのラベル別のテキストデータ総数は,「含む」が 3,096 件,「含まない」が 3,437 件であった．投票結果が付与されたデータセットと各ワーカの作業情報を比較し，3.1 節にて説明した正解作業数と不正解作業数を求める．

このデータセットを用いた実験は，一つのテキストデータあたり作業したワーカの人数は 5 人から 10 人であることが分かっているため，選出するワーカ人数を 1 人，3 人，5 人ワーカの 3 パターンで行った．これは，2 値分類において作業情報の多数決の結果が一意に決まる人数である．ベースラインである単純多数決は，無作為に抽出した数人のワーカの作業結果を用いた多数決によって作業結果を決定する．F 値によりワーカの作業精度を評価する集約手法では，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．本手法である，確率分布に従うサンプリング値を用いた集約手法では，サンプリング値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

4.4.2 結果・考察

4.2 節にて説明した 3 種類の手法を用いて集約した場合の正解データセットとの一致率を F 値で算出し，表 4.4.2 に示す．ベースラインである単純多数決による集約は，選出人数が増えるほど精度が高くなっている．単純多数決においては，作業情報が増えるほど，正解データセットとの一致率が高くなることが考えられる．これは，作業情報増えることによって，様々なワーカの意見が反映され，その中に有識者の作業情報が多く含まれようになるからだと考える．以後，ベースラインである単純多数決の集約精度は最も高かった 0.752 を採用し，他の手法の集約精度と比較を行う．

本手法である，サンプリング値を用いた集約手法の集約精度は，選出人数が 3 人の時に最も高く，0.756 であった．ベースラインの 0.752 と比べて本手法の集約精度が高いことから，サンプリング値を用いた集約は高品質な作業結果の出力に有効であることが分かる．これは，サンプリング値が高いワーカが実際に作業精度が高いワーカであり，作業精度の高いワーカの作業情報を用いた多数決によって集約精度が向上したと考えられる．

F 値により作業情報を評価する集約手法では，選出人数が 5 人の時に最も集約精度が高く，0.760 であった．本手法を用いた場合の作業精度は 0.756 であり，F 値により作業情報を評価する集約手法のほうが集約精度が高かった．しかし，F 値により作業情報を評価する集約手法を用いる際，選出人数が 3 人の

表 2　実験 2 における各手法を用いた際の F 値

	選出人数（人）			
	1	3	5	all
ベースライン	0.712	0.730	0.745	**0.752**
F 値上位	0.755	0.751	**0.760**	0.752
サンプリング	0.749	**0.756**	0.752	0.752

時の集約精度はベースラインよりも低い 0.751 であった．このことから，F 値により作業情報を評価する集約手法では，F 値の高い順に上位 4 番目と 5 番目に実際に作業精度の良いワーカおり，選出人数が 5 人の時にたまたま精度が良くなったと推測される．対して，本手法は，選出人数が 3 人の時点でベースラインよりも集約精度が高い．このことから実際に作業精度の高いワーカを優先的に選出するという点では，F 値により作業情報を評価する集約手法に比べて優れているといえる．また，本手法と F 値により作業情報を評価する集約手法における作業精度が高いと推測した上位 3 人のワーカに対して作業数の平均を求めた．本手法におけるサンプリング値が高い上位 5 人のワーカの平均作業数は 678.36 件であった．また，F 値により作業情報を評価する集約手法における F 値が高い上位 5 人のワーカの平均作業数は 674.61 件であり，本手法を用いた場合のほうが高い値であった．以上の結果から，サンプリング値を用いたワーカの選出は，作業精度が高いと考えられるワーカのうち，総作業数が多く，作業精度推定の信頼度が高いワーカを優先的に選出できると考える．

実験 2 の結果は，本手法を用いた場合の精度向上率があまり良くなかった．また，本手法の集約精度は F 値により作業精度を評価する手法の集約精度より低かった．本手法の集約精度があまり良くなかった理由について二つ考える．一つ目は，作業情報の集約に正解の作業結果との一致率が低い投票結果を用いていることである．正解の作業結果が一致しない投票結果を用いてワーカの作業情報を集計すると，誤った作業精度推定をしてしまう．その結果，実際は作業精度が高いワーカを取りこぼしてしまう可能性が高くなると考える．二つ目は，データセットを構築する際の質問が人によって回答がわかれやすいものであったことである．実験 2 で用いたデータセットを構築する際の質問は,「どのような感情を持っているか」という感覚的な内容であった．そのため，ワーカごとに判断基準が異なり，真面目に作業したワーカだとしても作業依頼者が求める作業結果を出力できない可能性が高くなる．これによって，推定される作業精度と本来の作業精度の誤差が大きくなり，集約精度があまり向上しなかったと考える．

5 おわりに

本研究は，ワーカの作業精度を考慮した結果集約によって，高品質な作業結果を出力することを目的としている．しかし，ワーカの作業精度を推定するうえで，総作業数が少ないワーカの正確な作業精度を推定できない問題がある．そこで我々は，作業精度が高いワーカの中でも，作業精度が多く，推定される作業精度の信頼度が高いワーカを優先的に選出する手法を提案する．

作業精度を Accuracy や F 値を用いて求める場合，ワーカの作業数を考慮していないため，推定される作業精度の不確かさを評価することができない．そこで我々は，ワーカごとに確率分布を生成し，その確率分布に従うサンプリング値が高い上位数人の作業情報を集約することによって推定される作業精度の不確かさを考慮した集約ができると考えた．実験では，手法の有効性を確認するため，三種類の集約手法を比較した．ベースラインとして，単純な多数決による方法と．ワーカの作業精度を F 値を用いて推定し，推定した作業精度が高いワーカの作業結果を用いた多数決によって作業結果を決定する方法である．それぞれの集約手法を用いて集約した結果と正解の作業結果を比較し，F 値を求めた．

実験の結果，確率分布に従うサンプリング値を用いた結果集約は，高品質な作業結果の出力に有効であることが確認できた．これは，本手法を用いることによって作業精度の高いワーカを選出できており，高品質な作業結果を出力できているからだと考える．また，確率分布に従うサンプリング値を用いたワーカの選出は，総作業数が多く，作業精度の高いワーカを優先的に選出できていることが確認できた．ワーカの作業数が多いほど，推定される作業精度の信頼度は高いため，安定した集約精度向上に繋がっていると考えられる．

今回行った二つの実験において，本手法を用いた集約がベースラインである単純多数決に比べて，高い集約精度であることを確認した．本手法を用いる際，確率分布に従うサンプリング値が高いワーカの選出人数によらず，集約精度はベースラインと比べて高かった．これは，作業精度の低いワーカのサンプリング値は低いため，選出されにくくなっていることが考えられる．作業数が多く，推定される作業精度の信頼度が高いワーカのうち作業精度の低いワーカは，確率分布に従うサンプリングで低い値が出やすく，選出される確率が低い，そのため，本手法を用いることによって作業精度の低いワーカの作業情報を高精度で排除することができる．

また，本手法を用いた集約は F 値による集約と比べて，作業数が多く，作業精度の高いワーカの選出精度が高いことを確認した．作業数が多く，作業精度が高いワーカの確率分布は，本来の作業精度に収束するため，サンプリング値が安定して高い値を出力する．そのため，サンプリング値を元にワーカを選出する際に，選ばれやすくなっているからだと考えられる．本手法を用いることによって推定される作業精度の推定誤差が小さいワーカを優先的に選出でき，推定誤差を考慮してない場合と比べて，安定した集約精度向上に繋がっていると考えられる．

本研究を通じ，確率分布に従うサンプリング値を用いた結果集約手法は，高品質な作業結果の出力に有効であることが分かった．我々は，確率分布に従うサンプリング値を利用したワーカの選出は，ワーカの作業精度を評価する方法の一つになりうると考える．また，本手法を用いる際に，サンプリング値の高いワーカの選出人数が，実験 1 では 5 人，実験 2 では 3 人の時に最も集約精度が高かった．これは，およそ半分の作業情報は多数決によって作業結果を決定する際のノイズになっていることを示している．そこで我々は，このようなノイズなりうる作業情報を事前に生み出さないように工夫することで，さらなる品質改善とコスト削減ができるのではないかと考える．

今後の展望として，データセットを変えたさらなる検証実験と手法の改善を考えている．本実験では，主にテキストデータに対する 2 値分類のデータセットを対象に実験を行った．今後

は本手法が，多値分類や画像の分類タスクを目的としたデータセットなどに対しても有効であるかを検証したいと思う．また手法の改善に関しては，主にワーカの作業情報の集計の際にさらなる工夫をしていきたい．作業情報の集計の際，本手法は正解作業数を投票結果と同じ作業情報であった作業数としている．しかし，タスクによっては難しいタスク，簡単なタスクが存在するため，一概に全てのタスクの価値を同じに考えるべきではない．適切な重みを付与した集計を行うことによって，集約精度が向上するのではないかと考える．

謝辞 本研究の一部は，JSPS 科学研究費 (19H04218, 23H03694），越山科学技術振興財団，およびソニー株式会社の助成を受けたものです．

文　献

[1] Jeff Howe, et al. The rise of crowdsourcing. *Wired magazine*, Vol. 14, No. 6, pp. 1–4, 2006.

[2] Xiaomei Zhang, Yibo Wu, Lifu Huang, Heng Ji, and Guohong Cao. Expertise-aware truth analysis and task allocation in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*, Vol. 20, No. 3, pp. 1001–1016, 2019.

[3] Chien-Ju Ho and Jennifer Vaughan. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 26, pp. 45–51, 2012.

[4] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, Vol. 90, pp. 123–133, 2013.

[5] Thomas Gillier, Cédric Chaffois, Mustapha Belkhouja, Yannig Roth, and Barry L Bayus. The effects of task instructions in crowdsourcing innovative ideas. *Technological Forecasting and Social Change*, Vol. 134, pp. 35–44, 2018.

[6] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28, 1979.

[7] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, Vol. 22, , 2009.

[8] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, Vol. 23, , 2010.

[9] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pp. 155–164, 2014.

[10] 小山聡, 馬場雪乃, 櫻井祐子, 鹿島久嗣. クラウドソーシングにおけるワーカーの確信度を用いた高精度なラベル統合. 人工知能学会全国大会論文集 第 27 回 (2013), pp. 2M5OS07b2–2M5OS07b2. 一般社団法人 人工知能学会, 2013.

理想と現実のボディイメージの数理モデル化

栗生紗希帆† 眞鍋 雅恵† 若宮 翔子† 荒牧 英治†

† 奈良先端科学技術大学院大学 〒 630–0192 奈良県生駒市高山町 8916-5
E-mail: †{kuriu.sakiho.ks0,mmanabe,wakamiya,aramaki}@is.naist.jp

あらまし 若年女性の過度なやせ願望や無理なダイエットが問題視されている．国際的に見ても日本人女性のやせ問題は顕著であり，成年女性のやせの割合は先進国の中で最上位である．本研究ではまず，クラウドソーシングを用いて，年代・性別・職業ごとに理想の BMI を調査した．次に，理想 BMI に到達すると，さらに，理想を更新するという仮説をたて，現在の若年女性が，どのようなボディイメージに至るかを数理的に検証した．この結果，女性はいずれの年代においても，標準体重の下限付近の BMI18 になるまで，収束しないことが観測され，特に若年女性においては，過激な目標を持ち，急速な収束が観測された．

キーワード ボディイメージ，BMI，やせ，若年女性，数理モデル，クラウドソーシング

1 はじめに

近年，やせを礼賛するメディアの影響や，成長に伴い生じるボディイメージの変化などにより，若年女性の無理なダイエットが増加している．ボディイメージとは，個人の身体的外見に関する考え方，感じ方，行動，さらに身体機能に関するもの [1] である．ボディイメージは幼少期から形成され始め，思春期に獲得されると考えられている．若年女性が無理なダイエットを行う要因として，ボディイメージの歪みが大きく関連している．ダイエットによる過度なやせがもたらす病気として，摂食障害があげられる．摂食障害とは，拒食や過食などの摂食に関する異常がみられる障害である．拒食は過度の食事制限で極端な低体重になる症状である．一方で，過食は大量の食事を摂取した後に食べ過ぎの埋め合わせのために，意図的に嘔吐して体重増加を防ぐ行動である．さらに，やせた若年女性の食後高血糖となる耐糖能異常が多く，その原因として主に肥満者に生じるインスリン抵抗性や脂肪組織の異常との関連性が明らかにされている [2]．

このように，若年女性のやせの問題が深刻となっている背景から，今後のやせの推移がどのように変化するのかという予測が求められている．2019 年における 20 代の BMI 平均値 [3] は，女性が 21.00，男性が 22.90 であった．なお，BMI は，肥満度を表す指標として国際的に用いられている体格指数である．厚生労働省が実施している国民健康・栄養調査 [4] によると，肥満度（BMI）が 18.5 未満の「やせ」が 20 歳代女性では 20.7 % であった．令和 14 年度における 20 歳〜30 歳代女性のやせの者の割合目標が 15% [5] であるのに対して，実際は上回っていることがわかる．このように，日本の若年女性の全体としては，国際的には，すでに十分やせている状態であり，多くの女性にとってダイエットの努力は必要ないとも思える．その一方で，ダイエットなどやせることへの関心は高く，高すぎる理想とのギャップが認められる．OECD 加盟国，人口 1 億人以上の国など 50 カ国を対象に実施された成人女性のやせ（BMI< 18.5）の割合比較 [6] によると，日本の成人女性のやせの割合は先進国の中で最も高かったため，本研究では対象を日本とする．

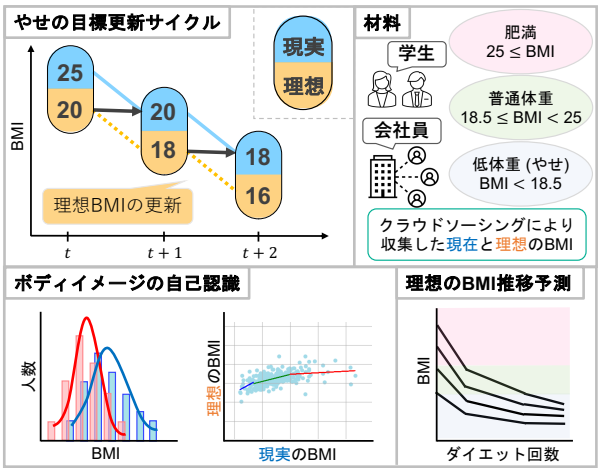


図 1: 本研究の概要．クラウドソーシングを用いて，年代・性別・職業ごとに，現実と理想の BMI を調査し，理想 BMI に到達すると，さらに，理想を更新すると仮定し，若年女性が，どのようなボディイメージに至るかを検証する．

このように，多くの女性がダイエットし続ける現状を鑑みて，本研究では，**若年女性の理想と現実の体重が最終的にどのように落ち着くのかを明らかにする**ことを試みる．そこで，クラウドソーシングを用いて，年代・性別・職業ごとに現実と理想の BMI を収集し，**ボディイメージの自己認識**，**理想と現実のギャップ**，**理想の BMI 推移**を調査する．特に，理想の BMI 推移の調査について，これを実現する直接的な方法は，ある女性がダイエットしていく過程を長期間観測し，理想体重および現実の体重の変化を継時的に得ることである．しかし，このような前向き観察研究は，大規模かつ長期の観測が必要であり難しい．そこで，我々は，一時点ではあるものの，多くの年代・性別，体重を網羅した大規模な人々から，現実の体重と理想とな

る体重を収集し，これをつなぎ合わせることで，擬似的に長期間のデータを作成した．具体的には，ダイエットに成功し，現時点 t で持っている理想体重 BMI_{t+1} に，次の時点 $t+1$ で到達するものとする．次に，理想体重 BMI_{t+1} 付近の人々が持っている現時点 t の理想体重 BMI'_{t+1} を，次々時点 $t+2$ の新たな理想体重 BMI_{t+2} として更新する．このようなシミュレーションにより，ダイエットに成功し続け（理想を更新し続け）ると，どのような体重に，どれくらいの速度で収束するかを調査する．

2 関連研究

ボディイメージややせに関する研究は多岐にわたり，若年層による SNS 利用の悪影響や BMI を用いたボディイメージの自己評価などがあげられる．若年層による SNS 利用の研究では，COVID-19 のロックダウンが SNS の利用に影響を与え，若年層において摂食障害のリスクとの関連性が示唆されている [7]．これは，Instagram の利用頻度が体型不満ややせ願望，自尊心の低さと正の相関があることを示している．さらに，SNS の過剰な使用が摂食障害のリスク上昇と関連することが立証されている [8]．このことは SNS 利用率が高い若年層にとって大きな問題であることが読み取れる．

BMI を用いたボディイメージの研究では，BMI とボディイメージの自己認識との関係の評価が行われている [9]．ボディイメージの自己認識の歪みが示され，適切な BMI を持つ女性でもより低い BMI を望むという事実が懸念されることを示唆している．さらに，BMI を用いたボディイメージの自己評価が分析されている [10]．自分の BMI が高すぎると過大評価することは，摂食障害の初期兆候であり，男性よりも女性に拒食や過食がみられることを述べている．

3 材料

BMI は，肥満度を表す指標として国際的に用いられている体格指数であり，肥満や低体重（やせ）の判定に用いられている．本研究でもこの指標を用いて「やせ」（BMI< 18.5 を判定する．なお，表 2 に BMI の基準を示す．本研究では，クラウドソーシングを利用し，2023 年 12 月 3 日から 12 月 5 日において 20 代，30 代，40 代の男女，各 300 人にアンケートを行った．1 タスクあたりの謝礼は 3 円とした．以下に質問例と回答例を示す．

質問 1：「あなたの身長を教えてください」
回答： 130cm〜200cm から一つ選択
質問 2：「あなたの体重を教えてください」
回答： 20kg〜100kg から一つ選択
質問 3：「あなたの理想体重を教えてください」
回答： 20kg〜100kg から一つ選択
質問 4：「あなたの職業を教えてください」
回答： 会社員，公務員，自営業／個人事業／自由業，会社役員，専業主婦・主夫，学生，パート／アルバイト，無職，その他から一つ選択
質問 5：「あなたは今，太っていると思いますか」
回答： 太っていると思う／太っていないと思うから一つ選択

クラウドソーシングで取得した，各個人の身長と体重のデータを用いて現実の BMI を，身長と理想体重のデータを用いて理想の BMI を計算し，データセットを作成した．なお，BMI の値は，[体重 (kg)] ÷ [身長 (m) の 2 乗] で算出される．

4 実験

4.1 手法

現実と理想の BMI について，以下の 3 つの調査により分析する．

調査 1：ボディイメージの自己認識

各個人が自身のボディイメージについて，どのように認識しているかを調査する．そのために，自身の現実の体重に対して「太っていると思うか」という意識調査を行い，年代・性別ごとに比較する．さらに，若年女性の特徴を探るため，20 代女性の職業ごとに比較し，属性による違いを明らかにする．これにより，各個人が持っているボディイメージの自己認識を調査する．

調査 2：理想と現実のギャップ

これは，現実の BMI と理想 BMI との関係性を分析するための調査である．そのために，年代・性別ごとの比較により，現実と理想の BMI の特徴を分析する．具体的には，年代・性別ごとの BMI 分散に加え，20 代女性については職業ごとの BMI 分散を求めて比較する．

調査 3：理想の BMI 推移

ダイエットを行った場合の理想の BMI 推移を予測し，やせの最終目標を明らかにする．ダイエットが成功したとき，人々はより高い理想を持つことが考えられる．つまり，理想の BMI が更新される．そこで，本研究では，ダイエットが成功したとき，理想が更新されることを前提とする．ある時点 t で持っている理想体重に，次の時点 t+1 でダイエット後の体重となる．体重が更新されるとその体重付近の人々が持つ理想体重の平均を新たな理想体重として更新し，t+2 に至る．このようにして，理想の BMI がどのように推移するかをシミュレーションする．理想の BMI 推移シミュレーションではダイエットを行ったときの BMI 推移を予測することを目指す．現実の BMI から理想 BMI への変換を 1 回のダイエットとする．現実の BMI と理想の BMI は，それぞれ整数となるように四捨五入を行う．現実の BMI が同じ人が存在している場合には，理想の BMI をそれぞれの平均を取ることにより算出する．本研究では，前提条件として，ダイエット成功確率を 100%と仮定する．

4.2 結果と考察

4.2.1 ボディイメージの自己認識

図 2 (a)˜(f) に，年代・性別ごとの「現在，太っていると思うか」という意識調査の分布を示す．近似分布は，「太っていると思う」がガンマ分布，「太っていないと思う」が正規分布に従っている．「太っていると思う」人の範囲は，男性と比較して女性

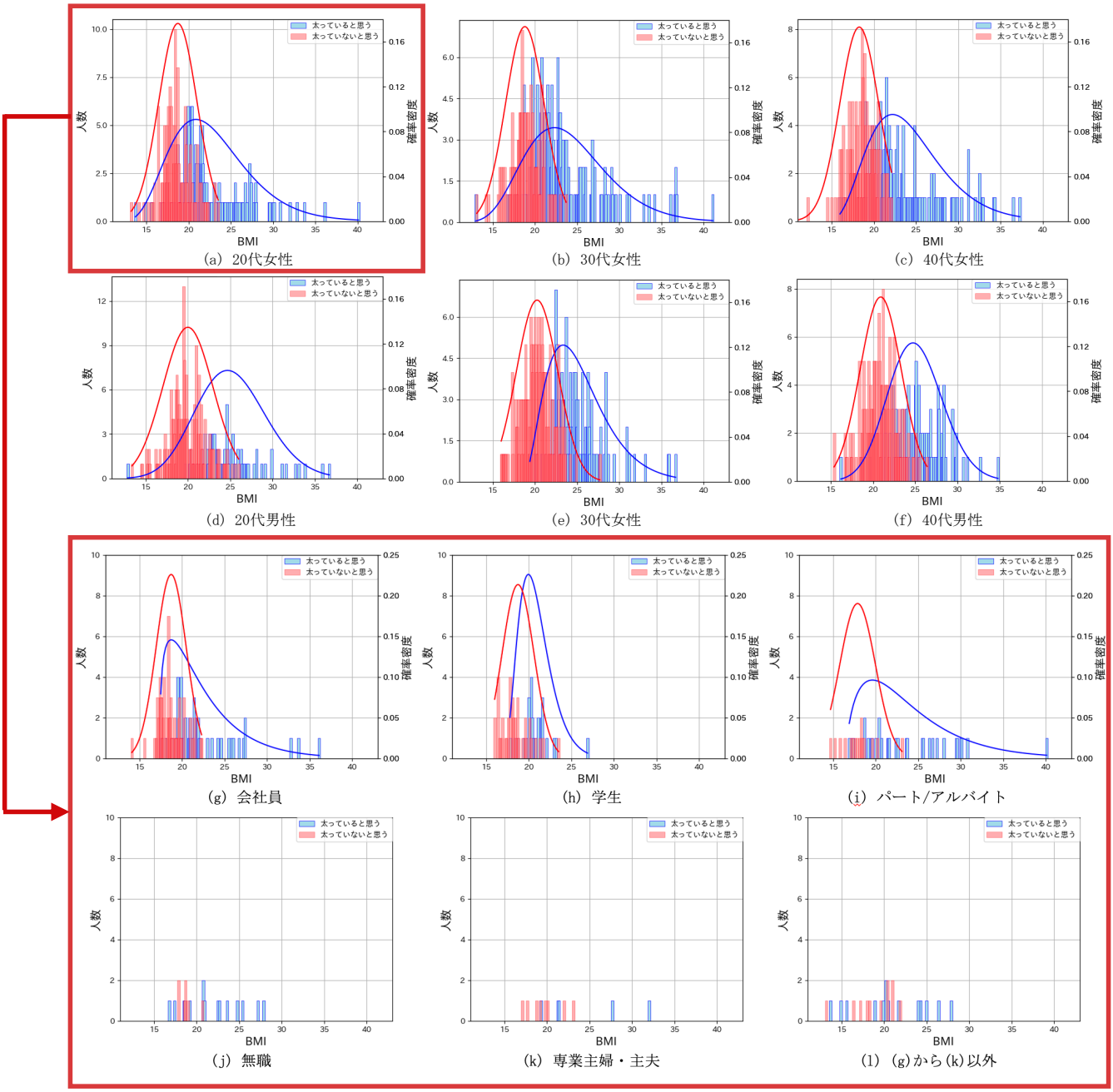


図 2: 現実の BMI 認識に関する意識分布．(a)˜(f) は，年代・性別ごと，(g)˜(l) は，20 代女性における職業別ごと．職業別の分布図の実線は，分布の近似曲線を表す．なお，サンプル数が少ない職業については，実線を引いていない．

はやや小さい値となっていることが読み取れる．ここで，現在の BMI が普通体重でかつ「太っていると思う」人の割合に着目する．普通体重である BMI18.5 以上 25 未満の人の中で,「太っていると思う」人の割合は，20 代女性が 60.7%，30 代女性が 63.9%，40 代女性が 57.9%，20 代男性が 24.5%，30 代男性が 34.4%，40 代男性が 28.4%であった．女性の割合は 60%前後，男性の割合は 30%前後であることから，性別による差が顕著に現れた．この結果から，普通体重であるにも関わらず，やせ願望が強いことが読み取れる．

一方で，年代による比較を行うと，男女共に 30 代が最も低い割合となっていた．さらに，低体重（やせ）である BMI18.5 未満の人の中で,「太っていると思う」人の割合は，20 代女性が 14.6%，30 代女性が 12.5%，40 代女性が 4.4%，20 代男性が 2.0%，30 代男性が 0.0%，40 代男性が 5.3%であった．ここでは特に，20 代と 30 代女性の割合が比較的大きいことが確認できる．現実の BMI は低体重（やせ）だが，自身の体型を「太っていると思う」人の割合では，20 代，30 代女性が比較的高く，若年女性を中心に過剰なやせ願望が存在していると考えられる．

さらに，若年女性の属性について詳しく調査するために，図 2 (g)˜(l) に，職業別の意識調査の分布を示す．職業内訳は，会社員が 123 人，学生が 70 人，パート／アルバイトが 47 人，無職が 22 人，公務員が 14 人，専業主婦・主夫が 13 人，自営業

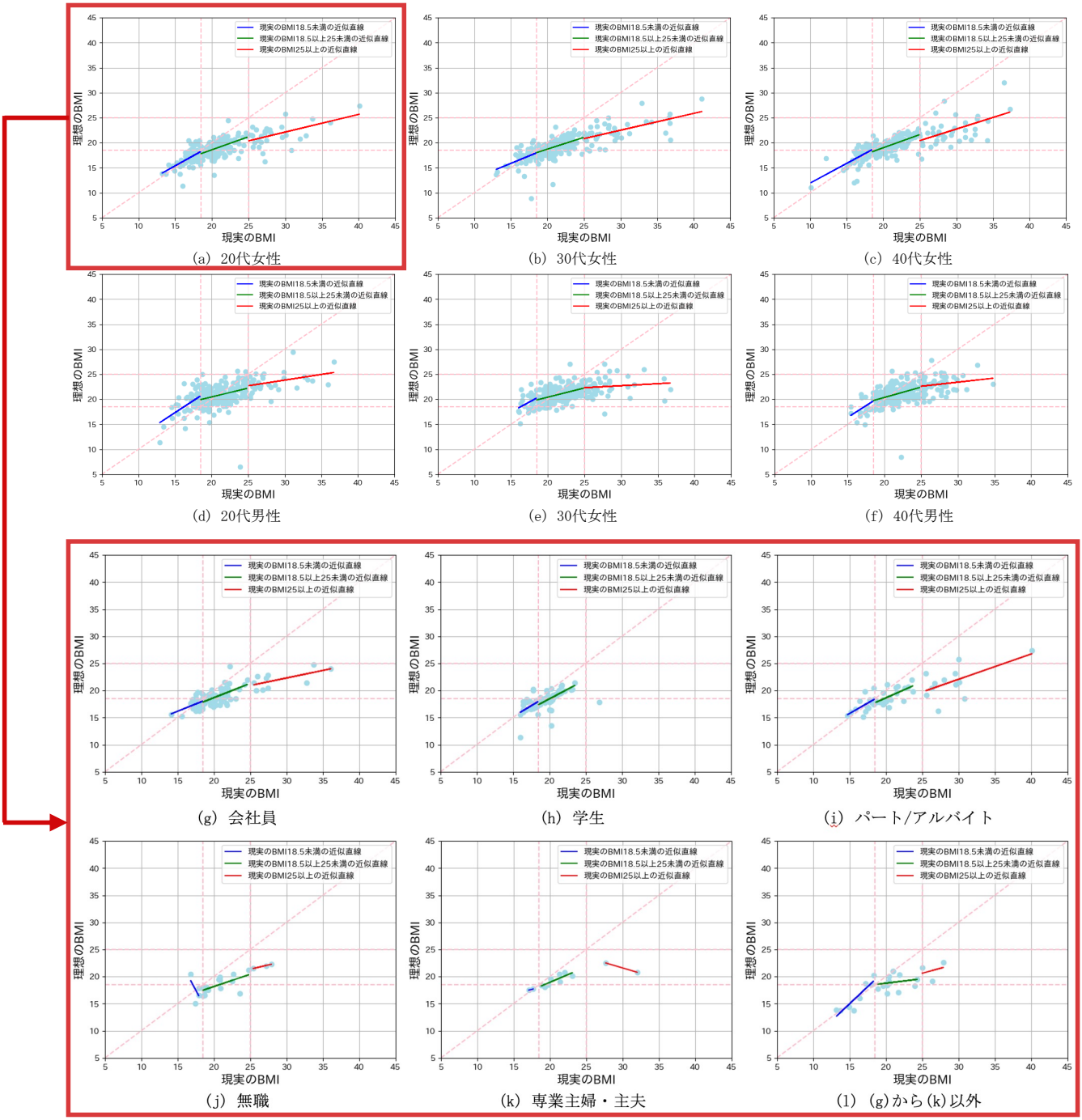


図 3: 現実と理想の BMI 分散．(a)˜(f) は，年代・性別ごと，(g)˜(l) は，20 代女性における職業別ごと．その他に含まれる職業は，公務員，自営業/個人事業/自由業，会社役員，その他である．分布図の実線は，分布の近似直線を表す．現実の BMI を低体重（やせ），普通体重，肥満の 3 つの領域に分け，それぞれ青線，緑線，赤線で引いている．なお，サンプル数が少ない職業については，近似直線を引いていない．

／個人事業／自由業が 8 人，会社役員が 2 人，その他が 1 人であった．特に人数が多い会社員，学生，パート／アルバイトと，他者と対面する時間が少ないと考えられる無職，専業主婦・主夫と，それ以外の職業をその他とし，それぞれの分布を確認する．会社員や学生は，普通体重では，太っていないと思う人が一定数いることが読み取れる．一方で，パート／アルバイトや無職では，普通体重である場合，太っていると思う人が多く存在する．ここから，会社員や学生は，多様な意見をもつのに対して，パート／アルバイトや無職では，やせ願望が強いと考えられる．無職やパート／アルバイトのように，他の職業と比較して，他者との対面時間が少ないと考えられる属性において，過剰なやせの意識が存在していることが示された．

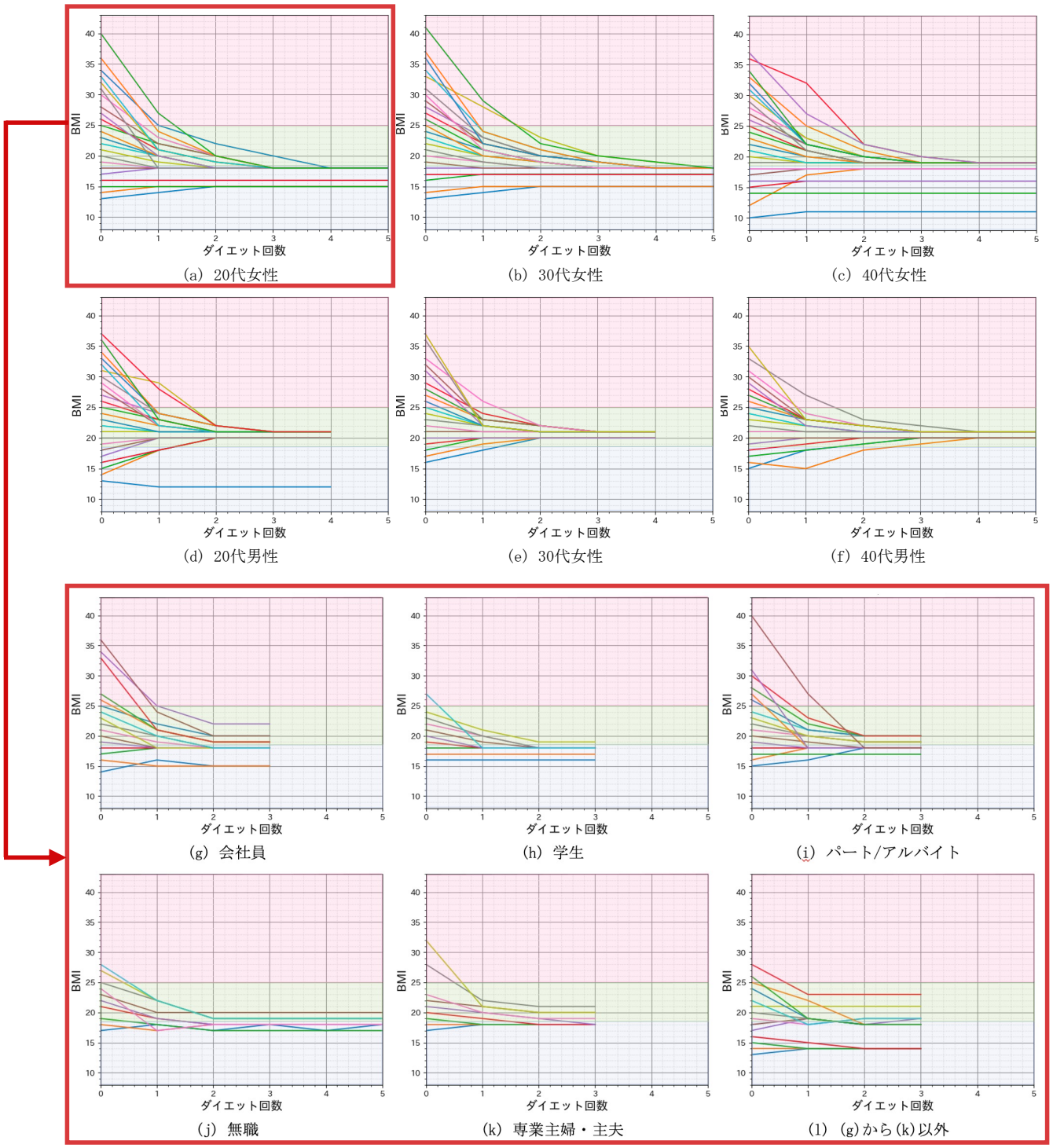


図 4: 年代・性別ごとの理想の BMI 推移．青の領域は低体重（やせ），緑の領域は普通体重，赤の領域は肥満を表す．

4.2.2 理想と現実のギャップ

年代・性別ごとの現実の BMI と理想 BMI の分散を図 3 (a)˜(f) に示す．$y = x$（現実と理想の BMI が等しい）に補助線を引いた．さらに，x 軸と y 軸それぞれに，低体重（やせ），普通体重，肥満のカットオフ値を示した．$y = x$ より下の領域は，現実の BMI より理想の BMI が低く，やせ願望がある領域である．この領域には女性に多く確認される．やせ願望のある領域に大多数が属している様子が読み取れる．さらに，この領域のうち，$y <$ 18.5 に属する領域は，理想 BMI が低体重（やせ）であり，過度なやせ願望をもっている．この領域には，特に 20 代女性が，多く存在していることが確認される．

表 1 に図 3 (a)˜(f) の分散図を，現実の BMI を低体重（やせ），普通体重，肥満の 3 つの領域に分け，それぞれに近似直線を引いた傾きを示す．近似直線の傾きは，現実の BMI が高くなるにつれて小さくなり，やせ願望が強くなっていることが読み取れる．現実の BMI が低体重（やせ）の場合は，傾き 1

表 1: 年代・性別ごとの近似直線の傾き

年代・性別	低体重（やせ）	普通体重	肥満
20 代女性	0.818	0.523	0.353
30 代女性	0.600	0.472	0.336
40 代女性	0.788	0.527	0.463
20 代男性	0.958	0.353	0.227
30 代男性	0.798	0.368	0.080
40 代男性	0.937	0.412	0.163

に近く，平均として現状のままを理想としていることが読み取れる．さらに，普通体重，肥満と BMI が高くなるにつれて傾きが緩やかになり，現実と理想が一致する $y = x$ との距離が大きくなっていることから，現実と理想の差が大きくなっていることが確認できる．次に，年代・性別ごとに比較する．男性と比べて女性はやせ願望のある領域に集まっており，女性全体でやせ願望が強いことが読み取れる．低体重（やせ）を理想とする人に着目すると，女性は年代が上がるにつれてやや減少傾向ではあるものの一定数存在している．一方で，男性は，20 代の場合では一定数存在しているが，30 代と 40 代の人数は少ないことが確認される．このことから，女性が男性と比べてやせ願望が強いことが読み取れる．普通体重以下の男性は，現実より高い BMI を理想とする人が一定数存在していることから，女性と比べてより健康志向であると考えられる．さらに，男性と女性を合わせた場合では，若年層にやせ願望が強い傾向にあることが考えられる．

図 3 (g)˜(l) に，若年女性の職業別分布を示す．現実の BMI 分布では学生が低い BMI に集まっている．一方で，理想の BMI ではあまり差がみられず，職業別における違いは確認されなかった．若年層の職業別分布では，大きな違いが確認できず，職業とやせとの相関がみられない結果となった．しかし，今回の実験では，参加者の統制はとっておらず，正しく評価できていない可能性がある．

4.2.3 理想の BMI 推移

図 4 に，ダイエット成功確率 100%の理想の BMI 推移の結果を示す．20 代女性の理想の BMI の収束は 15-18 の間に収まっており，最終的に低体重（やせ）を理想としていることが確認される．年代・性別ごとに比較すると，女性は低体重（やせ）を理想としているのに対して，男性は普通体重を理想としている．40 代女性では，理想の BMI が 19（普通体重）に収束する場合があることが確認された．次に，理想の BMI となるまでのそれぞれのダイエット回数を計算し，その平均収束回数を算出する．結果は 20 代女性が 1.21 回，30 代女性が 1.99 回，40 代女性が 1.11 回，20 代男性が 1.04 回，30 代男性が 1.26 回，40 代男性が 1.52 回であった．

20 代女性の推移では，収束がすべて低体重（やせ）であることから，理想 BMI の更新により普通体重では満足しないことが読み取れる．女性は低体重（やせ）を最終目標とし，男性は普通体重を最終目標としていることから，性別による違いが大きく存在する．さらに，収束回数で考えると，20 代女性は 30 代男性とほとんど同じ回数で収束しているが，それぞれの最終目標の違いから，20 代女性はより現実と理想の差が大きいと考えられる．40 代女性が最も少ない回数となっているが，これは理想の BMI が 19（普通体重）に収束する場合があることが影響していると考えられる．20 代女性と 30 代女性を比較すると，同じ収束域（15-18）にも関わらず 20 代女性の方が，収束回数が少ないことから，若年女性による現実と理想の差の大きさが読み取れる．一方で，男性は，年代が上がるごとに収束回数が増加していることから，理想が緩やかになっていると考えられる．それぞれの最終目標を比較すると，女性はいくつかの BMI に分かれているのに対して，男性は比較的近い値に固まっていることが確認できる．これは，女性は個人ごとに異なる理想のボディイメージを目指してダイエットをするが，男性は共通化された理想のボディイメージをもち，ダイエットをしていることが考えられる．

次に，理想の BMI が現実の BMI より高い人に着目する．これは，現在より太りたいと考えているケースである．20 代女性は BMI が 13, 14, 17 のとき，30 代女性は BMI が 13, 14, 16 のとき，40 代女性は BMI が 10, 12, 15, 17 のとき，20 代男性は BMI が 14-19 のとき，30 代男性は BMI が 16-19 のとき，40 代男性は BMI が 15, 17-19 のときにそれぞれ現実より高い BMI を理想としている．低体重（やせ）の BMI は 18.5 未満であることから，全体をみると正しい判断であると考えられる．しかし，最終目標では，女性が低体重（やせ），男性が普通体重となっていることから，大きな違いが読み取れる．普通体重であっても，女性は，ダイエットを行う必要があると捉えていることが懸念される．

4.3 限　　界

本研究では，ダイエットの成功確率を 100%と仮定してシミュレーションを行ったが，実際のダイエットには多くの要因が含まれる．具体的には，ダイエットが成功しない場合や，リバウンドする場合，途中で諦める場合などがあげられる．さらに，Instagram などの SNS の使用に伴い，やせた芸能人やインフルエンサーを見る機会が増えることにより，ボディイメージが変化し，より過度なやせ願望をもつことが考えられる．これらを考慮した上で，より現実的なダイエットのシミュレーションを行うことが，今後の課題となる．ダイエットが成功しない場合やリバウンドなどは，それらの発生確率を定義することで，現実的なダイエットに近づくことが考えられる．

次に，本実験で行った職業による比較では，サンプル数が少なく，正しく評価できない結果となった．しかし，職業による比較は，社会との関わりの強さにも関係するため，より追求することで，やせと社会との関係性が明らかになると考えられる．社会との関わりを考える上で，地域差も重要な違いである．そのため，都会と田舎に住んでいる人との比較を行うことも今後の課題である．

さらに，本研究では日本を対象に調査を行ったが，やせの問題は国ごとに大きな違いがあるため，国際比較を行い，世界と比較した日本のやせ問題を調査していくことが課題となる．本

研究では，過度なやせを問題として取り上げたが，体質などにより太れない人がいる場合も考慮しなければならない．そこで，本研究で用いた BMI 以外の健康指標に着目した調査も必要である．

5 おわりに

本研究では，BMI を用いてボディイメージについて分析し，ダイエットによる理想の BMI 推移のシミュレーションを行った．BMI を用いたボディイメージの分析では，若年女性の BMI 分布，各個人の現実と理想の BMI 分布を用いた年代・性別ごとによる比較，若年女性の職業別 BMI 分布の比較を行った．女性と男性を比較すると，女性は現実の BMI が普通体重の場合にも，より低い BMI を理想とし，やせたいと考えていることを確認した．一方で，女性の年代を比較すると，若年女性ほどやせ願望が強いことを確認した．理想の BMI 推移のシミュレーションでは，各年代・性別のやせの最終目標を明らかにし，やせ願望の詳細な緻密化を行うことができた．今後，日本人の BMI を長期的に観測すると，数十年後には本研究でのシミュレーションのような平均値となる可能性を示唆した．

今後の課題として，今回作成した理想の BMI のシミュレーションを実際のダイエットと比較し，有効性の評価検証する必要がある．シミュレーションの有効性を確認し，今後のやせの推移を予測することを目指す．さらに，やせに関係する要因を探るため，職業による比較を行い，やせのシミュレーションに反映することを目指している．

謝　辞

本研究は，花王株式会社共同研究費の支援を受けたものである．

文　献

[1] Bree D Abbott and Bonnie L Barber. Embodied image: Gender differences in functional and aesthetic body image among australian adolescents. *Body image*, Vol. 7, No. 1, pp. 22–31, 2010.

[2] Motonori Sato, Yoshifumi Tamura, Takashi Nakagata, Yuki Someya, Hideyoshi Kaga, Nozomu Yamasaki, Mai Kiya, Satoshi Kadowaki, Daisuke Sugimoto, Hiroaki Satoh, et al. Prevalence and features of impaired glucose tolerance in young underweight japanese women. *The Journal of Clinical Endocrinology & Metabolism*, Vol. 106, No. 5, pp. e2053–e2062, 2021.

[3] 厚生労働省. 政府統計の総合窓口 | 国民健康・栄養調査 15 bmi の平均値及び標準偏差 - 年齢階級別、人数、平均値、標準偏差 - 男性・女性、15 歳以上〔妊婦除外〕 — 統計表・グラフ表示, 2019. https://www.e-stat.go.jp/dbview?sid=0003224178.

[4] 厚生労働省. 国民健康・栄養調査, July 2020. https://www.mhlw.go.jp/content/001066903.pdf.

[5] 厚生労働省, May 2024. https://www.mhlw.go.jp/web/t_doc?dataId=00012990&dataType=0.

[6] 吉池信男, 小山達也, 三好美紀. 国内外の女性のやせの動向. 肥満研究= Journal of Japan Society for the Study of Obesity: 日本肥満学会誌, Vol. 24, No. 1, pp. 16–21, 2018.

[7] Helena Vall-Roqué, Ana Andrés, and Carmina Saldaña. The impact of covid-19 lockdown on social network sites use, body image disturbances and self-esteem among adolescent and young women. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Vol. 110, p. 110293, 2021.

[8] Juan Zhang, Yihui Wang, Qianru Li, and Chenggang Wu. The relationship between sns usage and disordered eating behaviors: a meta-analysis. *Frontiers in Psychology*, Vol. 12, p. 641919, 2021.

[9] Idalina Shiraishi Kakeshita and Sebastião de Sousa Almeida. Relationship between body mass index and self-perception among university students. *Revista de Saúde Pública*, Vol. 40, pp. 497–504, 2006.

[10] H Madrigal, A Sanchez-Villegas, MA Martinez-Gonzalez, J Kearney, MJ Gibney, J De Irala, and JA Martínez. Underestimation of body mass index through perceived body image as compared to self-reported body mass index in the european union. *Public health*, Vol. 114, No. 6, pp. 468–473, 2000.

[11] 日本肥満学会. 日本肥満学会 | jasso 表 肥満度分類. http://www.jasso.or.jp/data/magazine/pdf/chart_A.pdf.

付　録

表 2: 日本肥満学会が定めている基準 [11]

BMI	肥満の判定
18.5 未満	低体重（やせ）
18.5〜25 未満	普通体重
25〜30 未満	肥満（1 度）
30〜35 未満	肥満（2 度）
35〜40 未満	肥満（3 度）
40 以上	肥満（4 度）

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 13:30 ~ 15:40 | T5-A オンライン（Zoom Events)

医療・ヘルスケア (1)

座長:中島 伸介(京都産業大学)
コメンテータ:張 建偉(岩手大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T5-A-5-01] 睡眠ステージ判定を対象としたエンコーダ・デコーダモデルによる脳波信号オーグメンテーション

*大宮 直樹[1]、堀江 和正[2]、北川 博之[3] (1. 筑波大学大学院システム情報工学研究群情報理工学位プログラム、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構)

14:00 ~ 14:25
[T5-A-5-02] 睡眠ステージ自動判定における睡眠時無呼吸症候群の重症度を踏まえた深層学習処理の切替機構の開発

*川畑 竜平[1]、堀江 和正[1]、北川 博之[1] (1. 筑波大学)

14:25 ~ 14:50
[T5-A-5-03] 気象データを統合的に用いたクラウドソーシングに基づく頭痛発生傾向の検知

*飯田 静空[1]、若宮 翔子[1]、清水 伸幸[2]、藤田 澄男[2]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. LINEヤフー株式会社)

14:50 ~ 15:15
[T5-A-5-04] データ駆動型研究を促進する大規模かつ多種類の生体光学データのためのデータベース基盤の開発

*山中 卓也[1]、熊本 康昭[2,3]、西村 隆宏[2]、久保 俊貴[2,4]、Heqi Xi[2]、藤田 克昌[2]、松原 靖子[1]、櫻井 保志[1] (1. 大阪大学産業科学研究所産業科学AIセンター、2. 大阪大学大学院工学研究科、3. 大阪大学先導的学際研究機構、4. 大阪大学大学院医学系研究科)

15:15 ~ 15:40
[T5-A-5-05] 【技術報告】類型化と解釈性を考慮した可視化による自動車事故分析

桑元　凌 (三井住友海上火災保険株式会社ビジネスデザイン部)

睡眠ステージ判定を対象としたエンコーダ・デコーダモデルによる脳波信号オーグメンテーション

大宮　直樹†　　堀江　和正††　　北川　博之†††

† 筑波大学大学院ｼｽﾃﾑ情報工学研究群情報理工学位ﾌﾟﾛｸﾞﾗﾑ　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
†† 筑波大学計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
††† 筑波大学国際統合睡眠医科学研究機構　〒 305–8575 茨城県つくば市天王台 1 丁目 1–1
E-mail: †omiya@kde.cs.tsukuba.ac.jp, ††horie@cs.tsukuba.ac.jp, †††kitagawa@cs.tsukuba.ac.jp

あらまし　脳波などの複数の生体信号から睡眠の状態を判定する「睡眠ステージ判定」は，睡眠の臨床における必須の検査項目の一つである．この検査は多大な時間と労力を必要とするため，深層学習を用いた自動判定手法が数多く提案されてきた．判定モデルの精度改善にはオーグメンテーションが良く行われるものの，脳波のオーグメンテーションは特徴波の形状や位置を変化させてしまうものが多く，睡眠ステージ判定に向いていない．本研究ではエンコーダ・デコーダモデルをベースに特徴波の位置情報を利用するオーグメンテーション手法を開発し，その有用性を検証した．

キーワード　医療・ヘルスケア，生体信号処理，睡眠ステージ判定，ｴﾝｺｰﾀﾞ・ﾃﾞｺｰﾀﾞﾓﾃﾞﾙ，データオーグメンテーション

1　序　　論

アメリカ睡眠医学会は，ヒトの睡眠状態を全部で 5 つのステージ（W：覚醒，REM：レム睡眠，N1：ノンレム睡眠 1，N2：ノンレム睡眠 2，N3：ノンレム睡眠 3）に分類している [1]．これらステージの変遷や分布は，睡眠の質や睡眠障害の有無と関連があるため，ステージを把握することは睡眠臨床上重要である．

実際の睡眠の臨床では，医師や臨床検査技師によって「睡眠ステージ判定」が行われている．これは脳波等の生体信号に含まれる特徴的なパターン（特徴波）をもとに，30 秒（エポック）毎の睡眠がどのステージに割り当てられるか判断するものである．現在は睡眠ステージ判定の全行程が判定技師の目視で行われているため，多大な時間と労力を必要とする．

この問題を解決するため，深層学習を用いた自動睡眠ステージ判定手法が数多く提案された．DeepSleepNet [2] は判定技師が付けた正解ステージラベルとの一致率 86.2%を達成しており，技師間一致率が約 82.0% [3, 4] であることから，これは十分な判定精度を誇る睡眠ステージ判定モデルだといえる．しかし一部の被験者，例えば高齢者や睡眠障害の患者に対しては，判定一致率が 60%程度まで落ち込むことがある．これらの被験者は，健康な若者と睡眠ステージの分布や特徴波の性質が異なることが知られており，これが判定精度の低下につながっている．このような一部被験者に対する判定精度の低下は，自動睡眠ステージ判定手法の信用性・実用性を損ねてしまう．

上記の問題を解決する方法の一つは，高齢者や睡眠障害患者といった比較的特殊な被験者のデータを増やして判定モデルの学習に組み込むことである．しかし先に述べたように，判定自体に大きな労力・時間が必要であり，また原則入院検査となることから，特定の属性のデータを大量に収集するのは難しい．

一般的なタスクでは，十分なデータがない状況で深層学習モデルを学習する際にデータオーグメンテーションが行われる．しかし生体信号のオーグメンテーションは難度が高く，現に今まで提案されてきた生体信号オーグメンテーション手法の多くは特徴波の位置と形状を崩してしまう [5–7]．睡眠ステージ判定においては信号に含まれる特徴波の位置と形状が重要であることから，それらを保持したオーグメンテーションを行わなければならない．

一般的なエンコーダ・デコーダモデルでオーグメンテーションを行う場合．入力信号をエンコードして得た潜在表現ベクトルの一部を変化させて新たな信号を生成する．しかしこの際に特徴波を含むベクトルを変化させることで，生成信号に含まれる特徴波の位置と形状を崩す恐れがあった．それを防ぐために，特徴波の位置を推論できる Sleep-CAM [4] から得られる CAM ベクトルを教師信号として利用し提案モデルの学習を行うことによって，特徴波の位置と形状を保持したオーグメンテーションができると考えた．

2　睡眠ステージ判定

脳波，眼電図および顎筋電図の記録から 30 秒単位（エポック）毎に睡眠ステージのラベル付けを行う検査を「睡眠ステージ判定」と呼ぶ．アメリカ睡眠医学会（AASM）による判定マニュアル [1] には，この検査を行うために必要な技術的側面や，結果の解析・解釈に必要な知識が体系的にまとめられている [8]．

判定技師が 30 秒区間（エポック）の生体信号から睡眠ステージラベルを決定するとき，その判定根拠とする信号のパターンを特徴波という．本研究に特に関わりのある脳波の特徴波一覧を表 1 [4] に示し，その一部を図 1 に示す．

表 1　脳波に含まれる睡眠特徴波.

信号	特徴波	説明	睡眠ステージ
脳波	アルファ波	8-13Hz の波で，主に後頭部に現れる.	W
	LAMF	小さな振幅で，4-7Hz の波.	N1 & REM
	頭蓋頂鋭波	0.5 秒以内の鋭波．他の波とは明確に区別される.	N1
	K 複合波	負の鋭波と正の波．他の波とは明確に区別される．0.5 秒以上持続する.	N2
	紡錘波	11-16Hz の波で，0.5 秒以上持続する．最大振幅は波形の中央に位置する．他の波とは明確に区別される.	N2
	デルタ波	0.2-5Hz の波で振幅のピークは 75μV を超える．しばしば前頭部に現れる.	N3
	鋸歯状波	2-6Hz の鋸歯状の波.	REM

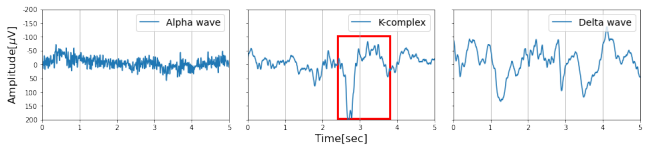


図 1　脳波の特徴波の例.（左）アルファ波.（中央）K 複合.（右）デルタ波.

3　関 連 研 究

以降，文中や数式中で脳波を EEG(Electroencephalogram)，眼電図を EOG(Electrooculogram)，筋電図を EMG(Electromyogram) と表記する.

3.1　自動睡眠ステージ判定モデル

3.1.1　Sleep-CAM

Sleep-CAM [4] は，6 チャンネルの EEG と，2 チャンネルの EOG，1 チャンネルの EMG から睡眠ステージ判定を行うと同時に，その判定根拠となった Class Activation Mapping (CAM) を提示する判定モデルである．既存手法が高い判定精度を達成するも臨床で使用されない理由の一つは，その説明可能性の欠如であり，信頼性のある判定ができなかったことである．そこで判定機構に組み込まれた畳み込み層から得た特徴量の活性値を参照することで，出力結果が入力信号のどの部分に大きく起因するかを明らかにした．図 2 に示すように提案手法の示した CAM が活性化する部分には特徴波が存在したケースがあったことから，判定技師が特徴波を見て睡眠ステージ判定を行うのと同様に，Sleep-CAM は特徴波を考慮して判定を行うモデルだといえる.

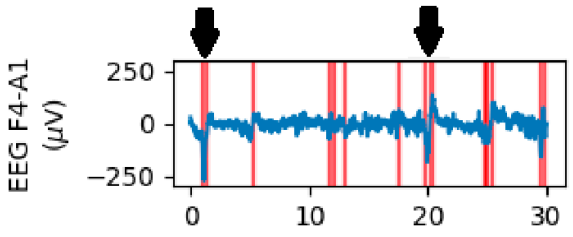


図 2　[4] より引用．赤くハイライトされた CAM 値の上位 10%部分に特徴波の存在が確認できる.

本研究においてオーグメンテーションによって生成した信号は特徴波を含んでいてほしいため，提案モデルに入力する参考情報として上記モデルの CAM 値を用いた.

3.1.2　DeepSleepNet

DeepSleepNet [2] は，1 チャンネル EEG から睡眠ステージ判定を行う畳み込みニューラルネットである．これは MASS データセットを使用して，判定技師がつけた正解ラベルと 86.2%の一致率を達成している.

近年提案される手法の多くは，時間周波数変換等を用いて変形した生体信号を判定モデルに入力する．一方，本モデルは 1 チャンネル EEG をそのまま入力して判定を行う手法であり，波形の情報を考慮した睡眠ステージ判定を行うモデルであることから，本研究の提案手法の効果を検証する際に使用した.

3.2　生体信号オーグメンテーション手法

3.2.1　GPT-2 を用いた EEG, EMG オーグメンテーション

GPT-2 [10] を用いて，3 値分類タスクを対象にした EEG, EMG データオーグメンテーションを行った研究がある [5]. GPT-2 は 800 万のウェブサイトをデータセットとして学習した 15 億パラメータの深層学習モデルである．その時系列信号を出力する特性を活かし，生体信号のオーグメンテーションに利用された.

同論文でも述べられている通り，実 EEG データに含まれる瞬きの波形等，突発的な波形の変化が合成 EEG データには見られない．また，合成 EMG データには実 EMG データよりも多くの高周波成分を含むことに加え，一部不自然な波形を生成することがあった．睡眠ステージの決定には突発的な特徴波の発現や，周波数の分布が重要な情報となるため，この手法を用いて睡眠ステージ判定を対象にしたオーグメンテーションを実装することは難しい.

3.2.2　周波数領域で変換を行う EEG オーグメンテーション

数ある生体信号オーグメンテーション手法を時間領域・周波数領域・時間周波数領域オーグメンテーションの 3 つに分類し，それら手法の有効性を評価したメタアナリシスがある [6]．この論文にて睡眠ステージ判定タスクにおける最も大きな精度向上を達成した FTSurrogate (Fourier Transform Surrogate) [6,7] は，EEG のフーリエ係数をランダム化することで新しい信号を生成する手法である．実装では，フーリエ係数にランダムノ

イズを加えることでオーグメンテーションを実現するため，以下のように定式化できる．

$$\mathcal{F}[FTSurrogate(X)](f) = \mathcal{F}[X](f)e^{i\Delta\phi}$$

ここで，i は虚数単位とし，$\mathcal{F}$ は FFT 関数，X は入力信号，f は周波数，$\Delta\phi$ は加えるランダムノイズの位相である．それぞれの周波数 f に加えるべき $\Delta\phi$ は一様分布 $[0, \Delta\phi_{max}]$ からサンプリングされ，$\Delta\phi_{max} \in [0, 2\pi)$ は FTSurrogate の影響度合いを決定するパラメータである．[6] によると，$\Delta\phi_{max} = \frac{9}{10}\pi$ が最も判定精度向上に適した値であると結論づけられている．

3.2.3 時間軸方向に逆転させる EEG オーグメンテーション

項 3.2.2 で紹介した文献 [6] において，2 番目に大きな精度向上を達成した Time reverse は，EEG の時間軸方向に沿って値を逆転させることで信号を生成する手法である．以下のように定式化でき，ここで，X は入力信号，t は時間，t_{max} は信号長を表す．

$$TRev[X](t) = X(t_{max} - t)$$

4 提案手法

提案手法の概略図を図 3 に，モデル図を図 4 に示す． このモデルはエンコーダ・デコーダによって構成され，オーグメンテーション実行時の入力は 1 チャンネル 1 エポック（30 秒）の EEG で，出力は入力と同型の再構成信号である．

一般的なエンコーダ・デコーダモデルでオーグメンテーションを行う場合，入力信号をエンコードして得た潜在表現ベクトルの一部を変化させてデコードする．しかしこの方法では，特徴波情報が含まれるベクトルまでも変化させてしまう恐れがあり，新しく信号を生成したときに特徴波が崩れてしまう可能性がある．そこで本手法では特徴波と潜在表現ベクトルの対応付けを行うために，Sleep-CAM から得られた CAM ベクトルを教師信号として利用して提案モデルの学習を行った．さらに潜在表現ベクトルに特徴波に関する情報を偏りなく学習させるため，16 本のうち 15 本にはどの睡眠ステージの情報を保存するかをあらかじめ決定しておき，後述する誤差関数で制御する．具体的には，5 本の潜在表現ベクトルにステージ W 情報を保存するようにし，別の 5 本にステージ REM，残った 5 本にステージ NREM を割り当て，残った 1 本には役割を割り当てずに全体を調整させるようにした．

学習時はエンコーダから出力される潜在表現ベクトルを全て使って信号再現を行い，入力信号との誤差を計算する．一方オーグメンテーション実行時は，潜在表現ベクトルを一部変化させてデコードする．このとき特徴波情報が含まれないベクトルを，他の同ステージ EEG を変換したときの潜在表現ベクトルから特徴波情報が含まれない成分で上書きすることで，形状が変化した信号の生成を試みた．

理想的な学習ができた場合，例えば入力信号にステージ W 特徴波が含まれていたとき，潜在表現ベクトル中ステージ W 担当ベクトルのうち一部（または全て）が他のベクトルに比べて活性化することが予想される．したがって，このベクトルを残しつつ波形再現の補佐として別の信号の潜在表現ベクトルと合わせてデコードすることによって，特徴波を残したまま別の信号を生成することができると考えた．

4.1 オーグメンテーション方法

実際のオーグメンテーションは，次の 4 ステップで実行する．

（1） 1 チャンネル EEG をエンコードし，16 本の潜在表現ベクトルを得る．

（2） 16 本の潜在表現ベクトルから，「16 本それぞれの max(| 最大振幅 |) を降順に並べて」順位付けしたベクトルの上位 n 本を残しつつ，それ以外の (16-n) 本のベクトルを他の同ステージ EEG のものに置き換える．

（3） 16 本の潜在表現ベクトルを，入力信号と同じ形にデコードして出力する．

（4） 入力信号の高周波成分を出力信号に加える．

ここで順位付けの基準を max(| 最大振幅 |) にした理由は，活性の大きな潜在表現ベクトルに特徴波情報が含まれると予想されるからである．

元の信号に残すべき潜在表現ベクトルの数 n の値による精度変化を第 6 章で検証する．また，ステップ 2 で参照する信号の選び方について第 7 章で考察する．

4.2 モデル学習のための前準備

ここでは，次節 4.3 で述べる誤差を計算するために必要な教師データである CAM ベクトルを，Sleep-CAM [4] を用いて獲得する方法について説明する．提案モデルと実験に使用する睡眠ステージ判定モデルには 1 チャンネル EEG を入力するが，Sleep-CAM には 6 チャンネル EEG，2 チャンネル EOG，1 チャンネル EMG を入力して CAM ベクトルを獲得する．

そもそも Sleep-CAM は，生体信号に含まれる睡眠ステージ判定に寄与する部分を明らかにする機能があった．現に，同モデルが示す CAM 値が高い部分には，特徴波が含まれていることが多い．また，提案モデルのエンコード・デコードの過程では，信号に含まれる特徴波の情報を保持していてほしい．そこで，提案モデルの学習に使うデータを事前に学習済み Sleep-CAM に入力し，その CAM 値を計算して CAM ベクトルを計算しておく．CAM ベクトルを教師信号として潜在表現ベクトルが CAM ベクトルに似るように提案モデルを学習することで，特徴波情報を保持したエンコード・デコードが実現できると考えた．そして，これを計算する誤差関数を $latent_cos_loss$ として実装した．

CAM ベクトルの計算方法は以下の 3 ステップからなる．ここで 5 クラスの睡眠ステージを 3 クラスに統合した理由は，周波数特性の似たステージでグルーピングすることで，学習を効率的に行うためである．

（1） ステージ N1, N2, N3 をステージ NREM として統合し，睡眠ステージ判定を 3 ステージ分類問題として再定式化する．

（2） Sleep-CAM に 6 チャンネル EEG, 2 チャンネル EOG, 1 チャンネル EMG を入力し，得られたステージ毎の CAM 値の

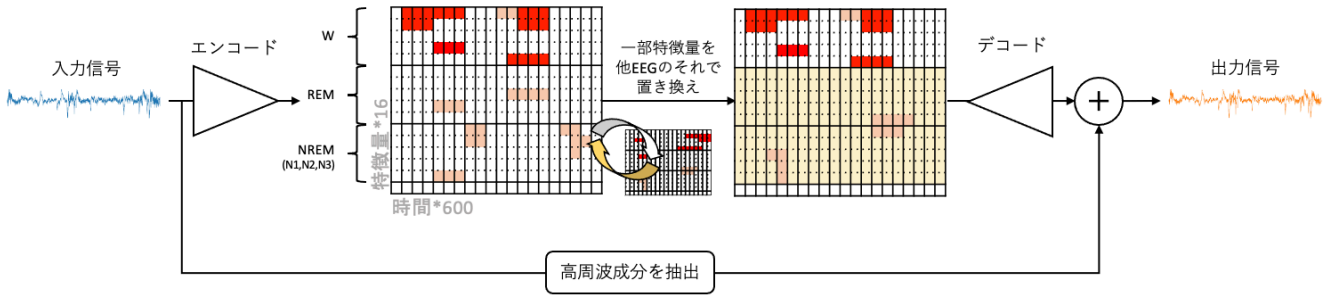


図 3　提案オーグメンテーション手法の概略図．図左に示す青色の入力信号をエンコードしたものを潜在表現ベクトルと呼称し，マトリクスで表現した．この潜在表現ベクトルに含まれる一部のベクトルを他のものに置き換えてデコードすることで，新たな信号を生成する．

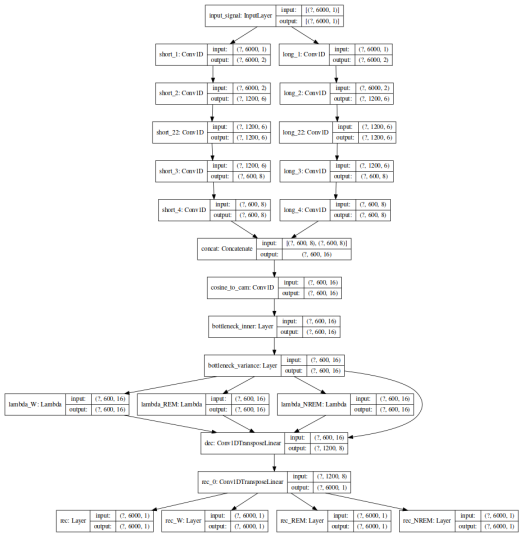


図 4　オーグメンテーションを実行するエンコーダ・デコーダモデル．1 秒と 0.1 秒の長さのフィルタで入力信号を畳み込み，エンコード・デコードの過程で 8 つの誤差を計算する．

大きさを降順に並べ替え，上位 5%の値を閾値として記録する．

（3） ステージ毎の CAM 値を参照し，閾値以上の値を持つインデックスに 1，そうでない所に 0 を持つような CAM 値と同型の CAM ベクトルを計算する．

4.3 誤差関数

提案モデルの学習に使う誤差関数の全体像を式 1 に示す．

$$\begin{aligned} loss &= w_{cos} \cdot latent_cos_loss + w_{red} \cdot reduce_reconly \\ &+ w_{inner} \cdot latent_inner_loss + w_{var} \cdot latent_var_loss \\ &+ w_{rec} \cdot RMSE_{rec} + w_W \cdot RMSE_W \\ &+ w_{REM} \cdot RMSE_{REM} + w_{NREM} \cdot RMSE_{NREM} \end{aligned} \quad (1)$$

ここで，w はそれぞれの誤差の影響を制御するハイパーパラメータであり，それらパラメータの値と誤差関数を追加した目的を表 2 に示す．

5 実験設定

実験は Python3.6.9, TensorFlow2.3.2, NVIDIA Tesla V100 上で行った．

5.1 データセット

本研究では 20-30 代の被験者の生体信号からなるデータセットを使用した．一般的に，若年の被験者から計測した生体信号は特徴波が比較的明確に発現しやすい [4] ため，提案手法のデータオーグメンテーションが有効に働きやすいと考えた．

本データセットは 49 夜分の生体信号データ（EEG6 チャンネル，EOG2 チャンネル，EMG1 チャンネル）を含み，判定技師によって 30 秒毎に睡眠ステージラベルが割り当てられている．また，サンプリングレートは 250Hz で，日本光電社 PSG-1100 を使用して計測された．

本研究では 6 チャンネルの EEG のうち 1 チャンネルのみを使用した．さらに前処理として，NS (Not Scored) ラベルがついた区間と実験に不適当な 1 データを取り除き，サンプリングレート 200Hz にダウンサンプリングした後，Z-score Normalization により正規化を行った．

5.2 実験手順

5.2.1 学習方法

実験では学習に使用する 1 チャンネル EEG のデータ数（以下，エポック数と表記）を変化させながらオーグメンテーションの効果を検証するため，学習用・検証用・テスト用データを用意する流れを図 5 に示す．

まず以下の 3 ステップによって，適切なエポック数に調整された学習用データを得る．

（1） 全データを Sleep-CAM に適用し CAM ベクトルを得ておく．

（2） 全データから 1 チャンネル EEG を取り出し，その順番をランダムに並べ替え，6:2:2 の割合で学習用 (仮)・検証用・テスト用に分割する．

（3） 学習用のサイズを $r \in (0, 1.0]$ 倍する．

（4） （3）で用意した 1 チャンネル EEG に対応した CAM ベクトルと睡眠ステージラベルを合わせ，最終的な学習用データとする．

次に，先の手順によって得られた学習用データの半数を生成データで置き換えるが，その置き換え方によって 3 パターンのデータを得た．

（i） 提案手法や FTSurrogate，Time reverse の効果を検

表 2　ハイパーパラメータと概要．ただし，L は潜在表現ベクトルを，C は CAM ベクトル，EEG^S_{in} は睡眠ステージ”S”($\in W, REM, NREM$) が割り当てられた入力信号，EEG^S_{out} は睡眠ステージ”S”担当潜在表現ベクトルから再構成した信号を表す．

誤差関数	重み	減らすもの	目的
$latent_cos_loss$	1	L と C のコサイン類似度の逆数	L 中に特徴波が含まれるようになること
$reduce_reconly$	0.1	睡眠ステージ担当がなされていない L 要素の値	睡眠ステージ担当がなされているベクトルの活性促進
$latent_inner_loss$	0.1	$L^{\{W,REM,NREM\}}$ 要素同士のコサイン類似度	それぞれに異なる（特徴波）情報を保存させること
$latent_var_loss$	2000	L 要素間の分散	提案モデルの推論に全てのベクトルを使わせること
$RMSE_{rec}$	100	RMSE(EEG_{in}, EEG_{out})	入出力を似通わせること
$RMSE_W$	0.2	RMSE(EEG^W_{in}, EEG^W_{out})	W 担当ベクトルが W の情報を保持すること
$RMSE_{REM}$	0.1	RMSE($EEG^{REM}_{in}, EEG^{REM}_{out}$)	REM 担当ベクトルが REM の情報を保持すること
$RMSE_{MREM}$	0.1	RMSE($EEG^{NREM}_{in}, EEG^{NREM}_{out}$)	NREM 担当ベクトルが NREM の情報を保持すること

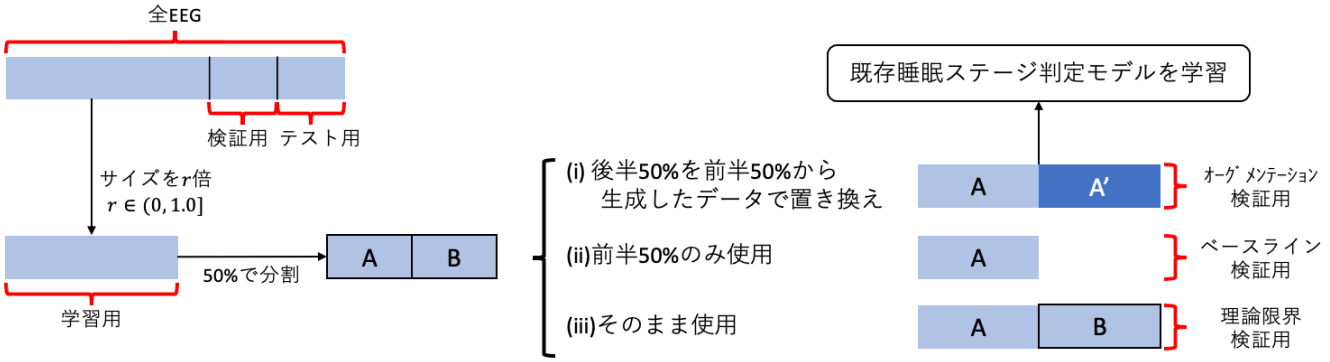


図 5　既存判定モデルの学習に使う学習用・検証用・テスト用データの獲得方法．

証するための学習用データ．

（ii）　ベースラインの測定を行うための学習用データ．

（iii）　上 2 つよりも多様なデータを使った，理論限界値を求めるための学習用データ．

また，学習の最適化アルゴリズムに Adam($\eta = 0.0005$) [11]，検証データの誤差が 3 回連続で改善しなくなるまで学習を継続した．

5.2.2　評 価 方 法

信号の波形を考慮した睡眠ステージ判定が行えること，主に畳み込み層で構成され軽量であることから，提案手法の効果検証をするために DeepSleepNet [2] を利用した．同モデルは 1 チャンネルの EEG から睡眠ステージ判定を行う．それぞれの層のパラメータは元論文を参照し，学習の最適化アルゴリズムに Adam($\eta = 0.0005$)，検証データの誤差が 3 回連続で改善しなくなるまで学習を継続した．

また，正しい結果を得るために 5 分割交差検証を行い，5 回の判定一致率の平均値を算出して記録した．

6　評 価 実 験

6.1　生成した信号の観察

元信号と，$n = 1, 2, 3, 4, 5, 6$ の場合における生成した信号を図 6 に示す．オーグメンテーションによって生成した信号は $n = 2$ あたりから明確な特徴波成分を持つようになった．また，$n = 3$ あたりから元信号に類似してくるのは，順位付けされた潜在表現ベクトルのうち上位 3 本の表現力が高く，それらのみで元信号の波形をほとんどを再現できてしまうからであると考えられる．よって，より脳波らしい信号を生成したい場合は n の値を大きく，より多様な信号を生成したい場合は n の値を小さく設定しオーグメンテーションを実行することで，状況に応じた使い分けが可能だといえる．

次に，FTSurrogate によって生成した信号を図 7 に示す．FTSurrogate は信号を周波数領域表現に変換した後，ランダムな値を加えることでその形を変形させている．つまり，信号に含まれる周波数の分布・比率は大きく変わらないため，睡眠ステージ判定に重要な周波数情報を保持したオーグメンテーション手法であるといえる．しかし，実際に生成された信号を見ると，判定に必要な特徴波が消失・全域化してしまっている．これでは同じ信号を技師が判定したとき，必ずしも生成元の信号と同じ睡眠ステージであると判定するとは限らないだろう．したがって，FTSurrogate は判定モデルの精度向上に一定の効果があるものの，その生成されたデータの質を担保することが難しいといえる．

また，Time reverse によって生成した信号を図 8 に示す．FTSurrogate は信号を時間軸方向に沿って逆転させることで新たな信号を生成するため，信号に含まれる周波数の分布・比率は変化しない．しかし，特徴波には左右非対称のものも存在する（K 複合等）ほか，特徴波の位置がずれることで，エポックに付与されるべき睡眠ステージが変わる場合がある．このオーグメンテーションによって自動睡眠ステージ判定手法の精度が向上する可能性はあるが，それが人の手によって行われたとき，必ずしも同じ睡眠ステージラベルを付与するとは限らない．そ

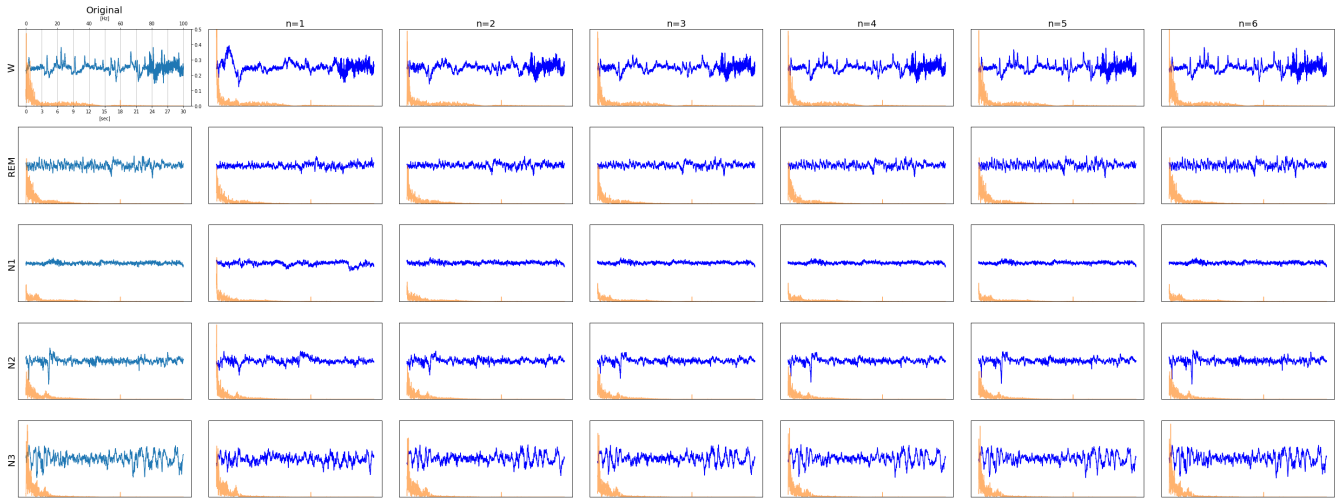


図 6　全ステージに渡る元信号（薄青）と提案オーグメンテーション手法によって生成された信号（紺）．橙色は周波数スペクトルを表す．

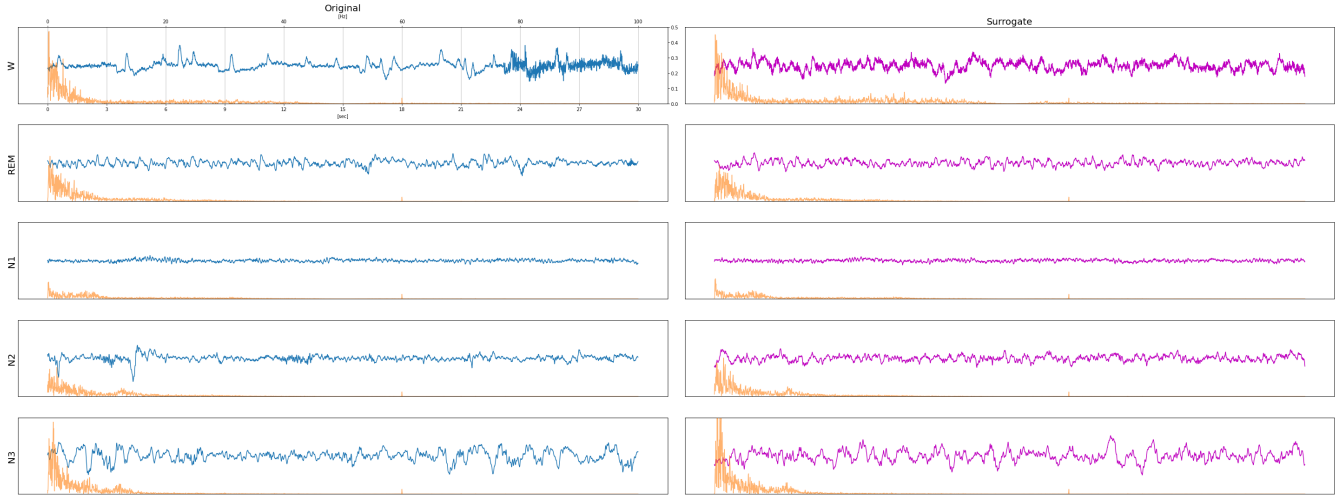


図 7　全ステージに渡る元信号（薄青）と FTSurrogate によって生成された信号（紺）．

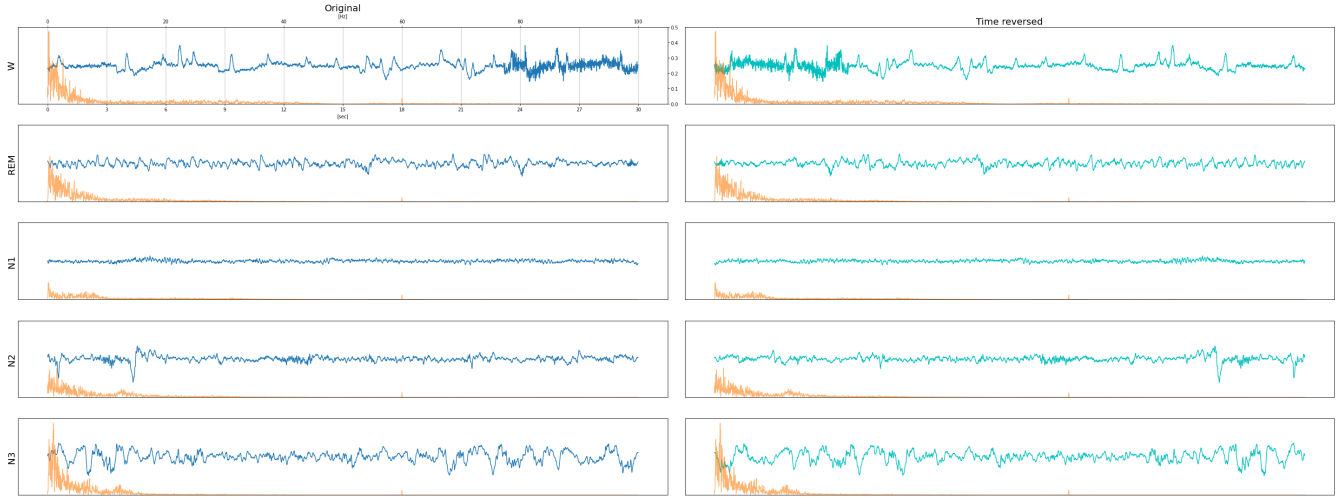


図 8　全ステージに渡る元信号（薄青）と Time reverse によって生成された信号（水色）．

のため，実際に臨床で運用することを考えると，Time reverse は最適なオーグメンテーション手法ではないと考えられる．

さらに，CAM ベクトルの値が大きいところ・小さいところにおける，元信号と生成した信号の類似度を表 3 に示す．CAM ベクトルの値が大きなところには特徴波が含まれることが多いため，その部分の類似度が大きくなることが望ましい．類似度はコサイン類似度 $\in [-1, 1]$ と RMSE $\in [0, \infty)$ によって測り，それぞれ 1 または 0 に近いほど元信号に類似しているといえる．この結果から，提案手法によって生成した信号は既存手法で生成したものに比べて，特徴波部分の波形を保持していることが定量的に示された．

表 3　元信号とオーグメンテーションによって生成した信号の類似度．CAM 値が大きなところで類似していることが望ましい．コサイン類似度は 1 に近いほど，RMSE は 0 に近いほど元信号に類似していることを表す．

元信号との比較対象	CAM 値	コサイン類似度	RMSE
提案手法 (n=2)	1	**0.7721**	**3.3637**
	0	0.7419	7.0660
FTSurrogate	1	-0.1248	7.7691
	0	-0.1009	18.3946
Time reverse	1	0.0118	7.3144
	0	-0.0003	17.5516

6.2　オーグメンテーション効果の検証

信号生成の過程で元の潜在表現ベクトルから上位 n 本のベクトルを残してオーグメンテーションを実行した．学習エポック数を変化させながら，$n = 1, 2, 3, 4, 5, 6$ の場合における提案手法と，オーグメンテーションを適用しなかったベースライン，理論限界の平均一致率の推移を図 9 に示す．提案手法はほとんどの場合においてベースラインの判定精度を凌駕し，特に $n = 5$ の判定精度向上が大きかった．

次に，提案手法 ($n = 5$) と FTSurrogate，Time reverse を重ねて表示した平均一致率の推移を図 10 に示す．判定精度の向上度合いは，学習エポック数が少ないと特に FTSurrogate，多いと提案手法が大きかった．

図 9 において学習エポック数を $r = 2^{-4}$ 倍したときに，提案オーグメンテーション手法による判定精度の改善割合が比較的大きかったことに着目した．ベースラインの判定による混合行列と，提案手法 ($n = 5$) を適用したときのベースラインからの変化を表す混合行列を図 11 に示す．図によると，睡眠ステージ W，REM，N1 の判定精度が向上したことが確認できた．そもそも睡眠ステージの分布には偏りがあり，ステージ N2 に至っては全体の約 40%を占める．これが原因して，一般的に睡眠ステージ判定モデルは入力信号の多くを N2 と判定してしまう傾向にある．しかし，特徴波を保持する提案手法によって判定モデルがステージ W や REM の特徴を効率的に学習したことにより，結果的にそれらステージの判定精度が向上したのだと考えられる．

6.3　FTSurrogate との併用

特徴波という時間的な情報に着目する提案手法と，信号の周波数情報に着目する FTSurrogate は，性質の違うオーグメンテーション手法であると考えられる．判定モデルの学習用データにそれぞれの手法によって生成した信号を半数づつ組み込むことで，さらなる精度向上ができないか調査した．

実験設定は同様に，提案手法と FTSurrogate を併用してオーグメンテーションを実行した．提案手法はハイパーパラメータ n の値を変化させるたびに異なる結果を得られるため，その一例として $n = 4$ の提案手法と FTSurrogate を統合して図 12 に示した．提案手法は学習エポック数が少ないときに効果が振るわなかったが，その弱点を FTSurrogate が補佐する形で安定した精度向上を記録した．しかし，学習エポック数が多くなったときの精度向上度合は提案手法と大きく変わらなかった．

7　考　　察

7.1　オーグメンテーション実行時に参照する信号による精度差の検証

提案手法はある EEG の潜在表現ベクトル上位数本を残しつつ，それ以外の成分を同ステージ別の EEG のものに置き換えることでオーグメンテーションを行なう．その際に参照する EEG との距離をステップ数と呼ぶことにすると，デフォルトではステップ数 1 の信号を生成している．ステップ数のみを変化させて提案手法を適用し，10 分割交差検証を行い一致率を表 4 に記録した．

参照する信号の選び方による一致率の差が有意かを確かめるため，有意水準 5%で両側検定の t 検定を行った．すると，$t(9) = 0.02, p = 0.98$ となり，それぞれの一致率の差に有意差は見られなかった．つまり，提案手法のハイパーパラメータ n_step の選び方によってオーグメンテーションの効果に差は現れないことが示された．

表 4　n_step を変化させて 10 分割交差検証を行ったときの一致率．

ステップ数	10 分割それぞれにおける一致率
$n_step = 1$	0.8395, 0.8454, 0.859, 0.8405, 0.8515, 0.839, 0.847, 0.8386, 0.8321, 0.8329
$n_step = 2$	0.8431, 0.8542, 0.8517, 0.8372, 0.8443, 0.8521, 0.8321 0.8235, 0.8452, 0.8412

7.2　提案手法の利点と欠点

図 10 を観察すると，学習エポック数が少ないときに FTSurrogate の効果が大きく，数が大きいときに提案手法の効果が大きいように見える．そもそも，睡眠ステージ判定に有用な情報は信号の周波数領域に多く存在すると言われている．学習エポック数が少ないときに FTSurrogate が効果的だった理由は，周波数領域の変換によって多様な信号が生成されたことで，睡眠ステージ判定モデルが不足している周波数情報を学習し，周波数分解能を得たからであると考えられる．一方，提案手法は元信号の特徴波に注目してオーグメンテーションを行うため，

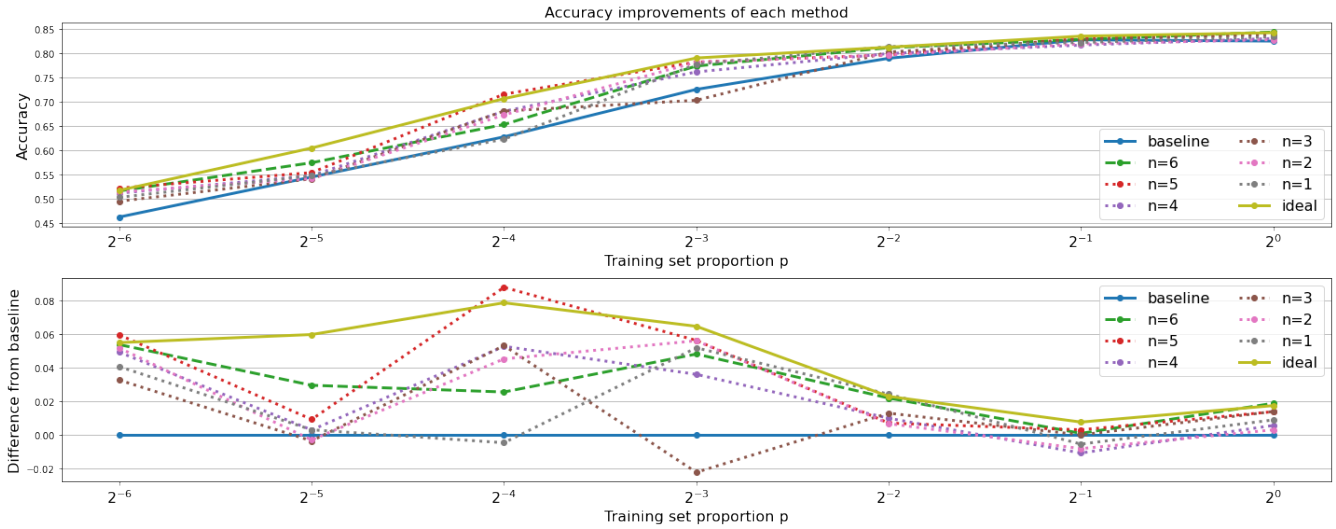


図 9　提案手法を適用した際の平均一致率の推移．（上）学習エポック数に対する判定一致率．（下）学習エポック数に対するベースラインの判定一致率との差．

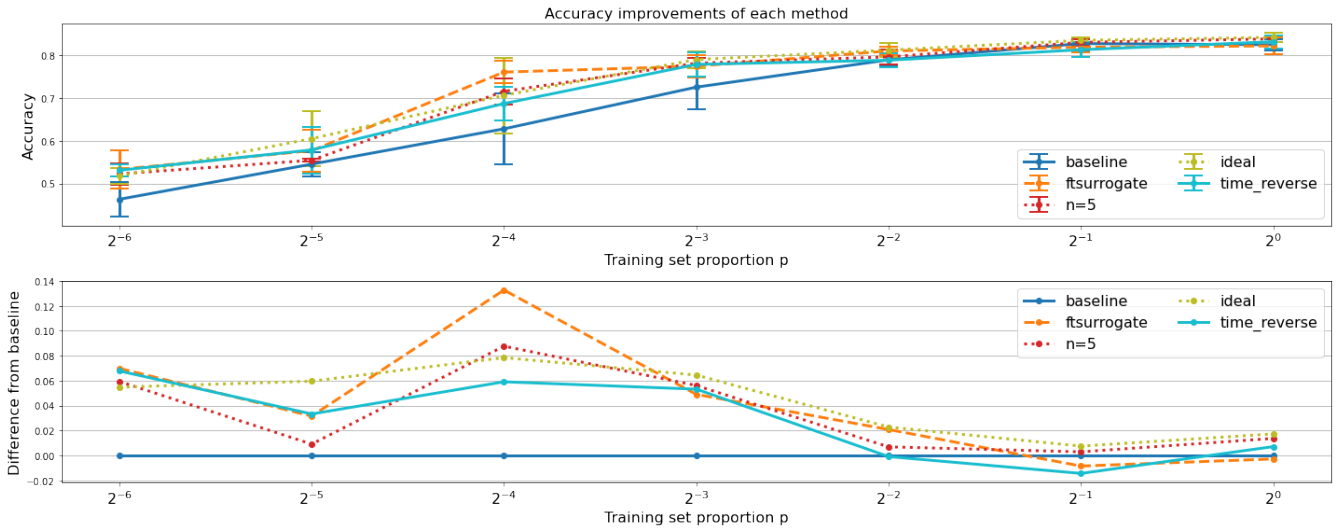


図 10　提案手法 ($n = 5$) または FTSurrogate を適用した際の平均一致率の推移．（上）学習エポック数に対する判定一致率．（下）学習エポック数に対するベースラインの判定一致率との差．

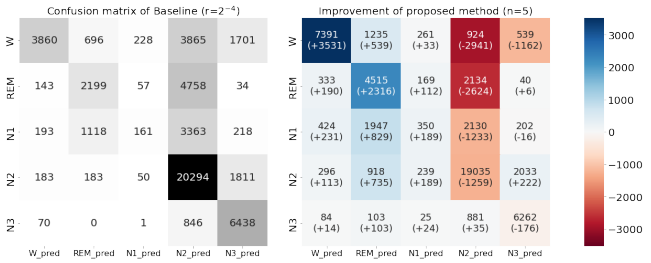


図 11　r=2^{-4} 時の判定結果．判定に使用したエポック数は同じ．（左）ベースラインの判定による混合行列．（右）提案手法 ($n = 5$) を適用したときの混合行列．括弧内は左の混合行列からの増減を表す．

こういった元となる信号が少ない状況下ではその効果を十分に発揮することはできない．しかし，FTSurrogate による判定精度向上度合いが提案手法のそれを大きく上回った $r = 2^{-4}$ の状況は，学習エポック数が 1966 件（≃ 睡眠 2 晩分）と非常に少ない．実際には，これほどデータが少ない状況で睡眠ステージ判定モデルを学習することは稀であると考えられる．

逆に言えば，提案手法が生成する信号には特徴波の波形が含まれることが多いことから，オーグメンテーション元の信号が十分にある場合には，安定した判定精度の向上が見込める．現に学習用データ数が多いときに提案手法が効果的だった理由は，学習 1 エポックあたりに使うデータに特徴波が多く含まれており，より質の高いデータで学習ができたからであると考えられる．さらに，提案手法はハイパーパラメータの調整によって元信号からの変化度合を調整することができ，用途に応じた信号を得ることができるという利点がある．本研究で示したように既存の睡眠ステージ判定モデルの判定精度を向上させる目的では，n の値を大きく設定した方が良い．しかし，例えば判

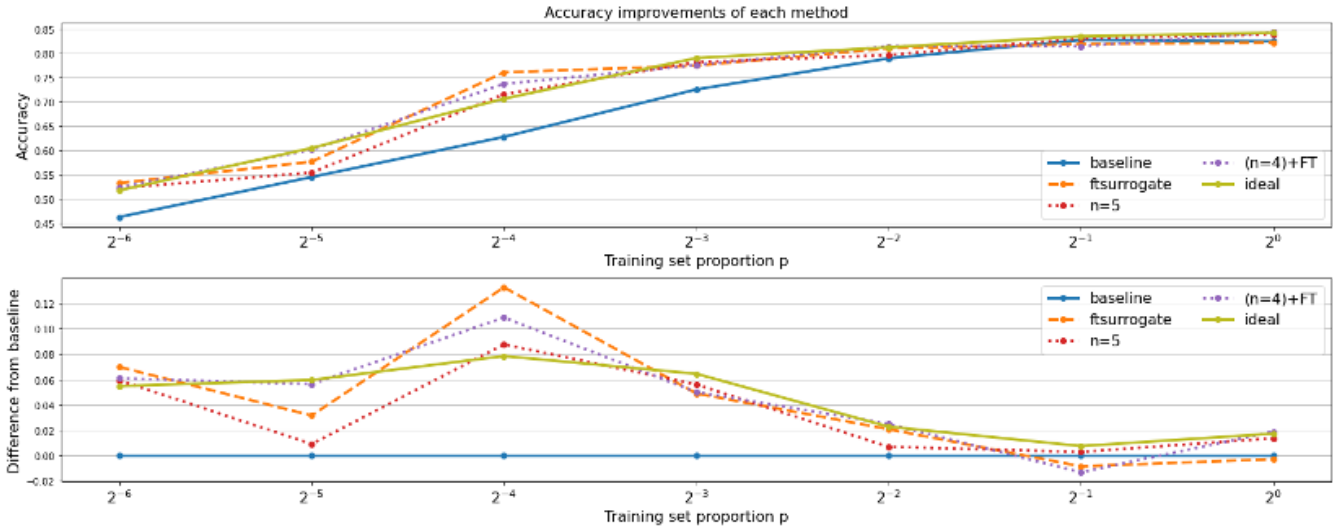


図 12　提案手法 ($n = 5$) に加えて，提案手法 ($n = 4$)+FTSurrogate を適用した際の平均一致率の推移．（上）学習エポック数に対する判定一致率．（下）学習エポック数に対するベースラインの判定一致率との差．

定技師の教育のため特徴波を含む多様な信号を用意したい場合には，n の値を小さく設定してオーグメンテーションを行うことで，入力信号とは異なる波形の信号を得ることができる．

また，Time reverse が特徴波を大きく崩していないにもかかわらず他の手法ほど判定精度を向上させなかった理由は，時間軸方向の反転によって特徴波の位置を大きく変えたことが判定に影響したためであると考えられる．時間軸方向に非対称な特徴波は K 複合や鋸歯状波程度であり，対称な特徴波の数の方が多いものの，睡眠ステージ判定には特徴波の位置情報も非常に重要であるという事実を裏付ける結果が得られた．

提案手法に残る問題点として，特徴波以外の部分においても生成される信号が元信号に似ていることが挙げられる．元信号と生成される信号の類似度は，CAM ベクトルの値が大きい部分で高く，そうでない部分は低い信号が生成されることが望ましい．しかし表 3 によると，それぞれのコサイン類似度は 0.7721, 0.7419 であることから，提案モデルは CAM ベクトルの値が小さい部分（=特徴波が存在しない部分) の波形も再現しようとしていることがわかる．つまり，現状では提案モデルが入力信号を潜在表現ベクトルにエンコードした段階で，順位付けされたベクトル上位数本の表現力が高く，それらが多くの情報を保持していると考えられる．潜在表現ベクトルに波形の情報が局所的に存在しないよう提案モデルの学習に使う誤差関数の設計を見直し，一見して元信号と異なると判断できるような信号を生成できるような手法を研究したい．

8　ま　と　め

本研究では，脳波に含まれる睡眠ステージ判定に有用な波形成分（特徴波）に着目しながら新たな信号を生成するオーグメンテーション方法を開発した．提案手法はハイパーパラメータ n の調整によって元信号からの変化度合を調整することができ，用途応じて様々な信号を生成することができる．学習エポック数を変化させながら提案手法を適用したところ，特に $n = 5$ の場合において安定した判定精度向上を記録した．

また，既存手法である FTSurrogate や Time reverse と比較して定量・定性評価を行った．それぞれの手法による判定精度改善度合の推移を観察したところ，学習に使用するエポック数が少ない場合には FTSurrogate がより効果的であった．しかしそのようなケースは非常に限定的であり，実際に睡眠ステージ判定モデルを学習するような状況では提案手法が安定した精度改善を達成した．

今後の展望として，潜在表現ベクトルそれぞれの表現力を均等に分散させる方法を見直したい．ハイパーパラメータ n（オーグメンテーション実行時に元の信号に残す潜在表現ベクトル数）を変えることで生成する信号の形状を変化させられることが提案手法の特徴の一つではあるが，波形情報の偏りが改善されれば最適な n を求めることが可能かもしれない．また，提案モデル学習時の誤差関数が最適化されていない可能性がある．誤差を構成するそれぞれの項の役割が重複・トレードオフの関係になっている部分があり，その係数を含めて最適な組み合わせを求めることができれば，より高速なオーグメンテーション実行が可能になると考えられる．

文　　献

[1] Berry RB, Brooks R, Gamaldo CE, Harding SM, et al. for the American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.5. Darien, IL:American Academy of Sleep Medicine, 2018.

[2] Supratak, Akara, et al. ”DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG.” IEEE Transactions on Neural Systems and Rehabilitation Engineering 25.11 (2017): 1998-2008.

[3] Danker‐Hopfe, Heidi, et al. ”Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders.” Journal of sleep research 13.1

(2004): 63-69.
[4] Horie, Kazumasa, et al. "Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability." Scientific reports 12.1 (2022): 12799.
[5] Bird, Jordan J., et al. "Synthetic biological signals machine-generated by GPT-2 improve the classification of EEG and EMG through data augmentation." IEEE Robotics and Automation Letters 6.2 (2021): 3498-3504.
[6] Rommel, Cédric, et al. "Data augmentation for learning predictive models on EEG: a systematic comparison." Journal of Neural Engineering 19.6 (2022): 066020.
[7] Schwabedal, Justus TC, et al. "Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates." arXiv preprint arXiv:1806.08675 (2018).
[8] 日本睡眠学会. 改訂版 臨床睡眠検査マニュアル. (株) ライフサイエンス, 2016.
[9] Phan, Huy, et al. "XSleepNet: Multi-view sequential model for automatic sleep staging." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.9 (2021): 5903-5915.
[10] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
[11] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

睡眠ステージ自動判定における 睡眠時無呼吸症候群の重症度を踏まえた深層学習処理の切替機構の開発

川畑　竜平† 　堀江　和正†† 　北川　博之†††

† 筑波大学情報科学類　〒 305-8574 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305-8577 茨城県つくば市天王台 1-1-1
††† 筑波大学国際統合睡眠医科学研究機構　〒 305-8575 茨城県つくば市天王台 1-1-1
E-mail: †kawabata@kde.cs.tsukuba.ac.jp, ††{horie,kitagawa}@cs.tsukuba.ac.jp

あらまし　深層学習による睡眠ステージ（睡眠状態，レム睡眠など）自動判定において，患者の属性によって判定精度が大きく変動してしまう課題がある．問題を解決する方法として患者の属性ごとに特化したモデルを作成することが考えられるが，一モデルあたりの学習サンプルが不足して精度に悪影響を与えることがある．そこで本研究では 1 つのモデル内で擬似的に複数の特化モデルを再現することを目指し，選択的不感化と呼ばれる浅層ニューラルネットの手法をもとにした内部処理の切替機構を開発した．年齢と睡眠時無呼吸症候群の重症度を切替機構の切り替えパラメータとして用い評価実験を行い，両属性に対して提案手法が有効であることを確認した．

キーワード　生体信号処理，医療・ヘルスケア，時系列データ処理，睡眠ステージ判定

1 序　　論

ヒトの睡眠は覚醒状態 (Wake) とレム睡眠 (REM), ノンレム睡眠 (N1, N2, N3) の 5 つの睡眠ステージからなっており，脳波や筋電位，眼電位などの生体信号の違いをもとに定義されている [1]．これらの睡眠ステージは睡眠時無呼吸症候群をはじめとした睡眠に関する病気の診断や治療に役立てられている．その一方で，睡眠ステージの検査には専門家が目視で生体信号を確認する必要があり，1 人あたり約 2 時間かかるなど，多くの時間と労力を必要としている．この問題を解決すべく，自動睡眠ステージ判定モデルがさまざま提案されており，深層学習を用いたモデルでは若年層では技師の精度を超えるものも存在する [2].

ところが，これらの自動判定システムは臨床の現場においては広く普及していない．その理由の一つに，患者の属性によって判定精度が大きく変動してしまうことが挙げられる．例えば，睡眠時無呼吸症候群の重症度 (AHI) が高い患者や高齢者においては，図 1 に見られるように睡眠ステージの分布が健常者や成人と異なっており，また判定の際の根拠として用いる「特徴波」も高周波成分の割合が現象する・振幅が減少するといったように特徴が変化することが知られている [3]．このように，成人と高齢者・健常者と睡眠障害の患者など，患者間の属性が異なる場合には，睡眠ステージ判定の精度が低くなる傾向が確認されている．臨床の現場では，高齢者や睡眠障害を持つ患者が多いことが予想されるため，こうした課題は臨床での実用化に向けて重要な課題である．

患者間の属性差を低減する方法としては，患者の属性ごとに複数の特化したモデルを作成することが考えられる．複数のモデルを用いることで，属性ごとの特徴・ステージ分布などを捉えることができ，精度の向上が期待できる．しかし，複数のモデルを用いる方法は，各モデルごとに異なる学習サンプルが必要となるため，特に特化モデルを必要とするグループにおいては学習サンプル数が不足することが懸念される．また，特化モデルを複数個運用することは，モデルの管理コスト・臨床での計算機環境を鑑みると現実的でない．

そこで本研究では，深層学習モデルの内部処理を属性によって切り替える機構を開発することで，1 つのモデルで「擬似的に」複数の特化モデルを再現することを考えた．患者の属性に関わらず共通している特徴は共通して学習し，属性ごとに特化した特徴は分離して学習することができれば，学習サンプル数不足の問題を解決しつつ，属性差の影響を最小限に抑えることができるだろう．

1 つのモデル内で複数のモデルを使い分ける手法を考えるうえで，筆者らはドロップアウト [4] に着目した．ドロップアウトはある 1 つのモデルの一部をランダムでマスクすることによって，オーバーフィッティングを防ぐ手法である．また，ドロップアウトは擬似的なバギングであることがこれまでの研究で知られている [5]．この特徴は，本研究で目指している，「複数のモデルを 1 つのモデルで再現する」というコンセプトと一致している．

一方で，ドロップアウトはマスクされる場所がランダムであるため，「特化モデルを使い分ける」という操作にはならない．そこで，固定することは簡単なので，「マスクを固定すれば」特化モデルを使い分けることができるのではないかと考えた．属性ごとに固定されたマスクを用いるにあたり，それらのマスクは属性が近いほど共通する要素が多く，属性が離れるほど共通する要素が少なくなるようなものであることが望ましい．このようなマスクを生成する手法として，選択的不感化ニューラルネットのパターンコーディング [6] のアイデアを用いた．選択的不感化ニューラルネットのパターンコーディングを用いてマ

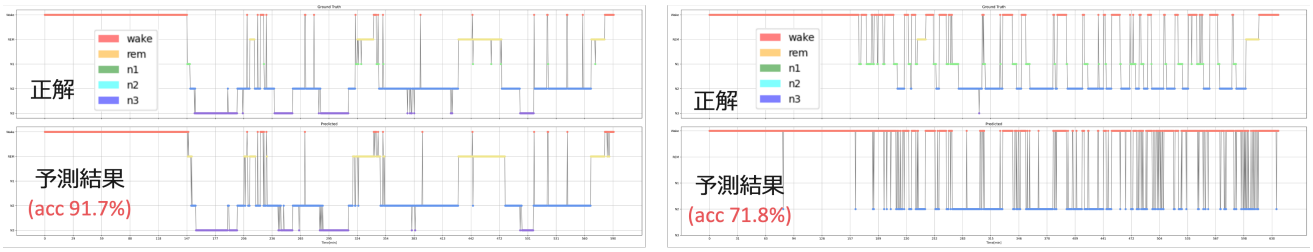


図 1　左が成人 (22 歳)・右が高齢者 (71 歳) の睡眠ステージ遷移図 (ヒプノグラム) の例.
成人に比べて高齢者の睡眠ステージは Wake が多く, 結果として予測精度が低くなることが確認できる.

スクを生成することで，属性間で共通する特徴量と，属性別で使用する特徴量を表現することができると考えた．

こうした背景を踏まえ，本研究では属性ごとにマスクを切り替える機構を開発し，提案手法の有効性を検証した．本論文では，患者の属性差のうち，特に年齢と睡眠時無呼吸症候群の重症度 (AHI) に対して提案手法を適用した結果について報告する．

2　睡眠ステージスコアリング

手動による睡眠ステージのスコアリングは次の手順で実行される．

PSG の測定: 患者に計測装置を取り付け，6 つの脳波 (EEG) 2 つの眼電図 (EOG)、1 つの筋電図 (EMG) を記録する．

信号の前処理: 測定された PSG 信号からノイズを除去し，信号を 30 秒間隔（1 エポックと呼ぶ）に分割する．技師は各エポックに 1 つの睡眠ステージラベルを割り当てる．

特徴波の検出: 技師は睡眠ステージを判断するための根拠となる「特徴波」を見つけ出す．

睡眠ステージの割り当て: 技師は米国睡眠学会 (AASM) のルール [1] に従って，睡眠ステージを割り当てる．このルールは主に以下の原則に基づいている．

- エポックの特に前半に特徴波が存在する場合，その関連する睡眠ステージを割り当てる
- 特徴波がない場合，先行するエポックのステージと特徴波を考慮してステージを割り当てる
- 複数の睡眠段階を示す特徴波が含まれる場合，もっとも多い割合を占めるステージを割り当てる

特徴波の検出と睡眠ステージの割り当ては，一人の患者に対して一晩で約二時間もの時間を要する作業であり，大きな労力と時間が必要である．これらの作業を自動化することで，医師や判定技師の負担を軽減できるため，多くの研究者が自動判定モデルの開発に取り組んでいる．

3　先 行 研 究

3.1　選択的不感化

選択的不感化 [7] [8] は浅層ニューラルネット向けの特徴抽出手法の一つであり，パターンコーディングと不感化操作の 2 つからなる．パターンコーディングはある特定のスカラー値を 1, -1 からなる 2 値パターンに変換する手法で，スカラー値の変化

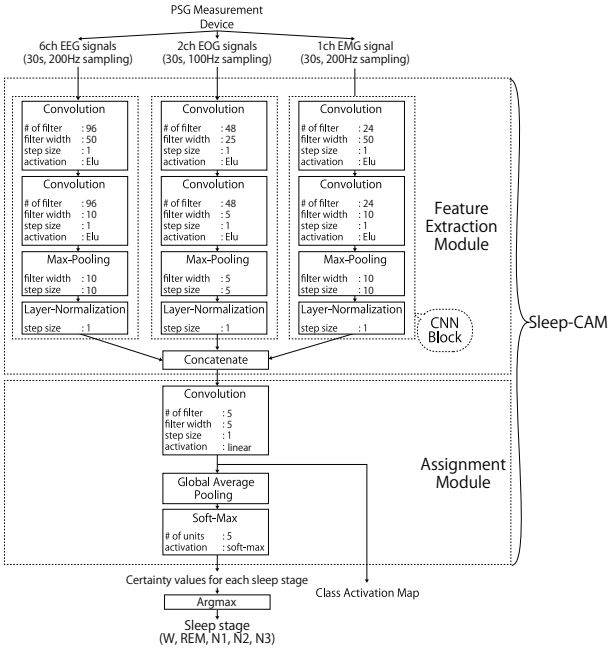


図 2　Sleep-CAM モデル概要図　([2] より)

に合わせて徐々にパターンが変化するようなコードパターンを作成する．選択的不感化は，2 つのコードパターンがあるとき，一方の値に合わせてもう一方をマスク (element を 0 にする) 操作である．

不感化の手順の説明のため，具体的に 2 種類の n 次元の 2 値 (±1) パターン $S = (s_1, \ldots, s_n)$ をパターン $C = (c_1, \ldots, c_n)$ で不感化する場合について考える．また S と C の各要素にはあらかじめランダムに決められた 1 対 1 の対応付けがなされているとする．s_i と c_j に対応があるとき， 不感化後の素子 s'_i は，

$$s'_i = \frac{1 + c_j}{2} s_i$$

と表される．$c_j = +1$ のとき s_i の値はそのまま保持され, $c_j = -1$ のとき s_i の値は中立値 (=0) に変換される．

年齢や睡眠時無呼吸症候群といったスカラー値の属性に対してパターンコーディングを行い，深層学習モデルの特徴量に対して選択的不感化することで，属性ごとに特化した特徴量を使い分けることを目指す．

3.2　SleepCAM

Sleep-CAM [2] は，脳波を用いた睡眠ステージ判定を行う深

層学習モデルである．Sleep-CAM のモデル概要は図 2 からわかるように，大きく特徴抽出部とステージ割り当て部から構成され，それぞれ睡眠ステージ判定における特徴波の検出とステージの割り当てを実現する．

特徴抽出部は 2 層の畳み込み層からならなる CNN ブロック 3 つからなり，それぞれのブロックが脳波・眼電位・筋電位の特徴を抽出する．

Sleep-CAM は説明可能性を獲得しながら，86.9%の精度で睡眠ステージ判定を行うことができる．これは睡眠の専門家の評価者間信頼性 (82.0%) [9] よりも高く，臨床での実用化に適している．本研究では Sleep-CAM をベースモデルとして使用する．

4 提案手法

切り替え機構を備えた提案モデルを図 3 に示す．本モデルは大きく特徴抽出部と切り替え部，ステージ判定部から構成される．特徴抽出部とステージ判定部はベースモデルの Sleep-CAM から脳波の特徴抽出部とステージ判定部を抜き出したものである．特徴抽出部とステージ判定部の間に切り替え部を挟むことで，年齢や睡眠時無呼吸症候群の重症度に応じて処理の切り替えを行うことができ，患者の属性を考慮した睡眠ステージ判定を実現する．

4.1 特徴抽出層

特徴抽出部はベースモデルの Sleep-CAM から脳波の特徴抽出部を抜き出したものである．主に 2 層の畳み込み層と Global Average Pooling 層からなる．

入力としてサンプリング周波数 200 Hz 30 秒の 6 チャンネル脳波を用い，長さ 96 のベクトルを抽出された特徴量として出力する．

4.2 切り替え部

提案手法では，コードパターンを用いて年齢をベクトルに変換して入力する．コードパターンとは図 4 のように，1 と 0 からなる行列である．

コードパターンの生成方法を説明する．まず，かけ合わせる特徴量の次元と同じ大きさのベクトルを用意する．このベクトルは 4 : 1 の割合で 1 と 0 が含まれるようにランダムに生成する．コードパターンに新たな行を追加する際に，前の行のベクトルから 1 の要素と 0 の要素 1 組をランダムに選択し入れ替えたものを新たな行とする．この操作を対応する世代の数だけ繰り返すことでコードパターンを生成する．

実験ではあらかじめ説明した方法でコードパターンを生成し，ある属性が入力として与えられたときテーブルルックアップを行い入力された属性に対応するコードパターンの行を選択し，選択されたコードパターンの行を切り替え部に入力する．

そのように入力されたコードパターンを特徴抽出部で得られた特徴量とかけ合わせることで，属性に応じて特徴量の一部を 0 にすることができる．コードパターンには年齢により変化する要素と年齢を超えて共有される要素があり，前者により年齢ごとの特徴を捉えることができ，後者により年齢を超えて共通する特徴を捉えることができる．

4.3 ステージ判定部

切り替え部で属性によって特徴量をマスクした後，特徴量を用いてステージ判定を行う．ステージ判定部は全結合層からなり，長さ 96 の特徴量を入力し 5 つの出力から softmax を用いて睡眠ステージの確率を出力する．

4.4 学習方法

学習率は，epoch が 6 回以下のときは 5.0×10^{-4}, epoch が 14 回以下のときは 1.0×10^{-4}, epoch が 15 回以上のときは 5.0×10^{-5}, 最適化手法は Adam(η = 0.005) を使用し，バッチサイズ 32 で 100 回の学習を行った．検証用データの誤差が 3 回連続で下がらなかったときに早期終了させ，37 回で学習が完了した．

5 評価実験

実験には米国国立衛生研究所が公開する CFS データセットを使用する [10] [11]. このデータセットは 735 人の実験参加者から睡眠中の脳波，筋電位，眼電位等を記録したものであり，図 5 にあるように，6 歳から 88 歳まで様々な属性を持つ実験参加者が含まれている．

本実験では EEG 信号 (6 チャンネル，サンプリング周波数 125Hz) を使用する．Sleep-CAM の EEG 信号は 200Hz でサンプリングされたものを使用するため，本実験ではサンプリング周波数を 200Hz にリサンプリングした後，30 秒ごとに正規化を行っている．

5.1 属性として年齢を用いた場合

5.1.1 単体モデル間での比較

まず，提案手法である切り替え機構の有効性を検証するために，単体モデルを複数用意し，それぞれのモデルに年齢を入力する方法を変えた場合の精度を比較した．

今回検証に用いたそれぞれのモデルについて説明する．既存手法とは先行研究である Sleep-CAM から眼球電位と表面筋電位に関する特徴抽出部を抜き出し，脳波の特徴抽出部のみを用いたモデルである．既存手法+年齢とは，既存手法に新たに年齢の入力を追加したモデルである．年齢は式 1 を用いて 0 から 1 に正規化したものを入力として与えている．DropOut は 50% の確率でランダムにマスクを行うため，コードパターンと生き残る素子数が異なり，そのままでは公平に比較できない．このため，この DropOut モデルではモデル構造を既存手法と同様にし，コードパターンと 1 と 0 の割合が等しく含まれたランダムなマスクを入力することで，DropOut の操作を行っている．

$$normalized_age = \frac{\min(age, 80)}{80} \tag{1}$$

表 1 に各モデルの評価結果を示す．各 fold を平均した結果を見ると，提案手法が一番精度が高いモデルとなっていることが確認できる．

年齢入力の有効性の観点から本実験の結果を比較すると，提案手法は既存手法と比べて平均 Accuracy が 0.44 % 高くなって

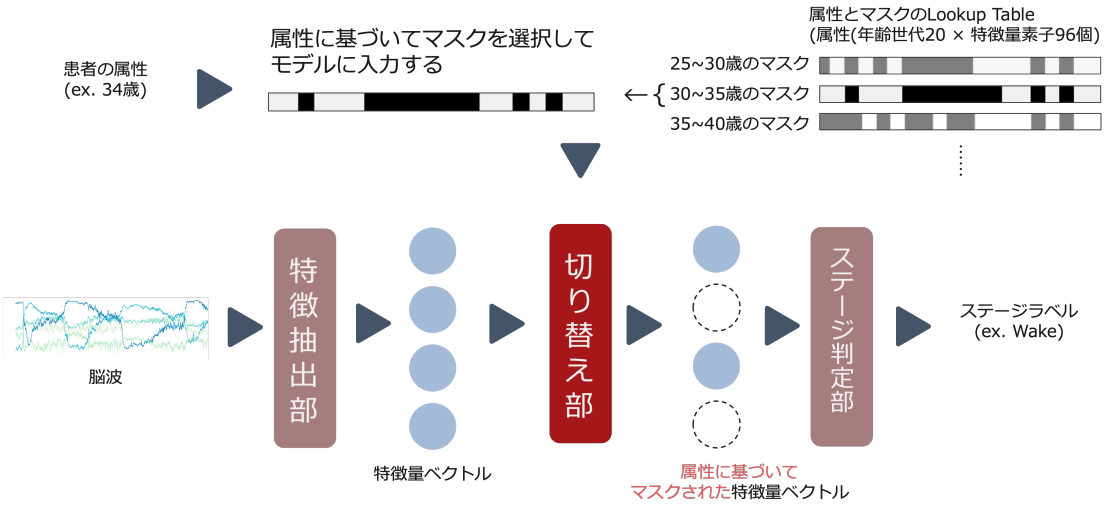


図 3　提案手法の概要

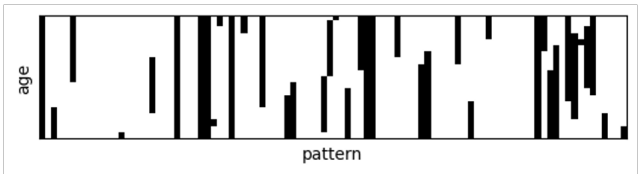


図 4　コードパターンの例

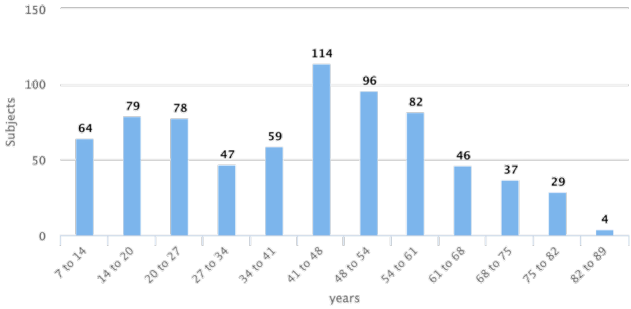


図 5　CFS データセットの年齢分布

fold	提案手法	既存手法	既存手法+年齢	DropOut
0	**0.8672**	0.8606	0.8632	0.8521
1	**0.8638**	0.8594	0.8504	0.8490
2	**0.8665**	0.8560	0.8572	0.8193
3	**0.8686**	0.8663	0.8629	0.8562
4	0.8644	**0.8661**	0.8619	0.8537
Avg.	**0.8661**	0.8617	0.8591	0.8461

表 1　年齢の入力を変えたモデル間での評価結果 (値は Accuracy)

いるが，既存手法+年齢は既存手法と比べて 0.26 % 低くなっている．このことから，単純に年齢を入力として加えるだけでは精度が向上しないことがわかる．年齢をもとに処理を切り替えている提案手法では，年齢を入力しない既存手法よりも精度が向上していることがわかる．

また提案手法は DropOut よりも平均 Accuracy が 2.00 % 高くなっている．これは，提案手法は DropOut とは異なる要因で精度が向上していることを示唆している．

世代	提案手法	複数モデル
20 歳未満	**0.7919**	0.7464
20 歳以上 60 歳未満	0.7975	**0.8367**
60 歳以上	**0.8258**	0.7340
世代間の最大精度差	**0.0339**	0.1027

表 2　年齢属性別グループ間での評価結果 (値は Accuracy)

fold	提案手法 (AHI)	既存手法	既存手法+AHI
0	0.8744	0.8736	**0.8762**
1	**0.8828**	0.8799	0.8816
2	0.8738	0.8716	**0.8796**
3	**0.8755**	0.8711	0.8699
4	**0.8827**	0.8820	0.8494
Avg.	**0.8778**	0.8756	0.8713

表 3　AHI の入力を変えたモデル間での評価結果 (値は Accuracy)

5.1.2　複数特化モデルとの属性別精度の比較

表 2 に年齢属性別グループ間での評価結果を示す．提案手法は複数モデルと比べて，20 歳未満と 60 歳以上の世代において精度が高くなっていることがわかる．また，世代間の最大精度差が提案手法の方が小さいことから，提案手法は年齢属性別グループ間での精度差が小さいことがわかる．

5.2　属性として睡眠時無呼吸症候群の重症度 (AHI) を用いた場合

5.2.1　単体モデル間での比較

年齢において提案手法が有効であることが確認できたため，次に睡眠時無呼吸症候群の重症度 (AHI) を用いた場合について検証を行った．また既存手法+AHI では，年齢のときと同様に，式 2 を用いて 0 から 1 に正規化したものを入力として与えている．

$$normalized_ahi = \frac{\min(ahi, 50)}{50} \tag{2}$$

表 3 に各モデルの評価結果を示す．各 fold を平均した結果を見ると，提案手法が一番精度が高いモデルとなっていることが

重症度	提案手法	複数モデル
重症者 (AHI > 30)	0.7358	**0.7373**
中症者 (30 ≧ AHI > 15)	0.8653	**0.8767**
軽症者 (15 ≧ AHI)	0.7659	**0.7899**
重症度別の精度差	**0.1394**	0.1295

表 4　睡眠時無呼吸症候群の属性別グループ間での評価結果 (値は Accuracy)

確認できる．

AHI 入力の有効性の観点から本実験の結果を比較すると，提案手法は既存手法と比べて平均 Accuracy が 0.22 % 高くなっているが，既存手法+AHI は既存手法と比べて 0.43 % 低くなっている．このことから，単純に AHI を入力として加えるだけでは精度が向上しないことがわかる．AHI をもとに処理を切り替えている提案手法では，AHI を入力しない既存手法よりも精度が向上していることがわかる．

5.2.2　複数特化モデルとの属性別精度の比較

表 4 に睡眠時無呼吸症候群の属性別グループ間での評価結果を示す．提案手法は軽症者と重症者において精度が高くなっていることがわかる．また，重症度別の精度差が提案手法の方が小さいことから，提案手法は睡眠時無呼吸症候群の重症度別グループ間での精度差が小さいことがわかる．

6　考　　察

6.1　伸び悩んだ属性について

表 2 の結果を見ると，提案手法の young での Accuracy が他の手法に比べて伸びていない．young の範囲が 20 歳から 60 歳までと広いが，child や old に比べて睡眠傾向のバラツキが小さい．その一方で，提案手法のパターンコーディングでは 40 歳分=8bin=16 箇所も反転しており，睡眠傾向の違いと比較し反転が多いため，精度が伸び悩んでいると考えられる．

今回の実験で評価に用いた young の 10 人の実験参加者について，1 人ずつの Accuracy を評価したところ表 5 のようになっている．表 5 から分かるように，実験参加者によっては Accuracy が 0.9 を超えているものもある一方で，一部の実験参加者では 0.5 を下回っているものもある．これは young の睡眠傾向のバラツキが大きいことを示している．

提案手法のみの Accuracy を見ると，それぞれ 79.19%, 79.75%, 82.58%となっており, 他の手法と比較して世代間のバラツキが軽減されていることが分かる．また提案手法では old の Accuracy が young と比較して高くなっており，これは他の手法では見られない傾向である．本実験結果から，提案手法は既存手法に比べて Accuracy が下がってしまったものの，世代間のバラツキを軽減するという当初の目的が達成できたと考えられる．

また child・old の結果や，交差検証での結果を踏まえると，学習サンプルに含まれるデータの傾向と評価に用いるデータの傾向が近いほど，精度が高くなることが示唆される．これより，より大規模なデータセットで学習した場合，より多様な睡眠傾向が含まれた学習サンプルである可能性が高いため，提案手法の精度は既存手法に比べさらに性能が向上すると予想される．

表 5　young 10 人それぞれでの評価

	Accuracy
患者 0	0.7672
患者 1	0.8603
患者 2	0.4741
患者 3	0.7105
患者 4	0.8737
患者 5	0.6309
患者 6	0.8093
患者 7	0.9239
患者 8	0.7919
患者 9	0.6628

6.2　パターンコーディング自体の有効性

属性ごとに固定されたマスクを用いるにあたり，それらのマスクは属性が近いほど共通する要素が多く，属性が離れるほど共通する要素が少なくなるようなものであることが望ましい．一方で，そのようなマスクを生成する手法としてパターンコーディング以外の手法も考えられる．

例えば，図 6 のように，属性ごとに異なるマスクを生成する手法が考えられる．図 6 は，行が増えるごとに値が 0 になる箇所が始まるインデックスを 1 ずつずらしていくことで生成されたマスクである．このようなマスクでも，属性が近いほど共通する要素が多く，属性が離れるほど共通する要素が少なくなっている．

先程のシンプルなマスクと，提案手法で用いたコードパターンを用いたマスクを用いた場合で，それぞれモデルを訓練しその精度を比較した．その結果をそれぞれ表 6, 7 に示す．

表 6 から，年齢を切り替え属性としてマスクの生成方法を変えたモデル間での評価結果を見ると，コードパターンを用いた

fold	Simple マスク	コードパターン
0	**0.8811**	0.8768
1	**0.8851**	0.8784
2	**0.8846**	0.8676
3	**0.8787**	0.8753
4	**0.8868**	0.8757
Avg.	**0.8833**	0.8747

表 6　年齢を切り替え属性としてマスクの生成方法を変えたモデル間での評価結果 (値は Accuracy)

fold	Simple マスク	コードパターン
0	0.8761	**0.8762**
1	**0.8821**	0.8816
2	**0.8796**	0.8796
3	**0.8729**	0.8699
4	**0.8845**	0.8494
Avg.	**0.8790**	0.8713

表 7　AHI を切り替え属性としてマスクの生成方法を変えたモデル間での評価結果 (値は Accuracy)

重症度	Simple マスク	コードパターン
重症者 (AHI > 30)	0.6993	**0.7155**
中症者 (30 ≧ AHI > 15)	**0.8427**	0.8405
軽症者 (15 ≧ AHI)	0.7466	**0.7540**
重症度別の精度差	0.1434	**0.1250**

表 8　AHI を切り替え属性としてマスクの生成方法を変えたモデル間での評価結果 (値は Accuracy)

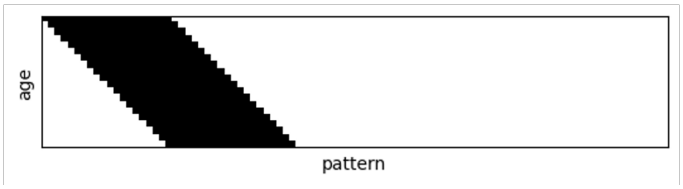


図 6　徐々に属性が変化するようなマスクの例

マスクを用いたモデルがもっとも高い精度を示していることがわかる．また AHI について適用した場合の結果を見ると，表 7 から，シンプルなマスクを用いたモデルがもっとも高い精度を示していることがわかる．

一方で，表 8 から，AHI を切り替え属性としてマスクの生成方法を変えたモデル間での評価結果を見ると，シンプルなマスクを利用した場合と比較し，コードパターンを利用した場合の方が精度が高いことがわかる．

以上の結果から，シンプルなマスクを用いる方法では，交差検証の結果としてはコードパターンを用いる方法に比べて精度が高いものの，コードパターンよりも汎化能力が低く，属性間の精度差が大きくなってしまうことが確認できる．これには

6.3 訓練時間

臨床での実用化を目指すにあたり，訓練にかかるコストは重要な要素の一つである．訓練時間について比較するため，年齢を切り替え属性として用いた際の各モデルの 5 交差検証にかかった時間をそれぞれ表 9 に示す．特化モデルについては child・young・old それぞれの訓練時間の合計を示している．表 9 より，提案手法は既存手法と比較して計算時間が増えているものの，特化モデルを作成する場合と比較しても差が 12 分 24 秒と，実用上問題になるほどではないと考えられる．

また各モデルでの 1 サンプルあたりの訓練時間をそれぞれ表 10 に示す．特化モデルは既存手法とモデル構造は同じであるため省略している．表 10 より，提案手法は既存手法と比較して 1 サンプルあたりの訓練時間が 30μs 減少しており，総訓練時間の増加は学習 epoch 数が増えた影響によるものであると考えられる．

表 9　各モデルでの総訓練時間

モデル	訓練時間
既存手法	1 時間 9 分 50 秒
既存手法+年齢	57 分 30 秒
特化モデル	1 時間 24 分 1 秒
提案手法	1 時間 32 分 25 秒

表 10　各モデルでの 1 サンプルあたりの訓練時間

モデル	1 サンプルあたりの訓練時間
既存手法	825 μs
既存手法+年齢	822 μs
提案手法	795 μs

7 結　論

本研究では，患者間の属性差を考慮した深層学習処理の切り替え機構を開発した．本研究で開発した切り替え機構を年齢・睡眠時無呼吸症候群の重症度に適用し，提案手法の有効性を評価した．提案手法を用いることで，既存手法と比較し精度が向上することが交差検証によって確認された．また年齢を用いた場合においては，提案手法が患者間の属性差を低減することに成功した．

提案手法は調整できるパラメータが複数あるため，最適なパラメータを探索することは今後の課題である．また特徴量ごとに利用頻度が大きく異なるため，その補正の必要性も検討している．さらに，提案手法は既存手法よりも個々人の傾向に合わせた特徴を抽出することができるため，より大きなデータセットを用いた検証を今後行うことも考えている．

謝　辞

本研究は JSPS 科研費・若手 23K16988 の助成を受けたものです。

文　献

[1] Berry, R. et al. “The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications” Version 2.5. In American Academy for Sleep Medicine (2018).
[2] Horie, K., Ota, L., Miyamoto, R., Abe, T., Suzuki, Y., Kawana, F., Kokubo, T., Yanagisawa, M., Kitagawa, H., “Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability” Sci Rep 12, 12799 (2022).
[3] 山﨑まどか， 松浦 雅人，“成人および高齢者の脳波” 臨床神経生理学 42 巻， 6 号， p. 387-392 (2014).
[4] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. “Improving neural networks by preventing co-adaptation of feature detectors” http://arxiv.org/abs/1207.0580
[5] D. Warde-Farley, I. J. Goodfellow, A. Courville, and Y. Bengio, “An empirical analysis of dropout in piecewise linear networks” Proceedings of the International Conference on Learning Representations (ICLR), 2014.
[6] 野中和明，田中文英，森田昌彦，“階層型ニューラルネットの 2 変数関数近似能力の比較” 電子情報通信学会論文誌 (D), J94-D, 12, 2114-2125, (2011).
[7] 森田昌彦，松沢浩平，諸上茂光，“非単調神経素子の選択的不感化を用いた文脈依存的連想モデル”, 電子情報通信学会論文誌 (D-II), J85-D-II, 1602-1612, (2002).
[8] 宮澤泰弘，末光厚夫，森田昌彦，“選択的不感化理論に基づく海馬ニューロン活動のモデル化”, 日本神経回路学会誌， 14, 3-12, (2007).
[9] Danker-Hopfe, H. et al. “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM”, J. Sleep Res. 18(1), 74–84 (2009).
[10] Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. “The National Sleep Research Resource: towards a sleep data commons” J Am Med Inform Assoc. 2018 Oct

1;25(10):1351-1358. doi: 10.1093/jamia/ocy064. PMID: 29860441; PMCID: PMC6188513.

[11] Redline S, Tishler PV, Tosteson TD, Williamson J, Kump K, Browner I, Ferrette V, Krejci P. “The familial aggregation of obstructive sleep apnea” Am J Respir Crit Care Med. 1995 Mar;151(3 Pt 1):682-7. doi: 10.1164/ajrccm/151.3_Pt_1.682. PMID: 7881656.

DEIM Forum 2024

気象データを統合的に用いたクラウドソーシングに基づく頭痛発生傾向の検知

飯田　静空†　　若宮　翔子†　　清水　伸幸††　　藤田　澄男††　　荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
†† LINE ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3
E-mail: †{iida.shizuku.iq0,wakamiya,aramaki}@is.naist.jp, ††{nobushim,sufujita}@lycorp.co.jp

あらまし　気象の影響を受けて体に痛みが生じる現象は天気痛と呼ばれており，最も多い症状は頭痛である．日本で天気痛の潜在患者は 1000 万人以上存在するため，頭痛が発生しやすいタイミングを事前に把握することは，効果的な先制医療に繋がる．一方で，気象データから頭痛発生傾向を検知するアプリケーションは，主に気圧の特徴量のみに依存しており，多様な気象データを用いる点で改善の余地が残されている．さらに，気象と頭痛発生の関係性を示す研究結果には相反が見られるため，気象の変化による頭痛発生の病態メカニズムは未だ不明な点が多い．本研究では，クラウドソーシングの回答に基づいた頭痛発生の指標である「頭痛率」を定義し，気圧，気温，湿度などの気象データを統合的に用いて頭痛発生傾向を検知する．実験の結果，気象データを統合的に用いることで頭痛発生傾向の検知性能を改善できること，気象状況が 1 日単位で大きく変化する時期は，気象データを用いた頭痛発生傾向の検知性能が向上することが分かった．

キーワード　医療・ヘルスケア，気象，機械学習，クラウドソーシング，ソーシャルセンシング，ソーシャルセンサ

1　はじめに

近年，頭痛は世界的な健康課題として大きく注目されている．2019 年の世界における頭痛の推定症例数は，15 億例以上に上り，1990 年から 10%以上の増加が確認されている [8]．日本では，頭痛の潜在患者は約 4000 万人存在すると推定されており [13]，この事実は，全人口の約 3 人に 1 人が頭痛を患っていることを表す．頭痛は緊急性が高い社会問題となっており [15]，それを解決するために，複雑化した頭痛発生のメカニズムの早期解明が求められている．特に，気象の影響を受けて体に痛みが生じる現象は天気痛と呼ばれており，最も多い症状は頭痛である [19]．日本の天気痛の潜在患者は，1000 万人以上存在することが示唆されており [5]，気象と頭痛発生の関係性に対する理解は，頭痛治療の革新的な進歩に繋がる．

気象要因の頭痛には，心理的要素が関係することが示唆されており [14]，頭痛発生のタイミングを自身で把握することで自己効力感を高める予防治療も近年注目されている [1,12]．そのため，気象データから頭痛発生のタイミングを検知し，頭痛がいつ起こりやすいかを自身で把握するためのアプリケーションがいくつか存在する．例えば，株式会社ベルシステム 24 が提供している頭痛ーる [1]がある．このシステムは 6 時間毎に，気圧低下量の年間平均値を算出し，頭痛が発生しやすい時間帯をユーザに事前通知する．他には，株式会社ウェザーニューズが提供している天気痛予報 [2]がある．これは気圧下降，微気圧変動，大気潮汐の 3 要因を基に，3 時間おきの頭痛の警戒情報を 4 段階の指標で発信している [14]．このように，頭痛発生傾向を検知するアプリケーションでは，気象データが用いられている．しかし，これらは主に気圧の特徴量のみに依存しており，多様な気象データを用いる点で改善の余地が残されている．主に気圧が用いられる理由は，気圧を感知するセンサーが内耳に存在し，頭痛が発生しやすい人ほど内耳が敏感であるという報告 [2] があるためと考えられる．一方で，気温や湿度も頭痛発生に関係すること [11,17] が示唆されており，これらを利用することで頭痛発生傾向の検知性能を改善することが期待される．

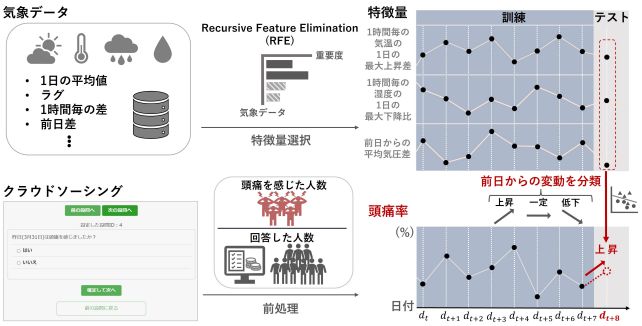


図 1　本研究の概要

本研究では，気象データを統合的に用いて頭痛発生傾向を検知する．図 1 に，本研究の概要を示す．具体的には，頭痛発生の指標としてクラウドソーシングを通して得られた回答から

1：株式会社ベルシステム 24 頭痛ーる
https://zutool.jp/
2：株式会社ウェザーニューズ 天気痛予報
https://weathernews.jp/s/pain/

「頭痛率」を定義し，この指標の前日からの変動を「上昇」,「一定」,「低下」の 3 クラスに分類する．このとき，気圧，気温，湿度のデータを統合的に用いることで，頭痛発生傾向の検知性能を改善できるか調査する．

2 先行研究

天気痛に関する最初の研究として，Hollander らは，気圧低下と湿度上昇の組み合わせが痛みを増強させることを示した [4]．この研究以降，特殊な人工環境装置が必要であることが原因で，十分な基礎研究が行われていない [14]．そのため，基礎研究の実験的エビデンスはないが，気象と頭痛発生の関係性を調査した研究が多く存在する．しかし，これらの研究結果には相反が見られるため，気象の変化による頭痛発生の病態メカニズムは未だ不明な点が多い．例えば，Li らの報告では，暖かい時期において相対湿度が高いほど頭痛発生が多い関係性がみられた [7] が，Yilmaz らの報告では，相対湿度と救急外来を受診した 1 日の頭痛患者数の間に負の相関がみられた [18]．本研究では，クラウドソーシングを用いて毎日 300 人を対象に頭痛発生のデータを収集し，気象と頭痛発生の関係性を分析する．これまで，頭痛発生の調査にクラウドソーシングを用いた研究は少なく，新たなソースに基づく分析は，相反する知見を統一するための新たな観点を提供すると考えている．

頭痛がいつ発生するかを把握する重要性とは裏腹に，気象データを用いて頭痛発生傾向を検知する可能性について調査した研究は，我々の知る限りほとんどない．勝木らは，頭痛ーるのユーザが記録した頭痛回数のデータを用いて，気象データから頭痛回数に回帰する統計モデルと深層学習ベースのモデルを構築した [6]．この研究は，気象データを用いて頭痛発生傾向を検知する可能性について評価した唯一の研究であるが，気象データの多様な組み合わせに応じた性能の違いを評価していない．さらに，結果のエラー分析に関する言及がないため，どのような場合に気象データが頭痛発生傾向を検知可能か詳細は不明である．本研究では，気圧，気温，湿度の特徴量の様々な組み合わせの利用とエラー分析を通して，気象データを用いた頭痛発生傾向の検知可能性を詳細に分析する．

3 データセット

3.1 頭痛率

本研究では，気象データを用いて頭痛発生傾向を検知し，気象と頭痛発生の関係性について分析する．そのために，人々の報告に基づく「頭痛率」指標を導入する．我々は，この指標を，「ある地域で 1 日における頭痛を感じた人の割合（%）」と定義する．

Yahoo!クラウドソーシング [3] を利用し，2022 年 7 月 2 日から 2023 年 10 月 1 日まで 457 日連続で，人々の頭痛についてのアンケートを行った．対象地域は「東京都」，1 回あたりのタスク数は 300 件，1 タスクあたりの謝礼は 5 円と設定した．質問した項目は以下の 3 項目である．

- お住まいの市区町村を教えてください．
  （回答）東京都の 62 市区町村から選択式
- 昨日（昨日の日付）は主に都内のどこにいましたか？
  （回答）東京都の 62 市区町村と「都内にはいなかった」からなる選択式
- 昨日（昨日の日付）は頭痛を感じましたか？
  （回答）「はい」か「いいえ」の選択式

1 日単位で収集したデータに対し，「昨日は主に都内のどこにいましたか？」という質問で「都内にはいなかった」と回答したユーザをそれぞれ除外した．その後，「昨日（昨日の日付）は「頭痛」を感じましたか？」という質問に「はい」と答えたユーザの割合を算出した．最終的に，2022 年 7 月 1 日から 2023 年 9 月 30 日における東京都の日々の頭痛に関するデータセットを作成した．1 日あたりに除外されたユーザ数の平均は 5.3，中央値 5.0 であった．2022 年 7 月 1 日から 2023 年 9 月 30 日の頭痛率の記述統計量は，平均値 10.8，標準偏差 2.1，最大値 17.9，中央値 10.5，最小値 5.8 であった．

3.2 気象データ

気象庁が提供している過去の気象データ検索 [4] を用いて気象データを集めた．頭痛率のデータは東京都在住者を対象としたため，東京都の「東京」地点のデータを取得した．気象データの特徴量に過去のデータも考慮するため，2022 年 6 月 24 日から 2023 年 9 月 30 日の期間で取得した．図 2 に収集した頭痛率，平均気温，平均湿度，平均現地気圧の日別推移を示す．

4 頭痛発生傾向の検知

本研究の目的は，気圧のみを用いるよりも気象データを統合的に用いた方が，頭痛発生傾向を高性能に検知できるかを検証することである．そのために，頭痛率の前日からの変動傾向を分類する問題に取り組む．

4.1 問題設定

この実験では，1 日の頭痛率の前日比を「上昇」,「一定」,「低下」のいずれかに分類する 3 クラス分類問題に取り組む．これは，前日の頭痛率との相対的な関係の検知に，気象データが寄与するか調査することを意味しており，本研究の目的を検証するアプローチとして妥当であると考えられる．

頭痛率の日毎のデータに対し前日比をとる前処理を行い，事前に決めた基準をもとにラベルを付与した．図 3 は，頭痛率の前日比のヒストグラムを表し，表 1 は頭痛率の前日比の分類基準とそのデータ数の内訳を表す．分類モデルを構築するために，データセットは 2022 年 7 月 1 日から 12 月 31 日を学習用，2023 年 1 月 1 日から 3 月 31 日を検証用，2023 年 4 月 1 日から 9 月 30 日をテスト用に分割した．学習データの偏りをでき

3：Yahoo!クラウドソーシング
https://crowdsourcing.yahoo.co.jp/

4：気象庁 過去の気象データ検索
https://www.data.jma.go.jp/obd/stats/etrn/index.php

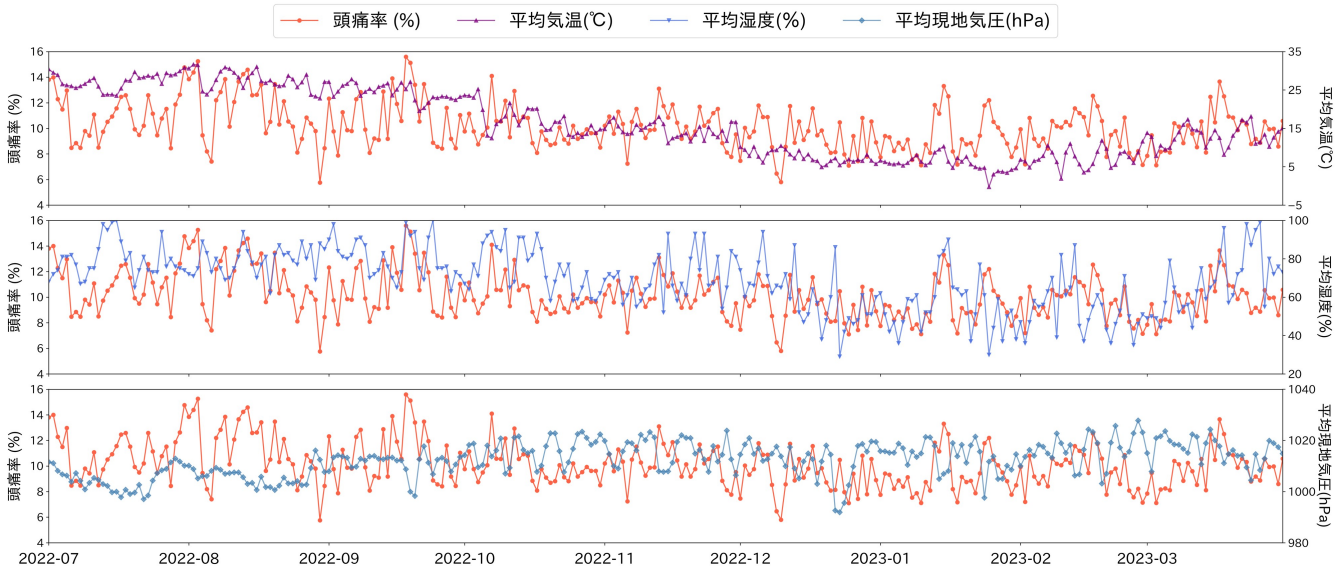


図 2　2022 年 7 月 1 日から 2023 年 3 月 31 日における頭痛率，平均気温，平均湿度，平均現地気圧の日別推移.

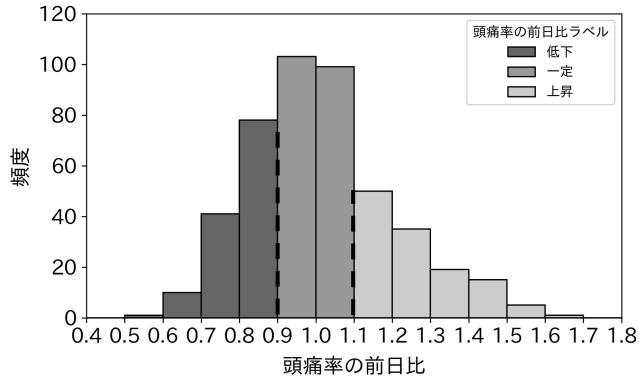


図 3　頭痛率の前日比のヒストグラム

表 1　頭痛率の前日比の分類基準とラベル数の内訳

頭痛率の前日比	ラベル	学習	検証	テスト
$1.1 < x$	上昇	50	23	52
$0.9 \leqq x \leqq 1.1$	一定	75	44	83
$x < 0.9$	低下	59	23	48

る限り少なくするために，ラベルの基準は，頭痛率の前日比 x が $1.1 < x$ を「上昇」，$0.9 \leqq x \leqq 1.1$ を「一定」，$x < 0.9$ を「低下」と設定した.

4.2　説明変数

本研究では，様々な気象データの組み合わせで構築した分類モデルの性能を比較する．説明変数には，分類モデルに必ず用いる固定変数と任意で用いる可変変数の 2 種類を用いた．さらに，可変変数は気温と湿度から成る気象特徴量と気圧特徴量の 2 種類を用いた.

4.2.1　固定変数

固定変数はラベルのダミー変数のラグであり，以下の 6 種類を用いた．これらを用いた理由は，過去の自己回帰的な特徴量は，頭痛率の前日からの変動傾向の分類に寄与すると考えられるためである．特に，分類対象日の直近の特徴量は分類への寄与が大きいと仮定し，ラグ 1 からラグ 3 までを対象とした.

- 上昇ラグ 1（1 日前が「上昇」の場合は 1，それ以外は 0）
- 上昇ラグ 2（2 日前が「上昇」の場合は 1，それ以外は 0）
- 上昇ラグ 3（3 日前が「上昇」の場合は 1，それ以外は 0）
- 低下ラグ 1（1 日前が「低下」の場合は 1，それ以外は 0）
- 低下ラグ 2（2 日前が「低下」の場合は 1，それ以外は 0）
- 低下ラグ 3（3 日前が「低下」の場合は 1，それ以外は 0）

4.2.2　可変変数

気象特徴量には，以下の 16 種類を用いた．気温と湿度は過去の情報や 1 日単位，前日からの変化などを特徴量の候補として挙げた.

- 1 日の平均気温
- 1 日の平均気温ラグ 1（1 日前の 1 日の平均気温）
- 1 日の平均気温ラグ 2（2 日前の 1 日の平均気温）
- 前日からの平均気温差
- 1 時間毎の気温の 1 日の最大値と最小値の差
- 1 時間毎の気温の 1 日の最大下降差
- 1 時間毎の気温の 1 日の最大上昇差
- 1 日の平均湿度
- 1 日の平均湿度ラグ 1（1 日前の 1 日の平均湿度）
- 1 日の平均湿度ラグ 2（2 日前の 1 日の平均湿度）
- 前日からの平均湿度比
- 1 時間毎の湿度の 1 日の最大値と最小値の比
- 1 時間毎の湿度の 1 日の最大下降比
- 1 時間毎の湿度の 1 日の最大上昇比

気圧特徴量には，以下の 9 種類を用いた．先行研究では，頭痛発生傾向の検知に大気潮汐 [14] や直近 1 週間の気圧パター

ン [6] が寄与することが示唆されている．そのため，大気潮汐と長期パターンを，気象特徴量の粒度に追加した．

- 1 日の平均気圧
- 1 日の平均気圧ラグ 1（1 日前の 1 日の平均気圧）
- 1 日の平均気圧ラグ 2（2 日前の 1 日の平均気圧）
- 前日からの平均気圧差
- 1 時間毎の気圧の 1 日の最大値と最小値の差
- 1 時間毎の気圧の 1 日の最大下降差
- 1 時間毎の気圧の 1 日の最大上昇差
- 1 日の 9 時と 15 時の気圧差（大気潮汐）
- 直近 1 週間の平均気圧の平均値に対する 1 日の平均気圧差（長期パターン）

4.3 実験設定

ベースラインには，学習データの多数派のラベルのみを分類結果とする Majority baseline を用いた．分類モデルにはロジスティック回帰 (LR)，決定木 (DT)，Support Vector Machine (SVM) を用いた．実装には，Python の scikit-learn のライブラリを用いた．LR は多項ロジスティック回帰で構築し，solver には lbfgs, 正則化には L2 正則化を指定した．DT の構築にはジニ不純度を用い，葉ノードに必要な最小サンプル数の min_samples_leaf を指定した．SVM は One-vs-Rest で構築し，カーネルには RBF，正則化には L2 正則化を指定した．さらに，不均衡データに対処するために，全てのモデルでクラス重み付けを指定した．これは以下の式に従って，分類の重みを各クラスのデータ数の分布に応じた割合で調整する方法である．

$$\mathrm{class}_{\mathrm{weight}} = \frac{\mathrm{n}_{\mathrm{samples}}}{\mathrm{n}_{\mathrm{classes}} \cdot \mathrm{n}_{\mathrm{samples}_{\mathrm{class}}}} \tag{1}$$

ここで，$\mathrm{n}_{\mathrm{samples}}$ はデータの総数，$\mathrm{n}_{\mathrm{classes}}$ はクラスの総数，$\mathrm{n}_{\mathrm{samples}_{\mathrm{class}}}$ は各クラスのデータの総数を表す．

4.2 節で挙げた全ての説明変数の組み合わせを比較することが理想的である．しかし，変数の数の増加とともに組み合わせの数は指数的に増加するため，現実的な時間内でモデルを比較することは不可能となる．そのため，モデルに用いる説明変数を効率的に削減するために，Recursive Feature Elimination (RFE) を用いた．これは，初期の特徴量の集合に対してモデルを学習し，推定された重要度を基に特徴量を再帰的に削除する手法である．可変変数の候補が 15 種類以上のモデルを比較する場合に，RFE を用いて候補を 14 種類にまで削除した．その後，候補の可変変数の全ての組み合わせを比較した．RFE の各イテレーションで削除する可変変数は 1 とし，全てのモデルでクラス重み付けを指定した．LR は L2 正則化パラメータ C = 1.0, DT は min_samples_leaf = 1 を指定した．SVM は RBF カーネルを用いた場合に，重要度を算出することができないため，線形 SVM で L2 正則化パラメータ C = 1.0 を指定した．

モデルの学習と評価には，Walk-Forward Validation を用いた．図 4 に，Walk-Forward Validation のイメージを示す．これは，1 時点毎に検証・テストに用いたデータを学習データに加えていき，次の時点のデータを検証・テストデータに用いる方法である．例えば，図 4 の Set_1 では，d_{t+2} までを学習データに用い，d_{t+3} 時点のラベルのみを分類する．次の Set_2 では，d_{t+3} までを学習データに用い，d_{t+4} 時点のみのラベルを分類する．今回は 1 セットあたり学習データの 1 時点先のみを分類していき，最終的に全セットの分類結果を総合して評価した．評価指標には，Accuracy と Macro-F1 を用いた．ハイパーパラメータの探索にはグリッドサーチを行い，検証用データの Macro-F1 の値が最も大きいパラメータを採用した．ハイパーパラメータの候補値は，LR で L2 正則化パラメータ C を {0.1 から 2.0 の範囲で 0.1 刻みの値 }，DT で min_samples_leaf を {1 から 30 の範囲で 1 刻みの値 }, SVM で L2 正則化パラメータ C を $\{0.1, 1, 10, 100\}$, RBF カーネルのパラメータ γ を $\{1, 5, 10\}$ とした．最終的なテストデータを用いた評価には，2022 年 7 月 1 日から 2023 年 3 月 31 日までを学習データに用い，Walk-Forward Validation で評価した．

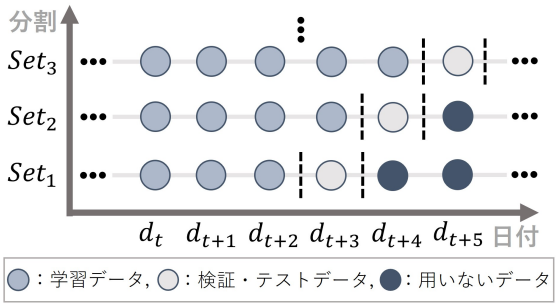


図 4　Walk-Forward Validation のイメージ

4.4 結　　果

表 2 に，頭痛率の前日比のラベル分類結果を示す．$*_{label}$ は固定変数のみで構築したモデル，$*_{weather}$ は気象特徴量のみで構築したモデル，$*_{pressure}$ は気圧特徴量のみで構築したモデル，$*_{all}$ は全ての特徴量で構築したモデルを表している．各モデルで考慮する説明変数の全組み合わせ中，最も Macro-F1 の大きいモデルの性能を示している．まず，Majority baseline を除く全てのモデルにおいて，Accuracy と Macro-F1 の間に大きな差が見られないため，モデルが不均衡データの分類に対処できていることが分かる．次に，3 つのモデルで，$*_{label}$，$*_{weather}$ もしくは $*_{pressure}$，$*_{all}$ の順に Accuracy と Macro-F1 が大きいことが分かる．全ての $*_{label}$ は Majority baseline 以下の分類性能だが，気象特徴量もしくは気圧特徴量を用いることで Majority baseline よりも分類性能が向上している．さらに，LR, DT, SVM の順に全体的な分類性能が良いことが分かる．特に，SVM と DT の分類性能に比べて，LR は格段に性能が良くなっている．各モデルにおいて，$*_{all}$ のモデルは最も分類性能が良く，$*_{all}$ の中で LR_{all} は Accuracy 0.53, Macro-F1 0.53 で最も性能が良い結果となった．気象特徴量か気圧特徴量を用いたモデルは，Majority baseline よりも分類性能は改善されたが，最高性能のモデルでも約 5 割を正確に分類可能である結果にとどまった．LR_{all} の可変変数の組み合わせは，1 時間毎の気温の 1 日の最大上昇差, 1 時間毎の湿度の 1 日の最大下降比，前日からの平均気圧差，1 時間毎の気圧の 1 日の最大下降

表 2　頭痛率の前日比のラベル分類結果．テストデータの評価を示す．各モデルが考慮する説明変数の全組み合わせ中，最も Macro-F1 の大きいモデルの性能を示す．

モデル	Accuracy	Macro-F1
Majority baseline	0.45	0.21
LR_{label}	0.46	0.45
$\mathrm{LR}_{weather}$	0.51	0.51
$\mathrm{LR}_{pressure}$	0.51	0.52
LR_{all}	**0.53**	**0.53**
DT_{label}	0.38	0.38
$\mathrm{DT}_{weather}$	0.50	0.50
$\mathrm{DT}_{pressure}$	0.48	0.48
DT_{all}	0.51	0.51
SVM_{label}	0.36	0.34
$\mathrm{SVM}_{weather}$	0.46	0.45
$\mathrm{SVM}_{pressure}$	0.48	0.46
SVM_{all}	0.49	0.46

差, 大気潮汐, 長期パターンの 6 種類であり，ハイパーパラメータの値は $C = 0.3$ であった．1 時間毎の気圧の 1 日の最大下降差, 大気潮汐, 長期パターンの気圧特徴量が重要であった本研究の結果は，先行研究の結果 [6, 14] と一致した．

5　考　　察

頭痛率の前日比ラベルの分類結果から，気象データを用いた頭痛発生傾向の検知可能性や気象と頭痛発生の関係性について分析する．

5.1　エラー分析

図 5 に，テストデータの正解ラベル（□）と LR_{all} の分類ラベル（◇）の推移を示す．上図は 2023 年 4 月 1 日から 6 月 30 日，下図は 2023 年 7 月 1 日から 9 月 30 日の推移を表している．これらの総合期間 (4 月から 9 月) の LR_{all} の Accuracy は 0.53, Macro-F1 は 0.53 であったが，時期で分けた場合，上図の方が下図よりも正しく分類できた日が多いことが分かる．実際に，上図の期間 (4 月から 6 月) の Accuracy は 0.60, Macro-F1 は 0.61 であるのに対し，下図の期間 (7 月から 9 月) の Accuracy は 0.46, Macro-F1 は 0.46 であった．これらの結果は，春先から梅雨の時期にかけて，気象データを用いることで頭痛発生傾向の検知が比較的可能になることを示している．

さらに，4 月から 6 月と 7 月から 9 月の期間を，細かな時系列の粒度で比較する．まず，4 月から 6 月の方が 7 月から 9 月よりも連続で正解している箇所が多い．例えば，4 月から 6 月では最長で 8 日間（5 月 19 日から 5 月 26 日）連続で正解することができている．次いで，連続 5 日間（4 月 6 日から 4 月 10 日，5 月 28 日から 6 月 1 日），連続 4 日間（4 月 1 日から 4 月 4 日，4 月 24 日から 4 月 27 日，5 月 4 日から 5 月 7 日，6 月 3 から 6 月 6 日）と続き，複数の該当箇所が見られる．一方で，7 月から 9 月では最長で 5 日間（8 月 3 日から 8 月 7 日）の正解にとどまっている．次に，正解と真逆の分類をした箇所に注目する．正解と真逆の分類とは，ある日付の正解ラベルが「上昇」,「低下」であるのに対し，分類結果が「低下」,「上昇」であった場合を指す．4 月から 6 月では 6 箇所（4 月 29 日，5 月 8 日，5 月 30 日，6 月 7 日，6 月 9 日，6 月 16 日，6 月 21 日）確認されたのに対し，7 月から 9 月では 8 箇所（7 月 2 日，7 月 6 日，7 月 25 日，8 月 19 日，8 月 23 日, 9 月 7 日, 9 月 10 日，9 月 23 日) 確認された．

これらの結果は，4 月から 6 月と 7 月から 9 月の期間で，頭痛発生傾向を検知する能力に違いがあることを示している．4 月から 6 月と 7 月から 9 月の期間の気象状況の違いを理解するために，各期間における平均気温と平均現地気圧の前日差，平均湿度と頭痛率の前日比の箱ひげ図を，図 6 に示す．7 月から 9 月に比べると，4 月から 6 月はどの前日差，前日比も中央値がほぼ同じで，四分位範囲が大きい傾向にあることが分かる．これは，4 月から 6 月の期間は，気象状況が前日から大きく変化する日が多いことを表している．気象と頭痛発生の時間的な関係性を調査した研究によれば，寒冷前線が頻繁にやってくる時期に気象と頭痛発生に関係性が見られたこと [17] が報告されている．気象と頭痛発生の関係性は未だ分かっていないことが多いが，急激な気象状況の変化に伴うストレスや物理的刺激が，自律神経の乱れに繋がりやすいため，頭痛発生を引き起こしやすいと考えられている [10]．この考えに基づくと，気象状況が 1 日単位で大きく変化するような時期では，気象の変化は頭痛発生に大きく影響を与えるようになるため，気象データを用いた頭痛発生傾向の検知が比較的可能になるのかもしれない．

5.2　気象と頭痛発生

5.1 節のエラー分析は，気象と頭痛発生の関係性が時期によって複雑に変化する可能性を示唆している．そこで，3ヶ月毎の気象データと頭痛率の相関係数を用いた関係性から，頭痛発生傾向の検知に，気象データを統合的に用いることが効果的な状況について分析する．表 3 に，LR_{all} の各説明変数と頭痛率の各期間の相関係数を示す．相関係数には，母集団の分布を仮定しないスピアマンの順位相関係数を用い，有意水準は 0.05 とした．全期間は 2022 年 7 月 1 日から 2023 年 9 月 30 日，summer22 は 2022 年 7 月 1 日から 2022 年 9 月 30 日，fall22 は 2022 年 10 月 1 日から 2022 年 12 月 31 日，winter23 は 2023 年 1 月 1 日から 2023 年 3 月 31 日，spring23 は 2023 年 4 月 1 日から 2023 年 6 月 30 日，summer23 は 2023 年 7 月 1 日から 2023 年 9 月 30 日を表している．

全期間で各説明変数に有意な相関を確認できるが，サンプル数 (n=457) の影響であることを考慮し，期間毎の相関について考察する．期間毎の傾向として，winter23 と spring23 では有意な相関が多く確認される．これは，気象が 1 日単位で大きく変化する時期は，気象と頭痛発生が関係する 5.1 節の考察を裏付けている．winter23 と spring23 で，1 時間毎の気温の 1 日の最大上昇差は -0.27 と -0.26 の有意な負の相関が確認された．これは，1 月から 6 月の 1 時間単位で気温上昇が大きい日に，頭痛発生が減少する可能性を示唆している．この条件に該当する日の特徴を理解するために，天気との関係性を調べる．

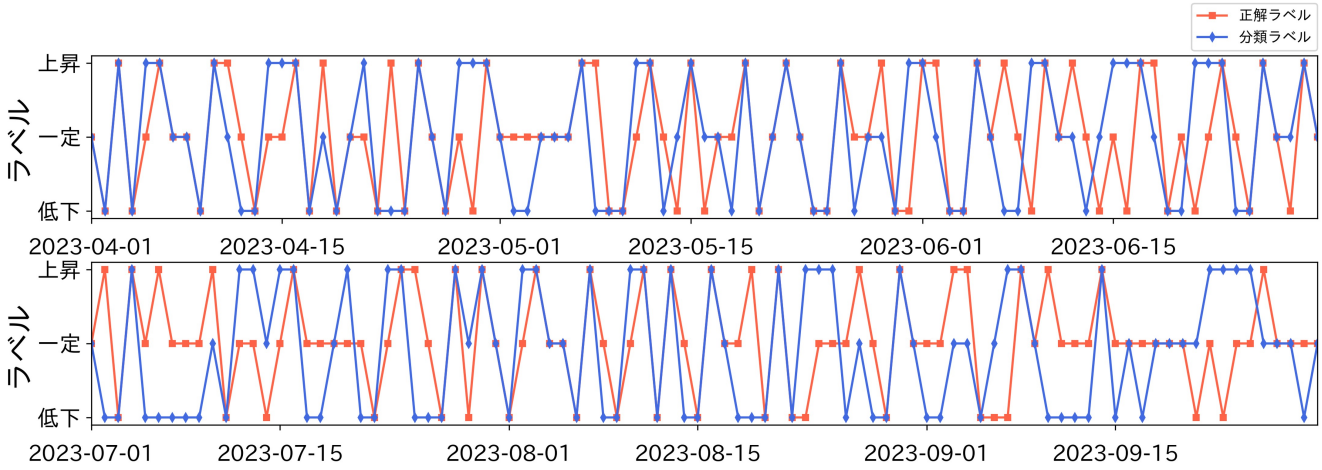


図 5　テストデータの正解ラベル（□）と LR_{all} の分類ラベル（◇）の推移．上図が 2023 年 4 月 1 日から 6 月 30 日，下図が 2023 年 7 月 1 日から 9 月 30 日の結果を表している．

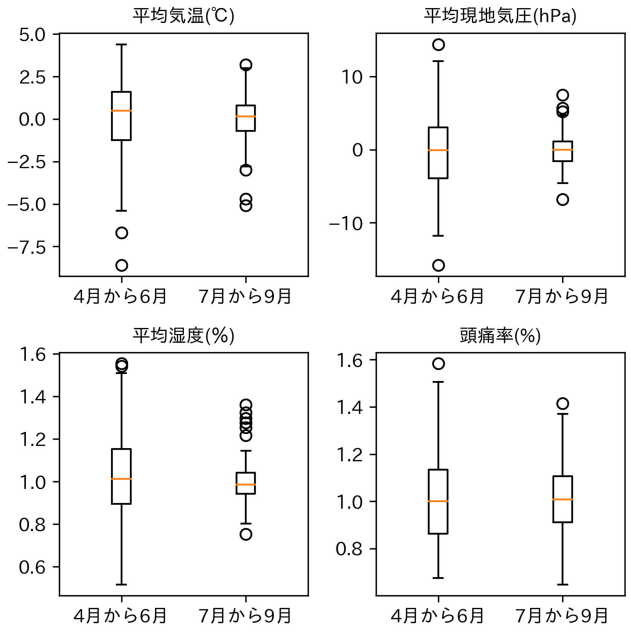


図 6　平均気温と平均現地気圧の前日差と平均湿度と頭痛率の前日比の箱ひげ図

気象庁の過去の気象データ検索から天気概況 (昼：06 時～18 時) のデータを取得し，1 時間毎の気温の 1 日の最大上昇差の中央値以上に該当する日数のうち，天気概況に ‘晴’，‘曇’，‘雨’ のそれぞれの文字を含む日数の割合を比較した．‘ ‘晴’，‘曇’，‘雨’ の割合は，それぞれ 86%, 63%, 12%であり，1 時間毎の気温の 1 日の最大上昇差が大きい日には，晴れの日が多く見られることが分かる．つまり，1 時間単位の気温上昇が大きい晴れの日に，頭痛発生が減少する傾向がある．spring23 で，1 時間毎の湿度の 1 日の最大下降比は，0.26 の有意な正の相関が確認された．気温と同様に，1 時間毎の湿度の 1 日の最大下降比でも天気の割合を比較した．‘晴’，‘曇’，‘雨’ の割合は，47%, 78%, 39%であり，1 時間毎の湿度の 1 日の最大下降比が大きい日には，曇りや雨の日が多く見られる．1 時間毎の湿度の 1 日の最大下降比の大きさは，1 時間後の湿度変化の小ささを意味している．この結果は，1 時間単位の湿度変化が小さい曇りや雨の日に，頭痛発生が増加する傾向であることを示している．

これらの 1 時間毎の特徴量に関する結果をまとめると，1 日の天気は頭痛発生に関係する可能性がある．この知見は，悪天候で頭痛発生が確認された多くの文献で主張されている [11, 17]．気象特徴量は，気圧特徴量だけでは十分に捉えられない天気の特徴を表現している可能性がある．一般的に，低気圧時は悪天候が多いため，天気を表す点で気圧特徴量は有効であるが，それだけで天気を説明することは難しい．例えば，晴れの日に地表が太陽光で温められることで上昇気流が生じ，一時的に気圧が下がる場合がある．一方で，このような状況では短時間の気温上昇が期待され，気象特徴量が天気の情報を補完する可能性が考えられる．気象の変化と天気の関係性は非常に複雑であるため，気象特徴量がどの情報を具体的に補完するかは今後調査されるべきである．

LR_{all} の各説明変数は，ある基準からの差や比を表す特徴量であるため，1 年を通したトレンドに関する情報は排除されている．一方で，1 日の気象の平均値がある閾値以上の場合に，頭痛が起こりやすい可能性があるため，各気象データの 1 日の平均値に関しても同様に分析する．表 3 の下段に，平均気温，平均湿度，平均気圧の各期間における相関係数を示す．winter23 と spring23 で各特徴量の相関が強まる傾向に関しては，5.1 節の考察と同様である．一方で，1 年を通して高い気温が観測されることが多い summer22 と summer23 において，平均気温と頭痛発生の間に有意な正の相関が見られる．さらに，summer23 の 1 時間毎の気圧の 1 日の最大下降差を除き，これらの期間で頭痛発生と有意な相関が見られる可変変数は確認できない．この結果は，高い気温が記録される時期の 1 日の気温は，頭痛発生に大きく寄与する可能性を示唆している．頭痛発生と暑さに焦点を絞った研究は極めて少ないが，山口らは夏季夜間の小児の頭痛患者数が，平均気温と有意な正の関係が見られたことを報告している [16]．頭痛と気温の詳細な病態メカニズムは不明

表 3　LR_{all} の各説明変数における 3ヶ月毎のスピアマンの順位相関係数と p 値．全期間は 2022 年 7 月 1 日から 2023 年 9 月 30 日，summer22 は 2022 年 7 月 1 日から 2022 年 9 月 30 日，fall22 は 2022 年 10 月 1 日から 2022 年 12 月 31 日，winter23 は 2023 年 1 月 1 日から 2023 年 3 月 31 日，spring23 は 2023 年 4 月 1 日から 2023 年 6 月 30 日，summer23 は 2023 年 7 月 1 日から 2023 年 9 月 30 日を表す．括弧内は p 値を表し，太字は $p \leqq 0.05$ の値を表す．

	全期間 $n = 457$	summer22 $n = 92$	fall22 $n = 92$	winter23 $n = 90$	spring23 $n = 91$	summer23 $n = 92$
1 時間毎の気温の 1 日の最大上昇差	**−0.09 (0.05)**	0.14 (0.17)	−0.15 (0.14)	**−0.27 (0.01)**	**−0.26 (0.01)**	−0.07 (0.49)
1 時間毎の湿度の 1 日の最大下降比	**0.23 (0.00)**	−0.09 (0.40)	0.15 (0.15)	0.14 (0.20)	**0.26 (0.01)**	0.16 (0.14)
1 時間毎の気圧の 1 日の最大下降差	**0.22 (0.00)**	−0.07 (0.49)	−0.02 (0.85)	0.00 (0.98)	−0.12 (0.27)	**0.21 (0.04)**
前日からの平均気圧差	**−0.15 (0.01)**	−0.08 (0.42)	−0.16 (0.14)	−0.15 (0.14)	**−0.21 (0.04)**	−0.05 (0.64)
大気潮汐	**−0.12 (0.01)**	−0.09 (0.38)	−0.03 (0.77)	0.03 (0.78)	−0.12 (0.24)	0.04 (0.70)
長期パターン	**−0.16 (0.00)**	−0.16 (0.12)	−0.15 (0.17)	**−0.29 (0.01)**	**−0.24 (0.02)**	−0.04 (0.68)
平均気温	**0.56 (0.00)**	**0.38 (0.00)**	0.18 (0.09)	0.11 (0.28)	**0.23 (0.03)**	**0.41 (0.00)**
平均湿度	**0.37 (0.00)**	0.01 (0.95)	**0.22 (0.04)**	**0.31 (0.00)**	**0.42 (0.00)**	−0.06 (0.57)
平均気圧	**−0.37 (0.00)**	−0.10 (0.32)	−0.02 (0.85)	**−0.23 (0.02)**	**−0.37 (0.00)**	−0.06 (0.59)

とされているが，夏季の時期に頭痛発生の要因が変化している可能性は考えられる．5.1 節で見られた LR_{all} の分類性能の低下に対し，この知見は性能改善の可能性を示唆している．

5.3 限　　界

気象データを統合的に用いて頭痛発生傾向を検知するモデルから，両者の関係性を分析した本研究のアプローチには，いくつかの限界がある．1 つ目は，今回用いた説明変数の最良な組み合わせは，別に存在する可能性がある．本研究では，RFE を用いてモデルが考慮する説明変数の数を最長 14 個に制限し，14 個の全組み合わせを調査した．RFE はモデルを構築した際の重要度が小さい変数を削除する方法であるため，実際の分類に有効な特徴量を削除してしまう可能性がある．候補として挙げた可変変数の探索範囲を拡大できれば，分類性能はさらに向上する可能性がある．しかし，現実的な時間内で全ての組み合わせを比較することは不可能であるため，RFE よりも最適な変数選択の方法を模索する必要がある．

2 つ目は，分析対象の集団内にバイアスが存在することである．本研究では，クラウドソーシングを用いて東京都在住者を対象とした上で，気象と頭痛発生の関係性を調査した．クラウドソーシングの手法上，回答者が特定の集団に偏ってしまう選択バイアスや，昨日のことを尋ねた時間感覚的な歪みに関する記憶バイアスなどが存在する．これらのバイアスを排除することはできないため，今回の知見が現実の集団に対しても適用できるか別アプローチからの検証を行うべきである．

3 つ目は，頭痛率のデータに気象以外の要因がどの程度含まれていたかは不明なことである．頭痛には，気象以外にストレス，疲れ，騒音など様々な要因や，片頭痛や緊張型頭痛などの様々な種類が存在することが分かっている．本研究では，クラウドソーシングの質問で気象条件に言及せずに，昨日頭痛があったかどうかのシンプルな質問をした．そのため，気象以外の要因による頭痛や，メカニズムが異なる様々な頭痛のノイズが混入している可能性がある．従って，気象が影響する頭痛に対象を絞ることができれば，気象と頭痛発生に関するより有益な知見を提供することに繋がると考えられる．

4 つ目は，アプリケーションを実現できる頭痛発生傾向の検知性能を，現状達成できていないことである．表 2 を見ると，Majority baseline と比較して気象データを用いたことで分類性能は改善したが，最高で約 5 割の分類性能にとどまっている．現状の分類問題に機械学習を用いる意味はあるが，アプリケーションの性能まで昇華させることは現時点では難しい．最終的に，気象状況から頭痛発生傾向を把握できるアプリケーションの構築を目指しているが，最低でも 8 割の分類性能を達成したいと考えている．解決策として，気象の影響を強く受ける特定の属性集団を対象とすることや，気象と頭痛発生の関係性をより表現できるモデルを検討することなどが考えられる．一方で，気象データを用いて頭痛発生傾向を検知することが，如何なる場合でも難しい可能性も考えられるため，多様なアプローチの観点から今後も研究されるべきである．

6 おわりに

本研究では，気象状況から頭痛発生傾向を把握するアプリケーションの構築を目標とし，気象データを統合的に用いて頭痛発生傾向を検知する可能性を調査した．そのために，クラウドソーシングで得られた回答から頭痛発生を表す「頭痛率」指標を導入し，頭痛率が前日から上昇したかを分類する 3 クラス分類を行った．気象データの様々な組み合わせを比較した実験結果から，気圧，気温，湿度の特徴量を統合的に用いたロジスティック回帰が最も高い分類性能を示した．さらに，構築した分類モデルの分析を通して，1 日単位の気象変化が大きい時期では，比較的高性能な分類が実現できることを示し，気温や湿度の気象データを用いて頭痛発生傾向を検知する重要性を示した．

今後の発展としては，大きく 3 点考えられる．1 点目は，限

定された属性の集団でどの程度気象データを用いた頭痛発生傾向の検知が可能かを調査することである．例えば，女性の方が男性よりも偏頭痛が多いことが報告されており [13]，性別の違いによって気象が頭痛発生に与える影響に違いが生じる可能性がある．本研究では，クラウドソーシングにおける東京都在住者の集団のみを対象としたため，様々な集団での頭痛発生に関する調査は，興味深い知見を提供すると考えている．

2 点目は，頭痛発生との関係性がより強く，収集が比較的容易なデータを検知に利用することである．5.3 節でも言及した通り，今回の頭痛発生傾向の検知性能は約 5 割である．気象データの収集は容易ではあるが，今回の結果から気象データを用いるだけでは頭痛発生傾向を十分に検知することは困難である．解決方法としては，頭痛発生とより関係性の強いデータを検知に用いることである．例えば,「頭痛」,「頭痛薬」,「天気」などの web 検索クエリの検索回数は頭痛発生と相関があるかもしれない [3, 9]．さらに，人間の暑さを定量化した暑さ指数は，気象データよりも頭痛発生の生体情報として有用かもしれない．これらのデータが頭痛発生とどう関係するかを調査することは，頭痛発生を将来的に抑制するための重要な研究になる．

3 点目は，過去の気象データから頭痛発生傾向をどの程度検知できるか調査することである．本研究では，主に同時点のデータから頭痛発生傾向を検知するアプローチを採用したが，数日前に入手可能なデータから検知できた方がより有用なアプリケーションを実現できる．今回の結果は，時間軸で最も近い気象データを用いた場合でも，約 5 割の検知性能にとどまった．頭痛発生傾向を検知するアプローチについて様々な観点から研究することは，頭痛発生を抑制するためのアプリケーションの実現に近づく．

謝　　辞

本研究は，LINE ヤフー株式会社共同研究費の支援を受けたものである．

文　　献

[1] Christopher J Elcik, Christopher M Fuhrmann, Scott C Sheridan, Kathleen Sherman-Morris, and Andrew E Mercer. Perceptions of weather-based pain forecasts and their effect on daily activities. *International Journal of Biometeorology*, pp. 1–15, 2023.

[2] Megumi Funakubo, Jun Sato, Takashi Honda, and Kazue Mizumura. The inner ear is involved in the aggravation of nociceptive behavior induced by lowering barometric pressure of nerve injured rats. *European Journal of Pain*, Vol. 14, No. 1, pp. 32–39, 2010.

[3] Kurubaran Ganasegeran, Alan Swee HockCh'ng, Zariah Abdul Aziz, Irene Looi. Population's health information-seeking behaviors and geographic variations of stroke in malaysia: An ecological correlation and time series study. *Scientific reports*, Vol. 10, No. 1, p. 11353, 2020.

[4] Joseph L Hollander and Sarantos J Yeostros. The effect of simultaneous variations of humidity and barometric pressure on arthritis. *Bulletin of the American Meteorological Society*, Vol. 44, No. 8, pp. 489–494, 1963.

[5] Shinsuke Inoue, Fumio Kobayashi, Makoto Nishihara, Young-Chang P. Arai, Tatsunori Ikemoto, Takashi Kawai, Masayuki Inoue, Tomomi Hasegawa, and Takahiro Ushida. Chronic pain in the japanese community—prevalence, characteristics and impact on quality of life. *PLOS ONE*, Vol. 10, No. 6, pp. 1–14, 2015.

[6] Masahito Katsuki, Muneto Tatsumoto, Kazuhito Kimoto, Takashige Iiyama, Masato Tajima, Tsuyoshi Munakata, Taihei Miyamoto, and Tomokazu Shimazu. Investigating the effects of weather on headache occurrence using a smartphone application and artificial intelligence: a retrospective observational cross-sectional study. *Headache: The Journal of Head and Face Pain*, Vol. 63, No. 5, pp. 585–600, 2023.

[7] Wenyuan Li, Suzanne M. Bertisch, Elizabeth Mostofsky, Catherine Buettner, and Murray A. Mittleman. Weather, ambient air pollution, and risk of migraine headache onset among patients with migraine. *Environment International*, Vol. 132, p. 105100, 2019.

[8] Xin-yu Li, Cheng-hao Yang, Jia-jie Lv, Hui Liu, Lu-yu Zhang, Min-yi Yin, Zhi-lin Guo, and Ru-hong Zhang. Global, regional, and national epidemiology of migraine and tension-type headache in youths and young adults aged 15–39 years from 1990 to 2019: findings from the global burden of disease study 2019. *The Journal of Headache and Pain*, Vol. 24, No. 1, p. 126, 2023.

[9] Marcelo D Mendonça, André Caetano, and Miguel Viana-Baptista. "Dr Google" will see you now–time trends in online searches on headache. *Cephalalgia*, Vol. 38, No. 2, pp. 407–408, 2018.

[10] Karl Messlinger, Megumi Funakubo, Jun Sato, and Kazue Mizumura. Increases in neuronal activity in rat spinal trigeminal nucleus following changes in barometric pressure—relevance for weather-associated headaches? *Headache: The Journal of Head and Face Pain*, Vol. 50, No. 9, pp. 1449–1463, 2010.

[11] Kayoko Ozeki, Tatsuya Noda, Mieko Nakamura, and Toshiyuki Ojima. Weather and headache onset: a large-scale study of headache medicine purchases. *International journal of biometeorology*, Vol. 59, pp. 447–451, 2015.

[12] Bianca Raffaelli, Jasper Mecklenburg, Lucas Hendrik Overeem, Simon Scholler, Markus A Dahlem, Tobias Kurth, Ana Sofia Oliveira Gonçalves, Uwe Reuter, and Lars Neeb. Determining the evolution of headache among regular users of a daily electronic diary via a smartphone app: observational study. *JMIR mHealth and uHealth*, Vol. 9, No. 7, p. e26401, 2021.

[13] F Sakai and H Igarashi. Prevalence of migraine in japan: a nationwide survey. *Cephalalgia*, Vol. 17, No. 1, pp. 15–22, 1997.

[14] J Sato, R Ueyama, K Morita, T Furuya, Y Otsuka, S Hatakeyama, et al. The epidemiological and clinical features of weather–related pain (tenkitsu) and development of prediction information service for the onset of pain. *Pain Res*, Vol. 36, No. 2, pp. 75–80, 2021.

[15] Toshihiko Shimizu, Fumihiko Sakai, Hitoshi Miyake, Tomofumi Sone, Mitsuhiro Sato, Satoshi Tanabe, Yasuhiro Azuma, and David W Dodick. Disability, quality of life, productivity impairment and employer costs of migraine in the workplace. *The journal of headache and pain*, Vol. 22, No. 1, pp. 1–11, 2021.

[16] Hiroshi Yamaguchi, Kandai Nozu, Shinya Ishiko, Hiroaki Nagase, Takeshi Ninchoji, Hiroki Takeda, Ai Unzaki, Kazuto Ishibashi, Ichiro Morioka, Kazumoto Iijima, et al. Multivariate analysis of the impact of weather and air pollution on emergency department visits for night-time headaches among children: retrospective, clinical observational study. *BMJ open*, Vol. 11, No. 4, p. e046520, 2021.

[17] Albert C Yang, Jong-Ling Fuh, Norden E Huang, Ben-Chang Shia, Chung-Kang Peng, and Shuu-Jiun Wang. Tem-

poral associations between weather and headache: analysis by empirical mode decomposition. *PLOS ONE*, Vol. 6, No. 1, p. e14612, 2011.

[18] Mustafa Yilmaz, Mehtap Gurger, Metin Atescelik, Mustafa Yildiz, and Sukru Gurbuz. Meteorologic parameters and migraine headache: Ed study. *The American Journal of Emergency Medicine*, Vol. 33, No. 3, pp. 409–413, 2015.

[19] ウェザーニュース. 【天気痛調査 2020】天気痛持ちは全体の 6 割？天気痛女子は約 8 割にも, July 2020. `https://weathernews.jp/s/topics/202007/070165/`.

データ駆動型研究を促進する大規模かつ多種類の生体光学データのためのデータベース基盤の開発

山中　卓也† 　熊本　康昭††,††† 　西村　隆宏†† 　久保　俊貴††††,†† 　Heqi　Xi††
藤田　克昌†† 　松原　靖子† 　櫻井　保志†

† 大阪大学産業科学研究所産業科学 AI センター　〒 567-–0047 大阪府茨木市美穂ケ丘 8–1
†† 大阪大学大学院工学研究科　〒 565–0871 大阪府吹田市山田丘 2-1
††† 大阪大学先導的学際研究機構　〒 565–0871 吹田市山田丘 1-1
†††† 大阪大学大学院医学系研究科　〒 565–0871 吹田市山田丘 2-2
E-mail: †t-yamanaka@sanken.oskaa-u.ac.jp, ††{kumamoto,kubo,xiheqi,fujita}@ap.eng.osaka-u.ac.jp,
†††nishimura-t@see.eng.osaka-u.ac.jp, ††††{yasuko,yasushi}@sanken.osaka-u.ac.jp

あらまし　フォトニクス (光工学) 技術の発展はめざましく，中でも近年その試料測定時の非侵襲性と前処理が不要であることから，生体・医療への応用が期待されている．フォトニクスによる測定データは物質からのラマン分光・吸収散乱スペクトルや，光学測定値，その空間画像情報など多種多様でかつ大量でデータサイズも大きいため，効果的なデータ活用のためにデータベース化が重要となる．測定条件や装置のパラメータ，試料種別等多数の情報を構造化し，適切なメタデータとして付与することで，大量の測定データから有用な情報を効率的に取得・分析できるだけでなく，データマイニングにより人の目では捉えることが難しい特徴や類似性を発見することで，近年さかんとなっているデータ駆動型の研究の推進も期待される．本研究では，多様なフォトニクス測定装置から集められた多数の生体光学データを管理し，付与されたメタデータによる検索機能，類似ピークの検索機能，ノイズ除去や特異値分解 (SVD) によるデータ前処理機能をもつデータベース基盤を開発したのでその設計と機能について報告する．

キーワード　データベース基盤, 生体光学データ, データ駆動型研究, データマイニング, スペクトルデータ

1　はじめに

フォトニクス (光工学) は光エレクトロニクスを背景に光の発生・制御・検出にかかわる現象・機能・技術であり，それを多角的に活用するデバイス・システムを含み，光通信，光波シンセシスや分光計測といった光科学，レーザー加工などの製造，医学など，極めて広い範囲に渡り，人類の発展に寄与するものである [1]．この中でも光を用いた生体の計測技術は，その非侵襲性や前処理が不要なことから医療への応用が期待されている．しかし，実際の医療に用いるためには課題も多く，特に計測時間が長くなることや，測定時に対象部位に標識をつけなければならない，などの問題がある．このような課題に対応するために藤田らは，ラマン分光技術を利用した生体のイメージング技術の研究を進めてきた [2] [3]．また，光を利用した光医療を安全に行うには，物質や組織ごとの光の各特性値を正確に測定する必要がある．このための研究として，西村らは双積分球を利用した生体組織の光学特性値の計測 [4] [5] を行っている．

これらのフォトニクスによる生体計測技術により得られるスペクトルデータは，試料の種類や前処理の加工技術，試薬の違い，測定時に用いる分光器 (spectrometer)，対物レンズの種類や露光時間といった多様な条件により測定される．また測定データ容量も大きく，画像の各 pixel に対応したスペクトルデータは数百 MB にもなる．研究者はスペクトルデータを解析することで，生体試料内に含まれる元素，分子，物質や組織などを同定するが，測定データは研究者ごとに管理されたり，共有フォルダ等に保存される．測定データに付随するメタデータは存在するものの，構造化されておらず，また測定者によって付与していない属性も多いため，単純なデータ検索も難しい状態となり，データが有効に活用されていない．

このようなことから，測定データをデータベース化することは重要である．メタデータが正確に付与されたデータは条件抽出による検索が容易となる．測定データに含まれたノイズを除去するためのデータ前処理を保存されたデータに対して，統一的に実施することができる．集約されたデータに機械学習手法を用いたデータマイニングを行うことで，研究者が事前に想定していない相関関係や特徴を発見する，データ駆動型研究を実現することも期待できる [6]．

本研究では 2 種類，あるいはそれ以上の生体測定データを扱う．それぞれデータ，メタデータ，測定装置も異なるので，データベース定義を測定日時や測定場所といった共通的なメタデータと，測定装置のパラメータ等の固有の条件を持つものに分類する必要がある．一方，データの検索や参照といったインタフェースは共通的に設計することが可能である．

この生体光学データベースを医療・診断へ活かすことは本研究の重要な目標の一つであるため，病院の医療データベースへの医療情報と連携できるように設計する．

以上をふまえて，本研究では次のような要求を満たすデータ

ベース基盤の開発を行う．

- 測定データのメタデータを構造化し，条件抽出，検索を可能とする．
- 測定装置からの日々の測定データをアップロードするための API を提供する．
- 測定データに対して必要な前処理を行う
- 測定データに対してデータマイニングを行うことができる．
- 他のデータベースとの連携する枠組みを提供する．
- 多種類の生体光学データに対して上記のすべての機能を可能とする．

2 関 連 研 究

科学研究においてデータベース利用は，研究におけるデジタル化やインターネットの普及とともに進展してきた．化学 [7], 生物学 [8], 材料科学 [9] といった各分野で化合物や遺伝子，物質の結晶構造などがデータベース化されており，広く研究者に用いられている．また，核磁気共鳴装置 (NMR), X 線解析装置 (XPS), 質量分析装置 (MS) 等，各種物理・化学作用を利用した測定装置によるスペクトルのデータベース [10] [11] [12] や，電子顕微鏡等を用いて撮像した生体画像データやクライオ顕微鏡を用いた 3 次元構造イメージなどのデータベース [13] [14] などが数多く存在する．今回扱うデータの 1 つは生体試料に対するラマン分光装置によるイメージング画像を含んだラマン分光スペクトルデータである．ラマン分光スペクトルデータを扱うデータベースは [15] [16] などがあるが，前者は無機物に対するものであり，生体試料に対するものではない．後者は有機化合物を対象としており，ラマンスペクトルのデータは一部の化合物しか存在しない．その他の前述した各種データベースにも対象の物質のラマンスペクトルの情報が含まれているものもあるが，生体試料に対するラマン分光データをメインとし，かつ 2 次元情報を扱うデータベースを見つけることはできなかった．

さらに，上述のようなデータベースは論文等の研究成果や，研究者からの測定データを収集して，主に人手を使って選別し，化合物，遺伝子，物質の構造等の各種リファレンスとして用いられる．測定装置で日々測定しているデータを構造化し，その解析を効率化するような使用を想定されたものではない．

日々の測定データ管理としては，Research Data Management(RDM) あるいは Science Data Management(SDM) を行うソフトウェアとして, [17] [18] など数多く存在するが，これらの多くは有償でかつ対象も単なる測定データの管理だけではなく，そのセキュリティや実験ノートとの連携，広範囲の装置を対象とするシステムであり，今回必要とするデータベース基盤とはスコープが異なる．オープンソースの RDM ソフトウェアとしては，研究データ管理だけでなく，実験ノートの機能といった包括的な実験管理を行うもの [19] や，顕微鏡画像を対象として，データを保存，分析，共有ができるもの [21]，バイオインフォマティクスをメインフィールドとして，遺伝子発現解析，チップシーケンス解析などを解析パイプラインの構築が可能なもの [20] がある．これらのソフトウェアをカスタマイズすることで生体光学データベース基盤を構築可能と考えられるが，対象が異なるため事前のサーベイと多数の機能追加・変更などが必要となる．そのため現時点では柔軟な対応と開発速度を優先するため，Python とフレームワーク Django による新規開発を行った．ただし，今後このような RDM ソフトウェアをうまく組み合わせてデータベース基盤を発展させていくことは検討の余地がある．

3 扱う生体光学データとその形式

今回扱う 2 つの生体光学データの形式について説明する．それぞれの測定手法の詳細については [2]〜[5] を参照のこと．

3.1 ラマン分光イメージングデータ

[2] [3] で使用されるラマン顕微鏡により，マウスの組織や，ヒト幹細胞などの生体試料に対して測定した $x_{max} \times y_{max} \times \lambda_{max}$ のデータとなる．x_{max} が 250 前後, y_{max} が 400 程度，λ_{max} は波数 (wave number: 単位当たりの波の数) で 800 - 1000 程度の 3 次元テンソルのデータとなる (図 1 上左)．データの数値は試料の平面座標 (x, y) における，波数 λ の強度を表す． 例としてある波数 $\lambda = 100$ における xy 平面画像 (図 1 上右) と，画像中の座標 (100, 100) におけるスペクトルデータを示す (図 1 下)．1 測定データにこの 3 次元テンソルデータが 1 つ生成され，200MB - 700MB 程度のサイズとなる．これらに測定時のメタデータを付与する．

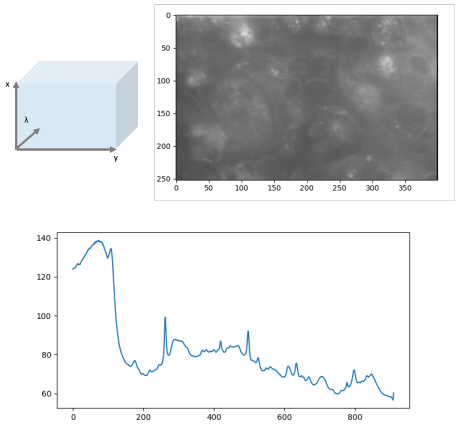


図 1 (上左) ラマン分光イメージングデータ構造 (上右) 波数 100 における xy 平面イメージ (下) $(x, y) = (100, 100)$ におけるスペクトル波形

3.2 双積分球データ

[4] [5] で用いられる双積分球光学系を用いて，マウスの組織，ヒトの皮膚組織等の生体試料について測定された光学特性値の $\lambda_{max} \times d$ の 2 次元の csv データとなる．λ_{max} は 1600 程度で，波数 400 - 1100 nm の範囲の波数，d は光学特性値の数で, 吸収係数 μ_a, 散乱係数 μ_s, 拡散反射率 R_d, 透過率 T_t で 4 となる．データの数値は各波数におけるそれぞれの光学特性値となる．

1 測定について，この csv データと測定時のメタデータを付与する．例として μ_a, μ_s のスペクトルデータを示す (図 2)．容量は 100KB 未満で小さい．

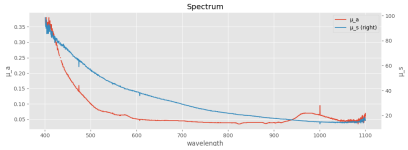


図 2　双積分球データの例: 吸収係数 μ_a, 散乱係数 μ_s のスペクトル

4　データベース基盤の構成

図 3 に本研究で開発するデータベース基盤の構成と，測定装置とデータの関係図を示す．測定装置から API を通じて各装置で測定した生体光学データをデータベースに保存・管理する Storage Module ．保存したデータを様々な条件で検索する Search Module．AI 機能にて基本的なデータ変換・ノイズ除去，発展的なデータ解析を行う AI Module．他のデータベースとの連携を行う Connection Module．の 4 つの Module を持つ．データの参照・検索等を行うための Web インタフェースも提供する．

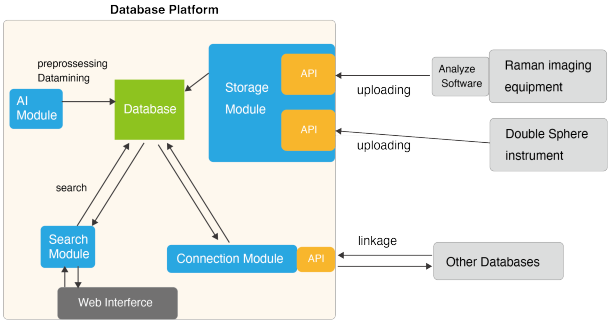


図 3　データベース基盤の構成および測定装置とデータの関係

4.1　Storage Module

測定したデータを保存，管理を行う．データそのものは物理ディスクもしくはオブジェクトストレージに保存し，メタデータ等は RDBMS である PostgreSQL を用いて管理する．測定時のメタデータとしては, 測定者, 場所 測定時刻といった全般的な情報, 分光器や検出器，対物レンズの種類，波長や露出時間等の機器についての情報, 生体試料の種類, 組織, 加工方法, コーティングや触媒の種類など試料についての情報などで，ラマン分光イメージングデータは 30 種類以上，双積分球データは 15 種類以上ある．これらを測定者が管理するのは困難であるため，測定装置と API で連携し, メタデータのチェックと画像を含むスペクトルデータ等の保存を行う．

4.2　Search Module

4.2.1　メタデータによる基本的な検索機能

データベース上に保存されたスペクトルデータから，様々な条件でデータの検索を行う．研究者は測定日や生体試料の種類・使用した測定装置の種類や条件で測定データを絞り込んで, 研究に有用なデータを探す必要があるため, この機能は非常に重要である．少ないキーワードでメタデータを検索できる簡易検索と, 複雑な条件を組み合わせて行う詳細検索を用意している．数値のデータは不等号による範囲指定で検索可能である．

4.2.2　高度な検索機能

データベース上に保存されたスペクトルデータから，より高度な検索を行う．研究者とミーティングを行い，研究の上で有用な検索機能について検討を行った．ラマン分光イメージングデータについては，保存されたスペクトルデータから, 共通のピーク検出を行う機能を開発した．研究者は特定の波数のピークを含むスペクトルから，試料中に含まれる物質，組織を同定する．ピークは通常目視で判別しているため，多数の測定データがある場合，そのすべてについて類似ピークを確認するのは難しい．この機能では, 指定した波数の範囲にピークがあるデータをすべて表示する．データベースに保存された大量のスペクトルデータからピーク位置を事前に計算しているため，効率よく複数のデータから類似ピークを持つデータを検出することができる．今後も有用な検索機能について良いアイデアがあれば，随時追加していくことが可能である．

4.3　AI(Analyze) Module

4.3.1　基本的なデータ処理機能

装置から保存された生体光学データに前処理を行う．測定した生データは多くのノイズが含まれているため，通常研究者はデータの前処理を行ってからデータの分析を行う．ラマン分光イメージングデータにおいて，主に用いられるのは特異値分解 (Singular Value Decomposition: SVD) であり，平面画像 $X \times Y pixels$ の各画素に対応する $X \times Y$ 本のスペクトルを上位 20 程度の特異ベクトルに集約する．研究者が個々に測定データに対して，手動でこの操作を行う必要があり，計算時間もかかっていたものが，データベースに保存されたデータについては自動でノイズ除去がされるので，作業時間を効率化できる．

4.3.2　AI データ解析機能

データベースに対して，より発展的なデータマイニング手法を用いて, 自動的にデータ同士の相関関係の抽出・特徴判別・予測を行う．これまで研究者の経験と知識と試行錯誤から測定データと組織の病変等の関係性を推論しなければならなかったところ，蓄積されたデータから特徴を抽出して教師なし分類や教師あり分類，あるいは研究者によるラベリングを加えた半教師あり分類などの機械学習手法を用いることで, データ駆動型の研究を推進する．本機能については具体的な機能は研究者との間でミーティングを行っており検討中である．

4.4　Connection Module

他のデータベースとの連携するための機能を提供する．病院

の患者データベースとの連携や，他のスペクトルデータベースとの連携なども視野に入れており，そのための API・機能を開発予定である．

5 開発環境

本研究のデータベース基盤は，以下表 1 の環境で開発を実施した．

表 1 開発環境

サーバ	Supermicro Super Server
CPU	Intel Xeon 5218 CPU @ 2.30GHz
メモリ	192GiB
OS	CentOS Linux release 7.9.2009
DBMS	Postgresql 14.7
Web サーバ	Apache 2.4
開発言語	Python3.10
Web フレームワーク	Django4.2

6 生体光学データベースの機能とインタフェース

実際に開発した生体光学データベースの画面構成と機能・インタフェースについて紹介する．

6.1 データアップロード API

データベースでは API を提供し，測定装置からのデータのアップロードを可能としている．図 3 にあるように，ラマン測定装置の場合は，専用の分析ソフトウェアがあるため，これと連携してデータのアップロードが行われる．双積分球装置の場合はそのまま API を用いてアップロードを行う．API では JSON 形式を用いた．通常実験後は装置から USB メモリ等を使用して手動でデータを取りだす必要があり，測定者それぞれで管理することになるが，本機能により省力化と統一的なデータの扱いが可能となる．また各メタデータについては，各測定装置に必要なデータについて十分なヒアリングを行い，仕様を取り決めた．図 4 に仕様の一部を示す．

Large-Attribute	attribute	json_keyname	units	value (example)	null accepted	float	possible value	input (client)	master	type (json)
General	StarDateTime	start_date_time		2023/5/5 10:23				select date and time	-	datetime
	EndDateTime	end_date_time		2023/5/5 12:30						
	MeasurementName	measurement_name		20230518160935.3839383.mat				Free writing	-	string
	Setup	setup		P3-413			Multi-line Dual P3-305 P3-207 P3-413 CoMIT	Select from the list	server	string
	Operator	operator		Menglu			Menglu Matt Xi Kuma	Select from the list	client	string
	Remarks	remarks			○			Free writing	-	string
	Sample Type	sample_type		Cell			Standard Live 2D Cell Live 3D Cell Fixed 2D Cell Fixed 3D Cell Fixed Tissue slice Frozen Tissue slice Fresh tissue block	Select from the list	server	string
	SampleMaterial	sample_material		Rat			Calibration sample Micro beads Cell line Mouse Rat Human being Others	Select from the list	server	string
	SampleName1	sample_name_1		MCF7	○			Free writing		string
	SampleName2	sample_name_2		passage time 7	○			Free writing		string
	CellType	cell_type		HeLa			None HeLa MCF-7	Select from	server	string

図 4 メタデータの仕様 (抜粋)

6.2 データの一覧表示・メタデータによる条件検索・一括ダウンロード

装置からアップロードしたデータを Web インタフェースにより，一覧表示 (図 5)，個別メタデータ，画像データ (図 6) の確認ができる．

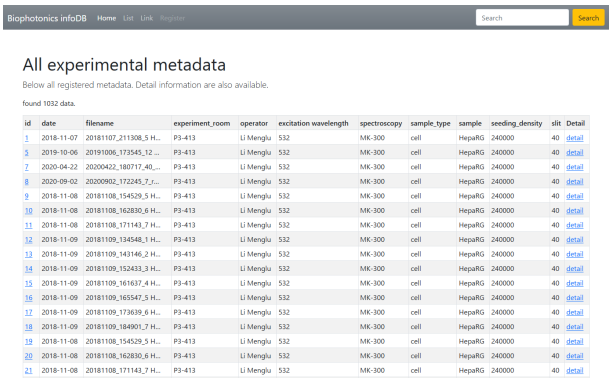


図 5 アップロードデータの一覧表示

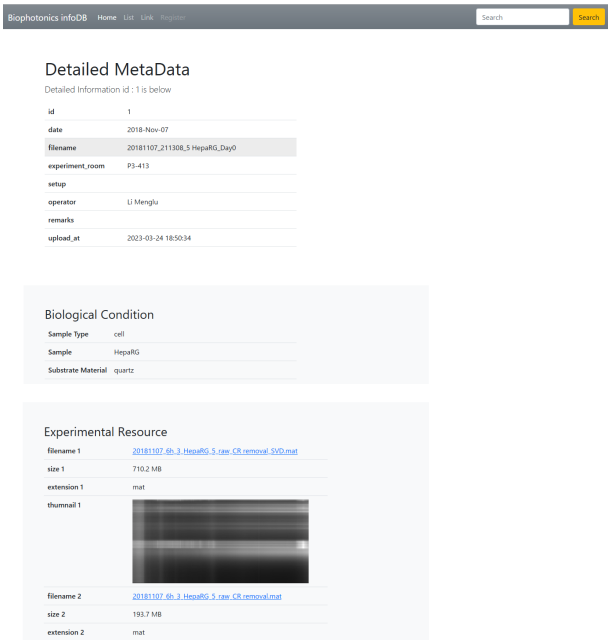


図 6 データの詳細表示

メタデータを複数条件指定することで，必要な測定データを検索できる．SQL 等を用いずに測定者，測定日，試料の種類，装置の測定パラメータ条件などを指定して必要とするデータを抽出し，該当した画像・スペクトルデータを同時にダウンロードも可能である．

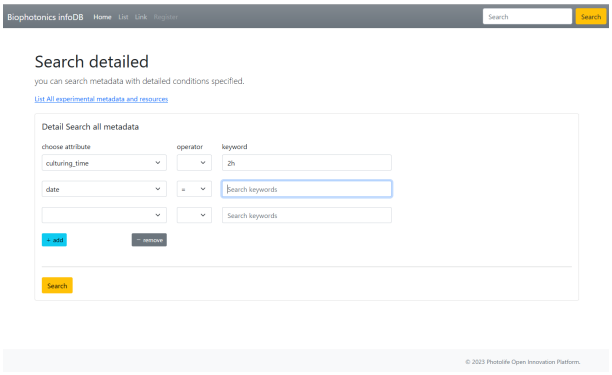


図 7　詳細な検索

6.3　データ前処理機能

データベースに格納されたデータからノイズを除去したデータを生成する．これは保存されたデータに対してバックグラウンド処理で自動的に行われる．ラマン分光イメージングデータでの例を図 8 に示す．図 8 上のように, 激しい振動ノイズが含まれたスペクトルが，図 8 下のように，解析容易なものとなる．この処理は 6.4 の類似ピークの検索時には，ノイズが多いとピークの判別が困難となることから必要不可欠である．

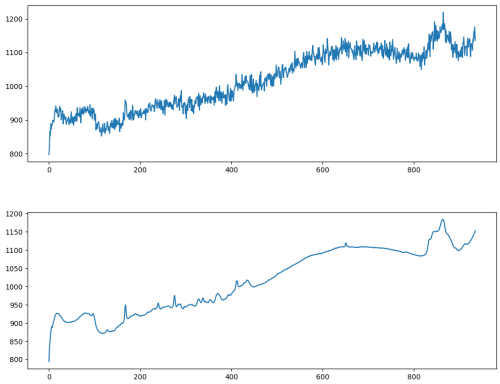


図 8　(上) ノイズを含んだ生データ (下) ノイズ除去後のスペクトル

6.4　類似ピークの検索

指定した波数周辺のピークを持つ測定データを検索し，一覧表示を行う．一覧表示には，その波数の xy 平面上のピーク位置を点で表示した画像と，その面に含まれるピーク数を表示する (図 9)．データベース内にある全測定データについて，この検索を行う．一覧表示された中で確認したい試料を選択すると，その試料でのピーク座標の一覧と各座標のスペクトル一覧を確認できる (図 10)．

7　ユースケース

実際にフォトニクス装置を用いた測定データに対して本データベース基盤を適用する例を説明する．HeLa 細胞株を測定し，その解析を行う。試料をラマンイメージング装置で測定すると，データは 6.1 節で示したように装置専用の分析ソフトウェアによりメタデータを付与されて，Storage Module による API を用いて本データベースにアップロードされる．データは図 11 上で示すように，3 次元テンソルの形式であり，今回の試料では $(x_{max}, y_{max}, \lambda_{max}) = (45, 399, 798)$ となる．アップロードされたデータは AI Module により，自動的に SVD を用いてノイズ除去される．45×399 個のスペクトルを上位か

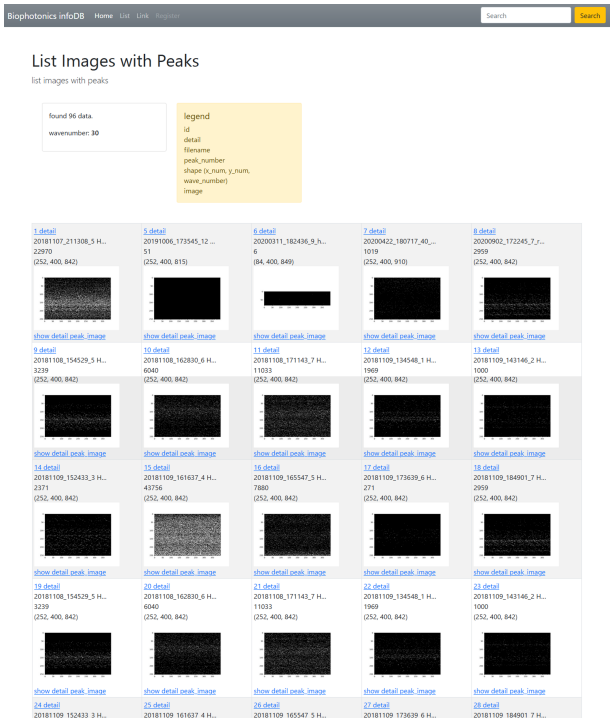


図 9　類似ピークの一覧表示

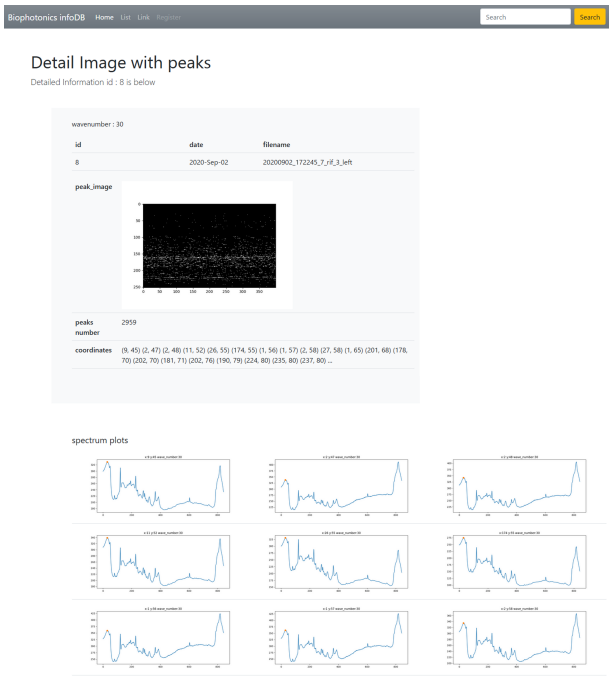


図 10　類似ピークの詳細表示

ら 8 個の特異値で再構成したデータとなっている。このように前処理されてデータベースに保存されたデータから，各生体物質で構成された画像を表示する．HeLa 細胞株のサイトクロム (Cytochrome)，たんぱく質 (Protein)，脂質 (Lipids)，および 3 つに色を割り当てて合成した (Merged) ラマンイメージを図 11 に示す。SVD 処理前と処理後でノイズが除去されて画像が鮮明となり，これらの画像や付与されたメタデータから研究者は効率的に研究を推進できる．以上のように，研究者によるデータ測定・解析の流れに本データベースが有用であることがわかる．

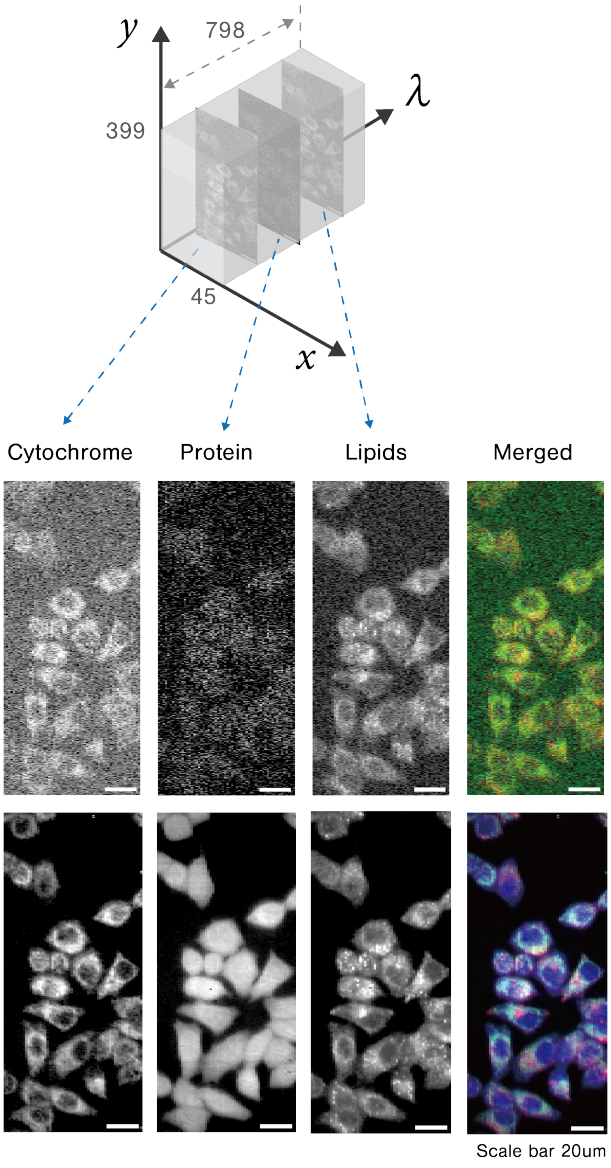


図 11　(上) HeLa 細胞株のデータイメージ (中) 各生体物質でのラマンイメージ．SVD 実施前 (下) SVD 実施後

8　今後の課題

現在データベースの利用はテスト運用を始めたばかりであり，実際に運用する中で，システムとしての堅牢さや処理性能などは，今後発展していく中での課題となると考えられる．現在は表 1 で述べたように，1 台のサーバによる Python, Django, Postgresql を用いたシンプルな構成であるが，データの増大により処理性能に限界が見られた場合, [22] や [23] など分散処理データベース技術の導入を想定している．

また今後は [7]〜[16] のようなリファレンスとしてのデータベースへの展開を考えている，その場合 SSBD [24] 等と同様，一定以上のデータが集まったものをリポジトリとして，その中から人手で有用なデータをキュレーション (選別) する手順の整備も課題となる．

9　おわりに

フォトニクス技術による測定装置からの複数種類の生体光学データの効率的なデータ利用を促進することを目指し，データベース基盤を開発した．複数の生体光学データについてメタデータを定義し，画像を含めた大容量データを保存する API を提供した．表示，検索といったデータベースとしての基本的な機能だけでなく，ノイズの除去，ビーク検索といった発展的な機能により研究者のデータ解析を効率化した．構造化されたデータが集まることで，新しい相関関係や特徴の抽出などをデータマイニングにより発見することも期待できる．特に，癌などの疾患と生体光学データの関係などを見つけることにより，医学や人類の発展に貢献することも大きな目標である．今後の展開としては，リファレンスとしてのデータベースの構築も想定している．

研究者に利用される中で改善の要望をフィードバックして，より利便性を向上させると共に，さらに光学分野・医学の研究者との共同研究をすすめて，データベースから新たな知見を得るための技術を発展させていきたい．

謝辞　本研究は，JST 共創の場形成支援プログラム JPMJPF2009 フォトニクス生命工学研究開発拠点 の支援を受けたものです．

文　献

[1] 尾松 孝茂, 植之原 裕行, 伊東 一良, 青木 泰彦, 小西 毅, 小林 功郎 , 西澤 典彦, 和田 尚也, 森田 隆二, 渡邉 恵理子, : フォトニクス応用物理, vol. 79, no. 8, pp.703-705, 2010.

[2] Kozue Watanabe, Almar F. Palonpon, Nicholas I. Smith, Liang-da Chiu, Atsushi Kasai, Hitoshi Hashimoto, Satoshi Kawata, and Katsumasa Fujita, : Structured line illumination Raman microscopy, Nature Communication Vol.6, 10095, 2015

[3] Kentaro Mochizuki, Yasuaki Kumamoto, Shunsuke Maeda, Masato Tanuma, Atsushi Kasai, Masashi Takemura, Yoshinori Harada, Hitoshi Hashimoto, Hideo Tanaka, Nicholas Isaac Smith, and Katsumasa Fujita High-throughput line-illumination Raman microscopy with multislit detection Biomedical Optics Express Vol. 14, no.3, pp. 1015-1026, 2023

[4] Takahiro Nishimura, Yusaku Takai, Yu Shimojo, Hisanao Hazama, Kunio Awazu: Determination of optical properties in double integrated sphere measurement by artificial neural network based method," Optical Review Vol. 28, No.

1, pp42-47, 2021.
[5] Yu Shimojo, Takahiro Nishimura, Hisanao Hazama, Toshiyuki Ozawa, Kunio Awazu: ”Measurement of absorption and reduced scattering coefficients in Asian human epidermis, dermis, and subcutaneous fat tissues in the 400- to 1100-nm wavelength range for optical penetration depth and energy deposition analysis,” Journal of Biomedical Optics, vol. 25 No.4 pp. 045002 ,2020.
[6] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Kristin Persson: New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, Journal of Materials Research vol. 31, pp. 977–994, 2016.
[7] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2023 update. Nucleic Acids Res. Vol.6, No.51, D1371-1380, 2023
[8] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. Vol. 41(Database issue), D36-42, 2013
[9] Anubhav Jain; Shyue Ping Ong; Geoffroy Hautier; Wei Chen; William Davidson Richards; Stephen Dacek; Shreyas Cholia; Dan Gunter; David Skinner; Gerbrand Ceder; Kristin A. Persson: Commentary: The Materials Project: A materials genome approach to accelerating materials innovation , APL Mater., Vol.1, pp.011002, 2013
[10] Jeffrey C Hoch; Kumaran Baskaran; Harrison Burr; John Chin; Hamid R Eghbalnia; Toshimichi Fujiwara; Michael R Gryk; Takeshi Iwata; Chojiro Kojima; Genji Kurisu; Dmitri Maziuk; Yohei Miyanoiri; Jonathan R Wedell; Colin Wilburn; Hongyang Yao; Masashi Yokochi; ”Biological Magnetic Resonance Data Bank”, Nucleic Acids Research, vol.51, Issue D1, Pages D368–D376, 2023
[11] NIST X-ray Photoelectron Spectroscopy Database, NIST Standard Reference Database Number 20, National Institute of Standards and Technology, Gaithersburg MD, 20899 (2000),
[12] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, Takaaki Nishioka : MassBank: a public repository for sharing mass spectral data for life sciences, Vol. 45, No. 7, pp. 703-714 2010
[13] Cell Image Library, http://www.cellimagelibrary.org/home
[14] Lawson CL, Patwardhan A, et al., Berman HM, Kleywegt GJ, Chiu W., “EMDataBank unified data resource for 3DEM.” Nucleic Acids Res. 44, D396-D403, 2016
[15] Lafuente B, Downs R T, Yang H, Stone N: The power of databases: the RRUFF project. In: Highlights in Mineralogical Crystallography, T Armbruster and R M Danisi, eds. Berlin, Germany, W. De Gruyter, pp 1-30, 2015
[16] SDBSWeb : https://sdbs.db.aist.go.jp (National Institute of Advanced Industrial Science and Technology,date of access)
[17] Labguru, https://www.labguru.com
[18] Labware LIMS, https://www.labware.com
[19] Bauch A., Adamczyk I., Buczek P., Elmer F.-J., Enimanev K., Glyzewski P., Kohler M., Pylak T., Quandt A., Ramakrishnan C., Beisel C., Malmström L., Aebersold R. and Rinn B. : openBIS: a flexible framework for managing and analyzing complex data in biology research BMC Bioinformatics, Vol.12, pp.468 2011
[20] The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351,
[21] Chris Allan, Jean-Marie Burel, Josh Moore, Colin Blackburn, Melissa Linkert, Scott Loynton, Donald MacDonald, William J Moore, Carlos Neves, Andrew Patterson, Michael Porter, Aleksandra Tarkowska, Brian Loranger, Jerome Avondo, Ingvar Lagerstedt, Luca Lianas, Simone Leo, Katherine Hands, Ron T Hay, Ardan Patwardhan, Christoph Best, Gerard J Kleywegt, Gianluigi Zanetti, Jason R Swedlow: OMERO: flexible, model-driven data management for experimental biology. Nature Methods Vol.9, pp.245–253. 2012
[22] Apache Hadoop, https://hadoop.apache.org
[23] Apache Spark, https://spark.apache.org
[24] Yukako Tohsato, Kenneth H. L. Ho, Koji Kyoda and Shuichi Onami : SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena, Bioinformatics, Vol. 32, No. 22, pp. 3471–3479, 2016

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

医療・ヘルスケア (1)

座長:中島 伸介(京都産業大学)
コメンテータ:張 建偉(岩手大学)
2024年2月29日(木) 13:30 〜 15:40 T5-A (オンライン（Zoom Events))

15:15 〜 15:40

[T5-A-5-05]【技術報告】類型化と解釈性を考慮した可視化による自動車事故分析

桑元　凌 (三井住友海上火災保険株式会社ビジネスデザイン部)

交通事故の防止および交通事故発生に伴う被害の低減を目指した取り組みの一つとして、当社事故データを活用し、事故パターンの類型化を行う。類型化した結果の解釈性を高めることで、事故の削減や適切な保険関連商品を開発するなど、社会的意義の大きい取り組みにつなげる。
上記目的を達成するため、下記研究結果を報告する。
・NMFを利用した新しいクラスタリング手法の提案
・ICA等を用いたクラスタリング結果の可視化手法の適用

©一般社団法人日本データベース学会

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 13:30 ~ 15:40 | T5-B オンライン（Zoom Events)

ルート探索

座長:勝沼 聡((株)日立製作所)
コメンテータ:若宮 翔子(奈良先端科学技術大学院大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T5-B-5-01] ランドマークを考慮した経路探索フレームワークの検証
*大滝 啓介[1]、吉村 貴克[1] (1. 豊田中央研究所)

13:50 ~ 14:15
[T5-B-5-02] 引力モデルを用いた経路探索手法の提案
*筒井 夏輝[1]、横山 昌平[1,2,3] (1. 東京都立大学、2. 国立情報学研究所、3. 東京大学)

14:15 ~ 14:40
[T5-B-5-03] 移動手段の混雑を考慮した観光ルートプランニング
*田中 寛之[1]、馬 強[1] (1. 京都工芸繊維大学)

14:40 ~ 15:05
[T5-B-5-04] 3D都市モデルを利用して移動中の景観を考慮する観光経路推薦手法の検討
*斎藤 康平[1]、横山 大作[1] (1. 明治大学)

15:05 ~ 15:30
[T5-B-5-05] 個人化レベルのメリハリを考慮した観光ルートプランニング
*廣澤 凛太郎[1]、宜 鯤[2]、馬 強[1] (1. 京都工芸繊維大学、2. 京都大学経済研究所)

ランドマークを考慮した経路探索フレームワークの検証

大滝　啓介†　　吉村　貴克†

† 株式会社豊田中央研究所　〒 112–0004 東京都文京区後楽 1-4-14 後楽森ビル 10F
E-mail: †{otaki,yoshimura}@mosk.tytlabs.co.jp

あらまし　徒歩移動時によく使われるナビアプリに着目し、最速・最短経路だけではなく、景観の良さや、賑わい、日照条件等を考慮した経路を案内することで、人々の街中での散策を促進することを目指している。本研究では、景観の良さとしてランドマークの見えやすさを採用し、3D 都市モデルを用いて、歩行地点からのランドマークの見えやすさを計算する枠組みを構築した。同時に、距離の短さとランドマークの見えやすさの 2 つの目的を考慮する経路を探索するアルゴリズムを実装した。実装したアルゴリズムで、東京タワー周辺の地図を元に、「距離が最小になる経路」と「東京タワーが見られる地点が最も多い経路」の 2 経路を出力し、実際に現地で散策体験した。期待通り「距離が最小になる経路」よりも「東京タワーが見られる地点が最も多い経路」の方が東京タワーがよく観測された。その一方で、ランドマークが進行方向とは逆方向にあり見えにくい、ランドマークを気にしながら歩くと不安全な道がある等、いくつかの課題が見つかった。

キーワード　歩行者ナビゲーション、ランドマーク、デジタルツイン、景観

1　はじめに

移動行動は、人との交流や、情報・モノの獲得のための基本的な行動であり、健康維持や幸福感とも関連する [2,3]。しかし、近年のオンライン会議サービスやフードデリバリーサービスの爆発的普及により、人々の移動行動が減少しつつあり、さらには人々が集う街の持続的な発展にも悪影響を与える可能性がある。我々は、徒歩移動時によく使われるナビアプリに着目し、最速・最短経路だけではなく、景観の良さや、賑わい、日照条件等を考慮した街を回遊する経路を案内することで、人々の街中での散策を促進することを目指している [5,6]。

街を回遊する経路を案内する回遊ナビでは、街中での移動を様々な評価軸で楽しめる経路を提示することが望まれる。複数の評価軸を考慮した経路を求めるためには、多目的な評価関数を持つ経路探索を行う必要がある [4,7,8]。また、主観的な要素を考慮する経路の評価システムが必要となる。既存の API などを用いる場合、主要な指標 (距離や時間など) を扱ったナビゲーションシステムを構築することができるが、複雑な枠組みに従うナビゲーションシステムを構築するには、全体のシステムを実装する必要がある。

本稿では、通りの景観の良さを評価する一つのアプローチとして、高い建造物等の地域の象徴的なランドマークの見えやすさに着目し、3D 都市モデル [1] を用いて、ある歩行地点におけるランドマークの見えやすさを評価する枠組みを提案する。ランドマークナビ自体はこれまでも研究されてきたが [9,10]、3D 都市モデルを活用した例は少ない。本稿では、客観的な評価指標として移動距離を、主観的な評価指標の代理指標としてランドマークの見えやすさを用いて、二目的最適経路探索を行い、2 地点間の様々な経路を計算してユーザーに提示する。ユーザーにランドマークが見える経路を歩かせることで、道の景観を印象付け、土地勘を身に着けさせる効果も同時に狙っている。回遊ナビのシステムとしては、さまざまな経路を計算することで、個人に適合した経路推薦を提供することを目標としている。本稿ではこれらのシステムの構成と、簡易的に実施した実地での回遊体験の結果、および考察を述べる。

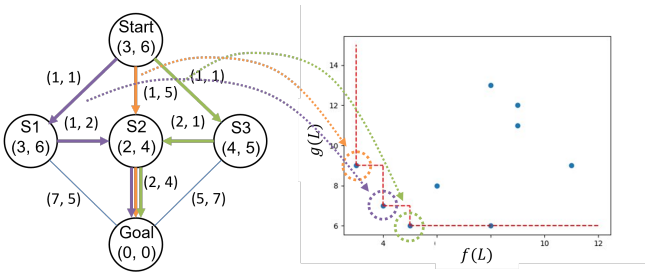


図 1: 目的関数が 2 つ (f, g) の場合の多目的経路探索の例。頂点内のペアは計算に用いるヒューリスティック関数値を示す。

2　多目的経路探索

辺コストが複数付与された無向グラフ $G = (V, E, \vec{c})$ において、2 点 $v_s, v_g \in V$ をつなぐ経路 π_{v_s,v_g} のうち、コストベクトルが他の解に支配されない経路を探索する問題を多目的経路探索と呼ぶ。図 1 に多目的経路探索の例を示す。図 1(左) に図示した頂点数 5 のグラフ G を対象とし、$v_s = Start$ から $v_g = Goal$ までの経路を探索する。辺 e には 2 次元のコストベクトル $(f(e), g(e))$ が付与されており、例えば $Start$ から s_1 への辺はコストベクトル $(1, 1)$ を要する。例えば紫色の経路 $(Start \to s_1 \to s_2 \to Goal)$ は、辺に付与されたコストベクトルの和によって、コストベクトル $(4, 7)$ を要する経路である。これを (f, g) の二次元プロットしたものが図 1(右) であり、図 1(右) の散布図の各点は何らかの経路を表す。多目的経路探索

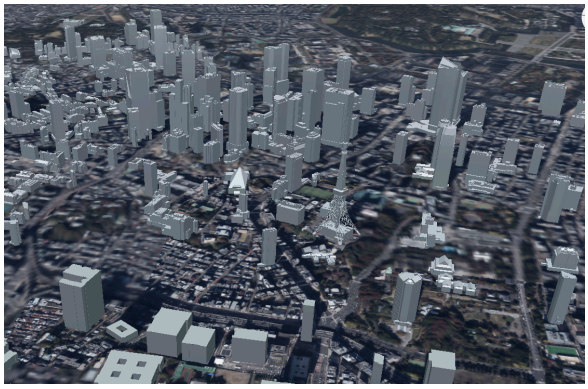


図 2: 東京タワー周辺で可視化したモデル

の技術的な詳細は [7] などを参照されたい。本稿では [8] で提案された BOA*アルゴリズムを実装して用いた。上記の例の場合、経路復元を実装した BOA*アルゴリズムは他の経路に支配されない解（図 1 に示す 3 色の経路）を出力する。

3 3D 都市モデルを用いた景観評価

近年、都市の設計や情報サービス等への応用を促進するために、3D 都市モデルが公開されている [1]。本稿ではこの 3D 都市モデルを用いて、景観評価を行うシステムを提案する。公開されているデータをビューワなどで可視化すると、図 2 に示すような建物オブジェクトがいくつかの詳細レベル [1]で得られる。景観には複数の要素が関与するが、本稿では特に大規模なランドマーク (例えば東京タワーなど) を対象とする。本稿では対象地域のモデルを用いて、以下の形で景観評価を行い、グラフ G の辺に情報を付与する。

定義 1. グラフ G 上の頂点 $v \in V$ から、3D 都市モデルで対象とするランドマークが見えるかどうか判定した数値を h_v とする。ただし 3D 都市モデルにおいてランドマークが見える場合 $h_v = 1$、見えない場合 $h_v = 0$ とする。辺 $e = \{u, v\} \in E$ の景観評価コストとして、$h_v = h_u = 1$ のときは $g(e) = 1$、それ以外の場合には $g(e) = 2$ とする。

本稿では東京駅周辺を対象環境として、ランドマークには東京タワー [2]を用いた。具体的な実装は Blender[3] を用いて以下のように実施した。図 3 は h_v を対象地域のグラフ構造で計算した結果を可視化したものである。

（1） 頂点 $v \in V$ の緯度経度 (lat_v, lon_v) を Plateau の 3D 都市モデル座標系に変換し、東京タワーの座標に向けて光線を打つ。
（2） 光線の当たったオブジェクトが東京タワーであるときに $h_v = 1$ と設定し、それ以外の場合に $h_v = 0$ とする。

1：LOD (Level of Detail) と呼ばれる。
2：Plateau データのうち、LOD2 モデルから東京タワーに属するモデルを抽出して一体化して作成
3：https://www.blender.org/

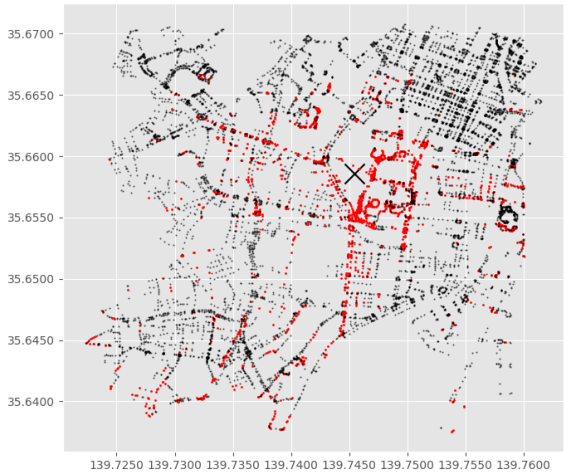


図 3: 東京タワー周辺の頂点評価値 (中央の × 印が東京タワー)

上記の定義に従うことで、評価値 $g(e)$ が最小となる経路を提示することで、ランドマークが見える回数が間接的に最大化されることが期待される。なお多目的経路探索の 1 つ目の評価値 $f(e)$ については、2 頂点の緯度経度から Haversine の公式から求めた距離を設定した。そのため f, g の二目的を考慮することで、距離とランドマークの見える度合いをそれぞれ考慮した経路が得られると期待される。

4 実験と評価

本稿で提案したランドマークを考慮する経路案内手法を用いて、東京タワー周辺の経路を 2 つの実験例で比較する実験を行った。

4.1 実 験 設 定

東京タワー周辺の道路ネットワーク構造を OpenStreetMap[4] より取得し、前処理を行って $|V| = 11816, |E| = 14239$ のグラフ構造を作成した。ナビゲーションの具体的な問題設定として、以下の 2 つの経路案内例を設定した。
（1） 三田 – 新橋
（2） 新橋 – 西久保八幡神社
ただし大まかな各地点の緯度経度は以下の通りである。

- 三田: $(35.6494, 139.7472)$ の地点。
- 新橋: $(35.6555, 139.7596)$ の地点。
- 西久保八幡神社: $(35.6612, 139.7434)$ の地点。

それぞれのナビゲーションシナリオについて、BOA*アルゴリズムを適用して解を求め、「距離が最小になるもの」と「最もランドマークが見られるもの」を両極端な経路を出力した。実際に出力された経路を 2 名で歩行して、対象ランドマークに設定した東京タワーが歩行中に見えるかどうか、またどのような景観の道になっているかを確認し、最後に感想をまとめた。

4.2 結　　果

設定 1 の実験における経路や景観をまとめたものを図 4 に、

4：https://www.openstreetmap.org/

図 4: 設定 1 の実験結果まとめ: 最短経路 (赤) とランドマークを考慮した経路 (青) およびランドマークの写真。

図 5: 設定 2 の実験結果まとめ: 最短経路 (赤) とランドマークを考慮した経路 (青) およびランドマークの写真。

設定 2 の実験における結果を図 5 にそれぞれ示す。図中では赤色の経路が「距離が最小になるもの」であり、青色の経路が「最もランドマークが見られるもの」である。

4.3 評価と感想のまとめ

多目的経路探索を実装したことで、図 4 や 5 に示すように、ランドマークが見える可能性が高い場所を通過する道を求めることができた。これを客観的に評価したり、数値評価へと一般することは今後の課題である。また実験参加者から、回遊経路、及び経路の評価方法に関する以下のような意見を得た。

- ランドマークが見えることにより、迷いなく移動することができた。
- ランドマークの見え方 (方向など) や時間など、改善することでよりよい体験が得られると思う。
- 歩きやすい道があると良い (例えば商店街の道や、自転車と分離されている歩道)。
- 歩行時の状況によって (例. 傘をさしている)、ランドマークを気にする余地がなく、効果を感じないことがあった。
- 長い距離を複数回歩いて比較することが体力的に難しい場合があると思う。
- また似ている経路を主観的に比較するのが難しい。

今後は複数のランドマークをを用意したり、定義 1 の評価関数を工夫することで、より覚えやすい経路や楽しめる経路などが出力できるようになると期待される。

5 ま と め

本稿では経路に沿った景観の良さを評価する一つのアプローチとして、ランドマークの見えやすさを採用し、3D 都市モデルを用いてこれを評価する枠組みを提案した上でコスト設計を行い、BOA*アルゴリズムを適用することで経路を出力した。また本アルゴリズムにより得られた 2 経路を実際に歩行し、ランドマークが観測しやすい経路が得られることと、二つの目的 (経路の短さ、ランドマークの見えやすさ) を同時に最適化する経路が得られることを確認した。

今後の課題として、今回の実地体験で課題として挙がったランドマークの見え方や道の歩きやすさ考慮がある。さらには今回は 2 つ目的関数の最適化に限定しているが、3 目的関数以上を考慮することや、ユーザ評価・実地評価などを行うことが考えられる。また今後の研究では、本研究で構築した手法を拡張し、主観的な評価を取り組んだナビゲーションシステムのサービス構築を目指したい。

謝　　辞

弊社のインターンシッププログラムを通じて本研究に協力頂いた、東京大学の住田圭吾氏に感謝いたします。

文　　献

[1] PLATEAU by MILT (オンライン) https://www.milt.go.jp/plateau (accessed 2023-10-18).
[2] Aaron S Heller, Tracey C Shi, CE Chiemeka Ezie, Travis R Reneau, Lara M Baez, Conor J Gibbons, and Catherine A Hartley. Association between real-world experiential diversity and positive affect relates to hippocampal–striatal functional connectivity. *Nature Neuroscience*, Vol. 23, No. 7, pp. 800–804, 2020.
[3] Richard Layard and Jan-Emmanuel De Neve. *Wellbeing: Science and Policy*. Cambridge University Press, 2023.
[4] Lawrence Mandow and José-Luis Pérez-de-la-Cruz. A new approach to multiobjective A* search. In *Proc. of IJCAI2005*, pp. 218–223, 2005.
[5] Keisuke Otaki, Tomosuke Maeda, Takayoshi Yoshimura, and Hiroyuki Sakai. Roaming navigation: Diverse constrained paths using heuristic search. *IEEE Access*, Vol. 11, pp. 75617–75627, 2023.
[6] Keisuke Otaki, Ai Nakada, Tomosuke Maeda, Takayoshi Yoshimura, and Hiroyuki Sakai. Roaming navigation for pedestrians (demo paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL' 23)*, 2023.
[7] Zhongqiang Ren, Richard Zhan, Sivakumar Rathinam, Maxim Likhachev, and Howie Choset. Enhanced multi-objective A* using balanced binary search trees. In *Proc. of SOCS2022*, pp. 162–170, 2022.
[8] Carlos Hernández Ulloa, William Yeoh, Jorge A Baier, Han Zhang, Luis Suazo, and Sven Koenig. A simple and fast bi-

objective search algorithm. In *Proc. of ICAPS2020*, Vol. 30, pp. 143–151, 2020.

[9] 若宮翔子, 森永寛紀, 岡山愛, 脇海晟, Adam Jatowt, 河合由起子, 秋山豊和, 川崎洋. Landmarknavi: マイクロブログを用いた効果的なランドマーク発見. 日本データベース学会和文論文誌, Vol. 15-J, pp. 1–8, 2017.

[10] 森永寛紀, 若宮翔子, 谷山友規, 赤木康宏, 小野智司, 河合由起子, 川崎洋. 点と線と面のランドマークによる道に迷いにくいナビゲーション・システムとその評価. 情報処理学会論文誌, Vol. 57, No. 4, pp. 1227–1238, 2016.

引力モデルを用いた経路探索手法の提案

筒井　夏輝†　　横山　昌平††,†††,††††

† 東京都立大学 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所　〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †tsutsui-natsuki@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　ソーシャルデータを利用した経路探索についてはさまざまな手法が存在する．しかし，これまでの研究では経由地を指定して経路探索を行う研究と経由地となりうるスポットの探索は別々の研究として行われていた．本論文では地図データとジオソーシャルビッグデータを併用することにより始点と終点およびユーザの好みを与えるだけで好みに応じた経由地を持つ経路を発見できる．具体的には，経由地検索と経路探索を同時に解くために，ユーザがソーシャルデータの集まる点群に引き寄せられるような引力モデルを提案する．また，実験にて OSM の任意の 2 つのノードについて，Flickr のジオラベル付き投稿データや人工の点群データを用いて経路の探索を行い，提案手法の有効性を示した．

キーワード　ジオソーシャルデータ，地理情報，経路推薦，Flickr，SNS

1 はじめに

スマートフォンの普及により，旅行者が自分の気に入ったスポットの写真を，気軽にソーシャルメディア上に共有することができるようになった．また，ソーシャルメディア上にアップロードされたソーシャルデータのうち位置情報が付与されたものをジオソーシャルデータと呼び，これらを用いた様々な分析や研究が行われている [5] [6] [7]．そのなかでも，それらのデータを用いてユーザの好みに応じた経路の探索，推薦を行う研究が盛んである．我々はこれまでに，ジオソーシャルデータを用いたユーザの好みに応じた経路探索手法を提案している [10]．しかし，これまでの研究では利用者の複雑な好みに対応する経路の探索は行われていない．例えば，「人ごみの多い地域を避けながらお寺を見て，その後，景色の良い場所を寄りたい．しかし途中の経路はなるべく海沿いを歩きたい」のようなユーザの複雑な好みには応じることができない．

そこで本論文では，経路探索に引力モデルを統合することにより，ユーザが与えた始点と終点，および上記のような複雑なユーザの好みに応じた経路探索手法の提案を行う．本論文で提案する引力モデルを用いた経路探索アルゴリズムは次の特徴を有している．

- 提案アルゴリズムでは経由地の検索と経由地を含む経路の探索に，DBSCAN とダイクストラ法のように別々のアルゴリズムを適用する必要がなく，これらの問題をひとつのアルゴリズムで解くことが可能である．
- 提案アルゴリズムはラベル付けをしたジオタグ付きの投稿を経路推薦に利用することができる．これにより，ユーザの好みを経路推薦に反映することが可能である．また，計算に必要な情報はラベルと座標情報のみであるため，多様なデータソースの結合が容易である．
- 提案アルゴリズムではジオソーシャルデータとユーザの現在位置から次の移動先を決定する．これにより，点群を通る，あるいは避けるといった経路をラベルごとに指定することが可能である．また，どのようなラベルの付いた点群をどの順番で通りたいかも指定することが可能である．

また，実験として人工データや flickr から取得したジオラベル付き投稿データ，OpenStreetMap(OSM) に基づいて引力モデルを構築し，経路の探索，有効性の検証を行う．

2 関連研究

ソーシャルデータを用いたスポット探索や経路探索の研究は古くから行われており [1] [2] [3]，これらのアルゴリズムを拡張しユーザの好みに合うような経路を推薦するような研究が多く行われている．

山下ら [10] はグリッドマップと Flickr に投稿された写真の緯度経度や時間などのメタデータを用いて，道路ネットワークに重みづけを行いダイクストラ法を用いてユーザの好みに最適な経路の推薦を行った．しかし，この手法ではユーザの到達順序の好みが指定できない点や，ユーザの通りたいという好みと避けたいという好みを同時に内包できない点で異なる．

夏ら [11] は Flickr に投稿されたメタデータを用いて，DBSCAN によりスポットを分け，ユーザの移動を道路ネットワークの重みへと変換し，ワーシャル-フロイド法により現在地と目的地の間の最短経路を推薦した．しかし，この手法ではユーザの好みを指定できないという点で異なる．

Kurashima ら [4] は Flickr での投稿のタイムスタンプ，移動記録を用いて，旅行経路を推薦している．ユーザの好みに応じた条件を設定することでランドマークを経由する経路を生成し，

経路の推薦を行う．この手法では位置情報の履歴が必要であるが，提案手法ではこれらの情報を用いない点で異なる．また，ユーザの到達順序の好みが指定できないという点でも異なる．

3 提案手法

本章では提案手法について説明する．引力モデルは，道路ネットワーク，ジオタグとラベルの付けられた点 (以降，投稿点と呼ぶ) の 2 種類のデータに基づいて構築される．ユーザは道路ネットワーク上のある始点から始め，各投稿点から力を受け，受けている力の総和から次の移動ノードを決定する．これを終点ノードにたどり着くまで繰り返す．

3.1 問題設定

Flickr や Twitter などのジオソーシャルデータでは，付与されたテキストや画像などのメタデータが，結び付けられたジオタグの地理的特徴を示すことがいくつかの研究で実証されている [8] [9]．具体的には，実際のビーチや海岸線では”beach”や”sea”などの単語が含まれたテキストが頻出する．これらの前提のもと，本論文では同一のラベルを持つ投稿点は同一の地点にまとまるように人工データの生成を行った．これらの投稿点と取得した道路ネットワークに基づいて，引力モデルを構築し，ユーザの複雑な好みに応じた経路の探索を行う．具体的にはユーザの避けたいラベルや通りたいラベル，さらにはどのような順番でどんなラベルを通りたいかなどの複合的な条件に応じた経路探索を行う．

本論文では引力モデルを用いた経路探索手法の提案に焦点をあてたものであり，各投稿点についてどのようにラベルを付与するかという点は本論文の対象外とする．

3.2 前処理

取得した道路ネットワークを無向グラフとして扱い，自己ループ，多重辺がなくなるようエッジを削除し単純グラフへと変換する．変換後，次数 1 となるようなノードと関連するエッジを削除することで，端点の存在しない単純グラフへと変換する．その後，終点からダイクストラアルゴリズムを適応し，各ノードにおける終点ノードを目的地とした最短経路方向のノード (以降，親ノードと呼ぶ) を取得する．

このような前処理をした後のグラフを $G = (V, E)$ で表す．V は道路ネットワークのノードの集合を，$E \subseteq V \times V$ はエッジの集合である．道路ネットワークのノード $v \in V$ の v は交差点を表す．道路ネットワークのエッジ $\{u, v\} \in E$ はノード u と v の間のエッジであり，道路を表す．また，次数 1 のノードと関連するエッジを削除しているため単純グラフ G のノード v は $deg(v) \geqq 2$ を満たす．

3.3 経路探索

引力モデルの経路探索では，始点ノードから始めて，終点ノードにたどり着くまで以下の手順を行う，

（1） 受けている力の計算
（2） 移動先のノードの決定
（3） 各投稿点，ラベルの重みの再計算
.

3.3.1 受けている力の計算

ユーザが受ける力 $\boldsymbol{F}$ は，親ノードと各投稿点から受ける力の総和として表す．また，今回の研究では各投稿点，ノードの緯度経度を 2 次元平面上の点として計算している．

$$\boldsymbol{F} = \boldsymbol{F}_s + \sum_{i=1}^{N} \boldsymbol{F}_{p_i} \tag{1}$$

この式において $\boldsymbol{F}_s$ は親ノードから受ける力を示しており，現在のノードから親ノード方向へ働く大きさ 1 の単位ベクトルである．N は投稿点の数を示している．また，$\boldsymbol{F}_{p_i}$ は各投稿点から受ける力を示しており，現在のノードから各投稿点方向へ働く力である．現在のノードの位置ベクトルを $\boldsymbol{x}$，投稿点の位置ベクトルを $\boldsymbol{p}$ とおくと，各投稿点から受ける力 $\boldsymbol{F}_p$ は以下の式で表される．

$$\boldsymbol{F}_p = GMm\frac{\boldsymbol{p} - \boldsymbol{x}}{|\boldsymbol{p} - \boldsymbol{x}|^3} \tag{2}$$

この式において G はパラメータであり，親ノードから受ける力と各投稿点から受ける力の比を調整するパラメータとして働く．M は投稿点の属するラベルの重み，m は投稿点の重みを表す．これらの重みは各ラベル，各投稿点の影響力を示しており初期値として 1 を与えられ，ユーザがノードを移動する度に増減していく．r は現在のノードと投稿点の距離を示しておりユーザが受ける力は現在のノードから各投稿点までの距離の 2 乗に反比例する．

3.3.2 移動先のノードの決定

移動先のノードは「受けている力のベクトル」と「現在のノードから隣接ノードへの方向ベクトル」の cosine 類似度が最も大きいノードとする．ここで，受けている力のベクトルを $\boldsymbol{F}$，ノード i から j の方向ベクトルを $\boldsymbol{v}_{i,j}$，現在のノードを i としたとき，移動先のノード k は以下の数式で表せる．

$$k = argmax_j \frac{\boldsymbol{F} \cdot \boldsymbol{v}_{i,j}}{|\boldsymbol{F}||\boldsymbol{v}_{i,j}|} \quad (\{i, j\} \in E) \tag{3}$$

3.3.3 各投稿点，ラベルの重みの再計算

提案手法では避けたいラベル，通りたいラベル，順番に通りたいラベルのそれぞれで重みの計算に違いがある．ここで「通りたいラベル」というのは経路中どの順序で通るかを指定しないラベルであり，「順番に通りたいラベル」というのはどういった順序で通りたいかを指定されたラベルであるとする．それぞれのラベルの重みの計算方法を以下に示す．

a） 避けたいラベル

避けたいラベルの場合，投稿とラベルの重みを変化させる必要がない．しかし，他のラベルと違い避けたいラベルの重み M の初期値に –1 を割り当てる必要がある．そうすることで，ユーザは避けたいラベルが付与された点群とは逆の方向に力を受けるため，移動先のノードに点群を避けるようなノードを選ぶようになる．

b） 通りたいラベル

通りたいラベルの場合，投稿の重みを減少させる必要がある．引力モデルの性質上，点群に近づくほどユーザが点群から受ける力が強くなる．そのため，投稿の重みを減少させない場合，ユーザが点群から抜けだせなくなってしまう．

投稿の重みは以下の数式を用いて再計算される．

$$d = c\frac{|\theta' - \theta|}{r^2} \tag{4}$$

$$m' = m - \min(m, d) \tag{5}$$

この式において d は投稿の重みの最大変化量， m, m' は移動の前後の各投稿の重み，θ, θ' を移動の前後のノードと各投稿の成す角，c は重みの減少速度を調整するパラメータである．m は初期値として 1 が与えられ， m, m' は 0 より小さくならない．これは，投稿の重みが減少してもユーザに力を与えなくなるよう変化していくだけで逆方向に力を働かせるわけではないということである．また，通りたいラベルの重み M は初期値として 1 を与える．

c） 順番に通りたいラベル

順番に通りたいラベルの場合，ラベルの重みを増減させることで指定した順番でラベルの点群を通るような経路を探索する．順番に通りたいラベルを n 個指定したとする．その際，ユーザが通りたい順番にラベルが並んでいるとし，ラベルの重みの初期値を以下のように与える．

$$M_i = \begin{cases} 1 & (i = 1) \\ 0 & (i = 2, \cdots, n) \end{cases} \tag{6}$$

各ラベル i に属する投稿の集合を L_i, ラベルに属する投稿の数を $n(L_i)$ とし，以下の数式に従って各ラベルの重みを増減させる．

$$D_i = \frac{1}{n(L_i)} \sum_{j \in L_i} (m_j - m'_j) \tag{7}$$

$$M'_i = M_i - \min(M_i, D_i) + \min(M_{i-1}, D_{i-1}) \tag{8}$$

この数式において，D_i はラベルに属する投稿点の重みの減少率とラベルの重みの最大変化量を表しており，M_i, M'_i はノードの移動前後の各ラベルの重みの大きさを表している．数式より，ラベルに属する投稿点の重みが減少するほどラベルの重みが減少していき，ラベルの重みが減少するほど次のラベルの重みが増加することが分かる．また，順番に通りたいラベルの重みは投稿点の重みと同様に負の値を取らない．

4 実　　験

提案手法は，引力モデルによって，ユーザの複雑な好みを反映させた経路を探索ができることが特徴である．経路探索によって提案された経路がユーザの好みを反映しているかについては，基本的にユーザの主観的な評価に依る．本論文では人工データを用いた 3 つのシナリオと flickr から取得した現実のソーシャルデータを用いた 2 つのシナリオにおいてを用いて経路探索を行い，提案手法の定性的な評価を行う．

また，投稿点の数に対して実行時間がどのように変化するかを人工データを用いて検証した。

シナリオ 1　避けたいラベルの投稿点の点群のみを生成し，始点，終点を設定した時，点群を避けるような経路を探索できるか．

シナリオ 2　避けたいラベルと通りたいラベルの投稿点の点群をそれぞれ生成し，始点，終点を設定した時，避けたいラベルの点群を避けながら通りたいラベルの点群を通るような経路を探索できるか．

シナリオ 3　避けたいラベルと通りたいラベル，順番に通りたいラベルの投稿点の点群をそれぞれ生成し，始点，終点を設定した時，避けたいラベルの点群を避けながら通りたいラベルの点群を通るような経路を探索できるか．

シナリオ 4　ニューヨーク州のマンハッタン区の一部における現実のソーシャルデータを用いて通りたいラベルの投稿点を取得し，始点，終点を設定した時，通りたいラベルの点群を通るような経路を探索できるか．

シナリオ 5　京都府京都市の一部における現実のソーシャルデータを用いて通りたいラベルの投稿点を取得し，始点，終点を設定した時，通りたいラベルの点群を通るような経路を探索できるか．

4.1 データセット

4.1.1 道路ネットワーク

本論文では OSM の道路ネットワークデータを利用しており，ニューヨーク州のマンハッタン区と京都府京都市の一部の道路ネットワークを取得した．正確には，シナリオ 1,2,3,4 では緯度 40.745 〜 40.767, 経度 -73.996 〜 -73.970 に含まれる道路ネットワークを用いている．経路探索の始点ノードの座標は (40.745, -73.997)，終点ノードの座標は (40.767, -73.971) とした．

シナリオ 5 では経度 34.984 〜 35.005, 緯度 135.753 〜 135.784 に含まれる道路ネットワークを用いている．経路探索の始点ノードの座標は (34.986, 135.758)，終点ノードの座標は (35.003, 135.778) とした．

4.1.2 投　稿　点

シナリオ 1,2,3 の投稿点には人工データを用いており，あらかじめラベルと中心座標を指定し，指定したバウンディングボックス内で点群を生成した．生成の際，中心座標を基準にして標準偏差 $\sigma = 0.0005$ のガウス分布に従うように点群を生成した。

シナリオ 4,5 の投稿点では，同シナリオで取得した道路ネットワークと同じ範囲の 2020-01-01 〜 2023-01-01 での投稿を flickr から取得している．投稿点として使用したデータ件数はシナリオ 4 では 819 件，シナリオ 5 では 501 件であった．取得した投稿のタグの中でも投稿件数が多く，”temple”や”cityscape”のようにユーザの好みを反映していると考えられるタグを選び，選んだタグが付いている投稿のみを投稿点とした．

4.2 シナリオ 1 における経路探索

4.2.1 問 題 設 定

本節ではシナリオ 1 として，最短経路に避けたいラベルの点

図 1　ダイクストラ法による最短経路 (ピンク色) と引力モデルによる負の引力を持つ点群を避けるような経路 (赤色)，避けたいラベルの投稿点 (赤色)

図 2　ダイクストラ法による最短経路 (ピンク色) と引力モデルによる避けたいラベルの点群を避けつつ通りたいラベルの点群を通るような経路 (赤色)，避けたいラベルの投稿点 (赤色)，通りたいラベルの投稿点 (青色)

群が存在するような始点と終点を設定し，点群を避けるような経路を探索できるかを評価する．

4.2.2　実験結果と分析

図 1 にダイクストラ法を用いて計算した最短経路，シナリオ 1 に基づいて生成したデータと引力モデルを用いて経路探索を行った結果を示す．

このように提案手法では，避けたいラベルの点群を通らない経路を推薦出来ており，シナリオ 1 に挙げた最短経路に避けたいラベルの点群を通るような始点と終点を設定した場合に，点群を避けるような経路を探索できたことが分かる．

4.3　シナリオ 2 における経路探索

4.3.1　問 題 設 定

本節ではシナリオ 2 として，最短経路に避けたいラベルの点群が存在するような始点と終点を設定し，さらに通りたいラベルの点群も存在するような場合に，避けたいラベルの点群を避けながら通りたいラベルの点群を通るような経路を探索できるかを評価する．

4.3.2　実験結果と分析

図 2 にダイクストラ法を用いて計算した最短経路，シナリオ 2 に基づいて生成したデータと引力モデルを用いて経路探索を行った結果を示す．

実験結果より，シナリオ 1 と違い今回は避けたいラベルの点群の北西方向に通りたいラベルの点群が存在するため違った経路を探索によって得ることができた．また，北東方向の通りたいラベルの点群についても経路に含まれていることが分かる．

結果から，シナリオ 2 に挙げた最短経路に避けたいラベルの点群を通るような始点と終点を設定し，さらに通りたいラベルの点群も存在するような場合に，点群を避けながら通りたいラベルを通るような経路を探索できていることが分かる．

4.4　シナリオ 3 における経路探索

4.4.1　問 題 設 定

本節ではシナリオ 3 として，最短経路に避けたいラベルの点群が存在するような始点と終点を設定し，さらに通りたいラベルの点群，順番に通りたいラベルの点群も存在するような場合に，避けたいラベルの点群を避けながら通りたいラベルの点群を通るような経路を探索できるかを評価する．

4.4.2　実験結果と分析

図 3 にダイクストラ法を用いて計算した最短経路，シナリオ 3 に基づいて生成したデータと引力モデルを用いて経路探索を行った結果を示す．

シナリオ 3 の実験結果より，シナリオ 2 で得られた結果に加えて緑色，黄色の投稿点の点群が経路に含まれていることが分かる．また，順番に通りたいラベルでは緑色を最初にしているため最初の青色の点群から近い黄色の点群ではなく，緑色の点群を通過してから黄色の点群を通過していることが分かる．

結果から，シナリオ 3 に挙げた最短経路に避けたいラベルの点群を通るような始点と終点を設定し，さらに通りたいラベル，順番に通りたいラベルの点群も存在するような場合に，点群を避けながら通りたいラベルを通るような経路を探索できていることが分かる．

4.5　シナリオ 4 における経路探索

4.5.1　問 題 設 定

本節ではシナリオ 4 として，ニューヨーク州のマンハッタン

図 3　ダイクストラ法による最短経路 (ピンク色) と引力モデルによる負の引力を持つ点群を避けつつ正の引力を持つ点群を寄るような経路 (赤色)，避けたいラベルの投稿点 (赤色)，通りたいラベルの投稿点 (青色)，順番に通りたいラベルの投稿点 (順番に緑色，黄色)

図 4　ダイクストラ法による最短経路 (ピンク色) と引力モデルによる正の引力を持つ点群を通るような経路 (赤色)，通りたいラベルの投稿点 (青色)

区の一部における現実のソーシャルデータを用いて通りたいラベルの投稿点を取得し，始点，終点を設定した時，通りたいラベルの点群を通るような経路を探索できるかを評価する．

4.5.2　実験結果と分析

図 4 にダイクストラ法を用いて計算した最短経路，シナリオ 4 に基づいて取得したデータと引力モデルを用いて経路探索を行った結果を示す．

結果から，シナリオ 4 に挙げたような現実のソーシャルを用いた場合でも，途中の通りたい点群を通って目的地に向かっていることがわかる．しかし，通っていない点群が多く経路探索として十分であるとはいえない．

4.6　シナリオ 5 における経路探索

4.6.1　問 題 設 定

本節ではシナリオ 5 として，京都府京都市の一部における現実のソーシャルデータを用いて通りたいラベルの投稿点を取得し，始点，終点を設定した時，通りたいラベルの点群を通るような経路を探索できるかを評価する．

4.6.2　実験結果と分析

図 5 にダイクストラ法を用いて計算した最短経路，シナリオ 5 に基づいて取得したデータと引力モデルを用いて経路探索を行った結果を示す．

結果から，シナリオ 5 に挙げたような現実のソーシャルを用いた場合，多くの通りたい点群を通りながら目的地に向かっていることがわかる．しかし，十分に近づかずに点群に到達する前にほかの点群に向かっている様子がわかり，これもシナリオ

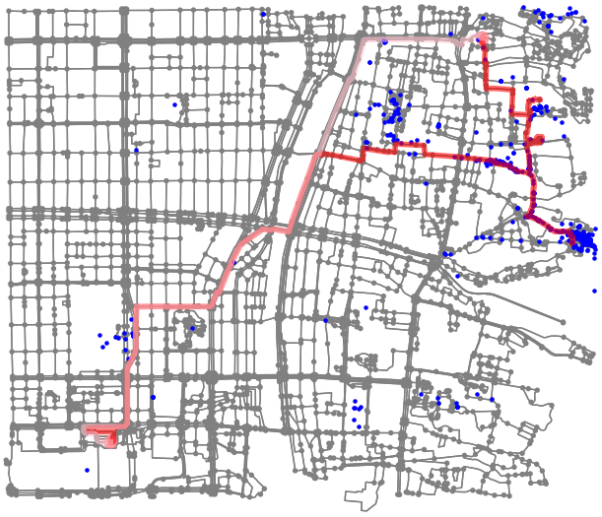


図 5　ダイクストラ法による最短経路 (ピンク色) と引力モデルによる負の引力を持つ点群を避けるような経路 (赤色)，通りたいラベルの投稿点 (青色)

4 同様十分な経路探索であるとはいえない．

4.7　実 行 時 間

実験として図 6 のように，マンハッタン区の道路ネットワークに対して投稿点を生成し、投稿点の数を増やしていくことで投稿点数に対する実行時間の関係を調べた．

また，投稿点数と実行時間の関係を図 7 に示す．

実験結果より，引力モデルでは投稿点数に比例して実行時間が伸びていくことが分かった．引力モデルでは，ノードの一回の移動に対してすべての投稿点からの引力を計算する．そのため投稿点に比例して計算時間が伸びたと考えられる．

図 6　投稿点数と実行時間比較のための道路ネットワークと生成した点群

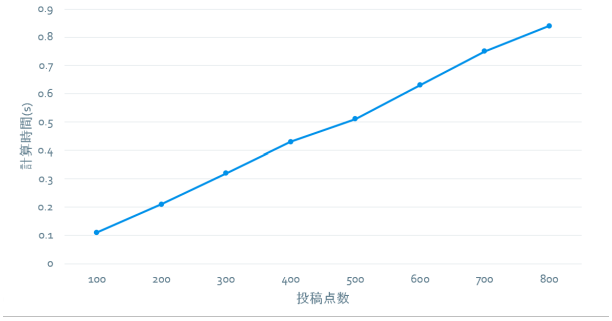


図 7　投稿点数と実行時間の関係

5　まとめと今後の課題

本論文は，OSM と人工データに基づいて，引力モデルを構築し，ユーザが与えた始点と終点，およびユーザの好みに応じた経路の探索を行うアルゴリズムの提案に焦点を当てたものである．提案手法では各投稿点がユーザに力を与え，次の移動先のノードを決定し，経路の探索を行うというものである．

実験として避けたいラベルを持つ投稿点や通りたいラベルを持つ投稿点に応じた経路の探索を行った．実験としてニューヨーク州のマンハッタン区，京都府京都市の一部の道路ネットワークと人工データ，ソーシャルデータを用いて提案手法と最短経路の比較を行い定性的な評価を行った．

実験の結果から，ガウス分布に従うように生成した人工データを用いた場合であれば，有効な結果が得られることが分かったが，現実のソーシャルデータに対しては改善の必要がある．また，今回の実験では結果に対して定性的な評価を行ったが，実験結果を定量的に評価する枠組みが必要であると考える．

提案手法では，各投稿点からの引力を移動ノードごとに計算している．このアプローチは投稿点の増加に伴い計算量が比例的に増加する．そのため，既存の経路探索アルゴリズムよりも計算量が多く生じる可能性があり，これに対する疑似アルゴリズムや計算量の改善が今後の研究課題として挙げられる．

さらに，本論文では各シナリオにおけるパラメータのチューニングを手動で行っている．そのため，先述したような実験結果に対する定量的な評価手法と．自動のパラメータチューニング手法が必要であると考える．特に現実のソーシャルデータに対してはパラメータの設定が難しく，今後の大きな課題であると考える．

また，今回の実験では焦点を当てなかったが，重みの減衰式，引力の計算式の改善や改善による比較実験の実施などが今後の方針として考えられる．

文　献

[1] Dijkstra, E. W. et al, ”A note on two problems in connexion with graphs”, Numerische mathematik, Vol.1, No.1, 269–271, 1959
[2] Peter E. Hart, Nils J. Nilsson, Bertram Raphael, ”A Formal Basis for the Heuristic Determination of Minimal Cost Paths”, IEEE Transactions on Systems Science and Cybernetics Vol.ssc-4, No.2, 100-107, 1968
[3] B. Golden, L. Levy, R. Vohra, ”The orienteering problem”, Naval Research Logistics Vol.34, 307-318, 1987
[4] T. Kurashima, T. Iwata, G. Irie and K. Fujimura, ”Travel Route Recommendation”, Knowledge and Information Systems, Vol.37 No.1 37-60, 2013.
[5] J. Gu, C. Song, W. Jiang, X. Wang, M. Liu, ”Enhancing personalized trip recommendation with attractive routes”, In Proceedings of the AAAI Conference on Artificial Intelligence, Vol.34, 662–669, 2020.
[6] P. Padia, B. Singhal, K. H. Lim, ”User-relative personalized tour recommendation”, In IUI Workshops, 2019
[7] D. Quercia, R. Schifanella, L. M. Aiello, ”The shortest path to happiness: Recommending beautiful, quiet,and happy routes in the city”, Proceedings of the 25th ACM conference on Hypertext and social media, p.116–125, 2014
[8] M. Omori, M. Hirota, H. Ishikawa, S. Yokoyama, ”Can Geo-tags on Flickr Draw Coastlines?”, International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL), 2014
[9] H. Zhang, M. Korayem, and D. J. Crandall, ”Mining Photo-sharing Websites to Study Ecological Phenomena”, In Proceedings of the 21st International Conference on World Wide Web, p.749-758, 2012
[10] 山下桜生, 横山昌平 ” ジオラベル付きツイートを用いた好みに応じた経路推薦手法” , 情報処理学会論文誌 データベース Vol.59 No.1 p.1–14, 2023.
[11] 夏思浩, 井嶋蒼, 横山昌平, ”大規模ソーシャルデータを用いた寄り道ルート推薦手法”, アカデミック・リソース・ガイド Web インテリジェンスとインタラクション研究会 No.16, p11-14, 2020

移動手段の混雑を考慮した観光ルートプランニング

田中　寛之† 　馬　　強†

† 京都工芸繊維大学　〒 606–0951 京都府京都市左京区松ケ崎橋上町
E-mail: †b0122035@edu.kit.ac.jp, ††qiang@kit.ac.jp

あらまし　特定の観光地について，観光客の著しい増加による混雑が生じている．この状況は，観光客自身の満足度低下や，観光地の住民の生活に悪影響を及ぼしている．本研究では，従来手法で用いたスポットの混雑状況に加えて，交通機関と道路状況を考慮した移動の混雑状況も考慮する動的報酬関数と制約条件を導入し，強化学習ベースの観光ルートプランニング手法を提案する．実験の結果，提案手法は既存手法と比べ観光スポットの混雑度が平均で約 6 ポイント低下し，観光スポット間の移動の混雑度が平均で約 4 ポイント低下したことが分かった．

キーワード　観光，強化学習，ルート生成，推薦，混雑

1　はじめに

特定の観光地では，観光客の著しい増加がオーバーツーリズム現象を引き起こし，観光地およびその周辺地域での交通渋滞が生じ，観光客の満足度が低下している．また，ごみや騒音などが増加し，観光地の住民の生活に悪影響を及ぼしている．観光中に予期せぬ滞在時間や移動時間が発生し，計画した観光活動が妨げられる可能性がある．

この問題に対処するための方法として，混雑状況を考慮した観光ルートプランニングが挙げられる [1] [2] [3]．これは，混雑回避や観光客の分散だけでなく，実際に移動可能な観光ルートを提案することも含まれる．

Gama ら [4] は強化学習と Pointer Network [5] を使用した，制約条件における最適な観光ルートを生成する手法を提案し，大規模な探索での計算のコストを削減している．また，前川ら [1] [6] は Gama らの手法に混雑の要素を組み込んだ手法を提案し，混雑を避けるルート生成を行う手法を提案している．Kong ら [2] [3] は，マルチエージェント強化学習ベースのルート生成手法を提案している．しかしながら，これらの手法では観光スポットの混雑を考慮しているが，観光スポット間の移動時に発生する観光客による混雑については考慮されていない．

そこで本研究では，前川ら [1] のモデルに移動時の混雑要素を導入する手法を提案する．移動中の時間帯と混雑から推定される「動的移動時間」と，混雑状況に応じた追加報酬機構「移動環境税メタファー」を導入して，観光ルート生成を行う．そのため，混雑の指標として「移動充足率」と「相対移動充足率」を定義する．移動充足率は，ある時刻におけるエリア間の移動人数にエリア間の移動確率を乗算した値と観光スポット間のバス本数の割合である．観光スポット間のバス本数は，移動可能なエリアから作成した交通ネットワークの最大フローの値を利用する．相対移動充足率は，ある時刻において移動充足率が正規分布に従うと仮定した場合に，移動充足率を標準化したものである．動的移動時間は，時間帯ごとに観光スポット間の混雑に応じて，観光スポット間の移動時間を変化させるものである．移動環境税メタファーが与える報酬は移動環境税と移動補助金であり，それぞれ混雑率の高い移動を抑制するための負の追加報酬と混雑率の低い移動を促進するための正の追加報酬である．

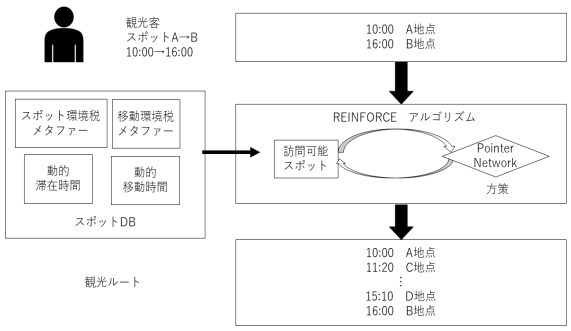


図 1　提案手法の概要図

図 1 にて，提案手法の概要について示す．入力には，観光の出発地と目的地とそれぞれの予定時刻を設定する．前川らの手法で提案された，環境税メタファーと動的滞在時間に加え，移動環境税メタファーと動的移動時間を追加し，訪問可能なスポットを求めていく．また，それらのスポットから強化学習と Pointer Network を利用し目的地に到達するまで学習を繰り返していき，混雑を考慮した価値の高い観光ルートを求めていく．

本論文の構成は次のとおりである．まず 2 節で関連研究を紹介し，3 節で問題提起を行う．続いて 4 節で本研究の提案手法を述べ, 最後に 5 節で結論を述べる．

2　関連研究

観光ルートとは，観光地をどの順番に巡っていくかを記したものである．観光ルート推薦の研究は数多く行われている．

2.1 混雑を考慮しない手法

Kong ら [2] は，観光客の多様性を考慮した動的な報酬機構を導入し，強化学習から複数のユーザに対して複数のルートを提示する手法を提案している．Gama ら [4] は，観光スポットの位置情報や設定した訪問時に得られるスコア，観光客の旅行時間から，強化学習と Pointer Network を利用した観光ルートの生成手法を提案している．加藤ら [7] は，観光スポットの位置情報やジャンルを設定し，観光客の観光動機によってその目的にあった観光ルートを最適化問題で求める手法を提案している．平野ら [8] は，ユーザが観光の際に所有している金銭や時間，体力と観光によって得られる満足度から，遺伝的アルゴリズムを利用し複数のルート推薦を提案している．武信ら [9] は，ユーザの観光スポットや移動経路に対する好みを入力として，その入力と観光地や移動経路の特徴を遺伝的アルゴリズムからルート推薦を行う手法を提案している．Sun ら [10] は，観光客の実際の観光ルートから頻出な観光スポットと移動ルートを求め，観光スポットの人気と観光客の移動時間を考慮した最適化問題を解くことで観光ルートを求める手法を提案している．

2.2 混雑を考慮した手法

Kong [3] らは，観光地の混雑と観光客の多様性を考慮し，観光地の局所的な混雑と世界的な観光客の分布を報酬とした観光ルート生成手法を提案した．前川ら [1] は，Gama らの手法に観光スポットの混雑状況や営業時間の要素を導入し，観光ルートを求める手法を提案した．田谷ら [11] は，ユーザ自身の情報と観光目的を利用し，他人と接触回数を減らせるような交通経路を遺伝的アルゴリズムに基づいたシステムを作成し推薦を行った．Li [12] らは，道路ネットワークと観光客の出発地，目的地，出発時間というデータから，旅行時間が最小となるルートを最適化問題から求める手法を提案した．移動時間は，道路の車両数や道幅，速度制限といった情報から決定している．

2.3 本研究について

既存研究では，ヒューリスティックを用いて最適化問題を解くことによる観光ルート推薦手法の提案が多数存在する．一方で，Gama [4] らのように深層学習や強化学習を用いた手法が登場している．混雑を考慮した手法という点においては，観光スポットの滞在時間を変化させる例 [1] や，移動時間を変化させる例 [12] がある．しかし，観光スポットと観光スポット間の移動に関する混雑の両方を考慮した研究は少ない．本研究では，現実世界における観光ルートのプランニングを行うため，混雑状況や報酬を導入した手法を提案する．

3 用語説明と問題定義

3.1 用語・記号

本論文で用いる用語と記号を定義し，説明を行う．

スポット ($P \in \mathcal{P}$) 一般に観光スポット，景勝地のことである．本研究では，京都市とその周辺の観光スポット 91 か所 [13] を対象とする．

エリア スポットとその周辺のバス停をまとめたものである．スポットと，徒歩 10 分以内のバス停をまとめている．本研究では，京都市を全 91 か所のエリアに分割している．

クエリ 観光客の (出発地・目的地・出発時間・時間予算) の組で，プランニング手法への観光客からの入力である．

インスタンス ある観光都市 (本研究では京都市) に存在するスポット情報と観光客のクエリの情報をまとめたものである．観光客のクエリに応じて，訪問するスポットをインスタンスから選定して最適なルートを生成する．

時間予算 (T) 旅程の開始時間から終了時間までの持ち時間である．ルートはこの時間予算内に抑える必要がある．

スポット充足率 (C_α) 前川らの手法の充足率にあたる部分である．スポットに滞在する観光客数と，スポットの収容人数の割合のことである．スポット充足率が高いほどそのスポットは観光客が集中している．

相対スポット充足率 (Z_{c_α}) 前川らの手法の相対混雑度にあたる部分である．あるスポットと，インスタンス内のほかのスポットを比べてどの程度混雑をしているかを表す．スポット充足率を利用し，インスタンス内のスポットの充足率が正規分布に従うという仮定の下で，スポットの充足率を標準化したものである．つまりスポットの充足率 C_α が $\mathrm{N}(\mu,\sigma^2)$ に従うと仮定したとき，相対スポット充足率 Z_{c_α} は

$$Z_{c_\alpha} = \frac{C_\alpha - \mu}{\sigma} \tag{1}$$

となる．

移動充足率 (C_τ) エリア間を移動する人数の，収容人数に対する割合のことである．移動充足率が高いほどそのエリア間の移動には人が集中している．以下の式 2 にて示す．

$$C_\tau = \frac{(l_{P_iP_j} + f_{P_i} \times 20) \times \pi_{P_iP_j}}{b_{P_iP_j}} \tag{2}$$

本研究では，エリア間の移動人数にエリア間の移動確率を乗算した値と，エリア間で運行されているバスの本数の割合を算出した．エリア間の移動人数は，スポットの訪問人数と観光地の地元の人数 l の和とする．スポットの訪問人数は Flickr の写真数 f に一定値を乗算した値を設定した．観光地の地元の人数は，正規分布に従う値とした．スポットの訪問人数で乗算された一定値や，観光地の地元の人数の平均，標準分布は，移動充足率が 0 から 1000 の間の値になるように決定した．エリア間の移動確率 π は，実際の観光の移動データと観光スポットの位置関係，スポットに来訪した時間の間隔から導出した [10]．エリア間で運行されているバスの本数 b は，エリアから作成したネットワークを利用した．ネットワークを利用し，直接エリア移動できる場合と間に 1 つのエリアを通ることで移動可能な場合の，総バス本数を最大フローから求めた値を設定した．移動可能でない場合は，0.5 とし移動可能な場合よりも移動充足率が高くなるように設定した．

相対移動充足率 (Z_{c_τ}) あるエリア間の移動が，観光都市における他の移動と比較してどの程度混雑しているかを表す．移動充足率をもとに計算され，スポット間の移動の充足率が正規

分布に従うという仮定の下で，スポット間の移動の充足率を標準化したものである．つまり移動充足率 C_τ が $\mathrm{N}(\mu,\sigma^2)$ に従うと仮定したとき，相対移動充足率 Z_{c_τ} は

$$Z_{c_\tau} = \frac{c_\tau - \mu}{\sigma} \tag{3}$$

となる．

報酬 (R)　スポットとエリア間の移動に対して用いる語で，本研究では訪問報酬，スポット環境税，スポット補助金，移動環境税，移動補助金の五種類が存在する．

訪問報酬 (R_v)　スポットを訪問した際に得られる正の報酬．時間帯ごとに，スポットで撮影された写真を数値化したものである．

スポット環境税 (R_p)　相対スポット充足率に基づく負の報酬．スポット充足率が一定値を超えると与えられ，最大は各スポットの訪問報酬の平均値の-1 倍である．

スポット補助金 (R_β)　相対スポット充足率に基づく正の報酬．スポット充足率が一定値を下回ると与えられ，最大は各スポットの訪問報酬の平均値である．

移動環境税 (R_τ)　相対移動充足率に基づく負の報酬．移動充足率が一定値を超えると与えられ，最大は各スポットの訪問報酬の平均値の-1 倍である．

移動補助金 (R_σ)　相対移動充足率に基づく正の報酬．移動充足率が一定値を下回ると与えられ，最大は各スポットの訪問報酬の平均値である．

ルート (route)　訪問するスポットを時系列順に並べた列 $((P_1,t_1),(P_2,t_2),\cdots,(P_k,t_k))$ のこと．提案手法での出力要素である．

価値 $(\mathcal{V})$　生成されたルートに対して用いる，ルートの価値はルートに含まれるスポットの訪問報酬，スポット環境税，スポット補助金，移動環境税，移動補助金をすべて足し合わせたものである．

滞在時間 (D, d)　前川らの手法の滞在時間にあたる部分である．スポットに観光客がどれくらい滞在するかを表す時間のこと．基準滞在時間 D はスポットが閑散としている場合の滞在時間である．動的滞在時間 d はスポット充足率によって滞在時間が変化するものである．

移動時間 (M, m)　スポット間をバスで移動するときにかかる時間のこと．基準移動時間 M は，スポットが閑散としておりかつ夜中から早朝の時間帯の場合の移動時間である．動的移動時間 m は，移動充足率と時間帯によって移動時間が変化するという仮定の下での移動時間を指す．

税率 (α,β)　提案手法にて示した 5 にて，スポット環境税メタファーと移動環境税メタファー，訪問報酬の重みを付けるためのパラメータである．

バス待ち時間 (W)　あるバス停で次に来るバスまでの時間間隔．

バスの平均到着間隔時間 (w)　時間帯ごとに，60 分をエリア間で運行されるバスの本数で割った値．

観光スポット在住の移動人数 (l)　観光客でないエリア間を移動する人数のこと．正規分布に従うと仮定して，時間帯ごとに値を設定した．

エリア間のバス移動本数 (b)　エリア間のネットワークを作成し，直接移動可能もしくは 1 スポットを経由して移動可能な場合の最大フローの値を設定した値．最大フローが設定できない場合は，0.5 とし移動充足率が高くなるように設定した．

本研究で用いる記号は表 3.1 のとおりである．

表 1　本研究で用いる記号

記号	意味
$\mathrm{P} \in \mathcal{P}$	スポット
C_α	スポット充足率
C_τ	移動充足率
Z_{c_τ}	相対移動充足率
R_v	訪問報酬
R_p	スポット環境税
R_β	スポット補助金
R_τ	移動環境税
R_σ	移動補助金
$route$	ルート
$\mathcal{V}$	ルートの価値
T	時間予算
D,d	D は基準滞在時間，d は動的滞在時間
M, m	M は基準移動時間，m は動的移動時間
α,β	α,βは税率
W	バス待ち時間
w	バスの平均到着間隔時間
l	観光スポットの地元の移動人数
f	観光スポットの写真数
tp	エリア間の移動人数
π	エリア間の移動確率
b	エリア間のバス移動本数
$\mathcal{A}$	訪問可能なスポットの集合

4 提 案 手 法

観光ルートプランニングの研究において，Gama ら [4] は強化学習と PointerNetwork を用いた手法を提案している．また，前川ら [1] は Gama 手法に動的滞在時間，環境税メタファーの二つを導入している．本研究では，前川手法をベースとして，混雑を表す動的な要素として動的移動時間，移動環境税メタファーの二つを導入する．従来の観光ルートプランニングの手法が式 (4) のような最適化問題としてモデリングできる．

$$\begin{aligned} \max \quad & \mathcal{V}_{route} = \sum_{P_i \in route} R_v(P_i) \\ \text{subject to} \quad & \sum_{P_i,P_j \in route} D_{P_i} + M_{P_iP_j} \leqq T \end{aligned} \tag{4}$$

提案手法は，動的移動時間と移動充足率の二つの動的要素を加えた式 (5) のような最適化問題として表すことが可能である．

$$\begin{aligned}\max \quad \mathcal{V}_{route} = & \sum_{P_i,P_j \in route} (1-\alpha-\beta)R_v(P_i,t) \\ & + \alpha(R_p(P_i,t) + R_\beta(P_i,t)) \\ & + \beta(R_\tau(P_i,P_j,t) + R_\sigma(P_i,P_j,t)) \quad (5) \\ \text{subject to} \quad & \sum_{P_i,P_j \in \text{route}} d_{P_i} + m_{P_iP_j} \leqq T \\ & \alpha \in [0,1], \quad \beta \in [0,1], \quad 0 \leqq \alpha+\beta \leqq 1\end{aligned}$$

本研究では，出発するスポットと目的地とするスポット，出発時刻．時間予算からなるクエリを入力として，観光ルートを出力する．

4.1 移動充足率の推定

移動充足率 C_τ は，Google Maps Platform の Places API で取得したデータと，京都市営バス [14] の時刻表データ，京都バス [15] の時刻表データ，およびスポット充足率に利用した GPS の軌跡データ，Flickr の写真データから求めた．Google Maps Platform でスポットから徒歩で移動可能である範囲内のバス停データ，本研究では徒歩 10 分以内のバス停のデータを取得した．そして，これらバス停に対して時間帯ごとのバス運行本数と系統別停留所を，京都市営バスの時刻表データから取得した．ここで，1 つのスポットに対し 1 つ以上の徒歩 10 分以内のバス停を割り当て，それを 1 つのエリアとした．そして，エリア間をバスで移動可能である部分をつないだ，エリアのネットワークを作成した．これらのデータから，エリア間の移動人数，エリア間の移動確率，エリア間のバス本数を推定した．以下の 4.1.1 節，4.1.2 節，4.1.3 節で説明する．

4.1.1 エリア間の移動人数の推定

エリア間を移動する人数 tp は，地元民と観光客を分けて考慮し，それらの和とした．地元民 l は，正規分布に従い人数を生成した．観光客は，Flickr の写真データ数の値 f から求めた．以下の式 6 にて示す．

$$tp_{Pi,Pj} = l_{P_iP_j} + f_{P_i} \times 20 \quad (6)$$

地元民の平均や標準偏差，Flickr の写真データ数の倍率の値は移動充足率が 0 から 1000 の間となる値に調整した．

4.1.2 エリア間の移動確率の推定

GPS の軌跡データからスポットの実移動確率を求め，スポット間の移動距離，実際のスポットの滞在時間間隔のデータを利用して [10]，スポット間の移動確率を求めた．以下の式 7 にて示す．

$$\begin{aligned}\min_{V_s,V_t,M} = \quad & ||I * (G' - V_s M^t V_t)||^2 \\ & + \lambda_c[||V_s - V_c||^2 + ||V_t - V_c||^2] \\ & + \lambda_d[||V_s - V_d||^2 + ||V_t - V_d||^2] \quad (7) \\ & + \lambda_r||M||^2 \\ C \approx V_c^t V_c, \quad & D \approx V_d^t V_d, \quad H \approx V_s M^t V_t\end{aligned}$$

$||\cdot||^2$ はフロベニウスノルム，I はスポット P_i, P_j 間で移動が観測されたかどうかを表すバイナリ行列，λ_r は正規化パラメータ，λ_c と λ_d はそれぞれスポット間の空間的影響と時間的影響を調整する重みである．G' は観光客移動行列 G の各成分を標準化した行列である．G は観光客の軌跡データから，実際にスポットで移動が行われた回数を抽出した行列である．D は各スポット P_i の来訪時間データを抽出し，朝，昼，夜に分割する値を設定し，スポット P_i, P_j 間の来訪時間分布の平均値の差を調べ，その値を来訪時間間隔分布に取り込んだ行列を正規化したものである．V_dは D を因子分解して求まる潜在変数である．C はスポット P_i, P_j間の地理座標の距離を $HaversineFormula$ を用いて求め，各要素の逆数を取った行列を最大値で正規化したものである．V_cは C を因子分解して求まる潜在変数である．V_s, V_t, M は推定重み付き移動行列 H を行列分解したときのそれぞれ移動前のスポットの潜在特徴行列，移動後のスポットの潜在特徴行列，スポットの位置間の関係を表す相互作用行列である．これらの行列は，N をスポット数，k を潜在特徴数として，$V_d \in \mathbb{R}^{N\times k}, V_c \in \mathbb{R}^{N\times k}, V_s \in \mathbb{R}^{N\times k}, V_t \in \mathbb{R}^{N\times k}, M \in \mathbb{R}^{N\times k}$ である．

本研究では，このスポット間の移動確率 H をエリア間の移動確率 π として利用する．

4.1.3 エリア間のバス本数の推定

エリアのネットワークから，直接移動できる場合の本数と，1 つのエリアを経由して移動できる条件での最大フローの値の和より移動におけるバス本数を設定した．ここで，スポットが密集している地域では，スポット間が徒歩 10 分以内で移動可能な箇所が多かったため，これらのスポット間は徒歩移動を行い．バス移動は発生しないと仮定した．徒歩移動中は混雑を考慮しないため，移動充足率は時間帯ごとに 0 としている．

4.1.4 移動充足率の計算

4.1.1 節，4.1.2 節，4.1.3 節で推定した値を利用して移動充足率の計算を式 8 にて行った．

$$C_\tau = \frac{(tp_{Pi,Pj}) \times \pi_{P_iP_j}}{b_{P_iP_j}} \quad (8)$$

4.2 動的移動時間

動的移動時間 m は，混雑によって変化するエリア間の移動時間である．実際，バスの移動時間はエリアや道路の混雑状況に影響される．道路は時間帯によって，朝夕のラッシュ時と，昼間，夜中から早朝などで混雑度合いが異なる．また，スポットの混雑によって付近のバス停に行列が生じ，次のバスを待機す

る可能性がある．これらから，混雑によって予想よりも移動時間が長くなることが発生する．

よって，エリア間の移動時間が混雑によって変化すると仮定してルート生成の要素に組み込む．ここでの混雑は，エリア間移動の混雑，つまり移動充足率と，時間帯ごとの道路の混雑の2種類を想定している．

4.3 移動充足率による混雑

移動充足率 C_τ と移動時間の関係に次のように仮定する．まず移動充足率に二つの閾値 L, H を設定する．$0 \leqq L < H \leqq 100$ とする．

- $0 \leqq C_\tau < L$ の場合，低度の混雑と呼び，混雑にする増加時間は生じないとする．
- $L \leqq C_\tau < H$ の場合，中度の混雑と呼び，混雑による増加時間はバスの平均到着間隔の時間 w とする．
- $H \leqq C_\tau$ の場合，高度の混雑と呼び，混雑による増加時間はバスの平均到着間隔の時間 w の2倍とする．

スポット P_i, P_j 間の時刻 t におけるバス待ち時間 $W_{P_i,P_j,t}$ は式 (9) と式 (10) のように表せる．

$$w_{P_i,P_j,t} = \frac{60}{b_{P_i,P_j,t}} \tag{9}$$

$$W_{P_i,P_j,t} = \begin{cases} 0 & 0 \leqq C < L \\ w_{P_i,P_j,t} & L \leqq C < H \\ 2w_{P_i,P_j,t} & H \leqq C \end{cases} \tag{10}$$

4.3.1 道路状況による混雑

道路には大きく分けて生活道路と幹線道路の二種類が存在する．生活道路は，地域住民が幹線道路に出るまでの道であり，道幅が比較的狭く歩行者が多いことが特徴である．幹線道路とは，主要な地点を結んだ道路網の一部であり，道幅が比較的広く車の交通量が多いことが特徴である．Ratwik ら [16] は，ある地点から別の地点への移動時に，出発地付近の生活道路から幹線道路に出て，幹線道路から目的地付近の生活道路へ向かい，そこから目的地へと向かうという考え方を示している．本研究では，生活道路をスポットからバス停までの道路，幹線道路をバスの走る道路とする．また，時間帯ごとの車移動時間について Ratwik らは次の式 (11) のように設定した．

$$M(t) = \begin{cases} M_{P_i,P_j} & t \in [0,7) \\ (1+\frac{t-7}{8-7})M_{P_i,P_j} & t \in [7,8) \\ 2M_{P_i,P_j} & t \in [8,9) \\ (2-\frac{3}{4}(\frac{t-9}{11-9}))M_{P_i,P_j} & t \in [9,11) \\ \frac{5}{4}M_{P_i,P_j} & t \in [11,15) \\ (\frac{5}{4}-\frac{3}{4}(\frac{t-15}{17-15}))M_{P_i,P_j} & t \in [15,17) \\ 2M_{P_i,P_j} & t \in [17,18) \\ (2-\frac{t-18}{20-18})M_{P_i,P_j} & t \in [18,20) \\ M_{P_i,P_j} & t \geqq 20 \end{cases} \tag{11}$$

本研究では，移動に使用する交通機関をバスに設定しており，バスを利用してエリア間を移動する時間はバス停からスポットまでの移動時間も含めて，経験的に車の移動時間に 10 分を加えた値とする．よって動的移動時間を

$$m_{P_iP_j}(t) = W_{P_i,t} + M(t) + 10 \tag{12}$$

と定める．

4.4 行動税メタファー

観光客を分散させることが，オーバーツーリズムの抑制方法の一つである [17]．本研究では，スポット間が混雑している箇所を避けたルートを生成することで，観光客を分散させる手段として提案する．そこで，動的報酬機関「移動環境税メタファー」を導入する．報酬の種類は，移動環境税と移動補助金の二つを提案する．

4.4.1 移動環境税

移動環境税 R_τ はあるスポット間の移動が観光都市内のほかのスポット間の移動と比較して混雑しているならば，スポットの移動時に負の報酬を与えるというものである．この報酬は移動環境税と呼び，訪問報酬とは独立している．

移動環境税は相対移動混雑度に依存して値を決める．スポット P_i, P_j の時刻 t における移動環境税は式 (13) のように表される．閾値を 2 つ設定し，それぞれ T_B, T_M(ただし, $0 \leqq T_B < T_M$) とする．T_Bは相対移動混雑度の許容値であり, T_Bを超えると移動環境税を課する. T_Mはこれを超えると移動環境税が最小値をとるようにする. 最小値はスポット訪問時刻の訪問報酬の平均値をマイナス 1 倍した値とする．

$$R_\tau(P_i,P_j,t) = -min(1, \frac{Z_{c_\tau P_i,P_j,t} - T_B}{T_M - T_B}) * \overline{R_v(P_*,t)} \tag{13}$$

4.4.2 移動補助金

移動補助金 R_σ はあるスポット間の移動が観光都市内のほかのスポット間の移動と比較して混雑していないならば，スポットの移動時に正の報酬を与えるというものである．この報酬は移動補助金と呼び，訪問報酬とは独立している．

移動補助金は相対移動混雑度に依存して値を決める．スポット P_i, P_j の時刻 t における移動補助金は式 (14) のように表される．閾値を 2 つ設定し，それぞれ T_B, T_L(ただし, $T_L < T_B \leqq 0$) とする．T_Bは相対移動混雑度の許容値であり, T_Bに届かないと移動補助金を与える. T_Lに届かないと移動補助金が最大値をとるようにする. 最大値はスポット訪問時刻の訪問報酬の平均値とする．

$$P_\sigma(P_i,P_j,t) = min(1, \frac{Z_{c_\tau P_i,P_j,t} - T_B}{T_M - T_B}) * \overline{R_v(P_*,t)} \tag{14}$$

4.5 モ デ ル

提案手法のモデルとその学習の仕組みは前川ら [1] や Gama ら [4] の手法をベースとする．入力はインスタンスとクエリの二つである．前者は訪問報酬，基準滞在時間，スポット充足率，

スポット間の移動時間, スポット間移動充足率の五つである．後者は観光客が与える情報で，旅行を開始するスポット (出発地), 終了するスポット (目的地)，出発時間と時間予算の四つをまとめたものである．出力は推定された最も価値の大きいルートである．

（1） インスタンスから，学習のためのクエリを発行する．
（2） 現在時刻と現在地から，訪問可能なスポットの集合 $\mathcal{A}$ を求める．訪問可能の基準は，現在地からそのスポットを訪問・滞在し，目的地に時間予算内に到着することができることである．また，動的滞在時間と動的移動時間の処理を行う．
（3） $\mathcal{A}$ を Pointer Network に通し，各スポットに訪問する確率分布を生成する．
（4） 生成された確率分布からスポットを抽出し，スポットを訪問し現在時刻と位置を更新する．
（5） 手順 2〜4 を目的地に到達するまで繰り返す．
（6） 生成されたルートの価値 $\mathcal{V}$ を計算し，モデルのパラメータを REINFORCE 手法を用いて更新する．また，スポット環境税メタファーと移動環境税メタファーの処理を行う．

ルートの生成は，与えられたクエリと学習されたモデルを利用し，ビームサーチによって価値が最大となるルートの生成 [6] を行う．

5 実　　験

5.1 概　　要

実験では，提案手法を前川手法および Gama 手法と比較することで，混雑を避けつつ観光体験の価値を失いにくいルートを生成可能かを確認する．

モデルは 30000 回学習したものを用いる．クエリは時間予算を 8 時間とし，出発地と目的地は同一とならないようにランダムに選択し，64 クエリ用意する．

評価基準は，訪問報酬，スポット充足率，移動充足率である．

5.2 比 較 手 法

比較する手法は，以下の 4 手法である．

- Gama 手法 (Gama)
- 前川手法 (Maekawa)
- 前川手法に動的移動時間を導入したもの (Tanaka-d)
- 本研究の提案手法 (Tanaka)

Gama 手法は提案手法のモデルと同じ条件にするために，出発地と目的地を異なる場所を指定できるようにすること，京都市のデータを参照して学習できるようにすること，の二点の変更を加えている．また，前川手法のパラメータは平均相対スポット充足率が低くなった値を利用している [1]．そして，提案手法のパラメータの一部は前川手法のパラメータと同一にしている．

5.3 データセット

本研究で用いるデータは下記のとおりである．

- 訪問報酬行列
- 基準滞在時間
- スポット充足率行列
- 移動時間行列
- 移動充足率行列

これらは以下のデータセットを用いて整理する．

- 京都観光マップ 2.0 [13] のデータ
- 京都市営バス [14] の時刻表データ
- 京都バス [15] の時刻表データ
- Yahoo Japan Corporation の GPS 統計データ

訪問報酬行列，基準滞在時間は前川らの手法で作成されたデータを利用した．これらは，京都観光マップ 2.0 [13] のデータをもとに作成された．これは Flickr に投稿された写真データをもとに，京都市を中心とした 91 か所のスポットの情報をまとめたものである．データにはスポットの名前と，地理的座標, 1 時間ごとの美的スコアが含まれている．美的スコアはそのスポットで撮影された写真を数値化したものである [18]．本研究ではスポットの訪問報酬として用いる．なお美的スコアが設定されていないスポットや時間帯は平均値を利用して補完し，訪問報酬行列が作成された．またこのスポットのデータとは独立した，GPS の軌跡データを用いられた．軌跡データがどのスポットに滞在しているかを座標から判定し，1 時間以内のログ更新は連続して滞在するとして，滞在した分数が計算された．その後滞在時間が 20 分以内と極端に短い場合，80 分以上と極端に長い場合はそれぞれ 40 分，70 分の値が採用された．この値を基準滞在時間として使用する．

スポット充足率行列は，前川らの手法で作成した充足率行列 [1] を利用した．基準滞在時間を求めたデータとは異なるデータを利用されている．ある期間に Yahoo! Japan のアプリを利用したユーザの GPS の軌跡から，スポットの緯度経度と比較し 1 時間ごとのスポット滞在人数が作成された．ここから，各スポットに収容人数を設定し滞在人数の収容人数に対する割合を計算された．収容人数は，特に有名で混雑が予想されるスポットが滞在人数が最大となる 2 時間前にスポット充足率が 100% を超えるように，それ以外のスポットでは 80%となるよう設定された．

移動時間行列は，Openrouteservice の matrix API [19] を用いて取得し作成した．前川手法では，公共交通機関での移動時間を経験的に自動車での移動時間を 2.5 倍としていた [1]．本研究では，バスでの移動時間を経験的に自動車での移動時間に 10 分を加算した値を用いて，移動時間行列を作成した．

移動充足率行列は，4.1 節で推定した移動充足率を用いて作成した．

5.4 パラメータの設定

パラメータは，前川手法で最も相対スポット充足率の低かった (L, H) = (40, 60), (T_L, T_B, T_M) = (-1.5, 0, 1.5) を利用する．また，(α, β) = ((1/3, 1/3), (2/5, 2/5), (1/2, 1/4), (2/3, 1/6) の中から最も低い高スポット充足率比率となった (2/3, 1/6) を利用した．

5.5 実験結果

実験結果を表 2 と表 3 に示す．表の項目は，すべて 64 クエリの平均値である．また項目のうち平均相対スポット充足率と平均相対移動充足率，訪問報酬は実数値で，その他は全て単位が百分率である．また，スポット・移動充足率は，ある移動の移動充足率とその移動後のスポット充足率の和の値である．生成されたルートの一例を図 2 にて示す．

5.5.1 動的移動時間

Maekawa と Tanaka-d を比較すると，Tanaka-d の方が平均移動充足率，平均相対移動充足率の値が良くなっていることがわかる．また，高スポット充足率比率と平均スポット充足率，平均相対スポット充足率に対しても値がそれぞれ 0.05 ポイント，約 4 ポイント値，約 1.2 ポイントが減少し，低スポット充足率比率とについては，約 0.008 ポイント上昇したため，動的移動時間の実装は，混雑を避けたルートの生成に貢献しているといえる．

5.5.2 移動環境税・移動補助金

Tanaka-d と Tanaka を比較すると，混雑に関する指標は平均移動充足率を除いて，Tanaka の方が Tanaka-d のよりも良いことがわかる．また平均移動充足率も 0.8 ポイントの差であり，平均相対移動充足率は Tanaka が Tanaka-d よりも低い値であるため，低い混雑のルートを選択していると言える．よって，移動環境税と移動補助金を導入することで，混雑回避のルート生成が可能になるといえる．

5.5.3 平均スポット充足率

移動充足率に関する基準を比較すると，Tanaka，Maekawa，Gama の順に優れた値を記録していることがわかる．Tanaka は，Maekawa と比べて 6 ポイント程，Gama と比べて 30 ポイント程度減少している．Gama と Maekawa の移動充足率にも差が生じているが，これは移動充足率の計算において Flickr の写真データを利用しており，スポット充足率にも同データが利用されていることからスポットの混雑と移動の混雑が連動していると考えられる．

5.5.4 高スポット充足率比率・低スポット充足率比率・平均相対スポット充足率

相対スポット混雑比率に関する基準を比較すると，Tanaka は Maekawa の値よりも，高スポット混雑比率は 8%程度減少していることがわかる．また，低スポット混雑比率に関しては 10%上昇，平均相対スポット充足率については同等であることがわかる．よって，Tanaka の方が混雑しないスポットを選択しているといえる．

5.5.5 平均移動充足率・平均相対移動充足率

平均移動充足率に関する基準を比較すると，Tanaka は Gama よりも 20 ポイント近く減少しており，Maekawa に関しても 5 ポイント近く減少していることがわかる．また，平均相対移動充足率の比較についても，Tanaka は Gama よりも 0.3 ポイント低く，Maekawa に関しても 0.15 ポイント程度低いことがわかる．このことから移動の混雑の要素を導入したことで，これらの混雑を回避したルートを生成することが可能となるといえる．

表 2　スポットに関する結果

モデル	高充足	低充足	平均充足	平均相対	訪問報酬
Gama	0.9777	0.007813	70.04	-0.004496	**42.71**
Maekawa	0.6509	0.01987	49.19	-1.507	34.16
Tanaka-d	0.6057	0.02734	44.45	-1.620	24.16
Tanaka	**0.5768**	**0.1391**	**42.78**	**-1.876**	26.65

表 3　移動に関する結果

モデル	平均充足	平均相対	平均スポット・移動充足
Gama	52.11	-0.1128	130.2
Maekawa	35.30	-0.3037	91.32
Tanaka-d	**30.00**	-0.3721	83.13
Tanaka	30.82	**-0.4553**	**81.15**

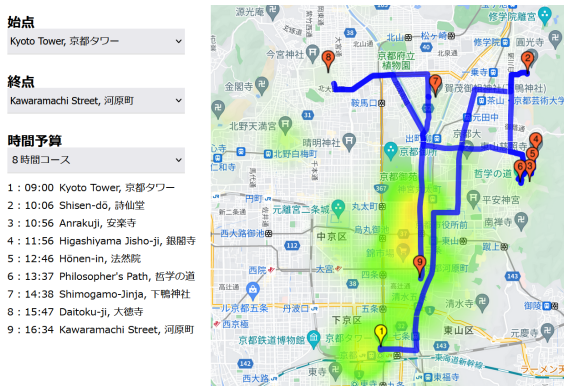


図 2　生成されたルートの例

5.5.6 平均スポット・移動充足率

平均スポット･移動充足率に関する基準を比較すると，Tanaka，Tanaka-d，Maekawa，Gama の順に良いことがわかる．このことから，スポットと移動の混雑の両方が関連した観光ルートを生成しているといえる．

5.5.7 訪問報酬

訪問報酬に関する基準を比較すると，Gama，Maekawa，Tanaka，Tanaka-d の順に良いことがわかる．Gama と Tanaka で値に差が出たのは，Gama は混雑や移動時間などの影響を全く受けない手法であり，訪問報酬のみを考慮したルートであることから訪問可能なスポットが他モデルに比べて多いことが影響していると考えられる．同様に，Tanaka の方が Maekawa よりも低い値となったのは，Tanaka が訪問報酬，スポット混雑，移動混雑の 3 つを考慮しており，ルート選択の幅が広くないことが原因である可能性がある．訪問報酬の改善を今後検討していく予定である．

6 おわりに

本研究では，混雑を考慮した観光ルートプランニングを行うために，観光ルート生成にスポットの混雑の要素を導入し提案した前川ら [1] の手法に，「動的移動時間」，「移動環境税メタファー」を導入したモデルを提案した．それぞれの要素の推定には，Flickr の写真データや観光客の観光軌跡データ，エリア間の移動ネットワークから最大フローの値の利用などを行い推

定した移動充足率を用いた．

提案手法の有用性を確認するために，Gama，Maekawa，Tanaka-d，Tanaka を比較する実験を行った．実験を行うにあたって，スポット環境税メタファーと移動環境税メタファー，報酬の割合のパラメータをチューニングした．その後，四つのモデルを相対充足率を中心とした評価基準で比較した．平均相対スポット充足率，平均相対移動充足率の共に，Tanaka，Tanaka-d, Maekawa, Gama の順に大きくなっており，Tanaka が混雑を回避したルートの生成を可能にしたことがわかった．

今後の展望として，バス以外の公共交通機関の移動を考慮することが挙げられる．

謝　　辞

本研究の一部は科研費（23H03404）による．

文　　献

[1] 前川丈幸, 笠原秀一, 馬強. 混雑を考慮した観光ルート生成に関する研究. *DEIM Forum 2023*, 2023. 5b-3-4.

[2] Wei Kun Kong, Shuyuan Zheng, Minh Le Nguyen, and Qiang Ma. Diversity-oriented route planning for tourists. In *in DEXA 2022*, Lecture Notes in Computer Science, pp. 243–255, 2022.

[3] Yuntao Kong, Peng Chen, Nguyen Minh Le, and Qiang Ma. Dual congestion-aware route planning for tourists by multi-agent reinforcement learning. In *in DEXA 2023*, Lecture Notes in Computer Science, pp. 331–336, 2023.

[4] Ricardo Gama and Hugo L. Fernandes. A reinforcement learning approach to the orienteering problem with time windows. *Comput. Oper. Res.*, Vol. 133, p. 105357, 2021.

[5] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *in NIPS2015*, pp. 2692–2700, 2015.

[6] Takeyuki Maekawa, Hidekazu Kasahara, and Qiang Ma. Tour route generation considering spot congestion. In *in DEXA 2023*, Lecture Notes in Computer Science, pp. 478–492, 2023.

[7] 加藤央昌, 早瀬光浩. 観光支援のためのルート提供システムの検討. 知能と情報, Vol. 32, No. 5, pp. 917–922, 2020.

[8] 平野陽大, 諏訪博彦, 安本慶一. ユーザのリソース消費を考慮した意思決定支援のための複数観光経路提示手法. マルチメディア，分散協調とモバイルシンポジウム 2019 論文集, Vol. 2019, pp. 830–839, 2019.

[9] 武信雄平, 奥野拓. 観光スポット間の類似度に基づく飽きにくい観光ルート生成手法の提案. マルチメディア, 分散, 協調とモバイルシンポジウム 2022 論文集, Vol. 2022, pp. 1570–1573, 2022.

[10] Haodong Sun, Yanyan Chen, Jianming Ma, Yang Wang, Xiaoming Liu, and Jiachen Wang. Multi-Objective Optimal Travel Route Recommendation for Tourists by Improved Ant Colony Optimization Algorithm. *Journal of Advanced Transportation*, Vol. 2022, pp. 1–14, 2022.

[11] 田谷瑛悟, 松田裕貴, 諏訪博彦, 安本慶一. 交通機関利用における接触を回避するための経路推薦手法の検討. 2020 年度 情報処理学会関西支部 支部大会 講演論文集, Vol. 2020, pp. 1–4, 2020.

[12] Ke Li, Lisi Chen, and Shuo Shang. Towards alleviating traffic congestion: Optimal route planning for massive-scale trips. In *in IJCAI 2021*, pp. 3400–3406, 2021.

[13] Taishan Li Jing Xu, Junjie Sun and Qiang Ma. Kyoto sightseeing map 2.0 for user-experience oriented tourism. *in MIPR2021*, pp. 239–242, 2021.

[14] 京都市. 京都市バス: ハイパー市バスダイヤ. `https://www2.city.kyoto.lg.jp/kotsu/busdia/bustime.htm`. 2024-2-7 閲覧.

[15] 京都バス株式会社. 時刻表. `https://www.kyotobus.jp/route/timetable/`. 2024-2-13 閲覧.

[16] Ritwik Raj and Chase C . Murray. The time - dependent multiple flying sidekicks traveling salesman problem : Parcel delivery with traffic congestion. `https://ssrn.com/abstract=3767870`, January 2021. pp.1–38.

[17] 国土交通省観光庁. オーバーツーリズムの未然防止・抑制に向けた対策パッケージ. `https://www.mlit.go.jp/kankocho/content/001705662.pdf`. 2024-2-7 閲覧.

[18] Bastian Goldlucke Vlad Hosu and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. *in ICPR2019*, pp. 9375–9383, 2019.

[19] Openrouteservice. `https://openrouteservice.org/dev/api-docs/v2/matrix/{profile}/post`. 2024-1-10 閲覧.

3D都市モデルを利用して移動中の景観を考慮する観光経路推薦手法の検討

斎藤　康平†　　横山　大作†

† 明治大学理工学研究科　〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1
E-mail: †{koh,dyokoyama}@meiji.ac.jp

あらまし　自動車は単に最短時間で移動するという目的だけでなく, 良い景観を眺めながら目的地まで移動したいという欲求がある。楽しみたい景観には自然や, 建築物, 都市などが含まれるが, 都市の中には建築物が密集しており, ある地点からの建築物の見え方を推測するのは困難である。そこで本研究では, 景観の良さを決定する要素の 1 つである都市内の建築物の見え方に重点をおき, 設定したランドマークが移動中に良く見える経路の推薦を目指す。ある地点からの景観の良さを推定するために, 国土交通省が提供する 3D 都市モデルの PLATEAU を利用しゲームエンジンを用いてレンダリングを行い, ターゲットとなる建築物の見え方を求め, その望ましさを数値化するモデルの構築を行った。また, 経路全体の満足度の推定のためには, その経路上の地点を均等にサンプリングして, 景観モデルから得られる満足度の平均値を用いた。実際の地図上で複数の経路を生成し, 提案手法の評価を行ったところ, 評価が高い経路は比較的良い景観を持っているという結果を得ることができた。今後, 景観モデルを移動経路の推薦システムに組み込み, 望ましい経路推薦が行えるかどうかの評価を行っていく。

キーワード　観光, 経路推薦, 地理データ活用

1　はじめに

平成 27 年の全国都市交通特性調査 [1] によると, 平日には移動を行う人のおよそ 45%が, 休日には 60%以上の人が自動車による移動をしており, 三大都市圏の中でも休日には半数以上が自動車を利用という集計結果がある. 自動車で移動する目的には, 通勤などの目的地まで早く移動したいという欲求や最短経路で移動したいという欲求だけではなく, 観光を目的とした良い景観を眺めながら移動をしたいという欲求が存在する. そのような欲求を持っている場合, 最短距離での移動経路がその欲求を満たすとは限らない. そこで, ユーザーが満足するような景観を持つ経路推薦システムを考える必要がある.

ユーザーが満足するような景観には, 自然景観と都市景観に大別される. 自然景観を基準にした経路推薦手法はいくつかある. 自然景観をベースに経路推薦を行う場合は, ユーザーが良いと考えるターゲットが田園や海岸などの風景となり, 景観が見えることができるかどうかが重要になる. それに加え, 都市景観をベースに経路推薦を行う場合は景観の対象が建築物になるため, 建築物がどのように見えるのかということを考慮する必要があり, より複雑になる. また, 都市内部は道路ネットワークが密に張り巡らされており, 想定外の地点からランドマークが見える可能性がある. 建築物も密集していることから, ランドマークがそれらの影になり十分な見え方ができない可能性があるため, 複雑な評価方法が必要となる.

一般的にターゲットの見え方を取得するソースとして画像があげられる.Flicker などのインターネットから画像を取得して景観を評価することが多い. それらの写真はフォトスポットからの撮影であることが多い. そのため決まった方向からの写真が多く, 位置情報もほとんど変わらないことからデータの自由度が低いと思われる. 例えば, 経路上の景観が, 取得した画像とは異なる方向からターゲットを見ている場合が考えられる. そのため, ターゲットを複数の方向から見ることができる, 新たな方法で景観画像を取得する必要がある.

本稿では, 景観の良さを決定する要素の 1 つである都市景観に着目し, 都市内部にあるランドマークの見え方がユーザーにとってどのくらい満足度の高い景観であるのかを画像から数値的に推定する. 推定された結果から, どの経路が満足度の高い結果であるのかを求めるシステムを作成する. 景観画像は時間などの情報の差を無くすため,3D 都市モデルを利用しレンダリングしたものを利用する. 本稿ではランドマークを観光の目的となる建築物やユーザーにとって印象に残る建築物のことを指す.

2　提案手法

2.1　レンダリングに基づく景観把握と経路推薦システム

本稿では自動車での移動を前提とした都市内部でのランドマークとなる建築物がより良い見え方となる経路の推薦を考える.

初めに, ランドマークの景観を考えるにあたり景観画像を取得する. 景観画像にはインターネットに投稿されたデータを用いる手法が数多くあるが, 時間帯のズレや天候の影響による画像差を無くすため, ゲームエンジン上でレンダリングされた画像を利用する. 国土交通省が主導する日本全国の都市を 3D モデル化している, 整備・活用・オープンデータ化プロジェクトである PLATEAU [2][1]を Unity にインポートし, レンダリングを行っ

1：mlit.go.jp/plateau

た.PLATEAU には建築物モデルの他に, 道路・都市整備・地形・都市計画決定情報・災害対策用のモデルなどが含まれており, 多岐にわたる分野で有用なデータが含まれている. PLATEAU の建築物モデルは地域によって LOD4 まで対応している.LOD とは Level of Detail の略でモデルの詳細さを表す. 今回は LOD1 の建築物モデルを利用した.LOD1 はデータが軽量であることや, どの都市でも提供されていること, 本研究の可用性などを考慮し LOD1 に決定した.LOD1 は建築物の概形を持った箱型のモデルとなっており, 屋根の形状表現はされず, 平屋根でない限りは省略される. そのため, モデルの高さは建築物の高さの中央値を採用しており, 低層部と高層部に分かれるような建築物のモデルには注意が必要である. モデルの位置が水平方向では 1.75 m, 高さは 0.66 m 以内の誤差になるように保証されている.

経路の推薦にはあらかじめいくつかの経路を取得し評価を行う. 得られた経路ごとに適当な間隔で図 1 のような画像をレンダリングし, その景観の良さを数値で評価する. 評価された値の平均を経路の景観の良さとし, 平均値が高い順に出力し推薦を行う. 景観の良さは景観画像に含まれるいくつかの要素をそれぞれ算出し, それらの要素に対し, 重回帰分析を行うことによって取得する. 正解データは人間が景観を見た際の定性的な評価を 0 から 5 の 6 段階の評価にしたものとする.

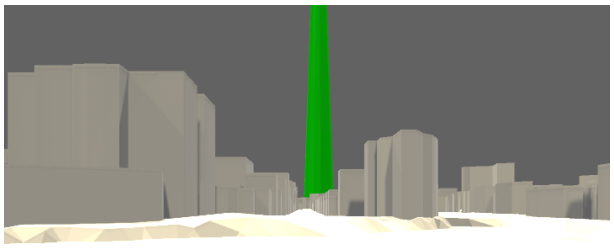


図 1: レンダリングされた画像

2.2 試験実装

本稿では景観の対象となるランドマークを 1 つとし, 経路は 2 地点間を結ぶものと考える. 都市内部のランドマークとして東京スカイツリーを採用した. 東京スカイツリーは東京都墨田区にある高さ 634m の電波塔として知られている. 東京都東部では多くの地点から視認することができ, 国内にあるランドマークの中では最適であると考えられる. 試験領域として, 図 2 のような東京スカイツリー周辺の約 3.8 km× 約 2.5 km の範囲を扱う.

2.2.1 使用データ

都市景観ベースの経路推薦を検討するにあたり, 景観を推定するための画像情報と道路ネットワークデータが必要となる. 前述の通り, 画像情報は Unity と PLATEAU を利用し,PLATEAU に含まれる建築物・道路・地形モデルをインポートした. しかし,PLATEAU には東京スカイツリーや隣にある東京スカイツ

図 2: 試験領域

リーイーストタワーのような高層の建築物モデルがふくまれていない. そのため,3D モデルを自作し Unity にインポートした. 手続き的なモデリングができることで知られている Houdini という 3DCG ソフトウェアを利用し, 図 3 のような錐体をモデリングした. 東京スカイツリーは底面が三角形で, 頂点に近づくにつれ円形となる珍しい形をしている. 他の建築物モデルは LOD1 で利用するため, それらある程度 LOD を合わせて底面が正 14 角形の錐体を作成した. これを FBX にエクスポートした後,Unity にインポートした. 作成したモデルは実際の大きさより小さいため,Unity 上でスケールを操作しおおよそ実際の大きさに整えた.

道路ネットワークデータは OpenStreetMap から取得したデータを利用した. 試験領域に含まれるネットワークデータには道路ノードが 3291 件, 道路リンクが 5282 件のデータが含まれる.OSM の位置情報は経緯度だが,Unity はインポートした 3D モデルの中心を原点としたメートル法になっている. そのため目視にて Unity の原点と経緯度を照らし合わせ, 位置情報の修正を行った.

レンダリングには Camera コンポーネントを用いた.Clipping Planes(オブジェクトを描画する範囲) の上限値を 5000 m, 下限値を 0.3 m に変更した. 視野角は 60 度, 画像サイズは 720×480 ピクセルとした. 撮影した画像は, ユーザーが自動車を運転中に見る景観を前提とし, ユーザーが自然に見えることができる正面から撮影を行う.

2.2.2 景観モデル

前述の通り, 都市景観ベースの経路推薦手法を考える上で景観の良し悪しを推定する必要がある. 推定する際はゲームエンジンでレンダリングされた画像から推定に必要な要素を算出し, 重回帰分析によって景観の満足度を推定する. 景観の満足度には 3 つの要素が大きくかかわっていると考える.

a) 画面内に占める割合

ユーザーが景観の良さを評価する上で最も重要な要素の 1 つである. 画像に占めるランドマークの面積が大きければ, 満足度の高い景観であると考えることができるので, ランドマークが写っている面積を算出した. 緑に塗りつぶされたターゲットを撮影し, 撮影した画像を OpenCV を用いて緑色とそれ以外の色で二値化した. 検出されたピクセルの総数を面積とする. 画像

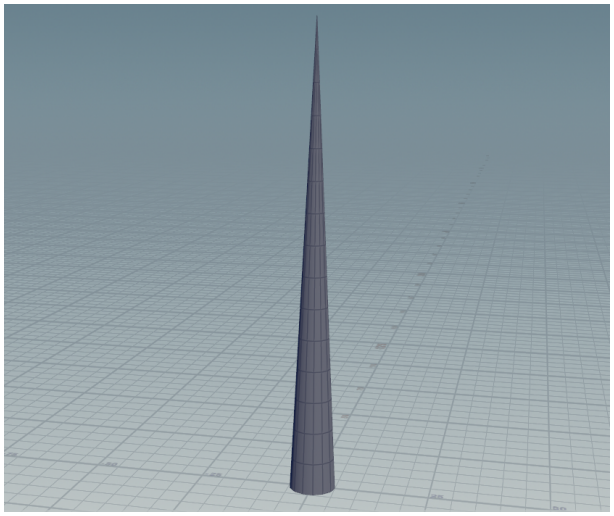


図 3: 作成したランドマークモデル

を二値化する際,Unity に初期値として設定されている Standard シェーダーは, 光源の位置によって影が入るため色が変化し, 正しく領域が検出できない可能性がある. そこで著者が常用している lilToon[2]を適用し, 影が入らないように設定した. これはアバターベースのサービス用に作成されたシェーダーで, 簡便な設定ができることで知られている.

例としてランドマークに東京スカイツリーを設定し,Google ストリートビューからそれが写っている画像を引用する. 図 4 のように中央に大きく写っており, 全体が画像内に収まっている. このような見え方ができているならば, 良い景観と考えることができる.

しかし画像内にランドマークが大きく写っていた場合でも, 一概に良い景観であるとすることは難しい. 図 5 に良くない景観の例を挙げる. 中央にスカイツリーが写っているが, タワーの下方しか見えておらず全体像を把握することが難しい. このような景観ではユーザーが満足することは難しいと考えられる. そのため, このようなケースを解決するために次の要素を追加した.

b） ランドマークの表面積の割合

ランドマークが写っていても景観が良くないケースがありその全体が見えているのかを判別するために, ランドマークが持つ表面積のうちどの程度見えているのかという数値を設定する必要がある. しかし, レンダリングされる地点が画像によって異なることなどから, 厳密に算出することが難しい. そこで, ランドマークモデルのテクスチャを縦に 14 分割し, それぞれに異なる色を塗った. 画像中に塗り分けられた色が識別できたのかによって, 大まかな表面積を推定した. テクスチャの分割を行ったランドマークモデルは図 6 に示す. 作成したモデルは通常とは異なるトポロジーであるが, 均等なテクスチャの分割が可能となっている. 作成したモデルは 15 分割になっているが, これはランドマークの下部のみが大きく写るケースを低く点数付けするために, 下から 2 つ目までを同じ色で着色することで 14 分割

2：lilxyzw.github.io/lilToon

図 4: 良い景観の例

図 5: あまり良くない景観の例

を実現した. そうすることで, ランドマーク全体が写っているが面積が小さいケースにも対応することができると考えられる.

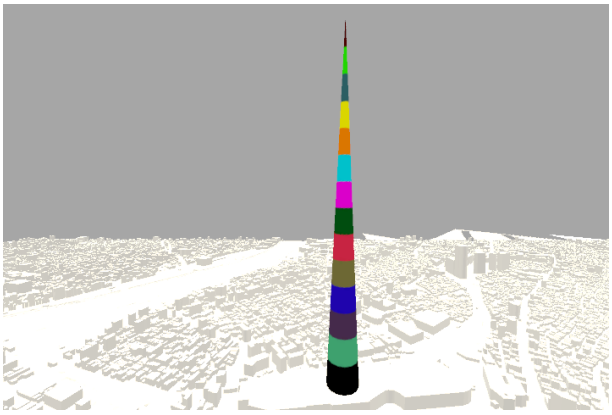


図 6: Unity 上で 14 分割されたランドマーク

c） 撮影地点からの距離

画像内に写るランドマークの大きさや表面積の割合は, 撮影地点からの距離が関わっていると考えている. レンダリングを行った地点からランドマークまでの距離を算出し, 以上 3 つの要素をそれぞれ正規化したものを説明変数とする.

2.2.3 景観モデルの学習

scikit-learn の LinearRegression クラスを利用し, 重回帰分析を行った. 本稿ではユーザーにとって良い景観を含む経路を推薦することを目的としているので, 目的変数をユーザーが評価した景観の満足度と設定する. ランドマークが写っていない画像の満足度は 0 点, ターゲットが写っている画像の満足度は 1 点から 5 点の計 6 段階で評価付けを行った. 評価付けを行う画像データは, 道路ノードの位置からランドマークの方向へカメラを回転させ, 道路メッシュからの高さ 1 m~2 m の位置でレンダリングをしたものである. 道路ノード上に道路メッシュが無い場合は海抜や橋などを考慮し,Unity 上の高さ 5m でレンダリングを行った.3291 枚撮影し, そのうち 536 枚にランドマークが写っていた. ランドマークが写った画像に評価付けを行った. 評価付けは著者の他に,3 名に景観画像の評価付けを目安となる評価基準を説明した上で依頼した. 計 4 件の評価データを収集し, 各満足度の平均値をユーザーが評価した景観の満足度とした. 著者が評価したデータを case1, 著者以外の評価データを case2,case3,case4 とする. 各評価の満足度の散らばりを図 7 に示す. プロットの大きさがその評価をしたデータの個数である. case1 から case3 は,1 点 3 点までは似たような点数付けととなっており,4 点 5 点は人によって評価が散らばっている.case4 は 5 点の評価がかなり少ないことから, 他の評価データとの散らばりが大きくなっている.

以上のデータを用いて学習を行い, モデルを作成した.

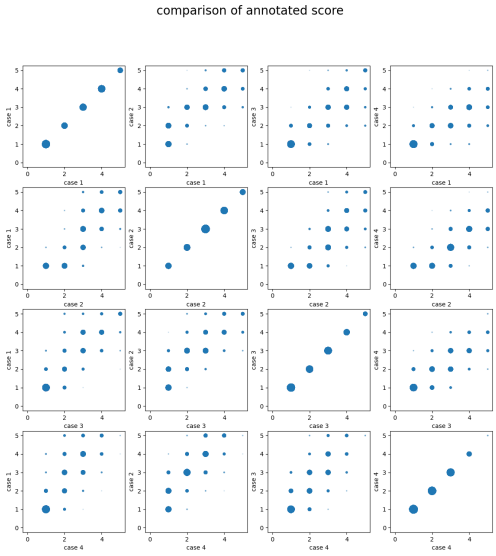


図 7: 各評価の分布図

2.2.4 経路モデル

経路の景観の満足度を評価するためには, ある地点の道路ノードから次の道路ノードの方向へ向いた景観画像が必要となる. これを出発地点から目的地点まで繰り返すことによって, 経路全体の景観を取得することができる. あらかじめ, 全ての道路ノードに対し隣接道路ノードへ向いた景観画像をレンダリングし, 収集した景観画像からそれぞれ説明変数を算出した. それを作成したモデルに与え, 得られたそれぞれの満足度の平均を取ることで経路全体の満足度とした. レンダリングした画像は 10564 枚だった.

経路上の景観を考える際, 道路ノードの中間地点でも撮影する方がより精密な評価ができると思われる. しかし, 図 8 のように直線ではない道路リンクが数多くあるため, 正しく撮影できない懸念がある. そのため, 本稿では単純化し上記の方法でレンダリングした.

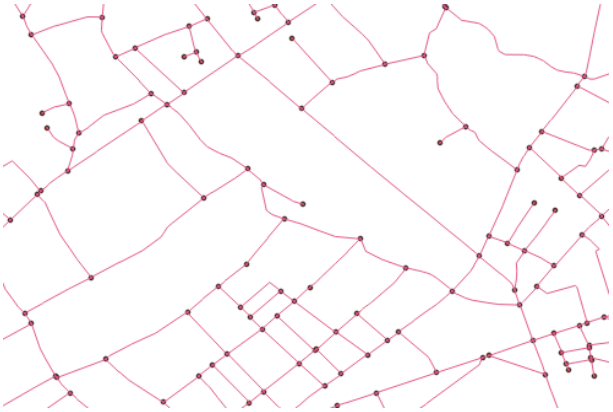


図 8: 道路リンクが直線でない例

2.2.5 評価用経路の作成

2 点間の経路探索アルゴリズムで良く知られているダイクストラ法 [3] を用いて経路を取得した. 複数の経路を比較検討する必要があるが, ダイクストラ法では最短経路以外の取得が難しい. ダイクストラ法に変更を加えることにより複数経路を取得した. 表 1 にそのアルゴリズムを示す. まず初めに, ナイーブなダイクストラアルゴリズムを実行することにより現在のグラフの最短経路を得る. 得られた経路の中に今まで通過したことが無い道路リンクがあれば, そのリンクに仮想の距離を加え, 実際の距離よりも長い道路リンクと見なす. その後, 通過したことがある道路リンクと得られた経路に含まれる道路リンクの和集合を取る. これを複数回繰り返すことで, 今までに得られた経路とは異なる経路を取得することが可能となる. こうすることで, ある程度最短距離に近い経路の中から景観が良い経路を推薦することができる. 本稿では一方通行などの交通法規を考慮せず, 評価用経路を作成した.

3 評　　価

3.1 景観画像評価モデル

初めに景観モデルの評価を行った. 学習に利用したデータを利用し, 著者らが評価した値とモデルから出力された値を比較した. その結果を図 9 に示す. 多少のばらつきはあるが,0.6 付近までは誤差± 0.1 程度で推定ができている.0.6 以上は正解の値より出力された値が低く推定されてしまった. これは正解データに偏りがあったことに起因すると考えられる.

表 1 に学習したモデルの決定係数 R^2 と平均二乗誤差を示す. 学習データと評価データの決定係数が共に 0.9 を超えており, 充分に目的変数を説明できている良いモデルが作成できている.

Algorithm 1 複数経路探索アルゴリズム

```
function SEARCH_MULTI_ROUTE(attempt_time, penalty)
    Graph ← OSM
    passed_link ← ∅
    TIME ← attempt_time
    p ← penalty
    node[TIME] ← Null
    path[TIME] ← Null
    for i = 0 ... TIME do
        node[i], path[i] = DIJKSTRA(start_node, target_node, Graph)
        link_need_penalty ← path[i].link ∩ ¬passed_link
        for link | link ∈ link_need_penalty do
            Graph.link.length ← Graph.link.length + p
        end for
        passed_link ← passed_link ∪ path[i].link
    end for
    return node, path
end function
```

平均二乗誤差は小さい値となっており, 十分だと考えられる.

作成したモデルから得られた出力データと正解データが乖離していないかを確かめるため, 統計的仮説検定を行った. 出力データと正解データに差が無いと仮定し, 統計的検定量と p 値を求めた. この時有意水準を 5%と設定した. 結果を表 2 に示す. 統計検定量は約-0.1,p 値は 0.9 となり, 有意水準の 0.05 を大きく超える. この結果により仮説が採択されることになる. これにより, 出力された値と正解データの値に差がなく, モデルから得られた値は尤もであると考えられる.

学習したモデルの偏回帰係数を表 3 に示す. ある地点から見えるランドマークが画像中に占める面積が最も評価に影響を与えていることがわかった. 景観を評価する際に, 前提として見えているかどうかが重要となる. ことから最も影響を与えているのではないかと思われる. 最も係数が小さかった要素は撮影地点からの距離となった. 撮影地点がランドマークから離れるほどランドマーク全体が見える可能性が増えるため, 画像に含まれていないが有用な説明変数になると考えていたが, 満足度の評価を画像で行っていることなどから出力に対する影響が少なかったと思われる.

表 1: モデルデータ

項目	学習データ	評価データ
決定係数	0.9489	0.9504
MSE	0.0024	0.0023

表 2: 評価データの検証

統計検定量	-0.1080
p 値	0.914

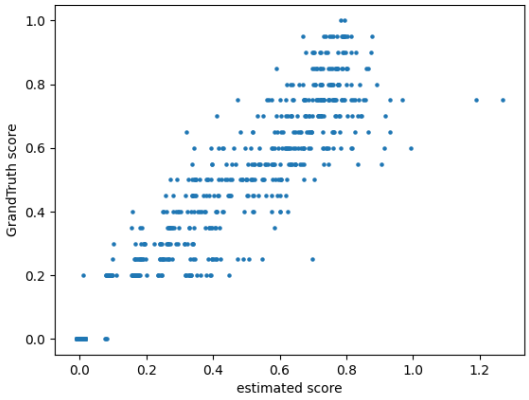


図 9: 正解の値と出力された値の比較

表 3: 偏回帰係数

ランドマークが写っている面積	1.18083507
ランドマークの表面積の割合	0.79403009
撮影地点からの距離	-0.02834543

3.2 経路推薦の評価

次にある経路の道路ノードごとの満足度を図 10, その経路上の道路ノードから撮影された画像を図 11 に示す. ①の景観では低い評価値が得られた. 中央からずれた位置にランドマークが見えており, 想定よりも低いが妥当な評価値であると考えられる.②の景観は中央からはずれているが,①よりも大きく写っておりある程度良い景観であるため, 高い評価値が得られていると思われる.③は②よりも少し小さく写っているが中央に見えていることや全体像がよりわかりやすいことから, 想定では高い評価値を付けていた. しかし, 実際のモデルから得られた評価値は②の方が高い評価を得ていた. アノテーションの偏りによる結果なのではないかと思われる.

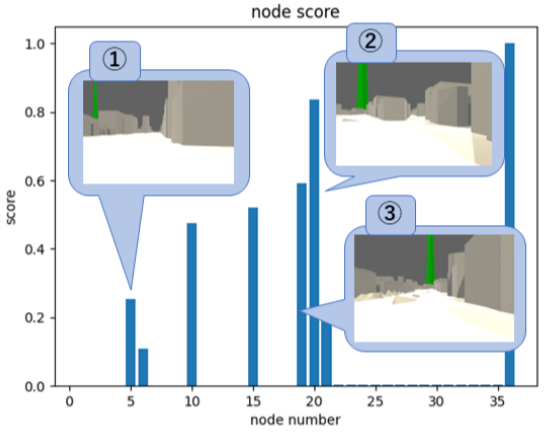


図 10: 道路ノードごとの満足度

経路に含まれる道路ノードから撮影された景観画像の推定値を学習モデルからそれぞれ算出し, その平均を経路の満足度とした. 最短経路と著者が良いと考える経路を含めた 8 種類の経

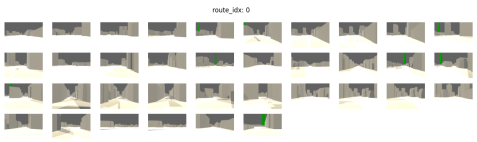


図 11: 道路ノードごとに撮影された経路画像

路を比較, 検討した. 満足度が高かった順に経路をソートし, その結果を経路上の道路ノードから撮影された画像の上位 4 経路を図 12 に示す.

著者が良いと考える経路が最も高い満足度が得られる経路であるという結果が得られ, 期待通りのモデルが作成できていると思われる. 著者が設定した経路の評価が高かった要因として, ランドマークが見えている景観が他の経路と比べ多かったことが考えられる. ルート 0 とルート 5 を比較したとき, 経路のノード数とランドマークが見えた枚数の差が少ないが, 満足度には大きな差が生まれている. これは景観の見え方に違いがあり, その違いを正しく評価できているからではないかと思われる.

表 4: 経路評価結果

順位	ルート	満足度	対象が見えた枚数	経路のノード数
1	7	0.2094	12	37
2	0	0.1123	7	36
3	4	0.1013	8	43
4	2	0.0837	9	37
5	6	0.0746	10	74
6	5	0.0609	6	38
7	1	0.0505	4	43
8	3	0.0383	6	55

上位の経路の中に, 自動車では通ることができないような非常に狭い道路が経路に含まれていた. 景観としては満足度が高くても, 観光を目的にした移動という点においては満足度が低くなってしまう懸念がある. また今回は道路ノードからの景観を元に評価を行ったが, その景観がどの程度の間見えるのかが不明である. 撮影された道路ノードから隣接ノードまで同じような景観が得られない可能性があるが, 見えていると仮定し, 学習に隣接ノードまでの距離を加えることでより満足度が高い経路が得られると考えられる. 今後はそれに加え道路の幅や車線数, 右左折数を考慮した経路を構築・推薦するモデルの作成を目指す.

4 関 連 研 究

最短経路探索アルゴリズムとして, ダイクストラ法 [3] や A* アルゴリズム [4] があげられる. これらは重みが付与されたグラフに含まれる, 特定の 2 点間を結ぶ重みが最小となる経路を求める手法である. これらのアルゴリズムを基に多くの経路探索システムが考案されている.Noto ら [5] は従来のダイクストラ法を拡張し, 限られた時間内に解を求める方法を提案した.

本稿では短時間で複数経路を得るためにダイクストラ法にペナルティを与えたものを利用したが, 利便性や妥当性が低い経路が得られやすい. 櫻場ら [6] らは粘菌の特性を利用した複数経路探索を行った. 粘菌は自身の身体を流動的に動かしネットワークを作成することで捕食行動を行う. 作成したネットワークは最終的に 1 本の経路になるが, これを電気回路と見なし数理モデル化した. そのネットワークに道路状況を反映する交通基本図を利用し拡張を行うことで複数経路を含む部分ネットワークを取得した.

Zhang ら [7] は 2 地点の間にあるランドマークの視認性に着目した観光経路探索システムを提案している. 撮影地点からランドマークの方向へレンダリングしたときに前方にオブジェクトが視認できた際に,Z-Buffer によって前方のオブジェクトまでの距離 (深度) を算出する. その値がランドマークまでの距離と一致したときにランドマークが視認できているとしている. ランドマークが視認できた道路ノードの数を基準にした経路推薦を行っているが, 深度情報を用いるためどの程度の景観であるのかなどを推定することが難しいと思われる.

大瀧ら [8] は位置情報付きの画像を基にした, 目的に適したルートを複数提案する手法を提案した. 収集した画像からどのような景観が含まれているのかを抽出し, 道路ノードにその景観が見える確率を付与している. 最短経路を取れる確率と景観が見える確率から確率値が極大となるノード集合を集め, それらを結ぶことで良い景観を含む経路を推薦している.

川俣ら [9] は土地利用分類データと標高データを基に、中景・遠景を考慮した経路推薦を行うための可視領域抽出手法を提案した。土地利用分類データと標高データは国土交通省から国土数値情報として提供されているものである。山などの遠くの景観が建築物に阻害されずに見ることができる領域を推定した。クラウドソーシングで収集した評価を利用した。その結果、可視領域が同じ大きさであっても遠景が可視である地点からの方が、評価が高いことが分かった。しかし、利用した標高データが 250m ごとの 3D モデルデータであるが、土地利用分類データは利用区分ごとの面積をまとめたデータである。その影響から視点領域内に建築物が考慮されないなどの問題がある。

5 ま と め

本稿では,3D 都市モデルを利用した移動中の景観を考慮する観光経路推薦手法の景観モデルの評価と経路評価を行った. 状況を自動車の運転時に限定をして景観評価を行ったが, 停車時などにフロントガラスなどから覗き込んで景観を見るケースが考えられる. また, ユーザーが徒歩や自転車を利用することもある. そのような状況では正面からの景観以外に左右や見上げた角度も景観として利用するできるため, それらを含めた評価を今後行う必要がある.

ダイクストラ法を改良したものを用いて得た経路に対して評価を行い, その中から良い景観の経路を高い評価として求めた. しかし, 本来ならば経路の探索と景観を分離するのではなく, 景観の良さを踏まえた経路の探索をするべきであると思われる. それが可能となる推薦手法を考慮する必要がある.

今後, 移動経路を考慮に入れた経路推薦手法を行い, ユーザーがより満足できる経路推薦システムの構築を目指す.

文　　献

[1] 国土交通省. 新型コロナ感染症による行動変化が露わに。〜第 7 回全国都市交通特性調査 (速報版) の公表〜, 11 2022. https://www.mlit.go.jp/report/press/toshi07_hh_000206.html.

[2] 裕弥 内山. 国土交通省による project plateau(プラトー) の取組について : 3d 都市モデルの整備・活用・オープンデータ化プロジェクト, 05 2021.

[3] E.W. DIJKSTRA. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[4] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[5] M. Noto and H. Sato. A method for the shortest path search by extended dijkstra algorithm. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0*, volume 3, pages 2316–2320 vol.3, 2000.

[6] 悠之 櫻場 and 康雄 須貝. 利便性を考慮した力学系に基づく複数経路探索. **第 *80* 回全国大会講演論文集**, 2018(1):309–310, 03 2018.

[7] Jianwei Zhang, Hiroshi Kawasaki, and Yukiko Kawai. A tourist route search system based on web information and the visibility of scenic sights. In *2008 Second International Symposium on Universal Communication*, pages 154–161, 2008.

[8] 大瀧 隼 and 伏見 卓恭. 位置情報付き写真に基づく景観と距離を考慮したドライブルート探索手法. **人工知能学会研究会資料 知識ベースシステム研究会**, 126:22–27, 2022.

[9] 奥健太 川俣光司. 景観ベースルート推薦のための標高データに基づく可視領域抽出手法の検討. **データ工学と情報マネジメントに関するフォーラム**, 2019.

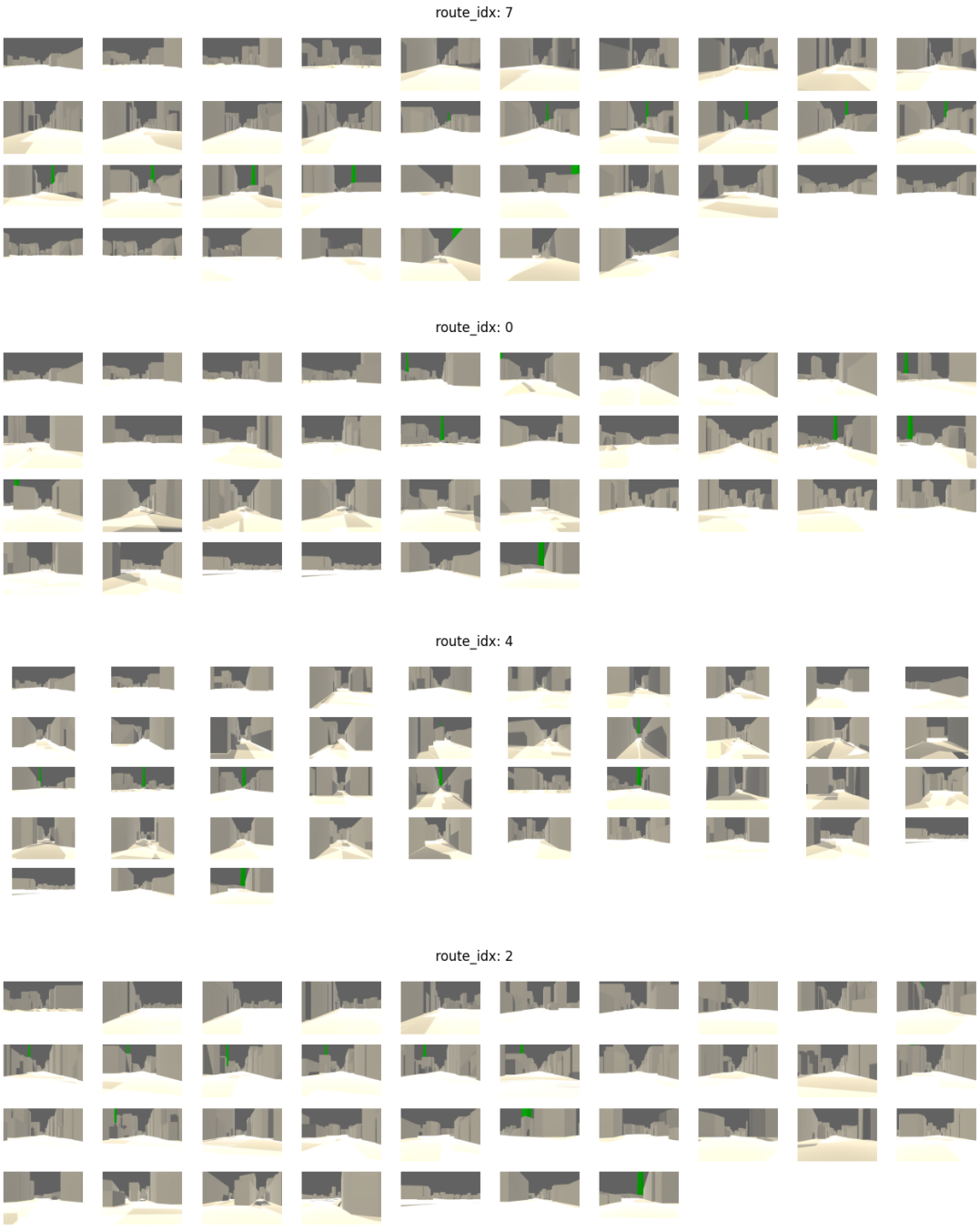


図 12: 満足度が高い上位の経路

個人化レベルのメリハリを考慮した観光ルートプランニング

廣澤凛太郎† 宜 鯤†† 馬 強†††

† 京都工芸繊維大学工芸科学部設計工学域情報工学課程 〒 606–8585 京都府京都市左京区松ヶ崎橋上町
†† 京都大学 経済研究所 〒 606–8501 京都府京都市左京区吉田本町
††† 京都工芸繊維大学 大学院工芸科学研究科 〒 606–8585 京都府京都市左京区松ヶ崎橋上町
E-mail: †b0122047@edu.kit.ac.jp, ††ykhnkf@outlook.com, †††qiang@kit.ac.jp

あらまし 推薦システムの個人化レベルは，非個人化，一時的個人化と永続的個人化の 3 段階に分類できる．観光において，個人に適応した推薦は観光客の満足度を向上させる効果があることが知られている．しかし，従来の観光推薦手法の多くは，これらの 3 段階の個人化レベルを固定して適用しており，個々の観光客のニーズを十分に満たすことが困難である．本研究では，スポットごとに個人化レベルを変化させて，スポット間での個人化レベルのメリハリがあるルートの実現を試みる．さらに，ルートの個人化レベルを人気スポットとユーザ好みスポットの割合で計算される連続値として表現し，個々の観光客の行動パターンに基づいて動的に調整できる観光ルートプランニング手法を提案する．また，本稿では，京都市における観光履歴データを用いた評価実験と考察について述べる．

キーワード 観光，ルート生成，プランニング，推薦，個人化レベル

1 はじめに

観光は，多くの国や地域にとって，経済的と社会的に重要な産業である．観光産業が世界の GDP に占める直接貢献の割合は 3.2 %．間接的および誘発的な影響を含めると 10.4 %にもなる [1]．このように観光は経済的に大きな影響を持っており，観光を促進することによって，地域の経済活動の活性化が期待できる．観光を促進するため，観光を行う際に訪れるスポットとそのルートを決定する観光ルートプランニングなどの観光支援システムが開発されている．これらのシステムは，初めての観光地における観光客の行動を支援して，満足度の高い観光体験の実現を試みる．

観光推薦において，各利用者に適応したスポットやルートを推薦する個人化の度合いは 3 段階に分類されるとされている [2] [3] [4]．第一段階は全観光客に対して同じ推薦を行う「非個人化」，第二段階は同じスポットに行った利用者に対しては同じ推薦を行う「一時的個人化」，第三段階は利用者の個人情報や過去の旅行履歴に応じて異なる推薦を行う「永続的個人化」である.「非個人化」による推薦では，全利用者に対して同じ推薦がされるため，一般的に人気なスポットが推薦され,「一時的個人化」,「永続的個人化」による推薦では，各利用者の好みに適応したスポットが推薦される．

京都を例に挙げると，伏見稲荷大社や金閣寺といった，全国的に知名度のあるスポットが「非個人化」，寺院などを好みとするユーザに対しては，延暦寺や慈善寺が「一時的個人化」，「永続的個人化」で推薦される可能性が高い．以下では，前者の低い個人化レベルで推薦されたスポットを人気スポット，後者の高い個人化レベルで推薦されたスポットをユーザ好みのスポットと記する．

高い個人化レベルで推薦されたスポット（ユーザ好みスポット）のみによるルートでは，自身の好みに適したスポットに訪れることはできるが，その観光地特有の代表的なスポットに訪れることができない可能性が生まれてしまう．逆に，低い個人化レベルで推薦されたスポット (人気スポット) のみによるルートでは，代表的なスポットは訪れることができるが，自身の好みに適していないスポットばかりになる可能性が生まれてしまう．このように，高い個人化レベルの推薦と，低い個人化レベルの推薦はトレードオフの関係にあり，推薦の際には個人化レベルを状況によって変化させる必要がある．また，観光スポットの中には人気でありながら，好みであるスポットも存在する．このようなスポットの優先的な推薦のために個人化レベルを柔軟に調整する仕組みが必要と考えられる．

したがって，ルートプランニングを行う際に観光客のニーズを満たすためには，ルート全体で個人化レベルを固定してスポットを選択するのではなく，異なる個人化レベルによる推薦を使い分けて，スポットを推薦する必要があると考えられる．しかし，既存の観光ルートプランニングとしては，与えられた時間予算内で，地域の知名度とユーザの好み度から算出される報酬の最大化を試みるものが多く，個人化レベルに関する考慮が不十分である．

そこで本研究では，それぞれの観光地に対して，人気や知名度による評価と，ユーザの好みによるかという評価の二種類の評価を行い，スポットごとにこれらの評価を使い分けることによって，異なる個人化レベルで推薦されたスポット含むルートの生成を試みている．提案手法の概要図を図 1 に示す．このように二種類の異なる評価法によって選ばれたスポットを用いることで，ルート内でのスポット間に個人化レベルのメリハリが生まれることを期待している．ここで言うスポット間のメリハリとは，ルート内のスポットの個人化レベルにメリハリがあり，人気スポットとユーザ好みのスポットが両立できている状態のことを指している．また，その際，ルート全体の個人化レ

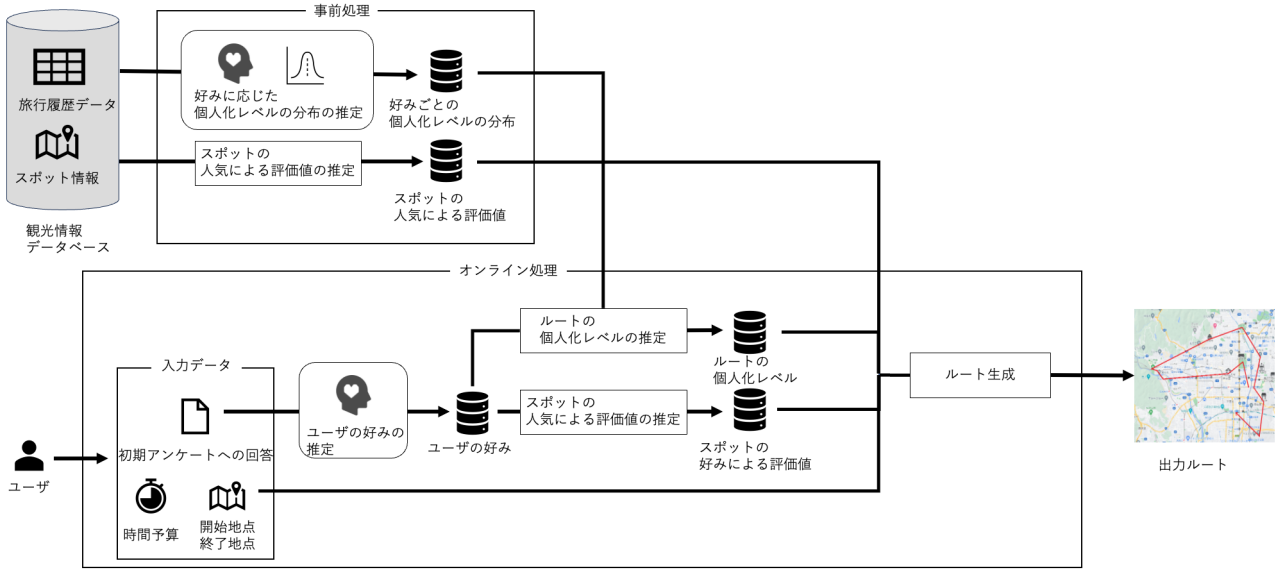


図 1: 提案手法の概要図

ベルを人気スポットと，ユーザ好みのスポットの割合で計算される連続値として表現し，この個人化レベルを個々の観光客の行動パターンに基づいて動的に調整することで，人気スポットとユーザ好みのスポットを適切な割合で含むルートの生成を試みている．

本研究における貢献を以下に示す．

（1） 個人化レベルを考慮したルートプランニング手法の提案 [4 節]．3 種類の個人化レベルによるスポットの推薦を組み合わせることで，任意の個人化レベルを持つ観光ルートを生成する手法を提案している．

（2） 観光の個人化を実現するため，スポット間のメリハリの概念を提唱し，それを実現する手法を提案 [4 節]．異なる個人化レベルの推薦によるスポットがルート内に存在することを，個人化レベルのメリハリと呼び，上記 3 つの個人化レベルによる推薦の割合を好みに応じて推定することで，ユーザに適したメリハリをつけた観光ルートを生成する手法を提案している．

（3） 京都市の旅行履歴データを用いて Ablation Study と比較実験を行い，提案手法が Precision，Recall と F1 スコアでの改善を確認している．

2 関 連 研 究

観光推薦に関する研究は数多く存在している．この節では，ルート推薦を中心に，個人化レベルごとに関連研究を紹介する．

2.1 個人化による観光推薦

個人の好みに対応する個人化を行った観光推薦の研究としては以下のものがある．Yi ら [5] [6] はユーザの行動を分析する確立生成モデルとそれに基づくユーザの求めるユーザ体験を推薦するシステムを提案している．ここで言うユーザ体験とは，行動の主体を表す Who，時間を表す When，場所を表す Where，そして行動自体を表す What からなるものであり，この 4 つの要素をトピックモデルを用いて分析することでその人の趣味嗜好および場所の特性を考慮して，ユーザ体験を推薦することを可能にしている．本研究では，これをスポットやルートの個人化レベルの評価に利用する．Lu ら [7] はユーザが撮影したジオタグ付き写真を用いて，ユーザの好みを推定し，ユーザの好みによってカスタマイズされたルートの生成を試みている．

2.2 非個人化による観光推薦

人気度を含む個人化を考慮しない指標による観光ルートプランニングの研究が多くある．Gama ら [8] は Pointer Network と強化学習による手法を用いて最適化問題を解くことで，少ない計算時間で満足度の高いルートの生成を試みている．森竹ら [9] は観光地を複数のエリアに分割し，認知度と固有度によって求められる代表スコアによって定められるコースをエリアごとに作成し，それらを組み合わせることで，エリア間に緩急のあるルートの生成を試みている．前川ら [10] や Kong ら [11] は観光地の混雑度合によって変わる滞在時間や，訪問への満足度を考慮することで，混雑を考慮したルートの生成を試みている．

2.3 両方を考慮した観光推薦

人気と個人化を考慮したルートプランニングの手法も数多く提案されている．Lim ら [12] はスポットをカテゴリごとに分け，全滞在時間のうちの滞在時間の割合を，そのカテゴリへの興味と捉え，全利用者の全滞在時間のうちのスポットへの滞在時間の割合を，そのスポットの人気と捉えている．そして，ユーザの旅行履歴から計算される重みを用いて両評価値の重み付き線形和を評価値として最適化問題を解くことで，二つの指標を両立したルートの生成を試みている．

Lu ら [13] は個人化を考慮していない評価値を各時間ごとのスポットの訪問数によって求めている．各カテゴリへの興味から求めたユーザ同士の類似度を用いた協調フィルタリングによる評価値と，システムを利用する観光客の友人のスポットへの訪問率による評価値を用いて，個人に適応した評価値を求めている．これらの評価値を用いることで，二つの指標を両立したルートの生成を試みている．

2.4 本研究について

これらの既存研究は，個人化レベルを固定して適用しており，異なる個人化レベルの推薦によるスポットを含むルートの生成には対応していない．そこで本研究では，異なる個人化レベルの推薦によるスポットをルートの中に組み込むことによって，さまざまなニーズを満たすことのできる個人化レベルのメリハリがあるルートの生成を目指す．

3 問題定義

3.1 用語・記号

本研究で使用する用語と記号の定義・説明を記述する．用語については以下の通りである．

- **スポット** p: 観光スポットのことであり，本研究では京都市とその周辺の観光スポット 91 か所のことを指す．
- **ルート** I: 訪問するスポットを時系列順に並べた列のことで，観光ルートプランニングにおける出力や旅行履歴に含まれる．$I = \{p_0, p_1, \ldots, p_n\}$
- **スポットの個人化レベル**: ルート内のスポットの個人化がどの程度行われているかを表す．本研究では人気，人気かつ好み，好みの三段階が低い順に存在すると定義する．
- **ルートの個人化レベル** $lel(u)$: ルート自体の個人化がどの程度行われているかを表す．ルート内に存在するスポットの個人化レベルの比率によって決まる値．3 種類のスポットの割合 r_1, r_2, r_3 を要素とする 3 次元のベクトル．$lel(u) = (r_1, r_2, r_3)$
- **旅行履歴** TH: あらかじめ処理しておく過去に実際に行われた観光のルートの軌跡データとその中で撮影された写真のデータ．
- **評価値 (スコア)** $S_{x,i}$: 各スポット p_x に個人化レベルごとに割り当てられる値で，ルート内に含まれるスポットの評価値を合計するとルート自体の評価値になる．
- **クエリ** $query$: 入力に使用する情報のうちユーザの好みを推定するための画像・スポットデータを除いた情報のことで，ルートの開始地 s，終了地点 d，時間予算 $Timebudget$ からなる．$query = (s, d, Timebudget)$ と記し，ユーザから指定されるルート生成の条件である．

記号は以下の表 1 の通りである．

表 1: 本研究で用いる記号

記号	意味
N	スポットの数，本研究では $N = 91$
U	履歴上のユーザの集合
G	UEM による趣味嗜好に応じたグループの集合
p_x	スポット x を示す変数
u	ユーザを示す変数，$u \in U$
g	UEM による趣味嗜好に応じたグループを示す変数，$g \in G$
$S_{x,i}$	スポット x の i 番目の理由による評価値 ($i = 1, 2, 3$)，$i = 0$ のとき，人気による評価値，$i = 1$ のとき，好みによる評価値，$i = 2$ のとき，それらの平均をそれぞれ示す．
r_i	ルート内での i 番目の理由の推薦によるスポットの割合 ($i = 1, 2, 3$)
$lel(u)$	ユーザ u の個人化レベル，r_1, r_2, r_3 からなる 3 次元のベクトル
$\hat{f}_i(x)$	グループ g_i のルートの個人化レベルの分布
$slel_i$	$\hat{f}_i(x)$ からサンプリングしたルートの個人化レベル (r_0, r_1, r_2)
$Time(p_x, p_y)$	スポット x から y への移動時間
$Timebudget$	ユーザの入力した時間予算
s	ユーザの入力したルートの開始地点
d	ユーザの入力したルートの終了地点
$query$	入力に必要な情報の一部をまとめたもの $(s, d, Timebudget)$
q_i	生成したルートの訪問した順を表す補助変数，部分巡回路を排除する制約条件に用いる

3.2 問題定義

観光ルートプランニングは，重要な産業である観光を促進するためのシステムとして盛んに研究されている．しかし，既存の観光ルートプランニングの研究では，固定した個人化レベルによる推薦によるルートの生成にしか対応していない．1 節でも述べたように固定された個人化レベルによる推薦のみでは，観光客のニーズを十分に満たすことはできない．したがって，ルート内でスポットごとに個人化レベルを変化させて推薦することで，個人化レベルでのメリハリがあるルートが生成できる．

そこで本研究では，各スポットを，観光における整数計画法の文脈 [12] [14] に基づいて以下の式 (1) をルートのスコアとして定義する．そして，ユーザの時間予算 $Timebudget$ とスポット s から開始し，スポット d で終了するという条件の中でスコアを最大化するルートを生成する．

$$\sum_{x=1}^{N}\sum_{y=1}^{N}\sum_{i=0}^{2}\sum_{j=0}^{2}(A(p_{x,i}, p_{y,j}) \times S_{x,i}) \qquad (1)$$

ここで，$A(p_{x,i}, p_{y,j})$ はユーザがスポット p_x を i 番目の理由で訪問した後，スポット p_y を j 番目の理由で連続して訪問する場合に 1 となるバイナリ変数である．i, j が 0 のときは人気を理由，1 のときは好みを理由，2 のときは両方を理由として訪問することを示す．この式 (1) を次のような制約のもと解く．

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{y=0\\ y\neq s}}^{N} A(p_{s,i},p_{y,j}) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{x=0\\ x\neq d}}^{n} A(p_{x,i},p_{d,j}) = 1 \tag{2}$$

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{y=0\\ y\neq s}}^{N} A(p_{d,i},p_{y,j}) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{x=0\\ x\neq s}}^{N} A(p_{x,i},p_{s,j}) = 0 \tag{3}$$

$$\forall w = 0,1,\ldots,N \quad w \neq s,d$$

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{x=0\\ x\neq w}}^{N} A(p_{x,i},p_{w,j}) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{\substack{y=0\\ y\neq w}}^{N} A(p_{w,i},p_{y,j}) \leqq 1 \tag{4}$$

$$\forall k = 0,1,2$$

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{x=0}^{N}\sum_{\substack{y=0\\ y\neq x}}^{N} A(p_{x,i},p_{y,j})\cdot r_k - \sum_{j=0}^{2}\sum_{x=0}^{N}\sum_{\substack{y=0\\ y\neq x}}^{N} A(p_{x,k},p_{y,j}) \leqq 1 \tag{5}$$

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{x=0}^{N}\sum_{\substack{y=0\\ y\neq x}}^{N} A(p_{x,i},p_{y,j})\cdot Time(p_x,p_y) \leqq Timebudget \tag{6}$$

$$q_i - q_j + 1 \leqq (N-1)\times(1-A(p_{x,i},p_{y,j}))$$
$$2 \leqq q_i, q_j \leqq N, \quad \forall i,j = 2\ldots N \tag{7}$$

[9] [12] [14] 等の既存の取り組みでは定義されていない制約，式 (5) を新たに考案し追加している．式 (2) と式 (3) は，ルートがスポット s から始まり，スポット d で終わるということをそれぞれ示す．式 (4) はルート内で経路が接続されており，同じスポットを 2 度以上訪れない制約を示す．式 (5) はルートの個人化レベルがあらかじめ決めておいたものに近くなる制約を表現している．式 (6) はルートに費やす時間が，ユーザの時間予算を超えない制約を，式 (7) は部分巡回路を排除する制約を表現している．

4 提案手法

本節では提案手法について説明する．4.1 節で事前処理について，4.2 節でシステムを利用する際の処理について説明する．

4.1 事前処理

この節では，システムを利用する前にあらかじめ行なっておく事前処理について説明する．

事前処理では既存の観光情報データベースから旅行履歴と観光スポットデータを取り出し，旅行履歴からは履歴上のユーザの好みと，人気スポットとユーザ好みのスポットの割合を推定する．観光スポットデータからは，人気による評価値を推定するといった処理を行う．

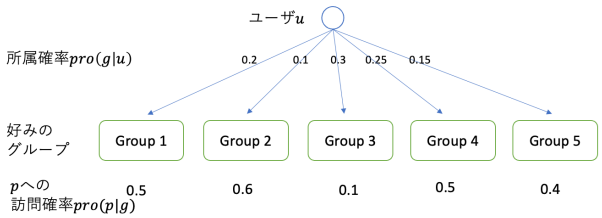


図 2: 所属確立と訪問確率の例

4.1.1 履歴上のユーザの好みの推定

好みの推定には，Yi らによるユーザ体験モデル (以下，UEM) を用いる [5] [6]．UEM では，確率生成モデルを用いてユーザの行動データからユーザごとの趣味嗜好を捉え，その人に合ったユーザ体験を推薦することができる．趣味嗜好を捉える際には，似ている趣味嗜好のユーザが同じグループに属するような複数のグループを生成し，各ユーザのそれらのグループに属する確率をユーザの趣味嗜好として捉えている．入力としては，ユーザの軌跡データに含まれるスポットの集合と画像データを使用する．出力は，ユーザの各グループへの所属確率と，各スポットへの訪問確率の二つとなる．グループへの所属確率は，入力にしようしたデータを持つユーザが UEM によって生成されたグループに所属する確率を示し，グループの数だけ生成される 0 から 1 の確率である．同じグループには似ている趣味嗜好のユーザが属するようなグループを生成しているため，本研究ではこの所属確率をユーザの好みとして扱っている．

各スポットへの訪問確率は，あらかじめグループへの所属確率と UEM の確率生成モデルから計算される，ユーザがそれぞれのスポットへ訪れる確率である．あるユーザ u のスポット p に対する訪問確率 $pro(p|u)$ は以下の式 (8) で計算される．

$$pro(p|u) = \sum_{g\in G} pro(p|g)pro(g|u) \tag{8}$$

図 2 はグループへの所属確立と各グループのスポット p への訪問確率の例である．この例の場合だと次式によってスポット p への訪問確率が求められる．

$$\begin{aligned} pro(p|u) &= \sum_{g\in G} pro(p|g)pro(g|u) \\ &= (0.2\times 0.5) + (0.1\times 0.6) + (0.3\times 0.1) \\ &\quad + (0.25\times 0.5) + (0.15\times 0.4) = 0.375 \end{aligned}$$

4.1.2 ルートの個人化レベルの分布の算出

観光客の旅行履歴上に存在するスポットを以下の 4 種類に分類する．

- P_{pop}: 人気によって選択したと推定されるスポット
- P_{pre} : 好みによって選択したと推定されるスポット
- P_{prp}: 人気と好みの両方によって選択したと推定されるスポット

- P_{oth}: 選択理由が明確でないスポット

その後，選択理由が明確でないスポット P_{oth} を除く三つのスポットの比率をその観光客の個人化レベルとする．分類基準として，後の 4.1.3 節で述べるスポットの評価値を用いる．人気による評価値が高いものをスポット P_{pop}，好みによる評価値が高いものをスポット P_{pre}，どちらの評価値も高いものをスポット P_{prp}，どちらの評価値も高くないものをスポット P_{oth} に分類する．

次に，履歴上のユーザ u のルートの個人化レベル $lel(u)$ と，i 番目のグループへの所属確率 $pro(g_i|u)$ から，グループごとに，3 種類のスポット P_{pop}, P_{pre}, P_{prp} のそれぞれの割合の同時確率分布 $\hat{f}_i(x)$ を生成する．分布の生成には，カーネル密度推定による手法を用いて，以下の式 (9) と式 (10) で行う．

$$\hat{f}_i(x) = \frac{1}{|U|h^3} \sum_{u \in U} pro(g_i|u) \phi \left(\frac{x - lel(u)}{h} \right) \tag{9}$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{10}$$

ここでは，h はバンド幅であり，この値が大きいほどデータ点が広範囲に影響を及ぼす．これにより，各グループそれぞれに対しての 3 種類のスポットの割合の同時確率分布 $\hat{f}_j(x)$ が生成できる．

4.1.3 スポットの評価値の推定

この節では，人気による評価値と好みによる評価値を生成する手法について述べる．

4 種類の人気による評価値の算出法を以下に示す．

- スポットの訪問率：全軌跡データにおけるそのスポットの訪問回数の割合である．この訪問率が高いほど，人気度が高いとする．以下の式 (11) によって訪問率 $Score_t(p_x)$ を求める．ここで，$visit(u, p_x)$ はユーザ u がスポット p_x に訪れた回数を表す．

$$Score_t(p_x) = \sum_{u \in U} \frac{visit(u, p_x)}{\sum_{i=1}^{N} visit(u, p_i)} \tag{11}$$

- ローカルリスティングによる評価：Google 検索や Google マップに表示されるスポットの情報をローカルリスティングと呼ぶ．スポットに対する評価の合計を得るため，平均評価と口コミの数の積を求め，その値が高いほど人気度が高いとする．以下の式 (12) によって評価値 $Score_l(p_x)$ を求める．ここで，$meanValuation$ は平均評価，$NumofComments$ は口コミ数を表す．

$$Score_l(p_x) = meanValuation \times NumofComments \tag{12}$$

- Normalized Google Distance(以下，NGD) [15]：NGD は Google 検索に駆けたときの検索結果の総数から，2 つの単語の意味のかかわりの強さ，共起度合いを示す指標である．NGD の求め方と，NGD を用いた評価値 $Score_g$ を生成する方法についてはこの節の最後で詳しく説明する．

- 検索ボリューム：検索ボリュームとはその単語が Google 検索によって検索された回数を示す数値である．スポット名に対する検索ボリュームが大きいほど，人気度が高いとする．以下の式 (13) によって評価値 $Score_v(p_x)$ を求める．ここで，$Volume$ は検索ボリュームを表す．

$$Score_v(p_x) = Volume \tag{13}$$

好みによる評価値は，4.1.1 節で求めた各スポットへの訪問確率を用いて求めた．

最後に，人気による評価値と好みによる評価値の両方に標準化を行なった．具体的には，Z スコア正規化を行い，どちらの評価値も平均 1，標準偏差 1 の正規分布となるような標準化を行なった．以下の式 (14) の $S_i(l)$ にスポット l に対する i 番目の理由による評価値を代入することで，標準化した値 $z(S_i(l))$ を求めることができる．

$$z(S_i(l)) = 1 + \frac{(S_i(l) - ave(S_i))}{sd(S_i)} \tag{14}$$

ここで $ave(S_i)$ は平均，$sd(S_i)$ は標準偏差を表している．なお，一般的な標準化では平均が 0 とするものが主流だが，今回使用するデータは評価値であり，なるべく正の値になることが望ましいため，平均を 1 とした．各スポットに対する評価値を $S_{x,i}$ で表す．$S_{x,i}$ は $i = 0$ のとき，人気による評価値，$i = 1$ のとき，好みによる評価値，$i = 2$ のとき，それらの平均をそれぞれ示す．

4.2 オンライン処理

この節では実際にシステムを利用する際の処理について説明する．ここでは，システムを利用する新しい観光客からの初期アンケートに対する回答と時間予算 $Timebudget$，ルートの開始地点 s と終了地点 d の希望を入力クエリ $(s, d, Timebudget)$ とし，観光ルートを出力としている．

まず，ユーザの初期アンケートの回答からユーザの好みを推定する．4.1 節で求めた軌跡データ・画像データから算出した人気スポットとユーザ好みのスポットの割合から，履歴上のユーザと新しい観光客の好みを比較し，生成するルート内のこれらのスポットの割合を決定する．それらをユーザの時間予算，開始地点，終了地点に応じてスポットの組み合わせを最適化することで，個人化レベルのメリハリを考慮した観光ルートを生成する．

4.2.1 新しい観光客の好みの推定

新しい観光客に対していくつかの画像を提示し，評価を行ってもらう初期アンケートを行う．この初期アンケートの回答を UEM [5] に入力し，4.1.1 節と同様にしてユーザの好みを求める．詳細については [5] を参照されたい．

4.2.2 ルートの個人化レベルの推定

この節では，新しい観光客の個人化レベルの推定手法につい

て述べる．新しい観光客に対してアンケートを実施して推定された好みと，あらかじめ旅行履歴から得た個人化レベルの分布から推定する．具体的に以下のように推定を行う．

（1） 4.1.2 節で生成した各グループ $g_i(i \in G)$ の分布 $\hat{f}_i(x)$ それぞれから値のサンプリングを行い，3 種類のスポットの割合からなるルートの個人化レベルを得る．グループ g_i から得たルートの個人化レベルをそれぞれ $slel_i = (r_1, r_2, r_3)$ とする．

（2） 初期アンケートによって得たシステム利用者 u のグループ g_i への所属確率 $pro(g_i|u)$ と，サンプリングした値の積の総和を求めることによって得られる割合を，ルートに存在する各 3 種類のスポットの割合とすることで，ルートの個人化レベル $lel(u)$ を決定する．人気によって選択するスポットの割合を r_1，好みによって選択するスポットの割合を r_2，両方を考慮して選択するスポットの割合を r_3 とする．具体的には以下の式 (15) によって求める．

$$lel(u) = \sum_{i=1}^{|G|}(slel_i \times pro(g_i|u)) \qquad (15)$$

4.2.3 ルートの生成

この節では，個人化レベルのメリハリを考慮したルートの生成手法について述べる．システム利用ユーザの入力した時間予算と開始・終了地点の条件と 4.2.2 節によるルートの個人化レベル，4.1.3 節と 4.2.1 節によるスポットの評価値から最も評価値の高くなるルートを 3.2 節で定義した最適化問題を既存のソルバー等 [16] を用いて生成する．

5 実　　験

本節では，4 節で述べた提案手法の有用性を立証するための実験について述べる．

5.1 データセット

京都市観光マップ 2.0 [17]，UEM [5] に用いられているデータセットを用いた．これは，Flickr に投稿された写真データをもとに京都府内の 91ヶ所の観光スポットの情報をまとめたものである．各スポットの滞在時間には，軌跡データから得られた平均滞在時間を用いたが，GoogleMap やスポットの公式 Web サイトに示されている滞在時間と比較したときに，2 倍以上の差がある場合は，後者の滞在時間を採用した．移動時間には，Openrouteservice [18] による移動時間を用いた．また，本研究では公共交通機関を用いた移動を想定しているため，車の移動時間に対しては 600 秒加算したものを移動時間として用いた．

5.2 実 験 方 法

5.2.1 ク　エ　リ

実験に用いるクエリとしては旅行履歴から生成した．具体的には，開始位置・終了位置として，旅行履歴内のルートの最初のスポットと最後のスポットを用い，時間予算として，開始位置から終了位置に移動するまでにかかった時間を用いた．

また，このとき移動にかかる時間が 8 時間以上 24 時間未満の移動があった場合，日を跨ぐ旅行である可能性や，間で別のことをしている可能性が高いため，5.1 節で述べた滞在時間と移動時間を用いて修正した．移動にかかる時間が 24 時間以上の場合，別日での旅行であると考えられるため，旅行履歴を分割することで各旅行のルートとなるように修正を行った．

好みの推定には 4.1.1 節で述べたとおり，旅行履歴に含まれるスポットと履歴内で撮影した写真を用いるが，生成したルートを旅行履歴と比較するため，推定には全てのデータを使わず，旅行履歴を 3 等分し，その内二つを入力として好みを推定し，ルートを生成する 3-fold の交差検証を行った．このとき，実験には平均旅行時間が約 3 時間 50 分の 221 個の旅行履歴から 663 個のクエリを生成した．

5.2.2 比 較 手 法

本研究の実験では，提案手法の Ablation Study，関連研究との比較の二つを行った．

Ablation Study では，個人化レベルに関する制約の有用性を確認するための比較を行う．具体的には，以下の四つのモデルを用いた．

- merihari：本研究の提案手法であるモデル
- merihari-nolevel：個人化レベルの制約を表す 3 節の式 (5) が含まれないモデル
- merihari-onlypop：人気による評価値のみによってルートを生成するモデル
- merihari-onlypref：好みによる評価値のみによってルートを生成するモデル

比較手法には，Gama ら [8] の Pointer Network と強化学習を用いた手法を用いた．また本研究のモデルとの比較を簡単にするため，以下の二つのモデルを用いた．

- Gama-pop：評価値として本研究の人気による評価値を用いたモデル
- Gama-pref：評価値として本研究の好みによる評価値を用いたモデル

また，ルート生成に利用した旅行履歴自体を Ground truth として用いた．

5.2.3 評 価 基 準

評価基準には以下の 4 種類の指標を使用した．

- 評価値
  ルートのスコア．各スポットの評価値を人気によるもののみとして計算したもの (pop)，好みによるもののみとして計算したもの (pref)，両評価値のうち高いものを選択して計算したもの (max)，両評価値の平均として計算したもの (mean) の 4 種類のスコアを算出した．
- 精度 (precision)
  入力に利用した旅行履歴と生成したルートの precision．生

表 2: 実験結果 (評価値の平均)

model	max	pop	pref	mean
original	9.6883	7.8247	8.4261	8.1254
merihari	16.9947	14.3643	12.9356	**13.6500**
merihari-nolevel	**17.2620**	14.7912	12.4796	13.6354
merihari-onlypop	16.6329	**15.4091**	10.8446	13.1268
merihari-onlypref	15.3003	11.7188	**14.0985**	12.9086
gama-onlypop	9.5881	8.3913	7.3015	7.8464
gama-onlypref	9.5336	7.2836	8.4197	7.8517

成したルートに含まれるスポットのうち，入力に使用した旅行履歴にも含まれるスポットの割合である．入力に使用したルートは実際にユーザが行った旅行のルートであるため，この値が高ければユーザに適したスポットを推薦できているということになる．ただしこの際，好みの推定に使用したスポット，開始地点，終了地点は既に入力として使われているため，precision の計算には使用しない．

- 再現率 (recall)
  入力に利用した旅行履歴と生成したルートの recall．入力に使用した旅行履歴に含まれるスポットのうち，生成したルートにも含まれるスポットの割合である．この値が高ければユーザの行きたかったスポットへの推薦が実際にできているということになる．precision と同様に，好みの推定に使用したスポット，開始地点，終了地点は，recall の計算には使用しない．

- F1-score
  precision と recall の調和平均を取った値．以下の式 (16) で求められる．

$$F\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \tag{16}$$

- 個人化レベルの MSE(Mean Squared Error)
  入力に利用した旅行履歴の個人化レベル $lel_{original}(u)$ と生成したルートの個人化レベル $lel_{generate}(u)$ の各要素の二乗誤差の平均 MSE．具体的には以下の式 (17) で求められる．

$$MSE = \frac{\|lel_{original}(u) - lel_{generate}(u)\|}{3} \tag{17}$$

5.3 実験結果

実験結果を表 2，3，4 に示す．表では生成した，または生成に利用した 221 個のルートの平均値を示している．また，表 2 の original は入力に利用した旅行履歴を表している．

5.3.1 評価値

最も評価値の平均が高いモデルは，pop による計算法では merihari-onlypop モデル，pref による計算法では merihari-onlypref モデル，max による計算法では merihari-nolevel モデル，mean による計算法では merihari モデルであった．ルート生成の際に使用した評価値による計算法のものが最も高くなるという結果となった．

表 3: 実験結果 (precision, recall, F1-score の平均)

model	precision	recall	F1-score
merihari	**0.0508**	**0.1474**	**0.0675**
merihari-nolevel	0.0503	0.1469	0.0667
merihari-onlypop	0.0447	0.1249	0.0587
merihari-onlypref	0.0512	0.1367	0.0663
gama-onlypop	0.0128	0.0343	0.0172
gama-onlypref	0.0325	0.0585	0.0358

表 4: 実験結果 (個人化レベルの MSE の平均)

model	personalize level MSE
merihari	**0.0833**
merihari-nolevel	0.0850
merihari-onlypop	0.0881
merihari-onlypref	0.0892
gama-onlypop	0.1224
gama-onlypref	0.1091

Gama 手法では，少ない計算時間で問題を解くために強化学習と Pointer Network を用いているため，数理最適化問題を解くための高性能なソフトウェアパッケージである gurobi を用いた本研究の手法に比べ，ルート生成に必要な時間は本研究による手法に比べ大幅に短くなっていたが，評価値が低くなってしまっていると考えられる．

5.3.2 precision/recall/F1-score

precision，recall，F1-score の平均値は本研究のモデル merirhari モデルが最も高いという結果が得られた．この結果は，入力に使用した元となる旅行履歴に近いルートを作る能力，すなわちユーザに適したスポットを推薦する能力が最も優れているのは本研究の merihari モデルであることを示している．

5.3.3 個人化レベル

入力に使用した元となる旅行履歴の個人化レベルと生成したルートの個人化レベルの平均二乗誤差は，本研究の merihari モデルが最も小さいという結果が得られた．この結果は，最もユーザに適した個人化レベルのルートを生成する能力に長けているのは本研究の merihari モデルであることを示している．

5.3.4 考察

これらの結果の merihari モデルと merihari-nolevel モデルを比較すると，precision/recall/F-score の値が高くなっていることからわかるように，個人化レベルによる制約はユーザが行きたいと感じるスポットを推薦することに有用であることが分かる．また，個人化レベルの平均二乗誤差が低くなっていることから，ユーザに適した個人化レベルのルートを生成することにも有用であると考える．評価値に関しては，max の評価値は merihari-nolevel に，pop の評価値は merihari-onlypop に，pref の評価値は merihari-onlypref に比べて，提案手法である merihari モデルは低いという結果となっている．これは，それぞれルート生成の際に使う評価値が異なるためであるからだと

考えられ，今後さらなる調査を行う予定である．

以上の結果より，個人化レベルによる制約を考慮した本研究の手法は各個人に適したルートを生成する際に有用であると考える．

6 ま と め

本研究では，異なる個人化レベルによって推薦されたスポットを同時に含む「個人化レベルのメリハリを考慮したルート」を生成する手法を提案している．ルート生成を行う前に，ルートの個人化レベルをユーザに合わせて決め，スポットごとに個人化レベルを変化させることで，各ユーザに適した個人化レベルのルートを生成することを可能とする．

実験では，入力として実際の観光客の観光履歴を用いて，個人化レベルによる制約を分析する Ablation Study と，ルート生成の既存手法と比較を行った．結果として，提案手法は他手法と比べて，ルート生成に利用した観光履歴と近い個人化レベルや訪問スポットを持つルートの生成が可能であるという結果が得られた．ルートの評価値についても，他手法と比べて遜色ない値を得ることができた．これらのことから，個人化レベルを考慮するルートプランニングは高い満足度のルートを生成することに有用であると考える．

今後の展望としては，新たな個人化レベルの実装とユーザの観光履歴の利用があげられる．

- 本研究ではスポットの個人化レベルとして，人気スポット，ユーザ好みのスポットと人気かつ好みのスポットという 3 段階のものを用いた．しかし，実際の観光では訪問したスポットの近くのスポットに対して好みとも人気とも言えないが，訪問することはあるだろう．このようなスポットの推薦に対応するため，推薦システムの個人化の度合いにおける,「一時的個人化」にあたる個人化レベルの実装を行う手法が考えられる．

- 生成するルートの個人化レベルを推定する際に，他の観光都市に対してのユーザの観光履歴を用いることを可能にすることで，よりユーザに適した個人化レベルを推定することができるようになるだろう．

謝　　辞

本研究の一部は科研費（23H03404）による

文　　献

[1] JICA 国際協力機構，“観光の今を知る，” 2018. https://www.jica.go.jp/Resource/publication/mundi/1807/201807_02_02.html. 2023 年 12 月 21 日 閲覧.
[2] 神嶌敏弘，“推薦システムのアルゴリズム (1)，” 人工知能学会誌，vol.22，no.6，pp.826–837，2007.
[3] 神嶌敏弘，“推薦システムのアルゴリズム (2)，” 人工知能学会誌，vol.23，no.1，pp.89–103，2008.
[4] 神嶌敏弘，“推薦システムのアルゴリズム (3)，” 人工知能学会誌，vol.23，no.2，pp.248–263，2008.
[5] K. Yi, R. Yamagishi, T. Li, Z. Bai, and Q. Ma, “Recommending pois for tourists by user behavior modeling and pseudo-rating,” CoRR, vol.abs/2110.06523, 2021.
[6] 山岸立，馬強，“個人適応型観光のためのユーザ体験推薦，” 日本データベース学会和文論文誌，vol.20-J，no.15，pp.1–8，2022.
[7] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, “Photo2trip: Generating travel routes from geo-tagged photos for trip planning,” ACM MM 2010., p.143–152, Jan. 2010.
[8] R. Gama and H.L. Fernandes, “A reinforcement learning approach to the orienteering problem with time windows,” Comput. Oper. Res., vol.133, p.105357, 2021.
[9] S. Moritake, H. Kasahara, and Q. Ma, “Merihari-area tour planning by considering regional characteristics,” DEXA(2), pp.49–64, 2023.
[10] T. Maekawa, H. Kasahara, and Q. Ma, “Tour route generation considering spot congestion,” DEXA(1), pp.478–492, 2023.
[11] K. Yuntao, P. Chen, N.M. Le, and M. Qiang, “Dual congestion-aware route planning for tourists by multi-agent reinforcement learning,” DEXA(2), pp.331–336, 2023.
[12] K.H. Lim, J. Chan, C. Leckie, and S. Karunasekera, “Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency,” Knowl.Inf.Syst, vol.54, no.2, pp.375–406, 2018.
[13] H.-C. Lu, C.-Y. Chen, and V. Tseng, “Personalized trip recommendation with multiple constraints by mining user check-in behaviors,” SIGSPATIAL/GIS 2012, pp.209–218, 2012.
[14] C.E. Miller, A.W. Tucker, and R.A. Zemlin, “Integer programming formulation of traveling salesman problems,” Journal of the ACM, vol.7, no.4, pp.326–329, 1960.
[15] S. Karve, V. Shende, and S. Hople, “Semantic relatedness measurement from wikipedia and wordnet using modified normalized google distance,” Data Analytics and Learning, pp.143–154, 2019.
[16] Gurobi Optimization, LLC, “Gurobi Optimizer Reference Manual,” 2023.
[17] J. Xu, J. Sun, T. Li, and Q. Ma, “Kyoto sightseeing map 2.0 for user-experience oriented tourism,” MIPR 2021, pp.239–242, 2021.
[18] “Openrouteservice”. https://openrouteservice.org/.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 16:00 ~ 18:10 | T1-A オンライン（Zoom Events)

構造データ処理

座長:神崎 映光(島根大学)
コメンテータ:伊藤 寛祥(筑波大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T1-A-6-01] マウス版Geneformerの構築と転移学習による細胞解析への応用

*伊藤 啓太[1]、平川 翼[1]、山下 隆義[1]、藤吉 弘亘[1] (1. 中部大学)

16:30 ~ 16:55
[T1-A-6-02] ペルソナ・モデリングによるソーシャルネットワークの将来予測の精度向上

*松本 寛智[1]、伊藤 寛祥[2]、三宅 健太郎[1]、森嶋 厚行[2] (1. 筑波大学情報学群、2. 筑波大学図書館情報メディア系)

16:55 ~ 17:20
[T1-A-6-03] 時間的特徴を制御可能な時間グラフ生成器

*中西 宏和[1]、菊地 良将[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学情報科学研究所マルチメディア専攻鬼塚研究室)

17:20 ~ 17:45
[T1-A-6-04] 時間グラフ生成に関する包括的な調査とベンチマーキング

*菊地 良将[1]、中西 宏和[1]、佐々木 勇和[1]、Yakovets Nikolay[2]、Fletcher George[2]、鬼塚 真[1] (1. 大阪大学情報科学研究科、2. Eindhoven University of Technology)

17:45 ~ 18:10
[T1-A-6-05] ExhaustiveTree: 網羅的な二次元座標木を用いた表の表現

*林 知司[1]、宮森 恒[2] (1. 京都産業大学大学院 宮森研究室、2. 京都産業大学大学院)

マウス版 Geneformer の構築と転移学習による細胞解析への応用

伊藤　啓太† 　平川　翼† 　山下　隆義† 　藤吉　弘亘†

† 中部大学　〒 487–0027 愛知県春日井市松本町 1200

E-mail: †{itokeita,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

あらまし　Single-cell RNA-sequence 解析は，細胞 1 つ 1 つの遺伝子発現量を次元圧縮やクラスタリングなどの数理的な手法を利用して解析する手法である．この解析により，新規の細胞種や異常な遺伝子の特定が可能となった．Single-cell RNA-sequence 解析手法には，機械学習や深層学習の技術を用いた手法も存在し，深層学習の技術を利用した解析手法として Attention ベースの事前学習モデル Geneformer がある．しかし，Geneformer は，人の Single-cell RNA-sequence データを用いて事前学習を行っているため，ファインチューニング時に，マウスの Single-cell RNA-sequence データを用いることはできない．そこで，本研究では，マウスの Single-cell RNA-sequence データを用いて事前学習したモデルとしてマウス版 Geneformer を構築する．これにより，従来解析手法よりも高精度に細胞型の分類を行うことが可能であることを示す．また，in silico 摂動実験により病気の原因遺伝子を同定することが可能であることを示す．

キーワード　自然言語処理応用，医療・ヘルスケア，Single-cell RNA-sequence 解析，Geneformer

1 はじめに

病気や障害の原因は，異常な細胞の遺伝子情報をタンパク質として発現していることである．病気や障害の原因究明には，異常なタンパク質を発現している異常な細胞を特定する必要がある．その代表的な解析手法として，Single-cell RNA-sequence 解析 [1] がある．Single-cell RNA-sequence 解析を用いることで，遺伝子の発現状態に基づいた細胞型の分類や，新規の細胞種の同定，異常な細胞の推定を行うことが可能である．遺伝子発現量の解析には，Cell Ranger [2] や Seurat [3] のような信頼性の高い解析ツールを使用することが主流となっている．インフルエンザや日本脳炎，風邪といった治すことが可能な病気は，現在の解析手法で原因遺伝子を発見することが可能である．しかし，アルツハイマー型認知症やパーキンソン病といった，疾患の原因が完全に判明していない病気は難病とされており，現在の解析手法では原因遺伝子を発見することができていない．特に，アルツハイマー型認知症は高齢者に多く，認知症全体のうち 60% を占めており，重要な疾患の 1 つであると言える．このことから，現在の Single-cell RNA-sequence 解析手法に代わる新しい解析手法の需要が高まりを見せている．

新しい Single-cell RNA-sequence 解析手法として，機械学習や深層学習などの技術を取り入れた手法が注目されている [4]．その中の 1 つに，Geneformer [5] がある．Geneformer は，自然言語処理の分野で広く用いられている Attention ベースの深層学習モデル BERT [6] のように Transformer Encoder を利用した事前学習モデルである．Geneformer は，Attention 機構により遺伝子間の関係を学習する．そのため遺伝子ネットワークの構造を理解することが可能となる．転移学習により，限られた環境下でのデータを用いて，事前学習で作成した遺伝子ネットワークをファインチューニングすることで，生物学実験と同様のシミュレーション実験が可能となり，病気の原因遺伝子を特定するまでのプロセスを効率化することが可能となる．さらに，治療法の発見や新薬の開発などにもつなげることも可能となる．

Geneformer [5] は，人の単一細胞データを事前学習に使用している．そのため，人以外の単一細胞データを転移学習に使用しては，Geneformer の性能を最大限に発揮することができない．また，遺伝子操作などの生物実験は，人よりもマウスを使用するのが一般的である．そのため，単一細胞データはマウスの方が豊富であり，疾患マウスの単一細胞データの取得も容易である．このことから，マウスの単一細胞データを事前学習に使用した Genformer であるマウス版 Geneformer は，Single-cell RNA-sequence 解析を用いた疾患研究の発展には，非常に有用であると言える．

そこで，本稿では，マウス版 Geneformer の構築を目的とする．マウス版 Geneformer の構築により，マウスの単一細胞データを正確に細胞型に分類することが可能となる．また，疾患マウスモデルを使用した，生物学実験をシミュレーション実験する in silico 摂動実験が可能になる．そして，疾患マウスモデルの in silico 摂動実験により，疾患モデルがマウスしか存在しない疾患の原因遺伝子を同定することが可能となる．

本稿では次の構成をとる．2 節で 従来の Single-cell RNA-sequence 解析と新しい解析手法の研究を説明し，3 節でマウス版 Geneformer の構築について述べる．続いて，4 節で評価実験の概要および実験結果について説明し，5 節で本稿をまとめる．

2 関連研究

本章では，Single-cell RNA-sequence 解析，従来の解析手法

について，機械学習と深層学習を取り入れた解析手法について，Attention ベースの解析手法についてまとめる．

2.1 Single-cell RNA-sequence 解析

Single-cell RNA-sequence 解析 [1] とは，個々の細胞が保持している messenger RNA (mRNA) 全体を質的，量的に網羅的に調べる手法である．これは，遺伝子の発現状態に基づいた細胞型の分類を，次元削減やクラスタリングとなどの数理学的解析を用いることにより可能としている．遺伝子の発現とは，細胞核内の DNA の塩基配列に基づき，細胞ごとにさまざまな生体機能を持つタンパク質を合成し，遺伝子情報が具体的に現れることである．Single-cell RNA-sequence 解析により，今まで知り得なかった希少な細胞の検出や，細胞の遺伝子発現量の変化，病気の原因となる異常な細胞，遺伝子の抽出が可能となった．

2.2 従来の Single-cell RNA-sequence 解析手法

1 つの細胞が持つ整体物質を解明し，定量しようとする試みは古くからあり，1960 年代に蛍光活性化セルソーティングが開発された．蛍光活性化セルソーティングは，識別抗体などのプローブとの組み合わせにより，多数の細胞集団の中で 1 つの細胞が保持している生体分子の種類や量について断片的な研究が可能となった．そして，2006 年に次世代シーケンサーが誕生した．次世代シーケンサーとは，数千から数百万もの DNA の塩基配列を同時に解析することが可能な技術である．この技術が開発されたことにより，塩基配列単位，細胞単位での遺伝子発現量の解析が容易になった．次世代シーケンサーの誕生により，希少な mRNA やノンコーディング RNA を含めた未知の転写産物の高感度検出が可能となった．また，mRNA 前駆体から不必要な塩基配列が切り離され，mRNA が熟成していく過程など，転写産物の種類だけでなく，転写産物の構造的差異の解析も可能となった．

次世代シーケンサーにより，個々の細胞が持つ遺伝子発現パターンを獲得することが可能となった．この技術を利用した Single-cell RNA-sequence 解析として遺伝子発現パターン解析技術 Cell Ranger [2] や，Seurat [3] が誕生した．これらは，個々の細胞の遺伝子発現パターンを次元削減やクラスタリングと言った数理学的解析を用いて，細胞の分類や異常な遺伝子の抽出を行うソフトウェアである．医学界では，信頼性の高い解析ツールとして様々な解析に広く用いられている．

2.3 機械学習や深層学習を取り入れた解析手法

機械学習や深層学習などの情報分野の技術も急速に発展している．機械学習は，教師あり学習と教師なし学習に分類され，教師あり学習として，サポートベクタマシン（SVM）や勾配ブースティング決定木が研究された．この技術を Single-cell RNA-sequence 解析に取り入れた手法が，CaSTLe [7] である．CaSTLe が誕生した時の単一細胞データのアノテーション手法は，手動でクラスタリングを行いアノテーションをするか，蛍光活性化セルソーティングを用いてアノテーションを行っていた．しかし，これらはアノテータの知識に依存したり，別で実験が必要なため，Single-cell RNA-sequence 解析に組み込むことができなかった．この課題を解決する手法として CaSTLe が誕生した．CaSTLe は，一般でも入手可能な 9 つの単一細胞データセットを用いて事前学習したモデルである．転移学習を用いて異なる単一細胞データセットのモデルを構築し，細胞型の分類を実現した．具体的には，一変量特徴選択手法を用いて特徴量を抽出し，弱学習器として決定木を用いた XGBoost 分類機 [8] を構築する．この手法は，単一細胞データのラベリングをより効率的かつ正確に行うことが可能となり，転移学習により事前学習時には存在しなかった細胞型の分類も行うことが可能である．

深層学習では，Convolution Neural Network (CNN) [9] が研究された．CNN は，深層学習モデルの 1 つで，畳み込み層，プーリング層，全結合層で構成され，画像認識やパターン認識などに利用されている．CNN を Single-cell RNA-sequence 解析に取り入れた手法として scDeepInsight [10] が存在する．この手法は，単一細胞データのような非画像データを画像に変換し，深層学習モデルに適用する手法が注目されるようになり，単一細胞データを視覚的に理解しやすくする手法が求められたことで誕生した手法である．scDeepInsight では，はじめに単一細胞データに対してフィルタリングと正規化を実施し，主成分分析や t-SNE [11] や UMAP [12] などで 2 次元に圧縮し，2 次元に埋め込んだ単一細胞データを画像に変換する．画像に変換した単一細胞データを EfficientNet-b3 [13] などの CNN モデルを用いて，特徴を抽出し，細胞型の識別を行う．また，scDeepInsight は，DeepInsight [14] という，遺伝子発現データを 2 次元の空間に埋め込むための手法に基づいて設計されている．scDeepInsight は，画像処理の分野で活躍している CNN ベースのモデルを活用しているため，他の従来手法に比べて細胞型の分類精度が高く，希少な細胞型や未知の細胞型を検出することも可能である．

また， グラフニューラルネットワーク (GNN) [15] の研究は 2010 年代から急速に発展していった．GNN の中でも重み付きグラフニューラルネットワークを用いた手法として，scDeepSort [16] が存在する．この手法は，従来の細胞型のアノテーションで必要であったマーカー遺伝子や RNA-sequence プロファイルなどを必要とせず，容易に細胞型のアノテーションを可能にした手法である．GNN を使用する理由は，単一細胞データは細胞と遺伝子発現量の行列として表現され，各行が細胞を表し各列が遺伝子を表す．このような行列データは，潜在的にグラフ構造を持っており，GNN を使用することで細胞間の関係性を考慮した細胞型のアノテーションが可能になるためである．scDeepSort は，単一細胞データのための事前学習済みの細胞型の分類手法であり，重み付きグラフニューラルネットワークを用いた深層学習モデルである．このモデルは，埋め込み層と重みグラフ集約層と線形分類層で構築されいる．埋め込み層では細胞と遺伝子を同じ次元に圧縮を行い，重みグラフ集約層では GraphSAGE [17] をバックボーンの GNN フレームワークとして使用し，新しい細胞表現を生成する．線形分類層では，新しい細胞表現を事前定義された細胞型に分類する．この手法

は，高い細胞型の分類性能と堅牢性を示し，細胞型のアノテーション手法として新しいアプローチを示した．

深層生成モデルでは，AutoEncoder (AE) [18]，Variational AutoEncoder (VAE) [19] や Gaussian-mixture VAE (GMVAE) [20] が研究された．これらの手法は，異常検知の分野で成功を収めている．深層生成モデルの VAE や GMVAE を Single-cell RNA-sequence 解析に利用した手法として，Single-cell VAE (scVAE) [21] が存在する．従来の解析手法では単一細胞データに対して次元削減を行い，情報を抽出する作法が一般的であった．しかし，それでは単一細胞データの希少で重要な情報を失う可能性があり，解析結果に影響を与える．そこで，次元削減前の生の単一細胞データを直接使用する手法として scVAE が誕生した．scVAE は，Single-cell RNA-sequence 解析用にカスタマイズされた VAE や GMVAE を構築した．通常の VAE や GMVAE との違いとして，使用する尤度関数をポアソン分布や負の二項分布などのカウント尤度関数を使用した点である．これは，生の遺伝子発現量データが離散的でかつスパースなデータであり，連続的なデータでないためである．負の二項分布を使用した GMVAE は，潜在空間でのクラスタリングが生物学的に適した細胞型の分布となっていることを示した．また，scVAE は，単一細胞データ内に存在する Human Error により生成されたデータをネットワーク内で除去するため，ノイズとなるデータを予め除去する必要がない．そのため，結果を得るためのプロセスが簡略化され，細胞の専門家でなくても解析が可能となった．

2.4 Geneformer

自己教師あり学習は，近年注目を集めている研究分野である．自己教師あり学習では，データから正解ラベルを自動で生成できるプレテキストタスクを用いて事前学習を行い，様々な下流タスクにファインチューニングすることで，より良い特徴量を抽出することが可能となる．そして，2018 年に自然言語処理の分野で Attention ベースの事前学習モデル BERT [6] が発表された．BERT は，Transformer Encoder [22] を使用した構造となっている．この Transformer Encoder を Single-cell RNA-sequence 解析に利用した手法が Geneformer [5] である．この手法は，遺伝子ネットワークのマッピングに焦点を当て，遺伝子間の関係を学習する手法である．Geneformer は，大規模な人の単一細胞データをテキストとして表現し，そのテキストデータを用いて事前学習した Attention ベースの深層学習モデルである．Geneformer は，Transformer Encoder [22] を使用した構造となっている．そのため，Attention 機構により遺伝子間の関係性の学習が可能となり，人の細胞の遺伝子ネットワークを正確に予測することが可能となる．正確に遺伝子間の関係を学習するために，遺伝子のトークン化には，細胞の状態を示す遺伝子を積極的にトークンとして使用する Rank Value Encoding を使用している．Geneformer は，転移学習することで限られた単一細胞データの遺伝子ネットワークを予測することが可能となり，従来の解析手法よりも細胞型の分類精度が向上することを示した．また，実際の生物学的実験をシミュレートすることが可能となる in silico 摂動実験が可能となり，実際の人の心筋症モデル [23] を用いて，心筋症の原因遺伝子を同定するなどして，有効性を示した．

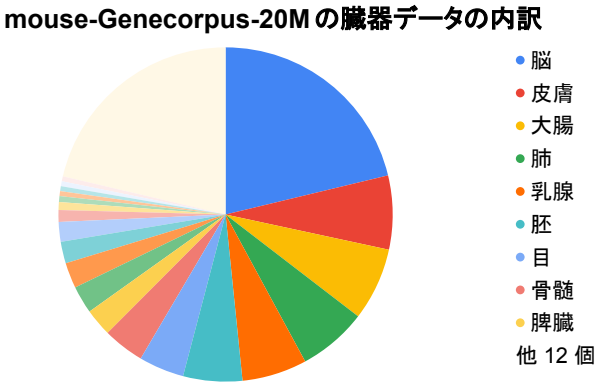


図 1　mouse-Genecorpus-20M の臓器データの内訳

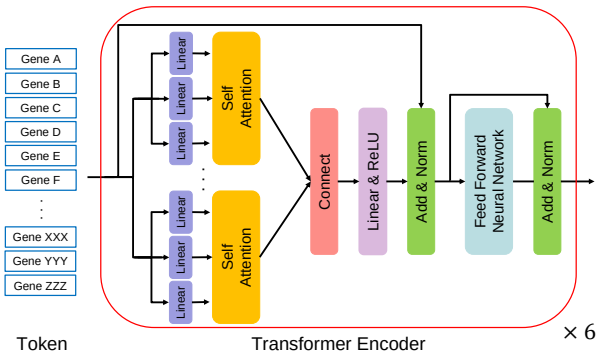


図 2　マウス版 Geneformer のネットワーク図

3　マウス版 Geneformer

本章では，マウス版 Geneformer の事前学習に使用したデータセットである mouse-Genecorpus-20M やアーキテクチャ，マウス版 Geneformer の事前学習について述べる．

3.1 mouse-Genecorpus-20M

mouse-Genecorpus-20M は，マウス版 Geneformer を事前学習するときに使用するデータセットである．このデータセットは，一般に公開されているマウスの Single-cell RNA-sequence データベースから，正常なマウスの全身の生の単一細胞データを約 2,300 万個取得し，構築している．mouse-Genecorpus-20M の構築に使用したマウスの臓器の内訳を図 1 に示す．また，主に使用した Single-cell RNA-sequence データベースを以下に示す．

- PanglaoDB
- Single Cell Expression Atlas
- Single Cell Portal
- ENCODE project
- 10x Genomics

生の単一細胞データは遺伝子発現量マトリックスとして表現されている．遺伝子発現量マトリックスは，生物学や遺伝学

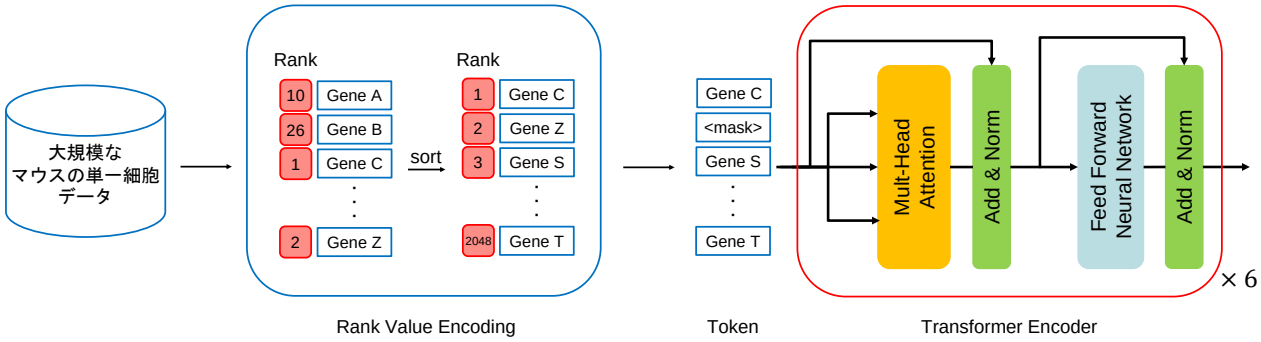


図 3　マウス版 Geneformer のアーキテクチャ

の研究において重要なデータ構造の一つである．行方向に異なる遺伝子を，列方向に異なる細胞を表し，対応する細胞における遺伝子の発現量が数値として記載されている．mouse-Genecorpus-20M は，正常なマウスの遺伝子ネットワークを学習させるための単一細胞データのコーパスなため，癌細胞などの遺伝子ネットワークを再編する原因となる単一細胞データは除外する必要がある．

3.2　アーキテクチャ

マウス版 Geneformer は，大規模なマウスの単一細胞データを用いて事前学習した，Attention ベースの深層学習モデルである．マウス版 Geneformer のネットワーク図を図 2 に示す．転移学習をすることで，希少な細胞や遺伝子が存在するデータの遺伝子ネットワークを正確に予測することが可能である．マウス版 Geneformer は，Attention 層とフィードフォワードニューラルネットワーク層を持つ Transformer Encoder を 6 つ使用した構成となっている．Attention Head の数は 4 つである．トークンには，細胞状態を正確に表現する 2,048 個の遺伝子を使用している．よって，1 層目の Transformer Encoder の入力次元数は 2,048 次元である．出力次元数で 256 次元であり，2 層目以降の Transformer Encoder の入力と出力の次元数は 256 次元である．

3.3　事 前 学 習

マウス版 Genformer の概要図を図 3 に示す．マウス版 Geneformer の事前学習には，大規模なマウスの単一細胞データを用いる必要がある．そのため，はじめに，mouse-Genecorpus-20M を作成した．mouse-Genecorpus-20M は，癌細胞などの遺伝子ネットワークを再編する原因となる単一細胞データは除外する必要がある．

次に，mouse-Genecorpus-20M 内のデータをフィルタリングする．フィルタリングは，細胞に由来しない RNA や死細胞，複数細胞が混じった細胞（Doublet），空の液滴等で取得したデータを除外するために行う．これらの単一細胞データは必ず存在する．これらの細胞データは，Single-cell RNA-sequence 解析においてノイズとなるデータであり，細胞状態を正確に表現する特徴量の獲得を困難にする．これらから，フィルタリングを行う必要がある．具体的なフィルタリング条件を以下に示す．

- 各データセットにおいて，遺伝子の総発現量の平均から 3 標準偏差以外の細胞を除去
- 各データセットにおいて，ミトコンドリア RNA の遺伝子発現量の平均から 3 標準偏差以外の細胞を除去
- 各データセットにおいて，1 細胞で検出する遺伝子数が 7 未満の細胞を除去
- 各細胞の遺伝子発現総数が 20,000 超過の細胞を除去

フィルタリング後の単一細胞データ数は，22,446,161 細胞となり，この単一細胞データを事前学習に使用する．

次に，フィルタリング後の mouse-Genecorpus-20M からトークンを作成する．トークンには，細胞の状態を示す遺伝子を使用する．トークン化には，トークンとして使用する遺伝子を抽出する Rank Value Encoding を使用する．Rank Value Encoding は，mouse-Genecorpus-20M 内での遺伝子発現量に基づいて，各細胞内に存在する遺伝子に対してランク付けを行う．このトークン化手法により，細胞状態を区別するための遺伝子を優先的に選択することが可能となる．具体的な手順を以下に示す．

1. mouse-Genecorpus-20M の全細胞の非ゼロ遺伝子発現量から，各遺伝子の非ゼロ中央値を計算する
2. 各細胞の各遺伝子発現量を，その細胞の総遺伝子発現量で除算し，正規化をする
3. 正規化した遺伝子発現量を，その遺伝子の非ゼロ中央値で除算し，さらに正規化をする
4. 正規化した遺伝子発現量を 10,000 倍し，大きい順にソートし，ランク値を設定する

そして，ランク値 2,048 までの遺伝子が細胞の状態を示す遺伝子であり，トークンとして使用される．ランク値は 1 に近づくにつれて細胞状態をよく示す．

作成した mouse-Genecorpus-20M のトークンを用いて Geneformer を学習させる．学習に使用するプレテキストタスクには，マスクトークンの予測タスクを使用する．このタスクは，入力されたトークンの一部をランダムにマスクし，そのマスクトークンを他のトークンを用いて予測するタスクである．このタスクによりマウス版 Geneformer は，マウスの遺伝子の関連性や発現パターンを学習することが可能であると考える．事前学習では，各細胞のトークンをランダムに 15% の確率でマスクし，残りのマスクされていないトークンを用いてマスク

したトークンを予測させるようにネットワークの最適化を行う．

4 評価実験

本章では，マウス版 Geneformer の有効性と頑健性を示すための評価実験を行う．

4.1 実験概要

構築したマウス版 Geneformer の評価するための実験を以下に 3 つ示し，実施する．

- マウス版 Geneformer と従来手法による細胞型の分類実験
- 事前学習の有無による細胞型の分類実験
- マウス版 Geneformer による in silico 摂動実験

マウス版 Geneformer と従来手法による細胞型の分類実験では，マウスの多様な臓器の単一細胞データを使用し，マウス版 Geneformer が従来手法よりも分類精度が向上するかを評価する．事前学習の有無による細胞型の分類実験では，マウス版 Geneformer の事前学習が意味のあるものなのかをマウスの胚，腎臓，尿道と前立腺が混ざった単一細胞データの細胞型を分類することにより評価をする．これらの評価には，Accuracy を用いて評価を行う．また，細胞分布を可視化することで，マウス版 Geneformer の性能の分析や，使用するデータの分析を行う．マウス版 Geneformer による in silico 摂動実験では，生物学実験である in vivo 実験で同定した遺伝子が in silico 摂動実験で検出することが可能かを評価する．in vivo 実験とは，人やマウスなどの生物の体内で，医薬品の効果や変異させた遺伝子の影響，生物学的な遺伝子発現のプロセスの調査，評価を行うための実験のことである．

4.2 in silico 摂動実験の概要

in silico 摂動実験とは，コンピュータ上で行う遺伝子の機能を解析する手法の 1 つである．遺伝子の欠失や活性化などの変異をシミュレーションすることで，その遺伝子が遺伝子ネットワークにどのような影響を与えるかを予測する．in silico 摂動実験では，はじめに遺伝子発現量データを解析し，各遺伝子の発現量を Rank Value Encoding を用いてランク付けを行う．次に，遺伝子をランダムに何回も摂動させる．遺伝子を欠失させる場合は，その遺伝子のランク値を下げ，他の遺伝子のランク値を上げる操作を行う．一方，遺伝子を活性化させる場合は，その遺伝子のランク値を上げ，他の遺伝子のランク値を下げる操作を行う．遺伝子の摂動により，摂動後の細胞状態が特定の細胞状態に近づいた時，摂動した遺伝子は重要な遺伝子であることが分かる．このとき，コサイン類似度を用いて摂動後の細胞状態が特定の細胞状態に近づいたかを数値的に評価する．

4.3 実験条件

マウス版 Geneformer の事前学習を mouse-Genecorpus-20M を用いて行う．事前学習に使用した条件を表 1 に示す．mouse-Genecorpus-20M を Geneformer に学習させるために，我々が使用した GPU は，32 GB の NVIDIA V100 を 8 枚使用し，約 2 日間学習を行った．

表 1　事前学習に使用したパラメータ

実験条件	パラメータ
入力最大次元数	2,048
Transformer Encoder の層数	6
attention head 数	4
埋め込み次元数	256
活性化関数	ReLU
Drop out 率	0.02
Epoch 数	20
最大学習率	1e-3
Warm up steps	10,000 step
スケジューラ	cosine スケジューラ
バッチサイズ	12
最適化手法	AdamW
Weight decay	1e-3

表 2　ファインチューニングに使用したパラメータ

実験条件	パラメータ
入力最大次元数	2,048
固定する層	0
Epoch 数	10
最大学習率	5e-5
Warm up steps	500 step
スケジューラ	cosine スケジューラ
バッチサイズ	12
最適化手法	AdamW
Weight decay	1e-3

マウス版 Geneformer と従来手法による細胞型の分類実験は，テキスト分類タスクを用いて細胞型の分類タスクへファインチューニングを行う．使用するデータは，マウスの舌，胸腺，乳腺，大腸，四肢の筋肉，脾臓，心臓，脳，腎臓の 9ヶ所の臓器の単一細胞データである．これらの臓器データを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件を表 2 に示す．

事前学習の有無による細胞型の分類実験は，テキスト分類タスクを用いて細胞型の分類タスクへファインチューニングを行う．使用するデータは，マウスの尿道と前立腺が混ざった単一細胞データ，胚，腎臓の単一細胞データである．胚は，動物や人間の発生学において受精卵のことを指し，様々な臓器の細胞型に変化する細胞が存在する．

このことから，胚の細胞は，細胞型の特徴を完全に保持していないと考えられる．尿道と前立腺が混ざったデータは，2 種類の臓器が混ざったデータである．これらから，マウス版 Geneformer と従来手法による細胞型の分類実験で使用する臓器データよりも，細胞型を分類するのが困難な臓器データと考えられる．これらの臓器データを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件は，表 2 の条件を使用する．

in silico 摂動実験は，テキスト分類タスクを用いて疾患型の分類タスクへファインチューニングを行う．使用するデータは，糖尿病性腎臓病と UMOD 腎臓病と正常な腎臓のデータ [24]，

表 3　マウス版 Geneformer と従来手法の分類精度の比較

臓器	細胞種数	マウス版 Geneformer	scDeepSort	scVAE
舌	3	93.08	76.69	80.44
胸腺	6	93.47	54.94	74.95
乳腺	7	98.16	47.76	74.25
大腸	7	90.91	49.78	59.00
四肢の筋肉	9	99.02	90.82	79.58
脾臓	10	97.83	81.01	76.47
心臓	11	95.55	79.55	79.42
脳	15	91.95	58.46	76.19
腎臓	18	92.82	58.01	56.25

タンパク質である COP1 をノックアウトしたデータと正常なデータ [25] である．前者のデータは，腎臓組織内の細胞型や疾患に特異的な遺伝子発現の変化を解析するためのデータである．後者のデータは，神経炎症を抑える役割のあるタンパク質 COP1 をノックアウトしたマイクログリアにおけるタンパク質 c/EBPβ を解析するためのデータである．これらのデータを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件は，表 2 の条件を使用し，疾患型の分類精度が 90.00 ポイント以上となったモデルを in silico 摂動実験に使用する．糖尿病性腎臓病の in silico 摂動実験では，正常な腎臓細胞を糖尿病性腎臓病の細胞に近づける実験を行う．UMOD 腎臓病の実験では，UMOD 腎臓病の細胞を正常な腎臓細胞に近づける実験を行う．COP1 ノックアウトの実験では，COP1 をノックアウトした細胞を正常な細胞に近づける実験を行う．

4.4　細胞型の分類結果

構築したマウス版 Geneformer と，従来手法である scDeepSort と scVAE の細胞型の分類精度の比較結果を表 3 に示す．表 3 から，マウス版 Geneformer は，従来手法である scDeepSort や scVAE よりも細胞型の分類精度が全てのマウスの臓器において大幅に向上していることが分かる．このことから，マウス版 Geneformer は，細胞型の分類タスクにおいて有効であると考える．そして，マウス版 Geneformer は，細胞型の種類数に依存することなく分類精度が 90 ポイントを超えている．従来手法では，細胞型の種類数ごとに分類精度が変動している．これらから，マウス版 Geneformer は，分類するクラス数に頑健であると言える．

また，scDeepSort や scVAE は，マウスの臓器ごとで細胞型の分類精度にばらつきがある．scDeepSort では最大 43.06 ポイントの分類精度差が存在し，scVAE では最大 24.19 ポイントの分類精度差が存在する．一方，マウス版 Geneformer は，マウスの臓器ごとで細胞型の分類精度のばらつきが抑えられ，最大分類精度差は 8.11 ポイントである．このことから，マウス版 Geneformer は，従来解析手法よりもマウスの臓器間の差に柔軟に対応することが可能であり，頑健であると言える．

さらに，舌と四肢の筋肉のデータにおいて細胞分布の可視化を行った．その結果を図 4 と図 5 に示す．舌の細胞分布の可視

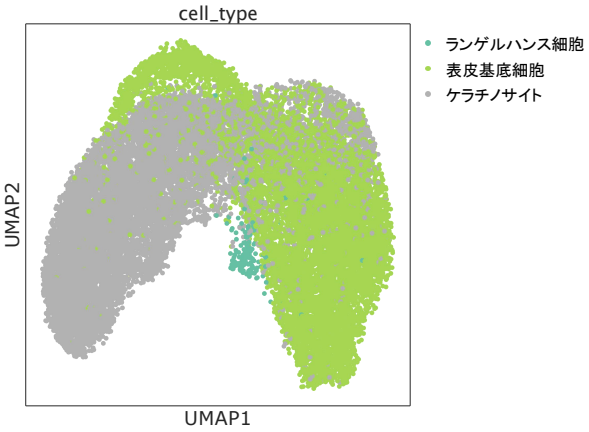


図 4　舌の細胞分布の可視化結果

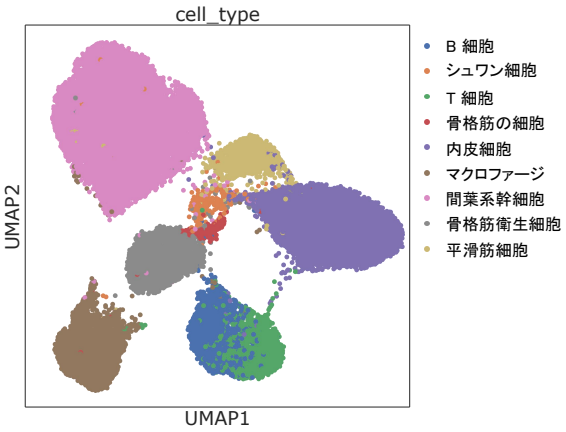


図 5　四肢の筋肉の細胞分布の可視化結果

化結果に注目すると，ケラチノサイトと表皮基地細胞が大まかに細胞型ごとに分かれていることが分かる．舌のデータの分類精度が 93.08 ポイントであることを考慮すると，細胞型を正確に予測している細胞はケラチノサイトと表皮基地細胞であると考える．一方，ランゲルハンス細胞が表皮基底細胞と被っている．これは，ランゲルハンス細胞と表皮基底細胞が非常に似た特徴量を持っていると考えられる．以上のことを踏まえるとマウス版 Geneformer は，細胞型の分布が重なるほど非常に似た特徴量を持つ細胞型の遺伝子ネットワークの予測は，困難であると考える．

次に，四肢の筋肉の細胞の可視化結果に注目する．こちらの細胞は，舌の細胞よりも細胞型に分けることができていることが分かる．また，シュワン細胞と骨格筋の細胞の分布が隣接し，B 細胞と T 細胞の分布が隣接し，細胞型の分類に困難と思われた．しかし，四肢の筋肉のデータのマウス版 Geneformer による分類精度が 99.02 ポイントであり，ほとんどの細胞を分類することが可能となっている．これらから，マウス版 Geneformer は，隣接している細胞型の特徴も詳細に捉える遺伝子ネットワークを予測することが可能であると考える．

4.5　事前学習の有無による細胞型の分類結果

事前学習ありのマウス版 Geneformer と事前学習なしのマウス版 Geneformer の細胞型の分類結果を表 4 に示す．表 4 から，事前学習したマウス版 Geneformer の分類精度は，事前学

表 4　事前学習の有無による分類精度の比較

臓器	細胞種数	事前学習あり	事前学習なし
尿道と前立腺	7	93.98	92.81
胚	9	77.22	69.78
腎臓	10	77.96	68.56

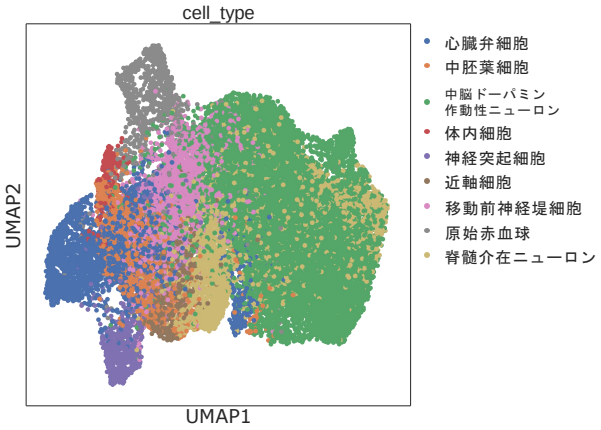


図 6　胚の細胞分布の可視化結果

習していないマウス版 Geneformer の分類精度よりもすべてのデータで高いことが分かる．最大の分類精度差は腎臓のデータで 9.40 ポイントで，最小の分類精度差は尿道と前立腺が混ざったデータで 1.17 ポイントである．このことから，事前学習モデルを使用することは，細胞型の分類を行うことが困難なデータになるにつれて効果的であると言える．よって，事前学習モデルの使用は，すべてのマウスの単一細胞データにおいて事前学習していないモデルよりも効率的に学習することが可能となり，分類精度の向上につながると考える．

さらに，胚のデータにおいて細胞分布の可視化を行った．その結果を図 6 に示す．胚の細胞分布の可視化結果から，四肢の筋肉の細胞分布のように細胞型がきれいに分布せず，全体的に 1ヶ所に集中したような分布となっていることが分かる．さらに，心臓弁細胞は 2ヶ所に分布し，中胚葉細胞や近軸細胞，脊髄介在ニューロンは，別の細胞型だが 1ヶ所に分布していることが分かる．これらから，胚のデータは細胞型の特徴を完全に保持していないと考えられる．また，胚のデータの事前学習ありの分類精度が 77.22 ポイントで事前学習なしの分類精度が 69.78 ポイントであることを考慮すると，マウス版 Geneformer は，事前学習を行うことで細胞型の特徴を完全に保持していないデータに対しても，事前学習で得た知識を使用して遺伝子ネットワークを予測することが可能であると考える．

これらから，マウス版 Geneformer は，事前学習を行うことで，多様なデータの遺伝子ネットワークを予測することが可能になると考える．

4.6　in silico 摂動実験の結果

4.6.1　in silico 摂動実験による糖尿病性腎臓病の結果

糖尿病性腎臓病の in silico 摂動実験の結果を示す．正常な腎臓細胞に存在する遺伝子 Slc12a3 を欠失させた時，糖尿病性腎臓病の細胞に最も近づくことが分かった．生物実験である in vivo 実験では，糖尿病性腎臓病にかかると顆粒細胞と黄斑細胞で構成する糸球体装置に変化があり，その顆粒細胞には遺伝子 Ren1 が存在し遺伝子 Slc12a3 が存在しないと同定している [24]．このことから，糖尿病性腎臓病における in silico 摂動実験は成功し，糖尿病性腎臓病で変化した遺伝子を特定したと考える．

4.6.2　in silico 摂動実験による UMOD 腎臓病の結果

UMOD 腎臓病の in silico 摂動実験の結果を示す．UMOD 腎臓病の細胞に存在する遺伝子 Slc35b1 を欠失させた時，正常な腎臓細胞に最も近づくことが分かった．in vivo 実験では，腎臓における変異型 UMOD タンパク質の蓄積により，Unfolded Protein Response (UPR) が変化し，UPR を活性化させる遺伝子 Slc35b1 と Slc3a2 の遺伝子の発現量が変化することが分かっている [24]．このことから，UMOD 腎臓病における in silico 摂動実験では，2 つの遺伝子のうち 1 つを特定したと考える．

4.6.3　in silico 摂動実験による COP1 ノックアウトの結果

COP1 をノックアウトしたミクログリアの in silico 摂動実験の結果を示す．COP1 をノックアウトした細胞に存在する遺伝子 Apoe を欠失させた時，正常な細胞に最も近づくことが分かった．遺伝子 Apoe を欠失させたときのコサイン類似度は 0.36 であった．また，遺伝子 Cebpb，Clec7a，Cst7 の遺伝子の欠失も正常な細胞に近づくことが分かった．in vivo 実験では，COP1 をノックアウトすることにより c/EBPβ が脳内に溜まることにより，神経炎症が起こることが言われている．また，COP1 のノックアウトにより遺伝子 Apoe と Cebpb の発現量が増加し，神経変性関連遺伝子 Clec7a，Cst7 の発現量も変化することが分かっている [25]．遺伝子 Cebpb はタンパク質 c/EBPβ を発現する遺伝子であり，遺伝子 Apoe は神経変性疾患であるアルツハイマー型認知症の危険因子と言われている遺伝子である．これらから，in silico 摂動実験は成功し，in vivo 実験を再現したと考える．

4.6.4　細胞分布の可視化結果

in silico 摂動実験で使用したデータの細胞分布の可視化結果を図 7，図 8 に示す．これらを見ると，正常な細胞と異常な細胞が別れて分類していることが分かる．特に，図 8 は正常な細胞である Cop1 WT と COP1 をノックアウトした細胞である Cop1 KO が完全に分離している．これらから，マウス版 Geneformer は，ファインチューニングしたことにより，疾患データ特有の特徴を詳細に捉えた遺伝子ネットワークを予測することが可能であると考える．

5　おわりに

本稿では，大規模なマウスの単一細胞データを用いて事前学習した，Attention ベースの深層学習モデルであるマウス版 Geneformer を構築した．マウス版 Geneformer と従来手法による細胞型の分類実験では，マウスの 9 つの臓器データを使用し，従来手法である scDeepSort，scVAE よりもマウス版 Geneformer の分類精度が向上していることを確認した．また，

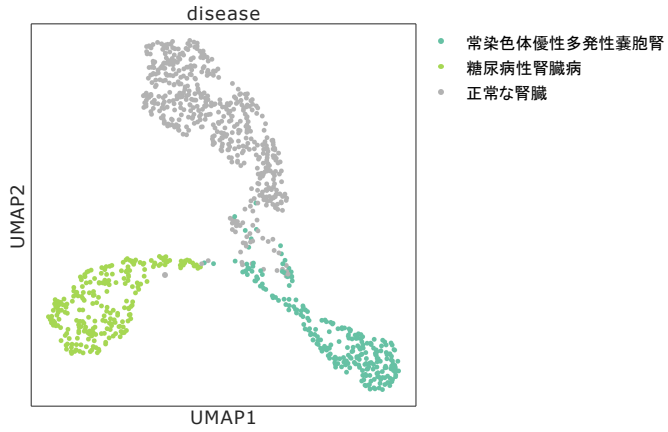


図 7　腎臓データの疾患分布の可視化結果

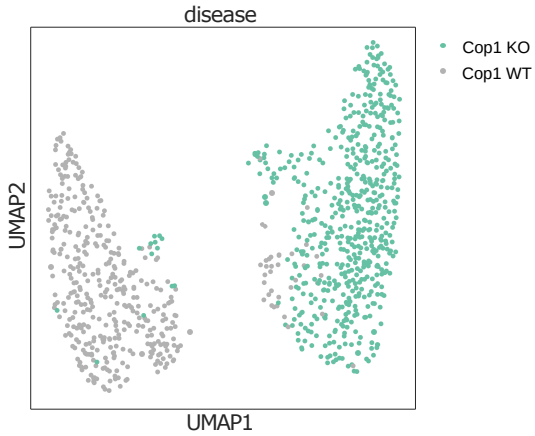


図 8　COP1 データの疾患分布の可視化結果

マウスの 9 つの臓器データで分類精度が一貫して 90 ポイントを超えていることから，マウス版 Geneformer は，マウスの臓器に依存しないことと，細胞型の種類数に依存しないことを確認した．これらから，マウス版 Geneformer の有効性と頑健性を示した．

また，マウス版 Geneformer の事前学習の有無による細胞型の分類実験では，2 つの臓器が混ざったデータと胚のデータ，腎臓のデータを使用し，事前学習していないマウス版 Geneformer よりも事前学習したマウス版 Geneformer の方が分類精度が向上することを確認した．また，胚の細胞分布の可視化結果から，事前学習を行うことで細胞型の特徴を完全に保持していないデータに対しても，事前学習で得た知識を使用して遺伝子ネットワークを予測することが可能であることを確認した．これらから，マウス版 Geneformer の事前学習の有効性を示した．

最後に，マウス版 Geneformer を使用した in silico 摂動実験では，糖尿病性腎臓病モデルと UMOD 腎臓病モデルと COP1 ノックアウトモデルを使用し，in vivo 実験を再現した．このことから，マウス版 Geneformer による in silico 摂動実験は，in vivo 実験と同等の結果を得られることを示した．また，疾患モデルの細胞分布の可視化結果から，正常な細胞と異常な細胞で別れて分布していることを確認した．これらから，マウス版 Geneformer による in silico 摂動実験の有効性を示した．

今後の展望として，Single-cell RNA-sequence データに特化したプレテキストタスクの開発や，本稿で提案したモデルを実際のマウスの疾患モデルに適用することが挙げられる．

文　献

[1] Saliba, Antoine-Emmanuel, et al. ”Single-cell RNA-seq: advances and future challenges.” Nucleic acids research 42.14 (2014): 8845-8860.

[2] Lepetit, Maxime et al. “scAN1.0: A reproducible and standardized pipeline for processing 10X single cell RNAseq data.” In silico biology vol. 15,1-2 (2023): 11-21. doi:10.3233/ISB-220252

[3] Gribov, Alexander, et al. ”SEURAT: visual analytics for the integrated analysis of microarray data.” BMC medical genomics 3 (2010): 1-6.

[4] Jiang, Peng, et al. ”MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.” Nucleic acids research 35.suppl_2 (2007): W339-W344.

[5] Theodoris, Christina V., et al. ”Transfer learning enables predictions in network biology.” Nature (2023): 1-9.

[6] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[7] Lieberman, Yuval, Lior Rokach, and Tal Shay. ”CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments.” PloS one 13.10 (2018): e0205499.

[8] Chen, Tianqi, and Carlos Guestrin. ”Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[9] LeCun, Yann, et al. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[10] Jia, Shangru, et al. ”scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning.” Briefings in Bioinformatics 24.5 (2023): bbad266.

[11] Van der Maaten, Laurens, and Geoffrey Hinton. ”Visualizing data using t-SNE.” Journal of machine learning research 9.11 (2008).

[12] McInnes, Leland, John Healy, and James Melville. ”Umap: Uniform manifold approximation and projection for dimension reduction.” arXiv preprint arXiv:1802.03426 (2018).

[13] Tan, Mingxing, and Quoc Le. ”Efficientnet: Rethinking model scaling for convolutional neural networks.” International conference on machine learning. PMLR, 2019.

[14] Sharma, Alok, et al. ”DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture.” Scientific reports 9.1 (2019): 11399.

[15] Kipf, Thomas N., and Max Welling. ”Semi-supervised classification with graph convolutional networks.” arXiv preprint arXiv:1609.02907 (2016).

[16] Shao, Xin, et al. ”scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network.” Nucleic acids research 49.21 (2021): e122-e122.

[17] Hamilton, Will, Zhitao Ying, and Jure Leskovec. ”Inductive representation learning on large graphs.” Advances in neural information processing systems 30 (2017).

[18] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. ”Reducing the dimensionality of data with neural networks.” science 313.5786 (2006): 504-507.

[19] Kingma, Diederik P., and Max Welling. ”Auto-encoding variational bayes.” arXiv preprint arXiv:1312.6114 (2013).

[20] Dilokthanakul, Nat, et al. ”Deep unsupervised clustering with gaussian mixture variational autoencoders.” arXiv preprint arXiv:1611.02648 (2016).

[21] Grønbech, Christopher Heje, et al. "scVAE: variational auto-encoders for single-cell gene expression data." Bioinformatics 36.16 (2020): 4415-4422.
[22] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[23] Chaffin, Mark, et al. "Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy." Nature 608.7921 (2022): 174-180.
[24] Marshall, Jamie L., et al. "High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways." Iscience 25.4 (2022).
[25] Ndoja, Ada, et al. "Ubiquitin ligase COP1 suppresses neuroinflammation by degrading c/EBP β in microglia." Cell 182.5 (2020): 1156-1169.

ペルソナ・モデリングによるソーシャルネットワークの将来予測の精度向上

松本　寛智† 　伊藤　寛祥†† 　三宅健太郎† 　森嶋　厚行††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s2011456@s.tsukuba.ac.jp, ††ito@slis.tsukuba.ac.jp, †††s2113591@klis.tsukuba.ac.jp,
††††morishima-office@ml.cc.tsukuba.ac.jp

あらまし

ソーシャルネットワークにおける将来予測の重要度は高まっている。これまでのソーシャルネットワークの将来予測の研究では、予測に使うデータの特徴量 (ノードの属性値やエッジ) の選択に焦点を当てた研究や、予測結果の説明可能性に焦点を当てた研究などが行なわれている。しかし、実際にグラフを用いてソーシャルネットワークの将来予測をする場合に十分なデータは存在しないことが多い。また、計算コストの面から全てのデータを利用できない場合もある。本研究では、ソーシャルネットワークの将来予測に “ペルソナ” を考慮することによりデータが不十分な時にある程度の水準の推定が可能な手法かどうかを明らかにする。本研究では、ペルソナをソーシャルネットワーク上での人間の行動パターンと定義する。属性値やエッジの接続情報が不足しているノードに対してペルソナを導入することで、ノードの行動をペルソナの行動方針で補うことができるため、精度が改善すると考えられる。本研究では、「ペルソナ」を組み込んだモデルを強化学習で学習する新しいアルゴリズムを提案する。提案手法により、現実世界のソーシャルネットワークにおけるネットワークの変化の予測において、既存手法と同等以上の精度を実現した。

キーワード　ソーシャルネットワーク予測, マルチエージェントシステム, 強化学習

1 はじめに

近年、日本におけるソーシャルメディアの利用者数は年々増加している。また、ヒット予測ランキングの結果からグッズの生産やイベント企画やソーシャルメディアでのトレンド予測など実世界で予測に基づいた計画が幅広く行われている。その結果、ソーシャルネットワークでの将来予測がより重要視されている。

ソーシャルネットワークとは人と人が関係を元に接続されているデータのことである。ソーシャルネットワークは、ノード (節点・頂点) 集合とノード間のつながりを表すエッジ (枝・辺) 集合からなるグラフ構造のデータで表現できる。例えば、ソーシャルメディアは人をノードで表し、ある人とある人の間の繋がりをエッジで表すことができる。グラフ構造のデータを入力としてソーシャルネットワークの将来予測を行う手法 [19] [21] [20] が提案されている。将来予測は、ノードの属性値の予測やエッジの予測がある。予測は、静的なネットワークを対象にしたものと動的なネットワークを対象にしたものがある。静的なネットワークは、グラフ構造の時系列的変化がないものと仮定して観測済みのグラフデータから未観測のグラフデータを予測する。一方、動的なネットワークは、グラフ構造は時間の経過と共に変化する可能性があるため観測済みのグラフデータから将来のグラフデータを予測する。本研究では、動的なネットワークを対象として研究を行う。

動的なグラフデータの将来予測に関しては、GNN などのネットワーク表現学習やマルチエージェント強化学習を使ったものなどがある。予測に使うデータの特徴量 (ノードの属性値やエッジ) の選択や更新方法に焦点を当てた研究 [19] [20] や、予測結果の説明可能性に焦点を当てた研究 [21] などが行なわれている。 しかし、実際にグラフを用いてソーシャルネットワークの将来予測をする場合に十分なデータは存在しないことが多い [3]。また、既存の提案手法では計算コストの面から全てのデータを利用できない場合もある [17] [18]。そのため、それらの場合に対応できるモデルが必要となっている。

人間の行動をいくつかの行動パターンに分ける研究がなされている。例えば、近年のソーシャルネットワークの発展にともなってソーシャルネットワークを対象にした人間の行動パターンを分析する研究 [2] [7] も行われている。Brandtzaeg&Heim らの研究 [2] では、ソーシャルネットワークの一つである複数のソーシャルメディアを対象に利用者を 5 つの類型で表せることを明らかにした。また、[7] は Instagram での利用者の投稿内容を分析し 5 つのユーザータイプがある事を発見した。このことから、ソーシャルメディアでの人々の行動はパターン化して考えることがてきると言える。

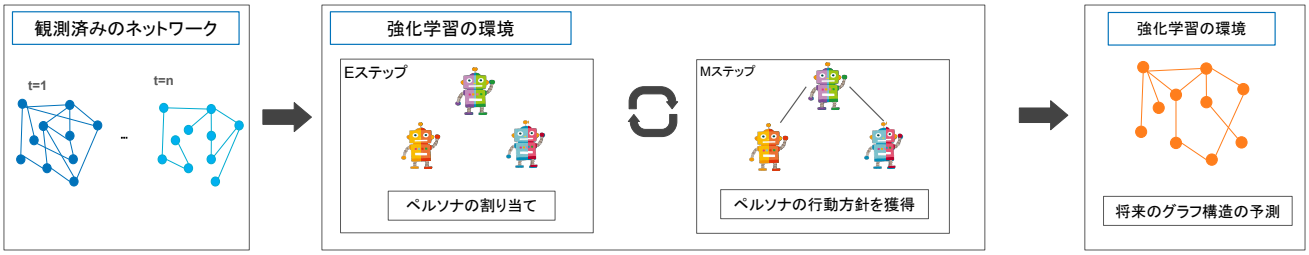


図 1: **提案手法の概要** 提案手法では、各ノードがもつ報酬関数と方策関数を定義し、観測済のネットワーク系列から報酬関数のパラメーターを学習する。そして、学習済の報酬関数を用いて強化学習の環境でペルソナの方策関数を学習する。学習済みの関数と強化学習の環境でマルチエージェントシュミレーションを行うことで未来のネットワーク系列を生成する。

そこで、本研究ではソーシャルネットワークでの人間行動パターンはいくつかに分類することができると仮定し、ペルソナの概念を強化学習の報酬関数・方策関数を使って表現するソーシャルネットワークの将来予測のための新しい手法を提案する。提案手法では、強化学習のアルゴリズムを使って探索・学習を行うことでペルソナの行動方策を獲得する。そして、学習したペルソナを将来予測に活用する。この方法を用いることで、例えば、あるノードのデータが不足している場合に、ペルソナの情報を使うことで行動方針を補うことで将来予測が可能になると予想される。本論文のリサーチクエスチョンは次である。

- (RQ1) ペルソナの導入によりソーシャルネットワークの将来予測の精度は向上するか
- (RQ2) どのようなデータの欠損であればペルソナの導入が効果的か

これらを検証するために，観測済みのグラフデータを入力として EM アルゴリズムとペルソナを導入した強化学習を融合した「ペルソナ・モデリング」を構築する。具体的には、次のような手法である。(1) まず、ペルソナの方策関数を利用して全てのノードに対してペルソナの割り当てを計算する。これが EM アルゴリズムの E-step に該当する。(2) 次に、MARL の強化学習手法の一つである COMA を用いてペルソナの行動方策のパラメータを求める。これが、EM アルゴリズムの M-step に該当する。(3) この (1)(2) を繰り返し行いパラメータの最適化を行う。

実験では研究者の共著者ネットワーク（NIPS、DBLP）データセットを対象にし、ソーシャルネットワークの将来予測における手法の有用性に関して，将来予測性能とデータ欠損の影響に関する検証を行なった。実験の結果、次のことがわかった.

- (1) ペルソナの数を調整することで既存手法と比較して高精度にソーシャルネットワークの将来予測することが可能である。
- (2) データ欠損に関しては，各ペルソナに割り当てられたデータが多い場合にはペルソナの導入が効果的である。一方、各ペルソナに割り当てられたデータが少なすぎる場合にはペルソナの学習が不十分となるため将来予測に悪影響を及ぼす。

本論文の構成は次の通りである。第 2 節では、関連研究についての紹介と提案手法との違いを説明する。第 3 節では、本論文でのペルソナの定義と本論文で使う記号について説明を行う。第 4 節では、本研究の先行研究について紹介と本研究との関連性について説明する。第 5 節ではペルソナと報酬関数・方策関数の関係について説明する。第 6 節では、実験の説明を行う。

2 関 連 研 究

動的に変化するソーシャルネットワークに対する予測手法は様々なアプローチで提案されてきた。本節ではそれら関連研究を示すとともに、本研究との差異および本研究の位置づけに関して述べる。

2.1 ネットワーク表現学習

表現学習とは、画像、音、自然言語、時系列データなどの要素を予測問題を解くことで分散表現（ベクトル）として抽象化する手法である。ネットワーク表現学習とは、ネットワーク（グラフ）内のノード（頂点）やエッジ（辺）を低次元の連続ベクトル空間に埋め込む手法である。

ネットワーク表現学習によって、ネットワーク内の複雑な関係性や構造を処理しやすい形式に変換することが可能になる。これにより、リンク予測、ノード分類、クラスタリング、推薦システムなど、様々なタスクに応用することが出来るようになる。ネットワーク表現学習の手法として、ランダムウォークベースの手法がある。例えば、word2vec を用いてネットワーク構造からノードの表現学習を獲得する手法の DeepWalk [11]、DeepWalk の改良でランダムウォークの探索戦略を工夫した手法の node2vec [5]、DeepWalk の改良でネットワークの近接性を考慮し大域的な構造にも着目した手法の LINE [12] などが提案されている。

近年では、ノードの変化と属性値の変化の両方を時系列的に埋め込む手法が提案されている。Twitter におけるユーザーの関心の変化を追跡する手法として [9]、テキスト情報を持つ動的なネットワークにおけるノードと単語の埋め込み手法 [1] がある。

また、ネットワーク表現学習の手法としてグラフニューラルネットワークを用いた表現学習手法が数多く提案されている。グラフ構造データに対する畳み込みニューラルネットワークの一種で、ノードの特徴とその隣接ノードの特徴を集約することで、ノードの新しい特徴表現を生成する手法 [8] がある。アテンションを利用して、隣接ノードからの特徴の集約方法を動的に決定することにより、ノード間の異なる重要性を考慮するこ

とができる手法 [13] がある。グラフに対する敵対生成ネットワークを用いた手法 [14] がある。動的グラフ構造に対応するために開発された手法 [15] など様々な方法が存在する。これらの研究は、観測済みのネットワークグラフに対するグラフやノードの特徴量を獲得するための手法で、本研究の目的である未観測の時刻におけるネットワークの予測についての検討は行なわれていない。

2.2 マルチエージェント強化学習

強化学習とは、エージェントが環境に対して行動を行い、その行動に対して環境から得られた報酬をもとに方策を改善する手法である。強化学習は、ゲームやロボット制御などのタスクで使われている。エージェントが単一の場合はシングルエージェント強化学習、エージェントが複数存在する場合はマルチエージェント強化学習と呼ばれる。実世界の問題に強化学習を適応する際に、エージェントが複数存在する設定が多くあるためマルチエージェント強化学習の重要さは年々増している。マルチエージェント強化学習（MARL）の問題設定は、学習と実行の方法に基づいて、主に 3 つのカテゴリーに分けられる。

（1） **集中学習集中実行（CTCE: Centralized Training and Centralized Execution）**：この設定では、複数のエージェントが環境内に存在するが、全てのエージェントは共通の方策関数に従って行動を選択する。学習は、全てのエージェントの観測情報を利用して行う。実行は、学習と同じ方法で行われる。一つの方策関数に従うため、結果が方策関数に左右されやすい。

（2） **分散学習分散実行（DTDE: Decentralized Training and Decentralized Execution）**：この設定では、環境内の各エージェントが個別の方策に基づいて行動を選択する。学習は、それぞれが自分自身の情報のみを利用して行う。実行は、学習と同じ方法で行われる。この方法は、基本的にシングルエージェント強化学習をマルチエージェント環境に適応させたものとなる。そのため、環境が非定常に変化するため困難な設定となりやすい。

（3） **集中学習分散実行（CTDE: Centralized Training and Decentralized Execution）**：この設定では、学習時には全エージェントの情報を含む環境情報を使用する。実行時（テスト時）には、各エージェントは自分自身の部分的な観測情報のみを利用する。

前述した 3 つの手法の中で、集中学習分散実行がマルチエージェント強化学習の手法の主流の考えであり COMA [4] や MADDPG [10] などで使われている。集中学習分散実行の強化学習のアルゴリズムを用いネットワークの将来予測をする手法として [6] などがある。この研究では、ある状況におけるエージェントの最適な行動やそれに伴うネットワーク構造の出現を予測する研究は行なわれているが、本研究が目指す観測済みネットワークから未観測の時系列ネットワークの生成を行なうタスクは行なわれていない。

3 問題設定

3.1 ペルソナ

ペルソナとは、人間の行動パターンである。ユーザプロフィリングという言葉があるように人間の行動をいくつかの行動パターンに分ける研究がなされている。例えば、推薦の分野では顧客の類似性をプロファイリング分析した書籍推薦システム [22] などがある。また、近年のソーシャルネットワークの発展にともなってソーシャルネットワークを対象にした人間の行動パターンを分析する研究 [2] [7] も行われている。Brandtzaeg&Heim らの研究 [2] では、ソーシャルネットワークの一つである複数のソーシャルメディアを対象に利用者を 5 つの類型で表せることを明らかにした。また、Hu et al. らの研究 [7] では Instagram での利用者の投稿内容を分析し 5 つのユーザータイプがある事を発見した。これらのことから、ソーシャルメディアでの人々の行動はパターン化して考えることが出来る。

そこで、本研究ではソーシャルネットワークでの人間の行動パターンを “ペルソナ” と定義し，ソーシャルネットワークの予測に活用する。ペルソナの例としては、インフルエンサーのペルソナやインフルエンサーを応援する人のペルソナといったものが考えられる。

本研究では、ソーシャルネットワークでの人間行動パターンはいくつかのペルソナに分類することができると仮定し、ペルソナを強化学習の報酬関数・方策関数を使って表現し、強化学習の学習アルゴリズムに基づく方法を用いて学習することでペルソナの方策関数を学習する。学習したペルソナを将来予測に活用することで、あるノードのデータが不足していて行動方針が分からない場合に、ペルソナの情報を使うことで行動方針を補うことで将来予測が可能になると期待される。ペルソナと報酬関数・方策関数の関係については、5 節の提案手法で説明する。また、ペルソナの効果については、6 節の実験で説明する。

3.2 記号・定義

本節では本研究で対象とする問題の定義について説明する。本研究では、入力となるソーシャルネットワークを属性値を持つグラフ構造として表現し、各ノード n は各人を表し、属性値 $\mathbf{x}_n$ は単語など各人の興味関心を表すベクトル、エッジは人と人との繋がりを表すものとする。すなわち、入力として離散的な時刻 $t = 1 \sim T$ におけるソーシャルネットワーク $G^{(t)} = (V^{(t)}, E^{(t)}, X^{(t)})$ と、各ノード $n \in V$ が与えられるとする。ここで $V^{(t)}$ はノード集合、$E^{(t)}$ はノード間のエッジ集合、$X^{(t)}$ は各ノードがもつ属性値ベクトルの集合とする。時刻 t に u と v の間にエッジが存在することを $e^{(t)}_{u,v} \in E^{(t)}$ と表現する。r_{ik} はあるエージェント i のペルソナ k に対しての割り当て率である。

本研究では新規ノードの出現や消滅は考慮せず、既存のノードの間に存在するエッジの存在確率の算出と、それに基づいた将来のネットワークを予測を行う。本研究では、新規ノード・単語の出現および消滅は考慮せず、時刻 $t' > T$ におけるソーシャルネットワーク中の既存のノード i, j 間に存在するエッジ

の存在確率 $P(e_{i,j}^{(t')} \in E^{(t')})$ を求めることが目的である。

表 1: 主な記号と定義

$G^{(t)} = (V^{(t)}, E^{(t)}), X^{(t)}$	時刻 t におけるグラフ
$V^{(t)} = \{i\}$	時刻 t におけるノード集合
$E^{(t)} = \{e_{i \to j}^{(t)}\}$	時刻 t におけるエッジ集合
$X^{(t)} = \{x\}$	時刻 t における属性値集合
r_{ik}	ペルソナの割り当て率
$k = \{R_k, \pi_k\}$	ペルソナ k
$n_i = \{n_i \mid n_i \in V^{(t)}\}$	ノード集合における i 番目のノード
$R^{(t)} = \{R^{(t)} \mid R^{(t)} \in \mathbb{R}\}$	時刻 t における報酬
$V_{i\text{neighbor}} \in V^{(t)}$	ノード i の隣接ノードの集合
$e_{i \to j}^{(t)} = (i, j)$	i から j への無向エッジ

4 Reinforce モデルについての前提知識

マルチエージェント強化学習を用いてソーシャルネットワークの将来予測をを行った先行研究 [21](以下 “Reinforce モデル” という。) がある。この研究では、ノードを強化学習でのエージェントと仮定しエージェントの報酬関数・方策関数にパラメータを導入している。また、観測済みソーシャルネットワークで各時刻で各ノードが最適な行動を選択していると仮定し、未来の報酬を最大化させる方策を獲得することで未観測の時系列のネットワークを予測することを可能にしている。先行研究は、(1) 観測済みのネットワークを用いて報酬関数・方策関数のパラメータの最適化 (2) 最適化された報酬関数・方策関数を用いて強化学習の環境下で将来の時刻のネットワークの予測、という 2 つの手順からなる。以下の項ではこれらの流れについて以下の順に詳説する。

（1） 観測済みネットワークにおける報酬関数の最適化手法
（2） 未観測時間における方策関数の最適化
（3） 未観測時間における時系列ネットワークの生成手法

4.1 観測済みネットワークにおける報酬関数の最適化手法

まず、報酬関数の定義について説明する。先行研究では、観測済みのネットワークを用いて各エージェントについて報酬関数を以下のように定義する。

$$R_i(G(t)|\psi_i) = \sum_{n_j \in N(n_i)} \alpha_i \mathrm{sim}(n_i, n_j) - \beta_i \mathrm{cost}(n_i, n_j) \tag{1}$$

- α: 類似するノードと接続する場合の報酬の大きさを表している。

このパラメータが大きい場合は、自分と同じ特徴を持つノードとエッジを張りやすくなる。

- β: 他のノードとの接続コストの大きさを表している。

このパラメータが大きい場合は、他のノードとエッジを張にくくなる。

$\psi_i = \{\alpha_i, \beta_i\}$ はノード i のパラメータ集合、$N(n_i)$ はノード n_i に隣接するノード集合を表す。$\mathrm{sim}(n_i, n_j)$ はノード間の類似度を示し、先行研究では以下のようなノードの属性値集合のコサイン類似度によって算出する。

$$sim(n_i, n_j) = \frac{dot(\mathbf{x}_i, \mathbf{x}_j)}{|\mathbf{x}_i||\mathbf{x}_j|} \tag{2}$$

$cost$ はノード i、j がソーシャルネットワーク内で接続していることのコストを示し、以下のように定義する。

$$cost(n_i, n_j) = \begin{cases} 1 & n_i, n_j \text{ にエッジが存在する} \\ 0 & n_i, n_j \text{ にエッジが存在しない} \end{cases} \tag{3}$$

次に、報酬関数の最適化について説明する。前述した、各エージェントについての報酬関数を使って、時刻 t のソーシャルネットワーク $G^{(t)}$ 全体における報酬関数の総和を以下のように定義する。

$$R(G^{(t)} \mid \Psi) = \sum_{n_i \in V} r_i(G^{(t)}|\psi_i) \tag{4}$$

全ての時刻でのソーシャルネットワーク報酬関数の総和をとり、その報酬が最大になるようなパラメーターを SGD で推定する。

$$\Psi^* \leftarrow argmax_\Psi \sum_{t=1}^{T} R(G^{(t)} \mid \Psi) \tag{5}$$

4.2 方策関数の最適化

まず、方策関数の定義について説明する。先行研究では、各ノードの属性値に基づく類似度関数 $sim(n_i, n_j)$ を用い、これを用いて、i 番目のノードについての方策関数 $\pi(e_{i,j}^{(t)}|G^{(t)}, \theta_i)$ は以下の式で定義する。

$$\pi^+(e_{i,j}^{(t)}|G^{(t)}, \theta_i) = \sigma\left\{\epsilon_i \exp\left(\frac{sim(n_i, n_j)}{\tau_i}\right)\right\} \tag{6}$$

$$\pi^-(e_{i,j}^{(t)}|G^{(t)}, \theta_i) = \sigma\left\{\epsilon_i \exp\left(\frac{1 - sim(n_i, n_j)}{\tau_i}\right)\right\} \tag{7}$$

- ϵ: エッジの生成・削除の頻度の大きさを表す。

このパラメータの値が高いほど、報酬を増やすことを目的として、エッジの作成や削除が頻繁に行われる。

- τ: エッジの生成・削除の乱雑さを表す。

このパラメータの値が小さい時、属性値の類似性に基づいてエッジの生成・削除をしやすくなる。

ここで、$\pi^+(\cdot)$ は新規エッジの生成する確率、$\pi^-(\cdot)$ は既存エッジを消滅させる確率を表現する。σ はシグモイド関数、$\theta_i = \{\epsilon_i, \tau_i\}$ はノード n_i のパラメータ集合である。

次に、方策関数の最適化について説明する。まず、観測済みソーシャルネットワークの最も新しい時刻でのソーシャルネットワークを $G^{(t)}$ とする。ノード n_i について、ランダム選択と推移性の考えに基づいてエッジの生成・削除を行なう対象を探索する。このとき、ノード n_i は自身の方策関数を用いて、各ノードに対するエッジ生成・削除の確率に基づき $G^{(t+1)}$ を生成し、報酬関数によって報酬を計算する。各ノードの方策関数に基づいて未観測の時刻のソーシャルネットワークを生成した際、「ソーシャルネットワークの状態、 ソーシャルネットワークの差分、報酬」からなる以下のような時系列データ τ を獲得したとする。

$$\tau =< \mathcal{G}^{(1)}, \Delta\mathcal{E}^{(1)}, R(\mathcal{G}^{(1)} \mid \Psi), \mathcal{G}^{(2)}, \Delta\mathcal{E}^{(2)}, R(\mathcal{G}^{(2)} \mid \Psi), ..., \mathcal{G}^{(T)}, \Delta\mathcal{E}^{(T)}, R(\mathcal{G}^{(T)} \mid \Psi) > \quad (8)$$

このとき、ネットワーク中のすべてのノードが未観測の時刻において獲得した累積を割引率 γ を用いて以下のように定義する。

$$R(\tau) = R(\mathcal{G}^{(1)} \mid \Psi) + \gamma R(\mathcal{G}^{(2)} \mid \Psi) + \gamma^2 R(\mathcal{G}^{(3)} \mid \Psi) + \cdots + \gamma^T R(\mathcal{G}^{(T)} \mid \Psi) \quad (9)$$

このとき、最適化された報酬関数の下で、自身の報酬を最大にするような方策関数のパラメータ学習を勾配降下法の 1 つである REINFORCE [16] によって学習する。目的関数 $J(\theta)$ は以下の式で表現する。

$$J(\Theta) = \mathbb{E}_{\tau \sim \pi(\Delta\mathcal{E}|\Theta)}[R(\tau)] \quad (10)$$

ここで目的関数におけるパラメーター θ に関する勾配を求める。

$$\nabla J(\Theta) = \mathbb{E}\ [R(\tau)\nabla \log(\pi(\Delta\mathcal{E} \mid \Theta)] \approx \frac{1}{N}\sum_{i=1}^{N} R(\tau^{(i)})\nabla \log(\pi(\Delta\mathcal{E} \mid \Theta) \quad (11)$$

$\tau^{(i)}$ は、方策関数より生成された i 番目の時系列データである。

これを用いてパラメーターを更新する。ここで α は学習率である。

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) \quad (12)$$

4.3 未観測時間における時系列ネットワークの生成手法

学習データから求めた報酬関数と方策関数のパラメーターを用いて以下の方法で将来のソーシャルネットワークを予測する。(1) エッジの生成予測 (2) エッジ生成後のノードの特徴量の更新 (3)(1)、(2) を交互に行なうことで複数ステップ先のソーシャルネットワークにおけるエッジとノードの状態を予測

4.4 先行研究と本研究の関わり

本研究では、ソーシャルネットワーク上での行動はいくつかのペルソナに割り当てることができると仮定する。ペルソナを先行研究の報酬関数・方策関数のパラメータを使って表現するペルソナの方策関数・報酬関数を用いる新しい強化学習の手法を提案する。具体的には、報酬関数・方策関数のパラメータを代表値とするペルソナをいくつか設定する。仮定に基づいて各エージェントの報酬関数、方策関数をいくつかのペルソナの報酬関数、方策関数を使って表現する。この報酬関数と方策関数を学習し予測に活用することで将来のネットワークの状態を予測する。より詳細については、5 節で説明する。

5 提 案 手 法

本節では提案手法の概要を述べる。提案手法では、ペルソナの特徴を表すパラメータを持つ報酬関数と方策関数を設定する。まず、観測済みのデータで報酬が最大になるように報酬関数のパラメータを最適化する。次に、方策関数のパラメータとペルソナの割り当て率の初期値を設定し強化学習の環境で行動を実施する。その後、その行動によって獲得した報酬をもとに価値関数と方策関数を更新する。最後に、更新した方策関数を用いて各ノードのペルソナの割り当て率を更新する。この、強化学習の環境での行動・価値関数・方策関数・ペルソナの割り当て率の更新繰り返し行うことでペルソナの最適な方策関数と各ノードのペルソナの割り当て率を学習をする。この計算結果をもとに、将来のソーシャルネットワークを予測する。

以降の節では以下の順に詳説する。

（1） 強化学習の環境とペルソナ
（2） 報酬関数・方策関数
（3） 方策関数・属性値の更新
（4） ペルソナの割り当て率の初期値と更新
（5） 将来予測

5.1 強化学習の環境とペルソナ

本研究では、時刻 t における状態をソーシャルネットワーク $G^{(t)}$ と表す。ペルソナ k でのパラメータ集合 ψ_k をもつ報酬関数 $R(G^{(t)} \mid \psi_k)$、ペルソナ k でのパラメータ集合パラメータ集合 θ_k をもつ方策関数 $\pi(e_{i,j}^{(t)} \mid G^{(t)}, \theta_k)$ とし、ペルソナ k を $k = \{R(G^{(t)} \mid \psi_k), \pi(e_{i,j}^{(t)} \mid G^{(t)}, \theta_k)\}$ と定義する。ここで、報酬関数は各ノードのソーシャルネットワークの好ましさを表現し、方策関数は現在のソーシャルネットワークの状態における各ノードのエッジと属性の変化の傾向を表現する。本研究では強化学習に基づき、$t = 1 \sim T$ のソーシャルネットワークの系列 $< G^{(1)}, G^{(2)}, ..., G^{(T)} >$ から、報酬関数のパラメータ $\Psi = \{\psi_k\}_{k \in M}$ と、行動方策のパラメータ $\Theta = \{\theta_i\}_{k \in M}$ を学習する。また、EM アルゴリズムを用い各エージェントのペルソナの割り当て率 r_{ik} を学習する。強化学習のモデルとしては COMA [4] を使用する。Actor には方策関数をモデリングしたものを使用し、Critic には、GCN を利用する。

5.2 報酬関数・方策関数

先行研究 [21] とは異なり、本研究では各エージェントについてに報酬関数・方策関数を定義するのではなくペルソナに対して報酬関数・方策関数を定義する。本研究では、ペルソナ k での報酬関数と方策関数を以下のように定義する。

$$R_{ik}(G^{(t)}|\psi_k, n_i) = \sum_{n_j \in N(n_i)} \alpha_k sim(n_i, n_j) - \beta_k cost(n_i, n_j) \quad (13)$$

$$\pi_{ik}(\Delta G^{(t)}|G^{(t)}, \theta_k) = \sigma\left\{\varepsilon_k exp\left(\frac{sim(n_i, n_j)}{\tau_k}\right)\right\} \quad (14)$$

ここで、報酬関数のパラメータ $\Psi = \{\psi_k\}_{k \in M}$ はペルソナ k でのパラメータ集合を表す。同様に、行動方策のパラメータ $\Theta = \{\theta_k^\pi\}_{n_k \in M}$ も、ペルソナ k でのパラメータ集合である。

各ノードの報酬関数・方策関数は、各ペルソナに対する報酬関数・方策関数に基づき、ペルソナの割り当て率を用いて計算する。この方法により、エージェント個別の報酬関数・方策関数を導出する。具体的には、各ノードの報酬関数と方策関数は次のように定義する。

$$R_i(G^{(t)}|\psi, n_i) = \sum_{k=1}^{M} r_{ik} R_{ik}(G^{(t)}|\psi_k, n_i) \quad (15)$$

$$\pi_i(\Delta G^{(t)}|G^{(t)}, n_i) = \sum_{k=1}^{M} r_{ik} \pi_{ik}(\Delta G^{(t)}|G^{(t)}, \theta_k^{\pi}) \quad (16)$$

ここで、r_{ik} はノード i のペルソナ k に対する割り当て率であり、$\sum_{k=1}^{M} r_{ik} = 1$ となる。

5.3 報酬関数の最適化

前述した各エージェントについての報酬関数を使って、時刻 t の観測済みソーシャルネットワーク $G^{(t)}$ 全体における報酬関数の総和を以下のように定義する。

$$\Psi^* \leftarrow argmax_{\Psi} \sum_{t=1}^{T} R(G^{(t)} \mid \psi, n) \quad (17)$$

全ての時刻での観測済みソーシャルネットワーク報酬関数の総和をとり、その報酬が最大になるようなパラメーターを SGD で推定する。

5.4 方策関数・属性値の更新

次に、観測済みのソーシャルネットワークの系列から強化学習に基づいてペルソナの方策関数のパラメータを更新する。具体的には、マルチエージェント強化学習の手法の一つである COMA [4] によって学習する。この手法は、Actor-Critic 手法の一種で、行動を行う「Actor」と行動を評価する「Critic」の二つの部分から構成されている。Actor は方策関数をもとに環境に対して行動をする。環境は行動に対しての報酬を返す。この Actor の行動と報酬を Critic に入力として渡すことで行動に対する評価を得ることができる。この評価値を使って方策関数を改善する。一方で、Critic は環境からの報酬に対して $TD(\lambda)$ 法を使うことで価値関数を更新する。上記の方法で、Actor と Critic の更新を交互に行うことで学習を進める。属性値の更新は、以下の式で行う

$$X_i^{t+1} = \sum_i \{\lambda_i \cdot x_i^t + (1-\lambda) \sum_{n_j \in N(n_i)} w_i \cdot x_i^t\} \quad (18)$$

5.5 ペルソナの割り当て率の初期値と更新

まず、ペルソナの割り当て率の初期値の設定について説明する。最適化した報酬関数のパラメータ α, β を用いて混合ガウス分布の EM アルゴリズムを利用して各ノードに対してペルソナの割り当て率を算出する。次に、ペルソナの割り当て率の更新について説明する。ペルソナの割り当て率の更新をペルソナの方策関数を用いて計算する。具体的には、以下の式を用いて計算する。

$$r_{ik} = \frac{\sum_t \pi_i(\Delta G^{(t)}|G^{(t)}, \theta_k)}{\sum_t \sum_k \pi_i(\Delta G^{(t)}|G^{(t)}, \theta_k)} \quad (19)$$

上記の、ペルソナの割り当て率の更新を EM アルゴリズムの E-step で行う。方策関数の更新を EM アルゴリズムの M-step で行う。この E-step と M-step を交互に行うことで学習を進める。E-step・M-step を 100 回繰り返すもしくは、対数尤度関数が収束条件 (対数尤度関数の更新前と更新後の差が 0.1 以下) を満たす場合には、繰り返しを終了する。

5.6 将来予測

学習データから求めたペルソナの報酬関数と方策関数のパラメーターを用いて将来のソーシャルネットワークを予測する。本手法ではエッジの生成予測と生成結果による各ノードの報酬と特徴量の更新を交互に行なうことで複数ステップ先のソーシャルネットワークにおけるエッジとノードの状態を予測する。具体的には、ペルソナの割り当て率を使って各ノードに対してペルソナを割り当てる。その後、割り当てられたペルソナの学習済み報酬関数 [式 13]、方策関数 [式 14]、ペルソナの割り当て率 [式 19]、属性値の更新式 [式 18] を用いてノード間のエッジの生成確率、ノードの属性値を計算し、それらによってネットワークグラフの状態を更新することで未観測時系列のネットワークを生成する。

6 実　　験

本節では、「実データを用いて提案手法がどれだけ正確にネットワークを予測できるか」と「実データをどのように欠損させた場合に提案手法が効果的か」を検証する。

6.1 データセット

データセットとして、論文 [21]、[19] で使われている NIPS、DBLP のデータセットを用いる。以下に、それぞれのデータセットの収集、作成方法を記述する。

6.1.1 NIPS

NIPS は研究者の共著者ネットワークである。ノードは著者、エッジは共著関係、時間区分は年単位である。ある年に二人の著者 u、 v が共著論文を発表したとき、その二人の著者間に双方向の有向エッジが張られる。各ノードはある年に発表した論文のタイトルに含まれる単語を属性値としてもつ。2008 年から 2017 年に発表された機械学習分野の 7 国際会議の論文データを取得し、2008 年から 2012 年までを学習データ、2013 年から 2017 年までをテストデータとして用いる。前処理として各単語はステミング処理を行い、全時間区分に出現しない著者はデータセットから除外した。ノード数は 32、属性値数は 2411、密度は 0.0806、平均次数は 2.16 である。

6.1.2 DBLP

DBLP は研究者の共著者ネットワークである。ネットワークデータは NIPS と同様の方法で作成した。2008 年から 2017 年に発表されたデータマイニング、データベース、自然言語処理、機械学習、人工知能、情報検索、コンピュータビジョンから構成されるの 47 国際会議の論文データを取得し、2008 年から 2012 年までを学習データ、2013 年から 2017 年までをテストデータとして用いる。ノード数は 500 で、属性値数は 3854、密度は 0.0086、平均次数は 0.08 である。

6.2 ベースライン

報酬関数と方策関数にペルソナを導入していないモデルと報

酬関数と方策関数にペルソナを導入しているがペルソナの割り当ての計算を変えたモデルをベースライン手法として用いることで、提案手法の有効性を検証する。

本研究では、比較手法として“Reinforce モデル”[21] を採用する。この手法は、報酬関数と方策関数にペルソナを導入していないモデルである。その他に、ペルソナの割り当てに k-means を使ったモデルを採用した。この手法は、報酬関数と方策関数にペルソナを導入しているが、各ノードを一つのペルソナにしか割り当てない。

6.3 実　　験

この項では、実施した 2 つの実験については説明する。実験 1 は、DBLP の 10 年分のデータセットの初めの 5 年分のデータを訓練データとし残りの 5 年分をテストデータとした。訓練データを用いて Reinforce モデル、提案手法を使って学習を行った。その後学習済みのモデルを使用して計算した予測値とテストデータを使って予測精度を算出することでモデルを評価した。NIPS についても同様の方法で行なった。実験 2 は、DBLP の 10 年分のデータセットの初めの 5 年分のデータを訓練データとし残りの 5 年分をテストデータとした。この時、学習データの 5 年目のデータの欠損を以下の方法で行った。

（1） 学習データの 5 年目のデータに対してエッジの接続情報を割合 (5%,15%,30%,50%,75%) を変えて欠損させる (元のデータを 0 とする)

（2） 学習データの 5 年目のノードの属性値を割合 (5%,15%,30%,50%,75%) を変えて欠損させる (元のデータを 0 とする。)

（3） 学習データの 5 年目のデータを全て欠損させる (学習データとして初めの 4 年分を使用)

(1),(2) では、[2], [7] の研究でユーザの行動が 5 つのパターンに分類できることから、ペルソナ数は 5 に固定した。(3) では、ペルソナ数は実験 1 で最も精度が高いモデルを採用した。また、本研究では予測精度の検証に AUC（Area Under Curve）を用いた。AUC が高いほど正確に将来のネットワークを予測していることを示している。結果の図では、縦軸が AUC、横軸が観測済みデータの 2013 年を 1 とした時刻を表す。また、濃い色の線は各条件下で 10 回の将来予測の平均を表し、薄い色の線は、各条件下で 10 回の将来予測での予測値の範囲である。

6.4 実験 1 結果

図 2 は、NIPS のデータを対象にエッジの生成の予測精度を提案手法と“Reinforce モデル”とで比較した図である。図 4、5 は図 2 を見やすいように横軸の時間を特定の時間に焦点をあてたものである。図 3 は、NIPS のデータを対象にノードの属性値の予測精度を提案手法と Reinforce モデルを比較したものである。エッジ生成の実験結果から、将来の時刻 $t = 1, 2, 5$ で提案手法のペルソナ数 4(persona4) モデルがエッジの生成予測で REINFORCE モデルより高い精度を出していることが分かる。しかし、ノードの属性値の予測に関してはペルソナ数 3(persona3) モデルが最も精度が高いが Reinforce モデルよりも高い精度を出すことはなかった。

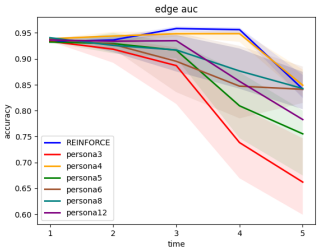


図 2: NIPS に対して
エッジ予測の AUC

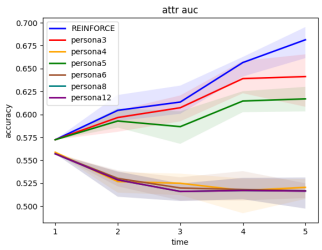


図 3: NIPS に対して
属性値予測の AUC

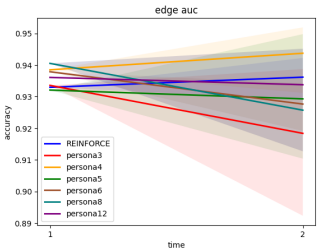


図 4: NIPS に対してエッジ予測の
AUC(時刻 t=1 2)

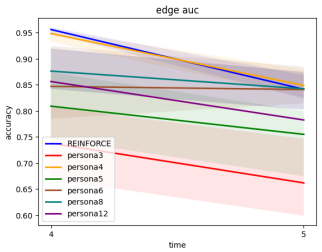


図 5: NIPS に対してエッジ予測の
AUC(時刻 t=4 5)

図 6 は、DBLP のデータを対象にエッジの生成の予測精度を提案手法と“Reinforce モデル”とで比較した図である。図 7 は、DBLP のデータを対象にノードの属性値の予測精度を提案手法と“Reinforce モデル”とで比較した図である。エッジ生成の実験結果から、ペルソナ数 8(persona8) モデルが最も精度が高く、将来の時刻 $t = 2 \sim 5$ で提案手法のエッジの生成予測で Reinforce モデルより高い精度を出していることが分かる。他にも、複数の提案手法のモデルで高い精度を出している。しかし、ノードの属性値の予測に関してはペルソナ数 4(persona4) モデルが最も精度が高いが Reinforce モデルよりも高い精度を出すことはなかった。

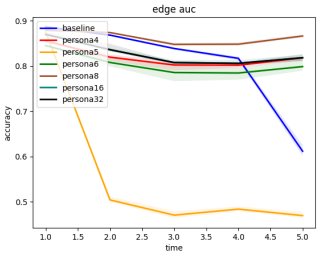


図 6: DBLP に対して
エッジ予測の AUC

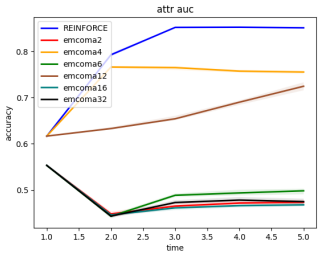


図 7: DBLP に対して
属性値予測の AUC

7 実験 2 結果

図 12 は、NIPS のデータを対象に学習データの 5 年目のデータを全て削除した場合の将来予測の結果である。図 17 は、DBLP のデータを対象に学習データの 5 年目のデータを全て削除した場合の将来予測の結果である。図 12,17 で Complete Proposed は、ペルソナ・モデリングで学習データの欠損を行わない場合である。InComplete Proposed は、ペルソナ・モデリングで学習データの欠損を行なった場合である。同様に、Complete Baseline は、Reinforce モデルで学習データの欠損を行わない場合である。InComplete Baseline は、Reinforce モデルで学習データの欠損を行なった場合である。図 8 は、NIPS の 5 年目の学習データから削除の割合を変えてエッジを削除したデータに対して Reinforce モデルを使った将来予測の結果である。図 9 は、図 8 と同様の方法で欠損させたデータに対してペルソナモデリングを使った将来予測の結果である。図 10 は、NIPS の 5 年目の学習データから削除の割合を変えて属性値を削除したデータに対して Reinforce モデルを使った将来予測の結果である。図 11 は、図 10 と同様の方法で欠損させたデータに対してペルソナモデリングを使った将来予測の結果である。

図 13 は、DBLP の 5 年目の学習データから削除の割合を変えてエッジを削除したデータに対して Reinforce モデルを使った将来予測の結果である。図 14 は、図 13 と同様の方法で欠損させたデータに対してペルソナモデリングを使った将来予測の結果である。図 15 は、DBLP の 5 年目の学習データから削除の割合を変えて属性値を削除したデータに対して Reinforce モデルを使った将来予測の結果である。図 16 は、図 15 と同様の方法で欠損させたデータに対してペルソナモデリングを使った将来予測の結果である。図 12, 図 17 より、ペルソナ・モデリングは 5 年目のデータを欠損させた場合に Reinforce モデルより精度の低下が小さいことが分かる。図 8-11, 図 13-16 より、属性値・エッジの一方を欠損させた場合は、欠損の割合を増やすにつれて Reinforce モデルとペルソナ・モデリングの双方で精度が落ちている。図 14、16 の結果より、学習データのエッジの情報を欠損させる場合より属性値を欠損させる場合の方が、データの欠損が無い学習データを用いる場合と比べて精度の低下が小さいことが分かる。

8 考　　察

まず、実験 1 の結果について考察する。NIPS、DBLP ともに将来の時刻 $t = 5$ でエッジの生成予測が、Reinforce モデルに比べて提案手法のモデルは精度が高いことが多い。このことから、ペルソナを用いることでより先の未来での行動を予測を高い精度で行うことができることが分かる。これは、入力データに含まれるデータに加えて、ペルソナという潜在的な情報を使うことができるため動的に変化する人の行動を観測済みのデータを使って学習できているためだと考えられる。

次に、実験 2 の結果を考察する。データの欠損のさせ方で精度が変わることから、データの欠損のさせ方によってペルソナが効果的に活用できるかが変わると考えられる。DBLP のデータに対してペルソナ・モデリングは、属性値を欠損させた場合にはモデルの精度低下が小さいのは、DBLP は NIPS に比べて平均次数が小さいためエッジの情報がより重要になっているため、DBLP のデータに対して属性値情報を欠損させた場合は影響が小さく精度の低下が小さいと考えられる。また、ペルソナ・モデリングが Reinforce モデルより精度が低いのは、ペルソナの割り当て・ペルソナの方策関数の学習が上手く行われていない場合はそのペルソナを用いる全てのノードのエッジの生成・属性値の予測に影響を与えるためだと考えられる。特に、割り当てられているデータが少ないペルソナに属するデータを欠損させた場合は、ペルソナの学習が進まず精度が低くなっていると考えられる。以上のことから、ペルソナを効果的に活用するには、学習データの性質に応じてペルソナが学習できるように学習データを欠損させる必要がある。また、少ない割合でデータを欠損させる場合は、割り当てられているデータが少ないペルソナに属するデータを欠損させないように注意する必要がある。今回用いたデータに関しては、NIPS はエッジ・属性値の双方のデータを消す場合が効果的であった。DBLP はエッジの予測には属性値を欠損させた場合が、ノードの属性値の予測の場合にはエッジ・属性値のの双方のデータを消す場合が効果的であった。また、属性値の予測では NIPS・DBLP の両方でエッジの生成に比べて精度が低いのは、誤差逆伝播での属性値の変化を表現するパラメータの勾配が小さいためうまく学習できなかったと考えられる。この点については、今後の研究でパラメータの調整や属性値の変化のモデリングを見直すことを検討している。

9 ま と め

本研究ではソーシャルネットワークの時間経過による変化を‘ペルソナ’を用いて予測する手法を提案した。提案モデルは、ソーシャルネットワークでの人間の行動はいくつかのパターンに分けることができるという仮定をし、ノードへのペルソナ割り当てと割り当てられたペルソナの方策関数を繰り返し学習する強化学習と EM アルゴリズムを融合したモデルで予測を行う手法を提案した。実データに基づく予測精度の比較で従来手法に対してペルソナの数を調整することで高い精度でエッジ生成確率を予測することができることを示した。また、ペルソナの数を調整することで従来手法と同程度のエッジの生成、ノードの属性値を予測することができるを示した。他にも、従来手法と提案手法に対してデータの欠損のさせ方を変えることで、ペルソナの学習ができるようにデータを欠損させることでペルソナの導入が効果的であることがわかった。これらの結果は、ペルソナを導入することによりデータが欠損が多い場合に従来手法と比較して高精度に将来のネットワークを予測できる可能性、全てのデータを計算量の問題で使えない場合にペルソナの学習ができる程度に欠損をさせたデータを使うことで将来予測ができることを示唆している。今後の展望としては、本研究ではクラスの数を事前に決めてクラスタリングを行うモデルを構築し

たが、ノンパラメトリックベイズなどの手法を用いてクラスの数を自動で決定するモデルの考案、ソーシャルメディア以外の異なるネットワークデータへモデルを適応や、生成されたペルソナを分析し異なるデータ間でのペルソナの接続関係の詳細な分析や、欠損のさせ方を変えた場合の分析を検討している。

10 謝　　辞

本研究の一部は、JSPS 科研費 (JP22H00508,JP22K17944) の支援を受けたものである。ここに謝意を示す。

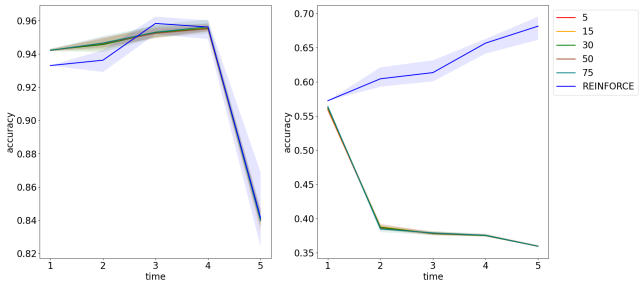


図 8: NIPS に対してエッジを削除 (Reinforce モデル)

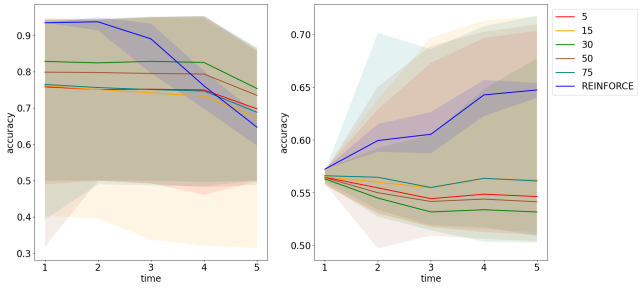


図 9: NIPS に対してエッジを削除 (ペルソナ・モデリング)

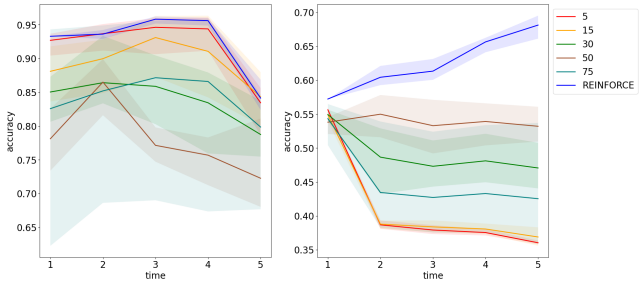


図 10: NIPS に対して属性値を削除 (Reinforce モデル)

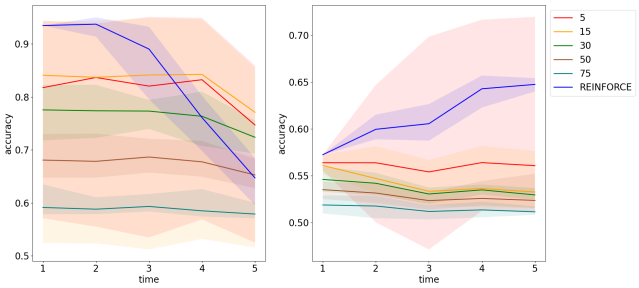
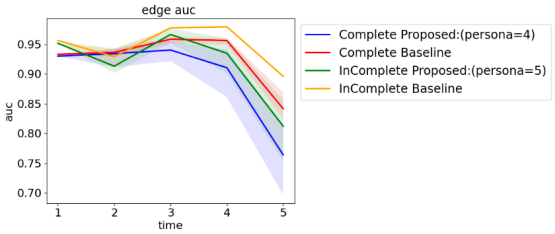


図 11: NIPS に対して属性値を削除 (ペルソナ・モデリング)



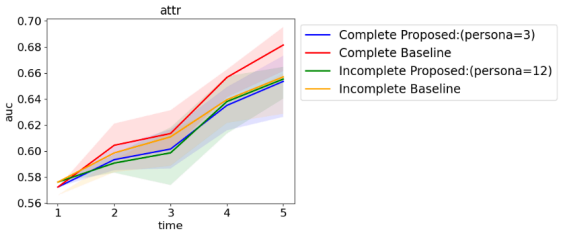


図 12: 学習データの 5 年目のデータを削除

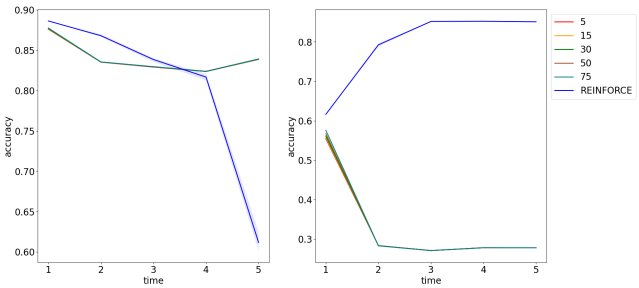


図 13: DBLP に対してエッジを削除 (Reinforce モデル)

文　　献

[1] Dynamic network embedding by modeling triadic closure process. Vol. 32, , Apr. 2018.

[2] Petter Bae Brandtzaeg and Jan Heim. A typology of social networking sites users. *Int. J. Web Based Communities*, Vol. 7, No. 1, p. 28–51, jan 2011.

[3] Varsha Chauhan, Alexander Gutfraind, and Ilya Safro. Multiscale planar graph generation, 2019.

[4] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32, 2018.

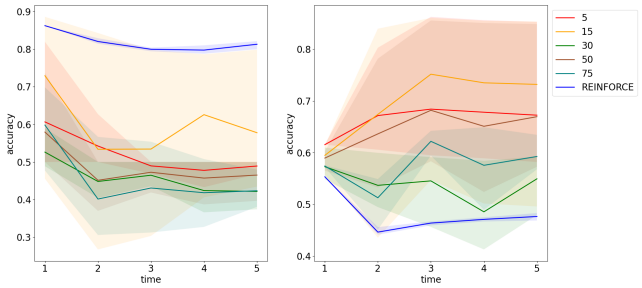


図 14: DBLP に対してエッジを削除 (ペルソナ・モデリング)

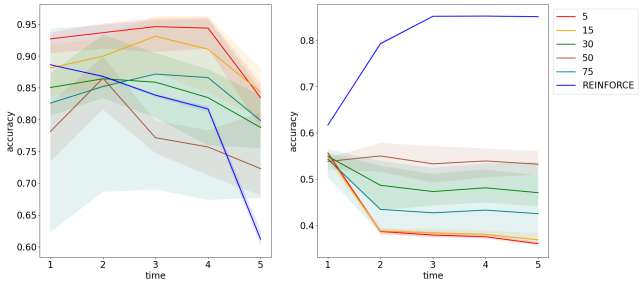


図 15: DBLP に対して属性値を削除 (Reinforce モデル)

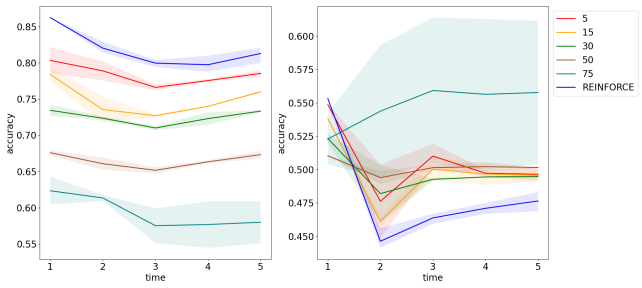


図 16: DBLP に対して属性値を削除 (ペルソナ・モデリング)

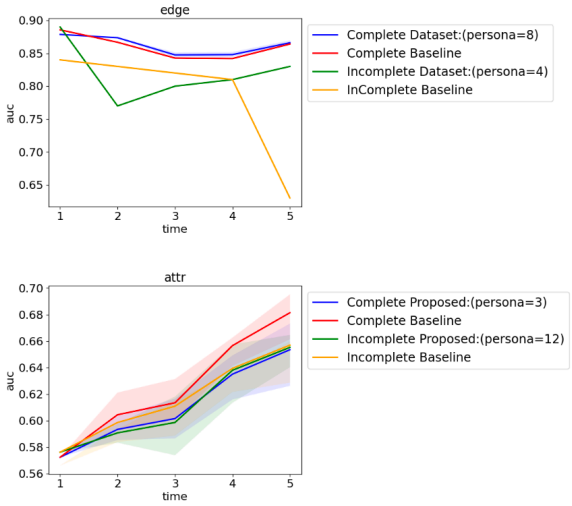


図 17: 学習データの 5 年目のデータを削除

[5] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016.

[6] Ziyang Guo, Zhenyu Chen, Peng Liu, Jianjun Luo, Xun Yang, and Xinghua Sun. Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks. *IEEE Journal on Selected Areas in Communications*, Vol. 40, No. 5, pp. 1587–1599, 2022.

[7] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. What we instagram: A first analysis of instagram photo content and user types. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8, No. 1, pp. 595–598, May 2014.

[8] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[9] Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, and Evangelos Kanoulas. Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, p. 1764–1773, New York, NY, USA, 2018. Association for Computing Machinery.

[10] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020.

[11] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ' 14. ACM, August 2014.

[12] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW ' 15. International World Wide Web Conferences Steering Committee, May 2015.

[13] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[14] Hongwei Wang, Jia Wang, Jialin Wang, Miao Zhao, Weinan Zhang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Graphgan: Graph representation learning with generative adversarial nets, 2017.

[15] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds, 2019.

[16] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, Vol. 8, No. 3–4, p. 229–256, may 1992.

[17] Guangrui Yang, Lihua Yang, Qing Zhang, and Zhihua Yang. An equivalent graph reconstruction model and its application in recommendation prediction, 2023.

[18] Chuxu Zhang, Dongkuan (KD) Xu, Mojan Javaheripi, Subhabrata Mukherjee, Lingfei Wu, Yinglong Xia, Jundong Li, Meng Jiang, and Yanzhi Wang. Relkd 2023: International workshop on resource-efficient learning for knowledge discovery. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, p. 5901–5902, New York, NY, USA, 2023. Association for Computing Machinery.

[19] Christos Faloutsos 伊藤 寛祥. 動的なソーシャルネットワークにおける興味関心の伝搬を考慮した将来予測モデル. pp. I25–3, 2021.

[20] 鹿島久嗣, 安倍直樹. ネットワーク構造の確率的な時変モデルに基づく教師ありリンク予測. 人工知能学会論文誌, Vol. 22, No. 2, pp. 209–217, 2007.

[21] Christos Faloutsos 森嶋 厚行三宅 健太郎. マルチイベント探知に学習におけるソーシャルネットワークの将来予測手法. pp. 1b–8–3, 2023.

[22] 武司山村, 澄雄大野, 正廣奥, 康民千種, 隆小池. 顧客像の類似性をプロファイリング分析した書籍推薦システム. 第 67 回全国大会講演論文集, 第 2005 巻, pp. 209–210, mar 2005.

時間的特徴を制御可能な時間グラフ生成器

中西　宏和†　　菊地　良将†　　佐々木　勇和†　　鬼塚　　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5
E-mail: †{nakanishi.hirokazu,kikuchi.ryosuke,sasaki,onizuka}@ist.osaka-u.ac.jp

あらまし　時間グラフは実世界における多くの複雑なシステムを表現することができる．最近，深層学習ベースの方法により時間グラフを生成するモデルが数多く提案されている．しかし，既存の手法は入力に類似したグラフを生成することを目的としており，ユーザはグラフの特徴を制御することができない．私たちは，ユーザが解釈可能な段階まで時間的・構造的特徴を抽象化することで，時間グラフの特徴を制御することができる ConTeG を提案する．具体的には，まずテンソル分解により入力グラフから時間グラフの時間的・構造的特徴を分離する．次に，時間的特徴を表す抽象表現を時系列データ分解手法により 3 つの要素に分解する．こうして得られた解釈可能な特徴にユーザの望む変更を施したあと，スケーラブルなグラフが生成可能な確率的エッジ生成手法により出力グラフを得る．実験では，特徴に変更を施したグラフの分析により，グラフ特徴の制御の妥当性を示した．また，意図的に特徴に変更を施した複数のグラフを既存のリンク予測手法に適用することで，各手法のメリット・デメリットを明らかにした．

キーワード　動的グラフ，グラフ生成，テンソル分解，時系列分析

1　はじめに (Introduction)

グラフは，様々な領域における実世界の複雑なシステムを抽象化することが可能なツールである．実世界のシステムは時間と共に変動する動的な性質を示すことが多いため，エッジに時間情報を付与した時間グラフとして扱うことがある．時間グラフには，ソーシャルネットワーク [1] や推薦ネットワーク [2] など様々なものが存在しており，これらを分析することで新たな知見を得ることができる．中でも，グラフの生成は，グラフ関連の手法のベンチマーク研究 [3] や代替データの共有 [4] を可能にするため，熱心に研究されている．

初期の伝統的なグラフ生成モデル [5] はグラフの Preferential Attachment [6] 等の構造的特徴を仮定した上でグラフを生成するヒューリスティックが主流であった．近年は，深層生成モデルの登場により，事前にグラフの構造的特徴の仮定を置かずとも，入力グラフから自動的にグラフの構造的特徴を学習し，実世界ネットワークをシミュレートした合成グラフを生成することが可能な手法 [7,8] が提案されている．これらヒューリスティックと深層学習モデルは，当初は静的なグラフを対象としていることが多かったが，最近になって時間グラフを生成することができる手法 [9–12] もいくつか登場してきている．

しかし，これらの時間グラフ生成器は多様なグラフを生成できないという機能の問題がある．ヒューリスティックは仮定した時間的特徴に従わないグラフを扱えない．深層学習モデルはモデルパラメータの解釈が難しいため，ユーザがグラフの特徴を自由に変更することができない．もし，ユーザによる生成グラフの時間的・構造的特徴の制御が可能であれば，意図的にある特徴を含んだ時間グラフを生成することができる．こうして生成した多様なグラフを時間グラフを扱う様々な分析手法 [13] に適用すれば，各手法の精度と実行時間に関するベンチマーク実験により，各手法の利点・欠点を明らかにすることができる．

上記の問題に対して，時間グラフの時間的・構造的特徴をユーザが解釈可能な段階まで抽象化することで，時間グラフの特徴を制御することができる高スケールな時間グラフ生成器 Controllable Temporal Graph Generator based on Tensor Factorization and Time Series Analysis (ConTeG) を提案する．本論文では以下の技術的課題にチャレンジする．

1) 解釈可能性：どのようにしてグラフからユーザが解釈可能な時間的特徴を得るか．

1-a) 時間・構造的特徴の分離：どのようにして時間・構造的特徴を分離するか．

1-b) 解釈可能な時間的特徴の抽出：どのようにして解釈可能な時間的特徴を得るか．

2) スケーラビリティ：どのようにしてスケーラブルなサイズのグラフを生成するか？

ConTeG では上記の技術的課題を以下の方針により解決する．具体的には，1-a) 時間的特徴と構造的特徴の分離については，グラフを Source node，Target node， Timestep の 3 つの軸で構成される動的隣接テンソルとみなし，テンソル分解手法により各軸ごとの潜在因子に分解する．1-b) 解釈可能な時間的特徴の抽出については，時間軸方向の抽象表現に対し，時系列分析手法を適用することで，解釈可能な 3 つの時間成分 （Seasonal，Trend，Residual） を得る．こうして得られた解釈性の高い抽象表現をユーザが変更することで，生成グラフの時間的特徴を制御可能とする．2) スケーラビリティについては，3 つの潜在因子からそれぞれ生成するエッジの要素 (Source node，Target node， Timestep) を確率的に選択することで，エッジを生成する．これによりテンソル分解の逆操作である高コストなテンソル積を経由しないため，エッジ生成の計算コストを削減できる．

実験では，ConTeG のグラフ生成能力と効率性に関する実験と，ConTeG により生成した多様なグラフを用いた時間グ

ラフニューラルネットワークのベンチマーク実験を実施する．ConTeG に関する実験では，グラフ特徴の制御が出力グラフに正しく反映されているかどうかと，生成グラフのノードサイズに対する実行時間のセンシティビティについて確認する．ベンチマーク実験では，ConTeG によりいくつかの時間的・構造的特徴を段階的にを変更したグラフを，7 つの時間グラフニューラルネットワークに適用することによりベンチマーク実験を実施する．

2 関連研究 (Related Work)

2.1 時間グラフ生成器

近年，時間グラフを生成できるヒューリスティックが提案されている．DYMOND [9] は，構造的な特徴であるモチーフを捉えることのできる生成モデルであり，モチーフにおける各ノードの役割にも注目して，時間グラフを生成することができる．MTM [10] は DYMOND よりも高次なモチーフを捉えられる手法であり，計算効率も大幅に改善されている．

近年，深層生成モデルの研究が進んだことで，グラフ構造を深層生成モデルによりモデリングすることに注目が集まっている．TagGen [11] は Self Attention により生成したウォークを GAN により検証し，時間グラフを構築する．TIGGER [12] は点過程と自己回帰モデリングを組み合わせることで，大規模で複雑な時間グラフを効率的に生成する．

2.2 グラフ特徴を制御可能なグラフ生成器

グラフ統計量に基づく古典的なヒューリスティックな方法として， Barabási–Albert モデル [14] と Kronecker Graph モデル [15] がある．Barabási–Albert モデルはノード間の接続に関するパラメータを制御することで，フリースケールな次数分布やハブのような構造的特徴を制御してグラフを生成することができる．Kronecker Graph モデルは行列のクロネッカー積を用いることで，次数分布の冪乗分布や小半径などの構造的特徴を制御してグラフを生成することができる．GenCAT [16] は，行列分解により得られる入力グラフの抽象表現を制御することで，ホモフィリー/ヘテロフィリーや次数分布といった構造的特徴を制御可能なグラフ生成器である．しかし，これらのグラフ生成器は時間グラフに対応しておらず，静的なグラフを生成することしかできない．

3 事 前 準 備

ここでは，時間グラフや時間的特徴，構造的特徴，問題定義，テンソル分解についての定義や説明を行う．

3.1 時間グラフ

本研究では時間グラフを離散時間動的グラフ （DTDG: Discrete Time Dynamic Graph） として扱う．DTDG を G，G のノードサイズを n，エッジサイズを e，タイムステップサイズを t とする．G には，エッジシーケンス $S \in \mathbb{N}^{(e\times 3)}$ と動的隣接テンソル $A \in \{0,1,2,\dots\}^{(n\times n\times t)}$ の 2 つの表現方法がある．エッジシーケンスは，各エッジの情報 (Source node, Target node, Timestep) をシーケンスとして保持する表現方法である．E の l 行目に (Source node, Target node, Timestep) $= (i,j,\tau), i,j,\tau \in \mathbb{N}$ というエッジが格納されている場合，S の l 行目は $S_{l\cdot} = [i,j,\tau]$ と表される．一方，動的隣接テンソルは，静的グラフにおける隣接行列を DTDG に拡張した表現方法である．タイムステップ τ において Source node が i，Target node が j であるエッジが $e \in \{0,1,2,\dots\}$ 個ある場合，$A_{ij\tau} = e$ である．

空間的計算量の観点では，動的隣接テンソルは $O(n^2t)$ であり，密な表現である．一方，エッジシーケンスは $O(e)$ であり，疎な表現であるためメモリ使用量が少ない．

3.2 時間的特徴

時間グラフにおいて，周期的なパターン （例えば，毎週月曜日の朝にメールを送信など） [17] や長期的な変動傾向 （例えば，フォロワー数の増加に伴うリツイート数の増加など） [18]，短期間だけイベントが増加するスパイク (例えば，システムの障害に伴う短期的のアクセス数の増加) [19] といった時間的特徴が確認されている．これら特徴はある時刻におけるリンクの発生数の分布により確認することができる．Kaijie らの研究 [20] では，これを CSS (Concurrent set size) と定義しており，時刻 τ における $CSS(\tau)$ は動的隣接テンソル A により以下のように定式化される．

$$CSS(\tau) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij\tau} \tag{1}$$

これら時間的特徴は時系列分析手法により抽出することが可能である．STL (Seasonal Decomposition Of Time Series By Loess) [21] に代表されるような分解手法では，時系列データを Seasonal と Trend，Residual に分解する．Seasonal は周期的なパターン [17] に，Trend は長期的な傾向 [18] に該当し，Residual は元の時系列データから Seasonal と Trend を取り除いたものである．元となる時系列データを $\boldsymbol{y} \in \mathbb{R}^t$，分解される要素のうち，Seasonal を $\boldsymbol{s} \in \mathbb{R}^t$，Trend を $\boldsymbol{\mu} \in \mathbb{R}^t$，Residual を $\boldsymbol{r} \in \mathbb{R}^t$ とすると，時系列データ $\boldsymbol{y} \in \mathbb{R}^t$ は以下のように 3 つの成分に分解される．

$$\boldsymbol{y} = \boldsymbol{s} * \boldsymbol{\mu} * \boldsymbol{r} \tag{2}$$

ただし，$*$ は要素積を表す．STL では，Seasonal と Trend は LOESS (Locally Estimated Scatterplot Smoothing) [22] という平滑化技術を使用して抽出される．また，STL には和と積の分解方法があるが，本研究では積の分解方法を活用している．Residual は元の時系列データから Seasonal と Trend を取り除くことによって得られる．

3.3 構造的特徴

時間グラフには，ノード数やエッジ数といった最も基本的なものから，冪乗則に従う次数分布やホモフィリー・ヘテロフィリーといった静的グラフの構造的特徴がある．また，近年は深層学習を活用することでより複雑なモチーフ [23] を捉える研究

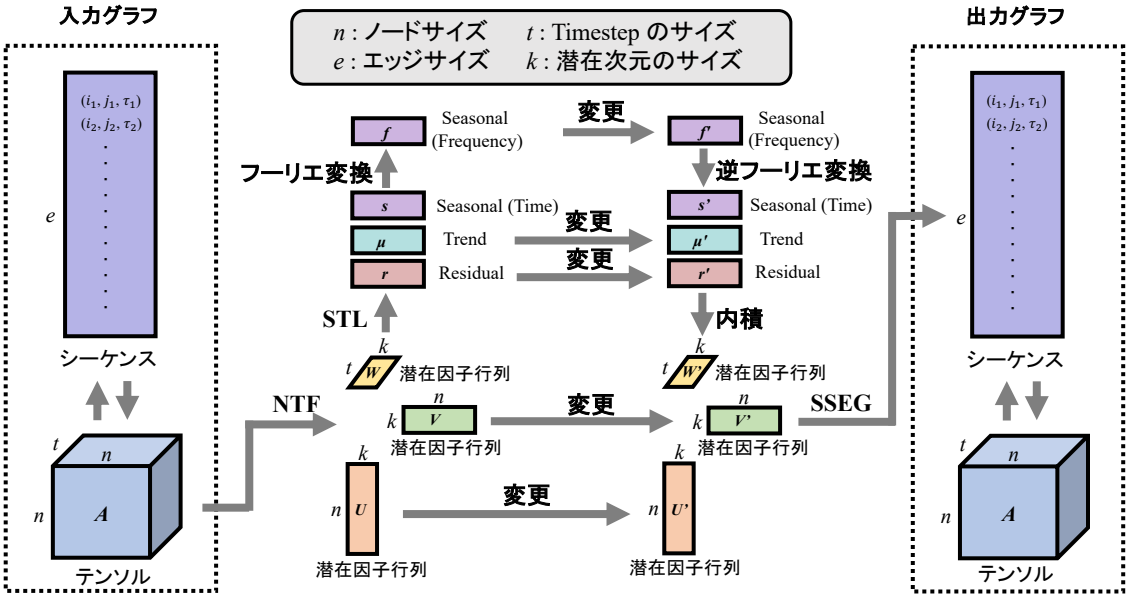


図 1: 提案手法 ConTeG の概要図．まず，入力グラフは NTF 分解により 3 つの潜在因子行列に分解される．そのうち Timestep 軸の潜在因子行列 W については STL 分解を施し，さらに Seasonal と Trend，Residual に分解し，Seasonal についてはフーリエ変換を施して時間領域から周波数領域に変換する．こうして抽出されたグラフ特徴はユーザによる変更を行なった後，分解の時と逆の順序で分解の逆操作を行い出力グラフを得る．ただし，潜在因子行列の合成については計算コストの高いテンソル積の代わりに，潜在因子行列から直接エッジを生成可能な SSEG という方法により最終的にはエッジシーケンスの形式で出力グラフを得る．

もなされている．

3.4 テンソル分解

テンソル分解とは，テンソルをより階数の少ないテンソルの積や和で表現する数学的な手法である．その中でも，非負値テンソル分解 NTF （Non-negative Tensor Factorization） [24] は，与えられた非負値テンソルを複数の非負値の行列の積に分解する手法である．トピックモデリングや画像処理，生体処理などの様々な分野で応用されている．

3 階のテンソル $X \in \mathbb{R}^{(d_1 \times d_2 \times d_3)}, d_1, d_2, d_3 \in \mathbb{N}$ は，潜在因子行列 $U \in \mathbb{R}^{(d_1 \times k)}$，$V \in \mathbb{R}^{(d_2 \times k)}$，$W \in \mathbb{R}^{(d_3 \times k)}$ に分解され，以下のように表される．

$$X \simeq U \otimes V \otimes W \tag{3}$$

ただし， $k \in \mathbb{N}$ は潜在次元数，$\otimes$ はテンソル積である．X' を $X' = U \otimes V \otimes W$ とすると，X'_{hij} は以下のように計算される．

$$X'_{hij} = \sum_{m=1}^{k} U_{hm} V_{im} W_{jm} \tag{4}$$

この分解は以下の目的関数 L に基づいて行われる．

$$\begin{aligned} L &= ||X - U \otimes V \otimes W|| \\ &= \sum_{h=1}^{d_1} \sum_{i=1}^{d_2} \sum_{j=1}^{d_3} (X_{hij} - \sum_{m=1}^{k} U_{hm} V_{im} W_{jm})^2 \end{aligned} \tag{5}$$

目的関数は，元のテンソル X と 3 つの潜在因子行列 U，V，W のテンソル積との差の二乗和のノルムを最小化する．つまり，目的関数により 3 つの潜在因子行列はそれらのテンソル積が元のテンソルに近づくように最適化される．

4 提 案 手 法

本章では，ユーザが時間的特徴を制御可能な時間グラフ生成器 ConTeG を提案する．ConTeG の概要図を図 1 に示す．以降の各節において，主たる機能を説明する．

4.1 グラフ特徴の抽出

グラフ特徴の抽出は，テンソル分解による抽出と時系列分析手法による抽出の 2 つのステップに分けられる．さらに，得られた抽象表現を制御する方法について簡単に説明する．

テンソル分解による抽出．時間的特徴と構造的特徴の分離を行うために，入力の時間グラフを動的隣接テンソルとみなしテンソル分解手法に入力することで各軸ごとの抽象表現を得る．テンソル分解手法には NTF を用いる．テンソル分解手法を用いる理由は，要素値が非負である動的隣接テンソルは NTF の入力の条件である非負値制約を満たしており，その制約により解釈性の高い表現の抽出が期待されるためである．

式 3，式 5 に基づいて NTF を動的隣接テンソル $A \in \mathbb{R}^{(n \times n \times t)}$ に適用すると，潜在因子行列 $U \in \mathbb{R}^{(n \times k)}$，$V \in \mathbb{R}^{(n \times k)}$，$W \in \mathbb{R}^{(t \times k)}$ に分解される．各潜在因子行列について，U，V，W は動的隣接テンソルにおける Source node，Target node，Timestep の軸をそれぞれ抽象化している．

NTF は入力されたテンソルから潜在次元サイズ k の数だけ潜在因子を抽出する．この k 個の潜在因子を**基底クラス**と定義する．各基底クラスは U，V，W の各列によって表現されている．例えば，l 番目の基底クラスは $U_{\cdot l}$，$V_{\cdot l}$，$W_{\cdot l}$ により表現されている．このベクトルを**基底クラスベクトル**と定義する．つまり，各基底クラスは対応する 3 つの基底クラスベクト

ルにより要約されている．

時系列分析手法による抽出．次に，時間的特徴の解釈性を高めるために，時系列分析手法として STL とフーリエ変換を Timestep 軸の潜在因子行列 W に適用する．まず，W の各列，つまり基底クラスベクトルに STL を適用することで Seasonal と Trend，Residual に分解する．式 2 に基づき基底クラス l の基底クラスベクトル $W_{\cdot l}$ を STL に適用すると，Seasonal $\boldsymbol{s}_l \in \mathbb{R}^t$，Trend $\boldsymbol{\mu}_l \in \mathbb{R}^t$，Residual $\boldsymbol{r}_l \in \mathbb{R}^t$ の 3 つの要素に分解される．こうすることで，基底クラスベクトル $W_{\cdot l}$ の周期的なパターンと長期的な傾向が分離され，時間的特徴の解釈性が増す．

更に，Seasonal にフーリエ変換 $\mathcal{F}[\cdot]$ を施し，時間領域から周波数領域に変換する．フーリエ変換後の Seasonal $\boldsymbol{f}_l$ は，以下のように表される．

$$\boldsymbol{f}_l = \mathcal{F}[\boldsymbol{s}_l] \tag{6}$$

これにより，Seasonal のスペクトル情報が得られる．

グラフ特徴の制御．得られた 5 つのグラフ特徴 $\boldsymbol{f}$,$\boldsymbol{\mu}$, $\boldsymbol{r}, U, V$ に対して，変更を施すことにより生成グラフの特徴の制御を行う．ConTeG では，時間的特徴に関しては特定周期の強調や長期的上昇/減少傾向やピークの追加が可能であり，構造的特徴に関してはノード数と平均次数を制御することができる．以降，グラフ特徴の変更を施した抽象表現 （変更しないものも含める） はハット付きの文字 $\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}, \hat{U}, \hat{V}$ により表現する．

4.2 グラフ生成

グラフ特徴の変更が反映された抽象表現 $\hat{U}$，$\hat{V}$，$\hat{\boldsymbol{f}}$，$\hat{\boldsymbol{\mu}}$，$\hat{\boldsymbol{r}}$ から時間グラフを生成する．グラフの生成は，基本的にはグラフ分解の逆操作を行うことにより実現する．まず，フーリエ変換と STL の逆操作を行い，$\hat{\boldsymbol{f}}$，$\hat{\boldsymbol{\mu}}$，$\hat{\boldsymbol{r}}$ から，時間的特徴の変更が反映された $\hat{W}$ を得る．その後，3 つの潜在因子行列 $\hat{U}$，$\hat{V}$，$\hat{W}$ の合成を行い，最終的な時間グラフ $\hat{G}$ を獲得する．

STL の要素の合成．まず，周波数領域における Seasonal $\hat{\boldsymbol{f}}$ に逆フーリエ変換 $\mathcal{F}^{-1}[\cdot]$ を施すことにより時間領域の Seasonal $\hat{\boldsymbol{s}}$ に戻す．基底クラス l の場合を以下に表す．

$$\hat{\boldsymbol{s}}_l = \mathcal{F}^{-1}[\hat{\boldsymbol{f}}_l] \tag{7}$$

次に，STL の要素の合成については，基底クラス l の変更後の基底クラスベクトル $\hat{\boldsymbol{s}}_l$，$\hat{\boldsymbol{\mu}}_l$，$\hat{\boldsymbol{r}}_l$ の要素積を計算することにより，潜在因子行列 $\hat{W}$ の $\hat{\boldsymbol{w}}_l$ が得られ，以下のように表される．

$$\hat{W}_{\cdot l} = \hat{\boldsymbol{s}}_l * \hat{\boldsymbol{\mu}}_l * \hat{\boldsymbol{r}}_l \tag{8}$$

これを全ての基底クラスについて計算し，$\hat{W}$ を得る．以上より，グラフの特徴変更後の潜在因子行列 $\hat{U}$，$\hat{V}$，$\hat{W}$ が揃う．

潜在因子行列の合成.

潜在因子行列の合成では，$\hat{U}$，$\hat{V}$，$\hat{W}$ から最終的な生成グラフとしてエッジシーケンス $\hat{S}$ を得る．単純には $\hat{U}$，$\hat{V}$，$\hat{W}$ のテンソル積を計算することで生成グラフとして動的隣接テンソル $\hat{A}$ が得られるが，この操作は時間的にも空間的にも計算量が大きい．特にグラフ生成のユースケースにおいては，小規模グラフを入力としてグラフの特徴を抽出し，出力時には大規模なグラフにスケールして出力することが多い [25]．そのため，計算の重い動的隣接テンソルを経由せずに生成グラフのクオリティを維持しながら高スケールにエッジを生成する方法として Scalable Stochastic Edge Generation (SSEG) を提案する．SSEG は，テンソル積を経由せずに Source node の潜在因子行列 U,Target node の潜在因子行列 V,Timestep の潜在因子行列 W 各々からエッジの構成要素を確率的に選択することで，エッジを生成する．さらに，SSEG は以下の定理により生成されるグラフの高い品質を保証する．

Algorithm 1 Scalable Stochastic Edge Generation(SSEG)

Input: $\hat{U},\ \hat{V},\ \hat{W}, n, e, t, k$
Output: $\hat{S}$

1: ### Initialize valuables ###
2: $\hat{S} = [\,]$
3: $prop = \text{zeros}(k)$
4: ### Calculate the probability of which base class to select ###
5: **for** $i = 0$ to $k-1$ **do**
6: $\quad prop[i] = Sum(\hat{U}[:,i]) * Sum(\hat{V}[:,i]) * Sum(\hat{W}[:,i])$
7: **end for**
8: ### Generate edges ###
9: **for** $i = 0$ to $e-1$ **do**
10: $\quad cls = Rand_{int}(\text{range} = (0, m-1), \text{weight} = prop)$
11: $\quad u_{id} = Rand_{int}(\text{range} = (0, n_n - 1), \text{weight} = \hat{U}[:, cls])$
12: $\quad v_{id} = Rand_{int}(\text{range} = (0, n_n - 1), \text{weight} = \hat{V}[:, cls])$
13: $\quad w_{id} = Rand_{int}(\text{range} = (0, n_t - 1), \text{weight} = \hat{W}[:, cls])$
14: $\quad \hat{S}.append([u_{id}, v_{id}, w_{id}])$
15: **end for**
16: **return** $\hat{S}$

定理 1. *SSEG* は、潜在因子行列のテンソル積によって計算される動的隣接テンソルと等価なエッジシーケンスを生成する．

この定理の証明は後に自身の Github にて公開する。SSEG の概要を図 2，アルゴリズムを Algorithm 1 に示す．

SSEG のアルゴリズムは 2 つのステップで構成される．ステップ 1 では，U,V,W に基づき，どの基底クラスを元にエッジ生成するかを決定する (line 10)．潜在次元 m，つまり基底クラス m が選択される事象を l_m とすると，$P(l_m)$ は以下のように表される．

$$P(l_m) = \frac{\left(\sum_{i=1}^{n} U_{im}\right)\left(\sum_{j=1}^{n} V_{jm}\right)\left(\sum_{\tau=1}^{t} W_{\tau m}\right)}{\sum_{m=1}^{k}\left(\sum_{i=1}^{n} U_{im}\right)\left(\sum_{j=1}^{n} V_{jm}\right)\left(\sum_{\tau=1}^{t} W_{\tau m}\right)} \tag{9}$$

この確率の式は，特定の基底クラスベクトルの要素値の全体に占める割合により算出される．ステップ 2 では，選択された基底クラスの基底クラスベクトルからエッジの構成要素を確率的に選択・組合せてエッジを構成してから，生成グラフに追加する (lines 11-14)．例えば，基底クラス l が選択された場合は，$U_{\cdot l}$,$V_{\cdot l}$, $W_{\cdot l}$ の各要素値を正規化する得られる確率分布により，それぞれ Source node, Target node , Timestep を選択する．

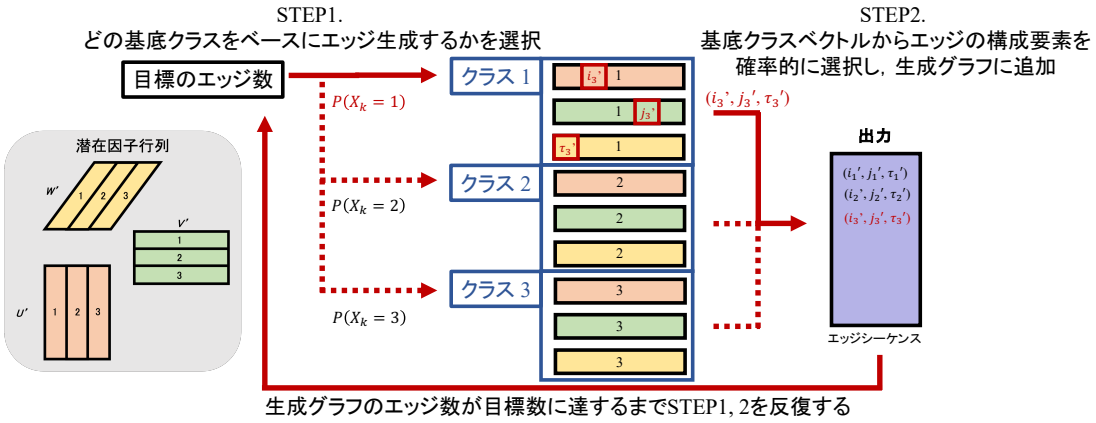


図 2: SSEG の概要図．まず，どの基底クラスからエッジを生成するかを確率的に選択する （STEP1）．次に，選択された基底クラスの基底クラスベクトルからエッジの構成要素を確率的に選択し，出力のエッジシーケンスに追加する （STEP2）．生成されるエッジの数が目標サイズに達するまで STEP1 と 2 を繰り返し，最終的な生成グラフを得る．

こうすることで，動的隣接テンソルを算出せず，3 つのベクトルからそれぞれ 1 要素を選択するだけでにエッジの構成要素を決定することができる．生成グラフのエッジ数が目標数に達するまで ステップ 1, 2 を反復することにより，最終的な生成グラフを獲得する (lines 9-15).

4.3 計 算 量

この章では，ConTeG の計算量について説明する．実際のグラフ生成のユースケースにおいては，小規模グラフを入力としてグラフの特徴を抽出し，出力時には大規模なグラフにスケールして出力することが多い [25]．そのため，ノードサイズとエッジサイズについて，入力グラフは n，e，出力グラフは n'，e' とそれぞれ区別する．さらに，出力グラフは入力グラフに比べて十分大きく，$n \ll n'$ と $e \ll e'$ を満たすと仮定する．時間的計算量についてはグラフ特徴の抽出では NTF 操作が最も大きく，$O(n^2t)$ である．SSEG の時間的計算量は， line 9-15 の for ループの計算量が最も大きく， line 11,12 ではノードサイズ n と Timestep サイズ t の $Rand_{int}$ 計算をエッジサイズ e の回数ループするので，$O((2n'+t)e')$ である．したがって，全体の時間的計算量は $O(n^2t+(2n'+t)e')$ である．空間的計算量については入出力グラフのサイズがボトルネックである．入力グラフの動的隣接テンソルは $O(n^2t)$ であり，出力グラフのエッジシーケンス $O(e')$ であるため，全体の空間的計算量は $O(n^2t+e')$ である．$n \ll n'$ 及び $e \ll e'$ を考慮すると，時間的計算量は $O((2n'+t)e')$ に，空間的計算量は $O(e')$ に簡略化される．

5 実験（Experiments）

ConTeG のグラフ生成能力と効率性に関する評価実験と，ConTeG により生成した多様なグラフを用いた時間グラフニューラルネットワークのベンチマーク実験の 2 つの実験を実施した．

データセット名	ノードサイズ	エッジサイズ	Timestep サイズ
Enron	184	125235	1317
Social Evo.	74	2099519	243

表 1: データセットの詳細．ノード数とエッジ数，Timestep サイズの値をそれぞれ示す．

5.1 ConTeG の評価実験

実グラフデータセットを用いて ConTeG のグラフ生成能力と効率性について評価する．グラフの生成能力については ConTeG の重要な特徴であるグラフ特徴の制御が正しく機能しているかどうかを検証する．具体的には，それらグラフの次数分布やイベント分布を可視化することにより定性分析を実施する．効率性については，出力グラフのノードサイズに対して ConTeG のグラフ生成にかかる時間がスケーラブルであるかどうかを検証する．

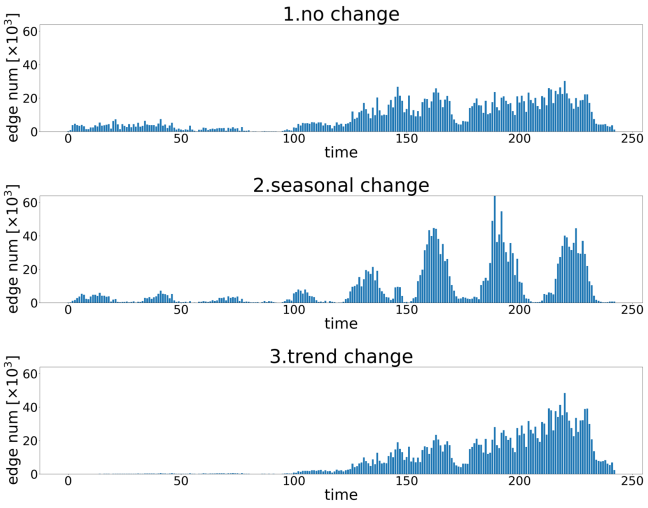
5.1.1 実 験 設 定

データセット．本実験で用いるデータセットの詳細を表 1 に示す．Enron は E-mail ネットワーク， Social Evo. はある寮における学生の近接ネットワークである．NTF を実行する上で入力を動的隣接テンソルの形式にする必要があるため， 1 day を最小粒度として DTDG に変換してる．

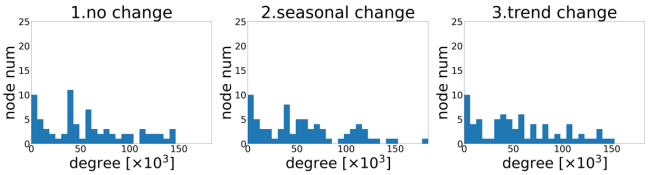
パラメータ．提案手法 ConTeG のパラメータは，NTF の潜在次元数が $k = 10$，STL の移動平均のウィンドウサイズについては各データセットごとの Timestep サイズを基準に決定しており， Social Evo. では 90 day ， Enron では 300 day とする．

5.1.2 グラフ生成能力

本実験では，グラフ特徴を変更したグラフと変更していないグラフを準備しそれぞれのグラフ指標を可視化することにより，特徴制御の機能に関する定性分析を実施する．分析対象として，3 つの異なる生成グラフ *1.no change* と *2.seasonal change*， *3.trend change* を用意する．*1.no change* はグラフ特徴の変更を行わずに生成したグラフである．*2.seasonal change*



(a) CSS の可視化の結果．横軸はエッジの生成時刻を，縦軸はその時刻に発生したエッジ数を示している．



(b) 次数分布の可視化の結果．横軸は次数を，縦軸はその次数のノードの数を示している．

図 3: 特徴制御の機能に関する実験結果．変更なしの *1.no change* と 1 month 周期を強調した *2.seasonal change*，二次関数の上昇傾向を加えた *3.trend change* についての CSS 分布と次数分布をそれぞれ示している．

は 1 month の周期に最も近い 5 個のスペクトルを強調するように Seasonal を変更し生成したグラフである．*3.trend change* は 二次関数的な上昇傾向を追加するように Trend を変更し生成したグラフである．このとき入力グラフとしては， Social Evo. データセットを使用した．グラフ指標には，時間的特徴を表すエッジの時間分布 （CSS） と構造的特徴を表す次数分布を用いた．

グラフ特徴の制御機能に関する定性分析の実験結果を図 3 に示す．まず CSS 分布の結果について検討する．結果の図 3a の横軸はエッジの生成時刻を，縦軸はその時刻に発生したエッジ数を示している．*1.no change* と比較して，*2.seasonal change* の CSS では 30 days 周期のピークが観察された．さらに，*3.trend change* においては時刻 100 から 250 day の期間で上昇傾向が確認された．したがって，時間的特徴の変更が生成グラフに正確に反映されていることが確認できた．次に次数分布の結果について検討する．結果の図 3b の横軸は次数を，縦軸はその次数のノードの数を示している．*2.seasonal change* と *3.trend change* の次数分布は，どちらも *1.no change* と類似した．このことから，変更されていない特徴に関しては，入力グラフの分布をある程度保持していることが確認できた．

5.1.3 効 率 性

出力グラフのノードサイズを増加させた際の実行時間を計測することで，効率性に関する定性分析を実施する．入力グラフとしては， Enron データセットを使用する．生成グラフのノードサイズは $10, 100, 1000, 10000$ とする．平均次数は入力グラフの値を元に 680 で固定しており，生成グラフのエッジサイズはそれぞれ $6800, 68000, 680000, 6800000$ である．実行時間はグラフ特徴の抽出と変更に要した時間は含まず，グラフ生成にかかった時間のみとする．

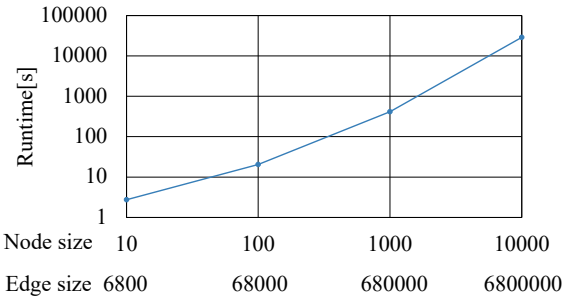


図 4: 効率性に関する実験結果．横軸はノードサイズとエッジサイズを表し，縦軸はグラフ生成にかかった時間を秒単位で表している．

効率性に関する実験結果を図 4 に示す．横軸はノードサイズとエッジサイズを表し，縦軸はグラフ生成にかかった時間を秒を単位として表している．結果としては，ノードサイズとエッジサイズに対して，ノード数×エッジ数の二次関数的な実行時間が得られた．これは，SSEG の計算量 $O((2n' + t')e')$ に対して，妥当な結果が得られたと考えられる．なお，ノードサイズが 10000，エッジサイズが 6800000 の最も大きなグラフの生成にかかった時間は，28892 秒，つまり約 8 時間であった．このことから，ある程度大規模なグラフでも現実的な時間でグラフ生成が可能であると考えられる．

5.2 ベンチマーク実験

5.2.1 実 験 目 的

ConTeG により特徴が制御された様々なグラフを，時間グラフニューラルネットワークに適用することで，精度と Timeste こうりつ p の両方の観点で各手法のメリット・デメリットを明らかにする．具体的には，時間的特徴や構造的特徴を段階的に変化させたグラフを複数生成し各時間グラフニューラルネットワークに適用することで，各手法がどのような特徴を持つグラフに対して有効な予測を行えるかを分析する．さらに，ノード数を段階的に変化させたグラフを複数生成し各時間グラフニューラルネットワークに適用することで，各手法の実行時間の検証も行う．

5.2.2 実 験 設 定

入力グラフ.

実行時間の都合により，エッジ数が比較的少ない Enron を ConTeG への入力グラフとして用いる．3 つの観点により Enron のグラフ特徴を段階的に変化させたグラフをそれぞれ複数生成する．時間的特徴に関しては，強調する周期，上昇傾向の関数型をそれぞれ段階的に変化させたグラフを生成する．強調する周期は，$1week$, $1month$, $3months$, $1year$ の 4 種類の異なる周期を用意する．上昇傾向の関数型については，*none*（上

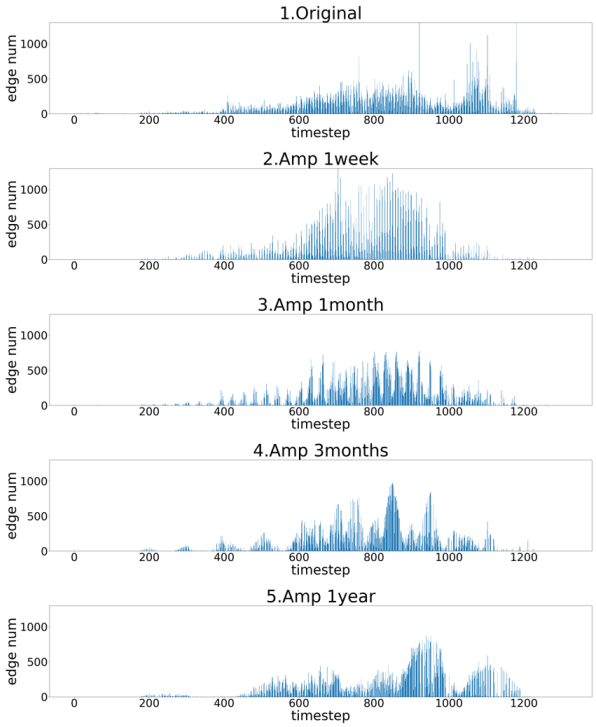


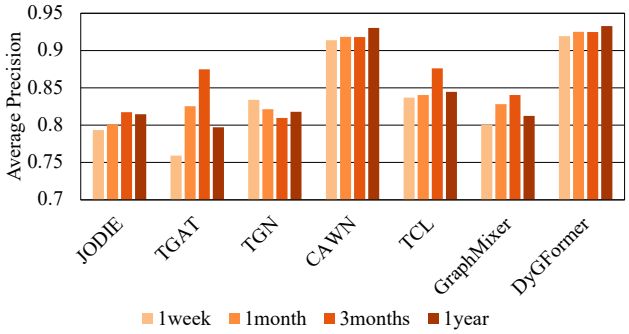
図 5: 異なる周期を強調して生成した 4 つのグラフと入力グラフの CSS 分布の可視化の結果．入力グラフは Social Evo. である．*1.Original* は入力グラフ，*2.Amp 1week* は *1week* 周期を，*3.Amp 1month* は *1month* 周期を，*4.Amp 3months* は *3months* 周期を，*5.Amp 1year* は *1year* 周期を強調した生成グラフである．

昇傾向の追加なし），*logarithm*, *linear*, *quadratic*, *cubic* の 3 種類の異なる関数型を用意する．構造的特徴に関しては，ノード数を $50, 100, 150, 200, 250, 300$ と段階的に変化させたグラフを生成する [1]．ただし平均次数は元グラフの値で固定しているため，エッジ数も線形的に増大する．生成グラフの例として，入力グラフと異なる周期を強調して生成した 4 つのグラフの CSS 分布の可視化を図 5 に示す．各々の生成グラフにおいて，強調周期の間隔で CSS 分布上においてピークが現れていることが確認できる．ノード数を変化させたグラフ群についてのみリンク予測タスクの実行時間を計測し効率性の結果を得る．それ以外のグラフ群についてはリンク予測タスクの精度を計測する．

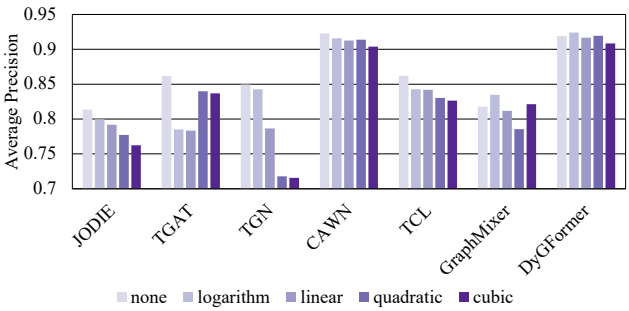
評価対象．評価対象として，時間グラフニューラルネットワークの統一ライブラリである DyGLib [13] を活用する．このライブラリに収録されている手法のうち， JODIE，TGAT，TGN，CAWN，TCL，GraphMixer，DyGFormer の 7 つの手法をベースラインとして用いる．

評価指標．タスクはリンク予測を用いる．精度の算出には Average Precision を用い，5 回実行した時の平均値を用いる．効率性の評価指標として，学習・推論にかかった時間 (total runtime) を用いる．ただし，学習時の max_epoch は 1 で固定する．

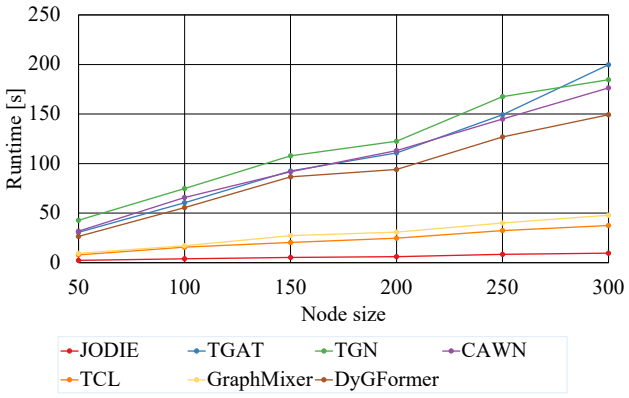
1：多くの GNN が実行できるようにするため，このような小規模なグラフを複数生成している．



(a) 異なる強調周期に対する精度の結果．縦軸は Average Presicion を示す．



(b) 異なる上昇傾向の関数型に対する精度．縦軸は Average Presicion を示す．



(c) 異なるノード数に対する実行時間．縦軸は Runtime を示す．

図 6: ベンチマーク実験の結果．

実行方法．実行環境として， NVIDIA A100 40 GB Memory (Driver Version: 515.43.04，CUDA Version: 11.7) を用い，高速な並列計算を実施する．

5.2.3 結果と考察

3 つの観点によりグラフ特徴を段階的に変化させたグラフを各リンク予測手法に対して適用した結果を以下に示す．

異なる強調周期に対する精度．強調する周期を変化させた際の実験結果を図 6a に示す．2 つの興味深い観察結果が得られた．第一に， CAWN と DyGFormer は一貫して精度で他を上回っている．第二に，ほとんどの手法は時系列分析タスクで典型的な，より長い期間においてより高い精度を達成している．これとは対照的に，TGN はより短い期間でより高い精度を達成しており，これはおそらくその記憶メカニズムが短期的な特徴を効果的に捉えているからだと考えられる．

異なる上昇傾向の関数型に対する精度．上昇傾向の関数型を

変化させた際の実験結果を図 6b に示す．ここでも CAWN と DyGFormer が一貫して他を上回っていることを確認した。また，修正なしや対数のような低次関数では，ほとんどの手法が高次関数よりも高い精度を示している．この結果は，機械学習モデルには高次関数ほど高い表現力が必要であることを示唆している．

異なるノード数に対する実行時間．ノード数を変化させた際の実験結果を図 6c に示す．全手法に共通する結果として，線形に増加するノード数 (エッジ数) に対して線形な実行時間が得られた．各手法を比較すると， JODIE が最速， TCL， GraphMixer も比較的実行時間は短く， TGAT， TGN， CAWN， DyGFormer は実行時間が長い．

以上から，精度を求めるなら CAWN か DyGFormer ，実行時間を短くするなら JODIE， 精度と実行時間の両方を求めるなら TCL を用いるのが良いと考えられる．

6 終わりに

時間グラフの時間的特徴をユーザが解釈可能な段階まで抽象化することにより，ユーザが生成グラフの特徴を制御することができる時間グラフ生成器 ConTeG を提案した．実験では，ConTeG による時間グラフの特徴の制御が出力グラフに正しく反映されていることが確認できた．さらに，ConTeG により段階的にグラフ特徴を変更したグラフを，7 つの時間グラフニューラルネットワーク手法に適用することにより，各手法の利点・欠点を明らかにした．

謝　　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです．

文　　献

[1] Huajie Shao, Shuochao Yao, Yiran Zhao, Chao Zhang, Jinda Han, Lance Kaplan, Lu Su, and Tarek Abdelzaher. A constrained maximum likelihood estimator for unguided social sensing. In *Proceedings of the IEEE*, pp. 2429–2437. IEEE, 2018.

[2] Xu Liu, Jingrui He, Sam Duddy, and Liz O'Sullivan. Convolution-consistent collective matrix completion. In *Proceedings of the CIKM*, pp. 2209–2212, 2019.

[3] Seiji Maekawa, Koki Noda, Yuya Sasaki, et al. Beyond real-world benchmark datasets: An empirical study of node classification with GNNs. *NeurIPS*, Vol. 35, pp. 5562–5574, 2022.

[4] Alessandra Stella, Peter Bouss, Günther Palm, and Sonja Grün. Comparing surrogates to evaluate precisely timed higher-order spike correlations. *Eneuro*, Vol. 9, No. 3, 2022.

[5] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 272, No. 1-2, pp. 173–187, 1999.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[7] Martin Simonovsky and Nikos Komodakis. GraphVAE: Towards generation of small graphs using variational autoencoders. In *Proceedings of the ICANN*, pp. 412–422. Springer, 2018.

[8] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. NetGAN: Generating graphs via random walks. In *Proceedings of the ICML*, pp. 610–619. PMLR, 2018.

[9] Giselle Zeno, Timothy La Fond, and Jennifer Neville. Dymond: Dynamic motif-nodes network generative model. In *Proceedings of the WWW*, pp. 718–729, 2021.

[10] Penghang Liu and Ahmet Erdem Sariyüce. Using motif transitions for temporal graph generation. In *Proceedings of the SIGKDD*, pp. 1501–1511, 2023.

[11] Dawei Zhou, Lecheng Zheng, Jiawei Han, and Jingrui He. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the KDD*, pp. 401–411, 2020.

[12] Shubham Gupta, Sahil Manchanda, Sayan Ranu, and Srikanta Bedathur. TIGGER: Scalable generative modelling for temporal interaction graphs. In *Proceedings of the AAAI*, 2022.

[13] Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. *NeurIPS*, 2023.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[15] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, Vol. 11, No. 2, 2010.

[16] Seiji Maekawa, Yuya Sasaki, George Fletcher, and Makoto Onizuka. GenCAT: Generating attributed graphs with controlled relationships between classes, attributes, and topology. *Information Systems*, Vol. 115, p. 102195, 2023.

[17] Hardeo Kumar Thakur, Anand Gupta, Anshul Garg, and Disha Garg. Periodic patterns in dynamic network: Mining and parametric analysis. In *Multidisciplinary Approaches to Service-Oriented Engineering*, pp. 244–264. IGI Global, 2018.

[18] Andry Alamsyah, Made Kevin Bratawisnu, and Puput Hari Sanjani. Finding pattern in dynamic network analysis. In *Proceedings of the ICoICT*, pp. 141–146. IEEE, 2018.

[19] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: Forecasting and control*. John Wiley & Sons, 2015.

[20] Kaijie Zhu, George Fletcher, Nikolay Yakovets, Odysseas Papapetrou, and Yuqing Wu. Scalable temporal clique enumeration. In *Proceedings of the SSTD*, pp. 120–129, 2019.

[21] Cleveland RB. STL: A seasonal-trend decomposition procedure based on loess. *J Off Stat*, Vol. 6, pp. 3–73, 1990.

[22] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, Vol. 74, No. 368, pp. 829–836, 1979.

[23] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the WSDM*, pp. 601–610, 2017.

[24] Tamir Hazan, Simon Polak, and Amnon Shashua. Sparse image coding using a 3d non-negative tensor factorization. In *Proceedings of the IEEE*, Vol. 1, pp. 50–57. IEEE, 2005.

[25] Sajad Darabi, Piotr Bigaj, Dawid Majchrowski, Artur Kasymov, Pawel Morkisz, and Alex Fit-Florea. A framework for large scale synthetic graph dataset generation. *arXiv*, 2022.

時間グラフ生成に関する包括的な調査とベンチマーキング

菊地　良将† 　中西　宏和† 　佐々木　勇和† 　NikolayYakovets†† 　GeorgeFletcher††
鬼塚　真†

† 大阪大学大学院情報科学研究科
†† Eindhoven University of Technology
E-mail: †{kikuchi.ryosuke,nakanishi.hirokazu,sasaki,onizuka}@ist.osaka-u.ac.jp,
††{n.yakovets,g.h.l.fletcher}@tue.nl

あらまし　時間グラフは，SNS や e-commerce サービスなど実世界の様々な事象を表現することができる．グラフ生成はベンチマーキングや非公開データの分析などに応用することができる．静的グラフにおける生成手法は長らく研究されてきた一方，時間グラフにおける生成手法は近年行われるようになったばかりであり，現在は分野全体での統一された知識・理解が不足している．そこで，本研究では時間グラフ生成手法の体系的な調査を通じて以下の 3 つ：1) 既存手法の包括的な調査・分類，2) 多様な実験による既存手法の性能分析，3) 既存手法の限界と時間グラフ生成分野の今後の展望．を提案する．性能分析においては既存研究では用いられていない temporal shortest path に関わる分布を使用した新たな評価方法を提案している．これにより，実世界における重要な情報である temporal shrortest path 情報がどれほど再現できているかを評価することが可能となる．我々の調査は時間グラフ生成の現状と可能性を明らかにし，時間グラフ生成分野のさらなる発展の促進を目的とする．本研究は，時間グラフ生成についての理解を深め，より効果的な応用に向けた道筋を示すものである．

キーワード　時間グラフ，グラフ生成

1　はじめに

グラフを用いることでソーシャルネットワーク [1,2]，学術論文の引用関係や e-commerce サービスなど [3] などの実世界の様々な事柄を表現することができる．グラフ生成はベンチマーキングや非公開データの分析などに応用することができる．そのほかにも，グラフ生成の過程で得られる情報を基に時間グラフの背後に隠れる核となる情報の分析も可能になる．

伝統的なグラフ生成では実データはある一定のルールに従っているとの仮定を基に発展してきた．代表的な例では Barabási-Albert モデル [4] が挙げられる．このモデルではノード次数に着目したグラフ生成を行っている．しかし，伝統的なグラフ生成ではコミュニティ構造といったグラフにおける複雑な関係性を捉えることができないという課題がある．そこで，近年の深層学習の発展に伴い， 複雑な関係性を捉えることを目的として RNN [5] などを用いた手法 [6] などが開発されてきた．近年では RNN 以外にもデータ生成のための深層学習モデルである GAN [7] を用いた手法 [8] の開発も進んでいる．

実際のグラフデータは時間情報を持つ時間グラフが多いのに対して，時間グラフでのグラフ生成はあまり盛んではなかった．そこで，時間グラフにおけるリンク予測手法 [9–12] などで得られた知見を基に，時間グラフ生成手法も提案されるようになってきた．代表的なリンク予測手法として CAW [11] が挙げられ，時間グラフにおける subgraph である Temporal motif を捉えることで高精度なリンク予測を実現している．そこで，時間グラフにおける最新のグラフ生成手法 [13] でも，Temporal motif の成長確率に着目することにより現実的なグラフ生成を行っている．しかし，時間グラフにおけるグラフ生成分野はいまだ黎明期であり，分野内での統一された知見を提供する研究は存在しない．そのため，同一分野の手法でも問題定義が様々であり，ある手法は別の手法の応用例には不向きであるといったことがある．

そこで，本研究では時間グラフにおけるグラフ生成手法分野における俯瞰的な知識を提供するために，包括的な調査を行った．調査を下に以下の 3 つを提案する．

- 既存手法の包括的な調査・整理 (3 章)
- 多様な実験による既存手法の性能分析 (4 章)
- 既存手法の限界と本分野の将来の展望 (5 章)

性能分析実験において，本稿では，時間グラフ生成手法を評価するための新しいフレームワークを示す．このフレームワーク内では，temporal shortest path 分布を使用する．これは，既存研究では用いられていない重要な指標である．temporal shortest path は，通信ネットワークにおけるパケットルーティング最適化や伝染病の拡散モデリングまで，様々な応用先において極めて重要である．その重要性にもかかわらず，temporal shortest path 分布は既存手法の評価として使用されていない．本稿では，2 章で 4 種類の temporal shortest path を定義し，4 章で既存手法の性能分析を行う．4 種類の temporal shortest path はそれぞれが時間グラフ上において異なる応用例が考えられる．Earliest-arrival Path は 2 ノード間で最も早い時刻

に到達する経路を表し，Latest-departure Path では 2 ノード間で最も遅い時刻に出発する経路を表す．Fastest Path は 2 ノード間を最短で到達する経路を表し，Least-wait Path では 2 ノード間を移動する際に経由ノードにおける待ち時間が短い経路を表す．続いて，既存手法の性能分析を行う．具体的には，すべての 2 ノード間での temporal shortest path を探索し， temporal shortest path に要する時間をもとに，既存手法の性能を評価する．

本研究では既存研究に使用されていない新たな評価観点を導入するだけでなく，時間グラフの動的性に関する新たな洞察を明らかにしている．本研究は既存研と現実データとのギャップを補完し，より現実的なグラフ生成を行ための可能性を提示する．

本論文の構成は以下のとおりである．2 章では，本研究で使用するための基礎知識や各種定義について述べる．次に，3 章では，既存手法の包括的な調査・整理について述べる．最後に，5 章では，既存手法の限界と本分野の将来の展望について述べる．

2 事前準備

2.1 時間グラフ

時間グラフはその時刻情報の扱い方により，離散時間グラフ (**DTDG**: Discrete Temporal Dynamic Graph) と連続時間グラフ (**CTDG**: Continuous Temporal Dynamic Graph) に分類される．DTDG は各時刻毎のグラフのスナップショットの集合として表現される．一方で，CTDG はタイムスタンプが付与されたエッジの集合として表現される．DTDG ではスナップショットの間に発生したエッジの順番が分からないため，CTDG のほうがより一般的な形式である．以下に本研究で使用する時間グラフの定義を示す．ここで，V はグラフ中に存在する全ノード集合，E はグラフ中に存在する全エッジ集合を表す．

Definition 1 (**離散時間グラフ**)**.** 離散時間グラフ DTDG は $G_{DTDG} = (G^{t_1}, G^{t_2}, \ldots, G^{t_n})$ として定義される．ここで，$G^{t_i} = (V^{t_i}, E^{t_i})$ は時刻 $t_i \in T$ におけるグラフのスナップショットであり，$V^{t_i} = (v_1^{t_i}, v_2^{t_i}, \ldots, v_n^{t_i})$ は時刻 t_i におけるグラフ中に存在する全ノード集合，$E^{t_i} = (e_1^{t_i}, e_2^{t_i}, \ldots, e_m^{t_i})$ は時刻 t_i におけるグラフ中に存在する全エッジ集合を表す．ただし，$e_i = (v_{src}, v_{tgt}, t_i)$ であり，$v_{src} \in V^{t_i}$ はエッジの始点，$v_{tgt} \in V^{t_i}$ はエッジの終点を表す．

Definition 2 (**連続時間グラフ**)**.** 連続時間グラフ CTDG は $G_{CTDG} = (V, E)$ として定義される．ここで，$e_i = (v_{src}^i, v_{tgt}^i, t_{start}^i, t_{end}^i) \in E$ は時刻 t_{start} から時刻 t_{end} まで存在するノード v_{src} からノード v_{tgt} へのエッジを表す．ただし，データセットには t_{end} を持たないものが存在する．その場合は $t_{end} = \infty$ とし，t_{start} から永遠に存在するエッジとして扱う．

2.2 時系列パス

ここでは，本研究における時系列パス (temporal path) の定義を示す．temporal path は時間グラフ中のエッジの集合である．本稿では，各手法を用いて生成したグラフについて性能分析を行う．その際に，temporal path を用いた分析をおこなうため，本節で定義を示す．ここで，temporal path 宙に含まれるエッジの個数を長さと呼び，L で表す．その際に，DTDG と CTDG でそれぞれ定義を示す．

Definition 3 (**DTDG における時系列パス**)**.** 離散時間グラフ $G_{DTDG} = (G^{t_1}, G^{t_2}, \ldots, G^{t_n})$ における時系列パス P_{DTDG} は $P_{DTDG} = (e_1, e_2, \ldots, e_L)$ として定義される．ただし，$e_i (i = 0, 1, \ldots L)$ について $t_i \leqq t_{i+1}$ が成り立つ．ここで，$e_i = (v_{src}, v_{tgt}, t_i)$ である．

続いて，CTDG における temporal path の定義を示す．その際，temporal path 内のエッジの時刻について制約の強い strongly temporal path と制約の緩い weakly temporal path の 2 種類を定義する．strongly temporal path は，各エッジの t_{start} が 1 つ前のエッジの時刻 t_{end} 以上でなければならない．一方で，weakly temporal path は，各エッジの t_{start} が 1 つ前のエッジの時刻 t_{start} 以上であればよい．これは，実データの特性によって temporal path も別々に定義することが妥当だと考えたからである．例えば，航空ネットワークを用いて，最終目的地までの経路を考える場合は strongly temporal path を用いることが妥当であると考えられる．なぜなら，経由地の空港に到着する前 (t_{end}) に経由地の空港から出発する (t_{start}) ことができないためである．逆に，リポストネットワークであれば，weakly temporal path を用いることが妥当であると考えられる．なぜなら，ユーザ A がユーザ B のツイートをリポストした後に，ユーザ B がユーザ C のツイートをリポストすることがあるためである．

以上を踏まえて以下のように 2 種類の temporal path を定義する．

Definition 4 (**CTDG における strongly temporal path**)**.** 連続時間グラフ $G_{CTDG} = (V, E)$ における strongly temporal path P_{str} は $P_{str} = (e_1, e_2, \ldots, e_L)$ として定義される．ただし，$e_i (i = 0, 1, \ldots L)$ について $t_{end}^i \leqq t_{start}^{i+1}$ を満たす．

Definition 5 (**CTDG における weakly temporal path**)**.** 連続時間グラフ $G_{CTDG} = (V, E)$ における weakly temporal path P_{weak} は $P_{weak} = (e_1, e_2, \ldots, e_L)$ として定義される．ただし，$e_i (i = 0, 1, \ldots L)$ について $t_{start}^i \leqq t_{start}^{i+1}$ を満たす．

Definition 6 (**temporal path の所要時間**)**.** 以上の temporal path の定義を基に，temporal path の開始時刻を $Start(P)$, 終了時刻を $End(P)$ として定義する．それらを用いて temporal path の所要時間を $Dura(P) = End(P) - Start(P)$ として定義する．また，temopral path の経由地における総待機時間を $Wait(P) = \sum_{i=1}^{L} (t_{start}^i - t_{end}^{i-1})$ として定義する．ただし，$Wait(P)$ は path 中のエッジが t_{end} を持つ場合にしか定義さ

れないので，特別な場合の weakly temporal path を除いて，strongly temporal path に対してのみ定義される．

以上の temporal path 定義をもとに時間グラフにおける最短経路 (temporal shortest path) を定義することができる．経路中の hop 数が最小である静的グラフにおける最短経路とはことなり，時間グラフにおいては複数の最短経路を定義することができる．本稿では各モデルにより生成されたグラフにおける最短経路に基づいた性能分析を行う．以下では文献 [14] を参考に，任意の 2 ノード間に以下の 4 種類の最短経路を定義する．

Definition 7. 時間グラフにおける最短経路

- Earliest-arrival Path: $Start(P)$ が最小の temporal path
- Latest-departure Path: $End(P)$ が最大の temporal path
- Fasterst Path: $Dura(P)$ が最小の temporal path
- Least-wait Path: $Wait(P)$ が最小の temporal path

ただし，2 ノード間に temporal path が存在しない場合は最短経路は定義されない．また，DTDG や wekly temporal path の場合は $Wait(P)$ が定義されないため，Least-wait Path は定義されない．

2.3 時間的中心性

ここでは，本研究における時間的中心性 (temporal centrality) の定義を示す．本稿では temporal centrality を用いて性能分析を行う．時間的中心性は時間グラフにおけるノードやエッジの重要性を示す指標であり，時間的中心性が高いノードやエッジは時間グラフにおいてハブといった重要な役割を果たすと考えられる．

Definition 8. temporal betweenness 時刻 t から t' が与えられたとき，ノードまたはエッジ e の temporal betweenness (時間的媒介中心性) は，t と t' の間に $Start(P), End(P)$ が存在する temporal path P における e の出現回数として定義される．すなわち，$betweenness(e,t,t') = |\{(P^*(v,u)) : e \in P^*(v,u), v,u \in V, v \neq u, Start(P) \leqq t, End(P) \geqq t'\}|$ である．ここで，$P^*(v,u)$ はノード v からノード u への temporal shortest path を表す．

Definition 9. temporal closenness 時間グラフにおけるノード v の temporal closenness (時間的近接中心性) は，ノード v から全てのノード u への temporal shortest path の時間的距離の和の逆数として定義される．すなわち，$closeness(v) = \frac{1}{\sum_{u \in V, u \neq v} Dura(P^*(v,u))}$ である．ここで，$P^*(v,u)$ はノード v からノード u への temporal shortest path を表す．ここで，ノード v からノード u への temporal shortest path が存在しない場合は，temopral closeness は定義されない．

3 時間グラフ生成手法の調査

本章では時間グラフ生成手法について調査した結果を図 1 に示す．本稿では時間グラフ生成手法を motif ベース手法，walk ベース手法，その他の手法の 3 つに分類した．グラフ生成において実データを入力に使用して，その特徴量利用する手法と実データを用いない手法に分け，実データを用いない手法をその他の手法に分類している．また，実データを入力に使用する手法においては motif を捉えるか walk を捉えるかによって分類している．

3.1 motif ベース手法

motif ベース手法では，時間グラフの発展の単位として temmporal motif を捉える．temporal motif とは，以下のように定義される時間グラフの部分グラフであり，時間グラフの発展過程で出現する．時間グラフにおけるリンク予測手法 [11, 16] でも，temporal motif に着目することで高精度なリンク予測が実現されており，時間グラフの重要な特徴であると考えられる．

Definition 10 (Temporal motif)**.** 時間グラフ G があるとき，大きさ l の temporal motif は $M_i^l = (V', E')$ は以下の条件を満たすサブグラフである．ただし，$|E'|$ は E' に含まれるエッジの数を表す．

- $V' \subset V, E' \subset E, |E'| = l$
- すべてのエッジはそのエッジ以前のノードのどれかのエッジと接続されている．つまり，$\{v_{src}^{j+1}, v_{tgt}^{j+1}\} \cap \{v_{src}^1, v_{tgt}^1, v_{src}^2, v_{tgt}^2, \ldots v_{src}^j, v_{tgt}^j\} \neq \emptyset$

motif ベース手法としては，STM [17], DYMOND [18], MTM [13] が挙げられる．

STM では入力データセット中の temporal motif の時間分布を抽出し，その分布を基にグラフを生成する．STM では最大 3 ノード間の temporal motif を考得ている．グラフ生成において既存の temporal motif にエッジを追加する際に，パラメータ α に基づいてエッジを追加するノードを決定する．α によって，Barabási-Albert グラフのような preferential attachment に基づくグラフ生成や power-law distribution に基づくグラフ生成が可能である．STM はこの点において生成グラフの特徴量を制御可能である点が他の motif ベース手法と異なる．

DYMOND では隣り合う時刻の離散時間グラフ間で temporal motif が変化する様子に一定のパターンがあるという観察を基にグラフ生成を行っている．DYMOND では 3 ノードの組合せについて 3 種類の temporal motif 間での遷移確率を学習し，その遷移確率を基にグラフ生成を行う．具体的には，3 ノードについてどの temporal motif に遷移するのかを決定し，その temporal motif 内の役割 (例えば hub など) を決定して，エッジを追加する．

MTM では DYMOND と同様に temporal motif 間での遷移確率を学習し，その遷移確率を基にグラフ生成を行う．MTM は STM や DYMOND と異なり，ユーザが指定するパラメータに応じて temoiral motif のサイズ (motif 中に含まれるエッジ数) を変更できる点が特徴である．

3.2 WALK ベース手法

WALK ベース手法では，元グラフから temporal walk をサ

表 1: 時間グラフ生成手法の分類．既存手法を motif ベース，walk ベース，その他の 3 つに分類した．時間情報が連続値であるグラフを扱える場合には連続時間の使用列に○をつけている．TagGen においては文献 [15] において最大でも 200 種類のタイムスタンプしか扱えないことが明らかにされているため，連続*としている．walk やクラスタ情報などの大域情報を用いる手法については大域的特徴列に○をつけている．エッジが発生するノード選択に関わるパラメータなどにより，ユーザが生成グラフの特徴量を制御可能な手法にはユーザ制御列に○をつけている．グラフ生成において，例えば元データにない時間情報についても出力できたり，元グラフと生成グラフのノード ID に 1 対 1 対応の関係が為されていない手法についてはプライバシ保護の列に○をつけている．

分類	手法	実データの特徴利用	時間値の種類	大域的特徴	ユーザ制御	属性値生成	プライバシ保護
motif ベース	STM	○	連続		○		
	DYMOND	○	離散				
	MTM	○	連続				
Walk ベース	TagGen	○	連続*	○			
	TGGAN	○	連続	○			○
	TIGGER	○	連続	○			○
その他	CDM		連続		○		-
	RT-GEN++		離散	○	○	○	-

ンプリングし，その temporal walk を基にグラフ生成を行う．walk ベース手法の特徴として深層学習手法を用いることが挙げられ，元グラフからサンプリングされた temporal walk を基に生成器を学習し，その生成器を基にグラフ生成を行う．walk ベース手法には，TagGen [19], TGGAN [20], TIGGER [15] がある．

TagGen では主に 4 つのステップで時間グラフ生成を行う．初めに，時間グラフから temporal walk をサンプリングする．続いて，temporal walk を基に temporal walk の生成器を学習する．続いて，temporal walk の真偽を判断する識別機を学習し，temporal walk 生成器から生成された temporal walk が元データに存在しそうかどうか判断する．ここで，生成器，識別機には Transformer [21] を参考にして，2 層の self-attention 層を用いたモデルを使用している．最後に，真と判断された temporal walk を基に時間グラフを生成する．TagGen では temporal walk のサンプリングにおいて，CTDNE [22] という手法を用いている．

Tigger も TagGen と同様に，元データから temporal walk をサンプリングし，その temporal walk を基に生成器 (RNN) を学習し，生成器を基にグラフ生成を行う．Tigger では temporal walk のサンプリングにおいて，リンク予測手法である CAW [11] と同様にエッジの発生間隔を考慮したサンプリングを行っている．Tigger では W-GAN [23] を用いることで，元データの ノード embedding を参考にして，生成グラフにおけるノード embedding を生成することもできる．そのため，ノード数を変更することが可能であり，元データと生成グラフのノード ID には 1 対 1 対応の関係が為されていない．その点において Tigger はプライバシ保護の観点からも有用である．

TGGAN でも他の手法と同様に元データから学習された temporal walk 生成器を基にグラフ生成を行う．TGGAN では temporal walk の生成器に GAN を用いており，生成器と識別機を交互に学習することで，生成器が元データに近い temporal walk を生成するように学習する．temporal walk のサンプリングにおいては，これまでの 2 手法と異なり隣接していないノードへの移動も考慮している．これは，元データにおいて隣接ノードがなくなったノードへ遷移した場合においても temporal walk をサンプリングできるようにするためである．

3.3 その他の手法

その他の手法では，静的グラフにおける Barabási-Albert モデルのように，パラメータを入力としてグラフを生成する手法である．この手法には CDM [24]，RTGEN++ [25] が挙げられる．

CDM [24] では Concurrent Set Size (CSS) [26] の時間分布を基にグラフを生成する．CSS とは各時刻において有効なエッジ数のことである．CDM では入力データをもとに CSS の時間分布や power value distribution を抽出する．抽出した CSS の分布を基に生成グラフにおいてエッジの発生する時刻と本数を決定する．そして，power value distribution を基にエッジが発生するノードを決定する．

RT-GEN++ では 2 種類のモデルを用いてグラフを生成する．その際に，入力された次数分布やコミュニティ構造パラメータを基にグラフ生成を行う．1 つ目のモデルはエッジが発生するノードを決定するモデルである．このモデルでは度数分布のとコミュニティ構造の両方を捉え，エッジが発生するノードを決定する．2 つ目のモデルはエッジの追加を制御するモデルである．こちらのモデルを基に追加するエッジの数などを決定する．

4 実　　験

本章では 3 章で紹介した実データを用いる手法のうちいくつかに対して，実データを用いた性能分析を行う．具体的には以下の 3 種類の実験を行った．

- 既存手法の定量的な性能分析
- temporal shortest path 分布を用いた既存手法の性能分析実験
- 各手法がグラフ生成に要した時間の比較実験

これらの性能分析は，時間グラフ生成の応用を踏まえ以下の 3 つの観点を基に設計された．1 つ目は実データの再現性であ

る．時間グラフ生成の応用例の 1 つである非公開データの分析においては，入力となる実データの特性を再現できることが重要である．そこで，ノード数や 2 章で定義した temporal centrality などの定量的な指標や，本研究で新たに提案した temporal shortest path 分布を用いて，各手法の性能を評価する．2 つ目はプライバシ保護の観点である．非公開データ分析においては元のグラフの情報を秘匿することが求められる．そこで，各手法が生成したグラフがどれほど元データと共通の情報を持つかを評価する．3 つ目は実行時間である．実データにおいては，ネットワークの規模が大きくなることが多く，そのような大規模なネットワークに対しても効率的に時間グラフを生成できることが望ましい．そこで，各手法においてグラフ生成に要した時間を比較することで，各手法の実行時間について評価する．この観点を基に 3 つ目の実験を設計した．

4.1 実験設定

実験は linux サーバ上で行っており，サーバの CPU は Intel(R) Xeon(R) Gold 5220R CPU (2.20GHz) を 96 コア，メモリは 377GB である．また，TIGGER などの深層学習を用いる手法には，NVIDIA Tesla V100S-PCIE-32GB を用いた．

temporal centrality や temporal shortest path の計算は tglib [27] を用いて行った．tglib は時間グラフに対して様々な分析を行うためのライブラリであり，c++ をベースに実装しており，高速な計算が可能である．ただし，tglib 内では Latest-departure Path が実装されていないため，本稿では Latest-departure Path を用いた分析は行っていない．

4.1.1 データセット

我々は 2 つの実データを用いて性能分析を行った．1 つ目は EU_email* というデータセットである．これは，ヨーロッパの研究者間でのメーリングネットワークである，EU_email [28] を基に作成された．最大のノード数となるサブグラフを取り出し，そのサブグラフ中の中で最も次数が高い 80 ノードを取り出して作成された．2 つ目は CollegeMsg [28] というデータセットである．こちらは，カリフォルニア大学アーバイン校の学生間でのメッセージネットワークである．

どちらのデータセットも各エッジが (v_{src}, v_{tgt}, t) という形式で表されており，それぞれのエッジはノード v_{src} からノード v_{tgt} へと時間 t にメールやメッセージが送信されたことを示している．これらについて，表 2 に基づくネットワークの特徴を示す．ここで Edges とはデータセット中の静的エッジの数，つまり，(v_{src}, v_{tgt}) の組合せの数である．T-edges とはデータセット中のエッジの数であり，つまり，(v_{src}, v_{tgt}, t) の組合せの数である．mean IET とはデータセット中のエッジの時刻順に並べた時の，隣接するエッジ間の平均時間間隔である．

表 2: 実データの統計量

Name	#Nodes	#Edges	#T-Edges	Timespan (days)	mean IET (sec.)
Email-EU*	80	1184	43.0K	500	514.2
CollegeMsg	1.90K	20.3K	59.8K	193	273.1

本稿で性能分析を行った手法は (v_{src}, v_{tgt}, t) という形式しか入力データとして受け付けない．つまり，$(v_{src}, v_{tgt}, t_{start}, t_{end})$ という時刻幅を持つデータセットを扱うことができない．そのため，Least-wait Path を計算することができず，本稿では Earliest-arrival Path と Fastest Path の 2 つのみを用いて temporal shortest path 分布を計算した．

4.1.2 分析対象手法

本稿では，3 章で紹介した 4 つの手法について性能分析を行った．MTM[1] では $l_max = 2, \Delta\ t = 86400(= 1day)$ のパラメータの下グラフ生成を行った．DYMOND[2], tigger[3], TagGen[4] では各手法の実装で示されているデフォルトのパラメータの下グラフ生成を行った．また，MTM, Tigger では入力に時間が連続値である CTDG を入力している．一方，DYMOND と TagGen では入力に時間を 1 日を単位としてデータセットを離散化した DTDG を入力している．TagGen においては文献 [15] において指摘されている通り，入力グラフのタイムスタンプの種類が 200 程度までしか対応していないため本稿ではデータセットを DTDG に変換して入力している．TagGen においては，入力の際に全く同じ edge が存在するとエラーが発生するため，そのような edge は削除している．

4.2 既存手法の定量的な性能分析

本節では，4.1 節で述べたデータセットを用いて，各手法の性能を定量的に評価する．再現性と元データの秘匿性の 2 つの観点から各手法の性能を評価する．再現性については，各手法が生成した時間グラフの特徴が元データに近いかどうかを評価する．これは，時間グラフ生成の応用例の 1 つである非公開データの分析においては，入力となる実データの特性を再現できることが重要であるためである．元データの秘匿性については，各手法が生成したグラフがどれほど元データと共通の情報を持つかを評価する．本稿では，ノードとエッジの数，平均 temopral betweenness と平均 temporal closenness を用いて評価する．これは，非公開データの分析においては，入力となる非公開データの秘匿性を保つことが重要である．こちらは，元データと生成グラフとの間でのエッジと時間の重複率を用いて評価する．エッジの重複率では，元データ中に含まれる (v_{src}, v_{tgt}, t) のうち，生成グラフ中にも含まれるものの割合を示す．時間の重複率では，元データ中に含まれる時間のうち，生成グラフ中にも含まれるものの割合を示す．

実験の結果を表 3 に示す．MTM ではノード数，エッジ数，closenness については元データとほとんど同じ値である．一方，betweenness はどちらのデータセットにおいても減少している．エッジ数がほとんど同じ値なことを鑑みると，ハブとなる役割を持つノードが減少していると考えられる．これは，MTM がグラフ生成の際にエッジが発生するノードをランダムに選択するため，出力グラフがランダムグラフのようになっているためであると考えられる．このことから，エッジ発生の際にノード

1：https://github.com/erdemUB/KDD23-MTM
2：https://github.com/zeno129/DYMOND
3：https://github.com/data-iitd/tigger
4：https://github.com/davidchouzdw/TagGen

の次数分布などを基に非ランダムに選択することで，ランダムグラフ性が解消され，さらに再現性の高いグラフ生成が可能であると考えられる．また，エッジの重複率はどちらのデータセットについても 1% 以下であるが，時間の重複率では 50% 以上である．これは，MTM がグラフ生成の際に元データの時刻情報を一部そのまま利用しているためであると考えられる．よって，MTM では元データの時刻情報に関しては完全に秘匿されていないことが分かり，非公開データの分析に MTM を用いる際には注意が必要であると考えられる．

DYMOND ではノード数が元グラフとほとんど同じ値となっているが，エッジ数についてはどちらのデータセットにおいても 40% 以上減少している．betweenness と closenness についてもどちらも 10 分の 1 以下に減少している．基本的な情報であるエッジ数の再現に失敗しているため，DYMOND は元データのグラフの発展傾向の再現に失敗していると考えられる．一方，エッジの重複率ではどちらのデータセットにおいても 1% 以下であり，元データの秘匿性を保てていることを示す．

TagGen ではノード，エッジ数，betweenness，closenness のすべてについて元データとほとんど同じ値になっている．これは，TagGen では元データのグラフの発展傾向を高精度で再現できていることを示している．一方，エッジの重複率はどちらのデータセットにおいても 60% 以上である．これは，TagGen が元データのグラフの発展傾向を高精度で再現できている一方で，元データの秘匿性を保つことができていないことを示している．よって，TagGen を用いて非公開データの分析を行う際には，MTM と同様に注意が必要であると考えられる．

以上の結果より，再現性と秘匿性についてどちらも高い精度を満たす手法は存在しないことが分かった．今後は，秘匿性を保ったままで再現性を高められる手法の開発が求められると考えられる．

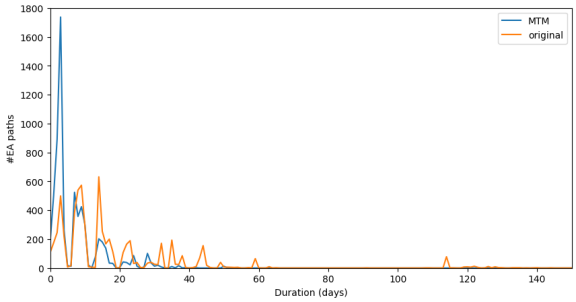
4.3 temporal shortest path 分布を用いた既存手法の性能分析実験

本節では，4.1 節で述べたデータセットを用いて，各手法の性能を temporal shortest path 分布を用いて評価する．temporal shortest path は，通信ネットワークにおけるパケットルーティング最適化や伝染病の拡散モデリングまで，様々な応用先において極めて重要である．その重要性にも関わらず，既存研究では用いられていない．そこで，本稿では temporal shortest path を再現できているか評価するための新しいフレームワークを使用する．具体的には，生成したグラフにおいてすべての 2 ノード間で temporal shortest path を探索し，所要時間をもとに評価を行う．紙面の都合上，Earliest-arrival Path を temporal shortest path として使用した場合の EU_email* についての結果を図 1 に示す．

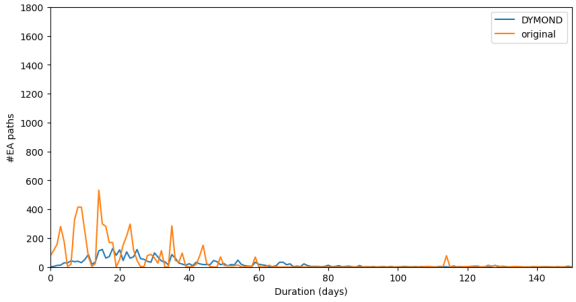
MTM で生成したグラフは元グラフと比較して，所要時間が短い temporal shortest path が多く存在していることが分かる．これは，4.2 節で述べたように，MTM がランダムグラフのようになっているためであると考えられる．ランダムグラフでは，元グラフにおける高・低次数のノードの次数が平滑化されるため，各ノード間の距離が近くなる．そのため，ノード間の距離が近くなったために，所要時間が短い temporal shortest path が多く存在していると考えられる．

続いて，DYMOND で生成したグラフでは元グラフと比較して temporal shortest path の本数自体が減少していることが分かる．これは，表 3 に示したように，DYMOND ではエッジ数が元データと比較して 40% 以上減少している．そのため，元データにおける temporal shortest path の本数も減少していると考えられる．

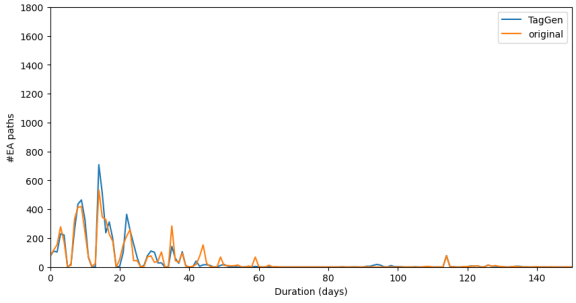
最後に，TagGen で生成したグラフは他の 2 手法と比較して，元グラフの分布を高い精度で再現できていることが分かる．これは，表 3 に示したように，TagGen では元データのグラフの発展傾向を高精度で再現できているためであると考えられる．また，MTM, DYMOND と異なり大域情報を保持しているため元データの temporal shortest path 分布を高い精度で再現できていると考えられる．



(a) MTM



(b) DYMOND



(c) TagGen

図 1: 各手法で生成したグラフにおける temporal shortest path 分布．青線が Earliest-arrival Path を temporal shortest path として使用した場合の EU_email* についての結果を示す．横軸は各 temporal shortest path (days) の所要時間であり，縦軸はその時間を持つ temporal shortest path の数である．

表 3: 元データと生成グラフにおける各種数値．diff とは生成グラフがどれほど元データか変化したかを表す．数値が低いほど生成グラフが元データに近い統計量を持っていることを示す．overlap 生成グラフが元データとどれほど共通の情報を持っているかを表す．数値が高いほど生成グラフが元データと共通の情報を持っていることを示す．- と表記している箇所は誤作動が生じているため値が得られていないことを示す．

Category		統計情報		大域情報		プライバシ保護	
dataset_name	**data**	**diff#Nodes**	**diff#edges**	**diff_betw(%)**	**diff_clos(%)**	**E_overlap(%)**	**T_overlap(%)**
EU_email*	MTM	0	11.86	16.04	90.21	1.39	66.87
	Tigger	-	-	-	-	-	-
	DYMOND	0	72.15	91.76	100.00	76.68	95.07
	TagGen	0	0.13	4.75	0.00	60.74	100.00
CollegeMsg	MTM	4.58	3.87	15.48	0.00	0.00	53.87
	Tigger	-	-	-	-	-	-
	DYMOND	0.32	42.66	93.98	95.58	5.01	99.93
	TagGen	3.48	0.10	7.62	18.22	65.63	100.00

4.4 各手法がグラフ生成に要した時間の比較実験

本節では，各手法がグラフ生成に要した時間を比較する．これは，実データにおいては，ネットワークの規模が大きくなることが多く，そのような大規模なネットワークに対しても効率的に時間グラフを生成できることが望ましいためである．実験の結果を表 4 に示す．

表 4 に示した結果より，MTM は数秒でグラフ生成を行うことができる一方で，DYMOND など他の手法では数十秒から数万秒要する．これは MTM のみが c++ で実装されていることや，計算量が $O(|E|)$ に近似されるため [13]，$O(N^3T)$ であるDYMOND よりも高速に動作しているためと考えられる

表 4: 各手法がグラフ生成に要した時間 (秒)．#T-edge はデータセット中のエッジの数を示す．- と表記しているのは実験中に実行時間がうまく取れなかったため現在実行中である．

Data	EU_email*	CollegeMsg
#T-edge	43.0k	59.8k
MTM	0.988	2.852
DYMON	50	47862
Tigger	-	-
TagGen	6391	-

5 既存手法の限界と時間グラフ生成分野の今後の展望

本章ではこれまでの得られた知見を踏まえて既存手法の課題を整理し，本分野の展望について述べる．

3 章の既存手法調査を通して以下のような限界が分かった．

- Inductive な手法が少ない
- 属性情報を扱える手法が少ない

Inductive なグラフ生成とは入力データセットからノード ID やエッジの発生時刻などを直接利用しない手法である．Inductive なグラフ生成により，知識転移やノードやエッジ数のアップ/ダウンスケーリングが可能になったり，入力データセットの秘匿性が向上といった利点がある．しかし，TIGGER や データセットを入力に使用しない RTGEN++ を除く多くの手法では入力データセットの情報を直接生成グラフに利用している．そのため，それらの手法では Inductive なグラフ生成は行えない．Inductive なグラフ生成に向けて，分布や深層学習モデルなどを実データから抽出することで実現できると考えられる．実際，TIGGER では深層学習モデルを使用して，ノードやエッジを embedding として扱うことにより，Inductive なグラフ生成を実現している．また，静的グラフにおける Barabási-Albert モデルのように入力パラメータのみをもとに生成を行う手法は RTGEN++ しか提案されていない．実データの観察を通して，時間グラフにおいて根底をなすパラメータを特定することにより実データを用いない手法の開発が可能であると考えられる．

実際の時間グラフデータにはノードやエッジに属性情報が付与されていることが多い．例えば，SNS であれば時間の経過によるユーザの所属機関の変化が考えられる．実際の状況を鑑みて，属性情報を扱うことができる手法を開発することでより現実に即した時間グラフが生成されると考えられる．しかし，属性情報を扱うモデルは現状では RTGEN++ に限られ，RTGEN++ においても属性情報はランダムに割り振るのみである．これでは，実世界の時間グラフデータにおける属性情報と構造・時間情報との関りを考慮できていない．今後の手法では，属性情報と構造・時間情報との関りを考慮したモデルの開発が求められる．

ここから，上記の課題と異なり時間グラフ生成分野における展望について以下に示す．

- 大規模言語モデルを用いた時間グラフ生成

現在様々な分野において大規模言語モデルが活用され，各分野において高精度な結果が報告されている [29, 30]．グラフ分野においても同様に LLM の活用が進んでいる [31]．時間グラフ生成分野においても RNN といった，自然言語処理分野で発達してきた深層学習モデルが活用されてきた．このような経緯を顧みると，時間グラフ生成分野においても大規模言語モデルの活用が進むと考えられる．

続いて，4 章の既存手法の性能分析を通して以下のような限界が分かった．

- 再現性と匿名性をどちらも高い精度で実現できていない
- temporal shortest path 分布を再現できていない手法が存

在する

6 まとめ

本研究では時間グラフ生成手法について体系的な調査を通じて，既存手法の分類，性能分析，分野の展望を提示した．既存手法の性に追う分析において，実世界において重要である temporal shortest path に関わる新たな評価方法を提案した．新たな評価方法により，時間グラフの動的性に関する新たな洞察を示した．

謝　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです。

文　献

[1] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of NeurIPS*, 2017.

[2] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

[3] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD*, pp. 135–144, 2017.

[4] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[5] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization, 2014. cite arxiv:1409.2329.

[6] Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models, 2018.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[8] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. NetGAN: Generating graphs via random walks. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 610–619. PMLR, 2018.

[9] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *Proceedings of ICLR*, 2019.

[10] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of KDD*, 2019.

[11] Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations*, 2021.

[12] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *Proceeding of the ICML*, 2020.

[13] Penghang Liu and Ahmet Erdem Sariyüce. Using motif transitions for temporal graph generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, p. 1501–1511, New York, NY, USA, 2023. Association for Computing Machinery.

[14] Huanhuan Wu, James Cheng, Yiping Ke, Silu Huang, Yuzhen Huang, and Wu Hejun. Efficient algorithms for temporal path computation. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, pp. 1–1, 11 2016.

[15] Shubham Gupta, Sahil Manchanda, Sayan Ranu, and Srikanta Bedathur. Tigger: Scalable generative modelling for temporal interaction graphs (to appear). In *Proc. of the 36th AAAI*, 2022.

[16] Ming Jin, Yuan-Fang Li, and Shirui Pan. Neural temporal walks: Motif-aware representation learning on continuous-time dynamic graphs. In *Advances in Neural Information Processing Systems*, 2022.

[17] Sumit Purohit. Temporal graph generation based on a distribution of temporal motifs. 2018.

[18] Giselle Zeno, Timothy La Fond, and Jennifer Neville. Dymond: Dynamic motif-nodes network generative model. In *Proceedings of WWW, 2021*, p. 718–729, New York, USA, 2021. Association for Computing Machinery.

[19] Dawei Zhou, Lecheng Zheng, Jiawei Han, and Jingrui He. A data-driven graph generative model for temporal interaction networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, p. 401–411, New York, NY, USA, 2020. Association for Computing Machinery.

[20] Liming Zhang, Liang Zhao, Shan Qin, and Dieter Pfoser. Tg-gan: Deep generative models for continuously-time temporal graph generation. *arXiv preprint arXiv:2005.08323*, 2020.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, 2017.

[22] Giang Hoang Nguyen, John Boaz Lee, Ryan A. Rossi, Nesreen Ahmed, Eunyee Koh, and Sungchul Kim. Continuous-time dynamic network embeddings. *In Proceedings of WWW*, 2018.

[23] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[24] Kaijie Zhu, G. Fletcher, and Nikolay Yakovets. Competition-driven modeling of temporal networks. *EPJ Data Science*, Vol. 10, pp. 1–24, 2021.

[25] Maria Massri, Zoltan Miklos, Philippe Raipin, Pierre Meye, Amaury Bouchra Pilet, and Thomas Hassan. Rtgen++: A relative temporal graph generator. *Future Generation Computer Systems*, Vol. 146, pp. 139–155, 2023.

[26] Kaijie Zhu, George Fletcher, Nikolay Yakovets, Odysseas Papapetrou, and Yuqing Wu. Scalable temporal clique enumeration. ACM International Conference Proceeding Series, pp. 120–129. Association for Computing Machinery, Inc, August 2019.

[27] Lutz Oettershagen and Petra Mutzel. Tglib: an open-source library for temporal graph analysis. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1240–1245. IEEE, 2022.

[28] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters, 2007.

[29] Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks, 2023.

[30] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Jellyfish: A large language model for data preprocessing, 2023.

[31] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. Large language models on graphs: A comprehensive survey, 2023.

ExhaustiveTree: 網羅的な二次元座標木を用いた表の表現

林　知司†　宮森　恒†

† 京都産業大学大学院 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{i2386111,miya}@cc.kyoto-su.ac.jp

あらまし　表形式のデータは，テキストとともに様々な文書中で一般的に使用されるデータ形式であり，多くのアプリケーションにとって重要で不可欠な要素である．しかし，表形式のデータは様々な構造で記述され，その内容理解やテキストとの関連付けは依然として容易ではない．従来研究では，表に含まれる階層的な情報を二次元座標木と呼ばれるツリーベースの構造で表現し，行と列の各インデックスと併用した埋め込みを用いることで，5 つのデータセットで最先端の結果を達成した．しかし，この手法では，表の上 (左) から下 (右) に向かって結合セルの大きさが徐々に小さくなることを前提としており，表内に階層構造がある場合，必ずしも座標を適切に割り当てることができない．本稿では，表中の結合セルの大きさを網羅的に把握した上で，その大きさの降順に二次元座標木を構築する手法を提案する．実験では，表のセルタイプ分類と，表とテキストとの関連付けの 2 種類のタスクにおいて，従来研究と性能を比較することにより，提案手法の有用性を検証する．

キーワード　表データ，表現学習，階層構造，座標木，セルタイプ分類

1　はじめに

表形式のデータは，Web ページやスプレッドシートなど様々な形式が存在し，データを管理するためには，不可欠な要素である．また，情報の根拠を明確にする目的で，テキストと併せて用いられることも多い．しかし，表形式のデータは様々な構造で記述され，その内容理解やテキストとの関連付けは依然として容易ではない．そのため，近年に渡っても表形式のデータを扱う研究が数多くなされている [13] [14] [15] [16] [17].

本稿では，The 17th NTCIR Conference Understanding of non-Financial Objects in Financial Reports [1](以下，UFO) タスクで提供されたデータセットと，deexcelerator(以下，DeEx) データセットを用いて検証する．UFO データセットは，HTML 形式であり，DeEx データセットは，スプレッドシート形式である．UFO タスクでは，有価証券報告書を対象に，そこに含まれる表形式データを理解する 2 つのサブタスクが設定された．Table Data Extraction(以下，TDE) サブタスクでは，与えられた有価証券報告書の表の各セルを，4 つのクラスに分類することを目的とし，Text-to-Table Relationship Extraction(以下，TTRE) サブタスクでは，有価証券報告書中の与えられたテキストに対して，それに関連する表中の該当するセルを選択することを目的としている．

図 1 は，有価証券報告書に含まれる表の例である．有価証券報告書に含まれる表の特徴として，表の右上に金額の単位が記載されるケースや，表の下に金額等の合計が結合セルの形で記載されるケースが多数見受けられた．我々は，有価証券報告書に含まれる表には，表中の結合セルやセル中のインデント等によって作られる複雑な構造が存在することに着目し，これをいくつかの方法で表の表現に取り入れる手法を開発した．

(Unit : One million yen)（単位：百万円）

ヘッジ会計の方法 Hedge accounting method	取引の種類 Type of transaction	主なヘッジ対象 Main hedged subject	契約額等 Contract amount and so on	うち 1 年超 In over one years	時価 Current market price
原則的処理方法 Processing method in general	商品スワップ取引 Swap transactions of Merchandise				
	受取変動・支払固定 Receive-floating・Pay-fixed 原油 Crude oil	営業未払金 Non-operating trade payables	69,132	27,452	△24,304
	商品オプション取引 Option transactions of Merchandise				
	売建 Going short プット Put Option 原油 Crude oil	営業未払金 Non-operating trade payables	33,120	15,468	△7,229
	買建 Going long コール Call Option 原油 Crude oil	営業未払金 Non-operating trade payables	42,798	20,103	△1,717
合計 Sum			145,051	63,025	△33,250

図 1　有価証券報告書に含まれる表の例

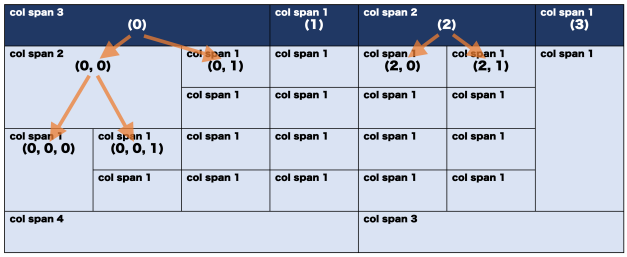


図 2　default vertical tree

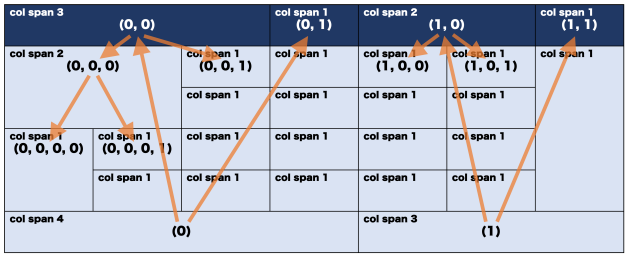


図 3　exhaustive vertical tree

2 関連研究

2.1 TaBERT

TaBERT [11] は自然言語と (半) 構造化テーブルの表現を共同で学習する事前学習済みモデルである．Wikipedia と WDC データセットによる 2,600 万のテーブルとその英語コンテキストからなる大規模なコーパスでトレーニングされている．困難な弱教師付きセマンティック解析ベンチマーク WikiTableQuestions で最先端の性能を達成した．提案手法とは，BERT のアーキテクチャを用いて学習した点において同じである．提案手法とは，表の階層構造を表現して学習した点において異なる．

2.2 TAPAS

TAPAS [11] は論理形式を生成せずにテーブル上で質問に応答する事前学習済みモデルである．BERT のアーキテクチャを拡張してテーブルを入力としてエンコードし，テキストセグメントと Wikipedia のテーブルの効果的な事前トレーニングから初期化し，エンドツーエンドでトレーニングされている．SQA では最先端の性能を達成し，WIKISQL では最先端の精能と同等のパフォーマンスを発揮した．提案手法とは，BERT のアーキテクチャを用いて学習した点において同じである．提案手法とは，表の階層構造を表現して学習した点において異なる．

3 データセット

実装するシステムの構築指針を明らかにするため，UFO タスクで提供された HTML 形式の有価証券報告書集合となるデータコレクションの諸元を分析した．

3.1 TDE

図 4 に TDE タスクの概要図を示す．TDE データセットは，有価証券報告書に含まれる表の各セルを，Metadata，Header，Attribute，Data の 4 つのクラスに分類するためのデータセットである．例では，Metadata は (単位：百万円) と書かれた主に表の外側にあるセルが該当し，Header は第 20 期，第 21 期，第 22 期と書かれた主に見出しのセルが該当し，Attribute は流動資産，固定資産，繰延資産，資産合計と書かれたセルが該当する．それ以外の主に数値などが書かれたセルは Data に該当する．

表 1 TDE データセットの統計

	文書	表	セル
test	190	1,660	45,499
train	252	2,530	66,369
合計	442	4,190	111,868

表 1 に，TDE データセットの分布を示す．TDE データセットは，HTML 形式の有価証券報告書が 442 件あり，test データ，train データにそれぞれ 190 件，252 件に分割されている．また，表は test データ，train データにそれぞれ 1660 件，2530 件あり，セルはそれぞれ 45499 件，66369 件あった．

(単位：百万円)

	第20期	第21期	第22期
流動資産	34,323	98,961	57,158
固定資産	158,763	208,933	189,903
繰延資産	73	241	180
資産合計	193,159	308,135	247,241

Metadata

	Header	Header	Header
Attribute	Data	Data	Data
Attribute	Data	Data	Data
Attribute	Data	Data	Data
Attribute	Data	Data	Data

図 4 TDE タスクの概要図

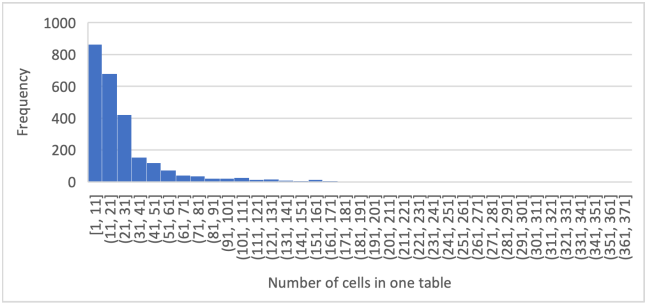


図 5 表に含まれるセル数のヒストグラム

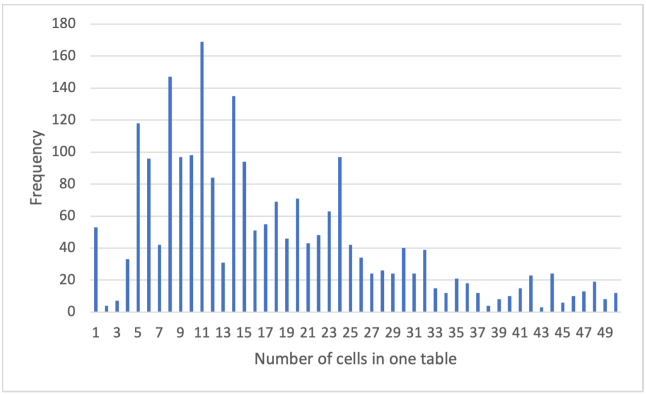


図 6 表に含まれるセル数の棒グラフ (50 個以下のみ)

図 5 に，1 つの表に含まれるセル数のヒストグラムを示す．図より，1 件の表に含まれるセル数はほとんどが 51 個以下であることがわかった．図 6 は，セル数が 50 個以下の表の頻度を

表すグラフである．図より，11 個のセルで構成された表が最も多いことがわかった．

3.2 TTRE

図 7 に TTRE タスクの概要図を示す．TTRE データセットは，有価証券報告書の与えられたテキストに対して，それに関連する表中の Name，Value に該当するセルを選択するためのデータセットである．Name は，所与のテキストが直接的に示しているセルであり，Value は，所与のテキストが間接的に示している数値などが該当する．

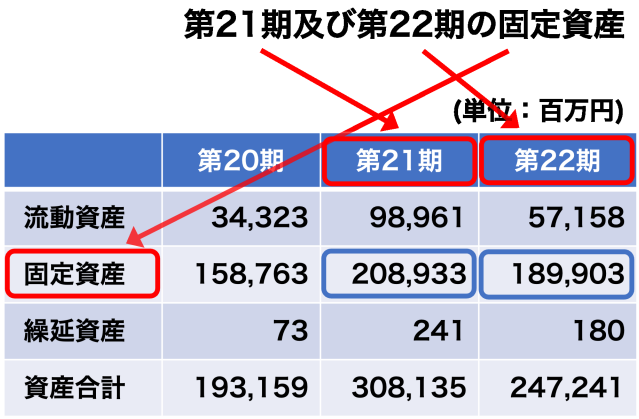


図 7　TTRE タスクの概要図

表 2 に，TTRE データセットの分布を示す．TTRE データセットは，HTML 形式の有価証券報告書が 67 件あり，test データ，train データにそれぞれ 25 件，42 件に分割されている．また，表は test データ，train データにそれぞれ 1125 件，1726 件あり，セルはそれぞれ 47517 件，80644 件あった．

表 3 に，TTRE データセットの train データにおける被検索セルが存在しないフレーズの上位 10 件を示す．数値または記号のみのフレーズが入力された場合，Name，Value に被検索セルが 1 つも該当しない事例が多数あることがわかった．なお，TTRE データセットの train データにおけるフレーズの個数は 3402 件であった．

表 2　TTRE データセットの統計

	文書	表	セル
test	25	1,125	47,517
train	42	1,726	80,644
合計	67	2,851	128,161

4 提 案 手 法

TUTA [2] は，表に含まれる階層的な情報を二次元座標木と呼ばれるツリーベースの構造で表現し，行と列の各インデックスと併用した埋め込みを用いることで，5 つのデータセットで最先端の結果を達成した．TUTA では，表内に階層構造がある場合，表の上 (左) から下 (右) に向かって，結合セルの大きさが徐々に小さくなることを前提としている．しかし，この方法では，図 2 に示すように，その前提に沿ったセルにしか座標を定めることができず，表の途中に出現する大きな結合セルには座標が割り当てられない．そこで，本稿では，表中の結合セルの大きさを網羅的に把握した上で，その大きさの降順に二次元座標木を構築する手法を提案する．

表 3　被検索セルが存在しないフレーズ

フレーズ	頻度	フレーズ	頻度
(1)	50	※ 1	18
2	34	(注)	18
(2)	34	1	18
3	25	(3)	18
4	18	※ 2	15

なお，以下では，TUTA で定まる二次元座標木を default tree と呼び，そのうち垂直方向 (列) と水平方向 (行) に関する木をそれぞれ “default vertical tree”，“default horizontal tree” と呼ぶこととし，提案手法で定まる二次元座標木を “exhaustive tree” と呼び，そのうち垂直方向 (列) と水平方向 (行) に関する木をそれぞれ “exhaustive vertical tree”，“exhaustive horizontal tree” と呼ぶこととする．

4.1 exhaustive tree

提案手法を図 3 を用いて説明する．まず，exhaustive vertical tree について，列方向の結合セル (col span) の最も大きいセルが含まれる 1 列を親ノードに設定する．ただし，セルのサイズが同一であった場合，より上部にあるセルを親ノードに設定する．例の表では，(0)，(1) と記されたの 2 つのセルが該当する．次に，親ノードに該当する各セルについて，列の範囲が内包されるセルのうち，列方向の結合セル (col span) のサイズが次に大きいセルが含まれる 1 列 (列の範囲が内包されるセルのみ) を親ノードに対する子ノードにする．ただし，セルのサイズが同一であった場合，より上部にあるセルを子ノードに設定する．例の表では，親ノード (0) の列方向の結合セル (col span) が 4 に対して，表の左側 4 列の中で，次に列方向の結合セル (col span) が大きいセルは，(0, 0) と記載されたセルであるため，当該の 1 列のうち，列の範囲が内包されるセルを親ノードに対する子ノードにする．ここでは，(0, 0)，(0, 1) と記されたの 2 つのセルが該当する．同様の手順を再起的に繰り返し，列方向の結合セル (col span) のサイズが 1 になったら終了とする．exhaustive horizontal tree についても，同様に，行方向の結合セル (row span) のサイズをもとに木構造を構築する．

4.2 データ変換

HTML 形式の有価証券報告書を，TUTA モデルの入力に適した形式に変換する方法を説明する．

4.2.1 表　抽　出

HTML 文書から table タグで囲まれた範囲を抽出した．ただし，アノテータにより cell-id が 1 つも付与されていない表は

```
<td data-ufo-tde-cell-id="S100LO2J-0102010-tab2-r5c6"
    data-ufo-tde-cell-type="data"
    style="border-left: 1px solid #000000;
        border-top: 1px solid #000000;
        border-right: 1px solid #000000;
        border-bottom: 1px solid #000000;
        vertical-align: middle">
    <p style="margin-right: 42px; text-align: right">
        <span style="font-size: 12px">812.6</span>
    </p>
</td>
```

図 8　HTML 形式のセルの例

表 4　使用したセルの特徴

特徴	説明	例
V	セルの値	812.6
DT	データタイプ	1
HF	数式であるか	0
LB	左部に罫線があるか	1
TB	上部に罫線があるか	1
BB	下部に罫線があるか	1
RB	右部に罫線があるか	1
BC	背景色	#fffff
FC	文字の色	#000000
FB	文字が太字であるか	0

除外した．

4.2.2　表 正 規 化

HTML 形式の表を NFKC(Normalization Form Compatibility Composition) 方式により正規化した．また，△記号は全て-記号に置換した．

4.2.3　Json 形式への変換

HTML 形式の表を TUTA モデルの入力に適した json 形式に変換した．使用したセルの特徴を表 4 に示す．また，HTML 形式のセルの例を図 8 に示す．V はセルの値を表しており，例では「812.6」となる．DT はセルの値のデータタイプを表しており，テキストであれば 0，数値であれば 1，空欄であれば 5 とした．数値であるかの判定は，セルの値の過半数が数値であるかという基準とした．HF は，セルの値に数式が含まれるかを表しており，全て 0 とした．LB，TB，BB，RB は，それぞれ，左部，上部，下部，右部に罫線があれば 1, なければ 0 とした．BC は背景色，FC は文字の色を設定した．FB はフォントが太字か否かを表しており，太字であれば 1，そうでなければ 0 とした．また，表の特徴として，各セルの座標，結合セルの座標，表の木構造も使用した．

4.2.4　英語への翻訳

Google 翻訳 API を用いて，セルの値を翻訳した．日本の通貨単位である「円」を「circle」と誤訳するケースがあったため,「circle」を全て「yen」に置換した．

4.3　TDE

事前学習済み TUTA-implicit モデルを Metadata, Header, Attribute, Data の 4 つのクラスに分類するように微調整した．TUTA には，TUTA-base モデル，TUTA-implicit モデル，TUTA-explicit モデルの 3 種類のモデルがあるが，TUTA-implicit モデルは，学習時に位置埋め込みを使用したモデルである．先行研究で，TUTA-implicit モデルの性能が最も高いことが示されているため，本研究では，TUTA-implicit モデルを使用した．モデルのアーキテクチャは変更せずに使用した．バッチサイズは 2 に設定し，学習率を 8e-6 として，最大 100 エポックまで学習させた．

4.4　TTRE

セルのクラスを考慮したセル検索手法を提案する．図 9 は TTRE サブタスクにおける提案手法の処理概要である．おおまかな流れとしては，まず，所与のフレーズと同一文書内の表の各セルのテキストを Text Encoder に入力し，それぞれの埋め込み表現を取得する．そして，フレーズとセルのテキストの類似度を計算し，Name を決定する．次に，表の特徴を Cell Type Classifier のためのモデルに入力し，表の各セルのクラスを取得する．最後に，Name の情報と，表の各セルのクラスをもとに Value を決定する．

4.4.1　Name

まず，所与のフレーズと同一文書内の表の各セルのテキストを multilingual-e5 [3] モデルにより埋め込み表現を取得した．そして，フレーズとセルのテキストの埋め込み表現のコサイン類似度を計算した．次に，類似度が高い順にランキングし，閾値以上の上位 K 件を Name とした．類似度と取得する件数の閾値は，学習データにより Name の性能が最も高くなる値に設定した．ただし，フレーズが数値または記号のみで構成されていた場合は除外した．また，出現頻度の多い日本語の “注” フレーズも除外した．

4.4.2　Value

表の特徴を TUTA モデルに入力し，表の各セルを Metadata, Header, Attribute, Data の 4 つのクラスに分類した．次に，Name と推定されるセルと同一の行または列に属するセルを抽出し，そのうち，Data クラスに分類されたセルを Value とした．

5　実　　験

本節では，exhaustive tree による手法が，表構造理解に有効であるかを TDE データセットと TTRE データセットを用いて検証する．

5.1　評 価 方 法

5.1.1　TDE

表の木構造を考慮しない手法 (no tree)，従来手法 (default tree)，提案手法 (exhaustive tree) の 3 種類の手法を比較することにより有用性を検証する．全ての手法において，5 回ずつ実行し，マクロ平均の適合率，再現率，F1 値のそれぞれの平均を比較した．

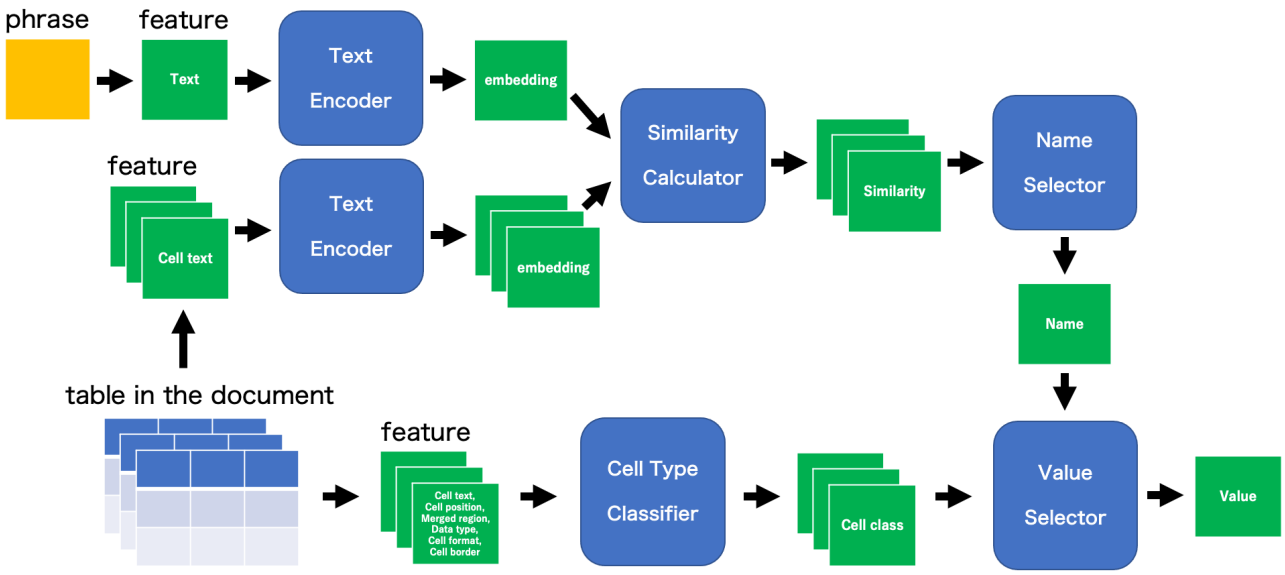


図 9　TTRE サブタスクにおける提案手法の処理概要

表 5　TDE データセットでの評価結果

Method	Metadata	Header	Attribute	Data	マクロ平均		
	F1	F1	F1	F1	適合率	再現率	F1
TUTA $_{\text{no tree}}$	**0.6744**	0.8620	**0.9886**	**0.9046**	**0.8863**	**0.8365**	**0.8574**
TUTA $_{\text{default tree}}$	0.6654	**0.8652**	0.9878	0.9002	0.8847	0.8330	0.8546
TUTA $_{\text{exhaustive tree}}$	**0.6743**	**0.8679**	**0.9901**	**0.9017**	**0.8887**	**0.8375**	**0.8585**

5.1.2　TTRE

表の木構造を考慮しない手法 (no tree)，従来手法 (default tree)，提案手法 (exhaustive tree)，セルのクラスを考慮しない手法の 4 種類の手法を比較することにより有用性を検証する．セルのクラスを考慮しない手法とは，Name と推定されるセルと同一の行または列に属する全てのセルを Value とする手法である．また，Text Encoder の違いによる影響についても検証する．Text Encoder は，multilingual-e5 モデル，sentence-luke モデル [4]，sentence-bert モデル [5]，deberta-v2 モデル [6]，mdeberta-v3 モデル [7] の 5 種類の性能を比較した．

最先端の性能を誇る汎用的なテキスト埋め込みモデルの 1 つに，E5(EmbEddings from bidirEctional Encoder rEpresentations) がある．E5 では，ラベルなしの大規模データセットである CCPairs を用いて弱教師ありの対照学習によって事前学習している．この事前学習済みモデルに対して，ラベル付きの小規模データセットでファインチューニングを施すことで，さらに高品質なテキスト埋め込みを獲得できる．multilingual-e5 は，xlm-roberta-base の事前学習済みパラメータを初期値として，多種多様な多言語データセット (Filtered mC4，CC News，NLLB，Wikipedia，Filtered Reddit，S2ORC，Stackexchange，xP3，Miscellaneous unsupervised SBERT data) を用いた対照学習による事前学習をしている．この多言語データセットで事前学習したモデルに対して，ラベル付きのデータセット (MS MARCO，NQ，Trivia QA，NLI from SimCSE，ELI5，DuReader Retrieval，KILT Fever，KILT HotpotQA，SQuAD，Quora，Mr. TyDi，MIRACL) を用いてファインチューニングしている．

LUKE は，BERT での事前学習タスクに加え，新たに単語とエンティティの文脈化表現を学習したモデルであり，エンティティに関する様々なタスクにおいて最先端の性能に到達している．このモデルは，与えられたテキスト中の単語とエンティティを独立したトークンとして扱い，それらの文脈化表現を出力する．ここではエンティティに着目するため，自己注意機構を拡張することで，単語およびエンティティを考慮したスコアを算出可能にしている．

SBERT は，従来の BERT よりも優れた文埋め込みを獲得するために，2 つの BERT のそれぞれの出力にプーリング層を追加したシャムネットワーク構造のモデルである．この SBERT は，従来の BERT よりも文埋め込みの獲得だけではなく，推論時間の大幅な削減も実現した．

DeBERTa は，BERT と RoBERTa を改良したモデルであり，各単語とその単語の相対的な出現位置をディスエンタング

表 6 TTRE データセットでの評価結果

Method	Name			Value			Total
	適合率	再現率	F1	適合率	再現率	F1	F1
multilingual-e5-small	**0.3687**	**0.4434**	0.3198	0.0898	0.4939	0.1154	0.2176
multilingual-e5-small + TUTA $_{\text{no tree}}$	**0.3687**	**0.4434**	0.3198	**0.2921**	0.4918	**0.2708**	0.2953
multilingual-e5-small + TUTA $_{\text{default tree}}$	**0.3687**	**0.4434**	0.3198	0.2682	0.4867	0.2562	0.2880
multilingual-e5-small + TUTA $_{\text{exhaustive tree}}$	**0.3687**	**0.4434**	0.3198	0.2794	0.4911	0.2604	0.2901
multilingual-e5-base	0.3556	**0.4574**	**0.3221**	0.0857	**0.5069**	0.1186	0.2204
multilingual-e5-base + TUTA $_{\text{no tree}}$	0.3556	**0.4574**	**0.3221**	0.2733	**0.5048**	**0.2742**	**0.2982**
multilingual-e5-base + TUTA $_{\text{default tree}}$	0.3556	**0.4574**	**0.3221**	0.2514	0.5001	0.2596	0.2909
multilingual-e5-base + TUTA $_{\text{exhaustive tree}}$	0.3556	**0.4574**	**0.3221**	0.2613	0.5044	0.2639	0.2930
multilingual-e5-large	**0.3728**	0.4355	**0.3212**	0.0885	0.4813	0.1153	0.2182
multilingual-e5-large + TUTA $_{\text{no tree}}$	**0.3728**	0.4355	**0.3212**	**0.2874**	0.4799	0.2704	**0.2958**
multilingual-e5-large + TUTA $_{\text{default tree}}$	**0.3728**	0.4355	**0.3212**	0.2647	0.4751	0.2566	0.2889
multilingual-e5-large + TUTA $_{\text{exhaustive tree}}$	**0.3728**	0.4355	**0.3212**	0.2757	0.4793	0.2613	0.2912
sentence-luke (Japanese)	0.3401	0.4213	0.3000	0.0827	0.4643	0.1120	0.2060
sentence-luke (Japanese) + TUTA $_{\text{no tree}}$	0.3401	0.4213	0.3000	0.2668	0.4633	0.2592	0.2796
sentence-luke (Japanese) + TUTA $_{\text{default tree}}$	0.3401	0.4213	0.3000	0.2436	0.4587	0.2446	0.2723
sentence-luke (Japanese) + TUTA $_{\text{exhaustive tree}}$	0.3401	0.4213	0.3000	0.2586	0.4629	0.2528	0.2764
sentence-bert (Japanese)	0.3237	0.3948	0.2805	0.0798	0.4493	0.1067	0.1936
sentence-bert (Japanese) + TUTA $_{\text{no tree}}$	0.3237	0.3948	0.2805	0.2567	0.4477	0.2463	0.2634
sentence-bert (Japanese) + TUTA $_{\text{default tree}}$	0.3237	0.3948	0.2805	0.2363	0.4434	0.2331	0.2568
sentence-bert (Japanese) + TUTA $_{\text{exhaustive tree}}$	0.3237	0.3948	0.2805	0.2470	0.4476	0.2381	0.2593
deberta-v2-tiny (Japanese)	0.2969	0.3922	0.2753	0.0735	0.4288	0.1029	0.1891
deberta-v2-tiny (Japanese) + TUTA $_{\text{no tree}}$	0.2969	0.3922	0.2753	0.2399	0.4273	0.2441	0.2597
deberta-v2-tiny (Japanese) + TUTA $_{\text{default tree}}$	0.2969	0.3922	0.2753	0.2204	0.4225	0.2299	0.2526
deberta-v2-tiny (Japanese) + TUTA $_{\text{exhaustive tree}}$	0.2969	0.3922	0.2753	0.2304	0.4267	0.2353	0.2553
deberta-v2-base (Japanese)	0.3013	0.3992	0.2810	0.0757	0.4254	0.1051	0.1931
deberta-v2-base (Japanese) + TUTA $_{\text{no tree}}$	0.3013	0.3992	0.2810	0.2443	0.4231	0.2475	0.2642
deberta-v2-base (Japanese) + TUTA $_{\text{default tree}}$	0.3013	0.3992	0.2810	0.2249	0.4194	0.2337	0.2574
deberta-v2-base (Japanese) + TUTA $_{\text{exhaustive tree}}$	0.3013	0.3992	0.2810	0.2334	0.4233	0.2373	0.2591
deberta-v2-large (Japanese)	0.3151	0.3823	0.2739	0.0765	0.4165	0.1015	0.1877
deberta-v2-large (Japanese) + TUTA $_{\text{no tree}}$	0.3151	0.3823	0.2739	0.2511	0.4152	0.2390	0.2564
deberta-v2-large (Japanese) + TUTA $_{\text{default tree}}$	0.3151	0.3823	0.2739	0.2323	0.4107	0.2260	0.2500
deberta-v2-large (Japanese) + TUTA $_{\text{exhaustive tree}}$	0.3151	0.3823	0.2739	0.2409	0.4146	0.2306	0.2523
mdeberta-v3-base	0.2345	0.2665	0.1983	0.0547	0.2791	0.0700	0.1342
mdeberta-v3-base + TUTA $_{\text{no tree}}$	0.2345	0.2665	0.1983	0.2014	0.2786	0.1822	0.1903
mdeberta-v3-base + TUTA $_{\text{default tree}}$	0.2345	0.2665	0.1983	0.1832	0.2773	0.1721	0.1852
mdeberta-v3-base + TUTA $_{\text{exhaustive tree}}$	0.2345	0.2665	0.1983	0.1898	0.2782	0.1729	0.1856

ル注意機構を用いることで，独立した 2 つのベクトルとしての表現を可能にしている．また事前学習時には，デコーダを単語の絶対的な出現位置をマスクできるように拡張している．それに加え，ファインチューニング時には仮想敵対的学習を行っている．DeBERTa V2 では，トークナイザーの語彙力が増強され，入力トークンの局所依存性を学習するために畳み込み層を追加し，注意機構では位置射影行列と内容射影行列は共有され，相対位置を符号化するためにバケットを使用している．DeBERTa V3 は，DeBERTa の改良版であり，Gradient Disentangled Embedding Sharing を使用している．mDeBERTa は，多言語データセットである CC100 を用いて DeBERTa を学習したモデルである．

sentence-luke モデル，sentence-bert モデル，deberta-v2 モデルは日本語のモデルを使用し，multilingual-e5 モデル，mdeberta-v3 モデルは多言語モデルを使用した．また，deberta-v2 モデルは，トークナイズの際に，Juman++ 2.0.0-rc3 [8] [9] を使用した．類似度と取得する件数の閾値は，学習データにより Name の性能が最も高くなる値に設定した．

5.2 結果と考察

5.2.1 TDE

表 5 に TDE データセットでの評価結果を示す．マクロ平均の適合率，再現率，F1 スコアにおいて，exhaustive tree による手法，no tree による手法，default tree による手法の順で高い性能を示す結果となった．全ての手法において，Header，Attribute，Data クラスに比べ，Metadata クラスに対しては

低い性能を示す結果となった．これは，Metadata クラスに分類される事例が極端に少なく，効率的に学習できなかったことが原因として挙げられる．

5. 2. 2 TTRE

表 6 に TTRE サブタスクの評価結果を示す．Text Encoder の違いによる影響については，multilingual-e5-base による手法が Name において最も高い性能を示す結果となった．Cell Type Classifier においては，no tree による手法が全ての Value，Total の F1 スコアにおいて最も高い性能を示す結果となった．また，Text Encoder に multilingual-e5-base，Cell Type Classifier に no tree による手法を用いた場合，29.82%と最も高い性能を示す結果となった．

no tree による手法が最も高い性能を示した理由として，TDE サブタスクでの，Data クラスを分類するためのモデルの性能が，TTRE サブタスクに大きな影響を与えるという仮説を立てた．TDE データセットを用いて，Data クラスを分類するためのモデルの性能を検証した結果，no tree による手法が最も高い性能を示した．この結果から，Data クラスを分類するためのモデルの性能を向上させることが重要であることがわかった．

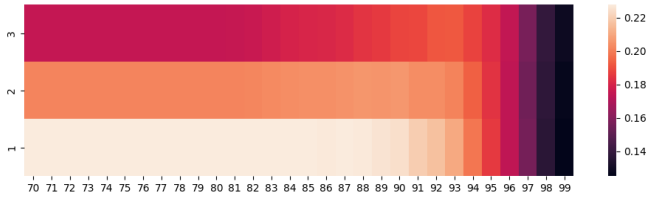


図 10　Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5)

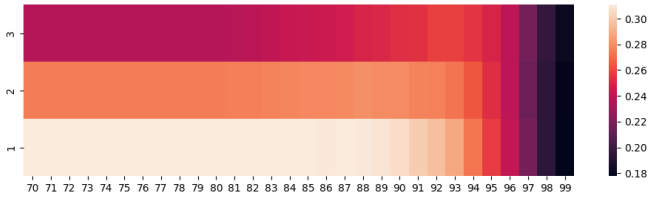


図 11　Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{no tree}}$)

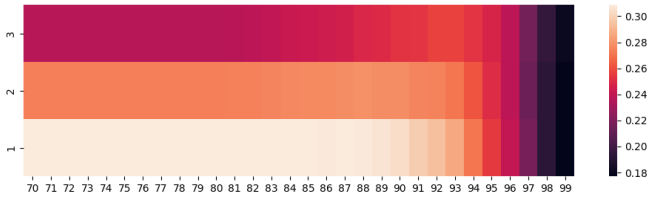


図 12　Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{default tree}}$)

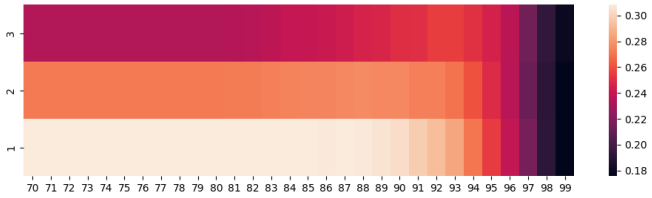


図 13　Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{exhaustive tree}}$)

図 10〜13 に Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ Total における類似度と上位 K 件による閾値ごとの評価結果のヒートマップを示す．縦軸は取得する被検索セルの個数の上限を表しており，横軸は取得する被検索セルのテキストと所与のテキストにおけるコサイン類似度の下限を表している．全ての手法において，所与のテキストに対して上位 1 件のセルを取得した場合に高い性能を示しており，複数セルの検索については，十分な性能を発揮できていないことが示唆された．

以下では，多言語モデルと日本語のモデルにおける性能の違いについて分析する．multilingual-e5-base は多言語モデルの中で最も高い性能を示したモデルであり，sentence-luke は日本語のモデルの中で最も高い性能を示したモデルである．TUTA を使用した手法と，TUTA を使用しなかった手法で，Value を比較すると，再現率には大きな差がなく，適合率には有意義な差があった．このことから，Value を決定する際に，セルのタイプを判定し，data クラス以外を除外することは非常に有力であるといえる．また，全ての手法で，適合率よりも再現率の方が高い結果となった．これは，Name が，所与のフレーズとセルの値の類似度のみで決められているため，実際に参照すべき表とは異なる表もまとめて取得してしまっていることが原因として考えられる．

表 7　多言語モデルと日本語モデルにおける予測結果の一致度合い

	頻度 (割合)
両方とも正解	441 (0.235)
多言語モデルのみ正解	40 (0.021)
日本語モデルのみ正解	33 (0.018)
両方とも不正解	677 (0.361)

表 7 に，Name における multilingual-e5-base を用いた手法と sentence-luke を用いた手法による予測結果の一致度合いを示す．両方の手法で正しく識別できた件数は 441 件であり，全体の 23.5%を占める結果となった．多言語モデルのみ正解できた件数は 40 件であり，全体の 2.1%を占める結果となった．日本語モデルのみ正解できた件数は 33 件であり，全体の 1.8%を占める結果となった．両方の手法で全く同じ誤りをした件数は 677 件であり，全体の 36.1%を占める結果となった．

6　追 加 実 験

本節では，exhaustive tree による手法が，表構造理解に有効であるかを DeEx データセットを用いて検証する．DeEx データセットは，セルタイプ分類のためのデータセットである．分類するクラスは，metadata, notes, data, attributes, header, derived の 6 種類であり，それぞれの分布を表 9 に示す．DeEx データセットは，data クラスに分類されるセルがほとんどであるといった特徴がある．一方で，notes クラスの分類されるセルはごく僅かであるといった特徴がある．

表 8　DeEx データセットでの評価結果

Method	metadata	notes	data	attributes	header	derived	macro-F1
TUTA $_{\text{no tree}}$	0.8574	**0.4825**	**0.9937**	**0.8114**	**0.8701**	**0.8092**	**0.8041**
TUTA $_{\text{default tree}}$	**0.8585**	0.4419	0.9930	**0.8382**	0.8547	0.7391	0.7876
TUTA $_{\text{exhaustive tree}}$	**0.8671**	**0.5589**	**0.9931**	0.8045	**0.8638**	0.7491	**0.8061**
TUTA $_{\text{raw tree}}$	0.8516	0.4271	0.9911	0.7888	0.8070	**0.7767**	0.7737

表 9　セルタイプの分布

セルタイプ	頻度 (割合)
metadata	20,020 (0.0154)
notes	5,314 (0.0041)
data	1,232,762 (0.9458)
attributes	7,024 (0.0054)
header	21,810 (0.0167)
derived	16,538 (0.0127)

6.1 評価方法

従来の default tree による手法，提案手法の exhaustive tree による手法，表の木構造を考慮しない no tree による手法，TUTA の著者らにより提供された raw tree の 4 つの手法を比較することにより有用性を検証する．事前学習済み TUTA-implicit モデルを metadata, notes, data, attributes, header, derived の 6 つのクラスに分類するように微調整した．モデルのアーキテクチャは変更せずに使用した．データセットは，DeEx データセットを無作為に 5 分割されたものを使用した．検証には，交差検証法を用い，平均マクロ F1 を算出した．バッチサイズは 2 に設定し，学習率を 8e-6 として，最大 200 エポックまで学習させた．

6.2 結果と考察

表 8 に DeEx データセットでの実験結果を示す．提案手法がマクロ F1 において，最も性能の高い結果となった．data クラスの F1 スコアが最も高く，notes クラスの F1 スコアが最も低い結果となった．data クラスの個数が最も多く，notes クラスの個数が最も少ないことから，セルタイプの分布が大きく影響していると考えられる．提案手法が notes クラスにおいて，最も性能の高い結果となったことから，提案手法が少数のデータからも効果的に学習する能力が高い可能性がある．

表 10　表データにおける木構造で表現した際の深度

木の種類	TDE		DeEx	
	平均	標準偏差	平均	標準偏差
default vertical tree	0.3658	1.0948	0.4444	2.2411
default horizontal tree	0.2162	0.6906	0.1995	1.0503
exhaustive vertical tree	0.3919	0.8152	0.5079	1.0413
exhaustive horizontal tree	0.2255	0.6488	0.1814	0.6380
合計	1.1994	-	1.3332	-

TDE データセットと DeEx データセットを用いて，表データを木構造で表現した際の深度を調査した．その結果，表 10 に示すように，TDE データセットと DeEx データセットは複雑な表が少ないことがわかった．そのため，no tree による手法と提案手法で，あまり性能の差が見られなかった原因として，データセットに複雑な表が少なかったためであると結論づけた．

7 まとめ

本稿では，表中の結合セルの大きさを網羅的に把握した上で，その大きさの降順に二次元座標木を構築する手法を提案した．表のセルタイプ分類と，表とテキストとの関連付けの 2 種類のタスクにおいて，従来研究と性能を比較することにより，提案手法の有用性を検証した．その結果，提案手法は，表が複雑な構造をしている際に最も活躍することがわかった．表とテキストとの関連付けのタスクにおいては，セルを検索する際に，Data クラスに該当するセルを除外する手法は有効的であった．ただし，セルタイプ分類の総合的な性能ではなく，Data クラスを抽出するための性能が重要であることがわかった．

謝　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　献

[1] Yasutomo Kimura and Hokuto Ototake and Kazuma Kadowaki and Takahito Kondo and Makoto P. Kato, “Overview of the NTCIR-17 UFO Task,” Proceedings of The 17th NTCIR Conference, 2023.

[2] Wang, Zhiruo and Dong, Haoyu and Jia, Ran and Li, Jia and Fu, Zhiyi and Han, Shi and Zhang, Dongmei, “TUTA: Tree-based Transformers for Generally Structured Table Pre-training,” Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1780–1790, 2021.

[3] Wang, Liang and Yang, Nan and Huang, Xiaolong and Jiao, Binxing and Yang, Linjun and Jiang, Daxin and Majumder, Rangan and Wei, Furu, “Text Embeddings by Weakly-Supervised Contrastive Pre-training,” 2022.

[4] Ikuya Yamada and Akari Asai and Hiroyuki Shindo and Hideaki Takeda and Yuji Matsumoto, “LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention,” EMNLP, 2020.

[5] Reimers, Nils and Gurevych, Iryna, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[6] Pengcheng He and Xiaodong Liu and Jianfeng Gao and Weizhu Chen, “DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION,” International Conference on Learning Representations, 2021.

[7] Pengcheng He and Xiaodong Liu and Jianfeng Gao and Weizhu Chen, “DEBERTA: DECODING-ENHANCED

BERT WITH DISENTANGLED ATTENTION," International Conference on Learning Representations, 2021.
[8] Morita, Hajime and Kawahara, Daisuke and Kurohashi, Sadao, "Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2292–2297, 2015.
[9] Tolmachev, Arseny and Kawahara, Daisuke and Kurohashi, Sadao, "Juman++: A Morphological Analysis Toolkit for Scriptio Continua," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 54–59, 2018.
[10] Koci, Elvis and Thiele, Maik and Rehak, Josephine and Romero, Oscar and Lehner, Wolfgang, "DECO: A Dataset of Annotated Spreadsheets for Layout and Table Recognition," 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1280-1285, 2019.
[11] Pengcheng Yin and Graham Neubig and Wen-tau Yih and Sebastian Riedel, "TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data", 2020.
[12] Herzig, Jonathan and Nowak, Pawel Krzysztof and Müller, Thomas and Piccinno, Francesco and Eisenschlos, Julian, "TaPas: Weakly supervised table parsing via pre-training", 2020.
[13] Jixiong Liu and Yoan Chabot and Raphaël Troncy and Viet-Phi Huynh and Thomas Labbé and Pierre Monnin, "From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods," Journal of Web Semantics, Volume 76, 2023.
[14] Gilbert Badaro, Mohammed Saeed, Paolo Papotti, "Transformers for Tabular Data Representation: A Survey of Models and Applications," 2023,
[15] Zilong Wang and Hao Zhang and Chun-Liang Li and Julian Martin Eisenschlos and Vincent Perot and Zifeng Wang and Lesly Miculicich and Yasuhisa Fujii and Jingbo Shang and Chen-Yu Lee and Tomas Pfister, "Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding," 2024,
[16] Tianshu Zhang and Xiang Yue and Yifei Li and Huan Sun, "TableLlama: Towards Open Large Generalist Models for Tables," 2023,
[17] Mehmed Kantardzic and Associate Editor and Witold Pedrycz and Editor-in-Chief, "Table understanding: Problem overview," 2022,

付　　録

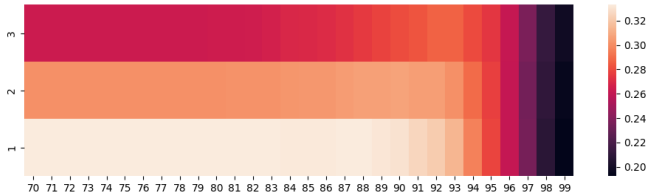


図 14　Name における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5)

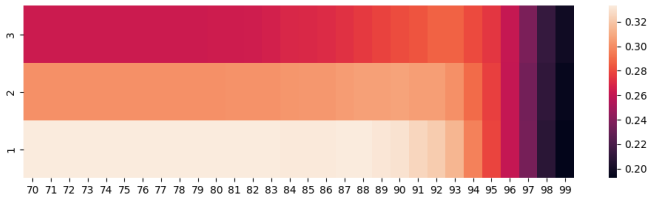


図 15　Name における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{no tree}}$)

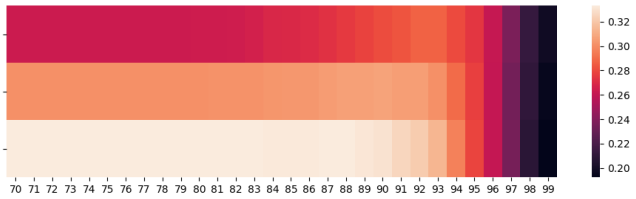


図 16　Name における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{default tree}}$)

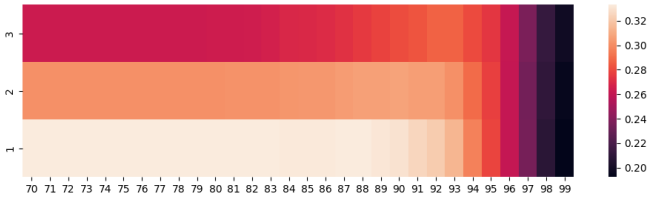


図 17　Name における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{exhaustive tree}}$)

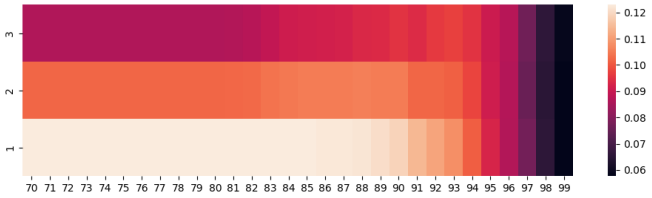


図 18　Value における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5)

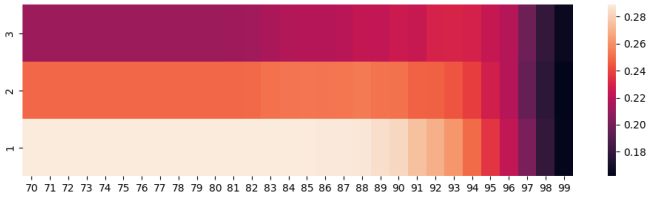


図 19　Value における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{no tree}}$)

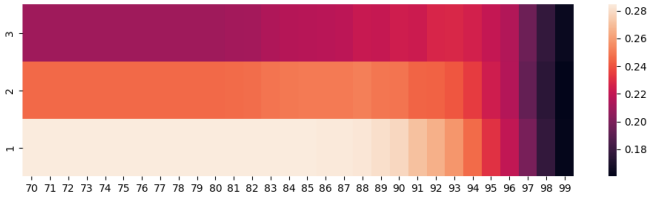


図 20　Value における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{default tree}}$)

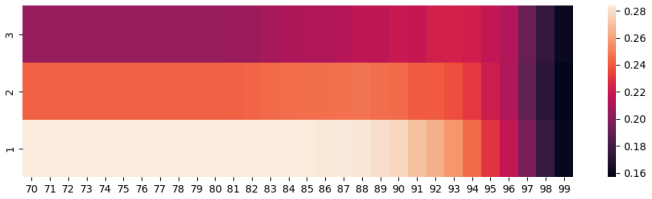


図 21　Value における類似度と上位 K 件による閾値ごとの評価結果のヒートマップ (multilingual-e5 + TUTA $_{\text{exhaustive tree}}$)

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月29日(木) 16:00 ~ 18:10 | T1-B オンライン（Zoom Events)

文生成

座長:佐々木 佑樹(富士通株式会社)
コメンテータ:酒井 哲也(早稲田大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T1-B-6-01] BERTを用いた感情推定に基づく発話スタイル付き応答文生成システム

*西 佑真[1]、梶岡 慎輔[1]、山本 大介[1] (1. 名古屋工業大学)

16:30 ~ 16:55
[T1-B-6-02] 経路単純化と案内文単純化を組み合わせた経路案内文要約システム

*鋤柄 知哉[1]、金 鎔煥[1]、山本 大介[1] (1. 名古屋工業大学)

16:55 ~ 17:20
[T1-B-6-03] 大規模言語モデルによる商品レビュー文のキャッチコピーへの変換

*沈 辰夫[1]、荘司 慶行[2]、山本 岳洋[3]、大原 剛三[1]、田中 克己[4] (1. 青山学院大学、2. 静岡大学、3. 兵庫県立大学、4. 福知山公立大学)

17:20 ~ 17:45
[T1-B-6-04] 特定選手に注目した試合結果記事見出しのパーソナライズ

*佐々木 泰河[1]、工家 昂之[1]、山本 岳洋[1]、荘司 慶行[2] (1. 兵庫県立大学、2. 静岡大学)

BERTを用いた感情推定に基づく発話スタイル付き応答文生成システム

西　佑真† 梶岡　慎輔† 山本　大介†

† 名古屋工業大学　〒 466-8555 愛知県名古屋市昭和区御器所町
E-mail: †y.nishi.744@stn.nitech.ac.jp, ††{kajioka,yamamoto.daisuke}@nitech.ac.jp

あらまし　本学にある MMDAgent というシステムとそれを拡張したシステムにより，システムが認識するキーワードとキャラクタの応答する文を設定するだけで簡単に対話を作成できるようになった．一方で一般ユーザにより登録された対話を見ると，依然として様々な問題点があった．そこで本研究では一般ユーザが簡単によりよい一問一答形式の対話を作成できるようにサポートすることを目的としている．提案手法では，認識キーワードから応答文の生成を行う機能，応答文から感情推定しモーションと声色を選択する機能，応答文中のモーションの開始時刻を決定する機能を作成した．これによりユーザは認識キーワードを考えるだけで，効率的に対話を作成することが可能となる．

キーワード　テキスト分類，対話システム，言語モデル，質問応答，応答生成

1 はじめに

1.1 研究の背景

近年，Siri [1] や Alexa [2] などのユーザがエージェントと対話することで情報を得る音声対話システムが普及している．本研究室では音声インタラクションシステム構築ツールキット MMDAgent [3] を用いた研究が行われている．MMDAgent では FST(Finite State Transducer) 形式で記述された対話シナリオに基づいて音声対話の処理を実行している．ユーザはこの対話シナリオを編集することで，対話シナリオを自由に構成することができる．しかし，一般ユーザにとって FST 形式は馴染みがなく，対話シナリオの編集が難しいという問題点があった．その問題を解決するため，MMDAgent EDIT [4] や Google Blockly を用いた対話シナリオ編集方法 [5] が研究・提案されてきた．これらの手法は専門知識を持たない一般ユーザでも簡単に対話シナリオを編集することを可能とした．

一方で、後述する MMDAgent EDIT に投稿された対話シナリオの登録データを確認するといくつかの問題点が存在することを確認した．まず最初に応答文を生成することが難しいということである．案内や分かりやすく説明する文章を作成することは難しく，結果として MMDAgent EDIT に投稿されたデータには認識キーワードをそのまま応答文とするデータや，対話として成立していないデータが存在する．次にモーションと声色が設定されていない場合があるということである．モーションはキャラクタが応答文を話す際に行う動作で，声色は応答文の話し方の調子である．これらは認識キーワードと応答文を設定する際に選択するだけで設定可能な項目であるが，デフォルトの普通となっている登録データがとても多くなっている．これは一般ユーザにとって，多くのモーションと声色の中から応答文に合った項目を選択することは面倒ということや，設定する優先度が認識キーワードと応答文に対して低いことが原因として考えられる．最後にモーションを開始するタイミングが考慮されていないということである．表 1 の上段に示すようにモーションが設定されている応答文では，キャラクタは話し始めると同時にモーションを開始してしまう．そこで，下段に示すように前半の文章ではモーションを開始せず，後半の「汗を流して引き締まった体を作りましょう。」の部分でガッツポーズを開始した方がよりよい対話になると考えられる．

表 1: モーション開始時刻の例

応答文	モーション
[開始] 名工大には三台のランニングマシンがあります。汗を流して引き締まった体を作りましょう。	ガッツポーズ
名工大には三台のランニングマシンがあります。	モーションなし
[開始] 汗を流して引き締まった体を作りましょう。	ガッツポーズ

1.2 本研究の目的

本研究の目的は，一般ユーザが簡単にモーションと声色が付与された，より良い一問一答形式の対話を作成できるようにサポートすることである．MMDAgent EDIT に投稿された対話シナリオの登録データでの問題点を解決するために，それぞれの観点から対話の作成をサポートする．

1.3 実現上の要件とアプローチ

本研究の目的を達成するためには以下のような要件があった．

要件 1　応答文の生成のサポートを行う必要がある．

要件 2　感情分類器で用いられる感情ラベルと，MMDAgent で使われるキャラクタのモーションと声色を対応させる必要がある．

要件 3　モーションの開始タイミングを決定する必要がある．

上記の要件に対して以下のアプローチによって問題を解決する．

アプローチ 1　MMDAgent EDIT に投稿されたデータを学習し，認識キーワードから応答文を生成する．

アプローチ 2　プルチックの感情の輪に基づき，感情を混合しモーション・声色を決定する．

アプローチ 3　感情分類器から出力される Attention の値からモーションの開始時刻を決定する．

2　関 連 研 究

2.1　音声インタラクションシステム構築ツールキット MMDAgent

本研究室では音声対話インタラクションシステム構築ツールキット MMDAgent を用いて研究を行っている．FST と呼ばれる対話スクリプトを編集することで，画面上の 3D キャラクタと対話が可能である．対話の登録には，システムが認識する認識キーワードと，キーワードを認識したときに話すキャラクタの応答文，応答する際のモーションや応答文を読み上げる声色の設定などが可能となっている．図 1 に，MMDAgent の動作画面を示す．

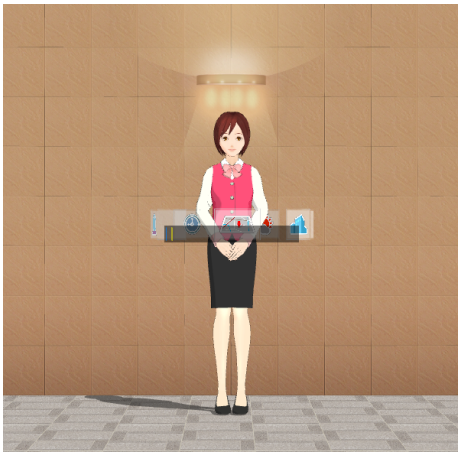


図 1: MMDAgent の動作画面

2.2　MMDAgent EDIT

MMDAgent EDIT は MMDAgent で構築可能なコンテンツを Web 上で作成可能なシステムである．図 2 に示すように，FST ファイルを編集する必要がなく，エージェントがユーザに対して応答する文章である応答文，応答文をエージェントから引き出すためにユーザが発話する認識キーワードの組を登録するだけで対話が作成できるため，一般のユーザでも簡単に一問一答形式の音声対話コンテンツを作成することができる．このシステムでも MMDAgent と同様に 3D キャラクタと対話する形式をとっており，ユーザがモーションや声色を設定することが可能となっている．

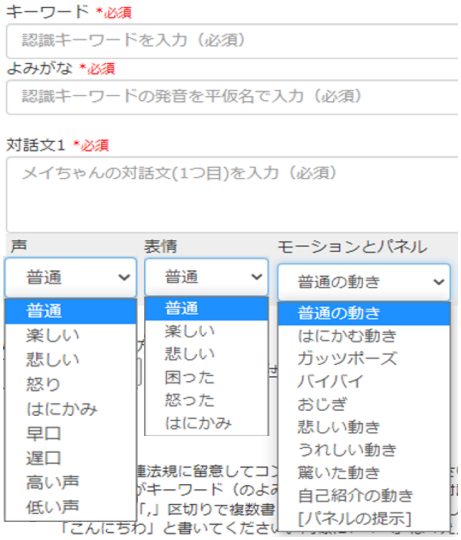


図 2: MMDAgent EDIT の設定画面

2.3　システム発話の感情分類による制御を行ったマルチモーダル対話システム

藤後ら [6] は，CG Erica のシステム発話を生成し，感情分類を行うことでジェスチャー及び音声制御を設定することによりマルチモーダルな対話システムの開発を行っている．感情分類結果に基づきジェスチャーの持つ 46 の変数を決定することで様々なジェスチャーを割り当てている．

3　提案システムの概要

提案システムの構成図を図 3 に示す．

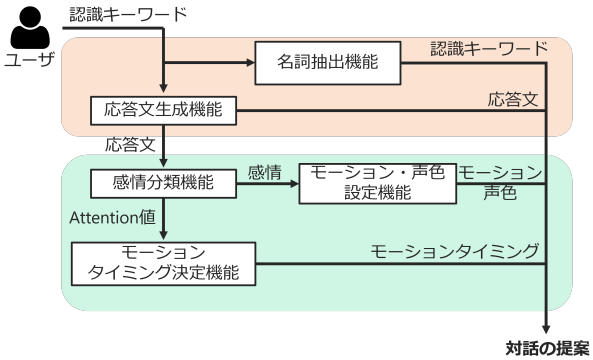


図 3: システムの構成図

まずシステムの流れとして，ユーザはシステムに認識キーワードを入力する．システムはそのキーワードから応答文生成機能によって応答文を生成し，その応答文に対して感情分類を行う．出力された感情に基づいて，対応付けされたモーションと声色がそれぞれ選択される．また，感情分類機能により感情分類された際に得られる Attention という値を用いることにより，応答文のどの部分に感情がこもっているかが可視化され，その値に基づいて応答文のモーションの開始時刻を決定する．最終的にユーザの入力による認識キーワードから名詞を抽出した認識キーワード，応答文，モーション・声色，モーションの

開始タイミングの内容をユーザが確認し，対話の作成を行うことができるというような流れになっている．

上記のアプローチにより，一般ユーザによる対話の作成をサポートすることができる．

4 提案システムの実現法

提案システムは，名詞抽出機能，応答文生成機能，感情分類機能，モーションタイミング決定機能，モーション・声色設定機能から成る．それぞれの実現方法を以下に示す．

4.1 名詞抽出機能

MMDAgent では，ユーザの発話内容から認識キーワードを認識したときに応答する仕組みとなっているため，名詞のような短いワードを認識キーワードとすることが多い．しかし，提案システムではどのようなキーワード・テキストでも認識キーワードとして受け付ける構造となっているため，認識キーワードとして入力されたテキストから名詞を抽出する．名詞の抽出には形態素解析器である Mecab を使用し，辞書には mecab-ipadic-NEologd [7] を使用した．mecab-ipadic-NEologd は週 2 回以上更新されているため，新語・固有表現に強いという特徴を持っているため採用した．

4.2 応答文生成機能

応答文生成機能は，入力としてユーザの認識キーワードを受け取り，応答文を生成する機能である．事前学習された汎用 GPT 言語モデルをファインチューニングすることで，実現している．

4.2.1 使用するモデル

rinna 社が公開している，36 億パラメータの GPT 言語モデル [8] を使用した．このモデルは日本語の Wikipedia などのオープンソースデータを使用して学習されている．

4.2.2 使用するテキストデータ

ファインチューニングに使用するデータは MMDAgent EDIT に投稿された，ユーザ生成の対話データである．データには認識キーワードや応答文，モーション，声色など様々な要素が含まれているが，今回は，認識キーワードと応答文のペアを抽出し，学習データとした．

4.2.3 テキストデータの前処理

ファインチューニングをする場合，データを学習させる前に処理しやすい形式に編集する必要がある．まず今回は認識キーワードと応答文の対話の形式になっている必要があるので，表 2 に示すような対話の形式になっていない認識キーワードと，誤字のある応答文を手作業で修正した．次に，機械学習に支障をきたすと考えられる文字列（“@”，“^”，“# ” 等）の削除を行い学習データを作成した．

表 2: 対話になっていない例

認識キーワード	応答文
名古屋工業大学	二十歳です
対話 1	一つ目の対話

4.2.4 学習の手順

ファインチューニングには，効率的に学習させることができる LoRA(Low-Rank Adaption of Large Language Models) [9] を使用して学習を行った．また，学習のエポック数と学習率に対して，optuna [10] を利用することでハイパーパラメータのチューニングを行った．

4.3 感情分類機能

感情分類機能は，応答文から感情を推定する機能である．本研究ではプルチックの感情の輪に基づき 8 クラス分類を行い，その出力に応じて応答文を 32 通りの感情に分類することで，モーションと声色を決定するというアプローチを取っている．

4.3.1 使用するモデル

東北大学の乾研究室により作成された，“bert-base-japanese-whole-word-masking” [11] を使用した．

4.3.2 使用するテキストデータ

ファインチューニングに使用するデータは日本語の感情分析の研究用データセットである，WRIME データセット [12] を使用した．このデータセットは SNS の投稿文書に対して，主観感情と客観感情のラベルが付与されたデータとなっている．主観感情は文章の書き手自身により付与されたラベルであり，客観感情は読み手 3 人によるラベルとなっている．本研究では，MMDAgent のキャラクタであるメイちゃんに行ってもらうモーションと声色のために感情推定を行うため，客観感情のラベルを利用することとした．また，ラベルはプルチックの感情の輪に基づき，喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼の 8 クラスとなっており，各感情の強度は 0 から 3 の 4 段階で与えられている．ここでプルチックの感情の輪は人間の感情を 8 つの基本的な感情と，その組み合わせで表現したものになっている．

4.3.3 テキストデータの前処理

基本的な処理は第 4.2.3 章と同様である．しかし，WRIME データセットに対して多く適用される処理として，三点リーダーと顔文字の除去がある．WRIME データセットは SNS から収集されたデータであるため，三点リーダーや顔文字が多く使用されている．三点リーダーは会話の印象を和らげる効果があると考えられるが，意図が伝わりにくくなる要因であると考えられたため，削除を行った．また顔文字については感情を表すことがあると思われるが，その顔文字が必ずしも文章と同じ感情を表しているとは限らないこと，句読点のように定型として付与されている可能性があることから，削除行った．表 3 に前処理の例を示す．

表 3: 修正した対話の例

修正前	修正後
人見知り治したい…(T_T)	人見知り治したい
三角チョコパイいちご…だと…… !?	三角チョコパイいちごだと!?

4.3.4 学習の手順

学習ではデータを学習とテストで 9:1 に分割した．その後学

習データを 5 分割して交差検証と，optuna によるハイパーパラメータのチューニングを行った．ハイパーパラメータのチューニングが終了したら，そのパラメータですべての学習用データで 1 つのモデルを作成するという手順で実験を行った．

4.3.5 感情選択の方法

感情分類機能では，応答文に対する 8 クラスの感情の確率値が出力される．本研究ではこの確率値に対して，最も大きい確率値が 50%を超えていた場合にはその基本感情を応答文の感情とし，50%以下の場合には 2 つの最大確率のラベルの感情を混合し，その混合感情を応答文に対応する感情として出力するように作成した．

4.4 モーションタイミング決定機能

モーションタイミング決定機能では，感情分類を行う際に得られる Attention の値を利用することで作成している．Attention は Tranformer で用いられる仕組みで，内積を計算することで入力されたデータのどこに注目するべきか動的に決定する仕組みである．感情分類を行うモデルの最終層の Transformer ブロックから，分類に使用している “[CLS]” トークンの Attention を取得する．この値は分類を行った際に文章のどの部分に注目していたかを示していると考えられるため，感情分類の根拠であり，感情がこもっている部分として使用している．図 4 にモーションタイミング決定機能の例を示す．

まず，応答文を句読点やエクスクラメーションマーク，クエスチョンマークで区切る．区切られたそれぞれのブロックで，Attention の値を合計し，最も値が大きいブロックからモーションを開始するというアプローチをとっている．図 4 では，応答文が「おめでとうございます。」と「明日の休日を楽しみにお仕事がんばってくださいね」の 2 つに分けられており，後半の方が Attention の値の合計が大きいため，その部分からモーションを開始するよう提案されている．

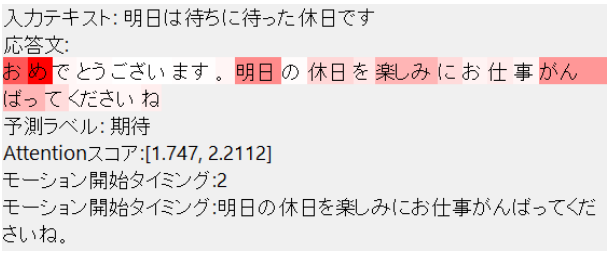


図 4: モーションタイミング決定機能の出力例

4.5 モーション・声色設定機能

プルチックの感情の輪で与えられている 8 つの基本感情，隣り合わせ，1 つおき，2 つおきの混合感情の計 32 種類の感情について，モーションと声色を設定した．設定する際には，MMDAgent で利用可能なモーションと声色を確認し，最もその感情を表していると考えられるモーションと声色を設定するという方法をとった．設定したモーションと声色を表 4 に示す．mei_act_aseri は焦っているモーションで，mei_greeting はお辞儀をするモーション，mei_act_gattu はガッツポーズを行うモーション，mei_look_down は下を見るモーションとなっている．用意されているモーションはポジティブな感情のモノが多く，ネガティブな感情の多くに mei_look_down を設定した．

表 4: 感情ラベルとモーション，声色の対応

感情	モーション	声色
喜び	mei_guide_happy	mei_voice_happy
喜び+信頼=愛	mei_guide_happy	mei_voice_happy
喜び+恐れ=罪悪感	mei_act_aseri	mei_voice_sad
喜び+驚き=歓喜	mei_guide_happy	mei_voice_happy
信頼	mei_guide_normal	mei_voice_normal
信頼+恐れ=服従	mei_greeting	mei_voice_normal
信頼+驚き=好奇心	mei_act_gattu	mei_voice_happy
信頼+悲しみ=感傷的	mei_guide_sad	mei_voice_sad
恐れ	mei_look_down	mei_voice_sad
恐れ+驚き=畏怖	mei_greeting	mei_voice_normal
恐れ+悲しみ=絶望	mei_look_down	mei_voice_sad
恐れ+嫌悪=羞恥心	mei_act_aseri	mei_voice_bashful
驚き	mei_surprise	mei_voice_high
驚き+悲しみ=拒絶	mei_look_down	mei_voice_sad
驚き+嫌悪=不信感	mei_guide_sad	mei_voice_normal
驚き+怒り=憤慨	mei_look_down	mei_voice_angry
悲しみ	mei_guide_sad	mei_voice_sad
悲しみ+嫌悪=後悔	mei_guide_sad	mei_voice_sad
悲しみ+怒り=嫉妬	mei_guide_sad	mei_voice_angry
悲しみ+期待=悲観的	mei_look_down	mei_voice_sad
嫌悪	mei_guide_normal	mei_voice_angry
嫌悪+怒り=軽蔑	mei_guide_normal	mei_voice_angry
嫌悪+期待=皮肉	mei_guide_normal	mei_voice_angry
嫌悪+喜び=不健康	mei_guide_normal	mei_voice_sad
怒り	mei_guide_normal	mei_voice_angry
怒り+期待=攻撃性	mei_guide_normal	mei_voice_angry
怒り+喜び=自尊心	mei_guide_happy	mei_voice_happy
怒り+信頼=支配	mei_guide_normal	mei_voice_happy
期待	mei_act_gattu	mei_voice_happy
期待+喜び=楽観	mei_guide_happy	mei_voice_happy
期待+信頼=希望	mei_act_gattu	mei_voice_happy
期待+恐れ=懸念	mei_act_aseri	mei_voice_normal

5 プロトタイプシステム

5.1 システムの概要

第 4 章で述べた提案システムを実現するため，Python を用いてプロトタイプシステムを実装した．また，開発環境として Visual Studio Code，miniconda，Windows10，機械学習ライブラリとして pytorch，ユーザインタフェースの作成に CustomTkinter を使用した．学習には RTX A5000 を使用した．

5.2 システムの実行画面

実際にユーザが使用する画面を図 5 に示す．

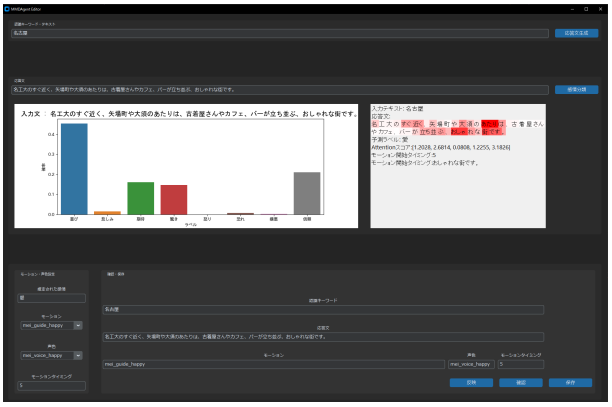


図 5: システムの実行画面

図 5 では，認識キーワードとして，名古屋を指定し提案システムを使用した例となっている．まず，応答文生成ボタンを押すと，応答文生成機能により応答文が出力される．次に，感情分類ボタンを押すと感情分類機能により，応答文の感情と，Attention の値を基に可視化された感情がこもっている部分の情報が表示される．応答文の感情は 8 クラスのそれぞれの感情の確率値が棒グラフで表示される．また，それらの情報から，画面下側の「モーション・声色設定」の枠にモーションや声色，モーションを行うタイミングが自動的に入力される．ユーザはそれらを確認・修正したら，反映ボタンを押すことで「確認・保存」の枠に反映される．確認ボタンを押すと，第 2 章の図 1 に示すようなキャラクタが表示され，「確認・保存」の枠にある応答の様子を確認することができる．そして保存ボタンを押すと FST 形式の対話が作成され，登録が行われるという仕組みになっている．

6 評価実験

本章では第 5 章で述べたプロトタイプシステムを使用した評価実験と，応答文生成機能やモーションタイミング決定機能など個別の機能ごとに評価実験を行った．なお，すべての評価実験は本学学生である 12 人を対象に実施した．アンケートの回答は 5 段階評価で，1（そう思わない），2（どちらかというとそう思わない），3（どちらともいえない），4（どちらかというとそう思う），5（そう思う）とした．

6.1 プロトタイプシステム全体の評価

6.1.1 実験の目的

従来手法である MMDAgent EDIT と提案システムを比較し，対話の作りやすさや生成した対話の品質について評価することを目的として実験を行った．

6.1.2 実験の方法

以下の 3 つの方法で対話を作成してもらい，作成時間と実験後のアンケートから評価を行う．

従来手法 MMDAgent EDIT

MMDAgent EDIT を使用して認識キーワード，応答文，モーション，声色などのすべての要素を手動で設定することで，対話が作成される．

提案手法 1 提案システムで生成された内容のみ

認識キーワードのみを考えて，応答文を生成するボタンを押し，出力された応答文を確認する．次に感情分類ボタンを押すことでモーション，声色，モーション開始タイミングが自動で設定され，対話が作成される．

提案手法 2 提案システム+被験者の手直し

認識キーワードのみを考えて，応答文を生成するボタンを押し，出力された応答文を確認する．満足できる応答文や，手直しすれば良いと思う応答文が出力された場合，応答文を手直しして設定する．次に感情分類ボタンを押すことで自動で設定されたモーション，声色，モーション開始タイミングを確認し，修正したい項目があれば修正して対話を作成する．

実験の手順，アンケート項目を以下に示す．

- 実験の手順

（1） 被験者に認識キーワードを考えてもらう．
（2） 被験者は手法 1 つにつき，3 回対話を生成する．
（3） 被験者は作成した対話を基に生成されたメイちゃんの応答を確認し，アンケートに回答する．

提案システムを用いる場合には，認識キーワードから文章を生成できる回数は 3 回までとした．

- アンケート項目

(a) 対話の生成が簡単だったか．
(b) モーションと声色の設定が面倒ではなかったか．
(c) またシステムを利用したいか．
(d) 生成された対話に満足できたか．
(e) 提案システムにおいて手直しは必要だと感じたか．
(f) 対話作成の満足度の順位．

対話作成の満足度の順位では 3 手法の比較のため 1 位から 3 位で回答する．

6.1.3 結果と考察

表 5 に被験者による対話の作成にかかった時間の平均を，図 6 にアンケート結果と t 検定を行った結果を示す．$p < 0.05$ の時に有意差があるとし，アスタリスクで表示している．図 7 には項目 (e) のアンケート結果について示し，表 6 に項目 (f) のアンケート結果を表示している．

対話の作成時間は，MMDAgent EDIT と手直しありの提案システムが同程度の時間で，手直しなしの提案システムが最も速い結果となった．手直しなしの場合，応答文を生成する部分でしか時間がかからないため，想定通りの結果である．また，提案システムでは作成時間が良い応答文が出力されるかに依存することが実験を通して分かった．

次にアンケート結果では項目 (a)，(b)，(c) では従来手法と提案システムの間で有意差が認められ，提案手法の方が良い結果となっていると考えられた．特に提案システムの手直しをしない場合では，認識キーワードを決定するだけで簡単に対話を作成することができるため，項目 (a)，(b) で高い評価を得られていると確認できた．一方で，項目 (d) では提案システムの手直しをしない場合の評価が低くなり，従来手法の方が評価が高くなっている．これは提案システムでは応答文を生成する都合上必ずしも良い応答文が出力されるとは限らず，3 回生成して

も良い応答文が得られないといったケースにより値が低くなってしまったと考えられる．この理由として，図 7 を見ると応答文が最も修正が必要な要素として高い値となっている事が挙げられる．しかし，提案システムで手直しありの手法が有意差は見られなかったものの，最も高い結果となったことが分かった．表 6 では対話作成の満足度の順位として手直しありの提案システムが最も良い評価となっている．被験者が当評価実験を行う際に，各項目で悩むケースが散見されたため，自動で多くの項目が設定され手軽に対話を作成できる提案システムの評価が高くなったのではないかと考えられた．

まとめとしては提案システムにより設定項目を自動で決定すれば簡単に対話を作成することができるが，満足できる対話を作成するためには人の目で確認をして編集を行う必要があると考えられた．

表 5: 対話の作成時間

手法	作成時間
MMDAgent EDIT	1 分 31 秒
提案システム (手直しなし)	1 分
提案システム (手直しあり)	1 分 35 秒

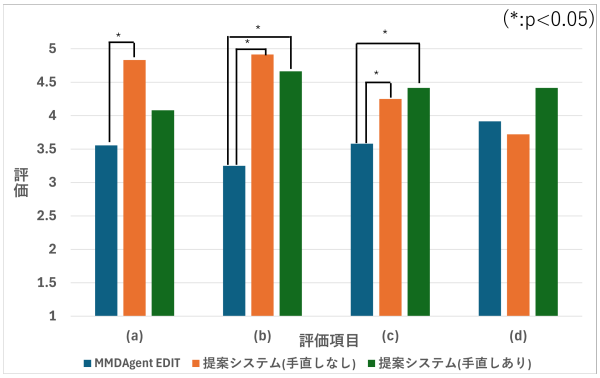


図 6: アンケート結果

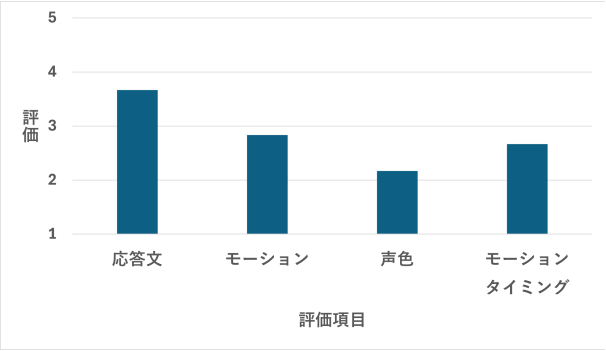


図 7: 項目 (e) のアンケート結果

表 6: 項目 (f) のアンケート結果

手法	評価
MMDAgent EDIT	2.67
提案システム (手直しなし)	2.17
提案システム (手直しあり)	1.17

6.2 応答文生成機能の評価

6.2.1 実験の目的

応答文生成機能では，MMDAgent EDIT で用いられるような案内や説明において分かりやすく，伝わりやすい応答文の生成を目的としている．そこで本実験の目的は，ファインチューニングして作成した応答文生成器が生成する応答文が日本語として正しいか，MMDAgent EDIT で用いられるような内容となっているか評価することである．提案システムで生成する都合上良い文章が出力されるとは限らないため，被験者のアンケートにより評価を行い調査する．

6.2.2 実験の方法

この評価実験の実験の手順，アンケート項目を以下に示す．

- 実験の手順

（1） 評価実験 1 で被験者が使用した認識キーワードを 30 個取得する．

（2） 取得された認識キーワードを作成した応答文生成機能に入力し，1 つの認識キーワードに対して 3 回文章を出力させる．

（3） 被験者を 4 人ずつ 3 グループに分け，1 人につき認識キーワード 10 個に対する 30 文を用意することで 1 文につき 4 人の評価が行われるように調整する．

（4） 被験者がアンケートに回答する．

- アンケート項目

(a) 応答文が日本語として自然か．
(b) 会話として文章の長さが適切か．
(c) メイちゃんらしさを感じるか．
(d) 会話として成立しているか．
(e) 認識キーワードに対応した応答となっているか．
(f) 文法は正しいか．
(g) 会話したいと思えるか．
(h) 応答文の内容は正しいか (地名など)．
(i) 修正が必要だと思うか．
(j) 3 つの文章のうち最も良いと感じた文章はどれか

アンケート項目 (j) は，最も良いと思われる文章に対して評価を行うための項目である．今回の実験では 1 つの認識キーワードに対して 3 回文章生成を行っている．これは提案システムで対話を作成する際には何度か文章を生成してもらうことを想定しているためであることから，1 つの認識キーワードに対して項目 (j) の評価が最も良かった文章に対し，被験者 4 人の評価の平均を取り評価を行った．また，項目 (j) で被験者 4 人による評価が割れた場合には，生成された回数が早い文章に対して評価を行った．

さらに,「メイちゃんらしさ」の項目ではメイちゃんのペルソナとして以下の項目を設定して被験者に伝えている．

（1） Siri や Alexa のような対話型エージェント
（2） 性別は女性で明るい性格
（3） 年齢は 10 代後半から 20 代前半程度
（4） 案内や説明をする応答をすることが多い
（5） 雑談などの会話も行うが、丁寧な応答を行う

6.2.3 結果と考察

用意した認識キーワードと応答文の組の例を表 7 に，12 人の被験者によるアンケート結果の平均を図 8 に示す．まず，項目 (a) と項目 (f) を見ると，生成された応答文は日本語として自然であり，文法的にも正しいと評価されていることがわかる．また，項目 (b)，項目 (d)，項目 (e) から，会話として適切な長さで，認識キーワードに対応した内容の応答が会話として成立していると確認できる．一方で平均の評価が 4 を下回っている項目として，(c) のメイちゃんらしさを感じるかと (g) の会話をしたいと思えるかがある．メイちゃんらしさがペルソナで示した 5 点で与えられているのみであるため，イメージが被験者によって異なってしまったことや，ペルソナの崩壊が発生していたことが原因として考えられる．表 7 を見ると，ペルソナで設定しているように丁寧な口調で案内や説明の文章を出力するように学習出来ていることがわかるが，認識キーワードを「自己紹介」としたとき，最初の 2 回の生成では自分をメイであると認識しているが，3 回目で全く異なる人物を自分の名前として挙げている．このようにしてメイちゃんらしさを損なうような文章が生成されることが確認された．さらに (g) の会話したいと思えるかという項目が低くなっている原因として，文法が正しく認識キーワードにも対応した文章であったとしても，応答が求めている内容ではないというケースが考えられる．認識キーワードが「おすすめのレシピ」の例では，応答内容として料理名を挙げてその紹介をして欲しいと思われる．しかし，出力された文章を見ると，応答としては正しいものの具体的な料理名を生成することはなかった．同様の具体例を求める入力に対して具体的な回答をすることができない場合が見られたので今後の課題として挙げられる．最後に青色で示している項目 (i) 修正が必要だと思うかの項目では 2.28 という値となっている．ここでは日本語として自然で会話として成立した文章が多く生成される一方で，メイちゃんらしさを損なっている文章や，求めているような内容ではないケースがあること，(h) の項目で評価しているように必ずしも地名や店名が存在する正しい内容を出力するとは限らないというような点から修正が必要であると感じた被験者が存在したと考察される．

表 7: 認識キーワードと応答文の組の例

認識キーワード	応答文
日本の観光地	1. 日本で人気のある観光地は、京都です。 清水寺や金閣寺、伏見稲荷大社など、 日本の有名な観光地がたくさんあります。 2. 日本で一番有名な観光地は京都です。 3. 京都には、京都御所や清水寺、金閣寺や 伏見稲荷大社、銀閣寺の庭園など、 日本の歴史や文化を感じられる観光地が たくさんあります。
自己紹介	1. 私の名前は、メイです。よろしくね。 2. はじめまして。私は、メイです。 名古屋工業大学の大学院生です 3. 私の名前は、松倉彩理です。 好きなものは、チョコレートです
おすすめのレシピ	1. 教えてください 2. こんな料理をお勧めします。 3. クックパッドにはお勧めの料理のアイデアが たくさんあります。

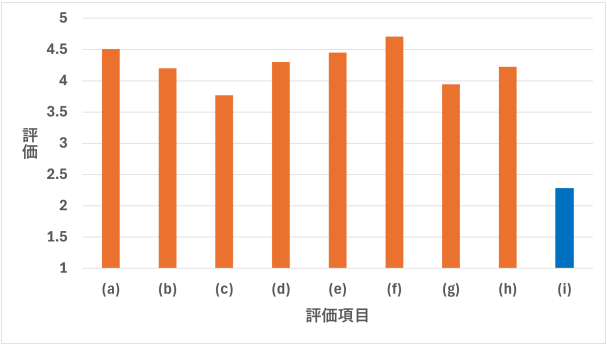


図 8: 応答文生成機能のアンケート結果

6.3 感情分類機能の評価

6.3.1 実験の目的

プルチックの感情の輪に基づいて 32 通りに分類された感情が応答文の感情を表現できているか評価することである．被験者のアンケートにより正解率を計る．また，正解できていない場合でも適切であったかアンケートを行い評価を行う．

6.3.2 実験の方法

この評価実験の実験の手順，アンケート項目を以下に示す．

- 実験の手順

（1） WRIME データセットのテストデータからランダムに文章データを 30 文抽出する．
（2） 被験者を 4 人ずつ 3 グループに分け，1 人につき 10 文を用意することで 1 文につき 4 人の評価が行われるように調整する．
（3） 被験者は 32 通りの感情ラベルから文章に合っていると思う感情ラベルを選択する．
（4） 4 人の予想した感情ラベルで最も票を集めたラベルを正解ラベルとして，感情分類機能の出力と比較し正解率を算出する．

（5） 被験者がアンケートに回答する．

4 人が予想した感情ラベルが 2 人と 2 人で票が分かれた時はどちらも正解ラベルとした．

- アンケート項目

(a) 分類された感情が適切だと思うか．

(b) 文章中に複数の感情もしくは感情の変化があったと思うか．

6.3.3 結果と考察

表 8 に正解率を集計した様子を示す．まず，用意した 30 文のうち 2 人以上の意見が一致した文章は 20 文であった．この 20 文でのシステムによる予測の正解率は 35.0% であった．このケースで正解出来たラベルは喜びや楽観，驚きが正解となった文章であり，高い正解率にはならなかった．次の計算方法では用意した 30 文に対して，人により付与された最大 4 種類の感情ラベルのすべてを正解として正解率を算出したところ，43.0% であった．この集計方法では，絶望や不信感，懸念といった複雑なラベルが正解として挙げられるようになった．計算方法 3 では，計算方法 2 のようにして正解ラベルを決定した 30 文のうち，アンケート項目 (b) での 4 人の評価の平均が 3 未満で評価された文章中で感情の変化が発生していないと思われる文章に対して計算した．この場合の正解率は 46.7%であり，計算した正解率の中では最も高い値となった．また，アンケート項目 (a) の値は 30 文の平均で 3.63 であった．図 9 にこの評価の件数の分布を示す．文章中に感情の変化があると思われるデータも存在するため，評価値が 3 以下となる件数も多く出力されていると考えられる．これらの実験の考察として，人間にとっても文章を 32 個の感情で分類することは困難でありばらつきがあるため，システムの正解率もあまり高くならなかったと考えられる．また，WRIME データセットからテストデータを抽出したため，SNS に投稿されているような複雑な文章が評価対象となっていることも，正解率があまり向上しなかった原因ではないかと感じた．また，人間が文章を見てどのような感情であると思うかは人それぞれであるため，単純な正解率で計算することができなかったと思われる．

表 8: 正解率

	文章数	正解数	正解率
計算方法 1	20	7	35.0%
計算方法 2	30	13	43.0%
計算方法 3	15	7	46.7%

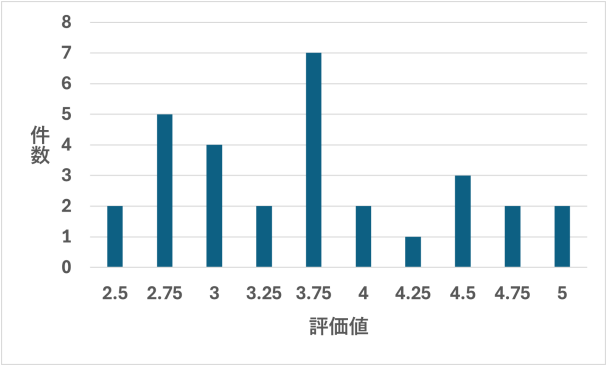


図 9: アンケート項目 (a) の結果

6.4 モーション開始タイミング決定機能の評価

6.4.1 実験の目的

システムから出力されるモーション開始タイミングが，文章中の感情表現が強い位置に着目できているか評価することである．文章中の感情表現が強いと感じる部分は個人差があると考えられるため，被験者のアンケートにより正解率を計る．また，正解できていない場合でも適切であったかアンケートを行い評価を行う．さらに，得られた Attention の値が感情表現が含まれている部分に着目できているかについても評価を行う．

6.4.2 実験の方法

この評価実験の実験の手順，アンケート項目を以下に示す．

- 実験の手順

（1） 感情分類機能の評価と同じ文章である 30 文を用意する．

（2） 被験者を 4 人ずつ 3 グループに分け，1 文につき 4 人の評価が行われるように調整する．

（3） 被験者は文章を見て，句読点やエクスクラメーションマーク，クエスチョンマークで区切られたブロックのどの部分からモーションを開始すべきか選択する．

（4） 4 人の予想したモーション開始タイミングで最も票を集めた位置を正解として，モーション開始タイミング決定機能の出力と比較し正解率を算出する．

（5） 被験者に文章に対して Attention の値を赤色で表示した画像を見せて，アンケートに回答してもらう．

4 人が予想した位置が 2 人と 2 人で票が分かれた時はどちらも正解とした．また，2 人と 2 人で分かれて文章のブロック分けが 2 つのみの場合も正解として計算した．

- アンケート項目

(a) 決定されたモーション開始位置が適切だと思うか．

(b) 感情がこもっている部分を認識できていると思うか．

6.4.3 結果と考察

表 9 に実験に使用した文章例を示す．スラッシュの位置で文章が区切られていることを表している．また，表 10 にブロック数ごとの正解率を示す．ブロック数が 2 の場合，ランダムに選んだ場合と同じ正解率となってしまっている．4 人中 3 人もしくは 4 人の意見が一致しているケースでも正解できていない事があったため，改善の余地があると思われた．今後の課題として，例えば今回の実験では句読点やエクスクラメーションマーク，クエスチョンマークで区切ってブロックを作成しているが，

Attention の値が最も大きい単語からモーションを開始するというような方法が挙げられた．ブロック数が 5，6 の場合は一つも正解できなかったという結果になった．ブロック数が増えると正解だと思われる位置も増えてしまうため，正解率は減少してしまうが，アンケート項目 (a) の結果を見ると 3 を超えており，的外れな位置を推定してはいないことが分かった．またアンケート項目 (a) の 30 文での平均は 3.73 であった．提案システム全体の評価結果やアンケート項目 (e) と合わせて考えると，修正する必要はあまりなく，対話の作成を妨げない程度の予測が可能となっているのではないかと考えられた．ブロック数が 8 と 11 のケースでは文章が長いが，文章中に「楽しい」や「難しい」というワードが含まれており人間とシステムの意見が一致しやすかったのではないかと考えられる．アンケート項目 (b) では評価の平均が 3.69 となった．またアンケート項目 (b) の評価件数の分布は図 10 のようになっており，感情表現が含まれない位置を赤く表示してしまうケースも存在することが分かった．原因としては，文章に感情があまり込められていないケースがあることが考えられた．

表 9: 実験に使用した文章例

文章	ブロック数
1. なんか、/本当に何もやる気なくなるくらい疲弊してる、/心が苦しいな。	3
2. ジェットスターで帰福する。/成田の第 3 ターミナルは普通に混雑してるな。/もっと自粛モードかと思ったが。/マスク率は高め。	4
3. 卒論するか、/寝るか、/どっちにする?	3

表 10: 正解率とアンケート項目 (a) の結果

ブロック数	文章数	正解数	正解率	アンケート項目 (a)
2	6	3	50.0%	3.67
3	10	5	50.0%	4.00
4	6	2	33.3%	3.71
5	4	0	0.0%	3.06
6	2	0	0.0%	3.63
8	1	1	100.0%	4.00
11	1	1	100.0%	4.25

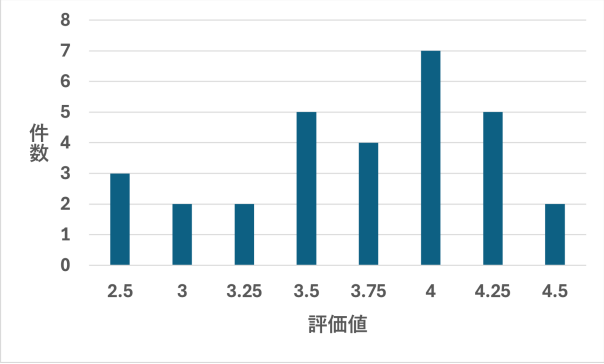


図 10: アンケート項目 (b) の結果

7 おわりに

本研究では応答文生成や感情分類により，一般ユーザが簡単にモーションと声色が付与された，より良い一問一答形式の対話を作成できるようにサポートするシステムを提案した．今後の課題としては，応答文生成の改良や，各機能の精度向上，感情に対応したモーションの作成，モーション開始タイミングの調整などが挙げられる．

文　献

[1] Siri, https://www.apple.com/jp/siri/ (2024.1.9 参照)
[2] Alexa, https://developer.amazon.com/ja-JP/alexa (2024.2.15 参照)
[3] A.Lee, K.Oura, and K.Tokuda, MMDAgent - A fully open-source toolkit for voice in-teraction systems, Proceedings of the ICASSP 2013, pp. 8382–8385, 2013.
[4] Ryota Nishimura, Daisuke Yamamoto, Takahiro Uchiya and Ichi Takumi, Web-based environment for user generation of spoken dialog for virtual assistants, EURASIP Journal on Audio, Speech, and Music Processing 2018, Article number17, 2018.
[5] 古市瑞希，山本大介，高橋直久，ブロック型ビジュアルプログラミング機能を有する音声対話シナリオ編集システム，情報処理学会論文誌デジタルコンテンツ（DCON） 8 巻 2 号，2020, pp.1–15
[6] 藤後英哲，大浦杏奈，菊池浩史，システム発話の感情分類による制御を行ったマルチモーダル対話システム， 第 96 回研究会言語・音声理解と対話処理研究会，pp.206–210，2022
[7] mecab-ipadic-NEologd, https://github.com/neologd/mecab-ipadic-neologd (2024.1.9 参照)
[8] rinna/japanese-gpt-neox-3.6b, Hugging Face, https://huggingface.co/rinna/japanese-gpt-neox-3.6b (2024.1.9 参照)
[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, LoRA:Low-Rank Adaptation of Large Language Models, International Conference on Learning Representations, 2022.
[10] optuna-A hyperparameter optimization framework, https://optuna.org/ (2024.1.10 参照)
[11] cl-tohoku/bert-base-japanese-whole-word-masking, Hugging Face, https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking (2024.1.9 参照)
[12] WRIME: 主観と客観の感情分析データセット， https://github.com/ids-cv/wrime (2024.1.9 参照)
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.4171–4186, 2019.
[14] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, Yoshua Bengio, A Structured Self-attentive Sentence Embedding, Conference paper in 5th International Conference on Learning Representations, 2017.

経路単純化と案内文単純化を組み合わせた経路案内文要約システム

鋤柄　知哉†　　金　　鎔煥†　　山本　大介†

† 名古屋工業大学大学院工学研究科　〒 466–8555 名古屋市昭和区御器所町
E-mail: †t.sukigara.939@stn.nitech.ac.jp, ††{kim,yamamoto.daisuke}@nitech.ac.jp

あらまし　近年, サイネージ型経路案内システムの利用が進んでいる. 目的地までの経路案内では, 具体的な方向転換の指示を逐一提供する案内文が主に利用されるため, 経路が複雑になるにつれて案内文も長く複雑になってしまう. これを解決するため, 本研究では複雑な案内文を要約する手法を提案する. 特に, 案内文の文数を削減し, 必要性の低い語句を省略することで, 問題の解決を目指す. 案内文の文数を削減するため, 経路から右左折地点を取捨選択することで単純化経路を生成し, その経路に基づいて案内文の生成を行う. 案内文の中で必要性の低い語句を省略するため, 案内文を意味の成す要素に分解し, 重みと価値を設定する. 重みの合計がユーザー指定の制限内に収まり, 価値の合計が最大化されるように要素を組み合わせることで要約案内文の生成を行う.

キーワード　文書要約, 地理データ活用, 経路案内, デジタルサイネージ, 観光

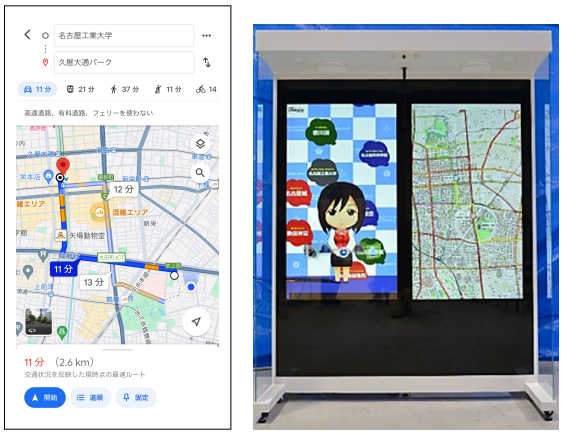
1　はじめに

近年, Google Maps [1] や Yahoo!地図 [2] などの地図アプリケーションが広く利用されている. これらのシステムは, 多くの人が所持しているスマートフォンで利用でき, 細やかな案内によって誰もが見知らぬ土地でも目的地へと容易に到達することが出来る.

一方で, デジタルサイネージを用いたサイネージ型経路案内システムも注目を集めている. これは地方創世や都市開発の活発化に伴い, デジタルサイネージ [3] といったデバイスが普及したことが影響している. 例えば, 東京都の国分寺市は国分寺駅にデジタルサイネージを導入し, 駅周辺の観光案内等に利用されている [4]. また, 公益財団法人 東京観光財団は都内 40ヵ所の街頭などにデジタルサイネージを設置し, 外国人観光客向けに多言語による東京観光案内を提供している [5]. 本研究室でも「しゃべるバス停」[6] といったシステムを開発し, 大画面で視覚的に目的地周辺の詳細を提供するとともに, 可愛らしいキャラクターと音声を併用することで, 誰でも親しみやすい経路案内を実現している. しかし, デジタルサイネージはスマートフォンのような持ち運びが出来るデバイスとは異なり, 固定の場所に設置して利用されるため, ユーザーはその場で案内文を覚える必要があるといった特徴を持つ.

このような経路案内システムで主に利用される案内文生成手法として, Turn-by-Turn 方式 [7] が挙げられる. この案内文生成手法では, 具体的な方向転換の指示を逐一提供するため, 目的地までの行動を明確に把握することが可能である. しかし, その特性上, 経路が複雑になるにつれて, 案内文も長く複雑になるといった特徴を持つ.

上記の 2 つの特徴を考慮すると, Turn-by-Turn 方式で生成された複雑な案内文をサイネージ型経路案内システムで利用する場合, ユーザーが案内文をその場で全て記憶することが非常に困難になるといった問題が生じる. また, スマートフォンの地図アプリではなくサイネージ型経路案内システムを利用するユーザーは, 目的地までの詳細な道順よりも, 目的地までの大まかなルートの概要を把握することに重きを置くと考えられる. そのため, サイネージ型経路案内システムで利用される経路について, 全ての詳細な案内文を音声で提供する必要性は低いとも考えられる.



(a) GoogleMaps　　**(b)** しゃべるバス停

図 1: 様々な経路案内システム

本研究の目的は, Turn-by-Turn 方式で生成される詳細な案内文をユーザー指定の文章量にまで要約を行う経路案内文要約手法を提案することである. これによって生成された要約案内文をユーザーに提示することで, 経路の概要を理解するために必要な情報のみを提供し, 迅速かつ容易な経路の理解を助けることが出来る. また, サイネージ型経路案内の一般的な課題である「その場で複雑な案内文を記憶する必要性」を克服し, より直感的で使いやすい案内文を提供することが可能となる.

このとき, 以下の 2 つのアプローチを取ることで, 要約案内文の実現を目指す.

＜アプローチ 1＞

経路を単純化することで，案内文の文数を削減する

Turn-by-Turn 方式の案内文の文数を削減する．これにより，経路の概要把握に役立ちにくい情報を出来るだけ減らし，ユーザーにとって重要だと判断される情報を優先的に提供する案内文を採用することが可能となる．

＜アプローチ 2＞

案内文を方角や距離などの要素に分解し，必要性の低い要素を削除する

生成された案内文を方角や距離などの要素に分解し，5 種類の情報カテゴリーに分類する．それぞれの要素に対して「重み」と「価値」を設定し，以下の 2 点を満たすように案内文の要素の取捨選択を行う．

- 指定した制限内に要素の「重み」の合計が収まる
- 選択された要素の「価値」の合計を最大化する

これにより，案内文の文章から重要な情報を優先的に選び出し，相対的に役立ちにくい情報を削除することが可能となる．

2 関 連 研 究

サイネージ型経路案内システムの充実を目指すため，多岐にわたる研究が行われているが，経路案内をより効率的に行うために不可欠な経路案内文の要約に関する研究は未だ十分でない．

原ら [8] は，経路単純化アルゴリズムである Visvalingam-Whyatt アルゴリズム [9] を用いて単純化経路を生成し，その経路に基づいて案内文の生成を行う．また，それによって生成された案内文を方角や距離などの要素に分解し，それらの要素を 5 つの情報カテゴリーに分類する．この 5 つの情報カテゴリー毎に価値の設定を行い，ユーザーの指定する文字数制限内に収まるようにして要約案内文の生成を行う．この研究は案内文を要約する上で非常に画期的なアプローチをとっているが，いくつか問題も保持している．例えば，Visvalingam-Whyatt アルゴリズムは隣接する三角形の面積に基づいて経路を単純化するため，要約案内文を生成する上で役立つ右左折地点の地理情報を考慮できないといった問題がある．また，情報カテゴリー毎に価値の設定を行っているため，情報カテゴリーに複数の要素を保持することとなり，案内文の要素の柔軟な取捨選択が行えないといった問題も生じている．

鋤柄ら [10] は，文献 [8] が抱える「経路を案内する上で重要な右左折地点の地理情報を考慮できない」といった問題を解決するため，「道路名」を経路案内における重要な情報とみなし，その経路に重点を置き案内文を生成することで要約案内文の実現を目指している．

一方で，経路案内文に限らず，一般的な文章の要約については広範囲にわたる研究が行われている．

Yang ら [11] は，BERT(Bidirectional Encoder Representations from Transformers) という高度な言語モデルを文章要約タスクに応用し，抽出型および生成型の要約モデルに適用している．また，エンコーダとデコーダに異なる最適化手法を用いることで，学習過程の調整を行い，要約の品質を向上させる手法を提案している．

Ramesh ら [12] は，リカレントニューラルネットワーク (RNN) を基盤とするエンコーダ・デコーダモデルに基づいた文章要約システムを提案している．このモデルは，テキスト要約の主要な課題であるキーワードのモデリングや文書の階層的構造の理解，稀な単語や未確認の単語への対応に重点を置いており，従来システムよりも精度の高い要約が可能となっている．

一般的に，生成型要約の実現法は大きく分けて，「ルールベース」と「深層学習を用いた文章の要約」の 2 種類に分類できる．関連研究として挙げた文献 [11] [12] では，深層学習を用いた文章の要約システムが提案されている．このようなシステムは，既存の文章に基づいてユーザー満足度の高い要約文の生成を実現する．しかし，経路案内文の要約においては，地理的な構造や要約時に役立つ特定の”地理情報”など，既存の文章に全く含まれない情報を考慮すべき状況がしばしば存在する．そのため，現時点では経路案内文を要約する上で，深層学習による文章要約を適用することが必ずしも最適であるとは限らない．一方で，ルールベースによる要約手法の利点は，重要と判断される特定の”地理情報”などに基づいて，独自の要約ルールを設定できることにある．例えば，地理的な構造や経路の概要把握に役立つ特定の”地理情報”を意図的に重視し，経路に関連性の低い情報を削除することで，ユーザーが経路の重要なポイントに注目できる要約案内文を生成することが可能となる．将来的には深層学習の技術が進化し，先述の課題を克服する可能性も考えられるが，現段階ではルールベースの要約手法を採用することが望ましいと考えられ，文献 [8], [10] などもルールベースの要約手法に該当する．上記を踏まえて，本研究でもルールベースを採用することで要約案内文の実現を目指す．

3 提案システム

3.1 使用するデータベース

道路データベース

道路データは OpenStreetMap の地図データから作成され，ポイント，アーク，ノードの 3 つのデータによって構成されている．ポイントは，道路上の点を表し，位置座標（緯度，経度）からなる．道路の形状は道路上の隣接ポイントを順番に繋ぐことで形成され，折れ線で近似的に表現される．アークとは，始点ノードから終点ノードに至る道路であり，道路上のポイントの系列で表す．ノードとは，道路上の交差点であり，隣接したノードとそのアークをまとめたものをリンクという．また，表 1 に示すように，リンクは「高速道路」や「国道」など，14 種類の異なった道路クラスに分類される．

ストロークデータベース

ストロークとは，認知心理学に基づき道路ネットワークをグルーピングしたものであり，道なりに続く道路を表す [13] [14]．本研究室が保持するストロークデータベースにはストロークテーブルが存在する．そのデータ構造を表 2 に示す．

表 1: 道路クラス

clazz	kmh	説明
11	120	motoway(高速道路)
12	30	motorway link(高速道路への連絡路)
13	90	trunk(国道)
14	30	trunk link(国道への連絡路)
15	70	primary(主要地方道)
16	30	primary link(主要地方道への連絡路)
21	60	secondary(一般地方道)
22	30	secondary link(一般地方道への連絡路)
31	40	tertiary(2 車線以上の一般道)
32	50	residential(住宅区域内道路)
41	30	road(その他の道路)
42	30	unclassified(2 車線未満の一般道)
51	5	service(敷地内道路)
63	7	living street(生活道路)

表 2: ストロークテーブル

カラム名	データ型	説明
stroke_id	int	ストロークの ID
stroke_length	double	ストロークの長さ
stroke_line	geometry(LineString)	ストロークの形状

3.2 提案システムの流れ

提案システムの流れを図 2 に示す.

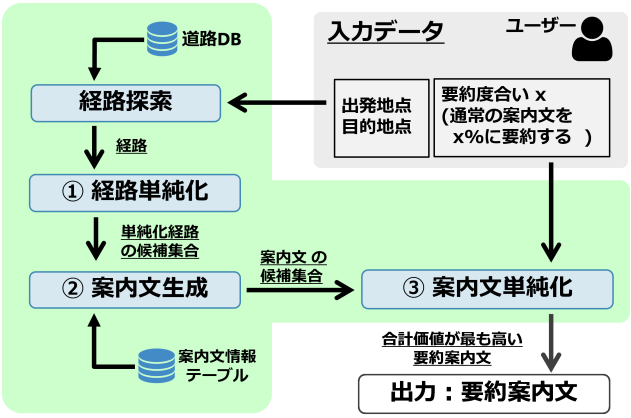


図 2: 提案システムの流れ

まず初めに, ユーザーが入力した出発地点と目的地点から道路データベースの情報を用いて, 経路の探索を行う. 提案システムでは, 目的地まで最も短い距離で到達できる最短経路を採用しているが, 目的に応じて経路は変更可能である. 次に, 獲得した経路に対して, 右左折地点の取捨選択を行う経路単純化機能を実行することで, 単純化経路を取得する. ここでは, 複数の単純化経路を取得し, 単純化経路の集合を出力とする. 次に, 獲得した単純化経路に対して, 案内文情報を保持したデータセット等を用いて, 案内文の生成を行う. ここでは, 全ての単純化経路について案内文を生成し, 案内文の集合を出力とする. 最後に, ユーザーから入力された文章量の制限を受けて, 案内文の集合に含まれる案内文それぞれに対して案内文単純化機能を実行し, 全ての案内文の要約案内文を獲得する. それらの要約案内文の中で価値ある情報を最も多く組み込むことが出来た要約案内文を最終的な出力とする.

3.3 経路単純化機能

経路単純化機能では, 経路探索機能で獲得した経路 (ノード集合 V) に含まれるノード v を取捨選択することによって単純化経路 (ノード集合 V') を獲得する. また, ここでは複数の単純化経路を取得し, 単純化経路の集合 $\mathcal{V}$ を最終的な出力とする. この機能と後に述べる案内文生成機能を合わせて実行することで, Turn-by-Turn 方式の案内文に含まれる「右左折地点の案内」を減らすことが可能となり, 全体の案内文の文数削減に繋がる. 本研究では, 単純化経路の集合 $\mathcal{V}$ の取得方法として, 以下の 2 種類の手法を提案する.

＜ **手法 1：$vScore$ による逐次探索** ＞

この手法では, ノード集合 V に含まれるノード v に対して, スコア $vScore$ を設定する. このとき, $vScore$ が最も小さい値を取るノード v_i を逐次的に削除することで単純化経路 V' を生成し, 単純化経路の集合 $\mathcal{V}$ に V' をその都度追加していく. これを単純化経路 V' のノード数が 3(始点, 終点を含む) となるまで繰り返す.

ここで, $vScore$ を説明する上で必要な前提知識として, 後に述べる案内文生成機能, 案内文単純化機能の概要を説明する. 案内文生成機能では, 単純化経路に基づいて案内文を生成する. 生成された案内文は方角や距離などの意味を成す要素 (例：”南西に”, ”300m”, ”進み”, ”久屋大通で”, ”左折.”) に分解され, 5 つの情報カテゴリー (”最低限の情報”, ”総距離, ”所要時間”, ”距離”, ”地理情報”) に分類される.

次に, 案内文単純化機能では分類した情報カテゴリー毎に「価値」を設定し, 「価値」の高い情報が要約案内文に優先的に含まれるように案内文の要素の取捨選択を行う. このとき, 分類した 5 つの情報カテゴリーの中でも, ユーザーが経路案内に含めて欲しいと考える程度はそれぞれ異なると考えられる. それを検討した実験の結果より, ”最低限の情報”を除いた 4 つのカテゴリーの中でも, 特に”地理情報”の重要度が高いことが明らかになっている. ここで, ”地理情報”とは右左折地点の地理的特徴に基づいて生成される情報であり, 先ほどの例では”久屋大通で”が該当する. 本研究では 5 種類の地理的特徴 (”つきあたりである”, ”道路名を持つ道路の一部”, ”施設名を持つ建物が近くにある”, ”広い道路の一部”, ”前の右左折地点から何個目の交差点である”) に基づき, ”地理情報”を生成している. この 5 種類の”地理情報”における「ユーザーが経路案内に含めてほしいと考える程度」も, 先ほどと同様の実験で検証し, 重要度として獲得している. 因みに, 「一つ前の右左折地点 v_{i-1} が存在する場合の”地理情報”」と「一つ前の右左折地点 v_{i-1} が存在しない場合の”地理情報”」は区別し, それぞれ重要度も異なっている.

他方で, 文献 [8] でも考慮していた通り, 経路の中でも微細な構造を持った右左折地点よりも, 経路の構造を特徴付ける大きな構造を持った右左折地点の方が経路概要を把握する上で案内

の優先度は高いはずである．そのため，経路構造の情報も考慮して価値を設定し，案内文の要素を取捨選択する必要がある．

これらを考慮して，”地理情報”に該当する要素の「価値」は，情報カテゴリー毎に比較することで獲得した重要度を基準に”地理情報”の重要度と経路構造の情報に基づいて調整される．

上記を踏まえて，「経路案内に役立ちにくい地理的特徴しか保持せず，経路構造が微細な右左折地点」を逐次的に削除することで，ユーザー満足度の高い要約案内文を作成することが出来ると考えられる．このとき，$vScore$ を以下のように定義する．

$$\mathbf{vScore(v_i)} = \frac{\mathbf{Y_i \cdot S_i}}{\mathbf{Z_i}}$$

ただし，

$$v_i : V \text{ に含まれる } i \text{ 番目のノード}$$

とする．

Y_i は「現在採用されている”地理情報”の重要度」を表す．この値が大きければ大きいほど，右左折地点 v_i は「案内文に残すべき”地理情報”を持った右左折地点」と考えることが出来る．

S_i は「v_i と隣接するノードによって構成される三角形の面積」を表し，経路単純化アルゴリズムの Visvalingam-Whyatt アルゴリズムを参考に作成している．この値が大きければ大きいほど，右左折地点 v_i は「経路の中でより大きな構造を持った右左折地点」と考えることが出来る．

Z_i は「v_i が削除されることで v_i の後に存在する v_{after} で採用される”地理情報”の重要度」を表す．この値が大きければ大きいほど，「右左折地点 v_i が削除されたとしても，v_{after} において経路案内に役立つ”地理情報”が保持される」と考えることが出来る．

このアプローチを取ることで，後に述べる手法 2 と比較して計算時間がほとんどかからないにも関わらず，案内文生成機能において，ユーザーが比較的満足のいく案内文を生成出来ると考えられる．しかし，経路単純化を行うことによって随時，経路構造や採用される”地理情報”が変化していき，$vScore$ もその都度変化していくため，手法 2 によって生成される合計価値の最も高い要約案内文を確実に生成することは困難となる．

＜手法 2：総探索＞

この手法では，ノードを取捨選択することで生成し得る全ての単純化経路を取得する．ただし，始点と終点を含んだ単純化経路のノード数は 3 以上であることを条件とする．

このアプローチを取ることで，全ての単純化経路に対する要約案内文を生成することが可能となる．これにより，ユーザーが指定した情報量の制限下において，合計価値の最も高い要約案内文の生成を確実に実現できる．しかし，このアプローチではノード数が極端に多い経路の要約案内文を生成する場合，計算量が指数関数的に増加してしまうといった問題が生じる．

3.4 案内文生成機能

案内文生成機能では，単純化経路の情報に基づき，案内文を生成する．ここでは，単純化経路の集合を入力とし，全ての単純化経路の案内文を最終的な出力とする．単純化経路に基づいて案内文を生成することで，案内文の文数そのものを削減し，経路の概要把握に役立つ案内文のみを採用することが可能となる．以下，案内文生成の流れを示す．

手順 1：　経路の基本情報を取得し，案内文を生成

経路から”総距離”や”所要時間”などの基本情報を取得し，案内文を生成する．

(例：総距離は 1.1km，所要時間は 14 分)

手順 2：　右左折地点 v_i の情報を取得し，案内文を生成

右左折地点 v_i の地理的特徴から”地理情報”，v_i の前に存在する v_{before} から v_i までの”方角”や”距離”等を取得し，案内文を生成する．

(a)　右左折地点 v_{i-1} が存在する場合

i. v_{i-1}，v_i 間の方角と距離情報を取得．(例：東，200m)

ii. Turn-by-Turn 方式の案内で役立つ v_i の地理情報を取得．(例：つきあたり)

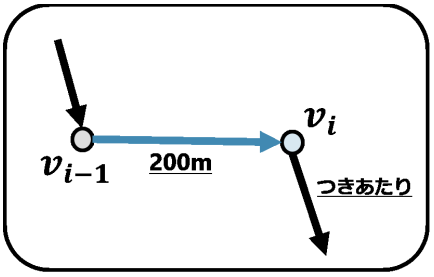


図 3: 右左折地点 v_{i-1} が存在する場合

(b)　右左折地点 v_{i-1} が存在しない場合

i. v_i の前に存在する v_{before}，v_i 間の方角と距離情報を取得．(例：南東，500m)

ii. v_{i-1} が省略時に役立つ v_i の地理情報を取得．(例：久屋大通の一部)

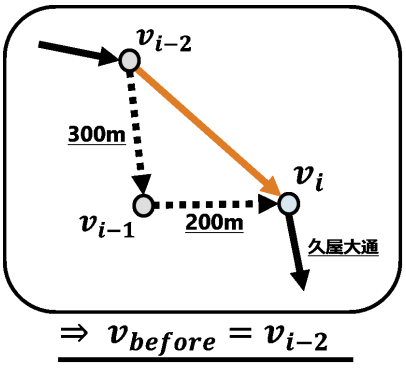


図 4: 右左折地点 v_{i-1} が存在しない場合

このように生成された案内文が「総距離は 1.1km．所要時間は 14 分．東に 200m 進み，つきあたりを右折．南東に 500m 進み，久屋大通で右折．まっすぐ 400m 進む．」であるとき，後に述べる案内文単純化機能を行うため，案内文を方角や距離などの要素に分解し，それぞれを表 3 に示すような 5 つの情報カテゴリーに分類する．

表 3: 案内文の要素を情報カテゴリー毎に分類

情報カテゴリー	案内文の要素
最低限の情報	東に, 進み, 右折, 南東に, 進み, 右折, まっすぐ, 進む
総距離	総距離は 1.1km
所要時間	所要時間は 14 分
距離	200m, 500m, 400m
地理情報	つきあたりを, 久屋大通で

3.5 案内文単純化機能

案内文単純化機能では，生成された案内文から必要性の低い語句を省略し, 要約案内文を生成する．ここでは入力として得られた全ての案内文に対して要約案内文を生成する．その中でも合計価値が最も高い要約案内文を最終的な出力としてユーザーに提示する．これを実現するため, 以下のような手順を踏む.

< 手順 1>
分解した案内文の要素に「価値」と「重み」を設定.

< 手順 2>
以下の 2 点を満たすように, 要素の組み合わせを求める.

- ユーザー指定の制限内に要素の「重み」の合計が収まる
- 選択された案内文の要素の「価値」の合計を最大化する

< 手順 3>
残った案内文の要素を組み合わせて, 要約案内文を構成.

3.5.1 重みの設定

要素の重みは, 意味の成す情報の区切りごとに 1 単位として設定する．例えば，「総距離は 1.1km．所要時間は 14 分．東に 700m 進み, つきあたりを右折．まっすぐ 400m 進む.」といった案内文があった場合, { 総距離は, 1.1km, 所要時間は, 14 分, 東に, 700m, 進み, つきあたりを, 右折, まっすぐ, 400m, 進む.} の区切り毎に 1 単位として設定する.

3.5.2 価値の設定

要素の価値は案内文生成時に分類した情報カテゴリー毎に設定する．ここで, 5 つの情報カテゴリーのうちの一つである”最低限の情報”は, 案内文に必須の情報であるため, 価値を設定することなく案内文に確定で導入するものとする．その他 4 つの情報カテゴリーである”総距離”, ”所要時間”, ”距離”, ”地理情報”は, ユーザーにアンケート調査を取ることで価値を設定する．ここでは, 定量的に選択肢の優劣を判断する手法として用いられる階層分析法 [15] を利用し, それによって獲得される重要度を情報カテゴリー毎の価値として設定する．また, 情報カテゴリーの”距離”と”地理情報”は複数の要素を保持する可能性があるため, それぞれ価値の調整を行う．表 4 に要素毎の価値設定についてまとめる.

”距離”や”地理情報”で用いた指数のパラメータは, ユーザー実験の結果から $\rho = \delta = \tau = 0.1$ を暫定解とする．X_i は「案内文に含まれる全ての”距離”の平均に対する i 番目の”距離”の大きさ」を示す．これにより, 平均よりも長い距離である要素に高い価値を設定することが出来る．Y_i は「案内文に含まれる”地理情報”の重要度の平均に対する i 番目の”地理情報”の重要度の大きさ」を示す．S_i は「隣接する右左折地点によって作成される三角形の面積の平均に対する i 番目の三角形の面積の大きさ」を示す．これらを用いて”地理情報”の価値を調整することで, 案内文に含まれる”地理情報”の中でも重要度が高く, 大きな構造を持つ”地理情報”に高い価値を設定することが出来る.

表 4: 情報カテゴリーに含まれる要素毎の価値設定

情報カテゴリー	価値
総距離	0.05
所要時間	0.13
距離	0.15 $(X_i)^{\rho}$
地理情報	0.67 $(Y_i)^{\delta} \cdot (S_i)^{\tau}$

3.5.3 案内文の要素の取捨選択

通常の Turn-by-Turn 方式によって生成された案内文の合計単位数 w を算出し, 入力として与えられたユーザー指定の要約度合い x を掛け合わせることで得られる単位数を $W = xw$ とする．この値 W を要約案内文として生成できる最大単位数として設定する．この制限内に収まることに加えて, 価値の合計を最大化するように動的計画法によって案内文の要素の取捨選択を行う．これにより, 案内文に含まれる要素全てに対して, 要約案内文に含まれる要素かどうかの判定を行う．以下, 目的関数と制約条件を示す.

目 的 関 数

$$\max \sum_{value_j^g \in g_j} value_j^g x_j \tag{1}$$

ここで, G を価値と重みが設定される案内文の要素 g の集合とし, j 番目の案内文の要素 g_j は $(item_j^g, value_j^g, weight_j^g) \in g_j$ と定義する．$item_j^g$ は j 番目の要素の文字列, $value_j^g$はその価値, $weight_j^g$ はその重みを表す．x_j は $x_j \in \{0, 1\}$ として定義される二値変数であり, 案内文の要素 g_j が要約案内文に含まれるかどうかを示す.

制 約 条 件

$$\sum_{weight_j^g \in g_j} weight_j^g x_j + \sum_{weight_k^m \in m_k} weight_k^m \leqq W \tag{2}$$

ここで, M を”最低限の情報”に含まれる要素 m の集合とし, k 番目の”最低限の情報”の要素 m_k は $(item_k^m, weight_k^m) \in m_k$ と定義する．$item_k^m$ は k 番目の要素の文字列, $weight_k^m$ はその重みを表す．W は要約案内文として生成できる最大単位数である.

ここで得られた結果を踏まえて, $x_j = 1$ である案内文の要素と”最低限の情報”に含まれる案内文の要素を組み合わせることで要約案内文を生成する．獲得している全ての案内文に対して要約案内文を生成し, 生成された要約案内文の中で合計価値が最も高い要約案内文を最終的な出力とする.

4 プロトタイプシステム

4.1 システムの概要

3 章で述べた提案システムを実現するため，Java を用いてプロトタイプシステムを実装した．開発環境として Eclipse, Windows11，データベースを扱う RDBMS として PostgreSQL を利用している．プロトタイプシステムで使用しているデータベースは，研究室サーバ内にある道路データベースとストロークデータベースの 2 種類である．

4.2 システムの動作

プロトタイプシステムを起動すると，地図画像が表示される．出発地点と目的地点をクリックして入力することで最短経路が取得される. 図 5 のように，赤い線で最短経路が地図上に描画され, 図, 6 のようにコンソール上には通常の案内文と要約案内文が生成される．

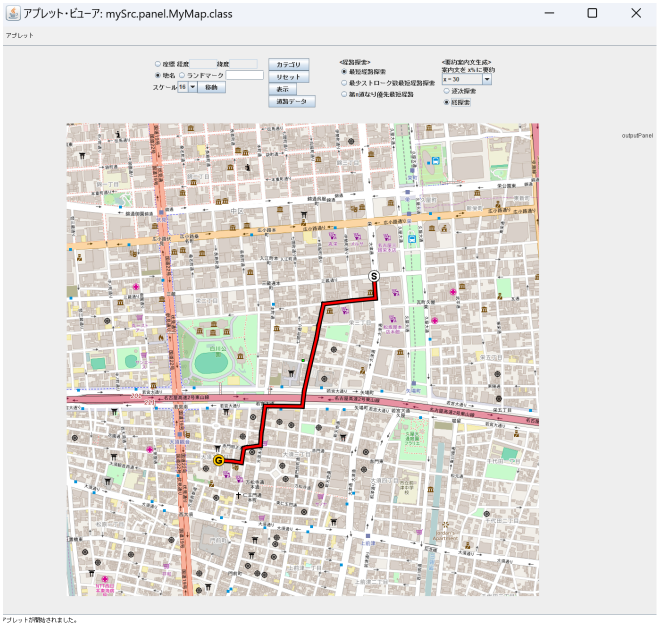


図 5: プロトタイプシステムの実行画面

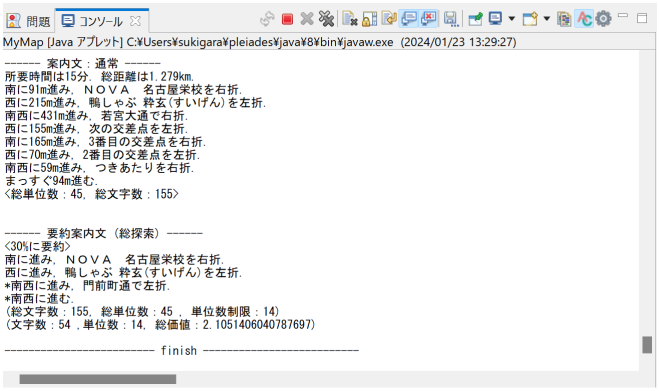


図 6: 通常の案内文と要約案内文

5 評 価 実 験

5.1 評価実験 1

実 験 概 要

この実験の目的は, 従来の経路単純化機能に対する改善が有効かどうかを検証することである．被験者 (本学学生 12 名) に対し, 経路と 3 種類の要約案内文を提示し, それらについて 5 段階評価のアンケートを実施する．このとき, 経路は始点と終点を含んだノード数が 10 の最短経路を用いる．この経路に対して, 従来の経路単純化機能と新たに提案した 2 種類の経路単純化機能を適用する．これによって生成された単純化経路に基づき, 案内文の生成を行い, さらに提案手法の案内文単純化機能を実行することで, 3 種類の要約案内文を生成する．本実験では, 8 種類の最短経路を取得し, 4 つの経路毎に案内文の要約度合いを 50%, 30%へと変更する．要約度合いの変化により, アンケート項目の評価にどのような影響が生じるかも合わせて調査する．具体的な要約手法を表 5, 比較対象を表 6, 経路と要約手法の内訳を表 7 にまとめる．

表 5: 要約手法について

種類	具体的な要約手法
従来手法	経路単純化 (Visvalingam-Whyatt アルゴリズム) ＋ 案内文単純化
提案手法 1	経路単純化 (逐次探索) ＋ 案内文単純化
提案手法 2	経路単純化 (総探索) ＋ 案内文単純化

表 6: 比較対象

パターン	経路	要約の度合い	要約手法
1 - 1	最短経路	50%	従来手法
1 - 2			提案手法 1
1 - 3			提案手法 2
2 - 1		30%	従来手法
2 - 2			提案手法 1
2 - 1			従来手法 2

表 7: 経路と要約手法の内訳

人数	経路 1 ～ 4	経路 5 ～ 8
12 人	パターン 1-1, 1-2, 1-3	パターン 2-1, 2-2, 2-3

アンケート項目

アンケート項目を以下に示す.

- 目的地に到達できると思う
- 経路の概要を把握できた
- 目的地へ向かい不安を解消できた
- この案内文を直感的に理解できた
- 経路を記憶するのに役立った
- 経路案内を利用時に知りたいと思う情報が得られた
- この案内文に満足した
- 再度、このような案内文を利用したい
- 通常の案内文よりも経路の概要把握に役立つ
- 通常の案内文よりも短時間で理解が出来た

結果と考察

アンケート項目に対する回答の平均値を図 7 から図 9 に示す．これらのグラフより，全ての項目で従来手法よりも提案された 2 種類の手法が高い評価を得ていることが分かる．要約度合いが 50%，30%の場合であっても，従来手法では”地理情報”の重要度を考慮せずに経路を単純化してしまうため，右左折地点 v_i を案内する際に v_{i-1} が存在しない場合に役立つ”地理情報”を考慮せずに要約案内文が生成される傾向にある．この点が全ての項目において，従来手法の評価が相対的に低くなる一因として考えられる．

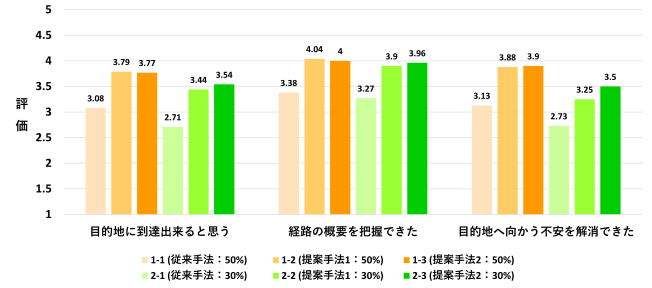


図 7: アンケート結果 項目：1-3

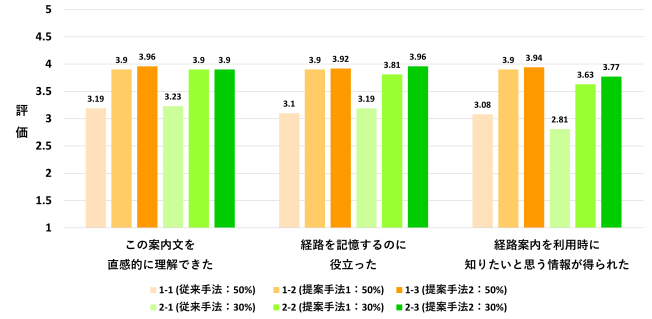


図 8: アンケート結果 項目：4-6

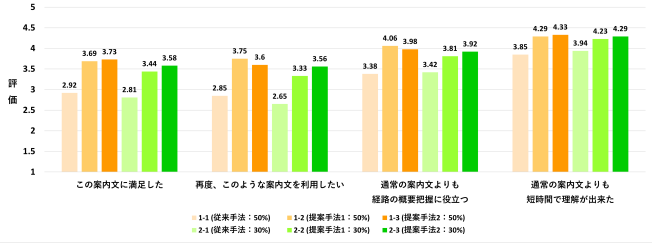


図 9: アンケート結果 項目：7-10

要約度合いが 50%の場合，どちらの提案手法も高い評価を得ているが，特に項目 2, 9, 10 では評価が 4 を超える結果が得られている．また，要約度合いが 30%の場合でも同様に，両方の提案手法は高い評価を維持し，特に項目 2, 5, 10 が良い評価を得ている．これらの結果から，どちらの提案手法を用いても，経路の概要把握や短時間での理解に繋がる可能性が高いと見られる．そのため，提案手法 1,2 によって生成される要約案内文は，その場で案内文を記憶する必要のあるサイネージ型経路案内システムで用いる案内文として適していることが示唆された．

また，要約度合いが 50%の場合，どの項目でも 2 種類の提案手法の評価に大きな差は見られなかった．これは，要約された案内文の文数が大幅に削減されなかったため，経路の概要把握に役立つ地理的特徴を保持しつつも，生成された要約案内文に大きな違いが生じなかったためだと考えられる．要約度合いが 30%の場合，提案手法間の差が徐々に顕著になり，特に経路単純化 (総探索) を用いた要約案内文の方が高い評価を得た．これは経路単純化 (逐次探索) では得られなかった価値の高い”地理情報”を経路単純化 (総探索) では獲得し，要約案内文に反映できたためだと考えられる．したがって，文章量をより減らしたい場合は，経路単純化 (総探索) を用いた提案手法 2 が適しているといえる．ただし，評価実験 2 で示している通り，ノード数が増えるにつれて実行時間が指数関数的に増加するため，アルゴリズムの効率化が今後の課題として挙げられる．

最後に，本研究ではどのような人でも利用出来るサイネージ型経路案内システムの充実を目指すため，道路名や施設名といった特定の名称を重要視して要約案内文を生成している．しかし，個人が利用するスマートフォンやカーナビゲーションなどのシステムにおいては，本研究のアプローチに加えて，過去に訪れた場所などの履歴に基づくパーソナライズされた情報を取り入れることで，システムのさらなる進化が期待できる．例えば，目的地へ向かう際，通勤途中によく通る一般道路を経由する場合，そこまでにどれだけ複雑な経路を進む必要があったとしても，そこまでの詳細な説明は不要となる．また，目的地の近くに馴染みのあるレストランがある場合，そこまでの経路が全体の大部分を占めていたとしても経路の概要を説明する必要性は低いはずである．このように，個人の利用履歴を考慮した要約案内文の生成が実現できれば，システムの普及と多くの人々による利用がさらに期待できる．

5.2 評価実験 2

実験概要

この実験の目的は，2 種類の提案手法について，それぞれの平均実行時間を計測することである．本研究では，2 種類の経路単純化機能と 1 種類の案内文単純化機能を提案している．これらを組み合わせた 2 種類の提案手法について，それぞれの実行時間を計測する．具体的には，始点と終点を含むノード数が 4 から 25 までの様々な経路を取得し，それらの経路に基づいて要約案内文の生成を行う．ノード数毎に 10 経路分の要約案内文を生成し，これらの平均を実行時間として記録する．比較対象となる手法を表 8 に示す．

表 8: 比較対象

種類	具体的な要約手法
提案手法 1	経路単純化 (逐次探索) + 案内文単純化
提案手法 2	経路単純化 (総探索) + 案内文単純化

結果と考察

グラフを図 10 に示す．提案手法 1 では，ノード数が増加するにつれて実行時間が徐々に増えていくが，ノード数が 25 を超えるような経路であっても，約 12 秒程度で要約案内文を生成することが可能である．一方，提案手法 2 ではノード数が 17 までの場合，提案手法 1 と実行時間に大きな違いは見られないが，ノー

ド数が 18 を超えると徐々に増加し, 最終的には指数関数的な増加を見せている. この結果は, 経路単純化機能 (総探索) によって非常に多くの単純化経路が生成されることに起因する. これによって, 案内文単純化機能での計算量が著しく増大し, 結果として全体の実行時間も大幅に長くなっていると考えられる.

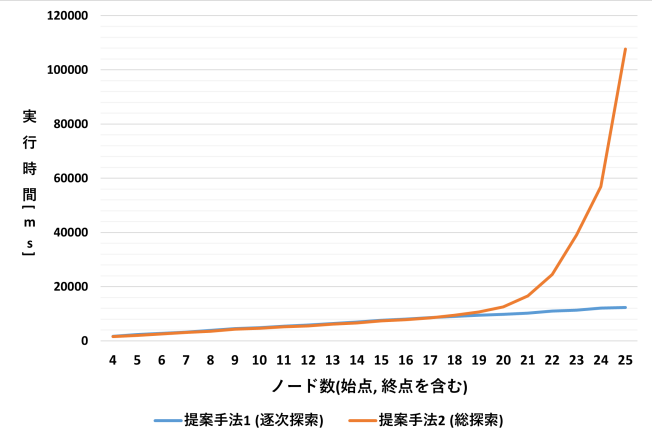


図 10: 2 種類の要約手法 実行時間

サイネージ型経路案内システムは, 主に徒歩や自転車を移動手段とするユーザーを対象としており, 徒歩圏内の経路では右左折地点が 20 を超えることはごく稀である. そのため, サイネージ型経路案内システムで要約案内文を生成する際には, 実行時間に大きな差が出ることはほとんどなく, どちらの手法も十分に利用可能であると考えられる. しかし, ノード数が多くなる長距離移動の場合, 例えばカーナビゲーションや地図アプリの使用時には, 提案手法 2 を使用すると非常に時間がかかる可能性がある. したがって, 長距離移動時において要約案内文を生成する際には, 提案手法 1 がより効果的な手段であると考えられる.

6 おわりに

本稿では, Turn-by-Turn 方式の経路案内文をユーザー指定の文章量にまで要約を行う経路案内文要約手法を提案した. このシステムは, 出発地点から目的地点までの経路を探索する経路探索機能, 獲得した経路から単純化経路を獲得する経路単純化機能, 単純化経路に基づいて案内文の生成を行う案内文生成機能, 案内文から必要性の低い語句を省略する案内文単純化機能を実装することで, 要約案内文の生成を実現している.

評価実験 1 では, 従来手法と提案手法によって生成される要約案内文について, ユーザーアンケートによりその有効性を評価した. 従来手法と 2 種類の提案手法の経路単純化機能を比較した結果, 全てのアンケート項目で従来手法よりも提案手法の方が高い評価を得ることが出来た. 従来手法のように経路構造のみを考慮するのではなく, 提案手法のように”地理情報”の重要度も考慮して経路を単純化することで, よりユーザー満足度の高い要約案内文を生成できることが示唆された. 評価実験 2 では, 2 種類の提案手法について実行時間を計測した. 実験結果より, 提案手法 2 の実行時間はノード数が増えるにつれて指数関数的に増加していくため, アルゴリズムのさらなる改良を検討していく必要があることが分かった.

今後の課題として, 要約案内文に適した経路描画や重要な情報の視覚的強調, 新たな”地理情報”の考案, 総探索による経路単純化を用いた提案手法 2 に要する実行時間を短縮するため, アルゴリズムの改良などが挙げられる.

文　献

[1] Google Maps
`https://www.google.co.jp/maps/` (2024.1.28 参照)
[2] Yahoo!地図
`https://map.yahoo.co.jp/` (2024.1.28 参照)
[3] デジタルサイネージとは? 仕組みやメリットをわかりやすく解説 — NTT コミュニケーションズ 法人のお客さま
`https://www.ntt.com/business/sdpf/knowledge/archive_86.html` (2024.1.28 参照)
[4] 国分寺市様 - 事例 - パナソニック コネクト (panasonic.com)
`https://connect.panasonic.com/jp-ja/case-studies/kokubunji` (2024.1.28 参照)
[5] 公益財団法人 東京観光財団様 - 事例 - パナソニック コネクト (panasonic.com)
`https://connect.panasonic.com/jp-ja/case-studies/tcvb` (2024.1.28 参照)
[6] 山本大介, 後藤直, 服部真由, 前田一樹, 山中勇弥, 竹内健佑, 林晃大, 福安浩明, 高橋直久, しゃべるバス路線案内システムの開発とその実証実験による評価, マルチメディア, 分散協調とモバイルシンポジウム論文集, pp.1872-1878, 2018.
[7] Caitlan A. Rizzardo, Herbert A. Colle, Elizabeth A. McGregor, Daniel Wylie, Using a single map display both for navigational planning and for turn-by-turn vehicle guidance: Configural spatial knowledge acquisition., Journal of Experimental Psychology, Vol.19, No.4, pp301-319, 2013
[8] 原 理緒 , サイネージ型音声案内システムのための経路と案内文の単純化による道案内手法の提案, 名古屋工業大学卒業論文, 2020.
[9] Visvalingam Maheswari and James D. Whyatt, Line generalisation by repeated elimination of points, The cartographic journal, Vol.30, No.1, pp.46-51, 1993.
[10] 鋤柄 知哉 , 道路名を用いた経路単純化による音声道案内文の要約手法, 名古屋工業大学卒業論文, 2021.
[11] Yang Liu and Mirella Lapata, Text Summarization with Pretrained Encoders, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp.3730–3740, 2019.
[12] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang, Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp.280–290, 2016.
[13] Daisuke Yamamoto, Masaki Murase and Naohisa Takahashi, On-Demand Generalization of Road Networks based on Facility Search Results, IEICE Transactions on Information and System, Vol.E102-D, No.1, pp.99-103, 2019.
[14] Masaki Murase, Daisuke Yamamoto, Naohisa Takahashi, On-demand Generalization of Guide Maps with Road Networks and Category-based Web Search Results, In Proceedings of the 14th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2015), Vol.19, pp.53-70, 2015.
[15] R. W. Saaty, The Analytic Hierarchy Process - What and how it is used, Pegamon Journals Ltd., Vol.9, No.3-5, pp.161-176, 1987

大規模言語モデルによる商品レビュー文のキャッチコピーへの変換

沈　辰夫† 荘司　慶行†† 山本　岳洋††† 大原　剛三† 田中　克己††††

† 青山学院大学 理工学部　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
†† 静岡大学 情報学部　〒 430 – 0807 静岡県 浜松市 中区 城北
††† 兵庫県立大学 社会情報科学部　〒 651 – 2197 兵庫県 神戸市 西区 学園西町
†††† 福知山公立大学 情報学部　〒 620 – 0886 京都府 福知山市 字堀
E-mail: †shen@sw.it.aoyama.ac.jp, ††shojiy@inf.shizuoka.ac.jp, †††t.yamamoto@sis.u-hyogo.ac.jp,
††††ohara@it.aoyama.ac.jp, †††††tanaka-katsumi@fukuchiyama.ac.jp

あらまし　本稿では，レビュー文を魅力的なキャッチコピーに変換する方法を提案する．例えば「ポケットに入れていても邪魔にならない」という単純なカメラのレビュー中の一文を入力すると,「ポケットにインして、気軽にスナップ。」のような創造的なキャッチコピーを出力する．そのために，大規模言語モデルをファインチューニングし，また Few-Shot プロンプティングによって，キャッチコピーの候補を多数生成した．そして，生成された候補を，入力レビュー文との内容の一致度・キャッチー度で再ランキングした．被験者評価を通して，生成されたキャッチコピーは商品選びの参考になることが示された．

キーワード　LLM，テキストマイニング，商品レビュー，キャッチコピー

1　はじめに

近年，amazon や楽天といったオンラインショッピングサイトの隆盛や，メルカリに代表される個人間通信販売アプリの普及を受けて，誰もが気軽にインターネットで商品を選び，購入する機会が増えてきている．総務省による調査では，現在の日本人の 73.4 パーセントは，インターネット上で買い物をしたことがあるとされる [1]．一般的に，膨大な数の商品の中から目的に沿った商品を探し出す際には，商品説明文やレビュー文を見ることが多い．このような状況下で，ユーザが投稿するレビュー文の数も増してきており，それらをすべて閲覧して商品を探し出すのは困難になってきている．そこで，画一的にそれぞれの商品の特長を把握したり，一覧したい局面が増えている．

一方，現時点では，商品の売り手側がそれぞれの商品のスペック表や商品説明文などを充実させるくらいでしか解決できない．さらに，スペック表はドメイン知識がないと理解しづらいことがあったり，商品説明文は無駄な記述が多く読みづらかったりするといった問題がある．実際に目的に沿った商品を探し出す際，意思決定に役立ちやすいものとしてユーザが投稿するレビュー文がある．例えば,「野鳥がきれいに撮影できるカメラ」を購入したい際には，実際に野鳥を撮影した体験を記述しているレビュー文があればそれが参考になりやすい．一方で，そのようなレビュー文を見つけるにはそれぞれのレビュー文を逐一読んで探さなければならない．また，このようなことを解決するために，複数文からなるレビュー文の中から，該当するレビュー文中の 1 文を抽出するという手法を考えても，その抽出してきた 1 文は文章の一部であり，そのままでは読みづらく理解しづらい可能性がある．

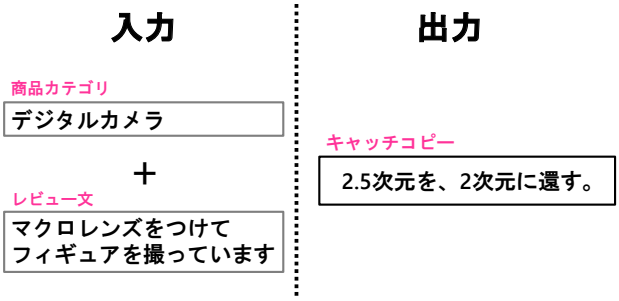


図 1　アプリケーションの入出力例．商品カテゴリとレビュー文を入力すると，キャッチコピーのランキングを出力する．

そこで，本研究では商品の特長や用途が記述されている任意のレビュー文中の 1 文を入力として与えると，そのレビュー文の内容を反映したキャッチコピーを出力する手法を提案する．実際のアプリケーションの動作例を図 1 に示す．

キャッチコピーは一般的には，商品やサービスを魅力的に見せるため，あるいは印象付けるために使われる簡潔なフレーズとして定義されるが，本研究では，商品の任意の特長を，簡潔に，キャッチーに言い表したものとして定義する．レビュー文をキャッチコピー化する利点として，キャッチコピーは元のレビュー文と比較すると，商品を魅力的に見せやすい，ユーザの印象に残りやすい，また並べて比べやすいといったことが挙げられる．

本研究では，このような処理を実現するために，生成的言語モデルを利用する．近年では，テキスト生成といった様々な自然言語処理タスクで大規模言語モデルが活用されている．ま

1：インターネットショッピングおよびオークションフリマの利用状況：令和 3 年版情報通信白書，第 1 部，第 1 章，第 1 節：
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/pdf/n1100000.pdf

た，大規模言語モデルをファインチューニングしたり，プロンプトを工夫することで特定のタスクに特化させることは一般的である．本研究において，商品カテゴリ，商品レビュー，そしてキャッチコピーの 3 つが揃った学習データを用いて生成的言語モデルをファインチューニングすることができれば，商品レビューとそれに対応するキャッチコピーの関係を学習させることができる．一方で，ファインチューニングにはそのような学習データが大量に必要であり，これを作成するには莫大なコストがかかってしまう．

そこで，ファインチューニング時には，商品カテゴリ，そしてキャッチコピーの 2 つを含む学習データを用いる．次に，商品カテゴリ，商品レビュー，そしてキャッチコピーの 3 つが揃ったデータを少数だけ人手で作成する．この少数だけ作成したデータを，Incontext Learning に基づくプロンプティング手法に活用する．生成的言語モデルの推論時に，このプロンプティング手法を用いることで，少数ではあるが商品レビューとそれに対応するキャッチコピーの関係を学習させることができる．このようにして，ファインチューニングを施した生成的言語モデルで，Incontext Learning に基づくプロンプティング手法を活用して推論させることで，商品レビューの内容を反映したキャッチコピーの生成を目指す．

また本研究では，前述の仕組みで多数のキャッチコピー候補を生成し，これらを元のレビュー文の内容を反映しているか，キャッチーであるかという基準でランキングする．これは，生成された複数のキャッチコピー候補の中から，より元の商品レビュー文の内容を正しく反映しているもの，よりキャッチコピーらしいものを抽出するために行う．このとき，元のレビュー文の内容を反映しているか，そしてキャッチーであるかを確かめるために，それぞれ別で用意した識別的言語モデルを用いてスコアを計算する．最終的に，それぞれのスコアを考慮してキャッチコピー候補はランキングされる．

このようにして，実際に，任意の商品の特長や用途が記述されているレビュー文中の 1 文を入力すると，キャッチーでありながら，元のレビュー文の内容を反映したキャッチコピーを出力するアルゴリズムが完成した．例えば,「ポケットに入れていても邪魔にならない」というレビュー文を入力すると,「ポケットにインして，気軽にスナップ.」のようなキャッチコピーを出力する．

実際に行った被験者実験を通して，

（1） キャッチコピーは確かにキャッチーであるか，
（2） キャッチコピーは元の商品レビュー文の内容を反映しているか，
（3） キャッチコピーは商品選びの参考になりそうか，

が明らかになった．本研究は，大規模言語モデルを用いて，ファインチューニングやプロンプティング，後処理といった工夫をすることで商品レビュー文の内容を反映したキャッチコピーを生成できることを示した．

本論文の構成は以下の通りである．本章では，本研究を行うにあたってどのような経緯や背景があったのかを述べた．第 2 章では，本研究に関連した研究について紹介し，本研究の位置づけを示す．第 3 章では，本研究が提案する手法を具体的に述べる．第 4 章では，本提案の有効性に関する評価実験について述べ，第 5 章でその結果を考察する．第 6 章でまとめと今後の課題について述べる．

2 関 連 研 究

本研究は，ユーザレビューを利用した研究，推薦フレーズの生成に関する研究，推薦理由の提示に関する研究の 3 つの分野に関連する．そのため，2.1 節ではユーザレビューを利用した研究について，2.2 節では推薦フレーズの生成に関する研究について，2.3 節では推薦時の根拠の提示に関する研究について，それぞれ関連研究を紹介する．

2.1 ユーザレビューを利用した研究

前述のように，ユーザが投稿するレビュー文の数は増してきている．これを解決するために，Hu [1] は，多数の商品レビューから商品の特長について述べたものをマイニングし，それらに基づいた要約を作成する手法を提案している．

また近年では，ユーザレビュー情報を用いてアイテムを推薦する手法が一般的になりつつある．ユーザレビューを用いた研究の例として，Zheng ら [2] は，ユーザレビューから商品やサービスの特性とユーザの行動を共同で学習する深層学習モデルを用いて，推薦を行う手法を提案している．江田ら [3] は，書籍レビューサイトにおいて対象とするレビュー文に「いいね」を押すユーザを予測する機械学習モデルを構築し，機械学習モデルを解釈する代表的なアルゴリズムの 1 つである LIME を用いて，ユーザに適した推薦フレーズを自動的に抽出する手法を提案している．

これらの研究は，商品購入の意思決定や商品推薦システムの構築にユーザレビューが有用であることを示している．

2.2 推薦フレーズの生成に関する研究

情報推薦の分野などで，推薦フレーズを自動生成する研究も行われてきている．Zhang ら [4] は，実際に JD.com の商品推薦プラットフォームにおいて，機械学習モデルを用いた商品コピーライティング自動生成システム（APCG）を導入し，7ヶ月で 253 万件の商品説明文を生成した結果を報告している．この研究では，機械学習モデルを用いた推薦フレーズが実用化できるものであり，その有効性があることを示している．

ほかに Zhang ら [5] は，条件付き変分オートエンコーダー（CVAE）を基にした Self labeling CVAE (SLCVAE) を提案し，商品説明文を入力に推薦フレーズを生成している．Li ら [6] は，生成的なアプローチを用いた推薦システムにおいて，文章の品質管理と表現力を両立させる試みとして Neural Template (NETE) 説明文生成フレームワークを提案している．Chan ら [7] は，複数商品にまたがる商品説明文や広告を作成する，S-MG Net と呼ばれる機械学習モデルを提案している．Deng ら [8] は，消費者のニーズに合わせて広告コンテンツを自動的にパーソナライズすることができる広告文の生成システムである SGS-PAC を提案している．大曽根ら [9] は，既存の広告文

から抽出した複数のキーワードやカテゴリを入力として，それらに適するような広告文を生成するように，GPT-2 を転移学習する手法を提案している．

これらの研究では商品広告やキャッチコピーを生成することを目的としているが，本研究では，生成されるキャッチコピーが商品の特長を反映しているか，という点に重きを置いている．

2.3 テキストのスタイル変換に関する研究

近年では，大規模言語モデルを活用したテキストのスタイル変換，および言い換え文の生成に関連する研究は盛んに行われている．その中でも，大量のラベル付き学習データを必要としないような教師なし学習手法が注目されている．liu ら [10] は，元の文に含まれる任意の単語を，言語モデルによって予測された別の単語に置換する手法によって言い換え文を生成し，言い換え文生成タスクを最適化問題として解くことで最適な言い換え文を得る手法を提案している．Reif ら [11] は，プロンプティング手法である Augmented zero-shot learning prompt を提案し，センチメントやフォーマルかインフォーマルかであったり，コミカルであるか，ドラマチックであるか等の様々なスタイル変換について評価している．

本研究では，レビュー文をキャッチコピーに変換するが，これは広義にテキストのスタイル変換であり，これらの研究に関連している．

3 提 案 手 法

本手法では，商品の特長について言及している任意の商品レビュー文中の 1 文と，その商品の商品カテゴリを入力すると，そのレビュー文を元にしたキャッチコピー候補を複数生成し，キャッチーであるか，元のレビュー文の意味を保持しているかという基準でキャッチコピーをランキングして出力する．

このような処理を実現するために，まずキャッチコピーコーパスから抽出したキャッチコピーと，人手でラベル付けした商品カテゴリを用いてファインチューニング用の学習データを作成する．そしてこの学習データを用いて生成的言語モデルをファインチューニングする．このようにファインチューニングすると生成的言語モデルは，商品カテゴリとそれに対応するキャッチコピーの関係を学習できる．また，生成的言語モデルの推論時に，Incontext Learning に基づくプロンプティング手法を用いることで，商品レビューとそれに対応するキャッチコピーの関係を学習させるとともに，より品質の高いキャッチコピー候補の生成を目指す．

最後に，生成されたキャッチコピー候補を，キャッチーであるか，また元の商品レビュー文の内容を反映しているかどうか判別するように学習された言語モデルを用いて算出したスコアを基に，ランキングして出力する．

3.1 キャッチコピーコーパスを用いたプロンプトを考慮したファインチューニング

本手法では，キャッチコピーコーパスを用いて生成的言語モデルのファインチューニングを行う．プロンプトを考慮したファインチューニングの例を，図 2 に示す．

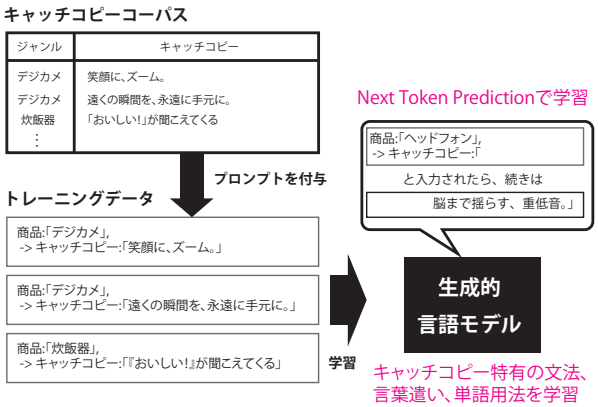


図 2 プロンプトを考慮したファインチューニングの例．書籍をもとに人手でキャッチコピーコーパスを構築し，商品ジャンルとキャッチコピーとプロンプトからなる学習データを作成した．Next Token Prediction タスクで言語モデルに続きの文を予測させることで，通常の日本語の文とは異なる，キャッチコピー特有の表現方法を学習させる．

ファインチューニングを行う際に，学習データを，推論時のプロンプトを考慮して，テンプレートに埋め込んだ形式にした．具体的には，“**商品:「【商品カテゴリ名】」 -> キャッチコピー:「【キャッチコピー本文】」**” という形式のテンプレートを用いた．たとえば商品が「デジタルカメラ」，キャッチコピーが「笑顔に，ズーム.」であった際には，学習用の 1 件のデータは，“**商品:「デジタルカメラ」 -> キャッチコピー:「笑顔に、ズーム。」**” となる．このようなテンプレートに当てはめることで，言語モデルは “**キャッチコピー:「** ” から始まるカッコの中には，日本語と似ているが文法や語彙の異なる，キャッチコピーが入ることを学習できると考えられる．

このような学習データでファインチューニングすることで，生成的言語モデルの推論時に，文の続きを推定するタスクで，任意の文をキャッチコピーに変換可能になる．たとえば，“**商品:「ヘッドホン」 -> キャッチコピー:「** ” のように，学習データを途中で打ち切ったような形式のプロンプトを入力する．すると，“**商品:「ヘッドホン」 -> キャッチコピー:「** 脳まで揺らす、重低音。」” のように，与えられたプロンプトの続きとして，キャッチコピー部分が生成される．

3.2 Incontext Learning に基づくプロンプトを用いたキャッチコピー候補の生成

本手法では，Incontext Learning に基づいたプロンプティング手法を用いてキャッチコピー候補を生成する．本手法で用いるプロンプティング手法は Reif ら [11] を参考にしている．具体的に，Reif ら [11] は，ターゲットとなるスタイルをゼロショットで言語モデルに解かせるのは困難であることから，様々なスタイルを解いた例を羅列してから，最後にターゲットとなるスタイルを解かせる手法をとっている．本手法でも同様な Few-Shot プロンプティング手法を採用する．ターゲットとなるドメインの商品のレビュー文に対するキャッチコピーを直接生成させる

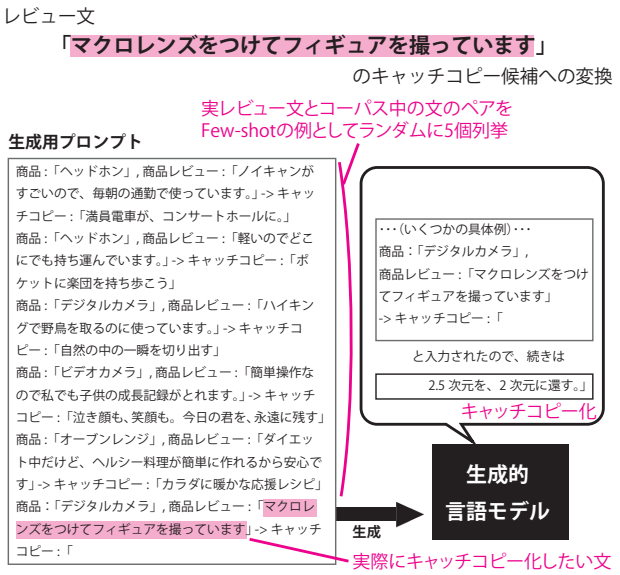


図 3　In-Context Learning に基づくキャッチコピー候補の生成．人手で作成した「商品カテゴリ，レビュー文，キャッチコピー」の組からなるデータセットから 5 件の例をランダムに抽出し，Few-Shot 学習の例としてプロンプトに含める．最後に，実際にキャッチコピーに変換したいレビュー文を入力し，その後の空白に来る文を生成させる．

のが困難であることから，様々なドメインの商品のレビュー文に対するキャッチコピー例を羅列してから，最後にターゲットとなるドメインの商品のレビュー文に対するキャッチコピー候補を生成する．Few-Shot プロンプティングによるキャッチコピー候補の生成の例を図 3 に示す．

さらに，Jiang ら [12] が述べているように，用いる一つのプロンプトは必ずしも最適化されているものではない．そのため，Jiang ら [12] の分析でも評価されている，プロンプトのアンサンブルを採用する．異なる Few-Shot のプロンプトを複数用意し，それぞれのプロンプトでキャッチコピー候補を生成する．

3.3　キャッチコピー候補のランキング

本手法では，別に用意した 2 種類の言語モデルを用いて，キャッチコピー候補をランキングする．生成したキャッチコピー候補のランキングの例を図 4 に示す．ランキングのために，それぞれのキャッチコピー候補に対して，入力レビュー文との内容一致度と，キャッチー度をそれぞれ計算する．そして，それらを合算したスコアで候補を並び替え，最終的なキャッチコピーランキングとして出力する．

はじめに，キャッチコピー候補と入力レビュー文の内容一致度を計算する．本手法では，内容一致度を計算するために，学習済み日本語言語モデルである Sentence-LUKE [13] モデルを用いた．これは，Sentence-BERT [14] と同じタスクで学習したモデルである．具体的には，ある文章とそれに類似する文章のペアを学習データとし，類似した文章から生成される文章ベクトルが似たベクトルになるように学習されている．このモデルに，元のレビュー文とキャッチコピー候補をそれぞれ入力し，それぞれの文章ベクトルを生成する．そして，生成したそれぞれの文章ベクトルのコサイン類似度を計算する．このようにして計算した値を，キャッチコピー候補と入力レビュー文の内容一致度のスコアとする．また，このモデルは，ほとんど元の商品レビュー文と同様な文章で生成されたキャッチコピー候補のフィルタリングにも利用する．

次に，それぞれのキャッチコピー候補のキャッチー度を計算する．本手法では，キャッチコピー候補のキャッチー度を計算するために，汎用言語モデルである BERT [15] をファインチューニングして用いた．具体的なファインチューニングの手順として，キャッチコピーと商品レビュー文が同量含まれる，ラベル付きの学習データを作成した．このデータで，ある文がキャッチコピーか商品レビュー文であるかを見分けるタスクで BERT をファインチューニングした．このようにしてファインチューニングした BERT は，任意の文章を入力すると，その文をキャッチコピー，もしくは商品レビュー文のどちらかに分類する．このモデルにキャッチコピー候補を入力文章として入力し，キャッチコピー，もしくは商品レビュー文のどちらかに分類する．このときに計算されるキャッチコピーへの分類確率をキャッチー度のスコアとする．

最後に，内容一致度とキャッチー度の両方を考慮した，最終スコアを計算する．最終スコアは，内容一致度とキャッチー度の重み付きの和として定義した．すなわち，入力されたレビュー文を r，あるキャッチコピー候補を c_i として，内容一致度を $s_{\mathrm{content}}(r, c_i)$，キャッチー度を $s_{\mathrm{catchy}}(c_i)$ と定義した際，最終スコア $s_{\mathrm{final}}(r, c_i)$ は，

$$s_{\mathrm{final}}(r, c_i) = \alpha s_{\mathrm{content}}(r, c_i) + (1 - \alpha) s_{\mathrm{catchy}}(c_i) \qquad (1)$$

である．ただし，この際，重み α は内容一致度を重視する度合いを表し，0 から 1 の値を任意で設定できる．この最終スコア $s_{\mathrm{final}}(r, c_i)$ の大きい順で候補を並び替えることで，キャッチコピーのランキングを作成する．

4　評　　価

本章では，提案手法の有用性を示すために用いたデータセットと，そのデータセットを用いて行った評価実験の詳細，そしてその結果について述べる．本研究では，提案手法の有効性を評価するために被験者実験を行う予定である．

4.1　データセット

本研究では，大規模なキャッチコピーコーパスと，実サービスのレビューデータを用いて実験を行った．キャッチコピーコーパスは，カタログ・チラシ キャッチコピー大百科 [16] をもとに，人手で作成した．カタログ・チラシ キャッチコピー大百科は，「食品」「生活用品」「ファッション」など，12 個のカテゴリに分かれているが，その中から「デジタル関連製品」および「家庭用電化製品」のカテゴリーに絞り，1,322 個のキャッチコピーを抽出した．さらに，それらの 1,322 個のキャッチコピーに対して，人手で「デジタルカメラ」「イヤホン」といったような商品カテゴリのラベル付けを行った．

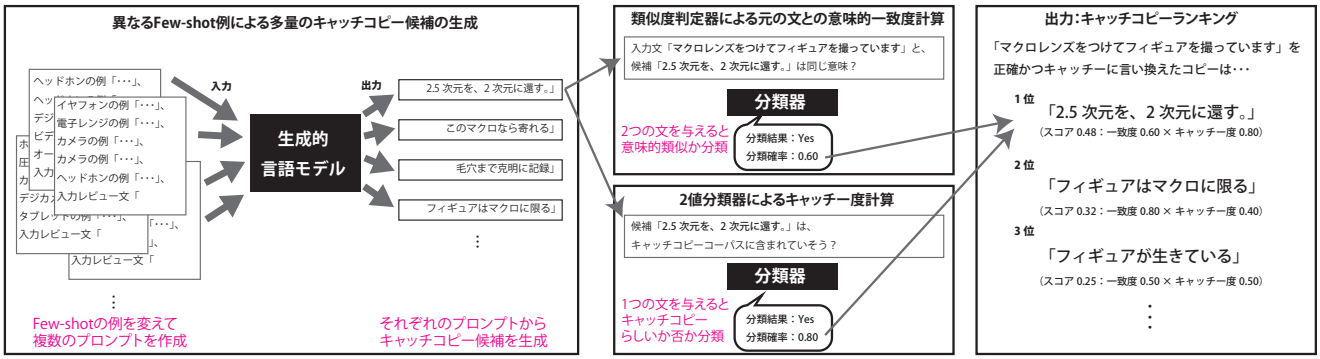


図 4　In Context Learning に基づいて生成したキャッチコピー候補のランキング．入力レビュー文との内容一致度・キャッチー度から算出される最終スコアを基にランキングされる．

実サービスのレビューデータとして，価格.com[2]からレビュー文を収集した．これらのレビューのうち一部は，キャッチコピーと人手で紐づけた．具体的には，キャッチコピーコーパスの中から，商品の特長や用途をアピールしているようなキャッチコピーを 56 個，人手で選出した．そして，そのキャッチコピーが付く商品にふさわしいような商品レビュー文と人手で紐づけ，レビュー文とキャッチコピーが対応する少数のデータを作成した．

4.2 実　　装

本研究で用いる大規模言語モデルの詳細や実装について述べる．キャッチコピー候補を生成する生成的言語モデルとして，サイバーエージェント社による OpenCALM[3]を用いた．このモデルについて，キャッチコピーコーパスを用いてプロンプトを考慮したファインチューニングを施し，Few-Shot のプロンプトをアンサンブルで用いてキャッチコピー候補を生成した．Few-Shot プロンプトのアンサンブルについて，人手で作成したレビュー文とキャッチコピーが対応する 56 個のデータの中から 5 つ選択して 1 つのプロンプトを作成した．また，そのような組み合わせのプロンプトを 10 個作成し，これら 10 個のプロンプトのアンサンブルでキャッチコピー候補を生成した．このとき，選んだ 5 つの例の中で商品カテゴリが 3 つ以上重複しないようにプロンプトを作成し，プロンプト内の商品カテゴリの比率が偏らないようにした．

キャッチコピー候補と入力レビュー文の内容一致度を計算するために用いる言語モデルとして，Studio Ousia 社の日本語 LUKE モデルを元に学習された Sentence-LUKE モデル [4]を用いた．このモデルは，内容一致度の計算だけでなく，キャッチコピー候補のクレンジングにも用いる．生成的言語モデルが生成したキャッチコピー候補の中には，元の商品レビュー文とほぼ同じ文章が含まれる場合がある．こういった候補を取り除くために，このモデルで内容一致度 $s_{\text{content}}(r, c_i)$ が 0.85 より大きかった場合，キャッチコピー候補から除外した．

キャッチコピー候補のキャッチー度を計算するために用いる言語モデルとして，東北大学が公開している BERT モデル [5]を用いた．このモデルを，キャッチコピーコーパスから抽出したキャッチコピーと，価格.com から抽出した商品レビュー文を同じ割合で含むラベル付きの学習データで，2 値分類タスクでファインチューニングした．

実験の際に，内容一致度とキャッチー度の両方を考慮した最終スコアの計算時に，内容一致度を重視する度合いである重み α を 0.2 とした．この値は，キャッチコピー候補内での内容一致度のスコアのばらつきが，キャッチー度のスコアのばらつきよりも比較的大きかったため，予備実験を通して経験的に設定したものである．

4.3 比 較 手 法

本節では，提案手法の有効性を調べるために提案手法と比較する手法について述べる．本実験では，ファインチューニングの効果と，Few-Shot プロンプティングの効果をそれぞれ確認するために，提案手法と比較手法をそれぞれ作成した．具体的には，

- **提案手法**：本研究で提案している，生成的言語モデルにファインチューニングを施し，Few-Shot プロンプティングでキャッチコピー候補を生成，入力レビュー文との内容一致度・キャッチー度でランキングする手法，
- **ファインチューニングなし**：生成的言語モデルはそのまま用いて，Few-Shot プロンプティングでキャッチコピー候補を生成，入力レビュー文との内容一致度・キャッチー度でランキングする手法，
- **Zero-Shot プロンプティング**：生成的言語モデルにファインチューニングを施し，Zero-Shot プロンプティングでキャッチコピー候補を生成，入力レビュー文との内容一致度・キャッチー度でランキングする手法，
- **ランキングなし**：生成的言語モデルにファインチューニ

2：価格.com - 「買ってよかった」をすべてのひとに．：
https://kakaku.com/
3：cyberagent/open-calm-3b：
https://huggingface.co/cyberagent/open-calm-3b
4：sonoisa/sentence-luke-japanese-base-lite：
https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite
5：cl-tohoku/bert-base-japanese-v3：
https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

ングを施し，Few-Shot プロンプティングでキャッチコピーを生成，別途ランキングは行わない手法，

- **ベースライン**：生成的言語モデルをそのまま用いて，Zero-Shot プロンプティングでキャッチコピーを生成する手法

の 5 つの手法である．

ここで使われている Zer-Shot プロンプティングを用いた手法は，キャッチコピー候補の生成時に，プロンプト中で例を与えない．つまり，“**商品：「デジタルカメラ」**, **商品レビュー**：**「マクロレンズをつけてフィギュアを撮っています」**-> **キャッチコピー**：「”のような短いプロンプトだけから，続きとしてキャッチコピーを生成する手法である．

4.4 実験方法

本提案手法によって生成されたキャッチコピーを，被験者実験によって評価する．具体的には，5 つのキャッチコピーの元となる商品レビュー文を用意し，それぞれ 1 つの商品レビュー文に対して生成された，各比較手法につき 5 つのキャッチコピーを評価する．このとき実験に用いる商品レビュー文については，そのレビュー文が確かにその商品の特長や用途を記述しているか，を基準に我々が選定した．実際に評価する項目は，

（1） キャッチー性：キャッチコピーは確かにキャッチーであるか，

（2） 内容反映性：キャッチコピーは元の商品レビュー文の内容を反映しているか，

（3） 参考性：キャッチコピーは商品選びの参考になりそうか，

の 3 項目である．それぞれの項目について，5 段階のリッカート尺度で評価した．

4.5 実験結果

本節では，4.4 節における被験者実験の結果を述べる．実験結果から，提案手法等と生成的言語モデルをそのまま用いたベースラインの手法とを t 検定を行い比較し，有意差が認められた項目については数値には*をつけて表記する．各比較手法のそれぞれの評価項目に対する平均評点をまとめて比較したものを表 1 に示す．実験の結果，提案手法とベースライン手法と比較して，すべての評価項目において有意差が認められた．ファインチューニングなしの手法においても，ベースライン手法と比較して，同様にすべての評価項目において有意差が認められた．Zero-Shot プロンプティングの手法やランキングなしの手法において，ベースライン手法と比較して，キャッチーな文章であるかの評価項目において有意差が認められた．

表 1　各比較手法のそれぞれの評価項目に対する平均評点（1 から 5，ベースラインと比べて ** $p < 0.01$, * $p < 0.05$）

評価項目	キャッチー性	内容反映性	参考性
提案手法	**3.45	****3.88**	*3.51
Zero-Shot	****3.89**	3.53	3.33
FT なし	**3.57	**3.81	****3.52**
ランキングなし	*2.77	3.11	3.32
ベースライン	2.21	3.15	2.99

5 考　　察

本章では，評価実験によって得られた結果を考察する．実験結果から，提案手法はベースライン手法と比較して，すべての評価項目において有意差が認められた．このことから，ファインチューニング，プロンプティング，ランキングによる後処理を行うことで，単純に生成的言語モデルをそのまま用いるよりも，キャッチーかつ，レビュー文の内容を反映し，商品選びの参考になるキャッチコピーを生成できることがわかった．さらに，いくつかの観点に注目して考察を述べる．

はじめに，キャッチコピーコーパスを用いたプロンプトを考慮したファインチューニング手法について考察する．まず，提案手法と，ファインチューニングなしの手法を比較すると，どの評価項目においても有意差は認められず，評点はほとんど同様な傾向に見える．一方で，Zero-Shot プロンプティングと，ファインチューニングなしの手法を比較すると，どの評価項目も有意差は認められないものの，Zero-Shot プロンプティングではよりキャッチーな文章であるかの評点が高く，ファインチューニングなしではよりレビューを反映しているかの評点が高い傾向に見える．このことから，プロンプティング手法による影響よりは小さいが，プロンプトを考慮したファインチューニング手法は生成文をよりキャッチーにするように影響していると考えられる．

次に，Incontext Learning に基づく Few-Shot プロンプティング手法について考察する．提案手法と，Zero-Shot プロンプティングの手法を比較すると，どの評価項目においても有意差は認められなかったが，提案手法ではよりレビューを反映しているかの評点が高く，Zero-Shot プロンプティングではよりキャッチーな文章であるかの評点が高い傾向に見える．このことから，Few-Shot プロンプティングによって，生成文はよりレビューを反映するように影響していると考えられる．

次に，レビューとの内容一致度およびキャッチー度でのランキング手法について考察する．提案手法と，ランキングなしの手法を比較すると，キャッチーな文章であるか，およびレビューを反映しているかの 2 項目において有意差が認められた．このことから，別に用意した 2 種類の言語モデルによる，レビューとの内容一致度，およびキャッチー度でのランキング手法は有効であると考えられる．

次に，プロンプトを考慮したファインチューニングや Few-Shot プロンプティング手法による生成文への影響について考察する．ベースライン手法の生成する文章の中には，商品カテゴリのようなものが含まれたりしていたが，これはプロンプトの形式を学習できていないからであると考えられる．他の手法では，そのようなことがほとんどなかったことから，プロンプトを考慮したファインチューニングや Few-Shot プロンプティング手法は，生成文を正しくキャッチコピーらしいものにすることに対して有効であったと考えられる．

最後に，各手法で被験者からの評点が相対的に高かった生成

表 2　各手法で被験者からの評点が相対的に高かった生成キャッチコピーの例

カテゴリ	レビュー文	手法	生成されたキャッチコピー
デジタルカメラ	ポケットに入れていても邪魔にならず快適です！	提案手法	ポケットにインして、気軽にスナップ。
		Zero-Shot	ずっと付き合いたくなるポケットサイズ。
		FT なし	いつでも、どこでも、どんなシーンでも。
		ランキングなし	旅行やレジャーにもぴったり。
		ベースライン	コンパクトなのに、高画質。
デジタルカメラ	運動会やアスレチックなどで激しく動く子どものような動体でもピントが合います。	提案手法	動きの速い被写体も、確実にキャッチ!
		Zero-Shot	動いているものを、ピンポイントで。
		FT なし	スポーツシーンでも、動く被写体でも、ピントが合う。
		ランキングなし	「動くもの」を撮るなら、このカメラ。
		ベースライン	動く被写体でも、ピントが合う。
イヤホン	かなりの重低音で、ライブ会場さながらの迫力で思わず体を揺らしてしまうようなノリの良いサウンドが聞けます。	提案手法	まるで、ライブ会場にいるような臨場感。
		Zero-Shot	もう、ポータブルでライブ会場です。
		FT なし	それは、ライブ会場さながらのサウンド。
		ランキングなし	音楽好きにはたまらない、ライブ会場さながらのサウンド。
		ベースライン	重低音が魅力のイヤホン
炊飯器	しっとり且つお米の一粒一粒が立っており、炊きあがりの香りや艶も問題なしでした。	提案手法	ふっくらと、つやつやと。お米の味わい方、無限大。
		Zero-Shot	ひと粒ひと粒に感動を。
		FT なし	ふっくら、つやつや、おいしいお米を炊く。
		ランキングなし	炊飯器なのに、まるで土鍋で炊いたような艶やかさ。
		ベースライン	ふっくら美味しいご飯が炊ける
エアコン	パワーも十分、音も静かで、大変満足しています。	提案手法	静かで、パワフルで、快適。
		Zero-Shot	静かな寝室にも、パワーあり。
		FT なし	省エネで、快適で、静か。
		ランキングなし	静かだから、寝室にもぴったり。
		ベースライン	快適空間を創造する、エアコン

キャッチコピーの例を表 2 に示す．提案手法，Zero-Shot プロンプティング，ファインチューニングなしの各手法については，キャッチーかつ，レビュー文の内容を反映したキャッチコピーが生成できていることがわかる．また，各手法で被験者からの評点が相対的に低かった生成キャッチコピーの例を表 3 に示す．提案手法，Zero-Shot プロンプティング，ファインチューニングなしの各手法については，生成文はキャッチーではあるがレビュー文を反映できていない，もしくはレビュー文の内容に忠実だがキャッチーではない例が見られる．ベースライン手法については，商品カテゴリや商品の一機能を生成してしまっている例が見られる．

6 ま と め

本稿では，レビュー文を魅力的なキャッチコピーに変換する方法を提案した．具体的には,「ポケットに入れていても邪魔にならない」のようなレビュー文を入力すると,「ポケットにインして、気軽にスナップ。」のような創造的なキャッチコピーを出力する．そのために，大規模言語モデルをプロンプト付きのデータセットでファインチューニングし，Few-Shot プロンプティングでキャッチコピーの候補を多数生成した．生成された候補を，入力レビューとの内容の一致度・キャッチー度を考慮して並び替え，キャッチコピーをランキング化した．実際に生成されたキャッチコピーを被験者に評価させることで，生成されたキャッチコピーがキャッチーで，内容を反映しており，商品選びの参考になることが示された．

今後の課題として，今後，より大規模な被験者実験を行う予定である．また，今回の研究では，任意のレビューを入力するとその文をキャッチコピーに言い換える部分だけを可能にした．実際に商品を比較検討する際には，その商品の特徴や魅力がよく表れたレビュー文を特定する必要がある．商品特徴を考慮したキャッチコピー生成などの，継続的な研究を行う予定である．

謝　　辞

本研究の一部は JSPS 科研費 21H03775，21H03774，22H03905 による助成を受けたものです．本研究では，楽天グループ株式会社が国立情報学研究所情報学研究データリポジトリを通じて提供している楽天市場データ [17] を使用しています．ここに記して謝意を表します．

文　　献

[1] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, p. 168–177, New York, NY, USA, 2004. Association for Computing Machinery.

表 3　各手法で被験者からの評点が相対的に低かった生成キャッチコピーの例

カテゴリ	レビュー文	手法	生成されたキャッチコピー
デジタルカメラ	ポケットに入れていても邪魔にならず快適です！	提案手法	カメラに装着したまま、そのまま持ち歩ける。
		Zero-Shot	夜景も、手ブレの心配なし。
		FT なし	どこでも、いつでも、どこでも。
		ランキングなし	手ぶれ補正のおかげで、夜景もきれいに撮れる。
		ベースライン	カメラ
デジタルカメラ	運動会やアスレチックなどで激しく動く子どものような動体でもピントが合います。	提案手法	「動き」に強い、という発想がすばらしい。
		Zero-Shot	動体への自信
		FT なし	動くものを撮るなら、動くものを撮る。
		ランキングなし	「動く」という被写体をいかに捉えるか。
		ベースライン	動体検知機能
イヤホン	かなりの重低音で、ライブ会場さながらの迫力で思わず体を揺らしてしまうようなノリの良いサウンドが聞けます。	提案手法	「臨場感」は、音楽にとって永遠の課題かもしれない。
		Zero-Shot	重低音好きは、見逃さない手強い勢いだ。
		FT なし	迫力のサウンドでライブ会場さながらの迫力を体感できる。
		ランキングなし	音漏れもほとんどなく、音漏れを気にせずに思いっきり音を楽しめる。
		ベースライン	イヤホン
炊飯器	しっとり且つお米の一粒一粒が立っており、炊きあがりの香りや艶も問題なしでした。	提案手法	土鍋で炊いたご飯は、格別です。
		Zero-Shot	土鍋で炊いたご飯は格別です。
		FT なし	ふっくらおいしいご飯が炊けます。
		ランキングなし	お米も炊飯器も、やっぱり日本製に限ります。
		ベースライン	炊飯器
エアコン	パワーも十分、音も静かで、大変満足しています。	提案手法	静かで、パワフルで、省エネで、しかもお財布に優しい。
		Zero-Shot	静かさ・省エネ・健康・環境への配慮。
		FT なし	それは、快適で、静かで、省エネで、そして、エコで、クリーンな、エアコン。
		ランキングなし	冷房病の予防に、一台購入しました。
		ベースライン	エアコンが効かない

[2] Lei Zheng, Vahid Noroozi, and Philip S Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 425–434, 2017.

[3] 怜央江田, 剛聡牛尼. ユーザレビューを用いた機械学習モデルの説明可能性に基づく推薦フレーズの抽出. 情報処理学会論文誌データベース（TOD）, Vol. 15, No. 3, pp. 50–62, 2022.

[4] Xueying Zhang, Yanyan Zou, Hainan Zhang, Jing Zhou, Shiliang Diao, Jiajia Chen, et al. Automatic product copywriting for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 12423–12431, 2022.

[5] Yuchi Zhang, Yongliang Wang, Liping Zhang, Zhiqiang Zhang, and Kun Gai. Improve diverse text generation by self labeling conditional variational auto encoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2767–2771. IEEE, 2019.

[6] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, pp. 830–831, New York, NY, USA, 2002. Association for Computing Machinery.

[7] Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. Selection and generation: Learning towards multi-product advertisement post generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3818–3829, Online, November 2020. Association for Computational Linguistics.

[8] Shasha Deng, Chee-Wee Tan, Weijun Wang, and Yu Pan. Smart generation system of personalized advertising copy and its application to advertising practice and research. *Journal of Advertising*, Vol. 48, No. 4, pp. 356–365, 2019.

[9] 大曽根宏幸, 張培楠. Gpt-2 の転移学習によるキーワードを考慮した広告文生成. 人工知能学会全国大会論文集, Vol. JSAI2021, pp. 2D4OS7b03–2D4OS7b03, 2021.

[10] Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. Unsupervised paraphrasing by simulated annealing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 302–312, Online, July 2020. Association for Computational Linguistics.

[11] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. A recipe for arbitrary text style transfer with large language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 837–848, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[12] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 962–977, 2021.

[13] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454, Online, November 2020. Association for Computational Linguistics.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, ed-

itors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] 久野寧子（編）. カタログ・チラシ キャッチコピー大百科. ピエ・ブックス, 2008.

[17] 楽天グループ株式会社. 楽天市場データ, 2020. https://doi.org/10.32130/idr.2.1 https://rit.rakuten.com/data_release/.

特定選手に注目した試合結果記事見出しのパーソナライズ

佐々木泰河† 工家 昂之†† 山本 岳洋†† 荘司 慶行†††

† 兵庫県立大学 社会情報科学部 〒 651-2197 兵庫県神戸市西区学園西町 8-2-1
†† 兵庫県立大学 大学院情報科学研究科 〒 651-2197 兵庫県神戸市西区学園西町 8-2-1
††† 静岡大学 情報学部 〒 432-8011 静岡県浜松市中区城北 3-5-1
E-mail: †fa20e036@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp,ad22l024@guh.u-hyogo.ac.jp,
†††shojiy@inf.shizuoka.ac.jp

あらまし 本研究では，試合結果記事における特定選手に注目した見出しを生成する手法を提案する．見出しの生成には T5 という文生成モデルを用いる．記事全文を入力として，T5 に見出しの最初の文字を注目している選手名にする制約をかけた単純な手法では，注目していない選手の活躍を，注目している選手の活躍のように見出し生成してしまうという問題がある．そこで我々は，試合結果記事から記事の主題に関する文，注目している選手の活躍に関する文のみを抽出して，T5 に入力する手法を提案する．このアプローチにより，指定した選手に注目した自然で事実に基づいた見出しを生成させる．評価実験により，提案手法は，記事全文を入力する単純な手法よりも，特定選手に注目した見出し生成の精度が良いことが示された．

キーワード 見出し生成，パーソナライズ，T5

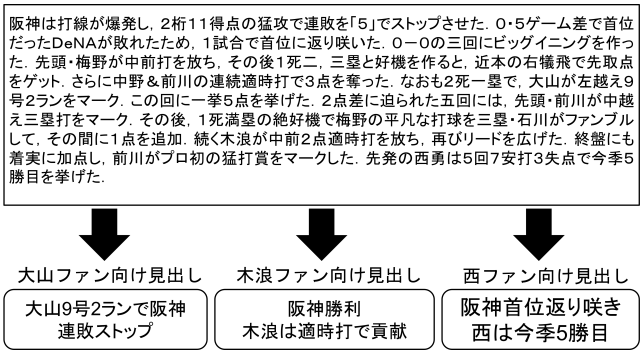


図 1 目標とする生成例.

1 はじめに

プロ野球では，試合が行われるとその試合結果を報じる記事が各新聞社から配信される．その記事には試合の勝敗や試合内容，活躍した選手に関する情報などが記される．この記事をクリックするかという判断基準として，見出しに自分の興味の引く内容が含まれているかが考えられる．例えば，興味のある選手の選手名が見出しに含まれていると，読者はその記事が目に止まり，クリックする可能性が高い．具体的には，「阪神が打線爆発で連敗ストップ！1試合で首位返り咲き 高卒2年目の前川がプロ初猛打賞で牽引 中日戦6連勝」という見出しの記事があるとき，前川選手のファンは記事をクリックする可能性が高い．しかし，見出しに自分の好きな選手の選手名が含まれていなかったとしても，記事の中身を読んでみるとその選手に関する情報が書かれていることもある．例えば，上記の例で示した記事本文には大山選手が 9 号 2 ランを放ったことが書かれている．この場合，読者が仮に大山選手のファンで見出しに「大山」と含まれていない理由により，記事をクリックしなければ，この情報を見逃すことになる．

そこで本研究では，試合結果記事から指定した選手に注目した見出しの生成を行う．本研究が目標とする生成例を図 1 に示す．ある試合結果記事があったときに仮に大山選手に注目したときには「大山 9 号 2 ランで阪神連敗ストップ」と生成する．木浪選手に注目したときには「阪神勝利 木浪は適時打で貢献」と生成する．また，西選手に注目したときには「阪神首位返り咲き 西は今季 5 勝目」と生成する．このような見出し生成をすることができれば，記事を表示する際に読者が興味を引きやすい見出しに書き換えて表示することも可能になる．

この技術があることにより，読者が自分の興味のある情報を見逃しにくくなる．また，記事の配信者がより多くのクリック数を集めることができる．

試合結果記事見出しの生成において, 生成型言語モデルの 1 つである T5 [9] を用いる．T5 は，文章を入力として与え，解いているタスクに応じた文章を出力するモデルである．大規模なデータで事前学習した T5 を少量のデータでファインチューニングすることで，様々なタスクに対して応用が可能である．本研究では，あるニュース記事本文を入力すると，その見出しを出力するように T5 をファインチューニングする．しかし，見出し生成の際，記事全文を入力すると，生成される見出しにおいて利用すべき情報を誤って選択してしまう問題がある．例えば，大山選手に注目して見出しを生成しようとしたとする．そして本文には「なおも 2 死一塁で，大山が左越え 9 号 2 ランをマーク.」という記述や「木浪が中前 2 点適時打を放ち，再びリードを広げた.」という記述がある場合を考える．このとき，仮に記事全文を入力したとすると，大山選手が適時打を放っていなかったとしても「大山の適時打で阪神連敗ストップ」というような見出し生成をしてしまう可能性がある．大山選手に注

目した見出しを生成するために，木浪選手が適時打を放ったことは余計な情報である．そこで，ニュース記事全文を入力するのではなく，理想的な見出しを生成するために必要な情報を含む文のみを抽出して，入力する．これにより，モデルが注目していない選手ついて言及している本文の情報を，見出し生成に利用することを防ぎ，見出しの精度を向上させることができる．

評価実験では，生成した見出しが本研究で目標とする見出しであるか評価する．

本論文の構成について説明する．第 2 節では，本研究と関連する研究を紹介する．第 3 節では，本論文で提案する手法について説明し，第 4 節では提案した手法の評価実験について述べる．第 5 節では実験で得られた結果をもとに提案手法による見出し生成の精度の変化について考察する．第 6 節ではまとめと今後の展望について述べる．

2 関 連 研 究

2.1 大規模言語モデルを用いた文生成

文生成に関する研究は，盛んに行われている．過去には RNN [2] や LSTM [3] による文生成が主流であったが，これらの手法は長い文章や長距離の依存関係をとらえることが難しいという課題がある．近年，このような課題に対処するために Transformer [12] が提案され，その登場により自然言語処理の様々なタスクにおいて革新的な進展が見られた．Transformer は 2017 年に Google によって発表され，Attention という機構を導入したモデルである．この Transformer をベースとして多くの大規模言語モデルが提案されており，その代表的な例には GPT-n [8]，BERT [1]，T5 がある．

江原らは，GPT-2 のファインチューニングを工夫することにより主題語に応じたニュース記事見出しを生成する手法を提案している [15]．この研究で提案されている手法は，GPT-2 が持つ「文の続きを推定する」という機能を利用している．黒木らは BERT2BERT を用いた広告文生成において特定のキーワードの出現率を高くする手法を提案している [17]．BERT2BERT とは事前学習済み BERT をテキスト生成タスクに応用したモデルである．福田らは文章の感情要約文を T5 によって生成する手法を提案している [19]．この研究では，学習時に入力とする文の接頭辞に感情タグをつけることにより精度を向上させた．

本研究では，見出しを生成するための大規模言語モデルとして，T5 を利用している．

2.2 要約タスク

要約タスクとは，与えられた文書や文章を簡潔にまとめ，主要な情報や要点を取り出す作業である．要約タスクは一般的に生成型要約と抽出型要約に分類される．

生成型要約とは，与えられた文書や文章から新しい表現で要約を生成するタスクである．このタスクでは大規模言語モデルが用いられることが多い．生成型要約は，元の文章には現れない新しい表現や言い回しで要約を生成する点に特徴がある．木村らはエンコードの際に原文の単語単位のベクトルに加え，原文の文単位のベクトルも導入することで，クエリ指向要約モデルを強化する手法を提案している [20]．また Zhang らは，異なる事前学習方法，プロンプト，モデルサイズにわたる 10 個のモデルで生成された要約文を人手で評価することで，自動要約において重要な要素を分析した [14]．Yang らは ChatGPT が多様なテキスト要約タスクにおいてどのような特性を持っているかについて分析した [13]．

抽出型要約は，元の文章にすでに存在する文や文の一部を選択して要約を形成する．小川らは，ニューラルネットワークを用いてニュース記事の各文の重要度を計算し，重要度の高い文を要約に利用する文として抽出する手法を提案している [18]．さらに Ruan らは，テキストが持つ階層構造の情報を抽出型要約モデルに注入することで，要約の精度を向上させる手法を提案している [10]．また高津らは，文の特徴量とアンケートで得られるプロフィール特徴量に基づいてニュース記事の各文に対するユーザの興味度を推定し，その興味度を用いて文を抽出することで，パーソナライズされた要約を作成する手法を提案している [16]．

生成型要約と抽出型要約は要約の目的や利用ケースによって使い分けられている．生成型要約は元の文章には現れない新しい表現で要約を生成するため，より柔軟で創造的な結果が得られる．また文のつながりも流暢性のあるものになりやすい．しかし，生成される要約が事実と合致していない可能性がある点に留意する必要がある．本研究は，要約タスクの結果として得られた要約文を，T5 に入力している．要約タスクを解いて入力文を作る目的は，生成した見出しと事実との不一致の可能性を下げることである．そのため本研究では，記事本文中の重要な文を抽出して入力文を作る抽出型要約を使用する．

3 提 案 手 法

3.1 問 題 定 義

本研究における問題定義は以下のとおりである．

- **入力**：注目する選手名，試合結果記事
- **出力**：指定した選手に注目した見出し

本研究では，注目する選手名と，試合結果記事を入力し，指定した選手に注目した見出しを生成することを目標とする．

また，生成したい見出しを明確にするために，以下の 5 つの特徴を持つ見出しを理想的な見出しとして定義する．

- **事実と異なっていない**：事実と異なった見出しを生成しては読者の誤解を招き良い見出しといえない．
- **指定した選手の選手名を含んでいる**：指定した選手の選手名が含まれていなければ，その選手に注目したニュース記事見出しを生成できているとはいえない．
- **記事全体の要約を含んでいる**：記事全体の要約も含まれていなければ，その記事で伝えていることが分からない．
- **日本語として自然である**：文法的に正しくなければ，良い見出しとはいえない．
- **見出しとして自然である**：見出しとして扱う情報として妥当でなければ，良い見出しとはいえない．例えば，「中野選手

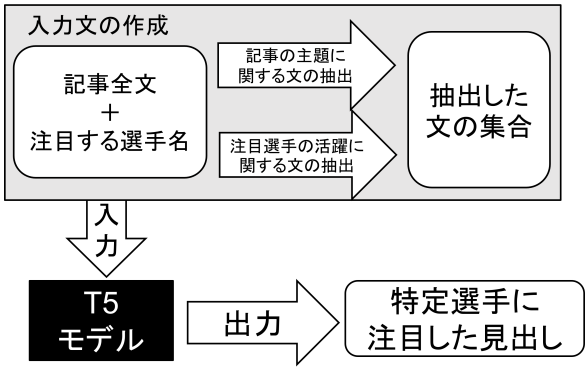


図 2　手法の全体像.

がカーブを打ってホームランにした」という情報があるとき，見出しで「カーブを打った」ことのみを述べて,「ホームランを打った」ことに触れていなければ良い見出しとはいえない.

本研究では，上記の 5 点の特徴を含んだ見出しを生成することを目標とする.

3.2　手法の全体像

本節では，試合結果記事から特定選手に注目した見出しを生成するために，抽出型要約と T5 を組み合わせた手法の全体像を説明する.

図 2 に本研究で提案する手法の全体像を示す．記事全文から，記事の主題に関する文の抽出と注目選手の活躍に関する文の抽出を行う．抽出した文を結合することにより，T5 への入力文を作る．完成した入力文を T5 に入力し，特定選手に注目した見出しを出力させる.

3.3　入力文の作成

通常，見出し生成をする際，モデルに記事全文を入力することが一般的である．しかし，記事全文をモデルに与えると，生成される見出しにおいて利用すべき情報を誤って選択する可能性がある．そこで，記事全文を入力しないことで，T5 に余計な情報を与えない．これにより，注目していない選手の活躍を，注目している選手の活躍であるかのように見出し生成する可能性を下げることができる.

実際に，どのような特徴を持つ文を抽出するかについて，説明する．本手法において抽出する特徴を持つ文は 2 種類ある．1 つ目は，記事の主題に関する文である．問題定義で定めた理想的な見出しが持つ特徴のひとつに「記事全体の要約を含んでいる」というものがある．記事の主題に関する文は，この特徴を生成した見出しに持たせるために抽出する．2 つ目は，注目選手の活躍に関する文である．これは，生成した見出しに「指定した選手の選手名を含んでいる」という特徴を持たせるために抽出する.

3.3.1　記事の主題に関する文の抽出方法

記事の主題に関する文の抽出方法を説明する．図 1 の試合結果記事を具体例として実際に文を抽出していく過程を説明する．本研究では 2 つのアプローチを組み合わせることで，記事の主題に関する文を抽出する.

a）　LEAD　法

1 つ目のアプローチは LEAD 法である．LEAD 法とは，文書の冒頭に重要な情報が存在するという仮定のもと，文書内の先頭に位置する文や，先頭のいくつかの文を選択する抽出方法である．石川らは LEAD 法は新聞記事の要約に特に効果的な手法として知られていると述べている [4]．試合結果記事においても記事の先頭にチームの勝敗など記事の主題となる重要な情報が書かれることが多い．そこで本手法では，記事の 1 文目を抽出し，入力文に含めることにする．図 1 の具体例に当てはめると「阪神は打線が爆発し，2 桁 11 得点の猛攻で連敗を「5」でストップさせた.」という文を抽出することになる.

b）　本物の見出しと類似度が高い文を抽出する手法

2 つ目のアプローチは，本物の見出しと類似度が高い文を抽出する手法である．本物の見出しは，記事の主題を読者に素早く伝えるために筆者が作成したものである．よって，本物の見出しと類似度が高い文は記事の主題を生成した見出しに含めるために必要な文であると考えられる．図 1 の本物の見出しは「阪神が打線爆発で連敗ストップ！1 試合で首位返り咲き 高卒 2 年目の前川がプロ初猛打賞で牽引 中日戦 6 連勝」である．この見出しと類似度が高い文を本文から抽出する．そのために全ての文について本物の見出しとの類似度を求める．類似度は本物の見出しをクエリ，記事本文内の各文を文書と見なして計算する．ここでは「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」という文を具体例として，類似度の計算方法を説明する.

まずは Mecab [5] を利用して記事本文と本物の見出しを形態素に分割する．その際，助詞と数値を省く処理をする．助詞を省くのは,「は」や「が」などの助詞は類似度をはかるのに必要無いためである．数値を省くのは，2 桁と 2 年目では同じ「2」という形態素であったとしてもまったく意味が異なるためである．その結果，本物の見出しの形態素は [‘阪神’,‘打線’,‘爆発’,‘連敗’,‘ストップ’,‘！’,‘試合’,‘首位’,‘返り咲き’,‘高’,‘卒’,‘年’,‘目’,‘前川’,‘プロ’,‘初’,‘猛’,‘打’,‘賞’,‘牽引’,‘中日’,‘戦’,‘連勝’] となる.「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」という文の形態素は [‘・’,‘ゲーム’,‘差’,‘首位’,‘ＤｅＮＡ’,‘敗れ’,‘試合’,‘首位’,‘返り咲い’] となる．そこから本物の見出しと類似度を計算している文で共起している形態素を抽出する．今回の例では [‘試合’,‘首位’] が抽出できる.

続いて，各形態素の IDF [11] を計算する．IDF は特定の単語が文書集合においてどれほど重要であるかを表す指標である．この指標を用いることにより，試合結果記事で現れることが少ない形態素が，本物の見出しの形態素と共起することの重要度を上げる．具体的には,「打」や「回」のような試合結果記事のあらゆる文で現れる形態素が，本物の見出しの形態素と共起することより,「爆発」や「ストップ」のような，試合結果記事で現れることが少ない形態素が本物の見出しの形態素と共起することの重要度を上げている．式 (1) を用いることで形態素 t の IDF 値である $\mathrm{idf_t}$ を計算する.

$$\mathrm{idf_t} = \log\left(\frac{10}{t\text{ が現れる文の数} + n}\right) \tag{1}$$

表 1　$\mathrm{idf_t}$ の計算結果（一部）.

形態素	値
爆発	0.51
ストップ	0.51
試合	0.51
首位	0.51
中前	0.36
打	0.11
回	0.11

式 (1) 中の n は定数である．n の値を大きくすると別の文で現れることが少ない形態素の重要度が上がる．今回は n に 5 を代入する．また通常の IDF では，対象記事の文の数が分子となる．しかし，記事によって文の数が異なることにより，形態素登場 1 回分の重要度が変動することを防ぐため，分子は 10 で固定する．

図 1 の例における $\mathrm{idf_t}$ を計算した結果の一部を表 1 に示す．1 つの文でしか登場しない形態素の $\mathrm{idf_t}$ の値は 0.51 であるのに対し，「打」や「回」など複数の文で登場する形態素は 0.11 と低い値である．

続いて，共起している形態素の $\mathrm{idf_t}$ の値の合計 $\mathrm{idf_{sum}}$ を計算する．$\mathrm{idf_{sum}}$ は式 (2) を用いて計算する．

$$\mathrm{idf_{sum}} = \sum_{i=1}^{k} \mathrm{idf_{t_i}} \tag{2}$$

この式における k は類似度を計算している文と本物の見出しで共起している形態素の数である．「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」という文と本物の見出しで共起している形態素は [‘ 試合 ’ , ‘ 首位 ’] である．よって k の値は 2 となる．式 (2) を用いて $\mathrm{idf_{sum}}$ の値を計算すると $0.51 + 0.51 = 1.02$ となる．

式 (2) で得られた $\mathrm{idf_{sum}}$ を用いて文の類似度を式 (3) で求める．

$$\mathrm{rank} = \alpha \cdot \mathrm{idf_{sum}} - \beta \cdot \mathrm{morphemes} \tag{3}$$

この式では長い文にペナルティをかけている．文が長ければ長いほど形態素が共起する可能性が上がるため，ペナルティをかける必要がある．morphemes は，類似度を計算している文の形態素の数である．また α と β は，値の重要度を表すパラメータである．今回は α に 5，β に 0.1 を代入する．

今回は式 (3) で値が 3 以上になったものを本物の見出しと類似度が高い文として抽出する．「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」という文の形態素は [‘・’ , ‘ ゲーム ’ , ‘ 差 ’ , ‘ 首位 ’ , ‘ ＤｅＮＡ ’ , ‘ 敗れ ’ , ‘ 試合 ’ , ‘ 首位 ’ , ‘ 返り咲い ’] である．したがって，形態素の数を表す morphemes の値は 9 である．式 (3) を用いて，「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」という文の類似度を計算すると $5 \cdot 1.02 - 0.1 \cdot 9 = 4.2$ となる．この値は 3 以上であるので，本物の見出しと類似度が高い文として「0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた.」は抽出対象の文となる．

しかし，生成する見出しで注目している選手が本物の見出しで注目している選手と同じではないときは，文の類似度が 3 以上であっても抽出するべきではない．具体的に図 1 の記事の例で説明すると「終盤にも着実に加点し，前川がプロ初の猛打賞をマークした.」という文の類似度は 10.85 で 3 以上である．しかし，生成した見出しを前川選手の活躍に注目させたい場合はこの文を入力文に含めても構わないが，生成した見出しを大山選手の活躍に注目させたい場合はこの文は入力文に含めるべきではない．そこで，文の類似度が 3 以上かつ，注目している選手以外の選手名が，その文に含まれない文を抽出対象の文とする．

3.3.2　注目選手の活躍に関する文の抽出方法

注目選手の活躍に関する文の抽出方法を説明する．試合結果記事においては，選手の活躍について述べる際にその文に選手名が入っていることが多い．そこで，注目している選手の選手名が含まれていたとき，その文を抽出することにする．図 1 の記事の例において，大山選手に注目した場合は「なおも 2 死一塁で，大山が左越え 9 号 2 ランをマーク.」，木浪選手に注目した場合は「続く木浪が中前 2 点適時打を放ち，再びリードを広げた.」という文が抽出対象の文となる．

3.3.3　抽出した文の結合

これまでに抽出した文章を結合して，T5 への入力文を作成する．結合する際には，文の順番を変えないようにする．文の位置関係は見出し生成の際に重要な情報の一部であり，これを変更すると生成される見出しの精度を低下させる可能性がある．したがって，抽出した文の位置関係は元の記事と同じになるようにする．以下に，図 1 の記事を例に大山選手に注目した場合と木浪選手に注目した場合の入力文を示す．

- **大山選手に注目した場合の入力文**：阪神は打線が爆発し，2 桁 11 得点の猛攻で連敗を「5」でストップさせた．0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた．なおも 2 死一塁で，大山が左越え 9 号 2 ランをマーク.
- **木浪選手に注目した場合の入力文**：阪神は打線が爆発し，2 桁 11 得点の猛攻で連敗を「5」でストップさせた．0・5 ゲーム差で首位だったＤｅＮＡが敗れたため，1 試合で首位に返り咲いた．続く木浪が中前 2 点適時打を放ち，再びリードを広げた.

これらの入力文は，理想的な見出しを T5 に生成させるため，記事全文から記事の主題に関する文と注目している選手の活躍に関する文のみを抽出している．

3.4　T5 による見出し生成

見出し生成に利用する文生成モデルである T5 について述べる．本研究では T5 の事前学習済みモデルとして sonoisa/t5-base-japanese[1]を用いる．このモデルを見出し生成に適した形にするためにファインチューニングをする．

見出しの生成には，ビームサーチを用いる．ビームサーチは

1：https://huggingface.co/sonoisa/t5-base-japanese

生成の各ステップにおいて次の単語として出現する可能性の高い，いくつかの単語を選択していく手法である．この手法を用いて，上位 100 件の見出しを T5 に出力させる．その後，1 位から順に生成した見出しを確認し，指定している選手名を含んでいる見出しを，指定した選手の活躍に注目した見出しとする．上位 100 件の中に指定している選手名を含んでいる見出しがひとつも存在しない場合は，生成失敗として処理する．

4 評価実験

生成した見出しの精度が実際に提案手法により向上したかどうかを確認するために，評価実験を行った．

4.1 評価基準

生成した見出しが良い見出しであるか判断するための評価基準を以下に示す．

- **事実と異なっていないか**
- **指定した選手の選手名を含んでいるか**
- **記事全体の要約を含んでいるか**
- **日本語として自然であるか**
- **見出しとして自然であるか**

このように，問題定義で定めた理想的な見出しが持つ５つの特徴を，生成した見出しが持っているかについて評価した．事実と異なっていないか，指定した選手の選手名を含んでいるか，記事全体の要約を含んでいるかについては，基準を満たしているものを 1，基準を満たしていないものを 0 として，2 値で評価を行った．日本語として自然であるか，見出しとして自然であるかについては，基準を満たしているものを 5，基準を満たしていないものを 1 として 5 段階で評価を行った．

4.2 ファインチューニング

ここでは，評価実験に利用する T5 のファインチューニングについて述べる．ファインチューニングに利用した訓練データは 2021 年のデイリースポーツの阪神タイガースに関する記事 2,738 件である．このデータは Ceek.jp News[2] というサイトにより収集した．このサイトは，様々なニュースサイトをクロールして Web ニュースを収集し，まとめているサイトである．

デイリースポーツが配信する Web ニュースは，阪神タイガースに関連するニュースの場合，そのニュースの URL に tigers という文字が含まれる．そこで，Web ニュースの中から URL に tigers という文字が含まれるニュースを抽出した．これらのデータを利用してファインチューニングすることで，T5 に見出し生成のタスクを解かせることができるようになる．

ファインチューニングを実行する際に定めた，各種ハイパーパラメータの設定を表 2 に示す．また本研究では patience を 10 とした早期終了によりエポック数を決定した．検証データには，訓練データと同様の方法で収集した 2021 年のデイリースポーツの阪神タイガースに関する記事 152 件を用いた．

2：https://news.ceek.jp/

表 2　ファインチューニング学習条件.

項目	内容
学習率	3.0×10^{-4}
バッチサイズ	4
最適化アルゴリズム	Adam Weight Decay [7]

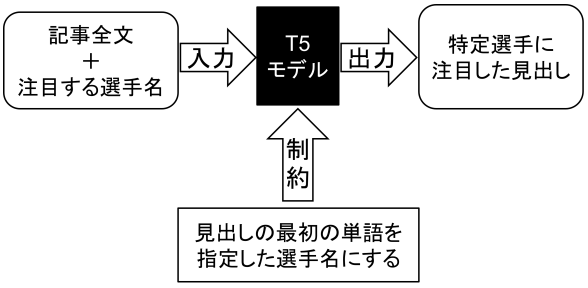


図 3　比較手法の全体像.

表 3　見出し生成時のハイパーパラメータ.

項目	内容
temperature	1
num beams	生成する見出しの数
diversity penalty	1
num beam groups	10
repetition penalty	10

4.3 比較手法

提案手法による見出し生成の精度の向上を確認するために，比較手法による見出し生成との比較を行った．比較手法では記事全文を入力文とした．文生成モデルの精度により生成される見出しの精度に差が出ないように，利用する文生成モデルは同じ訓練データで同じ学習条件によりファインチューニングした T5 とする．

比較手法では見出し生成時に，見出しの開始位置に注目する選手名のトークンを指定し，見出しを生成した．このように生成した見出しが，どの程度，評価基準を満たすことができるか確認した．比較手法の全体像を図 3 に示す．

4.4 見出し生成時のハイパーパラメータ

見出し生成をする際に定めた，ハイパーパラメータについて述べる．各種ハイパーパラメータの設定を表 3 に示す．temperature は生成にランダム性を加える温度パラメータである．値が高いほどランダム性が増す．num beams はビームサーチの探索幅である．diversity penalty は生成結果の多様性を生み出すためのペナルティである．num beam group はビームサーチのグループ数である．repetition penalty は同じ文の繰り返しを防ぐためのペナルティである．

4.5 評価用データセット

ここでは，評価用データセットの構築方法について説明する．評価用データの収集にも訓練データの収集に利用したサイトを用いた．訓練データでは 2021 年のデイリースポーツの阪神タイガースに関する記事を収集したが，評価用データでは 2022

年のデイリースポーツの阪神タイガースに関する記事を収集した．収集した記事から試合結果について報じている記事 50 件を抽出した．1 件の記事に対して 2 人の選手に注目する．注目する選手は記事本文に出てくる選手の中からランダムに決定した．提案手法で生成された最も上位の見出しと，比較手法で生成された最も上位の見出しを評価の対象とした．

評価している見出しが提案手法により生成した見出しなのか，比較手法により生成した見出しなのか評価者が分かってしまうと評価する際にバイアスがかかってしまう可能性がある．そこで，評価者がどちらの手法で生成された見出しなのか特定できないようにした．

4.6 評 価 者

見出しの精度を正確に評価するためには，プロ野球に関する知識や，プロ野球に関するニュース記事の見出しを頻繁に目にしている必要がある．そこで，評価者はプロ野球ファン 3 名とした．

4.7 評 価 結 果

本節では，評価者による評価の集計結果を示す．本実験は，人手による評価を行っている．そのため，評価者の評価の信頼性を確認する必要がある．評価者の信頼性を担保するために，各評価者の一致度を確認するため，Fleiss の κ 係数を各評価項目ごとに計算した．計算結果を表 4 に示す．κ 係数の評価に関して，本実験では Landis ら [6] によって提案された基準を使用した．この基準によれば，指定した選手の選手名を含んでいるかの項目は「Almost perfect」（ほとんど一致）と評価され，事実と異なっていないか，記事全体の要約を含んでいるかの項目は「Substantial」（かなりの一致）と評価された．また日本語として自然であるかの項目は「Fair」（ある程度の一致），見出しとして自然であるかの項目は「slight」（わずかな一致）であった．2 値評価である 3 項目については十分な一致が得られたといえる．5 段階評価である 2 項目については，2 値評価である 3 項目より κ 係数は下がるという結果が得られた．

評価の集計方法について説明する．2 値評価である場合，3 名の評価者の評価の多数決により集計した．5 段階評価である場合，3 名の評価者の評価の平均により集計した．その後，比較手法で生成された見出しと提案手法で生成された見出しを手法ごとに集計した．2 値評価のものは評価基準を満たした見出しの割合を計算し，5 段階評価のものは平均得点を計算した．

このように集計した結果を表 5，表 6 に示す．事実と異なっていないかという評価基準については提案手法のほうが精度が良いという結果が得られた．指定した選手名を含んでいるかという評価基準については比較手法のほうが精度が良いという結果が得られた．記事全体の要約を含んでいるかという評価基準については提案手法のほうがやや精度が良いという結果が得られた．日本語として自然であるか，見出しとして自然であるかという評価基準についても提案手法のほうがやや精度が良いという結果が得られた．また，提案手法と比較手法の結果の差が統計的に有意であるか確かめるため，有意水準 5% で，Welch の t 検定を行った．事実と異なっていないか，指定した選手名を含んでいるかという評価基準については，提案手法と比較手法の結果の差に統計的有意差が見られた．

表 4　κ 係数計算結果.

項目	κ 係数
事実と異なっていないか	0.788
指定した選手の選手名を含んでいるか	0.886
記事全体の要約を含んでいるか	0.621
日本語として自然であるか	0.212
見出しとして自然であるか	0.169

表 5　評価基準を満たした見出しの割合.

手法	事実か否か	指定した選手名	全体の要約
提案手法	0.56*	0.90	0.64
比較手法	0.25	0.99*	0.60

表 6　平均得点.

手法	日本語としての自然さ	見出しとしての自然さ
提案手法	3.59	2.99
比較手法	3.39	2.70

5 考　　察

本実験で得られた結果をもとに提案手法による見出し生成の精度の変化について考察する．

本実験により，提案手法は事実と異なっていない見出しを生成する可能性を上げることが分かった．提案手法と比較手法の事実と異なっていないかという評価基準における結果の差は統計的にも有意であった．これは，提案手法が記事全文ではなく理想的な見出しを生成するために必要な文のみを入力として使用していることが影響していると考えられる．これにより T5 が別の選手の活躍を注目している選手の活躍であるかのように見出しを誤生成するリスクを下げている．表 7 は実際に提案手法が別の選手の活躍を注目している選手の活躍であるかのように見出しを誤生成することを防いだ例である．記事全文を入力する比較手法では，実際に白星を挙げたのは藤浪投手であるにもかかわらず，佐藤選手が白星を挙げたかのような見出しを生成している．それに対して，提案手法では事実に即した見出しを生成することが出来ている．

しかし，提案手法においても事実と異なる見出しを生成することがある．これは注目選手の活躍に関する文を抽出する際に，別の選手の活躍に関する情報が含まれてしまう可能性があることが原因である．提案手法では注目選手の活躍に関する文を抽出するための基準を，注目選手の選手名が含まれているか否かとしている．そのため別の選手の活躍の情報も含まれてしまう場合がある．具体例として，「先頭の中野が右前打で出塁すると，続く熊谷は左翼線へ運ぶバスターエンドランを成功.」といった文が挙げられる．このような文は 1 つの文で 2 人の選手の活躍について述べられており，誤生成の原因となる可能性がある．

表 7　提案手法が別の選手の活躍を注目している選手の活躍であるかのように見出しを誤生成することを防いだ例.

記事全文：
阪神は３連勝を飾り，２年連続となる今季の巨人戦勝ち越しを決めた．先発・藤浪が２１年４月１６日・ヤクルト戦以来，４９１日ぶりとなる先発勝利で，今季初白星を挙げた．２試合連続の無四球で７回６安打１失点，５奪三振の好投．巨人戦の白星は１６年４月５日以来，２３２８日ぶりだ．この日は打線も藤浪を強力援護した．二回に２死一，二塁から梅野の右前適時打で先制．三回には２死からロハスが中前打で出塁すると，続く佐藤輝がバックスクリーン左へ飛び込む１８号２ランを放った．さらに３－０の四回には１死一，二塁で藤浪が三塁線への絶妙な犠打を決め，２死二，三塁と好機を演出．１番・中野が俊足を生かした三塁への適時内野安打で１点を追加した．なおも２死一，三塁では，島田がセーフティスクイズを決めて，五回までに５点のリードを奪った．
注目選手：
佐藤
提案手法の入力文：
阪神は３連勝を飾り，２年連続となる今季の巨人戦勝ち越しを決めた．巨人戦の白星は１６年４月５日以来，２３２８日ぶりだ．三回には２死からロハスが中前打で出塁すると，続く佐藤輝がバックスクリーン左へ飛び込む１８号２ランを放った．
提案手法：
阪神 2 年連続の巨人戦勝ち越し佐藤輝がバックスクリーン左へ 18 号 2 ラン
比較手法：
佐藤輝が今季初白星!藤浪が 491 日ぶり先発勝利

表 8　指定した選手名を含まない例.

記事全文：
阪神はまさかの逆転負け，横浜スタジアム７連敗となり，８連敗した９８年以来の屈辱となった．同球場では今季１０敗．これは７８年の開場以降，ワーストの数字そして最短で１１日にも優勝の可能性が完全消滅となる．６連戦の初戦を託された藤浪晋太郎投手（２８）が突如，崩れて大量失点された．通算１０００奪三振をこの試合で達成したが，素直に喜べない．２度の暴投が痛恨だ．１点差に詰め寄られた五回，先頭から連打を許し，その後１死二，三塁とピンチを招いた．９番の代打・オースティンに対する初球１５０キロ・スプリットがワンバウンドし，梅野が捕球できず，三走・楠本が生還し，同点を許した．なおも１死三塁で，２ボールから投じた１５５キロ直球が外角低めに外れ，梅野のグラブをかすめる暴投に．三走・伊藤光が生還し，まさかの形で勝ち越しを許した．結局，オースティンに四球を与えると，桑原，佐野に連打を浴びて１死満塁と窮地に陥った．ここで３番・佐野が高々と打ち上げた左翼線への打球が，左翼・ロハスの前にポトリと落ちる適時打に．ベンチはたまらず，交代が告げられた．序盤３回は，ＤｅＮＡ打線を１安打無失点に抑える圧巻の投球を見せていただけに，まさかの事態．４回１／３を今季ワーストの７失点でＫＯとなった.「ストレートに狙いを絞られて，うまく打たれてしまったという印象です．そんな中でも，なんとか粘りながら，もうひと踏ん張りしたかったなと思います」とコメントしている．打線は初回に近本の右翼線への先制打，大山の右犠飛で幸先良く先取点を奪ったが，二回以降は相手先発・上茶谷から決定打が出ず．三回までは毎回安打も，四回からはぱたっと安打も止まってしまう．打線の流れも悪く，藤浪が失点．悪循環が続いてしまい，敗戦につながってしまった．
注目選手：
伊藤
提案手法の入力文：
阪神はまさかの逆転負け，横浜スタジアム７連敗となり，８連敗した９８年以来の屈辱となった．これは７８年の開場以降，ワーストの数字そして最短で１１日にも優勝の可能性が完全消滅となる．三走・伊藤光が生還し，まさかの形で勝ち越しを許した．
提案手法：
三走・伊藤光がまさかの逆転負け横浜スタジアム 7 連敗
比較手法：
伊藤光が今季ワーストの 7 失点「もうひと踏ん張りしたかった」

生成された見出しに指定した選手名を含むかについて評価した結果では，比較手法のほうが良いという結果が得られた．この結果は，統計的にも有意であった．これは，提案手法は生成した上位 100 件の中に指定した選手名を含まない場合，生成失敗として処理されるためである．比較手法においても指定した選手名を含まない場合が存在する．その例を表 8 に挙げる．この例では伊藤将司投手と同じ名字の伊藤光選手に注目した見出しを生成してしまっている．このような同じ名字の選手をどのように対処するか検討する必要がある．

記事全体の要約を含む精度については提案手法のほうがわずかに高い結果が得られた．このことから，T5 に記事全文から記事の主題を見つけさせる手法よりも，記事の主題に関する文を事前に抽出しておき，その情報を T5 に渡すほうが精度がわずかに高くなることが分かる．しかし提案手法と比較手法の結果の差は統計的には有意でないため，今後より多くの評価データで精度を検証する必要がある．

日本語として自然さの生成精度についても提案手法のほうが比較手法よりわずかに高いという結果が得られた．比較手法が提案手法より日本語として不自然な見出しが多く生成された理由として，T5 に見出しの最初の単語を指定した選手名にする制約をかけたことが考えられる．ビームサーチによる生成で本来最初に配置されるべき単語ではなく，制約により強引に選手

名を最初の単語として配置している．これが日本語としての不自然さを生む原因になっていると考えられる．ただし，統計的な有意差は無いため今後より多くの評価データで精度を検証する必要がある．

また見出しとしての自然さの項目で提案手法が比較手法を上回ったことにより，提案手法による見出しのほうが比較手法による見出しより扱っている情報が見出しとして妥当であることが分かる．これは重要な文を事前に抽出することで T5 に与えられる情報が見出しとして扱う情報として，妥当なものにほぼ限定されることが要因であると考えられる．しかし，この項目についても統計的な有意差は無いため，今後より多くの評価データで精度を検証する必要がある．

6 まとめと今後の課題

本研究では，T5 による特定選手に注目したニュース記事の見出し生成に取り組んだ．文生成モデルが事実と異なっていない見出しを生成する可能性を上げるため，入力文を記事全文ではなく重要な文を抽出した抽出文とする手法を提案した．

今後の課題として，今回の実験ではすべて試合結果記事を対象とした見出しを生成している点がある．本研究では対象記事が野球の試合結果記事に絞られているが，重要文の抽出方法を工夫することで，政治的な記事や経済に関する記事などの別ジャンルの記事を対象としても汎用的に精度の高い見出し生成が可能になると考えられる．

また，現時点では，見出しの生成というタスクまでしか到達していない．本研究で提案した手法を用いたシステムを構築し，実用化するためには，これからも多くの検証が必要である．

謝辞 本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775，JP22H03905，による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[2] Jeffrey L Elman. Finding Structure in Time. *Cognitive science*, Vol. 14, No. 2, pp. 179–211, 1990.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[4] Kai Ishikawa, Shinichi Ando, and Akitoshi Okumura. Hybrid Text Summarization Method based on the TF Method and the Lead Method. *NTCIR Conference on Evaluation of Information Access Technologies*, 2001.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.

[6] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[10] Qian Ruan, Malte Ostendorff, and Georg Rehm. Histruct+: Improving Extractive Text Summarizationwith Hierarchical Structure Information. *arXiv preprint arXiv:2203.09629*, 2022.

[11] Karen Sparck Jones. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, Vol. 30, pp. 5998–6008, 2017.

[13] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization. *arXiv preprint arXiv:2302.08081*, 2023.

[14] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848*, 2023.

[15] 江原駿介, 荘司慶行, 山本岳洋, Martin J.Durst. スタンスに合わせたニュースタイトルの自動生成によるニュース記事本文へのアクセスを促す情報提示. 第 15 回データ工学と情報マネジメントに関するフォーラム, 3a–9–1, 2023.

[16] 高津弘明, 奥田真由, 松山洋一, 本田裕, 藤江真也, 小林哲則. 会話によるニュース記事伝達のための抽出型要約のパーソナライズ. 2020 年度人工知能学会全国大会（第 34 回）, 3Rin4-71, 2020.

[17] 黒木開, 岩井大志, 川上孝介, 石塚湖太, 中田和秀. キーワードを考慮した BERT2BERT による広告文生成. 2021 年度人工知能学会全国大会（第 35 回）, 2D3-OS-7a-02, 2021.

[18] 小川恒平, 永塚光一, 渥美雅保. ニューラルネットワークを用いたニュース記事の抽出的三行要約. 2019 年度人工知能学会全国大会（第 33 回）, 4Rin1-11, 2019.

[19] 福田芳樹, 伊東嗣功, 石井雅樹, 堂坂浩二. T5 を用いた感情要約文の生成. 言語処理学会 第 28 回年次大会, D8-2, 2022.

[20] 木村輔, 田上諒, 宮森恒. 文エンコーダによるクエリ指向要約モデルの強化. 情報処理学会論文誌, Vol. 12, No. 4, pp. 66–79, 2019.

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 16:00 ~ 18:10 | T2-A オンライン（Zoom Events)

先進ハードウェア活用・性能評価

座長:藤原 靖宏(日本電信電話株式会社)
コメンテータ:油井 誠(トレジャーデータ株式会社)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T2-A-6-01] サーバシステムの性能データ転送における効率化手法の評価と考察

*飯山 知香[1]、平井 聡[2]、山岡 茉莉[2]、福本 尚人[2]、小口 正人[1] (1. お茶の水女子大学、2. 富士通株式会社)

16:30 ~ 16:55
[T2-A-6-02] 量子ランダムアクセスメモリ上での関係データベースの構成法の一検討

*御手洗 陽紀[1]、合田 和生[2] (1. 東京大学 大学院情報理工学系研究科、2. 東京大学 生産技術研究所)

16:55 ~ 17:20
[T2-A-6-03] 永続メモリ向けロックフリー索引Bz木に関する研究

*中山 宗[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学大学院情報学研究科)

17:20 ~ 17:45
[T2-A-6-04] ラベルの出現頻度に着目したFPGAを用いた正規パス問合せの提案

*溝谷 祐大[1]、小林 諒平[2]、藤田 典久[2]、朴 泰祐[2]、天笠 俊之[2] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群情報学位プログラム　知識・データ工学研究室(KDE)、2. 筑波大学 計算科学研究センター)

17:45 ~ 18:10
[T2-A-6-05] 量子回路シミュレータの実行環境変化に伴う性能評価と特性比較

*青木 望美[1]、山崎 雅文[2]、平井 聡[2]、山岡 茉莉[2]、福本 尚人[2]、小口 正人[3] (1. お茶の水女子大学人間文化創成科学研究科博士前期課程理学専攻情報科学コース小口研究室、2. 富士通（株）、3. お茶の水女子大学理学部情報科学科)

サーバシステムの性能データ転送における効率化手法の評価と考察

飯山　知香† 　平井　聡†† 　山岡　茉莉†† 　福本　尚人†† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚２丁目１－１
†† 富士通株式会社　〒 211–8588 神奈川県川崎市中原区上小田中４丁目１－１
E-mail: †{chika,oguchi}@ogl.is.ocha.ac.jp, ††{ahirai,yamaoka.mari,fukumoto.naoto}@fujitsu.com

あらまし　近年，多数台サーバの共有利用などに関する需要が増加している．これらのサーバの負荷分散などをリアルタイムで行うためには，各サーバの性能データ分析を低オーバヘッドで行う必要がある．そこで本研究では，ターゲットサーバからデータ解析用サーバへのデータ転送時のオーバヘッドを考慮した効率的な転送手法の開発を目的としている．本報告では，転送効率化手法として，データ解析用サーバ性能の改善，転送タイミングの制御，および転送側で CPU 負荷の低いコアでの転送を行う．また，各手法に関して，データ転送時間や CPU 負荷率の計測を行い，その効果を評価する．さらに，CPU に負荷をかけるベンチマークと転送処理を同時実行し，それぞれに与える影響を分析する．

キーワード　時系列データ，データ転送，CPU 負荷率，転送タイミング制御，転送コア制御

Evaluation and Consideration of Improvement Plans to Increase the Efficiency of Performance Data Transfer for Server Systems

Chika IIYAMA†, Akira HIRAI††, Mari YAMAOKA††, Naoto FUKUMOTO††, and Masato OGUCHI†

† Ochanomizu University
2–1–1 Otsuka, Bunkyou-ku, Tokyo 112–8610, Japan
†† Fujitsu Limited
4–1–1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211–8588, Japan
E-mail: †{chika,oguchi}@ogl.is.ocha.ac.jp, ††{ahirai,yamaoka.mari,fukumoto.naoto}@fujitsu.com

Key words　Time-series data, data transfer, CPU load factor, transfer timing control, transfer core control

1. はじめに

近年，多数台サーバの共有利用や分散処理利用に関する需要が増えてきている．その際，各サーバのハードウェアレベルを含む詳細な性能データをリアルタイムで収集・分析・提示することで，より効果的な負荷分散やシステムのチューニングが可能になる．また，多数台サーバのデータを一元的にリアルタイムで分析する際，分析対象のサーバとその分析を行うサーバは分割され，サーバ間での性能データ転送が必要になる場合が多くある．さらに，時系列化された性能データはデータサイズが大きく，扱う際のオーバヘッドも大きくなる可能性がある．本研究では，この性能データとして，システム情報の中でも比較的詳細かつデータサイズの大きい，サーバの CPU コアごとに収集するデータを想定している．以上のことから，現在サーバ間でより効率的に性能データを転送する手法の実現が求められている．

2. 性能データ転送における課題と改善案

性能データ転送の効率化手法を検討するにあたり，[1] にてデータ収集を行うサーバと解析を行うサーバが分割された環境を想定して，データ収集・転送時のリソース利用量やオーバヘッドを計測した．その結果，複数コアの情報を 1 つのコアで転送する手法ではコア数に伴い転送処理の遅延が増大することが分かった．そこで [2] にて，1 つの CPU コアで行っていた転送処理を各 CPU コアで行うように並列化し，性能を評価・比較することで遅延の改善方法を検討した．その結果，解析を行うサーバの CPU コア数が少ないことが原因でデータ転送処理に遅延が発生している可能性が推察された．そのため，[3] にて，CPU コア数等の構成が異なる 2 台のサーバを使用して比較を行い，解析を行うサーバの性能差がデータ転送を行うサー

バでの処理に与える影響を分析した結果，解析を行うサーバの増強だけでは改善できないボトルネックがあることが推察された．また，解析を行うサーバを増強させた際の転送処理部分のCPU 負荷率が高く，転送処理と同一 CPU コア上で CPU 負荷ベンチマークを同時に実行させると少なからず互いに性能劣化が発生することが分かった．そこで本報告では，コアごとの転送タイミングをずらす転送タイミング制御を行い，転送効率が改善されるか分析した．加えて，転送側で CPU 負荷の低いコアでの転送を行う転送コア制御を行い，同時に動作しているベンチマークおよび転送処理において性能劣化が抑えられるか確認した．

3. 関 連 研 究

本研究と収集対象とする性能データが類似しているツールとして，Score-P と Vampir が挙げられる．Score-P(Scalable Performance Measurement Infrastructure for Parallel Codes) [4] というツールでは，主に並列 HPC アプリケーションの性能分析を行うことができる．Vampir(Visualization and Analysis of MPI Resources) [5] というツールでは，多数台の計算ノードのプロファイルデータや相互通信データなどを統合解析し可視化することができる．これらのツールでは，アプリケーションの実行後に，各ノードで収集した性能データをファイルベースで収集/分析/可視化するため，本研究が目指す実行時の性能データを時系列データとしてリアルタイムで収集・分析する手法とは異なる．既存研究 [6] では，スーパーコンピュータ京などの独自の CPU 性能データを汎用的なデータ形式に変換し，上記の Score-P や Vampir で扱えるようにするための手法が提案されている．分析しようとしている内容は本研究と類似しているが，データ収集と転送・分析を同時に行うことは困難である．

本研究とデータ収集時のオーバヘッド削減手法などの関連性が高い研究として，分散トレーシングという手法がある．クラウド環境で広く利用されているマイクロサービス・アーキテクチャに基づいたソフトウェアの詳細動作を解析することができる．分散トレーシングでは主にアプリケーションを対象領域として性能データを収集・分析しているのに対し，本研究では低レイヤを対象としている点が異なる．Google 社が 2010 年に発表した論文 [7] を発端に，Twitter 社が Zipkin [8] を，Uber 社が Jeager [9] を開発した．それぞれ OSS として公開され，多くのクラウド基盤で利用されている．

プロファイルデータをデータベースに取り込み分析する手法が本研究と類似している既存研究 [10] では，Google 社のデータセンターで行われているプロファイルデータの収集・分析手法が解説されている．既存研究 [11] では，上記手法を用いてデータセンターのワークロードが分析されている．数千台のサーバからランダムでプロファイルデータ収集を行い，その情報を元にアプリケーションチューニングを行うと共に，パフォーマンス監視やアプリケーション配備 (アフィニティ) 最適化にも利用している．本研究では CPU や OS 等の低レイヤの性能データをオープンなソフトウェア (OSS) を使用/改良して収集/転送/分析することを目指している．

4. 実　　験

4.1 評 価 環 境

環境構築として，データ収集用サーバ (以下収集サーバ) とデータ解析用サーバ (以下解析サーバ) を用意した．評価環境概要を図 1 に示す．収集サーバには，Linux Kernel の標準的なイベントデータ収集/トレース機能のフレームワークである perf [12] を使用した．また，CPU 負荷をかける性能ベンチマークとして，UnixBench [13] の Dhrystone，io 負荷をかける性能ベンチマークとして stress io を使用した．解析サーバには，収集した時系列データを管理する時系列 DB として InfluxDB [14] を使用した．InfluxDB については [15] で解説されている．事前に収集した stress [16] 実行時の perf record のデータをデータ転送プログラムを用いて転送し，解析サーバ上の InfluxDB に格納した．各使用技術のバージョンは，perf が 3.10.0，stress が 1.0.4，UnixBench が 5.1.3，InfluxDB が 1.8.3 である．

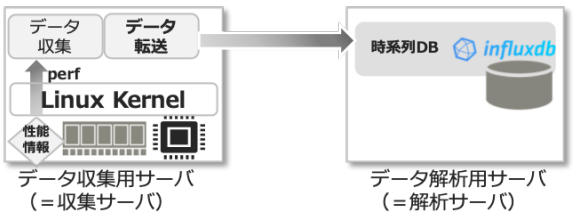


図 1　評価環境概要

解析サーバには，構成が異なる 2 台のサーバを用いて分析を行った．以降，CPU コア数，メモリ容量が多いサーバを高性能解析サーバ，少ないサーバを低性能解析サーバと記載する．収集サーバと高性能解析サーバの構成を表 1，低性能解析サーバの構成を表 2 に示す．各サーバのネットワーク構成図を図 2 に示す．

表 1　収集サーバ・高性能解析サーバ環境

機種名	Fujitsu PRIMERGY CX2550 M1
CPU	Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz (× 2CPU)
コア数	14
メモリ	128GB
OS	CentOS 7

表 2　低性能解析サーバ環境

機種名	Dell PowerEdge R620
CPU	Intel(R) Xeon(R) CPU E5-2603 v2 @ 1.80GHz (× 1CPU)
コア数	4
メモリ	16GB
OS	CentOS 7

収集した CPU 性能データを，InfluxDB に転送可能な形式に変換し，転送するまでの流れを以下に示す．まず収集データか

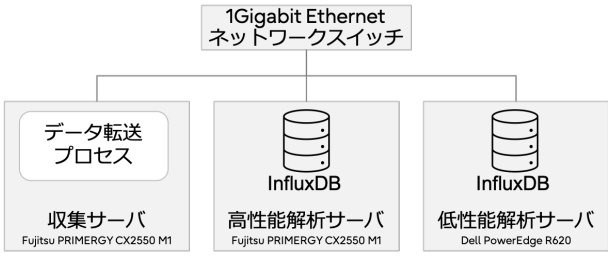


図 2　ネットワーク構成図

ら，必要なデータ (時刻，プロセス ID，スレッド ID，実行アドレス) のみを抽出する．1CPU コアの 1 秒分の収集データからは，100 マイクロ秒単位で収集した場合は 10000 レコードのデータが抽出される．この時，それぞれ 64bit のデータが 4 項目あるため，1CPU コアに関する転送データサイズは約 320KB となる．次に，抽出データを Python Pandas [17] を利用して DataFrame 形式に格納する．DataFrame 形式に格納したデータのサンプルを図 3 に示す．最後にこのデータを InfluxDB の Python Client モジュールを利用して解析サーバに転送する．上記モジュールには Pandas の DataFrame 形式データを InfluxDB サーバに書き込む API(Influxdb.DataFrameClient) [18] があり，今回はこれを利用している．

時刻(ns単位)	Process ID	Thread ID	EIP(実行アドレス)
2021-12-20T13:18:36.159557443	698085	698085	0x55e3e9dd1151
2021-12-20T13:18:36.160568219	698085	698085	0x7faa24bd7e02

図 3　DataFrame 形式データのサンプル

4.2　転送タイミング制御

4.2.1　実 験 内 容

転送タイミング制御の効果を検証するため，以下の実験 1〜3 の計測を行なった．実験 1 では，CPU コア数を変更しながらデータ転送時間を計測した．実験 1 で得られた結果の裏付けを取るため，CPU コア数 n=14 で実験 2〜3 を行なった．CPU コア数 n=14 で計測を行ったのは，コア数が多いほどオーバヘッドは分かりやすく，また表 1 に示すように収集サーバの 1CPU あたりのコア数は 14 であるためである．実験 2 では，データ転送処理部分の CPU 負荷率を計測した．実験 3 では，InfluxDB の CPU 負荷率を計測した．転送タイミング制御は，CPU コア数 n での転送において各コア i で転送前に i/n 秒 sleep し，コアごとの転送タイミングを 1/n 秒ずつずらすことで実装した．

4.2.2　実 験 結 果

実験 1 のデータ転送時間計測の結果を図 4 に示す．図 4 の各 CPU コア数における平均転送時間を表 3 に示す．図中の各点がコアを表している．表 3 ではスペースの都合上，低性能解析サーバはパターン 1，高性能解析サーバはパターン 2，低性能解析サーバ＋転送タイミング制御ありはパターン 3，高性能解析サーバ＋転送タイミング制御ありはパターン 4 と表記している．6 回計測を行い，初回転送のオーバヘッドによる影響を避けるため，2〜6 回目の結果のうち，全コアの平均のデータ転送時間が中間のものを用いた．低性能解析サーバおよび高性能解析サーバ使用時ともに，転送を実行するコア間で転送時間にばらつきがあり，CPU コア数の増加に伴いばらつきも増加していたが，転送タイミング制御を加えたことでそれらが減少し，CPU コア数の増加に伴う転送時間の増加が抑えられたことが読み取れる．また，転送タイミング制御を行うことで，低性能解析サーバ使用時においても，高性能解析サーバ使用時と同等の転送時間で転送できることが分かる．以上の結果から，データ転送時間の短縮において，転送タイミング制御は有用な手法であると推察される．

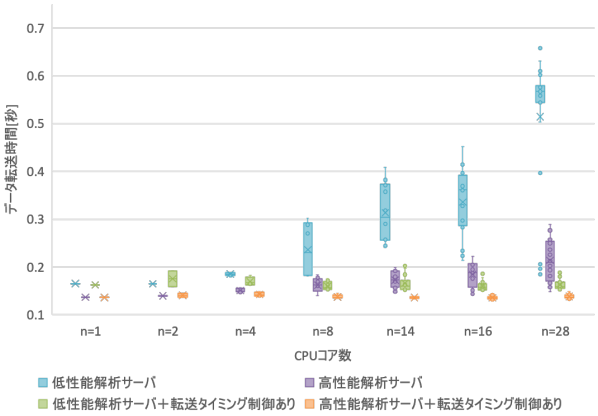


図 4　データ転送時間 (転送タイミング制御)

表 3　図 4 各 CPU コア数における平均転送時間

CPU コア数	パターン 1	パターン 2	パターン 3	パターン 4
n=1	0.17 秒	0.14 秒	0.16 秒	0.14 秒
n=2	0.17 秒	0.14 秒	0.18 秒	0.14 秒
n=4	0.19 秒	0.15 秒	0.17 秒	0.14 秒
n=8	0.24 秒	0.16 秒	0.16 秒	0.14 秒
n=14	0.31 秒	0.17 秒	0.17 秒	0.14 秒
n=16	0.34 秒	0.18 秒	0.16 秒	0.14 秒
n=28	0.51 秒	0.21 秒	0.16 秒	0.14 秒

実験 2 のデータ転送処理部分の CPU 負荷率計測の結果を図 5 と表 4 に示す．図 5 では各コアの CPU 負荷率とデータ転送時間を抽出した結果を示している．図中の各点がコアを表している．表 4 では表 3 と同様に，スペースの都合上，低性能解析サーバはパターン 1，高性能解析サーバはパターン 2，低性能解析サーバ＋転送タイミング制御ありはパターン 3，高性能解析サーバ＋転送タイミング制御ありはパターン 4 と表記している．6 回計測を行い，初回転送のオーバヘッドによる影響を避けるため，2〜6 回目の結果のうち，全コアの平均のデータ転送時間が中間のものを用いた．図 5 より，全てのパターンで共通して，転送時間が長くなるほど CPU 負荷率は低くなっていることから，コアごとの転送時間のばらつきは，転送処理中に発生した Idle 時間によるものであると推測できる．また，図 5 より，転送タイミング制御を加えることで，低性能解析サーバおよび高性能解析サーバ使用時ともに，転送時間および CPU 負荷率のばらつきが減少していることが読み取れる．以上の結果

から，転送タイミング制御による転送時間の短縮は，転送中の Idle 時間が減少したためであると推察される．表 4 より，最大 CPU 負荷率は 67.4%程度と高くなっていることが読み取れる．これによる同一 CPU コア上で動作する処理への影響が大きければ，転送用のコアを分けるなどの対応が必要になると考えられる．

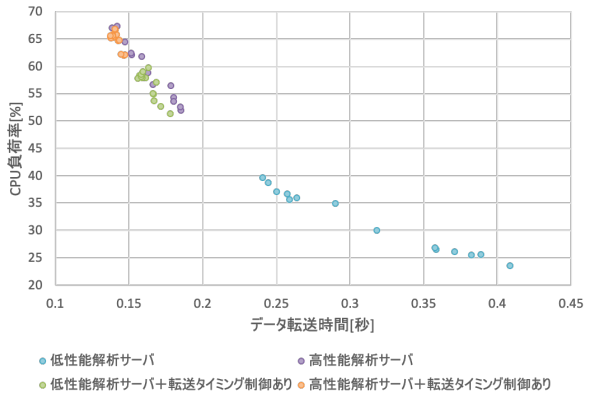


図 5　データ転送処理部分の CPU 負荷率 (n=14，並列転送版)

表 4　データ転送処理部分の CPU 負荷率 (n=14，並列転送版)

	パターン 1	パターン 2	パターン 3	パターン 4
平均	31.6 %	58.7 %	56.6 %	64.9 %
最大	40.0 %	67.4 %	59.8 %	66.8 %
最小	23.6 %	52.0 %	51.4 %	62.1 %

実験 3 の InfluxDB の CPU 負荷率計測について，低性能解析サーバでの結果 (パターン 1) を図 6，高性能解析サーバでの結果 (パターン 2) を図 7，低性能解析サーバ＋転送タイミング制御ありでの結果 (パターン 3) を図 8，高性能解析サーバ＋転送タイミング制御ありでの結果 (パターン 4) を図 9 に示す．図には 1 回計測の値を用いた．転送タイミング制御を加えたことで，低性能解析サーバおよび高性能解析サーバともに，CPU 負荷がかかる時間の範囲が広がり，各コアの CPU 負荷率の値は多少上下するが，CPU 負荷の総量はほとんど変化していないことが読み取れる．これは，転送元のコアにおいて転送タイミングが最大 1 秒程度ずれることで，転送先の InfluxDB でも負荷がかかるタイミングにずれが生じるが，転送されるデータ量に変化はないためであると推察される．よって，転送タイミング制御による転送時間の短縮は，InfluxDB 側の CPU 負荷率によるものではないと推察できる．

4.3　転送コア制御

4.3.1　実 験 内 容

転送コア制御の効果を検証するため，以下の実験 1〜5 の計測を行なった．実験 1 では，CPU コア数を変更しながらデータ転送時間を計測した．実験 2〜5 では，データ転送プログラムの CPU コア数は n=14 で計測を行なった．CPU コア数 n=14 で計測を行ったのは，転送タイミング制御の実験と同様，コア数が多いほどオーバヘッドは分かりやすく，また表 1 に示すように収集サーバの 1CPU あたりのコア数は 14 であるためである．実験 2 では，データ転送プログラムと CPU 負荷ベンチマークを同時に動作させた際の相互影響を計測した．実験 3 では，データ転送プログラムと CPU 負荷ベンチマークに加えて io 負荷ベンチマークも同時に動作させた際の相互影響を計測した．実験 4 では，実験 3 の計測時の CPU 負荷率を計測した．

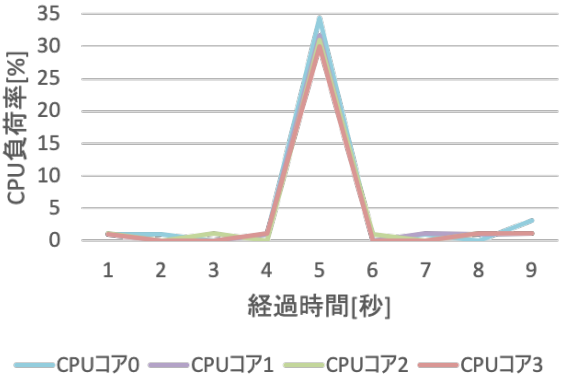


図 6　InfluxDB の CPU 負荷率 (n=14，並列転送版，低性能解析サーバ)

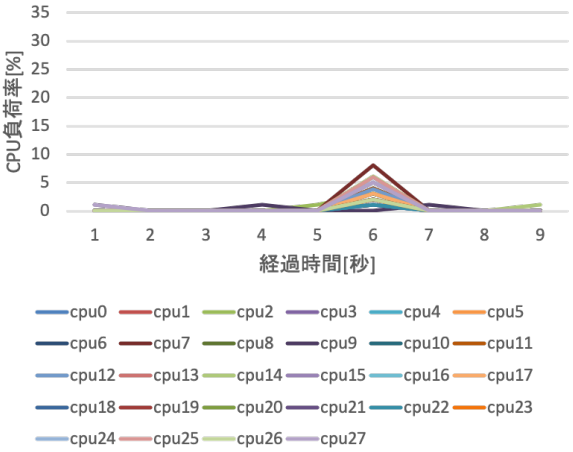


図 7　InfluxDB の CPU 負荷率 (n=14，並列転送版，高性能解析サーバ)

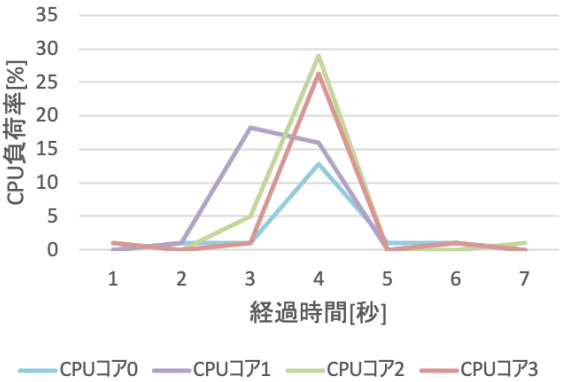


図 8　InfluxDB の CPU 負荷率 (n=14，並列転送版，低性能解析サーバ＋転送タイミング制御あり)

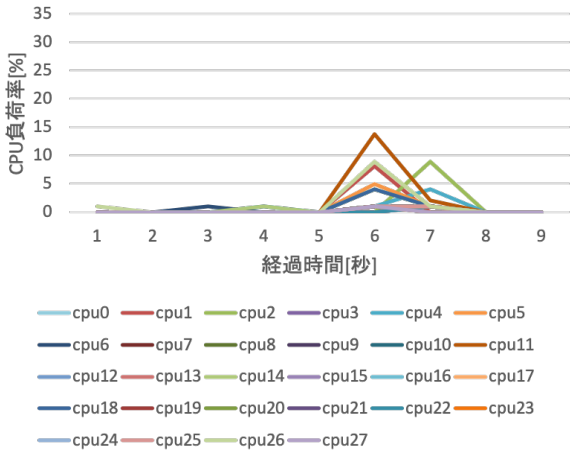


図 9　InfluxDB の CPU 負荷率 (n=14，並列転送版，高性能解析サーバ＋転送タイミング制御あり)

実験 5 では，実験 3 の計測時の各プロセスのコア遷移を計測した．転送コア制御は，転送側で毎秒 Idle 率の高いコアを検出し，該当コアで 4 コア分 (または端数分) のデータの転送を行わせることで実装した．そのため，転送は全コアではなく”n/4 の切り上げ”個のコアで実行される．該当コアで転送するデータ量は，1 コア転送版にて転送に 1 秒以上かからないコア数分とした．

4.3.2　実験結果

実験 1 のデータ転送時間の結果を図 10 に示す．図中の各点がコアを表している．CPU コア数 n=14 にて，3 つのコアでは 4 コア分，1 つのコアでは端数の 2 コア分のデータを転送しているため，一部転送時間が短くなっている．n が 4 以下の場合は 1 コア転送版と同等の転送時間になっていることが読み取れる．また，n が 4 より大きい場合はおおよそ n=4 の時と同等の転送時間になっているが, CPU コア数の増大に伴い転送時間および転送時間のばらつきが増大していることが読み取れる．これは，複数の Idle コアから同時に転送を行うようになることで，InfluxDB 側の負荷等により転送時の Idle 時間が増加するオーバヘッドが発生したためであると推察される．

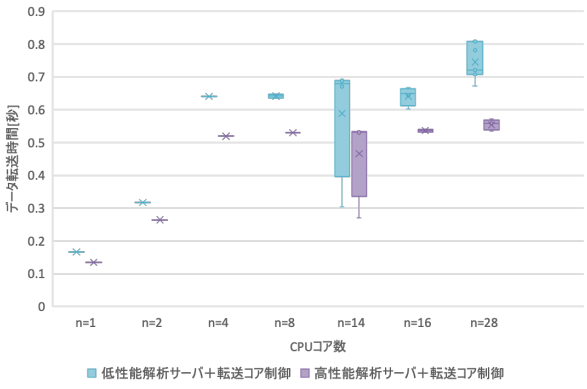


図 10　データ転送時間 (転送コア制御)

データ転送プログラムと CPU 負荷ベンチマークを同時に動作させた際の相互影響を計測した実験 2 の結果を表 5 に示す．データ転送プログラムの CPU コア数は n=14，つまり転送を実行するコアは 4 つで計測した．CPU 負荷ベンチマークとして，UnixBench のテストケースの一つで整数演算処理を行う dhry2reg を使用し，10 個のコアで実行した．データ転送プログラムとベンチマークプログラムをそれぞれ単独で動作させた場合と同時に動作させた場合を比較し，データ転送時間伸び率，ベンチマーク値劣化率，ベンチマーク実行時間伸び率を算出した．表には 3 回計測の平均の値を用いた．表 5 より，伸び率および劣化率が全て 1.00 であることから，同時動作においても互いにほとんど影響を与えていないことが分かる．これは，転送コア制御により CPU 負荷率の低いコア (ベンチマークの動作していないコア) が検出され，転送処理が実行されたためであると推察される．よって，転送コア制御を行うことで，転送処理とベンチマークの動作するコアを分割することが可能であり，その場合互いに性能劣化を起こさずに実行できることが分かる．

表 5　dhry2reg(10 コア) の動作時 (コア指定あり)

	データ転送時間	ベンチマーク値	ベンチマーク実行時間
単独動作	0.61 秒	26247.8	13.3 秒
同時動作	0.61 秒	26167.4	13.2 秒
伸び率/劣化率	1.00	1.00	1.00

データ転送プログラムと CPU 負荷ベンチマークに加えて io 負荷ベンチマークも同時に動作させた際の相互影響を計測した実験 3 の結果を表 6 に示す．CPU コア数 n=14 でのデータ転送プログラムと同時に，CPU 負荷ベンチマークである dhry2reg を 10 コア，io 負荷ベンチマークである stress io を 4 コアで実行させた場合の，dhry2reg とデータ転送プログラムの性能劣化を分析した．表には 3 回計測の平均の値を用いた．なお，ベンチマークが動作するコアは 14 コアの範囲内では指定せずに計測を行なった．表 6 より，転送処理は多少劣化しているが，dhry2reg ベンチマークではほとんど性能劣化が発生していないことが読み取れる．これは，CPU 負荷の低いコアを検出して転送を行わせる転送コア制御により，同一 CPU コア上で転送処理と stress io ベンチマークが同時に動作し，それらと異なるコア上で dhry2reg ベンチマークが動作したためであると推察できる．

表 6　dhry2reg(10 コア) と stress io(4 コア) の動作時 (コア指定なし)

	データ転送時間	ベンチマーク値	ベンチマーク実行時間
単独動作	0.61 秒	26247.8	13.3 秒
同時動作	0.62 秒	26187.5	13.2 秒
伸び率/劣化率	1.03	1.00	1.00

実験 4 では，実験 3 の計測時の各コアの CPU 負荷率を計測し，ベンチマーク特性を確認した．なお，ベンチマークが動作

するコアは 14 コアの範囲内では指定せずに計測を行なった．結果を図 11〜14 に示す．図 11 より，15〜24 秒でユーザレベルの CPU 負荷率が 100%近くなっていることが読み取れる．これは，dhry2reg ベンチマークは整数演算の繰り返し処理を行うためであると考えられる．また，図 12，13，14 より，システムレベルの CPU 負荷率やディスク I/O リクエスト待ち，idle 時間が多く発生していることが読み取れる．これは，stress io ベンチマークは sync システムコールによるディスクキャッシュ同期の繰り返し処理を行うためであると考えられる．

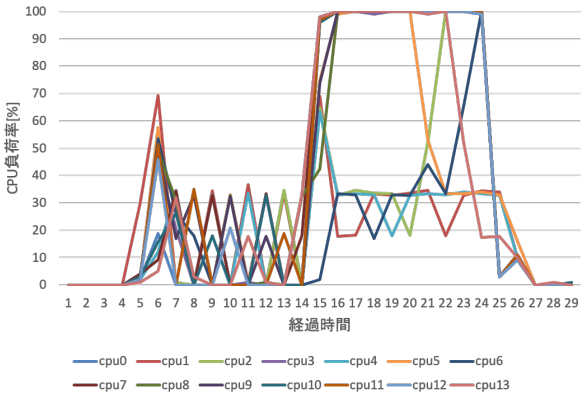


図 11　CPU 負荷率 (ユーザレベル)

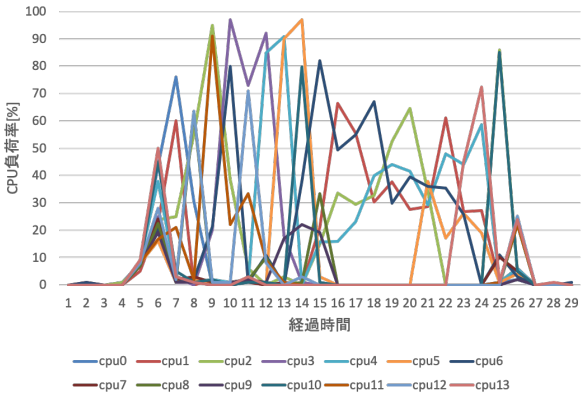


図 12　CPU 負荷率 (システムレベル)

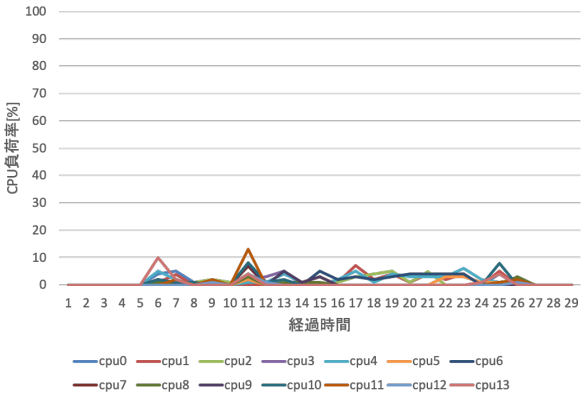


図 13　CPU 負荷率 (ディスク I/O リクエスト待ち)

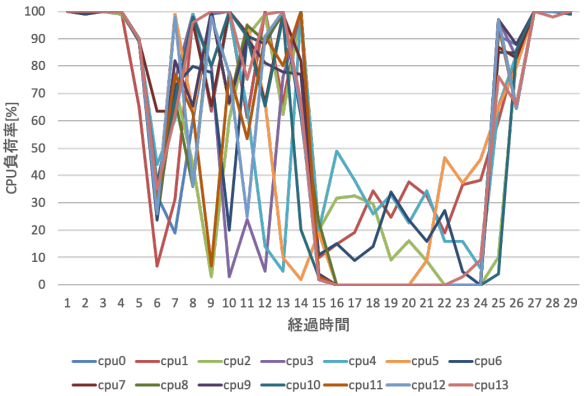


図 14　CPU 負荷率 (idle 時間)

実験 5 では，実験 3 の計測時の各プロセスのコア遷移を計測した．結果を表 7 に示す．dhry2reg ベンチマークはほとんどコア遷移していないことが読み取れる．これは，dhry2reg ベンチマークはユーザレベル 100%で動作し続ける (OS が遷移させない) ためであると考えられる．また，データ転送プログラムと stress io ベンチマークは比較的コア遷移の割合が高いことが読み取れる．データ転送プログラムは，実行した約 20 秒間のうち，dhry2reg ベンチマークが動作している 10 秒ほどの間は dhry2reg ベンチマークが動作していないコアを使用し，それ以外ではほとんど毎回使用するコアが遷移していたため，コア遷移の割合は平均 0.50 ほどになった．stress io ベンチマークも同様に，dhry2reg ベンチマークが動作している間は dhry2reg ベンチマークが動作していないコアを主に使用していた．また，全体を通して処理の繰り返し毎に 4 コア中 1 コアが遷移することが多く，コア遷移の割合は平均 0.25 ほどになった．これは，stress io ベンチマークはディスク I/O リクエスト待ちやアイドル時間が多く，コア遷移しやすいためであると考えられる．

表 7　コア遷移の割合

データ転送プログラム	0.50
dhry2reg	0.01
stress io	0.25

5.　まとめと今後の予定

本報告では，データ転送の効率化手法として，転送タイミング制御と転送コア制御を行い，その効果を評価した．その結果，低性能解析サーバ使用時においても，転送タイミング制御を行うことで，高性能解析サーバ使用時と同等の転送時間で転送できることが分かった．また，転送コア制御を行うことで，転送処理による同一 CPU 上で動作する処理の性能劣化を抑えられることが分かった．今後は，コアごとの転送タイミングをずらす転送タイミング制御において，より効率の良いずらし方の検討を行う．また，転送側で CPU 負荷の低いコアを検出して転送を行わせる転送コア制御にて，CPU 負荷の低いコアのより適切な予測・抽出方法についても検討していく．

謝　　辞

本研究の一部はお茶の水女子大学と富士通株式会社との共同研究契約に基づくものであり，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] 飯山知香, 平井聡, 山岡茉莉, 福本尚人, and 小口正人. サーバシステムの性能データ収集および転送における効率化手法の考察. In マルチメディア、分散、協調とモバイル *(DICOMO2022)* シンポジウム, 2022.
[2] 飯山知香, 平井聡, 山岡茉莉, 福本尚人, and 小口正人. サーバシステムの性能データ収集および転送効率化に向けた改善案の評価. In 第 *15* 回データ工学と情報マネジメントに関するフォーラム *(DEIM2023)*, 2023.
[3] 飯山知香, 平井聡, 山岡茉莉, 福本尚人, and 小口正人. サーバシステムの性能データ転送効率化に向けた改善案の評価. In マルチメディア、分散、協調とモバイル *(DICOMO2023)* シンポジウム, 2023.
[4] score-p. `https://www.vi-hps.org/projects/score-p`.
[5] vampir. `https://vampir.eu`.
[6] 阿部文武, 中村朋健, and 志田直之. プロファイラにおける汎用的な cpu 性能情報と表示機能. Technical Report 18, 富士通株式会社 次世代テクニカルコンピューティング開発本部, 2 2016.
[7] B. H. Sigelman, L. A. Barroso, M. Burrows, P. Stephenson, M. Plakal, D. Beaver, S. Jaspan, and C. Shanbhag. Dapper, a large-scale distributed systems tracing infrastructure. *Google Technical Report dapper-2010-1*, 2010.
[8] zipkin. `https://zipkin.io`.
[9] jeager. `https://www.jaegertracing.io/`.
[10] Gang Ren, Eric Tune, Tipp Moseley, Yixin Shi, Silvius Rus, and Robert Hundt. Google-wide profiling: A continuous profiling infrastructure for data centers. *IEEE Micro*, pages 65–79, 2010.
[11] Svilen Kanev, Juan Pablo Darago, Kim Hazelwood, Parthasarathy Ranganathan, Tipp Moseley, Gu-Yeon Wei, and David Brooks. Profiling a warehouse-scale computer. In *2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA)*, pages 158–169, 2015.
[12] perf. `https://perf.wiki.kernel.org/index.php/Main_Page`.
[13] Unixbench. `https://code.google.com/archive/p/byte-unixbench/`.
[14] influxdb. `https://www.influxdata.com/products/influxdb/`.
[15] Syeda Noor Zehra Naqvi and Sofia Yfantidou. Time series databases and influxdb, université libre de bruxelles. In *Advanced Databases Winter Semester 2017-2018*, 2017.
[16] stress. `https://linux.die.net/man/1/stress`.
[17] pandas. `https://pandas.pydata.org`.
[18] Dataframeclient. `https://influxdb-python.readthedocs.io/en/latest/api-documentation.html#dataframeclient`.

量子ランダムアクセスメモリ上での関係データベースの構成法の一検討

御手洗洗陽紀† 合田 和生††

† 東京大学大学院情報理工学系研究科電子情報学専攻 〒 116–8656 東京都文京区本郷 7–3–1
†† 東京大学生産技術研究所 〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{hmitarai,kgoda}@tkl.iis.u-tokyo.ac.jp

あらまし 量子コンピュータでは，量子的な重ね合わせ状態にあるデータに対して同時並列的に演算を行うことが可能であり，この性質は多くの量子アルゴリズムにより利用されている．しかし，量子的な重ね合わせ状態を準備する過程が効率的に行えなければ，量子アルゴリズムの古典アルゴリズムに対する優位性が失われる可能性がある．そのため，量子計算で用いるデータを量子的な重ね合わせ状態として効率的に準備する技術は非常に重要であり，近年盛んに研究されている．本研究では，ゲート型量子コンピューターにおける量子状態準備において関係データベースの概念を導入し，その基本的な実装や計算量の見積もりについて検討を行う．本研究で提案する量子版のデータベースシステムは，今後量子計算の基盤技術として重要な役割を果たすことが期待される．

キーワード データベース技術，量子コンピューティング，量子情報技術，QRAM，問い合わせ処理

1 はじめに

量子コンピュータでは，量子的な重ね合わせ状態にあるデータに対して同時並列的に演算を行うことが可能である．1992 年に提案された Deutsch–Jozsa のアルゴリズム [1] は，そのような量子コンピュータの性質を利用し，古典計算に対する量子計算の優位性を示した初めてのアルゴリズムである．Deutsch らは量子コンピュータに有利で実用性に乏しい問題設定ではあるものの，古典コンピュータで指数回数オラクル [1] に質問する必要があるのに対し，量子コンピュータでは 1 回の質問で問題を解くことができると示した．その後，1990 年代から 2000 年代にかけて，現在でも活発に議論される Grover のアルゴリズムや Shor のアルゴリズムといった量子アルゴリズムが多く提案された [2–4]．2010 年代以降は量子機械学習などのより応用的なアルゴリズムが提唱され，量子コンピュータの産業応用への期待が高まっている [5–12]．また，ハードウェアに目を向けてみると 1999 年に NEC の中村らが超伝導量子ビット (qubit) の基礎となるクーパー対箱を実証して以降，目覚ましい勢いで研究開発が進められている [13–19]．IBM は 2023 年に 1121 qubit の量子プロセッサーを発表し，さらに 2030 年頃には誤り訂正機能を持ったプロセッサを開発することを目標としている [2]．

ゲート型量子コンピュータでの計算は大きく分ければ 3 つの手順に分けられる．1 つ目は量子状態の準備，2 つ目は量子アルゴリズムを用いた計算，3 つ目は量子状態からの結果の読出しである．これら 3 つの手順のうち，ボトルネックになることが特に懸念されているのは，1 つ目の量子状態の準備と 3 つ目の量子状態からの結果の読出しである．例えば，線形方程式を解く HHL アルゴリズム [4] では，N 次元のベクトル $\boldsymbol{x}, \boldsymbol{b}$ と，$N \times N$ 次元の行列 A に対して，連立 1 次方程式 $A\boldsymbol{x} = \boldsymbol{b}$ の解 $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ を $O(\mathrm{poly}(N))$ の計算量で量子状態に埋め込むアルゴリズムである．しかし，このアルゴリズムではベクトル $\boldsymbol{b}$ が振幅符号化により量子状態埋め込まれていることを前提しているうえ，結果の読み出しについても，愚直に行った場合 $O(N)$ 回アルゴリズムを繰り返さなければならない．量子状態からの結果の読出しにおけるボトルネックは，結果を格納した量子状態の測定時に重ね合わせ状態が壊れることに由来しており，アプリケーションの工夫などで対応される [20, 21]．一方，量子状態の準備に関しては，効率的な量子ランダムアクセスメモリ (QRAM) [22–27] の存在が仮定されることが多い．しかしながら，QRAM の存在を仮定しただけでは実世界の複雑で互いに関係を持ったデータを，量子状態にエンコードするには非効率的である可能性がある．古典コンピューティングにおいてもデータが膨大になると，データベースでのデータの検索やそのデータのメモリ上への配置は大きなボトルネックとなる場合が多い．これらの課題を解決するうえで強力なツールとなるのはデータベース管理システム (DBMS) である [28]．我々は，古典的な関係データベースシステムの手法を量子状態準備に導入することで複雑で互いに関係を持ったデータを効率的に量子的な重ね合わせ状態として準備することができると考えた．

本研究では，ゲート型量子コンピューターにおける量子状態準備において，関係データベースの概念を導入する．DBMS の中でも，最も広く使われている関係データベース管理システム (RDBMS) での問い合わせ処理において重要な演算は，結合演算 (join) である [28, 29]．古典コンピュータにおける結合演算には，nested-loop join，hash join，sort-merge join など方法が用いられているが，量子コンピューター上での結合演算ではこれらの手法をそのまま流用することはできず，量子コンピューターで実現可能な方法を検討する必要がある．本稿では特に，結合演算に重きを置いて，ゲート型量子コンピューターにおける関係データベースの構成法について検討を行う．今回

1：本稿においてオラクルとは，正解を出力する何らかの量子回路を意味する．
2：https://www.ibm.com/quantum/blog/quantum-roadmap-2033, [Online; accessed 7-Jan-2024]

提案する構成法は大きく 2 つに分けられる．1 つ目は，QRAM 上での関係データベースの構成である．我々は，参照先のリレーションのタプルの QRAM 上でのアドレスを foreign key と一緒に保存しておくことにより，高速な結合演算が実行可能であることを示す．2 つ目は，近似的な振幅符号化 (Approximate Amplitude Encoding; AAE) を用いた関係データベースの構成である．AAE は QRAM と比べてゲート数が少なく済み，より near-term な技術であると考えられる．我々は，AAE に適した結合演算手法の候補として，Grover のアルゴリズムを用いた結合演算を提案する．また，QRAM，AAE いずれを用いた量子版関係データベースにおいても実行可能な汎用的な結合演算も提案する．この手法は，リレーションのタプルの数が非常に小さいケースでは，高速に結合演算が可能であり，事前に用意する QRAM などを削減できるといった利点がある．複雑で互いに関係を持ったデータを効率的に量子状態にエンコードすることができる量子版のデータベースシステムは，今後量子計算の基盤技術として重要になる可能性がある．

本稿の構成は以下の通りである．まず，第 2 節ではデータを量子的な重ね合わせ状態としてエンコードする手法である，QRAM と AAE を紹介し，第 3 節ではデータ検索を行う量子アルゴリズムである Grover のアルゴリズムを紹介する．そして，第 4 節において，本稿で提案する量子版のデータベースシステムの構成法について述べ，第 5 節では，その構成法について考察を行う．最後に，第 6 節で本稿のまとめを述べる．

2　データの量子状態へのエンコーディング

データのアドレスを量子状態 $|i\rangle$ で与えて，そのアドレスに対応する n ビットのデータ $D_i = \{0,1\}^n$ を量子状態 $|D_i\rangle$ として取り出す機能を量子ランダムアクセスメモリ (QRAM) と呼ぶ．QRAM は，具体的には以下で表される機構である [22].

$$\sum_i |i\rangle |0^n\rangle \xrightarrow{\text{QRAM}} \sum_i |i\rangle |D_i\rangle \tag{1}$$

古典的なランダムアクセスメモリとの違いは，アドレスは重ね合わせ状態で与えられ，取り出されるデータも重ね合わせ状態になっている点である．QRAM の機能はユニタリ変換として実現されると仮定されることが多く，その具体的な実装方法は現在盛んに研究されている．

しかしながら，厳密な QRAM では多くの量子ゲートが必要となるため，近い将来での実現は難しいと考えられる．そこで，データの量子的な重ね合わせ状態へのエンコーディング手法として，Parameterized Quantum Circuit (PQC) を用いた近似的な振幅符号化 (Approximate Amplitude Encoding; AAE) が提案されている [20]．この手法は，エンコードされた量子状態に誤差が生じることと引き換えに，少ないゲート数で実装可能であるため，QRAM と比べて near-term な技術であると考えられる．振幅符号化は，$N = 2^n$ 次元の実ベクトル $\boldsymbol{d} = (d_1, d_2, \cdots, d_N)^\top$ を以下のように n 量子ビット状態 $|D_{\text{AE}}\rangle$ の振幅として埋め込む符号化方法である．

$$|D_{\text{AE}}\rangle = \sum_{j=1}^{N} d_j |j\rangle \tag{2}$$

AAE では，ユニタリ変換 $U_{\text{AAE}}(\boldsymbol{\theta})$ で表される PQC を用いることで，式 (2) で表される状態を近似する．ただし，$\boldsymbol{\theta}$ は PQC のパラメータのベクトルである．

$$U_{\text{AAE}}(\boldsymbol{\theta}) |0^n\rangle = \sum_{j=1}^{N} a_j |j\rangle \tag{3}$$

ここで，AAE の目標は $a_j = d_j \; \forall j$ を $O(ln)$ のゲート数で達成することである．l は回路の深さであり，$O(\text{poly}(n))$ 程度の大きさであると想定されている．つまり，AAE を用いることで，$O(\text{poly}(n))$ 程度のゲート数で，データを効率的に量子状態にエンコードすることができると考えられている．ただし，事前に $U_{\text{AAE}}(\boldsymbol{\theta})$ のパラメータ $\boldsymbol{\theta}$ をデータ $\boldsymbol{d}$ に合わせて学習させる必要がある．

以上のように，データを量子的な重ね合わせ状態にエンコードする技術として QRAM と AAE を取り上げた．しかしながら，これらの技術だけでは実世界の複雑で互いに関係を持ったデータを，量子状態にエンコードするには非効率的である可能性がある．例えば図 1 に示すリレーション STUDENT，LAB，そして問い合わせ STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB を考える。これらの内ある 1 つのリレーションや問い合わせ結果を選び，そのタプルすべてを計算基底符号化により量子的な重ね合わせ状態としてエンコードする場合，ナイーブな実装では，すべてのリレーションや問い合わせに対して 1 つずつ QRAM あるいは PQC を用意する必要がある．

3　Grover のアルゴリズム

Grover のアルゴリズムはオラクル $f : \{0,1\}^n \to \{0,1\}$ が与えられたときに，$f(x_0) = 1$ を満たす x_0 を高確率で効率的に得ることができるアルゴリズムである [2]．ここでは簡単のため，$f(x_0) = 1$ を満たす x_0 はただ一つだけ存在するものとする．以下にアルゴリズムの概要を示す．

（1） まず，n 量子ビットの状態 $|0^n\rangle$ を準備する．そして，$|0^n\rangle$ に対しての各量子ビットに対して Hadamard 変換を作用させ，以下の量子状態 $|\psi\rangle$ を得る．

$$|\psi\rangle = H^{\otimes n} |0^n\rangle \tag{4}$$

（2） 続いて，オラクルが以下のようなユニタリ変換 U_f で実装されているものとして，この U_f を $|\psi\rangle$ 作用させる．

$$U_f |x\rangle = \begin{cases} -|x\rangle & \text{for } x = x_0 \\ |x\rangle & \text{otherwise} \end{cases} \tag{5}$$

（3） さらに，$|\psi\rangle$ に対する反転を行う以下のユニタリ変換 U_ψ を $U_f |\psi\rangle$ 作用させる．

$$U_\psi \equiv I - 2 |\psi\rangle \langle\psi| \tag{6}$$

ただし，I は単位行列である．なお，U_ψ はユニタリ変換

STUDENT

Sname	Address	Lname
Akari Hongo	…, Tokyo	Obara Lab
Yayoi Asano	…, Tokyo	Asuke Lab
Hajime Komaba	…, Tokyo	Obara Lab
Sora Kashiwa	…, Chiba	Obara Lab
Rin Tanashi	…, Tokyo	Asuke Lab

LAB

Lname	Professor
Obara Lab	Sakura Obara
Asuke Lab	Gohei Asuke

STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB

Sname	Address	Lname	Lname	Professor
Akari Hongo	…, Tokyo	Obara Lab	Obara Lab	Sakura Obara
Yayoi Asano	…, Tokyo	Asuke Lab	Asuke Lab	Gohei Asuke
Hajime Komaba	…, Tokyo	Obara Lab	Obara Lab	Sakura Obara
Sora Kashiwa	…, Chiba	Obara Lab	Obara Lab	Sakura Obara
Rin Tanashi	…, Tokyo	Asuke Lab	Asuke Lab	Gohei Asuke

図 1　リレーションの例．リレーション STUDENT の属性 Lab はリレーション LAB を参照する foreign key である．リレーション LAB の属性 Name はリレーション LAB の primary key である．何らかのアプリケーションで問い合わせ STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB の結果を量子的な重ね合わせ状態にエンコードする必要が生じた場合，ナイーブな実装では，STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB に対して 1 つの QRAM あるいは PQC を用意する必要がる．量子版の関係データベースシステムでは，リレーション STUDENT と LAB を保存した QRAM を用いて STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB の結果をを高速に得ることを目指す．

$U_{\mathrm{r}} \equiv |\psi\rangle\langle 0^n|$ を用いると，以下のように表すことができる．

$$U_\psi = U_{\mathrm{r}} \left(I - 2\left|0^n\right\rangle\left\langle 0^n\right|\right) U_{\mathrm{r}}^\dagger \tag{7}$$

本節の場合，$U_{\mathrm{r}} = H^{\otimes n}$ である．

(2), (3) を $\sqrt{N}$ 回繰り返したのちに状態を測定すると，$1-1/N$ 以上の確率で x_0 を得ることができる．

Grover のアルゴリズムではオラクルへの質問回数が $O(\sqrt{N})$ であり，古典的な全探索のアルゴリズムに対して 2 乗加速が実現されている．Grover のアルゴリズムを現実世界の問題に適用するには，まず，オラクルをユニタリ変換 U_f の形で量子回路で実装する必要がある．また，上述のアルゴリズムでは，$U_{\mathrm{r}} = H^{\otimes n}$ となっており，$|\psi\rangle$ は，n ビットで表現可能な $N = 2^n$ 通りの値すべてが等しく重ねあわせられた状態となっている．つまり，現実世界のデータベースに，Grover のアルゴリズムを適用するには，所望のデータのみを重ね合わせ状態としてエンコードするユニタリ変換 U_{r} が必要である．

4　提 案 手 法

量子デバイスにおいても古典コンピュータにおける関係データベースシステムと同様に，STUDENT と LAB の 2 つのリレーションのみを保存しておく．そのうえで，問い合わせ STUDENT $\bowtie_{\text{STUDENT.Lname=LAB.Lname}}$ LAB に対して，効率的にその結果を返すことを目指す．

QRAM は 2 節で述べたように，非常に多くのゲート操作が必要になるため，近い将来での実現は難しいと考えられる．一方，AAE を用いたエンコーディングは QRAM と比べて near-term な技術であると考えられる．ただし，あくまで近似的なエンコーディングしか行うことができず，エンコーディング結果に必ずエラーが生じるため，応用先に制約が生じる．本節では，QRAM 用いたときと，AAE を用いたときの量子版関係データベースシステムの構成法に関して，主に結合演算に焦点を当てて検討を行う．

4.1　QRAM を用いた関係データベースの構成

本項では図 2 に示すリレーションの QRAM 上での配置を例に，QRAM 上での関係データベースの構成法を議論する．図 2 では，4 つの QRAM を使用し，それぞれの QRAM に STUDENT と LAB のインデックスファイルとデータファイルが保持されている．各インデックスファイルにはデータファイルの QRAM 上のアドレスを指し示すポインタを保存する．なおインデックスファイルのレコードは，QRAM_1，QRAM_3 上でアドレスの 0 から隙間なく敷き詰められているものとする．QRAM_2，QRAM_4 にはデータファイルが保存されている．データファイルのレコードは，インデックスファイルのレコードのように QRAM 上でアドレス 0 から順に敷き詰めていく必要はない．ただし QRAM_2 では，STUDENT のレコードに加

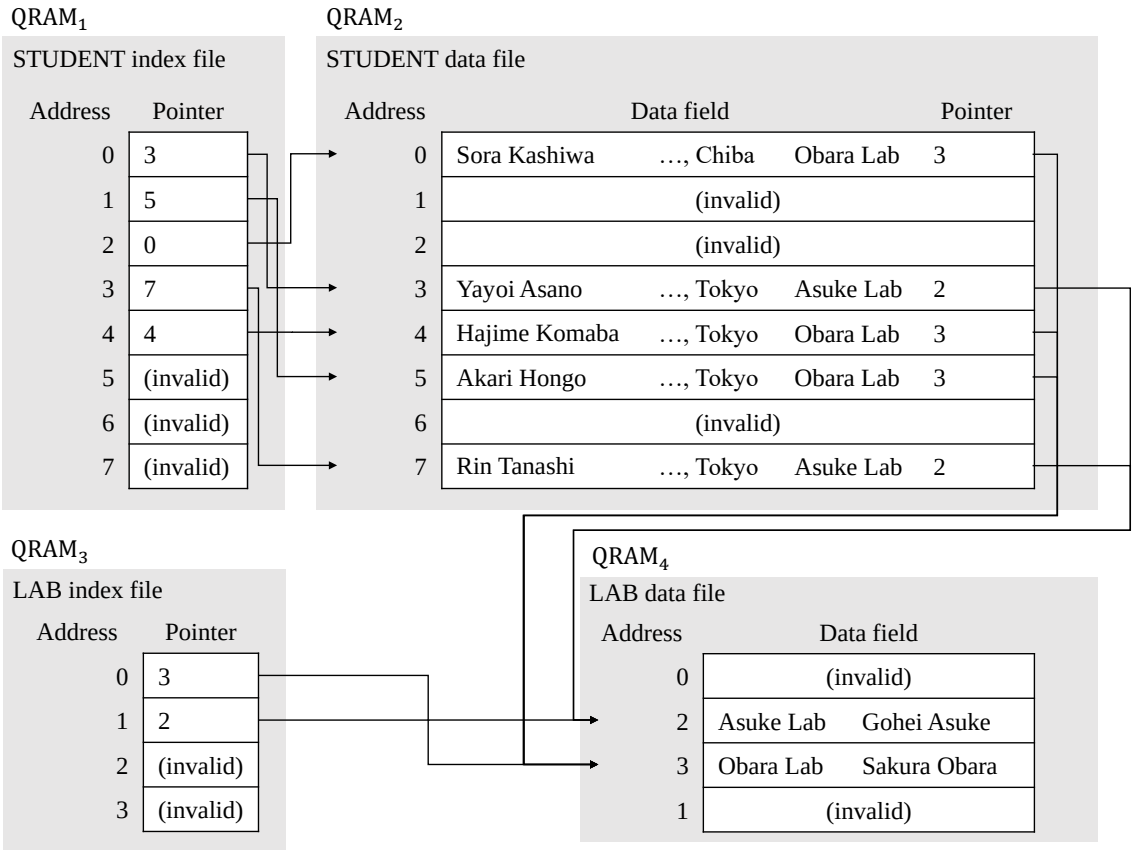


図 2　リレーション STUDENT と LAB の QRAM 上での配置の例．4 つの QRAM それぞれに，STUDENT と LAB のインデックスファイルとデータファイルが保持されている．各インデックスファイルにはデータファイルのレコードの QRAM 上のアドレスを指し示すポインタを保存する．なおインデックスファイルのレコードは，QRAM_1，QRAM_3 上でアドレスの 0 から隙間なく敷き詰める．QRAM_2，QRAM_4 にはデータファイルが保存されており，QRAM_2 では，STUDENT のタプルに加えて LAB のレコードを参照するためのポインタも保存する．データファイルのレコードはインデックスファイルのように QRAM 上でアドレス 0 から順に敷き詰めていく必要はない．なお，本稿では STUDENT と LAB を別々の QRAM 上に保存するものとしたが，同一の QRAM 上に保存することも可能であり，どちらがより効率的かは議論の余地がある．

えて LAB のレコードを参照するためのポインタも保存する．

a） データベースの更新

リレーション STUDENT(LAB) に対してタプルの挿入が生じた場合は，QRAM_2(QRAM_4) の空いているアドレスに新たなレコードを書き込む．またそれと同時に，QRAM_1(QRAM_3) の末尾に新たなレコードを書き込んだアドレスの書き込む．一方，STUDENT(LAB) のタプルの削除が生じた場合は，QRAM_1(QRAM_3) の対応するポインタを削除し，代わりに末尾のアドレス上に保存するポインタをその位置に書き込む．ただし，LAB のタプルを削除しそれを伝搬させる場合は，削除される LAB のタプルのレコードを参照する STUDENT のタプルのレコードを指し示すインデックスファイルのポインタも削除する．磁気ディスク等の古典記憶装置上でデータファイルを保存する場合，レコードの削除に際して効率的なデータの読み出しなどのためにデータファイルを物理的に再配置する必要が生じるケースがある．しかしながら，QRAM はあらゆるアドレスのデータを同時並列的に読み出す機能が備わっていると仮定されるため，我々はデータファイルのレコードを QRAM 上でアドレス 0 から順に敷き詰めるように再配置する必要がないと考えている．このようにレコードの QRAM 上のアドレスが変更されることがないため，QRAM_2 で LAB のレコードのアドレス (ポインタ) を保持していたとしても，それを高頻度で書き換えるといった不利益は生じない．

b） データベースからのデータの取得

まず，STUDENT の全タプルを計算基底符号化による重ね合わせとして量子状態にエンコードする手順を説明する．最初に，インデックスファイルにアクセスするため，QRAM_1 で有効なアドレスの重ね合わせ状態準備する．

$$|0\rangle|0\rangle|0\rangle \to \sum_{i=0}^{4} |i\rangle|0\rangle|0\rangle \tag{8}$$

この操作のナイーブな実装の一例として，以下の実装が考えられる．

（1） QRAM_1 に有効・無効を表す 1 量子ビットのフラグを用意する．

（2） QRAM_1 でアドレス 0 からアドレス α までが有効な

場合は，0 から $2^\beta - 1 \leqq \alpha$ を満たす最大の $\beta \in \mathbb{Z}$ について，0 から $2^\beta - 1 \leqq \alpha$ までの重ね合わせ状態 $\sum_{i=0}^{2^\beta-1} |i\rangle$ をアダマールゲートを用いて準備する．

（3） QRAM_1 に (2) で準備したアドレスを渡し，各アドレスに格納されたポインタと有効・無効フラグを得る．

（4） 有効・無効フラグを測定する．無効が測定された場合は，(2)–(3) を繰り返す．繰り返し回数の期待値は，最悪のケースでも 2 回未満である．

しかしながら，上述の手順はユニタリでない欠点があり，より優れたアルゴリズムは今後の研究課題である．次に，QRAM_1 を用いて，STUDENT データファイルへのポインタをエンコードする．

$$\sum_{i=0}^{4} |i\rangle |0\rangle |0\rangle \xrightarrow{\mathrm{QRAM}_1} \sum_{i=0}^{4} |i\rangle \left|P_i^{\mathrm{STUDENT}}\right\rangle |0\rangle \tag{9}$$

ただし，$P_i^{\mathrm{STUDENT}} \in \{0, 3, 4, 5, 7\}$ は STUDENT のインデックスファイルに含まれるポインタを表す．最後に QRAM_1 を用いて，STUDENT データファイルを取得する．

$$\sum_{i=0}^{4} |i\rangle \left|P_i^{\mathrm{STUDENT}}\right\rangle |0\rangle \xrightarrow{\mathrm{QRAM}_2} \sum_{i=0}^{4} |i\rangle \left|P_i^{\mathrm{STUDENT}}\right\rangle \left|t_i^{\mathrm{STUDENT}}\right\rangle \tag{10}$$

ただし，t_i^{STUDENT} は STUDENT リレーションのうちの 1 つのタプルを表す．なお，t_i^{STUDENT} には図 2 に示すように，LAB のタプルを参照するためのポインタも含まれている．式 (8) から式 (10) までの操作により，STUDENT の読み出しは完了する．LAB のリレーションの取得も，STUDENT でのケースとまったく同様の手順により行うことができる．さらに，STUDENT $\bowtie_{\mathrm{STUDENT.Lname=LAB.Lname}}$ LAB の結果は，STUDENT の取得の手順に 1 つステップを加えるだけで可能となる．式 (10) の状態を準備したうえで，$|t_i^{\mathrm{STUDENT}}\rangle$ の LAB のタプルを参照するためのポインタ部分を QRAM_4 に渡す．それにより，STUDENT と LAB の結合演算が自然と実行され，STUDENT $\bowtie_{\mathrm{STUDENT.Lname=LAB.Lname}}$ LAB の結果が得られる．

$$\sum_{i=0}^{4} |i\rangle \left|P_i^{\mathrm{STUDENT}}\right\rangle \left|t_i^{\mathrm{STUDENT}}\right\rangle |0\rangle \xrightarrow{\mathrm{QRAM}_2} \sum_{i=0}^{4} |i\rangle \left|P_i^{\mathrm{STUDENT}}\right\rangle \left|t_i^{\mathrm{STUDENT}}\right\rangle \left|t_i^{\mathrm{LAB}}\right\rangle \tag{11}$$

以上のように，は結合演算は単に QRAM の読み出し操作だけによって完結する．そのため，式 (8) の操作と QRAM の読み出しにかかる時間が量子ビット数 n に対して，いずれも $O(\mathrm{poly}(n))$ であれば，STUDENT $\bowtie_{\mathrm{STUDENT.Lname=LAB.Lname}}$ LAB の取得にかかる時間のオーダーも $O(\mathrm{poly}(n))$ となる．つまり，QRAM 上においても関係データベースを構成することで，結合されたリレーションも高速に取得することが可能となる．

4.2 AAE を用いた関係データベースの構成

本節では，図 1 に示すリレーションの具体的な例は用いず，より一般に議論を行う．まず，リレーション $\mathrm{R}(A_1, A_2, \cdots, A_{m_\mathrm{R}}), \mathrm{S}(B_1, B_2, \cdots, B_{m_\mathrm{S}})$ に対応した量子状態として $|\mathrm{R}\rangle$，$|\mathrm{S}\rangle$ を以下の式 (12)–(13) のように定義する．

$$|\mathrm{R}\rangle = \sum_{j=1}^{K_\mathrm{R}} \left|t_j^\mathrm{R}\right\rangle = \frac{1}{\sqrt{K_\mathrm{R}}} \sum_{j=1}^{K_\mathrm{R}} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \tag{12}$$

$$|\mathrm{S}\rangle = \sum_{j=1}^{K_\mathrm{S}} \left|t_j^\mathrm{S}\right\rangle = \frac{1}{\sqrt{K_\mathrm{S}}} \sum_{j=1}^{K_\mathrm{S}} \left|w_1^j, w_2^j, \cdots, w_{m_\mathrm{S}}^j\right\rangle \tag{13}$$

ここで，A_i, B_i はそれぞれ R,S の属性，$t_j^\mathrm{R} =< v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j >, t_j^\mathrm{S} =< w_1^j, w_2^j, \cdots, w_{m_\mathrm{S}}^j >$ は R,S のタプル，v_i^j, w_i^j はタプル $t_j^\mathrm{R}, t_j^\mathrm{S}$ の A_i, B_i に対応する値である．また，$K_\mathrm{R}, K_\mathrm{S}$ はそれぞれ R,S のタプルの数であり，$m_\mathrm{R}, m_\mathrm{S}$ は R,S の属性の数である．これらの $|\mathrm{R}\rangle$ と $|\mathrm{S}\rangle$ は以下の関係式を満たす AAE のユニタリ変換 $U_{r_\mathrm{R}}, U_{r_\mathrm{S}}$ で近似的に準備されるものとする．

$$U_{r_\mathrm{R}} |0^{n_\mathrm{R}}\rangle \sim |\mathrm{R}\rangle \tag{14}$$

$$U_{r_\mathrm{S}} |0^{n_\mathrm{S}}\rangle \sim |\mathrm{S}\rangle \tag{15}$$

ただし，n_R と n_S は R と S のタプルのビット長である．

a） データベースの更新

リレーション R,S に対してタプルの挿入・更新・削除などが生じた際には，PQC の再学習が必要となる．そのため，保存されたデータがわずかに更新された場合などに，効率的に $\boldsymbol{\theta}$ を学習するための手法を研究することが今後の課題になると考えられる．

b） データベースからのデータの取得

データベースから R と S を取得するには，基底状態に対して $U_{r_\mathrm{R}}, U_{r_\mathrm{S}}$ を作用させるだけでよい．一方，R $\bowtie$ S を取得するにはより複雑な操作が必要となる．まず，R と S の直積 $\mathrm{R} \times \mathrm{S}$ を量子状態で表す．これは単純に $|\mathrm{R}\rangle$ と $|\mathrm{S}\rangle$ のテンソル積 $|\mathrm{R} \times \mathrm{S}\rangle = |\mathrm{R}\rangle \otimes |\mathrm{S}\rangle$ で実現される．そして，結合条件を表すユニタリ変換 U_f を準備し，結合条件を満たした $\mathrm{R} \times \mathrm{S}$ のタプルに対応する量子状態の振幅を Grover のアルゴリズムにより大きくすることで，結合演算に相当する操作が実現される．なお，式 (7) 中の U_r は本節の例では $U_{r_\mathrm{R}} \otimes U_{r_\mathrm{S}}$ となり，$U_\psi = U_{r_\mathrm{R}} \otimes U_{r_\mathrm{S}} \left(I - 2 |0^{n_R+n_S}\rangle \langle 0^{n_R+n_S}|\right) U_{r_\mathrm{R}}^\dagger \otimes U_{r_\mathrm{S}}^\dagger$ となる．また，オラクルへの質問回数は $O(\sqrt{K_\mathrm{R} K_\mathrm{S} / \min(K_\mathrm{R}, K_\mathrm{S})})$ となる．したがって，全体としてのゲート数は，U_ψ と U_f で必要となるゲート数 G_ψ, G_f を用いると，$O\left((G_\psi + G_f)\sqrt{K_\mathrm{R} K_\mathrm{S} / \min(K_\mathrm{R}, K_\mathrm{S})}\right)$ となる．図 3 に，Grover のアルゴリズムを用いて結合演算に相当する操作を行う際の量子回路を示す．ただし，図 3 では $m_\mathrm{R} = m_\mathrm{S} = 2$ とし，A_1, A_2, B_1, B_2 の値は，いずれも 4 ビットであるとした．

4.3 汎用的な結合演算

本項では，QRAM と AAE，いずれを用いた関係データベースでも使用可能な汎用的な結合演算を提案する．前節と同様に，

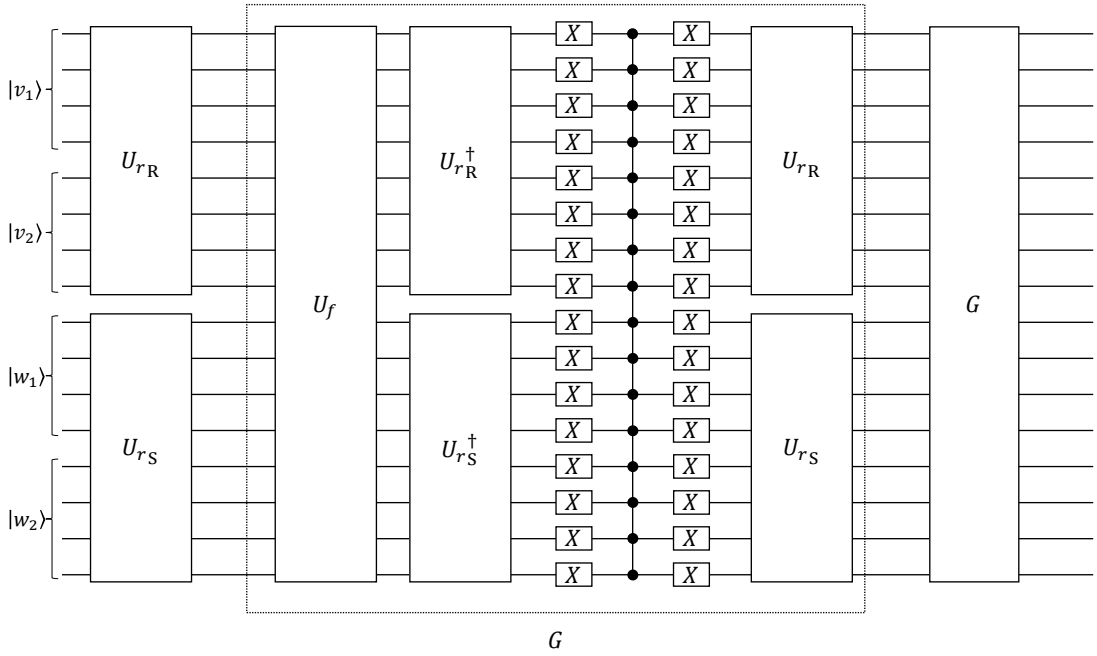


図 3　Grover のアルゴリズムを用いた結合演算を行うための量子回路．リレーション R, S は AAE のユニタリ $U_{r_\mathrm{R}}, U_{r_\mathrm{S}}$ で近似的にエンコードされる．図中の G は $G \equiv \left(U_{r_\mathrm{R}} \otimes U_{r_\mathrm{S}} \left(I - 2\left|0^{n_\mathrm{R}+n_\mathrm{S}}\right\rangle\left\langle 0^{n_\mathrm{R}+n_\mathrm{S}}\right|\right) U_{r_\mathrm{R}}^\dagger \otimes U_{r_\mathrm{S}}^\dagger\right) U_f$ として定義したユニタリ演算子であり，Grover のアルゴリズムの各繰り返しで 1 回ずつ作用させる．

2 つのリレーション $\mathrm{R}(A_1, A_2, \cdots, A_{m_\mathrm{R}}), \mathrm{S}(B_1, B_2, \cdots, B_{m_\mathrm{S}})$ を仮定する．ただし，本節では R の属性 A_{m_R} は，S の属性 B_{m_S} を参照する foreign key であるとし，$\mathrm{R} \bowtie_{A_{m_\mathrm{R}}=B_{m_\mathrm{S}}} \mathrm{S}$ に対応する量子状態を得ることを考える．

ここから，帰納法的に結合演算の手法を示す．なお，本項では状態の規格化定数は無視して表記する．まず，タプル t_j^R を保持する量子ビット $|v_1, v_2, \cdots, v_{m_\mathrm{R}}\rangle$，$t_j^\mathrm{S}$ を保持する量子ビット $|w_1, w_2, \cdots, w_{m_\mathrm{R}}\rangle$，そして，2 量子ビットの補助量子ビット (ancilla) $|a_0, a_1\rangle$ からなる量子システム $|v_1, v_2, \cdots, v_{m_\mathrm{R}}\rangle |w_1, w_2, \cdots, w_{m_\mathrm{R}}\rangle |a_0, a_1\rangle$ を考える．ここで，a_0 は結合演算の途中であることを表すフラグ，a_1 は結合演算のが完了していることを表すフラグとして用いる．このシステムにおいて，以下のように t_0^S から t_{h-1}^S までが t_j^R と結合されているものとする．

$$\sum_{j \notin \pi_1(\mathcal{J})} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle |0, 0, \cdots, 0\rangle |00\rangle + \sum_{(j,k)\in\mathcal{J}} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^k, w_2^k, \cdots, w_{m_\mathrm{R}}^k\right\rangle |01\rangle \quad (16)$$

ただし，

$$\mathcal{J} = \left\{(j,k) \in \mathbb{Z}^2 \mid 0 < k < h \wedge j > 0 \wedge v_{m_\mathrm{R}}^j = w_1^k\right\}$$

であり，$\pi_1(\cdot)$ は (j,k) の第一成分への射影である．また，集合 $\mathcal{J}'$ を以下のように定義する．

$$\mathcal{J}' = \left\{(j,k) \in \mathbb{Z}^2 \mid 0 < k < h+1 \wedge j > 0 \wedge v_{m_\mathrm{R}}^j = w_1^k\right\}$$

式 (16) に示す状態の $|w_1, w_2, \cdots, w_{m_\mathrm{R}}\rangle$ に対して，S の h 番目のタプル t_h^S を書き込んでいく．最初に，$v_{m_\mathrm{R}}^j = w_1^h$ であるときに a_0 を反転させ，以下の状態を得る．

$$\sum_{j\in\pi_1(\mathcal{J}')\setminus\pi_1(\mathcal{J})} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle |0, 0, \cdots, 0\rangle |10\rangle + \sum_{j\notin\pi_1(\mathcal{J}')} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle |0, 0, \cdots, 0\rangle |00\rangle + \sum_{(j,k)\in\mathcal{J}} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^k, w_2^k, \cdots, w_{m_\mathrm{R}}^k\right\rangle |01\rangle \quad (17)$$

次に，$a_0 = 1$ の場合，$|w_1, w_2, \cdots, w_{m_\mathrm{R}}\rangle$ に対して S の h 番目のタプル t_h^S を書き込み，以下の状態を得る．

$$\sum_{j\in\pi_1(\mathcal{J}')\setminus\pi_1(\mathcal{J})} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^h, w_2^h, \cdots, w_{m_\mathrm{R}}^h\right\rangle |10\rangle + \sum_{j\notin\pi_1(\mathcal{J}')} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle |0, 0, \cdots, 0\rangle |00\rangle + \sum_{(j,k)\in\mathcal{J}} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^k, w_2^k, \cdots, w_{m_\mathrm{R}}^k\right\rangle |01\rangle \quad (18)$$

さらに，$a_0 = 1$ の場合 a_1 を反転し，以下の状態を得る..

$$\sum_{j\in\pi_1(\mathcal{J}')\setminus\pi_1(\mathcal{J})} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^h, w_2^h, \cdots, w_{m_\mathrm{R}}^h\right\rangle |11\rangle + \sum_{j\notin\pi_1(\mathcal{J}')} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle |0, 0, \cdots, 0\rangle |00\rangle + \sum_{(j,k)\in\mathcal{J}} \left|v_1^j, v_2^j, \cdots, v_{m_\mathrm{R}}^j\right\rangle \left|w_1^k, w_2^k, \cdots, w_{m_\mathrm{R}}^k\right\rangle |01\rangle \quad (19)$$

最後に，$v_{m_\mathrm{R}}^j = w_1^h$ であるときに a_0 を反転させ，以下の状態を得る．

$$
\begin{aligned}
&\sum_{j\in\pi_1(\mathcal{J}')\setminus\pi_1(\mathcal{J})}\left|v_1^j,v_2^j,\cdots,v_{m_\mathrm{R}}^j\right\rangle\left|w_1^h,w_2^h,\cdots,w_{m_\mathrm{R}}^h\right\rangle|01\rangle\\
&+\sum_{j\notin\pi_1(\mathcal{J}')}\left|v_1^j,v_2^j,\cdots,v_{m_\mathrm{R}}^j\right\rangle|0,0,\cdots,0\rangle\,|00\rangle\\
&+\sum_{(j,k)\in\mathcal{J}}\left|v_1^j,v_2^j,\cdots,v_{m_\mathrm{R}}^j\right\rangle\left|w_1^k,w_2^k,\cdots,w_{m_\mathrm{R}}^k\right\rangle|01\rangle\\
&=\sum_{j\notin\pi_1(\mathcal{J}')}\left|v_1^j,v_2^j,\cdots,v_{m_\mathrm{R}}^j\right\rangle|0,0,\cdots,0\rangle\,|00\rangle\\
&+\sum_{(j,k)\in\mathcal{J}'}\left|v_1^j,v_2^j,\cdots,v_{m_\mathrm{R}}^j\right\rangle\left|w_1^k,w_2^k,\cdots,w_{m_\mathrm{R}}^k\right\rangle|01\rangle
\end{aligned}
\tag{20}
$$

以上のように，t_0^S から t_{h-1}^S までが t_j^R と結合されていれば，t_h^S を t_j^R と結合させることが可能である．また，t_k^S が 1 つも結合されていない状態に対して，t_0^S を結合する場合も，全く同じ手順において結合演算が可能である．したがって，上で述べた手順を繰り返すことで R と S の結合演算が実行可能である．この手法による時間計算量は，多入力 CNOT ゲートを ancilla を用いて分解することで，システム全体の量子ビット数 n，S のタプル数 K_S を用いると，$O\left(K_\mathrm{S}\mathrm{poly}\left(n\right)\right)$ となる．したがって，K_S が非常に小さいケースでは，高速に結合演算が可能である．また，S をエンコードする QRAM などを準備する必要がないといった利点がある．

5 考　　察

QRAM 上での関係データベースにおける結合演算は，4.1 節で述べたようにインデクシングによって QRAM 上で自然と実行でき，実行時間はリレーションのタプルのビット数を n とすると，$O(\mathrm{poly}(n))$ となる（ただし，QRAM の読出し時間が $O(\mathrm{poly}(n))$ であると仮定した）．古典の関係データベースシステムにおける結合では，少なくともリレーションのタプル数に比例する項がコストとして発生するため，本手法は極めて高速な結合演算である．しかしながら，この実装は（少なくとも現時点では）foreign key と primary key が等しいときに結合する EQUIJOIN でのみ適用できると考えられる．より一般の結合，さらには選択を行うには，測定を用いる必要になる可能性がある．このとき，Grover のアルゴリズムを用いることも考えられるが，QRAM のエラーの観点からも Grover のアルゴリズムの使用は近い将来では難しいと考えられる [24]．

一方，得られるデータにエラーが生じることが許容できる場合，4.2 節で述べた AAE を用いた関係データベースの構成も候補に挙がる．AAE では QRAM のようにアドレスに対応したデータを返す仕組みがないため，QRAM のような $O(\mathrm{poly}(n)$ 程度の時間での結合演算は不可能である．ただし，QRAM と比べてケート数が少なく，また誤差も許容されるケースを想定しているため，結合にグローバーのアルゴリズムが適用可能であると考えられる [6]．U_f は multi-controlled X ゲートで定数個の ancilla を用いるなどの工夫により，$O(\mathrm{poly}(n_\mathrm{R}+n_\mathrm{S}))$ のゲート数で実装できる．また，AAE に必要なゲート数は $O(\mathrm{poly}(n_\mathrm{R})+\mathrm{poly}(n_\mathrm{S}))$ で十分であると考えられている [20]．したがって，結合演算で必要なゲート数は $O((\mathrm{poly}(n_\mathcal{R}+n_\mathcal{S})+\mathrm{poly}(n_\mathcal{R})+\mathrm{poly}(n_\mathcal{S}))\sqrt{K_\mathrm{R}K_\mathrm{S}/\min(K_\mathrm{R},K_\mathrm{S})})$ となることが見込まれる．また選択に関しても，U_f に選択の条件を埋め込むことで Grover のアルゴリズムを適用することが可能となる．なお，Grover のアルゴリズムでは通常アルゴリズムの最後に測定を行うが，AAE を用いた実装ではもとから量子状態にエラーが含まれており，測定を行ったとしてもデータベースに含まれないデータの振幅が 0 にはならない．そのため，我々はアプリケーションによって，Grover のアルゴリズムでは通常アルゴリズムの最後に測定を行うかどうか，考慮する余地があると考えている．

選択と結合にかかわるクエリ最適化に関しては，今後の研究課題である．古典の関係データベースでは結合よりも選択を先に行うことで，I/O コストや CPU コストの削減につながるが，今回提案した量子版の関係データベースにおいても同様な最適化が可能であるかは，注意して考えなければならない．特に，Grover のアルゴリズムを用いた結合演算では，ユニタリ U_r を繰り返し作用させなければならず，選択を Grover のアルゴリズムを用いて先に行ったからといって，ゲート数や回路の深さの削減につながるとは限らない．また，選択に測定を用いると，その後 Grover のアルゴリズムを用いた結合演算を行うことができない点にも注意が必要である．結合演算の種類の選択も最適化の余地があり，例えば 4.3 節で提案した結合演算が他の手法よりも有利になるケースもあると考えられる．

そして，今回提案した量子版の関係データベースシステムは，量子デバイスだけで動作するものではなく，背後には古典コンピュータや古典ストレージの存在を前提としている．データベースのリレーションも，QRAM や PQC 上だけに保持されているのではなく，古典のストレージにも保存されていなければならない．そのため，我々は古典システムと量子システムの両方を合わせたシステム全体の最適化の余地も多く残されていると考えている．さらに，本研究ではデータベースとして関係データベースを用いたが，量子コンピューティングで実データを扱う際に最適なモデルが関係モデルであるかは不明である．関係モデルよりも量子計算に適したデータモデルも今後検討する必要があると考えられる．

6 おわりに

本稿では，量子回路上での関係データベースの実装方法について，特に結合に焦点を当てて検討を行った．我々は，主に 2 つの実装方法を提案した．1 つ目は，QRAM 上での関係データベースの構成である．我々は，参照先のリレーションのタプルの QRAM 上でのアドレスを foreign key と一緒に保存しておくことで，高速な結合演算が実行可能であることを示した．2 つ目は，AAE を用いた関係データベースの構成である．AAE は QRAM と比べてゲート数が少なく済み，より near-term な技術であると考えられる．我々は，AAE に適した結合演算手法の候補として，Grover のアルゴリズムを用いた結合演算を提案した．

また，QRAM，AAE いずれを用いた関係データベースにおいても実行可能な汎用的な結合演算も提案した．この手法は，リレーションのタプルの数が非常に小さいケースでは，高速に結合演算が可能であり，S をエンコードする QRAM などを準備する必要がないといった利点がある．

我々が知る限り，量子コンピューティングの分野に関係データベースの考え方を導入したのは本研究が初めてだと考えられる．そのため，この分野における研究の余地は大きく残されている．今後はクエリ最適化や，古典システムと量子システムを合わせたデータベースシステム全体の最適化に関する研究が必要となる．

文　献

[1] D. Deutsch and R. Jozsa, “Rapid solution of problems by quantum computation,” Proc. R. Soc. London, Ser. A, vol.439, no.1907, pp.553–558, 1992.
[2] L.K. Grover, “A fast quantum mechanical algorithm for database search,” Proc. STOC, p.212–219, 1996.
[3] P.W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” SIAM Review, vol.41, no.2, pp.303–332, 1999.
[4] A.W. Harrow, A. Hassidim, and S. Lloyd, “Quantum algorithm for linear systems of equations,” Phys. Rev. Lett., vol.103, p.150502, 2009.
[5] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, “Quantum circuit learning,” Phys. Rev. A, vol.98, p.032309, 2018.
[6] H. Tezuka, K. Nakaji, T. Satoh, and N. Yamamoto, “Grover search revisited: Application to image pattern matching,” Phys. Rev. A, vol.105, p.032440, 2022.
[7] Y. Sano, M. Norimoto, and N. Ishikawa, “Qubit reduction and quantum speedup for wireless channel assignment problem,” IEEE Trans. Quantum Eng., vol.4, pp.1–12, 2023.
[8] P. Niroula and Y. Nam, “A quantum algorithm for string matching,” npj Quantum Inf., vol.7, no.1, p.37, 2021.
[9] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P.J. Love, A. Aspuru-Guzik, Jeremy LO’brien, “A variational eigenvalue solver on a photonic quantum processor,” Nat. Commun., vol.5, no.1, p.4213, 2014.
[10] Y. Sato, R. Kondo, S. Koide, and S. Kajita, “Quantum topology optimization of ground structures using noisy intermediate-scale quantum devices,”arXiv:2207.09181, 2022.
[11] I. Kerenidis and A. Prakash, “Quantum recommendation systems,”arXiv:1603.08675, 2016.
[12] S. Gao, F. Hayes, S. Croke, C. Messenger, and J. Veitch, “Quantum algorithm for gravitational-wave matched filtering,” Phys. Rev. Res., vol.4, p.023006, 2022.
[13] Y. Nakamura, Y.A. Pashkin, and J. Tsai, “Coherent control of macroscopic quantum states in a single-Cooper-pair box,” Nature, vol.398, no.6730, pp.786–788, 1999.
[14] A. Blais, R.-S. Huang, A. Wallraff, S.M. Girvin, and R.J. Schoelkopf, “Cavity quantum electrodynamics for superconducting electrical circuits: An architecture for quantum computation,” Phys. Rev. A, vol.69, p.062320, 2004.
[15] J. Clarke and F.K. Wilhelm, “Superconducting quantum bits,” Nature, vol.453, no.7198, pp.1031–1042, 2008.
[16] L. DiCarlo, J.M. Chow, J.M. Gambetta, L.S. Bishop, B.R. Johnson, D.I. Schuster, J. Majer, A. Blais, L. Frunzio, S.M. Girvin, and R.J. Schoelkopf, “Demonstration of two-qubit algorithms with a superconducting quantum processor,” Nature, vol.460, no.7252, pp.240–244, 2009.
[17] J. Koch, T.M. Yu, J. Gambetta, A.A. Houck, D.I. Schuster, J. Majer, A. Blais, M.H. Devoret, S.M. Girvin, and R.J. Schoelkopf, “Charge-insensitive qubit design derived from the Cooper pair box,” Phys. Rev. A, vol.76, p.042319, 2007.
[18] M. Steffen, D.P. DiVincenzo, J.M. Chow, T.N. Theis, and M.B. Ketchen, “Quantum computing: An IBM perspective,” IBM J. Res. Dev., vol.55, no.5, pp.13:1–13:11, 2011.
[19] F. Arute, K. Arya, R. Babbush, D. Bacon, J.C. Bardin, R. Barends, R. Biswas, S. Boixo, F.G. Brandao, D.A. Buell, et al., “Quantum supremacy using a programmable superconducting processor,” Nature, vol.574, no.7779, pp.505–510, 2019.
[20] K. Nakaji, S. Uno, Y. Suzuki, R. Raymond, T. Onodera, T. Tanaka, H. Tezuka, N. Mitsuda, and N. Yamamoto, “Approximate amplitude encoding in shallow parameterized quantum circuits and its application to financial market indicators,” Phys. Rev. Res., vol.4, p.023136, 2022.
[21] P. Rebentrost and S. Lloyd, “Quantum computational finance: quantum algorithm for portfolio optimization,”arXiv:1811.03975, 2018.
[22] V. Giovannetti, S. Lloyd, and L. Maccone, “Quantum random access memory,” Phys. Rev. Lett., vol.100, p.160501, 2008.
[23] V. Giovannetti, S. Lloyd, and L. Maccone, “Architectures for a quantum random access memory,” Phys. Rev. A, vol.78, p.052310, 2008.
[24] S. Jaques and A.G. Rattew, “QRAM: A survey and critique,”arXiv:2305.10310, 2023.
[25] C.T. Hann, G. Lee, S.M. Girvin, and L. Jiang, “Resilience of quantum random access memory to generic noise,” PRX Quantum, vol.2, p.020311, 2021.
[26] R. Asaka, K. Sakai, and R. Yahagi, “Two-level quantum walkers on directed graphs. ii. application to quantum random access memory,” Phys. Rev. A, vol.107, p.022416, 2023.
[27] C.T. Hann, C.-L. Zou, Y. Zhang, Y. Chu, R.J. Schoelkopf, S.M. Girvin, and L. Jiang, “Hardware-efficient quantum random access memory with hybrid quantum acoustic systems,” Phys. Rev. Lett., vol.123, p.250501, 2019.
[28] R. Elmasri and S.B. Navathe, Fundamentals of database systems, Addison-Wesley, 2016.
[29] P. Mishra and M.H. Eich, “Join processing in relational databases,” ACM Comput. Surv., vol.24, no.1, p.63–113, 1992.

DEIM Forum 2024 T2-A-6-03

永続メモリ向けロックフリー索引Bz木に関する研究

中山　宗†　杉浦　健人†　石川　佳治†　陸　可鏡†

† 東海国立大学機構名古屋大学大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
Email: {nakayama, lu}@db.is.i.nagoya-u.ac.jp, {sugiura, ishikawa}@i.nagoya-u.ac.jp

あらまし　近年，主記憶に不揮発性メモリを用いた永続メモリ環境に注目が集まっている．永続メモリ環境では，障害時においても主記憶のデータが消失せず，残ったデータを用いた復帰が可能となる．永続メモリの扱いは難しく，利用を容易にするためのミドルウェアとして，永続メモリ向け索引構造の研究開発が行われている．永続メモリ向け索引の 1 つである Bz 木は，任意のデータ型を扱えるという大きな長所を持つが，既存研究にてその性能の低さが指摘されている．しかし，Bz 木の使用していたライブラリの性能が低いこと，アルゴリズムに誤りが存在していたことから，既存研究においては Bz 木の性能が正しく測定されていない可能性がある．そこで本研究では，修正と改善を加えた Bz 木の性能を改めて評価することで，既存研究において Bz 木の性能が過小評価されていたことを示す．

キーワード　データ構造・索引，並列・分散処理，永続メモリ．

1 はじめに

近年，主記憶に不揮発性メモリを用いた永続メモリ環境に注目が集まっている．大規模な容量を備えつつ，従来の揮発性メモリに迫る性能を実現した，Intel Optane DC Persistent Memory Modules（DCPMM）が 2019 年ごろに登場した．これにより，DCPMM を用いた永続メモリ環境が実現可能となった．

永続メモリの長所の 1 つは，主記憶に残ったデータを用いた復帰が可能なことであるが，このような復帰を行うためには常に復帰可能な状態を維持する必要がある．永続メモリ環境では障害発生時においても，不揮発性によりデータが消失せず，消失せずに残ったデータを用いた復帰が可能である．しかし，残ったデータを用いた復帰を行うためには，主記憶に残るデータを常に復帰可能な状態に維持する様々な工夫が必要となる．例えば障害によってロックを取ったスレッドの情報が失われると，どのスレッドがロックを解除すべきか不明となり，ロックをかけられたアイテムが残り続ける．このように，復帰可能な状態を維持するために注意すべき点が多く存在するため，永続メモリ上でのプログラムは非常に困難なものとなる．そこで，永続メモリの利用を容易にするためのミドルウェアとして，永続メモリ向け索引構造の研究開発が行われている．

永続メモリ向け索引の 1 つである Bz 木 [1] は，任意のデータ型を扱えるという大きな長所を持つが，既存研究では性能が低いと報告されている．FPTree [2] や LB$^+$Tree [3] は，他の索引と比較しより良い性能を示すことが報告されている．しかし，扱うデータに制約を課すことで性能を上げているため，扱えるキーの種類が固定長 8 バイトのみに限られる．また，PACTree [4] や ROART [5] ，DPTree [6] などのトライ木ベースの索引は，バイナリデータに特化している．一方，B$^+$ 木を拡張した Bz 木は任意のデータ型を扱うことができ，文字列型やユーザ定義型を扱える汎用性を持つ．しかし，Lersch らによる実験 [7] において，Bz 木の性能が他の永続メモリ向け索引 [2,8,9] と比較し劣ることが報告されている．

だが，既存研究においては，Bz 木本来の性能が検証されていない可能性がある．本研究グループでは，Bz 木が使用するライブラリの 1 つである Persistent Multi-Word Compare-and-Swap（PMwCAS）命令 [10] を改善した．結果として，最大約 12 倍の性能改善が見られた．また，提案手法における Bz 木のアルゴリズムには誤りが存在し，不整合が生じ得る．ゆえに，Lersch らによる実験では，Bz 木の性能が正しく評価されていないと考えられる．

そこで本研究では，Bz 木の元論文の手法に修正および改善を加え，改めて Bz 木の性能を評価する．そして，Bz 木の性能が既存研究において過小評価されていたことを示す．3 章にて，Bz 木の元論文での手法の詳細を述べる．4 章では，Bz 木の元論文での手法の問題点，改善の余地を述べ，改善方法を提案する．最後に 6 章において，改善した Bz 木の性能を検証する．

2 関連研究

本章では，関連研究として索引構造における同時実行制御の発展と，永続メモリ環境，Bz 木が使用する Persistent Multi-Word Compare-and-Swap（PMwCAS）命令について詳細を述べる．

2.1 索引構造における同時実行制御の発展

データベースにおける索引構造にも，マルチスレッド環境にてより良い性能を出すための工夫が凝らされてきた．B$^{\text{link}}$ 木 [11] は，ロックの専有期間を短くするため，葉ノード間だけでなく中間ノード間にも兄弟リンクを付与した．OLFIT [12] は，楽観的な排他処理を用いることで，読み取りのためのロックの必要性を排除した．Bw 木 [13] は，論理ページと物理ページを分離し，木の変更を差分レコードと呼ばれる追記型のログを用いて表すことでロックフリー化を実現した．本研究で扱う Bz 木もまた差分レコードによる更新を採用しており，PMwCAS 命令を用いることで永続メモリ環境におけるロックフリー化を簡

略化している．

2.2 永続メモリ環境

永続メモリ環境 [14] とは，主記憶に不揮発性メモリを用いた環境である．高性能な不揮発性メモリである，Intel Optane DC Persistent Memory Modules（DCPMM）の登場により，永続メモリ環境は実現可能となった．DCPMM の性能は複数の先行研究 [7, 15–18] によって報告されている．Lersch らによる実験 [7] では，従来の揮発性メモリと比較し，読み込み帯域幅が約 3–8 倍，書き込み帯域幅が約 11–14 倍低いことを観測した．

また，Compute Express Link（CXL） [19] を用いた永続メモリ環境の実現も注目されている．CXL はオープンインターコネクト規格の 1 つであり，様々なデバイスへの高速な CPU 接続のために提案された．CXL はコヒーレントかつ帯域幅の広い接続を実現するため，CXL で接続したメモリにバッテリーで不揮発性を付与する [20] ことで，永続メモリとしても利用できる可能性が示されている．Fridman らは DCPMM を模倣して CXL メモリにアクセスした際の性能を検証し，DCPMM と比較して CXL メモリが高い帯域幅を持つこと示した [21]．一方で，Persistent Memory Development Kit（PMDK）などを利用した永続メモリ特有のプログラミングは CXL メモリであっても必須であり，データ構造や手続きの改善は依然として求められている．

2.3 Persistent Multi-Word Compare-and-Swap

Persistent Multi-Word Compare-and-Swap（PMwCAS）命令 [10] は，Wang らによって提案された，永続メモリ上の複数ワードをアトミックかつロックフリーに更新する命令である．対象となるメモリのデータを書き換える直前に，書き換え前のメモリの持つデータが想定通りであることを確認することで，ロックフリーなメモリの書き換えを実現する．Wang らの PMwCAS 命令のアルゴリズムは Multi-Word Compare-and-Swap（MwCAS）命令の 1 つである CASN アルゴリズム [22] を拡張したものである．

CASN アルゴリズムおよび Wang らによる PMwCAS 命令のアルゴリズムは，確保したメモリ領域のガベージコレクションを必要とする欠点を持つ．本研究グループでは，既にガベージコレクションを必要としない PMwCAS 命令のアルゴリズム [23] を提案し，性能の改善を達成している．本研究では PMwCAS 命令の改善が，PMwCAS 命令を使用する Bz 木の性能に及ぼす影響についても評価する．

3 Bz 木 概 要

Bz 木 [1] は PMwCAS 命令を用いて B^+ 木を拡張したロックフリーの索引構造である．本章では，Bz 木の元論文での構造と操作の詳細を述べる．Bz 木の構造を述べた後，Bz 木がサポートするノード操作の一部，構造変更操作の詳細を述べる．

3.1 構　　造

Bz 木は，Bz 木の根を指すルートポインタと，複数の中間ノードと葉ノードからなる．全体像を図 1 に示す．ルートポインタが根となる中間ノードへのポインタ，中間ノードが子となる中間ノードまたは葉ノードへのポインタを持つ．これにより，全体として木構造を構築する．

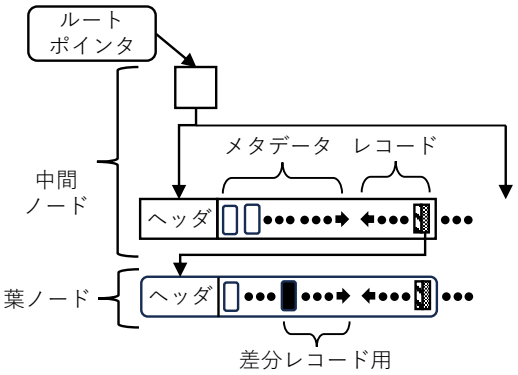


図 1　Bz 木の構造

図 1 に示すように，各ノードはヘッダと，複数のメタデータからなるメタデータ列，複数のレコードからなるレコード列を持つ．ヘッダは，ノードの大きさやレコード数，後述するステータスなど，ノード自身の情報が格納される．ヘッダが持つステータスは，表 1 に示す情報を持つ．メタデータは 1 つのレコードに対して 1 つ存在し，対応するレコードのメモリ上での位置など，対応するレコードの情報を格納する．レコードはキーとペイロードの組からなる．中間ノードのペイロードは葉ノードへのポインタである．

Bz 木は Bw 木 [13] と同様に，木への変更を差分レコードによって表す．差分レコードは対象ノードへの変更（レコードの追加・更新・削除）を表すレコードであり，葉ノードの空き領域に順に格納される．つまり，差分レコードの順序はノードに対して起きた変更の順序を表し，キーの順序とは一致しない．なお，Bw 木とは異なり，Bz 木において中間ノードは差分レコードを持たない．下の階層で構造変更が起きた際は後述する構造変更操作の手続きに従い，追加ないし削除されたものを含む全レコードが常にソートされた状態で格納される．

3.2 操　　作

本節では，Bz 木がサポートする操作の詳細を述べる．まず，Bz 木における操作の PMwCAS 命令による線形化方法を示すため，挿入操作の詳細を述べる．挿入操作は Bz 木内に存在しないキーを持つレコードを，特定のノードに挿入する操作である．その後，本研究の改善により変更を加える更新操作と削除操作，3 つの構造変更操作について説明する．更新操作は指定されたキーを持つレコードのペイロードを更新する操作，削除操作は指定されたキーを持つレコードを削除する操作である．

3.2.1 挿 入 操 作

挿入操作の挙動を図 2 に示す．手順としては，まずレコードの挿入対象となるノードを探索する．そして，挿入するレコードと同じキーを持つレコードがノード内に存在しないことを確認する．その後は 2 つの工程に大別される．

最初の工程では，挿入するレコードのための領域を確保する．

表 1　ステータスが持つ情報

名称	ビット数［bit］	詳細
record count	16	レコードの総数.
block size	22	レコードブロックが占有しているスペースのバイト数.
deleted size	22	レコードの削除により空いた状態となったスペースのバイト数.
frozen	1	ノードが凍結状態であることを示すフラグ.
control region	3	PMwCAS 命令のために使用される領域.

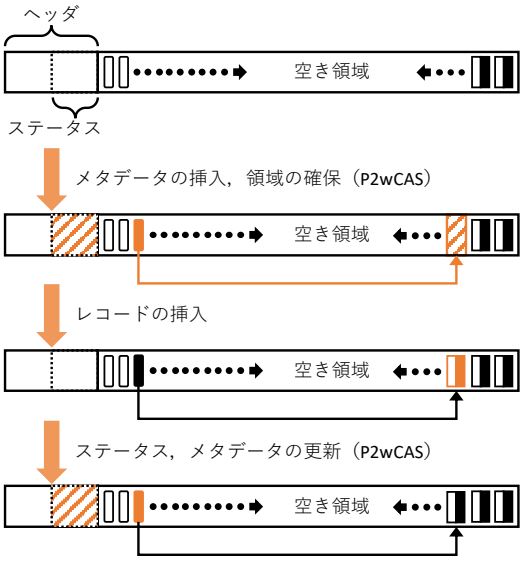


図 2　挿入操作におけるノード内の状態

指定されたキーを持つノードのステータスを読み取り，挿入するレコードに対応するメタデータと，新たなステータスを用意する．新たなステータスとは，変更前のステータスの record count と block size を挿入するレコードに合わせて変更したものである．用意するメタデータには，対応するレコードが書き込み途中である旨の情報が含まれる．そして，用意したメタデータとステータスを PMwCAS 命令を用いてノードに挿入することで，挿入するレコードのための領域を確保する．

そして最後の工程で，実際にレコードを挿入する．既に前の工程で領域を確保しているため，競合する操作を気にすることなくレコードを書き込む．レコードの書き込みが完了したら，ステータスと書き込み状態を示しているメタデータを PMwCAS 命令を用いて更新し，挿入操作を完了する．

3.2.2　更 新 操 作

更新操作は 2 つの場合によって挙動が異なる．1 つは，ペイロードが 8 バイトの固定長であり対象レコードが差分レコードでない場合である．この場合は，キーに手を加えずペイロードのみの変更によりレコードの更新を実現でき，置換更新と呼ぶ．もう 1 つは，ペイロードが可変長であるか，ペイロードが 8 バイトの固定長であるが対象レコードが差分レコードの場合である．この場合の更新操作の挙動は挿入操作と同じであり，追記更新と呼ぶ．

置換更新は，1 度の PMwCAS 命令にて実現する．手順としては，指定されたキーを持つレコードを含むノードを探索した後，ステータスと対象レコードのメタデータとペイロードを PMwCAS 命令を用いて更新する．メタデータを PMwCAS 命令の対象とすることで，対象ノードに対して同時に行われる削除操作を検知する．同様に，ペイロードを PMwCAS 命令の対象とすることで同時に行われる更新操作を，ステータスを対象とすることで同時に走る構造変更操作を検知する．

3.2.3　削 除 操 作

削除操作は，指定されたキーを持つレコードを削除する操作である．実際には，削除対象のレコードをノードから取り除くのでは無く，削除対象レコードを削除済み状態とする操作である．

具体的な手順としては，指定されたキーを持つノードを探索した後，PMwCAS 命令を用いて操作を実現する．PMwCAS 命令の対象は 2 つあり，1 つは削除対象となるレコードに対応するメタデータの，削除された状態への変更である．もう 1 つは，ノードのステータスの deleted size の，削除対象となるレコードとメタデータ分の増加である．PMwCAS 命令を用いることで更新操作と同様に，競合する書き込み操作や構造変更操作が行われていないことを確認しつつ，操作を完了する．

3.3　構造変更操作

Bz 木はノード再編成，ノード分割，ノードマージの 3 つの構造変更操作を持つ．ノード再編成は葉ノードが持つ全てのメタデータをソートする，葉ノード特有の操作である．ノード再編成は，葉ノードが持つ差分レコードのサイズが閾値を超えた場合に行う．ノード分割は保持するメタデータ列とレコード列の合計サイズが閾値を超えたノードを，2 つのノードに分割する操作である．ノード分割を行うことで，ノードの保持するデータがノードサイズを超えることを防ぐ．ノードマージは，2 つのノードを 1 つのノードへと統合する操作である．削除操作により，ノードが保持するデータのサイズが閾値を下回った場合に行う．データサイズが閾値を下回ったノードと，そのノードの左右に存在する兄弟ノードの 1 つを統合することで実現する．ノードマージを行うことで，保持するレコード数が少ないノードを減らし，索引構造全体におけるレコードの探索性能の低下を防ぐ．3.3.1 句にて構造変更操作に必要なメモリ管理について述べた後，3.3.2 句にて構造変更操作のアルゴリズムを紹介する．

3.3.1　構造変更操作におけるメモリ管理

構造変更操作は，索引外のノードを管理するコンポーネント（メモリマネージャと呼ぶ，図 3 参照）を必要とする．構造変更操作によるノードへの変更は，既存ノード内のデータを直接書き換えない．代わりに，変更結果となるノードを索引外に用意

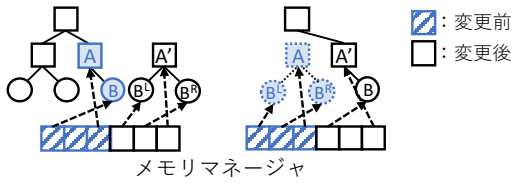


図 3　一時領域を用いた構造変更ノードの記憶

し，既存ノードと入れ替えることで実現する．索引内のノードは根ノードから辿ることが可能だが，索引外のノードは根ノードから辿ることができない．そのため，構造変更操作中の電源障害により，索引外に作られたノードはメモリリークの原因となり得る．ゆえに Bz 木は，永続メモリ上の一時的な使用を目的とする領域を要する．メモリマネージャは，永続メモリ上の一時領域とその領域を使用する索引外ノードを管理することで，メモリリークの発生を防ぐ．

メモリマネージャは，各スレッドに 6 つのノードを格納できる領域を提供する．構造変更操作に用いられるノードは，変更対象となるものと変更結果として作られるものに分けられ，それぞれ 3 種類存在する．1 つ目は構造変更操作の起点となるノード（以降，トリガノード）であり，構造変更前のこのノードを確認することで対象の構造変更操作が特定できる．分割・マージ操作であればトリガノードに加えて，その兄弟ノードおよび親ノードが構造変更の対象となり，一時領域に記憶される．一時領域となる配列ではこれら 6 種類のノードを固定の順番で格納し，リカバリ時は構造変更前のトリガノードを確認することで必要なリカバリ処理を決定する．

3.3.2　構造変更操作のアルゴリズム

3 つの構造変更操作の大まかな流れは共通しており，3 つの工程に大別される．各構造変更操作の流れを図 4 に示す．

1 つ目の工程では，対象ノードを凍結する．書き込み操作の対象となったトリガノードが，各構造変更操作の条件を満たす場合，構造変更操作を開始する．まず，構造変更操作を行うため，トリガノードを凍結（不変，immutable）状態にする．凍結状態とは，書き込み操作や削除操作，他の種類の構造変更操作を受け付けない状態である．トリガノードを凍結することで，トリガノードへのレコードの変更を防ぎつつ構造変更操作を行う．ノードの凍結は，ノードのステータスを対象とした PMwCAS

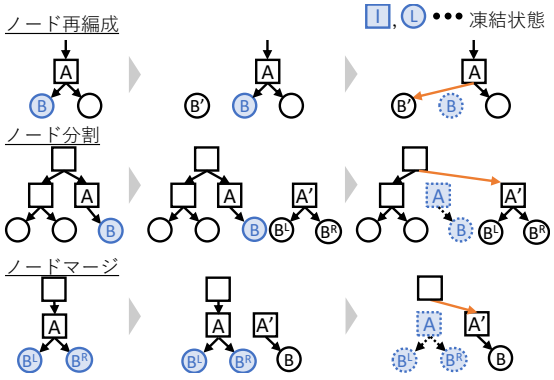


図 4　元論文における構造変更操作

命令によってアトミックに実現する．ステータス内の frozen フラグを立てることで，ノードは凍結状態になる．

2 つ目の工程では，トリガノードに構造変更操作を施した結果となるノードを索引外に用意する．ノード再編成では，トリガノードが持つレコードを全てソートしたノードを索引外に用意する．ノード分割やノードマージのようにノード数が変わる操作では，トリガノードだけでなくその親ノードも用意する．次の工程にて，交換対象となるポインタ数を減らすためである．ノード分割については，トリガノードを等分した 2 つのノードと，その親ノードを作製する．親ノードが持つトリガノードへのポインタは，新たに作製した 2 つのノードへのポインタへと変更する．親ノードに新たなポインタのためのスペースが無ければ，親ノードを対象としてノード分割を再帰的に開始する．ノードマージについても，トリガノードとその兄弟ノードを統合したノードに加え，その親ノードを作製する．ノード分割同様，親ノードが持つトリガノードと兄弟ノードへのポインタを，新たに作製したノードへのポインタへと変更する．親ノードが持つポインタの減少により，親ノードの持つデータサイズが閾値を下回った場合は，親ノードを対象としてノードマージを再帰的に開始する．

最後となる 3 つ目の工程にて，作製したノード（群）を索引に挿入することで，構造変更操作を完了する．作製したノード（群）の挿入は，対象ノード（群）との交換で実現する．ノード分割とノードマージの場合は，交換対象となるトリガノードの親ノードも凍結した後，ノードを交換する．ノードの交換は PMwCAS 命令を用いてアトミックに行う．

3 つ目の工程にて，ノードを交換する際の PMwCAS 命令の対象は 3 つである．1 つは，親ノードが持つトリガノードへのポインタであり，作製したノードへのポインタに変更することでノードを交換する．ノード分割とノードマージの場合は，親ノードへの索引内におけるポインタとなる．トリガノードへのポインタではなく，その親ノードへのポインタを交換対象とすることで，ノード数が変わる操作における PMwCAS 命令の対象の増加を防いでいる．2 つ目は，メモリマネージャが持つ作製したノードへのポインタであり，トリガノードへのポインタと交換する．交換済みの対象ノードへのポインタをメモリマネージャが持つことで，対象ノードをガベージ対象とする前に生じる電源障害にも対処可能である．3 つ目は，交換するポインタを持つノードのステータスである．交換するポインタを持つノードは，ノード編成の場合はトリガノードの親ノード，ノード分割とノードマージの場合は親ノードへのポインタを持つノードである．ステータスから，交換対象となるポインタを持つノードが凍結状態でないことを確認する．

4　修正による一貫性の保証

本研究では，Bz 木に 2 つの修正を加える．1 つは削除操作の修正，もう 1 つは構造変更操作の修正である．順に詳細を述べる．

4.1 削除操作の修正

元論文の手法では，追記更新と削除操作を同時に行った場合に，削除したはずのキー（レコード）がノードに残る不整合が生じ得る．不整合が生じ得る原因は，同一のキーを対象とする追記更新と削除操作が，お互いを検知できないことにある．置換更新と削除操作のお互いを検知する方法を示した後，追記更新と削除操作を同時に行うことで不整合が生じる例を示す．

置換更新と削除操作は，競合するお互いを検知できる．図 5 の左図は，置換更新と削除操作を同時に行った状態である．左下の状態では，削除操作によりメタデータやステータスが準備されてる際に，置換更新によりメタデータが更新されている．この後，削除操作を行うスレッドは，既存のメタデータが想定と異なることから競合する置換更新を検知する．そして，削除操作を中止することで不整合の発生を防ぐ．

置換更新と異なり，追記更新と削除操作は競合するお互いを検知できず，不整合が生じ得る．図 5 の右図は，追記更新と削除操作を同時に行った状態である．右下の状態では，削除操作によりメタデータやステータスが準備されてる際に，追記更新により領域の確保が行われている．対象メタデータに変更がないため，削除操作を行うスレッドは競合する追記更新を検知できず，操作を完了する．結果として，先行する削除操作によって削除されたはずのキー（レコード）が，ノード内に残り不整合が生じる．

元論文の手法にて追記更新と削除操作を同時に行った場合に不整合が生じる原因は，追記更新と削除操作の変更方法が異なる点にある．追記更新が対象キーの存在を確認した後に新規レコードを追加するのに対し，削除操作は対象キーの存在を確認した後，メタデータのみを更新する．ゆえに，追記更新と削除操作とで線形化の方針が異なるため，同時に行われるお互いの操作を検知できない．

追記更新と削除操作を同時に行った場合に生じる不整合を防ぐため，本研究では削除操作も差分レコードによって表す．つまり，削除操作をペイロードの空値への更新と捉え，追記更新と同様の手続きで空値を表す差分レコードを挿入することでレコードを削除する．なお，ペイロードが 8 バイトの固定長であり対象のレコードが差分レコードでない場合は，元論文と同様にメタデータへの削除フラグ埋め込みによりレコードの削除を実現する．

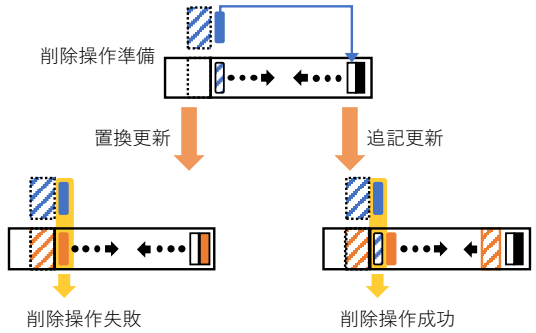


図 5　同時に削除操作と更新操作を行うことにより不整合が生じる例

4.2 構造変更操作の修正

元論文の手法では，複数の構造変更操作を同時に行うと操作結果が消失しうる．操作結果が消失しうるのは，大別して 2 つの場合である．1 つは，同階層にてノード分割（ないしマージ）とノード再編成を同時に行った場合である．もう 1 つは，ある階層でノード分割（ないしマージ）と同時に，1 つ上もしくは下の階層でノード分割やマージを行った場合である．

図 6 はノード分割と再編成を同時に行った結果，再編成を行ったスレッドによる書き込み操作の結果が消失する例である．図 6 の状態 1 ではあるスレッドがノード B を対象に再編成を行い，ノード B と新たに作成したノード B’ の交換を試みている．状態 2 はノード B の兄弟ノード C に対し他スレッドが同時に分割操作を行った状態であり，その後状態 3 のように再編成が先行したとする．このとき，状態 4 に示すように，再編成を完了したスレッドは当初の目的であるレコード書き込みを再編成後のノード B’ に対して行う．しかしノード分割を行うスレッドはノード B’ への書き込み操作を検知する術がないため，分割操作の完了後にはノード B’ に書き込まれたレコードが索引内から消失してしまう．

構造変更操作を同時に行う影響によるレコードの消失を防ぐため，本研究では構造変更操作におけるノードの凍結順序を変更する．図 7 のように，操作対象ノードの親ノードを交換するノード分割とノードマージについて，操作対象ノードの凍結と同時にその親ノードの凍結も行うよう変更する．これにより，操作対象ノードの親ノードが持つ子ノードへのポインタを対象とした PMwCAS 命令を防ぐ．すなわち，ノード分割やノードマージの対象ノードの兄弟ノードに対する構造変更操作は，PMwCAS 命令が成功せずに同時に走りえない．結果として，他の構造変更操作による操作対象ノードの兄弟ノードの差し替えを防ぎ，問題を解消する．

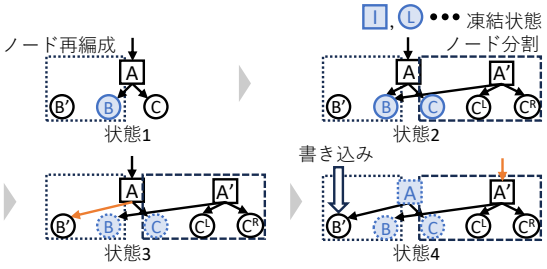


図 6　同時に行う構造変更操作により不整合が生じる例

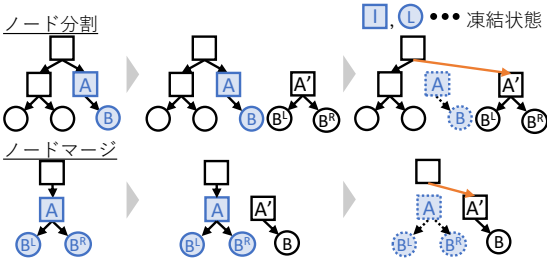


図 7　ノード分割とノードマージの改善手法

5 改　　善

本研究では，構造変更操作における追従処理を排除することで，Bz 木の改善を図る．改善内容を述べた後，追従処理を排除することに必要となるリカバリ処理について述べる．

5.1 構造変更操作における追従処理の排除

元論文の手法では，書き込み操作や削除操作の対象ノードが凍結状態であったスレッドは一旦操作を中断し，凍結状態の原因となった構造変更操作を追従する．操作対象ノードが凍結状態であった場合，まず，操作対象ノードの親ノードが他スレッドによる構造変更操作の対象となっていないか確認する．操作対象ノードの構造変更操作を追従した後に，削除する予定のノードにポインタを挿入することを防ぐためである．同様の理由から，操作対象ノードがその左右のノードに対して行われているノードマージの対象でないかも確認する．親ノードが他スレッドによる構造変更操作の対象になっている，もしくは操作対象ノードが左右のノードにおけるマージ操作の対象になっていた場合は，それらの操作を追従する．操作の対象になっていない場合は，操作対象ノードが持つレコード数やソート済みレコード数から，そのノードに行われている構造変更操作を特定する．その後，操作対象ノードに対して特定した構造変更操作を開始する．ゆえに，元論文に準ずる場合，複数のスレッドが同一の構造変更操作を行うこともあり得る．複数のスレッドが同一の構造変更操作を行った場合，最も速く操作を終了したスレッドの操作のみを反映し，その他のスレッドの操作は反映されない．

本研究グループでは既に揮発性メモリを対象とした Bz 木における性能測定 [24] にて，複数のスレッドが同一の構造変更操作を行う元論文の手法と比較し，構造変更操作の競合を検知したスレッドが対象ノードの再探索に戻る手法がより良い性能を示すことを観測している．不揮発性メモリ向け Bz 木を扱う本研究においても，元論文の手法ではなく，構造変更操作の競合を検知したスレッドが対象ノードの再探索に戻る手法を採用する．

5.2 リカバリ処理の追加

構造変更操作の追従処理を排除すると，リカバリ処理が必要となる．Bz 木の元論文にて一見非効率的な追従処理が採用されていた理由は，電源障害発生時のリカバリ処理を補うためである．構造変更操作中に電源障害が発生すると，凍結状態のノードが索引内に残り得る．電源障害により残された凍結状態のノードに対する構造変更操作は中止される．凍結状態ノードに対して書き込み操作は開始しないため，電源障害により残された凍結状態ノードは，特別な処置を施さない限り凍結状態であり続ける．この問題を解決するために元論文にて採用されていたのが，構造変更操作の追従処理である．書き込み操作の対象ノードが凍結状態である場合に追従処理が開始される．これにより，リカバリ後に対象ノードが凍結状態であり続けるために，操作が開始されない書き込み操作の発生を防ぐ．追従処理を排除した改善案では，代わりのリカバリ処理が必要となる．

リカバリ処理は，構造項変更操作で使用するメモリマネージャを活用する．全てのノードは，索引への挿入・削除時に必ず，メモリマネージャが管理する一時領域を経由する．そのため，一時領域上に残存するノード情報をログとして用いれば，追従処理は不要となる．

改善案におけるリカバリ処理の役割は 2 つに大別される．1 つは凍結状態ノードの除去である．電源障害により構造変更操作が中止された凍結状態ノードに対して，構造変更操作を再開することで，索引の再運用開始前に索引内の全ノードの凍結状態を解除する．もう 1 つは残留ノードによるメモリリークの防止である．構造変更操作にてメモリマネージャに管理される索引外のノードを，復帰時に解放することでメモリリークの発生を予防する．

リカバリ処理は 3 つの工程から成る．1 つ目の工程では，再開すべき構造変更操作を選定する．構造変更操作の最終工程では，トリガノードが新たに作製されたノードと交換される．ゆえに，メモリマネージャの管理する各トリガノードが，根ノードから辿れる場合に操作が終了していないと判断する．2 つ目の工程では，再開する構造変更操作に不必要なノードを解放する．解放するノードは，各トリガノード以外の全ノードである．新たに作製していたノードも，整合性の取れた状態であるか不明なため，解放した後構造変更操作を再開する．最後に，残ったトリガノードを対象に構造変更操作を再開する．構造変更操作の種類は，トリガノード内の状態から決定する．

6 性 能 評 価

改善を施した Bz 木の性能を，不揮発性メモリ向け B^+ 木の 1 つである FAST&FAIR [25] や，不揮発性メモリ向けスキップリストと比較することで評価する．Bz 木については，置換更新を採用した bz-inplace と，追記更新を採用した bz-append を用いる．それぞれの Bz 木は，本研究グループが改善を加えた PMwCAS 命令を使用する．FAST&FAIR については，読み取り操作の性能のみを比較する．

6.1 実 験 設 定

表 2 に本研究で用いるサーバの構成を示す．用いる構成は NUMA に対応しており，2 つの CPU ノードを持つ．本実験では，2 つの CPU の内 1 つのみを用いる．永続メモリを管理するメモリプールが，各ノードに分けて生成され，CPU ソケット間のリモートアクセスが生じることを避けるためである．

表 2　実験用サーバの構成

Item	Value
CPU	Intel(R) Xeon(R) Gold 6258R (two sockets)
DRAM	DIMM DDR4 (Registered) 2933 MHz (16GB × 12)
PMEM	DIMM Synchronous Non-volatile 2666 MHz (126GB × 8)
OS	Ubuntu 20.04.5 LTS
Compiler	GNU C++ ver. 9.4.0

表 3　ワークロード

Workload	Read/Write	Initial Keys	Access Pattern	Key Size
YCSB-A	0.5/0.5	1e7	random	8 byte
YCSB-C	1.0/0.0	1e7	random	8 byte

表 3 に本研究で扱うワークロードを示す．YCSB-A および YCSB-C は，YCSB（Yahoo! Cloud Serving Benchmark）[26] のワークロード A，C をそれぞれ表す．YCSB ワークロードは，一括挿入を用いて索引を構築した後に，読み取り操作と更新操作を行う．挿入済みのキーのみを扱うため，YCSB ワークロードにおける書き込み操作は，全て更新操作となる．ゆえに，YCSB ワークロードでは，構造変更操作の内，ノード再編成のみが発生し得る．各ワークロードでは各スレッドが最大 10^7 回操作し，総実行時間を総実行回数で割ってスループットを計算する．

各ワークロードにおけるキー分布は，式 (1) に示す Zipf の法則に従う．α は skew パラメータであり，N は要素の数を表す．α が 0 のとき，Zipf 分布は一様分布となる．α が 1 のとき，出現頻度が k 番目のキーは 1 番目のキーの出現頻度と比較して $\frac{1}{k}$ となる．すなわち，値 α が 0 から 1 に近づくほどキーに偏りが生じ，より競合の起きやすい設定となる．

$$f(k;\alpha,N) = \frac{1/k^{\alpha}}{\sum_{n=1}^{N} 1/n^{\alpha}} \qquad (1)$$

6.2　結　　果

YCSB-A および YCSB-C ワークロードにおいて，キー長を 8byte で固定しつつ，競合の生じやすい設定でスレッド数を変化させた時のスループットを図 8 および図 9 に示す．また図 10 は，YCSB-A においてスレッド数を 56 に固定した際の，各パーセンタイルに対応するレイテンシを示している．

YCSB-A の結果図 8 において，スキップリストが Bz 木よりも良い性能を示している理由は，スキップリストが 1 つのレコードのみを 1 つのノードに持つためであると考えられる．Bz 木は，1 つのノードに複数のレコードを持ち、かつ更新時に必ずステータスを PMwCAS の対象に含める．このため，Bz 木の競合の発生確率はスキップリストよりも高くなる．スキップリストも置換更新を採用しており，更新時には 1 ワードのみを PMEM へフラッシュする．これに対し、bz-inplace が合計 3 ワードのフラッシュが必要な点も，スキップリストと Bz 木の性能差の違いに影響していると考えられる．

読み取り操作のみからなる YCSB-C の結果図 9 から，bz-inplace の読み取り操作が，bz-append のものよりも高性能であることが伺える．この理由は，更新操作の違いによって生じる各読み取り操作の違いにあると考えられる．bz-inplace の読み取り操作は，置換更新を前提としてソート済みレコードの後に差分レコードを探索する．一方，bz-append の読み取り操作は，追記更新を考慮して差分レコードから探索する．このため，更新操作を含まない YCSB0-C においても，bz-inplace が bz-append より良い性能を示していると考えられる．また，YCSB-C におけるスキップリストの性能が，Bz 木や FAST&FAIR と比較して低

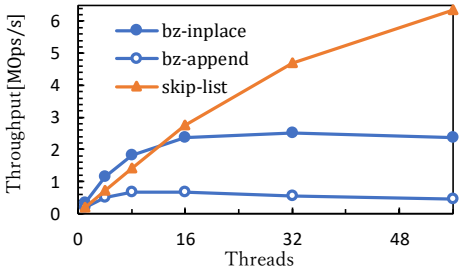


図 8　YCSB-A におけるスレッド数毎のスループット変化
（skew=1.0, key size=8 byte）

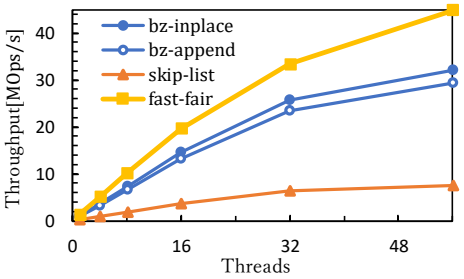


図 9　YCSB-C におけるスレッド数毎のスループット変化
（skew=1.0, key size=8 byte）

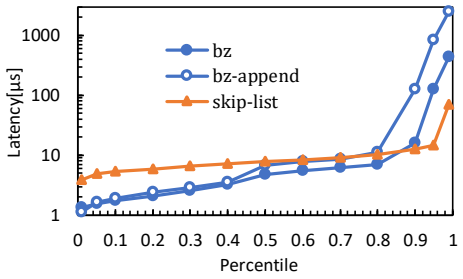


図 10　YCSB-A におけるテールレイテンシ
（skew=1.0, key size=8byte, threads=56）

い理由は，目的とするレコードにたどり着くまでに多くのリンクを経由する必要があるためであると推測される．FAST&FAIR は，扱うキーとペイロードが 8byte のみであることを前提に最適化を行っているため，Bz 木を上回る性能を出している．

2 つのワークロードにおいて，bz-inplace が bz-append と比較しより良い性能を示している理由は，読み取り操作と書き込み操作の違いにあると考えられる．読み取り操作の違いについては，先述した通りである．bz-inplace の扱う置換更新が必要とする PMwCAS 命令が 1 度のみであるのに対し，bz-append の扱う追記更新は 1 度に PMwCAS 命令を 2 回行う．更に，YCSB ワークロードは更新操作のみを行い，挿入操作を行わない．ゆえに，bz-append のみに，ノード再編成とそれによるノードの凍結が生じ得る．これらの要因により，bz-inplace と bz-append に性能差が生じていると考えられる．

更新操作自体の性能のみでなく，更新操作の再試行が bz-append の性能低下を招いていることが，図 10 から伺える．図 10 から，書き込み操作を含むワークロードにおいて，bz-

inplace や bz-append の性能が低下する理由は一部の高レイテンシな操作にあると考えられる．そして，レイテンシが高くなる理由は，競合による書き込み操作の中止や再試行であると推測される．この結果から，Bz 木の性能低下を防ぐためには，書き込み操作の中止や再試行を削減する工夫が必要となると考えられる．

6.3 考　　察

実験の結果から，Bz 木は読み取り操作の性能が高く，競合の高い状況での書き込み操作を苦手とすることが伺えた．ゆえに，Bz 木は読み取り操作主体のワークロードにて，より良い性能を発揮すると考えられる．そして，スキップリストとの比較により，Bz 木は実応用でも十分利用可能である可能性が示された．

Bz 木における書き込み操作の性能向上を図る方法の 1 つとして，PMwCAS 命令の排除が考えられる．Bz 木が使用する PMwCAS 命令は，複雑な永続メモリ上でのプロミングをより簡単に実現するための手法の 1 つであり，Bz 木に最適化された命令ではない．そのため，現状 PMwCAS 命令を用いている Bz 木の書き込み操作や構造変更操作に，PMwCAS 命令の代わりに独自のアルゴリズムを採用することが考えられる．PMwCAS 命令を使用せず，より各操作に最適化されたアルゴリズムへと変更することで，永続メモリに書き込む必要のあるデータ量を削減し得ると考える．

7 おわりに

本稿では，任意のデータ型を扱える永続メモリ向け索引である Bz 木に，修正と改善を施した．結果として，Bz 木本来の性能は既存研究で示されていたほど極端に悪いわけではなく，実応用でも十分利用可能であることが確認された．また，任意のデータ型を扱える Bz 木本来の性能を示すことにより，永続メモリ向け索引の性能基準の 1 つを示しすことができたと考える．今後の展望として，他の永続メモリ索引との性能比較が挙げられる．

謝　　辞

本研究の一部は JSPS 科研費 JP20K19804，JP21H03555，JP22H03594 の助成，日本電信電話株式会社との共同研究，および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

文　　献

[1] J. Arulraj, J. Levandoski, U. Minhas, and P. Larson, “BzTree: A high-performance latch-free range index for non-volatile memory,” *PVLDB*, vol. 11, no. 5, pp. 553–565, 2018.

[2] I. Oukid, J. Lasperas, A. Nica, T. Willhalm, and W. Lehner, “Fptree: A hybrid scm-dram persistent and concurrent b-tree for storage class memory,” 2016, pp. 371–386.

[3] J. Liu, S. Chen, and L. Wang, “Lb+trees: Optimizing persistent index performance on 3dxpoint memory,” *PVLDB*, vol. 13, no. 7, pp. 1078–1090, 2020.

[4] W.-H. Kim, M. K. Ramanathan, X. Fu, S. Kashyap, and C. Min, “Pactree: A high performance persistent range index using pac guidelines,” *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, 2021.

[5] S. Ma, K. Chen, S. Chen, M. Liu, J. Zhu, H. Kang, and Y. Wu, “ROART: Range-query optimized persistent ART,” in *19th USENIX Conference on File and Storage Technologies (FAST 21)*, 2021, pp. 1–16.

[6] X. Zhou, L. Shou, K. Chen, W. Hu, and G. Chen, “Dptree: differential indexing for persistent memory,” *PVLDB*, vol. 13, no. 4, pp. 421–434, 2019.

[7] L. Lersch, X. Hao, I. Oukid, T. Wang, and T. Willhalm, “Evaluating persistent memory range indexes,” *PVLDB*, vol. 13, no. 4, pp. 574–587, 2019.

[8] S. Chen and Q. Jin, “Persistent b+-trees in non-volatile main memory,” *PVLDB*, vol. 8, no. 7, pp. 786–797, 2015.

[9] J. Yang, J. Kim, M. Hoseinzadeh, J. Izraelevitz, and S. Swanson, “Nv-tree: Reducing consistency cost for nvm-based single level systems,” 2015, pp. 167–181.

[10] T. Wang, J. Levandoski, and P.-A. Larson, “Easy lock-free indexing in non-volatile memory,” in *Proc. ICDE*, 2018, pp. 461–472.

[11] P. L. Lehman and S. B. Yao, “Efficient locking for concurrent operations on B-trees,” *ACM TODS*, vol. 6, no. 4, pp. 650–670, 1981.

[12] S. K. Cha, S. Hwang, K. Kim, and K. Kwon, “Cache-conscious concurrency control of main-memory indexes on shared-memory multiprocessor systems,” in *PVLDB*, 2001, pp. 181–190.

[13] J. Levandoski, D. Lomet, and S. Sengupta, “The Bw-tree: a B-tree for new hardware platforms,” in *Proc. ICDE*, 2013, pp. 302–313.

[14] S. Scargall, *Programming persistent memory*. New York: Apress, 2020.

[15] Y. He, L. Duo, K. Huang, and T. Wang, “Evaluating persistent memory range indexes; part two,” *PVLDB*, vol. 15, no. 11, pp. 2477–2490, 2022.

[16] A. Van Renen, L. Vogel, V. Leis, T. Neumann, and A. Kemper, “Persistent memory i/o primitives,” in *Proceedings of the 15th International Workshop on Data Management on New Hardware*, 2019, pp. 1–7.

[17] S. Gugnani, A. Kashyap, and X. Lu, “Understanding the idiosyncrasies of real persistent memory,” *PVLDB*, vol. 14, no. 4, pp. 626–639, 2020.

[18] L. Benson, L. Papke, and T. Rabl, “Perma-bench: benchmarking persistent memory access,” *PVLDB*, vol. 15, no. 11, pp. 2463–2476, 2022.

[19] D. D. Sharma, R. Blankenship, and D. S. Berger, “An introduction to the compute express link (cxl) interconnect,” 2023.

[20] R. Kateja, A. Badam, S. Govindan, B. Sharma, and G. Ganger, “Viyojit: Decoupling battery and dram capacities for battery-backed dram,” in *International Symposium on Computer Architecture (ISCA’17)*, 2017.

[21] Y. Fridman, S. Mutalik Desai, N. Singh, T. Willhalm, and G. Oren, “Cxl memory as persistent memory for disaggregated hpc: A practical approach,” in *Proceedings of the SC ’23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, ser. SC-W ’23, 2023, p. 983–994.

[22] T. L. Harris, K. Fraser, and I. A. Pratt, “A practical multi-word compare-and-swap operation,” in *International Symposium on Distributed Computing*, 2002, pp. 265–279.

[23] 西村 学，杉浦 健人，石川 佳治，“永続メモリ向け multi-word compare-and-swap 命令の改善”，第 15 回データ工学と情報マネジメントに関するフォーラム（DEIM2023），2B–5，2023．

[24] 中山 宗，杉浦 健人，石川 佳治，“Bz 木におけるマルチスレッドでの構造変更操作に関する性能評価”，第 85 回全国大会講演論文集，vol. 2023，no. 1，2023，pp. 453–454．

[25] D. Hwang, W.-H. Kim, Y. Won, and B. Nam, “Endurable transient inconsistency in Byte-Addressable persistent B+-Tree,” in *16th USENIX Conference on File and Storage Technologies (FAST 18)*, 2018, pp. 187–200.

[26] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, “Benchmarking cloud serving systems with YCSB,” in *Proc. SoCC*. ACM, 2010, pp. 143–154.

ラベルの出現頻度に着目したFPGAを用いた正規パス問合せの提案

溝谷　祐大†　小林　諒平††　藤田　典久††　朴　泰祐††　天笠　俊之††

† 筑波大学大学院理工情報生命学術院　〒 305–8577 茨城県つくば市天王台 1 丁目 1-1
†† 筑波大学 計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1 丁目 1-1
E-mail: †mizotani@kde.cs.tsukuba.ac.jp, ††{kobayashi,amagasa}@cs.tsukuba.ac.jp,
††{fujita,taisuke}@ccs.tsukuba.ac.jp

あらまし　近年，グラフ分析は盛んに行われており，グラフから様々な情報が取得されている．グラフ分析の中でも，ユーザが望むデータを取得するための手法として，正規パス問合せ (RPQ) が存在する．RPQ とはエッジにラベルが貼られたグラフデータを対象とした問合せであり，指定されたラベルの並びを持つパスがグラフ中に存在するかどうかを探索し，存在する場合そのパスの始点・終点ノードを結果としてユーザに返す処理である．ここで課題となるのが，RPQ 評価の計算時間である．近年，データ分析において対象データの大規模化を受けてから，RPQ の対象となるグラフも大規模化が予想されており，現実世界に存在するような多種多様かつ大規模なグラフに対しては，実行に多大な時間を要することが想定される．そのような大規模なデータを処理するために FPGA (Field Programmable Gate Array) などのハードウェアアクセラレータの利用が注目されている．FPGA とは任意の回路をプログラミングによって繰り返し実装可能なハードウェアチップである．FPGA を用いた RPQ の高速化の既存研究では，FPGA の回路規模をすべて有効に利用できない場合が存在することや，複数 FPGA への拡張が困難といった課題点が存在する．そこで本研究では複数カーネルを利用して並列に RPQ 処理を行う手法を提案する．複数カーネルを用いることで，各カーネルが FPGA 内部で独立した回路として実装され並列動作が可能なため，FPGA の回路をより有効に活用できることや，今後複数 FPGA への手法の拡張が容易になることが利点として挙げられる．提案手法では，複数カーネルを用いた手法を実装するためにラベルの出現頻度に着目した．出現頻度が低いラベルをレアラベルを定義し，グラフとクエリをレアラベルを用いて分割することで，複数カーネルを用いた RPQ 処理が可能となる．評価実験では，レアラベルと定義するラベルの個数，クエリ中に出現するレアラベルの個数が多いときに RPQ 評価に要する時間が短くなることを確認した．また，一定の条件のもとで比較手法である，三浦らの手法よりも高速に RPQ 評価を行えることも確認した．

キーワード　FPGA, OpenCL，正規パス問合せ, ハードウェアアクセラレータ, グラフデータベース

1 はじめに

グラフ構造は，様々なデータをノードとエッジ，ラベルで表したデータ構造のことであり，我々の身の回りの多種多様なデータの関係性を表すのに有用である．実世界では，道路ネットワークや，分子構造，デジタル回路等様々なものがグラフとして変換され，保持されている．グラフの分析は盛んに行われており，グラフから様々な情報が取得されている [1].

このようなグラフデータからユーザが望むデータを取得するための手法として，幅優先探索や深さ優先探索，正規パス問合せ (RPQ : Regular Path Query) [2] などが存在する．中でも，RPQ とはエッジにラベルが貼られたグラフデータを対象とした問合せであり，指定されたラベルの並びを持つパスがグラフ中に存在するかどうかを探索し，存在する場合そのパスの始点・終点ノードを結果としてユーザに返すものである．

ここで課題となるのが，RPQ 評価の計算時間である．近年の対象データの大規模化を受け，RPQ の対象となるグラフも大規模化により，現実世界に存在する様な多種多様かつ大規模なグラフに対しては実行に多大な時間を要することが想定される．近年では，そのような大規模なデータを処理するために GPU (Graphics Processing Unit) や FPGA (Field Programmable Gate Array) といったハードウェアアクセラレータの利用が注目されている．FPGA とは任意の回路をプログラミングによって繰り返し実装可能なハードウェアチップである．FPGA は実装された各回路の並列性を利用した，並列度の高い処理が可能である．また，FPGA は GPU と比べて動作周波数が低く，消費電力を抑えることができる．そのため，単位電力あたりのスループットは GPU を上回ることもある．さらに，FPGA は CPU や GPU と比べると低レイテンシとなっている．このような利点から，近年，FPGA は機械学習や金融市場，ビッグデータ処理と様々な分野で利用され始めている [3]．しかし，FPGA は Verilog や VHDL のようなハードウェア記述言語によってプログラムされることが一般的であり，大規模なプログラムの開発には多大なコストを要していた．しかし近年，C/C++や OpenCL 等の高水準言語を利用して FPGA のプログラムを可能にした，高位合成と呼ばれる技術が確立し，開発コストを抑えることが可能となっている．

FPGA を用いた RPQ の高速化の既存研究では，FPGA の回路規模をすべて有効に利用できない場合が存在することや，複

数 FPGA への拡張が困難といった課題点が存在する．そこで本研究では複数カーネルを利用して並列に RPQ 処理を行う手法を提案する．複数カーネルを用いることで，各カーネルが FPGA 内部で独立した回路として実装され並列動作が可能なため，FPGA の回路をより有効に活用できることや，今後複数 FPGA への手法の拡張が容易になることが利点として挙げられる．提案手法では，複数カーネルを用いた手法を実装するためにラベルの出現頻度に着目した．提案手法では，複数カーネルを用いた手法を実装するためにラベルの出現頻度に着目した．出現頻度が低いラベルをレアラベルを定義し，グラフとクエリをレアラベルを用いて分割することで，複数カーネルを用いた RPQ 処理が可能となる．

評価実験では，レアラベルの定義数を増やすことで処理時間が短くなること，クエリ中に出現するレアラベルの個数が増えることで RPQ 評価に要する処理時間が短くなることを確認した．また，特定の条件のもとで比較手法である三浦らの手法よりも高速に RPQ 評価を行えることも確認した．

本稿の構成を以下に示す．第 2 節で前提知識，第 3 節で関連研究を紹介する．第 4 節で提案手法について述べ，第 5 節で実験結果を示す．最後に第 6 節で本研究のまとめを述べる．

2 前提知識

2.1 正規パス問合せ (RPQ)

正規パス問合せ (Regular Path Query) とは，ラベル付きグラフ中のユーザが指定したパスを任意の 2 ノード間に存在するかを調べる問合せである．処理対象となるラベル付きグラフを $G = (V, E, \mathcal{L})$ とすると，以下のような RPQ 評価のセマンティクス R が定められる．

$$R := \varepsilon \mid l \mid l^{-} \mid R \circ R \mid R \cup R \mid R^{i,j} \qquad (1)$$

ここで ε は空列，l はラベルを順方向に，l^{-} はラベルを逆方向に進むことを表し，$\circ$ は結合，$\cup$ は和，$R^{i,j}$ は R の $k\,(i \leqq k \leqq j)$ 回の繰り返しを表している．RPQ の例として，図 1 で示すグラフ G_s を探索対象とした場合，

$$RPQ\ :\ R\ =\ (a \circ b) \cup (a^{-} \circ c^{1,2}) \qquad (2)$$

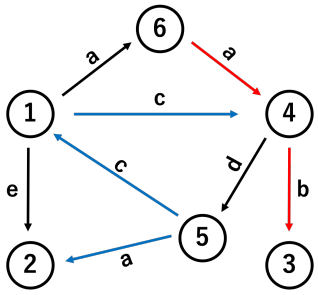
を適応すると

$$R(G_s)\ =\ \{(6,3),(2,1),(2,4)\} \qquad (3)$$

となる．前半のクエリ $(a \circ b)$ でパス $6 \to 4 \to 3$ が得られ，後半のクエリ $(a^{-} \circ c^{1,2})$ でパス $2 \to 5 \to 1, 2 \to 5 \to 1 \to 4$ が得られる．そのため，RPQ の結果として $\{(6,3),(2,1),(2,4)\}$ が得られる．

2.2 FPGA

Field Programmable Gate Array (FPGA) とは任意の回路をプログラミングによって繰り返し実装可能なハードウェアチップである．FPGA は実装された各回路の並列性を利用した，並列度の高い処理が可能である．



グラフG_s

図 1: RPQ の例 (式 (2) を適応)

また，FPGA は Graphics Processing Unit (GPU) と比べて低レイテンシであり，動作周波数が低く，消費電力を抑えることができる．このような利点から，近年，FPGA は機械学習やビッグデータ処理など様々な分野で利用され始めている．

従来，FPGA は Verilog や VHDL のようなハードウェア記述言語によってプログラムされることが一般的であり，大規模なプログラムの開発には多大なコストを要していた．しかし，近年 C/C++や OpenCL 等の高水準言語を利用して FPGA のプログラムを可能にした，高位合成 (HLS) と呼ばれる技術が確立し，開発コストを抑えることが可能となっている．

本研究では，OpenCL を用いて FPGA 上への回路の実装を行った．

3 関連研究

3.1 FPGA を利用した様々な分野での高速化

FPGA を利用した処理の高速化に関する研究は，様々な分野で行われている．グラフ分析の分野で Zhou ら [4] は，PageRank の高速化手法を提案した．また，機械学習の分野では，Elnawawy ら [5] がインターネット・トラフィックのリアルタイムでの分類について，ランダムフォレストを FPGA 上に実装した手法を提案している．さらに，複数 FPGA を用いたグラフ探索アルゴリズムの高速化事例として Dai ら [6] の研究がある．

3.2 RPQ の応用事例

RPQ の応用分野として Resource Description Framework (RDF) XML (Extensible Markup Language) やグラフデータベースなどが挙げられる．SPARQL1.1 [7] は RDF に対して単純な RPQ に対応している．近年，より複雑な RPQ に対応した SPARQL クエリ言語が研究されている [8].

3.3 RPQ 評価の高速化に関する研究

RPQ 評価を高速化する手法としてオートマトンベースの手法とインデックスベースの手法が挙げられる．

インデックスベースの手法として Path Index [9] が挙げられる．Path Index ではグラフ中の任意のパスを設定した長 k まで事前にインデクシングを行う．図 1 のグラフ G_s を例とすると $k = 2$ とした場合，作成されるインデックスは表 1 の様になる．このように事前にインデクシングを行うことで長さが k 以下の

パスの探索の場合，インデックスを参照するだけで即座に結果を取得することが可能となる．また，長さが k より大きいパスにおいてもインデックス同士を結合する事によって効率的に結果を取得することが可能となる．

表 1: Path Index の例 ($k = 2$ としてグラフ G_s (図 1) に適応)

(a) $k = 1$

パス	始点ノード	終点ノード
a	1	6
c	1	4
e	1	2
b	4	3
d	4	5
a	5	2
c	5	1
c	6	4

(b) $k = 2$

パス	始点ノード	終点ノード
$c \circ b$	1	3
$a \circ a$	1	4
$c \circ b$	1	5
$d \circ c$	4	1
$d \circ a$	4	2
$c \circ c$	5	4
$c \circ a$	5	6
$a \circ b$	6	3
$a \circ d$	6	5

また，オートマトンベースの高速化手法として Koschmieder らの手法 [10] も挙げられる．オートマトンベースの手法では与えられたクエリをもとにオートマトンを生成し，処理対象のグラフに適応していく．例として $R = (a \circ b) \cup (b^{1,2} \circ c)$ が与えられた場合，生成されるオートマトンは図 2 の様になる．この生成されたオートマトンをベースにグラフ中のすべてのノードを始点として探索が行われる．

Koschmieder らは，グラフ中に出現するラベルの頻度に着目した高速化を行った．グラフ中での出現頻度が低いラベルをレアラベルとして定義し，与えられたクエリをそのレアラベルで分割する．クエリの分割後，各クエリを満たすパスの探索を双方向探索アルゴリズムを用いて行う．それぞれのクエリを満たすパスとレアラベルを結合していき，結合が行われなかったパスは全体のクエリを満たすことがないためグラフ中から削除される．この処理を行うことによって，探索するクエリ長が短くなりクエリを満たすパスが探索を行うごとに少なくなっていくため，高速にパスの探索が可能となっている．結果として，従来のオートマトンベースの手法と比べ 2 倍から 6 倍の高速化を達成している．

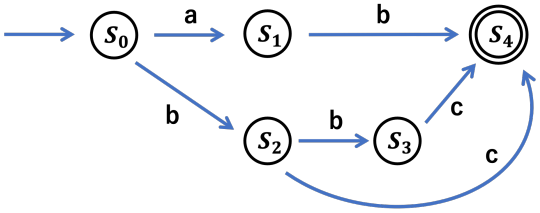


図 2: $R = (a \circ b) \cup (b^{1,2} \circ c)$ から生成されるオートマトン

4 FPGA を用いた RPQ 評価の高速化

FPGA を利用した RPQ 評価の高速化事例として，三浦らのオートマトンベースの手法 [11] がある．この研究では，NFA を FPGA 上に実装を行い RPQ 評価を行っている．Sidler らの研究 [12] において，FPGA 上に任意の NFA を実装するための実装方法が提案されている．三浦らは，その手法を参考に FPGA 上への NFA の実装を行っている．

三浦らが実装した NFA の構造を図 3 に示す．この NFA は Tokens，Trigger，State Transitions の 3 つの設定データを用いることで任意の NFA を実現している．Tokens は RPQ に現れるラベルの種類，Trigger はどのラベルがどの State へ状態遷移を起こすか，State Transitions は State 間の遷移をそれぞれ指定する．この手法の利点として，ユーザから与えられた RPQ に対してこれらの設定データを指定することで，FPGA 上の回路を変更する事なく任意の RPQ 評価を行うことができる点である．これは，一般的に FPGA の回路合成に数時間程度の時間を要することを考えると大きな利点であることがわかる．実際のパスの探索では，この NFA と Frontier を用いた幅優先探索を組み合わせた手法で行っている．

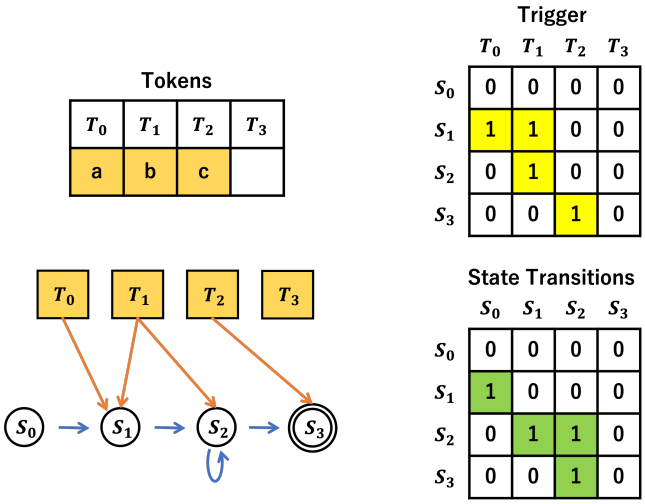


図 3: NFA の構造 ($R = (a \cup b) \circ b^{1,2} \circ c$)

5 提案手法

提案手法では，グラフデータを取得したあとでレアラベルを取得し，そのレアラベルのみを持つグラフデータとそれ以外のグラフデータに分割するステップ，与えられたクエリをレアラベルで分割するステップと FPGA 上で RPQ 評価を行うステップ，最後に得られたすべてのパスを結合し，元のクエリを満たすパスに変換するステップに分けて考える．

5.1 提案手法の処理の概要

図 4 に提案手法の処理の概要を示す．全体の処理は Host 側 (FPGA と通信を行う PC) で行われる処理と FPGA 側で行われる処理に分かれる．まず，処理対象のグラフデータを解析しどのラベルがレアラベルと成りうるかを決定する．その後，5.2 節で後述する方法でグラフデータを分割する．

また，ユーザから RPQ が与えられると，Host 側はそのクエリをレアラベルで分割し，すべてのクエリに対して 4 節で前述した NFA を生成する．その後，生成した NFA とすべてのグラフデータを隣接リスト形式で FPGA に送信する．NFA は FPGA に内蔵されている Block RAM に，グラフデータは FPGA ボー

ド上のグローバルメモリにそれぞれ配置される．FPGA は対象グラフ中のパスの探索と探索されたパスに対しての NFA を用いた RPQ 評価を行い，それぞれのクエリに対してそのクエリを満たすパスが FPGA から Host へ送信される．最後に，Host 側で受け取ったそれぞれのクエリを満たすパスに対して結合処理を行い，最終的にユーザから与えられた RPQ を満たすパスを結果とする．

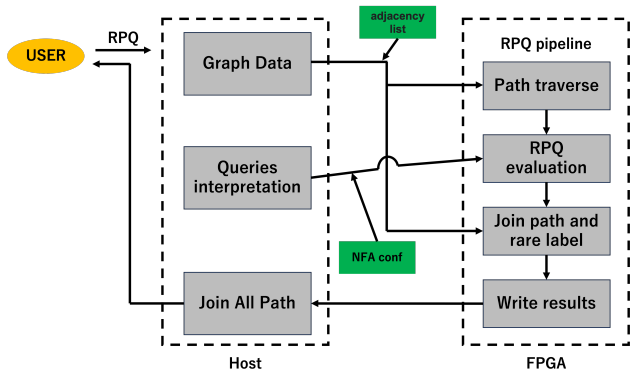


図 4: 全体の処理の概要

5.2 レアラベルの検出，グラフの分割

提案手法では，与えられたグラフデータからすべてのラベルが何回出現するかを確認する．その後，与えられたパラメータ (α) に基づき，レアラベルを決定する．α はユーザが指定するパラメータであり，グラフ中のラベルの中で出現率が低い α 個のラベルをレアラベルとして定義する．

図 5 の場合，$\alpha = 1$ のとき，グラフ中に現れるラベルの中で最も出現頻度の低いラベルがレアラベルとして定義される．図より，その条件を満たすラベルはラベル d であるため，グラフは図のようにラベル d を除いたグラフとラベル d のみのグラフの 2 つのグラフに分割される．

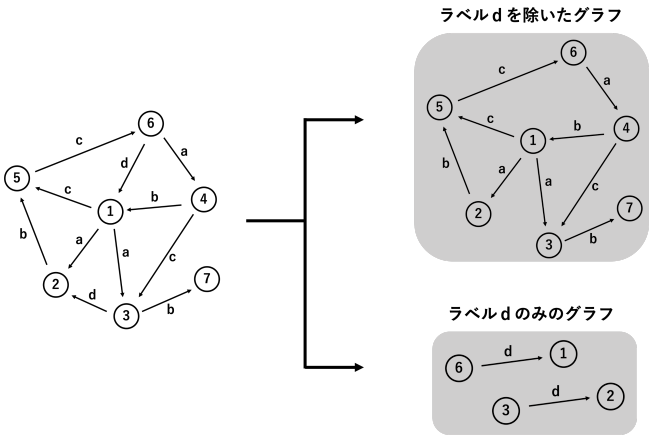


図 5: グラフ分割の例

5.3 クエリ分割

提案手法では，レアラベルの決定後，ユーザから与えられたクエリをレアラベルで分割し，それぞれのクエリに対する NFA を生成する．クエリ分割では，ユーザから与えられたクエリを前から探索していきレアラベルを発見すると，その前後でクエリを分割する実装とした．

ここで例として，ユーザから与えられた RPQ が $R = (a \cup c) \circ d \circ b$ であり，レアラベルが d の場合，レアラベルであるラベル d でクエリを分割するため，R は $R' = a \cup c$, $R'' = b$ の 2 つのクエリに分割される．その後，R', R'' に対応した NFA が生成される．

この実装方法では，クエリ中のレアラベルに和演算 ($\cup$) 等が含まれる場合について RPQ 評価の対象外としてる．つまり，クエリ中にレアラベルが出現したとき，その前後には単純な結合 ($\circ$) のみが現れることを前提とする．

しかし，レアラベルの前後に和演算 ($\cup$) 等が含まれる場合でも，それらのパスを事前に探索し，レアラベルのみのグラフに追加することで対応が可能となる．

5.4 レアラベルを用いた FPGA による正規パス問合せの提案

図 6 に FPGA 内の実装の概要を示す．提案手法では FPGA 上にカーネルを 4 つ定義して RPQ 処理を行う．複数のカーネルを利用することで，各カーネルが FPGA 内部で独立した回路として実装され並列動作が可能となり，FPGA の回路をより有効に活用できる．

5.4.1 FPGA の処理の概要

それぞれのカーネルの概要を以下に示す．

カーネル (1) : すべてのクエリに対するパスの探索，RPQ 評価を行い，RPQ 評価の結果に関わらず見つけたすべてのパスをカーネル 1 に送信

カーネル (2) : カーネル 1 から受け取ったすべてのパスから，RPQ 評価を満たすパスのみをカーネル 3 に送信

カーネル (3) : カーネル 2 から受け取ったパスに対してレアラベルとそのパスを結合し，その結果をカーネル 4 に送信

カーネル (4) : カーネル 3 から受け取ったパスをグローバルメモリにセット

各カーネル間のパスの送受信にはインテル FPGA SDK for OpenCL の拡張機能であるチャネルを利用する．チャネルを利用することで，外部メモリを通さずに FPGA 内部のみでデータの送受信が可能となり，カーネル間でのデータの受け渡しを高効率かつ低レイテンシに行うことが可能となる．

また，カーネル 1 で行うパスの探索，RPQ 評価は 4 節で前述した三浦らの手法をもとに実装を行った．

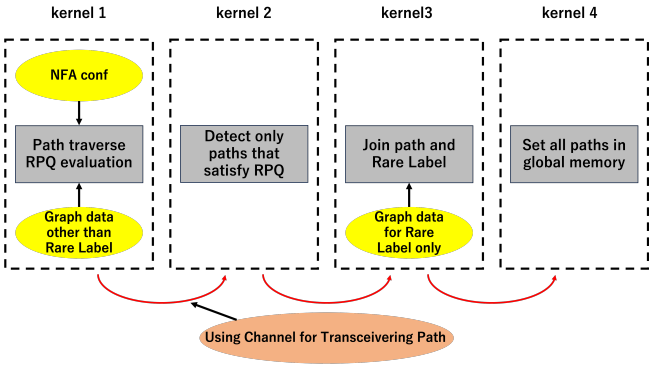


図 6: FPGA 内の実装の概要

5.4.2 RPQ 評価を満たすパスとレアラベルとの結合

提案手法では，ユーザから与えられた RPQ をレアラベルで分割するため，それぞれの RPQ を満たすパスとレアラベルを持つパスとを結合する必要がある．そこで，提案手法では前述した通りカーネル 3 で得られたパスとレアラベルを持つパスとの結合処理を行う．

まず，カーネル 3 ではグラフデータのうち，レアラベルのみで構成されるグラフデータで隣接リストに変換されたものをグローバルメモリに配置する．次に，カーネル 2 から受け取ったパスに対してレアラベルを持つパスとの結合ができるか判定する．例として，R' を満たすパスを (α, β) とするとレアラベルを持つパスとの結合は，R' を満たすパスの終点ノードである β が任意のノードとレアラベルが貼られたエッジを持つときに行う．したがって，パス (β, γ) に貼られているラベルがレアラベルのときパス (α, γ) をカーネル 4 に送信すればよい．

図 5 を例に取ると，パス $(1, 3)$ が R' を満たすパスであるため，カーネル 2 からカーネル 3 に送信される．そのパスをカーネル 3 が受け取ると，パスの終点ノードである 3 について，パス $(3, 2)$ がレアラベルが貼られたエッジであることがわかり，最終的にパス $(1, 2)$ をカーネル 4 に送信する．

また，R'' についてはレアラベルと結合する必要がないことがわかる．これは，与えられた RPQ の最後がレアラベルでないときにレアラベルと結合する必要がないためである．したがって，RPQ の最後がレアラベルでないときはカーネル 2 から受け取ったパスを，そのままカーネル 4 に送信する．

5.4.3 得られたパスの結合

提案手法ではクエリをレアラベルで分割しそれぞれのクエリに対して探索を行うため，それぞれの探索で得られるパスは元のクエリの一部を満たすパスとなる．そのため，最終的にクエリを満たすパスを結果として得るために，それぞれで得られたパスを結合する必要がある．提案手法では，得られたパスの結合は FPGA でなく CPU 上で行う実装とした．

図 5 の場合，R' を満たすパスと R'' を満たすパスのそれぞれを表 2 に示す．これらのパスを結合した結果得られるパスが最終的な RPQ を満たすパスとなる．

パス同士の結合は R' を満たすパスの終点ノードが R'' を満たすパスの始点ノードと一致している場合にそれらのパスが結合可能となる．$R' \circ d$ を満たすパス $(1, 2)$ と R'' を満たすパス $(2, 5)$ はそれぞれ上記の成約を満たしているため，これらのパスを結合することで R を満たすパス $(1, 5)$ が得られる．

表 2: $R' \circ d$，R'' をそれぞれ満たすパス

(a) $R' \circ d$ を満たすパス

始点ノード	終点ノード
1	2
5	1

(b) R'' を満たすパス

始点ノード	終点ノード
2	5
3	7
4	1

6 実　　験

本実験で使用した実験環境を以下の表 3 に示す．また，Host 側のコンパイルには g++ (GCC) 4.8.5 を，コンパイルオプションは-O3 を使用した．

表 3: 実験環境

(a) Host

CPU	Intel Xeon CPU E5-2690 v4 @ 2.60GHz (× 2)
OS	CentOS Linux (version 7)
Memory	DDR4 64GB (32GB / CPU)
Intel FPGA SDK for OpenCL	Version 19.4.0 Build 64 Pro Edition

(b) FPGA

Board	BittWare 520N
FPGA	Intel Stratix 10 GX 2800
Memory	Four banks of DDR4 SDRAM x 72 bits 8GB per bank (Total 32GB)
I/O	PCIe Gen3 x16
BRAM	229Mbits
MLAB	15Mbits

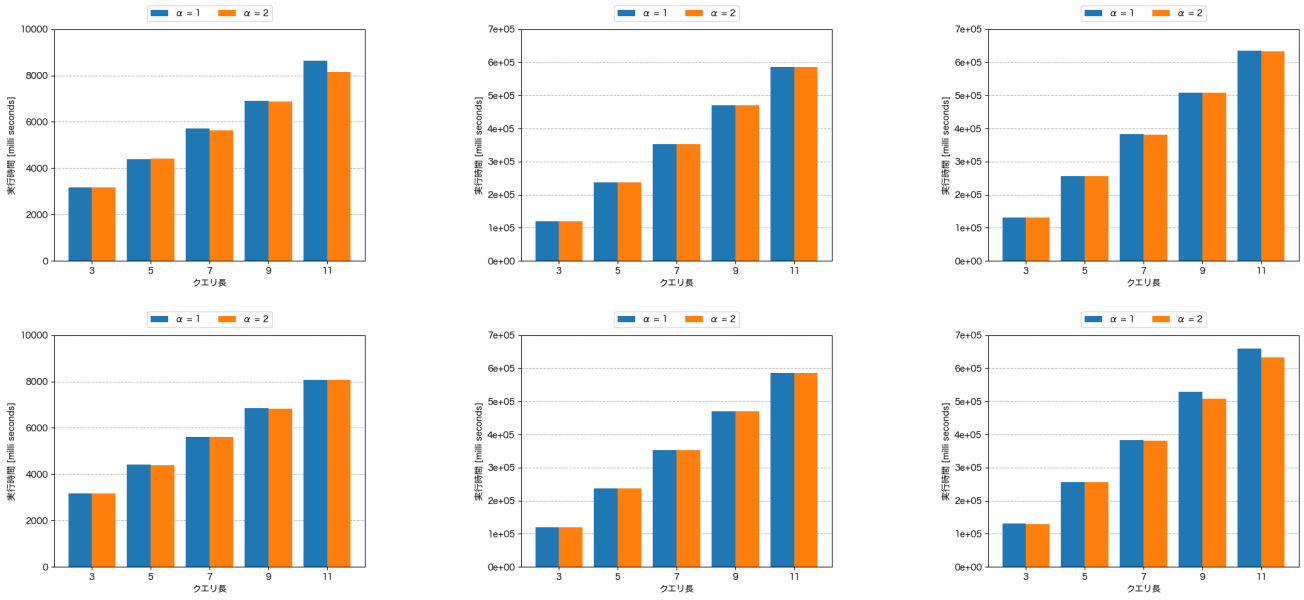
6.1 データセット

本実験では，Advogato [13]，DBLP Citation Network [14]，Reddit Hyperlink Network [15] の 3 種類のデータセットを用いて実験を行う．それぞれのデータセットの詳細は後述する．

また実験のために，それぞれのデータセットについて総ノード数，総エッジ数を変えずにラベルの情報のみを変更することとする．これは，レアラベルの出現率が提案手法にどのような影響を与えるかを調査するためである．ユーザが指定するパラメータ α について， α 個の任意のラベルの出現率が全体の RARE_PARAM%以下になるようにラベルの情報を変更する．実験では，$\alpha = 1, 2$， RARE_PARAM $= 5, 10$ の計 4 種類のラベルの変更を各グラフに行い，実験を行った．

6.1.1 Advogato

Advogato [13] は，フリーソフトウェアの開発者向けのオンラインコミュニティプラットフォームから作成された有向グラフで，ユーザ同士の信頼関係をラベル付きエッジで表したものである．advogato の情報を表 4 に示す．



(a) Advogato (b) DBLP (c) Reddit Hyperlink Network

図 7: パラメータによる影響

表 4: Advogato の詳細

ノード数	エッジ数	ラベルの種類
6541	51127	5

6.1.2 DBLP Citation Network

DBLP Citation Network [16] はコンピュータサイエンスに関する書誌サイト DBLP から収集した約 300 万件の論文データを集めたデータセットである．本実験では，この論文データから論文 ID，著者，引用論文 ID，開催地を抽出し，論文 (P)，著者 (A)，開催地 (V) をノードとする異種グラフデータを作成した．ラベルの種類は 6 種類とし，それぞれ $A \xrightarrow{writing} P$，$P \xrightarrow{writing\ by} A$，$V \xrightarrow{publishing} P$，$P \xrightarrow{published\ in} V$，$P \xrightarrow{citing} P$，$P \xrightarrow{cited\ by} P$ である．今回本実験が DBLP Citation Network から生成したグラフの詳細を表 5 に示す．

表 5: DBLP Citation Network の詳細

ノード数	エッジ数	ラベルの種類
65536	195039	6

6.1.3 Reddit Hyperlink Network

Reddit Hyperlink Network は英語圏のウェブサイトであり多種多様なウェブサイトへのリンクを収集し公開している Reddit [17] から作成されたグラフデータセットである．ここで，グラフ中の各ノードは subreddit と呼ばれる Reddit 内で特定のドメインに焦点を当てたコミュニティを表しており，各エッジは subreddit 間に貼られたハイパーリンクを表す．様々な属性がこのエッジに付与されており，その中の 1 つにエッジの始点コミュニティから終点コミュニティに対して何時にリンクが作成されたかを表す属性が存在する．この属性は 24 時間表記で表されており，本実験ではこの属性を 2 時間刻みで 12 分割し，12 個のエッジラベルに変換した．今回本実験が Reddit Hyperlink Network から生成したグラフの詳細を表 6 に示す．

表 6: Reddit Hyperlink Network の詳細

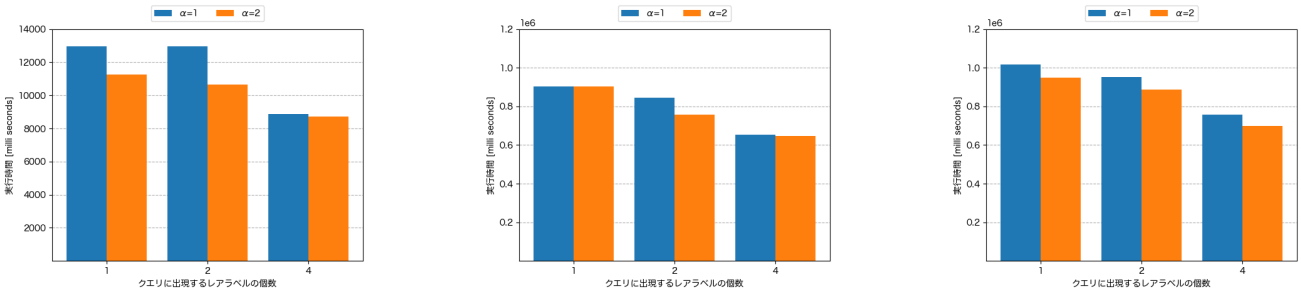
ノード数	エッジ数	ラベルの種類
67179	858488	12

6.2 パラメータによる性能比較

本実験では，パラメータ α，RARE_PARAM を変化させたときの性能比較を行う．図 7 に実験の結果の全体を示す．それぞれ上段の図が RARE_PARAM = 5，下段の図が RARE_PARAM = 10 に対する結果を示している．

α による影響：図より，クエリ長が 3，5，7，9 の場合では，α の値が変化しても，処理時間に大きな差は現れなかった．しかし，図 7 (a) の上段で，クエリ長が 11 の場合を見ると，$\alpha = 2$ のほうが処理時間が顕著に短くなっている．これは，クエリ長が長い場合，RPQ 評価のコストが大きくなるが，レアラベルと定義するラベルの個数が増えることで RPQ 評価の対象となるグラフデータの規模が小さくなるからである．また，レアラベルと定義するラベル数が多くなる場合でも，提案手法ではレアラベルとパスとの結合処理にかかるコストは変わらないことも理由である．

RARE_PARAM による影響：図より，ほとんどの場合で RARE_PARAM の違いによる影響は確認できなかった．これは，RARE_PARAM が小さいほどカーネル 3 で行うパスとレアラベルの結合に要するコストが減少する．しかし，グラフデータを分割する際，レアラベルを除いたグラフが元のグラフデータのサイズと大きく変わらない．一方で，RARE_PARAM が大きいとカーネル 3 の処理コストが増加するが，レアラベルを除いたグラフのサイズが小さくなるため，RPQ 評価に要する時間が短くなる．このような原因から RARE_PARAM の違いによる



(a) Advogato (b) DBLP (c) Reddit Hyperlink Network

図 8: クエリ中に出現するレアラベルの個数による影響

影響が確認できなかったと考える．

6.3 レアラベルの出現回数による実行時間の比較

本実験では前述した 3 つのデータセットを用いて，RPQ で指定するクエリにレアラベルの出現回数を変化させた際の提案手法の性能調査を行った．本実験の対象となるクエリは，和演算等を含まない結合演算のみの単純なクエリである．

この実験では，評価対象となる元のクエリ長を 16 に固定し，その中にレアラベルが 1 つ，2 つ，4 つの場合で実験を行う．また，実験で使用するクエリは，レアラベルの個数のみ指定し，それ以外の部分はランダムに生成したものを使用する．実験結果に示す処理時間はランダムに生成したクエリに対して RPQ 評価を 10 回行ったときの平均値を示す．

実験結果を図 8 に示す．図より，クエリ中に出現するレアラベルの個数が多くなると，RPQ 評価に要する時間が短くなっている．また，Advogato では α の変化による処理時間の変化は見られなかった．しかし，DBLP，Reddit Hyperlink Network では，すべての場合で $\alpha = 2$ のほうが高速になっている．

また，図 9 に，Advogato におけるカーネル 1 からカーネル 3 のそれぞれの処理時間を示した．この図からわかる通り，FPGA での処理時間はクエリに出現するレアラベルの個数が多くなるにつれて短くなっている．これは，より多くのレアラベルが出現するクエリでは，RPQ 評価の対象となるクエリの数が多くなるがその分，一つ一つのクエリ長が短くなり，1 つのクエリに対する処理時間が短くなるからだと考えられる．

また，図からカーネル 2 の処理時間が最も長いことがわかる．これは，カーネル 3 の RPQ 評価を満たすパスとレアラベルとの結合処理を行う．このとき，レアラベルのグラフ情報にアクセスするためにグローバルメモリへのアクセスが生じていることが原因と考えられる．通常 FPGA ではグローバルメモリへのアクセスは低速であるため，カーネル 2 ではカーネル 3 がパスを読み取ってから処理が終了するまで，チャネルへの書き込みを待つ必要があるため処理時間が長くなっている．

6.4 比較手法との性能評価

本実験では，比較手法である Path Index [9]，三浦らの手法 [11] との実行時間の比較を行う．今回使用したデータセットは Advogato と Reddit Hyperlink Network の 2 種類である．また，実行に 2,000 秒以上かかる場合はタイムアウトとした．

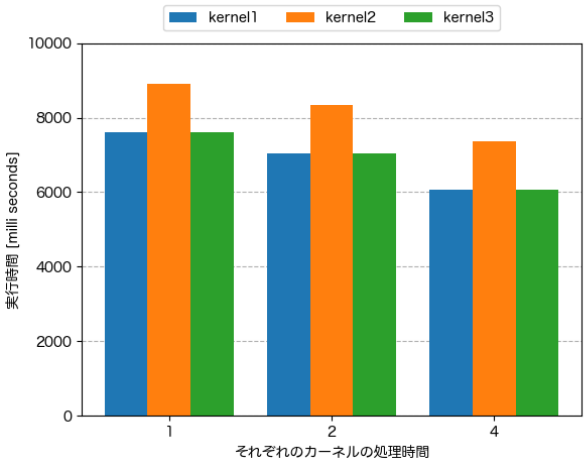


図 9: 各カーネルの処理時間 (Advogato)

図 11 に実験の結果を示す．図からわかる通り，Path Index がパスの長さが大きくなるに連れて処理時間も指数関数的に増加している．三浦らの手法ではパスの長さに比例した形で処理時間が増加している．本手法においてもパスの長さによる実行時間の増加具合は比例的になっている．しかし，提案手法は三浦らの手法と比べ低速になっている．これは三浦らの手法の特徴である，処理時間がクエリ長に比例することが原因である．提案手法では，クエリをレアラベルで分割し，それぞれのクエリに対して RPQ 評価を行っていく．そのため，クエリ長が短い場合，RPQ 評価を行う回数が増えることがボトルネックとなり低速になっている．

また，図 10 にクエリ長を 31 とした場合の三浦らの手法と提案手法の処理時間を示す．本実験では，クエリに出現するレアラベルの個数を 4 個とした．図より，この実験では，Advogato, Reddit Hyperlink Network に対しての RPQ 評価で，提案手法が三浦らの手法よりも高速に行えていることがわかる．

7 ま と め

本研究ではラベルの出現頻度に着目した FPGA を用いた正規パス問合せの手法を提案した．探索対象のグラフデータの中で出現頻度が低いラベルをレアラベルと定義して，そのレアラベルでクエリを分割する．FPGA 上では，カーネルを 4 つに分けることで各処理を並列的に処理できる実装とした．結果として，

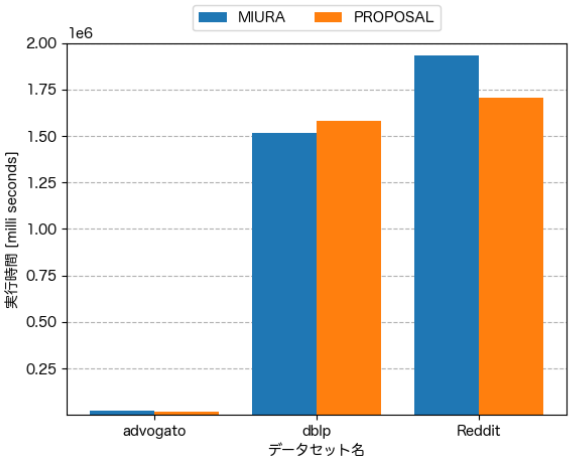
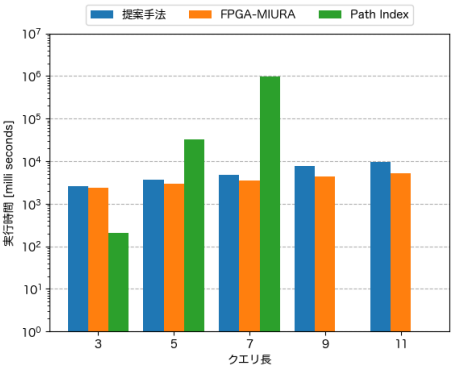
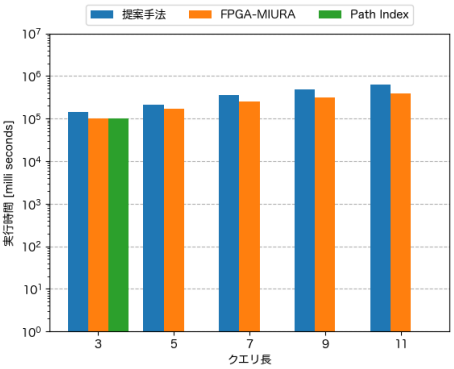


図 10: クエリ長が長い場合の処理時間の比較



(a) Advogato



(b) Reddit Hyperlink Network

図 11: 比較手法との処理時間の比較

レアラベルと定義するラベル数を増やすことでグラフを分割したときに，RPQ 評価の対象となるグラフデータのサイズを小さくすることができる，提案手法ではレアラベルとパスとの結合処理に要するコストはレアラベルの定義数に影響しないという点から元のクエリ長が長い場合に，レアラベルの定義数を増やすことで処理時間が短くなるという特徴が見られた．また，クエリ中に出現するレアラベルの個数が多くなることで，探索対象となるクエリ長が短くなり，一つ一つのクエリに対する RPQ 評価に要する時間が短くなる．これによって，クエリ中に出現するレアラベルの個数が増えることで RPQ 評価に要する処理時間が短くなるという特徴が見られた．評価実験では，クエリ長が長い場合に，三浦らの手法と比べ提案手法が高速に RPQ 評価を行えることもわかった．しかし，多くの点で提案手法は三浦らの手法と比べると低速となっているため，高速化が必要である．今後の課題として，複数 FPGA を用いた手法などが挙げられる．

謝　　辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業，JST CREST（JPMJCR22M2），科学研究費補助金（JP22H03694，JP22K17895）の支援によるものです．

文　　献

[1] Daniel E O'Leary. Artificial intelligence and big data. *IEEE intelligent systems*, Vol. 28, No. 2, pp. 96–99, 2013.
[2] Isabel F Cruz et al. A graphical query language supporting recursion. *ACM SIGMOD Record*, Vol. 16, No. 3, pp. 323–330, 1987.
[3] Shijie Zhou et al. Hitgraph: High-throughput graph processing framework on fpga. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 30, No. 10, pp. 2249–2264, 2019.
[4] Shijie Zhou et al. Optimizing memory performance for fpga implementation of pagerank. In *2015 International Conference on ReConFigurable Computing and FPGAs (ReConFig)*, pp. 1–6. IEEE, 2015.
[5] Mohammed Elnawawy et al. Fpga-based network traffic classification using machine learning. *IEEE Access*, Vol. 8, pp. 175637–175650, 2020.
[6] Guohao Dai et al. Foregraph: Exploring large-scale graph processing on multi-fpga architecture. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 217–226, 2017.
[7] Sparql 1.1 query language, December 4, 2020.
[8] Faisal Alkhateeb et al. Extending sparql with regular expression patterns (for querying rdf). *Journal of web semantics*, Vol. 7, No. 2, pp. 57–73, 2009.
[9] George HL Fletcher et al. Efficient regular path query evaluation using path indexes. 2016.
[10] André Koschmieder et al. Regular path queries on large graphs. In *Scientific and Statistical Database Management: 24th International Conference, SSDBM 2012, Chania, Crete, Greece, June 25-27, 2012. Proceedings 24*, pp. 177–194. Springer, 2012.
[11] Kento Miura et al. An fpga-based accelerator for regular path queries over edge-labeled graphs. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 415–422. IEEE, 2022.
[12] Sidler David et al. Accelerating pattern matching queries in hybrid cpu-fpga architectures. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 403–415, 2017.
[13] Paolo Massa et al. Bowling alone and trust decline in social network sites. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pp. 658–663. IEEE, 2009.
[14] Jie Tang et al. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990–998, 2008.
[15] Srijan Kumar et al. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 933–943. International World Wide Web Conferences Steering Committee, 2018.
[16] dblp: computer scinece bibliography. `https://dblp.org/`.
[17] reddit: the front page of the internet, December, 2020.

量子回路シミュレータの実行環境変化に伴う性能評価と特性比較

青木　望美† 山崎　雅文†† 平井　聡†† 山岡　茉莉†† 福本　尚人††
小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2 丁目 1 － 1
†† 富士通株式会社　〒 211–8588 神奈川県川崎市中原区上小田中 4 丁目 1 － 1
E-mail: †nozomi-a@ogl.is.ocha.ac.jp, ††{m.yamazaki,ahirai,yamaoka.mari,fukumoto.naoto}@fujitsu.com,
†††oguchi@is.ocha.ac.jp

あらまし　量子コンピュータは量子の性質を利用することで一部の問題に対して非常に高速な計算を実現可能として，現在も研究開発が盛んに行われている．しかし正確に動作する量子コンピュータの実現が期待されている中，量子コンピュータ上で動作するアプリケーションの開発などを同時進行で行うことが理想的である．量子回路シミュレータは従来型コンピュータ上で量子コンピュータの挙動を表現しているツールであり，量子コンピュータ実機上で動作する量子アプリの開発等にも貢献性が高い．本論文では数ある量子回路シミュレータの中でも Qiskit Aer シミュレータと Qulacs シミュレータについて，複数種の単一サーバ上における性能分析を行うとともに両者の計算機環境における量子回路シミュレータの性能比較を行う．

キーワード　性能分析，アプリ分析，プロファイラ，量子回路シミュレータ，HPC

1　はじめに

近年注目が集まっている量子コンピュータは一部の問題に対して従来型コンピュータと比べて非常に高速な計算を実現することができる可能性があるとされており，様々な機関で盛んに研究・開発がなされている．従来型コンピュータ上では情報の最小単位であるビットを 0 または 1 のどちらかの状態として表すが，量子コンピュータでは量子の持つ性質を利用して 0 と 1 が重なり合った状態を保持することができる量子ビットを扱うことが大きな特徴であり，この量子ビットを用い，量子ゲートなどを適切に組み合わせることで計算の解が出力される確率を高めて答えを導き出す．また，このとき複数通りの計算を同時に行うことと同様の挙動をするため高速な計算を実現することができるとされる（図 1 参照）．量子コンピュータにより高速化が期待されている計算問題の一例として，例えば量子化学計算などが挙げられる．しかし，現時点でもすでに複数の量子コンピュータが存在するものの，いずれもノイズの影響で計算結果にエラーが含まれるなど完全な挙動をするものとは言い難い．実際に理想的な挙動をする量子コンピュータが実現するまで，数年から数十年程度必要と言われている．

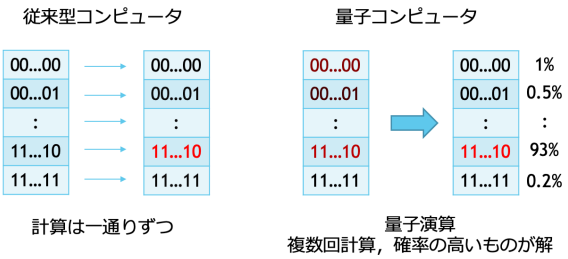


図 1　量子コンピュータの挙動

そして現在，従来型コンピュータ上で量子コンピュータの挙動を表現してシミュレートすることが可能な量子回路シミュレータが存在する．量子回路シミュレータを用いると計算結果は常に正確なものを得ることが可能である．量子コンピュータの実現まで数十年かかる可能性がある中で，量子コンピュータの開発と並行して量子回路シミュレータを使用して量子アプリ開発などを行うことは有用であり，このことから量子回路シミュレータに関する研究を行うことは有用性が高い．現在も IBM が開発した Qiskit Aer [1] や国立大学法人大阪大学と株式会社 Qunasys が共同開発した Qulacs [2]，Google が開発した Cirq [3] などの量子回路シミュレータが存在し，各シミュレータのユーザも多く有用であるが，それぞれの性能特性の違いは明らかになっていない．

そこで本研究では，複数の量子回路シミュレータを使用した複数の量子回路について性能分析を行い，性能特性の違いを明らかにする．前報 [4] では一種類のサーバ上で量子回路シミュレータの性能分析を行なった．本稿では性能が異なる複数種の実験用サーバ上でそれぞれ量子回路シミュレータを動作させ，それと同時にシステム稼働状況を収集し，実験環境の違いによる特性の違いを明らかにする．さらに収集したデータを用いて分析対象の量子回路シミュレータの動作改善方法について検討する．

2　関連研究

量子回路シミュレータには複数の方式が存在し，各方式の様々な量子回路シミュレータが存在する．その中でも State Vector 方式は，すべての量子状態をメモリに保持するため状態ベクトルをまとめて取得可能であり，計算量が量子回路の形状の影響を受けにくい．また，このような構造から動作途中の量子状態

を得ることも可能であるため，例えば量子アプリのデバッグ作業に有用である．Tensor Network 方式では量子回路をテンソルネットワークとして表す．また，シミュレートにあたって必要な量子状態のみを計算する．Density Matrix 方式では，量子状態を行列として表現しつつシミュレートを行う．しかしこのとき，必要なメモリ量は非常に多くなる．

例えば State Vector 方式の量子回路シミュレータに，Intel-QS がある [5]．Intel-QS は CPU ベースの分散型 State Vector 方式シミュレータであり，マルチコア CPU での使用のために OpenMP でマルチスレッド化されている．また，ひとつの大規模な回路だけでなく，複数の小/中規模回路のシミュレートも可能である．さらに Tensor Network 方式の量子回路シミュレータの例として，例えば TeNeS [6] が挙げられる．TeNeS は多体量子状態計算を行うためのオープンソースプログラムパッケージであり，OpenMP と MPI ハイブリッド並列に対応している．

続いて量子回路シミュレータの性能向上に関する関連研究を示す．今村らによって開発された mpiQulacs [7] は，State Vector 方式の量子回路シミュレータのひとつである Qulacs [2] を，A64FX CPU に対して最適化された量子回路シミュレータである．MPI を使用することで並列分散処理に対応している．演算を行う際は，CPU に搭載されている複数の演算を一度にまとめて実行することができる命令を使用することでメモリバンド幅を引き上げている．また，状態ベクトルを複数に分割し，各プロセスで均等に保持してローカルで計算を行い，他プロセスと状態ベクトルを共有する時のみ MPI 通信を使用することで並列処理を実現している．この他にも高速化のための工夫を複数行うことで，数十ノード単位の大規模シミュレーションに関して，他の量子回路シミュレータと比較して性能が高くなったことを確認している．

3 量子回路シミュレータ

前章でも述べたように，量子回路シミュレータには State Vector 方式と Tensor Network 方式など複数の方式が存在する．その中でも State Vector 方式は前述のとおり途中の量子状態を取得することが可能なため量子アプリ等の開発における有用性が高いという点から，今回の実験では State Vector 方式のシミュレータについて分析を行う．以下で分析に用いる量子回路シミュレータについて紹介する．

3.1 Qiskit Aer

Qiskit Aer [1] は量子コンピュータ用のフレームワークである Qiskit [8] に含まれる量子回路シミュレータである．Qiskit では量子回路を構築して実際の量子デバイス上で実行することが可能であり，モジュール構造のため最適化などの処理も行いやすく設計されている．さらに従来型コンピュータ上でも量子回路をシミュレートすることが可能となっており，量子回路シミュレータの方式は State Vector 方式と Tensor Network 方式の両方に対応している．

3.2 Qulacs

Qulacs [2] は国立大学法人大阪大学と株式会社 Qunasys によって開発された State Vector 方式の量子回路シミュレータである．主要部分は性能を重視しながら C++で実装されている．また使用時は Python インタフェースを介して使用することができるため，高速なシミュレーションと高い利便性を実現している．

さらに本稿では前章で紹介した mpiQulacs も実験対象シミュレータとして使用する．

4 性能分析ツール

本稿で使用した性能分析ツールについて説明する．

4.1 mpstat

mpstat は，CPU 使用率やハードウェア割り込みなどの情報を取得することが可能なコマンドである．通常各カテゴリについて，すべての CPU の平均値と各 CPU ごとの統計情報を得ることができる．

mpstat を用いることで各コアの挙動を詳細に確認することができるため，I/O 競合発生の可能性を発見することができるなどボトルネックの要因特定に有用である．本稿では特にユーザ時間割合を抜粋して紹介する．

4.2 Linux Perf

Linux Perf [9] は，主に Linux Kernel のイベントデータ収集，トレース機能を提供しているツールである．Linux Perf にはいくつかのサブコマンドが存在し，Linux Perf を使用する際は通常これらサブコマンドを併用する．

今回の実験では stat というサブコマンドを使用する．perf stat は，あるプログラムの実行中に指定したイベント値を出力することなどが可能である．取得することができるデータに，例えばページフォールトの回数やキャッシュヒット/ミス数などがある．

4.3 PCM Tools

PCM Tools [10] は CPU のリソース利用に関する情報を取得することができるツールである．PCM には複数のコマンドが含まれており，例えば pcm-memory を用いるとメモリバンド幅のモニタリングが可能である．実際に計測する際は必要に応じてこのようなコマンドを使用する．

なお，本稿では Primergy RX2540 M1 サーバにおけるデータ収集にてこの pcm-memory を用いている．

5 実　　験

5.1 実験概要

本稿では，Qiskit Aer と Qulacs それぞれのライブラリを使用した量子回路シミュレータの Python プログラムを実行し，実行中のシステム稼働状況を複数の性能分析ツールを用いて収集する．

使用する量子回路シミュレータのバージョンを表 1 に示す．

表 1　量子回路シミュレータのバージョン

qiskit	0.39.4
qiskit aer	0.11.2
qulacs	0.6.3
mpiQulacs	1.3.1

量子体積 (QV: Quantum Volume) は，量子コンピュータのシステムを評価する目的で，どの程度複雑な処理が可能かを示す指標として提案 [11] された．量子体積モデル回路はこの量子体積を評価するためのものであり，ランダムな組み合わせの量子ビットのペアに 2 量子ゲートを作用させる．以降この回路を QV 回路と記述する．

今回の実験には，Qiskit Aer，Qulacs，mpiQulacs それぞれを使用した各 Python プログラムに，この QV 回路を用いた．

測定時の量子ビット数は 30，回路の深さは 10 に設定した．

プログラム実行中の性能データ収集にあたり，bash スクリプトを使用した．動作は性能分析ツールの実行を開始した後シミュレーションプログラムの実行を開始し，シミュレーション実行が終わると性能分析ツールによるデータ収集も終了する形にした．また性能分析ツールの設定について，mpstat と vmstat は fflush() を用いることでデータ収集の間もバッファのフラッシュを行う．

また，Primergy RX2540 M1 サーバにおいて perf stat を使用して収集するイベントデータは以下に示すイベントに関して収集を行う．

- コアサイクル数，実行命令数，L1D キャッシュミス数，L1D キャッシュ参照数
- コアサイクル数，実行命令数，L2 キャッシュミス数，L2 キャッシュ参照数
- コアサイクル数，実行命令数，L3 キャッシュミス数，L3 キャッシュ参照数

さらに，同じく実験に使用する FX700 サーバにおいては以下に示すイベントに関してデータ収集を行う．

- コアサイクル数，実行命令数，L1D キャッシュミス数，L1D キャッシュ参照数，L2 キャッシュ参照数，L2 キャッシュミス数

そして Primergy RX2540 M1 サーバにおけるメモリバンド幅データ収集については PCM Tools の pcm-memory を用いる．また，今回使用する性能分析ツールは，すべて 1 秒間隔でデータの収集を行う．

5.2 実験環境

2 つの実験環境をそれぞれ表 2 と表 3 に示す．

以降，Primergy RX2540 M1 を RX サーバと記述する．また，mpiQulacs は A64FX CPU に対して最適化されたシミュレータであり，そのために FX700 サーバ上でのみ性能分析を行う．

表 2　実験環境 1

実験用サーバ	Primergy RX2540 M1
OS	Rocky Linux 8.6
CPU	Intel Xeon プロセッサー E5-2697v3(2.60GHz)
CPU コア数	14
L1d/L2/L3 キャッシュ	32 KB(core)/256 KB(core)/35 MB
メモリ	DDR4(2133) 128 GB

表 3　実験環境 2

実験用サーバ	FX700
OS	Rocky Linux 8.5
CPU	Fujitsu A64FX (2.0GHz)
CPU コア数	48
L1d/L2 キャッシュ	64KB(core)/8MBx4(CMG)
メモリ	HBM2 32 GB

5.3 実験結果

5.3.1 実行時間

はじめに量子回路の実行時間について表 4 に結果を示す．この表から，Qiskit Aer シミュレータと Qulacs シミュレータの比較において，今回の実験環境どちらも Qiskit Aer シミュレータと比べて Qulacs シミュレータの方が十数秒から数十秒程度量子回路の実行時間が短いことがわかる．そして FX700 サーバ環境において，mpiQulacs は Qulacs シミュレータよりさらに 30 秒程度回路実行時間が短いことがわかる．

表 4　量子回路実行時間（秒）

	RX	FX700
Qiskit Aer	129.1	77.5
Qulacs	119.0	49.5
mpiQulacs	-	19.5

さらに今回使用したシミュレーションプログラム内における処理の実行時間内訳を以下の図 2，図 3，図 4，図 5，図 6 にそれぞれ示す．なお，各グラフの右に処理内容を記述しており，プログラム中では記述されている処理内容のうち上に書かれた処理から順に実行している．

これらの図から Qiskit Aer シミュレータでは RX サーバ，FX700 サーバ共にサンプリング処理を含めた回路実行にかかる時間が最も長いことがわかる．同時に FX700 サーバ上では Qiskit ライブラリのインポートに 8 %程度の時間がかかっており，比較的目を引く結果であった．また，Qulacs シミュレータについては RX サーバ上では回路実行時間が最も長く，FX700 サーバ上ではサンプリング処理の時間が最も長いことがわかる．さらに mpiQulacs シミュレータについては，回路実行時間とサンプリング処理が実行時間のほとんどを占めており，各処理の時間割合の差は 2 %程度であった．同時に，Qulacs ライブラリ以外のライブラリのインポートに 5 %程度の時間を使用しており，こちらも比較的目を引く結果となった．

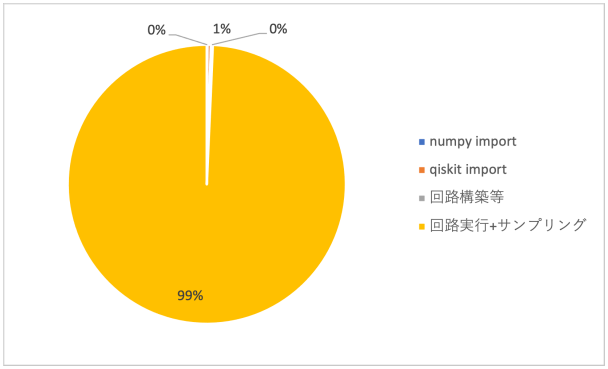


図 2　RX サーバ Qiskit Aer 実行時間内訳

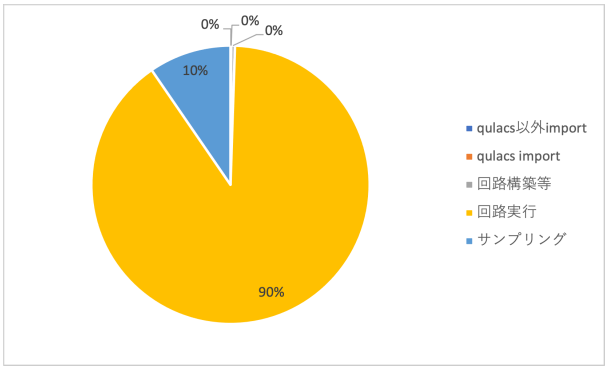


図 3　RX サーバ Qulacs 実行時間内訳

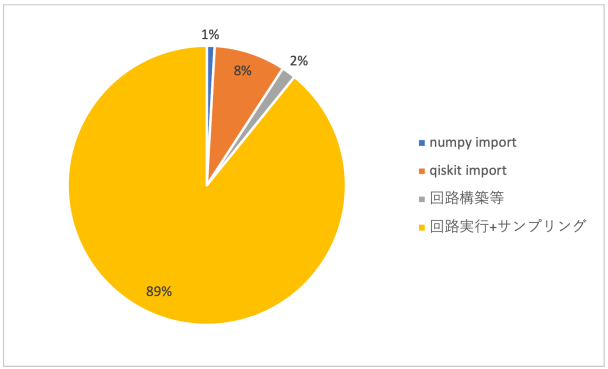


図 4　FX700 サーバ Qiskit Aer 実行時間内訳

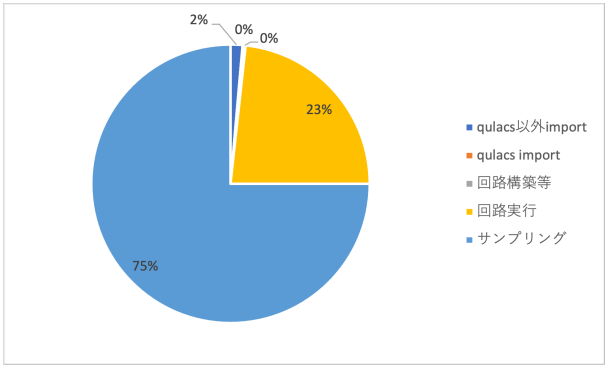


図 5　FX700 サーバ Qulacs 実行時間内訳

5.3.2　mpstat

続いて mpstat の結果について記述する．

mpstat で取得したプログラム実行中のユーザ時間割合の全コア平均値のグラフを図 7 に示す．

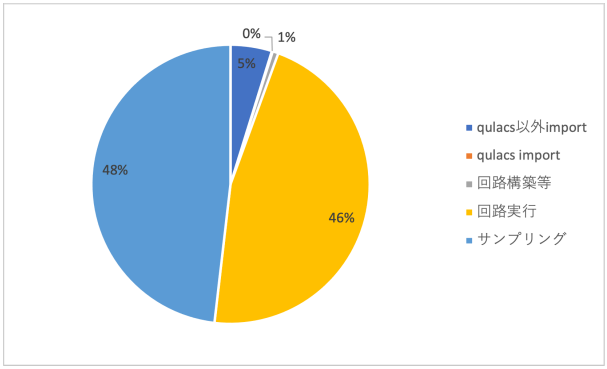


図 6　FX700 サーバ mpiQulacs 実行時間内訳

このグラフから，Qiskit の QV 回路は実行中ユーザ時間割合を約 100 %で保っていることがわかる．また，Qulacs QV 回路は，RX サーバ上では基本的に 100 %程度の値を保っているものの，終盤に数秒間ユーザ時間割合が低下した．また FX700 サーバ上では実行開始から 10 秒程度 100 %となっているが，その後著しく割合が低下した．Qulacs QV 回路のこれら現象については，図 3 と図 5 の結果からサンプリング処理が関係していると考えられる．さらに mpiQulacs は最大ユーザ時間割合が 90 %程度であり，図 6 よりこの部分は量子回路実行に対応，その後著しく値が低下する部分はサンプリング処理に対応していると考えられる．

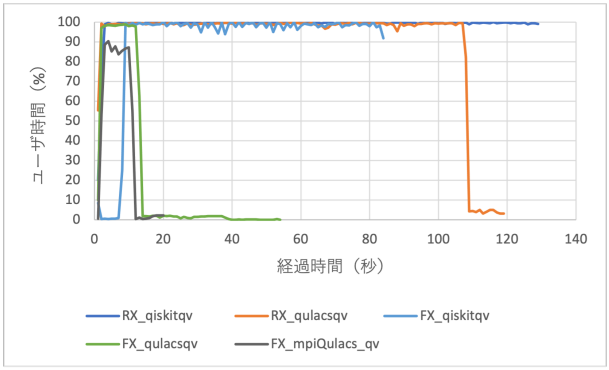


図 7　ユーザ時間割合（%）

5.3.3　Linux Perf

続いて Linux Perf における分析結果について記す．今回，サブコマンドとして stat を使用してキャッシュミス率等を取得した．

各環境におけるキャッシュミス率は以下の通りである．表 5 に RX サーバにおける L1D，L2，L3 キャッシュミス率を示す．さらに表 6 に FX700 サーバ上における L1D，L2 キャッシュミス率を示す．

また，図 8 に RX サーバにおける L1D キャッシュミス率，図 9 に L2 キャッシュミス率，図 10 に L3 キャッシュミス率を示す．さらに，図 11 に FX700 サーバにおける L1D キャッシュミス率，図 12 に FX700 サーバにおける L2 キャッシュミス率をそれぞれ示す．

表 5　RX サーバ キャッシュミス率（%）

	L1D	L2	L3
Qiskit Aer	1.5	45.9	67.9
Qulacs	2.7	54.3	38.4
mpiQulacs	-	-	-

表 6　FX700 サーバ キャッシュミス率（%）

	L1D	L2
Qiskit Aer	1.5	29.6
Qulacs	2.8	34.2
mpiQulacs	3.8	56.1

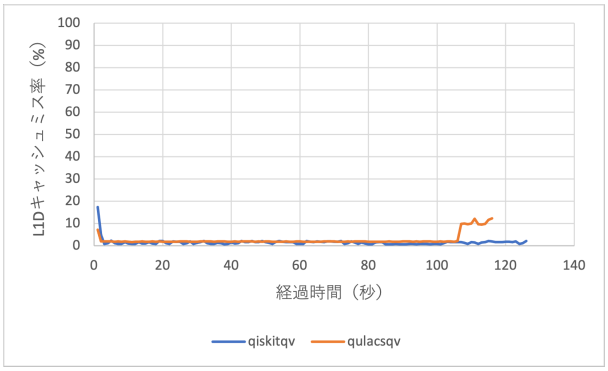


図 8　RX サーバ L1D キャッシュミス率（%）

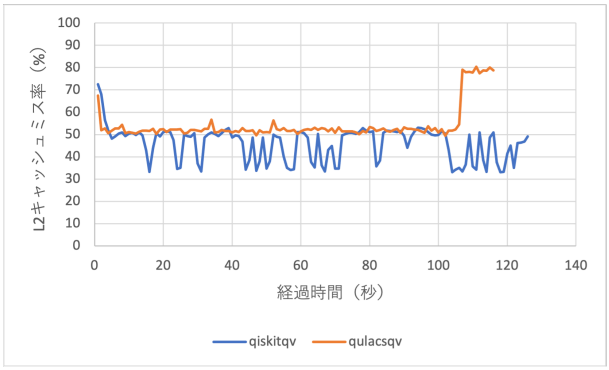


図 9　RX サーバ L2 キャッシュミス率（%）

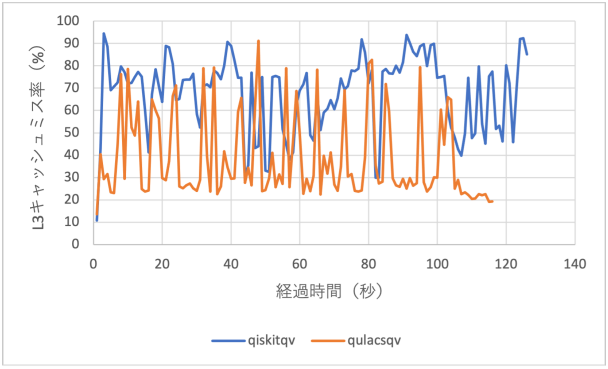


図 10　RX サーバ L3 キャッシュミス率（%）

表 5 から，L1D，L2 キャッシュミス率は Qulacs の方が高いが，L3 キャッシュミス率は Qiskit Aer より Qulacs の方が 43.4 %程度低い値となっていることがわかる．

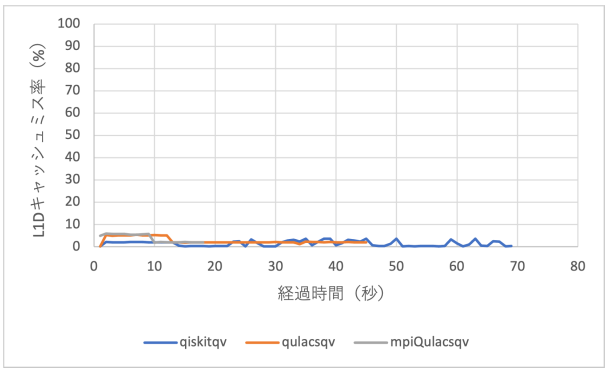


図 11　FX700 サーバ L1D キャッシュミス率（%）

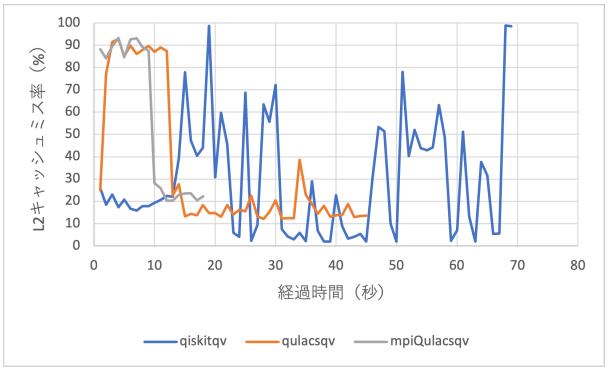


図 12　FX700 サーバ L2 キャッシュミス率（%）

さらに表 6 から，L1D キャッシュミス率，L2 キャッシュミス率の各項目についてすべて Qiskit Aer シミュレータよりも Qulacs シミュレータの方がキャッシュミス率が高く，そして Qulacs シミュレータよりも mpiQulacs の方がさらにキャッシュミス率が高い結果となった．

5.3.4　メモリバンド幅

続いて各サーバ上におけるメモリバンド幅計測結果について記述する．RX サーバのメモリバンド幅理論値は 68 GB/s，FX700 サーバにおけるメモリバンド幅理論値は 1024 GB/s である．

はじめに RX サーバについて，前述の通りメモリバンド幅の計測は PCM Tools のサブコマンドである pcm-memory を使用した．図 13 に結果のグラフを示す．このグラフから，プログラム実行中 Qulacs シミュレータは Qiskit Aer シミュレータと比べてメモリバンド幅を引き出していることがわかる．

続いて FX700 サーバ上におけるメモリバンド幅を図 14 に示す．このグラフから，Qulacs シミュレータは実行開始から 10 秒程度に渡りメモリバンド幅を 500 GB/s 近くまで引き出していることがわかる．また mpiQulacs も同様に実行開始から 10 秒程度 600 GB/s 近いメモリバンド幅を引き出していた．Qulacs シミュレータと mpiQulacs について，メモリバンド幅を高水準で引き出しているときの処理については，図 5 と図 6 から量子回路実行処理中であると考察される．

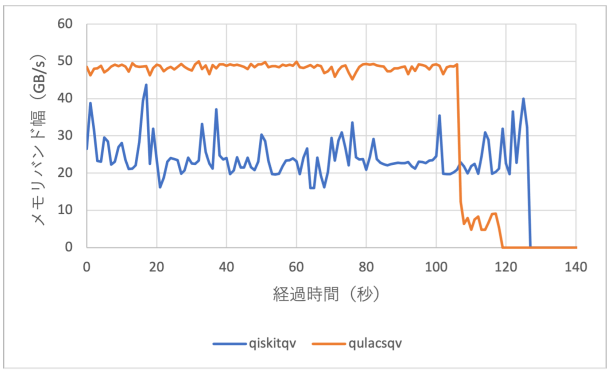


図 13　RX サーバ メモリバンド幅（GB/s）

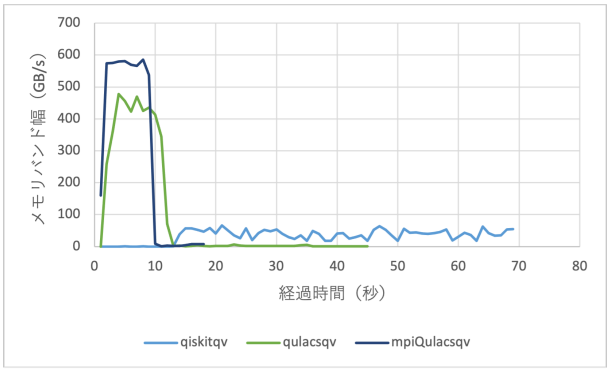


図 14　FX700 サーバ メモリバンド幅（GB/s）

6　まとめと今後の課題

本稿では異なる環境下で複数の量子回路シミュレータの性能データ収集を行い，性能の分析を行った．

その結果，量子回路の実行時間は両環境ともに Qiskit Aer シミュレータよりも Qulacs シミュレータの方が短い結果となった．さらに FX700 環境下において，mpiQulacs はこれら 2 つのシミュレータよりもさらに短い実行時間を記録した．そしてプログラム中の関数の実行時間割合の分析から，Qiskit Aer シミュレータは両環境において量子回路実行時間がその大半を占めていることがわかった．Qulacs シミュレータについては RX サーバ上では量子回路実行に消費する時間が比較的多く，FX700 サーバ環境下ではサンプリング処理に消費する時間が多いことがわかった．さらに mpiQulacs は量子回路実行とサンプリング処理にかかる時間割合が同程度であるとわかった．

また，mpstat の結果から Qiskit Aer シミュレータ実行時はプログラム実行中約 100 %のユーザ時間割合を保っていたことがわかるが，Qulacs シミュレータと mpiQulacs についてはサンプリング処理実行中と思われる瞬間著しく割合の低下が見られた．また，mpiQulacs の最大ユーザ時間割合は他 2 種類のシミュレータよりも 10 %程度低いことがわかった．

さらに perf stat を使用したキャッシュミス率の分析結果から，RX サーバでは L1D，L2 キャッシュミス率について Qiskit Aer シミュレータと比べて Qulacs シミュレータの方が高く，L3 キャッシュミス率は Qiskit Aer シミュレータの方が高い結果になった．また，FX700 サーバは L1D，L2 キャッシュミス率共に Qulacs シミュレータの方が高く，さらに mpiQulacs は Qulacs シミュレータよりも両キャッシュミス率が高い結果となった．

またメモリバンド幅の分析から，両環境共に Qulacs シミュレータは Qiskit Aer シミュレータと比べて明確にメモリバンド幅を引き出しており，FX700 サーバ環境下では mpiQulacs が最もメモリバンド幅を引き出していることがわかった．

これらの結果から，キャッシュミス率を抑えたとしても実行時間に与える影響は比較的小さく，メモリバンド幅を引き出すことが実行時間の短縮につながる可能性があるという考察が前報 [4] と比べさらに強固になった．さらに FX700 サーバにおいて，Qulacs シミュレータと mpiQulacs の処理のうちユーザ時間割合が低下しているサンプリング処理についても，より効率的な CPU 使用をするよう改善することで性能が高まる可能性も考えられる．

今後もさらに大規模システムを使用するなどし，量子回路シミュレータについて詳細な分析を進める．そしてその結果を基にシミュレータの動作改善方法を模索し，実装して検証を重ねていく．

謝　　辞

本研究の一部はお茶の水女子大学と富士通株式会社との共同研究契約に基づくものであり，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Qiskit aer simulator. https://qiskit.org/documentation/tutorials/simulators/1_aer_provider.html.
[2] Qulacs ドキュメンテーション — qulacs ドキュメント. http://docs.qulacs.org/ja/latest/index.html.
[3] Cirq. https://quantumai.google/cirq.
[4] 青木 望美, 山崎 雅文, 平井 聡, 山岡 茉莉, 福本 尚人, and 小口 正人. 量子回路シミュレータの性能向上のための分析と考察. マルチメディア、分散、協調とモバイル *(DICOMO2023)* シンポジウム, 2023.
[5] Intel quantum simulator — intel qs documentation. https://intel-qs.readthedocs.io/en/docs/getting-started.html#.
[6] Tenes - pasums - 東京大学. https://www.pasums.issp.u-tokyo.ac.jp/tenes.
[7] Imamura Satoshi, Yamazaki Masafumi, Honda Takumi, Kasagi Akihiko, Tabuchi Akihiro, Nakao Hiroshi, Fukumoto Naoto, and Nakashima Kohta. mpiqulacs: A distributed quantum computer simulator for a64fx-based cluster systems. *Distributed, Parallel, and Cluster Computing*, 2022.
[8] Qiskit. https://qiskit.org.
[9] Perf wiki. https://perf.wiki.kernel.org/index.php/Main_Page.
[10] Intel® performance counter monitor - a better way to measure cpu utilization. https://www.intel.com/content/www/us/en/developer/articles/technical/performance-counter-monitor.html.
[11] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta. Validating quantum computers using randomized model circuits. *Phys. Rev. A*, Vol. 100:p.32328(online), 2019. DOI: 10.1103/PhysRevA.100.032328.

一般発表 | Track 2: ビッグデータ基盤技術・セキュリティ・プライバシ

2024年2月29日(木) 16:00 ~ 18:10 | T2-B オンライン（Zoom Events)

時系列データ分類・予測

座長:天笠 俊之(筑波大学)
コメンテータ:杉浦 健人(名古屋大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T2-B-6-01] 階層的に細分化したマルチレベルイベントタイプからなる長時間マルチレベルエピソード抽出手法

*橋本 一輝[1]、新谷 隆彦[1]、大森 匡[1]、藤田 秀之[1] (1. 電気通信大学大学院情報理工学研究科)

16:30 ~ 16:55
[T2-B-6-02] 動的モード分解を活用した高速将来予測アルゴリズム

*千原 直己[1,2]、松原 靖子[1]、藤原 廉[1,2]、櫻井 保志[1] (1. 大阪大学産業科学研究所、2. 大阪大学大学院情報科学研究科)

16:55 ~ 17:20
[T2-B-6-03] 学習可能な長さを持つshapeletsに基づく時系列分類法

*山口 晃広[1]、植野 研[1]、鹿島 久嗣[2] (1. 東芝研究開発センター 知能化システム研究所 システムAIラボラトリー、2. 京都大学 大学院情報学研究科 知能情報学専攻)

17:20 ~ 17:45
[T2-B-6-04] 合成シェイプレットを用いた時系列データ分類

*岡部 正幸[1] (1. 県立広島大学)

階層的に細分化したマルチレベルイベントタイプからなる長時間マルチレベルエピソード抽出手法

橋本　一輝† 　新谷　隆彦† 　大森　匡† 　藤田　秀之†

† 電気通信大学情報理工学研究科　〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †h2231112@edu.cc.uec.ac.jp, {shintani, omori, fujita}@is.uec.ac.jp

あらまし　我々は, 総継続時間を評価値とするユーティリティエピソードマイニングとして長時間エピソードマイニングの研究を行っている. 長時間エピソードマイニングではシーケンスデータからイベントタイプのパターンを抽出するが, 個々のイベントタイプは何が出現したかのみを表し, 何時に開始し, どれだけの時間継続したかなどの情報を持たない. 本研究では, 開始した時刻と継続した時間を考慮することによって, 近い時刻に開始し同程度の時間継続した長時間エピソードを抽出することを目的とし, イベントタイプの階層的な細分化と, 階層的に細分化したイベントタイプからなる長時間エピソードの抽出手法を提案する.

キーワード　エピソードマイニング, ユーティリティマイニング, データマイニング, 時系列データ, ライフログ

1　はじめに

近年，スマートウォッチなどのリストバンド型デバイスの普及により，日常生活のライフログデータを容易に収集できるようになった. 蓄積したライフログデータを利用し, 生活の分析や健康管理などを行うための研究が行われている [1]. 我々は, ライフログから生活特性として長時間エピソードを抽出する研究を行っている [2]. 長時間エピソードマイニングは, 一つのイベントシーケンスデータから価値のあるエピソードを抽出するユーティリティエピソードマイニング [3] の一種である. 長時間エピソードマイニングでは総継続時間を評価値とし, 長い時間を占めた行動に相当するエピソードを抽出する.

長時間エピソードマイニングの処理対象データであるイベントシーケンスデータは, イベントタイプ, 開始日時, 終了日時の組を開始日時の順に並べたリストである. エピソードマイニングは, イベントシーケンスデータに出現したイベントタイプの順序パターンをエピソードとして抽出するが, 各エピソードはイベントタイプの出現順のみを表す. エピソード抽出の際, イベントシーケンスデータでのイベントタイプの出現順が同じであれば, 開始した時刻や継続した時間の長さが大きく異なる場合でも同一のエピソードとして扱われてしまう. また, エピソードを構成する個々のイベントタイプは何が出現したかのみを表し, 何時に開始し, どれだけの時間継続したかなどの情報を持たない.

本研究は, 近い時刻に開始し, 同程度の時間継続した長時間エピソードを抽出することを目的とする. イベントタイプを開始時刻と継続時間によって様々な粒度で階層的に細分化したマルチレベルイベントタイプとし, 近い時刻に開始し, 同程度の時間継続したマルチレベルイベントタイプからなる長時間マルチレベルエピソードの抽出を検討する. 異なる階層のマルチイベントタイプからなるエピソードを探索するために探索候補が膨大となるため, 階層を考慮した探索候補の枝刈りによって効率良く長時間マルチレベルエピソードを抽出するアルゴリズムを提案する. そして, 実データを用いた評価実験により, 提案手法を評価する.

2　関連研究

時系列データを処理対象とするデータマイニング技術として, エピソードマイニングの研究が広く行われている [2–6, 8]. Manila らは, 最初に頻出エピソードマイニングを定義し, 抽出アルゴリズムとして, WINEPI, MINEPI を提案した [4]. Zhu らは極小非重複オカレンスに基づく頻出エピソードマイニングを定義し, シリアルな頻出エピソードを効率的に抽出するアルゴリズム MANEPI を提案した [5]. 人の生活において同時に複数の行動を行うことはないため, 本研究におけるエピソードは極小非重複オカレンスで評価する. 頻出エピソードマイニングは, エピソードを頻度のみで評価し, それぞれのイベントが持つ重みは考慮しない. Wu らはイベントの重みを考慮する一般化したエピソードマイニングとして, ユーティリティエピソードマイニングを定義し, ユーティリティエピソードの抽出アルゴリズムとして UP-Span を提案した [3].

我々は, ユーティリティエピソードマイニングの一種である長時間エピソードマイニングの研究を行っている [2]. 長時間エピソードマイニングは, 開始日時と終了日時を持つイベントからなるイベントシーケンスデータを扱い, 終了日時と開始日時の差によって計算される継続時間を評価値とすることで, イベントシーケンスデータの中で長い時間を占めるエピソードを抽出する. 長時間エピソードマイニングにおいて, 総継続時間はエピソード毎のオカレンス集合から計算される. この際, 各オカレンスの開始した時刻や継続した時間の長さが大きく異なる場合でも同一のエピソードの出現として扱う問題点がある. 長時間エピソードマイニングで抽出された各エピソードについて, オカレンス集合を開始時刻と継続時間を基準に分割し, 近い時刻に開始し同程度の期間継続した細分化エピソードを抽出する

手法 [6] が提案されているが, この手法で抽出されるエピソードは未だ開始時刻と継続時間のばらつきがあるイベントタイプによって構成される. 本研究では, イベントタイプを細分化し, 細分化したイベントタイプからエピソードを構成する.

Srikant らは, 初めに階層の概念をパターンマイニングに導入し, 階層を持つアイテムからなる頻出エピソードマイニングを定義し, その抽出アルゴリズムとして EstMerge を提案した [7]. EstMerge では, あるアイテムセットの支持度が最小支持度未満の時, その下位のアイテムセットの支持度も最小支持度未満になることを利用して探索候補を削減する. Ao らは, 階層を持つイベントタイプからなる頻出エピソードマイニングを定義し, 分散フレームワーク上で階層を考慮したエピソードを抽出するアルゴリズム LA-FEMH を提案した [8]. LA-FEMH は, イベントタイプのオカレンス毎に射影データベースを作成し, 複数のプロセッサで並列に探索することで, 処理時間を短縮するアルゴリズムであるが, 本研究では階層構造を利用した探索候補の枝刈りによって全体の処理量を削減する手法を提案する.

3 長時間エピソードマイニング

3.1 定　　義

従来の長時間エピソードマイニングの定義を示す [2]. イベントタイプの集合を $M = \{m_0, ..., m_l\}$ と表す. イベントシーケンスデータ $D = \langle d_1, ..., d_n \rangle$ は, イベントを開始日時の順に並べたリストである. それぞれのイベントは $d_i = (m_i, t_{s_i}, t_{e_i})$ と表され (ここで, $1 \leqq i \leqq n$), 日時 t_{s_i} から t_{e_i} までの間, イベントタイプ $m_i \in M$ が発生していたことを意味する. このとき, 開始日時と終了日時は, $t_{s_i} < t_{e_i} (1 \leqq i < n)$, $t_{e_i} \leqq t_{s_{i+1}} (1 \leqq i \leqq n-1)$ を満たし, それぞれのイベントは重複して発生しないとする. イベント d_i の継続時間は $t_{e_i} - t_{s_i}$ と定義される.

エピソード $\alpha = \langle m_1, \ldots, m_k \rangle$ はイベントタイプのリストであり, $m_i \in M (1 \leqq i \leqq k)$ である. エピソード $\alpha = \langle m_1, ..., m_k \rangle$ は, イベントシーケンスデータ D のサブシーケンスとして $D' = \langle (m_1, t_{s_1}, t_{e_1}), ..., (m_k, t_{s_k}, t_{e_k}) \rangle$ が存在し, 全ての $i (1 \leqq i \leqq k-1)$ に対して $t_{s_i} < t_{s_{i+1}}$ であるとき, $[t_{s_1}, t_{e_k})$ の区間で出現する. このときの α の出現区間をオカレンスと呼び, $occ(\alpha)$ と表す. $occ(\alpha)$ の継続時間は, $t_{e_k} - t_{s_1}$ となる.

長時間エピソードマイニングでは冗長なオカレンスを排除するための制約として, 一つのオカレンスの継続時間の上限値 $maxspan$ と, オカレンスの間隔の上限値 $maxgap$ が設定される. エピソード $\alpha = \langle m_1, ..., m_k \rangle$ のオカレンス $occ(\alpha) = [t_{s_1}, t_{e_k})$ は, $t_{e_k} - t_{s_1} \leqq maxspan$ のとき, $maxspan$ の制約を満たす. また, 全ての $i (1 \leqq i < k-1)$ に対して, m_i の終了日時と m_{i+1} の開始日時の間隔が $maxgap$ 以下であるとき, $maxgap$ の制約を満たす. エピソード α の, $maxspan$, $maxgap$ の制約を共に満たすオカレンスの集合は $OCC(\alpha)$ と表される.

長時間エピソードマイニングでは, 極小非重複オカレンスに基づきエピソードを評価する. $OCC(\alpha)$ のオカレンスについて, $occ(\alpha)$ が別の α のオカレンスを含まないとき, 極小であると呼び, 極小オカレンスを $mo(\alpha)$ と表す. ここで, $[t_s, t_e), [t'_s, t'_e) \in OCC(\alpha)$ について, $t_s \geqq t'_s \wedge t_e \leqq t'_e$ のとき, オカレンス $[t'_s, t'_e)$ はオカレンス $[t_s, t_e)$ を含む. α の極小オカレンスの集合を $MO(\alpha)$ と表す. また, オカレンス $[t_s, t_e), [t'_s, t'_e) \in OCC(\alpha)$ について, $t_e < t'_s \vee t'_e < t_s$ のとき, 2 つのオカレンスの関係を非重複と呼ぶ. 極小と非重複を共に満たす α のオカレンスを $mano(\alpha)$ と表し, α の極小非重複オカレンスの集合を $MANO(\alpha)$ と表す. 複数の極小オカレンスが重複する場合, 開始日時が最も早い極小オカレンスを優先的に極小非重複オカレンス集合に選択する.

エピソードは頻度と総継続時間を評価値として持つ. エピソード α の頻度は $freq(\alpha) = |MANO(\alpha)|$ と定義される. また, エピソード α の総継続時間は $tdur(\alpha) = \sum_{[t_s, t_e) \in MANO(\alpha)} (t_e - t_s)$ と定義される. エピソードの総継続時間がユーザー指定の総継続時間の最小値 $mintdur$ 以上であるとき, そのエピソードを長時間エピソードと呼ぶ.

長時間エピソードマイニングは, イベントシーケンスデータ D, 総継続時間の閾値 $mintdur$, 一つのオカレンスの継続時間の上限値 $maxspan$, オカレンスの中のイベントタイプの間隔の上限値 $maxgap$ が与えられたときに, 全ての長時間エピソードを抽出する問題である.

3.2 抽 出 手 法

従来の長時間エピソードマイニング [2] では, 既存のエピソードの末尾に一つのイベントタイプを繋げることでエピソードを成長させ, 全ての長時間エピソードを抽出する. 初めにイベントシーケンスデータを読み取り, 各イベントタイプ m を長さ 1 のエピソードとして極小オカレンス $MO(m)$ を作成する. その後, 既に作成したエピソード α の末尾に長さ 1 のエピソード m を結合することで新しいエピソード $\beta = \langle \alpha, m \rangle$ を作成する. 作成したエピソード β について, 極小オカレンス集合 $MO(\beta)$, 極小非重複オカレンス集合 $MANO(\beta)$, 総継続時間 $tdur(\beta)$ を計算し, 総継続時間が $mintdur$ を満たすときは, 長時間エピソードとして出力する.

長時間エピソードの抽出において, 全ての探索候補を調べると探索候補数が膨大になり処理負荷が高くなる. 頻出エピソードマイニング [4] では Apriori の法則を用いた枝刈りをするが, 長時間エピソードマイニングは頻度でなく総継続時間を閾値とするため, 同様の枝刈りはできない. エピソード α と, α を成長させて作られたエピソード β について, β の総継続時間が α の総継続時間より長くなる場合があり, α が $mintdur$ を満たさない場合にも β について調べる必要がある. 長時間エピソードマイニングでは, 候補数を減らすための手法として, $mintdur$ と $maxspan$ から頻度の下限を求め, その値を閾値とすることで候補エピソードの枝刈りをする方法が提案されている. あるエピソードが長時間エピソードであるとき, そのエピソードの頻度はそれぞれのオカレンスの継続時間が $maxspan$ であるときに最小値となる. これにより, 長時間エピソードでは下限頻度 $lowfreq = \lceil \frac{mintdur}{maxspan} \rceil$ を設定し, 頻度が $lowfreq$ 未満のエピソードと, そのエピソードを成長させて作られるエピソードを枝刈りする.

Algorithm 1 ExtractLongDurationEpisodes(D, $mintdur$, $maxspan$, $maxgap$)

Input: a sequence data D, a total duration threshold of an episode $mintdur$, a duration constraint of an occurrence $maxspan$, a gap constraint of an occurrence $maxgap$
Output: all long-duration episodes

1: FI :=All event type that satisfy $lowfreq$
2: Generate minimal occurrence list MO for each event type in FI
3: **for** each event type $h \in FI$ **do**
4: **for** each event type $m \in FI$ **do**
5: ExtendEpisode(h, m)
6: **end for**
7: **end for**

Algorithm 2 ExtendEpisode(α, m)

Input: episode α, event type m
Output: long-duration episode β

1: Generate episode β by appending m to α
2: $MO(\beta) = \emptyset$
3: **for** each occ $[oa_s, oa_e) \in MO(m)$ **do**
4: Extract the occurrence $[om_s, om_e) \in MO(m)$ such that $om_s < om'_s \wedge om_s \geqq oa_e \wedge om'_s > oa_e$ for any $[om'_s, om'_e) \in MO(m)$
5: **if** $(om_e - oa_s \leqq maxspan) \wedge (om_s - oa_e \leqq maxgap)$ **then**
6: Append $[o_s, o_e)$ to $MANO(\beta)$
7: **end if**
8: **end for**
9: Delete occ $[o_s, o_e)$ from $MO(\beta)$ such that $o_e == o'_e \wedge o_s > o'_s$ for any $[o_s, o_oe), [o'_s, o'_e) \in MO(\beta)$
10: $MANO(\beta) = \emptyset$
11: **for** each occ $[o_s, o_e) \in MO(\beta)$ **do**
12: **if** $o_s \geqq o'_e$ for any $[o'_s, o'_e) \in MANO(\beta)$ **then**
13: Append $[o_s, o_e)$ to $MANO(\beta)$
14: **end if**
15: **end for**
16: **if** $|MANO(\beta)| \geqq lowfreq$ **then**
17: $tdur(\beta) = \sum_{[o_s,o_e) \in MANO(\beta)} (o_e - o_s)$
18: **if** $tdur(\beta) \geqq mintdur$ **then**
19: Output β as long-duration episode
20: **end if**
21: **for** each event type $m' \in FI$ **do**
22: ExtendEpisode(β, m)
23: **end for**
24: **end if**

従来の長時間エピソード抽出アルゴリズムを Algorithm 1, Algorithm 2 に示す. ExtractLongDurationEpisodes では, 初めにイベントシーケンスデータを読み取り, $lowfreq$ を満たす全てのイベントタイプを抽出し, 抽出されたイベントタイプを長さ 1 のエピソードとして極小オカレンスのリストを作成する (1-2 行). その後, 作成した長さ 1 のエピソードを別の長さ 1 のエピソードと組み合わせ, ExtendEpisode を呼び出す (3-7 行).

ExtendEpisode では, 入力として与えられたエピソード α の末尾にイベントタイプ m を繋げることでエピソード β を作り (1 行), α の極小オカレンスのリストと m の極小オカレンスのリストから β の極小オカレンスのリスト $MO(\beta)$ を作成する (2-9 行). その後, $MO(\beta)$ から重複しているオカレンスを除くことで β の極小非重複オカレンスのリスト $MANO(\beta)$ を作成する (10-15 行). 作成された $MANO(\beta)$ の要素数が下限頻度 $lowfleq$ を満たしていれば (16 行), β の総継続時間を計算する (17 行). β の総継続時間が継続時間の閾値 $mintdur$ を満たすとき, β を長時間エピソードとして出力する (18-20 行). $lowfreq$ を満たすエピソードを成長させたエピソードは長時間エピソードになる可能性があるため, β が $lowfreq$ を満たすとき, β とイベントタイプ m' を入力に与えて ExtendEpisode を呼び出す (21-23 行).

3.3 問 題 点

長時間エピソードマイニングはイベントタイプの順序パターンを長時間エピソードとして出力する. このとき, エピソードを構成するイベントタイプはイベントシーケンスデータに現れるイベントタイプであるため, イベントタイプは何が行われたかを表すが, 何時にどれくらいの期間行われたかなどの時間に関する意味は持たない.

図 1 は, 報告者のライフログにおける, イベントタイプ「デスクワーク」の開始時刻と継続時間の分布である. 横軸をそのイベントが開始した時刻, 縦軸をそのイベントが継続した時間とし, 全てのイベントをプロットした. 図 1 において, オレンジの点は午前 7 時に開始し 6 分継続したイベントを表し, 赤色の点は午後 5 時に開始し 664 分継続したイベントを表す. 従来の長時間エピソードマイニングではイベントタイプを開始時刻や継

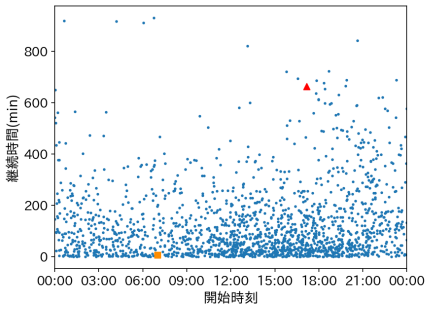


図 1: 「デスクワーク」の開始時刻と継続時間の分布

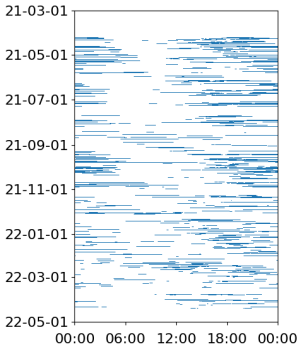


図 2: 「軽作業-デスクワーク」のオカレンス分布

続時間で区別せず, これらの開始時刻や継続時間が大きく異なるデスクワークを一つのイベントタイプとして扱う.

また, 報告者のライフログを入力として長時間エピソードマイニングを行った際に抽出された長時間エピソードの例として, 「軽作業-デスクワーク」のオカレンス分布を図 2 に示す. 図 2 において, 縦軸は日付を, 横軸は縦軸の日付における時刻を表し, 横棒が出現区間を表す. オカレンスの終了時刻が 24:00 を越えるとき, 24:00 を越えた区間は翌日の 00:00 に開始する区間として表す. 長時間エピソードマイニングにおいて, 抽出したエピソードは特定の行動を表すことが期待されるが, 図 2 の例では昼間に短時間継続したオカレンスと夜間に長時間継続したオカレンスが一つのエピソードに混在している.

4 長時間マルチレベルエピソードマイニング

4.1 長時間マルチレベルエピソードの抽出

従来の長時間エピソードマイニングでは, 開始した時刻や継続した時間が大きく異なるオカレンスが一つのエピソードとして扱われる場合がある. 近い時刻に開始し同程度の期間継続したオカレンスを持つエピソードを抽出する手法として, エピソードの細分化 [6] が提案されている. エピソードの細分化では, エピソードマイニングで抽出された各エピソードについて, エピソードのオカレンス集合を開始時刻と継続時間を基準に分割し, 分割されたオカレンス集合に一つのエピソードを割り当てることで細分化エピソードを抽出する. エピソードの細分化によって, 近い時刻に開始し同程度の期間継続した長時間エピソードを抽出できるが, この手法で抽出されるエピソードは開始時刻と継続時間のばらつきがあるイベントタイプによって構成される. 例えば, エピソードの細分化で「軽作業-デスクワーク」のオカレンス集合が, 5 時間程度継続したオカレンスの集合と 3 時間程度継続したオカレンスの集合に分けられたとき, 5 時間程度行った「軽作業-デスクワーク」のオカレンスと, 3 時間程度行った「軽作業-デスクワーク」のオカレンスを区別することができる. 一方で, 5 時間以上継続した「軽作業-デスクワーク」のオカレンス集合には, 4 時間程度の軽作業を行った後に 1 時間程度のデスクワークを行ったオカレンスと, 1 時間程度の軽作業を行った後に 4 時間程度のデスクワークを行ったオカレンスが混在してしまい, これらを区別することができない.

本研究では, 開始時刻と継続時間によってイベントタイプを細分化し, 細分化したイベントタイプからエピソードを構成することで, 近い時刻に開始し同程度の時間継続した長時間エピソードの抽出を図る.

4.2 イベントタイプの階層的な細分化

クラスタリングによって個々のイベントタイプを開始時刻と継続時間が類似したグループに分割するイベントタイプの細分化手法が検討された [9] が, エピソードマイニングを行うにあたりイベントタイプ毎の分割数を事前に決定できない問題がある. 分割数を過度に大きく設定した場合, 個々のイベントタイプの出現が少なくなり, そのイベントタイプを含むエピソードは長時間エピソードとして抽出されなくなる. 分割数を事前に決定できない中で, 長時間エピソードとして抽出される範囲で最も細かいイベントタイプからなるエピソードを抽出するため, 本研究では, 様々な粒度で細分化を行い, 細分化した全てのイベントタイプから候補エピソードを作成する.

様々な粒度で分割されたイベントタイプ全ての組み合わせから候補エピソードを作成する場合, 候補の数が増大し, 処理量が膨大になる. 階層構造を持つアイテムからなる頻出アイテムセットマイニング [7] において, 下位のアイテムが上位のアイテムの一部であること利用した枝刈りが提案されていることから, イベントタイプの細分化を階層的に行うことで長時間エピソードを抽出する際の処理量を削減できると考えた.

本研究では, イベントタイプの階層的な細分化手法の一種として, 二分割を繰り返すクラスタリング [10] を使用し, 階層的に様々な粒度のイベントタイプを作成する. イベントシーケンスデータ中のイベントタイプ毎に, 出現する区間の開始時刻と継続時間を特徴量としたクラスタリングによる二分割を繰り返し行い, 各分割で作成されたクラスタ毎に新しいイベントタイプを割り当てる. イベントタイプの階層構造を木で保存し (図 3), イベントシーケンスデータ中のイベントタイプを葉のイベントタイプで置換する. 長時間エピソードとして抽出される範囲で最も細かく分割されたイベントタイプからなるエピソードを抽出するため, 細分化されたイベントタイプの頻度が下限頻度以上である限り, 分割を繰り返し行う.

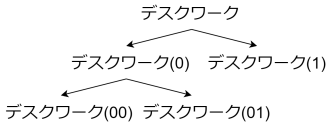


図 3: イベントタイプの細分化の例

図 1 のイベントタイプ「デスクワーク」に対し, 二分割を 2 回繰り返した結果を図 4 に示す. 一度目の分割で「デスクワーク (0)」と「デスクワーク (1)」が作成され, 二度目の分割でデスクワーク (0) のオカレンス集合を二分割することで赤のクラスタの「デスクワーク (00)」と青のクラスタの「デスクワーク (01)」が作成される. 図 4 の赤のクラスタでは, 200 分未満の継続時間を持つオカレンスが排除されており, 長く継続した「デ

スクワーク」が「デスクワーク (00)」として区別される.

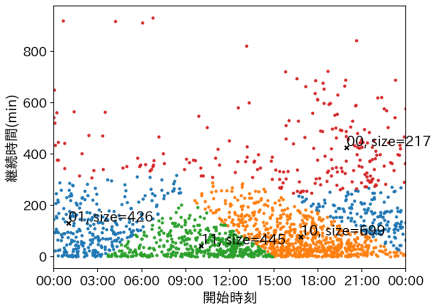


図 4: 「デスクワーク」レベル 2

4.3 定　　義

従来の長時間エピソードマイニング抽出問題を拡張し, 階層を持つイベントタイプにより構成される長時間エピソードの抽出問題を定義する.

階層を持つイベントタイプをマルチレベルイベントタイプと呼び, その集合を $M = \{m_0, ..., m_l\}$ と表す. 階層構造 T はマルチレベルイベントタイプを節点とする根つき木の集合であり, マルチレベルイベントタイプの階層構造を表す. マルチレベルイベントタイプ p が q の子であるとき, $p \Rightarrow q$ と表す. また, マルチレベルイベントタイプ p が q の子孫であるとき, $p \Rightarrow^* q$ と表す. $p \Rightarrow^* q$ であるとき, q は p の上位のマルチレベルイベントタイプであり, p は q の下位のマルチレベルイベントタイプである.

マルチレベルイベントシーケンスデータ $D = \langle d_1, ..., d_n \rangle$ は, マルチレベルイベントを開始日時の順に並べたリストである. それぞれのマルチレベルイベントを $d_i = (m_i, t_{s_i}, t_{e_i})$ と表し (ここで, $1 \leqq i \leqq n$), 日時 t_{s_i} から t_{e_i} までの間, 葉のマルチレベルイベントタイプ m_i と, m_i の先祖が発生していたことを意味する.

マルチレベルイベントタイプのリスト $\alpha = \langle m_1, ..., m_k \rangle$(ここで, $m_i \in M(1 \leqq i \leqq k)$) をマルチレベルエピソードと定義する. マルチレベルエピソード $\alpha = \langle m_1, ..., m_k \rangle$ は, マルチレベルイベントシーケンスデータ D のサブシーケンスとして $D' = \langle (m'_1, t_{s_1}, t_{e_1}), ..., (m'_k, t_{s_k}, t_{e_k}) \rangle$ が存在し, 全ての $i(1 \leqq i \leqq k-1)$ に対して $t_{s_i} < t_{s_{i+1}}$ であり, 全ての $i(1 \leqq i \leqq k)$ に対して $m'_i \Rightarrow^* m_i$ であるとき, $[t_{s_1}, t_{e_k})$ の区間で出現する.

マルチレベルエピソードの総継続時間がユーザー指定の総継続時間の最小値 $mintdur$ 以上であるとき, そのマルチレベルエピソードを長時間マルチレベルエピソードと呼ぶ.

長時間マルチレベルエピソードマイニングは, 総継続時間の閾値 $mintdur$, マルチレベルイベントシーケンスデータ D, 階層構造 T が与えられ, オカレンスの制約条件が $maxspan$, $maxgap$ のとき, 全ての長時間マルチレベルエピソードを抽出する問題である.

Algorithm 3 BasicMethod(D, T, $mintdur$, $maxspan$, $maxgap$)

Input: a sequence data D, a taxonomy T, a total duration threshold of an episode $mintdur$, a duration constraint of an occurrence $maxspan$, a gap constraint of an occurrence $maxgap$

Output: all long-duration multi-level episodes

```
1: FI :=All multi-level event type that satisfy lowfreq
2: Generate minimal occurrence list MO for each multi-level event type in FI
3: for each multi-level event type h ∈ FI do
4:    for each multi-level event type m ∈ FI do
5:       ExtendEpisode(h, m)
6:    end for
7: end for
```

5 長時間マルチレベルエピソード抽出手法

5.1 基 本 手 法

従来の長時間エピソード抽出アルゴリズムを拡張して全ての長時間マルチレベルエピソードを抽出する手法を基本手法とする. マルチレベルイベントシーケンスデータに表れるマルチレベルイベントタイプと, そのすべての先祖のマルチレベルイベントタイプの極小オカレンス集合を作成することで, 全ての長時間マルチレベルエピソードを抽出することができる.

基本手法のアルゴリズムを Algorithm 3 に示す. 基本手法では, 初めにマルチレベルイベントシーケンスデータ D を読み取り, 極小オカレンス集合の要素数が下限頻度 $lowfreq$ 以上となるマルチレベルイベントタイプの集合 FI を作成し, 極小オカレンス集合を保持する. 従来の長時間エピソードマイニング抽出アルゴリズムでは, イベントシーケンスデータに直接現れるイベントのみを FI として保持するが, 基本手法では, マルチレベルイベントシーケンスデータ中に直接表れない上位のマルチレベルイベントタイプも FI に保持する. その後, 全ての $h, m \in FI$ に対し, Algorithm2 に示した ExtendEpisode(h, m) を呼び出し, エピソードを成長させながら全ての長時間マルチレベルエピソードを探索する.

5.2 提 案 手 法

基本手法では, 階層を考慮せずに全てのマルチレベルイベントタイプを探索済みエピソードに結合するため, 探索候補数が膨大になる. 提案手法では, 深さ優先でエピソードを探索する際, 上位のマルチレベルイベントタイプにより構成されるエピソードを優先的に探索することで下位のマルチレベルイベントタイプにより構成されるエピソードの冗長な探索を回避する.

マルチレベルイベントタイプ c が m の子孫であるとき, $MO(m) \supset MO(c)$ となる. そのため, エピソード $\langle \alpha, c \rangle$ の

Algorithm 4 ProposedMethod(D, T, $mintdur$, $maxspan$, $maxgap$)

Input: a sequence data D, a taxonomy T, a total duration threshold of an episode $mintdur$, a duration constraint of an occurrence $maxspan$, a gap constraint of an occurrence $maxgap$

Output: all long-duration multi-level episodes

1: FI :=All multi-level event type that satisfy $lowfreq$
2: Generate minimal occurrence list MO for each event type in FI
3: FI^0 :=All root event types in FI
4: **for** each multi-level event type $m \in FI$ **do**
5: **for** each root event type $r \in FI^0$ **do**
6: ExtendEpisode$'(m, r)$
7: **end for**
8: **end for**

任意の極小オカレンス $[t_s, t_e)$ に対して, $t'_e \leqq t_e$ となるエピソード $\langle \alpha, m \rangle$ の極小オカレンス $[t_s, t'_e)$ が存在し, 常に $|MANO(\langle \alpha, m \rangle)| \geqq |MANO(\langle \alpha, c \rangle)|$ が成立する. よって, $\langle \alpha, m \rangle$ が下限頻度を満たさないとき, 末尾のマルチレベルイベントタイプをその子孫に置き換えたエピソード $\langle \alpha, c \rangle$ は長時間エピソードにならないことが確定し, 枝刈りすることができる.

提案手法では, エピソード $\beta = \langle \alpha, m \rangle$ について調べた後, β に根のイベントタイプ r を結合することで β を成長させたエピソード $\langle \beta, r \rangle$ を作成する. また, β の末尾のマルチレベルイベントタイプ m をその子イベントタイプ c に置き換えることで下位のマルチレベルイベントタイプにより構成されるエピソード $\langle \alpha, c \rangle$ を作成する. β の頻度が下限頻度未満のとき, $\langle \beta, r \rangle$ と $\langle \alpha, c \rangle$ を調べる処理を省略する. 提案手法のアルゴリズムを Algorithm 4, Algorithm 5 に示す.

基本手法から変更した点を説明する. 提案手法では, FI の作成後, 下限頻度を満たす根のマルチレベルイベントタイプの集合 FI^0 を保持する (3 行). 基本手法では, 全ての $h, m \in FI$ の組に対し ExtendEpisode の呼び出しを行うが, 提案手法では, $m \in FI$ と, $r \in FI^0$ の全ての組に対して, 根に近いマルチレベルイベントタイプを優先的に探索する ExtendEpisode$'(h, m)$ を呼び出す (4-8 行).

ExtendEpisode$'(\alpha, m)$ では, ExtendEpisode と同様に $MO(\beta)$, $MANO(\beta)$ を計算し長時間エピソードを出力する (1-20 行). β の頻度が $lowfreq$ 以上の時, 新しく作成したエピソード β と全ての $m' \in FI^0$ に対し ExtendEpisode$'(\beta, m)$ を呼び出す (21-23 行). また, 成長前のエピソード α と全ての m の子 $c \in FI$ に対し, ExtendEpisode$'(\alpha, c)$ を呼び出す (24-26 行). β が下限頻度を満たさない場合, これら ExtendEpisode$'$ の呼び出しを省略し, β を成長させたエピソードと, β の末尾のマルチレベルイベントタイプをその子に置き換えたエピソードの探索を回避する.

Algorithm 5 ExtendEpisode$'(\alpha, m)$

Input: episode α, multi-level event type m

Output: long-duration multi-level episode β

1: Generate episode β by appending m to α
2: $MO(\beta) = \emptyset$
3: **for** each occ $[oa_s, oa_e) \in MO(m)$ **do**
4: Extract the occurrence $[om_s, om_e) \in MO(m)$ such that $om_s < om'_s \wedge om_s \geqq oa_e \wedge om'_s > oa_e$ for any $[om'_s, om'_e) \in MO(m)$
5: **if** $(om_e - oa_s \leqq maxspan) \wedge (om_s - oa_e \leqq maxgap)$ **then**
6: Append $[o_s, o_e)$ to $MANO(\beta)$
7: **end if**
8: **end for**
9: Delete occ $[o_s, o_e)$ from $MO(\beta)$ such that $o_e == o'_e \wedge o_s > o'_s$ for any $[o_s, o_oe), [o'_s, o'_e) \in MO(\beta)$
10: $MANO(\beta) = \emptyset$
11: **for** each occ $[o_s, o_e) \in MO(\beta)$ **do**
12: **if** $o_s \geqq o'_e$ for any $[o'_s, o'_e) \in MANO(\beta)$ **then**
13: Append $[o_s, o_e)$ to $MANO(\beta)$
14: **end if**
15: **end for**
16: **if** $|MANO(\beta)| \geqq lowfreq$ **then**
17: $tdur(\beta) = \sum_{[o_s, o_e) \in MANO(\beta)} (o_e - o_s)$
18: **if** $tdur(\beta) \geqq mintdur$ **then**
19: Output β as long-duration episode
20: **end if**
21: **for** each root event type $r \in FI^0$ **do**
22: ExtendEpisode$'(\beta, r)$
23: **end for**
24: **for** each child event type c of m **do**
25: ExtendEpisode$'(\alpha, c)$
26: **end for**
27: **end if**

6 評価実験

提案手法の処理性能と, 細分化したイベントタイプからなる長時間マルチレベルエピソードの有用性を評価するため, 実データを用いた実験を行った.

6.1 データセット

本実験では, 報告者が収集した 372 日分の運動状態データと, オンラインストアのウェブサーバログ [11] からマルチレベルイベントシーケンスデータを作成した.

運動状態データは, リストバンド型センサで取得した, いつか

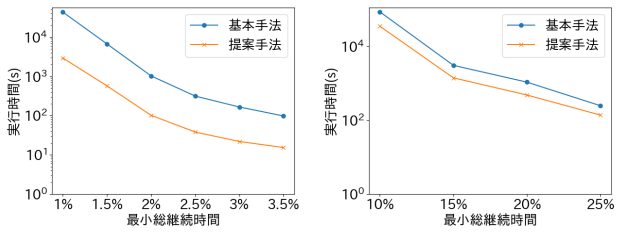
らいつまでどの程度の運動状態が継続したかを表すデータで, 静止, 安静など, 運動の程度を表す 8 種類の運動状態が開始日時, 終了日時と共に記録される.

本稿で提案した手法はライフログへ適応することを意図しているが, アルゴリズムの性能を評価するため, e コマースログとして, Chodak らにより公開された e コマースにおけるアクセスログ [11] も使用した. 長時間マルチレベルエピソードマイニングを行うため, 特定の IP からの GET リクエストのみを選択し, 次のリクエストまでの期間を継続時間とした. 同時刻に発生するイベントが複数存在する場合は最初に記録されたイベントのみを選択した. 継続時間が *maxspan* を超える場合は, そのイベントは排除した.

6.2 提案手法の性能評価

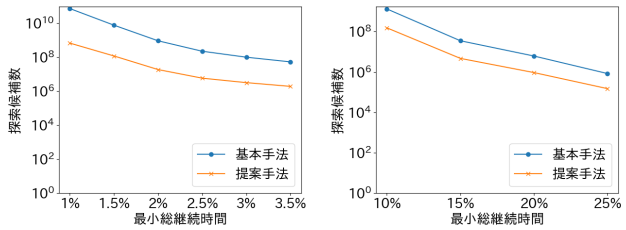
基本手法と提案手法を C 言語で実装し, 抽出アルゴリズムの性能評価を行った. Intel(R) Xeon(R) W-2235 3.80GHz プロセッサと 512GB のメインメモリを搭載し, CentOS Linux 7 を導入した計算機を実験に使用した. 実験は *maxspan* と *maxgap* を固定し, 最小総継続時間を変化させて行った. 報告者の運動状態データでは *maxspan* を 600 分, *maxgap* を 60 分とし, e コマースログでは *maxspan* を 30 分, *maxgap* を 10 分とした. 最小総継続時間はイベントシーケンスデータ全体の継続時間における割合で指定した. 前処理として, 指定した最小総継続時間毎に下限頻度までイベントタイプを細分化したマルチレベルイベントシーケンスデータを作成して長時間マルチレベルエピソードマイニングの入力とした.

異なる最小総継続時間での実行時間を図 5 に, 探索候補数を図 6 に示す. 図 5, 6 において, 横軸は最小総継続時間を表し, 縦軸はそれぞれ実行時間, 探索候補数を表す.



(a) 報告者の運動状態データ　　**(b)** e コマースログ

図 5: 基本手法と提案手法による実行時間



(a) 報告者の運動状態データ　　**(b)** e コマースログ

図 6: 基本手法と提案手法による探索候補数

図 5, 6 より, 両方のデータ, 全ての最小総継続時間で, 提案手法により実行時間, 探索候補数を削減できていることが分かる. 実験結果より, 提案手法によって探索候補数を削減し, 実行時間を短縮できることを確認した.

6.3 抽出された長時間マルチレベルエピソード

細分化したイベントタイプからなるエピソードの有用性を示す例として, 長時間マルチレベルエピソードマイニングにより新たに抽出された意味のあるパターンを報告する. 報告者の運動状態データを処理対象とし, *mintdur* を 3.5%(353 分/週), *maxspan* を 600 分, *maxgap* を 60 分としたときに抽出された長時間マルチレベルエピソードは 4958 件であった. 分割されていないマルチレベルイベントタイプのみで構成されるエピソードは 371 件であり, 92%以上の長時間マルチレベルエピソードは従来の長時間エピソードマイニングでは抽出されないエピソードであった.

抽出された 4958 個の長時間マルチレベルエピソードの内, 2 回以上分割されたマルチレベルイベントタイプのみで構成される全 18 個のエピソードを表 1 に示す. ここで, マルチレベルイベントタイプ「静止 (10)」は, 一度目の分割で 1 のクラスタに, 二度目の分割で 1 を分割して作られた 10 のクラスタに割り当てられたことを意味する. 表中に表れる, 「デスクワーク」と「軽作業」を分割して作成されたレベル 2 のマルチレベルイベントタイプの分布を図 4, 7 に示す. 図 4, 7 において, 横軸はそのイベントが開始した時刻を表し, 縦軸はそのイベントが継続した時間を表す. 色分けされた各点はクラスタ毎のオカレンスを表し, バツ印はクラスタの中心点を表す.

表 1: 2 回以上分割されたマルチレベルイベントタイプのみで構成される長時間マルチレベルエピソード

長時間マルチレベルエピソード	頻度	総継続時間 (分)
静止 (10)-デスクワーク (11)	85	23908
静止 (11)-デスクワーク (11)	46	22744
軽作業 (01)-デスクワーク (00)	50	21092
デスクワーク (01)-静止 (10)	60	21713
デスクワーク (10)-軽作業 (11)	243	21364
デスクワーク (00)-デスクワーク (01)	47	21333
デスクワーク (000)-デスクワーク (01)	46	20769
デスクワーク (01)-デスクワーク (01)	153	29536
デスクワーク (10)-デスクワーク (00)	59	26055
デスクワーク (10)-デスクワーク (000)	54	23256
デスクワーク (10)-デスクワーク (10)	349	40532
デスクワーク (10)-デスクワーク (10)-軽作業 (11)	137	18892
デスクワーク (10)-デスクワーク (10)-デスクワーク (10)	153	22837
デスクワーク (10)-デスクワーク (10)-デスクワーク (100)	134	20962
デスクワーク (10)-デスクワーク (100)	289	31977
デスクワーク (100)-デスクワーク (00)	45	19209
デスクワーク (101)-デスクワーク (10)	136	27739
デスクワーク (101)-デスクワーク (100)	100	21223

表 1 の 3 行目のエピソード「軽作業 (01)-デスクワーク (00)」は, 図 7 のオレンジのクラスタに属する, 夕方から朝にかけて開始した軽作業が行われた後, 図 4 の赤のクラスタに属する長時間のデスクワークが行われたことを表している. このエピソードは, 報告者が不定期に行う, 夜間, 長時間の PC 作業に相当する. このエピソードのオカレンス分布を図 8a に, 分割前のイベントタイプからなる従来の長時間エピソードで抽出されるエピソード「軽作業-デスクワーク」のオカレンス分布を図 8b に示す. 図 8a, 8b において, 縦軸は日付を, 横軸は縦軸の日付にお

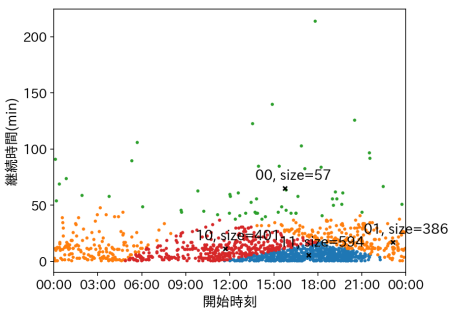
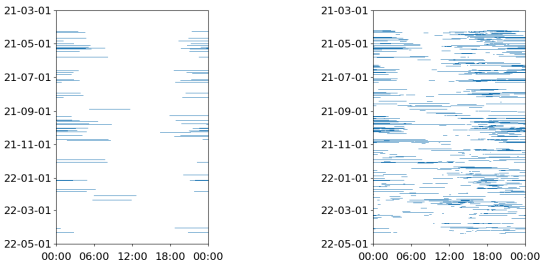


図 7: 「軽作業」レベル 2

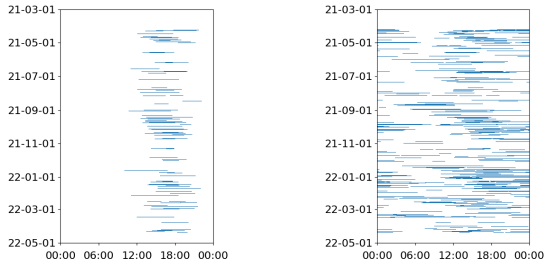
ける時刻を表し, 横棒が出現区間を表す. 従来の長時間エピソードマイニングでは, 軽作業の後にデスクワークを行う行動が長時間行われた行動であることを知ることができるが, 細分化したイベントタイプからなるエピソードの抽出により, 夜間に短時間の軽作業を行った後, 長時間のデスクワークを行う行動が長時間行われた行動であることを知ることができる. また, 図 8a のオカレンス分布では図 8b のオカレンス分布に存在する, 昼間のオカレンスや継続時間が短いオカレンスが排除されており, 開始時刻と継続時間のばらつきが小さいオカレンス集合をもつエピソードが抽出されていることが確認できる.



(a) マルチレベルエピソード **(b)** 従来のエピソード

図 8: 軽作業-デスクワークのオカレンス分布

また, 表 1 の 13 行目のエピソード「デスクワーク (10)-デスクワーク (10)-デスクワーク (10)」は, 図 4 のオレンジのクラスタに属する午後, 短時間のデスクワークが一時間以内の間隔で 3 回行われたことを表している. このエピソードは, 報告者の生活における, 午後に散発的に行う卓上作業に相当する. このエピソードのオカレンス分布を図 9a に, 分割前のイベントタイプからなるエピソード「デスクワーク-デスクワーク-デスクワーク」のオカレンス分布を図 9b に示す. 従来の長時間エピソードマイニングでは, デスクワークを連続で行う行動が長時間行われた行動であることを知ることができるが, 細分化したイベントタイプからなるエピソードの抽出により, 午後に短時間のデスクワークを連続して行う行動が長時間行われた行動であることを知ることができる. また, 図 9a より, エピソード「デスクワーク (10)-デスクワーク (10)-デスクワーク (10)」のオカレンス集合は, 午後に行われたオカレンスのみで構成されていることが確認できる.



(a) マルチレベルエピソード **(b)** 従来のエピソード

図 9: デスクワーク-デスクワーク-デスクワークのオカレンス分布

7 結　　論

本研究では, 近い時刻に開始し同程度の時間継続した長時間エピソードを抽出することを目的とした. クラスタリングによる二分割の繰り返しによって異なる粒度で階層的にイベントタイプを細分化し, 細分化したイベントタイプからなる長時間マルチレベルエピソードを抽出する手法を提案した. 長時間マルチレベルエピソード抽出において, イベントタイプの階層構造を利用する枝刈りによって探索候補数を削減する手法を提案した. 実データを用いた実験により, 提案アルゴリズムで探索候補数と実行時間を削減できることを示した. また, 長時間マルチレベルエピソードマイニングにより, 近い時刻に開始し同程度の時間継続するイベントタイプからなる, 従来の長時間エピソードマイニングでは抽出されないエピソードを抽出できることを示した.

今後の課題として, 冗長な長時間マルチレベルエピソードが大量に抽出される問題が挙げられる. 有用なエピソードを効率的に抽出するための手段として, 抽出されるエピソードに有用性を担保するための制約条件を与え, 条件を満たさない冗長な候補エピソードの探索を省略する方法が考えられる.

文　　献

[1] A. Ksibi, A. S. D. Alluhaidan, A.Salhi, and S. A. El-Rahman, "Overview of Lifelogging: Current Challenges and Advances," IEEE Access, 2021.
[2] T. Shintani, T. Ohmori and H. Fujita, "Method for Comparing Long-term Daily Life using Long- duration Episodes," DARLI-AP, 2019.
[3] C. Wu, Y. Lin, P. Yu, and V. Tseng, "Mining high utility episodes in complex event sequences," ACM SIGKDD, 2013.
[4] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovering frequent episodes in event sequences," KDD, 1997.
[5] H. Zhu, P. Wang, X. He, Y. Li, W. Wang, and B. Shi. "Efficient episode mining with minimal and non-overlapping occurrences," IEEE ICDM, 2010.
[6] 中山恭明, 新谷隆彦, 大森匡, 藤田秀之, "細分化エピソードを用いた生活比較手法に関する一考察," DEIM, 2019.
[7] R. Srikant and R.Agrawal, "Mining generalized association rules." VLDB, 1995.
[8] X. Ao, H. Shi, J. Wang, L. Zuo, H. Li, and Q. He, "Large-Scale Frequent Episode Mining from Complex Event Sequences with Hierarchies," ACM Transacctions on Intelligent Systems and Technology, 10(4), 36, 2019.
[9] 柳川 凜太朗, 新谷 隆彦, 大森 匡, 藤田 秀之, "リストバンド型センサで取得した運動状態データの開始時刻と継続時間を考慮した細分化の試み," FIT2021, 2021.
[10] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," ICML, 2000.
[11] G. Chodak, G. Suchacka, and Y. Chawla, "EClog: HTTP-level ecommerce data based on server access logs for an online store," Computer Networks, vol. 183, 2020. doi:10.7910/DVN/Z834IK

動的モード分解を活用した高速将来予測アルゴリズム

千原　直己[†,††]　松原　靖子[†]　藤原　廉[†,††]　櫻井　保志[†]

† 大阪大学産業科学研究所　〒 567–0047 大阪府茨木市
†† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市
E-mail: †{naoki88,yasuko,r-fujiwr88,yasushi}@sanken.osaka-u.ac.jp

あらまし　本論文では，動的モード分解を活用した大規模時系列データストリームの高速予測手法 ModeCast を提案する．ModeCast はセンサデータや，Web データ等，多種多様な時系列パターンにより構成される大規模時系列データストリームが与えられた時，その中から潜在的なダイナミクスに基づいた重要な時系列パターンを発見することで将来予測を行う．より具体的には，このようなパターンを発見するために動的モード分解（DMD）を活用する．提案手法は，(a) 大規模データストリームの中から重要なダイナミクスを発見し，リアルタイムかつ長期的な予測を可能とする．また，(b) 様々なデータに対して予測を行うことが可能であり，実用的である．さらに，提案手法は (c) データストリームの長さに依存せず，非常に高速である．実データを活用した実験により，提案手法が時系列データストリームの将来予測のための既存手法と比較して高精度であること，計算時間についてデータサイズに依存せず，より高速なリアルタイム予測を達成していることを明らかにした．

キーワード　時系列予測，データストリーム処理，動的モード分解

1　はじめに

時系列データストリームは，IoT 関連の応用技術 [1, 2] や，web 活動 [3, 4]，疫病の感染 [5]，ユーザの行動 [6] など，多くの事象から大量に生成されている．このようなデータの応用のための重要な技術の一つとして大量に発生し続ける時系列データストリームをリアルタイムに解析する技術が挙げられる．リアルタイム処理の中でも，時系列データストリームのリアルタイム予測は重要かつ挑戦的な課題である．

実際に，時系列データストリームには周期性やその増幅や減衰などといった様々な潜在的なダイナミクスが含まれており，高精度な将来予測を実現するためには，複数のダイナミクスから構成された重要な時系列パターンを効果的にとらえることは必要不可欠である．このために，我々は動的モード分解 [7, 8] を活用する．動的モード分解とは，時系列データを動的モードと呼ばれる要素に分解することにより，時間的なダイナミクスを抽出する数値解析手法である．具体的に，動的モードとは，発振周波数や固有値によって決定される減衰率/振動数と密接な関わりがあるため，動的モード分解では周期性やその増幅や減衰といったダイナミクスを効率的に捉えることができる．

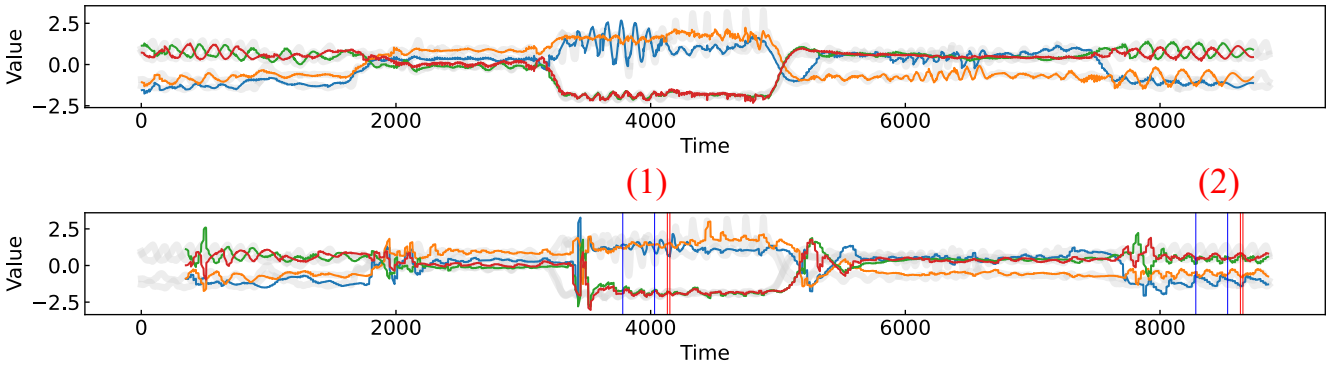
また，時系列データストリームに含まれる様々な潜在的ダイナミクスは刻一刻と変化する．例えば，web の検索イベントデータには 1 年を通じた特徴的な時系列パターンの遷移（e.g., ブラックフライデー，クリスマスなど）が見られることが多い．また，時期に応じて検索数には増減や減衰が見られる．そのため，より高精度に動的モード分解を活用した時系列データストリームのリアルタイム予測を実現するためには，潜在的なパターンの遷移も捉える必要がある．本論文ではこのような潜在的時系列パターンのことを “レジーム”と呼ぶ．

本研究では，上記で述べたような要件を満たした時系列データストリームのリアルタイム予測手法 ModeCast を提案する．本論文では次の問題を扱う．

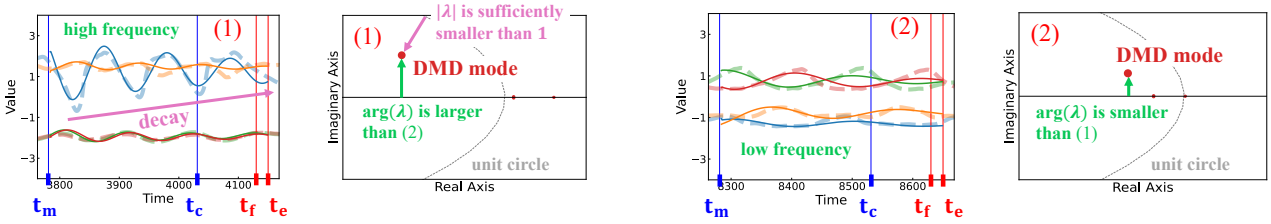
問題:. d 次元の観測データで構成される時系列データストリーム $X = \{\boldsymbol{x}(1), ..., \boldsymbol{x}(t_c), ...\}$ が与えられ，現時刻を t_c とした時，l_s ステップ先の $\{\boldsymbol{v}(t_c + l_s), ..., \boldsymbol{v}(t_c + l_s + l_p - 1)\}$ を予測する．

1.1　具体例

図 1 はモーションイベントストリーム (#2) *house-cleaning* における ModeCast の予測結果の様子を示している．このデータセットは 4 つのモーションキャプチャセンサ (左右の腕及び足) から生成された “house cleaning”に関する複数のモーション (e.g., dragging a mop, wiping a window) から構成される．図 1(a) は ModeCast のオリジナルデータストリーム (灰色の曲線) に対するフィッティング結果及び予測結果 (それぞれの色付きの曲線) を示している．具体的には，各時刻 t_c において $\{\boldsymbol{x}(t_c + 100), ..., \boldsymbol{x}(t_c + 119)\}$ の 20 点を予測し続けている (上記で定義した**問題**の $l_s = 100, l_p = 20$ のケース)．図中の青い線はそれぞれ $\{t_m, t_c\}$，赤い線はそれぞれ $\{t_f (= t_c + l_s), t_e (= t_c + l_s + l_p - 1)\}$ を示しており，t_c は現時刻を示す．この時系列データストリームには様々な明らかに異なる時系列パターンの変化が含まれているが，提案手法はその変化をよく捉えることができており，それを適切に予測に活用できていると言える．図 1(b) は，2 つの異なる時刻における ModeCast の予測結果のスナップショット及び予測のために使用されている複素平面上の動的モードの固有値 (点線は単位円) を示している．具体的には，スナップショットは $t_c = 4031, 8531$ における出力の様子を示している．また，図中の固有値の絶対値はダイナミクスの減衰率を，偏角が振動数を表している．特に，減衰率に関しては，1 より小さい場合は減



(a) モーションイベントストリーム全体 (灰色の曲線) に対するフィッティング結果（上）及び予測結果（下）



$t_c = 4031$ $t_c = 8531$

(b) 各時刻における $l_s = 100$ ステップ先の予測結果のスナップショット及び動的モードに対応した固有値

図 1　モーションイベントストリームに対する ModeCast による出力例：提案手法は継続的かつ自動的に l_s ステップ先の予測を行う．(a) 提案手法は時系列データストリームに対して良いフィッティング及び予測を実現する．(b) 異なる 2 点の l_s ステップ先の予測の様子及び予測に活用している動的モードに対応した固有値を示している．後者の動的モードに比べて前者の固有値の偏角の方が大きいことが，より高周波な成分を表現できていることを意味する．

衰，1 より大きい場合は増幅に対応している．図を見てわかる通り，提案手法は時系列パターンの変化を捉え，適切な予測を連続的かつ自動的に行っていると言える．また，2 枚のスナップショットを比較すると，周波数の観点から違いが明らかである．具体的には，$t_c = 4031$ 時点での時系列データは高周波であり，これを正確に表現できていることが確認できる．この観察を裏付ける形で，それぞれのスナップショットに対応する動的モードを比較したところ，固有値の偏角の大きさに顕著な違いがあることがわかる．さらに，前者のスナップショットに示される青色の曲線は減衰傾向にあるが，この点も固有値の絶対値を考慮することで，正確に表現できていると言える．

1.2　本論文の貢献

本論文では，大規模データストリームを効果的に表現する時系列モデル，及び，その将来予測アルゴリズムである ModeCast を提案する．ModeCast は次の特長を持つ．

- 時系列データに含まれる特徴的なダイナミクスを動的モード分解により抽出することで，柔軟にレジームを発見し，最適かつリアルタイムな予測を実現する．
- 様々なデータに対して，実用的な予測を行うことが可能である．
- 計算コストはデータストリームの長さに依存することなく，高速に処理が可能である．

2　関 連 研 究

時系列のモデリング及びその予測に関する研究は多岐にわたる．具体的な既存手法の例としては，自己回帰モデル（AR: autoregressive model），カルマンフィルタ（KF: Kalman filters）などがある [9]．加えて，このような古典的時系列解析の改良のための手法も多数提案されてきた [10–13]．RegimeCast [14] は，生成され続ける観測データに対応した潜在的なパターンを非線形動的システムを適用することでリアルタイムに推定し，将来予測を行う．しかし，RegimeCast は周期性やその減衰率に基づいた時系列パターンを抽出することが困難である．また，深層学習を活用した時系列予測のための研究も盛んに行われている [15–19]．DeepAR [18] は，RNN ベースの確率的予測手法であり，複数系列のダイナミクスを考慮した予測が可能である．しかし，このような深層学習を基とした時系列予測手法は，モデルの学習の際に膨大な時間を要するため，リアルタイムにモデルを更新させ，現時刻の情報を取り入れながら予測を連続的に行い続けることは困難である．

動的モード分解に関する研究も盛んに行われている [20–24]．この手法は数値流体力学の分野で複雑な流体動力学データからコヒーレント構造を抽出するための手法として提案された [24]．また，動的モード分解には，非線形動力学系の線形であるが無限次元の表現を提供する Koopman 作用素 [25–27] のスペクト

ル解析と密接に関連しているという理論的背景がある．加えて，動的モード分解はデータ駆動型手法である，すなわち，潜在的なダイナミクスの抽出のために入力データ以外必要としない．そのため，たとえ，動的システムの物理モデルが不明だとしても，潜在的なダイナミクスを把握することができる．このような特性から動的モード分解は，非線形力学系から生成される時系列データの解析に有効であると考えられており，実際に，流体力学 [24]，動画処理 [28,29]，脳波解析 [30]，疫学 [31] など，多くの分野で分析手法として広く採用されている．また，動的モード分解は時系列予測のためにも活用されている [32]．しかし，この手法は単次元データにしか対応していない．

3 動的モード分解

本節では，動的モード分解について詳細に述べる．動的モード分解とは，時間発展するデータに含まれる潜在的な非線形ダイナミクスを抽出するための数値解析手法であり，出力結果は特定の条件下において Koopman 演算子 [25–27] によるモーダル分解と一致する．即ち，対象のシステムの固有の減衰率/振動数を持つ特徴的な構造（i.e., 動的モード）を捉えることが可能である．以下の式 (1) は動的モード分解の対象となる動的システムを表している．

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k \tag{1}$$

上式における $\boldsymbol{A}$ は動的システムの時間発展を表現するパラメータであり，これを解析することで重要な動的モードを抽出する．入力時系列データ $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n} \in \mathbb{R}^{d\times n}$ に対して以下のようなデータ行列を考える．

$$\boldsymbol{X}_0 = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{n-1} \end{bmatrix} \in \mathbb{R}^{d\times(n-1)}$$
$$\boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{n} \end{bmatrix} \in \mathbb{R}^{d\times(n-1)}$$

式 (1) で表現される式が，最も正しく表現される $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ は上記のデータ行列を用いて以下のように記述できる．

$$\boldsymbol{X}_1\boldsymbol{X}_0^{\dagger} = \arg\min_{\boldsymbol{A}} \sum_{k=1}^{n-1} ||\boldsymbol{x}_{k+1} - \boldsymbol{A}\boldsymbol{x}_k||$$
$$= \arg\min_{\boldsymbol{A}} ||\boldsymbol{X}_1 - \boldsymbol{A}\boldsymbol{X}_0||$$

但し，$\boldsymbol{X}_0^{\dagger}$ は $\boldsymbol{X}_0$ の Moore-Penrose の擬似逆行列を示す． 以下では $\boldsymbol{A}$ を効率的に解析するための動的モード分解によるアルゴリズムを述べる．

I. $\boldsymbol{X}_0$ に対して特異値分解（SVD）を計算する．

$$\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$$

但し，$\boldsymbol{U} \in \mathbb{R}^{m\times k}, \boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}, \boldsymbol{V} \in \mathbb{R}^{(n-1)\times k}$ であり，k は特異値の数を表している．一般的に k は $\boldsymbol{X}_0$ のランクで与えられるが，一部削減を施すことにより観測される時系列データに通常含まれているノイズの影響を除去することが可能である．

II. ここで，より効率的な演算を行うために，線形作用素 $\boldsymbol{A}$ を左特異値ベクトル $\boldsymbol{U}$ が張る k ($\leqq \min(d,n)$) 次元部分空間への射影へを考える．

$$\tilde{\boldsymbol{A}} = \boldsymbol{U}^{\top}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}^{\top}\boldsymbol{X}_1\boldsymbol{V}\boldsymbol{\Sigma}^{-1}$$

III. $\tilde{\boldsymbol{A}}$ の固有値分解を計算する．ここで，II. で行った射影は同相写像であるため，ここで求めた固有値は $\boldsymbol{A}$ の支配的な k 個の固有値と一致する．

$$\tilde{\boldsymbol{A}}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda}$$

IV. III. で求めた $\boldsymbol{W}$ 及び $\boldsymbol{\Lambda}$ を用いて $\boldsymbol{A}$ の固有値分解を計算する．具体的には，III. で述べた通り，$\boldsymbol{A}$ の固有値は $\boldsymbol{\Lambda}$ で与えられ，固有ベクトルは $\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{W}$ で与えられる [1]．
以上の手続きにより得られた $\boldsymbol{\Phi}$ のことを動的モードと言う．また，$\boldsymbol{\Lambda}$ は Δt（1ステップ）間の動的モードの減衰率/振動数を表現している．具体的には，一般的な線形動的システムの解が $e^{a+bi}(a,b \in \mathbb{R})$ の形式で表現できることに着目すると，i 番目の動的モードの減衰率及び振動数はこれに対応する固有値 λ_i を用いて，それぞれ以下のような形式で表現できることがわかる．

$$\sigma_i = \frac{\mathrm{Re}(\log(\lambda_i))}{\Delta t}$$
$$\omega_i = \frac{\mathrm{Im}(\log(\lambda_i))}{\Delta t}$$

ここで，Re, Im はそれぞれ複素数の実部及び虚部である．

表 1 主な記号と定義

記号	定義
t_c	現在の時刻
$\boldsymbol{X}$	時系列データストリーム
$\boldsymbol{x}(t)$	時刻 t における観測値
$\boldsymbol{X}^C$	カレントウィンドウ：$\boldsymbol{X}^C = \boldsymbol{X}[t_m : t_c]$
$\boldsymbol{X}^F$	予測ウィンドウ：$\boldsymbol{X}^F = \boldsymbol{X}[t_f : t_e]$
h	遅延座標系への射影のための遅延時間
$\boldsymbol{X}_{\mathrm{aug}}$	遅延座標系で表現された時系列データストリーム
$\boldsymbol{\Lambda}$	固有値行列 $\boldsymbol{\Lambda} \in \mathbb{C}^{k\times k}$
$\boldsymbol{\Phi}$	動的モード $\boldsymbol{\Phi} \in \mathbb{C}^{dh\times k}$
$\boldsymbol{s}(t)$	時刻 t における潜在値
$\boldsymbol{v}(t)$	時刻 t における推定値
Θ	全レジームのパラメータ集合：$\Theta = \{\theta_1, ...\theta_r\}$

4 ModeCast モデル

本章では，本研究で扱う問題を定義し，その問題を解決するための ModeCast について述べる．まず，本手法において必要な概念の定義について説明する．また，**表 1** に主な記号と定義を示す．

［定義 1］（カレントウィンドウ $\boldsymbol{X}^C$） $\boldsymbol{X}^C = X[t_m : t_c]$，すなわち時刻 t_m から t_c までの区間で与えられる部分シーケンスを時系列予測に使用する最新のストリームデータとし，長さ $l_c = t_c - t_m$ のカレントウィンドウと呼ぶ．

1：このように $\boldsymbol{\Phi}$ を与える動的モード分解を投影動的モード分解 (projected DMD) と言う．

［定義 2］（予測ウィンドウ $\boldsymbol{X}^F$） 時刻 t_c から l_s ステップ先の区間で与えられる $\boldsymbol{X}^F$ を予測対象のデータ区間とし，予測ウィンドウと呼ぶ．ここで，$t_f = t_c + l_s$ を満たし，予測ウィンドウの長さが $l_p = t_e - t_f + 1$ の時，l_p ステップごとに $\boldsymbol{X}^F$ の値を予測する．

本論文では以下のような問題を取り扱う．

［問題 1］（l_s ステップ先予測） t_c を最新の時刻とし，長さが l_c のカレントウィンドウ $\boldsymbol{X}^C = \boldsymbol{X}[t_m : t_c]$ が与えられる間，l_s ステップ先の長さ l_p の時系列データ $\boldsymbol{X}^F = \boldsymbol{X}[t_f : t_e]$ を予測し続ける．

4.1 提案モデル

本研究の目的は，時系列データストリームが与えられたとき，その中から重要かつ潜在的な動的モードに基づいた時系列パターン（レジーム）を発見し，リアルタイム予測を実現することである．このような目的を達成するために ModeCast に対して以下の 2 つのアイデアを取り入れる．

(P1) 遅延座標への射影

(P2) 動的モードで表現されるレジーム

4.1.1 遅延座標への射影 (P1)

本項では初めに，時系列データの状態空間について述べる．本手法では時間遅延座標系を用いた状態空間を利用する．時間遅延座標系とは，非線形方程式を推定するベースとなる技術であり，より長期的な潜在的ダイナミクスを把握するためにしばしば用いられている．具体的には，$\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ に対して，以下のような行列 $\boldsymbol{X}_{\text{aug}}$ を考える．

$$\boldsymbol{X}_{\text{aug}} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{n-h+1} \\ \boldsymbol{x}_2 & \boldsymbol{x}_3 & \cdots & \boldsymbol{x}_{n-h+2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_h & \boldsymbol{x}_{h+1} & \cdots & \boldsymbol{x}_n \end{bmatrix} \tag{2}$$

但し，h は遅延時間を意味する．上記の行列の構造にある通り，$\boldsymbol{X}_{\text{aug}}$ の状態ベクトルに過去の情報を付与できているため，位相情報を付与することが可能になり，これにより，より高精度にダイナミクスを表現することが可能になる．尚，Takens の埋め込み定理 [33] によると，このように遅延座標系で表された運動軌道は，時系列データを生み出した現実の位相空間上の運動軌道と同相な軌道となるため，第 3 章で述べた動的モード分解のアルゴリズムに関して，$\boldsymbol{X}$ を $\boldsymbol{X}_{\text{aug}}$ に置き換えた結果得られる動的モードは適切なダイナミクスを表現する．

4.1.2 動的モードで表現されるレジーム (P2)

次に，時系列パターンの表現方法について述べる．本研究では，潜在的ダイナミクスに基づいた時系列パターンを表現するために，複数の動的モードから構成されるレジームを定義する．このモデルにおけるレジーム $\boldsymbol{\theta}$ は次の 2 種類の状態によって構成されると考える．

- $\boldsymbol{s}(t) \in \mathbb{C}^k$: 時刻 t における k 次元部分空間の潜在値
- $\boldsymbol{v}(t) \in \mathbb{R}^{dh}$: 時刻 t における dh 次元の観測値

これらを考慮した線形動的システムを次のように定義する．

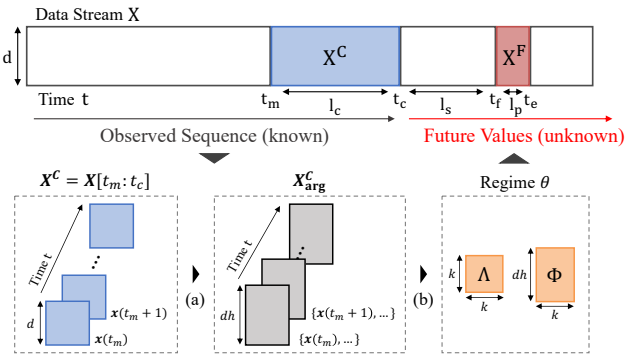


図 2 ModeCast のモデル概要図：(a) カレントウィンドウ $\boldsymbol{X}^C$ を遅延座標系へ射影し $\boldsymbol{X}^C_{\text{aug}}$ を得る．その後，(b) $\boldsymbol{X}^C_{\text{aug}}$ からレジーム $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}, \boldsymbol{\Phi}\}$ を計算し，l_s ステップ先を予測する．

$$\begin{aligned} \frac{d\boldsymbol{s}(t)}{dt} &= \boldsymbol{\Lambda}\boldsymbol{s}(t) \\ \boldsymbol{v}(t) &= \boldsymbol{\Phi}\boldsymbol{s}(t) \end{aligned} \tag{3}$$

但し，初期条件は $\boldsymbol{s}(0) = \boldsymbol{\Phi}^\dagger \boldsymbol{v}(0)$ で与えられる．また，$\boldsymbol{\Lambda}, \boldsymbol{\Phi}$ はそれぞれ動的モード分解で得られる固有値及び動的モードである．まとめると，以下を得る：

［定義 3］（レジーム $\boldsymbol{\theta}$） 単一のレジームにおけるパラメータ集合 $\boldsymbol{\theta}$ は右の通りとする：$\boldsymbol{\theta} = \{\boldsymbol{\Lambda}, \boldsymbol{\Phi}\}$.

さらに，我々は複数のレジーム $\boldsymbol{\theta}$ が時々刻々と変化する様子を捉えたい．そのため時系列データストリームを複数のレジームで表現をしたい．そのために複数のレジームで構成されたレジームセット $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_r\}$ も同時に定義しておく．

提案モデルの全体図を図 2 に示す．提案モデルはカレントウィンドウ $\boldsymbol{X}^C$ に最も適したレジーム $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}, \boldsymbol{\Phi}\}$ を遅延座標系及び動的モードを活用することで推定し，それを用いて将来予測を行う．

5 アルゴリズム

本章では，大規模時系列データストリームの将来予測アルゴリズムである ModeCast について述べる．

初めに，本アルゴリズムに必要な概念について定義を行う．

［定義 4］（更新用パラメータ $\boldsymbol{\omega}$） 遅延座標系で表現されたカレントウィンドウ $\boldsymbol{X}^C_{\text{aug}}$ に対して，$\boldsymbol{P} = (\boldsymbol{U}^\top \boldsymbol{X}^C_{\text{aug}_0} (\boldsymbol{U}^\top \boldsymbol{X}^C_{\text{aug}_0})^\top)^{-1}$ とする時，$\boldsymbol{\omega} = \{\boldsymbol{P}, \boldsymbol{U}\}$ を更新用パラメータと呼ぶ．但し，$\boldsymbol{U}$ は $\boldsymbol{X}^C_{\text{aug}}$ の左特異ベクトルである．

［定義 5］（全パラメータ集合 $\boldsymbol{M}$） 全レジームの数を r とした時，全レジームの集合 $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_r\}$，に対応した更新用パラメータの集合を $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_r\}$ とすると，ModeCast のための全パラメータ集合 $\boldsymbol{M}$ は $\{\boldsymbol{\Theta}, \boldsymbol{\Omega}\}$ となる．

また，我々の最終目標としては，**問題** 1 の l_s ステップ先予測を逐次的，かつ高速に行うことである．ここで，現時刻 t_c におけるレジームを $\boldsymbol{\theta}^c$，それに対応した更新用パラメータ集合を $\boldsymbol{\omega}^c$ とする．また，l_s ステップ先予測のために，時刻 t_c における潜在値が必要であり，これを $\boldsymbol{s}^c_{out}$ とする．これらをまとめて候補パラメータ $\boldsymbol{C} = \{\boldsymbol{\theta}^c, \boldsymbol{\omega}^c, \boldsymbol{s}^c_{out}\}$ として保持する．

Algorithm 1 ModeCast $(\boldsymbol{x}(t_c), \boldsymbol{M}, \boldsymbol{C})$

Input: (a) New value $\boldsymbol{x}(t_c)$ at time points t_c
(b) Parameter set $\boldsymbol{M} = \{\boldsymbol{\Theta}, \boldsymbol{\Omega}\}$
(c) Candidate $\boldsymbol{C} = \{\boldsymbol{\theta}^c, \boldsymbol{\omega}^c, \boldsymbol{s}^c_{out}\}$
Output: (a) Updated parameter set $\boldsymbol{M}'$
(b) Updated candidate $\boldsymbol{C}'$
(c) Estimated variables $\boldsymbol{V}^F$
$\boldsymbol{X}^C \leftarrow \boldsymbol{X}[t_m : t_c]$ // Update current window $\boldsymbol{X}^C$
/* Embed into time-delay coordinates */
Compute $\boldsymbol{X}^C_{\text{aug}}$ according to Eq. (2)
/* Estimate optimal $\boldsymbol{\theta}$ */
$\boldsymbol{M}', \boldsymbol{C}' \leftarrow$ ModeEstimator $(\boldsymbol{X}^C_{\text{aug}}, \boldsymbol{M}, \boldsymbol{C})$
/* Forecast l_s-step-ahead estimated value */
$\boldsymbol{V}^F \leftarrow$ ModeGenerator $(\boldsymbol{C}')$
return $\{\boldsymbol{M}', \boldsymbol{C}', \boldsymbol{V}^F\}$

5.1 提案アルゴリズム

ModeCast は次のアルゴリズムのとおり (Algorithm 1).

- カレントウィンドウ $\boldsymbol{X}^C$ から遅延座標系で表現されたカレントウィンドウ $\boldsymbol{X}^C_{\text{aug}}$ を生成する.
- ModeEstimator ：遅延座標系で表現されたカレントウィンドウ $\boldsymbol{X}^C_{\text{aug}}$，パラメータ集合 $\boldsymbol{M}$，候補パラメータ $\boldsymbol{C}$ が与えられとき，現在のデータに最も適したレジーム $\boldsymbol{\theta}^c$ を抽出する．また，パラメータ集合 $\boldsymbol{M}$ 及び候補パラメータ $\boldsymbol{C}$ を更新する (Algorithm 2).
- ModeGenerator ：候補パラメータ $\boldsymbol{C}$ が与えられた時，l_s ステップ先の予測ウィンドウ $\boldsymbol{V}^F$ を予測する.

5.2 ModeEstimator

遅延座標系で表現されたカレントウィンドウ $\boldsymbol{X}^C_{\text{aug}}$，パラメータ集合 $\boldsymbol{M}$，候補パラメータ $\boldsymbol{C}$ が与えられた時を考える．ここで，損失関数を $f(\boldsymbol{s}^c_0, \boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}})$ と定義する．但し，$\boldsymbol{s}^c_0$ はレジーム $\boldsymbol{\theta}^c$ の初期条件であり，$f(\cdot)$ は遅延座標系で表現されたカレントウィンドウ $\boldsymbol{X}^C_{\text{aug}}$ と，遅延座標系で表現された推定ウィンドウ $\boldsymbol{V}^C_{\text{aug}}$ のフィッティング精度を示す. 具体的には，最小二乗誤差にて計算を行う (i.e., $f(\boldsymbol{s}^c_0, \boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}}) = \sum_{t=t_m}^{t_c} ||\boldsymbol{x}(t) - \boldsymbol{v}(t)||$). 続いて，具体的な $\boldsymbol{V}^C_{\text{aug}}$ の計算方法について述べる．式 (3) で与えられる線形動的システム的には，レジーム $\boldsymbol{\theta}^c$ に対して初期値 $\boldsymbol{s}^c_0$ として $\boldsymbol{\Phi}^\dagger \boldsymbol{X}^C_{\text{aug}}[0]$ を与えれば計算自体は可能である．しかし，初期値 $\boldsymbol{X}^C_{\text{aug}}[0]$ に過度なノイズが含まれていた場合，適切な $\boldsymbol{V}^C_{\text{aug}}$ を計算することができない．そこで，我々は LM (Levenberg-Marquardt) アルゴリズム [34] を用いて $\boldsymbol{s}^c_0$ を最適化することで，ノイズの影響の軽減により，より適切な $\boldsymbol{V}^C_{\text{aug}}$ の計算を行う．ここで，上記で述べたような初期値 $\boldsymbol{s}^c_0$ を最適化するための関数を $g(\cdot)$ と定義する．また，この一連の過程を計算するための関数，すなわち，$f(\cdot)$ と $g(\cdot)$ を組み合わせた関数を $\mathcal{F}(\cdot)$ と定義する.

ModeEstimator は以下の手順に従う.

(I) $\mathcal{F}(\boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}}) > \rho$ となるとき，レジームセットからフィッティング精度を最小化するレジームを探索し，$\boldsymbol{\theta}^c$ を更新する.

Algorithm 2 ModeEstimator $(\boldsymbol{X}^C_{\text{aug}}, \boldsymbol{M}, \boldsymbol{C})$

Input: (a) Current augmented window $\boldsymbol{X}^C_{\text{aug}}$
(b) Parameter set $\boldsymbol{M}$
(c) Candidate $\boldsymbol{C}$
Output: (a) Updated parameter set $\boldsymbol{M}'$
(b) Updated candidate $\boldsymbol{C}'$
if $\mathcal{F}(\boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}}) > \rho$ **then**
  /* Find better regime in $\boldsymbol{\Theta}$ */
  $\{\boldsymbol{s}^c_0, \boldsymbol{\theta}^c\} \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{F}(\boldsymbol{\theta}, \boldsymbol{X}^C)$
  if $\mathcal{F}(\boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}}) > \rho$ **then**
    /* Create new regime */
    Compute $\boldsymbol{\Lambda}, \boldsymbol{\Phi}$ from $\boldsymbol{X}^C_{\text{aug}}$ according to Chapter 3
    $\boldsymbol{\theta}^c \leftarrow \{\boldsymbol{\Lambda}, \boldsymbol{\Phi}\}$; $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} \cup \boldsymbol{\theta}^c$
    $\boldsymbol{\omega}^c \leftarrow \{\boldsymbol{P}, \boldsymbol{U}\}$; $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} \cup \boldsymbol{\omega}^c$
    $\boldsymbol{s}^c_0 \leftarrow g(\boldsymbol{\theta}^c, \boldsymbol{X}^C)$
  else
    $\{\boldsymbol{\theta}^c, \boldsymbol{\omega}^c\} \leftarrow$ RegimeUpdate$(\boldsymbol{\theta}^c, \boldsymbol{\omega}^c, \boldsymbol{X}^C_{\text{aug}})$
  end if
else
  $\boldsymbol{s}^c_0 \leftarrow g(\boldsymbol{\theta}^c, \boldsymbol{X}^C)$
  $\{\boldsymbol{\theta}^c, \boldsymbol{\omega}^c\} \leftarrow$ RegimeUpdate$(\boldsymbol{\theta}^c, \boldsymbol{\omega}^c, \boldsymbol{X}^C_{\text{aug}})$
end if
$\boldsymbol{M}' \leftarrow \{\boldsymbol{\Theta}, \boldsymbol{\Omega}\}$
$\boldsymbol{C}' \leftarrow \{\boldsymbol{\theta}^c, \boldsymbol{\omega}^c, \boldsymbol{s}^c_{out}\}$
return $\boldsymbol{M}', \boldsymbol{C}'$

(II) $\mathcal{F}(\boldsymbol{\theta}^c, \boldsymbol{X}^C_{\text{aug}}) > \rho$ となるとき，新たなレジームを生成し，レジームセット $\boldsymbol{\Theta}$ へ追加する.

(III) (I), (II) の過程で得られた $\boldsymbol{V}^C_{\text{aug}}$ の終端の潜在値 $\boldsymbol{s}^c_{out}$ を l_s ステップ先の予測のための初期値として，候補パラメータ $\boldsymbol{C}$ に保持する.

(IV) 最適なレジームとして既存のレジームが採択された場合，RegimeUpdate に従って更新する.

ここで，遅延座標系で表現された推定ウィンドウ $\boldsymbol{X}^C_{\text{aug}}$ に未知のパターンが出現した時，新たなレジーム $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}, \boldsymbol{\Phi}\}$ を推定する必要がある．この推定にはそれぞれ第 3 章で述べたアルゴリズムに従って推定する.

オンラインレジーム更新. 新たな観測データ $\boldsymbol{x}(t_c)$ 及び 1 ステップ前のデータ $\boldsymbol{x}(t_c - 1)$ から，既存レジームの要素 $\boldsymbol{\Lambda}$，$\boldsymbol{\Phi}$ を更新用パラメータ集合 $\boldsymbol{\omega}^c = \{\boldsymbol{P}, \boldsymbol{U}\}$ を活用して，以下の流れで更新する.

I'. 新たなデータ及びその一時刻前のデータを左特異ベクトル $\boldsymbol{U}$ が張る k 次元部分空間へ射影する

$$\tilde{\boldsymbol{x}}_1 \leftarrow \boldsymbol{U}^\top \boldsymbol{x}(t_c)$$
$$\tilde{\boldsymbol{x}}_0 \leftarrow \boldsymbol{U}^\top \boldsymbol{x}(t_c - 1)$$

II'. $\tilde{\boldsymbol{A}}$ を $\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}$ から再構成し，以下の式に従って更新する.

$$\tilde{\boldsymbol{A}} \leftarrow (\boldsymbol{U}^\top \boldsymbol{\Phi}) \boldsymbol{\Lambda} (\boldsymbol{U}^\top \boldsymbol{\Phi})^\dagger$$
$$\tilde{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{A}} + \gamma (\tilde{\boldsymbol{x}}_1 - \tilde{\boldsymbol{A}} \tilde{\boldsymbol{x}}_0) \tilde{\boldsymbol{x}}_0^\top \hat{\boldsymbol{P}}$$

但し，$\gamma, \hat{\boldsymbol{P}}$ とはそれぞれ以下の通り．

$$\hat{\boldsymbol{P}} = \frac{\boldsymbol{P}}{\delta}, \quad \gamma = \frac{1}{1 + \tilde{\boldsymbol{x}}_0^\top \hat{\boldsymbol{P}} \tilde{\boldsymbol{x}}_0}$$

III'. 次の更新のために $\boldsymbol{P}$ を更新する．

$$\boldsymbol{P} \leftarrow \frac{1}{\delta}(\hat{\boldsymbol{P}} - \gamma \hat{\boldsymbol{P}} \tilde{\boldsymbol{x}}_0 \tilde{\boldsymbol{x}}_0^\top \hat{\boldsymbol{P}})$$

IV'. 第 3 章の III. IV. の手順通りに更新後の $\tilde{\boldsymbol{A}}$ から $\boldsymbol{\Lambda}$，$\boldsymbol{\Phi}$ を求める．

また，現時刻 t_c に近ければ近いほど，カレントウィンドウ $\boldsymbol{X}^C$ の l_s ステップ先の予測に効果的なレジームを発見するために重要であるため，更新の際，新たな情報を付与するだけでなく，過去の情報をモデルから忘却させる必要がある．そこで，忘却係数 δ を導入する．[35] に従い，δ は特別に明記しない限り $\delta = 0.999$ とする．

5.3 ModeGenerator

次に，ModeEstimator で推定された候補パラメータ $\boldsymbol{C}$ が与えられた時を考える．

ModeGenerator は以下の手順で行われる．

(I) 式 (3) に従って，$\boldsymbol{s}_{out}^c$ を初期値として，時刻 t_e までの推定値をレジーム $\boldsymbol{\theta}^c$ を利用して遅延座標系で表現された推定ウィンドウ $\boldsymbol{V}_{\text{aug}}^F$ を生成する．

(II) 遅延座標系で表現されたカレントウィンドウ $\boldsymbol{V}_{\text{aug}}^F$ から推定ウィンドウ $\boldsymbol{V}^F$ を復号する．

6 評価実験

本論文では，ModeCast の有効性を検証するため，実データを用いた実験を行った．本章では以下の項目について検証する．

Q1 データストリームの予測に対する提案手法の有効性

Q2 リアルタイム予測に対する提案手法の精度の検証

Q3 データストリームの予測に対する計算時間の検証

また，実験は 512GB のメモリ，Xeon Gold 6258R 2.7GHz の 28 コア CPU および RTX A6000 48GB GPU を搭載した Linux マシン上で実施した．各データセットは平均値と分散で正規化 (z-normalization) して使用した．

比較手法. 有効性を検証するために，以下の比較手法と比較を行った．

- RegimeCast [14]：非線形動的システムに基づく将来予測アルゴリズム．提案論文に従い，潜在状態の数 $k = 4$，モデルの階層数 $h = 2$，モデル推定の閾値 $\epsilon = 0.5 \cdot ||\boldsymbol{X}^C||$ とした．
- ARIMA：古典的な時系列予測手法の一つ．適切なパラメータ数は [14] に従い，AIC を用いて決定した．
- DeepAR [18]：時系列予測のための最新のニューラルネットワーク手法．リカレントニューラルネットワーク（RNN）の層の数の候補は $\{1, 2\}$，ユニット数の候補は $\{8, 16, 32, 64\}$ とした．また，最適化アルゴリズムとしては Adam [36] を使用し，学習率を 0.01 で固定した．最大エポック数 100 で，適切な early stopping を利用し学習を行った結果，最もフィッティング性能が良いモデルを採用した．

表 2 データセット詳細

データセット	属性・クエリ	次元
センサデータ：CMU Motion Capture Database [37] から取得		
(#1) *exercise*	左右の腕と足の動き	4
(#2) *house-cleaning*	”	4
(#3) *chicken-dance*	”	4
Web データ：GoogleTrends [38] から取得		
(#4) *G-outdoor*	skiing, fishing, cycling	3
(#5) *G-sports*	tennis, American football, Major League Baseball (MLB)	3
(#6) *Beer*	Corona, Keystone, Coors, Modelo, SierraNevada	5
(#7) *OnlineTV*	Netflix, Hulu, YouTube, Amazon(P)	4
(#8) *Social media*	Tumblr, Facebook, LinkedIn, SoundCloud, Yelp	5
(#9) *Software*	HTML, Java, SQL, JavaScript, HTML5	5

データセット. 実験に使用した 2 種類，計 9 つの実データセットは次のとおりである．また，データセットの詳細を**表 2** に示す．

- (#1) *exercise*, (#2) *house-cleaning*, (#3) *chicken-dance*：様々な動きをする人間の様子をモーションキャプチャによって収集したセンサデータ．
- (#4) *G-outdoor*, (#5) *G-sports*, (#6) *Beer*, (#7) *OnlineTV*, (#8) *Social media*, (#9) *Software*：GoogleTrends [38] から毎週収集されたさまざまなクエリ (i.e., キーワード) の検索数によって構成されている．

6.1 Q1：提案手法の有効性

本節では，多次元時系列データストリームに対する ModeCast の予測性能を検証する．実際のモーションイベントストリームにおける ModeCast の予測結果については第 1 章の図 1 で説明したとおりであり，提案手法はレジームの時間変化を捉えることによって，高精度なリアルタイム将来予測を実現している．特に，図 1(a) のとおり，ModeCast は与えられた時系列データから適切なレジームを捉え，それを予測に活かすことができている．また，図 1(b) では，各時刻における ModeCast のリアルタイムに $l_s = 100$ 先の予測をしている様子及びその予測のために使用されている複素平面上の動的モードの固有値 $\boldsymbol{\Lambda}$ が描画されている．この図は，寄与率の高い固有値が，時系列データストリームに含まれる周期性及びその増幅や減衰を適切に表現できていることを明らかにしている．

6.2 Q2：提案手法の精度

本節では，ModeCast の l_s ステップ先の予測精度を検証するため，比較手法である RegimeCast，ARIMA 及び DeepAR と精度比較を行った．長期のモーションイベントストリーム (#1, #2) においては $l_s = 100$，短期のモーションイベントストリーム (#3) においては $l_s = 30$，web の検索イベントデー

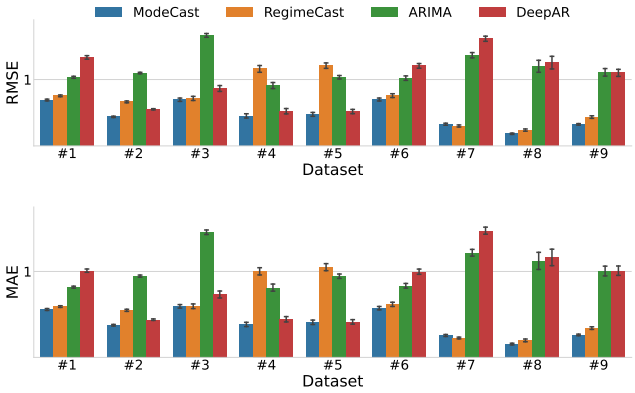


図 3　ModeCast の平均予測精度の比較

タ (#4 − #9) においては $l_s = 13$ (i.e., 3ヶ月先) と設定した．全ての手法において，パラメータチューニングのためにシーケンス全体の 3 分の 1 を訓練データとして与え，残りのデータに対して将来予測を行った．図 3 は，多次元時系列データストリームにおける ModeCast 及び比較手法の予測誤差の平均値を示している．また，評価指標としては推定値の二乗平均誤差（RMSE：root mean square error）および平均絶対誤差（MAE：mean absolute error）の 2 種類を利用している．図に示す通り，提案手法は比較手法である RegimeCast，ARIMA 及び DeepAR と比較し，2 つの評価指標のどちらにおいても高い予測精度を持つ．RegimeCast は固有値に基づいた潜在的なダイナミクスを捉えられていないため，予測精度が低下している．特に，web の検索イベントデータのうち，(#4, #5) のような固有の周期をもつと考えられるデータに対して予測精度が低下する．ARIMA は時系列データ間の線形関係を前提としたモデルであるため，複雑かつ非線形なデータに対応できず予測精度が低下する．また，DeepAR は逐次的なモデルの更新ができないため，レジームの遷移に対応できない．このため，予測精度が低下しており，特に (#1) のような複数の明らかに異なるレジームを持つデータに対する予測精度が低下する．

6.3　Q3：提案手法の計算時間

続いて，提案手法の計算コストについて検証する．図 4 の左図は，(#1) *exercise* に対する各時刻 t_c における計算コストを，提案手法の ModeCast と比較手法である RegimeCast, ARIMA 及び DeepAR と比較したものである．また，図 4 の右図は，データストリーム全体の計算時間の平均値を示している．これらの図の y 軸は対数スケールで表示している．図 4 が示す通り，ModeCast はデータストリームの長さに依存せずに，高速に動作することが可能である．特に，提案手法が高速である主な要因は，固有値分解のために部分空間への射影を行っている点である．実際に，固有値分解の計算コストは $O(n^3)$ であるため，行列のサイズを小さくするフェーズは本手法の高速化に大きく貢献していると考えられる．一方、各比較手法は適切なレジームの学習のための処理に時間を要するため，ModeCast と比べて低速である．RegimeCast については新たなレジームを推定するフェーズに多くの時間を要するためである．実際，図 4 が示す通り，RegimeCast の予測の大半が 1〜5 秒程度で処理が完了しているのに対して，新たなレジームを推定しているであろう予測段階の処理時間が 100 秒弱に達している．これが全体の平均が約 10 秒となっている原因である．また，ARIMA 及び DeepAR は逐次的なモデルの更新ができないため，各時刻におけるストリーム処理に対する計算に一定の時間を要する．そのため，高速な時系列予測が実現できていない．以上より，提案手法は時系列データストリームのリアルタイム予測に適した性能を有すると言える．

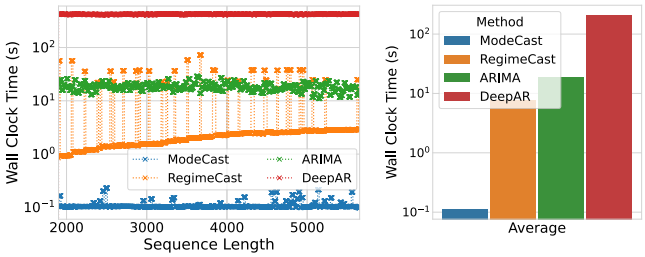


図 4　各時刻 t_c における計算コスト (左) と平均値 (右)：これらの図の y 軸は対数スケールで表示されている．

7　むすび

本論文では，動的モード分解を活用した大規模時系列データストリームの高速予測手法として ModeCast を提案した．ModeCast は冒頭で確認した以下の優れている特性を全て達成している．

- 時系列データに含まれる特徴的なダイナミクスを動的モード分解により抽出することで，柔軟にレジームを発見し，最適かつリアルタイムな予測を実現する．
- 様々なデータに対して，実用的な予測を行うことが可能である．
- 計算コストはデータストリームの長さに依存することなく，高速に処理が可能である．

また，実データを用いた実験では，提案手法の性能を評価し，既存の時系列データストリームの将来予測のための既存手法と比較して予測精度，計算時間ともに改善していることを確認した．

謝辞　本研究の一部は JSPS 科研費，JP21H03446，国立研究開発法人情報通信研究機構委託研究（NICT）委託研究 JPJ012368C03501，JSTAIP 加速課題 JPMJCR21U4，JST CREST JPMJCR23M3, の助成を受けたものです．

文　献

[1] Gianmarco De Francisci Morales, Albert Bifet, Latifur Khan, Joao Gama, and Wei Fan. Iot big data stream mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2119–2120, 2016.

[2] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks*, 4(3):161–175, 2018.

[3] Koki Kawabata, Yasuko Matsubara, Takato Honda, and

Yasushi Sakurai. Non-linear mining of social activities in tensor streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2093–2102, 2020.

[4] Kota Nakamura, Yasuko Matsubara, Koki Kawabata, Yuhei Umeda, Yuichiro Wada, and Yasushi Sakurai. Fast and multi-aspect mining of complex time-stamped event streams. In *Proceedings of the ACM Web Conference 2023*, pages 1638–1649, 2023.

[5] Tasuku Kimura, Yasuko Matsubara, Koki Kawabata, and Yasushi Sakurai. Fast mining and forecasting of co-evolving epidemiological data streams. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3157–3167, 2022.

[6] Yasuko Matsubara, Yasushi Sakurai, Christos Faloutsos, Tomoharu Iwata, and Masatoshi Yoshikawa. Fast mining and forecasting of complex time-stamped events. In Qiang Yang, Deepak Agarwal, and Jian Pei, editors, *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 271–279. ACM, 2012.

[7] Jonathan H Tu. *Dynamic mode decomposition: Theory and applications*. PhD thesis, Princeton University, 2013.

[8] J Nathan Kutz, Steven L Brunton, Bingni W Brunton, and Joshua L Proctor. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

[9] James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Number 9780199641178 in OUP Catalogue. Oxford University Press, 2012.

[10] Lei Li, B Aditya Prakash, and Christos Faloutsos. Parsimonious linear fingerprinting for time series. *Proceedings of the VLDB Endowment*, 3(1-2):385–396, 2010.

[11] Qiquan Shi, Jiaming Yin, Jiajun Cai, Andrzej Cichocki, Tatsuya Yokota, Lei Chen, Mingxuan Yuan, and Jia Zeng. Block hankel tensor arima for multiple short time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5758–5766, 2020.

[12] Spiros Papadimitriou, Anthony Brockwell, and Christos Faloutsos. Adaptive, hands-off stream mining. In *Proceedings 2003 VLDB Conference*, pages 560–571. Elsevier, 2003.

[13] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association*, 106(496):1513–1527, 2011.

[14] Yasuko Matsubara and Yasushi Sakurai. Regime shifts in streams: Real-time forecasting of co-evolving time sequences. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054, 2016.

[15] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers.

[16] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987, 2022.

[17] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[18] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[19] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[20] Yoshinobu Kawahara. Dynamic mode decomposition with reproducing kernels for koopman spectral analysis. *Advances in neural information processing systems*, 29, 2016.

[21] Naoya Takeishi, Yoshinobu Kawahara, Yasuo Tabei, and Takehisa Yairi. Bayesian dynamic mode decomposition. In *IJCAI*, pages 2814–2821, 2017.

[22] N Benjamin Erichson, Lionel Mathelin, J Nathan Kutz, and Steven L Brunton. Randomized dynamic mode decomposition. *SIAM Journal on Applied Dynamical Systems*, 18(4):1867–1891, 2019.

[23] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

[24] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

[25] Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[26] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41:309–325, 2005.

[27] Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S Henningson. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641:115–127, 2009.

[28] N Benjamin Erichson and Carl Donovan. Randomized low-rank dynamic mode decomposition for motion detection. *Computer Vision and Image Understanding*, 146:40–50, 2016.

[29] Jacob Grosek and J Nathan Kutz. Dynamic mode decomposition for real-time background/foreground separation in video. *arXiv preprint arXiv:1404.7592*, 2014.

[30] Jan Rohleff. An incremental approach to dynamic mode decomposition for time-varying systems with applications to a model for erythropoiesis. Master's thesis, Universität Konstanz, Kontanz, 2020.

[31] Joshua L Proctor and Philip A Eckhoff. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International health*, 7(2):139–145, 2015.

[32] Santosh Tirunagari, Samaneh Kouchaki, Norman Poh, Miroslaw Bober, and David Windridge. Dynamic mode decomposition for univariate time series: analysing trends and forecasting. 2017.

[33] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, pages 366–381. Springer, 2006.

[34] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977*, pages 105–116. Springer, 2006.

[35] Hao Zhang, Clarence W Rowley, Eric A Deem, and Louis N Cattafesta. Online dynamic mode decomposition for time-varying systems. *SIAM Journal on Applied Dynamical Systems*, 18(3):1586–1609, 2019.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[37] Carnegie Mellon University Motion Capture Database. `http://mocap.cs.cmu.edu/`.

[38] Google Trends. `https://trends.google.com/trends/`.

学習可能な長さを持つshapeletsに基づく時系列分類法

山口　晃広† 植野　研† 鹿島　久嗣††

† 株式会社東芝 研究開発センター 知能化システム研究所 システム AI ラボラトリー
〒 212–8582 川崎市幸区小向東芝町 1
†† 京都大学 大学院情報学研究科 知能情報学専攻　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †akihiro5.yamaguchi@toshiba.co.jp

あらまし　Shapelets とは時系列データのクラス分類に用いられる複数の局所的な波形パターンである．従来研究では，shapelets の形を連続最適化により学習する手法が提案されているが，shapelets の長さはハイパーパラメータとして固定され学習できなかった．本研究では，shapelets の形だけでなく長さ自体も連続最適化により同時に学習可能にする．特に，分類の判定根拠を提示する説明性を保ちつつ，クラス間の波形の違いをより明確に捉えるように，shapelets の長さが学習されることを理論的に示す．実験評価では，71 種類の UCR 公開時系列データセットを用いて AUC の向上や合計学習時間の短縮を示すとともに，3 種類のケーススタディを通して学習された shapelets の説明性を示す．

キーワード　Shapelets，クラス分類，時系列データ，Interpretable ML，説明可能性

1 はじめに

IoT の普及に伴い，機械学習を用いた時系列データのクラス分類手法（時系列分類法）が研究されている．2 クラス分類の場合は，学習時に正または負のクラスラベルが付与された複数の時系列インスタンスを与えて，テスト時に未知の時系列インスタンスのクラスラベルを予測する問題となる．一般的なクラス分類とは異なり時系列分類では，時系列の並びや波形パターンの形状が重要であり，波形パターンの出現位置のずれや時系列データのノイズにロバストな時系列分類法がこれまでに提案されている [1]．

機械学習による意思決定に対してその判定根拠を提示する説明性が，欧州をはじめ日本でも重要視されている [2, 3, 4]．説明性の実現には，意思決定を行うモデル自体を透明で説明可能とするアプローチと，元のモデルを別の説明可能なモデルで事後的に解釈するアプローチがある [5]．後者のアプローチでは元のモデルを近似するためその解釈は不完全で忠実ではない場合があり，適用分野によってはモデル自体の説明性が求められる [6]．

Shapelets とは時系列分類に用いる複数の局所的な波形パターンである．時系列分類法の中で shapelets を機械学習により発見する手法（shapelet 法）が注目されている [7, 8]．Shapelet 法では，分類に有効な特徴は長い時系列全体ではなく局所的な波形パターンに表れるというアイデアに基づく．特長として，学習が終わればテスト時の分類は高速であり，良好な分類性能を達成する．より重要なことは，特徴ベクトルが shapelets と時系列インスタンスとの間の非類似度（距離）に対応しておりモデル自体が透明で説明性を持つ．医療・インフラ・製造などの産業分野の専門家は波形パターンに関する専門知識を持っている場合が多いため，特に少数の shapelets を発見できれば，専門家は学習モデルに用いられる特徴を全て理解しその判定根拠を専門知識と比較することで納得して利用できる．そのため，このような産業分野の専門家に比較的受け入れられやすい [9, 10]．

Shapelets の研究は，学習データを構成する部分時系列を探索してその中から shapelets を選択する探索型の手法から始まったが [8]，shapelets が学習データの部分時系列に制限されない任意の波形パターンを表せるように shapelet の各点を連続変数とみなして確率的勾配降下法（SGD）で分類器と同時に学習する学習型の手法も提案された [7]．探索型の手法 [8, 11, 12] と比べて，学習型の手法（shapelet 学習法）は分類性能を向上するだけでなく学習時の計算量を削減しノイズにロバストな shapelets を生成できる [7, 13, 14, 15, 16]．

しかしながら，従来手法では shapelets の長さはハイパーパラメータとして固定されており学習できない．例えば，文献 [17] では shapelets を学習する前に shapelets の長さをヒューリスティックに選定する方法が提案されたが，shapelets の長さは学習時には固定されたままである．各 shapelet の形と長さは相互に依存しているため本来ならば同時に学習されるべきである．

殆どの従来研究では，shapelets を学習する際にはそれらの長さや個数はハイパーパラメータであり交差検証でチューニングされる．各 shapelet に対してその長さをチューニングしようとすると，組み合わせが増えてしまい学習を何度も繰り返す必要があるため，多くのユーザーにとっては現実的に難しい．また，交差検証ができるほど学習データを十分に収集できない場合もある．それゆえ，従来のアプローチから脱却して，shapelets の長さを効率的かつ効果的に学習することが重要な課題である．

本研究では shapelets の長さ自体を学習するという新しい視点を提案する．本提案の shapelet 学習法（Time-series Shapelets with Learnable Lengths; TSL という技術）では，各 shapelet の長さを連続変数としその形と一緒に連続最適化で学習する．

図 1 は shapelet の長さを学習することによる時系列分類への効果を示している．図 1 (a) は，上に尖った局所波形の周辺を拡大した時系列インスタンスである．上に尖った波形の傾きは負例では正例よりも緩やかに上昇するため，上に尖った波形に着目することで正例と負例を正しく分類できる．従来手法（LTS） [7]

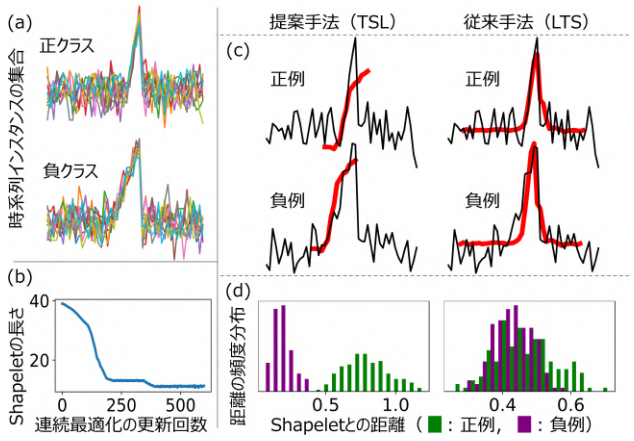


図 1　学習可能な長さを持つ shapelets は分類に寄与する局所波形を効率的かつ効果的に捉えクラス間の特徴量も明確に分離する．

では，shapelet の長さはハイパーパラメータであり最適値よりも長く設定されたとする [1]．一方，提案手法（TSL）では，図 1 (b) に示すように連続最適化の更新中に shapelet の長さを最適値に近づけることができる．その結果，図 1 (c) で TSL と LTS とで学習された shapelets（太赤線）を比較すると，TSL の方が分類に寄与する局所波形をより適切な長さで捉えていることが分かる．これにより，shapelet と時系列インスタンスとの間の距離に基づく特徴量が，図 1 (d) に示すように 2 つのクラス間でより明確に分離される．このように，shapelets の長さを学習することで分類性能が向上することが，この簡単な例から確認できる．

5 章と 6 章での実験評価では，UCR アーカイブ [18] から 71 種類のデータセットを用いて Area Under the Curve (AUC) 及び合計学習時間における提案アプローチの有効性を示すとともに，3 つのケーススタディを通して shapelets の説明性も示す．

本論文の主な貢献を以下に示す．なお，本内容は文献 [19] に準じ，関連研究や式の導出の詳細は文献 [19] を参照されたい．

- Shapelets の長さを学習可能とする最初の研究である．
- Shapelets と 2 クラス分類器だけでなく各 shapelet の長さも同時に学習する 1 つの連続最適化問題を定式化する．
- Shapelet の長さを学習することで，shapelet が対象クラスの波形に近づきそれ以外の波形から離れる（つまり，shapelet の説明性とクラス間分離性が強化される）．
- 上述の性質は数学的に保証でき，計算量の増加もない．
- オプションとして，分類損失とモデルの複雑さを考慮して shapelets の個数を自動的に特定する方法についても提案し，実験的にその個数が少ない傾向になることを示す．

2　準　　備

2.1　記号と定義

本研究では，時系列インスタンスを $\{1,-1\}$ の 2 クラスに分類する問題を扱う．学習データのインスタンスの数を I 個として，i 番目の真のクラスラベルを $y_i \in \{1,-1\}$ とする．図 1 のように，長い数列として表される時系列インスタンスに対して shapelet は短い数列として表される．長さ Q の時系列データセットを $\mathbf{T} \in \mathbb{R}^{I \times Q}$ とし，i 番目の時系列インスタンス $\mathbf{t}_i \in \mathbb{R}^Q$ の j 番目の値を $t_{i,j} \in \mathbb{R}$ と記述する．また，shapelets の数を K 個とし，k 番目の shapelet の長さを L_k として，K 個の shapelets を $\mathbf{S} := [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_K]$ とする．

本研究では，shapelet の長さ L_k は連続変数（つまり，$L_k \in \mathbb{R}_+$）であり，連続最適化の中で shapelet $\mathbf{s}_k$ と一緒に学習される．

2.2　固定された長さを持つ shapelets の学習

従来の shapelets 学習法では，shapelets の個数 K と k 番目の shapelet の長さ L_k は固定された整数であり，各 shapelet の形のみが学習される．本節では，殆どの従来手法の基になっている文献 [7] の定式化を説明する．時系列インスタンス $\mathbf{t}_i$ と shapelet $\mathbf{s}_k$ の間の距離は，$\mathbf{s}_k$ の時系列上の位置 j を 1 から $Q-L_k+1$ までの範囲で移動させたときに部分時系列 $[t_{i,j}, t_{i,j+1}, \cdots, t_{i,j+L_k-1}]$ と $\mathbf{s}_k$ の間の最小の二乗ユークリッド距離として定義される．

$$x_{i,k}(\mathbf{s}_k, \mathbf{t}_i) := \min_{j=1,2,\cdots,Q-L_k+1} \frac{1}{L_k} \sum_{l=1}^{L_k} (t_{i,j+l-1} - s_{k,l})^2. \quad (1)$$

ここで，$\mathbf{x}_i := [x_{i,1}, x_{i,2}, \cdots, x_{i,K}] \in \mathbb{R}^K$ が i 番目のインスタンスの特徴ベクトルである．分類器には線形予測モデルを用いる．

$$\hat{y}_i(\mathbf{x}_i, \mathbf{w}) := \sum_{k=1}^{K} w_k x_{i,k} + w_0. \quad (2)$$

ここで，$\mathbf{w} \in \mathbb{R}^{K+1}$ はバイアス項を含むモデルパラメータである．クラスラベルが $\{1,-1\}$ であることに注意し，（インスタンスごとの）交差エントロピー損失関数 $\mathcal{L}(\cdot,\cdot)$ を用いる．

$$\mathcal{L}(y_i, \hat{y}_i) := -\frac{1+y_i}{2} \ln(\sigma(\hat{y}_i)) - \frac{1-y_i}{2} \ln(1-\sigma(\hat{y}_i)),$$
$$\sigma(\hat{y}_i) := \left(1 + e^{-\beta \hat{y}_i}\right)^{-1}. \quad (3)$$

文献 [14] では動的にスケールパラメータ β を調整するが，ここでは β を 1 に固定する．最終的には，各 k 番目の shapelet $\mathbf{s}_k \in \mathbb{R}^{L_k}$ の長さ $L_k \in \mathbb{N}$ が固定して与えられ，特徴ベクトルの集合 $\mathbf{X}$ を通して shapelets $\mathbf{S}$ はモデルパラメータ $\mathbf{w}$ と同時に最適化される．

$$\underset{\mathbf{S} \in \prod_{k=1}^{K} \mathbb{R}^{L_k}, \mathbf{w} \in \mathbb{R}^{K+1}}{\text{minimize}} \sum_{i=1}^{I} \mathcal{L}(y_i, \hat{y}_i) + \lambda \sum_{k=1}^{K} w_k^2. \quad (4)$$

ここで，λ は ℓ_2 正則化パラメータであり非負の固定値をとる．

このように従来の定式化では，shapelet の長さ L_k は整数値で固定されており学習できない．

3　方　法　論

Shapelets の長さを学習することは基本的な問題であり，特徴量に式 (1) を用いる従来手法 [7, 13, 14, 15, 16, 17, 21, 22, 23, 20] には提案する方法論を直接適用できる．本章では従来手法 [7] に適用するかたちで提案する方法論を説明する．3.1 節で shapelets の長さを連続変数に緩和した後，連続最適化問題を定式化し勾配を導出する．3.4 節では shapelets の長さを学習する際の重要な性質を理論的に示し，最後にアルゴリズムをまとめる．

1：長い時系列データで交差検証を用いて最適な shapelet の長さを特定することは難しく，ここでは最適値よりも長い shapelet 長が学習前に設定されたと仮定する．

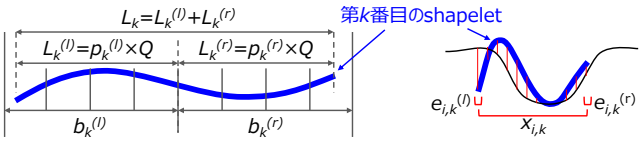


図 2　新しく導入した変数.

3.1　連続変数を用いた shapelet の長さ

2.2 節で述べたように，従来手法では $k=1,2,\cdots,K$ に対して各 shapelet $\mathbf{s}_k$ の長さ L_k は整数であったが，本節では L_k を連続最適化で学習できるように連続変数に緩和する（つまり，$L_k\in\mathbb{R}_+$ とする）．このために導入する変数を図 2 (左) に示す．

Shapelet $\mathbf{s}_k$ の両端点でその長さ L_k を別々に変更できるように，L_k を左側 $L_k^{(\mathrm{l})}$ と右側 $L_k^{(\mathrm{r})}$ に分割する（つまり，$L_k=L_k^{(\mathrm{l})}+L_k^{(\mathrm{r})}$ が成立する）．そして，連続最適化で学習する前に $L_k^{(\mathrm{l})}=L_k^{(\mathrm{r})}=\frac{L_k}{2}$ と初期化する．以降では，上付き文字 (l) と (r) は左側と右側を表し，上付き文字 $(*)$ は (l) と (r) の両側を表す．

一般に，センサの解像度で時系列の長さ Q は異なり，shapelet の長さ L_k は Q に依存するため，共通の学習率で L_k を直接最適化することは非効率かもしれない．そのため，Q への比率 $p_k^{(*)}\in\mathcal{R}_+$ を用いて（両側の）shapelet の長さ $L_k^{(*)}$ を最適化する．

$$p_k^{(*)}:=L_k^{(*)}/Q. \tag{5}$$

次に，$L_k^{(*)}$ より大きい最小の整数を $b_k^{(*)}$ とし，$L_k^{(*)}$ 以下の最大の整数を $\tilde{b}_k^{(*)}$ とし，$b_k^{(*)}$ を用いて L_k の上界を b_k で定義する．

$$b_k^{(*)}:=\lfloor L_k^{(*)}\rfloor+1,\qquad \tilde{b}_k^{(*)}:=\lfloor L_k^{(*)}\rfloor,\qquad b_k:=b_k^{(\mathrm{l})}+b_k^{(\mathrm{r})}. \tag{6}$$

3.2　連続最適化問題の定式化

本節では，長さを学習可能な shapelets に基づく特徴ベクトルを定義した後，本提案の最適化問題を定式化する．

Shapelet の学習可能な長さ L_k は連続変数だが，shapelet $\mathbf{s}_k$ を構成する実際の点数は整数である．このギャップを埋めるために，$L_k\in\mathbb{R}_+$ の上界 $b_k\in\mathbb{N}$ を用いて $\mathbf{s}_k$ が b_k 個の点から構成されるとまずは考える．次に，$k=1,2,\cdots,K$ と $i=1,2,\cdots,I$ と $j=1,2,\cdots,Q-b_k+1$ に対して，重み付き二乗ユークリッド距離 $d_{i,k,j}$ を計算し，$d_{i,k,j}$ を用いて特徴ベクトルの各要素 $x_{i,k}$ を定義する．

$$\begin{aligned} x_{i,k}(p_k^{(*)},\mathbf{s}_k,\mathbf{t}_i)&:=\min_{j=1,2,\cdots,Q-b_k+1} d_{i,k,j},\\ d_{i,k,j}(p_k^{(*)},\mathbf{s}_k,\mathbf{t}_i)&:=\frac{1}{L_k}\sum_{l=1}^{b_k} v_{k,l}(t_{i,j+l-1}-s_{k,l})^2. \end{aligned} \tag{7}$$

ここで，shapelet $\mathbf{s}_k$ の両端点 $(s_{k,1},s_{k,b_k})$ だけが $L_k^{(*)}-\tilde{b}_k^{(*)}$ つまり $p_k^{(*)}Q-\tilde{b}_k^{(*)}$ で重みづけられる．

$$v_{k,l}(p_k^{(*)}):=\begin{cases} p_k^{(\mathrm{l})}Q-\tilde{b}_k^{(\mathrm{l})} & \text{if } l=1,\\ 1 & \text{if } l=2,3,\cdots,b_k-1,\\ p_k^{(\mathrm{r})}Q-\tilde{b}_k^{(\mathrm{r})} & \text{if } l=b_k. \end{cases} \tag{8}$$

読者は shapelet $\mathbf{s}_k$ の各端点の二乗誤差 $(t_{i,j+l-1}-s_{k,l})^2$ を $L_k^{(*)}-\tilde{b}_k^{(*)}$ で重みづけるだけのこの定義の適切さに疑問を持つかもしれない．しかし，3.3 節および 3.4 節でそれぞれ後述するように，勾配の計算を容易にするだけでなく，shapelets の学習において重要な性質も保有する．ここでは，式 (8) を用いて定義される式 (7) が重み付き二乗ユークリッド距離であることを証明する．

命題 1. 各 $k=1,2,\cdots,K$ と $l=1,2,\cdots,b_k$ に対して，式 (8) から (a)$0\leqq v_{k,l}\leqq 1$ と (b)$v_{k,1}+v_{k,2}+\cdots+v_{k,b_k}=L_k$ を共に満たす．

証明. 条件 (a) は $p_k^{(*)}Q-\tilde{b}_k^{(*)}=L_k^{(*)}-\lfloor L_k^{(*)}\rfloor\in[0,1)$ から成立する．条件 (b) は式 (5) と式 (6) を用いて $\sum_{l=1}^{b_k}v_{k,l}=(p_k^{(\mathrm{l})}Q-\tilde{b}_k^{(\mathrm{l})})+(b_k-2)+(p_k^{(\mathrm{r})}Q-\tilde{b}_k^{(\mathrm{r})})=L_k$ から成立する．　□

最終的には，shapelets の個数 K と学習データセット $(\mathbf{T},\mathbf{y})$ が与えられて，本最適化問題は，式 (5) で計算される shapelets の長さ $\mathbf{L}^{(*)}$ の比率 $\mathbf{p}^{(*)}$，特徴ベクトル集合 $\mathbf{X}$ を通して shapelets $\mathbf{S}$，モデルパラメータ $\mathbf{w}$ を同時に最適化するよう定式化される．

$$\underset{\substack{\mathbf{p}^{(\mathrm{l})}\in\mathbb{R}_+^K,\mathbf{p}^{(\mathrm{r})}\in\mathbb{R}_+^K\\ \mathbf{S}\in\prod_{k=1}^K\mathbb{R}^{b_k},\mathbf{w}\in\mathbb{R}^{K+1}}}{\text{minimize}} \sum_{i=1}^{I}G_i,\quad G_i:=\mathcal{L}(y_i,\hat{y}_i)+\frac{\lambda}{I}\sum_{k=1}^{K}w_k^2. \tag{9}$$

ここで，分類損失 $\mathcal{L}$ は式 (2)，(7) と (8) を用いて式 (3) で定義され，λ は非負の ℓ_2 正則化パラメータである．

本節の最後に，従来研究 [14, 24] と同様にモデルパラメータ w_k を用いて，shapelet $\mathbf{s}_k$ の（対象）クラスを定義する．

定義 1. 各 $k=1,2,\cdots,K$ に対して shapelet $\mathbf{s}_k$ が所属するクラスは $w_k<0$ と $w_k>0$ の場合にそれぞれ 1 と-1 とする．

Shapalet の長さ L_k を学習せずに固定長の shapelet $\mathbf{s}_k$ を学習する場合には，文献 [14] の Property 3.1 のように，$\mathbf{s}_k$ は対象クラスの波形に類似しそれ以外の波形からは逸脱する傾向がある．我々はこの望ましい性質が shapelet の長さ L_k を学習する際にも成立することを 3.4 節で証明する．つまり，shapelet の形と長さを同時に学習することで，この性質が更に強化されてより高品質な shapelet が生成されることを理論的に示す．

3.3　解　　法

Shapelets の長さが固定された従来手法と同様に，式 (9) の最適化問題は非凸だが，その目的関数はインスタンスごとの目的関数 G_i の和に分解できる．また，G_i は shapelet の長さ L_k に対しても殆どいたるところで微分できる．これらの理由から，式 (9) を SGD アルゴリズムで解く．以降，時系列インスタンス $\mathbf{t}_i$ に shapelet $\mathbf{s}_k$ がマッチングする時系列上の位置を j^* とする．

$$j^*(p_k^{(*)},\mathbf{s}_k,\mathbf{t}_i):=\underset{j=1,2,\cdots,Q-b_k+1}{\arg\min}\ d_{i,k,j}. \tag{10}$$

ここで $d_{i,k,j}$ は式 (8) を用いて定義される式 (7) のものと等しい．

Shapelet の長さの比率 $p_k^{(*)}$ に対する G_i の勾配は微分の連鎖率を用いて分解できる．

$$\frac{\partial G_i}{\partial p_k^{(*)}}=\frac{\partial\mathcal{L}(y_i,\hat{y}_i)}{\partial p_k^{(*)}}=\frac{\partial\mathcal{L}(y_i,\hat{y}_i)}{\partial\hat{y}_i}\frac{\partial\hat{y}_i}{\partial x_{i,k}}\frac{\partial x_{i,k}}{\partial p_k^{(*)}}. \tag{11}$$

図 2 (右) のように，k 番目の shapelet $\mathbf{s}_k$ の端点と i 番目の時系列インスタンス $\mathbf{t}_i$ の対応する点との間の二乗誤差を $e_{i,k}^{(*)}$ とする．

$$e^{(\mathrm{l})}_{i,k}(p^{(\mathrm{l})}_k,\mathbf{s}_k,\mathbf{t}_i) := (t_{i,j^*}-s_{k,1})^2,$$
$$e^{(\mathrm{r})}_{i,k}(p^{(\mathrm{r})}_k,\mathbf{s}_k,\mathbf{t}_i) := (t_{i,j^*+b_k-1}-s_{k,b_k})^2. \quad (12)$$

これを用いて勾配 $\partial x_{i,k}/\partial p^{(*)}_k$ を書き下す．

$$\frac{\partial x_{i,k}}{\partial p^{(*)}_k} = \frac{e^{(*)}_{i,k}-x_{i,k}}{p^{(\mathrm{l})}_k+p^{(\mathrm{r})}_k}. \quad (13)$$

他は従来研究 [7, 14] と同様だが，勾配 $\partial x_{i,k}/\partial s_{k,l}$ の導出には若干の違いがある．また，文献 [14, 23] と同様に劣勾配を用いて時系列の長さ Q に対する計算量を線形オーダーまで削減する．

$$\frac{\partial x_{i,k}}{\partial s_{k,l}} = \frac{2v_{k,l}}{L_k}(s_{k,l}-t_{i,j^*+l-1}). \quad (14)$$

3.4 Shapelets の長さの学習における理論解析

本節では，shapelets の長さを学習することで，shapelets が対象クラスの波形に類似しそれ以外の波形からは逸脱することを理論的に示す．これは，shapelet が説明性を維持しつつクラス間で特徴量をより効果的に分離することを意味する．この理論解析が本研究の技術的な最大の強みである．

まず準備として，式 (11) における勾配の符号に影響する変数を補題 1 で特定し，$p^{(*)}_k$ と $x_{i,k}$ の増減関係を補題 2 で関連付ける．

補題 1. 式 (11) における勾配 $\partial\mathcal{L}(y_i,\hat{y}_i)/\partial p^{(*)}_k$ の符号は $\left(x_{i,k}-e^{(*)}_{i,k}\right)y_i w_k$ の符号に一致する．

証明. $y_i=1$ または -1 で $\sigma(\hat{y}_i)-\frac{y_i+1}{2}$ はそれぞれ負または正の値をとり，$p^{(\mathrm{l})}_k+p^{(\mathrm{r})}_k>0$ であるため，以下の関係式より成立する．

$$\frac{\partial\mathcal{L}(y_i,\hat{y}_i)}{\partial p^{(*)}_k} = \left(\sigma(\hat{y}_i)-\frac{y_i+1}{2}\right)w_k\frac{e^{(*)}_{i,k}-x_{i,k}}{p^{(\mathrm{l})}_k+p^{(\mathrm{r})}_k}. \quad \square$$

補題 2. 式 (8) を用いて定義される式 (7) において，$x_{i,k}<e^{(*)}_{i,k}$ であれば，$p^{(*)}_k$ が減少または増加すると $x_{i,k}$ は減少または増加する．$x_{i,k}>e^{(*)}_{i,k}$ の場合は，逆の結果が成り立つ．

証明. ここでは $p^{(\mathrm{l})}_k$ が変化する場合を証明するが，$p^{(\mathrm{r})}_k$ が変化する場合も同様である．式 (10) と (12) を用いて式 (7) を書き直す．

$$v_{k,1}\left(x_{i,k}-e^{(\mathrm{l})}_{i,k}\right) = \sum_{l=2}^{b_k} v_{k,l}\left((t_{i,j^*+l-1}-s_{k,l})^2-x_{i,k}\right). \quad (15)$$

式 (8) より $p^{(\mathrm{l})}_k$ が増減すると $v_{k,1}$ も増減する．また，各 $l=1,2,\cdots,b_k$ で $v_{k,l}$ が非負であることにも注意する．まずは $x_{i,k}<e^{(\mathrm{l})}_{i,k}$ の場合を背理法で証明する．式 (15) において，$v_{k,1}$ が変化しても $x_{i,k}$ が変化しないと仮定すると，左辺は変化するが右辺は変化しないという矛盾が生じる．また，$v_{k,1}$ が減少（または増加）すると $x_{i,k}$ は増加（または減少）すると仮定すると，式 (15) において左辺が減少（または増加）して右辺が増加（または減少）するという矛盾が生じる．これらの理由から，$p^{(\mathrm{l})}_k$ ひいては $v_{k,1}$ が減少（または増加）すると，$x_{i,k}$ も減少（または増加）する．次に，$x_{i,k}<e^{(\mathrm{l})}_{i,k}$ の場合も同様に背理法で証明できる． $\square$

補題 1 と 2 から次の主定理が成立する．

Algorithm 1 提案手法のアルゴリズム

Input: 学習データ: $(\mathbf{T},\mathbf{y})$; 繰り返し回数: M;
ℓ_2 正則化パラメータ: λ; [オプション：shapelets の個数: K]
Output: Shapelets の長さ: $(L_1,L_2,\cdots,L_K)$，ここで各 $k=1,2,\cdots,K$ で $L_k=p^{(\mathrm{l})}_k Q+p^{(\mathrm{r})}_k Q$ と計算; Shapelets: $\mathbf{S}$; モデルパラメータ: $\mathbf{w}$

1: 変数 $\mathbf{p}^{(*)}$ と $\mathbf{b}^{(*)}$ と $\mathbf{S}$ と $\mathbf{w}$ を初期化する．
2: **for** $m=1,2,\cdots,M$ **do**
3: [オプション：文献 [14] のように式 (3) の β を調整する．]
4: **for** $i=1,2,\cdots,I$ **do**
5: 3.3 節に従い，$\mathbf{p}^{(*)}$ と $\mathbf{S}$ と $\mathbf{w}$ をそれらの勾配で更新する．
6: **for** $k=1,2,\cdots,K$ **do**
7: **if** $p^{(*)}_k Q \geq b^{(*)}_k$ **then**
8: $b^{(*)}_k$ をインクリメントして $\mathbf{s}_k$ の端点を広げる．
9: **end if**
10: **if** $p^{(*)}_k Q < b^{(*)}_k - 1$ **then**
11: $b^{(*)}_k$ をデクリメントして $\mathbf{s}_k$ の端点を縮める．
12: **end if**
13: **end for**
14: **end for**
15: **end for**
16: **return** $(L_1,L_2,\cdots,L_K)$, $\mathbf{S}$, $\mathbf{w}$

定理 1. 各 $k=1,2,\cdots,K$ と $i=1,2,\cdots,I$ に対して，shapelet $\mathbf{s}_k$ と時系列インスタンス $\mathbf{t}_i$ の間の距離を式 (7) の $x_{i,k}$ で測り，$w_k \neq 0$ 及び $x_{i,k} \neq e^{(*)}_{i,k}$ を仮定する．式 (11) における勾配を用いて shapelet の長さ $L^{(*)}_k$ を更新すると，$\mathbf{s}_k$ のクラスが y_i であれば $\mathbf{s}_k$ は $\mathbf{t}_i$ に近づき，そうでなければ $\mathbf{s}_k$ は $\mathbf{t}_i$ から遠ざかる．

証明. 式 (5) より $L^{(*)}_k$ が増減すればその比率 $p^{(*)}_k$ も同じ方向に増減することに注意する．まず，shapelet $\mathbf{s}_k$ のクラスが y_i である場合，すなわち定義 1 から $y_i w_k<0$ の場合を証明する．$x_{i,k}<e^{(*)}_{i,k}$ であれば，式 (11) の勾配更新で補題 1 より $p^{(*)}_k$ は減少し，補題 2 より $\mathbf{s}_k$ と $\mathbf{t}_i$ の距離 $x_{i,k}$ も減少する．逆に $x_{i,k}>e^{(*)}_{i,k}$ の場合，式 (11) の勾配更新で補題 1 より $p^{(*)}_k$ は増加し，補題 2 より距離 $x_{i,k}$ は減少する．次に，shapelet $\mathbf{s}_k$ のクラスが y_i でない場合，すなわち $y_i w_k>0$ の場合，式 (11) の勾配更新が補題 1 より逆向きになり，距離 $x_{i,k}$ は $x_{i,k}<e^{(*)}_{i,k}$ と $x_{i,k}>e^{(*)}_{i,k}$ で増加する． $\square$

Shapelets の研究からは離れるが，畳み込みフィルタのサイズを勾配降下法で学習する従来研究がある [25]．しかし，そのフィルタサイズを shapelets の長さに置き換えたとしても，式 (13) の勾配の導出が異なるため補題 1 は成立せず，その結果定理 1 も成立しない．このように，この性質は提案手法に固有である．

3.5 アルゴリズムとその計算量

Algorithm 1 に TSL の疑似コードを示す．Shapelets の個数 K を入力するかはオプションであり，1 行目で 4 章で後述する方法を用いれば K を自動的に発見できる．なお，$p^{(*)}_k$ と $b^{(*)}_k$ の初期化は 3.1 節に従う．オプションとして 3 行目では，文献 [14] のように式 (3) の β を調整してもよい．5 行目では，3.3 節の勾配更新を適用し，$b^{(*)}_k-1.5 \leq p^{(*)}_k Q \leq b^{(*)}_k+0.5$ となるように $p^{(*)}_k$ をクリッピングする．6–13 行目では文献 [25] のように，更新された shapelet $\mathbf{s}_k$ の長さ L_k に応じて上界 $b^{(*)}_k$

と下界 $\tilde{b}_k^{(*)}(=b_k^{(*)}-1)$ を調整する．8 行目で $\mathbf{s}_k$ を広げる場合は，文献 [25] のように元の $\mathbf{s}_k$ の端点の値でパディングする．

提案手法 TSL の計算量を文献 [7] に従い分析する．元の計算量 $O(IMQ^2)$ [7] を増加することなく，本方法論による追加の計算量は $O(IMQ)$ であり，文献 [14, 23] と同様に時系列の長さ Q に対して線形である．3 行目で β を設定する場合の計算量も $O(IMQ)$ である [14]．従って，4 章のオプションを利用するか否かでアルゴリズム全体の計算量はそれぞれ $O(IMQ+e)$ か $O(IMQ)$ になり，e は後述する Algorithm 2 の計算量である．

4 オプション：BIC を用いた shapelets 数の特定

3 章で提案した方法論は shapelets の長さを最適化するが shapelets の個数を特定できない．そのため，本章ではこの個数を特定する方法をオプションとして提案する．また，式 (9) は非凸最適化問題であり，shapelets 学習法では shapelets の初期値の選択が重要である [7]．そのため，文献 [20] のアイデアを基に良質な初期 shapelets を併せて選択する．文献 [20] との主な違いは，モデルの複雑さを考慮する点と，shapelets の個数を自動的に特定する点である．なお本章でのみ，初期 shapelets の候補を $\mathbf{S}:=[\mathbf{s}_1,\mathbf{s}_2,\cdots,\mathbf{s}_K]$ と記述しその候補の個数を K とする．

4.1 解　　法

文献 [20] と同様に，初期 shapelets の候補を時系列データセットの部分時系列に限定し，式 (3) の損失関数の最小化に効く少数の部分時系列を初期 shapelets として選択する． 一方，文献 [20] とは異なり，この方法ではベイズ情報量規準（BIC）に基づいて shapelets の個数を自動的に特定する．

時系列データセット $\mathbf{T}$ には多くの部分時系列があり，それら全てを初期 shapelets の候補とみなすと，特徴ベクトルを計算するだけで $O(I^2Q^2)$ の計算量になる．そのため，文献 [20] と同様に本章の方法もオンライン特徴選択に基づく．基本的には，$\mathbf{T}$ から様々な長さの部分時系列をランダムに選んで初期 shapelets の候補に追加しながら，BIC が最小になるように不要な候補を削除することを繰り返すことで，初期 shapelets を得る．クラスラベルが 1 と予測される確率を式 (3) の $\sigma(\cdot)$ と解釈すると，式 (3) の $\mathcal{L}(\cdot,\cdot)$ で定義される損失関数は負の対数尤度関数である．また，自由パラメータの個数は $K+1$ である．従って，BIC は次の最適化問題の最適値として求まる．

$$\underset{\mathbf{w}\in\mathbb{R}^{K+1}}{\text{minimize}} \sum_{i=1}^{I}\mathcal{L}(y_i,\hat{y}_i)+\frac{1}{2}(K+1)\log(I). \qquad (16)$$

ここで，$\mathbf{S}$ が初期 shapelets の候補であることを除き，$\hat{y}_i$ は式 (1) を用いて定義される式 (2) である．モデルの複雑さ（つまり式の第 2 項）は $\mathbf{w}$ に依存しないので，最適解は標準的なロジスティック回帰と等しく，ソルバーで効率的に解ける [26]．

更に文献 [20] と同様に 2 つの方法で計算コストを削減する．1 つ目では，文献 [27] のように非冗長な部分時系列のみを処理する．2 つ目では，$k=1,2,\cdots,K$ の中で重み $|w_k|$ が最小となる部分時系列 $\mathbf{s}_k$ に絞って不要か否かを調べる．これは，重み $|w_k|$ が小さいと，式 (2) の $\hat{y}_i$ への影響が小さく分類損失 $\mathcal{L}(y_i,\hat{y}_i)$ への影響も小さい傾向があるためである．

Algorithm 2 Shapelets の個数特定と初期化のアルゴリズム

Input: 学習データ: $(\mathbf{T},\mathbf{y})$; 繰り返し回数: M
Output: Shapelets の個数: K; Shapelets と分類器の初期値: $(\mathbf{S},\mathbf{w})$

1: 部分時系列間の冗長性の閾値を決定して，$K\leftarrow 0$, $\mathbf{S}\leftarrow\emptyset$, $\mathbf{R}\leftarrow\emptyset$.
2: **for** $m=1,2,\cdots,IM$ **do**
3: 任意の長さの部分時系列 $\mathbf{s}$ をランダムに選ぶ.
4: **if** $\mathbf{s}$ が $\mathbf{R}$ に含まれるどの部分時系列とも類似しない **then**
5: $\mathbf{R}$ と $\mathbf{S}$ に $\mathbf{s}$ を追加し，$K\leftarrow K+1$, $(o,\mathbf{w})\leftarrow$calcBIC($\mathbf{S}$).
6: **while** $K\geqq 2$ **do**
7: $k\leftarrow \underset{k=1,2,\cdots,K}{\arg\min}\ |w_k|$ として，$(o',\mathbf{w}')\leftarrow$calcBIC($\mathbf{S}\backslash\{\mathbf{s}_k\}$).
8: もし $o<o'$ であれば while ループを抜ける.
9: $\mathbf{S}\leftarrow\mathbf{S}\backslash\{\mathbf{s}_k\}$, $K\leftarrow K-1$, $\mathbf{w}\leftarrow\mathbf{w}'$, $o\leftarrow o'$.
10: **end while**
11: **end if**
12: **end for**
13: **return** K, $(\mathbf{S},\mathbf{w})$

式 (16) の BIC に基づく初期 shapelets の個数は，式 (9) における shapelets の最適な個数とは正確には一致しない．代わりに，5 章と 6 章で AUC，特定された shapelets の個数，shapelets の説明性の観点から実験的にこのアプローチの有効性を検証する．

4.2 アルゴリズムとその計算量

Algorithm 2 に 4.1 節の疑似コードを示す．関数 calcBIC(·) は初期 shapelets の候補を入力して式 (16) を解きその目的関数（つまり BIC）の値 o と最適解 $\mathbf{w}$ を出力する．集合 $\mathbf{R}$ には繰り返し処理で選ばれた部分時系列を格納する．1 行目では文献 [27] のように $\mathbf{T}$ からランダムに選んだ部分時系列のペア間の距離の 25%タイルを閾値として設定し，4 行目で冗長性を判定する．5 行目では新しい非冗長な部分時系列 $\mathbf{s}$ を初期 shapelets の候補 $\mathbf{S}$ に追加し，6–10 行目で不要な候補を特定し削除する．

この計算量を分析する．4 行目で非冗長な部分時系列が見つかる回数を r として，式 (16) の 1 回分のロジスティック回帰の計算量を u とする．文献 [27] と同様に $r\ll IM$ が成立する．部分時系列の冗長性を判定する計算量は $O(IM|\mathbf{R}|)$ であり，特徴ベクトルと calcBIC(·) を計算する計算量はそれぞれ $O(rIQ)$ と $O(ru)$ である．従って，全体の計算量は $O(IM|\mathbf{R}|+r(IQ+u))$ となる．

5 定量的な実験

本章では 5.1 節で実験設定を述べた後で定量的に TSL を評価する．まずは 71 種類の UCR 公開時系列データセットで AUC，発見された shapelets の個数と長さの傾向，合計学習時間を評価し，次に人工時系列データでその長さを変えた場合も評価する．

5.1 実 験 設 定

提案手法である **TSL** は Algorithms 1 と 2 で実装されており Algorithm 1 の 3 行目で β を調整する．比較手法は大きく 2 つのカテゴリに分けられる．1 つ目は，時系列の長さ Q に線形または 2 次のオーダーの計算量を持つ shapelet 手法であり，TSL やその変種を含めて 6 種類の手法からなる．**woScale**

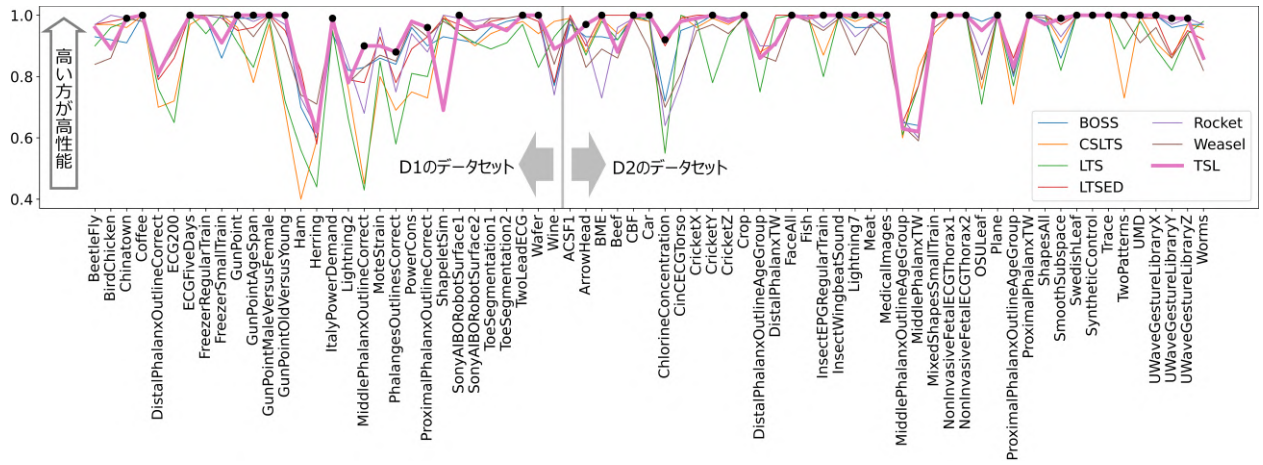


図 3　各テストデータにおける TSL と従来手法の AUC（41 個の黒点で TSL が最良）.

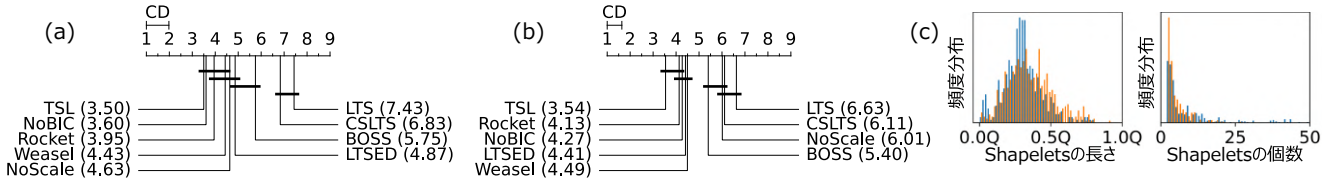


図 4　AUC の CD ダイアグラム (a) D1 and (b) D1&2.　(c) Shapelets の長さと個数の分布.

は TSL において Algorithm 2 の 3 行目で β を 1 に固定する．**woBIC** は TSL において Algorithm 2 を用いない．**LTS** [7] は良く知られた shapelet 学習法で高い正解率を持つ [28, 29]．**CSLTS** [13] はコスト考慮型の損失関数を導入して不均衡な 2 クラス分類で高い F 値を達成する．**LTSED** [14] は式 (3) の β を調整しつつ教師有特徴選択により shapelets を初期化することで少数の shapelets でも 2 クラス分類で高い AUC を達成する．

残りのカテゴリは，通常の PC でも実行できる 3 種類のよく知られた shapelets 以外の手法である．多数の畳み込みカーネルから特徴ベクトルを作る **Rocket** [30] は，正解率の高いアプローチの 1 つとして知られる．**BOSS** [31] や **WEASEL** [32] は，時系列データを記号で近似して特徴量を作る手法の中で正解率が高い．

UCR 公開時系列データセット [18] に含まれる 30 種類の 2 クラス分類向けデータセットは文献 [14, 20, 33] のものと同じである．更に，UCR 公開時系列データセットに含まれる 41 種類の多クラス分類向けデータセットから 2 つの最小クラスのみを使用することで，合計 71 種類の 2 クラス分類向けデータセットを使用する．これら欠損値のない 30 種類，71 種類，41 種類のデータセットをそれぞれ **D1**，**D1&2**，**D2** とする．D1 は各データセットのサイズが $IQ^2 \leqq 30{,}000{,}000$ であるように選ばれた．D2 は各データセットで各クラスに 5 個以上の学習データがあり $IQ^2 \leqq 50{,}000{,}000$ と $IQ \leqq 100{,}000$ を満たすように選ばれた．学習データとテストデータの分割にはデフォルトの設定を使用し [18]，実験を 5 回繰り返す．CSLTS では正クラスの見逃しをより厳しく罰するため小さいクラスを 1 とし大きいクラスを-1 とする．全ての shapelet 手法の繰り返し回数 M は 600 である．Shapelets の個数と長さの設定が必要な場合はそれぞれ {2,5,10} と $\{0.2Q, 0.3Q, 0.4Q\}$ から選ぶ．LTSED を除く shapelet 手法と Rocket では正則化パラメータを {0.1,1,10} から選ぶ．CSLTS では追加のハイパーパラメータを $\theta \in \{1{,}100\}$ と $D \in \{0.1, 10\}$ から選ぶ．これらのハイパーパラメータを設定するために学習データの 20%を事前に確保する．AUC は分類器の閾値に依存せず均衡データ及び不均衡データの両方に適している．文献 [14, 20, 33] と同様に各手法の AUC を比較する．

5.2　AUC の比較

図 3 は，各テストデータにおける TSL と従来手法の AUC 値を表し，TSL が最良となるデータセットを黒点でプロットしている．TSL は 71 種類のデータセット（D1&2）のうち 41 種類のデータセットで最良であり分類性能が高いことが分かる．図 4 (a) と (b) は，それぞれ D1 のテストデータと D1&2 のテストデータにおける Critical Difference（CD）ダイアグラムである．括弧内の値は平均順位であり，太棒で繋がれた手法は 95%信頼水準で互いに有意差がない [34]．

どちらの CD ダイアグラムも次の共通する結果を示している．TSL は，従来の shapelet 手法（LTSED，CSLTS，LTS）及び BOSS を有意に上回り，Rocket と（D1 では WEASEL とも）差は有意ではないが最良の順位である．TSL と woScale の比較から式 (3) の β の調整が AUC を改善すること，TSL と woBIC の比較から 4 章の方法が AUC を悪化させないことも分かる．

本節の結果から，shapelets の学習において提案手法の AUC への改善を確認した．また，学習したモデルに説明性がない Rocket, WEASEL, BOSS とは異なり，TSL のモデルは専門家に理解しやすい少数個の shapelets に基づき説明性がある．

5.3　発見された shapelets の長さと個数

図 4 (c) は，実験全体で TSL で発見された shapelets の長

表 1　全実験にわたる合計学習時間（単位：時間）.

CSLTS	LTS	LTSED	TSL （提案手法）
5783.8	1204.0	1217.8	313.0

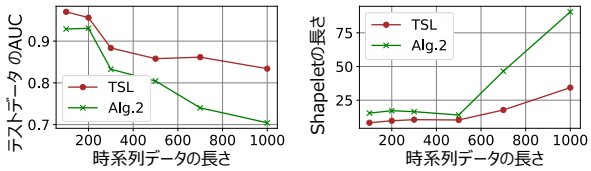


図 5　時系列データの長さを変えた場合への影響.

さと個数の分布を示し，青色と橙色がそれぞれ D1 と D2 のデータセットに対応する．71 種類のデータセット（D1&2）全体で学習された shapelets の長さの最頻値と中央値は，それぞれ $Q\times0.25$ と $Q\times0.32$ である．更に，71 種類のデータセット全体で特定された shapelets の個数の最頻値と中央値は，それぞれ 2 と 4 で少ない．この結果は，2 クラス分類のみで確認されたことではあるが，BIC に基づく最適な shapelets の個数は本質的に少ない傾向があることを示している．6 章で示すように，この傾向は shapelets の説明性という観点から望ましい．

5.4　合計学習時間への効果

表 1 は，Xeon 2.9 GHz CPU のシングルスレッドで実行された各 shapelet 手法の 71 種類のデータセット（D1&2）全体にわたる合計学習時間（単位：時間）を示している．TSL は従来の shapelet 手法である CSLTS，LTS，LTSED と比較してそれぞれ 95%，74%，74%の時間を短縮している．これにより，TSL ではハイパーパラメータを設定するコストを削減しながら shapelets の長さを効率的に学習できることが分かる．

5.5　時系列データの長さを変えた場合への影響

3 章で提案した連続最適化による shapelets の長さの学習効果を確認するため，shapelets の長さを離散的に探索する Algorithm 2（Alg.2）と提案手法（TSL）を比較する．図 1 の時系列データからノイズ部分の長さを変化させた場合で，各手法におけるテストデータでの AUC と発見された shapelets の長さとを測る．図 5 から，時系列データが長くなるにつれて，Alg.2 は TSL よりも AUC と shapelets の長さが悪化する．これにより，提案する連続最適化は特に長い時系列データに有効であることが分かる．

6　ケーススタディ

本章では，UCR 公開時系列データセット [18] の中で特に shapelets の正解がある程度分かっている 3 つのケーススタディを通して TSL で学習された shapelets の説明性を評価する．

6.1　GunPoint データセットの場合

このデータセットでは，人が銃を向ける動作（クラス 1）と指をさす動作（クラス -1）を分類する．TSL により発見された shapelets の個数は 2 であるため，式 (7) の距離 $x_{i,k}$ を各軸とする特徴空間は図 6 (a) に示すように 2 次元となる．クラスが

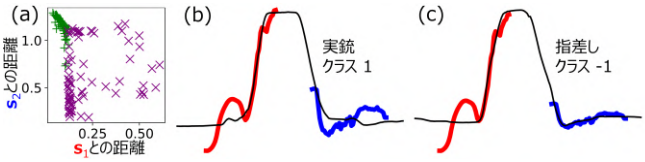


図 6　GunPoint データセットにおける TSL の学習結果.

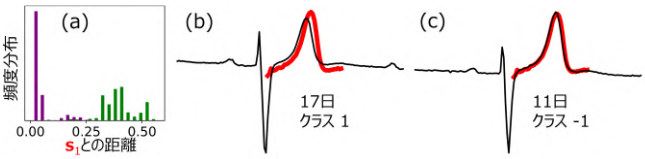


図 7　ECGFiveDays データセットにおける TSL の学習結果.

1 と-1 のテストデータの特徴ベクトルをそれぞれ緑の “+” と紫の “×” でプロットしている．AUC も 1.0 であり，TSL はたった 2 個だけの shapelets で殆どのテストデータを分離できる．

図 6 (b) と (c) では，それぞれクラス 1 と-1 の時系列インスタンスを黒線で描き，それらにマッチングする位置に TSL により学習した shapelets を重ね描きしている．以降では，$\mathbf{s}_1$ と $\mathbf{s}_2$ とをそれぞれ赤太線と青太線で描く．この波形には，ホルスターから銃を抜く時に突起が現れ，指差しを終えて腕を下げる時にオーバーシュートにより凹みが現れることが知られている [8, 11]．Shapelets $\mathbf{s}_1$ と $\mathbf{s}_2$ はそれぞれこの突起と凹みを強調して捉えていることが分かる．モデルパラメータ w_1 と w_2 の符号はそれぞれ負と正であることから，定義 1 より $\mathbf{s}_1$ と $\mathbf{s}_2$ のクラスはそれぞれ 1 と-1 である．実際に突起と凹みの波形はそれぞれクラス 1 と-1 に現れるため，この結果は定理 1 と一致する．更に，従来の shapelets 学習法では，より多数の shapelets を用いるか [7, 13, 17, 35]，或いは shapelets の個数や長さを事前に設定する必要があるが [14]，TSL では 2 つの shapelets を過不足なくその長さや個数とともに発見できる．

6.2　ECGFiveDays データセットの場合

このデータセットでは，ある男性の心電図が 11 日（クラス -1）か 17 日（クラス 1）に計測されたかを分類する．TSL により発見された shapelet の個数は 1 であり，図 7 (a) はテストデータにおけるクラス 1（緑）とクラス -1（紫）の特徴量の分布を示す．AUC も 1.0 であり，TSL は殆どのテストデータを単一の shapelet で分離できることが分かる．図 7 (b) と (c) は，それぞれクラス 1 と-1 の時系列インスタンスとそれらにマッチングする shapelet $\mathbf{s}_1$ を示している．モデルパラメータ w_1 の符号は正であり，定義 1 から $\mathbf{s}_1$ のクラスは-1 である．実際に $\mathbf{s}_1$ はクラス -1 の波形に似ており，この結果は定理 1 と一致する．このデータセットで医学的に重要な違いは専門医が特定した T 波だけであることが知られている [36]．従来手法 [7, 13, 14, 17, 35, 37] では多数の冗長な shapelets を学習するが，TSL では適切な長さを持つ 1 つの shapelet として T 波を発見できる．

6.3　Chinatown データセットの場合

このデータセットでは，Chinatown の 1 時間毎の歩行者数を

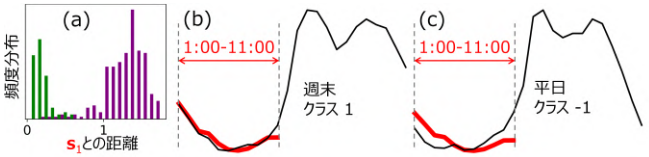


図 8　Chinatown データセットにおける TSL の学習結果.

表す時系列インスタンスが平日（クラス -1）のものか週末（クラス 1）のものかを分類する．TSL により発見された shapelet の個数は 1 であり，図 8 (a) はテストデータにおける特徴量の 2 クラス間の分布を示す．AUC も 0.99 であり，TSL は殆どのテストデータを単一の shapelet だけで分離できる．図 8 (b) と (c) は，それぞれクラス 1 と-1 の時系列インスタンスとそれらにマッチングする shapelet $\mathbf{s}_1$ を示す．モデルパラメータ w_1 の符号は負（つまり $\mathbf{s}_1$ のクラスは 1）であり，実際に $\mathbf{s}_1$ はクラス 1 の波形に似ているため，この結果は定理 1 と一致する．このデータセットでは，週末の深夜は歩行者が多く平日の深夜は人が少ない [38]．TSL はこの違いを 1 個の shapelet として発見できており，それは午前 1 時から午前 11 時頃の傾向によく一致する．

7　関 連 研 究

本章では関連研究とそれらの提案手法との違いを紹介する．

7.1　Shapelets を用いない時系列分類法

1 つ目のアプローチでは，動的時間伸縮法（DTW）などを使用し時系列インスタンス間で距離を計算する [39, 40, 41, 42]．しかし，このアプローチは，最近の他のアプローチに比べて計算効率が悪く正解率も低い傾向がある [43]．HIVE-COTE とその亜種 [44, 45, 46] は，様々な分類手法のアンサンブル学習で高い正解率を達成するが，計算コストが高く説明性を提供しない．近年，分類に効く区間を特定することで説明性を提供するアンサンブル学習法が提案されたが [47]，区間の境界は曖昧であり各区間で多様な特徴量を計算するため説明性は低い．深層学習によるアプローチでは文献 [48, 49] が高い正解率を達成し，時系列データから分類に効く箇所も特定できる [48]．しかし，文献 [48] の説明性では元のアーキテクチャを変更する必要がある [48]．また，学習したモデルを事後的に解釈する研究では [50, 51, 52, 53, 54]，元のモデルを別の説明可能なモデルで近似する．そのため，これらの説明性は元のモデルの予測プロセスを忠実に捉えられない．Rocket [30] のような多数の固定値の畳み込みカーネルから高次元特徴量を作る手法は正解率が高く高速だが，これらのモデルには説明性がない．時系列データを記号で近似するアプローチでは BOSS [31] や WEASEL [32] が良く知られた手法であるが，これらのモデルにも説明性はない．

一方，高品質な少数の shapelets を発見する TSL では，分類性能を低下させずに元のモデルに忠実で完全な説明性を提供する．

7.2　Shapelet 手法

Shapelet 手法は大きく探索型と学習型に分けられる．探索型 [8, 11, 12, 27, 36] では，時系列データセットの部分時系列を網羅的またはランダムに探索する．ランダム探索では長い時系列に対して最適な shapelets とそれらの長さを見つけることが難しく，網羅的探索では時系列の長さに対して計算量が著しく増大する．また，shapelets の形は学習データに現れる波形に制限され元の波形のノイズの影響も受けやすい．学習型では [7, 13, 14, 16, 17, 20, 21, 22, 23, 24, 33]，shapelets が任意の形状をとれるように連続最適化により学習することで，クラス間で分離性の高い shapelets を効率的に探索する．

しかし，shapelets の長さを学習することは基本的な課題であるにも関わらず，直接挑戦した従来研究はない．学習前にヒューリスティックに shapelets の長さを選択する従来手法 [17] とは異なり，提案手法では各 shapelet の長さと形をそれらの相互依存性を考慮した連続最適化問題からエンドツーエンドで学習する．

8　ま　と　め

本研究では長さを学習可能な shapelets を導入した．提案手法では，shapelets とそれらの長さと分類器を同時に学習する 1 つの連続最適化問題を数学的に定式化し，SGD を用いて計算量を増加させずに解く．特に，この shapelets の長さの学習は shapelets を対象クラスの波形に類似させながらそれ以外の波形から乖離させる．これにより，説明性とクラス間分離性において高品質な shapelets が生成されることを理論的に保証する．また，オプションとして BIC に基づき（大抵少数個となる）shapelets の個数を特定する方法も提案した．UCR 公開時系列データセットを用いて，AUC，合計学習時間，shapelets の説明性における提案手法の有効性を示した．

文　　献

[1] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. “Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures”. In: *Proc. VLDB Endow.* (2008).

[2] B. Goodman and S. R. Flaxman. “European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation””. In: *AI Mag.* (2017).

[3] 総務省. **国際的な議論のための AI 開発ガイドライン案**. `www.soumu.go.jp/main_content/000490299.pdf`. 2017.

[4] 人間中心の AI 社会原則検討会議. **人間中心の *AI* 社会原則**. `www8.cao.go.jp/cstp/tyousakai/humanai/`. 2019.

[5] M. Du, N. Liu, and X. Hu. “Techniques for Interpretable Machine Learning”. In: *Commun. ACM* (2019).

[6] C. Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nat. Mach. Intell.* (2019).

[7] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme. “Learning Time-series Shapelets”. In: *KDD*. ACM, 2014.

[8] L. Ye and E. Keogh. “Time Series Shapelets: A New Primitive for Data Mining”. In: *KDD*. ACM, 2009.

[9] E. Matsumoto, K. Uchida, M. Saito, A. Yamaguchi, and T. Maekawa. “Recent Digitization of GIS and Sophistication

of Equipment Condition Monitoring and Diagnosis applying AI Technologies". In: *CIGRE Paris*. CIGRE, 2022.

[10] A. Yamaguchi, K. Ueno, K. Uchida, E. Matsumoto, and T. Saida. "Development of advanced AI technologies for condition diagnosis of high voltage switchgear in substations". In: *CIGRE Science and Engineering* (2022).

[11] J. Lines, L. M. Davis, J. Hills, and A. Bagnall. "A Shapelet Transform for Time Series Classification". In: *KDD*. ACM, 2012.

[12] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. J. Bagnall. "Classification of time series by shapelet transformation". In: *Data Min. Knowl. Discov.* (2014).

[13] S. Roychoudhury, M. Ghalwash, and Z. Obradovic. "Cost Sensitive Time-Series Classification". In: *ECML PKDD*. Springer, 2017.

[14] A. Yamaguchi, K. Ueno, and H. Kashima. "Learning Time-series Shapelets Enhancing Discriminability". In: *SDM*. SIAM, 2022.

[15] Q. Zhang, J. Wu, H. Yang, Y. Tian, and C. Zhang. "Unsupervised Feature Learning from Time Series". In: *IJCAI*. AAAI Press, 2016.

[16] Q. Zhang, J. Wu, P. Zhang, G. Long, and C. Zhang. "Salient Subsequence Learning for Time Series Clustering". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2018).

[17] Z. Fang, P. Wang, and W. Wang. "Efficient Learning Interpretable Shapelets for Accurate Time Series Classification". In: *ICDE*. IEEE Computer Society, 2018.

[18] H. A. Dau, E. Keogh, K. Kamgar, C.-C. Yeh Michael, Y. Zhu, et al. *The UCR Time Series Classification Archive*. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`. 2018.

[19] A. Yamaguchi, K. Ueno, and H. Kashima. "Time-Series Shapelets with Learnable Lengths". In: *CIKM*. ACM, 2023.

[20] A. Yamaguchi and K. Ueno. "Learning Time-series Shapelets via Supervised Feature Selection". In: *SDM*. SIAM, 2021.

[21] S. Roychoudhury, F. Zhou, and Z. Obradovic. "Leveraging Dependencies among Learned Temporal Subsequences". In: *SDM*. SIAM, 2022.

[22] H. Wang, Q. Zhang, J. Wu, S. Pan, and Y. Chen. "Time series feature learning with labeled and unlabeled data". In: *Pattern Recognit.* (2019).

[23] A. Yamaguchi, S. Maya, K. Maruchi, and K. Ueno. "LTSpAUC: Learning Time-series Shapelets for Optimizing Partial AUC". In: *SDM*. SIAM, 2020.

[24] A. Yamaguchi, K. Ueno, and H. Kashima. "Learning Evolvable Time-series Shapelets". In: *ICDE*. IEEE Computer Society, 2022.

[25] S. Han, Z. Meng, Z. Li, J. O'Reilly, J. Cai, et al. "Optimizing Filter Size in Convolutional Neural Networks for Facial Action Unit Recognition". In: *CVPR*. IEEE Computer Society, 2018.

[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. "LIBLINEAR: A Library for Large Linear Classification". In: *J. Mach. Learn. Res.* (2008).

[27] J. Grabocka, M. Wistuba, and L. Schmidt-Thieme. "Fast classification of univariate and multivariate time series through shapelet discovery". In: *Knowl. Inf. Syst.* (2016).

[28] L. Hou, J. T. Kwok, and J. M. Zurada. "Efficient Learning of Timeseries Shapelets". In: *AAAI*. AAAI Press, 2016.

[29] X. Li and J. Lin. "Evolving Separating References for Time Series Classification". In: *SDM*. SIAM, 2018.

[30] A. Dempster, F. Petitjean, and G. I. Webb. "ROCKET: Exceptionally Fast and Accurate Time Series Classification Using Random Convolutional Kernels". In: *Data Min. Knowl. Discov.* (2020).

[31] P. Schäfer. "The BOSS is concerned with time series classification in the presence of noise". In: *Data Min. Knowl. Discov.* (2015).

[32] P. Schäfer and U. Leser. "Fast and Accurate Time Series Classification with WEASEL". In: *CIKM*. ACM, 2017.

[33] A. Yamaguchi, S. Maya, and K. Ueno. "RLTS: Robust Learning Time-series Shapelets". In: *ECML PKDD*. Springer, 2020.

[34] J. Demšar. "Statistical Comparisons of Classifiers over Multiple Data Sets". In: *J. Mach. Learn. Res.* (2006).

[35] Q. Ma, W. Zhuang, S. Li, D. Huang, and G. Cottrell. "Adversarial Dynamic Shapelet Networks". In: *AAAI*. AAAI Press, 2020.

[36] E. Keogh and T. Rakthanmanon. "Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets". In: *SDM*. SIAM, 2013.

[37] Q. Ma, W. Zhuang, and G. Cottrell. "Triple-Shapelet Networks for Time Series Classification". In: *ICDM*. IEEE Computer Society, 2019.

[38] W. Tang, L. Liu, and G. Long. "Interpretable Time-series Classification on Few-shot Samples". In: *IJCNN*. IEEE Computer Society, 2020.

[39] P.-F. Marteau. "Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2009).

[40] A. Stefan, V. Athitsos, and G. Das. "The Move-Split-Merge Metric for Time Series". In: *IEEE Trans. Knowl. Data Eng.* (2013).

[41] E. J. Keogh and M. J. Pazzani. "Derivative Dynamic Time Warping". In: *SDM*. SIAM, 2001.

[42] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. J. Keogh. "Indexing Multidimensional Time-Series". In: *VLDB J.* (2006).

[43] A. J. Bagnall, J. Lines, A. Bostrom, J. Large, and E. J. Keogh. "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances". In: *Data Min. Knowl. Discov.* (2017).

[44] J. Lines, S. Taylor, and A. J. Bagnall. "HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles for Time Series Classification". In: *ICDM*. IEEE Computer Society, 2016.

[45] J. Lines, S. Taylor, and A. Bagnall. "Time Series Classification with HIVE-COTE: The Hierarchical Vote Collective of Transformation-Based Ensembles". In: *ACM Trans. Knowl. Discov. Data* (2018).

[46] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall. “HIVE-COTE 2.0: A New Meta Ensemble for Time Series Classification”. In: *Mach. Learn.* (2021).

[47] N. Cabello, E. Naghizade, J. Qi, and L. Kulik. “Fast and Accurate Time Series Classification Through Supervised Interval Search”. In: *ICDM*. IEEE Computer Society, 2020.

[48] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. “Deep Learning for Time Series Classification: A Review”. In: *Data Min. Knowl. Discov.* (2019).

[49] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, et al. “InceptionTime: Finding AlexNet for Time Series Classification”. In: *Data Min. Knowl. Discov.* (2020).

[50] J. Crabbé and M. van der Schaar. “Explaining Time Series Predictions with Dynamic Masks”. In: *ICML*. PMLR, 2021.

[51] J. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro. “TimeSHAP: Explaining Recurrent Models through Sequence Perturbations”. In: *KDD*. ACM, 2021.

[52] P. S. Parvatharaju, R. Doddaiah, T. Hartvigsen, and E. A. Rundensteiner. “Learning Saliency Maps to Explain Deep Time Series Classifiers”. In: *CIKM*. ACM, 2021.

[53] R. Doddaiah, P. S. Parvatharaju, E. A. Rundensteiner, and T. Hartvigsen. “Class-Specific Explainability for Deep Time Series Classifiers”. In: *ICDM*. IEEE Computer Society, 2022.

[54] A. Sood and M. Craven. “Feature Importance Explanations for Temporal Black-Box Models”. In: *AAAI*. AAAI Press, 2022.

合成シェイプレットを用いた時系列データ分類

岡部　正幸†

† 県立広島大学　〒 734–8558 広島県広島市南区宇品東
E-mail: †okabe@pu-hiroshima.ac.jp

あらまし　時系列データ分類方法の 1 つに，シェイプレット (shapelet) と呼ばれるデータ中の特徴的な変動を示す部分的なデータ系列を利用する方法がある．この方法では，データ内における複数の shapelet の出現情報を利用して分類を行うが，shapelet 間の出現順序については考慮されない．本研究では，L1 正則化付き分類器を用いて選定された shapelet を組み合わせて新たな shapelet を合成することで，shapelet 間の出現順序および出現間隔を特徴として取り入れる方法を提案する．合成対象となる shapelet の組み合わせ数を抑制するため，提案手法では 2 つのことを行う．1 つは同一データから選定された shapelet のみを合成対象とすることである．もう 1 つは同一データから選定されたすべての shapelet を 1 つの新たな shapelet として合成することである．これにより shapelet 数の増大を抑制する．実験では，UCR Time Series Classification Archive から選んだ 112 個のデータセットを用いて提案手法による分類を行い，合成 shapelet の導入により分類精度が向上することを示す．

キーワード　時系列データ分類，L1 正則化，合成シェイプレット

1　はじめに

時系列データ分類 (Time Series Classification, TSC) は，数値列に対しクラスラベルを付与する教師あり機械学習問題の 1 つであり，多くの場合任意の長さを持つ数値列から一定長に切り出した部分系列の集合を学習用データとして用いる．TSC の適用対象となるデータは，脳波，心電図，行動認識用センサデータ，画像の輪郭，スペクトログラム，光度曲線，交通量や電力使用量の推移など広範囲に渡る [1]．

TSC に有効なアプローチには，様々な統計量による数値列の要約 [2]，シェイプレット (shapelet) に基づく手法 [3]，畳み込み特徴量を用いた手法 [4]，ハイブリッド手法 [5]，ディープラーニングを用いた手法 [6] などがある．本研究では，このうちの shapelet に基づく手法を扱う．shapelet に基づく手法では，データ中に現れる特徴的な変動を利用して分類を行う．この特徴的な変動を示す部分系列のことを shapelet と呼び，様々な変動パターンをもつ複数の shapelet を生成し，それらを用いて各データを特徴付けることで分類を実現する．具体的には，各データの特徴ベクトルは，そのデータと各 shapelet との間で計算された最小距離を要素として並べたものであり，データ中における各 shapelet の出現の有無または出現する度合いを表したものと捉えることができる．つまり，shapelet に基づく手法では各 shapelet の出現の有無および共起関係を特徴として分類を行っていると解釈できる．

一方，分類時に利用可能な特徴は各 shapelet の出現の有無および共起関係だけでなく，出現場所，出現順序，出現間隔といった特徴も利用可能である．例えば，行動認識のようなタスクでは，shapelet が示す動作の順序 (例えば，右手を上げた後に右手を下げるなど) によって全体の動作が異なる場合，出現の有無および共起関係だけでは特徴として不十分で，出現順序を考慮することで分類がうまく行えるようになると考えられる．このうち，出現場所を特徴として利用した研究 [7] はあるが，出現順序と出現間隔を陽に考慮した方法はこれまでに提案されていない．本研究では，この出現順序と出現間隔を特徴として取り入れる方法を提案する．

具体的には，まず各データから shapelet 候補をランダムに抽出し，L1 正則化付き分類器を用いて分類に有効なものを選別する．次に，同じデータから選別された shapelet を集め，それらを合成することで新たな shapelet を生成する．一般に，合成可能な shapelet の出現状態は選別された shapelet 数に応じて増大するため，本研究では，合成対象となる shapelet を同じデータから選別されたものに限定する．更に同じデータから選別されたすべての shapelet を 1 つの新たな shapelet として合成することで，shapelet 数の増大を抑制する．合成対象となる shapelet は同じデータから選別されているためデータ内における出現場所が分かっており，shapelet 間の出現順序および出現間隔を保ったまま合成処理を行うことができる．

shapelet の合成は，2 つのパターンで行われる．1 つ目は，合成対象となる 2 つの shapelet の出現場所が重複している場合である．このパターンでは 2 つの shapelet の併合処理が行われ，より長さの大きな shapelet が新たに生成される．2 つ目は，合成対象となる 2 つの shapelet の出現場所が重複していない場合である．このパターンでは 2 つの shapelet の連結処理が行われ，間に空白区間を持つより長さの大きな shapelet が新たに生成される．空白区間を持つ合成 shapelet の最小距離の計算では，空白区間の距離の加算は行わない．空白区間は合成前の shapelet の出現順序と出現間隔を表しており，従前と同様の最小距離の計算を行うことで，複数の shapelet 間の順序関係を伴う出現パターンを特徴として取り込むことが可能となる．

shapelet 集合の生成は，大きく分けて訓練データ集合から分類に有効な変動を示す箇所を部分的な数値列として切り出す

アプローチ [11] と，事前に用意した shapelet 集合に対し，各 shapelet の変動パターンに修正を加え分類精度を向上させるアプローチ [9] の 2 つに分けられる．本研究のアプローチは，基本的には前者の分類に有効な変動を切り出すアプローチをベースとしているが，それらを連結して空白区間を伴う新たな shapelet の生成も行なっている点がこれまでのアプローチと異なる．空白区間を伴う最小距離の計算は，dilation を用いた shapelet ベースの手法 [7] と似ているが，dialtion では shapelet を構成する数値の間に一定の長さの空白区間を設けることにより，shapelet を拡張表現するのに対し，本研究が提案する合成 shapelet では空白区間の長さは様々であり，その目的も shapelet 間の出現順序と出現間隔を表す点が異なる．

2 Shapelet に基づく時系列データ分類

本章では，まず各データを表す記号について説明し，次に shapelet の選定と shapelet を利用した分類学習法について説明する．

2.1 時系列データと shapelet

時系列データの集合 T および各データ T_i を以下のように表す．

$$T = \{T_1, T_2, \ldots, T_N\}, \qquad T_i = (t_{i,1}, t_{i,2}, ..., t_{i,Q})$$

とする．ここで，N はデータ数，Q はデータの長さを表す．T_i にはそれぞれクラスラベルが与えられており，TSC におけるタスクは各 T_i のクラスラベルを予測することである．また，shapelet 集合 S とその要素 S_i を以下のように表す．

$$S = \{S_1, S_2, ..., S_M\}, \qquad S_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,L})$$

ここで，M は選定された shapelet の数，L は shapelet の長さ，$s_{i,j}$ は S_i の j 番目の値を表している．S_i は任意のデータ T_i から切り出した長さ L の部分系列の中で分類に有効なものとして選定されたものである．また，長さ L は一定とは限らず，手法によっては様々な長さの shapelet を混在して利用するものもある．shapelet と同様に，時系列データ T_i の任意の部分系列 $T_{i,j}$ を以下のように表す．$T_{i,j}$ は T_i の j 番目の値から切り出した長さ L の部分系列を表す．

$$T_{i,j} = (t_{i,j}, \ldots, t_{i,j+L-1})$$

shapelet は分類における特徴であり，その選択方法として一般にフィルター法とラッパー法がある．フィルター法では，特徴の有効性を様々な尺度を用いて事前に評価するが，shapelet に基づく方法では，情報利得，クラスカル・ウォリス検定，F 統計量，Mood のメディアン検定などの評価尺度が使用されている [3]．これらの尺度の多くは各 shapelet と各データ間の最小距離および各データのクラス属性を利用することで，全データ集合と各クラスのデータ集合における最小距離 (またはその順位) の分布の偏りを定量化しており，偏りが大きいものを有効性が高いものとして評価する．

一方，ラッパー法では実際に学習を行うことによって特徴の有効性を判定する．例えば，Grabocka らによる方法 [8] では，shapelet 候補を 1 つ追加するごとに k-近傍法を実行し，精度向上の有無を確認することにより，その shapelet 候補の有効性を判定する．この作業をすべての候補について行い，shapelet 集合を選別する．一般にラッパー法はフィルター法に比べ計算コストが高いため，k-近傍法のように学習速度が高速なものが望ましいとされている．Grabocka らによる方法はラッパー法による shapelet の選別を高速に行うことを可能にしたものの，切り出した shapelet 候補集合の有効性を検証する際の順番について考慮されておらず，順番を変更すると選択される shapelet 集合が変化し，それに伴い分類精度も大きく変化し安定しないという現象が我々が行った予備実験において確認された．

一方，L1 正則化付きロジスティック回帰はラッパー法に分類される特徴選択方法としてよく知られている．この方法は L1 正則化を伴うことで予測式における各特徴の重みが 0 になりやすいという性質を利用して，重みが 0 になった特徴を削除するというものである．L1 正則化では正則化係数の設定によって選定される特徴集合が変わるため，その調整作業が必要となるが，他の手法と比較して高速かつ精度が安定しているため，本研究ではこれを特徴選択方法として採用する．アルゴリズムの詳細については，第 3 章で述べる．

2.2 Shapelet Transform Classification

Shapelet Transform Classification (STC) [3] は，Hills らによって提案された時系列データ分類方法であり，この方法では，shapelet 集合を用いて時系列データを特徴ベクトルに変換する．ここでは，shapelet を用いた特徴量の計算方法について説明する．

時系列データ T_i の変換後の特徴ベクトルを X_i を以下のように表す．X_i の次元数は shapelet 数と同じである．

$$X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,M})$$

X_i の各要素 $x_{i,m}$ の計算式は以下の通りである．

$$x_{i,m} = \min_{j=1\sim Q-L+1} \sqrt{\sum_{l=1}^{L} (T_{i,j+l-1} - s_{m,l})^2} \qquad (1)$$

$x_{i,m}$ は，T_i と S_m の最小距離である．この値が 0 に近ければ，S_M が示す値の変動が T_i に出現することを意味する．STC は各 shapelet が示す変動の有無に基づいて判定を行う方法であるため，クラス間の分離性能の高い shapelet の選別が重要となる．

特徴ベクトルが生成できれば，分類器には任意のものを用いることができる．Hills らは，など 7 つの分類手法を比較し，SVM(線形カーネル)，ベイジアンネットワーク，ローテーションフォレストが高い性能を示したと報告している．また，計算コストを重視するアルゴリズム [4] では，ロジスティック回帰が用いられることも多い．本研究では，分類精度と計算時間の観点からロジスティック回帰の有用性が高いと判断し，特に L2 正則化付きロジスティック回帰によって分類器の生成を行う．

3 合成 shapelet による時系列データ分類

ここでは，本研究で提案する合成 shapelet の生成手順について説明する．

3.1 データ分類の流れ

まず最初に，本研究で提案する時系列データ分類の流れについて説明する．処理手順は以下の通りである．

（1） 基本 shapelet 集合の選定

（2） 合成 shapelet 集合の生成

（3） L2 正則化付きロジスティック回帰によるデータ分類

(1) では，データ集合から shapelet 候補をランダムに切り出し，L1 正則化付きロジスティック回帰を用いて分類に有効なものを基本 shapelet 集合として選定する．(2) では，基本 shapelet 集合の各要素の抽出元データに関する情報を元に，同一データ内から選定された shapelet 集合に対して合成処理を行う．(3) では，(2) の合成処理によって生成された shapelet 集合と基本 shapelet 集合の両方を用いて各データを Shapelet Transform により特徴ベクトルに変換する．この特徴ベクトルから L2 正則化付きロジスティック回帰による分類器を生成し，データ分類を行う．以下では，(1) と (2) の各処理について詳しく説明する．

3.2 基本 shapelet 集合の選定

基本 shapelet 集合の選定アルゴリズムを Algorithm 1 に示す．このアルゴリズムは，時系列データ集合 T を入力とし，選定された基本 shapelet 集合 S^b および S^b の各要素の各クラス分類時における重みを格納した行列 W^b を出力として返す．まず最初に shapelet の候補となる部分系列を各データからランダムに抽出する．抽出は重複しないように行われる．また，抽出開始位置の前後 2 つ分の位置は後の抽出開始位置として選択できないようにすることで類似の shapelet 候補が抽出されないようにする．本研究では，合成 shapelet 生成時に各基本 shapelet がどのデータのどの位置から抽出されたものであるかについての情報を利用するため，これを保存しておく．

各データから抽出する shapelet 候補の数は，shapelet 候補数の最大値をデータ数で割った値 $\lfloor M_{max}/N \rfloor$($\lfloor x \rfloor$ は x を超えない最大の整数とする) とデータ長を基本 shapelet の長さで割った値の 2 倍 $\lfloor Q/L^b \rfloor * 2$ の大きい方とする．M_{max} については，値を大きくし特徴数を増やすことで分類精度の向上が期待できるが，その分最小距離の計算と学習に要するコストが増えるため，データ数とデータ長および計算コストを考慮して設定する必要がある．

shapelet 候補集合 C の抽出が終了すると，これを用いて各データ T_i を特徴ベクトル X_i に変換する．X_i の各要素 $x_{i,j}$ は (1) 式によって計算される T_i との最小距離である．各データの特徴ベクトルへの変換が終了すると，これを用いて分類学習を行う．この分類学習は基本 shapelet 集合選定のためのもので，L1 正則化付きロジスティック回帰を用いて行う．L1 正則化付きロジスティック回帰では，正則化係数 λ の値によって重みが

Algorithm 1 基本 shapelet 集合選定アルゴリズム

入力: 時系列データ集合 T，クラス数 K，
　　正則化係数 λ，
　　最大 shapelet 候補数 M_{max}，
　　基本 shapelet の長さ L
出力: 基本 shapelet 集合 S^b，重み行列 W^b
$S^b \leftarrow \emptyset$
$C \leftarrow \emptyset$
for $i = 1$ to N **do**
　$n \leftarrow \max\left(\lfloor M_{max}/N \rfloor, \lfloor Q/L^b \rfloor * 2\right)$
　for $j = 1$ to n **do**
　　T_i から長さ L の部分系列をランダムに抽出し，C に加える
$X \leftarrow \emptyset$
for $i = 1$ to N **do**
　$X_i \leftarrow \emptyset$
　for $j = 1$ to $|C|$ **do**
　　C の要素 c_j と T_i との最小距離 $x_{i,j}$ を (1) 式を用いて
　　計算し，X_i の要素として加える
　$X \leftarrow X \cup \{X_i\}$
X を用いて，L1 正則化付ロジスティック回帰学習を実行する
学習結果から各 shapelet 候補の重み行列 W を取得する
for $j = 1$ to $|C|$ **do**
　if W の j 列の値に 0 より大きいものが 1 つ以上ある **then**
　　$S^b \leftarrow S^b \cup \{c_j\}$
W から shapelet として選定された候補に対応する列だけを残した行列を W^b とする
return S^b, W^b

0 となる特徴の数が変化する．一般に，λ の値を小さくすると正則化の影響が小さくなり重みが 0 となる特徴数が減る．つまり，選別される基本 shapelet 数は増加する．逆に，λ の値を大きくすると選別される基本 shapelet 数は減少する．λ の適正値はデータに依存して決まるため，本研究では正則化係数 λ と基本 shapelet の長さ L^b については事前に交差検証を用いて設定値を決定する．

基本 shapelet の選別は L1 正則化付きロジスティック回帰によって生成されたモデルの重み行列 W を用いて行う．W は以下に示すように，クラス数 K を行数，shapelet 候補数を列数とする行列である．ただし，$K = 2$ の場合は行数は 1 である．

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1|C|} \\ w_{21} & w_{22} & \cdots & w_{2|C|} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1} & w_{K2} & \cdots & w_{K|C|} \end{pmatrix}$$

W の列ベクトルは，データのクラス所属確率を計算する際の重みを表している．列ベクトルの値がすべて 0 である場合は対応する shapelet 候補が分類判定時に利用されないことを意味するため，shapelet として選別されない．列ベクトルの値に 0 より大きいものが 1 つ以上あれば，その shapelet 候補は分類に有効な shapelet として選択される．

すべての shapelet 候補の選別終了後，W の中から shapelet として選別されたものに対応する列のみを残し，W^b として保

存しておく．本研究では W^b を合成 shapelet の生成時に利用する．

各 shapelet は shapelet 合成のことを考え，以下の 4 つの情報で表現される．

（1） 抽出元のデータ番号
（2） 出現場所
（3） 長さ
（4） マスク情報

(4) のマスク情報は，0 と 1 を要素とする (3) と同一の長さを持つベクトルで表現され，当該 shapelet がその位置に値を持つかどうかを表している．値を持てば 1 を，値を持たなければ 0 がマスク値となる．次章で導入する合成 shapelet では，複数の基本 shapelet が間を空けて連結されたものが新たな shapelet になる場合もあるため，shapelet 内の値のない部分をこのマスク情報で表現している．

3.3 合成 shapelet 集合の生成

基本 shapelet 集合の選別が完了すると，続いて合成 shapelet の生成を行う．shapelet の合成は同一データ内に出現する基本 shapelet 集合に対して行われるが，すべての出現パターンについて合成 shapelet を生成することは現実的でないため，本研究では，以下の 2 つのアプローチによって生成数を抑制する

- 合成対象となる shapelet を同じデータから選別されたものに限定する．
- 同じデータから選別されたすべての shapelet を 1 つの新たな shapelet として合成する

具体的なアルゴリズムを Algorithm 2 に示す．このアルゴリズムでは，クラスごとに各データから 1 つずつ合成 shapelet を生成する．ただし，2 クラス分類問題の場合には 5 行目のループは 1 回のみとする．各クラス k に対する処理では，まずデータがクラス k に属する確率を計算する際に利用する基本 shapelet 集合 C^k を求める．C^k の各要素は重み行列 W^b の第 k 行目の重みのうち $w^b_{kj} > 0$ となる基本 shapelet S^b_j である．この C^k の中から更にデータ毎に，データ T_i を抽出元とする基本 shapelet 集合 C^{ki} を求める．C^{ki} の要素はクラス k に属するデータを判定する際の重みが 0 より大きな値を持ち，かつ抽出元となるデータが T_i となる基本 shapelet である．この C^{ki} の要素をすべて合成し，1 つの新たな shapeletS^c_i として加える．

shapelet の合成は，図 1 に示すように隣り合う 2 つの shapelet に重複箇所がある場合とそうでない場合の 2 つのパターンが存在する．前者の場合の合成では，2 つの shapelet を合併して長さのより大きい shapelet になる．また後者の合成では，2 つの shapelet を間隔を空けて連結された shapelet が生成される．このように，新たに生成される合成 shapelet では途中に値を持たない区間を持つ場合もあるため，各データとの最小距離を計算する際は，以下のようにマスク情報を用いる．

$$x_{i,k} = \min_{j=1\sim Q-L+1} \sqrt{\sum_{l=1}^{L} mask[l] * (T_{i,j+l-1} - s_{k,l})^2} \quad (2)$$

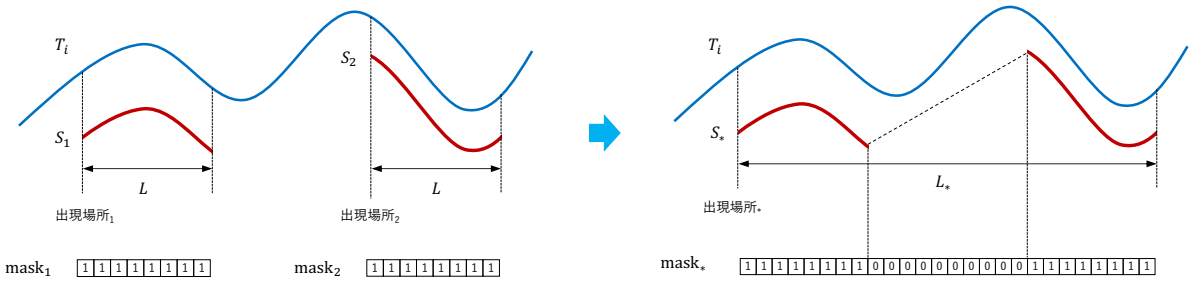
ここで，$mask$ はマスク情報を格納したベクトルで，$mask[l] \in \{0, 1\}$ はその l 番目の値を表している．この間に空白区間をもつ合成 shapelet は空白区間を含めて 1 つの shapelet として扱うため，最小距離の計算は連結前の shapelet それぞれの最小距離を足せば良いわけではなく，新たに計算する必要がある．

Algorithm 2 合成 shapelet 生成アルゴリズム

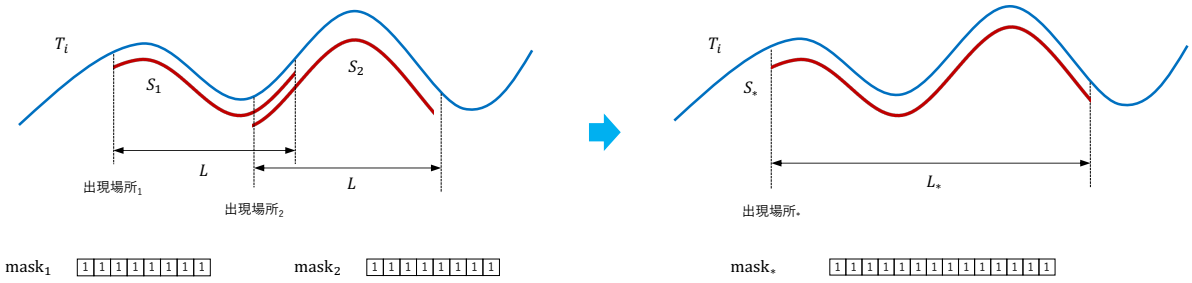
1: **入力**: 基本 shapelet 集合 S^b
2: 重み行列 W^b
3: **出力**: 合成 shapelet 集合 S^c
4: $S^c \leftarrow \emptyset$
5: **for** $k = 1$ to K **do**
6: $C^k \leftarrow \emptyset$
7: **for** $j = 1$ to $|S^b|$ **do**
8: **if** $w^b_{kj} > 0$ **then**
9: $C^k \leftarrow C^k \cup \left\{S^b_j\right\}$
10: **for** $i = 1$ to N **do**
11: $C^{ki} \leftarrow \emptyset$
12: **for** $j = 1$ to $|C^k|$ **do**
13: **if** C^k_j の抽出元データが T_i である **then**
14: $C^{ki} \leftarrow C^{ki} \cup \left\{C^k_j\right\}$
15: C^{ki} から要素 C^{ki}_j を任意に 1 つ取り出す
16: $S^c_{ki} \leftarrow C^{ki}_j$
17: **while** $|C^{ki}| > 0$ **do**
18: C^{ki} から要素 C^{ki}_j を任意に 1 つ取り出す
19: S^c_{ki} と C^{ki}_j を合成したものを新たに S^c_{ki} とする
20: $S^c \leftarrow S^c \cup S^c_{ki}$
21: **return** S^c

以上のように，合併処理によって生成された合成 shapelet では基本 shapelet とは異なる長さの shapelet を新たに生成することができる．また，連結処理によって生成された shapelet により shapelet の出現順序と出現間隔を表現することが可能となる．さらに，最小距離の計算時にマスク情報を追加する点以外は従来の Shapelet Transform と同じ要領で特徴ベクトルへの変換を行うことができる．

合成 shapelet は同一データ内に出現する基本 shapelet の出現パターンに対して生成することが可能であるが，各データにおける shapelet の出現の有無は，shapelet とデータ間の最小距離が 0 であるかどうがで判定できる．この条件を緩和し最小距離が 0 に近い場合も出現しているとすることもできるが，本研究では合成 shapelet 生成数を抑制するためこれを行わない．よって，合成 shapelet の生成対象となるのは同一データ内で最小距離が 0 となる基本 shapelet 集合，つまり同じデータから抽出された基本 shapelet 集合となる．ただし，同一データから抽出された基本 shapelet の数が大きい場合には合成対象となる組み合わせが増大していまうという問題が発生する．例えば，基本 shapelet 数を n とすると合成対象となる組み合わせの数は $2^n - n - 1$ となり，$n = 10$ の場合には 1 つのデータから 1000 以上の新たな shapelet が追加されてしまうことになる．合成対象とする基本 shapelet の数を限定することもできるが，本研究ではアルゴリズムを単純にして高速に実行できることを優先し，



(a) 合成パターン 1



(b) 合成パターン 2

図 1　shapelet 合成時における 2 つのパターン

同一データから抽出された基本 shapelet すべてを合成して新たな shapelet を生成することとした．必要のない基本 shapelet がノイズとして混在してしまう恐れはあるものの，shapelet 数をさらに抑制できるため，このアプローチを採用する．

4　実　　験

本章では，提案手法の有効性を示すために行なった実験について述べる．

4.1　データセット

実験には，文献 [1] で使用された 112 種類の UCR Time Series Classification Archive [1] データセットを用いた．このデータセットは 13 種類のデータタイプから構成されており，データ数とクラス数およびデータ長が大きく異なるものが混在する．表 1 に概要を示す．すべてのデータセットは訓練データとテストデータに予め分割されており，本実験ではこの分割をそのまま用いて学習と評価を行った．

表 1　データセットの概要

	最大値	最小値	平均値
訓練データ数	8926	16	507
テストデータ数	16800	20	1108
クラス数	60	2	8.3
データ長	2844	15	551

4.2　比 較 手 法

まず合成 shapelet の効果を検証するため，以下の 3 つの手法を比較する．

- **Base**: 提案手法における合成 shapelet の生成を行わず，基本 shapelet のみを用いて分類を行う方法．
- **Comp**: 提案手法における合成 shapelet のみを用いて分類を行う方法．
- **CSTC**: 提案手法における基本 shapelet と合成 shapelet の両方を用いた方法．基本 shapelet の長さ (L^b) と L1 正則化係数 (λ) については，各データセットの訓練データを用いて 2 分割交差検証を用いて有効な組み合わせを事前に求めた．L^b は $\{8, 16, 32, 64, 128\}$，λ は $\{0.001, 0.01, 0.1\}$ の中から 1 つずつ選択した．最大 shapelet 候補数 M_{max} には 10000 を設定した．また 3 手法の評価結果は乱数の初期値を変えて行った 10 回の試行の平均値とした．

次に，shapelet に基づく既存手法との比較を行うため以下の 2 つの手法を用いた．

- **STC**: Lines ら [10] が提案した Shapelet Transform に基づく方法である．この方法は，求める shapelet の長さの範囲を設定し，shapelet の候補となる部分系列をランダムに探索する．
- **RDST**: Guillaume ら [7] が提案した dilation を伴う shapelet を利用する方法である．この方法は，dilation の導入により畳み込み処理を用いた方法と同様の高い効果を得ている．モデル構築時に利用する shapelet 数 M に対して，$3M$ 個の特徴量が生成される．

STC と RDST の評価結果は，aeon-toolkit [2]で提供されている get_estimator_results 関数を使用して得られるデータを用いた．aeon-toolkit で提供されている STC モジュールのデフォルト設定値は，shapelet 候補の最大数が 10000，モデル構築の制限

1：https://www.cs.ucr.edu/~Eeamonn/time_series_data_2018/

2：https://www.aeon-toolkit.org/en/stable/

時間は 10000 分となっている．RDST モジュールのデフォルト設定は，最大 shapelet 数が 10000 となっている．

4.3 実験結果

図 2 はデータセット毎に 5 つの手法の分類精度に順位を付けその平均値を比較したものである．太字で示された横棒は，ウィルコクソンの符号順位検定によって 2 つの平均順位に有意差があることを示している．有意基準は 5%である．

まず，提案手法における 3 つの手法については，CSTC が有意差を伴って Base と Comp を上回っており，本研究で提案した合成 shapelet の効果が確認できた．Base と Comp の比較ではほとんど差が見られない結果となった．データによっては Comp のみでも良好な結果を示すものも見られた．次に提案手法と既存手法の比較では，CSTC が STC を上回っているものの有意差は見られない結果となった．また RDST との比較では有意差を伴って下回る結果となった．

図 3 は各データセットにおける分類精度の分布をバイオリンプロットにより示したものである．縦軸の値は各データセットにおける分類精度の中央値からの偏差を表している．CSTC は中央値付近になめらかに分布しており，比較的ばらつきは少ない．ただし，Comp の分布を見て分かるように一部のデータセットにおいて合成 shapelet 単体による結果が著しく悪いものが存在しており，その影響を受けて CSTC にも下方向の外れ値がいくつか存在する．Comp についてはばらつきが大きくデータセットによって効果があるものとないものに分かれていることが確認できる．Base については下方向の外れ値がやや目立つのが確認できる．Base と似ている STC では下方向の外れ値はそれほど確認できないため，パラメータの調整により改善できる可能性がある．CSTC と STC，RDST との差を詳しく調べるため，2 つの手法を個別に比較した．

図 4 は 112 個の各データセットにおける 2 つの手法の優劣をペアワイズ散布図によって示したものである．この図において各点がデータセットを表しており，その座標が 2 つの手法のそのデータセットにおける分類精度を表している．傾き 1 の直線の上側または下側にある点の数が多い方が優位な手法となる．(a) の CSTC と RDST の比較では，半数以上のデータセットにおいて RDST が優位であるとの結果になったが，35 個のデータセットでは RDST を上回る精度を示した．(b) の STC との比較では，CSTC が約半数のデータセット数で分類精度を上回っているものの，分類精度の平均値は STC の方が高い結果となった．これは先に述べたように，CSTC は一部のデータセットにおいて極端に精度の悪いものが見受けられたためと考えられる．ただ，反対に CSTC の方が分類精度を大きく上回ったデータセットも存在しており，合成 shapelet が有効に機能したと考えられる．(c) の Comp との比較では，CSTC が優位である結果となったがそれほど大きな違いは見られなかった．一方，(d) の Base との比較では，CSTC が大きく上回っているデータセットが目立っており，STC との比較と同様に合成 shapelet が精度を押し上げたと考えられる．

図 4(e) は Base と Comp が同等の性能を持つことを示している．後で述べるように，Comp で利用する合成 shapelet 数は Base で利用する基本 shapelet 数と比べて少ないため，shapelet の合成により必要がなくなる基本 shapelet が存在することが推測される．

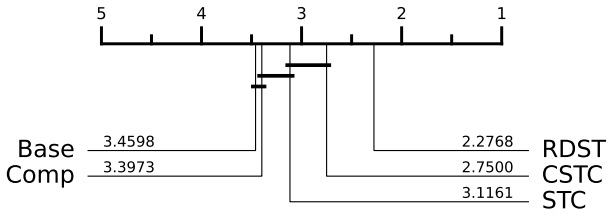


図 2　平均順位の比較

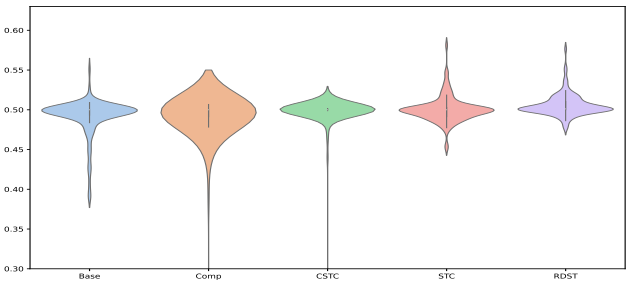


図 3　分類精度のバイオリンプロット

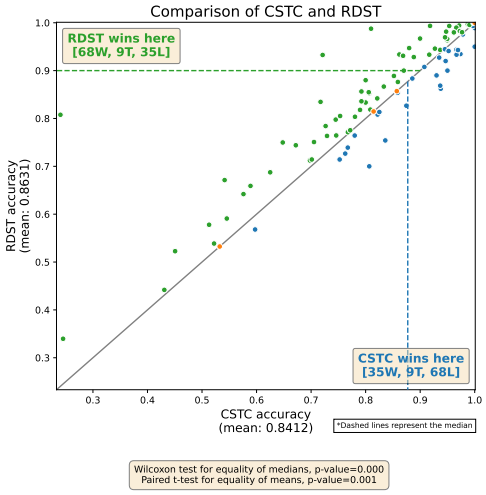
図 4(f) は Base と STC を比較したものである．STC では 1 つのデータセットに対し，様々な長さの shapelet を探索し活用する．これに対し，Base で利用した shaepelt の長さはデータセットによって変えてはいるものの，1 つのデータセットに対しては同じ長さの shapelet しか用いていない．それでも STC の分類精度を上回っているデータセットが少なからずあるため，L1 正則化による特徴選択の効果であるといえる．Base と STC は基本的には同じ処理を行なっているため，Base の精度向上により CSTC の性能の底上げが期待できる．

4.4 合成 shapelet の効果

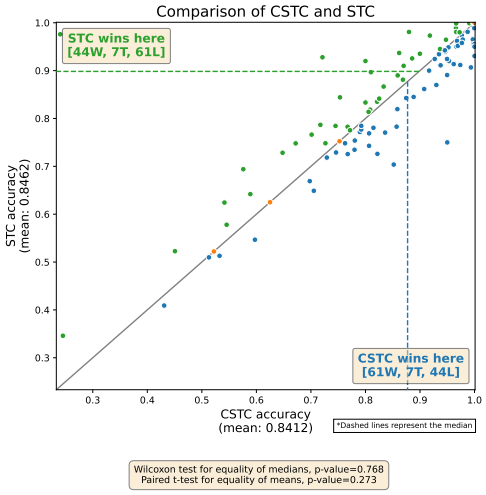
表 2 はデータタイプ別に CSTC とその他の手法の分類精度の勝敗数を調べたものである．例えば，CSTC と Base の Image タイプデータでの比較は，32 個のデータセットのうち 15 個で CSTC の精度が高く，7 個では Base，残りは同値ということを示している．Base との比較では Image，Motion，Simulated，Spectrum などで比較的優位な結果を示している．STC との比較ではそれほど大きな違いはないが，Motion，Spectrum では比較的良い結果となっている．RDST との比較では特に ECG，Spectrum などで若干の優位性が見られるものの，STC との比較で有利だった Motion では水をあけられている．ただし，分類精度の差はそれほど大きくはなく改善の余地は残されている．

4.5 shapelet 数について

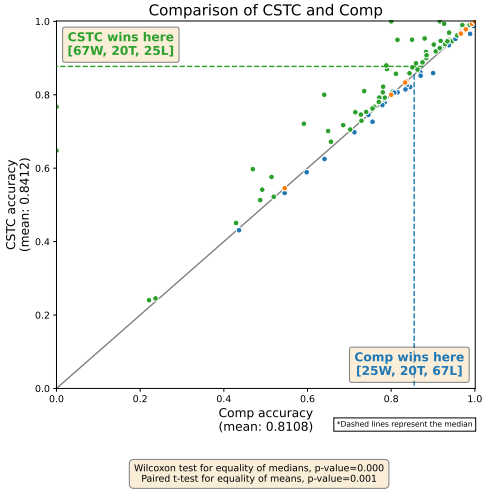
今回の実験では CSTC の基本 shapelet 数の最大値を 10000 に設定した．ただし，CSTC では L1 正則化による特徴選択でモデル構築に利用される shapelet 数は減少するため，実際に各データセットにおいて使用された shapelet 数について調べた．



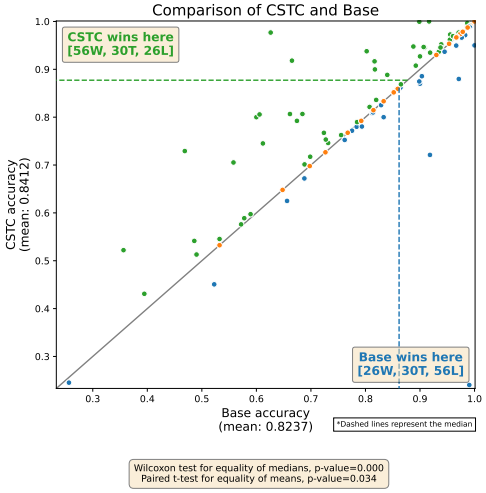
(a) CSTC vs RDST



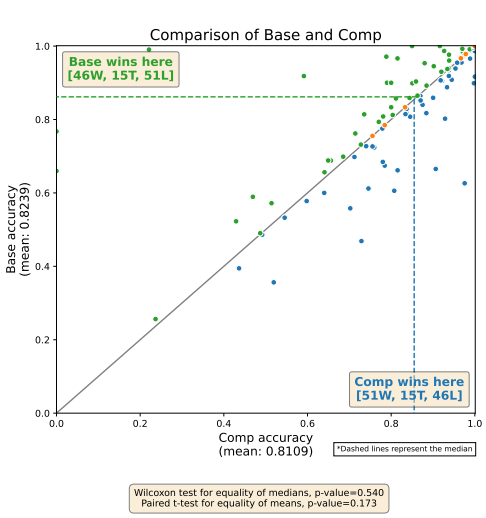
(b) CSTC vs STC



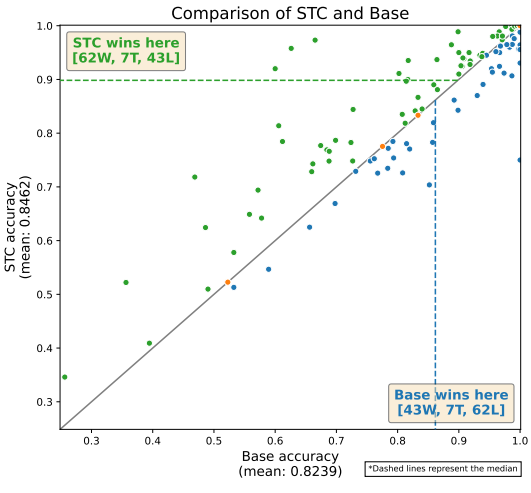
(c) CSTC vs Comp



(d) CSTC vs Base

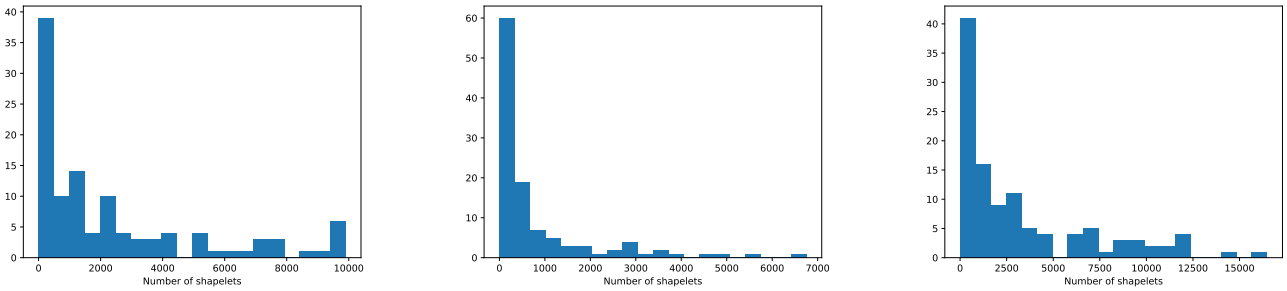


(e) Base vs Comp



(f) STC vs Base

図 4　ペアワイズ散布図による個別比較



(a) Base (b) Comp (c) CSTC

図 5　提案手法のモデル構築時に使用された shapelet 数の度数分布

表 2　データタイプ別の比較

	データ数	vs. Base	vs. STC	vs. RDST
Image	32	15/7/10	19/12/1	6/23/3
Sensor	20	9/8/3	7/11/2	8/10/2
Motion	17	14/0/3	12/5/0	2/14/1
Spectro	8	1/2/5	4/3/1	3/4/1
Simulated	8	4/0/4	6/1/1	4/3/1
Device	8	2/5/1	3/5/0	2/6/0
ECG	6	3/1/2	3/2/1	4/2/0
Spectrum	4	4/0/0	3/1/0	3/1/0
Hemodynamics	3	0/3/0	0/3/0	0/3/0
EOG	2	2/0/0	0/1/1	0/2/0
EPG	2	0/0/2	2/0/0	1/0/1
Traffic	1	1/0/0	1/0/0	1/0/0
Power	1	1/0/0	1/0/0	1/0/0

図 5 は Base，Comp，CSTC それぞれのモデル構築時に利用された shapelet 数の度数分布を示したものである．

まず，Base について使用された shapelet 数の平均値は 2378 であり，度数分布からも多くのデータセットにおいて 3000 以下であることが分かる．データ数の大きな一部のデータセットでは最大数近くの shapelet が使用されていた．次に，Comp で使用された shapelet 数の平均値は 813 であり，Base よりも大幅に数が減少している．度数分布からも大半のデータセットが 1000 以下に収まっていることが分かる．Comp では基本的に訓練データ数と同じ数の合成 shapelet が生成されるが，1 つのデータから抽出された基本 shapelet 数が 1 つだけの場合には合成 shapelet は生成されない．このため，データ数が比較的少ないデータセットでは L1 正則化によって基本 shapelet 数が削減され，合成 shapelet の生成が抑制されたと考えられる．CSTC は基本 shapelet と合成 shapelet を併用するため，Base と Comp で使用された shapelet 数を足した数になることが度数分布からも分かる．このため，一部のデータセットでは 10000 を超える数の shapelet が使用されていた．

5　ま　と　め

本研究では，shapelet に基づく時系列データ分類手法において考慮されていない shapelet 間の出現順序と出現間隔に関する特徴を導入するため，L1 正則化付き分類器を用いて選定された shapelet を組み合わせて新たな shapelet を合成する方法を提案した．合成対象となる shapelet の組み合わせ数は一般に膨大であるため，提案手法では，合成対象となる shapelet の組み合わせ数を抑制するため，提案手法では，同一データから選定された shapelet のみを合成対象とすること，同一データから選定されたすべての shapelet を 1 つの新たな shapelet として合成することの 2 つを行うことで shapelet 数の増大を抑制した．実験では，UCR Time Series Classification Archive から選んだ 112 個のデータセットを用いて提案手法による分類を行い，合成 shapelet の導入による効果を確認した．

文　　献

[1] M. Middlehurst, P. Schafer and A. Bagnall, “Bake off redux: a review and experimental evaluation of recent time series classification algorithms,” ArXiv, 2304.13029, 2023.
[2] C. Lubba, S. Sethi, P. Knaute, S. Schultz, B. Fulcher and N. Jones, “catch22: canonical time-series characteristics,” Data Mining and Knowledge Discovery, Vol. 33, No. 6, pp. 1821-1852, 2019.
[3] J. Hills, J. Lines, E. Baranauskas, J. Mapp and A. Bagnall, “Classification of time series by shapelet transformation,” Data Mining and Knowledge Discovery, Vol. 28, No. 4, pp. 851-881, 2014.
[4] A. Dempster, F. Petitjean and G. Webb, “ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels,” Data Mining and Knowledge Discovery, Vol. 34, pp. 1454–1495, 2020.
[5] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom and A. Bagnall, “HIVE-COTE 2.0: a new meta ensemble for time series classification,” Machine Learning, Vol. 110, pp. 3211-3243, 2021.
[6] Z. Wang, W. Yan and T. Oates, “Time series classification from scratch with deep neural networks: A strong baseline,” In 2017 International joint conference on neural networks (IJCNN), IEEE, pp. 1578–1585, 2017.
[7] A. Guillaume, C. Vrain and E. Wael, “Random Dilated Shapelet Transform: A New Approach for Time Series Shapelets,” Lecture Notes in Computer Science, Springer, vol. 13363, pp. 653–664, 2022.
[8] J. Grabocka, M. Wistuba and L. Schmidt-Thieme, “Fast classification of univariate and multivariate time series through shapelet discovery,” Knowledge and Information Systems, Vol. 49, pp. 429–454, 2016.
[9] J. Grabocka, N. Schilling, M. Wistuba and L. Schmidt-Thieme, “Learning time-series shapelets,” In Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 392-401, 2014.
[10] J. Lines and A. Bagnall, “Time series classification with ensembles of elastic distance measures,” Data Mining and Knowledge Discovery, Vol. 29, pp. 565-592, 2015.
[11] G. Li, B. Choi, J. Xu, S. S. Bhowmick, K. P. Chun and G. L. H. Wong, “Efficient Shapelet Discovery for Time Series Classification (Extended Abstract),” In Proc. of the IEEE 37th International Conference on Data Engineering (ICDE), pp. 2336-2337, 2021.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 16:00 ~ 18:10 | T3-A オンライン（Zoom Events)

レビュー・コメント分析 (3)

座長:野宮 浩揮(京都工芸繊維大学)
コメンテータ:有次 正義(熊本大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T3-A-6-01] レビュー分析に基づくユーザ目線でのシーン別スポットスコアリング手法の提案

*大岡 芙実[1]、福嶋 亮太朗[1]、張 建偉[2]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学)

16:20 ~ 16:45
[T3-A-6-02] 直感的動画検索における単語分散表現を用いた評価項目別スコアリング手法

*笹原 彰斗[1]、山口 史弥[2]、上田 真由美[3,4]、中島 伸介[1] (1. 京都産業大学、2. 京都産業大学大学院、3. 流通科学大学、4. 大阪大学)

16:45 ~ 17:10
[T3-A-6-03] 生成系言語モデルによる架空のレビューを介した映画検索

*宮下 天祥[1]、荘司 慶行[2]、藤田 澄男[3]、大原 剛三[1] (1. 青山学院大学、2. 静岡大学、3. LINEヤフー株式会社)

17:10 ~ 17:35
[T3-A-6-04] SNSコメントを利用した対照学習によるニュースの潜在的な関係性の抽出と可視化

*高橋 陸[1]、牛尼　剛聡[2] (1. 九州大学大学院芸術工学府牛尼研究室、2. 九州大学大学院芸術工学研究院)

DEIM2024

レビュー分析に基づくユーザ目線での シーン別スポットスコアリング手法の提案

大岡　芙実† 福嶋亮太朗† 張 建偉†† 中島　伸介†

† 京都産業大学 情報理工学部 〒 603–8555 京都府京都市北区上賀茂本山
†† 岩手大学 理工学部 〒 020-8551 岩手県盛岡市上田 4-3-5
E-mail: †{g2153227，g2054155，nakajima}@cc.kyoto-su.ac.jp，††zhang@iwate-u.ac.jp

あらまし 既存のグルメサイトなどでは目的・シーン別の検索があり，ユーザの条件や目的に応じてタグを選び，飲食店などの店舗情報を絞り込めるようになっている．しかし，サイト側で用意されたタグだけでは，ユーザの要望に対して柔軟に答えることができないと考えた．本研究ではスポットごとのレビュー情報を集め，ユーザ側に自由文の入力を求め，入力された自由文に類似するスポットを探索することで，ユーザの要望に柔軟に対応したスポットの推薦を行う．

キーワード スポット推薦，利用シーン，レビュー分析，自動スコアリング

1. はじめに

我々が好みのカフェやレストランを検索する際，食べログ [1] や Retty [2] をはじめとするグルメサイトを利用して，ユーザが自身の目的や目的のシーンなど，こだわりの条件に基づく店舗検索が可能である．例えば，目的やシーン別の検索では，グルメサイト側が設定した「接待・会食」や「記念日」,「デート」といったシーンを選択することができる．しかしながら，これらのシーンはあくまでもサイト側もしくは店舗側が設定した限定的なシーンであり，例えば...

- 勉強や PC 作業が可能なカフェ
- SNS 映えする写真が撮れるカフェ
- 静かにまったりリラックスできるカフェ
- 子供連れの友達とゆっくりおしゃべり出来るカフェ

等の様々なユーザ要求に柔軟に対応することは難しい．

これらのユーザ要求は，ユーザ目線での評価に基づくものであり，必ずしも店舗側で把握できるものではない．すなわち，各店舗がこれらのユーザ要求に合致するかを判断するためには，利用したユーザが投稿したレビューを適切に分析する必要がある．

そこで本研究では，上記のようにユーザが想定する利用シーンに対して合致するかを数値化することが可能な，レビュー分析に基づくユーザ目線でのシーン別スポットスコアリング手法を提案する．ただし,「勉強や PC 作業が可能なカフェ」を評価する際，単純に「勉強」や「PC 作業」というキーワードマッチングで評価することは十分ではない．なぜなら,「レポート作成がはかどった」や「空き時間に報告書が書けた」等の「勉強」や「PC 作業」というキーワードを含まない形でレビューを記述することが大いにあり得るためである．したがって，本研究では，各種評価表現に対応する評価表現辞書に基づくスコアリング手法 [3] や fastText による分散表現を用いた手法を用いて，様々な形で記述されたレビューテキストを適切に分析することでユーザ目線でのシーン別スポットに対するスコアリングを可能にする手法を提案する．なお，本研究の最終目標としてはカフェやレストラン等の飲食店に限らない様々なスポットに対するシーン別スコアリングを目指しているが，手始めとしてカフェを対象としたスコアリング手法を提案し，今後の研究において各種スポットに対応可能となるよう手法の拡張を行う予定である．

2. 関連研究

レビューなどのユーザの書き込みを分析し, その結果をもとに情報を推薦する研究は数多く行われている.

奥ら [4] は, 位置情報付きツイートを用いた観光スポットの特徴化を行う手法について提案している. 観光スポットと位置情報付きツイートの対応付けを領域判別問題の解法である OC-SVM を用いて行っている. また, 観光スポットの時間的特徴及び特徴語句を抽出し, 各種推薦手法を適用した観光スポット推薦システムを提案している. この提案手法では観光スポットとツイートの対応付けに着目しており, ユーザのスポットに対する評価などはあまり考慮していない.

熊本ら [5] は, 新聞記事を読んだ人々が感じる印象の強さを数値的に求めるための印象マイニング手法について提案している. 印象辞書を構築し, 記事の印象値とその記事を読んだ人が感じる印象の強さを定式化することで, 印象の強さを定量化している.

住友ら [6] は, 口コミ情報に含まれる感情語に基づいて, 選択したスポットの感情スコアと類似するスポットを推薦手法について提案している.「感情語辞書」を作成し, 感情スコアを算出している. 対象スポットの各口コミ毎に形態素解析を行い, その基本形と「感情語辞書」のマッチングにより, 感情スコアを取得している. 感情別にスコアを合計し, 各口コミの感情スコアを平均し, データベースに記録している. そして, 比較元のスポットとデータベース内に記録された比較元以外の全てのスポットで各感情スコアの差を求め, 差分の少ない類似スポットを抽出

図 1　スコアリング方式の例

している.

中野ら [7] は, 商品レビューから, その商品の機能や特徴に関する意見を抽出する手法について提案している. レビューを形態素解析, 係り受け解析を行い, 属性と意見の候補を含む文節を抽出し, 素性ベクトルを作成し, 正解ラベルの付与を行い, 機械学習により属性・意見を抽出している. また, 属性と意見のペアについては, 係り受け関係を用いて抽出している.

櫛見ら [8] は, ユーザレビューから特徴ベクトルを作成し, 類似度から映画を推薦する手法について提案している. 特徴ベクトルを作成には,Doc2vec を利用することで, 文脈や単語の語順を考慮している. また, 一つの映画の特徴ベクトルを類似度を計測する際にも, ある映画について, その映画の全レビューを用いて, 特徴ベクトルを生成する全体類似度と映画のレビューを文ごとに分割し, その内容や意味に基づいた特徴ベクトルを作成する部分類似度の 2 つの観点で考慮している.

3. 提 案 手 法

3.1 概　　要

飲食店を探す際には，食べログ [1] や Retty [2] などのグルメサイトを利用することが多い．これらのグルメサイトでは「接待・会食」や「記念日」などの特別なシーンに適した店舗を検索することが可能となっている．しかし,「勉強や PC 作業が可能なカフェ」や「子供連れの友達とゆっくりおしゃべり出来るカフェ」のような，ユーザの要求に柔軟に対応した検索をすることができない．そこで本研究では，シーン別に検索が可能なシステムを構築し，ユーザの要望に対応した店舗の推薦を目的とする.

3.2 デ　ー　タ

本研究ではレストランなどの店舗と比較して,「勉強をしている人」や「子供連れの友達とゆっくりおしゃべりができる」など，様々なシーンで利用されているためカフェのレビューを収集する．データは食べログ [1] のレビュー情報を収集する.

3.3 手動による評価表現辞書の作成方法

本節では手動による評価表現辞書の作成方法について示す. 評価表現辞書には，評価項目，キーワード，程度語がある．評価項目とは表 1 節に示している「勉強・作業」や「子供連れ」などのことである．キーワードとは表 1 に示している「Wifi」や「親子」などである．程度語とは,「とても」や「少し」など副詞のことを示している.

本研究でスコアリングする評価項目と対応するキーワードを表 1 に示す．評価項目は 2 名の学生で検討し，日常的なシーンに対応できるように決定した.

表 1　評価項目と対応するキーワード例

評価項目	キーワード
勉強・作業	wifi, コンセント, 静か,PC, 捗る
子供連れ	親子, 子連れ, キッズスペース, ベビーカー
リラックス・気分転換	リラックス, くつろげる, 混んでない
景色がいい	自然, ウッドデッキ, テラス
SNS 映え	映える, かわいい, 推し
ランチ	美味しい, 香りが良い, 香ばしい
待ち合わせ	分かりやすい, 目印になる, 駅から近い

次に，評価表現辞書によるスコアリング方法を説明する.「とてもリラックスできた」というレビューがあると,「リラックス・気分転換」の項目に 5 点のスコアを付与する．このように，レ

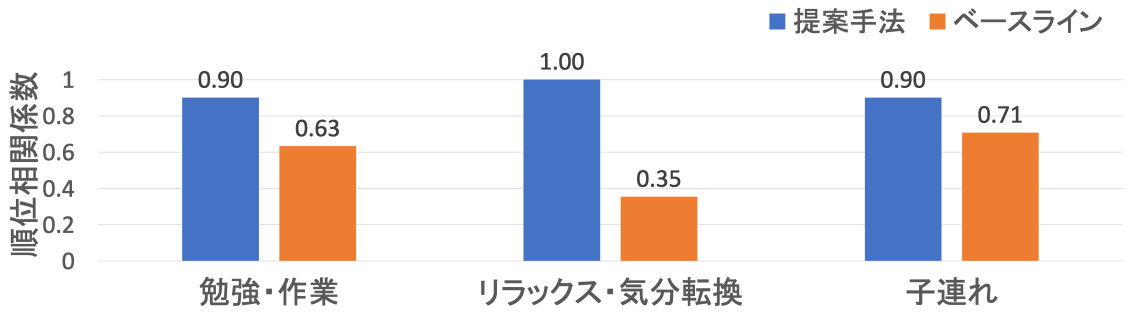


図 2　正解データとの順位相関係数

ビュー中に指定したキーワードと一致すると評価項目にスコアを付与する．

また,「少しだけリラックスできた」というレビューがあると,「リラックス・気分転換」の項目に 3 点のスコアを付与する．このように，キーワードが一致していても，程度語によって付与するスコアが異なる．レビューごとにスコアを付与し，店舗ごとに平均スコアを算出する．また，平均スコアをもとにユーザの要望に対応した店舗を推薦する．

3.4　分類器によるラベル別スコアリング

本節では, 分類器によるラベル別スコアリングの方法について述べる．評価項目に該当するレビューにラベルを付与し, 正解データを作成する．この正解データを用いて, 分類器を構築する．構築した分類器を用いて, 図 1 のように各店舗ごとのレビューを評価し, 評価項目ごとのスコアを合計し, 平均をその店舗のスコアとする．出力されたスコアをもとに，ユーザの要望に対応した店舗を推薦する．

3.5　店舗の推薦手法

本節では, ユーザの要望に応える推薦方法について述べる．ユーザが選択した評価項目のスコアが大きい店舗を降順に提示し, 推薦する．ユーザが「リラックス・気分転換」の項目を選択したとすると,3.3,3.4 節の手法で算出した「リラックス・気分転換」のスコアをもとに各店舗を降順に並べ, ユーザに提示する．

4.　評 価 実 験

本節では,3.5 節で説明した提案手法に基づく評価実験について述べる．

4.1　実験目的および方法

提案手法は, 従来のグルメサイトの目的・シーン別の検索よりもユーザ目線でのシーン別検索を可能にすることを目的としている．そのため評価実験では, 評価項目を「勉強・作業」「リラックス・気分転換」「子連れ」といったユーザの目的や利用シーンに応じたものに設定した．既存のグルメサイトでこれらの項目について検索することはできない．したがって, 被験者に各店舗ごとのレビューを読んでもらい, 先ほど挙げたそれぞれ評価項目についてランキングを作成して貰った．被験者の評価をもとに各評価項目について総合的なランキングを作成し, これを正解データとした．また, ベースラインは単純にレビューの単語数を数え, 降順に店舗を並べることによってランキングを作成した．ベースラインを作成するための単語として,「勉強・作業」では「勉強」「作業」,「リラックス・気分転換」では「リラックス」「気分転換」,「子連れ」では「子ども」「子供」を選んだ．提案手法については,3.3 節の手法を用いて算出したスコアをもとにランキングを作成した．

4.2　実験結果と考察

正解データとベースライン、提案手法のそれぞれで作成したランキングの順位相関係数を求めた結果を図 2 に示す．

すべての評価項目で順位相関係数の値がベースラインよりも大きくなった．特に「勉強・作業」と「リラックス・気分転換」の評価項目についてはベースラインよりも差が付いた．

先ず,「勉強・作業」の項目で, ベースラインよりも良い結果となったのは, 評価辞書で「Wi-Fi」や「ノマドワーク」といった単語を評価したことにより, 被験者の回答と近いものとなったことが考えられる．

次に「リラックス・気分転換」の項目で, ベースラインよりも良い結果となったのは, 評価辞書で「落ち着いた雰囲気」や「隠れ家」といった単語を評価したことにより, 被験者の回答と近いものとなったことが考えられる．

次に「子連れ」の項目で, ベースラインとあまり変わらない結果となったのは, そもそもキッズメニューや子供についてのレビューが少なく, 極端に子供が好みそうな店舗と親が子供を連れていくのを避けるような店舗以外でのランキングが被験者にとって作成しにくかったことが考えられる．ここで言う, 子供が好みそうな店舗とは子供向けのイベントがあったり, キッズメニューがある店舗だ．また, 親が子供を連れていくのを避けるような店舗とは, 分煙されていないような店舗だ．それ以外の子供向けのメニューが特にないような店舗では被験者にとって評価が難しいと考えられる．

今回,3 つの評価項目について実装した．今後は表 1 で実装しなかったものも順次実装していき, よりユーザ目線であらゆるシーンでの検索を可能にしていきたい．また，様々な評価に対応するために今回実装しなかった 3.4 節の分類器によるラベル別スコアリングも実装していきたい．

5. ま と め

本稿では, 既存のグルメサイトなどのキーワード検索では検索できないユーザの要望に沿った新たな推薦手法について提案した. 今後は, 提案したシステムの構築に取り組んでいく.

謝　　辞

本研究の一部は，JSPS 科研費 (課題番号：20H04293) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す.

文　　献

[1] 食べログ，https://tabelog.com/
[2] Retty [レッティ]，https://retty.me/
[3] Mayumi Ueda, Yuki Matsunami, Panote Siriaraya and Shinsuke Nakajima, "Developing Evaluation Expression Dictionaries for the Cosmetic Review Recommendation", Proc. IAENG International MultiConference of Engineers and Computer Scientists (IMECS 2019), pp.236-241, Hong Kong, China, March 2019.
[4] 奥健太, 橋本拓也, 上野弘毅, 服部文夫, "位置情報付きツイート対応付けに基づく観光スポット推薦システムの開発", Web インテリジェンスとインタラクション研究会, No2, pp. 7-12, 2013.
[5] 熊本忠彦, 河合由起子, 張建偉, "ユーザ印象評価データの観察と分析に基づくテキスト印象マイニング手法の設計", 研究報告自然言語処理 (NL), Vol.2012, No.4, pp. 1-10, 2012.
[6] 住友千将, 石野拓也, 久保洸貴, 岳五一, "口コミ情報に含まれる感情語に基づく類似スポット推薦システムの構築と実証実験", パーソナルコンピュータ利用技術学会論文誌, pp.29-35, 14(1), 2020.
[7] 中野裕介, 湯本高行, 新居学, 上浦尚武,"機械学習による商品レビューの属性-意見ペアの抽出", 情報処理学会研究報告, Vol.2015-DBS-162, No.14, 2015
[8] 櫛見圭司, 北山大輔, "ユーザレビューを用いた全体的・部分的観点の類似に基づく映画推薦",DEIM Forum 2017, B5-4, 2017.

直感的動画検索における単語分散表現を用いた評価項目別スコアリング手法

笹原　彰斗† 　山口　史弥†† 　上田真由美†††,†††† 　中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
††† 流通科学大学 経済学部　〒 651–2188 兵庫県神戸市西区学園西町 3-1
†††† 大阪大学 サイバーメディアセンター　〒 560–0043 大阪府豊中市待兼山町 1 番 32 号
E-mail: †{g2054263,nakajima}@cc.kyoto-su.ac.jp, ††i2286211@cc.kyoto-su.ac.jp,
†††,††††Mayumi_Ueda@red.umds.ac.jp

あらまし　近年，動画共有サイトには膨大な動画が存在し，多くのユーザはタイトルやサムネイルから好みの動画を判断して視聴する．しかし，既存の情報では実際に再生するまでユーザが真に求めている動画か判断することが困難である．我々は，動画に対して投稿されたコメントを分析し，その動画に対するユーザの直感的評価である評価項目別スコアの算出に取り組んできた．本稿では，提案手法である評価項目別スコアリングにおける分類器構築に使用する学習データの収集コスト削減に取り組んだので報告する．

キーワード　コメント分析，直感的動画検索，自動スコアリング

1　はじめに

近年，動画共有サイトには膨大な動画が存在し，多くのユーザはキーワード検索をした後に，タイトルやサムネイルから好みの動画を判断して視聴する．しかしながら，ユーザが直感的に観たいと思える動画を検索することは，従来のキーワード検索では限界がある．例えば，印象を表す表現である「エモい」というキーワードで動画を検索するケースを考える．「エモい」という言葉でキーワード検索した場合，タイトルや説明文に「エモい」というキーワードを含む動画がヒットするが，タイトルや説明文に「エモい」というキーワードが含まれていても必ずしも「エモい」と感じるとは限らない．それと同時に，「エモい」という言葉が含まれていなくても「エモい」と感じる動画は多数存在している．

そこで我々は，ユーザが直感的に観たいと思える直感的動画検索を可能とする，動画に対する評価項目別スコアを用いた直感的動画検索システム [5] を提案した．この動画検索システムは，動画の印象を表す 8 つの各評価項目ごとに分類器モデルを構築し，動画の特徴をレーダーチャートでユーザに提示するシステムである．直感的動画検索システムでは，評価項目名（例：エモい）のキーワードをタイトルや説明文に含んでいない動画に対しても，有効な検索を可能にしている．本稿では，評価項目別スコアを算出する際に使用する分類器の構築方法を再検討することで，分類器構築に使用する学習データの収集コスト削減に取り組み，評価実験を行ったため報告する．

本稿の構成は以下の通りである．2 章では，関連研究を紹介する．3 章では，これまでに開発してきた「直感的動画検索システム」について述べる．4 章では，評価項目別スコアリングに使用する分類器の構築方法を再検討し，学習データ収集コストの削減について述べる．5 章では評価実験を行い，4 章で述べた学習データ収集コストの削減手法の有効性を確認する．最後に 6 章でまとめと今後について述べる．

図 1　従来の検索手法の問題点

2　関連研究

以下に，動画の分析およびテキスト分析に関する関連研究をあげ，本研究との差異を示す．近年の動画共有サイトではユーザが自由にコメントすることができる，YouTube [1] やニコニコ動画 [4] などが存在する．また，上記のサイトには膨大な数の動画が日々アップロードされており，タグやキーワードを用いた検索が可能である．

動画分析に基づく関連研究として，村上らは動画に付与されるコメントの「w」という文字に着目し，その動画の「笑い」の度合いを数値化した [6]．中村らは，感性に基づく正規表現辞書を構築し，ユーザが書き込んだコメントを分析して，positive，negative，happy，sad，surprise といった感性に基づくランキング検索を可能とした [9]．本研究では，これらの感性分析には留まらず，ユーザのより直感的な動画検索を可能とする技術の開発を目指している．

また，レビュー分析に基づく評価表現別自動スコアリング手法に関する従来研究として，松波らは，コスメアイテムに特化した評価表現辞書を構築し，レビュー自動スコアリング手法を開発した [10]．同様に，安部らは，ホテル予約サイトに投稿されたレビュー文を分析することで，ホテルに特化した評価表現辞書を構築し，ホテルレビュー自動スコアリング手法を開発した [11]．

本研究ではこれらの辞書構築手法を参考にしつつも，動画に対するより直感的な評価項目の選定と各評価項目に対する評価表現の効率的な抽出を目指すものである．上述の通り，動画のコメント分析による感性検索の研究や，レビュー解析による評価表現別のスコアリング手法に関する研究は数多く行われているが，本研究にて実施している直感的動画検索のための評価表現別自動スコアリング手法に関する研究開発は行われていない．本研究では，直感的動画検索のための評価表現別スコア算出に使用する分類器構築手法を再検討し，必要な学習データの収集コスト削減に取り組む．

3　直感的動画検索システム

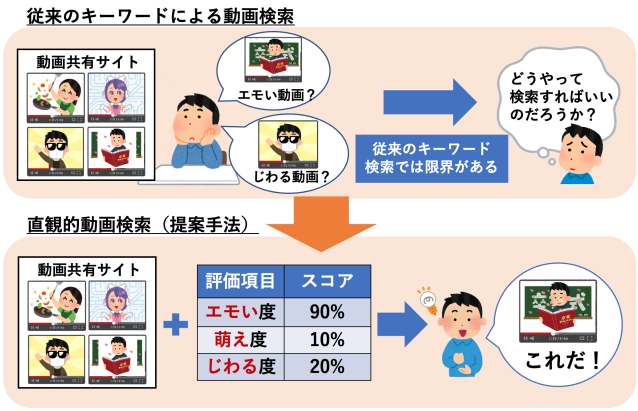


図 2　直感的動画検索システムの概要

3.1　直感的動画検索システムの概要

直感的動画検索システムでは，動画に付与されるコメントを分析し，評価項目別自動スコアリングを行う．これにより，動画を再生する前に，ユーザが真に求めている動画か否かを判断可能となる．図 2 のように，従来手法では動画共有サイトの一覧から主にタイトルとサムネイルを見て視聴する動画を選択するため，実際に動画を視聴してみると，予想していた内容と異なることがある．一方，提案手法である直感的動画検索システムでは，各動画に従来のキーワード検索では拾えないような，ポジティブやネガティブと言った単純な感性検索ではない動画に特化した評価項目にスコアを付与することで，視聴前からユーザが真に求めている動画か否かを判断可能としている．

3.2　各評価項目の分類器モデルの構築

本節では，動画に対する評価項目別スコアリングのための分類器モデルの従来の構築方法について述べる．動画共有サイト

図 3　直感的動画検索システムの検索画面

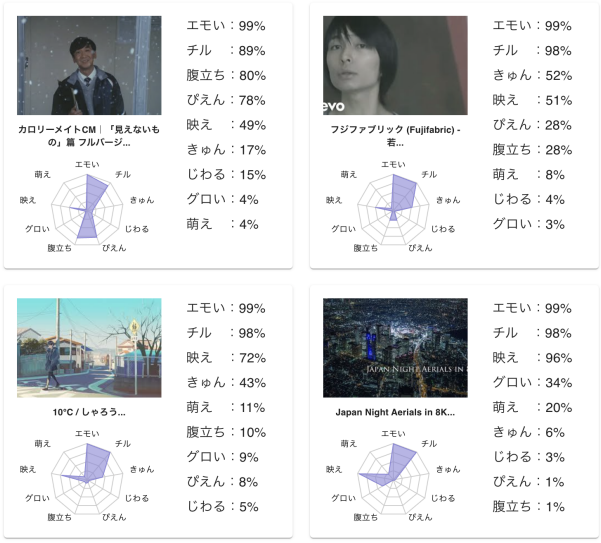


図 4　「エモい」で検索した場合の検索結果画面

において動画を視聴するユーザは様々な直感的表現を用いたコメントを投稿している．そこで，実際に投稿されている直感的表現を用いたコメントを収集し，学習データとして使用することで評価項目ごとの分類器モデルを構築する．使用する評価項目は「エモい」「チル」「きゅん」「じわる」「ぴえん」「イラつく」「映え」「萌え」の計 8 項目である．

学習データとして使用するコメントデータは，動画共有サイト YouTube が提供する YouTube Data API [2] から取得する．分類器モデルの構築には fastText [3] を用いている．以下に，例としてエモい分類器モデル構築についての手順を示す．

エモい分類器構築手順

（1）　YouTube で評価項目（例：エモい）だと感じる動画を収集し，各評価項目ごとに 5000 コメントづつ収集する．
（2）　fastText を用いて評価項目 8 つに対してのマルチラベル分類器モデルを作成する．
（3）　「エモくないコメント」を収集するため，(2) で構築したマルチラベル分類器を用いて動画をスコアリングする．
（4）　エモいスコアが 0.05 以下の動画をエモくない動画と定め，

そこから 5000 コメントを収集する．

（5） 収集した「エモいコメント」5000 件と「エモくないコメント」5000 件を用いて fastText で学習させエモい分類器モデルを構築する．

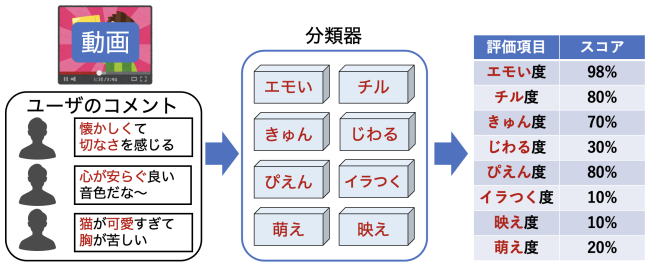


図 5　評価項目別スコアリングの手順

3.3　動画に対する評価項目別スコアの算出方法

本節では，動画に対する評価項目別スコアの算出方法について述べる．以下に，ある動画における「エモいスコア」の算出方法を示す．(図 6)

ある動画における「エモいスコア」の算出方法

（1） ある動画に投稿されているコメントを全て取得する．

（2） 取得したコメントを高評価数の多い順に並び替える．

（3） 上位 5%のコメントを 1 つずつエモい分類器に通していく．

（4） それぞれのコメントに対して「エモい確率」と「エモくない確率」が算出される．

（5） 「エモい確率」の平均をその動画の「エモいスコア」とする．

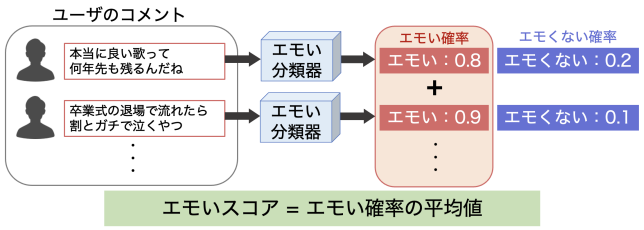


図 6　「エモいスコア」の算出方法

4　学習データ収集コストの削減

本章では，評価項目別スコアリングに使用する分類器構築の際に使用する学習データ量の削減について述べる．後述する手法で分類器を構築することによって，使用する学習データ量の削減・学習データ収集の際に使用している YouTube Data API の呼び出し回数の削減等が期待できる．

4.1　学習データ収集コスト削減手法

fastText を使用して分類器を構築する際は，学習データとして「エモいコメント」と「エモくないコメント」それぞれ 5000 件ずつ使用していた．そこで，学習データ量を削減するため，BERT を使用して「エモいコメント」と「エモくないコメント」それぞれ 200 件ずつまで減少させて分類器を構築する．(図 7)

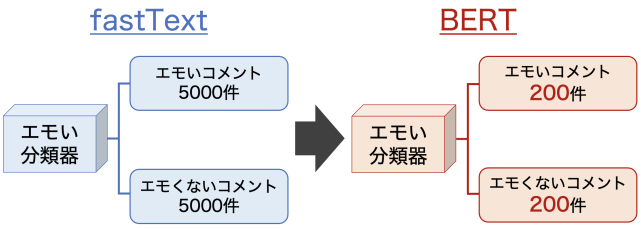


図 7　学習データ数の削減

4.2　BERT を使用した分類器の構築手順

以下で，BERT を使用して分類器を構築する際の「エモいコメント」と「エモくないコメント」の収集方法・分類器の構築方法について述べる．

エモい分類器構築手順

（1） YouTube で評価項目（例：エモい）だと感じる動画を選択する．

（2） その動画に付与されているコメントを高評価数順に並び替える．

（3） 高評価数上位 5%のコメントのみを収集する．

（4） (1)～(3) の手順をコメントが 200 件集まるまで繰り返す．

（5） BERT を使用して評価項目 8 つに対してマルチラベル分類モデルを作成する．

（6） エモくないコメントを収集するため，(5) で構築したマルチラベル分類器を用いて動画をスコアリングする．

（7） エモいスコアが 0.05 以下の動画をエモくない動画と定め，そこから「エモくないコメント」を 200 件収集する．

（8） 収集した「エモいコメント」200 件と「エモくないコメント」200 件を用いて BERT でエモい分類器を構築する．

5　ランキング精度比較実験

5.1　提案手法 1（fastText）と提案手法 2（BERT）のランキング精度比較実験

本節では，提案手法 1（fastText）で生成したランキングと提案手法 2（BERT）で生成したランキングの精度比較実験について述べる．(図 8)

順位	fastTextランキング
1	動画A
2	動画D
:	:
10	動画F

比較

順位	BERTランキング
1	動画B
2	動画C
:	:
10	動画D

図 8　nDCG で比較するランキング

5.1.1　実 験 手 順

提案手法 1 と提案手法 2 のランキング精度の比較を行うにあたり，評価指標として nDCG@k を用いた．以下で，ランキング精度の評価方法を評価項目「エモい」を例にして説明する．

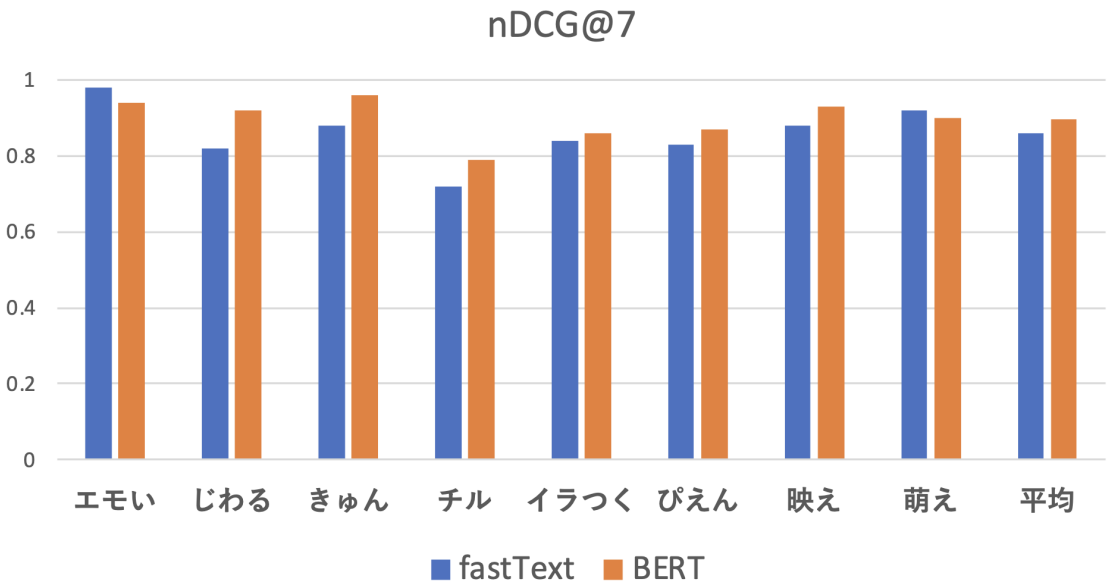


図 9　nDCG@7 の結果

ランキング精度の評価手順

（1） YouTube で「エモい」と感じる動画を 7 本,「エモくない」と感じる動画を 3 本の計 10 本の動画を収集する.

（2） 10 本の動画を「エモい順」に並び替えて,「理想エモいランキング」を作成する.

（3） 提案手法 1（fastText）で構築した分類器と提案手法 2（BERT）で構築した分類器の 2 つを使用して，10 本の動画をスコアリングし,「エモいスコア」の高い順に並び替えて nDCG@7 でランキング精度を評価する.

5.1.2　実験結果

図 9 は評価項目ごとの nDCG@7 の結果である．提案手法 2（BERT）によって構築された分類器を用いることで,「エモい」と「萌え」を除く全ての評価項目において，fastText よりも優れたランキング精度となった．この結果から，BERT を用いて分類器を構築することで学習データを 10000 件から 400 件まで削減することが可能であり，以前の 10000 件の学習データを使用した場合と同等以上のランキング精度となることが示された.

5.2　従来手法と提案手法のランキング精度比較実験

本節では，従来手法（YouTube）で生成したランキングと，提案手法 1（fastText）と提案手法 2（BERT）で生成したランキング精度の比較実験について述べる.（図 10）

5.2.1　実験手順

従来手法と提案手法のランキング精度の比較を行うにあたり，評価指標として Probabilistic Interleaving 手法を用いた．Interleaving を用いた評価手法は加藤誠ら [7] の研究を参考にした．Probabilistic Interleaving では，2 つのランキングを混ぜ合わせた混合ランキングを作成し，混合ランキングを被験者に提示することでどちらのランキングが優れているかを判断する手法（図 11) である．以下で，2 つのランキング（従来手法，提案手法）の作成手順と，ランキング精度の評価方法を評価項目「エモい」を例にして説明する.

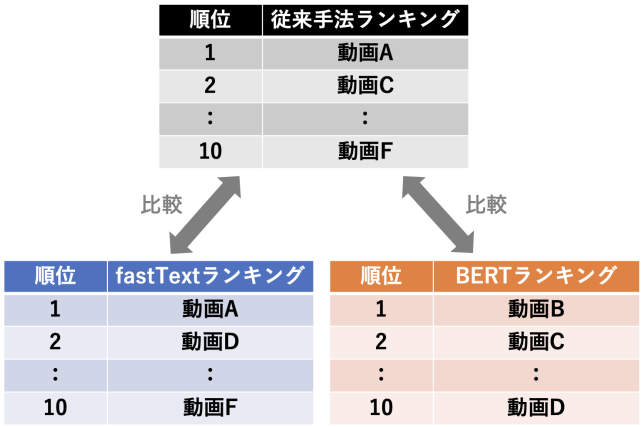


図 10　Interleaving で比較するランキング

従来手法の「エモいランキング」については YouTube で評価項目名「エモい」で検索し，関連度順 TOP10 を従来手法の「エモいランキング」としている.

提案手法の「エモいランキング」作成手順

（1） YouTube から約 700 本の動画を収集する.

（2） 構築済みの「エモい分類器」を使用して全ての動画をスコアリングし,「エモいスコア」の高い順に並び替え，TOP10 を提案手法の「エモいランキング」とする.

ランキング精度の評価手順

（1） 作成した「従来手法のエモいランキング」と「提案手法のエモいランキング」を Probabilistic Interleaving 手法により混ぜ合わせ，混合ランキング TOP10 を生成する.

（2） 混合ランキングを被験者に提示し，直感的にエモいと感じる動画をクリックして視聴してもらう.（提示する情報はサムネイルとタイトルのみ）

表 1　インターリービングで fastText が支持された割合（従来手法 vs fastText）

評価項目	エモい	チル	きゅん	じわる	ぴえん	イラつく	映え	萌え
支持された割合	100%	95%	97%	89%	89%	87%	97%	84%

表 2　インターリービングで BERT が支持された割合（従来手法 vs BERT）

評価項目	エモい	チル	きゅん	じわる	ぴえん	イラつく	映え	萌え
支持された割合	85%	89%	96%	93%	96%	100%	100%	82%

（3） 動画を視聴後，その動画を見て実際にエモいと感じたかどうかを 5 段階で評価してもらう．
（4） 「はい」または「ややはい」の場合，credit を与える．
（5） (1)〜(4) の手順を，直感的に「エモい」と感じる動画がなくなるまで繰り返す．

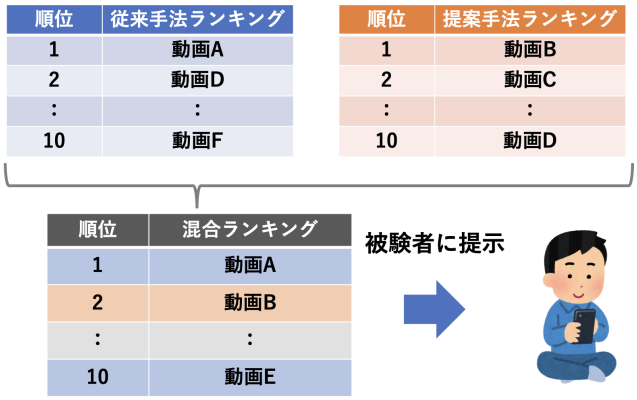


図 11　インターリービング実験の概要

5.2.2　従来手法と提案手法 1（fastText）の実験結果

表 1 は評価項目ごとの Probabilistic Interleaving の結果である．は評価項目ごとの Probabilistic interleaving の結果である．引き分けの場合，従来手法・提案手法どちらにも 0.5 人分の credit を付与する．

分類器構築に提案手法 1（**fastText**）を使用した場合

エモいランキングの結果は，被験者全員が提案手法 1（fastText）を支持した．チルランキングの結果は，95%の被験者が提案手法 1（fastText）を支持した．チルランキングの結果は，97%の被験者が提案手法 1（fastText）を支持した．きゅんランキングの結果は，97%の被験者が提案手法 1（fastText）を支持した．じわるランキングの結果は，89%の被験者が提案手法 1（fastText）を支持した．ぴえんランキングの結果は，89%の被験者が提案手法 1（fastText）を支持した．イラつくランキングの結果は，87%の被験者が提案手法 1（fastText）を支持した．映えランキングの結果は，97%の被験者が提案手法 1（fastText）を支持した．萌えランキングの結果は，84%の被験者が提案手法 1（fastText）を支持した．よって，全ての評価項目において提案手法 1（fastText）の方が優れているという結果になった．

5.2.3　従来手法と提案手法 2（BERT）の実験結果

表 2 は評価項目ごとの Probabilistic Interleaving の結果である．引き分けの場合，同様に従来手法・提案手法どちらにも 0.5 人分の credit を付与する．

分類器構築に提案手法 2（**BERT**）を使用した場合

エモいランキングの結果は，85%の被験者が提案手法 2（BERT）を支持した．チルランキングの結果は，89%の被験者が提案手法 2（BERT）を支持した．きゅんランキングの結果は，96%の被験者が提案手法 2（BERT）を支持した．じわるランキングの結果は，93%の被験者が提案手法 2（BERT）を支持した．ぴえんランキングの結果は，96%の被験者が提案手法 2（BERT）を支持した．イラつくランキングの結果は，被験者全員が提案手法 2（BERT）を支持した．映えランキングの結果は，被験者全員が提案手法 2（BERT）を支持した．萌えランキングの結果は，82%の被験者が提案手法 2（BERT）を支持した．よって，全ての評価項目において提案手法 2（BERT）の方が優れているという結果になった．

5.2.4　実験結果のまとめ

表 1 と表 2 を比較してみると，提案手法が支持されている割合が，fastText（学習データ 10000 件）の場合と，BERT（学習データ 400 件）の場合で，大きな差がないことが示された．以上の結果により，従来手法のランキングと提案手法のランキングを比較すると，BERT を活用することにより，学習データを 10000 件から 400 件まで減少させても，10000 件の学習データを使用していた場合とと同程度のランキング精度が確認できた．

6　おわりに

本稿では，分類器の構築方法を再検討することにより，従来の分類器構築方法に比べて，少ない学習データで分類器を構築し，以前と同程度のスコアリング精度を出すことができた．今後は，直感的動画検索システムのユーザインタフェースや評価項目の自動追加などについても検討する．

謝　辞

本研究の一部は，JSPS 科研費 (課題番号：19K12243, 20H04293, 22K12281) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] YouTube https://www.youtube.com/
[2] YouTube Data API https://developers.google.com/youtube/v3/getting-started
[3] fastText https://fasttext.cc/
[4] ニコニコ動画 https://www.nicovideo.jp/
[5] 鵜田 和士ら，動画に対する評価項目別スコアを用いた直感的動画検索システム，DEIM Forum 2022 C21-5, 2022.
[6] 村上 直至ら，動画コンテンツの視聴者コメントに基づくランキングとその評価，DEIM Forum 2012.
[7] 加藤 誠ら，コミュニティQ&A サイトにおける質問検索システムの大規模オンライン評価，DEIM Forum 2018 G2-2，2018.
[8] Katja Hofmann, Shimon Whiteson and Maarten de Rijke，A Probabilistic Method for Inferring Preferences from Clicks，Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011
[9] 中村 聡史ら，ソーシャルアノテーションに基づく動画検索手法，DEIM Forum 2009 D6-1，2009.
[10] Mayumi Ueda, Yuki Matsunami, Panote Siriaraya and Shinsuke Nakajima, "Developing Evaluation Expression Dictionaries for the Cosmetic Review Recommendation", Proc. IAENG International MultiConference of Engineers and Computer Scientists (IMECS 2019), pp.236-241, Hong Kong, China, March 2019.
[11] 安部 克ら，レビュー評価項目別スコアを用いたホテル推薦手法の提案，DEIM Forum 2019 P1-134，2019.
[12] S. Siersdorfer, S. Chelaru, W. Nejdl, J. San Pedro, How useful are your comments? analyzing and predicting youtube comments and comment ratings, in: Proceedings of the 19th international con
[13] M. Naomichi, I. Eisuke, Emotional video ranking based on user comments, in: iiWAS ’11: Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, 2011, pp. 499–502.

生成型言語モデルによる仮想映画レビューを介した映画検索

宮下　天祥†　　荘司　慶行††　　藤田　澄男†††　　大原　剛三†

† 青山学院大学大学院 理工学研究科　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
†† 静岡大学 情報学部　〒 430 – 0807 静岡県 浜松市 中区 城北 3 - 5 - 1
††† LINE ヤフー株式会社　〒 102 – 8282 東京都 千代田区 紀尾井町 1 - 3
E-mail: †miyashita@sw.it.aoyama.ac.jp, ††shojiy@inf.shizuoka.ac.jp, †††sufujita@lycorp.co.jp,
††††ohara@it.aoyama.ac.jp

あらまし　本論文では，映画への感想のような抽象的な検索要求に対し，それに近い映画を発見する手法を提案する．近年主流の生成型言語モデルは，文の続きを予測することに長けている．そこで，映画レビューを用いて追加学習し，書き出しを与えると続きのレビューを生成するモデルを構築した．このモデルに「北野武の撮ったスターウォーズみたい」のような抽象的な要求を与えると，その文に続く，具体的で現実的なレビュー文が生成される．生成されたレビューと類似したレビューをもつ映画をランキングすることで，抽象的な検索要求に合致する映画が検索できる．実レビューサイトのデータを用いた被験者実験における定量的，定性的評価を通して，生成されたレビューに具体的かつ検索に有用な語が多く含まれ，実際に映画を検索可能であることが明らかになった．

キーワード　情報検索，映画レビュー，GPT-2，生成型言語モデル

1　はじめに

近年，ストリーミングによる映画配信サービス，サブスクリプションなどの流行を受けて，映画の視聴方法が多様化している．従来であれば，映画を観る際の選択肢として，映画館で鑑賞するか，DVD レンタルショップで DVD をレンタルするしかなかった．いずれの場合も，あらかじめ観たい映画を決めるか，数少ない候補の中から一番見たい映画を選択するのが一般的であった．しかし，インターネット上には昔の映画から最新の映画まで数多くの映画が存在し，ストリーミングといったサービスを利用することでこれらを自由に選んで視聴することが可能である．このように現代のユーザは，日々膨大な数の映画の中から次に観たい映画を選んで視聴している．そのため，より柔軟でユーザの細かい要求を反映可能な検索アルゴリズムへの要求が高まっている．

このようなアイテム検索においては，メタデータ，アイテムの画像，ユーザの購買行動などが利用されている [1–3]．しかし，映画は映像作品であるため，こうした情報は多くない．そのため，現状の映画検索アプローチは，限られたメタデータからの検索，協調性フィルタリング [4] などに限られる．ここで，まだ活用しきれていない外部リソースとして，レビューが挙げられる．レビューには，ユーザの映画に対する感想，どのような内容だったかなどの情報が含まれるため，映画検索に用いる情報として活用可能であることが期待される．杉木ら [5] の研究では，レビュー文からアイテムの意見情報を抽出することで，メタデータのみからでは検索できないアイテムを検索可能にしている．Yang ら [6] は，Online Customer Review（OCR）を用いて検索した意見文から製品特徴を抽出する手法を提案している．このように，レビューを情報検索に活用することの有効性はすでに示されている．そのため，こうしたレビューの中から，ユーザの検索要求に近い記述が含まれるレビューを検索できれば，ユーザの求める映画を発見できる可能性がある．しかし，レビューは多数のユーザによる自由記述の文章であり，文体，語彙なども書き手によってさまざまである．したがって，ユーザの入力する検索クエリと実際のレビューでは語彙空間が異なるため，直接クエリからレビューを検索することは困難である．

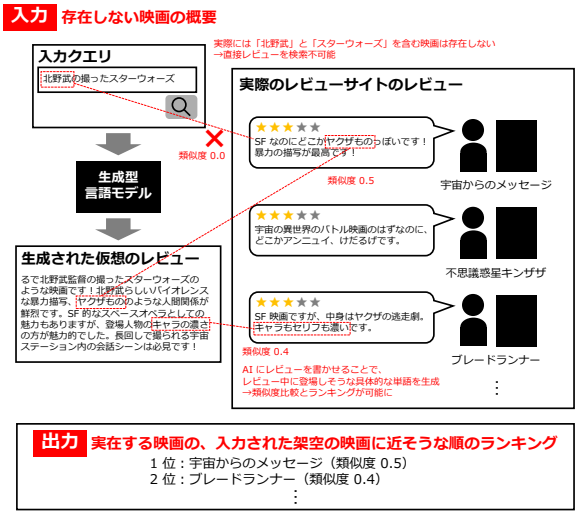


図 1　提案する検索モデルの入出力例と生成レビューの使用方法

そこで本研究では，生成型言語モデルを用いて仮想レビューを生成し，生成されたレビューから映画を検索する手法を提案する．提案手法では，生成型言語モデルによって映画に対する仮想レビューを生成させることで，実際のレビューと検索クエリとの語彙空間のギャップを軽減する．図 1 に示すように，生

成型言語モデルにレビューの書き出しを与えることで，その続きに来そうなレビュー文を生成させる．そして，その続きのレビュー文を拡張クエリとして扱うことで，映画を検索可能にする．

このような，書き出しを与えると具体的な続きのレビュー文を生成する言語モデルを作成するために，実際の映画レビューを教師データとして用いる．レビュー文の続きのレビューを生成するというタスクで言語モデルを学習させる．この際，実際のレビューには映画の内容に言及しない文が多く含まれる．たとえば,「友達と観ました.」や「期待外れだった.」など，映画の内容とは関係ない文が多く含まれる．また実際のレビューの書き出しは,「主人公が自身を鼓舞するために叫ぶシーンがかっこよかった.」といった具体的すぎる記述や,「ヒロインのひたむきな姿がかわいかった.」といった映画全体を評さない書き出しを持つものが多く含まれる．

そこで本研究では，学習時の工夫として，

（1） 映画内容に関連する記述の抽出，

（2） 重要部分の書き出しへの移動

という，2 つの処理を施した学習用のレビューデータを作成し，言語モデルを学習させた．こうすることで，抽象的な映画の説明文が入力された際に，その続きとして，具体的なレビューを生成可能にした．

提案手法が実際に映画の発見に役立つかを検証するために，実際の映画レビューを用いて学習した生成型言語モデルの生成レビューの質を定量的に評価し，仮想レビューを利用して検索した映画を定性的に評価した．これにより，生成型言語モデルを用いることでクエリとレビューという異なる言語空間をまたいだクエリ拡張が可能であるかとレビューの重要部分を書き出しに移動することの効果を明らかにした．

本論文は本章を含めた 6 章から構成される．第 2 章では，本研究と関連のあるアイテム検索，クエリ拡張，大規模言語モデルを用いた情報検索に関する先行研究について述べる．第 3 章では，提案手法の詳細について述べる．第 4 章では，提案手法の生成した仮想レビューに対する評価実験の結果について述べる．第 5 章では，実験結果に対する考察ついて述べる．最後に，第 6 章で本研究から得られた結論を述べる．

2 関 連 研 究

本研究は，映画というアイテムを検索することを目的としている．これを実現するために，生成型言語モデルを用いて仮想レビューを生成し，この生成レビューを拡張的なクエリと見なして映画を検索するという手法を取る．このように生成型言語モデルから出力された文書を用いた情報検索に関する研究は近年盛んにおこなわれている．そのため本章では，提案手法と関連する，アイテム検索，クエリ拡張，大規模言語モデルを用いた情報検索について述べる．

2.1 アイテム検索

オンラインショッピングサイトや特定のサービスに関する情報サイトにおいては，無数に存在する商品やサービスの中からユーザの求めるアイテムを精度よく検索することが求められる．本節では，こうしたアイテム検索に対する先行研究について述べる．

Karmaker ら [1] は，ランキング学習を用いて E コマース上のアイテム検索を最適化する手法を提案している．実際の E コマースデータセットを用いた実験から，ランキング学習手法のうち， LambdaMART [7] を適用すると最も検索の精度が高くなることを明らかにしている．Gao ら [2] は，ファッション分野の商品のテキスト-画像ペアで BERT をファインチューニングすることで，オンラインショッピング上のアイテムをテキストと画像の両方から検索可能にする手法を提案している．ファッション製品の画像とプロのスタイリストによる説明が含まれる Fashion-Gen データセットを用いたアイテム検索タスクから，Image-to-Text， Text-To-Image の両方で提案手法が最も高精度な検索を可能にしたことを明らかにしている．Lu ら [3] は，Transformer ベースの多言語モデルとグラフニューラルネットワークアーキテクチャを組み合わせた，E コマース検索のためのフレームワークを提案している．この手法では，アイテムを特徴量ベクトルにエンコードする際に，ユーザの購買行動に基づくグラフを用いて類似アイテムを推定し，類似アイテムの特徴量も同時に元のアイテムの特徴量に含めることで，クエリとアイテム間の語彙のギャップを軽減している．複数の国で展開している E コマースプラットフォームの検索ログからサンプリングされたデータセットを用いたアイテム検索タスクに対して，提案手法がアイテムの検索精度を大幅に向上させたことを明らかにしている．

しかし，映画は映像作品であるためメタデータ，購買行動などといったアイテムの特徴を表す情報が多くない．そこで，本研究ではユーザ投稿のレビューをアイテム検索のための情報として用いる．このようにアイテム検索にレビューを活用している研究について述べる．杉木ら [5] は，レビュー文の係り受け解析によってそれぞれのアイテムの意見情報を抽出し，検索クエリの要求を満たすアイテムを検索可能にする手法を提案している．実際の宿泊予約サイトに投稿されたレビューを用いた検索タスクから，宿泊施設の提供する情報のみでは検索が困難なクエリからの検索が可能となったことを明らかにしている．Yang ら [6] は，OCR（Online Customer Review）をセンチメント分析することで，クエリに関連する製品特徴に言及したレビューの検索精度を向上させる手法を提案している．実際のオンラインショッピングサイトのアイテムデータを用いたレビュー検索タスクから，提案手法が関連レビューの検索精度を大幅に向上させたことを明らかにしている．そのため，本研究では映画検索のための情報としてレビューを対象にする．

2.2 クエリ拡張

情報検索分野において，インターネット上に存在する情報量の指数関数的な増加とクエリに含まれる情報量の少なさによる検索文書のミスマッチが長らく問題視されている．このような問題に対処するためにユーザの入力したクエリを拡張するとい

うアプローチを選択している手法が数多く存在する．

例として，Voorhees ら [8] は WordNet を用いて入力クエリに含まれる単語の同義語を検索し，拡張クエリとして用いる手法を提案している．TREC [9] と呼ばれるテキスト検索用データセットを用いた実験から，提案手法によるクエリ拡張が特に短いクエリからの検索に有効であることを明らかにしている．Carpineto ら [10] は，クエリ拡張のための外部コーパスから単語の共起に基づいてクエリを拡張する単語を選択する手法を提案している．TREC を用いた情報検索タスクを通して，単語の共起に基づく拡張クエリ選択手法が，いくつかのタスクで他のクエリ拡張手法より効率の良い検索を可能にしたことを明らかにしている．Cui ら [11] は，ユーザの過去の検索クエリログからクエリに含まれる単語と検索に文書含まれる単語の相関関係を抽出し，拡張クエリの選択に活用する手法を提案している．TREC を用いた情報検索タスクから，ユーザの検索ログを用いたクエリ拡張によって情報検索の精度を向上させることを明らかにしている．Salton ら [12] は，適合性フィードバック手法を用いたクエリ拡張による情報検索の精度向上への効果を検証している．

また，クエリから検索された上位 k 件の文書をクエリと関連した文書として使用する疑似適合性フィードバックを用いたクエリ拡張手法も存在する．Xu ら [13] は，擬似適合性フィードバックと単語の共起を組み合わせたクエリ拡張手法を提案している．TREC を用いた情報検索タスクから，提案手法によるクエリ拡張がグローバル分析手法と比較して高精度な情報検索を可能にしたことを明らかにしている．そのため本研究では，クエリとユーザ投稿のレビュー間のギャップ軽減のためにクエリを拡張するというアプローチをとる．

2.3 大規模言語モデルを用いた情報検索

大規模言語モデルは Transformer [14] をベースとした，大規模なパラメータ，データ量が特徴の自然言語処理モデルである．大規模言語モデルは様々なタスクに対して，モデルのパラメータの更新なしに応用可能なことから様々な分野で用いられている．本節では，大規模言語モデルを情報検索に応用した既存研究について述べる．

Zhu ら [15] の論文によれば，大規模言語モデルを用いた情報検索は大きく Rewriter，Retriever，Reranker および，Reader に分けられる．本研究は映画レビューを検索するためのクエリを大規模言語モデルに生成させる研究であるため，大規模言語モデルを用いた Rewriter の研究にフォーカスする．Dai ら [16] は，大規模言語モデルに文書に対するクエリを生成させるタスクを few-shot で学習させることで，ラベルなし文書に対する正解クエリを生成し，検索器を訓練するためのデータセットを作成する手法を提案している．BEIR ベンチマークを用いた検索性能評価の結果から，提案手法によって訓練された検索器が人手によってラベル付けされたデータから訓練された検索器と同等の性能を発揮することを明らかにしている．

Gao ら [17] は，大規模モデルが生成した回答から実際の文書を検索することで信憑性のある回答を提示する手法を提案

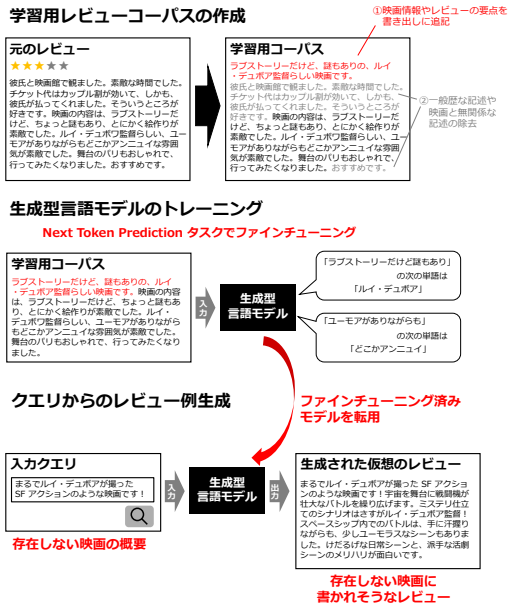


図 2　提案手法の概要

している．TREC DL19，DL20，BEIR データセットを用いた情報検索タスクから，この手法が zero-shot によるものであるにも関わらず，豊富な学習用データを用いた手法と比較して遜色ない性能を発揮したことを明らかにしている．Wang ら [18] は，大規模言語モデルにクエリとクエリに対する回答の文書の例をいくつか与えることで，未知のクエリに対する回答の文書を生成させる手法を提案している．こうして大規模言語モデルによって生成されたクエリと文書のペアを用いてリトリーバを学習させることで情報検索タスクに応用させることが可能である．TREC DL19，DL20，BEIR データセットを用いた情報検索タスクから，提案手法が従来の検索手法による検索精度を大幅に改善したことを明らかにしている．このように，大規模言語モデルを用いた情報検索手法は数多く提案されている．そこで本研究では，これを映画分野に応用して，映画検索の精度を向上させるための手法を提案する．

3 生成レビューを介した映画検索手法

本研究では，生成型言語モデルを用いて仮想レビューを生成し，生成レビューから類似する実際のレビューの投稿された映画を検索する手法を提案する．

3.1 手法の流れ

図 2 に，本手法で提案する生成型連語モデルを用いて仮想レビューを生成する一連の流れを示す．はじめに，レビューデータから，学習用レビューコーパスを作成する．ここでは，レビューデータに含まれるレビューに対して単語登場頻度による映画内容に関連する文の抽出（クレンジング）と，レビュー中の重要部分の書き出しへの移動を行う．次に，作成した学習用コーパ

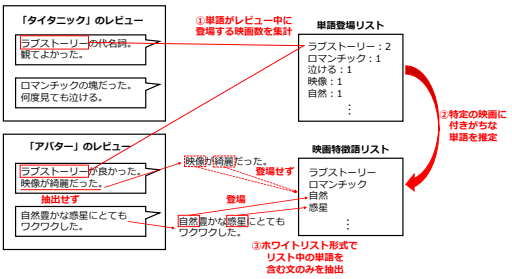


図 3　映画内容に関連する記述の抽出の流れ．それぞれのレビューから多くの映画に登場しすぎない特徴的な表現を含むレビューを抽出

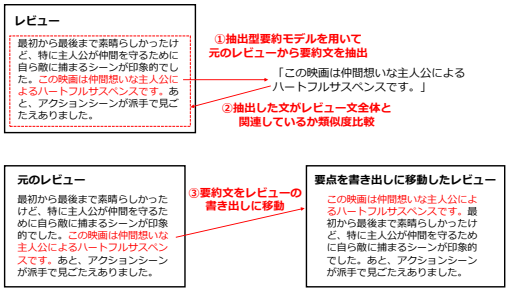


図 4　抽象型要約モデルを用いた重要部分の書き出しへの移動の流れ

スを用いて，生成型言語モデルを Next Token Prediction タスクでファインチューニングする．こうすることで，与えられた書き出しからレビューを生成する言語モデルを作成する．最後に，生成されたレビューと類似したレビューを持つ映画を，実際に検索する．

3.2　学習データ用の映画内容に関連する記述の抽出

本研究では，生成型言語モデルに仮想レビューを生成させるために，実際のレビューデータを用いてモデルを学習する．しかし，実際のレビューデータはユーザの自由記述による文章であるため，一般的な記述，映画の内容と無関係な記述などのノイズが多く含まれる．そのため，映画に関連する内容のレビューを書くようにモデルを学習可能にするため，クレンジングが必要である．そのため，学習用のレビューデータセットから映画内容に関連する記述だけを抽出する．

映画内容に関連する記述の抽出の手順は，

（1）　レビューデータに含まれるすべての単語の登場映画数を計算し，

（2）　登場映画数により映画特徴語を抽出し，

（3）　映画特徴語を含むレビュー文を抽出する

という 3 工程からなる．

図 3 に，具体的な映画内容に関連する記述の抽出の流れを示す．はじめに，単語がレビュー中に登場する映画数を集計し，単語登場リストを作成する．次に，これらの単語登場リストの中から登場映画数が一定以下の単語を映画特徴語として抽出する．そして，それぞれのレビュー文の中で映画特徴語を含む文を抽出する．普遍的な表現を含む文を除去するという方法を取らなかった理由としては，具体的な映画の内容に言及している文であっても普遍的な表現を含む場合は多く存在するからである．たとえば，「面白い」や「すごい」といった表現は普遍的であるが，「二人の恋の駆け引きが面白い」「銃撃戦の迫力がすごい」などは具体的な映画の内容に言及した文と言える．

映画特徴語を抽出する際に，登場映画数の上限と下限を設けてクレンジングを行った．あまりにも少ない映画にしか登場しない単語は，固有名詞や作中に登場する人物名ばかりになる．逆に，あまりに多くの映画に登場する単語は，普遍的な単語である可能性が高い．そのため，予備実験を通して，複数の映画への感想に含まれ，かつ 1 %以上の映画へのレビューに含まれる語だけを残した．

最後にレビューデータから映画特徴語を含む文を抽出する．それぞれのレビューを文単位で分割し，映画特徴語リストの単語を 1 つでも含む文を抽出し，映画特徴語リストの単語を 1 つも含まない文は除去する．こうして映画内容に言及している文を抽出する．

3.3　学習データ用の重要部分の書き出しへの追加

実際のレビューは，必ずしも書き出しがレビュー全体の内容を表した一文であるとは限らない．具体的には，「主人公が自身を鼓舞するために叫ぶシーンがかっこよかった.」などの具体的すぎる記述，「ヒロインのひたむきな姿がかわいかった.」などの映画全体を評さない記述からレビューが書き出されている場合が多い．こういった書き出しを持つデータで言語モデルを学習した場合，入力された文の内容に対応する続きのレビューを予測するというタスクを言語モデルが正しく学習できない恐れがある．そこで，本節ではレビュー中の重要部分を書き出しへ移動する処理を施す．図 4 に，重要部分の書き出しへの移動の流れを示す．

はじめに，抽出型要約モデルを用いて，元のレビューからもっとも重要そうな部分を，要約文として抽出する．そして，この抽出された要約文が，レビュー全体と関連しているかを判定する．要約文と本文について，類似度がある一定以上の場合，この要約文はレビュー全体を表した文であるとみなせる．つまり，検索システム全体の入力である架空の映画の概要と近いものであると考えられる．そのため，抽出した要約文をレビューの書き出しに移動し，学習データとして用いる．

この際，要約文とレビュー全体の類似度が高いレビューのみを使用したのは，レビュー中にレビュー全体を表した文が存在しない場合を考慮したためである．たとえば，1 文目で映画のキャスティングについて言及し，2 文目で印象的なワンシーンについて言及し，3 文目で映画を観た後の感想を述べているようなレビューの場合，レビューで説明しているすべてのことをまとめた 1 文は存在しない．このようなレビューの要約文を書き出しへ移動しても，映画の概要からその映画に付きそうなレビューを生成するというタスクを言語モデルが学習できない可能性が高いため，このようなレビューは除いた．

このような処理を実現するために，はじめに，レビューを抽出型要約モデルに入力し，要約文を得た．具体的には，近年主流の大規模言語モデルを用いた抽象型要約器である BERTSUM [19]

を用いた．次に，要約文がレビュー文全体と意味的に類似するかを判定する．本手法の学習では，書き出しがレビュー文全体を表すような，概要的な文から始まるレビューだけを学習に用いたいと考えた．そのために，要約文が，レビュー文全体と意味的に類似するかを判定した．

意味的類似度の判定には，近年主流の類似度判定用モデルである Sentence-BERT [20] を用いた．Sentence-BERT では，2 つの文章を入力すると，その意味的類似度を 0 から 1 で計算できる．類似度が閾値（実験の際には 0.6 とした）以下だった場合は，そのレビューは学習用データから除外した．一方，類似度が閾値より高かった場合，学習用に，レビュー文から要約文と同じ文章を除去したうえで，先頭に移動した．これは，元のレビューから抽出された文を除かないと，1 つのレビュー文中に全く同じ文が 2 回登場すると，言語モデルがレビュー内で同じことを繰り返し記述する場合があるためである．こうして，重要部分を書き出しに移動したレビューからなる，学習用データセットが作成できた．

3.4 書き出しから続きのレビューを予測させるタスクによるモデルの学習

入力された書き出しに対する続きのレビューを生成可能にするために生成型言語モデルを追加学習する．学習には Next Token Prediction タスクを用いる．このタスクは，入力されたトークン列の次に来る確率が高いトークンを予測する機械学習タスクである．

言語モデルの事前学習では，Next Token Prediction を繰り返すことで文章を生成する．このタスクはラベルを必要としない，自己教師あり学習である．このタスクでは，入力トークン列に対して次に来る確率が最も高いと予測されたトークンを選択する．この際利用される損失関数 $L(U)$ は，言語モデルの語彙を $U = u_1, u_2, ...u_N$，言語モデルのパラメータを θ，言語モデルの語彙中に含まれる i 番目のトークンが，入力されたトークン列 $u_k, ...u_{-1}$ の次に来る確率を $p(u_i|u_k, ...u_{-1}, \theta)$ とした際に，

$$L(U) = \sum_{i=1}^{N} \log(p(u_i|u_k, ...u_{-1}, \theta)) \quad (1)$$

と定義される．

4 評　　価

提案手法の生成したレビューが映画検索における有用性を向上させているかを検証するために，生成レビューの定量的評価と仮想レビューを利用して検索した映画の定性的評価の 2 つの評価実験を実施した．

4.1 データセット

本実験における生成型言語モデルの学習および仮想レビューを利用した検索した映画の定性的評価には，IMDb Review Dataset - ebD[1]を利用した．このデータセットは 453,528 件の映画等に対して投稿された 5,571,499 件のレビューデータから構成されている．この中から，232,622 件の映画に対するレビューを使用した．また，投稿されたレビューの数が 10 件を超える映画に関してはその映画に投稿されたレビューの中から 10 件をランダムサンプリングして使用した．サンプリング後のレビューの総数は 1,007,151 件である．

4.2 比 較 手 法

提案手法の生成レビューに対する効果を検証するために，

- **ベースライン**：生成型言語モデルに対してレビューデータを用いた追加学習をせず，そのままのモデルでレビューを生成する手法，
- **追加学習のみ**：処理を施していない生のレビューデータで追加学習した生成型言語モデルでレビューを生成する手法，
- **映画に関連する文の抽出＋追加学習**：映画内容に関連する記述の抽出を施したレビューデータで追加学習した生成型言語モデルでレビューを生成する手法，
- **提案手法**（書き出し移動＋映画に関連する文の抽出＋追加学習）：重要部分の書き出しへの移動と映画内容に関連する記述の抽出を施したレビューデータで追加学習した生成型言語モデルでレビューを生成する手法

の 4 つの方法でモデルをトレーニングし，それぞれの生成レビューを人手で評価した．

4.3 実　　装

実験のために，実際に生成型言語モデルにレビューデータを学習させ，任意の書き出しの続きを生成できるようにした．生成型言語モデルとして，近年主流である GPT-2 [21] を用いた．学習データ用の重要部分の推定に，抽出型要約モデルである BERTSUM を用いた．また，文の類似度比較のために，SentenceBERT の paraphrase-MiniLM-L6-v2[2] モデルを用いた．

追加学習のみの手法の学習には，1,007,151 件のレビューから 10 万件のレビューをランダムにサンプリングして使用した．映画に関連する文の抽出＋追加学習の手法の学習には，10 万件のサンプリングされたレビューから，映画内容に関連する記述のみを抽出して用いた．提案手法の学習は，映画内容に関連する記述を抽出した 10 万件のレビューで追加学習をしてから，さらに重要部分の書き出しへの移動した 10 万件のレビューで追加学習を行った．

GPT-2 の追加学習時はすべての場合で，エポック数を 3，ブロック数を 512，訓練時のバッチサイズを 1 とした．fast tokenizer の使用は False に設定し，その他のハイパーパラメータはデフォルト値のままで学習と生成を行った．

4.4 生成レビューの定量的評価

本節では，提案手法が生成した仮想レビューの映画検索にお

1：「IMDb Review Dataset - ebD」：
https://www.kaggle.com/datasets/ebiswas/imdb-review-dataset
2：haggingface「paraphrase-MiniLM-L6-v2」：
https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

ける有用性検証のための被験者実験について述べる．本実験では，あらかじめ設定したクエリから提案手法および比較手法の生成したレビューの映画検索における有用性を 4 つの評価項目で被験者に評価させた．

仮想のレビューを生成するための入力クエリを 48 件，あらかじめ人手で作成した．これらのクエリは，すべて，「This is a movie like STAR WARS taken by Takeshi Kitano」のような，存在しない映画の概要である．1 件のクエリに対して 4 つの比較手法からそれぞれ 3 件ずつ，計 12 件のレビューを生成した．そのため，生成レビューの総数は 576 件である．

今回はこれらの生成レビューを，12 人の被験者に評価させた．1 件のレビューにつき 3 人の被験者で評価し，その平均値を評価値として用いた．1 人の被験者には 12 件のクエリに対する 144 件の生成レビューについて評価させた．なお，レビューが英語であるのに対し，被験者は全員，日本人である．そのため，実際の実験では元の英語のレビューと Google 翻訳によるレビューの日本語訳を同時に提示した．提案手法および比較手法の生成したレビューを，

- 映画内容言及率,
- クエリ説明性,
- 視聴促進度,
- クエリ説明単語量

の 4 項目で被験者に評価させた．

映画内容言及率は生成レビューを構成する文のうち，具体的な映画の内容に言及している文の割合がどのくらいであるかという評価項目である．被験者には，レビュー各文に対して，この文が具体的な映画の内容に言及しているかどうかを，言及しているなら 1，していないなら 0 で評価させた．そして，レビューを構成する文の数に対する映画内容に言及する文の数の割合で評価した．

クエリ説明性は生成レビューがクエリをどのくらい説明できているかという評価項目である．被験者には，生成レビューがクエリをどのくらい説明できているかを 1 を最低評価，5 を最高評価とした 1 から 5 の 5 段階で評価させた．

視聴促進度は，仮にクエリで表されるような映画が観たいと思っているとき，この生成レビューが投稿された映画をどれくらい観たくなるかという評価項目である．被験者には，生成レビューがどのくらい映画の視聴を促進するかを 1 を最低評価，5 を最高評価とした 1 から 5 の 5 段階で評価させた．

クエリ説明単語量は，生成レビューを構成する単語のうち，クエリを説明している単語の割合がどのくらいであるかという評価項目である．被験者には，生成レビューの中からクエリを説明していると思う表現をすべて抽出させた．なお，抽出する表現は 1 単語に限定しない．このようにして 1 件の生成レビューから抽出された表現すべてを構成する単語数を求めることで，生成レビューを構成する単語数に対する割合を算出し，クエリ説明単語量を評価した．

4.5 生成レビューの定量的評価の結果

各手法の生成したレビューに対する被験者実験の結果を表 1 に示す．表 1 から，提案手法はすべての評価項目において，ベースライン，追加学習のみの 2 つの手法と比較して有意な精度向上が見られた．しかし，提案手法の精度と映画に関連する文の抽出+追加学習の精度の間にはすべての評価項目で有意差が見られなかった．映画内容言及率，視聴促進度といった評価項目では，映画に関連する文の抽出+追加学習が提案手法と比較して高精度という結果となった．クエリ説明単語量においては有意差はなかったものの，提案手法が映画に関連する文の抽出+追加学習を上回る結果となった．

表 1　提案手法（書き出し移動＋映画に関連する文の抽出+追加学習）と比較手法の生成したレビューに対する被験者評価の結果（Student t 検定, **: $p < 0.01$, *: $p < 0.05$）

手法	映画内容言及率	クエリ説明性	視聴促進度	クエリ説明単語量
提案手法	0.29	0.24	0.21	**3.28%**
映画に関連する文の抽出+追加学習	**0.30**	**0.24**	**0.22**	2.74%
追加学習のみ	0.20**	0.16**	0.16**	2.18%*
ベースライン	0.21*	0.09**	0.09**	1.41%**

表 2　提案手法のクエリ説明性が最も高かったクエリと最も低かったクエリから生成されたレビューの例

	クエリ	生成レビュー	平均クエリ説明性
クエリ説明性の最も高かった生成レビューの例	This is a movie of western, but samurai	This movie has been made a decade before samurai movies which were made for the Japanese audience. The samurai movies and movies from that era did have the use of a lot of stylized weapons in their games, which is a nice feature that you could expect from samurai movies. Although there are some flaws in the movie, I think it is good and watch it for the style of the movies. My only criticism is the movie lacks realism in its depiction of the samurai at first. Also, I would have liked to see the movie take place in some kind of western town, but it would have been better if the movie has a more Western setting and that was the case. If you like samurai movies and you like the movies more than samurai movies, it is definitely worth to watch this one. I am not kidding: If you liked samurai movies, you will like this one.	0.75
クエリ説明性の最も低かった生成レビューの例	This is a movie that makes you unbelievable everything	The movie shows you the world of the main character's life as he travels around the world with a different purpose. At the beginning it is obvious that the plot is extremely complicated due to the fact that the main character is not at all human. But then you understand that he is also a normal person because when he goes to the movies he is all dressed in black and a strange woman comes across him. His acting is extremely believable especially in the scenes where he tries to escape from the strange woman by using a wheelbarrow and then through a window.The story was amazing and it was very interesting to think about it, even if you don't know much about Thailand. The way of seeing the movie is a combination of watching the movie with your typical Thai movie experience, as if you are going through a similar movie experience. I know that there are Thai movies that does not use the same genre as the movie	0.0

次に，提案手法から生成されたレビューが被験者から高い評価を得たクエリと低い評価を得たクエリについて議論する．提案手法のクエリ説明性が最も高かったクエリと最も低かったクエリから生成されたレビューの例を表 2 に示す．クエリ説明性が最も高かった「This is a movie of western, but samurai」というクエリから生成されたレビューには，「samurai」や「Western」といった表現は生成されているが，クエリに含まれる以上の情報はほとんど存在しなかった．クエリ説明性が最も低かった「This is a movie that makes you unbelievable everything」というクエリから生成されたレビューには，映画の内容に言及した文が含まれているがクエリと関連するような表現は存在しなかった．

4.6 仮想レビューを利用して検索した映画の定性的評価

実際に生成されたレビューを用いることで，どのような映画が検索可能になるかを検証するために，被験者実験を実施した．本実験では，提案手法から生成された仮想レビューを用い

表 3　提案手法が生成した仮想レビューによる映画検索が有効に働いた例（クエリ：日本版の騎士道物語）

クエリ	This is a movie like a Japanese version of Tales of Chivalry
クエリを説明している表現	‘some interesting side stories about two men who are both friends’, ‘a movie about two lovers who are fighting over Chivalry in the Japanese colonial war.’, ‘The story of The Sword and Sandal is a very modern one, as it concerns the life and death of a knight from a noble family.’, ‘One day a girl comes to meet him and she starts looking for him, but a samurai comes after her and takes her away.’, ‘samurai’
クエリを説明している表現から検索された映画	Shikonmado- Dai tatsumaki(1964) Junko’s Shamisen (2010) Bushi no ichibun(2006) Otsuyu:Kaidanbotan-dôrô(1998) Za samurai (1987) Kakushi-torideno san-akunin(1958) Bushidô zankokumonogatari(1963) The Sea Is Watching (2002) Miyamoto Musashi (1954) The Twilight Samurai (2002)

表 4　提案手法が生成した仮想レビューによる映画検索が有効に働かなかった例（クエリ：マトリックスのようなホラー）

クエリ	This is a movie of Matrix-like horror
クエリを説明している表現	‘attacking humans’, ‘The movie isvery violent’, ‘raped and murdered hundreds of women’, ‘enemy’, ‘using thelatest technology’, ‘fear’, ‘kill’, ‘horror’, ‘Matrix’, ‘The film is full of a very real sense of fear and horror.’, ‘murdered’
クエリを説明している表現から検索された映画	The Houses October Built (2014) 咒怨 2 (2003) Voodoo Dolls (1991) Un uomo, un cavallo, una pistola (1967) Take My Eyes (2003) 駭客任務完結篇：最後戰役 (2003) O fovos (1966) Jûsan-nin renzoku bôkôma (1978) Napasta (1982) Village of the Damned (1995)

て映画を検索し，検索された映画とクエリの関連度を主観で評価した．被験者は 1 名で，検索結果の映画を知らなかった場合，ウェブでその映画について調べ，その映画がクエリと適合するかを判断した．また，主観的に，検索結果の映画や，抽出されたキーワードが正しかったかを分析した．

4.4 節の実験で被験者が抽出した「クエリを説明している表現」を，元のクエリを拡張するクエリとして使用する．BM25 を用いて，被験者が抽出した「クエリを説明している表現」を入力し，関連度の高いレビューをランキングする．検索対象は，4.1 節で紹介した 232,622 件の映画に対する 1,007,151 件のレビューである．ランキングされたレビュー上位 10 件の投稿先の映画を検索結果として，それぞれの映画の内容を確認した．実際の検索には，提案手法が生成した 3 件レビューに対して 3 人の被験者が抽出した「クエリを説明している表現」すべてを用いた．また，検索モデルとして，Python 用ライブラリである Rank BM25[3] の Okapi BM25 を用いた．

4.7　仮想レビューを利用して検索した映画の定性的評価の結果

提案手法が生成した仮想レビューによる映画検索が有効に働いた例を表 3 に示す．この結果では，「This is a movie like a Japanese version of Tales of Chivalry」というクエリに対して，武士や侍が主役の映画が多数登場している．武士や侍は，クエリ中の「Japanese」「Tales of Chivalry」といった要素と関連がある上に，これらを直接的に表す表現はクエリに登場していないため，仮想レビュー生成が映画検索に有効に働いた例と言える．また，実際に生成されたレビューから被験者が抽出した「クエリを説明している表現」には，「samurai」という単語が登場している．

被験者からの評価が高かったにもかかわらず提案手法が生成した仮想レビューによる映画検索が有効に働かなかった例を表 4 に示す．この結果では，「This is a movie of Matrix-like horror」というクエリに対して，「Matrix」らしさのないホラー映画ばかりが登場している．実際に生成されたレビューから被験者が抽出した「クエリを説明している表現」は，ほぼすべて「horror」に関係する表現であった．

5　考　　察

生成レビューの定量的評価の結果から，映画内容に言及した文の抽出は生成レビューの映画検索における有用性向上に大きく貢献することが明らかとなった．このことから提案手法では，映画内容に言及した文を精度よく抽出できていたため，言語モデルが与えられた書き出しの内容を表す続きの文を生成するというタスクを正しく学習できたと考えられる．

一方，重要部分の書き出しへの移動は生成レビューの映画検索における有用性向上にあまり効果がないことが明らかとなった．このことについて考察するために，実際に提案手法で使用した抽出型要約モデルが抽出したレビュー要約の例を提示する．抽出型要約モデルが抽出したレビュー要約の中に「Skye insists that it was some sort of evil some kind of “darkness” which was the cause of her boyfriend’s death.」というものがあった．これは，映画に登場する人物の主張を要約したものである．本研究では，映画の内容を端的に表した概要からのレビュー生成を目的としているため，映画に登場する人物の主張を要約したこの文は目的と齟齬がある．このように，学習に使用したレビューの書き出しすべてが映画の内容を端的に表した概要となっていなかったため，生成レビューの映画検索における有用性向上にあまり効果がなかったことが考えられる．

仮想レビューを利用して検索した映画の定性的評価の結果から，クエリによっては映画検索に仮想レビューを利用することが有効であることが明らかとなった．このことから，提案手法で学習した生成型言語モデルが，クエリに含まれる要素に関連する表現を学習できていたこと，入力された書き出しの内容を説明するレビューを生成するというタスクを理解できていたことが考えられる．しかし，多くのクエリからは，クエリを網羅的に説明できる表現が生成されないという結果になった．このことから，生成型言語モデルはあくまで入力文に対する続きを予測するモデルであるため，入力文に要素が複数存在する場合，特定の一部の要素に対する続きを予測しやすいと考えられる．

3：「Rank BM25」：
https://pypi.org/project/rank-bm25/

6 まとめと今後の課題

本研究では，生成型言語モデルを用いて仮想レビューを生成し，生成されたレビューと類似する実際のレビューの投稿された映画を検索する手法を提案した．生成型言語モデルに，実際のレビューを用いて，書き出しの続きとなる文を予測させることで仮想レビューを生成可能にした．この際，生成レビューが与えられた書き出しの内容を具体的に説明したものになるように，学習用レビューから映画内容に関連する記述を抽出し，重要部分を書き出しへ移動した．提案手法の生成したレビューが映画検索においてどれだけ有用であるかを検証するために，生成レビューの定量的評価と仮想レビューを利用して検索した映画の定性的評価を実施した．

生成レビューの定量的評価の結果として，映画内容に関連する記述の抽出は，生成レビューの映画内容言及率，クエリ説明性，視聴促進度の 3 つの評価値を有意に向上させた．一方，重要部分の書き出しへの移動は，今回の評価において有意な精度向上をもたらさなかった．これらのことから，映画内容に関連する記述の抽出が生成レビューの映画検索における有用性向上に大きな影響があることが明らかになった．仮想レビューを利用して検索した映画の定性的評価の結果として，一部のクエリでは映画検索に生成レビューを利用することで，クエリの情報を拡張可能であった．一方，多くのクエリでは，クエリの全ての要素を網羅的に説明する表現が生成されなかった．これらのことから，まだ精度に改善の余地があるものの，生成型言語モデルを用いた仮想レビューが映画検索において有用であることを示した．

今後の課題として，生成レビューに含まれる情報量の少なさの改善，生成レビューを有効に活用するための検索アルゴリズムの提案が挙げられる．提案手法が生成したレビューには，クエリに含まれる表現がそのまま登場してしまうことが多く，それらを具体的に説明するような表現はあまり登場しなかった．これに対しては，映画内容に言及するレビューの抽出精度の改善，レビュー本文に対する書き出しを説明する表現の追記などの対策が考えられる．また，今回の実験では，映画を検索するための拡張クエリを人手で抽出した．生成レビューからの映画検索手法について，今後検討する必要がある．

謝　　辞

本研究の一部は JSPS 科研費 21H03775，21H03774，22H03905 による助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. On application of learning to rank for e-commerce search. In *Proc. ofthe 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 475–484, 2017.

[2] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proc. ofthe 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2251–2260, 2020.

[3] Hanqing Lu, Youna Hu, Tong Zhao, Tony Wu, Yiwei Song, and Bing Yin. Graph-based multilingual product retrieval in E-commerce search. In *Proc. ofthe 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 146–153, 2021.

[4] Taghi M. Khoshgoftaar Xiaoyuan Su. A survey of collaborative filtering techniques. In *Advances in Artificial Intelligence*, Vol. 2009, p. 19, 2009.

[5] 杉木健二, 松原茂樹ほか. 消費者の意消費者の意見に基づく商品検索見に基づく商品検索. 情報処理学会論文誌, Vol. 49, No. 7, pp. 2598–2603, 2008.

[6] Chin-Sheng Yang, Chih-Ping Wei, and Christopher C. Yang. Extracting customer knowledge from online consumer reviews: a collaborative-filtering-based opinion sentence identification approach. In *Proc. ofthe 11th International Conference on Electronic Commerce*, pp. 64–71, 2009.

[7] Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, Vol. 11, No. 23-581, p. 81, 2010.

[8] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proc. ofthe 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 61–69, 1994.

[9] Donna K Harman. *The first text retrieval conference (TREC-1)*, Vol. 500. US Department of Commerce, National Institute of Standards and Technology, 1993.

[10] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, Vol. 19, No. 1, p. 1–27, 2001.

[11] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Probabilistic query expansion using query logs. In *Proc. ofthe 11th International Conference on World Wide Web*, p. 325–332, 2002.

[12] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, Vol. 41, No. 4, pp. 288–297, 1990.

[13] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proc. ofthe 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11, 1996.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. ofthe 31st International Conference on Neural Information Processing Systems*, p. 6000–6010, 2017.

[15] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

[16] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proc. ofthe 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1762–1777, 2023.

[18] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query

expansion with large language models. In *Proc. ofthe 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423, 2023.

[19] Yang Liu. Fine-tune bert for extractive summarization. *arXiv*, 2019.

[20] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. ofthe 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

[21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

SNS コメントを利用した対照学習による ニュースの潜在的な関係性の抽出と可視化

高橋　陸†　牛尼　剛聡††

† 九州大学大学院芸術工学府　〒 815-8540 福岡県福岡市南区塩原４丁目９－１
†† 九州大学大学院芸術工学研究院　〒 815-8540 福岡県福岡市南区塩原４丁目９－１
E-mail: †takahashi.riku.110@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年 SNS の普及に伴い，ニュースに関する情報収集を行う際に SNS を用いることが一般化した．SNS を用いた情報収集では，リアルタイムで更新される情報や投稿者から直接意見が得られるなどの利点がある．一方，キーワード検索による情報収集では個別の投稿を探すことは可能であるが，関連する投稿やニュース間の関係を把握することは困難である．本研究では，X(旧 Twitter) 上に投稿されたニュース記事と，それに対するリプライに着目し，機械学習を用いてニュースとリプライの関係を学習することで，ニュースに対する社会的反応の特徴を抽出する手法を提案する．本手法では，ニュース記事とリプライの文章ベクトルから，同一ニュースに関する，関連性の高い投稿間の文章ベクトルの類似度が高くなるように学習を行う．各投稿のベクトル間の類似度から，関連性の高いニュースや投稿を求め，ユーザが興味を持ったニュースに関連する情報を提示することで情報収集を支援する．

キーワード　SNS，対照学習，ニュース, ユーザ支援, 可視化

1　はじめに

近年，個人が手軽に情報を発信できる SNS(Social Networking Service) の普及に伴い，X(旧 Twitter) や Instagram など，多くの SNS で様々なトピックに関する投稿が行われている．SNS では過去の投稿に対して返信することが可能であることから，一つのトピックに対して様々なユーザによる意見や反応が集まることも珍しくない．また，SNS 上にはテレビ局や新聞社などの報道機関が，取材などで得た情報をリアルタイムにニュース記事として投稿している．こうした背景から，SNS は情報収集の手段として用いられることも多くなってきている．

新聞通信調査会による「第 13 回メディアに関する全国世論調査」[1] では，インターネットは情報源としての信頼度は新聞やテレビに劣る一方で，情報源として最も欠かせないメディアとして支持されている．また，NTT ドコモのモバイル社会研究所による「週 1 回以上アクセスして日常的に生活情報を得ているメディア」に関する調査 [2] では，10 代 20 代においてソーシャルメディアの情報収集としての利用率ははテレビや新聞による情報収集の割合を上回っており，特に若い世代を中心に SNS が情報源として広く利用されていることが分かる．

SNS を情報収集に用いることで，ユーザは多様な視点からの反応や意見を得ることが可能である．一方で，SNS 上に存在する情報は整理されておらず，関連する内容など，特定のトピックに関する情報を網羅的に取得することは困難である．また，SNS 上に投稿されたニュース記事では、ニュースの詳細情報が記載されている一方で，その因果関係や他のニュースとの関連といった情報は少ない．そのため，ニュース間の関連性の判断は，現時点ではユーザ個人の持つ知識や経験に依存している．

現在，ユーザが SNS を用いてニュースに関する情報収集を行う際に一般的に用いられている方法は，対象とするニュースに関する代表的なキーワードを用いたキーワード検索である．キーワード検索では，対象とするニュースに対してユーザが関係があると考える単語や，知りたい内容についてのキーワードをクエリとして投稿を検索する．例えば，新型コロナウイルスに関する情報を探す際には,「コロナ」や「感染状況」,「ワクチン」といった単語がクエリとして用いられることが多い．しかし，一般的に対象のニュースにおいて，そのニュースに関する SNS 投稿の全てに必ず含まれるような単語は存在しない．したがって，キーワードを用いた検索では，ユーザが欲しい情報に直接アクセスすることは困難であり，関連する情報を網羅的に集めるにはキーワードを変えて何度も検索を行う必要がある．

SNS 上に投稿されるニュース記事では，一般的に本文でニュースに関する詳細情報が述べられる．しかし，ニュース記事の本文内でニュースとなる事柄が起こった経緯や他のニュースとの関係について言及されることは少なく，単一のニュース記事から得られる情報から他のニュースとの関係を推察することは難しい．ここで，物価高騰に関するニュースを具体例として挙げる．物価高騰を扱ったニュース記事の本文では，現在の物価について過去のデータとの比較などが行われているが，なぜ物価高騰が起こっているのかについて十分に説明されていない．物価高騰の背景について，例えばウクライナにおける戦争の影響や日本の金融政策，その年の気候など様々な要因が考えられる一方，こうしたトピックと物価高騰との関係について，ニュース記事の本文で直接言及されることは少ない．以上のことから，従来のキーワード検索では特定のトピックに関する情報は取得できる一方で，複数のトピックにまたがるような話題，関係性について調べる際には適していない．現在，こうしたニュース間の関係性の判断は各ユーザが持つ経験や知識に委ねられており，SNS を用いた情報収集においても解決できていないことか

ら，現状では SNS を情報源として有効に活用できているとは言い難い．

本研究では、代表的な SNS の一つであった Twitter を用いて，Twitter に投稿されたニュース記事とそれに対するリプライの関係に着目する．投稿されたニュース記事に対するリプライコメントをそのニュースに関する情報をして用い，機械学習を用いてニュースコメントを介したニューストピック間の関係を学習する．その後，学習したモデルを用いてニュースやリプライの文書ベクトルを生成し，各ニュースや投稿の関係性を可視化，抽出することでユーザが情報収集を行う際の支援とすることを目的とする．

本研究の貢献は，以下の通りである．

- Sentence-BERT を用いた対照学習により，単純な文章の類似性に寄らないニュースや SNS 投稿の反応を考慮した文書ベクトルを生成する手法を開発した．
- 生成したニュースコメントのベクトルを用いてニュースに関する関連コメントを抽出，可視化することで情報収集に関して従来手法とは異なったユーザ支援を実現する手法を開発した．

本論文の構成は以下のようになっている．第 2 章では，関連研究の紹介を行う．第 3 章では，提案手法について述べる．第 4 章では，提案手法の有効性を評価するための実験内容について述べ，実験結果について述べる．第 6 章では，まとめと今後の課題について述べる．

2 関 連 研 究

これまでにも，文章の特徴から，機械学習を用いて関連する情報を抽出する手法が提案されている．ここでは，それらの手法のうち，提案手法に関連する手法について述べる．

2.1 ニュース理解支援に関する研究

神谷 [3] らは，ウェブ上から社会問題に関連する情報を抽出する手法を提案している．この手法では，対象とするウェブ上のテキストに対して，それが社会問題に関連する内容であるかを判断し，関連する社会問題に分類している．この手法では，日本語 wikipedia 中の，社会問題に関する記事を教師データとして，自然言語処理の代表的な機械学習モデルの一つである BERT [4] を用いたテキスト分類モデルを構築している．この手法では，対象とするテキストが社会問題に関連するか否かを判定した後に，どの社会問題に関連するかを分類している．評価実験では，社会問題であるか否かの判定に関しては良い結果が得られなかった一方で，社会問題であると判定された記事を分類するタスクでは高い精度であった．

この手法では，社会問題との関係が明示的に示されていないテキストを，社会問題に分類するという点が本研究の提案手法と類似している．しかし，本論文の提案手法では SNS 上の投稿を学習データとして用いるため，Wikipedia に掲載されていないニュースに関しても対応可能である．

2.2 文書の特徴抽出に関する研究

大倉 [5] らは，ニュース記事に対する関連記事を判定するためのキーフレーズ抽出方法を提案している．この手法では，記事内容の特徴を捉え，かつ同一トピック内で広く使用されるフレーズを抽出するために，記事内での単語の出現頻度と類似記事における出現数からフレーズを抽出し，RNN [6] を用いて学習を行うことで関連記事を判定するためのキーフレーズを抽出している．

この手法では，文章中に現れるフレーズを用いて機械学習モデルに類似記事の学習を行っている．関連する記事を得るという点で本研究の目的を類似している．しかし，本研究では SNS 投稿を用いて間接的にニュース記事間の関係性を求めるため，記事中に含まれるフレーズに関係なくニュース記事の特徴を抽出するという点で異なる．

3 提 案 手 法

3.1 概　　要

本研究では，入力として用いる文章間の関係性を学習するため，代表的な自然言語処理モデルの一つである BERT [4] に対して，文書ベクトルの生成に関して改良を行った Sentence-BERT [7] を用いる．Sentence-BERT は，事前学習を行った BERT モデルに加え，Siamese Network [8] を用いることで BERT よりも高精度な文章ベクトルを生成可能である．BERT によるベクトル生成ではトークンごとにベクトル化を行い，利用する際には最終層の cls トークン部分を文章ベクトルとして用いるなどことが多い．一方，Sentence-BERT では各トークンのベクトルを pooling 層で平均化することで文章全体を一つのベクトルとして表現する．その後，図 1 に示す様に，入力として与えられた二つの文章のベクトル化について類似する文章同士が近いベクトルになるように学習を行うことで，文章間の関係を文書ベクトルに反映する．

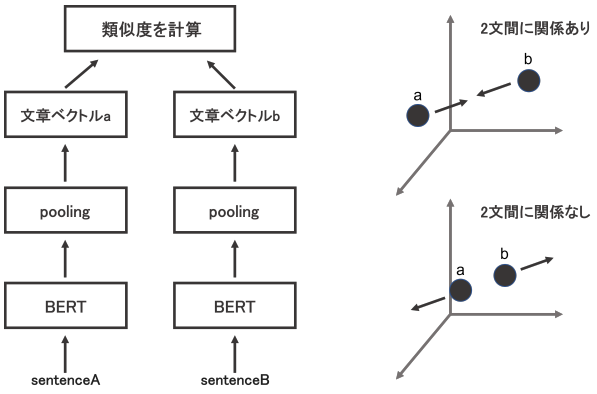


図 1　Sentence-BERT の学習

本研究の概要を図 2 に示す．Sentence-BERT による学習方法として CrossEncoder を用いる．CrossEncoder は 2 つの文章を入力として用い，付与されたラベルに基づいて，それぞれの文章のベクトルのベクトル空間上での位置関係を学習する．本研究では，入力文として SNS に投稿されたニュース記事に対

するリプライを用いる．リプライ関係の有無を基にニュースコメント間の関係を学習することで，ニュースやコメント間の関係を反映したニュースコメントベクトル空間を生成し，ニュースコメントのベクトルを生成する．その後，各ニューストピックごとに階層クラスタリングを行うことでニュースコメントに含まれる類似意見をクラスタ化し，クラスタ間の距離を計算することで近いクラスタのコメントを関連情報として抽出する．

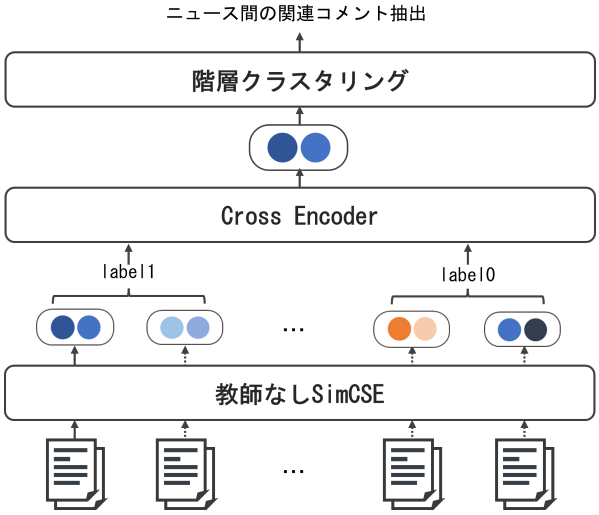


図 2　提案手法の概要

3.2　対照学習

本研究では SNS に投稿されたニュースコメントを用いてニュース間の関連性を考慮したベクトルを生成するため対照学習を用いる．対照学習とは，自己教師あり学習の一種であり，データ同士の比較によってデータ間の関係を学習する手法である．一般的に，類似したデータや関連データはベクトル空間上で近い位置に，異なるデータは遠くなるよう学習を行う．マルチクラス分類を行うモデルと異なる点として，入力データから具体的なクラスラベルを予測するのではなく，データ同士の関係から目的に合うベクトル表現を学習する．そのため，無数に存在する各ニュースとそのコメントに対してラベルを付与する必要がない点で本研究の目的に適している．

本研究では，ニュースやコメント間の関係を反映したニュースコメントベクトル空間の生成が目的であることから，ニュースコメントの適切な文表現を取得するため，Sentence-BERT を用いて教師なし SimCSE [9] 学習を行った後，CrossEncoder [10] モデルの学習を行う．教師なし SimCSE では適切なニュースコメントのベクトル表現について学習を行い，ニュースコメント間の類似度計算などの精度向上を目的とする．CrossEncoder モデルでは，SNS 上のリプライ関係の有無を基にニュースコメント間の関係を学習することで，ニュースやコメント間の関係を反映した文章ベクトル生成を目的とする．

3.2.1　教師なし SimCSE

図 3 に教師なし SimCSE の仕組みを示す．教師なし SimCSE とは，文の埋め込み表現について学習を行う対照学習の一種である．同じ文章のベクトルを 2 つ生成し，そのペアを正解とし，異なる文章のベクトルのペアを不正解として学習を行う．本研究において入力はニュースコメントであり，入力ニュースコメントを $c_1, c_2, ..., c_n$ とするとき，BERT モデルを適用してベクトル $h_1, h_2, ..., h_n$ を得る．

$$h_i = vecBERT(c_i) \tag{1}$$

次に，ドロップアウトを適用してノイズを加えた各ニュースコメントのベクトル
$h'_1, h'_2, ..., h'_n$ を得る．

$$h'_i = vecDropout(h_i) \tag{2}$$

このとき，h_i と h'_i は同じニュースコメント c_i から作られたベクトルであるのの，ドロップアウトの働きにより若干異なったベクトル表現となっている．これらのベクトルを用いて，式 (3) で示すようにモデルは同じ文から生成されたベクトルが近くなるように学習を行う．また，式 (4) に示すように，異なる文のベクトルはベクトル間の距離がマージン m より大きくなるように学習することで，より良い文表現を獲得することができる．

$$Loss_{textsame} = sum_{i=1}^{n} |h_i - h'_i|^2 \tag{3}$$

$$Loss_{textdiff} = sum_{i \neq j} max(0, m - |h_i - h'_j|^2) \tag{4}$$

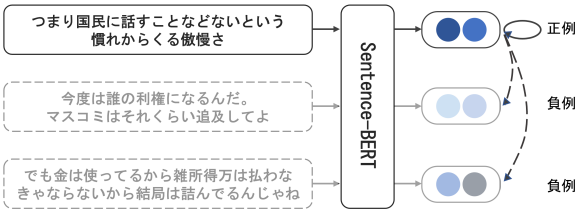


図 3　教師なし SimCSE

3.2.2　CrossEncoder モデル

CrossEncoder とは，文章のペアを入力として，与えられたラベルに基づき 2 つの入力文章が一致する確率を計算，学習する手法である．図 4 に示すように，学習時には，入力文章ペアのラベルが 1 の場合は 2 つの文が一致確率が高いものとして学習を行い，ラベルが 0 の場合は異なる確率が高いものとして学習を行う．本研究では，SNS 上に投稿されたニューストピックに対するコメントを用いることから，コメントの集合を

$$N = \{n_1, n_2, \cdots, n_{|N|}\} \tag{5}$$

$$C = \{c_1, c_2, \cdots, c_{|C|}\} \tag{6}$$

とする．このとき，コメント $c_i (i = 1, 2, \ldots, |C|)$ が，対象としているニューストピックを $news(c_i)$ と表記する．CrossEncoder では，入力文章ペアに対して 0 または 1 のラベルを付与する必要がある．そこで本研究では，ラベルを式 (8)，(9) のように決定する．

$$Pair_{label0} = (c_i, c_j | news(c_i) \neq news(c_j)) \quad (7)$$

$$Pair_{label1} = (c_i, c_j | news(c_i) = news(c_j)) \quad (8)$$

こうして作成した入力文章ペアを Sentence-BERT に渡し，コメントのペアに対するスコア y_{pred} を計算する．

$$y_{pred} = CrossEncoder(c_i, c_j) \quad (9)$$

ここで y_{pred} は入力されたコメントのペアの類似度を示す値であり，0 から 1 の間の値を取る．入力文章ペアに付与されたラベルを y_{true} とするとき，CrossEncoder では式 (10) に示すように，y_{pred} と入力文章ペアに付与されたラベル間の差が最小になるよう学習を行う．これにより，単純な文章の類似性に留まらない，ニュースやコメント間の関係を考慮した文章ベクトルの生成を目指す．

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_{true,i}\log(y_{pred,i}) + (1 - y_{true,i})\log(1 - y_{pred,i}) \quad (10)$$

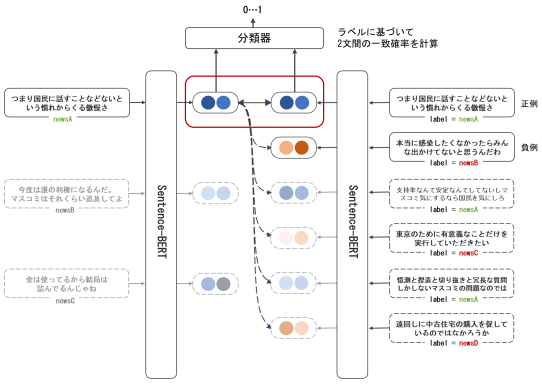


図 4　CrossEncoder の仕組み

3.3　ニュースコメントベクトルを用いた関連情報の抽出と可視化

図 5 に関連情報抽出の流れを示す．学習を行った CrossEncoder モデルにより生成されたニュースコメントベクトルを用いて，ニュースに関する関連情報の抽出と可視化を行う．ニューストピック $n_i (i = 1, 2, \ldots, |N|)$ に対するコメント集合を

$$C_{n_i} = \{c | news(c) = n_i\} \quad (11)$$

とする．この C_{n_i} について階層クラスタリングを行い，各ニューストピックに含まれる類似意見をクラスタにまとめる．

$$K_{n_i} = HierarchicalClustering(C_{n_i}) \quad (12)$$

ここで得られた各クラスタ $k \in K_{n_i}$ について，式 (13) でクラスタの中心 μ_k を計算する．

$$\mu_k = \frac{1}{|k|}\sum_{c \in k} v_c \quad (13)$$

各クラスタの中心 μ 間の類似度について計算し，類似度が高いクラスタを関連クラスタとして扱う．最後に，関連クラスタに含まれるコメントを関連コメントとして抽出を行う．

$$sim(\mu_k, \mu_l) = \frac{\mu_k \cdot \mu_l}{||\mu_k||_2 ||\mu_l||_2} \quad (14)$$

続いて，可視化図の作成について説明する．提案手法により生成されたニュースコメントベクトルは 768 次元で表現されており，このままでは可視化に用いることができないから，次元削減を行うことで 2 次元に変換する．一般に次元削減に用いられる手法として PCA [11] などが存在するが，本研究では他に，t-SNE [12] および，UMAP [13] を用いてニュースコメントの次元削減を行い，比較する．

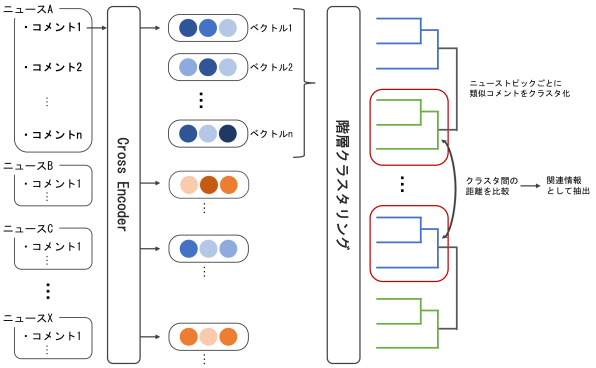


図 5　関連コメント抽出の流れ

4　評価実験

4.1　実験方法

本実験では，提案手法を用いてニュースコメント間の関係性の学習を行うほか，比較のため BERT によるマルチクラス分類モデルの学習を行った．2 つのモデルに関して，同一のデータセットにより学習を行ない，ニュースコメントベクトルを生成した．BERT によるマルチクラス分類モデルでは，ニュースコメントとリプライ関係にあるニューストピックを正解ラベルとして学習を行った．学習中に検証 loss の値が最も低かった際のモデルを保存し，最終層の cls トークン部分を文章ベクトル生成に用いた．各モデルに対する評価として，階層クラスタリングを用いた関連情報抽出を行った．その後，ニュースコメントベクトルを 2 次元に変換し，散布図に可視化することで各ニュース間の関係について分析した．

4.2　追加事前学習

先行研究において，Gururangan [14] らは，既に一般的なコーパスで事前学習を行ったモデルに対して，追加で解かせたいタスクに関連するデータセットで事前学習を行うことで精度が向上することを示している．本研究ではニュースに関する SNS 投稿を用いることから，文語的な整った文章だけでなく口語的な文章が使用するデータに含まれる．そのため，ベースとして用

いる日本語 BERT モデル [15] の事前学習に使用されたデータセットでは提案手法で学習を行うタスクに対して十分ではないと考えられる．そこで，本研究では日本語 Wikipedia による事前学習に加え，Twitter から取得したニュース記事に対するリプライ計 99243 件を用いて追加事前学習を行った．追加事前学習には Masked Language Model と Next Sentence Prediction を用いた．

4.3 教師なし SimCSE による類似度計算

より適切なニュースコメントベクトルを生成するため，教師なし SimCSE による学習を行った．学習データには追加事前学習と同じデータを用いた．学習中のスピアマン係数の値が最も良いモデルを保存し，追加事前学習を行っていない日本語 BERT モデル，追加事前学習を行った BERT モデルと比較を行った．図 6 から図 8 に学習に用いていないニュースコメントのベクトルを生成し，類似度を計算した結果を示す．CrossEncoder モデルの学習に用いるニュースコメント 6703 件に関してすべての組み合わせのニュースコメントペアを作成し，類似度を計算した結果を見ると，ニュースコメントを用いた追加事前学習なしの日本語 BERT モデルと追加事前学習有りの日本語 BERT モデルでは全体的に類似度の値が高く出ていることが分かる．一方，教師なし SimCSE による学習を行ったモデルでは類似度の最頻値は約 0.2 であり，全体的に低いことが分かる．類似度計算に用いたニュースコメントは，16 件の異なるニューストピックに対するリプライから構成されており，内容が類似しているコメントはそうでないコメントと比較して少ないことから教師なし SimCSE による学習は適切なニュースコメントのベクトル生成において効果的であると考えられる．教師なし SimCSE 以外の 2 つのモデルによって計算した類似度が高いことに関して，この 2 つのモデルでは文章中に出現する単語の頻度が類似度に強く影響されていると考えられる．ニュースコメントでは他の文章と異なり，文章としては異なる内容であっても出現する単語は類似している場合が多く存在する．このようなコメントについて，教師なし SimCSE では正しく類似度を計算できていると考えられる．

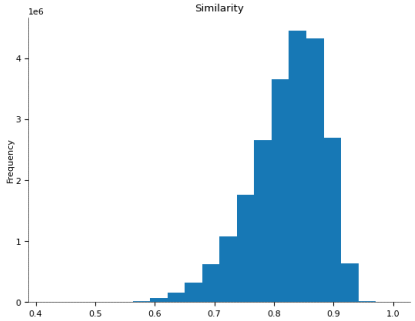


図 6　類似度計算の結果（追加事前学習なし）

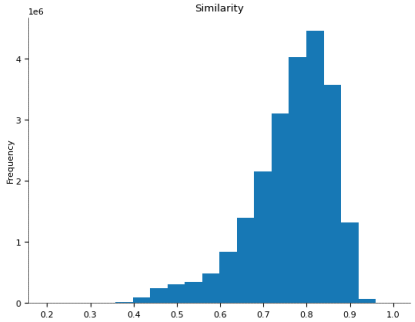


図 7　類似度計算の結果（追加事前学習あり）

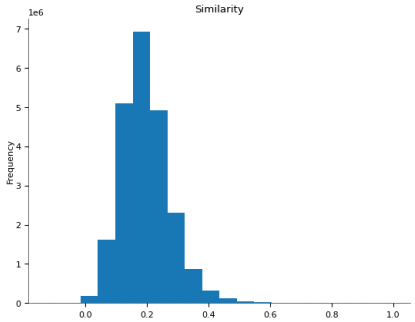


図 8　類似度計算の結果（SimCSE）

4.4 データセットの作成

データセットとして，本研究では代表的な SNS の一つである Twitter を用いた．収集したニューストピックについて，表 1 に示す，Twitter API を利用して 2022 年 5 月 9 日から同月 24 日の間に Twitter に投稿されたニューストピック 16 件と各ニューストピックに対するニュースコメントの取得を行った．取得したニュースコメントを用いて，同じニューストピックに対するコメントのペアを作成し，学習を行った教師なし SimCSE モデルを用いてベクトルを生成，類似度を計算した．その結果から，同じニューストピックに対するコメントかつ類似度が 0.5 以上のペアを正例としてラベル 1 を付与した．次に，異なるニューストピックに対するコメントのペアを作成し，正例と同じ件数をランダムに選択し負例としてラベル 0 を付与することで合計 20828 件のデータセットを作成した．

表 1　データセットに用いたニュース一覧

No.	ニュース記事見出し
0	岸田首相は当初は積極的に取材に応じる姿勢を示していた…
1	先進 7 カ国 7 は 49 日オンラインで首脳会議を開きロシア産…
2	日経平均株価は先週末の終値より 684 円安い 2 万 6319 円…
3	新型コロナウイルスの水際対策について緩和すべきと..
4	財務省は 10 日税収で将来返済する必要がある国の借金長期…
5	新型コロナウイルスについて東京都は 8 日新たに 4711 人の…
6	東京都医師会は 101 日午後に行われた定例会見で段階的に…
7	小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を…
8	ロシアの国連代表部に勤務してきた露外交官、ボリス…
9	近年、「体育座り」は子供の体に負担がかかるとの指摘が…
10	マイナンバーカードと健康保険証の機能を併せ持つ「マ…
11	国民審査“ 在外の日本人が投票できないのは違憲 ”最高裁
12	東京都は、一戸建て住宅を含む都内の新築建物に、太陽…
13	「他意はなく、不用意な発言だった。心からおわび…
14	性行為伴うＡＶ禁止する法制定を別途検討　立憲民主党
15	【誤送金問題】「約 4299 万円を法的に確保した」阿武町…

表 2　BERT による関連ニュースと関連コメント抽出の例

ニュースコメント	ニュース	平均関連度
当然ガソリン税無くすんですよね〜	G7 は 8 日、ロシア産油輸入の「段…	3.6470
	東京都は、一戸建て住宅を含む都…	1.6764
予言あるある言いたい. 電力不足で一部地域で使えません震災で使えません	マイナンバーカードと健康保険証…	1.9117
	東京都は、一戸建て住宅を含む都…	3.8235
なんというか接客を伴う飲食店やポケカ爆買いとかに使ってなくてよかったよね	【誤送金問題】阿武町長が会見で…	4.1470
	マイナンバーカードと健康保険証…	1.2647
全額税収で回収して返済したら日本円何処にも流通しないことになるんですが…	財務省は税収で将来返済する必要…	4.5882
	【誤送金問題】阿武町長が会見で…	1.7352
電気代が上がると今年は建物内も冷房の設定温度を上げて暑くなりそう	東京都は、一戸建て住宅を含む都…	3.2647
	東京都医師会は段階的に感染対策を…	1.3235
祟りで大地震がこないことを祈ります	小池氏、今度は「神宮外苑再開発…	4.3823
	【誤送金問題】阿武町長が会見で…	1.2647

表 3　提案手法による関連ニュースと関連コメント抽出の例

ニュースコメント	ニュース	平均関連度
結論ありきで物事話し合うなよ。民意を無視して政治してたら支持出来ないよ。	東京都は、一戸建て住宅を含む都…	3.4705
自称有識者も会議の内容を公表しないとクリーンではないよな	マイナンバーカードと健康保険証…	3.5882
こんなアホな政策を実行する首領を選んだのは都民だぞ。分かってんか	東京都は、一戸建て住宅を含む都…	4.0588
	小池氏、今度は「神宮外苑再開発…	4.3823
有識者が知識を有してるとは思えないアホみたいな政策	東京都は、一戸建て住宅を含む都…	4.4411
	小池氏、今度は「神宮外苑再開発…	3.6176
感染対策怠ってた奴等がなってるだけ。最近マスクしてない奴結構見るようになったしね	新型コロナウイルスについ東京都…	4.6470
	東京都医師会は段階的に感染対策を…	3.9117
結局エコとか抜かす奴はこんな事を言うワケだ。全て利権とカネだ	小池氏、今度は「神宮外苑再開発…	4.3529
	東京都は、一戸建て住宅を含む都…	3.9411

4.5 結　　果

BERT によるマルチクラス分類モデルの学習結果について表 4 に示す．BERT の学習結果を見ると，SNS 投稿に多い主語や目的語が省略された文章であってもニューストピックの予測ができていることが分かる．データセット中にはニューストピックと関係のない内容のコメントの含まれていることから，BERT を用いることでニュースコメントからニューストピックの予測が可能であると言える．

表 4　コメントのニューストピック予測 (BERT)

	val acc	test acc	
BERT	0.734375	0.7058	-
random	-	-	0.0625

BERT によるマルチクラス分類モデルによる関連情報抽出の例を表 2，提案手法による関連情報抽出の例を表 3 に示す．表 3 は階層クラスタリングの結果から，異なるニューストピックに属するクラスタ間で類似度の高いクラスタのペアを取り出し，それぞれのクラスタが属するニューストピックと各クラスタに含まれるコメントの例である．

また，表 9 にニュースコメントのニューストピックに対する関連度評価の結果を示す．ニュースコメントのニューストピックとの関連度の評価は，クラウドソーシングを用いて募集した 34 名の被験者が行った．提案手法により抽出された関連コメントと BERT により抽出された関連コメント各 15 件からなる計 30 件のニュースコメントを提示し，被験者は各コメントに関してリプライ関係のある正解ニュースと各手法によって関連があると予測されたニュースそれぞれとの関連度を 5 段階で評価した．

表 9 より，提案手法と BERT により抽出したニュースコメントの正解ニュース，関連予測ニュースに対する関連度の平均を見る．正解ニュースに対する平均関連度では，提案手法と BERT の両方においてニュースコメントの関連度が高く出ており，2 郡の平均値に差がないとした場合の片側 t 検定の結果を見ても正解ニュースに対する関連度では有意差が存在しないことが分かる．一方，関連予測ニュースとニュースコメントの関連度について見ると，片側 t 検定の結果から提案手法の結果はと BERT の結果を優位に上回っていることが分かる．この結果から，BERT と比較して，提案手法による関連情報の抽出では，異なるニュース間で関連したコメントが抽出可能であると言える．また，被験者実験の結果から，抽出したコメントの内容から，ニューストピック間においてどのような視点から関連性があるのかをユーザに提示することが可能であると言える．一方で，抽出したコメントが，ユーザにとってニュース理解の参考になるかどうかについては課題が存在する．提案手法による関連コメント抽出の例を見ると,「岸田を擁護する気にもマスコミに肩入れする気にもならない。勝手にやっててくれ。まともに日本のこと考えてない者同士」というコメントのように，関連予測ニュースに対する関連度スコアが低いコメントには，

特定ニュースを強く想起させる具体的な単語が含まれていることが多い．こうしたコメントの場合は，コメント全体の主旨は関連予測ニュースに対しても関係がありそうな内容であっても，関連度スコアは低い結果となった．また，「他に優先すべきこと山盛りやろがい」というコメントのように，コメントが抽象的すぎる場合は，具体的なコメントと比較して関連予測ニュースに対する関連度スコアが高い傾向にあった．こうした抽象度の高すぎるコメントは，ユーザにとってニュース理解の参考になるとは言えないことから，より適切なユーザのニュース理解支援をおこなうには，具体的すぎるコメントの抽象化や，過度に抽象的なコメントを取り除くなどの改善の必要があると考えられる．

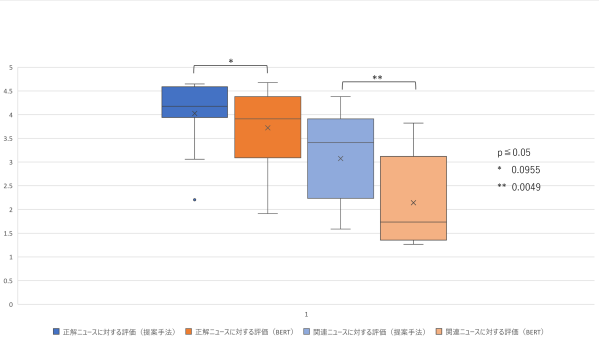


図 9　コメントの関連度評価（$p \leqq 0.05$*）

続いて，マルチクラス分類を行った BERT によるニュースコメント可視化の結果と提案手法による可視化の結果について図 10，図 11 に示す．それぞれ，提案手法と BERT によって生成されたニュースコメントのベクトルを t-SNE によって 2 次元に変換し，可視化したものである．提案手法では，各ニューストピックに対するコメントの分布がある程度まとまっており，複数のニュースにまたがる関連コメントはニューストピック間の境界付近に存在していることが分かった．一方，BERT によるベクトル生成と可視化では，ニューストピックごとにコメントの分布が独立しており，各ニューストピックごとにコメントは固まって存在していることが分かる．しかし，提案手法による可視化と異なり，複数のニュース間に存在する類似意見などのコメントは可視化図では確認できない．このことから，ニュース理解支援問目的において，提案手法は BERT による分類モデルと比較して適していると言える．

最後に，提案手法により生成したニュースコメントのベクトルに対して次元削減を行い，散布図に可視化した結果について図 12 に示す．右の可視化図で強調されているデータ点は，赤がニュース 7 に属するコメント，青はニュース 12 に属するコメントである．この 2 つのコメント群は関連コメント抽出の結果から異なるニュースに属するクラスタの中から類似度が高い結果となったクラスタのコメントである．この結果から，t-SNE と UMAP においては次元削減前のベクトルの類似性をある程度保ったまま次元削減できているものの，PCA では様々なニューストピックのコメントが混在しており，本研究の目的に適さない結果となった．また，関連コメント抽出において，関連性が高いと判断されたクラスタについて，可視化図上での位置を確認したところ，t-SNE と UMAP ではそれぞれ可視化図上でも近い位置に関連情報が存在していることが分かった．

5　おわりに

本論文では，SNS 上のニュースコメントについて，リプライ関係を基に対照学習を行うことで文章の類似性に寄らないニュースや SNS 投稿の反応を考慮した文書ベクトルを生成し，ニュース間の関連情報を抽出する手法を提案した．実験の結果から，対照学習を用いた，リプライ関係に基づいたニュースコメント間の関係性学習を行うことで，複数ニュースにまたがる関連コメントの抽出や可視化において，提案手法が BERT によるマルチクラス分類モデルと比較してユーザのニュース理解支援に適していることが分かった．提案手法では，ニュース間の関連性やコメント中に含まれる類似意見などをコメントのベクトル表現で適切に表すことができた．一方，BERT モデルによるコメントのベクトル生成では，ニュース間の関連コメントの抽出や可視化の結果から，ニュース間やコメント間の関係性を反映したベクトル表現の生成には適していないことが分かった．今後の課題として，本研究の結果では，関連すると予測されたニュースとの関連スコアが低いコメントも抽出する結果となったことから，関連コメント抽出について，よりユーザのニュース理解に有用なコメントを抽出するための機械学習モデルの性能向上が必要である．また，より多くのニュートピック，コメントを対象とした際の提案手法の有効性や，未知のニュースに対する文書ベクトルの生成と関連コメント抽出について調査する必要がある．

文　献

[1] 新聞通信調査会. 第 13 回メディアに関する全国世論調査. 2021. https://www.chosakai.gr.jp/oshirase/●第 13 回メディアに関する全国世論調査（2020 年）報告書.pdf.
[2] NTT ドコモモバイル社会研究所. 2021 年一般向けモバイル動向調査.
[3] 白松 俊神谷 晃. 市民協働のための web 記事上の社会問題の自動タグ付けと関連事例抽出手法. 人工知能学会全国大会論文集 JSAI2020, pp. 1C3OS6a03–1C3OS6a03, 2020.
[4] Lee Kenton Toutanova Kristina Devlin Jacob, Chang Ming-Wei. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.
[5] 大倉 俊平小野 真吾. 関連記事判定のためのニュース記事キーフレーズ抽出.
[6] Williams Ronald J and Zipser David. *A learning algorithm for continually running fully recurrent neural networks*, Vol. 1, pp. 270–280. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ⋯, 1989.
[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
[8] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
[10] Cross-encoders. https://www.sbert.net/examples/appli-

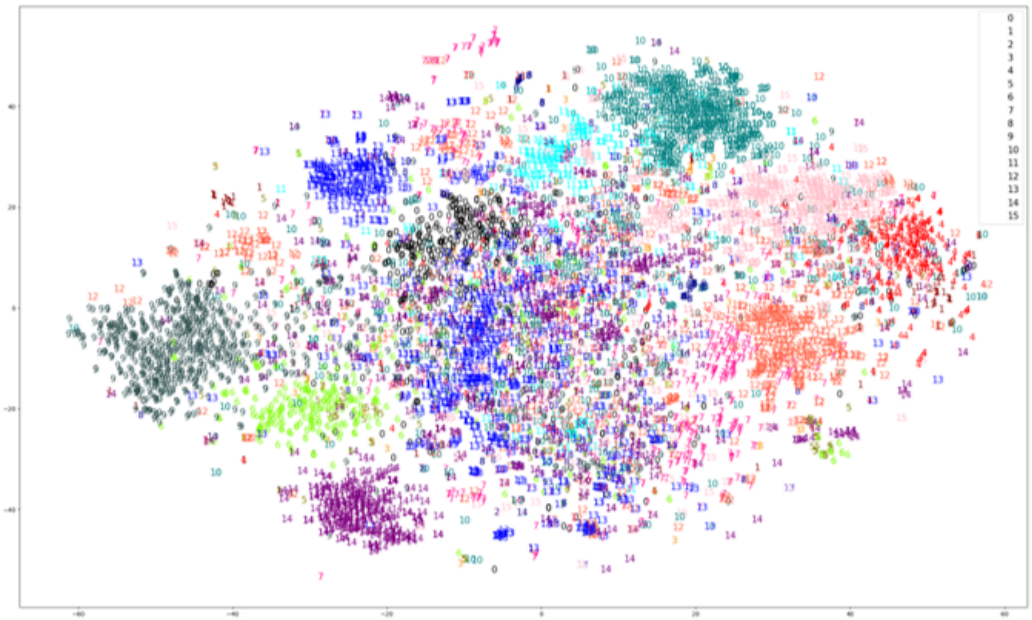


図 10　提案手法の可視化図

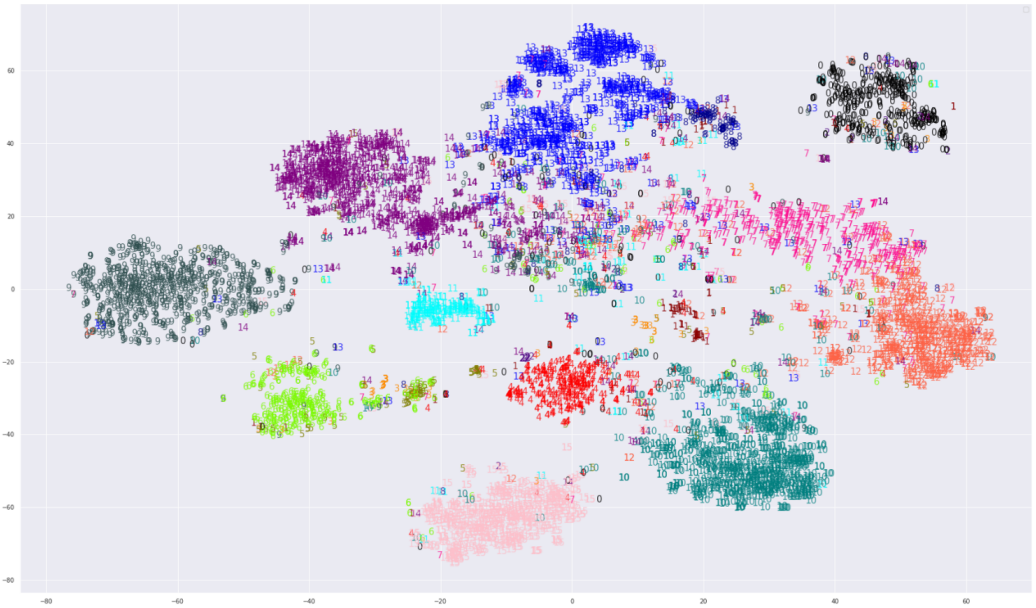


図 11　BERT による可視化

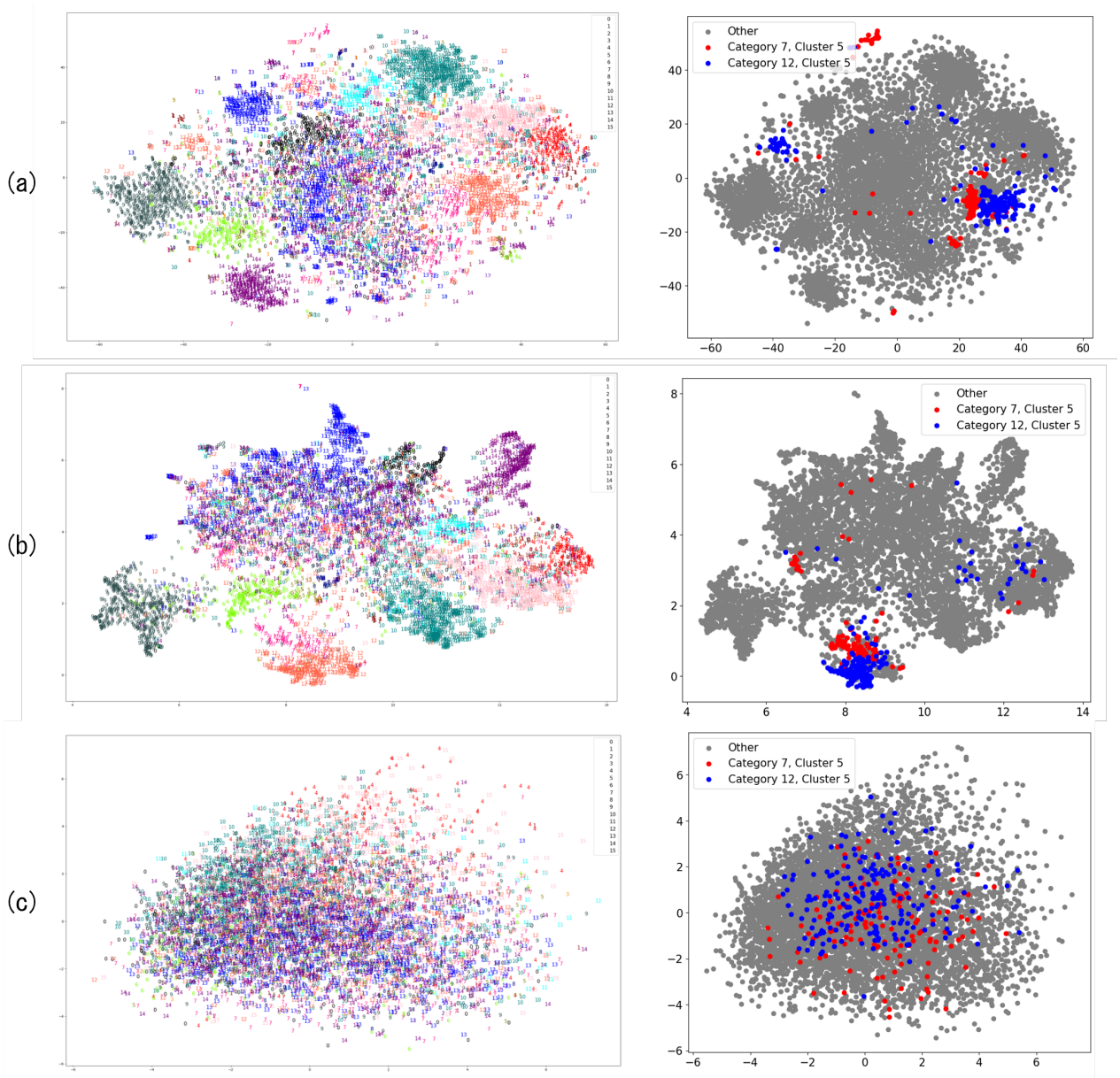


図 12　次元削減手法による可視化図の比較 (a)t-SNE，(b)UMAP，(c)PCA

cations/cross-encoder/README.html.

[11] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[12] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. In *Journal of Machine Learning Research 9(2605)*, pp. 2579–2605, 2008.

[13] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[14] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

[15] 東北大学乾研究室. Pretrained japanese bert models / 日本語 bert 訓練済みモデル. https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月29日(木) 16:00 ~ 18:10 | T3-B オンライン（Zoom Events)

Beyond Accuracy

座長:江口 浩二(広島大学)
コメンテータ:山名 早人(早稲田大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-B-6-01] 音楽推薦におけるセレンディピティ向上のための
アーティストエンコーダ

*佐藤 諒平[1]、SITKRONGWONG PADIPAT[1]、高須 淳宏[3]、相原 健朗[2] (1. 総合研究大学院大学　複合科学研究科 情報学専攻　高須淳宏研究室、2. 東京都立大学 大学院 都市環境科学研究科 観光科学域 教授国立情報学研究所 客員教授(兼任)社会データ構造化センター 客員教授、3. 国立情報学研究所 コンテンツ科学研究系 教授,総合研究大学院大学 複合科学研究科 情報学専攻)

16:30 ~ 16:55
[T3-B-6-02] empty-answer問題に対する多様性を考慮した効率的な探索手法

*池田 悠人[1]、肖 川[1]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科)

16:55 ~ 17:20
[T3-B-6-03] 相互近接性に基づくCollaborative Metric Learningの意外性向上手法

*中嶋 太一[1]、陳 漢雄[1]、古瀬 一隆[2] (1. 筑波大学、2. 白鴎大学)

17:20 ~ 17:45
[T3-B-6-04] δ-spannerを使用した位置匿名化によるモバイル端末からのイベント検索

*石神 京佳[1]、榎 美紀[2]、小口 正人[1] (1. お茶の水女子大学、2. 日本アイ・ビー・エム)

17:45 ~ 18:10
[T3-B-6-05] 好奇心に基づいたユーザ適応型セレンディピティ推薦手法の提案

*徐 哲林[1]、王 旭[1]、松村 敦[2] (1. 筑波大学人間総合科学学術院人間総合科学研究群情報学学位プログラム、2. 筑波大学図書館情報メディア系)

音楽推薦におけるセレンディピティ向上のための アーティストエンコーダ

佐藤　諒平† Padipat SITKRONGWONG†† 相原　健郎†††,†† 高須　淳宏††

† 総合研究大学院大学　〒 101-8430 東京都千代田区一ツ橋 2-1-2
†† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2
††† 東京都立大学　〒 192-0397 東京都八王子市南大沢 1-1
E-mail: †{ryo_1997,padipat,takasu}@nii.ac.jp, ††kenro.aihara@tmu.ac.jp

あらまし　推薦分野において，セレンディピティを考慮した推薦が精度のみを追求した推薦よりもユーザの満足度が上昇することから注目されている．音楽推薦においては人気度バイアスが特に顕著で，Spotify を調査した Rolling Stone 誌 [1] のデータによれば，上位 1 %のアーティストが再生数の 90 %を占めており，深刻なロングテール問題を引き起こしている．またユーザの過去の履歴に類似した推薦は，フィルターバブル現象を引き起こし，ユーザの満足度と予測精度の不一致を招く可能性がある．本研究では，セレンディピティを関連性，多様性，意外性の 3 つの要素を定義とし，マルチチャンネル型のアーティスト自己アテンションを提案する．関連性チャンネルではアーティスト間の類似性を，多様性チャンネルではプレイリスト内の未出現アーティストを考慮し，意外性チャンネルではアーティストに関する記述を利用している．これらチャンネルを統合しセレンディピティの向上とロングテール問題の解決を試みる．

キーワード　推薦システム，セレンディピティ，ロングテール, 音楽推薦

1 はじめに

推薦システムは，膨大な情報をフィルタリングしユーザの効果的な意思決定を支援するためのツールとして広く利用されている．協調フィルタリングやコンテンツベースフィルタリングを利用した従来の推薦手法は主に，推薦対象の予測精度を最大化することに比重が置かれている [1]．そしてユーザの過去の履歴に類似し過ぎた推薦は，ユーザにとって目新しさや意外性が失われやすくなり，ユーザを「フィルターバブル」と呼ばれる現象に陥らせる可能性があり，結果としてユーザの満足度と予測精度に乖離が生じることがある．[2]．音楽推薦システムにおいては，フィルターバブル問題に並んで，ロングテール効果が大きな問題となっている．ロングテールとは，「売れ筋商品」と「それ以外の商品」を軸に並べたとき，その売上の少ないほうの商品群（テール）が，低く長く図示される状況を指す [3]．音楽推薦においては，再生された楽曲の回数をアーティスト毎にまとめ，アーティスト単位の再生頻度グラフを作成した際に発生する人気のないアーティスト群を「ロングテールアーティスト」と呼ぶ．本研究では，ロングテールのアーティスト推薦においてセレンディピティを適用することの利点を研究することを目的としている．既存の研究とは異なり，本研究はセレンディピティの 3 つの重要な側面：関連性，多様性，意外性を定量的に定義し，同時にモデル化する．具体的には，音楽アーティストに新しいアーティストを推薦するためのアテンションベースのモデルである Multi-Channel Self-Attention for Serendipitous Artist Recommendations (MCAttnSAR) を提案する．MCAttn-SAR は，前述の 3 つのセレンディピティの側面を考慮するために，3 つの対応するチャンネルから構成され，それぞれ異なるタイプの入力とスコアリング関数を持つ．すべてのチャンネルにおいて，まずはプレイリストをアーティストリストに変換したのち，各アーティストリストから順序を除いたアーティスト集合 P へと変換される．関連性チャンネル は，アーティスト ID（一意識別子）の埋め込みを入力とし，アーティスト集合と関連性の高いアーティストに高得点を与える，精度に特化した標準的なモデルである．一方，多様性チャンネル は，アーティストのジャンル埋め込みを入力とし，アーティスト集合内のアーティストが持つジャンル集合を考えた時に，ジャンル集合に存在しないジャンルを持つアーティストに高いスコアを生成することを目的とする．最後に，意外性チャンネル はアーティストの説明の埋め込みを入力とし，説明はアーティスト集合と一致するが，関連性が低いアーティストに高いスコアを割り当てることを目的とする．本研究は，Transformer [4] で用いられるマルチヘッド自己アテンションを採用し，アーティストの埋め込みからアーティスト集合表現を生成する．そして，すべてのチャンネルからのスコアを集約することにより，適切なアーティストを推薦する．さらに，既存手法との比較やアブレーション分析を行うことにより，MCAttn-SAR が精度の点でより効果的であるだけでなく，関連性の高いロングテールのアーティストを推薦できることを示す．

本研究による貢献は，以下の 3 つの点となる．

（1） セレンディピティの概念をマルチヘッド自己アテンションに導入することで，新しいセレンディピティ推薦手法を提案する．本手法は，セレンディピティの複数のチャンネルを考慮し，異なるタイプの入力データを設計することで，正確な推薦を実現するだけでなく，音楽推薦において困難な課題であるロングテールアーティストのプロモーションを可能にする．

（2） セレンディピティの 3 大特性である「関連性」「多様性」「意外性」を同時にモデル化し，エンドツーエンドの推薦モデルを構築したのは，知る限り初めてである．

（3） ジャンルの関係性と LLM が生成した説明文という 2 種類の入力に基づき、アーティストの表現を学習する効果的な方法を設計し、セレンディピティ・推薦とシームレスに統合する．

2 関 連 研 究

推薦システムのアプローチとして，大きく分けてコンテンツベースと協調フィルタリングの 2 つがある．協調フィルタリングは多数のユーザから嗜好や嗜好情報を収集することにより，ユーザの興味・関心を自動的に予測するアプローチとなっており，協調フィルタリングアプローチの基本的な前提は，あるユーザ A が特定のドメインにおけるアイテムに関する評価が他のユーザ B と同じ評価をしている場合，A は別のアイテムについて，無作為に選ばれた他のユーザの評価よりもユーザ B の評価を持っているという考え方である．音楽推薦においては，主にユーザの持つプレイリストに着目し，特定のプレイリストに対して追加しうる楽曲を推薦するネクストソング推薦と呼ばれるタスクの際に使用される [5]．推薦システムでは，ユーザの満足度を向上させるためにはユーザにとって適しているアイテムを高い精度で提供することが重要と考えられてきた [6] [7]．しかし近年では，推薦システムとしての精度が高いことだけが，ユーザの満足度を決定する要因ではないことが分かってきており [8]，新規性や意外性といった推薦精度以外の評価指標をもって満足度を決定する研究が多く行われている [9]．推薦分野では "Serendipity" と呼ばれ，多くは関連性（Relevance），新規性（Novelty），多様性（Diversity），意外性（Unexpectedness）の 4 つを含むようなアイテムとされる [10] [11] [12]．ロングテールアクセスを可能にする研究はいくつか発表されており，セレンディピティを考慮したもの [13] や，JUST と呼ばれるトピックベースのモデル [14] などが提案されているが，セレンディピティを考慮した研究においてはセレンディピティの定義とされている関連性，新規性，多様性，意外性の 4 つの内，関連性と意外性を考慮するにとどまっている．一方でトピックベースモデルに関してはユーザから暗黙的に得られる履歴データの他に，ユーザの持つ他のユーザとの関連性を示すコンテキストを必要としており，ユーザの持つ積極性に依存しているという問題が存在する．本研究では，セレンディピティの 3 つの重要な側面を同時に推薦モデルに組み込み，音楽推薦におけるロングテール・アーティストのプロモーション効果を研究する．

3 MCAttnSAR

3.1 前　　提

本研究においては, [15] に従って，以下のようにセレンディピティの要素を定義し活用する．

- 関連性はアイテム間，若しくはアイテムとユーザ間において関連があること．
- 多様性はユーザの作成したある特定のアーティスト集合に対して，出現していないアイテムであること．
- 意外性とは既存の履歴依存の推薦手法やユーザ自身の検索能力では辿り着くことのできないアイテムであること．

表 1　各チャンネルの概要

	関連性	多様性	意外性
入力	ID	ジャンル	記述
エンコーダ	ID ルックアップ	行列因数分解	SentenceBERT
アーティストリストエンコーダ	マルチヘッド自己アテンション		
出力（スコア）	関連性スコア	多様性スコア	意外性スコア

3.2 概　　要

MCAttnSAR のモデル・アーキテクチャを図 1 に示す．ユーザーが作成したプレイリストをアーティストの集合に変換した P(アーティスト集合) に含まれる M 組のアーティスト $\{a_1, \ldots, a_M\}$ と候補アーティスト a_c を入力とし，タスクは，a_c を P に推薦する尤度スコアを計算することである．MCAttnSAR は，セレンディピティの 3 つの主要な側面をモデル化するために 関連性チャンネル，多様性チャンネル，意外性チャンネルの 3 つのチャンネルを用いる各チャンネルは異なるタイプの入力とスコアを持つが，同じような目的を持つ 2 つのコンポーネント，アーティスト・エンコーダ と アーティスト集合・エンコーダ を共有している．まず，アーティスト・エンコーダ は，対応する入力データを *artist embedding* と呼ぶ各アーティストの密なベクトル表現にエンコードする．次に，アーティスト集合・エンコーダ は，各アーティスト集合内のアーティスト・埋め込み間の相互作用をモデル化するために，マルチヘッド自己アテンション [4] を採用する．そして，各アーティスト集合のアーティスト・埋め込み間の相互作用をモデル化し，それらを集約してアーティスト集合・埋め込みを作成する．その後，各チャンネルは，そのスコアリング関数に応じて，候補 a_c をターゲットアーティスト集合 P に推薦するスコアを出力する．最後に，すべてのチャンネルからのスコアが結合され，モデルは負サンプリングによる自己教師付きで学習される．

3.3 セレンディピティな推薦のための各チャンネルの目的

正確な推薦を提供することに主眼を置く従来の推薦システム（RS）とは異なりセレンディピティの 3 つの性質に従うように設計されている．各チャンネルの概要を表 1 に示す．

3.3.1 関連性チャンネル

単純に協調フィルタリング（CF）アプローチの考え方を踏襲している．アーティスト集合とアーティストの推薦に関しては，アーティストがアーティスト集合に何度も一緒に追加されている場合（つまり，アーティスト集合での共起率が高い場合），それらは何らかの共通の意味的関係を共有しており，将来また一緒に追加される可能性が高い．アーティスト ID エンコーダの目的は，入力アーティスト（P 内の全 M アーティスト，および候補アーティスト a_c）を，関連性チャンネル上の対応する埋め込みにマッピングすることであり，入力は各アーティストの

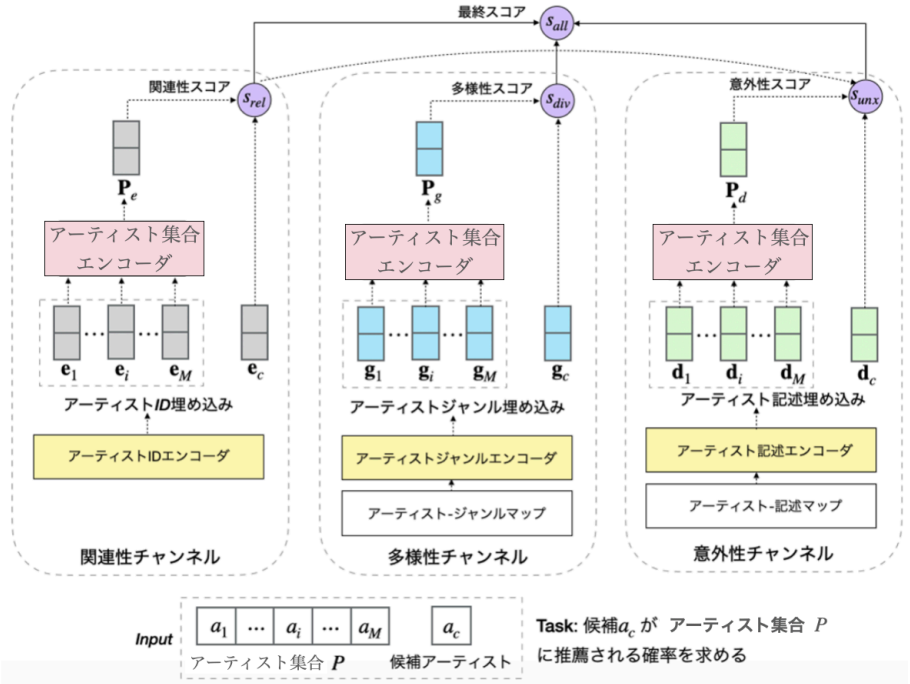


図 1　MCAttnSAR モデルの概要.

ID である．具体的には N をデータセット内の全アーティスト数とし，$\mathbf{E} \in \mathbb{R}^{N \times k_e}$ で表されるアーティスト ID の埋め込み行列を取得する．k_e は ID の埋め込みの次元を表し，アーティスト a_i に対しての出力は，$\mathbf{e}_i \in \mathbf{E}$ となる．

3.3.2　多様性チャンネル

関連性チャンネルとは対照的に，コンテンツまたはメタデータに基づいて候補アーティストのスコアを計算することによるコンテンツベースフィルタリング（CBF）のアプローチに従う．ジャンルが類似しているアーティストを優先的に推薦するのではなく，アーティスト集合に登録されているアーティストと比較して，より「多様性」のあるジャンルのアーティストを推薦することを目的とする．アーティストのジャンル埋め込みがアーティスト集合とは異なる場合，このチャンネルをより高いスコアが与えられる．関連性チャンネルと組み合わせ，ジャンルの異なる候補アーティストを追加することで，アーティスト集合に関連するアーティストを考慮しつつ，フィルタリングバブル問題の軽減を目的とする．アーティストジャンルエンコーダの目的は，入力アーティスト（アーティスト集合 P 内の全 M アーティスト，および候補アーティスト a_c）を，多様性チャンネル上の対応する埋め込みにマッピングすることであり，入力は各アーティストのジャンルである．具体的には N をデータセット内の全アーティスト数とし，$\mathbf{G} \in \mathbb{R}^{N \times k_g}$ で表される事前に学習されたアーティストジャンルの埋め込み行列を取得する．k_g は ID の埋め込みの次元を表し，アーティスト a_i に対しての出力は，$\mathbf{g}_i \in \mathbf{G}$ となる．

3.3.3　意外性チャンネル

関連性チャンネルと多様性チャンネルはそれぞれ CF アプローチと CBF アプローチを踏襲した．意外性チャンネルは，2 つのアプローチのトレードオフを考慮し，アーティスト集合に頻繁に追加されないロングテールのアーティストを推薦することを目的とする（人気度バイアス）．本モデルではアーティスト集合とアーティストの推薦において，関連性チャンネルのみに頼ると，頻繁にアーティスト集合に追加される人気アーティストが優先される可能性がある．すべてのチャンネルの目的を同時に達成することで，MCAttnSAR は，精度（＝関連性）だけにこだわるモデルよりも，ロングテールのアーティストの推薦を促進する．アーティスト記述エンコーダの目的は，入力アーティスト（アーティスト集合 P 内の全 M アーティスト，および候補アーティスト a_c）を，意外性チャンネル上の対応する埋め込みにマッピングすることであり，入力は各アーティストの記述である．具体的には N をデータセット内の全アーティスト数とし，$\mathbf{D} \in \mathbb{R}^{N \times k_d}$ で表される事前に学習されたアーティスト記述の埋め込み行列を取得する．k_d は ID の埋め込みの次元を表し，アーティスト a_i に対しての出力は，$\mathbf{d}_i \in \mathbf{D}$ となる．

3.4　アーティスト集合エンコーダ

アーティスト集合埋め込みを生成する．アーティスト集合埋め込みは，そのアーティスト集合内のアーティストの特性と関係を表す．

3.4.1　マルチヘッド自己アテンション層

アーティスト集合内のアーティスト間の意味的な関係を捉えるために，マルチヘッド自己アテンション層を使用してアーティスト埋め込み間の相互作用をモデル化する [4]．そして，これらの変更されたアーティスト埋め込みはアーティスト集約層を通過し，別のアテンションメカニズムを実行してアーティスト集合埋め込みを生成する．アーティスト集合 P からの関連するチャンネルの入力として，M 個のアーティスト ID 埋め込み $\mathbf{e}_1, \ldots, \mathbf{e}_M$ が与えられた場合，マルチヘッド自己アテンション

層は次のようにして，P 内の他のアーティスト埋め込みとの相互作用に基づいて各 $\mathbf{e}_i \in \mathbf{e}_1, \ldots, \mathbf{e}_M$ の *weighted artist embedding* を出力する：

$$\alpha_e^h(i,j) = \frac{\exp((\mathbf{Q}_e^h \mathbf{e}_i)^T \mathbf{K}_e^h \mathbf{e}_j)}{\sum_{m=1}^{M} \exp((\mathbf{Q}_e^h \mathbf{e}_m)^T \mathbf{K}_e^h \mathbf{e}_m)} \tag{1}$$

$$\mathbf{e}_i^h = \mathbf{V}_e^h (\sum_{j=1}^{M} \alpha_e^h(i,j)\, \mathbf{e}_j) \tag{2}$$

$\alpha_e^h(i,j)$: h 番目ヘッド内のアーティスト a_i と a_j のアテンションスコア

最終的な a_i の重み付き埋め込み $\mathbf{e}_i'$ は，

$\mathbf{e}_i' = [\mathbf{e}_i^{(h=1)}, \ldots, \mathbf{e}_i^{(h=H)}]$ (H はヘッド数)

ここで，$\mathbf{e}_i^h$ は a_i の h 番目の自己アテンションヘッドでの重み付きアーティスト ID 埋め込みであり，$\alpha_e^h(i,j)$ は h 番目のヘッド内のアーティスト a_i と a_j のペア間のアテンションスコアである．また，$\{\mathbf{Q}_e^h, \mathbf{K}_e^h, \mathbf{V}_e^h\}$ は h 番目のヘッド内のクエリ，キー，値行列のセットとなっている．最終的に，関連性チャンネル内での a_i の出力重み付きアーティスト ID 埋め込み $\mathbf{e}_i'$ は，すべてのヘッドからの $\mathbf{e}_i^h$ を連結して得られ，$\mathbf{e}_i' = [\mathbf{e}_i^{h=1}, \ldots, \mathbf{e}_i^{h=H}]$ であり，ここで H は自己アテンションヘッドの数に対応している．この重み付きアーティスト埋め込みは，アーティスト a_i の情報だけでなく，P 内の他の M アーティスト $\{a_1, \ldots, a_M\}$ との意味的な関係も表している．また，重み付きアーティストジャンル埋め込み $\mathbf{g}_i'$ と説明埋め込み $\mathbf{d}_i'$ も同様の方法で計算され，それぞれのチャンネルごとに異なるセットのマルチヘッド自己アテンション（すなわち，$\mathbf{Q}$，$\mathbf{K}$，および $\mathbf{V}$ のパラメータ）を使用している．

3.4.2 アーティスト集約層

与えられたアーティスト集合 P 内の M 個の修正されたアーティスト ID 埋め込み $\{\mathbf{e}_1', \ldots \mathbf{e}_M'\}$ を入力として，目的は集約関数 f_{agg} を適用し，すべての $\mathbf{e}_i' \in \{\mathbf{e}_1', \ldots \mathbf{e}_M'\}$ を集約し，P のアーティストの特徴を表すアーティスト集合埋め込み $\mathbf{P}_e$ を生成することである．f_{agg} の選択肢は平均関数など単純なものでは各アーティストの異なる情報レベルを考慮することができないため，[16] に従い，異なるアーティストからの異なる重みをモデル化するために *additive attention* 方法を適用している．形式的には，アーティスト a_i の他のアーティスト $\{a_1, \ldots, a_M\}$ に対する注目の重みは次のように計算される：

$$\gamma_i = \mathbf{h}^T \tanh(\mathbf{W}_1 \mathbf{e}_i' + \mathbf{b}_1) \tag{3}$$

$$\beta_i = \frac{\exp(\gamma_i)}{\sum_{i=1}^{M} \exp(\gamma_i)} \tag{4}$$

ここで，$\mathbf{W}_1$，$\mathbf{b}_1$，および $\mathbf{h}$ はモデルのパラメータである．アーティスト集合 P のアーティスト集合埋め込みは，次のようにしてすべての $\{\mathbf{e}_1', \ldots \mathbf{e}_M'\}$ からの加重平均として計算される：

$$\mathbf{P}_e = \sum_{i=1}^{M} \beta_i \mathbf{e}_i' \tag{5}$$

なお，ジャンル埋め込みと説明アーティスト集合埋め込みは，$\mathbf{P}_g$ と $\mathbf{P}_d$ も同様に計算される．

3.5 スコア推定

アーティスト集合 P に追加される各アーティスト候補 a_c の尤度スコアを計算するために，まず a_c を $\mathbf{e}_c, \mathbf{g}_c, \mathbf{d}_c$ と対応付ける．$\mathbf{d}_c$ はそれぞれ ID, ジャンル, 記述埋め込みを表す．次に，各チャンネルの候補スコアを以下のように計算する：

- **関連性スコア**: このスコアは，候補のアーティスト a_c が与えられたアーティスト集合 P にどれだけ「関連している」かを測定する．関連スコアを計算するためには，候補のアーティスト ID 埋め込み $\mathbf{e}_c$ とアーティスト集合 ID 埋め込み $\mathbf{P}_e$ との間でドット積演算を適用する．

$$s_{rel} = \mathbf{P}_e^T \mathbf{e}_c \tag{6}$$

- **多様性スコア**: このスコアは，候補のアーティスト a_c が与えられたアーティスト集合 P とジャンルの観点でどれだけ「多様」かを測定する．多様性スコアを計算するためには，候補のアーティストのジャンル埋め込み $\mathbf{g}_c$ とアーティスト集合のジャンル埋め込み $\mathbf{P}_g$ とのユークリッド距離を適用する．

$$s_{rel} = dist(\mathbf{P}_e^T, \mathbf{e}_c) \tag{7}$$

ここで，dist(・) はユークリッド距離.

- **意外性スコア**: このスコアは，候補のアーティスト a_c が与えられたアーティスト集合 P に対してどれだけ「意外であるか」かを測定する．このスコアを計算するためには，2 つの要因のトレードオフを考慮する必要がある：1) a_c がコンテンツ（すなわち説明）の観点で P とどれだけマッチするか（つまり，候補のアーティストの説明埋め込み $\mathbf{d}_c$ とアーティスト集合の説明埋め込み $\mathbf{P}_d$ とのドット積操作），および 2) a_c がアーティスト集合 P にとってどれだけ一般的でないかであり，これは逆関連スコアと見なすことができる：

$$s_{unx} = \frac{\mathbf{P}_d^T \mathbf{d}_c}{s_{rel}} \tag{8}$$

そして，すべてのチャンネルからのスコアを組み合わせて，プレイリスト P にアーティスト a_c を推薦するための総合スコアを計算する：

$$s_{all} = \lambda_{rel} s_{rel} + \lambda_{div} s_{div} + \lambda_{unx} s_{unx} \tag{9}$$

ここで，λ_{rel}，λ_{div}，および λ_{unx} は，各チャンネルから最終スコアへの寄与を制御するハイパーパラメータである．最後に，アーティスト a_c をプレイリスト P に推薦する確率は $p_{a_c} = \sigma(s_{all})$ として計算される．また σ はスコアを範囲 $[0,1]$ に変換するためのシグモイド関数である．

表2　人気アーティストとロングテールアーティストにおけるジャンルデータの例.

Artist Name	Longtail	Artist Genre
Drake	No	canadian hip hop, canadian pop, hip hop, pop rap, rap
Rihanna	No	barbadian pop, pop, urban contemporary
Justin Bieber	No	canadian pop, pop
Ed Sheeran	No	pop, singer-songwriter pop, uk pop
FINNEAS	Yes	alt z, la indie, pop, pov: indie
Vita Chambers	Yes	barbadian pop, canadian pop
Sam Bailey	Yes	talent show, uk pop

3.6　モデル学習

ネガティブサンプリング技術を採用している．各アーティスト集合 $P = \{a_1, \ldots, a_{|P|}\}$ について，アーティスト集合 P 内のアーティストのセット $A_P \subseteq P$ をサンプリングし，$y_{a_c} = 1$ としてその正例を割り当てる．そして，ラベル $y_{a_c} = 0$ で P に含まれないアーティストのセット $A_{(P')^c} \cap P = \emptyset$ をランダムにサンプリングする．その後，二項交差エントロピー（BCE）損失を用いてモデルをトレーニングする式を以下に示す：

$$\mathcal{L} = -\frac{1}{N_C}\sum_{c=1}^{N_C}(y_{a_c}\log(p_{a_c}) + (1 - y_{a_c})\log(1 - p_{a_c})) \qquad (10)$$

ここで，N_C はすべてのアーティスト集合からのすべての正例および負例の候補アーティストの数を指す.

3.6.1　アーティスト-ジャンル埋め込み

データセットには ACM Recsys ２０１８にて開催された際に Spotify から提供されたデータセットである Spotify Million Playlist Dataset2018（SMPD ２０１８）を利用した [17]．また，ジャンル情報については *GlennMcDonald* らによって開発された Web サイトで音楽サブスクリプションサービスの 1 つである Spotify の持つ５０００を超えるジャンル情報を取得できる Every Noise at Once(ENAO)[2]を使用した．本研究では，ENAO の持つジャンルのリスト情報を取得し，各ジャンルについて類似度の高い上位 10 %である 600 件のジャンル情報を用いてすべてのジャンル間の関係性をマッピングした.

3.6.2　アーティストとジャンルの関係性

ENAO によって取得したジャンル情報を活用するために，Spotify API[3]を用いて各アーティストのジャンル情報を取得した．Spotipy API から得られたジャンル情報の例を表 2 に示す．この時人気アーティストとロングテールアーティストの間には"pop"，"canadian pop"，"barbadian pop"などの共通のジャンルがある．これにより，アーティストの持つ人気によってジャンル情報が左右されていないということが分かる.

取得したジャンル情報はリストで管理されており，各ジャンル間の距離については分からず，順序だけが分かる状態であった．そこで，1 − 600 位まで取得しているジャンルランクを，１０段階のスコアに変換した．変換は G 分布に従い，以下の図 2 ようにランクをスコアに変換した.

各ジャンルのペア（g_i，g_j）ごとのスコアを計算した後，行列

2：https://www.everynoise.com/

3：https://developer.spotify.com/documentation/web-api

Genre 1	Genre 2	Rank	Score
pop	dance pop	1	10
pop	post-teen pop	2	10
pop	boy band	3	10
pop	talent show	14	9
pop	japanese r&b	31	9
pop	brazilian rock	415	4
pop	rap siciliano	421	4
pop	new wave	595	1
pop	emo rap	599	1

図 2　ランクスコア変換プロセス

因子分解（MF）技術（latent factor model）[18] を適用して，各ジャンルの低ランク表現を学習する．具体的には，$\mathbf{p}_{g_i}$ と $\mathbf{q}_{g_j}$ をそれぞれジャンル g_i と g_j の潜在表現としたとき，g_i と g_j の関係スコアは以下のように計算される：

$$score_{g_i,g_j} = \mathbf{q}_{g_j}^T \mathbf{p}_{g_i} \qquad (11)$$

MF でジャンルの埋め込みを学習した後，各ジャンルは対応する $\mathbf{p}_{g_i}$ または $\mathbf{q}_{g_j}$ のいずれか一方または両方にマッピングされる可能性がある．この時，表 2 から分かるように，各アーティストの持つユニークなジャンルリストの数は必ずしも等しくならない．その後，上位セット列からのみジャンルの埋め込みを選択して，アーティストのジャンルの埋め込みの学習のための入力とする.

最終的に，1 人のアーティストが複数のジャンルを含む可能性があるため，1 つのアーティストのジャンル埋め込みを生成するために，関連するすべてのジャンルの埋め込みを集約または結合する．結合に当たっては，すべてのジャンル埋め込みを平均してアーティストのジャンル埋め込みを構築する．具体的には，アーティスト a_i の N_g のジャンルセットが与えられた場合，アーティストのジャンル埋め込み $\mathbf{g}_i$ は次のように計算される：$\mathbf{g}_i = \sum_{k=1}^{N_g} \mathbf{p}_k$.

3.6.3　アーティスト-記述埋め込み

個々のアーティストの情報を取得するために，大規模言語モデル（LLM）を使用して，各個別のアーティストの情報を生成するデータ拡張技術を利用した．記述を取得した後，各アーティストの記述を密な表現，つまりアーティストの記述埋め込みにテキストエンコーダとしての事前学習済み言語モデルを用いて変換する.

個々のアーティストの情報を取得する方法としては, Wikipedia などのウェブサイトからクロールする方法や，そのアーティストのホームページから取得する方法などがある．しかし，複数のソースからデータを取得することは，記述の質だけでなく，テキスト構造にも矛盾をもたらす．さらに，人気アーティストとロングテール・アーティストの間で情報量に偏りや大きな差がないことが，これらの記述から生成されるアーティスト埋め込みが比較可能であり，高い品質であることを保証するために重要となる.

Prompt	① ② ③ ④ ⑤ Provide a short factual description of the music artist {artist_name}.
Response	artist_name: <input_name>

図 3　プロンプトの応答例

表 3　人気アーティストとロングテールアーティストにおけるジャンルデータの例.

プレイリスト数	20,724
アーティストの総数	1,723,704
人気アーティスト	1,533,766 (89%)
ロングテールアーティスト	189,938 (11%)
プレイリストに含まれるアーティストの平均	97.83
プレイリストに含まれるロングテールアーティストの平均	9.17

3.6.4　プロンプト-エンジニアリング

記述の信頼性と有用性を確保するために，プロンプトを設計する際には何度かの試行錯誤が行われた．その中で，図 3 が最も効果的だと判断し使用するプロンプトを決定した．決定にあたっての理由としては，以下の点が挙げられる：

（1） 本実験では，"provide "という言葉を使うことで，"generate "のような他の言葉よりも，ChatGPT がより信頼できる回答を生成することがわかった．

（2） "short": LLM が明確で簡潔な応答を生成するため（あまり冗長にならないように）．

（3） "factual": LLM が架空の情報ではなく事実に基づいた情報を生成するようにする（"fake 情報を生成しないでください"という指示を追加するよりも簡略化されている）．

（4） "description": "description"は，ChatGPT の回答が明確で簡潔であることを保証すると共に，アーティストに関する冗長で重要でない詳細を含まないことを保証する．

（5） "music": LLM が音楽アーティストの説明のみを提供し，他の種類のアーティストではないように制限する．

また，取得したアーティスト記述のエンコーディングには，Sentence-BERT [19] を用いた．

4　データ準備

4.1　データクリーニング

SMPD2018 データセットには，100 万件のプレイリストに対し 287,741 種類のユニークなアーティストが存在する．まず，100 万件のプレイリスト内で，10 件未満のプレイリストでしか確認されなかったアーティストを削除し，ジャンル情報，及び記述情報の得られなかったアーティストも同様に削除した．その上で，人気アーティストとロングテールアーティストとの閾値として，本実験に使用するアーティストの中で，50 回以下のアーティストをロングテールアーティストとし，それ以上のアーティストをポピュラーアーティストとした．閾値を 50 とした点として，図 4 に示すように 90 %分位数であることが挙げられる．次に，アーティスト集合内に存在するロングテールアーティストの数が 5 未満のプレイリスト及びアーティスト集合長が 40 未満のアーティスト集合を削除した結果を表に示す．

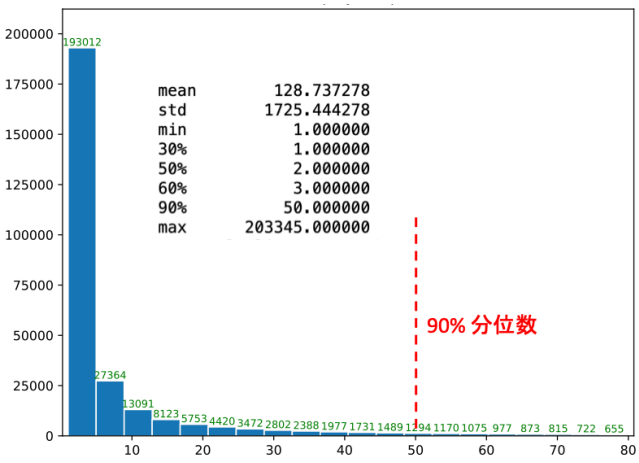


図 4　各アーティスト集合に出現したアーティストの種類

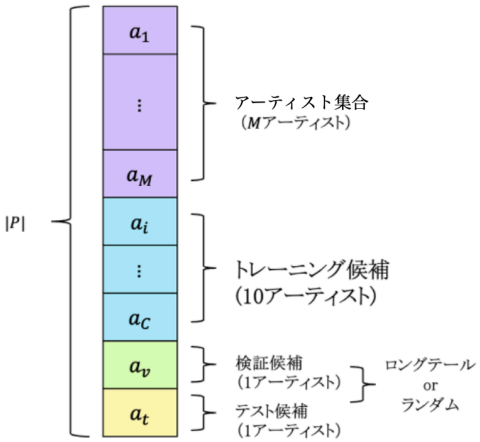


図 5　データ分割概要

4.2　データ分割

図 5 のように 4 つのデータセットに分割する．トレーニング，検証，テストで使用される候補データセットをそれぞれトレーニングは 10，検証は 1，テストは 1 でランダムに選定し，残ったデータセットを全てプレイリストとしてデータの分割を行うここで，検証用データセット及びテストデータセットでは実験の際に用いられる 2 つのロングテール，ランダムシナリオで決定される．

5　実験設定

本研究では，2 つのシナリオから実験が行われている．

- ランダム: 候補アーティストを制限しない
- ロングテール：候補アーティストが必ずロングテールアーティストとなるように制限する

実験の際に比較したベースラインは以下のモデルであり，実験の公平性のために実装は Recbole[4]を使用して実装した．

- GRU4Rec (2016)：畳み込みニューラルネットで用いられるアーキテクチャの 1 つである GRU を用いて順番にエンコードすることで，アイテム間の類似性や相補性を学習する [20].

4：https://recbole.io/

- STAMP (2018)： アテンション機構を持つ手法で，ユーザの興味を長期的な観点と短期的な観点から考慮する [21].
- NARM (2017)：ユーザのシーケンシャルな情報をモデル化する際に，最新のセッションにおけるユーザの目的についてを考慮するためにアテンション機構が用いられている [22].

また評価手法は，*nDCG*、*Hit Rate* (*HR*) の 2 つに加えて、カバー率及びロングテール率もモデルがフィルタリングバブルや人気度バイアスをどの程度緩和できるかを評価するために評価する．カバー率とはアイテムと比較した推奨アイテムの比率を測定する．

$$Coverage\ Rate = \frac{N_{recommend}}{N_{item}}$$

ここで、$N_{recommend}$ は全てのユニークな推薦アーティストの数であり、N_{item} は全てのプレイリストの全てのユニークなアーティストの数である．カバー率の高さは、フィルタリングバブルを回避し幅広いアーティストの推薦を可能にすることを示す．ロングテール率とは全推薦アイテムに対するロングテールアイテムの比率を示す．

$$Long\ Tail\ Rate = \frac{N_{longtail}}{N_{interaction}}$$

ここで、$N_{longtail}$ は推薦リストに含まれる全てのロングテールアーティストの数であり、$N_{interaction}$ は評価データに含まれる全てのアーティストの数を示す．ロングテール率の高さは、ロングテールアクセスを可能にし人気度バイアスの緩和がなされていることを示す．

6 結　　果

ランダムシナリオにおける実験の結果を表 4 に示す．ランダムなシナリオでは，HR の値や一部 nDCG の値で既存の手法よりも低くなっているものの，他の指標で全て上回る結果となった．HR はランキング内の正解ラベルの順位は考慮されないが，NDCG はランキング内の正解ラベルの順位が考慮されるものとなっている．これより，本モデルは既存の手法よりも正解ラベルをより上位に促進する効果があると考えられる．また，本研究では HR，nDCG は関連性，カバー率は多様性，ロングテール率は意外性との相関と見ることもでき，その観点からは一定の関連性を確保しつつ，多様性や意外性をユーザの推薦に追加する，という目的は達成されたとも言える．次に，ロングテールシナリオにおける結果を表 5 に示す．表からも分かる通り，ロングテールを推薦する場合においては HR,nDCG 共に既存の手法を大幅に上回る結果となった．これは，既存の手法には存在しない多様性，及び意外性を定義しジャンルと記述を用いてスコアリングを行ったことによって推薦性能が向上したと考えられる．また，既存の手法ではやはりロングテールアーティストは人気バイアスなどによって正しい評価を受けにくいということも本結果からより鮮明になったと言える．本モデルのカバー率が既存の手法と比較して低くなっているが，アーティスト全体の 90 %がロングテールに属しており，全体の 50 %をカバーしながらも高い評価を獲得している上，ロングテール率から判断するに，むしろ既存の手法よりも高い精度でロングテールアイテムの推薦を行うことができていると言える．一方で，元々ロングテールへのアクセスを目的としたベースラインとの比較ではないので，ベースライン自体をロングテールアクセス可能なモデルに変更することが今後の課題だと言える．次に，アブレーション研究の結果を表に示す．ここで，MCAttnSAR の後ろでそれぞれ’Rel’は関連性チャンネルを,’Div’は多様性チャンネルを,’Unx’は意外性チャンネルを示す．具体的には,’MCAttnSAR-RelDiv’は関連性，多様性の 2 つのチャンネルを用いた結果であり，意外性チャンネルは含まれていない．表の一番下にある’MCAttnSAR’はすべてのチャンネルを使用しているものである．表からも分かるように，関連性チャンネルをベースとして他のチャンネルを追加することでカバー率やロングテール率が上昇しており，本研究の目的であるロングテールへのアクセスが可能になっていることが分かる．

7 ま と め

本研究によって，既存の手法では使われてこなかったアーティストの記述埋め込みを用いることで効果的にロングテールアクセスを行うことができることがわかった．また，記述埋め込み及びジャンル間の関係とアーティストの記述埋め込みを用いることで従来の手法よりも高い評価を獲得することができたということは，ユーザに満足度の高い推薦を行うためには従来の協調フィルタリングなどで引き起こされる‘人気度バイアス‘や,’フィルターバブル’,‘ロングテール’問題の解決がユーザにとってより良い推薦である，ということがデータセットからも証明されたまた，今後の研究課題を以下に示す．

- より新しいベースラインとの比較を行う．
- 他の種類の損失関数を使用する．
- センシティビティ分析を行う．
- 異なる集約関数を適用する．
- ジャンルクラスタリングの再考．

特に，ベースラインについては新しさだけでなく，本研究と同じようにロングテールアクセスを考慮したベースラインとの比較を行う．

文　　献

[1] Paul Resnick and Hal R. Varian. Recommender systems. *Commun. ACM*, Vol. 40, No. 3, p. 56 – 58, mar 1997.
[2] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.
[3] Chris Anderson, Christopher Nissley, and Chris Anderson. The long tail, 2006.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000 – 6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
[5] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 13–22, 2012.
[6] Dietmar Jannach, Bamshad Mobasher, and Shlomo Berkovsky. Research directions in session-based and sequential recommendation: A preface to the special issue. *User Modeling and User-Adapted*

表 4　ランダムシナリオの結果

モデル	HR@10	HR@20	HR@50	NDCG@10	NDCG@20	NDCG@50	カバー率@10	ロングテール率@10
GRU4Rec	0.8216	0.8910	0.9504	0.6240	0.6417	0.6536	0.7209	0.2392
NARM	**0.8481**	**0.9022**	0.9552	0.6579	**0.6717**	0.6823	0.7002	0.2229
STAMP	0.6972	0.8022	0.9181	0.4866	0.5132	0.5365	0.6637	0.1879
MCAttnSAR	0.6288	0.6288	**0.9854**	**0.6677**	0.6677	**0.6974**	**0.7800**	**0.5540**

表 5　ロングテールシナリオの結果

モデル	HR@10	HR@20	HR@50	NDCG@10	NDCG@20	NDCG@50	カバー率@10	ロングテール率@10
GRU4Rec	0.3647	0.4680	0.6537	0.2249	0.2509	0.2874	**0.7060**	0.2470
NARM	0.3974	0.4960	0.6846	0.2504	0.2752	0.3121	0.6995	0.2487
STAMP	0.2685	0.3719	0.5610	0.1574	0.1834	0.2207	0.6565	0.2081
MCAttnSAR	**0.4940**	**0.4940**	**0.7612**	**0.4940**	**0.4940**	**0.5421**	0.5027	**0.5332**

表 6　MCAttnSAR のアブレーション結果

モデル	HR@10	HR@20	HR@50	nDCG@10	nDCG@20	nDCG@50	カバー率@10	ロングテール率@10
MCAttnSAR-Rel	0.4624	0.5967	0.7963	0.2766	0.3105	0.3502	0.2273	0.2032
MCAttnSAR-RelDiv	0.3834	0.5173	0.7881	0.2184	0.2501	0.3064	0.4018	0.3317
MCAttnSAR-RelUnx	0.4931	0.5317	0.7631	0.3485	0.5317	0.5733	0.5004	0.4797
MCAttnSAR	**0.6113**	**0.6697**	**0.8759**	**0.6207**	**0.6207**	**0.6224**	**0.7207**	**0.6301**

Interaction, Vol. 30, pp. 609–616, 9 2020.

[7] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. Meta-graph based recommendation fusion over heterogeneous information networks. Vol. Part F1296, pp. 635–644. ACM, 8 2017.

[8] Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, p. 1097 – 1101, New York, NY, USA, 2006. Association for Computing Machinery.

[9] Zhenyu Zhang, Lei Zhang, Dingqi Yang, and Liu Yang. Kran: Knowledge refining attention network for recommendation. *ACM Transactions on Knowledge Discovery from Data*, Vol. 16, pp. 1–20, 4 2022.

[10] Joseph Corneli, Anna Jordanous, Christian Guckelsberger, Alison Pease, and Simon Colton. Modelling serendipity in a computational context. *arXiv preprint arXiv:1411.0440*, 2014.

[11] Leo Iaquinta, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. Introducing serendipity in a content-based recommender system. In *2008 eighth international conference on hybrid intelligent systems*, pp. 168–173. IEEE, 2008.

[12] Paul André, m.c. schraefel, Jaime Teevan, and Susan T. Dumais. Discovery is never by chance: Designing for (un)serendipity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, C&C '09, p. 305 – 314, New York, NY, USA, 2009. Association for Computing Machinery.

[13] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, p. 13 – 22, New York, NY, USA, 2012. Association for Computing Machinery.

[14] Chunhua Sun and Yinjie Xu. Topic model-based recommender system for longtailed products against popularity bias. In *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pp. 250–256, 2019.

[15] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, Vol. 111, , 08 2016.

[16] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6389–6394, Hong Kong, China, November 2019. Association for Computational Linguistics.

[17] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. Recsys challenge 2018: Automatic music playlist continuation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, p. 527 – 528, New York, NY, USA, 2018. Association for Computing Machinery.

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, Vol. 42, No. 8, p. 30 – 37, aug 2009.

[19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[20] Yong Kiam Tan, Xinxing Xu, and Yong Liu. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, p. 17 – 22, New York, NY, USA, 2016. Association for Computing Machinery.

[21] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. Stamp: Short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, p. 1831 – 1839, New York, NY, USA, 2018. Association for Computing Machinery.

[22] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conf. on Information and Knowledge Management*, CIKM '17, p. 1419 – 1428, New York, NY, USA, 2017. ACM.

DEIM2024 X00-0

empty-answer 問題に対する多様性を考慮した効率的な探索手法

池田　悠人† 　肖　　川† 　鬼塚　　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5
E-mail: †{ikeda.yuto,chuanx,onizuka}@ist.osaka-u.ac.jp

あらまし　empty-answer 問題とは，ユーザがクエリを与えた際に，その条件の多さや厳しさのために条件を満たすレコードが得られないことである．この問題の解決に向け，既存手法においては，主にランキング関数を利用する方法と，レコードを得られるような緩和したクエリを探索する手法の 2 つが提案されている．しかしこれらの手法においては，ユーザが得られるレコード群の多様性を担保した手法が存在しない．そこで本研究では，ランキング関数に多様性を考慮した項を組み込み，多様性とクエリからの類似性の双方が高いレコード群を探索する高速化手法を提案する．本研究では，多様性とクエリからの類似性の双方が高い緩和クエリ群を探索することで，クエリから得られるレコードの類似性と多様性を確保するアルゴリズムを提案する．本研究における実験により，精度を維持しながら 100 倍以上の高速化の達成を確認した．

キーワード　データベース，クエリ緩和，多様性

property	fee	distance	area	garage
1	50000	20	35	No
2	40000	30	40	No
3	60000	18	30	No
4	100000	5	25	Yes

表 1: 賃貸物件のデータ例

1 はじめに

様々なアプリケーションにおいて，ユーザが望む条件を設定しデータの検索を行うことは基本的な操作である．ユーザが検索を行う場合，求める条件に合致したレコードが一定数存在するケースが理想的である．しかし，ユーザの望む条件によって，得られるレコード数が多すぎる場合や，逆に全くレコードが得られない場合が容易に想定される．前者を many-answer 問題と呼び，後者を empty-answer 問題 [1] と呼ぶ．

これらのうち，本研究では，潜在的にデータベース内の全レコードが探索範囲となることから，より解決が困難であると考えられる empty-answer 問題の解消を目的とする．empty-answer となる要因としては，条件を多くすることにより，個々の条件を満たすレコードは存在しているものの，それら全ての条件を満たすレコードが得られないケースと，事前にデータに設定されている制約を満たさないレコードの検索など条件が不正であるケースの 2 つが考えられる．本研究では前者のケースを想定する．

ユースケース

表 1 のようなテーブルを想定し，ユーザが

$$fee < 40000 \wedge area > 40 \wedge disdance < 20 \wedge garage = \text{"}No\text{"} \quad (1)$$

のようにクエリ条件を設定し，物件検索を行った場合を仮定する．この場合，ユーザが設定したクエリの条件を全て満たす物件は存在しないため，ユーザは空のクエリ実行結果を得る．

上に示したような条件設定における empty-answer の解決手法としては，既存手法では大きく 2 つのアプローチに分類される．まず 1 つはクエリ緩和手法である [2]．具体的には，ユーザが設定した連言の条件が多数にわたる場合に，その条件のうち一部を取り除くことにより，empty-answer 問題の解消を目指す手法である．この手法の利点として，ユーザが指定した条件に変更を加える手法であるため，ユーザとのインタラクションが容易であり，ユーザの意思をリアルタイムで反映しやすい点が挙げられる．このメリットを活用し，既存研究ではユーザとのインタラクションなしで直接クエリ緩和を行う手法に加え，ユーザとのインタラクションを行うことで漸進的にクエリ緩和を行う手法 [3, 4] も提案されている．例えばユースケースの例においては，式 1 のクエリに対して，*distance* の条件を取り除いた緩和クエリを生成することで，property が 1 である物件が新たなクエリの条件を全て満たすようになる．

もう 1 つのアプローチはランキング手法 [5] である．これはユーザの嗜好を特定の関数に落とし込むことで，全てのレコードに対してその関数の評価を実施し，スコアの高いレコードをユーザに提示する手法である．この手法は，ユーザ自身がランキング関数を設計する能力がある場合において，ユーザの嗜好を反映しやすい点が強みとして挙げられる．ユーザがスコア関数として満たしている条件数を返す関数を設定し，ユーザが 2 件のレコードを要求している場合，ユースケース例においては物件 1, 2 はクエリの条件数を 3 つ満たしている．これより多く条件を満たす物件は存在しないため，この場合ユーザには property が 1 と 2 である物件が推薦される．

ここでユーザが得るレコードセットにおいては，クエリからの近似性だけではなく推薦レコードセット内での多様性も重要視される．これは，推薦レコード間での多様性を確保することでユーザの好みを捉えたレコードを選択しやすいことがその理

由として挙げられる [6]．ユーザが検索をかける際に設定するクエリ条件には，クエリには現れない潜在的な優先度が存在することがある．上に挙げた賃貸物件を例にとると，ユーザは賃料に関して，設定したクエリ条件の遵守を求めるが，部屋の広さに関してはある程度の妥協の余地が存在するかもしれない．しかし，このようなユーザの希望はクエリにおいては表現されない．一方で推薦レコードセットを探索するときに多様性を考慮しない場合，データセットの分布によってはユーザが暗に重要と考える条件を満たさないレコードセットを推薦する可能性が高まる．これによりユーザの潜在的な好みと推薦レコードセットとのミスマッチの発生が考えられる．これを防ぐ方法として，推薦レコードセットにおいて多様性を考慮することが挙げられる．

既存研究では，データベース分野における empty-answer 問題の文脈において，問題の解消後にユーザが得られるレコードセットにおける多様性に着目した手法は提案されていない．一方で，データベースの分野外に目を向けると，ユーザの嗜好をクエリからの近似性とし，近似性と多様性の両立を図ったランキング関数である Maximal Marginal Relevance (MMR) が提案されている [7]．

そこで，本研究では，MMR を ランキング関数として活用することで，多様性を考慮した empty-answer 問題の解消を実現する．MMR をデータベース分野に応用可能な形で定式化を行うとともに，多様かつ条件を多く残した緩和クエリを探索することで，MMR を最大化するレコードセットの探索を高速に行う手法を提案する．また，実世界でのデータセットを利用した実験においては，MMR スコアの低下を抑えながら，最大で 50 倍以上の高速化を達成したことを確認した．

2 予 備 知 識

2.1 事 前 定 義

ユーザが与えるクエリを q とし，そのクエリに $|q|$ 個の条件が含まれているとする．この時，これらの各条件を $c_i (0 \leqq i \leqq |q|-1)$ とすると，クエリは

$$q = \wedge_{i=0}^{|q|-1} c_i$$

と表現される．つまり，クエリはその条件の連言として表現される．

次に，クエリ q の条件の集合を C_q とし，その冪集合を $Pow(C_q)$ と定義する．q の緩和クエリ q' は，クエリ q の条件の真部分集合 $C \in Pow(C_q) \setminus C_q$ を用いて以下のように表現される．

$$q' = \wedge_{\{\forall c_i \in C\}} c_i \quad (2)$$

また，$Pow(C_q)$ の要素数は $2^{|}q|$ 個であることから，そこから C_q を除いた q' の候補数は $2^{|}q| - 1$ 個あることがわかる．この他，データベースが含むレコードの総数を n，レコードに含まれるカラムの数を M とし，ユーザが推薦レコードセットのサイズとして指定する数を k とおく．

Algorithm 1: greedy 手法の疑似コード

Input : dataset D, query Q
Output : recommended dataset RD

```
RD ← ∅;
foreach i ∈ [0, k − 1] do
    RD[i] ← arg max_{x∈(D\RD)}(similarity(RD ∪ {x}) +
      λdiversity(Q, RD ∪ {x});
return RD
```

2.2 既存の MMR 探索方法

MMR はレコードの集合に対して定義される list-wise な評価指標である．そのため，MMR を最大化する推薦レコードセットを探索するためには，全ての k 個のレコードの組み合わせに対して MMR を計算して比較する必要がある．これを実現するためには計算量 $O(|q|n^k)$ を必要とする．これは特に k を大きくとる場合に，探索の実行は現実的ではなくなる．そこで，貪欲法に推薦レコードセットを探索する方法が提案されている [8]．

アルゴリズム 1 に，貪欲法による MMR を最大化するレコードセットの探索手法の疑似コードを掲載する．この手法において，1 つめのレコード選択時にはクエリからの近似性が最も高いレコードが選択される．2 つ目以降のレコード選択時には，すでに選択されているレコード群に追加候補レコードを加えたセットに対する MMR を計算し，そのスコアが最も高くなる追加候補レコードを推薦レコードセットとして選択する．これらを推薦レコードが k 件選択されるまで反復を行う．

ここで，既存手法の素朴な実装では，計算量は $O(nMk^2)$ となる [8]．しかしこの手法では，MMR の計算の際に，同じレコードペアに対して複数回の距離比較を行なっている．そこで，この計算の重複を回避し計算量を削減した手法が提案されており [9] 推薦レコードセットにかかる計算量は $O(nMk)$ と見積もられている．

3 提 案 手 法

3.1 各 種 定 義

3.1.1 MMR の定義

MMR はユーザが与えたクエリからの推薦レコードセットに含まれる各レコードへの近似性と，推薦レコードセット間の多様性に対して一定の重みをつけ，その和をスコアとする指標である．推薦レコードセットを D' と定義したとき，MMR の定義は以下の通りである [7]．

$$MMR(q, D') = relevance(q, D') + \lambda diversity(D') \quad (3)$$

但し，式 3 における λ は近似性と多様性の間のバランスを調整するパラメータである．

この relevnce と diversity の具体的な定義に関して，推薦システム分野においてはアイテムベクトルとユーザベクトルとの最大内積探索問題の文脈で定義が行われている [9]．そこで，本研究で MMR を活用するにあたり，近似性と多様性についての

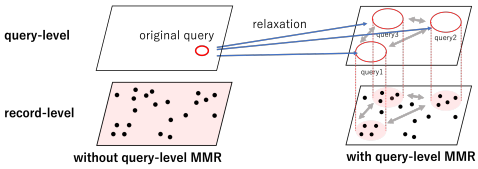


図 1: クエリレベル MMR のイメージ図

具体的な定義を行う必要がある．本研究においてはこれらを以下の形で定義する．

$$similarity = \frac{1}{k} \sum_{r \in D'} \frac{\sum_{j=0}^{|q|-1} \delta(c_j, r)}{|q|} \tag{4}$$

$$diversity = \begin{cases} 1(k = 1) \\ \min_{r,r' \in D', r \neq r'} (\frac{dist(r,r')}{l})(k \neq 1) \end{cases} \tag{5}$$

式 4 は，本研究における近似性を定義している．この式における $\delta(c_i, r)$ は，レコード r がクエリ条件 c_i を満たしている場合に 1 を返し，そうでない場合には 0 を返す．直感的には，式 4 は合致する条件の数を全ての条件の数で割り算した値をレコード毎で平均を取ったものである．また，式 5 では本研究における多様性について定義を行なった式である．多様性は種々の定義がなされている [10] が本研究においては， min-max アプローチと min-sum アプローチの 2 つの形で定義している．式 5 は min-max アプローチに基づく定義であり，推薦レコードセットの多様性を，全てのレコードペアの距離の最小値と定義している．この定義は [11] に従っている．式 5 に含まれる $dist$ 関数は，レコード間の距離を定義する関数である，$dist$ 関数には任意の関数が適用可能であり，直感的には，レコード間の違いを表す指標となっている．本研究においては，数値データに対してはマンハッタン距離を採用し，カテゴリデータやバイナリデータに対してはハミング距離を適用する．カテゴリデータと数値データとの間において距離の定義が異なる理由として，カテゴリデータにおける値の差異の大小がその違いの大きさを示すとは限らないためである．

3.1.2 問 題 定 義

データセット D に対し empty-answer を返すクエリ q が与えられた際に，$MMR(q, D')$ を最大化する，k 個からなる推薦レコードセット D' を探索することを，本研究の目的とする．

3.2 提案手法の着想・達観

MMR スコアを厳密に最大化する推薦レコードセットを特定するためには，式 5 における定義に従い，全てのレコードペア間の距離を計算する必要がある．これは計算コストが極めて大きく，実行は現実的ではない．そこで既存手法 [8, 9] においては，2.2 章で説明したように，近似解を見つけるために貪欲法によるヒューリスティックなアプローチを採用している．これには，最先端の手法 [9] も含まれ，この手法は近似解を見つけるために $O(nMk)$ の時間複雑性を要する．この手法を適用することで，多様性と近似性を確保したレコードセットを探索し empty-answer 問題を解決することが可能である．しかし，4 章において詳しく述べる通り，これら既存手法は依然応答時間の長さに課題を抱えている．

そこで本研究では，既存研究と比較してより高速に近似解を見つけるためにヒューリスティックなアプローチを提案する．その着想は，クエリをその実行結果のレコードセットの代表値として見ることである．具体的には既存手法である，レコードレベルの MMR の前処理ステップとしてクエリレベルの MMR の概念を本研究において導入する．図 1 はクエリレベル MMR の概要を示している．初めに，図の右上隅に示されているように，元のクエリから複数の緩和クエリを生成し，クエリレベルの MMR を最大化することを目指す．その後，図の右下隅に示されているように，これらの緩和クエリの結果を取得し，得られたレコード集合を候補として，レコードレベルの MMR を最大化する推薦レコードセットを選択する．

クエリレベルの MMR は，式 3 におけるレコードレベル MMR 推薦レコードセット (D') を緩和クエリセット (Q') に置き換えることで定義される．具体的な定義は以下のとおりである．:

$$MMR(q, Q') = relevance(q, Q') + \lambda\, diversity(Q'). \tag{6}$$

さらに，クエリの近似性と多様性に関する指標を導入する．クエリの近似性は，元のクエリと緩和クエリセットとの類似性を計算し，クエリの多様性は緩和クエリセット内の多様性を測定する．これらの指標は以下のように定義される

$$relevance(q, Q') = \frac{1}{k} \sum_{q' \in Q'} \frac{len(q')}{len(q)} \tag{7}$$

$$diversity(Q') = 1 - \min_{q'', q' \in Q', q' \neq q''} \frac{len(q' \cap q'')}{len(q' \cup q'')} \tag{8}$$

ここで，$len(q)$ はクエリ q に含まれる条件の数を表す．クエリレベルにおけるオリジナルクエリから緩和クエリの近似性は，緩和クエリが持つ条件数をオリジナルクエリの条件数で割ったものとして定義される．一方で緩和クエリセットの多様性は，全ての緩和クエリペアのジャカード距離を計算したものの最小値と定義される．

アルゴリズム 2 に提案手法の疑似コードを掲載する．提案手法は 2 つの段階に分かれている．1 つ目の段階では，クエリレベルの MMR を最大化する緩和クエリを探索し，その結果を得る．ここで得られる緩和クエリは図 1 における，右上のクエリにあり，緩和クエリの実行結果が，図の右下の赤色で着色された領域に含まれるレコードセットである．2 つ目の段階では，これらの結果からレコードレベルの MMR を最大化するレコードを推薦レコードセットに 1 つ追加する．提案手法では，これ

Algorithm 3: 提案手法の疑似コード

Input : dataset D, query Q
Output : recommended dataset RD

$RD \leftarrow \emptyset$;
$PairMatrix \leftarrow$ an $m \times m$ matrix, initialized by 0;
$CountArray \leftarrow [0] * k$, $SubQuery(SQ) \leftarrow [\emptyset] * k$;
foreach $i \in [0, k-1]$ **do**
  $SubQuery \leftarrow [\emptyset] \times k$;
  $SubQuery[0] \leftarrow$ the constraint with the lowest value in CountArray;
  foreach $j \in [1, m-1]$ **do**
    $SQ[j] \leftarrow \arg\min_{C_j \in (Q \setminus SQ)} \Sigma_{C_l \in SQ} PairMatrix[l][j]$;
  $DS \leftarrow D$;
  foreach $j \in [m-1, 0]$ **do**
    remove C_j from $SubQuery$;
    $DS_new \leftarrow ExecuteQuery(DS, SubQuery)$;
    if $(DS_new \setminus RD) = \emptyset$ **then break**;
    $DS \leftarrow DS_new$;
  $RD[i] \leftarrow \arg\max_{x \in (DS \setminus RD)} MMR(RD \cup \{x\})$;
  /* choose record */
  foreach $i \in len(SubQuery)$ **do**
    $Index_C_1 \leftarrow GetIndexOfConstraint(Q, C_1)$;
    $CountArray[Index_C_1] += 1$;
    foreach $C_2 \in SubQuery$ such that $C_1 \neq C_2$ **do**
      $Index_C_2 \leftarrow GetIndexOfConstraint(Q, C_1)$;
      $PairMatrix[Index_C_2][Index_C_2] += 1$;
return RD

らの 2 つの段階を k 回繰り返すことで，k 個のレコードからなる推薦レコードセットを得る．

以下，3.3 章において緩和クエリの探索手法について説明を行い，3.4 章においてレコードレベルの MMR を最大化する手法を説明する．また，3.5 章において一定の仮定のもとにおける提案手法の計算量について議論する．さらに，3.6 章では提案手法をもとに，さらなる改良を加えた手法についての説明を行う．

3.3 緩和クエリ探索

ここでは，ユーザが与えたクエリを緩和することで，緩和クエリを満たすレコードを推薦レコードセットの候補とする．この時候補のレコードは，ユーザが与えたクエリに含まれる条件をより多く満たしつつ，すでに推薦レコードセットに加えられたレコード群とは異なる特徴を持つことが求められる．そのようなレコードを候補に加えるため，緩和クエリが

（1） empty にならない範囲で条件を極力残す
（2） 緩和クエリ群での多様性を確保する

という特性を満たすように条件が選択されることを提案手法での戦略とする．(1) は近似性を高める狙いがあり，(2) は多様性を高める狙いがある．しかし，2.1 章で述べた通り，緩和クエリの候補は $2^{|}q| - 1$ 個存在する．このため，これらの候補緩和クエリ全てから実直に探索を行うことは困難である．そこで提案手法では，これまでの緩和クエリ群において利用された条件のペアを記録しておき，新たな緩和クエリ生成の際には greedy に条件を一つずつ決定する．これにより，緩和クエリ群で異なるクエリ条件が選択されることにより，その緩和クエリから得られるレコードにも多様性が確保されることが期待されるとともに，条件の選択を計算量 $O(|q|^2)$ で達成可能である．

具体的な処理方法として，まず初期クエリでの条件の中から，これまでの緩和クエリ群で最も利用されていない条件を 1 つ選択し，新たな緩和クエリの条件とする．2 つ目以降の条件は，すでに決定された条件の集合を $C_{q'}$ としたとき

$$\sum_{c \in C_{q'}} count(c, c_new)$$

を最小とする条件 c_new が選択される．ただし $count$ は，過去の反復で利用された緩和クエリに選択された条件のペアの登場回数を記録した関数である．例としてユーザのクエリに含まれる条件を $c1, c2, c3$，現在の 2 個目の推薦レコードを探索中で，1 個目の推薦レコード探索に利用した緩和クエリに $c1, c2$ が含まれている条件を考える．このとき，$count(c1, c2) = 1$ であり $count(c1, c3) = 0$ である．また，$count$ 関数を最小化する条件が複数存在した場合，これまでの緩和クエリ群で最も利用されていない条件が優先される．最終的に緩和クエリが empty となる直前のクエリ条件を全て満たすレコードを候補とする．緩和クエリを決定した後，この緩和クエリにおける全ての条件のペアを記録する．

また緩和クエリを探索する際には，条件を決定するごとにレコードの絞り込みを行う．このとき，2 回目以降の条件決定時には，その以前の条件を全て満たすレコードセットに対してのみ条件の評価を行う．これにより，条件を評価するレコードの数を削減することが可能である．

具　体　例

ここではクエリ探索の実行例を示す．ユーザが 7 個の条件を持つクエリを与え，その結果が empty となっている．この中で，決定済みの推薦レコードセット決定に利用された緩和クエリが $q'_1 = c_1 \wedge c_2 \wedge c_3$ ，$q'_2 = c_3 \wedge c_4 \wedge c_5$ である場合を仮定する．この時 q'_3 は次のように決定される．

まず，1 つ目の条件にはこれまで最も選ばれていない条件とすることから，c_7 が選ばれる．次に選ばれるのは c_7 と最も同時に選ばれていない条件のうち，選択回数が最も少ないものである．このケースでは c_1 から c_6 の全てが該当する．実装においては，同条件の場合にはより早く先に登場する条件が優先されることから c_1 を選択される．同様に 3 つ目の条件は c_1, c_7 と最も共に選ばれた回数が少ないもののうち，選択回数が最も少ないものである．c_3, c_4, c_5 がこれらに該当するため，ここから 1 つを選択する．このような操作を緩和クエリによる検索結果が空になるまで繰り返す．

3.4 レコード検索

3.3 で推薦レコードセットの候補を決定した後，そのレコードセットの中から MMR を最大化するレコードを探索する．こ

のとき推薦レコードセットに選択されたレコード間での距離の最小値を保持することで計算量を削減可能である．しかし，既存手法とは異なり提案手法ではレコードの全探索を行わないことから，追加レコード候補に対する計算の重複を避ける手法を適用することは不可能である．よって提案手法においては，全ての候補レコードに対して実直に MMR スコアの計算を行う必要がある．

3.5 計算量に関する考察

提案手法の正確な計算量は，クエリが含む条件ごとの満たすレコードの割合や複数の各条件を満たすレコード集合間の依存関係に依存するため，算出が困難である．以下では，各カラム間の値の分布に依存関係が存在しないという一般的な条件を仮定した場合の計算量に関して考察する．

c_i のカーディナリティを $\sigma_i(0 \leqq \sigma_j \leqq 1)$ とした場合，提案手法の計算量は

$$O(k(|q|^2 + kl + n\sum_{i=0}^{|q|-1}\prod_{j=0}^{i}\sigma_j))$$

である．全ての項にかかる k はユーザのレコード要求量分の反復を示す．その中の第 1 項は緩和クエリの条件選択と q' に利用した条件の記録にかかる計算量であり，疑似コードにおける $7, 8$ 行目と $16-21$ 行目に対応する．これはクエリに含まれる条件のペアの集合の全要素数によるものである．同様に第 2 項は q' から得られたレコードセットに対して MMR を計算し，推薦レコードセットを決定するのに必要な計算量であり，$15, 16$ 行目に対応する．これは最大 $k-1$ 個の推薦レコードセットに対して計算量 $O(l)$ の追加候補レコードとの距離を計算するためである．なお，ここで追加レコードの個数は考慮していない．第 3 項は q' の探索にかかる計算量であり，$10-14$ 行目に対応する．これは含まれる条件数が最大の q' を発見するまでに評価されるレコードの内の値の数を意味し，$\sum$ 内は各条件に対して評価対象となるレコード内の数を表す．

これらの各パラメータに関して通常 $k, |q|$ は高々2 桁程度である一方で，n はそれ以上であることが多い．よって，$k, |q|, l \ll n$ としたとき，計算量は近似的に

$$O(kn\sum_{i=0}^{|q|-1}(\prod_{j=0}^{i}\sigma_j)) \tag{9}$$

と表現される．ここで計算量の見積もりのために，全ての条件において選択率が同一である場合について検討する．このとき式 9 における σ_j が σ とすると

$$O(kn\sum_{i=0}^{|q|-1}(\prod_{j=0}^{i}\sigma_j)) = O(kn\sum_{i=0}^{|q|-1}(\sigma^j)$$
$$\leqq O(kn\sum_{i=0}^{\infty}(\sigma^j) = O(kn\frac{1}{1-\sigma}) \tag{10}$$

と計算される．式 10 から，ユーザが設定した条件のカーディナリティが低い状況において，より低い計算量が達成されることがわかる．これは，ユーザが設定した条件が厳しいことから発生する，empty-answer 問題の解決という条件設定に合致したものと言える．例として，$\sigma = \frac{1}{2}$ である場合，計算量は $O(kn)$ となり，既存手法と比較して計算量を可能である．

Algorithm 4: カーディナリティ推定を活用した提案手法

Input : dataset D, query Q
Output : recommended dataset RD

1 $RD \leftarrow \emptyset$;
2 $PairMatrix \leftarrow$ an $m \times m$ matrix, initialized by 0;
3 $CountArray \leftarrow [0] * k$, $SubQuery(SQ) \leftarrow [\emptyset] * k$;
4 **foreach** $i \in [0, k-1]$ **do**
5 $SQ \leftarrow [\emptyset] \times k$;
6 $FirstCondition \leftarrow$ the condition with the lowest value in CountArray;
7 $SQ[0] \leftarrow FirstCondition$;
8 $DS_new \leftarrow EvaluateCondition(DS, FirstCondition)$; `/* EvaluateCondition is hereafter denoted as EC */`
9 **foreach** $j \in [1, m-1]$ **do**
10 $Candidates \leftarrow$ $\mathrm{argmin}_{C_j \in (Q \setminus SQ)} \Sigma_{C_l \in SQ} PairMatrix[l][j]$;
11 $ChosenCondition \leftarrow$ $\arg\max_{C_l \in (Candidates)} len(EC(DS_new, C_l))$; `/* If len(DS_new) is larger than the threshold, use cardinality estimation */`
12 $DS_temp \leftarrow EC(RD, ChosenCondition)$;
13 **if** $(DS_temp \setminus RD) = \emptyset$ **then break**;
14 $DS_new \leftarrow DS_temp$;
15 $SQ[j] \leftarrow ChosenCondition$;
16 $RD[i] \leftarrow \arg\max_{x \in (DS \setminus RD)} MMR(RD \cup \{x\})$; `/* choose record */`
17 **foreach** $i \in len(SQ)$ **do**
18 $Index_C_1 \leftarrow GetIndexOfcondition(Q, C_1)$;
19 $CountArray[Index_C_1] += 1$;
20 **foreach** $C_2 \in SubQuery$ such that $C_1 \neq C_2$ **do**
21 $Index_C_2 \leftarrow GetIndexOfcondition(Q, C_1)$;
22 $PairMatrix[Index_C_2][Index_C_2] += 1$;
23 **return** RD

3.6 カーディナリティ推定手法による改善手法

ここで，提案手法について改めて検討を行う．3.3 章におけるクエリの条件決定を行う際，該当する条件が複数存在する場合についての定義が行われていないため．この時，得られる緩和クエリに求められる要件は上に示した通り，「クエリ間の多様性を高い」ことと，「ユーザクエリの条件を多く残していること」の 2 点である．このことを踏まえたとき，共起行列における優先順位が同じであることは，該当するそれぞれの条件を選択した時，クエリ間の多様性は同様に高いことを意味する．よって，これらの条件に優先順位をつけるとき，考慮されるべき事項はユーザクエリの条件を多く残した緩和クエリを得られることである．そこで，提案手法のさらなる改善方法として，共起行列

における優先度が等しい条件の中から，評価実施後にレコードが多く残るようなものを選択する手法を提案する．

条件のレコード数を探索する方法としては，

- 条件を実際に評価し，正確な値を計算する方法
- カーディナリティを利用し，概数を得る方法

の 2 つが存在する．これらの方法にはそれぞれ利点と欠点が存在する．条件を実際に評価する方法は，正確な cardinality を得られる利点を有するが，特に選択された条件が少ない状況においては，条件の評価対象が増大し，実行時間の増加の原因となる．一方で，カーディナリティ推定を利用する方法では，条件の数による実行時間の変化が前者より小さい利点を有するが，条件が増えるに従い同時確率を推定する際の真値との誤差が大きくなるため，正確な条件を得られない可能性が高まる．

そこで提案手法においては，これらの手法を併用することで各手法の欠点を補い，精度と実行時間の両立を図る．具体的には，選択された条件が少なく評価対象が多く残っている場合にはカーディナリティ推定を利用し，評価対象のレコードが一定数を下回った時には，条件を直接評価する方式に切り替える手法を提案する．疑似コードをアルゴリズム 4 に示す．この疑似コードにおいて，6, 9 行目でカーディナリティ推定や緩和クエリ実行による条件の決定が行われる．MMR を利用したレコードの決定は提案手法と同様である．

4 実験結果

本研究では提案手法の有効性を実証するため，既存手法との比較実験を行なった．この章では，実験方法とその実験結果を掲載するとともに，その結果の考察を行う

4.1 実験環境と利用データセット

評価指標

本実験において用いる評価指標は，2.1 章で説明した MMR スコアと実行時間を利用する．

比較手法

本研究において，3.2-3.5 章で説明した比較手法と，3.6 章で説明した改良手法に加え，比較用の手法としてデータセット全体を対象とした貪欲法による手法と，データセットからランダムに選択する手法を実装し，比較実験を実施した。

実験環境

本研究で実施した実験は，全て Apple M2 CPU 及び，メインメモリ 24 GB を搭載した MacOS Ventura 13.2.1 マシン上で実験を行った．

データセット

本研究では，2 つの実データセットにより実験を行なった．実データセットはそれぞれ真偽値から構成されるデータセットと，数値型からなるデータセットを利用している

実験で用いたアルゴリズムはすべて，Python3.8 を実装や実験に利用している．データの格納には PostgreSQL を利用し，提案手法と全探索手法との実験環境の統一のため，実験の際には一度データフレームとしてメインメモリにロードを行ってから，レコードの探索を実施した．以下，本研究で実施した実験の結果の説明と考察を行う．

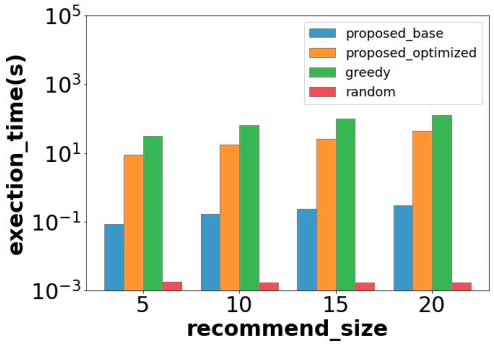


図 2: クエリ 1 における実行時間の計測結果

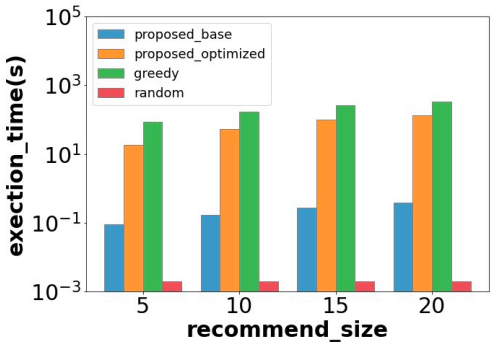


図 3: クエリ 2 における実行時間の計測結果

4.2 真偽値の実データに対する実験結果

本研究で Cars データセットを利用した．これは，クエリリラクゼーションについて扱った既存研究 [4] において利用されたものである．但し，重複したレコードを取り除き使用している．データセットは 31 個のカラムからなる 128443 個のレコードを含み，全てのカラムは真偽値で構成されている．条件数 m が与える影響の比較のため，実験には 2 種類のクエリを利用した．1 つは既存研究に利用された 11 個の制約からなるクエリであり，もう 1 つは実験用に用意した，31 個のカラム全てに対して制約を設けたクエリである．以下，前者のクエリをクエリ 1，後者をクエリ 2 と呼称する．

実行時間

各クエリに対して k の値を変更し各手法の実行時間を計測した結果を記載する．これらの結果は，全て 5 回計測を行った平均値である．この結果によれば，全探索手法と比較して，提案手法は最大で 100 倍以上の高速化を達成した．

3.2 章での正確な計算量の見積もりにおいては，提案手法に $O(|q|k^2)$ の項が存在するため，k を大きく取るほど提案手法と全探索手法との差が縮まる結果が得られると予想される．しかしそのような予想に反して，実験結果によれば，全探索手法と比較して k を変更した際の実行時間の倍率には大きな変化が見られない。この理由として $|q|, k$ が n と比較して十分に小さいからであると考えられ，見積もりにおける近似の有効性を示していると言える．また，クエリ 1 と 2 においては，カーディナ

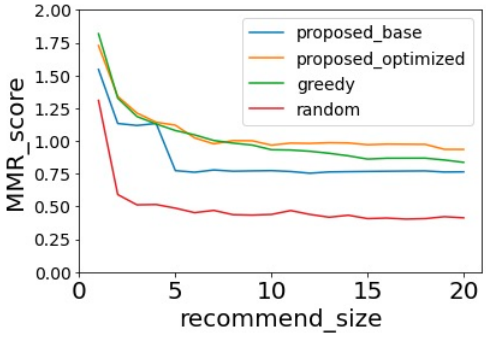


図 4: クエリ 1 における MMR スコアの結果

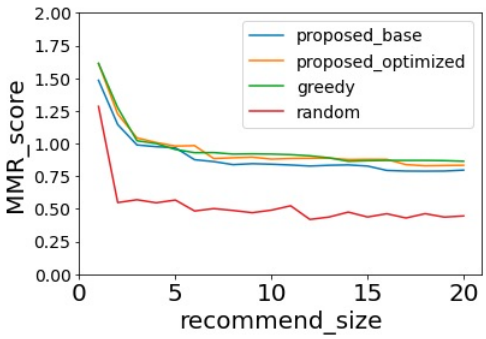


図 5: クエリ 2 における MMR スコアの結果

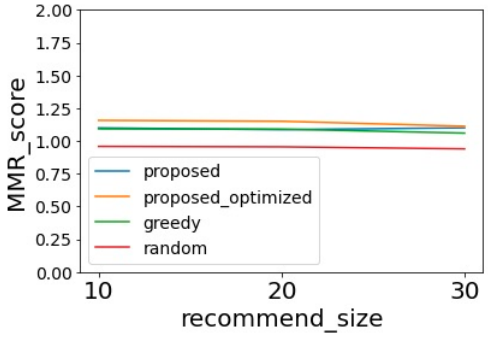


図 6: m を変化させた時の MMR スコア計測結果

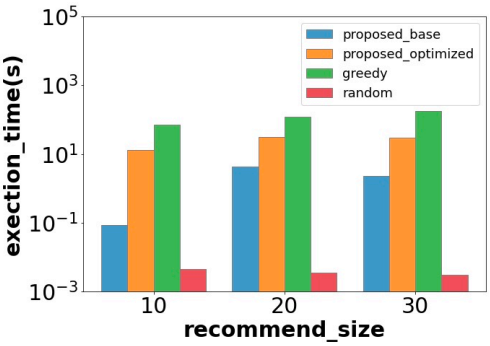


図 7: m を変化させた時の実行時間計測結果

リティを利用しない提案手法は既存手法と比較して実行時間の相対的な増大量が少ないことが見て取れる．これは，提案手法が $|q|$ の変化による実行実行時間への影響が少ないことを示している．

MMR のスコア

図 4，5 に MMR スコアの計測結果を記載する．こちらも実行時間と同様に，5 回実験を行い得られた結果の平均値を掲載する．この実験では多様性と近似性の重み付け項 $\lambda = 1$ と設定している．また，多様性項と近似性項ともに $[0, 1]$ で正規化されている．このため，MMR スコアの値域は $[0, 2]$ である．

実験の結果から，クエリ $1, 2$ の双方において，提案手法は全探索手法に接近し，ベースラインであるランダム手法と比較すると高い MMR スコアを示した．クエリ間の比較では，条件数の小さいクエリ 1 がどの手法においても，比較的高いスコアも示している．また，手法間の比較では，提案手法のうちカーディナリティ手法を活用した方法は素朴な方法と比較して特に貪欲法に接近することが明らかになった．このことから，カーディナリティ推定を活用する手法では，実行時間の増大とトレードオフで，さらなる MMR スコアの向上が見込めることが明らかになった．

また，図 4 におけるカーディナリティ推定なし手法に着目すると，$k = 4$ から $k = 5$ にかけて大幅なスコアの落ち込みが見られる．このことは，貪欲法と比較して，提案手法がデータセットの分布によるスコアの影響を受けやすいことを示唆していると考えられる．

4.3 数値データによる実験

4.2 章では，真偽値からなるデータに対する実験を行った．しかし提案手法は真偽値の他に数値型のデータセットにも適用可能である．そこで本研究では，提案手法が様々なデータセットに適用可能なことを示すため，数値型のデータセットを対象とする実験を実施した．利用したデータセットは KDD2011 である．このデータセットは数値型やカテゴリ型の値を含むデータセットであるが，その中から 31 個の数値型のカラムのみを抜き出して利用している．このデータセットから 200000 件のレコードをサンプリングし，実験を実施した．なお，このデータセットにはクエリが与えられていないため，実験においてはクエリを 5 つランダムに生成し実験に利用した．なお，自動生成したクエリは全ての実験間で共通のものであり，実験条件の統一を図っている．このデータセットにおける全ての実験は，生成した全てのクエリの結果の平均値を取っている．また，MMR スコアの λ の値は 1 に固定しており，スコアの値域は [0,2] である．

以下，実験結果を示す．

4.3.1 クエリサイズに対する実験結果

スコアに関しては，クエリサイズを変更しても結果への影響はほとんど見られなかった．これは，cars データセットに対する実験におけるクエリ 1,2 の結果と同様である．一方で，実行時間に着目した場合，クエリサイズが増えるに従い，実行時間も増大する傾向が見られる．これも cars におけるクエリ $1, 2$ における実験結果と同様である．

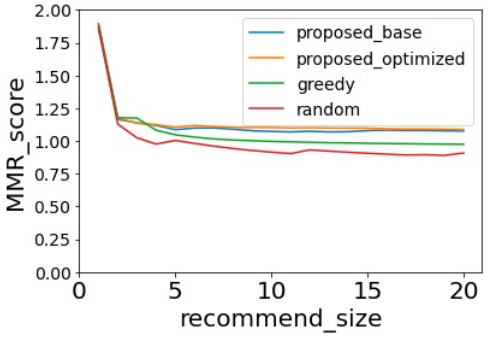


図 8: k を変化させた時の MMR スコア計測結果

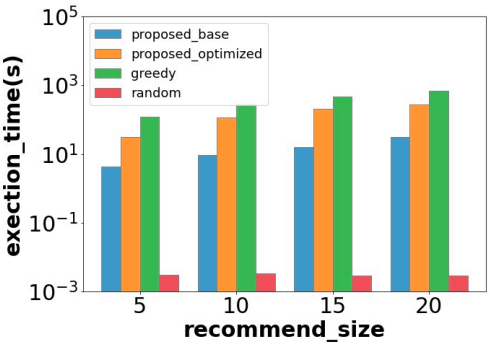


図 9: k を変化させた時の実行時間計測結果

4.4 推薦データセットサイズに対する実験結果

k を変更しての実験結果を図 8,9 に掲載する．実行時間においては概ね k の値に比例した値が得られた．これはカテゴリデータにおける実験結果での傾向と同様である．また，スコア関数の違いによる実行時間への顕著な影響は見られなかった．一方で，手法間の比較では，カテゴリデータの結果と比べて，貪欲法と比較した提案手法の実行時間の倍率が大きい結果となった．

MMR スコアの結果においても真偽値データと同様な傾向の結果が得られた．

一方でクエリサイズを変更した実験において，MMR スコアではクエリサイズを変更した場合の結果への影響はほとんど見られなかった．これは，cars に対する実験におけるクエリ 1,2 の結果と同様である．一方で，実行時間に着目した場合，クエリサイズが増えるに従い，実行時間も増大する傾向が見られる．これも cars におけるクエリ 1, 2 における実験結果と同様である．

ここで，カーディナリティ推定を利用しない提案手法に着目すると，$|q| = 20$ の時の実行時間が $|q| = 30$ の時の実行時間を上回る結果となっている．このように実行時間のばらつきが生じる原因として，極端にカーディナリティの低い条件が含まれているために，レコード MMR の評価対象レコードセットの絞り込みが設計通りにいかないケースが存在することが示唆される．

5 ま と め

本研究において，データベース分野における多様性と近似性の双方を考慮した評価指標を定式化し，また既存手法と比較して高速に提示するレコードセットを探索し，empty-answer 問題を解消する手法を提案した．

また，既存研究で利用されたデータセットとクエリに対して評価実験を行い，手法の設計の通りに精度の低下を抑えながら，大幅な高速化を達成することを確認した．

文 献

[1] Senjuti Basu Roy, Haidong Wang, Gautam Das, Ullas Nambiar, and Mukesh Mohania. Minimum-effort driven dynamic faceted search in structured databases. In *CIKM*, pp. 13–22, 2008.

[2] Nick Koudas, Chen Li, Anthony KH Tung, and Rares Vernica. Relaxing join and selection queries. In *VLDB*, pp. 199–210, 2006.

[3] Chaitanya Mishra and Nick Koudas. Interactive query refinement. In *EDBT*, pp. 862–873, 2009.

[4] Davide Mottin, Alice Marascu, Senjuti Basu Roy, Gautam Das, Themis Palpanas, and Yannis Velegrakis. A holistic and principled approach for the empty-answer problem. *The VLDB Journal*, Vol. 25, No. 4, pp. 597–622, 2016.

[5] Sanjay Agrawal, Surajit Chaudhuri, Gautam Das, and Aristides Gionis. Automated ranking of database query results. *CIDR*, 2003.

[6] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, Vol. 7, No. 1, pp. 1–42, 2016.

[7] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pp. 335–336, 1998.

[8] Ilio Catallo, Eleonora Ciceri, Piero Fraternali, Davide Martinenghi, and Marco Tagliasacchi. Top-k diversity queries over bounded regions. *ACM Transactions on Database Systems*, Vol. 38, No. 2, pp. 1–44, 2013.

[9] Kohei Hirata, Daichi Amagata, Sumio Fujita, and Takahiro Hara. Solving diversity-aware maximum inner product search efficiently and effectively. In *CIKM*, pp. 198–207, 2022.

[10] Francisco Parreño, Ramón Álvarez Valdés, and Rafael Martí. Measuring diversity. a review and an empirical analysis. *European Journal of Operational Research*, Vol. 289, No. 2, pp. 515–532, 2021.

[11] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pp. 22–32, 2005.

謝 辞

本研究は日本学術振興会基盤研究 (B) 22H03903，23H03406，科学技術振興機構 JPMJCR22M2 の助成を受けて実施されています．

相互近接性に基づくCollaborative Metric Learningの意外性向上手法

中嶋　太一†　　陳　　漢雄††　　古瀬　一隆†††

† 筑波大学情報学群情報メディア創成学類　〒305–8577 天王台つくば市天王台１丁目１－１
†† 筑波大学システム情報系　〒305–8577 天王台つくば市天王台１丁目１－１
††† 白鴎大学経営学部　〒323–8585 栃木県小山市駅東通り２丁目２－２
E-mail: †s2010780@s.tsukuba.ac.jp, ††chx@cs.tsukuba.ac.jp, †††furuse@fc.hakuoh.ac.jp

あらまし　コンテンツが絶えず増え続ける現代のWeb空間において，ユーザが受動的に情報を探索できる推薦システムは必須の技術であり，その精度もかなり高くなっている．しかし近年では，ユーザの嗜好に合ったアイテムを推薦し，推薦精度を高めるだけでなく，新たな発見や期待以上の推薦内容を提供することで，推薦システムの魅力を高めることが課題であるとされている．本研究では，Collaborative Metric Learningという埋め込みベースの推薦手法において，ユーザへの意外性を実現する，Mutual Proximity Collaborative Metric Learningという推薦手法を提案する．この手法は，既存手法に対して埋め込み空間の探索方法を改良し，推薦の人気アイテムへの偏重を低減することで，ユーザへ意外性を実現する．また，実験結果によりUnpopularityやSerendipityなどの様々な評価指標において，既存手法を上回ることを示す．

キーワード　セレンディピティ，新規性，意外性，バイアス，協調フィルタリング

1　はじめに

近年のWeb空間では，膨大なコンテンツが存在し，ユーザが好みのアイテムを能動的に発見するのは困難になっており，ユーザは推薦システムを用いて，アイテムを受動的に探索することが一般的になりつつある [1]．多くの推薦モデルは，ユーザやアイテムの属性情報を利用したり [2]，ユーザの過去の取引情報を利用したりする [3] ことで，ユーザの嗜好情報や過去の取引に関連度の高いアイテムを推薦することが可能になっているが，一方で，関連度の高いアイテムの推薦はユーザが事前に予測することが容易であったり，推薦システム特有のバイアスの影響で人気度の高いアイテムに偏った推薦を行ってしまったりする [4] ため，推薦システムを利用するユーザに肯定的な意外性をもたらさない．この肯定的な意外性はセレンディピティと呼ばれており，ただ関連度の高いアイテムを推薦することだけでなく，セレンディピティを実現することが推薦システムにとっての課題となっている [5]．

例えば，ユーザとアイテムを，それらの類似性や嗜好関係が反映されるように共通のベクトル空間に埋め込むことによって推薦を行うCollaborative Metric Learning（以下「CML」という.）[6] という推薦モデルが存在するが，このモデルでも人気アイテムに偏ったアイテムの推薦が行われてしまい，ユーザのセレンディピティを実現できないということが指摘されている [7]．本研究では，このCMLにおいて，ユーザとアイテムの埋め込み空間の探索方法を改良することで，推薦によるユーザのSerendipityを実現する．

本論文の構成は以下のとおりである．まず，第2章において本研究の背景にある関連研究と，先行研究に対する本研究の優位性を説明する．次に，第3章において本研究の提案手法について説明をし，第4章にて提案手法に関する実験を行い，提案手法の有用性を確かめる．そして，第5章において実験の結果を踏まえた考察と今後の課題を示しつつ，本研究のまとめを述べる．

2　関連研究

2.1　Serendipity

Serendipityという語には,「思いもよらなかった偶然がもたらす幸運」というような意味があり，推薦システムの文脈においては，ユーザにとって有用であり，かつ目新しく意外性が高いという意味で用いられるのが一般的である．Serendipityの定義や評価の仕方にはさまざまな方法が提案されているが, [8] の調査によれば，近年のSerendipityに関する研究においては，Serendipityは3つのコンポーネントからなると言われている．1つ目のコンポーネントは関連度である．一般的な推薦システムはユーザが過去に消費したアイテムと関連度が高いアイテムを推薦したり（コンテンツベースフィルタリング [2]），推薦対象のユーザと嗜好情報が似ている他のユーザが過去の消費したアイテムを推薦したりする（協調フィルタリング [9]）ことによって，ユーザにパーソナライズされた推薦を実現している．つまり，関連度が高い推薦ほどユーザにとって有用性が高いと言い換えることができ，Serendipity以外の文脈での推薦システムの研究においても，これを「精度」としてモデルを評価する場合が多い．2つ目のコンポーネントは新規性である．Serendipityの文脈における新規性は，推薦されるアイテムがそのユーザにとって未知であるかどうかを意味し，ユーザが未知であるアイテムが推薦されるほど新規性が高いと言える．そして，新規性を高めるための条件に関してもさまざまな研究が存在するが，アイテム全体の中で人気度の低いアイテムを推薦することで新

規性を高めることができるとされている [10], [11]．これはアイテム全体の中で人気度が低いアイテムは，ユーザの目に触れる機会も少なく，新規性が高まるためである．3 つ目のコンポーネントは予測不可能性である．これはユーザがどれだけそのアイテムの推薦を事前に予測することができないかを意味する．推薦されるアイテムが，ユーザの嗜好情報プロファイルと異なるほど予測不可能性が高くなるとされている [12], [13]．

本研究では，これらのうち，関連度と新規性に注目し，推薦の Serendipity を実現することを目指す．すなわち，推薦されるアイテムが，関連度が高く，かつ人気度の低くなるような推薦を目指す．ただし，推薦の関連度と非人気度はトレードオフの関係にあり，推薦の多様性をあげ，人気アイテムに偏らない幅広い推薦を行おうとすると推薦の関連度は下がってしまう [5]．本研究では，第 4.3.2 項において，関連度と非人気度がトレードオフになっていることを実証しつつ，両者を総合的に評価するための評価指標（第 4.2.3 項）によって提案手法を評価する．

2.2 Collaborative Metric Learning (CML)

CML は協調フィルタリング [9] と Metric Learning [14] を組み合わせた手法である．CML はまず，ユーザのアイテムに対する暗黙的フィードバック [15] を学習することで，ユーザとアイテムを共通のベクトル空間に埋め込む．このとき，あるユーザがあるアイテムを好んでいる場合は，両者のユークリッド距離が近くなるように埋め込みを行う．そして，その埋め込み空間を最近傍探索することによってユーザに推薦を行う．

CML 埋め込み空間には，ユーザとアイテムの関連性が反映されているだけでなく，ユーザ同士・アイテム同士の関連性も反映されているという特徴がある．例えば，距離の近いユーザ同士・アイテム同士は互いに関連度や類似性が高くなっているといえる．これは三角不等式の関連性伝播によるものである [6]．

2.3 Mutual Proximity (MP)

Mutual Proximity（以下「MP」という.）はベクトル空間におけるハブ性 [16] を低減するスケーリング手法である [17]．ハブ性とは，高次元ベクトル空間において，特定のオブジェクトが，他のオブジェクトの最近傍として選ばれやすくなってしまう現象であり，いわゆる「次元の呪い」の一種である．他のオブジェクトと比べて，最近傍として選ばれやすくなっているオブジェクトを「ハブ」という．ハブはデータのドメインや特徴に関わらず，次元の増加とともに出現しやすくなることが知られている [16]．また，ベクトル空間において，ハブとなっているオブジェクトと，そのオブジェクトを最近傍に持つオブジェクトの近傍関係は非対称になることが知られている [16]．例えば，ベクトル空間上に 2 点のオブジェクト x, y があるとする．y が x の最近傍であるとき，x も y の最近傍であれば，この近傍関係は対称であり，両者の親和性が確かに高い状態である．しかし，y が x の最近傍であるが，x が y の最近傍でなければ，この近傍関係は非対称であり，両者の親和性が必ずしも高いとはいえない状態になってしまう．MP はベクトル空間の非対称な近傍関係が対称になるように，ベクトル空間のメトリック（ユークリッド距離やコサイン類似度）をスケーリングすることでハブ性を低減する．

本研究では，非人気アイテムの推薦が行われやすくするために MP の考え方を利用する．埋め込み空間における人気アイテムは，多くのユーザから最近傍として選ばれるオブジェクトであるといえるので，ハブ性の低減手法は，推薦システムにおける人気アイテムへの偏りの低減にも用いることができると考えた．提案手法の詳細は第 3 章にて述べる．

2.4 先行研究に対する優位性および新規性

CML の改良手法は，多くの研究がなされている [7], [18], [19] が，これらの手法はいずれも目的に応じて，CML の損失関数に調整を加えることにより目的の達成を試みている．例えば, [7] の研究では，CML を改良し，Serendipity を実現するための手法として，損失関数を式 (1) として埋め込み空間の学習を行うことで目的の達成を試みた．

$$L_{m2}(d) = \sum_{(i,j)\in S} \alpha_j \sum_{(i,k)\notin S} w_{ij}[m + d(i,j)^2 - d(i,k)^2]_+ \quad (1)$$

ただし，α_j はアイテム j に対する重みであり，j に対する暗黙的フィードバックが多いほど損失が大きくなる．全体における j に対する暗黙的フィードバックの数 X_j とハイパーパラメータ a, b を用いて式 (2) で定義される．

$$\alpha_j = aX_k^{-b} \quad (2)$$

このように，Serendipity を実現するために，CML の損失関数を改良し，埋め込み空間自体を調整している．しかし，第 2.2 節で述べたように，CML の埋め込み空間にはユーザとアイテムの関連性だけでなく，ユーザ同士・アイテム同士の関連性も反映されており，埋め込み空間をユーザへの Top-K 推薦以外のさまざまなタスクにできるという利点があるが，特定の目的のためだけに埋め込み空間を調整してしまうと，この利点が損なわれてしまいかねない．したがって，本研究では，オリジナルの CML の埋め込み空間への調整は行わず，埋め込み空間の探索方法のみを工夫することによって Serendipity を実現する手法を提案し，CML の埋め込み空間のさまざまなタスクへの応用を妨げずに目的を達成することができる．

また，本研究では埋め込み空間の探索方法に MP の考え方を取り入れるが，MP は本来，推薦システムの文脈とは関係なく，高次元ベクトル空間のハブ性を低減するために提案された手法であり，推薦システムにおける人気アイテムへの偏重を低減するための手法ではない．MP を推薦システムに適応した研究はいくつか存在する [20], [21] が，ハブ性の低減手法を，推薦システムにおける人気アイテムへの偏重を低減する手法として応用した研究は，我々の知る限り，本研究を除いて他にない．

3 提 案 手 法

以下では，ユーザ i があるアイテム j に対して暗黙的フィードバックを与えた場合，j を i のポジティブアイテム，(i, j) を

ポジティブペアと呼ぶものとする．また，本研究で用いる記号の内容は表 1 の通りである．

表 1　手法の提案で用いる記号

記号	説明
i, j, k, l	ユーザまたはアイテムに対する添字
$\mathbf{u}_i, \mathbf{v}_j$	ユーザ i，アイテム j の埋め込み表現
S	ポジティブペアの集合
$d(i,j)$	$\mathbf{u}_i, \mathbf{v}_j$ 間のユークリッド距離
$L_m(d)$	損失関数 (4)
m	損失関数に対するハイパーパラメータ
w_{ij}	Weighted Approximate Rank Pairwize 損失 [22]
$P(X)$	あるユーザから全てのアイテムまでの距離の分布
$P(Y)$	あるアイテムから全てのユーザまでの距離の分布
$\mathcal{F}_X, \mathcal{F}_Y$	$P(X), P(Y)$ に対する累積分布関数
α	$MP_{i,j}$ に対するハイパーパラメータ

3.1　埋め込み空間の調整

提案手法ではまず，CML と同様にして，データセット内のユーザとアイテムをベクトル空間に埋め込む．ユーザを起点として，ポジティブアイテムを近づけ，そうでないアイテムを遠ざけるように埋め込み位置を調整する．ユーザ i，アイテム j の埋め込み表現をそれぞれ $\mathbf{u}_i, \mathbf{v}_j$ とし，埋め込み空間上での $\mathbf{u}_i$ と $\mathbf{v}_j$ の距離はユークリッド距離 (3) を用いて定義される．ポジティブペアの集合を S としたとき，損失関数は式 (4) で定義される．

$$d(i,j) = ||\mathbf{u}_i - \mathbf{v}_j|| \tag{3}$$

$$L_m(d) = \sum_{(i,j)\in S} \sum_{(i,k)\notin S} w_{ij}[m + d(i,j)^2 - d(i,k)^2]_+ \tag{4}$$

ただし，$[x]_+ = \max(0, x)$ である．ユーザ i とポジティブアイテム j の距離にマージン m を加えた距離が，そうでないアイテム k との距離よりも大きい場合に損失とみなす．また，w_{ij} はポジティブペアに対する重みであり，Weighted Approximate Rank Pairwize 損失 [22] を用いる．w_{ij} は以下の式 (5) で定義される．

$$w_{ij} = \log(\mathrm{rank}_d(i,j) + 1) \tag{5}$$

ただし，$\mathrm{rank}_d(i,j)$ は，$\mathbf{u}_i$ と $\mathbf{v}_i$ の距離が，u にとっての全ポジティブアイテムのなかで，何番目に近いかの順位を表す．CML では，式 (4) を最小化しながら学習を行うことで埋め込み空間を得る．また，学習はミニバッチ確率勾配降下法をで行い，学習率の更新には Adam [23] を用いる．

3.2　埋め込み空間の探索と推薦

本研究では，ユーザにとって最も関連度が高い，かつ全体の中で人気度が低いアイテムを K 個推薦するという Top-K 推薦をタスクとして設定する．まず，ユーザにとって関連度が高いアイテムを推薦するためには，埋め込み空間上において，ユーザとの距離が近いアイテムを推薦すればよい．これは，CML の埋め込み空間には，ユーザのアイテムに対する嗜好情報が反映されており，ユーザがあるアイテムを好んでいる場合，それらの距離が近づくように学習を行なっているためである．また，人気度が低いアイテムを推薦するためには，埋め込み空間上において多くのユーザから最近傍になっているアイテムの推薦を低減すればよい．これを実現するために，本研究では MP に着想を得た埋め込み空間の探索で推薦を行う，Mutual Proximity Collaborative Metric Learning（以下「MPCML」という.）を提案する．

埋め込み空間上におけるオブジェクト同士の関連度は，埋め込み空間でのユークリッド距離に基づいて決まる．しかし，あるオブジェクト同士の関連度の高さは，他のオブジェクトとの位置関係によって相対的である．つまり，ユークリッド距離の値が小さかったとしても，そのオブジェクトの近くに他のオブジェクトが多数あれば，関連度は必ずしも高いとは限らない．したがって，MPCML ではオブジェクト間の関連度を，MP 同様に，確率として解釈可能な値に変換する．ユーザ i からみた，アイテム j に対する関連度を，i と j の距離 $d(i,j)$ と，i から埋め込み空間中の他の全てのアイテムへの距離の分布 $P(X)$ を用いて，式 (6) として定義する．これは，アイテムからみたユーザに対する関連度を考える場合も同様である．

$$P(X > d(i,j)) = 1 - P(X \leqq d(i,j)) = 1 - \mathcal{F}_X(d(i,j)) \tag{6}$$

例えば，$P(X)$ が図 1 のような分布に従っていた場合，$P(X > d(i,j))$ は図 1 における，灰色で塗られた部分の面積に相当する．この図からも分かる通り，関連度 $P(X > d(i,j))$ は，i と j の距離 $d(i,j)$ が他のアイテムに比べて相対的に近いほど大きくなる．

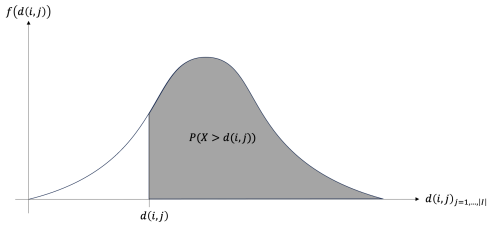


図 1　i から埋め込み空間中の他の全てのアイテムへの距離分布 $P(X)$ が与えられたときの，$d(i,j)$ と $P(X > d(i,j))$ の関係

また，ユーザとアイテムの関連度は必ずしも対称的であるとは限らず，一方のオブジェクトから見て関連度が高かったとしても，もう一方から見ると関連度が低い場合がある．このような非対称な関連性は，アイテムが人気アイテムである場合に現れやすくなる．これは人気アイテムの近くには多数のユーザが集中するためである．MPCML では，このような非対称な関連性を低減し，双方にとって関連度が高いペアに関して推薦を行うようにすることで，関連度が高い，かつ人気アイテムを避けた推薦を実現する．MPCML における，ユーザ i とアイテム j の双方の関連度 $MP_{i,j}$ は，i と j の距離 $d(i,j)$，i から埋め

込み空間中の他の全てのアイテムへの距離分布 $P(X)$，j から埋め込み空間中の他の全てのユーザへの距離分布 $P(Y)$，ハイパーパラメータ α を用いて式 (7) で定義される．

$$\begin{aligned} MP_{i,j}(d(i,j)) &= P(X > d(i,j))^{\alpha} \cdot P(Y > d(i,j))^{(1-\alpha)} \\ &= (1 - \mathcal{F}_X(d(i,j)))^{\alpha} \cdot (1 - \mathcal{F}_Y(d(i,j)))^{(1-\alpha)} \end{aligned} \tag{7}$$

MPCML では，CML による埋め込み空間が与えられた時，以下のようなアルゴリズムで推薦を行う．ただし，ユーザとアイテム全体の集合をそれぞれ U, I とし，距離分布 $P(X), P(Y)$ を [24], [25] での研究内容に基づき，ガンマ分布として仮定する．

（1） ユーザ i から埋め込み空間上の任意のアイテム j までの距離 $d(i,j)[\forall j \in I]$ を計算する．

（2） 1 で求めた $d(i,j)$ に対して，式 (9) を計算することにより，i からすべてのアイテムへの距離分布 $P(X)$ のパラメータを推定する．

（3） 各 j に関して，T 個のユーザをサンプリングし，その集合を U'_j とする．

（4） 各 j に関して，j から U'_j に含まれる任意のユーザまでの距離 $d(l,j)[\forall l \in U'_j]$ を計算する．

（5） 各 j に関して，4 で求めた $d(l,i)$ に対して，式 (9) を計算することにより，j からすべてのアイテムへの距離分布 $P_j(Y)$ のパラメータを推定する．

（6） 1で求めた $d(i,j)$, 2で求めた $P(X)$, 5で求めた $P_j(Y)$ に対して，式 (7) により，u から埋め込み空間上の任意のアイテム i までの $MP_{i,j}[\forall j \in I]$ を計算し，$MP_{i,j}$ が高い順に K 個のアイテムを推薦する．

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} d(x,i), \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(d(x,i) - \mu_x)^2 \tag{8}$$

$$\kappa_x = \frac{\mu_x^2}{\sigma_x^2}, \qquad \theta_x = \frac{\sigma_x^2}{\mu_x^2} \tag{9}$$

第 4.3.3 項でも示す通り，パラメータ推定のためのサンプル数は $T = 100$ 程度で十分で，実行時間に大きな影響はなく，推薦タスクを行う上での時間計算量は $O(|U| + |I|)$ となり，埋め込み空間中のオブジェクトの数に対して線形時間で推薦を行うことができる．

4 実　　験

4.1 データセット

本実験ではデータセットとして，MovieLens データ [26] を用いた．このデータセットは映画に対するユーザの評価履歴の情報を含んでおり，推薦システムの性能評価に広く用いられている．このデータセットには，ユーザから映画（以下「アイテム」という.）に対して 1〜5 の 5 段階で評価を行った結果が含まれているが，先ほども述べた通り，CML は暗黙的フィードバックを前提とするモデルであるため, [6] で行われた実験と同様に，評価値 4, 5 のみを「ポジティブフィードバック」とし，ポジティブフィードバックのレコードのみを抽出した．また，データセットを 3:1 で教師データとテストデータに分割することを考慮し，ポジティブフィードバックを 4 つ以上のアイテムに与えたユーザのみを抽出し実験を行った．

データセットのサイズに関して，サイズの違いによる性能の差を検証するために，MovieLens 100K Dataset（以下「MovieLens 100K」という.）と MovieLens 20M Dataset（以下「MovieLens 20M」という.）の 2 種類のデータセットで検証を行った．先ほど述べた前処理を行った後の状態において，MovieLens 100K はユーザ数が 940，アイテム数が 1447，ポジティブフィードバック数が 55369 であり，MovieLens 20M はユーザ数が 137330，アイテム数が 20720，ポジティブフィードバック数が 9993294 であった．

4.2 評 価 指 標

提案手法を定量的に評価するために，以下のオフライン評価指標を用いる．

4.2.1 Recall

Recall は，ユーザが実際にポジティブフィードバックを与えたアイテムの集合を，推薦アイテム集合がどれだけカバーできたかを表す指標であり，推薦の関連度を評価するために広く用いられている．以下のように定義される．

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

ただし，TP はユーザが実際にポジティブフィードバックを与えたアイテムであり，かつ，推薦システムもユーザがポジティブフィードバックを与えると予測したアイテムの数を表し，FN はユーザが実際にポジティブフィードバックを与えたアイテムであり，かつ，推薦システムもユーザがポジティブフィードバックを与えると予測しなかったアイテムの数を表す．

4.2.2 Unpopularity

推薦システムがどれだけ非人気アイテムを推薦することができるかを表す指標を定義する．その上でまず，あるアイテム i の人気度を以下のように定義する．

$$\text{Popularity}_i = \frac{|U_i|}{|U|} \tag{11}$$

ただし，$|U_i|$ はアイテム i にポジティブフィードバックを与えたユーザの数を表し，$|U|$ は全ユーザ数を表す．この上で，推薦システムの非人気度の評価指標を以下のように定義する．

$$\text{Unpopularity}(R) = \sum_{i \in R} \frac{1}{\text{Popularity}_i} \tag{12}$$

ただし，R は推薦システムがユーザに推薦したアイテムの集合を表す．

4.2.3 Serendipity

第 2.1 節で述べた通り，推薦システムにおいて Serendipity を実現するためには，関連度と非人気度の両方が高いことが必要であるが，4.3.2 項で示すように両者はトレードオフの関係にある．そこで，Serendipity を 2 つの観点から総合的に評価

するために以下の指標を用いる．

$$\text{Serendipity}(R) = \sum_{i \in R} \frac{1}{\text{Popularity}_i} \times 1_A(i) \tag{13}$$

$$1_A(i) = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{if } i \notin A. \end{cases} \tag{14}$$

ただし，A はユーザが実際にポジティブフィードバックを与えたアイテムの集合を表す．

4.2.4 Long-tail Rate

推薦システムがどれだけロングテールアイテム [27] を推薦することができるかを Long-tail Rate [28] という指標を用いて評価する．Long-tail Rate は推薦リスト中に含まれるロングテールアイテムの割合である．ただし，推薦システムにおける全アイテムのうちフィードバック数下位 80%のアイテムをロングテールアイテムとする．

4.3 実験結果

4.3.1 人気度の偏り

MovieLens データにおいて，アイテムの人気度に偏りがあるのかを検証した．データセット内のすべてのアイテムについて，各アイテムが，何人のユーザからポジティブフィードバックを得たのかカウントする．図 2, 3 は，フィードバックの数を縦軸にとり，アイテムをフィードバックが多い順にソートした結果を表す棒グラフである．また，グラフの青い部分は，全アイテムのうちフィードバック数上位 20%のアイテムのフィードバック数を表す．ただし，MovieLens 20M のデータセット（図 3）に関しては，人気アイテムのフィードバック数とそうでないアイテムのフィードバック数の差が非常に大きかったため，対数グラフで描画した．この実験結果より，MovieLens データにおいて，人気アイテムと非人気アイテムには大きな偏りが存在することがわかった．

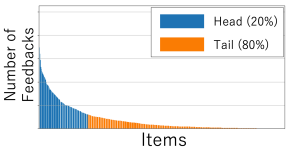


図 2　MovieLens 100K

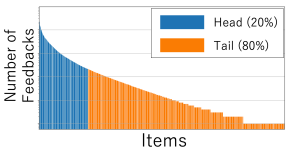


図 3　MovieLens 20M

4.3.2 既存手法と提案手法の比較

第 4.2 節で述べた評価指標を用いて，通常の CML と MP を適応した CML を比較した．式 (7) におけるハイパーパラメータ α の値を 0〜1 の間の 0.1 刻みで変化させながら，各評価指標のスコアを計測した．いずれの評価指標に関しても，推薦するアイテムの数 k は $k = 10$ とした．第 4.1 節で述べた方法でデータを前処理し，データセットを 3:1 で教師データとテストデータに分割した．さらに，テストデータには，ユーザがポジティブフィードバックを与えなかったアイテムのレコードをランダムに追加した．そのとき，ユーザがポジティブフィードバックを与えたアイテムとそうでないアイテムの数が 1:10 になるようにした．第 3.2 節で述べた，確率分布のパラメータ決定のためのサンプリング数 T の値は第 4.3.3 項で示す実験結果に基づき $T = 110$ とした．また，各 α において，各評価指標のスコアを 10 回計算し，その平均値をスコアとした．実装に関しては，[29] のソースコードをベースにして，提案手法を検証できるように拡張した．

実験結果は図 4〜9 に示す通りとなった．まず，図 4,5 からわかるように，Recall と Unpopularity はトレードオフの関係になった．ハイパーパラメータ α の値を 1 に近づけるにつれ，Recall が高くなり，0 に近づけるにつれ Unpopularity が高くなった．ただし，$\alpha = 1$ のとき，MPCML は実質的に，オリジナルの CML と同じ方法での探索が行われる．Unpopularity，Serendipity，Long-tail Rate に関しては，いずれの指標も，オリジナルの CML を大きく上回った．これらのことから，MPCML はオリジナルの提案手法に対し，ある程度ユーザへの関連度を担保しながら，人気度の低いアイテムの推薦を可能にしていると言える．また，トレードオフである関連度と人気度に関しては，その比重をハイパーパラメータによって，調整可能であることがわかった．

4.3.3 距離分布推定のためのサンプリング数

第 3.2 節で述べた，あるアイテム i からすべてのユーザへの距離分布 $P_i(Y)$ のパラメータを推定するためのサンプリング数 T の値を検証した．T の値を 10〜150 の 10 刻みで変化させながら，各評価指標のスコアを計測した．また，各 T において，各評価指標のスコアを 10 回計算し，その平均値をスコアとした．ただし，式 (7) におけるハイパーパラメータ α は，$\alpha = 0.5, 0.8$ とし，その他の実験条件は第 4.3.2 節と同様である．

図 10〜17 が示す通り，MovieLens20M に対する Serendipity 以外の全て評価指標において，T の大きさは $T = 50$ 程度あれば，十分安定したスコアを出すことができ，それ以上大きくしてもスコアは上昇しない．MovieLens20M に対する Serendipity のスコアに関しても，$T = 110$ 程度あれば十分であり，それ以上大きくしてもスコアは大きく上昇しない．このことから，MPCML では，距離分布にガンマ分布を仮定することにより，$O(|U| + |I|)$ のデータセットのサイズに対して線形な時間計算量で計算を行うことができるといえる．

5 まとめ

本研究では，CML における埋め込み空間に対して，MP の考え方に基づいた探索方法を適用する MPCML を提案した．これにより，人気アイテムに偏った推薦を低減し，ユーザにとって関連度が高い，かつ人気度が低いアイテムの推薦が可能になり，推薦によるユーザにとっての肯定的な意外性（Serendipity）が実現されやすくなった．また，MPCML は既存の手法と異なり，意外性向上のために CML の埋め込み空間自体を変更しないため，CML の他タスクへの応用可能性を損なうことなく目的を達成することができる．

今後の課題としては，異なるデータサイズやドメインに対しても本研究と同程度の性能を出すことができるかを検証する必

要があると考えられる．本研究では MovieLends 100K データセットと 20M データセットにおいて，評価結果に差があったことから，MPCML がデータセットのサイズに対してスケーラブルな手法であるかは疑問が残る．ドメインに関しても，映画の評価データ以外での実験は行なっていないので，幅広いデータに対して MPCML が有効であるかを検証する必要がある．また，評価方法に関しても課題があると考えられる．本研究で用いた評価指標である Unpopularity と Serendipity に関しては，直感に沿った評価式ではあるものの，評価指標の有効性の検証は行われていない．そして，本研究ではオフライン評価のみを行なったが，より正確に有用性を検証する上では，オンライン評価も必要であると考えられる．特に，推薦が Serendipity を生み出すかどうかは実際にはユーザの感覚によって決まるものであり，個人差も存在するので，オフラインの評価のみでの完全な検証は難しいと考えられる．

参 考 文 献

[1] 神嶌敏弘. 推薦システム-情報過多時代をのりきる (<特集>情報のフィルタリング). 情報の科学と技術, Vol. 56, No. 10, pp. 452–457, 2006.

[2] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pp. 995–1000, 2010.

[3] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, Vol. 42, No. 8, pp. 30–37, 2009.

[4] Òscar Celma and Pedro Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *NETFLIX '08*, 2008.

[5] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, p. 257–260, New York,NY,USA, 2010. Association for Computing Machinery.

[6] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. Collaborative metric learning. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 193–201, Republic and Canton of Geneva,CHE, 2017. International World Wide Web Conferences Steering Committee.

[7] Akiko Yoneda, Ryota Matsunae, Haruka Yamashita, and Masayuki Goto. A method to improve serendipity of recommendation lists based on collaborative metric learning. *Total Quality Science*, Vol. 9, No. 2, pp. 62–73, 2023.

[8] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, Vol. 111, pp. 180–192, 2016.

[9] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, p. 175–186, New York,NY,USA, 1994. Association for Computing Machinery.

[10] Òscar Celma Herrada. *Music recommendation and discovery in the long tail*. PhD thesis, Universitat Pompeu Fabra, 2009.

[11] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, p. 109–116, New York,NY,USA, 2011. Association for Computing Machinery.

[12] Panagiotis Adamopoulos and Alexander Tuzhilin. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Trans. Intell. Syst. Technol.*, Vol. 5, No. 4, dec 2014.

[13] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. Metrics for evaluating the serendipity of recommendation lists. In Ken Satoh, Akihiro Inokuchi, Katashi Nagao, and Takahiro Kawamura, editors, *New Frontiers in Artificial Intelligence*, pp. 40–46, Berlin,Heidelberg, 2008. Springer Berlin Heidelberg.

[14] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, Vol. 10, p. 207–244, jun 2009.

[15] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, 2008.

[16] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, Vol. 11, No. 86, pp. 2487–2531, 2010.

[17] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Local and global scaling reduce hubs in space. *J. Mach. Learn. Res.*, Vol. 13, No. 1, p. 2871–2902, oct 2012.

[18] Xiao Zhou, Danyang Liu, Jianxun Lian, and Xing Xie. Collaborative metric learning with memory network for multi-relational recommender systems. *CoRR*, Vol. abs/1906.09882, , 2019.

[19] Chanyoung Park, Donghyun Kim, Xing Xie, and Hwanjo Yu. Collaborative translational metric learning. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 367–376, 2018.

[20] Peter Knees, Dominik Schnitzer, and Arthur Flexer. Improving neighborhood-based collaborative filtering by reducing hubness. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, p. 161–168, New York,NY,USA, 2014. Association for Computing Machinery.

[21] Arthur Flexer and Jeff Stevens. Mutual proximity graphs for improved reachability in music recommendation. *Journal of New Music Research*, Vol. 47, No. 1, pp. 17–28, 2018. PMID: 29348779.

[22] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, Vol. 81, pp. 21–35, 2010.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[24] Michael Casey, Christophe Rhodes, and Malcolm Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio,Speech,and Language Processing*, Vol. 16, No. 5, pp. 1015–1028, 2008.

[25] Rui Cai, Chao Zhang, Lei Zhang, and Wei-Ying Ma. Scalable music recommendation by search. pp. 1065–1074, 09 2007.

[26] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, Vol. 5, No. 4, dec 2015.

[27] Chris Anderson. *The long tail : why the future of business is selling less of more*. Hyperion, 2008.

[28] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. Explainable fairness in recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM, July 2022.

[29] Matsui Ryo. Pytorchcml. `https://github.com/hand10ryo/PyTorchCML`, 2021.

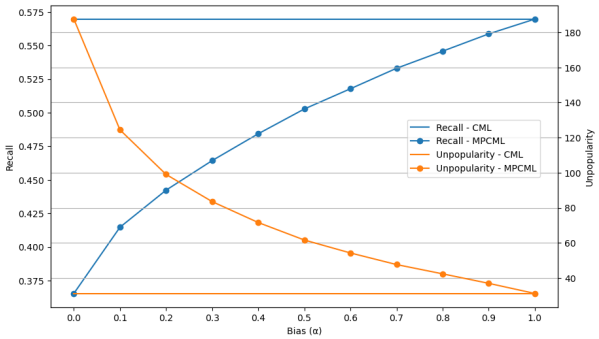


図 4　Recall vs Unpopularity (MovieLens 100K)

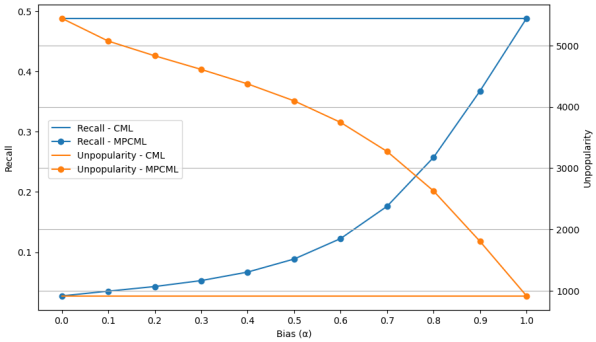


図 5　Recall vs Unpopularity (MovieLens 20m)

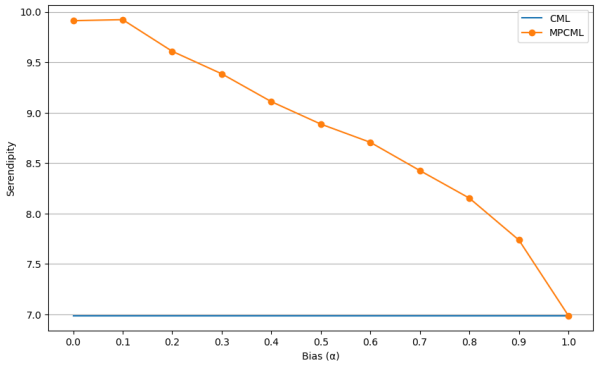


図 6　Serendipity (MovieLens 100K)

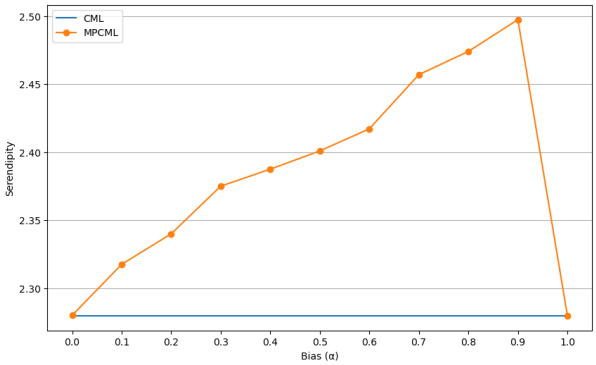


図 7　Serendipity (MovieLens 20M)

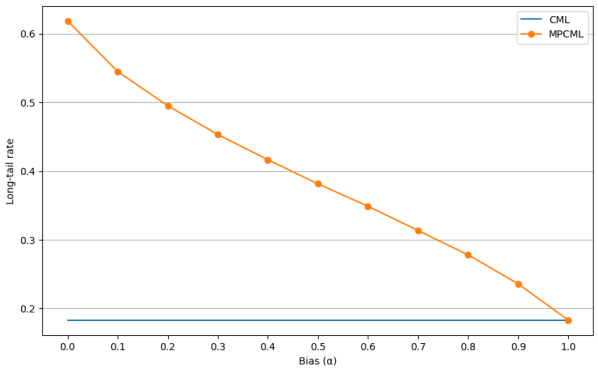


図 8　Long-tail Rate (MovieLens 100K)

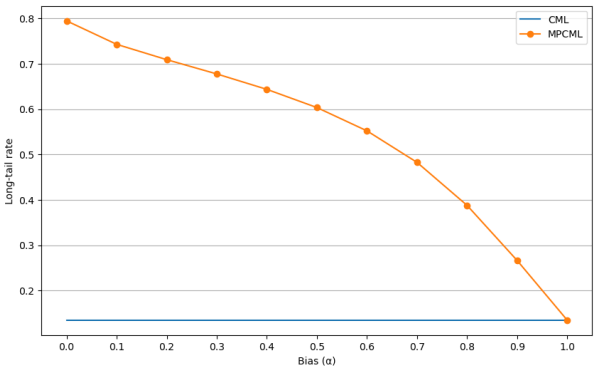


図 9　Long-tail Rate (MovieLens 20M)

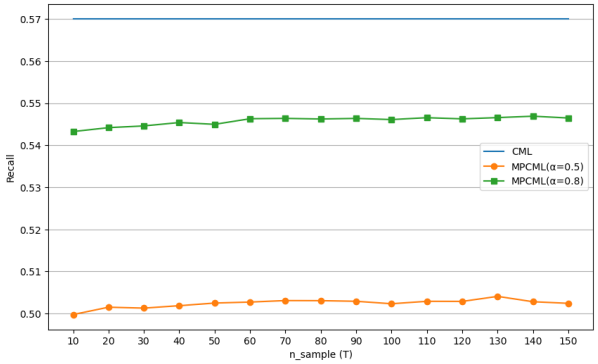


図 10　Recall (MovieLens 100K)

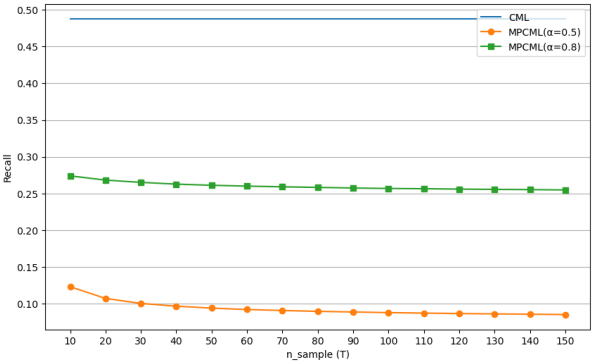


図 11　Recall (MovieLens 20M)

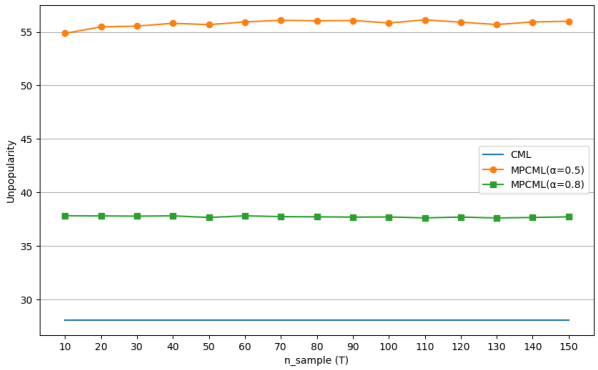


図 12　Unpopularity (MovieLens 100K)

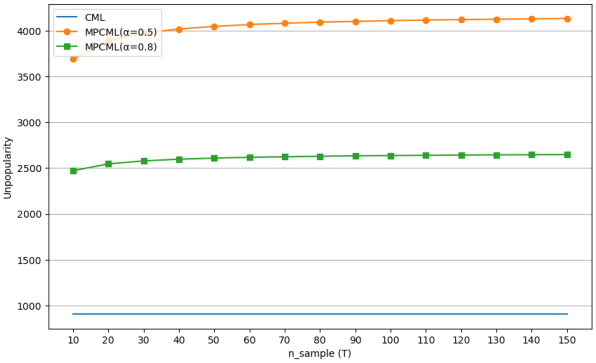


図 13　Unpopularity (MovieLens 20M)

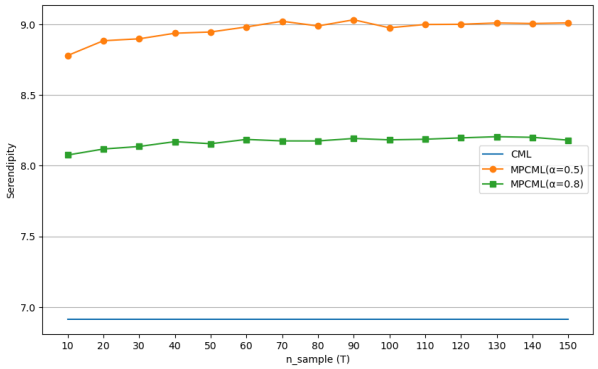


図 14　Serendipity (MovieLens 100K)

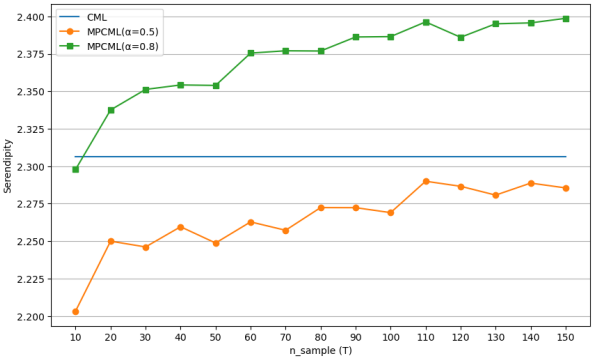


図 15　Serendipity (MovieLens 20M)

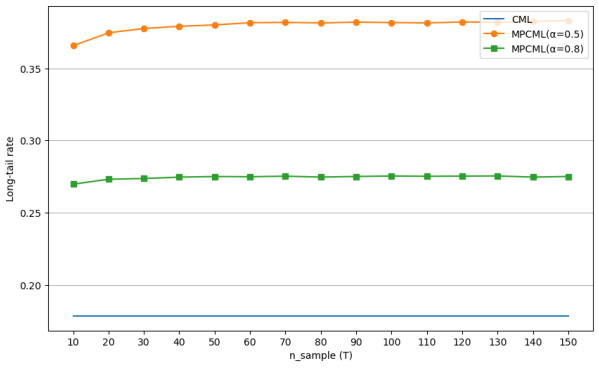


図 16　Long-tail Rate (MovieLens 100K)

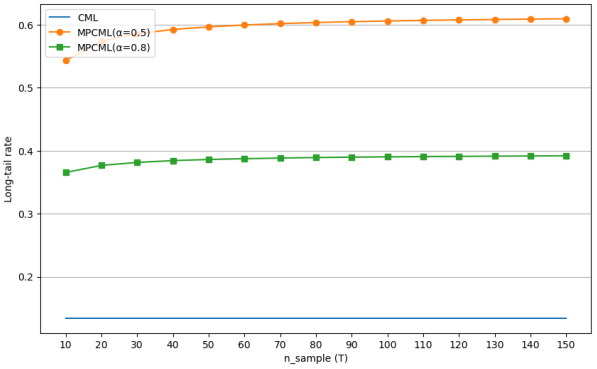


図 17　Long-tail Rate (MovieLens 20M)

δ-spannerを使用した位置匿名化による モバイル端末からのイベント検索

石神　京佳† 　榎　 美紀†† 　小口　正人†

† お茶の水女子大学　〒112–8610 東京都文京区大塚２丁目１－１
†† 日本アイ・ビー・エム株式会社　〒103–8510 東京都中央区日本橋箱崎町 １９－２１
E-mail: †{kyoka,oguchi}@ogl.is.ocha.ac.jp, ††enomiki@jp.ibm.com

あらまし　ユーザがモバイル端末の位置情報を提供して，現在地付近の情報を得るようなサービスは近年多く存在するが，位置情報をそのままサーバに送信して情報分析に使用されることはプライバシ上の懸念がある．滞在人口数のリアルデータを利用し，差分プライバシを位置情報匿名化に拡張した技術である Geo-Indistinguishability を満たすようなユーザ位置のダミー位置をサーバへの問い合わせに使用することで，ユーザの位置情報プライバシを考慮したイベント推薦システムを実装する．本稿では実装範囲の拡大のため，ユーザ位置とノイズ付与後位置（ダミー位置）の距離の期待値が最小になるように最適化を行う匿名化メカニズムにおいて，δ-spanner を使用し，計算時間の短縮と有効性の評価を行う．

キーワード　位置情報プライバシ，情報推薦，SNS データ

1　はじめに

近年，ソーシャルネットワーキングサービス（以下 SNS とする）の普及に伴い，SNS 上にはローカルイベントや開催中のイベントをはじめ，大小様々な規模のイベントに関する情報が投稿されるようになった．同研究室では，SNS 上の膨大なイベント情報をサーバ側に収集し，問合せユーザの位置情報や趣味趣向を利用することで，ユーザに対し特定の場所や時間で開催されるイベントを推薦するシステムが研究されてきた [1], [2]．しかし，問い合わせユーザにとって，位置情報をはじめとする個人情報をサーバに送信して情報分析に使用されることはプライバシ上の懸念がある．この問題を解決するため，サーバを信頼しないようなユーザデータのプライバシ保護に関する研究が行われており，保護対象データを位置情報データに拡張する動きも盛んである．

本研究では，SNS の一種である Twitter [3] 上からイベント情報を取得し，ユーザの必要とするイベント情報の個数的な条件，地理的な制約条件を満たすイベント情報検索システムの実装における，ユーザの位置情報匿名化手法を提案する．滞在人口数のリアルデータ [4] を利用し，差分プライバシ [5] を位置情報匿名化に拡張した技術 Geo-Indistinguishability（以下 Geo-I とする）[6] を満たすようなユーザ位置のダミー位置を算出し，サーバへの問い合わせに使用する手法 [7] では，リアルデータに対し理論式であるプライバシ保護技術を適応させ，実際にイベント情報の問い合わせを行った．ユーザ位置 x とダミー位置 z のユークリッド距離 $d_Q(x,z)$ の期待値が最小になるような最適化問題を解くことで Geo-I を満たすメカニズム K を得ていた．この最適化問題は線形計画法で解くことができるが，計算量が $O(|X|^3)$ と大きく，広範囲での実装に時間がかかるため，限定的な範囲での実装に留まっていた．

そこで本稿では，Bordenabe [8] らが提案した δ-spanner を用いた位置情報の簡略化を行うことで計算量を $O(|X|^2)$ に減らしたメカニズムを提案手法に適応させた場合の計算時間の短縮と有効性の評価を行う．

2　先行研究

本章では，本研究の先行研究について紹介する．同研究室工藤ら [1] が，Twitter のツイートデータからのイベントに関する情報の抽出に関する研究を行っている．工藤らの提案システムは，ツイートの抽出パート，イベントのカテゴリ分類パート，ユーザへの情報配信準備パートの大きく分けて 3 つで構成されている．また，同研究室今井ら [2] が，ユーザの過去のツイートデータから趣味趣向を判別し，工藤らのシステムを用て収集したイベント情報のデータを，取得したユーザ趣向に基づいて順位付けし推薦するシステムを提案している．

先行研究ではユーザの位置情報やユーザの過去のツイート情報をそのままサーバに付与しているため，プライバシ上の課題があった．本研究ではユーザの位置情報プライバシ保護のためのデータ処理を行うことで，先行研究のプライバシを考慮したモデルへの拡張を行い，よりユーザが安心して SNS 上のイベント情報を検索できるようなシステムの実装を目指す．

3　プライバシ保護に関する研究

本章では，関連研究として，近年盛んに行われているユーザのプライバシを保護しながらデータを活用するためのプライバシ保護技術と，位置情報データの保護に関する研究について紹介する．

3.1　k-匿名化

代表的なプライバシ保護手法の一種に，Sweeney によって提

案された k-匿名化 [9] が挙げられる．k-匿名化とは，k-匿名性というプライバシ保護指標を満たすようにデータを処理する手法である．k-匿名性は，データ内に同じ準識別子の組み合わせを持つデータが k 件以上存在するようにデータを変換し個人のデータが特定される確率を k 分の 1 以下に低減させるという指標である．この k-匿名化モデルは，位置情報，特に軌跡情報の保護 [10] などに応用されている．

3.2 差分プライバシと局所差分プライバシ

代表的なプライバシ保護保証の評価手法の一種に，Dwork らにより発表された差分プライバシ [5] が挙げられる．これは，あるユーザのデータが含まれるデータベースと含まないデータベースの差分を少なくすることで，攻撃者が統計的結果を得ても，結果がどちらのデータベースから得たものなのか見分けを付きにくくするという概念である．差分プライバシの概念では，データ収集者であるサーバを信頼し，データ解析結果を第三者に提供する際のプライバシ保護を想定している．近年ではユーザがデータ収集者を信頼せず，ユーザ自身が各々データに差分プライバシを満たすようなノイズを加える，局所差分プライバシに関する研究 [11] も盛んである．図 1 に差分プライバシと局所差分プライバシの概要を示す．

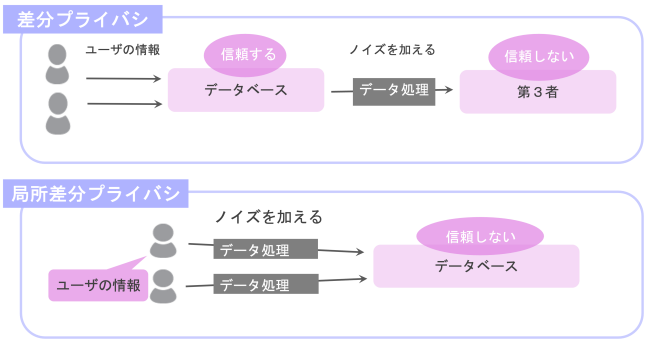


図 1: 差分プライバシ・局所差分プライバシの概要

3.3 Geo-Indistinguishability（Geo-I）

Geo-I は，Andrés ら [6] によって提案された，差分プライバシにおけるデータを位置情報，ハミング距離をユークリッド距離に置き換え，位置情報データの保護に応用したプライバシ保護基準である．局所差分プライバシと同様に，データ収集者を信頼しないようなモデルで，攻撃者に大まかな位置が予測されても，具体的な位置は予測されないようにするという概念である．X をユーザの位置の集合，Z をメカニズム K によって曖昧化した結果得られる位置 (ダミー位置) の集合とする．プライバシ指標（一般的にはユークリッド距離）d_χ とプライバシを制御するパラメータ ϵ が与えられたとき，任意の二つの点 $x, x' \in X$ において以下が成り立てば，位置 x が位置 $z \in Z$ に属する確率 K_{xz} を出力するメカニズム K は ϵ-Geo-I を満たすという．

$$K_{xz} \leqq e^{\epsilon d_\chi(x,x')} K_{x'z}$$

メカニズム K が ϵ-Geo-I を満たすとは，ユーザが確率分布 K_{xz} に従ってダミー位置 z を出力したときに，攻撃者が z を観測しても確率分布 K_{xz} から出力されたのか，$K_{x'z}$ から出力されたのか見分けがつかない状態で，ϵ が小さいほど匿名性は高くなる．

また，高木ら [12] は Geo-I の弱点として，攻撃者が道路ネットワークを考慮した場合，ユーザの予測の精度が上がってしまうことを挙げ，道路ネットワークにおける位置情報プライバシという題で道路ネットワークを考慮した基準 Geo-Graph-Indistinguishability(GGI) を提案している．

このように，有効性のあるユーザデータのプライバシ保護基準が研究されている．プライバシ侵害の判定基準はユーザによって異なる主観的な概念であるため，明確な基準を定めるのは非常に困難であるが，これらの数式で示された保護基準を用いることで，プライバシ保護度合いを値 ϵ で表現でき，統計的なデータの保護に応用することが可能となった．

4 提案システムの概要

本研究ではユーザの位置情報プライバシを保護しつつ，膨大な SNS 上のイベント情報から，ユーザに適したイベント情報を推薦するシステムの構築を目指す．提案するシステムモデルを図 2 に示す．

クライアントは，ユーザのモバイル端末を想定している．サーバは先行研究で紹介した工藤ら [1] が提案する手法を参考にし，パブリックなツイートデータからのイベント情報抽出・分析処理を行う．クライアントはユーザのプライバシに関わるような情報の保護処理をしてからサーバへイベント情報の問い合わせをする．本提案手法ではメッシュ単位の滞在人口数データを使用し算出した確率分布を満たすダミー位置を問い合わせに使用する．クライアントは問い合わせ結果で得られたイベント情報から，ユーザが必要としている範囲や個数などの制約を満たすイベント情報を選択し，ユーザへのイベント情報推薦を行う．

4.1 ツイートデータからのイベント情報抽出

X(旧 Twitter) API [13] を使用しキーワード (地名) を含むツイートを収集する．尚，同一内容のツイートの重複した収集を防ぐため，リツイート機能を用いて再投稿されたツイートは除いて収集する．続いて，取得したツイートデータから，正規表現を用いて日付・イベント名・イベント開催地（スポット名 [1] または住所）が含まれるツイートを抽出し，それらをイベント情報としてイベント名が重複するものを除きデータベースに格納する．表 1 にデータベースとして格納されるイベント情報の一例を示す．

1：東京 23 区内のアミューズメント施設，ミュージアム，ショッピング施設，エンターテインメント施設，温泉，劇場，ホールを東京ウォーカー [14] とナビタイム [15] から抜粋しイベントスポット辞書を作成する．ツイートにイベントスポット辞書内の要素が存在する場合は，その要素をイベント開催地として保持

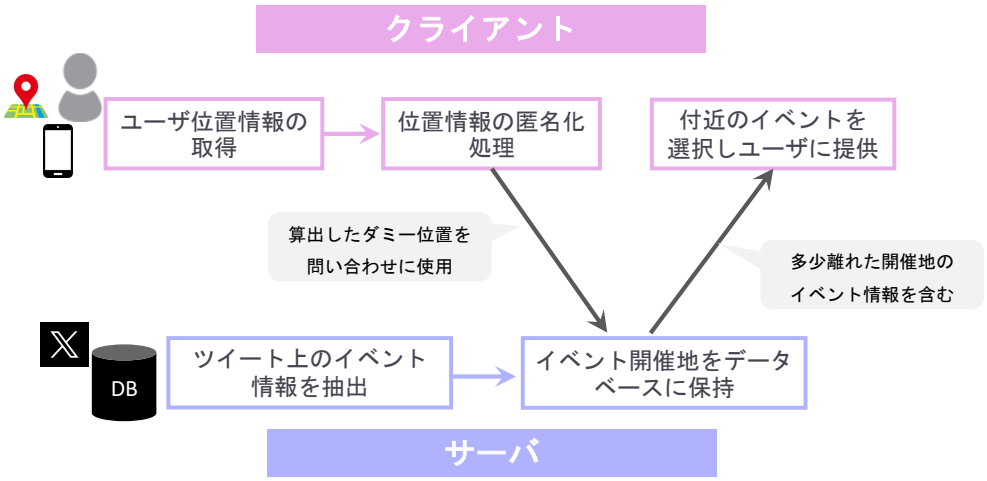


図 2: 提案システムモデル

表 1: データベースに格納されているイベント情報 (一部抜粋)

Event Name	Location	Start date	End date
ナイアガラ盆踊り	Yoyogi National Stadium	2023/7/16	2023/7/16
mode.展	Design Festa Gallery Harajuku	2023/8/25	2023/8/27
うまみちゃん風似顔絵	Laforet Harajuku	2023/7/30	2023/7/30
#台湾フェスティバル™TOKYO2023	Ueno Onshi Park	2023/6/17	2023/6/17
エリザベート 1878	Bunkamura	2023/6/16	2023/6/16

4.2 地域メッシュ情報の付与

本研究では，地域メッシュ [16] 情報を用いて，ユーザの位置情報のプライバシ保護処理を試みる．地域メッシュは，統計に利用するために緯度・経度に基づいて地域を隙間のない網の目のような区画としたもので，それぞれの区画に地域メッシュコードが割り振られている．4.1 節で作成したイベント情報データベースにおいて，イベント開催地にスポット名または住所が保存されているイベント情報に対し，Geolocation API [17] を用いてイベント開催地の緯度経度を取得し，地域メッシュコードを算出し保持しておく．本研究では，現在日本で用いられている，昭和 48 年 7 月 12 日行政管理庁告示第 143 号に基づく「標準地域メッシュ・コード」から以下の区画を使用する．

第 1 次メッシュ (第 1 次地域区画)：
緯度の間隔 40 分，経度の間隔 1 度，一辺の長さ約 80km，メッシュコード 4 桁
第 2 次メッシュ (第 2 次地域区画)：
第 1 次メッシュを緯線方向及び経線方向に 8 等分してできる区域，一辺の長さ約 10km，メッシュコード 6 桁
第 3 次メッシュ (基準地域メッシュ)：
第 2 次メッシュを緯線方向及び経線方向に 10 等分してできる区域，一辺の長さ約 1km，メッシュコード 8 桁
分割地域メッシュ (2 分の 1 地域メッシュ)：
第 3 次メッシュを緯線方向及び経線方向に 2 等分してできる区域，一辺の長さ約 500m，メッシュコード 9 桁

4.3 プライバシを考慮した問い合わせ

4.3.1 Geo-I を用いた位置匿名化

地域メッシュと滞在人口数データを利用し，3.3 で紹介した Geo-I を満たすようなユーザ位置のダミー位置を算出し，サーバへの問い合わせに使用することで，ユーザ位置の匿名化を保護度合い ϵ で行う．

Geo-I を満たすメカニズム K として，Bordenabe [8] らが提案した，ユーザ位置が $x \in X$ である事前確率 π_x を用いつつ，ユーザ位置 x とダミー位置 z のユークリッド距離 $d_Q(x,z)$ の期待値が最小になるように最適化を行う手法を採用する．このメカニズム K を使用したときのユーザ位置 x とノイズ付きの位置 z との距離 $d_Q(x,z)$ の期待値は以下のように表せる．

$$\sum_{x,z \in \chi} \pi_x k_{xz} d_Q(x,z)$$

従って，Geo-I を見たしつつ，$d_Q(x,z)$ が最小となるメカニズムは，以下の最適化問題を解くことで得られる．

$$\begin{aligned} Minimize: & \sum_{x,z \in \chi} \pi_x k_{xz} d_Q(x,z) \\ Subject\ to: & \ k_{xz} \leqq e^{\epsilon d_\chi(x,x')} k_{x'z} \qquad x,x',z \in \chi \\ & \sum_{z \in \chi} k_{x,z} = 1 \qquad x \in \chi \\ & k_{xz} \geqq 0 \qquad x,z \in \chi \end{aligned}$$

携帯端末の位置情報データを集計した，単位時間あたりの分割地域メッシュ単位の滞在人口数データ [18] を用いた重みづけをユーザ位置の事前確率 π_x とし，ユーザ位置の匿名化を保護度合い ϵ で行う．求めた ϵ-Geo-I を満たす確率分布に従ってダミー位置を算出し，問い合わせの位置として使用する．今回の実装では図 3 に表すように，分割地域メッシュ区画に対し，滞在人口数リアルデータを事前確率 π_x とし，ダミー位置として選択された各メッシュの中心位置の緯度経度をサーバに送信し，イベント情報の問い合わせを行う．

4.3.2 δ-spanner を用いた位置情報の簡略化

4.3.1 節で述べた，ユーザ位置 x とダミー位置 z のユーク

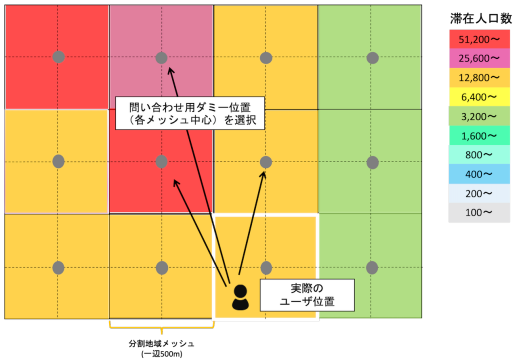


図 3: Geo-I を用いた位置匿名化手法の概要

リッド距離 $d_Q(x,z)$ の期待値が最小になるような最適化問題を解くことで Geo-I を満たすメカニズム K を得る手法において，この最適化問題は線形計画法で解くことができるが，計算量が $O(|X|^3)$ と大きいため，広範囲での実装に時間がかかる．

そこで本稿では，Bordenabe [8] らが提案した δ-spanner を用いた位置情報の簡略化を行うことで計算量を $O(|X|^2)$ に減らした場合の計算時間の短縮と有効性の評価を行う．この δ は拡張係数と呼ばれ，1 に近いほどメカニズム K の有効性は最適なものに近づき，拡張係数 $\delta = 1$ のとき，得られるメカニズムは δ-spanner を用いない場合と同じである．

δ-spanner を用いた位置情報簡略化について詳しく説明する．図 4 のような，ノードが位置の集合 X，エッジの重みが距離関数 d_x で与えられている格子状の無向グラフを考える．

全ての位置間のエッジを増加する順に検討し，2 つの位置間の最小パスの重みが，距離の δ 倍より大きい場合，そのエッジを spanner に追加し，δ 倍以下の場合は最小パスの重みを spanner とする．G がノード集合 X の重み付きの無向グラフであり，G によって誘導される距離関数を d_G と表し $d_G(x,x')$ は G 中のノード x と x' 間の最小パスの重みを表すとする．d_x を d_G で近似することができる場合，G は X の spanner（またはスパニング・グラフ）であるといい，以下が成り立つ．

$$d_G(x,x') \leqq \delta d_\chi(x,x') \quad \forall x,x' \in \chi$$

拡張係数 δ を持つ X の spanner を，X の δ-spanner と呼ぶ．図 4 青色の実線は，拡張変数 $\delta = 1.08$ の X の spanner の例である．このように青色の実線が破線の $\delta = 1.08$ 倍以下の場合，実線として地理ネットワークの距離を近似することで位置情報の簡略化を行う．

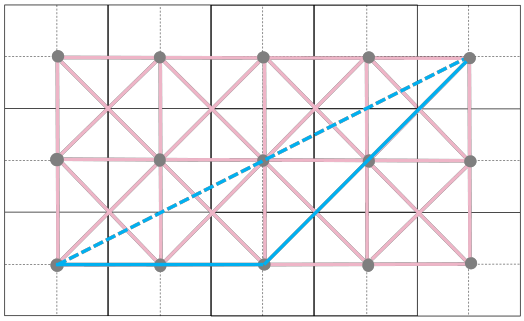


図 4: ノードが位置の集合 X，エッジの重みが距離関数 d_x で与えられている無向グラフ

5 実　　験

5.1 計算時間の比較

実装の区画を図 5 のように増加させた場合の，δ-spanner を用いたメカニズムの使用有無によるダミー位置出力に掛かる計算時間の比較を図 6 に示す．それぞれ 10 回実行し spanner 未使用を 1 としたときの計算時間の比を区画ごとに表している．なおクライアント端末として表 2 のラップトップを今回の比較実験で使用した．

表 2: 実験に使用したマシン性能

クライアント MacBook Air	OS	macOS Ventura バージョン 13.2.1
	CPU	Apple M2
	Number of Cores	8
	Main Memory	16GB

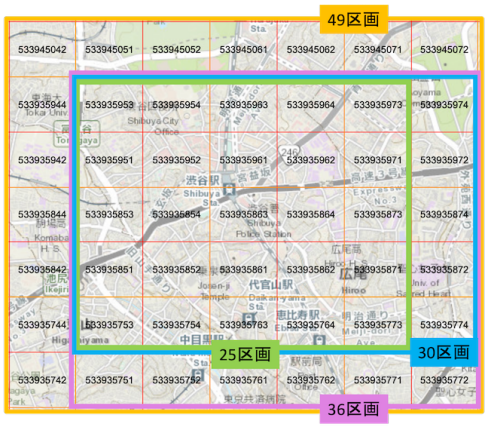


図 5: 実験で使用した区画（渋谷駅周辺）

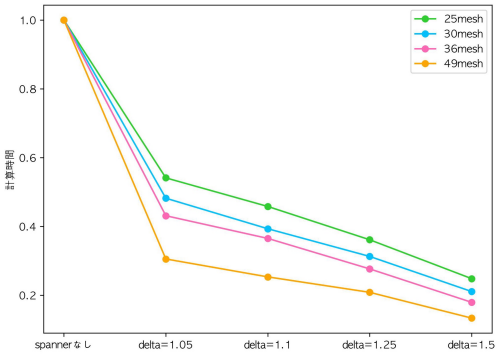
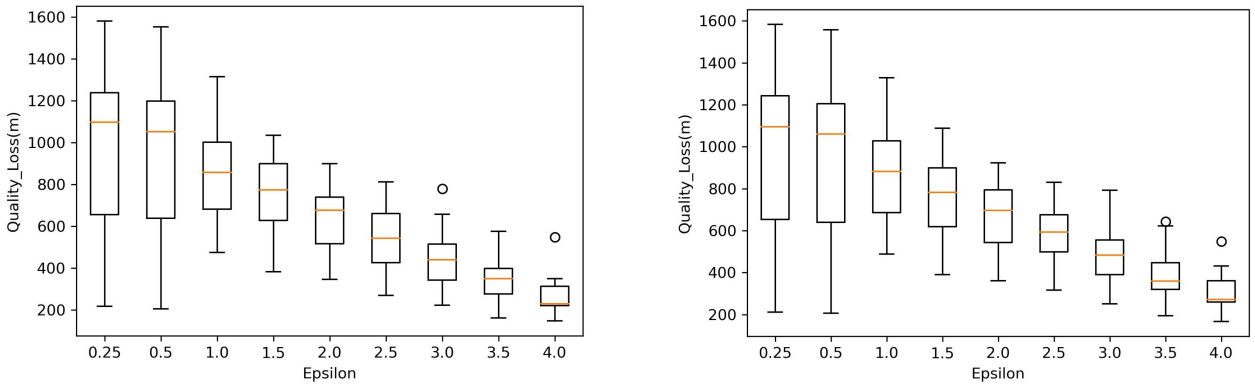


図 6: δ-spanner 使用有無による計算時間の比較

実装範囲を広げた場合，spanner を用いない場合，36 区画の実装では約 4.20 秒，49 区画の実装では約 11.15 秒の計算時間が掛かるのに対し，spanner を用いた場合は $\delta = 1.05$ における 36 区画の実装では約 1.81 秒，49 区画の実装では約 3.40 秒となり，計算時間が 1/2 以下に短縮される．



(a) spanner 未使用 (平均計算時間約 1.39 秒)　　(b) spanner 使用 $\delta = 1.1$(平均計算時間約 0.64 秒)

図 7: δ-spanner 使用有無による $QL(m)$ の比較

5.2　有効性の比較

ϵ の値と実際のユーザ位置とダミー位置のずれ (m) を Quality Loss(以下 QL) とする．

QL=(各メッシュ区画がダミー位置として選ばれる確率) × (実際のユーザ位置とダミー位置の属するメッシュ区画中心点の距離 (m))

とし，δ-spanner を用いたメカニズムの使用有無による，ϵ の値と実際のユーザ位置とダミー位置のずれ QL(m) の比較は図 7 のようになる．

ϵ はノイズ付きの位置から元の位置に関して漏洩する情報量を制御するパラメータであり，その値が小さいほどプライバシ保護度合いは高くなるが，ノイズは大きくなり，ユーザ位置に関わらず人口が多い区画が選ばれやすくなる．ϵ が大きい値のときはプライバシ保護度合いが下がり，よりユーザ位置からの距離が近い区画が選ばれる．ϵ の値が低いときはユーザが人口が多い区画にいるか，少ない区画にいるかによって QL が大きく異なるため分散が大きくなり，ϵ の値が大きくなっていくにつれ，匿名性が下がりユーザ位置付近が選ばれやすくなるため QL は小さくなる．図 7(a)(b) を比較したところ，QL の分散や中央値に大きな差異は見られず，十分な有効性を保ちながら計算時間の短縮が行えていると推測できる．次に，spanner を用いたメカニズムを使用して出力されるダミー位置の確率分布を，δ の値を変化させ詳しく調査する．ユーザ位置とプライバシ保護度合い $\epsilon = 2.0$ で固定し，spanner 未使用の場合（図 8(a)），spanner を使用し $\delta = 1.05$ としたときの結果（図 8(b)），$\delta = 1.1$ としたときの結果（図 8(c)），$\delta = 1.5$ としたときの結果（図 8(d)）を示す．今回の実装は渋谷駅周辺（図 5）における 500m 四方の分割地域メッシュ 36 区画の，2023 年 1 月 21 日 (土) の午前 11 時から正午における滞在人口数を事前確率 π_x として使用した．画内に各メッシュ区画が選ばれる確率を百分率で記載し，各メッシュ区画の色は図 3 に表した，10 段階の滞在人口数を表している．

δ の値が 1 に近いほど位置情報の簡略化は行われていないため，δ の値を 1.05, 1.1 としたときの結果（図 8(b)(c)）を見ると，ダミー位置の確率分布は spanner 未使用の場合とそれほど大きな差は無く，ユーザ位置に近い区画から少し離れた区画までが選ばれるようになっていることが分かる．δ の値を 1.5 まで大きくしたときの結果（図 8(d)）を見ると，実際の距離より 1.5 倍まで距離関数の重みが大きくなる影響で，ユーザ位置から離れた区画が選ばれにくくなっていることが分かる．

今回の比較実験を通して，δ の値を 1.1 付近に設定することで，位置情報の簡略化を行った場合も十分な有効性を保ちながら計算時間の短縮を実現でき，本手法における実装範囲の拡大に適応できることが分かった．

6　まとめと今後の課題

ユーザ位置周辺という地理的な制約条件を満たすような，SNS データからのイベント情報の検索をする際の，ユーザの位置情報プライバシを保護するためのデータ処理方法について Geo-I を用いた位置情報匿名化手法 [7] では，ユーザ位置とダミー位置のユークリッド距離の期待値が最小になるような最適化問題を解くことで Geo-I を満たすメカニズム K を得る手法を採用していた．この手法は計算量が大きいため，δ-spanner を用いた位置情報の簡略化を行うことで計算量を減らした場合の計算時間の短縮と有効性の評価を行った．実装範囲を広げた場合の比較実験では，δ の値を 1.1 付近に設定すると，位置情報の簡略化を行った場合も十分な有効性を保ちながら計算時間の短縮を実現できることが分かった．今後，spanner を用いたメカニズム K を適応させ，実装範囲の拡大を進めたい．

7　謝　　辞

本研究は一部，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Ruriko Kudo, Miki Enoki, Akihiro Nakao, Shu Yamamoto, Saneyasu Yamaguchi, and Masato Oguchi. Real-time event search corresponding to place and time using social stream. *the 3rd IEEE International Conference on Big Data Intelligence and Computing (DataCom2017)*, pages 1047–1053, 2017.

(a) spanner 未使用 (平均計算時間約 4.20 秒)

(b) spanner 使用 $\delta = 1.05$(平均計算時間約 1.81 秒)

(c) spanner 使用 $\delta = 1.1$(平均計算時間約 1.53 秒)

(d) spanner 使用 $\delta = 1.5$(平均計算時間約 0.75 秒)

図 8: $\epsilon = 2.0$ 固定, 各区画がダミー位置として選ばれる確率 (%)

[2] Miki Imai, Miki Enoki, Ruriko Kudo, and Masato Oguchi. Personalized local event search based on sns data analysis. *the 14th IEEE International Conference on Ubiquitous Information Management and Communication*, pages 6–14, 2020.

[3] X(formerly twitter). `http://twitter.com/`.

[4] Ministry of land, infrastructure, transport and tourism national open data on human flow. `https://www.geospatial.jp/ckan/dataset/mlit-1km-fromto`.

[5] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Differential privacy. *Proc. 33rd International Colloquium on Automata, Languages and Programming (ICALP' 06)*, pages 1–12, 2006.

[6] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *Proceedings of the 2013 ACM SIGSAC conference on Computer and Communications Security (CCS'13))*, pages 901–914, 2013.

[7] Kyoka Ishigami, Miki Enoki, and Masato Oguchi. Event retrieval from mobile devices by location anonymization satisfying geo-indistinguishability using real population data. *Computer Security Symposium 2023(CSS2023)*, pages 109–115, 10 2023.

[8] Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. pages 251–262, 2014.

[9] Sweeney L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[10] Tianen Liu, Yingjie Wang, Zhipeng Cai, Xiangrong Tong, Qingxian Pan, and Jindong Zhao. A dynamic privacy protection mechanism for spatiotemporal crowdsourcing. *Security and Communication Networks*, 2020, 2020.

[11] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. *Proc. IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS' 13)*, pages 429–438, 2013.

[12] Shun Takagi, Yang Cao, Yasuhito Asano, and Masatoshi Yoshikawa. Geo-graph-indistinguishability: Location privacy on road networks based on differential privacy. *CoRR*, abs/2010.13449, 2020.

[13] Twitter search api. `https://dev.twitter.com/rest/public/search`.

[14] Walker+ list of places to go in tokyo. `https://www.walkerplus.com/spot_list/ar0313/`.

[15] Navitime. `https://www.navitime.co.jp/category/`.

[16] Ministry of internal affairs and communications information bureau regional mesh statistics overview. `https://www.stat.go.jp/data/mesh/m_tuite.html`.

[17] Google maps geolocation api. `https://developers.google.com/maps/documentation/geolocation/overview`.

[18] Mobile geospatial statistics. `https://mobaku.jp`.

好奇心に基づいたユーザ適応型セレンディピティ推薦手法の提案

徐　哲林† 王　旭† 松村　敦††

† 筑波大学人間総合科学学術院人間総合科学研究群情報学学位プログラム　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s2030511@u.tsukuba.ac.jp, †wang.xu.lm@alumni.tsukuba.ac.jp, ††matsumur@slis.tsukuba.ac.jp

あらまし　精度重視によって推薦範囲が狭くなる問題の解決のために，有用性だけではなく，意外性も考慮したセレンディピティのある推薦が注目されている．しかしながら，ユーザごとに推薦に求める有用性と意外性の割合が異なるため，セレンディピティのある推薦の実現が難しい．そこで本研究では，好奇心に基づいてユーザごとの有用性と意外性を求める割合を推定し，その傾向にあわせた推薦を行う手法を提案する．本手法では，ユーザの好奇心を長期的な興味と短期的な興味に基づいて推定する．MovieLens-1m を用いてオフライン実験を行い，提案手法の有効性を検証した．実験の結果，提案手法による推薦は既存手法と比較し，同程度の精度が得られるため，セレンディピティのある推薦が達成した．

キーワード　セレンディピティ，情報推薦，好奇心

1　はじめに

インターネットの発展によって，Web 上のデータは爆発的に増加している．それに伴い，ユーザが Web 上で得る情報の選択肢が豊富になる一方で，価値のある情報を見つけるためのユーザの意思決定までの時間は長くなった．その結果，ユーザ体験に悪影響を及ぼしていることが指摘されている [1]．この問題を解決するため，1990 年代中期以降，データマイニング技術を応用した推薦システムが登場した [2]．以降，現在までに多くの推薦システムが提案され，ユーザに価値のある情報を提供することが試みられている．

これまでの推薦システムの研究では，ユーザが自らの嗜好を正確に予測できるシステムを好むことを前提条件とし，この条件を満たすことで，ユーザの満足度を向上させることを目指してきた [3]．したがって，伝統的な推薦システムでは，予測精度を重視しているものが多い [4]．しかしながら，これらの推薦システムは，ユーザのプロファイル（行動履歴や嗜好など）と類似しているアイテムしか推薦しないという傾向がある [5]．これに伴い，filter bubble [6] や over-specialization 問題 [7] が深刻化し，ユーザが自らのプロファイルに関連しないアイテムに触れることは難しくなっている [8]．その結果，興味を広げようとするユーザの満足度は低くなってしまっている．

このような状況に対して，推薦システムの構築において，単に予測精度を重視するだけでは不十分であり，他の評価指標によって評価する必要があると指摘されている [9]．現在，推薦システムに対して予測精度以外の評価指標，特にセレンディピティ (*Serendipity*) が注目されている [10]．セレンディピティとは，ユーザへの推薦に対して，関心がある（有用性）だけではなく，意外性という要素が加わった評価指標である [11]．セレンディピティのある推薦は，人々を自分の興味に合うもののみに囲まれる空間から解放し，今まで見つからなかったアイテム，または，予想外の新しいアイテムとの遭遇を増やすことができる．そのため，予測精度を重視した場合に発生する問題を改善でき，ユーザの満足度を向上させられることが指摘されている [12], [13]．

以上に述べたように，セレンディピティのある推薦は，ユーザにとって有用性だけでなく意外性もある必要がある．このようなセレンディピティのある推薦を行うために，一部の研究は，reranking algorithms に基づく推薦手法を提案している．Reranking algorithms とは，予測精度を重視するアルゴリズムにセレンディピティのあるアイテムを推薦するためのアルゴリズムを加えたものである [14]．具体的には，まず，予測精度を重視する推薦手法を用いて各アイテムの有用性の予測値を算出し，ランキングリストを作成する．その後，アイテムごとの意外性の予測値を算出し，有用性の予測値に加算することで，セレンディピティの予測値を求める．最後に，セレンディピティの予測値に基づいて，リストを再度ランキングし，上位のアイテムをセレンディピティのあるアイテムとしてユーザに推薦する．

また，推薦候補のアイテムに関する有用性と意外性の重要度が異なると想定し，2 つの予測値を加算する際に重み付けをする先行研究もある．例えば，Li ら [15] は，ユーザが意外性のある推薦アイテムを受け入れる程度を重みとして，アイテムの意外性の予測値に付与した．しかしながら，この手法では各アイテムの意外性の重要度は考慮するが，有用性の重要度は考慮されていない．その結果，意外性の高いアイテム，すなわち，新奇な情報を積極的に探索するユーザに自身の興味に関連する情報を推薦してしまう場合もある．これに対して，Kotkov ら [16] は，有用性の予測値と意外性の予測値の両方に重みを付与することを試みてはいるが，全てのユーザに対して事前に決定した同じ重みを使う．しかしながら，ユーザごとに推薦に求める有用性と意外性の割合は異なると考えられる．つまり，知識を求める傾向には個人差があるため，全く新しい知見を求めるユーザもいれば，自身の興味と全く同じものを望むユーザもいて，

様々なレベルのユーザが存在する．そのため，ユーザごとのセレンディピティ体験を向上させるためには，各推薦候補のアイテムに関する有用性と意外性の予測値に重みを動的に付与することによって，ユーザの知識を求める傾向に応じる推薦を提供する必要があると考えられる．

ここで，ユーザの知識を求める傾向はユーザの好奇心と関わりがあるとされている [17]．すなわち，好奇心の高い人は新奇な情報を受け入れやすく，好奇心の低い人は自身の興味に関連する情報を求めている．このことから，ユーザ自身の好奇心を重みとして，各推薦候補のアイテムに関する有用性と意外性の予測値に付与することで，ユーザの知識を求める傾向に応じた推薦が実現できる可能性がある．これによって，ユーザの好奇心が高いなら，自身の興味と遠いアイテムを推薦し，ユーザの好奇心が低いなら，自身の興味に近いアイテムを推薦することができるようになる．

これまで，一部のセレンディピティのある推薦に関する研究は，ユーザの好奇心の高さを推定するために心理学分野で開発された Curiosity and Exploration Inventory–II（CEI-II）を利用している [18], [19]．しかしながら，CEI-II はアンケート分析手法であるため，ユーザが複数の質問項目に回答する必要があり，大量のユーザの好奇心を自動的に推定することができないという問題が発生する．

以上を踏まえて，本研究では，推薦候補のアイテムごとの有用性と意外性の予測値を算出したうえで，自動的に推定されたユーザそれぞれの好奇心を考慮することによって，ユーザ適応型のセレンディピティのある推薦を行う手法を提案し，セレンディピティのある推薦を実現することを目的とする．

2 関 連 研 究

2.1 セレンディピティのある推薦

予測精度を重視する推薦手法による filter bubble などの問題を改善するために，多くのセレンディピティのある推薦手法が提案されている．一部の研究は，予想外の新しい情報との出会いを狙うために，意外性の高いアイテムをユーザに推薦している [20]．しかしながら，このようなアイテムはユーザ自身の嗜好と大幅に異なるため，その多くはユーザが全く知らないアイテムとなってしまう．その結果，推薦されたアイテムはユーザの固有知識の範囲を大きく超えた関心のないものとなり，セレンディピティのある推薦を実現できない [19]．

また，前節で述べた reranking algorithms に基づく推薦手法も多く提案されている．例えば，推薦候補のアイテムに対して，意外性と有用性の予測値を算出し加算することで，セレンディピティの予測値を求める手法が提案されている [11]．また，各アイテムの意外性と有用性に関する重要度を考慮した上で，その重要度を重みとして付与する手法も提案されている [15], [16]．しかしながら，これらの方法はユーザ自身の好奇心を考慮しないため，ユーザごとの知識を求める傾向に対応できない．その結果，セレンディピティのある推薦がうまく行えないと考えられる．

これに対して，ユーザ自身の好奇心を考慮したセレンディピティのある推薦手法 [18], [19] が提案されている．しかしながら，これらの手法はアンケート調査によるユーザからのフィードバックを使うため，自動的に好奇心の高さを推定できないという問題が残っている．

2.2 本研究の位置づけ

本研究では，ユーザ自身の好奇心に基づいた適応型セレンディピティ推薦を行う手法を提案している．Reranking algorithms に基づいく既存手法と比べて，まず，本研究でも既存手法のように推薦候補のアイテムに関する有用性と意外性の予測値をそれぞれ算出する．次に，各ユーザ自身の好奇心を重みとして 2 つの要素に付与することによってユーザの知識を求める傾向に対応できる点で大きく異なる．また，ユーザ自身の好奇心の推定については，本研究では，ユーザの暗黙的なフィードバック（例えば，クリックや共有など）[21] に基づいて自動的好奇心を推定することは従来の好奇心を推定する手法と異なる点であると考えられる．

3 提 案 手 法

本章では，好奇心に基づいたユーザ適応型セレンディピティ推薦手法を説明する．まず，3.1 節では，本研究で着目した好奇心について述べる．次に，3.2 節では，ユーザの拡散的好奇心の高さを推定する手法を述べる．その後，3.3 節では，各推薦候補のアイテムに関する有用性と意外性の予測値を算出する手法を説明する．最後に，3.4 節では，アイテムのセレンディピティに関する予測値を算出した上で，これらの予測値に基づくセレンディピティのある推薦を行う手法について述べる．

3.1 好 奇 心

ユーザが新しく，かつ，意外性のあるものを探索することにどの程度喜びを感じるかは，好奇心が大きく影響することが指摘されている [17]．例えば，複数人のユーザに同じ意外性のありそうなものを推薦しても，各ユーザの好奇心の強さが異なるため，それを得た時の喜びの程度も異なってくる．つまり，ユーザが新しい知識を求めているかどうかは，ユーザの好奇心によって判断できると考えられる．このことから，本提案手法では，推薦候補のアイテムに関する有用性と意外性の重要度をユーザ自身の好奇心によって調整することで，ユーザの知識を求める傾向に対応できる適切な推薦を提供する．これにより，セレンディピティ体験の向上につなげる．

心理学分野において好奇心は個人特性であり，大きく分けると，知的好奇心（*epistemic curiosity*）と知覚好奇心（*perceptual curiosity*）の 2 種類がある [22]．知的好奇心とは知的活動を動機づけることである．また，知覚好奇心とは知識や情報ではなく，新奇な視覚刺激や聴覚刺激に対する動機づけることである．本研究では，映画などの商品を推薦の対象となるため，知的好奇心に着目する．

知的好奇心をさらに分類すると，拡散的好奇心（*diversive curiosity*）と特殊的好奇心（*specific curiosity*）

の 2 種類がある．拡散的好奇心とは方向性を定めずに新奇な情報や知識を幅広く求める探索行動である．具体的な例 [23] としては，図 1 に示すように，あるユーザの読書する際の行動において，従来から常に青春小説を読むが，刺激に乏しい状況で退屈して新奇な対象を求め，たまたま手にした他の種類の小説（例えば，推理小説や歴史小説）を読もうとする行動である．一方で，特殊的好奇心とは矛盾あるいは情報の不整合に対して，方向性を定めて探索行動を行うことを動機づけることである．

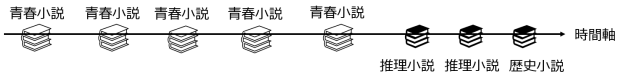


図 1　拡散的好奇心の具体事例

1 節に述べた通り，セレンディピティのある推薦はユーザを自分の興味に合うものだけに囲まれる空間から解放することを目標とするため，本研究において，ユーザの拡散的好奇心に着目する．その後，各ユーザの拡散的好奇心の高さを推定した上で，拡散的好奇心が高いユーザに，新奇な情報，すなわち，意外性の高いアイテムを推薦する．一方で，拡散的好奇心が低いユーザに，自身の興味に近い情報，すなわち，意外性の低いアイテムを推薦する．これに伴い，推薦されるアイテムはユーザの知識を求める傾向に対応できると考えられる．

3.2　拡散的好奇心の高さの推定

3.1 節に述べたどおり，拡散的好奇心とは，明確な方向性はもたずに幅広く新奇な情報を探し求める動作である．この定義に基づいて，ユーザが情報を探索する際に，これらの情報に関する新規性（新奇な情報であるか）と多様性（幅広い情報であるか）を測ることによって，ユーザ自身の拡散的好奇心の高さを推定できると考えられる．

ここで，ユーザの興味は大きく分けると，短期的な興味と長期的な興味があると考える．短期的な興味とは最近ユーザがやっていることである．一方で，長期的興味とはユーザが従来からやってきたことである．これら長期的な興味と短期的な興味との差分によって，ユーザが求めている情報の新規性を判断できると考えられる．具体的には，ユーザの長期的な興味と短期的な興味が大幅に異なるなら，常にやることに対して，刺激に乏しい状況を改善するために，ユーザが積極的に新奇な情報を探索していると考え，このユーザは新規性が高い情報を探していると判断する．また，短期的な興味の多様性を測ることでユーザが幅広く情報を探索するかを判断できる．

図 1 は，あるユーザが読んできた本を時系列で並べたものである．ここにに示したユーザの場合，最近読んでいる「推理小説」と「歴史小説」が短期的な興味となり，従来から読んでいた「青春小説」は長期的な興味となる．長期的な興味と短期的な興味が大きく異なっているため，新規性を求める傾向が強いと判断できる．また，短期的興味が複数に渡っていることから，多様性もあると判断でき，拡散的好奇心が強いユーザであることが推定できる．

本研究では，ユーザが求めている情報の新規性と多様性は同じ重要度を持つと想定し，式 1 によって，ユーザの拡散的好奇心の高さを測る手法を提案する．

$$Curosity_u = \frac{Diff_u^{long-short} + Div_u^{short}}{2} \qquad (1)$$

ここで，$Diff_u^{long-short}$ はユーザ u の長期的な興味と短期的な興味の差分，すなわち，ユーザが探索する情報の新規性に関する予測値である．また，Div_u^{short} はユーザ u の短期的な興味の多様性である．以下では,「長期的な興味と短期的な興味の差分の計算」と「短期的な興味の多様性の判定」について詳細を説明する．

3.2.1　長期的な興味と短期的な興味の差分の計算

長期的な興味と短期的な興味は，時間情報を考慮して推定するユーザの興味であるため，以下の手順でこれを推定する．まず，各ユーザの行動履歴から，タイムスタンプによって図 1 に示すような時系列データを作成する．ここで，長期的な興味はユーザが従来から常にやることであるため，あまり変わらない性質を持つことが指摘されている [24]．したがって，図 2 に示すように，緑枠で囲まれるこのユーザの全行動履歴を長期的な興味を推定するために利用する．本研究では，伝統的な手法である singular value decomposition（SVD）[25] を利用して，ユーザの長期的な興味を推定する．

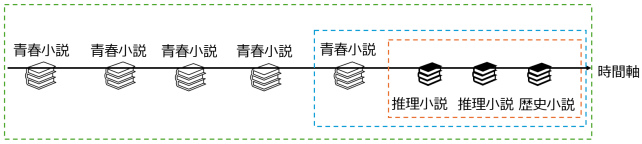


図 2　ユーザの行動履歴に基づく作成された時系列データ

一方で，短期的な興味は最近にやっていることであるため，一部の最新の行動履歴だけを利用して，推定できることが指摘されている [26]．ただし，短期的な興味を推定するために，どの程度の割合の行動履歴を利用すべきかは先行研究の中でも統一されていない．そのため，例えば図 2 に示す青枠の 4 つの履歴を利用する手法もオレンジ枠の 3 つの履歴を利用する手法も可能であり，それによって結果が変わる．本研究では，短期的な興味を推定するために利用する行動履歴の割合をパラメータとして扱う．

さらに，短期的な興味を推定するための手法は大きく分けると，伝統的な手法であるマルコフ連鎖とディープラーニング手法である recurrent neural networks（RNN）の 2 種類がある．マルコフ連鎖手法より，RNN 手法はユーザの行動履歴の時系列を考慮するため，興味の動的変化や購買動機を把握できることが指摘されている [27]．したがって，本研究では，Yu ら [24] が提案した Long Short Term Memory（LSTM）の改善モデルを利用してユーザの短期的な興味を推定する．

LSTM は RNN の一種であり，Yu らは以下の 2 つの改善を行っている．

（1） **ユーザの行動時間の不規則性を考慮する．**例えば，あるユーザが 2023 年 12 月 1 日に商品 A，2023 年 12 月 2 日に商品 B，2024 年 1 月 10 日に商品 C を購入した場合，商品 A と B は購入日が近いため関連性が高いと想定できる．そのため，短期間で連続購入した商品の関連性が高くなるための仕組みを導入する．

（2） **ユーザの行動履歴の性質を考慮する．**例えば，あるユーザの購買履歴が {iphone, airpods, 傘, 鉛筆 } であった場合，これらの商品はそれぞれのカテゴリに属している．ユーザが次に購入する IT 製品を予測する際には，IT 製品のカテゴリに属している iphone と airpods の情報だけが有用である．したがって，行動履歴の性質を考慮することで，次回の行動を予測する際に，役に立つ行動履歴の重要度が高くなる仕組みを導入する．

以上に述べた 2 つの手法でユーザの長期的な興味と短期的な興味をそれぞれ 20 個を推定し，長期的興味集合と短期的な興味集合を作成する．ここで，各集合にある興味は 80 次元のベクトルで表現する．文章間の類似度を測る際に，文章を表すベクトルは，文章にある全ての単語ベクトルの平均値で表現できることが指摘されている [28]．これによって，文章ベクトル間の類似度に基づいて，文章間の類似度を測ることができると考えられる．そのため，各興味集合を文章と見なした上で，式 2 によってユーザの長期的な興味集合を表すベクトルを算出する．

$$Vector_u^{long} = \frac{\sum_{i \in \boldsymbol{LP^u}} vector_u^i}{20} \tag{2}$$

ここで，LP^u はユーザ u の長期的な興味集合，$vector_u^i$ はユーザ u の長期的な興味集合にある興味を表すベクトルである．

同様に，式 3 によってユーザの短期的な興味集合を表すベクトルを算出する．

$$Vector_u^{short} = \frac{\sum_{j \in \boldsymbol{SP^u}} vector_u^j}{20} \tag{3}$$

ここで，SP^u はユーザ u の短期的な興味集合である．$vector_u^j$ はユーザ u の短期的な興味集合にある興味を表すベクトルである．

最後に，2 つのベクトルを正規化した上で，式 4 によってユークリッド距離に基づいてユーザの長期的な興味と短期的な興味の差分を算出する．

$$Diff_u^{long-short} = \sqrt{\sum_k^{80} (Vector_k^{long} - Vector_k^{short})} \tag{4}$$

ここで，$Vector_k^{long}$ は式 2 によって求められるユーザ u の長期的な興味集合を表すベクトルの k 次元の数値である．同様に，$Vector_k^{short}$ は式 3 によって求められるユーザ u の短期的な興味を表すベクトルの k 次元の数値である．

3.2.2 短期的な興味の多様性の判定

本研究では，式 5 を利用し，Intra-List 類似度 [29] に基づいて短期的な興味の多様性を算出する．

$$Div_u^{short} = 1 - \overline{Intra - List\ Similarity} \tag{5}$$

ここで，$\overline{Intra - List\ Similarity}$ は式 6 によって算出する．

$$\overline{Intra - List\ Similarity} = \frac{1}{|S|} \sum_{u \in S} \sum_{\substack{i,j \in SP_{u,20}, \\ j<i}} CosSim(i,j) \tag{6}$$

ここで，S は全ユーザ数である．SP_u はユーザ u の短期的な興味集合であり，20 個の興味を含む．i と j はユーザ u のそれぞれの短期的な興味であり，これらの興味の類似度 $CosSim(i,j)$ を式 7 によって算出する．

$$CosSim(i,j) = \frac{\#\text{users who like both } i \text{ and } j}{\sqrt{\#\text{ prefs in } i} \times \sqrt{\#\text{ prefs in } j}} \tag{7}$$

3.3 アイテムに関する有用性と意外性の予測値の算出

推薦候補のアイテムに関する有用性と意外性の予測値は Li ら [15] が提案した手法をそのまま利用して，算出する．

有用性の予測値の算出手法　有用性の予測値とは推薦されたアイテムに対して，ユーザがどのぐらい関心があるかを予測する数値である．式 8 によって推薦候補のアイテムに関する有用性の予測値を算出する．

$$useful_u^i = MLP(R_u; E_u; E_i) \tag{8}$$

ここで，E_u と E_i はユーザ u と推薦候補のアイテム i を埋め込んだ上で，求められたそれぞれのベクトルである．R_u はユーザ u の行動履歴を埋め込んで，求められたベクトルである．ここで，双方向 Gated Recurrent Unit ニューラルネットワーク（Bidirectional GRU）と self-attentive メカニズムを組み合わせるモデルを利用し，ユーザの行動履歴を埋め込む．双方向 GRU も RNN の一種であり，ユーザの行動履歴の時間情報と行動の順序を考慮する．通常の GRU に比較し，過去の行動を学習し，未来の行動を予測するだけではなく，未来の行動を学習し，過去の行動も予測する．これによって，学習済モデルを利用して予測を行う際に，前後情報を同時に考慮できるようになる．また，GRU は LSTM と比べて，一個のゲートを削除するため，構造はシンプル化でき，計算量も減らすことができる．さらに，self-attentive メカニズムを組み合わせことによって，各行動の重みを付与するため，ユーザの長期的な興味と短期的な興味を同時に把握することができる [30]．3 つの要素の埋め込みを行った上で，MLP に入力することで推薦候補のアイテム i の有用性の予測値を算出する．

また，3.2 節において，ユーザの短期的な興味を推定する際に，本研究では双方向 GRU と self-attentive メカニズムを組み合わせるモデルを利用しない．その理由は以下の 2 つがある．

（1） Yu ら [24] が使った実世界データセットに基づいて評価を行った結果，GRU より LSTM の性能が高かった

（2） Self-attentive メカニズムを利用しても，ユーザの短期的な興味を高精度で推定できない [30]

以上の理由から，ユーザの短期的な興味を推定する際は，GRU と self-attentive メカニズムを組み合わせる手法ではなく，より性能が良い LSTM の改善モデルを利用する．

意外性の予測値の算出手法　意外性の予測値とは推薦され

たアイテムに対して，ユーザがどのぐらい意外性を感じるかを予測する数値である．特徴空間におけるアイテムとユーザの行動履歴との距離に基づいて，意外性の予測値をより正確に算出できる [31]．その理由は推薦候補のアイテムとユーザの行動履歴を特徴空間，すなわち，潜在的な空間（*latent space*）に埋め込むと，アイテムとユーザの行動履歴との潜在的，かつ，複雑な関係を把握できるからである [32]．以上に述べた推薦候補のアイテムの有用性の予測値を算出する手法を利用することで，アイテムとユーザの行動履歴をそれぞれ潜在的な空間に埋め込んで，ベクトルに変換できる．そのため，これらのベクトルを利用し，ユーザ u への推薦候補のアイテム i に対して，式 9 によって意外性の予測値を算出する．

$$unexp_u^i = \sum_{k=1}^{N} d(w, C_k) \times \frac{|C_k|}{\sum_{k=1}^{N} |C_k|} \tag{9}$$

ここで，w はアイテム i を潜在的な空間に埋め込んだ後のベクトルである．C_k はユーザの行動履歴を埋め込んだ後，mean shift 法でクラスタリングを行い，求められた各クラスタである．$|C_k|$ は各クラスタの中で，行動履歴であるアイテムの数である．$d(w, C_k)$ は潜在的な空間において，推薦候補のアイテムとクラスタとのユークリッド距離である．

3.4 アイテムに関するセレンディピティの予測値の算出

3.1 節に述べたとおり，拡散的好奇心が高いユーザは知識を求める傾向も高いため，ユーザの興味と大幅に異なり，意外性の高いアイテムを推薦する．一方で，拡散的好奇心が低いユーザは知識を求める傾向も低いため，ユーザの興味に近い，少し意外性のあるアイテムを推薦する．このような推薦を実現するために，ユーザの拡散的好奇心を重みとしてアイテムに関する有用性と意外性の両方に付与する．これに伴い，式 10 によって推薦候補のアイテムに関するセレンディピティの予測値を算出する．

$$\begin{aligned} serendipity_u^i = (1 - curiosity_u) \times useful_u^i \\ + curiosity_u \times unexp_u^i \end{aligned} \tag{10}$$

4 評価実験

本研究の提案手法の有効性を評価するために，評価実験を実施した．

4.1 評価用データセット

本研究では，MovieLens-1m [33] を利用し，提案手法の有効性を検証した．MovieLens-1m の基本情報は表 1 に示している．このデータセットは情報推薦分野において，よく使われる公開された実世界データセットであり，映画名やユーザの評価点数などの情報が含まれている．

また，leave-last-out データ分割方法 [30] を利用し，データセットにあるユーザの行動履歴に基づいて訓練データ，検証データ，テストデータを作成した．例えば，一人のユーザの行動履歴は N 個のアイテムがある．これらのアイテムをタイムスタンプに基づいて時系列データにした上で，最新の 1 個をテストデータ，次の 1 個を検証データ，残った N-2 個を全て訓練データに入れる．ここで，検証データとテストデータに一個のポジティブサンプルに対して，9 個と 49 個のネガティブサンプルをそれぞれランダムで生成する．

表 1　MovieLens-1m の基本情報

Number of Users	Number of Items	Number of Ratings	Rating Scale	Release Time
6040	3952	1,000,209	1-5	2/2003

訓練データにあるユーザの行動履歴を利用してモデルの学習を行う．その後，学習済モデルを利用し，テストデータにあるアイテムごとの有用性と意外性の予測値を算出する．その上で，算出されたユーザごとの拡散的好奇心を重みとして 2 つの要素に付与することで，ポジティブサンプルを予測できるかを検証する．

4.2 比較手法

本研究では，PURS モデル [15] と比較する．PURS は reranking algorithms に基づく推薦手法であり，推薦候補のアイテムに関する有用性と意外性の予測値を算出した上で，ユーザが意外性のあるアイテムを受け入れる程度を重みとして意外性の予測値に付与する手法である．

4.3 評価指標

4.1 節に述べたように，MovieLens-1m を利用し，提案手法のセレンディピティに関する性能評価を行う．しかしながら，このデータセットは，ユーザの嗜好を予測できるかどうかを検証できるが，セレンディピティに関する情報がないため，セレンディピティを間接的に評価する必要がある．具体的には，セレンディピティの定義によって，有用性と意外性の 2 つの要素があるため，モデルによって生成された推薦アイテムに対して有用性と意外性をそれぞれ客観的に評価する．precision@N と recall@N を利用し，推薦アイテムの有用性を評価する．また，式 9 によって推薦アイテムの意外性を評価する．

4.4 ユーザの短期的な興味を推定するための行動履歴

3.2 節に述べたどおり，ユーザの短期的な興味を推定するための行動履歴の割合が統一されていない．そのため，本実験では，最新の 30%の行動履歴に基づいてユーザの短期的な興味を推定する．

4.5 実験結果と考察

推薦の有用性に関する評価結果を表 2 に示す．これらの結果から，PURS と比べて，precision や recall の評価値が低くなるため，提案手法を用いて有用性が低下することがわかった．一方で，推薦の意外性に関する評価結果を表 3 に示す．これらの結果に基づいて，推薦の意外性を少し向上することがわかった．提案手法により十分な性能が得られない原因は，短期的な興味を推定するための行動履歴の割合は固定しているため，ユーザ自身の拡散的好奇心をうまく推定できなかったからであると考

表 2　推薦の有用性に関する評価結果

Model	precision@5	precision@10	precision@15	precision@20	recall@5	recall@10	recall@15	recall@20
Proposed method	0.7500	0.7151	0.6886	0.6664	0.0762	0.1405	0.1970	0.2463
PURS	**0.7545**	**0.7185**	**0.6919**	**0.6683**	**0.0766**	**0.1409**	**0.1978**	**0.2465**

えている．

本研究では，ユーザの短期的な興味と長期的な興味を利用し，ユーザ自身の拡散的好奇心を推定する．短期的な興味を推定するための行動履歴の割合が統一されていなく，本研究においてユーザの最新の 30%の行動履歴に基づいて短期的な興味を推定する．しかしながら，使用した行動履歴の割合が変わると，ユーザの短期的な興味を変更する可能性があるため，ユーザの拡散的好奇心の高さに影響を与えると推測する．これを確認するために，30%以外に，最新のユーザの行動履歴の割合を {5%, 10%, 15%, 20%, 25%} に設定し，拡散的好奇心を推定した．それぞれのユーザ行動履歴の割合に基づいて推定された拡散的好奇心は図 3 に示す．これらの結果に基づいて，ユーザの最新の行動履歴（current session）の割合が変わると，ユーザの拡散的好奇心の分散図も変わるため，一部のユーザの拡散的好奇心の高さも変わることがわかった．また，行動履歴の割合が増えると，拡散的好奇心が低くなったユーザ数が多くなることも明らかにした．したがって，セレンディピティのある推薦をより多く提供するために，各ユーザの適切な行動履歴を利用して拡散的好奇心の高さを推定する手法を検討する必要があると考えている．

5　まとめと今後

伝統的な推薦システムは予測精度を重視し，ユーザの嗜好に合うものを推薦することを目標として設定するため，filter bubble などの問題が発生した．この状況を改善するために，好奇心に基づいたユーザ適応型セレンディピティのある推薦手法を提案した．具体的には，まずユーザの長期的な興味と短期的な興味に基づいてユーザの拡散的好奇心を自動的に推定する．その上で，推定された拡散的好奇心を重みとして推薦候補のアイテムに関する有用性と意外性に付与した上で，セレンディピティのあるアイテムを推定する．

実世界データセットである movielens-1m に基づいて評価実験を実施した．その結果，提案手法は既存手法に比較して，同程度の精度が得られるため，セレンディピティのある推薦が達成したと言えた．今後の課題は，ユーザの拡散的好奇心をより正確に推定する手法を検討することである．

文　献

[1] F. Ricci, L. Rokach, and B. Shapira, “Recommender systems: Techniques, applications, and challenges,” *Recommender Systems Handbook*, pp. 1–35, 2021.

[2] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, “An algorithmic framework for performing collaborative filtering,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 230–237.

[3] R. J. Ziarani and R. Ravanmehr, “Serendipity in recommender systems: a systematic literature review,” *Journal of Computer Science and Technology*, vol. 36, pp. 375–396, 2021.

[4] B. Alhijawi, A. Awajan, and S. Fraihat, “Survey on the objectives of recommender systems: measures, solutions, evaluation methodology, and new perspectives,” *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–38, 2022.

[5] A. M. Silva, F. H. da Silva Costa, A. K. R. Diaz, and S. M. Peres, “Exploring coclustering for serendipity improvement in content-based recommendation,” in *Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, November 21–23*. Springer, 2018, pp. 317–327.

[6] G. M. Lunardi, G. M. Machado, V. Maran, and J. P. M. de Oliveira, “A metric for filter bubble measurement in recommender algorithms considering the news domain,” *Applied Soft Computing*, vol. 97, pp. 1–12, 2020.

[7] M. Manca, L. Boratto, and S. Carta, “Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system,” *Information Systems Frontiers*, vol. 20, pp. 825–839, 2018.

[8] V. Maccatrozzo, “Burst the filter bubble: using semantic web to enable serendipity,” in *International Semantic Web Conference*. Springer, 2012, pp. 391–398.

[9] S. M. McNee, J. Riedl, and J. A. Konstan, “Being accurate is not enough: how accuracy metrics have hurt recommender systems,” in *CHI’06 extended abstracts on Human factors in computing systems*, 2006, pp. 1097–1101.

[10] D. Kotkov, A. Medlar, and D. Glowacka, “Rethinking serendipity in recommender systems,” in *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 2023, pp. 383–387.

[11] X. Peng, H. Zhang, X. Zhou, S. Wang, X. Sun, and Q. Wang, “Chestnut: Improve serendipity in movie recommendation by an information theory-based collaborative filtering approach,” in *Human Interface and the Management of Information. Interacting with Information, Copenhagen, Denmark, July 19–24*. Springer, 2020, pp. 78–95.

[12] L. Chen, Y. Yang, N. Wang, K. Yang, and Q. Yuan, “How serendipity improves user satisfaction with recommendations? a large-scale user evaluation,” in *The world wide web conference*, 2019, pp. 240–250.

[13] Z. Fu, X. Niu, and M. L. Maher, “Deep learning models for serendipity recommendations: A survey and new perspectives,” *ACM Computing Surveys*, pp. 1–26, 2023.

[14] D. Kotkov, S. Wang, and J. Veijalainen, “A survey of serendipity in recommender systems,” *Knowledge-Based Systems*, vol. 111, pp. 180–192, 2016.

[15] P. Li, M. Que, Z. Jiang, Y. Hu, and A. Tuzhilin, “Purs: personalized unexpected recommender system for improving user satisfaction,” in *Proceedings of the 14th ACM Conference on Recommender Systems*, 2020, pp. 279–288.

[16] D. Kotkov, J. Veijalainen, and S. Wang, “How does serendipity affect diversity in recommender systems? a serendipity-oriented greedy algorithm,” *Computing*, vol. 102, pp. 393–411, 2020.

[17] N. Wang, L. Chen, and Y. Yang, “The impacts of item features and user characteristics on users’ perceived serendipity of recommendations,” in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*,

表 3　推薦の意外性に関する評価結果

model	unexp@5	unexp@10	unexp@15	unexp@20
Proposed method	**0.1839**	**0.1820**	**0.1807**	**0.1799**
PURS	0.1815	0.1803	0.1793	0.1785

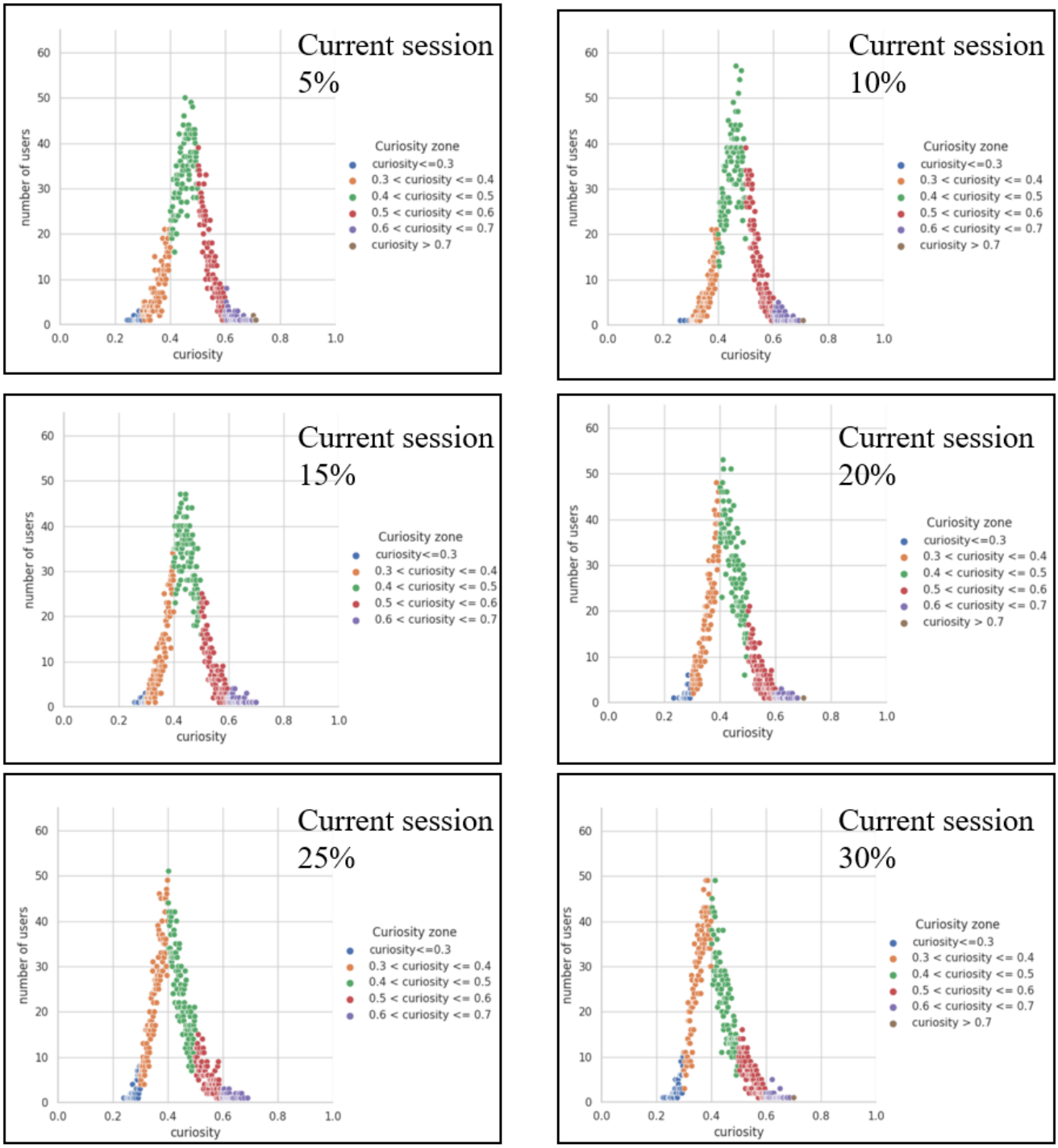


図 3　ユーザの拡散的好奇心の分散図

2020, pp. 266–274.

[18] A. Menk, L. Sebastia, and R. Ferreira, "Curumim: a serendipitous recommender system based on human curiosity," *Procedia computer science*, vol. 112, pp. 484–493, 2017.

[19] V. Maccatrozzo, M. Terstall, L. Aroyo, and G. Schreiber, "Sirup: Serendipity in recommendations via user perceptions," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017, pp. 35–44.

[20] X. Li, W. Jiang, W. Chen, J. Wu, G. Wang, and K. Li, "Directional and explainable serendipity recommendation," in *The world wide Web Conference*, 2020, pp. 122–132.

[21] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," in *Acm Sigir Forum*, vol. 37, no. 2. ACM New York, NY, USA, 2003, pp. 18–28.

[22] D. E. Berlyne, “Conflict, arousal, and curiosity.” 1960.
[23] 西川一二, 雨宮俊彦, “知的好奇心尺度の作成—拡散的好奇心と特殊的好奇心—”, 教育心理学研究, vol. 63, no. 4, pp. 412–425, 2015.
[24] Z. Yu, J. Lian, A. Mahmoody, G. Liu, and X. Xie, “Adaptive user modeling with long and short-term preferences for personalized recommendation.” in *IJCAI*, 2019, pp. 4213–4219.
[25] D. Yang, T. Chen, W. Zhang, and Y. Yu, “Collaborative filtering with short term preferences mining,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1043–1044.
[26] G. Zhu, Y. Wang, J. Cao, Z. Bu, S. Yang, W. Liang, and J. Liu, “Neural attentive travel package recommendation via exploiting long-term and short-term behaviors,” *Knowledge-Based Systems*, vol. 211, p. 106511, 2021.
[27] C. Ma, L. Ma, Y. Zhang, J. Sun, X. Liu, and M. Coates, “Memory augmented graph neural networks for sequential recommendation,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5045–5052.
[28] M. Farouk, “Measuring sentences similarity: a survey,” *arXiv preprint arXiv:1910.03940*, 2019.
[29] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor, “Auralist: introducing serendipity into music recommendation,” in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 13–22.
[30] Z. He, H. Zhao, Z. Lin, Z. Wang, A. Kale, and J. McAuley, “Locker: Locally constrained self-attentive sequential recommendation,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 3088–3092.
[31] P. Adamopoulos and A. Tuzhilin, “On unexpectedness in recommender systems: Or how to better expect the unexpected,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 4, pp. 1–32, 2014.
[32] P. Li and A. Tuzhilin, “Latent unexpected recommendations,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 6, pp. 1–25, 2020.
[33] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *Acm transactions on interactive intelligent systems*, vol. 5, no. 4, pp. 1–19, 2015.

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 16:00 ~ 18:10 | T4-A オンライン（Zoom Events)

AIとの対話

座長:灘本 明代(甲南大学)
コメンテータ:山本 祐輔(静岡大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T4-A-6-01] AIは「思いやり」を持った反応ができるか？

*川内野 穂香[1,3]、牛尼 剛聡[2,3] (1. 長崎県立長崎東高等学校、2. 九州大学 大学院芸術工学研究院、3. 九州大学 未来創生科学者育成プロジェクト)

16:20 ~ 16:35
[T4-A-6-02] タスク遂行促進のための作業リアクションの自動生成手法の有効性評価

*堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

16:35 ~ 17:00
[T4-A-6-03] 小説の感想を語り合えるエージェント構築のためのレビューをもとにした小説解釈の再現

*笹原 雄太[1]、牛尼 剛聡[1] (1. 九州大学)

17:00 ~ 17:25
[T4-A-6-04] 対人スキルと対AIスキルの関係

*西岡 竜生[1]、若宮 翔子[1]、清水 伸幸[2]、藤田 澄男[2]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. LINEヤフー株式会社)

AIは「思いやり」を持った反応ができるか？

川内野 穂香†§　　牛尼 剛聡‡§

†長崎県立 長崎東高等学校 〒850-0007 長崎県長崎市立山 5-13-1
‡九州大学 大学院芸術工学研究院 〒815-8540 福岡市南区塩原 4-9-1
§九州大学 未来創生科学者育成プロジェクト 〒819-0395 福岡市西区元岡 744
E-mail: †honopy5588@gmail.com, ‡ushiama@design.kyushu-u.ac.jp

あらまし　近年、ChatGPT や Gemini などの高性能な対話型生成 AI が多くの注目を集め、多様なサービスでの応用が期待されている。本研究では、対話型生成 AI による悩み相談の応用に焦点を当てる。一部の対話型生成 AI は、正しい回答を提供する一方で、悩みを抱える質問者に対して適切でないと考えられる回答をすることがある。本論文では、ユーザに対して思いやりのある回答を生成する「内省的質問応答法」を提案する。この手法では、AI が生成した初期の回答を、質問者の視点に立って評価し、その後の回答を修正する。提案手法の効果を検証するために行った被験者実験では、有意水準 1%で、「思いやり」、「親しみ」、「受容性」という 3 つの指標において提案手法が 3 種類の比較手法よりも優れていることが示された。

キーワード　生成 AI、巨大言語モデル、メンタルヘルス、思いやり

1. はじめに

近年、人工知能（Artificial Intelligence: AI）技術は目覚ましい発展を遂げ、日常生活、ビジネス、教育、学術研究等、様々な分野において重要な役割を果たすようになってきている。近年の飛躍的な AI 技術の最も重要な側面の一つに、深層ニューラルネットワークを利用した自然言語処理がある。2022 年 11 月に OpenAI が公開した ChatGPT[1]は、Transformer[2]に基づいた代表的なニューラル言語モデルである GPT に対して、高度な会話を可能とするアライメントを行った代表的な巨大言語モデル（Large Language Model: LLM）である。ChatGPT は人間と AI の対話の可能性を大きく広げ、今後も様々なサービスにおいて活用が期待されており、現在も活発に研究と開発が行われている。

一方、現代の社会における重要な課題の 1 つにメンタルヘルスのサポートがある。図 1 に 2023 年から 2022 年の日本における年代ごとの自殺率の推移を示す。この調査結果からわかるように、10～19 歳においては、調査期間中常に増加傾向が見られる。メンタルヘルスの分野は需要が増えながらも人手不足が懸念されている。

こうした背景の下、我々は、メンタルヘルスサポートの分野において、膨大な知識を持った ChatGPT 等の対話型生成 AI が活用できるのではないかと考えた。現在提供されている ChatGPT は、ユーザの発した多種多様な問題に対して高度な回答を行うことが可能である。したがって、ユーザはシステムに関して自分の悩みをプロンプトとして指定することで、自分が相談したい悩みを ChatGPT に与え、回答を得ることは可能である。

人は自分の個人的な悩みを他者に話すことに抵抗を保つ場合も少なくない。2023 年 10 月 20 日から 22 日にかけて、18 歳から 69 歳の 2971 名を対象に実施された調査において、「ChatGPT は良き相談相手になるか」についての調査が行われた[3]。この結果、人間よりもチャット GPT に対して秘密を打ち明けやすいと感じている人の割合が全体の約 42%に上ることが明らかになった。この結果は ChatGPT が相談相手として有用であることを示唆している。さらに、調査参加者の約 24.4%がチャット GPT を「会話相手・相談相手として」利用しており、ChatGPT が既に日常生活や個人的な悩みに対するサポートに利用されていることがわかる。「専門家の人と ChatGPT のような生成 AI とどちらに悩み相談をしたいと思いますか」という質問項目においては「専門家のみ」と回答した人が 43.4%で最多であったが、「AI のみ」や「必要に応じて専門家にも」といった選択をした回答者も存在し、AI への相談受け入れが一定数存在することが確認された。

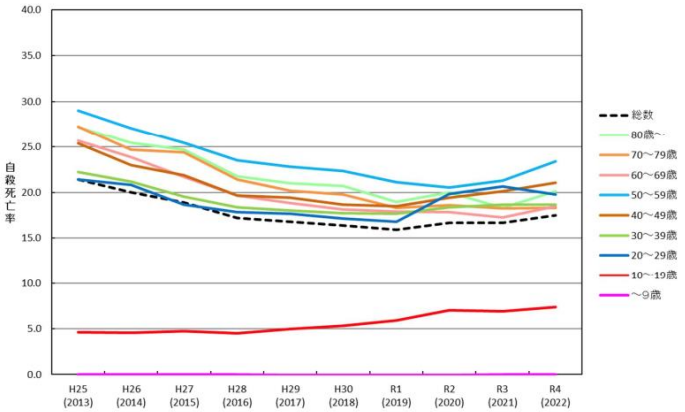


図 1: 日本における年代ごとの自殺率の推移[3]

そうしたことから、AI に対して自分の悩みを相談することができれば、これまで他人に話すことができずに一人で悩みを抱え込むような状況を少しでも改善できる可能性がある。

人が悩みを相談できる AI を考える場合、一般的な質問に対する回答とは異なる側面が存在する。一般的には、AI はユーザが与えた問いに対して正しい回答を行うことが重要と考えられる。しかし、悩みを相談された際の回答としては、単に正しい回答を行うだけでは不十分である場合がある。ユーザが相談を行うとき、本人単独では解決が困難な事象に関して、その問題に対する意見を求める。その場合、一般的で正確さを重視した回答ではユーザの「相談する」という目的に応えられない場合がある。悩みの相談に対する回答としては、回答者が相談者に寄り添い、ユーザにとって受容性の高い回答を行うことが重要であると考えられる。

しかし、現在 ChatGPT が行う回答は、一般論や必要な事実のみ端的に伝える内容が主であり個々の状況に寄り添っていない、既知の事実の提示してくる、相談者に共感しない等の特徴から、相談に対する回答としては適切ではないと考えられる。

上記の問題を解決するために本研究では相談に対する適切な回答を「思いやりが感じられるもの」「相談者の役に立つもの」の二つを満たすものであると定義した。この二つの要素の中で、相談者に与える印象としてより大きな割合を占め、現在の ChatGPT に欠けていると思われる「思いやりが感じられるか」について着目し、ユーザからの相談に対して思いやりを持った回答を回答できる手法を開発することを本研究の目的とする。ChatGPT を活用するためには、ユーザの要求や質問に対して AI が適切な応答を生成するための指示であるプロンプトの最適化が不可欠である。この研究により、特に心理的支援やカウンセリングの分野での AI の利用が拡大する可能性があり、また AI によるメンタルヘルスサポートのアクセシビリティを高める効果も期待できる。

2. 関連研究

Kosinsk ら[5]は、大規模な言語モデルにおいて、心の理論（ToM）と呼ばれる他人の観察できない精神状態を理解する能力が自然発生的に現れる可能性があることを述べている。これは、言語モデルの発展が人間の社会的相互作用、コミュニケーション、共感、自意識、道徳といった重要な側面を反映する可能性があることを示しており、ToM タスクを解決する能力が言語モデルの発展とともに向上している可能性を指摘している。また、ToM のような精神的な状態を他者に帰属させる能力が、AI が人間と対話・コミュニケーションする上で大きな進展をもたらす可能性があることを示唆している。

Chen[6]らは、メンタルヘルス分野で LLM を利用する際に、既存の LLM のシステムは感情分析や共感に特化している傾向があり、対象者の「思考や認知」をモデル化して診断する方法論が欠けているとして、LLM に精神的な健康問題を診断させるフレームワークを開発した。フレームワークにより生成された診断の根拠は、人間の専門家によって高く評価され、専門家による精神療法の補助ツールとして有用であるとされた。

3. 提案手法

3.1 対話型生成 AI を利用した相談に対する回答

本研究では ChatGPT を対象とする。ChatGPT を利用する場合は、ユーザは ChatGPT にプロンプトと呼ばれるテキストを与える必要がある。いま、ユーザが考える相談内容を表すテキストをqとするとき、質問に回答するためのプロンプトを以下に示す。

「$\{q\}$」という相談に回答してください、回答は文章で行ってください。

例として「親が頑張りを認めてくれません。毎日のようにやりたくもない勉強をさせられる日々がつらいです。」という相談に対する ChatGPT の回答を図 2 に示す。

> 親御さんからの認識を得られないことは、心に大きな重荷となるでしょう。まず大切なのは、自分自身の努力を認め、自己価値を大事にすることです。次に、親御さんとの関係に目を向けましょう。困難かもしれませんが、冷静に自分の感情を整理し、それを親御さんに伝えることが大切です。あなたのストレスやプレッシャー、そして勉強に対する本心を、平穏な環境で話し合う機会を設けることを考えてみてください。親御さんも、あなたのことを思うあまり、あなたの感じていることに気づかないことがあります。もし直接話すことが難しい場合は、信頼できる第三者に助けを求めるのも一つの方法です。学校のカウンセラーや信頼できる親族、友人などに相談して、あなたの気持ちを伝える手助けをしてもらうことができます。そして、勉強だけが人生ではないことを忘れないでください。自分の興味や趣味に時間を割くことで、精神的なバランスを保つことができます。自分を大事にし、自分に合ったペースで物事を進めることが、長期的に見て良い結果をもたらすでしょう。どんな状況でも、自分自身の幸せと健康を最優先に置くことが何よりも大切です。自分の頑張りを認め、自信を持って進んでいくことが、あなた自身の成長につながるはずです。

図 2:「親が頑張りを認めてくれません。毎日のようにやりたくもない勉強をさせられる日々がつらいです。」という相談に対する ChatGPT の回答の例

3.2 思いやりのある回答の生成

「思いやり」とは、『保育用語辞典 第 5 版』では、「相手の立場に立って、相手の感情を感じ取り、共感できる能力」であると定義されている。一方、『広辞苑』では、「相手の立場で考える」と定義されている。これらの定義を参考にすると、対話において「思いやり」をもった回答を生成するために、以下の 3 段階からなる処理が有効ではないかと仮説を立てた。

1. ユーザの相談に対して初期回答（回答①）を生成する。
2. 相手の立場に立って、相手が回答①を受け取った際に生じる相手の感情を予測する。
3. 予測した感情内容に基づき、回答（回答②）を修正する

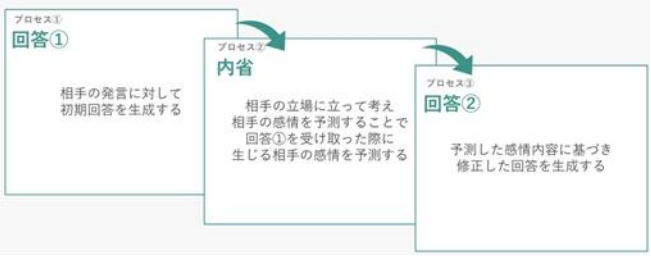


図 3:提案手法で仮定する思いやりを持つ回答をするためのプロセスの流れ

3.1 に示したプロンプトによる回答文の生成でのような、通常の ChatGPT の対話応答では、上記ようなプロセスをどの程度経ているかを検討した結果、AI がこのプロセスを完全には経ていないか、あるいは情報不足により相手の感情を適切に予想できていない可能性が考えられた。先行研究により、ChatGPT が ToM タスクを解決する能力が言語モデルの発達とともに向上している可能性が示唆されていることから情報不足により相手の感情を適切に予想できていない可能性は低いと考えられる。

そこで、AI に対して「自らの回答が相手にどのように受け取られるかを予想する」という指示を与えたところ、回答①における共感の不足や寄り添う姿勢の欠如などの問題点を適切に指摘できることが明らかになった。これに基づき、AI が本プロセスを経ていない可能性が高いという結論に至った。

以上の仮説に基づき、AI の応答を相手の立場に立った、思いやりのある回答へと改善するための手法として、「内省的質問応答法」を提案する。この手法は、以下の二段階からなるプロンプトに基づく指示により、ChatGPT に内省を促すものである。

1．「○○」という相談に対する回答を行ってください回答は文章で行ってください。

2．この回答を受け取ることで生じる相手の感情を考慮して、修正を行ってください、回答は文章で行ってください。

本手法を「内省的質問応答法」と称し、AI の対話応答の質の向上に資することを目指す。

図 4: 内省的質問応答法の概要

図 1 に示したものと同じ相談内容に対して内省的質問応答法により生成した回答の例を図 5 に示す。

> この状況についてお話しいただき、本当にありがとうございます。親御さんからの理解や支持を得られず、望まない勉強を強いられることによって生じるストレスは、非常に厳しい状況であることが伝わってきます。まず、自分の感情に耳を傾け、それらを受け入れることが大事です。親御さんとのコミュニケーションは難しいかもしれませんが、あなたの感じていることを落ち着いて、はっきりと伝えることが重要です。自分の気持ちを伝える時は、落ち着いて、感情的にならずに話すようにしてください。また、あなた自身の努力を自分で認めることも忘れないでください。他人からの評価だけでなく、自分自身で自分の頑張りを認めることは、自尊心を育て、精神的な健康にとっても良いことです。勉強に関しては、今はつらいかもしれませんが、この経験が将来的にどのように役立つかを考えてみると良いかもしれません。しかし、勉強だけが人生ではありません。勉強以外の時間を有意義に過ごし、自分の趣味や好きなことに時間を割くことも大切です。バランスの良い生活を送ることで、ストレスを減らし、日々をより楽しく過ごすことができます。一人で抱え込むのは難しいですから、信頼できる友人や教師、カウンセラーに相談するのも良いでしょう。他人の意見や助けが、問題解決の手助けになることがあります。親御さんとの関係を改善し、自分自身の幸福を高めるために、適切なサポートを探すことをお勧めします。あなたは一人ではないので、必要な時には手を差し伸べてもらうことを恐れないでください。自分自身を大切にして、少しずつでも前に進んでいきましょう。

図 5: 内省的質問応答法による回答の例

4. 評価実験

4.1 実験設定

提案手法の有効性を評価するために、評価実験を行った。本論文では提案手法の他に 3 種類の比較手法を考えた。比較手法を以下に示す。

- **比較手法 1**:「{q}」という相談に回答してください、回答は文章で行ってください。
- **比較手法 2**:「{q}」という相談に思いやりを持って

回答してください、回答は文章で行ってください。

- **比較手法 3**:「{q}」という相談に相談者にとって有益と思われる回答をしてください、回答は文章で行ってください。

ここで、比較手法 1 は、ChatGPT に直接相談をしたときの回答である。比較手法 2 は、思いやりを持つことをプロンプトに明示的に指示をしたものである。比較手法 3 は、相談者にとって役に立つことをプロンプトに明示的に指示したものである。

今回の実験で利用した相談は以下の 5 つである。

- **相談 1**（親からの否定）
  「親が頑張りを認めてくれません。毎日のようにやりたくもない勉強をさせられる日々がつらいです。」
- **相談 2**（親の介護）
  「親の介護が負担です。独身の自分が親の介護を自分が一手に担っており、近い将来老々介護になりそうで不安です。」
- **相談 3**（同僚への嫉妬）
  「優秀な同僚に嫉妬してしまいます。自分より早く出世していった人を見て自己嫌悪で死にそうです。」
- **相談 4**（同性愛の不安）
  「長年片思いしている同性に告白したいと考えています。ですが関係が壊れてしまわないか不安です。」
- **相談 5**（父親との関係）
  「父親がウザイです。説教じみた話を毎日のようにしてきます。大嫌いになりそうです。」

10 名に対して、アンケート形式の調査を行った。まず、被験者に対して、悩みを持つ 5 名がそれぞれの相談内容についてチャットで相談を行っているという状況を説明したのち、以下 4 項目について、5 段階で評価した。

1. 思いやりがある回答だと思いますか（**思いやり**）
2. この回答が相談者の問題解決に役立つと思いますか（**問題解決**）
3. 相談者にとってこの回答は受け入れられやすいと思いますか（**受容性**）
4. この回答者に親しみを感じますか（**親しみ**）

なお、ユーザに提示する質問の順序をランダムにすることによって、評価者が、各結果がどの手法に起因するのかを識別できないようにした。

4.2 実験結果

アンケートの結果に関して、それぞれの手法で生成した回答に対して、4 つの評価項目に関する被験者の回答の平均値を表 1 に示す。また、それぞれの評価項目の評価値分布を表す箱ひげ図を図 7〜9 に示す。

表 1: 評価項目ごとの平均値

	比較手法1	比較手法2	比較手法3	提案手法
思いやり	3.55	3.71	3.65	**4.39**
問題解決	3.82	3.61	3.55	**3.94**
受容性	3.53	3.22	3.39	**4.06**
親しみ	3.24	3.22	3.24	**4.2**

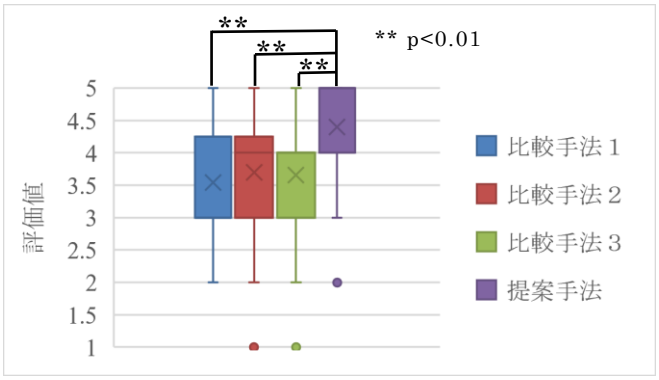


図 6: 「思いやり」に関する評価値の箱ひげ図

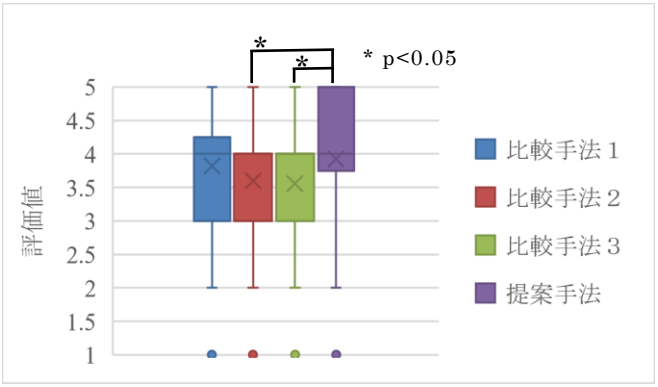


図 7: 「問題解決」に関する評価値の箱ひげ図

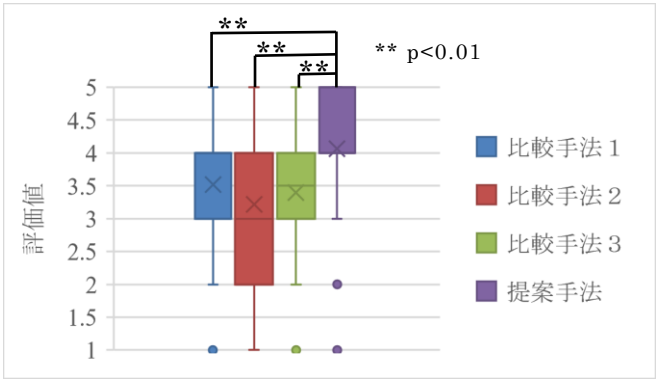


図 8: 「受容性」に関する評価値の箱ひげ図

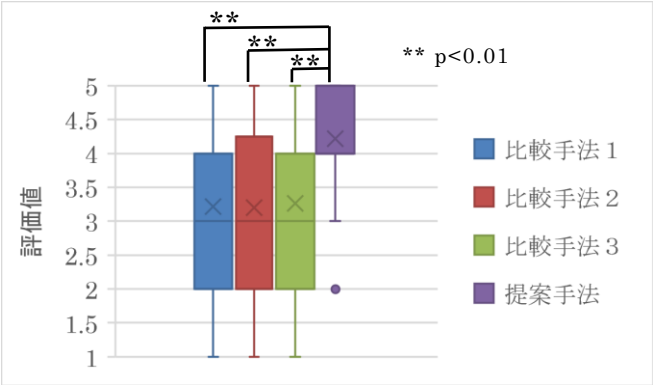


図 9: 「親しみ」に関する評価値の箱ひげ図

本手法では「提案手法は比較手法 i と評価項目 j において評価に差がない」を帰無仮説とする場合、12 個の検定を行う多重比較となる。そこで BH 法を用いて多重検定を行う。目標とする偽発見率（False Discovery Rate: ）FDR=5%と FDR=1%としたとき、それぞれの帰無仮説に対する t 検定の p 値と、それらを昇順に並べたときの満たすべき p 値の値をまとめた表を表 2 に示す。この表から有意水準 1%で思いやり、親しみ、受容性に関してはすべての帰無仮説が棄却され、提案手法の有用性が示された。さらに有意水準を 5%とした場合、比較手法 2 と比較手法 3 については、提案手法の優位性が示された。

表 2:BH 法による多重検定

順位	評価項目	比較対象	t検定のp値	p*（FDR=0.05）	p*（FDR=0.01）
1	思いやり	比較手法1	0.0000002	**0.0041667**	**0.0008333**
2	親しみ	比較手法1	0.0000004	**0.0083333**	**0.0016667**
3	親しみ	比較手法2	0.0000009	**0.0125000**	**0.0025000**
4	思いやり	比較手法2	0.0000020	**0.0166667**	**0.0033333**
5	受容性	比較手法2	0.0000037	**0.0208333**	**0.0041667**
6	親しみ	比較手法3	0.0000076	**0.0250000**	**0.0050000**
7	思いやり	比較手法3	0.0000195	**0.0291667**	**0.0058333**
8	受容性	比較手法3	0.0002585	**0.0333333**	**0.0066667**
9	受容性	比較手法1	0.0016095	**0.0375000**	**0.0075000**
10	問題解決	比較手法2	0.0155187	**0.0416667**	0.0083333
11	問題解決	比較手法3	0.0162239	**0.0458333**	0.0091667
12	問題解決	比較手法1	0.2256682	0.0500000	0.0100000

なお、相談項目ごとの各手法の評価平均値を表す棒グラフを図 10～14 に示す。なお以下のグラフでは、手法 1～手法 3 は比較手法 1～3 を表し、手法 4 が提案手法を表す。

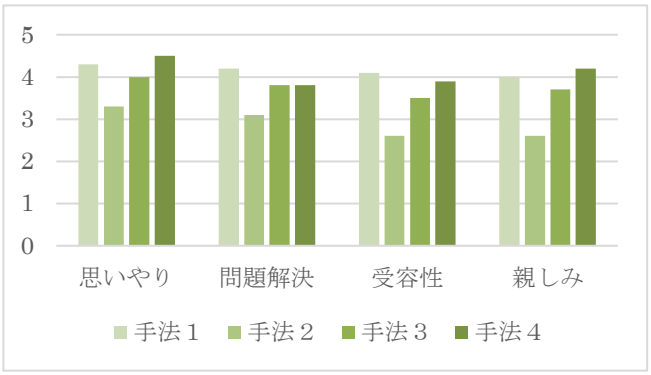


図 10: 「親からの否定」に関する結果

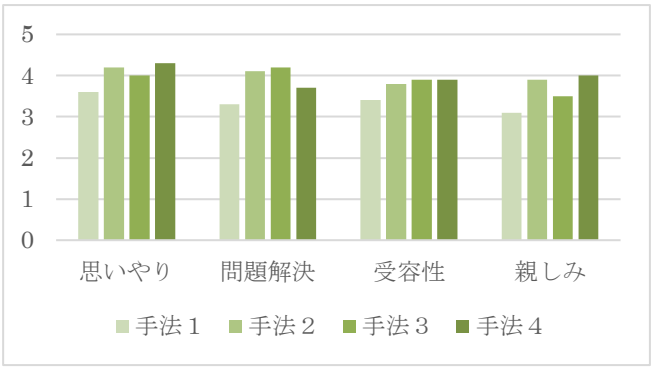


図 11: 「親の介護」に関する結果

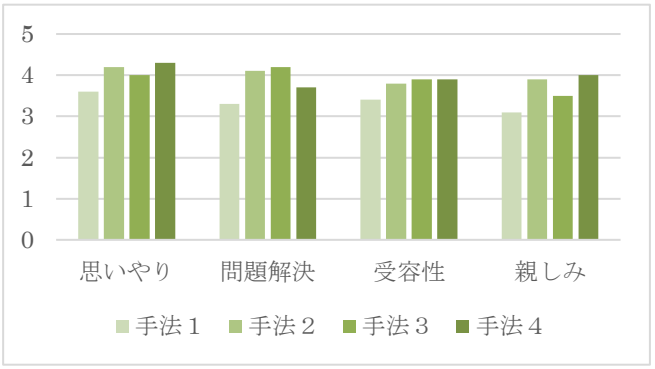


図 12: 「同僚への嫉妬」に関する結果

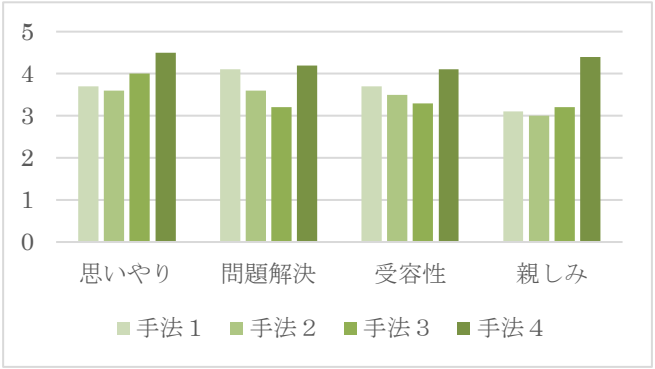


図 13: 「同性愛の不安」に関する結果

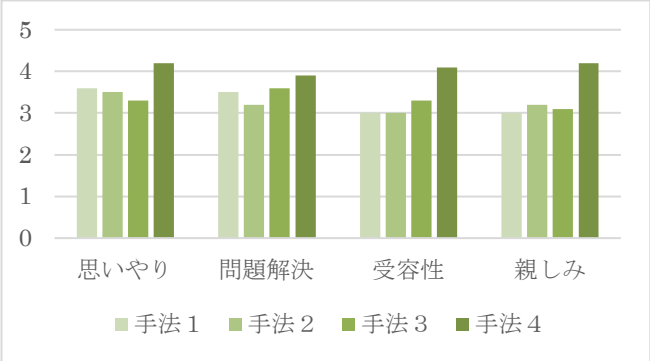


図 10: 「父親との関係」に関する結果

4.3 考察

「思いやり」、「親しみ」、「受容性」の評価項目において有意差が見られるため、思いやりを持たせるという面においては提案手法が効果的であることがわかる。

本研究により、AI 応答における思いやりの要素が相談者の感情への影響に及ぼす効果が確認された。ChatGPT が単なる情報提供だけでなく、相談者が抱える感情や状況に寄り添う応答が可能であることがわかる。提案手法が他の手法に比べて高い評価を受けたことから、相談者の感情に焦点を当てることで、相談者に受け入れられやすい思いやりを持った回答を生成することができることが分かった。

しかし、本研究においてアンケート対象のサンプルサイズが小さいため、結果の一般化は難しい。また、個々の相談者の特性や文化的背景によって受け入れられる回答の性質が異なる可能性があり、より広範な層に対する調査と研究が求められる。

5. おわりに

本論文では対話型生成 AI を利用した相談に対してより適切な回答を行うための内省的質問応答法を提案した。被験者実験により、「思いやり」、「親しみ」、「受容性」に関して、比較手法よりも良好な声質を持つ回答を生成可能であることが明らかになった。

今後の研究では、より大規模なサンプルを用いた調査、多様な特性を考慮した応答スタイルの開発を行う必要である。

将来的には、ChatGPT が人間の感情や心理状態をより深く理解できることになることで、メンタルヘルスサポートの分野において ChatGPT を活用することが期待される。

参 考 文 献

[1] L. Ouyang et al., “Training language models to follow instructions with human feedback,” arXiv, 2203.02155, 2023.

[2] A. Vaswani et al., “Attention Is All You Need,” arXiv, 1706.03762, 2017.

[3] 厚生労働省自殺対策推進室, “令和４年中における自殺の状況,” https://www.mhlw.go.jp/content/R4kakutei01.pdf

[4] “ＡＩに相談、１年で日常に　夫婦の悩み「モヤモヤを言語化”, 朝日新聞, 2 月 12 日朝刊, 2024

[5] M. Kosinski, “Theory of Mind May Have Spontaneously Emerged in Large Language Models,” arXive, 2302.02083, 2023.

[6] Y. Y. Chiu, A. Sharma, I. W. Lin, T. Althoff, “A Computational Framework for Behavioral Assessment of LLM Therapists,” arXive, 2401.00820, 2024

6. 付録：質問に対する回答

相談 1-手法 1 よる回答

「親御さんからの認識を得られないことは、心に大きな重荷となるでしょう。まず大切なのは、自分自身の努力を認め、自己価値を大事にすることです。次に、親御さんとの関係に目を向けましょう。困難かもしれませんが、冷静に自分の感情を整理し、それを親御さんに伝えることが大切です。あなたのストレスやプレッシャー、そして勉強に対する本心を、平穏な環境で話し合う機会を設けることを考えてみてください。親御さんも、あなたのことを思うあまり、あなたの感じていることに気づかないことがあります。もし直接話すことが難しい場合は、信頼できる第三者に助けを求めるのも一つの方法です。学校のカウンセラーや信頼できる親族、友人などに相談して、あなたの気持ちを伝える手助けをしてもらうことができます。そして、勉強だけが人生ではないことを忘れないでください。自分の興味や趣味に時間を割くことで、精神的なバランスを保つことができます。自分を大事にし、自分に合ったペースで物事を進めることが、長期的に見て良い結果をもたらすでしょう。どんな状況でも、自分自身の幸せと健康を最優先に置くことが何よりも大切です。自分の頑張りを認め、自信を持って進んでいくことが、あなた自身の成長につながるはずです。」

相談 1-手法 2 による回答

「ご相談ありがとうございます。親御さんからの認識や支持が得られないと感じるのは難しい状況です。自分の意志に反して勉強をさせられることは、ストレスを引き起こす原因となり得ます。感じている感情については、それを認識し、適切に対処することが必要です。信頼できる人に相談することを考えてみると良いでしょう。他者の意見は新しい視点を提供し、問題

解決に役立つことがあります。親御さんとのコミュニケーションも重要です。彼らの意図を理解し、あなたの立場を伝えることで、相互の理解が深まるかもしれません。また、自己肯定感を高めることも重要です。自分自身を大切にし、自分の幸せや健康に優先順位を置くことは、ストレスの多い状況を乗り越える上で助けになります。これらのステップを踏むことで、状況が改善することを期待しています。自分自身の強さと能力に信頼を置き、一歩ずつ前進してください。」

相談 1-手法 3 による回答

「このような状況は非常に難しいものですが、親御さんとのコミュニケーションを改善することで、あなたの気持ちが理解される可能性があります。まず、冷静になり、自分の気持ちを正直に伝えることが大切です。例えば、勉強に対する自分の感情や、頑張りが認められないことに対するフラストレーションを表現することが重要です。この際、非難や怒りの言葉は避け、どうしてそう感じるのかを説明することが効果的です。また、親御さんがあなたの勉強についてどのように考えているのかを理解することも大切です。彼らがなぜ勉強を強く推奨するのか、その背景や理由を聞くことで、お互いの立場を理解しやすくなります。そして、あなた自身の将来に対する考えや、勉強以外に力を入れたいことがあれば、それを伝えるのも良いでしょう。コミュニケーションの際には、お互いの意見を尊重し、解決策を一緒に考える姿勢を示すことが重要です。例えば、勉強の方法や時間を調整することで、ストレスを減らし、より効果的な学習ができるかもしれません。また、カウンセリングや学校のカウンセラーとの相談を通じて、第三者の意見を聞くことも一つの方法です。最後に、自分自身を大切にすることも忘れないでください。趣味やリラクゼーションの時間を持つことで、ストレスを軽減し、心のバランスを取ることができます。あなたが健康で幸せであることが最も重要です。」

相談 1-手法 4 による回答

「この状況についてお話しいただき、本当にありがとうございます。親御さんからの理解や支持を得られず、望まない勉強を強いられることによって生じるストレスは、非常に厳しい状況であることが伝わってきます。まず、自分の感情に耳を傾け、それらを受け入れることが大事です。親御さんとのコミュニケーションは難しいかもしれませんが、あなたの感じていることを落ち着いて、はっきりと伝えることが重要です。自分の気持ちを伝える時は、落ち着いて、感情的にならずに話すようにしてください。また、あなた自身の努力を自分で認めることも忘れないでください。他人からの評価だけでなく、自分自身で自分の頑張りを認めることは、自尊心を育て、精神的な健康にとっても良いことです。勉強に関しては、今はつらいかもしれませんが、この経験が将来的にどのように役立つかを考えてみると良いかもしれません。しかし、勉強だけが人生ではありません。勉強以外の時間を有意義に過ごし、自分の趣味や好きなことに時間を割くことも大切です。バランスの良い生活を送ることで、ストレスを減らし、日々をより楽しく過ごすことができます。一人で抱え込むのは難しいですから、信頼できる友人や教師、カウンセラーに相談するのも良いでしょう。他人の意見や助けが、問題解決の手助けになることがあります。親御さんとの関係を改善し、自分自身の幸福を高めるために、適切なサポートを探すことをお勧めします。あなたは一人ではないので、必要な時には手を差し伸べてもらうことを恐れないでください。自分自身を大切にして、少しずつでも前に進んでいきましょう。」

質問 2-手法 1 による回答

「親御さんの介護を一人で担うというのは、大変な心労となることは間違いありません。そんな中での老々介護への不安も理解できます。まず大切なのは、自分自身の健康と心のケアです。介護に追われる中で自分を犠牲にしてしまうと、結果的に親御さんへのサポートも難しくなります。可能であれば、地域の介護支援センターや社会福祉協議会などに相談してみましょう。これらの機関は介護のプロフェッショナルであり、介護サービスの利用方法や費用の支援、介護に関する相談に応じてくれます。また、独身であることによる孤独感や負担感も重要な要因です。友人や同僚、地域の人々とのコミュニケーションを大切にし、心の支えとしてください。また、自分だけでなく他の家族や親戚、近隣の方々にも協力を求めることも一つの手です。介護は一人で行うものではなく、周囲と協力し合うことが重要です。最後に、将来の老々介護に関しては、今から備えておくことが大切です。貯蓄計画を立てたり、将来的に必要となる介護サービスについて情報を集めたりすることが重要です。また、自分自身の健康管理にも注意を払い、できるだけ健康を保つ努力をしましょう。」

質問 2-手法 2 による回答

「介護の重圧を一人で担っているという状況は、非常に大変なことであり、あなたが感じている不安や負担について共感します。お一人で全てを背負い込むことは、精神的にも肉体的にも大きな負担になります。介護は一人ではなく、周囲のサポートを得ながら行う

ことが大切です。まず、自分自身の健康とウェルビーイングを守ることが何より重要です。自分を労わり、適切な休息を取ることは、長期的な介護において非常に大切なことです。また、心理的なサポートを受けることも、ストレスを軽減するのに役立ちます。地域の支援団体やカウンセリングサービスを利用して、感情を話す機会を持つことをお勧めします。次に、地域のリソースやサービスを活用することが重要です。介護サービス､デイサービス､ホームヘルプサービスなど、多くの支援が利用可能です。これらのサービスを利用することで、あなたの負担を軽減し、親御さんにも適切なケアを提供できます。そして、将来に向けての計画を立てることも重要です。老々介護の可能性に備えて、早めに情報を収集し、必要なサポートを確保することが大切です。可能であれば、財務計画も含め、将来のケアについての計画を立てることをお勧めします。このような状況は誰にでも起こり得ることで、あなたが感じているような不安や圧倒感を持つのは自然なことです。どうか、一人で全てを抱え込まず、周囲の支援を積極的に求めてください｡あなたの健康と幸福も、親御さんのケアと同じくらい重要です。」

質問 2-手法 3 による回答

「親の介護を一人で担うことは、非常に大きな負担となることがあります。独身でこの重責を背負うことは、特に将来の老々介護の可能性を考えると、不安を感じるのも自然です。このような状況においては、まず、自分自身の健康と精神的なウェルビーイングを大切にすることが重要です。自分が倒れてしまっては、介護は成り立ちません。次に、介護に関するサポート体制を探ることが大切です。地域の社会福祉協議会や市町村の福祉窓口などに相談し、利用できる介護サービスや支援制度について情報を得ることをお勧めします。例えば、訪問介護やデイサービス、ショートステイなどの利用を検討することで、あなた自身の負担を軽減できる可能性があります。また、家族や親族、友人などに相談し、可能であれば支援を求めることも一つの手段です。全てを一人で背負い込むのではなく、他者の協力を得ることで、精神的な負担を減らすことができます。将来のことを考えると、老々介護のリスクを減らすためにも、早めに相談し、計画を立てることが大切です。自分自身の将来の生活設計も含め、専門家のアドバイスを受けながら、適切な支援体制を整えることが望ましいでしょう。最後に、精神的なケアも忘れずに。カウンセリングやセルフケアの方法を学ぶことで、心の負担を和らげることができます。介護は肉体的な負担だけでなく、心理的な負担も大きいため、自分自身の心のケアを大切にしてください。」

質問 2-手法 4 による回答

「このような状況での介護は、本当に心身ともに大変な努力が求められますね。あなたの親御さんへの深い愛情と、その責任感には、心から敬意を表します。しかし、同時にあなた自身の健康や未来に対する不安を感じるのは、とても自然なことです。自分だけで全てを担うことの大変さや、将来への不安は、誰もが感じるものです。まず大切なのは、自分自身の健康と精神的な安定を保つことです。介護の負担を軽減するために、地域の支援サービスや専門家のアドバイスを活用することを考えてみてはいかがでしょうか。また、親族や友人といった周囲の人々にも、あなたの状況を共有し、協力を求めることも重要です。一人で抱え込まず､周りのサポートを受け入れる勇気を持つことが、長期的な介護生活を支える鍵となります。そして、介護に関する知識や技術を学ぶことは、あなた自身の負担を減らすだけでなく、介護の質を高めることにも繋がります。地域の介護講座やワークショップに参加することで、効率的な介護方法やストレスの管理方法を身につけることができます。どうか、あなたが一人で全てを解決しようとしないでください。あなたの負担を少しでも軽減できるよう、周囲の支援を受け入れ、適切なサービスを利用することをお勧めします。あなたの心身の健康が、介護生活の持続可能性にとっても非常に重要です。心から、あなたの負担が軽くなることを願っています。」

質問 3-手法 1 による回答

嫉妬は人間関係においてよくある感情ですが、それが自己嫌悪につながるとき、自分自身と向き合う重要な機会となります。まず、他人の成功に集中するのではなく、自分自身の長所や成功を認めることが大切です。他人と自分を比較するのではなく、自分の成長と進歩に焦点を当てることで、自己評価を高めることができます。また、同僚の成功を刺激として、自分自身の目標を見つめ直す良い機会と捉えることも有効です。彼らがどのような努力や戦略で成功を収めたのかを分析し､自分のキャリアプランに活かすことができます。嫉妬をポジティブな動機付けに変えることで、自分自身の能力を最大限に引き出すことが可能になります。最後に、自己嫌悪の感情はしばしば過小評価から来るものです。自分の価値を正しく評価し、自己肯定感を高めるために、自分の成果を振り返り、それを認めることが重要です。必要であれば、専門家の助けを借りて､自己受容の道を歩むことも検討してみてください。自己受容とは、完璧でなくても自分自身を認めることであり、それによって自己嫌悪の感情を乗り越え、よ

り健全な自己評価を築くことができます。

質問 3-手法 2 による回答

「職場での他人の成功は、時に自分自身の進歩や価値を疑問視させることがありますが、この感情はあなたが自分自身と向き合う機会を提供しています。まず、嫉妬や自己嫌悪を感じること自体を否定せず、自分の感情を受け入れましょう。感情は私たちが直面する状況への自然な反応であり、それを認識することは自己理解への第一歩です。自分自身を責めるのではなく、どうしてこのように感じるのかを静かに探求することが大切です。次に、自己比較のサイクルから抜け出すためには、自分自身の長所に目を向けることが重要です。他人の成功に焦点を当てるのではなく、自分自身がこれまでに達成したこと、自分のユニークな才能や強みに注目してみましょう。自分自身の成長と進歩に焦点を当て、他人との比較ではなく、自己達成に喜びを見出すことが大切です。また、同僚の成功を学びの機会として捉えることも有効です。彼らがどのようなスキルや戦略を用いて成功を収めたのかを観察し、それを自分の成長に役立てることができるかもしれません。こうすることで、嫉妬をポジティブな動機付けに変換することが可能です。自己嫌悪の感情に直面したときは、自己受容の重要性を思い出してください。完璧でなくても自分を受け入れ、自分の価値を認識することが自己肯定感を高める鍵です。自分自身の小さな成功を認め、それに基づいて自己評価を築くことが大切です。もし感情が圧倒的であれば、信頼できる友人や家族、または専門家に相談することも考えてみてください。あなたの感情を共有し、支援を求めることは強さの表れです。自分を大切にし、自分の感情と向き合う勇気を持つことで、新たな自己理解と成長が生まれます。あなたは一人ではないので、必要なサポートを求めてください。」

質問 3-手法 3 による回答

「優秀な同僚に対する嫉妬は、職場でよくある感情ですが、これを乗り越えることはあなたの個人的および職業的成長にとって非常に重要です。まず、自己嫌悪や嫉妬の感情は自然なものであると認識することが重要です。これらの感情は、自分自身の欠点や不満を映し出していることが多いので、自己反省の機会として捉えることができます。他の人の成功に焦点を当てるのではなく、自分自身の成長と進歩に集中することが重要です。自分自身の強みや達成したことに目を向け、それらをさらに伸ばす方法を探してください。また、他人の成功から学び、それを自分自身のスキルや知識の向上に役立てることも重要です。プロフェッショナルな成長には時間がかかることを理解し、短期間での大きな成果を求めずに、小さな成功を積み重ねることに焦点を当ててください。また、メンターや同僚との良好な関係を築くことも、職場での成功には不可欠です。彼らからのフィードバックやアドバイスを積極的に求め、それを自己改善に活かすことで、より良い職業人に成長できるでしょう。最後に、自己嫌悪や嫉妬の感情があまりにも強い場合は、プロのカウンセリングを受けることも検討してください。プロの助けを借りることで、これらの感情を健康的に処理し、自己成長に向けた道を開くことができます。」

質問 3-手法 4 による回答

「嫉妬や自己嫌悪に苦しんでいるあなたにとって、こうした感情は非常に辛く、深刻なものでしょう。私の回答があなたの心に寄り添えるよう努めます。同僚への嫉妬や自己嫌悪を感じることは、自然な感情です。このような感情は、自分自身に厳しい目を向けがちな時によく起こります。まずは、自分の感情を否定せずに受け入れてみましょう。自分を他人と比べることは、誰にでも起こり得ることですが、それを自己成長のきっかけとして捉えることも可能です。具体的なアドバイスとしては、日記をつけることをお勧めします。毎日の感情や考えを記録することで、自分の内面をより深く理解することができます。また、自分がこれまでに成し遂げたこと、小さな成功でも良いので、それを記録してみてください。自分自身の成果や強みに目を向け、それを認めることは、自己評価を高める助けになります。同僚の成功を見て、それを自分自身のキャリアや目標を見直す機会として活用しましょう。彼らの成功から学び取ることで、あなた自身の成長に繋げることができます。また、自分の得意分野や興味を深めるためのコースやワークショップに参加することも有効です。感情が辛い場合、専門家の助けを求めることも一つの方法です。カウンセリングやコーチングを通じて、自己嫌悪と向き合い、自己受容の道を見つけることができます。自分に合ったサポートを探すことが重要です。大切なのは、自分自身を受け入れ、自分だけのペースで成長していくことです。他人の成功はあなたの失敗ではなく、それぞれの人が異なる道を歩んでいることを忘れないでください。自分自身に優しくあり、自分の成長を信じてください。あなたもまた、自分だけの成功を築くことができるのです。」

タスク遂行促進のための作業リアクションの自動生成手法の有効性評価

堀川　達平†　　北山　大輔††

† 工学院大学大学院工学研究科情報学専攻　〒 163–8677 東京都新宿区西新宿 1–24–2
†† 工学院大学情報学部システム数理学科　〒 163–8677 東京都新宿区西新宿 1–24–2
E-mail: †em23040@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　近年 YouTube などの動画配信プラットフォームにおいて，作業配信というジャンルの配信が行われている．自己説明や視聴者の反応によって仕事の生産性を向上させる可能性がある一方で，秘匿性や相当数の視聴者が必要であるという課題に直面している．これらの課題を克服するために，著者らは以前にもユーザの発話に対して視聴者のコメントを自動生成することで，リアルタイムのインタラクションをシミュレートする新しい手法を提案しているが，実験の対象を大学生 4 名としており，検証の余地があった．そこで本研究ではクラウドソーシングを利用しより大規模な実験を行い，提案手法の有効性，ひいては思考発話法の有効性を検証した．成果物の質的評価においては提案手法による質の向上は確認されず，俳句作成というタスクにおいては思考発話法の有効性を認めることはできなかったが，発話率などを比較した結果，提案手法は有意に発話の活性化を促すことが確認された．

キーワード　ユーザ支援，思考発話，LLMs，HCI，作業配信

1 はじめに

我々は論文執筆や報告資料の作成など，日常的にさまざまな作業を抱えて生活しているが，時には集中力の低下やモチベーションの喪失により，作業を円滑に進められない場合がある．そのような中，近年 YouTube などの動画配信プラットフォームにおいて，作業配信と呼ばれる自身の作業状況を実況しつつ視聴者とコミュニケーションを行い作業を進める「作業配信」というジャンルの配信が行われている．

作業配信では，作業内容を実況するという自己説明による課題解決促進効果と，応援やアドバイスといった視聴者からのリアクションによる発話促進効果という 2 つの効果の相乗効果によりタスク遂行が促進されていると我々は考えた．他者からのリアクションが存在する場合自己説明の理解が促進されるという研究結果があることからも，作業配信の効果が伺える [1]．しかしこのような利点のある作業配信は，自身の行動をインターネット上に配信する必要があり，機密性が高い作業を行うことができないという秘匿性の問題と，リアクションを得るためある程度の規模のリスナーを獲得する必要があるという知名度の問題を抱えており，実施の際の障壁となっている．

このことから，我々は以前に視聴者からのリアクションとしてコメントを自動生成する手法を提案したが [2]，実験の対象を大学生 4 名としており，さらなる検証の余地が残されていた．そこで本稿ではクラウドソーシングサービスを用いて 80 名の被験者に協力を依頼し，より大規模な実験を行うことで提案手法の有効性検証を改めて実施する．また，成果物である俳句に対し，別途依頼した被験者により各俳句について 3 回の質的評価を実施し，思考発話法についての有効性の検証も実施する．

2 関連研究

ここで，本稿に関連する研究を紹介し，本研究を位置付ける．まず，本研究では作業配信に着想を得てそれを模倣することを狙うが，配信に関する調査を行った先行研究が複数存在する．

Lee らの研究は勉強配信を対象に，人々が配信をどう利用しているかを調査した [3]．その結果，視聴者は配信者の存在感やコメント欄にいる他者からの刺激や精神的な支えを求めており，一緒に勉強しているという雰囲気により勉強を進めていることを明らかにした．この研究では視聴者を対象としているが，我々は配信者側の視点に立っている点で異なっている．Fraser らの研究では創作配信を対象に，視聴者と配信者がそれぞれどのような動機を持っているのかを調査している [4]．その結果，配信者は主にコミュニティ形成や視聴者の獲得，創作過程の共有，視聴者との有意義な会話を求めて配信を行っていることが明らかとなった．また，視聴者からインスピレーションを得たときや，感謝されたときにやりがいや楽しさを感じることがわかった．我々の取り扱う作業配信は創作配信の範囲を広げたものと考えられるため，これらの結果を取り込むことができると考える．

本研究の提案手法は生成されたコメントにより仮想的な視聴者を作り，配信者に見立てたシステム利用者にやりがいや楽しさを与えると考えることができる．一方で Fraser らの研究では分割注意の問題により配信者がチャットについていけていないことを示しているが，コメント生成量に限界があることから本研究では特別な対応を行っていない．

作業配信を他者の影響を受けながらの作業と捉えると，その作業の例としてペアプログラミングが考えられる．ペアプログラミングを取り扱った研究として Williams らの研究があり，ここではペアがいることでコーディングが楽しくなり，チェック

図 1　被験者に表示される作業画面

者がいることで自分の解に自信が持てるという結果が得られている [5]．また，作業者はパートナーがいることで良いプレッシャーを得られ，タスクに集中できるようになるとも触れられている．この研究で得られた結果は他者から見られることによる作業効率の向上を示していると言える．本研究の提案手法は仮想的に多数の他者を再現することで，作業者に対し良いプレッシャーを与えることが可能であると考える．一方ペアプログラミングではペアは能動的な内容の確認を行っているが，本研究の提案手法はそれを行っていないため，作業効率の向上効果は限定される．

3　提 案 手 法

本研究の提案手法は，被験者が発話した作業内容の実況に基づき，大規模言語モデルを活用してコメントを生成し，それを表示することにより実現される．今回の検証に用いたシステムは以前の研究で用いたものに改良を加えているため，本稿では変更点を中心に説明する．

3.1　システム概要

今回用いたシステムはクラウドソーシングによって集められる被験者がオンラインで実験を実施できるよう，Web アプリケーションとして構築を行った．作業画面は図 1 のようになっている．画面は主に 3 つのエリアに分かれており，参考サイト部には作業に活用する参考情報を得るサイトが，コメント部には自身の発言の書き起こしと生成されたコメントが，作業領域部には作業内容のメモ書きや結果を記載する入力欄がそれぞれ表示される．

本稿では音声認識に，JavaScript より利用可能な Web Speech API の SpeechRecognition[1] を用いている．

3.2　視聴者コメントの生成

視聴者コメントの生成には OpenAI Chat Completions API[2] を利用している．用いたモデルは gpt-3.5-turbo-1106 である．

Chat Completions API では大規模言語モデルと自然言語でのやりとりが可能となっており，モデルのふるまいを指定する system プロンプト，ユーザからの入力を表す user プロンプト，モデルの出力を表す assistant プロンプトでやりとりが行われる．生成に使用したプロンプトは以下の通りである．

> User は自身が俳句の作成を行っている状況を実況しています。あなたは作業配信の視聴者というラフな立場です。あなたは User の発言に対し、視聴者コメントを日本語で複数個生成してください。なお、生成するコメントは短く簡潔である必要があります。

また，上記の system プロンプトに加え，プログラムでの連携のために API 機能の 1 つである function_calling[3] を利用するため，オリジナルの関数 send_to_comments を定義した．function_calling は JSON スキーマを用いて関数の引数を定義することで，API から構造化データをより確実に取得することができる．作成した JSON スキーマを以下に示す．function_calling の利用により，user の発言に対して JSON 形式で複数のコメントを返す機能を実現している．スキーマの中には生成されたコメントが応援 (encourage) かアドバイス (advice) かをを表す “type” が定義されているが，これは今後の研究のための予備的な定義であり，本稿では利用していない．

```
[ { "name": "send_to_comments",
    "description": "Send comments to the user's terminal. You can
     send more than one at a time by separating them with commas
    .",
    "parameters": { "type": "object",
      "properties": {
        "comments_list": { "type": "array",
          "description": "Multiple comments to send",
          "items": { "type": "object",
            "properties": {
              "type": { "type": "string",
                "description": "Comment type",
                "enum": ["encourage", "advice"] },
              "comment": { "type": "string",
                "description": "Comment to send" } } } } },
      "required": [ "comments_list" ] }
} ]
```

最後に，本提案では User プロンプトを入力する際，30 秒前までの User プロンプトの履歴を同時に入力することで，発話の文脈をモデルが認識できるようにしている．

4　実　　験

4.1　実 験 方 法

クラウドワークス [4] を利用して 80 名の被験者を募集し，得られた結果から正常に発話時間が記録できなかった結果と発話率が 2.5%未満の結果を除外することで，コメント生成機能無しで 21 名，コメント生成機能有りで 16 名分の有効な実験結果を取得した．システムは前述の通り著者らが Web アプリケーションとして構築しており，被験者らは自身のブラウザからアクセスする形で実験に参加する．

対象とするタスクは以前と同様に俳句の作成とした．タスク

1：https://developer.mozilla.org/ja/docs/Web/API/SpeechRecognition
2：https://platform.openai.com/docs/api-reference/chat/create
3：https://platform.openai.com/docs/guides/function-calling
4：https://crowdworks.jp/

図 2　練習に用いたテーマ

図 3　提示した 3 種類のテーマ

表 1　作業時間 [s], 発話率 [%], 発話間隔 [s] の実験結果

	コメント生成なし		コメント生成あり	
	平均	標準偏差	平均	標準偏差
作業時間	438.8	186.2	420.6	183.9
発話率	9.6*	7.6	17.8*	13.2
発話間隔	11.7*	34.0	7.4*	13.4

に取り組む前には発話思考の練習として，30 秒間，図 2 に示す画像のストーリーを自身の考えを発話しながら考えるよう依頼した．練習は 5 秒以上の無言を 2 回検知した場合やり直しとした．

本番では，テーマとして図 3 に示す 3 種類の画像のうちどれか 1 つをランダムに提示した上，季語の参考となるサイト [5] を提示し，俳句を作成してもらった．

さらに，得られた俳句に対し別途募集した被験者による 2 群間での相対評価を実施した．俳句は著者らがチェックし，短歌や短文などの俳句以外のものを除外した結果，コメント生成なしで 20 句，コメント生成ありで 13 句となった．20 句× 13 句の各組合せを 3 人で比較し，多数決によりどちらの俳句が優れているかを確認した．

4.2　結果と考察

コメント生成のありなしで，作業時間，発話した時間，ならびに発話間隔がどう変化したかを比較する．各被験者の作業時間，発話率，発話間隔の平均，標準偏差を表 1 に示す．発話率は発話時間を作業時間で割り百分率にした値である．発話間隔は初回の発言終了から，各発言の終了と開始の間隔を取得し，被験者全体で平均・標準偏差を求めている．ウェルチの t 検定を実施したところ，発話率と発話間隔の平均値に 5%水準の有意差が見られた．発話毎の間隔が短縮されている場合，提案手法が本来発話しなかった被験者に発話を促していると考えることができる．これらの結果は，提案手法である視聴者コメント生成機能が被験者の発話を活性化していることを示している．

次に，作業の成果物である俳句の評価について確認する．260 対の組み合わせの中で 60%以上の評価を獲得した俳句は，コメント生成なしで 7 句，コメント生成ありで 5 句であった．また，

5：https://kigosai.sub.jp/001/

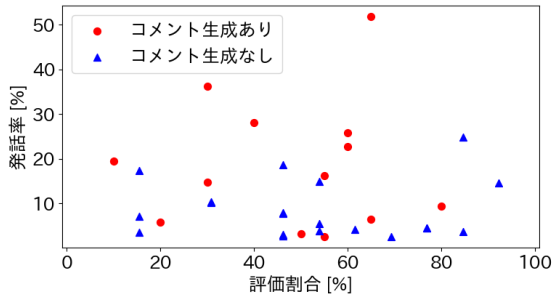


図 4　評価と発話率の関係

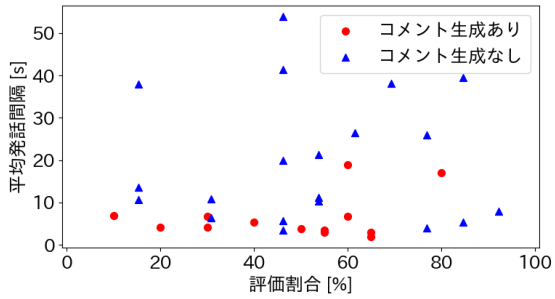


図 5　評価と発話間隔の関係

横軸を評価，縦軸を発話率としてプロットした結果を図 4 に，横軸を評価，縦軸を発話間隔としてプロットした結果を図 5 にそれぞれ示す．これらの結果からは，発話量の増加に伴う成果物の質の向上を確認できなかった．

今回の検証は俳句の作成というタスクで実験を行ったが，これは「比較的短時間で作業を完了できる」，「音声認識による誤認識が比較的少ない，日常会話の範疇の語彙が使われる」などの理由から選定したものであった．しかしながら伊藤らの研究 [1] によると，説明が効果を発揮するのは意味付けに対する説明やその繰り返しであることが示されており，俳句作成自体の過程ではそのような意味付けの説明が誘発されにくいと考えられる．このことから，成果物の質の向上を確認するためには，「俳句の推敲過程を説明する」といったような，より質に影響を与えるタスクを設定する必要があると考える．

5　おわりに

本論文では，我々が以前提案した作業に対するリアクションの自動生成手法について，実験規模を拡大して有効性の評価を行った．

今回行った実験の結果は，我々の提案手法が被験者の発話活性化を引き起こすことを示しており，本手法の有効性が確認できたと考える．一方で前提としていた発話量と成果物の質との関係が確認できなかったことから，本手法の用途は限定的なものであるといえる．

今後の展望として，発話思考に適切な実験タスクを検討した上での検証やよりよいプロンプトの設計，閲覧しているページの内容や作業内容の入力など発話以外の情報を用いた作業内容

認識によるコメント生成により応答内容の適切性を高めることなどが挙げられる．他にも、応援だけではなく叱責や批判などのコメントの導入や，盛り上がりに緩急をつけるといった，より実際の作業配信に即したコメント生成を行うことも考えられる．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 伊藤貴昭, 垣花真一郎. 説明はなぜ話者自身の理解を促すか. 教育心理学研究, Vol. 57, No. 1, pp. 86–98, 2009.

[2] Tappei Horikawa and Daisuke Kitayama. Task execution promotion method by automatic generation of viewer comments for live commentary of work contents. In *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 1–4, 2024.

[3] Yoonjoo Lee, John Joon Young Chung, Jean Y. Song, Minsuk Chang, and Juho Kim. Personalizing ambience and illusionary presence: How people use “Study with me” videos to create effective studying environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, pp. 1–13, 2021.

[4] C. Ailie Fraser, Joy O. Kim, Alison Thornsberry, Scott Klemmer, and Mira Dontcheva. Sharing the Studio: How creative livestreaming can inspire, educate, and engage. In *Proceedings of the 2019 Conference on Creativity and Cognition*, C&C ’19, pp. 144—155, 2019.

[5] Laurie Williams, Robert R. Kessler, Ward Cunningham, and Ron Jeffries. Strengthening the case for pair programming. *IEEE Softw.*, Vol. 17, No. 4, pp. 19—-25, jul 2000.

小説の感想を語り合えるエージェント構築のための レビューをもとにした小説解釈の再現

笹原　雄太†　　牛尼　剛聡††

† 九州大学 芸術工学部　〒 815–8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学院 芸術工学研究院　〒 815–8540 福岡県福岡市南区塩原 4-9-1
E-mail: †sasahara.yuta.745@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年, LLM を用いたユーザ行動の再現やパーソナライズが注目されている. 小説の感想を語り合えるエージェントをパーソナライズするためには, ユーザごとに異なる小説の解釈を LLM に理解させることが必要である考えられ, ユーザが小説を読んで感情形成が行われるまでの思考過程を明らかにすることが重要である. しかし, この思考過程が明示的に与えられていることは多くない. そこで本研究では, LLM を用いてレビューに対して感情を予測した後, LLM による自己質問応答によって思考過程を明らかにする枠組みである RUPIN を提案する. さらにそれを用いた, ユーザが小説の感想を語り合えるエージェントの構築手法を提案する. 評価実験において, 提案手法を用いることで, ユーザの考えに合ったエージェントの構築ができることが示唆された.

キーワード　LLM, ユーザ理解, 質問応答, 自己学習, 対話システム

1　はじめに

小説は多くの人が体験したことのあるコンテンツの 1 つである. 小説体験を通して, 感動を味わったり, 体験したことのない環境や生活を擬似的に経験することができる. また, 小説を読むだけでなく, レビューの投稿・閲覧や人と感想を語り合うことを通して, 小説の世界を楽しんだり, その小説に対する理解を深めたりする人も少なくない. このような行為はコンテンツ体験の拡張であると考えられる.

このようなコンテンツ体験の拡張においては, 例えばレビューの閲覧ではそのレビューの投稿者, 感想を語り合うときには語り合う相手のように, 同じ作品を体験したことのある他人が必要である場合が多い. しかし, 近年では電子書籍の普及とともに, 書籍の出版数は増加傾向にあり [1]（図 1）, 多くのコンテンツが流通する中で, 自身が体験した作品について語り合える人と出会うことは容易ではない.

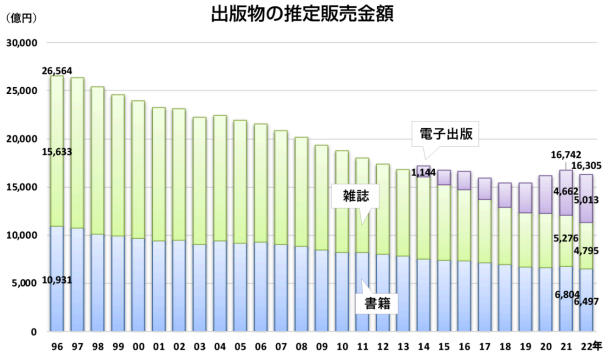


図 1　出版販売額の推移

この問題を解決するために, 小説の感想を語り合う相手として AI を用いることが考えられる. AI の中でも, LLM は自然言語に特化しており, ユーザと小説の感想を語り合うという対話を行う場合に適していると考えられる. Chiu ら [2] は心理療法の手法に基づいて LLM を感情や問題を理解するセラピストとして振る舞わせることで, 優秀な人間のセラピストに近い受け答えを生成することが可能であり, 特にユーザの発言や感情に対して適切に反応することを得意としていることを示した. 小説を読み終えた後の感想や感情を語り合うエージェントには, ユーザの発言や感情に対して適切に反応することが必要であると考えられる. したがって, 感想を語り合うエージェントとして LLM を使用することは有効な手段の 1 つであると考えられる.

LLM（Large Language Model）とは大量のテキストデータを用いて訓練された自然言語処理のモデルである. LLM を構築する際には, Vaswani ら [3] が提案したニューラルネットワークアーキテクチャである Transformer が標準的に用いられる. Transformer には, エンコーダ・デコーダ, エンコーダのみ, デコーダのみの 3 種類の構成が存在し, 構成によって得意なタスクが異なるが, 特にデコーダ構成の Transformer は文章生成タスクなどに用いられる. 例えば, Radford ら [4] が提案した GPT にも Transformer が採用され, 文章生成を中心とした下流タスクにおいて高い性能が示された. また, GPT は定期的にバージョンが更新されており, 2019 年に GPT-2 [5] , 2020 年に GPT-3 [6] , 2022 年に Instruct GPT [7] , 2023 年に GPT-4 [8] が発表され, バージョンの更新に伴ってモデルに含まれるパラメータ数および事前学習に使用されるコーパスの容量が増加することで, 性能が急速に向上した. この背景には, Kaplan ら [9] が提案した, モデルの規模を大きくすることで性能が比例して改善していくという経験的な法則であるスケール則が存在している. したがって, 今後もモデルの大規模化に伴って性能が向上すると考えられるため, GPT をはじめとした LLM は様々な分野への応用が期待されている.

LLM は事前に学習した内容に基づいて尤もらしい回答を生

成することに優れている一方で, 未学習の内容に関する回答や特定のユーザに最適化された回答を単純な生成方法で獲得することは難しい. しかし, 特に LLM を用いて小説の感想を語り合えるエージェントを構築するためには, ユーザの感想や, 抱いている感情に基づいて回答を生成する必要がある. そのため, ユーザごとに LLM のパーソナライズを行うことが重要である. そこで本研究では, 小説に対する感想の生成における LLM のパーソナライズタスクに注目する. LLM のパーソナライズとは, 対象とするユーザの特徴や好みなどの情報を分析し, それらをもとに LLM を調整することで, 出力をユーザごとに最適化することである.

本タスクでは, ユーザの特徴が反映された感想が生成されることが望ましい. 小説に対する感想におけるユーザの特徴として, 小説の内容に対する考えや価値観, 読んだ際に抱く感情等の要素が異なることが挙げられる. したがって, これらの要素が反映された感想を生成するためのシステムを構築する必要がある.

近年, LLM のパーソナライズタスクを解決するための手法として, Lewis ら [10] が提案した RAG（図 2）を用いた生成内容の専門特化が注目されている. RAG は, 外部の知識ベースに対してユーザの入力をクエリとして検索を行い, 検索結果の情報に基づいて LLM に回答を生成させることで, 事前知識として含まれていない情報を反映した回答を獲得可能にするフレームワークであり, データベースからクエリに基づいて関連性の高い情報を抽出する Retriever, Retriever で抽出した情報をもとに回答を生成する Generator, Retriever と Generator に対して追加情報を与える Augmentation の 3 つの要素から構成される [11] . RAG を用いることで, LLM が事前知識として有していない外部知識に基づいた LLM の調整と回答の生成が可能である. また, RAG の構成要素を任意のフレームワークに追加することで, 外部知識に最適化された手法の構築が可能である.

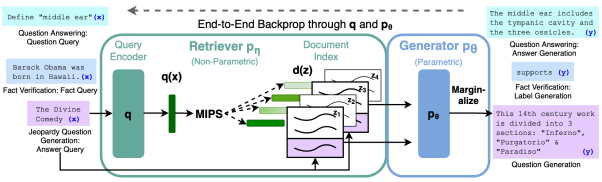


図 2　RAG の概要

LLM のパーソナライズタスクでは, 入力としてパーソナライズ先の情報が与えられる. 例えば, 企業の業務に特化した LLM を構築する場合, 対象とする企業の情報やこれまでの業務履歴を入力とすることでパーソナライズ可能である. また, ユーザに特化した LLM を構築する場合には, 対象となるユーザのプロフィールや行動履歴を入力とすることでパーソナライズ可能である. このように, パーソナライズ対象に関する情報を入力することで, パーソナライズされた LLM を構築できる.

提案手法では, LLM のパーソナライズタスクにおける入力を小説本文およびレビュー, 出力を小説に対する感想とすることで, ユーザの小説解釈に基づいた, 小説に対する感想の生成を実現する. ここで入力するレビューは, パーソナライズを行う対象であるユーザと考え方や立場が近いものを利用する. 本研究では考え方や立場が近い複数のレビューを, あるユーザを仮定したレビュー集合として扱い, 入力として用いる.

ここで, LLM にレビュー集合を与えるだけでは表層的な感情が反映されるだけであり, 小説を解釈したとは言い難い. そこで, 本研究では人間の小説解釈過程を再現した LLM システムを構築し, 人間らしい解釈に基づいたパーソナライズを行う.

本研究の貢献は, 次の通りである.

- 人間の小説解釈過程を再現した, LLM による小説解釈のための新しいフレームワークを示した.
- 任意のレビュー集合をもとに LLM による質問-回答ループを行い, 質問-回答集合を構築することで, LLM のパーソナライズが可能であることを示した.
- 提案手法を用いることで, LLM が前提として有するバイアスを変更して, パーソナライズされた文章を生成することが可能であることを示した.

2　関 連 研 究

2.1　人の感情形成

本研究では, 人が小説体験をしてから感情が形成されるまでの過程を LLM に再現させることによって, ユーザ情報を作成し, パーソナライズを行う. これまでに, 人の感情形成に関して, いくつかの研究が行われている. Ellis [12] が提案した心理療法理論である ABC 理論では, Activating events（出来事）をどのような Belief（思考・信念）を通して受け取るかによって, Consequences（感情・行動）が決定すると定義されている. すなわち, 出来事から直接的に感情が形成されるのではなく, 出来事を人間がどう認知するのかによって形成される感情が異なることを示している. これは小説の読書において, 小説を読むという体験を読み手がどう認知して解釈するのかによって, そこで形成される感情が異なることを示唆している. また, 三和 [13] は小説を対象とした読後の感情形成において, 読者のパーソナリティ特性, 認知的評価, 読後の感情状態の 3 つを構成概念として提案し, これらの構成概念の間に因果関係が存在していることを示している. このことから, 小説を読んだ後に抱く感情を明らかにする上で, 読み手の感情や思考における特徴や, 登場人物の感情をどのように解釈しているかということを明らかにすることは重要であると考えられる.

本研究では, ABC 理論をもとに小説体験から感情形成までの過程を, 小説の読書（Activating events）, 小説の解釈（Belief）, 感情・レビュー（Consequences）に分解し, この過程を LLM を用いてレビューから遡ることで, ユーザが小説をどう解釈したのかということを明らかにするためのシステムを提案する.

2.2　LLM による文章解釈

本研究では, ユーザが小説を読んだときに, その内容を解釈する過程を再現させることで LLM のパーソナライズを行う. これまでに, LLM の文章解釈能力や文章解釈における特性を明らかにするための研究がいくつか行われている. Zhang ら [14] は文章中に含まれる単語の重要度を分類する実験を人間と LLM のそれぞれに対して行い, 人間と LLM の重要度分類に相関があ

ることを示した. また, Chakrabarty ら [15] が提案した言葉の持つ暗喩的なイメージを画像と結びつけるためのフレームワークである LLM-Diffusion では, Chain of Thought を用いることで言葉の持つ暗喩的な意味を LLM に説明させている. 生成した画像が暗喩的なイメージを含んでいるかという評価において, 合意データセットだけで Fine-tuning した場合よりも優れていることを示した. したがって, LLM は暗喩的な表現が含んでいる意味を解釈し, 説明することが可能であると考えられる.

これらの先行研究から, LLM による文章の解釈方法には人間における解釈方法と類似する点が存在しているため, 人間が小説という文章を読む際の解釈過程を LLM で再現できる可能性があると考えられる.

2.3 LLM による人間の思考や行動の再現

人間の価値観に沿う文章や個人にパーソナライズされた文章を LLM に生成させるための研究が行われている. Christiano ら [16] が提案した RLHF では, LLM が生成した文章に対して人間が人手でランク付けを行い, 報酬が最大化するように強化学習を行うことで LLM を Fine-tuning している. Jang ら [17] が提案した Personalized Soups では, 個人のフィードバックに対して RLHF を用いることで, 個人の好みに適応した文章を生成している. Shao ら [18] が提案した Character-LLM では, 特定の人物に関するプロフィールを収集し, そこから抽出される対話や行動の展開が記述されている文章を用いて LLM を Fine-tuning することで, その人物を模倣した LLM を構築している. これらのフレームワークでは, 共通してパーソナライズの対象である個人に関する情報をもとに LLM を調整することで, 生成内容の最適化を可能にしており, 学習データとしてパーソナライズ対象のユーザ情報を用いることは有効であると考えられる.

また, Lee ら [19] は複数の心理療法理論に基づいて, 感情の識別, 感情への共感, ユーザ理解の探求, 対話という 4 つのステップに分解した段階的な推論を通して LLM に回答を生成させる CoE を提案し, 共感性を有する回答生成が可能であることを示した. したがって, ユーザと語り合う対話相手として LLM エージェントを構築する上で, 心理療法理論に基づいた Chain of Thought による段階的な推論は有効であると考えられる.

本研究では, パーソナライズを行う対象であるユーザの情報として, そのユーザの考え方や価値観に近いレビューを用いて LLM を調整することで, 生成内容のパーソナライズを行う. しかし, 小説の感想を語り合えるエージェントを構築する上で, ユーザごとに LLM を Fine-tuning するのは, 計算機資源および時間資源の観点で容易ではない. したがって, 本研究では Fine-tunning ではなく, LLM に入力するプロンプト中に few-shot でユーザ情報を与える in-Context Learning を用いることで, LLM のパーソナライズを行う. また, 心理療法理論の 1 つである ABC 理論に基づき, 感情の予測, 自己質問応答によるユーザの考え方や価値観の理解, 対話の 3 ステップに分解して感想を語り合うための LLM エージェントを構築する.

3 提 案 手 法

3.1 提案手法の概要

本研究では, LLM に外部知識を反映してパーソナライズを行う手法である RAG [10] を拡張し, LLM における小説解釈のパーソナライズを行う手法として Replication of User's Process to Interpret Novels（RUPIN）を提案する. さらに, RUPIN を用いた下流タスクの 1 つとして, 小説に対するユーザの考え方を反映した感想の生成を定義し, ユーザが小説の感想を語り合えるエージェントの構築手法を提案する.

ここで提案するフレームワークは, 扱うコンテンツの種類や入力するユーザ情報を変更することで, 小説以外のコンテンツに対する利用や, 個人単位ではなく, ある特徴を持った集団に対するパーソナライズを行うことも可能であると考えられる.

3.2 RUPIN

RUPIN（図 3）は感情予測部, 質問生成部, 回答生成部の 3 つのモジュールから構成される. 感情予測部でレビューをもとに予測した感情に対して, 質問生成と回答生成を繰り返し行うことで解釈過程に対応する質問-回答対を獲得する. ここで, この質問生成と回答生成を繰り返し行うことを質問-回答ループと呼ぶ.

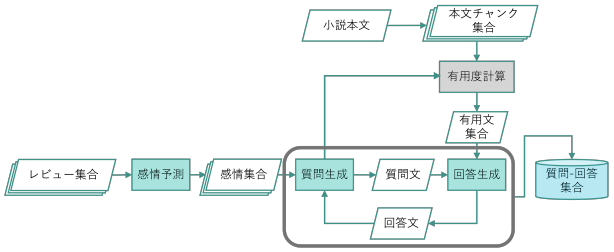


図 3　RUPIN の概要

RUPIN への入力は, レビュー集合 $R = \{rvw_1, ..., rvw_{|R|}\}$ と小説本文 t である. 小説本文 t を任意の長さで分割したものを本文チャンク集合 $T = \{txt_1, ..., txt_{|T|}\}$ とする. ここで, レビュー集合 R に含まれるレビューは, 小説本文 t に対するレビューの中で, パーソナライズを行う対象であるユーザの考え方や立場に近い内容が書かれているレビューであるとする.

3.2.1 感情予測部

感情予測部では, レビューから読み取ることのできるユーザの感情とその根拠の抽出を行う. レビュー $rvw \in R$ から読み取れる感情からなる集合を感情集合 $E = \{emt_1, ..., emt_{|E|}\}$ とする. レビュー $rvw \in R$ を含んだプロンプト（図 4）を LLM に入力し, レビュー感情集合 $RE_{rvw} = F(rvw)$ を獲得する. ここで, $F(x)$ はレビューから感情を予測可能な関数であり, $RE_{rvw} \subseteq E$ である. これをレビュー集合 R に属する全てのレビュー rvw に対して行うことで, 感情集合 $E = \sum_{rvw \in R} RE_{rvw}$ を獲得する.

<system>
あなたはプロの心理学者です。

<user>
命令文:
以下のレビューから読み取れる感情とその根拠を 5 つ教えてください。

レビュー:
{レビュー}　#該当のレビューを入力

出力形式：
[番号]. [感情]：[根拠]

図 4　感情予測部で入力するプロンプト

3.2.2　質問生成部

質問生成部では, 入力文に対して「なぜそう考えるのか」という解釈過程を明らかにするための質問文の生成を行う.

質問-回答ループにおいて, 1 回目の質問生成では, レビュー $rvw \in R$ とレビュー感情集合 RE_{rvw} を含んだプロンプト（図 5）を LLM に入力し, 質問文 $q_{rvw,1}$ を生成する.

いま, 質問文 q に対する回答文を $ans(q)$ とする. 2 回目以降の質問生成では, 回答生成部で生成された回答文 $ans(q_{rvw,i})$ を含んだプロンプト（図 6）を LLM に入力し, 質問文 $q_{rvw,i+1}$ を生成する.

<system>
あなたはプロのインタビュアーです。

<user>
命令文:
以下のレビューから読み取れる感情について、なぜそういう感情になるのかを簡潔に 1 つ質問してください。

レビュー:
{レビュー}　#該当のレビューを入力

感情：
{感情}　#該当の感情を入力

図 5　質問生成部 (1 回目) で入力するプロンプト

<system>
あなたはプロのインタビュアーです。

<user>
命令文:
以下の説明について、なぜそういう考えるのかを簡潔に 1 つ質問してください。

説明:
{回答}　#該当の回答を入力

図 6　質問生成部 (2 回目以降) で入力するプロンプト

3.2.3　回答生成

回答生成部では, 入力文に対して小説本文を根拠とした回答文の生成を行う. まず, 質問文 q と本文チャンク $txt \in T$ に対してコサイン類似度を計算し, コサイン類似度の高い本文チャンク k 個を有用文集合 $UT(q,k)$ とする. その後, レビュー rvw , 有用文集合 $UT(q,k)$, 質問文 q を含んだプロンプト（図 7）を LLM に入力し, 回答文 $ans(q, rvw, k)$ を生成する.

<system>
あなたは以下のレビューを書いたユーザです。

レビュー：
{レビュー}　#該当のレビューを入力

<user>
この小説を読んでどう思いましたか。

<assistant>
私はこの小説を読んで以下のように感じました。

{レビュー}　#該当のレビューを入力

<user>
命令文:
小説本文の一部を参考にして、質問に対してとても簡潔に答えてください。

小説本文の一部:
・{本文チャンク 1}
・{本文チャンク 2}
・{本文チャンク 3}
・{本文チャンク 4}　#該当の本文チャンクを入力

質問：
{質問}　#該当の質問を入力

図 7　回答生成部で入力するプロンプト

3.2.4　質問-回答ループ

質問生成と回答生成を繰り返すことで質問-回答ループを行う. すなわち, 1 つの感情に対して質問生成（1 回目）→有用度計算→回答生成→質問生成（2 回目）→有用度計算→回答生成→... というループを行い, これを感情 $emt \in RE_{rvw}$ それぞれについて行う. さらに, この操作をレビュー $rvw \in R$ それぞれについて行う. したがって, 質問-回答ループにおけるループ回数を $Roup$ とすると, 質問-回答ループを行うことによって, 1 つの感情に対して $Roup$ 組の質問-回答対 $(q, ans(q))$ が得られるため, 全体では $Roup \times |RE| \times |R|$ 組の質問-回答対 $(q, ans(q))$ が得られる.

3.2.5　下流タスクへの利用

質問-回答ループによって獲得した全ての質問-回答対 $(q, ans(q))$ を含む質問-回答集合 $QA = \sum_{rvw \in R} \sum_{i=1}^{|RE|} \sum_{j=1}^{Roup} \{q_{rvw,i,j}, a_{rvw,i,j}\}$ をパーソナライズの対象であるユーザの小説解釈過程を表現したものであると仮定することで, 種々の下流タスクへ利用することが可能である.

3.3　感想を語り合えるエージェント

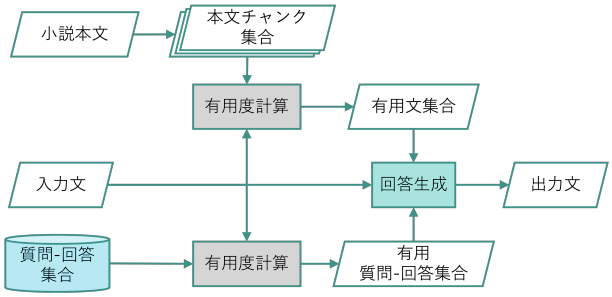


図 8　感想を語り合えるエージェントの概要

入力文に対して, レビュー集合から読み取れる考え方や立場を反映した文章の生成を行う（図 8）. まず, 入力文 $input$ と本文チャンク $txt \in T$ に対してコサイン類似度を計算し, コサイン類似度の高い本文チャンク k 個を有用文集合 $UT'(input, k)$ と

する. また, 入力文 *input* と回答文 $ans(q, rvw, k)$ のコサイン類似度を計算し, コサイン類似度の高い回答文に対応する質問-回答対 l 個を有用質問-回答集合 $UQA(input, l)$ とする. その後, 有用文集合 $UT'(input, k)$, 有用質問-回答集合 $UQA(input, l)$, 入力文 *input* を含んだプロンプト（図 9）を入力し, 目的の文章を生成する.

```
<system>
あなたはプロの書評家です。

<user>
命令文:
小説本文の一部とあなたが今までしてきた受け答えを参考にして、入力に対してとても簡潔に答えてください。

小説本文の一部：
・{本文チャンク 1}
・{本文チャンク 2}
・{本文チャンク 3}
・{本文チャンク 4}  #該当の本文チャンクを入力

あなたが今までしてきた受け答え:
・質問-{質問 1}, 回答-{回答 1}
・質問-{質問 2}, 回答-{回答 2}
・質問-{質問 3}, 回答-{回答 3}
・質問-{質問 4}, 回答-{回答 4}
・質問-{質問 5}, 回答-{回答 5}
  #該当の質問-回答対を入力

入力：
{入力}  #エージェントへの入力
```

図 9　エージェントに入力するプロンプト

4 評価実験

4.1 データセット・前処理

青空文庫[1]に掲載されている太宰治の『走れメロス』, カクヨム[2]に掲載されている岡本圭地の『俺の悪友』の 2 作品を対象として実験を行った. 『走れメロス』は 9,855 字, 『俺の悪友』は 9,716 字の短編小説作品である. なお, 小説本文の抽出精度による性能への影響を最小化するために, 前述した 1 万字程度の短編小説作品である 2 作品を対象とした. また, 『走れメロス』に関するレビューは, 小説レビューサイトである読書メーター[3]に掲載されているレビュー, 『俺の悪友』に関するレビューは, カクヨムに掲載されているレビューから収集した.

本実験で使用したモデルの設定は表 1 の通りである. 感情予測, 質問生成, 回答生成, エージェントの構築には OpenAI[4] が提供している LLM である GPT-4 を使用し, 類似度計算時に用いる分散表現の獲得には OpenAI が提供している Embedding モデルである text-embedding-ada-002 を使用した. なお, 今回使用した GPT-4 は 2022 年 1 月までのデータを学習しているため, 『走れメロス』の知識は有するが, 『俺の悪友』の知識は有していないと考えられる.

小説本文と収集したレビューに対して前処理として, 空白, 改行, 絵文字の削除を行った. また, 小説本文は句点区切りで 1 チャンクあたり 500 字以下となるようにチャンク分けを行った. その結果, 『走れメロス』と『俺の悪友』はともに 21 個のチャンクに分けられた. 『走れメロス』と『俺の悪友』に関するレビューそれぞれについて, 考え方や立場が近いものを 5 つ集めたレビュー集合を 2 つずつ人手で作成した. なお, 作成の際には同一作品の 2 つのレビュー集合間で考え方や立場が異なるようにした. また, 質問-回答ループにおけるループ回数は $Roup = 3$ とした.

表 1　使用モデル

LLM モデル	gpt-4
temperature	0.7
Embedding モデル	text-embedding-ada-002

4.2 実験方法

Lee ら [20] は文章生成タスクにおいて手動評価を行う上で重要な指標として, Fluency, Naturalness, Quality, Meaning preservation, Relevance などを示した. 本研究における文章生成では, Quality および Relevance として任意のレビュー集合に対してパーソナライズされた文章生成ができているか, Meaning preservation として小説本文の内容に基づいた文章生成ができているかが重要な評価基準であるといえる. 以上を考慮して, 本実験では事前に人手で作成した対象作品に関する入力文各 8 個に対して, エージェントが生成した回答文について,

- どちらのレビュー集合に属するか
- ハルシネーションが含まれているか

の 2 項目を被験者 10 人に評価させた. 各項目の評価は 5 段階のリッカート尺度によって行った.

また, 以下のベースラインとの比較を行った.

- PRM：有用文集合のみを含むプロンプトを与えた場合
- PRM+REV：有用文集合とレビュー集合を含むプロンプトを与えた場合

4.3 『走れメロス』を対象とした評価実験の結果と考察

4.3.1 定量的結果

「どちらのレビュー集合に属するか」という評価項目について, 「A に属する」および「どちらかというと A に属する」を A, 「B に属する」および「どちらかというと B に属する」を B と分類した場合の正解率は表 2 の通りであった. レビュー全体およびレビュー集合 B の正解率では, PRM+REV および提案手法のどちらも, 無作為に分類した場合とカイ二乗検定を行った結果, それぞれ $\chi^2(1) = 16.2$, $\chi^2(1) = 9.8$ であり, 有意水準 5%で有意差があった. 一方, レビュー集合 A の正解率では, 提案手法のみが無作為に分類した場合と, カイ二乗検定において有意水準 5%で有意差があった.

表 2　『走れメロス』における分類評価の正解率 (%)

手法	全体	レビュー集合 A	レビュー集合 B
PRM+REV	63.13	53.75	72.50
提案手法	62.50	57.50	67.50

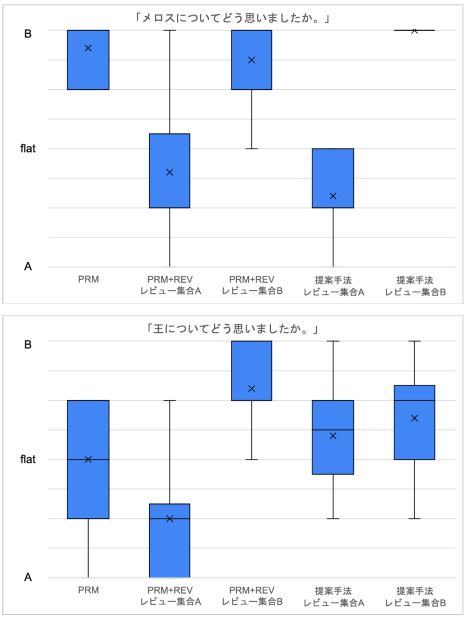
PRM では, レビュー集合 A に分類されたものが 14.4%, レ

1：https://www.aozora.gr.jp/
2：https://kakuyomu.jp/
3：https://bookmeter.com/
4：https://openai.com/

ビュー集合 B に分類されたものが 66.9%であったことから, 本研究で使用した GPT-4 にはレビュー集合 B に属する考え方や価値観のバイアスがかかっていたと考えられる. そのため, PRM+REV および提案手法のどちらについても, レビュー集合 A よりレビュー集合 B の正解率の方が高かったと考えられる. また, レビュー集合 A の分類について, 提案手法を用いた場合の正解率が有意に高かったことから, 提案手法によって GPT-4 が前提として有するバイアスを変更し, レビュー集合が有するバイアスを適用することができたといえる.

また, 質問ごとに分類評価の分布を箱ひげ図を用いて表すと図 10 の通りであった. 「結婚式の場面についてどう思いましたか。」や「メロスは勝手な行動が目立つと思いませんか。」などの分布から, PRM+REV よりも提案手法の方が良かったものは提案手法における分布の範囲がレビュー集合 A の方向に狭くなっており, PRM+REV よりも提案手法の方が悪かったものは提案手法における分布の範囲がレビュー集合 A の方向に広がっている傾向がある. このことから, 文章を生成する際に GPT-4 が有するバイアスが, few-shot で与えた有用質問-回答集合以上に影響したことで, 提案手法において十分なパーソナライズ精度が得られなかったことが考えられる. したがって, LLM にパーソナライズの対象であるユーザの情報を与えるだけではなく, LLM が有するバイアスやそのバイアスを形成している知識を取り除くことが必要であると考えられる.

学習済みの知識を LLM から取り除く手法として, Eldan ら [21] は取り除きたい対象のデータを用いて LLM に忘れさせるべきトークンを特定し, そこに含まれる特有の表現を一般的なもの置き換えて Fine-tuning することで, 対象に関する回答を生成する LLM の能力を消去できることを示している. また, Zhang ら [22] は LLM が有する知識の中で編集したい知識を決め, それに関連するデータを収集した後, モデルの編集を行うことで知識編集を行うことが可能であることを示している. これらの手法を取り除きたいバイアスに対して適用することで, パーソナライズ精度を向上することができると考えられる.



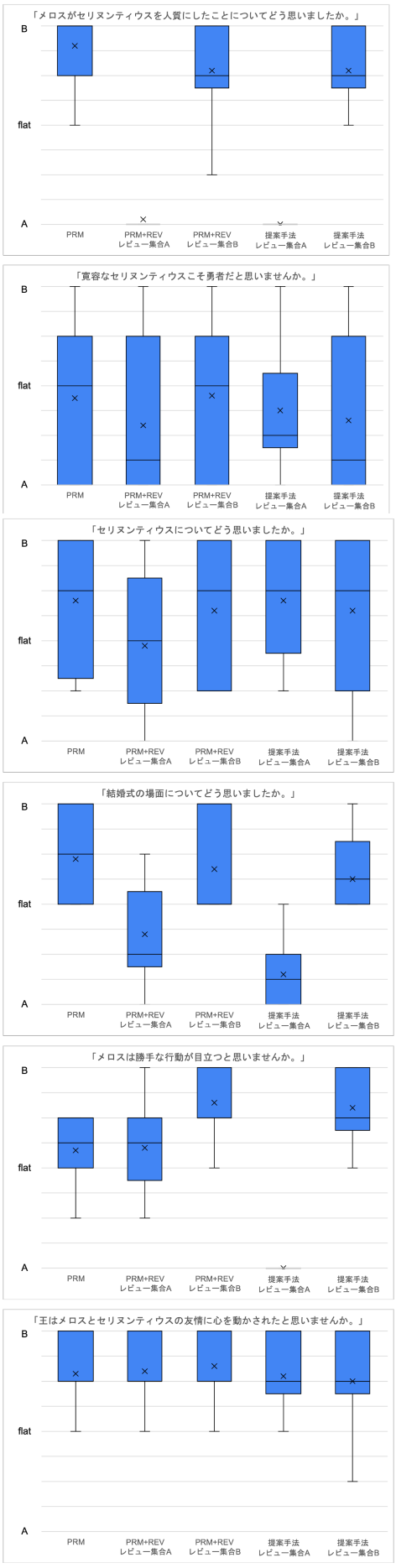


図 10　『走れメロス』における分類評価の分布

次に,「ハルシネーションが含まれているか」という評価項目について, 5 段階評価の平均スコアと分散は表 3 の通りであった. なお, スコアは 1〜5 の間の値を取り, スコアが低いほどハルシネーションが含まれていないことを表す. すべての生成文に対するスコアでは, PRM, PRM+REV, 提案手法の間で, 対応のある t 検定において有意差は見られなかった. 一方, レビュー集合 A のスコアでは, 対応のある t 検定の結果, $p = 0.047$ であり, PRM+REV より提案手法の方が有意水準 5%で有意に低かった.

表 3　『走れメロス』におけるハルシネーション評価の平均スコアと分散

手法	全体		レビュー集合 A		レビュー集合 B	
	μ	σ^2	μ	σ^2	μ	σ^2
PRM	2.994	2.206	-	-	-	-
PRM+REV	3.138	2.194	3.188	2.327	3.088	2.055
提案手法	3.006	2.231	2.988	2.037	3.025	2.424

PRM と提案手法のスコアに有意差が見られなかったことから, 少なくとも RUPIN において GPT-4 自身が生成した質問-回答対を GPT-4 に対して few-shot で与える場合に, この処理によるハルシネーションの増大が発生しているとは言えない. また, レビュー集合 A のスコアについて, PRM+REV より提案手法の方が有意に低く, 上で述べた通り GPT-4 が前提としてレビュー集合 B に属するバイアスを有していたことを考慮すると, GPT-4 が有していないバイアスがかかったレビュー集合に対してパーソナライズを行う場合, 提案手法が有効であるといえる. 一般に LLM を構築する際, 巨大言語コーパスを利用して損失関数を最小化するように学習を行うことで, 尤もらしい文章を生成するようにされていることを踏まえると, 特に対象の作品に関するレビューとして少数派であるものや, 広く言われているわけではない個人的な見解であるものに対するパーソナライズにおいて, 提案手法が有効であると考えられる.

4.3.2　定性的結果

「どちらのレビュー集合に属するか」という評価項目について, 正しく分類された生成文と誤って分類された生成文の例を図 11 に示す. 正しく分類された生成文は, レビュー集合 A をもとに生成した文章とレビュー集合 B をもとに生成した文章の内容に明らかな差が見られる一方, 誤って分類された生成文は, レビュー集合 A, B のそれぞれをもとに生成した文章の内容の類似度が高かった. また, 誤って分類された生成文に対応する入力文の内容がレビューの中であまり言及されていない人物またはレビュー集合 A, B 間で意見や立場に差がない人物に関するものであったことから, レビューで直接言及されていないまたはレビュー集合間で差異がない内容に関する文章生成のパーソナライズは失敗しやすいと考えられる.

レビュー集合間で差異がないことに起因する誤分類は, 本研究で扱っていないレビュー抽出をユーザの考え方や価値観の情報に基づいて適切に行うことで改善できると考えられる. また, レビューの中で直接言及されていないことに起因する誤分類は, 質問生成部に作品内に登場する人物や場面に着目したコンテンツベースの質問生成機構を追加し, 質問-回答集合に含まれる内容の網羅度を高くすることで改善できると考えられる.

正分類例

入力：メロスについてどう思いましたか。
出力：メロスはその信念と義務感を貫き、友人のセリヌンティウスを救うために全力を尽くす姿に感銘を受けました。彼が困難に立ち向かい、約束を果たすために命を危険にさらす強さは、読者自身にも勇気と信念を持つことの重要性を教えてくれ、非常に魅力的なキャラクターだと感じました。

入力：メロスは勝手な行動が目立つと思いませんか。
出力：はい、メロスの行動は自分の信念を通そうとするあまり、他人を無視したり危険に晒す側面があり、それは自己中心的な行動と受け取れます。

誤分類例

入力：セリヌンティウスについてどう思いましたか。
出力：セリヌンティウスは、友人であるメロスへの深い信頼と忠誠心を備えた人物として描かれており、その姿勢には深い尊敬の念を抱いています。彼はメロスを無条件で信じ、自身が人質となる運命を受け入れました。このような彼の行動は、友情と信頼の強さを象徴しており、その強い精神力に感銘を受けました。

入力：王はメロスとセリヌンティウスの友情に心を動かされたと思いませんか。
出力：はい、王は最終的に二人の友情と忠誠に感銘を受け、彼らの仲間になりたいと願っています。

図 11　『走れメロス』における分類評価の正分類例と誤分類例

4.4 『俺の悪友』を対象とした評価実験の結果と考察

4.4.1　定量的結果

『走れメロス』の場合と同様にして分類した場合の正解率は表 5 の通りであった. レビュー集合 A の正解率のみ, PRM+REV および提案手法のどちらも, 無作為に分類した場合とカイ二乗検定を行った結果, それぞれ $\chi^2(1) = 18.05$, $\chi^2(1) = 22.05$ であり, 有意水準 5%で有意差が見られた. 一方, PRM+REV と比較して提案手法は, レビュー全体およびレビュー集合 B の正解率に対する対応のある t 検定を行った結果, それぞれ $p = 0.0152$, $p = 0.0024$ であり, 有意水準 5%で有意に分類評価が良かった.

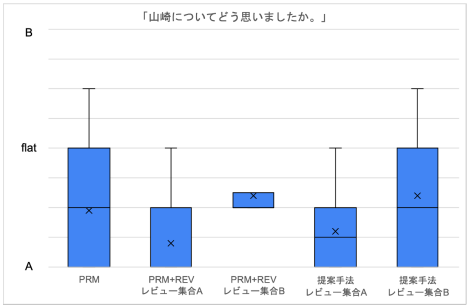
表 5　『俺の悪友』における分類評価の正解率 (%)

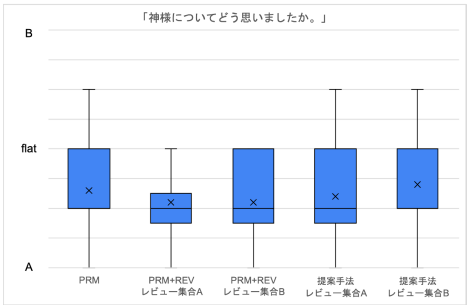
手法	全体	レビュー集合 A	レビュー集合 B
PRM+REV	45.00	73.75	16.25
提案手法	55.63	76.25	35.00

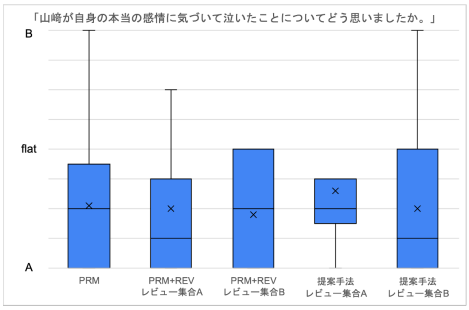
PRM では, レビュー集合 A に分類されたものが 61.3%, レビュー集合 B に分類されたものが 16.9%であったことから, 本研究で使用した GPT-4 にはレビュー集合 A に属する考え方や価値観のバイアスがかかっていたと考えられる. そのため『走れメロス』の場合と同様に, PRM+REV および提案手法のどちらについても, レビュー集合 B よりレビュー集合 A の正解率の方が高かったと考えられる. レビュー集合 B の分類について, 提案手法を用いた場合の正解率が有意に高かったことから, GPT-4 が知識を有していない作品を対象とした場合でも, 提案手法によって GPT-4 が有するバイアスをレビュー集合が有しているバイアスに変更することができたといえる.

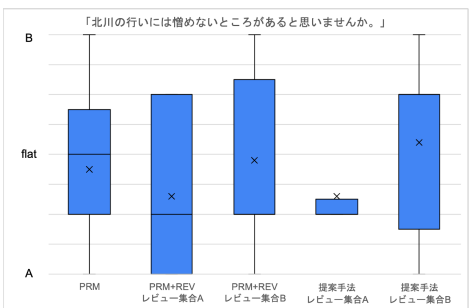
また, 質問ごとに分類評価の分布を箱ひげ図を用いて表すと図 13 の通りであった. 『走れメロス』の場合と比較して, 分類

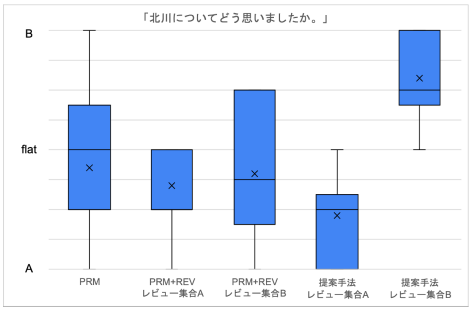
評価の分布が GPT-4 が有しているバイアスの方向へより偏っていることがわかる. したがって, LLM が知識を有している作品を対象とした場合よりも知識を有していない作品を対象とした場合の方が, LLM が有するバイアスの影響が大きいことが予想される. このことから, 文章を生成する際に LLM が有するバイアスが few-shot で与えた有用質問-回答集合以上に影響したことで, 提案手法において十分なパーソナライズ精度が得られなかったことが考えられる. 絶えず新しい小説が誕生し, ユーザが新作の小説を読む機会も多いことを考慮すると, 小説の感想を語り合うエージェントにおいて扱う対象となる小説は LLM が知識を有していない場合も多いと考えられる. したがって,『俺の悪友』における結果は前項で述べた LLM が有しているバイアスを消去または編集することの必要性を支持しているといえる.

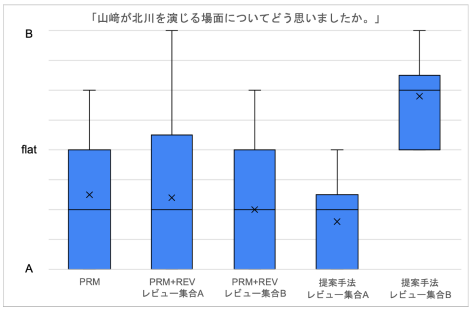


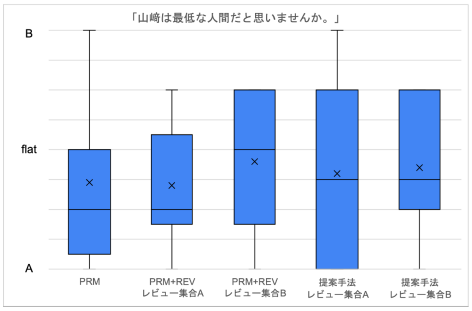












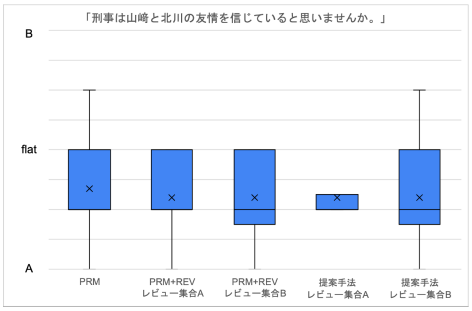


図 13　『俺の悪友』における分類評価の分布

次に,「ハルシネーションが含まれているか」という評価項目についての平均スコアと分散は表 6 の通りであった. 生成文全体に対するスコアでは, PRM, PRM+REV, 提案手法で, 対応のある t 検定において有意差は見られなかった. また, レビュー集合 A およびレビュー集合 B のスコアでは, PRM+REV と提案手法の間で, 対応のある t 検定において有意差は見られなかった.

PRM と提案手法に有意差がなかったことから,『走れメロス』の場合と同様に, 質問-回答対を few-shot で与える場合に, ハルシネーションの増大が発生しているとはいえない.

表 6　『俺の悪友』におけるハルシネーション評価の平均スコアと分散

手法	全体		レビュー集合 A		レビュー集合 B	
	μ	σ^2	μ	σ^2	μ	σ^2
PRM	3.000	1.863	-	-	-	-
PRM+REV	2.981	1.931	3.050	1.873	2.913	1.980
提案手法	2.975	1.562	3.038	1.561	2.913	1.555

4.4.2 定性的結果

「どちらのレビュー集合に属するか」という評価項目について, 正しく分類された生成文と誤って分類された生成文の例を図 14 に示す. 正しく分類できた生成文は, レビュー集合 A をもとに生成した文章の内容とレビュー集合 B をもとに生成した文章の内容の間で注目している観点に大きい差が見られる一方, 誤って分類された生成文は, レビュー集合 A, B のそれぞれをもとに生成した文章の内容の間で細かい立場や価値観の違いは存在しているが, 全体としての観点の類似度が高いものが多かった. また, 『走れメロス』の場合と同じく, 誤って分類された生成文に対応する入力文の内容がレビューの中であまり言及されていない人物に関するものであったことから, レビューで直接言及されていない内容に関する文章生成のパーソナライズでは性能が悪くなると考えられる.

細かな立場や価値観の違いを反映できなかったことに起因する誤分類はレビュー集合の特徴よりも, 一般的な視点に基づいた内容が生成した文章に含まれていたため発生したと考えられる. この誤分類は, エージェントの最終的な回答として出力する前に, 生成した内容にパーソナライズの対象であるレビュー集合の特徴が大きく反映されているかを LLM 自身にフィードバックさせ, そのフィードバック内容をもとに生成内容に変更を加えるような一種の自己改善を行うことによって改善が可能であると考えられる. また, LLM が知識を有していない作品を対象にした場合にも, レビューの中で直接言及されていないことに起因する誤分類が発生しているため, 質問生成部にコンテンツベースの質問生成機構を追加し, 質問-回答集合に含まれる内容の網羅度を高くすることの重要性が示唆された.

正分類例
入力：北川についてどう思いましたか。 出力：北川は一貫して自己中心的で無責任な行動をとるキャラクターで、その行動が読者の道徳観に反するものであったため、彼に対する評価は低かったです。それにもかかわらず、彼の人間性が物語に緊張感と予測不能な要素をもたらし、読者を引きつける一方で、倫理的な観点からは彼を容認することはできませんでした。 入力：山崎が北川を演じる場面についてどう思いましたか。 出力：山崎が北川を演じる場面は物語に深みを加え、予測不能な展開が読者を引きつける要素となりました。また、山崎の視点から見た北川のキャラクターがより鮮明に描かれ、二人の友情が深く描写される独特の手法だと感じました。
誤分類例
入力：山崎が自身の本当の感情に気づいて泣いたことについてどう思いましたか。 出力：それは非常に感動的な瞬間で、山崎自身が彼の本当の感情、つまり友人である北川に対する深い愛情と悲しみに気づいた瞬間でした。これは彼のキャラクターの深みを増し、読者に強い感情的な反応を引き出す力があります。 入力：山崎は最低な人間だと思いませんか 出力：それは主人公の視点から見た北川の評価であり、彼の人間性全体を示すものではないと感じます。

図 14 『俺の悪友』における分類評価の正分類例と誤分類例

5 ま と め

本研究では, 小説に対する感想の生成における LLM のパーソナライズタスクを扱った. 任意のレビュー集合に対して感情の予測を行った後, 質問-回答ループによって小説解釈過程を明らかにすることで LLM のパーソナライズを行う手法である RUPIN と小説の感想を語り合えるエージェントの構築手法を提案した.

評価実験の結果, レビュー集合の考え方や価値観を反映した文章の生成が可能であることが示された. 特に, LLM が前提として有していないようなバイアスのかかったレビュー集合に対するパーソナライズにおいて, レビュー集合を few-shot で与えるよりも性能が良いことが示された. また, 任意のレビュー集合に対してパーソナライズを行う下流タスクにおいて, RUPIN を用いたシステム構築が有効であることが示唆された.

今後は, コンテンツベースの質問生成機構, レビュー抽出機構, レビューの反映度や感想を語る相手としての自然さに基づいた自己改善機構の追加による性能改善に取り組みたい. また, RUPIN を利用した他の下流タスクとして紹介文生成のパーソナライズ, 説明性を有する推薦システム, 未読小説に対するレビュー予測などが考えられるため, それらのシステム構築および評価実験にも取り組みたい.

文 献

[1] 全国出版協会出版科学研究所. 2023 年版 出版指標 年報, 2023.
[2] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. A computational framework for behavioral assessment of llm therapists, 2024.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.
[4] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.
[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.
[7] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.
[8] OpenAI. Gpt-4 technical report, 2023.
[9] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.

[11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[12] Albert Ellis. Rational psychotherapy. *TACD Journal*, Vol. 17, No. 1, pp. 67–80, 1989.

[13] 三和義秀. 小説を対象とした読後の感情状態形成モデルの研究-読者のパーソナリティ特性と認知的評価に基づいて. 情報知識学会誌, Vol. 23, No. 1, pp. 92–110, 2013.

[14] Yuhong Zhang, Qin Li, Sujal Nahata, Tasnia Jamal, Shih kuen Cheng, Gert Cauwenberghs, and Tzyy-Ping Jung. Integrating llm, eeg, and eye-tracking biomarker analysis for word-level neural state classification in semantic inference reading comprehension, 2023.

[15] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors, 2023.

[16] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[17] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023.

[18] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing, 2023.

[19] Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models, 2023.

[20] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355–368, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.

[21] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023.

[22] Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models, 2024.

対人スキルと対 AI スキルの関係

西岡　竜生†　　若宮　翔子†　　清水　伸幸††　　藤田　澄男††　　荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
†† LINE ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3
E-mail: †{nishioka.ryuki.np4,wakamiya,aramaki}@is.naist.jp, ††{nobushim,sufujita}@lycorp.co.jp

あらまし　人と AI が協働することで，AI 単体を上回るパフォーマンスが発揮されることが注目されている．近年においては，大規模言語モデル (Large Language Model; LLM) やチャットボットなどの対話型 AI が急速に発展し，教育分野等への応用が検討されている．人と LLM の協働において，LLM の性能のみならず LLM を利用するユーザの能力も考慮する必要があると考えられる．しかしながら，ユーザのスキルと LLM のパフォーマンスとの関係は明らかにされていない．そこで，本研究では対人スキルと対 AI スキルが相関するかどうかを分析した．具体的には，ユーザのコミュニケーションスキルの高さと ChatGPT を用いたタスクの成果の関係を分析した．その結果，一部の対人スキルと一部のタスクにおける対 AI スキルに関係が見られた．

キーワード　大規模言語モデル (LLM)，ChatGPT，HCI，対人スキル，対 AI スキル

1　はじめに

人と AI が協働することで，AI 単体を上回るパフォーマンスが発揮されることが注目されている．このような現象はしばしば，フリースタイルチェスにおいて「ケンタウロス」という人と AI の混成チームが AI 単体に勝利したことから「ケンタウロス現象」と呼ばれる [1][2]．このことから，AI の効果的な活用には人間の介入が不可欠であると考えられる．

近年においては，大規模言語モデル (Large Language Model; LLM) を用いた ChatGPT のようなチャットサービスが普及し，教育分野等への応用が検討されている．このような LLM の応用に関して，LLM の性能を評価する研究や，LLM の利用による可能性を探求した研究がなされている．Bubeck らは様々なタスクを通して GPT-4 の性能を調査している [1]．また，Big Five [2] や Myers–Briggs Type Indicator (MBTI) [3], [4] といった心理測定尺度を用いて LLM を評価し，LLM が疑似的なパーソナリティを持つ可能性や，LLM の挙動を理解するためにはこのような疑似的なパーソナリティの理解が必要であることが示されている [5], [6]．このように，LLM の性能に着目した研究は多いが，LLM を利用する人間の能力を考慮した研究は少ない．

そこで，本研究では人間の能力（**対人スキル**）と AI との協働によって高いパフォーマンスを発揮する能力（**対 AI スキル**）の関係について分析を行う．ここで，ChatGPT は対話形式で利用するモデルであることから，人の能力として対人スキルに着目する．具体的には以下の手順で分析を行う．はじめに，対 AI スキルを測定するための，ChatGPT を利用して取り組む知識タスク，言語タスク，計算タスク，問題解決タスクの 4 種類のタスクを設計する．実験では，心理測定尺度を用いて 6 つの対人スキルを測定した後，タスクに取り組む．この際，設問ごとに ChatGPT を利用可能か否かを指定する．その後，タスクごとに人間が独力で取り組んだ場合と ChatGPT を利用して取り組んだ場合の解答スコアに基づき，ChatGPT を利用した際のスコアの向上度合いを求め，対 AI スキルとする．これらの結果から，6 つの対人スキルと 4 つの対 AI スキルの関係を分析する．本研究のコンセプトと可視化のイメージを図 1 に示す．

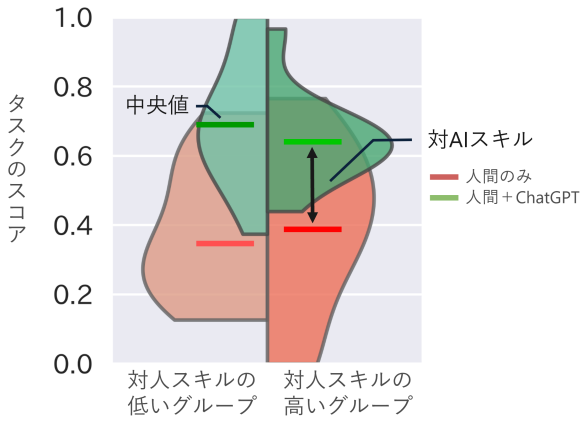


図 1　対人スキルと対 AI スキルの関係分析

2　関連研究

人と LLM の協働によるパフォーマンスが教育分野などで探求されている．Caballé らはソフトウェア教育における対話エージェントの介入がオンライン学習に与える影響を調査している [7]．その結果，対話エージェントによって受講者のオンラ

1：”AI（人口知能）が引き起こす「ケンタウロス現象」に備えていますか？”，株式会社ラーニングデザインセンター，清宮普美代，https://ldcjp.com/event-report/ai（人口知能）が引き起こす「ケンタウロス現象」（2024/2/16 参照）

2：”AI との協調が最適解生む　マルチエージェントシステムが示唆する人間と AI の関係”，立命館大学，https://shiruto.jp/technology/2389/（2024/2/16 参照）

イン学習への関与が向上する同時に脱落率が低下する可能性を示している．Ding らはニュースの見出し文の生成を通してライティングタスクにおける人と AI の協働を分析している [8]．人間が手動で生成した場合，LLM が生成した場合，人間と LLM が協働した場合を比較した結果，LLM 単体でも十分な見出しは生成可能だが，出力の修正等における人間の介入が必要であることを示している．Zhao らは物語生成を通して人と LLM の協働を探求している [9]．LLM のみによって生成された物語と，LLM によって生成された文と人間が生成した文を交互に挿入した物語を比較した結果，LLM のみによって生成された物語の方が理解しやすく好まれる傾向が示されている．

AI の性能に着目した人との協働パフォーマンスの向上が検討されている．Li らは LLM が人間のような推論能力や専門スキルを備えるためのフレームワークを提案し，就職フェアのシミュレーションによって評価している [10]．その結果，タスクワークフローの提案等において有望な性能を示している．Rastogi らは人による LLM の動作の監査・評価をサポートする LLM を提案している．実験の結果，ツールの有効性とより良い監査結果に関連する人間の行動が明らかにされている．

以上のように，人と LLM が協働することによる可能性や，人と協働する LLM の性能に関する研究は多数ある．しかしながら，LLM と協働する人の特徴を考慮した研究は少ない．本研究は LLM ではなく LLM を利用する人に着目し，人と AI の協働による可能性を探求している点において以上の研究と異なる．

3 手　　法

3.1 対人スキル

対人スキルの自己評定尺度として，包括的なコミュニケーションスキルを測定する ENDCOREs [11] を用いる．本尺度はコミュニケーションスキルを対人スキルと基本スキルの階層構造で定義しており，国内外の複数の心理測定尺度を基にコミュニケーションスキルを多面的に評価する．回答には「とても苦手」を 1 点,「とても得意」を 7 点とするリッカート尺度を使用する．使用する質問項目を表 1 に示す．なお，本研究では ENDCOREs の 6 つのメインスキルを対人スキルとし，それぞれ**自己統制スキル**，**表現スキル**，**解読スキル**，**自己主張スキル**，**他者受容スキル**，**関係調整スキル**と呼称する．各対人スキルは，ENDCOREs の各スキルごとの平均によって評価する．

表 1　ENDCOREs [11] の質問項目．回答は「とても苦手」,「苦手」,「やや苦手」,「普通」,「やや得意」,「得意」,「とても得意」の 7 つから当てはまるものを選択する．

メインスキル	サブスキル	項目文
自己統制	欲求抑制	1. 自分の衝動や欲求を抑える
	感情統制	2. 自分の感情をうまくコントロールする
	道徳観念	3. 善悪の判断に基づいて正しい行動を選択する
	期待応諾	4. まわりの期待に応じた振る舞いをする
表現力	言語表現	5. 自分の考えを言葉でうまく表現する
	身体表現	6. 自分の気持ちをしぐさでうまく表現する
	表情表現	7. 自分の気持ちを表情でうまく表現する
	情緒伝達	8. 自分の感情や心理状態を正しく察してもらう
解読力	言語理解	9. 相手の考えを発言から正しく読み取る
	身体理解	10. 相手の気持ちをしぐさから正しく読み取る
	表情理解	11. 相手の気持ちを表情から正しく読み取る
	情緒感受	12. 相手の感情や心理状態を敏感に感じ取る
自己主張	支配性	13. 会話の主導権を握って話を進める
	独立性	14. まわりとは関係なく自分の意見や立場を明らかにする
	柔軟性	15. 納得させるために相手に柔軟に対応して話を進める
	論理性	16. 自分の主張を論理的に筋道を立てて説明する
他者受容	共感性	17. 相手の意見や立場に共感する
	有効性	18. 友好的な態度で相手に接する
	譲歩	19. 相手の意見をできるかぎり受け入れる
	他者尊重	20. 相手の意見や立場を尊重する
関係調整	関係重視	21. 人間関係を第一に考えて行動する
	関係維持	22. 人間関係を良好な状態に維持するように心がける
	意見対立対処	23. 意見の対立による不和に適切に対処する
	感情対立対処	24. 感情的な対立による不和に適切に対処する

3.2 対 AI スキル

対 AI スキルを測定するために，**知識タスク**，**言語タスク**，**計算タスク**，**問題解決タスク**の 4 種類のタスクを設計した（表 2）．なお，解答はすべて自由記述にて行われるものとする．

知識タスク　一般的な知識を問うものであり，解答数により評価する．

言語タスク　単語の意味や概念の理解や言語による説明能力を問うものであり，知能検査に用いられる設問を参考に設計した．表 2 の設問例 1 のような場合は，解答数により評価し，例 2 のような単語の意味を問う設問は，解答と広辞苑に記載された該当単語の意味の類似度に基づき評価する．具体的には，Sentence BERT を用いてベクトル化して求めたコサイン類似度を用いる．

計算タスク　論理的思考力や計算能力を問うものであり，中学

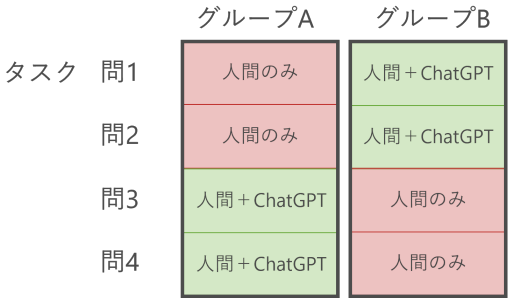


図 2　実 験 設 定

入試における文章題を想定して設計し，1 問 1 点として評価する．

問題解決タスク　明確な正答のない設問を用いて，創造的な思考力や発想力を問うものであり，書籍 [12] を参考に設計し，解答数により評価する．

これら 4 種類のタスクに人間が独力で取り組んだ場合と ChatGPT を利用した場合の解答スコアに基づき，ChatGPT の利用によるスコアの向上度合いを求め，対 AI スキルとする．具体的には，独力で取り組んだ設問の平均と ChatGPT を利用して取り組んだ設問の平均の差を対 AI スキルとして評価する．

3.3　対人スキルと対 AI スキルの関係

対人スキルと対 AI スキルの関係は，スピアマンの順位相関係数を用いて評価する．相関係数は，対人スキルが高いと対 AI スキルも高いのか（AI をうまく使えるかどうか）を意味している．なお，前述したとおり，対人スキルは 6 つのスキルからなり，対 AI スキルは 4 つのタスクごとに定義されるため，24 (=4×6) の相関係数が得られる．考察においては，これは個別に議論していくことにする．

4　実験と考察

4.1　実 験 設 定

対人スキルと対 AI スキルの関係を分析するために実験を行った．本実験には，奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室に所属する成人 12 名（男性 3 名，女性 9 名）が参加した．実験協力者は，心理測定尺度 ENDCOREs（3.1 節）に回答し，4 種類のタスク（3.2 節）として，著者が選定した合計 16 の設問（各タスクにつき 4 問ずつ）に解答した．このとき，設問ごとに ChatGPT を利用せず独力で解くか，必要に応じて ChatGPT を利用可能かを指定する．今回は，設問の難易度による差の影響を軽減するため，図 2 のように実験協力者を 6 名ずつ 2 グループに分け，グループ A は各タスクの設問 1 と 2 を解く際は独力 (w/o AI)，設問 3 と 4 を解く際は ChatGPT を利用可 (w AI) とし，グループ B はこれと反対の設定とした．なお，ChatGPT の利用知識を統一するため，実験協力者はプロンプトガイドラインを確認してからタスクに取り組んだ．このプロンプトガイドラインは，OpenAI の提供す

図 3　プロンプトガイドラインと解答フォームとして用いた Google Form のスクリーンショット．

るドキュメント [3]を参考に著者らが作成したものであり，プロンプトの種類や工夫の仕方を記載している．プロンプトガイドラインと解答フォームの例を図 3 に示す．

これらに加え，実験協力者は ChatGPT の利用に関するアンケートに回答した．このアンケートでは，ChatGPT の利用頻度，各カテゴリのタスクにおいて自身と ChatGPT のどちらが優れていると思うか，各設問の解答に要した時間，各設問の解答に対する自信，各設問に利用した ChatGPT との対話ログを収集する．

4.2　相 関 分 析

ENDCOREs における 6 つの対人スキルと各タスクにおける対 AI スキルとの相関係数をスピアマンの順位相関係数によって求めた．対人スキルの自己評定の結果を表 3 に，各タスクの結果を表 4 に，相関分析の結果を表 5 に示す．タスクの結果は，

3：https://github.com/openai/openai-cookbook/（2023/12/20 参照）

表 2　タスクの意図と設問例

カテゴリ	意図	設問例	評価方法
知識タスク	一般的な知識を問う	哺乳類の動物をできるだけ多く挙げてください。	解答数
言語タスク	単語の意味や概念の理解，言語による表現力を問う	例 1.「机」と「椅子」はどのような点で似ていますか。 例 2.「従う」とはどういう意味ですか。	解答数 広辞苑に記載されている意味との類似度
計算タスク	論理的思考力，計算能力を問う	A は部屋の掃除に 3 時間かかり、B は同じ部屋を掃除するのに 2 時間かかる。A と B が一緒に掃除すると、どれだけの時間で終わらせることができますか。	1 問 1 点
問題解決タスク	思考力，発想力を問う	10 個のミカンを 3 人に平等に分配する方法をできるだけ多く挙げてください。	解答数

表 3　ENDCOREs による対人スキルの自己評定の結果．値はすべて小数第 3 位で四捨五入している．対人スキルは 1 点から 7 点で評価され，4 点が「普通」であることを示すことから，4 点以下の値を “Low”，4 点より大きい値を “High” とした．“High” の値を太字で示す．

実験協力者	自己統制スキル	表現スキル	解読スキル	自己主張スキル	他者受容スキル	関係調整スキル
H1	**4.50**	**5.50**	**5.50**	**5.00**	**5.75**	**6.00**
H2	**5.00**	3.00	**4.75**	2.00	**4.75**	**4.50**
H3	**5.25**	2.25	**4.25**	3.75	**6.00**	**4.75**
H4	**4.50**	3.75	3.75	3.75	**4.50**	4.00
H5	4.00	**4.75**	3.25	**4.25**	2.00	3.50
H6	2.50	2.50	2.25	3.25	3.50	3.25
H7	4.00	4.00	4.00	4.00	4.00	4.00
H8	4.00	3.00	4.00	3.75	**4.25**	4.00
H9	3.50	1.75	1.25	3.50	4.00	2.50
H10	**5.00**	3.50	3.50	2.00	**6.00**	4.00
H11	**5.25**	4.00	**4.25**	**4.25**	**5.25**	**4.75**
H12	3.75	3.50	**5.50**	4.00	4.00	4.00
平均	4.27	3.46	3.85	3.63	4.50	4.10

解答を表 2 に従って評価した後，各設問のスコアを Min-Max Normalization によって [0, 1] の範囲にスケーリングしている．なお，表 4 の 2 名の実験協力者（H4 と H6）は，計算タスクにおいて ChatGPT を利用しなかった．そのため，計算タスクにおける対 AI スキルと対人スキルとの相関係数を算出する際は除外した．表 5 より，言語タスクと自己統制スキル (r=0.61, $p < 0.05$)，言語タスクと他者受容スキル (r=0.77, $p < 0.01$)，言語タスクと関係調整スキル (r=0.63, $p < 0.05$) の間に有意な正の相関が見られた．

4.3　対人スキルと対 AI スキルの関係

対人スキルと各タスクの結果の関係を図 4 に，ChatGPT を利用した際の対話のステップ数と各ステップにおけるプロンプトの平均の長さを図 5 に示す．結果として，タスクによって対人スキルとの関係に違いが見られた．言語タスクは，すべてのメインスキルにおいて，対人スキルの高い人の方が大きく向上している．ChatGPT との対話の特徴として，設問をそのまま投げかける場合が多くみられたが，知識タスクのように追加の出力を促す場合や，例 2 のような単語の意味を解答する設問では「簡単に説明して」や「辞書のように説明して」といったような指示を追加する場合も見られた．しかしながら，対人スキルによる大きな差は見られなかった．他者受容スキルと強い正の相関が見られた理由として，他者受容スキルの高い人の方が ChatGPT の生成した解答を受け入れやすい傾向にある可能性が考えられるが，今回の結果においては確かでないため，さらなる調査が必要である．一方，知識タスクは，すべてのメインスキルにおいて，対人スキルの低い人の方が大きく向上している．図 5 より，対話の特徴としてはじめに設問をそのまま投げかけたのち「他には？」といった追加の出力を促すことや，「重複のないように」といった指示を追加する傾向にあった．関係調整スキルなどの対人スキルが低い人の方が対話のステップ数が大きいことから，対人スキルの低い人の方が遠慮なく指示を出せたため，より多くの解答を得られたのではないかと考えられる．問題解決タスクは，表現スキルなどの対人スキルの低い方がより大きく向上する傾向にあるが，対人スキルの高い人でも他のタスクと比較して大きく向上している．対話の特徴とし

表 4　タスクごとの対 AI スキルの結果．値はすべて小数第 3 位で四捨五入している．“w AI” は ChatGPT を利用した場合，“w/o AI” は ChatGPT を利用せず独力で解答した場合，“対 AI スキル” は “w AI” と “w/o AI” の差を示す．H4 および H6 の実験協力者は計算タスクにおいて ChatGPT を利用しなかったため，“w AI” および “対 AI スキル” の値は空欄となっている．分析対象である対 AI スキルを太字で示す．

	知識タスク		言語タスク		計算タスク		問題解決タスク	
実験協力者	対 AI スキル	（w AI - w/o AI）	対 AI スキル	（w AI - w/o AI）	対 AI スキル	（w AI - w/o AI）	対 AI スキル	（w AI - w/o AI）
H1	**0.15**	(0.15 - 0.00)	**0.32**	(0.77 - 0.44)	**0.75**	(1.00 - 0.25)	**0.19**	(0.25 - 0.06)
H2	**0.51**	(0.68 - 0.18)	**0.49**	(0.54 - 0.06)	**-0.75**	(0.25 - 1.00)	**0.68**	(0.78 - 0.10)
H3	**0.08**	(0.15 - 0.07)	**0.43**	(0.88 - 0.44)	**0.83**	(1.00 - 0.17)	**0.40**	(0.55 - 0.15)
H4	**0.87**	(0.97 - 0.10)	**0.10**	(0.63 - 0.53)	-	(- - 0.50)	**0.41**	(0.71 - 0.30)
H5	**0.74**	(0.91 - 0.16)	**0.04**	(0.75 - 0.71)	**-0.50**	(0.50 - 1.00)	**0.31**	(0.50 - 0.19)
H6	**0.71**	(0.94 - 0.23)	**-0.18**	(0.40 - 0.58)	-	(- - 1.00)	**0.59**	(0.79 - 0.20)
H7	**0.85**	(0.91 - 0.06)	**-0.03**	(0.50 - 0.53)	**0.00**	(0.00 - 0.00)	**1.00**	(1.00 - 0.00)
H8	**0.62**	(0.62 - 0.00)	**0.13**	(0.44 - 0.32)	**-0.50**	(0.00 - 0.50)	**0.69**	(0.79 - 0.10)
H9	**0.51**	(0.58 - 0.07)	**0.09**	(0.60 - 0.51)	**0.00**	(0.00 - 0.00)	**0.30**	(0.65 - 0.35)
H10	**0.02**	(0.11 - 0.10)	**0.23**	(0.73 - 0.50)	**0.50**	(0.50 - 0.00)	**0.39**	(0.50 - 0.11)
H11	**-0.01**	(0.06 - 0.07)	**0.05**	(0.45 - 0.40)	**-0.75**	(0.00 - 0.75)	**0.55**	(0.80 - 0.25)
H12	**0.01**	(0.05 - 0.03)	**0.10**	(0.37 - 0.27)	**0.00**	(0.00 - 0.00)	**0.24**	(0.24 - 0.00)
平均	**0.42**	(0.51 - 0.09)	**0.15**	(0.59 - 0.44)	**-0.04**	(0.33 - 0.43)	**0.48**	(0.63 - 0.15)

表 5　ENDCOREs の各メインスキルと成果の向上度合いの相関．*は $p < 0.05$，**は $p < 0.01$ となった数値を表す．

	ENDCOREs におけるメインスキル					
タスクカテゴリ	自己統制スキル	表現スキル	解読スキル	自己主張スキル	他者受容スキル	関係調整スキル
知識タスク	-0.44	0.07	-0.52	-0.06	-0.56	-0.50
言語タスク	**0.61***	-0.17	0.54	-0.25	**0.77****	**0.63***
計算タスク	0.03	-0.05	0.08	0.01	0.44	0.19
問題解決タスク	0.09	-0.16	-0.11	-0.32	-0.10	-0.02

て，はじめに設問をそのまま投げかけたのち，必要に応じて指示を追加する場合が見られたが，対人スキルによる差はあまり見られなかった．計算タスクにおいては対人スキルの高い人の方が向上する場合もあるが，より大きく低下する場合もある．このことから，対人スキルに関係なく ChatGPT の利用が適していないと考えられる．

解答時間については，全タスクにおいて，対人スキルの低い人の方が ChatGPT の利用によって短縮する傾向がみられた．対人スキルと各タスクの平均解答時間の関係は付録にて示す．

全体として，ChatGPT を利用した場合の方が成果は向上したが，計算タスクにおいては独力で解答した場合の方が評価が高くなる場合があった．計算タスクにおいて，自身の方が ChatGPT よりも優れていると認識しているにも関わらず，ChatGPT の誤った解答を信用する場合があった．例えば，「5m の木を 1m ずつに切り分けたい。1 回切るのに 5 分かかり、1 回切るごとに 1 分休憩すると、何分で切れますか。」という設問の正しい解は「23 分」であるが，ChatGPT の出力した「26 分」という解を解答している場合があった．また，問題解決課題においては，表 2 の「10 個のミカンを 3 人に平等に分配する方法をできるだけ多く挙げてください。」という設問に対し，ChatGPT が「まず、各人に 2 個ずつ配り、残り 2 個を 3 人で交互に 1 個ずつ取り合う。」という解答を生成した例が見られた．この場合，合計 8 個のミカンにのみ言及されており，残り 2 個に関しては考慮されていない．明確な正答もないため誤りであるという判定も困難であるが，評価の際に注意が必要である．

5　おわりに

本研究では，LLM を利用する人の対人スキルに着目し，人と ChatGPT の協働によるパフォーマンスについて調査を行った．具体的には，ChatGPT を利用して行うタスクとして 4 つのカテゴリのタスクを設計し，実験を実施した．その結果，次の事実が明らかになった．

- 対人スキルの低い人は，知識タスクにおいて ChatGPT の恩恵を受けやすい可能性がある．
- 自己統制スキル，他者受容スキル，関係調整スキルの高い人は言語タスクにおいて ChatGPT の恩恵を受けやすい傾向にある．
- 対人スキルの高さに関わらず，計算タスクに ChatGPT を用いることは有効でない．
- 対人スキルの高さに関わらず，問題解決タスクに ChatGPT を用いることは有効である．

以上のように，一部のタスクと対人スキルには関係があり，人と協働する LLM の開発を行う際は人の能力を考慮した評価

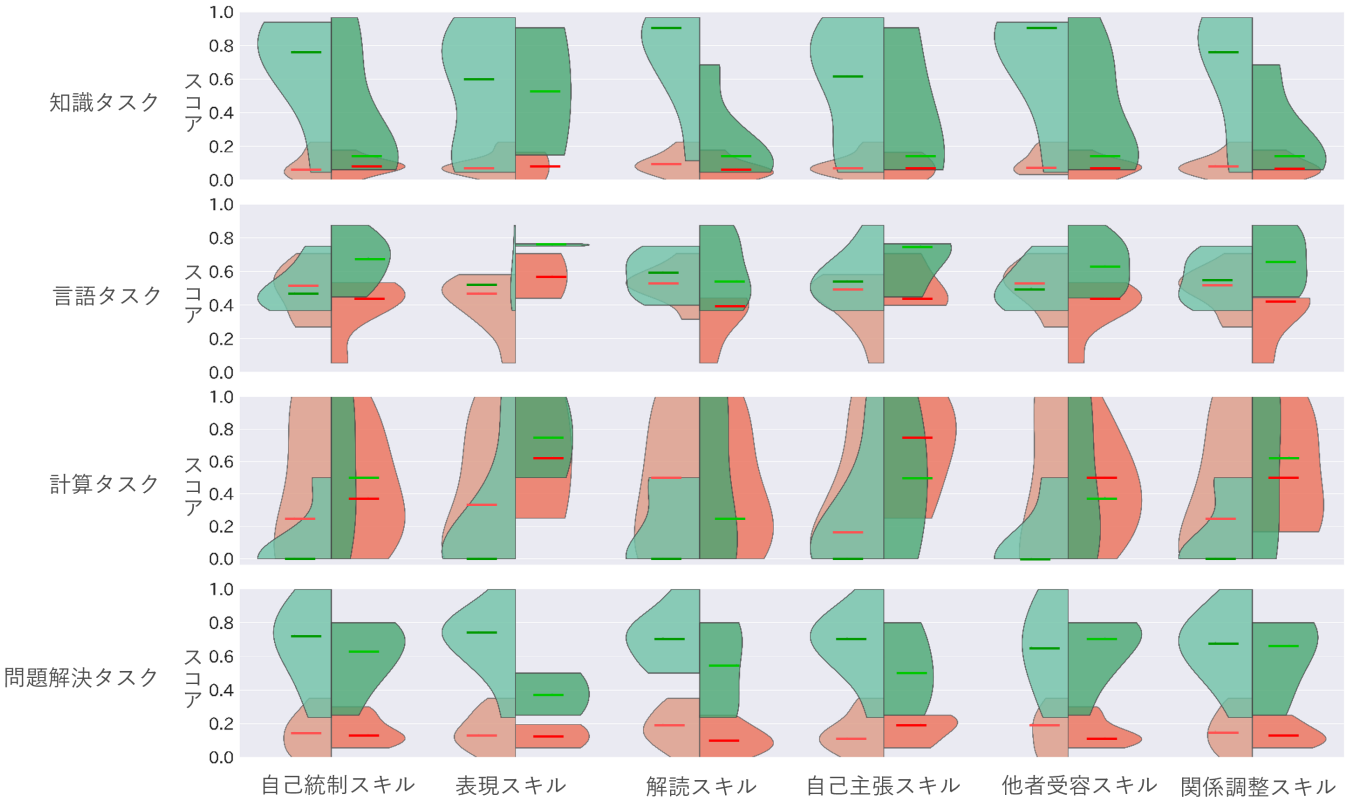


図 4　対人スキルとタスクの成果の関係．横軸は ENDCOREs のメインスキル，縦軸は各タスクのスコアを示す．バイオリンプロット左側に対人スキルの低い場合，右側に対人スキルの高い場合を示す．それぞれ赤色のグラフは ChatGPT を利用せず独力で取り組んだ場合，緑色のグラフは ChatGPT を利用して取り組んだ場合を示す．図中の線は中央値を示す．

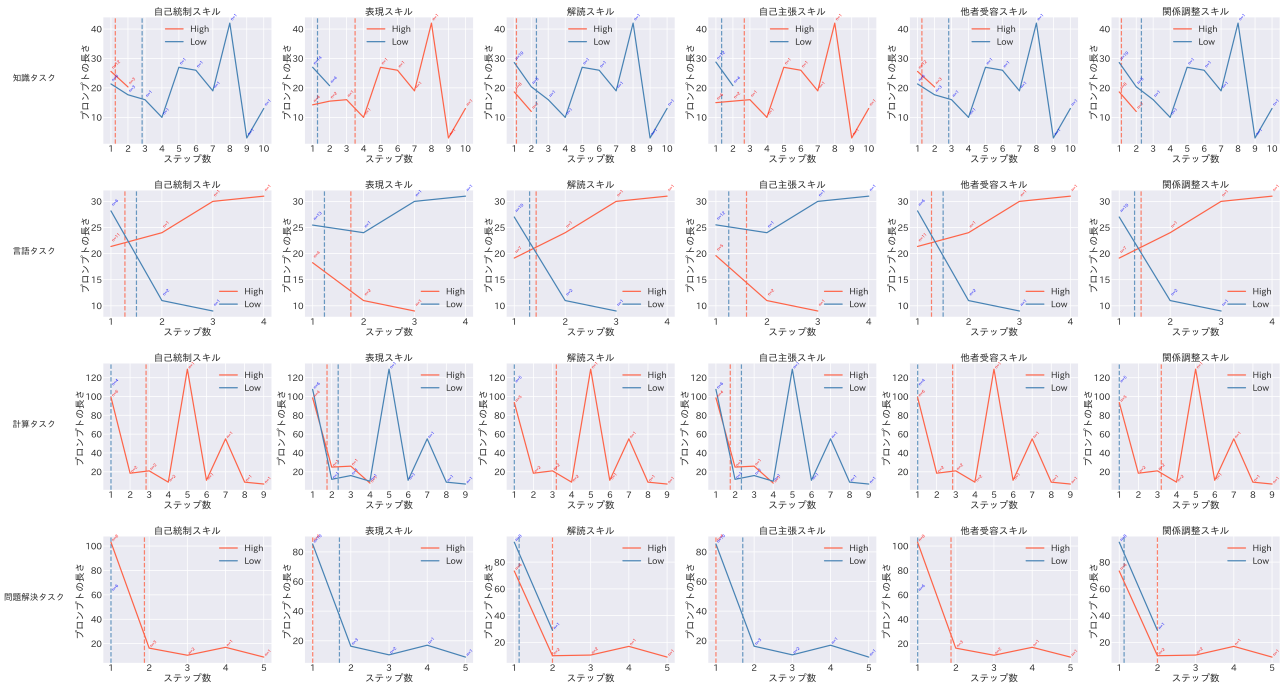


図 5　ChatGPT を利用した際の対話のステップ数と各ステップにおけるプロンプトの平均の長さ．各列に ENDCOREs のメインスキル毎，各行にタスクカテゴリ毎の結果を示す．破線は平均ステップ数を示す．各点における数字は該当するデータ数を示す．

が有用である可能性が示された．しかしながら，タスクに対する評価は十分でなく，タスクの設計や評価方法にさらなる検討が必要である．例えば，問題解決タスクにおいては明確な答えが存在しないため，すべての解答を認めているが，中には類似した解答が複数存在する場合や，問題の解決に繋がるか不明な解答も存在している．そのため，単に解答数のみで評価するのではなく，解答の質を考慮する必要がある．また，本研究ではコミュニケーションスキルを対人スキルとして定義したが，設計したタスクでは ChatGPT とのコミュニケーションが見られない場合があった．そのため，ディベートや意思決定といった，よりコミュニケーションの必要性が高いと考えられるタスクの設計が必要である．さらに，実験データも小規模であるため，追加で実験を行う必要がある．

今後の展望として，クラウドソーシングを用いて大規模な実験を実施する．これにより，偏差値等を用いて 6 つの対人スキル，4 つのタスクにおける対 AI スキルをそれぞれ 1 つの対人スキル，対 AI スキルとして総合して評価する．加えて，アンケート結果を考慮した分析を行う．

謝　　辞

本研究は，LINE ヤフー株式会社共同研究費の支援を受けたものである．

文　　献

[1] Sébastien Bubeck, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

[2] John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, Vol. 41, No. 3, pp. 417—440, 1990.

[3] Isabel Briggs Myers. *The Myers-Briggs Type Indicator: Manual*. Consulting Psychologists Press, 1962.

[4] Isabel Briggs Myers and Mary H. McCaulley. *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, 1985.

[5] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. MPI: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*, 2022.

[6] Haocong Rao, Cyril Leung, and Chunyan Miao. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1184–1194, 2023.

[7] Santi Caballé and Jordi Conesa. Conversational Agents in Support for Collaborative Learning in MOOCs: An Analytical Review. In *In Proceedings of the 10th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2018)*, pp. 384–394. Springer International Publishing, 2019.

[8] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3321–3339, 2023.

[9] Zoie Zhao, et al. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition (C&C '23)*., pp. 368–370. Association for Computing Machinery, 2023.

[10] Yuan Li, Yixuan Zhang, and Lichao Sun. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *arXiv preprint arXiv:2310.06500*, 2023.

[11] 藤本学, 大坊郁夫. コミュニケーション・スキルに関する諸因子の階層構造への統合の試み. パーソナリティ研究, Vol. 15, No. 3, pp. 347–361, 2007.

[12] 木村尚義. ずるい考え方 ゼロから始めるラテラルシンキング入門. あさ出版, 2011.

付　　録

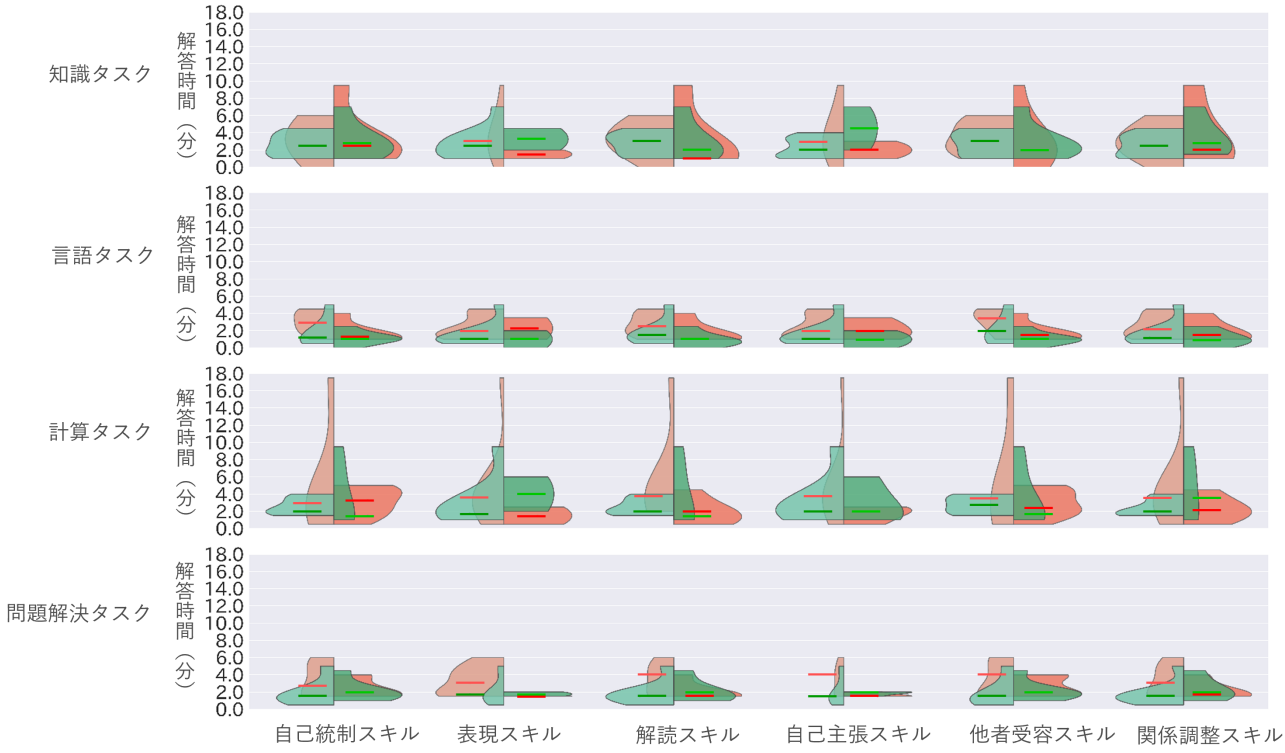


図 6　対人スキルとタスクの平均解答時間（分）の関係．横軸は ENDCOREs のメインスキルを示し，縦軸は各タスクのスコアを示す．バイオリンプロット左側に対人スキルの低い場合，右側に対人スキルの高い場合を示す．それぞれ赤色のグラフは ChatGPT を利用せず独力で取り組んだ場合，緑色のグラフは ChatGPT を利用して取り組んだ場合を示す．図中の線は中央値を示す．

一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2024年2月29日(木) 16:00 ~ 18:10 | T4-B オンライン（Zoom Events)

クラウドソーシングと人間・社会理解 (2)

座長:吉川 正俊(京都大学)
コメンテータ:シリアーラヤ パノット(京都工芸繊維大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T4-B-6-01] 単語埋め込みのジェンダーバイアスとデバイアスの影響の可視化

*杉野 有咲[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:20 ~ 16:45
[T4-B-6-02] TVMLを用いた講師映像が受講生に対して与える影響の分析

*廣岡 采華[1]、熊本 忠彦[1] (1. 千葉工業大学)

16:45 ~ 17:10
[T4-B-6-03] プロジェクト型クラウドソーシングにおけるワーカのスキル推薦

*大石 啓登[1]、伊藤 寛祥[2]、豊田 正史[3]、森嶋 厚行[2] (1. 筑波大学 情報学群、2. 筑波大学 図書館情報メディア系、3. 東京大学 生産技術研究所)

17:10 ~ 17:35
[T4-B-6-04] 映像集中を目的とした第二言語字幕における音声変化を考慮した字幕提示手法

*西 優己[1]、中村 優吾[1]、福嶋 政期[1]、荒川 豊[1] (1. 九州大学)

単語埋め込みのジェンダーバイアスとデバイアスの影響の可視化

杉野　有咲†　　伊藤　貴之†

†お茶の水女子大学大学院人間文化創成科学研究科　112-8610 東京都文京区大塚 2-1-1

E-mail:　†{g1920521, itot}@is.ocha.ac.jp

あらまし　本稿では，日本語の Wikipedia から学習した単語埋め込みを用いて，ジェンダーバイアスの分布を可視化し，可視化結果からバイアス緩和を支援する手法を提案する．本研究では「辞書では性別を意味に含んでいないが，単語埋め込みが「男」または「女」の一方に偏っている単語」を「ジェンダーバイアスが生じている」と定義する．バイアス緩和のためのデバイアス処理がもたらすモデル全体への影響を目で確認しながらデバイアス処理のパラメータを操作すべきという観点から，本研究では可視化を導入する．我々は学習済み Word2vec のデータセットからジェンダーバイアスを検出し，バイアスの分類に取り組んだ．本稿はデバイアス前後のモデルの性能の差を可視化した点を新たに報告する．

キーワード　情報可視化，データ分類，ジェンダーバイアス

1. はじめに

近年の AI 技術の進化に伴い，OpenAI 社が提供する ChatGPT のような AI チャットボットや，画像解析技術を用いた Google 翻訳など，AI を活用したサービスが日々増加し，AI は我々の生活のあらゆる場面で活躍している．中でも，自然言語処理は機械翻訳や AI アシスタントなどに活用される重要な役割を果たしている．

単語埋め込みとは，自然言語処理において，単語を数値に変換してコンピュータ上で扱えるように技法である．しかし，単語埋め込みには，言語モデルを作成する際の事前学習に用いたデータセットに内包されたバイアスが，そのモデルに影響を与える欠点がある．Farkas ら[1]は，性別に依存しない代名詞を持つハンガリー語と英語の職業名を，Google 翻訳を用いて比較し，Google 翻訳にジェンダーバイアスが生じていることを示した．Motoki ら[2]は，ChatGPT の政治的バイアスを測定した．単語埋め込みのバイアス及びデバイアスは研究途上分野である．特に日本語の単語埋め込みは先行研究が少なく[3]，そのほとんどの研究対象が英語の単語埋め込みである．

また「ジェンダー平等の実現」が SDGs の目標に含まれているように，近年世界的にジェンダー問題への意識が高まっている．世界経済フォーラム(WEF)が発表した 2023 年の日本のジェンダーギャップ指数(GGI)[1]は 0.647，順位は 146 カ国中 125 位であり，この順位は 2006 年の GGI 公表開始以来最低である．これは先進国の中で最低水準かつ，2006 年からスコアはほぼ横ばいであり，日本のジェンダーギャップ解消は足踏みが続いていることがわかる．日本では長年にわたり STEM 分野全体に男女格差が存在する．特に工学分野において，ジェンダーギャップの解消への研究が遅れている[4][5].

そこで我々は，日本語版 Wikipedia の本文全文を元に学習した，学習済み Word2vec のデータセットから，ジェンダーバイアスを検出し，それらをデバイアスした結果を可視化し，バイアスの分類に取り組んだ[6]．分類に着手した理由は，モデルの性能劣化を抑えたバイアス緩和を実現するためである．本報告はその研究の続報として新たに，デバイアス前後のモデルの性能差を可視化した点を報告する．

2. 関連研究

3 種類に分けて関連研究を紹介する．1 つ目は単語埋め込みのバイアスの評価方法に関する研究である．Bolukbasi ら[7]は，「女王=王様-男+女」という事例をもって，単語埋め込みはベクトルの加算と減算で単語の類推を表現することが可能であり，単語埋め込みはジェンダーバイアスを内包していることを示した．そして，このようなジェンダーバイアスの内包により例えば，「Computer Programmer」が男性，「Homemaker」が女性の名詞であると類推してしまう危険性があることを指摘した．本研究では，Bolukbasi ら[7]の加算・減算による単語埋め込みの類推方法を参考にして，本

[1]https://www.weforum.org/publications/global-gender-gap-report-2023/

来「男」または「女」から中立であるべき単語が,「王様」と「女王」のように共起している単語をジェンダーバイアスが生じていると定義する．Hutchison ら[8]は,精神･身体的疾患に関するバイアスに着目し,BERT による空間補充分析を行うことでバイアスを評価した.Dixon ら[9]は，性的嗜好・人種・宗教に関するバイアスを対象として,「バイアスを持つ有害な文章」と「バイアスを持たない無害な文章」を事前学習してバイアスを評価した．これらの研究は，特定の単語のバイアスしか考慮されていない.

2 つ目はデバイアスに関する研究である．デバイアス手法は多数存在している．Bolukbasi ら[7]は部分空間の射影を用いた Hard Debias を示した．Hard Debias は，竹下ら[10]や Wang ら[11]や Liang ら[12]など，広く応用されているデバイアス方法である．Hard Debias は主に英語の単語埋め込みに対して使用されているが,竹下ら[11]によって日本語の単語埋め込みにも応用可能であると示されている．Elazar ら[13]は，敵対的学習を用いるデバイアス手法を示した．Lu ら[14]は，学習済みデータセットに性別を反転させたデータを追加し，CDA (Counterfactual Data Augmentation)と呼ばれる，文法的な矛盾が生じないジェンダーバイアスの緩和方法を示した．また Sun ら[15]によって，デバイアス手法は，再学習でデバイアスする方法と，既存のモデルにデバイアスプログラムを適用して出力を調整しデバイアスする推論法の 2 種類に分類できることが示された.現段階では,本研究は推論法で議論を進める.これらの研究では，デバイアスによるモデルへの影響が考慮されていない.

3 つ目はデバイアスによるモデルへの影響を評価する研究である．前述の単語埋め込みのデバイアス方法には,いずれもモデルの性能が劣化する問題点がある.そこで現在も，モデルの性能劣化とデバイアスの関係性は研究対象になっている[16]. 小林ら[16]は，単語埋め込みの「表現の曖昧さの増大」によりモデルの性能が劣化する可能性を示している．具体的には，デバイアス前後の単語埋め込みのL^2-ノルムおよびそれを用いて作成した分類モデルについて，SHAP 値の差を比較した．この研究では，Bolukbasi ら[7]の既存のデバイアス方法を用いているが，デバイアス方法の評価および改善は実施されていない．本研究では，Bolukbasi ら[7]のデバイアス方法のパラメータを操作することで，モデルへの影響が小さくなるようにデバイアスを改善することを目標とする.

以上の研究はいずれも単語埋め込みモデルを一括で扱っており，特定の単語に注目した可視化方法やデバイアス方法は見当たらない．また，全て英語の単語埋め込みが対象である.

3. 提案手法

本章では提案手法の前提となるデータについて述べたのち，処理手順を示す．なお，当初の予定では単語埋め込みのモデル全体を k-means 法を用いてクラスタリングし，クラスタごとにデバイアスの度合いを調節することで，モデル全体の性能劣化を抑えようと試みた．しかし k-means 法を適用するにあたりエルボー法を適用したところ，適切なクラスタ数を算出できなかった．本報告では，モデル全体に一律のデバイアスを適用している.

3.1. 使用したデータ

本研究では，デバイアス対象として東北大学が公開している「日本語 Wikipedia エンティティベクトル」[2]を使用した．これは日本語の Wikipedia の本文から学習した Word2vec のデータセットである．Unicode で CJK またはカタカナまたはひらがな以外の名詞を除外した．英単語や記号の除外が目的である.

3.2. バイアス検出

バイアス検出のために，まずデータの単語全てに対して「-男+女」及び「+男-女」を計算し，コサイン類似度上位 10 位をそれぞれ出力する.出力結果を比較して，単語が重複かつコサイン類似度が一定値離れているペアを,バイアスが生じているとみなして抽出する.

例として「家事」という単語で説明する.「家事-男+女」「家事+男-女」の出力結果には「仕事」が共通している.「家事-男+女」と「仕事」のコサイン類似度は 0.5804643,「家事+男-女」と「仕事」のコサイン類似度は 0.7221109 である．この場合「仕事」が「男」寄り,「家事」が「女」寄りにバイアスが生じているので,「仕事」と「家事」を抽出する.また,「兄」や「女優」など性別を意味に含む単語はバイアス検出対象から除外した．除外する単語リストの作成は，Lu ら[14]が公開していた「Gender Pairs」と，Schmdit[17]を参考にした.

3.3. デバイアス

Bolukbasi ら[7]の Hard Debias を応用する.「男」または「女」の一方に偏った単語が,「男」「女」両方から等距離の位置に近づくほど，その単語のジェンダーバイアスが緩和される.

処理手順は以下の通りである.

1. 性別を表す単語群をもとに，主成分分析で性別の基準となる軸を計算する．今回は{男，女，男性，女性，彼，彼女，父，母，息子，娘，少年，少女}を用いて軸を計算した.これらの単語は,Bolukbasi ら[7]の性別を含む単語群を日本語に訳したもので

[2]https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

ある．

2. バイアスを除去する単語から，性別の基準となる軸成分を減算する．
3. 性別を含む単語群を，軸から等距離になるように移動する．

減算する軸成分の割合を調整することで，バイアスの緩和度を操作できる．

3.4. デバイアス前後のモデルの評価

2 つの方法でデバイアス前後のモデルを評価する．

1 つ目はデバイアス前後の単語ベクトルの大きさの差で評価する．デバイアス後の単語ベクトルの大きさが小さい場合，デバイアスによって情報を損失したとみなすことができる．本研究では，デバイアス前後の単語同士の関係の変化の確認を目的として，Tensor Board を用いてモデル全体を可視化する．

2 つ目は，デバイアス前後の単語のカテゴリを比較して評価する．まずバイアスを含む単語全体に対して，science · sports · politics · business · entertainment の 5 種類のカテゴリに分類する．次にデバイアス済みのデータに対して，同様のカテゴリ分類を実施し，デバイアス前後のカテゴリの分布の差を比較する．

4. 実行結果· 考察

4.1. デバイアス前後の L2 ノルムの比較

図 1 はデバイアス前後の単語ベクトルの L2 ノルムの差(デバイアス前-デバイアス後)のヒストグラムである．横軸が L2 ノルムの差であり，ほとんど正の値であることから，デバイアスにより単語ベクトルが全体的に縮小していることが分かる．

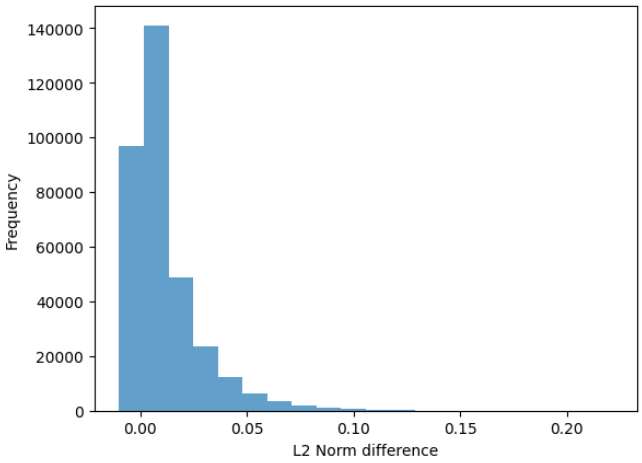


図 1:デバイアス前後の L2 ノルムの差．

4.2. デバイアス前後のカテゴリ分類の比較

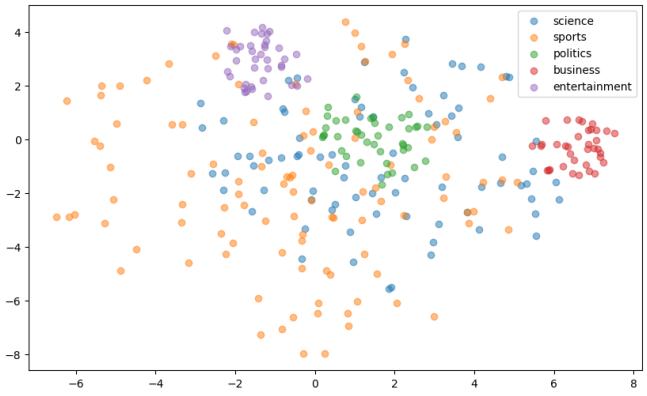


図 2:デバイアス前．

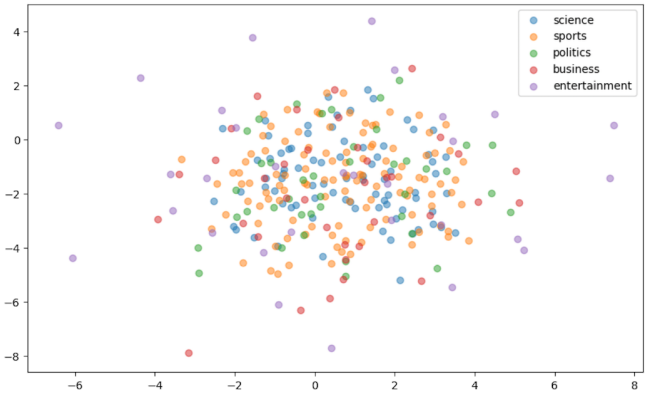


図 3:デバイアス後．

図 2 および図 3 は，ジェンダーバイアスが強い上位 500 語の散布図である．図 3 のカテゴリには，デバイアス前のカテゴリを適用している．これらの散布図は主成分分析で次元削減した結果を示している．デバイアス後は，デバイアス前よりベクトル空間が小さくなっていることがわかる．デバイアス前は確認できたカテゴリのクラスタが，デバイアス後に，デバイアスにより単語の意味が変化していることがわかる．

図 4 はカテゴリ分類の結果である．undebiased(青色)はデバイアス前の単語のカテゴリ分布，debiased(橙色)はデバイアス後の単語のカテゴリ分布，difference(赤色)はデバイアス前のカテゴリ分布と比較して，デバイアス後の単語のうち，異なるカテゴリに分類されてしまった単語である．全てのカテゴリに分類ミスが生じており，デバイアスによりモデルの性能が低下していることがわかる．カテゴリに着目すると，entertainment カテゴリの単語が最もデバイアスの影響を受けていることが確認できる．

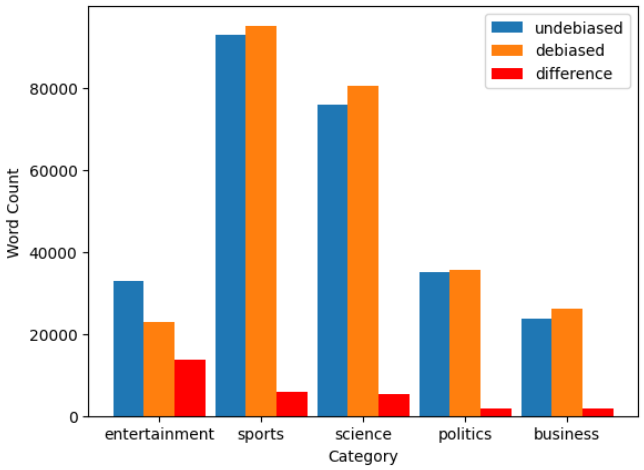


図 4:カテゴリ分類の結果.

5. まとめと今後の課題

本稿では，デバイアスによって単語埋め込みの性能が低下することを，可視化によって示した．デバイアスによって単語埋め込みのモデル全体が縮小し，カテゴリごとにデバイアスの影響度が異なることが確認できた．

現在，各カテゴリのデバイアスの度合いを調整可能な形で，カテゴリ分類の結果を可視化するシステムを実装している．この実装により，モデルの性能劣化を抑制したバイアス緩和の実現を目標とする．

今後の予定として，他のタスクによるモデルの性能を比較することで，複数のタスクを考慮して最適なデバイアス度を検討できる可視化システムの実装があげられる．最終的には，AI 対話システムのようなアプリケーションのジェンダーバイアスを緩和する助力となるシステム開発を目標とする．

参 考 文 献

[1] Anna Farkas, Renata Nemeth, “How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points”, Social Sciences & Humanities Open 5, 2022.

[2] Fabio Motoki， Valdemar Pinho Neto， Victor Rodrigues，”More human than human: measuring ChatGPT political bias”，Public Choice2023，2023.

[3] Shintaro Sakai， Yasuhiro Suzuki，”Evaluation of Gender Bias in Japanese Word Embedding Model”，The 37th Annual Conference of the Japanese Society for Artificial Intelligence，2023.

[4] 渡邊智子，北川尚美，田中真美，“日本の工学系女性研究者の現状-八大学工学系連合会に着目して-”，公益社団法人日本工学教育協会 2022 年度工学教育研究講演会，pp.116-117，2022.

[5] Yuko Ikkatai， Azusa Minamizaki， Kei Kano, Atsushi Inoue， Euan McKay, Hiromi M. Yokoyama,” Gender-biased public perception of STEM fields, focusing on the influence of egalitarian attitudes toward gender roles”， Journal of Science Communication ,2020.

[6] 杉野有咲，伊藤貴之，”単語埋め込みのジェンダーバイアスの可視化”，DEIM Forum 2023，2023.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4356‐4364, 2016.

[8] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton , Kellie Webster , Yu Zhong , Stephen Denuyl, ”Social Biases in NLP Models as Barriers for Persons with Disabilities”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，pp.5491-5501，2020.

[9] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, Lucy Vasserman, ” Measuring and Mitigating Unintended Bias in Text Classification” , Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society，pp.67-73，2018.

[10] 竹下昌志，ジェプカ・ラファウ，荒木健治，“日本語の単語埋め込みにおける文字種による性別バイアスの相違の分析”，電気・情報関係学会北海道支部連合大会，pp. 129-130，2020.

[11] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez1 Caiming Xiong, “Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5443-5453, 2020.

[12] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, Louis-Philippe Morency, “Towards Debiasing Sentence Representations”, The 58th Annual Meeting of the Association for Computational Linguistics, pp. 5502-5515, 2020.

[13] Yanai Elazar, Yoav Goldberg, ”Adversarial Removal of Demographic Attributes from Text Data” , Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.11-21, 2018.

[14] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, Anupam Datta, “Gender Bias in Neural Natural Language Processing”, Logic, Language, and Security, pp. 189-202, 2019.

[15] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, William Yang Wang, “Mitigating Gender Bias in Natural Language Processing”, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1630‐1640, 2019.

[16] 小林一樹，脇田建，“バイアス除去がもたらす NLP モデルの性能劣化”，The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022.

[17] Schmidt Ben， ”Rejecting the gender binary: a vector-space operation”， Ben’s Bookworm Blog， 2015.

TVML を用いた講師映像が受講生に与える影響の分析

廣岡 采華† 熊本 忠彦‡

†, ‡千葉工業大学 情報科学部 〒275-0016 千葉県習志野市津田沼 2-17-1

E-mail: †s2032127GM@s.chibakoudai.jp, ‡kumamoto@net.it-chiba.ac.jp

あらまし 本稿では, フリーの CG アニメーション作成ツールである TVML(TV program Making Language)を用いて作成した講師キャラクタが受講者のオンライン授業(オンデマンド配信)に対する印象や理解度にどのような影響を及ぼすかをアンケート調査に基づいて調べ, 明らかにする. 具体的には, 講師キャラクタとして 2 種類の上半身画像(男体, 女体)と 2 種類の声(男声, 女声)を組み合わせ, 4 種類の授業動画を用意し, インターネット上のアンケート調査において回答者(18～22 歳の学生 743 人)に授業動画(1 動画の再生時間は約 1 分)を視聴してもらった後, 授業の印象や理解度に関して回答してもらった. その結果,「女声・男体」の組合せが授業や講師の印象, 授業の理解度に関して高評価であることが分かった. また, 先行研究(24 歳・女性が講師役を務めた授業動画)と比較してみた結果, 全体的な印象は人間の女性の方が高評価であったが, 授業の理解度は同程度であった.

キーワード オンライン授業, オンデマンド配信, CG キャラクタ, 合成音声, 授業動画, 授業の印象

1. はじめに

近年, 情報通信技術の発達と一般社会への普及により, インターネットを介したオンライン授業が行われるようになった. さらに 2020 年には新型コロナウイルス感染症の蔓延により全大学の約 90%においてオンライン授業が行われるようになった[1]. コロナ禍後となった現在, オンライン授業の割合は減ったが, 今なおオンライン授業に対する需要は高い.

オンライン授業で視聴する授業動画には, 授業コンテンツを説明する音声だけでなく, それを説明する講師映像を含める場合もある. しかしながら, 講師映像を含めてしまうと画面内の貴重なスペースを占有してしまい, 無駄が生じる. にもかかわらず, 講師映像の必要性についてはあまり議論されていない. この問題に対し, 四戸ら[2]は, 同一の授業コンテンツからなる講師映像を含まない授業動画(音声のみ), 講師の上半身画像が配置された授業動画(音声＋静止画), 講師の上半身が映った映像が配置された授業動画(音声＋映像)の 3 種類を用意し, インターネット上でのアンケート調査(回答者は 18～22 歳の学生 622 人)により「音声のみ」に比べ「音声＋静止画」や「音声＋映像」の方が高評価であることを明らかにしている.

一方, 授業動画の作成・管理コストという面から見てみると, 講師映像(音声)の撮影(録音)は容易でないため, 授業動画の新規作成だけでなく, 追加や修正, 削除にも高いコストがかかってしまう. 簡単に作成できる CG キャラクタを講師として用いることができれば, 低コスト化を図ることが可能となる.

そこで本稿では, テキスト形式の台本からテレビ番組風の CG アニメーションを作成することができるフリーツールである TVML(TV program Making Language)[3]を用いて作成した講師キャラクタが受講者のオンライン授業に対する印象や理解度にどのような影響を及ぼすかをアンケート調査に基づいて調べるとともに, 四戸らの分析結果[2]と比較することで, CG キャラクタを導入することの有効性を検証する.

具体的には, 四戸ら[2]が用意した授業コンテンツを用い, 講師の映像と音声を TVML で作成した講師キャラクタに差し替えることで授業動画を用意する. このとき, 2 種類の上半身映像(男体, 女体)と 2 種類の声(男声, 女声)を組み合わせて講師キャラクタを作成し, 全部で 4 種類の授業動画を用意する. その後, 18～22 歳の学生 743 人を対象とするアンケート調査をインターネット上で行い, 授業動画(1 動画あたりの再生時間は約 1 分)を視聴してもらった後, 授業や講師キャラクタの印象と授業に対する理解度を回答してもらう. なお, 授業の印象を評価するための評価項目には「丁寧な/丁寧でない, 熱意のある/熱意のない, 説明がわかりやすい/説明がわかりにくい, 面白い/面白くない, メリハリがある/メリハリがない, 内容を理解できた/内容を理解できなかった」の 6 つを用い, 講師キャラクタの印象を評価するための評価項目には「温かい/冷たい, 外向的な/内向的な, 信頼できる/信頼できない, 真面目な/不真面目な, 知的な/知的でない, 明るい/暗い, 自信のある/自信のない, 好感が持てる/好感が持てない」の 8 つを用いる.

2. 関連研究

上間ら[4]は, 対面授業と映像方式のオンライン授業において教員の表情や声, マスクの有無などが受講生に与える印象を調べ, その結果, 声のトーンを高くし, 表情を作ることが教員の印象にポジティブな影響を与えることを明らかにした. これに対し本研究では, 授業に関する様々な印象と理解度を分析対象としている点が異なっている.

柿澤[5]は, 学生と教師の性別の組合せが同性であるとき, 成績が向上する傾向にあることを明らかにした. そこで本稿でも, 講師キャラクタの性別(外見と声)と回答者の性別との一致・不一致が与える影響についても検証する.

斉藤ら[6]は, プレゼンテーション時に音声合成エンジン

により生成した合成音声を用いた場合と講師自身の声を用いた場合を比較し，プレゼンテーションの内容理解には大きな影響がないことを明らかにした．これに対し本研究では，合成音声と CG キャラクタの相乗効果を検証する点が異なっている．

3. アンケート調査による印象評価の実施

本節では，TVML（TV program Making Language）[3]により作成した授業動画を用いて 18～22 歳の学生 743 人に対してアンケート調査を実施する．

3.1. アンケート調査の準備

まず，アンケート調査に用いる授業動画を作成した．授業コンテンツには，四戸ら[2]が用いた「整数の基数変換」と「小数の基数変換」を採用した（いずれも 10 進数を 2 進数に変換する）．これは文系か理系かに関係なく，ある程度の知識があることが期待され，かつ四戸らの分析結果[2]との比較を行うためである．また，講師映像の作成には TVML を用いた．このとき，CG キャラクタには General_F（女体）と General_M（男体）を，合成音声には AquesTalk2[7]の aq_f3a（女声）と aq_mf1（男声）を採用した．TVML で作成した映像に対し，マイクロソフト社が提供するビデオ作成編集ツール Clipchamp を用いて画面録画と編集を行い，約 1 分程度の講師映像を作成した．この講師映像を授業コンテンツに配置し，マイクロソフト社が提供する PowerPoint のエクスポート機能を用いて mp4 形式の動画ファイルに変換した．なお，講師映像は，四戸ら[2]と同様，「整数の基数変換」では右上に，「小数の基数変換」では右下に配置した．

アンケート調査に用いる授業動画として合成音声（2 種類）と CG キャラクタ（2 種類）と授業コンテンツ（2 種類）を組み合わせ，全部で 8 本の授業動画を用意した．ここで参考のために，「整数の基数変換」と女体の組合せを図 1 に示し，「小数の基数変換」と男体の組合せを図 2 に示す．

また，質問紙には四戸ら[2]が用いたものを再利用した．すなわち，Q1 では整数と小数の基数変換に関する学習経験を回答してもらうために，「10 進数の整数を 2 進数に変換する方法を学んだことがある」，「10 進数の整数を 2 進数に変換する方法を覚えている」，「10 進数の小数を 2 進数に変換する方法を学んだことがある」，「10 進数の小数を 2 進数に変換する方法を覚えている」という 4 つの小問を用意し，それぞれ「はい」（1 点），「多分はい」（2 点），「多分いいえ」（3 点），「いいえ」（4 点），「わからない」（点なし）の中から 1 つを選んでもらうことにした．

次に，Q2 と Q3 では授業の印象を回答してもらうために，「丁寧な/丁寧でない」，「熱意のある/熱意のない」，「説明がわかりやすい/説明がわかりにくい」，「面白い/面白くない」，「メリハリがある/メリハリがない」，「内容を理解できた/内容を理解できなかった」という 6 個の評価項目を用意し，各評価項目の左側の項目を A，右側の項目を B として「A に当てはまる」（1 点），「A にやや当てはまる」（2 点），「どちらとも言えない」（3 点），「B にやや当てはまる」（4 点），「B に当てはまる」（5 点）の 5 段階で評価してもらうことにした．なお，Q2 は「整数の基数変換」に対する印象評価であり，Q3 は「小数の基数変換」に対する印象評価である．

図 1 「整数の基数変換」の授業動画（女体の場合）

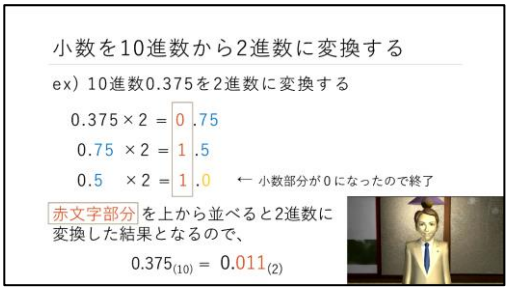


図 2 「小数の基数変換」の授業動画（男体の場合）

次に，Q4 で講師キャラクタの印象を回答してもらうために，「温かい/冷たい」，「外向的な/内向的な」，「信頼できる/信頼できない」，「真面目な/不真面目な」，「知的な/知的でない」，「明るい/暗い」，「自信のある/自信のない」，「好感が持てる/好感が持てない」という 8 個の評価項目を用意し，Q2・Q3 と同様，「A に当てはまる」（1 点）～「B に当てはまる」（5 点）の 5 段階で評価したもらうことにした．

最後に，Q5 で授業の理解度を回答してもらうために，「2 つの授業動画を見て，10 進数を 2 進数に変換できるようになったと思いますか？」という設問を用意し，「整数の基数変換」と「小数の基数変換」のそれぞれについて「そう思う」（1 点），「ややそう思う」（2 点），「どちらとも言えない（動画を見る前から変換できた）」（3 点），「あまりそう思わない」（4 点），「全くそう思わない」（5 点），「わからない」（点なし）の中から 1 つを選択してもらうことにした．

3.2. アンケート調査の実施

アンケート調査はインターネット調査会社[8]を介して実施した．このとき，回答者の条件として職業を「学生」，年齢を「18～22 歳」と指定し，講師キャラクタの外見（男体，女体）と合成音声（男声，女声）の組合せ毎に実施した．結果，計 4 回のアンケート調査を実施し，回答者は全員で 743 人であった．なお，回答者に重複はない．

3.3. データクリーニング

アンケート調査では回答者が不適切な回答を行う場合がある．そこで，以下のような基準でデータクリーニングを行った．すなわち，授業の印象を問うQ2・Q3と講師の印象を問うQ4において，選択した回答に偏りがある場合（具体的には，選択した回答が2種類以下の場合）にその回答者のデータを全て削除した．その結果，204人分の回答者データが削除され，539人分が残った．表1にデータクリーニング後の回答者の性別・年齢構成を示す．

表1 アンケート調査の回答者の性別・年齢構成（データクリーニング後）

（a）女声・女体のとき

	男性	女性	小計
18～19歳	43	42	85
20～22歳	41	14	55
小計	84	56	140

（b）女声・男体のとき

	男性	女性	小計
18～19歳	35	32	67
20～22歳	12	15	27
小計	47	47	94

（c）男声・女体のとき

	男性	女性	小計
18～19歳	34	42	76
20～22歳	41	35	76
小計	75	77	152

（d）男声・男体のとき

	男性	女性	小計
18～19歳	39	39	78
20～22歳	36	39	75
小計	75	78	153

4. 講師映像が受講生に与える影響の分析

4.1. 基本統計量（平均）の分析

データクリーニング後のデータから平均を求めた．結果を表2～表5に示す．また，回答者を男女別に分けた場合の平均を表6～表9に示す．各表において，青色のセルはその値がそれぞれの評価項目（A/B）において最小（最もA寄り）であることを示し，赤色のセルは最大（最もB寄り）であることを示している．

表2の基数変換に関する学習経験を見てみると，平均の低い講師キャラクタと平均の高い講師キャラクタに若干の偏りが見られるが，いずれの評価項目でも差は0.26以下であり，問題にならない範囲と言える．また，「整数の基数変換を学んだことがある」では平均が1.461～1.725であり，「多分はい」（2点）より点数が低いことから，学んだことがある人がそれなりに多いことがわかる．一方，「整数の基数変換の方法を覚えている」では平均が2.164～2.362であり，「多分はい」（2点）と「多分いいえ」（3点）のほぼ中間であることから，丁度良いレベルの問題になっていると言える．一方，「小数の基数変換を学んだことがある」では平均が1.786～1.936であり，整数の基数変換よりは学習経験が少ないことがわかる．「小数の基数変換の方法を覚えている」では平均が2.343～2.500であり，整数の基数変換より難易度が高めになっていることがわかる．

表2 学習経験（平均）

	女声 女体	女声 男体	男声 女体	男声 男体
整数の基数変換を学んだことがある	1.493	1.617	1.461	1.725
整数の基数変換の方法を覚えている	2.164	2.362	2.217	2.222
小数の基数変換を学んだことがある	1.786	1.936	1.849	1.895
小数の基数変換の方法を覚えている	2.343	2.500	2.447	2.431

表3 授業の印象（平均）

（a）「整数の基数変換」の場合

	女声 女体	女声 男体	男声 女体	男声 男体
丁寧さ	2.314	2.138	2.428	2.510
熱意	3.607	3.340	3.638	3.588
分かりやすさ	2.279	2.043	2.461	2.693
面白さ	3.657	3.457	3.553	3.477
メリハリ	3.507	3.074	3.395	3.451
内容理解	2.100	1.862	2.105	2.444

（b）「小数の基数変換」の場合

	女声 女体	女声 男体	男声 女体	男声 男体
丁寧さ	2.414	2.266	2.520	2.706
熱意	3.714	3.383	3.513	3.595
分かりやすさ	2.471	2.404	2.625	2.680
面白さ	3.543	3.500	3.296	3.588
メリハリ	3.571	3.138	3.382	3.386
内容理解	2.286	2.404	2.461	2.542

表3の「整数の基数変換」の授業における印象を見てみると，すべての評価項目で「女声・男体」が最も高評価だったのに対し，6項目のうち3項目（「丁寧さ」「分かりやすさ」「内容理解」）で「男声・男体」が低評価であり，声の違いで高評価と低評価に分かれているのがわかる．一方，「小数の基数変換」の授業における印象では，整数の基数変換

の場合と同様，6 項目のうち 4 項目(「丁寧さ」「熱意」「分かりやすさ」「メリハリ」)で「女声・男体」の評価が最も高評価だったのに対し，6 項目のうち 4 項目(「丁寧さ」「分かりやすさ」「面白さ」「内容理解」)で「男声・男体」が最も低評価であった．また，表 3 によれば，「整数の基数変換」でも「小数の基数変換」でも「丁寧さ」「分かりやすさ」「内容理解」の平均は「どちらとも言えない」(3 点)より低く，高評価であったが，「熱意」「面白さ」「メリハリ」の平均は「どちらとも言えない」(3 点)より高く，低評価であった．CG キャラクタと合成音声からなる講師キャラクタの弱点とも言えるが，キャラクタの作りや合成音声の声質を工夫することで改善する余地もあると考えている．

表 4 講師キャラクタの印象(平均)

	女声 女体	女声 男体	男声 女体	男声 男体
温かさ	3.921	3.798	3.750	3.732
外向的	3.557	3.309	3.401	3.373
信頼	2.821	2.840	3.066	2.961
真面目さ	2.107	2.128	2.250	2.320
知的さ	2.507	2.500	2.513	2.667
明るさ	3.800	3.649	3.836	3.719
自信	3.007	2.851	3.039	3.026
好感	3.507	3.447	3.421	3.497

表 5 授業の理解度(平均)

	女声 女体	女声 男体	男声 女体	男声 男体
整数の基数変換	2.007	1.840	2.112	2.248
小数の基数変換	2.050	2.245	2.250	2.373

表 6 学習経験(男女別平均)

	女声・女体		女声・男体	
	男性	女性	男性	女性
整数の基数変換を学んだことがある	1.464	1.536	1.617	1.617
整数の基数変換の方法を覚えている	2.048	2.339	2.362	2.362
小数の基数変換を学んだことがある	1.762	1.821	1.894	1.979
小数の基数変換の方法を覚えている	2.202	2.554	2.426	2.574
	男声・女体		男声・男体	
	男性	女性	男性	女性
整数の基数変換を学んだことがある	1.427	1.494	1.947	1.897
整数の基数変換の方法を覚えている	2.160	2.273	2.173	2.846
小数の基数変換を学んだことがある	1.827	1.870	2.133	2.372
小数の基数変換の方法を覚えている	2.427	2.468	2.360	3.141

表 4 の講師キャラクタの印象を見てみると， 8 項目のうち 4 項目(「外向的」「知的さ」「明るさ」「自信」)で「女声・男体」が最も高評価であり，他の 4 項目も割と高い評価であった．しかしながら，平均が「どちらとも言えない」(3 点)より低かったのは，「信頼」「真面目さ」「知的さ」「自信」の 4 項目のみであり，全体として評価が良かったとは言えない．これも講師キャラクタならではの特徴と言える．

表 7 授業の印象(男女別平均)

(a) 「整数の基数変換」の場合

	女声・女体		女声・男体	
	男性	女性	男性	女性
丁寧さ	2.262	2.393	2.319	1.957
熱意	3.571	3.661	3.532	3.149
分かりやすさ	2.167	2.446	2.000	2.085
面白さ	3.583	3.768	3.660	3.255
メリハリ	3.250	3.893	3.319	2.830
内容理解	2.060	2.161	1.830	1.894
	男声・女体		男声・男体	
	男性	女性	男性	女性
丁寧さ	2.400	2.455	2.613	2.410
熱意	3.480	3.792	3.667	3.513
分かりやすさ	2.413	2.506	2.827	2.564
面白さ	3.533	3.571	3.360	3.590
メリハリ	3.267	3.519	3.453	3.449
内容理解	2.200	2.013	2.560	2.333

(b) 「小数の基数変換」の場合

	女声・女体		女声・男体	
	男性	女性	男性	女性
丁寧さ	2.333	2.536	2.298	2.234
熱意	3.583	3.911	3.617	3.149
分かりやすさ	2.298	2.732	2.447	2.362
面白さ	3.440	3.696	3.681	3.319
メリハリ	3.452	3.750	3.319	2.957
内容理解	2.143	2.500	2.383	2.426
	男声・女体		男声・男体	
	男性	女性	男性	女性
丁寧さ	2.493	2.545	2.773	2.641
熱意	3.373	3.649	3.680	3.513
分かりやすさ	2.573	2.675	2.680	2.679
面白さ	3.227	3.364	3.467	3.705
メリハリ	3.253	3.506	3.400	3.372

内容理解	2.293	2.623	2.640	2.449

表 5 の授業の理解度を見てみると,「整数の基数変換」で最も高評価だったのは「女声・男体」であり,最も低評価だったのは「男声・男体」であった.一方,「小数の基数変換」で最も高評価だったのは「女声・女体」であり,最も低評価だったのは「男声・男体」であった.ただし,「女声・男体」も「女声・女体」とほぼ同じくらい高評価であり,どちらの授業でも「女声」の方が「男声」より高評価であったと言える.

表 6 の学習経験における男女別の平均を見てみると,「男声・男体」の「整数の基数変換を学んだことがある」を除いて男性の方が女性より学習経験が多いことがわかる.男女間で最も差があったのは,「男声・男体」の「小数の基数変換を覚えている」の 0.781 であった.ただし,男性内での傾向と女性内での傾向は似ており,①「整数の基数変換を学んだことがある」の平均はいずれも「多分はい」(2 点)より低く,学んだことがある人がそれなりに多いこと,②「整数の基数変換の方法を覚えている」の平均はいずれも「多分はい」(2 点)と「多分いいえ」(3 点)の間にあり,丁度良いレベルの問題になっていること,③「小数の基数変換を学んだことがある」と「小数の基数変換の方法を覚えている」の平均はいずれも「整数の基数変換」の場合より高く,難易度が高めになっていることがわかる.

表 8 講師キャラクタの印象(男女別平均)

	女声・女体		女声・男体	
	男性	女性	男性	女性
温かさ	3.726	4.214	3.979	3.617
外向的	3.548	3.571	3.362	3.255
信頼	2.690	3.018	2.830	2.851
真面目さ	2.155	2.036	2.255	2.000
知的さ	2.583	2.393	2.511	2.489
明るさ	3.714	3.929	3.660	3.638
自信	2.988	3.036	2.787	2.915
好感	3.369	3.714	3.553	3.340
	男声・女体		男声・男体	
	男性	女性	男性	女性
温かさ	3.667	3.831	3.653	3.808
外向的	3.267	3.532	3.253	3.487
信頼	2.987	3.143	2.933	2.987
真面目さ	2.307	2.195	2.520	2.128
知的さ	2.467	2.558	2.787	2.551
明るさ	3.707	3.961	3.760	3.679
自信	2.933	3.143	3.227	2.833
好感	3.293	3.545	3.560	3.436

表 7 の授業の印象における男女別の平均を見てみると,「整数の基数変換」でも「小数の基数変換」でも,女性に対しては「女声・男体」が高評価だったが,男性に対しては「女声・男体」「女声・女体」「男声・女体」がほぼ同じくらい高評価だった.なお,男女関係なく高評価だった「女声・男体」を見てみると,男女間で評価が分かれたのは「メリハリ」であり,その差は 0.49 と最も大きかった.このとき,男性が「メリハリがない」寄りの評価だったのに対し,女性は「メリハリがある」寄りの評価となっていた.

表 8 の講師キャラクタの印象における男女別の平均を見てみると,「女声・男体」に対する女性の評価が他の講師キャラクタに対する評価に比べ比較的良いことがわかる.また,女声/男声と男体の組合せでは女性の方が高評価であり,女声/男声と女体との組合せでは男性の方が高評価と言える.講師キャラクタの外見に関しては受講者と性別が異なる方が高評価になると考えられる.

表 9 授業の理解度(男女別平均)

	女声・女体		女声・男体	
	男性	女性	男性	女性
整数の基数変換	1.893	2.179	1.851	1.830
小数の基数変換	1.845	2.357	2.319	2.170
	男声・女体		男声・男体	
	男性	女性	男性	女性
整数の基数変換	2.133	2.091	2.333	2.397
小数の基数変換	2.213	2.286	2.547	2.513

表 10 t 検定の結果(p 値)

(a)「整数の基数変換」に関する授業の印象

	女声女体 VS 男声男体	女声女体 VS 女声男体	女声女体 VS 男声女体	男声男体 VS 女声男体	男声男体 VS 男声女体	女声男体 VS 男声女体
丁寧さ	0.116	0.205	0.354	0.006	0.485	0.030
熱意	0.886	0.091	0.808	0.113	0.692	0.051
分かりやすさ	0.002	0.098	0.165	0.000	0.072	0.003
面白さ	0.193	0.202	0.435	0.898	0.560	0.523
メリハリ	0.677	0.006	0.408	0.013	0.662	0.035
内容理解	0.015	0.104	0.968	0.000	0.012	0.081

(b)「小数の基数変換」に関する授業の印象

	女声女体 VS 男声男体	女声女体 VS 女声男体	女声女体 VS 男声女体	男声男体 VS 女声男体	男声男体 VS 男声女体	女声男体 VS 男声女体
丁寧さ	0.029	0.305	0.410	0.003	0.145	0.070
熱意	0.370	0.031	0.131	0.164	0.536	0.390
分かり	0.134	0.674	0.255	0.083	0.682	0.156

やすさ						
面白さ	0.734	0.783	0.073	0.567	0.032	0.196
メリハリ	0.159	0.004	0.154	0.095	0.975	0.103
内容理解	0.071	0.494	0.221	0.413	0.550	0.739

表 9 の授業に対する理解度の男女別平均を見てみると，「整数の基数変換」の場合，男性に対しても女性に対しても「女声・男体」が最も高評価であり，続いて「女声・女体」が高評価であった．一方，「小数の基数変換」の場合，男性の受講生に対しては「女声・女体」が最も高評価であり，続いて「男声・女体」が高評価であった．これに対し，女性の受講生に対しては「女声・男体」が最も高評価であった．次点は「男声・女体」であったが，「女声・女体」や「女声・女体」とほぼ同じ評価であった．

4.2. t 検定による分析

まず，「整数の基数変換」と「小数の基数変換」のそれぞれにおいて，講師キャラクタの性別（外見と声）の組み合わせ方により受講者（回答者）が受ける授業の印象や講師の印象，授業の理解度に統計的に有意な差が生じるかどうかを確かめるために，データクリーニング後のデータを用いてウェルチのt検定（両側分布，非等分散の 2 標本を対象とするt検定）を行った．結果を表 10～表 12 にまとめる．なお，各表に示した数値はt検定を行ったときのp値であり，p値が1%未満の場合は赤いセルで，5%未満の場合は水色のセルで表している．

表 11 講師キャラクタの印象に対するｔ検定の結果(p値)

	女声 女体 VS 男声 男体	女声 女体 VS 女声 男体	女声 女体 VS 男声 女体	男声 男体 VS 女声 男体	男声 男体 VS 男声 女体	女声 男体 VS 男声 女体
温かさ	0.166	0.441	0.159	0.650	0.989	0.653
外向的	0.165	0.089	0.219	0.615	0.879	0.528
信頼	0.215	0.799	0.041	0.393	0.388	0.114
真面目	0.086	0.952	0.257	0.162	0.535	0.372
知的さ	0.195	1.000	0.918	0.251	0.220	0.927
明るさ	0.651	0.397	0.675	0.633	0.366	0.221
自信	0.883	0.268	0.835	0.205	0.958	0.167
好感	0.889	0.825	0.558	0.730	0.470	0.763

表 12 授業の理解度に対するｔ検定の結果(p値)

	女声 女体 VS 男声 男体	女声 女体 VS 女声 男体	女声 女体 VS 男声 女体	男声 男体 VS 女声 男体	男声 男体 VS 男声 女体	女声 男体 VS 男声 女体
整数	0.328	0.261	0.214	0.039	0.746	0.023
小数	0.146	0.973	0.263	0.170	0.748	0.293

表 10 の(a)「整数の基数変換」によると，表 3(a)で最も高評価だった「女声・男体」と他の組合せを比較した結果，「男声・男体」や「男声・女体」とは複数の評価項目で有意な差が認められたが，「女声・女体」に対しては「メリハリ」のみだった（p=0.009）．声が同じ「女声」であるにもかかわらず．有意な差が出たのは，声と外見との一致/不一致が影響を与えている可能性がある．一方，表 10 の(b)「小数の基数変換」では，あまり有意な差が見られなかった．

表 11 の講師キャラクタの印象に関しては，ほとんど有意な差が見られなかった．唯一有意な差が出たのは「女声・女体」と「男声・女体」の間の「信頼」に関してのみであり，声のみが異なる組み合わせ方なので，声と外見の性別の一致が良い影響を及ぼした可能性がある．

表 12 の授業の理解度によると，「整数の基数変換」で「女声・男体」と他の組合せとの間で有意な差が観測され，「女声・男体」の優位性が明らかになった．一方，「小数の基数変換」では有意差は認められなかった．

次に，受講者（回答者）の性別によって授業の印象や講師の印象，授業の理解度に統計的に有意な差があるかどうかを調べるために，男女間でウェルチのt検定（両側分布，非等分散の 2 標本を対象とするt検定）を行った．結果を表 13～表 15 にまとめる．

表 13 男女間のt検定（p値）

(a) 整数の基数変換に関する授業の印象

	女声 女体	女声 男体	男声 女体	男声 男体
丁寧さ	0.490	0.083	0.740	0.233
熱意	0.651	0.127	0.069	0.405
分かりやすさ	0.159	0.687	0.602	0.158
面白さ	0.359	0.092	0.830	0.225
メリハリ	0.001	0.041	0.168	0.980
内容理解	0.618	0.769	0.290	0.264

(b) 小数の基数変換に関する授業の印象

	女声 女体	女声 男体	男声 女体	男声 男体
丁寧さ	0.305	0.772	0.764	0.481
熱意	0.093	0.049	0.139	0.371
分かりやすさ	0.040	0.736	0.578	0.998
面白さ	0.181	0.141	0.489	0.200
メリハリ	0.131	0.119	0.174	0.877
内容理解	0.102	0.879	0.090	0.323

表 14 講師キャラクタの印象における男女間のt検定（p値）

	女声 女体	女声 男体	男声 女体	男声 男体
温かさ	0.005	0.109	0.348	0.399
外向的	0.893	0.599	0.098	0.135
信頼	0.085	0.925	0.370	0.751

真面目さ	0.444	0.231	0.495	0.024
知的さ	0.303	0.926	0.602	0.195
明るさ	0.232	0.929	0.116	0.638
自信	0.804	0.545	0.187	0.028
好感	0.052	0.348	0.181	0.493

表 13 の授業の印象(男女間)によれば，統計的に有意な差が出た項目は，整数の基数変換では「メリハリ」のみであり，有意水準は「女声・女体」の場合で1%，「女声・男体」の場合で 5%であった．それぞれの場合で平均を比較してみると，「女声・女体」では男性の方が高評価で，「女声・男体」では女性の方が高評価だった．本来，声に関する印象と考えられる「メリハリ」だが，回答者と異なる性別の外見の方が高評価だった．一方，「小数の基数変換」では「女声・女体」の場合の「分かりやすさ」と「女声・男体」の場合の「熱意」において有意な差(有意水準 5%)が見られた．いずれも講師キャラクタの外見の性別と異なる性別の回答者の方が高評価だった．

表 14 の講師キャラクタの印象(男女間)によれば，「女声・女体」の場合の「温かさ」と「男声・男体」の場合の「真面目さ」および「自信」で有意な差が見られた．有意水準は「温かさ」に関しては 1%，「真面目さ」と「自信」に関しては5%であった．平均を比較してみると，「女声・女体」の場合の「温かさ」では男性の方が高評価で，「男声・男体」の場合の「真面目さ」と「自信」では女性の方が高評価であった．いずれの場合でも講師キャラクタの性別(外見と声)と異なる性別の回答者の方が高評価であった．

表 15 の授業の理解度(男女間)によれば，「女声・女体」の「少数の基数変換」でのみ有意な差が見られた．有意水準は5%であった．

表 15 授業の理解度における男女間のt検定(p値)

	女声 女体	女声 男体	男声 女体	男声 男体
整数の基数変換	0.151	0.923	0.943	0.922
小数の基数変換	0.016	0.662	0.643	0.739

4.3. 先行研究との比較

次に，四戸らの先行研究[2]から講師の上半身が映った映像が配置された授業動画「音声＋映像」(講師の外見も声も女性)に対する印象評価データを入手し，条件を等しくするため，「女声・女体」の場合の印象評価データとの間でウェルチのt検定(両側分布，非等分散の 2 標本を対象とするt検定)を行った．「音声＋映像」に関するアンケート調査に参加した回答者の性別・年齢構成を表 16 に示し，t検定の結果を表 17 に示す．

表 16 先行研究[2]でアンケート調査「音声＋映像」に参加した回答者の性別・年齢構成(データクリーニング後)

	人数
男性	72
女性	81
計	153

表 17 のt検定の結果によれば，「知的さ」を除くすべての印象項目に対し，「音声＋映像」の方が高評価であり，多くの項目(「整数の基数変換」の「熱意」と「メリハリ」，「小数の基数変換」の「丁寧さ」「熱意」「メリハリ」，講師の印象の「真面目さ」「知的さ」を除く)で有意な差も見られた．したがって，CG で作られた講師キャラクタより生身の講師の方が好印象であると言えるが，授業の理解度という点では有意な差は見られなかった．

表 17 先行研究[2]の「音声＋映像」と「女声・女体」の間でのt検定の結果(p値と平均)

(a)「整数の基数変換」に関する授業の印象

	p値	音声＋映像	女声・女体
丁寧さ	0.175	2.144	2.314
熱意	0.000	2.915	3.607
分かりやすさ	0.255	2.131	2.279
面白さ	0.064	3.418	3.657
メリハリ	0.000	2.889	3.507
内容理解	0.521	2.013	2.100

(b)「小数の基数変換」に関する授業の印象

	p値	音声＋映像	女声・女体
丁寧さ	0.027	2.137	2.414
熱意	0.000	2.863	3.714
分かりやすさ	0.728	2.425	2.471
面白さ	0.090	3.333	3.543
メリハリ	0.000	2.935	3.571
内容理解	0.936	2.275	2.286

(c) 講師(キャラクタ)の印象

	p値	音声＋映像	女声・女体
温かさ	0.000	2.405	3.921
外向的	0.000	2.791	3.557
信頼	0.010	2.503	2.821
真面目さ	0.701	2.065	2.107
知的さ	0.939	2.516	2.507
明るさ	0.000	2.536	3.800
自信	0.000	2.575	3.007
好感	0.000	2.275	3.507

(d) 授業の理解度

	p値	音声＋映像	女声・女体
整数部分	0.462	2.065	2.007

小数部分	0.292	2.150	2.050

4.4. 重回帰分析

各印象評価項目が授業の理解度に及ぼす影響を分析するために，重回帰分析を行った．具体的には，データクリーニング後のデータを講師キャラクタの外見や声に関係なくすべて統合し，Q1 や Q5 において「わからない」と回答した人のデータをすべて削除した上で．授業の理解度を目的変数，各印象評価項目を説明変数とする重回帰分析（変数増加法）を行った．その結果となる偏回帰係数と標準偏回帰係数およびp値を表 18 に示す．なお，自由度修正済み決定係数は，「整数の基数変換」の場合で 0.368，「小数の基数変換」の場合で 0.022 であり，良好な結果とは言えない．

表 18 重回帰分析の結果

(a)「整数の基数変換」の場合

	偏回帰係数	標準偏回帰係数	p値
内容理解	0.272	0.251	0.000
整数の基数変換を学んだことがある	0.337	0.297	0.000
知的さ	0.141	0.124	0.002
分かりやすさ	0.137	0.123	0.004
信頼	0.092	0.080	0.044
定数項	0.092		0.491

(b)「小数の基数変換」の場合

	偏回帰係数	標準偏回帰係数	p 値
小数の基数変換方法を覚えている	0.120	0.128	0.005
真面目さ	-0.135	-0.115	0.025
丁寧さ	0.103	0.095	0.063
定数項	2.043		0.000

表 18 によれば，「整数の基数変換」に関する授業の理解度に正の影響を与えている評価項目は，「整数の基数変換」に対する「内容理解」と「分かりやすさ」，学習経験を問う「10 進数の整数を 2 進数に変換する方法を学んだことがある」，講師に対する印象の「知的さ」と「信頼」であり，負の影響を与えている評価項目はなかった．授業の内容を分かりやすいと捉えるレベルの人が内容を理解できたということと授業内容の信頼性に関わる「知的さ」と「信頼」が授業の理解度に影響を及ぼすのは，ある意味当然と言える結果であった．一方，「小数の基数変換」に関する授業の理解度に正の影響を与えている評価項目は，学習経験を問う「10 進数の小数を 2 進数に変換する方法を覚えている」と講師に対する印象の「丁寧さ」であった．逆に負の影響を与えている項目は，講師に対する印象の「真面目さ」であった．しかしながら，自由度修正済み決定係数が 0.022 と極めて低く，意味のある結果が出ているとは言えない．

5. まとめ

本稿では，TVML[3]を用いて作成した授業動画が受講生に与える印象を男女 743 人が参加するアンケート調査により調べ，分析した．TVML はテキスト形式の台本からテレビ番組風の CG アニメーションを作成することができるフリーツールであり，本稿では CG キャラクタの外見（男体と女体）と合成音声（男声と女声）を組み合わせ，4 種類の講師キャラクタを作成した．

アンケート調査の結果得られたデータに対し，4 種類の講師キャラクタの間でt検定を行った結果，講師キャラクタの外見が「男体」，合成音声が「女声」のとき，授業の印象や講師の印象，授業の理解度において高評価であった．

一方，回答者を男女に分け，講師キャラクタの性別（外見と声）が回答者に及ぼす影響を調べたところ，講師キャラクタの性別と回答者の性別が異なるとき，高評価になりやすいことがわかった．ただし，授業の理解度に関しては，統計的に有意な差は見られなかった．

また，TVML を用いて作成された授業動画と人間の女性が講師役を務めた授業動画を比較してみたところ，全体的な印象は人間の女性の方が高評価であったが，授業の理解度に関しては．統計的に有意な差は見られなかった．

さらに，授業の理解度を目的変数，各評価項目を説明変数とする重回帰分析（変数増加法）を行ってみたが，自由度修正済み決定係数が低く，あまり良好な結果は得られなかった．

最後に本研究の課題について述べ，その改善を今後の課題とする．まず本稿では，授業動画の印象に着目したため，TVML を用いて作成した授業動画が学習効果に与える影響についてはあまり詳しく調査していない．また，授業コンテンツを説明する音声には TVML に付属した音声合成エンジンで生成される音声をそのまま使用しており，抑揚や間の取り方を制御できなかった．講師キャラクタの外見や声についても標準的なものをそのまま用いた．

謝辞

本研究の一部は，福田将治奨学寄付金による助成の成果であり，ここに記して深く感謝の意を表す．

参考文献

[1] 大学 ICT 推進協議会 ICT 利活用調査部会，2020 年度 高等教育機関における ICT の利活用に関わる調査研究報告（速報版），https://ict.axies.jp/sig/52/，2022．（2024 年 1 月 4 日閲覧）

[2] 四戸なつみ，熊本忠彦，オンライン授業において講師映像が受講生に及ぼす影響の分析，DEIM Forum 2021，C31-1，2021．

[3] TVML，https://tvmllab.com/ （2024 年 1 月 4 日閲覧）

[4] 上間朱梨，栂ちかこ，教師役の声と表情が学習者に

与える印象，鹿屋体育大学学術研究紀要，Vol.62，pp.153-166，2023.
[5] 柿澤寿信，生徒と教員の性別の組み合わせが成績に与える影響の検証，NIER Discussion Paper Series，No.5，pp.1-33，2017.
[6] 斉藤雄路，松本慎平，寺西大，動画教材における音声合成の有用性評価に関する基礎的研究，教育システム情報学会第46回全国大会，Vol. 46，P1-01，pp. 1-2，2021.
[7] Aquestalk2 ， https://www.a-quest.com/products/aquestalk_2.html（2024 年 1 月4日閲覧）
[8] インターネット調査会社「FastAsk」，https://monitor.fast-ask.com/ （2024 年 1 月 4 日閲覧）

プロジェクト型クラウドソーシングにおけるワーカのスキル推薦

大石　啓登†　　伊藤　寛祥††　　豊田　正史†††　　森嶋　厚行††

† 筑波大学 情報学群　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
††† 東京大学 生産技術研究所　〒 153—8505 東京都目黒区駒場 4–6–1
E-mail: †s2213502@u.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp, †††toyoda@tkl.iis.u-tokyo.ac.jp

あらまし　クラウドソーシングは新しい働き方の一つとして広く受け入れられている．その中でもプロジェクト型クラウドソーシングはポピュラーな形態の一つである．これまでのクラウドソーシングの技術的研究は，雇用者側の利益を考えた研究が多く，クラウドワーカを支援する研究は限定されている．本論文では，クラウドワーカがどのようなスキルを身につけるべきかという問題に注目し，現在のワーカの登録スキルと，タスク市場のデータを元に期待報酬を増加させるような関連スキルを推薦する手法を構築し，評価する．具体的には，クラウドソーシングタスク市場のデータ分析と推薦アルゴリズムの評価により，次の結果が得られた (1) タスク市場でワーカが登録しているスキルは必ずしも期待報酬を上げるようなスキルになっておらず，スキル推薦が有効である可能性があること，(2) 現在の登録スキルに関連するスキルを推薦することにより，期待報酬が上がること (3) 3 つまでのスキルの推薦については，貪欲型の探索アルゴリズムでも全探索アルゴリズムと遜色なく，高速に計算できること．

キーワード　クラウドソーシング，スキル推薦，ワーカ支援，リスキリング

1 はじめに

クラウドソーシングは新しい働き方の一つとして広く受け入れられている．2020 年の調査では，日本国内のクラウドソーシングサービスの市場は約 2950 億円に達している [14]．特に，クライアントのタスクに対して興味を示した複数のワーカからクライアントが委託先を選択する「プロジェクト型」は重要なカテゴリの一つであり，多くのクラウドソーシングサービスで提供されている．

このように，クラウドソーシングは社会に受け入れられている一方で，改善が望まれる様々な課題を持つ．Gross-Amblard らの調査 [3] によればクラウドワークス [1] のワーカの 68%がプラットフォームに将来望む機能やサポートとして「新しいスキルを学ぶこと」を挙げており，ワーカに対して新しいスキルを推薦したり，その習得をするためのサポートをしたりといった支援は重要である．この調査に加え，Kim ら [8] は Upwork などを活用して働く所謂プラットフォームワーカがスキル構築について困難に直面していることを指摘している．Kim らによればプラットフォームワーカは基本的に制度的な支援なしにスキルを習得する必要があったり，クライアントのスキル要求を予測して備える必要があったりとスキル習得のための障害が存在する．

このような背景のもと，本稿では，タスク市場の需要を考慮して，ワーカが新たに習得すべきスキルを推薦する問題に取り組む．習得すべきスキルとは何かという問題は難しいが，本論文では現在のタスク市場において報酬を増やすことができるスキルとする (詳細は 2.3 節で定義する)．

1：https://crowdworks.jp

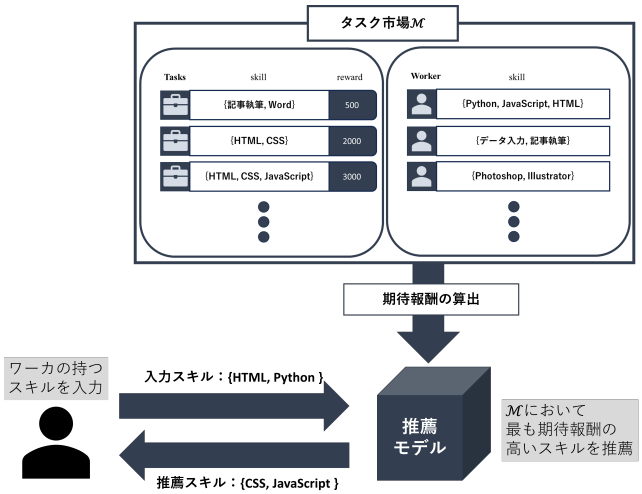


図 1: 推薦システム概要

図 1 に本稿で扱う問題の概要を示す．タスク市場 $\mathcal{M}$ にはタスクとワーカが登録されており，各ワーカはそれぞれが異なるスキルの集合を持ち，また，各タスクはそのタスクを遂行するのに必要なスキルの集合と，そのタスクの実施に対する報酬の情報を持つ．本稿ではこのタスク市場にある情報を考慮して，あるワーカが新たに身に着けるべきスキル集合を推薦する．

まず，本稿では，実際のクラウドソーシングサービスのタスク市場データを利用して，ワーカのスキルとタスクの関係について調査し，提案するようなスキル推薦システムの有効性について議論を行う．次に，現在持っているスキルと関係する他のスキルを身につけることが実際に市場でどれぐらい意味があるかを検証する．最後に，大量のタスク，ワーカ，スキル情報が存在するような実際のタスク市場において，効率的なアルゴリ

ズムによる上記のようなスキル推薦が可能かどうか検証する．

まとめると，本稿におけるリサーチクエスチョンは以下のとおりである．

(RQ1) 実際のタスク市場において，ワーカが宣言しているスキルと，タスク実施のために求められているスキルの関係はどのようなものか

(RQ2) 実際のタスク市場において，関連するスキルを推薦することはどれぐらい意味があるか

(RQ3) 効率の良いスキル推薦アルゴリズムは実現可能か

これらに関して，実験を通じて得られた発見は次のとおりである．(1) 実際のタスク市場において，ワーカはどのようなスキルが求められているかをあまり考慮せずにスキルを登録しており，スキル推薦は有効であると考えられる．(2) 現在のスキルに関連する推薦されたスキルをワーカが身につけることによって，報酬の期待値は大幅に増加する．(3) 特定の条件において，貪欲法に基づく効率的なアルゴリズムが存在する．

本稿の構成は次のとおりである．2 節では関連研究について説明する．3 節では用語と問題を定義する．4 節ではスキル集合の推薦アルゴリズムを説明する．5 節では実験と結果について示す．6 節はまとめである．

2 関 連 研 究

本研究に関連する研究として，①オンラインフリーランシングプラットフォームを題材としたフリーランス支援の研究，②マイクロタスク型クラウドソーシングにおけるタスク割り当ての研究，③求人サービスのデータを活用した研究が存在する．それぞれのアプローチについて説明したうえで，本研究との違いを述べる．

2.1 オンラインフリーランシングプラットフォームを題材としたフリーランス支援の研究

プロジェクト型クラウドソーシングプラットフォームは Amazon Mechanical Turk[2](以下 MTurk) などのマイクロタスク型と区別してオンラインフリーランシングプラットフォームとも呼ばれ，Upwork を中心としたオンラインフリーランスワーカ支援の研究が多く行われている．

例えば，Jarrahi ら [7] は Upwork を使用するギグワーカーへのインタビューやウォークスルー分析を通して Upwork における管理機能について調査している．また，Blyth ら [1] は Upwork のフリーランスワーカとクライアントにインタビューを行い，その分析からフリーランスワーカが「プロフィールを充実させる」「実際にスキルを披露する」などの計 5 つの戦略をとることで，より多くの仕事を獲得することができると結論付けている．

以上の様な研究のほとんどでオンライン労働市場での仕事探しが困難であることやワーカのセルフブランディングの重要性が説かれており，オンライン労働市場で働く人々が新しいスキルや資格を習得する所謂リスキリングが継続的に仕事を獲得していくために必要不可欠であることが分かる．

その一方で Kim ら [8] はオンラインフリーランシングプラットフォームのワーカがスキル構築のパラドックスに直面している点を指摘している．Kim らはスキル構築の上で直面する「責任化」と個人の社会的立場がスキル要求やスキル構築行動に影響を与える「異質化」について特に着目している．特に「責任化」は本研究と関連が深く，スキル開発の個人の経験が (a) 従業員と雇用者の共同作業から労働者中心の責任へのシフト (b) 職務経歴書ベースのスキル構築からクライアントと市場を介したスキル構築へのシフト (c) 構造化された Formal な学習から Informal な学習へのシフト (d) 職務中心のスキル習得から職務に必要なスキルと職務に必要な補助的活動の両方のスキル習得へのシフトの 4 つのシフトによってプラットフォームを介した仕事に必要なスキルがより複雑で無秩序になっている点が指摘されている．また，日本のクラウドソーシングプラットフォームに視点を当てた調査では，Gross-Amblard ら [3] の実施したアンケートではクラウドワークスを利用するワーカの 68%が「スキルを学ぶ機能」が将来プラットフォームに実装されることが好ましいと回答している．

以上のような研究によりワーカが収入を上げるための戦略やワーカがスキル構築をする上での問題が指摘されスキル習得の需要が存在する一方で，ワーカのスキル取得・構築をサポートするような研究は我々の知る限り存在しない．フリーランスワーカがオンラインプラットフォームを活用して働くうえでスキル構築に関する様々な問題があるからこそ，ワーカが自分の収入を上げることができるスキルを知ることが重要だと我々は主張する．

2.2 マイクロタスク型クラウドソーシングにおけるタスク割り当ての研究

マイクロタスク型クラウドソーシングにおいて，タスク割り当てに関する研究が多く行われている [6] [5]．例えば，Zheng らは MTurk などのマイクロタスク型クラウドソーシングプラットフォーム上でワーカの能力やタスクの難易度を推定しながらタスクを割り当てるフレームワーク”QASCA(A Quality-Aware Task Assignment System for Crowdsourcing Applications)” [12] を提案している．しかし，これらの研究は画像のラベル付けなどの「特定の問題を解く能力を推定し，適切な難易度の問題を割り当てる」というものであり，本研究とは問題設定が異なるため，これらの研究を応用することは不可能である．

その他のタスク割り当てに関する研究として Fan らは”iCrowd” [2] という手法を提案している．iCrowd では「自動車」や「食品」といったドメイン知識が解く問題によって異なるようなタスクにおいて，ワーカがどのドメインについてどのくらいの精度で回答できるかを推定し，適切にタスクを割り当てるといったアプローチを行っている．このような手法においても，ワーカのこなしたタスクという成果物があることが前提であり，1 種類のタスクが大量に発行されたり，クライアントが自動的・効率的にワーカにタスクを割り振るといったアプローチになっている．しかし，本研究においてスキルの推薦を

2：https://www.mturk.com

行うのはワーカやクライアントが利用するプラットフォームという第三者である．そのため，ワーカの行った仕事の成果物を得ることは不可能であり，成果物をもとにスキルの推定や推薦を行うことはできない．

以上のようにマイクロタスク型クラウドソーシングプラットフォームにおける研究は問題設定や当事者ではなく第三者が推薦を行うという点から応用することが困難である．

2.3 求人サービスのデータを活用した研究

コンピュータとインターネットの普及により，知識経済の時代が到来したと言われる現代において，個人の知識やスキルが極めて重要とされている．このような背景から LinkedIn や Lagou などの求人サービスのデータを活用することで，スキルの市場価値を推定や，ワーカの持つ特定スキルの能力の推定，ワーカの履歴書と求人情報からマッチングを行う研究などが行われている．特にその人に適した仕事を推薦する重要性から履歴書と求人のマッチングは活発に行われている．

まずスキルの市場価値の推定に関して，Sun ら [9] はスキル価値をそのスキルのみを必要とする仕事の給与とみなし，求人情報から得られる必要なスキルと職務給を入力として，一つのスキルを与えた際のスキル価値と複数のスキルを与えた際の職務給を推定している．

次にワーカの持つ特定スキルの能力推定に関して，Yan ら [11] は「あるスキルに精通している」という事のコンテキストが求人により異なる (例えば「機械学習ができる」人材を探す場合，大学の教授の様な理論的な知識が豊富な人材が必要か，限られたリソースで大規模なモデルを動かすことが出来る開発者が必要かは求人によって違う) という前提の上で，ワーカのプロフィールや求人広告，LinkedIn の独自の機能である Skill Endorsements[3] を活用することでコンテキストに応じたスキルレベルの推定を可能にしている．

履歴書と求人情報のマッチングの研究として，Wang ら [10] は履歴書と求人情報を与えた際に，与えられた 2 つの情報に加え，求人を出しているクライアントの過去の採用実績 (採用された人の履歴書と求人情報) を使用することで，80 %以上の精度でワーカがその求人に採用されるかどうかを予測することに成功している．

ここまで，求人サービスのデータを使用した主な先行研究について述べてきたが，これらの研究を本研究に応用するにはいくつかの障害が存在する．まず，スキルの市場価値の研究について Ying らのアプローチはあくまでワーカの能力ではなく市場価値に着目しており，「時給 600 円のスキル A と時給 500 円のスキル B の両方を使用できる場合，そのワーカの時給が 2000 円になる」というようなスキルとスキルの相乗効果については考慮されていない．次に Yan らのアプローチは Skill Endorsements という LinkedIn 独自の機能に依存したものであり，他のプラットフォームのデータでは手に入らない可能性が高い．最後に，履歴書と求人情報のマッチングに関しても，プロジェクト型クラウドソーシングプラットフォームではワーカの求人が手に入らないといった問題が存在し履歴書のデータに基づくワーカのスキル推定ができないため，スキル推薦によるワーカ支援という本研究の問題に適応することが出来ない．

3：https://www.linkedin.com/help/linkedin/answer/a565106

3 用語定義

問題を説明するために，いくつかの用語を導入する．まず，タスク市場 $\mathcal{M} = (T, W, S, r, ts, ws)$ を定義する．ここで利用する記号は次の通りである．

- タスク集合 T，ワーカ集合 W，スキル集合 S
- タスクを実施すると得られる報酬 $r : T \to \mathbb{R}$
- 各ワーカが持つスキル集合が $ws : W \to 2^S$，各タスクの実施に必要なスキル集合 $ts : T \to 2^S$

そのほかの記号や関数の定義については表 1 に示す．

タスクに関する期待報酬 あるタスク $t \in T$ に関する期待報酬とは，そのタスクの報酬を，そのタスクを実施可能なワーカの数で割った値 $r(t)/pw(t)$ である．ここで，$pw(t)$ は，t を実施できるワーカの人数 (すなわち，$pw(t) = |\{w|w \in W \wedge ts(t) \subseteq ws(w)\}|$ である) である．

スキル集合の期待報酬 タスク市場 $\mathcal{M}$ において，あるスキル集合 $s \in 2^S$ の期待報酬 $rs(s)$ とは，そのスキル全てを必要とするタスクに関する期待報酬である．具体的には次の通りとなる．

$$rs(s) = \sum_{\forall_t \ s.t. \ ts(t)=s} \frac{r(t)}{pw(t)}$$

ワーカの期待報酬 タスク市場 $\mathcal{M}$ において，あるワーカ $w \in W$ の期待報酬 $rw(w)$ とは，そのワーカのスキルがあればできる全てのタスクの期待報酬の和である．すなわち，

$$rw(w) = \sum_{\forall_t \ s.t. \ ts(t) \subseteq ws(w)} \frac{r(t)}{pw(t)}$$

スキル推薦問題 本論文で取り扱う問題は，あるワーカ $w \in W$ と新しく身につけたいスキル数 n が与えられたとき，現在のタスク集合 T においてそのワーカの期待報酬を最大化するようなスキルの集合 $s'_w \subseteq S - ws(w)$ を推薦することである．式では次のように表される．

$$\arg\max_{s'_w} \sum_{\forall_t \ s.t. \ ts(t) \subseteq (s'_w \cup ws(w))} \frac{r(t)}{pw(t)}$$

ここまで説明した定義に基づき，ワーカがこの後身に付ける価値があるスキル集合 s'_w を求めるアルゴリズムを次章で説明する．

4 推薦アルゴリズム

3 章で定義したスキル推薦問題は探索問題であるため，解を求めるためには探索アルゴリズムを構築する必要がある．本章では探索空間を表現するグラフ (skill-dev graph) を説明した後に，その空間を探索する複数のアルゴリズムについて説明する．

表 1: 本研究における記号と定義

記号	定義
T	タスクの集合
W	ワーカの集合
S	スキルの集合
$ts(t)$	あるタスク t の実施に必要なスキル集合
$ws(w)$	あるワーカ w の持つスキル集合
$r(t)$	あるタスク t を完了すると得られる報酬
$pw(t)$	あるタスク t を実施可能なワーカの数
n	新しく学ぶスキル数
s'_w	ワーカがこの後身に付ける価値のあるスキル集合

4.1 スキル開発グラフ (skill-dev graph)

あるタスク市場におけるにおけるスキル開発グラフとは，そのタスク市場において，ワーカがスキルを増やすと期待報酬がどれぐらい増えるかを表したものである．図 2 に例を示す．ここで，各ノードはスキル集合と，そのスキル集合の期待報酬が書かれている．エッジは，あるノードから別のノードが表すスキル集合にふやすためにはいくつのスキルを増加させる必要があるかを示している．例えば，グラフ中の $(\{html, English\}, 5000) \xrightarrow{2} (\{html, javascript, css\}, 1000)$ は，現在スキルとして html, English を持つワーカが新たに 2 個のスキル javascript, css を学ぶことで，追加で 1000 の期待報酬が得られることを表す．

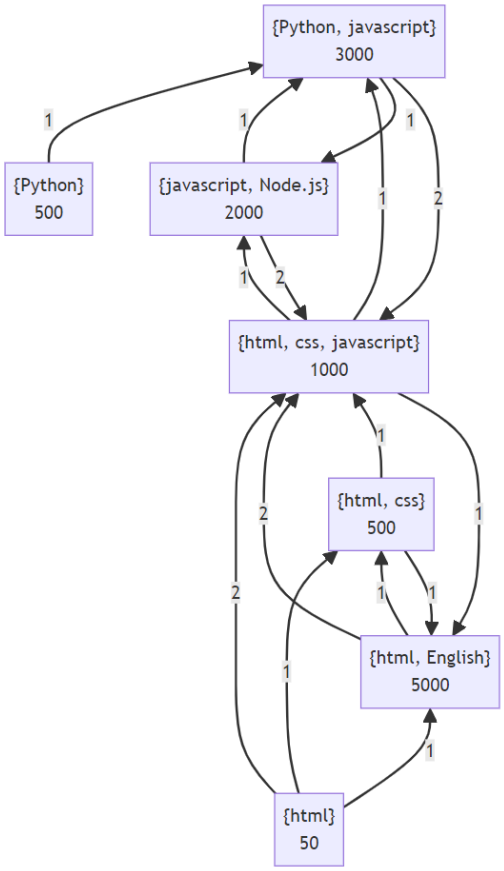


図 2: skill-dev graph の例

Algorithm 1 推薦アルゴリズムの全体像

Input: $\mathcal{M}, w$
Output: s'_w
1: Construct $G_{\mathcal{M}}$.
2: $v_0 \leftarrow G_{\mathcal{M}}.update(ws(w))$
3: return $recommend(G_{\mathcal{M}}, v, n, V' \cup \{v\})$

フォーマルには，あるタスク市場 $\mathcal{M}$ のスキル開発グラフ $G_{\mathcal{M}} = (V, E, attr)$ は次のよう定義する．

$$V = \{(ts : ts(t), r : rt(t)) \mid t \in T\}$$

$$(v_1, v_2) \in E \; if \; v_2.ts - v_1.ts \neq \phi \; \wedge v_1.ts \cup v_2.ts \neq \phi$$

$$attr : E \rightarrow \mathbb{N} \; where \; attr(v_1.ts, v_2.ts) = |v_2.ts - v_1.ts|$$

すなわち，タスク市場に存在するタスクを実施するのに必要なスキル集合毎に，ノードが作られる．各ノードは，そのスキル集合を必要スキルとするタスクの期待報酬を属性 $attr$ として持つ．また，エッジに関しては，それぞれのスキル集合に重複が存在して，追加のスキルが必要な場合のみ，それらの間に有効エッジがはられる．各エッジには，追加が必要となるタスクの数が表現される．

表 2: skill-dev graph の記号と定義

記号	定義
V	ノード集合
E	エッジの集合
$attr(u, v)$	ノード u からノード v へのエッジが持つ値

スキル開発グラフが表す探索空間． スキル開発グラフでは，定義により，ノード同士で①重複するスキルが存在し，かつ②そのノードに存在しないスキルも存在する場合にのみエッジを張っている．つまり，ワーカが持つスキルと関連性がある程度あるスキルのみが推薦候補となる．例えば「HTML」のスキルを持つワーカに「JavaScript」を推薦するといったことである．一方，「公認会計士」や「普通自動車運転免許」のように関連性が薄いと考えられるスキルは，期待報酬を大幅に上げるとしても探索空間に含まれない可能性が上がる (たまたまそれらを組み合わせたタスクがあれば探索空間に現れる)．このように絞られた探索空間がどの程度の期待報酬の増加を実際にもたらすかは本論文の RQ2 であり，実験によって検証する．

4.2 スキル推薦アルゴリズムの全体像

スキル推薦のアルゴリズムの全体像をアルゴリズム 1 に示す．まず，1 行目でスキル開発グラフを構築する．ここで，$G_{\mathcal{M}}$ の中に，スキル集合 $ws(w)$ に対応するノードが必ずしも存在しないことに注意して欲しい．なぜなら，これらのノードが対応するスキル集合は，$\mathcal{M}$ に存在するタスクが必須スキル集合としてあげているスキル集合であり，ワーカが持ちうるスキル集合ではないからである．2 行目では，ワーカ w が持つスキル集合 $ws(w)$ を用いて，そのスキル集合に対応するノード v_0 を $G_{\mathcal{M}}$ に追加するように更新する．具体的には次のようになる．

（1） $ws(w)$ に含まれる任意のスキル集合に対応する G_M 中のノード群 V' を構築する．

（2） v_0 からノード v' のリンク先ノード全てにリンクを張る．ノード v を追加する．

最後に，3 行目で G_M に対して 推薦アルゴリズムを適用する．その引数は，スキル推薦グラフ G_M, スタートノード v_0, 推薦スキル数 n, 訪問済み (とみなす) ノードの集合 $V' \cup \{v_0\}$) である．推薦アルゴリズムは，訪問済みノード以外のノードを探索して，推薦するタスク集合を一つ求める．

4.3 3 種類のスキル推薦アルゴリズム

本稿では $recommend(G_M, v_0, n, V' \cup \{v\})$ の具体的な実装として，次の 3 つのアルゴリズムを検討する．

- **All**: 全探索アルゴリズム．厳密に，最も大きな期待報酬をもたらす n 個の新しいスキル集合を返す．
- **Greedy**: あるノードから次に探索するノードのエッジを選択するときに，エッジの宛先の中から n 以内のスキルの追加で，最も大きな期待報酬をもたらすノードを選ぶ．
- **Greedy-Ratio**: あるノードから次に探索するノードのエッジを選択するときに，エッジの宛先の中から n 以内のスキルの追加で，追加スキル 1 つあたりの平均の期待報酬が最も大きいノードを選ぶ．

全探索アルゴリズム All: このアルゴリズムは，グラフを全探索し，次の条件を満たすパスを求め，そのパスが表す新たなスキル集合を返すものである．

- v_0 から開始する．このパスを $v_0.v_1, \ldots, v_m$ とする．
- すべてのエッジ $v_i \xrightarrow{k} v_j$ の attr k の総和が n 以下である．
- 「パス上のノードのスキル集合がもたらす期待報酬の総和」が最大である．

ここで，「パス $v_0, \ldots, v_i, \ldots, v_m$ 上のノードのスキル集合がもたらす期待報酬の総和」は次のように計算する．

（1） v_0 から始まって v_{i-1} までのノード上のスキル集合がもたらす期待報酬の総和を $rsum'$ とし，この時点でのパス上のノード集合を V' とする．ただし，$i = 0$ の時は，$rsum' = 0$, $V' = \phi$ とする．また，v_0 を V' に足すときは，そのもととなるノード集合 (アルゴリズム 1 で説明) も加える．

（2） このとき，パス $v_0, \ldots, v_i$ 上のスキル集合がもたらす期待報酬の総和は，$rsum'$ に次を足したものである．

- v_i が表す期待報酬 $v_i.r$
- v_i が表すスキル集合の部分集合に対応するノードであって，V' に含まれないノードの期待報酬．例えば，図 2 の例において v_{i-1} がスキル集合 {javascript, Node.js} に対応するノードであり，v_i が {Python, javascript} のノードである場合，v_i の期待報酬だけでなく，その部分集合である {Python} ノードの期待報酬も加算する．

貪欲アルゴリズム Greedy: この実装では，n 以下のスキルを追加するすべてのパスを探すのではなく，探索時に今のノードから次に探索するノードのエッジを選択するときに，1 ホップ先のノードに向かうエッジの宛先となるノードの中から，n 以内のスキルの追加で，最も大きな期待報酬をもたらすノードを選ぶ．アルゴリズム 2 に疑似コードを示す．アルゴリズム 1 の 3 行目で recommend 関数のところで greedy 関数を呼び出す．各ノードがもたらす期待報酬の計算方法は，全探索で説明したものと同じである．

Algorithm 2 貪欲法による探索アルゴリズム greedy

Input: $G_{\mathcal{M},ws}, v, n, V'$
Output: s'_w
  $s'_w \leftarrow \phi$
  if there is $v' \in v.dest$ s.t. $v'.r$ is the highest among $v.dest$ and $v' \notin V'$ and $attr(v, v') \leqq n$ **then**
    add $v'.s - v.s$ and $greedy(v', n - m, V' \cup \{v'\})$ to s'_w
  end if
  return s'_w

貪欲アルゴリズム Greedy-Ratio: この実装は，Greedy と基本的には同じであるが，1 ホップ先の宛先ノードの選択の基準が異なる．すなわち，次に探索するノードのエッジを選択するときに，エッジの宛先の中から n 以内のスキルの追加で，単純にそのノードの追加がもたらす期待報酬が最も大きいものを選ぶのではなく，その値を追加が必要なスキル数で割る．すなわち，追加スキル 1 つあたりの平均の期待報酬 $v'.r/attr(v, v')$ が最も大きいノードを選ぶ．このアルゴリズムは，アルゴリズム 2 の if 分の条件の $v'.r$ をその式で置き換えるだけであるので本稿では省略する．

5 実　　験

本稿の 3 つのリサーチクエスチョンに答えるための実験を行う．本実験の概要を図 3 に示す．本実験では，クラウドワークス社のデータを利用し，RQ1 のスキルに関する分析をまず行う．次に，スキル推薦のアルゴリズムを利用してその結果の分析を行う．その際，全探索 All, 貪欲法 (Greedy, Greedy-Ratio) の結果と実行時間の違いを比較し，RQ2 と RQ3 について検証する．

5.1 利用データ

本実験で利用するデータの概要を表 3 に示す，クラウドワークス社の 2022 年の 12 月に発行されたタスクの一部と 2022 年 12 月におけるアクティブワーカ情報を利用した．タスクについては全てのタスクを利用せず，タスクの募集・応募欄に必須スキルの記載がある 3157 件だけを対象とした．したがって，本実験で出てくる期待報酬は，現実の市場における期待収入に比べて大幅に小さくなることに注意が必要である．

利用データはそのままでは表記揺れや省略があるため，次のデータの前処理を行った．

（1） **ワーカ登録スキルの拡張** 応募データを見て，あるワーカがタスクを実施した記録がある場合には，そのタスクの必須スキルを，ワーカの登録スキルに追加する．

（2） **スキル表記の統一** スキルの表記揺れ等を統一する．

ワーカ登録スキルの拡張 ワーカがタスクに応募して実施した場合には，そのタスクの必須スキルを持っていると考えることが自然である．したがって，表 3 の「応募データ」を見て，

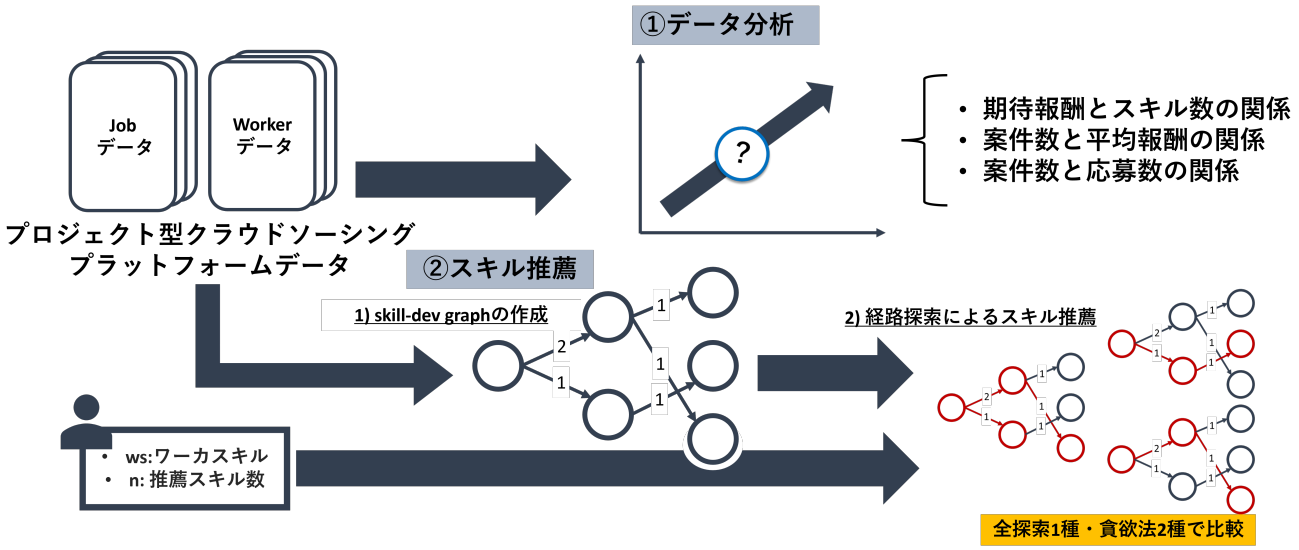


図 3: 実験の流れ

表 3: 使用したデータ．タスクに関しては全てのデータを利用したわけではなく，スキルが明示的に支持された一部のタスクだけを利用したことに注意．したがって実験結果では，期待報酬は現実よりも大幅に小さくなっている．

種類	内容	件数
タスクデータ	2022 年 12 月に発行されたタスクの ID，タスク実施に必須と書かれているスキル，タスク実施の報酬	22853 件のうち 3157 件
ワーカデータ	2022 年 12 月のアクティブワーカの ID，ワーカが登録しているスキル	49145 件
応募データ	2022 年 12 月の応募データ．タスク ID とワーカ ID の組と，実際に実施したワーカのフラグ	270688 件

ワーカがあるタスクを実施した実績があれば，そのタスクの必須スキルをワーカの登録スキルに追加した．

スキル表記の統一 データ中の必須スキルや登録スキルは自由入力のため，「記事執筆」と「記事作成」，「k8s」と「Kubernetes」のように同一の意味を持つスキルや略語を同定できるように前処理を行う．その際，前処理を容易にするため，各スキルは単純な文字列ではなく，スキルトークンの集合としてモデル化する．具体的には”Microsoft Excel”というスキルは {Microsoft, Excel}，”ソフトウェア開発”というスキルは { ソフトウェア, 開発 } と表現する．

スキル s に対する前処理の手順は次の通りである．

（1） s に対して Unicode 正規化を行う

（2） s のアルファベットを大文字にする

（3） s を形態素解析器 Sudachi [13] によりトークンに分割し，スキルトークンの集合 *skill_tokens* を生成する

（4） 語彙統制辞書により *skill_token* のトークンの語彙統制を行う．語彙統制辞書には，スキルトークンレベルの統制 (エクセル →Excel 等 42 件) と，スキルレベル (集合レベル) の統制 ({EXCEL} → {MICROSOFT, EXCEL} 等 13 件) がある．

以上の前処理により，元データにおける distinct なスキル数

表 4: 利用データにおけるタスクの必須スキル集合と，ワーカの登録スキル集合に含まれる異なるスキルの総数と，前処理後の各種統計量

前処理前スキル	5606 件
前処理後スキル	577 件
前処理後スキル集合	611 件
前処理後スキルターム	419 件

である 5606 件がスキル数 577 件に削減された (そこに含まれる異なるスキルターム数は 419 件となった) (表 4)．結果として，タスクの必須スキル集合と，ワーカの登録スキル集合に現れる異なるスキル集合の数は 611 件となった．

5.2 指標の計算

スキルをスキルタームの集合で表現することにした結果，ワーカの登録スキル集合や，タスクの必須スキルは，スキルタームの集合の集合で表現されることになる．したがって 3 節で定義した期待報酬などの計算で，ワーカ w がタスク t を実施するスキルがあるかどうかの判断として，w の登録スキル集合が t の必須スキル集合のスーパセットになっているか否かの判断をする必要がある．本実験では，異なる二つのスキル (スキルタームの集合)S_1 が S_2 の部分集合であるという関係を次のように定義する．

- For a pair of skill sets S_1, S_2, $S_1 \subseteq S_2$ holds if for each $s_1 \in S_1$, there exists $s_2 \in S_2$ s.t. $s_1 \subseteq s_2$.

たとえば，以下のスキル集合 $S1$ はスキル集合 $S2$ の部分集合である．すなわち，S_2 を登録スキル集合としてもつワーカは，S_1 が必須スキル集合であるようなタスクを実施することが可能である．

$$S_1 = \{\{EXCEL, VBA\}, \{PYTHON\}\},$$
$$S_2 = \{\{EXCEL, VBA\}, \{PYTHON, FLASK\}\}$$

5.3 タスク市場のスキルに関する分析

タスクデータにおける必須スキル数の分布と，スキル数と報酬の関係. 図 4 は，タスクデータ中の全てのタスクを点として，そのタスク実施の必須スキルに含まれるスキル数と，そのタスク実施で含まれる報酬の関係を示したものである．スキルが登録されているタスクの中で 1 件しか登録されていない場合が半分以上を占めた．ここからわかるように，必須スキル数が多数書かれているタスクは，報酬が高額でないものが多い．この理由は，必須スキルの数はタスクの難易度を表すものではないことを示している．クラウドワークスには必須スキルでタスクを検索する機能があるため，むしろインデックスとしての利用を意図して多数記述していると推察される．例えば，最も大きな必須スキル数 17 のタスクには「パソコン, 操作」「パソコン, 入力」等といったスキルが列挙されていた．

ワーカの持つスキル数と期待報酬の関係. 図 5 は，ワーカデータ中の全てのワーカを点として，そのワーカが登録しているスキル集合に含まれるスキル数と，スキル集合から計算されるワーカの期待報酬をプロットしたものである．スキル数が 50 以下のワーカを対象とし，実際のタスクの必須スキルに存在しないスキルは除外している．赤点はそのスキル数を持つワーカの平均期待報酬である．この図から，むやみに登録スキルを増えても必ずしも期待報酬は高くならないことがわかる．25 を超えたあたりで平均期待報酬が上がるが，これはワーカの数が大きくへっており，値が高いワーカの平均期待報酬の影響が大きくなっている結果である．

タスクが持つスキル集合ごとの期待報酬と平均報酬の関係. 図 6 は，タスクの必須スキル集合として現れる異なるスキル集合を各点で表し，そのスキル集合を必須とするタスクの平均報酬と平均期待報報酬の関係をプロットしたものである．この図からわかるように，各タスクの報酬を見ても必ずしも期待報酬がわからないため，期待報酬をより高くするスキルの推薦の有効性が示唆される結果が得られた．5.1 節でも説明したとおり，本研究ではタスクの募集・応募データについて必須スキルが登録されている 3157 件のみを対象に実験を行っており，これは総データ数の約 14 %である．それに対しワーカデータに関しては 5 万人弱のデータをすべて使用しており，これは「そのスキル集合を使用する案件の合計報酬/そのスキル集合を保有するワーカ数」で算出される期待報酬が著しく低くなっている．

5.4 推薦アルゴリズムの実行結果

推薦アルゴリズムの比較においては，ワーカデータからランダムに 50 人の選択を行い，それぞれに対して，3 つの推薦アルゴリズム (All, Greedy, Greedy-Ratio) を適用して，n 個 ($1 \leqq n \leqq 5$) のスキル推薦を行い，実装し，実行時間とスキル追加によって得られるワーカの期待報酬について比較した．表 5 は実行環境である．

表 5: 推薦アルゴリズムの実行環境

CPU	Core i7 14 世代
メモリ	32GB
OS	Ubuntu20.04.6(Windows Subsystem for Linux)
実装言語	Python3.12

推薦されたスキルの追加によって得られる期待報酬. 図 8 の実線は，50 人の各ワーカに対して，3 つの推薦手法により推薦されたスキルを追加されることによって得られる平均期待報酬を描いた物である．横軸の推薦スキル数が 0 の箇所は，推薦スキルを追加する前のワーカの平均期待報酬を指す．

期待報酬に関しては，推薦スキル数が 3 までは 3 つの手法がほぼ同じ平均期待報酬をもたらしたが，それ以降は All が最も期待報酬が高い結果となった．推薦前の平均期待値報酬は 1396 円であったのが，推薦された 2 つのスキルを追加するといずれの手法も約 27000 円に増加した．これは，今回の「現在もっているスキルに関連するスキルのみを探索空間とする」という手法であっても，期待報酬の大幅な増大を可能にすることを示唆している．

図 7 に n=2 で Greedy によるスキル推薦を行った後のワーカの期待報酬の上昇率の分布を示す．最小の上昇率は 14.2%であり，50 名中 42 名のワーカは 100%の上昇率 (2 倍の期待報酬) 以上の結果が得られた．また，表 6 に 5 人のワーカの推薦前のスキルと推薦したスキルと期待報酬の変化を示す．いずれも大幅な期待報酬の増加を示している．case id 1 のケースでは，Microsoft Word だけではなく，文書作成や校正をいったスキルがあれば期待報酬が大幅に増加することを示している．2 のケースでは，画像を編集するだけではなく，イラストを描くというスキルを推薦している．3 のケースは，動画編集だけではなく，CG やモデリングのスキルの追加を推薦している．4 のケースは，Microsoft Excel やライティングのスキルを持つワーカに，script のスキルを追加することを推薦している．5 のケースは最も期待報酬が上昇したケースであるが，元のスキルと比較的関連が薄いと思われる会計・簿記のスキルを推薦している．これは，このような組み合わせの必須スキルをもつタスクが存在した事によって探索空間に現れたからである．したがって，必ずしも関連が深いスキルが推薦される場合ばかりではないが，多くの場合は元のスキルとある程度関連のあるスキルが推薦されていることがわかる．

スキル開発グラフの形状. 上記の結果のように，スキル開発グラフが表す探索空間の範囲で推薦スキルを追加することによって期待報酬が全体的にあがる理由を調べるために，作成したスキル開発グラフの形状について調査を行った．表 7 にスキル開発グラフの統計量を示し，図 9 に out degree の分布を示す．これらが示すように，スキル開発グラフは密なグラフになっており，ワーカの登録スキルに関連するスキルが数多く存在する可能性が高い．したがって，登録スキルに関連するスキルだけを探索空間に含むだけでも期待報酬を向上できる事が出来ていることがわかった．

推薦アルゴリズムの実行時間の比較. 図 8 の点線は，3 つの推薦アルゴリズムの実行時間を示したものである．実行時間については All の時間が n が増加するごとに指数関数的に長くなり，貪欲法の 2 つについては概ね同じように増加を抑えること

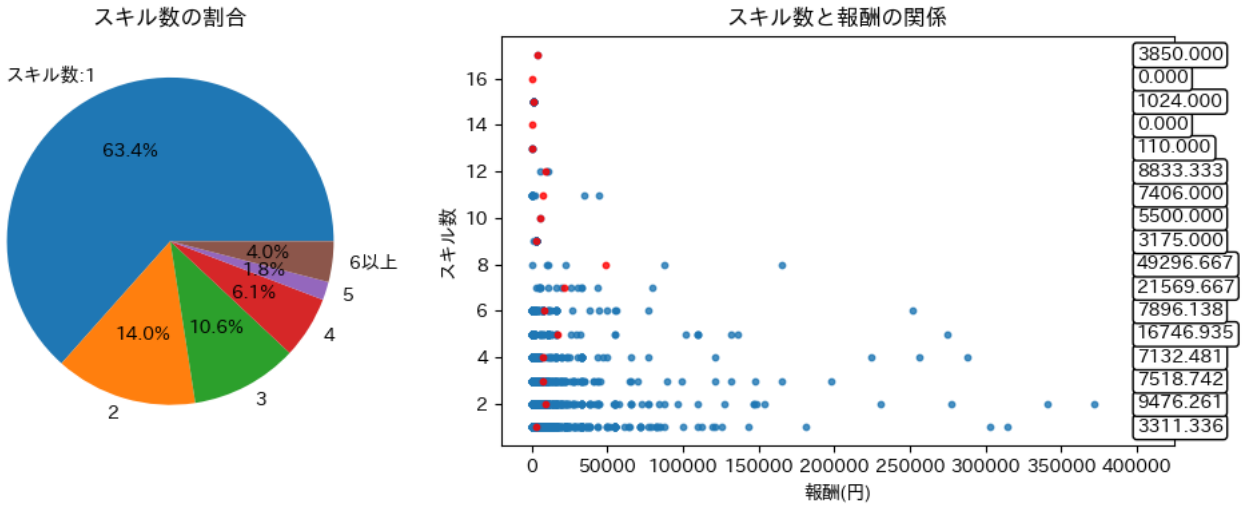


図 4: タスクデータにおける必須スキル数の分布と，スキル数報酬の関係 (右の値と赤点はそのスキル数における報酬の平均値)

表 6: n=2 の greedy に基づく推薦スキルの例

case id	元の登録スキル	推薦スキル	追加前後の期待報酬変化
1	{{'WORD', 'MICROSOFT'}}	{{' 文章', ' 作成', ' 編集', ' 校正'}, {'EXCEL', 'MICROSOFT'}}	32.61 → 8514.62
2	{{'ILLUSTRATOR', ' イラストレーター', 'ADOBE'}, {' 画像', ' 編集', ' 加工'}, {'PHOTOSHOP', 'ADOBE'}}	{{' イラスト', ' デジタル'}, {'CLIP', 'STUDIO', 'PAINT'}}	293.67 → 56882.72
3	{{'YOUTUBE'},{' 動画', ' 編集', ' 加工'}, {' 動画', ' 制作'},{' データ', ' 入力'}}	{{'CG', '3D'}, {' モデリング', '3D'}}	503.46 → 121564.72
4	{{'EXCEL', 'MICROSOFT'},{' データ', ' 入力'}, {' ライティング', ' 記事'},{' 作成', ' 原稿'}, {' 文章', ' 作成', ' 原稿'}}	{{'GOOGLE', 'SPREADSHEET'}, {'GOOGLE', 'SCRIPT', 'APPS'}}	287.21 → 37589.35
5	{{'YOUTUBE'},{'MICROSOFT', 'EXCEL'}, {'ADOBE', 'ILLUSTRATOR', ' イラストレーター'},{'ADOBE', 'PHOTOSHOP'}, {' 加工', ' 動画', ' 編集'}, {'ADOBE', 'PREMIERE', 'PRO'},{'WORD', 'MICROSOFT'}}	{{'SHAREPOINT'}, {' 簿記', ' 会計'}}	22.85 → 35211.26

表 7: skill-dev graph の統計量

nodes	611
edges	20339
avg out degrees	33.26

が出来ている．All は追加スキル数 n 以下の到達可能な全ての経路を探索するため確実に最も期待報酬が高くなるような組み合わせを見つけることができる．しかし図 8 の通り，貪欲法と比較してかなり長い探索時間を要することが分かる．n が増えるごとに時間も大幅に増加しており，n の値をさらに増やすと時間の増加はさらに顕著なものになると考えられる．

5.5 結果に関する適用範囲の制限

本研究では，タスクの件数が限定されており実際よりも期待報酬が少ないという点以外にも，結果を解釈する際に考慮すべき点がいくつか存在する．まず，「ワーカがそのスキルを持っていればそのスキルを必要とする仕事ができる」という仮定の下で期待報酬の計算を行ったが，一概に「スキルを持っている」といってもワーカごとにその熟練度は異なり，案件によっても必要となる熟練度は違う．実際に Sun らの研究 [9] ではスキル価値をそのスキルのみで可能な仕事の報酬と定義したときに，スキルの熟練度がスキル価値に影響を及ぼすことを示している．一方，本論文の期待報酬の計算は，熟練度を考慮していないため，スキルの熟練度が高いワーカにとっては実際より低い期待報酬となり，逆に低いワーカにとっては実際より高い期待報酬となる可能性がある．

加えて，Ha-Thuc ら [4] が取り組んでいる通り，ワーカは持っているスキルを全て登録しているとは限らず，登録されていないが持っていると考えられるスキルを推定する必要がある．今回は，5.1 節で説明したように，期待報酬を計算する際に「そのスキル集合を登録しているか，そのスキル集合を使った仕事をしたことがあるワーカの数」を使用しているが，実際には更にワーカの数が増える可能性がある．

6 ま と め

本論文では，クラウドワーカがどのようなスキルを身につけるべきかという問題に注目し，現在のワーカの登録スキルと，タスク市場のデータを元に期待報酬を増加させるような関連ス

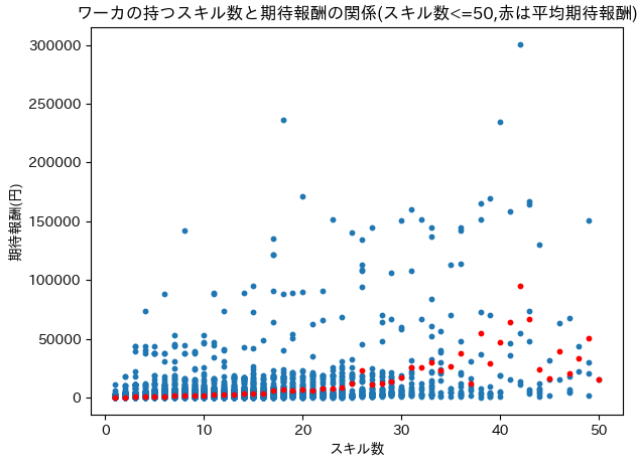


図 5: ワーカの持つスキル数と期待報酬の関係．各点はワーカに対応する．赤点はそのスキル数を持つワーカの平均期待報酬である．スキル数が増えても必ずしも期待報酬は高くならないことがわかる．

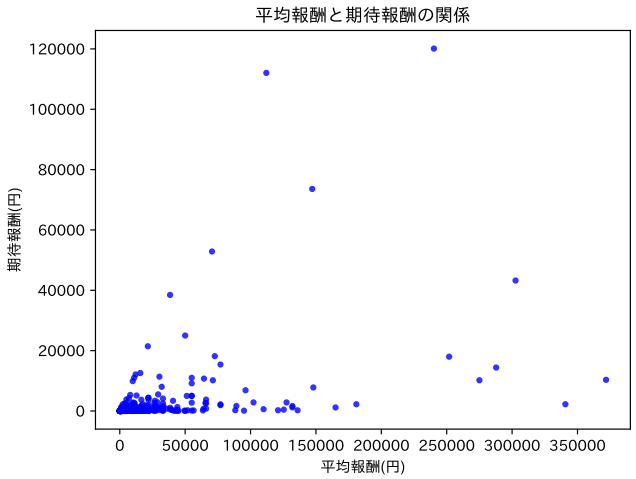


図 6: タスクが持つスキル集合の期待報酬と平均報酬の関係．各点は，タスクデータに表れる異なる必須スキル集合のそれぞれに対応する．平均報酬はその必須スキル集合をもつタスクの報酬の平均であり，期待報酬はそれらのタスクに対する平均期待報酬の値である．

キルを推薦する手法を構築し，評価した．具体的には，クラウドソーシングタスク市場のデータ分析と推薦アルゴリズムの評価により，次の結果が得られた (1) タスク市場でワーカが登録しているスキルは必ずしも期待報酬を上げるようなスキルになっておらず，スキル推薦が有効である可能性があること，(2) 現在の登録スキルに関連するスキルを推薦することにより，期待報酬が上がること (3) 3 つまでのスキルの推薦については，貪欲型の探索アルゴリズムでも全探索アルゴリズムと遜色なく，高速に計算できること．今後の課題としては，スキルが書かれていないタスクやワーカのスキル推定，熟練度推定，推薦アルゴリズムにおけるこれらの考慮，等が挙げられる．

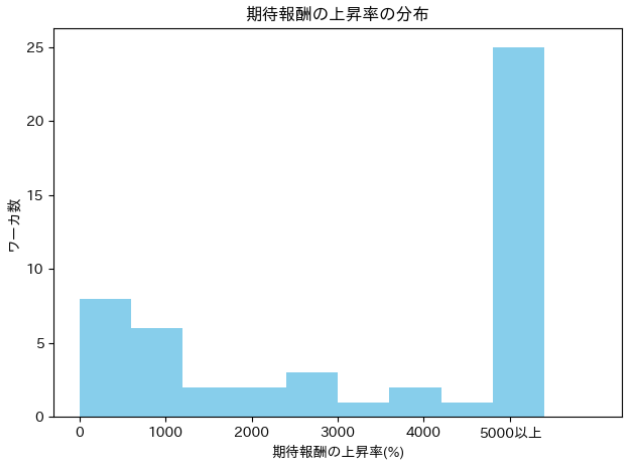


図 7: n=2 の greedy に基づく推薦後の期待報酬の上昇率の分布．各範囲の幅は 500 である．

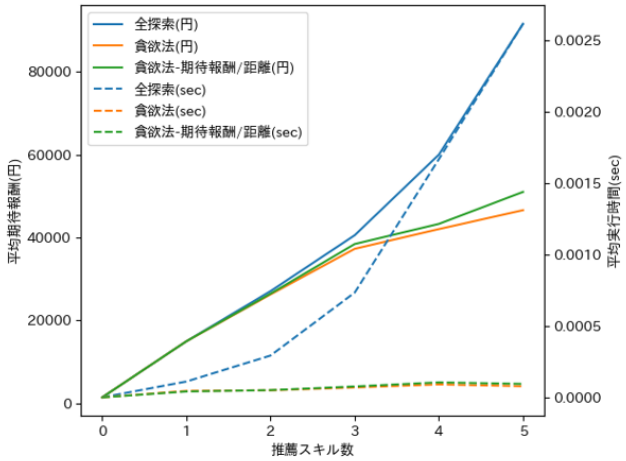


図 8: 推薦スキル数ごとの期待報酬の変化

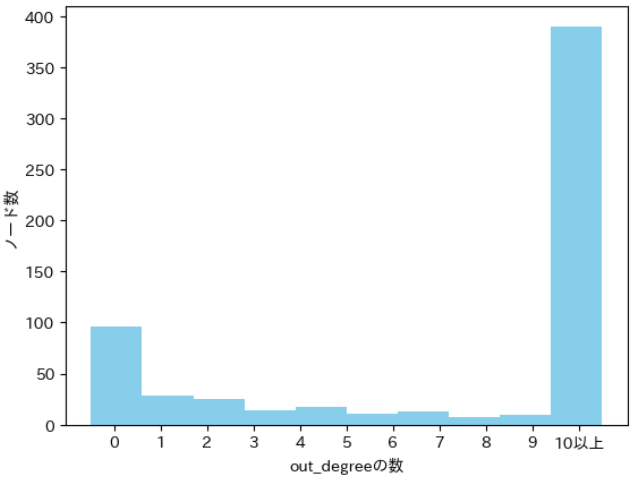


図 9: skill-dev graph の out degree の分布

謝　　辞

本研究の一部は JSPS 科研費 (22H00508) の支援をうけたもの

である．本研究で使用したデータは株式会社クラウドワークスから提供をいただいたものである．ここに謝意を示す．

文　献

[1] Dorothy Lee Blyth, Mohammad Hossein Jarrahi, Christoph Lutz, and Gemma Newlands. Self-branding strategies of online freelancers on upwork. *new media & society*, p. 14614448221108960, 2022.

[2] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. iCrowd: An Adaptive Crowdsourcing Framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 1015–1030. Association for Computing Machinery, May 2015.

[3] David Gross-Amblard, Atsuyuki Morishima, Saravanan Thirumuruganathan, Marion Tommasi, and Koichiro Yoshida. Platform design for crowdsourcing and future of work. *IEEE Data Eng. Bull.*, Vol. 42, pp. 35–45, 2019.

[4] Viet Ha-Thuc, Ganesh Venkataraman, Mario Rodriguez, Shakti Sinha, Senthil Sundaram, and Lin Guo. Personalized expertise search at LinkedIn. In *2015 IEEE International Conference on Big Data (Big Data)*, pp. 1238–1247, Santa Clara, CA, USA, October 2015. IEEE.

[5] Danula Hettiachchi, Vassilis Kostakos, and Jorge Goncalves. A Survey on Task Assignment in Crowdsourcing. *ACM Computing Surveys*, Vol. 55, No. 3, pp. 1–35, March 2023.

[6] Chien-Ju Ho and Jennifer Vaughan. Online Task Assignment in Crowdsourcing Markets. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26, No. 1, pp. 45–51, 2012.

[7] Mohammad Hossein Jarrahi, Will Sutherland, Sarah Beth Nelson, and Steve Sawyer. Platformic Management, Boundary Resources for Gig Work, and Worker Autonomy. *Computer Supported Cooperative Work (CSCW)*, Vol. 29, No. 1, pp. 153–189, April 2020.

[8] Pyeonghwa Kim and Steve Sawyer. Many Futures of Work and Skill: Heterogeneity in Skill Building Experiences on Digital Labor Platforms. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pp. 1–9, Oldenburg Germany, June 2023. ACM.

[9] Ying Sun, Fuzhen Zhuang, Hengshu Zhu, Qi Zhang, Qing He, and Hui Xiong. Market-oriented job skill valuation with cooperative composition neural network. *Nature Communications*, Vol. 12, No. 1, p. 1992, March 2021.

[10] Ziyang Wang, Wei Wei, Chenwei Xu, Jun Xu, and Xian-Ling Mao. Person-job fit estimation from candidate profile and related recruitment history with co-attention neural networks. *Neurocomputing*, Vol. 501, pp. 14–24, August 2022.

[11] Xiao Yan, Ada Ma, Jaewon Yang, Lin Zhu, How Jing, Jacob Bollinger, and Qi He. Contextual Skill Proficiency via Multi-task Learning at LinkedIn. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 4273–4282, Virtual Event Queensland Australia, October 2021. ACM.

[12] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1031–1046, Melbourne Victoria Australia, May 2015. ACM.

[13] ワークスアプリケーションズ徳島人工知能 NLP 研究所. Worksapplications/sudachi: A japanese tokenizer for business. https://github.com/WorksApplications/Sudachi, 2023. (2024/1/5 閲覧).

[14] 岩本隆. ポスト・コロナ社会におけるニューノーマルな働き方〜「worktech ×フリーランス」がもたらす潜在的労働力活用と経済成長〜. https://www.j-paa.or.jp/wp/wp-content/uploads/2020/09/%E3%83%9D%E3%82%B9%E3%83%88%E3%83%BB%E3%82%B3%E3%83%AD%E3%83%8A%E7%A4%BE%E4%BC%9A%E3%81%AB%E3%81%8A%E3%81%91%E3%82%8B%E3%83%8B%E3%83%A5%E3%83%BC%E3%83%8E%E3%83%BC%E3%83%9E%E3%83%AB%E3%81%AA%E5%83%8D%E3%81%8D%E6%96%B9%EF%BD%9E%E3%80%8CWorkTech%C3%97%E3%83%95%E3%83%AA%E3%83%BC%E3%83%A9%E3%83%B3%E3%82%B9%E3%80%8D%E3%81%8C%E3%82%82%E3%81%9F%E3%82%89%E3%81%99%E6%BD%9C%E5%9C%A8%E7%9A%84%E5%8A%B4%E5%83%8D%E5%8A%9B%E6%B4%BB%E7%94%A8%E3%81%A8%E7%B5%8C%E6%B8%88%E6%88%90%E9%95%B7%EF%BD%9E.pdf, 9 2020. (2024/02/15 閲覧).

映像集中を目的とした第二言語字幕における音声変化を考慮した字幕提示手法

西　優己† 中村　優吾†† 福嶋　政期†† 荒川　豊††

† 九州大学大学院システム情報科学府　〒 819–0395 福岡県福岡市西区元岡 744
†† 九州大学大学院システム情報科学研究院　〒 819–0395 福岡県福岡市西区元岡 744
E-mail: †nishi.yuki@arakawa-lab.com, ††{y-nakamura,shogo,arakawa}@ait.kyushu-u.ac.jp

あらまし　映画やテレビ番組などを見る際，字幕は映像の内容理解を促進する役割を担っている．一般的な字幕は画面下部に提示されるが，字幕と映像の視線の切り替えが多く発生し，認知負荷の要因になっている．著者らは，第二言語の映像視聴時に映像集中を促す字幕提示として，一般的な字幕に比べて情報量を削減し，提示時間を短くした字幕を画面中央に提示する手法を提案している．本研究では著者らの従来手法と比べて英語の音声変化を考慮し，提示する字幕の情報量を増加した手法を提案する．評価実験では，21 歳から 25 歳の 18 名の参加者に英語の映像を視聴してもらい，提案手法が映像視聴時の視線分布，理解度，認知負荷に与える影響を調査した．その結果，一般的な字幕に比べて提案手法による字幕では画面中央付近に視線が分布し，映像の理解度が同程度であることが明らかとなった．映像視聴時の認知負荷は一般的な字幕よりも大きかったが，有意差は確認されなかった．

キーワード　字幕，第二言語，認知負荷，視線，ユーザインタフェース

1 はじめに

第二言語の映画やテレビ番組を理解する上で，母国語の字幕は重要な役割を担っている．外国語の映像を視聴する際，母国語の字幕が内容理解に有益であることが報告されている [1] が，字幕と映像の視線の切り替えが眼精疲労の原因となったり，字幕を注視することにより映像内の重要なシーンを見逃したりする可能性がある [2] などのデメリットも存在する．そのため，ユーザが字幕付きの映像を視聴する際の認知負荷を適度に減らしつつ，映像に集中できる状態を作ることが望ましい．さらに，音声の内容が分かっている場合でも，テキストを読む傾向があることが分かっており [3]，視聴する人にとって不要な情報を減らすことが望ましいと考える．これに対し，著者らは，第二言語の映像視聴時の映像集中を妨げないための字幕として，字幕の提示位置，提示時間，情報量の 3 つに着目した手法を提案している [4]．具体的には，一般的な字幕に比べて情報量を削減し，提示時間を短くした字幕を画面中央に提示した手法である．著者らが過去に行った実験では，一般的な字幕と比較して，提案手法の字幕では映像の理解度を同程度に保ちつつ，映像集中を妨げない可能性が示唆された．しかしながら，提案手法において，参加者は認知負荷が高い状態で映像を視聴していることが明らかとなり，提案手法よりも一般的な字幕を好む参加者が多かった．したがって，本研究では，提案する字幕提示手法において，映像の理解度を維持しつつ，認知負荷をより下げることを目的とする．過去の実験で参加者から，字幕の情報量が想定より少なかったという意見が多く得られた．また，理解度テストの分析の結果から，映像視聴時に，特に英語の音声変化が生じている箇所の理解が困難である可能性が示唆された．そのため，認知負荷を下げるためには，従来の提案手法に加え，音声変化を考慮し情報量を増やす必要があると考えた．また，本研究では，英語の映像を日本語字幕で視聴することを想定している．実験では，18 名の参加者に 3 種類の字幕で映像を視聴してもらい，提案手法が映像視聴時の視線分布，理解度，認知負荷に与える影響を調査した．1 つ目は，画面下部に文章単位で提示される一般的な字幕である．2 つ目は，画面中央に単語単位で提示される字幕であり，提示する字幕の選定に CEFR と音声変化のルールを適用したもの (提案手法) である．3 つ目は画面中央に単語単位で提示される字幕であり，字幕の選定に CEFR ルールのみを適用したもの (著者らの従来手法) である．視線データはスクリーンベースのアイトラッカーである Tobii Pro Nano[1]で計測し，視線分布を表すヒートマップを作成した．映像の理解度は，映像視聴後に実施する理解度テストの点数で評価し，認知負荷は NASA-TLX(NASA Task Load Index) [5] による参加者の主観的な評価とした．実験の結果，その結果，一般的な字幕に比べて提案手法による字幕では画面中央付近に視線が分布し，映像の理解度が同程度であることが明らかとなった．映像視聴時の認知負荷は一般的な字幕よりも大きかったが，有意差は確認されなかった．

本稿の構成は以下の通りである．2 章では字幕提示手法に関する関連研究を示し，3 章で提案手法について示す．4 章で実験について説明し，5 章で実験結果を示す．最後に 6 章でまとめと今後の展望について述べる．

2 関連研究

字幕提示手法に関する研究は様々なものがある．Hu ら [6] は，映像内の話者付近に字幕を表示させる話者追従型字幕 (Speaker-

1：https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-nano

following Subtitles) を提案した．Kurzhals ら [7] の研究において，話者追従型字幕では，画面下部に提示された一般的な字幕に比べて映像視聴時の視線が画面中央付近に集中し，字幕よりも話者に注意が向けられたことが報告された．しかし，字幕の位置が動的に変化することから，一般的な字幕よりもテキストを探すのに苦労したことが報告されている．提示する文量や時間に着目した Rapid Serial Visual Presentation(RSVP) [8] に関する研究も行われている．RSVP は高速逐次視覚提示とも呼ばれ，単語などを一定の速度で同じ位置に連続的に提示する手法である．Rzayev ら [9] は，スマートグラスを用いた読書におけるテキスト位置と提示方法について研究している．座った状態で読書をした場合に，文章単位で提示される方法と比べて RSVP 条件でのテキスト理解度が高かったことが報告された．Erin ら [10] は，スマートウォッチを用いた読書における RSVP の影響について調査している．RSVP 条件での読書は，一般的な方法と比較して同程度の理解度を保ちつつ，短い時間で読書することができたが，満足度に関するアンケートでは RSVP 条件よりも従来手法での読書が好まれた．さらに，映像に関する研究において，音声の内容が分かっている場合でも，テキストを読む傾向があることが分かっており [3]，視聴する人にとって不要な情報はなるべく減らすことが望ましいと考える．

3 提 案 手 法

本章では，著者らが提案した従来の手法 [4] について述べ，その後に今回提案する字幕提示手法について述べる．著者らは，映像集中を目的とした字幕提示手法として，字幕の提示位置，提示時間，情報量の 3 つに着目しており，一般的な字幕に比べて情報量を削減し，提示時間を短くした字幕を画面中央に提示する手法を提案している．

3.1 字幕の提示位置

字幕の提示位置に関して，一般的な字幕は画面下部に提示されるが，提案手法では，映像集中を促すために，字幕の提示位置を画面中央に固定した．

3.2 字幕の提示時間

一般的な字幕は 1〜3 秒程度提示されるが，字幕の提示時間は，RSVP に基づいて各語 300ms 提示する．RSVP を用いた読書では，200WPM(Words Per Minute) で内容をよく理解できることが示唆されており [11]，これを一単語あたりに換算すると 300ms であることから，この値を用いた．

3.3 字幕の情報量

字幕の情報量は，文章単位でなく単語単位で提示し，かつ提示する語を語彙レベルに応じて選定した．提示する語彙の選定には New Word Level Checker[2] というツールを用いた．New Word Level Checker は，日本人英語学習者に向けて提供されているツールの一つであり，入力された英文に含まれる語彙のレベルを分類するものである．本ツールの用途として，教師が英語教材の難易度を評価したり，学生が書いた英文の語彙カバー率を調査したりすることなどが挙げられている [12]．本ツールに登載されている語彙リストとして New JACET8000, JET2000，SVL12000 などがある．今回は CEFR（Common European Framework of Reference for Languages）[13] の語彙リストを使用した．CEFR は，外国語の運用能力を測る国際基準であり，6 段階 (A1，A2，B1，B2，C1，C2) に分類されている．A1 が最も基礎的なレベルとされ，C2 はネイティブレベルに相当するとされている．実験用の映像に含まれる単語を New Word Level Checker で分類したところ，約 6〜7 割が A1 に分類され，C1，C2 に該当する単語は含まれていなかった．今回は，最も基礎的な A1 レベルの単語を除外し，A2 レベル以上の単語 (A2，B1，B2) を提示対象とし，語彙の選定を行った．A2 レベルでは，簡単な文章を理解したり，日常的な範囲での会話ができるが，外国語のメディア (映画やラジオなど) を視聴するのは難しいとされている．映像の内容理解に必要であると想定される語彙のみ残すことで，情報量を減らしつつ，従来の字幕と同程度の理解度を保つことができると考えた．このように，理解が容易だと想定される語彙の情報量を削減し，提示時間を短くした字幕を中央に提示することで，字幕の役割である内容理解の促進に加え，映像の視聴をより促すことができると考えた．図 1 に，一般的な字幕と提案手法による字幕レイアウトの比較を示す．著作権の関係で作品そのものの画像は記載しておらず，生成した類似の画像を記載している．図 1a の一般的な字幕では，字幕が文章単位で画面下部に提示されており，図 1b では，字幕が単語単位で画面中央に提示されていることが分かる．

3.4 本研究で検証する手法

上述した字幕提示手法を用いて，著者らが過去に行った実験では，一般的な字幕と比較して理解度は同程度であったが，認知負荷が高いことが明らかとなった．また，実験参加者から得られた意見として,「字幕の情報量が少ないと感じた」「字幕がいつ出てくるかわからないのでとても頭を使った」などがあり，認知負荷を下げるためには，情報量を増やし，提示時間を長くすることが必要であると考えた．さらに理解度テストの誤答の要因として，英語特有の音声変化が生じることが大きく影響している可能性が示唆された．そこで，英語特有の音声変化に着目し，情報量を増やしつつ，提示時間を長くした手法について提案する．英語の音声変化に関して，Carler ら [14] は，音声変化 (Connected speech form) を，“A pronunciation of a word influenced by the sounds in surrounding morphemes, syllables, or words” と 2019 年に定義しており，北米発音 (General American) で見られる変化の例として脱落 (Elision) と同化 (Assimilation) の 2 つを挙げている．また，同著者らの 2021 年の文献 [15] では，連結 (Liaison: when a sound that doesn’t occur in the citation form of a word is inserted in connected speech) が加えられている．これらの定義については，書籍間等で大きな違いはないが，現象をどのように分類し，どのよう

2：https://nwlc.pythonanywhere.com/

(a) 一般的な字幕

(b) 提案手法による字幕

図 1: 字幕レイアウトの比較

な用語を使うかについては完全に一致していない．本研究では，連結 (Liaison)，同化 (Assimilation)，脱落 (Elision) の 3 つの音声変化を扱うこととした．本研究で扱う音声変化ルールの説明を表 1 に示す．

表 1: 本実験で扱う音声変化

現象	説明	例
連結	単語の最後の子音と次の語の母音が繋がる	take on, afraid of
同化	連結が起こりその音が別の音に変化する	get you, right away
脱落	同じ子音が重なりあるべき音が発音されない	some more, get tired

以上をまとめると，今回検証する字幕提示手法は，CEFR による語彙レベルに加え，英語の音声変化が生じている箇所を選定したものである．以降，これを CEFR+音声変化条件と略すことがある．また，これと区別するために著者らの従来手法を CEFR 条件と略す．上述した各字幕提示手法の比較を表 2 に示す．これらのルールを基に，実験で用いる映像の字幕を作成した．また，表 3 に今回作成した字幕の情報量の比較を，表 4 に一度に提示される字幕の比較を示す．表 3 は，通常字幕の情報量を 100%としたときの各字幕の情報量の割合を示している．CEFR のみを適用した字幕 (CEFR 条件) の情報量は，通常字幕の 2 割から 3 割程度であり，これに音声変化を適用した字幕 (CEFR+音声変化条件) は通常字幕の 6 割程度の字幕が提示される．CEFR 条件に比べ，情報量が増えており，映像を視聴する人が期待する量の字幕が提示できると考える．表 4 は，各映像において，一度に提示される字幕の字数を比較したものであり，字幕の合計字数を提示回数で割ることにより算出している．CEFR+音声変化条件では，通常字幕の半分以下である，約 3〜4 文字の字幕が一度に提示される．

表 2: 字幕提示手法の比較

手法	提示位置	提示時間	情報量
一般的な字幕	画面下部	1〜3s	文章単位で提示
従来手法 [4]	画面中央	300ms	CEFR による選定で単語を提示
提案手法	画面中央	500ms	CEFR と音声変化による選定で単語を提示

表 3: 作成した字幕の情報量の比較 (%)

映像	通常字幕	CEFR+音声変化	CEFR
A	100	55.7	34.2
B	100	59.3	18.4
C	100	61.3	27.3

表 4: 一度に提示される字幕 (字)

映像	通常字幕	CEFR+音声変化	CEFR
A	7.6	3.5	3.4
B	7.5	3.4	2.4
C	7.7	3.4	2.5

4 実　　験

提案する字幕提示手法が映像視聴時の視線分布，理解度，認知負荷に与える影響を調査するため，評価実験を行った．

4.1 仮　　説

字幕の提示位置，提示時間，情報量に着目した提案手法が，映像視聴時の視線分布，理解度，認知負荷に影響を与えるという仮説を立てた．

- **仮説 1: 提案手法による字幕では，一般的な字幕に比べて映像に集中できる．**

提案手法では字幕を画面中央に固定しており，画面下部に提示される字幕に比べて話者との距離が近く，映像に集中できると考える．

- **仮説 2: 情報量を増やした字幕 (CEFR+音声変化条件) では，CEFR 条件の字幕に比べて高い理解度を保つことができる．**

提示する字幕を語彙レベルと音声変化に応じて選定し，情報量を通常字幕の 6 割程度に増やすことで，CEFR 条件の字幕よりも高い理解度を保つことができると考えられる．

- **仮説 3: 情報量を増やした字幕 (CEFR+音声変化条件) では，CEFR 条件の字幕に比べて映像視聴時の認知負荷を減らすことができる．**

音声変化を考慮し，字幕の情報量を増やすことで，音声変化を

考慮していない字幕に比べて認知負荷の低い状態で映像を視聴することができると考える．

4.2 実験内容

実験は 21 歳から 25 歳の 18 名 (男性 16 名，女性 2 名) の参加者に協力していただいた．参加者は英語の映像を視聴し，その後映像の内容に関する設問に回答した．実験条件は，画面下部に提示される一般的な字幕 (通常字幕)，画面中央に提示される字幕 (CEFR+音声変化)，画面中央に提示される字幕 (CEFR) の 3 つに設定した．実験の様子を図 2 に示す．主観的な認知負荷の評価には NASA-TLX(NASA Task Load Index) を用いており，参加者は，英語の映像を各条件で視聴し，理解度テストに回答するというタスクについて NASA-TLX の各項目を評価した．また，ユーザビリティの評価として System Usability Scale(SUS) [16] を用い，各字幕提示手法の使いやすさなどを評価してもらった．実験にはアメリカのコメディドラマ作品であるフレンズの映像を使用し，2 人または複数人の人物が会話しているシーンを切り取ったものを使用した (表 5)．映像は，時間が 3 分から 3 分半程度の長さで WPM(Words Per Minute) が比較的同じものを選定している．実験手順について示す．始めに，参加者は椅子に座り，視聴する映像と提示される字幕の説明を受けた．次に，アイトラッカーのキャリブレーションを行い，視線データが計測できていることを確認した．その後，イヤホンを装着した状態で映像を視聴し，映像に関する理解度テスト (4 択 11 問) と NASA-TLX(7 段階リッカート尺度)，SUS(5 段階リッカート尺度) に回答した．理解度テストは，登場人物がなんと言ったか，どんな行動をしたかなど，映像内の出来事に関する質問を含んでいる．映像の視聴と回答を 1 セットとしこれを 3 映像分行った．各参加者が視聴する映像と提示する字幕の組み合わせはランダムにしている．参加者は最後にアンケートに回答した．実験にかかった時間は 1 人あたり 30 分程度であった．実験条件を以下に示す．

- 通常の日本語字幕：一般的な文章の日本語字幕が画面下部に提示される．字幕の提示時間は文章の長さによって異なり，1〜3 秒程度提示される．
- CEFR+音声変化条件：CEFR と音声変化ルールで選定された日本語字幕が画面中央に提示される．各単語の提示時間は 500ms である．
- CEFR 条件：CEFR ルールで選定された日本語字幕が画面中央に提示される．各単語の提示時間は 300ms である．

表 5: 実験で使用した映像

映像	タイトル	時間 (分: 秒)
A	The Friends Pretend To Like Rachel's English Trifle	3:38
B	Ross Had His Teeth Whitened	3:00
C	Rachel Hires Tag As Her Assistant	3:35

5 実験結果

評価実験の結果を視線分布，映像の理解度，認知負荷ごとに示す．本実験において視線分布は被験者間比較とし，映像の理解度，認知負荷，ユーザビリティは被験者内比較としている．

図 2: 実験の様子

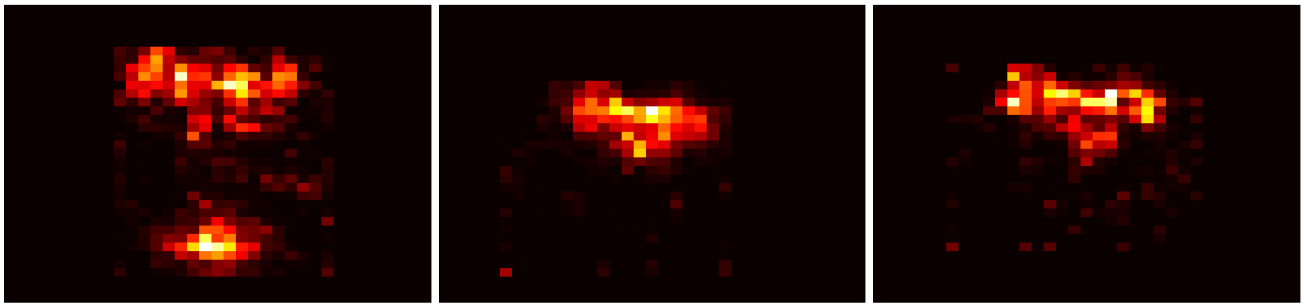
5.1 視線分布

Tobii Pro Nano で計測した視線データから視線の分布を示すヒートマップを作成した．映像視聴時の参加者の視線の分布を図 3 に示す．色が明るい部分に視線がより集中していることを示す. 通常字幕では画面中央付近に加え，画面下部に視線が集中していることが分かる．一方，CEFR+音声変化条件と CEFR 条件では，分布が比較的似ており，画面の中央付近に視線が集中している．したがって通常字幕に比べてより映像に集中できることが示唆された．

5.2 映像の理解度

理解度テストの結果を図 4，図 5 に示す．図 4 は映像ごとの比較，図 5 は手法ごとの比較である．図中の緑の三角形は平均値，黒線は中央値を表しており，$*$ は有意水準 5%で理解度に差があることを示している．映像ごとの理解度テストの点数は映像 C(M=10.11)，映像 A(M=10.05)，映像 B(M=9.21) の順に高い結果となった．一元配置分散分析 (ANOVA) を実施したところ，映像間の理解度に有意差は生じなかった ($F(2, 51)$=3.82, p >.05)．したがって，映像による難易度調整は問題ないと考えられる．

手法ごとの理解度テストの点数は通常字幕 (M=10.26)，CEFR+音声変化 (M=10.0)，CEFR(M=9.11) の順に高い結果となった．一元配置分散分析 (ANOVA) を実施したところ，手法間の理解度に有意差が見られた ($F(2, 51)$=6.20, p <.01)．Bonferroni 法により，調整化された有意水準 α=0.0167 を求めた (補正前 p=0.05)．Bonferroni の多重比較検定を実施した結果，通常字幕と CEFR 条件，CEFR+音声変化条件と CEFR 条件の間で有意差が確認された．今回の提案手法である CEFR+音声変化の字幕では，CEFR 条件に比べて理解度が有意に高かった．また，通常字幕と CEFR+音声変化の条件間での有意差は生じなかったことから，提案手法は通常字幕と同程度の理解度が保たれたと考える．



(a) 通常字幕　(b) CEFR+音声変化条件　(c) CEFR 条件

図 3: 映像視聴時の参加者の視線分布

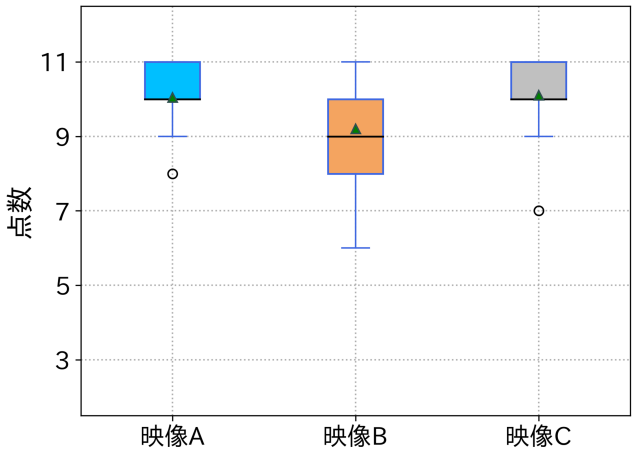


図 4: 映像ごとの理解度

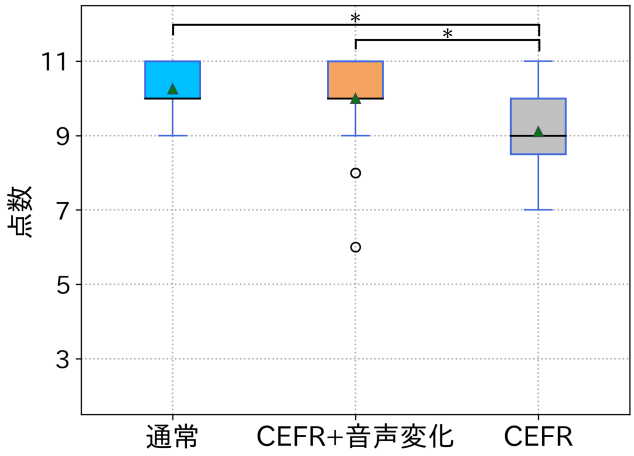


図 5: 手法ごとの理解度

5.3 認知負荷

タスク時の主観的な認知負荷の結果を図 6 から図 8 に示す．図中の緑の三角形は平均値，黒線は中央値を示しており，$*$ は有意水準 5%で差があることを示している．図 6 は知覚的負荷を示したものであり，数値が高いほど，見る，考える，記憶するなどの負荷を要したといえる．

参加者の主観的な知覚的負荷は CEFR 条件 (M=4.2)，CEFR+音声変化条件 (M=3.7)，通常字幕 (M=2.7) の順になった．一元配置分散分析 (ANOVA) を実施した結果，手法間の知覚的負荷に有意差が生じた ($F(2,51)$=6.82, p <.01)．. Bonferroni 法により，調整化された有意水準 α=0.0167 を求め (補正前 p=0.05)，Bonferroni の多重比較検定を実施した結果，通常字幕と CEFR 条件の条件間で有意差が生じた．CEFR+音声変化の条件では，通常字幕に比べてタスク時にかかる知覚的負荷は少し大きかったが，有意差は生じなかった．

図 7 は，参加者がタスクの達成にどの程度努力を要したかを示したものであり，数値が高いほど，英語の映像を視聴し，理解度テストに回答するというタスクの達成に苦労したといえる．参加者の主観的な努力は CEFR 条件 (M=4.3)，CEFR+音声変化条件 (M=3.7)，通常字幕 (M=2.94) の順になった．一元配置分散分析 (ANOVA) を実施した結果，手法間のタスクに要した努力に有意差が生じた ($F(2,51)$=5.04, p <.05)．Bonferroni の多重比較検定を実施した結果，通常字幕と CEFR 条件間で有意差が生じた．CEFR+音声変化条件では，通常字幕に比べてタスクの達成に努力を要したものの有意差は生じなかった．

図 8 は，参加者がタスク時に感じたストレスを示したものであり，数値が高いほど，ストレスが大きかったといえる．参加者による主観的なストレスは CEFR 条件 (M=3.0)，CEFR+音声変化条件 (M=2.8)，通常字幕 (M=2.2) の順になった．一元配置分散分析 (ANOVA) を実施した結果，手法間のタスク時のストレスに有意差は生じなかった ($F(2,51)$=1.71, p >.05)．本実験では，通常字幕，CEFR 条件，CEFR+音声変化条件のいずれの条件間においても，参加者の主観的なストレスに有意差は確認されなかった．

5.4 ユーザビリティ

システムユーザビリティスケールの評価結果について示す．スコアの平均は，通常字幕が 83.3 点，CEFR+音声変化条件が 66.1 点，CEFR 条件が 57.9 点となり，通常の字幕が最も使いやすいことが示された．一般的なシステムの平均値は 68 点とされており，CEFR+音声変化条件はわずかに平均を下回った．

5.5 アンケート結果

実験後に実施したアンケートの結果を示す．今後洋画などを視聴する際，実験で用いた手法 (通常の日本語字幕，CEFR 条件，CEFR+音声変化条件) のどれを用いたいか質問したところ，通常字幕が 8 名，CEFR+音声変化条件が 8 名，CEFR 条件が 2 名と回答した．その字幕提示手法を選んだ理由について，

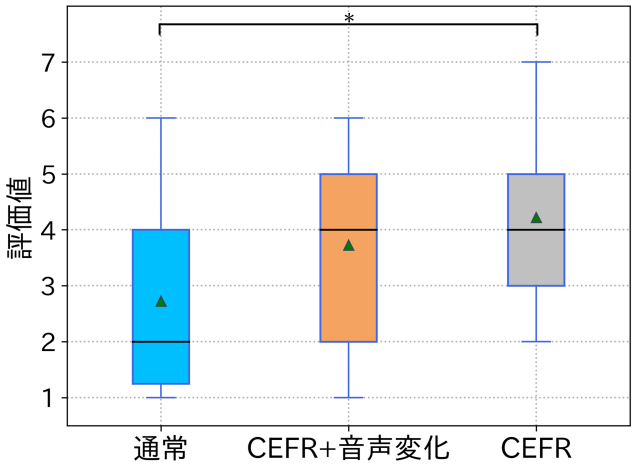


図 6: 知覚的負荷

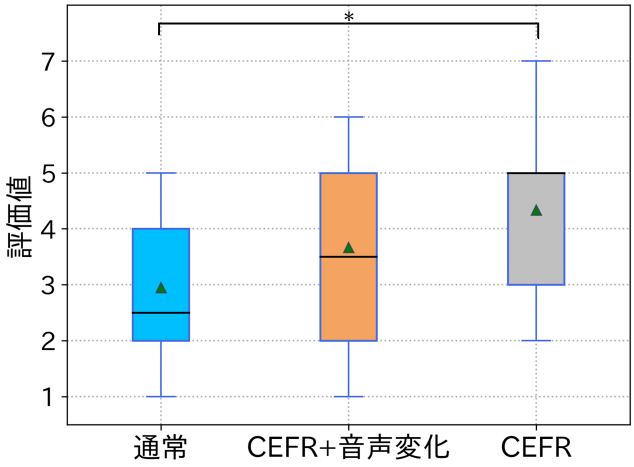


図 7: タスク達成に要した努力

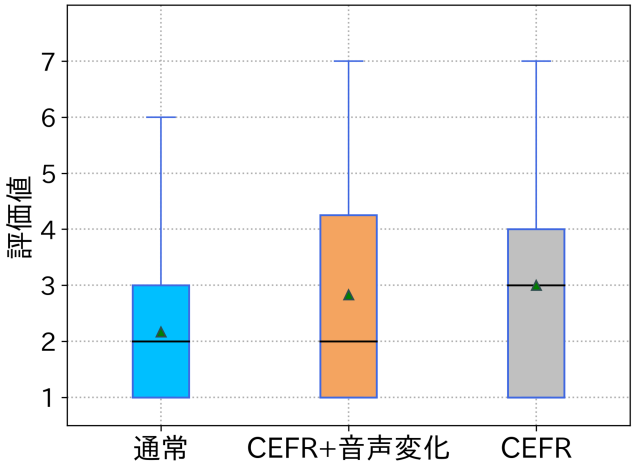


図 8: タスク中に感じたストレス

通常字幕では,「いつも見ている形式で理解しやすいから」,「他の提示手法だと全てが説明されているわけではないので所々意味を理解できないところがあるため」などの回答が得られた. CEFR+音声変化条件では,「字幕がシンプルであったし，ある程度多くの単語をカバーしていて，英語の聞き取りと字幕を読むバランスが取れていると感じた」,「視線移動が少なく，聞き取れなかったところをちょうど補完してくれるような字幕だったから」などの回答が得られた. CEFR 条件では,「映画に集中できそうだから」,「内容を理解しつつ，リスニングの勉強になると思ったから」などの回答が得られた.

6 おわりに

本研究では，映像集中を目的とした第二言語字幕において，音声変化に着目し，情報量を増加した字幕提示手法を提案した. 評価実験では，提案手法による字幕提示が映像視聴時の視線分布，理解度，認知負荷，ユーザビリティに与える影響を調査した. 以下に，本研究で設定した仮説に対する結論を示す.

- **仮説 1: 提案手法による字幕では，一般的な字幕に比べて映像に集中できる.**

一般的な字幕の場合，画面下部に視線が集中していたが，提案手法では画面中央付近に視線が集中していたため，映像に集中できていた可能性が示唆される.

- **仮説 2: 情報量を増やした字幕 (CEFR+音声変化条件) では，CEFR 条件の字幕に比べて高い理解度を保つことができる.**

映像の理解度は高い順に一般的な字幕，CEFR+音声変化条件の字幕，CEFR 条件の字幕であった. 検定の結果，CEFR+音声変化条件では，CEFR 条件より有意に理解度が高く，一般的な字幕と CEFR+音声変化の条件間では理解度に有意な差が生じなかった. したがって，音声変化を考慮した提案手法では，一般的な字幕と同程度の理解度を保つことができたと考える.

- **仮説 3: 情報量を増やした字幕 (CEFR+音声変化条件) では，CEFR 条件の字幕に比べて映像視聴時の認知負荷を減らすことができる.**

CEFR+音声変化条件では，知覚的負荷，タスク達成に要した努力，タスク中に感じたストレスの全ての項目において，CEFR 条件より認知負荷を下げることができたが，統計的な有意差は確認されなかった.

課題と今後の展望について述べる. 本研究では，英語特有の音声変化に着目し，情報量を増加した字幕提示手法を提案した. 一方で，情報量を増加したものの，一般的な字幕に比べ約 6 割程度の字幕の量であることから，この字幕提示手法を使いこなすにはある程度英語力があることが前提となる. 本実験の 18 名の参加者の TOEIC スコアは全員 600 点以上であり，比較的英語の基礎知識が備わっていたことが考えられる. また，TOEIC のスコアが 800 点以上の参加者の一人は，情報量のより少ない従来の字幕提示手法 (CEFR 条件) を好んだことから，それぞれの参加者にとって最適な字幕の情報量が異なると考えられる. 今後の展望として，個人の英語レベルに応じて提示する情報量を変化させることが可能な字幕提示手法が必要であると考える. また，本研究では 3 分から 3 分半程度の映像を用いたが，実際の洋画やドラマは 30 分を超えるものが多いため，より長い映像での実験が必要であると考える. さらに，話し言葉の多いフレンズだけでなく，堅い雰囲気のスピーチなど，映

像の特徴によって手法の有効性が異なると想定されるため，他の種類の映像を用いて調査を進めたいと考える．

謝　　辞

本稿で示した研究の一部は，Society 5.0 実現化研究拠点支援事業 (グラント番号: JPMXP0518071489) の支援のもと実施されている．

文　　献

[1] G. Dizon and B. Thanyawatpokin, “Language learning with netflix: Exploring the effects of dual subtitles on vocabulary learning and listening comprehension,” *Computer Assisted Language Learning*, vol. 22, no. 3, pp. 52–65, 2021.

[2] L. Bergen, T. Grimes, and D. Potter, “How attention partitions itself during simultaneous message presentations,” *Human Communication Research*, vol. 31, no. 3, pp. 311–336, 2005.

[3] G. d’Ydewalle, J. Van Rensbergen, and J. Pollet, “Reading a message when the same message is available auditorily in another language: The case of subtitling,” in *Eye movements from physiology to cognition*. Elsevier, 1987, pp. 313–321.

[4] 西優己, 中村優吾, B. Dawton, 福嶋政期, and 荒川豊, “第二言語の映像視聴時の認知負荷を考慮した字幕提示手法の提案,” in **マルチメディア、分散、協調とモバイル** *(DICOMO2023)* **シンポジウム**, 2023.

[5] S. G. Hart, “Nasa-task load index (nasa-tlx); 20 years later,” in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[6] Y. Hu, J. Kautz, Y. Yu, and W. Wang, “Speaker-following video subtitles,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2, pp. 1–17, 2015.

[7] K. Kurzhals, E. Cetinkaya, Y. Hu, W. Wang, and D. Weiskopf, “Close to the action: Eye-tracking evaluation of speaker-following subtitles,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 6559–6568.

[8] M. C. Potter, “Rapid serial visual presentation (rsvp): A method for studying language processing,” in *New methods in reading comprehension research*. Routledge, 2018, pp. 91–118.

[9] R. Rzayev, P. W. Woźniak, T. Dingler, and N. Henze, “Reading on smart glasses: The effect of text position, presentation type and walking,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–9.

[10] E. Gannon, J. He, X. Gao, and B. Chaparro, “Rsvp reading on a smart watch,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2016, pp. 1130–1134.

[11] C.-H. Chen and Y.-H. Chien, “Effects of rsvp display design on visual performance in accomplishing dual tasks with small screens,” *International Journal of Design*, vol. 1, no. 1, 2007.

[12] B. Milliner, “Evaluating the lexical difficulty of teaching materials with nwlc,” *ELF 2*, p. 49, 2022.

[13] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.

[14] P. Carley and I. M. Mees, *American English phonetics and pronunciation practice*. Routledge, 2019.

[15] P. Carley and I. M. Mees, *British English phonetic transcription*. Routledge, 2021.

[16] J. Brooke, “Sus: a “quick and dirty’ usability,” *Usability evaluation in industry*, vol. 189, no. 3, pp. 189–194, 1996.

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 16:00 ~ 18:10 | T5-A オンライン（Zoom Events)

医療・ヘルスケア (2)

座長:横田 治夫(城西大学)
コメンテータ:白石 優旗(筑波技術大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T5-A-6-01] 福祉支援施設の支援記録からのインシデントの予兆検出手法の改良

*山本 隆太[1]、上野 史[2]、太田 学[2] (1. 岡山大学大学院環境生命自然科学研究科、2. 岡山大学学術研究院環境生命自然科学学域)

16:30 ~ 16:55
[T5-A-6-02] レム睡眠のリアルタイム自動検出に関する研究

*佐野 史弥[1]、国生 泰資[1]、堀江 和正[2]、古木 淳也[4]、鈴木 陽子[3]、阿部 高志[3]、北川 博之[3] (1. 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構、4. 筑波大学大学院人間総合科学学術院 人間総合科学研究群)

16:55 ~ 17:20
[T5-A-6-03] リアルタイム睡眠ステージ判定システムの開発と評価

*国生 泰資[1]、堀江 和正[2]、鈴木 陽子[3]、阿部 高志[3]、北川 博之[3] (1. 筑波大学システム情報工学研究群知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構)

17:20 ~ 17:45
[T5-A-6-04] 特定健診データに含まれる自由記述に着目した予後予測

*大塚 皇輝[1]、池之上 辰義[2]、福間 真悟[3]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学　ソーシャルコンピューティング研究室、2. 滋賀大学　データサイエンス・AIイノベーション研究推進センター、3. 京都大学 医学研究科 人間健康科学科系専攻)

17:45 ~ 18:10
[T5-A-6-05] AI 音楽療法士（さつきさん）の開発研究

*小杉 尚子[1]、石井 健太郎[1]、児玉 直樹[2] (1. 専修大学、2. 新潟医療福祉大学)

福祉支援施設の支援記録からのインシデントの予兆検出手法の改良

山本　隆太†　　上野　　史††　　太田　　学††

† 岡山大学大学院環境生命自然科学研究科　〒 700–8530 岡山市北区津島中 3-1-1
†† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山市北区津島中 3-1-1
E-mail: †p0pm8dxf@s.okayama-u.ac.jp, ††{uwano, ohta}@okayama-u.ac.jp

あらまし　上野らは，福祉支援施設の入所者が起こすトラブル（インシデント）の予兆検出のため，施設の支援記録本文を用いた BERT によるインシデントの予兆検出手法を提案した．また実験により，支援記録に含まれる日付などの付随情報を使用することで分類精度が向上する可能性も示唆した．しかし，入所者毎に個別に予兆検出器を学習したため，インシデントをあまり起こさない入所者に対して検出精度が悪かった．そこで本研究では，他の入所者の支援記録を学習に使用した分類器による実験を行い，入所者本人の支援記録のみで学習した分類器と精度を比較する．さらに本研究では，付随情報を利用することで分類精度が向上する可能性について検討した．実験の結果，分類対象者を除く全ての入所者の支援記録で学習する汎用分類器の精度が最もよく，施設 A で平均 F 値 0.790，施設 B で平均 F 値 0.753 となった．またその汎用分類器にインシデントの有無の付随情報を利用した場合，さらに精度が向上し施設 A で平均 F 値 0.825，施設 B で平均 F 値 0.772 となった．

キーワード　インシデント予兆検出，BERT，付随情報，医療・ヘルスケア，福祉支援施設

1　はじめに

近年，高齢者や障がい者が入所する福祉支援施設の需要が高まってきている．福祉支援施設では，入所者の人数に対する職員の数は限られており，支援には限度がある．そのため，入所者による器物破損や他者への傷害などといったインシデントが施設内で発生すると，業務の妨げとなり，入所者は十分な支援が受けられなくなる．しかし，インシデントを事前に察知することはベテランの職員でも難しいため，入所者が起こすインシデントの予兆を発見し未然に防止することは重要な課題である．

本研究では，福祉支援施設の支援記録からインシデントの予兆を検出する [1]．本研究では，実際に現場で働く施設の職員から見て，支援記録本文に利用者の衝動性が高い状態，あるいはネガティブな心理状態が現れているものをインシデントの予兆と定義している．本研究で利用する手法は，上野らが [1] で提案したものであり，上野らは福祉支援施設の利用者 10 人の支援記録を用いて，各入所者のデータのみで個別に学習した分類器の予兆検出性能を検証した．予兆の検出は，予兆の有無を示すラベルを用いた 2 クラス分類により実現しており，その際上野らは入力として，支援記録の記録本文のみでなく支援記録に含まれる日付などの付随情報を使用することで，分類精度が向上する可能性を示唆している．本研究では，上野らの手法を予兆の程度に応じた 4 値分類に拡張し，予兆を検出する．具体的には，支援記録の本文に加えて日付データなどの付随情報も利用した分類器を作成し，分類器に入力された支援記録を「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 クラスに分類することで予兆の検出を実現する．4 値分類への拡張により，施設の職員が予兆の程度に応じてより適切な対応ができるようになる．検出には，自然言語処理モデルの 1 つである Bidirectional Encoder Representations from Transformers (BERT)[2] を利用した上野らの予兆検出器を拡張して利用する．上野らの研究では，分類器の学習に分類対象となる入所者の記録しか用いておらず，作成した分類器は，他の入所者の記録分類には向かない．そのため本研究では，他の入所者の支援記録を学習に使用した分類器による実験を行い，学習に入所者本人の記録しか用いない場合と精度を比較し，分類器の汎用性について考察する．さらに本研究では，分類精度向上のため，付随情報の有効な利用方法を検討する．

本稿の構成を述べる．2 節では関連研究を紹介する．3 節では，本実験で使用する支援記録と付与されたインシデントと予兆のラベルについて説明する．4 節では，本稿で用いるインシデントの予兆検出手法を説明する．5 節で実験について説明し，結果と考察を述べる．6 節では，まとめと今後の課題について述べる．

1：本研究は, 岡山大学の研究倫理審査専門委員会の承認を得て実施した.(研 2104-005)

2　関連研究

2.1　BERT

Bidirectional Encoder Representations from Transformers (BERT)[2] は，Devlin らが提案した自然言語処理モデルであり，双方向 Transformer[3] というニューラルネットワークを利用している．BERT は大量の教師なしデータによる事前学習を行ったモデルを，比較的少量の教師ありデータでファインチューニング (FT) して個々のタスクに合わせて学習する．これにより事前学習済みモデルをそのまま適用するよりも性能の向上が期待できる．

BERT の特徴として，Transformer を利用し，文章を文頭

と文末の双方向から学習することで，文脈を考慮できることが挙げられる．BERT ではこの双方向からの学習を，Masked Language Model (MLM) と Next Sentence Prediction (NSP) の 2 つのタスクを行うことで可能としている．MLM では，まず入力文章を単語ごとにトークンに分割し，その中からランダムに選ばれた 15 %のトークンを確率的に別のトークンで置き換える．具体的には，選択された 15 %のトークンのうち，その 80 %を [MASK] という特殊なトークンに，10 %をランダムな別のトークンに置き換え，残りの 10 %はそのままのトークンとする．その後 [MASK] トークンに置き換えられたトークンを予測し学習するのが MLM のタスクである．NSP は 2 つの入力文を受け取り，そのペアにおいて 2 つ目の文が 1 つ目の文の後続の文となっているかを予測するタスクである．具体的には，文の片方を 50 %の確率で他の文に置き換え，それらが隣り合っているか隣り合っていないか判別することで学習する．BERT は発表時に，General Language Understanding Evaluation (GLUE) ベンチマーク [4] の 8 つのタスク，Stanford Question Answering Dataset (SQuAD)[5] の v1.1 と v2.0，Situations With Adversarial Generations (SWAG)[6] の計 11 の自然言語処理タスクで最高精度を記録した．

2.2 インシデント検出

介護や福祉，医療分野におけるインシデントやトラブルへの対応は大きな医療事故や介護事故を防ぐためにも重要であり，事例の分析や予防手法の提案，支援システムの開発など様々な研究がされている．

例えば田中らは，大量に蓄積されている医療事故情報やインシデント情報をトラブルの未然防止に活用するための仕組みとして，機械学習を利用した階層型クラスタリングの方法を提案した [7]．具体的には，まず初めに日本医療機能評価機構が WEB 公開している医療事故データベースから必要情報をダウンロードして，ヒヤリハットデータベースを作成した．その後，それらに階層型クラスタリングを適用し，ヒヤリハットデータベース全体のクラスタ構造を各事例の類似性という視点で階層的に表示することで，インシデントをいくつかのグループに分類することを可能とした．このような自動分類により必要なクラスタを深堀できるなど，各医療機関でのインシデントへの事前対応が可能になることが期待できると報告されている．

峯崎らは，複数介護施設における事故報告書を使用して，テキストマイニングおよび機械学習によって事故の起こりうる状況を分析した [8]．テキストデータ部分について単語文書行列を作成して元のデータと結合し，事故の結果を目的変数としてランダムフォレスト法で重要な単語クラスタ及び重要な単語を抽出し，決定木学習によって可視化した．その結果，1. 出血は目に見える外傷が発見されて介護士が初めて気づく，2. 移乗などの際に転倒による打撲が発生する可能性がある，3. 食事時に誤嚥が起こりうる，4. 無断外出につながるいくつかの行動がある，などのいくつかの事故パターンを発見した．

中沢は，高齢者介護施設において，施設職員の支援や負担軽減を目的として，夜間徘徊者の発見，通報，声掛け，足止めを総合的に実行する見守りロボットを開発した [9]．ロボットに搭載された赤外線カメラにより施設廊下を監視し，HOG 特徴量とフレーム間差分法による動体検知を併用することで徘徊者を検出した．また，その後施設職員へ無線連絡をすると同時に，ロボットが徘徊者に近づき，職員の音声で呼びかけることで職員が到着するまでの時間稼ぎが可能なシステムを構築した．

青木らは，介護 AI 記録分析ツール “けあらぽ” を開発した [10]．これは同社の開発した介護 AI 入力予測ツールによって作成された介護記録を対象とした記録分析ツールであり，施設利用者の日常生活における基本動作の評価値を推定することを可能とした．今後は介護記録から利用者の事故予兆や状況悪化の予測を行うことを目指していると報告されている．

上野らは，福祉支援施設で問題となるインシデントの予兆検出のために，BERT を利用した支援記録を予兆の有無で分類する手法を提案した [1]．具体的には，支援記録の記録本文とその付随情報を入力として，予兆の有無の 2 クラス分類を行うことで予兆の検出を実現している．実験では，利用者 10 人の支援記録を用いて提案する予兆検出器の精度を検証し，平均で 0.87 の F 値を確認した．また，8 人の利用者は付随情報を活用することで検出精度が向上することを確認した．

本研究は，支援記録を利用して施設利用者の介助の助けとする点では [10] と類似しているが，[10] は文例を選択することで介護記録が生成できる入力支援ツールがあるため，入力形態が統一化されている記録を利用している．一方本研究では，文章の記述形式といったものが定まっていない自由記述形式の記録を利用する．また，画像処理による支援システムを構築する [9] に対し，本研究では主に自然言語処理によって支援記録の文章を解析して用いる．さらに，[7] や [8] が介護施設や医療現場の，事故やインシデントの記録を分析するのに対し，本研究では記録からインシデントの予兆の検出をする．

3 支援記録と付与ラベル

3.1 支 援 記 録

本研究で用いる支援記録は，福祉支援施設の職員が各入所者に対して実施した支援について記録したものである．記録が書かれるタイミングは様々であり，支援の直後に書かれたものや業務終了時に一日分の支援をまとめて記録したものなどがある．また支援記録の本文は，文の長さや文字数，記述形式といったものが定まっていない自由記述形式の文章となっている．使用する支援記録の例を表 1 に示す．表 1 の通り，支援記録には先に説明した自由記述の本文以外にも複数の情報が含まれている．入所者 ID は施設入所者を判別するための 6 桁の数値である．処理日付はその記録の入力された年月日を意味する．入力者 ID は支援をして記録を書いた支援者を判別するための 6 桁の数値である．インシデントラベルはその記録に含まれるインシデントの情報を，予兆ラベルはその記録に含まれる予兆の情報をそれぞれ表している．これらのラベルは元の支援記録には含まれておらず，後述の基準に従って人手で付与されたものを本研究で使用している．

表 1　支援記録とラベルの例

入所者 ID	入力者 ID	処理日付	記録本文	予兆ラベル	インシデントラベル
000111	000011	2020/04/12	寝ていることが多い。奇声もなく静かに過ごされている。	予兆なし	なし
000111	000012	2020/04/12	作業：朝礼を終え、イオン水を飲んだあとに職員への他害がある。クールダウンまでは早かったが、原因は不明である。	予兆（強）	暴力#職員
000122	000011	2020/04/13	舎活動：居室で穏やかに過ごされている。(14：30）大きな声が聞こえ、見に行くと笑っている。	予兆（弱）	なし

表 2　インシデントとそのラベル

種類	対象	ラベル
器物破損	自分の所有物	器物破損#自分の所有物
	他入所者の所有物	器物破損#他入所者の所有物
	職員の所有物	器物破損#職員の所有物
	施設の所有物	器物破損#施設備品
	その他	器物破損#その他
暴力	職員	暴力#職員
	施設の入所者	暴力#施設の入所者
	その他	暴力#その他

3.2　インシデントの定義とインシデントラベル

本研究では，周囲に大きな損害を与える状況をインシデントと定義する．周囲に大きな損害を与える状況の例としては，暴力全般や水や唾をかけられるなどの職員や他の入所者が不快な思いをするものや，器物破損のような修復コストがかかるものなどが挙げられる．なお本研究ではインシデントは，暴力や器物破損が自己以外の対象に害を与える他害に限定する．そのため，自傷行為のような自身に害を与えるものはインシデントとしていない．この定義に従って，インシデントには表 2 のようなラベルが付与されている．一つの記録に対して複数のインシデントラベルが付与されることもある．このインシデントは福祉支援施設での支援業務を妨げるものであり，現場の施設職員の意見を参考に定められたものである．また本研究では，ここに定めたインシデントの種類を，予兆検出の分析には用いるが，インシデントの分類は行わない．

3.3　予兆の定義と予兆ラベル

本研究では，支援者である施設職員の判断に基づいて，支援記録本文に施設入所者の衝動性が高い，あるいはネガティブな心理状態が現れているものを予兆と定義する．予兆ラベルには「予兆 （弱)」「予兆（強)」の 2 種類がある.「予兆 （弱)」は，衝動性が高い，あるいはネガティブな心理状態が現れている状態,「予兆（強)」は衝動性あるいはネガティブな感情を抑えることができず，結果として周囲へ影響が出てしまっている状態と定義している.「予兆 （弱)」の例としては，自傷行為，情緒が乱れている，大声を発する，などがあげられ,「予兆（強)」の例としては，他者への暴力，器物破損などのインシデントや，物に当たる行為などが挙げられる．一方で，やる気がない，暴言，騒ぐ，などはそれだけでは予兆とはせず，記述内容の文脈によって予兆か予兆でないかを判断している．例えば，気分が悪い，つらそうな表情，などは精神面の話であれば「予兆 （弱)」であるが，体調面の話であればこれらは「予兆なし」となる．また一つの記録に対して「予兆（弱)」と「予兆（強)」の予兆ラベルが同時に付与されることもあり，本研究においてはそういった記録を「予兆 (両方)」として扱う．本研究では，この予兆ラベルを用いて，支援記録を「予兆なし」「予兆（弱)」「予兆（強)」「予兆（両方)」の 4 クラスへ分類を行う．

4　インシデントの予兆検出

4.1　インシデントの予兆の検出手法

図 1 に使用する予兆検出器の概略を示す．この検出器ではまず，入力された記録本文を BERT により文書表現ベクトルに変換する．また，支援記録から支援者 ID や日付情報などの本文以外の付随情報を取り出し，付随情報ベクトルに変換する．次に，これらの文書表現ベクトルと付随情報ベクトルを入力とする分類器によって支援記録の予兆を,「予兆なし」「予兆（弱)」「予兆（強)」「予兆（両方)」の 4 クラスに分類する．その際に使用する分類器は図 1 の右側に示す 4 層のニューラルネットワークである．この各層は全結合層であり，左から順に 256 次元，128 次元，32 次元，4 次元ベクトルへと線形変換される．活性化関数にはシグモイド関数，出力層には softmax 関数を用いている．損失関数には交差エントロピー誤差を採用し，Adam[11] によって最適化する．分類器のエポック数は 20，バッチサイズは 32 とする．

図 1 の右にある分類器の 1 層目では，文書表現ベクトルを 256〜0 次元，付随情報ベクトルを 0〜256 次元の全結合層へ入力して，ベクトルの次元を変換し，2 つのベクトルを合わせて 256 次元のベクトルとしている．また各層間では Dropout[12] を実行している．Dropout はニューラルネットワークの過学習防止のための手法の一つで，各層ごとに一定確率でランダムにノードの出力を 0 にすることで無効化して学習を進める．

分類器の出力結果は 0 から 1 の範囲で正規化された各ラベルに対する確率であり，その総和は 1 になる．本稿で行う 4 クラス分類では「予兆なし」「予兆（弱)」「予兆（強)」「予兆（両方)」の 4 つのクラスに対する確率が出力され，確率が最大のクラスに分類する．

図 1 の左の BERT の部分は，事前学習済みモデルを支援記録を用いてファインチューニングし分類に利用する．また本研究では，BERT 日本語 Pretrained モデル [2]を利用する．

また本研究で入力として使用する付随情報は表 3 に示す 5 種類である．日付情報として利用する月，日，曜日はそれぞれ 12 次元，31 次元，7 次元，入力者 ID は 100 次元，インシデント

2：https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

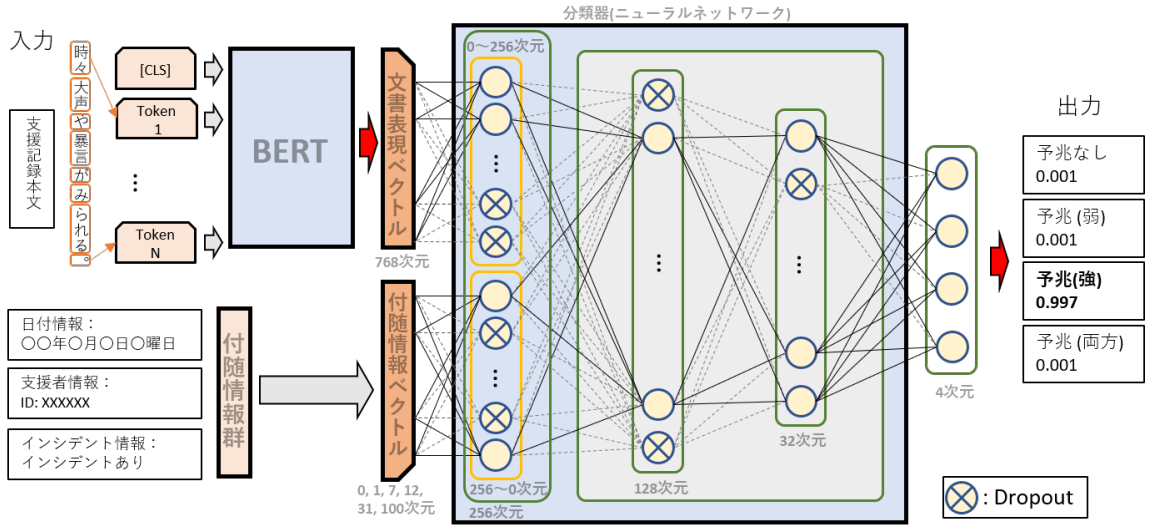


図 1　予兆検出器の概略図 [1]

表 3　付随情報と次元

付随情報	ベクトルの次元
月	12 次元
日	31 次元
曜日	7 次元
入力者 ID	100 次元
インシデントの有無	1 次元

の発生の有無は 1 次元の One-hot ベクトルで表現する．

4.2　個人専用分類器と汎用分類器

上野らは，BERT のファインチューニングに分類対象となる入所者本人の記録のみを用いたため，本稿ではこの分類器を個人専用分類器と呼ぶ．本稿では分類器の汎用性について考察するため，この個人専用分類器の他に，他の入所者の支援記録をファインチューニングに用いた汎用分類器を使用する．汎用分類器は，分類対象者と同施設に入所している他の入所者全員の支援記録で学習をする汎用分類器 (同施設)，分類対象者とは別の施設に入所している入所者全員の支援記録で学習をする汎用分類器 (他施設)，分類対象者を除く全ての入所者の支援記録で学習をする汎用分類器 (本人以外) の 3 つを使用する．

5　インシデントの予兆分類実験

5.1　実験に用いるデータ

インシデント予兆分類実験には，2 つの福祉支援施設の入所者 10 名ずつ，計 20 名の支援記録を用いる．この支援記録には，3 節で説明した「予兆なし」「予兆（弱）」「予兆（強）」の 3 つの予兆ラベルが付与されており，「予兆（弱）」と「予兆（強）」の両方のラベルが付与されている記録は「予兆 (両方)」ラベルが付与されているものとして扱う．これらの支援記録は図 1 の BERT のファインチューニングや，分類器の学習に利用する．表 4 に支援記録の統計情報を示し，図 2 に入所者ごとの支援記録の各予兆ラベルの割合をまとめる．表 4 には，入所者ごとに，各予兆ラベルの付与されている記録件数と全記録に占める割合を示している他，インシデントラベルが付与されている記録の件数とその全記録に占める割合についても記している．

5.2　実 験 内 容

本研究では，以下の (1),(2) について実験する．

(1)　汎用分類器による予兆分類

(2)　付随情報を利用した予兆分類

まず実験 (1) では，他の入所者の支援記録を学習に用いた汎用分類器により予兆分類をし，その精度の評価，考察をする．この際，4 節で述べた個人専用分類器と汎用分類器 (同施設)，汎用分類器 (他施設)，汎用分類器 (本人以外) の 3 つの汎用分類器を使用し分類実験をする．

個人専用分類器は，分類対象の入所者本人のデータを用いて五分割交差検定により分類精度を求め，その他 3 つの汎用分類器は五分割交差検定を行わず，各支援記録集合で学習した分類器に対して，テスト用データとなる分類対象者の記録全件の分類精度を求める．

実験 (2) では，日付などの付随情報を利用して予兆分類をし，その精度を評価する．上野らの作成した付随情報を入力とする分類器は，入力の文書表現ベクトルと付随情報ベクトルをそれぞれ 128 次元のベクトルに変換し，利用している．実験 (2) では，このベクトルの次元を変更して，付随情報の有効性を比べる．具体的には，分類器 1 層目の 256 次元のうちの文書表現ベクトルと付随情報ベクトルの重み割合を，16:0〜0:16 まで変化させ，その精度を比較する．重み割合が 16:0 の場合は文書表現ベクトルのみ，0:16 の場合は付随情報ベクトルのみを利用した分類となる．

また実験 (2) では，付随情報ベクトルの次元に文書表現ベクトルの次元を合わせた実験も行う．つまり，各付随情報ベクトルの次元はそのまま，文書表現ベクトルを付随情報ベクトルの次元まで削減し入力とする．例えば，付随情報が「入力者 ID」の場合は，文書表現ベクトルと付随情報ベクトルがそれぞれ

表 4　支援記録の統計情報

施設	人物	性別	記録件数 (割合)					インシデント件数
			予兆なし	予兆（弱）	予兆（強）	予兆（両方）	合計	
A	a	女性	1,519(62.9%)	646(26.7%)	88(3.6%)	163(6.7%)	2,416	106(4.4%)
	b	女性	2,207(64.2%)	999(29.0%)	83(2.4%)	150(4.4%)	3,439	205(6.0%)
	c	女性	1,830(55.5%)	607(18.4%)	503(15.2%)	360(10.9%)	3,300	630(19.1%)
	d	女性	1,510(78.6%)	304(15.8%)	14(0.7%)	93(4.8%)	1,921	87(4.5%)
	e	男性	2,800(73.7%)	832(21.9%)	83(2.2%)	82(2.2%)	3,797	156(4.1%)
	f	男性	2,702(81.3%)	551(16.6%)	40(1.2%)	30(0.9%)	3,323	52(1.6%)
	g	男性	2,564(75.7%)	599(17.7%)	67(2.0%)	157(4.6%)	3,387	99(2.9%)
	h	男性	3,034(88.2%)	367(10.7%)	31(0.9%)	8(0.2%)	3,440	15(0.4%)
	i	男性	3,216(84.3%)	428(11.2%)	94(2.5%)	77(2.0%)	3,815	73(1.9%)
	j	男性	3,320(85.6%)	532(13.7%)	14(0.4%)	13(0.3%)	3,879	6(0.2%)
B	k	男性	1,781(90.0%)	63(3.1%)	121(6.0%)	37(1.8%)	2,002	149(7.4 %)
	l	男性	1,927(96.3%)	59(2.9%)	10(0.5%)	7(0.3%)	2,003	11(0.5%)
	m	男性	1,443(72.1%)	157(7.8%)	379(18.9%)	22(1.1%)	2,001	385(19.2%)
	n	男性	1,621(81.0%)	277(13.8%)	78(3.9%)	26(1.3%)	2,002	56(2.8%)
	o	男性	1,667(83.3%)	120(6.0%)	202(10.1%)	14(0.7%)	2,003	174(8.7%)
	p	男性	1,774(88.7%)	168(8.4%)	54(2.7%)	4(0.2%)	2,000	58(2.9%)
	q	男性	1,885(94.3%)	95(4.8%)	17(0.9%)	3(0.2%)	2,000	18(0.9%)
	r	男性	1,546(77.3%)	251(12.5%)	144(7.2%)	62(3.1%)	2,003	206(10.3%)
	s	男性	744(97.4%)	19(2.5%)	0(0.0%)	1(0.01%)	764	1(0.01%)
	t	男性	1,747(87.3%)	91(4.5%)	135(6.7%)	28(1.4%)	2,001	133(6.6%)

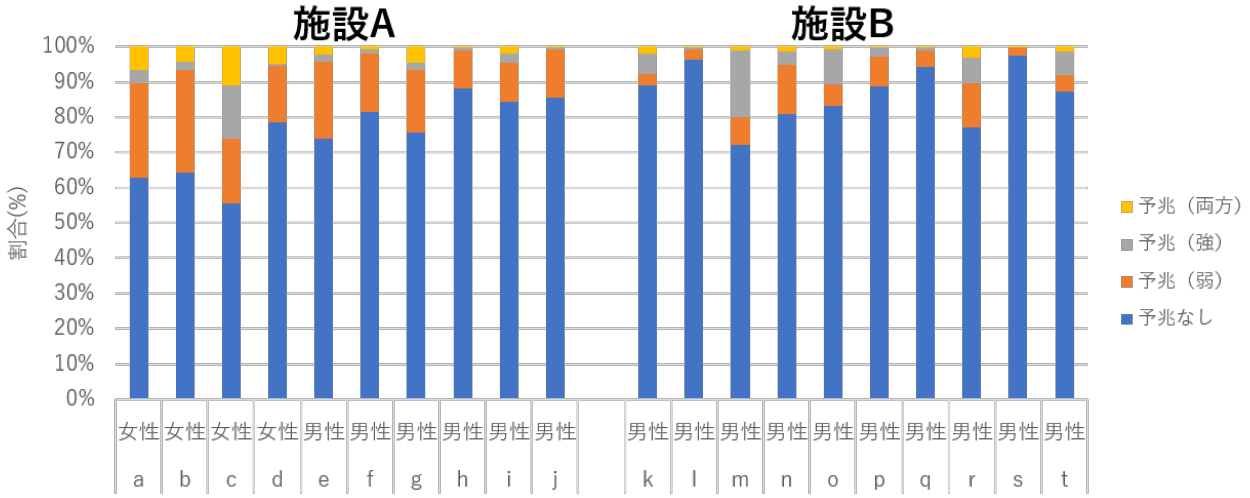


図 2　支援記録の予兆ラベルの割合

100 次元となり，図 1 の分類器 1 層目の次元は 256 次元ではなく 200 次元となる．この際，付随情報が「インシデントの有無」である場合のみ，文書表現ベクトルを付随情報ベクトルに合わせて 1 次元にせず，10 次元に削減して利用する．具体的には，図 1 の右側の分類器の一層目の次元を指定の次元とすることで実現している．これらの付随情報を用いた実験は，先の汎用分類器による予兆分類実験における汎用分類器 (本人以外) を用いて行う．

5.3　評 価 指 標

分類器の評価には，以下に示す再現率 (Recall)，適合率 (Precision)，F 値 (F-measure) を利用する．

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{F - measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3}$$

例えば，予兆ラベル「予兆（強）」についての TP, TN, FP, FN は下記のようになる．

- TP :「予兆（強）」だと分類された記録のうち，実際に「予兆（強）」である記録の件数
- TN :「予兆（強）」ではないと分類された記録のうち，実際に「予兆（強）」でない記録の件数

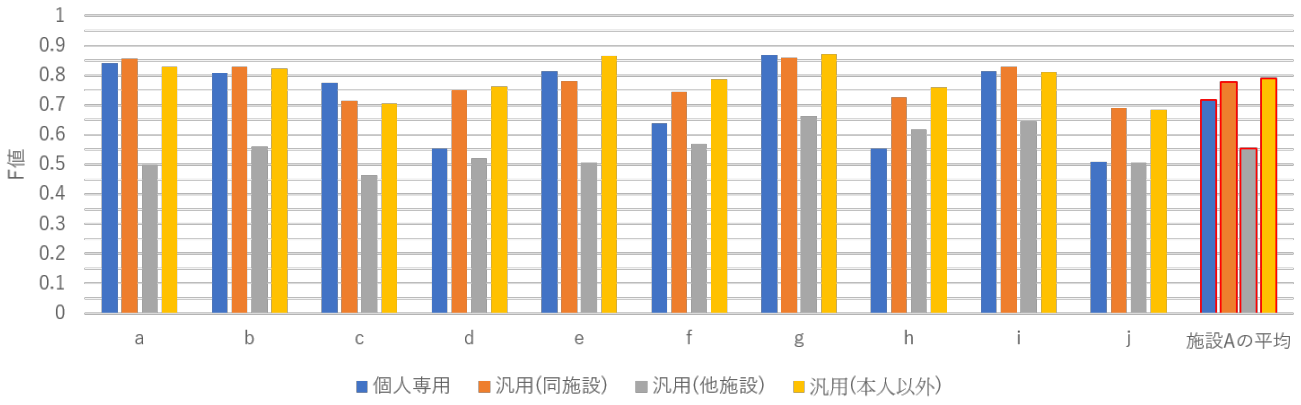


図 3　各入所者の分類器ごとの F 値 (施設 A)

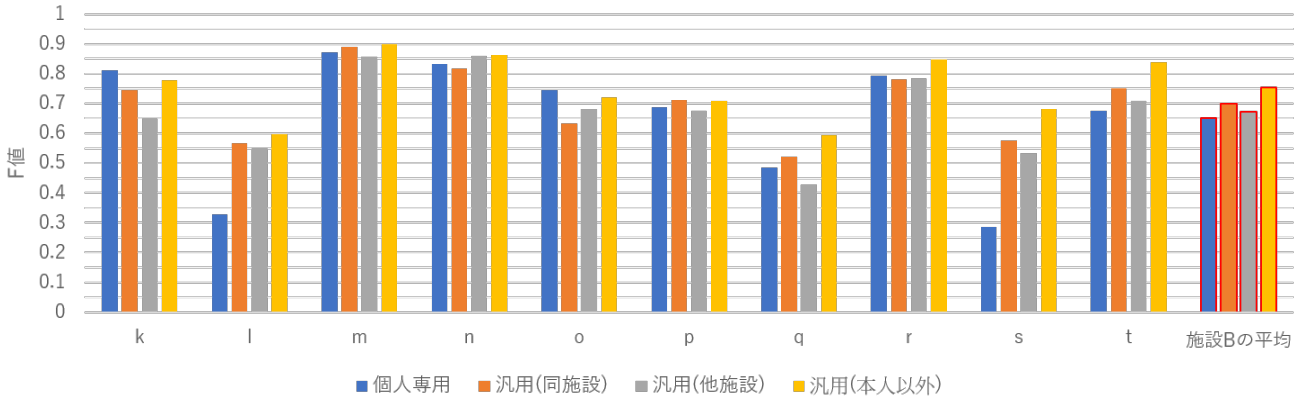


図 4　各入所者の分類器ごとの F 値 (施設 B)

- FP：「予兆（強）」だと分類された記録のうち，「予兆（強）」ではない記録の件数
- FN：「予兆（強）」ではないと分類された記録のうち，「予兆（強）」である記録の件数

よって再現率は全ての「予兆（強）」である文章に対し，分類器が正しく「予兆（強）」と分類したものの割合であり，適合率は分類器が「予兆（強）」と分類した文章のうち，実際に「予兆（強）」であるものの割合である．また F 値は再現率と適合率の調和平均となる．

以下の評価実験では，4 つの予兆ラベルについて個別にこの値を計算した後，その平均を求める．

5.4　汎用分類器による予兆分類実験

5.4.1　実 験 結 果

汎用分類器による予兆分類結果の F 値を図 3（施設 A の各入所者）と図 4（施設 B の各入所者）に示す．

図 3 と図 4 のいずれにおいても，入所者ごとの差異はあるが，各施設の平均で見ると汎用分類器 (本人以外) の精度が最も良く，次いで汎用分類器 (同施設) の精度が良いことが分かる．各入所者について見ると，施設 A の入所者 d，h や施設 B の入所者 l，q のように個人専用分類器の F 値が他の入所者に比べて小さくなっている人物は，個人専用分類器に比べて，汎用分類器 (同施設) や汎用分類器 (本人以外) の F 値が大きいことが分かる．また，図 3 の施設 A では汎用分類器 (他施設) の F 値が小さいのに対し，図 4 の施設 B の汎用分類器 (他施設) の F 値は，他の分類器とあまり差がないことが分かる．

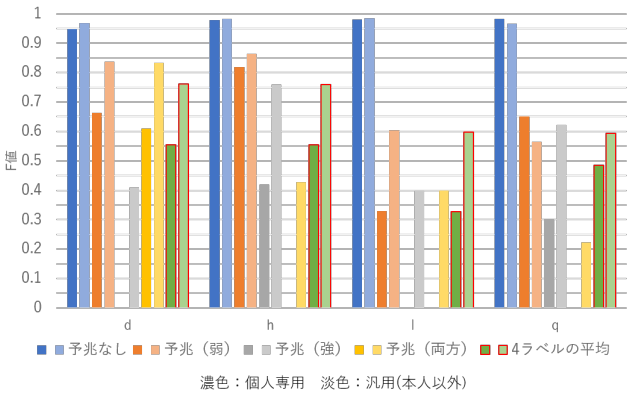


図 5　個人専用分類器と汎用分類器 (本人以外) の F 値の比較

5.4.2　考　　察

両施設の平均の F 値で，汎用分類器 (本人以外) 及び，汎用分類器 (同施設) の値が大きいのは，先に述べた施設 A の入所者 d，h や施設 B の入所者 l，q のような人物の分類精度の影響がある．当該入所者の分類結果では，汎用分類器 (本人以外)，汎用分類器 (同施設) の F 値が大きい．これらの人物は，表 4 や図 2 の支援記録の統計情報を見ると分かるように，「予兆（強）」

表 5 「予兆（強）」の少ない入所者の支援記録の分類事例

	記録本文	予兆ラベル	分類結果	
			個人専用	汎用 (本人以外)
(1)	朝の着替えの際、自身のパジャマを割く。	予兆（強）	予兆なし	予兆（強）
(2)	脱衣場で○○さんの頭を叩く。注意する。 突かれた事に怒った様子。	予兆（両方）	予兆（弱）	予兆（両方）
(3)	朝食後□□さんが食器を片付けないことに 腹を立て食堂内で追いかけ掴みかかろうとする。	予兆（両方）	予兆（強）	予兆（強）

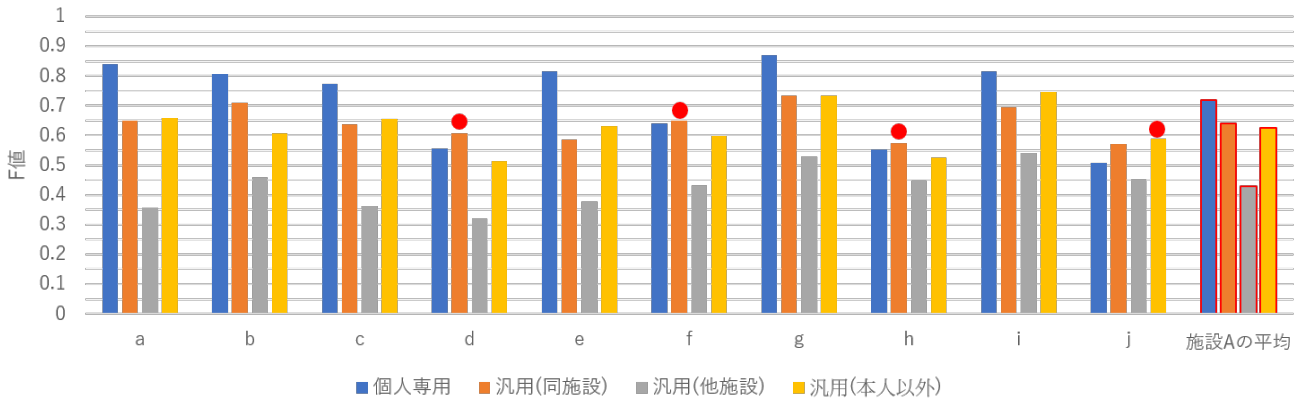


図 6　学習回数を揃えた実験における各入所者の分類器ごとの F 値 (施設 A)

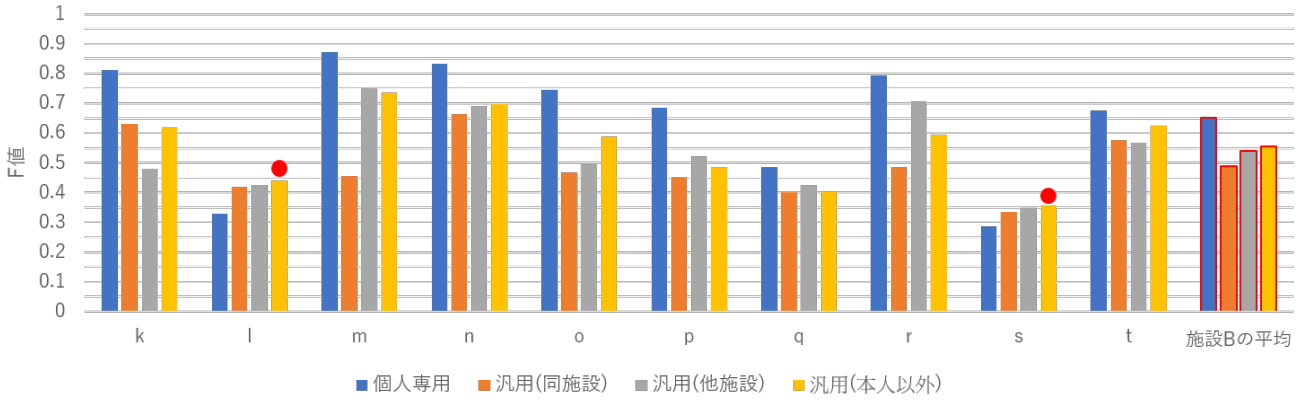


図 7　学習回数を揃えた実験における各入所者の分類器ごとの F 値 (施設 B)

を含む支援記録の割合が他の入所者と比べて小さい．そのため，自身の記録のみでの学習では「予兆（強）」や「予兆（両方）」ラベルの分類が困難である．図 5 に個人専用分類器と汎用分類器 (本人以外) の予兆ラベル毎の F 値を示す．色の濃い棒が個人専用分類器，淡い棒が汎用分類器 (本人以外) の F 値である．個人専用分類器の当該の入所者の F 値を見ると，「予兆（強）」や「予兆（両方）」ラベルの分類精度が悪いことが分かる．一方，汎用分類器 (本人以外) の当該入所者の予兆ラベルごとの F 値は，個人専用分類器と比べて「予兆（強）」や「予兆（両方）」の値が相対的に大きくなっていることが分かる．

本実験において，予兆の少ない入所者について個人専用分類器で誤分類され，汎用分類器 (本人以外) では正しく分類された記録本文の例を表 5 の (1)，(2) に，汎用分類器 (本人以外) でも誤分類となってしまった例を (3) に示す．表 5 の (1) の例では，「自身のパジャマを割く」という「予兆（強）」である器物破損の記述があるが，個人専用分類器では「予兆（弱）」に分類されている．(2) の例では，「頭を叩く」という「予兆（強）」の記述と「怒った様子」という「予兆（弱）」の記述があるため「予兆（両方）」であるが，個人専用分類器では，「予兆（弱）」に分類されている．汎用分類器 (本人以外) では，このような「予兆（強）」を含む記録の誤分類が少なくなっている．

一方いずれも誤分類となってしまった (3) の例では，「追いかけ掴みかかろうとする」という「予兆（強）」の記述と「腹を立て」という「予兆（弱）」の記述があるため，「予兆（両方）」であるが，分類結果はどちらも「予兆（強）」となった．

また，各汎用分類器で学習に使用する記録の件数を，五分割交差検定を行う個人専用分類器の学習で使用する件数まで削減し，学習回数を揃えた分類実験を行った．分類結果の F 値を図 6（施設 A の各入所者）と図 7（施設 B の各入所者）に示す．図 6 と図 7 のいずれにおいても，各施設の平均を見ると個人専

表 6　施設 A と施設 B の記録本文の統計情報

施設	文字数			「予兆（強）」を含む記録の文字数			人物の名前		「：」区切りの表記	
	最大値	最小値	平均	最大値	最小値	平均	件数	全記録に対する割合	件数	全記録に対する割合
施設 A	1,047	0	56.00	1,047	12	130.73	2,671	8.16% (2671/32,717)	9,567	29.24% (9,567/32,717)
施設 B	809	3	44.80	448	6	53.91	4,689	24.97% (4689/18,779)	188	1.00% (188/18,779)

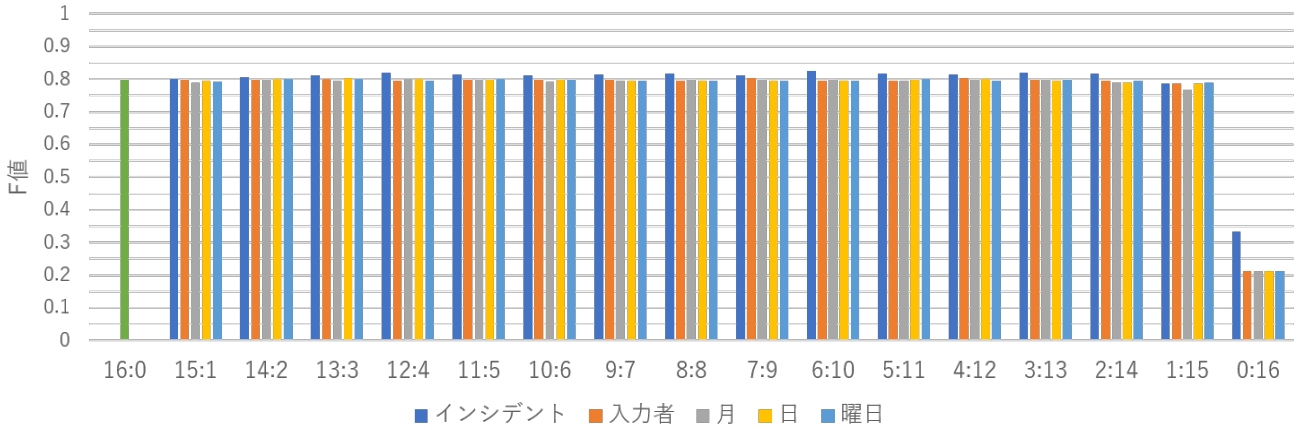


図 8　文書表現ベクトルと付随情報ベクトルの割合と施設 A の入所者の F 値の平均

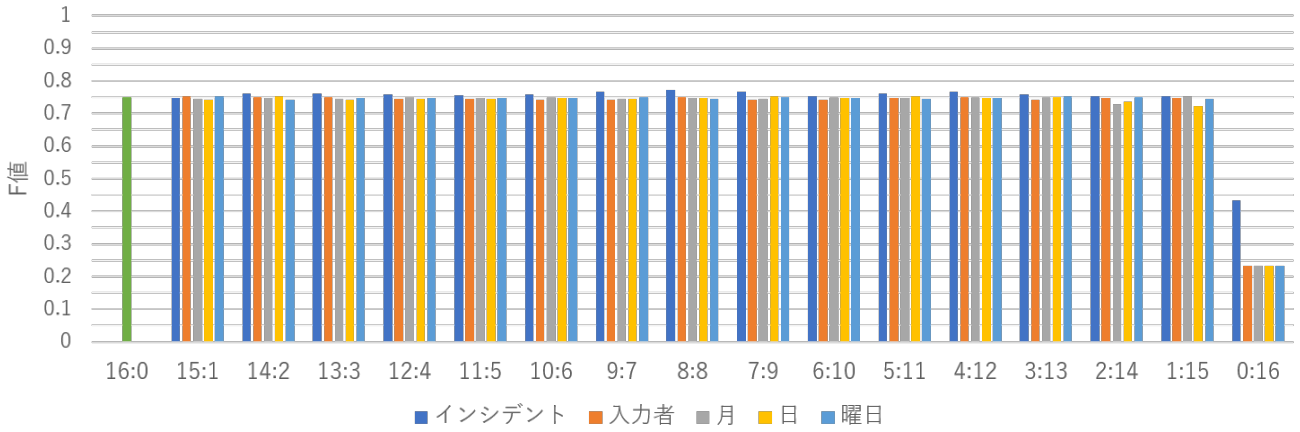


図 9　文書表現ベクトルと付随情報ベクトルの割合と施設 B の入所者の F 値の平均

用分類器の F 値が最大であることが分かり，各入所者についても個人専用分類器の F 値が最大となる人物が殆どである．よって，学習に利用する支援記録の件数が等しい場合は，学習に利用する記録は，本人の記録が良いことが分かる．しかし，図 6 と図 7 でグラフに赤丸を付けた入所者については，個人専用分類器と比べて汎用分類器 (本人以外) や汎用分類器 (同施設) の F 値が大きくなっている．これらの入所者は，先に述べた「予兆（強）」を含む支援記録が他の入所者と比べて少ない人物である．このことから，そのような「予兆（強）」を含む支援記録の少ない人物には，他者の記録を用いた学習も有効である．

また，図 3 と図 4 の比較でも，図 6 と図 7 の比較でも，汎用分類器 (他施設) の分類精度が，両施設で大きく異なっていたため，施設ごとに記録本文の内容に差があると考え，分析を行った．具体的には，両施設の支援記録の文字数の最大値，最小値，平均値を調べ，さらに「○○さん」というような，人物の名前を変換した記述や，記録の初めに「舎活動：～」や「作業：～」といった固有の表現を含む支援記録の件数を調べた．その結果を表 6 に示す．文字数から，施設 A の記録の方が文章が長い傾向にあることが分かる．中でも「予兆（強）」を含む記録の文字数については，施設 A は施設 B と比べ，最大値，平均値共に 2 倍以上の長さとなっている．また，人物名は施設 B の記録本文に多く出現し，「：」区切りの表記は施設 A の記録本文に多く出現しているように，施設間で記録の文章に差があることが分かった．

5.5　付随情報を利用した予兆分類実験

5.5.1　実 験 結 果

図 1 の分類器 1 層目の 256 次元の文書表現ベクトルと付随情報ベクトルの重み割合を変化させた予兆分類実験の結果の F 値を図 8（施設 A の入所者の平均）と図 9（施設 B の入所者の平均）に示す．なおこれらの図のグラフの 16:0 は，付随情報を用いない文書表現ベクトルのみで分類した結果の F 値である．

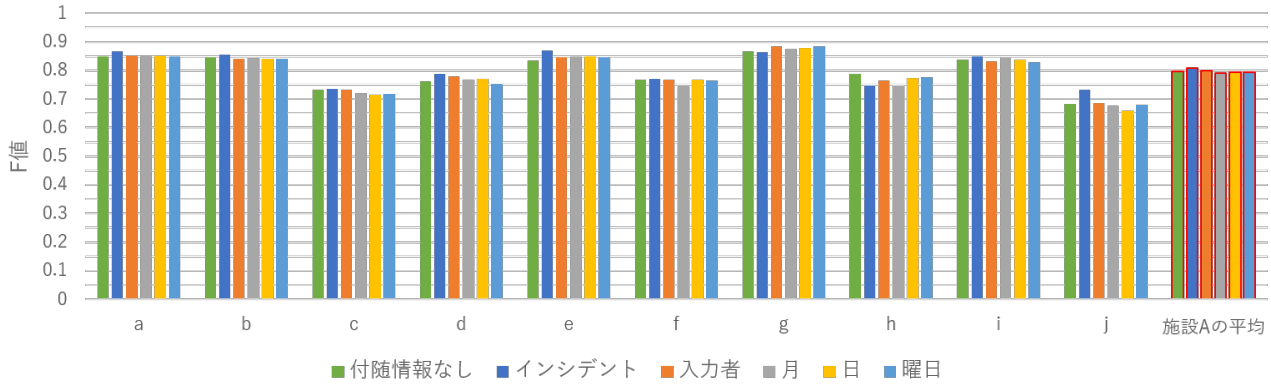


図 10　文書表現ベクトルを付随情報ベクトルに合わせ削減した実験の F 値 (施設 A)

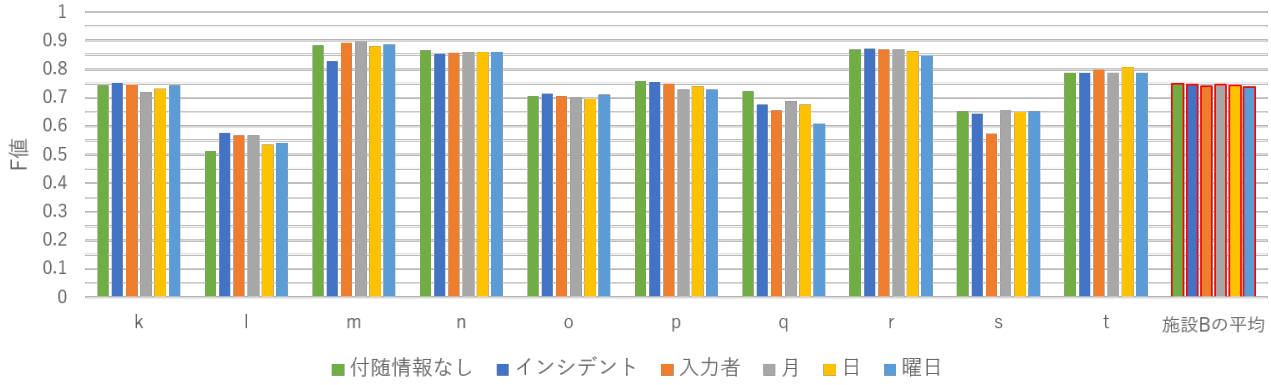


図 11　文書表現ベクトルを付随情報ベクトルに合わせ削減した実験の F 値 (施設 B)

0:16 はその反対に付随情報のみで分類した結果の F 値である．図 8 と図 9 のいずれにおいても，インシデントの有無以外の付随情報については，ベクトルの重み割合を変更しても，F 値がほとんど変化せず，0:16 においては F 値が大きく下がる．月，日，曜日の付随情報ベクトルのみを使用した場合は，分類結果として「予兆なし」のみしか出力しておらず，入力者 ID の付随情報ベクトルのみを使用した場合もほぼ「予兆なし」のみを出力した．一方，インシデントの有無の付随情報ベクトルを用いた場合は，F 値が大きくなっている場合がある．また，付随情報ベクトルのみを使用する場合においても，他の付随情報と比べると相対的に F 値が大きくなっている．

次に，文書表現ベクトルを付随情報ベクトルの次元まで削減して入力とした予兆分類の結果の F 値を図 10（施設 A）と図 11（施設 B）に示す．比較のため付随情報なし 16:0 の F 値も示している．図 10 と図 11 のいずれも，入所者ごとに結果の差異はあるものの，平均でみると付随情報なしの結果とあまり差がないことが分かる．

5.5.2　考　　察

図 8 と図 9 より，インシデントの有無以外の付随情報はほとんど有効性が確認できなかった．インシデントの有無については，インシデントの付随情報ベクトルを使用した場合に F 値が 0.03 以上大きくなることのある入所者が施設 A で 10 名中 5 名，施設 B で 10 名中 4 名の計 9 名存在しており，分類精度の向上に有効な場合があった．

図 8，図 9 の予兆分類実験において，インシデントの付随情報ベクトルをベクトル割合 1:15 で使用した場合に分類精度が向上した入所者の分類事例を表 7 に示す．表 7 の (1) が付随情報ベクトルを使用した場合に正しく分類ができるようになった例，(2) が付随情報ベクトルを使用した場合に誤分類となった例である．表 7 の (1) の例では,「他利用者の手をはねたり押して」という「予兆（強）」の記述があるにもかかわらず，文書表現ベクトルのみの場合では「予兆なし」に分類されているが，インシデントの付随情報ベクトルを用いた場合では正しく分類されている．(1) の支援記録は,「インシデントあり」となっており，正しく分類ができた．また例には示していないが,「予兆（弱）」のみを含む記録を，文書表現ベクトルのみの場合は「予兆（両方）」と誤分類，インシデントの付随情報ベクトルを用いた場合は「予兆（弱）」に正しく分類ができた事例がある．この例の支援記録は,「インシデントなし」となっており，正しく分類ができた．

一方 (2) は誤分類の例で「××さんを引っ張ろうとする」という「予兆（強）」の記述があり，文書表現ベクトルのみの場合は「予兆（強）」と正しく分類出来ているが，インシデントの付随情報を使用した場合は「予兆なし」の誤分類となった．この例は,「インシデントなし」の支援記録で，その上 1:15 と付随情報ベクトルの比重が大きいため,「予兆なし」に誤分類した可能性がある．

文書表現ベクトルを付随情報ベクトルの次元まで削減し入力

表 7　インシデントの付随情報ベクトルを利用した分類事例

	記録本文	予兆ラベル	分類結果	
			文書表現ベクトルのみ	付随情報あり (1:15)
(1)	職員（△△）に笑顔で付いて歩き、他の利用者が 職員の側に来たり手を引いたりすると、他利用者の手を はねたり押して移動させたりしていた。	予兆（強）	予兆なし	予兆（強）
(2)	起床し血圧測定を行う。その後××さんを 引っ張ろうとする事が何度もあった。	予兆（強）	予兆（強）	予兆なし

とした図 10 と図 11 の予兆分類実験では，入所者毎に差はあったものの，両施設ともに平均の値は，付随情報があってもなくてもあまり差はなかった．そのため，本実験のようにベクトルの次元を揃えるだけでは，分類精度の向上はあまり期待できない．

6 ま と め

本稿では，上野らの提案した福祉支援施設の支援記録からのインシデントの予兆検出器を，本人以外の入所者の支援記録で学習して予兆検出実験を行った．予兆の検出は，支援記録の本文を利用した分類器を作成し，対象の支援記録を「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 クラスに分類することで実現した．また，この分類に支援記録の本文以外の付随情報が有効かどうか実験した．

分類器の学習に本文以外の入所者の支援記録を利用する実験では，施設 A 入所者の全ラベルの平均 F 値は，個人専用分類器で 0.718，汎用分類器 (同施設) で 0.778，汎用分類器 (他施設) で 0.555，汎用分類器 (本人以外) で 0.790，施設 B 入所者のそれは，個人専用分類器で 0.652，汎用分類器 (同施設) で 0.699，汎用分類器 (他施設) で 0.673，汎用分類器 (本人以外) で 0.753 となり，汎用分類器 (本人以外) の分類精度が最も高かった．また，学習に本人以外の入所者の支援記録を利用した分類器は，予兆やインシデントの少ない入所者の予兆検出に有効であることが分かった．

分類器の学習に付随情報を利用する実験では，インシデントの有無の情報を利用した場合の予兆分類の F 値が最も大きく，施設 A の入所者では文書表現ベクトルと付随情報ベクトルの割合が「6:10」の場合に全ラベルの平均 F 値が 0.825，施設 B の入所者では，ベクトルの割合が「8:8」の場合に全ラベルの平均 F 値が 0.772 だった．インシデントの有無の付随情報は分類精度の向上に有効である可能性がある．

今後は，両施設間の支援記録文章の違いを調査し，分類器の汎用性の向上に有効な文章の特徴の発見を目指す．付随情報については，有用な可能性のあるインシデントの有無などの付随情報の有効な利用方法を検討する．

文　献

[1] 上野史, 松本典久, 太田学, “福祉支援施設の支援記録からの BERT によるインシデントの予兆検出,”情報処理学会論文誌, Vol. 64, No. 11, pp. 1561–1566, 2023.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pretraining of deep bidirectional transformers for language understanding,” arXiv preprint-arXiv:1810.04805, 2018.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,”Advances in Neural Information Processing Systems 30, pp. 5998–6008, 2017.
[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” arXiv preprint arXiv:1804.07461, 2018.
[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,”arXiv preprint arXiv:1606.05250, 2016.
[6] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference,” arXiv preprint arXiv:1808.05326, 2018.
[7] 田中健次, 坂東幸一, 津本周作, 佐藤直樹, “医療界でのインシデント情報を活用したリスク対応,”横幹, Vol. 13, No. 2, pp. 84–90, 2019.
[8] 峯崎智裕, 松木萌, 井上創造, “介護サービス向上に向けた介護事故事例テキストの分析,”情報処理学会論文誌, Vol. 58, No. 10, pp. 1701–1711, 2017.
[9] 中沢信明, “高齢者介護施設における夜間徘徊者の検出,” 社会安全とプライバシー, Vol. 3,No. 2, pp. 8–12, 2019.
[10] 青木三重子, 撫中達司, 豊川絢也, 竹内銀河, “科学的介護を支援する介護 AI 記録分析ツール“けあらぽ”,”三菱電機技報, Vol. 95, No. 8, pp. 506–509, 2021.
[11] Diederik P Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,”arXiv preprint arXiv:1412.6980, 2014.
[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, Vol. 15, No. 1, pp. 1929–1958, 2014.

レム睡眠のリアルタイム自動検出に関する研究

佐野　史弥† 　国生　泰資† 　堀江　和正†† 　古木　淳也††† 　鈴木　陽子††††
阿部　高志†††† 　北川　博之††††

† 筑波大学大学院 理工情報生命学術院 システム情報工学研究群　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
†† 筑波大学 計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
††† 筑波大学大学院人間総合科学学術院人間総合科学研究群　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
†††† 筑波大学 国際統合睡眠医科学研究機構　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
E-mail: †{f.sano,kokusho}@kde.cs.tsukuba.ac.jp,
††horie@kde.cs.tsukuba.ac.jp,† † ††{suzuki.yoko.gz,abe.takashi.gp}@u.tsukuba.ac.jp,
† † ††kitagawa@cs.tsukuba.ac.jp

あらまし　本研究では，レム睡眠のリアルタイム自動検出の機械学習モデル及びシステムアーキテクチャの提案を行う．既存の手法では計測済の睡眠データに対してレム睡眠を後から検出する手法を用いている．そのため，睡眠の計測中にレム睡眠を検出することが不可能であった．しかし，睡眠研究において，レム睡眠の重要性が認識されており，レム睡眠時に音や匂いなどを刺激を与えるなど，リアルタイムで被験者のレム睡眠状態を検出したいという要求が存在する．そのため，本研究では，睡眠ステージ判定に移動窓を適用させ，かつレム睡眠の検出に特化するようにファインチューニングしたモデルを提案し，評価を行う．

キーワード　ヘルスケア，レム睡眠，機械学習

1　はじめに

睡眠は人間の生活において極めて重要な役割を果たし，全生活時間の約 1/3 を占める．健康や日々の活動に直接的な影響を与えるこの生理現象は，睡眠障害により労働効率の低下，生活習慣病 (例えば糖尿病)，疫力の低下などを引き起こす可能性がある．したがって，睡眠パターンを正確 に把握し，改善を促進することは非常に重要である．睡眠の評価には PSG(睡眠ポリグラフ検査) が用いられる．この検査では，被験者の脳波 (EEG)，眼電位 (EOG)，筋電位 (EMG)，血中酸素飽和度などの生体信号を一晩通して記録し， 睡眠サイクルを評価する．実際に測定された生体信号を図 1 に示す．測定された生体信号から睡眠状態を推定する方法として睡眠ステージ判定がある．睡眠は大きく，覚醒状態，レム睡眠，ノンレム睡眠に分けられる．米国睡眠学会によると，睡眠は 5 つのステージ (W: 覚醒，REM: レム睡眠，ノンレム睡眠:N1，N2，N3) に分類される [1]．これらのステージ判定は生体信号における波の周波数や波形などの特徴から判断される．

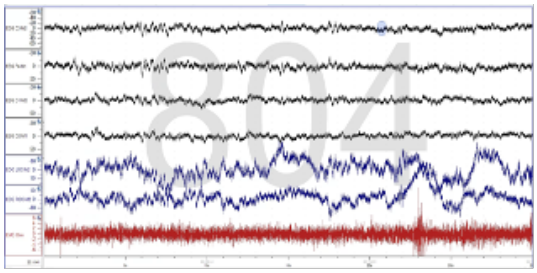


図 1: 実際に測定された生体信号

従来の睡眠ステージの判定は，生体信号を通常 30 秒単位で分割し，専門家が目視で各エポックにおける睡眠ステージを決定することが行われてきた．しかしながら一夜分のデータを分析する作業は専門的なスキルを要し，大きな労力と専門家に対する負担が伴う．この課題に対処するため，睡眠ステージの自動判定に関する研究が行われてきた．近年では，機械学習の進展により専門家と同等の精度を持つ判定モデルが提案されている．これを受け，近年では，睡眠中の被験者の睡眠状態をリアルタイムで分析したいというニーズも生じている．

本研究では具体的な応用先として，睡眠の基礎研究，特に，睡眠中の記憶や夢に関する介入研究に着目する．このような研究では睡眠中の被験者に対し，医師，技師がリアルタイムの判定を実行，レム睡眠中の被験者に対して刺激を与えるなどの実験が行われている．ここでは，このようなリアルタイムでのレム睡眠の自動検出 (レム睡眠の開始点，終了点の検出) を行う深層学習モデルの開発に取り組む．目標とする深層学習モデルは以下の条件を満たすようなモデルである．

- 確定レムの検出
  - レム睡眠において特徴波である急速眼球運動を伴わないトニックレムと特徴波を伴う確定レムがある．以下の三つの全ての条件を満たした時確定レムであるという．
    - K 複合あるいは睡眠紡錘波を伴わない LAMF が混在する EEG 活動
    - 顎 EMG 低緊張
    - 急速眼球運動 (REM)
  - 確定レムの際に刺激などを与えることが一般的である

ので，本研究では確定レムを検出する．これより先レム睡眠は確定レムを意味する．

- 誤検出が少ない
  - 非レム睡眠状態に刺激を与えることは良くないため，検出を頻繁に行うより誤検出を少なくさせることを重視．
- 誤差が少ない
  - レム睡眠開始時および終了時をできるだけ誤差なく判定することが求められる．

2 関連研究

リアルタイム性を重視した睡眠ステージ自動判定の関連研究として，Patanaik らは EEG1ch を用いたリアルタイムステージ判定システムをウェアラブル上で構築している.[2] 正答率は 0.83 と高精度を記録しているが，多チャンネルの生体信号に 対応できないという問題がある．

Koushik らは，信号計測と睡眠ステージの判定処理を別々のプロセスとして分離し，CFS 形式のファイルを介してデータをやり取りするアーキテクチャを提案している．このシステムでは約 5 ミリ秒ごとに 1 エポ ック分の睡眠ステージの判定が可能である. しかし，特定の時点でのステージ判定には，その直後の 5 エポック分のデータを反映させる必要があるため，実際にはリアルタイムでのステージ判定は 150s の遅延があるため行えない.[3]

3 提案手法

本研究の目的は，リアルタイムで高精度生体電位計測器 (Biosemi 社製 ActiveTwo) からデータを取得し，レム睡眠を検出するシステムアーキテクチャおよび，レム睡眠の検出モデル，モデルの適用方法の提案である．
対象の高精度脳波計は Biosemi 社製 ActiveTwo である．

3.1 システムアーキテクチャ

リアルタイムで生体電位測定器からデータを取得し，レム睡眠を検出するシステムアーキテクチャは現段階では，図 2 のようなアーキテクチャを想定している．リアルタイム判定を行うサーバと生体電位計測器から被験者の生体信号を常時計測しサーバに送信するクライアントを用意する．生体信号をクライアント側から送信し，サーバ側で受け取りリアルタイムにレム睡眠かの判定を計測中行い続ける．

図 2: 想定システムアーキテクチャ

3.2 移動窓を用いた基本判定モデル

基本判定モデルをリアルタイムに適用できるよう，移動窓を基本判定モデルに適用させた．

3.2.1 基本判定モデル

基本判定モデルは，本研究室によって提案された Sleep-CAM[4] である．Sleep-CAM は CNN ベースのモデルであり，図 3 に示すような特徴抽出層とステージ割り当て層で構成されている．このモデルは 30 秒の生体信号に対してステージ判定を行う．

入力は脳波 $EEG \times 6$，眼電位 $EOG \times 2$，顎筋電位 $EMG \times 1$ の 9 つの信号からなる．入力信号のサンプリング周波数は EEG と EMG は 200Hz，EOG は 100Hz である．

使用した重みは高精度脳波計 ActiveTwo で計測された 36 件のデータセットと簡易脳波計 PSG1100 で計測された 49 件のデータセットの計 75 件でトレーニングした重みである．ステージ判定精度はおよそ 80%と高精度である．しかし，30s 単位で判定をおこなうため，リアルタイムに適していないという問題がある．それを解決する方法を次節で説明する．

3.2.2 移 動 窓

リアルタイム性を重視するためのアプローチとして移動窓を用いた．移動窓を使用することで，図 4 のように入力窓を任意の時間でずらし，これを判定間隔と捉る．入力窓を 1s ずつずらすことで，基本判定モデルの問題点であった 30s ごとの判定を任意の短い時間で判定することを可能とさせる．入力窓 30 秒のステージ判定を行うそのステージ判定を「レム睡眠（REM）」か「非レム睡眠（W，N1，N2，N3）」に分け出力する．しかし，リアルタイムでレム睡眠を検出するには入力の最新の時点が「レム睡眠」か「非レム睡眠」かを判定することが求められるため判定の部分が適していない．判定の問題を解決する方法を次節で説明する．

3.3 ファインチューニングモデル

基本判定モデルとファインチューニングモデルの判定の違いを図 5 に示す．入力の最新の時点が「レム睡眠」か「非レム睡眠」かを判定するようにファインチューニングしたモデルである．基本判定モデルは入力窓に対する睡眠ステージ判定を行うモデルであった．この基本判定モデルの出力としては睡眠ステージの 5 値の出力となり，それを「レム睡眠」か「非レム睡眠 (Wake,N1,N2,N3)」としてレム睡眠か判定していた．そこで，図 5 のように入力窓の最新点が「レム睡眠」か「非レム睡眠」かを判定するモデルになるよう基本判定モデルのステージ割り当て層の Softmax のユニット数を 5 から 2 に変更しファインチューニングした．ファインチューニングモデルの構造は図 6 に示す．

3.4 重み層の導入

判定が最新点が「レム睡眠」か「非レム睡眠」かを判定するので最新点の重要性を強調することに重点を置いた．時間的な情報をより考慮させるためファインチューニングモデルのステージ割り当て層の GlobalAveragePooling 層の前に重みづけ

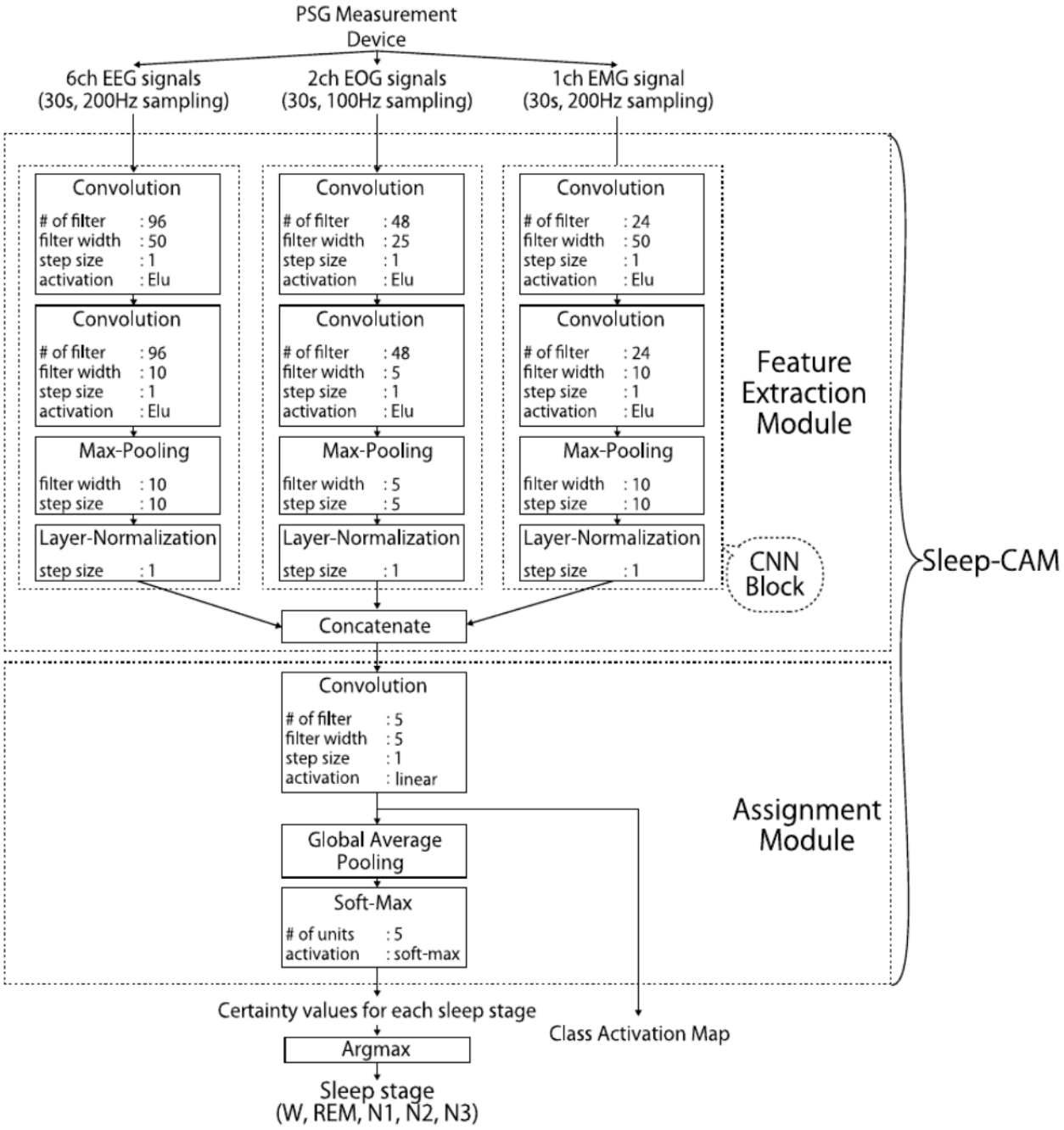


図 3: Sleep-CAM[4]

の層を導入した．時間的な情報を考慮するために線形重み層 (LinearWeight-Layer) とシグモイド重み層 (SigmoidWeight-Layer) を実装した．これらの層は，最新の状態ほど大きな重みを与えることで，最新の状態をより重視することが可能となる．

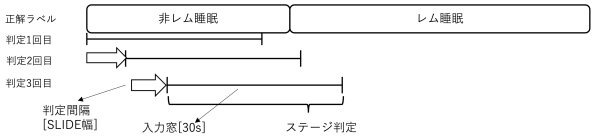


図 4: 移動窓

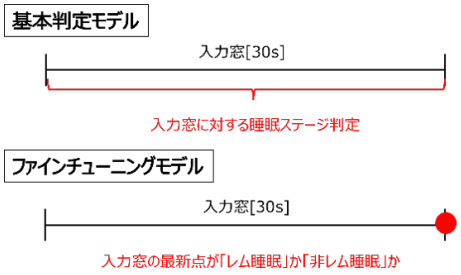


図 5: 判定の違い
(上) 基本判定モデルの判定
(下) ファインチューニングモデルの判定

3.4.1 線形重み層 (LinearWeight-Layer)

線形重み層は重み付けに最小値を 0，最大値 1 とする線形関数を使用し．入力の各点の重要性を線形的に調整する層である．この線形重み層は，図 7b のように入力の最初の点を重み 0 を割り当て，最新点の重みを 1 とし，時間の経過と共に重要度が線形的に増加するようになる．これにより，モデルは時系列の進行に伴って最新の情報をより重視し，古い情報の影響を徐々に薄めることが可能となる．

3.4.2 シグモイド重み層 (SigmoidWeight-Layer)

シグモイド重み層は重み付けにシグモイド関数を使用し．入力の各点の重要性を非線形的に調整する．シグモイド関数を用いることで、図 7a 入力の最新点に対しては重要性を高め、より古いデータ点に対してはその重要性を徐々に低下させるような重み付けを行う．

このアプローチにより，モデルは最新のデータ点をより重視しつつ，入力データを通じた情報の流れを捉える能力を保持できる．

4 実　　験

移動窓のスライド幅を 1s とし，基本判定モデルとファインチューニングモデルの比較を行った．使用したデータセットは提案手法のファインチューニングで使用したデータセットと同じである．実験環境について，OS は Ubuntu LTS 16.04.07，実装は Python3.8.0,TensorFlow 2.3.0 を使用した．また，CPU については Intel(R) Xeon(R) Gold 5122 CPU @3.60GHz，GPU については Nvidia Quadro P6000 を使用した．

4.1 データセット

実験に用いるデータセットは，計高精度生体電位計測器 ActiveTwo で測定され，専門家がレム睡眠の正解ラベルがついた 5 夜分のデータである．5 夜分の合計長は 41 時間 5 分であり，うちレム睡眠の時間は 9 時間 47 分 10 秒であった．各データの睡眠時間，レム睡眠の時間，非レム睡眠時間を表 1 に示す．

なおこのデータセットの前処理として，簡易脳波計 PSG1100 の電極間の電位差のチャネルとなるよう ActiveTwo の単一チャネルを用いて PSG1100 と同様の電極間の電位差のチャネルになるように変換し，ダウンサンプリング，ノイズ除去，正規化が行われている．

表 1: データセットの詳細

データ名	合計時間 [s]	レム睡眠 [s]	非レム睡眠 [s]
Data1	29671	5577	24123
Data2	29431	7034	22426
Data3	29731	6634	23126
Data4	29611	9451	20189
Data5	29311	7034	22306

4.2 ファインチューニング

異なる時間窓 (15s 及び 30s) を入力として用い，基本判定モデルをそのまま用いたモデル，線形重み層を追加したモデル．シグモイド層を追加したモデルの 3 つの異なるモデルに対してファインチューニングを行なった．全てのファインチューニングモデルにおいて図 6 に示されるようにファインチューニングの対象として基本判定モデルのステージ割り当て層の畳み込み層のみとし，特徴抽出層は基本判定モデルで重みを固定した．

モデルの汎化能力の評価を行うため，5 分割交差検証 (5-fold cross-validation) を実施した．各交差検証のサイクルで 1 件のデータがテストデータとして，別の 1 件が検証データとして，残りの 3 件が訓練データとして使用した．このアプローチにより，各データ点がモデル評価のために均等にテストデータとして利用され，同時にモデルの訓練と検証にも等しく寄与させた．モデルの学習において，学習率は 5.0×10^{-6} とし，損失関数はバイナリクロスエントロピーを，最適化アルゴリズムとして Adam を用いた．バッチサイズは 32 である．

4.3 モデルの精度比較

4.3.1 実 験 方 法

リアルタイムな判定を想定し，入力窓を 1s ずつずらし，レム睡眠かどうかの判定を行う．モデルの評価指標は，分類問題で用いられる，Accrucacy(正解率)，Precision(適合率)，Sensitivity(感度)，F1-score とした．Accuracy はモデルの全体的な正しさを表し，Sensitivity はレム睡眠を正しく識別するモデルの能力の指標であり，Precision はモデルが検出したレム睡眠が実際に正しかった割合を，F1-score は精度と感度の調和平均で，両者のバランスを取った単一スコアである．

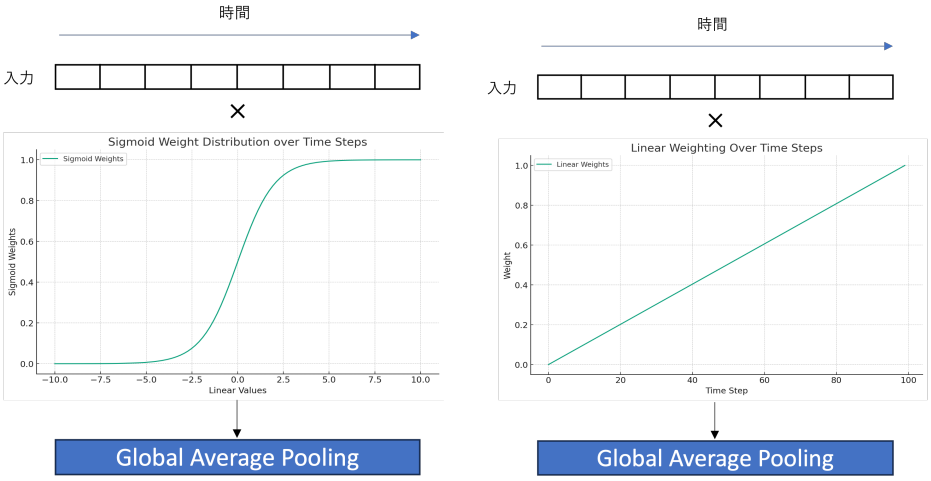
4.3.2 実 験 結 果

実験結果を表 3 に示す．表 3 におけるモデル状態はファインチューニング前後かを表し，モデルバリエーションにおいて，Normal は基本判定モデルをそのまま用いたモデル，LinearWeight は線形重み層を追加したモデル，SigmoidWeight はシグモイド重み層を追加したモデルである．正解率において全てのモデルで 90%を超えているため高精度であることがいえる．また，ファインチューニング後のモデルにおいて，入力 30s で線形重み層を追加したモデルが最も F1score が高くなっている．

ファインチューニング後の全てのモデルにおいて，ファインチューニング前のモデルと比較して，F1-scor が低下してしまっていることが分かる．しかし，感度と適合率を見てみると，感度が低下し，適合率が向上していることがわかる．これは，ファインチューニング前より，レム睡眠を見逃す確率は高くなってしまったが，誤って非レム睡眠をレム睡眠と識別することが少なくなったことを示唆している．これより，ファインチューニングを行ったことでモデルは全体的な評価としてはファインチューニング前より劣ってしまうが，レム睡眠の判定がより慎重になり，正確にレム睡眠と判定できるようになり誤検出が少なったと考察できる．

次に一夜分の正解ラベルと判定結果を図 9 に示す．これはヒプノグラムという睡眠中の異なる睡眠段階を可視化したグラフであり，横軸は時間，縦軸は睡眠段階 (「レム睡眠」,「非レム睡眠」) である．

図 6: ファインチューニングモデル



(a) シグモイド重み層　　(b) 線形重み層

図 7: 重み層 (a) シグモイド重み層, (b) 線形重み層

図 9 に示される結果から，30 秒の入力窓を持つモデルにファインチューニングを施した結果，より細かいレム睡眠段階を識別できるようになったことが明らかとなる．これは，ファインチューニングによりモデルのレム睡眠検出能力が向上し，睡眠サイクル中の短時間のレム睡眠も捉えられるようになったことを示唆している．一方で，15 秒の入力窓を用いたモデルでは，レム睡眠と非レム睡眠の区別において一定の問題が確認された．特に，短期間のレム睡眠の識別が不確かとなり，レム睡眠と非レム睡眠の間での分類が短時間で不安定に振動する傾向が観察された．

4.4 レム睡眠の開始点及び終了点の検出率

4.4.1 実験方法

正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるか及びモデルによる判定の開始点と終了点から、判定を基準として正解ラベルが許容誤差内に存在するかを評価した．今回の実験において許容誤差は 30s とした．正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるかについて図 8 を用いて説明する．図 8 はヒプノグラム (睡眠遷移図) の一部を切り抜いたものである．上のヒプノグラムが正解ラベル，下のヒプノグラムがモデルの判定である．正解ラベルから前後 30s 以内に開始点または終了点があれば，その正解ラベルの開始点または終了点を検出できたとみなす．同様に基準を逆にしモデルによる判定から前後 30 秒以内に正解ラベルが存在するかを検出した．

正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるかの評価指標として，感度 (Sensitivity), 検出回数 (Det.Count), 正解ラベルからの平均時間誤差 (Average)，標準偏差 (Std) を用いる．感度は，モデルが正解ラベルのレム睡眠の開始点及び終了点をどれだけ正確に検出できているのかを示す．判定のレム睡眠の開始点と終了点を基準として，許容範囲内に正解ラベルがあるかどうかの判定の評価指標として，適合率 (Precision)，判定回数 (Count)，前後 30 秒以内に正解ラベルが含まれている判定回数 (Count<30s)，判定から正解ラベルの平均時間誤差 (Average)，標準偏差 (Std) を用いる．適合率は，レム睡眠の開始点及び終了点の検出においてモデルが正解ラベルと一致しているかを示す．

今回用いたデータセットの各データの開始点及び，終了点の回数を表 2 に示す．データセットにおける総開始点は 279 回あり，総終了点は 277 回であり，1 データあたりの平均開始点および終了点の回数は 55.8 回，55.4 回であった．

表 2: データセットにおける開始点及び終了点の回数

	開始点 [回]	終了点 [回]
Data1	60	59
Data2	37	37
Data3	65	65
Data4	64	64
Data5	53	52
Average	55.8	55.4

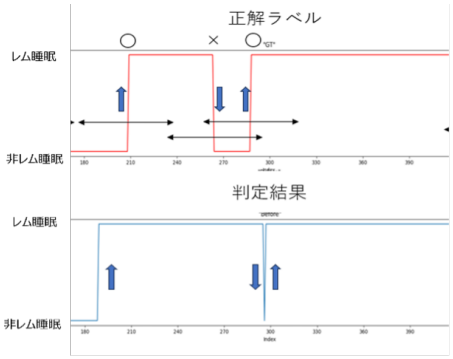


図 8: 判定方法

4.4.2 実験結果

実験結果を表 4，6，5，7 に示す．正解ラベルを基準とした結果から，開始点，終了点共に，ファインチューニングを行ったことで感度は向上していることがわかり，特に 15s の Sigmoid-Weight が 0.9426,0.9458 と高い値を示した．また入力窓を狭くすることで平均誤差も短くなり標準偏差も小さくなっていることが分かる．一方で，判定基準の結果を見てみると，ファインチューニングを行ったことで，開始点及び終了点の判定回数が増加していることがわかる．これは短期間のレム睡眠の識別が不確かとなり，レム睡眠と非レム睡眠の間での分類が短時間で不安定に振動する傾向が要因として考えられる．それにより，開始点と終了点共に適合率が低下してしまっていることがわかった．これはモデルがレム睡眠に過剰に反応してしまっているということが考えられる．

5 まとめ

本研究では，リアルタイムにレム睡眠を検出するようなモデルを提案した．目標とする深層学習モデルに対して今回の結果は以下のようにまとめられる．

- 確定レムの検出

  ファインチューニングを行ったことで確定レムの全体的な検出はあまり向上しなかったが，より細かな確定レムは捉えることができるようになった

- 誤検出が少ない

  ファインチューニングを行ったことで，全体的なレム睡眠全誤検出は少なくなった．ただし，開始点と終了点における誤検出は向上しなかった．

- 誤差が少ない

  開始点と終了点の誤差はファインチューニングを行うことで改善された．

今後の課題として，以下が挙げられる．

- レム睡眠の開始点と終了点の検出精度のさらなる向上及びモデルの評価方法

- モデル判定におけるレム睡眠と非レム睡眠の間での短時間で不安定な振動の抑制
- レム睡眠を検出するシステムアーキテクチャの設計及び実装

6 謝　　辞

本研究の一部は科研費: 基盤研究 C(JP20K03493), 挑戦的研究 (萌芽)(JP22K19802), 基盤研究 B(JP23H03399), 若手研究 (JP23K16988), AMED ムーンショット型研究開発事業 (JP21zf0127005) による.

文　　献

[1] American Academy of SLEEP MEDICINE, The AASM Manual for the Scoring of Sleep and Associated Events, “AASM Manual for the Scoring of Sleep and Associated Events”, (株) ライフサイエンス，2018.

[2] Amiy Patanaik, Ju Lynn Ong, Joshua J Gooley, So-nia Ancoli-Israel and Michael W L Chee, “An end-to-end framework for real-time automatic sleep stage classification,Sleep“, Vol. 41, No. 5, 2018.

[3] Abhay Koushik, Judith Amores, Pattie Maes, “Real-Time Sleep Staging using Deep Learning on a Smartphone fora Wearable EEG“, IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN),2019.

[4] Kazumasa Horie, Leo Ota, Ryusuke Miyamoto, TakashiAbe, Yoko Suzuki, Fusae Kawana, Toshio Kokubo, MasashiYanagisawa and Hiroyuki Kitagawa, “Automated sleep stagescoring employing a reasoning mechanism and evaluation ofits explainability“, Scientific Reports 12, No. 12799, 2022

表 3: ファインチューニング前後のモデル性能比較

モデル状態	モデルバリエーション	input window	Accuracy	Sensitivity	Precision	F1-score
ファインチューニング前	N/A	30s	0.9233	0.967	0.773	0.859
ファインチューニング後	Normal	15s	0.9161	0.845	0.815	0.830
		30s	0.9222	0.845	0.835	0.840
	LinearWeight	15s	0.9170	0.842	0.820	0.831
		30s	0.9213	0.861	0.822	0.841
	SigmoidWeight	15s	0.9108	0.817	0.814	0.816
		30s	0.9170	0.850	0.815	0.832

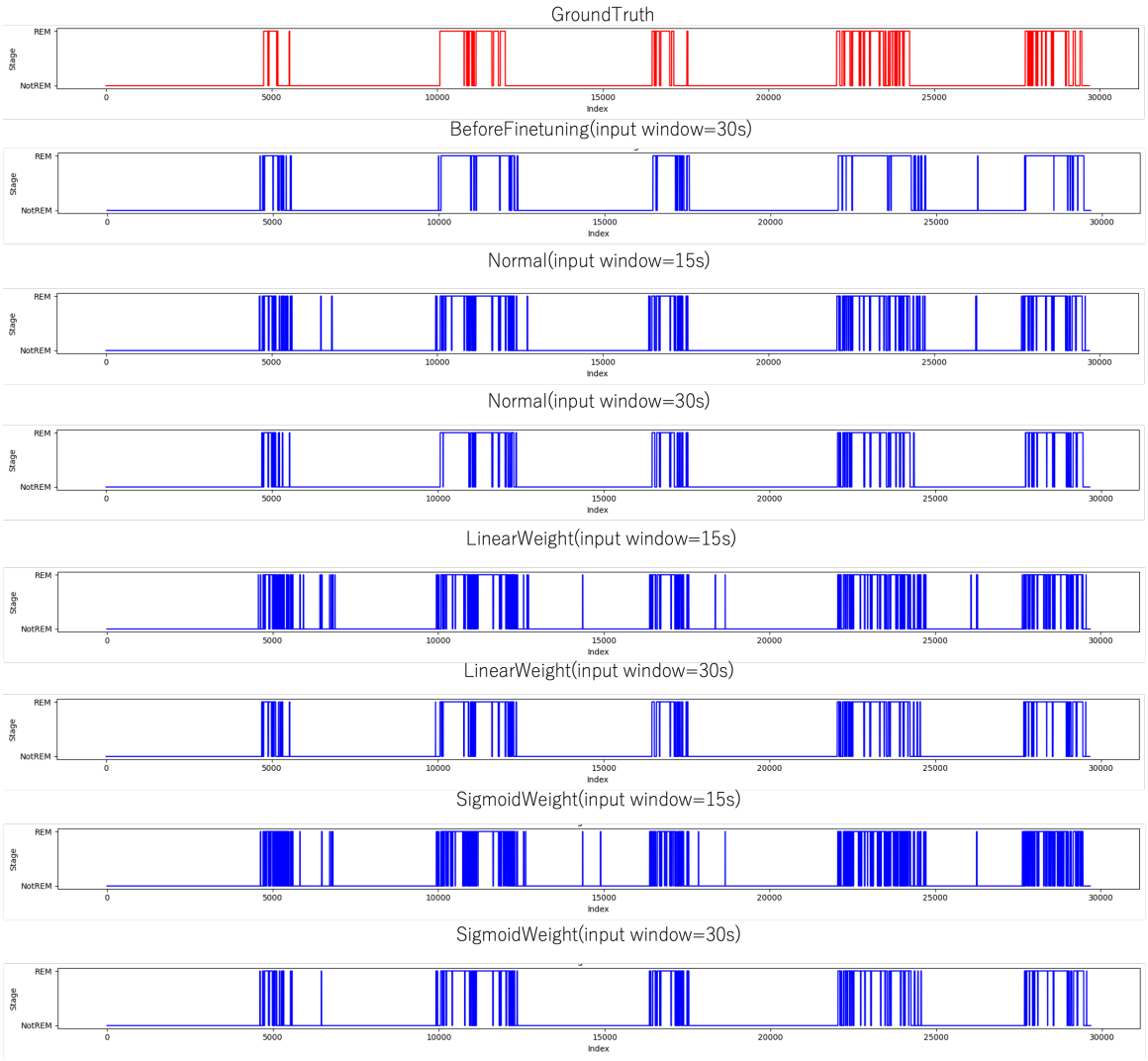


図 9: ヒプノグラム (睡眠遷移図)

表 4: 正解ラベルを基準としたレム睡眠の開始点における結果

モデル状態	モデルバリエーション	input window	Sensitivity	Det.Count	Average	Std
ファインチューニング前	N/A	30s	0.4767	133	+1.40	14.55
ファインチューニング後	Normal	15s	0.8781	245	-0.61	10.31
		30s	0.6308	176	+4.92	13.88
	LinearWeight	15s	0.8924	249	-0.90	9.88
		30s	0.6953	194	+0.45	14.14
	SigmoidWeight	15s	0.9426	263	-0.77	9.46
		30s	0.7311	204	-0.05	12.71

表 5: 判定基準としたレム睡眠の開始点における結果

モデル状態	モデルバリエーション	input window	Precision	Count	Count<30s	Average	Std
ファインチューニング前	N/A	30s	0.4120	381	157	-0.09	14.91
ファインチューニング後	Normal	15s	0.3011	1448	436	+0.75	15.15
		30s	0.3618	644	233	-2.45	15.37
	LinearWeight	15s	0.2995	1636	490	+1.56	15.27
		30s	0.3409	880	300	+2.71	15.22
	SigmoidWeight	15s	0.3015	1804	544	+0.11	15.84
		30s	0.3290	933	307	+2.19	15.57

表 6: 正解ラベルを基準としたレム睡眠の終了点におけるの結果

モデル状態	モデルバリエーション	input window	Det.Rate	Det.Count	Average	Std
ファインチューニング前	N/A	30s	0.5126	142	+9.40	12.61
ファインチューニング後	Normal	15s	0.9061	251	+0.67	8.46
		30s	0.6714	186	+4.78	12.31
	LinearWeight	15s	0.9061	251	+0.13	7.19
		30s	0.7075	196	+2.58	10.15
	SigmoidWeight	15s	0.9458	262	-0.43	6.94
		30s	0.7509	208	+1.39	9.68

表 7: 判定基準としたレム睡眠の終了点における結果

モデル状態	モデルバリエーション	input window	Precision	Count	Count<30s	Average	Std
ファインチューニング前	N/A	30s	0.4120	381	157	-9.15	12.64
ファインチューニング後	Normal	15s	0.3035	1446	439	-0.37	13.36
		30s	0.3841	643	247	-3.61	13.30
	LinearWeight	15s	0.2990	1635	489	+1.54	13.68
		30s	0.3788	879	333	-0.35	13.65
	SigmoidWeight	15s	0.298	1802	537	+1.19	14.68
		30s	0.3401	932	932	-0.47	12.37

リアルタイム性を考慮した睡眠ステージ判定システムの開発

国生　泰資† 　堀江　和正†† 　鈴木　陽子††† 　阿部　高志††† 　北川　博之†††

† 筑波大学大学院 理工情報生命学術院 システム情報工学研究群　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
†† 筑波大学 計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
††† 筑波大学 国際統合睡眠医科学研究機構　〒 305–8573 茨城県つくば市天王台 1 丁目 1-1
E-mail: †{kokusho,horie}@kde.cs.tsukuba.ac.jp, ††{suzuki.yoko.gz,abe.takashi.gp}@u.tsukuba.ac.jp,
†††kitagawa@cs.tsukuba.ac.jp

あらまし　睡眠ステージ判定とは被験者の脳波や筋電位などの生体信号に現れる特徴をもとに，一定間隔で睡眠状態をラベル付していく作業であり，被験者の睡眠の質やサイクルなどを評価することを可能とする．これまで睡眠ステージ判定とは被験者ごとに測定された睡眠データ (生体信号) に対して，専門技師の目視による 30s 間隔のラベル付が行われてきたが，このアナログ的作業は時間及び人的コストを多く要し，睡眠研究を促進していく上で課題となっていた．これに対し近年，様々な睡眠ステージ判定の自動化手法が研究され，特にニューラルネットワークを用いた判定手法が注目を集めている．一方，被験者の睡眠の質の向上や睡眠障害への治療といった睡眠研究においては，睡眠ステージをリアルタイムに判定し被験者の睡眠状態に応じたフィードバックを行う等のニーズが存在する．そこで本研究では睡眠中の患者の睡眠状態を把握できるようなリアルタイムなステージ判定手法を提案する．具体的には移動窓を用いた連続的なステージ判定手法について実験・評価を行う．

キーワード　ヘルスケア, 睡眠ステージ判定, 機械学習

1 はじめに

睡眠は人間の健康と活動に大きく関わっている．睡眠の質はヒトの日中活動のパフォーマンスに大きく影響し, 不眠症やナルコレプシーなどの睡眠障害を伴うことも珍しくない．したがって, 現在に至るまで睡眠に関する研究が盛んに行われてきた．睡眠の検査手法の一つに PSG(Polysomnography) 検査がある．PSG 検査を通して, 被験者の終夜にわたる脳波や眼電図, 筋電図などを記録することができ, そうして得られた生体信号を元に睡眠ステージ判定が行われる．従来のステージ判定手法は専門技師による波形の目視により行われ, 人的及び時間的コストが多くかかることが課題となっている．そのような課題に対し, 機械学習を用いた自動判定に関する研究が多くなされ, 近年ではニューラルネットワークモデルを中心とした判定モデルを用いて高い精度で判定することが可能となってきている．一方で, このような睡眠ステージ判定をリアルタイムに行いたいといったニーズが存在する．リアルタイムに睡眠ステージ判定を行うことは, 被験者の睡眠状態に対応した即座のフィードバックを可能にし, 睡眠障害の治療や睡眠の質の向上などを目指す睡眠研究への活用が期待されている．しかし, これまでに提案された睡眠ステージ判定の自動化に関する研究は測定済の生体信号に対しステージ判定を行うことを想定しており, 睡眠中の被験者の睡眠ステージをリアルタイムに判定する手法は未だ確立されていない．本研究では, 機械学習に基づくステージ判定モデルを用いたリアルタイムなステージ判定システムを提案する．実験では, リアルタイムシステムの処理時間の検証や, 従来の睡眠ステージ判定モデルと新たにリアルタイム用に学習した判定モデルの 2 つについて判定精度を検証する．

2 前提知識

2.1 PSG 検査

PSG 検査は, 睡眠脳波や筋電位, 呼吸サイクルなど睡眠状態に関わる生体信号の計測を目的に行われる．PSG 検査は終夜を通して行われ, 被験者には頭部や目元等の複数箇所に電極を取り付け, 生体信号を記録していく．こうして得られた 1 夜分にわたる生体信号の時系列データを元に睡眠のサイクルや睡眠中の異常などを分析することができる．

2.2 睡眠ステージ判定

PSG 検査により得られた生体信号を元に睡眠ステージ判定が可能となる．睡眠ステージは大きく分けて REM 睡眠と Non-REM 睡眠に別れる．REM 睡眠とは急速眼球運動 (Rapid Eye Movement) が伴う睡眠状態であり, 脳が活発に動き記憶の定着などが行われる．Non-REM 睡眠とは体, 脳共に休んでいる状態であり成長ホルモンの分泌や脳の休息が行われる. [3]

さらに Non-REM 睡眠には睡眠の深さが浅い場合と深い場合が存在し, 米国睡眠医学会 (AASM) が示すガイドライン [4] では Non-REM 睡眠を 3 段階に分割して定義している．ヒトは 1 夜を通してこれらのステージを往来しながら, 睡眠を行う．このような睡眠ステージは PSG 検査より取得した生体信号に現れる, 特徴波や振幅及び持続時間などの特徴を元に判別可能である．これらの判定基準についても, 米国睡眠医学会 (AASM) が判定基準を定義しており, その基準を元に専門技師が生体信号に対し, ステージをラベリングしていく．

REM ステージでは, 脳波信号 (EEG) の電位が低く 2-6Hz ののこぎり波が出現する. また, 筋電位 (EMG) では顎筋電位が低く, 眼球電図 (EOG) では急速眼球運動 (REMs) が発生する. Non-REM1 ステージでは, EEG では頭蓋頂鋭波が発生し, EOG では低速眼球運動が見られる. Non-REM2 ステージでは, EEG にて K-complex 波や紡錘波が発生する. Non-REM3 ステージでは, EEG にて 0.2-5Hz で高振幅なデルタ波が発生する. [4]

睡眠ステージ	特徴
Wake	α波(8-13Hz), まばたき, 急速眼球運動
REM	急速眼球運動, のこぎり波, 低顎筋電位
N1	低速眼球運動, 頭蓋頂鋭波
N2	K-complex波, 紡錘波
N3	低周波数波の活発化(0.5-20Hz)

図 1　睡眠ステージの判定基準となる生体信号の特徴例

一般にこのラベリングはエポックと呼ばれる通常 30s 幅の時間帯の生体信号に対し 1 ステージを付与する形式で行われる. ラベリング作業は睡眠検査技師による波形の目視を必要とし, 被験者一人の 1 夜分の記録は約 8 時間, すなわち約 1000 エポックほどの長さにもわたり複数夜分の記録に対し睡眠ステージ判定を行うことはかなりの労力と時間を要することがわかる.

このような課題に対して, 睡眠ステージ判定の自動化についての研究が多くなされている.

3 関 連 研 究

3.1 ニューラルネットワークを用いた睡眠ステージ判定手法

現在に至るまでさまざまなヒトを対象とした睡眠ステージ判定モデルが提案されてきたが, その中でもニューラルネットワークを用いたモデルが高い精度を記録しており, 注目が集まっている. Tsinaliset ら [8] により 2 層の畳み込み層を用いた CNN モデルが発表されている. しかし, 1ch 分の EEG のみを入力とする簡素なモデルであるため正答率 0.71-0.76 にとどまっている. Supratak ら [9] により提案された DeepSleepNet は CNN と Bi-directional LSTM を組み合わせたモデルを用い, 前後情報を考慮した推測を可能とした. 結果として正答率は 0.81-0.86 を記録しているが, ステージの判定に 25 エポック分の情報が必要となる. Phan ら [10] により提案された SeqSleepNet についても正答率 0.87 を記録しており, 高精度であるが, 10-30 エポックの情報を必要とする. ヒト (専門技師) によるステージ判定の正答率としては, 複数の専門技師間での判定結果の一致率である検者間信頼性が目安となる指標であり, その値は約 80% [5] ほどであるとされている. これらのことより, 上記の研究はヒトと同等位以上の判定精度を達成できているといえるが, 実際には被験者の年齢や睡眠障害の有無等により判定精度がバラついたり, 深層学習であるため判定の根拠がブラックボックス化されているなどの課題は未だ残っている.

3.2 Sleep-CAM

Sleep-CAM モデル [13] とは, 本研究グループ内にて開発された深層学習を用いたステージ判定モデルであり, 判定根拠のブラックボックス化という課題に対し, その CAM 層 (Class Activation Map) を用いて判定根拠を提示することができるモデルである. そのアーキテクチャについて図 2 に示す. 脳波信号 (EEG), 眼電図 (EOG), 顎筋電位 (EMG) に対しそれぞれ 2 層の畳み込み層を用いて, 各信号に現れるステージを判定するための特徴量を抽出する (Feature Extraction Module). 得られた特徴量に対し, 5 つのステージに対する確信度を出力すると共に CAM を用いて, 30s エポック内の各時点での 5 ステージ確信度も出力する (Assignment Module). CAM の出力を見ることで 30s エポック内のどの時点が判定に対し強い影響を及ぼしているか確認することができる.

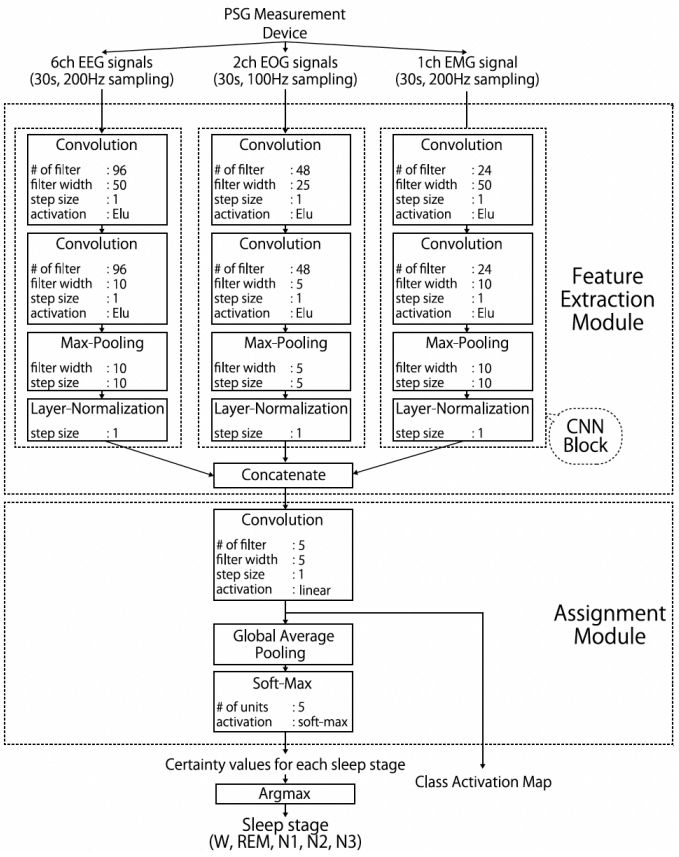


図 2　Sleep-CAM のモデルアーキテクチャ [13] より

3.3 リアルタイムなステージ判定手法

ステージ判定の自動化に関する研究は数多くなされてきたが, リアルタイム性を考慮したステージ判定に関する研究はまだ少ない. Koushik ら [11] は信号計測と睡眠ステージ判定処理それぞれの機能をプロセスレベルで分離し, 両者間のデータ転送に CFS 形式のファイルを使用したアーキテクチャを提案しており, およそ 5ms 毎に 1 エポック分のステージ判定を下すことができる. しかし, ある時点でのステージ判定には直後 5 エポック分のステージ判定結果を反映させるモデルであるため, 現実的にはリアルタイムにステージ判定を行うことはでき

ない. Patanaik [12] は 1ch の EEG を用いたリアルタイムなステージ判定システムをスマートフォン上で構築している. 1ch の EEG のみであるが正答率は 0.83 と高い精度を記録している. しかし, 多チャンネルの信号に対応できないという問題点があり, 頑健性に疑問が残る.

4 提案手法

4.1 リアルタイム判定システムの概要

リアルタイム性を考慮した自動睡眠ステージ判定システムを開発する上で, "PSG 計測機器からデジタル信号を受け取り, ステージ判定をフィードバックするシステムの設計"と"機械学習モデルを用いてリアルタイムにステージ判定を行う判定する仕組み"の 2 つを考慮する必要がある.

4.2 PSG 計測機器と判定システムとのリアルタイム通信

一般的に PSG 検査は被験者の各部に電極を取り付け, 得られた電気アナログ信号をデジタル変換し, USB 等の有線を伝ってユーザ PC へと送信される. ユーザ PC では計測器専用のソフトウェアが付属されており, そのソフトウェアを通して, 生体信号のモニタリングやディスク保存などを行えるようになっている. こうした環境に対して, ステージ判定システムを組み込むことを考えたとき PSG 計測機器から直接信号を受け取ることは計測器の組み込みプログラムを修正する必要があり, 計測器の種類ごとにその操作を行うことは現実的ではない. また, SoC(separation of conerns) に基づき, リアルタイムな生体信号を送信する機能とステージ判定を行う機能を分離させ, 様々な計測機器に対応しやすいシステムを目指した. 以上を踏まえ, ユーザ PC 上のソフトウェアをサーバとし, ユーザ PC 上またはインターネット上の別ノード (クライアント) に対し随時計測した生体信号を送信するようなサーバクライアントシステムを採用した.

ユーザ PC 上のソフトウェアでは PSG 計測器から送信され, USB や WiFi 等を通して受け取った生体信号をクライアントプログラムへと TCP/IP を用いて送信する. クライアントプログラムではサーバから送られてくるパケットを遅延なく受け取り, リアルタイムにステージ判定をおこない, その結果をサーバプログラムまたは任意のノードへ送信する.

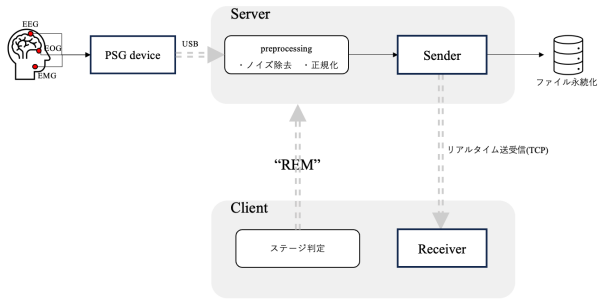


図 3 PSG 計測器からステージ判定結果提示までのアーキテクチャ図

4.3 リアルタイム判定を行う仕組み

本研究ではこれまで開発されてきた睡眠ステージ自動判定モデルを活用したリアルタイム判定手法を目指した. Sleep-CAM では入力サイズは 1 エポックであり, 出力は入力エポックに対する 5 ステージの確信度となる. このモデルをリアルタイムに拡張するため, エポック長の移動窓を用意する. この移動窓には随時最新の生体信号が追加され, 直近 30s までの生体信号を保存しておく. その移動窓に対し, あらかじめ設定した判定間隔ごとにステージ判定をおこなう. 以下に判定間隔を slide[s] と設定した場合のアルゴリズムを図 4.3 に示す. サーバから計測信号を受け取り, 判定に必要な 1 エポック (30s) まで受け取ると判定結果をサーバ等へ送信する. それ以後もパケットを受け取り続け, スライド幅 (slide[s]) ごとにステージ判定を行うことを繰り返す.

Algorithm 1 Mechanism of realtime sleep scoring in client program

```
S = slide[s], t = 0.0[s]
socket.connect()
data = []
while socket.isalive() == True do
    r = socket.receive(length = 1s)
    data = data.append(r)
    t = t + 1.0
    if t%S == 0 & data.length == 30.0s then
        stage = model.predict(data)
        socket.send(stage)
        data = data[t − 30 + S :]
    end if
end while
```

図 4 スライド窓を用いた判定手法

4.4 リアルタイム判定モデル

4.4.1 Sleep-CAM に基づくモデル

4.3 節で述べた通り, 今回設計した判定システムでは既存の自動判定モデルを組み込むことができる. Sleep-CAM モデル [13] は 1 エポックの入力に対し, 当該エポックのステージ判定を行うことができる. 本研究では, 従来の 30s エポックに対するステージ判定モデルを直接的にリアルタイム判定に適用してものとして, Sleep-CAM モデルを採用した自動判定システムを開発し, そのリアルタイム性を検証した.

4.4.2 Realtime-Sleep-CAM モデル

上に述べた Sleep-CAM モデルは, 従来の 30s のエポックを単位としてラベル付けされた正解データを学習データとして用いている. したがって, このデータでは, エポックの途中で睡眠状態の変化が起きているような場合に, その変化が正確には記述できていない. そこで, 1 秒単位でステージラベルが付与された正解ラベルを用意し, このラベルを用いて Sleep-CAM モデルを学習させた Realtime-Sleep-CAM モデルを作成した.

ただし, 1 秒単位の正解ラベルの作成には多大な時間を要するためステージラベルは”Wake または N1”, ”REM”, ”N2 または N3”の 3 ステージとした. さらに検証のため, モデルへの入力サイズを 5s, 15s, 30s とした場合についてそれぞれ学習を行い, 実験をおこなった. 30s エポックごとにしか存在しない従来の正解エポックでは 30s より短い区間を入力とする判定モデルを作成することは難しかったが, 1s 刻みの正解ラベルを用意したことで入力サイズを短くすることが可能となった.

5 実　　験

5.1 実験の概要と目的

提案手法の評価のため, 実際の PSG 検査を想定したリアルタイム判定を行う. この実験では, リアルタイム性の評価と判定精度の評価を目的としている.

実験 5.3.1 では, 一定間隔でサーバより送信される生体信号に対し遅延なく判定が行えているかを検証する目的で, 実際の PSG 検査を想定したリアルタイム測定実験を行う.

一方, 判定モデルの精度検証については 30s エポック単位での判定精度と移動窓を用いたリアルタイムでのステージ判定精度の両方を考慮する必要がある. 実験 5.3.2 では, 30s エポックに対する Sleep-CAM モデルおよび Realtime-Sleep-CAM モデルの判定精度を検証する. この指標を用いて, 作成したモデルの精度を関連研究などと比較することができる. 本研究の主題であるリアルタイム判定 (30s より短い移動幅) での精度の検証については, 実験 5.3.4 で行う. 実験 5.3.3 では, Realtime-Sleep-CAM モデルについて, ステージ遷移の有無を区別したエポックについての判定精度, 及びその影響について検証する. 実験 5.3.4 では, 実際に Sleep-CAM モデルと Realtime-Sleep-CAM モデルを用いてリアルタイム判定を行い, 今回新たに用意した 1s 刻みの正解ラベルを用いてその判定精度を検証し, リアルタイム判定において Realtime-Sleep-CAM モデルに優位性があるかを検討する.

5.2 データセット

今回, モデルの学習及び精度検証では PSG1100(日本光電) を用いて測定されたオリジナルデータセットを用いる. 計測は 20〜30 歳の健常者を対象に行われ, サンプル数は 49 夜分 (53630 エポック) となる. 取得したデータのうち, 脳波信号 (EEG) を 6ch(F3-M2, F4-M1, C3-M2, C4-M1, O2-M1, O1-M2), 眼球運動信号 (EOG) を 2ch(E1-M2, E2-M2), 顎筋電信号 (EMG) を 1ch(EMG1-Z) をモデルへの入力として使用する.

	SleepCAM
# of Night	49
# of epoch	53630
Stage	W, N1, N2, N3 REM
channel	EEGx6, EOGx2, EMGx1

図 5　Sleep-CAM モデルの学習に用いるデータセットの概要

		INPUT = 5s	INPUT = 15s	INPUT = 30s
# of Night			5	
# of epoch	Overall		28749(original=511)	
	without transition	27938	26206	24555
	with transition	811	2543	4194
Stage			WorN1, N2orN3, REM	
channel			EEGx6, EOGx2, EMGx1	

図 7　Realtime-Sleep-CAM モデルの学習に用いるデータセットの概要

さらにこれら信号データセットに対し, 検者間信頼性 92.4% の専門技師による AASM ルールに基づいた全エポックに対するラベル付が行われた. Sleep-CAM モデルの学習ではこのエポック単位のラベルを用いて学習を行う.

一方, 49 夜分のうち 5 夜分について専門技師による 1 秒単位のラベルデータを新たに用意し, これらを用いて Realtime-Sleep-CAM モデルの学習をおこなう. 1s 単位のラベルデータから学習およびテストデータを作成するために Realtime-Sleep-CAM モデルの入力サイズに合わせたエポックへと切り出しを行う必要がある. 1 秒単位のラベルデータから入力サイズに合わせた正解ラベルを作成する手順を図 6 に示す.

エポックフリーなラベルから一定間隔で予測モデルの入力サイズ長のエポックをサンプリングするとき, ステージ遷移にまたがるエポックについては 2 つ以上のステージラベルが混在する状態となる. これについては, エポックに占める割合が大きいステージラベルを正解ラベルとしてサンプリングした.

さらに, 5 夜分のラベルデータでは学習が不十分だと考えられるため, 位置を 5 秒, 10 秒, 15 秒とずらした正解ラベルを用意し, データ拡張を行った.

また, Realtime-Sleep-CAM への入力エポックのサイズについては 5s, 15s, 30s の場合についてそれぞれ学習をおこなった. それに伴い, 設定した入力サイズ, 5s, 15s, 30s ごとに学習, テストデータを上記の手順で作成した.

5.3 実 験 結 果

5.3.1 リアルタイム性の評価

リアルタイム性の評価では Sleep-CAM モデルを使用し, サーバからリアルタイムで送信されてくる信号データに対し, ステージ判定を遅延なく行えているか検証する. しかし, 実際に被験者を用意し実験することは難しいため, あらかじめ計測済みの信号データを 1 夜分用意し, リアルタイム計測と同様な速度でクライアントプログラムへ信号データを送信するようにして実験をおこなった. また, 実験条件として判定間隔であるスライド幅を 1s, 5s, 30s の場合でそれぞれ実験し計測を行った.

クライアントプログラムではスライド幅分の信号データをパケットで受け取り, ステージ予測を行い予測結果を送信することを繰り返す. そこで, パケット受け取りにかかる時間を TCP Time, ステージ予測と予測結果の送信にかかる時間を Scoring Time, TCP Time と Scoring Time を合わせたステージの判定間隔に相当する時間を Iteration Time としてそれぞれ計測した. 実験に用いた 1 夜分のデータ長は約 9 時間 6 分にわたり, 30s 幅では 1093 回, 15s 幅では 2186 回 1s 幅では 6558 回の TCP Time 及び Scoring Time の計測が行われ, それらの平

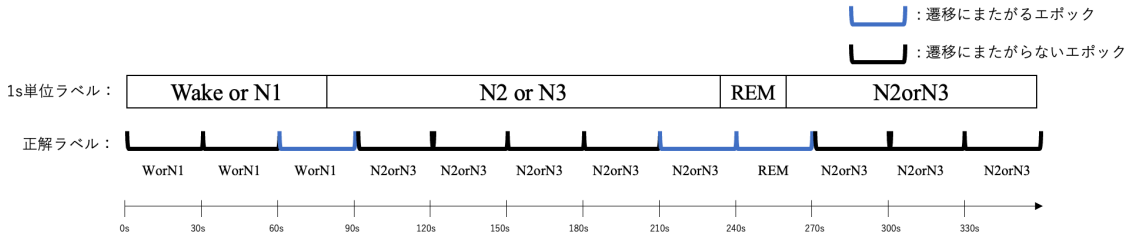


図 6　入力サイズ=30s の場合の, 1 秒単位のラベルデータを元にした正解データの作成

均と標準偏差を算出した．実験に用いた OS は Ubuntu LTS 20.0.4 を使用し, 実装は Python の Tensorflow-gpu-2.1.0 を使用した．また, CPU は Intel Core i3-12100, GPU は Nvidia GeForce GTX 1660 SUPER を使用した.

実験結果を図 8 に示す．結果を見てみるとまずどのスライド幅においても, Iteration Time は判定間隔がそれぞれ 30s, 5s, 1s のとき, 30.128, 4.998, 0.978 となっているため, 遅延なく判定が行えていることがわかる．さらにその内訳を見てみると, ステージ判定にかかる時間 Scoring Time は比較的小さく, TCP パケットを受け取る時間 TCP Time が支配的であることがわかる．TCP Time はサーバプログラムから送られてくる, 設定したスライド幅分のパケットを受け取るまでの時間に相当することから, スライド幅分に近い時間を要する．Scoring Time については, どのスライド幅においても約 0.6s の判定時間を必要としていることがわかる．この時間は判定間隔の下限であり, 0.6s より短い判定間隔でのリアルタイム判定は行えないことがわかる.

これらのことから, 最大で判定間隔 1s のリアルタイムなステージ判定が遅延なく行えていることがわかる.

	Iteration		TCP		Scoring	
	Mean	Std	Mean	Std	Mean	Std
SLIDE = 30s	30.128	0.122	29.512	0.085	0.618	0.072
SLIDE = 5s	4.998	0.058	4.408	0.043	0.631	0.042
SLIDE = 1s	0.978	0.046	0.389	0.033	0.619	0.041

図 8　リアルタイム性の評価結果

5. 3. 2　モデル単体での判定精度

Sleep-CAM モデルと Realtime-Sleep-CAM モデルの学習をおこなった．Sleep-CAM モデルでは 49 夜分のレコード (53630 エポック) に対し, 7 分割交差検証をおこなった. 正解ステージ数は 5 つで, 'Wake', 'REM', 'N1', 'N2', 'N3' である．損失関数はクロスエントロピー関数とし, 最適化手法は Adam(lr=5×10^{-6}) を用いた.Realtime-Sleep-CAM モデルではデータ拡張をおこなった 5 夜分 (28749 エポック) のレコードに対し, 5 分割交差検証をおこなった．正解ステージ数は 3 つで, 'Wake または N1', 'REM', 'N2 または N3' である．損失関数及び最適化手法は Sleep-CAM モデルと同様である.

入力サイズ長のエポックに対する判定精度は, より細かな粒度でのリアルタイムな判定精度を直接評価することはできないが, 関連研究と比較することでモデルの性能を判断することができる．リアルタイムな判定精度については, 5. 3. 3 節で述べる.

まず Sleep-CAM モデルについて正答率 (Accurcy) と Kappa 係数及びを示す.

Overall	
ACC	Kappa
0.83	0.76

図 9　30s エポックに対する Sleep-CAM モデルの正答率, Kappa 係数

混同行列と再現率, 適合率は図 10 の通りとなった.

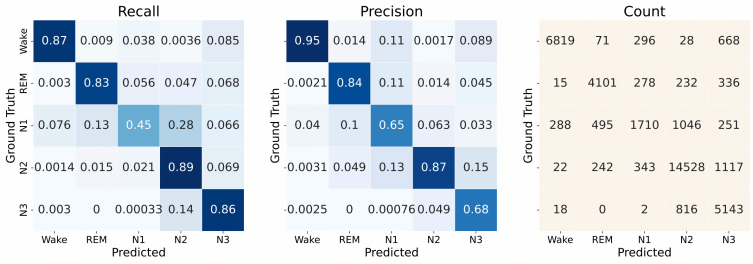


図 10　Sleep-CAM モデルの混同行列 (右) および再現率 (左), 適合率 (中)

関連研究と精度を比較したとき, 1 エポック (30s) を用いた軽量なモデルとしては十分な判定精度をもっているといえる．また専門技師間の一致率が約 82% [5] であることから, 専門技師と同等の正答率を持つといえる．各ステージについて、再現率及び適合率を見てみると N1 ステージの精度が低いが, これはサンプルの中の N1 ステージの占める割合が少ないことに起因する.

次に Realtime-Sleep-CAM モデルについて正答率 (Accurcy) と Kappa 係数及びを示す.

	Overall	
	ACC	Kappa
5s	0.84	0.66
15s	0.85	0.69
30s	0.86	0.72

図 11　各入力サイズ長のエポックに対する Realtime-Sleep-CAM モデル (入力サイズ=5s, 15s, 30s) の正答率, Kappa 係数

混同行列と再現率, 適合率は図 12, 13, 14 の通りとなった.

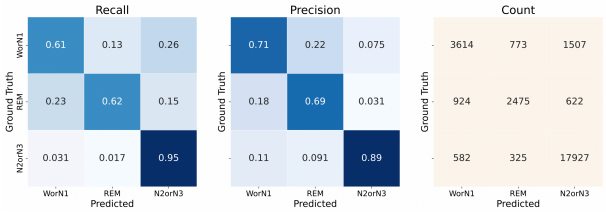


図 12　Realtime-Sleep-CAM モデル (入力サイズ=5s) の混同行列 (右) および再現率 (左), 適合率 (右)

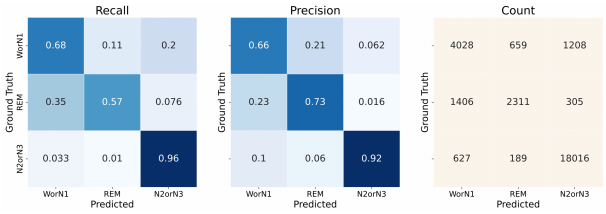


図 13　Realtime-Sleep-CAM モデル (入力サイズ=15s) の混同行列 (右) および再現率 (左), 適合率 (右)

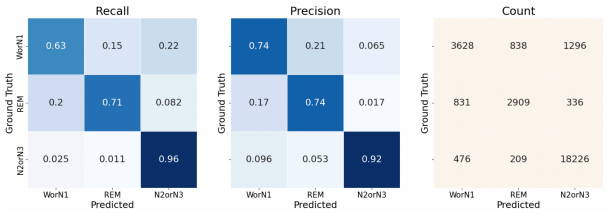


図 14　Realtime-Sleep-CAM モデル (入力サイズ=30s) の混同行列 (右) および再現率 (左), 適合率 (右)

正答率を見てみると 3 ステージ判定であるため, Sleep-CAM より数%高い判定精度となった．図 12,13,14 の混同行列よりステージ学習に用いた正解ラベルのステージの分布が偏っており, 特に”N2orN3”ステージのラベル数が多いため, 結果として”N2orN3”ステージについて高い判定精度をもつモデルへと学習が行われたと考えられる．また, 図 11 の 30s ごとのエポック単位での判定精度を見てみると, 入力サイズが大きいほどモデルの精度は高くなった.

5.3.3　ステージ遷移の有無に対する Realtime-Sleep-CAM モデルの判定精度

5.2 節で述べた正解ラベルの作成方法では, エポック内で 1 つのステージが継続するものとエポック内でステージ遷移が生じるエポックが存在する．そこで, 複数ステージにまたがらないエポックに対する判定精度を”Without transition”, 複数ステージにまたがるエポックに対する判定精度を”With transition”として, それぞれ判定精度を求めた. 結果を図 15 に示す.

	Overall		Without transition		With transition	
	ACC	Kappa	ACC	Kappa	ACC	Kappa
5s	0.84	0.66	0.84	0.67	0.56	0.26
15s	0.85	0.69	0.87	0.73	0.59	0.32
30s	0.86	0.72	0.9	0.78	0.64	0.42

図 15　ステージ遷移の有無による Realtime-Sleep-CAM モデルの判定精度

結果を見てみるとステージ遷移にまたがるエポックに対する判定精度が低いことがわかる．これについてはエポック内に 2 つ以上のステージが混在し, それに伴い Realtime-Sleep-CAM モデルの 3 ステージ確信度がバラついたため, ステージ継続エポックと比較して精度が下がることが予想される．一方で, ステージが継続するエポックに対してはさらに高い判定精度をもつことがわかる.

5.3.4　リアルタイムな判定精度

今回提案したサーバクライアントシステムを用いて, Sleep-CAM モデル及び Realtime-Sleep-CAM のリアルタイムな判定精度を検証する．ただし, 正解ラベルの都合上”Wake or N1”, ”REM”, ”N2 or N3”についての判定精度を検証する．判定間隔は 1s とし, relatime-Sleep-CAM モデルについては入力サイズが 5s, 15s, 30s の場合それぞれでリアルタイム判定をおこなった.

		Overall	
		ACC	Kappa
SleepCAM		0.85	0.75
realtime-SleepCAM	5s	0.86	0.76
	15s	0.86	0.76
	30s	0.84	0.74

図 16　Sleep-CAM モデル, Realtime-Sleep-CAM モデル (入力サイズ=5s, 15s, 30s) のリアルタイム判定に対する正答率, Kappa 係数

結果を見てみると, 入力サイズが 5s の場合の Realtime-Sleep-CAM モデルが最も高い正答率をもっていることがわかった．しかし, その差は 1%ほどであり, リアルタイム判定において Realtime-Sleep-CAM モデルが優位であると言い難い.

一方でステージが切り替わるタイミングに注目してみるとある一定の傾向が見られた．図 17, 18 にその傾向が現れる睡眠経過図を示す．図 17 では 02：00：10 に”N2 または N3”から”REM”への遷移が発生しているが, Realtime-Sleep-CAM モデルの入力サイズが 5s の場合誤差 2s, 15 の場合誤差 9s,30s の場合 24s となり, Sleep-CAM モデルでは誤差 20s となり Realtime-Sleep-CAM モデルの方が誤差が少ないといえる．さらに Realtime-Sleep-CAM モデルの入力サイズが小さいほどその遷移に対して誤差なく判定できていることがわかる．図 18 では 01：52：36 におよそ 10s の短い”Wake または N1”ステージへの遷移が発生しているが, 入力サイズが 5s の場合のみ, その遷移を判定することができている.

【このようなステージ遷移に関する判定精度を検証するため

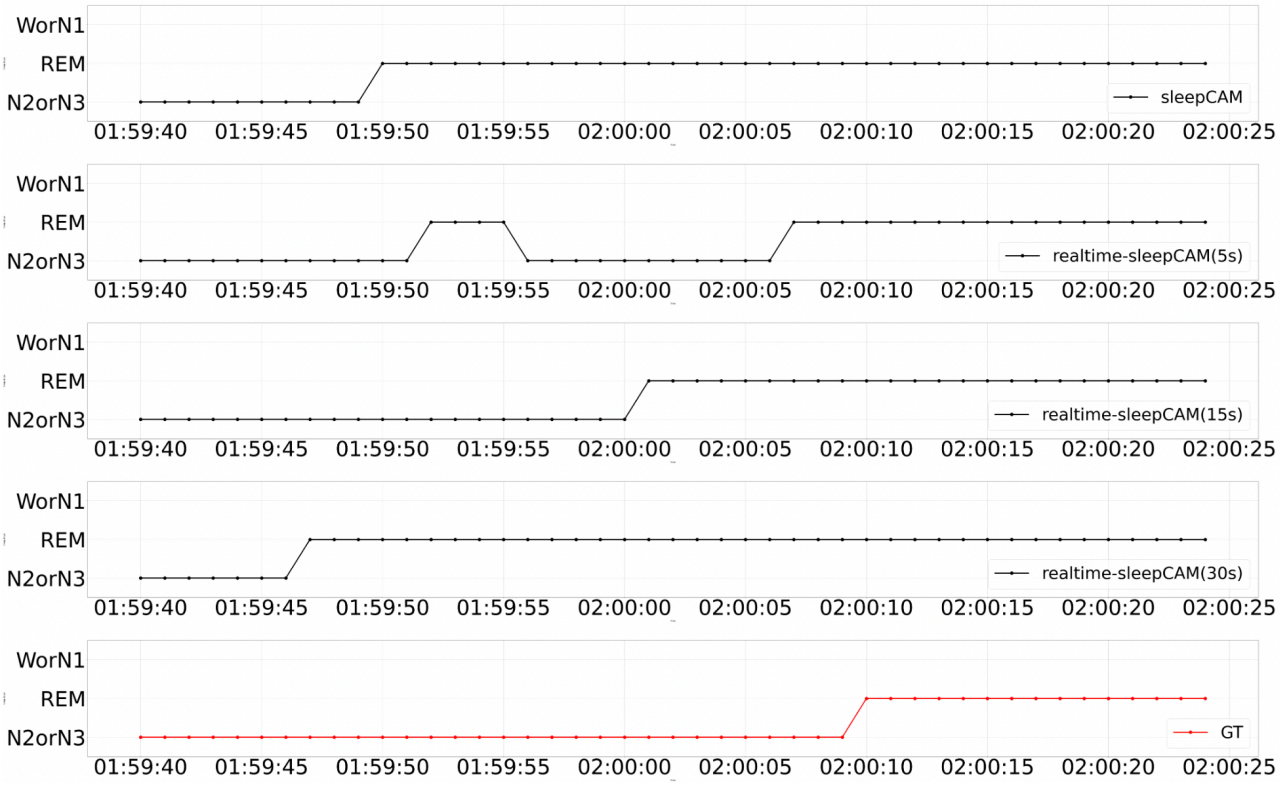


図 17　ステージ遷移 (02:00:10) に着目した正解ラベル (GT) と各入力サイズごとの判定結果

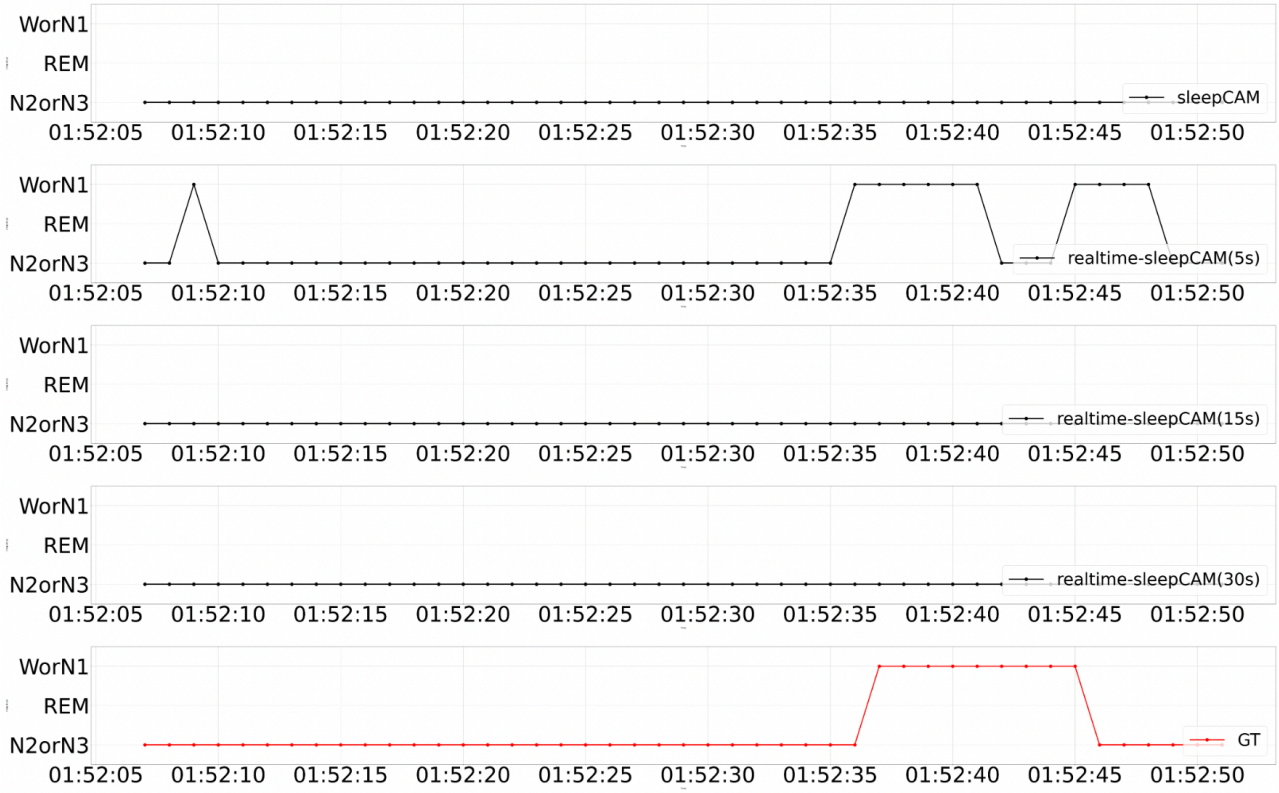


図 18　ステージ遷移 (01:52:36) に着目した正解ラベル (GT) と各入力サイズごとの判定結果

に正解ラベル (GT) 内に存在するステージ遷移の前後区間で各モデルが同様なステージ遷移判定を行えている確率を $Recall^T$, 各モデルの判定結果のうちステージ遷移判定の前後区間で正解ラベル (GT) にも同様なステージ遷移が見られる確率を $Precision^T$, $Precision^T$ と $Recall^T$ の調和平均を $F1^T$ として算出した. また前後区間の幅としては, 10s, 30s, 60s の場合それぞれについて各指標を求めた. 結果を図 19 にしめす.

結果を見てみると, 区間幅 10s で見たとき $Recall^T$,

		Sleep-CAM			Realtime-Sleep-CAM(5s)			Realtime-Sleep-CAM(15s)			Realtime-Sleep-CAM(30s)		
		$Recall^T$	$Precision^T$	$F1^T$	$Recall^T$	$Precision^T$	$F1^T$	$Recall^T$	$Precision^T$	$F1^T$	$Recall^T$	$Precision^T$	$F1^T$
window	10s	0.069	0.064	0.066	**0.399**	**0.076**	**0.128**	0.168	0.071	0.100	0.079	0.053	0.063
	30s	0.243	**0.222**	0.232	**0.658**	0.159	0.256	0.470	0.214	**0.294**	0.296	0.201	0.239
	60s	0.491	**0.430**	**0.458**	**0.783**	0.243	0.371	0.644	0.316	0.424	0.515	0.363	0.426

図 19　各区間でのステージ遷移に関わる判定精度．各区間での最良スコアを色付 (Recall: 緑, Precision: 青, F1: 赤)

$Precision^T$, $F1^T$ 共に入力サイズが 5s の Realtime-Sleep-CAM モデルが最良となった．しかし区間幅を広げていくにつれ, $Precision^T$, $F1^T$ について, 入力サイズが 15s の Realtime-Sleep-CAM(15s) や Sleep-CAM モデルなどが優れていることがわかる．$Precision^T$ について Realtime-Sleep-CAM(5s) モデルが比較的低い値である理由として, 1s ごとのステージ判定結果が振れやすくなってしまうことが挙げられる．図 20 に示す通り, 入力サイズが小さいモデルほどステージ遷移の判定回数が多いことがわかる．

Sleep-CAM	Realtime-Sleep-CAM(5s)	Realtime-Sleep-CAM(15s)	Realtime-Sleep-CAM(30s)
1257	6019	2689	1752

図 20　各モデルのリアルタイム判定結果ステージ遷移判定の回数

これらのことから, 入力サイズの大きい Sleep-CAM モデルや Realtime-SleepCAM(30s) モデルについては, 前後 30s の 60s という広い区間でみると判定精度が高い一方, 前後 5s の 10s という狭い区間でみると判定精度が低いことからリアルタイムな判定において 30s 以内の遅延が発生する可能性はあるが, 最もステージ遷移を捉えられているといえる．一方, 入力サイズの小さい Realtime-SleepCAM(5s), Realtime-SleepCAM(15s) モデルについては, 短い区間 (10s, 30s) においては最も高い判定精度を持っていることから, ステージ判定結果が頻繁に振れてしまうため, Sleep-CAM や Realtime-SleepCAM(30s) よりステージ遷移の判定精度は低いが, その遅延は 5s, または 15s 以内に収まる可能性が高いことを示唆している．したがって, ステージ遷移の誤検出が少ないかつ, ステージ遷移判定の遅延が短いモデルを今後開発していくことが今後の課題として挙げられる．

6　ま　と　め

本研究ではヒトを対象としたリアルタイムなステージ判定システムを提案した．PSG 計測器側とステージ判定側をサーバ, クライアントとして分離し, 両者間で TCP/IP による生体信号の送受信を行うようなシステムを設計した．一方で本システムの判定の仕組みとしては, 機械学習に基づく自動判定モデルを組み込み, 移動窓による判定手法を用いてリアルタイム判定を可能とした．このリアルタイム判定について処理時間を実験により確認したところ, 最小 1s でのリアルタイム判定が可能だと確認できた．

さらに新たに用意した 1s 刻みでのステージラベルデータを用いて, 既存のステージ判定モデル (Sleep-CAM モデル) を学習させ, Realtime-Sleep-CAM を作成した．1s 刻みのラベルデータを用いることで従来の 30s エポック単位でのラベルデータでは検証できなかった, リアルタイムなステージ判定に対する精度を算出することができた．実験を通して, Realtime-Sleep-CAM モデルは 30s エポックに対しても Sleep-CAM モデルと同等の判定精度を確認でき, リアルタイムな判定精度では入力サイズを短くすることで, 最も高い判定精度を達成できることを確認できた．

今後の課題として, リアルタイム判定精度の改善が挙げられる．本研究では, 5 夜分の 1s 刻みのラベルデータを作成し, リアルタイム判定モデルを学習させたが, そのラベルの事前分布が偏っていたため, ある特定のステージを優先して判定してしまう学習となってしまった．これについては, ラベルの分布を均一にするようなデータ拡張が必要だと考えられる．また, 汎用的な学習を目指すためにはより多くのラベルデータを作成することも重要である．

7　謝　　辞

本研究の一部は, JSPS 科研費 JP23H03399, JP22K19802, AMED Grant Number JP21zf0127005, F-MIRAI 共同研究, (株) S'UIMIN 共同研究による．

文　　献

[1] Ferrie JE, et al, Sleep epidemiology - a rapidly growing field. Int J Epidemiol, 40(6) 1431-1437, 2011.
[2] Andrew M. Gleason, A curvature formula, Amer. J. Math. 101 (1979) 86–93.
[3] Roth Thomas, Characteristics and determinants of normal sleep, The Journal of clinical psychiatry vol. 65 Suppl 16 pp.8-11, 2004.
[4] Iber, C., Ancoli-Israel, S., Chesson, A. L., and Quan, S., "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, 2007.
[5] Danker-Hopfe, H. et al, Interrater reliability for sleep scoring according to the Rechtschaffen & Kales ans the new AASM, J. Sleep Res. 18(1) 74-84, 2009.
[6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, Going deeper with convolutions, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1-9, 2015.
[7] Matthew D Zeiler and Rob Fergus, Visualizing and understanding convolutional networks, In European conference on computer vision, Springer, pp.1097-1105, 2014.
[8] Orestis Tsinalis, Paul M. Matthews, Yike Guo, Stefanos Zafeiriou, Automatic Sleep Stage Scoring With Single-Channel EEG Using Convolutional Neural Networks, 2016

[9] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo, Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 25, No.11, pp. 1998-2008, 2017.

[10] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chen, and Maarten De Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 27, No. 3, pp.400-410, 2019.

[11] Abhay Koushik, Judith Amores, Pattie Maes, Real-Time Sleep Staging using Deep Learning on a Smartphone for a Wearable EEG, IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2019.

[12] Amiya Patanaik, Ju Lynn Ong, Joshua J Gooley, Sonia Ancoli-Israel and Michael W L Chee, An end-to-end framework for real-time automatic sleep stage classification, Sleep, Vol. 41, No. 5, 2018.

[13] Kazumasa Horie, Leo Ota, Ryusuke Miyamoto, Takashi Abe, Yoko Suzuki, Fusae Kawana, Toshio Kokubo, Masashi Yanagisawa and Hiroyuki Kitagawa, Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability, Scientific Reports 12, No. 12799, 2022.

DEIM Forum 2023 XX-Y

特定健診データに含まれる自由記述に着目した予後予測

大塚　皇輝† 　池之上辰義†† 　福間　真悟††† 　若宮　翔子† 　荒牧　英治†

† 奈良先端科学技術大学院大学 先端科学技術研究科情報領域　〒 630-0192 奈良県生駒市高山町 8916 番地の 5
†† 滋賀大学 データサイエンス AI 推進センター　〒 522-8522 滋賀県彦根市馬場 1 丁目 1-1
††† 京都大学 医学研究科 人間健康科学科系専攻　〒 606-8507 京都府京都市左京区聖護院川原町 53
E-mail: †{otsuka.koki.om2,wakamiya,aramaki}@is.naist.jp, ††saikawa@n-univ.ac.jp,
†††fukuma.shingo.3m@kyoto-u.ac.jp

あらまし　日本は少子高齢化の影響によって医療財源の逼迫が問題となっており，予防医療や実診療での適切な医療資源の配分が求められている．このような背景から，近年では機械学習を用いて患者の予後を予測することで，医療分野における意思決定を補助するような試みも登場してきた．一方で，機械学習はその予測過程がブラックボックスであり説明可能性という観点から医療分野での導入の障壁となっている．そこで本研究では自然言語の持つ説明可能性の高さに着目し，特定健診データから言語モデルを用いて予後の予測を試みた．特に，減量の予後予測に対する言語モデルの予測精度と説明可能性について検証を行った．

キーワード　自然言語処理，医療情報学，予後予測，解釈可能性，特定健診

1　はじめに

日本は少子高齢化や人口減，労働人口現象に伴う財政難など様々な問題を抱えている．これらに起因して日本の税収に占める社会保険料の増加が問題となっており，医療の人的・物的資源の効率的な利用が課題となっている．特に，日本では悪性新生物（がん）や心疾患といった一般的には生活習慣に起因する疾患が死因の上位を占めている [1]．これに対して，日本では 2008 年より生活習慣病の予防を目的として 40 歳から 75 歳までを対象に特定保険審査，及び生活習慣病リスクの高い患者に対しては特定保健指導を実施している．しかし，指導を行なっても生活習慣が改善されない場合も多く，効果を疑問視する声もある．例えば運動や禁酒を薦められ，しばらく努力しても諦めてしまうといった事例も多い．どのような患者にどのように指導を行うかは，保健師・栄養士などの指導員の経験によるところが多く，指導の質の向上と均質化が求められる．

このような中，近年急速に発達している機械学習技術に注目し，この技術を医療分野に応用するような研究も進んできた．特に，患者予後を予測することができれば，予防医療や実診療での医療資源の最適な利用に繋げることが可能となる．しかし，機械学習によって導き出された予測結果は，解釈性に乏しいという問題がある．基本的に機械学習モデルの予測過程はブラックボックスとなっており，人間が解釈を行うことが難しい．医療分野では治療をする上で解釈性の乏しい機械学習モデルの結果を信頼することは難しく，このことが医療分野における機械学習を用いた予測モデルの導入の障壁となっている [2]．

そこで今回は自然言語を有効に利用することで，医療現場における説明可能性という障壁を取り除けないかと考えた．医療データには様々な種類のデータが含まれており，検査値のような数値データに加え，医師の診断所見といった自由記述が多く含まれている．しかし，従来ではこれらの自然言語を機械学習モデルで扱うことが難しく，医学研究の材料としてあまり用いられてこなかった．数値データ単体を人間が解釈するためには十分な経験が必要とされるが，自然言語は比較的人間には解釈しやすいため，この特性を有効活用することで医療分野における機械学習の導入の一助となる．

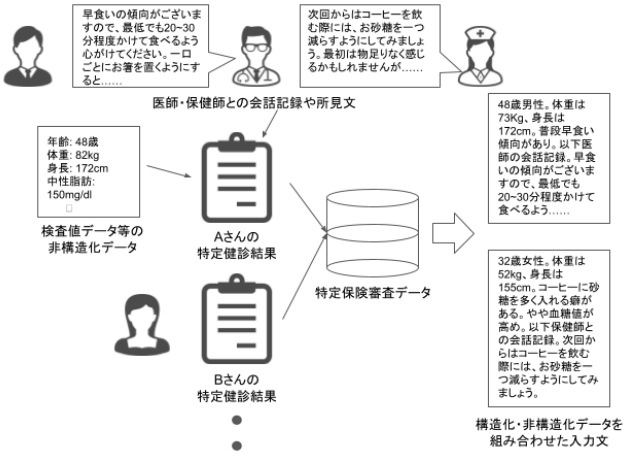


図 1　本研究の概要: 特定健診データに含まれる様々な種類のデータを言語モデルが解釈できるような自然言語に変換し，予後を予測する．

このような背景から，本研究では自然言語の持つ解釈性に注目した予後の予測, 及びその根拠となる情報を提示することにより，解釈可能性の検証を行った．予後予測の研究では検査値等の数値データや，カテゴリカルな検査データを入力として用いることが一般的となっている．しかし，このようなデータを入力に用いた場合，どの項目が予測結果に寄与したのかを判定

することしか出来ず，その数値に対する解釈はそれを見た医師の経験に左右される場合がある．一方で，自然言語の場合数値データよりもコンテキストが含まれているため，解釈を行う上では有効に作用する．加えて，電子カルテに代表される診断記録には医師の自由記述による所見も多く含まれているものの，この部分の利活用はあまり進んでいない．

本研究では，特定保険審査データと言語モデルを用いて，予後が良好か不良かを当てる分類問題を設定し，予測性能の検証をした．本研究で定義する予後とは，次回健診時までの腹囲と体重の差分から判定をする．具体的には，予後良好は腹囲が 2cm 以上減少し，同時に体重が 2kg 以上減少している場合とし，予後不良はそれ以外の状態としている．加えて，分類結果の判断根拠となる自然言語部分を可視化することで，診断の判断根拠として使用できるかどうか検証を行った．

2 関連研究

2.1 健診データを用いた研究

健康診断は 1 年に 1 回定期的に実施されることから時系列的な変化を追いやすい．そのため，このデータを軸にして様々な他媒体のデータと組み合わせた研究が盛んに行われている．

恒川ら [3] は生活習慣病の発症を予測するにあたり，健診データとレセプトを利用した．レセプトとは，医療機関が医療保険者に対して行った検査や診断，処方した薬剤等の情報が記された診療報酬明細書のことである．彼らの研究では，生活習慣病の発症を予測するにあたり，1 年に 1 回しかない健診データだけを用いた場合，生活習慣病になった時期の詳細が不明であるためレセプトを同時に使用した．作成されたデータには数値データや医師の自由記述など様々な形式のデータが含まれることになる．これらを解析した結果，生活習慣病の中でも糖尿病が識別しやすい病気であることが発見され，HbA1c や糖代謝判定といった項目が予測に大きく寄与していることがわかった．一方で，この研究では心電図の結果等に関する医師の自由記述項目がデータに含まれていたが，自然言語の処理を組み込むことができず今後の課題として紹介されている．

大場ら [4] も同様に健康診断データを利用して，生活習慣病の中でも特に糖尿病を予測するという研究を行った．彼らの研究では，予測精度だけではなく，医療分野での使用を想定し解釈可能性といった点にも注目している．健診データに加えて質問票への回答から予測に寄与するデータを明らかにしている．これを調査するにあたり，構築したモデルの入力として使われている属性をランダムに置き換えて変化を計測する Permutation importance [5] と，摂動を与えてその前後の予測結果の違いを分析する Sensitivity analysis といった手法を利用している．前者は変化が大きかった場合，その属性が予測結果に与える影響が大きかったと判定することができるが，その解釈を行うことはできない．一方で，後者は摂動を与えた際の前後の変化を観察することで，どの程度の変化がどのような予測結果に紐づいているか解釈することが可能となる．

2.2 マルチモーダルなデータ活用

健診データのように一度に複数項目の検査を実施すると，検査値データや心電図等の時系列データ，医師の所見などの自然言語等様々な種類のデータが作成される．それぞれ検査目的が異なるため，どの項目を用いて予測を行うべきかは医師などのドメインエキスパートによって選ばれる．そこで，例えば，モデルが数値データだけではなく，医師が診断時に感じ取った患者の状態等を記した自由記述データを同時に処理することができれば，より高精度な予測が可能になる．こういったことから，Agarwal ら [6] はマルチモーダルなデータを一つのモデルで処理することで，COVID-19 の患者の予後を予測する研究を行った．具体的には，一定期間後にまだ入院しているかどうか，人工呼吸器を装着することになるかどうかの 2 点を予測項目とした．これらの項目を予測する際に，性別や身長といったデータから飲酒喫煙の状態，検査値等の連続データを一つのベクトルとして埋め込むことでモデルの入力として利用している．

3 データセット

3.1 特定保険審査・特定保健指導データ

日本では生活習慣の変化により，高血圧症や糖尿病等のいわゆる生活習慣病の有病者が増加している．日本人の死因上位における悪性新生物や心疾患の原因には生活習慣病があるとされ，これらの対策が急務となっている．このようなことから日本は平成 20 年 4 月より生活習慣病の予防を目的として，年に一度 40 歳から 75 歳までを対象に内臓脂肪肥満に着目した特定保険審査を実施している．これには身体計測や血液検査に加え，質問票を用いた問診も含まれる．この検査で生活習慣病の発症リスクが高く，生活習慣の改善で予防ができると判断された場合，医師や保健師による特定保健指導により，生活習慣の改善に取り組むことになる．今回は，全国土木建築国民健康保険組合加入者に対して，2013 年 3 月から 2017 年 3 月の間に行われた特定保健師指導の電子的報告記録から抽出した 28954 件を，データセットとして用いた．

このデータには「年齢」や「身長」といった 54 種類の構造化データと，「既往歴」や「生活習慣病の改善状況」といった 11 種類の非構造化データが含まれている．表 1 に含まれている項目の一部を示す．

3.2 予測ラベルの作成

3.1 節で示した通り，特定保険審査は生活習慣病の予防を目的とし，内臓脂肪型肥満に着目した健康診断となっている．そのため，今回は内臓脂肪の数値に関係すると考えられる腹囲と体重に注目した．なお，これら 2 つの項目については，次回健診時の数値も入力されていた．そのため，これらの数値を利用し，次回健診から腹囲が 2cm 以上減少しかつ，体重が 2kg 以上減少していた場合予後が良好であると判定することとした．

3.3 不均衡データの扱い

表 2 に示す通り予測対象となる予後が良好となっているデー

表 1　特定健診データに含まれる項目の一部．下線は入力文章の作成に用いた項目を表す．

種別	項目名
構造化データ	患者 ID，喫煙有無，年齢，性別，生年月日， 歩行速度，貧血，飲酒頻度，一年間の体重変化 30 分以上の運動習慣，腹囲，BMI，身長，体重 収縮期・拡張期血圧，GOT(AST)，HbA1c， LDL コレステロール，HDL コレステロール 中性脂肪，空腹時血糖，尿糖，尿蛋白 血清クレアチニン，次回健診時体重， 次回健診時腹囲，診断回数，ALT(GPT)
非構造化データ	メタボリックシンドローム判定，保健指導レベル， 既往歴，自覚症状，自覚症状初見， 食べ方，飲酒量，保健指導時の会話記録， 飲酒習慣詳細，睡眠習慣詳細，生活習慣改善意思

タが 2698 件，不良となっているデータが 28954 件となっており大きな偏りが生じている．このまま言語モデルを学習すると予後不良となっている多数派の影響を強く受けることでモデルの出力がほぼ予後不良となるように学習してしまう危険性がある．そのため今回はデータのダウンサンプリングを実施した．予後不良となっているデータからランダムに予後良好となっている件数と同じ 2698 件を抽出して用いることにした．これにより予後良好が 2698 件，予後不良が 2698 件で同数となるデータセットを作成した．

4　手　　法

4.1　モ　デ　ル

予後の良好・不良を予測する分類問題を解くに当たって，今回は Bidirectional Encoder Representations from Transformers (BERT) [7] をベースとした 2 つの言語モデルを用いて実験を行った．1 つ目は 2019 年 9 月までの日本語版 Wikipedia に含まれる，約 1700 万行の文書を用いて事前学習した bert-base-japanese-whole-word-masking（以下 Tohoku-BERT）[8]，2 つ目は東京大学附属病院の電子カルテに保存されていた約 1 億 2000 万行の文書をコーパスとして事前学習した UTH-BERT-BASE-512-WWM (以下 UTH-BERT) [9] を用いた．これらのモデルは両方とも日本語を学習したモデルであるが，前者が汎用的な知識を獲得しているのに対し，後者は医療分野に特化した知識を獲得している．今回使用するデータに含まれる自由記述部分には医師・保健師と患者による対話が収められた部分が多く存在している．そのため医療知識をあまり持たない患者向けに，噛み砕いた説明になっていることが予想されたため，医療特化のモデルだけではなく，汎用的な知識を保持した日本語モデルも用いることとした．

構造化データを日本語に変換して結合した入力文

これは 4 回目の診察です．性別は男性．GPT(ALT) は正常．HbA1c の値は要注意．

非構造化データを日本語に変換して結合した入力文

先日はご多忙の中，お時間を頂きましてありがとうございました．〇〇〇〇〇〇です．足の痛みが和らいだこと，痛風の治療と服薬は医師の指示どおり行えていることを伺い，安心しました．今回のお電話では，...

構造化・非構造化データを結合した入力文

これは 4 回目の診察です．性別は男性．GPT(ALT) は正常．HbA1c の値は要注意．先日はご多忙の中，お時間を頂きましてありがとうございました．〇〇〇〇〇〇です．足の痛みが和らいだこと，痛風の治療と服薬は医師の指示どおり行えていることを伺い，安心しました．今回のお電話では，...

4.2　入力文章の作成

今回は言語モデルを用いて予測を行うことから，入力は自然言語による記述となるため，データセット内の特に自由記述部分の項目を主に扱うこととした．主な自由記述項目としては「生活改善指導時の発話」「継続配慮に関する申し送り」「飲料習慣」「睡眠習慣」「雑記」の 5 項目となっていたため，まずこの文章同士を繋げて一つの入力文章とした．加えて構造化データ部分に関しても予測結果に寄与すると考えられる項目が複数含まれていることから，これらの項目に関しても入力として用いることとした．図 2 はこのようにして作成された入力文章の一例である．特定保険審査データには 54 種類の構造化データが含まれていることは 3.1 節で紹介した通りである．これらの項目すべてを入力文章として変換して用いることは，今回使用するモデルの入力トークン数の最大が 512 トークンとなっていることから困難であると考えた．そのためこの 54 項目のうち予測に寄与すると考えられる「性別」「ALT(GPT)」「HbA1c」「中性脂肪」「診断回数」の 5 項目を入力文章として追加することとした．この中で ALT(GPT) は肝臓脂肪に関連する数値である．HbA1c はヘモグロビンのうち糖化したものの割合を示しており，主に糖尿病に関わる数値となっている．これらの数値を単純に文章化すると「性別は男性．ALT(GPT) は 50IU/L．HbA1c は 4.3 %．中性脂肪は 69mg/dl．診断回数は 2 回」といったなる一方で，「ALT(GPT)」「HbA1c」「中性脂肪」に関してはその数値が異常なのか判断がつかない．そのため，今回はこの 3 項目に関しては基準値を調べ，ALT(GPT) の数値が 4 未満であれば「低い」，HbA1c の数値が 5.8%であれば「正常」，中性脂肪の数値が 149mg/dl より大きければ「多い」という形で変換して文章を作成した．作成した文章とその数を表 2 に示す．また，作成した文章のトークン数を調べた結果を図 3 に示す．これによりデータのほとんどがモデルの入力トークンの最大値である 512 トークン内に収まっていることを確認した．

表 2　カテゴリを一定の基準でグルーピングした際の基準と該当数

項目		該当数
予後		
	良好	26356
	不良	2598
性別		
	男性	27841
	女性	1113
GPT(ALT)		
	低い ($< 4IU/L$)	1
	正常	23546
	高い ($44IU/L <$)	5407
HbA1c		
	正常 (< 5.6 %)	12532
	要注意	13418
	極めて高い (6.4 % <)	3002
	測定不能	2
中性脂肪		
	少ない ($< 30mg/dl$)	6
	正常	15175
	多い ($149mg/dl <$)	13773

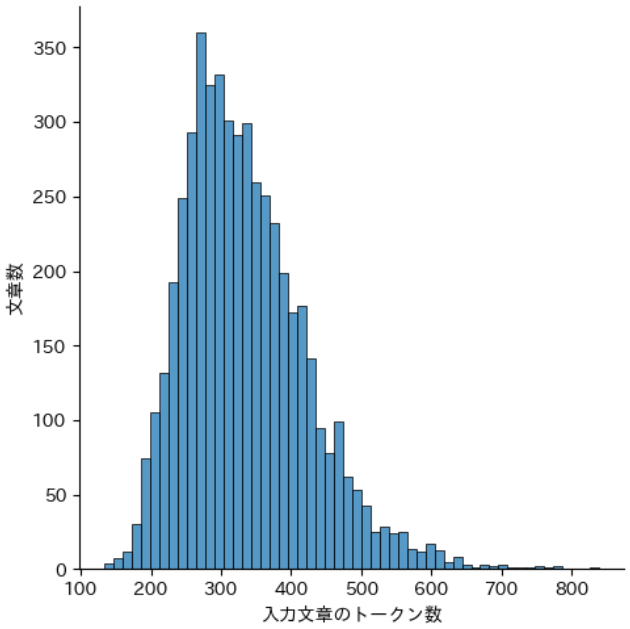


図 2　作成した入力文数とトークン数の分布

5　実　　験

5.1　設　　定

本研究では前回健診時からの患者予後が良好か不良かを判定する分類問題を設定した．データに関しては予後の数が不均衡になっていたため，これを解消するためにダウンサンプリングを実施し，最終的に 5396 件のデータを用いて実験を行った．モデルに関しては一般テキストを用いて事前学習した Tohoku-BERT と，医療分野のテキストで事前学習した UTH-BERT を用いた．入力文章に関しては構造化データのみを文章化したもの，自由記述欄のみを結合して文章化したもの，これら 2 つの文章を結合させて一つの文章にしたものの 3 種類の入力文章を作成した．評価指標に関しては分類タスクということを考慮し F 値，AUROC，Accuracy の 3 種類を用いた．加えて予後判定時に用いる閾値を 0.1 ごとに区切って感度と特異度を計算した．感度とは実際に予測対象の疾病に罹患している人のうち予測結果が罹患しているとされた人の割合を示し，特異度とは予測対象の疾病に罹患していない人のうち予測結果が罹患していないとされた人の割合を示している．本研究の場合は，予後が良好となっている患者に対して過剰な医療リソースの投下を抑制するといった状態を想定している．そのため予後良好のラベルがつけられた人のうち予測結果が予後良好となっている人の割合を感度，予後不良のラベルがつけられた人のうち予測結果が予後不良となっている人の割合を特異度として計算している．実験のハイパーパラメータとしてはバッチサイズは 32, エポック数は 1000, アーリーストップは 50, 学習率は 2e-5 とし，消費した GPU のメモリは約 29GB, 約 100epoch で停止し, 計算時間は約 2 3 時間ほど要した．

5.2　結　　果

表 2 に実験結果を示す．まず入力データセット間における精度は，構造化データのみ，自由記述欄のみ，構造化データ+自由記述の順に全体として精度が向上することが確認できた．構造化データのみを文章化したものを入力として用いた場合は UTH-BERT のほうが性能が高く，自由記述欄のみを結合して文章化したものを入力として用いた場合は Tohoku-BERT のほうが精度が高くなった．検査値データが含まれた文章と自由記述欄を組み合わせて作成した文章の場合，F 値に関しては UTH-BERT のほうが精度が高く，AUROC と Accuracy に関しては Tohoku-BERT のほうが精度が高いという結果となった．

表 4 は感度・特異度の結果を表している．医療資源の最適化といった場面で過剰な医療リソースを投下しすぎないということを目的とした場合, 閾値を 0.4 とした時の感度が Tohoku-BERT と UTH-BERT 共にが 85%ほどと十分な精度となっている．

図 4 は入力文章に Attention を表示して可視化した結果である．Attention が全体的にかかっており，ひと目見て注目するべき点がどこかを判定することは難しい．

5.3　考　　察

5.3.1　モデルの特性について

モデル別の性能を比べてみると，検査値データのみを入力文章とした場合は UTH-BERT のほうが全体的な精度が高くなっている．一方で，自由記述部分のみを入力文章とした場合は Tohoku-BERT のほうが全体的な精度が高くなっている．UTH-BERT は医療テキストを用いて事前学習されていること

表 3　入力とモデル別に測定した性能の一覧．構造化データと非構造化データを組み合わせて作成した文章を入力に用いた際に一番精度が高くなる．

	構造化データのみ		非構造化データのみ		構造化+非構造化データ	
評価指標	Tohoku BERT	UTH-BERT	Tohoku BERT	UTH-BERT	Tohoku BERT	UTH-BERT
F 値	0.630	0.629	0.646	0.015	0.658	**0.683**
AUROC	0.428	0.621	0.644	0.523	**0.662**	0.642
Accuracy	0.478	0.605	0.619	0.501	**0.630**	0.615

表 4　閾値をずらして感度と特異度を計算した結果
表 5　予後ラベル判別の閾値をずらして，感度と特異度を計算した結果

	感度		特異度	
閾値	Tohoku BERT	UTH-BERT	Tohoku BERT	UTH-BERT
0.0	1	1	0	0
0.1	0.9769	0.9961	0.0961	0.0653
0.2	0.9269	0.9961	0.2461	0.1038
0.3	0.9038	0.9615	0.3	0.1807
0.4	0.8576	0.9	0.3846	0.3346
0.5	0.7115	0.8307	0.55	0.4
0.6	0.2461	0.5769	0.8384	0.5961
0.7	0.0038	0.1807	1	0.8884
0.8	0	0	1	1
0.9	0	0	1	1
1.0	0	0	1	1

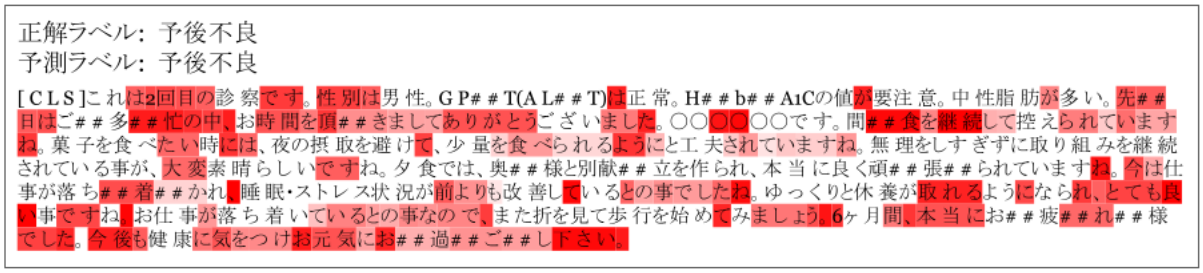


図 3　入力文章に対する予測根拠部分の Attention 可視化．

から，検査値データに関する認識精度が高かったと考えられる．一方で，自由記述欄の内容は医師・保健師と患者の会話時の文章が多く含まれており，非医療従事者向けの語彙が多く含まれている．そのため，一般的な語彙を多く含む Tohoku-BERT の方が認識できる単語が多く含まれていることから，UTH-BERT よりも Tohoku-BERT の方が精度が高くなったと考えられる．検査値と自由記述文の両方を含む内容を入力とした場合，入力文に含まれる情報としては自由記述欄の情報の方が多く含まれるようになっている．そのため Tohoku-BERT の方が若干精度が高くなったと考えられる．

5.3.2　分類性能について

実験では入力文を検査値と自由記述文から作成したとき，AUROC が 0.662，Accuracy が 0.630 程度にとどまっている．医療現場で使用することを想定すると，精度改善の余地が十分に残されている．

まず，本研究では，使用するデータセットを大きくダウンサンプリングしているため，予後不良ラベルが付与された未使用データが多く存在する．これを有効に利用するため，予後ラベルの出現割合が等しくなるようにリバランスしたデータセットを複数作成し，データセットごとにモデルを訓練した後，そのモデルの出力結果を多数決する方法が考えられる [10]．これにより，全データをモデルに投入することが可能となり，より広範囲な知識が獲得できると考えられる．一方で，今回の言語モデルのようにメモリを多く使用する場合，データセットごとのモデルを同時に学習するというのは難しい．そのため，一つのデータセットとモデルごとの結果を保存し，すべての計算が終わった後に集計処理を行うといった実装上の工夫が必要となる．

また不均衡データを，基準とする項目の個数が均衡するようにデータセットを改変して学習した後，不均衡データ全体を使って学習を行うと精度が向上するといった研究結果が存在する [11]．今回の実験では，データを均衡状態にしたため，作成したモデルを全体のデータを使ってもう一度学習することで，

精度が向上する可能性がある．

加えて，今回は，体重 2kg 以上減少かつ腹囲 2cm 以上減少した場合に予後良好とみなしたが，この条件が厳しかったという問題が考えられる．全データのうち，この条件を達成していたのは約 9%であったため，今回はこの数値に対して大きくダウンサンプリングを実施することになり，十分な精度を出すことができなかった．これに関しては，実際の現場で予後が良好と判断される人の割合と相違が無いのか調査する必要があると同時に，実験を実施する上では，条件を緩和する方法も考えられる．例えば今回の予後良好ラベルの設定を体重 1kg 以上減少，腹囲 1cm 以上減少とすることにより予後良好判定となるデータを増やすことができる．その状態での精度まず確認した上で，より厳しい条件の精度を確認するといった方法も考えられる．

5.3.3 予測確率の分布について

本研究ではデータをダウンサンプリングして予後の良好と不良が同数になったものをデータセットとして使用している．それに伴って閾値を 0.5 としてた場合にモデルの出力も予後の良好と不良の出現割合がおよそ 1 対 1 になるようになっている．しかし現実には予後が良好となる人の割合は少なく，今回のデータセットの場合約 9%しか存在していない．仮に本研究で作成したモデルが実際の医療現場で使えるだけの精度が出た場合でも，実験に使ったデータセットの分布が現実の患者予後の分布と異なるため，今回の場合では予後良好となる確率が高くなってしまう．この場合予後が不良の患者を予後良好と誤診する場合が想定され，重大な医療過誤につながる恐れがある．そのため予測確率の補正が必要になってくる．この場合現場で使用するといった際にはデータセット上の分布と現実の分布の乖離に注意し，Isotonic Regression [12] などの手法で現実のデータに適応するように確率補正を行うという処理が必要となってくる．

5.3.4 説明可能性について

本研究では従来の手法では確認することが難しい予測結果の根拠という部分に着目した．そこで可視化結果に注目すると，全体的に Attention がかかっているようになっており，特に際立った根拠部分を見つけるというのは難しい．今回は Attention 部分の可視化を行ったが，他には Integrated Gradients [13] といった手法が存在している．一方で Attention がかかっている部分が根拠として解釈してもよいのかという問題が存在しており，自然言語を用いた解釈性といった部分にはまだ検討の余地が残されている．

5.3.5 限　　界

特定健診の予後予測は経験を積んだ指導員にとっても難しい課題となっている．生活習慣の改善が求められるため，血糖値等の数値データよりも本人の家庭環境や改善への意思といったデータのほうが，予測する上では大きく寄与すると考えられる．例えば日々の仕事に追われておりほとんど家に帰宅しないといった人に対して，食生活の改善を促したとしても，それを実行するのが困難である場合が多いと考えられる．しかし従来の数値データを用いた手法ではこのコンテキストを把握することは難しく，そこに医師や保健師が診察を通じて記した自由記述データを有効に活用する価値は多く残されている．一方で本研究の結果の通り精度や解釈性には多くの課題が残されており，今後どのように活用していくか検討の余地が残されている．

5.3.6 今後の展望

近年 ChatGPT に代表されるように，BERT よりも多くのパラメータを持つ言語モデルが登場してきた．医療分野に特化したものでは Med-PaLM [14] というモデルが登場している．モデル自体はまだ一般公開されているものでは無いが，既に米国の医師国家試験と同等レベルとされる MedQA タスクにおいて約 86.5%もの精度を達成したと報告されている．このような医療分野に特化したより大規模なモデルは日々開発させており，自然言語を用いた病院内データの利活用はより進んでいくと考えられる．

このような巨大な言語モデルが発達していくに連れて，人間がこれらを扱う方法についても変わってくる．本研究で用いた BERT のようなモデルは，一定のタスクを設定してそれに対してモデルを学習させて使用するといった形が一般的であった．そのため入力は文章，出力は数値といった形になるとその数値の算出過程を明らかにすることが難しいため解釈方法といった別の問題も出てきていた．一方で近年の巨大な言語モデルは，一つのモデルをあらゆる大量のデータで学習することで，そのモデルですべてのタスクを解くという方向に変わってきている．このようなモデルは入力，出力ともに文章となっているため，入力時の命令に「出力時には予測根拠も出力すること」といったようなプロンプトを入力しておくことでモデルの予測根拠のようなものを同時に出力することができる．また，ChatGPT のようにチャット形式で使用することで人間が補完しきれていない知識を取り出すことが可能となっている．そのため人間側が機械の出力結果を使用できる程度のタスクに落とし込むことさえできれば，解釈性といった問題を無視することができる．例えば出力に対する信頼性が低くとも，人間がその出力に対して裏取り調査を行った上で利用する等が考えられる．このような変化が起こる中で，言語モデルの進化に伴い人間側もその使用方法を改善することで導入の障壁を取り除くことができるのではと考えられる．

6 おわりに

本研究では自然言語の持つ解釈性に着目し，特定健診データを使用して患者の予後予測を行った．データには構造化データも含まれていたため，これらを活用するべく 3 種類の文章を作成し，言語モデルの入力文章とた．予測ラベルに関しては次回健診時の体重と腹囲の変化から作成した．言語モデルは一般分野のテキストで事前学習済みの Tohoku-BERT と，医療分野のテキストで事前学習済みの UTH-BERT の 2 種類を使用した．これらを組み合わせた 6 種類の条件から予後の予測精度を調べた結果，約 63%程度の正解率となることが判明した．加えて医療分野で使用するためには説明可能性が重要になってくるため，今回はこの予測モデルの Attention を可視化することで根拠部分の判定が可能か調査を行った．その結果，現段階

では Attention のかかった部分を見ることで判断の根拠となる部分を判別することは難しかったが，一部診断回数には比較的 Attention が強くかかっていることが判明し改良次第では実応用可能なのではないかと考えられる．

7 謝　　辞

本研究の内容の一部は戦略的創造研究推進事業（CREST），及び JSPS 科研費 JP 19K10620 の助成を受けたものである．また本研究に使用したデータは，SOMPO ヘルスサポート株式会社および全国土木建築国民健康保険組合より提供を受けたものである．なお，本研究は京都大学医の倫理委員会（R0817-3）で承認を受けている．

文　　献

[1] 令和 4 年 (2022) 人口動態統計月報年計 (概数）の概況. p. 10, 2022.
[2] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 11, pp. 4793–4813, 2021.
[3] 荒木雅弘 新谷元司 吉川昌孝恒川充. 健診データを用いた生活習慣病の発症予測. 2019.
[4] 讃岐勝 我妻ゆき子大場勇貴. 健康診断データを用いた疾患予測における解釈可能なモデルの構築. 2020.
[5] Oliver Sander Thomas Lengauer André Altmann, Laura Tološi. Permutation importance: a corrected feature importance measure. Bioinformatics, Vol. 26, No. 10, pp. 1340–1347, 2010.
[6] Sutanay Choudhury Sindhu Tipirneni Pritam Mukherjee Colby Ham Suzanne Tamang Matthew Baker et al Agarwal, Khushbu. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multimodal bert: a study on covid-19 outcome prediction. Scientific Reports, Vol. 12, , 2022.
[7] Ming-Wei Chang Kenton Lee Devlin, Jacob and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
[8] M Suzuki. cl-tohoku/bert-japanese: Bert models for japanese text. Accessed: 3-Nov-2023.
[9] Daisaku Shibata Emiko Shinohara Eiji Aramaki Kawazoe, Yoshimasa and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. PLoS One, Vol. 16, No. 11, 2021.
[10] Kevin Small Carla E. Brodley Wallace, Byron C. and Thomas A. Trikalinos. Class imbalance, redux. In 2011 IEEE 11th international conference on data mining, pp. 754–763, 2011.
[11] Simon Wouters Ruben Cartuyvels Erfan Ghadery Malik, Farjad and Marie-Francine Moens. Two-phase training mitigates class imbalance for camera trap image classification with cnns. p. 10, 2021.
[12] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. pp. 625–632, 2005.
[13] Ankur Taly Sundararajan, Mukund and Qiqi Yan. Axiomatic attribution for deep networks. PMLR, pp. 3319–3328, 2017.
[14] Tao Tu Juraj Gottweis Rory Sayres Ellery Wulczyn Le Hou Kevin Clark et al Singhal, Karan. Towards expert-level medical question answering with large language models. 2023.

AI 音楽療法士（さつきさん）の開発研究

音楽療法のデジタルセラピューティクスにむけて

小杉　尚子†　　石井健太郎†　　児玉　直樹††

† 専修大学ネットワーク情報学部　〒 214–8580 神奈川県川崎市多摩区東三田 2-1-1
†† 新潟医療福祉大学医療技術学部　〒 950–3198 新潟県新潟市北区島見町 1398 番地
E-mail: †naonaoan@gmail.com, ††kenta@pc.fm.senshu-u.ac.jp, †††kodama@nuhw.ac.jp

あらまし　本論文では、認知症高齢者を対象とする、人工知能（AI）を用いた音楽療法士 (AI 音楽療法士) の開発研究について述べる。音楽療法は、認知症の「行動・心理症状」に効果があるため、介護負担の軽減に資することが期待されている。しかし、日本における音楽療法士の数は認知症高齢者の数に比べて非常に少なく、日本全国の認知症高齢者が誰でも気軽に音楽療法を受療できる状況ではない。そこで認知症高齢者がいつでもどこでも何回でも音楽療法を受療できるようにするために、パソコンの画面 に現れるバーチャルな音楽療法士（AI 音楽療法士）「さつきさん」を開発している。本論文では、最新の「さつきさん」について報告すると共に、その「さつきさん」を、実際に認知症高齢者の介護に従事する介護スタッフに試用して頂いた際に得た意見をまとめ、本研究の方向性について考察する。わたしたちは、 将来的には、AI 音楽療法士による音楽療法の受療者から多様なデータを収集し、受療者の状態の客観評価や認知症の進行予測なども可能な音楽療法のデジタルセラピューティクスを確立したいと考えている。
キーワード　AI、音楽療法、認知症、デジタルセラピューティクス

1　はじめに

令和 5 年版高齢社会白書 [4] によると、令和 4 年 10 月 1 日現在、65 歳以上人口は 3,624 万人、75 歳以上人口は 1,936 万人となり、総人口に占める割合（高齢化率）はそれぞれ 29.0%、15.5%となったとのことである。令和 4 年版高齢社会白書 [3] によると、令和 3 年 10 月 1 日現在、65 歳以上人口は 3,621 万人、75 歳以上人口は 1,867 万人で、総人口に占める割合（高齢化率）はそれぞれ 28.9%、14.9%であったので、1 年間で高齢化率はそれぞれ 0.1%、0.6%上昇したことになる。また令和 4 年版高齢社会白書によると、65 歳以上の要介護者数は増加しており、介護が必要となった主な原因としては、「認知症」が最も多く、18.1%を占めているとのことである。一方令和 5 年版高齢社会白書では、要介護者等と同居している主な介護者の年齢が 60 歳以上である割合が、男性では 72.4 %、女性では 73.8 %となっており、「老老介護」のケースが相当数存在していることや、介護を理由にした離職者の約 3/4 が女性であったことなどを報告しており、認知症高齢者数の増加と介護人材の不足が、重要な社会課題となっていることが分かる。

認知症とは、正常に発達した認知機能が、後天的な何らかの障害により、生理的老化の範囲を超えて広汎かつ慢性的に低下した状態のこと [1] で、記憶、言語、視空間認知などの認知機能の障害と、BPSD (Behavioral and Psychological Symptoms of Dementia）と呼ばれる徘徊や無気力などの「行動・心理症状」からなる [2]。認知症には、現状では有効な治療法がないため、認知症高齢者への対応はケアが中心となるが、認知症高齢者のケアは非常に困難であるため、介護者の負担を軽減することが重要である。

認知症高齢者の介護を困難にするのは徘徊、不安、抑うつなどの BPSD であるが、これは適切なケアを行うことで発症を抑制したり、症状を軽減したりすることができるため、非薬物療法で対応することが推奨されている。認知症疾患診療ガイドライン 2017 には、非薬物療法として運動療法、回想法、音楽療法などが挙げられている。

音楽療法は認定資格を持つ音楽療法士によって実施されるが、日本で活動する音楽療法士の数は 2,461 名（2021 年 3 月末日現在）[5] と少なく、その 4〜5 割が関東域で活動している可能性が高いため、絶対数不足と首都圏偏在が問題になっている。また音楽療法士には女性が多いため、出産・子育てや介護などが音楽療法士として経験を重ね技量を磨く時期に重なる場合も多く、そのような時期には十分な臨床実践が難しくなるという問題も抱えている。これは結果的に、音楽療法士は自身の成長機会を失い、療養者も音楽療法の受療機会を失うという、音楽療法士と療養者の両者の問題になっている。

そこで私達は、認知症高齢者が「いつでもどこでも何回でも」気軽に音楽療法を受療できる社会の確立を目指して、AI(人工知能) を用いた音楽療法士（以下、AI 音楽療法士）の開発研究を推進している [8]。私達が開発している AI 音楽療法士は「さつきさん」という名前で、認知症高齢者の「発話サポート」と「歌唱伴奏」に特化した AI 音楽療法士である。本論文では「さつきさん」の研究開発の最新の進捗について述べると共に、それを認知症介護従事者に試用してもらってインタビューしたのでその結果について報告する。そしてそれに基づいて、今後の展開について議論する。本論文の構成は以下である。第 2 章では音楽療法とデジタルセラピューティクスについて説明する。

第 3 章で AI 音楽療法士について説明し、第 4 章で実験と評価結果を報告し考察を述べる。最後に第 5 章で本論文をまとめる。

2　音楽療法とDTx(デジタル・セラピューティクス)

音楽療法とは「音楽の持つ生理的、心理的、社会的働きを用いて、心身の障害の回復、機能の維持改善、生活の質の向上、行動の変容などに向けて、音楽を意図的、計画的に使用すること」である [6]。週に 1 回から月に 1,2 回の頻度で、歌唱（合唱）や楽器演奏などを通して、受療者のストレス発散や不安の軽減 [11]、認知機能の改善 [15], [16], [17] などを支援する。主な対象者は、精神疾患患者や高齢者などの投薬や手術だけでは状態の改善が難しい方々である。日本では 5〜10 名を対象に、1 時間前後のプログラムが行われることが多いが、受療者の主訴や人数等に応じて柔軟に調整可能である [18], [19], [20]。

音楽療法は、第 1 章でも述べた通り、認知症に対する有効性が既知であるが、日本では音楽療法士が不足しているため、誰でも気軽に受療できる状況ではない。そこで著者らは、インターネットを利用して音楽療法を提供・受療する遠隔音楽療法システム（Music-Telepy）を開発した [9]。Music-Telepy は、音楽療法士と認知症高齢者施設を光回線で結び、パソコン、Web キャスティング・ミキサー、マイク、スピーカ、モニタを用いて音楽療法を実施するもので、特に音楽療法士のいない離島や冬場の外出が困難になる豪雪地では大変好評だった [10]。そして著者らは、このような臨床実践・臨床研究を通して、遠隔音楽療法が、認知症高齢者に対して、対面方式の音楽療法とほぼ同等の臨床効果があることと、認知症高齢者が、モニタに映し出された音楽療法士と自然に会話できることも確認した [11]。

「デジタル・セラピューティクス（Digital Therapeutics: DTx）」は、日本語では「治療用アプリ」と訳される。医療現場における ICT 技術の活用は、主に診断の場で進められてきた。しかし昨今、ICT 技術の新しい活用の場として注目を浴びているのが「治療の場」である。日本では 2020 年 8 月 21 日に禁煙を支援するためのニコチン依存症治療用アプリ「CureApp SC」と呼気中の一酸化炭素濃度を計測する「ポータブル CO チェッカー」が薬事承認を得て注目を集めた [12]。アメリカでは約 10 年前にアメリカ食品医療品局が糖尿病の治療補助アプリ「BlueStar」を承認し、医療保険の適用が認められている。現代社会では、医療技術の進化に伴う超高齢化や、重大事故などによる死亡が減少している。これ自体は好ましいことであるが、超高齢者や重大事故などで障害を負った方々が、その人らしく生きる・生活するためには、病院での治療に加えて様々な医療や介護サポートが必要であることは想像に難くない。また社会のグローバル化・複雑化等で精神疾患や慢性疾患を抱えながら社会生活を送る人々が増えている。そのような傷病では「行動変容」や「自己管理」が治療の過程で欠かせないため、いつでもどこでも利用できる DTx に期待が高まるのは必然である。さらに DTx ではセラピストの力量に左右されない均質なセラピーや、患者の応答内容・表情や声のトーン等の様々な情報を利用・学習することで、患者の状態のタイムリーな把握、治療効果の客観評価、個々の患者に合わせたアプローチや、病気・症状の進行予測、それに基づいた予防措置の提案等も可能になると考えられるため、今後益々関心が高まることは間違いない。

3　AI 音楽療法士

著者らは音楽療法の DTx(第 2 章参照) として AI 音楽療法士を研究開発している。

AI によって代替できない職種として常に挙げられるものの 1 つに「セラピスト（Therapist, 療法士）」があり、AI 療法士の開発は容易ではないと考えられている。音楽療法士（ミュージック・セラピスト）は、音楽を使って対象者に寄り添う専門職であるため、これを完全に AI で実現するには相当困難であると考えられる。しかし音楽療法士の業務は多様で多岐に渡るため、AI が実施可能、むしろ AI の方が得意な業務も存在することは想像に難くない。そこで我々は、音楽療法を「AI 音楽療法士と生身の音楽療法士の協業」と捉え直し、AI 音楽療法士が実施可能な業務を厳選してその完成度を高め、生身の音楽療法士との協業の形を新しく定義するという方針で研究を進めている。

3.1　概　　要

著者らは AI 音楽療法士が実施可能な業務として、私達は「会話」と「歌唱伴奏」を選定した。認知症高齢者に対する音楽療法では、歌唱、体操、楽器演奏などが行われることが多く、それらの活動の間や、セッションの最初と最後では、音楽療法士との楽しい会話が人気だからである。

特に「会話」については認知症高齢者の「発話サポート」にフォーカスし、高齢者に安心して楽しく好きなだけおしゃべりしてもらえる環境構築を目指している。そのため、「〜と言えば?」や「〜の思い出は?」等の「問いかけ」や、「○○が大変だったのですね」や「私も○○が好きです」等の「復唱」を多用する、シナリオベースの会話のデザインを進めている。これは「イライザ効果」[26] として知られる、「復唱と簡単な質問を繰り返すだけで、あたかも自分が理解されている・受け容れられていると勘違いし、ユーザは対話に没頭した」ことを参考にしたもので、私達もこのような対話デザインを採用することで、話好きな高齢者の発話を促進し、対話による満足感を高め、AI 音楽療法士の定期的かつ長期的な利用に繋げられるのではないかと考えている。AI 音楽療法士は音楽療法のデジタルセラピューティクスとして開発するため、利用者の定期的かつ長期的な利用は療養データの収集には欠かせない重要な特性である。

「歌唱」は音楽療法において高齢者に人気の高い最頻活動であるため必須だと考えている。将来的には発話音声を AI を用いて分析し、対象者の歌声のキーやテンポを推定しながら [13], [14]、目の前の高齢者の歌唱に合わせた伴奏をすることを目指している。現段階では音楽療法士に指定されたキーとテンポで、「幸せなら手をたたこう」の伴奏音源を再生している。本研究で開

図 1　AI 音楽療法士が高齢者に音楽療法を提供しているイメージ

発する AI 音楽療法士のイメージを図 1 に示す。

3.2　システム構成

AI 音楽療法士のシステム構成を図 2 に示す。

システムは、1) 音声認識モジュール、2) 発話伴奏モジュール、3) バーチャルエージェントモジュール、4) 対話シナリオ管理モジュールからなる。

音声認識モジュールは、対話シナリオ管理モジュールからのトリガーにより、「療法対象者」の発話音声を認識し、テキストへの変換を行う。音声からテキストへの変換には外部の音声認識エンジンを利用しており、変換されたテキストは対話シナリオ管理モジュールに転送される。

発話伴奏モジュールは、対話シナリオ管理モジュールからのトリガーにより、「さつきさん」の発話音声と音楽の伴奏を再生する。再生する発話音声と伴奏はファイルとしてシステムの実行環境に予め用意してあり、システムの実行時には指定のファイルが再生される仕組みである。今後は、システムの実行時に音声合成や伴奏音源の微調整を行いながら再生することも検討している。

バーチャルエージェントモジュールは、対話シナリオ管理モジュールからのトリガーにより、「さつきさん」の身体動作を再生する。再生する身体動作はファイルとしてシステムの実行環境に予め用意してあり、システムの実行時には指定のファイルが再生される仕組みである。今後は、まばたきなどの不随意動作をシステムの実行時に自動で追加する仕組みを導入することも検討している。

対話シナリオ管理モジュールは、音声認識テキストに基づき、対話の状態と次の話題を決定する。決定した話題から、「さつきさん」に何と発話させるのか、あるいは、音楽の伴奏を再生するのかも決定し、音声認識モジュール・発話伴奏モジュール・バーチャルエージェントモジュールにそれぞれのトリガーをかける。

バーチャルエージェントモジュールの画面例を図 3 に示す。画面に現れる女性が AI 音楽療法士の「さつきさん」である。

3.3　実　　装

本節では、「さつきさん」の実装について説明する。システムは JavaScript によるウェブアプリとして実装されており、PC・タブレット・スマートフォンのウェブブラウザから利用可能である。

音声認識モジュールは、まず音声区間検出（Voice Activity Detection; VAD）を行い、マイク入力中の音声が入力されている区間を切り出してから、外部の音声認識エンジンによる音声認識を実行する。VAD には「Silero VAD [21]」の JavaScript 実装 [22] を利用し、音声認識エンジンには OpenAI が開発した「Whisper [23]」を利用している。

発話伴奏モジュールは、指定の発話音声ファイルや伴奏ファイルを再生する機能を持つ。発話音声ファイルは、第 3.4 節に後述する SSML 形式のシナリオテキストを、Google Cloud の Text-to-Speech [24] を用いて音声合成して作成している。MIDI ファイルは、音源としては MIDI データの利用を想定しているが、現段階では対象となる楽曲の MIDI 音源から、キーの異なる音源を MP3 データの形式で予め複数作成し、伴奏再生時にキーを指定して、該当する伴奏音源を再生することができるように実装されている。

バーチャルエージェントモジュールは、「さつきさん」や仮想空間周囲の物体を表示するためのモジュールで、three.js [25] を利用している。「さつきさん」を含む仮想空間のオブジェクトは glTF 形式のファイルとして予め作成してあり、身体動作用のモーションも同じ glTF ファイルに含めている。バーチャルエージェントモジュールは、指定の身体動作用のモーションを再生する機能を持ち、対話シナリオ管理モジュールからのトリガーによりモーションを再生する。

対話シナリオ管理モジュールは、以上の各モジュールを制御する JavaScript プログラムとして実装されている。

3.4　シナリオ

さつきさんのシナリオは、音声合成マークアップ言語 SSML（Speech Synthesis Markup Language）で記載している。以下に実際のシナリオの一部を示す。さつきさんは、(1)〜(5) の各発話の後、対象者からの回答を待ち、回答を得た後は、音声認識を行って次の発話に進む。

（1）　こんにちは。

（2）　AI 音楽療法士の <s>「さつき」</s> です。5 月生まれなので名前は「さつき」と言います。今日、<s> 音楽 </s> をご一緒させて頂きますね 。<s> よろしくお願いします。</s> さて、今日のご気分はいかがですか？

（3）　そうですか。<break time=”2s”/> 私は朝ごはんをたくさん食べましたので <s> とても元気 </s> です。今日は、そちらは、どんなお天気ですか？

（4）　そうですか。<break time=”2s”/> 晴れ、曇、雨などありますが、どんなお天気がお好きですか？

（5）　そうなんですね。<break time=”1s”/> 私は晴れの

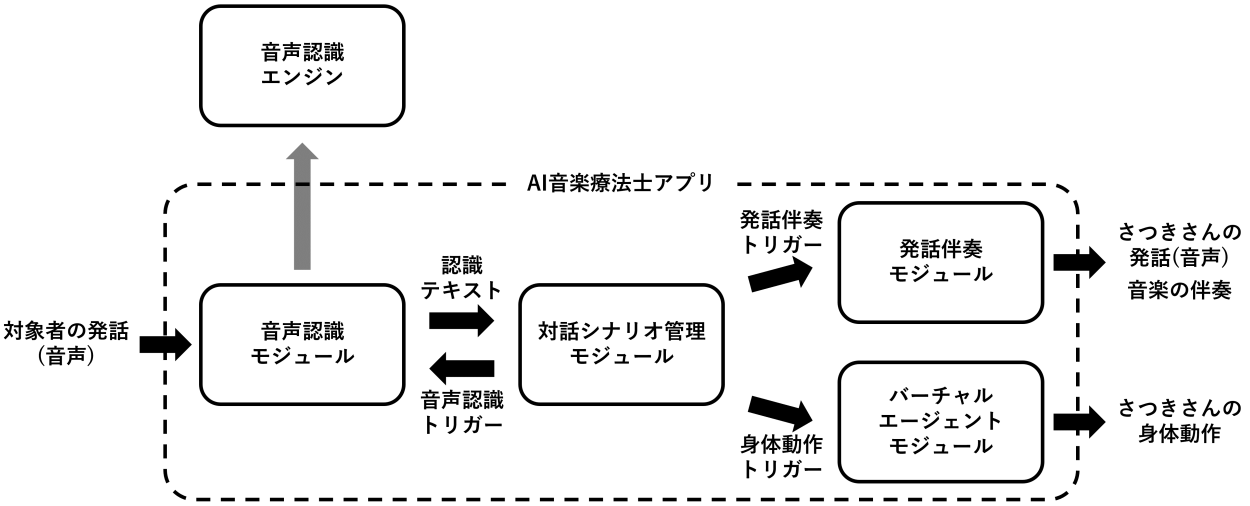


図 2　システム構成図

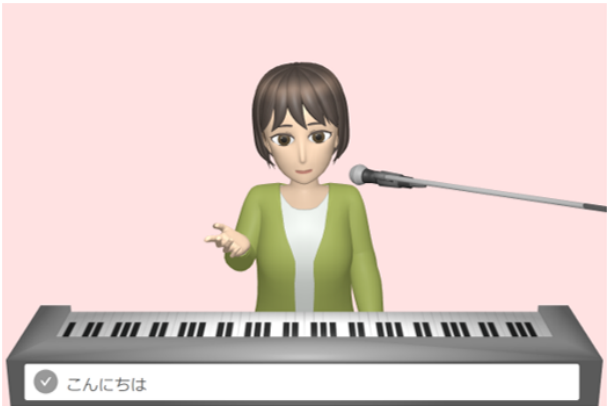



図 3　AI 音楽療法士の「さつきさん」

日が好きです。青空の中にふわふわの白い雲を見つけると、嬉しくなり、わたあめが食べたくなります。

4　実験と評価、および考察

「さつきさん」を、認知症高齢者を対象とした介護施設で従事する 5 名の介護スタッフに見てもらい、その動作や使用感についてヒアリングを行った。本章ではヒアリングで得られた指摘事項をまとめ、各指摘事項に対する考察と今後の進め方について述べる。

4.1　質 問 項 目

以下に今回のヒアリングの項目を列記する。

（1）　「さつきさん」との会話は楽しかったですか？「さつきさん」との会話についてご意見をお聞かせください。

（2）　「さつきさん」とどんなことを話してみたいですか？

（3）　「さつきさん」の伴奏による歌唱について、ご意見をお聞かせください。

（4）　「さつきさん」のどんなところが良いと思いましたか？特に認知症高齢者にとって良いと思うのはどのようなことですか？

（5）　「さつきさん」のどんなところが問題だと思いましたか？それに対する改善策としては何が考えられると思いますか？

（6）　「さつきさん」に今後、どのようなことを期待しますか？

（7）　将来的にどのような利用方法が良いと思いますか？「さつきさん」をどのように利用してみたいですか？

（8）　その他、ご意見・ご要望などございましたら、なんでも自由にお答えください。

これらの質問を通して得られた評価を以下にまとめる。

4.2　さつきさんとの会話について

さつきさんとの会話について、複数の人から得た評価は (1) もっと言ったことに反応して欲しい、(2) 名前を呼んでくれると良い、の 2 点である。言ったことに反応したり、名前を呼ぶためには、音声認識結果を形態素解析するなどして、受療者の発話から、主要な単語や名前を切り出す・認識する必要がある。特に、受療者を名前で呼ぶことは音楽療法の現場でも大切に考えられていることなので、できるだけ早く対象者の名前を聞いて復唱できるように修正したい。その上で、今後は音楽療法の現場で交わされる会話データを収集して、反応パターンのバリエーションを増やし、シナリオを充実させていきたいと考える。

さつきさんとの会話について、各介護スタッフから得られた良い評価は以下である。

（1）　自分の言ったことを広げてくれる感じが良いと思う。

（2）　身近な質問を選んでいるところは良い。

（3）　私はこういうのが好きですよ、などと言っていたのは親しみやすくて良かった。

身近な話題を選んでいることや、さつきさんからの質問に対する回答を受けて、さつきさんが自分のことを話すというやりとりも好意的に捉えられたようである。

一方、各介護スタッフから得られた指摘事項は以下である。

（1）　感情が足りない。事務的にしゃべっているように感じる。

（2）　共感するという感覚が足りない感じがする。「私もです」って言ってもらえると良い。

（3） 一方的に聞かれるだけではなくて、こちらからも話せるようになると良い。話している人が、まだ心を許せない感じがしてしまう。

（4） もっとコミュニケーションの取り方に時間をかけた方が良い。

これらの指摘に共通するキーワードは「共感」だと考えられる。もっと「寄り添ってくれている」という印象を与えられるような会話ができるように改良したい。

なお、「音楽のことを話題できると良いと思う」「だじゃれとか、ただの会話というよりも、笑いが出ちゃう感じだと楽しいと思う」といった、会話の内容に関するアドバイスも得ることができたので、これらの指摘はシナリオの改善に役立てたいと考えている。

4.3 さつきさんの声、見た目、動きについて

さつきさんの声や外見等について、複数の人から得た評価は「表情が乏しい」だった。認知症を含めて、脳に障害がある場合は、認知機能の中の情報処理機能が低下している可能性があるため、生身の人間の複雑な表情の読み取りが困難なケースがある。そのため、さつきさんの表情があまり豊かでは無い方が良いという可能性が考えられる。しかし介護スタッフによると、認知症高齢者は非常によくスタッフや周囲の人の表情を見ている、とのことだったので、今後はどの程度の表情の豊かさが適切であるか調査が必要だと考えている。

さつきさんの声については、「聞き取りやすい声だった」という良い評価も、「もっと抑揚があった方が良い」「語尾を伸ばしたりすると親しみやすいかもしれない」などといった改善点も指摘された。抑揚は聞き取りやすさに関連するので調整を進めていきたい。

一方、さつきさんの外見や動作については、上述の「表情が乏しい」以外は、「目もはっきりしていて良い」「拍手する動作が良い」「かわいくて髪型も良いと思う」など、総じて評価が高かった。今後は動作のバリエーションを増やしていきたいと考えている。

4.4 さつきさんの伴奏による歌唱について

さつきさんの伴奏による歌唱については、さつきさんも一緒に歌ってくれた方が良い、という指摘があった。ボーカロイド技術の利用を検討していきたい。

それ以外は、

（1） こちらのリクエストに応えてもらえると良い

（2） 地域ごとの季節の歌を聞ける・歌えると良い

（3） 歌詞も画面に出ると良いと思う

（4） これから、何が始まるんだろう続いていくんだろうといった、どきどき感とかわくわく感が足りない。今日は唱歌にしましょうかとか、ちょっと期待感を持てるようなのがいいと思う。

（5） 歌い終わったら「上手ですよ」と言ってくれると良いなど、曲目や歌詞の表示、歌う前と歌い終わった後のコミュニケーションに関する指摘事項だった。

歌唱は音楽療法セッションの中でも主要なアクティビティであるため、今後、様々な改善点やアイデアが出てくる可能性が高いと考えられる。ぜひ AI 音楽療法士ならではの歌唱アクティビティの提供に繋げられるように手を広げていきたい。

4.5 さつきさんへの更なる期待について

さつきさんへの更なる期待については、主に (1) さつきさん自身の発展、(2) さつきさんの利用方法、(3) さらに将来的な期待、の 3 つに分類できる。

4.5.1 さつきさん自身の発展について

「○○さんいいですね、お口開いてますね、いい笑顔ですねのような、さつきさんがこっちを見てくれている感じが出るといいと思う」や、「季節感があると良い。クリスマスだったら着るものを変えてみるなど。」といった意見を頂いた。

認知症高齢者の介護施設でのアクティビティでは、「見てくれている」と高齢者が感じられるようにすることに対する要望は高いので、妥当な指摘だと考える。表情認識技術の導入を検討したい。また、認知症高齢者は自力での外出が困難なため、屋内で過ごす時間が長く、季節感を感じるのが難しい状況である。音楽療法士は季節感のある曲を選ぶと共に、季節感のある服装をしたり、小道具を用意したりして、屋内で過ごす時間の長い療養者に、季節感を感じてもらえるように配慮するので、さつきさんもそのようなことができるように改良していきたい。

4.5.2 さつきさんの利用方法について

さつきさんの利用方法については多数の意見を頂いた。

（1） 利用者さんにとっては、（音楽療法の）機会が増えるのは良いこと

（2） 施設など、複数の利用者さんがいるところで会話できるようになると良い

（3） 日々のデイサービスの中で、毎日の朝の会とかで活用できるとよい。

（4） スタッフが記録を残したりしている間に AI が利用者さんの相手をしてくれるとよい

（5） 独居の高齢者の方が受け入れやすいと思う

音楽療法に接する機会が増えることで、毎日利用することが想定されたり、忙しい介護スタッフをサポートする形での利用など、こちらが思いつかない利用方法を聞くことができた。

複数の高齢者に同時に対応するような、高齢者のグループでの利用はかなり難しいと考えられるので、まずは受療者 1 人を対象とする設定で完成度を高めていきたいと考えている。その上で、人数分のパソコンがあれば、さつきさんがそれぞれの高齢者に対応するので、そのような形での利用であれば施設などでもサービス展開することが可能と考えられる。

4.5.3 さらに将来的な期待について

「違う人がでてくると、利用者さんにも刺激があってよい。」「利用者さんの反応の記録もさつきさんができるようになると良い」といった意見を聞くことができた。特に後者は、私達が目指している音楽療法のデジタルセラピューティクスに関連するものなので、研究の方向性が適切であることを確認することができた。

5 ま と め

本論文では、AI 音楽療法士の「さつきさん」の開発研究について説明した。AI 療法士の開発は容易ではないと考えられているが、我々は、音楽療法を「AI 音楽療法士と生身の音楽療法士の協業」と捉え、AI 音楽療法士が実施可能な業務を厳選してその完成度を高め、生身の音楽療法士との協業の形を新しく定義するという方針で研究を進めている。私達は AI 音楽療法士が実施可能な業務として、発話促進に特化した対話と歌唱伴奏に絞り込んだ。今後は 4 章に記載した認知症高齢者の介護に従事する介護スタッフによる評価に対する対応を中心に開発を進めると共に、さつきさんの利用データを収集し、認知症の重症度の客観評価や進行予測などに活かすための環境整備も進め、デジタルセラピューティクスとしての完成度を追究する予定である。そして、さつきさんと生身の音楽療法士との新しい協業の形を考案していく予定である。「さつきさん」により「会話に飢えている」と言われる高齢者が、いつでもどこでも何回でも気軽にさつきさんとの会話と歌唱を楽しみ、それによって認知症高齢者の BPSD が軽減し、それが介護者の負担軽減に繋がって、高齢者と介護者双方の QOL が向上することが期待される。

謝　　辞

AI 音楽療法士の開発に協力して下さった株式会社 dott の清水俊之介氏、AI 音楽療法士のシナリオ作成に協力して下さった NAT 音楽療法事務所の認定音楽療法士の相川直子さんに感謝いたします。評価実験にご協力いただいた、「福祉ネットワーク・やえやま」の當山房子先生と、「あかゆら」の介護スタッフの皆さまに感謝いたします。本研究は JSPS 科研費 JP20K11059, JP23H03407、新潟医療福祉大学学長裁量研究費、AMED の課題番号:JP21dk0207056、専修大学情報科学研究所共同研究助成の助成を受けました。

文　　献

[1] 「臨床精神医学」編集委員会編集, 精神科臨床評価検査法マニュアル, 2004
[2] 日本神経学会監修, 認知症疾患臨床ガイドライン 2017
[3] 内閣府, 令和 4 年版高齢社会白書, http://www8.cao.go.jp/kourei/whitepaper/w-2022/zenbun/04pdf_index.html
[4] 内閣府, 令和 5 年版高齢社会白書, http://www8.cao.go.jp/kourei/whitepaper/w-2023/zenbun/05pdf_index.html
[5] 一般社団法人 日本音楽療法学会 入会のご案内, https://www.jmta.jp/publish/pdf/information.pdf
[6] 一般社団法人 日本音楽療法学会, https://www.jmta.jp
[7] 認知症ちえのわ net, https://chienowa-net.com/
[8] 小杉尚子, 石井健太郎, 児玉直樹, AI 音楽療法士の開発研究–音楽療法のデジタルセラピューティクスにむけて–, 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023
[9] 小杉尚子，児玉直樹，清水幸子，數井裕光, 遠隔音楽療法サービスの開発研究, 情報処理学会論文誌, 60(1), 193-202, 2019
[10] 琉球新報, 音楽療法遠くてもＯＫ　ネット活用最新機器で音ずれず, https://ryukyushimpo.jp/news/prentry-210864.html
[11] 小杉尚子，児玉直樹，清水幸子，數井裕光, 認知症高齢者に対する遠隔音楽療法の効果, 日本遠隔医療学会雑誌, 15(2), 145-148, 2019, 優秀論文賞受賞
[12] 承認番号：30200BZX00271000, 管理医療機器 禁煙治療補助システム（JMDN コード：71087002）
[13] 小杉尚子, 小島明, 片岡良治, 串間和彦, 大規模音楽データベースのハミング検索システム, 情報処理学会論文誌 43(2), 287-298, 2002
[14] 小杉尚子, 櫻井保志, 山室雅司, 串間和彦, SoundCompass: ハミングによる音楽検索システム, 情報処理学会論文誌, 45(1), 333-345, 2004
[15] Kosugi N, Oshiyama C, Kodama N, Niwa S, Incorporating music therapy into cognitive remediation to improve both cognitive dysfunction and negative symptoms in schizophrenia, Schizophrenia Research, 204, 423-424, 2018
[16] Kosugi N, Oshiyama C, Kodama N, Niwa S, Introduction of Music Therapy Incorporated into Cognitive Remediation: A New Approach to Cognitive Dysfunction in Psychiatric Disorders and a Preliminary Report on Its Effects in Schizophrenia, Open Journal of Psychiatry, 9(1), 23-38, 2019
[17] Kosugi N, Oshiyama C, Kodama N, Niwa S, Predictability of a favorable outcome in schizophrenia associated with positive effects of “music therapy incorporated into cognitive remediation”, Schizophrenia Research, 238, 52-53, 2021
[18] 日野原重明監修, 篠田知璋・加藤美知子編集, 標準音楽療法入門 (上) 理論編, 春秋社, 1998
[19] 日野原重明監修, 篠田知璋・加藤美知子編集, 標準音楽療法入門 (下) 実践編, 春秋社, 1998
[20] 遠山文吉, 阪上正巳, 岡崎香奈, 中野万里子, 屋部操, 国立音楽大学音楽研究所音楽療法研究部門編著, 音楽療法の現在, 株式会社人間と歴史社, 2007
[21] https://github.com/snakers4/silero-vad
[22] https://github.com/ricky0123/vad
[23] https://openai.com/blog/whisper/
[24] https://cloud.google.com/text-to-speech
[25] https://threejs.org
[26] 東中竜一郎, 稲葉通将, 水上雅博, Python でつくる対話システム, 株式会社オーム社, 2020

一般発表 | Track 5: 高度なデータ利活用・ドメイン応用（医療情報、教育、地理情報等）

2024年2月29日(木) 16:00 ~ 18:10 | T5-B オンライン（Zoom Events)

ルート推薦

座長:小林 亜樹(工学院大学)
コメンテータ:角谷 和俊(関西学院大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T5-B-6-01] Happyスポットを考慮したウォーキング経路推薦手法の提案

*荒木 賢斗[1]、Siriaraya Panote[2]、角谷 和俊[3]、河合 由起子[1]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学、3. 関西学院大学)

16:20 ~ 16:45
[T5-B-6-02] 安全性を考慮した夜間の徒歩経路推薦システムの提案と評価

*横井 茉紘[1]、石原 靖哲[1] (1. 南山大学)

16:45 ~ 17:10
[T5-B-6-03] 公共交通機関の利用を想定した観光ルート推薦の一手法

*田中 俊[1]、上野 史[1]、太田 学[1] (1. 岡山大学)

17:10 ~ 17:35
[T5-B-6-04] 人流データに基づく観光スポットの周辺混雑度の予測と旅行計画への影響分析

*星 結人[1]、小竹 神[1]、北山 大輔[1] (1. 工学院大学)

17:35 ~ 18:00
[T5-B-6-05] ユーザの性格タイプの差異に基づく災害時行動促進情報理解の分析

*若杉 広介[1]、鈴木 優[2]、灘本 明代[3] (1. 甲南大学大学院自然科学研究科、2. 岐阜大学工学部、3. 甲南大学知能情報学部知能情報学科)

Happy スポットを考慮したウォーキング経路推薦手法の提案

荒木　賢斗† Panote　Siriaraya†† 角谷　和俊††† 河合　由起子† 中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学 情報工学・人間科学系　〒 606–8585 京都府京都市左京区松ケ崎橋上町
††† 関西学院大学 総合政策学部　〒 662-8501 兵庫県西宮市上ケ原一番町１－１５５
E-mail: †g2153056@cc.kyoto-su.ac.jp, ††spanote@kit.ac.jp, †††sumiya@kwansei.ac.jp,
†{kawai,nakajima}@cc.kyoto-su.ac.jp

あらまし　近年の健康ブームや，コロナ禍による運動不足により，ユーザのウォーキングに対する意識が向上傾向にある．しかし，ウォーキングに楽しさを見出せず継続を困難と感じるユーザもいると考えられる．そこで我々は，ユーザがウォーキングに対して前向きになれる経路推薦システムの開発を目指している．先行研究にて，グルメスポットの口コミ分析から評価を算出し，この値に基づいたウォーキング経路推薦システムの開発を行なった．しかし，口コミがない場所は考慮されていない．一方で，HappyButton を活用することにより，口コミのない場所であっても楽しさや幸せを感じるスポットを記録することができる．そこで本研究では，楽しさや幸せを感じるウォーキングを通じて，ユーザがウォーキングに対して前向きになれるように，Happy スポットを考慮したウォーキング経路推薦手法を提案する．

キーワード　ウォーキング支援，ルート推薦，Happy スポット

1. はじめに

近年の健康ブームによるダイエットや体力維持のため，ウォーキングに対する意識が高まっている [1] [2]．しかし，ウォーキングに楽しさを見いだせず，継続して実施することを困難に感じるユーザも少なくない [3]．歩行ルートを推薦するような既存の歩行者ナビゲーションシステムも存在するが，基本的にユーザが指定した目的地へ最短で移動するルートを推薦するものであり，ウォーキングそのものを楽しくさせる事を目的とするものではない．また，我々は先行研究 [4] で，グルメスポットおよび観光スポットの口コミを分析し，スポットの評価算出アルゴリズムを提案すると同時に，これに基づいて算出した評価値によるウォーキング経路推薦を行うシステムの構築を行った．しかし，口コミがない場所は考慮されていない．一方で，HappyButton [5] を活用することにより，口コミのない場所であっても楽しさや幸せを感じるスポットを記録することができる．

そこで本研究では，楽しさや幸せを感じるウォーキングを通じて，ユーザがウォーキングに対して前向きになれるように，HappyButton を活用して収集した Happy スポットを通るウォーキング経路推薦システムを提案する．

2. 関連研究

幸せに関する研究として，Asai らの HappyDB [6] がある．この研究では米国人が答えた幸せな瞬間を収集してデータベース化するという取り組みがなされている．データには，幸せな瞬間を説明したテキストやユーザの識別 ID などが含まれており，これらの瞬間が直近 24 時間以内と直近 3ヶ月以内に起きたものとして区別されている．

Seligman らの研究 [7] では，Three good things という毎晩寝る前にその日あった幸せな出来事を 3 つとその理由を書くエクササイズがある．このエクササイズを行うことで幸福度の向上と抑うつ度の低下が起き，この効果の持続性を報告している．

Quercia らの研究 [7] では，ソーシャルデータ（Flicker のメタ付き情報）を使用し，最短経路ではなく，ユーザが静かで，美しく，幸せに感じるであろうルートを推薦するというものである．この研究では，美しさ，静かさなどの全てのユーザに当てはまる感情をもとに推薦を行っている．

Johnson らの研究 [9] では，美しさを重視したシーニックルート，安全性を重視したセーフティルート，そして経路の簡単さを重視したシンプルルートを作成し，従来の経路推薦手法との比較を行っている．この研究は，それぞれ特定の要素を達成するための経路を作成している．

Joy らの研究 [10] では、歩行者の視点からエジプトのカイロの街並みを歩きやすい場所にするための要素を考察した．その結果，良い歩行場所には，利便性，安全性，清潔さが必要であることがわかった．そのため歩行には安全面や視覚的に認識できるものが動機となると結論づけている．

Daniele らの研究 [11] では，二枚の写真を基にどちらが好ましいのかを判定し，どの様な道を好むのかを機械学習させ，従来のナビの様に最短経路を推薦するだけでなく，綺麗に感じるルートを推薦している．

すなわち，本研究のように，実際に幸せを体験した記録をもとにした Happy スポットを通るウォーキング経路を推薦する取り組みの新規性は高い．

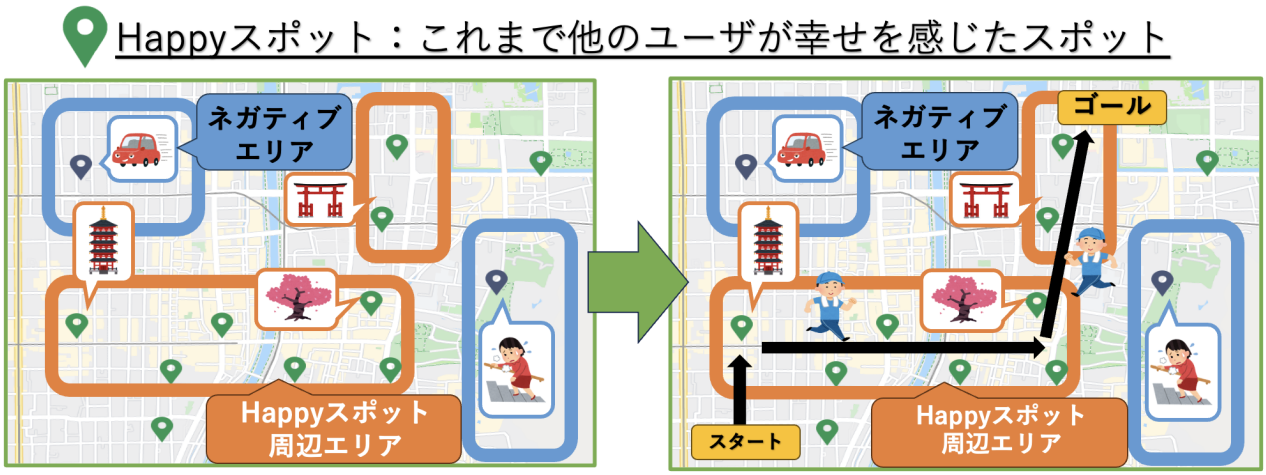


図 1　ウォーキング経路推薦の概要

3. 提 案 手 法

3.1 概　　要

図 1 は Happy スポットを考慮したウォーキング経路推薦の概要図である．本研究の提案手法では，HappyButton を活用し，ユーザが幸せを感じた時間と場所を収集する．そして，幸せを感じた場所を Happy スポットと定義し，収集された場所をマップ上に表示する．そして，Happy スポットや Happy エリア，ネガティブエリアを考慮して経路コストの更新を行い，ダイクストラ法を用いて経路を推薦する．

3.2 Three good things および Happy スポット

Three good things [7] とは，ポジティブ心理学を考案した心理学者 Seligman が提唱するエクササイズである．このエクササイズは毎晩その日の良かったこと 3 つとその原因を紙に書き留めるという習慣を 1 週間続けるというものである．

Happy スポットとは，ユーザが幸せを感じるであろうスポットのことである．経路推薦を行うために，Happy スポットを収集する必要がある．そこで，本研究では Happy スポットの収集に HappyButton システムを活用する．図 2 に HappyButton システムの概要図を示す．このシステムは，ユーザがその日幸せと感じる体験をした際に HappyButton を押すと，HappyButton からその体験をした時間と位置情報をスマートフォンのアプリに送信して記録を行う．HappyButton システムでは 1 日にボタンを押す回数を制限していないので，ユーザは幸せと感じた体験をするたびにボタンを押すことが可能である．そして，寝る前にスマートフォンのアプリの記録を参考にして，幸せ体験の詳細を入力する．本研究では，HappyButton が押された場所を Happy スポットとし，Happy スポットとそのスポットの情報を収集する．

3.3 経 路 推 薦

本研究では，楽しさや幸せを感じるウォーキング経路の推薦

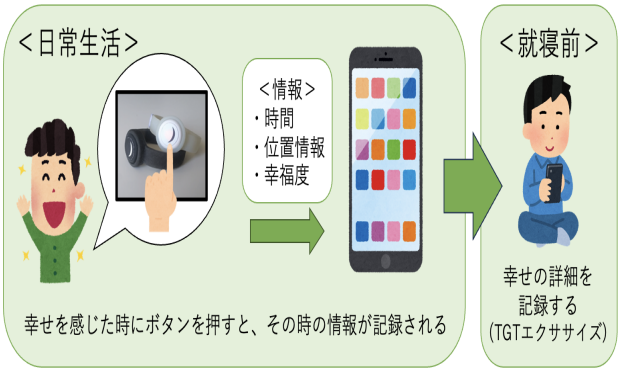


図 2　HappyButton システム

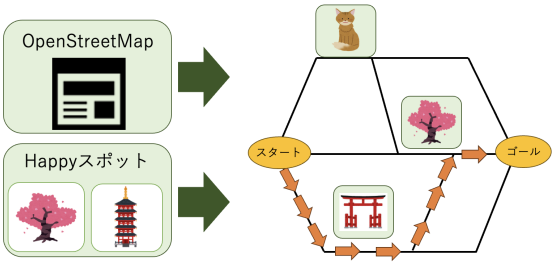


図 3　ダイクストラ法による経路推薦

を行う．使用データ例としてスポットの口コミや，ジオタグ付きツイートの内容，交通情報，事故データなどが挙げられる．本研究では，Happy スポットに着目し，それをもとに経路推薦を行う．図 3 に経路推薦の概要を示す．はじめに，OpenStreetMap から経路推薦を行う場所の歩道ノードを取得する．取得したデータには緯度経度情報の他に道路を識別する ID やノードの隣接判定ができるものが付与されており，それらを用いて道路の構築する．そして，Happy スポットをもとに道路のコストを更新する．その結果を用いてダイクストラを適用し，経路推薦を行う．

図 4　最短経路を表示した Google Maps

図 5　Happy スポットを通る経路を表示した Google Maps

4. 予 備 実 験

大学生 14 名を対象に，最短経路と Happy スポットを考慮した経路の比較とアンケート調査を行った．

4.1 経　　路

本実験で使用する経路は，出発地を京都市内の四条烏丸，目的地を京都市内の四条河原町とする．図 4 は最短経路を表示した Google Maps である．最短経路は平坦な経路であり，距離は 900m，ウォーキング時間は約 13 分である．図 5 は Happy スポットを考慮した経路を表示した Google Maps である．Happy スポットを考慮した経路は平坦な経路であり，距離は 1200m，ウォーキング時間は約 17 分である．経路を提示するにあたり，それぞれの経路をウォーキングするような動画を Google Street View を用いて生成した．

4.2 Happy スポット

Happy スポットは，出発地と目的地付近にあり，Google Maps Score [12] の値が高いスポットを選び出した．Google Maps Score とは，Google Maps に投稿されているスポットのレビュー値であり，0 から 5 の値を取る．このスコアは Google Maps に投稿されている情報のみを使用しているため，ユーザの好みは考慮されていない．本実験で使用する Happy スポットは，錦市場（Google Maps Score : 4.2）と寺町京極商店街（Google Maps Score : 4.0）とする．

4.3 評 価 方 法

まず，それぞれの経路において，5 段階のリッカート尺度により評価する．質問項目は下記の 5 つである．

質問 1 ：実際に歩きたいか
質問 2 ：風景や景色を楽しむことができたか
質問 3 ：瞬間的な驚きがあったか
質問 4 ：新しい発見や気付きがあったか
質問 5 ：行ってみたい場所や気になる場所があったか

とてもそう思うを 5 点，全くそう思わないを 1 点とし，平均点を算出し，比較を行う．

次に，A を最短経路，B を Happy スポットを考慮した経路とし，どちらの経路を歩きたいかを A または B で選択してもらう．集計後，割合を算出し，比較を行う．

4.4 実 験 結 果

表 1 は，リッカート尺度を用いたそれぞれの経路の評価結果である．A は最短経路，B は Happy スポットを考慮した経路である．5 つの質問においてそれぞれ平均点を算出した．算出した結果，全ての質問において，Happy スポットを考慮した経路の方が平均点が高い結果が得られた．

表 1　リッカート尺度を用いたそれぞれの経路の評価結果

	質問 1	質問 2	質問 3	質問 4	質問 5
A	2.4	2.7	2.1	2.0	2.8
B	3.9	4.2	3.7	3.6	4.1

表 2 は，どちらの経路を歩きたいかの選択結果である．14 名中 12 名が Happy スポットを考慮した経路を歩きたいと解答した．

表 2　どちらの経路を歩きたいかの選択結果

A	B
14.3%	85.7%

これらの二つのアンケートにより，Happy スポットを考慮した経路の方が最短経路より高い評価を得られた．

5. ま　と　め

本研究では，ユーザが楽しさや幸せを感じることで，継続的にウォーキングを行うための経路推薦手法の提案を行った．また，被験者実験により，提案手法の有効性を示す結果を得ることができた．

今後は，プロトタイプシステムの開発を行う．また，Happy スポットと呼ばれる幸せを感じるであろうスポットを収集し，複数の Happy スポットを通る経路推薦を行う．さらに，パーソナライゼーション手法の検討を行い，屋外での実証実験の実施を行う．

December 2021.

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22K12274, 22H03700, 22K19837) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] スポーツ庁 ”https://www.mext.go.jp/sports/b_menu/sports/mcatetop05/jsa_00010.html”

[2] 水野映子：”コロナ禍”としての運動不足”http://group.dai-ichi-life.co.jp/dlri/pdf/ldi/2020/wt2005b.pdf”，LIFE DESIGN REPORT 2020.5

[3] お客様生活文化研究所”https://www.asahigroup-holdings.com/company/research/hapiken/maian/bn/200611/00159.html”

[4] Katsuyuki Yamauchi, Da Li, Panote Siriaraya, Yukiko Kawai and Shinsuke Nakajima, ”Enhancing Walking Experience: A Walking Route Recommendation System Considering Nearby Spots”, International Workshop on Computing, Networking and Communications (CNC 2024), Feb. 2024.(to appear)

[5] Panote SIRIARAYA, Kenta SUZUKI, Wan Jou SHE, Ryotaro TANAKA, Da LI, Jin NARUMOTO, Noriaki KUWAHARA, Kazutoshi SUMIYA, Shinsuke NAKAJIMA, ”Palm Happiness: A Location-Based Application to Promote Awareness of Everyday Happy Moments”, The ACM International Conference on Mobile Human-Computer Interaction (ACM MobileHCI 2022), Article No.: 9, Pages 1–4, September 2022.

[6] Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy,Vivian Li, Andrei Lopatenko, Daniela Stepanov, YoshihikoSuhara, Wang-Chiew Tan, Yinzhan Xu, HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments,2018.

[7] Martin E. P. Seligman and Tracy A. Steen, NansookPark, Christopher Peterson. Positive psychology progress:Empirical validation of interventions, American Psychologist(60,410–421), 2005.

[8] Quercia, D., Schifanella, R., and Aiello, L. M. (2014,September). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In Proceedings of the 25th ACM conference on Hypertext and social media (pp. 116-125). ACM.

[9] I. JOHNSON，J. HENDERSON，C. PERRY，J. SCHÖNING，B. HECHT：”Beautiful...but at What Cost? An Examination of Externalities in Geographic Vehicle Routing”，Proceedings of the ACM on Interactive, Mobile，Wearable，and Ubiquitous Technologies，2017.

[10] Joy Maged, Rania Rushdy Moussa, Usama Konbr：”An Investigation into the Causes of Pedestrians' Walking Difficulties in Cairo Streets”, Architectural Engineering 41, 2022.

[11] Daniele Quercia，Rossano Schifanella，Luca Maria Aiello:”The Shortest Path to Happiness：Recommending Beautiful，Quiet，and Happy Routes in the City”，HT'14 Proceedings of the 25th ACM conference on Hypertext and social media，Pages 116-125，2014.

[12] Kenta Suzuki, Panote Siriaraya, Wan Jou She, Ryotaro Tanaka, Da li, Shinsuke Nakajima, ”HappyRec: Evaluation of a “Happy Spot” Recommendation System Aimed at Improving Mental Well-Being”, 2021 International Conference on Data Mining Workshops (ICDMW), pp.889-892,

安全性を考慮した夜間の徒歩経路推薦システムの提案と評価

横井　茉紘†　　石原　靖哲†

† 南山大学理工学部　〒 466–8673 愛知県名古屋市昭和区山里町 18
E-mail: 20sc108@m.nanzan-u.ac.jp, yasunori.ishihara@nanzan-u.ac.jp

あらまし　現在，地図アプリケーションでは最短距離による経路推薦が普及しているが，その他に様々な観点に基づく経路推薦手法も多く提案されている．土地の安全性を考慮した経路推薦に関する研究もその 1 つである．しかし，経路を推薦する時間帯を考慮した経路推薦手法は少ない．特に，夜間は見通しが悪い道路もあるため安全性を十分に考慮する必要がある．このような背景のもと，本研究では，夜間の安全性を考慮した徒歩経路推薦システムの提案と評価を行う．木梨らは，歩行者が夜間において経路を選択する際に距離と時間だけでなく明るさも考慮することを明らかにした．そこで，本研究では各地域に点在するスポットの営業時間を用いて経路の明るさを推定し，それに加えて警察が提供する不審者目撃情報を利用して経路推薦を行う．そして，推薦した経路が明るく，また安全性を考慮できているかを，それらの経路を夜間に歩行する被験者実験により評価する．

キーワード　地理データ活用，観光，ソーシャルメディア，徒歩経路推薦，夜間の安全性

1　はじめに

現在広く普及している Google Map などの地図アプリケーションは，一般に，距離や所要時間に基づいて経路を推薦する．その一方で，単に最短経路を推薦することを目的とせず，観光に特化した経路推薦手法 [1] [2] [3] など，様々な観点に基づく経路推薦手法が提案されている．土地の安全性を考慮した経路推薦手法に関する研究もその 1 つである．これには，ネガティブなスポットを避け，安全かつポジティブなスポットを推薦する手法 [4] や，犯罪・事故データをもとに経路推薦を行う手法 [5] などがある．しかし，これらの手法は，経路推薦を行う時間帯を考慮しておらず，夜間という限られた時間帯に起こり得る経路の見通しの悪さを加味していない．

このような背景のもと，本研究では，夜間における安全性を考慮した徒歩経路推薦システムの提案と評価を行う．木梨ら [6] は，歩行者が夜間において経路を選択する際に交通の観点だけでなく防犯の観点を考慮しているとし，距離や所要時間だけでなく明るさを考慮することを明らかにした．これより，明るさが夜間の安全性の指標になると捉えられる．夜間における道路の明るさに大きく影響を与えるものとして，飲食店やコンビニエンスストアなどのスポットによる明かりが考えらえる．実際に，図 1 のように，ほぼ同時刻でも営業時間中と営業時間外とで道路の明るさに大きな違いがある場所もある．そこで本研究では，Google Places API で取得できるスポットデータの営業時間を用いて経路の明るさを推定する．さらに，警察が提供する不審者の目撃情報データを利用し，ダイクストラのアルゴリズムによって安全性を考慮した経路推薦を行う手法を提案する．具体的には，経路推薦を行う時間帯にスポットが営業している場合には，スポットには明かりがありそのスポット付近の道路も明るさを保持していると考える．そして，営業中のスポット付近の道路をより通りやすくなるような経路推薦を行う．また，

(a) 営業時間中のスポット (2023 年 12 月 30 日 19 時ごろ撮影)　(b) 営業時間外のスポット (2024 年 1 月 8 日 19 時ごろ撮影)

図 1: スポットの営業時間による道路の明るさの違い (撮影場所：名古屋市中区栄)

警察が提供する不審者目撃情報も用いることで，危険性のある経路を通りづらくなるようにコスト定義を行う．そして，本研究で提案するシステムが推薦した経路が明るく，また安全性を考慮できているかについて被験者実験を行い評価する．これは，指定した本システムの推薦経路と最短経路の 2 経路を，夜間に歩いてもらい比較することで評価する．被験者実験の結果，本システムの有用性を確認できた．

なお，本研究は，南山大学研究審査委員会による「人を対象とする研究」倫理審査を受け，実施の承認を得ている（承認番号：23-080）．

2　関連研究

湯浅ら [7] は，位置情報付き写真に基づく景観ルート推薦システムを提案した．この研究では，位置情報付き写真のデータ数に基づき道路リンクに対してスコア付けを行い，スコア付き道路ネットワークに対してダイクストラ法を適用することで景観の良い経路を優先した経路推薦を行った．これに対し，本研究では Google Places API より取得するスポットの営業時間データと不審者目撃データをダイクストラのコスト定義に活用

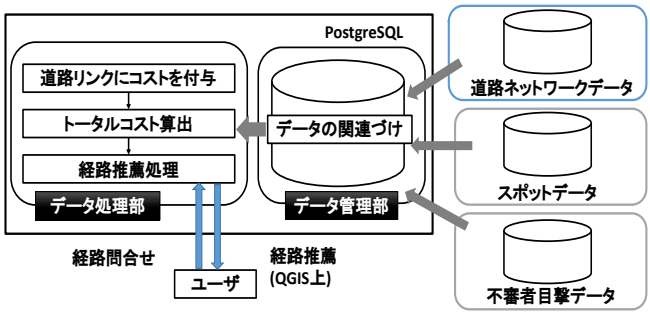


図 2: 夜間における徒歩経路推薦システム

する．また，山下ら [8] は Flickr の Exif 付き写真を用いて夜間の賑わいに基づいた経路推薦を行った．この研究では，夜間の定義を 19:00〜5:00 として Flickr の Exif 付き写真の明度から賑わいを定義し，それらをクラスタリングをすることで賑やかなエリアを抽出した．しかし，この抽出手法は経路推薦を行う時間帯を考慮しておらず，どの時間帯においても等しく推薦経路から賑わいが得られるとは限らない．そこで，本研究では Google Places API よりスポットの営業時間を取得することで，経路推薦を行う曜日や時間帯に応じた経路の明るさを推定し経路推薦を行う．さらに，不審者の目撃データを取得することで，夜間の安全性を考慮した経路推薦システムを提案する．

3　夜間における徒歩経路推薦システムの提案

この節では，本研究の徒歩経路推薦システムの提案手法について示す．まず，図 2 にシステムの概要を示す．

本研究の夜間における徒歩経路推薦システムはダイクストラによるアルゴリズムを用いる．そのためには，経路を形成する道路ネットワークデータと，コスト定義に必要なデータを取得する必要がある．

本研究では，名古屋駅から栄駅間を経路推薦の対象地域として定め，その地域の道路ネットワークデータを OpenStreetMap データにより取得する．また，夜間の安全性を考慮するために，夜道の明るさの指標として用いるスポットの営業時間データと，不審者目撃データをコスト定義に用いる．スポットデータは Google Places API より取得し，不審者目撃データは愛知県警察が提供する安心・安全マップ [9] より取得する．

そして，これらのデータを PostgreSQL 上に読み込み，スポットデータによる経路の明るさを表すパラメータ値と，不審者目撃データによる経路の危険性を表すパラメータ値を道路リンクに付与する．付与したこれらのパラメータ値と道路ネットワークデータより得られる各道路リンクの長さに基づいてコスト定義を行う．定義したコストに基づいて，PostgreSQL の拡張機能で経路推薦が行える pgRouting のうちダイクストラ法を適用する pgr_dijkstra を道路ネットワーク上で用い，経路推薦を行う．最後に，実行結果を GIS ソフトである QGIS 上で可視化する．

3.1　道路ネットワークデータ

道路ネットワークデータとは，道路ノードと道路リンクで構成される，主にナビゲーション向けのベクタデータである．道路ノードは交差点や曲がり角，行き止まりなどの特徴点を表し，道路リンクはノードとノードを結び，国道や県道といった道路種別などを表す線を表す．

表 1: 取得する道路ネットワークデータ

データタイプ	意味	データ型
gid	ID	integer
geometry	緯度・経度	geometry
distance	道路リンクの長さ	double precision
source	道路リンクの始点ノード	integer
target	道路リンクの終点ノード	integer

本研究では，道路ネットワークデータとして Geofabrik [10] から OpenStreetMap データを取得する．Geofabrik とは，国や地方単位で OpenStreetMap データを無料で提供するサイトである．本研究では，名古屋駅から栄駅の範囲で評価実験を行うため，中部地方のデータを取得する．ただし，取得した中部地方の地域全てのデータを経路推薦の度に読み込むことは無駄であるため，読み込む範囲を限定した上で，徒歩経路推薦に必要のない車道データなどの削除を行う．また，実際の道路との相違が生じている箇所があればその都度道路リンクを追加する．このような処理を行った後に，経路推薦に必要な道路ノードと道路リンクの関連と接続の管理をするためのトポロジー構築を行う．以上の処理により，経路推薦に利用可能なデータとなる．表 1 は，取得した道路ネットワークデータを表す．source と target は，各道路リンクを結ぶ 2 点のノードに割り当てた数値を格納する．これは，トポロジー構築の過程で付与する値であり，道路リンク間の関連付けに必要である．

3.2　スポットデータ

経路推薦において夜間の安全性を考慮するために，経路の明るさを表す指標として経路の周辺スポットの営業時間データを用いる．これは，経路推薦を行った夜間の時間帯にあるスポットが営業中であった場合に，そのスポットには明かりがありそのスポット付近の道路も明るさを保持していると考えられるためである．

3.2.1　Google Places API

Google Places API とは，Google Maps Platform の 1 つで Goole Map 上のデータを取得することができる API である．本研究では，Google Places API の中でも，特に Place Details API と Place Search API を用いる．Place Details API はある特定のスポットの詳細な情報を取得することができる API であり，表 2 に示すパラメータの値を指定することで取得する．ただし，パラメータの 1 つである place_id はスポットを一意に定めるパラメータであり，この place_id の取得には Place Search API を用いる必要がある．Place Search API は，表 3 に示すパラメータの値を指定することで place_id を含むその周辺のスポットの基本的なデータを取得することが可能である．

3.2.2　スポットデータの取得手順

Google Places API を用いたスポットデータの取得手順とし

表 2: Place Details API

データタイプ	意味
key	API キー
place_id	一意にスポットを定める ID
field	取得したいデータのタイプ

表 3: Place Search API

データタイプ	意味
key	API キー
location	場所情報を取得するための緯度・経度
radius	エリアの半径
type	場所の種類

表 4: 取得するスポットのデータ

データタイプ	意味	データ型
name	店舗名	char
geometry	緯度・経度	geometry
opening_hours	営業時間	time
rating	5 段階評価	double precision
user_ratings_total	評価したユーザの人数	integer

て，まずは Place Search API で place_id を取得する．本研究では，名古屋駅間から栄駅間のスポットデータを取得する．この範囲内のデータを取得するために，Place Search API でエリアの指定を行う．一度に取得できる place_id の数は限られているため，名古屋駅から栄駅間でいくつか location を指定しエリアを細分化して API を利用する．そして，本研究では，飲食店やドラッグストア，コンビニエンスストア，ホテルを各地域に点在しやすくかつ明かりを放ちやすいスポットとしてデータを取得する．そのため，場所の種類を定める type パラメータには，これら 4 つに該当するキーワードを指定する．

Place Search API によって place_id を取得した後，Place Details API を用いて表 2 に示すパラメータの値を指定し，より詳細なデータを取得する．取得したいデータのタイプを定める field パラメータには，表 4 に示すデータタイプを指定する．そして，Place Details API を用いて取得したスポットデータを，PostGIS を用いて PostgreSQL データベースに読み込む．PostGIS とは PostgreSQL の拡張モジュールであり，PostgreSQL データベース上に緯度・経度による位置情報をもったデータを格納することを可能にする．

本研究では，以上の処理により，名古屋駅から栄駅間のスポットを 505 個取得した．

3.3 道路ネットワークデータとスポットデータの関連付け

取得したスポットデータは，経路推薦のコスト定義に必要となる．そしてコスト定義に用いるためには，ダイクストラ法を適用する際のエッジにあたる道路リンクとスポットデータを関連づける必要がある．これにより，経路推薦を行った時間帯と，スポットデータテーブルに格納される各スポットの営業時間データに応じて，各道路リンクのコストを決定することを可能にする．

図 3: 名古屋駅から栄駅間の不審者目撃情報領域

この関連付けは，PostgreSQL データベース上に読み込んだ，道路ネットワークデータとスポットデータの位置情報を比較して実現する．まず，各スポットデータについて，そこからの距離が 30m 以内の道路リンクを求める．そして，道路ネットワークテーブルにおいて，求まった道路リンクのレコードにそのスポットを一意に表す ID を格納することでデータの関連付けを実現する．各道路リンクに関連付けるスポットの数は 5 つまでとする．

3.4 不審者目撃データ

スポットデータを用いることによって，経路の明るさを推定する他に，過去に不審者が目撃された領域を不審者目撃データとして扱うことで，危険性のある経路を避けるような経路推薦を行う．本研究では，愛知県警察が提供する安全・安全マップより，2022 年と 2023 年 11 月時点の名古屋駅から栄駅間の不審者の目撃領域を活用する．そして，この領域内の道路リンクを不審者目撃データとして取得する．この取得には，QGIS を用いる．

まず，安心・安全マップ上に表示される不審者の目撃情報が存在する場所と同様の領域に QGIS でポリゴンを作成する．図 3 には，ポリゴンで作成した名古屋駅から栄駅間における不審者目撃領域を示す．次に，PostgreSQL データベース上にある道路ネットワークテーブルをレイヤとして QGIS 上に表示する．そして，作成したポリゴンに道路リンクが含まれる，または交差する場合に属性の空間結合を行う．これにより，不審者が目撃された道路リンクを取得することができ，その道路リンクのコストを高く設定することで通りづらくすることが可能になる．

3.5 道路リンクのコスト定義

取得したスポットデータと不審者目撃データに基づいて安全性と距離を考慮したコスト定義を行う．本研究では道路リンクのコスト TC を定義する式として以下の式を用いる．

$$TC = D \times \prod_{i=1}^{5} TS_i \times SP$$

ここで，D はその道路リンクの長さ，TS_i はその道路リンクに関連付けられた i 番目のスポットにより定まる明るさの評価値，SP は不審者目撃データより定まる評価値である．

各 TS_i は，経路推薦を行う時間とスポットの営業時間を比

較して決定される値である．スポットが営業時間中の場合は，スポットの 5 段階評価により定める値 RE と，スポットを評価した総人数により定める値 TU の積 $TS_i = RE \times TU$ と定義する．ただし，$0 < TS_i < 1$ となるように RE と TU を定める．そして，スポットが営業時間外の場合は，$TS_i = 1$ とし，道路リンクに対して 1 つも関連付くスポットがない場合は，$\prod_{i=1}^{5} TS_i = 1$ とする．不審者目撃領域に依存する値 SP は，道路リンクが不審者目撃領域内であれば $SP > 1$，領域外であれば $SP = 1$ とする．

4　コスト定義式におけるパラメータ値の検討

3.5 節のコスト定義式に基づいて経路推薦を行うには，TS_i を定めるためのパラメータである RE および TU と，SP の具体的な値を定める必要がある．これらの値によって推薦される経路が決まるが，本節は，それらをどのように定めたかについて述べる．また，パラメータの値で道路リンクのコストの増減の度合いを変えることの他に，経路推薦を行う曜日や時間帯の違いによっても異なる推薦経路が得られる場合がある．これは，経路推薦時にスポットが営業中であれば $TS_i = RE \times TU$，そうでなければ $TS_i = 1$ をとるためである．そこで，まずパラメータの定め方とパラメータ値の違いによる経路の違いについて示し，その後に経路推薦を行う曜日と時間帯の違いによる経路の違いを示す．

4.1　パラメータ値による推薦経路の違い

まず，TS_i は営業中のスポットが関連付く道路リンクのコストを低減する値である．$0 < TS_i < 1$ の範囲内で，0 に近い値の場合，コストを大幅に低減する一方で，1 に近い値の場合，前者と比べてコスト低減への影響は少ない．そこで，RE の値を固定し，コスト低減への影響が大きい場合と小さい場合の 2 通りの TU の値を用いて，推薦される経路にどのような違いが生じるかを比較・検討した．表 5 に RE の値を，表 6 には TU の値のうち一方を $TU1$，もう一方を $TU2$ として示す．RE と TU の具体的な数値について，まず RE は，スポットの 5 段階評価を 0.5 点単位で階級分けし，それぞれの階級ごとに値を定めた．TU については，スポットを評価した総人数を基本的に 50 人単位で階級分けし，それぞれの階級ごとに値を定めた．ただし，本研究で取得したスポットデータのうち評価した総人数が 300 人以上であるスポットは取得した全てのデータのうちの 1/4 以下で，1〜50 人であるスポットは全体の 1/4 以上を占めた．また，全てのスポットデータうち，スポットを評価した総人数の中央値は 140 人であった．このように，取得したスポットデータのうち，評価した総人数は全体的に少ない傾向であったことから，評価の総人数 300 人以上のスポットは，相対的に多くのユーザに評価されたと判断できる．これより，評価した総人数が 300 人以上である場合は，50 人単位で階級を区切ることなく値を定めることにした．そして，定めた TU のうち，$TU1$ と $TU2$ では $TU2$ の方が階級ごとの値が小さく，よりコスト低減への影響が大きい．

表 5: RE の値

5 段階評価	RE
1.0 〜 1.5	1.14
1.6 〜 2.0	1.12
2.1 〜 2.5	1.10
2.6 〜 3.0	1.08
3.1 〜 3.5	1.06
3.6 〜 4.0	1.04
4.1 〜 4.5	1.02
4.6 〜 5.0	1.00

表 6: TU の値

評価の総人数	$TU1$	$TU2$
1 〜 50	0.85	0.26
51 〜 100	0.82	0.25
101 〜 150	0.79	0.24
151 〜 200	0.76	0.23
201 〜 250	0.73	0.22
251 〜 300	0.70	0.21
301 〜 500	0.67	0.20
501 〜	0.64	0.19

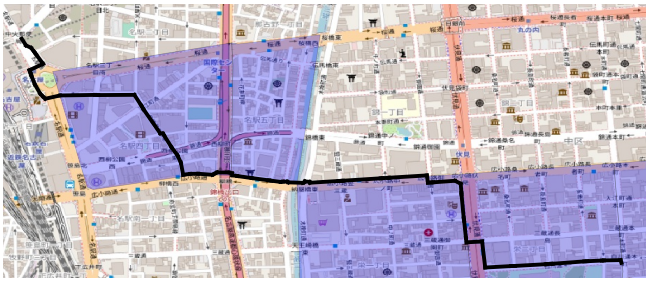


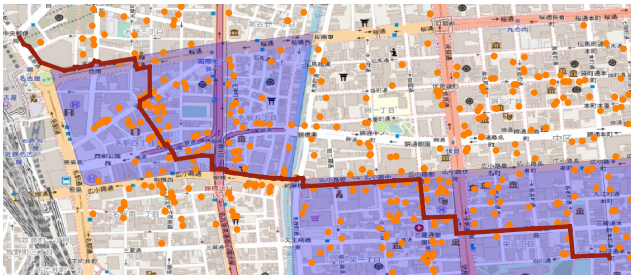
図 4: 距離のみに基づく経路 A

次に，SP は不審者目撃領域内の道路リンクのコストを増加させる値であり，$SP > 1$ の範囲内で定める．本研究では，2 通りの TU に組み合わせる SP の値として 1.3，1.6，1.9 の 3 通りを比較・検討した．その結果，$TU1$ には $SP = 1.3$ を，$TU2$ には $SP = 1.9$ を組み合わせて用いることとした．以下，TU と SP のパラメータ値の組み合わせによる経路の違いを詳細に示す．ここで，比較対象の経路として，距離のみに基づく経路 A を図 4 に示す．図中のブルーのポリゴンは，不審者目撃領域を表す．

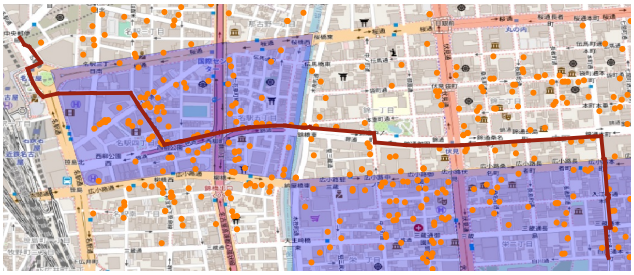
4.1.1　$TU1$ を用いた経路推薦

まず，$TU1$ に 3 通りの SP を組み合わせた場合，$TU1$ かつ $SP = 1.6$ と $TU1$ かつ $SP = 1.9$ では同経路が得られる傾向にあった．そのため，ここでは $TU1$ かつ $SP = 1.3$ と $TU1$ かつ $SP = 1.6$ の経路を示す．ただし，不審者目撃データを考慮しなかった場合，すなわち $SP = 1$ とした場合の経路を図 5a に経路 B として示す．そして，図 5b に $SP = 1.3$ とした経路 C を，図 5c に $SP = 1.6$ とした経路 D を示す．これらは水曜 20 時を条件に推薦した経路である．オレンジのポイントは水曜 20 時以降に営業しているスポットデータを表す．

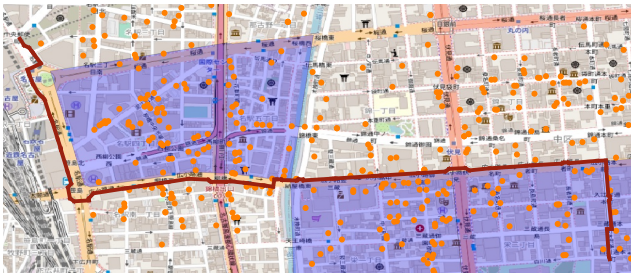
経路 B の全経路長当たりのスポットが関連付く経路長は 56.7%，経路 C は 51.8%，経路 D は 30.1%となり，$SP = 1.3$ と $SP = 1.6$ では $SP = 1.3$ の方がスポットの関連長は長くなった．また，全経路長当たりの不審者目撃領域内の経路長は経路 B が 90.7%，経路 C が 49.8%，経路 D は 19.6%となり，不審者目撃データを考慮していない経路 B は経路長のほとんどが領域内を通った．一方で，$SP = 1.3$ と $SP = 1.6$ では $SP = 1.6$ の方が不審者目撃領域を通りづらい．ただし，$SP = 1.3$ の方が全経路長は短く，全経路長当たりの不審者目撃領域長は最短経路の 73.8%に対して 24.0%減少している．これは，時間と曜日の条件を変えた場合でも同様の結果が得られた．以上より，



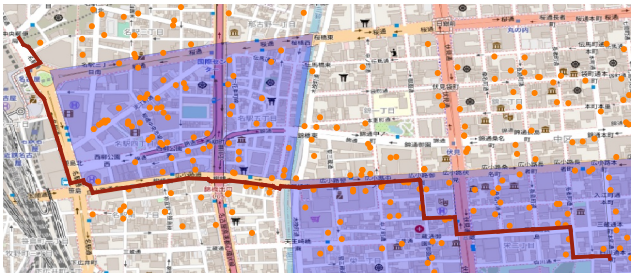
(a) $SP = 1$ とし，水曜 20 時を条件に推薦した経路 B



(b) $SP = 1.3$ を用い，水曜 20 時を条件に推薦した経路 C

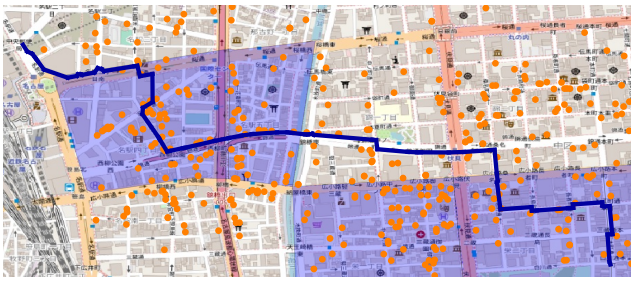


(c) $SP = 1.6$ を用い，水曜 20 時を条件に推薦した経路 D

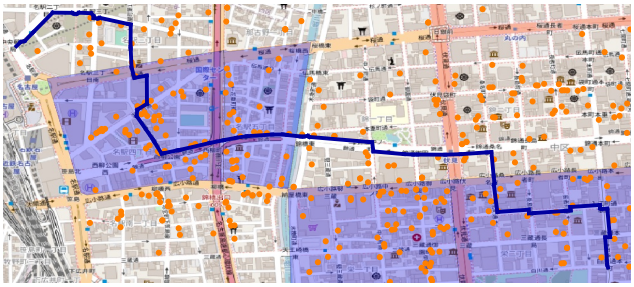


(d) $SP = 1.3$ を用い，日曜 20 時を条件に推薦した経路 E

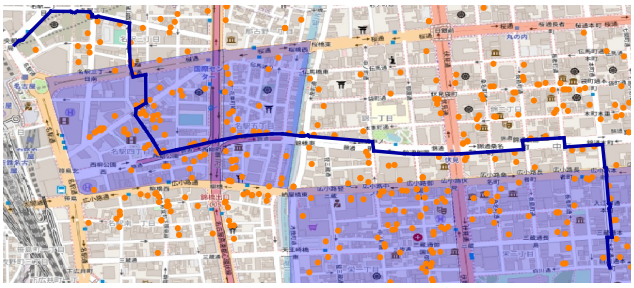
図 5: $TU1$ を用いた場合のシステム推薦経路



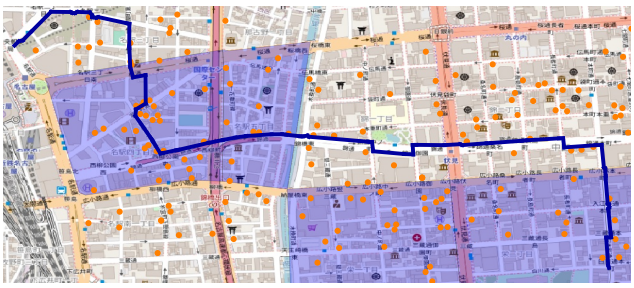
(a) $SP = 1$ とし，水曜 20 時を条件に推薦した経路 F



(b) $SP = 1.6$ を用い，水曜 20 時を条件に推薦した経路 G



(c) $SP = 1.9$ を用い，水曜 20 時を条件に推薦した経路 H



(d) $SP = 1.9$ を用い，水曜 23 時を条件に推薦した経路 I

図 6: $TU2$ を用いた場合のシステム推薦経路

TU1 には $SP = 1.3$ を組み合わせて用いる．

4.1.2 $TU2$ を用いた経路推薦

$TU2$ に 3 通りの SP を組み合わせた場合は，SP の値の違いによる全経路長当たりのスポットが関連長に大きな差はなく，全経路長当たりの不審者目撃領域長は SP の値が大きくなるほど短くなる傾向にあった．これより，$TU2$ かつ $SP = 1.6$ と $TU2$ かつの経路 $SP = 1.9$ を示す．ただし，$TU1$ と同様に不審者目撃データを考慮しない，$SP = 1$ とした場合の経路を図 6a に経路 F として示し，図 6b に $SP = 1.6$ とした場合の経路 G を，図 6c に $SP = 1.9$ とした場合の経路 H を示す．これらの経路は水曜 20 時を条件に推薦した経路である．

経路 F の全経路長当たりのスポット関連長は 67.2%，経路 G が 68.7%，経路 H が 64.6%となり，SP の値による大きな値の差は見られなかった．また，全経路長当たりの不審者目撃領域長は経路 F は 69.3%，経路 G は 57.4%，経路 H は 44.3%であり，SP が増加するほど領域内を通る経路は短くなった．SP の値の違いによるこれらの結果は時間と場所を変えても同様であった．これより，$SP = 1.9$ は，$SP = 1.6$ に比べてスポット関連長に大きな差を与えることなく，不審者目撃領域長を短くすることができることから，$TU2$ には $SP = 1.9$ を組み合わせて用いる．

4.2 曜日や時間帯による推薦経路の違い

4.1.1 節で定めたパラメータ値を用いて，曜日や時間帯による推薦経路の違いを示す．

まず，曜日による推薦経路の違いを示す．パラメータ値には，RE かつ $TU1$ かつ $SP = 1.3$ を用いた．そして，図 5b の水曜 20 時に推薦した経路 C と比べて，図 5d は日曜 20 時に推薦した経路 E を示す．オレンジのポインタはそれぞれ経路推薦を行った曜日の 20 時以降に営業しているスポットを示している．同じパラメータ値でも，経路推薦を行う曜日を変えることで営業中のスポットが変化し，異なる経路を推薦する．

次に，時間帯による推薦経路の違いを示す．パラメータ値には，RE かつ $TU2$ かつ $SP = 1.9$ を用いた．そして，図 6c の水曜 20 時に推薦した経路 H と比べて，図 6d は水曜 23 時に推薦した経路 I を示す．オレンジのポインタはそれぞれの経路推薦を行った時間帯以降に営業しているスポットを示している．経路 H と経路 I で大きな違いはないものの，経路推薦を行う時間帯に応じて異なる経路を推薦する．

5 評価実験

5.1 実験方法

本研究で提案する徒歩経路推薦システムが夜間の安全性に配慮した経路を推薦できているかについて，被験者実験を行い評価する．実験方法として，被験者 16 名を対象に日没から 20 時までの間の時間帯で指定した 2 つの経路を歩いてもらい，2 つの経路の比較評価とそれぞれの経路に対する評価を行った．2 つの経路のうち，一方は本研究のシステムが推薦する経路であり，もう一方は最短距離に基づく経路である．経路は，出発地と目的地に御園座，丸の内駅付近を指定し，その間の経由地として下園公園を通るように設定した．被験者には，どちらが本研究のシステムによる推薦経路であるかを事前に知らせない．その上で，道の明るさや安全性，そしてそれらの要因に着目しながら経路を歩いてもらい，安全性を評価する．なお，本研究で提案する経路推薦システムは，経路推薦を行う時間帯や曜日によって異なる経路が得られる場合がある．そのため，本実験の経路には全ての被験者が行う時間帯や曜日といった条件の変化に伴って経路が変わらないよう出発地，目的地，経由地を選択した．さらに，4 節で述べたようにコスト定義におけるパラメータ値の違いによって，異なる推薦経路が得られる．そこで，実験 1，実験 2 として 2 つの異なるシステム推薦経路を用いて評価を行う．

まず，実験 1 では被験者として 20 代女性 5 名，20 代男性 5 名を対象に経路の評価を行った．図 7 に距離のみに基づく最短経路 A を，図 8 にシステム推薦経路 B を示す．なお，経路 B は，表 5 の RE と表 6 のうち $TU1$，そして $SP = 1.3$ を用いて推薦した．

実験 2 では，被験者として 20 代の女性 6 名を対象に経路の評価を行った．最短経路には，実験 1 と同様に経路 A を用い，システムが推薦する経路には図 9 に示す経路 C を用いる．経路 C は，実験 1 と同様の RE と，表 6 のうち $TU2$，$SP = 1.9$ を用いて推薦した．さらに，表 7 に，実験に用いた 3 経路の全経路長，スポット関連長，不審者目撃領域長を示す．3 経路の全経路長は経路 A，経路 B，経路 C の順で約 100m ずつ長くな

図 7: 距離のみに基づく経路 A

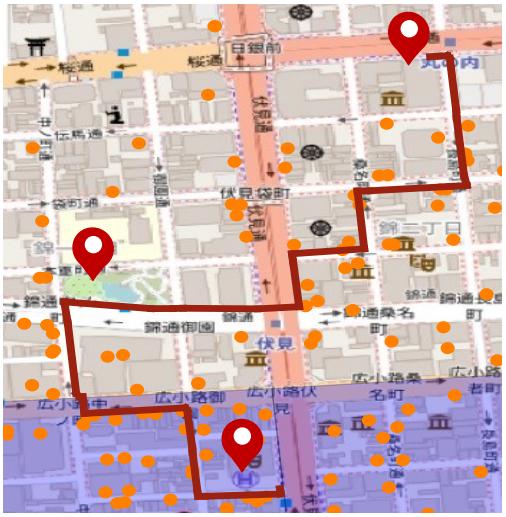


図 8: 実験 1 におけるシステム推薦経路 B

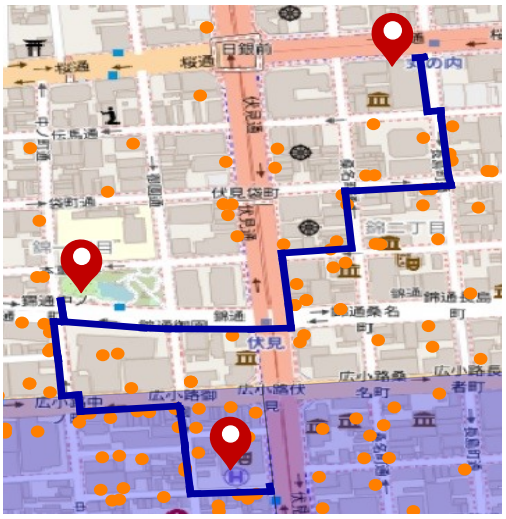


図 9: 実験 2 におけるシステム推薦経路 C

表 7: 実験で用いた 3 経路の全経路長，スポット関連長，不審者目撃領域長

経路名	全経路長	スポット関連長	不審者目撃領域長
経路 A	1280.7m	270.4m (21.1%)	167.1m (13.0%)
経路 B	1381.7m	745.1m (53.9%)	374.7m (27.1%)
経路 C	1485.0m	1086.0m (73.1%)	374.7m (25.2%)

る．その一方で，全経路長当たりのスポット関連長は，経路 A，経路 B，経路 C の順に長くなる．このように，経路 C は経路 B よりもわずかに遠回りするが，よりスポット付近を通る経路である．また，全経路長当たりの不審者目撃領域長は，経路 A が最も短く，経路 B と経路 C は同程度である．

5.2 評価結果

経路の評価は，被験者に歩行した経路の安全性に関する設問に回答してもらうことで行う．この回答方式には 2 通り用いる．1 つは，システム推薦経路と最短経路を 2 択の回答とし，相対的に経路を比較評価する方法であり，もう一方は，5 段階のリッカート尺度 (そう思う，どちらかと言えばそう思う，どちらとも言えない，どちらかと言えばそう思わない，思わない) を用いて評価する方法である．そして，リッカート尺度による評価は 5 段階の尺度のうち，そう思わないを 1 点，どちらかと言えばそう思わないを 2 点，どちらとも言えないを 3 点，どちらかと言えばそう思うを 4 点，そう思うを 5 点，というように選択肢を点数化して行う．

5.2.1 実験 1 の評価結果

最短経路 A とシステム推薦経路 B を用いた実験 1 の評価結果について示す．図 10 は，経路 A と経路 B を 2 択の回答とし，経路の比較評価を行った結果である．経路の目的地までのわかりやすさは，経路 A と B で同程度の評価であった．また，経路の総合的な明るさ，店舗照明による明かり，賑わいについては経路 B がわずかに経路 A よりも上回る結果となった．そして，どちらの経路がより安全に配慮されているかについては，10 名中 4 名が経路 A，10 名中 6 名が経路 B と回答した．これは経路 A とわずかな差であり，推薦した経路 B の方がより安全であるとは言えない．

5.2.2 実験 2 の評価結果

次に，最短経路 A とシステム推薦経路 C を用いた実験 2 の評価結果について示す．図 11 は，経路 A と経路 C を 2 択の回答とし，経路の比較評価を行った結果である．経路の目的地までのわかりやすさは，実験 1 と同様に経路 A と経路 C は同程度の評価を得た．そして，どちらの経路がより経路の総合的な明るさがあるか，安全に配慮されているか，という 2 つの設問に対して 6 名の被験者全員が経路 C と回答した．以上の結果より，経路 C の方が経路 A よりも総合的に明るく，安全な経路だと言える．

次に，5 段階評価による経路 A，経路 C それぞれの評価結果を図 12 に示す．被験者は 6 名であるため，各設問の満点を 30 点として評価点を算出する．経路の明るさを感じた要因として，経路 A と経路 C ともに店舗照明による明るさが一番高い評価

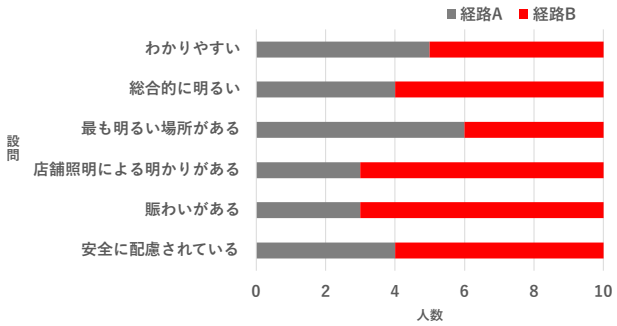


図 10: 実験 1 における経路 A と経路 B の比較評価

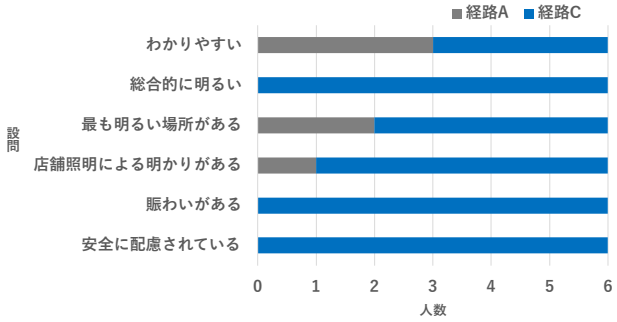


図 11: 実験 2 における経路 A と経路 C の比較評価

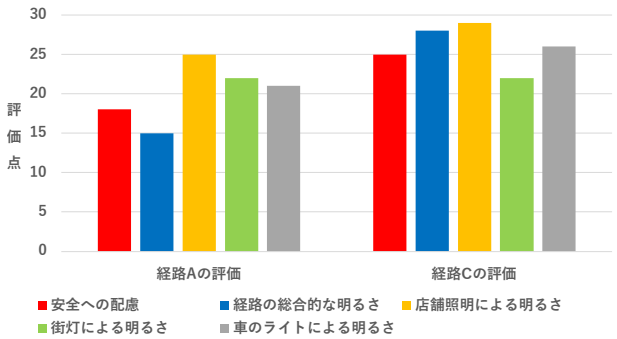


図 12: 実験 2 における経路 A と C の評価

を得た．また，3 つの明るさの要因の評価は経路 A と経路 C で評価点に大きな差はなかったのに対して，安全への配慮と総合的な明るさの評価点は差がつく結果となった．これは，安全への配慮の評価で 5 つの尺度のうち何故その回答を選択したのか理由を記述してもらったところ，経路 A は暗い道があったため，という記述や具体的に暗いと感じた場所の記述がみられた．このことから，経路 A の安全への配慮や総合的な明るさの評価点が低くなったと考えられる．

5.3 考　　察

被験者実験の結果，少しでも暗い道を通る経路は，5 段階のリッカート尺度の評価において安全への配慮と経路の総合的な明るさで低い評価点を得る傾向にあった．これより，ある明るさ以上を保持する道のみを通るような経路を推薦する必要があると考えられる．

また，被験者 16 名それぞれについて，比較評価において，より総合的に明るい経路とより安全に配慮された経路に同一経路を回答した．これより，経路の安全性と総合的な明るさの評価

には，正の相関があると考えられ，[6] の報告内容とも矛盾しない結果となった．

さらに，システム推薦経路において，明るさを感じたのは店舗照明によるものであるか，という設問に対して実験 1，実験 2 ともに，そう思う，どちらかと言えばそう思うの 2 つ回答のみであった．この結果より，取得したスポットデータは経路の明るさとして十分に反映されたと言える．また，賑わいを感じる経路について，図 10 より実験 1 では 10 名中 7 名が経路 B と回答し，図 11 より実験 2 では，6 名の全ての被験者が経路 C と回答した．これより，推薦経路システムの構築に用いたスポットデータは経路の明るさだけでなく，賑わいの指標にもなり得ると考えられる．

なお，本実験は 12 月中旬から 1 月上旬に行ったため，イルミネーションによる季節限定の明るさが存在していた．そのため，経路の評価にはイルミネーションの明かりも少なからず反映されていると考えられる．

6 まとめと今後の課題

6.1 ま と め

本研究では，スポットの営業時間データから経路の明るさ，不審者目撃領域から危険性の推定を行うことで，夜間の安全性を考慮した経路推薦システムを提案し，その評価を行った．経路推薦にはダイクストラのアルゴリズムを利用し，推定する経路の明るさや危険性を道路リンクのコストとして算出した．これにより，営業中のスポットを通りやすく，不審者目撃領域を通りづらくなるような経路推薦を行った．評価結果として，実験 1 では最短経路とシステム推薦経路の安全性に関する評価は同程度であり，システム推薦経路がより安全であるとは言えない結果となった．ただし，実験 1 からコスト定義式におけるパラメータ値を変えて行った実験 2 では，被験者 6 名全員がシステム推薦経路がより安全に配慮されていると回答した．これより，わずかに遠回りしてもスポット付近を通ることを優先する方が最短経路と比較してより安全に配慮された経路だと言える．以上の結果より，経路推薦のコスト定義に $TU2$，$SP = 1.9$ を用いた場合に，本システムは有用であると言える．

6.2 今後の課題

本研究では，スポットデータを明るさの指標としてスポットデータから 30m 以内の道路リンクに対してスポットを関連づけた．このスポットデータを関連づけるフェーズでは，スポットデータが道路リンクのどの付近に位置しているのかを考慮していない．つまり，スポットデータが道路リンクに対して端に位置している場合も，中心に位置している場合も同様のコストとなり，端だけ明るく全体的に暗い経路も明るさを保持している経路としてコストを低減してしまう．これより，スポットデータが道路リンクのどこに位置しているのかを考慮してコストを与えることでより明るさを保持する経路を得られると考える．

また，本研究で提案するシステムは営業中のスポットデータから経路の明るさを推定し，スポット付近を通りやすくするような経路を推薦する．しかしながら，評価結果より，暗い道を避け，ある明るさ以上の経路を推薦することが望ましいと考えられることから，暗い区間を通るのを最短化するような経路を求める手法でも，夜間の安全性を考慮できると考える．

さらに，本研究で経路の危険性の推定に用いた不審者目撃データは，2022 年と 2023 年 11 月時点のデータから取得しており，リアルタイム性に欠けている．また、不審者が目撃された領域内のうち，具体的にどの地点で目撃されたのかについても，曖昧である．そのため，リアルタイムに不審者が目撃された地点を得ることができれば，経路の危険性をより正確に推定できると考える．

最後に，本研究ではスポットの営業時間を用いて安全性を考慮した徒歩経路推薦システムの提案を行った．しかし，安全と感じる基準には個人差があり，人通りの多さや街灯の多さ，道幅，歩きやすさなど個人によって判断基準も様々である．そのため，ユーザが夜間に歩く際に重視する条件に応じて，ユーザごとのカスタマイズを可能にすることでより個人に特化した安全な経路推薦が行えると考える．

謝　　辞

不審者目撃情報の利用について示唆くださった南山大学理工学部河野浩之教授に深謝します．また，発表を通じて貴重なコメントをくださった皆様に深謝します．本研究の一部は JSPS 科研費 JP19K11912，JP23K17456 の助成を受けたものです．

文　　献

[1] 角上直哉, 新妻弘崇, 太田学. 観光スポットの訪問目的を考慮した観光ルート推薦の一手法. 第 12 回データ工学と情報マネジメントに関するフォーラム，J1-1, 2020.
[2] 磯田祥吾, 日高真人, 松田裕貴, 諏訪博彦, 安本慶一. 観光ナビにおいて必見スポットをよりよい時間帯に訪問するオンサイトプランニング手法. マルチメディア，分散，協調とモバイル (DICOMO2020) シンポジウム, pp. 1487–1498, 2020.
[3] 前川丈幸, 笠原秀一, 馬強. スポットの混雑を考慮した観光ルート推薦. 第 14 回データ工学と情報マネジメントに関するフォーラム，B41-3, 2022.
[4] 山内克之, 桐生拓海, Panote Siriaraya, 河合由起子, 中島伸介. 安全性や楽しさを考慮したウォーキング経路推薦方式の提案. 第 12 回データ工学と情報マネジメントに関するフォーラム，P2-10, 2020.
[5] Shivangi Soni, Venkatesh Gauri Shankar, and Chaurasia Sandeep. Route-The Safe: A Robust Model for Safest Route Prediction Using Crime and Accidental Data. Vol. 28, pp. 1415–1428, 12 2019.
[6] 木梨真知子, 金利昭. 光環境に着目した歩行者の夜間経路選択構造に関する研究. 都市計画論文集, Vol. 45.3, pp. 451–456, 2010.
[7] 湯浅智紀, 奥健太. 位置情報付き写真に基づくルート推薦システム. 第 11 回データ工学と情報マネジメントに関するフォーラム, P2-127, 2019.
[8] 山下桜生, 横山昌平. ジオソーシャルデータを用いた夜間の賑わいに基づいた経路推薦. 第 14 回データ工学と情報マネジメントに関するフォーラム，D41-3, 2022.
[9] 愛知県警察 安心・安全マップ. https://map.police.aichi.dsvc.jp/ (Last access: November 19, 2023).
[10] Geofabrik Download Server. Download OpenStreetMap data for this region : Chubu region. https://download.geofabrik.de/asia/japan/chubu.html (Last access: July 31, 2023).

公共交通機関の利用を想定した観光ルート推薦の一手法

田中　俊†　上野　史††　太田　学††

† 岡山大学工学部情報系学科　〒 700–8530 岡山県岡山市北区津島中 3-1-1
†† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山県岡山市北区津島中 3-1-1
E-mail: †pvbo4gku@s.okayama-u.ac.jp, ††{uwano, ohta}@okayama-u.ac.jp

あらまし　近年，e-Tourism と呼ばれる情報学を利用した観光に関する研究が加速している．中野らは，ユーザが訪問先として選択した観光スポットと出発地，到着地の計 3 地点の近傍のスポットを優先して訪問する移動効率の良い観光ルートを推薦する手法を提案した．しかし，彼女らは自動車以外の移動手段を想定しておらず，実際に観光ルート推薦システムとして利用する上で課題が残る．そのため，本稿では，徒歩と電車，バスの公共交通機関を利用する効率的な観光ルート推薦手法を提案する．実験では，複数の入力条件に基づいて岡山県内のスポットを巡る観光ルートを生成し，ルートの良さに加えて，移動時間や移動手段，移動にかかる料金などを評価した．中野らの提案したルート生成手法の移動手段を徒歩，公共交通機関に変更したモデルと提案手法を比較すると，推薦ルートの評価値，移動時間の点で改善が見られた．一方，中野らの提案した自動車を利用する観光ルート推薦と比較すると，推薦ルートの評価値の低下や移動時間の増大が見られた．

キーワード　観光，公共交通機関，ルート推薦

1 はじめに

近年観光は多くの関心を集める研究分野である．日本政府観光局の推計 [1]によると，2023 年の 1 月から 11 月の累計訪日外国人数は 2,233 万人となり，COVID-19 の感染拡大後 4 年ぶりに年間 2,000 万人を超え，今後さらに観光に対する関心の高まりが予想される．特にニューラルネットワークを用いた個人の好みや傾向を考慮した観光の研究や，ソーシャルメディアに投稿された観光に関する情報を利用した研究などが盛んである．また，COVID-19 の感染拡大をきっかけにサステナブル・ツーリズム (持続可能な観光) を推進する動きも活発化しており，各地域で暮らす生活者と観光客の双方にとってより良い観光を目指す機運などから新たな研究分野の開拓なども予想される．

観光においてソーシャルメディアの役割は大きい．観光客は観光スポットでの自身の体験を Twitter (現 X[2]) などで共有し，それを見た人に魅力を発信している．新井らは [1] は Twitter に投稿された観光スポットでの体験情報を含む投稿を収集し，そのスポットを訪れる時間帯や他の観光スポットとの関係性を考慮した観光ルート推薦手法を提案した．一方で，新井らの推薦手法は移動距離の長いルートを推薦することがあり，中野ら [2] はユーザが入力した出発地，到着地，そして訪問先としてユーザが選択した観光スポットの計 3 地点のいずれかから近い観光スポットを優先してルートを生成する手法を提案し，移動距離の短縮に成功した．

上記の観光ルート推薦手法は自動車での移動を想定しているが，観光客の利便性と地方の活性化を考慮すると公共交通機関を利用した観光を促進するほうが望ましい．そこで本稿では移動時間を基準に徒歩，電車，バスのうち適切な移動手段を用いる観光ルート推薦手法を提案する．また，実験では推薦した観光ルートをルートの良さや移動時間などの点から評価する．

本稿の構成は次の通りである．まず 2 節で関連研究について述べる．3 節で中野らが提案したルート生成手法を説明し，4 節で提案する公共交通機関の利用を想定した観光ルート推薦手法を説明する．5 節で実験において生成した推薦ルートの評価値や移動時間を比較し，考察する．最後に 6 節ではまとめと今後の課題を述べる．

2 関連研究

2.1 観光ルートの推薦研究

中嶋ら [3] は，位置情報を含む投稿を利用して観光ルートを推薦する手法を提案した．中嶋らは，観光客の Twitter での投稿で頻繁に用いられる特徴と Foursquare[3]や Instagram[4]のサービスを利用した Twitter の投稿をもとに観光体験情報を収集し，観光体験情報を投稿したユーザのタイムラインから観光ルートを取得した．また，収集した投稿を, 手がかり語や品詞の特徴から「食事」,「景観」,「行動」の 3 つのカテゴリに分類し, カテゴリを考慮した観光ルートを生成することで，観光客の好みに合う観光ルートを推薦した．しかし，中嶋らの手法はユーザのタイムラインからのみ観光スポットを取得するため，得られる観光スポットが少なくなるという問題があった．

新井ら [1] は，この問題を解決するために Google Places API[5] を用いて観光スポットの候補である施設の情報を取得した．そ

1：https://www.tourism.jp/tourism-database/stats/inbound/
2：https://twitter.com/?lang=ja
3：https://foursquare.com/
4：https://www.instagram.com/
5：https://developers.google.com/maps/documentation/places/web-service/overview?hl=ja

して，Yahoo！知恵袋[6]の質問を用いて取得した施設が観光スポットとして適しているかを判断した．また，新井らは取得した観光スポット名を手がかり語として，観光体験情報を含むTwitterの投稿を収集した．そして，収集した投稿から算出した観光スポットのスコアを用いて観光ルートを推薦した．しかし，新井らの観光ルート推薦は，スコアの高い観光スポットを優先して訪問することで移動効率の悪い観光ルートが推薦されることがあった．

中川ら [4] は，移動効率を改善するため，新井らの観光ルートの推薦に用いる観光スポットのスコアとルート生成アルゴリズムを変更した．具体的には，観光ルート内のスポット同士の関係性の評価である共起スコアを定義した．この共起スコアは，訪問順序が近いスポットにのみ設定した．また，生成した観光ルートからのスポットの削除と追加を繰り返す遺伝的アルゴリズムを利用して，局所最適解を回避するルート生成手法を提案した．

中野ら [2] は，さらなる観光ルートの移動効率の改善のため，推薦ルートで訪問する可能性のある候補スポットを取得して，取得した候補スポットからなる観光ルートの推薦を提案した．候補スポットは，ユーザが入力した出発地，到着地，そして訪問先として選択した観光スポットの計3地点のいずれかから近い観光スポットで，これにより移動効率を改善した．

前川ら [7] は，観光スポットの混雑が観光に与える影響を考慮した観光ルート推薦手法を提案した．具体的には動的滞在時間と呼ばれる観光スポットの混雑度によって変化する滞在時間や，環境税メタファーと呼ばれる混雑した観光スポットを避けるためのスコアを利用することにより，観光スポットの混雑による観光客の満足度の低下を抑える観光ルートを推薦した．

住友ら [9] は，ユーザが入力した観光における3種類の感情に対する重要度と，5種類の観光カテゴリに対する重要度からユーザの観光への興味の傾向を数値化した．また観光スポットに対して感情ベクトルと観光カテゴリベクトルを生成した．そしてユーザの観光への興味の傾向を表す数値と観光スポットの感情ベクトル，観光カテゴリベクトルを用いて各スポットのユーザにとっての魅力度を算出し，ユーザにとって魅力度の高い観光スポットを優先して訪問するルートを生成する手法を提案した．

Mouら [6] は，Long Short Term Memory networks (LSTM) を用いたPersonalized recurrent neural network (P-RecN) により，ユーザの過去の観光の軌跡データを入力にして観光データの意味ベクトルを生成した．そして生成した意味ベクトルをもとにユーザが気に入るであろう観光ルートを推薦した．ベースラインであるDeepMoveによる推薦結果と比較すると，推薦スポットの良さと訪問順序の妥当性を両立した観光ルートを推薦した．

本稿では観光での移動手段に着目し，移動手段として徒歩と公共交通機関を用いる観光ルート推薦手法を提案する．

6：https://chiebukuro.yahoo.co.jp/

2.2 観光スポット推薦研究

馬 [8] は，ブログやSocial Networking Services (SNS) などユーザが自ら情報を発信することで生成されるUser Generated Contents (UGC) から地域，観光に関するデータを収集した．そして，ユーザのプロフィールや発信履歴をもとにしたユーザモデリングの結果を利用して，観光客に知られていないが地元住民が頻繁に訪問するスポットの発見や個人に適した観光スポットの推薦や発見を可能にした．

中野ら [5] はユーザが撮影した写真にはユーザ固有の好みが含まれていると仮定し，マイクロソフトのクラウドプラットフォームであるMicrosoft Azure[7]のComputer Vision API[8]を利用して，撮影した写真からユーザの好みを反映したタグセットを生成した．そして，ユーザの好みを反映したタグセットと観光予定のルートを入力することで，Google Maps API[9]，4travel.jp[10]の口コミ，Flickr[11]の位置情報付きの写真のデータベースから，ユーザが立ち寄るのに適した観光スポットを推薦した．アンケート調査の結果，Flickrの位置情報付きの写真をデータベースに用いた場合の推薦の評価が低下した．

これらは個別の観光スポットを推薦する手法である．本稿では観光スポットの訪問順序とスポット間の移動手段を含めたルートを推薦するための手法を提案する．

3 中野らの推薦ルート生成手法

本節では観光体験情報を含むTwitterの投稿から算出した観光スポットの3つのスコアを用いた中野ら [2] の観光ルート推薦手法について説明する．本稿で提案する公共交通機関を利用する観光ルート推薦システムはこの手法を基に実装する．

3.1 観光に関する投稿の収集

中野らは新井ら [1] の手法を用いて観光スポット名を収集し，収集した観光スポットを用いて観光体験情報を含むTwitterの投稿を収集した．収集した観光スポット名を手掛かりと語としてTwitter API (現X API[10]) を用いて，以下の手順で各観光スポットについて観光体験情報を含む投稿を収集した．

（1） 観光スポット名と収集した観光スポットが存在する県名を含む投稿をしたユーザのIDを取得する．

（2） 取得したユーザIDから1ユーザ当たり最大3,200件の投稿を取得する．

（3） 取得した投稿から本文に観光スポット名と県名を含むものを抽出し，ツイートの本文や投稿日時などの情報を得る．なお，手順3において投稿が得られなかった観光スポットは，観光ルート推薦に利用しない．

7：https://azure.microsoft.com/
8：https://learn.microsoft.com/ja-jp/azure/architecture/data-guide/cognitive-services/vision-api
9：https://developers.google.com/maps?hl=ja
10：https://4travel.jp/
11：https://www.flickr.com/
10：https://developer.twitter.com/en/products/twitter-api

3.2 候補スポット集合の取得

中野らは推薦ルート生成のため，出発地，到着地，選択スポットの 3 地点のいずれかからの自動車による移動時間が短い X 個の観光スポットを観光ルートに含める可能性がある候補スポット集合として取得する．そして，ユーザが指定する出発地，到着地，選択スポットと候補スポット集合の観光スポットを含む観光ルートを生成する．この際候補スポットの個数 X は以下の式で求める．

$$X = \alpha \cdot \frac{T_G - T_S}{\frac{1}{m}\sum_{j=1}^{m} V_j} \tag{1}$$

ここで T_G は観光の終了時刻，T_S は観光の開始時刻，m は 3.1 節で収集した全ての観光スポットの数，V_j は Twitter の投稿時間から算出したスポット j の滞在時間を表す．つまり式 (1) の分母は全観光スポットの滞在時間の平均値であり，観光時間が長くなるほど候補スポットの個数 X は多くなる．また α は移動効率を調整するための係数であり，α が小さければ候補スポットが少なくなり，移動を制限した観光ルートが生成される．

3.3 ユーザの入力からルートの推薦までの流れ

中野らが提案したルート生成手法の，ユーザの条件の入力からルートの推薦までの流れを図 1 に沿って説明する．

（1） ユーザに出発地，到着地，ユーザが訪問先とする選択スポットを指定させ，出発地 → 選択スポット → 到着地を初期ルートとする．

（2） X 個の観光スポットからなる観光ルートに含める可能性がある候補スポット集合を取得する．

（3） 候補スポット集合のスポットを初期ルートに 1 つ追加したルートの集合を生成する．

（4） ルートの集合を出発地から観光スポットを経由して到着地までたどり着くまでの所要時間が短い順にソートする．

（5） 所要時間の最も短いルートに観光時間の制約を満たさなくなるまで候補スポット集合の観光スポットを追加する．

（6） 観光スポットを追加できない場合，推薦する観光ルートの候補として保存する．

（7） 例えば出発地 → スポット 2→ スポット 1→ 選択スポット → 到着地のルートに観光時間の制約を満たしつつ観光スポットを追加することができない場合，出発地 → スポット 3→ スポット 1→ 選択スポット → 到着地のルートに候補スポットから観光スポットを追加する．

（8） 制限時間内に生成された観光ルートのうち観光ルートの評価値が最も大きい観光ルートを推薦する．

3.4 観光スポットとルートの評価

3.4.1 観光スポットの評価

中野らは 2019 年に 3.1 節の手法で収集した観光体験情報を含む投稿から，岡山県内の 51 個の観光スポットについての時間帯スコア，共起スコア，人気スコアの 3 種類のスコアを算出した．

時間帯スコアは，観光スポットを訪れる時間帯の評価であり，24 時間を 3 時間ごとに区切った 8 つの時間帯ごとにスコアを設定する．例えば，ある観光スポットに対する投稿が全 10 件であった場合，そのうち 5 件の投稿が 12 時から 15 時の間に投稿されていた場合，この観光スポットの 12 時から 15 時の時間帯の時間帯スコアは $5/10 = 0.5$ となる．

共起スコアは，観光スポット同士の関係性の評価であり，あるユーザの同日の投稿に出てくる他の観光スポットとの共起確率を表す．例えば観光スポット A を訪れたユーザが 10 人いて，そのうち 3 人が同じ日に観光スポット B を訪れていた場合，観光スポット A における観光スポット B の共起スコアは $3/10 = 0.3$ となる．

人気スコアは，観光スポットの人気度であり，観光スポットの人気度を 1 から 5 の 5 段階で評価する．具体的には，投稿総数が下位 20 %の観光スポットの人気スコアを 1，上位 20 %の人気スコアを 5，その間を 20 %ごとに 2，3，4 としている．

3.4.2 観光ルートの評価値

中野らは，3.4.1 項で説明した観光スポットのスコアを用いて以下に示す評価値 v を生成した観光ルートごとに算出した．

$$v = \sum_{i=1}^{n} f_i \tag{2}$$

$$f_i = f_i^P (f_i^T + f_i^{CO}) \tag{3}$$

ここで n は観光ルートに含まれるスポットの総数，f_i は観光ルートにおける i 番目に訪れる観光スポットのスコアを表す．f_i^T，f_i^{CO}，f_i^P はそれぞれ i 番目に訪れる観光スポットの時間帯スコア，共起スコア，人気スコアを表す．

4 提案手法：公共交通機関を利用した観光ルート推薦手法

本研究では移動手段に徒歩と公共交通機関 (電車，バス) を利用する観光ルートを推薦する．システムの流れを岡山県を例に以下に示す．

（1） ユーザに岡山県の鉄道交通の要所と考えられる岡山駅，倉敷駅，東岡山駅，総社駅，津山駅，茶屋町駅の中から出発地と到着地を選択させる (図 2)．

（2） ユーザにスライドバーを用いて観光の開始時刻と終了時刻を指定させる．

（3） ユーザに図 3 のように表示される観光スポットの中から必ず訪れるスポットを選択させる．

（4） 制限時間内に生成された観光ルートのうち，観光ルートの評価値が最も大きい観光ルートを推薦する．

中野ら [2] は自動車による移動時間を用いてルートを生成し，スポットを追加する位置に関係なく 1 つの候補スポット集合からスポットを追加した．しかし，本研究では公共交通機関を用いるため，駅やバス停までの移動時間や，駅，バス停での待ち時間を考慮した候補スポット集合を取得する必要がある．そこで 3.3 節で説明した中野らのルート生成の流れの手順 (2) において候補スポット集合の取得し，手順 (3) において候補スポット集合からルートへスポットを追加する方法を 2 つ提案する．

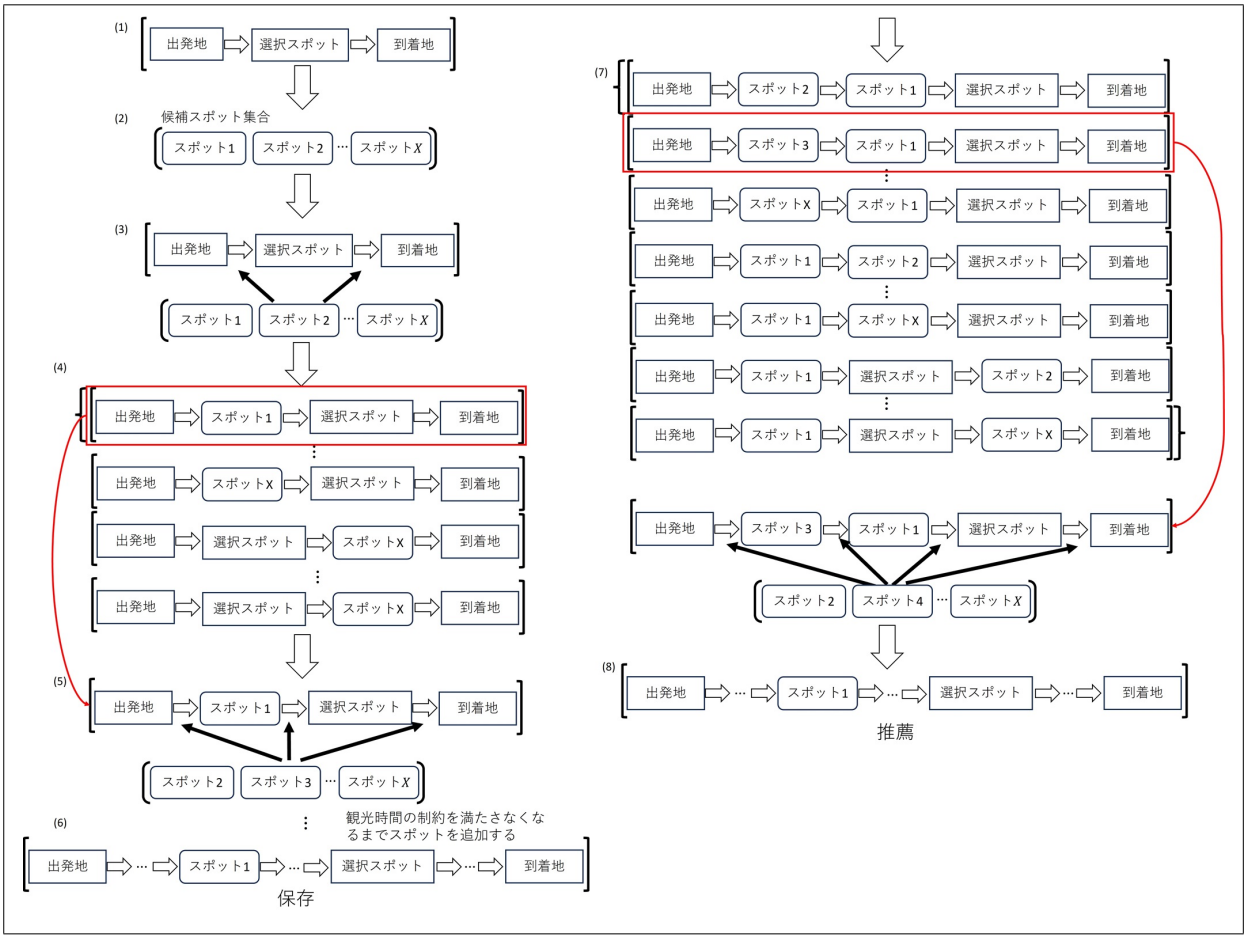


図 1　中野らのルート生成アルゴリズム

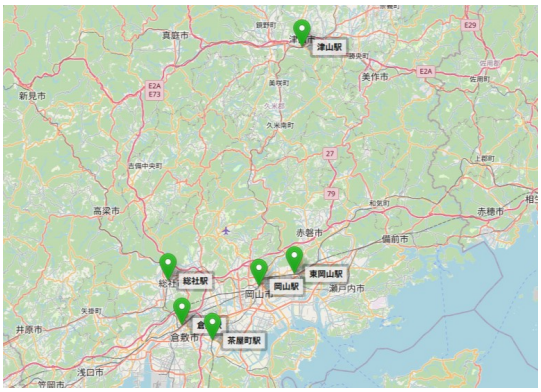


図 2　出発地，到着地，観光時間決定画面 (岡山県)

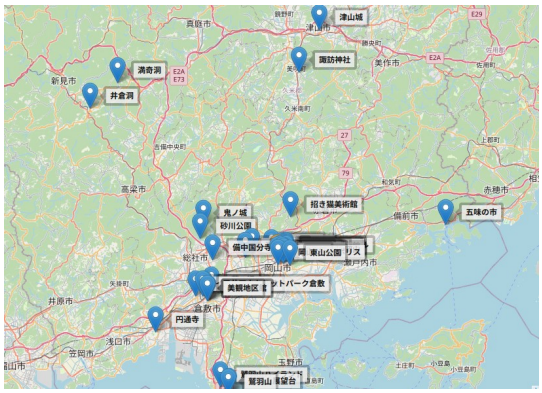


図 3　選択スポット決定画面 (岡山県)

4.1　2 スポット間の移動手段と移動時間

本研究では推薦する観光ルートの移動手段に徒歩，電車，バスを想定する．本節では移動手段の決定方法について説明する．まず HeartRails Express[11]を用いて本研究で使用する岡山県内の 51 件の観光スポットの最寄り駅を取得する．また，Google Directions API[12]を用いて観光スポット間を徒歩で移動した際の移動距離も取得する．電車の時刻表データは Yahoo!Japan 路線情報 [13]から，バスの時刻表データを NAVITIME[14]から取得する．移動手段を以下の条件に従って決定する．

（1） 徒歩による移動速度を分速 80 mとし，2 スポット間の徒歩による移動時間が 20 分以内の時，移動手段を徒歩とする．

（2） 2 スポット間の徒歩による移動時間が 20 分以上で，2 スポットの最寄り駅が同じ場合移動手段をバスとする．

（3） (1)，(2) の条件をどちらも満たさない場合，電車を移動手段とする．

11：https://express.heartrails.com/api.html
12：Google https://developers.google.com/maps/documentation/directions/overview?hl=ja
13：https://transit.yahoo.co.jp/
14：https://www.navitime.co.jp/

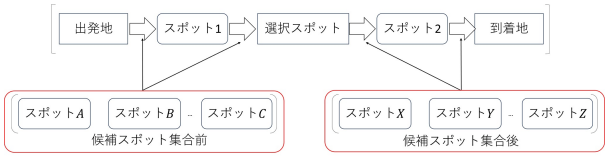


図 4　選択スポットの前後を意識したルート生成手法

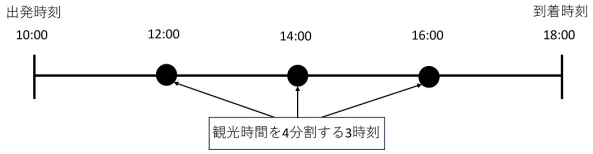


図 5　移動時間の想定

（4） 電車を用いる場合，現在地とその最寄り駅，目的地とその最寄り駅の移動手段は，徒歩による移動時間が 20 分以内の時は徒歩，20 分以上の時バスとする．

4.2　選択スポットの前後で異なる候補スポット集合を利用するルート生成手法

中野らは推薦ルート生成において，スポットを追加する位置に関係なく 1 つの候補スポット集合から観光スポットを追加する手法を提案した．本研究では，ユーザの選択スポットの前後で候補スポット集合を分ける手法を提案する．具体的には，出発地と選択スポットの間にスポットを追加する場合と，選択スポットと到着地の間に追加する場合で異なる候補スポット集合からスポットを追加する (図 4)．まず，出発地と選択スポットのいずれかからの移動時間が短いスポットの集合を候補スポット集合前として取得する．そして到着地と選択スポットのどちらかからの移動時間が短いスポットの集合を候補スポット集合後として取得する．そして，ルートの出発地と選択スポットの間にスポットを追加する場合には，候補スポット集合前からスポットを追加し，ルートの到着地と選択スポットの間にスポットを追加する場合には，候補スポット集合後からスポットを追加する．中野らの手法は候補スポット集合が 1 つであるため，到着地に近いスポットを出発地と選択スポットの間に訪問するルートが生成される可能性がある．しかし，この提案手法はそのようなルートの生成を抑制し，またルートの計算コストも下げる．ただし，公共交通機関による移動時間は観光スポットの出発時刻によって変動するため，候補スポットの取得には推定移動時間を使用する．

この移動時間の推定のため，出発時刻から到着時刻までの観光時間を 4 分割した 3 つの時刻を選択スポットの出発時刻と仮定する．3 つの時刻に選択スポットから他の観光スポットに移動するのに要する移動時間それぞれ計算し，3 つの移動時間のうち最も時間の要するものを，候補スポット集合の取得に用いる推定移動時間とする．例えば図 5 のように出発時刻が 10 時，到着時刻が 18 時の観光ルートを推薦する場合，観光時間を 4 分割した 12 時，14 時，16 時を選択スポットの出発時刻として計算した移動時間のうち最も長い時間を推定移動時間とする．

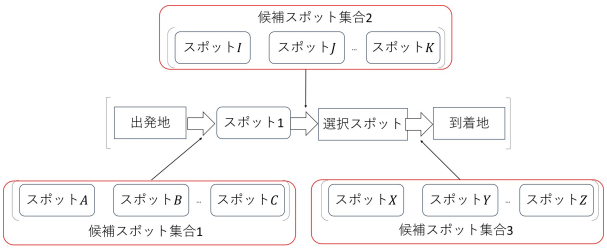


図 6　前後のスポットとの移動時間を考慮したルート生成手法

4.3　前後のスポットとの時間的近傍のスポットを優先したルート生成手法

本稿では，スポットを追加する位置ごとに，前後のスポットと時間的近傍にある候補スポット集合を取得し，ルートにスポットを追加する手法も提案する．この提案手法ではまず，スポットを追加する位置ごとに，前後 2 つの観光スポットのいずれかからの移動時間が短いスポットを候補スポット集合として取得する．そして，スポットを追加する位置ごとに時間的近傍の候補スポット集合からスポットを追加する．図 6 においては出発地とスポット 1 のいずれかからの移動時間が短い候補スポット集合 1，スポット 1 と選択スポットのいずれかからの移動時間が短い候補スポット集合 2，選択スポットと到着地のいずれかからの移動時間が短い候補スポット集合 3 を取得している．ルート生成では，追加する位置ごとに候補スポット集合 1，2，3 のうち対応する候補スポット集合から観光スポットを追加する．4.2 節で説明した候補スポット集合よりも細分化した候補スポット集合を取得するため，より移動効率の良いルートの生成が期待される．しかし，ルートにスポットを追加するたびに候補スポット集合を計算するため計算コストはかかる．なお，候補スポット集合の取得で利用する推定移動時間は，4.2 節と同じ方法で計算する．

5　観光ルートの推薦実験

5.1　実 験 内 容

実験では，中野ら [2] の観光ルート推薦手法 (以下,「中野らの手法 (自動車)」)，中野らの手法の移動手段のみを徒歩，公共交通機関に変更した手法 (以下,「中野らの手法 (公共交通機関)」)，選択スポットの前後で異なる候補スポット集合を利用するルート生成手法 (以下,「提案手法 (2 分割)」)，前後のスポットとの時間的近傍のスポットを優先したルート生成手法 (以下, 提案手法 (細分化)) の推薦ルートを比較する．ルート推薦の実験に用いる 3 つの入力を表 1 に示す．この 3 つの入力により推薦される観光ルートの訪問するスポット数，観光ルートの評価値 v，公共交通機関の乗車時間，徒歩による移動時間，公共交通機関の待ち時間，総移動時間，公共交通機関の利用に要した料金を比較する．

本実験では表 2 に示す新井ら [1] の方法を利用して中野らが 2019 年に収集した岡山県内の 51 件の観光スポットを推薦に使用する．

表 1　実験に用いる入力

	出発地	到着地	選択スポット	観光時間
(i)	岡山駅	岡山駅	岡山城	10:00-18:00
(ii)	岡山駅	倉敷駅	三井アウトレットパーク倉敷	10:00-18:00
(iii)	岡山駅	岡山駅	鷲羽山ハイランド	8:00-18:00

表 2　岡山県の推薦の候補スポット一覧

ひるぜん	鷲羽山展望台
石山公園	おもちゃ王国
岡山後楽園	鬼ノ城
湯郷温泉	池田動物観
岡山プラザホテル	アリオ倉敷
備中松山城	招き猫美術館
宝来軒	後楽ホテル
湯原温泉	祇園神社
讃岐の男うどん	イオンモール倉敷
醍醐桜	満奇洞
夢二郷土美術館	鷲羽山ハイランド
イオンモール岡山	備中国分寺
蔵の湯	文の助茶屋
大原美術館	神庭の滝
渋川海岸	吉備津彦神社
クレド岡山	東山公園
岡山県立美術館	岡山ジョイポリス
砂川公園	岡山県立美術館
円通寺	井倉洞
倉敷珈琲館	吉備津神社
宇野港	渋川動物公園
五味の市	鷲羽山
藤公園	美観地区
三井アウトレットパーク倉敷	岡山城
お城茶屋	諏訪神社
津山城	

5.2 実験結果

計算時間に制限を設けなかった場合の入力 (i)，(ii)，(iii) に対する推薦ルートをそれぞれ表 3，4，5 にまとめる．また，計算時間を 60 秒にして実行した場合の推薦ルートを表 6，7，8 にまとめる．制限時間なしの場合，提案手法 (2 分割) と提案手法 (細分化) は入力 (i) に対して同じルートを推薦した．また，制限時間なしの場合と 60 秒の制限時間を設けた場合の推薦ルートを比較すると，入力 (i) に対しては中野らの手法 (自動車) 以外の 3 つの手法が異なるルートを推薦したが，入力 (ii)，入力 (iii) に対してはすべての手法が同じルートを推薦した．入力 (i) に対する提案手法 (細分化) の制限時間なしの場合と 60 秒の制限時間を設けた場合の推薦ルートは他 3 つの手法と比較すると，評価値 v が大幅に低下している．これは提案手法 (細分化) は計算コストがかかるため，制限時間の影響が大きかったといえる．以下では 60 秒の制限時間を設けた場合の推薦ルートを比較する．

中野らの手法 (自動車) と徒歩と公共交通機関を利用した他 3 つの手法の推薦ルートを比較する．表 6 のように出発地と到着地，選択スポットの 3 地点が距離的に近い場合，公共交通機関を利用するルートは，総移動時間は増加するものの，訪問スポット数はほぼ同じで，観光ルートの評価値 v もあまり変わらなかった．表 7 や 8 のように 3 地点のいずれかが距離的に離れている場合，訪問スポット数が減少し，ルートの評価値 v も低下した．総移動時間の増加も顕著だった．

表 3　制限時間なしの場合の入力 (i) に対する推薦ルートの評価値等

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
計算時間 (秒)	1,341	10,733	663	579
訪問スポット数	5	5	4	4
ルートの評価値 v	8.66	7.69	10.08	10.08
乗車移動時間 (分)	16	0	9	9
徒歩移動時間 (分)	-	65	51	51
待ち時間 (分)	-	0	9	9
総移動時間 (分)	16	65	69	69
運賃 (円)	-	0	120	120

表 4　制限時間なしの場合の入力 (ii) に対する推薦ルートの評価値等

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
計算時間 (秒)	7	395	119	536
訪問スポット数	5	4	4	4
ルートの評価値 v	6.84	4.96	4.91	5.01
乗車移動時間 (分)	51	26	17	17
徒歩移動時間 (分)	-	62	47	48
待ち時間 (分)	-	30	14	7
総移動時間 (分)	52	118	78	72
運賃 (円)	-	450	330	330

表 5　制限時間なしの場合の入力 (iii) に対する推薦ルートの評価値等

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
計算時間 (秒)	229	285	135	299
訪問スポット数	6	4	4	4
ルートの評価値 v	10.73	4.99	3.40	4.99
乗車移動時間 (分)	106	91	91	91
徒歩移動時間 (分)	-	45	40	45
待ち時間 (分)	-	78	79	78
総移動時間 (分)	106	214	210	214
運賃 (円)	-	1580	1580	1580

中野らの手法 (自動車) を除く，公共交通機関を利用する他 3 つの手法の推薦ルートを比較する．表 6 の入力 (i) に対する推薦ルートの中で最も評価値 v の高い提案手法 (2 分割) の推薦ルートを図 7 に示す．選択スポット周辺の人気スコアの高いスポットを訪問するルートを推薦した．表 7 の入力 (ii) に対する推薦ルートの中で最も評価値 v の高いルートを推薦した提案手法 (細分化) の推薦ルートを図 8 に示す．出発地周辺のスポットを訪問せず，選択スポットと到着地周辺のスポットのみで構成されるルートを推薦した．表 8 の入力 (iii) に対して同じルートを推薦し，最も評価値 v が高かった中野らの手法 (公共交通機関) と提案手法 (細分化) の推薦ルートを図 9 に示す．出発地周辺のスポットを訪問したのち，選択スポットを訪問するルートを推薦した．図 7，8，9 のタイムテーブルのスポットの数字は出発地と到着地，観光スポットの訪問順序を示す．数字のないスポットは観光ルート内で経由する駅やバス停を示す．また，移動手段は次のスポットへ向かう際の移動手段を示す．

表 6 の入力 (i) に対する推薦ルートを比較する．中野らの手法 (公共交通機関) は評価値 v が低く，総移動時間も長い観光ルートを推薦した．提案手法 (2 分割) は 3 つの手法のうち最も

表 6　制限時間 60 秒の場合の入力 (i) に対する推薦結果

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
訪問スポット数	5	5	5	4
ルートの評価値 v	8.66	7.50	9.98	7.93
乗車移動時間 (分)	16	9	9	6
徒歩移動時間 (分)	-	63	53	47
待ち時間 (分)	-	8	11	12
総移動時間 (分)	16	80	73	65
運賃 (円)	-	120	120	120

表 7　制限時間 60 秒の場合の入力 (ii) に対する推薦結果

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
訪問スポット数	5	4	4	4
ルートの評価値 v	6.84	4.96	4.91	5.01
乗車移動時間 (分)	52	26	17	17
徒歩移動時間 (分)	-	62	47	48
待ち時間 (分)	-	30	14	7
総移動時間 (分)	52	118	78	72
運賃 (円)	-	450	330	330

表 8　制限時間 60 秒の場合の入力 (iii) に対する推薦結果

	中野らの手法 (自動車)	中野らの手法 (公共交通機関)	提案手法 (2 分割)	提案手法 (細分化)
訪問スポット数	6	4	4	4
ルートの評価値 v	10.73	4.99	3.40	4.99
乗車移動時間 (分)	106	91	91	91
徒歩移動時間 (分)	-	45	40	45
待ち時間 (分)	-	78	79	78
総移動時間 (分)	106	214	210	214
運賃 (円)	-	1580	1580	1580

評価値 v が高くなり，総移動時間も中野らの手法 (公共交通機関) より短くなった．また，提案手法 (細分化) は他 2 つの手法と比較すると訪問したスポット数が 1 つ少ないものの，中野らの手法 (公共交通機関) よりも評価値 v が高くなり，移動時間が短くなった．

表 7 の入力 (ii) に対する推薦ルートを比較する．中野らの手法 (公共交通機関) は提案手法 (2 分割) よりも評価値 v が高いものの，電車の待ち時間が長く，運賃も高い観光ルートを推薦した．提案手法 (2 分割) は 3 つの手法の中で最も評価値 v が低くなり，提案手法 (細分化) よりも総移動時間が長くなった．提案手法 (細分化) は 3 つの手法の中で最も評価値 v が高くなり，総移動時間も短くなった．

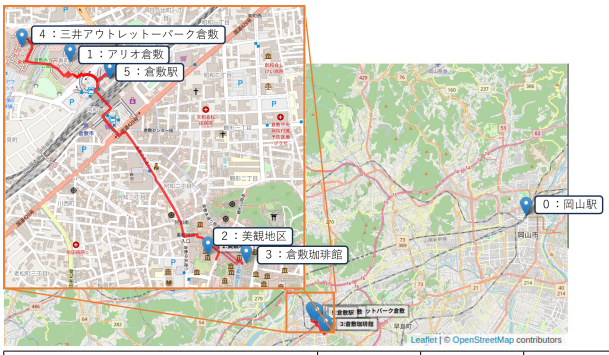
表 8 の入力 (iii) に対する推薦ルートを比較する．中野らの手法 (公共交通機関) と提案手法 (細分化) は同じルートを推薦し，提案手法 (2 分割) よりも評価値 v が高くなった．提案手法 (2 分割) は他 2 つの手法の推薦ルートと比較すると総移動時間がわずかに短くなったものの，評価値 v が大きく低下した．

5.3　考　　察

徒歩，公共交通機関を利用する 3 つの手法の推薦ルートと中野らの手法 (自動車) を比較すると，表 7，8 の入力 (ii)，(iii) については評価値 v が大きく低下した．これは移動手段に自動車を用いることで，待ち時間などが発生せず短時間での移動が可能なため，より多くの観光スポットを訪問することができるためである．しかし，評価値 v は観光スポットのスコアから式 (2) で算出されるため，自動車による観光ルートと公共交通機関による観光ルートを比較するには，さらに自動車の燃料費や環境への影響などをスコアに含める必要がある．

スポット	移動手段	到着時刻	出発時刻
0: 岡山駅	徒歩		10:00
1: 岡山県立美術館	徒歩	10:20	11:20
2: 石山公園	徒歩	11:25	12:25
3: 岡山県立博物館	徒歩	12:35	13:40
4: お城茶屋	徒歩	13:50	15:00
5: 岡山城	徒歩	15:05	16:40
城下駅	路面電車	16:48	16:59
岡山駅前駅	徒歩	17:08	17:08
6: 岡山駅		17:11	

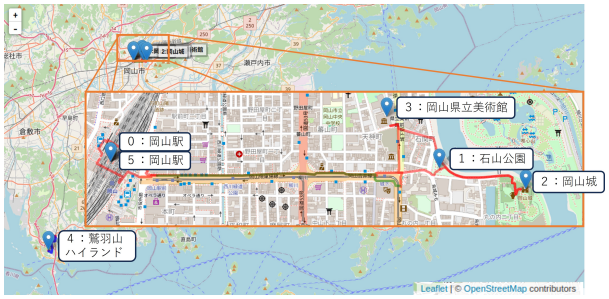
図 7　提案手法 (2 分割) の入力 (i) に対する推薦ルート



スポット	移動手段	到着時刻	出発時刻
0: 岡山駅	電車		10:07
倉敷駅	徒歩	10:24	10:24
1: アリオ倉敷	徒歩	10:30	11:50
2: 美観地区	徒歩	12:05	14:10
3: 倉敷珈琲館	徒歩	14:15	15:15
4: 三井アウトレットパーク倉敷	徒歩	15:40	17:00
5: 倉敷駅		17:10	

図 8　提案手法 (細分化) の入力 (ii) に対する推薦ルート

提案手法 (2 分割) は入力 (i) に対して 4 つの手法の中で最も評価値 v の高い観光ルートを推薦した．移動効率の悪い観光ルートが生成される可能性を減少させ，制限時間内で効率的にルートを生成したことが要因として挙げられる．一方で入力 (iii) に対してはスコアの高い観光ルートを推薦できなかった．これは入力 (i) は選択スポットである「岡山城」周辺に多くの観光スポットが存在するが，入力 (iii) の選択スポットである「鷲羽山ハイランド」周辺は観光スポットが少なく，「岡山城」や「岡山後楽園」などのスコアの高いスポットを候補スポットとして取得できなかったことが原因である．スコアの高いスポットを候補スポットとして取得できない場合，候補スポットの個数を決定する式 (1) の移動効率を調整する係数 α を自動で調整する処理などの追加が必要である．



スポット	移動手段	到着時刻	出発時刻
0: 岡山駅	徒歩		8:00
1: 石山公園	徒歩	8:20	9:20
2: 岡山城	徒歩	9:25	11:10
3: 岡山県立美術館	徒歩	11:10	12:10
城下駅	路面電車	12:13	12:23
岡山駅前駅	徒歩	12:32	12:32
岡山駅	電車	12:35	12:42
児島駅	徒歩	13:04	13:04
児島駅前	バス	13:07	13:30
鷲羽山ハイランド前	徒歩	13:40	13:40
4: 鷲羽山ハイランド	徒歩	13:40	16:10
鷲羽山ハイランド前	バス	16:10	16:40
児島駅前	徒歩	16:49	16:49
児島駅	電車	16:52	17:19
5: 岡山駅		17:41	

図 9　中野らの手法 (公共交通機関)，提案手法 (細分化) の入力 (iii) に対する推薦ルート

提案手法 (細分化) は入力 (i)，入力 (ii) に対して中野らの手法 (公共交通機関) や提案手法 (2 分割) と比較すると総移動時間の短い観光ルートを推薦した．これは候補スポット集合を細分化して取得することにより，移動時間の短いスポットをより優先して訪問する観光ルートが生成されたためである．

移動手段は，スポット間の徒歩の移動時間が 20 分以内の時に徒歩，20 分以上の時に電車となる．そのため，現在地と訪問先のスポットのいずれかが最寄り駅から離れている場合や，待ち時間が長い場合，運賃がかかるにも関わらず，徒歩とほとんど移動時間が変わらない場合があった．徒歩の移動時間のみを考慮して移動手段を決定するのではなく，徒歩や電車，バスによる移動時間を複合的に比較して，妥当な移動手段を採用する必要がある．

6　ま　と　め

本稿では，推薦観光ルートに含める可能性のある候補スポット集合を利用して，ルートを生成する中野らの手法を基に，移動手段として徒歩，公共交通機関の利用を想定した観光ルート推薦手法を提案した．提案手法は，徒歩，公共交通機関による移動時間を考慮して候補スポットを取得することにより，ルートの評価値と移動効率のバランスの取れた観光ルートが生成できる．しかし，中野ら [2] が提案した自動車によるを想定した推薦観光ルートと比較すると，移動に多くの時間を要し，訪問できる観光スポットが減少するなどの原因でルートの評価値が低下した．

徒歩，公共交通機関による観光ルートの比較では，選択スポットの前後で異なる候補スポット集合を利用するルート生成手法は，出発地，到着地，選択スポットの周辺に多くの観光スポットが存在する場合，制限時間内で評価値の高い観光ルートを推薦した．一方で周辺に観光スポットが存在しない場合，評価値が大きく低下した．前後のスポットとの時間的近傍のスポットを優先したルート生成手法は，スポットを追加する位置の前後のスポットとの徒歩，公共交通機関による移動時間を考慮するため，総移動時間が短く，評価値も高いルートを推薦した．

また，実験では選択スポットが岡山県の中心地から離れている場合，電車やバスの出発時刻の間隔が 1 時間以上のことも多かった．そのためそのような地域を訪問するルートを生成する場合，待ち時間が長くなる傾向が見られた．観光において待ち時間の長さは観光客の満足度の低下にもつながるため，今後の課題として公共交通機関の乗り換えを考慮したルート生成手法の検討が挙げられる．

文　献

[1] 新井晃平, 新妻弘崇, 太田学, “Twitter を利用した観光ルート推薦の一手法,” 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM 2015), G7-6, 2015.
[2] 中野翔子, 新妻弘崇, 太田学, “ユーザが入力した地点の近傍を優先する観光ルート推薦の一手法,” 第 12 回データ工学と情報マネジメントに関するフォーラム (DEIM 2020), J3-4, 2020.
[3] 中嶋勇人, 新妻弘崇, 太田学, “位置情報付きツイートを利用した観光ルート推薦,” 情報処理学会研究報告データベースシステム (DBS), Vol. 2013-DBS-158, No. 28, pp. 1-6, 2013.
[4] 中川智也, 新妻弘崇, 太田学, “マイクロブログを利用した観光ルート推薦における移動効率の改善,” 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), H1-3, 2016.
[5] 中野広貴, 荒澤孔明, 渡邊稜平, 服部峻, “観光写真から抽出した撮影者の好みに基づく観光スポット推薦,” 電子情報通信学会技術研究報告, Vol. 118, No. 408, pp. 45-50, 2019.
[6] Naixia Mou, Qi Jiang, Lingxian Zhang, Jiqiang Niu, Yunhao Zheng, Yanci Wang, Tengfei Yang, “Personalized tourist route recommendation model with atrajectory understanding via neural networks,” International Journal of Digital Earth, Vol. 15, No. 1, pp. 1738–1759, 2022.
[7] 前川丈幸, 笠原秀一, 馬強, “混雑を考慮した観光ルート生成に関する研究,” 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM 2023), 5b-3-4, 2023.
[8] 馬強, “観光情報学の最前線 観光の分散化と個人化を促進する集合知活用情報技術,” 情報処理, Vol. 58, pp. 220-226, 2017.
[9] 住友千将, 岳五一, “観光者の嗜好性を考慮した観光経路構成アルゴリズムの実装と実証実験,” パーソナルコンピュータ利用技術学会論文誌, Vol. 15, No. 1, pp. 13-20, 2021.

人流データに基づく観光スポットの周辺混雑度の予測と旅行計画への影響分析

星　結人† 小竹　神†† 北山　大輔†

† 工学院大学情報学部システム数理学科　〒 163-8677 東京都新宿区西新宿 1 丁目 24-2
†† 工学院大学工学研究科情報学専攻　〒 163-8677 東京都新宿区西新宿 1 丁目 24-2
E-mail: †{jx20246,em22009}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　観光スポット周辺では混雑が生じ，観光客に影響を与えることが予想される．そのため，観光スポット推薦の研究では，混雑を回避するような観光スポット推薦手法が提案されている．しかし，ユーザの計画時において，周辺の混雑状況はどのように影響を与えているのかは明らかにされていない．そこで本研究では，被験者に 1 日の旅行計画を立ててもらう実験を行う．この時，被験者には観光スポット情報に加え，周辺混雑度やレビュー数などの補足情報の表示を制御して表示する．補足情報の種類や表示の有無を変更した実験により，被験者が旅行計画で選択した観光スポットの周辺混雑度とレビュー数を基に，観光スポットの選択において，観光客には観光スポットの周辺の混雑状況が影響を与えているのかを検証した．その結果，観光客はより混雑している観光スポットを選択することがわかった．

キーワード　混雑度，観光スポット検索，行動分析

1 はじめに

観光において，近年ではオーバーツーリズムやコロナウイルス感染症の流行により，観光スポットの混雑が問題となっている．混雑の問題に関する研究として，混雑を回避する観光スポット推薦などが提案されている．前川ら [1] は観光スポットの選択において，混雑を回避して観光する人が多いと仮定し，観光スポットの混雑状況を考慮したルート生成の手法を提案した．Hidaka ら [2] は観光において，観光客に満足のいく体験を提供するには，観光スポットの混雑を回避することも重要であると仮定し，混雑状況を含めた上で観光客の嗜好に合った観光スポットを推薦するシステムを提案した．しかし，ユーザの観光スポットの選択において，観光スポットの周辺の混雑状況は影響を与えているのかが明らかにされていない．

そこで本研究では，観光スポットの選択において，観光客には観光スポットの周辺の混雑状況が影響を与えているのかを検証する．本稿では，観光スポットに対する周辺混雑度，レビュー数，相対混雑度，相対レビュー数を定義する．観光地検索システムを構築し，被験者に 1 日の旅行計画を立ててもらう実験を行う．この時，検索結果として観光スポット，観光スポットの説明文，観光スポットの公式サイトのリンクの他に，追加情報なし，周辺混雑度のみ追加，レビュー数のみ追加，両方追加のどれかのパターンで表示される．周辺混雑度，レビュー数は相対値に変更したパターンも行い，計 7 パターンの出力で実験を行う．この実験により，ユーザが旅行計画で選択した観光スポットの情報を基に，観光スポットの周辺の混雑状況の影響度合いを評価する．

本稿の構成を以下に述べる．2 章では関連研究について述べる．3 章では提案手法について述べる．4 章では実験について述べる．5 章では結果と考察について述べる．6 章ではまとめと今後の展望について述べる．

2 関連研究

混雑度やレビュー数を始めとするさまざまな観点に基づいて観光スポットや観光ルートを推薦する研究が行われている．まずは，それぞれを 2.1 節，2.2 節にて紹介する．しかしながら，それらの観点がユーザの選択に及ぼす影響を明らかにした研究は少ない．本研究では，これらの研究の暗黙的な仮定となっている，周辺混雑度やレビュー数の影響を明らかにするものである．このような旅行に関する行動に影響を与える要素を分析した研究を 2.3 節にて紹介する．本研究もこれらの分析に類するものであるが，旅行計画時における選択の影響の分析を目的としている点で異なる．

2.1 混雑状況を考慮した観光スポット推薦

混雑状況を考慮した観光スポット推薦に関する研究は，[1], [2] の他にも数多く行われている．野本ら [3] は観光レビュー文を用いて，観光スポットを「食事」,「景観」,「購買」,「体験」,「設備」,「混雑」,「交通」の 7 つの観点で評価し，レビュー数を基に穴場スポットを発見した．ユーザ実験を行った結果，発見した穴場スポットの 66.7 %が穴場スポットとして妥当であることを明らかにした．野本ら [4] は観光レビュー文を用いて，観光スポットを「飲食」,「景観」,「学び」,「購買」,「レジャー」,「エンタメ」,「リラックス」,「設備」,「混雑」,「交通」,「応対」,「ファミリー」の 12 の観点で評価した．さらに，ユーザの各観点への興味を表す興味ベクトルを定義し，これらを用いて算出したスコアが大きい観光スポットを推薦した．ユーザ実験を行った結果，提案した観点マッチングによって，認知度が低くても魅

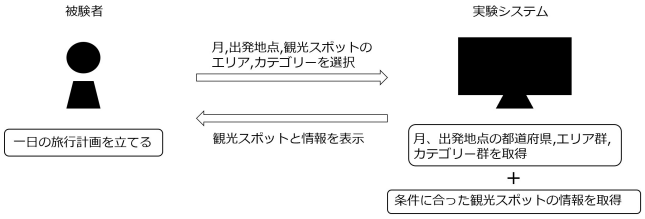


図 1　システム概要

力があり，ユーザが興味を持てるスポットを推薦できることを明らかにした．

2.2　混雑状況を考慮した観光ルート生成

混雑状況を考慮した観光ルート推薦に関する研究を紹介する．前川ら [5] は動的な要素である混雑を考慮した観光ルート生成を行うために，強化学習と Pointer Network を用いたモデルに,「動的滞在時間」,「環境税メタファー」を導入したモデルを提案した．ユーザ実験を行った結果，提案手法のモデルが観光体験価値を維持し，混雑を回避するルート生成できることを明らかにした．Isoda ら [6] は観光スポットの特徴を用いてスコアを設定し，動的計画法を用いて，もっとも高いスコアを持つルートを探索する手法を提案した．観光スポットの特徴として，スポットそのものの価値である静的な特徴と，時間や季節，混雑状況によって変動する動的な特徴を用いた．ユーザ実験を行った結果，提案手法のアルゴリズムは，もっとも高いスコアを持つルートを推薦できることを明らかにした．栗山ら [7] は大規模道路網における混雑状況を考慮し，多数のユーザがそれぞれ，気着時間制約を満たしつつ，需要度の高い目的地を巡回できるようなスケジューリング手法を提案した．ユーザ実験の結果，20000 人のユーザについて 8 分程度でスケジュールの算出が可能であることを明らかにした．

2.3　旅行客の行動分析

行動分析に関する研究は，数多く行われている．武ら [8] は東京都新式根島の海水浴場を事例に，混雑度の変化を測定し，混雑度が利用者の混雑感および満足感に与えた影響を分析した．アンケート調査を行った結果，混雑度が利用者の混雑感および満足感に与えた影響には相関関係がないことを明らかにした．大村ら [9] はスポット属性を考慮した実空間のユーザ行動分析に基づく潜在的興味分析について提案し，提案手法の有効性を検証した．藤井ら [10] は旅行ブログエントリを利用し，訪日外国人旅行者や各都市の特徴を明らかにした．具体的には，分析の対象となる属性「性別」や「使用言語」,「観光タイプ」の自動判定を行い，判定された属性に基づいて分析した．

3　実験システム

3.1　システムの概要

実験システムの概要を図 1 に示す．被験者は，実験システムを用いて一日の旅行スケジュールを立てる．まず，訪れる月を選択する．次に，被験者は出発地点を札幌駅，東京駅，横浜駅，名古屋駅，京都駅，大阪駅，博多駅の計 7 地点から 1 つを選択する．選択された出発地点から，実験システムは出発地点の都道府県と，その都道府県内をわけるエリア群を取得する．さらに，被験者は訪れたい観光スポットのエリアとカテゴリを 1 つずつ選択する．実験システムは，選択された月と観光スポットのエリアとカテゴリから，条件に合った観光スポットを表示する．被験者はこの操作を繰り返しながら，旅行計画に加えたい観光スポットを選択していく．

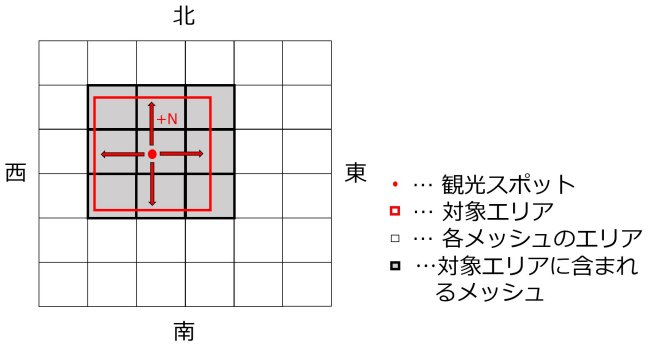


図 2　観光スポットの周辺混雑度の算出

3.2　対象となる観光スポットのデータ

じゃらん net[1]から収集したデータを用いてさまざまなデータを取得する．じゃらん net のデータとして，47 都道府県の観光スポット，観光スポットの属するカテゴリ群，各観光スポットの属する都道府県，エリア群，市町村，各観光スポットの緯度・経度，レビュー数などがある．これを利用し，被験者が選択した出発地点の都道府県と，その都道府県のエリア群，観光スポットのカテゴリ群を取得する．

3.3　周辺混雑度，相対混雑度の計算

G 空間情報センターが提供している 2019 年における全国の 1km メッシュ別の滞在人数データ [2]を用いて周辺混雑度の計算をする．この滞在人数データには月ごとに平日，休日の滞在人数が記録されている．また，メッシュに対する終日（7 時から 8 時台，18 時から 20 時台），深夜帯（1 時台から 4 時台），日中（11 時台から 14 時台の平均）ごとの滞在人数が記録されている．そのため，終日，深夜帯，日中の滞在人数データを元に近似値を求め，時間帯ごとの周辺混雑度を算出した．時間帯ごとの周辺混雑度で使用した滞在人数データの近似値を表 1 に示す．

観光スポットの周辺混雑度の算出方法を図 2 に示す．ある観光スポット s を入力地点とし，これを図 2 では赤丸で表す．入力地点から N 度で拡張した図 2 の赤枠を対象エリア $t(s)$ とする．対象エリアの最大緯度・経度，最小緯度・経度を取得し，対象エリアに含まれるメッシュ集合 M を取得する．ここで，対象エリアとメッシュ集合 M に属する各メッシュ m の重複領域を計算することで，対象エリア内の人口の近似値を求める．メッシュの人口に重複領域の割合をかけることで，そのメッシュの

1：https://www.jalan.net/
2：https://front.geospatial.jp/

表 1　時間帯ごとの周辺混雑度で使用した滞在人数データ

時間帯	用いる滞在人数データ
0 時から 4 時台	深夜帯の人口
5 時から 6 時台	深夜帯と終日の人口の平均
7 時から 8 時台	終日の人口
9 時から 10 時台	日中と終日の人口の平均
11 時から 14 時台	日中の人口
15 時から 17 時台	日中と終日の人口の平均
18 時から 20 時台	終日の人口
21 時から 23 時台	深夜帯と終日の人口の平均

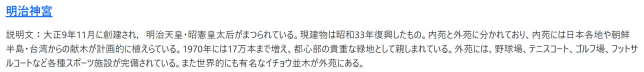


図 3　基本情報のみ

近似値とする．対象エリア内のメッシュの近似値を合計することで，この値を周辺混雑度 $conf(s)$ とする．

$$conf(s) = \sum_{m \in M} overlap(t(s), m) \times pop(m) \qquad (1)$$

ここで，$overlap(t(s), m)$ はあるメッシュと対象エリアの重複割合，$pop(m)$ はメッシュの人口である．

周辺混雑度において，捉え方は人それぞれである．たとえば，周辺混雑度が 100 の場合，混雑していると捉える人もいれば，混雑していないと捉える人もいる．そのため，相対混雑度は，対象スポットが近辺の観光スポットの周辺混雑度に対してどの程度混雑しているのかを計算する．具体的には，じゃらん net のデータを用いて，対象スポットがある市町村内の観光スポット集合 A を取得する．取得した観光スポット集合 A に属する各観光スポット a の周辺混雑度を計算し，その平均を算出する．最後に対象スポットの周辺混雑度を近辺の観光スポットの周辺混雑度の平均で除算し，これを相対混雑度 $r_conf(s)$ とする．

$$r_conf(s) = conf(s) \times \frac{|A|}{\sum_{a \in A} conf(a)} \qquad (2)$$

3.4　レビュー数，相対レビュー数の計算

レビュー数は周辺混雑度同様に，じゃらん net のデータを用いて取得する．被験者が訪れた観光スポットのレビュー数を取得し，この値をレビュー数 $rev(s)$ とする．

相対レビュー数は相対混雑度と同様，対象スポットが近辺の観光スポットのレビュー数に対してどの程度数が多いかを計算する．対象スポットがある市町村内の観光スポット集合 A を取得する．取得した観光スポット集合 A に属する各観光スポット a のレビュー数の平均を計算する．最後に対象スポットのレビュー数を近辺の観光スポットのレビュー数の平均で除算し，これを相対レビュー数 $r_rev(s)$ とする．

$$r_rev(s) = rev(s) \times \frac{|A|}{\sum_{a \in A} rev(a)} \qquad (3)$$

4　実　　験

4.1　実 験 方 法

今回実験を行う目的は，以下の仮説を検証することである．

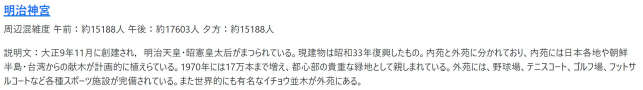


図 4　基本情報と周辺混雑度

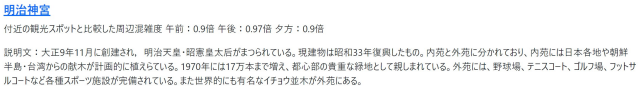


図 5　基本情報と相対混雑度

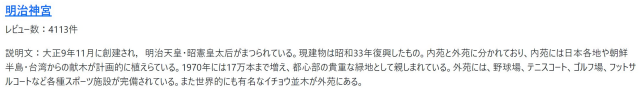


図 6　基本情報とレビュー数

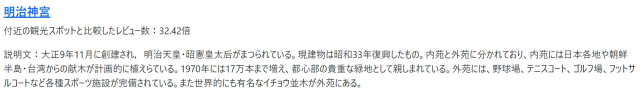


図 7　基本情報と相対レビュー数

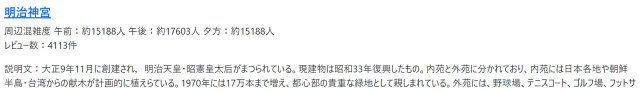


図 8　基本情報と周辺混雑度，レビュー数

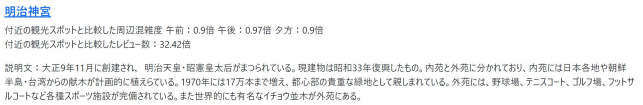


図 9　基本情報と相対混雑度，相対レビュー数

仮説 1　周辺混雑度を示すことで，旅行計画時の行き先に影響を与える
仮説 2　レビュー数を示すことで，旅行計画時の行き先に影響を与える
仮説 3　周辺混雑度とレビュー数は，旅行計画時の行き先への影響に差がある

仮説を検証するために，実験システムの UI において以下の 7 パターンの表示で実験を行う．被験者には，7 パターンから 1 つを振り分け，旅行計画を立ててもらう．

- 基本情報: スポット名と説明文 (図 3)
- 基本情報と周辺混雑度 (図 4)
- 基本情報と相対混雑度 (図 5)
- 基本情報とレビュー数 (図 6)
- 基本情報と相対レビュー数 (図 7)
- 基本情報と周辺混雑度，レビュー数 (図 8)
- 基本情報と相対混雑度，相対レビュー数 (図 9)

基本情報のみを表示する目的は，仮説を検証するためである．基本情報と周辺混雑度を表示する目的は，基本情報のみを表示した場合の周辺混雑度と，基本情報と周辺混雑度を表示した場合の周辺混雑度を用いて仮説 1 を検証するためである．基

表 2　選択された出発地点の数

	札幌駅	東京駅	横浜駅	名古屋駅	京都駅	大阪駅	博多駅
基本情報のみ	0	6	0	3	0	1	0
周辺混雑度	0	1	2	3	4	0	0
相対混雑度	1	3	1	0	3	2	0
レビュー数	1	6	2	0	0	1	0
相対レビュー数	0	3	2	0	1	2	2
周辺混雑度とレビュー数	2	2	2	1	0	2	1
相対混雑度と相対レビュー数	1	4	0	1	2	2	0
合計	5	25	9	8	10	10	3

本情報と相対混雑度を表示する目的は，被験者によって周辺混雑度の捉え方が異なるためであり，仮説 1 を検証するためである．基本情報とレビュー数を表示する目的は，基本情報のみを表示した場合のレビュー数と，基本情報とレビュー数を表示した場合のレビュー数を用いて仮説 2 を検証するためである．基本情報と相対レビュー数を表示する目的は，基本情報と周辺混雑度を表示した場合と同様である．基本情報と周辺混雑度，レビュー数を表示する目的は，基本情報のみ表示した場合の周辺混雑度とレビュー数，基本情報と周辺混雑度，レビュー数を表示した場合の周辺混雑度とレビュー数を用いて仮説 3 を検証するためである．基本情報と相対混雑度，相対レビュー数を表示する目的は，基本情報と周辺混雑度を表示した場合と同様である．

被験者には，実験システムを用いて 1 日の旅行計画を立ててもらう．旅行計画として，9 時から 11 時台で訪れたい観光スポット，12 時から 14 時台で訪れたい観光スポット，15 時から 17 時台で訪れたい観光スポットをそれぞれ 1 つずつ選択してもらう．被験者はクラウドソーシングサイトを用いて 70 名を募集し，7 パターンの表示のいずれかに振り分けた．

4.2 評価方法

評価の対象として，被験者が選択した観光スポットの周辺混雑度，レビュー数の順位の中央値を用いる．順位を用いた理由は，出発地点によって観光スポットの周辺混雑度やレビュー数に偏りがあり，被験者が選択した出発地点がパターンごとでばらつきがあったからである．また，1 パターンあたりの被験者数が 10 人であるため，外れ値の影響が少ない中央値を用いた．被験者が選択したパターンごとの出発地点の数を表 2 に示す．被験者が選択した出発地点の周辺混雑度とレビュー数の順位として，まず被験者が選択した条件である月，観光スポットのエリア，カテゴリを取得する．次に，同じ条件を持つ観光スポットと，観光スポットの周辺混雑度，レビュー数を取得し，周辺混雑度とレビュー数ごとで降順に並べ，上から順位をつける．最後にパターンごとの被験者が選択した観光スポットの周辺混雑度とレビュー数の順位の中央値を算出する．

仮説 1 を検証するために，実験システムにおいて，基本情報のみ表示の場合と基本情報と周辺混雑度を表示の場合，それぞれのパターンで選択された観光スポットの周辺混雑度の順位の中央値より，観光スポットの混雑状況が被験者の観光スポットの選択に影響を与えているかを検証する．

仮説 2 を検証するために，実験システムにおいて基本情報のみ表示の場合と基本情報とレビュー数を表示の場合，それぞれのパターンで選択された観光スポットのレビュー数の順位の中央値より，観光スポットのレビュー数が被験者の観光スポットの選択に影響を与えているかを検証する．

表 3　周辺混雑度の順位の中央値，レビュー数の順位の中央値

	周辺混雑度の順位の中央値	レビュー数の順位の中央値
基本情報のみ	22	2.5
周辺混雑度	12	4
相対混雑度	10	5
レビュー数	16	4.5
相対レビュー数	12	5
周辺混雑度とレビュー数	7	5.5
相対混雑度と相対レビュー数	9	3.5

仮説 3 を検証するために，実験システムにおいて基本情報のみ表示の場合と基本情報と周辺混雑度，レビュー数を表示の場合，それぞれのパターンで選択された観光スポットの周辺混雑度とレビュー数の順位の中央値より，観光スポットの周辺混雑度とレビュー数は被験者の観光スポットの選択への影響に差があるのかを検証する．

5 結　　果

被験者実験の結果，合計 70 名の被験者から，選択した観光スポットの周辺混雑度，相対混雑度，レビュー数，相対レビュー数のデータを，7 種の表示パターン別で収集した．

それぞれのパターンによって選択された観光スポットの周辺混雑度の順位の中央値，レビュー数の順位の中央値を表 3 に示す．

表 3 より，基本情報のみ表示した場合と，基本情報と周辺混雑度を表示した場合では，周辺混雑度を表示した場合の方が周辺混雑度の順位の中央値が低くなっている．これより，周辺混雑度を表示すると，より混雑している観光スポットが選択されるということがわかる．基本情報のみ表示した場合と，基本情報とレビュー数を表示した場合では，レビュー数を表示した場合の方がレビュー数の順位の中央値が高くなっている．これより，レビュー数を表示すると，よりレビュー数の少ない観光スポットが選択されるということがわかる．基本情報と周辺混雑度を表示した場合と，基本情報と相対混雑度を表示した場合では，相対混雑度を表示した場合の方が周辺混雑度の順位の中央値が低くなっている．また，基本情報とレビュー数を表示した場合と，基本情報と相対レビュー数を表示した場合では，相対レビュー数を表示した場合の方がレビュー数の順位の中央値が高くなっている．これらより，相対化することでより被験者の選択に影響を与える可能性があるということがわかる．

仮説 1 の検証について，表 3 より，基本情報のみ表示したときに選択された観光スポットの周辺混雑度の順位の中央値は 22 位に対して，基本情報と周辺混雑度を表示したときに選択された観光スポットの周辺混雑度の順位の中央値は 12 位であった．これより，周辺混雑度は被験者により混雑している観光スポットを選択させる影響がある．これは混雑している観光スポットは被験者にとって魅力的に見えるからと考える．また，基本情報のみ表示したときに選択された観光スポットの周辺混雑度の順位の中央値と，基本情報と周辺混雑度を表示したときに選択

された観光スポットの周辺混雑度の順位の中央値の結果に対して,「2 標本の母中央値に差はない」という帰無仮説を立て，有意水準 0.05 で中央値検定を行った．その結果，P 値は 0.027 と有意水準を下回り，帰無仮説が棄却され，有意な差があることが確認できた．

仮説 2 の検証について，表 3 より，基本情報のみ表示したときに選択された観光スポットのレビュー数の順位の中央値は 2.5 位に対して，基本情報とレビュー数を表示したときに選択された観光スポットのレビュー数の順位の中央値は 4.5 位であった．これより，レビュー数は被験者によりレビュー数が少ない観光スポットを選択させる影響がある．しかし，これらのレビュー数の順位の中央値の差は 2 位であり，差が小さい．これは，被験者はレビュー数が少なすぎない観光スポットを選択するからではないかと考える．また，基本情報のみ表示したときに選択された観光スポットのレビュー数の順位の中央値と，基本情報とレビュー数を表示したときに選択された観光スポットのレビュー数の順位の中央値の結果に対して,「2 標本の母中央値に差はない」という帰無仮説を立て，有意水準 0.05 で中央値検定を行った．その結果，P 値は 0.44 と有意水準を上回り，帰無仮説が棄却されず，有意な差は確認できなかった．

仮説 3 の検証について，表 3 より，基本情報のみ表示したときに選択された観光スポットの周辺混雑度の順位の中央値は 22 位に対して，基本情報と周辺混雑度，レビュー数を表示したときに選択された観光スポットの周辺混雑度の順位の中央値は 7 位であり，順位の中央値に 3 倍以上の差があった．基本情報のみ表示したときに選択された観光スポットのレビュー数の順位の中央値は 2.5 位に対して，基本情報と周辺混雑度，レビュー数を表示したときに選択された観光スポットのレビュー数の順位の中央値は 5.5 位であり，順位の中央値の差は 3 倍以下であった．これより，周辺混雑度がレビュー数より旅行計画時の行き先に影響を与える．これは混雑している観光スポットは被験者にとって魅力的に見え，また，被験者はレビュー数よりも周辺混雑度を優先するからと考える．基本情報のみ表示したときに選択された観光スポットの周辺混雑度の順位の中央値と，基本情報と周辺混雑度，レビュー数を表示したときに選択された観光スポットの周辺混雑度の順位の中央値の結果に対して,「2 標本の母中央値に差はない」という帰無仮説を立て，有意水準 0.05 で中央値検定を行った．その結果，P 値は 0.017 とと有意水準を下回り，帰無仮説が棄却され，有意な差があることが確認できた．基本情報のみ表示したときに選択された観光スポットのレビュー数の順位の中央値と，基本情報と周辺混雑度，レビュー数を表示したときに選択された観光スポットのレビュー数の順位の中央値の結果に対して,「2 標本の母中央値に差はない」という帰無仮説を立て，有意水準 0.05 で中央値検定を行った．その結果，P 値は 0.19 とと有意水準を上回り，帰無仮説が棄却されず，有意な差は確認できなかった．

6 おわりに

本研究では，被験者に実験システムを用いて 1 日の旅行計画を立ててもらい，選択された観光スポットの周辺混雑度とレビュー数の順位の中央値を用いて，観光スポットの混雑状況が被験者の観光スポットの選択に影響を与えているかを検証した．実験結果より，周辺混雑度を表示するとより混雑している観光スポットが選択されることがわかり，レビュー数はどの場面においてもレビュー数が少なすぎない観光スポットが選択されることがわかった．また，周辺混雑度がレビュー数より旅行計画時の行き先に影響を与えることがわかった．しかし，今回は被験者数が少なく，分析内容が限定されていた．今後の課題として，実際に旅行計画を立てるシチュエーションに近い実験環境を用意した上でデータ数を増やし，被験者や観光スポットの特徴を用いて，被験者の観光スポット選択における混雑状況の影響についてさらに分析していく．また，混雑回避志向，知名度優先志向などの被験者の選択の特性について調査可能な設問と合わせた実験を設計する．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 前川丈幸, 笠原秀一, 馬強. スポットの混雑を考慮した観光ルート推薦. 第 14 会データ工学と情報マネジメントに関するフォーラム, No. B41-3, pp. 1–6, mar 2022.

[2] Hidaka Masato, Kanaya Yuki, Kawanaka Shogo, Matsuda Yuki, Nakamura Yugo, Suwa Hirohiko, Fujimoto Manato, Arakawa Yutaka, and Yasumoto Keiichi. On-site trip planning support system based on dynamic information on tourism spots. *Smart Cities*, Vol. 3, No. 2, pp. 212–231, 2020.

[3] 野本輝, 上野史, 太田学. 観光レビュー文を用いた穴場スポットの発見. 第 14 会データ工学と情報マネジメントに関するフォーラム, No. B43-3, pp. 1–6, mar 2022.

[4] 野本輝, 太田学. 観点マッチングによる認知度に依存しない観光スポット推薦. 第 14 会データ工学と情報マネジメントに関するフォーラム, No. 14, pp. 1–7, mar 2022.

[5] 前川丈幸, 笠原秀一, 馬強. 混雑を考慮した観光ルート生成に関する研究. 第 15 会データ工学と情報マネジメントに関するフォーラム, No. 5b-3-4, pp. 1–9, mar 2023.

[6] Isoda Shogo, Hidaka Masato, Matsuda Yuki, Suwa Hirohiko, and Yasumoto Keiichi. Timeliness-aware on-site planning method for tour navigation. *Smart Cities*, Vol. 3, No. 4, pp. 1383–1404, 2020.

[7] 栗山恭嘉, 村田佳洋, 柴田直樹, 安本慶一, 伊藤実. 俊哉観光地における混雑状況を考慮した多数ユーザ同時巡回スケジューリング手法. 情報処理学会論文誌, pp. 885–898, mar 2010.

[8] 武正憲, 和田茂樹. 東京都式根島の海水浴場における混雑度が利用者の混雑感及び満足感に与える影響. 環境情報科学 学術研究論文集 34, pp. 31–36, mar 2020.

[9] 大村貴信, 鈴木健太, Panote Siriaraya, 栗達, 河合由起子, 中島伸介. スポット属性を考慮した実空間のユーザ行動分析に基づく潜在的興味分析の検証. 第 14 会データ工学と情報マネジメントに関するフォーラム, No. D33-4, pp. 1–20, mar 2022.

[10] 藤井一輝, 難波英嗣, 竹澤寿幸, 石野亜耶. 旅行ブログエントリの属性に基づいた旅行者の行動分析. 第 7 会データ工学と情報マネジメントに関するフォーラム, No. G7-5, pp. 1–8, mar 2015.

ユーザの性格タイプの差異に基づく災害時行動促進情報理解の分析

若杉　広介†　　鈴木　　優††　　灘本　明代†††

† 甲南大学大学院自然科学研究科　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
†† 岐阜大学工学部　〒 501-1193 岐阜県岐阜市柳戸 1-1
††† 甲南大学知能情報学部知能情報学科　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †m2324004@s.konan-u.ac.jp, ††suzuki.yu.r4@f.gifu-u.ac.jp, †††nadamoto@konan-u.ac.jp

あらまし　災害時 SNS 上ではユーザに対して行動を促進したり抑制したりする情報（以下，行動促進情報）が多数投稿されている．このような情報はユーザの行動に様々な影響を与えていると考えられる．SNS は多数のユーザが閲覧しておりその性格は様々であるため，ユーザの性格によっては情報の捉え方が異なると考えらえる．そこで本研究ではユーザの性格タイプに着目し，その性格により行動促進情報のとらえ方がどのように異なるかを分析する．ここではユーザの性格タイプとして BigFive 尺度を用いる．具体的には，行動促進情報を内容毎に「推奨」「行動抑制」「励まし」「願望」の 4 種類に分類する．そしてユーザをアンケート調査により BigFive 尺度の性格タイプに分類し，ユーザの性格タイプ毎にこれら 4 種類の行動促進をどのようにとらえたかの実験を行い，結果を分析する．

キーワード　SNS，災害，BigFive 性格特性

1　はじめに

近年，台風や豪雨・大雨，地震といった災害が多く発生している．災害時には「〜してください」や「〜しないでください」など様々な呼びかけが行われている．このような他者からの呼びかけは人の避難行動を促すにあたりとても重要な役割を果たすことが指摘させている [1]．しかしながら，呼びかけが逆効果になる場合がある．防災心理学者の木村ら [2] によると，例えば,「河川が氾濫しました危ないので近寄らないでください」のような人に行動を抑制する情報が発信された場合，一般的な人はこの情報を見た時「危ないから近寄らないでおこう」と考える．一方，好奇心や責任感の強い人はこの情報を見た時「どうなっているのか気になるから見にいこう」や「非常時だからこそ行って様子を確かめないと」と考え，呼びかけられた内容と真逆の行動をし呼びかけがむしろ被害を拡大させ逆効果になる場合があることがわかっている．

近年，災害発生時 SNS 上には様々な情報が投稿されており，SNS より情報を取得する人が増えている．大規模災害発生時，ユーザに的確かつ必要な情報を迅速に提供することは重要である．様々な SNS が存在するが，その中でも X(旧 Twitter) は一般ユーザや被災者のみならず，市区町村などの行政機関からの情報発信がなされている．X 上には様々な情報が発信されているが，その情報の中には閲覧ユーザに「河川が氾濫しましたので，河川に近寄らないでください」や「今は大変な時ですが，一緒に頑張りましょう」など行動を促進するような情報がある．本研究ではこのように人に行動を促している情報や抑止しようとしている情報をまとめて行動促進情報と呼ぶ．我々は災害時にこのような行動促進情報を抽出して提示することはユーザの行動に影響を与えるため防災や減災に重要であると考え，行動促進情報の抽出に関する研究を行っている [1] [2] [3]．しかしながら，行動促進情報には「〜してください」などのように行動を呼びかけているものや「〜しないでください」などの行動を抑止しているものなど様々な趣旨がある．また，SNS は様々なユーザが閲覧しており，そのユーザの性格は様々である．そのため木村らが指摘したように，ユーザに提示した行動促進情報がそのユーザにとって適切なのかはわからない．例えば「避難所に避難してください」といった情報が発信された場合,「避難しよう」と考えるユーザがいる一方で，ユーザの性格によっては「避難しなくても大丈夫」と考え避難しない可能性もある．そこで，本研究では，ユーザの性格に着目し，行動促進情報を「推奨」「行動抑制」「励まし」「願望」の４つのラベルからなる促進軸に分類し，ユーザの性格毎に適切な行動促進情報の促進軸を分析する．この時，ユーザの性格測定の方法として，BigFive 性格特性を用いる．BigFive 性格特性とは人間の性格を表 1 に示すような「外向性」「協調性」「勤勉性」「神経症傾向」「開放性」の５つの性格特性（因子）によって説明する理論である．

表 1　ビックファイブの 5 つの性格特性（因子）

外向性	興味や関心が外界に向けられ刺激に敏感に反応し決断が速く，行動的で社交的な傾向
協調性	洞察力に優れていて、人間性を瞬時に把握し，仕事を上手くこなす傾向
勤勉性	責任感があり勉強や仕事に精力的に取り組むまじめな傾向
神経症傾向	落ち込みやすいなど感情面・情緒面で不安定な傾向
開放性	知的、美的、文化的に新しい経験に開放的な傾向

本研究の手順は以下の通りである．

（1）　行動促進情報を災害時のツイートから抽出する．

1：https://www.pref.hiroshima.lg.jp/uploaded/attachment/526989.pdf
2：https://president.jp/articles/-/71423

（2） クラウドソーシングを用いて行動促進情報の促進軸をアノテーションする．
（3） 行動促進情報を促進軸で分類する．
（4） 被験者ユーザの性格特性テストを行う．
（5） ユーザの性格特性毎に促進軸の影響を調査する．
（6） 調査結果を分析する．

本研究での災害は，大地震のような大規模災害ではなく，毎年起こる可能性が高い台風を対象とする．本研究により災害時にユーザの持つ性格に応じてどのような行動促進情報を提示すれば良いのかがわかり，ユーザへのより的確な情報提供が可能となることが期待できる．

以下，2 章で関連研究について述べ，3 章で行動促進情報の抽出手法について説明する．4 章で促進軸について述べて，促進軸分類モデルの作成の手法について説明する．5 章で実験手法について説明し実験を行う．6 章で実験結果を分析しその結果と考察を述べる．最後に，7 章で本論文のまとめと今後の課題について述べる．

2 関 連 研 究

災害時にツイートから重要な情報を抽出する研究は数多く行われている．Xiaodong ら [7] はツイート特有の言語的，感傷的，感情的特徴に着目し，ツイートを災害に関連するツイートとそうでないツイートの分類するモデルを提案している．Paul ら [5] は，2012 年から 2018 年に発生した台風に関するツイートを分析し，機械学習を用いて停電や通信障害に関するツイートのカテゴリ分類を行っている．そして，BERT を用いることで短時間かつ高精度でカテゴリ分類できることを示している．Yasin ら [16] は災害関連ツイートを機械学習を用いて「救助が必要」「DECW（障害者，高齢者，子供，女性の略）」「水が必要」「負傷」「病気」「洪水」の 6 つのラベルに分類している．この分類をすることでツイート発信者を支援するのに必要な情報を特定している．これらの研究はツイートの特徴を分析することで,「救助」や「寄付」,「津波」といった災害情報や被害情報でカテゴリ毎に分類している．これらの研究に対して本研究では人に行動を促進や抑制を促す情報を分類して，分析を行なっている点で異なる．

ユーザの性格に着目した研究も多く行われている．Volkel ら [9] はテキストメッセージにおいて，ユーザの性格によって絵文字の選択と解釈に違いをもたらすことを示している．Dey ら [10] はユーザに異なる感情的な要素を持つビデオ広告を見せて，実験を行なっている．その結果，ユーザの好むビデオ広告の感情はそのユーザによって大きく異なることを示している．これらの研究は様々な対象となったコンテンツでユーザの性格により解釈や反応の違いを分析している．それに対し本研究では災害時の情報を対象に分析を行なっている点が異なる．

Toriumi ら [11] はリツイートに着目した研究を行なっている．同じツイートをリツイートしているユーザは同じトピックに興味を持っていると仮定することで，災害時のユーザの要求に基づいた情報を効果的に分類できることを示した．この研究ではリツイートに着目して情報の分類している．それに対して本研究では行動促進情報に着目して分類している点で異なる．

さらに，災害時のツイートを分析する研究も数多く行われている．Lu ら [6] はユーザの削除したツイートを分析するとこで，不適切な内容を 10 項目で分類できることを示している．この研究はユーザの評判やプライバシーに影響を与える情報を分析しているのに対して，本研究では災害時にユーザの行動に影響を与える情報を分析している点で異なる．David ら [12] はツイートを 1) 助けが必要，2) 誰かを探しているなどの 11 項目に分類して感情分析を行うことで，災害時のツイートは主に支援や提案の内容が中心となっていることを示している．Yamada ら [17] は，2018 年の西日本豪雨災害時のツイートを対象に，ツイート数の推移やハッシュタグ・絵文字の利用傾向，リツイート数，ニュース記事の URL 付きツイート数に着目した分析を行っている．Nishikawa ら [18] は，2018 年西日本豪雨災害時に投稿された救助要請のハッシュタグが付いたツイートの内容や傾向を分析している．そして，救助要請のハッシュタグが付いたツイートは地域名やランドマーク，住所が含まれており，救助が必要な場所を特定する際に有用であることを示している．これらの研究は，ハッシュタグ，単語や感情に着目して，ツイートの分析を行っている．それに対し本研究では，促進軸毎にツイートを分類して性格特性毎に比較分析している点で異なる．

3 行動促進情報の抽出手法

本研究では，これまで我々が提案してきた行動促進情報の抽出 [1] [3] で最も良い精度を出した RoBERTa [19] を用いる．RoBERTa の実装には機械学習ライブラリである PyTorch [3]を採用する．Pre-training モデルには，RoBERTa 日本語 Pretrained モデル [4]を用いる．単語の特徴ベクトルは，URL と X のユーザ名を除いたツイートに対して Juman++ [5]を用いて形態素解析を行い，RoBERTa 日本語 Pretrained モデルに入力して，RoBERTa の最終層から取得した分散表現を用いる．取得した分散表現は全結合層への入力となり，Fine-Tuning を行い，行動促進情報か否かを判別する．このとき，RoBERTa の各種ハイパーパラメータはグリッドサーチにより決定し，隠れ層は 12，ベクトルサイズは 768，バッチサイズは 32，エポック数は 5，学習率は 0.0005，ドロップアウト率は 0.1 とする．また，オプティマイザーには AdamW [20] を採用する．Fine-tuning には 2019 年に関東地方（特に千葉県の房総半島）を襲った大型台風である台風 15 号 [6]のデータ（2019 年 9 月 6 日〜9 月 18 日）から，行動促進情報，非行動促進情報ともに 12,215 件で合計 24,430 件を用いる．テストデータに行動促進情報，非行動促進情報からそれぞれ 20%のツイートを無作為に抽出し，その結果を表 2 に示す．これにより，本研究ではこのモデルを使用して行動促進情報を災害時のツイートでたから抽出する．

3：https://pytorch.org/
4：https://huggingface.co/nlp-waseda/roberta-base-japanese
5：https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
6：https://ja.wikipedia.org/wiki/令和元年房総半島台風

表 2　行動促進情報分類モデルの性能

model	precision	recall	F 値	AUC
RoBERTa	0.900	0.949	0.924	0.973

表 3　各台風データ数

No	災害	行動促進情報	実験で使用する行動促進情報
1	2019 年台風 15 号	12,215	5,000
2	2022 年台風 14 号	67,378	5,000

4　行動促進ラベル分類

4.1　行動促進ラベルの定義

山本ら [2] は大規模災害時の行動促進情報を抽出してその特徴を分析を行っている．その結果，行動促進の促進方法により下記に説明する 5 つのラベルからなる促進軸に分類できることを提案している．以下に山本らの提案している 5 つのラベルと述べる．

- 推奨ラベル

推奨ラベルとは「〜してください」や「〜に気をつけてください」といったユーザに対して行動を促している情報である．例えば,「近所の河川が増水しているので、今すぐ避難してください。」は推奨ラベルである．

- 行動抑制ラベル

行動抑制ラベルとは「〜に近づかないようにしましょう」や「〜に近づいてはいけません」といったユーザに対して行動の抑制や制止，禁止している情報である．例えば,「大雨で河川が増水しているので近づかないようにしましょう。」は行動抑制ラベルである．

- 励ましラベル

励ましラベルとは「〜に頑張りましょう」や「元気出して行きましょう」といったユーザに対して元気づける，力づけている情報である．例えば,「停電や食糧不足が続きますが、一緒に頑張りましょう。」は励ましラベルである．

- 願望ラベル

願望ラベルとは「〜なので助けてほしい」や「支援をお願いします」といった情報の発信者が願い望んでいる情報である．例えば,「食料が足りてません。早く助けにきてください。」は願望ラベルである．

- その他のラベル

その他のラベルとは上記のどのラベルにも当てはまらない情報のことである．例えば,「〜をお待ちください」はその他のラベルである．

本研究では山本らの提案する 5 つのラベルの内，その他を除いた「推奨」「行動抑制」「励まし」「願望」の 4 つのラベルを用いる．

4.2　行動促進ラベルの分類手法

本研究では行動促進情報を 4 つの行動促進ラベルに分類する．この時，一つの行動促進情報（ツイート）に対して，複数の行動促進ラベルが含まれている場合がある．例えば,「台風が近づいてきています。避難してください。またお近くにこの情報を必要としている方に情報提供をお願いします。」といった情報はユーザに対して避難の行動を促している「推奨」ラベルとユーザに情報の拡散を要請している「願望」ラベルが含まれている．そのため，作成する分類モデルは行動促進情報に複数のラベルを割り当てることのできるマルチラベル分類が適していると考える．次に以下 2 つのマルチラベル分類モデルを作成する．尚，2 つのモデルとも，事前学習モデルには RoBERTa 日本語 Pretrained モデルを使用する．特徴ベクトルは URL と X のユーザ名を除いたツイートに対して Juman++ を用いて形態素解析を行い，RoBERTa 日本語 Pretrained モデルに入力して，RoBERTa の最終層から取得した分散表現を用いる．

（1）　単体モデル

「推奨」「行動抑制」「励まし」「願望」それぞれの分類モデルを作成し，行動促進情報に対して各々の単体のモデルで判定することでマルチラベル分類を行う．単体モデルでは RoBERTa 日本語モデルの最終層から CLS トークンに対応した分散表現を用いる．取得した分散表現は全結合層への入力となり，Fine-Tuning を行い，対象ラベルか非対象ラベルかの 2 値分類を行う．ここでは，単体モデルをすべての 4 つのラベルに対して作成し，マルチラベルに対応する．

（2）　マルチモデル

一つのモデルで一度に行動促進情報に対して 4 つのラベルからなる促進軸のマルチラベル分類を行う．マルチモデルではモデルの最終出力層の CLS トークンに対応する出力に対して，出力が各ラベルに対応する 4 次元の全結合層を追加してシグモイド関数を適応し各ラベルが付く確率値を算出する．そして，この確率値が 0.5 を超えたもの全てをその情報のもつラベルとする．

4.3　行動促進ラベル分類モデルの評価実験

提案した 2 種類の分類モデルの評価実験を行う．

データ

分類モデルの評価実験データは（1）2019 年に発生した大型台風である台風 15 号（抽出期間 2019 年 9 月 6 日〜18 日）と（2）2022 年に発生した大型台風である台風 14 号（2022 年 9 月 16 日〜27 日）の 2 つの台風から，提案した行動促進情報抽出手法を用いて行動促進情報を抽出した．これにより抽出された行動促進情報の数を表 3 に示す．このうち各々のデータから無作為に 5,000 件ずつ抽出した合計 10,000 件の行動促進情報を用いる．

次に，クラウドソーシングにより，抽出した 10,000 件の行動促進情報を対象に促進軸の 4 つのラベルのアノテーションを行う．被験者は 10 人で評価項目は促進軸である「推奨」「行動抑制」「励まし」「願望」の 4 種類とする．この時，被験者は 1 つの情報に対して複数評価が可能とする．ここで，1 つの行動促進情報（ツイート）に対して 10 人の被験者が判定し，6 人以上が同じラベルをアノテーションした情報をその行動促進情報がもつラベルとする．この条件を満たさない情報を「その他」のラベルに分類する．分類結果を表 4 に示す．

表 4　促進軸の分類結果

ラベル	個数
推奨	8,078
行動抑制	122
励まし	230
願望	617
推奨と行動抑制	132
推奨と励まし	249
推奨と願望	163
願望と励まし	10
行動抑制と願望	6
推奨と願望と励まし	2
その他	391

表 5　アンダーサンプリング後のデータ数

ラベル	対象ラベル	非対象ラベル	合計
推奨	985	985	1,970
行動抑制	260	260	520
励まし	491	491	982
願望	798	798	1,596

実験：単体モデルとマルチモデルの比較実験

単体モデルとマルチモデルを比較する実験を行う．この時，実験データをアンダーサンプリングした場合とオーバーサンプリングした場合の比較も同時に行う．

アンダーサンプリング

単体モデルを作成するために，上記のアノテーションで取得したデータのうち促進軸毎に対象の促進軸とその対象の促進軸を含まない情報に分類する．例えば,「推奨」モデルに入れるデータは「推奨」を含む情報と「推奨」を全く含まない情報に分類する．この時，情報はマルチラベルかどうかを考慮しない．例えば,「推奨」と「行動抑制」の 2 つを含むデータは「推奨」を含むデータとして扱い,「行動抑制」と「励まし」を含むデータは「推奨」を含まないので「推奨」を含まない情報として扱う．そして，対象の促進軸データと非対象の促進軸データを同程度の数にするためにアンダーサンプリングを行う． この操作を各ラベル毎に行い，結果を表 5 に示す．

オーバサンプリング

元々のデータが少ないため，アンダーサンプリング後のデータも少なくなっている．そこで，我々はデータのオーバーサンプリングを行う．具体的には，オーバーサンプリング手法の一つである EDA [13] を用いる．EDA のハイパーパラメータ α は 0.05，生成するデータ数 n_{aug} は 16 とする．EDA の実行には daaja[7]を用いる．

実験は Fine-Tuning に 80%の学習データを用い，残りの 20%をテストデータとした 5 交差検証を行う．

結果と考察

accuracy，precisionl，recall，F 値，AUC の結果表 6，7，8 と 9 に示す．アンダーサンプリングのみの場合（表 6,7）とオーバーサンプリングを行った場合（表 8,9）を比較すると，オーバーサンプリングを用いたモデルの方がアンダーサンプリングのみのモデルよりも高い精度で分類できていることがわかる．その結果，オーバーサンプリングを行った手法がよいことがわる．さらに単体モデル（表 8）とマルチモデル（表 9）を比較するとまた，単体モデルの方がマルチモデルよりも若干精度が高いことがわかる．しかしながら，単体モデルはそれぞれの 4 つのモデルで 1 つの情報を判定しているため，計算コストがかかる．そのため本研究では，高い分類精度を示したオーバーサンプリングありのマルチモデルを促進軸分類モデルとして採用する．

7：https://github.com/kajyuuen/daaja

表 6　(1) 単体モデル

model	accuracy	precision	recall	F 値	AUC
推奨	0.861	0.847	0.888	0.866	0.928
行動抑制	0.758	0.714	0.918	0.798	0.863
励まし	0.777	0.74	0.855	0.793	0.845
願望	0.894	0.902	0.882	0.891	0.952
平均	0.823	0.801	0.886	0.837	0.897

表 7　(2) マルチモデル

label	accuracy	precision	recall	F 値	AUC
推奨	0.878	0.893	0.949	0.920	0.926
行動抑制	0.961	0.512	0.515	0.512	0.930
励まし	0.914	0.650	0.500	0.545	0.914
願望	0.927	0.889	0.762	0.819	0.963
平均	0.920	0.736	0.682	0.699	0.933

表 8　(1) 単体モデル (オーバーサンプリングあり)

model	accuracy	precision	recall	F 値	AUC
推奨	0.966	0.961	0.972	0.966	0.992
行動抑制	0.99	0.989	0.992	0.99	0.998
励まし	0.969	0.961	0.978	0.969	0.993
願望	0.986	0.983	0.989	0.986	0.995
平均	0.978	0.973	0.983	0.978	0.994

表 9　(2) マルチモデル (オーバーサンプリングあり)

label	accuracy	precision	recall	F 値	AUC
推奨	0.958	0.973	0.969	0.971	0.981
行動抑制	0.992	0.93	0.952	0.94	0.996
励まし	0.971	0.886	0.912	0.898	0.981
願望	0.984	0.968	0.956	0.962	0.994
平均	0.976	0.939	0.947	0.943	0.988

5　ユーザの性格特性毎の行動促進情報の特徴分析

ユーザの性格特性に応じた適切な情報の抽出を行うために，まずユーザの性格特性を分類し，その後性格特性毎に行動促進情報の影響調査の為にアンケートを行う．このアンケート調査には，クラウドソーシングを用いる．

5.1　ユーザの性格特性の測定

性格特性測定方法

本研究ではユーザの性格特性を取得する為に，BigFive 性格特性によるアンケートを用いたユーザの性格特性の測定を行う．

表 10　性格特性判定の質問内容

番号	質問内容
1	活発で、外交的だと思う
2	他人に不満を持ち、もめ事を起こしやすいと思う
3	しっかりしていて、自分に厳しいと思う
4	心配性で、うろたえやすいと思う
5	新しいことが好きで、変わった考えを持つと思う
6	ひかえめで、おとなしいと思う
7	人に気をつかう．やさしい人間だと思う
8	だらしなく，うっかりしていると思う
9	冷静で，気分が安定していると思う
10	発想力に欠けた．平凡な人間だと思う

ここで本研究は被験者の回答負担を減らすために，小塩ら [4] が作成した 10 項目の質問で性格特性を測定できる尺度を用いる（表 10 参照）．この尺度において各性格特性は 1 から 14 の値を取る．

分析対象の性格特性の決定

スクリーニング調査を行い，実験対象者を決定した．まず，全体的な性格特性の傾向を把握するために，被験者 1,000 人を対象に表 10 で示すアンケート調査を行い，性格特性の判定を行った．

その結果を表 11 に示す．ここで，被験者の性格に偏りがあることがわかる．その為，表 11 から実際にアンケート調査の対象となる被験者を決定する．この時，アンケート結果に公平性を持たすために，被験者の年齢と性別を考慮する．そこで，一般的にマーケティングで用いられている 6 分類を用いる．この 6 分類は，20 歳から 34 歳の男性を M1 層，女性を F1 層，35 歳から 49 歳の男性を M2 層，女性を F2 層，50 歳以上の男性を M3 層で女性を F3 層としている．さらに，ある性格特性を持っている人と持っていない人との比較分析を行うために，性格特性毎にその性格を持っている人と持っていない人を被験者とする．そこで，各層とも基本的に対象の性格特性の値が 10 以上の場合はその性格特性を持っているとし，4 以下の場合は対象の性格特性を持っていないとする．ここで,「協調性」「勤勉性」「神経症的傾向」が 4 以下の人数が極端に少ない為，この 3 つの性格特性の持っていない条件を 6 以下とした．その結果を表 12 に示す．表 12 に示す被験者をアンケート調査対象とする．

5.2　分析対象の決定

上記で決定した条件をもとに，再度アンケートを実施し，有効な回答を得られた 248 人を対象に分析を行う．各性格特性の持っている・持っていない条件を満たすデータの数を表 12 に示す．

実験のデータは X を対象とし，クエリとして「台風」を含む情報を抽出し，提案している行動促進抽出手法を用いて行動促進情報を抽出する．抽出した行動促進情報を我々の提案する分類手法を用いて分類を行なった．分類結果を表 13 に示す．

対象被験者に対し，それぞれの性格特性の人がどの分類の促進軸から影響を受けやすいかを分析するために，248 名の被験者に対しアンケートをおこなった．具体的には，2022 年台風 15 号を対象とし，我々の提案する手法を用いてラベル付けされた行動促進情報（表 13 参照）から，各促進軸 50 ツイートをランダムに抽出し被験者に提示する．そして被験者は各々の行動促進情報を読んだときに，推奨ラベルの情報は「行動したいと思ったか」，行動抑制ラベルの情報は「行動をやめたいと思ったか」，励ましラベルの情報は「励まされたか」，願望ラベルの情報は「要請に応えたいと思ったか」という質問に対し，それぞれ評価値として表 14 に示す 4 段階のリッカート尺度 [14] を用いる．4 段階中 3 以上の評価を得た情報を「適切な情報」とし，被験者毎に各促進軸の情報 50 件に対する「適切な情報」の数を求める．この時，被験者には事前に大規模台風時に行動に影響を及ぼす情報の分析を目的にアンケートを行っている旨を伝え，その上で，提示された情報を見て大規模発生時に台風被災者としてこの情報を読んだ人の気持ちになって各々の質問項目に答えるように指定した．対象の被験者の「適切な情報」の程度を算出し，性格特性を持っている，持っていない毎にその平均を表 15 に示す．

6　ユーザの性格特性毎の行動促進情報の特徴分析

本研究ではユーザの性格特性毎の行動促進情報の特徴分析を行うにあたり，Research Question を立てて分析を行う．また，分析手法には統計検定を用いる．本研究における調査結果は標本の母集団が正規分布に従うと仮定できない為，母集団の特徴に捉われず使用できるノンパラメトリックな検定を行う．

6.1　予 備 実 験

各ユーザーの性格特性に関する行動促進情報の特徴分析を行うための予備実験として，次の 2 つ検証を行う．(1) 性格特性の分析において複数の性格特性を考慮すべきか，(2) 性格特性に関わらず，促進軸がユーザに与える影響はどのようなものか．

(1) 性格特性の分析において複数の性格特性を考慮すべきか？

人は一つの性格特性を持つだけでなく，複数の性格特性を持つ場合がある．そこで，特徴分析をするにあたり．性格毎に分析をするのか，それとも複数の性格を考慮した分析をするのかのいずれかを決定する．そのために，各性格特性における相関係数を求めた．結果を表 16 に示す．表 16 に示すように性格特性ごとの相関は大きいものでも 0.42 であり弱い相関しかないことがわかる．そこで本研究では各性格特性間の影響を考慮せずに，各性格特性単体ごとの分析を行う．

(2) 性格特性に関係なく促進軸の影響はあるのか？

本研究では性格特性毎の行動促進情報の促進軸の影響について分析することを目的としている．そこで，そもそも性格特性に関係なく，誰にとっても行動促進情報の促進軸の影響があるかどうかを明らかにする．そのため，ここではユーザの性格特性は考慮しない．

4 つの促進軸間に適切だと思う情報の程度に差があるかどうかを調べるために分散分析を行う．分散分析の手法として，対応のある 3 群以上の比較を行えるノンパラメトリック検定であるフリードマン検定を用いる．両側 5%で検定を行い，帰無仮説

表 11　スクリーニング時の性格特性毎の人数（合計 1000 人）

	外向性		協調性		勤勉性		神経症的傾向		開放性	
	持っている	持っていない	持っている	持っていない	持っている	持っていない	持っている	持っていない	持っている	持っていない
区分	(10 以上)	(4 以下)	(10 以上)	(6 以下)	(10 以上)	(6 以下)	(10 以上)	(6 以下)	(10 以上)	(4 以下)
F1	19	29	51	21	19	28	48	28	17	24
F2	26	40	81	15	18	38	54	43	24	32
F3	36	25	125	11	18	37	47	59	20	32
M1	18	33	42	15	17	19	34	18	23	12
M2	19	34	61	18	13	35	40	37	23	23
M3	33	36	109	11	12	43	33	54	39	24

表 12　性格特性毎のアンケート対象の人数 (合計 248 人)

	外向性		協調性		勤勉性		神経症的傾向		開放性	
	持っている	持っていない	持っている	持っていない	持っている	持っていない	持っている	持っていない	持っている	持っていない
区分	(10 以上)	(4 以下)	(10 以上)	(6 以下)	(10 以上)	(6 以下)	(10 以上)	(6 以下)	(10 以上)	(4 以下)
F1	11	14	16	9	9	13	19	16	10	13
F2	11	13	19	11	9	16	17	15	10	9
F3	15	10	29	10	10	10	15	17	9	11
M1	11	14	13	12	10	10	16	10	9	8
M2	11	13	18	10	8	12	14	14	9	11
M3	14	11	24	9	8	13	13	12	11	10

表 13　実験データ数

災害	行動促進情報	推奨	行動抑制	励まし	願望
2022 年台風 15 号	51,599	22,291	588	1,983	2,959

表 14　評 価 項 目

そう思う	少しそう思う	あまり思わない	思わない
4	3	2	1

表 15　性格特性を持っている・持っていない評価の平均

性格特性	条件	推奨	行動抑制	励まし	願望
外向性	持っている	2.14	2.46	2.49	2.31
	持っていない	1.96	2.37	2.23	2.07
協調性	持っている	2.07	2.45	2.42	2.26
	持っていない	1.92	2.38	2.13	1.95
勤勉性	持っている	1.98	2.45	2.23	2.16
	持っていない	1.92	2.22	2.18	2.03
神経症傾向	持っている	2.04	2.45	2.30	2.15
	持っていない	2.14	2.45	2.41	2.25
開放性	持っている	2.22	2.44	2.54	2.35
	持っていない	1.80	2.31	2.16	1.96

は「促進軸間において適切だと思う情報の程度に差はない」とし，対立仮説には「促進軸間において適切だと思う情報の程度に差がある」とする．

結果を表 17 に示す．分散分析の結果は P 値が 5.00×10^{-16} となり，有意差ありとなった．よって帰無仮説が棄却され，促進軸間において適切だと思う程度には差があると言える．この結果からユーザの性格特性を考慮しない場合，促進軸の種類によって適切だと思う程度には差あることが分かった．これにより，性格特性毎の行動促進情報の促進軸の影響について分析することは妥当であると言える．

6.2　性格特性の有無による比較

我々は先ほど示した表 15 の結果から，我々は性格特性を持っているユーザはその性格特性を持っていないユーザよりも適切だと思う情報の程度が大きくなるのではないかと考えた．そこ

表 16　被験者 1,000 人の性格特性の相関係数

	外向性	協調性	勤勉性	神経症傾向	開放性
外向性	1	0.03	0.08	-0.40	0.42
協調性	0.03	1	-0.13	-0.39	0.09
勤勉性	0.08	-0.13	1	0.39	0.13
神経症傾向	-0.40	-0.39	0.39	1	-0.30
開放性	0.42	0.09	0.13	-0.30	1

表 17　促進軸間の分散分析の結果

被験者数	248
比較する群数	4
検定統計量	74.35
P 値	5.00×10^{-16}

で性格特性毎に対象の性格特性を持っているユーザのアンケート結果を持っていないユーザの結果を比較し促進軸毎に差があるのかを分析する．分析手法として，等分散でないデータに対して対応のない対応のない 2 群間の比較を行えるブルンナー・ムンツェル検定を用いる．片側 5%で検定を行い，帰無仮説は「対象特性の有無において適切だと思う程度に差はない」とし，対立仮説には「対象特性を持っている方が持っていない方より適切だと思う程度が大きい」とする．

結果と考察

結果を表 18 に示す.「外向性」を持っているユーザは「励まし」「願望」の P 値が 0.05 以下であり有意差が認められた．これは「外向性」を持っている人は積極的にコミュニケーションをとる傾向があるため「外向性」を持っていないユーザよりも「〜頑張りましょう」といったユーザ同士のコミュニケーションを図るような情報や「〜お願いします」といった要請に対して応えやすいからであると考える．

「協調性」を持っているユーザにおいて「励まし」「願望」の P 値が 0.05 以下であり有意差が認められる．このことから「協調性」を持っているユーザは「協調性」を持っていないユーザよりも「励まし」と「願望」において適切だと思う情報が多い

表 18　特性を持っている・持っていないユーザ間における比較検定

性格特性	統計	推奨	行動抑制	励まし	願望
外向性	統計量	-1.101	-0.680	-1.922	-1.754
	P 値	0.136	0.249	0.028	0.041
協調性	統計量	-1.060	-0.681	-2.148	-2.739
	P 値	0.146	0.248	0.017	0.004
勤勉性	統計量	-0.501	-1.708	-0.346	-1.166
	P 値	0.309	0.045	0.365	0.123
神経症傾向	統計量	0.476	0.030	0.798	0.543
	P 値	0.682	0.512	0.787	0.706
開放性	統計量	-2.380	-0.819	-2.340	-2.509
	P 値	0.010	0.207	0.011	0.007

ことが分かる．これは「協調性」を持っているユーザは周りと協力しながら物事を進めていく傾向があるため,「〜お願いします」や「〜頑張りましょう」といった情報を見ると周りの人に応えないといけないと感じたからだと考える．

「勤勉性」を持っているユーザは「行動抑制」の P 値が 0.05 以下であり有意差が認められる．これは「勤勉性」を持っているユーザは慎重に行動する傾向があるので,「〜をしないで」といった情報で行動することにより慎重になりやすいためであると考える．

「神経症的傾向」では各促進軸において有意差が認められなかった．このため「神経症傾向」の有無は行動促進情報を適切だと思う程度に差がない事がわかる．

「開放性」を持っているユーザは「推奨」「励まし」「願望」の P 値がそ 0.05 以下であり有意差が認められる．これは「開放性」を持っているユーザは新しい経験を受け入れやすいく行動力が高い傾向があるため,「開放性」を持っていないユーザよりも促進軸の情報に対して行動を起こしやすかったからだと考える．

7　まとめと今後の課題

本研究では，SNS 上の大規模台風時におけるユーザの性格特性に着目し,「推奨」「行動抑制」「励まし」「願望」の促進軸間で適切だと思う情報に反応の違いがあるのかの分析を行った．具体的には，災害時に X 上の行動促進情報を抽出し，取得した行動促進情報を促進軸の分類モデルを提案して分類を行った．そして，分類した情報をユーザに提示して，性格特性を持っている場合とその性格特性を持っていない場合とで適切だと感じた情報の評価に違いがあるのかを分析を行った．その結果，促進軸毎にユーザの適切だと思う情報に違いがある事がわかった．性格特性毎にどの情報を適切だと評価しやすいのかが分かった．具体的には「外向性」もしくは「協調性」を持っているユーザには「励まし」「願望」の情報,「勤勉性」を持っているユーザには「行動抑制」,「開放性」を持っているユーザには「推奨」「励まし」「願望」の情報を提示した場合，適切な情報だと評価しやすい事がわかった．一方で,「神経症傾向」を持つユーザは持たないユーザと行動促進情報に対して反応の違いを認める事ができなかったため，どの情報を適切な情報だと評価しやすいのかが分からなかった．今後の課題として,「神経症的傾向」を持つユーザが適切だと思う情報の分析を行う．また，今回は性格特性を持っているユーザを対象としたが，対象の性格特性を持たないユーザにとっての適切な情報も分析する．

謝　　辞

論文の一部は私学助成金（大学間連携研究助成金）の助成によるものである．ここに記して謹んで感謝の意を表する．

文　　献

[1] Yoshiki Yoneda, Yu Suzuki, Akiyo Nadamoto “Detection of Behavioral Facilitation information in Disaster Situation ” The 21st International Conference on Information Integration and Web-based Applications and Services (iiWAS2019), pp.255–259, 2019.

[2] 山本楓登，鈴木優，灘本明代，“災害被災者にとって有用な SNS 上の口コミ情報の分類と分析”，第 14 回データ工学情報管理フォーラム（DEIM2022）”，B33-2、12pages，2022

[3] Kosuke Wakasugi, Futo Yamamoto, Yu Suzuki and Akiyo Nadamoto, “Feature analysis of Regional Behavioral Facilitation Information based on Source Location and Target People in Disaster”, The 25th International Conference on Big Data Analytics and Knowledge Discovery (DAWAK 2023), pp.538–543, 2023.

[4] Atsushi Oshio and Shingo Abe and Pino Cutrone, “Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J),” The Japanese Journal of Personality 2012,pp.40–52, 2012.

[5] U. Paul, and A. Ermakov, M. Nekrasov, V. Adarsh and E. Belding, “#Outage: Detecting power and communication outages from social networks,” Proc. of the Web Conference 2020, no.11, pp.1819–1829, 2020.

[6] Zhou, Lu and Wang, Wenbo and Chen, Keke, “Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones,” Proceedings of the 25th International Conference on World Wide Web, no.10, pp.603-612, 2016.

[7] Ning, Xiaodong and Yao, Lina and Benatallah, Boualem and Zhang, Yihong and Sheng, Quan Z. and Kanhere, Salil S., “Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization,” Association for Computing Machinery, no.3, 2019.

[8] Kabir, M. Yasin and Gruzdev, Sergey and Madria, Sanjay, “STIMULATE: A System for Real-time Information Acquisition and Learning for Disaster Management,” 2020 21st IEEE International Conference on Mobile Data Management (MDM), pp.186-193, 2020.

[9] Völkel, Sarah Theres and Buschek, Daniel and Pranjic, Jelena and Hussmann, Heinrich, “Understanding Emoji Interpretation through User Personality and Message Context,” Association for Computing Machinery, 2019.

[10] Dey, Sanorita and Duff, Brittany RL and Chhaya, Niyati and Fu, Wai and Swaminathan, Vishy and Karahalios, Karrie, “Recommendation for Video Advertisements Based on Personality Traits and Companion Content,” Proceedings of the 25th International Conference on Intelligent User Interfaces, pp.144–154, 2020.

[11] Toriumi, Fujio and Baba, Seigo, “Real-Time Tweet Classification in Disaster Situation,” Proceedings of the 25th International Conference Companion on World Wide Web, pp.117–118, 2016.

[12] Valle-Cruz, David and López-Chau, Asdrúbal and Sandoval-Almazán, Rodrigo, “Impression Analysis of Trending Topics in Twitter with Classification Algorithms,”

Association for Computing Machinery, no.11, pp.430–441, 2020.

[13] Wei, Jason and Zou, Kai and Inui, Kentaro and Jiang, Jing and Ng, Vincent and Wan, Xiaojun, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), no.11, pp.6382–6388, 2019.

[14] Likert, Rensis, “A technique for the measurement of attitudes.,” Archives of psychology, 1932.

[15] T. H. Nguyen and K. Rudra, “Towards an interpretable approach to classify and summarize crisis events from microblogs,” Proc. of the Web Conference 2022, pp.3641–3650, 2022.

[16] Md. Yasin Kabir and Sanjay Madria, “A Deep Learning Approach for Tweet Classification and Rescue Scheduling for Effective Disaster Management, ” In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems ,pp.269–278, 2019.

[17] S. Yamada, K. Utsu and O. Uchida, “An analysis of tweets posted during 2018 western japan heavy rain disaster,” 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), pp.1–8, 2019.

[18] S. Nishikawa, O. Uchida and K. Utsu, “Analysis of rescue request tweets in the 2018 japan floods,” Proc. of the 2019 International Conference on Information Technology and Computer Communications, pp.29–36, 2019.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 13pages, 2019.

[20] Ilya Loshchilov and Frank Hutter, “Decoupled Weight Decay Regularization,” International Conference on Learning Representations, 2019.