2024年2月28日(水)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

汎用機械学習技術 (1)

座長:天方 大地(大阪大学)
コメンテータ:新井 淳也(日本電信電話株式会社)

10:05 ~ 10:20
[T1-A-1-01] SVD Training を用いた継続学習の検討と評価
*松本 一則[1]、小西 達也[1]、鶏内 朋也[1]、黒川 茂莉[1]、橋本 慧志[1] (1. KDDI総合研究所)

10:20 ~ 10:45
[T1-A-1-02] 人工データを活用したグラフ深層学習におけるプライバシー保護と精度向上
*三谷 貴宏[1]、菊地 良将[1]、中西 宏和[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学)

10:45 ~ 11:10
[T1-A-1-03] Approach to Efficient Neural Architecture Search on the Latent Space of VGAE Models with Multiple Datasets
*逸見 一喜[1,2]、谷垣 勇輝[2]、大西 正輝[2,1] (1. 筑波大学、2. 産業技術総合研究所)

11:10 ~ 11:35
[T1-A-1-04] モンテカルロ木探索を用いたグラフニューラル構造探索
*佐々木 勇和[1] (1. 大阪大学)

11:35 ~ 12:00
[T1-A-1-05] 【技術報告】LayerXにおけるAI・機械学習技術の活用事例とその裏側
松村 優也 (株式会社LayerX)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

言語モデル

座長:喜田 拓也(北海学園大学)
コメンテータ:若林 啓(筑波大学)

10:05 ~ 10:20
[T1-B-1-01] Evaluating saliency methods by LLMs
*LU XIAOTIAN[1] (1. 京都大学大学院情報学研究科知能情報学コース鹿島研究室)

10:20 ~ 10:45
[T1-B-1-02] 確信度を用いたBERTによる参考文献書誌情報抽出の誤り検出の一手法
*中山 竣平[1]、金澤 輝一[2]、上野 史[3]、太田 学[3] (1. 岡山大学大学院環境生命自然科学研究科　太田研究室、2. 国立情報学研究所、3. 岡山大学学術研究院環境生命自然科学学域)

10:45 ~ 11:10
[T1-B-1-03] LLMにおけるAttention機構の役割の分析
*大塚 空来[1]、湯本 高行[1] (1. 兵庫県立大学)

11:10 ~ 11:35
[T1-B-1-04] k近傍法を用いた事前学習済みBERTモデル検索
*ファム フーロン[1]、三林 亮太[1]、山本 岳洋[1]、加藤 誠[2]、山本 祐輔[3]、荘司 慶行[4]、大島 裕明[1] (1. 兵庫県立大学、2. 筑波大学、3. 名古屋市立大学、4. 静岡大学)

11:35 ~ 12:00
[T1-B-1-05] 【技術報告】NECのLLM技術とデータ分析高速化技術
小山田昌史、董于洋、山崎智史、石坂一久 (日本電気株式会社)

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

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

問合せ処理・データベース管理

座長:金政 泰彦(富士通株式会社)
コメンテータ:三島 健(株式会社SRA)

10:05 ~ 10:20

[T2-A-1-01] RDBMSにおけるベクトル類似検索と関係演算を複合した問合せ処理の実験と考察

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

10:20 ~ 10:45

[T2-A-1-02] 誤差保証付き近似的問合せ処理におけるシノプシス構築の高速化

*倪 天嘉[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学)

10:45 ~ 11:10

[T2-A-1-03] 単一表クエリを対象とするシノプシス埋込みによる近似問合せ処理の評価

*髙田 実佳[1]、合田 和生[2] (1. 東京大学情報理工学研究科電子情報学専攻、2. 東京大学)

11:10 ~ 11:35

[T2-A-1-04] データマネジメントとデータベース管理

*増永 良文[1] (1. お茶の水女子大学名誉教授)

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

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

準同型暗号

座長:上土井 陽子(広島市立大学)
コメンテータ:天笠 俊之(筑波大学)

10:05 ~ 10:30

[T2-B-1-01] Intel SGX上で動作する準同型暗号の秘密鍵に対するサイドチャネル攻撃

*横平 正熙[1]、大西 隆太郎[2]、鈴木 拓也[2]、山名 早人[3] (1. 早稲田大学　基幹理工学部　情報理工学科　山名研究室、2. 早稲田大学大学院　基幹理工学研究科　山名研究室、3. 早稲田大学　理工学術院)

10:30 ~ 10:55

[T2-B-1-02] 効率的な暗号処理に向けた完全準同型暗号方式・ライブラリの評価

*辻 有紗[1]、小口 正人[1] (1. お茶の水女子大学)

10:55 ~ 11:20

[T2-B-1-03] 辞書を用いた圧縮センシングによる準同型暗号化データの通信量削減

*泉 湖雪[1]、松本 茉倫[1]、小口 正人[1] (1. お茶の水女子大学)

11:20 ~ 11:45

[T2-B-1-04] 準同型暗号と隔離実行環境の組合せによるプライバシを保護したBERT文章分類手法

*橋本 太志[1]、大西 隆太郎[2]、鈴木 拓也[2]、山名 早人[3] (1. 早稲田大学基幹理工学部、2. 早稲田大学大学院基幹理工学研究科、3. 早稲田大学理工学術院)

11:45 ~ 12:10

[T2-B-1-05] 【技術報告】不動産情報サービスの研究開発における共有データ資源

清田陽司 (株式会社LIFULL)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

SNS分析 (1)

座長:関 洋平(筑波大学)
コメンテータ:平手 勇宇(楽天株式会社)

10:05 ~ 10:30
[T3-A-1-01] 画像付きSNS投稿における画像ーテキスト間の相互関係分析

*鍋岡 琢渡[1]、馬 強[2] (1. 京都大学、2. 京都工芸繊維大学)

10:30 ~ 10:55
[T3-A-1-02] コロナ禍の話題表す俳句生む

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

10:55 ~ 11:20
[T3-A-1-03] 同一ハッシュタグ投稿に着目した対照学習によるトピック類似性を捉えた文ベクトルの獲得

*仲田 明良[1]、狩野 芳伸[1] (1. 静岡大学　情報学部　狩野芳伸研究室)

11:20 ~ 11:45
[T3-A-1-04] SNS実データ分析に基づくユーザのフォロー行動モデルの評価

*山川 衛[1]、田島 敬史[1] (1. 京都大学大学院情報学研究科)

11:45 ~ 12:10
[T3-A-1-05] イベント発生の不均衡性に適合した人口変動予測

*塚田 涼太郎[1]、豊田 正史[2]、梅本 和俊[2]、是津 耕司[3] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所、3. 情報通信研究機構)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

検索モデル

座長:義久 智樹(大阪大学)
コメンテータ:黒川 茂莉(株式会社KDDI総合研究所)

10:05 ~ 10:30
[T3-B-1-01] リスティング広告閲覧タイミングに基づくリランキングにおける適合フィードバック算出手法の提案

*濱口 滋久[1]、須山 瑠万[1]、栗 達[2]、中島 伸介[1]、河合 由起子[1] (1. 京都産業大学、2. 福岡大学)

10:30 ~ 10:55
[T3-B-1-02] 多様性リランキングにより栄養バランスを考慮したレシピ気分検索システムの評価

*三塚 麟太郎[1]、高久 雅生[2] (1. 筑波大学情報学群知識情報・図書館学類、2. 筑波大学図書館情報メディア系)

10:55 ~ 11:20
[T3-B-1-03] 思い出せない映画に特化した情報検索システムの作成

*吉越 玲士[1]、酒井 哲也[1] (1. 早稲田大学大学院 情報理工・情報通信専攻 酒井研究室)

11:20 ~ 11:45
[T3-B-1-04] ドメイン特徴に基づくランキング学習モデルのドメイン適応

*伊藤 拓誠[1]、丸田 敦貴[3]、加藤 誠[2] (1. 筑波大学 知識情報・図書館学類、2. 筑波大学 図書館情報メディア系、3. 筑波大学大学院 人間総合科学学術院)

11:45 ~ 12:10
[T3-B-1-05] List-wise手法における順位に基づいた多段階分割と学習データの上位遷移による能動学習

*エルゲン 瑛夏[1]、鈴木 優[1] (1. 岐阜大学)

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

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

画像の理解

座長:宮森 恒(京都産業大学)
コメンテータ:風間 一洋(和歌山大学)

10:05 ~ 10:20
[T4-A-1-01] アンサンブル学習を用いた植物標本の分類

*小柳 亮輔[1]、白井 匡人[1] (1. 島根大学)

10:20 ~ 10:35
[T4-A-1-02] 複雑な劣化画像を学習させた Vision Transformer による
単一画像超解像

*林 亮佑[1]、河野 浩之[1] (1. 南山大学大学院　理工学研究科　機械電子制御工学専攻　河野研究室)

10:35 ~ 10:50
[T4-A-1-03] CNNとTransformer を用いた画像誘導型深度補完システムの実装

*大極 康平[1]、河野 浩之[1] (1. 南山大学大学院　理工学研究科　機械電子制御工学専攻　河野研究室)

10:50 ~ 11:15
[T4-A-1-04] 画像復元のための空間方向Attention

*沖津 永享[1]、白井 匡人[1] (1. 島根大学)

11:15 ~ 11:40
[T4-A-1-05] レシピの手順説明画像認識における食材視覚特徴分析

*張 訳心[1]、山肩 洋子[2]、田島 敬史[1] (1. 京都大学、2. 東京大学)

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

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

Accessibility・インタフェース

座長:白石 優旗(筑波技術大学)
コメンテータ:鷹野 孝典(神奈川工科大学)

10:05 ~ 10:20
[T4-B-1-01] テーマパークにおけるUX向上を目的とした情報提供システムの実用性評価

*糸賀 千晴[1]、本多 賢[2]、服部 哲[2]、青柳 西蔵[2]、山口 浩[2] (1. 駒澤大学大学院グローバル・メディア研究科、2. 駒澤大学)

10:20 ~ 10:35
[T4-B-1-02] 視覚情報と知識を利用したマルチモーダル対話における曖昧な言語指示対象の同定

*石井 里奈[1]、宮森 恒[2] (1. 京都産業大学 宮森研究室、2. 京都産業大学)

10:35 ~ 11:00
[T4-B-1-03] 複数のレイアウトアルゴリズムの選択的適用による階層型グラフのレイアウト最適化

*片岡 春菜[1]、村上 綾菜[1]、伊藤 貴之[1] (1. お茶の水女子大学　伊藤研究室)

11:00 ~ 11:25
[T4-B-1-04] チーム対戦ゲーム中のボイスチャットにおける音声を用いた発言促進インタフェースの検討

*重堂 孝太[1]、服部 峻[1]、砂山 渡[1] (1. 滋賀県立大学)

11:25 ~ 11:50
[T4-B-1-05] Multi-layer warning recognition system　駅構内における聴覚障害者サポートのためのマルチレイヤワーニング認識システムの実現

*須藤 光喜[1]、中西 崇文[2]、岡田 龍太郎[2]、峰松 彩子[2] (1. 武蔵野大学データサイエンス学部データサイエンス学科 transmedia-tech-lab、2. 武蔵野大学データサイエンス学部データサイエンス学科)

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

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

教育・学習 (1)

座長:張 建偉(岩手大学)
コメンテータ:林 康弘(武蔵野大学)

10:05 ~ 10:20

[T5-A-1-01] Web閲覧および発言履歴を用いた大規模言語モデルによる学習アドバイスの生成手法

*嶋津 瑛[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

10:20 ~ 10:35

[T5-A-1-02] 大規模言語モデルを活用したノーコードツールによるAI人材育成での学習体験の向上

*丹野 良介[1]、坂本 悠記[1]、中村 耀[1]、杉本 海人[1]、切通 恵介[1] (1. NTTコミュニケーションズ株式会社)

10:35 ~ 11:00

[T5-A-1-03] アニメーションおよびソースコード表示を活用したソートアルゴリズム学習アプリケーションの開発

*渡辺 大智[1]、石井 幹大[2]、伊藤 一成[3] (1. 青山学院大学 社会情報学部 伊藤一成研究室、2. 青山学院大学 理工学部、3. 青山学院大学 社会情報学部)

11:00 ~ 11:25

[T5-A-1-04] ランダムサーファーモデルの学習アプリケーション「人型ピクサーフィング」の実装と評価

*高橋 伶奈[1]、伊藤 一成[1] (1. 青山学院大学社会情報学部社会情報学科伊藤一成研究室)

11:25 ~ 11:50

[T5-A-1-05] 講義資料作成支援のための科目用語体系の階層構造と教科書の説明順序を用いたスライド分析手法

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

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

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

観光・行動

座長:大野 成義(職業能力開発総合大学校)
コメンテータ:熊本 忠彦(千葉工業大学)

10:05 ~ 10:30

[T5-B-1-01] 旅行時の対応能力向上のための大規模言語モデルを用いたトラブル当てゲームの評価

*斉藤 さつき[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

10:30 ~ 10:55

[T5-B-1-02] ジオタグつき投稿を用いたコロナ禍の影響の地域差の分析

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

10:55 ~ 11:20

[T5-B-1-03] 移動軌跡とそれに対応づく地理空間データを活用した人の移動特性の理解

*須藤 龍一[1]、戸田 浩之[1] (1. 横浜市立大学)

11:20 ~ 11:45

[T5-B-1-04] 機械学習を用いた旅行活動ツイート判定精度の可視化視認性への影響

*藤大 友都[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学 知能情報工学分野研究室、2. 滋賀県立大学)

11:45 ~ 12:10

[T5-B-1-05] 近接関係ネットワークに影響を与える歩行者行動の可視化

*森越 彩楓[1]、丹羽 了[2,3]、鷹見 竣希[3]、大西 正輝[3]、伊藤 貴之[1] (1. お茶の水女子大学、2. 筑波大学、3. 産業技術総合研究所)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

汎用機械学習技術 (2)

座長:小山田 昌史(NEC)
コメンテータ:古賀 久志(電気通信大学)

13:35 ~ 14:00

[T1-A-2-01] ミスラベルデータの忘却による学習済みモデルの汎化性能の向上手法の提案

*杉浦 一瑳[1]、岡村 真吾[2,1]、山下 恭佑[1]、矢内 直人[1] (1. 大阪大学、2. 奈良工業高等専門学校)

14:00 ~ 14:25

[T1-A-2-02] 教師あり対照学習に基づく次数公平なグラフ深層学習

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

14:25 ~ 14:50

[T1-A-2-03] 機械学習のためのアクティブ・データプール拡張手法

*京谷 辰哉[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学 情報学群 情報メディア創成学類、2. 筑波大学 図書館情報メディア系)

14:50 ~ 15:15

[T1-A-2-04] AGATE : 属性付き時間発展グラフの汎用予測フレームワーク

*山崎 翔平[1]、佐々木 勇和[2]、Karras Panagiotis[3]、鬼塚 真[2] (1. 株式会社野村総合研究所、2. 大阪大学、3. Aarhus University)

15:15 ~ 15:40

[T1-A-2-05] 複数タスクを対象としたクラス知識のメタ学習にもとづくゼロショット学習

*南部 優太[1]、幸島 匡宏[1]、岩田 具治[2]、山本 隆二[1] (1. 日本電信電話株式会社 NTT人間情報研究所、2. 日本電信電話株式会社 NTTコミュニケーション科学基礎研究所)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

論理・意味分析

座長:鷹野 孝典(神奈川工科大学)
コメンテータ:上土井 陽子(広島市立大学)

13:35 ~ 14:00

[T1-B-2-01] 俳句･短歌を対象とした短文の意味分析

*杉田 祥太郎[1]、田中 克己[1] (1. 福知山公立大学情報学部)

14:00 ~ 14:25

[T1-B-2-02] 対照学習による慣用句を考慮した文埋め込みの獲得

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

14:25 ~ 14:50

[T1-B-2-03] 命題論理における言語の構成性に着目した言語モデルの汎化能力の調査

*井上 綾介[1]、宮森 恒[1] (1. 京都産業大学)

14:50 ~ 15:15

[T1-B-2-04] 金融数学問題解法におけるChatGPTによる言語処理と記号処理機能の統合

*白田 由香利[1]、桑名 杏奈[2]、Chakraborty Basabi[3,4] (1. 学習院大学経済学部、2. 和洋女子大学全学教育センター、3. マダナパレ工科大学、4. 岩手県立大学)

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

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

問合せ最適化・多次元データ処理

座長:鬼塚 真(大阪大学)
コメンテータ:神崎 映光(島根大学)

13:35 ~ 14:00

[T2-A-2-01] 動的空間データに対する高速学習型カーディナリティ推定

*Ji Yuchen[1]、天方 大地[1]、佐々木 勇和[1]、原 隆浩[1] (1. 大阪大学大学院情報科学研究科)

14:00 ~ 14:25

[T2-A-2-02] 和文に対する文字列述語の深層学習を用いた選択率推定

*大谷 蒼太[1]、板野 凌太[1]、土田 隼之[1] (1. 明石工業高等専門学校電気情報工学科)

14:25 ~ 14:50

[T2-A-2-03] グラフデータベースにおけるWorst-case Optimal Joinの最適化

*伊藤 寿浩[1]、塩川 浩昭[1] (1. 筑波大学)

14:50 ~ 15:15

[T2-A-2-04] 関係データベースにおける動的複数問合せ最適化の試作実装と初期評価

*木村 元紀[1]、合田 和生[1] (1. 東京大学)

15:15 ~ 15:40

[T2-A-2-05] 多次元ソートのための整数値変換手法の評価

*高尾 大樹[1]、黒川 能毅[1]、西川 記史[1]、茂木 和彦[1]、藤原 真二[1] (1. 株式会社 日立製作所)

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

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

データ活用におけるプライバシ

座長:小口 正人(お茶の水女子大学)
コメンテータ:早水 悠登(東京大学)

13:35 ~ 14:00

[T2-B-2-01] m-不変性に基づいた連続的データ公開における安全性問題と小規模データ改ざんによる解決法

*張 扶蘇[1]、上土井 陽子[1]、若林 真一[1] (1. 広島市立大学)

14:00 ~ 14:25

[T2-B-2-02] 準同型暗号を活用したプライバシ保護型ニュース推薦手法

*中野 瑛心[1]、鈴木 拓也[2]、矢田 宙生[2]、大西 隆太郎[2]、山名 早人[3] (1. 早稲田大学基幹理工学部、2. 早稲田大学基幹理工学研究科、3. 早稲田大学理工学術院)

14:25 ~ 14:50

[T2-B-2-03] 解析処理単位の認可および監査に基づく柔軟なデータ流通処理基盤の提案

*松本 直樹[1]、川西 智也[2]、大平 健司[3]、小塚 真啓[4] (1. 京都大学、2. セーラネットワークス、3. 大阪大学、4. 岡山大学)

14:50 ~ 15:15

[T2-B-2-04] 階層構造・多重解像度による差分プライベートな深層人流生成

*高木 駿[1]、加藤 郁之[1]、曹 洋[2]、吉川 正俊[3] (1. 京都大学、2. 北海道大学、3. 大阪成蹊大学)

15:15 ~ 15:40

[T2-B-2-05] 顔画像プライバシ保護技術の比較検証

*馬 世昂[1]、曹 洋[1]、中村 篤祥[1] (1. 北海道大学大学院情報科学院)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

SNS分析 (2)

座長:荘司 慶行(静岡大学)
コメンテータ:岩井原 瑞穂(早稲田大学)

13:35 ~ 13:50
[T3-A-2-01] XとMastodonへ投稿されるトピックの比較

林 駆[1]、*廣中 詩織[1]、首藤 一幸[1] (1. 京都大学)

13:50 ~ 14:15
[T3-A-2-02] 意外性のある意見がもたらすエコーチェンバー現象の分析

*中川 啓[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

14:15 ~ 14:40
[T3-A-2-03] 日米間におけるTwitterユーザのCOVID-19ワクチンに対する接種スタンス形成過程の比較

*張 翔[1]、久光 祥平[1]、金 洪善[2]、豊田 正史[2]、吉永 直樹[2] (1. 東京大学、2. 東京大学生産技術研究所)

15:05 ~ 15:30
[T3-A-2-05] 受験向けSNSの記事分析に基づく性別推定に関する研究

*寺口 敏生[1]、道谷 輝[1]、丸山 亜希子[1]、上田 真由美[1] (1. 流通科学大学)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

検索クエリ

座長:清水 敏之(九州大学)
コメンテータ:横田 治夫(城西大学)

13:35 ~ 14:00
[T3-B-2-01] 系列変換モデルを利用したフィールド検索のためのクエリ書き換え

*仲地 優登[1]、加藤 誠[2]、百合草 陽介[3] (1. 筑波大学 大学院人間総合科学学術院 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系、3. 株式会社ミスミグループ本社)

14:00 ~ 14:25
[T3-B-2-02] クエリ自動補完のための文書コレクションからのクエリログ生成

*染谷 瑛進[1]、加藤 誠[2] (1. 筑波大学 知識情報・図書館学類 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系)

14:25 ~ 14:50
[T3-B-2-03] 音声対話型検索における階層型検索意図を用いたクエリ推薦の戦略

*平位 遼[1]、飛岡 憲[1]、山本 岳洋[1] (1. 兵庫県立大学)

14:50 ~ 15:15
[T3-B-2-04] A Token-Free Approach to Entity-Based Keyword Search in Classical Chinese

*蒋 中慶[1]、呉 増青[2]、肖 川[1,3]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学、2. 京都大学、3. 名古屋大学)

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

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

音楽

座長:佃 洸摂(産業技術総合研究所)
コメンテータ:富井 尚志(横浜国立大学)

13:35 ~ 13:50
[T4-A-2-01] 料理レシピと音楽の融合によるレシピ閲覧支援の提案

*浦川 航貴[1]、宮本 拓実[1]、王 元元[1] (1. 山口大学)

13:50 ~ 14:05
[T4-A-2-02] 最適輸送理論を用いたメロディー楽譜からの作曲家推定

*八木 皓平[1]、中村 篤祥[1] (1. 北海道大学　工学部　情報エレクトロニクス学科　情報理工学コース　アルゴリズム研究室、2. 北海道大学 情報科学研究院 情報理工学部門 知識ソフトウェア科学分野)

14:05 ~ 14:20
[T4-A-2-03] 教師なし学習による主成分分析を用いた流行楽曲の特徴量の相関

*小俵 亜紀[1]、木津 美咲[1]、内田 瑞貴[1]、原 香奈江[1]、齋藤 優莉[1]、金井 友希美[1] (1. 奈良女子大学)

14:20 ~ 14:45
[T4-A-2-04] 人体キーポイントを用いて映像に整合する楽器音を生成するモデルの提案

*岡野 日翔[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

14:45 ~ 15:10
[T4-A-2-05] 楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式

*佐伯 綾亮[1]、岡田 龍太郎[1]、峰松 彩子[1]、中西 崇文[1] (1. 武蔵野大学データサイエンス学部)

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

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

XR

座長:伊藤 貴之(お茶の水女子大学)
コメンテータ:大塚 真吾(神奈川工科大学)

13:35 ~ 14:00
[T4-B-2-01] 視覚障がい者のためのソーシャルVR空間の開発

*横山 皓祐[1]、Siriaraya Panote[2]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学)

14:00 ~ 14:25
[T4-B-2-02] 立体音響ARを用いた リラクゼーション支援手法の有効性評価

*大塚 智貴[1]、Siriaraya Panote[3]、中島 伸介[2] (1. 京都産業大学大学院、2. 京都産業大学、3. 京都工芸繊維大学)

14:25 ~ 14:50
[T4-B-2-03] MR空間におけるホワイトボードに対するインタラクションを伴う商品情報の探索

*津田 裕哉[1]、山本 岳洋[1]、大島 裕明[1] (1. 兵庫県立大学)

14:50 ~ 15:15
[T4-B-2-04] 物理的距離によるLoD制御をともなうVR環境における不動産情報探索

*中山 裕紀[1]、津田 裕哉[1]、荘司 慶行[2]、大島 裕明[1] (1. 兵庫県立大学、2. 静岡大学)

15:15 ~ 15:40
[T4-B-2-05] 空港手荷物受取所のUX向上を目指した拡張現実型情報提示サービス

*池田 呼人[1]、松原 克弥[1] (1. 公立はこだて未来大学)

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

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

教育・学習 (2)

座長:熊本 忠彦(千葉工業大学)
コメンテータ:飯沢 篤志(リコーITソリューションズ株式会社)

13:35 ~ 13:50

[T5-A-2-01] 3段階の学習過程を前提とした自己学習環境における講義動画閲覧履歴による学習状況分析

*齋藤 光貴[1]、渡辺 知恵美[2] (1. 筑波技術大学 産業技術学部 産業情報学科 情報科学コース 渡辺研究室、2. 筑波技術大学 産業技術学部 産業情報学科)

13:50 ~ 14:05

[T5-A-2-02] 教材ビデオにおける学習者の視聴操作に基づくシーン選択支援

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

14:05 ~ 14:30

[T5-A-2-03] ピクトグラミーティング - ピクトグラムコンテンツの共有及び相互評価システム -

*松村 貫司[1]、石井 幹大[2]、伊藤 一成[1] (1. 青山学院大学社会情報学部、2. 青山学院大学理工学部)

14:30 ~ 14:55

[T5-A-2-04] 学生向け学務情報のシステム透過なオープンデータ化

*中川 匠海[1]、松原 克弥[1] (1. 公立はこだて未来大学)

14:55 ~ 15:20

[T5-A-2-05] ユニバーサルピクトグラミング　ー人型ピクトグラムを用いたプログラミング学習環境の多ヶ国語化ー

*石井 幹大[1]、伊藤 一成[2] (1. 青山学院大学 理工学部、2. 青山学院大学 社会情報学部)

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

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

地理情報推薦

座長:廣田 雅春(岡山理科大学)
コメンテータ:小林 亜樹(工学院大学)

13:35 ~ 13:50

[T5-B-2-01] ジオタグSNSに基づくプロンプトおよび賑やかさストリートビュー生成手法の検討

*松石 侑樹[1]、Siriaraya Panote [2]、河合 由起子[1] (1. 京都産業大学、2. 京都工芸繊維大学)

13:50 ~ 14:05

[T5-B-2-02] 旅行体験との関係抽出に基づく土産話を引き出す土産物推薦システム

*山田 樹生[1]、斉藤 さつき[1]、北山 大輔[1] (1. 工学院大学)

14:05 ~ 14:30

[T5-B-2-03] Food Approach Bias軽減のための移動中の店舗推薦手法の提案

*ヨウ シンニー[1]、國松 音和[1]、王 元元[2]、Panote Siriaray[3]、河合 由起子[1,4] (1. 京都産業大学　、2. 山口大学、3. 京都工芸繊維大学、4. 大阪大学サイバーメディアセンター)

14:30 ~ 14:55

[T5-B-2-04] 観光スポットに関する口コミの感情に基づく観点の抽出

*飯濱 航平[1]、戸田 浩之[1] (1. 横浜市立大学)

14:55 ~ 15:20

[T5-B-2-05] 博物館の展示物を用いたビンゴゲームによる鑑賞軸の習得

*西本 海生[1]、相原 健郎[2,3]、神門 典子[2,4]、荘司 慶行[5]、山本 祐輔[6]、山本 岳洋[1,2]、大島 裕明[1,2] (1. 兵庫県立大学、2. 国立情報学研究所、3. 東京都立大学、4. 総合研究大学院大学、5. 静岡大学、6. 名古屋市立大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

汎用機械学習技術 (3)

座長:塩川 浩昭(筑波大学)
コメンテータ:北川 博之(筑波大学)

16:05 ~ 16:20
[T1-A-3-01] マルチエージェント強化学習におけるMA-POCAモデルのFFNの改善

*藤江 拓郎[1]、白井 匡人[1] (1. 島根大学)

16:20 ~ 16:45
[T1-A-3-02] マルチエージェントゲームを通じた取引ネットワークにおけるより良いルールの自動生成手法

*沢木 翔太[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学情報学群知識情報・図書館学類、2. 筑波大学図書館情報メディア系)

16:45 ~ 17:10
[T1-A-3-03] GANを用いた異常検知における異常度算出方法についての考察

*森 仁美[1]、中野 美由紀[2]、丸 千尋[3]、小口 正人[1] (1. お茶の水女子大学理学部情報科学科小口研究室、2. 津田塾大学学芸学部情報科学科、3. お茶の水女子大学文理融合AI・データサイエンスセンター)

17:10 ~ 17:35
[T1-A-3-04] Empowering Zero-Shot Extreme Multi-Label Text Classification via Weighted Contrastive Learning and Semantic Label Augmentation

*Zhao Yanan[1]、Iwaihara Mizuho [1] (1. 早稲田大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

言語表現認識

座長:片山 薫(東京都立大学)
コメンテータ:金子 邦彦(福山大学)

16:05 ~ 16:20
[T1-B-3-01] 日本語における大規模言語モデルの生成文検出

*丸井 渚生[1]、曹 洋[1]、中村 篤祥[1] (1. 北海道大学)

16:20 ~ 16:45
[T1-B-3-02] ポジティブな単語を含んだ煽り・誹謗中傷目的のコメント検出方法の提案

*佐藤 豪[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

16:45 ~ 17:10
[T1-B-3-03] 動画に対する否定的コメント・フィルタリングにおけるfew-shot学習の評価

*三田寺 聖[1]、宇津呂 武仁[1] (1. 筑波大学)

17:10 ~ 17:35
[T1-B-3-04] Conceptual Commonsense-aware Humor Recognition with Various Pretrained Masked Language Models

*佐々木 裕多[1]、張 建偉[2]、白石 優旗[3] (1. 東京工業大学、2. 岩手大学、3. 筑波技術大学)

17:35 ~ 18:00
[T1-B-3-05] 皮肉検出モデルにおける適切なコモンセンスの選別と性能評価

*皆川 翔哉[1]、佐々木 裕多[2]、張 建偉[1] (1. 岩手大学、2. 東京工業大学)

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

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

並列処理・分散データベース

座長:合田 和生(東京大学)
コメンテータ:金政 泰彦(富士通株式会社)

16:05 ~ 16:30

[T2-A-3-01] データ検証可能な分散DB実現手法の提案と実現に向けたHyperledger Irohaのコマンド開発

*堀 遥[1]、小口 正人[1] (1. お茶の水女子大学)

16:30 ~ 16:55

[T2-A-3-02] SparkSQLによる効率的な問合せ処理のためのワークロードに基づくRDFデータ分割手法

*山﨑 昂輔[1]、天笠 俊之[2] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群情報理工学位プログラム 知識・データ工学研究室、2. 筑波大学計算科学研究センター)

16:55 ~ 17:20

[T2-A-3-03] Fast Parallel Computation for Random Walk based t-SNE

*三木 武志[1]、藤原 靖宏[2]、川島 英之[1] (1. 慶應義塾大学、2. 日本電信電話株式会社)

17:20 ~ 17:45

[T2-A-3-04] ブロックチェーン技術を活用したデータ検証可能な分散データベースの性能に関する検討

*坂本 明穂[1]、小口 正人[2] (1. お茶の水女子大学理学部情報科学科小口研究室、2. お茶の水女子大学基幹研究院自然科学系)

17:45 ~ 18:10

[T2-A-3-05] PostgreSQLの外部連携機能におけるレコード変換処理の高速化

*柿村 直拓[1]、立床 雅司[1]、柴田 秀哉[1] (1. 三菱電機株式会社 情報技術総合研究所 データマネジメント技術部 データ最適化技術グループ)

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

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

機械学習とプライバシ

座長:高橋 翼(LINEヤフー株式会社)
コメンテータ:佐々木 勇和(大阪大学)

16:05 ~ 16:30

[T2-B-3-01] ソースフリードメイン適応設定における重み付きクラスター連合学習

*森 隼基[1,2]、木原 孝輔[1]、寺西 勇[1] (1. NEC セキュアシステムプラットフォーム研究所、2. 京都大学)

16:30 ~ 16:55

[T2-B-3-02] 個人化されたユーザレベルの差分プライバシ保護型連合学習

*加藤 郁之[1]、高木 駿[1]、Xiong Li[2]、曹 洋[3]、吉川 正俊[4] (1. 京都大学、2. Emory University、3. 北海道大学、4. 大阪成蹊大学)

16:55 ~ 17:20

[T2-B-3-03] 非独立同分布の学習データに対する分割学習の性能向上に関する研究

*池奥 裕太[1]、曹 洋[2]、吉川 正俊[3] (1. 京都大学大学院情報学研究科、2. 北海道大学大学院情報科学研究院、3. 大阪成蹊大学データサイエンス学部)

17:20 ~ 17:45

[T2-B-3-04] エンコーダを用いた不均一モデル構造間の連合学習

*松永 悠斗[1]、籾山 悟至[1]、ENKHTAIVAN BATNYAM[1]、寺西 勇[1] (1. 日本電気株式会社)

17:45 ~ 18:10

[T2-B-3-05] 【技術報告】安全なLLMの実現を目指して - ストレステストとセーフアラインメントの取り組み

綿岡晃輝 (LINEヤフー株式会社)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

SNS分析 (3)

座長:平手 勇宇(楽天株式会社)
コメンテータ:大野 成義(職業能力開発総合大学校)

16:05 ~ 16:30

[T3-A-3-01] 都市別のソーシャルメディア投稿を利用した市民の性格特性分析

*岩﨑 大晟[1]、関 洋平[2]、柏野 和佳子[3]、欅 惇志[4]、神門 典子[5,6] (1. 筑波大学 情報学群 知識情報・図書館学類、2. 筑波大学 図書館情報メディア系、3. 国立国語研究所、4. 一橋大学 ソーシャル・データサイエンス研究科、5. 国立情報学研究所、6. 総合研究大学院大学)

16:30 ~ 16:55

[T3-A-3-02] コロナ禍における草の観察

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

16:55 ~ 17:20

[T3-A-3-03] 降雨データを用いた SNS ユーザーの居住地推定

*山下 暁天[1] (1. 京都大学工学部情報学科)

17:20 ~ 17:45

[T3-A-3-04] ソーシャルメディアテキストからのユーザ行動の分類

*中岡 明義[1]、若宮 翔子[1]、荒巻 英治[1] (1. 奈良先端科学技術大学院大学)

17:45 ~ 18:10

[T3-A-3-05] 機械学習とクラウドソーシングを用いた顧客の声の自動分類

*鈴木 優[1]、城所 祥太[1]、北村 拓斗[1]、太田 奈那[1]、三島 惇也[1]、本田 文隆[2]、長徳 弘一[2]、中村 勇太[2]、佐々木 和彦[2]、笠丸 陽一[2] (1. 岐阜大学、2. ソニー（株）)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

データ検索・解析

座長:飯沢 篤志(リコーITソリューションズ株式会社)
コメンテータ:國島 丈生(追手門学院大学)

16:05 ~ 16:30

[T3-B-3-01] 鏡映を識別可能な3次元CADモデル検索のためのバイスペクトルの利用

*小島 優希也[1]、片山 薫[1] (1. 東京都立大学)

16:30 ~ 16:55

[T3-B-3-02] 流通を考慮した研究データのメタデータ付与支援

*篠原 桐[1]、清水 敏之[2]、冨浦 洋一[3] (1. 九州大学大学院統合新領域学府冨浦研究室、2. 九州大学附属図書館研究開発室、3. 九州大学大学院システム情報科学研究所)

16:55 ~ 17:20

[T3-B-3-03] スプレッドシートを用いたデータ解析における情報要求分析

*丸田 敦貴[1]、田貝 奈央[1]、加藤 誠[1] (1. 筑波大学)

17:20 ~ 17:45

[T3-B-3-04] 研究内容からの研究データ検索

*山﨑 基数[1]、冨浦 洋一[1]、清水 敏之[1] (1. 九州大学)

17:45 ~ 18:10

[T3-B-3-05] 大規模言語モデルによる文書補強とリランキングによる統計データ検索

*黒川 博生[1]、宮森 恒[1] (1. 京都産業大学 情報理工学部)

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

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

生成系マルチメディア

座長:井手 一郎(名古屋大学)
コメンテータ:野宮 浩揮(京都工芸繊維大学)

16:05 ~ 16:30

[T4-A-3-01] FPSゲームにおけるチート対策のためのMistによる物体検出阻害効果の検証実験

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

16:30 ~ 16:55

[T4-A-3-02] 表構造情報を利用した棒グラフの自動生成の一手法

*田上 歩夢[1]、金澤 輝一[2]、上野 史[3]、太田 学[3] (1. 岡山大学工学部情報系学科、2. 国立情報学研究所、3. 岡山大学学術研究自然科学学域)

17:20 ~ 17:45

[T4-A-3-04] 条件付きトランスフォーマーとテキストリフレーサーに基づくクワッドローター用視覚言語ナビゲーション

*Chen Zhe[2,1]、Li Jiyi[1]、Fukumoto Fumiyo[1]、Liu Peng[2]、Suzuki Yoshimi[1] (1. 山梨大学、2. 杭州電子科技大学)

17:45 ~ 18:10

[T4-A-3-05] 【技術報告】ヒトとAIの協調によるデザイン創出支援

山口喬弘 (株式会社豊田中央研究所)

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

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

可視化によるユーザ支援

座長:北山 大輔(工学院大学)
コメンテータ:荘司 慶行(静岡大学)

16:05 ~ 16:20

[T4-B-3-01] レビュー動画内の商品特徴量分析システムにおける情報提示方法の検討

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

16:20 ~ 16:35

[T4-B-3-02] 日中同形異義語の意味分析と可視化：言語横断的意味研究の一環として

*韓 禾[1]、伊藤 貴之[1] (1. お茶の水女子大学　人間文化創成研究科　情報コース　伊藤研究室)

16:35 ~ 17:00

[T4-B-3-03] 散布図選択による多次元データ可視化における配置の最適化

*松尾 深景[1]、伊藤 貴之[1] (1. お茶の水女子大学)

17:00 ~ 17:25

[T4-B-3-04] 粒度の異なる商品比較を実現する評判情報可視化インタフェースの提案

*板垣 紫音[1]、上野 史[1]、太田 学[1] (1. 岡山大学大学院　環境生命自然科学研究科　太田研究室)

17:25 ~ 17:50

[T4-B-3-05] Webブラウザのタブ群のタスクごとへのグループ化手法

*久冨 瑞稀[1]、田島 敬史[1] (1. 京都大学情報学研究科)

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

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

教育・学習 (3)

座長:木村 昌臣(芝浦工業大学)
コメンテータ:橋本 敦史(オムロンサイニックエックス)

16:05 ~ 16:20

[T5-A-3-01] 生成AIを用いた教え合い学習のための間違いを模倣するAgent

*森部 想水[1]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府、2. 九州大学大学院芸術工学研究院)

16:20 ~ 16:45

[T5-A-3-02] ピクタグラミング −ピクトグラムの作成を通じたマークアップ言語学習アプリケーションの開発とその評価−

*木下 倭[1]、高橋 伶奈[1]、伊藤 一成[1] (1. 青山学院大学 社会情報学部 伊藤一成研究室)

16:45 ~ 17:10

[T5-A-3-03] ピクトグラムを用いた教育向けネットワークシミュレータ「Pictocol Simulator」の開発

*前田 祐杜[1]、伊藤 一成[2] (1. 青山学院大学社会情報学部　伊藤一成研究室、2. 青山学院大学社会情報学部)

17:10 ~ 17:35

[T5-A-3-04] 図形への書き込みと手書き解答文字の自動認識を用いた数学学習者の解答戦略推定

*林 凱博[1]、宮﨑 公彦[2]、工藤 雅士[2]、山名 早人[3] (1. 早稲田大学 基幹理工学部、2. 早稲田大学 基幹理工学研究科、3. 早稲田大学 理工学術院)

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

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

地理情報分析・シミュレーション

座長:大塚 真吾(神奈川工科大学)
コメンテータ:義久 智樹(大阪大学)

16:05 ~ 16:30

[T5-B-3-01] 平均二乗誤差を利用した明るさが変動する天体の周期解析手法

*小野 晴樹[1]、中西 宏和[1]、藤原 靖宏[2]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科マルチメディア工学専攻鬼塚研究室、2. NTTコミュニケーション科学基礎研究所)

16:30 ~ 16:55

[T5-B-3-02] 風況を考慮したEVのエネルギーベースラインマップの作成と
長期間実走行データによる精度検証

*劉 屹[1]、曽我 紗代子[1]、何 新[1]、田中 悠斗[1]、富井 尚志[2] (1. 横浜国立大学　大学院環境情報学府　情報環境学専攻、2. 横浜国立大学　大学院環境情報研究院)

16:55 ~ 17:20

[T5-B-3-03] PVとEVを統合する負荷平準指向VGIDBを用いたスマートグリッド導入効果の多角的定量化

*廣居 樹[1]、片平 昇輝[1]、鈴木 博人[1]、小松 虎太郎[2]、本藤 祐樹[3]、富井 尚志[3] (1. 横浜国立大学 大学院環境情報学府 情報環境専攻、2. 横浜国立大学 理工学部 数物・電子情報系学科、3. 横浜国立大学 大学院環境情報研究院)

17:20 ~ 17:45

[T5-B-3-04] 水害時における市民の避難行動に関する意識の分析

*石井 樹[1]、妹尾 考[2]、関 洋平[3]、川島 宏一[4]、鈴木 誉幸[5] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群社会工学学位プログラム博士後期課程、2. 筑波大学大学院人間総合科学学術院人間総合科学研究群情報学学位プログラム博士前期課程、3. 筑波大学図書館情報メディア系、4. 筑波大学システム情報系、5. 筑波大学大学院理工情報生命学術院システム情報工学研究群社会工学学位プログラム博士前期課程)

17:45 ~ 18:10

[T5-B-3-05] 人口動態を基盤としたソーシャルセンサのパフォーマンス分析

*大西 駿太朗[1]、矢田 竣太郎[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室)

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 10:00 ~ 12:10 | T1-A オンライン（Zoom Events)

汎用機械学習技術 (1)

座長:天方 大地(大阪大学)
コメンテータ:新井 淳也(日本電信電話株式会社)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T1-A-1-01] SVD Training を用いた継続学習の検討と評価
*松本 一則[1]、小西 達也[1]、鶏内 朋也[1]、黒川 茂莉[1]、橋本 慧志[1] (1. KDDI総合研究所)

10:20 ~ 10:45
[T1-A-1-02] 人工データを活用したグラフ深層学習におけるプライバシー保護と精度向上
*三谷 貴宏[1]、菊地 良将[1]、中西 宏和[1]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学)

10:45 ~ 11:10
[T1-A-1-03] Approach to Efficient Neural Architecture Search on the Latent Space of VGAE Models with Multiple Datasets
*逸見 一喜[1,2]、谷垣 勇輝[2]、大西 正輝[2,1] (1. 筑波大学、2. 産業技術総合研究所)

11:10 ~ 11:35
[T1-A-1-04] モンテカルロ木探索を用いたグラフニューラル構造探索
*佐々木 勇和[1] (1. 大阪大学)

11:35 ~ 12:00
[T1-A-1-05] 【技術報告】LayerXにおけるAI・機械学習技術の活用事例とその裏側
松村 優也 (株式会社LayerX)

SVD Training を用いた継続学習の検討と評価

松本　一則†　　小西　達也†　　鶏内　朋也†　　黒川　茂莉†　　橋本　慧志†

† KDDI 総合研究所　〒 356–8502 埼玉県ふじみ野市大原 2-1-15

E-mail: †{mz-matsumoto,mo-kurokawa,st-hashimoto}@kddi.com, ††{tt-konishi,to-kaichi}@kddi-research.jp

あらまし　継続学習の枠組みに，学習済みタスクの重要パラメータが変更されないよう，新タスクでは既存タスクの部分空間に直交するようにパラメータ更新を行う手法が提案されている．一方，コンパクトな学習モデルを構築することを目的とし，全結合層の重みパラメータに代わり，同パラメータを特異値分解の逆過程で合成する SVD-training が提唱されている．本稿では，SVD-Training の仕組みを継続学習に組み入れることで継続学習で問題となる忘却を抑止する手法を提案する．評価の結果，提案手法が忘却を起こさずに継続学習を行えること，継続学習が進んだ状態で提案手法が既存手法の精度を上回ることが示された．本手法は忘却を抑止するために多くの手法で採用されているリプレイ用のデータや部分モデルの集合を記憶する必要がなく，メモリ省力化の点でも優位である．

キーワード　DEIM，フォーマット，論文執筆の注意事項

1 はじめに

近年注目を集める AI 研究分野の一つに継続学習 (Continual Learning) という分野がある．継続学習では, タスクと呼ばれる新たなデータセットをモデルが順次学習していく状況を前提とする．新タスクの学習が完了した時点で以前のタスクに対する性能劣化を忘却 (Forgetting) と呼び，この忘却を抑止することが継続学習での大きな課題となっている．忘却の抑止には種々の手法が提案され，文献 [1] によれば下記に分類される．

Memory-based Method:　学習済みタスクの学習例をメモリバッファに保存し，新タスクの学習で再利用 (=リプレー) する．

Architecture-based Method:　旧タスクで得た知識を保持するためモデルの構造 (=アーキテクチャ）を更新する．

Regularization-based Method:　忘却を抑止するための損失関数を工夫する．

Subspace-based Method:　学習済みタスクの重要パラメータが変更されないよう，新タスクでは既存タスクの部分空間に直交するようにパラメータ更新を行う．

Bayesian Method:　不確実性の評価や正規化を使って忘却を減らす戦略をベイズ推定で決定する．

一方，別分野の研究では，コンパクトな学習モデルを構築することを目的とし，全結合層の重みパラメータを更新するのではなく，特異値分解の逆過程で低次元行列から重みパラメータを合成する SVD-training が提唱されている [2].

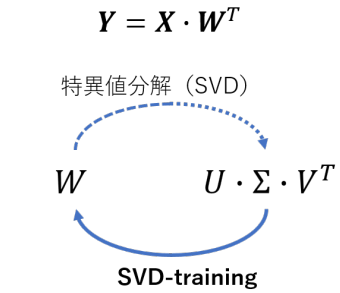
本稿では，SVD-Training の仕組みを継続学習に組み入れることで継続学習で問題となる忘却を抑止する手法を提案する (2 章)．そして，提案手法の有用性とその再現性を主張するため，評価実験については詳しく報告 (3 章) し，最後に結論を述べる (4 章).

2 提案手法

提案の手法では，文献 [2] の手法同様に学習過程では全結合層 $\mathbf{W}$ に対して SVD-trainig を行う (図 1)．つまり，同図中の $\mathbf{W}$ ではなく，$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V^T}$ を学習時に更新していく．ただし，本提案では単に学習時に SVD-training を行うだけではなく，各タスク終了時に重要な (=大きい) 特異値 $\sigma_i, \cdots$ を決定し，固定対象とする．固定対象になった特異値は以降のタスクの学習で更新されないように，そのインデックスを記憶していく．タスクが増えるごとに固定対象となった特異値の数は単調増加するが，後述の実験 (3 章) でも $\mathbf{W}$ のランクにまで至ることはなかった．よって現実的にはこの戦略は機能している．

学習が終わり，過去タスクを含めての評価が行われる場合，未知データのクラスが所属する $task-N$ をまず推定する．ただし Task-Incremental と呼ばれる継続学習の著名シナリオの場合，推定時における未知データに対する所属タスクは既知であるとの前提であるため，このタスク推定については不要となる．そして，$task-N$ において固定対象でなかった特異値の値を 0 に設定した $\mathbf{\Sigma}^*$ を使って，$\mathbf{W} = \mathbf{U} \cdot \mathbf{\Sigma}^* \cdot \mathbf{V}^T$ を合成し，出力 $\mathbf{Y}$ を得る．この時，$\mathbf{Y}$ は以前 $task-N$ の学習終了時に $\mathbf{X}$ から得られたものと一致することが保証される．

上記の戦略で特異値を固定していけば，忘却は生じないことになる．この提案手法では，新タスクでは既存タスクの部分空間に直交するようにパラメータ更新を行うので，継続学習向け手法の中では部分空間法 (Subspace-based Method) に属する．ただし，既存の部分空間法は $\mathbf{W}$ の更新履歴に基づいて更新されるパラメータの傾きを制限したり，入力データ $\mathbf{X}$ を直行する空間に分割している．本提案のように特異値分解の逆過程を用いて効率よく部分空間法の目的を達しているものはない．また，継続学習向け手法のうち，Memory-based Method のようにリプレイ用のデータを各タスクごとに保管したり，新タスクにお



ただし、
・U は正規直交行列 ⇔ $\| U^T \cdot U - I \| = 0$
・V は正規直交行列 ⇔ $\| V^T \cdot V - I \| = 0$

図 1　通常の特異値分解と SVD-training の関係

いて保管したリプレイ用データを使って学習を行う必要もない．一般的にリプレイ用データを保管するためのメモリに比べ，提案手法で保管する特異値のインデックス情報のメモリは少なくて済むので，省メモリの点で本手法は優位である．同様に，既存知識を部分モデルの集合で保持する Architecture-based Method の方法に比べても，特異値のインデックス情報の方が省メモリである．

3　実験と評価

実験では継続学習の性能比較用プラットフォームである FACIL を使用し，Task-Incremental のシナリオで継続学習向け各種手法による画像分類精度を比較した．

3.1　測 定 条 件

使用するデータセットやモデルの構造の違いによる結果への影響を少なくするため，2 種類のデータセットと 2 種類のアーキテクチャを用いて実験を行った．使用した画像データセットとモデルの構造は下記の通り．

画像データセット

CIFAR100:　100 種類のクラスを持つ 32x32 画素のカラー画像の集合．FACIL のデータ管理機能を使って，各クラス 450 件，50 件，100 件の訓練用，検証用，テスト用に分けて使用した．

Tiny ImageNet:　200 種類のクラスをもつカラー画像の集合．本来は 64x64 画素の画像であるが，32x32 に縮小して使用した．訓練用，検証用，テスト用のデータ件数は各クラスそれぞれ 500 件，50 件，50 件である．

モデルの構造

AlexNet-like:　Feature Extractor は AlexNet と同じ構成であり，5 個の畳み込み層と 3 個のプーリング層を持つ．Classifier は全結合層もしくは SOW 層を 2 個持つ．そして他の継続学習研究同様，各タスク専用のクラス識別器（全結合層）が Classifier に繋がっている（図 2）．

ResNet-50-like:　Feature Extractor は畳み込み層と ReLU で構成される residual block を繰り返したものであり，Resnet-50 と同じ構成である．clasifier は全結合層もしくは SOW 層が 1 個を含み，Head に繋がっている（図 2）．

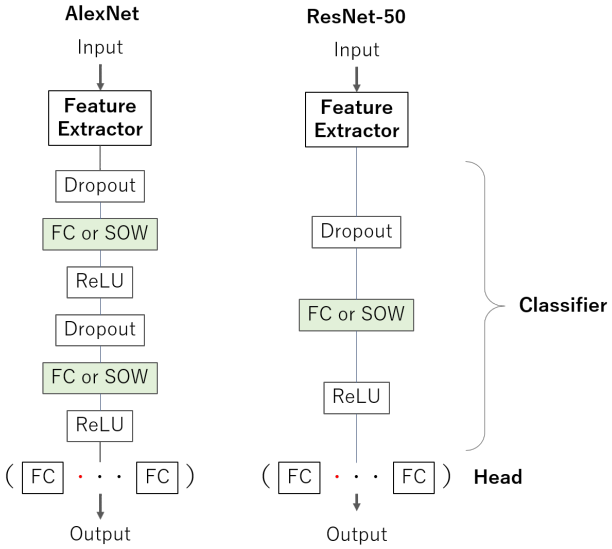


図 2　実験に使用した 2 種類のモデルの構造

なお，本実験は全結合層を提案の SOW 層に置き換えた時の精度を比較することを目的とするため，Feature Extractor は事前に構築した．つまり実験内では学習中でも FeatureExtractor のパラメータは固定されたままで更新されず，Classifier と Head のパラメータのみが更新される．固定された Feature Extractor を使うことで，Classifier への入力は従来手法と提案手法で一致するため，全結合層と SOW 層の比較が明確になる．事前構築の Feature Extractor のパラメータ学習と実験のパラメータ学習で同じデータセットが使われることがないように，実験で CIFAR100 を使う場合は Tiny ImageNet でパラメータ学習を行った Feature Extractor を使用した．また，Tiny ImageNet を実験で使用する場合は，CIFAR100 で構築した FeatureExtractor を使用した

さらに一般的な結果を得るため，Task-Incremental で想定するタスクの数は 2 種類にした．ここまでの説明で示した実験設定をまとめて図 3 に示す．

この各設定下で，提案手法と継続学習手法として著名な EWC，Path Integration，ならびに継続学習の性能の上限と下限を示す fine-tuning と joint の精度を測定した．モデルが十分収束するように各手法で 400 epoch の訓練を行い，検証用データを用いて最善のモデルを決定した．パラメータ更新に使用したアルゴリズムはモメンタムを考慮した SGD である．dropout 層に使用するハイパーパラメータはベイズ最適化ツールである optuna を用いて決定した．そして，学習時における画像データの順序による影響を減らすため，各設定下で各種手法を乱数系列を変えて 3 回測定した．

以上により，提案手法と従来手法を各種条件下で網羅的に比較した．

図 3　実験の種別（手法を区別していない）

Classifier の学習・評価に使用したデータセット	FeatureExtractor の作成に使用したデータセット	NW構造	タスク数	実験名（略称）
CIFAR100	Tiny ImageNet	AlexNet	10	CA-10
			20	CA-20
		Resnet50	10	CR-10
			20	CR-20
Tiny ImageNet	CIFAR100	AlexNet	10	TA-10
			20	TA-20
		Resnet50	10	TR-10
			20	TR-20

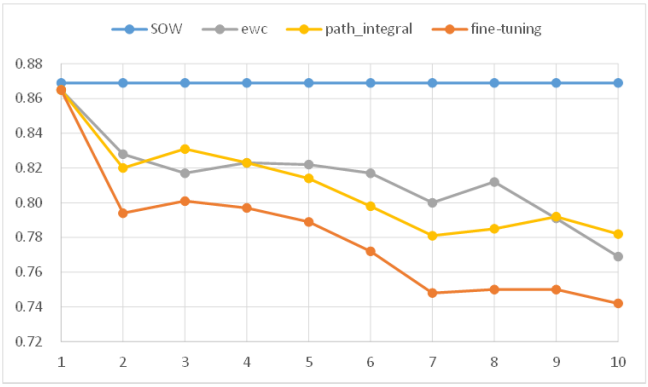


図 4　Task1 用テストデータに対する各種手法の忘却の様子

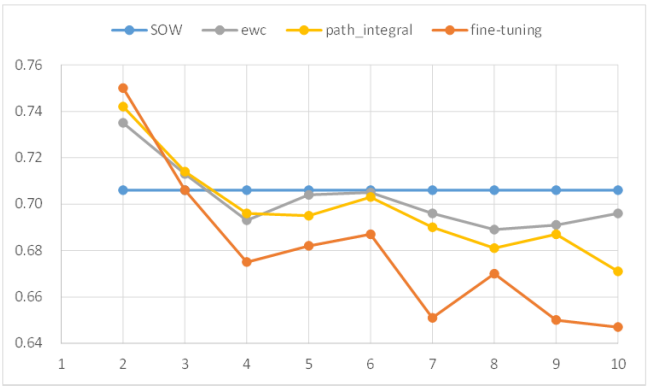


図 5　Task2 用テストデータに対する各種手法の忘却の様子

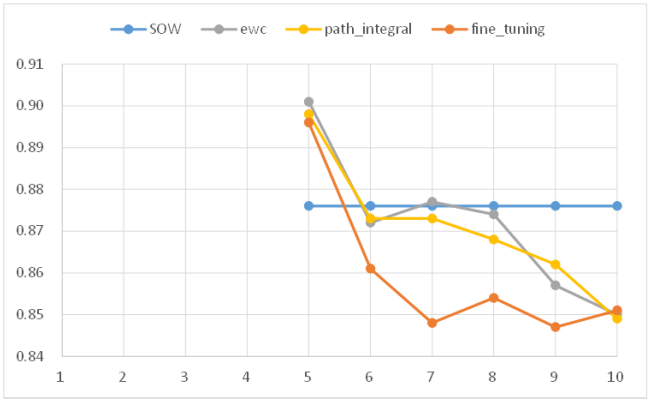


図 6　Task5 用テストデータに対する各種手法の忘却の様子

3.2　実験結果とその評価

ここでは，紙面の都合から，CA-10 の状況下における各種手法の結果に絞って報告を行う．ただし他の実験設定でも同様の結果を得ている．

3.2.1　忘却の状況

図 4,5,6 は Incremental-Task のシナリオに従って各タスクの学習を順次行っていた際の特定タスクのテストデータに対する各種手法の精度を表している．図 4 では忘却を抑止する目的で提案された ewc や path integral であっても，タスク数が増えるにつれ，タスク 1 の評価データに対する忘却が生じている（=精度低下の傾向がにある）ことがわかる．

一方，提案の SOW を使った手法では，いずれのテストデータにおいても想定通り忘却が起こっていない（=精度低下が生じていない）．このため，図 5,6 が示すよう，タスク数が増えると SOW の精度が他手法の精度を上回っている．

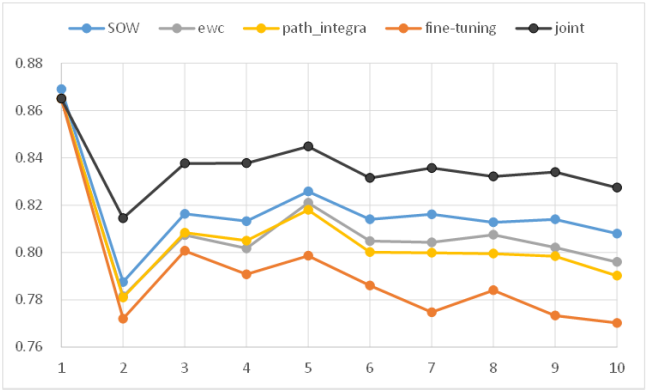


図 7　各タスク終了時における手法別分類精度

3.2.2　各タスク終了時での総合的な精度

図 7 は，Incremental-Task のシナリオに従って各タスクの学習を順次行っていた際，あるタスク（$task-N$）の学習完了時において，終了済みタスク $(Task-1, \cdots, Task-(N-1))$ のテスト用データと最期に終了したタスクのテストデータの分類精度を示したものである．同図からいずれのタスク完了時においても，提案手法の精度が ewc や path integral の精度を上回っていることがわかり，提案手法の優位性が確認できる．

4　おわりに

本稿では，全結合層を SOW(Self-Organized Weight) 層と呼ぶ層に置き換え，各タスク終了時に大きな値を持つ特異値のインデックスを追加記憶していく継続学習向き手法を提案した．忘却を抑止するために多くの手法で採用されているリプレイ用のデータを必要とせず，本手法は使用メモリ量の点で優位である．また，全結合層だけでなく畳み込み層も原理的に置き換えることができる．本提案手法を複数のデータセット，複数のモデル構造，複数種類のタスク数で網羅的に提案手法と既存手法 ewc, path integral の手法を比較した．その結果，提案手法が忘却を起こさずに継続学習を行えること，継続学習が進むと提案手法が既存手法の精度を上回ることが示され，本手法の優位性が確認できた．

文　献

[1] Zhenyi Wang and Enneng Yang and Li Shen and Heng Huang, ”A Comprehensive Survey of Forgetting in Deep Learning Beyond Continual Learning,” https://arxiv.org/abs/2307.09218, 2023.

[2] Huanrui Yang and Minxue Tang and Wei Wen and Feng Yan and Daniel Hu and Ang Li and Hai Li and Yiran Chen, ”Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification”, https://arxiv.org/abs/2004.09031, 2020.

人工データを活用したグラフ深層学習におけるプライバシー保護と精度向上

三谷　貴宏† 　菊地　良将† 　中西　宏和† 　佐々木　勇和† 　鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5

E-mail: †{mitani.takahiro,kikuchi.ryosuke,nakanishi.hirokazu,sasaki,onizuka}@ist.osaka-u.ac.jp

あらまし　グラフ深層学習を扱う上で，グラフデータのプライバシーの保護や利用許諾の遵守は重要な課題である．グラフデータを秘匿したまま学習できる既存手法として，連合学習や転移学習が存在する．しかし既存手法は学習済みモデルを共有する必要があり，学習済みモデルを共有できない環境では使用できない．提案手法では，ソースドメインを用いてモデルを作成した後，ターゲットドメインに対して推論を行う．学習と推論に別のグラフを使用することで，学習データのデータ量を確保する．学習時にソースドメインを模倣した人工データを使用することでプライバシーを保護し，学習済みモデルを共有可能にする．人工データの生成時にターゲットドメインの統計情報を共有し使用することでドメインシフトを軽減し，プライバシーを保護しつつ GNN の性能向上を目指す．実験では提案手法がプライバシーを保護しない手法と同程度以上の性能を達成した．

キーワード　グラフ深層学習，人工グラフ，ノード分類，プライバシー保護

1　はじめに

グラフとはノードとエッジで構成されたデータ構造であり，ノードが物事を，エッジがその関係を表すことで，様々なデータを表現することができる．グラフデータを解析することは様々な分野に応用することができるため，盛んに研究がおこなわれている．

グラフデータの健全な利用とデータ量の確保による GNN の性能向上の両立は重要な課題である．利用許諾や商業的な問題，プライバシー保護の観点等から，グラフデータあるいはグラフデータを学習したモデルを共有することができない状況が発生することは容易に考えられる．

1.1　問題と方針

グラフデータを秘匿しつつ GNN で学習をする手法としては，連合学習が高い精度を達成している [1]．連合学習では複数のグラフについてそれぞれ GNN で学習し，学習済みモデルのみを共有し統合することで，一つの学習済みモデルを作り出す．また，共有不可能なグラフデータの所有者が他のグラフで訓練されたモデルに対して転移学習 [2] を使用することでも，データを秘匿しつつ高い精度が得られる．転移学習では学習データであるソースドメインで学習したモデルを元に，推論したいデータであるターゲットドメインに合うようにモデルを調整する．

しかし既存手法の問題点として，学習済みモデルを共有できない場合，連合学習や転移学習を適用できないことが挙げられる．学習済みモデルは統計データの一種として捉えることができるが，学習済みモデルの権利的な扱いについては定説が定まっておらず，事例ごとに契約や利用規約が異なるというのが現状である [3]．元データの所有者の許可が無いままでは学習済みモデルの利活用および第三者との共有ができない規約や状況が存在し，このような場合は連合学習や転移学習などのの既存技術を適用できない．

既存技術を適用できない状況について，データの所有者の許可を取ることが現実的でない状況は容易に考えられる．利用目的がデータの集計時と異なるため再度データセットに含まれる全員の許可を取りなおす必要がある場合などが一般的な例である．学習済みモデルを共有するためにはデータの所有者の許可が必要であるとする考え方に基づく規約は多く存在する．例として ChatGPT 等を提供する OpenAI 社の Privacy policy [4] が挙げられ，データの所有者であるユーザが許諾しない意思を示した場合，サービスとして公開するモデルの学習データにそのデータを使用することができない．

ここで，統計情報はプライバシー情報を含んでおらず利活用および第三者との共有は自由であるとする考えがあり，日本の個人情報保護法などもこの立場である．本研究はこの考え方を前提とする．この場合，データを統計情報に加工した後，その統計情報のみを参照し生成された人工データであればプライバシーを含むデータではなくなり，人工データはプライバシー保護の観点から安全であると言うことができ，人工データを学習したモデルであれば他者と共有することができる．

1.2　提案手法

本研究では，ターゲットドメインのデータ量が小さい場合のため，ターゲットデータと類似する大規模なソースデータを活用してモデルを学習する．本研究では 2 つの技術的課題に取り組む．1) ソースデータのプライバシーを保護するため，ソースデータを模倣した人工データを生成して学習に使用する．具体的には，既存の人工データを生成する手法 [5] を活用する．2) 精度を向上させるため，ターゲットデータとソースデータ間のドメインシフトを軽減するよう，人工データを調整する．具体

的には，精度向上に効果が高いと考えられプライバシーを含まない統計情報として，クラスサイズ分布および次数分布の分布関数をターゲットデータから計算し，人工データの生成に使用する．

本研究では GNN の性能向上を目指すため，人工データとターゲットドメインとの間のドメインシフトを軽減することに着目した．ドメインシフトを軽減するため，ターゲットドメインの統計情報を共有してターゲットドメインの特徴を人工データに反映する．共有可能な統計情報として，クラスサイズ分布および次数分布の分布関数のパラメータを提案する．

実験では人工グラフデータの利用，ターゲットドメインの統計情報の利用がそれぞれ GNN の性能に与える影響を調査した．結果として 6 通りのうち 4 通りのデータセットで提案手法が最も良い精度を達成し，すべての場合でプライバシーを考慮しない手法に近い精度を記録した．

本論文の貢献点は以下の 3 つである．1) ソースデータを人工データ化しプライバシーを削除することによって，転移学習のように学習時と推論時で異なるデータを使用し学習データ量を確保する方法を，既存技術を適用できない状況でも適用できるようにした．2) プライバシーを保護しながらターゲットデータとソースデータ間のドメインシフトの問題を軽減するため，ターゲットデータの統計情報をソースドメインの人工データに反映した．3) 人工データ生成器の使用による性能への影響，および統計情報を人工データに反映しドメインシフトを軽減することによる性能への影響を実験によって確認し，プライバシーを保護しない手法と比較した．

2 事 前 準 備

2.1 グ ラ フ

本論文で用いるグラフは重み無しの無向グラフであり，以後これをグラフと呼ぶ．グラフ G のノード数を n とすると，グラフ G は $G = (\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{C})$ と表される．グラフの隣接行列 $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ とは，ノード i とノード j の間にエッジが存在する場合は $S_{i,j} = 1$，エッジが存在しない場合には $S_{i,j} = 0$ と定義される行列である．属性情報 $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ について属性情報の次元数は d であり，$\boldsymbol{X}_i$ は，ノード i に対して $\boldsymbol{X}_i \in \mathbb{R}^k$ として付与される k 次元の属性情報である．ラベル $\boldsymbol{C} = \{C_1, C_2, \cdots, C_k\}$ について，ラベルの種類は k であり，k をクラス数と呼ぶ．$\boldsymbol{\Omega}_i$ はクラスが C_i であるノードの集合である．

2.2 ノード分類

グラフデータを解析する主な手法の一つにグラフ深層学習 (GNN) がある．本論文では，グラフ深層学習のタスクとしてノード分類を扱う．ノード分類とは，グラフにおいてクラスが不明になっている一部のノードに対して，そのノードが属するクラスを予測するタスクである．あるグラフ G に対して，一部ノードのクラスが不明であるとする．このときノード分類タスクでは，G の情報を用いてクラスが不明なノードのクラスを推論することを目的とする．

GraphSAGE [6] はノード分類における最先端のグラフ深層学習の手法であり，ノードが属性情報を持つグラフにおいて特に高い精度を実現している．GraphSAGE は未知のノードに対して推論することが可能である．学習時と推論時で使用するグラフが異なるような実験設定は inductive な設定と呼ばれ，GraphSAGE は inductive な実験設定が可能である．従来の GNN ではあるグラフについて各ノードの埋め込み表現を学習しているのに対し，GraphSAGE では近傍情報を取得する集約関数を学習する．これにより未知のグラフに対して汎化した学習をすることが可能となる．GraphSAGE による inductive な実験設定では，学習時と推論時に使用するグラフで属性情報が表現するものが同じである必要があり，本論文でもこれを前提としている．

2.3 グラフの統計情報

統計情報とはデータセット内のデータが集約化されたものである．グラフデータの統計情報の例として，ノード数やエッジ数，クラスサイズ分布，次数分布などがある．

統計情報を他の情報と組み合わせたとしても，統計情報から個人のデータを特定することができなくなっている必要がある．この統計情報は個人情報を含んでおらず，統計情報を他の情報と照合しても個人を特定することができないため，プライバシーに関する情報を含んでいないことが分かる．そのため，元データの所有者の許可が無くても統計情報の利活用および第三者との共有が可能である [1]．

統計情報のソースドメインとターゲットドメイン間での受け渡しが可能であるため，統計情報を後述するドメインシフトの軽減に利用することができると考えられる．本節では特にドメインシフトの軽減に有効であると考えられる，クラスサイズ分布および次数分布の分布関数について説明する．そして，統計情報を扱う上で注意する必要がある非構造的ゼロによる開示についても説明する．

2.3.1 クラスサイズ分布

本論文では，グラフデータのクラスサイズ分布 $\boldsymbol{\rho} \in \mathbb{R}^k$ を扱う．クラスサイズ分布とは，グラフにおいて各クラスに属するノードの数を集計した統計情報である．データセットにおいて，クラスサイズ分布が偏っていることは珍しくない [7]．

クラスサイズ分布は統計情報であるため，クラスサイズ分布を与えられても，各ノードのラベルなどのグラフデータの詳細を知ることはできない．

2.3.2 次 数 分 布

本論文では，次数分布およびその分布関数を扱う．次数とはノードにつながっているエッジの数を表し，ノード v の次数は

1：条件が揃うと，後述する非構造的ゼロによる開示によって個人情報が特定される可能性がある．統計情報の内容や量によっては共有することができない場合があるため，注意が必要である．また，統計情報がプライバシーを含んでおらず，第三者への提供に本人の許諾が必要ないとする立場は普遍的なものではなく，日本の個人情報保護法を含む一部の文脈で共有されている考え方であることに注意が必要である．本論文ではこの考え方を前提とする．

$deg(i)$ と表記される．ノード数を n とすると，次数分布 $\boldsymbol{\theta} \in \mathbb{R}^n$ について $\theta_i = deg(i)$ である．

次数分布の分布関数とは，次数分布を降順に並べたとき，その分布に従うような関数である．現実世界のネットワークグラフ等の実データにおいて，一般的に次数分布はべき乗則に従うことが知られており [8]，この性質はスケールフリー性と呼ばれている [9]．べき乗則の確率密度関数は次のように表すことができる．

$$f(x, \alpha) = \alpha x^{\alpha - 1} \tag{1}$$

α はべき指数と呼ばれるパラメータである．この確率密度関数をパラメータ loc だけシフトする，またはパラメータ $scale$ だけ倍率をかけることで実データに近い分布関数を作成することができる．以降，この α，loc，$scale$ を次数分布の分布関数のパラメータと呼ぶ．

2.3.3 非構造的ゼロによる開示

統計情報からプライバシーが開示される場合が存在する．統計情報においてあるグループに該当するデータ数が極端に小さい場合，他の情報と照合することでそのグループに関するデータがリークしてしまうことがある．例として，あるグループ A に属する人について，ある質問に「はい，いいえ」と答えた人数がそれぞれ公表されたとする．このときいいえと答えた人数が 0 であった場合，グループ A に属する人間は全員がその質問にはいと答えたことが分かり，これはプライバシーのリークとなる．このように統計情報上の 0 のような極端に小さい値からデータがリークすることを，非構造的ゼロによる開示と呼ぶ．統計情報を第三者と共有する場合は，非構造的ゼロによる開示に注意する必要がある．

統計情報からプライバシーが開示されるリスクがあるため，共有する統計情報の内容に注意し，量を最小限にする必要がある．

2.4 ドメインシフト

ドメインシフトがある状態とは，ソースドメインとターゲットドメインの間でデータの分布が異なる状況である．本研究のようにソースドメインのデータで学習した GNN をターゲットドメインのデータの推論に適用する場合，ドメインシフトの存在によって GNN の性能が低下してしまう．これは GNN がソースドメインに基づいて学習をし，ターゲットドメインの分布が異なっていた場合は誤った分布に基づいた推論をすることになるためである．一般に異なるデータセット間では分布は偏っているため，ドメインシフトが存在する状況は生じ得ると言える [10]．

3 提案手法

提案手法では，ターゲットデータと類似する大規模なソースデータを活用してモデルを学習する．ターゲットデータはデータ量が小さく，単体では十分な学習ができないとする．非公開データセットのデータ量は小さいことが多い．学習データのデータ量の確保は重要な課題となっており，転移学習 [11] など様々な研究が行われている．提案手法ではターゲットデータと類似するソースデータを用いることで，十分な量の学習データを確保する．

提案手法の概要図を図 1 に示す．提案手法ではソースドメインの統計情報を計算し，統計情報を元にして人工データを生成する．GNN の性能向上のため，人工データの生成時に使用する統計情報の一部をターゲットドメインの統計情報で上書きする．生成した人工データを GNN で学習し，学習済みモデルを共有する．最後に，学習済みモデルを用いてターゲットドメインを推論する．

本研究では 2 つの技術的課題として，プライバシーの保護および精度の向上に取り組む．

提案手法ではソースデータのプライバシーを保護するため，ソースデータを模倣しプライバシーを含まない人工データを生成して学習に使用する．統計情報はプライバシーを含まず，自由に利活用や共有が可能であるとされている．よって，統計情報を元に生成した人工データにはプライバシーは含まれておらず，データの所有者は存在しないと考えられる．このような人工データを学習したモデルであれば，学習済みモデルの共有に学習データの所有者の許可を必要とする規約や契約の中でも自由に共有することができる．したがって，提案手法ではソースデータの統計情報を計算し，統計情報だけを参照して学習のための人工データを生成する．人工データの生成器には，提案手法で必要とされる機能を持つ既存の人工データ生成手法 [5] を活用する．

学習時と推論時で使用するグラフデータが異なることや，統計情報から生成した人工データの使用による情報量の減少によって，GNN の性能が低下することが考えられる．精度を向上させるため，ターゲットデータとソースデータ間のドメインシフトを軽減することを考える．提案手法ではターゲットデータとの間のドメインシフトを軽減するよう，人工データを調整する．具体的には，ターゲットデータの統計情報を計算し，人工データの生成に使用するソースデータの統計情報をターゲットデータの統計情報で一部上書きすることで，人工データにターゲットデータの情報を反映する．ソースドメインおよびターゲットドメイン間で共有するものは統計情報のみであるため，プライバシーは保護される．本論文では精度向上に効果が高いと考えられる統計情報として，クラスサイズ分布および次数分布の分布関数をターゲットデータから計算し，人工データの生成に使用する．

なお，共有する統計情報の情報量は可能な限り抑えるべきである．統計情報の正確な値を共有することには非構造的ゼロによる開示のリスクがあり，規約や契約で共有を制限されることも考えられる．統計情報に含まれる情報量が大きいほどリスクは高まると考えられるため，共有する統計情報の内容は制限する必要がある．ソースドメインやターゲットドメインのグラフデータの全ての統計情報を共有しないのはこのためである．

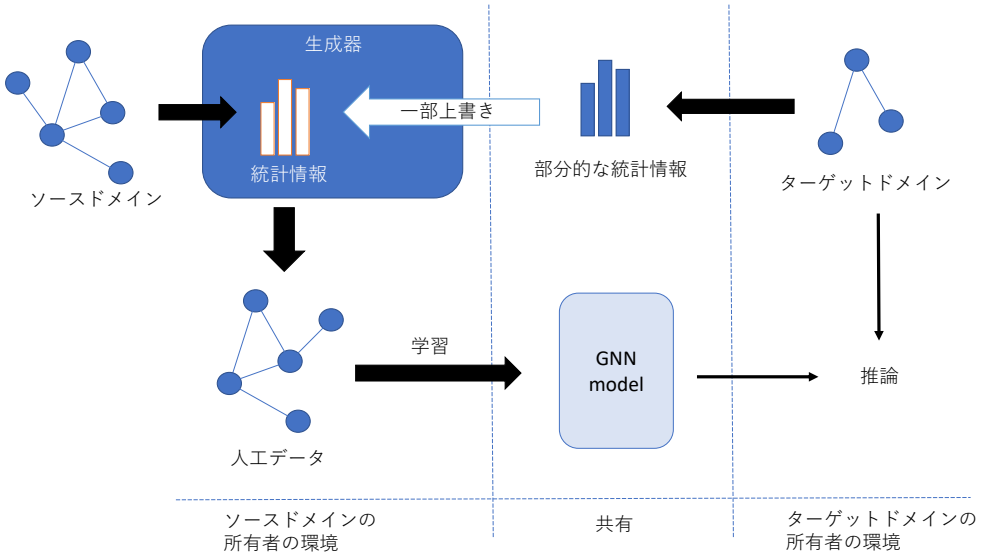


図 1: 提案手法の概要図．ターゲットドメインから抽出したクラスサイズ分布とソースドメインを GenCAT に入力し，人工グラフデータを生成する．その後 GNN でこれを学習し，学習したモデルでターゲットドメインを推論する．

3.1 共有可能な情報

ソースドメインとターゲットドメイン間にはドメインシフトが存在するため，ソースドメインを模倣して生成した人工データとターゲットドメイン間にもドメインシフトが存在する．これにより GNN の性能が低下するため，ドメインシフトを軽減することを目指す．人工データの生成時にターゲットドメインの情報を入力して反映することでドメインシフトを軽減することができるが，ターゲットドメインを共有することはできないため，グラフデータのリークが発生しない程度に加工された統計情報を使用する．

ここで，ドメインシフトの問題はいくつかの問題設定に分類することができる [12]．本研究で想定するドメインシフトは共変量シフトと呼ばれ，各ドメインが本質的に表すものは同じであるが，入力変数の分布が異なるというドメインシフトである．しかし，共有する統計情報の内容を工夫することで，様々なドメインシフトに対応できるようになることが期待される．

本論文では，ターゲットドメインの詳細な情報を秘匿しつつ推論の精度向上に寄与する共有可能な情報として，次数分布の分布関数のパラメータおよびクラスサイズ分布を提案する．ターゲットドメインが共有可能な加工された統計情報はユースケースによって変化することが考えられる．次数分布の分布関数のパラメータとクラスサイズ分布は GNN の性能向上に貢献しつつ，多くのユースケースで共有を許容できると考えられる一例である．

次数分布の分布関数およびクラスサイズ分布の調整は常には GNN の性能向上に貢献せず，それぞれの分布についてドメインシフトが大きい場合に特に性能向上に貢献すると考えられる．

提案手法ではターゲットドメインから次数分布の分布関数またはクラスサイズ分布を計算し，共有可能な情報のみをソースドメインの所有者と共有する．ソースドメインの所有者はソースドメインのグラフデータと共有された情報を生成器に入力し，ターゲットドメインの情報を反映させた人工データを生成する．

なお，データを集約し統計情報を共有する際は，非構造的ゼロによるリークに気を付ける必要がある．

3.2 次数分布の分布関数の推定

提案手法では，グラフデータから次数分布の分布関数を抽出する必要がある．次数分布の分布関数はターゲットドメインの次数分布の分布関数を共有する場合や，人工データのノード数を操作する場合に必要である．

本論文では，グラフデータに次数分布の分布関数をフィットさせ，これのパラメータを共有することを提案する．一般的に次数分布はべき乗則に従う [8] ことが知られているため，べき乗則をグラフデータにフィットさせることを考える．

次数分布の分布関数を実データにフィットさせる方法として，最尤推定を使用する．最尤推定では，データの確率分布がパラメータ λ を持つと仮定し，そこから n 個のデータが得られたとき，λ を以下のような式で推定することができる．

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\lambda) \tag{2}$$

ここで $x_1, x_2, \cdots, x_n$ はフィットさせたいデータであり，$p(x_i|\lambda)$ はパラメータ λ を持つ分布関数から値 x_i が得られる確率である．

よって，最尤推定を使用することで，実データの次数分布から次数分布の分布関数のパラメータを推定することができる．このパラメータは加工された情報であり，次数分布の分布関数

を与えられても，具体的な次数分布などのグラフデータの詳細を知ることはできない．

3.3 GenCAT による人工データ生成

提案手法のグラフ生成器に求められる機能は 4 つあり，既存のグラフを模倣し，新たなグラフデータを生成できること，模倣の過程で元となったデータを統計情報に加工し，実データをそのまま人工データに用いないこと，属性情報を扱うことができること，構造情報や属性情報を考慮してノードにラベルを付与できることである．グラフの模倣時に一度統計情報へ加工することはプライバシー保護のために必要である．属性情報を扱えること，それに基づいてノードにラベルを付与できることは inductive な実験設定において精度を確保するために必要である．

GenCAT [5] はこれらの要求を満たす唯一の手法である．よって，提案手法ではグラフ生成器として GenCAT を使用する．GenCAT は最先端の属性付きグラフ生成器であり，クラスサイズ分布，クラス間のエッジ接続割合，属性情報，グラフサイズを考慮してグラフデータを柔軟に生成することができる唯一の手法である [13].

GenCAT を使用することで，実データの統計情報を用いて構造情報および属性情報を模倣した人工グラフを新しく生成することができる．本論文ではソースドメインのグラフデータを GenCAT に入力することで，ソースドメインを模倣した人工データを生成する．

ここで，GenCAT が実データの模倣に使用するパラメータのうち，クラスサイズ分布および次数分布をターゲットドメインのものに変更することを提案する．クラスサイズ分布については，ターゲットドメインが共有したものをそのまま入力することができる．次数分布は分布関数のパラメータが共有されるため，パラメータに基づいて分布関数から新しく次数分布を生成し，これを入力する．これによりターゲットドメインの特徴を反映した人工データを生成し，ドメインシフトの軽減による GNN の性能向上を目指す．

人工データのノード数を操作する場合は，データから計算した次数分布の分布関数を基にして新たな次数分布を生成する．次数分布の分布関数が既知であれば任意のノード数についての次数分布を生成することができる．

3.4 GNN による学習と推論

提案手法の GNN に求められる機能は 2 点あり，inductive な問題設定が可能であることと，属性情報を扱えることである．inductive な問題設定が可能であることでソースドメインとターゲットドメインを分けることができる．これはターゲットドメインが小さすぎて学習できないという条件では必要な機能である．精度のため GenCAT と同様に，GNN も属性情報を扱うことができる必要がある．

GraphSAGE [6] は最先端のグラフ深層学習の手法であり，これらの条件を満たしている．よって，提案手法では GNN として GraphSAGE を使用する．

なお inductive な問題設定においては，使用するデータセット間で属性情報の次元数が等しい必要があり，属性情報の各次元が同じものを表している必要がある．

4 実　　験

本論文では提案手法を様々な設定で実験し，他の単純な方法と比較する．これにより，人工データの使用およびターゲットドメインの統計情報の使用が GNN の性能に及ぼす影響について検証する．

4.1 サンプリングによる人工データ生成

人工データを生成する方法について，GenCAT の比較対象としてサンプリングを用いた方法で実験する．この方法ではソースドメインをサンプリングして新たな人工データとするが，GenCAT を使用する場合と異なりソースドメインの情報がそのまま使用される．サンプリングは性能評価のための比較対象であり，人工データの生成器の候補ではないことに注意する必要がある．

サンプリングではクラスサイズ分布を使用し，クラスごとに目標のノード数になるまでランダムにノードをコピーまたは削除する．使用するクラスサイズ分布は，ターゲットドメインのクラスサイズ分布を利用できる場合はこれを利用し，利用できない場合は全クラスのノード数が平均値となるような一様分布をクラスサイズ分布として使用する．

4.2 データセット

表 1: データセットの詳細情報.

データセット名	ノード数	エッジ数	平均次数	属性情報の次元数	クラス数
Wisconsin	251	458	3.649	1703	5
Cornell	183	278	3.038	1703	5
Texas	183	287	3.137	1703	5

本実験で用いるデータセットとして，WebKB データセットのサブセットである Cornell，Texas，Wisconsin データセットを用いる．このグラフデータではノードは Web ページを表しており，エッジはページ間のハイパーリンクである．ノードの属性情報は各ページにある単語が含まれるかどうかであり，ラベルはそのページの属するカテゴリーに応じて 5 つに分けられている．train データ，validation データ，test データの割合について，毎回ランダムな 6 割のノードを train データに，ランダムな 2 割のノードを validation データに，残りの 2 割のノードを test データに割り振る．

データセットの統計情報を表 1 に，クラスサイズ分布を図 2 に示す．

これらのデータセットを本実験で用いる理由は，互いに似通ったデータセットであり，属性情報の次元数が等しく，かつ広く使われているデータセットであるためである．

提案手法には，ソースドメインを人工データで模倣する

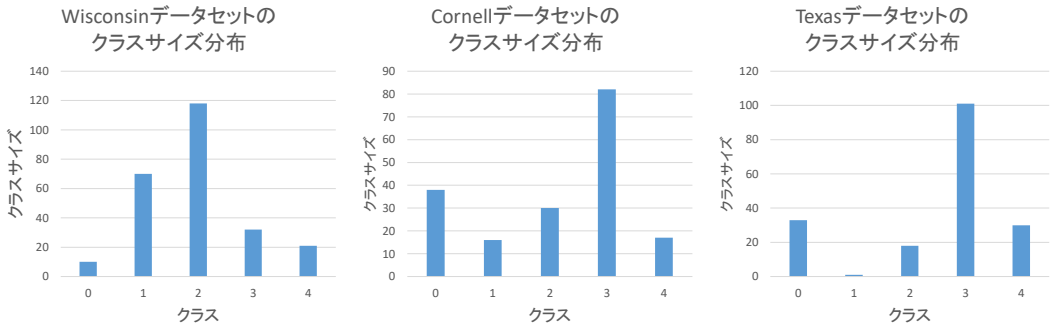


図 2: データセットごとのクラスサイズ分布を．左:Wisconsin データセット，中:Cornell データセット，右:Texas データセット．

ことにより，ターゲットドメインのデータ量が小さい場合でも高品質なデータを準備できるという利点がある．Cornell,Texas,Wisconsin データセットはグラフサイズが比較的小さいデータセットである．これらのデータセットをソースドメインとして用いた場合，このような利点を活用することはできないことに留意する必要がある．

4.3 実験設定

本実験では Cornell，Texas，Wisconsin データセットを用いて，ソースドメインとターゲットドメインの 6 通りの組み合わせについて実験する．実験設定は人工データの生成方法によって 8 通りに場合分けする．

実験設定について、人工データを生成する手法と使用する統計情報によって設定した 8 通りの場合を比較する．具体的には，ソースドメインをそのまま学習に用いる場合，サンプリングでクラスサイズ分布をターゲットドメインに合わせる場合，提案手法でクラスサイズ分布および次数分布をソースドメイン，ターゲットドメインに合わせる場合の組み合わせ 4 通りに加えて，サンプリングと提案手法でクラスサイズ分布に一様分布を用いる場合を加えた計 8 通りである．クラスサイズ分布を一様分布にする場合について，これはターゲットドメインの情報が一切使用できない場合でも使用することができる実験設定の一例である．一様分布を使用することで，クラスサイズ分布のドメインシフトが大きい場合に影響を軽減することができると考えられる．

公平な比較をするため，サンプリングおよび GenCAT を用いて生成する人工データのノード数は，ソースドメインのノード数に合わせるものとする．

4.4 比較実験

比較実験の結果のうち精度を表 2 に，macro-f1 の値を表 3 に示す．各実験設定についてシードをランダムに変更しながら実験を 10 回試行し，その結果の平均値を記載し，± 以下に標準偏差を記載する．データセット間の類似度の指標の一つとして，クラスサイズ分布の比率のコサイン類似度を類似度として表 4 に記載している．

表内のサンプリングの行では人工データの生成器としてサンプリングを使用した場合の，提案手法の行では GenCAT を使用した場合の数値を記載している．クラスサイズ分布が一様分布となっている行は目標とするクラスサイズ分布として一様分布を使用した場合を，ターゲットとなっている行はターゲットドメインのクラスサイズ分布を使用した場合を，ソースとなっている行はソースドメインのクラスサイズ分布を使用した場合をそれぞれ表している．次数分布についても同様に，ターゲットとなっている行はターゲットドメインの次数分布の分布関数を使用した場合を表しており，ソースとなっている行はソースドメインの次数分布を使用した場合を表している．

ターゲットドメインの統計情報を共有した場合の提案手法の精度について他の手法と比較すると，データセットの組み合わせ 6 通りのうち 4 通りで最も良い精度を達成した．単純な手法が高い精度を達成する 2 通りの場合でも，提案手法は人工データを使わない手法の 96%および 80%の精度を達成した．ここで表 4 より，提案手法が最も良い精度を達成したデータセットの組み合わせはドメインシフトが大きく，単純な手法が高い精度を達成するさ組み合わせではドメインシフトが小さいと考えられる．この結果より，提案手法はドメインシフトが大きいデータセットの組み合わせでは高い精度を達成することができ，ドメインシフトが小さいデータセットの組み合わせでもプライバシーの保護を考慮しない手法に近い精度を出すことができると言うことができる．

実験結果より，提案手法でソースドメインの情報のみを使用した場合，単純な手法の精度が低い場合は提案手法で精度が向上し，単純な手法の精度が高い場合は提案手法で精度が減少した．これについて，提案手法でソースドメインの情報のみを使用した場合の推論結果は，学習できていない時の結果に近づくと考えられる．今回使用したデータセットのクラス数は 5 であるため，推論を完全にランダムに出力した場合の精度は 0.2 に近づくと考えられる．表 2 の結果より，提案手法でソースドメインの情報のみを使用した場合の精度は 0.2 から 0.3 程度の値に近づいていると考えられ，適切な学習ができていないことが

表 2: 比較実験の精度の結果．各データセットにおいて精度が最も良いものは太字で，次いで精度が良いものは下線で強調している．

ソースドメイン			Cornell	Wisconsin	Wisconsin	Texas	Cornell	Texas
ターゲットドメイン			Wisconsin	Cornell	Texas	Wisconsin	Texas	Cornell
手法名	クラスサイズ分布	次数分布						
生成器無し	-	-	0.1431 ± 0.0370	0.1162 ± 0.0286	0.1189 ± 0.0364	0.0764 ± 0.0397	0.4973 ± 0.1020	0.4216 ± 0.0676
サンプリング	一様分布	-	0.0941 ± 0.0355	0.1270 ± 0.0624	0.0765 ± 0.0437	0.1725 ± 0.0783	0.5081 ± 0.0925	0.1919 ± 0.0725
サンプリング	ターゲット	-	0.1960 ± 0.1169	0.1162 ± 0.0637	0.1756 ± 0.0875	0.2960 ± 0.0730	$\mathbf{0.5513 \pm 0.0676}$	$\mathbf{0.4378 \pm 0.1200}$
提案手法	ソース	ソース	0.1999 ± 0.1478	0.1621 ± 0.0569	0.1486 ± 0.0997	0.1372 ± 0.0852	0.3026 ± 0.1501	0.2837 ± 0.1060
提案手法	一様分布	ソース	0.2509 ± 0.1299	0.1810 ± 0.0953	0.2108 ± 0.1193	0.2097 ± 0.0867	0.1216 ± 0.0808	0.1892 ± 0.0900
提案手法	ターゲット	ソース	$\mathbf{0.3019 \pm 0.1074}$	$\mathbf{0.2918 \pm 0.1041}$	0.4594 ± 0.1103	$\mathbf{0.3274 \pm 0.0929}$	0.4783 ± 0.1621	0.3162 ± 0.1480
提案手法	ソース	ターゲット	0.1932 ± 0.0907	0.1838 ± 0.0889	0.0973 ± 0.0626	0.2353 ± 0.1401	0.4675 ± 0.1167	0.3369 ± 0.0993
提案手法	ターゲット	ターゲット	0.2901 ± 0.1512	0.2756 ± 0.1358	$\mathbf{0.5081 \pm 0.1017}$	0.2862 ± 0.1021	0.4513 ± 0.1714	0.2513 ± 0.1181

表 3: 比較実験の macro-F1 の結果．各データセットにおいて macro-F1 が最も大きいものは太字で，次いで大きいものは下線で強調している．

ソースドメイン			Cornell	Wisconsin	Wisconsin	Texas	Cornell	Texas
ターゲットドメイン			Wisconsin	Cornell	Texas	Wisconsin	Texas	Cornell
手法名	クラスサイズ分布	次数分布						
生成器無し	-	-	0.1309 ± 0.0425	0.1000 ± 0.0332	0.0674 ± 0.0210	0.0631 ± 0.0370	$\mathbf{0.3421 \pm 0.0621}$	0.2604 ± 0.0571
サンプリング	一様分布	-	0.0937 ± 0.0242	0.0869 ± 0.0456	0.0498 ± 0.0307	0.0880 ± 0.0358	0.2825 ± 0.0648	0.1515 ± 0.0782
サンプリング	ターゲット	-	0.0935 ± 0.0537	0.0696 ± 0.0307	0.1154 ± 0.0419	0.1322 ± 0.0473	0.2238 ± 0.0658	$\mathbf{0.2899 \pm 0.1317}$
提案手法	ソース	ソース	0.1023 ± 0.0660	0.1139 ± 0.0771	0.0745 ± 0.0467	0.0963 ± 0.0467	0.1330 ± 0.0434	0.1498 ± 0.0423
提案手法	一様分布	ソース	0.1151 ± 0.0358	0.1066 ± 0.0556	0.1276 ± 0.0556	0.1105 ± 0.0437	0.0695 ± 0.0448	0.1363 ± 0.0854
提案手法	ターゲット	ソース	$\mathbf{0.1431 \pm 0.0341}$	0.1073 ± 0.0349	0.1659 ± 0.0360	$\mathbf{0.1501 \pm 0.0454}$	0.1816 ± 0.0452	0.1323 ± 0.0417
提案手法	ソース	ターゲット	0.1033 ± 0.0487	0.0930 ± 0.0455	0.0515 ± 0.0344	0.1081 ± 0.0357	0.1801 ± 0.0561	0.1292 ± 0.0280
提案手法	ターゲット	ターゲット	0.1254 ± 0.0654	$\mathbf{0.1291 \pm 0.0479}$	$\mathbf{0.1685 \pm 0.0206}$	0.1250 ± 0.0364	0.1611 ± 0.0472	0.1129 ± 0.0461

表 4: データセット間のクラスサイズ分布の比率のコサイン類似度．

ソースドメイン	Cornell	Wisconsin	Texas
ターゲットドメイン	Wisconsin	Texas	Cornell
類似度	0.5706	0.4076	0.9684

分かる．学習ができていない理由としては，統計情報を計算し人工データを生成する過程で情報が失われるためであると考えられる．

よって，人工データを使用することによる影響は，情報が失われ，GNN の性能が下がることであると言うことができる．

表 2 の結果より，提案手法でターゲットドメインの統計情報を共有した場合について，すべての場合においてソースドメインの情報のみを用いた場合よりも精度が向上し，ほとんどの場合において macro-F1 も向上した．ターゲットドメインの統計情報を人工データの生成時に使用することでドメインシフトを軽減でき，それにより GNN の性能が向上したと考えられる．

よって，ターゲットドメインの統計情報を使用することによる影響は，ドメインシフトを軽減し GNN の性能を向上させることであると言うことができる．

共有する統計情報ごとの影響について実験結果より，クラスサイズ分布は精度の向上に大きく貢献することが分かる．しかし次数分布の分布関数は精度の向上に必ずしも貢献せず，精度が落ちることがあった．この原因は，実験に使用したデータセット間でクラスサイズ分布のドメインシフトが大きく，次数分布のドメインシフトが小さかったためであると考えられる．クラスサイズ分布のドメインシフトは図 2 よりクラス 1 やクラス 2 で特に大きいことが分かり，これを調整することで精度が向上すると考えられる．しかし次数分布について，使用したデータセットは全て Web ページを基にしたデータセットであり，べき乗則に従うと考えられる．よって，これを調整することにより向上する精度は小さいと考えられる．

統計情報の代わりに一様分布を使用することについて表 2 より，クラスサイズ分布に一様分布を使用することで，6 通りのデータセットの組み合わせのうち 4 通りの組み合わせにおいて精度が向上し，そのうち 3 通りの組み合わせにおいて macro-F1 も向上した．他 2 通りの組み合わせでは精度，macro-F1 ともに低下した．この実験結果と図 2 より，一様分布の使用は使用するデータセット間のドメインシフトが大きい場合にのみ有効であることが分かる．

5 関 連 研 究

5.1 プライバシー保護

グラフデータの匿名化の程度に関する研究は多くはない．一般に広く使用されるプライバシー保護の度合いの評価指標のようなものはなく，匿名化の程度について議論されていない研究が多いというのが実情である．

匿名化の指標について，ノードの次数に着目した k 次匿名 [14] という指標が提案され，この研究では同じ次数を持つノードが複数あることで匿名性が保たれるとした．しかし，データが偏っており攻撃者が背景情報を持っている場合，k-匿名化を適用した場合であっても個人情報が特定される可能性がある．この問題に対応するため，l-多様性 [15] が提案された．l-多様性ではプライバシー情報に一定以上の多様性があることを保証することでデータの偏りを解消し，プライバシーを保護する．t-近接性 [16] は k-匿名性について，あるグループ内の多様性がデータ全体の多様性に一定以上近いことを保証することで，データの偏りによる個人情報の特定を防ぐ．

5.2 人工グラフデータ

人工グラフデータは様々な目的で活用されることが期待されており，GNN の学習用データの拡張や改変，プライバシー保護などがその一例である．

人工グラフの生成において，コミュニティ構造はグラフの重要な特徴の一つである．構造情報を再現することでグラフの情報量が増加し，GNN の性能向上に貢献する．ノードの属性情報も重要な要素であり，グラフの情報量を増加させる．ANC [17] はコミュニティ構造を考慮したグラフ生成を可能とし，ノードに属性情報を付与することができる人工グラフ生成器である．

近年，深層学習を用いて実データを模倣し，人工グラフを生成する手法も提案されている．GraphVAE [18] は GCN を用いて隣接行列を学習し，生成モデルを構築した．NetGAN [19] はグラフ上のランダムウォークを GAN で学習することで特徴を捉え，人工グラフを生成した．

本研究では提案手法の一部として GenCAT [5] を使用した．GenCAT は提案手法に必要な機能を備えた唯一の手法である．

6 終わりに

本論文では，ソースドメインで学習しターゲットドメインについて推論できる共有可能なモデルを人工データを用いて訓練する手法を提案した．提案手法では推論をしたいグラフのデータ量が小さい場合を考慮して，ターゲットドメインに類似したソースドメインを用いて学習する．プライバシー保護のため，ターゲットドメインの実データは推論時のみに使用する．プライバシーの問題を回避するため，ソースドメインの統計情報を用いて人工データを生成し，これを GNN の学習に使用した．ドメインシフトを解消するため，ソースドメインを模倣した人工データを生成する際，ターゲットドメインの統計情報を共有し，人工データに反映した．実験では，提案手法がプライバシーを保護しない手法と同程度以上の精度を達成することを確認した．

謝　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです．

文　献

[1] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Tao Qi, Yongfeng Huang, and Xing Xie. A federated graph neural network framework for privacy-preserving personalization. *Nature Communications*, Vol. 13, , 2022.

[2] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *In Proceedings of the ICLR*, 2020.

[3] 板倉陽一郎, 熊谷雄介, 猪谷誠一. 学習用データセットおよび学習済みモデルのライセンスに関する法的問題. 情報法制研究, Vol. 11, , 2022.

[4] OpenAI. Privacy policy. https://openai.com/policies/privacy-policy, 2023.

[5] Seiji Maekawa, Yuya Sasaki, George Fletcher, and Makoto Onizuka. GenCAT: Generating attributed graphs with controlled relationships between classes, attributes, and topology. *Information Systems*, Vol. 115, , 2023.

[6] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *In Proceedings of the NIPS*, 2017.

[7] Aida Ali, Siti Mariyam Hj. Shamsuddin, and Anca L. Ralescu. Classification with class imbalance problem: A review. In *Soft Computing Models in Industrial and Environmental Applications*, 2015.

[8] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[9] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *IEEE Circuits and Systems Magazine*, Vol. 3, No. 1, 2003.

[10] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, 2009.

[11] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, Vol. 109, No. 1, 2021.

[12] Jose G. Moreno-Torres, Troy Raeder, Rocio Alaiz-Rodriguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, Vol. 45, No. 1, 2012.

[13] 前川政司, 野田昂希, 佐々木勇和, 鬼塚真. 多様な人工グラフを用いた GNN によるノード分類の実証研究. データ工学と情報マネジメントに関するフォーラム, 2023.

[14] Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In *In Proceedings of the SIGMOD*, 2008.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. In *In Proceedings of the ICDE*, 2006.

[16] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *In Proceedings of the ICDE*, 2007.

[17] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, Vol. 78, No. 5, 2008.

[18] Martin Simonovsky and Nikos Komodakis. GraphVAE: Towards generation of small graphs using variational autoencoders. In *In Proceedings of the ICANN*, 2018.

[19] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. NetGAN: Generating graphs via random walks. In *In Proceedings of the ICML*, 2018.

Approach to Efficient Neural Architecture Search on the Latent Space of VGAE Models with Multiple Datasets

逸見　一喜[†,††]　谷垣　勇輝[††]　大西　正輝[††,†]

† 筑波大学　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 産業技術総合研究所　〒 305–8560 茨城県つくば市梅園 1-1-1
E-mail: †kazuki.one.jack.work@gmail.com, ††tanigaki.yuki@aist.go.jp, †††onishi@ni.aist.go.jp

あらまし　Neural Architecture Search（NAS）は，データや目的に応じてニューラルネットワークのアーキテクチャを自動で最適化する AutoML 手法の一つである．しかし，NAS は膨大な探索空間により高い計算コストが必要であり，新たなデータセットに適したアーキテクチャを 0 から探索することは非効率である．そこで本研究では，既存データセットに適したアーキテクチャを潜在空間にマッピングすることで，未知のデータセットに適したアーキテクチャを生成するフレームワークを提案する．また，従来の NAS は主に単一データセットの精度を最適化するが，実世界では異なる特性を持つ複数のデータセットに対して適応するアーキテクチャが重要であるため，複数のデータセットに適したアーキテクチャのトレードオフを特徴量空間に落としこむ方法も検討する．NAS のアーキテクチャがグラフ構造で表現できることから，潜在特徴量のマッピングには Variational Graph Auto-Encoders に基づく Encoder-Decoder 変換モデルを用いる．実験の結果，提案手法は従来法に比べて高い探索効率と優れた性能を示した．

キーワード　Neural Architecture Search，AutoML，多目的最適化，Graph Neural Network

1　序　　論

近年，データセットやタスクに合わせて最適なニューラルネットワークのアーキテクチャを自動的に探索する Neural Architecture Search (NAS) は AutoML (自動機械学習) の領域において重要な技術となっている．NAS が登場する以前はアーキテクチャの構築は基本的にハンドクラフトな方法で行われており，性能向上のためには試行錯誤を繰り返す必要があった．特に，ハンドクラフトなアーキテクチャの構築は深層学習の専門的な知識が必要であり，膨大な時間が必要であった．しかし，NAS の登場によりアーキテクチャ探索の試行錯誤を自動的に行うことで，深層学習の専門的な知識を有していなくても高性能なネットワークアーキテクチャが容易に構築可能となった．一方で NAS はネットワークアーキテクチャの探索にかかる時間が大きいという欠点がある．

NAS を用いた最適なアーキテクチャの探索は大量の計算リソースと時間を必要とする．よって，新たなデータセットが与えられるたびに適したアーキテクチャを 0 から探索することは非効率である．今後予想されるデータセットの増加に対応するためには，未知のデータセットに最適なアーキテクチャを直接推測する技術が必要となる．そこで，私たちは既存のデータセットに適したアーキテクチャを潜在特徴量空間にマッピングすることで，NAS を行わずに未知のデータセットに適したアーキテクチャを生成するフレームワーク Variational Graph Auto-Encoders Neural Architecture Search (VGAE–NAS) を提案する．潜在特徴量のマッピングには，VGAE に基づいた Encoder-Decoder 変換モデルを用いる．VGAE は Encoder と Decoder を用いてグラフの特徴量抽出を行う手法であり，グラフ構造で表現される NAS のアーキテクチャのマッピングに適している．入力するアーキテクチャと再構成されたアーキテクチャの誤差を減らすことで，潜在空間の潜在特徴量がアーキテクチャの特徴を表現する．マッピング後の潜在特徴量空間から VGAE の Decoder により直接アーキテクチャを生成することで，NAS を実行せずとも新たなデータセットに対して有望なアーキテクチャを生成することができる．新たなデータセットに対して 0 からアーキテクチャを探索する計算コストを大幅に削減できるため，NAS におけるアーキテクチャ情報の転移は非常に重要である．提案手法は多様なデータセットに対する最適なアーキテクチャの構築を目指した手法である．また，従来の NAS は主に単一データセットの精度を最適化するが，実世界では異なる特性を持つ複数のデータセットに対して適応する

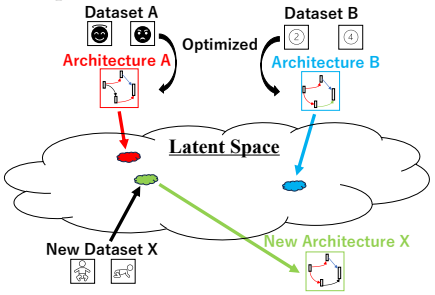


図 1: 本研究のコンセプト，データセット A，B を基に，それぞれアーキテクチャA，B を最適化し構築する．最適化されたアーキテクチャA，B は潜在特徴量空間にマッピングされる．新しいデータセット X が生成されたとき，そのデータセットの特徴を考慮した最適なアーキテクチャX が潜在空間から瞬時に生成される．

アーキテクチャが重要であるため，複数のデータセットに適したアーキテクチャのトレードオフを特徴量空間に落としこむ方法も検討する．本研究の概要を**図 1** に示す．

本研究の構成は以下の通りである．第 2 節はベースとなる NAS や GNN，関連手法について説明し，第 3 節では提案手法の詳細について述べる．第 4 節は提案手法を用いた実験設定とその結果について考察を行う．そして第 5 節にて本研究の結論を述べる．

2 関 連 研 究

本節では，本研究で扱う NAS 手法や GNN について述べる．

2.1 Neural Architecture Search (NAS)

NAS は AutoML (自動機械学習) の一分野であり，深層学習で扱う最適なニューラルネットワークのアーキテクチャを自動的に探索する最適化手法である．NAS では主に層（Layer）の繋ぎ方と演算（Operation）の選択といった 2 つの問題を対象に最適化することでネットワークアーキテクチャを探索する．進化的アルゴリズムや強化学習で最適化する手法 [1] [2] [3] [4] は NAS が登場した頃に提唱されたが，計算コストが大きく時間がかかる手法であるため工夫が必要となる．NAS を用いたアーキテクチャの探索はコストのかかるアプローチであるため，効率的なアーキテクチャ表現が求められており，Cell ベースの NAS 手法が主流な解決方法として登場した．Cell ベースの NAS 手法の中でも効率の良い構造探索を可能にした DARTS [5] や PC-DARTS [6] は現在でも主流な NAS 手法である．ここで，ネットワークアーキテクチャは，コンパクトなモジュールの集合体として定義され，各モジュールは有向非巡回グラフとして表現される．これにより，単純な探索空間内で大規模なアーキテクチャを作成することが容易となる．近年，Cell ベースの NAS アルゴリズム [7] による高精度なアーキテクチャが数多く発見されており，探索された高精度なアーキテクチャを再利用することが今後必要になる．

2.2 Graph Neural Network (GNN)

GNN [8] はグラフ構造のデータに深層学習の仕組みを組み込んだニューラルネットワークである．GNN の中にはグラフ分類やエッジ予測といったタスクがある．ノードとエッジで表現可能な構造であれば GNN の手法で扱うことが可能であるため，GNN は化学の分子構造 [9] [10] や物理の多体粒子シミュレーション [11] [12] などに応用されている．本研究では NAS の出力がグラフ構造で表現できるため，入力として扱えると考え GNN 手法を採用する．多くの GNN では Graph Convolution Network (GCN) [13] と呼ばれる畳み込み演算が使用される．特に GCN では**式（1）**を用いて，隣接行列や次数行列を扱うことで局所的なグラフの構造学習に成功している．

$$\mathbf{F_N}^{(l+1)} = h(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{F_N}^{(l)}\mathbf{W}^{(l)}) \qquad (1)$$

ここで，$\mathbf{F_N}$ が層の入出力でありノードの特徴量を表す．$\mathbf{A}$ はエッジの繋ぎ方を示す無向グラフの隣接行列と単位行列を足し合わせた行列，$\mathbf{W}$ は学習可能な重み行列，$\mathbf{D}$ はノードの関係を示す次数行列，h は活性化関数（Softmax と ReLU を使い分ける）となっている．その他の代表的な GNN の演算として Graph Attention Network (GAT) [14] が挙げられる．GAT は GCN に複数の Attention 機構を組み込むことにより，ノード間の重要度に対して重みづけを行い表現力が向上した演算子である．Variational Graph Auto-Encoders (VGAE) [15] は GNN の一手法であり，Encoder と Decoder を用いてグラフの特徴量抽出を行うことができる．画像を扱う VAE [16] を拡張することで，グラフの表現を可能とした手法となっている．詳しい Encoder と Decoder の詳細については次節で示す．

2.3 関連手法と提案手法の特徴

本研究に類似したような手法はいくつか存在する．提案手法は以下の 5 つの特徴を持っている:

(a) Alleviating computational cost and search time
(b) Building the latent feature space using VGAE
(c) Capable of handling continuous architectures
(d) Directly generated architectures
(e) Enabling information transfer in existing datasets

NAS は高い探索時間と計算コストが問題点として取り上げられることが多く，(a) のような大幅に計算コストを削減する研究 [17] [18] [19] が多く取り組まれている．また，アーキテクチャに対して特徴量を取り出し，潜在特徴量空間を構築する関連手法 [20] [21] [22] [23] [24] [25] [26] [27] [28] が数多く存在する．その中でも提案手法のように VGAE を使用し特徴量を取り出す手法 [21] [26] [27] [28] は一部であり，他と差別化できる点 (b) である．VGAE を用いることでグラフを介して細かいアーキテクチャの特徴量や性質を捉えることが可能となる．関連手法としては，精度等を予測する機構である predictor を構築する手法 [29] [30] [22] [23] [24] [31] [32] [33] もあるが，提案手法は直接潜在特徴量空間からアーキテクチャを生成 (d) する．また，提案手法は演算子の埋め込み表現も同時に学習することで離散的なアーキテクチャだけでなく，連続的なアーキテクチャも対象として扱うことが可能 (c) である．有望な機知情報を利用 (e) し性能を上げる点では，転移学習のアイディア [7] [34] [35] [36] も類似していると言える．転移学習の手法では，主にデータセット情報の転移に焦点を当てているが，提案手法ではアーキテクチャそのものの情報転移に特化している．関連手法の多くは主に精度向上を目指した手法であるのに対し，本研究では様々なデータセットに対する構造の調査を目的の 1 つとする．

単一の特徴を持つ手法だけでなく，複数の特徴を持つ手法も関連手法の中には存在する．しかし，提案手法は (a)〜(e) の 5 つの特徴を全て持っており，これは関連研究と大きく異なる部分である．5 つの特徴はアーキテクチャの生成において全て重要である．さらに本研究の特徴として，改良した VGAE を使うことで，NAS で探索したアーキテクチャの特徴を詳しく捉えることが可能となる．複数のデータセットにおいて高い精度を達成可能なアーキテクチャの構築を目指す部分も本研究の特徴である．

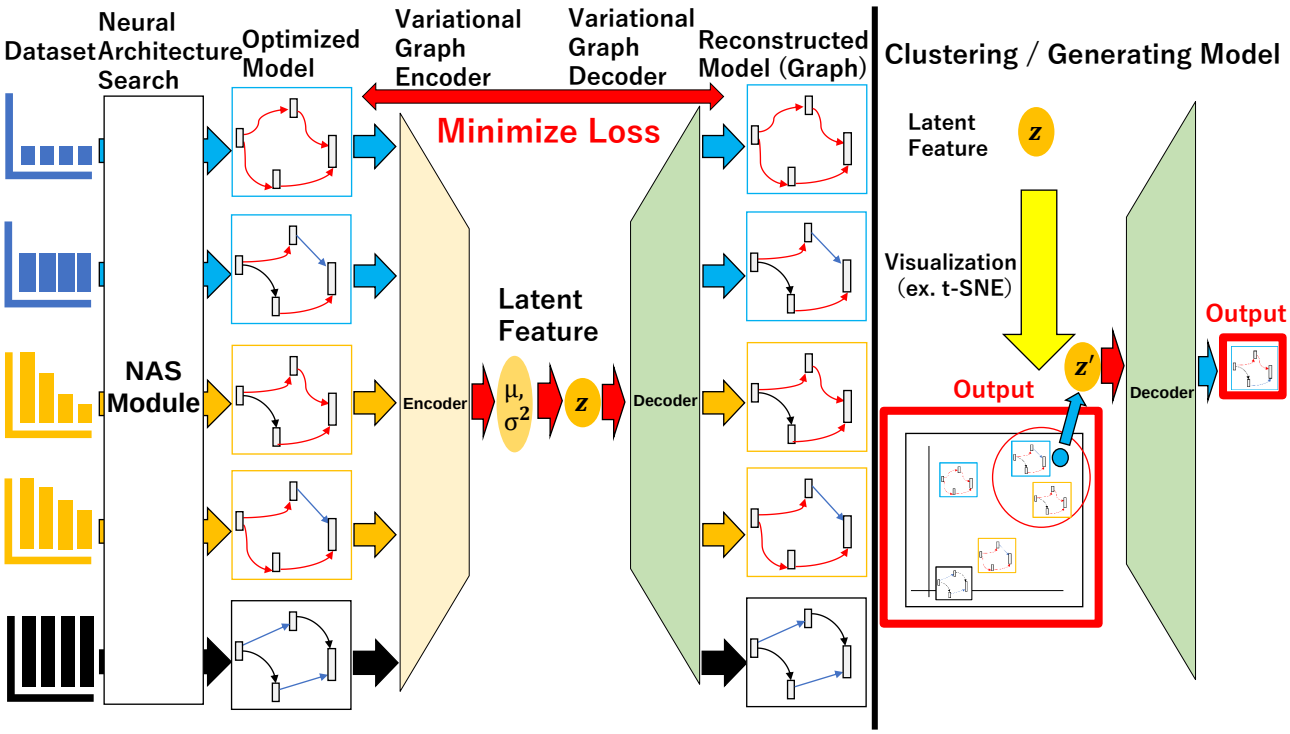


図 2: 提案手法：VGAE–NAS の概要

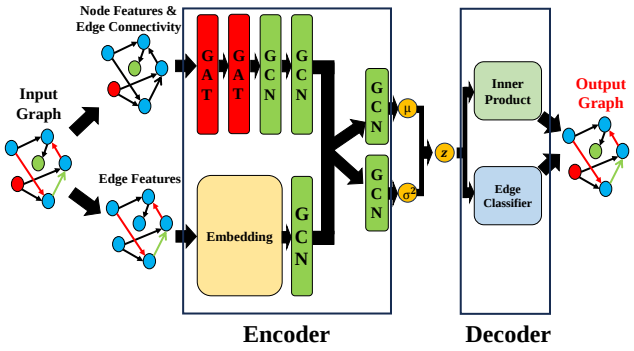


図 3: Encoder と Decoder の詳細．GAT: Graph Attention Network，GCN: Graph Convolution Network，Edge Classifier: ReLU と 2 つの全結合層

3 Variational Graph Auto-Encoders Neural Architecture Search

本研究では，NAS と VGAE を組み合わせて潜在特徴量を取り出し，アーキテクチャの特徴を獲得・可視化する手法である Variational Graph Auto-Encoders Neural Architecture Search (VGAE–NAS) を提案する．VGAE–NAS は潜在特徴量空間からアーキテクチャを生成する機構も含む．

3.1 提案手法の概要

VGAE–NAS の概要図を**図 2**，Encoder と Decoder の詳細を**図 3**, VGAE–NAS のアルゴリズムを **Alg. 1** に示す．

VGAE–NAS は NAS，VGAE, Generating Model の 3 種類のコンポーネントで構成されており，**Alg. 1** の手順（NAS：Line 1–5, VGAE：Line 6–15, Generating Model：Line 16–19）で実行される．**図 2** に示すように，潜在特徴量を獲得した後は可視化手法によりクラスタリングすることや新たなモデルを生成することが可能となる．VGAE–NAS では NAS モジュールとして勾配法の PC-DARTS を用いる．PC-DARTS は短時間で高精度な結果が出せるため，複数の構造を活用する提案手法にて採用した．VGAE–NAS で扱う NAS モジュールは PC-DARTS 以外の NAS 手法を扱うことも可能である．可視化手法には，次元削減アルゴリズムである t-SNE [37] や PCA [38] を採用した．

3.2 Encoder

Encoder は**式（2）**を用いて μ と $\log\sigma$ を算出する．提案手法の入力は NAS 手法で探索したネットワークアーキテクチャまたは NAS の探索空間からランダムサンプリング等で取り出したアーキテクチャである．

$$\mu = \mathrm{GNN}_\mu(\mathbf{F_N}, \mathbf{A}, \mathbf{F_E})$$
$$\log\sigma = \mathrm{GNN}_\sigma(\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}) \quad (2)$$

ここで，$\mathbf{F_N}$ はノードの特徴量，$\mathbf{A}$ はエッジの隣接行列，$\mathbf{F_E}$ はエッジの特徴量が含まれる．ノードの特徴量 $\mathbf{F_N}$ は NAS の入力ノード，中間ノード，出力ノードで構成されており，3 種類のノードの種類が特徴量として入力される．PC-DARTS の 2 種類の Cell を組み合わせる場合，入力ノードは 4 つ，中間ノードは 8 つ (ハイパパラメータ)，出力ノードは 2 つとなる．また，入力ではグラフのエッジ特徴量 $\mathbf{F_E}$ も同時に学習する．エッジ特徴量 $\mathbf{F_E}$ は NAS の探索空間に依存し，PC-DARTS では $|\mathbf{O}| = 8$ の演算候補 (Skip Connection，None，Average Pooling 3×3,

Algorithm 1 VGAE–NAS (1 試行)

Input: Dataset $\boldsymbol{D}$
Output: Latent Feature Space $\boldsymbol{z}$, New Architecture $\mathbf{A'_r}$

Prepare dataset $\boldsymbol{D}$
for Epoch = 1 to Specified Epochs in NAS **do**
  Update architecture and evaluate on $\boldsymbol{D}$
  Retain architecture $\mathbf{A_r}$ with best performance
end for
Convert A_r into graph format $\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}$
for Epoch = 1 to Specified Epochs in VGAE **do**
  Encoder $\leftarrow \mathbf{F_N}, \mathbf{A}, \mathbf{F_E}$
  $\mu_i, \sigma_i^2 \leftarrow$ *Encoder*
  $\mathbf{z} \leftarrow$ Sample from μ_i, σ_i^2
  Decoder $\leftarrow \mathbf{z}$
  $\mathbf{A'}, \mathbf{F'_E} \leftarrow$ *Decoder*
  $\boldsymbol{Loss}$ = Difference $(\mathbf{A}, \mathbf{F_E} \quad , \quad \mathbf{A'}, \mathbf{F'_E})$
  Update weights $\boldsymbol{w}$ of *Encoder* and *Decoder* by descending $\nabla_{\boldsymbol{w}}$ $\boldsymbol{Loss}$
end for
$\mu_i, \sigma_i^2 \leftarrow$ Trained *Encoder* with $\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}$
Latent Feature Space $\mathbf{z} \leftarrow$ Sample from μ_i, σ_i^2
Generate $\mathbf{z'}$ using $\mathbf{z}$ and the generating model algorithm (**Section 3.5**)
New Architecture $\mathbf{A'_r} \leftarrow$ Trained *Decoder* with $\mathbf{z'}$

Max Pooling 3×3, Separable Convolution 3×3, Separable Convolution 5×5, Dilated Convolution 3×3, Dilated Convolution 5×5) を使用する．各演算子がエッジ特徴量 $\mathbf{F_E}$ に対応しており，このエッジ特徴量 $\mathbf{F_E}$ に Embedding Layer を通す．訓練中はこの Embedding Layer も対象に学習していく．エッジ特徴量 $\mathbf{F_E}$ をもとに Embedding Layer を使用して埋め込むことによって演算子間での関係性を考慮するようになり，各演算子の情報がより潜在空間に反映されやすくなる．提案手法ではグラフ構造だけでなくエッジの演算を潜在空間で表現するため，元の VGAE にエッジ特徴量を埋め込むなど大きな変更を加えている．$\mathbf{F_N}$，$\mathbf{A}$ を入力とする GNN の詳細は GCN のみを用いた層と**図 3** のように GAT と GCN を組み合わせた層の 2 種類になっている．2 種類を組み合わせることで序盤の GAT Layer で近傍ノード間の具体的な関係性の情報を学習し，後半に GCN Layer を通すことで近傍ノード間の局所的なグラフ情報を学習するため表現力が向上する仕組みとなっている．

潜在特徴量 $\mathbf{z}$ は算出した μ と $\log\sigma$ で表現された正規分布に従うと仮定し，**式（3）**で $\mathbf{z}$ を出力する．

$$q(\mathbf{z}|\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}) = \prod_{i=1}^{N} q(\mathbf{z}_i|\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}) \quad \text{with} \quad q(\mathbf{z}_i|\mathbf{F_N}, \mathbf{A}, \mathbf{F_E}) = Norm(\mathbf{z}_i|\mu_i, \boldsymbol{\sigma}^{\mathbf{2}}{}_i) \tag{3}$$

ここで，N はノード数である．Encoder で正規分布のパラメータに落とし込み，各ノードに対する正規分布の積を取ることで全体の分布としている．

3.3 Decoder

Decoder では潜在特徴量 $\mathbf{z}$ から**式（4）**の Inner Product でグラフ構造を再構成する．ここで，$\hat{\mathbf{A}}$ は生成されたグラフの隣接行列である．$\hat{\mathbf{F}}_\mathbf{E}$ は生成されたエッジの特徴量であり，エッジの種類を表す．また，潜在特徴量 $\mathbf{z}$ から Edge Classifier を通すことによりエッジの種類を予測する．

Inner Product と Edge Classifier が含まれる Decoder で再構成したグラフと Encoder に入力する元のグラフが同一の形になるよう Loss を減らして最適化することで，対象となるグラフの潜在特徴量を取り出すことが可能となる．

$$p(\hat{\mathbf{A}}, \hat{\mathbf{F}}_\mathbf{E}|\mathbf{z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\hat{A}_{i,j}, \hat{F}_{Ei,j}|\mathbf{z}_i, \mathbf{z}_j) \quad p(\hat{A}_{i,j} = 1|\mathbf{z}_i, \mathbf{z}_j) = Sigmoid(\mathbf{z}_i^T, \mathbf{z}_j) \tag{4}$$

3.4 Loss Function

VGAE–NAS の Loss Function は，VGAE の Reconstruction Loss (Recon Loss)，Edge Class Loss と Kullback-Leibler Divergence Loss (KL Loss) の和で構成する．具体的には以下の**式（5）**で表す．W_{Edge} は Edge Class Loss に対する重要度，W_{KL} は KL Loss に対する重要度を調整するハイパパラメータ（基本的には 1.0）である．

$$Loss = L_{Recon} + W_{Edge} * L_{Edge} + W_{KL} * L_{KL} \tag{5}$$

ここで，L_{Recon} は VGAE の Recon Loss であり，Positive Loss と Negative Loss を用いた以下の**式（6）**で表す．L_{Recon} はグラフを指標にして構造を近づける役割である．W_P は Positive Loss に対する重要度，W_N は Negative Loss に対する重要度を調整するハイパパラメータである．

$$L_{Recon} = \sum_{(i,j)\in\mathbf{A}} W_P * A_{i,j}\log(\hat{A}_{i,j}) + W_N * (1 - A_{i,j})\log(1 - \hat{A}_{i,j}) \tag{6}$$

ここで，$A_{i,j}$ は真のグラフ，$\hat{A}_{i,j}$ は生成されたグラフの隣接行列の i 行 j 列要素である．L_{Edge} は，エッジの種類を予測する Edge Class Loss を表している．Edge Class Loss は，Cross-Entropy Loss を用いて計算する．L_{Edge} はエッジの種類を指標にして構造を近づける役割である．その他の Loss は元の VGAE でも使用されるが，L_{Edge} は NAS の細かい特徴を適用するため提案手法にて追加している．L_{KL} は，潜在変数の正規分布と事前分布の KL Divergence であり，以下の**式 (7)** で表される．

KL Divergence とは，二つの確率分布がどれだけ異なるかを測る指標である．L_{KL} は潜在空間を連続的にし，学習を安定させる役割をもつ．

$$L_{KL} = -\frac{1}{2}\sum_{i=1}^{D}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2) \tag{7}$$

式 (7) で，D は潜在変数の次元数，μ_i は潜在変数の平均，σ_i^2

は潜在変数の分散を表す（**式（2）**で算出）．これらの Loss の合計が提案手法の目的関数である Loss Function となる．

3.5 Generating Model

アーキテクチャは以下の手順に沿って生成する:

STEP 1: 良い結果を示すネットワークアーキテクチャの潜在特徴量 $\mathbf{z}$ を特定

STEP 2: 特定した潜在特徴量 $\mathbf{z}$ を他の潜在特徴量を考慮し各次元で [0,1] の範囲に正規化

STEP 3: 生成した方向ベクトルをスケーリングし，対応する STEP 2 の正規化した潜在特徴量に加算し，元の良い点の近傍にある新たな潜在特徴量を生成

STEP 4: STEP 3 で生成した新たな潜在特徴量を STEP 2 の逆変換を使用して逆正規化し，新たな潜在特徴量の学習済みの VGAE の Decoder に入力

STEP 5: 新たなネットワークアーキテクチャを生成

ここで，STEP 2–4 のスケーリングでは以下の**式（8）**のように係数をかけて小さくした新たな潜在特徴量 $\mathbf{z}'$ を生成する．Sgn は符号関数であり，入力値の符号を返すため，一様乱数 $\mathcal{U}(-0.5, 0.5)$ から各次元に対して -1 または 1 の値を持つ方向ベクトルが生成される．

$$\begin{aligned} \mathbf{z}_{\text{normalize}} &= \frac{\mathbf{z} - \mathbf{z}_{\min}}{\mathbf{z}_{\max} - \mathbf{z}_{\min}}, \quad \mathbf{d} = \text{Sgn}\left(\mathcal{U}(-0.5, 0.5)\right) \\ \mathbf{z}' &= (\mathbf{z}_{\text{normalize}} + \gamma \cdot \mathbf{d})(\mathbf{z}_{\max} - \mathbf{z}_{\min}) + \mathbf{z}_{\min} \end{aligned} \tag{8}$$

STEP 3 では元の潜在特徴量の近傍で新しい潜在特徴量の取得するために微小な摂動を表すため，一様乱数と γ を用いる．γ はハイパパラメータである．正規化した潜在特徴量空間の中で方向ベクトルの距離を一定にすることで，γ の値に応じた絶対距離を取る．生成した新たなネットワークアーキテクチャは元の良い点に対応するネットワークアーキテクチャと近い特性を持つと考える．次節の実験では生成したアーキテクチャを評価し，提案手法を検証する．

4 実　　験

本研究では，複数の異なるデータセットを用いた実験によって提案手法 VGAE–NAS の有用性を評価する．探索したアーキテクチャの潜在特徴量を獲得し可視化できるかを評価するために実験 A を，潜在特徴量空間から新たに生成したアーキテクチャが高い性能を持つか評価するために実験 B を行った．実験では 1 試行に対して NVIDIA Tesla V100 SXM2 16GB を 1 つ用いる (ImageNet では 4 つの GPU を用いて並列計算).

4.1 実験 A-I: CIFAR-10 と MNIST で探索したアーキテクチャの潜在特徴量可視化実験の設定

Experiment A では中身として異なる特性を持つ種類のデータセットに対して NAS で最適なアーキテクチャを探索し，それぞれのデータセットに適したネットワーク構造の潜在特徴量を可視化する．データセットとしては CIFAR10 [39] と MNIST [40] を用いる．CIFAR10 は深層学習や NAS の実験でよく用いられるデータセットであり，60000 枚の画像がデータセットに含まれている．MNIST は手書き数字の画像をまとめたデータセットである．各データセットや構造の特徴を細かく捉えることを目的とするため，2 種類の Cell を繋げた構造を入力するアーキテクチャとして用いる．NAS や VGAE の設定は**表 1** に示す．5 種類の Seed を用いるため，NAS は個別に 5 回探索しており，各 1 回は NAS アルゴリズムを最後まで実行し良い性能を示した構造を取り出している．VGAE の入力として，2 種類のデータセットと 5 回の試行で探索した 10 個のアーキテクチャを使用した．各アーキテクチャにはエッジの接続関係とエッジの種類が含まれており，ノードの数は固定されている．

表 1: 実験設定（↑ : NAS, ↓ : VGAE）

Number of Initial Channels	16
Number of Training Epochs	50 Epochs
Loss Function	CrossEntropyLoss
Initial Learning Rate	0.1
Learning Rate Decay	→ 0 (Cosine Schedule)
Learning Rate for Architecture	0.0003
Data Augmentation	Random Crop, Horizontal Flip
Optimizer for Training Parameter	SGD
Optimizer for Architecture Weights	Adam
Number of Training Epochs	500000 Epochs
Initial Learning Rate	0.001
Learning Rate Decay	→ 0.00025 (Cosine Scheduler)
Optimizer	Adam
Dimension of Latent Features	4
Seed	2024
W_P , W_N , W_{Edge} , W_{KL}	2.0 , 1.0 , 0.5 , 1.0

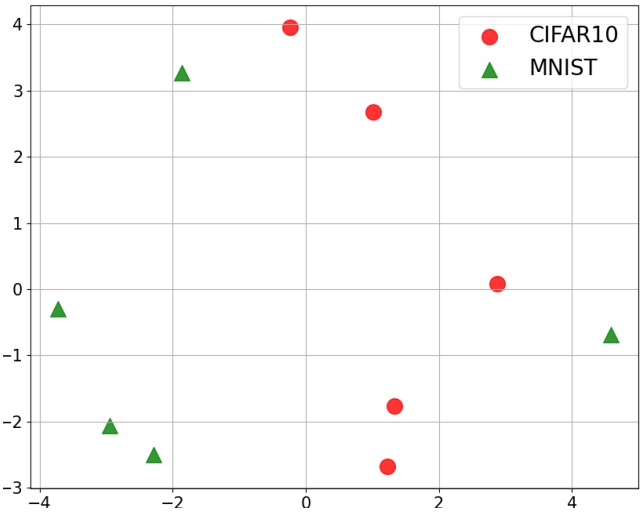


図 4: 提案手法で獲得した潜在特徴量の可視化図，軸：2 次元の PCA プロット上の座標軸，色：各データセットの種類

4.2 実験 A-II: CIFAR10 と MNIST の実験結果と考察

図 4 に提案手法で獲得した潜在特徴量の可視化図を示す．

表 2: USPS データセットを用いた実験 B-II の実験結果

Architecture	Model Size (Mean ± SD) (M)	Best Test Acc. (Max ± SD) (%)	Search Time (h)
CENAS [35]	0.57	89.94	7.4
ASED [41]	-	97.75	153.6
NASDA [42]	2.70	98.00	7.2
VGAE–NAS (Random Model)	2.15 ± 0.28	97.96 ± 0.10	0
VGAE–NAS (CIFAR10 Model, $\gamma = 0.1$)	3.32 ± 0.63	98.01 ± 0.11	0
VGAE–NAS (MNIST Model, $\gamma = 0.0$)	**3.37 ± 0.60**	**98.01 ± 0.09**	**0**
VGAE–NAS (MNIST Model, $\gamma = 0.1$)	**3.16 ± 0.66**	**98.11 ± 0.13**	**0**

図 4 は PCA によって次元削減したグラフであり，軸は潜在特徴量の多次元データを圧縮した結果の座標軸に相当する．**図 4** より，MNIST の右側の 1 点を除き，各データセットに対して潜在特徴量が座標軸として近傍に集まっている傾向が確認できる．CIFAR10 と MNIST では画像サイズやデータセットの特性（CIFAR10 は RGB 画像，MNIST はグレースケール）が違うため，NAS ではよりデータセットの特徴を捉えたアーキテクチャが探索できていると考える．また，CIFAR10 は画像サイズが 32 × 32 であるのに対して MNIST は 28 × 28 であり入力する画像サイズも異なる．従って，データセットに対して最適な近い潜在特徴量がそれぞれ獲得できていることから，CIFAR10 と MNIST で傾向が異なっていることが確認できる．

4.3 実験 B-I: 提案手法で生成したアーキテクチャの性能評価実験の設定

実験 B は提案手法で生成したアーキテクチャとランダムに生成したアーキテクチャを USPS データセット [43] と ImageNet データセット [44] で比較し，提案手法が有用なアーキテクチャを生成できるか，そしてランダムなアーキテクチャと比較して性能向上が見られるか評価することを目的とする．新たに使う USPS (United States Postal Service) は，手書き数字の画像データセットで，16 × 16 のグレースケール画像で内容は 0 から 9 までの手書きの数字が含まれる．USPS は MNIST からのドメイン適用としてもよく用いられる類似したデータセット [45] [46] [47] であるため，類似したデータセットに最適化されたアーキテクチャは同様の性能を持つかを実験 B-II により検証する．

本実験では，4 種類のアーキテクチャに対して再学習を行い性能を評価する．第 1 に，獲得した MNIST 周辺の潜在特徴量（$\gamma = 0.1$）から作ったアーキテクチャ．第 2 に CIFAR10 周辺の潜在特徴量（$\gamma = 0.1$）から作ったアーキテクチャ．第 3 に獲得した MNIST の潜在特徴量（$\gamma = 0$），第 4 にランダムな潜在特徴量から作ったアーキテクチャである．$\gamma = 0.1$ とした理由としては，$\gamma = 0.0$ のアーキテクチャの特徴を捉えつつ，一部異なるアーキテクチャであることから周辺の潜在特徴量を構築するパラメータとして設定する．MNIST と CIFAR10 はそれぞれ 5 構造の潜在特徴量が存在し，方向ベクトル生成時に発生する乱数を 4 回生成させアーキテクチャを構築した．4 種類で合計 50 構造 (5 Random Model, 5 MNIST Model ($\gamma = 0.0$), 20 CIFAR10 Model ($\gamma = 0.1$), and 20 MNIST Model ($\gamma = 0.1$)) を評価した．その他の実験設定は**表 1** に記載する．

また，ImageNet は画像識別でよく用いられる大規模なデータセットであり，ベースとしている PC-DARTS でも CIFAR10 で探索したアーキテクチャの Cell を転移して ImageNet で評価している．実験 B-III では CIFAR-10 周辺の潜在特徴量から作ったアーキテクチャ（20 構造）を使用して ImageNet で評価する．ImageNet は元の画像サイズが大きいことから 224 × 224 にクロップし，ランダムに左右反転を行う RandomHorizontalFlip とランダムに明るさやコントラストなどを変化させる ColorJitter を Data Augumentation として採用した．Initial Learning Rate を 0.5，再学習の Epochs を 250 と設定するなど，ImageNet 用のハイパパラメータを用いて 4 つの GPU で並列計算を行った．

4.4 実験 B-II: USPS の実験結果と考察

表 2 に各 NAS 手法の結果を示す．**表 2** より，精度が一番高いのは Best Test Acc.(Max) が 98.11 % を示した MNIST model ($\gamma = 0.1$) であることが確認できた．実験結果より，類似したデータセットで最適化された潜在特徴量を参考に構築したアーキテクチャが最も良い結果を示した．MNIST model が Random Model や CIFAR 10 Model より良い性能を示したのはもちろん，MNIST Model の中でも $\gamma = 0.0$ よりも $\gamma = 0.1$ のモデルが高精度であった．これは $\gamma = 0.0$ が MNIST で最適化されたアーキテクチャであり，USPS データセットと類似しつつも異なる内容であるためであると考える．新しいデータセットに対しても同様の手順で学習が可能となるため，NAS で 0 から探索しなくて済むため，探索時間を大幅に削減することができる．

4.5 実験 B-III: ImageNet の実験結果

表 3 に各 NAS 手法の結果を示す．**表 3** より，精度が一番高いのは Best Test Acc.(Max) が Top1 で 75.43%，Top5 で 92.60% を示した提案手法の CIFAR10 model であることが確認できた．さらに，上記の結果が探索時間をかけずに示した結果であることを考慮すると，非常に効率の良い性能を示せたと言える．提案手法は画像サイズが大きなデータセットや Cell を転移させた大きなアーキテクチャでも有効に働くことが実験結果から明らかになった．

表 3: ImageNet データセットを用いた実験 B-III の実験結果

Architecture	Model Size (Mean ± SD)	Best Test Acc. (Top1) (Max ± SD) (%)	Best Test Acc. (Top5) (Max ± SD) (%)	Search Time (h)
Inception-v1 [48]	6.60	69.80	89.90	0
NASNet-A [17]	5.30	74.00	91.60	43200
AmoebaNet-A [49]	5.10	74.50	92.00	75600
PNAS [50]	5.10	74.20	91.90	$\sim$ 5400
DARTS (CIFAR10 Model) [5]	4.70	73.30	91.30	96
PC-DARTS (CIFAR10 Model) [6]	5.30	74.90	92.20	2.4
iDARTS (CIFAR10 Model) [51]	5.10	74.80	92.10	4.8
VGAE–NAS (CIFAR10 Model, $\gamma = 0.1$)	**4.61 ± 0.75**	**75.43 ± 1.12**	**92.60 ± 0.64**	**0**

5 結　　論

本研究では，NAS と VGAE を組み合わせたフレームワーク VGAE–NAS を用いることで，NAS の探索結果の解釈性の向上を目指した．本研究の課題として，事前に用意した NAS 手法のアーキテクチャや探索空間にのみ対応している点がある．そのため，対応可能なアーキテクチャや探索空間を拡張して汎用性の高い手法に改良していく必要がある．提案手法の応用により，ネットワークアーキテクチャの構造や性能に関する新たな知見を提供できると考える．将来的には，提案手法を用いた複数のアーキテクチャの評価をさらに容易にするために，最適化された構造の精度以外の潜在的な特徴（例えば，モデルサイズや特定の演算子を多く含むアーキテクチャ）を抽出する予定である．また，データセット間の距離を測る手法 [52] [53] も存在するため，提案手法と組み合わせてデータセット間の距離を定量化し，潜在特徴量空間で最適化を行う手法へ改良したいと考える．この先進的なアプローチが，AutoML や Neural Architecture Search の分野を発展させ，ネットワーク設計と最適化の領域により深い洞察をもたらすことを期待する．

文　　献

[1] Barret Zoph and Quoc V Le, "Neural architecture search with reinforcement learning", *ICLR*, 2016.

[2] Yan Xu, Zhipeng Jia, Liang-Bo Wang, Yuqing Ai, Fang Zhang, Maode Lai, Eric I Chang, et al., "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features", *BMC bioinformatics*, vol. 18, no. 1, pp. 1–17, 2017.

[3] Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu, "Hierarchical representations for efficient architecture search", *arXiv preprint arXiv:1711.00436*, 2017.

[4] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin, "Large-scale evolution of image classifiers", in *International conference on machine learning*. PMLR, 2017, pp. 2902–2911.

[5] Hanxiao Liu, Karen Simonyan, and Yiming Yang, "Darts: Differentiable architecture search", *ICLR*, 2018.

[6] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong, "PC-DARTS: Partial channel connections for memory-efficient architecture search", in *ICLR*, 2020.

[7] Catherine Wong, Neil Houlsby, Yifeng Lu, and Andrea Gesmundo, "Transfer learning with neural automl", in *Neural Information Processing Systems*, 2018.

[8] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini, "The graph neural network model", *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[9] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou, "Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models", *Journal of cheminformatics*, vol. 13, no. 1, pp. 1–23, 2021.

[10] Xiao Wang, Sean T Flannery, and Daisuke Kihara, "Protein docking model evaluation by graph neural networks", *Frontiers in Molecular Biosciences*, vol. 8, pp. 647915, 2021.

[11] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia, "Learning to simulate complex physics with graph networks", in *International conference on machine learning*. PMLR, 2020, pp. 8459–8468.

[12] Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant, "Graph neural networks in particle physics", *Machine Learning: Science and Technology*, vol. 2, no. 2, pp. 021001, 2020.

[13] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks", *arXiv preprint arXiv:1609.02907*, 2016.

[14] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., "Graph attention networks", *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[15] Thomas N Kipf and Max Welling, "Variational graph auto-encoders", *arXiv preprint arXiv:1611.07308*, 2016.

[16] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes", *arXiv preprint arXiv:1312.6114*, 2013.

[17] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le, "Learning transferable architectures for scalable image recognition", *CVPR*, pp. 8697–8710, 2018.

[18] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley, "Neural architecture search without training", in *Proceedings of the 38th International Conference on Machine Learning*, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 7588–7598, PMLR.

[19] Arjun Krishnakumar, Colin White, Arber Zela, Renbo Tu, Mahmoud Safari, and Frank Hutter, "Nas-bench-suite-zero: Accelerating research on zero cost proxies", *Advances in Neural Information Processing Systems*, vol. 35, pp. 28037–28051, 2022.

[20] Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu, "Neural architecture optimization", *Advances in neural information processing systems*, vol. 31, 2018.

[21] David Friede, Jovita Lukasik, Heiner Stuckenschmidt, and Margret Keuper, "A variational-sequential graph autoencoder for neural architecture performance prediction", *arXiv preprint arXiv:1912.05317*, 2019.

[22] Shen Yan, Yu Zheng, Wei Ao, Xiao Zeng, and Mi Zhang, "Does unsupervised architecture representation learning help neural architecture search?", in *NeurIPS*, 2020.

[23] Jian Li, Yong Liu, Jiankun Liu, and Weiping Wang, "Neural architecture optimization with graph vae", *ArXiv*, vol. abs/2006.10310, 2020.

[24] Xuefei Ning, Yin Zheng, Tianchen Zhao, Yu Wang, and Huazhong Yang, "A generic graph-based neural architecture encoding scheme for predictor-based nas", in *European Conference on Computer Vision*. Springer, 2020, pp. 189–204.

[25] Hayeon Lee, Eunyoung Hyung, and Sung Ju Hwang, "Rapid neural architecture search by learning to generate graphs from datasets", in *International Conference on Learning Representations*, 2021.

[26] Michail Chatzianastasis, George Dasoulas, Georgios Siolas, and Michalis Vazirgiannis, "Graph-based neural architecture search with operation embeddings", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 393–402.

[27] Jovita Lukasik, David Friede, Arber Zela, Frank Hutter, and Margret Keuper, "Smooth variational graph embeddings for efficient neural architecture search", in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[28] Gabriela Suchopárová and Roman Neruda, "Graph embedding for neural architecture search with input-output information", 2022.

[29] Wei Wen, Hanxiao Liu, Hai Helen Li, Yiran Chen, Gabriel Bender, and Pieter-Jan Kindermans, "Neural predictor for neural architecture search", in *European Conference on Computer Vision*, 2019.

[30] Yaofo Chen, Yong Guo, Qi Chen, Minli Li, Wei Zeng, Yaowei Wang, and Mingkui Tan, "Contrastive neural architecture search with neural architecture comparators", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9502–9511.

[31] Colin White, Willie Neiswanger, and Yash Savani, "Bananas: Bayesian optimization with neural architectures for neural architecture search", in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 10293–10301.

[32] Jovita Lukasik, Steffen Jung, and Margret Keuper, "Learning where to look–generative nas is surprisingly efficient", in *European Conference on Computer Vision*. Springer, 2022, pp. 257–273.

[33] Andrea Agiollo and Andrea Omicini, "Gnn2gnn: Graph neural networks to generate neural networks", in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 32–42.

[34] Martin Wistuba, "Xfernas: Transfer neural architecture search", in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer, 2021, pp. 247–262.

[35] Mohan Singamsetti, Anmol Mahajan, and Matthew Guzdial, "Conceptual expansion neural architecture search (cenas)", *arXiv preprint arXiv:2110.03144*, 2021.

[36] Zhichao Lu, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, and Vishnu Naresh Boddeti, "Neural architecture transfer", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 9, pp. 2971–2989, sep 2021.

[37] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.", *Journal of machine learning research*, vol. 9, no. 11, 2008.

[38] Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[39] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images", 2009.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] Anton Muravev, Jenni Raitoharju, and Moncef Gabbouj, "Neural architecture search by estimation of network structure distributions", *IEEE Access*, vol. 9, pp. 15304–15319, 2021.

[42] Yichen Li and Xingchao Peng, "Network architecture search for domain adaptation", 2021.

[43] Jonathan J. Hull, "A database for handwritten text recognition research", *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[44] Olga Russakovsky, Jia Deng, et al., "Imagenet large scale visual recognition challenge", *IJCV*, vol. 115, pp. 211–252, 2015.

[45] Stefan Schrod, Jonas Lippl, Andreas Schäfer, and Michael Altenbuchinger, "Fact: Federated adversarial cross training", 2023.

[46] Jing Wang, Jiahong Chen, Jianzhe Lin, Leonid Sigal, and Clarence W de Silva, "Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by gaussian-guided latent alignment", *Pattern Recognition*, vol. 116, pp. 107943, 2021.

[47] Geoffrey French, Michal Mackiewicz, and Mark Fisher, "Self-ensembling for visual domain adaptation", *arXiv preprint arXiv:1706.05208*, 2017.

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[49] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le, "Regularized evolution for image classifier architecture search", in *Proceedings of the aaai conference on artificial intelligence*, 2019, vol. 33, pp. 4780–4789.

[50] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy, "Progressive neural architecture search", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[51] Huiqun Wang, Ruijie Yang, Di Huang, and Yunhong Wang, "idarts: Improving darts by node normalization and decorrelation discretization", *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[52] David Alvarez-Melis and Nicolo Fusi, "Geometric dataset distances via optimal transport", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 21428–21439, Curran Associates, Inc.

[53] Bjørn Magnus Mathisen, Agnar Aamodt, Kerstin Bach, and Helge Langseth, "Learning similarity measures from data", *Progress in Artificial Intelligence*, vol. 9, 10 2019.

モンテカルロ木探索を用いたグラフニューラル構造探索

佐々木勇和†

† 大阪大学大学院情報科学研究科
E-mail: †sasaki@ist.osaka-u.ac.jp

あらまし グラフ深層学習（GNN）は，多様な領域のデータサイエンスタスクにおける強力なツールであり，幅広い応用シナリオで GNN は活用される．しかし，多様なグラフに対して最適な GNN 構造を設計および選択するのは困難な作業である．そのため，人間の労力を削減するために，グラフニューラル構造探索（Graph NAS）が既存のコンポーネントを組み合わせた準最適な GNN 構造を探索するために使用されている．しかし，効率性と様々なグラフへの適応性を満たすグラフニューラル構造探索手法は現状存在しない．そこで，本論文では効率的かつ多様なグラフに適応可能なグラフニューラル構造探索手法を提案する．提案手法は (i) 様々なグラフに適応可能なシンプルな探索空間と (ii) モンテカルロ木探索を用いた探索アルゴリズムで構成されている．探索アルゴリズムでは，ニューラルモデルを使用しないモンテカルロ木探索により，GNN 構造を効率的に探索することが可能である．評価実験により，提案手法は最先端のグラフニューラル構造探索手法と比較してノード分類の正解率を最大 12.3 増加させ，実行時間を最大 88 %削減することを示す．

キーワード グラフ深層学習，ニューラル構造探索，モンテカルロ木探索

1 はじめに

グラフ深層学習 (GNN) はデータサイエンスタスクにおいて強力なツールであり，様々な応用にて用いられている．例えば，化学 [1,2] や物理 [3]，ソーシャル科学 [4]，神経科学 [5] などが代表的な例である．多様なグラフに対して GNN 構造が開発されている一方で，全てのグラフに最適な GNN 構造は存在しない．そのため，グラフの特性に合わせて新しい GNN 構造の開発や既存の GNN 構造を選択することが必要であるが，多くの労力を要する作業である．

GNN の設計や選択における人間の労力を削減するために，グラフニューラル構造探索が準最適な GNN 構造の探索に用いられている [6–8]．GNN 構造の増加とグラフ分析の需要の増大から，グラフニューラル構造探索の重要性は増している．

グラフ構造探索の課題. グラフ構造探索には二つの技術的な課題，(1) 探索空間と (2) 探索アルゴリズムがある．まず，探索空間はグラフ構造のパターンを決定し，探索空間内のコンポーネントの組み合わせにより GNN 構造は生成される．そのため，探索空間が適切に設計されていない場合，与えられたグラフに対して適切な GNN 構造が存在しない可能性がある．次に，探索アルゴリズムは GNN 構造の探索順を決定する．膨大なパターンの GNN 構造を全て探索することは困難であるため，いかに優先的に高精度な GNN 構造を発見するかが重要である．探索アルゴリズムが洗練されていない場合，探索に時間がかかることや適切な GNN 構造が発見できない可能性がある．

既存のグラフニューラル構造探索技術は精度と効率を改善するために様々なアプローチが用いられている．複雑なコンポーネントを含む探索空間からニューラルモデルに基づく探索アルゴリズムにて GNN 構造が探索されているため，探索時間が長いことおよび発見された GNN 構造の分析が困難であることがあげられる．加えて，探索空間は主としてホモフィリックグラフを対象として設計されており，ヘテロフィリックグラフへの適用性は低いという課題がある．

研究動機. グラフニューラル構造探索を研究者と利用者の二つの観点から再考する．まず，研究者においては，グラフニューラル構造探索は，多様なグラフに対する高精度なベースラインとして利用できること，および GNN 構造の設計における重要なコンポーネントを把握できる必要である．利用者においては，GNN の知識や，グラフの特性，重いハイパーパラメータチューニングを必要とせずに，効率的かつ簡単に準最適な GNN 構造を多様な構造に対して発見できる必要がある．これらの必要性から，シンプル，高速，かつ多様なグラフに適応可能なグラフニューラル構造探索技術が必要である．

筆者らの知る限り，これらの必要性を満たすグラフニューラル構造探索は存在しない．例えば，GraphGym [9] は GNN 構造の重要なコンポーネントを理解することを目的としており，シンプルな探索空間を設計している．しかし，探索アルゴリズムはランダム探索である点およびヘテロフィリックグラフは対象としていない．Auto-HeG [10] はヘテロフィリックグラフを対象としたグラフニューラル構造探索技術である．しかし，探索が非効率である点と探索空間に複雑なコンポーネントが含まれているため重要なコンポーネントの理解が難しい．また，ホモフィリックグラフとヘテロフィリックグラフにおける重要なコンポーネントを実践的に比較した研究は存在しない．そのため，ホモフィリックグラフとヘテロフィリックグラフの両方に適用可能かつ高速なグラフニューラル構造探索技術の開発が必要である．

貢献. 本研究はシンプル，高速，かつ多様なグラフに適用可能なグラフニューラル構造探索技術を提案する．提案するグラフ

ニューラル構造探索技術はシンプルな探索空間とニューラルモデルを伴わないモンテカルロ木探索を活用した探索アルゴリズムから構成される．まず，探索空間はホモフィリックグラフとヘテロフィリックグラフの両方で高精度となるように設計されている．MLP，活性化関数，ジャンピングナレッジなどの基礎的なコンポーネントのみが含まれており，最新の GNN 構成は含まれていない．そのため，シンプルな GNN 構造が構築され，容易に構造が理解可能である．次に，モンテカルロ木探索を活用した探索アルゴリズムはニューラルモデルを用いないため，学習の必要がなく高速である．探索済み GNN 構造の平均精度を用いて，次に検証する GNN 構造の優先度を決定する．

評価実験では，既存のグラフニューラル構造探索と比較して，最大で正解率が 12.3 向上し，88%の探索時間が減少することを示す．さらに，提案手法で発見した構造を分析することで，ホモフィリックグラフとヘテロフィリックグラフにおける GNN 構造の違いを分析可能であることを示す.

再現性. コードを公開する．[1]

2 事前準備

2.1 グラフ

本研究は節点属性およびラベル付きの無向グラフ $\mathcal{G} = (\mathbf{S}, \mathbf{X}, \mathbf{Y})$ を対象とする．$\mathbf{S} \in \{0,1\}^{n \times n}$ は隣接行列，$\mathbf{X} \in R^{n \times d}$ は節点属性の行列，および $\mathbf{Y} \in \{0,1\}^{n \times y}$ は節点のラベルを表す行列であり， $n, d,$ および y はそれぞれ節点数，各節点の属性数，およびラベルの種類数を表す．節点 v と u 間に枝が存在する場合, $\mathbf{S}_{(v,u)}$ と $\mathbf{S}_{(u,v)}$ は 1 となる．$\mathbf{Y}_v$ はワンホットベクトルであり，節点 v のラベルを表す.

2 つのグラフ種類，ホモフィリックグラフとヘテロフィリックグラフを定義する．枝ホモフィリー率 [11] は，$H(\mathcal{G}) = \frac{\sum_{0 \leqq v,u<n} \mathbf{S}_{(u,v)} \delta(\mathbf{Y}_v, \mathbf{Y}_u)}{\sum_{0 \leqq v,u<n} \mathbf{S}_{(u,v)}}$ で計算される．$\delta(\mathbf{Y}_v, \mathbf{Y}_u)$ は $\mathbf{Y}_v = \mathbf{Y}_u$ の場合，1 を返し，それ以外は 0 を返す関数である．直感的には，枝でつながってる節点が同じラベルをもつ場合に枝ホモフィリー率は増加する $H(\mathcal{G})$ が大きいグラフはホモフィリックグラフ，小さいグラフはヘテロフィリックグラフとなる．ホモフィリックグラフを想定した GNN はヘテロフィリックグラフにおいて有効ではないことが知られている．

2.2 ニューラル構造探索

ニューラル構造探索 (Neural Architecture Search, NAS) は与えられたデータセットに対して最も高精度を達成する深層学習モデルを見つけることを目的とする．モデルは，モデル構造 α とパラメータ $\mathbf{W}$ のペア，$(\alpha, \mathbf{W})$ で表現される．探索空間 $\mathcal{A}$ はモデル構造のパターンを定義する．ベストモデル α^* は下記で定義される．

$$\alpha^* = \arg\max_{\alpha \in \mathcal{A}} \mathrm{E}_{\mathrm{Val}}(\alpha, \mathbf{W}^*_\alpha). \quad (1)$$

$$\mathbf{W}^*_\alpha = \arg\min_{\mathbf{W}} \mathrm{L}_{\mathrm{Train}}(\alpha, \mathbf{W}). \quad (2)$$

1：https://github.com/OnizukaLab/AutoGNN_mcts.

$\mathrm{E}_{\mathrm{Val}}$ および $\mathrm{L}_{\mathrm{Train}}$ は開発データにおける評価指標の値と訓練データにおけるロス関数をそれぞれ表す.

本研究では，既存研究にならいノード分類タスクに取り組む (例えば，[7,8,12–17]). ノード分類タスクは，一部の節点にラベル付けされたグラフが与えられて，グラフ内の他のラベルを予測するタスクである.

3 関連研究

グラフニューラル構造探索. グラフニューラル構造探索手法は，与えられたグラフに対して最適な構造を見つけるために，探索空間と探索アルゴリズムの両面から研究されている [6–10,12–27]. 既存のグラフニューラル構造探索手法の探索空間と探索アルゴリズムを紹介する.

探索空間. 探索空間は二つのパート，構造空間とハイパーパラメータ空間に大別される．構造空間は GNN 構造のパターンを表し，例えば，統合関数や，活性化関数，層数などが含まれ，ハイパーパラメータ空間は，学習率やウェイトディケイなどのハイパーパラメータの値が含まれている．グラフニューラル構造探索手法では，固定のハイパーパラメータで構造空間のみが含まれる探索空間 [7,8,10,12–17] と構造とハイパーパラメータ空間の両方が含まれている探索空間 [6,9,12,18–20,22] が提案されている．どちらの探索空間がよいかは自明ではなく，例えば，ハイパーパラメータを探索した方がよいモデルが見つかる可能性もあるが，ハイパーパラメータを固定した方がより多くの GNN 構造を探索できるためよいモデルが見つかる可能性もある．本研究では，固定のハイパーパラメータで構造空間のみが含まれる探索空間に着目する．これは，グラフの種類に合わせてどのように GNN 構造が異なるかの分析することも目的の一つであるためである．一方で，提案手法はハイパーパラメータ空間も探索アルゴリズムの変更なし組み込むことができる.

構造空間は既存研究により異なるため，この構造空間の設計も技術的な貢献の一つである．探索アルゴリズムを提案していない論文もあることを明記しておく [9,27]. 多くの研究では，基礎的なコンポーネントを組合せて GNN 層を生成し，GNN 層をスキップコネクションなどを用いてつなげることで GNN 構造を構築する [9,13,14,18]. GraphNAS [6] は活性化関数や，アテンション関数，ヘッドの数などの関数を探索するが，層数は固定であるため，柔軟性が低い．GraphGym [9] は GNN 層の前後に MLP を含み，層数やスキップコネクションも構造空間に含まれている．AutoGraph [14] は，スキップコネクションに特化した探索空間であり，層数も自動で決定される．GAUSS [13] は，大規模グラフに対応するためにサンプリングに着目している．AutoGT [27] は，知識グラフにおけるグラフ Transformer に着目した構造空間である．GAP [12], DFG-NAS [15], PDNAS [17], NAS-Bench-Graph [21], および Auto-HeG [10] は，より最新の GNN 層をどのような任意の組合せで GNN 構造を構築するかに着目している．上記において，ヘテロフィリックグラフに着目したグラフニューラル構造探索手法は Auto-HeG [10] のみであり，様々なグラフへの対応性は未だ深く調査されていない.

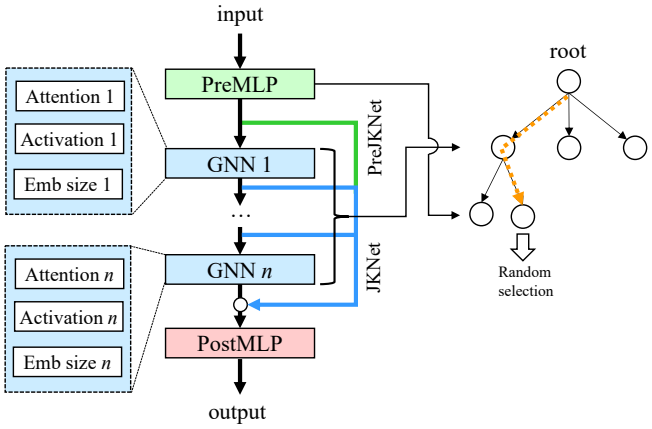


図 1: 探索空間とモンテカルロ木探索

探索アルゴリズム. 探索アルゴリズムは探索空間から効率的に最適な GNN モデルを見つけることを目的としている．大きく分けて，強化学習 [6–8, 28], 進化アルゴリズム [8, 14, 22, 27], および微分可能探索 [10, 12, 17–20, 25] の 3 つに大別できる．モンテカルロ木探索は強化学習に分類することができる．既存の強化学習ベースの探索アルゴリズムは GNN モデルの予想される性能を最大化するようにニューラルモデルが活用されている．進化アルゴリズムベースの探索アルゴリズムは，探索済みのモデルの多くを引き継ぎ，一部を探索したモデルの探索を繰り返す．微分可能探索ベースの探索アルゴリズムは，探索空間を学習パラメータとして，徐々に最適なモデルに近づくようにモデル構造が選ばれている．EGNAS [28] はモンテカルロ木探索と深層強化学習を用いた探索アルゴリズムを採用している．この手法はモンテカルロ木探索により探索空間を分割し，その後深層強化学習により構造を探索している．

4 提 案 手 法

本章では，グラフニューラル構造探索手法を説明する．提案手法は，ホモフィリックグラフとヘテロフィリックグラフの両方に対応可能かつ高速な探索が可能である．以下では,，シンプルかつ効果的な探索空間と探索アルゴリズムについて紹介する．

4.1 探 索 空 間

探索空間は GNN 構造のパターンを定義する．もし探索空間の設計が稚拙な場合，多様なグラフに対する効果的な GNN 構造は含まれないため，探索空間は極めて重要である．我々の探索空間は Micro 構造と Macro 構造の二つの観点が含まれている．Micro 構造は独立した GNN 層の内部を決定し，Macro 構造は GNN 層間の関係性を決定する．図 1 の左図が GNN 構造の探索空間を表す．直感的には， GNN 1 – GNN n が Micro 構造を示し，全体の構造が Macro 構造を示す．

設計方針. 最新の複雑な技術ではなく，基礎的なコンポーネントのみが含まれる探索空間を設計する．これは，複雑ではない GNN 構造のみが候補となるため，与えられたグラフに対してどのようなコンポーネントが重要なのかを理解することに役立つ．ヘテロフィリックグラフに対応するために，有効と知られている (i) 自身の属性の強調および (ii) Jumping knowledge が含まれている [11, 29–32].

探索空間は構造パラメータのセットとして表現される．スーパーネットが NAS では頻繁に用いられるが，グラフニューラル構造探索では一般的ではない．構造パラメータのセットは，重要なパラメータの分析に効果的である．これは，生成された GNN 構造間の比較が容易であり，一方でスーパーネットは複雑な構造が生成されるため比較が容易ではないためである．

Micro 構造. 多層のメッセージパッシングモデルは標準的な GNN 構造である．このモデルはノード特徴量を層毎に畳み込みをしながら学習する．正式には，l 番目の GNN 層は下記で定義される．

$$\mathbf{z}_u^{(l)} = \sigma\left(\sum_{v\in\mathcal{N}_u} e_{(u,v)}\mathbf{W}^{(l)}\mathbf{z}_v^{(l-1)}\right). \tag{3}$$

$\mathbf{z}_u^{(l)}$ は l 層目の節点 u のエンベッディング，$\mathbf{W}^{(l)} \in R^{|\mathbf{z}^{(l)}|\times|\mathbf{z}^{(l-1)}|}$ は学習パラメータ，および $\mathcal{N}_u$ は節点 u の隣接節点を表す．ここでは，$\mathbf{z}_u^{(0)}$ は節点 u の属性である．$e_{(u,v)}$ と σ は節点 u と v 間のアテンションと活性化関数をそれぞれ表す．

Micro 構造はアテンション，活性化関数，エンベッディングサイズの 3 つの設計次元がある，これらは GNN 層における標準的なコンポーネントである．モデル特有のコンポーネント，例えば GAT におけるヘッド数など，は GNN 構造のパターン数削減のために含まれていない．

Macro 構造. Macro 構造は GNN 層のつながりを全体のグラフ深層学習としてどのように構成するかを決定する．最も簡単な方法は GNN 層を単に重ねる場合となる．提案手法はそれぞれの GNN 層の出力をつなげる jumping knowledge network (*JKNet*) を用いる．また，GNN 層の前後に MLP 層を用いる．我々はそれぞれ *preMLP* と *postMLP* と呼ぶ．加えて，preMLP の出力を最終層の出力の入力とする．*PreJKNet* と呼ぶ．

この Macro 構造の利点はヘテロフィリックグラフに対して高精度なモデルの構築が期待できる点である．ヘテロフィリックグラフにおいて，自身の節点属性は隣接節点からの特徴量の統合に比べて効果的であることが多い [9, 11, 33–36]．PreMLP と PreJKNet の組合せはヘテロフィリックグラフにおいて効果的であるが，既存のグラフニューラル構造探索ではまだ用いられていない．

構造パラメータ. 探索空間の構造パラメータについて述べる．表 1 に構造パラメータをまとめる．提案手法はそれぞれの GNN 層に異なる構造パラメータを選択できる．例えば，層数が 2 の場合，第一層のアテンションが GCN，第二層のアテンションが GAT のような構造が可能である．この構造パラメータでは，探索空間には 2 億以上の構造パターンが含まれている．ここで，構造パラメータには依存関係がある．例えば，JKNet が Max の場合，全てのエンベッディングサイズが同じ必要がある．エンベッディングサイズはグラフサイズと GPU メモリに合わせて設定することができる．

表 1: 構造パラメータ．d_u と || はそれぞれ節点 u の次数と連結関数を示す．

Function	Parameters
The number of GNN layers	1, 2, 3
Attention	Constant: $e_{(u,v)} = 1$ GCN: $e_{(u,v)} = 1/\sqrt{d_u d_v}$ GAT: $e_{(u,v)} = leakyReLU(\mathbf{W}_l\mathbf{z}_u \| \mathbf{W}_r\mathbf{z}_v)$
Activation	None, Relu, Sigmoid, Tanh
The embedding size of of GNN layer	16, 32, 64, 128, 256, y
JKNet	None, concat, max
PreJKNet	None, use
PreMLP	None, use
The embedding size of preMLP	16, 32, 64, 128, 256
The number of layers in postMLP	0, 1, 2
The hidden size of postMLP	64, 128, 256

4.2 探索アルゴリズム

提案手法における探索アルゴリズムの特徴はモンテカルロ木探索を利用し，ニューラルモデルを利用しないことである．これにより，高速かつスケーラブルな探索可能となる．このアプローチはシンプルである一方で，GNN 構造の決定に有効である．

図 1 の右図はモンテカルロ木探索による構造パラメータの決定方法を表している．モンテカルロ木を辿ることで，構造パラメータを決定する．その際，それぞれの木の深さにコンポーネントを割り当て，モンテカルロ木の節点に構造パラメータを割り当てる．例えば，この図では層数が深さ 1 に割り当てられている．

設計方針．探索アルゴリズムでは，微分可能アルゴリズム (例，[37]) や深層強化学習 (例，[6]) などのニューラルモデルが頻繁に用いられる．このようなニューラルモデルを用いるアプローチは高精度な GNN 構造を選択するのに有効であることが多い，一方で学習コストが高く，非効率であることが多い．そこで，ニューラルモデルを用いないモンテカルロ木探索を用いることで，高速かつスケーラブルな探索を可能とする．加えて，構造パラメータ選択やそのパラメータを用いた場合の平均精度がわかるため，モデルの分析にも有効である．

構造選択. 下記の手順により GNN 構造を選択する．(1) 葉ノードを選択する， (2) ルートから葉ノードの経路上にある節点に割り当てられている構造パラメータを決定する，そして (3) それ以外の構造パラメータをランダムに決定する．

葉ノードの決定において，UCB (Upper Confidence Bound) [38] を拡張したスコアを節点毎に与える．節点 i の ucb スコアは下記で計算される．

$$ucb(i) = \frac{\sum_{(\alpha,\mathbf{W})\in\mathcal{M}_i} \mathrm{E}_{\mathrm{Val}}(\alpha,\mathbf{W})}{m_i} + c\sqrt{\frac{\ln M}{m_i}} \qquad (4)$$

$\mathcal{M}_i$ は評価済みのモデル，m_i は i が利用された回数，および M は探索数である．c は探索と活用のバランスを調整するハイパーパラメータであり，c が大きいほどより未知のものを探索する．m_i が増加すると，$usb(i)$ が低下するため，モンテカルロ木探索は局所探索に陥ることなく探索を行える．

木構造の更新. モデル構造を選択した後は，選択されたモデル構造の性能を確認するために学習と検証が行われる．その後，ルートから葉ノードの経路上にある節点の ucb を更新する．節点 i が θ 以上選ばれたら，新しい子節点を作成する．これらの子ノードは $m = 0$ (つまり, $ucb = \inf$) であるため，次の探索時には優先的に選択される

この木構造には平均性能や，モデルが選択された回数，平均学習時間も格納するため，その構造パラメータが精度と効率性に与える影響も調査できる．

ハイパーパラメータ. この探索アルゴリズムには c と θ という二つのハイパーパラメータがある．既存研究にならい，$c = \sqrt{2}$ とする [38]．加えて，これらのハイパーパラメータの影響が大きくないことを本実験にて示す．そのため，提案手法はハイパーパラメータチューニングの必要はない．

コンポーネント順. 高精度な GNN 構造を効果的に発見するためにモンテカルロ木のコンポーネント順を決定する．小さい木の深さに割り当てられたコンポーネントは早めに決定されるため，その構造パラメータの精度への影響が早めに精査される．そのため，重要なコンポーネントを小さい木の深さに割り当てる必要がある．

提案手法のモンテカルロ木では，GNN の層数, preMLP の有無, preJKNet の有無, および JKNet の種類の順に割り当てる．これらは，性能に大きく影響することが知られている [11,32–36]. その後，最初の GNN 層の活性関数とアテンション関数，PreMLP のエンベッディングサイズ，PostMLP の中間層の数，最初の GNN 層のエンベッディングサイズを決定する．最後に，性能の影響を予測するのが難しいその他のコンポーネント，つまり 2 層目以降の活性化関数，アテンション関数，エンベッディングサイズを決定する．

時間計算量. GNN 構造の選択では，モンテカルロ木を辿り，その後ランダムに構造のパラメータを決定する．この処理は，$O(|\mathbb{F}||\mathbb{A}|)$ となる．ここで，$|\mathbb{F}|$ と $|\mathbb{A}|$ はそれぞれコンポーネントと構造パラメータの数であるが，これらは定数と考えることができるため $O(1)$ である．木構造の更新においても，モンテカルロ木を辿るだけであるため $O(1)$ である．そのため，GNN 構造を決定するための探索アルゴリズムは $O(1)$ となる．

疑似コード. アルゴリズム 1 に提案手法の疑似コードを示す．このアルゴリズムは GNN 構造 α の選択と訓練を繰り返す（3–6 行目）．モデルの訓練が完了後に，モンテカルロ木を更新する（7–9 行目）．L 回のモデル探索を繰り返す（2 行目）．

5 評 価 実 験

本章では，ノード分類タスクにおける提案手法の有効性を検証する．本実験では，提案手法の精度と効率性を評価し，発見した GNN 構造の分析が設計に役立つかを示す．

提案手法は Python3 で実装し，AWS の p3.x2large インスタンス (NVIDIA V100 Tensor Core GPU および 16 GB の GPU メモリを搭載) を用いる．

Algorithm 1: Ours

input : $\mathcal{G}$, L
output: α^*, $\mathbf{W}^*$, MCT

Initialize MCT, $best$;
for $1, \ldots, L$ **do**
  $i \leftarrow$ leaf node with the maximum ucs;
  $\alpha \leftarrow$ functions of n_i with random parameters;
  Initialize $\mathbf{W}$ of α;
  Train model $(\alpha, \mathbf{W})$ on $\mathcal{G}$;
  Update MCT;
  if $m_i \geqq \theta$ **then**
    Generate child nodes of i;
  if $best < \mathrm{E}_{\mathrm{VAL}}(\alpha, \mathbf{W})$ **then**
    $\alpha^*, \mathbf{W}^* \leftarrow \alpha, \mathbf{W}$;
    $best \leftarrow \mathrm{E}_{\mathrm{VAL}}(\alpha, \mathbf{W})$;
return α^*, $\mathbf{W}^*$, MCT;
end procedure

5.1 実験環境

データセット. GNN のタスクでよく用いられている 12 個のグラフを用いる．これらのグラフは，様々な応用から作成され，異なる枝ホモフィリー率を有する．表 2 にグラフの統計を示す．

表 2: データセットの概要

Dataset	# nodes	# edges	# features	# labels	$H(\mathcal{G})$
Cora	2,708	5,429	1,433	7	0.81
CiteSeer	3,327	4,732	3,703	6	0.73
Amz-P	7,650	238,162	745	8	0.82
Amz-C	13,752	491,722	767	10	0.77
Co-CS	18,333	163,788	6,805	15	0.79
PubMed	19,717	44,338	500	3	0.81
Cornell	195	304	1,703	5	0.13
Wisconsin	265	530	1,703	5	0.20
Chameleon	2,277	36,101	2,325	5	0.23
Squirrel	5,201	217,073	2,089	5	0.23
Actor	7,600	30,019	932	5	0.23
Penn94	38,815	2,498,498	4,772	2	0.53

比較手法. コードが公開されている 4 つのグラフニューラル構造探索 GraphNAS [6], GraphGym [9], DFG-NAS [15], および Auto-HeG [10] を比較手法として用いる. GraphNAS, DFG-NAS, および Auto-HeG はそれぞれ深層強化学習，進化アルゴリズム，および微分可能探索を探索アルゴリズムに用いている．GraphGym は探索アルゴリズムが提案されていないため，既存手法に合わせて一様サンプリングを探索アルゴリズムとして用いる [15].

ハイパーパラメータ. ニューラル構造探索のモデル探索数は 1000 とする [39]. 提案手法における θ と c は，それぞれ 10 および $\sqrt{2}$ とする．GraphNAS, GraphGym, DFG-NAS, および Auto-HeG においては，GitHub に記載されているデフォルトパラメータを用いる．また，公平な評価のためにハイパーパラメータチューニングは行わないが，GraphGym と Auto-HeG は探索空間にハイパーパラメータが含まれている．訓練，開発，および検証データの割合は，それぞれ 0.6/0.2/0.2 とする．

5.2 性能評価

精度比較. 表 3 および表 4 にそれぞれ AUC と正解率の実験結果を示す．各手法において正解率と AUC の最大化のために異なる GNN モデルを構築している．

提案手法はヘテロフィリックグラフとホモフィリックグラフの両方で高い平均正解率と AUC を達成している．提案手法は各グラフで頻繁に最高性能ではないが，最低性能の場合であっても，他の手法との差は非常に小さい．ヘテロフィリックグラフにおける GraphNAS とホモフィリックグラフにおける Auto-HeG を除く既存のグラフニューラル構造探索手法は平均的に高い性能を達成している．ホモフィリックグラフとヘテロフィリックグラフ間の GraphNAS の性能差は，ホモフィリックグラフ用に設計された探索空間がヘテロフィリックグラフには適していない可能性を示している．GraphGym と DFG-NAS は一部のグラフにおいて最高性能を達成している一方で，ヘテロフィリックグラフのいくつかに対しては低い性能となっている，例えば，GraphGym の Chameleon と Squirrel や DFG-NAS の Penn94 である．Auto-Heg は主にヘテロフィリックグラフに焦点を当てているため，ホモフィリックグラフに対するは性能が低い．また，Auto-Heg は多数の枝を持つグラフにおいて複雑な GNN 構造を構築するためにスケーラブルではないことがわかる．

結果として，提案手法は 12 のグラフの中で最も高い平均 AUC と正解率を達成しているため，提案手法がホモフィリックグラフとヘテロフィリックグラフの両方に高い適応性を持っていることが確認できる．

探索時間の比較. 図 2 にグラフ構造探索の実行時間を示す．この探索時間にはモデルの訓練時間も含まれている．実験結果より Ours, GraphGym, および DFG-NAS が効率的であることがわかる．これらの手法はニューラルモデルを構造選択に利用していない．提案手法はヘテロフィリックグラフで高速，一方 GraphGym と DFG-NAS はホモフィリックグラフで高速であることが多い．GraphNAS と Auto-HeG は探索アルゴリズムに訓練時間が含まれるため実行時間が低速である．この結果より，提案手法は高速であることを確認できる．

モデルサイズ比較. 表 5 に出力したモデルのパラメータ数を示す．提案手法が出力したモデルはモデルサイズが小さいことがわかる．特に，ヘテロフィリックグラフにおいて全てのグラフで最小のモデルを出力している．この結果より，提案手法のモデルは高精度かつ小モデルであることがわかる．

5.3 ハイパーパラメータセンシティビティ

提案手法は c と θ の二つのハイパーパラメータがある．図 3 はそれぞれのハイパーパラメータの AUC への影響を示している．これらの実験結果より，c と θ は AUC に大きく影響しないことがわかる．そのため，提案手法はハイパーパラメータチューニングを実施することなくデフォルトパラメータで十分であることを示しており，計算コストの低下と使いやすさが向上しているといえる．

表 3: AUC の比較．青と赤のハイライトは各列におけるそれぞれ最大と最小を表す．DNF は 48 時間以内で実行できなかったもの，OOM は ‘Out of Memory.’ を表す．

	Heterophilic							Homophilic						
	Cornell	**Wisconsin**	**Chameleon**	**Squirrel**	**Actor**	**Penn94**	Average	**Cora**	**CiteSeer**	**Amz-P**	**Amz-C**	**Co-CS**	**PubMed**	Average
GraphNAS	69.2±0.3	74.7±0.2	57.5±0.1	36.7±0.7	33.2±0.0	DNF	54.3	87.1±0.0	76.0±0.0	95.8±0.0	**91.9**±0.0	94.2±0.0	87.8±0.0	88.8
GraphGym	74.9±5.9	87.2±6.3	54.0±2.0	37.2±1.3	**38.3**±1.0	**85.6**±0.4	62.9	82.4±1.6	74.2±1.5	**96.2**±0.3	**91.9**±0.3	95.3±0.4	**89.7**±0.7	88.3
DFG-NAS	**83.1**±0.4	**91.7**±0.1	67.0±0.1	47.2±0.0	38.1±0.0	76.1±0.0	67.2	**88.5**±0.0	**77.1**±0.0	95.2±0.0	88.0±0.0	**95.8**±0.0	87.6±0.0	88.7
Auto-HeG	74.9±14.2	86.8±39.9	OOM	OOM	38.0±1.0	OOM	66.6	75.5±19.9	70.5±3.5	OOM	OOM	OOM	88.3±0.1	78.1
Ours	72.3±0.3	84.2±0.1	**69.9**±0.0	**59.5**±0.0	37.9±0.0	82.5±0.0	**67.7**	**88.5**±0.0	74.7±0.0	95.6±0.0	91.2±0.0	95.6±0.0	89.3±0.0	**89.2**

表 4: 正解率の比較．青と赤のハイライトは各列におけるそれぞれ最大と最小を表す．DNF は 48 時間以内で実行できなかったもの，OOM は ‘Out of Memory.’ を表す．

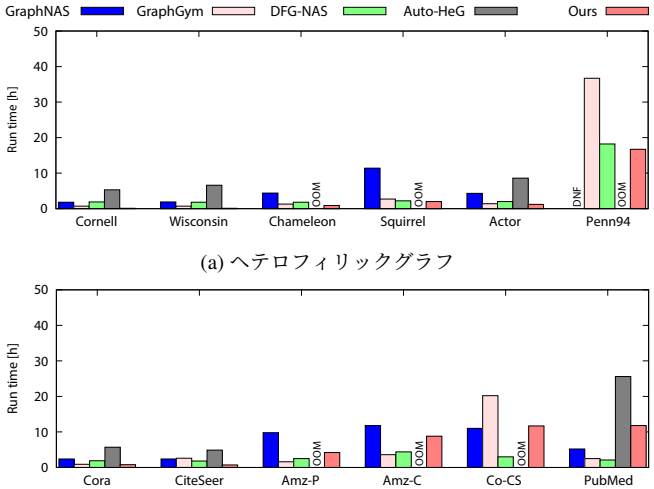
	Heterophilic							Homophilic						
	Cornell	**Wisconsin**	**Chameleon**	**Squirrel**	**Actor**	**Penn94**	Average	**Cora**	**CiteSeer**	**Amz-P**	**Amz-C**	**Co-CS**	**PubMed**	Average
GraphNAS	86.1±0.2	89.8±0.1	83.9±0.0	73.8±0.0	65.1±0.0	DNF	79.7	97.2±0.0	92.1±0.0	99.5±0.0	**99.2**±0.0	99.7±0.0	97.0±0.0	97.5
GraphGym	93.0±2.1	95.4±3.8	83.4±1.5	71.4±1.0	**72.4**±0.6	**93.4**±0.2	84.8	96.7±0.5	92.6±0.6	**99.6**±0.2	**99.2**±0.1	99.8±0.0	**97.5**±0.3	97.6
DFG-NAS	**94.5**±0.0	**95.8**±0.0	87.3±0.0	77.2±0.0	70.0±0.0	84.8±0.0	84.9	**98.2**±0.0	**93.6**±0.0	99.4±0.0	98.9±0.0	**99.9**±0.0	96.8±0.0	**97.8**
Auto-HeG	89.3±0.3	**95.8**±0.1	OOM	OOM	70.0±0.0	OOM	85.0	92.9±0.0	91.0±0.0	OOM	OOM	OOM	96.8±0.0	93.6
Ours	91.0±0.1	93.6±0.2	**87.6**±0.0	**81.4**±0.0	70.1±0.0	91.9±0.0	**85.9**	98.0±0.0	93.4±0.0	99.1±0.0	98.8±0.0	99.7±0.0	97.3±0.0	97.7

表 5: モデルサイズ．青のハイライトは各列におけるそれぞれ最小値を表す．

	Heterophilic						Homophilic					
	Cornell	**Wisconsin**	**Chameleon**	**Squirrel**	**Actor**	**Penn94**	**Cora**	**CiteSeer**	**Amz-P**	**Amz-C**	**Co-CS**	**PubMed**
GraphNAS	1288.4K	2214.7K	1806.0K	1584.9K	1470.4K	OOM	1590.7K	1213.1K	224.6K	239.7K	1346.3K	320.1K
GraphGym	491.0K	491.0K	572.4K	541.5K	390.0K	891.9K	456.1K	753.2K	366.3K	369.7K	1161.6K	332.9K
DFG-NAS	430.1K	459.9K	314.9K	284.7K	272.0K	611.2K	191.1K	1699.8K	96.5K	99.6K	873.1K	94.2K
Auto-HeG	357.7K	442.5K	OOM	OOM	529.3K	OOM	667.3K	930.2K	OOM	OOM	OOM	69.7K
Ours	136.5K	377.0K	100.5K	241.6K	53.3K	102.2K	317.1K	545.6K	143.5K	111.0K	1812.8K	69.6K

表 6: アブレーションスタディ．

	GraphGym	**Uniform**	**Max**	**Ours**
Homophilic	97.8	97.3	97.3	97.7
Heterophilic	84.9	85.5	85.4	85.9



(a) ヘテロフィリックグラフ

(b) ホモフィリックグラフ

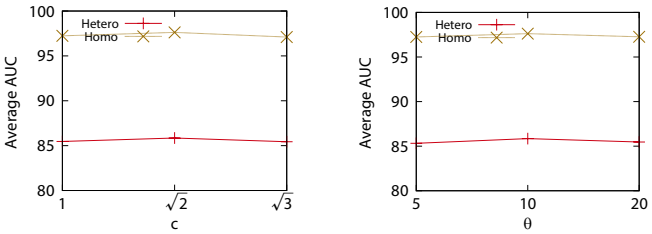
図 2: 実行時間 [h].

5.4 アブレーションスタディ

表 6 は GraphGym と提案手法に加えて, 提案手法の探索アルゴリズムを一様サンプルに変更した手法 (Uniform) と最大性能であるモデル構造を優先的に探索した手法 (Max) の実験結果を示す. GraphGym と Uniform は同じ探索アルゴリズムを用いているため探索空間を比較することができ，Uniform，Max，および提案手法を比べることで探索アルゴリズム間の比較ができる．まず，Uniform はヘテロフィリックグラフにおいて GraphGym より性能が高く，ホモフィリックグラフにおいて GraphGym より性能が低い．これは提案手法の探索空間がヘテロフィリックグラフにより適していることを示している．次に，提案手法は Uniform と Max よりも性能が良いため，モンテカルロ木探索が優先的に良いモデルを探索できていることがわかる．

5.5 GNN 構造分析

提案手法が GNN 構造の分析に効果的であることを示す．ここでは，特にホモフィリックグラフとヘテロフィリックグラフの違いについて着目する．既存のグラフニューラル構造探索では，コンポーネントが複雑であることや両方のグラフタイプにて高精度な構造を発見できないため，構造の分析では効果的ではない．



(a) c (b) θ

図 3: ハイパーパラメータ c と θ の影響

選択された GNN 構造の例. 図 4 に Citeseer, Amz-c, Cornell, および Squirrel で選択されて GNN 構造を示す．これらの例から選択された GNN 構造が大きく異なることがわかる．まず，ホモフィリックグラフにおける GNN 構造（Citeseer と Amz-c）は単純であり，ヘテロフィリックグラフにおける GNN 構造（Cornell と Squirrel）は複雑であることがわかる．特に，Citeseer における GNN 構造は極めて単純であり，ノードラベルが隣接節点情報から推測できることがわかる．一方で，Cornell と Squirrel では，より広範囲な節点情報と自身の属性情報が重要であることがわかる．これらの例から，ヘテロフィリックグラフとホモフィリックグラフでは最適なモデル構造が大きく異なることがわかる．

選択された構造パラメータ分析. 表 7 に本実験にて選択された構造パラメータの割合を示す．まず，ヘテロフィリックグラフではより多くの層を用いることがわかる．GNN 構造の層数は一般的に 2 層であるが，ヘテロフィリックグラフでは最適ではないことが多いことがわかる．次に，ハイパボリックタンジェントが活性化関数で最も多く選ばれている．一方で，GNN では ReLU を活性化関数として用いることが一般的であるが，最適ではないことが多いことを示している．三番目に，JKNet はヘテロフィリックグラフでは効果的であることがわかる．最後に，アテンションはホモフィリックおよびヘテロフィリックグラフともに大きな差がないことがわかる．そのため，constant, GCN，および GAT をグラフタイプに合わせて事前に選択することが難しいことがわかる．これらの洞察は GNN 構造の設計に有益である．

上記の分析にあわせて，GCN の性能を Tanh と ReLU を用いた場合で比較する．標準的な GCN は ReLU を活性化関数として用いており，多くの既存研究でも活性化関数を変えた評価はしていない．表 8 に Tanh と ReLU を用いた場合の GCN の AUC を示す．Cornell, Wisconsin, および Actor では，Tanh を用いた方が AUC が増加し，一方他のグラフでは低下した．Wisconsin では AUC が 6.0 増加している．この実験により，活性化関数はグラフに合わせて選択する必要があることを表している．提案手法は，重要なコンポーネントの分析と既存の GNN 構造の再設計に有効であることがわかる．

6 ま と め

本論文はグラフニューラル構造探索を提案した．評価実験において，提案手法はホモフィリックグラフおよびヘテロフィリックグラフの両方において，効果的かつ効率的に GNN 構造を探索できることを示した．今後は，構造パラメータを自動的に生成する手法およびより効果的な探索空間と探索アルゴリズムを開発する．

謝　　辞

本研究は JST さきがけ JPMJPR21C5 および JSPS 科学研究費 JP20H00583 の支援よって行われた．

文　　献

[1] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NeurIPS*, 2015.

[2] Victor Fung, Jiaxin Zhang, Eric Juarez, and Bobby G Sumpter. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, Vol. 7, No. 1, pp. 1–8, 2021.

[3] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, pp. 8459—-8468, 2020.

[4] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[5] Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 2020.

[6] Yang Gao, Hong Yang, Peng Zhang, Chuan Zhou, and Yue Hu. Graph neural architecture search. In *IJCAI*, pp. 1403–1409, 2020.

[7] Huan Zhao, Lanning Wei, and Quanming Yao. Simplifying architecture search for graph neural network. *arXiv preprint arXiv:2008.11652*, 2020.

[8] Kaixiong Zhou, Qingquan Song, Xiao Huang, and Xia Hu. Auto-gnn: Neural architecture search of graph neural networks. *arXiv preprint arXiv:1909.03184*, 2019.

[9] Jiaxuan You, Zhitao Ying, and Jure Leskovec. Design space for graph neural networks. *NeurIPS*, 2020.

[10] Xin Zheng, Miao Zhang, Chunyang Chen, Qin Zhang, Chuan Zhou, and Shirui Pan. Auto-heg: Automated graph neural network on heterophilic graphs. *WWW*, pp. 611—-620, 2023.

[11] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *NeurIPS*, 2020.

[12] Shaofei Cai, Liang Li, Jincan Deng, Beichen Zhang, Zheng-Jun Zha, Li Su, and Qingming Huang. Rethinking graph neural architecture search from message-passing. In *CVPR*, pp. 6657–6666, 2021.

[13] Chaoyu Guan, Xin Wang, Hong Chen, Ziwei Zhang, and Wenwu Zhu. Large-scale graph neural architecture search. In *ICML*, pp. 7968–7981, 2022.

[14] Yaoman Li and Irwin King. Autograph: Automated graph neural network. In *NeurIPS*, 2020.

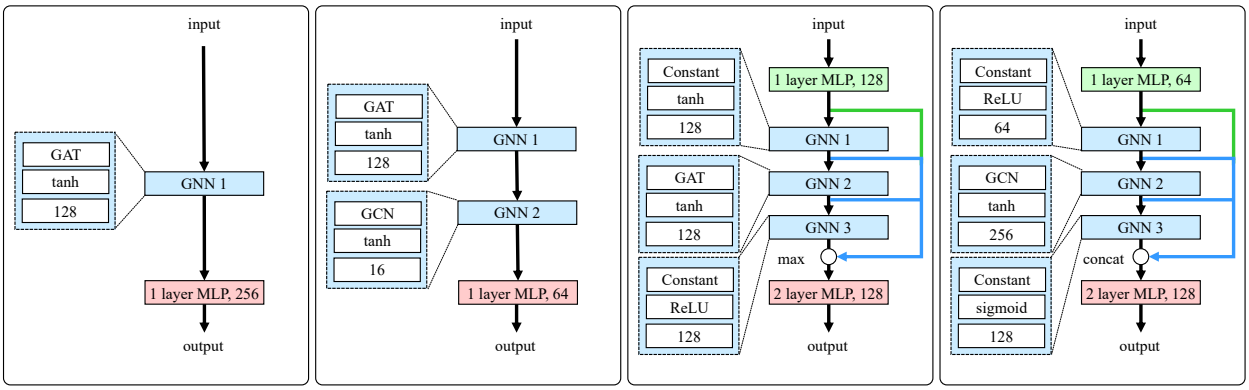
[15] Wentao Zhang, Zheyu Lin, Yu Shen, Yang Li, Zhi Yang, and Bin Cui. Dfg-nas: Ddeep and flexible graph neural architecture search. In *ICML*, pp. 26362–26374, 2022.

[16] Wentao Zhang, Yu Shen, Zheyu Lin, Yang Li, Xiaosen Li, Wen Ouyang, Yangyu Tao, Zhi Yang, and Bin Cui. Pasca: A graph neural architecture search system under the scalable paradigm. In *WWW*, pp. 1817–1828, 2022.

[17] Yiren Zhao, Duo Wang, Xitong Gao, Robert Mullins, Pietro Lio, and Mateja Jamnik. Probabilistic dual network architecture search on graphs. *arXiv preprint arXiv:2003.09676*, 2020.

[18] Yuhui Ding, Quanming Yao, Huan Zhao, and Tong Zhang. Diffmg: Differentiable meta graph search for heterogeneous graph neural networks. In *KDD*, pp. 279–288, 2021.

[19] ZHAO Huan, YAO Quanming, and TU Weiwei. Search to aggre-



(a) Citeseer (b) Amz-c (c) Cornell (d) Squirrel

図 4: 選択された GNN 構造の例

表 7: 選択された構造パラメータの割合

	層数			活性化関数				JKNet			アテンション		
	1	2	3	none	ReLU	Sigmoid	Tanh	none	concat	max	constant	GCN	GAT
Hom	0.14	0.63	0.23	0.32	0.10	0.04	0.55	0.29	0.54	0.17	0.30	0.48	0.22
Hetero	0	0.47	0.53	0.11	0.21	0.17	0.51	0.03	0.67	0.30	0.25	0.51	0.24

表 8: GCN における活性化関数の影響

	Cornell	Wisconsin	Chameleon	Squirrel	Actor	Penn94
GCN (ReLU)	70.2	67.6	85.7	71.9	56.4	89.3
GCN (Tanh)	71.1	73.5	80.3	69.8	58.1	87.9

gate neighborhood for graph neural network. In *ICDE*, pp. 552–563, 2021.

[20] Yanxi Li, Zean Wen, Yunhe Wang, and Chang Xu. One-shot graph neural architecture search with dynamic search space. In *AAAI*, pp. 8510–8517, 2021.

[21] Yijian Qin, Ziwei Zhang, Xin Wang, Zeyang Zhang, and Wenwu Zhu. Nas-bench-graph: Benchmarking graph neural architecture search. 2022.

[22] Min Shi, David A Wilson, Xingquan Zhu, Yu Huang, Yuan Zhuang, Jianxun Liu, and Yufei Tang. Evolutionary architecture search for graph neural networks. *arXiv preprint arXiv:2009.10199*, 2020.

[23] Zhen Wang, Zhewei Wei, Yaliang Li, Weirui Kuang, and Bolin Ding. Graph neural networks with node-wise architecture. In *SIGKDD*, pp. 1949–1958, 2022.

[24] Ziwei Zhang, Xin Wang, and Wenwu Zhu. Automated machine learning on graphs: A survey. *arXiv preprint arXiv:2103.00742*, 2021.

[25] Peng Xu, Lin Zhang, Xuanzhou Liu, Jiaqi Sun, Yue Zhao, Haiqin Yang, and Bei Yu. Do not train it: A linear neural architecture search of graph neural networks. In *ICML*, 2023.

[26] Guanghui Zhu, Wenjie Wang, Zhuoer Xu, Feng Cheng, Mengchuan Qiu, Chunfeng Yuan, and Yihua Huang. Psp: Progressive space pruning for efficient graph neural architecture search. In *ICDE*, pp. 2168–2181, 2022.

[27] Zizhao Zhang, Xin Wang, Chaoyu Guan, Ziwei Zhang, Haoyang Li, and Wenwu Zhu. Autogt: Automated graph transformer architecture search. In *ICLR*, 2023.

[28] TianJin Deng and Jia Wu. Efficient graph neural architecture search using monte carlo tree search and prediction network. *Expert Systems with Applications*, Vol. 213, p. 118916, 2023.

[29] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *ICML*, pp. 21–29, 2019.

[30] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *ICML*, pp. 1725–1735, 2020.

[31] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *ICLR*, 2021.

[32] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

[33] Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? 2021.

[34] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Is heterophily a real nightmare for graph neural networks to do node classification? 2022.

[35] Xiang Li, Renyu Zhu, Yao Cheng, Caihua Shan, Siqiang Luo, Dongsheng Li, and Weining Qian. Finding global homophily in graph neural networks when meeting heterophily. In *ICML*, pp. 13242–13256, 2022.

[36] Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *NeurIPS*, 2021.

[37] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. 2019.

[38] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *ECML*, pp. 282–293, 2006.

[39] Linnan Wang, Yiyang Zhao, Yuu Jinnai, Yuandong Tian, and Rodrigo Fonseca. Neural architecture search using deep neural networks and monte carlo tree search. In *AAAI*, pp. 9983–9991, 2020.

一般発表 | Track 1: 自然言語処理・機械学習基礎

汎用機械学習技術 (1)

座長:天方 大地(大阪大学)

コメンテータ:新井 淳也(日本電信電話株式会社)

2024年2月28日(水) 10:00 〜 12:10 T1-A (オンライン（Zoom Events))

11:35 〜 12:00

[T1-A-1-05]【技術報告】 LayerXにおける AI・機械学習技術の活用事例とその裏側

松村 優也 (株式会社LayerX)

©一般社団法人日本データベース学会

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 10:00 ~ 12:10 | T1-B オンライン（Zoom Events)

言語モデル

座長:喜田 拓也(北海学園大学)
コメンテータ:若林 啓(筑波大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T1-B-1-01] Evaluating saliency methods by LLMs

*LU XIAOTIAN[1] (1. 京都大学大学院情報学研究科知能情報学コース鹿島研究室)

10:20 ~ 10:45
[T1-B-1-02] 確信度を用いたBERTによる参考文献書誌情報抽出の誤り検出の一手法

*中山 竣平[1]、金澤 輝一[2]、上野 史[3]、太田 学[3] (1. 岡山大学大学院環境生命自然科学研究科　太田研究室、2. 国立情報学研究所、3. 岡山大学学術研究院環境生命自然科学学域)

10:45 ~ 11:10
[T1-B-1-03] LLMにおけるAttention機構の役割の分析

*大塚 空来[1]、湯本 高行[1] (1. 兵庫県立大学)

11:10 ~ 11:35
[T1-B-1-04] k近傍法を用いた事前学習済みBERTモデル検索

*ファム フーロン[1]、三林 亮太[1]、山本 岳洋[1]、加藤 誠[2]、山本 祐輔[3]、荘司 慶行[4]、大島 裕明[1] (1. 兵庫県立大学、2. 筑波大学、3. 名古屋市立大学、4. 静岡大学)

11:35 ~ 12:00
[T1-B-1-05] 【技術報告】NECのLLM技術とデータ分析高速化技術

小山田昌史、董于洋、山崎智史、石坂一久 (日本電気株式会社)

Evaluating saliency methods by LLMs

Xiaotian Lu †

† Graduate School of Informatics, Kyoto University 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

E-mail: † lu@ml.ist.i.kyoto-u.ac.jp

Abstract Deep learning models are black boxes, i.e. it is difficult to obtain the relationship between the input and output of deep learning models. One solution is to use saliency methods to highlight the important parts of the input. However, as well as several saliency methods have been proposed, how to evaluate which saliency method is better remains open question. In this study, we provide the important parts of the input given by different saliency methods to Large Language Models (LLMs) and evaluate saliency methods by observing whether LLMs could recognize the correct label of the input.

Keyword NLP、Saliency method、Large Language Model

1. Introduction

Deep learning models (DNNs) have performed well in many fields recently, but they are often considered black boxes. In other words, it is challenging for humans to understand the relationship between the output and input of DNNs. In some crucial application fields, such as medical care, DNNs maybe not trustable.

To solve this problem, opening the black box of the DNNs, many saliency methods have been extensively investigated. Saliency methods estimate the influence of input feature subsets on model predictions. Saliency methods give each input feature a saliency score to form an importance ranking of the input feature. More significant features will receive higher scores, and important features in the input can be identified by ranking contribution scores. We can enhance trust in the network and bolster confidence in its decision-making by assessing the coherence of the importance ranking.

The explanations provided by various saliency methods differ from one another. It is challenging for us to determine which saliency method is superior. Therefore, evaluating saliency methods addresses an important issue.

Large Language Models (LLMs), such as ChatGPT, exhibit excellent performance in various tasks. In this study, we utilize ChatGPT to evaluate saliency methods on text dataset.

2. Preliminary

Saliency methods can be categorized into two types: gradient-based and perturbation-based. Gradient-based saliency methods are unsuitable for non-differentiable models like random forests, while perturbation-based saliency methods can be applied to any model that maps an input to an output.

Gradient-based saliency methods operate on the assumption that more important features exhibit larger gradients. Vanilla Gradient[1] represents the most basic method, measuring the direct gradient of input features to assess the impact of small changes in each input dimension on the network's output. InputXGrad[2] involves multiplying the input features by the gradient with respect to those features. However, a zero gradient does not necessarily imply zero contribution to the output. To address this, Integrated Gradient[3] is calculated by accumulating gradients along a straight-line path from specified baselines to inputs. Deep Learning Important FeaTures (DeepLIFT) [4] decomposes the contribution to the prediction output of a network into the contribution of each hidden neuron to a given input feature.

Perturbation-based saliency methods operate under the assumption that perturbing more important features results in larger influences on the output. LIME[5] samples points around inputs and trains a simpler, interpretable surrogate model, such as a linear model, based on the samples. The weights of the simple model are then used as the contribution scores for inputs.

In addition to common saliency methods, hidden states—variables or outputs of hidden layers—are sometimes used as explanations. In computer vision tasks, the output of convolutional layers may highlight important parts of the

input image. In this study, we define 'All Attention' as the mean of all layers and heads of self-attention layers for a given word, and 'Last Attention' as the mean of the last layer and all heads.

3. Proposed Method

We assessed saliency methods by presenting ChatGPT with the top important words identified by the saliency method using the IMDB[6] dataset and the BERT model. IMDB is a binary text classification dataset comprising negative and positive movie reviews. In the evaluation process, ChatGPT is only provided with the most important words in a sentence, with all other words replaced by '.'. We then had ChatGPT infer the category of the sample based on the displayed information. The experiment involved settings with varying numbers of top important words, specifically, 5, 10, 20, 30, and 40. For example, consider the movie review: '*This movie made it into **one** of my **top** 10 most **awful** movies. **Horrible**. There wasn't a **continuous** minute where there wasn't a fight with one monster or another.*' If the Saliency method identifies the words in bold as the five most important words, we will ask ChatGPT to determine, based on the following sentence '*... one ... top ... awful. Horrible ... continuous ...*', whether this movie review should be classified as negative or positive. ChatGPT might infer a negative sentiment for this movie review based on the presence of words such as 'awful' and 'Horrible' in the sentence. We believe that the better a saliency method performs, the higher its accuracy should be in the experiment. This implies that the important words it identifies are more meaningful.

4. Results

Table 1 shows the results of our experiment.

Table 1: Accuracy of different saliency methods on the IMDB dataset by ChatGPT. The numbers in brackets show the ranks of the saliency method. The bold numbers show the best results, and the underlined numbers show the top-3 methods.

	5 Words	10 Words	20 Words	30 Words	40 Words
Random	47 (7)	57 (7)	64 (7)	72 (7)	76 (7)
All Attention	88 (2)	86 (4)	92 (3)	93 (3)	87 (4)
Last Attention	73 (5)	89 (3)	89 (4)	93 (3)	91 (3)
Vanilla Gradient	88 (2)	91 (2)	**97 (1)**	**96 (1)**	95 (2)
InputXGrad	73 (5)	65 (6)	78 (5)	75 (6)	79 (6)
Integrated Gradient	**94 (1)**	**97 (1)**	95 (2)	95 (2)	**98 (1)**
DeepLIFT	78 (4)	73 (5)	76 (6)	79 (5)	81 (5)
LIME	39 (8)	54 (8)	64 (7)	68 (8)	75 (8)

A random baseline, i.e., assigning input features random contribution scores, was included in the experiment.

According to the experimental results, it is observed that as more words are displayed, the accuracy of all methods increases. Integrated Gradient performs the best among all methods. This may be attributed to the property of completeness inherent in Integrated Gradient, which may highlight potential top important words that the model can learn.

5. Conclusion

In this study, we propose a method for evaluating saliency methods using LLMs. Empirically, we have found that Integrated Gradient performs the best and LIME performs worse than the random baseline on the IMDB dataset.

References

[1] Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions[J]. The Journal of Machine Learning Research, 2010, 11: 1803-1831.

[2] Shrikumar A, Greenside P, Shcherbina A, et al. Not just a black box: Learning important features through propagating activation differences[J]. arXiv preprint arXiv:1605.01713, 2016.

[3] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C]//International conference on machine learning. PMLR, 2017: 3319-3328.

[4] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences[C]//International conference on machine learning. PMLR, 2017: 3145-3153.

[5] Ribeiro M T, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.

[6] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011: 142-150.

確信度を用いたBERTによる参考文献書誌情報抽出の誤り検出の一手法

中山　竣平† 　金澤　輝一†† 　上野　史††† 　太田　学†††

† 岡山大学大学院環境生命自然科学研究科　〒 700–8530 岡山市北区津島中 3-1-1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2
††† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山市北区津島中 3-1-1
E-mail: †pb8y7087@s.okayama-u.ac.jp, ††tkana@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　参考文献書誌情報は電子図書館における文書間リンクの生成等に役立ち，著者やタイトルなどの書誌情報をメタデータとして精度良く抽出する手法が研究されている．中でも荒川らは汎用的な自然言語処理の手法であるBERT を利用し，電子情報通信学会英文論文誌の参考文献文字列を用いた実験で 93.37%の抽出精度を示した．しかし，実サービスとして利用するには，書誌情報の抽出誤りを人手で修正する作業が必要となり，そのコストが非常に大きい．そこで，本稿ではその修正コストの削減のため，参考文献書誌情報抽出の誤りの自動検出を試みる．具体的には，BERT 抽出器によって参考文献文字列のトークンに付与される書誌情報の予測確率，そしてその分散とエントロピーを抽出の確信度として定義し，抽出誤りを検出した．

キーワード　書誌情報抽出，参考文献文字列，BERT，電子図書館

1　はじめに

図書館情報学では，文書から抽出したメタデータのサービスへの展開が研究されている．中でも参考文献書誌情報は，電子図書館等において蓄積された多数の学術論文の文書間リンクを生成し，様々なサービスに役立つ重要なメタデータである．近年では自然言語処理技術を利用した手法が研究され，参考文献の著者名やタイトルといった書誌情報をメタデータとして，参考文献文字列から自動抽出する技術が提案されている．

荒内ら [1] は Conditional Random Fields（CRF）[2]，荒川ら [3] は Bidirectional Encoder Representations from Transformers (BERT) [4] による参考文献書誌情報抽出を行ったが，その抽出精度はそれぞれ 90.37%，93.37%と高水準を示している．そのため，電子図書館の運用への支援を想定すると，参考文献書誌情報の抽出誤りを適宜人手で修正した上での利用が可能となっているが，修正作業のコストは無視できない．そこで，本研究では荒内ら [1] が CRF 抽出器において定めた確信度と同様に，トークンに付与される書誌情報の予測確率，そしてその分散とエントロピーを確信度として BERT 抽出器の結果に適用することで，書誌情報抽出結果の誤りの自動検出を試みる．実験では，3 種類の確信度それぞれの抽出誤りの自動検出の精度を評価し，修正によって得られる書誌情報の精度とそれにかかる修正コストについて考察する．さらに，高性能なチャットボットとして知られる ChatGPT を用いて書誌情報抽出及び抽出誤りを検出し，その結果について考察する．

本稿の構成は次の通りである．まず，2 節で学術論文からの書誌情報抽出及び抽出誤りの検出に関する研究を紹介し，3 節で本研究で用いる BERT による参考文献書誌情報抽出法について説明する．つづく 4 節で，3 種類の確信度による参考文献書誌情報抽出の誤りの自動検出について説明する．5 節で参考文献書誌情報結果及び抽出誤りの検出結果とその考察を述べる．6 節で ChatGPT による実験の説明とその結果及び考察を述べる．最後に 7 節でまとめる．

2　関連研究

2.1　参考文献書誌情報抽出

機械学習による参考文献書誌情報抽出には，CRF [2] を用いた書誌情報抽出に関する研究が多く，例えば Do ら [5]，Cuong ら [6] の研究がある．Do ら [5] は CRF により著者とその所属機関を識別する情報抽出システムである Enlil を開発した．3 つのデータセットを対象に文字列から著者名を抽出し，その所属機関を照合する実験を行った．ACM Digital Library を対象とした著者名の抽出の F 値は 0.946 であり，著者とその所属機関の照合における F 値は 0.889 であった．また，Cuong ら [6] は CRF を拡張した higher order semi-Markov CRFs (HO-SCRFs) を用いて参考文献文字列から書誌情報を抽出し，合計 1,384 件の参考文献文字列を対象に，13 項目の書誌情報を抽出する実験を行ったところ，平均 F 値は 0.943 であった．

ニューラルネットワークによる参考文献書誌情報抽出では，Prased らの開発した Neural ParsCit [7] が，単語の分散表現と単語を構成する文字の分散表現を Bi-directional LSTM [8] へ入力し，その出力を CRF に入力することで，参考文献文字列から書誌情報を抽出する．Cora [9] データセットに対する実験の結果，その平均 F 値は 0.914 であった．浪越ら [10] は Bi-directional LSTM-CNN-CRF を利用して 2000 年の電子情報通信学会英文論文誌の参考文献文字列から書誌情報を抽出する実験を行い，書誌情報抽出精度は 92.8%であった．荒川ら [3] は BERT を利用して浪越らと同じ参考文献文字列データから参考文献書誌情報を抽出し，その書誌情報抽出精度は 93.4%であった．

2.2 情報抽出の誤り検出

荒内ら [1] は CRF で参考文献文字列から書誌情報を抽出した結果から，トークンに付与される書誌要素ラベルの周辺確率を用いた 3 種類の確信度を利用して，抽出誤りを自動検出する手法を提案した．実験により情報処理学会論文誌の参考文献文字列で，自動抽出後にその 35.7%を人手で確認して修正することで，99%の抽出精度となることを示した．村田ら [11] は，日本語作文の形態素解析の誤り検出を行い，形態素自体の現れにくさである生起コスト及び形態素同士の連続しにくさである連接コストを合わせた累積コストを利用し，累積コストのランキングに基づいた 10 通りのデータを機械学習させた手法を提案した．実験により，形態素解析失敗の検出精度の正解率は 82%であった．Rei ら [12] は，ニューラルネットワークモデルによる文法誤り検出を行った．誤用文と正用文の文対を比較し，一致する部分と一致しない部分で別のラベルを付けることで誤りを検出する手法を提出した．実験では，Bi-LSTM が最も良い結果を示し，文法誤りの検出精度は 44%であった．我々 [13] は，BERT の最大予測確率を利用し，設定した閾値以下の最大予測確率のトークン列を含む参考文献文字列を，抽出誤りのある参考文献文字列と定義し参考文献書誌情報抽出の誤り検出実験を行った．閾値が 0.95 の時，再現率が最も高く 0.7896 で，検出した参考文献文字列を確認して修正した書誌情報抽出精度は 98.9%であった．

3 BERT による参考文献書誌情報抽出

3.1 参考文献書誌情報抽出の流れ

本研究の参考文献書誌情報抽出は，荒川ら [3] と同様に BERT を利用する．図 1 に示す順で参考文献文字列から著者名やタイトルといった参考文献書誌情報抽出の流れを説明する．具体的には，入力された参考文献文字列をデリミタで分割して得られるトークン列に書誌情報 BI ラベルを付与する．この書誌情報 BI ラベルは，先頭（B）とそれ以外（I）を分ける BI ラベルと書誌情報の種類を組み合わせたものである．各トークンに対してトークンが書誌要素の先頭であれば「書誌要素 B」，先頭以外にあれば「書誌要素 I」というラベルを付与する．これらを書誌要素 BI ラベルと呼ぶ．同様に，トークンがデリミタの先頭であれば「デリミタ B」，先頭以外であれば「デリミタ I」というラベルを付与する．これらをデリミタ BI ラベルと呼ぶ．この書誌要素 BI ラベルとデリミタ BI ラベルを合わせたものが書誌情報 BI ラベルである．

抽出する書誌要素の一覧と，それに対応する書誌要素ラベルを表 1 にまとめる．表 1 の Other は他のどの書誌要素にも分類されない書誌要素であり，具体的には所属機関や特許情報などが含まれる．また，抽出するデリミタの一覧と，それに対応するデリミタラベルを表 2 にまとめる．本研究の参考文献書誌情報抽出は，図 1 の結合処理によって同種の連続する BI ラベルを結合し，書誌情報を抽出する．すなわち，個々の書誌要素やデリミタの塊を表すトークンとその書誌情報の種類を同時に推定する．

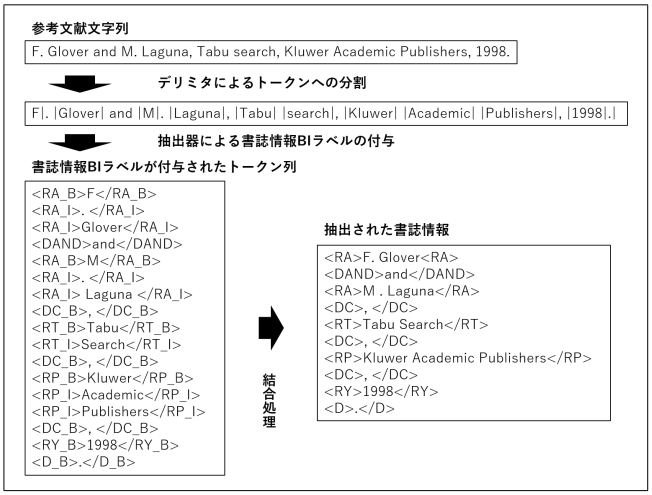


図 1　本研究における参考文献書誌情報抽出の流れ

表 1　抽出する書誌要素とそのラベル

書誌要素	ラベル
Author	RA
Editor	RE
Translator	RTR
Author Other	RAOT
Title	RT
Booktitle	RB
Journal	RW
Conference	RC
Volume	RV
Number	RN
Page	RPP
Publisher	RP
Day	RD
Month	RM
Year	RY
Location	RL
URL	RURL
Patent	RPTNT
Other	ROT

3.2 BERT とその利用

BERT [4] は大量の教師なしデータで事前学習し，少量の教師ありデータでファインチューニングすることで多様なタスクに対応できる自然言語処理モデルである．本研究では，英語 Wikipedia と BookCorpus [15] で事前学習した $\text{BERT}_{\text{base}}$ を，教師データとする書誌情報 BI ラベルが付与された参考文献文字列によりファインチューニングして，参考文献文字列から書誌情報を抽出する．$\text{BERT}_{\text{base}}$ は Multi-Head Attention を利用した Transformer [14] のエンコーダが 12 層重なった構成となっている．Transformer では入力を前後の双方向から読み込むことで入力トークン間の前後関係を重みとして計算し，ベクトルを更新する．これにより複雑な文脈や単語間の関係を考慮することができる．参考文献文字列では，大文字と小文字の違いも書誌情報抽出の手がかりとなるため，本研究で用いる BERT には大文字と小文字を区別する bert-base-cased を使用

表 2　デリミタとそのラベル

デリミタ	デリミタラベル
.（ピリオド）	D
,（カンマ）	DC
and	DAND
Eds., eds., Ed., ed., editors, 訳, 編, 編著, 監修, 監訳, 編集, （訳）, （編）, （編著）, （監）, （監修）, （監訳）, （邦訳）, （共訳）	DED
”（二重引用符）	DE
Vol., vol.	DV
No., no.	DN
Nos., nos.	DNS
pp.	DPP
p.	DP
:, ;（コロン，セミコロン）	DCL
/（スラッシュ）	DSL
-（ハイフン）	DHY
(, [, {（各種括弧・始）	DLBR
),], }（各種括弧・終）	DRBR
その他	DUN

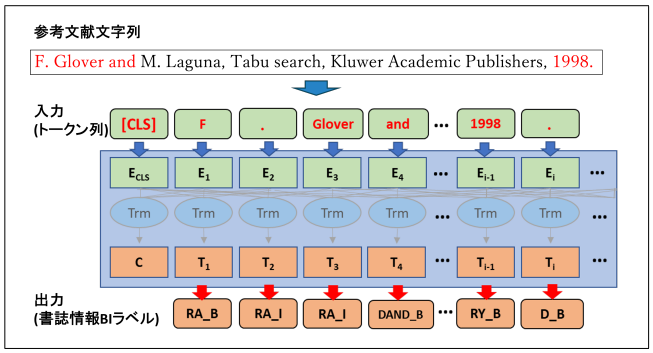


図 2　BERT による書誌情報 BI ラベル付与の概略

する．BERT が参考文献のトークン列に書誌情報 BI ラベルを付与する様子を図 2 に示す．BERT への入力は参考文献文字列を表 2 のデリミタで区切ったトークン列，出力は各トークンに対する各ラベルの予測確率である．その予測確率から表 1 と表 2 で示したラベルに BI ラベルを付加した書誌情報 BI ラベルを特定する．BERT の入力の先頭にある [CLS] はトークン列の先頭につける特殊なトークンである．[CLS] の出力は分類タスクの集約シーケンス表現として使用されるが，本研究で行う参考文献書誌情報抽出では利用しない．なお，図 2 の中の Trm は Transfomer を表している．

4　参考文献書誌情報抽出の誤り検出

本研究では，BERT による参考文献書誌情報抽出で得た，参考文献文字列のトークンに付与された書誌情報ラベルの予測確率を用いて，書誌情報抽出誤りの自動検出手法を提案する．図 3 は本研究で用いる BERT の最終層の一部を表した例である．図 3 に示すように，BERT の最終層には参考文献文字列の各トークンに対する書誌情報 BI ラベルとその予測確率がある．例えば，先頭のトークンである“ F ”は，Author の先頭を表す

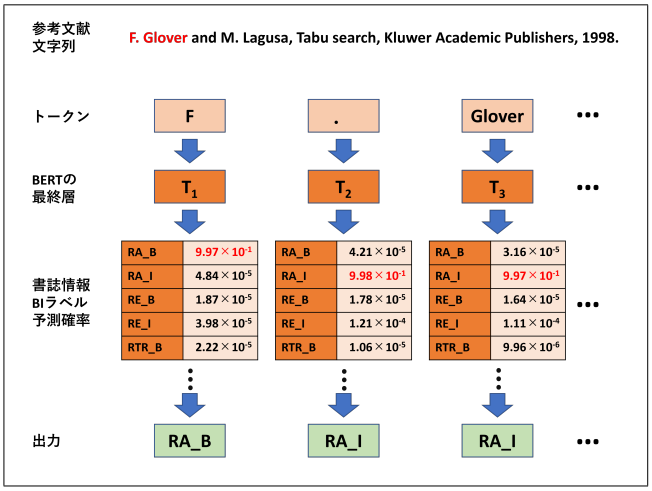


図 3　参考文献書誌情報抽出を行う BERT の最終層とその出力

書誌情報 BI ラベルである RA_B の予測確率が 0.997 で最大であるため，参考文献書誌情報抽出の出力が RA_B となる．本稿では，この書誌情報 BI ラベルの予測確率を利用して 3 種類の確信度を定義し，この確信度が低いトークンを書誌情報抽出誤りを含む可能性が高い参考文献文字列として検出する．

4.1　書誌情報抽出の確信度

本研究では荒内ら [1] の CRF 抽出器の提案に基づき 3 種類の確信度にならって BERT 抽出器の確信度を定義する．この確信度は，参考文献文字列中の各トークンの書誌情報の確からしさを表す書誌情報 BI ラベルの予測確率に基づいて以下の 3 種類を定める．

- **Marginal Probability of Assigned Label**

1 つ目の確信度は，トークンに付与された書誌情報ラベルの最大予測確率のみを利用して定める．具体的には式 (1) で定義する

$$c_{MP}(x_i) := \max_{l \in L} P(Y_i = l) \tag{1}$$

ここで入力トークン列 $\mathbf{x}$ に対して，Y_i は参考文献文字列中の i 番目のトークン x_i に対して参考文献書誌情報抽出により付与された書誌情報ラベルである．L はデリミタを含む書誌情報ラベルの集合であり，その詳細は表 1 と表 2 の通りである．確率 $\max_{l \in L} P(Y_i = l)$ は i 番目のトークンに注目したラベル付与の確信度とみなすことができる．この確信度を MP とする．

- **Variance of Marginal Distribution**

2 つ目の確信度は，トークンに付与された書誌情報ラベルの予測確率だけでなく他の候補ラベルの予測確率も考慮して算出する．多くのラベル候補の予測確率がほとんど同等であるならば，付与された書誌情報ラベルの確信度が低いとみなすことができる．式 (2) のように正規化した分散の値を確信度として利用する．

$$c_{Var}(x_i) := \frac{n}{n-1} \sum_{l \in L} \left(P(Y_i = l) - \frac{1}{n} \right)^2 \tag{2}$$

n はデリミタを含む書誌情報ラベルの種類数を表している．こ

の確信度を Var とする．

- **Entropy of Marginal Distribution**

分散だけでなくエントロピーも周辺確率の分布の特徴を表す．よって 3 つ目の確信度は，全候補ラベルの予測確率のエントロピーを利用して定める．

エントロピーは値が大きいほど多くの候補ラベルに確率が分散していることを表す．よって 1 と正規化したエントロピーの差を確信度として利用する．

$$c_{Ent}(x_i) := 1 - \frac{\sum_{l \in L} -P(Y_i = l) \log P(Y_i = l)}{\log n} \qquad (3)$$

この確信度を Ent とする．

4.2 確信度に基づく誤り検出

定義した 3 種類の確信度を用いて参考文献書誌情報抽出の誤りを含む可能性が高いと判断する参考文献文字列を自動検出する．具体的には 4.1 節で定義したトークンの確信度を利用した書誌情報抽出誤りの検出で，確信度に閾値を設定し，閾値以下のトークンを 1 つでも含む参考文献文字列を書誌情報抽出誤りを含む可能性が高い参考文献文字列として検出する．ただし，表 2 のデリミタ BI ラベルが付与されたトークンとその予測確率は誤りの検出に利用しない．

5 評 価 実 験

参考文献書誌情報抽出の誤りの自動検出精度を検証するため，評価実験を行う．まず．5.1 節で実験の諸条件を説明し，5.2 節で BERT の書誌情報抽出精度を示す．5.3 節で書誌情報抽出誤りの検出実験を行い，その検出精度を評価する．各確信度を用いた修正コストや修正によって保証できる精度をまとめる．

5.1 実験の諸条件

実験データとして，2000 年の電子情報通信学会英文論文誌 (IEICE-E) に含まれる参考文献文字列 4,497 件を利用する．なお，書誌情報抽出精度を 5 分割交差検証で算出するため，4,497 件の参考文献文字列を 5 つに分割し，そのうち 4 つをファインチューニングデータ，残りの 1 つをテストデータとする．本実験における BERT のハイパーパラメータを表 3 にまとめる．また，抽出した書誌要素の正解判定は類似している書誌要素をまとめた表 4 の大分類で行う．つまり，この大分類が同じであれば正解と判定する．なおデリミタトークンの抽出誤りは正誤判定において無視するため，デリミタトークンしか誤っていない参考文献文字列は抽出に成功したものとなる．

5.2 BERT による参考文献書誌情報抽出

表 5 に，BERT 抽出器による参考文献書誌情報抽出結果として，その精度と書誌要素の推定誤りを含む参考文献文字列数とそうでない参考文献文字列数を示す．この参考文献文字列数は 5.1 節で 5 分割した電子情報通信学会英文論文誌の合計である書誌要素の抽出誤りを含む参考文献文字列数は 189 件であり，これはデータ全体の 4.2%にあたる．

表 3　bert-base-cased のハイパーパラメータ

ハイパーパラメータ	値
max sequence length	300
training batch size	10
evaluation batch size	64
warmup propotion	0.1
training rate	5.0×10^{-4}
weight decay	1.0×10^{-2}

表 4　抽出する書誌要素とその大分類 [3]

書誌要素 (書誌要素ラベル)	評価実験における大分類
Author (RA)，Editor (RE) Translator (RTR)，Author Other (RAOT)	AUTHOR
Title (RT)，Booktitle (RBT)	TITLE
Journal (RW)，Conference (RC)	JOURNAL
Volume (RV)，Number (RN)，Page (RPP)	VOLUME
Publisher (RP)	PUBLISHER
Day (RD)	DAY
Month (RM)	MONTH
Year (RY)	YEAR
Location (RL)，URL (RURL) Patent(RPTNT), Other (ROT)	OTHER

表 5　BERT による書誌情報抽出結果

書誌情報抽出精度	0.9580
参考文献文字列数（抽出誤り有り）	189
参考文献文字列数（抽出誤り無し）	4,308

5.3 参考文献書誌情報抽出の誤り検出実験

3 種類の確信度の閾値を 0.1〜1.0 の間で 0.1 ずつ変えて，閾値以下の確信度の書誌要素トークンを含む参考文献文字列を抽出誤りとして検出した際の再現率を図 4，適合率を図 5，F 値を図 6 に示す．なお，再現率と適合率は以下に定義する通りで，F 値はそれらの調和平均である．

再現率

参考文献書誌情報抽出の誤りを含む参考文献文字列のうち正しく検出できた参考文献文字列の数

適合率

検出した参考文献文字列のうち実際に誤りを含む参考文献文字列の数

図 4 から，再現率は閾値とともに単調増加し，閾値が 0.2 から 0.9 では Var が高いことがわかる．閾値 0.9 以下を検出したときの再現率は MP が 0.7354，Var が 0.7937，Ent が 0.7672 であり，Var は MP，Ent よりそれぞれ 5.83 ポイント，2.65 ポイント高いことがわかる．図 5 において，閾値 0.9 以下の抽出誤りの検出では，MP が最も高く 0.4947，Ent が 0.4850，Var が 0.4587 となっている．再現率が高かった Var は抽出誤りを含む参考文献文字列を MP や Ent に比べて多く検出するが，図 6 の F 値では，閾値 0.1〜0.7 の間は Var が最も高かったが，閾値 0.8 では Ent が最大で 0.6212 となり，閾値 0.9 でも Ent が最大だった．0.8 や 0.9 といった閾値が大きい場合，F 値では Ent が 3 種類の確信度の中で最良であることがわかる．

次に，人手による修正コストを評価するため，検出した参考

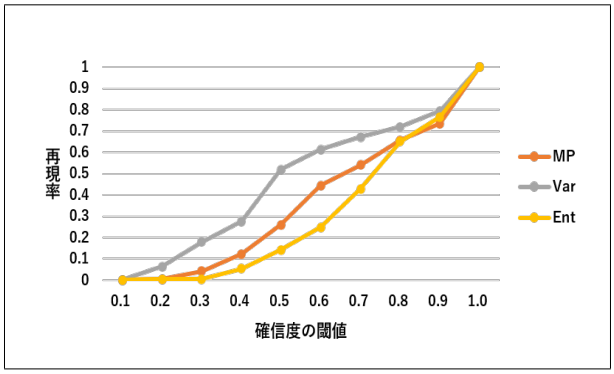


図 4　書誌情報抽出誤り検出の再現率

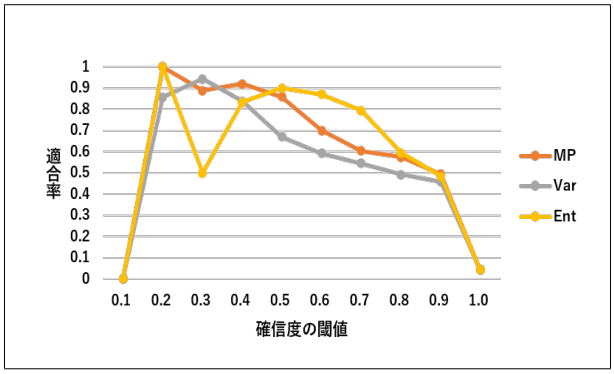


図 5　書誌情報抽出誤り検出の適合率

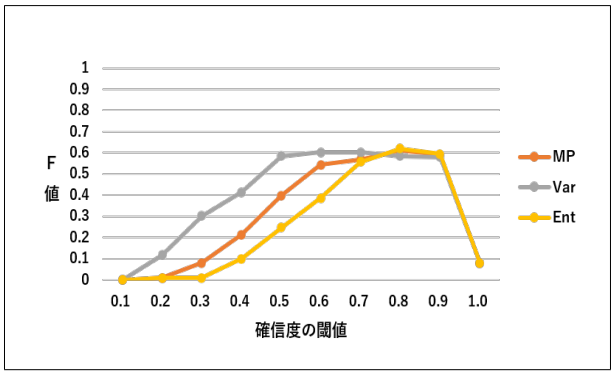


図 6　書誌情報抽出誤り検出の調和平均 F 値

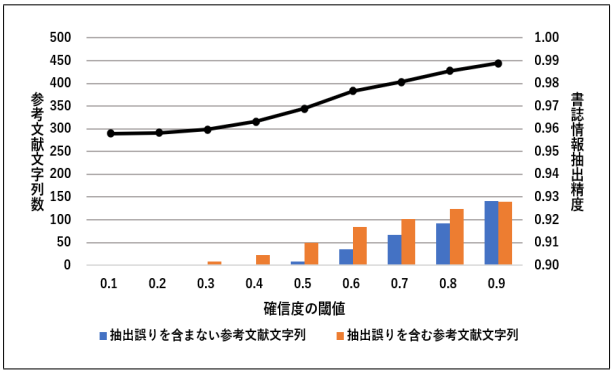


図 7　抽出誤りの検出数と修正後の抽出精度（MP）

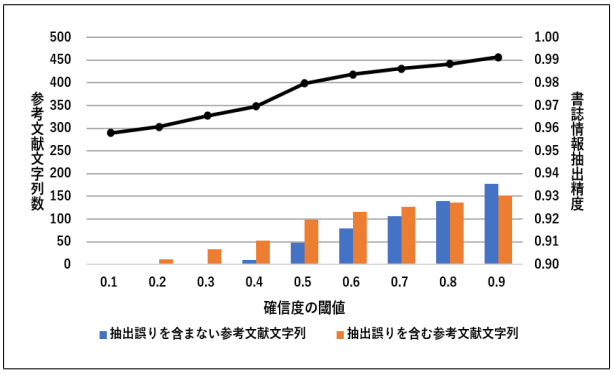


図 8　抽出誤りの検出数と修正後の抽出精度（Var）

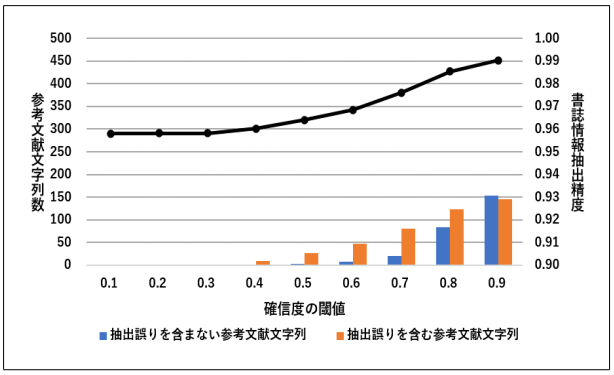


図 9　抽出誤りの検出数と修正後の抽出精度（Ent）

文献文字列をすべて人手で修正したとみなした時の書誌情報抽出精度を示す．確信度 MP の結果を図 7，Var のそれを図 8，Ent のそれを図 9 に示す．これらの図の横軸は確信度の閾値を示し，縦軸の参考文献文字列数は棒グラフ，書誌情報抽出精度は折れ線グラフで表している．棒グラフはその閾値によって検出される参考文献文字列数で，橙は抽出誤りを含む参考文献文字列数，青は抽出誤りを含まない参考文献文字列数である．一方で，折れ線グラフは検出した参考文献文字列を人手で全て修正したとみなした時の書誌情報抽出精度を示している．なお，閾値 1.0 の場合は全参考文献文字列を誤りとして検出し，この書誌情報抽出精度も 1 となるため，図 7，図 8，図 9 では省略している．図 7 から，確信度 MP による抽出誤りを含む参考文献文字列の検出数は閾値 0.9 で 139 件，抽出誤りを含まない参考文献文字列の検出数は 142 件で，検出数は合計 281 件である．図 8 では，確信度 Var による抽出誤りを含む参考文献文字列の検出数は，閾値 0.9 で 150 件，抽出誤りを含まないそれは 177 件で，検出数は合計 327 件である．図 9 の確信度 Ent による抽出誤り含む参考文献文字列の検出数は，閾値 0.9 で 145 件，抽出誤りを含まないそれは 154 件で，抽出誤りの検出数は合計 299 件である．一方図 9 では，閾値 0.3 以下の確信度では参考文献文字列の検出数は 0 件である．

以上より，MP では電子情報通信学会英文論文誌の 4,497 件の実験データの 281 件，すなわちデータ全体の 6.2%を人手で確認して修正することで 98.9%の書誌情報抽出精度が得られている．また，Var では 4,497 件全体の 7.2%を人手で確認して修正することで 99.1%の書誌情報抽出精度，Ent の場合は全体の 6.6%を人手で確認して修正すれば，99.0%の書誌情報抽出精度が得られている．いずれの確信度でも，閾値を 0.9 とすることで約 99%の参考文献文字列の書誌情報を正しく抽出できることになる．

5.4 BERT による書誌情報抽出とその誤り検出の考察

本実験の参考文献書誌情報抽出精度は表 5 の通り 0.9580 だったが，デリミタトークンを含めて抽出の正誤判定をすると書誌情報抽出精度は 0.9331 となり，2.49 ポイント下がる．例えば，Author と Author の間にあるカンマを Author の一部として検出することがあるが，正誤判定にデリミタトークンまで含めると，そのデリミタ以外の書誌要素が正しく抽出されていても誤りとなる．

書誌情報抽出誤りの検出では，確信度の閾値が大きいほど，多くの抽出誤りを含む参考文献文字列を抽出する．しかし，同時に誤検出も増加する．そのため，今後は 0.9 以上という比較的大きな閾値で誤り検出を行っても誤検出，すなわち抽出誤りを含まない参考文献文字列の検出数を減少させることが重要である．また，今回の手法で検出できなかった参考文献文字列として，ある書誌要素に頻出しているトークン列をそれとは別の書誌要素としてもつ参考文献文字列があげられる．例えば，Journal に頻出していたトークン列が Publisher の書誌要素内に含まれた時，Journal の予測確率が大きくなることがある．この原因として過学習の可能性がある．そのため，今後は少ないエポック数で学習するなど，過学習を防ぐ手法の効果を検証したい．

6 ChatGPT の利用

6.1 ChatGPT による参考文献書誌情報抽出実験

BERT と同様に自然言語処理 AI であり，近年話題になっている ChatGPT による参考文献書誌情報抽出を行う．ChatGPT による書誌情報抽出では，ChatGPT に図 10 のプロンプトと参考文献文字列を入力し，書誌要素を抽出する．抽出した要素の正解判定は，BERT と同様に類似した書誌要素をまとめた表 4 の大分類で行う．5 節の実験と同様に 2000 年の電子情報通信学会英文論文誌 (IEICE-E) の参考文献文字列 4,497 件で，書誌情報抽出精度を計算する．本実験では ChatGPT-3.5 を利用し，ファインチューニングは行わない．表 6 に書誌情報抽出精度及び抽出誤りを含む参考文献文字列数と全ての書誌要素が正しく抽出された参考文献文字列数を示す．書誌情報抽出精度では表 5 の BERT より 12.32 ポイント低い結果となった．その原因として ChatGPT の抽出する書誌情報の不安定さがあげられる．具体的には，Volume，Number，Page，Day，Month，Year の 5 種類の書誌要素が Title や Conference などの他の書誌要素に包含されることや ChatGPT が独自の書誌要素ラベルを付与することがある．図 11 にその例を示す．この図の上の参考文献文字列では，Volume の書誌要素である「volume SSE95 49-54」が Journal の書誌要素に包含されている．下の参考文献文字列では，ChatGPT が「18-22 Nov. 1996」を“ Date ”という独自の書誌要素と推定している．本稿での定義ではそれぞれ Day，Month，Year が正しい書誌要素であり，別の参考文献文字列では Day，Month，Year が正しく抽出されることもある．このように Volume や Year などの書誌要素が Journal や Title などの書誌要素と一緒にされたり，ChatGPT が定義

```
Extract the bibliographic information from the references
following the labels and outputs.

#label
Author, Editor,Title, Booktitle,Journal,Conference,
Report,Publisher, Volume, Number,Page,Other, Location, URL.

#output
Author:
Title:
Journal:
Author:
Publisher:
```

図 10　参考文献書誌情報抽出を行う ChatGPT のプロンプト

表 6　ChatGPT の抽出精度及び誤りの総数

書誌情報抽出精度	0.8348
参考文献文字列数（抽出誤り有り）	743
参考文献文字列数（抽出誤り無し）	3,754

表 7　ChatGPT が抽出する書誌要素とその評価分類

書誌要素	評価実験における大分類
Author，Editor	AUTHOR
Title，Booktitle	TITLE
Journal，Conference，Report	JOURNAL
Publisher	PUBLISHER
上記以外の書誌要素（定義していないものも含む）	OTHER

書誌情報VolumeがJournalに包含される参考文献文字列

Author	H. Miwa and J. Yamada
Title	Robust VP reconfiguration method and VP bandwidth control method
Journal	Technical Journal of IEICE, volume SSE95 49-54
Year	1995

書誌情報Day，Month，YearがDateラベルが付与される参考文献文字列

Author	J. Puate and F. Jordan
Title	Using fractal compression scheme to embed a digital signature into an image
Journal	Proc. SPIE Photonics East Symposium
Other	Boston, USA
Date	18-22 Nov. 1996

図 11　ChatGPT による参考文献書誌情報抽出誤りの例

にない書誌情報ラベルを付与したりすることで書誌情報抽出精度が下がっている．

6.2 ChatGPT を利用した確信度による抽出誤りの検出

Volume や Year などの書誌要素の影響で ChatGPT による書誌情報抽出は BERT よりも低い結果となったが，Author や Title といった書誌要素は正しく抽出されていることが多かった．そこで，本稿では BERT による書誌情報抽出誤り検出の課題である誤検出を減らすために ChatGPT を利用する．ChatGPT による書誌情報抽出では，Volume，Number，Page，Day，Month，Year の 5 種類の書誌要素が Title や Conference などの他の書誌要素に包含されることが多いため，その 5 種類の書誌要素を小さな書誌要素，それ以外の書誌要素を大きな書誌要素と呼ぶことにする．そして大きな書誌要素に限定し，ChatGPT による書誌情報抽出結果と BERT の確信度により検出した参考文献文字列の書誌情報抽出結果を比較することで

誤検出を減らす．具体的には以下の手順で行う．

（1） BERT の書誌情報抽出誤り検出で検出された参考文献文字列を対象に，ChatGPT により書誌情報を抽出する．このとき，ChatGPT が抽出する書誌要素と評価のための大分類を表 7 に示す．

（2） ChatGPT による書誌情報抽出結果と BERT による書誌情報抽出結果を比較する．この時書誌要素の比較は評価のための大分類で行う．

（3） 大きな書誌要素のトークン列で，ChatGPT の抽出した書誌要素と BERT が抽出した書誌要素が同じ，もしくは ChatGPT の抽出した書誌要素に BERT の抽出した書誌要素がすべて包含されていれば，その参考文献文字列を誤りとして検出しない．

ChatGPT による書誌情報抽出は 6.1 節で説明したように不安定であり，小さな書誌情報が大きな書誌情報に包含されていることがある．そのため，ChatGPT の抽出した各書誌要素トークン列に BERT が抽出した書誌要素トークン列がすべて包含されるという条件にした．

6.3 ChatGPT を用いた抽出誤りの検出結果

表 8 に 5 節の評価実験で使用した参考文献文字列 4,497 件から，確信度 Var を用いて書誌情報抽出誤りを検出した結果と，その結果を 6.2 節で説明した ChatGPT による処理によってフィルタリングした結果を示す．表 8 に示す通り，抽出誤りの検出において ChatGPT の処理により抽出誤りを含む参考文献文字列数は 17 件減少したが，誤検出，すなわちすべての書誌要素が正しく抽出された参考文献文字列の検出は 88 件減少した．よって，誤検出を減少させる手法として ChatGPT は有効である可能性がある．なお，この ChatGPT によるフィルタリング結果の参考文献文字列数はデータ全体の 4.9%であり，これをすべて人手で修正すると，書誌情報抽出精度は 0.9876 となる．

7 ま と め

本稿では，3 種類の確信度を用いて BERT による参考文献書誌情報抽出の誤りの自動検出を行い，実験によりその検出の精度ならびに見込まれる修正コストを評価した．電子情報通信学会英文論文誌の参考文献文字列 4,497 件を対象とした実験において，BERT 抽出器による書誌情報抽出精度は 95.8%であった．この抽出結果に対し書誌情報ラベルの予測確率に基づいた MP，Var，Ent の 3 種類の確信度を用いて参考文献書誌情報抽出誤りを検出する実験を行った．具体的には確信度が閾値以下となる参考文献文字列を，書誌情報抽出誤りが含まれる可能性が高いものとして検出した．実験の結果，閾値 0.9 のとき誤り検出の F 値は Ent が最も高く 0.5943 であった．また，閾値 0.9 で確信度 MP が検出した全参考文献文字列の 6.2%に相等する参考文献文字列を人手で確認して修正することで 98.9%，Var では閾値 0.9 で全参考文献文字列の 7.2%，Ent では閾値 0.9 で参考文献文字列の 6.6%を人手で確認して修正することで，

表 8　BERT の抽出結果を用いた ChatGPT による事後処理結果

	Var による検出結果	ChatGPT による処理結果
参考文献文字列数（抽出誤り有り）	150	133
参考文献文字列数（抽出誤り無し）	177	88

いずれも 99%以上の書誌情報抽出精度となることを示した．

さらに，本稿では確信度による誤り検出結果を ChatGPT によりフィルタリングして検出精度が改善できるか実験した．実験の結果，検出した抽出誤りを含む参考文献文字列数は 133 件となり，確信度による誤り検出結果から 17 件減少したが，誤検出されていた抽出誤りを含まない参考文献文字列数は 177 件から 88 件となり，89 件減少した．今後の課題として，書誌情報抽出誤りの検出数を減らさずに誤検出のみを減らす方法の検討があげられる．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 (B)(課題番号 22H03904) および 2023 年度国立情報学研究所公募型共同研究 (23FC02) の援助による．

文　　献

[1] 荒内大貴，“CRF による参考文献文字列からの自動書誌情報抽出に関する研究，” 岡山大学大学院自然科学研究科修士論文, 2013.
[2] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data, ” in Proc. of ICML 2001, pp. 282-289, 2001.
[3] 荒川瞭平，金澤輝一，高須淳宏，上野史，太田学，“BERT による参考文献書誌情報抽出における擬似学習データの有効性の評価，” ARG 第 17 回 WI2 研究会予稿集, pp. 25-28, 2021.
[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. of NAACL-HLT, pp. 4171-4186, 2019.
[5] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan, “Extracting and matching authors and affiliations in scholarly documents, ” in Proc. of JCDL 2013, pp. 219-228, 2013.
[6] N. V. Cuong, M. K. Chandrasekaran, M. Y. Kan, and W. S. Lee, “Scholarly Document Information Extraction Using Extensible Features for Efficient Higher Order,” in Proc. of JCDL 2015, pp. 61-64, 2015.
[7] A. Prased, M. Kaur, and M. Y. Kan, “Neural ParsCit: a deep learning based reference string parser, ” International Journal on Digital Libraries (IJDL), vol. 19, no. 4, pp. 323-337, 2018.
[8] X. Ma, and E. Hovy, “End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF, ” in Proc. of ACL 2016, pp. 1064-1074, 2016.
[9] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, “Automating the construction of internet portals with machine learning, ” in Proc. of Information Retrieval, vol. 3, no. 2, pp. 127-163, 2000.
[10] 浪越大貴, 太田学, 高須淳宏, 安達淳, “Bi-directional LSTM-CNN-CRF による参考文献書誌情報抽出,” 信学技報, vol. 118, no. 377, pp. 17-22, 2018.
[11] 村田紘基，“日本語学習者の作文からの誤り検出ーラティスに着目した形態素解析失敗の検出ー，” 三重大学大学院工学研究科

博士前期課程電気電子工学専攻修士論文，2022.
[12] M. Rei, and H. Yannakoudakis, ”Compositional sequence labeling models for error detection in learner writing,” Proc. of ACL, pp. 1181-1191, 2016.
[13] 中山竣平，金澤輝一，高須淳宏，上野史，太田学，“ BERT による参考文献書誌情報抽出の誤り検出の評価, ” 第 15 回データ工学と情報マネジメントに関するフォーラム（DEIM2023），1b-6-1, 2023.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “ Attention is All You Need,”in Proc. of NeurIPS, vol. 30, pp. 5998–6008, 2017.
[15] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “ Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, ” in Proc. of the ICCV 2015, pp.19–27. 2015.

LLMにおける自己注意機構の役割の分析

大塚　空来†　　湯本　高行††

† 兵庫県立大学社会情報科学部　〒651-2197 兵庫県神戸市西区学園西町8丁目2-1
†† 兵庫県立大学大学院情報科学研究科　〒651-2197 兵庫県神戸市西区学園西町8丁目2-1
E-mail: †sora.judo.1026@gmail.com, ††yumoto@sis.u-hyogo.ac.jp

あらまし
本研究ではLLMを用いた生成タスクにおいて自己注意機構が生成文にどのような影響を及ぼしているのかを調査する。調査方法として、LLM内の自己注意機構のパラメータ内の値を変更し、自己注意機構の重みを全て同じ値にした時の生成結果を比較することによって、LLMにおける自己注意機構の役割を考察する。具体的な操作方法としては、対象とするDecoder層の1層目の自己注意機構の重みを均一に0にし、1層目の自己注意機構の重みを全て同じ値にする。また、生成タスク用と質問応答タスク用のプロンプトの2つを用いて、それぞれの場合での品詞別のトークンに対する自己注意機構の重みの傾向を比較する。同時に、自己注意機構の重みだけではなく、それらを変換したベクトルのノルムからトークンの重要度を調査する。

キーワード　自然言語処理, Transformer, LLM, Attention

1　はじめに

近年の自然言語処理分野 (以下 NLP とする) において、BERT [1] や T-5 [2] といった事前学習済み言語モデルが幅広い NLP タスクにおいて成功を収めている。そのような NLP タスクにおいて高い精度を実現しているモデルの多くは Transformer [3] の Encoder, もしくは Decoder を多層重ねたモデルである。また、そのような Transformer をベースとしたモデルの成功に寄与していると考えられているのが Transformer モデル内部の**自己注意機構構 (self-attention mechanism)** である。それらのモデルが成功を重ねるとともに、モデルの出力や成功理由を解釈するための自己注意機構の説明性に関する研究が盛んに行われきた [4] [5]。

自己注意機構構 (self-attention mechanism) とは、入力となるベクトル列の中から特に重要だとされるベクトルに注目し、トークンごとの重要度の加重和を取ることによって、トークンの情報を重要性を考慮しながら文脈に沿った意味を算出する機構である。

また、現在は前述した事前学習済み言語モデルに加えて、ChatGPT を始めとする文章生成に特化した**大規模言語モデル (Large Language Models)** が成功を収めている。具体例として、ChatGPT は 1 つのモデルで既存の分類や要約、質問応答などの複数の NLP タスクに対して高い成功を収めるとともに、Zero-Shot でのソースコードや文章生成を行うことが可能となっている。

本研究では、既存の attention weight に関する分析を文書生成を行う大規模言語モデルに対しても行うことで、LLM における自己注意機構の役割を分析する。具体的には、対象とする Decoder 層を変化させながら自己注意機構のパラメータの均一化と入力プロンプトの変更による自己注意機構の変化を見る。また、同時に入力ベクトルを関数によって変換したベクトルのノルムを見ることによって、自己注意機構以外の側面からトークンの重要性を評価する。

本研究の主な貢献は以下の 2 つである。

- LLM における自己注意機構の重みと出力との関係を明らかにする。
- LLM のプロンプト中のどのトークンが重要なのかを明らかにする。

2　自己注意機構の構造

2.1　自己注意機構の概要

Transformer には**自己注意機構**と呼ばれる、入力文における単語間の関連性を抽出するための機構が存在する。言語モデルへ文章が入力される際は、各モデルに対応した tokenizer によって文章がトークン列 $\{t_1, \cdots, t_n\}$ に分割される。その後、自己注意機構へ入力される際は各トークンがさらに埋め込みベクトル列 $\{x_1, \cdots, x_n\}(x)$ に変換されたものが入力される。

ここで、自己注意機構からの出力ベクトル $\boldsymbol{y_i}$ は以下のように計算される。

$$\boldsymbol{y_i} = \left(\sum_{j=1}^{n} \alpha_{i,j} v(x_j) \right) W^O \tag{2.1.1}$$

$$\alpha_{i,j} = \underset{j}{softmax} \left(\frac{q(x_i) k(x_j)^T}{\sqrt{d'}} \right) \tag{2.1.2}$$

$$q(x) = xW^Q + B^Q \tag{2.1.3}$$

$$k(x) = xW^K + B^K \tag{2.1.4}$$

$$v(x) = xW^V + B^V \tag{2.1.5}$$

ただし、バイアス項 B は実装によって存在するものとしないも

のとがある。

ここで計算された attention weight$\alpha_{i,j}$ は文章内のあるトークン x_i と x_j との関係性が数値で表されている。この attention weight が元の入力データに足し合わされることによって、Transformer は入力データの情報を保持しつつ、各層で新しい情報を統合し、より複雑な文脈情報を元の入力データに付け加える形で獲得する事ができる。

2.2 マルチヘッド自己注意機構

また、Transformer 内部のの自己注意機構には**マルチヘッド自己注意機構 (Multi-Head-Attention)** が用いられている。マルチヘッド自己注意機構では、前節での $\alpha_{i,j}$ を計算する際に、$q(x), k(x), v(x)$ を h 個に分割した $q_l(x), k_l(x), v_l(x)$ を用いてそれぞれのヘッドごとに自己注意を算出した後に、それらを結合し全結合層 W^O を通して出力を行う $\left(q_l(x), k_l(x), v_l(x) \in R^{d \times d'/h}\right)$。つまり、マルチヘッド自己注意機構は数式を用いて以下のように表される。

$$\boldsymbol{y_i} = concat(head_1, \cdots, head_h)W^O \tag{2.2.1}$$

$$head_l = \left(\sum_{j=1}^{n} \alpha_{i,j,l} v_l(x_j)\right) \tag{2.2.2}$$

$$\alpha_{i,j,l} = \underset{j}{softmax}\left(\frac{q_l(x_i)k_l(x_j)^T}{\sqrt{d'}}\right) \tag{2.2.3}$$

ここで計算された attention weight$(\alpha_{i,j,l})$ は、あるヘッド l において獲得された文章内のあるトークン x_i と x_j との関係性が数値で表されている。このような構造によって、各ヘッドがそれぞれ違う文脈表現を獲得することで、より複雑な意味を捉えることができているとされている。

2.3 Decoder ベースの LLM における自己注意機構

Transformer は Encoder と Decoder からなるモデルであり、Encoder はクラス分類などのための特徴抽出、Decoder は入力文の続きとなる単語の生成を行うことができる。LLM はこの Decoder を重ねて構成されており、Transformer の Decoder を用いて、ある単語列 $\{x_1, x_2, \cdots, x_L\}$ が与えられたときに、$p(x_{L+1}|x_1, x_2, \cdots, x_L)$ の確率が最も大きい x_{L+1} を次単語として予測し、再帰的に文章を生成する**自己回帰言語モデル**である。

ここで Transformer の Decoder においては Encoder と同様に単語間の関連性を自己注意機構によって抽出し、それを元に単語の生成確率をモデル化している。しかし、このような Transformer の Decoder 部分において自己注意機構の仕組みは Encoder のものと少し異なっている。具体的に言うと、Decoder については入力テキスト内のあるトークンから見て未来にあるトークンについては Attention を計算しないようにマスクをしている。これは、Decoder は入力から未来のテキストを予測する機構であるため、いわゆる「カンニング」を防ぐ目的で行われている。実際の内部処理では、α を計算する処理の中で softmax を取る前に該当する要素に大きな負の値を足すことによって、マスクを実現している。

また、Transformer が Encoder-Decoder モデルであるのに対して、LLM では Decoder のみが使用されているため、Transformer の Decoder とは異なった 2 つの特徴を持っている。

1 点目は、入力テキストと出力テキストをまたがっている Cross-Attention が LLM には存在せず、Self-Attention と Feed Forward で構成されているという点である。そのため、LLM においては Decoder に文章が入力されると、まずはその文章内のトークン間の関連性のみが Self-Attention で計算される。その後、その値を Feed Forward によってさらに高次元な表現に変換することで複雑な特徴を学習し、次の Decoder 層に渡すという流れが Decoder 層の数だけ繰り返される構造となっている。

2 点目は、モデルへの入力はプロンプトを含めたテキストであるということである。LLM においてはモデルへの命令をプロンプト内に記述して入力することによって、そのプロンプトに続く単語予測を行いそのプロンプトに対する出力が生成される。このような構造であるため、LLM は Decoder 単体でもプロンプトと出力との間で Cross-Attention のように文脈を表現できている。

2.4 自己注意機構の分析に関する先行研究

既存の研究において、Transformer モデルをベースとしたモデルにおける自己注意機構を解析する方法は盛んに行われている。自己注意機構を解釈することを目的した典型的な分析方法は、自己注意機構が割り振る「重み」に着目する分析である。この分析では、BERT などの事前学習済み言語モデルを用いて何らかのタスクを解き、その際の attention weight をもとに「モデルがシーケンスの中のどのトークンに注目しているのか」を解析し、解釈する。

ただし、この attention weight による解釈可能性には未だに議論の余地が残っている。Jain らは、 attention weight と勾配ベースの特徴量の重要度との相関を調査することによって、自己注意機構の解釈可能性を調査した [4]。その調査の結果、両者の間に相関は見られなかったために、attention weight は解釈可能性を持たないという考察をした。また、元の attention weight と最大限に異なるが予測にほとんど変化をもたらなさい自己注意機構の敵対的な重みを生成することによって、解釈可能性を否定した。

この手法では、特定の NLP タスクに特化したデータを用いて学習させることで、1 つのモデルを 1 つのタスクに特化させる Fine-Tuning ベースのモデルが使用されていた。その場合においては、自己注意機構以外のフィードフォワードネットワークや残差接続といった機構もそのタスクに特化するように学習されることから、自己注意機構が異なるが予測にはほとんど影響がないような重みが存在することには異論はない。しかし、LLM のような In-Context-Learning のモデルでは、必ずしも 1 つのタスクに最適化されるように学習がなされる訳ではない。そのため、「In-Context-Learning モデルでは文章内のある単語へ注目する機構である自己注意機構の役割が大きい」と著者ら

は考える。そこで、この研究では第一段階として自己注意機構のパラメータを変化させることによる生成内容の精度変化を調査することによって、Fine-Tuning ベースのモデルとの差異を調査する。

また、Kobayashi らは attention weight に注目するのではなく、変換したベクトルのノルムの大きさを基に重要度を分析することで、重要度を分析する方法を提案している [6]。この分析によると、attention weight による分析結果では，出力ベクトルに対する特殊トークンの寄与が大きいという直感に反する傾向が見られていたが，ベクトルの大きさに注目した分析では，特殊トークンの寄与は小さいという結果が得られた。

この研究では、Encoder ベースの事前学習済み言語モデルである BERT が用いられていた。今回研究対象とする LLM は Decoder ベースであるため、その内部の自己注意機構についても Encoder ベースのものとは異なる。そこで、第二段階として LLM における自己注意機構の役割をその重みではなく、ベクトルの大きさからも調査を行う。それによって、Encoder ベースのモデルとの差異を測定するとともに、LLM における自己注意機構内の重み以外の観点からの寄与についても調査する。

3 LLM における Attention Weight の分析

3.1 分析手法

自己注意機構を分析する先行研究 [6] では，Encoder ベースのモデルにおいてある特定のタスクにおける自己注意機構の大きさを調査した。そこで、本研究では In-Context-Learning で推論を行う Decoder ベースの LLM に対して、その自己注意機構の値が及ぼす影響を調査する。具体的な手法としては、まずは質問応答タスクに対して自己注意機構を変化させずに推論を行う。その後、同じタスクに対して各層の自己注意機構の Query, Key に対応する重みを W^Q, W^K 一律な重みに付け替えた状態で再度推論を行う。具体的には W^Q, W^K を 0 にすることで任意の i, j について、$q(x), k(x)$ の値を 0 にする。この際の数式は以下のように表される。

$$\boldsymbol{y_i} = \left(\sum_{j=1}^{n} \alpha_{i,j} v(x_j) \right) W^O \tag{3.1.1}$$

$$\alpha_{i,j} = \underset{j}{softmax} \left(\frac{q(x_i) k(x_j)^T}{\sqrt{d'}} \right) \tag{3.1.2}$$

$$W^Q = O \tag{3.1.3}$$

$$W^K = O \tag{3.1.4}$$

$$q(x) = \mathbf{0} \tag{3.1.5}$$

$$k(x) = \mathbf{0} \tag{3.1.6}$$

ここで、任意の i, j について、$q(x), k(x)$ の値が 0 である時、(4.2) の softmax に入る値が全て 0 になることから、この場合の attention weight は全て同じ値となる。

この実験を通して、まずは自己注意機構における生成精度の違いを確認する。次に、プロンプト内の命令だけを要約や質問応答などの別タスクに用に変更した上で、その出力結果と attention weightα を確認する。これによって、プロンプトの命令によって attention weight が大きい単語に有意な変化が見られたら、自己注意機構に一定の解釈可能性が存在するといえる。

3.2 attention weight の変更に伴う生成内容の変化の分析

3.2.1 実験方法

自己注意機構の生成への寄与を確認するために、次の 2 通りの方法で実験を行う。

まずは、事前学習された重みをそのまま用いて SQuAD データセットからランダムに選んだ 300 文について質問応答タスクを解き、LLM の出力に正解が含まれる割合を正解率として精度を算出する。次に、自己注意機構における Query, Key に対応する重み W^Q, W^K をすべて 0 に変更した後に、同様にタスクに対して出力することで精度を算出する。ここで変更対象とする層は、1 層目と最終層とし、これらの重みの変更はそれぞれ別で行う。1 層目を変更させた場合は入力に最も近い 1 層目の変更は 2 層目以降の自己注意機構によって補正され、結果にあまり影響しないという仮説に基づいている。これによって、自己注意機構のパラメータを変更したときに生成の質 (精度) にどの様な変化があるのか、及びどれほど変更すれば生成に影響が出るのかを調査する。

3.2.2 実験設定

実験には LLM モデルとして Llama2 [7] のモデルの 1 つである Llama2-7b-chat を用いる。Llama2-7b-chat は Meta 社によって提供されているオープンソースのモデルであり、パラメータ数は 70 億、1 層あたりのヘッドは 32 個で、層は 32 層から成る。モデルの事前学習には 2 兆個のトークンが使用され、さらに人間のフィードバックからの強化学習によって、人との対話に特化した形に改良されている。

また、データとして SQuAD データセット [8] を用いる。SQuAD データセットは、英語の質問応答タスクに対応したデータセットで、それぞれのデータは context, question, answer の 3 つからなっている。モデルは与えられた自然文の質問である question に対して、回答自体を context から検索して出力する。その回答精度でモデルの評価を行う。具体例としては以下の通りである。

> **context**: Beyoncé Giselle Knowles-Carter ((略) / bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child,
> ··· (中略) ···
> Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles ”Crazy in Love” and ”Baby Boy”.
> **question**: When did Beyonce start becoming popular?

answer: in the late 1990s

また、入力するプロンプトは以下の通りである。

<s>[INST]<<SYS>>You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. <</SYS>>
Extract from the following context only the minimal span word for word that best answers the question. You must not explain your way to think.
Context: {context}
Question: {question} [INST] </s>

3.2.3 実験結果

事前学習された重みを用いて質問応答タスクを解いた際の精度と、LLM の 1 層目のみの重みを一律に変更した際の結果を表 1 に示す。

表 1　SQuAD データセットにおける LLM の生成精度

変更点	正解数	全データ数	正解率
変更なし	187	300	62.3%
1 層目	23	300	7.7%
最終層	180	300	60.0%

実験結果より、LLM において 1 層目の重みのみを変更した場合は最終層の重みのみを変更した場合に比べて大幅に精度が低下した。特に、最終層のみを変更した場合は変更なしの結果に比べて正解率がわずか 2.3%しか低下しなかった。この結果から、LLM においては 1 層目に不具合があってもその後の層で調整できるような機構は存在せず、むしろ序盤の層で得られた自己注意を更に後ろの層で補足していくという意味合いが強いと考えられる。

3.3 プロンプト内の指示の変化による attention weight の変化傾向の分析

3.3.1 実験方法

本実験では同じ context, question に対して、質問応答タスクと要約タスクという 2 つの指示を対応付けてそれぞれの自己注意に重さの違いを比較する。具体的には、入力プロンプトとして、質問応答タスクに対応するプロンプトと、要約タスクに対応するプロンプトの 2 通りを用意して、それぞれをモデルに入力することによって、2 つの違いを検証する。この章で具体的に入力するプロンプト内の指示分はタスク別で以下のとおりである。

質問応答

Extract from the following context only the minimal span word for word that best answers the question. You must not explain your way to think.

要約

Summarize from the following context only the minimal span word for word. You must not explain your way to think.

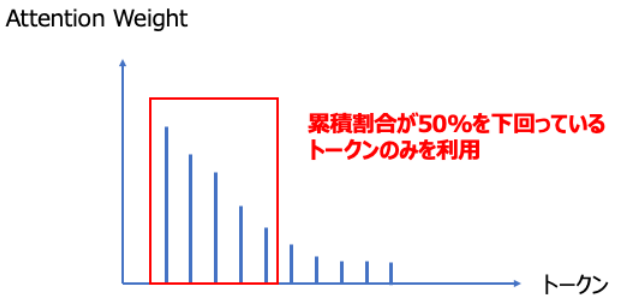


図 1　トークンの累積和のイメージ

この実験において分析対象とするトークンは </s>である。このトークンに注目した理由は、このトークンが入力プロンプトの一番最後のトークンであり、このトークンを元にモデルがその続きを出力するからである。質問応答タスクと要約タスクという 2 つの異なるタスク間での違いを確認するために、より生成内容と関連が高い </s>トークンを今回は分析の対象とする。

まずはプロンプトを入力して、各層のヘッドごとに入力末尾のトークンである </s>と各単語との attention weight を抽出する。次に、attention weight を層ごとにヘッドで平均を取ることで、各層でのトークンごとの attention weight の平均値を算出する。次に、トークンごとに品詞を付与し、名詞、形容詞、副詞、前置詞、動詞の単語以外を削除する。その後、層ごとに attention weight でトークンを降順に並び替え、累積和を取り、累積和が全体の 50%以内であるトークンのみを抜き出し、その中における各品詞の割合を算出する。ここで累積和が全体の 50%以内であるトークンを抜き出している理由は、attention weight の値が大きく、より生成に対する特徴が現れていると考えられるトークンを重点的に分析対象とするためである。最後にすべての層における品詞の割合の平均値を取る。

上記処理を行った後に、質問応答プロンプトと要約プロンプトのそれぞれでに全ての入力プロンプトにおける品詞の割合の平均値を算出し、それらを比較する。また、全ての入力プロンプトにおける質問応答プロンプトと要約プロンプトの差分の平均値を算出し、差分の傾向を明らかにする。

3.3.2 実験設定

本実験では実験 1 と同様に、LLM モデルは Llama2-7b-chat、データは SQuAD データセットを用いる。

また、この章では品詞を以下のように定義する。

- 単語が ['</s>', '<s>', '<<', '>>', 'INST', '[', ']', '/'] に含まれる場合は「特殊トークン」
- 単語が ['%', '&', '!', '?', ':', ';', '$', '-'] に含まれる場合は「記号」
- 単語が ['', '\n', ' '] に含まれる場合は「空白」
- 単語が Python のライブラリである 'nltk'のストップワード辞書に含まれる場合は「ストップワード」
- それ以外の単語は nltk の品詞分析で割り当て

られた品詞

3.3.3 実験結果

質問応答における割合の平均、要約における割合の平均、差分の平均の結果を表 4.2 に示す。

表 2　品詞ごとの分布割合の平均 (%)

品詞	質問応答タスク	要約タスク	差
前置詞	0.43	0.41	-0.02
動詞	5.67	9.19	3.52
名詞	87.94	84.06	-3.88
形容詞	4.27	4.81	0.53
副詞	1.68	1.53	-0.15

この結果より、質問応答タスクでは要約タスクに比べて名詞に対して attention weight がより大きく割り振られていることがわかる。逆に、要約タスクでは質問応答タスクに比べて動詞に対して attention weight がより大きく割り振られていることがわかる。

これは、質問応答タスクにおいては質問に対する解答となる名詞が重要で、逆に要約タスクにおいては誰が何をしたかという動詞までが重要であるということがそのまま attention weight に反映されていると考えることができる。

3.4 考　　察

1 つ目の実験において、LLM の重みを一律に変更したところ、1 層目を変更した場合は大幅に精度の減少が見られたが、最終層を変更した場合はそのような大幅な精度の減少が見られなかった。この結果から、LLM においてはトークンを埋め込んだ入力ベクトルである $\boldsymbol{x}$ の情報から、入力に近い層で生成の方向性を定め、それを補足する形で出力に近い層でより詳細な文脈情報を獲得するという考察が得られた。

また、2 つ目の実験において、入力プロンプトの指示文を変更することによって、その自己注意のかかり方に違いが発見された。具体的には、質問応答タスクに対応した入力の場合はようやくタスクに比べて名詞に多く自己注意がかかり、要約タスクに対応した入力の場合は、質問応答に比べて動詞に多く自己注意がかかった。ここから、LLM においてはプロンプトに応じて自己注意機構が別の部分から文脈情報を獲得している可能性が示唆された。

この 2 つの結果から、LLM においてはその自己注意機構のパラメータを変更すると精度に増減が見られるとともに、入力プロンプトの差によって自己注意機構の品詞別のかかり方に違いが出た。ここから、LLM における自己注意機構はそれ単体である程度の解釈性があると考えることができる。ただし、今回のモデルについては入力に近い層によりその傾向が強いと考察できる。

4 LLM における Attention Vector の分析

4.1 分析手法

1 つ目の実験において、1 つ目の層における自己注意機構のパラメータを変更すると、その精度は大幅に低下した。このことから、入力に近い部分の層が大まかな LLM の出力の方向性を決定していると考えることができ、自己注意機構のパラメータが生成精度に大きな影響を及ぼすことがわかる。

一方で、attention weight についてはその説明性については明確になっておらず、特に「,」や「.」などの記号が高い重みを持つことが知られている [6]。

そこで、次は自己注意機構が与える影響をその重み以外の側面から調査する。具体的な手法としては、前述した質問応答タスクに対応するプロンプトを入力し、次の関数 f で 1 層目への入力である埋め込みベクトル $\boldsymbol{x}$ を変換する。関数 f は Kobayashi らによって提案された次の関数を用いる [6]。

$$\begin{aligned}\boldsymbol{y_i} &= \left(\sum_{j=1}^{n} \alpha_{i,j,l} v_l(x_j)\right) W^O \\ &= \sum_{j=1}^{n} \alpha_{i,j,l} f_l(x) \qquad (4.1.1)\end{aligned}$$

$$\boldsymbol{f_l(x)} = v_l(x_j) \cdot W^O \qquad (4.1.2)$$

次に、$\boldsymbol{f(x)}$ のノルムの大きさをトークンごとに比較することによって、attention weight ではなく、そのノルムからトークンの重要度を調査する。

4.2 実　　験

4.2.1 実験方法

前章の context, question を前章と同様のプロンプトに埋め込んで、LLM に入力する。その際の attention weight と、Query, Key, Value, Output 行列の重み W^Q, W^K, W^V, W^O を取り出して、そこから 1 つ目の層におけるヘッドごとの $\alpha, ||\boldsymbol{f(x)}||$ を算出し、それらの平均値を算出する。そこで、各トークンにおけるノルムの大きさをと α の大きさを比較することによって、トークンの重要性の考察を行う。

4.2.2 実験設定

前章の実験と同様に LLM モデルは Llama2、データセットは SQuAD データセットを用いた。データは SQuAD からランダムに 50 件を抜き出したものとし、プロンプトについては全章で用いたものを使用する。

また、この章では品詞を以下のように定義する。

- 単語が ['</s>', '<s>', '<<', '>>', 'INST', '[', ']', '/'] に含まれる場合は「特殊トークン」
- 単語が ['%', '&', '!', '?', ':', ';', '$', '-'] に含まれる場合は「記号」
- 単語が ['', '\n', ' '] に含まれる場合は「空白」
- 単語が「Python のライブラリである ‘nltk‘のストップワード辞書から「前置詞」「副詞」「人称代名詞」「関係代名詞」を抜いたもの」に含まれる場合は「ストップワード」
- それ以外の単語は nltk の品詞分析で割り当てられた品詞

この実験においては品詞は全 13 種類である。

4.2.3 実験結果

attention weight とノルムにおいて値が下位 5 件であるトークンを表 5.1 に示す。これをみると、attention weight では most や?, what といったものの値が小さくなっているのに対して、ノルムの方では「改行記号」や「,」「.」といった区切り文字の値が小さくなっている。このように、ノルムでみると文中でそこまで意味を持たない区切り文字のノルムが小さくなることがわかる。

表 3 attention weight とノルムの大きさ下位 5 件

順位	attention weight	ノルム
1	most	\n
2	pl	.
3	iful	,
4	?	ate
5	what	ers

次に、attention weight とノルムにおいて品詞別で値を平均したときの値の降順での上位 5 件を表 5.2 に示す。

表 4 値の降順上位 5 件での品詞

順位	attention weight	ノルム
1	特殊トークン	関係代名詞
2	空白	特殊トークン
3	区切り文字	記号
4	副詞	人称代名詞
5	ストップワード	副詞

表 5.2 より、attention weight の大きさで比較すると特殊トークンや空白、区切り文字が大きさの平均で上位に来ているのに対して、ノルムの方では関係代名詞や人称代名詞、記号が上位に来ている。例えば、自己注意機構の方を見るとストップワードが上位 5 件に入っているのに対して、ノルムでは入っていない。これは、attention weight が大きくなる傾向にある情報量が少ない単語は、ノルムを計算したときには値が小さくなるということが言える。

最後に、品詞別に attention weight、ノルム別の平均の大きさ降順で並べて順位を比較したものを表 5.3 に示す。

表 5 各品詞の自己注意の重みとノルムを降順に並び替えたときの順位

順位	attention weight	ノルム
特殊トークン	1	4
空白	2	12
区切り文字	3	13
ストップワード	5	10
記号	12	5

この表をみると、attention weight においては空白や区切り文字、ストップワードなどの情報量が少ないとされる文字の大きさが他の品詞に比べて大きいのに対して、ノルムではストップワード、空白、区切り文字における値が全ての品詞の種類が 13 種類中、下位 5 件以内に入っている。

このように、LLM においても $\boldsymbol{f}(\boldsymbol{x})$ によって変換したベクトルのノルムは、自己注意機構に比べてより単語の情報量を表していると考えられる。

4.3 考察

本手法では、LLM の Decoder 層の 1 層目を対象として、attention weight ではない方法で単語の重要性を表す方法を検証した。結果を比較すると、自己注意機構においては空白や区切り文字、副詞、ストップワードといった情報量が少ないであろう品詞の単語に対して自己注意が多くかかるという結果が得られた。それに対して、入力ベクトル $\boldsymbol{x}$ を変換した $\boldsymbol{f}(\boldsymbol{x})$ のノルムを見ると、それらの情報量が少ない単語のノルムは他の品詞に比べてかなり低いという結果が得られた。このことから、attention weight は情報量が少ない単語に値を多く割り振ることによって他に情報量が多い単語がない場合の目印としている可能性が考えられる。また、それに対してベクトル $\boldsymbol{f}(\boldsymbol{x})$ を計算することによって、文章内で重要ではない区切り文字や空白などの単語を特定できると考えられる。

5 おわりに

本稿では、過去の自己注意機構に対する解釈手法を元に、In-Context-Learning で推論を行う Decoder ベースの LLM に対してその自己注意機構の役割を分析した。

質問応答タスクにおける生成精度の検証では、自己注意機構のパラメータを層別に一定にすることによって、出力結果に差異が出て、質問応答タスクに対する出力精度の低下が確認された。具体的には、1 層目の重みを変更した場合は出力精度が大幅に低下し、逆に最終層の重みを変更した場合は出力精度に有意な低下が見られなかった。つまり、LLM においては入力層に近い部分が文章の意味と出力の方向性を決定し、最終層に向かうにつれて、それを補正していく役割があると結論づけることができる。

また、プロンプトを質問応答、要約と 2 種類用意してその attention weight を確認した際は、その自己注意のかかり方に品詞ごとの違いが見られた。これは、LLM が事前学習の段階でプロンプトの指示に対応できる形で自己注意の重みを学習しているからであると考えられる。

また、自己注意機構における変換後ベクトルのノルムについての解析では、「改行記号」や「,」といった意味を持たないトークンにおけるベクトルのノルムが小さいことは確認できた。また、品詞別に見たときにその様な情報量の少ない品詞に対して自己注意は大きい値を出すが、ノルムで見ると小さい値になるということも確認できた。そのため、LLM においてもこのノルムの大きさで単語の重要性を測ることによって、自己注意機構だけでは測れなかった単語の重要性を見ることができると結論づけられる。

文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova “BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding," arXiv:1810.04805 [cs.CL], 2018.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv:1910.10683 [cs.LG], 2019.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin "Attention Is All You Need," arXiv:1706.03762 [cs.CL], 2017.

[4] Sarthak Jain, Byron C. Wallace "Attention is not explanation," arXiv:1902.10186 [cs.CL], 2019.

[5] Sarah Wiegreffe, Yuval Pinter "Attention is not not explanation," arXiv:1908.04626, 2019.

[6] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui "Attention is Not Only a Weight: Analyzing Transformers with Vector Norms," In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7057–7075, Online. Association for Computational Linguistics, 2020.

[7] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288 [cs.CL], 2023.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang "SQuAD: 100,000+ Questions for Machine Comprehension of Text," In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics, 2016.

k近傍法を用いた事前学習済みBERTモデル検索

ファムフーロン† 三林 亮太† 山本 岳洋† 加藤 誠†† 山本 祐輔†††
荘司 慶行†††† 大島 裕明†

† 兵庫県立大学 情報科学研究科 〒651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 筑波大学 図書館情報メディア系 〒305–8550 茨城県つくば市春日 1–2
†††† 静岡大学 情報学部 〒432–8011 静岡県浜松市中区城北 3–5–1
††† 名古屋市立大学 データサイエンス学部 〒467–8501 愛知県名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †{af23a009,af22h007}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp,
†††mpkato@slis.tsukuba.ac.jp, ††††yusuke_yamamoto@acm.org, †††††shojiy@inf.shizuoka.ac.jp,
††††††ohshima@ai.u-hyogo.ac.jp

あらまし 本研究では，特定の文書分類タスクにおいて良い性能を出すBERT事前学習済みモデルの検索を行う手法を提案する．自然言語処理の問題において，新たにモデルを学習する代わりに，既存の事前学習済みモデルをファインチューニングして用いることが一般的である．しかし，多数の事前学習済みモデルからタスクに適したモデルを選択することは困難である．多数のモデルを用いてファインチューニングを行うには，時間と計算資源が多く必要となる．そこで，我々は，k近傍法を用いることで，ファインチューニングを行うことなく，BERT事前学習済みモデルを検索する手法を提案する．17の文書分類タスクにおいて，20のBERTモデルを対象とし たベンチマークデータセットを構築し，実験を行った．

キーワード BERTモデル検索，言語モデル，検索モデル，テキスト分類

1 はじめに

事前学習済みBERTモデルは，文書分類，質問応答など，多岐にわたる自然言語処理のタスクで用いられている．新たにモデルを学習する代わりに，既存の事前学習済みモデルをファインチューニングして用いることが一般的である．しかし，特定のタスクに最も適した事前学習済みのBERTモデルを選択することは簡単ではない．多数の事前学習済みモデルの中から，タスクに適したモデルを選択する必要がある [1]．特定のタスクに対する事前学習済みBERTモデルの適合性は，ファインチューニングを通じて評価することが一般的であるが，ファインチューニングは計算コストが高い．特定のタスクに適したモデルを見つける一般的なアプローチは以下の通りである．

（1） 多数の事前学習済みモデルの中からいくつかの候補を選択する

（2） 1で選択したモデルをタスクの一部のデータ（例えば訓練データ）を用いてファインチューニングを行う

（3） タスクの残りのデータ（例えばテストデータ）を用いて，2でファインチューニングを行ったモデルの性能を評価する

（4） 最も性能が良いモデルを採用する

上記のアプローチにはいくつかの課題がある．第一に，ファインチューニングは計算リソース及び時間のコストが高い．第二に，事前学習済みモデルの候補を選択する段階で，最もタスクに適しているモデルが含まれてない可能性がある．そのため，事前学習済みBERTモデルを効率的に検索する必要がある．

本研究では，事前学習済みモデルの検索問題を定義する．例えば，図1に示されているように，医療文書の分類タスクの場合，事前学習済みBERTモデルの中から，最も性能が良いMed-BERTを検索する．また，ファインチューニングを行うことなく，k近傍法を用いることで，BERT事前学習済みモデルを検索する手法を提案する．さらに，17の文書分類タスクにおいて，20のBERTモデルを対象としたベンチマークデータセットを構築し，実験を行った．本研究の貢献は以下の通りである．

（1） 事前学習済みBERTモデル検索問題の問題定義

（2） 上記問題の評価用ベンチマークデータセットの構築

（3） k近傍法を用いた事前学習済みBERTモデルの検索手法の提案

2 関連研究

事前に大規模なデータセットで訓練された事前学習済みモデルの利用は，現代の機械学習の基盤である．このような事前学習済みモデルは，タスク固有のパラメータでのファインチューニングから大きな恩恵を受けることが示されており，様々なタスクの性能を向上させている．たとえば，CIFAR-10のような

1：訳注：2023年11月11日時点で，Hugging Faceには19,503の事前学習済みBERTモデルが利用可能である

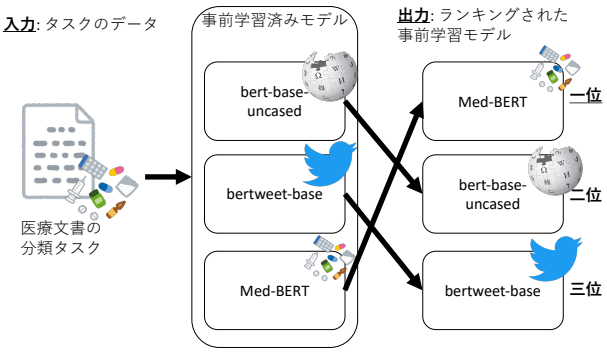


図 1　事前学習済み BERT モデル検索の例．解決すべきタスクは医療文書の分類である．最高の性能を持つ Med-BERT が検索結果の先頭に表示されるべきである．

画像データセットで事前学習されたモデルは，画像分類タスクでの効果が増大していることが示されている．近年では，特定分野の事前学習済みモデルの構築が，特定の分野内のタスクで顕著な性能向上につながっている．これらのモデルは，分野関連のデータセットで事前学習されており，一般的なデータセットで訓練されたモデルよりも優れた性能を発揮する．

BERT [7] は，Transformer アーキテクチャ[22] のエンコーダーの部分を利用する言語モデルであり，大規模なテキストデータセットでの広範な事前学習の重要性を示している．BERT の改良版である RoBERTa [18] は，事前学習中に次の文章予測タスクを取り除き，事前学習用のテキスト量を増やすことで，元のモデルを改善し，後続のタスクの性能向上が実証されている．一方で，ALBERT は，元の BERT と比較してモデルサイズを削減しつつ，性能を維持し向上させることに焦点を当てている [16]．DeBERTa は，BERT および RoBERTa モデルを，解きほぐされた注意メカニズム，強化されたマスクデコーダー，および新しい仮想敵対訓練法を使用して改善している [10]．さらに，この分野ではモデル蒸留技術への関心が高まっている．Sanh ら [21] は，事前学習フェーズ中に知識蒸留を成功裏に実装し，BERT モデルのサイズを大幅に削減しながら言語理解能力を維持し，速度を向上させることに成功した．

さらに，ドメイン固有のデータを用いて追加学習を行うことで，特定のドメインに特化したモデルに関するいくつかの提案がなされている．例えば，大規模な多分野の科学論文コーパスを使って事前学習された SciBERT [2] は，下流の科学的 NLP タスクでの性能が向上したことが示されている．金融データで事前学習およびファインチューニングされた FinBERT [1] は，金融センチメント分析データセットにおいて大幅に性能を改善した．Caselli ら [4] は，英語の Reddit コメントの大規模データセット RAL-E を使用して HateBERT を再トレーニングし，悪用言語検出タスクで一般的な BERT モデルを上回る性能を発揮した．Chalkidis らは，LEGAL-BERT [6] を複数分野の英語法律テキストで事前学習することで，法律分野の 3 つのエンドタスクで最先端の結果を達成した．医療分野では，BioBERT [17] や ClinicalBERT[[11] のような提案があり，それぞれドメイン固有の文献と匿名化された患者記録を利用している．医療コーパスで事前学習された Med-BERT [20] は，一般的なモデルと比較して病気予測能力が優れていることが示された．さらに，インターネットベースの言語のニュアンスへの事前学習済みモデルの適応は，成長する関心の領域であり，オンライン意見用の IMHO [5] や Twitter データ用にファインチューニングされた RoBERTa ベースのモデル BERTweet [19] などのモデルの作成につながっている．これらのモデルは，モデルトレーニングにおけるドメイン固有の学習への傾向を強調し，そのような集中的な強化が特定のアプリケーションに対するより効果的なツールにつながることを示している．

Han ら [9] は，シーケンスラベリングタスクを新しい難しいドメインに教師なしで適応させるために，ドメイン適応ファインチューニングが単純だが効果的な手法であることを示した．一般に，事前学習およびファインチューニングは通常，かなりの量のデータを必要とする．近年，データセットの検索に焦点を当てた研究が増えている．例えば，加藤らは，アドホックなデータセット検索用に特別に設計された新しいテストコレクションを導入した [14]．Kacprzak らは，4 つの国のオープンデータポータルからのクエリログに関する詳細な分析を行った [12]．Gregory らは，観察研究データの検索と評価における一般的なユーザー行動を特定するために，さまざまな分野のデータ検索文献を調査した [8]．さらに，Kern らは，文献検索とデータセット検索の間の明確な違いを強調し，データセット検索では文献検索に比べてメタデータの量と品質がより重要な役割を果たすことを指摘した [15]．

データセット検索に関する広範な研究にもかかわらず，モデル検索に関する研究には顕著なギャップが存在する．コンピュータビジョン分野では，Karimi らは *Model Picker*，オンラインモデル選択のアプローチを導入した [13]．この手法は，ラベルなしのデータ例のストリームを使用して，事前学習済み分類器のセットから最も適したモデルを選択するという課題に巧みに対処する．You らは，特定のタスクのための事前学習済みモデルの評価を扱い，これらのモデルによって抽出された特徴からのラベル証拠の最大値を推定することに焦点を当てた [23]．NLP 分野では，Zou らは，ウェブ検索での意味マッチングのための効果的な検索モデルの構築と実装の複雑さを調査した [24]．

3　問題定義

3.1　問題定義

事前学習済み BERT モデルの検索問題を次のように定義する．テキストとラベルのペアの訓練データ $D^{train} = \{x_i^{train}, y_i^{train}\}_{i=1}^{M}$ と，検証データ $D^{val} = \{x_i^{val}, y_i^{val}\}_{i=1}^{N}$ と，テストデータ $D^{test} = \{x_i^{test}, y_i^{test}\}_{i=1}^{O}$ を持つ特定のタスクをクエリとする．訓練データ D^{train} と検証データ D^{val} のみを利用し，事前学習済み BERT モデルのリスト $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ を $eval(D^{test}, fine\text{-}tune(m_i))$ の高い順で検索する．$eval$ 関数は，ファインチューニングを行った BERT モデル m_i の D^{test} セットに対する性能を示す．本研究では，$eval$ として正解率を使用する．

3.2 事前学習済み言語モデル検索技術の分類

事前学習済み BERT モデルの検索問題は以下の 3 つのカテゴリーに分類することができる．

ファインチューニングあり: このカテゴリーの手法では，特定のタスクに適した事前学習モデルを検索するためにファイチューニングを実施する．通常，ファインチューニングを検証データにおける損失が最小になるまで行うことが一般的であるが，ある程度性能がわかるタイミングでファイチューニングを止めることで，計算コストや時間の削減ができる．

Inference あり：このカテゴリーでは，検索対象の事前学習済みモデルにタスクのデータを入力し，それぞれのモデルから得られたベクトルを用いてモデル間の性能を比較する．第 5 章で詳述する提案手法は，このカテゴリーに属している．

Inference なし：このカテゴリーでは，ファインチューニングも，事前学習済みモデルににタスクデータを入力することも実施しない．そのため，他のカテゴリーと比べて効率が良いが難易度が最も高い．

3.3 事前学習済み BERT モデル検索のための提案手法

本研究は，**Inference あり**カテゴリーの手法を提案する．ファインチューニングは行わず，各事前学習済みモデルを使用して訓練データ D^{train} と検証データ D^{val} のベクトルを計算し，これらのベクトルを用いて検索対象モデルのランキングを行う．

ファインチューニングしていない状態でも，特定のタスクにおいて良い性能を示す事前学習済みモデルは，そのタスクに適している可能性が高いと思われる．本研究は特定のタスクの訓練データ D^{train} および検証データ D^{val} を用いてファイチューニングを行うことなく，事前学習済みモデルの性能を評価し，その評価結果を元にモデルのランキングを行う．詳細については第 5 章にて説明する．

4 ベンチマークデータセット

本研究では，17 の文書分類タスクと 20 の異なる事前学習済み BERT モデルから，事前学習済み BERT モデル検索問題向け評価用ベンチマークデータセットを構築した．それぞれのタスクに対する事前学習済み BERT モデルのランクリスト $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ を以下の通り作成した．

（1） タスクの訓練データと検証データを用いて，それぞれの事前学習済みモデルのファインチューニングを行う．

（2） タスクのテストデータを用いて，ファインチューニングされた各モデルの性能を正解率で評価する．

（3） テストデータにおける各モデルの正解率を元にモデルのランキングを行う．

ファインチューニングは 4 つの Tesla V100 GPU で行った．バッチサイズは 8 から 32 で，損失関数にはクロスエントロピーを使用し，学習率は 2×10^{-5} とした．エポック数は 100 と設定しているが，10 エポックが経過しても検証データにおける損失が下がらなかった場合，ファイチューニングを中止させる．

4.1 文書分類タスク

ベンチマークデータセットの文書分類タスクの一覧を表 1 で示している．これらのタスクは Hugging Face で公開されており，次の URL からアクセスすることができる：https://huggingface.co/datasets/< タスクの名前 > ．それぞれのタスクはラベル数とデータ数が異なる．各タスクのデータを次の 3 つのデータセットに分割した：**訓練データ**，**検証データ**，**テストデータ**．第 3 章に記載しているように，特手のタスクにおけるモデルの実際の適切さはテストデータで評価される．一方，事前学習済みモデルの検索は，訓練データと検証データのみを使用する．各タスクのデータは以下のように分割した．

- タスクが Hugging Face で既に訓練データ，検証データ，テストデータに分割されている場合，そのまま利用する．
- 訓練セットとテストセットのみに分割されている場合，訓練データの 10%を検証データとして利用する．
- 訓練データのみの場合，訓練データの 10%づつを検証用とテスト用として利用する．

4.2 事前学習済み BERT モデル

検索対象の事前学習済み BERT モデルの一覧を表 2 で示している．これらのモデルも Hugging Face で公開されており，次の URL でからアクセスすることができる：https://huggingface.co/< モデルの名前 >．これらの事前学習済みモデルは，アーキテクチャ（BERT，RoBERTa など），パラメータの数，事前学習およびファインチューニングに使用されたデータが異なっており，特定の文書分類タスクに対するモデルの適合性に影響を与える．例えば，医療文書の分類タスクでは，医療文書で訓練されたモデルが，映画レビューなど他の種類のデータで訓練されたモデルと比較して，そのタスクにより適している可能性がある．

4.3 ベンチマークデータセットの特徴

図 2 は，各モデルの平均ランクを昇順で示している．モデルの間には明確な性能の違いが示されている．例えば，$\text{BERT}_{\text{LARGE}}$ のようなパラメータ数が多いモデルは，よりパラメータ数が少ない $\text{BERT}_{\text{BASE}}$ と比較して，より高いランクを示す傾向にある．さらに，同じパラメータサイズを持つモデルの中でも，RoBERTa や DeBERTa のような新しいアーキテクチャのモデルは，従来の BERT モデルと比較してより性能が良い．

図 3 は，各タスクのテストセットにおけるファインチューニングされたモデルの相対的な正解率のヒストグラムを示している．この図で示している正解率は，特定のタスクにおける一番高い正解率が 1.0 になるように正規化されている．各タスクにおけるモデル間の性能には顕著な差異がある．例えば，sms_spam タスクの場合，ほとんどのモデルの性能が類似しており，このタスクにおいてはモデルの選択が大きな影響を与えない可能性があることを示している．一方で，ethos/binary のようなタスクでは，モデル間の正解率にに大きな差異があり，最適な結果を得るために適切なモデルを選択することが重要であることを示している．

表 1　ベンチマークデータセットにおける文書分類タスク

タスク	サビセット	ラベル数	データ数			Hugging Face での分割		
			訓練	検証	テスト	訓練	検証	テスト
tweet_eval	emotion	4	3,200	374	1,420	✓	✓	✓
tweet_eval	hate	2	9,000	1,000	2,970	✓	✓	✓
tweet_eval	irony	2	2,860	955	784	✓	✓	✓
tweet_eval	offensive	2	11,900	1,320	860	✓	✓	✓
tweet_eval	sentiment	3	45,600	2,000	12,300	✓	✓	✓
dair-ai/emotion		6	16,000	2,000	2,000	✓	✓	✓
poem_sentiment		4	892	105	104	✓	✓	✓
rotten_tomatoes		2	8,530	1,066	1,066	✓	✓	✓
sst2		2	67,349	872	1,821	✓		✓
emo		4	27,144	3,016	5,509	✓		✓
banking77		77	9,003	1,000	3,080	✓		✓
imdb		2	22,500	2,500	25,000	✓		✓
ethos	binary	2	800	99	99	✓		
ade_corpus_v2	classification	2	18,800	2,350	2,350	✓		
onestop_english		3	455	56	56	✓		
sms_spam		2	4,456	557	557	✓		
snips_built_in_intents		10	264	32	32	✓		

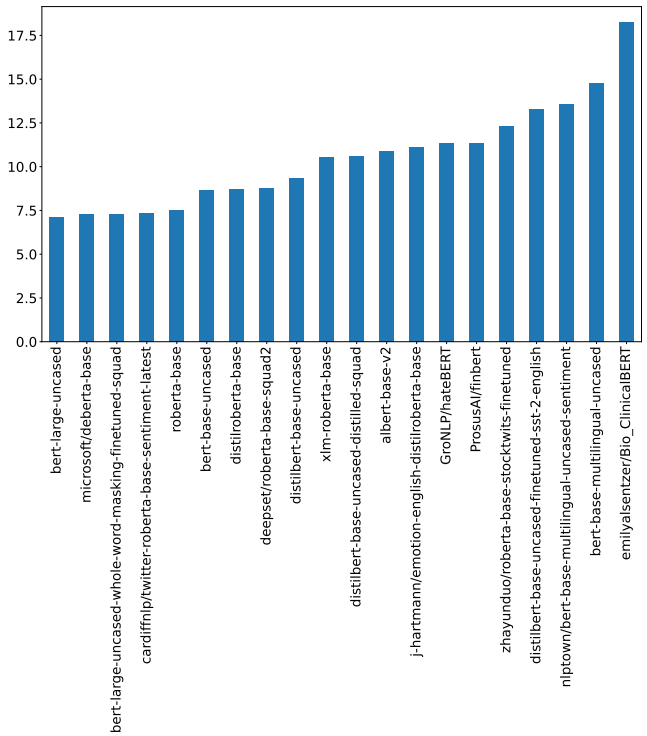


図 2　モデルの平均ランク順．低い平均ランクは高い性能を意味する．

5　k 近傍法を用いた事前学習済み BERT モデルの検索手法

k 近傍法を用いた事前学習済み BERT モデルの検索手法は **Inference あり**カテゴリー属している．検索時に各事前学習済みモデルを用いてタスクの訓練データと検証データのベクトルを計算し，得られたベクトルを使いモデルのランク付けを行う．ファインチューニングは行わないため，大幅に計算コストの削減ができる．

特定のタスクの訓練データ $D^{train} = \{(x_i^{train}, y_i^{train})\}_{i=1}^{M}$ と検証データ $D^{val} = \{(x_i^{val}, y_i^{val})\}_{i=1}^{N}$ を入力とし，タスクに対して性能が良い順で並べた事前学習済み BERT モデルのリスト $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ を出力するために，各事前学習済み BERT モデルに対し以下の処理を行う．

（1）　訓練データの文書 $\{x_i^{train}\}_{i=1}^{M}$ のベクトル $\{\mathbf{d}_i^{train}\}_{i=1}^{M}$ 及び検証データの文書 $\{x_i^{val}\}_{i=1}^{N}$ のベクトル $\{\mathbf{d}_i^{val}\}_{i=1}^{N}$ を計算する．

（2）　訓練ペア $\{(\mathbf{d}_i^{train}, y_i^{train})\}_{i=1}^{M}$ を使用して 1 近傍モデルを構築する．このモデルは，新しい文書のラベルを予測する際に最も近い訓練データのベクトルのラベルを出力するものである．ベクトル間の距離はコサイン類似性を使用する．

（3）　検証データ $\{(\mathbf{d}_i^{val}, y_i^{val})\}_{i=1}^{N}$ に対する，上記の 1 近傍モデルの正解率を計算する．

最終的に，第 3 ステップで得られた正解率の降順に事前学習済み BERT モデルを並べ替え，ランクリスト $\mathbf{m}$ を出力．ファインチューニングや学習のを必要としないため，提案手法は時間と計算リソースの削減が可能である．

6　評　　価

ベンチマークデータセットを用いて提案手法の評価を行った．ベースラインとして，図 2 の平均ランクの昇順を採用する．例えば，bert-large-uncased の平均ランクが一番高く，その次が microsoft/deberta-base のため，ベースラインは単純に全てのタスクに対して (bert-large-uncased, microsoft/deberta-base, ...) のランクリストを出力しているものである．

モデルの適合性が高いものの検索検索結果を強調するために，

表 2　ベンチマークデータセット内の事前学習済み BERT モデル．各モデルはアーキテクチャ，パラメータ数，事前学習およびファインチューニングに使用されたデータが異なっている

モデル	アーキテクチャ	事前学習データ	ファイチューニング
bert-base-uncased	BERT	BookCorpus, English Wikipedia	
bert-large-uncased	BERT	BookCorpus, English Wikipedia	
bert-base-multilingual-uncased	BERT	BookCorpus, Wikipedia	
GroNLP/hateBERT	BERT	BookCorpus, English Wikipedia, RAL-E	
ProsusAI/finbert	BERT	BookCorpus, English Wikipedia, FiQA Sentiment	Financial PhraseBank, TRC2-financial
bert-large-uncased-whole-word-masking-finetuned-squad	BERT	BookCorpus, English Wikipedia	SQuAD
nlptown/bert-base-multilingual-uncased-sentiment	BERT	BookCorpus, Wikipedia	Product Reviews
emilyalsentzer/Bio_ClinicalBERT	BERT	BookCorpus, English Wikipedia, PubMed, PMC, MIMIC III	NCBI Disease, 2010 i2b2/VA BC5CDR, BC4CHEMD, BC2GM, JNLPBA, LINNAEUS, Species-800, GAD, EU-ADR, CHEMPROT, BioASQ 4b-factoid, BioASQ 5b-factoid, BioASQ 6b-factoid
distilbert-base-uncased	Distil-BERT	BookCorpus, English Wikipedia	
distilbert-base-uncased-finetuned-sst-2-english	Distil-BERT	BookCorpus, English Wikipedia	SST-2
distilbert-base-uncased-distilled-squad	Distil-BERT	BookCorpus, English Wikipedia	SQuAD
xlm-roberta-base	RoBERTa	CommonCrawl, Wikipedia	
roberta-base	RoBERTa	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories	
cardiffnlp/twitter-roberta-base-sentiment-latest	RoBERTa	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories, Tweets	TweetEval
zhayunduo/roberta-base-stocktwits-finetuned	RoBERTa	BookCorpus, English Wikipedia CC-News, OpenWebText, Stories,	StockTwits
deepset/roberta-base-squad2	RoBERTa	BookCorpus, English Wikipedia CC-News, OpenWebText, Stories	SQuAD
distilroberta-base	Distil-RoBERTa	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories, Tweets	
j-hartmann/emotion-english-distilroberta-base	Distil-RoBERTa	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories, Tweets	Crowdflower, Emotion, GoEmotions, ISEAR, MELD, SemEval-2018
albert-base-v2	ALBERT	BookCorpus, English Wikipedia	
microsoft/deberta-base	DeBERTa	BookCorpus, English Wikipedia OpenWebText, Stories	

[3] で定義された nDCG スコアを使用した：

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{1}$$

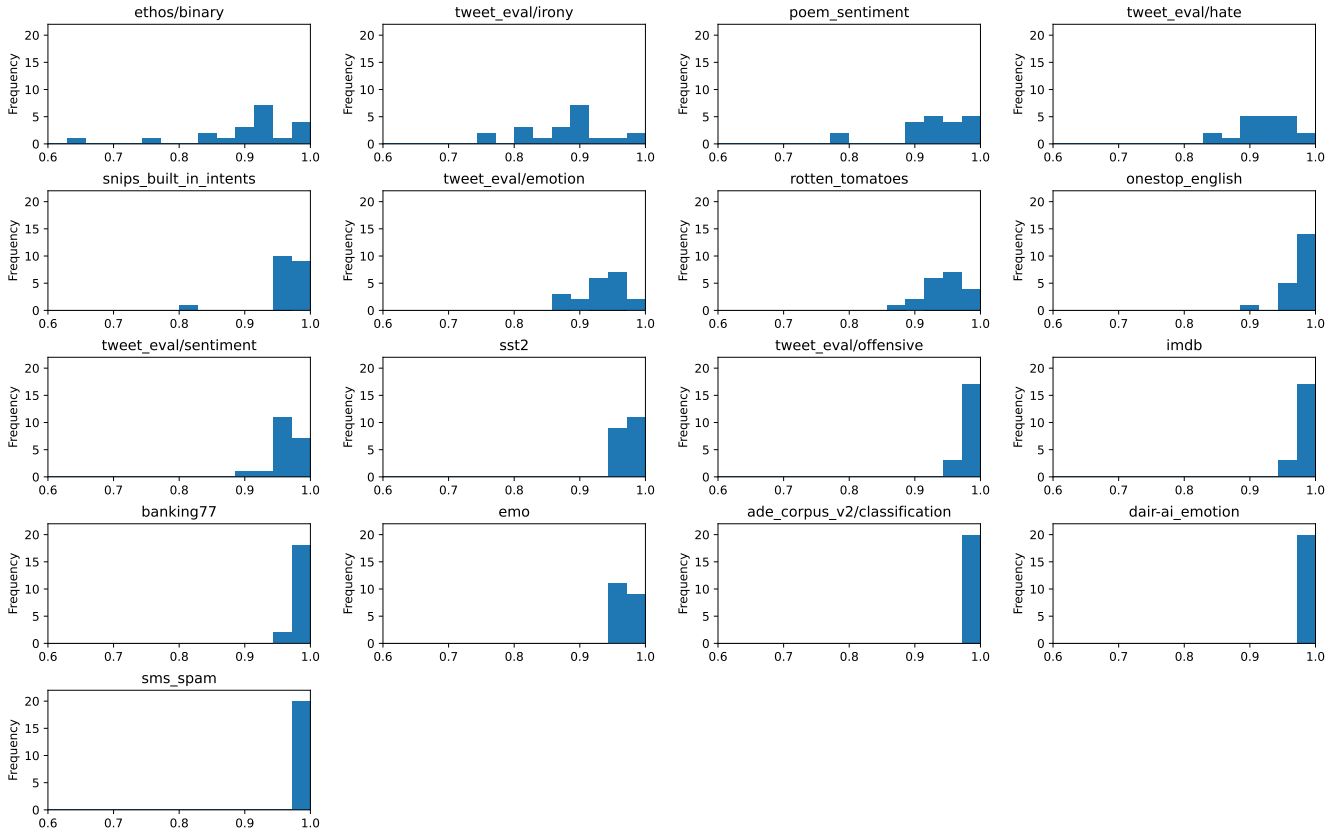


図 3　それぞれのタスクのテストデータにおけるファインチューニングされたモデルの正解率のヒストグラム．正解率は，一番のものが 1.0 になるように正規化されている．sms_spam, dair-ai/emotion，ade_corpus_v2/classification のようなタスクは，モデル間の性能に違いが小さいため，モデルの選択による影響も小さいと考えられる．一方，ethos/binary や tweet_eval/irony のようなタスクでは，モデル間の性能に大きいな違いがあるため，モデル選択がより重要である．

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2}$$

理想的な DCG_p（$IDCG_p$）は正解ランキングと完全に一致している場合のスコアである．rel_i は，テストデータ $D^{test} = \{(x_i^{test}, y_i^{test})\}_{i=1}^{M}$ に対するファインチューニングされたモデルの相対的正解率（$relacc$）から計算する．

$$rel = \begin{cases} 0, & \text{if } 0.00 \leqq relacc < 0.90 \\ 1, & \text{if } 0.90 \leqq relacc < 0.91 \\ 2, & \text{if } 0.91 \leqq relacc < 0.92 \\ \vdots & \\ 9, & \text{if } 0.98 \leqq relacc < 0.99 \\ 10, & \text{if } 0.99 \leqq relacc \leqq 1.00 \end{cases}$$

上記の rel_i の付け方により，特定のタスクに対して低い正解率を持つモデルの影響を減らしながら，高い適合性を持つモデルの性能を強調することができる．

評価結果を表 4 にて示している．提案手法は殆どのタスクにおいてベースラインを上回っている．さらに，提案手法は，高い nDCG スコアだけでなく，タスク間の nDCG スコアの分散も低いので，より安定した検索能が期待できる．第 4 章に記載しているように，一部のタスクではモデル間の性能に大きな違いがなく，モデルの選択が結果に大きく影響しない可能性がある．逆に，一部のタスクではモデルの性能に大きな差があり，モデル選択が重要である．テストデータ対するファインチューニングされたモデルの相対的正解率の分散に基づいて，タスクを 4 つのグループに分類した：分散が最も大きいグループ，分散が大きいグループ，分散が小さいグループ，分散が最も小さいグループ．分散が最も大きいグルーと大きいプグループのタスクでは，後者 2 つに比べてモデル選択による影響が大きい表 3 は，各グループの平均 nDCG スコアを示している．提案手法は分散が大きい二つのグループにおいて，ベースライン手法とよりも性能が良い．

7　結　　論

本研究では，事前学習済みモデル検索問題の定義を行い，この問題のためのベンチマークデータセットを構築した．また，k 近傍法を用いることで，ファインチューニングを行うことなく，BERT 事前学習済みモデルを検索する手法を提案した．ファインチューニングを行わないことで，計算リソースと時間を大幅

表 3　モデル間の性能の分散に基づいてグループ化されたタスクの $nDCG$ スコアの平均

グループ	$nDCG_1$		$nDCG_3$	
	ベースライン	提案手法	ベースライン	提案手法
分散が最も大きい	0.269	**0.421**	0.361	**0.413**
分散が大きい	0.196	**0.640**	0.424	**0.677**
分散が小さい	**0.687**	0.640	**0.702**	0.546
分散が最も小さい	0.725	**0.800**	**0.805**	0.756

に削減することができる．

今後の研究では，文書分類タスクの他に，他のタイプのタスクにおける事前学習済みモデルの検索を行いたいと思う．また，**Inference なし**カテゴリーの方手法も検討する予定である．

謝　　辞

本研究は，JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成，ならびに，2023 年度国立情報学研究所公募型共同研究（23S0502）の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Dogu Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*, 2019.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620, November 2019.

[3] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, p. 89–96, 2005.

[4] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, August 2021.

[5] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 558–563, June 2019.

[6] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2898–2904, November 2020.

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

[8] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. Searching data: A review of observational data retrieval practices. *Journal of the Association for Information Science and Technology*, Vol. 70, , 07 2017.

[9] Xiaochuang Han and Jacob Eisenstein. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4238–4248, November 2019.

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[11] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342*, 2019.

[12] Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. A query log analysis of dataset search. In *Web Engineering*, pp. 429–436, 2017.

[13] Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online active model selection for pre-trained classifiers. In *International Conference on Artificial Intelligence and Statistics*, pp. 307–315, 2021.

[14] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. A test collection for ad-hoc dataset retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2450–2456, 2021.

[15] Dagmar Kern and Brigitte Mathiak. Are there any differences in data set retrieval compared to well-known literature retrieval? In *TPDL*, Vol. 9316, pp. 197–208, 09 2015.

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*, 2019.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Vol. 36, pp. 1234–1240, September 2019.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

[19] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9–14, October 2020.

[20] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, Vol. 4, , 2021.

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6000–6010, 2017.

[23] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pp. 12133–12143, 2021.

[24] Lixin Zou, Weixue Lu, Yiding Liu, Hengyi Cai, Xiaokai Chu, Dehong Ma, Daiting Shi, Yu Sun, Zhicong Cheng, Simiu Gu, Shuaiqiang Wang, and Dawei Yin. Pre-trained language model-based retrieval and ranking for web search.

表 4　各タスクにおける提案手法とベースラインの評価結果

グループ	タスク	サブセット	*relacc* の分散	$nDCG_1$		$nDCG_3$	
				ベースライン	提案手法	ベースライン	提案手法
分散が最も大きい	ethos	binary	0.00542	0.001	**0.500**	0.207	**0.369**
	poem_sentiment		0.00299	0.015	**0.062**	0.024	**0.208**
	tweet_eval	irony	0.00244	0.062	**1.000**	0.419	**0.743**
	snips_built_in_intents		0.00161	**1.000**	0.124	**0.794**	0.330
分散が大きい	tweet_eval	emotion	0.00098	0.030	**1.000**	0.564	**0.849**
	rotten_tomatoes		0.00073	0.249	**1.000**	0.280	**0.816**
	tweet_eval	hate	0.00054	0.007	**0.062**	0.294	**0.395**
	onestop_english		0.00051	0.500	0.500	0.560	**0.648**
分散が小さい	sst2		0.00029	**1.000**	0.062	**0.734**	0.191
	tweet_eval	sentiment	0.00019	0.249	**1.000**	0.529	**0.671**
	imdb		0.00012	**1.000**	0.500	**1.000**	0.543
	tweet_eval	offensive	0.00011	0.500	**1.000**	0.543	**0.778**
分散が最も小さい	banking77		0.00010	0.500	0.500	**0.648**	0.617
	emo		0.00008	0.124	**0.500**	**0.642**	0.459
	ade_corpus_v2	classification	0.00004	1.000	1.000	**0.883**	0.852
	dair-ai/emotion		0.00002	1.000	1.000	0.852	0.852
	sms_spam		0.00001	1.000	1.000	1.000	1.000
平均				0.484	**0.636**	0.587	**0.607**
分散				0.171	**0.141**	0.073	**0.057**

ACM Transactions on the Web, Vol. 17, No. 1, dec 2022.

一般発表 | Track 1: 自然言語処理・機械学習基礎

言語モデル

座長:喜田 拓也(北海学園大学)
コメンテータ:若林 啓(筑波大学)
2024年2月28日(水) 10:00 〜 12:10 T1-B (オンライン（Zoom Events))

11:35 〜 12:00

[T1-B-1-05]【技術報告】 NECの LLM技術とデータ分析高速化技術

小山田昌史、董于洋、山崎智史、石坂一久 (日本電気株式会社)

©一般社団法人日本データベース学会

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

2024年2月28日(水) 10:00 ~ 12:10 | T2-A オンライン（Zoom Events)

問合せ処理・データベース管理

座長:金政 泰彦(富士通株式会社)
コメンテータ:三島 健(株式会社SRA)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T2-A-1-01] RDBMSにおけるベクトル類似検索と関係演算を複合した問合せ処理の実験と考察
*堀内 美聡[1]、渡 佑也[1]、山室 健[1] (1. NTTソフトウェアイノベーションセンタ)

10:20 ~ 10:45
[T2-A-1-02] 誤差保証付き近似的問合せ処理におけるシノプシス構築の高速化
*倪 天嘉[1]、杉浦 健人[1]、石川 佳治[1]、陸 可鏡[1] (1. 名古屋大学)

10:45 ~ 11:10
[T2-A-1-03] 単一表クエリを対象とするシノプシス埋込みによる近似問合せ処理の評価
*髙田 実佳[1]、合田 和生[2] (1. 東京大学情報理工学研究科電子情報学専攻、2. 東京大学)

11:10 ~ 11:35
[T2-A-1-04] データマネジメントとデータベース管理
*増永 良文[1] (1. お茶の水女子大学名誉教授)

RDBMS におけるベクトル類似検索と関係演算を複合した問合せ処理の実験と考察

堀内　美聡† 　渡　佑也† 　山室　健†

† NTT ソフトウェアイノベーションセンタ　〒 108–8224 東京都港区港南 1-2-70 品川シーズンテラス 29F
E-mail: †{ms.horiuchi,yuya.watari,takeshi.yamamuro}@ntt.com

あらまし　表現学習の進展によって，画像や文章など様々な情報が高次元空間上のベクトルとして埋め込まれることが一般的になってきている．このベクトルの類似検索は，従来は困難であった意味論的な検索を可能にし，推薦や質問応答などの分野では必須の技術となっている．これらの動向を背景に，企業データが多く蓄積されている RDBMS 上で，属性にベクトルを含むテーブルへの問合せ処理の重要性が増している．このような問合せでは，ベクトルの類似検索と選択や射影などの既存の関係演算を複合した処理（複合問合せ）が必要である．通常，ベクトルの類似検索は Top-k 検索であるため，複合問合せも Top-k 検索である．しかし，ベクトル類似検索の探索範囲が十分でない場合に，出力数が k を満たさない可能性があるという課題がある．これは，単に探索範囲を大きくすることで解決するが，それに応じて実行時間が長くなるため探索範囲は小さい方が望ましく，そのような適切な設定は困難である．そこで本稿では，関係演算の対象となる属性が均一分布であるような簡単な場合に対し，複合問合せにおいて出力数が k を満たすようなベクトル検索の探索範囲の理論・実験の両面での考察を行う．

キーワード　データベース，ベクトル類似計算，RDBMS，埋め込み，PostgreSQL

1 はじめに

深層学習による表現学習の進展により，画像や文章などあらゆる情報を高次元空間上のベクトルに変換した埋め込みが一般的になってきている [1–3]．埋め込み間の距離は情報の意味的な近さを表し，埋め込みに対するベクトル類似検索は，従来では困難であった非構造データへの問合せや，意味論的な検索を可能にした．応用分野は，推薦や質問応答など様々である [4,5]．

加えて，実務のデータ処理では，ベクトルの類似検索と，選択や射影，結合などの既存の関係演算を複合した処理（複合問合せ）がより有効である．これは以下の理由のためである．(1) 実世界の要求により適した，柔軟な検索が実現できる．例えば，商品説明の文章の埋め込みと価格の列を含むテーブルに対し，欲しい商品の説明文と予算を与えた問合せは，与えた説明に類似し，かつ予算内の商品を抽出することができる．(2) ベクトルで表される非構造データに加え，正確な問合せが可能な構造データの検索を組み合わせることにより，ベクトル類似検索精度の向上が期待される．例えば，ベクトル類似検索のみの場合，高次元ベクトルを扱うユースケースが多く，次元の呪いによって精度が低下する課題がある．一方で，複合問合せでは，構造データの情報を活用し，特に大規模データに対し，より本来の問合せ意図に即した検索実現が可能である [6]．

従来はベクトルと構造データが別の DBMS に保存されていたため，複合問合せの際は，それぞれに対して問合せを行い，後で結合する必要がある．しかし，そのような手法では，問合せそのものが非効率であることに加え，アプリケーションにて問合せと結果の管理を行う必要があり，開発コストも大きい．そこで，RDBMS 上で処理を行うことで，ベクトル類似検索と関係演算をまとめて効率的に処理できるようになり，導入時の実務上のコストの増加も低減できる．

RDBMS 上での複合問合せは，ベクトル類似検索と関係演算の処理を単に複合する手法が考えられる．例えばベクトル類似検索を先に行う場合，事前に構築された索引を用いてベクトルを Top-k' 近似検索し，得られた結果に対して選択処理を実行する．反対に，ベクトル類似検索を後に行うような実行順序も存在する．問合せによっていずれの実行順序が適しているか異なり，RDBMS では既に持つプランナによってコスト推定され，決定される．しかし，ベクトル類似検索を先に行う実行計画は，その後に関係演算によっていくつかの行が取り除かれるため，ユーザが指定した k 個の結果を出力できないという課題がある．これはベクトル類似検索の探索範囲を大きくすることで解決するが，探索範囲に沿って実行時間が長くなるため探索範囲は小さい方が望ましく，そのような適切な探索範囲の特定は困難である．

そこで本稿では，特にクラスタベースのベクトル索引を対象とし，複合検索においてベクトル類似検索を先に実行するような実行計画について，ベクトルの適した探索範囲を理論・実験の両面で考察する．実験では，ビジネス用途で広く使用されており，かつベクトル類似検索が可能な RDBMS である PostgreSQL を使用する．

本稿は以下の通りに構成される．2 節にて関連研究を，3 節にて事前知識と課題を説明する．4 節で理論による考察をし，5 節では実験結果を示す．最後に 6 節で本稿の結論を述べる．

2 関連研究

本節で複合問合せの関連研究を述べる．RDBMS にベクトル類似検索を組み合わせた複合問合せが提案されている [7,8]．これは RDBMS にてベクトル類似検索を行うことが適当であることを示す．いくつかのベクトル DB はベクトルに加えてカテゴリ値のようなメタデータを属性として持つことができ，それらに対するフィルタ処理（属性フィルタ）が可能である．そのようなベクトル DB に属性フィルタを組み合わせた複合問合せが提案されている [6,9–12]．なお，フィルタ処理とベクトル類似検索の併用は，関係演算としては選択演算のみが可能であるような複合問合せと同一である．

3 事前知識

本節で事前知識を説明する．複合問い合わせは，ベクトル類似検索と関係演算を組み合わせた Top-k 検索である．複合問合せを下記のように定義する．

問題定義（複合問合せ） ベクトルで構成される列を含むデータ集合を $\mathbb{D}$ とし，ベクトル v, u の距離を $dist(v,u)$ とする．複合問合せは，自然数 k，クエリベクトル v，関係演算による選択 f を与えたとき，$\forall d_1 \in \mathbb{D}', \forall d_2 \in f(\mathbb{D}) \setminus \mathbb{D}'$ に対し，$dist(v,d_1) \leqq dist(v,d_2)$ を満たす $\mathbb{D}' \subset f(\mathbb{D}), (|\mathbb{D}'| = k)$ を抽出する．

なお，通常，複合問合せは上記の定義に従い，$\mathbb{D}'$ の近似解を求める．SQL 1 は複合問合せの例である． *recipies_tb* というテーブルに対し， 説明文の埋め込みである *embedding* が *target_embedding* と類似しており，さらに *price* は 200 以下であるものを $k = 10$ 個取得する．

ベクトル類似検索は，多くの場合で Top-k 近似検索手法であり，事前に構築された索引を用いて，与えたベクトルに対して類似度が高い上位 k' 件を近似計算する．ベクトル索引の手法は，クラスタベース，グラフベースなど多数存在し，本稿ではクラスタベースの手法について議論を行うが，他のアプローチに基づく手法にも応用可能である．クラスタベースの手法は，データ D を n 個のクラスタ $\mathbb{C}$ にクラスタリングし，各クラスタには類似したデータが含まれる．探索時は，ハイパーパラメータ m を与え，クラスタを m 個選択する．次に，それぞれに含まれる全てのデータを探索し，ベクトルの類似度を計算する．そして，ベクトルの類似度が高い上位 k' 件を出力する．

関係演算は，演算結果のデータ点を全て出力する．なお，簡単のため関係演算は選択演算とするが，結合演算も結合キーが一致しないものを取り除く選択演算のように扱うことで，同様の結果が得られる．

3.1 複合問合せの課題

複合問合せは，事前ベクトル類似検索，事後ベクトル類似検索が存在する．事前ベクトル類似検索は，ベクトル類似検索を先に実行し，その後関係演算を実行する順序である．事後ベクトル類似検索は，先に関係演算を実行し，その後ベクトル類似検索を実行する順序である．いずれの実行計画がより効率的であるかは，問合せによって異なる．SQL 2 は， SQL 1 について事前ベクトル類似検索の実行順序を表した例である．SQL 2 では， *recipies_tb* というテーブルに対し， まず *target_embedding* と類似する *embedding* を $k' = 100$ 件検索し，その後 *price* は 200 未満であるという選択演算を実行し，最後に $k = 10$ 個のみを抽出する．

SQL 1　実行する SQL

```
SELECT * FROM recipies_tb WHERE price < 200 ORDER BY
    embedding <-> target_embedding LIMIT 10;
```

SQL 2　事前ベクトル類似検索の SQL

```
SELECT * FROM (SELECT * FROM recipies_tb ORDER BY
    embedding <-> target_embedding LIMIT 100)
    WHERE price < 200 LIMIT 10;
```

このような事前ベクトル類似検索の実行順序であるとき，複合問合せは k 個の結果を出力しない場合が多く存在するという課題がある．これは 2 つのことに起因する．(1) ベクトル類似検索の結果である k' 個のデータは，その後に選択演算などの関係演算によって一部が取り除かれる場合があるため．(2) ベクトル類似検索の m, k' はハイパーパラメータであり，これらはベクトル類似検索の出力数を決定するが，値が小さいと，出力数が k を満たさないため．また，これは (1) の理由のため，ベクトル類似検索のみ，関係演算のみの場合では発生せず，複合問合せ特有の課題である．したがって，k 個の出力を満たすような k'，m を指定することが必要であるが，これらの値は探索空間を示しているため，値を大きくすると探索空間が大きくなり，線形に実行時間が増加してしまう．そのため，k 個の出力を満たすことに加え，できるだけ小さな m, k' が求められるが，そのような値の特定は困難である．

4 ベクトル類似検索の適切な探索範囲

本節では，3 節で述べた課題を解決するため，ベクトル類似検索の適切な探索範囲について考察する．簡単のため，関係演算の対象である属性の値の分布を，均一分布とする．

まず，複合問合せの出力数が k 以上になる最も小さい k' について考える．s を関係演算の選択率とすると，ベクトル類似検索で得られる出力数 k' に対し，関係演算の対象の属性は均一分布であるので，関係演算実行後に得られる出力数は $k's$ 個である．関係演算実行後に得られる出力数が k 以上であるとき，

$$k's \geqq k \tag{1}$$

である．この式を満たす最も小さい k' を $\tilde{k'}$ とする．すると，k が与えられたとき，$\tilde{k'}$ の期待値 $E[\tilde{k'}]$ は

$$E[\tilde{k'}] = \frac{k}{s} \tag{2}$$

である．実際の問合せで用いる値は，$\tilde{k'}$ は自然数であることに留意し，$\left\lceil E[\tilde{k'}] \right\rceil$ である．

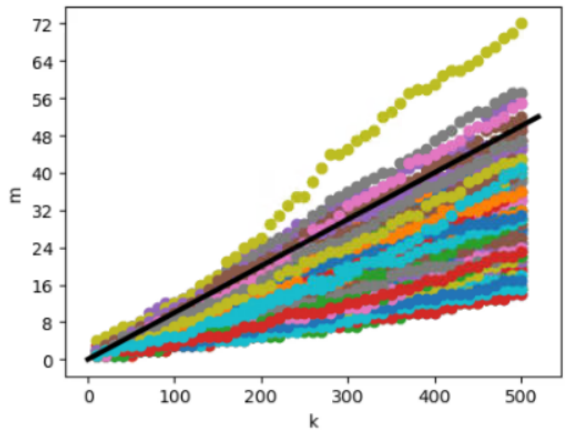


図 1　k を変化させた際の $\tilde{m}$

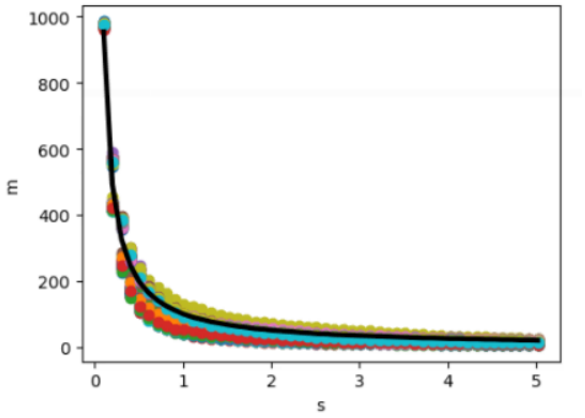


図 2　s を変化させた際の $\tilde{m}$

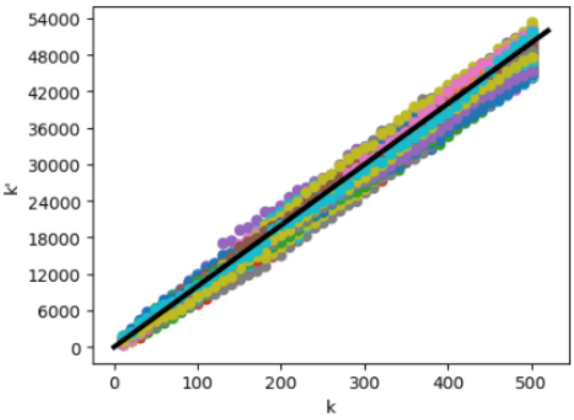


図 3　k を変化させた際の $\tilde{k'}$

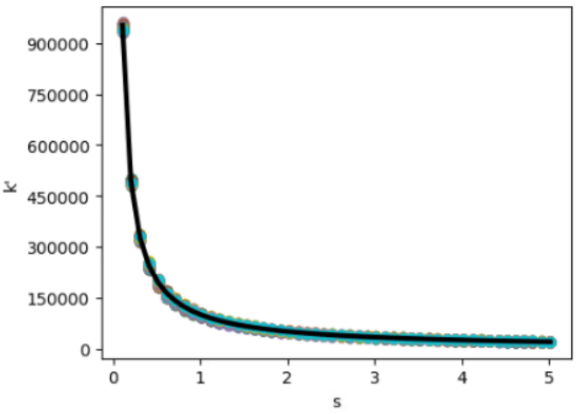


図 4　s を変化させた際の $\tilde{k'}$

次に，複合問合せの出力数が k 以上になる最も小さい m である $\tilde{m}$ について考える．索引によってクエリベクトルの近傍のクラスタからたどるように探索される．その探索順序を $C_1, C_2, ..., C_n, (\forall i \in [1,n] \forall C_i \in \mathbb{C})$ とする．$|C_i|$ をクラスタ C_i に含まれるデータ数とし，

$$\sum_{i\in[1,m-1]} s|C_i| \leqq k \leqq \sum_{i\in[1,m]} s|C_i| \tag{3}$$

を満たす m が求めたい $\tilde{m}$ である．$|C_i|$ が未知のとき，クラスタの個数 m とクラスタ当たりのデータ数の期待値 $E[|C|]$ を用いると，

$$\sum_{i\in[1,m]} |C_i| \approx E[|C|]m \tag{4}$$

と近似できる．したがって，k が与えられたとき，式 (3), (4) より，$\tilde{m}$ の期待値 $E[\tilde{m}]$ は，

$$E[\tilde{m}] = \frac{k}{sE[|C|]} \tag{5}$$

である．実際の問合せで用いる値は，$\tilde{m}$ は自然数であることに留意し，$\lceil E[\tilde{m}] \rceil$ である．

5　実　　験

本節では，4 節で理論的に考察した式に対し，実験によって考察を行う．実験は，m, k' の値をそれぞれ式 (5), (2) に基づいて求めた際に，実際の $\tilde{m}, \tilde{k'}$ とどの程度乖離しているかを調べる．

5.1　実 験 環 境

実験は Intel(R) Core(TM) i9-12900 CPU, 62 GB メモリを搭載した Ubuntu22.04 のマシン上で実行した．評価手法は，実務で幅広く利用されている RDBMS の PostgreSQL 16.1 と，その拡張である pgvector 0.5.1 を使用した．データセットには SIFT-1M を使用する [13]．このデータセットは， 128 次元のベクトルからなる列のみが存在し，データサイズは $|D| = 1,000,000$ 行である．クラスタに含まれるデータ数の期待値は $E[|C|] = |D|/m$ と計算する．複合問合せの実験のため，このデータセットに $[0, 1000)$ の値の範囲で均一分布の人工データをランダムに生成して付与した．ベクトル類似検索の索引は，クラスタベースの手法である IVFFlat [14] を使用し，索引構築時のハイパーパラメータ $lists = 1000$ とし，これは $\mathbb{C}$ の要素数を表す． SQL は SQL 2 の形で与えた．クエリベクトルは，データセットに含まれるベクトルをランダムに抽出した 100 個を使用した．

5.2　実 験 結 果

5.2.1　探索クラスタ数 m

図 1，2 にそれぞれ k, s を変化させた際の最適な探索クラス

タ数 $\tilde{m}$ を示す．横軸はそれぞれ k, s, 縦軸は最適な探索クラスタ数 $\tilde{m}$，色はそれぞれ異なるクエリベクトルを表す．黒色の線が式 (5) を表し，この線よりも下側にプロットされた問合せの点は，式 (5) より導出した m の値を利用すると k 個の出力を満たすことを意味し，線より上側にプロットされた点は満たさないことを意味する．なお，実験は $k' = |D|$ とし，$\tilde{m}$ を二分探索によって取得した．結果より，式 (5) は k, s を変化させた際の傾向を捉えられていることを示している．k, s を変化させた際に k 個の出力を満たしている割合は，それぞれ 85.8%, 88.1% であった．これは，クラスタに含まれるデータ数はクラスタによって偏りがあり，データ数が少ないクラスタが多く探索範囲に入るとき，式 (4) による近似の誤差が大きくなるためである．また，k が大きい場合は $\tilde{m}$ の値の幅が大きくなっている．これは，式 (3) で定義される $\tilde{m}$ が $|C_i|$ の値に依存し，C_i はクラスタの探索順序によって決定されるが，それはクエリベクトルによって異なるためである．

5.2.2 ベクトル類似検索の出力数 k'

図 3，4 にそれぞれ k, s を変化させた際のベクトル類似検索の最適な出力数 $\tilde{k'}$ を示す．横軸はそれぞれ k, s, 縦軸は最適な探索クラスタ数 $\tilde{k'}$，色はそれぞれ異なるクエリベクトルを表す．黒色の線が式 (2) を表し，この線よりも下側にプロットされた問合せの点は，式 (2) より導出した k' の値を利用すると k 個の出力を満たすことを意味し，線より上側にプロットされた点は満たさないことを意味する．なお，実験は $k = 1000$ とし，m は図 1，2 で求めた $\tilde{m}$ を用いて，$\tilde{k'}$ を二分探索によって取得した．結果より，式 (2) は k, s を変化させた際の傾向を捉えられていることを示している．k, s を変化させた際に k 個の出力を満たしている割合は，それぞれ 45.1%, 68.6% であった．これは，式 (2) より，値がクエリベクトルに依存せず，一様分布の条件下で期待値を計算しているため，実際に取りうる値の平均値近くに黒線が位置するためである．

5.3 考　　察

上記の結果より，$\tilde{m}$ の値は $\tilde{k'}$ に比べて，取りうる値の範囲が大きい．これは，式 (3) より，$\tilde{m}$ はクラスタの探索順序や，クラスタごとのデータの偏りの影響を受ける一方で，式 (1) より，$\tilde{k'}$ はそれらの影響を受けないためである．したがって，$\tilde{m}$ をより正確に見積もるには，正確な $|C_i|$ の値を知る必要がある．これに対し，クラスタに含まれるデータ数とクラスタの探索順序を取得して，適応的に探索クラスタ数を決定する，のような方法が考えられる．

6 ま と め

本稿では，RDBMS における構造データとベクトル化した非構造データを同時に検索する複合問合せに取り組んだ．複合検索は事前ベクトル類似検索の実行計画が指定の出力数未満を出力する課題が存在し，その解決のために，適切な探索範囲について理論と実験の両面からの考察を行った．今後は，考察結果をもとにさらに実験を行い，適用的に探索範囲を変化させる手法の開発や，実際の PostgreSQL に組み込んでコスト推定に適用することを検討する．

文　　献

[1] Seongwon Lee, Suhyeon Lee, Hongje Seong, and Euntai Kim. Revisiting self-similarity: Structural embedding for image retrieval. In *Proceedings of the CVPR*, pp. 23412–23421, 2023.

[2] Guanghui Qin and Benjamin Van Durme. Nugget: Neural agglomerative embeddings of text. In *Proceedings of the ICML*, pp. 28337–28350. PMLR, 2023.

[3] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the SIGKDD*, pp. 855–864, 2016.

[4] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the SIGIR*, pp. 639–648, 2020.

[5] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *Proceedings of the WSDM*, pp. 105–113, 2019.

[6] Chuangxian Wei, Bin Wu, Sheng Wang, Renjie Lou, Chaoqun Zhan, Feifei Li, and Yuanzhe Cai. Analyticdb-v: A hybrid analytical engine towards query fusion for structured and unstructured data. *Proceedings of the VLDB Endowment*, Vol. 13, No. 12, pp. 3152–3165, 2020.

[7] Qianxi Zhang, Shuotao Xu, Qi Chen, Guoxin Sui, Jiadong Xie, Zhizhen Cai, Yaoqi Chen, Yinxuan He, Yuqing Yang, Fan Yang, et al. Vbase: Unifying online vector similarity search and relational queries via relaxed monotonicity. In *Proceedings of the OSDI*, 2023.

[8] Wen Yang, Tao Li, Gai Fang, and Hong Wei. Pase: Postgresql ultra-high-dimensional approximate nearest neighbor search extension. In *Proceedings of the SIGMOD*, pp. 2241–2253, 2020.

[9] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. In *Proceedings of the SIGMOD*, pp. 2614–2627, 2021.

[10] Mengzhao Wang, Lingwei Lv, Xiaoliang Xu, Yuxiang Wang, Qiang Yue, and Jiongkang Ni. Navigable proximity graph-driven native hybrid queries with structured and unstructured constraints. *arXiv preprint arXiv:2203.13601*, 2022.

[11] Siddharth Gollapudi, Neel Karia, Varun Sivashankar, Ravishankar Krishnaswamy, Nikit Begwani, Swapnil Raz, Yiyong Lin, Yin Zhang, Neelam Mahapatro, Premkumar Srinivasan, et al. Filtered-diskann: Graph algorithms for approximate nearest neighbor search with filters. In *Proceedings of TheWebConf*, pp. 3406–3416, 2023.

[12] Wei Wu, Junlin He, Yu Qiao, Guoheng Fu, Li Liu, and Jin Yu. Hqann: Efficient and robust similarity search for hybrid queries with structured and unstructured constraints. In *Proceedings of the CIKM*, pp. 4580–4584, 2022.

[13] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, Vol. 87, p. 101374, 2020.

[14] Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. Learning passage impacts for inverted indexes. In *Proceedings of the SIGIR*, pp. 1723–1727, 2021.

誤差保証付き近似的問合せ処理におけるシノプシス構築の高速化

倪　天嘉† 杉浦　健人† 石川　佳治† 陸　可鏡†

† 東海国立大学機構名古屋大学大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
Email: {ni, lu}@db.is.i.nagoya-u.ac.jp, {sugiura, ishikawa}@i.nagoya-u.ac.jp

あらまし　近年，分析要件の高度化に伴い，データベースにおける効率的な問合せ処理はより重要になってきている．近似的問合せ処理（approximate query processing, AQP）は問合せ処理の中でも解析処理の高速化に取り組む研究分野の 1 つであり，処理結果にいくらかの誤差を許すことでより高速な処理を実現する．近似問合せ処理においてシノプシス（synopsis）は対象となるデータをコンパクトに集約したデータを指す一般的な概念であり，さまざまなアプローチがある．近年提案された BAQ（bounded approximate query processing）は数値属性のデータを適切にバケット化することで，近似的問合せ処理において厳密な誤差保証を提供するフレームワークを開発した．しかし，BAQ では数値属性のバケット化にデータのソートを必要とする他，元テーブルからシノプシスを抽出する際にも複数回のテーブル走査が必要となるなど，シノプシスの構築時間が大きいという問題がある．そこで本稿ではシノプシス構築の高速化に焦点を当て，新たなシノプシスの定義およびアルゴリズムを提案する．複数のデータセットにおける対象問合せを用いた実験により既存研究と比較し，提案手法の有効性を示す．

キーワード　近似的問合せ処理，問合せ処理，集計計算．

1 はじめに

近年，データ量の増加と分析の要求の高度化に伴い，データベースにおける効率的な問合せ処理はより重要になってきている．近似的問合せ処理（approximate query processing, AQP）は問合せ処理の中でも解析処理の高速化に取り組む研究分野の 1 つであり，処理結果にいくらかの誤差を許すことでより高速な処理を実現する．近似問合せ処理においてシノプシス（synopsis）は対象となるデータをコンパクトに集約したデータを指す一般的な概念であり，さまざまなアプローチがある [1–3]．スキャンが一定時間に完了できない大規模なデータベースを対象として，AQP は分析問合せの結果の精度と問合せ処理の待ち時間の間のトレードオフを提供する．

近年提案された bounded approximate query processing（BAQ）は，シノプシスに基づき厳密な誤差保証を提供する近似的問合せ処理フレームワークである [4]．BAQ では，ユーザが設定する誤差の閾値と典型的なワークロードを示す問合せ集合を用いてオフライン処理でデータベースからシノプシスを生成し，そのシノプシスを使用して効率的にオンライン問合せに回答する．他の要約データを使用するアプローチ（サンプリング，分位数など [1,5]）と比べて，BAQ は一般的な集計関数（COUNT, SUM, AVG, MIN, MAX）に対して厳密な誤差上限を保証する．一方，BAQ では対象データや問合せによって誤差保証が成立しない場合があるため，著者らはそれらの問題を改善した BAQ± を先行研究として提案した [6]．BAQ± では BAQ を基にシノプシスの構成要素であるバケット（数値属性の値をグループ化したもの）の生成手法を改善し，より多様なデータ・問合せにおいても誤差保証を提供すると共に，バケットの生成数を減少させることで最終的なシノプシスのサイズも削減した．これにより，元の BAQ よりも多様な問合せへの対応とオンラインの問合せ処理時間の削減とを両立した．

しかし，BAQ± には前処理におけるシノプシスの構築時間が大きいという問題がある．BAQ± は BAQ と同様に部分シノプシスの構築と統合シノプシスへの統合の 2 手順でシノプシスを構築する．部分シノプシスを構築する際には数値属性のバケット化のためににデータのソートを必要とする他，統合シノプシスを構築する際にも複数回の元テーブルとシノプシスの走査が必要となる．さらに，部分シノプシスの構築はワークロードの各問合せに対して必要なため，ワークロード問合せの数が増えるほど構築時間が増大する．

また，BAQ と BAQ± は時間属性などのようにカーディナリティの大きい属性を効率的に処理できないという問題も持つ．BAQ の主な貢献は相対誤差に基づく数値属性のバケット化，およびそれに伴うシノプシスサイズの削減であり，カテゴリ属性に対しては単純なグルーピングしか行わない．そのため，時間属性などの粒度が細かいデータも愚直にグルーピングされ，シノプシスサイズおよび問合せ処理時間の増大が発生しうる．

本稿では以上の問題を解決するために，シノプシス構築の改善に焦点を当て，新たなシノプシスの定義およびアルゴリズムを提案する．複数のデータセットにおける対象問合せを用いた実験により既存研究と比較し，提案手法の有効性を示す．

2 準　　備

BAQ± は BAQ を基にした手法であり，近似的問合せ処理における厳密な誤差保証を提供する．BAQ± では高精度かつ高効率な近似的問合せ処理を目標としており，シノプシス構築における数値属性のバケット生成とオンラインの近似問合せ処理を改善した．図 1 に BAQ± 全体のフレームワークを示す．

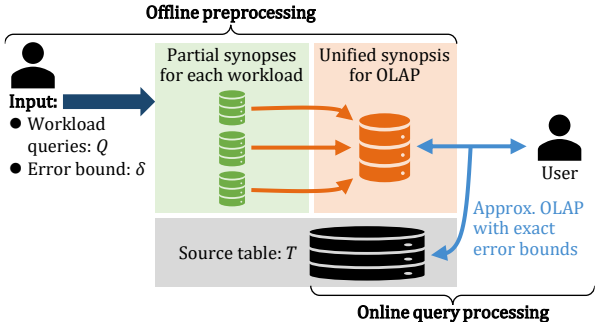


図 1: BAQ± のフレームワーク

BAQ± では以下の演算を用いた OLAP 問合せを想定する．

- COUNT，MIN，MAX，SUM，および AVG での集約
- カテゴリ属性に対する =，≠ を用いた選択
- 数値属性に対する =，≠，>，≧，<，および ≦ を用いた選択
- カテゴリ属性によるグルーピング

BAQ± では，BAQ と同様に与えられた問合せに対し，集約結果の真値と近似値との誤差 *err* として相対誤差を用いる．2 つの値 $x, y \in R$ の相対誤差を以下の式で定める．

$$\mathrm{err}(x,\hat{x}) = \begin{cases} \left|\frac{x-\hat{x}}{x}\right| & (|x| \geqq \epsilon) \\ \frac{|x-\hat{x}|}{\epsilon} & (\text{otherwise}) \end{cases} \qquad (1)$$

ただし，x を真値，$\hat{x}$ を近似値とし，ϵ を非常に小さな正の値とする．

2.1 前処理におけるシノプシスの構築

BAQ± では，BAQ と同様にユーザからワークロードとなる問合せの集合が与えられると想定しており，与えられた問合せ集合の情報を使用してシノプシスを構築する．各ワークロード問合せ q は *query column set*（QCS）π に変換され，この QCS を基に各問合せに対応した部分シノプシスを構築する．なお，他の QCS の部分集合となるようなものについては，上位集合となる QCS から生成されたシノプシスによって問合せへの回答が可能となるため，部分シノプシスの生成は行われない．

QCS π がカテゴリー属性のみを持つ場合，元テーブル内のタプルを π に含まれるカテゴリ属性でグループ化することで部分シノプシスを構築する．例えば QCS $\pi_1 = \{\mathrm{region}, \mathrm{date}\}$ を用いる場合，表 1 のテーブルから表 2 の部分シノプシスが構築される．なお，シノプシスでは各グループのタプル数，つまりグループごとの COUNT 集約の結果が CNT として保持される．

QCS π が数値属性を含む場合，BAQ± では以下の式 (2) を満たすよう各値をバケット $b \subset \mathcal{R}$ へ分割する．なお，$\overline{b}$ はバケット内の値の平均値を表し，δ はユーザが指定する誤差上限である．

$$err(\min b, \overline{b}) \leqq \delta \wedge err(\max b, \overline{b}) \leqq \delta \qquad (2)$$

まず，QCS π がカテゴリ属性を含む場合，テーブル中のタプルをカテゴリ属性でグループ化する．次に，各グループにおいて数値属性の値を昇順にソートし，各値を式 (2) を満たすバケット b に割り当てる。例えば，QCS $\pi_2 = \{\mathrm{region}, \mathrm{temperature}\}$，誤差

表 1: 入力テーブル “temperature”

ID	region	date	temperature
1	Middle East	2006-06-27	98.5
2	Middle East	2007-01-10	103.5
3	Europe	2006-12-23	-13.1
4	Asia	2008-01-02	-10.8
5	Europe	2006-12-21	-12.7
6	Middle East	2009-08-01	103.1
7	Asia	2008-01-02	-10.5
8	Europe	2012-02-01	-12.6
9	Middle East	2009-08-01	110.1
10	Asia	2008-05-03	23.2

表 2: 表 1 における $\pi_1 = \{\mathrm{region}, \mathrm{date}\}$ のシノプシス

region	date	CNT_1
Middle East	2006-06-27	1
Middle East	2007-01-10	1
Europe	2006-12-23	1
Asia	2008-01-02	2
Europe	2006-12-21	1
Middle East	2009-08-01	2
Europe	2012-02-01	1
Asia	2008-05-03	1

表 3: 表 1 における $\pi_2 = \{\mathrm{region}, \mathrm{temperature}\}$ のシノプシス

region	temperature (min., max., mean)	CNT_2
Middle East	(98.5, 103.5, 101.7)	3
Middle East	(110.1, 110.1, 110.1)	1
Asia	(-10.8, -10.5, -10.7)	2
Asia	(23.2, 23.2, 23.2)	1
Europe	(-13.1, -12.6, -12.8)	3

表 4: 表 2 および表 3 を統合したシノプシス

region	date	temp. (min., max., mean)	CNT_1	CNT_2
Middle East	2006-06-27	(98.5, 103.5, 101.7)	1	3
Middle East	2007-01-10	(98.5, 103.5, 101.7)	1	0
Europe	2006-12-23	(-13.1, -12.6, -12.8)	1	3
Asia	2008-01-02	(-10.8, -10.5, -10.7)	2	2
Europe	2006-12-21	(-13.1, -12.6, -12.8)	1	0
Middle East	2009-08-01	(110.1, 110.1, 110.1)	2	1
Europe	2012-02-01	(-13.1, -12.6, -12.8)	1	0
Asia	2008-05-03	(23.2, 23.2, 23.2)	1	1

上限 $\delta = 0.05$，および表 1 のテーブルを用いた場合，表 3 のバケットおよび部分シノプシスが構築される．

ワークロードから複数の部分シノプシスを構築した後，BAQ± は BAQ と同様にスキャンベースもしくは貪欲ベースの方法に基づいて表 4 のような**統合シノプシス**を生成する．各方法の詳細はここでは省くが，基本的なアイデアは各部分シノプシスの実体として利用可能なタプルの，元テーブルからの抽出である．

例えば表 2 および 3 の部分シノプシスの場合，表 1 の ID 1 のタプルが各シノプシスの 1 行目とカテゴリ属性が一致し，かつ数値属性がバケット範囲内に収まっている．したがって，ID 1 のタプルを元テーブルから抽出し，対応する部分シノプシスの CNT 値と共に表 4 の統合シノプシスへ挿入する．このように元テーブル中のすべてのタプルを確認し，統合シノプシスへ挿入していくのが BAQ ないし BAQ± におけるシノプシス構築である．入力テーブル T に N 個のタプル，τ 個の部分シノプシスに合計で M 個のタプルがあるとしたとき，スキャンベースおよび貪欲法ベースの方法ではそれぞれ $O(MN)$ および $O(\tau MN)$ の時間計算量が必要となる．

2.2 OLAP 問合せへの近似的な回答

BAQ± では統合シノプシスを用いて OLAP 問合せを処理する．数値属性による選択がある問合せの場合，数値属性に関する各バケットは**条件を満たす**，**部分的に条件を満たす**および**条件を満たさない**の 3 通りに分けられる．図 2 は属性 “temperature” に対し条件 $-5 <$ temperature < 15 を与えた際の例である．全てのバケットが条件を満たす場合，誤差なしで問合せを処理できる．一方部分的に条件を満たすバケットが存在するとき，シノプシスのみでは厳密な誤差を保証できない場合がある．そのため BAQ± では条件を部分的に満たすバケットが持ちうる誤差の最大値からシノプシスのみで問合せに回答可能か判定し，誤差を保証できない場合は元テーブルを使用する．なお，最小・最大値集約に関してはバケット定義から問合せの書き換えなしで厳密な誤差を保証できるため，以下では説明を省く．

数値属性 A に対する集約問合せでは，条件を完全に満たすバケットについては以下のルールに基づき統合シノプシス上で処理する．この場合，誤差は発生しない．

- $\texttt{COUNT}(*) \rightarrow \texttt{SUM}(\texttt{CNT})$
- $\texttt{SUM}(A) \rightarrow \texttt{SUM}(\text{mean}(A) \cdot \texttt{CNT})$
- $\texttt{AVG}(A) \rightarrow \dfrac{\texttt{SUM}(A)}{\texttt{SUM}(\texttt{CNT})}$

一方で，部分的に条件を満たすバケットが含まれ，かつシノプシスのみで誤差保証可能な場合は以下の変換ルールを用いる．ただし，n と m をそれぞれ条件を満たすバケットにおけるカウントおよび合計値であるとし，t，$a\,(>0)$，および $b\,(<0)$ をそれぞれ部分的に条件を満たすバケットのカウント，正数に関する合計値，負数に関する合計値とする．

- $\texttt{COUNT}(*) \rightarrow n + t/2$
- $\texttt{SUM}(A) \rightarrow m + (a+b)/2$
- $\texttt{AVG}(A) \rightarrow (m + (a+b)/2)\,/n$

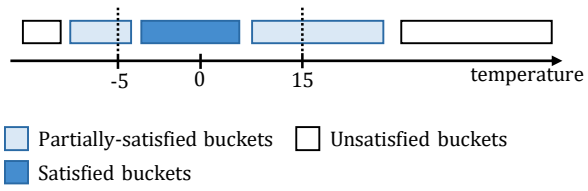


図 2: バケットの分類（$-5 <$ temperature < 15 の場合）

3 BAQ± フレームワークの改善

前述したように，BAQ± のフレームワークには 1) 部分・統合シノプシスに基づく 2 段階の構築手続きによる前処理時間の悪化，および 2) カーディナリティの大きい属性に対する問合せ処理時間の悪化という 2 つの課題がある．これらの課題に対し，まず 1) に対してはバケット定義を改善し，元テーブルから直接最終的なシノプシスを構築する手法を提案する．更に，2) に対してはデータウェアハウスにおける次元テーブル [7] を導入し，次元属性として扱われる値の集約を実施する．

提案する BAQ± の新たなフレームワークを図 3 に示す．本研究ではユーザが入力した次元テーブル $T_D \in \mathcal{D}$ に応じて，シノプシスを階層的に構築する．図 3 では例として，時間に関して {day, month, quarter, week, year}からなる次元テーブルが入力された際の階層的シノプシスを示している．なお，各要素は以下のような順序関係を持つとする．

$$\text{day} \prec \text{month} \prec \text{quarter} \prec \text{year}$$

$$\text{day} \prec \text{week} \prec \text{year}$$

つまり，この順序関係において前のものほど詳細な，後のものほど集約された情報がシノプシスで保持される．

本章の残りでは次元テーブルの導入によって拡張されるオフライン・オンライン手続きについて述べ，続く次章でシノプシス構築の高速化について説明する．なお，本研究では次元テーブルを導入するが，伝統的なデータウェアハウスにおいてサポートされる roll-up などの汎用的な OLAP 問合せは扱わない点に注意する．従来の BAQ と同様に本フレームワークではユーザが対象とするワークロードを問合せ集合として与えることを想定し，与えられたワークロードに合致した OLAP 問合せを近似処理により高速で処理することを目指す．

3.1 オフライン処理の拡張

次元テーブルの導入およびシノプシス構築の高速化にあたり，まず数値属性を近似的に集約するためのバケット定義を変更する．その後，構築されたシノプシスを基にした次元テーブルによる階層的シノプシスの構築について述べる．

3.1.1 バケット定義の改善

前述したように，BAQ において数値属性の各バケットは元

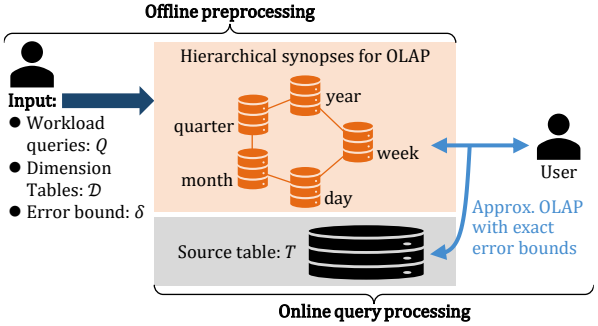


図 3: 改善後の BAQ± フレームワーク

テーブル内の値に応じて動的にその範囲が決定されるが，この範囲の動的な決定がシノプシスの構築時間を悪化させる原因の 1 つである．バケットの生成時には元テーブル内の値は未知であるため，まずすべてのタプルを対象の数値属性によってソートし，正数および負数をそれぞれ昇順・降順で走査することでバケットを生成していく．つまり，ワークロード問合せに含まれる数値属性の数だけ元テーブルのソートが必要となる上，数値属性同士が相関を持つ（i.e., 1 つの問合せ内で 2 つ以上の数値属性が使用される）場合はある数値属性をバケット化した後に残りの数値属性を用いてバケット内の値を再ソートしなければならない．

このようなソート処理を除くために，提案手法では各数値属性 A_i に対して最小単位 ϵ_i を導入し，各バケットの範囲およびその一意な ID を静的に決定する．例として，図 4 にこのバケット分割の概略図を示す．まず最小単位として ϵ_i を持つ数値属性 A_i に対し，絶対値が最小単位未満の値（$|a_i| < \epsilon_i$）を ID 0 のバケットとしてまとめる．次に，正の値を含むバケットを以下の式によって ID j にまとめる．

$$[\epsilon_i(1+\delta)^{j-1}, \epsilon_i(1+\delta)^{j}) \tag{3}$$

同様に負の値を含むバケットは ID $-j$ としてまとめる．

$$(-\epsilon_i(1+\delta)^{j}, -\epsilon_i(1+\delta)^{j-1}] \tag{4}$$

この分割において最小単位は式 (1) における ϵ に対応し，言い換えればゼロとみなせるしきい値を表す．

このバケット分割の利点として，ある数値属性の値 $a_i \in A_i$ から対応するバケット ID を以下の式で求められる点が挙げられる．

$$j = \operatorname{sgn} a_i \left\lceil \log_{1+\delta} \frac{|a_i|}{\epsilon_i} \right\rceil \tag{5}$$

なお，sgn は与えた値がゼロのときは 0，正負いずれかの場合は対応する符号（+1 or −1）を返す符号関数である．つまり，元テーブルに含まれる各数値属性の値は静的にバケット化された ID へ変換できる．次章ではこの静的な ID への変換を利用したシノプシス構築について述べる．

なお，ϵ_i は数値属性 A_i において許容される最小相対誤差のしきい値とする．このしきい値の決定は対象の数値属性に関する背景知識に依存するため，本研究では基本的にユーザによって各数値属性のしきい値が指定されることを想定する．ただし，ユーザによる指定が困難である場合は，計算処理の面からシステム側でしきい値を決定することも考える．例えば対象の数値属性が整数型であれば，最も悲観的なしきい値として $\epsilon = 1$ の利用が考えられる．もしくは，対象の型が浮動小数点数であれば，生成可能（i.e., 64 ビット整数で管理可能）なバケット数および対象数値属性の定義域に応じたしきい値の決定も考えられる．

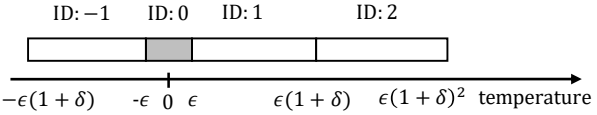


図 4: バケット範囲および ID の静的な生成

3.1.2 階層的シノプシスの構築

本研究では前述したバケット定義および後述する構築アルゴリズムによりシノプシスを構築するが，それによって構築されるのは次元テーブル上で最も粒度の細かい（図 3 における day）シノプシスである．そのため，次元テーブル集合 $\mathcal{D}$ の情報を基に階層的なシノプシス群を構築する．この構築処理は基本的に下層のシノプシスから順に上層のシノプシスへと集約することで実現できる．

例えば，表 4 に示した既存の BAQ± によって生成した統合シノプシスに対し，提案手法において時間粒度 year で集約したシノプシスは表 5 のようになる．このように，タプル数が肥大化しやすい属性を次元テーブルに基づき集約することで，特に次元テーブルに関与しないワークロード問合せに対する処理高速化が期待できる．この例で言えば，$\pi_2 = \{\text{region}, \text{temperature}\}$ は時間属性を QCS に含んでおらず，各地域における温度に関する集約が目的である．しかし，表 4 のように時間属性をまとめない場合，CNT_2 列に見られるように頻度が 0 となるタプルが頻出する．このようなタプルは最終的な結果に影響を与えない一方で問合せ処理時には読み取りの対象として含まれてしまうため，不要な走査による性能悪化が発生してしまう．それに対し，提案手法では表 5 のように最大粒度で集約されたシノプシスを利用できるため，このような無駄なタプルの走査を抑制できる．

なお，表 5 において temperature 属性は $\epsilon = 0.1$ としてバケット化している．また，静的なバケット化を採用したことで，各バケットの最小値・最大値を保持する必要がない点にも注意する．改善前の BAQ± では各バケットの最小値・最大値もシノプシス上で保持しなければならないためその分シノプシスサイズが肥大化しまう恐れがあったのに対し，本手法ではよりコンパクトなシノプシスを実現している．

3.2 オンラインの問合せ処理の拡張

オンライン問合せの処理方法は前章で述べた既存研究と概ね同様であるが，次元テーブルによる階層的シノプシスにあわせて処理を拡張する．BAQ± の問合せ処理では，与えられた問合せに対しまずシノプシス上で指定された選択条件を満たすバ

表 5: $\pi_1 = \{\text{region}, \text{date}\}$ および $\pi_2 = \{\text{region}, \text{temperature}\}$ に対する year 層のシノプシス

SID	region	year	temp (ID,mean)	CNT_1	CNT_2
1	Middle East	2006	(142, 98.5)	1	1
2	Middle East	2007	(143, 103.3)	1	2
3	Europe	2006	(-100, -12.8)	2	3
4	Asia	2008	(-96, -10.7)	3	2
5	Middle East	2009	(144, 110.1)	2	1
6	Europe	2012	(112, 23.2)	1	1

ケットと部分的に満たすバケットとを区別しつつその結果を計算する．その後，各計算結果から部分的に条件を満たすバケットを近似処理した際に発生しうる最大の誤差を見積もり，それがユーザ指定の誤差上限を下回っていればそのまま回答を返す．一方で，最大誤差を上回る場合は元テーブルの部分的に条件を満たすバケットにあたるタプルを読み取り，正確な値を計算することで厳密な誤差保証を提供する．

次元テーブルの導入後は，このような誤差の見積もりと詳細なデータの読み取りを階層的シノプシス群でも行う．例えば，図 3 であれば，問合せ処理時にまず読み取られるのは year のシノプシスとなる．その後，year のシノプシスで誤差上限が満たされなかった場合は，選択条件に応じて quarter ないし week のシノプシスを読み取り，year シノプシスにおいて部分的に条件を満たしたバケットの値を精査する．この処理を繰り返し，day シノプシスにおいても誤差が満たされなかった場合は，次元テーブル導入前と同様に元テーブルを読み取り，正確な値を計算する．このように次元テーブル導入後は複数回のシノプシス読み取りが必要となるが，上層のシノプシスは下層のシノプシスに比べタプル数が十分小さくなるため，その悪影響もある程度抑制できると考えられる．

4 シノプシス構築・更新の高速化

本章ではシノプシス構築を高速化するため，新たなフレームワークをベースとしてバケットを再定義してシノプシスの構築アルゴリズムを提案する．BAQ± と BAQ におけるシノプシス構築は部分シノプシスの構築と統合シノプシスへの統合の 2 手順に分けられるが，特に部分シノプシスの構築がボトルネックとなる．部分シノプシスの構築は各ワークロード問合せに対して必要となるため，ワークロードとなる問合せが増えるほど処理時間が増大する．特に，全ての QCS がカテゴリ属性しか含まない場合は元テーブルを一度走査するだけで部分シノプシスを構築できるが，数値属性を含む場合はソート処理のために複数回の走査が必要となる．

以上の問題を解決するために，4.1 節では再定義したバケットに基づくシノプシスの構築手法を述べる．また，4.2 節では元テーブルに対する更新操作への追従について説明する．

4.1 シノプシス構築アルゴリズム

本論文では BAQ± と同様にワークロード問合せの存在を想定し，シノプシスを高速的に構築する手法について提案する．主なアイデアはワークロードの QCS に基づいて元のテーブルのタプルをスキャンし，関連する QCS の対応する CNT と平均値を保存する．QCS 間に同じ属性が存在する場合がある，つまりターゲット QCS にはいくつかの値がすでにシノプシスに存在している可能性がある．言い換えれば，QCS に関する全て値を挿入せずに直接更新できる．また，シノプシスにターゲット QCS 値が含まれていない，同様に他の QCS 値を含むレコードがある場合，挿入操作を行わずにレコードを更新することもできる．シノプシスに以上の更新可能性がない場合，タプルを新

Algorithm 1: GenerateSynopsis(T, Q)

Input: T, Q // 元テーブル T と T の QCS
Output: S // 構築されるシノプシス

1 $S \leftarrow \emptyset$
2 **foreach** $t \in T$ **do**
3 　t の数値属性を ID へ変換
4 　**foreach** $\pi \in Q$ **do**
　　// シノプシスに T の QCS を確認
5 　　**if** $\exists r \in S, r.\pi = t.\pi$ **then**
　　　// update numerical attributes A's mean
6 　　　$\mathrm{mean}(r.\pi)_A = \frac{\mathrm{mean}(r.\pi)_A \times \mathrm{SF}_{r.\pi} + t.\pi_A}{\mathrm{SF}_{r.\pi}+1}$
7 　　　r の CNT_π を 1 増やす
8 　　　**continue** // 次の QCS
　　// シノプシスに部分 QCS 値が存在可能
9 　　**foreach** $\pi' \in Q \setminus \pi$ **do**
　　　// $|\pi \cap \pi'|$ の降順で QCS 値を確認
10 　　　**if** $\exists r \in S, \{r.\pi'! = NULL \wedge r.\pi = NULL\} \vee$
　　　$\{r.\pi \cap \pi' = t.\pi \cap \pi' \wedge \forall a \in r.\pi \setminus \pi', a = NULL\}$
　　　then
　　　　// update categorical attributes
11 　　　　update $a \in r.\pi \setminus \pi'$ by $t.\pi \setminus \pi'$
　　　　// update numerical attributes A's mean
12 　　　　$\mathrm{mean}(r.\pi)_A = t.\pi_A$
13 　　　　r の CNT_π を 1 に設定
14 　　　　**break**
　　// シノプシスに T の QCS がない場合タプルを追加
15 　　**if** $\nexists r \in S, r.\pi = t.\pi$ **then**
16 　　　$t.\pi$ で r を S に挿入
17 　　　$\mathrm{mean}(r.\pi)_A = t.\pi_A$
18 　　　r の CNT_π を 1 に設定

たなレコードとして挿入する．提案手法では以上のアイデアによるデータ処理の優先順位を付け，対応属性の値，CNT および平均値を更新/挿入する．

具体的には，アルゴリズム 1 に示すように，T を元テーブルとする．QCS 集合 q で構成される Q およびデータ T を入力データとし，シノプシス S を初期化する（1 行目）．続いて，順番に T のタプルについて，各数値データを式 (eq:renew) によるバケット ID に変換し，各 QCS q によるシノプシス S へ追加する（2–4 行目）．シノプシス S にレコード r の π 値がタプル t の π 値と同じ場合，対応する mean と CNT を更新する（5–8 行目）．π' が π と同じ値をいくつか持つ場合，提案手法は挿入操作の代わりに残りの値を格納する空きスペースを見つけて更新する（9–14 行目）．タプル t の π 値がシノプシス S に存在しない場合，シノプシス S に π 値，CNT および mean を追加する（15–18 行目）．

例として，表 1 において，2 つのワークロード問合せに対応する QCS π_1 = {region, year} と π_2 = {region, temperature} が与えられたとする．QCS の π_1 と π_2 は属性 `region` が共通であるため，表 1 の `ID=9` のタプルを処理する場合がある．シノプ

シスの 5 番目のレコードは既に π_2 の region 値を持っている．temperature が空きであるため，対応するレコードの tempID，平均値および CNT を更新する．同様に，表 1 の ID=10 を持つ π_2 の値に対してはシノプシスの 6 番目のレコードを更新して保存する．以上のアルゴリズムにより，シノプシスのレコード数は表 5 まで削減できる．

4.2 元テーブルの更新への追従

本研究では静的なバケット ID を導入したため，元テーブルの更新に対して効率的にシノプシスの内容を追従させられる．BAQ および改善前の BAQ± では，元テーブルの更新があった際にシノプシスの完全な再構築が必要であった．そのため，特に更新頻度が高い場合にシノプシスの情報を最新に保つことは困難であり，BAQ を基にした手法の課題の 1 つであった．一方で，本研究では静的なバケット化により，更新が発生したタプルがシノプシスのどの箇所に影響を及ぼすかが一意に特定できる．つまり，既存研究のようにシノプシスを再構築する必要がなく，影響が発生する箇所のみ増分更新することでシノプシスを最新の状態に保てる．

更新操作について詳細を説明する．まず式 (5) を用いて更新する数値データのバケット ID を計算する．提案手法では 1 の 4-7 行目の操作と同様に，同じカテゴリ値とバケット ID を持つレコードを更新する．挿入操作を処理する場合，QCS の値がすでに概要に存在すれば，対応するレコードの平均値と CNT を更新するだけで完了する．QCS の値がすでに概要に存在しない場合，提案手法は 1 の 14-17 行目に示すように，新たな情報をレコードとしてシノプシスに挿入する．削除操作に関しては，削除条件に該当するレコードの平均値と CNT を更新することで同様に対処可能である．

5 実験および評価

本章では，提案手法に関する実装と評価する．BAQ± および BAQ に対し，オフラインの構築時間，メモリ使用量，オンラインの遅延時間および最大相対誤差の 4 点で分析する．

5.1 準　　備

a） データセット

本研究では 3 つのデータセットを使用した．1 つ目は 7 列（6 つのカテゴリ列と 1 つの数値列）を持つ Kaggle の temperature データセットである．この 280 万レコードがあるデータセットには，321 個都市の 36 年間の気温の統計が含まれている．2 つ目はよく知られている TPC-H ベンチマークからの scale=1 の 900 万のレコードを含むデータセットであり，5.2 節の実験では 5 列（3 つのカテゴリ列，1 つの数値列と 1 つの時間列）の orders テーブルを使用する．3 つ目は scale=1 の TPC-H ベンチマークからの lineitem テーブルであり，6 列（5 つのカテゴリ列と 1 つの数値列）について実験する．

b） 問　合　せ

各データセットのワークロードについて説明する．order の場合が{status, priority, clerk}, {status, price, date}, {priority}, lineitem の場合が{discount, tax, price}, {flag, status}, {flag, shipmode}，temperature の場合が{country, city, temperature}, {country, region}, {city, year}で設定する．オンラインの実験について，各データセットについて 310 個の問合せを生成した．これには，選択条件がない MIN/MAX 問合せ 2 個，SUM/AVG 問合せ 2 個，条件がカテゴリ属性のみの COUNT 問合せ 50 個，MIN/MAX 問合せ 50 個，SUM/AVG 問合せ 50 個，選択条件に数値属性を含む COUNT 問合せ 52 個，MIN/MAX 問合せ 52 個，SUM/AVG 問合せ 52 個である．orders データには時間属性が存在するので，各粒度について COUNT，SUM と AVG 問合せを追加する．各数値属性の ϵ を 0.1 とする．

c） 評 価 項 目

提案手法と既存手法を 4 つの角度から分析する：(1) オフラインのシノプシスの生成時間（分），(2) メモリ使用量，(3) オンラインの遅延時間（ミリ秒），(4) 最大相対誤差．評価実験では，提案手法，BAQ± と BAQ をそれぞれ 30 回実行し，平均値を結果とする．既存手法 BAQ± および BAQ では，部分シノプシスの構築時間および統合シノプシスの生成時間の合計値をオフラインのシノプシスの生成時間とする．統合シノプシスのサイズをメモリ使用量とする．

5.2 誤差上限に対する性能比較

誤差限界 δ を 0.05，0.1，0.15，0.2，0.25 を設定して，提案手法と BAQ±，BAQ を比較する．

5.2.1 シノプシス生成時間

オフラインにシノプシスを生成する時間について，orders，lineitem と temperature データセットに対する計測結果を図 (5a)，図 (5b) と図 (5c) に示す．

BAQ± と BAQ について，orders データを使用するとき提案手法は BAQ± に対して 35–50%，lineitem の場合が 55–57%，temperature の場合が 39–48%となっている．BAQ より提案手法は orders データの場合が最大 45%，lineitem が最大 65%，temperature が最大 58%の時間で計算できる．

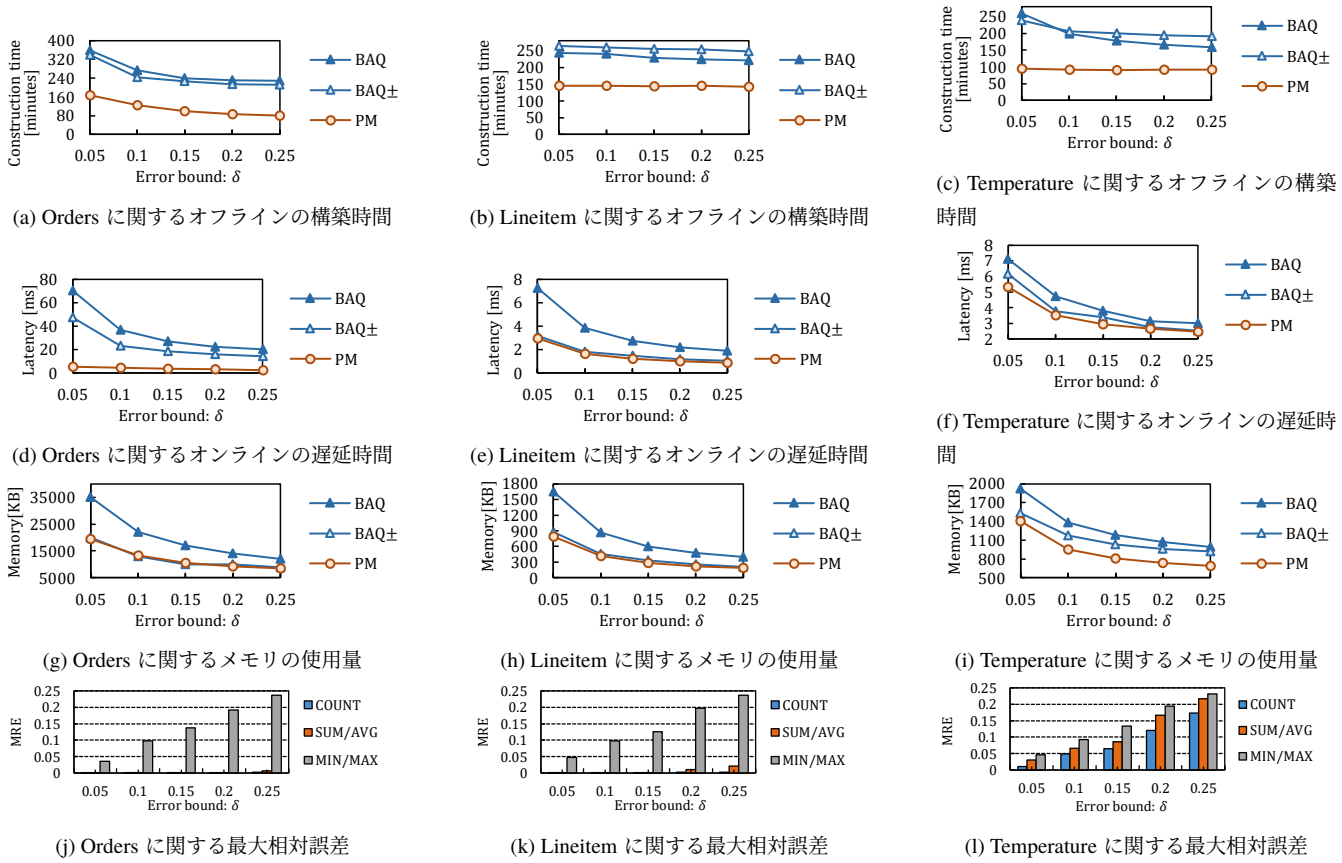
ここで簡単な分析をする．BAQ および BAQ± 手法ではテーブルと各シノプシスのスキャンが複数回必要なため，構築時間が長い．提案手法はテーブルの各タプルについて各属性集合 π によるシノプシスをスキャンするため，構築時間が BAQ と BAQ± より小さい．また，orders のような時間属性がある場合，複数のシノプシスに生成時間が必要になるが，各粒度のシノプシスの生成時間が短いので，全体の構築時間は既存手法より小さい．

5.2.2 遅 延 時 間

提案手法，BAQ± と BAQ でオンラインの平均遅延時間の実験結果を図 (5d)，図 (5e) と図 (5f) に示す．

orders データについて提案手法のオンラインの遅延時間は BAQ± に対して 11–19%，lineitem の場合が 83–93%，temperature の場合が 86–97%となっている．BAQ に対して orders データの場合が 8–14%，lineitem の場合が 40–47%，temperature の場合が 74–84%の遅延時間で問合せを回答する．

具体的に，lineitem と temperature の場合，時間属性がないの



(a) Orders に関するオフラインの構築時間
(b) Lineitem に関するオフラインの構築時間
(c) Temperature に関するオフラインの構築時間
(d) Orders に関するオンラインの遅延時間
(e) Lineitem に関するオンラインの遅延時間
(f) Temperature に関するオンラインの遅延時間
(g) Orders に関するメモリの使用量
(h) Lineitem に関するメモリの使用量
(i) Temperature に関するメモリの使用量
(j) Orders に関する最大相対誤差
(k) Lineitem に関する最大相対誤差
(l) Temperature に関する最大相対誤差

図 5: 実験結果

で，提案手法はシノプシスのみスキャンして短い遅延時間で処理できる．シノプシスのサイズは BAQ と BAQ± より小さいので，提案手法の遅延時間はより短くなる．orders の場合，時間属性に関して選択また集約計算する場合，最大 2 つのシノプシスへアクセスする．各シノプシスは BAQ と BAQ± より短くなり，全体の遅延時間は小さい．

5.2.3 メモリ使用量

提案手法，BAQ± と BAQ で使用したメモリの実験結果を図 (5g)，図 (5h) と図 (5i) に示す．

orders データには時間属性が存在するので，提案手法では複数の階層的シノプシスのサイズの合計値をメモリ使用量とする．orders データについて提案手法の使用するメモリは BAQ± に対してほぼ同じ，lineitem の場合が 84–92%，temperature の場合が 74–91%となっている．BAQ に対して orders データの場合が 55–71%，lineitem の場合が 45–48%，temperature の場合が 68–73%のメモリでシノプシスを構築できる．

以下の分析が行える．提案手法は，実際の問合せ処理において必要な集約のみ考慮してシノプシスを生成する．シノプシスに各 QCS 対応する π の数に最大数を M_π とし，提案手法は各 QCS に対応する列を順番に埋めるため，シノプシス全体のレコード数は M_π である．しかし BAQ± および BAQ では，複数の部分シノプシスに関して元テーブルのタプルを順に走査し統合シノプシスのレコードとして保存するため，レコード数は M_π より大きい可能性が高い．さらに，誤差の限界を増加させると，各バケット ID の範囲が式 (2) により広くになるため，シノプシスのサイズが減少する．

5.3 最大相対誤差

提案手法の最大相対誤差の実験結果を図 (5j)，図 (5k) と図 (5l) に示す．orders，lineitem と temperature データについて提案手法は誤差しきい値以内で問合せを処理できる．提案手法は誤差限界の増加に伴い，最大相対誤差は増加するが，指定された誤差限界 δ よりも小さくなる．

6 関 連 研 究

近年，分析要件の高度化に伴い，増加し続けるデータ量に直面して効率的な問合せ処理の要求が高まっているため，AQP が注目を集めている [8,9]．問題を解決するために，[10,11] のような研究では，AQP に基づいた問合せの処理時間と近似結果の正確さのバランスを取るシステムを開発した．サンプリング，ヒストグラム，シノプシスなどの主要な技術が AQP の基礎である [12]．各技術は AQP の効率を向上させる上で独自の役割を果たし，最新のデータ管理戦略の重要な部分となっている．

Ioannidis ら [13] によると，ヒストグラムはデータ分布の要約と推定に効果的であり，大規模なデータベースシステムでの問合せの最適化と実行を行う．ヒストグラムには等幅ヒストグラム，等深度ヒストグラム，V 最適ヒストグラムなどの形式があり，これらはさまざまな種類のデータや問合せに適用される．データベースの問合せ処理では，ヒストグラムを利用してデータをコンパクトにし，効率的な問合せ処理を可能になる．ただ

し，ヒストグラムはコンパクトであるという利点がある一方で，問合せ処理の精度の保証には問題がある．

また，サンプリングに基づく近似的問合せ処理（SAQ）はAQP に広く研究されている．サンプリングに基づく近似的問合せ処理では，オンラインに問合せの列によるサンプルを抽出し，そのサンプルを使用して結果を推定する．サンプリングに基づく近似的問合せ処理では，パフォーマンスと精度の間でトレードオフが実現されるが，いくつかの問題が存在する．まず，サンプリングベースの手法には厳密な誤差保証がない．加えて，サンプリングベースの手法はロングテールなどの歪んだデータに対してパフォーマンスが低下になる可能性がある．つまり，事前に対象データの分布を把握しておくことが必要にある．

最近では，近似結果の精度を向上させるために事前計算とサンプリング手法を組み合わせる研究がある [5, 14–16]．Sample+Seek [5] は，計算されたサンプルを増強するために反転インデックスと統合することで問合せの精度を高めるアプローチとして注目されている．Sample+Seek は，問合せの処理時間と誤差のトレードオフの重要性を強調している．しかし，Sample+Seek は，サンプルと反転インデックスにおいて使用するスペースの膨大さと複雑さをもたらす可能性がある．

BAQ [4] は，これらの課題に対して新しいフレームワークを提案し，オフラインのシノプシス構築とオンラインの近似処理を通じて精度を向上させ，誤差保証を確保できる．ただし，データの均一分布を仮定する BAQ は，分布が異なるデータセットに適用する場合に不正確さを引き起こす可能性がある．さらに，シノプシスのサイズと遅延時間には改善の余地がある．BAQ に基づいて，BAQ± [6] はバケット計算手法を改善し，新しいオンライン近似計算方法を提案する．BAQ± の提案手法により，厳密な誤差保証，より小さなシノプシスおよび短縮された遅延時間が実現される．しかし，BAQ と BAQ± はデータを数値属性とカテゴリ属性として定義および処理しているため，カーディナリティが大きい属性を処理する際には効率が悪い．さらに，部分と統合シノプシスに基づく構築プロセスによる前処理時間が問題になる．

7 まとめと今後の課題

本稿では，近似的問合せ処理におけるシノプシス構築の高速化に対する提案手法を述べた．また，上記のフレームワークを実装し，3 つのデータセットを用いて評価実験を行った．実験結果により，提案手法は既存手法 BAQ と BAQ± よりシノプシスサイズ，構築時間および遅延時間において優れていることを示した．今後の課題としては，今回議論した提案手法に基づくより広いデータ範囲をサポートできるフレームワークの改善について詳細な検討を行う予定である．

謝　　辞

本研究の一部は JSPS 科研費 JP20K19804，JP21H03555，JP22H03594 の助成，名古屋大学および JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2120 による「名古屋大学融合フロンティアフェローシップ」の支援，および国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）の結果得られたものである．

文　　献

[1] B. Walenz, S. Sintos, S. Roy, and J. Yang, “Learning to sample: Counting with complex queries,” *Proc. VLDB Endowment*, vol. 13, no. 3, pp. 389–401, 2019.

[2] S. Agarwal, H. Milner, A. Kleiner, A. Talwalkar, M. I. Jordan, S. Madden, B. Mozafari, and I. Stoica, “Knowing when you’re wrong: Building fast and reliable approximate query processing systems,” in *Proc. SIGMOD*, pp. 481–492, 2014.

[3] Q. Ma and P. Triantafillou, “DBEst: Revisiting approximate query processing engines with machine learning models,” in *Proc. SIGMOD*, pp. 1553–1570, 2019.

[4] K. Li, Y. Zhang, G. Li, W. Tao, and Y. Yan, “Bounded approximate query processing,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2262–2276, 2019.

[5] B. Ding, S. Huang, S. Chaudhuri, K. Chakrabarti, and C. Wang, “Sample+Seek: Approximating aggregates with distribution precision guarantee,” in *Proc. SIGMOD*, pp. 679–694, 2016.

[6] 倪 天嘉，杉浦 健人，石川 佳治，陸 可鏡, “シノプシスの最適化に基づく近似問合せ処理の高効化,” in **第 *13* 回データ工学と情報マネジメントに関するフォーラム**, 2022.

[7] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, ch. 3. Morgan Kaufmann, 4 ed., 2022.

[8] S. Chaudhuri, B. Ding, and S. Kandula, “Approximate query processing: No silver bullet,” in *Proc. SIGMOD*, pp. 511–519, 2017.

[9] B. Mozafari and N. Niu, “A handbook for building an approximate query engine,” *IEEE Data Eng. Bull.*, vol. 38, no. 3, pp. 3–29, 2015.

[10] S. Agarwal, A. Panda, B. Mozafari, S. Madden, and I. Stoica, “BlinkDB: Queries with bounded errors and bounded response times on very large data,” in *Proc. EuroSys*, pp. 29–42, 2013.

[11] Y. Park, B. Mozafari, J. Sorenson, and J. Wang, “VerdictDB: Universalizing approximate query processing,” in *Proc. SIGMOD*, pp. 1461–1476, 2018.

[12] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, *Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches*, pp. 1–294. Delft: Now Publishers, 2011.

[13] Y. Ioannidis, “The history of histograms (abridged),” *Proc. VLDB Endowment*, pp. 19–30, 2003.

[14] J. Peng, D. Zhang, J. Wang, and J. Pei, “Aqp++: Connecting approximate query processing with aggregate precomputation for interactive analytics,” in *Proc. SIGMOD*, pp. 1477–1492, 2018.

[15] X. Li, J. Han, Z. Yin, J.-G. Lee, and Y. Sun, “Sampling cube: A framework for statistical olap over sampling data,” in *Proc. SIGMOD*, pp. 779–790, 2008.

[16] B. Babcock, S. Chaudhuri, and G. K. Das, “Dynamic sample selection for approximate query processing,” in *Proc. SIGMOD*, pp. 539–550, 2003.

単一表クエリを対象とするシノプシス埋込みによる近似問合せ処理の評価

高田　実佳†　　合田　和生†

† 東京大学情報理工学系研究科　〒 153–8503 東京都目黒区駒場 4-6-1
E-mail: †{mtakata,kgoda}@tkl.iis.u-tokyo.ac.jp

あらまし　業務データを収集し，蓄積したビッグデータを分析することによる業務の改善・新たな知識発見への期待が高まっている．そのような分析では集計が頻繁に実施される．集計には，必ずしも従来データベースが求める正確な値は必要ではなく，それよりもレスポンスの早さが求められることがある．本論文では、索引構造に付加情報を埋込むことによって高速に近似問合せ処理を実行する方式を既存 RDBMS で実装し，検証する．

キーワード　データベース，索引，近似問合せ

1　はじめに

多くの企業や組織では，日々様々な業務データが生成され，そして収集・蓄積された大量データを分析することによって業務の改善や新たな知識発見に期待が高まっている．業務データの蓄積・管理には構造データベースが広く利用されており，様々な合計値や最大値，最小値など集計値の問合せ処理が分析には必要とされる．例えば，小売業では，日々の売り上げが期待できる製品，製品数や人員を適切に調整する為，製品の購買履歴を用いて，一日の売り上げ総数，在庫数，来客数等の定期的な集計が必要とされている [1]．この様な分析はインタラクティブに実施されることが多く，その為にはデータベースに蓄積されたデータに対し，高速に様々な集計値計算が実行されることが求められる．

高速な集計値計算に向けて，正確な値を高速に求めるデータベースの検索技術はこれまで多数提案されてきた [2]．代表的なものの一つには索引があり，中でも B+Tree [3] は多くの関係データベースで利用されている． B+Tree は範囲検索を効率的に行えるメリットがあり集計計算に必要な範囲のデータを高速に検索する為に有効である．しかし，日々業務データが生成・蓄積されることで，そうして増大した業務データを対象とした集計計算の場合，索引のサイズが大きくなり，最下層のリーフページまでノードを辿ると検索時間が増大するという問題が生じている．

インタラクティブな分析における高速な集計値計算を考慮した時に，正確性は必ずしも必要ではない．こうした観点に着目し，[4] らは，既存の B+Tree など索引にシノプシスを埋込むことで似問合せの高速化手法 (Synopsis-aware search; SAS, Synopsis-aware search plus inter-attribute correlation; SAS+) を提案している．SAS は，B+Tree の中間ノードに下位ノードの最大値，最小値，合計値，総数といった統計情報をシノプシスとして持たせておく事で B+Tree のリーフページまで辿らずとも，近似値を求めることができるという近似問い合せ手法であり，SAS+は SAS を発展させ，複数カラム間の相関を考慮してシノプシスの利用可否を決定する近似問合せ手法である．本論文では，[4] の研究を発展させ，業務データの蓄積・管理に広く利用されている関係データベースシステムである PostgreSQL [5] にシノプシスを埋込んだ索引による近似問合せ処理 (SAS, SAS+) を組込み，その試作による評価結果を示す．

本論文の構成は，以下の通りである．第 2 章では，既存の近似問合せとその課題について言及する．第 3 章では，シノプシス埋込み索引による近似問合せ処理について説明し，第 4 章では，その評価を示す．第 5 章では，関連文献について言及し，第 6 章では，今後に向けた課題と将来展望について纏める．

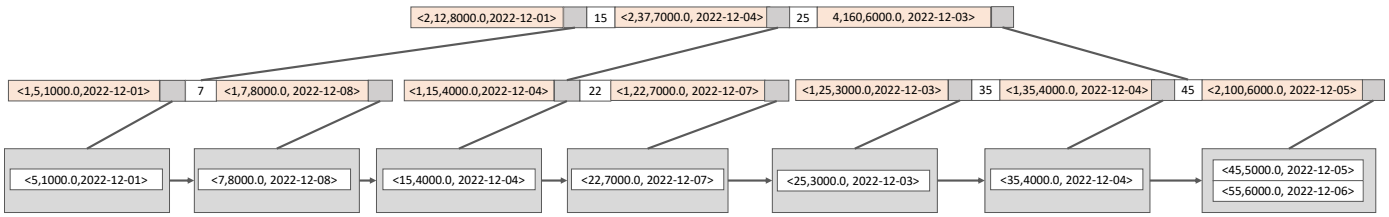
2　近似問合せ処理

本章では，近似問合せ処理へのニーズと，既存の近似問合せ処理およびその問題点について述べる．近年，ビックデータを用いた分析に期待が高まり，広くインタラクティブな集計計算が実施されている．例えば製造業では，IoT センサ等の発達により，原材料の調達，設計，製造，流通，といった製品のライフサイクルの管理のデータを計測でき収集・蓄積が進んでいる [6]．蓄積された大量のデータを利用し，日々の受注製品数，最大受注製品，売上金額なと集計値を定期的に計算し，その値に基づき，調達すべき原材料の種類，個数や，製造すべき生産量，出荷量，販売量の適切な決定に活用している [7]．このようにデータを活用した業務改善・効率化には，統計的な集計値の計算が頻繁に実施される必要がある．この様な集計計算には，必ずしも一円単位の売り上げ金額や一桁単位の生産数といった正確な値は必要ではなく，それ以上に様々な種類の集計値をインタラクティブに実行できることに価値が求められる場合がある．即ち，合計値，最大・最小値，平均値など様々な集計値の問合せに対してインタラクティブなスピードで結果を返すことが求められる．

集計値の近似値を求める近似問合せ処理はこれまで長年議論されてきた．代表的な手法は 2 通りに分類され，蓄積されたデータからサンプリングしたデータのみを用いて問合せ時に即座に集計値を計算・推定する方式であり，もう一つは事前に

ORDERKEY	PRICE	SHIPDATE
5	1000.0	2022-12-01
7	8000.0	2022-12-08
15	2000.0	2022-12-02
22	7000.0	2022-12-07
25	3000.0	2022-12-03
35	4000.0	2022-12-04
45	5000.0	2022-12-05
55	6000.0	2022-12-06

(a) テーブル例



(b) シノプシス埋込み B+Tree の例

図 1: シノプシス埋込み索引の例

データベースに蓄えられたデータを用いて集計値など付加情報を計算し保存しておくことで，問合せ時にその値を用いて集計値を返答する方式である．しかし，前者は事前準備が不要のため事前準備に時間や付加情報を保村しておくためにストレージ負荷がかからないが応答時間が長いというデメリットがあり，後者は前者に比べ応答時間が高速だが，付加情報保存のためのストレージ負荷がかかるというデメリットがある．その上，多くの既存提案は近似問合せの為に新たなデータ構造や処理を備える必要がある．

この様な問題に対し，[4] は既存のデータ構造に付加情報であるシノプシスを埋込むことによって近似問合せの高速化する手法 (Synopsis-aware search; SAS, Synopsis-aware search plus inter-attribute correlation; SAS+) を提案している．SAS は，事前にシノプシスを埋め込んだ索引を生成しておき，問合せクエリが入力されると，その索引に埋め込まれたシノプシスを利用する事で探索範囲を狭め，高速に近似解を返す検索方式であり，ストレージ負荷が約 2%程度に抑えながら最大 25% 以上近似解の応答時間を高速化を達成できる場合があることを示している．

図 1 を用いて SAS の有効性を説明する．図 1(a) は関係データベースで管理されたテーブルの例であり，そのテーブルの ORDERKEY を索引キーとし，付加情報を埋め込んだ B+Tree 索引の例を図 1(b) に示す．図 1(b) では，一般的な B+Tree に加え，その各中間ノードには，下位ノードの集計値をシノプシスと呼ばれる付加情報として埋め込まれている．図 1(b) の例では，付加情報として COUNT(ORDERKEY)， SUM(ORDERKEY)，MAX(PRICE)，MIN(SHIPDATE) を保有している．B+Tree はリーフノードと中間ノードをもつが，一般的にリーフノードにテーブルレコードを保有するためデータサイズが大きくなるが，中間ノードは軽量である．その特性を利用することで，中間ノードに下位ノードの集計値をシノプシスとして埋め込んでも，シノプシスによるストレージオーバーヘッドは軽量になると見込まれる．そして SAS による問合せ処理では．一般的な B+Tree を用いた探索同様，ルートから深さ優先探索を実施し，問合せクエリに対してマッチする集計値を中間ノードで発見した時点で，それより下位のノード検索をスキップする．これにより，問合せ検索において索引のリーフノードの探索時間を削減し，問合せ処理時間を大幅に短縮できることが検証されている．しかし，SAS は事前に準備した付加情報が利用できる場合においては非常に有効であるが，ユーザの多様な問合せニーズに対応可能や広く利用されている既存のデータベースシステムとの併用について検討が十分行われていない．本論文では，既存のデータベースシステムへの近似問合せの組込んだ高速な近似問合せ処理を提案する．

3 既存関係データベースシステムへの近似問合せ手法の組込み

本章では，ユーザの多様な問合せニーズに対応可能や広く利用されている既存のデータベースシステムとの併用に向けて，[4] らの提案した高速な近似問合せ処理方式を既存データベースに取り込む事によって，多様な問合せに対して高速に近似解を求める問合せ方式を提案する．

まず，ユーザの多様な集計値計算ニーズに応える為，広く利用されている既存データベースシステムである PostgreSQL [5] へのシノプシスを埋め込み索引を用いた近似問合せを組込む実装方式を提案する．通常，データベース管理システムは，データベース構築時に指定したデータ領域に存在しているデータにアクセスすることが一般的である．PostgreSQL はそれに加え，外部データラッパ (FDW; Foreign Data Wrapper) をサポートしており，その機能により外部サーバ (Foreign Server) やその外部サーバに存在している外部テーブル (Foreign Table) を登録しておくことで，PostgreSQL インターフェースを用いて

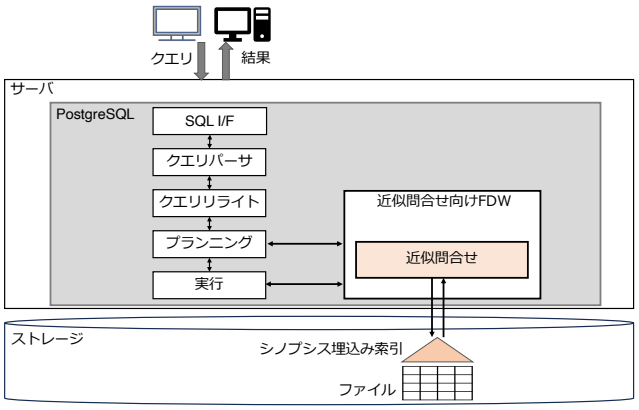


図 2: PostgreSQL 外部データラッパを用いたアーキテクチャ

外部サーバにある外部テーブルを参照できる機能である [8]. 様々な種類の外部データラッパは既に公開されており，例えば，外部のファイルにアクセスして検索する File FDW や別の PostgreSQL サーバにアクセスする Postgre FDW などが既に公開されている．その様な公開済み FDW に加え，PostgreSQL は独自の FDW を開発できるよう API を提供しており，我々はその API を利用する事で近似問合せ処理を実施する独自 FDW を開発する方式を検討に取り入れた．以下その独自 FDW を Approximate_FDW とよぶ.

PostgreSQL は FDW 作成に必要な API を提供しており，本提案ではその中から下記の 7 種類の API を用いる事とする．各 API の概要は以下の通りである.

- GetForeignRelSize: テーブルサイズを見積もる
- GetForeignPaths: アクセスパスを生成する
- GetForeignPlan: プランノードを生成する
- BeginForeignScan: スキャン開始時に呼ばれる
- IterateForeignScan: 上位ノードが 1 行必要とした時に呼ばれる
- ReSacnForeignScan: スキャン位置を先頭に戻したい時に呼ばれる
- EndForeignScan: スキャン終了時に呼ばれる

これら API は PostgreSQL に実行計画と実行処理の 2 種類に大別できる．GetForeignRelSize, GetForeignPaths, GetForeignPlan が実行計画に関連する API であり，BeginForeignScan, IterateForeignScan, ReSacnForeignScan, EndForeignScan が実行処理に関連する API である．ユーザは問合せ前に，PostgreSQL サーバに Approximate FDW，Approximate FDW を用いる外部サーバ，外部テーブルの登録を行うことで，外部テーブルへの問合せを PostgreSQL が認識した際，Approximate FDW を呼び出す事ができる．下記に例を示す様に，外部テーブルは通常テーブル定義と同様に SQL を用いて登録が可能である.

- Approximate_FDW(e.g., approximate_fdw) 登録

```
CREATE EXTENSION approximate_fdw;
```

- Approximate_FDW を用いる外部サーバ (e.g., approximate_server) 登録

```
CREATE SERVER approximate_server
    FOREIGN DATA WRAPPER approximate_fdw;
```

- 外部テーブル (e.g., foreign_table) 登録

```
CREATE FOREIGN TABLE foreign_table (val int)
    SERVER approximate_server;
```

この様に．FDW 機能を用いて，シノプシスを埋め込んだ索引を用いた検索を実現するアーキテクチャを図 2 に提案する．ユーザが PostgreSQL のインターフェースを介してクエリを投入すると，図 2 に示すように，クエリはパース，クエリリライトを経て，プランニング，実行を行う．事前に登録した外部テーブルへの問合せがクエリパーサ検知されると，PostgreSQL は対応する FDW(e.g., Approximate_FDW) を呼び出し，FDW 作成時に登録した実行計画，実行処理を行うので，実行処理内で，シノプシスを埋込み索引を参照し，近似問合せ処理が可能となる．これによりユーザが与えた問合せクエリに対し，独自の処理方式を用いて求めた近似解を返す事ができる．以上の仕組みを用いて，我々は問合せ時にシノプシス埋込み索引を用いて集計計算に対する近似値を求めるプロトタイプを開発し，評価を行った.

4 評　　価

本章では，[4] らの提案した近似問合せ処理（SAS）が外部データラッパ (FDW) 上で既存データベースがもつ索引と同様に実行可能か，および既存データベースがもつ索引に比べ優位性の有無を検証する.

4.1 実 験 環 境

本実験では，ベンチマークセットである TPC-H [9] の dbgen を用いて生成したデータセットを利用する．データセットのスケールファクタ (SF) は，1,10,100 を用意した.

次に，問合せ処理に用いる索引として，一般に多くのデータベースシステムで適用されている B+Tree [2, 3] を用い，以下の索引キーをもつ B+Tree 索引を用意した．全てノードサイズは 8192Byte とした.

- ORDERS (ORDERKEY): ORDERS 表の O_ORDERKEY を索引キーとしてもつ索引
- LINEITEM (ORDERKEY, LINENUMBER): LINEITEM 表の (L_ORDERKEY, L_LINENUMBER) を索引キーとする索引
- LINEITEM (SHIPDATE): LINEITEM 表の L_SHIPDATE を索引キーとする索引.

索引に埋込む付加情報であるシノプシスとしては，以下を用意した.

- **nosyn**: シノプシス埋込みなし
- **syn1**: 索引キーの集計値 (SUM, MAX, MIN, COUNT) (e.g., ORDERS (ORDERKEY) の場合 O_ORDERKEY の集計値)
- **syn2**: syn1 に加え，索引キーとは異なる属性の集計値 (SUM, MAX, MIN, COUNT)

表 1: 単一制約クエリ

Q1	SELECT SUM(L_EXTENDEDPRICE) FROM LINEITEM BETWEEN 5,4000,001 AND 6,000,0001;
Q2	SELECT MAX(L_EXTENDEDPRICE) FROM LINEITEM WHERE L_ORDERKEY BETWEEN 5,4000,001 AND 6,000,0001;
Q3	SELECT MAX(L_SHIPDATE) FROM LINEITEM WHERE L_ORDERKEY BETWEEN 5,4000,001 AND 6,000,0001;

表 2: 複合制約クエリ

Q4	SELECT SUM(L_EXTENDEDPRICE) FROM LINEITEM WHERE L_ORDERKEY BETWEEN 36,000,000 AND 42,000,000 AND L_SHIPDATE BETWEEN 1992-01-01 AND 1992-12-31;

- **syn3**: syn2 に加え，索引キーと索引キーとは異なる属性との相関係数および相関係数計算に必要な集計値

相関係数の計算には，分散の定理

$$V_A = \mu_{A^2} - \mu_A^2 \tag{1}$$

および，共分散の定理

$$V_{AB} = \mu_{AB} - \mu_A * \mu_B \tag{2}$$

を用いて，以下の様に相関係数 r を計算可能である．

$$r = \frac{V_{AB}}{\sqrt{V_A} * \sqrt{V_B}} \tag{3}$$

以上より，**syn3** には，索引の各ノード以下の属性間の相関係数を計算する為，各属性の和，二乗和，属性積の和を相関係数計算に必要な集計値として保有した．

本実験では評価クエリとして，単一表の問合せに対して単一制約をもつクエリ (Q1-Q3) と複合制約をもつクエリ (Q4) を用いた．

本実験では，索引は事前に生成しファイルとして保存し，問合せ時に索引を呼び出し，索引を用いた検索は，IterateForeignScan API 内で実施した．評価クエリでは ORDERS 表および LINEITEM 表を対象としていることから，ORDERS 表および LINEITEM 表を外部表として事前に create する．これにより，外部表に対する問合せクエリを PostgreSQL インターフェース (I/F) から入力すると，近似問合せ処理を実行する FDW(Approximate_FDW) が呼び出される様にした．

評価軸には，各種シノプシスを埋込み索引のリーフノード数，中間ノード数を計測し，シノプシス埋込みによるストレージオーバーヘッドを検証する．また，異なる問合せに対して，一般的な B+Tree の検索と SAS，および相関係数を考慮した SAS+による検索を比較し，エラー率 [10] および問合せに要する実行時間を用いて検証する [4]．なお，全ての実験は，OS は Amazon Linux release 2 (Karoo)，PosgreSQL はバージョン

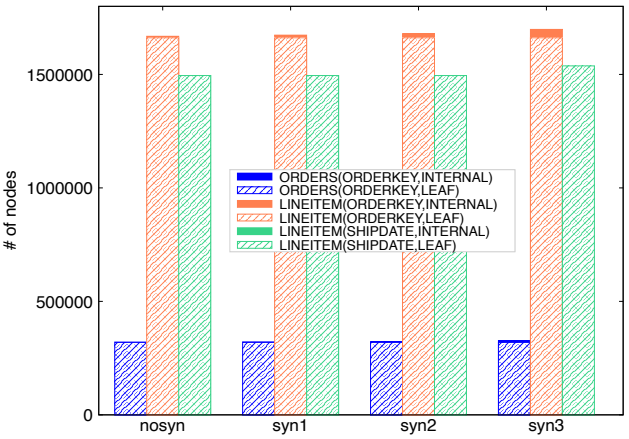


図 3: シノプシス埋込みによるストレージオーバーヘッド

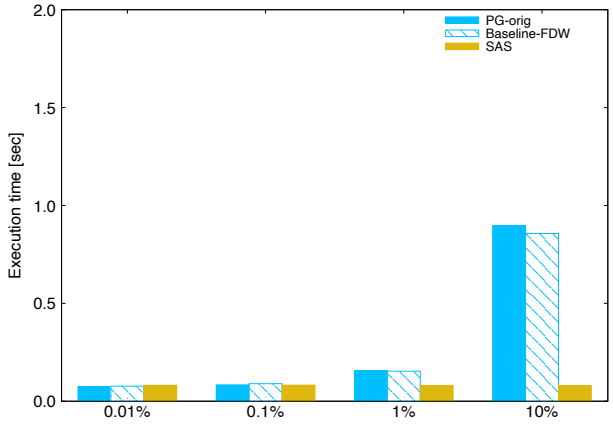


図 4: 異なる選択率に対する問合せ実行時間

14.4 を用い，2CPU コア，14GB メモリ，640GB ストレージの環境下で行なった．

4.2 実験結果

初めに，シノプシス埋込みによる，索引生成に要するオーバーヘッドを検証する．図 3 に，各索引およびシノプシスをもつデータ構造を保有することによるストレージオーバーヘッドを示す．[4] らの示した結果と同様であり，リーフノードが中間ノードに比べ圧倒的にサイズが大きいため，中間ノードはわずかに増加はしているものの，全体のノード数は，シノプシスによるストレージオーバーヘッドは **syn3** でも約 1.83% 程度であることを示している．

次に，PostgreSQL がもつ既存の B+Tree を用いた正解値を求める問合せ処理 (PG-orig) と，外部表 (FDW) に **nosyn** の場合の B+Tree を用いた正解値を検索する方法 (Baseline-FDW)，対して，B+Tree 索引に対して **syn2** のシノプシスを保持し，SAS を Approximate_FDW で実行する方式 (SAS) の実行性能を比較検証する．

まず，異なる選択率に対する問合せ処理の実行時間を比較する．図 4 にスケールファクタ 10 の ORDERS 表に対して 0.01%, 0.1%,1%,10% の選択率で ORDERKEY を取得し COUNT を計算する集計計算に対する実行時間を示す．PG-

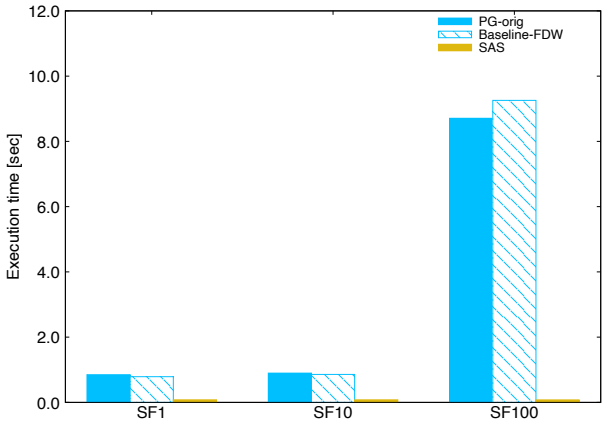


図 5: 異なるデータサイズに対する問合せ実行時間

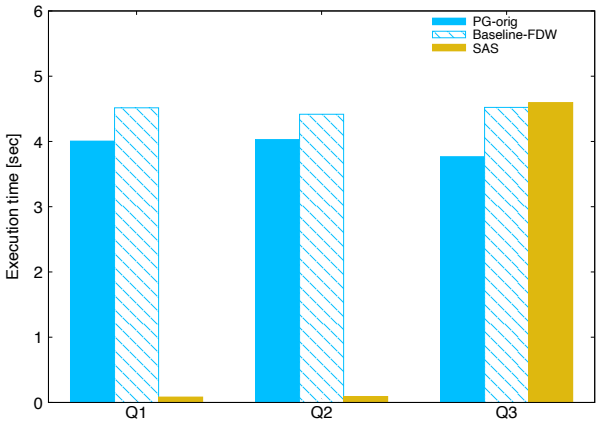
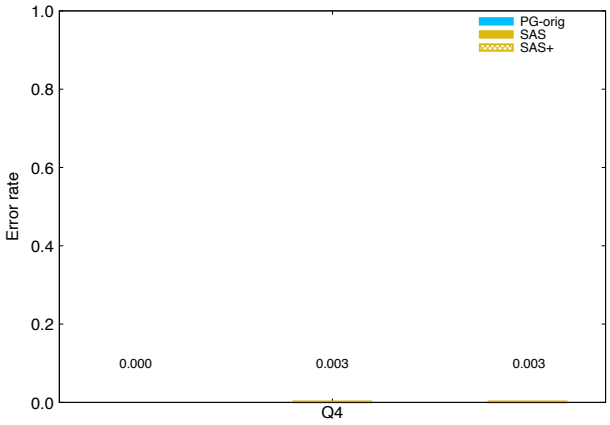
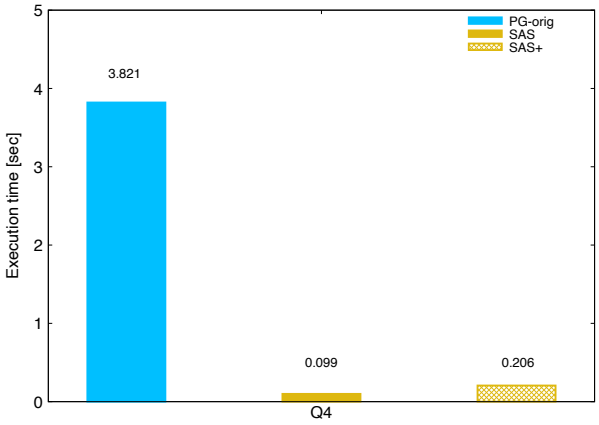


図 6: 異なるクエリにおける問合せ実行時間



(a) エラー率



(b) 実行時間

図 7: 複数制約による近似問合せ処理

orig と Baseline-FDW は選択率によらずほぼ同等性能を確認できており，FDW 上での実装が現実的であることが確認できる．また，選択率が大きくなるにつれ SAS の有効性が大きくなり，最大 91.2%PostgreSQL に比べ高速であることが言える．

続いて，異なるスケールファクタをもつデータセットに対する問合せ処理の実行時間を比較する．図 5 に，SF 1,10,10 の ORDERS 表に対して選択率 10% となる問合せを実行した時の実行時間である．選択率が変わらない場合，スケールファクタが大きい場合，PostgreSQL に比べ最大 99.1% 高速であることが確認できる．これは大きなデータサイズの方がリーフノードが多く，リーフノードの検索範囲が大きくなる為と考えられる．

次に，異なるクエリにおける近似問合せの有効性を検証する．図 6 に単一制約クエリ (Q1-Q3) における問合せ実行時間を比較する．**syn2** がシノプシスとして保持している為，Q1，Q2 では SAS が 97% 以上高速を達成している．一方，**syn2** は Q3 のクエリに対して十分なシノプシスを保持せず従来 B+Tree と同等の実行処理となっているのが確認できる．

さらに，複合制約クエリ（Q4）に対して，近似解のエラー率および実行時間を検証する．図 7(a) にエラー率，図 7(b) に実行時間を示す．PostgreSQL に比べ，エラー率は無視できる程度でありつつ，SAS では約 97.4%，属性間の相関を考慮した SAS+で 94.6% 高速であることが確認できる．本結果は，本実験で用いたデータセットが属性間にほぼ相関を持たず一様に分布している為であると考えられる．

5 関 連 文 献

大量データを収集・蓄積できるようになるに従い，蓄積した大量データから集計値を高速に計算する手法は長く研究されてきた．分析のための高速検索手法は OLAP [11] 問合せとして提案されてきた．代表的なものに索引があり，B+Tree [2,3] をはじめ検索キーを検索に効率的なデータ構造で保持することで高速検索を実現する．Data cube [12] は Group-by など集計に必要な処理をオペレータとして用意することによる処理の効率化を提案している．Materialized View [13] は過去の問合せ結果やその中間データを保存しておくという手法であり，類似する問合せを多く実行する分析系の問合せにおいて大幅な高速化が可能である．Materialized View は大規模データを対象とした検索時間を大幅に削減できるが，その効果の反面，大規模な Materialized View を保存するとストレージコストやメンテナンスコストが増加してしてしまうため，高速性とのバランス調整が求められる．

一方，正確な値ではなく近似値を求める方法として近似問合せ方式 [14,15] があり，それは 2 通りに大別される．1 つは蓄

積されたデータからサンプリングしたデータのみを用いて問合せ時に即座に集計値を計算・推定するオンライン処理型方式であり，もう 1 つは事前にデータベースに蓄えられたデータを用いて集計値など付加情報を計算し保存しておくことで，問合せ時にその値を用いて集計値を返答する事前処理型方式である．前者には，問合せ時にサンプリングしたデータを用いて近似値を推定し，問合せをインタラクティブに繰り返す毎にその近似解の精度を高めていく方式 [16] がある．この方式では事前処理が不要な為，事前準備の時間や事前に処理した結果を保持するストレージのオーバーヘッドが不要であるというメリットがあるが，問合せ時にサンプリングする処理が追加される為問合せ時間が増大するという問題がある．

後者には，事前にシノプシスと呼ばれる付加情報を計算し保持しておき，それらの情報を用いて問合せ時に近似値を高速に返す方式がある．代表的なものには，全体のデータセットを圧縮した情報を保有する wavelet [17,18] 方式がある．また，データセットの分布を示すヒストグラムを保持して近似解を推定する histogram 方式 [19] がある．この様な事前処理型方式は，オンライン処理型方式に比べ問合せ時の計算コストを抑え近似値の精度が高いというメリットがあるが，事前に付加情報を計算し保持しておくストレージオーバーヘッドが問題である．

上記の様なメリット・デメリットから，オンライン処理型方式と事前処理型方式を組み合わせた近似問合せ方式も提案されている．BlinkDB [20] はサンプリングを事前に実施することで問合せ時のサンプリングコストを軽減させながら，ワークロードを元に定期的に適切なサンプリングを実施する為，より正確な近似値を算出している．ストレージオーバーヘッドを抑制する為，問合せ処理をグルーピングし，保持するサンプルデータを低減させている．[21] らは，シノプシスとサンプリングを組み合わせて，インタラクティブな応答時間で集計値の近似値を返す AQP++を提案しており，[10] らは問合せに応じてどちらを使って近似値を求めるかを決定する方式で高い精度を達成している．この様な研究が実施されているが，特定のワークロードにおいて近似問合せの精度を重視する内容が中心であり，インタラクティブに多様な分析を実施したいユーザニーズに十分応えるものはこれまでの調査では見つかっていない．

6 おわりに

本論文では，多様な問合せに対応可能な近似問合せとして，シノプシスと呼ばれる付加情報を埋込んだ B+Tree 索引索引を利用した問合せ処理によって集合値の問合せに対する近似解を高速に検索する方式を既存の関係データベース管理システムである PostgreSQL に組込み，検証した．実験では，PostgreSQL インターフェースを介して集合値の問合せを実行し，シノプシスを埋込んだ索引によるストレージオーバーヘッドが無視できる程度であることと，従来の B+Tree を要する PostgreSQL の検索に比べ，シノプシスを埋め込んだ索引を用いた集合値の計算処理が最大 99.1% 高速である事を確認した．今後は，複雑な問合せに対する近似問合せ処理方式について検討を進める．

文　献

[1] 川村晃一. 売上分析. 計測と制御, Vol. 28, No. 1, pp. 17–22, 1989.
[2] Ramez Elmasri. Fundamentals of database systems seventh edition. 2021.
[3] David J Abel. A b+-tree structure for large quadtrees. *Computer Vision, Graphics, and Image Processing*, Vol. 27, No. 1, pp. 19–31, July 1984.
[4] Hiroki Yuasa, Kazuo Goda, and Masaru Kitsuregawa. Exploiting embedded synopsis for exact and approximate query processing. In *Database and Expert Systems Applications*, pp. 235–240. Springer International Publishing, 2022.
[5] Postgressql. `https://www.postgresql.org/`. 2024-01-10 参照.
[6] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin-driven product design, manufacturing and service with big data. *Int. J. Adv. Manuf. Technol.*, Vol. 94, No. 9-12, pp. 3563–3576, February 2018.
[7] Gerald Rebitzer, Tomas Ekvall, Rolf Frischknecht, Davis Hunkeler, Gregory Norris, Tomas Rydberg, W-P Schmidt, Sangwon Suh, B Pennington Weidema, and David W Pennington. Life cycle assessment part 1: framework, goal and scope definition, inventory analysis, and applications. *Environ. Int.*, Vol. 30, No. 5, pp. 701–720, July 2004.
[8] Postgressql insider. `https://www.postgresql.fastware.com/postgresql-insider-fdw-ove`. 2024-01-10 参照.
[9] Tpc-h. `https://www.tpc.org/tpch/`. 2024-01-09 参照.
[10] Xi Liang, Stavros Sintos, Zechao Shang, and Sanjay Krishnan. Combining aggregation and sampling (nearly) optimally for approximate query processing. In *Proceedings of the 2021 ACM SIGMOD International Conference on Management of Data*, SIGMOD '21, pp. 1129–1141, New York, NY, USA, June 2021. Association for Computing Machinery.
[11] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and OLAP technology. *SIGMOD Rec.*, Vol. 26, No. 1, pp. 65–74, March 1997.
[12] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing Group-By, Cross-Tab, and Sub-Totals. *Data Min. Knowl. Discov.*, Vol. 1, No. 1, pp. 29–53, March 1997.
[13] Imene Mami and Zohra Bellahsene. A survey of view selection methods. *SIGMOD Rec.*, Vol. 41, No. 1, pp. 20–29, April 2012.
[14] Surajit Chaudhuri, Bolin Ding, and Srikanth Kandula. Approximate query processing: No silver bullet. In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data*, SIGMOD '17, pp. 511–519, New York, NY, USA, May 2017. Association for Computing Machinery.
[15] Kaiyu Li and Guoliang Li. Approximate query processing: What is new and where to go? *Data Sci. Eng.*, Vol. 3, No. 4, pp. 379–397, December 2018.
[16] Joseph M Hellerstein, Ron Avnur, Andy Chou, Christian Hidber, Chris Olston, Vijayshankar Raman, Tali Roth, and Peter J Haas. Interactive data analysis: the control project. *Computer*, Vol. 32, No. 8, pp. 51–59, August 1999.
[17] Ioannis Mytilinis, Dimitrios Tsoumakos, and Nectarios Koziris. Distributed wavelet thresholding for maximum error metrics. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*, SIGMOD '16, pp. 663–677, New York, NY, USA, June 2016. Association for Computing Machinery.
[18] Abdul Naser Sazish and Abbes Amira. An efficient archi-

tecture for HWT using sparse matrix factorisation and DA principles. In *APCCAS 2008 - 2008 IEEE Asia Pacific Conference on Circuits and Systems*, pp. 1308–1311. ieeexplore.ieee.org, November 2008.

[19] Graham Cormode, Antonios Deligiannakis, Minos Garofalakis, and Andrew McGregor. Probabilistic histograms for probabilistic data. *Proceedings VLDB Endowment*, Vol. 2, No. 1, pp. 526–537, August 2009.

[20] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. BlinkDB: queries with bounded errors and bounded response times on very large data. In *Proceedings of the 8th ACM European Conference on Computer Systems*, EuroSys '13, pp. 29–42, New York, NY, USA, April 2013. Association for Computing Machinery.

[21] Jinglin Peng, Dongxiang Zhang, Jiannan Wang, and Jian Pei. AQP++: Connecting approximate query processing with aggregate precomputation for interactive analytics. In *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data*, SIGMOD '18, pp. 1477–1492, New York, NY, USA, May 2018. Association for Computing Machinery.

データマネジメントとデータベース管理

増永 良文†

†お茶の水女子大学名誉教授 〒112-8610 東京都文京区大塚 2-1-1
E-mail: †masunaga.yoshifumi@ocha.ac.jp

あらまし データサイエンスの興隆とともにデータマネジメント（data management）への関心が高いが，そもそもデータマネジメントは従来培われてきたデータベース管理（database management）とは何が違うのか？あるいは両者はどのように関係しているのか？これらの疑問点に関して筆者は寡聞にして明快な見解を目にしたことがない．データサイエンティストがデータ分析のために使用するデータ分析基盤（data analytics framework）を構築する際，両者の違いや関係性を明確に認識しておかないと，その構築に支障をきたすであろう．このような認識から，本稿では DAMA International が示した DMBOK2（データマネジメントの知識体系第２版）を検証し，それを再構成することにより DMBOK キーボードを提案する．さらに，これに基づいて両者の関係性を論じ，「データマネジメントという概念はデータベース管理という概念とデータガバナンスいう概念との概念融合である」ことを明らかにする．その結果，データ分析基盤構築で DBMS 技術の果たすべき役割が明確化する．

キーワード データマネジメント，データベース管理，データ分析基盤，データサイエンス，データサイエンティスト，DMBOK2，DAMA ホイール，DMBOK ピラミッド，DMBOK キーボード，概念融合，認知言語学

1. はじめに

データ分析（data analytics）が注目を集めているもっとも大きな要因は組織体がデータ中心の組織に変革しなければならないと努力しているからである．データ中心，あるいはデータ駆動とは，データを分析してその結果に基づき組織体の意思決定を下していくことをいう．

その根底にあるのが，データを**資産**（assets）として見なすという考え方である．データを**資源**（resources）ではなく，預貯金，有価証券，不動産，IT 資産などと同じように資産と見なすことの意味は，データに**価値**を認めているということである．そうなると，データは組織体のミッション（mission，使命）・ビジョン（vision，理念）・バリュー（value，行動指針），以下，**MVV** と略記，に照らし合わせて適切に統治されていかねばならない．それが**データガバナンス**（data governance）であり，データガバナンスのもとで，データをどのようにマネージメントするか，これが**データマネジメント**（data management）である．

データサイエンティストは組織体の意思決定に関与し，資産としてのデータを分析する責務を負っているから，まず，データマネジメントについての知識をしっかりと持ち合わせている必要がある．データマネジメントとは何か，そして，とりわけ大事なことは，データマネジメントはこれまで我々が営々と培ってきた**データベース管理**（database management）と**何が違うのか**，あるいは**どう関係しているのか**，についてしっかりとした見識を持ち合わせていなければならないということである．

2. データ・データ資源・データ資産

2.1 データと情報

そもそも「**データ**は記号の集まりであって，それ以上でもそれ以下でもない」[1]．このようにデータの定義は至極シンプルであるにも関わらず，それが石油に代わる資源であり，組織体の掛け替えのない資産となる．これはどういうことかを順を踏んで論じておく．

我が国では巷間，「データ」と「情報」という用語は相当に混用されている．英語で書けば，データは data，情報は information であって，明確に異なる概念なのに，本来データというべきところを情報といったり，その逆であったり，はたまたデータベースを「情報の格納庫」と説明する国語辞書があったりと，この混用には根深さを感じる．

さて，データの定義は上で与えたので，次に「データと情報の関係」を論じる．**情報**はデータの受け手（receiver）あるいはユーザの存在を前提として成り立つ概念である．つまり，データはその受け手に情報を与えることもあり，そうでないこともある．どういうことか．

データはまずその受け手により「意味解釈」される．これはデータを記述するために用いられている語彙を収録している辞書，構文規則，意味規則，つまり**データモデル**（data model）を用いてなされよう．たとえば，リレーショナルデータモデルで，(花子, 20)という 2 項タップル（binary tuple）がデータだとする．このデー

タをどう解釈すればよいかはこのタップルを眺めているだけでは分からない．もし，このタップルがリレーション 年齢(名前，年齢) のタップルであるなら，「花子は二十歳である」と意味解釈されるし，もし，このタップルがリレーション 給与(名前，給与) のタップルであれば，「花子の給与は 20（万円）である」と意味解釈できる．これがデータの持つ**意味**（meaning）である．

次いで，データの持つ意味は受け手がその時点で有している**知識**（knowledge）と比較される．ここに，知識とは知っていることの総体をいう（広辞苑）．もし，それが受け手の知識の増加をひきおこせば，このときデータは受け手に**情報**（information）をもたらしたことになる．たとえば，(花子，20)というデータの意味が「花子は二十歳である」としたとき，このデータの受け手がすでにそのことを知っていれば，このデータは受け手の知識の増分を引き起こさないから情報とはならないが，それを知らなければ情報となる．

さらに，受け手は人であれ組織体であれ，情報に対する価値観を有しよう．その結果，情報に価値の付与が行われ，**価値付情報**（value-added information）になる．たとえば，受け手が「花子は二十歳である」ということを知らなかったとしても，(花子，20)というデータから得られる情報は，花子に関心のない人にとっては三文の値打もないだろう．しかし，花子の二十歳の誕生日に薔薇の花束を贈りたいと夢見ていた受け手にとってはかけがえのない情報となろう．このように，データは受け手の存在のもとに（とても価値のある）情報となり得たり，そうでなかったりする．これが価値付情報の概念である．

図 1 にデータ，意味，情報，価値付情報の関係性を示す．

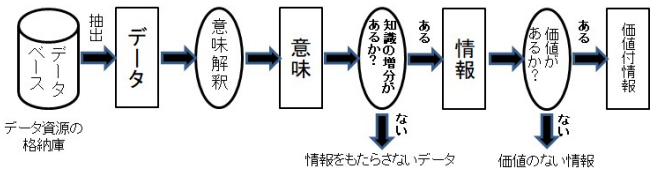


図 1　データ，意味，情報，価値付情報の関係性

2.2　データ資源とデータベース管理

「世界で最も価値のある資源はもはや石油ではなくデータである」というようなキャッチコピーをよく目にするようになって久しい．データをこれからの時代の最も重要な資源とみなすという考え方である．かつて，鉄，レアメタル，石油，天然ガスなどの天然資源がそうであったように．

さて，**データ資源**（data resources）はどこかを掘れば出てくるというようなものではない．確かに，データ自体はあらゆる時空に存在して多様な発生源からさまざまな形で泉のごとく湧き出ているのかもしれないが，明確な意図をもって汲み上げない限りは有用なデータ資源とはならないであろう．

そもそも，「資源としてのデータ」は石油などの天然資源と顕著に異なる性質を有しており，それらを列挙すれば次のようになろう[2].

(a) 人々の営為のあるところ，必ずやデータは発生する．もちろん自然の営みによってもデータは時々刻々発生している．
(b) データ資源は大量のそれも生データ（raw data，加工されていないデータ）であることが多い．
(c) データ資源は量的に単調に増加する．天然資源が単調に減少するのとは真逆である．
(d) データ資源のコンピュータによる利活用を考えるとき，それはデジタル化されていなければならない．
(e) データ資源の管理・運用にはコンピュータとネットワーク（＝通信回線）が必須である．
(f) データ資源はネットワークにより転送可能である．データ資源の転送にはあまりコストが掛からない．
(g) データ資源の保管には一般に大容量の電子記録媒体（＝データストレージ）を必要とする．
(h) データ資源は複製可能である．複製（＝コピー）の作成にほとんどお金がかからない．ただし，無断で複製されても（すなわち，盗まれても）痕跡が残らないことが多い．
(i) データ資源は暗号化できる．
(j) データ資源は多種多様である．テキストデータ，音声データ，音響データ，静止画像データ，動画像データ，時系列データ（センサーデータ），といった分類．あるいは，位置情報データ，生体情報データ，POS データ，といった分類．さらに構造化データ，半構造化データ，非構造化データ，といった分類．加えて，オープンデータといった分類もある．
(k) データ資源は多種多様である（続）．たとえば，テキストデータといっても，テキスト形式なのか，XML 形式なのか，CSV 形式なのか，PDF 形式なのかなどとさまざまであろう．また，画像データといっても，JPEG 形式なのか，PNG 形式なのか，BMP 形式なのか，TIFF 形式なのかといった具合にデータは文字通り多様である．さらに言えば，データがファイルとして管理されているのかデータベースとして管理されているのかで，その利活用の仕方はまるで異なってくる．

なお，(k)項末で指摘したデータ資源の管理について補足しておくと，組織体の共有資源としてのデータはファイルではなくデータベースとして管理されるべきである．なぜならば，ファイルはアプリケーションプログラムに隷属したデータなので，たとえば，異なるファイルに存在する同じデータが同一であることを保証するデータの整合性[1]（data integrity）の維持が困難である．一方，データベースはデータベース管理システム（database management system, DBMS）のもとで一元管理されたデータ群であって，トランザクションの同時実行や障害時回復などの機能に長けており，データベースが実世界を正確に写し込んでいるというデータベースの一貫性（database consistency）が保証され，組織体の共有資源としてのデータ資源の管理・運用を行うのに適している．

したがって，データベース管理（database management）を定義すると，次のようになろう．

【定義 1】（データベース管理）

資源としてのデータの管理をデータベース管理という．　□

2.3　データ資産とデータマネジメント

2.1 節で述べたように，データは意味解釈ルールのもとで意味を持つこととなり，それが利用者の知識の増分をもたらしたときに情報となり，それは利用者の価値観により価値付情報となった．

さて，利用者が**組織体**，つまり企業のような経済的組織体や行政組織のような非経済的組織体であったとしよう．組織体はMVVを有していると想定されるが，このときデータがもたらす情報がそれらに照らして戦略的意思決定や業務的意思決定（以下，単に**意思決定**という）に資すると判断されたとき価値を有することとなろう．そうすると，問題は組織体がこの価値をどのように見なすかである．組織体がこのような価値を生み出す基となるデータに対して資産的価値を認めたとき，データは**データ資産**（data assets）となる．

ただ，データを資産と見なすことにはデータが無形物であるが故に難しさがある．そもそも，組織体の資産とはそれが有する有形無形の財産を指すが，現金や物品・不動産といった有形物に限らず権利といった無形物まで含め，財産となるものすべてというのが一般的である．組織体が有しているデータが現時点でその組織体の意思決定に役立っていてその価値が認められていれば資産と見なし易いかもしれないが，そのような価値がその時点ではあらたかではない場合もあろう．では，そのようなデータはゴミとして捨ててよいのかといえば，そうではないかもしれない．なぜならば，そのようなデータが将来大きな価値を生みだすことがあるかもしれないからである．

データの資産的価値をどう評価するか，これは難問であろう．この問題は何も一組織体が抱え込むようなレベルの問題ではないように思われる．たとえばの話であるが，もし国や地方公共団体が，組織体が所有している不動産（土地・家屋・償却資産）に対して課している固定資産税とは別に，データ資産に対しても税金を課そうとした場合にその資産的価値をどう評価するのか？データ中心の組織体運営が常態化した暁にはこの問題は避けては通れないように思われるが，そのときにはその評価額算出の方程式をぜひ見てみたい．

さて，組織体がデータ中心の運営を行っていく場合，避けて通れないのが，データを組織体のかけがえのない資産と見なすということであった．一体，誰がどのような仕組みで組織体の資産としてのデータの管理を行うのか？これを**データマネジメント**（data management）というが，この概念は明らかにデータを資源と見なしてそれを管理するデータベース管理とは異なり，組織体の意思決定プロセスに深く関わるので極めて人的な匂いの濃い管理の概念であり，次のように定義するのが的確であろう．

【定義 2】（データマネジメント）

資産としてのデータの管理をデータマネジメントという．　□

2.4　実世界のデータモデリングの観点から見たデータベース管理とデータマネジメントの違い

データベース管理とデータマネジメントの定義はそれぞれ定義 1 と定義 2 で与えたが，両方の概念の違いを**実世界のデータモデリング**（data modeling）の観点から示してみよう．

図 2 に実世界のデータモデリング[1]の概念を示す．この観点から見たデータベース管理とデータマネジメントとの違いは以下のようである．

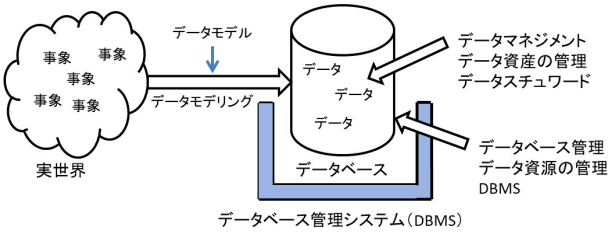


図 2 実世界のデータモデリングの観点から見たデータマネジマントとデータベース管理の違い

[1] たとえば，花子の年齢があるファイルでは二十歳と記録されているが，別のファイルでは十九歳となっているというようなことがないということ．

まず，データモデリングについて補足すると，実世界で生起している様々な事象（event, 出来事）を記述するためには，何らかの記号系（symbol system）が必要で，これを**データモデル**（data model）という．たとえば，実世界がビジネスワールドであればリレーショナルデータモデルが広く受け入れられている．データモデルに基づいて実世界をデータベース化する過程を**データモデリング**（data modeling）という．データモデリングの結果，データベースが構築される．図 2 はそのようにして構築されたデータベースをデータベース管理システムが(DBMS)が管理している様子を表している．DBMS はメタデータ管理，質問処理，トランザクション管理を行ってユーザがデータベースを安心して利用できるようにする．定義 1 でデータベース管理の定義を与えたが,それは上記の事柄を表している．

では，データベースのコンテンツとしてのデータそのものは誰がどのように管理してくれるのであろうか？換言すれば，データが組織体のかけがえのない資産としての価値を有するためには，データベースに格納されているデータ自体に資産としての価値のあることが前提となるが，それはどのような仕組みで誰が保証してくれるのであろうか？それがデータマネジメントであって，それが定義 2 として規定された．データマネジメントは組織体の資産としてのデータを管理するわけだから，それは MVV に基づく多様なアクションを伴い DBMS が行えるわけではない．データマネジメントの根本はデータガバナンスであり，それを実施する主体は**データスチュワード**（data steward）と称される組織体の人々やグループである．データガバナンスという大局観のもとでデータスチュワードが執り行うべき職務を挙げれば次のようである．

- 主要なメタデータの作成と管理：ビジネス用語，有効なデータ値，およびその他の重要なメタデータの定義と管理を行う．
- ルールと標準の文書化: ビジネスルール，データ標準，データ品質ルールの定義と文書化を行う．
- データ品質の管理
- データガバナンスの実行：データガバナンスのポリシーと戦略が遵守されているか，責任を持って確認する．

3. データマネジメント

3.1 データマネジメントとは

データサイエンスの興隆とともに脚光を浴びることとなったデータマネジメントであるが，近年それを新たな学問分野と捉えて,「データマネジメントの知識体系[2]」（data management body of knowledge, DMBOK）を明らかにしてみようという取組が世界のいくつかの組織でなされてきた．その中で，DAMA International が刊行した DAMA-DMBOK Second Edition（以下，**DMBOK2**）[3]は邦訳もされデータマネジメントを語る上で参照されることが多いのでその骨子を見てみる．

DMBOK2 は「データマネジメントとは，データおよび情報資産の価値をライフサイクル全体にわたって提供，制御，保護，および強化するための計画，ポリシー，プログラム，および実践の開発，実行，および監督である」と定義し，データマネジメントの知識体系を**図 3** に示すように定義している．

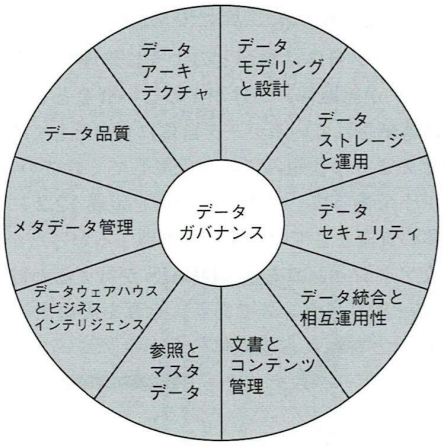


図3 DMBOK2 [3]

図 3 を補足すると, DMBOK2 はデータガバナンスという知識領域を中核とし，それを 10 個の知識領域が取り囲んでいる．その様子から **DAMA ホイール**（The DAMA Wheel）とも呼ばれる（ホイールは車輪の意）．DMBOK2 を構成する 11 個の知識領域の概要は次の通りである[3]（順不同）．

1. **データガバナンス**：組織体のデータ資産の管理に対する権限，制御，共有の意思決定法 (計画立案，監視，施行) の行使を規定する知識領域．具体的には，組織体がデータを資産として管理できるようにすること，データマネジメントの原則，ポリシー，手順，指標，ツール，および責任を定義，承認，伝達，実装すること，ポリシーの遵守，データの使用，管理活動を監視および指導することなどを含む．そのために，チーフデータオフィサー（CDO）を置き，組織体の各部門の責任者からなるデータガバナンス委員会を設け活動する．

(2) 知識体系（body of knowledge, BOK）はそれを特徴づける幾つかの知識領域（knowledge area）からなる．コンピュータサイエンスを規定した CSBOK，ソフトウェア工学を規定した SWEBOK，プロジェクト管理を規定した PMBOK など多数の BOK が公表されている．

2. **データアーキテクチャ**：組織体のデータ要求を部門横断的（cross-functional）に特定し，それらの要求を満たすマスタブループリントを設計しメンテナンスする．マスタブループリントを使用して，データ統合と相互運用性に指針を与え，データ資産を管理し，データ投資をビジネス戦略に合わせて調整することなどを規定する知識領域．
3. **データモデリングと設計**：データ要件を発見，分析，有効範囲を決定し，データモデルと呼ばれる正確な形式でこれらのデータ要件を表現して伝達するプロセスである．このプロセスは反復的であり，概念的，論理的，物理的なモデルを含んでよい．データモデリングと設計はこれを規定する知識領域．
4. **データストレージと運用**：保存されたデータの価値を最大化するための設計，実装，およびサポートなどを規定する知識領域．
   ANSI/X3/SPARCのDBMSの標準アーキテクチャ[1]の用語に従えば，データベース管理者（database administrator, DBA）が執り行う．
5. **データセキュリティ**：データ資産の適切な認証，認可，アクセス，監査を提供するためのセキュリティポリシーと手順の定義，計画立案，開発，実行などを規定する知識領域．
6. **データ統合と相互運用性**：アプリケーションと組織体内およびアプリケーションと組織体間でのデータの移動と統合の管理などを規定する知識領域．
7. **文書とコンテンツ管理**：あらゆる形式またはメディアのデータと情報のライフサイクル管理のための計画立案，実装，および制御活動などを規定する知識領域．
8. **参照とマスタデータ**：参照データとマスタデータという組織体の共有データを管理して，組織体の目標を達成し，データの冗長性に関連するリスクを軽減し，高品質を保証し，データ統合のコストを削減することなどを規定する知識領域．ここに，参照データとは他のデータを分類するために使用されるデータで，通常，それらは静的であるか時間の経過とともにゆっくりと変化する．参照データの例としては，測定単位，国番号，企業コード，固定換算率（重量，温度，長さなど）などがあげられる．一方，マスタデータとはビジネスエンティティに関するデータを言い，組織体のヒト・モノ・カネに関するデータ，たとえば，顧客データや商品データなどをいい，たとえば，受注といったトランザクションのデータ内にのみ含まれる．
9. **データウェアハウスとビジネスインテリジェンス**：意思決定支援のためのデータを提供し，レポート作成，問合せ，分析に従事するナレッジワーカを支援するための計画立案，実装，および制御プロセスなどを規定する知識領域．
10. **メタデータ管理**：高品質で統合されたメタデータへのアクセスを可能にするための計画立案，実装，および制御活動などを規定する知識領域
11. **データ品質**：データが利用に適しており，データ利用者の要求を満たしていることを保証するために，データに品質管理の手法を適用する活動の計画立案，実装，および制御などを規定する知識領域．

3.2 DMBOK ピラミッド

DMBOK2 はデータガバナンスを中核にして 10 個の知識領域がそれを取り巻いているだけなので，計 11 個の知識領域間の関係性が見えない．このことは，データ資産を最大限に利用して意思決定に役立てたいとの思いから**データ分析基盤**（data analytics framework）を構築したいと考えている組織体にとって，何から手を付けるとそれを構築できるのか，これだけではその道筋が見えないという難点と認識される．

これに対して，元 DAMA 会長の P. Aiken は DMBOK2 で規定された 11 の知識領域を基にして，それをピラミッド構造で再構成することで，組織体が段階的に究極の目的である高度なデータ分析が可能なデータマネジメント達成の道筋を示した[3]．これはその形状から**DMBOK ピラミッド**と称されるが，それを**図 4** に示す．

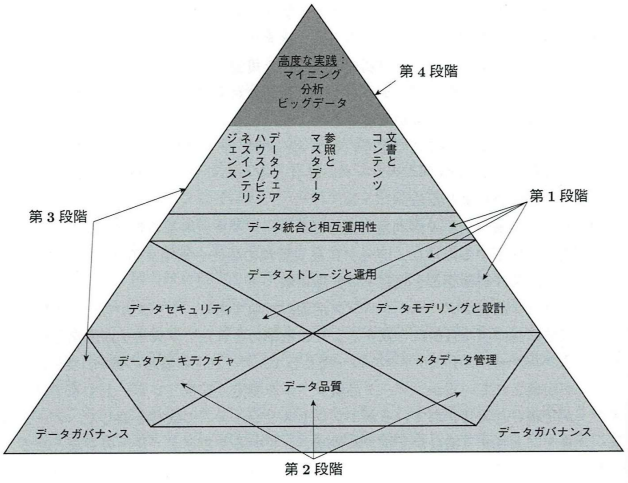


図 4　DMBOK ピラミッド[3]

DMBOK ピラミッドの意味しているところは，多くの組織体は次に示す「Aiken のフレームワーク」の 4 つ

の段階（phase）を踏むことで高度なデータ分析を行えるデータ分析基盤を開発できるということであろう．

【Aiken のフレームワーク】

第 1 段階　組織体はデータモデリングと設計，データストレージと操作，データセキュリティに関するノウハウを前提にデータベース機能を有するアプリケーションを購入する．システムを機能させるためにデータ統合と相互運用性にも取り組む．

第 2 段階　アプリケーションの使用を開始すると，データ品質に関する課題が見つかるが，より高品質のデータを取得するには信頼できるメタデータと一貫したデータアーキテクチャが必要となることを知る．これらにより，多様なシステムのデータがどのように連携するかが明確となる．

第 3 段階　データ品質，メタデータ，アーキテクチャの管理をきちんと実践するためには，データマネジメント活動に構造的なサポートを提供するデータガバナンスが必要であることを知る．データガバナンスにより，文書とコンテンツ管理，参照とマスタデータ管理，データウェアハウスとビジネスインテリジェンスといった戦略的イニシアティブの実行も可能となり，黄金のピラミッド内で高度なデータ分析が十二分に行えるようになる．

第 4 段階　組織体は適切に管理されたデータの利点を活用してデータ分析能力を向上させる．

なお，文献[4]は DMBOK2 を参照しつつ DX（digital transformation）成功のために組織体が何を行えばよいのかを具体的に記しているので実務者の参考になろう．

3.3　DMBOK キーボード

図 3 に示された DMBOK2，いわゆる **DAMA ホイール**は，中核にデータガバナンスという知識領域がありそれを 10 個の知識領域が取り巻いているが，その 10 個の知識領域はスポーク状に配置されているので，そのどれもがお互いにデータマネジメントの知識体系の構成要素としては対等である．一方，図 4 に示された **DMBOK ピラミッド**は，データベース管理に本質的な技術をベースにしつつ，それを基にしてデータ分析基盤実現への道筋を DMBOK2 で与えられた 11 個の知識領域を用いて段階的に表現している．この 4 段階にわたる道筋は，データマネジメントを実現しようと取組んでいる担当者にとって指針となるであろうことは間違いない．

しかしながら，この DMBOK ピラミッドでも，筆者の抱いた「データマネジメントとデータベース管理はどう関係しているのか？」という問への解答とは言い難い．なぜならば，両者の関係性を明らかにすることが両者の目的ではないからである．

では，この問題に対して単刀直入な解答は示せないのか？これが示されれば，データ中心を掲げる組織体の DX 推進の現場に対して，データマネジメントを実現するためにはこれまで先達が営々と築き上げてきた DBMS 技術に加えて何にどのようなことに注力すればよいのかが明確になり大いに役立つのではないかと考えられる．本節と次節ではそのような観点から DMBOK2 を再構成し，それに基づき両者の関係性を明らかにしてみる．

まず，データガバナンスという知識領域はその定義から明らかにデータマネジメントに固有な知識領域である．次に，データアーキテクチャであるが，3.1 節で説明した通りデータ資産の管理・運用を規定する知識領域である．したがって，これはデータガバナンスの延長上にあり，データマネジメントに固有色の強い領域と考えてよいであろう．加えて，データ品質であるが，これはデータ資源というよりはデータ資産としての価値に直結している．したがって，それをいかに保証するかは情報技術（IT）を越えている．従来の DBMS 技術はデータベースの一貫性，つまりデータベースと実世界の事象との間に齟齬がないことを保証することが目的であって，データ品質を保証することが主眼ではなかったから，データ品質はデータガバナンスの延長線上にあり，データマネジメントに固有色の強い知識領域と考えてよいであろう．言うまでもなく，品質の悪いデータ，たとえば，正しくないデータ，古いデータ，整合性のないデータ，あるいは信頼性のないデータなどを使ったデータ分析の結果に信用性はない．**データアーキテクチャとデータ品質はともにデータスチュワードの活躍するべき領域と考えられる**．

一方で，上記のデータガバナンス，データアーキテクチャ，データ品質以外の 8 個は元々はデータベース管理に固有な知識領域と考えられる．まず，メタデータ管理であるが，これは問合せ処理やトランザクション管理と並んで DBMS が持ち合わせないといけない 3 大機能の一つとして古くから認識され実装されてきた機能である．確かに，データマネジメントにとってメタデータ管理は部門横断的で多様なデータを統合管理する上でも，そしてデータが組織体の意思決定に役立つような形で構成されていることをデータガバナンスを介して保証するためにも，他の知識領域に比べて格段にデータマネジメントと組する大事な知識領域であることには間違いないが，データマネジメントに固有とは言い難い．続いて，データセキュリティであるが，DMBOK2 の説明によればデータマネジメントに固有色の強い知識領域と読めるが，そもそもデータセキュリティも暗号化など DBMS の管理・運用技術として鋭意育まれてきた経緯があり，それは DBMS 技術の延長

線上で実現されるべき事項として捉えられるから，これもデータベース管理に固有の知識領域とした．残る 6 個の知識領域：データモデリングと設計，データストレージと運用，データ統合と相互運用性，文書とコンテンツ管理，参照とマスタデータ管理，データウェアハウスとビジネスインテリジェンスについては，いずれもデータベース管理の知識体系[3]に固有の知識領域と考えてよいであろう．なぜならば，それらはこれまでの DBMS 構築（分散型 DBMS を含む）の過程で鋭意育まれてきた知識領域であるからである．

以上の議論から，データガバナンス以外の 10 個の知識領域が，データマネジメントに固有色が強いか，あるいはデータベース管理に固有かという旗色が鮮明となったが，もう一つ大事な点は，データマネジメントにおいては 10 個の知識領域はその旗色を問わず何れもがデータガバナンスによる統治の対象とならねばならないということである．これは，データ分析ではシステム的な要件に加えて，組織体の MVV に合致した資産としてのデータ管理が徹底されないといけないから当然のことである．

上記の議論に基づき DMBOK を再構成した結果を**図 5** に示す．その構造から，これを **DMBOK キーボード**（DMBOK keyboard）と呼ぶ（キーボードは鍵盤の意）．

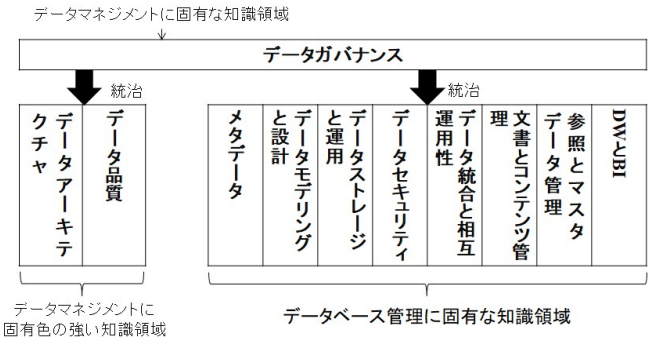


図5　DMBOKキーボード

DMBOK キーボードの意味するところを，データマネジメントの実現，すなわちデータ分析基盤の構築に向けた観点からまとめると次の通りである．

① 組織体内でデータガバナンス並びにデータスチュワードシップの体制を作り上げて，データアーキテクチャとデータ品質に求められるビジネス的並びにシステム的要件をことごとく洗い出す作業を行う．そこで得られた結果は組織体のデータガバナンスのあり方を示していると考えられるが，このとき何を統治するのかに特に留意しなければならない．この作業ではデータ資産の同定とその運用法が主題となるが，そのためにはビジネス側とシステム側のスタッフ両者の協同作業が必然となる．

② 情報システム部門はこれまで培ってきた DBMS 構築に関する経験とノウハウを基に，データ分析基盤構築のためのメタデータ管理，データモデリングと設計，データストレージと運用，データセキュリティ，データ統合と相互運用性，文書とコンテンツ管理，参照とマスタデータ管理，データウェアハウスとビジネスインテリジェンスのあり方を検討し，それらを反映したデータ分析基盤の構築を推進する．

③ データ分析基盤が組織体の意思決定支援に役立つためには，その開発に①と②の協調（＝擦り合わせ）作業が必須であることは言うまでもなく，中でもメタデータ管理とそれに関連してデータ統合と相互運用性をどのように実現していくかには特に注力していく必要があろう．加えて，ビジネス側とシステム側双方に，いわゆる PDCA サイクル[4]の実施を求めることになろう．その結果，データガバナンスとデータベース管理の融合が達成されるはずである．

4. データマネジメントとデータベース管理の関係性

さて，DMBOK キーボードを基にして，本稿冒頭に掲げた素朴な疑問「データマネジメントとデータベース管理はどう関係しているのか？」について定式化できないかと考えた．その解を**命題 1** に示す．ここに，＜データマネジメント＞，＜データベース管理＞，＜データガバナンス＞はそれぞれデータマネジメント，データベース管理，データガバナンスという「概念」を表す．⊗は概念融合を表す記号で，概念融合[5]（conceptual blending）とは 2 つのメンタル・スペース

[3] ただし，データベース管理知識体系（DBMBOK）はまだどの機関からも公表されてはいない．

[4] Plan（計画），Do（実行），Check（検証），Action（対策）からなる仮説と検証のサイクルを循環させて業務を継続的に改善していく手法のこと．

[5] 概念融合の身近な例としてよく引き合いに出されるのが「折る刃式カッターナイフ」OLFA の誕生エピソードである．戦後間もないころ，路上の靴職人たちは当時，靴底を削るのにガラスの破片を使い，切れ味が鈍るとまた割って使っていたという．その姿をみていた OLFA の創業者はふと，敗戦後，進駐軍の兵隊がかじっていた板チョコを思い出し，「そうだ，板チョコのように刃に折り筋を入れておき，切れなくなったら，ポキポキと折っていくと 1 枚の刃で何回も新しい刃が使えるぞ！」と折る刃式カッターナイフを発想したという．（出典：https://www.olfa.co.jp/birth_of_olfa_cutter.html）

（＝入力概念）に作用して第 3 のスペース（＝融合スペース．融合概念）が創発されるという認知言語学（cognitive linguistics）における一連の思考過程をいう[5]．融合スペースは，入力スペースからは与えられない創発構造（emergent structure）を持つ．これは互いに関連した次に示す 3 つのやり方（合成，完成，精緻化）で起こる．すなわち，合成（composition）：2 つの入力スペースからの投射が組み合わさることで，新たな関係が利用可能になる．完成（completion）：入力スペースから融合スペースに投射された合成構造が，融合スペースのより大きな自足的構造の一部と見なせるようになる．精緻化（elaboration）：融合スペース内部で，それ自身の創発論理に従って行われる認知作業で，精緻化できる．

【命題 1】(データマネジメントとデータベース管理の関係性)

データマネジメントとデータベース管理の関係性について次式が成立する．ここに，⊗は概念融合を表す．

＜データマネジメント＞＝
　＜データベース管理＞⊗＜データガバナンス＞

（証明の概略）＜データマネジメント＞は図 3 に示された計 11 個の知識領域からなる知識体系を表す概念である．一方，＜データベース管理＞は図 5 に示されたように 8 個のデータベース管理に固有な知識領域からなる知識体系を表す概念であり，＜データガバナンス＞はデータガバナンスという唯一の知識領域からなる知識体系を表す概念である．このとき，概念融合のもと，2 つの入力概念＜データベース管理＞と＜データガバナンス＞に作用して，第 3 の融合概念（＝融合スペース）である＜データマネジメント＞が創発される．その様子を**図 6** に示す．つまり，既存の 8 個の＜データベース管理＞に特有な知識領域は，それぞれが＜データガバナンス＞を構成する唯一の知識領域であるデータガバナンスと対応して融合概念である＜データマネジメント＞に投射されてそこでの知識領域として「合成」される．また，＜データガバナンス＞の知識領域であるデータガバナンスは＜データマネジメント＞に投影されて，融合概念である＜データマネジメント＞の知識領域として「完成」する．さらに，概念融合以前は潜在的であった「データアーキテクチャ」と「データ品質」という 2 つの知識領域が融合概念である＜データマネジメント＞で「精緻化」され創発される．その結果，計 10 個の知識領域をデータガバナンスという知識領域が統治するという新たな知識体系を表す概念である＜データマネジメント＞が融合概念として創発された．

□

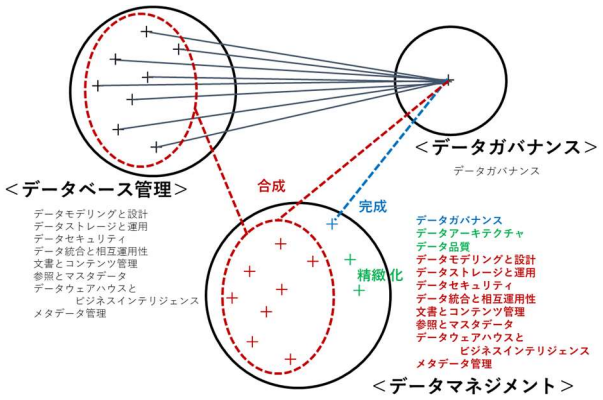


図 6　概念融合による
＜データマネジメント＞の創発

命題 1 をデータ資産管理とデータ資源管理の観点から表現し直すと**系 1** が成立しよう．この成立は上述の議論からほとんど明らかであろう．

【系 1】(データ資産管理とデータ資源管理の関係性)

＜データ資産管理＞＝
　＜データ資源管理＞⊗＜データガバナンス＞

5. おわりに

データマネジメントとデータベース管理の違いをデータ資産の管理とデータ資源の管理という側面から規定するとともに，両者の関係性を概念融合という認知言語学での手法を用いて明らかにした．これらの結果はデータ分析基盤の構築にあたり DBMS 構築技術がどのような役割を果たすべきかについて明快な指針を与えるものである．

なお，本稿で明らかにした事柄は筆者の近著[6, 7]に収録予定である．

参　考　文　献

[1] 増永良文．リレーショナルデータベース入門[第 3 版] ―データモデル・SQL・管理システム・NoSQL―．サイエンス社，2017.

[2] 増永良文．コンピュータサイエンス入門[第 2 版] ―コンピュータ・ウェブ・社会―．サイエンス社，2023.

[3] DAMA International. *DAMA-DMBOK Data Management Body of Knowledge*, Second Edition, Technics Publications, 2017.

[4] 小川康二，伊藤洋一．DX を成功に導くデータマネジメント．翔泳社，2021.

[5] Gilles Fauconnier (著), 坂原 茂，三藤 博，田窪行則 (訳). 思考と言語におけるマッピング―メンタル・スペース理論の意味構築モデル．岩波書店，2000.

[6] 増永良文．リレーショナルデータベース特別講義 ―データモデル・SQL・管理システム・データ分析基盤―．サイエンス社，2024.（出版予定）

[7] 増永良文．データサイエンティストのためのデータベース基礎．サイエンス社，2024.（出版予定）

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

🏛2024年2月28日(水) 10:00 ~ 12:10 | 🏛T2-B オンライン（Zoom Events)

準同型暗号

座長:上土井 陽子(広島市立大学)
コメンテータ:天笠 俊之(筑波大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T2-B-1-01] Intel SGX上で動作する準同型暗号の秘密鍵に対するサイドチャネル攻撃

*横平 正熙[1]、大西 隆太郎[2]、鈴木 拓也[2]、山名 早人[3] (1. 早稲田大学　基幹理工学部　情報理工学科　山名研究室、2. 早稲田大学大学院　基幹理工学研究科　山名研究室、3. 早稲田大学　理工学術院)

10:30 ~ 10:55
[T2-B-1-02] 効率的な暗号処理に向けた完全準同型暗号方式・ライブラリの評価

*辻 有紗[1]、小口 正人[1] (1. お茶の水女子大学)

10:55 ~ 11:20
[T2-B-1-03] 辞書を用いた圧縮センシングによる準同型暗号化データの通信量削減

*泉 湖雪[1]、松本 茉倫[1]、小口 正人[1] (1. お茶の水女子大学)

11:20 ~ 11:45
[T2-B-1-04] 準同型暗号と隔離実行環境の組合せによるプライバシを保護したBERT文章分類手法

*橋本 太志[1]、大西 隆太郎[2]、鈴木 拓也[2]、山名 早人[3] (1. 早稲田大学基幹理工学部、2. 早稲田大学大学院基幹理工学研究科、3. 早稲田大学理工学術院)

11:45 ~ 12:10
[T2-B-1-05] 【技術報告】不動産情報サービスの研究開発における共有データ資源

清田陽司 (株式会社LIFULL)

Intel SGX 上で動作する準同型暗号の秘密鍵に対するサイドチャネル攻撃

横平 正熙[†,*]　大西 隆太郎[‡]　鈴木 拓也[‡]　山名 早人[§]

† 早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡ 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§ 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: *mark-chan1@yama.info.waseda.ac.jp

あらまし クラウドコンピューティング上での機密情報漏洩への対策の一つである準同型暗号は，耐量子暗号の一つであり，データを暗号化したまま安全に計算できる．しかし，準同型暗号上での計算は，実行時間が長いため，実行時間短縮を目的に，隔離実行環境での実行と組み合わせる手法が提案されている．Intel SGX に代表される隔離実行環境では，データを隔離及び保護できるため，一旦データを復号し平文上で計算を実行できる．しかし，サイドチャネル攻撃によって，隔離実行環境内の平文が漏洩する恐れがある．本稿では，Intel SGX を対象に，隔離実行環境（エンクレーブ）内にある準同型暗号の秘密鍵をサイドチャネル攻撃によって漏洩させることを試みて，どの程度のサイズの秘密鍵がどの程度の時間で漏洩するのかを検証した．検証の結果，6.3 分で 64KiB の秘密鍵が窃取できることを確率的に算出した．

キーワード Intel SGX，準同型暗号，サイドチャネル攻撃，クラウドコンピューティング

1. はじめに

近年，クラウドコンピューティングへの関心の高まりから，クラウドサーバ上で機械学習や深層学習を行うケースが増えている．クラウドコンピューティングでは，利用者は機密データをクラウドサーバへ送信する必要がある．そして，クラウドサーバが外部からの攻撃を受けた場合や，クラウドサーバ管理者に悪意がある場合は，情報漏洩の危険性がある．

情報漏洩への対策として，準同型暗号[1]と隔離実行環境 [2]が挙げられる．準同型暗号は，データを暗号化したまま計算を実行できる一方で，評価できる演算に制限がある問題や準同型演算のレイテンシが長い問題，暗号文の改竄有無を判断できない問題がある．隔離実行環境（Trusted Execution Environment : TEE）はアクセス制御や暗号化によってデータを隔離し，安全な環境で計算を行う技術である．隔離実行環境の一つである Intel SGX [3]（以下，SGX）はメインメモリ上にエンクレーブと呼ばれる，ルート権限でもアクセス不可な暗号化された領域を確保し，データやコードの完全性と機密性を保護することができる．しかし，SGX はサイドチャネル攻撃に対する脆弱性が指摘されている[4].

以上に示した準同型暗号と SGX の欠点を補うために，両者を組み合わせて活用する手法が研究されている[5, 6, 7]．これらの研究では，(1) 準同型暗号の欠点である暗号文の改竄と SGX の欠点であるサイドチャネル攻撃を防ぐために，準同型暗号の演算をエンクレーブ内で行う手法と，(2) 条件分岐等の準同型暗号では評価できない処理や高負荷な処理を行うために，エンクレーブ内で暗号文を復号して平文上で処理を実行する手法がある．しかし，(2)では，エンクレーブ内で暗号文を復号するために，準同型暗号の秘密鍵をエンクレーブ内に配置する必要があり，サイドチャネル攻撃によって準同型暗号の秘密鍵が漏洩する危険性がある．そのため，準同型暗号の秘密鍵をエンクレーブ内に配置しても安全かどうかを検証する必要がある．

本研究では，SGX に対しサイドチャネル攻撃の一つである SgxSpectre 攻撃[10]を行い，エンクレーブ内にある準同型暗号の秘密鍵窃取を試みることで，SGX 上で準同型暗号の復号を行う際の安全性を検証する．SGX に対するサイドチャネル攻撃は，様々な種類が提案されており，データを読み取ることができる攻撃は Foreshadow[8], SGX-Step[9], SgxSpectre[10]が挙げられる．Foreshadow は，キャッシュ上のデータはエンクレーブ外のプログラムからも読めるという脆弱性を使い，L1 キャッシュを不正に読み取ることができる不具合を利用した攻撃である．しかし，Foreshadow は既にハードウェアベンダによって上記の不具合が改善されており，現実的な攻撃ではない．SGX-Step は入力値によってプログラム中で実行される命令数が変化するという挙動を利用して機密データを推測する攻撃手法である．しかし，準同型暗号の演算や復号は，処理内容が入力値に依存して変化することなく，攻撃者は秘密鍵の情報を取得できない．したがって，エンクレーブ内の準同型暗号の秘密鍵に対しての SGX-Step を用いた攻撃は不可能である．一方で，SgxSpectre[10]は CPU の投機的実行の機能により機密データをキャッシュに載せ，キャッシュヒット／ミスから機密データを推測

する攻撃であり，SGX を搭載するコンピュータ全てに有効である．また，SgxSpectre が可能な要因は，エンクレーブ内外において分岐予測の結果が共有される点にある．対策として，同一スレッド・同一プロセス以外の分岐予測結果は，エンクレーブ内外で共有されないという対策が施されている．また，分岐予測結果を削除するアセンブリ命令の追加により，分岐命令の実行前に分岐予測結果を削除する対策が可能となる．一方，同一スレッド・同一プロセス内の分岐予測結果はエンクレーブ内外で共有されるため，分岐予測結果を削除しない場合，SgxSpectre は有効である．そのため，本論文では SgxSpectre[10]を使った攻撃を対象とする．

本稿の貢献は，SGX に対するサイドチャネル攻撃において，準同型暗号の秘密鍵が攻撃者に完全に漏洩するまでの時間を算出し，安全な鍵再生成のタイミングについて考察した点である．ただし，考察はエンクレーブ内のプログラムがエンクレーブ内に入る前，事前に書き換えられているという条件のもとにおいて行う．

本稿は次の構成をとる．2 節で SGX と準同型暗号に関する背景知識を解説し，3 節で関連研究を紹介する．4 節で検証目的と方法を述べ，5 節で実験方法と実験結果を示す．6 節で考察を行い，7 節で結論を述べる．

2. 背景知識

2.1 Intel SGX

SGX はメインメモリ上にエンクレーブと呼ばれる AES 暗号化された領域を確保し，情報を秘匿する．さらに，アテステーションと呼ばれる認証技術によりエンクレーブ内のコード及びデータの改竄を防ぎ，完全性を保証する．また，エンクレーブはルート権限でもアクセスは不可能なため，悪意のあるサーバ管理者やルート権限を攻撃者に奪取されたコンピュータからもエンクレーブ内の情報を保護することができる．しかし，SGX はサイドチャネル攻撃への脆弱性が多く指摘されている[4, 9, 10]．そのため，SGX に対するサイドチャネル攻撃対策の研究が進められている[11]．サイドチャネル攻撃対策を行い，さらにオーバーヘッドを改善している研究も存在する[12]．

2.2 準同型暗号

準同型暗号は，暗号文を入力とする加算や乗算を復号せずに評価できる．本研究が対象とする準同型暗号の方式の一つである Cheon-Kim-Kim-Song（CKKS）方式[13]は，固定小数点数・複素数計算をサポートしており，機械学習や統計計算に適している．

CKKS 方式での演算は多項式環上で行われる．この際，計算量を削減するために，数論変換（Number Theoretic Transform: NTT）が用いられる．NTT により，畳み込み演算を要素ごとの乗算に置き換えることができ，計算量を削減できる．また，CKKS 方式では各項の係数の法Qは 200〜900 ビットと大きく，多倍長整数として扱う必要がある．そのため，中国剰余定理によって$Q=\prod_{i=0}^{l} q_i$とする（q_iは互いに素）．$q_i < 2^{62}$とすることで，和や差を 62 ビットに収めることができる．すなわち，1 つの要素を 8[Byte]で表すことができる．

次に，CKKS 方式における鍵生成と暗号化及び復号のアルゴリズムについて述べる．

- 秘密鍵：$\{-1,0,1\}$の中からランダムにN個選び，それを係数とするN次多項式の要素を秘密鍵sとする．なお，秘密鍵は通常 NTT 形式で保持されるため，データサイズは$8\times N$[Byte]である．
- 公開鍵：N次多項式空間の中からランダムに 1 つaを選ぶ．分散σ^2に基づく正規分布に従う乱数（準同型暗号ライブラリの一つである Microsoft SEAL[14]ではσ^2のデフォルト値は 10 である）を係数とするN次多項式eを選び，$b := -as+e$とする．$pk=(b,a)$が公開鍵である．
- 暗号化：平文（N次多項式）をmとする．$\{-1,0,1\}$の中からランダムに選んで係数としたN次の多項式をv，分散σ^2に基づく正規分布に従う乱数を係数とするN次の多項式をe_0, e_1とする．暗号文$Enc(m)$は式(1)で表される．
$$Enc(m) := (vb+m+e_0, va+e_1) \tag{1}$$
- 復号：暗号文cを$c=(c_0,c_1)$とする．復号$Dec(c)$は秘密鍵sを用いた式(2)で表される．
$$Dec(c) := c_0+c_1 s = m+(ve+e_0+e_1 s) \tag{2}$$
なお，v, e, e_0, e_1, sの係数はmよりも十分小さいため，$Dec(c)\approx m$となる．

3. 関連研究

本節では，メモリアクセス時のキャッシュヒット／ミスによるメモリアクセス時間差を利用したサイドチャネル攻撃手法である Spectre[15]を紹介する．本稿では，Spectre で用いられる手法を SGX に適用した攻撃である SgxSpectre[10]を評価実験で用いる．なお，電圧操作による破壊攻撃，コードやデータを改変するインジェクション攻撃などの，データの読み取りを目的としないサイドチャネル攻撃は本稿の対象外である．

2018 年に Kocher らは Spectre と呼ばれる，CPU の投機的実行とキャッシュヒット／ミスにおけるアクセス時間差を悪用したサイドチャネル攻撃を提案した[15]．

現代のコンピュータは高速化のためにパイプライン方式を採用している．そして，実行性能を最大限発揮するには，パイプラインを埋められるように演算器に連続して命令を流し込む必要がある．一方，入力依存データの存在によって分岐条件の評価に時間がかか

り，次の命令を決定できない際には，パイプラインを一度停止する必要があり，遅延が発生する．この遅延を回避し高速化を図るため，現代の CPU には，分岐結果が出るより先に，過去の分岐結果に基づく分岐予測器の予測結果から次の命令を実行する投機的実行の仕組み（Pattern History Table (PHT) と Branch Target Buffers (BTB) により実現されている[16]）が採用されている．一方で，投機的実行を行う際に，分岐条件の評価が誤りであった場合には，CPU はロールバックを行い，レジスタ値を実行前に戻すが，キャッシュの情報はロールバックされないという性質がある．

Spectre はこの性質を悪用し，攻撃対象データをキャッシュに載せ，メモリ内の機密データを盗み見る．具体的には次のような流れでメモリの内容を盗み見る（図 1）．ただし，(1) 攻撃者は攻撃対象とするプログラム（以下，標的プログラム）の内容を把握しており，(2) 攻撃者は標的プログラムと同一のプロセス内で動作させることのできる攻撃用のコード（以下，攻撃コード）を埋め込んでいることを前提とする．これにより，攻撃プログラムは標的プログラムと同一のメモリ空間を共有できる．

1. 図 1①に示す標的プログラムに対し，攻撃者はエンクレーブ外から同一プロセス内にある任意のメモリ領域 a（以下，監視領域）へ繰り返しアクセスし，メモリアクセス時間の変化を監視する．
2. 監視領域 a に対して標的プログラムからのアクセスがあった直後に攻撃者が監視領域へアクセスした場合，アクセス時間が短くなる．これはキャッシュヒットが生じるためである．すなわち，キャッシュヒットの発生有無を監視することで，監視領域 a の x 番目にアクセスがあったことを攻撃者は知ることができ，攻撃対象データの内容（配列 a[b[i]]の添字 b[i]）が分かる．
3. 図 1 左の標的プログラムには a[b[i]]が含まれている．攻撃者が b[i]を操作できる場合，攻撃者は a[b[i]]へのアクセスを通じて同一プロセス内の任意のアドレス上のデータを読み取ることができる（配列の範囲外アクセスを利用）．上記手順 2 において，キャッシュヒットが判明した場合，x=b[i]となり，攻撃対象データの値である b[i]が明らかになる．なお，b[i]を攻撃者が操作できたとしても図 1 左の標的プログラム中の分岐条件が真になるかどうかは制御できない．そのため，別途用意した攻撃プログラムにより分岐予測器を学習させ，分岐条件が偽の場合であっても，投機的に a[b[i]]への配列アクセスは実行され，a[b[i]]はキャッシュに載る．
4. 手順 1 に戻り，繰り返しデータを特定する．

上記方法により，攻撃者が攻撃対象データの格納されているアドレスを把握しているという条件下では，メモリから任意の情報を漏洩させることができる．Spectre の原理を利用した攻撃は一般的に「キャッシュ攻撃」と呼ばれ，SGX に対しても有効である．Spectre の亜種として SpectrePrime[17]や，SGX を攻撃対象としているものとして SgxPectre[4]や SgxSpectre[10]が提案されている．[4]及び[10]はどちらもキャッシュ攻撃である．SgxSpectre では同一のプロセス内から分岐予測器のトレーニングを行って投機的実行によりキャッシュにエンクレーブ内の機密データを載せている．一方，SgxPectre では他プロセスから分岐予測器のトレーニングを行い，SGX SDK（SGX を動作させるためのランタイム）中に存在する脆弱性コードへジャンプし，レジスタを操作して機密情報をキャッシュに載せているという点で異なる．SgxPectre では 184 バイトを窃取するのに最速で 414 秒，最大で 3,677 秒，1 バイトの窃取に平均 6.6 秒かかることが示されている[4]．なお，他プロセスからの分岐予測器のトレーニングは行うことができないように対策されており，現在は他プロセスから SgxPectre 攻撃を行うことはできない．

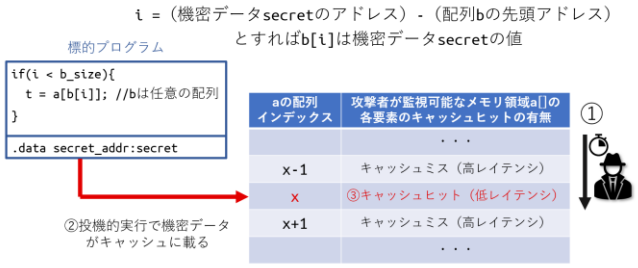


図 1. キャッシュヒット／ミスによるメモリアクセス時間差を利用した機密データの特定

4. 検証目的と方法

4.1 検証目的

本研究では，エンクレーブ内にある準同型暗号の秘密鍵をエンクレーブ外（同一プロセス内）から窃取することを試み，どの程度のデータサイズの秘密鍵がどの程度の時間で漏洩するか検証する．さらに，鍵寿命（攻撃者が秘密鍵を特定できるまでの時間）を定義し，得られた結果から算出した鍵寿命から，鍵再生成のタイミングについて考察することを目的とする．

4.2 前提条件

Kocher ら[15]や Bulck ら[10]のサイドチャネル攻撃での仮定と同様に，「攻撃者は，標的のエンクレーブ外の任意のプログラムに攻撃者のプログラムを埋め込め，標的と同一のプロセス内のコードを任意のタイミングで呼び出せる」と仮定する．

検証においては，以下を仮定する．

- 公開鍵は公開情報であるとし，攻撃者は正当に（攻撃を用いずに）公開鍵を入手可能である．
- 攻撃者は公開情報である準同型暗号パラメータから，秘密鍵のデータサイズを算出できる．
- 秘密鍵の正しさは暗号文に対して復号処理をした結果から判断できる．
- 攻撃者はSgxSpectre[10]によるサイドチャネル攻撃を行う．
- 標的プログラムがエンクレーブ内において実行される前に，攻撃者によって標的プログラムが改変され，秘密鍵が格納されているアドレスはエンクレーブ外から取得できるとする．
- 標的プログラム内にはエンクレーブ外から入力された値を引数として受け取り，エンクレーブ内からエンクレーブ外の配列へアクセスを行うコードを含むとする（例えば，暗号文をエンクレーブ内にコピーするために，エンクレーブ外の暗号文配列へのアクセスが必要である）．

4.3 鍵寿命

攻撃者が秘密鍵を特定するまでの時間を鍵寿命と定義する．ただし，鍵寿命は攻撃対象マシンのCPU性能，メモリアクセス時間，攻撃以外のプロセスの動作状況など，攻撃の実行環境に依存する．

鍵寿命の計算方法を説明する（図2）．攻撃者が長さMバイトの秘密鍵の窃取を試みたとき，窃取したデータの各バイトは確率pで正しく窃取できるとする．なお，攻撃者が1回で窃取するデータ量が1バイトであるため，バイト単位で考える．また，pは標的データのメモリアドレスによらず独立であるとする．このとき，pを精度と定義する．攻撃者は秘密鍵全体を窃取する試行をn回繰り返すとn個の窃取データが得られる．n個の窃取データのiバイト目をチャンクiと呼ぶとすると，図2中に示す1つのチャンク内において，正しい値がk個である確率は，式(3)で求められる．

$$P(X=k)=\frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k} \tag{3}$$

攻撃者は窃取データの各チャンク内で多数決を取り，秘密鍵を推測する．試行回数がn回の時，1つのチャンク内について，正しい値を得る確率は式(4)で表される．ただし，$x\in\mathbb{R}$ に対して$[x]$はxの整数部分を表す．

$$P\left(X\geq\left[\frac{n}{2}\right]+1\right)=\sum_{k=\left[\frac{n}{2}\right]+1}^{n}P(X=k) \tag{4}$$

Mバイトすべてについて多数決が成功すればよいので，n回の試行で多数決によって正しい秘密鍵を得る確率$P_1(n)$は，式(5)で表される．

$$P_1(n)=\left(P\left(X\geq\left[\frac{n}{2}\right]+1\right)\right)^M \tag{5}$$

したがって，n回の窃取試行のあと多数決によって得られた秘密鍵は平均して$M(1-P_1(n))$バイトが誤っていると考えられる．窃取データの誤りが1バイト未満，つまり，$P_1(n)>1-1/M$となるまでの最小の試行n回の窃取を行うためにかかる時間を鍵寿命と定義する．一度の試行にT秒かかるとき，$n=n_1$が$M\left(1-P_1(n)\right)<1$を満たすnの最小値であるとすると，鍵寿命はn_1Tとなる．秘密鍵は1ビットでも異なると暗号文を正しく復号できない上，秘密鍵のどの位置のビットが誤っているか攻撃者は分からない．そのため，攻撃者はデータ窃取を繰り返し行い，各バイトに対して多数決をとることで秘密鍵を推測することになるため，精度が低いほど鍵寿命が長くなると考えられる．

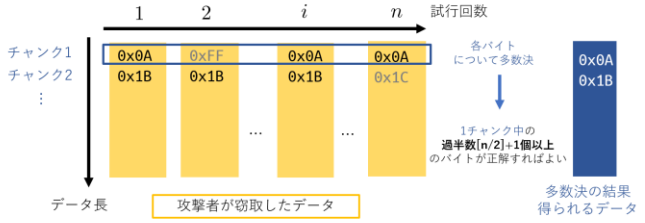


図2. 長さMバイトの秘密鍵をバイトごとの多数決で推測する方法

4.4 検証手順

4.4.1 検証の流れ

検証は以下の手順で行う．

1. 秘密鍵をエンクレーブ内に配置する．
2. 準同型暗号の暗号文を復号する処理を繰り返すプログラム（リスト1）を標的プログラムとしてエンクレーブ内で動作させる．このプログラムが標的プログラムとなる．
3. エンクレーブ内のデータにはアクセスできず，キャッシュヒット有無は検知できない．そのため，リスト1のプログラムを実行中にエンクレーブ外のプログラム（リスト2）からリスト1のプログラムへ不正な値を入力し，エンクレーブ内からエンクレーブ外部のアクセス可能な領域（監視領域）へのアクセスを繰り返させ，秘密鍵を漏洩させる（SgxSpectre[10]攻撃）．
4. メモリアクセス時間を監視することによって秘密鍵の窃取を試みる．
5. 平均の窃取時間Tを算出する．

動作の概要を図3に示す．検証で用いるリスト2の

```
1 ecall_decrypt(x){ //x:Index from outside of
  enclave
2   if(x < N) {
3     junk = outside_array[encl_arr[x]];
  //junk:any value
4     pt[x] = (ct[0][x] + ct[1][x]*sk[x])%Q;
5   } //pt:plaintext, sk:secretKey,
  ct:cipherText, Q:modulus
6 }
```

リスト1. 復号関数の疑似コード

コードは，Spectre[15]の Appendix C (Spectre Example Implementation) 68 行目の標的関数（投機的実行により，データがキャッシュに漏洩する脆弱性が存在する関数）を，準同型暗号の復号関数 ecall_decrypt()へ変更したものであり，リスト 1 のコードは 26 行目の後に準同型暗号の復号計算処理を追加したものである．以下でリスト1のコードとリスト2のコードについて解説する．

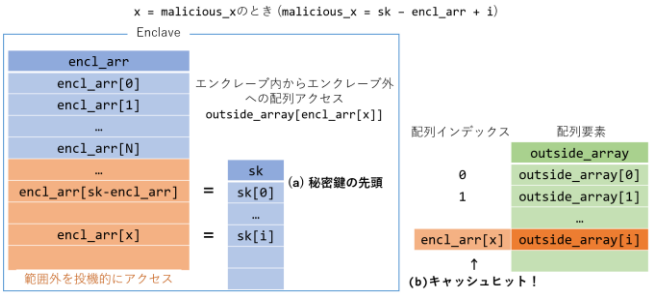


図 3. キャッシュヒットの有無で秘密鍵 sk の要素の値を特定する仕組み

4.4.2 復号関数（リスト 1）

リスト 1 に示す関数は，エンクレーブ内で準同型暗号の暗号文の復号を行う関数である．前提として，関数内に 3 行目の配列アクセスを含むとする（例えば，暗号文がエンクレーブ外に配置されている場合など，エンクレーブ外の配列へアクセスを行う状況は考えられる）．1 行目で引数として受け取る x は配列のインデックスを指定するための数値である．長さ N の配列としてエンクレーブ内に格納されている秘密鍵配列及び要素数$2\times N$の多次元配列として格納されている暗号文配列のインデックスを x により指定する．2 行目で x が多項式環の次数 N より大きい不正な値かどうか，範囲外アクセスのチェックを行う．一方，攻撃者は x<N を満たす x である training_x を復号関数に渡すことを繰り返し，分岐予測器を学習させる．この学習により，x<N を満たさない不正な x である malicious_x が復号関数に渡されても，CPU は投機的に 3 行目の配列アクセスを行う（encl_arr はエンクレーブ内の任意の配列である）．この操作により，攻撃者は malicious_x により指定した任意のアドレスへメモリアクセスできる．例えば，図 3 中の(a)に示すように，エンクレーブ内で参照されている秘密鍵が保存されているアドレス sk（攻撃対象アドレス）と encl_arr の先頭アドレスの差 sk-encl_arr を malicious_x とすれば，encl_arr [malicious_x]への配列アクセスで秘密鍵の先頭アドレスへアクセスできる．図 3 の(b)に示すように，攻撃者の監視領域であるエンクレーブ外の配列 outside_array について，outside_array[encl_arr[x]]の配列アクセスを行うと，outside_array 配列の先頭から encl_arr[x]番目がキャッシュに載る．3 行目～4 行目では配列のインデックスを指定して復号計算を行う．

```
1  readMemoryByte(malicious_x){
2    //Repeat line3~21 to improve precision
3    cflush(); //Flush cache
5    x //Index of  vector to decrypt
6    for(j=TRAIN_LOOP; j>=0; j--) {
7      if(j%ACCESS_FREQ==0) x = malicious_x;
8      else x = training_x;
9      ecall_decrypt(x); //Modified from Spectre
10   }
11   for(i=0; i<MEMACCESS_LOOP; i++) {
12     addr = ATTACKER_OBSERVING_ADDR + i;
13     time1 = getTime();
14     junk = *addr; // Memory access
15     time2 = getTime();
16     mem_access_time = time2 - time1;
17     if(mem_access_time <= CACHE_HIT_TIME)
18       result[i]++;
19   }
20   //Find index of most cache hits
21   largest_index = largestIndex(result);
22   return largest_index;
23 }
```

リスト 2. SgxSpectre[10]におけるデータを窃取する関数の疑似コード（Spectre[15]の Appendix C (Spectre Example Implementation) をもとに改変）

以上により，エンクレーブ内からエンクレーブ外へアクセスした後，エンクレーブ内の秘密鍵を用いて暗号文の復号を行う．

4.4.3 データ窃取関数（リスト 2）

リスト 2 に示す関数は，リスト 1 の復号関数を繰り返し呼び出すことにより，エンクレーブ内の機密データをキャッシュに載せ，監視領域へのメモリアクセスを繰り返して機密データを推定する関数である．1 行目の引数 malicious_x を受け取り，6～10 行目では，秘密鍵の長さをはみ出さないような 0 や 1 などの値をとる training_x を指定し，復号関数内の分岐結果を分岐予測器に学習させる．7 行目で ACCESS_FREQ 毎に 1 回，malicious_x を指定して復号関数を呼び出し，投機的実行によって攻撃者の監視領域へのアクセスを行う．次に，11～19 行目で攻撃者は監視領域へのアクセス時間を計測し，21 行目でキャッシュヒットが最も多く起こったアドレスを抽出する．キャッシュヒット数が最大のアドレスにはエンクレーブ内からのアクセスがあることが推測できる．

5. 実験

5.1 実験方法

準同型暗号ライブラリである Microsoft SEAL [18]を用いて鍵ペアと暗号文の生成を行い，測定においては同一の秘密鍵と暗号文を用いる．鍵ペアと暗号文の生成に使用した暗号化パラメータについて，多項式環の

次数Nは2^{13}，暗号文の法は 40 bit，スケールは2^{35}とした．なお，検証プログラム内では準同型乗算を評価しないため，暗号文のレベルlは 0 とした．また，プログラムの実行はすべてシングルスレッドで行う．

実験プログラムにおいて，プログラム内での攻撃開始から秘密鍵と同じデータサイズのデータを窃取するまでの時間（窃取時間）を計測する．そして，この実験プログラムを逐次的に 10 回実行し，(回数は任意に設定した)平均をとることで平均窃取時間を算出する．また，窃取データの精度も算出する．秘密鍵のデータサイズをMバイト，誤ったバイト数の合計をBバイトとすると，精度は$100 \times (M-B)/M$ [%]となる．この際，リスト 2 に示す関数 `readMemoryByte()`中の 6 行目の `TRAIN_LOOP` 及びリスト 2 の 2 行目～21 行目のループ回数を変化させる．例えば，`TRAIN_LOOP` 及び 2 行目～21 行目のループ回数を減少させると，`readMemoryByte()`内における合計ループ回数が減少し，単位時間当たりの `readMemoryByte()`の呼び出し回数が減少するため，平均窃取時間の減少が期待される．一方，`TRAIN_LOOP` の値を減少させると分岐予測器の学習が不十分になる可能性が高まり，2 行目～21 行目のループ回数を減少させると試行回数が減少するため，エンクレーブ内のデータを正確に窃取できない可能性が高まる．したがって，平均窃取時間と精度の関係を明らかにするため，リスト 2 の 6 行目の `TRAIN_LOOP` の値を 1, 3, 5, 10, 20 と変化させ，リスト 2 の 2 行目～21 行目のループ回数も 10 回，20 回，50 回，100 回，500 回と変化させて窃取時間の記録を行う．

5.2 実験環境

実験環境を以下に示す．使用する CPU は，Intel® Core™ i9-9900K であり，動作周波数は 3.0GHz で固定とした．また，物理コア数は 8 コアで，ハイパースレッディングは有効とする．CPU キャッシュは，L1 キャッシュはデータ用及び命令用がそれぞれ 32[KiB/core]，L2 キャッシュは 256[KiB/core]，L3 キャッシュは 16[MiB/CPU]である．なお，ラインサイズが 64[Byte]である．メインメモリは 32[GB]であり，エンクレーブ容量は 128[MB]である．

OS は Ubuntu 22.04 であり，カーネルバージョンは 6.2.0-39-generic である．使用する SGX ドライバのバージョンは 2.14 であり，SGX SDK のバージョンは 2.16.100.4 である．なお，コンパイラは gcc 11.4.0 を使用する．BIOS は，American Megatrends Inc.製であり，バージョンは E7C24IMO.106 である．なお，SGX 脆弱性に関する BIOS アップデートによるマイクロコードパッチは適用していない．

5.3 実験結果

表 1～3 に実験結果を示す．なお，表中のループ回数とはリスト 2 の 2～21 行目のループ回数であり，1 バイト窃取のための試行回数である．`TRAIN_LOOP` は分岐予測器の学習と監視領域へのアクセスを行うループの回数である．ループ回数と `TRAIN_LOOP` を変化させると，リスト 2 の関数 `readMemoryByte()`の単位時間あたりの呼び出し回数が変化し，平均窃取時間が変化することが推測される．窃取データの精度は，正しい秘密鍵データに対して誤ったバイト数の合計をBバイトとすると，$100 \times (M-B)/M$ [%]と計算する．また，4.3 鍵寿命項で定義した鍵寿命算出における窃取試行回数nを大きくすると計算量が大きくなり，鍵寿命の評価に時間がかかるため，nの上限を 300 とした．

また，多項式環の次数$N = 8192$であるから，秘密鍵のデータサイズ$M = N \times 8$[Byte] = 65536[Byte]である．平均窃取時間T, Mを用いて算出した鍵寿命を表 3 に示す．なお，表 3 中の「計算不可」は，4.3 鍵寿命項における窃取試行回数nが上限の300に達しても$P_1(n) > 1 - 1/M$を満たさなかったことを意味する．また，ループ回数と `TRAIN_LOOP` の組み合わせごとにおける，経過時間に対する平均窃取データ量を図 4～図 8 に示す．

表 1. 64KiB 秘密鍵の平均窃取時間[s]

ループ回数	平均窃取時間T[s]				
	TRAIN_LOOP =1	TRAIN_LOOP =3	TRAIN_LOOP =5	TRAIN_LOOP =10	TRAIN_LOOP =20
10	12.44	94.25	94.31	94.22	94.00
20	33.15	94.69	94.81	94.91	94.87
50	32.44	95.32	95.41	95.27	95.19
100	57.89	95.53	95.53	95.30	95.28

表 2. 1 窃取データの精度[%]

ループ回数	窃取データの精度p[%]				
	TRAIN_LOOP =1	TRAIN_LOOP =3	TRAIN_LOOP =5	TRAIN_LOOP =10	TRAIN_LOOP =20
10	84.69	94.25	94.31	94.22	94.00
20	67.95	94.69	94.81	94.91	94.87
50	87.78	95.32	95.41	95.27	95.19
100	76.40	95.53	95.53	95.30	95.28

表 3. 64KiB 秘密鍵の鍵寿命[s]

ループ回数	鍵寿命$n_1 T$[s]				
	TRAIN_LOOP =1	TRAIN_LOOP =3	TRAIN_LOOP =5	TRAIN_LOOP =10	TRAIN_LOOP =20
10	771.3	375.0	478.8	766.8	1,425
20	計算不可	705.4	894.1	1,291	3,060
50	1,525	745.4	944.8	1,813	3,671
100	計算不可	1,378	1,850	2,852	6,043

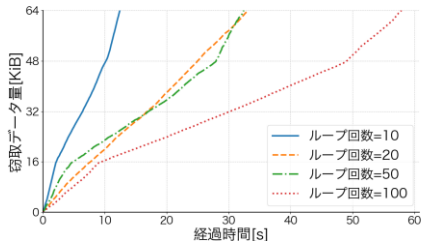


図 4. TRAIN_LOOP=1 のときの経過時間に対する窃取データ量

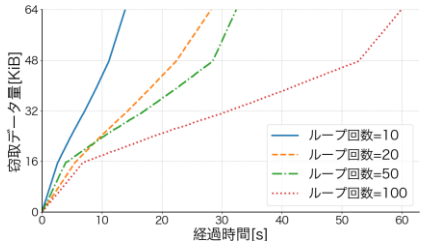


図 5. TRAIN_LOOP=3 のときの経過時間に対する窃取データ量

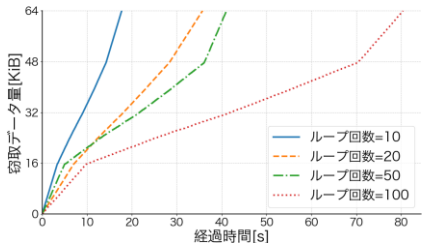


図 6. TRAIN_LOOP=5 のときの経過時間に対する窃取データ量

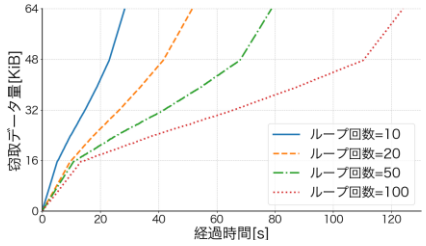


図 7. TRAIN_LOOP=10 のときの経過時間に対する窃取データ量

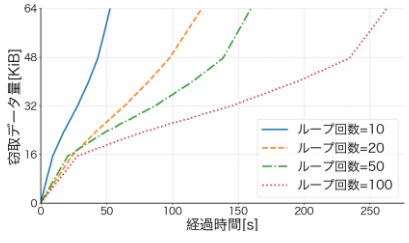


図 8. TRAIN_LOOP=20 のときの経過時間に対する窃取データ量

6. 考察

6.1 平均窃取時間と精度の関係

平均窃取時間（秘密鍵データを全て窃取するまでにかかった時間の平均）はリスト 2 に示す関数 `readMemoryByte()`の単位時間あたりの呼び出し回数に反比例するため，単位時間当たりの `readMemoryByte()` の呼び出し回数が平均窃取時間を左右する．例えば，リスト 2 の 6 行目の `TRAIN_LOOP` 及びリスト 2 の 2 行目～21 行目のループ回数を減少させると，`readMemoryByte()`内における合計ループ回数が減少し，単位時間当たりの `readMemoryByte()`の呼び出し回数が減少するため，平均窃取時間の減少が期待される．実験の結果，表 1 に示した通り，`TRAIN_LOOP` の値とループ回数が小さいほど，平均窃取時間は短くなった．

一方，`TRAIN_LOOP` の値を減少させると分岐予測器の学習が不十分になる可能性が高まり，2 行目～21 行目のループ回数を減少させると試行回数が減少する．したがって，エンクレーブ内のデータを正確に窃取できない可能性が高まる．つまり，窃取データが秘密鍵に対して正しい割合である精度（例えば，秘密鍵を M バイト，誤ったバイト数の合計を B バイトとすると，精度は $100 \times (M - B)/M[\%]$ である）が低下すると考えられる．計測の結果，表 2 に示した通り，`TRAIN_LOOP` の値が 1 以外ではループ回数と `TRAIN_LOOP` の値を増加させても精度は増加せず，精度の変化は 1%以内であった．したがって，`TRAIN_LOOP` の値が 1 のときには，分岐予測器の学習が不十分であったと考えられる．一方，`TRAIN_LOOP` の値を変化させても精度に影響はなく，分岐予測器の学習は 3 回で十分であったと考えられる．したがって，平均窃取時間と精度の間にはトレードオフの関係があると考えられる．

6.2 鍵寿命から考えられる安全な鍵再生成タイミング

`TRAIN_LOOP` の値が 3，ループ回数が 10 での鍵寿命 375 秒（6.3 分）が最も短い結果となった．このパラメータは，平均窃取時間が 2 番目に短く，精度が最大の精度と比べて 1.3%以内の差であったため，精度が高く，平均窃取時間が短いほど鍵寿命が短いことが分かる．また，攻撃者が最短の鍵寿命である 6.3 分で秘密鍵を特定すると考えると，エンクレーブ内に秘密鍵を配置する際には，鍵の再生成を 6.3 分毎に行うのが最も安全であると言える．しかし，クラウドサーバにおいて鍵の再生成を行う場合，クラウドサーバはユーザと鍵共有をする必要があり，通信のコストが生じる．

一方，SgxSpectre 攻撃を行うためには，秘密鍵が格納されたアドレスの特定が必要である上，リスト 1 のような SgxSpectre 脆弱性のあるプログラム（エンクレーブ内からエンクレーブ外への配列アクセスを伴うプログラムや，データをアドレスに変換するようなプ

ログラム）が存在する必要がある．例えば，エンクレーブ外に配置された暗号文配列にアクセスする場面など，エンクレーブ外のデータが必要な場合は，エンクレーブ内からエンクレーブ外の配列にアクセスする必要があると考えられる．エンクレーブ内の秘密鍵が配置されているアドレスをエンクレーブ外に漏洩させないことや，`outside_array[encl_arr[i]]`といった配列の添字に配列要素の値を使用する記述をしないことによって回避が可能となり，本研究で行った評価実験と同様な攻撃が可能な条件は限定的であると言える．

本稿では復号関数を対象としたが，復号以外の処理をエンクレーブ内で行う場合には，処理内容が増加し，それに伴って単位時間当たりの攻撃関数の呼び出し回数が減少するため，平均窃取時間は長くなると考えられる．また，キャッシュに載った攻撃対象データが他の処理によって退避させられる場合が考えられ，精度が低くなる可能性がある．

7. 結論

本稿では，SGX のエンクレーブ内に秘密鍵を配置することに対する安全性を検証するため，SGX に対するサイドチャネル攻撃である SgxSpectre を用いて秘密鍵の漏洩を試み，平均窃取時間を算出した．さらに，得られた鍵寿命をもとに，エンクレーブ内で準同型暗号の復号を伴う処理を行うことの安全性について検証した．その結果，標的プログラムにサイドチャネル攻撃に対する脆弱性があり，秘密鍵の格納アドレスが既知であるという条件下では，64KiB の秘密鍵の場合，鍵の更新を 6.3 分ごとに行わなければ，秘密鍵が漏洩する危険性があることが明らかになった．

今後の課題は，SGX と準同型暗号を組み合わせた実用的なアプリケーションに対してサイドチャネル攻撃を行い，当該アプリケーションが安全性を保証できる要件を明らかにすることである．

参 考 文 献

[1] C. Gentry, “Fully Homomorphic Encryption Using Ideal Lattices,” in Proc. of STOC 2009, pp.169–178, 2009.

[2] GlobalPlatform, “Introduction to Trusted Execution Environment,” https://globalplatform.org/wp-content/uploads/2018/05/Introduction-to-Trusted-Execution-Environment-15May2018.pdf, Last Access: Jan 5th, 2024.

[3] F. McKeen, I. Alexandrovich,A. Berenzon, C. V. Rozas, H. Shafi, V. Shanbhogue, and U. R. Savagaonkar, “Innovative instructions and software model for isolated execution,” Hardware and Architectural Support for Security and Privacy, vol.10, no.1, pp.1–8, 2013.

[4] G. Chen, S. Chen, Y. Xiao, Y. Zhang, Z. Lin, and T. H. Lai, “SgxPectre: Stealing Intel Secrets From SGX Enclaves via Speculative Execution,” IEEE Security & Privacy, vol.18, no.3, pp.28–37, 2020.

[5] C. Yakupoglu, K. Rohloff, “PREFHE, PREFHE-AES and PREFHE-SGX: Secure Multiparty Computation Protocols from Fully Homomorphic Encryption and Proxy ReEncryption with AES and Intel SGX,” Proc. of SecureComm 2022, pp.819–837, 2022.

[6] P. Liu, and W. Zhang, “Towards Practical Privacy-Preserving Solution for Outsourced Neural Network Inference,” Proc. of CLOUD 2022, pp.357–362, 2022.

[7] Fischer, B. Fuhry, J. Kußmaul, J. Janneck, F. Kerschbaum, and E. Bodden, “Computation on Encrypted Data Using Dataflow Authentication,” ACM Transactions on Privacy and Security, vol.25, no.21, pp 1–36, 2022

[8] J. V. Bulck et al., “Foreshadow: Extracting the Keys to the Intel SGX Kingdom with Transient Out-of-Order Execution,” Proc. of USENIX Security 18, pp.991–1008, 2018.

[9] J. V. Bulck, M. Minkin, O. Weisse, D. Genkin, B. Kasikci, F. Piessens, M. Silberstein, T. F. Wenisch, Y. Yarom, and R.l Strackx, "SGX-Step: A Practical Attack Framework for Precise Enclave Execution Control,” Proc. of SysTEX '17, pp.1–6, October 2017.

[10] D. O’Keeffe, D. Muthukumaran, P. Aublin, F. Kelbert, C. Priebe, J. Lind, H. Zhu, and P. Pietzuch , “Spectre attack against SGX enclave,” https://github.com/lsds/spectre-attack-sgx, Last modified: Jan 15th, 2018.

[11] M. Orenbach, A. Baumann, and M. Silberstein, “Autarky: Closing Controlled Channels with Self-Paging Enclaves,” Proc. of EuroSys '20, pp.1–16, 2020.

[12] P. Zhang, C. Song, H. Yin, D. Zou, E. Shi, and H. Jin , “Klotski: Efficient Obfuscated Execution against Controlled-Channel Attacks,” Proc. of ASPLOS 2020, pp.1263–1276, 2020.

[13] J. H. Cheon, A. Kim, M. Kim, and Y. Song, “Homomorphic encryption for arithmetic of approximate numbers,” Proc. of Asiacrypt 2017, pp.409-437, 2017.

[14] P. Borrello, A. Kogler, M. Schwarzl, M. Lipp, D. Gruss, and M. Schwarz , “ÆPIC Leak: Architecturally Leaking Uninitialized Data from the Microarchitecture,” Proc. of USENIX Security 22, pp.3917–3934, 2022.

[15] P. Kocher, J. Horn, A. Fogh, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz, and Y. Yarom , “Spectre Attacks: Exploiting Speculative Execution,” Proc. of S&P '19, pp.1–19, 2019.

[16] 斎藤史子，山名早人，“BTB のエントリ有無を参照した分岐予測器,” 情報処理学会論文誌：コンピューティングシステム，vol.45, no.SIG 11(ACS 7), pp.1–9, 2004.

[17] C. Trippel, D. Lustig, and M. Martonosi , “MeltdownPrime and SpectrePrime: Automatically-Synthesized Attacks Exploiting Invalidation-Based Coherence Protocols,” arXiv, arXiv:1802.03802, pp.1–11, 2018.

[18] Microsoft SEAL (release 4.1), “https://github.com/Microsoft/SEAL,” January 2023. Microsoft Research, Redmond, WA, Last Access: Jan 5th, 2024.

効率的な暗号処理に向けた 完全準同型暗号方式・ライブラリの評価

辻　有紗† 小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2–1–1
E-mail: †{arisa-t,oguchi}@ogl.is.ocha.ac.jp

あらまし　クラウドサーバ上のデータ保護のために暗号化が重要であり，暗号化された状態で任意の演算が可能な完全準同型暗号 (以下 FHE: Fully Homomorphic Encryption) はその目的に多いに期待されている．FHE の実用化に向けて，複数の暗号方式が提案されており，それぞれの暗号方式を実行可能なライブラリが複数存在する．本稿では，実行環境や処理内容に基づき適切な暗号方式・ライブラリを選択するための比較を行った．具体的には，初めに OpenFHE, Lattigo，TFHEpp ライブラリにそれぞれ実装された BFV， BGV，CKKS，TFHE 亜種方式の時間空間計算量や得意な演算を整理した．ここで，FHE を用いてデータの収集・分析を行う場合，搭載されている DRAM 容量や CPU に制約がある IoT クライアント端末で FHE 処理で用いる鍵の作成やデータの暗号化を行う必要がある．そこで二番目に，TFHE 亜種暗号方式が実装された OpenFHE と TFHEpp について，低性能な IoT クライアント端末における実行時間とボトルネックとなる処理の分析を行なった．

キーワード　完全準同型暗号，FHE 暗号方式，FHE 暗号ライブラリ

1 はじめに

近年，IoT 機器やセンサで収集したデータをクラウドに送信し，分析する機会が増加している．個人情報保護のため，クライアントデバイスでデータを暗号化しサーバに送信するが，従来の暗号技術では分析中にデータを復号する必要があり，クラウドサーバにおける個人情報の漏洩や不正利用のリスクが高い．そこで，暗号化されたデータの任意の演算が可能な完全準同型暗号 (以下 FHE: Fully Homomorphic Encryption) [1] が注目を集めている．FHE では，BFV 方式 [2]，BGV 方式 [3]，CKKS 方式 [4]，TFHE 亜種方式 [5] など複数の暗号方式が提案されており，それぞれ暗号処理の方法が異なる．また，これらの暗号方式を実行可能なライブラリが複数存在し，各ライブラリは細かい実装が異なる．実行環境や暗号化された状態で行う処理に応じて適切な暗号方式とライブラリを選択する必要があるが，暗号の研究者以外が FHE ライブラリ内で適切なパラメータを設定し，暗号方式やライブラリの比較を行うことは容易でない．

本稿では，複数の暗号方式とライブラリの計測を行い，FHE を使用してシステムを構築する際の参考情報を提供することを目的とする．具体的には，初めに OpenFHE [6]，Lattigo [7]，TFHEpp ライブラリ [8] でそれぞれ実装された BFV， BGV，CKKS，TFHE 亜種方式において，乗算実行時の時間空間計算量や効率的に実行可能な演算を整理した．実験から，整数の暗号処理を行う場合，Lattigo ライブラリの BGV 方式が最速であることが分かった．浮動小数点の暗号処理が可能な CKKS 方式では，連続した乗算回数，すなわち乗算深度が小さい場合は Lattigo が高速であり，多い場合は OpenFHE が高速であった．ここで，BFV，BGV，CKKS 方式と TFHE 亜種方式では基本的な暗号処理の単位や bootstrapping の仕組みが異なるため，FHE で実行される演算の種類により適切な暗号方式を選択することが重要である．CKKS 方式と TFHE 亜種方式の比較では，任意の複数の値の乗算を行う際は TFHE 亜種方式の時間空間計算量が小さいが，ベクトル内積では CKKS 方式が効率的であることを示した．

実用的なアプリケーションを考慮した際の全ての FHE 暗号方式に共通の課題は，時間空間計算量が大きいことである．特に，FHE 処理で用いる鍵の作成やデータの暗号化を行う必要がある IoT クライアント端末は，搭載されている DRAM 容量が少なく組み込み機器用の CPU を使用するため，FHE 暗号処理に適していない．そこで二番目に，IoT クライアント端末における TFHE 暗号処理の高速化に向けて，クライアント側の処理を行う機器の性能による実行時間の変化と，IoT クライアント端末での実行に向けて改善する必要がある実装の分析を行なった．実験から，クライアント側で行う処理の中で key-switching key の作成時間が長く，また，IoT クライアント端末の性能の低下により，key-switching key の作成時間が大幅に増加することがわかった．key-switching key の作成では，DRAM 容量が減少することでストレージへのアクセスが増加し，I/O 処理がボトルネックとなる．OpenFHE の実装は，メモリ使用量が少ないが，ストレージにデータが置かれた場合の読み込みの効率が低下した．TFHEpp は，メモリ使用量が多いが，ストレージへのまとまったアクセスが可能であることがわかった．また，OpenFHE，TFHEpp 共に，key-switching key を保存するためのメモリの確保や key-switching key を作成する際に用いる疑似乱数の生成で実行時間が大幅に増加することがわかった．

本稿の貢献を以下に示す．

- FHE 暗号スキーム（BFV，BGV，CKKS，TFHE 亜種方式）およびライブラリ（OpenFHE，Lattigo，TFHEpp）の時間

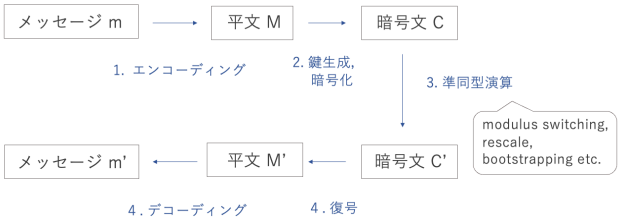


図 1: FHE の暗号処理の流れ

空間計算量を測定し，実行環境と FHE を用いて行う操作に応じて適切なものを選択するための比較を行った．

- IoT クライアント端末における TFHE 暗号処理の高速化に向けて，OpenFHE と TFHEpp ライブラリを用いた際の IoT クライアント端末における実行時間とボトルネックとなる処理の分析を行なった．

2 準　　備

2.1 完全準同型暗号

FHE は暗号化された状態で任意回加算・乗算が可能な暗号方式である．FHE は常に (1)，(2) を満たす．

Fully Homomorphic Encryption

$$Encrypt(m) \oplus Encrypt(n) = Encrypt(m+n) \ldots (1)$$
$$Encrypt(m) \otimes Encrypt(n) = Encrypt(m \times n) \ldots (2)$$

図 1 に FHE の暗号処理の流れを示す．各処理は以下のように行われる．

- **エンコーディング**: メッセージ m を平文 $\mathcal{M}$ に変換する．BFV，BGV，および CKKK 方式（2.2.1 を参照）では，平文は次数 $N-1$ の円分多項式であり，$R_t = Z_t[x]/(x^N+1)$ と表現される． 複数のメッセージ ms を 1 つの平文 $\mathcal{M}$ にパッキングすることができる．TFHE 亜種方式 (2.2.1 を参照）では，平文 $\mathcal{M}$ はトーラス上で定義され，1 つのメッセージ m が 1 つの平文 $\mathcal{M}$ に変換される．
- **鍵生成**: 平文 $\mathcal{M}$ から暗号文 C に変換するための鍵を生成する．FHE は公開鍵暗号方式を用いており，秘密鍵と公開鍵の組を作成する．秘密鍵では暗号化，復号共に可能であり，公開鍵では暗号化のみ可能である．
- **暗号化**: 平文 $\mathcal{M}$ から暗号文 C を生成する．BFV，BGV，および CKKS 方式では，暗号文 C は多項式 $R_q = Z_q[x]/(x^N+1)$ で表現される．平文 Rt は，ノイズ，スケーリングパラメータ，および公開鍵を使用して暗号文 Rq にマッピングされる．TFHE 亜種方式では暗号文は各要素がトーラスである $(n+1)$ 次のベクトルで表現される．
- **準同型演算**: BFV，BGV，CKKS 方式では準同型加算・乗算，TFHE(亜種) 方式では準同型論理演算を行う．
- **Modulus switching 又は Rescale**: (R)LWE 問題に基づきメッセージ m にノイズを加えて暗号化されるため，計算中に暗号文 C にノイズが蓄積する．ノイズは暗号文のモジュラスを q から q' $(q', q \in \mathbb{N}, q' < q)$ に変換することで削減される．この操作は，BFV， BGV 方式では modulus switching，CKKS 方式では rescale と呼ばれる．FHE 処理を開始する前に modulus switching 又は rescale の実行回数を決定する必要がある．
- **Bootstrapping**: 暗号文 C に蓄積したノイズを削減する．ノイズが閾値を超えた暗号文には bootstrapping を適用できない．適切なタイミングでブートストラップを行うことにより，任意回連続して乗算することが可能である．
- **復号**: 秘密鍵を使用して暗号文 C を復号し，平文 Rt' を出力する．
- **デコーディング**: エンコーディングの逆の手順を行い，平文 Rt' をメッセージ m' に変換する．

表 1: FHE 暗号方式の比較

FHE 暗号方式	特徴			
	算術演算	論理演算	ドメイン	パッキングサイズ
BFV	✓		integer	N
BGV	✓		integer	N
CKKS	✓		floating-point	$N/2$
TFHE		✓	binary	1
TFHE 亜種	✓	✓	floating-point	1

2.2 FHE 暗号方式とライブラリ

FHE には BFV，BGV，CKKS，TFHE [9]，TFHE 亜種方式など，複数の暗号方式が提案されており，それぞれ暗号処理の仕組みが異なる．また，各暗号方式が実行可能な複数のライブラリが公開されている．

2.2.1 暗 号 方 式

表 1 に BFV，BGV，CKKS，TFHE，TFHE 亜種方式の特徴を示す．BFV，BGV，および CKKS 方式は，bootstrapping の理論など，基本的な暗号処理の流れは同じである．BFV，BGV，CKKS 方式では，安全な状態で暗号文を復号し，復号結果を再び暗号化することで bootstrapping を実現する．また，算術演算を使用して暗号文を処理する．ここで，BFV および BGV 方式は RLWE 問題 [10] を使用して整数を評価し，1 つの平文に N 個のメッセージを格納可能である．一方で，CKKS 方式は Fourier 変換を使用して浮動小数点を評価し，1 つの平文に $N/2$ 個のメッセージを格納可能である．

TFHE 方式は，BFV，BGV，CKKS 方式と基本的な暗号処理が異なる．TFHE 方式では，暗号文を構成する各ビットが論理回路を使用して評価される．加算や乗算などの算術演算を行う場合は，論理ゲートを組み合わせる必要がある．bootstrapping は，ノイズが一定量に削減された暗号文が格納された LUT（Lookup Table）と呼ばれるベクトルの適切な位置を参照することで行われる．bootstrapping は回路を構成する各論理ゲートで実行される．

TFHE 亜種方式は TFHE 方式を拡張したもので，programmable bootstrapping と呼ばれる実装により，論理演算だけでなく単変数関数の bootstrapping も可能である．また，エンコーダの実装により，浮動小数点の評価が可能である．将来の課題の一つに計算中のドリフトエラーを削減することが挙げられる [5]．本研

表 2: FHE 暗号方式が実装されているライブラリ

FHE 暗号方式	ライブラリ		
	OpenFHE	Lattigo	TFHEpp
BFV	✓	✓	
BGV	✓	✓	
CKKS	✓	✓	
TFHE	✓		✓
TFHE 亜種	✓		✓

究では，BFV，BGV，CKKS，TFHE 亜種方式の比較を行った．

2.2.2 ライブラリ

表 2 に各暗号方式が実装されているライブラリを示す．BFV，BGV，CKKS 方式は OpenFHE，Lattigo などのライブラリで実装されている．OpenFHE はオープンソースのライブラリであり，2021 年に Google によって発表された，C++のプログラムを FHE 暗号を用いた処理に変換するトランスパイラ [11] 内で使用されている．Lattigo は Go 言語のライブラリであり，秘密分散処理も実装されている．

TFHE，TFHE 亜種方式は OpenFHE，TFHEpp などのライブラリで実装されている．TFHEpp はオープンソースの C++のライブラリで，オリジナルの TFHE の実装 [12] より約 10 ほど高速である．

2.3 TFHE 亜種方式の暗号処理

TFHE 亜種方式では，bootstrapping 処理のため，データの暗号化を行うクライアントで gate key(bootstrapping key と key-switching key) も作成する必要がある．本節では，TFHE 亜種方式の暗号化，bootstrapping の仕組みと，OpenFHE，TFHEpp ライブラリにおける key-switching key の作成に関する実装を紹介する．

2.3.1 暗　号　化

メッセージ $m \in [a, b]$ を Torus 上の点 $\mathcal{M} \in [0, 1]$ にマッピングする．また，平文 $\mathcal{M} \in \mathbb{T}$ から暗号文 $C = (\mathbf{a}, \mathrm{b})(\mathbf{a} \in \mathbb{T}^n, \mathrm{b} \in \mathbb{T})$ を求める．$\mathbf{a} \in \mathbb{T}^n$ を $\mathbb{U}_{t^n}$ を元にサンプルし，$e \in \mathbb{T}$ を $\mathbb{D}_{t,\mathbf{a}}$ を元にサンプルする．暗号文 C は，秘密鍵 $\mathrm{SK} \in \mathbb{B}^n$ を用いて以下のように表される．

$$(\mathbf{a}, b) = (\mathbf{a}, \mathbf{a} \cdot \mathrm{SK} + \mathcal{M} + e)$$

2.3.2 bootstrapping

FHE 暗号方式では LWE 問題を基に暗号文にノイズを加えることで暗号化しているため，演算を繰り返すことで暗号文にノイズが蓄積する．ノイズが閾値を超えると正しい値に復号できなくなるため，適切なタイミングで暗号文のノイズを削減する bootstrapping と呼ばれる処理を行う．

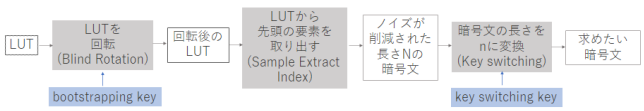


図 2: TFHE 亜種方式の bootstrapping の流れ

図 2 に TFHE 亜種方式の bootstrapping の流れを示す．TFHE 亜種方式の bootstrapping では，x 次の係数に x を暗号化した値が格納された LUT と呼ばれる多項式を用いる．LUT に対して，$V \cdot X^{-\Delta m+e} = V \cdot X^{-b+\sum_{i=0}^{n-1} a_i sk_i}$ を計算することで，Δm に対応する暗号文が 0 次の項に移動する．その後，0 次の項を取り出すことでノイズが削減された m の暗号文を得ることができる．

ここで V・X の計算で，Torus 係数多項式である $\mathbf{a}$ と整数係数多項式 SK の乗算はそのままでは行うことができず，SK を TRGSW 形式に変換する必要がある．SK を TRGSW 形式で暗号化したものは bootstrapping key とよばれる．

また，bootstrapping により得られる暗号文の $\mathbf{a}$ は要素数 N のベクトルであり，元の暗号文の $\mathbf{a}$ の要素数 n と異なる．そこで，要素数 N の暗号文の秘密鍵 SK1 を要素数 n の秘密鍵 SK で暗号化した鍵 SK2 を用いて，要素数 N の暗号文に対して復号する $(b - \sum_{i=0}^{N-1} a_i sk2_i)$ ことで，要素数 n の暗号文を作成する．復号では丸め誤差を削減するため，$SK1$ をパラメータ $digitCount, baseSK$ を用いて $\frac{sk1_i}{2^{digitCount \cdot baseSK}}$ の形で暗号化し，計算を分割する．$\frac{sk1_i}{2^{digitCount \cdot baseSK}}$ を暗号化したものは key-switching key と呼ばれる．

2.4 クライアントによる key-switching key の作成

key-switching key の作成では，要素数 N の暗号文の秘密鍵を要素数 n の秘密鍵を用いて暗号化する．一方で，TFHE 暗号ライブラリ毎に key-switching key の作成に関する実装が異なる．

図 3 に OpenFHE の key-switching key の作成に関する実装を示す．OpenFHE では Si を $digitCount$ で分解した値の暗号文を vector2A, vector2B に保存し，$sk1_0$〜$sk1_N$ に関する vector2A, vector2B を vector1A, vector1B に保存する．vector 型の変数を用いて必要なメモリが動的に確保される．vector1A, vector1B の作成後，key-switching key を保存するためのオブジェクトを作成し，vector1A, vector1B を移動する．

図 4 に TFHEpp における key-switching key の作成アルゴリズムを示す．TFHEpp では，key-switching key を保存するためのオブジェクトを事前に作成し，key-switching key を構成する配列に順番にアクセスすることで，$sk1_0$〜$sk1_N$ をそれぞれ $digitCount$ で分解した値の暗号文を保存する．

3 関 連 研 究

Sathya ら [13] は複数の FHE 暗号方式を分類し，Editor ら [14] は各 FHE ライブラリで実行できる操作を比較している．しかしながら，これらの研究は実際の実行時の評価を含んでいない．本研究では，実行時の複数の FHE 暗号方式とライブラリの時間空間計算量を比較した．Gouert ら [15] は Terminator 2 Benchmark Suite と呼ばれるベンチマークを作成し，複数のライブラリ間で高速に処理できる計算ドメインとアプリケーションを比較している．しかしながら，異なるスキームを使用してライブラリ間の比較を行っている．本研究では，使用されるライブラリと暗号方式による影響を別々に検討している．Fawaz ら [16] は Microsoft SEAL ライブラリを使用して，BFV，BGV，

```
std::make_shared<LWESwitchingKeyImpl>
KeySwitchGen(const std::shared_ptr<CryptoParams> &params,
PrivateKey &sk, PrivateKey &skN) const {
(1) key-switching keyの作成で使用する変数の宣言
    set n, N, baseKS, digitCount
    set qSK to the modulus used in key-switching
    set vector skn to the secret key whose length is n
    set vector skN to the secret key whose length is N
    // digitsKSを宣言
    int value{1};
    vector<int> digitsKS;
    digitsKS.reserve(digitCount);
    for (size_t i = 0; i < digitCount; ++i) {
        digitsKS.emplace_back(value);
        value *= baseKS;
    }
(2)  key-switching keyを作成
   // 離散一様分布に従った乱数の生成器を作成
    DiscreteUniformGeneratorImpl<vector<int>> dug(qKS);

    // key-switching keyの要素を保存するための変数resultVecA, resultVecBを宣言
    vector<vector<vector<vector<int>>>> resultVecA(N);
    vector<vector<vector<int>>> resultVecB(N);

    for (size_t i = 0; i < N; ++i) {
        // sk1_iをskで暗号化した値を保存するための変数vector1A, vector1Bを宣言
        vector<vector<vector<int>>> vector1A;
        vector1A.reserve(baseKS);
        vector<vector<int>> vector1B;
        vector1B.reserve(baseKS);

        for (size_t j = 0; j < baseKS; ++j) {
            // sk1_iをbaseSK, digitCountで分解した値をskで暗号化した値を保存するための
            // 変数vector2A, vector2Bを宣言
            vector<vector<int>> vector2A;
            vector2A.reserve(digitCount);
            vector<int> vector2B;
            vector2B.reserve(digitCount);

            for (size_t k = 0; k < digitCount; ++k) {
                vector2A.emplace_back(dug.GenerateVector(n));
                  // sk1_iをbaseSK, digitCountで分解した値をskで暗号化
                vector<int>& a = vector2A.getLastElement();
                int b =
             (params->GetDiscreteGaussianGenerator().GenerateInteger(qKS))
                                .ModAdd(svN[i].ModMul(j * digitsKS[k], qKS), qKS);
                for (size_t i = 0; i < n; ++i) {
                    b += a[i].ModMulFast(sv[i], qKS);
                }
                b.ModEq(qKS);
                vector2B.emplace_back(b);
            }
            vector1A.push_back(std::move(vector2A));
            vector1B.push_back(std::move(vector2B));
        }
        resultVecA[i] = std::move(vector1A);
        resultVecB[i] = std::move(vector1B);
    }
    return std::make_shared<LWESwitchingKeyImpl>
              (LWESwitchingKeyImpl(std::move(resultVecA), std::move(resultVecB)));
}
```

図 3: OpenFHE の key-switching key の作成に関する実装

```
// bootstrapping keyとkey-switching keyを保存するためのオブジェクトを事前に作成
std::unique_ptr<TFHEpp::GateKey> gk =
        std::make_unique<TFHEpp::GateKey>(*sk);

// key-switching keyを構成する配列に順番にアクセスし,
// sk1_0~sk1_Nをそれぞれ baseSK, digitCountで分解した値の暗号文を保存
for (int i = 0; i< N; i++) {
        for  (int j=0; j<baseSK; j++) {
                for (int k = 0; k< (1 << digitCount - 1); k++) {
                        ksk[i][j][k] = Encrypt(
                        skn[i] * (k + 1) * (1ULL << digits -(j + 1) * digitCount), skn);
                      }
        }
}

//平文を秘密鍵を用いて暗号化する関数
array<n+1> Encrypt(Plaintext p, const std::array<n> &skn) {
        // 離散一様分布に従った疑似乱数生成器を作成
        std::uniform_int_distribution  dug;
        // res is the ciphertext of plaintext
        res[P::n] = ModularGaussian<P>(p);
        for  (int i = 0; i<n; i++) {
                res[i] <-dug(linear_congruential_engine);
                res[P::n] += res[i] * skn[i];
        }
        return res;
}
```

図 4: TFHEpp の key-switching key の作成に関する実装

および CKKS 方式の各暗号処理の実行時間を比較している．しかしながら，高速な暗号方式はライブラリによって異なる．本研究では，複数のライブラリを使用して暗号方式を比較した．Acar ら [17] は SHE [17]，LHE [18]，および FHE の理論の説明と，一般に公開されていない暗号方式の実行時間を評価している．本研究では，最新のオープンソースライブラリの実行時間を比較した．Doan ら [19] は複数の暗号方式とライブラリの実行時の評価を行っている．この評価では，BFV，BGV，CKKS 方式について暗号文多項式の次数 N を変化させた際のライブラリ間の実行時間を比較しているが，本研究では乗算深度を変化させた際の時間空間計算量を比較した．

表 3: BFV，BGV，CKKS 方式の比較で使用するパラメータ

(a) Parameters

暗号方式	t	N	batch
BFV	65537	32768	-
BGV	65537	32768	-
CKKS	65537	65536	32768

(b) Ciphertext modulus q

暗号方式	ライブラリ	乗算深度				
		2	4	6	8	10
BFV	Lattigo	118	177	235	294	352
	OpenFHE	118	177	236	295	354
BGV	Lattigo	118	235	352	410	583
	OpenFHE	118	236	354	472	590
CKKS	Lattigo	146	236	326	416	506
	OpenFHE	145	235	325	415	505

表 4: TFHE 亜種方式の評価で使用するパラメータ

ライブラリ	N	n	q	gadgetBase	baseSK
OpenFHE	2048	512	2^{27}	2^7	2^7
TFHEpp	2048	500	2^{32}	2^6	2^4

FHE の時間計算量が大きい課題に対して，ハードウェアを用いた高速化も検討されている．Jung ら [20] は GPU に適した FHE アルゴリズムを検討し，Riazi ら [21] と Beirendonck ら [22] は FPGA，Kim ら [23] は ASIC を使用した新しいアーキテクチャを提案している．これらの研究は，サーバにおける FHE 処理の高速化を検討している．本研究では，TFHE 亜種暗号方式を用いた際の IoT クライアント端末の処理の高速化に向けて，IoT クライアント端末における実行時間とボトルネックとなる処理の分析を行なった．

4 FHE 暗号方式とライブラリの評価

OpenFHE，Lattigo，および TFHEpp ライブラリでそれぞれ実装された BFV，BGV，CKKS，TFHE 亜種方式を比較した．評価は以下の 3 つのステップで行った．

- 暗号文に対して算術演算を行い，複数のメッセージを 1 つの平文にパッキング可能な BFV，BGV，CKKS 方式．Lattigo および OpenFHE ライブラリに実装された BFV，BGV，および CKKS 方式の時間空間計算量を乗算深度を変化させながら比較した．
- 暗号文を構成する 1bit ずつに対して論理演算を行う TFHE 亜種方式．OpenFHE と TFHEpp ライブラリの時間空間計算量を比較した．
- CKKS 方式と TFHE 亜種方式．CKKS 方式では OpenFHE ライブラリ，TFHE 亜種方式では FHEpp ライブラリを使用して暗号方式毎の適切な操作を調査した．

表 5: CKKS 方式の式 (1)，式 (2) の評価で使用するパラメータ

n	t	N	batch	q	乗算深度
式 (1)					
2	65537	8192	4096	183	2
20	65537	65536	32768	955	20
40	65537	65536	32768	1475	26
60	65537	65536	32768	1640	29
80	65537	65536	32768	1750	31
100	65537	65536	32768	1805	32
式 (2)					
2	65537	4096	2	78	1
20	65537	4096	32	78	1
40	65537	4096	64	78	1
60	65537	4096	64	78	1
80	65537	4096	128	78	1
100	65537	4096	128	78	1

4.1 実行環境

表 3 に BFV, BGV, CKKS 方式を比較する際に使用したパラメータ，表 4 に TFHE 亜種方式を評価する際に使用したパラメータを示す．TFHE 亜種方式では，暗号文は LWE 問題 [24] に基づくベクトルと RLWE 問題に基づく多項式環の 2 つの形式を用いる．gadgetBase は，RLWE に基づく暗号文間の乗算のための decomposition と呼ばれる操作において，暗号文のノイズを削減するために使用されるパラメータである．baseSK は，RLWE に基づく暗号文を LWE 問題に基づく暗号文に変換するための key-switching と呼ばれる操作で使用する．表 5 に CKKS 方式と TFHE 亜種方式で得意な演算を比較する際に，式 (1) および式 (2) (4.B を参照) で使用したパラメータを示す．各評価において，128 ビットのセキュリティを満たす最速のパラメータを使用した．CPU は AMD Ryzen 7 5700G@3.80GHz を使用し，8 つのコアと 16 の論理 CPU を備えている．DRAM は 16GB の DDR4 を使用し，ストレージは容量が 238GB の SK hynix PC711 SSD を用いた．

4.2 評価結果

4.2.1 BFV，BGV，CKKS 方式の比較

図 5 に OpenFHE と Lattigo を使用して BFV，BGV，および CKKS 方式の乗算深度を変化させた際の実行時間の比較を示す．暗号方式の比較では，全ての乗算深度で BGV，BFV，CKKS 方式の順に高速である．全ての暗号方式とライブラリの組み合わせの中で，Lattigo の BGV 方式が最速である．BFV と CKKS 方式では，乗算深度が小さい場合は Lattigo が高速であり，大きい場合は OpenFHE が高速である．図 6 に OpenFHE と Lattigo を使用して BFV，BGV，および CKKS 方式の乗算深度を変化させた際のメモリ使用量を示す．CKKS 方式は BFV と BGV 方式に比べてメモリ使用量が多い．すべての暗号方式および乗算深度に対して，OpenFHE はより少ないメモリ容量で実行可能である．また全ての暗号方式とライブラリにおいて，実行時間とメモリ使用量は乗算深度または暗号文多項式の次数 N の増加に伴い線形に増加する．

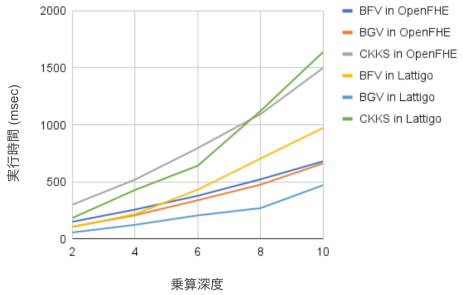


図 5: OpenFHE と Lattigo に実装された BFV，BGV，および CKKS 方式の乗算深度を変化させた際の実行時間の比較

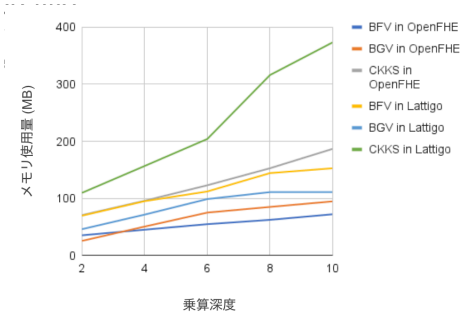


図 6: OpenFHE と Lattigo に実装された BFV，BGV，および CKKS 方式の乗算深度を変化させた際のメモリ使用量の比較

表 6: TFHE 亜種方式が実装された OpenFHE と TFHEpp の実行時間の比較

ライブラリ	sk の生成 (ms)	gk の生成 (ms)	暗号化 (ms)	bootstrapping (ms)	復号 (ms)
OpenFHE	0.801	3652.380	0.040	463.729	0.007
TFHEpp	0.041	6442.310	0.018	26.865	0.003

4.2.2 TFHE 亜種方式の時間空間計算量

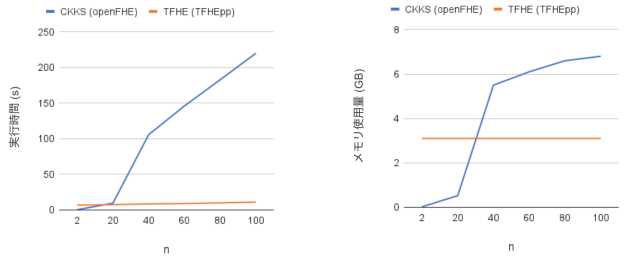
表 6 に OpenFHE と TFHEpp ライブラリを用いて TFHE 亜種方式を実行した際の実行時間を示す．gate key の生成と bootstrapping の実行時間が長いことが分かる．gate key の生成において，OpenFHE は TFHEpp よりも 1.764 倍高速である．一方で，bootstrapping では TFHEpp は OpenFHE よりも 17.261 倍高速である．したがって，bootstrapping が複数回実行される場合は TFHEpp が高速である．両ライブラリとも，gate key の生成中にメモリ使用量が線形に増加し，完了時には約 3.3GB に達する．その後実行される bootstrapping では，メモリ使用量は約 3.3GB で推移する．5 章で詳細な評価を行う．

4.2.3 CKKS 方式と TFHE 亜種方式の比較

この実験では，(1) で表される任意の複数の値の乗算，(2) で表されるベクトル内積について，OpenFHE ライブラリに実装された CKKS 方式と TFHEpp ライブラリに実装された TFHE 亜種方式の時間空間計算量を比較した．$x \in \mathbb{R}$，$\boldsymbol{a} = (a_0, a_1, ..., a_{n-1})$，$\boldsymbol{b} = (b_0, b_1, ..., b_{n-1})$ とすると，(1) と (2) は以下のように定義される．

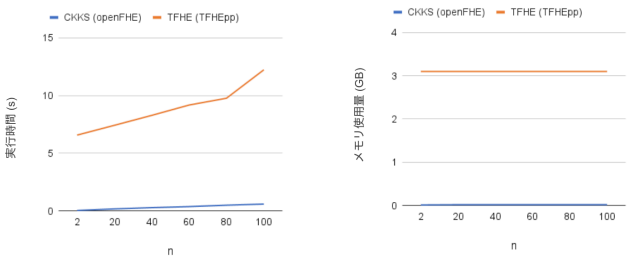
$$f(x) = \prod_{i=0}^{n-1} x_i \tag{1}$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=0}^{n-1} a_i b_i \tag{2}$$



(a) 実行時間　(b) メモリ使用量

図 7: CKKS 方式と TFHE 亜種方式における式 (1) の時間空間計算量の比較



(a) 実行時間　(b) メモリ使用量

図 8: CKKS 方式と TFHE 亜種方式における式 (2) の時間空間計算量の比較

式 (1) の評価　(1) を実行する際，CKKS 方式では，n 個のメッセージがパッキングされた暗号文多項式に対して Rotation と呼ばれる操作を行い，平文ベクトルが一つずつシフトされた新しい暗号文を作成する．それらの暗号文を掛け合わせることで，暗号化された n 個のメッセージの乗算が完了する．乗算が連続して行われるため，一定の乗算深度に達するたびに bootstrapping を実行する必要がある．一方で，TFHE 亜種方式では，各メッセージに対応する n 個の暗号文を作成し，それらを乗算する．毎回の乗算後に programmable bootstrapping を実行する．

図 7 に CKKS 方式および TFHE 亜種方式を使用して式 (1) を実行する際のデータ数 n の変化に伴う時間空間計算量の比較を示す．データ数 n が 20 未満の場合，CKKS 方式の時間空間計算量が小さく，40 以上の場合，TFHE 亜種方式の時間空間計算量が小さい．CKKS 方式ではパッキングされた暗号文に対して n 回，Rotation と乗算を行う必要がある．乗算回数が増加すると，必要な bootstrapping の回数も増加する．データ数 n が 0，20，40，60，80，100 の場合，それぞれ 0 回，2 回，3 回，4 回，5 回の bootstrapping が必要である．1 回の bootstrapping に必要な時間空間計算量が大きいため，必要な bootstrapping 回数が増加すると，TFHE 亜種方式と比較して効率が著しく低下する．TFHE 亜種方式では，n 回 bootstrapping が必要であり，時間空間計算量はデータ数 n に比例して増加する．ただし，bootstrapping に必要な時間空間計算量が小さいため，著しい増加は見られない．

式 (2) の評価　(2) を実行する際，CKKS 方式では，n 個のメッセージをそれぞれパッキングした 2 つの暗号文を乗算する．その後，乗算結果の暗号文にパッキングされている要素間の加算を行う．一方で，TFHE 亜種方式では，各ベクトルの要素のセット毎に乗算と bootstrapping を実行する必要がある．その後，n 個の乗算結果の加算を行う．図 12 に CKKS および TFHE 亜種方式を使用して式 (2) を実行する際のデータ数 n の変化に伴う時間空間計算量の比較を示す．全てのデータ数 n において，CKKS 方式は TFHE 亜種方式より小さい時間空間計算量で演算が可能である．CKKS 方式では，n 個の 2 要素間の乗算結果は 2 つの暗号文間の 1 回の乗算のみで得られ，bootstrapping は不要である．TFHE 亜種方式では，(1) と同様に，各ベクトル要素のセット毎に n 回の乗算と bootstrapping を行う必要がある．

表 7: クライアント側の評価で使用した機器

	使用した機器			
	hpc5000-xsl	MacBook Pro	Raspberry Pi 4	Raspberry Pi Zero 2W
CPU	Intel(R) Xeon(R) Gold 5115 CPU@2.40GHz 10Cores × 2	Apple M2 @3.49GHz 12Cores × 1	Cortex-A72 @2.50GHz 4Cores × 1	Cortex-A53 @1GHz 4Cores × 1
アーキテクチャ	AMD64	ARMv8	ARMv8	ARMv8
DRAM 容量	192GB	16GB	8GB	512MB
ストレージ	SSD	SSD	SanDisk Ultra microSD	Samsung EVO Plus microSD

5　IoT クライアント端末における TFHE 亜種方式の評価

FHE の実用に向けた課題は時間空間計算量が大きいことである．ここで，FHE を用いて IoT デバイスで収集されたデータの分析を行う場合，DRAM 容量や CPU に制約がある IoT クライアント端末で FHE 処理で使用する鍵の作成やデータの暗号化を行う必要がある．本章では，IoT クライアント端末における TFHE 暗号処理の高速化に向けて，クライアント側の処理を行う機器の性能による実行時間の比較と，Raspberry Pi Zero 2W のような低性能な IoT クライアント端末での実行に向けて改善が必要な実装の特定を行なった．

5.1　実 行 環 境

表 7 にクライアント側の処理の評価で使用した機器を示す．hpc5000-xsl は 20 個の論理コアと 192GB の DRAM を備えたサーバである．IoT クライアント端末との比較のために hpc5000-xsl で同じ処理を実行した．MacBook Pro の CPU は Apple M2 で，アーキテクチャは ARMv8 である．Raspberry Pi 4 と Raspberry Pi Zero 2W は，IoT クライアント端末と同様の性能を持ち，CPU は組み込み機器で用いられる Cortex-A72 または Cortex-A53 である．Raspberry Pi 4 の DRAM 容量は 8GB，Raspberry Pi Zero 2W の DRAM 容量は 512MB であり，ストレージは micro SD カードを用いる．

5.2　実 験 結 果

5.2.1　クライアント側の処理を行う機器による実行時間の比較

図 9 に使用する機器によるクライアント側の処理時間 (秘密鍵

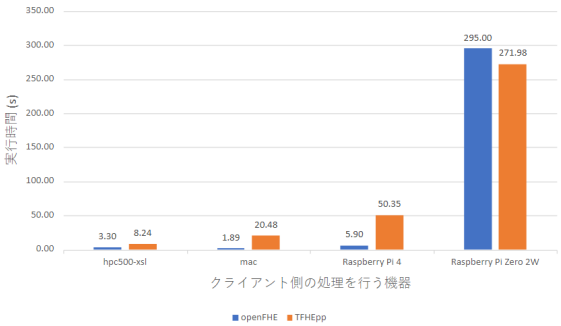


図 9: 使用する機器によるクライアント側の処理時間の比較

表 8: クライアント側の処理の実行時間の内訳

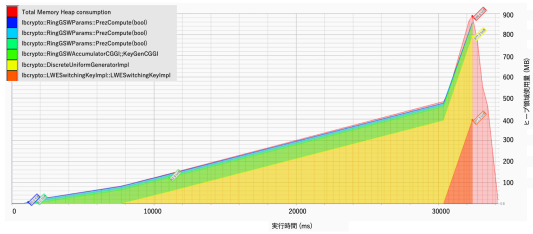
機器	ライブラリ	クライアント側で行う処理			
		秘密鍵の作成 (ms)	bootstrapping key の作成 (ms)	key-switching key の作成 (ms)	暗号化 (ms)
Raspberry Pi 4	OpenFHE	0.15	757.17	5798.69	5.12
	TFHEpp	0.26	2602.71	47775.20	2.69
Raspberry Pi Zero 2W	OpenFHE	2.94	1540.30	293460.08	150.12
	TFHEpp	2.27	7458.45	216317.21	43.409

の作成，bootstrapping key の作成，key-switching key の作成，1bit の暗号化の合計) の比較を示す．Raspberry Pi 4 から Raspberry Pi Zero 2W に変更することで実行時間が増加し，OpenFHE では約 50 倍，TFHEpp では約 5.4 倍となる．DRAM 容量が 8GB から 512MB に制限されることが原因であり，512MB 以上 8GB 以下に現在の OpenFHE，TFHEpp の実装で高速に実行可能なメモリ容量の最小値があることがわかる．また，TFHEpp ライブラリでは hpc5000-xsl から MacBook Pro に変更することで実行時間が約 2.48 倍増加する．TFHEpp では実行時のアーキテクチャを AMD64 から ARMv8 に変更することで，処理の効率が低下することが推測される．

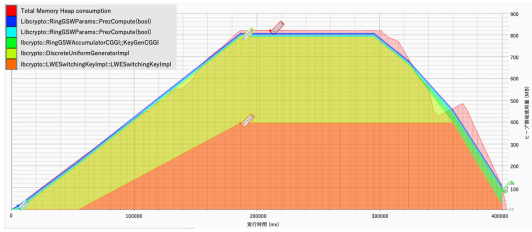
表 8 にそれぞれの機器におけるクライアント側の処理の実行時間の内訳を示す．両ライブラリで，最も高速に実行可能な hpc5000-xsl において key-switching key の作成時間が 90 %以上となる．また，hpc5000-xsl から Raspberry Pi Zero 2W に変更することで，key-switching key の作成時間が OpenFHE では約 156 倍，TFHEpp では約 30.2 倍に増加する．処理を行う機器の性能に合わせた key-switching key の作成方法の検討が必要である．

5.2.2 Raspberry Pi Zero 2W における key-switching key の作成時間の増加の原因

Raspberry Pi Zero 2W のような低性能な IoT クライアント端末における key-switching key の作成の高速化に向けて，Raspberry Pi 4 と Raspberry Pi Zero 2W で実行した際の CPU 使用率とストレージ帯域幅の比較を行なった．OpenFHE における key swtching key の作成では，複数のコアを用いて並列処理を行い，Raspberry Pi 4 使用時の平均 CPU 使用率は 40%程度で推移する．一方で，Raspberry Pi Zero 2W を用いることで，iowait が約 50%，idle 値が約 40%で推移し，平均 CPU 使用率は約 10%に低下する．TFHEpp における key-switching key の作成では，並列処理を行わず，Raspberry Pi 4 使用時の平均 CPU 使用率は約 21%で推移する．Raspberry Pi Zero 2W を用いることで，iowait



(a) Raspberry Pi 4



(b) Raspberry Pi Zero 2W

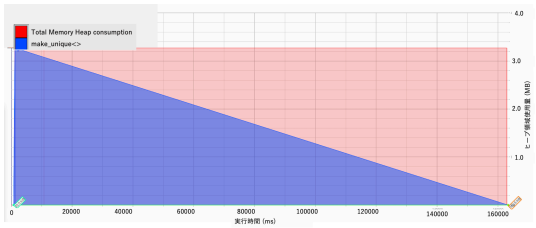
図 10: OpenFHE におけるヒープ領域の使用量

が約 80%，idle 値が約 0%で推移し，平均 CPU 使用率は約 20%に低下する．両ライブラリとも，io 待ち時間が増加することで CPU 使用率が低下し，処理が低速になることがわかる．

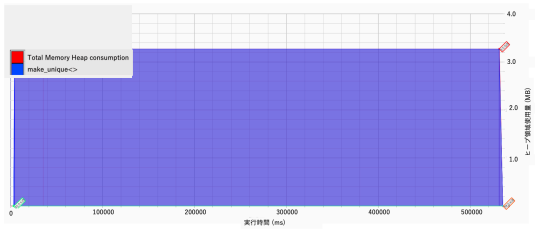
Raspberry Pi Zero 2W を用いた際のストレージ帯域幅は，OpenFHE では key-switching key 作成時に Read, Write 共に約 16MB/s で推移し，帯域幅使用率は 100%である．TFHEpp では Read が約 2MB/s，Write が約 15MB/s 以上で推移し，帯域幅使用率は 100%以上である．OpenFHE の実装は，TFHEpp に比べて Read の帯域幅が上昇し，ストレージからの読み込みの効率が悪いことがわかる．一方で，ストレージへの書き込みは TFHEpp ライブラリの方が高く，広くメモリを使用する実装であることがわかる．以上から，Raspberry Pi Zero 2W における OpenFHE，TFHEpp を用いた TFHE 亜種方式の実行では，ストレージへの Read/Write 量が多いことで I/O 待ち時間が発生することがわかる．ストレージへの Read/Write の量を削減することで高速に実行できることができる．

5.2.3 key-switching key 作成時の処理毎のヒープ領域の使用量

メモリ使用量を削減し，DRAM 上で処理を行うことで IO 待ち時間を削減するため，OpenFHE，TFHEpp ライブラリにおける key-switching key 作成時の関数毎のメモリ使用量を計測した．図 10 に OpenFHE における Raspberry Pi Zero 2W または Raspberry Pi 4 を用いた際の key-switching key 作成時のヒープ領域の使用量を示す．OpenFHE では，最大で 800MB 程度のヒープ領域を使用する．key-switching key を保存するクラスのコンストラクタである LWESwitchingKeyImpl::LWESwitchingKeyImpl と離散一様分布に従った疑似乱数生成器を作成する DiscreteUniformGeneratorImp::DiscreteUniformGeneratorImpl のヒープ領域の使用量が多く，それぞれ約 377MB を用いる．また，実行時のメモリが 512MB に制限されることで，LWESwitchingKeyImpl::LWESwitchingKeyImpl と DiscreteUniformGeneratorImp::DiscreteUniformGeneratorImpl の実行時間が増加する．

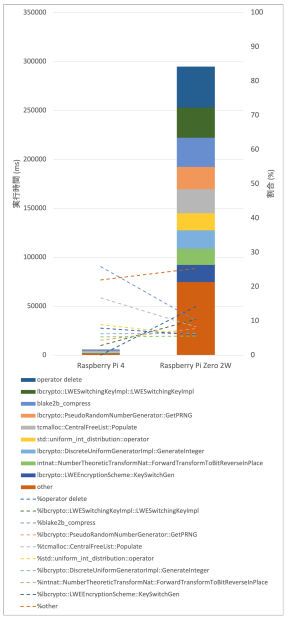


(a) Raspberry Pi 4

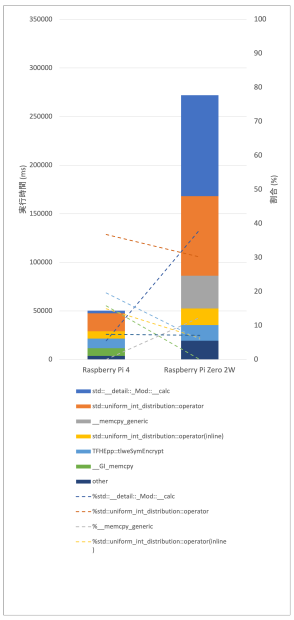


(b) Raspberry Pi Zero 2W

図 11: TFHEpp におけるヒープ領域の使用量



(a) OpenFHE



(b) TFHEpp

図 12: Raspberry Pi 4 または Raspberry Zero 2W を用いた際の処理毎の実行時間の比較

OpenFHE のように，メモリ使用の再利用や確保の工夫を行うことで，ストレージにデータが置かれた場合の読み込みの効率が低下するため，実行時のメモリ使用量に合わせてメモリを確保する方法を検討する必要がある．

図 11 に TFHEpp における Raspberry Pi Zero 2W または Raspberry Pi 4 を用いた際の key-switching key 作成時のヒープ領域の使用量を示す．TFHEpp では，3.3GB 程度のヒープ領域を使用し，処理の後半でメモリリークが発生する．gate key 作成前に gate key を保存するためのオブジェクトを作成する際にヒープ領域が確保される．

5.2.4 I/O 待ち時間の改善が必要な処理

DRAM 容量が 8GB から 512MB に制限されることで I/O 待ち時間が増加する key-switching key 作成時の処理の分析を行なった．図 12(a) に OpenFHE における Raspberry Pi 4 または Raspberry Zero 2W を用いた際の処理毎の実行時間の比較を示す．OpenFHE では Raspberry Pi 4 から Rasberry Pi Zero 2W に変更することで，key-switching key の作成時間が約 156 倍増加する．最も実行時間の増加の割合が大きい関数は，bootstrapping key, key-switching key の作成時にヒープ領域に確保したメモリを解放する operater:delate である．二番目に実行時間の増加の割合が大きい関数は，key-switching key を保存するためのクラスのコンストラクタである lbcrypto:LWESwitchingKeyImpl::LWESwitchingKeyImpl である．key switching key を格納するためのサイズが大きい vector 型の変数がヒープ領域に確保される．三番目に，離散一様分布に従って key-switching key の一部を作成するための疑似乱数を生成する lbcrypto:RseudoRandomNumberGenerator である．

図 12(b) に TFHEpp における Raspberry Pi 4 または Raspberry Zero 2W を用いた際の処理毎の実行時間の比較を示す．TFHEpp では Raspberry Pi 4 から Rasberry Pi Zero 2W に変更することで，key-switching key の作成時間が約 30.2 倍増加する．最も実行時間の増加の割合が大きい関数は，離散一様分布に従って key-switching key の一部を作成するための疑似乱数を生成する linear_congruential_engine で呼び出される std::__detail_Mod__calc である．二番目に実行時間の増加の割合が大きい関数は，gate key オブジェクトを作成する際や key-swiching key に値を格納する際にメモリをコピーするために呼び出される__memcpy_generic である．

OpenFHE，TFHEpp 共に，key-switching key を保存するためのメモリの確保の方法や key-switching key を作成する際に用いる疑似乱数の生成方法の改善が必要であることがわかる．

6 ま と め

本稿では初めに，複数の暗号ライブラリ（OpenFHE，Lattigo，TFHEpp）および暗号方式（BFV，BGV，CKKS，TFHE 亜種）の時間空間計算量を比較した．整数を暗号処理する場合，Lattigo の BGV が全ての暗号方式，ライブラリの組み合わせの中で最速であった．浮動小数点を評価できる CKKS 方式では，乗算深度が小さい時は Lattigo が高速であり，乗算深度が大きい時は OpenFHE が高速であることを示した．また，全ての暗号方式で OpenFHE は Lattigo より少ないメモリで実行可能であった．CKKS，BFV，BGV 方式と TFHE 亜種方式は異なる基本的な暗号処理の単位と bootstrapping の仕組みを備えているため，得意な演算が異なる．任意の値を 40 回以上乗算する場合，TFHE 亜種方式は CKKS 方式より時間空間計算量が小さい．一方で，ベクトル内積を計算する場合，CKKS 方式は TFHE 亜種方式より時間空間計算量が小さい．

二番目に，TFHE 亜種暗号方式が実装された OpenFHE と TFHEpp について，低性能な IoT クライアント端末における実行時間とボトルネックとなる処理の分析を行なった．クライア

ント側で行う処理の中で key-switching key の作成時間が長く，また，IoT クライアント端末の性能の低下により，key-switching key の作成時間が大幅に増加することがわかった．key-switching key の作成では，DRAM 容量が減少することでストレージへのアクセスが増加し，I/O 処理に関するボトルネックが発生することが原因である．帯域幅を削減するため，実行時のメモリ使用量と，ストレージにデータが置かれた場合の読み込みの効率をする改善する必要がある．特に改善が必要な処理は，OpenFHE, TFHEpp 共に，key-switching key を保存するためのメモリの確保や key-switching key を作成する際に用いる疑似乱数の生成であることがわかった．

今後は，IoT クライアント端末における TFHE 暗号処理の高速化に向けて，メモリ使用量が少ないアルゴリズムや効率的なストレージアクセスの方法の検討を行う．

謝　　辞

本研究は一部，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Gentry, C. 2009.: A FULLY HOMOMORPHIC ENCRYPTION SCHEME, PhD Thesis, Stanford University (2009).
[2] Fan, J. and Vercauteren, F.: Somewhat practical fully homomorphic encryption, Cryptology ePrint Archive, Report 2012/144 (2012).
[3] Brakerski, Z., Gentry, C. and Vaikuntanathan, V.: (leveled) fully homomorphic encryption without bootstrapping. *In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (ITCS’12), pp.309-325. Association for Computing Machinery (2012).
[4] Cheon, H.J., Kim, A., Kim, M. and Song, Y.: Homomorphic encryption for arithmetic of approximate numbers, *International Conference on the Theory and Application of Cryptology and Information Security* (Asiacrypt’17), pp.409-437, Cryptology ePrint Archive, Report 2017/421 (2017).
[5] Chillotti, I., Joye, M. and Paillier, P.: Programmable Bootstrapping Enables Efficient Homomorphic Inference of Deep Neural Networks, *Cyber Security, Cryptology, and Machine Learning* (CSCML’21), pp.1-19, Cryptology ePrint Archive, Report 2021/091 (2021).
[6] OpenFHE: https://github.com/openfheorg/openfhe-development.
[7] Lattigo: https://github.com/tuneinsight/lattigo.
[8] TFHEpp: https://github.com/kenmaro3/TFHEpp.
[9] Chillotti, I., Gama, N., Georgieva, M., and Izabach^^c3^^a8ne, M.: TFHE: fast fully homomorphic encryption over the torus. *Journal of Cryptology*, Vol.33, No.1, pp.34-91. (2020).
[10] Brakerski, Z., Gentry, C. and Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (ITCS’12), pp.309-325, Association for Computing Machinery (2012).
[11] Google: https://github.com/google/fully-homomorphic-encryption.
[12] TFHE: https://github.com/tfhe/tfhe.git.
[13] Sathya, S.S., Vepakomma, P., Raskar, R., Ramachandra, R. and Bhattacharya, S.:A Review of Homomorphic Encryption Libraries for Secure Computation, arXiv preprint arXiv:1812.02428 (2018).
[14] Editor, I. and El-Yahyaoui, A.: Fully Homomorphic Encryption: State of Art and Comparison, *International Journal of Computer Science and Information Security* (IJCSIS’16), Vol.14, No.4 (2016).
[15] Gouert, C., Mouris, D. and Tsoutsos, G.N.: SoK: New Insights into Fully Homomorphic Encryption Libraries via Standardized Benchmarks, *Proceedings on Privacy Enhancing Technologies* (PoPETs’23), Issue 3, pp.154-172, Cryptology ePrint Archive, Report 2022/425 (2022).
[16] Fawaz, M.S., Belal, N., Elrefaey, A. and Fakhr, W.M.: A Comparative Study of Homomorphic Encryption Schemes Using Microsoft SEAL, *Journal of Physics Conference Series*, Vol.2128, 012021 (2021).
[17] Acar, A., Aksu, H., Uluagac, S.A. and Conti, M.: A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys*, Vol.51, Issue 4, No.79, pp.1-35 (2018).
[18] Armknecht, F., Boyd, C., Carr, C. et al.: A Guide to Fully Homomorphic Encryption, *International Association for Cryptologic Research* (IACR), Report 1192, pp.1-35 (2015).
[19] Doan, T.V.T., Messai, M., Gavin, G. and Darmont, J.: A survey on implementations of homomorphic encryption schemes . *The Journal of Supercomputing*, Vol.79, pp.15098–15139 (2023).
[20] Jung, W., Kim, S., Ahn, H.J., Cheon, H.J., and Lee, Y.: Over 100x Faster Bootstrapping in Fully Homomorphic Encryption through Memory-centric Optimization with GPUs, *IACR Transactions on Cryptographic Hardware and Embedded Systems*, Issue 4, pp.114^^e2^^80^^93148 (2021).
[21] Riazi, S.M., Laine, K., Pelton, B. and Dai, W.: HEAX: An Architecture for Computing on Encrypted Data. *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems* (ASPLOS’20), pp.1295-1309, arXiv preprint arXiv:1909.09731 (2020).
[22] Beirendonck, V.M., D’Anvers, J., Turan, F., Verbauwhede, I.: FPT: a Fixed-Point Accelerator for Torus Fully Homomorphic Encryption, *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (CCS’23), pp.741-755, Cryptology ePrint Archive, Paper 2022/1635 (2022).
[23] Kim, S., Kim, J., Kim, J.M. et al.: BTS: An accelerator for bootstrappable fully homomorphic encryption, *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp.711-725, arXiv preprint arXiv:2112.15479 (2021).
[24] Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM* (JACM),Vol.56, Issue 6, pp.1-40 (2009).

辞書を用いた圧縮センシングによる準同型暗号化データの通信量削減

泉　湖雪†　松本　茉倫†　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †koyuki@ogl.is.ocha.ac.jp, marin@ogl.is.ocha.ac.jp, oguchi@is.ocha.ac.jp

あらまし　準同型暗号は暗号文同士の加算や乗算が可能であり，データを秘匿したままで分析し活用可能な技術の一つである．準同型暗号の応用例としては暗号化データベースが挙げられる．クライアントが暗号化したデータをデータベースサーバで集計し，データ分析者は集計結果だけを得られるようになる．しかしながら，このようなアプリケーションを想定した場合，準同型暗号は平文に比べ暗号文サイズが大きいため，クライアント・データベースサーバ間の通信量が大きくなる．本研究では辞書を用いた圧縮センシングによる通信量削減を提案する．提案手法では，事前にクライアントの一部のデータを辞書学習して辞書を作成し，サーバに送信しておく．そして，クライアントが次元圧縮したレコードを暗号化し，データベースサーバが圧縮前のデータを暗号化したままで辞書を利用して復元することで，クライアント・データベースサーバ間の通信量を抑えてデータ集計に利用できる．評価の結果，圧縮率を 50% に設定した場合は提案手法は圧縮せずに送信した場合に比べて 50%通信量を削減できることを示した．復元精度については POS データの収集を想定し，カウントクエリの誤差を RMSE を用いて評価した．また，復元した POS データのレコードの誤差と圧縮率/学習量の関係性についても調査した．

キーワード　準同型暗号, 圧縮センシング, 辞書学習

1　はじめに

スマートフォンなどの IoT デバイスが普及したことで，クラウドなどのサーバにデータが蓄積され，分析に利用されるようになった．収集したデータは個人に関する情報を含む場合があるため，外部からの攻撃などによるデータの漏洩に備える必要がある．そこで，暗号文同士の加算や乗算が可能であり, データを秘匿したままでデータ活用利用な技術の一つである準同型暗号が有用である．準同型暗号の応用例としては暗号化データベースが挙げられる．クライアントにより暗号化されたデータはデータベースサーバに蓄積されるが，準同型暗号化されていると，個々のレコードを公開せずに集計結果のみを分析者に共有できる．しかしながら，暗号文は平文よりもサイズが大きいという準同型暗号の性質により，このようなアプリケーションを想定した場合，クライアント・データベースサーバ間の通信量が大きくなってしまう．本研究では，図 1 に示すような圧縮センシングを使った通信量削減を提案する．提案手法では多次元の平文を低次元に圧縮してから準同型暗号化することで，クライアント・サーバ間の通信量を削減する．そしてサーバで，暗号化された状態で低次元データから多次元データを辞書を用いて復元する．復元時には，クライアントの一部のデータを辞書学習して作成した辞書を用いている．復元時に辞書を用いることによって，スパースではないデータに対しても復元可能となっている．

圧縮センシングは，スパース性を利用することにより，少ないサンプル数からもとのデータを復元する技術である．圧縮行列を掛けることで低次元データに圧縮し，得られた低次元データと圧縮行列から元の多次元データを推定して復元する．また，

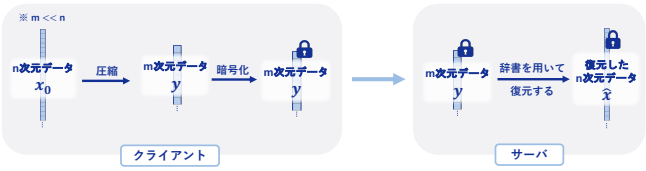


図 1: 提案手法.

辞書学習とは一部のデータから最適なスパース変換を学習することである．これらの技術を用いてクライアントが次元圧縮したレコードを暗号化し，データベースサーバが元データを暗号化したままで辞書を用いて復元することで，クライアント・データベースサーバ間の通信量を抑えてデータ集計に利用できる．

提案手法では，クライアント側で圧縮センシングによる次元圧縮と準同型暗号ライブラリ Microsoft SEAL による準同型暗号化を行う．また，事前にクライアントの一部のデータを使用して K-Singular Value Decomposition(以下 K-SVD) で辞書学習を行い辞書を作成し，サーバに送信する．サーバ側では作成した辞書を用いて，暗号化したまま圧縮センシングによる復元を行う．代表的な再構成アルゴリズムとして Orthogonal Matching Pursuit(以下 OMP) があるが，OMP で復元する際には比較演算を行う．準同型暗号では比較演算が不得意であるため，比較演算を使用しない復元アルゴリズムを用いる．今回は一般逆行列による L_2 ノルムの最小化によって圧縮前のデータを推定する．実験では，POS データのデータセットを用いて，クライアント・サーバ間の通信量，サーバでの辞書を用いた圧縮センシングによる復元結果/カウントクエリ結果の精度，実行時間の比較を Baseline と提案手法で行う．ここでは単純に，スパースな多次元データをクライアントで準同型暗号化して，圧縮せずにサーバに送信する方法を Baseline とする．

評価の結果，圧縮率を 50%に設定した場合は提案手法は圧縮せずに送信した場合に比べて 50%通信量を削減できることを示した．復元精度については，圧縮率 50 % ($n = 256, m = 128$), 学習量 25 %としたときの 256 個の $\hat{\boldsymbol{x}}$ の RMSE は，おおよそ 0.025 となった．POS データの収集を想定したカウントクエリの誤差の平均値は 0.22 だった．また，復元した POS データのレコードの誤差と圧縮率/学習量の関係性についても調査した．

2 関 連 研 究

ここでは，準同型暗号文のサイズ削減，準同型暗号とデータベースに関する研究について説明する．

2.1 準同型暗号文サイズの削減

準同型暗号の最適化についての研究は主に準同型演算を高速化するものが多く [1] [2] [3]，暗号文サイズの削減について取り組んだ研究は少ない．ここでは，Key Switching による暗号文サイズの削減を紹介する．松本ら [4] は，IoT デバイスにおける Leveled 準同型暗号 (以下 LHE: Homomorphic Encryption) の負荷を軽減するために，RLWE 暗号と LHE を組み合わせることを提案した．この方法では，IoT デバイス側では LHE よりも軽い RLWE 暗号で平文を暗号化し，代わりにクラウドサーバー側で RLWE 暗号文から LHE 暗号文に切り替える．提案手法では既存手法に比べて約 73〜233 倍暗号文サイズが小さくなることを示した．一方で，この手法は管理する鍵の数が増えるというデメリットもある．本研究で提案する手法では，管理する鍵の本数を増やすことなく暗号文サイズを削減できる．

2.2 準同型暗号とデータベース

準同型暗号を使用したデータベースの代表例には Crypt ϵ [5] や HEDA [6] がある．

Crypt ϵ は準同型暗号化されたデータベースに対し，クエリを実行し，差分プライバシ [7] を満たす分析結果を出力するシステムである．Crypt ϵ ではデータベースのレコードは one-hot エンコーディングによって表現される．one-hot エンコーディングを使った条件一致の簡単な例を示す．例えば，r を $Age = 24$ のレコード，q を $Age = 50$ のクエリとする．

$$r = [\underbrace{0, 0, ..., 0}_{23}, 1, \underbrace{0, 0, ..., 0}_{76}]$$

$$q = [\underbrace{0, 0, ..., 0}_{49}, 1, \underbrace{0, 0, ..., 0}_{50}]$$

r と q が暗号化された状態での内積は暗号化された条件一致かどうか (0 または 1) を表す．

one-hot エンコーディングを使用しない場合の例として，Ren ら [6] による FHE に基づいたデータベース分析システムである HEDA が挙げられる．比較と集計は理論的には FHE でサポートされているが，効率を最大化するために慎重にアルゴリズムを設計する必要があり，これまで検討されてこなかった．このような中で，HEDA では one-hot エンコーディングを使わずに条件一致を抽出するために，2 種類の FHE の方式を採用している．同時に，FHE 方式間の暗号文の変換など，one-hot エンコーディングに比べて多くの処理を必要とする．

また，Rasoul ら [8] は準同型暗号化された DB 上で条件一致を取り出す際の演算方法を比較した．条件一致の演算として，Folklore と Constant-weight Codewords という方法が紹介されている．クエリが平文の場合は Constant-weight のほうが Folklore より速く，クエリが暗号文の場合はより小さなパラメータが使用される場合に Constant-weight が優れている．本研究では POS データの収集を想定し，カウントクエリの誤差を評価しているが，その際 Folklore を用いてカウントクエリを実行している．

3 準　　備

本節では，本研究で使用した準同型暗号と圧縮センシング, 辞書学習について説明する．

3.1 準同型暗号

3.1.1 特　　徴

準同型暗号とは，以下のような特徴を持つ公開鍵暗号である．

$$Decrypt(Encrypt(m) \oplus Encrypt(n)) = m + n \quad (1)$$

$$Decrypt(Encrypt(m) \otimes Encrypt(n)) = m \times n \quad (2)$$

暗号文同士の加算が可能であるという特徴を表す (1) を加法準同型性といい，暗号文同士の乗算が可能であるという特徴を表す (2) を乗法準同型性という．準同型暗号には多くの種類が存在し，加法準同型性のみを持つ加法準同型暗号には Paillier 暗号 [9]，乗法準同型性のみを持つ乗法準同型暗号には RSA 暗号 [10] が挙げられる．

また，加法準同型性と乗法準同型性の両方を満たす準同型暗号は演算可能回数によって 3 つに分類できる．演算可能回数に制限がある場合は Somewhat Homomorphyc Encryption(以下 SHE) といい, パラメータにより演算可能回数が決まる場合は Leveled Homomorphic Encryption(以下 LHE) という [11]．演算可能回数に制限がない準同型暗号は FHE という．ブートストラップを行い, 演算による誤差の蓄積をリセットすることにより, 加算と乗算の任意回数の演算を実現している．FHE は，1978 年に Rivest ら [12] によって構想され，2009 年に Craig Gentry ら [13] によって格子暗号を用いた実現手法が提案された．

SHE は LHE よりも演算にかかる時間は短いが乗算可能回数に制限があり，LHE は演算可能回数をあらかじめ設定できるが，演算可能回数の増加と演算にかかる時間は比例関係にあるという課題がある．FHE には暗号文同士の加算と乗算が任意回数可能であるが，暗号文のサイズが大きく，クライアントとサーバ間の通信量が大きくなってしまうといった課題がある [14]．本研究では LHE を使用する．

3.1.2 暗号化方式 [15] とライブラリ

準同型暗号は，第 1 世代 イデアル格子による方式，第 2 世代 BGV [16] や BFV 方式 [17]，第 3 世代 GSW 方式や TFHE [18] [19], 第 4 世代 CKKS 方式 [20] の 4 つの世代に分類

される．特に，CKKS 方式は整数だけではなく実数や複素数も計算することができるという特徴を持ち，機械学習に応用可能であることから近年注目されている．スケールというパラメータによって実数を整数に見立てて計算するため，計算結果に若干のノイズが生じ，計算を繰り返すとノイズが増加するため，リスケール処理を行う．

また，現在では多くの準同型暗号ライブラリが公開されており，有名なライブラリの一つとして，Microsoft SEAL [21] がある．SEAL は Microsoft が開発したライブラリで，BFV 方式と CKKS 方式の 2 つを利用でき，C++で実装されている．他にも，BGV 方式，C++で実装され，ブートストラップもサポートしている HElib [22] も有名である．

3.2 圧縮センシング [23]

3.2.1 概　　要

圧縮センシング (compressed sensing) は，スパース性 (零成分が多いという性質) を持つ高次元の信号を少ない観測から復元する枠組みである．未知の n 次元ベクトル $\boldsymbol{x_0}$ が，m 行 n 列の行列 A を用いて以下のように線形変換されているとする．このとき，$m < n$ とする．

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1n} \\ \vdots & \ddots & & & \vdots \\ a_{i1} & & a_{ii} & & a_{2n} \\ \vdots & & & \ddots & \vdots \\ a_{m1} & \cdots & a_{mi} & \cdots & a_{mn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\therefore \boldsymbol{y} = A\boldsymbol{x_0} \tag{3}$$

元のベクトル $\boldsymbol{x_0}$ の零ではない要素が k 個あると仮定して，低次元な m 次元ベクトル $\boldsymbol{y}$ と圧縮行列 A から (3) を満たす $\boldsymbol{x_0}$ を求める．一般には，未知数の数が方程式の数より多いため不定となり，解を一意に定めることができない．しかし，$\boldsymbol{x_0}$ の要素の多くが零であったとすると，解を一意に推定できる．

3.2.2 再構成アルゴリズム - OMP

圧縮センシングに使用される再構成アルゴリズムは多数存在し，代表的なアルゴリズムには，OMP(Orthogonal Matching Pursuit), IHT(Iterative Hard Tresholding), IST(Iterative Soft Tresholding) などがある．ここでは OMP のアルゴリズムを紹介する．

OMP とは貪欲法の一種であり，$\boldsymbol{y}$ を列ベクトル a_i の成分のみで近似する．残差を最も良く近似する順に A の列ベクトルの添字を添字集合 T に加えていくことで復元する．OMP の詳細は以下の通りである．

入力：観測信号 $\boldsymbol{y} \in \mathbb{R}^m$, 観測行列 $A \in \mathbb{R}^{m \times n}$
(非零要素数 $k \in \mathbb{N}$)
初期化：推定値 $\hat{\boldsymbol{x}} = \boldsymbol{0}$, 添字集合 $T = \phi$
終了条件を満たすまで以下を繰り返す.
（1） 残差 $\boldsymbol{y} - A\hat{\boldsymbol{x}}$ と最も相関の高い (内積の絶対値の大きい)A の列ベクトルの添字を T に追加する.
（2） $\hat{\boldsymbol{x}} = \underset{\hat{\boldsymbol{x}}}{\operatorname{argmin}} ||\boldsymbol{y} - A_T\boldsymbol{x_T}||_2^2$ を一般逆行列で評価することによって $|T|$ スパースな推定値 $\hat{x}$ を求める (一般逆行列で評価する).
出力：推定値 $\hat{\boldsymbol{x}} \in \mathbb{R}^n$

「残差を最も良く近似する」というのは,「内積の絶対値が最も大きい」と言い換えられる．また，$\underset{\hat{\boldsymbol{x}}}{\operatorname{argmin}} ||\boldsymbol{y} - A_T\boldsymbol{x_T}||_2^2$ の解である $\hat{\boldsymbol{x}}$ は，$\hat{\boldsymbol{x}} = (A_T^\top A_T)^{-1} A_T^\top \boldsymbol{y}$，つまり，$\hat{\boldsymbol{x}} = (A_T$ の一般逆行列) と $\boldsymbol{y}$ の積で求められることが一般に知られている．

準同型暗号は比較演算が不得意であることや，シンプルなアルゴリズムではない場合準同型演算が増えると同時に，演算可能回数を表すパラメータの Level も大きく設定する必要がある．したがって復元のアルゴリズムは慎重に選択しなくてはならない．本研究では一般逆行列による L_2 ノルムの最小化によって圧縮前のデータを推定したが，詳細については 4.3 章で説明する．

3.3 辞 書 学 習

辞書学習とは，最適なスパース変換をデータから学習することである [24]．したがって，圧縮センシングに辞書を用いることによってスパースではないデータに対しても復元可能となる．

圧縮センシングに辞書学習を用いる場合，以下のように用いられる [25].

$$x_0 = Dc \tag{4}$$

$$A' = AD \tag{5}$$

$$\therefore y = Ax_0 = ADc = A'c \tag{6}$$

式 (3) のように表されているとするとき，辞書学習による圧縮センシングでは，辞書（n 次元基底ベクトルの集合）D と未知の結合係数 c を用いて，(4) のように表現する．ここで辞書 D は，結合係数がスパースになるように訓練データを用いて予め学習されているものとする．(5) のように行列 A と辞書 D の積を新たな行列 $A' = AD$ と考えると (6) のようになり，$y = A'c$ に対して圧縮センシングによる復元を行うことで結合係数 c を推定することができる．そして (4) より，推定した結合係数 c に辞書 D を掛けることで，x_0 を復元することができる．(6) に対して復元アルゴリズムを適用した場合の詳しい計算は 4.3 章で説明する．

辞書学習のアルゴリズムはいくつか存在し，代表的なものには MOD(Method of Optimal Direction) や K-SVD(K-Singular Value Decomposition) [26] などが挙げられる．本研究では K-SVD による辞書学習を使用した．

4 システム概要

本節では, 提案手法を用いたシステムの流れや，使用した圧縮・復元アルゴリズムについて説明する．

4.1 プロトコル

提案手法を用いたシステムの流れを図 2 と以下に示す.

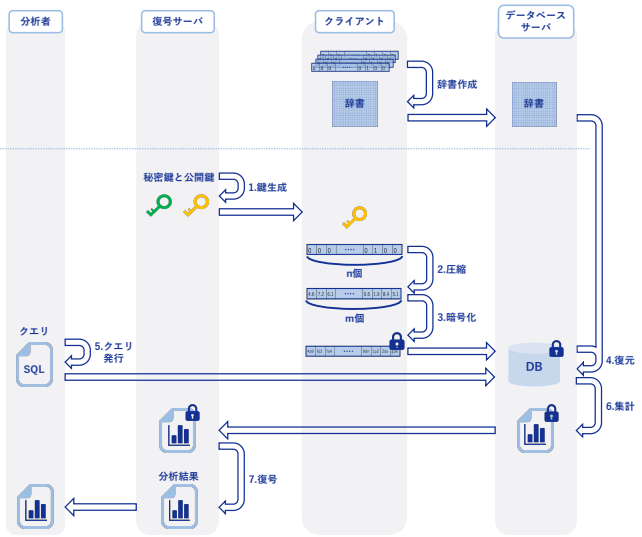


図 2: 提案手法を用いたシステムの流れ

辞書作成：クライアントの一部のデータを辞書学習して辞書を作成し，データベースサーバに送信する．

（1） **鍵生成：**復号サーバが準同型暗号の秘密鍵と公開鍵を生成し，公開鍵をクライアントとデータベースサーバに送信する．

（2） **圧縮：**クライアントが圧縮する．

（3） **暗号化：**クライアントが準同型暗号で平文を暗号化して，データベースサーバへ送信する．

（4） **復元：**サーバが辞書を用いて暗号文を復元する．

（5） **クエリ発行:** 分析者がクエリを投げる．

（6） **集計：**サーバが，復元した暗号文を用いてデータベースサーバ上で暗号状態のまま結果を集計する．

（7） **復号：**復号サーバが分析結果を復号し，分析者が復号された分析結果を得る．

このうちの (2) の圧縮と (4) の復元で圧縮センシングを使用している．次元圧縮してクライアントからデータベースサーバへ送信するため，クライアント・サーバ間の通信量削減が可能となる．(3) 暗号化では，圧縮後のベクトルを複数用意して，一つの暗号文に詰め込むことができる平文のベクトルの長さ ($slot_count$) と同じになるようにする．すなわち，圧縮後のベクトルの長さが m であれば，$slot_count/m$ 個のベクトルを一つの暗号文に詰め込む．$slot_count$ の決定については 5.2 章で説明する．

4.2 圧縮アルゴリズム

本研究ではスパースではないデータに対しても提案手法を行えるように，辞書学習による圧縮センシングを用いて圧縮を行う．辞書を用いない場合，式 (3) のように n 次元のベクトルを m 次元に圧縮するが，辞書を用いる場合には (4)，(5) から (6) のように式を変換するため，(6) における圧縮の式を改めて考える必要がある．ここで，(4) より，$c = D^{-1}x_0$ であり，これを (6) に代入すると，$y = ADD^{-1}x_0 = Ax_0$ となる．したがって，本研究では $y = Ax_0$ のように n 次元のベクトルを m 次元に圧縮することとする．このとき，m 行 n 列の圧縮行列 A はランダムに生成された公開情報とする．

表 3 にあるように, 本研究では一つの暗号文に 4096 の要素をパッキングして扱う．したがって，クライアントはできるだけ一つの暗号文に複数の平文 (ベクトル) を詰めてサーバに送信する．そのため図 5 のように，スパースな n 次元データ $\boldsymbol{x_0}$ を圧縮したベクトル $\boldsymbol{y}$ を，$slot_count \div m$ 個詰めて一つのベクトルにしてパッキングを行う．例えば $slot_count = 4096, m = 128$ の場合，32 個の $\boldsymbol{y}$ が一つの暗号文に含まれる．そして，復元する際は分割して元の m 次元データ $\boldsymbol{y}$ に戻す．

4.3 復元アルゴリズム

準同型暗号は比較演算が不得意である．また，本研究では LHE を使用しており，LHE はパラメータにより演算可能回数が決まるため，シンプルなアルゴリズムではない場合 Level(暗号文同士の乗算可能回数を示す) が大きくなってしまう．これらの点を踏まえ，本研究では一般逆行列による L_2 ノルムの最小化によって圧縮前のデータを推定した．一般逆行列での L_2 ノルムの最小化による復元アルゴリズムでは，一般逆行列によって最小ノルム解，つまり L_2 ノルムが最小となる推定結果を得る．ここで辞書を用いているため，(6) に対して復元アルゴリズムを適用すると，$encrypt(\boldsymbol{c}) = (A'^{\top}A')^{-1}A'^{\top} \times encrypt(\boldsymbol{y})$ となる．これは (4) より，$encrypt(\hat{\boldsymbol{x}}) = D((AD)^{\top}(AD))^{-1}(AD)^{\top} \times encrypt(\boldsymbol{y})$ となる．このようにして，辞書 D と AD の一般逆行列の積と，暗号化された低次元行列との積を求める．

準同型暗号を用いた行列積の実装方法．$slot_count/m$ 個の $\hat{\boldsymbol{x}}$ を一度に復元する方法を図 3 と以下に示す．ここでは，一度に複数の $\hat{\boldsymbol{x}}$ を復元するため，n 行 m 列の $D((AD)^{\top}(AD))^{-1}(AD)^{\top}$ を $slot_count/m$ 個横に繋げて n 行 $slot_count$ 列にしておく．本研究で採用した CKKS 方式では，暗号文のスロットを左方向または右方向に指定したステップ数だけ回転させる rotation 操作が可能であり，復元アルゴリズムの行列積の実装に用いる．

（1） $D((AD)^{\top}(AD))^{-1}(AD)^{\top}$ の i 行目と $encrypt(\boldsymbol{y})$ の積を求める．

（2） (1) を rotation して足し合わせる ($= D((AD)^{\top}(AD))^{-1}(AD)^{\top}$ の i 行目と $encrypt(\boldsymbol{y})$ の内積).

（3） フィルターをかけて，$slot_count/m$ 個の各内積の値を取り出す．

（4） 各内積の値が i 番目に来るように rotation する．

(1)〜(4) を n 回繰り返した合計が $slot_count/m$ 個の $\hat{\boldsymbol{x}}$ となる．

このとき，(2) の rotation によってベクトルの合計を得る場合，単純には一つ要素をずらす rotation が m 回必要になるが，図 4 のように 2^i ずつ要素をずらすと，$\log_2 m$ 回の rotation でベクトルの合計を計算できる．

復元の準同型演算回数．ここでは，本研究の復元アルゴリズムの rotation と準同型乗算の回数について分析する．圧縮前のベク

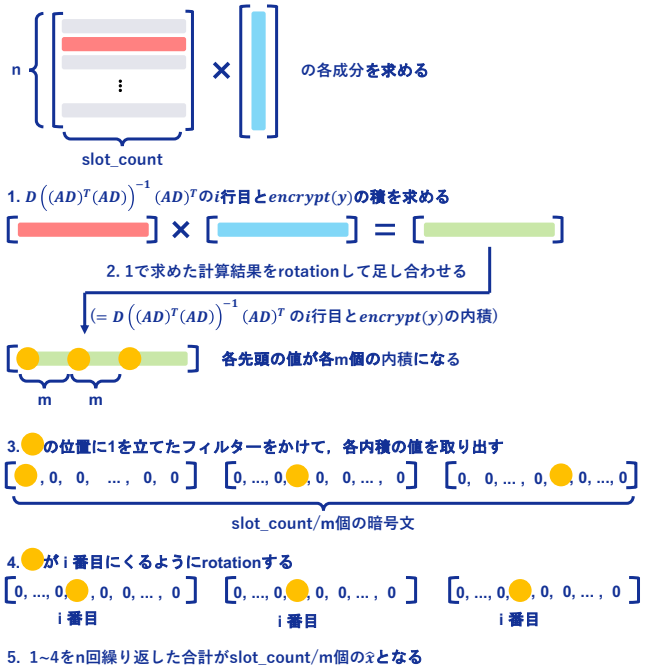


図 3: 準同型暗号を用いた行列積の実装方法.

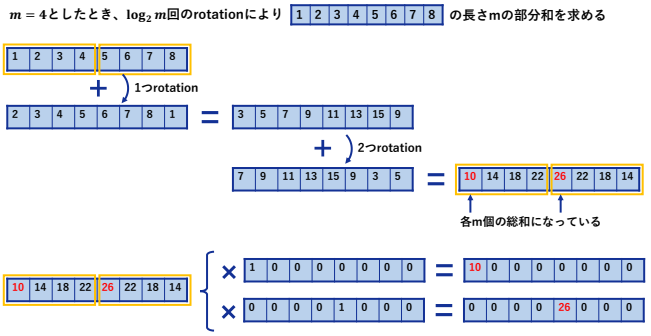


図 4: $\log_2 m$ 回の rotation により長さ m の部分列和を求める方法.

トルの長さを $n = 2m$，すなわち圧縮率を 50%とする．復元アルゴリズムの (2)，(4) の rotation 呼び出し回数を $num_rotation$ とすると，m に比例することが分かる．

$$\begin{aligned} num_rotation &= n(\log_2 m + slot_count/m) \\ &= 2m(\log_2 m + slot_count/m) \\ &= 2m\log_2 m + 2slot_count \end{aligned}$$

また，復元アルゴリズムの (1)，(3) に含まれる平文と暗号文の乗算回数を num_mult とすると，$num_mult = n(1 + slot_count/m) = 2m + 2slot_count$ 回であり，m に比例する．よって m を小さくする，すなわち短い $\hat{x}$ を複数回復元するほうが準同型演算の回数は少なくすることができる．

5 評　　価

本節では，提案手法の評価について説明する．

5.1 実 験 概 要

実験ではクライアントには Raspberry Pi，サーバには

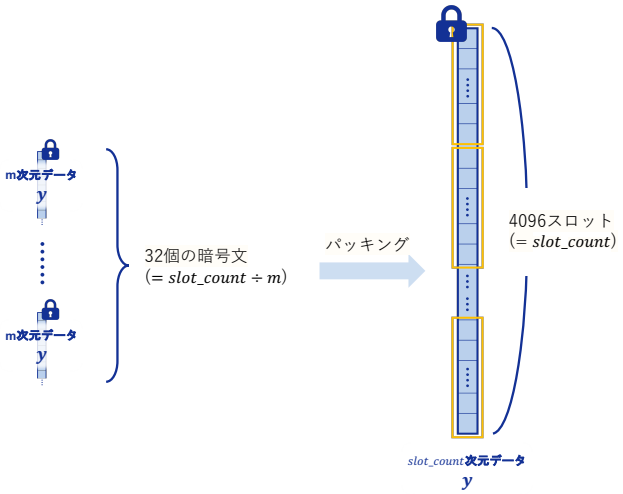


図 5: パッキングの様子.

Ubuntu のマシンを使用し，Microsoft SEAL ライブラリを使用して Baseline と提案手法を C++で実装した．Baseline，提案手法ともに POS データのデータセットを利用した．一部 POS データを使用した K-SVD による辞書学習の実装はこちら [27] を参考にしてコードを実装した．また，評価指標を以下に示す．

評価指標.

- クライアントからサーバに送信する準同型暗号文のサイズ
- 圧縮センシングの再構成アルゴリズムによるサーバでの復元結果/カウントクエリの精度
- クライアントの実行時間とサーバの復元の実行時間

5.2 実 験 設 定

データセット.　実験ではスーパーマーケットの POS データ [28] を使用した．POS データは表 1 のように加工し，合計が n の 256bit になるように設定した．具体的には以下のような加工を行った．BeginDateTime を BeginYear, BeginMonth, BeginDay, BeginHour, BeginMinute, BeginSecond に分割し，EndDateTime を EndYear, EndMonth, EndDay, EndHour, EndMinute, EndSecond に分割した．TNcash, TNcard はそれぞれ TRUE を 1，FALSE を 0 に置き換えた．Amount は四捨五入して整数にし，負の値は 0 に置き換えた．このように加工した POS データの各カラムの値を 2 進数に変換したデータを使用した．ただし，2 進数にしたとき足りない桁数部分は 0 でパディングした．

実験環境.　実験に使用したマシンの性能を表 2 に示す．クライアントには IoT デバイスとして Raspberry Pi を使用し，サーバには Ubuntu のマシンを使用した．

準同型暗号に関するパラメータ [29].　実験で使用した SEAL の CKKS 方式のパラメータを表 3 に示す．暗号文の長さ $slot_count$ は使用する多項式の次元を示す $poly_modulus_degree$ に 0.5 をかけた値となる．ここでは，$poly_modulus_degree = 8192$ と

表 1: POS データのデータ加工方法

WorkstationGroupID	サービスかセルフサービスか	6 bit
TranID	トランザクションID (日付, ストアID, Checkout ID, シーケンス番号)	55 bit
BeginYear	取引開始時の年	12 bit
BeginMonth	取引開始時の月	4 bit
BeginDay	取引開始時の日	5 bit
BeginHour	取引開始時の時	5 bit
BeginMinute	取引開始時の分	6 bit
BeginSecond	取引開始時の秒	6 bit
EndYear	取引終了時の年	12 bit
EndMonth	取引終了時の月	4 bit
EndDay	取引終了時の日	5 bit
EndHour	取引終了時の時	5 bit
EndMinute	取引終了時の分	6 bit
EndSecond	取引終了時の秒	6 bit
OperatorID	一意のレジID	15 bit
TranTime	取引にかかった時間 [秒]	15 bit
BreakTime	休憩時間 [秒]	15 bit
ArtNum	購入したアイテムの数	15 bit
TNcash	現金支払いかどうか	1 bit
TNcard	カード支払いかどうか	1 bit
AmountGroup	取引額	57 bit

表 2: マシン性能

	Raspberry Pi	サーバ
OS	Raspbian 11	Ubuntu 22.04.1
CPU	ARM Cortex-A72	Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz
コア数	4	10
プロセッサ速度	600MHz	1GHz
RAM	4GB	192GB

したため，$slot_count = 4096$ となる．また，暗号文同士の乗算可能回数を示す Leveled や復号時の精度，復号時の実数値の精度は，チェインの設定によって決まる．実験ではチェインは次のように設定した．

$$modulus_chain = \{60, 40, 40, 60\} \tag{7}$$

(7) は最初の数値 (以下 プライマリビット) である 60，最後の数値 (以下 ラストビット) である 60, それ以外の (以下 スケーリングビット)2 つの 40 に分けることができる．Leveled はスケーリングビットの個数に依存するため，$Leveled = 2$ となり，2 回乗算可能となる．また，プライマリビットは復号時の精度と対応しており，復号時の実数値の整数部分の精度はプライマリビットとスケーリングビットの差と，復号時の実数値の小数部分の精度はスケーリングビットと復号時の実数値の整数部分の精度ビットの差と対応している．

表 3: SEAL の CKKS 方式に関するパラメータ

	値
暗号文の長さ $slot_count$	4096
セキュリティ	128 bit
Leveled	2
復号時の精度	60 bit
復号時の実数値の精度	整数部分　20 bit 小数部分　20 bit

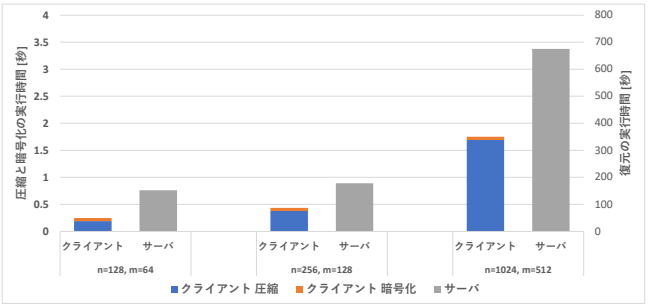


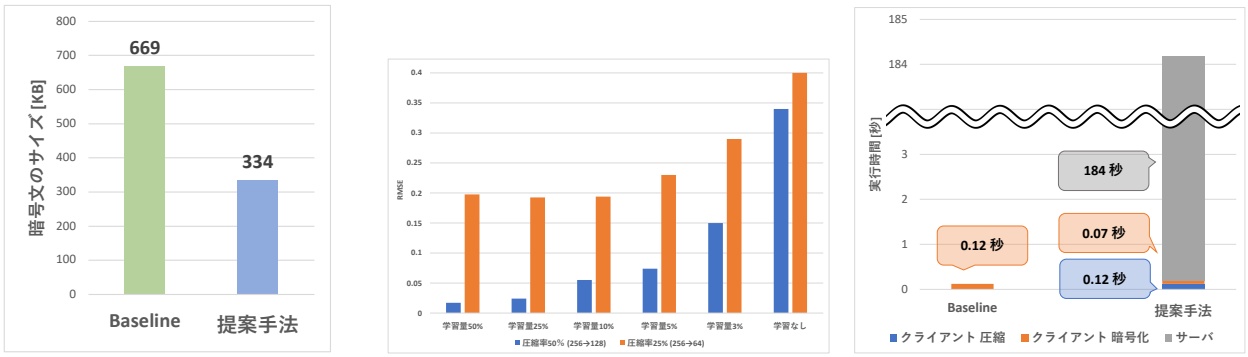
図 6: 圧縮率 50%時の実行時間 (辞書未使用時).

圧縮センシングに関するパラメータ． 実験での圧縮センシングのパラメータは，図 6 に示した圧縮率 50 %で n, m の大きさを変化させた時の実行時間, 圧縮率と精度がトレードオフの関係にあることを参考に決定した．図 6 では，扱うデータの合計が同じにもかかわらず，n と m の大きさが小さいほど実行時間が短くなっている．これは圧縮行列 A が $n \times m$ であり，n, m の大きさと行列計算の時間には正の関係があるということと，4.3 節でも説明したように，復元アルゴリズムの準同型演算の回数が m に比例するためであると考えられる．よって長いベクトルを復元するよりも短いベクトルを複数復元する方が高速であり適しているといえる．また，辞書学習をする際に与えるデータからパッチを抽出する際に平方数である必要がある．したがって $n = 256, m = 128$ とした．

5.3 通信量についての評価

まず圧縮センシングによる圧縮の効果を評価する前に，Zip 形式に圧縮した場合にどのくらいサイズが変わるかを評価する．暗号化した $\boldsymbol{y}$ を Zip 形式に圧縮した場合，圧縮前は 334.447KB だったのに対し，圧縮後は 332.398KB となり，ほとんどサイズ減少は見られなかった．この理由は，Zip 形式による圧縮は同じ文字が何回も出てくるものが圧縮されやすいからであると考えられる．準同型暗号文はそのような特徴を持たないため，Zip 形式による圧縮は向いていない．したがって，暗号化する前の平文の状態で圧縮する必要がある．

図 7(a) にクライアントとサーバ間の通信量，つまり，クライアントからサーバに送信する準同型暗号文のサイズを示す．Baseline に比べ提案手法では約半分の通信量に抑えられている．これにより，従来の Baseline での通信コストで 2 倍のデー



(a) 圧縮の有無と通信量　(b) 圧縮率/学習量と RMSE の関係　(c) 圧縮の有無と実行時間

図 7: (a) 縦軸は圧縮率 50 %の場合のクライアントとサーバ間の暗号文のサイズを表す．提案手法では圧縮を行わない Baseline に比べて約半分の通信量に抑えられており，通信量の削減比率は n と m の比率に対応している．(b) 圧縮率を高める/学習量が減るほど圧縮前と復元後の誤差は大きくなっており，圧縮率のほうが精度に与える影響が大きい．(c) 圧縮率 50 % ($n = 256, m = 128$) のときの Baseline と提案手法での実行時間は，クライアントからサーバに送信する暗号文のサイズの減少に伴い暗号化にかかる時間は減っている．一方で，サーバで復元する際の再構成アルゴリズムの計算に時間がかかっている．

タ量をクライアントからサーバに送信することが可能となる．Baseline と提案手法の通信量の比率は n と m の比率に対応しており，n と m の比率が一定であれば大きさは関係ない．例えば，$n = 1024, m = 512$ にしたときと $n = 128, n = 64$ にしたときの通信量削減量は同等である．したがって，n/m の値が大きくなるようなパラメータにすることで通信量をより少なくすることができると考えられる．しかし，図 7(b) に示したように，n/m の値が大きくなる，つまり圧縮率を大きくすると精度が下がってしまう．したがって，どの程度精度が必要で，どの程度通信量を少なくする必要があるのかによって，n, m のパラメータを選択する必要がある．

5.4　復元精度についての評価

精度について，Baseline と提案手法での圧縮センシングの再構成アルゴリズムによるサーバでの復元結果，それらの復元結果に対してカウントクエリを実行した結果を RMSE を用いて評価する．式 (8) 中の x_i は圧縮前の値，$\hat{x_i}$ は復元後の値，n はデータの総数を表す．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(x_i - \hat{x_i})^2} \qquad (8)$$

図 7(b) に圧縮率/学習量とサーバでの復元結果の RMSE との関係性を示す．圧縮率が低い/学習量が多いほど精度が高いことがわかる．そして，学習データが多くても圧縮率が高いと精度が下がる，即ち圧縮率のほうが精度に与える影響が大きい．今回の実験環境である圧縮率 50 % ($n = 256, m = 128$), 学習量 25 %としたときの 256 個の $\hat{\boldsymbol{x}}$ の RMSE は，おおよそ 0.025 となった．このとき，辞書を使用しないと 256 個の $\hat{\boldsymbol{x}}$ の RMSE はおおよそ 0.34 となる．このことから，辞書によりかなり精度を高められることがわかる．ここで，圧縮率 50 %，25 %ともに学習量が増えるほど精度は高まっているが，圧縮率 25 %のとき，学習量 50 %，25 %，10 %で RMSE 値がほとんど変化していない．現時点では正確な理由は分かっていないが，圧縮率が高いと情報量が少ない状態から復元するため，復元時の辞書による効果が少なくなるからではないかと考えている．

また，圧縮率 50%において，Baseline, 提案手法で各カラムごとにカウントクエリを実行した結果を図 8 に示す．カウントクエリの計算は Folklore [8] という演算方法で計算した．図 8 内の 5 つのクエリは，それぞれ 2 つの条件を AND で組み合わせたクエリである．具体的には以下のクエリを実行した．

- Q1：有人レジでカード支払いした人の数
  SELECT COUNT(*) FROM T
  WHERE WorkstationGroupID = 1 AND TNcard = 1
- Q2：2019 年に現金支払いした人の数
  SELECT COUNT(*) FROM T
  WHERE BeginYear = 2019 AND TNcash = 1
- Q3：4 月に 7 つの商品を購入した人の数
  SELECT COUNT(*) FROM T
  WHERE BeginMonth = 4 AND ArtNum = 7
- Q4：15 時台にある特定のレジ (ID=131) でお会計をした人の数
  SELECT COUNT(*) FROM T
  WHERE BeginHour = 15 AND OperatorID = 131
- Q5：2019 年 4 月にお会計をした人の数
  SELECT COUNT(*) FROM T
  WHERE BeginYear = 2019 AND BeginMonth = 4

このときの 5 つそれぞれのカウントクエリの RMSE の平均は，学習量 50%で約 0.2, 学習量 25%で約 0.3, 学習量 10%で約 0.7, 学習量 5%で約 1.2, 学習量 3%で約 5.4 であり，学習量を増やすほど精度が高くなることがわかる．

5.5　実行時間についての評価

Baseline と提案手法における実行時間を図 7(c) に示す．

クライアントによる圧縮と暗号化．　提案手法ではクライアントからサーバに送信する暗号文のサイズの減少に伴い暗号化に

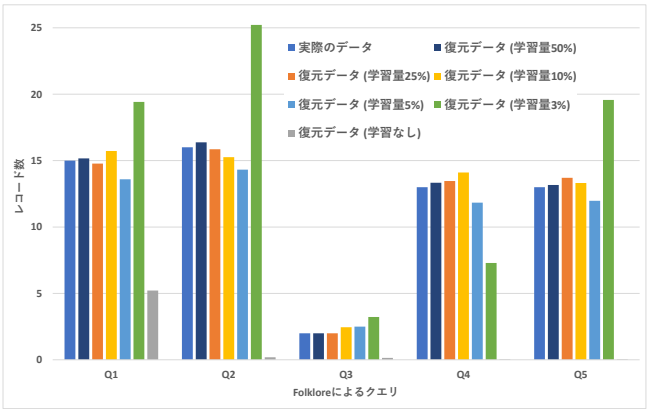


図 8: 圧縮率 50%において Baseline, 提案手法でカウントクエリを実行した結果.

かかる時間は減っている．クライアント全体での実行時間 (クライアントでの圧縮と暗号化) を見ると，$n = 256, m = 128$ では Baseline:0.12 秒，提案手法:0.19 秒であり，圧縮する場合と圧縮しない場合でも実行時間はそこまで変わらないといえる．図 6 に示したように，n, m の大きさを大きくするほど，トータルの平文の長さは同じでも，圧縮に要する時間が長くなる．圧縮する際の行列計算では $m \times n$ の行列の積を求めている．したがって図 6 からも，n, m の大きさを大きくするほど，今回の $n = 256, m = 128$ で測定した結果よりも圧縮に要する時間が長くなるといえる．

サーバによる復元. サーバでの復元では 184 秒かかっており，サーバで復元する際の再構成アルゴリズムの計算に時間がかかってしまっていることが分かる．こちらも復元する際には $encrypt(\hat{\boldsymbol{x}}) = D((AD)^\top (AD))^{-1}(AD)^\top \times encrypt(\boldsymbol{y})$ のように辞書 D と AD の一般逆行列の積と，暗号化された低次元行列との積を求めている．したがって図 6 からも，n, m の大きさを大きくするほど，今回の $n = 256, m = 128$ で測定した結果よりも復元に要する時間が長くなるといえる．

6 結　　論

準同型暗号の応用例である暗号化データベースでは，データ分析者に集計結果のみを知らせることができるという利点があるが，暗号文は平文よりもサイズが大きいため，クライアントとサーバ間の通信量が大きくなってしまう．そこで本研究では辞書を用いた圧縮センシングによる通信量削減を提案した．

実験の評価の結果，圧縮率を 50%に設定した場合は提案手法は圧縮せずに送信した場合に比べて 50%通信量を削減できることを示した．精度については，復元前と復元後のベクトル $\hat{\boldsymbol{x}}$ の誤差，POS データを収集した場合のカウントクエリの誤差を RMSE で評価した．今後は精度や実行時間をより改善していきたい．

謝　　辞

本研究は一部，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Nikola Samardzic, Axel Feldmann, Aleksandar Krastev, Srinivas Devadas, Ronald Dreslinski, Christopher Peikert, and Daniel Sanchez. F1: A fast and programmable accelerator for fully homomorphic encryption. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 238–252, 2021.

[2] Sangpyo Kim, Jongmin Kim, Michael Jaemin Kim, Wonkyung Jung, John Kim, Minsoo Rhu, and Jung Ho Ahn. Bts: An accelerator for bootstrappable fully homomorphic encryption. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 711–725, 2022.

[3] M Sadegh Riazi, Kim Laine, Blake Pelton, and Wei Dai. Heax: An architecture for computing on encrypted data. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 1295–1309, 2020.

[4] Marin Matsumoto and Masato Oguchi. IoT Device Friendly Leveled Homomorphic Encryption Protocols. In *2022 IEEE International Conferences on Smart Data (SmartData)*, pp. 525–532. IEEE, 2022.

[5] Amrita Roy Chowdhury, Chenghong Wang, Xi He, Ashwin Machanavajjhala, and Somesh Jha. Crypte: Crypto-assisted differential privacy on untrusted servers. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 603–619, 2020.

[6] Xuanle Ren, Le Su, Zhen Gu, Sheng Wang, Feifei Li, Yuan Xie, Song Bian, Chao Li, and Fan Zhang. Heda: Multi-attribute unbounded aggregation over homomorphically encrypted database. *Proceedings of the VLDB Endowment*, Vol. 16, No. 4, pp. 601–614, 2022.

[7] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pp. 1–12. Springer, 2006.

[8] Rasoul Akhavan Mahdavi and Florian Kerschbaum. Constant-weight PIR: Single-round keyword PIR via constant-weight equality operators. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1723–1740, 2022.

[9] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pp. 223–238. Springer, 1999.

[10] Ronald L Rivest, Adi Shamir, and Leonard Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, Vol. 21, No. 2, pp. 120–126, 1978.

[11] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, Vol. 51, No. 4, pp. 1–35, 2018.

[12] Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. On data banks and privacy homomorphisms. *Foundations of secure computation*, Vol. 4, No. 11, pp. 169–180, 1978.

[13] Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 169–178, 2009.

[14] 佐藤宏樹, 馬屋原昂, 石巻優, 今林広樹, 山名早人. 完全準同型暗号のデータマイニングへの利用に関する研究動向. 第 15 回情報科学技術フォーラム F-002, 2016.

[15] 準同型暗号の世代区分. `https://www.eaglys.co.jp/news/column/homomorphicencryption/`.

[16] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, Vol. 6, No. 3, pp. 1–36, 2014.

[17] Zvika Brakerski. Fully homomorphic encryption without modulus switching from classical gapsvp. In *Advances in Cryptology–CRYPTO 2012: 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*, pp. 868–886. Springer, 2012.

[18] Ilaria Chillotti, Nicolas Gama, Mariya Georgieva, and Malika Izabachène. TFHE: fast fully homomorphic encryption over the torus. *Journal of Cryptology*, Vol. 33, No. 1, pp. 34–91, 2020.

[19] TFHE. https://tfhe.github.io/tfhe/.

[20] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yongsoo Song. Homomorphic encryption for arithmetic of approximate numbers. In *International conference on the theory and application of cryptology and information security*, pp. 409–437. Springer, 2017.

[21] Microsoft SEAL. https://github.com/microsoft/SEAL.

[22] HElib. https://github.com/homenc/HElib.

[23] 圧縮センシング: 疎情報の再構成とそのアルゴリズム. https://www.kurims.kyoto-u.ac.jp/~kyodo/kokyuroku/contents/pdf/1803-03.pdf.

[24] 中尾 愛寺田 康彦, 中込真優. 圧縮センシング mri の基礎, 2018.

[25] 岡部 孝弘小林 直人. 辞書学習を用いた圧縮センシングによる多波長・多方向光源下画像の効率的な獲得, 2015.

[26] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, Vol. 54, No. 11, pp. 4311–4322, 2006.

[27] K-svd の実装. https://github.com/takeshan/KSVD/tree/master.

[28] Point of sale (POS) data from a supermarket: Transactions and cashier operations. https://www.mdpi.com/2306-5729/4/2/67.

[29] 格子暗号の CKKS 形式のパラメータ、精度ビットなどの解説（SEAL ライブラリ）. https://qiita.com/kenmaro/items/931835476ca8781de9f9.

準同型暗号と隔離実行環境の組合せによるプライバシを保護した BERT 文章分類手法

橋本 太志† 大西 隆太郎‡ 鈴木 拓也‡ 山名 早人§

†早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: †‡§{hashimotof, ryutaro634, t-suzuki, yamana}@yama.info.waseda.ac.jp

あらまし 現在，BERT を用いた文章分類はスパム検出といった場面で活用されている．しかし，外部のサーバを用いる場合では，入力データに含まれるプライバシ情報の漏洩が問題となる．そこで，プライバシ保護のために準同型暗号の適用が考えられる．しかし，準同型暗号を適用した従来手法では，準同型暗号上では実行できない処理を多項式近似するか，ユーザ側で実行しなければならず，精度低下やユーザが担う計算コストの増加を招く．本稿では，これらの問題を解決するために，準同型暗号上では実行できない処理を隔離実行環境内で平文で実行する．これにより，プライバシを保護した BERT による文章分類手法を実現する．評価実験の結果，机上計算で算出したベースライン手法と比べ，提案手法は約 53 倍の高速化を達成した．

キーワード 準同型暗号，プライバシ保護機械学習，BERT，文章分類，隔離実行環境

1. はじめに

近年，自然言語処理を対象とした深層学習の一つである BERT[1]による文章分類が，スパム検出やニュース分類で活用されている．こうした処理は，一般に，BERT の「モデル提供者」，BERT を実行する「サーバ」，利用者であり BERT への入力をサーバへ送る「ユーザ」の 3 パーティから構成される．

しかし，サーバ内で扱われる文章データやモデルは暗号化されておらず，サーバ上での文章データや学習モデル自体（具体的には，モデルの構成とモデル内の重み）の機密漏洩が問題となっている．そこで，本稿では，BERT をクラウド上で動作させる場合の，入力データとモデルの漏洩を防ぐことを目的とする．

こうした問題を解決する手段として，データを暗号化したまま復号することなく演算できる準同型暗号（HE: Homomorphic Encryption）[2]やハードウェア技術で隔離された環境下でデータやプログラムの機密性と完全性を保護する隔離実行環境（TEE: Trusted Execution Environment）[3]の活用が挙げられる．しかし，HE は加算や乗算以外の演算が評価できない問題や実行時間が平文での実行と比べて 10^6 倍以上必要となる実利用上の問題がある．一方，TEE にはサイドチャネル攻撃によって機密性が損なわれる危険性がある．

そこで，本稿では，上記の問題を解決するために，BERT に対して HE と TEE を組み合わせ，モデルとデータを秘匿する手法を提案する．具体的には，1）モデルパラメータとユーザからの入力データを HE で暗号化して扱い機密性を担保する、2）HE で実現できない処理を TEE 内で一旦復号して実行する．ただし，TEE へのサイドチャネル攻撃への対応のため，TEE 内での復号を最小限に抑える手法をとる．これにより，従来問題となっていた多項式近似による計算誤差の問題を解決すると共に，ユーザ側で行っていた処理を削減することができることからユーザ・サーバ間での途中結果の通信量を削減できる．

本稿の貢献を以下に示す．

- HE を適用した大規模言語モデルをサーバ上で実行し，入力データとモデルの重みの保護を実現すること．
- HE に TEE を組み合わせることで，BERT における活性化関数を平文で実行し，正解率を維持しながらユーザの負担を減らすこと．

本稿は以下の構成である．2 節で HE と TEE，BERT について説明し，3 節で BERT に HE を適用した関連研究についてまとめる．そして，4 節で提案手法を説明する．その後，5 節で実験方法を示し，6 節で実験結果と考察を述べる．最後に，7 節でまとめを述べる．

2. 背景知識

2.1 準同型暗号

準同型暗号（HE: Homomorphic Encryption）は，暗号化されたデータに対し，復号することなく加算（準同型加算）や乗算（準同型乗算）を評価できる暗号技術である．2009 年に Gentry[2]によって提案された，格子暗号ベースの完全準同型暗号は，任意回の準同型加算と準同型乗算を評価できる．格子暗号ベースの暗号方式では，解読を困難にするためにノイズが付加されており，このノイズは準同型演算が行われるごとに増加

する．そして，暗号文に含まれるノイズ量が許容量を超えると正しく復号できない．そこで，Gentry が提案した Bootstrapping と呼ばれるノイズ削減手法をノイズ量が許容値を超える前に暗号文に適用する必要がある．また，暗号文のレベルは，その暗号文に対して適用可能な準同型乗算回数を表し，値が大きいほどその暗号文を入力とする準同型演算のレイテンシが長い．そして Bootstrapping によりレベルをリセットできる．しかし，Bootstrapping は他の準同型演算よりも実行時間が長い（数秒）ため，Bootstrapping の多用はアプリケーション全体の実行時間を増大させる．

本稿では，機械学習の推論に HE を適用するために，完全準同型暗号の中でも固定小数点数を扱える Cheon-Kim-Kim-Song（CKKS）方式[4]を採用する．CKKS 方式はパッキング[5]と呼ばれる，複数の値を一つの暗号文に収めることができ，これらの値に一斉に同一の演算を実行（SIMD 計算）できる．また，CKKS 方式が対応している rotation は，暗号文にパッキングされた値の位置（スロット）を移動させる．本稿では，暗号文Xを$step$分 rotation することを rotation(X, $step$)と表す．図 1 に rotation(X, 1)を実行したときの位置の変化を示す．ただし，暗号文Xはデータの要素数が 8,192 のベクトルをx，それを暗号化したものである．また，図 1 における四角は暗号文のスロットを表す．

入力	x[1]	x[2]	x[3]	…	x[8191]	x[8192]
出力	x[2]	x[3]	…	x[8191]	x[8192]	x[1]

図 1　Rotation(X, 1)に対する入出力の対応

HE には以下に示す欠点が存在する．

- 平文と比較し，実行時間が約$10–10^7$倍増大する[6].
- 加算や乗算以外の，条件分岐や平方根といった演算を評価できない．
- 暗号文のサイズが平文と比較して約$10–10^6$倍大きく[6]，メモリ消費量が増える．
- 計算結果を書き換えられたことを検出できない（完全性が保証されない）．

2.2 隔離実行環境

本稿では，隔離実行環境（TEE：Trusted Execution Environment）の一種である Intel SGX[7]を用いる．Intel SGX は，メモリ上に保護された領域を生成し，その内部でデータやプログラムを暗号化して格納することで，格納されたデータやプログラムを OS から保護しながらプロセスを実行する技術である．

Intel SGX では，メモリ上にエンクレーブと呼ばれる保護領域を生成する．エンクレーブ内ではデータとプログラムが暗号化されたままメモリに保存され，プログラムの実行時にはデータとプログラムが CPU 内の専用回路で復号され実行される．したがって，HE と比較して計算コストが低い．

Intel SGX では，エンクレーブが安全に生成され，意図したプログラムがロードされているかを判別するリモートアテステーションと呼ばれる機能が提供されている．上記の機能により，エンクレーブ内のデータとプログラムについての完全性が保証される．

しかし，Intel SGX は，コンピュータの物理的特性から情報を盗むサイドチャネル攻撃を受ける危険性があり[8]，機密性の観点で問題が残る．

2.3 BERT

2018 年に Google が大規模言語モデルの一つである BERT を提案した[1]. BERT は，文脈理解を達成したことで，検索エンジンやニュース分類，チャットボットといった用途に活用されている．BERT は，2017 年に発表された Transformer[9] という自然言語処理モデルのエンコーダ部分をベースにしており，教師無し学習によって学習された多くの事前学習モデルが公開されている．これにより，事前学習に要するデータと比較して少数の教師ありデータで軽量な追加学習を行うことで，高精度のモデルを構築できる．

BERT の構造を図 2 (a)に示す．まず one-hot ベクトル化された入力に Embedding を行う．Embedding は，行列積（Matmul）に Layer Normalization を行う処理である．その後，Multi-Head Attention, Linear, Layer Normalization, Feed Forward Network を N 回（N はモデルサイズに依存）繰り返す．図 2 (b)に示すように，Linear は Matmul と重みの加算で構成される．

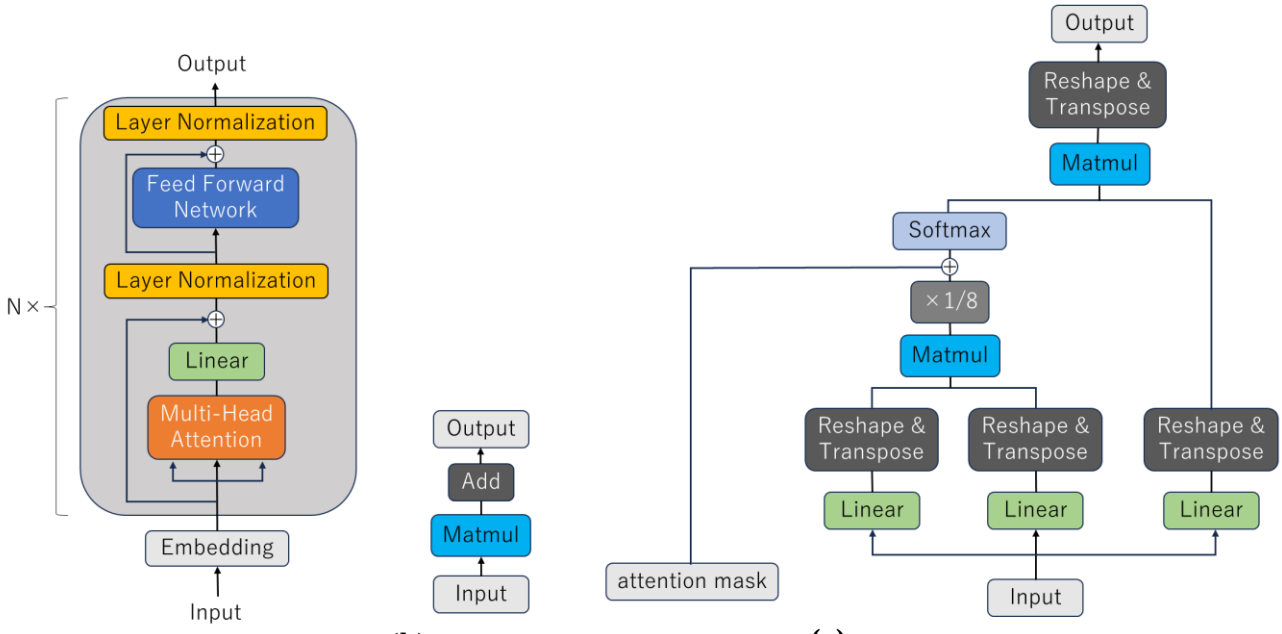
Multi-Head Attention の構造を図 2 (c)に示す．Multi-Head Attention は，Linear と Softmax 関数，Matmul，Reshape&Transpose（例えば，サイズが[1, 150,128]の 3 次元行列をサイズが[1, 64, 150, 2]の 4 次元行列に変換する）で構成される．n個の要素を持つベクトルxに softmax 関数を適用する時，i番目の値y_iは次の式(1)で表される．

$$y_i = \frac{exp(x_i)}{\sum_{k=1}^{n} exp(x_k)} \tag{1}$$

n個の要素を持つベクトルxに Layer Normalization を適用する時，i番目の値z_iは以下の式(2)で表される．ここで，μは要素平均を，Var(·)は分散を，εはモデルのパラメータを，γ及びβは重みをそれぞれ表す．

$$z_i = \gamma \frac{(x_i - \mu)}{\sqrt{\mathrm{Var}(x_i) + \varepsilon}} + \beta \tag{2}$$

Feed Forward Network は式(3)に示す GELU 関数[10]が Linear に挟まれる形で構成される．ここで，erf(x)は式(4)で表される．また，GELU 関数は，シグモイド関数$\sigma(\cdot)$を用いて式(5)と近似することができる．



$$\mathrm{GELU}(x) = \frac{x}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \tag{3}$$

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{4}$$

$$\mathrm{GELU}(x) \fallingdotseq x\sigma(1.702x) \tag{5}$$

3. 関連研究

本節では BERT に HE を適用した研究について紹介する．機械学習モデルに HE を適用する際の問題を以下に示す．

- 実行時間が増大する（準同型演算単体としての処理時間大，bootstrapping によるオーバヘッド）
- 一部の処理が HE 上で評価できない（HE が評価できるのは加算と乗算のみ）

3.1 BERT に接続する線形回帰モデルの処理のみに準同型演算を適用する方式

2022 年に，Garam ら[11]は，自然言語処理モデルの一つである Sentence BERT[12]に線形回帰モデルを接続した文章分類モデルに HE を適用した．ユーザが Sentence BERT を平文上で実行し，その出力を暗号化した上でサーバに送信する．サーバは HE 上で線形回帰モデルの推論を行う．しかし，Garam らの手法は，BERT 処理の全てをユーザが実行するため，ユーザの負担が大きい．

3.2 一部処理をユーザ側で実行する方式

2022 年に，Chen ら[13]は，BERT_tiny [14]の推論処理に HE を適用した．Chen らは，HE では評価できない Layer Normalization や softmax 関数を近似している．近似ができず，HE では評価できない処理は，サーバからユーザへ送信し，ユーザが一旦復号し，平文で実行を行っている．しかし，近似処理を含むため精度は低下している．また，途中結果をユーザ・サーバ間で通信しているため，サーバで処理が完結する場合と比較して通信量とユーザ負担が増加している．

3.3 秘密分散による方式

2022 年に Hao ら[15]，2023 年に Lu ら[16]，Pang ら[17]，Liu ら[18]は，ユーザとモデルを提供するサーバの二者を想定し，HE と秘密分散[19,20]を組み合わせて BERT に適用した．秘密分散とは，1 つの値を複数の値に分割し、それぞれの値に対する計算の出力を統合することで，本来の値に対する計算の出力を得る手法である．秘密分散はユーザ・サーバ間で行われ，入力データはサーバに対して秘匿され，モデルはユーザに対して秘匿される．

Hao らは，平文で実行した場合と同等の精度を保ちつつ，BERT_tiny のエンコーダ部分の推論実行時間はおよそ 26 秒であった．Lu らは BERT_base モデルが可能な限り多くのスレッドを用いて 2.77 分，Pang らは BERT_base モデルが 32 スレッドで 1.52～29.1 分，Liu らは GPU を用いた実装を行い BERT_tiny の 1 つのエンコーダ部分でおよそ 2.9 秒であった．これらの実行時間の違いは近似の導入方法や GPU 実装，マルチスレッド実装の違いによって現れる．

推論処理がサーバで完結する場合と比較してユーザの負担が増加する．また，推論過程においてユーザが常にオンラインである必要がある．

3.4 多項式近似による方式

2023 年，Zimerman ら[21]は，BERT がベースとする Transformer モデルに対して，HE で実行できない演算

や関数を多項式近似することで HE を適用した．推論に GPU やマルチスレッドを用いる実装をすることや近似で，実行時間を短縮している．実行時間は，BERT like な 6 層のモデルで 305 秒だった．ただし，Bootstrapping を用いるため，途中で復号する場合と比較して実行時間が長くなるという問題が生じる．

3.5 関連研究における課題

関連研究のまとめを表 1 に示す．BERT に HE を適用する先行研究では，以下の問題点がある．

問題点1. ユーザ側での処理が必要不可欠でありユーザ側の負担が大きい

問題点2. ユーザ・サーバ間の通信量が大きい

問題点3. 活性化関数の近似を利用した場合，平文モデルと比較して精度が低下する

表 1　関連研究

論文	概要	問題点
Garam ら[11]	●線形回帰を除きユーザ処理で実現 ●実行時間短縮 ●高精度	●ユーザ処理大 ●モデルの大部分が不保護
Chen ら[13]	●ユーザ・サーバ間のインタラクションにより実現	●ユーザ処理大 ●精度低下 ●通信量大
Hao ら[15] Lu ら[16] Pang ら[17] Liu ら[18]	●秘密分散により実現 ●高精度	●ユーザ処理大 ●通信量大
Zimerman ら[21]	●多項式近似により非線形関数を実行	●実行時間の増大

4. 準同型暗号と隔離実行環境の組み合わせによるプライバシを保護した BERT 文章分類手法の提案

4.1 概要

本節では，大規模言語モデルの一つである BERT をクラウド上で実行するにあたり，ユーザのデータと学習モデルを第三者から守るための手法を提案する．具体的には，HE と Intel SGX を組み合わせることで，HE の欠点を防ぐ. BERT のモデルとして，先行研究[13,15]と同様に，実行時間の観点から，BERT_tiny に線形層を接続したモデルを用い，HE を適用する．HE の鍵管理を Intel SGX 内で行うと共に，平方根や転置，softmax 関数や GELU 関数といった HE を適用できない演算を平文で実行する．その際，Intel SGX 内での平文処理を最小限に抑える．平文処理を減らすことで，攻撃者がサイドチャネル攻撃で途中結果を取得され，入力を逆算される可能性や出力を推測される可能性を減らす．ここで，HE の秘密鍵は，秘密鍵が格納されているメモリアドレスを TEE 外である REE に漏らすようなコードがある場合，サイドチャネル攻撃によって盗まれる[22]. よって，本稿では，そのようなコードが埋め込まれていないことを前提とする．また，全ての処理を Intel SGX 内で行うため，完全性が保証される．

以上の手法により，HE に適さない処理を TEE 内で平文実行することで，通信およびユーザ負担に頼ることなく推論処理をサーバ上で完結させ，3.5 項で示した問題点 1 及び問題点 2 を解決する．同様に，HE を適用できない演算を近似せずに TEE 内において平文で計算することにより，問題点 3 を解決する．

4.2 推論過程のプロトコル

「ユーザ」,「モデル提供者」,「サーバ」の三者を想定し，いずれも semi-honest であることを想定する．なお，通信経路上では AES 暗号を用いて暗号化する．以下に前処理及び推論処理のプロトコルを示す．

前処理：

① サーバ：TEE 内で HE の鍵ペアを生成.
② サーバ：生成した公開鍵をユーザとモデル提供者に送信.
③ モデル提供者：モデルの重みを暗号化.
④ モデル提供者：暗号化した重みとプログラムを TEE に送信.
⑤ ユーザ：AES 暗号の鍵を生成.
⑥ ユーザ：生成した AES 鍵をサーバへ送信.

推論処理（図 3）：

① ユーザ：BERT への入力する文章に対して辞書を用いて one-hot ベクトル化．出力をマスクするための二つのランダム値（マスク）を生成.
② ユーザ：①で生成したベクトルとマスクを HE により暗号化.
③ ユーザ：②をサーバに送信.
④ サーバ：TEE 上で推論処理を実行.
⑤ サーバ：復号時に推論結果が第三者に漏れないように，得られた二値にユーザから受け取ったマスクをそれぞれ加算[1].
⑥ サーバ：TEE 内で処理結果を復号した上，ユーザから受け取った AES 鍵を用いて暗号化.
⑦ サーバ：⑥をユーザに送信.
⑧ ユーザ：受け取った二つの値を復号した後，送信したマスクを減算し，分類結果を取得.

[1] オーバーフロー，アンダーフローは考慮しない.

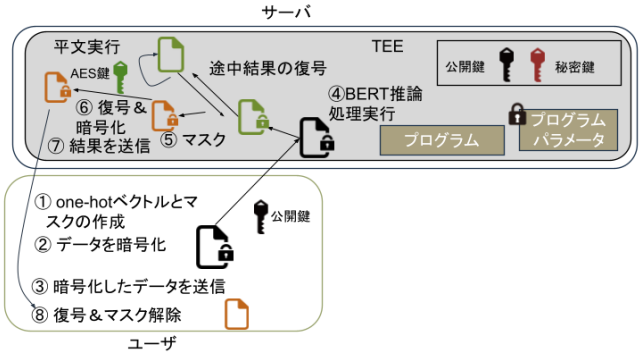


図 3　提案手法のプロトコル

4.3　準同型暗号と TEE の適用手法

2.3 項で説明した BERT の構造を元に，提案手法での BERT への HE の適用方法について説明する．

4.3.1 Matmul

図 4 に示すように，サイズ$[A,B]$の行列Pとサイズ$[B,C]$の行列Qの積を考える．この時，行列Pは行方向に，Qは列方向にそれぞれパッキングする．これは，パッキングを利用した行列積の最もナイーブな手法である．ナイーブな手法を用いたのは，汎用的に適用できるためである．

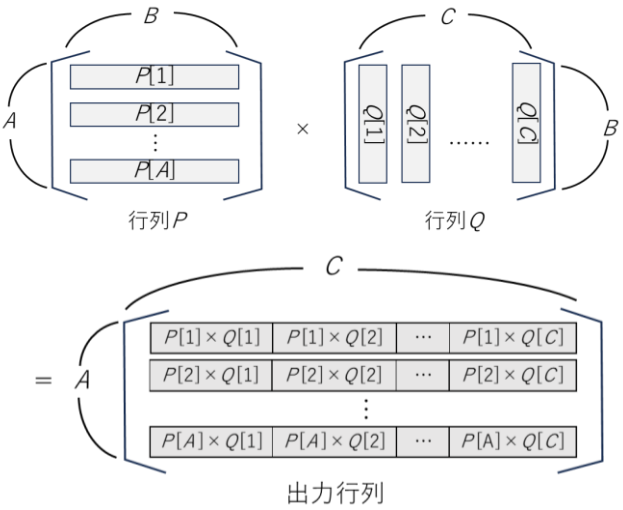
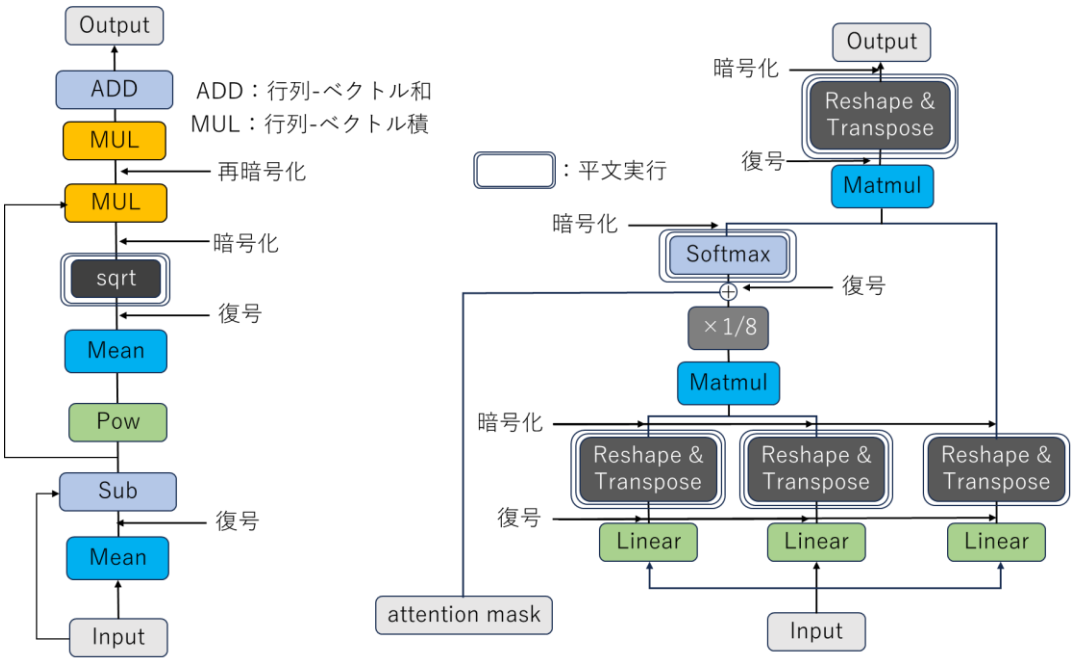


図 4　Matmul 時のパッキング手法

図 4 における$P[1]$と$Q[1]$で準同型乗算を行い，各スロットの値を合計したものが，出力行列の 1 行 1 列目となる．各スロットの値の合計は，rotate-sum（rotation と加算を繰り返す）で求める．サイズ 4 のベクトルRを格納した暗号文に rotate-sum を適用する処理の流れを図 5 に示す．

	$R[1]$	$R[2]$	$R[3]$	$R[4]$	-	-	…
			+				
rotate	$R[2]$	$R[3]$	$R[4]$	-	-	-	…
	$R[1]+R[2]$	$R[2]+R[3]$	$R[3]+R[4]$	$R[4]$	-	-	…
			+				
rotate	$R[3]+R[4]$	$R[4]$	-	-	-	-	…
	$R[1]+R[2]+R[3]+R[4]$	$R[2]+R[3]+R[4]$	$R[3]+R[4]$	$R[4]$	-	-	…

図 5　rotate-sum の処理の流れ（「-」は暗号文の使用しないスロットを表す）



(a) Layer Normalization　　(b) Multi-Head Attention

図 7　提案手法におけるモデル構造

以上の手法でP[1]に対しQ[1]～Q[C]まで同様の操作を行い，求めた暗号文を図 6 のように one-hot ベクトルをかけ合わせることでフィルタリングする．得られたベクトル列を足し合わせることで，出力行列の 1 行目のベクトルの暗号文を得る．この操作をP[2]～P[A]に対して同様に行い，行列積の結果を得る．

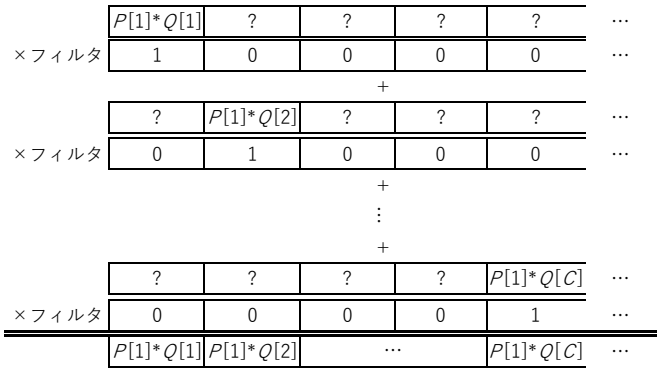


図 6　フィルタリングを使用した行ベクトルの作成（?は出力で使用しない値，…は使用しないスロット）

4.3.2 TEE の適用方法

Layer Normalization における復号及び暗号化の流れを図 7 (a)に示す．ここで，MUL は行列ベクトル積，ADD は行列ベクトル加算を表す. Embedding では，ユーザから受け取った暗号化された one-hot ベクトルに対して，Matmul 及び Layer Normalization を適用する．Multi-Head Attention（図 2 (c)）に HE を適用する手法を図 7 (b)に示す．4.3.2 項と同様に，非線形関数の実行やレベル削減のため，平文実行を行う．Feed Forward Network で用いられる GELU 関数は HE 上で評価できないため，GELU の前に復号し平文で実行する．他の Linear は HE 上で処理する．

5. 評価実験

TEE による，レベル削減と bootstrapping を行わないことの効果を測定するため，提案手法の推論レイテンシを評価する．具体的には，推論プログラムを 10 回実行し，推論レイテンシの平均を算出する．なお，ユーザ・サーバ間の通信は測定対象外とする．分類モデルの最終段である線形層は，入力サイズが[1,128]で出力サイズが[1,2]である．

また，比較対象として，HE に適さない演算を平文上で実行する代わりに，TEE を使わず bootstrapping 及び多項式近似を適用する手法をベースライン手法とする．

5.1 データセット

BERT 文章分類の推論対象として，IMDb（Internet Movie Database）[23]を用いる．IMDb は，7～2,952 語映画の感想文データと，各感想文がポジティブかネガティブかのラベルで構成される．評価実験では，評価対象のデータとして，IMDb に含まれている 62 語の感想文を用いる．

5.2 実験方法

実験環境を表 2 に示す．評価実験は，numactl コマンドを用いて，使用する CPU コアを固定してシングルスレッド実行で実施する．HE パラメータは，多項式環の次数を 16,384，スロット数を 8,192，scale factor を 40，最大初期レベルを 3，rescaling のモードを FIXEDMANUAL（rescaling を手動で行う OpenFHE の設定），準同型演算の方式を CKKS-RNS とした．また，TEE での測定には Gramine-SGX[25]を用いる．Gramine-SGX とは実行バイナリを修正することなく TEE で実行することを可能にするライブラリ OS である．

表 2　実験環境

構成要素		値
CPU	型番	Xeon Gold 6430@2.1GHz
	ソケット数	2
	SGX	SGXv2
	ターボブースト	無効
メモリサイズ		256 GB
エンクレーブ容量		64 GB/システム
OS		Ubuntu 20.04.6 LTS
HE ライブラリ		OpenFHE v1.1.1
Gramine-SGX		バージョン 1.6
カーネル		5.15.0-89-generic

5.3 ベースライン手法の推論レイテンシ算出方法

ベースライン手法では，Bootstrapping や多項式近似の活用により，推論レイテンシが長いことが予測される．そこで本実験では，準同型演算ごとの実行回数と 1 回あたりの平均レイテンシを用いて推論レイテンシを机上計算する．

実行は bootstrapping で 100 回，他は 1,000 回行う．Bootstrapping ではランダムに生成した複数の暗号文それぞれについて演算を行い，他は同様の暗号文にランダム値（暗号文もしくは平文）との準同型演算を行うことで，平均レイテンシを得る．また，多項式環の次数を 131,072，スロット数を 65,536，最大レベルを 31，rescaling のモードを FLEXIBLEAUTO（rescaling を自動で行う OpenFHE の設定）に設定する．

GELU 関数の近似は，Kim ら[26]の先行研究を参考にシグモイド関数を近似し，それを式(5)に適用する．また，$1/\sqrt{x}$の近似として，Panda[27]の近似法を利用する．そして，Lee ら[28]の近似法を用いて sofmax 関数を近似する．Bootstrapping は level 1 で実行する．そのため，Lee らを参考にした softmax に合わせ，そのために必要な最大レベル 10 の範囲で，レベルが枯渇したタイミングで行う．行列積など，一つの処理に 2 レベル以上消費する演算があり，かつその処理中にレベルが不足する場合は，処理を行う前に bootstrapping を行う．ベースライン手法における各準同型演算の回数を表 3 に示す．ここで，add_many(T)は T 個の暗号文の

表 3　ベースライン手法における各準同型演算の回数

level	mult	mult_plain	add	add_plain	rotate	bootstrap	add_many (128)	add_many (512)	add_many (62)	add_many (64)
10	1,922	930	0	0	0	0	0	0	0	0
9	4,774	2,418	0	1,302	0	0	0	0	0	0
8	194,246	930	0	2,232	0	0	0	0	0	0
7	4,340	169,694	1,770,968	992	1,911,646	0	0	0	0	0
6	6,820	4,488	3,038	992	2,554	0	806	124	0	0
5	5,582	11,718	5,270	992	2,170	0	0	0	0	0
4	35,404	3,286	4,960	992	9,978	0	0	0	256	0
3	5,954	36,084	241,864	1,302	269,762	0	0	0	0	0
2	3,224	18,044	1,874	2,046	1,502	0	0	0	0	248
1	0	0	5,953	15,872	0	4,992	1	0	0	0

表 4　準同型演算の実行時間（「—」はベースライン手法で使用されないことを表す.）

level	mult [ms]	mult_plain [ms]	add [ms]	add_plain [ms]	rotate [ms]	bootstrap [ms]	add_many (128) [ms]	add_many (512) [ms]	add_many (62)[ms]	add_many (64) [ms]
10	425.8	122.7	11.4	14.5	383.4	—	—	—	—	—
9	395.8	112.5	10.5	13.6	271.8	—	—	—	—	—
8	366.3	102.1	9.2	13.0	255.0	—	—	—	—	—
7	336.8	92.3	8.1	12.4	239.7	—	—	—	—	—
6	306.9	82.2	6.8	11.6	222.1	—	1,095.5	4,265.5	—	—
5	277.2	72.1	5.7	10.9	206.1	—	—	—	—	—
4	247.5	61.7	4.4	10.0	188.6	—	—	—	355.2	—
3	218.6	51.9	3.3	9.0	172.5	—	—	—	—	—
2	190.0	41.7	2.3	8.0	157.0	—	—	—	—	236.1
1	—	—	1.6	7.1	—	108,958.0	—	—	—	—

総和を求める操作のことである.

6. 実験結果と考察

6.1 TEE による実行時間への効果

提案手法の平均推論レイテンシは 22,333 秒であった．また，表 4 に示す各レベルでの準同型演算の実行時間の結果と表 3 で示した演算回数から，ベースライン手法の推論レイテンシは 1,178,581 秒と算出した．以上の結果から,提案手法はベースライン手法の 52.77 倍の実行速度となることが期待できる.

6.2 問題点の解決とレイテンシ短縮法

提案手法は，推論処理がサーバで完結し，ユーザ負担は入力ベクトルの作成と暗号化，出力の復号とマスク除去のみとなる.また,通信も入出力時のみとなる.そのため，3.5 項における問題点 1 及び 2 が解決できる．そして，近似を行っていないため，近似による精度低下（問題点 3）が起こらないのは明らかである．しかし，ベースライン手法と比較するとレイテンシの短縮は期待できるが，提案手法によるレイテンシは先行研究と比較して実用性に欠ける.

提案手法では，行列積アルゴリズムとそれに伴うパッキング法がナイーブであった．複数の行を一つの暗号文に格納するなど，より準同型演算の実行回数を減らせる余地がある．提案手法において，行列積は推論時間の 99.8%を占める処理であり，行列積に関する処理を改善できれば,推論レイテンシ短縮につながる.

6.3 提案手法のセキュリティレベル

ベースライン手法では，推論処理の途中で復号を行わないため，機密性が保証される．対して提案手法では，平文実行箇所の途中結果がサイドチャネル攻撃によって盗まれる可能性がある．その場合，モデルの重みを使用する箇所は全て HE 上で行っているため，モデル提供者以外が途中結果単体から入出力を割り出すことは困難である．しかし，提案手法では秘密鍵を Intel SGX 上で管理している.この秘密鍵がサイドチャネル攻撃によって盗まれた場合，HE の復号ができてしまうため，機密性が失われる.

上記の通り，提案手法は，モデルを暗号化しない手法と比較して，サイドチャネル攻撃によって途中結果が盗まれた場合に，入力データを復元されるリスクや出力を推測されるリスクが低い．しかし，途中結果から入力値が逆算できるかどうかなど，漏洩しても入力が逆算できない漏洩許容量を求めることで，平文で実行する箇所を増やすことができ，より現実の需要に即

したセキュリティレベルと実行速度を追求できると考えられる．

7. おわりに

BERT を用いた文章分類がクラウドコンピューティングによって行われる事例において，入力データやモデルの重みといった機密データの漏洩が懸念されている．そこで，本稿では，HE と Intel SGX を組み合わせることで，入力データと分類結果，分類モデルを秘匿しながら推論を行う手法を提案した．TEE 内で平文実行した結果，ユーザ・サーバ間の通信を分類モデルの入力時や出力時に限定することができ，ユーザ負担も軽減された．また，平文実行により，精度低下をもたらす近似を行わず実行時間を削減することが可能となった．また，推論レイテンシは 6.2 時間（22,333 秒）であり，ベースライン手法と比較して約 53 倍の高速化を達成することが期待できる．

今後の課題は，より大きなモデルに提案手法を適用することである．また，提案手法は関連研究と比較すると実用的な推論レイテンシではないため，行列積アルゴリズム改善によるレイテンシ短縮が課題である．

参 考 文 献

[1] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, 2019.

[2] C. Gentry, “Fully Homomorphic Encryption Using Ideal Lattices,” Proc. of the 41st annual ACM symposium on Theory of computing, pp. 169–178, 2009.

[3] GlobalPlatform, “Introduction to Trusted Execution Environment,” https://globalplatform.org/wpcontent/uploads/2018/05/Introduction-to-TrustedExecution-Environment-15May2018.pdf, (参照 2024-01-10).

[4] J. H. Cheon, A. Kim, M. Kim, Y. Song, “Homomorphic encryption for arithmetic of approximate numbers,” Proc. of International Conference on the Theory and Application of Cryptology and Information Security, pp. 409–437, 2017.

[5] N.P. Smart, F. Vercauteren, “Fully homomorphic SIMD operations,” Des. Codes Cryptogr. Vol. 71, pp. 57–81, 2014.

[6] 大西隆太郎, 鈴木拓也, 山名早人, “準同型暗号 と隔離実行環境の組み合わせに関するサーベイ,” コンピュータセキュリティシンポジウム 2022（CSS2022）, 2A2-I-1, pp.1–8, 2022.

[7] F. McKeen, I. Alexandrovich, A. Berenzon, C. V. Rozas, H. Shafi, V. Shanbhogue, U. Savagaonkar, “Innovative instructions and software model for isolated execution,” Proc. of the International Workshop on Hardware and Architectural Support for Security and Privacy, pp. 1–8, 2013.

[8] 横平正熙, 大西隆太郎, 鈴木拓也, 山名早人, “Intel SGX に対するサイドチャネル攻撃手法に関するサーベイ,” コンピュータセキュリティシンポジウム 2023（CSS2023）, 2E4-1, pp.1–8, 2023.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. “Attention is all you need,” Proc. of Neural Information Processing Systems, pp. 6000–6010, 2017.

[10] D. Hendrycks, K. Gimpel, “Gaussian Error Linear Units (GELUs),” arXiv preprint arXiv: 1606.08415, pp.1–10, 2016.

[11] G. LEE, M. KIM, H. J. Park, S. W. Hwang, J. H. Cheon, “Privacy-Preserving Text Classification on BERT Embeddings with Homomorphic Encryption,” Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3169–3175, 2022.

[12] N. Reimers, I. Gurevych. “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” Proc. of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992, 2019.

[13] T. Chen, H. Bao, S. Huang, L. Dong, B. Jiao, D. Jiang, H. Zhou, J. Li, F. Wei, “The-x: Privacy-preserving transformer inference with homomorphic encryption,” in Findings of Association for Computational Linguistics, pp. 3510–3520, 2022.

[14] I. Turc, M.W. Chang, K. Lee, K. Toutanova, “Well-read students learn better: On the importance of pre-training compact models,” arXiv preprint arXiv:1908.08962, pp.1–13, 2019.

[15] M. Hao, H. Li, H. Chen, P. Xing, G. Xu, T. Zhang, “Iron: Private Inference on Transformers,” Proc. of Neural Information Processing Systems, pp. 15718–15731, 2022.

[16] W. J. Lu, Z. Huang, Z. Gu, J. Li, J. Liu, K. Ren, C. Hong, T. Wei, W. Chen, “BumbleBee: Secure Two-party Inference Framework for Large Transformers,” Cryptology ePrint Archive, cryptoeprint:2023/1678, pp.1–18, 2023.

[17] Q. Pang, J. Zhu, H. Möllering, W. Zheng, T. Schneider, “BOLT: Privacy-Preserving, Accurate and Efficient Inference for Transformers,” Cryptology ePrint Archive, cryptoeprint:2023/1893, pp.1–19, 2023.

[18] X. Liu, Z. Liu, “LLMs Can Understand Encrypted Prompt: Towards Privacy-Computing Friendly Transformers,” arXiv preprint arXiv: 2305.18396, pp.1–15, 2023.

[19] R. Cramer, I. B. Damgård, J. B. Nielsen, “Secure multiparty computation and Secret Sharing,” Cambridge University Press, pp. 1–373, 2015.

[20] D. Demmler, T. Schneider, M. Zohner, “Aby-a framework for efficient mixed-protocol secure two-party computation.” Proc. of The Network and Distributed System Security Symposium, pp. 8–11, 2015.

[21] I. Zimerman, M. Baruch, N. Drucker, G. Ezov, O. Soceanu, L. Wolf, “Converting Transformers to Polynomial Form for Secure Inference Over Homomorphic Encryption,” arXiv preprint arXiv: 2311.08610, pp.1–13, 2023.

[22] 横平正熙, 大西隆太郎, 鈴木拓也, 山名早人, “Intel SGX 上で動作する準同型暗号の秘密鍵に対するサイドチャネル攻撃,” データ工学と情報マネジメントに関するフォーラム, 2024.

[23] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts. “Learning Word Vectors for Sentiment Analysis,” Proc. of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150, 2011.

[24] A. A. Badawi, J. Bates, F. Bergamaschi, D. B. Cousins, S. Erabelli, N. Genise, S. Halevi, H. Hunt, A. Kim, Y. Lee, Z. Liu, D. Micciancio, I. Quah, Y. Polyakov, Saraswathy. R.V, K. Rohloff, J. Saylor, D. Suponitsky, M. Triplett, V. Vaikuntanathan, V. Zucca, “OpenFHE: Open-Source Fully Homomorphic Encryption Library,” Proc. of the 10th Workshop on Encrypted Computing & Applied Homomorphic Cryptography, pp. 53–63, 2022.

[25] “Introduction to Gramine,” Gramine SGX - Read the Docs, https://gramine.readthedocs.io/en/stable/

[26] M. Kim, Y. Song, S. Wang, Y. Xia, X. Jiang, “Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation,” JMIR Medical Informatics, pp. 1–12, 2018.

[27] S. Panda, “Polynomial Approximation of Inverse sqrt Function for FHE,” Proc. of Cyber Security, Cryptology, and Machine Learning, pp. 366–376, 2022.

[28] J. W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, E. Lee, J. Lee, D. Yoo, Y. S. Kim, J. S. No, “Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network,” IEEE Access, vol. 10, pp. 30039–30054, 2022.

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

準同型暗号

座長:上土井 陽子(広島市立大学)
コメンテータ:天笠 俊之(筑波大学)
2024年2月28日(水) 10:00 〜 12:10 T2-B (オンライン（Zoom Events))

11:45 〜 12:10

[T2-B-1-05]【技術報告】不動産情報サービスの研究開発における共有データ資源

清田陽司 (株式会社LIFULL)

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 10:00 ~ 12:10 | T3-A オンライン（Zoom Events)

SNS分析 (1)

座長:関 洋平(筑波大学)
コメンテータ:平手 勇宇(楽天株式会社)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T3-A-1-01] 画像付きSNS投稿における画像ーテキスト間の相互関係分析

*鍋岡 琢渡[1]、馬 強[2] (1. 京都大学、2. 京都工芸繊維大学)

10:30 ~ 10:55
[T3-A-1-02] コロナ禍の話題表す俳句生む

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

10:55 ~ 11:20
[T3-A-1-03] 同一ハッシュタグ投稿に着目した対照学習によるトピック類似性を捉えた文ベクトルの獲得

*仲田 明良[1]、狩野 芳伸[1] (1. 静岡大学　情報学部　狩野芳伸研究室)

11:20 ~ 11:45
[T3-A-1-04] SNS実データ分析に基づくユーザのフォロー行動モデルの評価

*山川 衛[1]、田島 敬史[1] (1. 京都大学大学院情報学研究科)

11:45 ~ 12:10
[T3-A-1-05] イベント発生の不均衡性に適合した人口変動予測

*塚田 涼太郎[1]、豊田 正史[2]、梅本 和俊[2]、是津 耕司[3] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所、3. 情報通信研究機構)

画像付きSNS投稿における画像ーテキスト間の相互関係分析

鍋岡　琢渡†　　馬　　強††

† 京都大学 大学院情報学研究科　〒606–8501 京都府京都市左京区吉田本町
†† 京都工芸繊維大学 大学院工芸科学研究科　〒606–8585 京都府京都市左京区松ヶ崎橋上町
E-mail: †nabeoka.takuto.45e@st.kyoto-u.ac.jp, ††qiang@kit.ac.jp

あらまし　SNS には目を惹く画像を伴う投稿が多数存在する．その画像とテキストのペアにより異なる意図を与えるため，画像ーテキスト間の情報の冗長性や相互の情報付加のパターンを分類し，その関係性をふまえた情報提供や理解が大切である．本研究では，画像から想起されるテキストの集合と，テキストから想起される画像の集合の双方向の傾向を利用して画像ーテキスト間の関係を分類することを試みる．具体的には，画像から検索や生成されるテキストの集合とペアデータのテキストによるクラスターと，テキストから検索や生成される画像の集合とペアデータの画像によるクラスターの分布を利用する．実験の結果，関連には複数のパターンがあることが確認され，本手法による分類の有用性が示唆された．最後には提案手法を用いた応用例についても議論する．

キーワード　ソーシャルメディア, 関係分析, 補完関係, 生成 AI, クラスター分析, 情報検索

1　はじめに

世間では X(旧 Twitter) や Instagram を始めとするソーシャル・ネットワーキング・サービス (以下，SNS) が多数活用され，誰でも情報の送受信が手軽になっている．SNS 投稿では，テキストのみならず画像を伴う投稿をされることも多い．特に X 上ではテキストのみの投稿だけではなく，複数の画像を伴う投稿や画像のみの投稿といったように，複数の投稿手段が存在し，画像を伴う投稿はテキストのみの投稿に比べて，目を惹き，興味をそそるものが多い．実際に画像を添付された投稿は，約 42%存在し，画像のないものに比べて約 22.8%多くエンゲージメント [1]を獲得しているという報告もある [1]．すなわち，投稿の際に画像を用いることで興味を惹かせることが可能となる．

しかし画像付き投稿では，画像とテキストの組み合わせに応じて意図する内容が変わり，投稿の印象を変える可能性がある．例えば，あるラーメン屋 A におけるラーメンの画像を伴う投稿を想定し，図 1 の 4 つのテキストがそれぞれ添付されている場合をそれぞれ考える．

（1）「美味しい！」

この例の場合，テキストから美味しいという感想が伝わるが，美味しさの対象が明示されておらず，情報として不十分である．一方で画像が存在することで初めて，具体的にラーメン屋 A のラーメンを指していることが分かる．すなわち，画像がテキストを補足する関係になっていると考えられる．

（2）「これって，どこのラーメン？」

この例の場合，テキストに「これ」という指示語が含まれていることで，テキスト内で画像の存在を示唆しており，画像の存在の意図を明確にしている．すなわち，テキストにより画像の役割を示唆する関係になっていると考えられる．

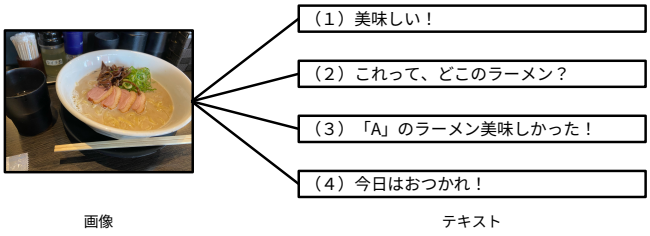


図 1　画像とテキストの関連の具体例

（3）「A のラーメン美味しかった！」

この例の場合，テキストは具体的な店名を指す内容で，添付画像の種類も想像しやすい一方で，画像についても，テキストに含まれる“「A(店名)」ラーメン“の具体的な視覚的情報として，画像がテキストの補足を行っており，相互に情報を補足し合う関係になっていると考えられる．

（4）「今日はおつかれ！」

この例の場合，テキストのみでも内容の意図は伝わる一方で，ラーメン屋の画像のみからでは伝わらない内容になっており，ラーメンの画像がない場合でも，内容を伝える意図に対して相違はなく，画像によりテキストを強調している関係になっていると考えられる．

以上のように一つの画像からにおいても添付されるテキストの内容によって，テキストや画像の役割がそれぞれ変化すると考えられ，効果的に内容を伝えるためには，意図に合う画像とテキストのペアであることが重要である．本研究では SNS 投

1：ここでは，ユーザーが投稿した反応や行動数を指し，具体的にはリツイート数 (令和 6 年 3 月 21 日現在はリポストと呼ばれることもある)，いいね数，そしてメンション数が含まれる．

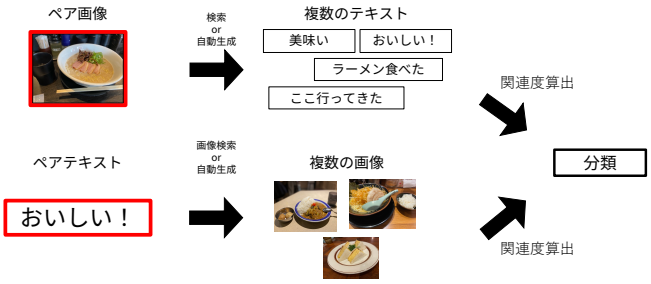


図 2　提案手法の概要

稿の画像ーテキストのペアを対象にしているが，投稿で用いられる使用方法をふまえたマルチモーダルの生成モデルのデータセット作成や，資料作成支援の連想検索にも役立つと考えられる．応用例については，第 5 節で詳しく議論する．

しかしこれらの応用例を扱う際には，テキストと画像のペアが必ずしも互いに説明し合うものになっているとは限らなく，単純に画像を説明するラベルになっているデータのみでは不十分である．またラベル付けに対しても，ラベル付与者の主観によるバイアスが含まれる可能性があり，コストもかかる．

そこで本研究では，画像を伴うテキスト付きの投稿を対象とし，ラベル付きデータセットを用いずに，画像とテキストの関連性の調査，ならびに分類することを試みる．ここでペアデータとなっているテキストから画像が想起可能であるか，またペアデータとなっている画像からテキストが想起可能であるかに注目する．この時，複数人が想起した際の画像やテキストの集合の類似性が小さい場合は曖昧性があると考えられ，曖昧性が高い元の画像やテキストの重要性はペアデータとして高いと考えられる．またテキストからペアデータの画像に類似する画像が想起されたり，画像からペアデータのテキストに類似するテキストが想起されたりする場合，情報が冗長であるとみて，ペアデータとして重要性が低くなると考えられる．これらを基にした分類システムを提案する．

図 2 に提案手法の概要を示す．具体的には，次の 2 つの評価を総合した結果をふまえて分類する．一つ目は，画像から**想起できる**複数のテキスト間の類似性で，二つ目は，テキストから**想起できる**複数の画像間の類似性である．ここで「**想起できる**」について,「検索」や「自動生成」を活用して分析する．すなわち，画像から検索や自動生成することにより得られる複数の説明テキスト間のクラスタリングと，テキストから検索や自動生成することにより得られる複数の画像間のクラスタリングを分析し，そこから情報冗長度や曖昧性を考慮した分類を行う．ここで，検索手法は実在する要素を基に分析するアプローチであり，リアルタイム性があることに加えて，流行や一般的に使用されやすい画像やテキストが得られることを期待する．自動生成手法は与えられた要素からの想像を基に分析するアプローチであり，一般的に想像されやすい画像やテキストが得られることを期待する．両者の性質は異なり，それぞれの手法について調査する．

本稿の構成は以下の通りである．まず第 2 節で，関連研究を述べる．次に第 3 節で,「想起できる」について検索を用いた場合と自動生成を用いた場合の各提案手法を述べる．次に第 4 節では，図 1 の例や実例を用いて提案手法により分類を行い，結果の可視化ならびに分析を行い，本手法の有用性を議論する．そして第 5 節で，本研究の他の応用例について議論する．最後に第 6 節で本稿のまとめ，今後の展望を述べる．

2　関 連 研 究

この節では SNS における画像のようなマルチメディアを分析する研究や，画像やテキストを利用した研究について述べ，本研究の位置付けをまとめる．

2.1　SNS 投稿上のマルチメディアに関する研究

SNS 投稿における画像とテキストの関連度についての研究は数多く行われている．

SNS 投稿におけるタグに注目して画像の説明を行うことを試みる研究がある．Wang らは，画像と紐づけられたユーザーが与えたタグを活用して説明文を生成する方法を提案し，注意すべき単語にアテンションさせる手法を提案している [2]．Maenishi らは，Instagram のタグは必ずしも画像内の写真に関連ないことに注目し，どのタグが関連する画像コンテンツを記述している可能性が高いかを決定する方法を提案している．具体的には，画像に添付されているタグ 1, タグ 2 に対して，タグ 1+2 の類似度がどちらの方により近いかを計算し，より高い方を画像への関連度が高いものとして判断する方法をとっている [3]．また確率的アプローチを用いた研究もある．Katsurai らは，画像データベースから概念の意味的関係を抽出するためのクロスモーダルアプローチを提案している．彼らは正準相関分析 (CCA) を用いて，データベース中の視覚的特徴とタグ特徴の相関を捉えて，CCA の確率的解釈により，画像の分布に注目して分析することを行っている [4]．一方で本研究では，タグではなくテキストを用いて画像との関連性に追求する．

SNS 上のマルチモーダル名前付きエンティティ認識（MNER）に注目した研究がある．画像を活用して名前付きエンティティ認識 (NER) の精度を向上させることが可能とされている．Zhang らは，X における画像を用いた NER を，LSTM や Co-attention ネットワークを使用して初めて行なった [5]．Zhao らは，異なる画像ーテキストペア間のマッチング関係を用いた NER のための R-GCN を提案した [6]．Sun らは，X におけるツイートに無関係なテキスト画像ペアが大きな割合を占めていることに注目し，テキストと無関係な画像によって引き起こされる視覚的なアテンションの影響を抑えることを目標とし，マルチモーダル BERT モデルにテキストと画像の関連性を伝播する RpBERT を提案した [7]．一方本研究はその関連性自体に注目して分類を行うため，関連の有無に留まるわけではない．

また画像とテキストの関連性を分類する研究もある．Vempala らは，画像がツイートに意味を与えているかどうかとツイートの文章が画像内で表現されているかどうかを分類するタスクを行った．この 2 つの軸について各々二値分類を行なった計 4 つの分類軸により，人為的にラベル付けされた X のデー

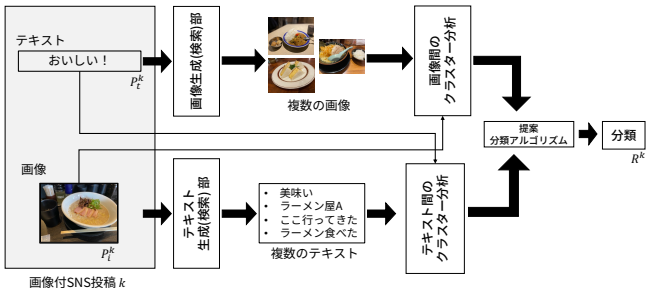


図 3　提案手法の全体像

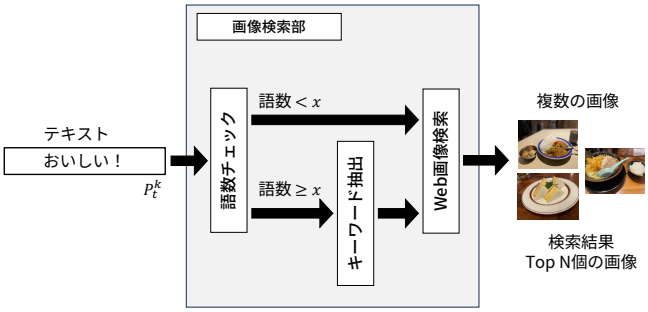


図 4　画像検索部の詳細

タセットを準備し，Bag of Words や LSTM，InceptionNet を用いて自動分類を行った．LSTM+InceptionNet を組み合わせたモデルを用いて画像が必要かどうかのタスクで優れた結果を示し，画像が必要かのタスク，特に画像が冗長であってツイート内容に意味を与えていない場合に対して，性能が高いことが示されている [8]．しかし，これは分類するのみであるタスクであるため，具体的にどの部分に影響を受けて分類されたのか明確ではない．また，この研究ではラベル付けされたデータを用いて分類を行っているため，ラベル付与者の主観によるバイアスが含まれる可能性がある．一方で本研究ではデータセットに対してラベル付けを行うことなく分類を行うことを試みており，検索や生成を用いることで，既に蓄積された一般的な傾向をふまえた分類が可能になると考えられる．

2.2　画像とテキストのペアを用いた研究

画像説明や画像生成として次のような研究がある．画像を入力し妥当な英語説明文を生成するニューラルネットワークを用いた研究は既に盛んに行われている [9]．また，RNN を用いて画像とその領域の自然言語記述を生成するモデル [10] がある．近年では，生成 AI と呼ばれるものを用いて，画像説明や画像生成が行われることが多い．

画像とテキストを同時に扱う方法についての研究もある．Li らは，マルチヘッドアテンションネットワークを利用して，画像とテキストに対して共通のハッシュコードを生成することで，クロスメディアハッシュ検索に活用する研究を行なった [11]．

しかし，今回対象とする SNS 投稿においては，必ずしも画像内の内容を説明したり補足したりする内容ではなく，第 1 節の (4) のように関連が不明瞭なものも含まれる．そのため，単に画像説明により検索・生成されたテキストや，テキストから検索・生成された画像のみでは，必ずしも意図に相応した結果が得られるわけではなく，今回のような画像とテキストの関連度を考慮した機構を含むシステムが必要とされる．

3　提 案 手 法

この節では提案手法について，問題定義として提案手法の入出力と，提案手法における各部分の機構の順に説明を行う．

3.1　提案手法の全体像

全体像としては，図 3 である．各画像付き SNS 投稿 k のテキストと画像に対して，並列して各々処理を行う．そして，それらの結果を総合分析して画像とテキストの関係を分類する．

まずここでは，提案手法を説明するにあたって必要となる入出力や記号を定義する．入力は，画像付き SNS 投稿 k における画像 P_i^k とテキスト P_t^k である．複数の画像が添付されている場合はその中から一つの画像を選ぶものとする．出力は，それらの画像とテキストの関連度 R^k である．この関連度はラベルとして出力され，詳細は 3.5 節で述べる．

次にテキスト P_t^k の処理の流れについてである．テキスト P_t^k に対して画像生成 (検索) 部により，複数の画像を取得する．この詳細は検索を用いたものは 3.2.1 節，自動生成を用いたものは 3.3.1 節で説明を行う．次にそれらの画像間に対して関連度を算出する．この際にクラスター分析により傾向を可視化しつつ計算を行う．この詳細は 3.4.1 節で説明を行う．

そして画像 P_i^k の処理の流れについてである．画像 P_i^k に対してテキスト生成 (検索) 部により，複数のテキストを取得する．この詳細は検索を用いたものは 3.2.2 節，自動生成を用いたものは 3.3.2 節で説明を行う．次にそれらのテキスト間に対して関連度を算出する．この際にクラスター分析により傾向を可視化しつつ計算を行う．この詳細は 3.4.2 節で説明を行う．

最後に画像間のクラスター分析とテキスト間のクラスター分析の結果をもとに分類を行い，R^k を出力する．この詳細は 3.5 節で説明を行う．以上の流れで説明を行う．

3.2　検索を用いての変換

この節では，図 3 の初めの部分にあたる画像検索部やテキスト検索部についての説明を行う．

3.2.1　画像検索部

画像検索部では，テキスト P_t^k から検索エンジンを用いて画像を検索する．これは図 3 の画像生成 (検索) 部において，検索を選択したものに値する．図 4 に，画像検索部全体の処理を示す．

与えられたテキスト P_t^k が長文である場合，検索を行う際に語数制限が存在し，重要語が抜けてしまう可能性がある．そのため，語数チェックとして長文の場合，すなわちテキスト P_t^k のキーワード語数が制限語数 x を超える場合を判定し，超えた場合，キーワード抽出にて重要語を y(本研究では $y = 5$ とする) 個取り出す．本研究では，後述する Web 検索において Google 検索を用いているが，Google 検索では，検索キーワードが 32 語に制限されているため，$x = 32$ として語数を判定し，この

$x(=32)$ 語を超えた場合に Google Bard[1]を用いてキーワード抽出を行った．Google Bard には，

> 「(ペアテキスト P_t^k)」ここから重要順に語句を y 個列挙してください．

というプロンプトを与えて，y 個の語句を取得する．この処理により，可能な限りテキスト内の重要語を削ることない状態で検索を行う．

Web 検索では，語数が x より少ない場合はテキスト P_t^k，それ以上の場合はキーワード抽出により取得した y 語のキーワードを用いて，Top N 個の検索結果を利用する．本研究では，Google 検索を用いて，画像の検索結果の Web サイトから初めの N(本研究では $N=20$) 枚の見出しの画像を取り出す．

3.2.2 テキスト検索部

テキスト検索部では，画像 P_i^k からテキストを検索する．これは図 3 のテキスト生成 (検索) 部において，検索を選択したものに値し，図 4 における画像検索の処理を行う．

画像検索では，ペア画像 P_i^k に対して画像を検索し，その結果を用いて得られた複数のテキストを使用する．ここでは，画像検索により得られた Web サイトの見出しを Top N 個取得する．これらを画像の説明テキストとする．本研究では，Google 画像検索を用いて画像 P_i^k に関連するページを取得し，一番初めの N(本研究では $N=20$) 個のページの見出しのテキストを取得する．

3.3 自動生成を用いての変換

この節では，画像生成部とテキスト生成部についての説明を行う．これは，3.2 節において検索の代わりに自動生成を用いた場合の手法である．

3.3.1 画像生成部

画像生成部では，テキストから画像を検索する．これは図 3 の画像生成 (検索) 部において，生成を選択したものに値する．

画像生成では，ペアテキスト P_t^k をプロンプトとして画像を複数生成し，その得られた複数の画像を使用する．この方法により，テキストに対して関連度が高い画像を取得することが可能となると考えられる．

今回の研究では，Stable Diffusion を用いて，N(本研究では $N=20$) 個生成するまで繰り返し行い，生成された画像を使用する．

3.3.2 テキスト生成部

テキスト生成部では，画像からテキストを検索する．これは図 3 のテキスト生成 (検索) 部で，生成を選択した場合である．

テキスト生成では，ペア画像 P_i^k に対して，その画像を説明するような複数の説明文を N(本研究では $N=20$) 個生成する．今回の研究では，Google Bard を用いて，ペア画像 P_i^k とプロンプトとして次のものを与える．

> この画像を説明する文章を N 個作成してください．

1：令和 6 年 3 月 21 日現在，Google Gemini に名称変更され，アップデートされた．

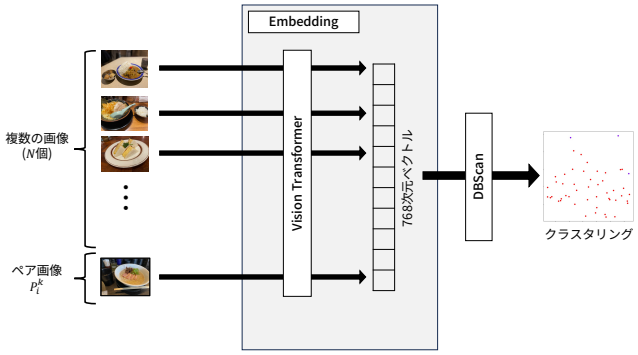


図 5　画像間のクラスタリング手法の詳細

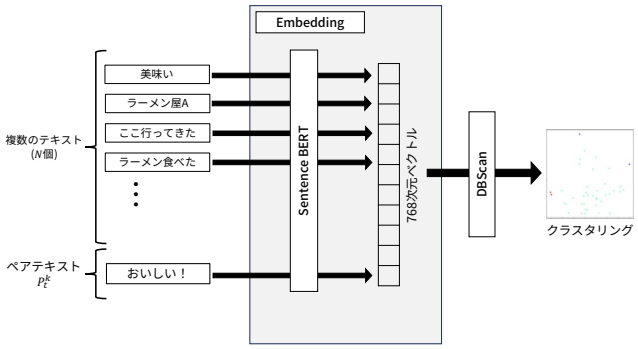


図 6　テキスト間のクラスタリング手法の詳細

これにより，そのペア画像に対する説明文が N(本研究では $N=20$) 個生成される．

3.4 クラスター分析

この節では図 3 のクラスター分析の詳細の説明を行う．

ここでクラスター分析は，DBScan [12] を用いて行う．DBScan では，密度に基づくクラスタ概念に依存するクラスタリングアルゴリズムであり，K-means 法のような事前のクラスターの数の設定が不要である．本研究では，カテゴリー数が明確ではない一方で，カテゴリーの分布を調べたい性質を持つことから DBScan を用いてクラスター分析を行う手法を選択する．

Embedding の手法，並びに具体的なクラスター分析の手法は，3.4.1 節で画像間，3.4.2 節でテキスト間，についてそれぞれ詳細の説明を行う．

3.4.1 画像間のクラスタリング

図 5 に画像間のクラスタリングの処理を示す．画像に対してクラスタリングを扱えるようにするため，各画像をベクトル表現にする．すなわち，Embedding の処理を行う．本研究では，Vision Transformer [13] を用いて各画像をベクトルに変換する．その際に，特徴抽出のための事前学習モデルである”vit-base-patch16-224-in21k” [14] を利用する．このモデルは，解像度 224x224 ピクセルの ImageNet-21k [15] に対して事前学習された変換エンコーダー・モデルである．そのため画像の入力サイズは全て 224x224 ピクセルにリサイズして実験を行う．そして最終隠れ層から，768 次元のベクトルを取得して利用する．

表 1　提案分類表 (R^k)

		画像 P_i^k ✓	画像 P_i^k ×
テキスト P_t^k	✓	(a)	(b)
	×	(c)	(d)

3.4.2　テキスト間のクラスタリング

図 6 にテキスト間のクラスタリングの処理を示す．テキストに対してクラスター分析を扱えるようにするため，各テキストについて Embedding 処理を行う．本研究では，SentenceBERT [16] を用いて Embedding を行う．この Transformer は複数の文章のクラスタリングに向く事前学習済み BERT ネットワークの修正であり，単語単位に分割することなく文単位でベクトル表現が可能となる．今回取得したテキストは説明文であるため，テキスト全体で処理を行えるように，この Transformer を用いている．また，”all-mpnet-base-v2”の事前学習モデルを使用する．このモデルにより，与えられた文や段落が 768 次元の密なベクトル空間にマッピングされる．これにより画像とテキストの Embedding がともに 768 次元ベクトルにマッピングされる．

3.5　分　　類

3.4 節の結果を利用して分類を行う．画像間のクラスタリングでは，入力のペアデータの画像 P_i^k と，ペアデータのテキスト P_t^k から 3.2.1 節や 3.3.1 節により変換した複数の画像を考慮した指標が望ましい．テキスト間のクラスタリングでは，入力のペアデータのテキスト P_t^k と，ペアデータの画像 P_i^k から 3.2.2 節や 3.3.2 節により変換した複数のテキストを考慮した指標が望ましい．

これらの条件をふまえて本研究では，以下の方法で行う．

（1） 画像 P_i^k が，テキスト P_t^k から変換した複数の画像により形成される，いずれかのクラスターに属している

（2） テキスト P_t^k が，画像 P_i^k から変換した複数のテキストにより形成される，いずれかのクラスターに属している

これら 2 つの軸による二値分類の総合の結果で，表 1 のように 4 つのパターン ((a), (b), (c), (d)) に分類する．この表では，もとのペアデータがいずれかのクラスターに属している場合にチェックマーク (✓) となる．上記の (1) の判断軸で，クラスターに属している場合は (a)(c)，そうでない場合は (b)(d) に分類する．上記の (2) の判断軸で，クラスターに属している場合に (a)(b)，そうでない場合に (c)(d) に分類する．

4　実　　験

この節では，提案手法の有用性を確認するため，図 1 における 4 つの例に対して実験を行った後，X 上の実際の投稿のデータを用いた実験も行なった．なお (3) の例で”A”と書かれている部分については，実際の具体的なラーメン屋の名前に置き換えて実験を行なった．

4.1　方　　法

3.2 節の検索手法と 3.3 節の自動生成手法のそれぞれの手法に従って，画像からテキストに変換したもの，テキストから画像に変換したものを各 20 個準備した．検索手法では，4 つの例のテキストのキーワード語数は $x(=32)$ 語を超えていないため，キーワード抽出を行わずに検索を行なった．画像生成手法では，今回画像生成では StabilityAI の Stable Diffusion 2.1 を用いて行ったが，各日本語のテキストを英訳したものを用いてプロンプトとした．具体的には，(1) は”Delicious!” (2) は”Where is this ramen from?” (3) は”A’s ramen was delicious!” (4) は”Good work today!” である．また望ましくない要素を画像内に描写しないようにするための指示文章である Negative Prompt として，worst quality(最低品質), ugly(醜い), bad anatomy(おかしな人体構造), jpeg artifacts(JPEG での圧縮による劣化) の 4 つの英単語を与えて，計 20 個の画像が得られるまで，各文に対して繰り返し生成を実行した．

次に 3.4 節によりクラスター分析を行なった．本研究ではクラスター分析の実験を行う際に，Python の機械学習ライブラリの一つである scikit-learn.cluster の DBSCAN を用いてクラスタリングを行なった．ここで予め定めておくパラメータとして eps と min_samples がある．eps は同じクラスターとするデータ点 p を中心とした半径の距離を指し，この実験では各手法により値を変えた．検索手法においては，テキストから画像への変換 (すなわち画像間のクラスタリング) の実験では eps=4.0，画像からテキストへの変換 (すなわちテキスト間のクラスタリング) の実験では eps=0.9 を選択した．生成手法においては，テキストから画像への変換 (すなわち画像間のクラスタリング) の実験では eps=3.0，画像からテキストへの変換 (すなわちテキスト間のクラスタリング) の実験では eps=0.7 を選択した．min_samples は半径 eps の内側に含むデータポイントの数の最小値を指し，この実験では，全て min_samples=2 に統一した．

また結果を可視化する際には，高次元データを低次元に落とし込むための次元削減アルゴリズムの一つである t-SNE を用いて行なった．クラスタリングを行なったベクトルの次元数が 768 次元であり，2 次元に次元削減した．また同じクラスターは同色になるようにし，それぞれ変換後の 20 個のデータにもとのペアデータを含めた 21 個に対して行なった．なお図の紫色で書かれた数字は各々外れ値となるように可視化した．

4.2　予備調査結果

4.2.1　検 索 手 法

検索手法によりクラスタリングした結果を，図 7 に示す．表 1 の規則により分類した結果は表 2 であり，(1)(4) が (b)，(2) が (d)，(3) が (a) となっている．これらのテキストを分析すると，(a) の (3) は具体的なラーメン屋の名前を与えたグループ，(b) の (1)(4) は画像中に存在する物に言及していないグループ，そして (d) の (2) は画像に写っている物のジャンル (ラーメン) を言及しているグループとなるように分類されていると考えられる．ここから，各クラスタリングは比較的特徴をもってクラスタリングされている可能性を示している．

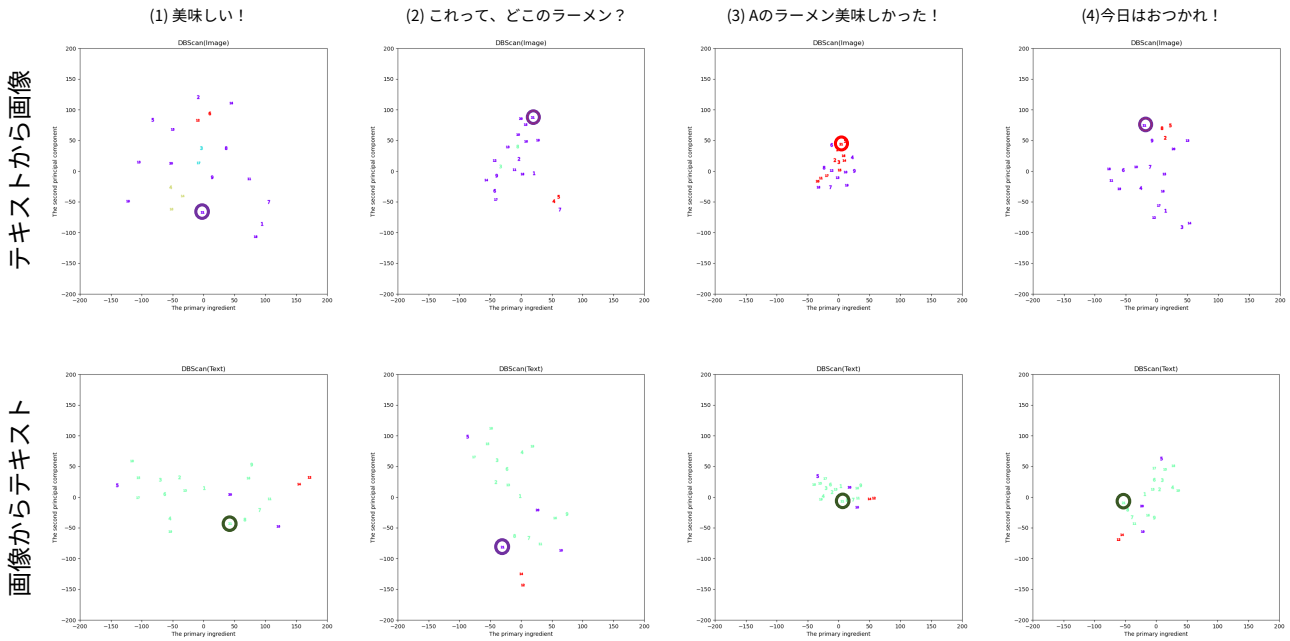


図 7　検索手法による結果 (図中の丸はペアデータを示し，紫色はクラスターの外れ値を示す．また，実験のスクリプトでは (3) の”A”には実際の具体的な店名が与えられている)

表 2　結果 (検索手法)

		画像 P_i^k ✓	画像 P_i^k ×
テキスト P_t^k	✓	(3)	(1)(4)
	×		(2)

以上が提案手法による分類であるが，ここで結果のクラスタリング自体について考察を行う．分類軸では上記のようなクラスタリングが行われたが，テキストから画像のクラスタリング結果に注目すると，ペアデータを含めた 21 個のデータセットに対して，同じパラメータの条件下であるものの異なる傾向が見られる．(1) や (2) は複数個のクラスターが見られ，(4) ではクラスターがほとんど見られない．また (3) は他と比べて属する要素数が多いクラスターが形成されている．同一クラスター間に属している要素同士は距離が近いと考えられるため，(4) のように紫色 (外れ値) が多かったり，(1) や (2) のように要素数の少ないクラスターの数が多かったりするほど，何らかの部分で曖昧な部分が多くなっていると考えられ，(3) のように特定可能な要素がテキスト内に含まれていると，共通した想起を行いやすいことを示している結果となっていると考えられる．

一方でこの結果の改善として検討すべき点もある．一つ目は検索結果の画像サイズのばらつきである．検索結果を使用する場合，各画像により解像度やサイズが異なる一方で，入力サイズは統一したサイズで入力する必要がある．そのため実験結果が解像度の影響を受けた結果となっている可能性がある．二つ目はクラスタリングやその可視化の方法である．データ数が少ない一方で，使用データの次元が大きいため，図中の位置で明確にクラスターが分かれていないように見える部分も多い．また DBSCAN や t-SNE において予め設定するパラメータの調整により結果が変化することが判明した．これらの設定により結果に影響を与えるため，適切なパラメータを調査していく必要があると考えられる．

4.2.2　自動生成手法

自動生成手法によりクラスタリングした結果を，図 8 に示す．表 1 の規則により分類した結果は表 3 であり，(1)(2)(4) が (d)，(3) が (b) となっている．すなわち (3) のみ画像から Google Bard を用いて説明文を生成した際に，ペアデータのテキストと類似するような説明文を複数生成したことを示している．よって，固有名詞を含むような特定しやすい要素が含まれているかどうかの判定については，こちらの手法で判別できる可能性がある．

ここで結果のクラスタリング自体について考察を行う．テキストから画像のクラスタリングの結果に注目すると，全て元のペアデータの画像は外れ値として扱われているものの，それ以外で生成されているクラスターに着目すると (1)(4) と (2)(3) で大きく 2 つに分けられるように見える．(1)(4) は全てが外れ値，すなわち互いに異なるような画像が生成される傾向がある一方で，(2)(3) は共に一つの大きなクラスター (図中の赤色) が形成され，似ている画像が生成されていることが読み取れる．今回の例の場合，(2) と (3) は”ラーメン”という単語が入っており，一般的に用いられて共通で想起しやすい用語がテキストに含まれている場合は，似た画像が生成されやすいことを，同手法により定量的に示唆することが可能であると考えられる．

こちらも結果改善のために検討すべき点がある．パラメータ調整は同じく検討していく必要がある．他にも自動生成手法では，アプリケーションの制約で，生成や認識できる対象に制限

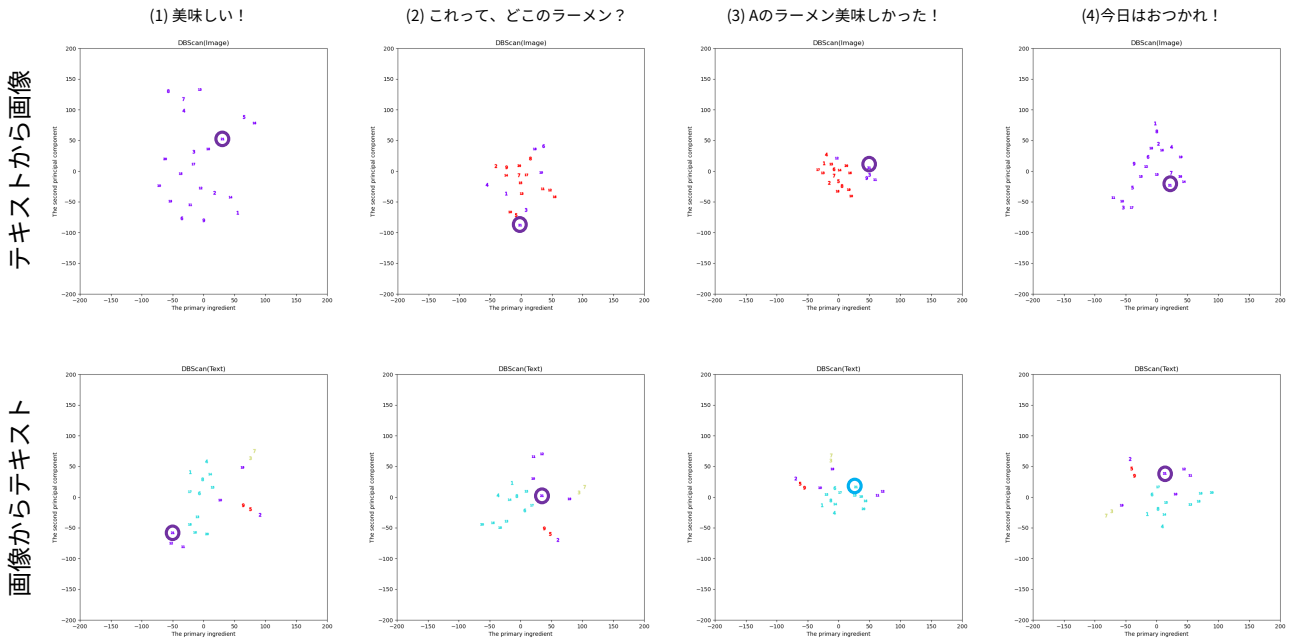


図 8　自動生成手法による結果 (図中の丸はペアデータを示し，紫色はクラスターの外れ値を示す．また実験のスクリプトでは (3) の”A”には実際の具体的な店名が与えられている)

表 3　結果 (自動生成手法)

		画像 P_i^k ✓	画像 P_i^k ×
テキスト P_t^k	✓		(3)
テキスト P_t^k	×		(1)(2)(4)

が存在したり，プロンプトの与え方により生成物に変化が生まれたりする問題があるため，アプリケーションの性能に左右される可能性がある．この部分について検討していくことが必要になると考えられる．

4.3　実際にポストされたデータでの分類結果

図 9 の左に挙げた X 上に実際にポストされているデータ [2]を用いて，提案手法による分類を行なった．4.1 節の方法で各パラメータを設定し，Stable Diffusion における英語のプロンプトは，”Cats who witnessed a stray cat fight.”を与えた．

結果は図 9 の右である．検索手法による結果が (b)，自動生成手法による結果が (d) となり，4.2 節の (1)(4) と同じ分類となった．クラスターの分布に注目すると，テキストから画像について，元のペアデータ以外に一つのクラスターを生成している．このことからテキストの曖昧性は比較的低いと考えられるが，そのクラスターに元の画像が含まれていないために，想起することは難しいと考えられる．画像からテキストについては，2 個以上のクラスターが存在しており，曖昧性がテキストに比べたら存在するが，検索からは想起しやすいという結果となった．

5　応　　用

提案手法を用いた応用例を複数挙げて議論する．分類結果に応じて，例えば画像 P_i^k が，テキスト P_t^k から想起される画像のクラスターに属する場合に，画像情報がなくてもテキストのみで情報が伝わる可能性が考えられる．これらを応用し，ここでは 2 例提案する．

1 つ目は，SNS 投稿上の条件に応じて画像を省略する画面 UI である．これは画像はテキストよりも通信量を消費したり，目立ったりするため，今回のシステムを利用して不必要であると判断された画像を自動的に省略したり，読み込まないようにしたりして，最低限の通信量で，効率よくデータを収集可能にするものである．また，教育上不適切な画像に対してのフィルタリングも応用例として考えられる．

2 つ目は，投稿支援や資料作成補助システムである．これは，スライド作成の際に挿入画像あるいは挿入テキストを提案するものである．挿入画像の提案は，短文のテキストから内容の意図に沿って，効果的に伝えることが可能となる適切な画像を提案するというものである．テキストの内容に対して補足する画像といった関連度が高いものを挿入したい時と，単にスライドの画面を強調したい時で，使用すべき画像が変化すると考えられ，今回のシステムはこのようなニーズに対して，解決できるようなシステムとなる可能性を示唆している．一方で，添付の画像から意図に沿ったテキスト作成を行う応用も考えられる．これは添付した画像の意図に応じて，書くべきテキストの内容を提案するものである．Microsoft 社が開発・販売している Powerpoint では，デザイナーという機能が 令和 6 年 3 月 21

2：https://twitter.com/catmotimaru/status/1660934386043297801?s=61&t=4-VnYoikuUZUGrtFBOOmYQ(最終閲覧日: 令和 6 年 3 月 21 日)

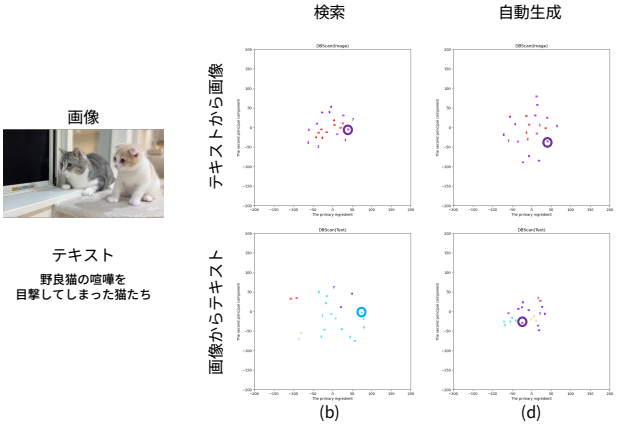


図 9　実例による結果

日現在存在し，加工パターンとなるデザインを複数推薦するということを可能にしているが，今回，根本となる挿入テキストや画像自体を推薦するという点で，目的が異なると考えられる．

6 ま と め

本稿では，SNS 投稿における画像とテキストの関連度を分類するため，既存の検索や生成 AI を活用して分類する手法を提案している．その際に SNS 投稿でよく使用される画像ーテキストの例や実際のデータを用いて実験を行い，画像に含まれているものがテキストに単語として含まれている場合に，特に異なるクラスター結果が得られ，テキストの種類によって異なる結果が得られるような本手法の有用性について示唆した．

しかし画像とテキストの関連性を扱う上で，多くの課題が残されている．検索手法と自動生成手法の併用法の検討や，定量的評価として既存の定量的尺度を活用し，各データがもつ情報を評価する方法を通して，分類手法の確立と更なる分析を行う必要がある．この際にデータ対象を絞りつつ，効率的な実装と実例での実験の充実の検討も考えられる．また別のアプローチとして，SNS には画像とテキストのペアが数多くあり，それらの関連性をより明確化するため，画像やテキストをノード，ペアとなっているそれらの間をエッジとして，グラフニューラルネットワーク (GNN) [17] を用いて分類する手法も考えられる．画像ーテキスト間のソーシャルメディア分析の GNN の適用は，感情分類 [18] に使用される等の例があり，マルチモーダルのデータに対しても注目されている手法である．これらの手法については，今後の展望として取り組んでいきたいと考えている．

謝　　辞

本研究の一部は科研費（23H03404）による．

文　　献

[1] Kevan Lee. What 1 Million Tweets Taught Us About How People Tweet Successfully. Buffer, 2015.
[2] Leiquan Wang, Xiaoliang Chu, Weishan Zhang, Yiwei Wei, Weichen Sun, and Chunlei Wu. Social Image Captioning: Exploring Visual Attention and User Attention. *Sensors*, 18(2):646, 2018.
[3] Taka Maenishi and Keishi Tajima. Identifying Tags Describing Image Contents. In *HT*, pages 297–298. ACM, 2019.
[4] Marie Katsurai, Takahiro Ogawa, and Miki Haseyama. A Cross-Modal Approach for Extracting Semantic Relationships Between Concepts Using Tagged Images. *IEEE Trans. Multim.*, 16(4):1059–1074, 2014.
[5] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *AAAI*, pages 5674–5681. AAAI Press, 2018.
[6] Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER. In *ACM Multimedia*, pages 3983–3992. ACM, 2022.
[7] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In *AAAI*, pages 13860–13868. AAAI Press, 2021.
[8] Alakananda Vempala and Daniel Preotiuc-Pietro. Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts. In *ACL (1)*, pages 2830–2840. Association for Computational Linguistics, 2019.
[9] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pages 3156–3164. IEEE Computer Society, 2015.
[10] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.
[11] Zhixin Li, Xiumin Xie, Feng Ling, Huifang Ma, and Zhongzhi Shi. Matching Images and Texts with Multi-head Attention Network for Cross-media Hashing Retrieval. *Eng. Appl. Artif. Intell.*, 106:104475, 2021.
[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, pages 226–231. AAAI Press, 1996.
[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net, 2021.
[14] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *CoRR*, abs/2006.03677, 2020.
[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.
[16] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics, 2019.
[17] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.
[18] Wenxiong Liao, Bi Zeng, Jianqi Liu, Pengfei Wei, and Jiongkun Fang. Image-Text Interaction Graph Neural Network for Image-Text Sentiment Analysis. *Appl. Intell.*, 52(10):11184–11198, 2022.

コロナ禍の話題表す俳句生む

金子　卓矢†　　横山　昌平†,††,†††

† 東京都立大学　〒 191–0065 東京都日野市旭が丘 6-6
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2-1-2
††† 東京大学　〒 153-8505 東京都目黒区駒場 4-6-1
E-mail: †kaneko-takuya@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　新型コロナウイルス感染症（COVID-19）の流行により日常の生活は大きく変化した．そして，マスクの不足や外出自粛など様々な話題が発生した．コロナ禍における日常の話題は変化する．そこで，ある一日の話題を SNS から見つけ出し短くまとめることにより，一目で一日の出来事を知ることが出来るようにする．そして，ただ短い文にまとめるのではなく，俳句とする事で印象に残りやすくなる．筆者らはこれまでに Twitter（現 X）へ投稿された文の単語をトピック分類し，トピックごとに五七五の生成を行った．しかし，季語は含まれていなかったため俳句とはならなかった．本研究では，Twitter の全量データを使用し，ツイートから単語の共起ネットワークを作成してクラスタリングし季語と関連する単語を見つける．季語を含めて単語を組み合わせることにより，一日の特徴を示す俳句を生成した．

キーワード　SNS，Twitter，新型コロナウイルス，文書要約，情報抽出

1 はじめに

2019 年 12 月に中国の武漢市で新型コロナウイルス感染症（COVID-19）の感染者の第一例目が報告された後，世界的に流行し日本でも 2020 年初頭から流行したことにより日常の生活は大きく変化した．例えば，2020 年 4 月に 1 回目の緊急事態宣言が発令され，2020 年 7 月に開催予定だった 2020 東京オリンピックは 2021 年に延期された．そして，日常の生活においても，マスクの不足や外出自粛など様々な話題が発生した．

5G やスマートホンの普及により，SNS の利用者は増加している．Facebook や Instagram，そして Twitter（現 X）など，さまざまな利用者がいる．それに伴い，SNS へ多くの情報が投稿されている．

俳句は，季語を含み，五拍・七拍・五拍の十七音を基本とした日本の詩である．しかし，季語を含まない無季の俳句や，五七五の定型にとらわれない，自由律なども存在する．

SNS には毎日膨大なデータが存在する．例えば「コロナ リモートワーク オンライン授業 オリンピック」というキーワードで検索を行うと，1 回目の緊急事態宣言が発令された 2020 年 4 月 7 日の一日のデータ量は 670MB となる．我々は，そのような膨大なデータから，ある一日の話題を SNS から見つけ出し短くまとめることにより，一目で一日の出来事を知ることが出来るようにする．そして，ただ短い文にまとめるのではなく，俳句という十七音で表現する事で印象に残りやすくなる．筆者らはツイートの単語をトピック分類し，それぞれのトピックの名詞を組み合わせ，名詞の間に単語を当てはめることによって五七五の生成を行った [1]．しかし，季語が含まれていなかったため俳句とはなっていなかった．本研究では，Twitter の全量データを使用して季語と共起する単語を組み合わせることによって一日の話題を表す俳句の生成を行っている．

2 関連研究

Rafal らは Web の文書を資源とし，オリジナルのテンプレートに単語を当てはめることによって俳句の生成を行った [2]．Yael らは俳句コーパスを言語資源とし，連想語ネットワークを用いて俳句の単語を決定し，品詞 N-gram によって俳句の生成を行った [3]．本研究は，Tweet のデータからネットワークを作成し，N-gram モデルを用いて俳句の生成を行っている．

安部ら [4] は論文の要約となる 575 の生成を行った．特徴語を抽出し，575 のテンプレートに当てはめることにより生成し，タイトルやアブストラクトとの cos 類似度，係り受けスコア，ネガポジスコアを用いてランク付けしている．本研究でも，ツイートからある一日の特徴語を抽出し，俳句のテンプレートに当てはめることで生成している．異なる点として，一日のツイートのタイトルや要約などは存在しないため，一日分のデータから偶然俳句になっているものを取得し，テキストランクで重要なものをランク付けしたものとの cos 類似度を利用している．

山根ら [5] は Facebook のファンページの投稿を取得し，モデルとなるキャッチコピーと投稿に含まれる単語からテーマを選択し，カテゴリが合致する単語を探索し候補を生成している．生成された候補に対して，Web 日本語 N グラムを用いて文の自然さをスコア付け，モデルキャッチコピーに含まれる単語がどれだけ多いかで意味の関連性をスコア付け，Facebook へ投稿された文章に対して bag-of-words モデルを用いて線形回帰を行うことで恰好度合いをスコア付けし生成したキャッチコピーをランク付けしている．本研究でも，文の自然さと cos 類似度を用いた関連性をスコア付けしている．

3 提案手法

1 日の話題は日々変化するため，新たな話題にも対応する必要がある．言語モデルによる生成はすでに学習された結果を用いるため，時事問題に対応するには適さないと考えた．本研究では，ツイートから取得した単語を組み合わせることで俳句の生成を行った．処理の概要を図 1 に示した．

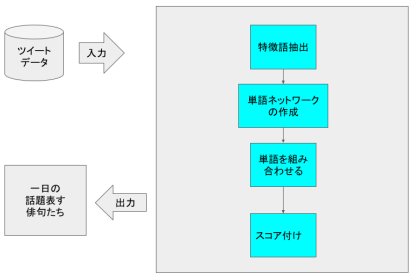


図 1　処理の流れ

3.1 データの取得と前処理

はじめに，株式会社 NTT データ「なずきのおと」を用いて Twitter からキーワードを指定してデータの取得を行う．取得する際，リツイートは同じツイートのデータのため除外する．

次に取得したツイートの前処理を行う．アフィリエイトのツイートは同じような投稿をたくさんしており，出来事を表すツイートではないため除外する．本研究では，URL が含まれているツイートを取り除くことでアフィリエイトのツイートを除外している．ツイートの絵文字や記号を削除し，数字を 0 に置き換える．

3.2 単語のネットワークの作成

ツイートの形態素解析を行い，ある単語の次にどの単語が来るのかというネットワークを作成する．形態素解析器は janome，辞書は Neologd [6] を使用した．

次に単語の取得を行う．DF-IDF（Document Frequency-Inverse Document Frequency）を用いて名詞，動詞，形容詞，副詞のそれぞれで上位の単語を取得する．DF-IDF は一つのツイートを一つのドキュメントとして扱い，一日分のツイートにその単語が出現するドキュメントの数の DF（Document Frequency），他の日と比較したときのレア度 IDF（Inverse Document Frequency）をかけ合わせた値で，TF-IDF（Term Frequency-Inverse Document Frequency）のような指標である．ある単語 w の DF-IDF 値 $DF-IDF_w$ は $N_w(t)$ を期間 t における単語 w が出現するドキュメントの数，$N(t)$ を期間 t のドキュメントの数，T を期間の数として式 (1) のように表される．

$$DF-IDF_w = \frac{N_w(t)}{N(t)} \times \log\left(\frac{\sum_{i=1}^{T} N(i)}{\sum_{i=1}^{T} N_w(i)}\right) \tag{1}$$

TF-IDF との違いは，単語 w を含むドキュメントの数だけをカウントすることである．1 つの短いツイートでは同じイベントに関して複数出現するためである [7]．

取得した単語を用いて共起ネットワークを作成する．単語間のノードの重みは jaccard 係数を使用し，しきい値を定め，しきい値以上のノードを作成する．jaccard 係数はキーワード間の関係の強さを表す係数で，$n(w)$ を単語 w を含むツイートの数として単語 w_1, w_2 の jaccard 係数 $J(w_1, w_2)$ は式 (2) で示される．

$$J(w_1, w_2) = \frac{n(w_1 \cap w_2)}{n(w_1 \cup w_2)} \tag{2}$$

作成した共起ネットワークをクラスタリングをすることで単語のコミュニティを検出する．クラスタリングのアルゴリズムは Clauset-Newman-Moore greedy modularity maximization [8] を使用した．

3.3 俳句の候補の生成

俳句の候補を生成する際，はじめに名詞の組み合わせを見つける．各クラスタから名詞を取得し，実際のツイートに出現する組み合わせを見つける．このときに，季語が含まれない組み合わせは俳句とならないため除外する．季語のデータは現代俳句協会の現代俳句データベース [1]から取得した．

取得した名詞の間にどのような単語が入るか，単語のネットワークから出現数の多いものから当てはめ，五拍または七拍となる場合は五七五の候補とする．不自然な文を除外するため，累積確率を設定し，しきい値を超えない範囲で生成する．

3.4 スコア付け

生成した俳句を MLM(Masked Language Model) [10] を用いてスコア付けし，文として成り立っているかをスコア付けする．使用した MLM は PLL（pseudo-log-likelihood scores）と呼ばれるもので，各単語をマスクしたときの単語の予測確率の対数尤度の和で文 W の PLL の値 $PLL(W)$ は式 (3) で表される．

$$PLL(W) = \sum_{t=1}^{|W|} \log(P_{MLM}(w_t|W_{\setminus t})) \tag{3}$$

MLM は現代俳句データベースの俳句を使用し，transformers [2] を用いてファインチューニングを行う．

次に実際のツイートから偶然俳句になっている部分を抜き出す．そして，取得した俳句についてテキストランク [9] で上位のものを取り出す．取得した偶然俳句になっているものと生成した俳句の類似度を求める．BoW（Bag of Words）でベクトル化して cos 類似度を求める．

文の自然さを正規化した値 $S_{natural}$ と類似度 S_{sim} を用いて式 (4) のように調和平均を生成した俳句のスコアとしてランキングする．

$$score = \frac{2}{\frac{1}{S_{natural}} + \frac{1}{S_{sim}}} \tag{4}$$

1：https://haiku-data.jp/index.php

2：https://github.com/huggingface/transformers/blob/main/examples/tensorflow/language-modeling/run_mlm.py

4　俳句の生成結果

本節では，俳句の生成結果について述べる．はじめに，検索する季語について，出現頻度を調べた．出現数が多い季語を用いてツイッターの全量データからツイートデータの取得をした．

取得したデータを用いて俳句の生成を行った．本実験では春と冬でいくつか生成を行った．

4.1　検索する季語のキーワード

はじめに，取得するツイートのキーワードを決める．2020 年 4 月 1 日から 2021 年 3 月 31 日までの期間で「コロナ，リモートワーク，オンライン授業，オリンピック」のキーワードで検索し，形態素解析を行い出現する頻度をそれぞれの月の 1 日，10 日，20 日について数え，1 年で出現頻度の多い上位 10 個を図 2,3 にまとめた．データの件数は 102871199 である．冬の季語は冬に多く出現し，春の季語は春に多く出現するなど季語がその季節に多く出現していることを確認した．本研究では，ツイートから検索する際はこれらの季語を使用することにした．

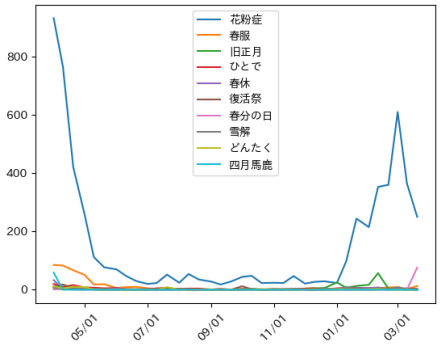


図 2　春 の 季 語

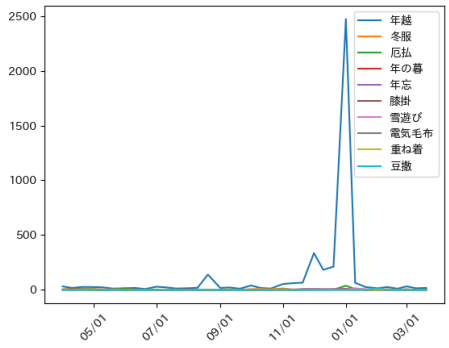


図 3　冬 の 季 語

4.2　春 の 俳 句

ツイッターの全量データから 2020 年 2 月 1 日 0:00 から 2020 年 6 月 1 日 0:00 までの期間で「花粉症，春服，旧正月，ひとで，春休，復活祭，春分の日，雪解，どんたく，四月馬鹿」をキーワードとして指定しデータを取得し，2020 年 3 月 30 日と 2020 年 4 月 7 日で俳句の生成を行った．ツイートの件数は 4011000 である．

4.2.1　2020 年 3 月 30 日の俳句

2020 年 3 月 30 日の俳句の生成を行った．ツイートの件数は 37649 である．上位の単語について，共起ネットワークを作成すると図 4 のようになり，クラスタリングした単語を用いて生成された俳句を表 1, 表 2 にまとめた．

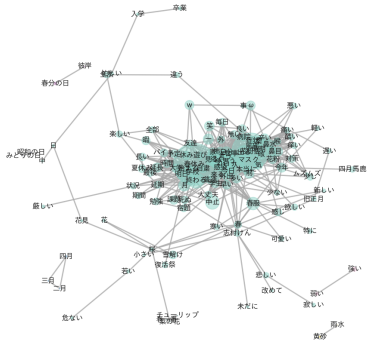


図 4　2020 年 3 月 30 日の共起ネットワーク

表 1　2020 年 3 月 30 日の俳句（community0）

俳句	スコア
春休み延長ですよ春なので	0.27728778356135786
友達や春服予定中止とか	0.11775884662136407
春なので今日まででした春服か	0.08353131691431036
春なので今日まででした春服よ	0.04582628182258173
春なので今日まででした春服や	0.045773929677399974
春なので今日まででした春服は	0.04389439533999852
春なので今日まででした春服に	0.043741773067803455
春なので今日まででした春服が	0.04363402535235429
春なので今日まででした春服と	0.04343341384959162
春なので今日まででした春服だ	0.04164007201955587

表 2　2020 年 3 月 30 日の俳句（community1）

俳句	スコア
咳だけの花粉症とか風邪だって	0.5966854732395146
花粉症対策だけど薬なの	0.5932885501316374
鼻水や花粉症とか薬なの	0.5902790654213662
花粉症対策だけどマスクなの	0.5891888609972559
鼻水や花粉症だろ薬なの	0.5855290414478035
花粉とは花粉症だろ薬なの	0.5778421986411314
咳だけの花粉症です鼻水や	0.5743360697046738
病院や花粉症とか薬なの	0.5723071991865469
咳だけの花粉症とか症状か	0.5714055441124398
咳だけの花粉症やら風邪だって	0.569325554332995

4.2.2 2020 年 4 月 7 日の俳句

2020 年 4 月 7 日の俳句の生成を行った．ツイートの件数は 31568 である．上位の単語について，共起ネットワークを作成すると図 5 のようになり，クラスタリングした単語を用いて生成された俳句を表 3 にまとめた．

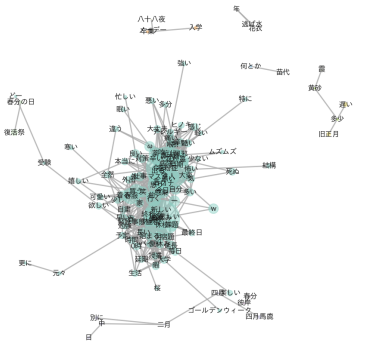


図 5　2020 年 4 月 7 日の共起ネットワーク

表 3　2020 年 4 月 7 日の俳句

俳句	スコア
症状は花粉症とかアレルギー	0.6434226935495598
喉からの花粉症です薬には	0.6337799564994161
アレルギー対策じゃない花粉症	0.6260175738393889
喉からの鼻水なのに花粉症	0.6222742042252423
アレルギー症状ですな花粉症	0.6214147612032571
喉からの花粉症です薬かよ	0.6189108731519113
アレルギー花粉症です薬には	0.617808922970466
花粉症症状だった喉からの	0.617544297440762
喉からの花粉症かな症状は	0.6170565860952101
症状は花粉症です薬だけ	0.6163229803137907

4.3 冬の俳句

ツイッターの全量データから 2020 年 11 月 7 日 0:00 から 2021 年 2 月 3 日 0:00 までの期間で「年越，冬服，厄払，年の暮，年忘，膝掛，雪遊び，電気毛布，重ね着，豆撒」をキーワードとして指定しデータを取得し，2020 年 12 月 24 日で俳句の生成を行った．ツイートの件数は 4487790 である．

また，ツイッターの全量データから 2022 年 11 月 7 日 0:00 から 2023 年 2 月 4 日 0:00 までの期間で「年越，冬服，厄払，年の暮，年忘，膝掛，雪遊び，電気毛布，重ね着，豆撒」をキーワードとして指定しデータを取得し，2022 年 12 月 24 日で俳句の生成を行った．ツイートの件数は 4119218 である．

4.4 2020 年 12 月 24 日の俳句

2020 年 12 月 24 日の俳句の生成を行った．ツイートの件数は 34993 である．上位の単語について，共起ネットワークを作成すると図 6 のようになり，クラスタリングした単語を用いて生成された俳句を表 4 にまとめた．

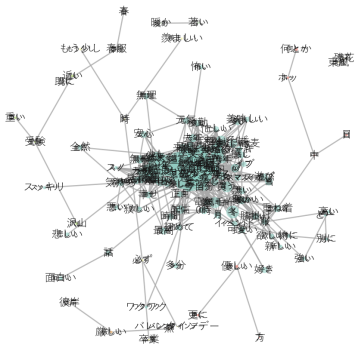


図 6　2020 年 12 月 24 日の共起ネットワーク

表 4　2020 年 12 月 24 日の俳句

俳句	スコア
クリスマス休みないので年越しは	0.5994959184357518
クリスマス年末年始年越しは	0.5805451412672724
クリスマス休みないので年越しも	0.5142379134510321
クリスマス年末感じ年越しは	0.5037275859874106
クリスマス自分だけです年越しは	0.5028922211493061
年越しも夜勤今年はクリスマス	0.5012318627911254
年越しよ夜勤今年はクリスマス	0.4986505146841664
年越しや夜勤今年はクリスマス	0.4946108787130022
クリスマス休みないので年越しよ	0.4944346930058053
クリスマス休みないので年越して	0.49246946117236795

4.4.1 2022 年 12 月 24 日の俳句

2022 年 12 月 24 日の俳句の生成を行った．ツイートの件数は 42176 である．上位の単語について，共起ネットワークを作成すると図 7 のようになり，クラスタリングした単語を用いて生成された俳句を表 5 にまとめた．

表 5　2022 年 12 月 24 日の俳句

俳句	スコア
人じゃない雪遊びたの今年だよ	0.5030908243062516
人じゃない雪遊びたの笑だから	0.47969302543005365
前だった雪遊びたの今日ですね	0.4710351150653959
人だった雪遊びたの今年だよ	0.46986814761822787
人じゃない雪遊びたの今年には	0.46213037862785167
人じゃない雪遊びたの今年だな	0.4613846684119634
人だった雪遊びたの笑だから	0.4511193903882096
人ですね雪遊びたの今年には	0.44448446803282454
人じゃない雪遊びたの今年では	0.44302039750469024
人かでは雪遊びたの笑だから	0.4397810864170769

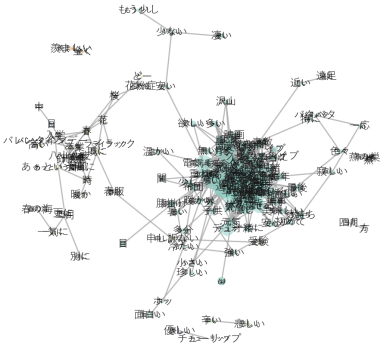


図 7　2022 年 12 月 24 日の共起ネットワーク

5　生成した俳句の評価

生成した俳句について定性的な評価と定量的な評価を行った．定性的な評価ではいくつかの日付で生成した結果について，それぞれの日を表す内容の俳句となっているか，スコア付けについて自然さだけでスコア付けしたときと，偶然俳句との類似度も考慮したときの違いが表れているかを確認した．

定量評価ではベースラインを作り，生成した俳句についてパープレキシティを測定し，ベースラインと提案手法で比較し，有効性があるかを確認した．なお，2020 年 3 月 30 日のベースラインは花粉症に関する話題のみ生成されたため，比較も花粉症の話題のみを使用している．

5.1　日ごとに特徴が表れているか

本節では，生成された俳句に特徴が表れているかを確認する．

生成された俳句について確認すると，表 2，3 より花粉症の季節は花粉症の症状についての俳句が生成されている．また，表 1 を確認すると春休みが延長という俳句が確認された．この時期は新型コロナウイルスが流行していることにより，緊急事態宣言が発令されるのではということや，春休みが延長になるのではないかという話題があったため，それらが俳句として表れている．

冬の季節について確認すると，表 4 よりクリスマスだが休みがないという俳句が生成されている．また，表 5 では雪遊びに関する俳句が生成された．2022 年 12 月 24 日の天気を確認すると，東北南部や北陸を中心に雪雲が発生し，広い範囲で積雪が確認されている．クリスマスイブだが年によって異なる俳句が生成されている．

同じ季節でも異なる内容の俳句が表れているため，その日の特徴を表していると考えられる．

5.2　スコア付けが適切か

本節では，スコア付けについてを確認する．生成された俳句について，cos 類似度は使用せず，自然さだけでランク付けした結果を表 6，7，8，9 にまとめた．

表 6　2020 年 3 月 30 日の自然さでランク付けした俳句

俳句	スコア
花粉症対策だけど風邪でしょう	-5.774660581424007
花粉症対策だけど風邪だろう	-6.468786114842715
目だけでも花粉症とか風邪でしょう	-6.830243430755627
目だけでも花粉症やら風邪でしょう	-7.21975187804139
目だけでも花粉症とか風邪だろう	-8.346620436924573
症状が花粉症とか風邪だって	-9.38896544972069
目だけでも花粉症やら風邪だろう	-10.730286038810883
症状が花粉症とか風邪だろう	-11.06299117569503
症状が花粉症とか風邪でしょう	-11.343168413808051
目だけでも花粉症かな風邪でしょう	-11.976600829244818

表 7　2020 年 4 月 7 日の自然さでランク付けした俳句

俳句	スコア
花粉症症状じゃない大丈夫	-1.9946475411494755
アレルギー症状だった花粉症	-2.5144346084253386
ヒノキでも花粉症とか大丈夫	-2.8888955132586878
アレルギー症状じゃない花粉症	-3.732864956633372
花粉症対策じゃない人だった	-3.7659878829067215
花粉症症状じゃない人だった	-4.000706258310481
花粉症症状じゃない人だけに	-4.53657587336771
花粉症対策じゃない人だけに	-4.693200081388015
ヒノキでも花粉症かな大丈夫	-4.7119130217206635
花粉症症状だった今じゃない	-4.838320406276575

表 8　2022 年 12 月 24 日の自然さでランク付けした俳句

俳句	スコア
クリスマス年越しだって仕事です	-3.103599599403534
クリスマス年末年始仕事です	-4.049652765785788
クリスマス年末年始仕事かよ	-4.284872844040652
クリスマス年末年始仕事って	-4.4257436692073515
クリスマス年越しだろう仕事です	-4.427822040938859
クリスマス休みないので年越しよ	-5.2977495152051794
クリスマス年末年始年越しよ	-5.415800938344219
クリスマス年末年始年越しか	-5.792534622278801
クリスマス休みないので年越して	-6.1482848455065815
クリスマス年越し仕事予定です	-6.2432122223886966

表 3 では花粉症について「喉からの症状」という内容が表れているが，表 7 ではこの内容は表れていない．また，自然さだけでランク付けした表 8 と表 9 を比較すると，どちらもクリスマスや年越しの話題となっているが，cos 類似度を考慮することによって表 4 と表 5 のように，クリスマスに関することや，雪に関することなど異なる話題が表れている．

表 9　2022 年 12 月 24 日の自然さでランク付けした俳句

俳句	スコア
クリスマス年越しでした今じゃない	-7.083073015602705
クリスマス年越しでした人じゃない	-8.068679159350353
クリスマス年越しでした安心だ	-8.59893204245911
クリスマス年越しでした今年だな	-10.370784581819981
クリスマス年末だった年越しぞ	-10.435715100380552
年末や仕事ないから雪ですね	-10.878887007693177
クリスマス年越しでした今年だよ	-10.937678110146361
クリスマス年越しでした今年には	-11.419863474434871
クリスマス仕事年越し休みかな	-11.501446887919371
クリスマス年越しでしょう今年では	-11.523642087232881

5.3　ベースラインとの比較

提案手法で生成した俳句の有効性を調べるためにベースラインを作成した．ベースラインとなる俳句は単語の頻度のみで単語を取得し，共起ネットワークは作成せずに単語を組み合わせることで生成した．生成した俳句について式 (4) を用いてスコア付けした．

生成した俳句のパープレキシティを測定した．パープレキシティとは式 (5) で表され，正しい単語を正しく予測できていれば値が小さくなる．

$$ppl = \exp(-\frac{1}{N}\sum_{n}\sum_{k} t_{n,k}\log(p_{model}(y_n,k))) \qquad (5)$$

n はデータセット中の n 番目の単語を表し，$t_{n,k}$ は n 番目の単語の正解ラベル，$p_{model}(y_n,k)$ は n 番目の単語に k を予測する確率を表す．

2020 年 3 月 30 日のツイートでベースラインで生成した俳句のパープレキシティを表 10 に，提案手法で生成した俳句のパープレキシティを表 11 にまとめた．使用したモデルは GPT2 である．

表 10　2020 年 3 月 30 日のベースラインの俳句とパープレキシティ

俳句	パープレキシティ
人多い花粉症とかのお知らせ	70.80729675292969
人多い花粉症とかの弊害	204.97067260742188
人多い花粉症とかのはじまり	142.12899780273438
人多い花粉症とかの戦い	217.99453735351562
人多い花粉症とかの質問	89.52516174316406
人多い花粉症とかの問題	119.3707504272461
人多い花粉症とかの心配	142.9925994873047
人多い花粉症とかの限界	147.67391967773438
人多い花粉症とかの感想	154.92129516601562
人多い花粉症とかの楽しみ	126.67095184326172

2020 年 4 月 7 日のツイートでベースラインで生成した俳句のパープレキシティを表 12 に，提案手法で生成した俳句のパープレキシティを表 13 にまとめた．

2020 年 12 月 24 日のツイートでベースラインで生成した俳句のパープレキシティを表 14 に，提案手法で生成した俳句のパープレキシティを表 15 にまとめた．

2022 年 12 月 24 日のツイートでベースラインで生成した俳句のパープレキシティを表 16 に，提案手法で生成した俳句のパープレキシティを表 17 にまとめた．

表 11　2020 年 3 月 30 日の提案手法の俳句とパープレキシティ

俳句	パープレキシティ
咳だけの花粉症とか風邪だって	52.73542404174805
花粉症対策だけど薬なの	178.2971649169922
鼻水や花粉症とか薬なの	72.9620132446289
花粉症対策だけどマスクなの	211.55319213867188
鼻水や花粉症だろ薬なの	60.17658615112305
花粉とは花粉症だろ薬なの	58.56978988647461
咳だけの花粉症です鼻水や	53.96765899658203
病院や花粉症とか薬なの	133.95779418945312
咳だけの花粉症とか症状か	58.740623474121094
咳だけの花粉症やら風邪だって	50.38547897338867

表 12　2020 年 4 月 7 日のベースラインの俳句とパープレキシティ

俳句	パープレキシティ
花粉症コロナ対策の一環	949.025634765625
花粉症コロナ対策のお知らせ	748.3439331054688
花粉症コロナ対策の学校	824.938720703125
花粉症コロナ対策の我が家は	1095.0545654296875
花粉症コロナ対策の宿題	1513.056396484375
花粉症コロナ対策のあやめは	1132.646728515625
花粉症マスク着用のお知らせ	240.11126708984375
花粉症コロナ対策の疑い	1067.248779296875
花粉症マスク着用の我が家は	427.941162109375
花粉症マスク着用のカップル	436.80560302734375

表 13　2020 年 4 月 7 日の提案手法の俳句とパープレキシティ

俳句	パープレキシティ
症状は花粉症とかアレルギー	47.05629348754883
喉からの花粉症です薬には	60.10420608520508
アレルギー対策じゃない花粉症	99.69890594482422
喉からの鼻水なのに花粉症	52.45611572265625
アレルギー症状ですな花粉症	130.3199005126953
花粉症症状今年大丈夫	71.59452056884766
喉からの花粉症です薬かよ	57.400753021240234
アレルギー花粉症です薬には	356.3583984375
花粉症症状だった喉からの	75.50884246826172
喉からの花粉症かな症状は	44.21377944946289

表 14　2020 年 12 月 24 日のベースラインの俳句とパープレキシティ

俳句	パープレキシティ
さんどうかクリスマスの日年越しは	86.77295684814453
さんこれはクリスマスの日年越して	31.076602935791016
さんこれはクリスマスの日年越しだ	39.959205627441406
さんこれはクリスマスの日年越しよ	31.853546142578125
さんこれはクリスマスの日年越してぇ	37.15984344482422
さんこれはクリスマスの日年越しかぁ	38.56959915161133
さんこれはクリスマスの日年越しやぁ	35.793922424316406
さんでしたクリスマスの日年越しは	39.83354187011719
さんこれはクリスマスの日年越しじゃ	27.3111572265625
さんこれはクリスマスの日年越しも	41.1404914855957

ベースラインと提案手法のスコアが上位 10 個の俳句のパープレキシティに有意差があるか t 検定を行った．帰無仮説を

表 15　2020 年 12 月 24 日の提案手法の俳句とパープレキシティ

俳句	パープレキシティ
クリスマス休みないので年越しは	46.67136764526367
クリスマス年末年始年越しは	306.1989440917969
クリスマス休みないので年越しも	47.536865234375
クリスマス年末感じ年越しは	604.076171875
クリスマス自分だけです年越しは	37.278438568115234
年越しも夜勤今年はクリスマス	48.932037353515625
年越しよ夜勤今年はクリスマス	45.092891693115234
年越しや夜勤今年はクリスマス	48.10463333129883
クリスマス休みないので年越しよ	37.35902404785156
クリスマス休みないので年越して	35.60931396484375

表 16　2022 年 12 月 24 日のベースラインの俳句とパープレキシティ

俳句	パープレキシティ
クリスマス服着せました雪遊び	94.67350006103516
んえしたいの道具だった雪遊び	41.53044891357422
んみたいなのひとつだった雪遊び	30.29187774658203
んえしたいのひとつだった雪遊び	33.091270446777344
遊び中のひとつだった雪ですか	33.9489860534668
遊び中の道具だった雪ですか	41.86832046508789
遊び中の道具だった雪ですね	35.04990768432617
遊び中のひとつだった雪ですね	28.50712776184082
んみたいなの道具だった雪遊び	39.330230712890625
んとこ人の道具だった雪遊び	48.21969223022461

表 17　2022 年 12 月 24 日の提案手法の俳句とパープレキシティ

俳句	パープレキシティ
人じゃない雪遊びたの今年だよ	24.082517623901367
人じゃない雪遊びたの笑だから	45.18269729614258
前だった雪遊びたの今日ですね	25.513792037963867
人だった雪遊びたの今年だよ	32.03742980957031
人じゃない雪遊びたの今年には	24.315196990966797
人じゃない雪遊びたの今年だな	29.329195022583008
人だった雪遊びたの笑だから	82.39366912841797
人ですね雪遊びたの今年には	32.76719284057617
人じゃない雪遊びたの今年では	24.896682739257812
人かでは雪遊びたの笑だから	90.53710174560547

「ベースラインと提案手法についてパープレキシティに差がない」，対立仮説を「提案手法のほうがベースラインよりパープレキシティの値が小さい」，有意水準を 0.05 として片側検定を行った．パープレキシティの t 統計量と p 値を表 18 にまとめた．表 18 より 2020 年 3 月 30 日と 2020 年 4 月 7 日の p 値は

表 18　t 統計量と p 値

日付	t 統計量	p 値
2020 年 3 月 30 日	2.0542242208991324	0.02791457411401389
2020 年 4 月 7 日	5.8667758858385755	0.00007762823029457419
2020 年 12 月 24 日	-1.424581301144645	0.9062485528576858
2022 年 12 月 24 日	0.1557686091064528	0.43902771860006284

どちらも有意水準の 0.05 より小さいため帰無仮説は棄却し，提案手法のほうがベースラインよりパープレキシティの値が小さいと判断される．

一方，2020 年 12 月 24 日と 2022 年 12 月 24 日は p 値がどちらも有意水準の 0.05 より大きいため，帰無仮説は棄却されず，提案手法のほうがベースラインよりパープレキシティの値が小さいとは言えないと判断される．しかし，生成された俳句を確認したところ，2020 年 12 月 24 日の俳句はベースラインでは「さん」という単語が文頭にあるが，これは不自然である．定量的な評価では有意差を確認できなかったが，定性的に評価すると提案手法のほうが文は自然であるといえる．

6　おわりに

本研究では，一日の話題を表す俳句の生成を行い，コロナ禍の特徴を調べた．ツイートから単語を取得し，共起ネットワークを作成して俳句に使う単語を決定した．単語のネットワークから，名詞の間に入る単語を見つけることで，俳句の生成を行った．その結果，花粉症に関する話題や梅雨開けに関する話題の俳句が生成された．

生成された俳句について，パープレキシティを計算してベースラインと比較を行った．t 検定を行うとベースラインと提案手法のパープレキシティに有意差があることを確認した．

謝　　辞

本研究は，内閣府 COVID-19 AI・シミュレーションプロジェクトの一環として実施され，JST，CREST，JPMJCR19A4 の支援を受けたものである．

文　　献

[1] 金子卓矢，横山昌平．一日のトピック示す五七五: ソーシャルビッグデータを用いた一日の出来事を表す五七五の生成．第 22 回情報科学技術フォーラム（FIT2023）．pp263–264
[2] Rzepka, Rafal, and Kenji Araki. ”Haiku generator that reads blogs and illustrates them with sounds and images.” Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI 2015). IJCAI, 2015.
[3] Yael Netzer et al.“Gaiku: Generating Haiku with Word Associations Norms”. Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, pp. 32 – 39, 2009.
[4] 安部 文紀 寺田 実，575 の音韻的読みやすさを付与した学術論文の要約文自動生成手法，DEIM 2018
[5] 山根 宏彰，萩原 将文、ウェブにおけるユーザの嗜好を反映させたキャッチコピー自動生成システム、日本感性工学会論文誌、13 巻（2014 年）4 号
[6] 佐藤敏紀, 橋本泰一, 奥村学．単語分かち書き用辞書生成システム NEologd の運用 — 文書分類を例にして —．自然言語処理研究会研究報告（2016）．NL-229-15.
[7] Weng, Jianshu, and Bu-Sung Lee. ”Event detection in twitter.” Proceedings of the international aaai conference on web and social media. Vol. 5. No. 1. 2011.
[8] Clauset, A., Newman, M. E., Moore, C. “Finding community structure in very large networks.” Physical Review E 70(6), 2004
[9] Erkan, Günes, and Dragomir R. Radev. ”Lexrank: Graph-based lexical centrality as salience in text summarization.” Journal of artificial intelligence research 22 (2004): 457-479.
[10] Julian Salazar, Davis Liang, Toan Q. Nguyen, Katrin Kirchhoff. ”Masked Language Model Scoring”, ACL 2020.

同一ハッシュタグ投稿に着目した対照学習による トピック類似性を捉えた文ベクトルの獲得

仲田　明良†　　狩野　芳伸†

† 静岡大学 情報学部　〒 432–8011 静岡県浜松市中区城北 3 丁目 5–1
E-mail: †{anakada,kano}@kanolab.net

あらまし　SNS 上の投稿に付随するハッシュタグは，投稿者がコンテンツに関連付けたり，トピックを指定したりする手段として利用されている．本研究では，投稿文に加え付随したハッシュタグの収集を行い，投稿文ペアが同じハッシュタグのものか異なるハッシュタグのものか，ハッシュタグ同一性に基づいた対照学習を用いて，SNS 投稿文のトピック類似性を捉えた文ベクトルを学習させることを提案する．本手法によって得られる投稿文のベクトル表現は，単純なコサイン類似度の比較のみで高い精度でハッシュタグの同一性を識別することができ，既存の文ベクトルモデルと比較してトピックの違いを効果的に捉えることを可能にする．

キーワード　SNS, 文ベクトル, BERT, 対照学習

1　はじめに

現代社会において，ソーシャルメディア（以下，SNS）の普及により，ユーザは日常的に大量の情報に触れるようになった．これに伴い，ユーザは情報を整理し，特定のトピックや自身の興味に基づいてコンテンツを検索・閲覧する必要が生じている．ハッシュタグは，このような特定のテーマやトピックを追跡するための有用な手段として利用されている．しかし，全ての投稿にハッシュタグが付与されているわけではなく，ハッシュタグが付与されていない投稿は，トピックを特定することが困難である．そこで，我々は，ハッシュタグの有無に関わらず，投稿文のテキストからトピックの類似性を捉えられる文ベクトルを学習することを目的とする．

本研究では，このような投稿に現れるトピックの類似性を捉えた文ベクトルを非常にシンプルな手法で学習させることを提案する．具体的には，同一のハッシュタグを含む投稿文ペアを学習済みエンコーダに入力して得られたベクトル表現を正例ペアとし，異なるハッシュタグを含む投稿文ペアのベクトル表現を負例ペアとした対照学習を行う．このように学習したベクトル表現は，同じようなトピックの投稿文ペアは近いベクトル表現を持ち，異なるトピックの投稿文ペアは離れたベクトル表現を持つようになることを期待する．

本手法によって得られたベクトル表現がトピックの類似性を捉えられていることを確認するために，既存のデータセットを活用することが考えられる．意味的類似度計算（Semantic textual similarity, STS）[1] [2] は，2 つのテキストが与えられた時に，意味が似ている度合いを人間が 0 から 5 の間で評価したタスクである．本来であればこの意味的類似度にトピックの類似性も組み込まれているべきだが，そもそも意味が似ているとはどういうことなのだろうか．例えば，「python が好きです。」という文章があった時に，「java が嫌いです。」という文章と「ゲームが好きです。」という文章のどちらが意味的に類似しているのだろうか．この質問に対して，明確な答えを出せる人はおそらく少ない．トピック的には，「python が好きです。」と「java が嫌いです。」の方が類似していると考えられるが，対象物に対する意見が異なるため，意味的には類似していないとも考えられる．このように，STS タスクにおける人手評価は，注釈者によって意味に対する解釈が異なる可能性があると考えた．実際に，STS タスクにおいて高性能を達成している既存の文埋め込みモデル [1]で，これらの文章のベクトル表現を比較したところ，「python が好きです。」と「ゲームが好きです。」という異なるトピックであると考えられる文章のコサイン類似度の方が高くなった．以上の点より，STS タスクはトピックの類似性を測るためのデータセットとしては適切ではないと考えた．そこで，我々は，SNS における投稿間のトピックの類似性を直感的かつ定量的に評価するための新たなデータセットを自作した．共通のハッシュタグを持つ投稿を接続することにより，投稿間の関連性をネットワークとしてモデリングし，そのネットワーク内でのホップ数の最短経路を用いて類似性を評価することを提案する．

本手法によって得られるベクトル表現は，単純なコサイン類似度の比較のみで高い精度でハッシュタグの同一性を識別することができた．また，自作のトピック類似度の自動評価タスクにおいても，既存の文埋め込みモデルと比べて，文のトピックの類似性をより効果的に捉えられていることを示した．

2　関連研究

2.1　SNS 投稿のハッシュタグ予測

SNS 投稿のハッシュタグ予測を行った研究として，Zhang らが行った研究がある [3]．彼らは，Twitter で作成された BERT ベースの多言語言語モデルを用いて，投稿文から各言語ごとに人気のある 500 のハッシュタグを予測するタスクを行った．し

1：https://huggingface.co/cl-nagoya/sup-simcse-ja-base

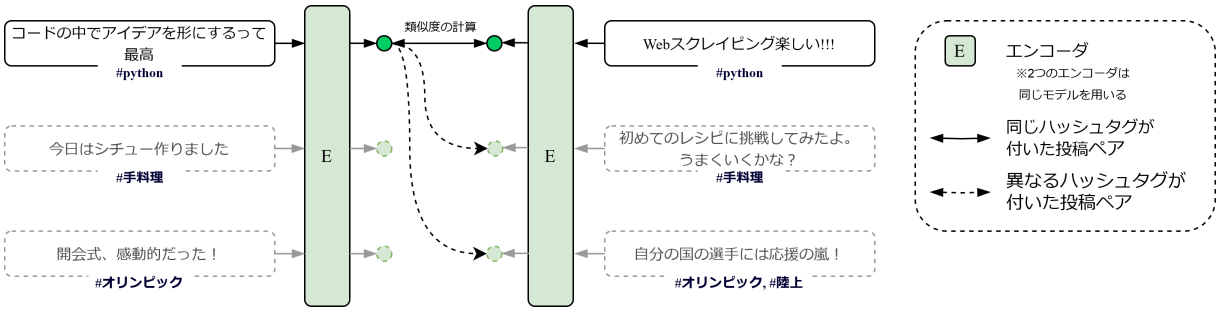


図 1　学習イメージ．ペアの各文を左右に示している．図中の各ツイート例は，筆者による作例．

かし，この研究では，ハッシュタグを 500 個に限定しているため，ハッシュタグが無数に存在しうる SNS におけるトピックを正しく測れているとは言えない．ハッシュタグが同一か否かという情報のみを用いることで，訓練，評価に用いるハッシュタグの種類を限定することなく，SNS 投稿のトピック類似性を捉えることが出来ると考えた．

2.2　教師あり対照学習を用いた文ベクトルの学習

対照学習（Constrastive Learning）[4] は，ラベル付けされていないデータの中から，互いに類似した事例のペア（正例）と異なる事例のペア（負例）を取り出し比較を行うことでモデルの学習を行う手法である．自然言語処理においては，対照学習を用いて文埋め込みの学習に利用することがある．

Gao ら [5] が提案した教師あり SimCSE は，対照学習を用いた文埋め込みの学習に教師ありの設定を適用したモデルである．自然言語推論（Natual Language Inference，NLI）データセットの前提文に対して,「含意」のラベルが付いた仮説文とのペアを正例のペアとし，ミニバッチ内の正例以外の事例を全てバッチ内負例とし，対照学習を行う．この手法によって得られるベクトル表現は，STS タスクにおいて高い性能を達成した．また,「矛盾」のラベルが付いた仮説文をハード負例（hard negative）として定義することでさらなる性能の向上が見られたと報告している．しかし，この手法では，正例とするペアが NLI データセットにおける前提文と仮説文のペアにおける「含意」のペアに限定されている．NLI データセットにおける「含意」のラベルが付いた前提文と仮定文のペアは，同じ画像に対するキャプションであるため，内容が過度に類似しており，違いはわずかである．そのため，この手法によって得られるベクトル表現は，文の意味的な類似性を捉えることに特化しているが，トピックの違いを捉えることには向いていないと考えられる．

3　ハッシュタグ同一性に基づく対照学習

本研究では，ハッシュタグの同一性に基づく教師あり対照学習を用いて，トピックの類似性を捉えた文ベクトルを学習する．図 1 に，提案手法の学習のイメージを示す．

まず，アンカーとなる投稿文 x_i を抽出する．次に，アンカーと同じハッシュタグを含む投稿文を正例のペアとして定義する (x_i^+)．そして，モデルに入力するミニバッチ内の正例以外の投稿文はすべて異なるハッシュタグを含む投稿文で構成し，これらをバッチ内負例として対照学習を行う．また，過度に類似している 2 つの投稿文は，キャンペーン投稿や bot による投稿などの可能性があるため，正例，バッチ内負例ともに含めない．類似度の計算には，編集距離（Levenshtein 距離）[6] を用い [2]，類似度が 0.7 を超える投稿文は除外する．

アンカーとなる投稿文 x_i，正例となる投稿文 x_i^+ から成るデータセット $D = \{(x_i, x_i^+)\}_{i=1}^m$ が与えられた時，h_i，h_i^+ がそれぞれのベクトル表現だとすると，対照学習の損失関数は以下のようになる．

$$\ell_i = -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N}(e^{sim(h_i, h_j^+)/\tau})} \tag{1}$$

ここで，N はミニバッチサイズ，τ は温度付きソフトマックス関数の温度パラメータである．また，ベクトル同士の類似度を計算する関数 $sim(\cdot)$ には，コサイン類似度を用いる．

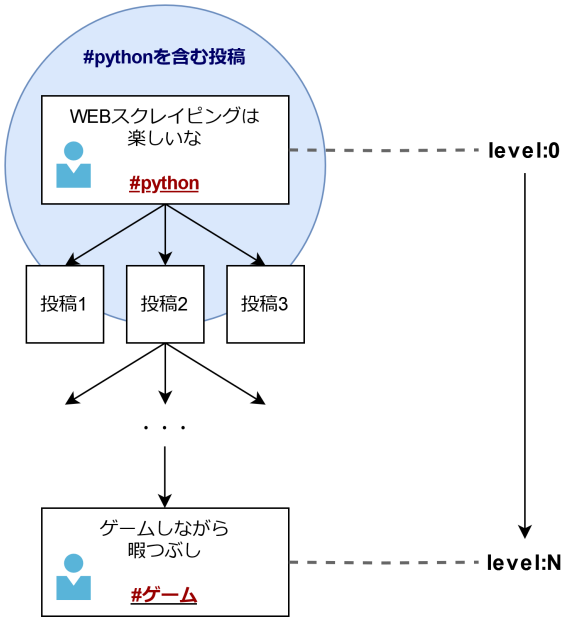


図 2　投稿間のトピック類似性評価タスクのイメージ

2：2 つの文字列間の編集距離を計算し，最大文字列長で正規化する．

4 タスク設計

ハッシュタグ同一性分類の自動評価

評価として，ハッシュタグ同一性分類タスクを設計し，データセットを構築した．アンカーとなる投稿文，投稿文 1，投稿文 2 の 3 つ組が与えられた時，アンカーと同じハッシュタグを含む投稿がどちらかを予測する 2 値分類タスクをハッシュタグ同一性分類タスクとして設計した．予測ラベルは，モデルが出力したベクトル表現のコサイン類似度によって決定さした．

トピック類似度計算の自動評価

本手法によって得られるベクトル表現が，投稿に含まれるハッシュタグの同一性だけでなく，投稿同士のトピックがどれくらい近いかという類似性も捉えられているかを評価するために，トピック類似性の評価タスクを設計し，データセットを構築した．我々は，投稿間の関連性をネットワークとしてモデリングし，トピックの類似性を評価することを提案する（図 2）．

以下に，トピック類似性の評価タスクの詳細を示す．

まず，ハッシュタグ間の類似度を定義する．各ハッシュタグをネットワークのノードと見なし，共通のハッシュタグを持つ投稿によってこれらのノードを接続する．このネットワークを用いて，特定のハッシュタグから別のハッシュタグまでの最短経路を求める．この最短経路の長さは，ハッシュタグ間の関連性の強さを反映すると考えられ，経路が短いほどハッシュタグ間の関連性は強く，逆に経路が長いほど関連性は弱いと考えらる．そこで，ハッシュタグ間の類似度 $s(h_1, h_2)$ を以下のように定義する．

$$s(h_1, h_2) = \frac{1}{d(h_1, h_2)} \tag{2}$$

ここで，$d(h_1, h_2)$ は，ハッシュタグ h_1 と h_2 の間の最短経路の長さである．

次に，投稿間のトピック類似度を定義する．投稿 P_1 に含まれるハッシュタグの集合を H_1，投稿 P_2 に含まれるハッシュタグの集合を H_2 とする．任意のハッシュタグ $h_1 \in H_1$ と $h_2 \in H_2$ に対し，それらのハッシュタグ間の類似度を $s(h_1, h_2)$ とする．この時，投稿 P_1 と投稿 P_2 の間のトピック類似度を以下のように定義する．

$$\text{トピック類似度}\,(P_1, P_2) = \frac{\sum_{h_1 \in H_1} \sum_{h_2 \in H_2} d(h_1, h_2)}{|H_1| \times |H_2|} \tag{3}$$

このトピック類似度は，投稿 1 と投稿 2 の各ハッシュタグの類似度の平均値と等価である．

アンカーとなる投稿文とランダムに抽出した投稿文を与えられた時に，ラベルとして投稿間のトピック類似度を付与し，モデルが出力する 2 文の類似度スコアとラベルの間の相関係数を評価指標として用いた．

5 実　　験

5.1 SNS 投稿の収集と前処理

本研究で使用するデータは，Twitter API V2 の Academic Research アクセス [3]を用いて取得し、ランダムに抽出した 15,000 ユーザの 2006 年 3 月 21 日〜2021 年 8 月 31 日の期間に投稿された投稿文を用いた．すべての投稿文に共通する前処理として，リツイート，画像・動画を含むツイートは除外した．また，他のアカウントに対するリプライ及びメンションを含む投稿を除外する．さらに文字フィルタとして ”RT ”という単語を含む投稿を除外し，URL は削除した．投稿文中に含まれるハッシュタグは，データセット構築時のみ利用し，モデルの学習には用いない．これらの前処理を行った投稿のうち，文字数が 15 文字以上，60 文字以下の投稿を抽出した．

5.2 学習用データセット

前章で述べた 15,000 ユーザの投稿文を、アカウント単位で訓練: 検証: テストが 7:2:1 になるよう分割した．結果，10,500 ユーザが訓練，1,500 ユーザが検証，3,000 ユーザが評価データセットとなった．なお，投稿数が非常に多いユーザに偏らないように各ユーザの最大投稿数は 20,000 件とし，これを超える投稿は除外した．各セットの投稿数を表 1 に示す．

表 1　各データセットの投稿数

データセット	投稿数
訓練データセット	14,348,691
検証データセット	2,050,260
評価データセット	4,140,965

訓練データ

訓練には，アンカーとなる投稿文，正例となる投稿文の 2 つ組を 1,000,000 ペア抽出して用いた．

検証データ

検証では，文ペアを入力した時に訓練モデルが出力する類似度スコアを用いた．アンカーとなる投稿文とランダムに抽出した投稿文の 2 つ組を 20,000 ペア抽出し，検証データとして使用した．この時，ラベルとして，アンカーと同じハッシュタグを含む投稿文であれば 1，異なるハッシュタグを含む投稿文であれば 0 を付与した．そして，モデルが出力する 2 文の類似度スコアとラベルの間の Peason 相関係数を評価指標として用いた．

評価データ

評価では，前章で述べたハッシュタグ同一性分類タスク，トピック類似性の評価タスクを行った．ハッシュタグ同一性分類タスクでは，アンカーとなる投稿文，投稿文 1，投稿文 2 の 3 つ組を 40,000 ペア抽出し，使用した．トピック類似性の評価タスクでは，アンカーとなる投稿文とランダムに抽出した [4]投

3：https://developer.twitter.com/en/docs/twitter-api

4：過度にばらつきがでないように，類似度 0.0 から 1.0 の範囲で 4 つのクラスに分割し，それぞれのクラスの投稿文を均等に抽出する．

表 2　タスクの精度比較

	ハッシュタグ同一性	トピック類似性	
Model	Accuracy	Pearsonr	Spearmanr
Publicly Released Pretrained Model			
$\text{BERT}_{\text{base}}$	0.688	0.156	0.161
$\text{TwhinBERT}_{\text{base}}$	0.691	0.388	0.429
Publicly Released Embedding Model			
Unsupervised $\text{SimCSE-BERT}_{\text{base}}$	0.746	0.260	0.232
Supervised $\text{SimCSE-BERT}_{\text{base}}$	0.757	0.283	0.253
Open AI Embedding Model			
text-embedding-ada-002	0.777	0.292	0.266
Proposed Model	**0.877**	**0.602**	**0.597**

稿文の 2 つ組を 40,000 ペア抽出し，使用した．

また，クエリ検索を用いたトピック類似性の評価を行った．検索文に対し，評価データセットのために割り当てられた 3,000 ユーザの 4,140,965 件の投稿文の中から，コサイン類似度が最も高い投稿文を出力させ，トピックの類似性を捉えられていることを確認する．

5.3　モデルの学習

提案モデル

本研究では，東北大学が公開している日本語 $\text{BERT}_{\text{base}}$ モデル [5]をベースモデルとして対照学習に用いた [7]．訓練時の各種パラメータ設定を，表 3 に示す．出力されるベクトル表現の次元数は，ベースモデルと同じ 768 次元である．

表 3　訓練時の各種パラメータ

項目	値
バッチサイズ	128
エポック数	3
検証ステップ	250
学習率	$5e-5$
温度パラメータ	0.05

比較モデル

提案モデルの比較対象として，下記の五つのモデルを用意した．一つ目として、提案モデルのベースモデルである日本語 $\text{BERT}_{\text{base}}$ モデルの [CLS] トークンに該当するベクトル表現を用いた．二つ目として，Twitter 社が公開している多言語 SNS 事前学習モデルである $\text{TwhinBERT}_{\text{base}}$ モデル [6]を用いた．二つ目と三つ目として，名古屋大学が公開している教師あり SimCSE モデル [7]と教師なし SimCSE モデル [8]を用いた．これらは，提案モデルのベースモデルと同じ日本語 $\text{BERT}_{\text{base}}$ モデルに対して対照学習を行ったモデルであり，出力されるベクトル表現の次元数は，768 次元である．

さらに，OpenAI が公開している埋め込み特化モデル [9]も四つ目として比較対象として用いた [8]．このモデルが出力するベクトル表現の次元数は，1,536 次元である．

6　実 験 結 果

ハッシュタグ同一性分類の評価

表 2 に，モデルごとのハッシュタグ同一性分類タスクの結果を示す．既存の事前学習済みモデルや SimCSE モデルと比較して，提案モデルは，得られるベクトル表現の単純な類似度比較のみで高い精度でハッシュタグを識別することが可能であった．さらに，出力するベクトルの次元数が大きい OpenAI の埋め込み特化モデルと比較しても，提案モデルの方が高い精度でハッシュタグを識別できた．

トピック類似度計算の自動評価

表 2 に，モデルごとのトピック類似性の評価タスクの結果を示す．既存の事前学習済みモデルや SimCSE モデルと比較して，提案モデルによって得られるベクトル表現の類似度は，投稿に含まれるハッシュタグの同一性だけでなく，投稿同士のトピックがどれくらい近いかという類似性も捉えられていることを示した．

クエリ検索によるトピック類似性の評価

表 4 に，提案モデルを用いたクエリ検索の結果を示す．検索文に対し，評価データセットのために割り当てられた 3,000 ユーザの 4,140,965 件の投稿文の中から，コサイン類似度が最も高い投稿文を 5 つ出力させた．比較モデルには，前章で述べた教師なし SimCSE モデル [10]を用いた．

既存の埋め込みモデルは,「言語」という単語に反応して,「語彙力」や「日本語」などの投稿文を上位に出力しており,「C 言語」がプログラミング言語であることを捉えられていない．一方，提案モデルは,「C 言語」がプログラミング言語であることを捉えられており,「PHP」や「javascript」などのプログラミング言語に関する投稿文を上位に出力している．

5：https://huggingface.co/cl-tohoku/bert-base-japanese-v3
6：https://huggingface.co/twitter/twihinbert-base
7：https://huggingface.co/cl-nagoya/sup-simcse-ja-base
8：https://huggingface.co/cl-nagoya/unsup-simcse-ja-base
9：text-embedding-ada-002
10：https://huggingface.co/cl-nagoya/unsup-simcse-ja-base

表 4　クエリ検索結果 (Top5)．表中の各ツイートは，匿名化加工済み．

検索文：C 言語はすごい言語だな	
比較モデル：Unsupervised SimCSE-BERT$_{\text{base}}$	
#1	すごい...（こんなに語彙力が限られているのに、一体何を話しているんだろう...。
#2	すごい...（それしか言えないほどの語彙力）
#3	まさにこれ……言語化って難しいよね
提案モデル	
#1	PHP は読みやすい言語だな。
#2	JavaScript は面白そうだな。
#3	自然言語を作ったら、それは人工言語になるよね。

7 考　　察

表 2 に示したように，提案手法は，既存の文埋め込みモデルと比較し，高い精度でハッシュタグの同一性を識別できた．また，自作のトピック類似度の自動評価タスクにおいても，提案手法によって得られるベクトル表現の類似度は，トピックの類似度を正しく反映していることを示した．

既存の文埋め込みモデルは，文の意味的類似度を捉えるように学習が行われるため，本来であればより広い意味での類似度であるトピックの類似度も反映できると考えられる．しかし，SNS のようにトピックが無数に存在し，増え続ける可能性があるデータに対して，既存の文埋め込みモデルは，トピックの類似度を反映することができないのではないかと考えた．ハッシュタグが同一か否かという情報のみを用いる本手法は，トピックが無数に存在し，増え続ける可能性があるデータに対しても，半自動的に学習データを増やすことが可能であり，高い精度でトピックの類似度を反映することができる．

表 4 に示したように，提案手法は，クエリ検索による評価においても，トピックの類似性を捉えていることを確認できた．1 つの投稿には，複数のハッシュタグが含まれる可能性があり，その同一性を学習することで，検索文に含まれる「C 言語」という単語に対して,「PHP」や「javascript」といった，単語は異なるが共通のトピックである投稿文を上位に出力することができたと考えられる．一方，既存の文埋め込みモデルは，同じ画像に対するキャプションの同一性を学習するため,「すごい」や「言語」といった同じ単語が含まれる投稿文を上位に出力してしまったのだと考えられる．今回は，プログランミングに対する投稿文を例に検索を行ったが，提案手法によって得られるベクトル表現を利用することで，他の様々なトピックに対しても同様の結果が得られると考えられる．

8 おわりに

本研究では，SNS 投稿文のトピックの類似性を捉えた文ベクトルを，ハッシュタグの同一性に基づく対照学習を用いて，学習する手法を提案した．提案手法によって得られるベクトル表現は，文章に含まれるトピックの類似性を捉え，高い精度でハッシュタグの同一性を識別できた．この手法は，ハッシュタグが無数に存在しうる SNS 投稿文のトピックを捉えるための有効な手法であると考えられる．

今後は，提案手法によって得られたベクトル表現を用いて，トピックモデリングやユーザの意図を反映したクエリ検索などのタスクに応用していきたい．

謝　　辞

本研究は，JSPS 科研費 JP22H00804，およびセコム科学技術財団特定領域研究の助成を受けたものです．

文　　献

[1] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[2] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. pp. 2023–2028, 2022.

[3] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, p. 5597–5607, New York, NY, USA, 2023. Association for Computing Machinery.

[4] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, Vol. 9, No. 1.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, Vol. 10, pp. 707–710, 1965.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein,

Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas A. Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David P. Schnurr, Felipe Petroski Such, Kenny Sai-Kin Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *ArXiv*, Vol. abs/2201.10005, , 2022.

SNS実データ分析に基づくユーザのフォロー行動モデルの評価

山川　衛† 田島　敬史†

† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町 36-1
E-mail: †yamakawa.mamoru@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　ソーシャルネットワークサービス (SNS) 上のフォローネットワークの構造分析は様々な側面から行われてきたが，その中でも特にユーザの三つ組についての分析はユーザのフォロー行動を分析することにおいてとても重要である．そこで本論文では日本の X ユーザのフォローネットワークにおけるユーザの三つ組について分析を行い，SNS におけるユーザのフォロー行動の詳細を明らかにする．また分析結果から以前提案したフォロー行動モデルをさらに妥当性のあるものに拡張し，モデルにより生成したネットワークと実データとの定量的比較検証を行う．
キーワード　ソーシャルメディア，ユーザモデル，ソーシャルネットワーク生成モデル

1 はじめに

近年は X（旧 Twitter）をはじめとした SNS 上のフォローネットワークの分析が盛んにおこなわれているが，ユーザによる情報の匿名化や実データのクロールを行うサービス等の変化により実データを入手することが困難になってきている．そのため，SNS 上のそれに近しいフォローネットワークを人工的に生成する技術は情報プライバシーの観点からも非常に有用であると考える．以前筆者らは文献 [1] で X の日本人ユーザ 2750 万人からなるフォローネットワークについての分析を行い，SNS で起こるフォロー行動に基づいた有向ネットワークを生成するモデルを提案したが，分析結果を踏まえたモデルの妥当性の議論や生成したネットワークの数値的な評価が十分に行えていなかった．そこで本論文ではフォローネットワークをサンプリングした小規模のネットワークを用いて計算量の大きいネットワーク指標について分析を行った．文献 [1] で提案したモデルの特徴である 2-hop 先へのフォローについての分析を行うために特定のフォロー関係を満たすユーザの三つ組について分析を行ったが，その他の三つ組に対して 2-hop 先のフォローが起こりやすいというモデルの妥当性を高める結果が得られた．また分析結果を受けて改良版のモデルを提案し，パラメータを変化させ定量的な評価を行った結果，改善モデルは従来のモデルから次数分布を損なわずクラスター係数を高めることが出来るという結果という結果が得られた．今後はパラメータの変化に対する生成したネットワークの性質の変化についての数学的説明や実際に生成ネットワーク上で感染モデルなどを動かした際のふるまいを観察する実験などを行っていきたい．

2 実データの解析

文献 [1] で述べたように，今回は，2015 年に収集した lang=ja の約 4 千万の X ユーザのうち，その当時のフォロワー数が多い方から 2750 万ユーザについて，2021/12（から 2022/1 にかけて）時点でのフォロワー，フレンドのリストを X 社から購入したデータからサンプリングした 10000 ノードからなるフォローネットワークについて解析を行った．なお今回はサンプル手法としてサンプル前のネットワークで計算が困難であったクラスター係数を保存しやすい [3] ランダムウォークによるサンプリング [4] を採用した．フォローネットワークは有向グラフ $G = (V, E)$ で定義され，各 $i \in V$ はユーザを，$(i, j) \in E$ は i が j をフォローしているという関係を表す．

2.1 関連研究

今回は X のフォロー行動についてネットワーク構造を用いて分析した研究についていくつか述べる．
X ではユーザが了承を得ることなく (鍵アカウントをのぞき) 一方的に他方をフォローすることができ，フォローされたユーザがフォローバックするかどうかは自由となっている．[6] ではこのフォローされたユーザのフォローバックに起因する特徴について研究しており，主に以下の六つを挙げている．

（1） 地理的距離は相互フォロー自体の数に影響を与えるが，相互フォローになる割合には少ししか影響しない
（2） ユーザは共通フレンド，共通相互フォロワーを持つユーザに対してフォローバックしやすい傾向にある
（3） エリートユーザ同士は一般ユーザ同士よりも相互フォローになりやすい傾向にある
（4） リツイートやリプライによって形成される潜在的なネットワークが相互関係の形成と強い相関関係にある
（5） 相互フォローで形成されたネットワークの多くは Balance Theory を満たしており，一方で一方向フォローで形成されたネットワークはほとんど Balance Theory を満たしていない
（6） エリートユーザが中間点となった場合に三角形が形成される確率はランダムの 10 倍高い

Balance Theory [7] とはノードの三つ組についての理論であり，各エッジに端点同士の好感度の符号をあてはめ，負の符号をもつ枝の数が偶数であれば均衡 (balanced) とするものである．[6] では好感度として相互フォロー (一方向フォロー) であるかどうかを採用しており，一方向フォローで形成されたネットワークが balance theory を満たさない理由の一つとして同じ映画スターをフォローしている 2 ユーザはお互いを知らないことが多

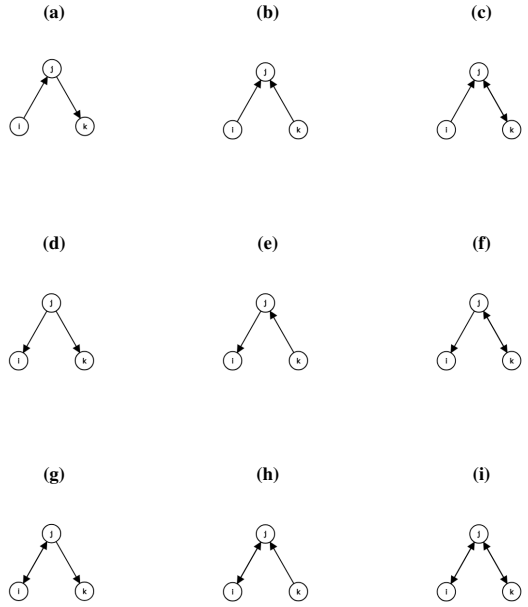


図 1: (i, j, k) の三つ組

パターン	$\|\{i,j,k \in T(x)\}\|$	$\|\{(i,k) \in E \cap i,j,k \in T(x)\}\|$	prob(x)
(a)	6942265	1004918	0.144754
(b)	60172354	2221049	0.036911
(c)	12376578	1676327	0.135443
(d)	27325772	1962089	0.071804
(e)	5420334	81734	0.015079
(f)	14300291	719156	0.050290
(g)	16661957	1935287	0.116150
(h)	12214003	719156	0.058880
(i)	49632496	9186729	0.185095

表 1: i, j, k の組数と (i,k) の存在確率 prob

いことを挙げている．このように，X のフォローネットワークにおけるユーザの三つ組はフォロー形成に非常に重要な役割を果たしている．

2.2 フォローネットワークにおける三つ組の解析

フォローネットワークにではノードはフォロー関係によって結ばれており，その中でも三つ組はフォロー行動の解析に非常に有用であると考える．そこで，今回はまず i と j，j と k がどのようなフォロー関係の場合に (i,k) が存在するかについて調査を行った．i と j，j と k のフォロー関係によって構成される i, j, k の三つ組のパターンは図 1 の 9 種類となっている．以後，i, j, k の三つ組がパターン x に属していることを $i, j, k \in T(x)$ で表す．

各パターンについて，ネットワークに存在する (i, j, k) の組み合わせの総数とそのうち有向辺 (i,k) が存在する組の総数，そして (i,k) が存在する確率 prob(x) を調べた．結果は表 1 のようになっており，三つ組の数が多いパターンは (b)，(d)，(i)，(i,k) の存在確率が高いものが (a)，(c)，(g)，(i) となっている．(b) と

パターン	corr	$\text{corr}_{\log}$
(a)	0.716293	0.877356
(b)	0.861798	0.704398
(c)	0.803216	0.930994
(d)	0.758131	0.606555
(e)	0.597674	0.282978
(f)	0.617362	0.432271
(g)	0.750171	0.804053
(h)	0.787018	0.596138
(i)	0.609262	0.871556

表 2: j の数と (i,k) の存在確率

(d) は三つ組の数に対して (i,k) の存在確率が非常に低い結果となった．(b) は i，k が j を一方的にフォローしている形であり，i と k は有名人である j をフォローしているユーザ同士である可能性が高く，また (d) は j が i，k を一方的にフォローしている形であり，i，k が j の属している別のトピックにそれ属しているも可能性が高い．加えて，両パターンで i，k はリツイートなどで j を通じてお互いを観測するのが難しい構造をしている．そのためこの 2 パターンではユーザの組数は多いがお互いを発見しフォローする確率が小さくなるのではないかと考えられる．また (a)，(c)，(g) は三つ組の数はそこまで多くはないが (i,k) の存在確率が高い結果となったが，この 3 パターンにおいて，(i,k) は全てフレンドのフレンド，つまり 2-hop 先のノードをフォローしている形となっている．2-hop 先のノードは自身と同コミュニティに属している可能性が高く，またリツイートによる k から i 情報伝播によって i が k を観測することが可能になっているためフォロー行動を行いやすく，(i,k) の存在確率が高くなる結果になったと考えられる．そして (i) は三つ組の数自体も非常に多く，(i,k) の存在確率も非常に高くなる結果となった．これは i, j, k が全て相互フォローで結ばれており i と k が同コミュニティに属している可能性が非常に高いためであると考えられる．この結果から，2-hop 先のノードへのフォローはその他の起こりやすいものであると考える．

次に，j の数と (i,k) の存在確率について関係があるかどうかを調べるために，以下の手順で分析を行った．

（1） 各 i,k の組について，$i, j, k \in T(x)$ となる j の数 $N_j(i,k,T(x))$ を求める

（2） 各 n について (i,k) の存在確率
$\frac{|\{i,k|(i,k) \in E \cap N_j(i,k,T(x)) = n\}|}{|\{i,k|N_j(i,k,T(x)) = n\}|}$ を求める

（3） (i,k) の存在確率と $n, \log(n)$ の相関係数 $\text{corr}, \text{corr}_{\log}$ をそれぞれ求める

結果は図 2 となっており，図 1 で (i,k) の存在確率が高かった (a)，(c)，(g)，(i) では corr，$\text{corr}_{\log}$ ともに高い値を取っており，特に $\text{corr}_{\log}$ が非常に高い値となっていることわかる．また，残りの 5 パターンでは前述の 4 パターンに比べて corr，$\text{corr}_{\log}$ の値が小さく，大小結果が逆となった．

2.2.1 フォローバック確率に関する分析

ユーザのフォローバック行動についての分析を行うために，共通フレンド数とフォローバック数に着目した分析を行った．

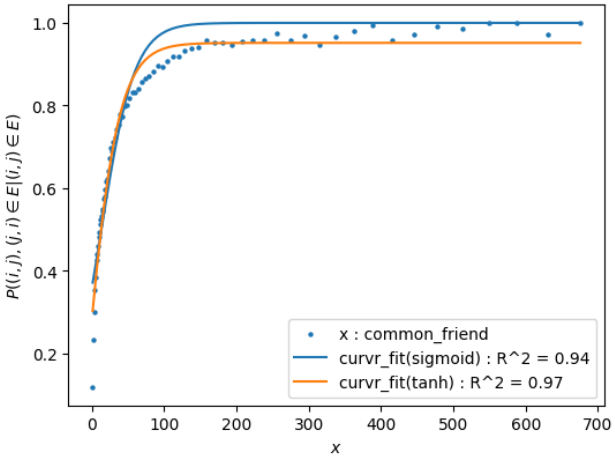


図 2: フォローバック確率のプロットと回帰曲線．今回は共通フレンド数が 2^x の際のフォローバック確率を共通フレンド数が $2^{x+0.05}$ 以上 $2^{(x+1)+0.05}$ 未満である i, j のうち相互フォローである組の割合としている (x は自然数).

今回はフォローバック確率を $(i, j) \in E$ である i, j の組のうち，$(i, j), (j, i) \in E$ である組の割合で定義する．

$$P((i, j), (j, i) \in E | (i, j) \in E) = \frac{|\{i, j \in V | (i, j), (j, i) \in E\}|}{|\{i, j \in V | (i, j) \in E\}|}$$

図 2 は $(i, j) \in E$ である i, j の共通フレンド数とフォローバック確率をプロットしたものであり，フォローバック確率は共通フレンド数が 150 程度までは単調に増加していき，200 以上になると一定の値に落ち着くふるまいを見せた．そこで tanh と sigmoid 関数を用いて共通フレンド数とフォローバック確率による曲線のフィッティングを行ったところ，それぞれ決定係数 0.97，0.95 で以下の関数が得られた．

$$\begin{aligned} P((i, j), (j, i) \in E | x, (i, j) \in E) &= 0.95 \times \tanh(0.02 \times (x + 14.24)) \\ &= 0.95 \times \frac{e^{0.02\times(x+14.24)} - e^{-0.02\times(x+14.24)}}{e^{0.02\times(x+14.24)} + e^{-0.02\times(x+14.24)}} \\ P((i, j), (j, i) \in E | x, (i, j) \in E) &= \frac{0.775}{0.775 + \exp(0.160 \times (x - 2.492))} \end{aligned} \quad (1)$$

2.3 その他の指標の解析

2.3.1 その他の指標について

三つ組の解析の他にも，以下の指標について分析を行った．

In-degree

ユーザ i の In-degree $I(i)$ は i の入次数であり，フォロワー数に該当する．

Out-degree

ユーザ i の Out-degree $O(i)$ は i の出次数であり，フレンド数に該当する．

Reciprocal-degree

ユーザ i の Reciprocal-degree $R(i)$ は $(i, j), (j, i) \in E$ となるような $j \in V$ の数で定義する．これは X における相互フォロワー数に該当する．

Degree

ユーザ i の次数 $deg(i)$ は $I(i) + O(i)$ で定義する．

密度

有向グラフの密度 D は $\frac{|E|}{|V|(|V|-1)}$ で定義される．

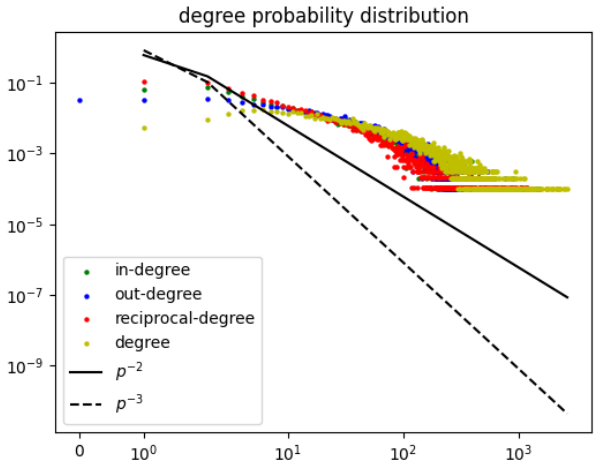


図 3: サンプルグラフの次数分布

平均経路長

平均経路長は任意の 2 ノード間の経路を考慮した最短経路長の平均で定義される．

クラスター係数

本論文ではノード i のクラスター係数 [5] を以下で定義する．

$$C(i) = \begin{cases} \frac{2t(i)}{k(i)(k(i)-1)} & (k(i) \geqq 2) \\ 0 & (k(i) = 0, 1) \end{cases}$$

ここで，$k(i), t(i)$ はそれぞれフォローネットワークを単純無向グラフ化した際の i の次数，i を頂点とする三角形の数である．また分析の際には全ノードのクラスター係数の平均をとったものを用いる．

In-Out 相関

In-Out 相関は In-degree と Out-degree の相関係数で定義する．

2.3.2 分析結果

In-degree，Out-degree，Reciprocal-degree，Degree の四つの次数について，次数分布とその累積分布関数は図 3，4 のように，その他の指標については表 3 のようになった (図 3 の ρ^{-2}, ρ^{-3} はそれぞれスケールフリーネットワークの特徴を表すべき乗分布である).

サンプリングを行った実データの次数分布は，次数の小さい部分では従来のスケールフリーネットワークで見られるべき乗分布とは異なる丸みを帯びた分布とであり，大きい部分ではべき乗分布で見られるロングテールとなっている．これは [1] の実データ全体の次数分布における低次数帯，高次数帯の分布の特徴と一致している．また，次数の小さいところではべき乗分布よりも頂点数が少ないことが多い [2] というのはスケールフリーネットワークの特徴であり，このサンプリンググラフもスケールフリーネットワークであると考えられる．

3 モデルの改善

ここでは筆者ら文献 [1] で提案したネットワーク生成モデルについての説明と，分析結果によるその改善を行う．

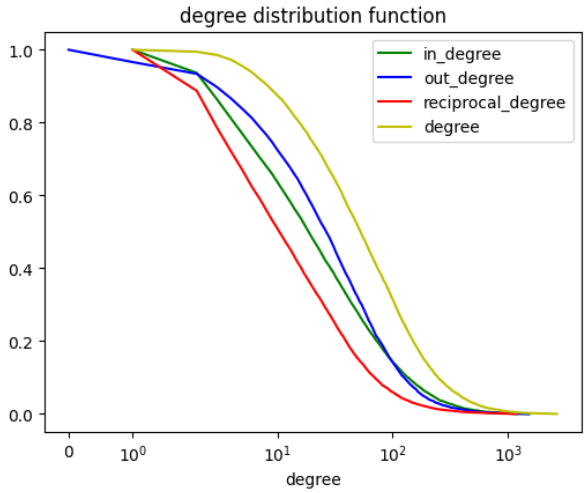


図 4: 累積分布関数

$\|V\|$	10000
$\|E\|$	535725
密度	0.005358
平均経路長	3.0100
クラスター係数	0.1836
In-Out 相関	0.6056

表 3: その他の指標

3.1 文献 [1] における提案モデルについて

文献 [1] で提案したモデルは入力として頂点数 N とパラメータ μ_1，μ_2，α が与えられ，ネットワークの頂点数が N になるまで確率的に以下を繰り返すモデルである．

（1） 確率 μ_1 でフォロー行動を行う既存ノード i をランダムに選択する．その後，i は入次数に応じた確率でノード j をランダムに選択し接続を行う．その後，j は確率 α で i と接続を行う．

（2） 確率 μ_2 でフォロー行動を行う既存ノード i をランダムに選択する．その後，i は $(A^2)_{ij}$ に応じた確率でノード j をランダムに選択し接続を行う．その後，j は確率 α で i と接続を行う．

（3） 確率 $1-(\mu_1+\mu_2)$ で新規ノード i をネットワークに加える．この時，新規ノードは入次数に応じた確率で既存ノード j を選択し接続する．その後，j は確率 α で i と接続を行う．

ここで，A は各行動時点でのネットワーク全体の隣接行列で，A_{ij} は枝 (i,j) が存在する場合に 1，存在しない場合には 0 となる行列であり，μ_1，μ_2，α は以下の条件を満たす実数である．

- $0 \leqq \mu_1 + \mu_2 < 1$
- $\mu_1 \geqq 0$
- $\mu_2 \geqq 0$
- $0 \leqq \alpha \leqq 1$

また，(1)，(3) でノード i が接続先としてノード j を選択する確率，(2) でノード i が接続先としてノード j を選択する確率はそれぞれ以下の値となる．

- $\dfrac{I(j)+1}{\sum_{j\neq i}(I(j)+1)}$

Algorithm 1 generate_graph(N, μ_1, μ_2)

Require: $N \in \mathbf{N}, \mu_1, \mu_2 \geqq 0, 0 \leqq \mu_1+\mu_2 < 1, 0 \leqq \alpha_{ji} \leqq 1$

$V = \{0\}$
$E = \emptyset$
$G = (V, E)$
$n = 1$
while $n < N$ **do**
  p = rand()
  A = adjacency matrix of G
  if $p \leqq \mu_1$ **then**
    randomly select i from V
    select j with probability $\frac{I(j)+1}{\sum_{j\neq i}(I(j)+1)}$
  else if $p \leqq \mu_1+\mu_2$ **then**
    randomly select i from V
    select j with probability $\frac{\log((A^2)_{ij}+1)}{\sum_{j\neq i}\log((A^2)_{ij}+1)}$
  else
    $i = n$
    randomly select j from V
    $V = V \cup i$
    $n = n+1$
  end if
  $\alpha_{ji} = \frac{I(i)+1}{I(i)+I(j)+2}$
  add_edge(G, i, j, α_{ji})
end while

Algorithm 2 add_edge(G, i, j, α_{ji})

Require: Directed graph $G = (V, E), i, j \in V, 0 \leqq \alpha_{ji} \leqq 1$

if $(i,j) \notin E$ **then**
  $E = E \cup (i,j)$
  p = rand()
  if $p \leqq \alpha_{ji}$ and $(i,j) \notin E$ **then**
    $E = E \cup (j,i)$
  end if
end if

- $\dfrac{(A^2)_{ij}}{\sum_{j\neq i}(A^2)_{ij}}$

3.2 モデルの改善

2 節での結果から，今回実験に使用するモデルを [1] で提案したモデルから以下のように数点改良する．なお，改善後のモデルのアルゴリズムは Algorithm 1，2 となっている．以後，このモデルを改善モデルと記載する．

（1） (2) で行う優先的選択について，$I(i)$ に比例した確率ではなく $\log I(i)$ に比例した確率で優先的選択を行うようにする．

（2） フォローバック確率に式 1 を用いる．

4 実　　験

4.1 概　　要

今回はまず改善モデルにおける行動の割合を決定するパラメータの値の変化によるモデルのふるまいの変化を観察するために，パラメータを以下の値の中から組み合わせて設定した合計 63 個のネットワークを生成した．

- $|V| = 10000$
- $1-(\mu_1+\mu_2) \in [0.05, 0.10, 0.15]$
- $\mu_1 : \mu_2 \in [1:0, 0.9:0.1, 0.75:0.25, 0.5:0.5, 0.25:0.75, 0.1:0.9, 0:1]$
- $\alpha_{ij} = \alpha_{ij}^1, \alpha_{ij}^2, \alpha_{ij}^3$

なお，$\alpha_{ij}^1, \alpha_{ij}^2, \alpha_{ij}^3$ は以下で定義する．

- $\alpha_{ij}^1 = 0.95 \times \tanh(0.02 \times (CF(i,j) + 14.24))$
- $\alpha_{ij}^2 = 0.5$
- $\alpha_{ij}^3 = \frac{I(j)+1}{I(i)+I(j)+2}$

ここで，$CF(i,j)$ は i と j の共通フレンド数である．それぞれのフォローバック確率において，α_{ij}^1 は 2 において共通フレンドとフォローバック確率の関係についてフィッティングを行った結果，決定係数が最も高かった tanh 関数によるフィッティング結果である式 1 を採用したものであり，実データにおけるフォローバック行動を再現した場合のネットワークとしている．α_{ij}^2 は全てのノードがランダムにフォローバックするネットワークに対応し，α_{ij}^3 はフォローされたユーザがフォローしたユーザとの次数比でフォローバック確率を決定するネットワークに対応している．これらのネットワークについて各指標を計算し，まずは行動 1 と行動 2 の割合変化によるネットワークのふるまいの変化について分析を行い，次にノード参入確率によるネットワークのふるまいの変化について分析する．その後フォロバック関数の違いによるネットワークのふるまいの変化について分析を行い，最後に実データとの比較を行い，提案モデルの最適なパラメータ，改善点について議論を行う．今回比較する指標は 2 で実データの解析を行った際に用いた指標に加えて，次数分布の忠実度を定量的に評価する指標として以下で定義されるカルバック・ライブラー情報量 (KL 情報量) を用いる．

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

本実験では，$P(i)$，$Q(i)$ はそれぞれサンプリングした実データ，生成したネットワークの次数分布において次数が i である確率とする．KL 情報量が小さいほど P と Q の次数分布が似ていることになり，KL 情報量が 0 であるときに 2 つの分布は完全に一致している．なお，今回は $P(i) = 0$ となる i が存在した場合に $D_{\mathrm{KL}}(P\|Q) = -\infty$ となってしまう問題を解決するために，以下のように各 $P(i)$ に微小量 $\varepsilon = 10^{-5}$ を加えた値を用いて評価する．

$$D_{\mathrm{KL}}(P\|Q) = \sum_i (P(i)+\varepsilon) \log \frac{P(i)+\varepsilon}{Q(i)+\varepsilon}$$

5 実験結果・考察

5.1 実験結果

三つ組の指標について，各パターンにおけるフォロー確率は図 8〜10 のように，各パターンの総数については図 5〜7 のようになった． また，その他の指標について，結果は図 11，12，13 のようになった．

フォローバックを $\alpha_{ij}^1, \alpha_{ij}^2, \alpha_{ij}^3$ と変化させて生成したネット

図 5: 行動割合による各パターンの数の変化 $(1-(\mu_1+\mu_2) = 0.05)$

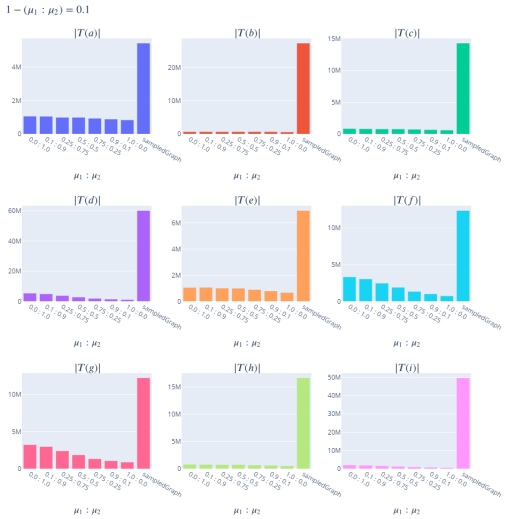


図 6: 行動割合による各パターンの数 $1-(\mu_1+\mu_2) = 0.10)$

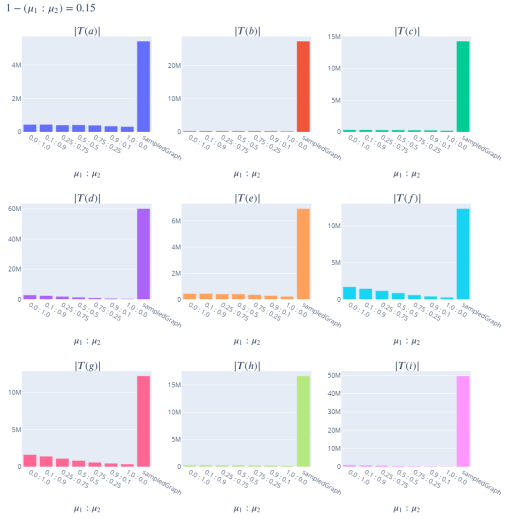


図 7: 行動割合による各パターンの数の変化 $(1-(\mu_1+\mu_2) = 0.15)$

図 8: 行動割合による各パターンのフォロー確率の変化 ($1-(\mu_1+\mu_2)=0.05$)

図 9: 行動割合による各パターンのフォロー確率の変化 ($1-(\mu_1+\mu_2)=0.10$)

ワークについては表 14〜16 のようになった．なお今回は本文中には代表して $1-(\mu_1+\mu_2)=0.05, \mu_1:\mu_2=0.5:0.5$ のみを掲載している．

5.2 行動 1 と行動 2 の割合変化によるふるまいの変化

ここでは行動 1 と行動 2 の割合に対するふるまいについての分析を行う．まずは三つ組の数について，図 5〜7 から行動 2 の割合を増加させるにつれて総数が増加しているものはパターン (d)，(f)，(g)，(i) であり，その他のパターンについてはわずかに増加傾向であるか横ばいという結果であった．改善モデルの行動 2 はパターン (a)，(e)，(g)，(i) の状態にある i から k へのフォローを積極的に行うモデルであるため，それらの 4 パターンに (i,k) を追加した三角形がネットワークに増加すると考えられる．そしてこの四つの三角形がネットワークに新たにできることで，それぞれ以下のパターンを増加させる．

- パターン (a)：パターン (b)×2，(d)×2

図 10: 行動割合による各パターンのフォロー確率の変化 ($1-(\mu_1+\mu_2)=0.15$)

図 11: 行動割合によるその他の指標の変化 ($1-(\mu_1+\mu_2)=0.05$)

図 12: 行動割合によるその他の指標の変化 ($1-(\mu_1+\mu_2)=0.10$)

図 13: 行動割合によるその他の指標の変化 $(1-(\mu_1+\mu_2)=0.15)$

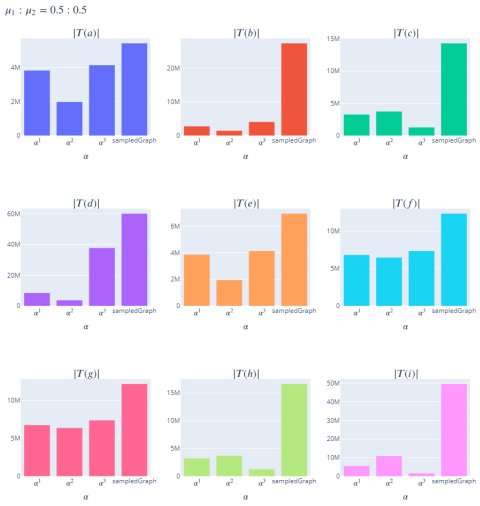


図 14: フォローバック確率による各パターンの数の変化 $(\mu_1:\mu_2=0.5:0.5)$

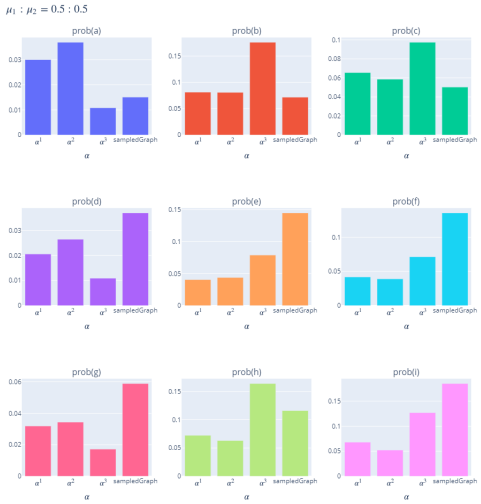


図 15: フォローバック確率による各パターンのフォロー確率の変化 $(\mu_1:\mu_2=0.5:0.5)$

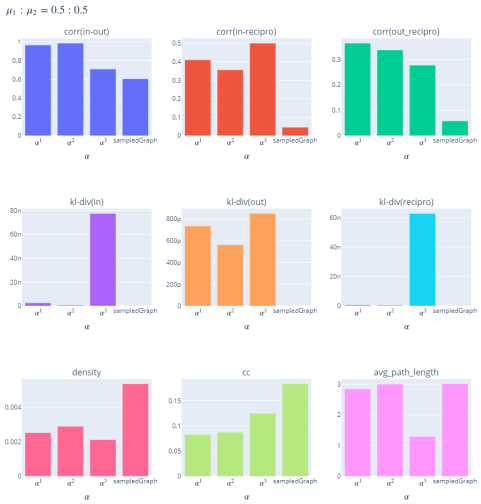


図 16: フォローバック確率によるその他の指標の変化 $(\mu_1:\mu_2=0.5:0.5)$

- パターン (c)：パターン (c)，(d)×2，(f)
- パターン (g)：パターン (b)×2，(f)，(g)
- パターン (i)：パターン (c)，(d)×2，(f)

また，フォローバックが起こった場合には以下の各パターンが増加する

- パターン (a)：パターン (c)，(f)，(g)，(h)
- パターン (c)：パターン (c)，(f)，(i)×2
- パターン (g)：パターン (c)，(g)，(i)×2
- パターン (i)：パターン (i)×4

また，行動 1 は有名人をフォローする行為であり，フォロワー数 (In-degree) が大きいノードを優先的に選択する行動であるため，パターン (b) やそれを含む (c)，(h)，(i) が増加しやすいと考えられる．このことから，行動 2 の割合を増加させることによって，行動 2 のみで増加しやすいパターンである (d)，(f)，(g) が顕著に増加したと考えられる．また (i) については行動 1 によって有名人をフォローする場合に有名人からフォローバックが返ってこない場合に増加しないため行動 1 より行動 2 の方が (i) を増加させやすいのではないかと考えられる．

また各パターンにおけるフォロー確率については図 8〜10 を見ると，行動 2 の増加とともにフォロー確率が増加したのは (a)，(b)，(c)，(g)，(h)，(i) であった．行動 2 によってパターン (a)，(c)，(h)，(i) に着目しているため，フォロー確率 (a)，(c)，(h)，(i) が増加傾向にあるのは想定した挙動であるが，パターン (b) と (h) については直接的に着目したわけではないが，i による k のフォローによって，二つのパターンにおいて i が k をフォローした三角形も増加し，また先述のように行動 1 と 2 の変化によってパターン (b) と (h) の数はあまり変化しないため，フォロー確率のみが増加したのではないかと考えられる．

その他の指標について，図 11〜13 の corr(in-out)，corr(in-recipro)，corr(out-recipro) を見ると，次数同士の 3 つの相関係数については行動 1 と 2 の割合変化による変動はあまり見られなかった．

これはどちらの行動もフォローするノードはランダムに選んでおり，また優先的選択についてもフォロワー数に大きく左右されやすいため各行動で相関傾向は変わりづらかったのではないかと考えられる．KL 情報量については，各次数分布で行動 2 を増加させるにつれて大きくなる傾向がみられたため，次数分布の良さという観点においては行動 1 のみを採用したモデルが適していると考えられる．密度については行動 1 と 2 の割合変化による変動は見られなかったが，クラスター係数については行動 2 との増加と共に顕著な増加を見せており，行動 2 によって密度を保ちながらクラスター係数を増加させることができることが分かる．モデルの特徴でも述べた通り，改善モデルではネットワーク内の三角形をより多くするノードをフォローしやすい傾向にあるためクラスター係数を増加させることができると考えられる．平均経路長については，行動 2 の増加と共に微妙に減少傾向がみられた．

5.3 ノード参入確率によるふるまいの変化

次に，ノード参入確率 $1-(\mu_1+\mu_2)$ の変化によるモデルのふるまいの変化について述べる．まず各パターンの数について，すべてのパターンについてノード参入確率の増加とともに総数が減少していた．これは単純にノード参入確率が増加したため行動 1，2 の選択確率が減少し，エッジが増えにくくなったためである．しかしながらフォロー確率については興味深い傾向がみられ，パターン (a)，(d)，(g) はノード参入確率の増加と共に減少傾向を見せており，パターン (b)，(e)，(h) は増加傾向，パターン (c)，(f)，(i) についてはどちらともいえない結果となった．各パターン数がノード参入確率と反比例したことを考えると，パターン (a)，(d)，(g) は行動 2 の増加によって増加しづらいパターンであるか，行動 2 が増えたことによって積極的にフォローが起こったパターンであると考えられる．逆にパターン (b)，(e)，(h) は行動 2 の増加と共にパターン数の増加がフォローの増加よりも急激に起こったのではないかと考えられる．

その他の指標について，KL 情報量に関してはノード参入確率の増加と共に増加しており，密度とクラスター係数についてはノード参入確率の増加と共に減少，相関係数，平均経路長については顕著な変化は見られなかった．この結果については，ノード参入確率によって影響されるネットワークの要素がエッジ数のみであるからだと考えられる．改善モデルにおいて，ノードが目標数 N に達するまでに必要な行動の総回数 T は以下で計算できる．

$$T = \frac{N}{1-(\mu_1+\mu_2)}$$

また，一度の行動で増加するエッジ数 ΔE を考えると

$$\Delta E = 1+\alpha_{ij}$$

これらから，目標ノード数 N に達したときのエッジ数と密度は以下で計算できる

$$\begin{aligned}|E| &= T\Delta E \\ &= \frac{N(1+\alpha_{ij})}{1-(\mu_1+\mu_2)} \\ D &= \frac{|E|}{N(N-1)} \\ &= \frac{N(1+\alpha_{ij})}{1-(\mu_1+\mu_2)} \cdot \frac{1}{N(N-1)} \\ &= \frac{(1+\alpha_{ij})}{(1-(\mu_1+\mu_2))(N-1)}\end{aligned}$$

よって，エッジ数と密度はノード参入確率 $1-(\mu_1+\mu_2)$ に反比例し，実験結果の傾向と一致することが分かる．また，ノード参入確率とクラスター係数が比例関係にあるという結果になったが，これはノード参入の時のフォローも行動 1 と同様のフォローであるため，行動 1 と行動 2 の割合が実質的には $1-\mu_2:\mu_2$ であることが原因ではないかと考えられる．相対的に行動 1 が増えることはハブとなるノードがさらにフォロワーをもらうようになるため，2-hop 先にあるハブをフォローした場合，ハブはフォロワー数が多いことから，2-hop のパスがたくさんあることが多く，三角形がたくさんできることが多い．改善モデル内では，行動 2 の回数増加よりもフォローするハブノードのフォロワー数増加の方がクラスター係数の増加に大きくつながったのだと考えられる．

5.4 フォローバック確率の設定方法によるふるまいの変化

まず α^1_{ij} を用いたネットワークとランダムなフォローバック確率 α^2_{ij} を用いたネットワークとを比較する．各パターンの総数について，パターン (a)，(b)，(d)，(e) では α^1_{ij} が多くなっており，パターン (c)，(h)，(i) では α^1_{ij} を用いたネットワークの方が小さくなった．各パターンを構成する 3 ノード i,j,k について，i と j，もしくは j と k がフォロー関係を作った際にランダム以上の確率でフォローバックが起こるために必要な共通フレンド数は式より 16 と多く，今回のエッジ生成割合では多くのノードでランダムなフォローバックの方が起こる確率が高いと考えられる．そのため α^2_{ij} が作るネットワークの方がパターンを構成する 3 ノード間でフォローバックが起こりやすかったため，相互フォローを含まないパターン (a)，(b)，(d)，(e) の総数は α^1_{ij} を用いたネットワークの方が大きくなったと考えられる．また，フォロー確率についてはパターン (c)，(f)，(h)，(i) では α^1_{ij} を用いたネットワークが多くなっており，パターン (a)，(b)，(d)，(e) では α^1_{ij} を用いたネットワークの方が小さかった．この確率の差についてはおもにパターン数の差が主な原因となっていると考えられる．先述の通り，パターン (a)，(b)，(d)，(e) の総数は α^1_{ij} を用いたネットワークの方が多いため，行動 2 で同じ回数優先的選択を行ったとしても確率は α^1_{ij} を用いたネットワークが小さくなるからである．

また，その他の指標について，相関係数とクラスター係数は各次数で同程度であり，密度と平均経路長は α^1_{ij} を用いたネットワークの方が小さく，KL 情報量は α^1_{ij} を用いたネットワークの方が大きくなってしまう結果となった．相関係数が同程度

であることから，ランダムなフォローバックによる次数への影響が実は小さく，各行動におけるフォロワー数に応じた優先的選択が相関係数に大きく寄与している可能性がある．また，密度が小さいことについてはフォローバックが全体的に起こりづらかったことを示唆しており，その結果を踏まえたうえでクラスター係数が同程度であることを考えると，共通フレンドを用いたフォローバックはクラスター係数を高めることに寄与している行動だと考えられる．たとえば行動 2 でフォローバックを行った場合，2-hop 経路が共通フレンドの数ぶん必ず増えるので，フォローバック確率が小さいとしても 2-hop 経路の数自体はランダムよりも大きく増える可能性が高かったため，クラスター係数の上昇につながったのではないかと考えられる．

次に，フォローバック確率をフォローした相手との比によって定めた α^3_{ij} を用いたネットワークと比較する．まず各パターンの数について，α^3_{ij} を用いたネットワークには α^1_{ij} を用いたネットワークと比べてパターン (d) の数が非常に多いこと，またパターン (c)，(h)，(i) が非常に少ないことがわかる．改善モデルではフォロー行動を行うノードはランダムに選ぶが，フォローするユーザを選択するときはフォロワー数に応じた確率で優先的選択を行うため，フォローするユーザがフォロー先よりもフォロワー数が大きい場合が少ないと考えられる．このため α^3_{ij} を用いたネットワークはフォローバックが他のネットワークに比べて起こりづらくなっており，相互フォローが存在するパターンの総数が極端に少なくなると考えられる．相互フォローを含まないパターンのうち (a)，(b)，(e) についても (パターン (a) の総数については大小の逆転が起こるが) α^1_{ij} と同程度の多さを持っており，またパターン (d) は非常に多いため α^3_{ij} も同様にランダムな確率よりも小さい確率でフォローすることが多く相互辺が現れにくいネットワークであると考えられる．また，パターン (d) の形が多いということは，フォローバック確率が α^1_{ij} を用いたネットワークよりもさらに小さいのではないかと考えられる．

フォロー確率については, パターン (a), (d), (g) では α^1_{ij} を用いたネットワークの方が高く，その他のパターンについては α^3_{ij} を用いたネットワークの方が高くなる結果となった．パターン (d) と (h)，(i) についてはそれぞれパターン数が多すぎることと少なすぎることによって α^1_{ij} と差が出ていると考えられる．　またその他の指標について，次数相関は corr(in-out)，corr(out-recipro) は α^1_{ij} を用いたネットワークの方が高く，corr(in-recipro) については α^3_{ij} を用いたネットワークの方が高くなった．これは α^1_{ij} が共通フレンド数という Out-degree に着目した確率であり，α^3_{ij} がフォロワー数の比という In-degree に着目した確率であるからである．自分がフォロー行動を行った際にフォローバックされる，つまり In-degree と Reciprocal-degree を増やす要因が out-degree である α^1_{ij} を用いたネットワークでは corr(in-out)，corr(out-recipro) が高くなり，要因が In-degree である α^3_{ij} を用いたネットワークでは corr(in-recipro) が高くなるの．

α^3_{ij} を用いたネットワークは α^1_{ij} を用いたネットワークに比べてクラスター係数が大きくなった．α^3_{ij} を用いたネットワークではフォローバックされるノードは自身のフォロワー数に影響されやすいため，ハブノードがフォローバックでさらにフォロワーを増やし成長しやすくなる．そのため前述のノード参入確率を増やした時と同じクラスター係数の増加につながったのではないかと考えられる．また平均経路長は α^1_{ij} を用いたネットワークに比べ小さくなった．これはフォローバックの際に 2 ノード以外の情報のみを用いるため，フォローされたノードからの距離によらずフォロワーのみでフォローバックする結果，今まで遠くにあったノードが相互フォローになりやすく，両方から短い距離でたどり着けるようになるからであると考える．密度は α^1_{ij} を用いたネットワークに比べ小さくなっており，α^1_{ij} に比べ α^3_{ij} の方がフォローバックが起きにくくなっていることが分かる．

5.5　実データとの比較による問題点

本節では実データと照らし合わせながらモデルのふるまいと現状の改善点に関しての比較考察を行う．

まずは密度に関して，ノード参入確率に反比例する密度が $1-(\mu_1+\mu_2)=0.05$ で実データの 0.5 倍ほどの値を取っているため，ノード参入確率を今の 0.5 倍にする必要があると考えられる．ノード参入確率との増加によって不足している各パターンの総数も増加させることができると考えられるが，フォロー確率も増加してしまう可能性があり，既にいくつかのフォロー確率は実データよりも高いため，現状の優先的選択をさらに改良する必要がある．α^1_{ij} を用いたネットワークにおいて行動 2 で着目したパターン (a)，(c)，(g)，(i) のうちフォロー確率が実データより高いものはパターン (a)，(c) であり，(g)，(i) ではフォロー確率が実データより低くなっていた．パターン (a) は相互辺を持たないパターンであることと，実データとの確率の差が一番大きいパターンが (i) であることを考えると，優先的選択で選ぶ経路に i と j，j と k が相互フォローであるかどうかを考慮した重みを導入し，よりパターン (g)，(i) のような三つ組を優先的選択しやすくなるようのが良いと考えられる．またその他のパターンも確率が少ないものも存在するため，別の三つ組に着目した行動を増やす解決策も考えられる．

ノード参入確率を減らした場合クラスター係数の減少につながるため，クラスター係数をより高める行動を導入する必要性がある．これに対しては，一回のフォロー行動でユーザが新規接続するノードを複数個にすることで解決できると考えられる．またフォローバック確率はフィッティングによって求めたものであったが，ランダムに設定したものと次数分布の忠実度があまり変わらずむしろ KL 情報量が大きくなる結果となっているため，別の観点からの分析も必要なのではないかと考えられる．全パラメータを通して実データよりも三つの相関係数が高くなってしまう問題がみられた．これはフォローバックやフォロー行動を行う際の優先的選択に共通フレンド数やフォロワー数といったユーザの情報を利用しているためと考えられる．その中でも α^2_{ij} も高い相関係数を持つことを考えると，フォロー行動でノードの次数に関する情報のみに依存しているのが大きな問題であると考えられる．そのためフォロー行動やフォロー

バックを行う際にランダムな要素や次数に依存しない要因を少し混ぜることによって相関係数を抑えることができる可能性があると考えられる．

6 結論と今後の展望

本論文では [1] で用いたデータからサンプリングしたネットワークを用いてフォローを行うノード i・フォロー先となるノード k・二つのノードを繋ぐ中間点となるノード j から成る三つ組に関しての分析を行った結果，ノード j による拡散行動によって i が k を観測しやすい構造の場合に i のフォロー確率が高いこと，そして同じ構造を持つ三つ組の場合は j の数が多いほど i のフォロー確率が高くなることが判明した．また，ユーザ j が i をフォローバックする確率は共通フレンド数と相関があり tanh 関数を用いて高い精度で近似できることを示した．

次に改善モデルにおけるパラメータの変化による次数分布の変化の観察を行った結果，行動 2 の割合を変化させることでネットワークの密度を保ちながらクラスター係数を増加させることが可能なモデルであり，ネットワークの三つ組の数を調整できることを示した．しかしながら，行動 2 の増加によって KL 情報量も増加してしまい，行動 2 の増加によって次数分布をより忠実に再現することは達成できなかった．三つのフォローバック確率を用いたネットワークについて比較したところ，分析結果からフィッティングしたフォローバック確率よりはランダムなフォローバック確率で生成したネットワークに次数分布の忠実度は少し劣るが，パターン数やその他の指標で違った特徴を見せた．改善モデルでは実データすべての特徴を再現することが出来なかったため，さらなる分析によるパラメータの調整やフォローバック確率の推定，モデルの改善や拡張の必要性があると考えられる．

モデルの拡張については様々存在するが，その中でも現実のユーザ行動に即したものとして，以下のような例が考えられる．

ユーザの成長によるポリシー変化

改善モデルは，初期条件として行動の選択割合を示すパラメータ μ_1, μ_2 を設定したが，ユーザはネットワーク上での地位の変化とともにポリシーが変化する可能性がある．

ネットワークの成長によるノード参入確率変化

改善モデルではノード参入確率がネットワークの大きさに対して一定であり，ノード数の成長速度に対してネットワークの平均次数の成長速度が徐々に小さくなるという実世界の SNS の成長ではあまり考えづらい状態となっていた．これを解消するために，ネットワークの成長とともに一度のノード参入までにネットワークに増加するエッジ数をネットワークの大きさに応じて増加させることで，ノード数の増加に対して平均次数の増加度合が小さくなる問題を解決できる可能性がある．

フォロー行動を行うユーザの選択

改善モデルでは行動 1，2 においてフォロー行動を行うユーザをランダムに選択していた．しかしユーザは様々な目的で X を利用しているため，フォロー行動を行いやすいユーザや行いづらいユーザが存在すると考えられる．適応度をノードに導入することで，このような多様性を実現可能であると考える．

行動の追加

改善モデルでは特定のパターンについてのみ着目したが，その他のパターンについての着目も考えられる．また，優先的選択についてもさらなる拡張ができると考えられる．例えば有名人をフォローする際も hop 数を制限することで，あまり遠くのユーザとつながるといった現実の行動に反するようなフォローを制限できると考える．

以上のような拡張を行うにあたってはユーザの行動を分析する必要があるが，その際は動的なデータを扱う必要がある．例えばユーザのポリシー変化については時刻ごとの新規フォロー先を集計し，その変化を観察する必要があり，フォロー行動を行うユーザの発見には単位時間当たりのフォロー数の集計などが必要であるからである．そのため今後は時間的なデータを取得することでより忠実にユーザ行動を分析し，より正確なネットワーク生成モデルに拡張していきたい．

7 謝　　辞

本研究は JSPS 科研費 21H03446 の助成を受けたものです．

文　　献

[1] 山川衛・田島敬史,「ソーシャルネットワークの特徴に基づいた有向ネットワーク生成モデルの提案」, 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023.

[2] 増田直紀, and 今野紀雄. 複雑ネットワーク: 基礎から応用まで. 近代科学社, 2010.

[3] Leskovec, Jure, and Christos Faloutsos. ”Sampling from large graphs.” Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006.

[4] HENZINGER, Monika R., et al. On near-uniform URL sampling. Computer Networks, 2000, 33.1-6: 295-308.

[5] Watts, Duncan J., and Steven H. Strogatz. ”Collective dynamics of ‘small-world’ networks.” nature 393.6684 (1998): 440-442.

[6] Lou, T., Tang, J., Hopcroft, J., Fang, Z. and Ding, X.: Learning to predict reciprocity and triadic closure in social networks, ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 7, No. 2, pp. 1–25 (2013).

[7] Heider, F.: The psychology of interpersonal relations, Psychology Press (2013).

イベント発生の不均衡性に適合した人口変動予測

塚田涼太郎† 豊田 正史†† 梅本 和俊†† 是津 耕司†††

† 東京大学大学院情報理工学系研究科 〒 113–8656 東京都文京区本郷 7-3-1
†† 東京大学生産技術研究所 〒 153–8505 東京都目黒区駒場 4-6-1
††† 情報通信研究機構 〒 184–8795 東京都小金井市貫井北町 4-2-1
E-mail: †{tsukada,toyoda,umemoto}@tkl.iis.u-tokyo.ac.jp, ††zettsu@nict.go.jp

あらまし 多くの参加者が集まるイベントは，その発生場所周辺において人口の急変動を引き起こす．こうした人口変動に起因する混乱を避けるためには，人口変動の発生時間や規模を事前に知ることが必要である．しかしながら，イベントの発生時間や参加者数は様々であり，またイベント日は非イベント日と比べて少数の事例しかないことから，イベントに起因する人口変動をその発生場所の過去の人口情報のみに基づいて予測することは難しい．本研究では，都市における多様な種別のイベント発生場所の数日先までの人口予測のために，未来の日付とイベント発生場所の名称の両方に言及しているマイクロブログ上の投稿の内容およびその場所の過去の人口情報を融合し， クラスタリングに基づくオーバーサンプリングによって学習データの不均衡性を解消する手法を提案する．混雑統計 ® データと Twitter データを用いた実験の結果，東京・神奈川の主要なスポーツ会場・コンサート会場・展示会場・公園・デモ会場・観光地の 24 時間先までの人口予測について，提案手法が過去の人口情報のみを用いるベースライン手法の予測誤差を改善することを確認した．

キーワード 人口変動予測，マイクロブログ，不均衡データ

1 はじめに

都市において野球の試合・コンサート，花見・デモ行進などのイベント[1]発生時には，参加者が発生場所付近に集まることで突発的に人口が増加する．こうした人口の急増は参加者自身だけでなく，周囲を巻き込んで様々な悪影響を及ぼし得る．例えば，通常は空いている時間帯の道路で大量のイベント帰りの客に起因する渋滞が発生すると，その道路を利用しているバスやトラックの交通網に想定外の遅延が引き起こされることになる．また，イベントによる混雑の発生を知らずに観光地を訪れた観光客にとっては，予定通りの電車に乗れないなどの事態に遭遇して旅行の満足度が低下する．

イベント発生場所の人口変化の予測は，人口の急増に伴う諸問題を解決する上で重要な役割を果たす．例えば，道路の利用者や観光客は予測される混雑を避けて予定を立てることでそのリスクを低減できる．より積極的な利用方法として，コンビニなどの小売店の販売網において人口の急増に伴う需要増加に応じ適切に商品を配置することで，売上を増やすことも考えられる．こうした動機のもと，我々のこれまでの研究 [2–4] では，マイクロブログ上に未来のイベントに言及している投稿が存在し，イベントの発生時間・場所・規模を予測する上で有用な情報となることを利用して，都市における多種多様なイベント発生場所の人口の数日先までの予測の問題に取り組んだ．以下の投稿は，そのアイデアを端的に表している．

> 明日の東京ドームのコンサート、チケット取れなかった

すなわち，これが 3 月 1 日の投稿であった場合，翌 3 月 2 日に東京ドームでコンサートが開催されること，およびチケットが取れなかったことから来場者数は満員に近いことが読み取れる．実際に，このような投稿の内容のテキストを解析して当該会場の過去の人口の時系列のデータと融合するアプローチによって，過去の人口情報のみを用いるベースライン手法より高い精度の予測が可能であることを確かめた．

しかしながら，より高精度な予測を実現するためには，イベント発生に係る以下 2 つの不均衡性が問題になる：

（1） イベント日が非イベント日に比べて非常に少ない
（2） イベントが多い会場と少ない会場が極端に分かれている

そこで本研究では，(1) k-means クラスタリングに基づくオーバーサンプリングによりイベント日の事例を水増しし，(2) 同一種別の会場のデータを利用した同時学習を行う手法を提案する（第 3 節）．実世界の混雑統計 ® データと Twitter（現 X）データを用いた実験の結果，提案手法が当該会場の人口情報のみを用いる場合と比較して予測誤差を最大 30%以上改善することを示した（第 4 節）．

2 関連研究

2.1 位置情報やマイクロブログを用いた人口変動予測

イベントの発生によって突発的に変化する都市の人口を予測するために，携帯電話などから収集された位置情報を含むデー

1：イベント検出に関する既存の研究 [1] における定義を参考に，本研究で扱うイベントは時空間上のある点で発生する人の自発的な集まりとする．これには，主催者が日程を定めるもの（例：野球の試合，コンサート）ならびに参加者が自然発生的に集合するもの（例：花見，デモ行進）の 2 種類がある．

タを未来の人の動きの手がかりとして活用する研究が行われている．Fan ら [5] は，大規模イベント開催期間における都市全体の人々の動きのモデル化に取り組んだ．Fan らの手法は，大規模イベントの開催時には会場に早く到着する人と遅れて到着する人がいるという仮定の下で直近の人々の移動経路を含むデータを活用するものであり，都市全体の規模での人口をイベントの影響下においても予測可能であると報告されている．未来の人の動きを把握するためのより直接的なアプローチとして，乗換案内や地図のアプリケーションに入力される検索クエリのログデータを予測に活用する取り組みも行われている．Konishi ら [6] は，乗換案内アプリケーションの検索クエリログを活用することで，花火大会などのイベントに伴う都市全体規模の駅の混雑を 1 週間前の時点で予測した．Liao ら [7] は，地図アプリケーションにおけるイベント会場を目的地とした検索回数がイベント開始時刻直前に増加する現象に着目して，イベント会場周辺の交通速度の数時間先までの予測を行った．

一方で，マイクロブログ上の投稿をソーシャルセンサーとして用い，実世界で起きているイベントを把握する試みもなされてきた．Yamada ら [8] は，旅行の計画に役立つローカルなイベント情報を Twitter 上の投稿から抽出する研究を行った．Yamada らの手法は，イベント会場名を含む投稿からイベントの開催期間を取得し会場名の表記揺れを考慮して集約する．この手法により，Yamada らは従来手法より正確なイベント情報抽出ができることを示した．He ら [9] は，Twitter 上の投稿に交通状況への言及があることに着目し，投稿に含まれる単語の情報を用いて自動車の交通量を比較的長期にわたって予測する手法を提案した．He らの手法は，投稿に付与されている位置情報を用い，交通量の予測対象となる区画内で発信された投稿のみをフィルターして利用している．しかしながら，2019 年に Twitter は投稿に正確な位置情報（ジオタグ）を付与する機能を廃止[2]したため，この手法は現在適用不可能である．

これらの研究の流れを合流させ，我々のこれまでの研究 [2–4] では，マイクロブログ上で未来のイベントに言及している多様な投稿の中からイベント発生場所の名称と日付に言及している表現の両方を含む投稿に着目し，未来のイベント参加者数の手がかりとして利用した．自然言語の形で未来のイベントの位置と時間の情報が記述されているマイクロブログ上の投稿は，幅広い利用者から容易に収集可能であるという利点がある．さらに，予測した人口変化の原因について投稿内容による説明を加えられる．予測された人口変化の原因が分かればより適切な対策の実施が可能になるため，予測結果の説明可能性は実用上重要な利点である．しかしながら，第 1 節で述べたように，この手法はイベント発生の不均衡性に起因する学習データの品質の制約の影響を受けてしまう．

2.2 不均衡データに対する機械学習

主に分類問題について，一方のクラスのデータが他方のクラスのデータと比べて極端に少ないデータのことを不均衡データと呼ぶ．例えば学習データの中で，予測の対象として注目したい正例が負例よりも極端に少ないような不均衡データの場合，分類結果が負例に偏ってしまい精度が落ちてしまうため，不均衡データに対する機械学習の手法は従来から研究されている．

その手法の一つとして，オーバーサンプリングがある．オーバーサンプリングは，少ない方のクラスのデータを水増しすることで不均衡データのバランスを改善する手法である．SMOTE [10] は，少ない方のクラスのデータの近傍データを補完して作り出すオーバーサンプリングを行う．SMOTE の拡張版も数多く提案されている [11]．Borderline-SMOTE [12] は，クラス境界付近のデータを重点的にオーバーサンプリングすることで，分類を容易にする手法である．ADASYN [13] は，オーバーサンプリング対象となる少数データの近傍のローカルなデータ分布を考慮して補完データを作成する手法である．K-Means SMOTE [14] は，SMOTE によるオーバーサンプリングを行う前に k-means クラスタリングによるクラスター分けを行う手法で，少数データのバリエーションを保ったまま補完データを作成できる．

オーバーサンプリングは，都市における人口変化予測において，イベント発生の事例が通常時の事例と比べて著しく少ない不均衡データになる問題に対しても適用が試みられている．Anno [15] らは，乗換案内アプリケーションの検索クエリログと携帯端末の位置履歴の情報を組み合わせて，非日常的なイベントについても安定的な混雑予測が行えるためのオーバーサンプリング手法を提案した．

3 提案手法

3.1 マイクロブログ投稿の集合と人口の時系列を入力とした予測モデル

予測モデルのアーキテクチャは我々の以前の研究 [4] のモデルから，簡単のために時刻表現エンコーダを取り除いたものである．以下に概要を説明する．

イベント発生場所で将来発生する人口変化を予測する手がかりとして，未来の日付とイベント発生場所の名称の両方に言及しているマイクロブログ上の投稿を用いる．ここで，イベント発生場所 v のある未来の日付 d の人口予測に有用な投稿は，以下の条件（以降，**抽出条件**）をすべて満たすと考えた：

- d に言及する日付の表現を含む
- イベント発生場所 v の名称を含む
- d の前日以前に投稿されている

詳細は第 4.1 節で述べるが，抽出条件を満たすイベント関連投稿のカバレッジを上げるために，検索クエリを拡張して日付表現ならびにイベント発生場所の名称の表記揺れをカバーした．

投稿の集合を集約する部分は，Set Transformer（以降，**ST**）[16] を中心とした構造になっている．ST を用いる動機を以下に示す：

- ST は個別の投稿を一つ一つ処理できる．投稿の集合を単一のベクトルで表現する Bag-of-Words のアプローチとは

2：https://twitter.com/TwitterSupport/status/1141039841993355264

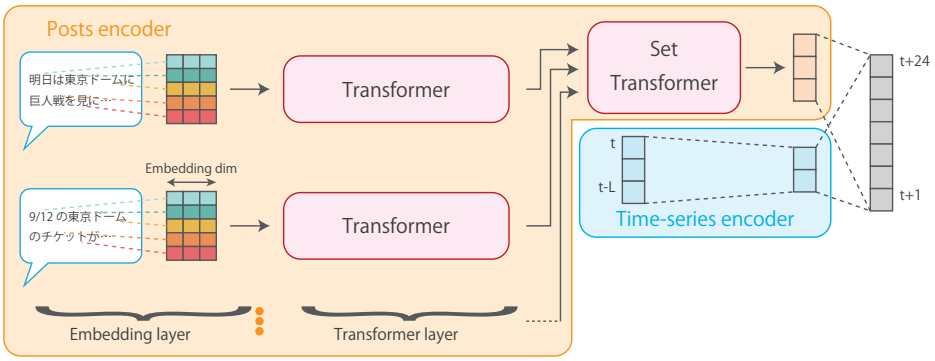


図 1　提案手法のモデルの概要.

表 1　各イベント発生場所の種別，テストデータ中の年間イベント日数，訓練データ中の投稿数.

種別	名称	イベント日数	投稿数
サッカースタジアム	日産スタジアム	15	6,229
	味の素スタジアム	29	15,842
野球場	東京ドーム	131	82,629
	明治神宮野球場	87	7,842
	横浜スタジアム	72	20,892
展示会場	パシフィコ横浜	53	21,310
	東京国際展示場	80	79,237
コンサート会場	横浜アリーナ	106	52,115
	日本武道館	114	41,507

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

対照的に，ST では集合に含まれる投稿間の複雑な関係性を考慮できることが期待される．

- 各種のデータの特徴を考慮した表現を学習できる．GBR は異なる種類の入力を対等に扱うが，NN モデルは異種データのそれぞれを効果的に扱うための事前知識を反映した個別のエンコーダを使える．データ融合を用いた近年の交通予測の研究 [7,17] でその有効性が示されている．
- 多くの投稿で言及されている場所については，NN モデルが高い性能を発揮することが期待される．

図 1 に ST モデルの概要を示す．ST は，set-input 問題に適用可能な自己注意の機構に基づく順序不変なモジュールである．モデルは，投稿・時系列の 2 つのエンコーダからなる．投稿エンコーダでは，ST が投稿ベクトルの集合を単一のベクトルにエンコードする．ここで，各投稿は Transformer [18] でエンコードされる．各投稿の先頭に付与した<CLS>トークンを，Transformer でエンコードした時の埋め込みをその投稿のベクトルとして用いる．これは，既存の研究 [19] において分類タスクを解く際に文の表現を得るための方法である．ここで，訓練時に出現しない単語には未知語トークン（<UNK>）を割り当てる．時系列エンコーダでは，イベント発生場所の人口変化の時系列における短期的トレンドを捉えるために，直近の 1 時間ごとの人口の時系列 $24 \times n$ ステップ分のベクトルが全結合層に入力される．ここで，n は過去の人口情報を何日前まで遡って用いるかを決めるハイパーパラメータである．データ融合による交通予測の既存のアプローチ [7] を参考に，これら 2 種類の入力（投稿，時系列）の埋め込み表現を連結し，後続の全結合層に入力する．

3.2　オーバーサンプリングと同時学習による不均衡性の緩和

表 1 に，本研究で予測対象とするイベント発生場所の年間イベント日数を記載した．この表からは第 1 節で問題提起したイベント発生の不均衡性を確認できる：

（1）　イベント日が非イベント日に比べて非常に少ない
（2）　イベントが多い会場と少ない会場が極端に分かれている

具体的には，(1) 最もイベント日数の多い東京ドームであっても過半数は非イベント日であることが分かる．また，(2) 例えば東京ドームと横浜スタジアムはどちらも野球場であるが，イベント日数としては東京ドームのほうが横浜スタジアムの 2 倍近い値になっている．このような 2 種類のイベント発生の不均衡性の問題は，イベント発生場所の人口変化予測を難しくする要因である．最も単純な解決策は，過去の人口情報を収集する期間を長くすることでイベント日の事例を学習データに多く含めることであるが，特定の場所の人口情報を長期間・一定品質で入手することは実用上困難である．そこで，本研究ではオーバーサンプリングと同時学習という 2 つの学習時の工夫により，この不均衡性の問題に対処する．

オーバーサンプリング：予測の対象として特に注目したいのはイベント日であるため，オーバーサンプリングによってイベント日のデータを水増しすることを考える．ただし，第 3.1 節で述べたように，入力データはマイクロブログ上の投稿の集合と人口の時系列からなる複合的なデータであるため，SMOTE [10] のような離散データに対する補完手法をそのまま適用することは難しい．そこで，人口の時系列に対して k-means クラスタリングを行い，各クラスターから等確率でサンプリングすることで，イベント日と非イベント日が均等に混在した学習データを作成する．具体的には，以下の操作を繰り返して学習データを拡張していく．

（1）　正規化した混雑度データに対して k-means でクラスタリ

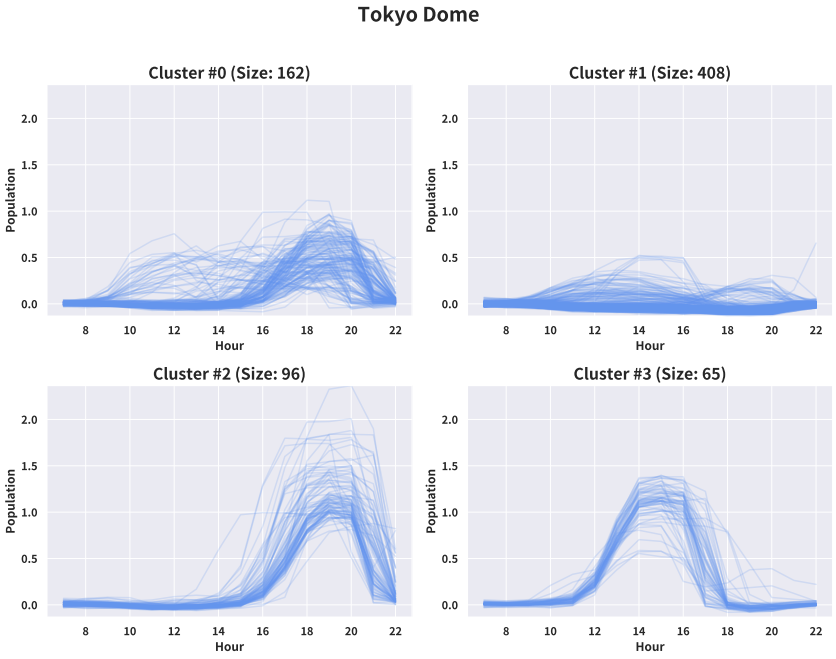


図 2　東京ドームのクラスタリング結果（$k = 4$）．人口は正規化されている．
「混雑統計 ®」©ZENRIN DataCom CO., LTD.

ングする

（2） 各クラスタを $\frac{1}{k}$ の確率で選択する

（3） 選択したクラスタの中から 1 日分をランダムに取り，学習データに追加する

これは，イベント発生場所の人口の時系列に対して k-means クラスタリングを行うと，イベント日と非イベント日のクラスターが分かれるという観察に基づいている．例として，東京ドームにおける $k = 4$ の場合のクラスタリング結果を図 2 に示す．このようにオーバーサンプリング結果の解釈が視覚的に容易であることも，k-means クラスタリングを用いる利点である．ここで，k はハイパーパラメータになる．

同時学習：東京ドームと横浜スタジアムを野球場という種別で見たときに，入力データと予測される人口の時系列との関係は類似していると考えられる．例えば「プロ野球のデーゲームに関する投稿が多数あれば，昼過ぎにピークのある人口変化が予想される」ことは，東京ドームでも横浜スタジアムでも同じであるはずである．こうした仮定のもと，イベント日数の多い会場のデータを有効活用するために，同一種別の会場をまとめて学習する同時学習の手法を導入する．具体的には，以下の手順で予測モデルを学習させる．

（1） 同一種別の全会場の学習データ（オーバーサンプリング済み）を結合して同時学習用の学習データとする

（2） 上記の学習データを利用して同時学習を行う

（3） ターゲット会場のデータのみを使って再学習（ファインチューニング）を行う

4　評価実験

提案手法の有効性を実世界データを用いて評価した．実験は 3 種類行い，それぞれ下記の問いに答えられるように設計した：

Q1　提案手法はイベント発生日と非発生日を区別できるか？（第 4.2 節）

Q2　異なるタスク設定において提案手法はどれほどの精度で人口変化を予測できるか？（第 4.3 節）

Q3　提案手法はイベント発生の不均衡性の問題をどのように改善しているか？（第 4.4 節）

4.1　実験設定

データセット：実験に用いたデータセットは，時空間人口統計データとマイクロブログ投稿データからなる．

時空間人口統計データについては，混雑統計 ® データ[3]を使用した．混雑統計 ® データには，約 250m 四方のメッシュごとに，そのメッシュ内の滞在人数の推定値を 1 時間単位で集計した値（混雑度）が含まれる．このうち 2014 年 12 月から 2018 年 11 月までの 4 年分について，最初の 2 年分を訓練，次の 1 年分を開発，最後の 1 年分をテストに用いた．表 1 に示す通り，

3：「混雑統計 ®」データは，NTT ドコモが提供するアプリケーション（※）の利用者より，許諾を得た上で送信される携帯電話の位置情報を，NTT ドコモが総体的かつ統計的に加工を行ったデータ．位置情報は最短 5 分毎に測位される GPS データ（緯度経度情報）であり，個人を特定する情報は含まれない．またデータの加工には「非特定化」「集計処理」「秘匿処理」がなされており個人が特定されることはない．※ドコモ地図ナビサービス（地図アプリ・ご当地ガイド）等の一部のアプリ．

表 2　イベント有無の予測性能（Precision，Recall，F1）．

	ベースライン手法						提案手法								
	AR			LSTM			単一・OS なし			単一・OS あり			同時・OS あり		
イベント発生場所	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
日産スタジアム	0.33	0.07	0.11	0.00	0.00	0.00	0.58	0.47	0.52	0.48	0.87	**0.62**	0.73	0.53	0.62
味の素スタジアム	0.00	0.00	0.00	1.00	0.04	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.57	**0.70**
東京ドーム	0.85	0.24	0.37	0.81	0.20	0.32	0.65	0.66	0.65	0.80	0.71	0.75	0.81	0.77	**0.79**
明治神宮野球場	0.64	0.31	0.42	0.87	0.15	0.26	0.77	0.51	0.62	0.72	0.51	0.60	0.80	0.81	**0.81**
横浜スタジアム	0.65	0.21	0.32	0.83	0.07	0.13	0.97	0.54	0.70	0.90	0.50	0.64	0.92	0.96	**0.94**
パシフィコ横浜	0.39	0.32	0.35	0.33	0.13	0.18	0.67	0.51	**0.58**	0.67	0.51	0.58	0.62	0.51	0.56
東京国際展示場	0.45	0.67	0.54	0.56	0.41	0.47	0.52	0.70	0.59	0.52	0.70	0.59	0.51	0.81	**0.62**
横浜アリーナ	0.43	0.21	0.28	0.55	0.21	0.30	0.54	0.47	0.50	0.54	0.47	0.50	0.54	0.62	**0.58**
日本武道館	0.43	0.27	0.33	0.42	0.15	0.22	0.62	0.52	**0.57**	0.55	0.50	0.53	0.64	0.49	0.55
マクロ平均	0.46	0.25	0.30	0.60	0.15	0.22	0.59	0.49	0.52	0.57	0.53	0.53	0.72	0.67	**0.68**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

予測対象のイベント発生場所には，異なる種別のイベントが発生する東京・神奈川の 9 地点を選んだ．各イベント発生場所についてその場所を囲むメッシュを選び，その混雑度をその場所周辺の人口とした．ここで，過去の人口変化の時系列に現れる週毎のパターンを捉えるために，第 3.1 節のハイパーパラメータ n は $n = 7$ とした．この時系列データは，通勤・通学に起因する日毎や週毎の周期的なトレンドを含む．時系列データの解析に先立っては，こうしたトレンドの除去が重要である [20]．そこで，1 日の各時間ごとに過去の人口から算出した定常人口を時系列データから引くことでこれらのトレンドを除去した．なお，定常人口は平日・祝休日のそれぞれに対して以下の要領で算出した値である．

（1） 各日の人口を $k = 3$ で k-means クラスタリングする
（2） 重心の最大人口が最小のクラスタ重心を定常人口とする

マイクロブログ投稿データについては，研究室で収集している Twitter アーカイブを利用した．[4]第 3.1 節の抽出条件を満たすイベント関連投稿のカバレッジを上げるために，Wikipedia のリダイレクトデータから名寄せ辞書を構築して，イベント発生場所の名称の表記揺れ（通称や旧称など）に対応した．抽出条件を満たした投稿は，JUMAN[5]で分かち書き処理を行いストップワードを除去した．抽出条件を満たす場所の名称と日付表現は，それぞれ特別なトークン（<TARGET_PLACE>および<TARGET_DATE>）で置換することで表記揺れを解消した．

ベースライン手法：提案手法を 2 つのベースライン手法と比較した．具体的には，AR および LSTM モデルを用いる手法である．LSTM のハイパーパラメータは，PyTorch[6]の実装のデフォルト値とした．これらの手法はマイクロブログ投稿を使わない手法であり，24 時間先までの人口を自己回帰的な方法（前の時点の出力を次の時点の入力にする）で予測する．

評価指標：提案手法の有効性の評価は，イベント有無のレベルと人口変化のレベルの 2 つの粒度で行った．

イベント有無のレベルの評価では，各手法がイベント発生日と非発生日を分類できるかを確認した．しかしながら，多くのイベント発生場所でイベント予定表は公開されていない．そこで，場所ごとに 1 日の人口の時系列に対して k-means クラスタリングを行い，要素数が最大のクラスタ（以降，最大クラスタ）に属する日をイベント非発生日，そうでない日を発生日とした．ここで，クラスタ数 k は $k = 3$ とした．これは，訓練データを用いた予備実験において $k = 3$ とした場合，イベント発生日が属する 2 つの小さいクラスタと，非発生日が属する 1 つの大きいクラスタに分かれる傾向にあるという分析に基づく．表 1 に各イベント発生場所のテストデータ中のイベント日数を示す．テスト時には，回帰である各手法による出力の予測値がどのクラスタの平均値に最も近いかという基準で，その出力をイベント発生日または非発生日のクラスに分類して扱った．具体的には，予測値が最大クラスタの平均値に最も近い場合はイベント非発生日，そうでない場合は発生日とした．この基準により，各手法の性能をイベント有無の 2 値分類問題として Precision，Recall，F1 で評価した．

人口変化のレベルの評価では，実際の人口の規模をどの程度正確に予測できるかを確認した．評価には式 (1) で定義される平均絶対誤差（MAE）を用いた．

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N} |\hat{X}_t - X_t| \tag{1}$$

ここで，$X_t, \hat{X}_t$ はそれぞれ時刻 t における実測値，予測値であり，N は評価対象のデータの総数である．

4.2 Q1：イベント有無の予測

最初に，イベント有無のレベルでの予測結果を表 2 に示す．

4：このアーカイブは，Twitter API によるクローリングを 2011 年 3 月から継続的に行うことで構築されており，約 250 万人の公開ユーザーのタイムラインからなる．このクローリングは，日本の有名人の 30 ユーザーから開始して，そのタイムライン上のリツイートとメンションを追っていくことで繰り返しユーザーの集合を拡張してきたものである．
5：http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
6：https://pytorch.org/

表 3　$n = 7$ の場合の全期間の予測誤差（MAE）.

	ベースライン手法		提案手法		
イベント発生場所	AR	LSTM	単一・OS なし	単一・OS あり	同時・OS あり
日産スタジアム	465.11	415.86	356.00	358.40	**343.80**
味の素スタジアム	723.68	690.36	605.50	558.30	**493.60**
東京ドーム	6,819.39	6,681.91	4,471.80	4,471.80	**2,940.70**
明治神宮野球場	2,041.24	1,926.30	1,419.80	1,482.60	**1,285.90**
横浜スタジアム	1,872.04	1,716.61	1,422.10	1,353.60	**1,058.30**
パシフィコ横浜	2,106.85	1,981.06	1,471.80	1,471.80	**1,467.20**
東京国際展示場	2,471.88	2,361.67	1,911.00	1,911.00	**1,901.60**
横浜アリーナ	1,449.69	1,400.66	1,272.00	1,272.00	**1,240.40**
日本武道館	1,618.61	1,727.46	1,272.70	1,272.70	**1,195.20**
マクロ平均	2,174.28	2,100.21	1,578.08	1,572.47	**1,325.19**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

表 4　イベント発生日の予測誤差（MAE）.

	ベースライン手法		提案手法		
イベント発生場所	AR	LSTM	単一・OS なし	単一・OS あり	同時・OS あり
日産スタジアム	2,543.48	2,797.59	**1,737.70**	1,945.70	1,848.00
味の素スタジアム	3,811.12	3,813.54	4,141.50	3,126.50	**2,893.40**
東京ドーム	10,124.14	11,877.79	6,465.10	6,465.10	**4,702.50**
明治神宮野球場	4,113.13	4,822.52	3,199.50	2,996.10	**2,594.70**
横浜スタジアム	4,025.14	4,769.67	3,860.60	3,276.20	**2,136.40**
パシフィコ横浜	4,786.35	5,216.44	2,932.00	2,932.00	**2,925.80**
東京国際展示場	3,116.17	4,018.55	2,329.50	2,329.50	**2,216.30**
横浜アリーナ	2,307.84	2,396.94	1,938.70	1,938.70	**1,769.60**
日本武道館	2,405.64	2,596.76	2,151.50	2,151.50	**2,050.90**
マクロ平均	4,137.00	4,701.09	3,195.12	3,017.92	**2,570.84**

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

オーバーサンプリングと同時学習を行った提案手法は，多くの場所で一貫して高い性能を達成した．特に，味の素スタジアムのようなイベント日数が少ない場所では，単一場所のデータをオーバーサンプリングせずに利用する場合と比較して予測精度が大きく向上した．このような場所では極端に少ない数のイベントしか発生しないため，入力の時系列は人が集まらない日の系列になることが多く，イベント発生日の時系列を予測するために有用な手がかりがほとんど得られない．提案手法はオーバーサンプリングによりイベント日のデータを水増しし，同時学習により同一種別の他の場所のデータを活用することで，イベント発生の不均衡性に対処してイベント発生日を判別できた．

イベント発生場所の種別ごとに Precision, Recall を見ると，東京ドーム・明治神宮野球場・横浜スタジアム（野球場），横浜アリーナ・日本武道館（主にコンサートが開催される会場）において，提案手法による予測精度の改善度合いが大きい．原因として，これらの場所における発生イベントが特徴的で，人口の時系列変化も類型化しやすいことが考えられる．例えば，プロ野球の試合は野球場の発生イベントの多数を占めるイベントだが，未来の野球の試合に言及する投稿にはチーム名などの特徴的な単語が頻出する．また，コンサート会場ではコンサートに言及する投稿が出演者の名称などを含んでいる．そのような情報は会場によらず有用な情報であるため，同時学習の過程で汎化されて，イベント発生日の予測精度向上に寄与したと考えられる．

4.3　Q2：人口変化の予測

次に，人口変化のレベルでの予測結果を表 3 に示す．表 3 はテストデータの全期間で測定された人口変化の予測誤差を示している．すべての場所において，提案手法は単一場所のデータをオーバーサンプリングせずに利用する場合より小さい誤差を達成した．誤差の削減の度合いは特に野球場で大きく，第 4.2 節のイベント有無の予測結果と同様の傾向である．

テストデータのほとんどの日は非イベント日であるため，テストデータの全期間での評価は予測誤差を過小評価することになる．そこで，予測誤差をイベント日だけで評価した（表 4）．実応用上重要なイベント日における人口変化の予測はより難しいタスクだが，その場合も提案手法はベースライン手法と比較してほぼすべての場所で最良の結果となった．

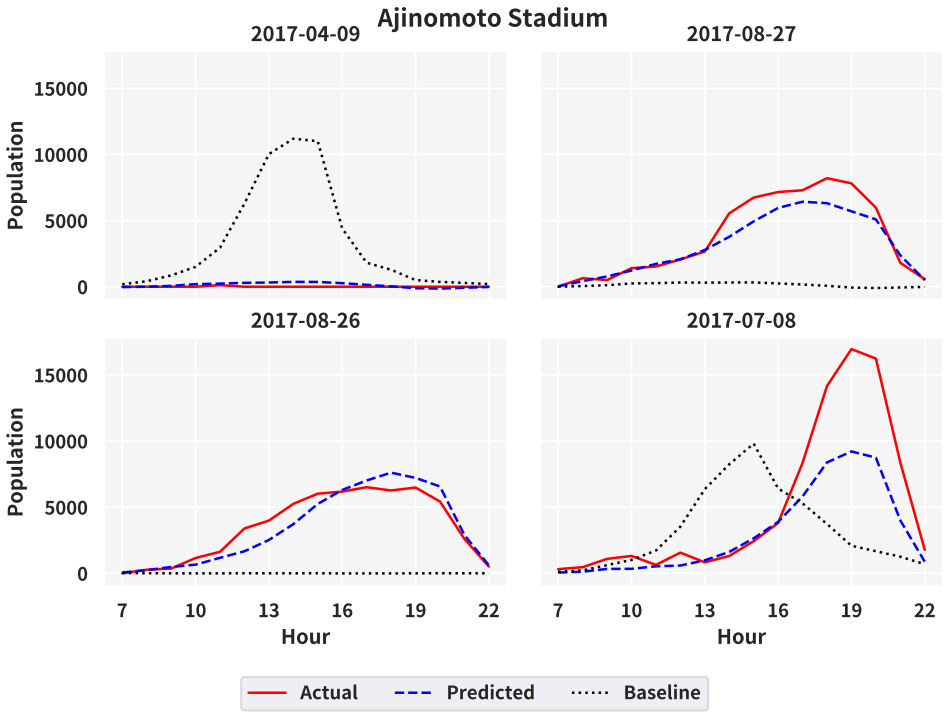


図 3　予測精度の改善幅が上位の日の実測値と予測値（味の素スタジアム）．

「混雑統計 ®」©ZENRIN DataCom CO., LTD.

4.4　Q3：事例分析

最後に，提案手法のオーバーサンプリングと同時学習が，イベント発生の不均衡性に適合できているかを実際の予測事例から分析する．

図 3 は，味の素スタジアムにおける予測精度の改善幅上位 4 例の日について，実測値と予測値（提案手法とベースライン）をプロットしたものである．2017 年 4 月 9 日の例では，実際にはイベントがないにも関わらずベースライン手法ではイベントがあるような時系列を予測しているが，提案手法ではイベントがないことを予測できている．2017 年 8 月 26 日，8 月 27 日の例では，実際にイベントが発生しており，ベースライン手法ではそのことを予測できていないが，提案手法では人口の規模・タイミングともに概ね予測できている．実際，この 2 日間は味の素スタジアムを会場としたコンサート[7]が開催されていた．2017 年 7 月 8 日の例では，提案手法もベースライン手法もイベントがあることを予測できているが，提案手法の方がより実測値に近いタイミングで予測できている．この日は 19 時キックオフの J リーグの試合[8]が開催されていた．

従来の手法は，イベント発生の不均衡性の影響を大きく受けるため，実際にはイベント日だったとしても非イベント日に近いような時系列を予測してしまう傾向にある．その問題が，オーバーサンプリングと同時学習を行う提案手法では改善されていることが分かった．特に，コンサートやサッカーの試合のような他会場の情報を汎化して利用できるイベントについて，想定通り予測精度が改善されていることが確かめられた．

7：https://www.ajinomotostadium.com/schedule/2017/0827_33.php
8：https://www.ajinomotostadium.com/schedule/2017/0708_53.php

5　おわりに

本研究では，イベント発生場所周辺の人口変化の予測における，イベント発生の不均衡性の問題に取り組んだ．イベント発生には (1) イベント日が非イベント日に比べて非常に少ない (2) イベントが多い会場と少ない会場が極端に分かれているという 2 つの不均衡性があり，このことが予測精度を悪化させる原因になっていた．そこで，(1) k-means クラスタリングに基づくオーバーサンプリングによりイベント日の事例を水増しし，(2) 同一種別の会場のデータを利用した同時学習を行うという 2 つのアプローチを提案した．実世界の混雑統計 ® データと Twitter データを用いた実験によって，提案手法が従来手法より優れていることを示した．

今後の研究の方向性として，より小規模なイベント発生場所についても本研究の提案手法を適用することを検討している．小規模なイベント発生場所ではイベント日自体が少なかったり，イベントに関する投稿が少なかったりすることが予想されるが，本研究の提案手法によりある程度緩和できると考えている．

謝　辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（222C02）および，国立研究開発法人情報通信研究機構（NICT）の委託研究（JPJ012368C05401）により得られたものです．

文　献

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: Real-time event detection by social sensors,” in *WWW*, pp. 851–860, 2010.
[2] 塚田涼太郎, 詹浩森, 石渡祥之佑, and 豊田正史, “マイクロブログおよび携帯電話人口統計を用いた大規模イベント会場における人口変化の長期予測,” in *DEIM*, 2020.
[3] R. Tsukada, H. Zhan, S. Ishiwatari, M. Toyoda, K. Umemoto, H. Shang, and K. Zettsu, “Crowd forecasting at venues with microblog posts referring to future events,” in *BSD*, 2020.
[4] 塚田涼太郎, 詹浩森, 石渡祥之佑, 豊田正史, 梅本和俊, 商海川, and 是津耕司, “未来のイベントに言及するマイクロブログ投稿を用いた人口変化の予測,” in *DEIM*, 2021.
[5] Z. Fan, X. Song, R. Shibasaki, and R. Adachi, “CityMomentum: An online approach for crowd behavior prediction at a citywide level,” in *UbiComp*, pp. 559–569, 2015.
[6] T. Konishi, M. Maruyama, K. Tsubouchi, and M. Shimosaka, “CityProphet: City-scale irregularity prediction using transit app logs,” in *UbiComp*, pp. 752–757, 2016.
[7] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu, “Deep sequence learning with auxiliary information for traffic prediction,” in *KDD*, pp. 537–546, 2018.
[8] W. Yamada, D. Torii, H. Kikuchi, H. Inamura, K. Ochiai, and K. Ohta, “Extracting local event information from micro-blogs for trip planning,” in *ICMU*, pp. 7–12, 2015.
[9] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, “Improving traffic prediction with tweet semantics,” in *IJCAI*, pp. 1387–1393, 2013.
[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[11] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, “SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.
[12] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” in *International Conference on Intelligent Computing*, pp. 878–887, 2005.
[13] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.
[14] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on K-Means and SMOTE,” *Information Sciences*, vol. 465, pp. 1–20, 2018.
[15] S. Anno, K. Tsubouchi, and M. Shimosaka, “CityOutlook+: Early crowd dynamics forecast through unbiased regression with importance-based synthetic oversampling,” *IEEE Pervasive Computing*, 2023.
[16] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, “Set Transformer: A framework for attention-based permutation-invariant neural networks,” in *ICML*, pp. 3744–3753, 2019.
[17] F. Rodrigues, I. Markou, and F. C. Pereira, “Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach,” *Information Fusion*, vol. 49, pp. 120–129, 2019.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, vol. 30, pp. 5998–6008, 2017.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, pp. 4171–4186, 2019.
[20] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, “On the trend, detrending, and variability of nonlinear and nonstationary time series,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 38, pp. 14889–14894, 2007.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 10:00 ~ 12:10 | T3-B オンライン（Zoom Events)

検索モデル

座長:義久 智樹(大阪大学)
コメンテータ:黒川 茂莉(株式会社KDDI総合研究所)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T3-B-1-01] リスティング広告閲覧タイミングに基づくリランキングにおける適合フィードバック算出手法の提案

*濱口 滋久[1]、須山 瑠万[1]、栗 達[2]、中島 伸介[1]、河合 由起子[1] (1. 京都産業大学、2. 福岡大学)

10:30 ~ 10:55
[T3-B-1-02] 多様性リランキングにより栄養バランスを考慮したレシピ気分検索システムの評価

*三塚 麟太郎[1]、高久 雅生[2] (1. 筑波大学情報学群知識情報・図書館学類、2. 筑波大学図書館情報メディア系)

10:55 ~ 11:20
[T3-B-1-03] 思い出せない映画に特化した情報検索システムの作成

*吉越 玲士[1]、酒井 哲也[1] (1. 早稲田大学大学院 情報理工・情報通信専攻 酒井研究室)

11:20 ~ 11:45
[T3-B-1-04] ドメイン特徴に基づくランキング学習モデルのドメイン適応

*伊藤 拓誠[1]、丸田 敦貴[3]、加藤 誠[2] (1. 筑波大学 知識情報・図書館学類、2. 筑波大学 図書館情報メディア系、3. 筑波大学大学院 人間総合科学学術院)

11:45 ~ 12:10
[T3-B-1-05] List-wise手法における順位に基づいた多段階分割と学習データの上位遷移による能動学習

*エルゲン 瑛夏[1]、鈴木 優[1] (1. 岐阜大学)

リスティング広告閲覧タイミングに基づくリランキングにおける適合フィードバック算出手法の提案

濱口　滋久† 　須山　瑠万† 　栗　達††
中島　伸介† 　河合 由起子†,†††

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†† 福岡大学　〒 814-0133 福岡県福岡市城南区七隈８－１９－１
††† 大阪大学サイバーメディアセンター　〒 567-0047 大阪府茨木市美穂ケ丘５－１
E-mail: ††{g2054029,i2286220,nakajima,kawai}@cc.kyoto-su.ac.jp, †††lida@fukuoka-u.ac.jp

あらまし　Web 検索結果には検索語に連動したリスティング広告が表示されるが，広告とそれ以外の検索結果を選択閲覧するタイミングは検索語の特性により異なる．我々は先行研究 [1] において，検索語が固有名詞等の場合，検索結果に含まれる広告閲覧のタイミングは検索語が一般的な名詞等の場合より早い傾向であることを明らかにした．これは検索前に，固有名詞等の一意の検索語，ユーザは検索結果に含まれる広告を潜在的に期待していると我々は考えた．すなわち，広告閲覧のタイミングに基づき，広告とそれ以外の検索結果の閲覧履歴の学習利用価値は異なると仮定できる．そこで，本研究では，適合フィードバックに用いる閲覧履歴を広告とそれ以外の検索結果に分類し，リスティング広告を選択閲覧するタイミングに基づき閲覧履歴の重要度を決定する手法を提案する．具体的には，検索結果の閲覧履歴となる rocchio の式の第 3 項に広告を検索結果として追加し，sigmoid の重みを付与し検索結果をリランキングする．本稿では，既存の検索結果と先行研究の検索結果，提案手法によるリランキング結果の正答率を比較検証する．

キーワード　情報検索，リスティング広告，リランキング，適合フィードバック

1 はじめに

検索結果には検索語に連動したリスティング広告が多く表示されている．また，多くの研究で検索結果をクラスタリング・リランキングすることで，ユーザの情報検索を支援する研究が行われている．しかし，リスティング広告に焦点を当てた検索支援手法は提案されていない．そこで，我々はユーザによるリスティング広告の閲覧という行動に注目し，リスティング広告の閲覧をトリガとする検索結果のリランキング手法を提案する．本研究ではユーザと広告掲載主双方にとってリスティング広告が有用となることを目的にとして研究を進めている．

先行研究では検索結果の閲覧履歴をポジティブフィードバック，閲覧履歴を除くユーザに提示された検索結果の履歴・リスティング広告の閲覧履歴をネガティブフィードバックと仮定し，未提示の検索結果をリランキングした [1]. しかし，先行研究によるリランキングは，リスティング広告を閲覧したタイミングによって，ポジティブフィードバックなる可能性を考慮していない．そこで，本研究ではリスティング広告を閲覧するタイミングによって場合分けを行うことで，さらなる精度向上を目指す．具体的な場合は，例えば「iPhone」について詳細に知りたい場合の検索では「iPhone」に関する広告がトップに提示された場合，ユーザはリスティング広告を早い段階で閲覧する．この場合，リスティング広告に興味がありポジティブフィードバックとなる．一方で，長い時間検索を行い何度もリスティング広告が提示されることで閲覧した場合は，興味が少ないと考えられる．

このことから，早いタイミングでリスティング広告を閲覧した場合ポジティブフィードバック，遅いタイミングでリスティング広告を閲覧した場合小さいポジティブフィードバックとする．

本稿では，提案手法と先行研究における手法 [1] との正答率を比較し，提案手法の制度実験を行う．本論文は以下の構成とする．2 章では関連研究を紹介する．3 章ではユーザの広告閲覧を基にしたリランキング手法の概要について説明する．4 章ではリスティング広告閲覧時にリランキングを行う提案手法の有効性検証実験について述べる．最後に 5 章でまとめを記述する．

2 関連研究

2.1 検索結果のリランキング手法

和田ら [2] は，適合フィードバックにおいて不適合文章のみを利用する手法を提案している．実験によって，既存の適合フィードバックよりも早く適合文書を発見可能であることが明らにしている．Zhao ら [3] は，検索結果を取得した HTML を解析，名詞を抽出し X-means 法を利用してクラスタリングを行い，ウェブ検索結果のクラスタリング手法を提案している．実験によって，複数の語義を持つ検索キーワードにおいて提案手法が有効であることを明らかにしている．Khaled [4] らは，検索結果の信憑性や有用性を基準として，検索結果のリランキングする手法を提案している．信憑性と有用性を判定し再ランキングするモデルを作成し，実験により提案手法の有効性を明らかにしている．枝 [5] らはユーザの検索支援を目的に，ユーザ

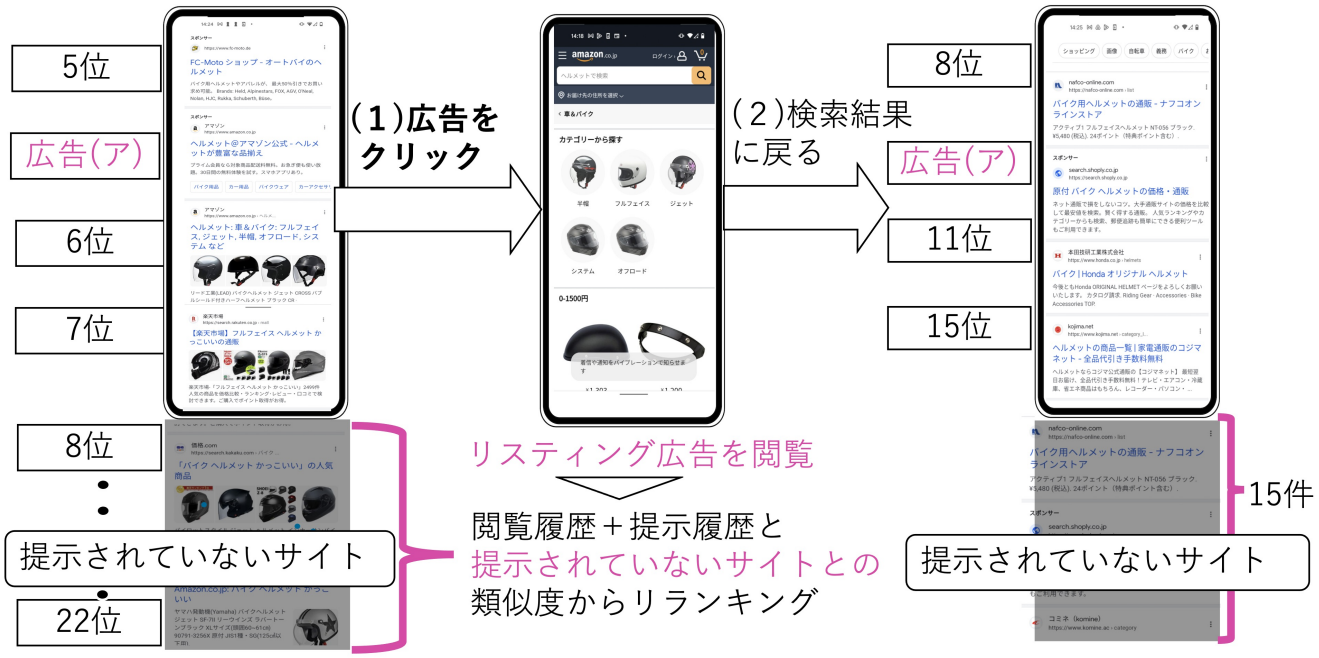


図 1　動的リランキング手法の概要

のページ移動から変わる話題のキーワード遷移を表すグラフを作成し，グラフを基にユーザの興味を推定，Web ページを推薦する手法を提案している．実験によって，提案手法がキーワード遷移グラフがユーザの探索を表すことができており，有効性を明らかにしている．Roul ら [6] は検索を効率化するため，検索結果から意味を考慮したクラスタリングを作成するために，検索キーワードと関係深い話題語を抽出し手法を提案している．実験結果から提案手法により適切な話題語を発見出来ることを証明している．Fernandez ら [7] は情報検索を支援するために，検索キーワードに類似した単語に対してシステムが情報を収集し，閲覧履歴を元にランキングしユーザに提示するシステムを提案している．

2.2　広告推薦に関する研究

Omura ら [8] [9] は，潜在的なユーザの興味に合った広告を推薦するために，ウェブ検索における閲覧履歴の順序から，ユーザの興味分析方式について提案している．実験結果からサイトによって精度が高くなることがわかった．Bleier ら [10] は，ユーザに合った広告を推薦するために，ユーザの状況に合わせて広告を推薦する手法を提案した．コンテンツサービスと広告配信サーブスとで連携し，ユーザが行った行動に対して広告を配信する． この手法により既存の広告配信サービスとは違い，納得性のある広告を推薦することができると説明している．Sobitha ら [11] は，効率よく運用する方法を提案するために，リスティング広告に関するログデータを分析している．分析の結果から，検索キーワードと関係のない URL にも訪れてもらう工夫が必要だと説明している．また, [12] 山口らはこれまで潜在的な購買者に対して広告を推薦するために，特定の web サイトを訪れたユーザの閲覧履歴から特徴を学習し，ユーザに対してその web サイトを推薦する手法を提案した．実験から閲覧履歴の取得期間が長い方が効果的であることを証明している．

3　リスティング広告のクリック閲覧に基づくリランキング

3.1　提示と閲覧と広告閲覧によるリランキング手法の概要

Web サイトが提供するリスティング広告は，特定の検索語に連動して検索結果が提示される．そのため，ユーザが特定の検索語から検索する場合はリスティング広告提示をすることが，ユーザにとっても広告主にとっても効果的な検索結果の提示方法である．一方で，検索語を特定できない検索では多様なリスティング広告が提示され，検索の網羅性が下がる．

そこで，本研究では検索結果提示後に検索語を含むリスティング広告をクリック閲覧する回数が，リスティング広告以外の検索結果をクリック閲覧する回数より少ない場合を一意の検索とし，多い場合を任意の検索として，リスティング広告のクリック閲覧により検索結果をリランキング提示する手法を提案する．

図 1 に，リランキング提示までの流れを示す．まず，ユーザは検索を行い提示された検索結果を探索し，7 位までの検索結果が提示されたと仮定する．次に，ユーザがリスティング広告をクリック閲覧し（図中（1）），リスティング広告のサイトが表示される．この広告閲覧をトリガとし，検索結果のリランキングが開始する．リランキングの対象は，提示されていない検索結果の内上位 n 件までが対象となる．本項で n =15 件となり，提示されていないサイトの 8 位から 15 位までをリランキングする．リランキング対象の検索結果は，閲覧履歴・提示履歴・リスティング広告を利用してリランキングを行う．リランキングでは，閲覧履歴・提示履歴・リスティング広告のタイトルとスニペットからベクトルを作成し，章 3.3 で記述する式を用

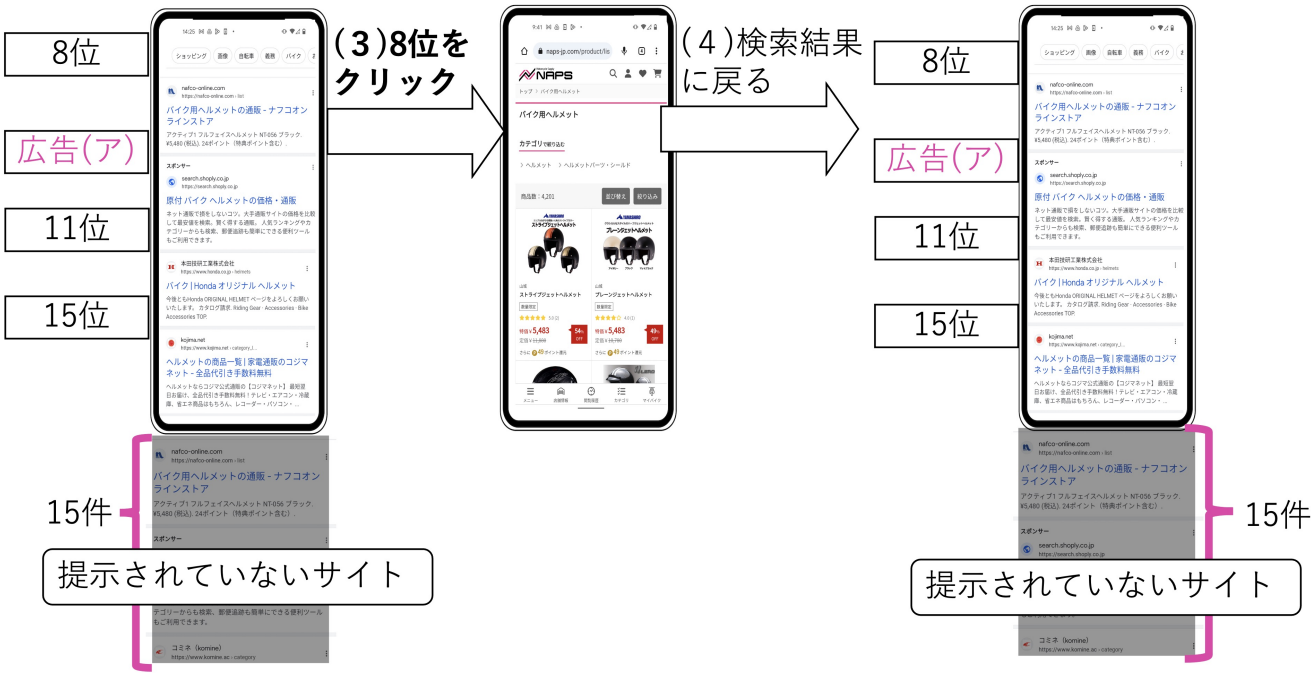


図 2　リスティング広告以外を閲覧した際の流れ

いてリランキングする．また，リランキングではリスティング広告は対象外であるため，広告（ア）はリランキングされない．最後に，ユーザが広告サイト閲覧から検索結果の一覧に戻った際（図中（2））, 検索結果がリランキング提示される．この時，リスティング広告はリランキング対象外であるため，リランキングされず広告（ア）は同じ位置で提示される．

3.2　リスティング広告以外をクリック閲覧した際の流れ

図 2 にリスティング広告以外をクリック閲覧した場合の例を示す．ユーザは図 1 でリランキングされた結果の内，リスティング広告ではない 8 位のサイトをクリック閲覧する（図中（3））．この時，リスティング広告をクリック閲覧していないため，提示されていない検索結果のリランキングは行われない．よって，8 位のサイトを閲覧後に検索結果に戻ると元の検索結果（図 1 でリランキングされた結果）が提示される（図中（4））．

3.3　リランキング方法について

情報検索において検索結果を改善しユーザの検索を支援するシステムとして適合フィードバックが利用される．本研究では，適合フィードバックを利用してリランキング対象の検索結果をリランキングする．提案手法では rocchio の式 [13] を元に，リランキング時に利用する式を作成する．

$$\overline{Q}_l + 1 = \overline{Q}_l + \frac{1}{n_1}\sum_i \overline{R}_i - \frac{1}{n_2}\sum_i \overline{S}_i \qquad (1)$$

上式は本提案手法で利用する rocchio の式である．$\overline{Q}_l$ は l 回目のクエリベクトルであり，$\overline{Q}_{l+1}$ はそれを修正したクエリベクトルである.n_1，n_2 はそれぞれ l 回目の検索における適合文章数と非適合文書数である．$\overline{R}_i$，$\overline{S}_i$ は適合文章と非適合文書に対応するクエリベクトルである．上記式から，修正されるクエリベクトルは適合文章のベクトルはより大きく，非適合文章のベクトルはより小さくなるようにクエリベクトルを修正しユーザの興味をベクトル化している．

本研究では上式を改変してリランキングする式を作成する．式（1）を改変した式が式（2）である．

$$\overline{Q}_l = \overline{Q}_{l-1} + \frac{1}{n_1}\sum_i \overline{R}_i - \frac{1}{n_2}\sum_i \overline{S}_i + \alpha\left(\overline{A}_l\right) \qquad (2)$$

$\overline{Q}_{l-1}$ は l-1 回広告を閲覧した際のユーザの興味を表すベクトルであり，$\overline{Q}_l$ はそれを修正したベクトルである．この $\overline{Q}_l$ とリランキング対象のタイトルとスニペットの類似度を算出しリランキングを行う．n_1，n_2 はそれぞれ 1 回目に広告を閲覧するまでの適合文章数と非適合文書数である．$\overline{R}_i$ は適合文章であり，閲覧済みの検索結果のタイトルとスニペットのベクトルとなる．閲覧済みの検索結果はタイトルとスニペットに興味があったため，ユーザが検索結果を閲覧しているとして適合文章としている．また，$\overline{S}_i$ は非適合文書として提示された検索結果の内，閲覧されていない検索結果のタイトルとスニペットのベクトルとする．提示された検索結果の内，閲覧されていない検索結果はタイトルとスニペットに興味がないため，ユーザが閲覧しなかったとして非適合文章としている．また，式（2）では式（1）にない $\alpha\left(\overline{A}_l\right)$ を追加した．$\overline{A}_l$ は閲覧された広告のタイトルとスニペットのベクトルを表しており，α の値は以下の *sigmoid* 式となる．

$$\alpha = \frac{1}{1 + e^{(k-t)}} \qquad (3)$$

k は広告を選択閲覧した際のそれまでの選択閲覧回数であり，t は任意の値とする．つまり，α の値は k の値が大きくなるにつれて減少するため，例えば 1 回目に広告を選択閲覧した場合は α=1 となり，検索結果から広告を選択閲覧する回数が増加するに従い α は減少し，閾値以降（例えば 5 回目以降）では α=0

図 3　左は手順 1 で提示される検索結果
手順 4 で提示される検索結果結果

となる．1 章で記述したように，ユーザがリスティング広告を早い段階で閲覧した場合興味が強く，長い時間検索を行い何度もリスティング広告が提示されることで閲覧した場合は興味がないことを考慮でき，リスティング広告を閲覧するタイミングを考慮したリランキングが可能となる．

式（4）から求めたユーザの興味を表すベクトル $\overline{Q_l}$ と未提示の検索結果のタイトルとスニペットのベクトルとの類似度を測り，検索結果をリランキングする．

$$\mathrm{score}(j) = \mathrm{sim}(\overline{Q_l}, \overline{Site_j}) \qquad (4)$$

$\overline{Q_l}$ は式（2）を利用して求めたユーザの興味を表すベクトルであり，$\overline{Site_j}$ は未提示の検索結果 j 番目である．本研究では上記の式を利用してリランキングし，ユーザに提示する．

4　リスティング広告がクリックされたタイミングを考慮したリランキングの比較検証

本章では，リスティング広告がクリックされたタイミングを考慮したリランキングの妥当性を検証する．評価方法は，提示した検索結果に対するランキング結果を nDCG より検証する．

4.1　検索語と正解データ

まず，ユーザが検索結果に含まれるリスティング広告を閲覧した後，検索結果がリランキング提示されることの妥当性を調査した．表 3 の検索語を用いて，クラウドソーシングより 101 人を対象にアンケート調査を行った（期間：2023 年 5 月 13 日から 2023 年 5 月 24 日）．アンケート方法を下記に示す．

A：検索結果の一覧　　B：下位と上位を入れ替えた検索結果の一覧

図 4　左は提示される検索結果の一覧 A
右は下位と上位を入れ替えた結果の一覧 B

手順 1：　検索結果（タイトルとスニペットと画像）と広告の 3 件を提示（図 4 の左側の検索結果 A）．

提示される検索結果は上からリスティング広告，検索結果が 2 件の合計 3 件を提示．

手順 2：　提示されたタイトルとスニペットから最も興味がありクリック閲覧したい検索結果（リスティング広告含む）と興味度（5 段階）を質問．

手順 3：　2 番目にクリック閲覧したい検索結果と興味度を質問．

手順 4：　最後に手順 1 と同じ検索結果 A と，未提示の検索結果を入れ替えて提示した検索結果 B のどちらが良いかを質問．

表 1 に，4 つの検索語に対して，2 回目に広告を閲覧したユーザの内，B を選択したユーザの割合を示す．

このうち最も値が低い「スマートフォン」と最も値が高い「英会話教材」を検証の対象とした．スマホケース・英会話教材の 2 つを用いて，クラウドソーシングより 102 人に，リスティ

表 1　検索における異なる検索語に対する検索結果の評価

広告閲覧有無での A または B 選択状況	電子レンジ	英会話教材	ヘルメット	スマホケース
”2 回目に広告を閲覧し，B を選択”	39.29 %	79.55 %	53.33 %	19.05 %

表 2　正解データの例：英会話教材

	検索結果 A	検索結果 B	検索結果 C	検索結果 D	検索結果 E
検索結果（1）	3	5	3	7	12
検索結果（2）	0	0	1	2	12

表 3　検　索　語

検索語
電子レンジ
英会話教材
ヘルメット
スマートフォンケース

表 4　スマホケースにおける nDCG@5 結果

	検索結果（1）	検索結果（2）
リランキング 1	0.589	0.863
リランキング 2	0.688	0.863
リランキング 3	0.615	0.819

表 5　英会話教材における nDCG@5 結果

	検索結果（1）	検索結果（2）
リランキング 1	0.818	0.929
リランキング 2	0.818	0.895
リランキング 3	0.818	0.790

ング広告の選択を基に検索結果をリランキングすることを評価してもらった．ユーザ評価の手順を以下に示す．

手順 1：　ユーザに対して検索語を用いて検索した結果を 3 つ提示（既存の検索結果）．3 左の画面が実際に提示する検索結果．提示される検索結果は，検索結果上位 1，2 位のサイトとリスティング広告を提示．

手順 2：　ユーザに対してタイトルとスニペットからサイトをクリック閲覧したいか質問．

手順 3：　リスティング広告が選ばれた場合，手順 4 に進む．広告以外のサイトをクリックした場合手順 2 を繰り返す．

手順 4：　3 右の検索結果のように，6 つの検索結果を提示し，クリック閲覧したいサイトを選択させる（複数選択可能）．

手順 4 で提示される検索結果は以下の 6 つを提示した．

- 手順 3 までに選択されなかった検索結果または広告
- 検索結果の 3，4 位
- 提案手法よりリランキングした検索結果上位 1，2，3 位

この評価結果から nDCG の正解データを作成した．以下に作成手順を示す．

（1）　アンケート結果から，アンケート手順 3 においてリスティング広告を選択したユーザを抽出

（2）　アンケート手順 2 で検索結果（1）·検索結果（2）を選択したユーザ毎に場合分け

（3）　各場合ごとに，アンケート手順 4 における検索結果 A～E の選択結果を集計

（4）　集計結果を基に検索結果 A～E をランキング

作成結果の例が 2 となる．これは英会話教材に対して作成した正解データとなる．検索結果（1）の場合，検索結果検索結果 E が 1 位，検索結果 D が 2 位，検索結果 B が 3 位，検索結果 A · C が 4 位，5 位となった．また，検索結果（2）の場合，検索結果 E が 1 位，検索結果 D が 2 位，検索結果 C が 3 位，検索結果 A · B が 4 位，5 位となる．この正解データをもとに nDCG@5 を求める．

4.2　nDCG による評価結果

本研究では，提案手法によって提示した検索結果に対するリランキング結果を評価するために，異なるリランキング方法の結果から nDCG を算出し比較した．結果が表 4 · 5 である．リランキング 1 は提案手法・リランキング 2 は α の値を 1 に固定したリランキング方法・リランキング 3 は α を 0 にしたリランキング方法である．スマホケースの検索結果（1）では，リランキング 2 が一番高い値，リランキング 1 は 0.589 で一番低い値となった．また，検索結果（2）ではリランキング 1 · リランキング 2 が 0.863，リランキング 3 が一番低い値となった．英会話教材の検索結果（1）では，リランキング 1 · 2 · 3 すべて同じ値となった．検索結果（2）ではリランキング 1 が 0.929 で最も高く，リランキング 3 が 0.790 で最も低い値になった．スマホケースでは nDCG@5 の値はほとんど変化していないが，英会話教材ではリランキング 1 の提案手法で最も高い値が出た．このことから，広告の閲覧タイミングを考慮することで，リランキング精度が向上することが示唆された．

5　終 わ り に

本研究では，リスティング広告閲覧タイミングを考慮し，ユーザの興味に合わせたリランキング結果を推薦するシステムを提案した．ことなるリランキング方法の結果から nDCG を求め比較検証した．検証結果より，リスティング広告がクリックされたタイミングを考慮してリランキングすることが，ユーザにとって妥当であることが明らかとなった．今後は提案手法を実装し，実際にユーザに使用してもらい既存の検索サイトとの比較検証を実施する予定である．

謝　　辞

本研究の一部は，JSPS 科研費 JP19K12240，JP20H00584 および京都産業大学先端科学技術研究所（ヒューマン・マシン・

データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　献

[1] 濱口滋久, 須山瑠万, 栗達, 中島伸介, 河合由起子, et al. 検索結果におけるリスティング広告閲覧に基づく動的リランキング手法の提案. **研究報告情報基礎とアクセス技術（*IFAT*）**, (32):1 – 6, 2023.

[2] 和田貴映, 永井秀利, and 中村貞吾. 適合フィードバックにおける不適合文書の活用法. In **電気関係学会九州支部連合大会講演論文集 平成 *26* 年度電気・情報関係学会九州支部連合大会 *(*第 *67* 回連合大会*)* 講演論文集**, pages 358–358. 電気・情報関係学会九州支部連合大会委員会, 2014.

[3] Ziliang Zhao, Zhicheng Dou, Jiaxin Mao, and Ji-Rong Wen. Generating clarifying questions with web search results. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 234–244, 2022.

[4] Khaled Yasser, Mucahid Kutlu, and Tamer Elsayed. Reranking web search results for better fact-checking: a preliminary study. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1783–1786, 2018.

[5] 枝隼也 and 佐藤哲司. ユーザの web 探索履歴におけるキーワード遷移に基づく web ページ推薦システム. *DEIM2010 , Mar*, pages A3–3, 2010.

[6] Rajendra Kumar Roul. An effective approach for semantic-based clustering and topic-based ranking of web documents. *International Journal of Data Science and Analytics*, 5:269–284, 2018.

[7] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. Semantically enhanced information retrieval: An ontology-based approach. *Journal of Web Semantics*, 9(4):434–452, 2011.

[8] Takanobu Omura, Felix Dollack, Panote Siriaraya, Da Li, Katsumi Tanaka, Yukiko Kawai, and Shinsuke Nakajima. Verification of a method for latent interest estimation based on user behavior analysis and poi attributes. In *2023 International Conference on Computing, Networking and Communications (ICNC)*, pages 529–535. IEEE, 2023.

[9] 大村貴信, 鈴木健太, 栗達, 河合由起子, and 中島伸介. 実空間のユーザ行動分析に基づく潜在的興味分析方式. **情報処理学会論文誌データベース *(TOD)***, 15(1):1–11, 2022.

[10] Alexander Bleier and Maik Eisenbeiss. Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5):669–688, 2015.

[11] S Sobitha Ahila and KL Shunmuganathan. Role of agent technology in web usage mining: homomorphic encryption based recommendation for e-commerce applications. *Wireless Personal Communications*, 87:499–512, 2016.

[12] 山口由莉子, 森下民平, 稲垣陽一, 中本レン, 張建偉, 青井順一, and 中島 伸介. ユーザの潜在的興味に基づく web 広告推薦方式の検討. In **第 *8* 回データ工学と情報マネジメントに関するフォーラム *(DEIM Forum 2016)***, pages B1–2, 2016.

[13] JJ Rocchio and Gerard Salton. Information search optimization and interactive retrieval techniques. In *Proceedings of the November 30–December 1, 1965, fall joint computer conference, part I*, pages 293–305, 1965.

多様性リランキングにより栄養バランスを考慮したレシピ気分検索システムの評価

三塚麟太郎† 高久 雅生††

† 筑波大学情報学群知識情報・図書館学類 〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s2010048@klis.tsukuba.ac.jp, ††masao@slis.tsukuba.ac.jp

あらまし 様々な観点を考慮して献立を考えることは負担である。そこで本研究では、気分と関連付けたレシピデータをもとに、気分に合ったレシピを検索し、主食・主菜・副菜の割合が偏らないようにランキングを並び替えたシステムによる献立作成支援の手法を提案する。手法の評価として、提案手法と結果をランダムに出力するベースライン手法に対して、2 種類のシナリオの場合とシナリオのない場合で献立を考えるユーザ実験を行った。提案手法は、シナリオなし課題で献立作成の参考になるとの回答が有意に多く、適合性と多様性を考慮した指標 D♯-nDCG が有意に高かった。結果として、気分と合ったレシピを選びつつも自然に栄養バランスの良い献立を作成でき、有用性は高いと考える。シナリオに合わせてパラメータを設定できる仕組みや他の比較手法との実験などが今後の課題である。

キーワード 多様性, 推薦システム, レシピ・料理, 評価・データセット, ユーザ支援

1 はじめに

1.1 食事の必要性と問題点

栄養バランスの取れた健康的な食事が極めて重要である。しかし、食事を考える際には、栄養バランスだけでなく他の観点も考慮する必要があり、それが負担となることがあると考える。マルハニチロが行った調査 [1] によると、週に 1 回以上自宅で夕食を作る 20 歳から 59 歳の女性 1,000 名の約 7 割が「栄養バランスをとる」「調理時間を短くする」「材料費を抑える」といった項目について、日ごろの料理で「改善できたらいいな」と思っていることとして回答している。また、日ごろの料理で実際に工夫していることとして「材料費を抑える」「栄養バランスをとる」「調理時間を短くする」の 3 つの回答が約 4 割半で上位であった。このように、食事を考える際には栄養バランスに加えて様々な観点を考慮する必要がある。

また、マイボイスコムの調査 [2] では、料理についての困りごと・不満として「レパートリーが少ない」の約 4 割に次いで、「メニューがなかなか決まらない」「毎日作らなければならない」こともそれぞれ 3 割弱で上位に含まれた。

このことから、人々の料理に対するニーズは多様であるため、様々な観点を考慮する必要のある料理を考えることが負担となっていると考えられる。

1.2 レシピ検索サイトの活用と限界

料理を新たに考えるという負担を軽減するために、今日では様々なレシピ検索サイトがあり、広く利用されている。例えば、クックパッド [3] や楽天レシピ [4] などの検索サイトを通じて、知っている料理名や食材名を用いて簡単に検索し、料理を作ることができる。

しかし、適切なレシピを探すためには料理名や材料だけでは不十分と考える。なぜならばレシピ選択に関わる「栄養バランスが良い」「気分と合うか」といった要素は、材料や手順、独自のタグやタイトルから自分で判断・検索しなくてはならないからである。岩上ら [5] は、健康を意識したレシピを提供するウェブサイトについて、「生活改善のための献立を決めるには栄養素に関する専門的な知識が必要であり、目的の料理を検索することはユーザーにとって負担が大きい」と述べている。

したがって、「栄養バランスが良い」「気分と合うか」といった人のニーズの合うような適切なレシピを見つけるには、料理や栄養に関する専門的な知識が必要であり、そうしたことは負担となると考えられる。また、前述の通り、「材料費を抑えたい」「好みにあったものを食べたい」などの料理を考えるための複数の観点が、レシピ選択に大きな影響を与える。複数の観点を考慮しながら、適切なレシピを選ぶことは、負担となる場合があると言える。そこで、本研究では複数の観点すなわち料理を決定づける様々な思いを気分として定義する。気分を基にレシピを検索するシステムを作成することで、料理を作る人が負担を感じることなく複数の観点を考慮しながら料理を選択できるのではないかと考える。

1.3 献立と栄養バランスの重要性

一方で、1 つの料理のレシピを選ぶだけでは不十分である。複数の料理を組み合わせてバランスの取れた献立を考える必要もある。文部科学省、厚生労働省、農林水産省 [6] は、「主食・主菜・副菜」という料理の分類を基本とすることにより、多様な食品を組み合わせ、必要な栄養素をバランスよくとることを推奨している。しかし、農林水産省の調査 [7] によれば、「主食・主菜・副菜」を 3 つそろえて食べることが 1 日に 2 回以上あることがほぼ毎日であると回答した人の割合は約 4 割に過ぎず、

半分にも満たないことが明らかになっている。そのため、本研究では、検索結果に表示される複数のレシピの中で「栄養バランス」を考慮するために「主食・主菜・副菜」の割合が偏らないように支援する。すなわち、レシピが複数表示される検索結果から気分に合ったレシピを選びながらも、「主食・主菜・副菜」からそれぞれのレシピを自然と選ぶことができるようにすることで、バランスの良い献立作成を支援する。

1.4 研究の目的

本研究の目的は、複数の観点から気分に合わせて簡単にレシピを検索できるようにし、その検索結果のランキングを栄養バランスを考慮して並び替えるシステムを構築することで、バランスの良い献立作成の支援をすることである。具体的には、気分検索を用いて複数の観点から気分に合ったレシピを推薦するためのランキングを作成する。次に、栄養バランスの観点から多様性リランキングを適用することで、気分と合ったレシピを選びつつ、栄養バランスの観点から多様なレシピの選択肢を提供するシステムを提案する。この手法により、人々の様々な気分に合ったレシピを提供し、自然と栄養バランスの取れた食生活を促すことが期待できる。

2 関 連 研 究

2.1 関連システム

クックパッド [3] や楽天レシピ [4] はレシピ検索サービスである。食材やレシピ名により検索することができる。しかし、第 1 章で述べたように、料理を考える際には複数の観点がある。例えば、「栄養バランスが良い」や「手間が掛からない」、「気分と合う」レシピを探したい場合は、料理名や食材名で検索したのち、レシピのタイトルやレシピの調理手順、レシピ検索サービスが独自に用意するタグ・キーワードを基に、自分で判断する必要がある。

2.2 レシピ検索に関する研究

レシピの様々な特徴を利用することで検索を行うことができる。Yajima ら [8] は、ユーザの条件を考慮して、簡単なレシピを推薦する研究を行っている。調理工程が少ない、レシピに使用する調味料や食材の種類が少ない、追加で用意する調味料が少ない、調理時間が短いことを簡単なレシピであると定義している。ユーザによる小規模な実験を行い、かなり検索結果に満足しているとの結果が得られたが、提案手法が定量的に有用であると確認することは今後の課題であると述べている。

2.3 気分検索に関連する研究

中村と森下 [9] は、気分と料理の関連付けとレシピの気分検索システムの評価を行っている。インターネット上のレシピの中からオーソドックスなレシピ名を 480 件抽出し、それぞれの料理名に対する気分との関係を − 5 点から＋ 5 点の間で得点付けしたデータセットを作成した。そのデータセットをもとにしたプロトタイプの評価実験を通して、気分による献立検索機能の受容性が高いことを明らかにした。さらに、森下と中村 [10] は、中村と森下 [9] が提案した 6 つの検索軸の重要度に偏りがあると述べ、気分による献立検索システムは、入力した気分検索軸のみを気分と料理との一致度の計算の対象にしており、入力した気分検索軸が増えるほど 1 つの検索軸の重要度が減り検索結果の納得度が減る傾向にあることから検索軸を厳選する必要性を指摘した。

2.4 多様性リランキングに関する研究

情報検索における多様性は、関連性を表す指標と多様性を表す指標の線形結合として定義される。Carbonell と Goldstein [11] は関連度と新規性を独立して測定し、線形結合した指標を周辺関連性と呼び、多様性を評価する関数として定義している。この手法によるランキングは、冗長性を最小化することでユーザに有益な情報を提供できると述べている。Kelly と Bridge [12] は、検索候補となるランキングに対して、あるアイテムと既に推薦リスト内にあるアイテムとの平均距離を用いて多様性を定義した。仮のユーザによって評価が行われたが、実際のユーザによる評価は今後の課題とした。

2.5 本研究の位置づけ

本研究では、複数の観点を厳密に定義するのではなく、ユーザの印象によってレシピを検索する。さらに、レシピを検索するために栄養バランスを直接的には考慮せず、検索結果全体で栄養バランスを考慮する。また、ユーザ実験による評価を行うことで、本研究の提案手法の有用性を確かめる。

3 提 案 手 法

3.1 提案手法の全体像

本研究で提案する手法は主に、レシピ検索システム部分とレシピデータ部分の 2 つに分かれる。さらに、レシピ検索システム部分には、気分検索と多様性リランキングの 2 つの処理が含まれる。レシピデータ部分には、元のレシピデータと気分と関連付けるためにアノテーションされたレシピデータが含まれる。

3.2 レシピデータ部分

3.2.1 元となるレシピデータの取得

本研究においては、専門家がオーソドックスなレシピを体系的にまとめている、ぐるなびの「プロから学ぶ簡単家庭料理 シェフご飯」[13] というレシピサイトから料理名を 227 件抽出した。また、料理名と併せて調理時間も抽出した。その際、調理時間の相対的な違いを得るために、標準化を行った。

3.2.2 料理名に対する特徴付け

献立の中での位置づけである主食・主菜・副菜をレシピの特徴として付与する。本研究ではレシピ同士の非類似度に平均距離を用いたことから、複数の特徴を持ったレシピとの類似度を表現できるために、表 1 のように合計が 1 になるように付与する。

表 1　料理の特徴付けの例

料理名	主食	主菜	副菜
唐揚げ	0	1	0
親子丼	0.5	0.5	0
ビビンバ	0.33	0.33	0.33

3.2.3　料理名に対する気分との関連付け

中村と森下 [9] を参考に、気分と料理の関連付けに用いる単語を選択した。本研究においては、気分と関連付けたレシピデータの検討を行い、気分を『精神』『身体』『経済』『時間』からなるものとした。また、それぞれの代表的な気分ワードは表 2 に示す。

表 2　気分と料理の関連付けに用いる項目

項目	気分ワードと点数
精神	お疲れと思う (5 点) ⇔お疲れとは思わない (-5 点)
身体	がっつりと思う (5 点) ⇔がっつりとは思わない (-5 点)
経済	節約とは思わない (5 点) ⇔節約と思う (-5 点)
時間	手軽と思う (5 点) ⇔手軽とは思わない (-5 点)

また、筆者を含めた学生 4 名で 227 件の料理名に対する気分との関係を－ 5 点〜＋ 5 点の 11 段階で得点付けを行った。学生 4 名のうち、料理名を知らない学生がいる場合は、その学生を除いた残りの学生の評価の平均値を用いて、料理名に対してアノテーションを行った。得点付けの際の評価項目は表 2 の内、『精神』『身体』『経済』である。227 件の料理名に対する『時間』との関係は、3.2.1 節で取得した標準化した調理時間を用いて、料理名に対してアノテーションを行った。このように、227 件の料理名に対して、それぞれ『精神』『身体』『経済』『時間』の得点をメタデータとして付与したものを、本研究のシステムで用いるレシピデータとする。

3.3　レシピ検索システム部分

3.3.1　気 分 検 索

a）　ユーザの気分の値を取得

ユーザの気分として取得する評価項目は表 2 と同じ項目を利用する。それぞれ－ 5 点〜＋ 5 点の 11 段階で入力する。

b）　気分検索のランキングを作成

レシピデータに含まれるあるレシピの気分と、ユーザの気分とのユークリッド距離を計算する。そのユークリッド距離を気分検索におけるレシピのスコアとして、スコアの昇順にレシピを並び替えたランキングを、検索結果に出力する候補となるランキング、すなわち候補ランキングとする。あるレシピの気分とユーザの気分とのユークリッド距離を求める式は式 1 に示す。ただし、森下と中村 [10] は、「操作をしていない気分検索軸は気分とスコアの一致度の計算に入れないように制御している. つまり, 「味」のみを操作した場合は, 「味」のみの気分の入力結果と 480 種類の料理のスコアの一致度を計算している」と述べていることから、本研究においてもユーザが入力しなかった、すなわちゼロを入力した観点は計算しない。

あるレシピの気分とユーザの気分とのユークリッド距離

$$= \sqrt{\sum_k (\text{レシピの値}_k - \text{ユーザから取得した値}_k)^2} \quad (1)$$

$$k \in \{ \text{精神, 身体, 経済, 時間} \} \quad (2)$$

3.3.2　多様性リランキングの手法

a）　多様性リランキングの前処理

候補ランキングにおける n 番目のレシピのスコア x_n を Min-Max 法を用いて、式 3 のように正規化する。正規化後の候補ランキングにおけるレシピのスコアを x'_n、候補ランキングにおけるレシピ x の最大値のスコアを $\max(x)$、候補ランキングにおけるレシピ x の最小値のスコアを $\min(x)$ とする。

また、候補ランキングにおけるレシピのスコアを逆数変換したものを $\mathrm{rel}(x)$ とする。候補ランキング中のレシピのスコアが 0 の場合、逆数変換ができないため、スコアが 0 でないレシピの中で最小のスコア $\min'(x)$ を用いて、その半分の値を 0 と置き換える。

$$x'_n = \begin{cases} \frac{x_n - \min(x)}{\max(x) - \min(x)} & (x_n > 0) \\ \frac{\min'(x)}{2} & (x_n = 0) \end{cases} \quad (3)$$

b）　栄養バランスの観点による多様性リランキング

リランキングのアルゴリズムとして貪欲なリランキングを用いる。候補ランキングのリストを C、リランキング後の最終的に検索結果として返すリストを R、検索結果のリスト R の件数 N をとする。貪欲なリランキングアルゴリズム [14] をアルゴリズム 1 に示す。

Algorithm 1：貪欲なリランキングアルゴリズム (リストの連結を表すのに"++"、集合の差を表すのに"\"を用いる。)(筆者翻訳)

Data: N; 候補アイテムの集合、ただし $|C| > N$
Result: 結果のリスト R、ただし $|R| = N$

```
R ← [];
while |R| < N do
    i ← arg max_{i ∈ C} f_obj(i, R);
    R ← R++[i];
    C ← C \ [i];
end
return R;
```

また、多様性を評価したスコアを求める目的関数 $f_{obj}(x, R)$ を、あるアイテム x があったとき、既にランキング中に出ているアイテムとの平均距離と候補ランキングのスコアとの線形結合を求める関数とする。α は関連性と多様性のバランスを表し、α が 0 に近づくほど多様性寄りになる。目的関数 $f_{obj}(i, R)$ を式 4 に示す [14] [12]。

$$f_{obj}(x,R) = \alpha \cdot \mathrm{rel}(x) + (1-\alpha) \cdot \frac{1}{|R|} \sum_{y \in R} \mathrm{dist}(x;y) \quad (4)$$

関数 $\mathrm{dist}(x;y)$ は、3.2.2 節で付与した主食・主菜・副菜をレシピの特徴として、レシピ同士の特徴のユークリッド距離を求める関数とする。レシピ x とレシピ y の特徴のユークリッド距離を求める関数 $\mathrm{dist}(x;y)$ を式 5 に示す。

$$\mathrm{dist}(x;y) = \sqrt{\sum_i (\text{レシピ } \mathrm{x}_i - \text{レシピ } \mathrm{y}_i)^2} \quad (5)$$

$$i \in \{ \text{主食, 主菜, 副菜} \} \quad (6)$$

4 ユーザ実験

レシピを気分によって検索し、その検索結果を栄養バランスを考慮して並び替えたランキングの有用性を評価するために、学生 24 名を対象とした実験参加者による本研究の提案手法とベースライン手法の比較実験を行う。

4.1 ユーザ実験の流れ

実験前アンケートで、実験参加者の基本情報と、料理経験やレシピ検索サービスの利用状況、日ごろの料理の現状などについて回答してもらう。

ユーザ実験はシナリオなどの条件のもとでレシピ検索を行い、献立を考えるという課題を 4 回実施する。各課題は 1 回 5 分の時間制限を設けた。1 回目、2 回目は、実験参加者にシナリオを提示し、シナリオの主人公になりきってレシピ検索を行ってもらう。検索結果から、シナリオの主人公の状況を想定して、料理名を選び、献立を考えてもらう。各課題で選んだ料理名などをシナリオあり課題後アンケートに記入する。3 回目、4 回目は、実験参加者自身の昨日の気分に基づいて、レシピ検索を行ってもらう。検索結果から、実験参加者自身の昨日の夕食を想定して、料理名を選び、献立を考えてもらう。各課題で、選んだ料理名などをシナリオなし課題後アンケートに記入する。

4 回目の課題終了後に、気分検索で用いた検索観点の評価、実験全体を通してのコメントを実験後アンケートに記入してもらう。

4.2 ユーザ実験に用いるシナリオとシステム

システムに慣れてもらうために、1 回目、2 回目は、具体的な状況や値などが設定されたシナリオがある課題を行う。その後の 3 回目、4 回目に、シナリオのない実験参加者自身の昨日の気分という状況を想定して、レシピ検索を行ってもらう。順序効果による影響を無くすため、1 回目、2 回目および、3 回目、4 回目に用いるシナリオとシステムの組み合わせは実験参加者ごとに変えた。

シナリオは 1、2 があり、シナリオなしを便宜上、3 と表記する。シナリオは「Automatic Recipe Metadata Generating Method by Considering Users ’ Various Moods」[15] を参考にした。シナリオ 1 とシナリオ 2 の詳細は表 3、表 4 に示す。

表 3　シナリオ 1

年齢	33 歳
性別	女性
立場	主婦
内容	子どもの誕生日パーティーにあったレシピを探している。
値	お疲れ：-5 点、がっつり：5 点、節約：-3 点、手軽：-3 点

表 4　シナリオ 2

年齢	22 歳
性別	女性
立場	大学生
内容	授業を終えた後に、夕食を作りたいと思っている。
値	お疲れ：-2 点、がっつり：-2 点、節約：0 点、手軽：1 点

また、システムは提案手法（A）と検索結果をランダムに返すベースライン手法（B）とし、UI は同一としたものを用いた。提案手法について、事前評価でシナリオによって適切なパラメーターはそれぞれ異なることが分かった。そこで、ユーザ実験ではパラメーター α=0.5 を用いた。また、検索件数 20 件 >16 件 >10 件の順に評価指標の D-nDCG が高いことが分かったため、ユーザ実験における検索結果は 20 件を提示した。

4.2.1 評価方法

結果は実験前アンケート、シナリオあり課題後アンケート、シナリオなし課題後アンケート、実験後アンケートの 4 つのアンケートの回答結果を基に示す。また、各課題後アンケートの回答結果から献立品数を算出するとともに、評価指標の D♯-nDCG を求める。

課題後アンケートに記入してもらった料理名を、実験参加者にとって適合するレシピとして D♯-nDCG を算出した。酒井 [16] によると、D♯-nDCG は式 7 の通り D-nDCG と I-rec の線形結合により定義される（γ=0.5 とする）。式 7 の @l は上位 l 位までの D♯-nDCG であり、本研究は $l = 20$ である。

$$\text{D♯-nDCG@}l = \gamma \text{I-rec@}l + (1-\gamma)\text{D-nDCG@}l \quad (7)$$

また、式 7 において D-nDCG は、式 8 のように定義される [16]。$\mathrm{GG}^*(r)$ はグローバル利得により適合文書をソートした理想的リストの第 r 位の適合文書のグローバル利得で、$\mathrm{GG}(r)$ は評価対象の第 r 位の適合文書のグローバル利得である [16]。第 r 位文書 d の検索意図 i に適合する確率 $\Pr(rel = 1|i,d)$ が文書 d の意図 i に対する利得 $g_{i,d}$ に比例すると仮定したとき、グローバル利得は式 9 のように表わされる [16]。

$$\text{D-nDCG@}l = \frac{\text{D-DCG@}l}{\text{D-DCG}^*\text{@}l} = \frac{\sum_{r=1}^{l} \mathrm{GG}(r)/\log(r+1)}{\sum_{r=1}^{l} \mathrm{GG}^*(r)/\log(r+1)} \quad (8)$$

$$\sum_i \Pr(i|q) g_{i,d} \quad (9)$$

また、式 7 において I-rec は、検索結果がどれくらい多くの意図をカバーできているかを示す意図再現率であり、式 10 にように定義される [16]。酒井 [16] によると、式 10 における $\mathrm{isnew}_i(r)$ は、検索結果のランク r と各検索意図 i について、$\mathrm{I}_i(r) = 1$ かつ $\mathrm{C}_i(r-1) = \sum_{k=1}^{r-1} \mathrm{I}_i(k) = 0$ のとき $\mathrm{isnew}_i(r) = 1$、それ以外のとき $\mathrm{isnew}_i(r) = 0$ という意味である。

$$\text{I-rec@}l = \frac{\sum_i \sum_{r=1}^{l} \mathrm{isnew}_i(r)}{\sum_i} = \frac{\sum_{r=1}^{l} \sum_i \mathrm{isnew}_i(r)}{\sum_i} \tag{10}$$

4.3 ユーザ実験の結果

4.3.1 課題から得られた結果

課題から得られた結果を以下に示す。回答形式が択一回答の場合は、選択肢のうち評価の高い順に 4、3、2、1 のようにスコアを付与し、システムごとシナリオごとの箱ひげ図で示す。

a） 実験参加者とシナリオの親和性

図 1 は「実験参加者とシナリオの親和性」の回答を箱ひげ図にしたものである。A1、A2、B1、B2 について被験者間 2 要因分散分析を行った結果、シナリオ（$F(1,44) = 21.368, p = 0.000, partial\eta^2 = 0.327$）に有意差が得られた。システム（$F(1,44) = 0.126, p = 0.724, partial\eta^2 = 0.003$）、および交互作用（$F(1,44) = 0.126, p = 0.724, partial\eta^2 = 0.003$）に有意差は得られなかった。

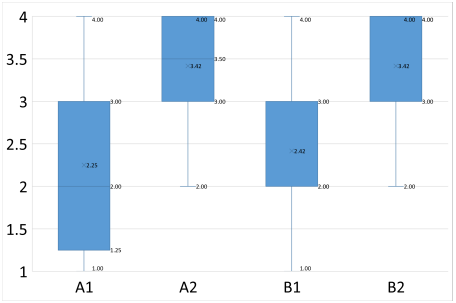


図 1　実験参加者とシナリオの親和性

b） 検索結果から主食・主菜・副菜をバランスよく選べたか

図 2 は「検索結果から主食・主菜・副菜をバランスよく選べたか」の回答を箱ひげ図にしたものである。A1、A2、B1、B2 について被験者間 2 要因分散分析を行った結果、シナリオ（$F(1,44) = 7.308, p = 0.010, partial\eta^2 = 0.142$）、および交互作用（$F(1,44) = 5.690, p = 0.021, partial\eta^2 = 0.115$）に有意差が得られた。システム（$F(1,44) = 0.632, p = 0.431, partial\eta^2 = 0.014$）に有意差は得られなかった。Bonferroni を用いた多重比較の結果、システム A の場合において、シナリオ 1 と 2 の間に有意差が得られた（$2 > 1, p = 0.030$）。また、シナリオ 1 の場合において、システム A と B の間に有意差が得られた（$B > A, p = 0.009$）。また、A3、B3 について対応のある t 検定を行った結果、システム間に有意な差は得られなかった（$t(23) = 0.972, p = 0.341$）。

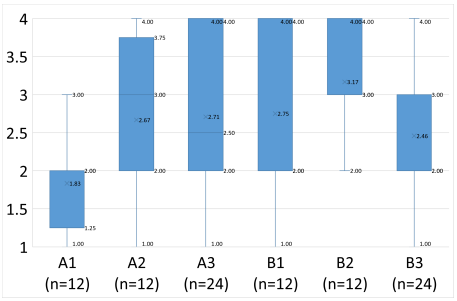


図 2　検索結果から主食・主菜・副菜をバランスよく選べたか

c） 検索結果は献立を考える時の参考になるか

図 3 は「検索結果は献立を考える時の参考になるか」の回答を箱ひげ図にしたものである。A1、A2、B1、B2 について被験者間 2 要因分散分析を行った結果、シナリオ（$F(1,44) = 4.622, p = 0.037, partial\eta^2 = 0.095$）、および交互作用（$F(1,44) = 6.655, p = 0.013, partial\eta^2 = 0.131$）に有意差が得られた。システム（$F(1,44) = 1.664, p = 0.204, partial\eta^2 = 0.036$）に有意差は得られなかった。Bonferroni を用いた多重比較の結果、システム A の場合において、シナリオ 1 と 2 の間に有意差が得られた（$1 > 2, p = 0.002$）。シナリオ 2 の場合において、システム A と B の間に有意差が得られた（$B > A, p = 0.009$）。また、A3、B3 について対応のある t 検定を行った結果、システム間に有意差が得られた（$t(23) = 2.099, p = 0.047$）。

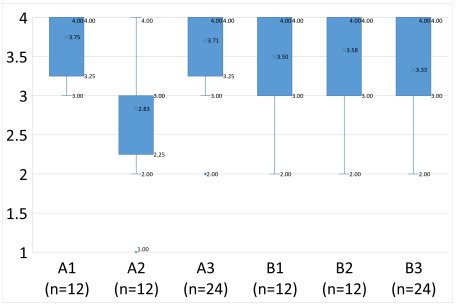


図 3　検索結果は献立を考える時の参考になるか

d） 検索結果の満足度

図 4 は「検索結果の満足度」の回答を箱ひげ図にしたものである。A1、A2、B1、B2 について被験者間 2 要因分散分析を行った結果、システム（$F(1,44) = 0.173, p = 0.679, partial\eta^2 = 0.004$）、シナリオ（$F(1,44) = 0.693, p = 0.410, partial\eta^2 = 0.016$）、および交互作用（$F(1,44) = 0.693, p = 0.410, partial\eta^2 = 0.016$）のいずれにおいても有意差は得られなかった。また、A3、B3 について対応のある t 検定を行った結果、システム間に有意な差は得られなかった（$t(23) = 0.894, p = 0.380$）。

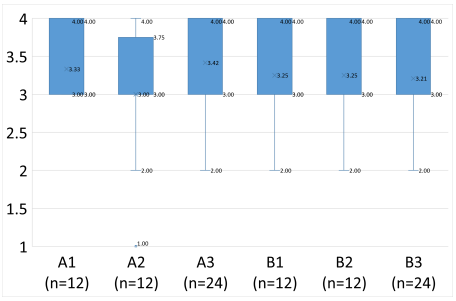


図 4　検索結果の満足度

e）　D♯-nDCG

図 5 は D♯-nDCG を箱ひげ図にしたものである。A1、A2、B1、B2 について被験者間 2 要因分散分析を行った結果、システム（$F(1,44) = 0.992, p = 0.325, partial\eta^2 = 0.022$）、シナリオ（$F(1,44) = 0.105, p = 0.747, partial\eta^2 = 0.002$）、および交互作用（$F(1,44) = 1.238, p = 0.272, partial\eta^2 = 0.027$）のいずれにおいても有意差は得られなかった。また、A3、B3 について対応のある t 検定を行った結果、システム間に有意差が得られた（$t(23) = 2.524, p = 0.019$）。

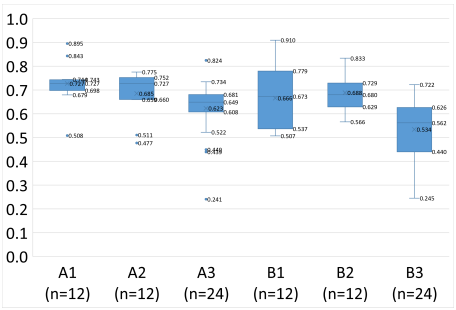


図 5　D♯-nDCG

4.4　ユーザ実験の考察

4.4.1　実験参加者の傾向（実験前アンケートより）

料理の頻度では、「ほとんどしない」または、「週に 1~2 回」料理をする人が半分を占めている。料理が好きかどうかについては、半分の人が「あまりそう思わない」と回答している。さらに、料理が得意かどうかについては、4 分の 3 が「そう思わない」「あまりそう思わない」を合わせた料理を得意と思っていないグループに属することが分かった。実験参加者の半数が料理が好きかどうかについて「あまりそう思わない」と回答したこと、4 分の 3 が料理を得意と思わないことから、実験参加者の半数程度は料理することに対して消極的であると言える。

メニューを決める時に重視することでは、「手間がかからない」ことが最も重視され、次いで「短時間で作れる」「自分や家族の好みにあう」「費用がかからない」の順に重視されている。また、「手間がかからない」「短時間で作れる」ことをまとめて、手軽に作れることを重視していると言える。東京広告協会 [17] が大学生 816 名に対して行った調査によると、金銭的余裕が「ない（計）」と回答した人が約 6 割、時間的余裕が「ない（計）」と回答した人が約 7 割であったと述べている。時間や節約を重視することは、大学生の実態に沿った結果であると考える。

厚生労働省 [18] は「バランスの良い食事を摂っている者の増加」という目標を測る指標に「主食・主菜・副菜を組み合わせた食事が 1 日 2 回以上の日がほぼ毎日の者の割合」と定義している。しかし、バランスの良い食事のために主食・主菜・副菜について意識している人は少ないと推測される。実際に「栄養のバランスがとれている」ことを重視していると回答した人は 8 名に対して、「主菜、副菜、汁物など、複数のメニューを用意する」ことを重視すると回答した人は 3 名であり、重視度に違いがあることが分かった。さらに、そもそも「栄養のバランスがとれている」、「主菜、副菜、汁物など、複数のメニューを用意する」ことを重視している人は少なかった。

4.4.2　課題における考察

a）　実験参加者とシナリオの親和性

図 1 に示される通り、「実験参加者とシナリオの親和性」についてシナリオ間に有意差が得られた。シナリオ 1 よりもシナリオ 2 の親和性が高いと言える。これは、ユーザ実験に用いたシナリオの内容が影響していると考えられる。表 3 のシナリオ 1 の内容は、主婦が子どもの誕生日パーティに合ったレシピを検索するものであった。また、表 4 のシナリオ 2 の内容は、大学生で授業を終えた後の夕食の献立を考えるものであった。実験参加者が大学生であることを踏まえると、シナリオ 2 の方が実験参加者との親和性が高かったものと考えられる。

b）　検索結果から主食・主菜・副菜をバランスよく選べたか

次に、図 2 に示される通り、「検索結果から主食・主菜・副菜をバランスよく選べたか」についてシナリオ間および交互作用に有意差が得られた。システム A の場合において、シナリオ 1 よりもシナリオ 2 でバランスよく選べたと回答した人が多い。図 2 では、有意な差があるとは言えないが、システム B の場合においても、シナリオ 1 よりもシナリオ 2 の平均が高い。これは、シナリオの内容が影響していると考えられる。回答理由を聞いた自由記述からシナリオ 1 の内容がパーティを想定したものであったため、栄養バランスを考慮しなかったと回答が得られた。一方で、シナリオ 2 では、手軽さを重視してバランスが悪くなったと回答した人がいたが、夕食として複数品選ぶことでバランスが取れたという回答も見られた。これらから、パーティを想定したシナリオ 1 の内容により、シナリオ 1 はバランスよく選ぶことを重視しなかったため、シナリオ 1 よりもシナリオ 2 でバランスよく選べたと回答した人が多くなったと考えられる。また、システム A の場合に強い影響を受けている理由は、シナリオ 2 とシステム A との組み合わせと比較して、栄養バランスを重視しないシナリオ 1 と気分検索の結果を踏まえつつも栄養バランスを考慮したランキングの並べ替えを行っているシステム A の組み合わせが適切でなかったためと考えられる。また、シナリオ 1 の場合において、システム A よりもシステム B でバランスよく選べたと回答した人が多い。有意な差があるとは言えないが、図 2 では、シナリオ 2 の場合においても、システム A よりもシステム B の平均が高い。システム A の場合に強い影響を受けている理由も、栄養バランスを重視し

ないシナリオ 1 とシステム A の組み合わせが適切でなかったためと考えられる。さらに、実験参加者の偏りによるものであった可能性も考えられる。シナリオ 1 の場合かつシステム A の条件で課題を行った実験参加者の回答理由には、そもそもバランスを意識することがなかったという回答が得られた。4.4.1 節で述べた通り、料理に対して消極的な実験参加者が半数を占めている。そうした実験参加者は、パーティを想定したシナリオ 1 という条件において栄養バランスまで考慮しなかった可能性が考えられる。システム B の条件で課題を行った実験参加者の回答理由にもバランスを意識していなかったという回答が見られた一方で、栄養バランスを意識して献立を作成したという回答も見られた。ユーザ実験では、24 名の実験参加者にはシステム A・シナリオ 1、システム A・シナリオ 2、システム B・シナリオ 1、システム B・シナリオ 1 の 4 種類の組み合わせのうち 2 種類の組み合わせで課題を実施したため、実験計画上 1 つの組み合わせあたり 12 名を割り当てている。そうしたサンプルサイズの影響により実験参加者の偏りが生じたことも、シナリオ 1 の場合にシステム B でより多くの実験参加者がバランス良く選べたと回答した理由の 1 つであると考える。

シナリオなし課題ではシステム間に有意な差が得られなかった。これは、システムが主食・主菜・副菜を提案しているかどうかに関わらず、実験参加者が献立作成の際にバランスよく選ぶことを意識しなかった可能性が考えられる。システム A の条件の課題に対する回答理由には、疲れているために栄養バランスを意識しなかった、バランスよりも手軽さを重視した、食べたいものを選んだという回答が得られた。一方で、システム B の条件の課題に対する場合の回答理由にも同様の内容が得られた。こうした、料理に消極的な傾向のある実験参加者における献立作成のためのレシピ検索という場面では、栄養バランスよりも手軽さや好みを重視していることが分かる。結果として、「検索結果から主食・主菜・副菜をバランスよく選べたか」という主観的な質問に対しては、システム A とシステム B の間に有意差が得られなかったと考えられる。

c） 検索結果は献立を考える時の参考になるか

また、図 3 に示される通り、「検索結果は献立を考える時の参考になるか」についてシナリオ間および交互作用に有意差が得られた。システム A の場合において、シナリオ 1 よりもシナリオ 2 で参考になると回答した人が少ない。加えて、シナリオ 2 の場合において、システム A よりもシステム B で参考になると回答した人が多い。これは、シナリオ 2 の大学生で授業を終えた後の夕食という条件を踏まえると、手軽さを表す時間の値が－ 1 では検索結果の絞り込みが十分ではないため、上位のレシピでも授業後に求める手軽さを満たさないレシピが表示されていたことが理由であると考えられる。実際に、実験参加者らも手軽度の点で絞り込みが不十分であったことを理由に挙げていた。大学生で授業を終えた後の夕食という状況という手軽さを重視しているシナリオ 2 の場合には、システム A において手軽度の点で絞り込みが不十分であったため、参考になるかという点でネガティブな影響を与えたと考えられる。

シナリオなし課題ではシステム間に有意差が得られた。システム B に比べてシステム A で参考になると回答した人が多い。これは、システムが検索において手軽さや気分を考慮するかどうかが影響していると考えられる。システム A の条件の課題に対する回答には、気分に合ったレシピがあったという回答が見られた。しかし、システム B の条件の課題に対する場合の回答理由には、気分に合ったレシピがあったという回答に加えて、気分に合った献立が少ない、手軽さの点で不十分だったという回答が得られた。これは、実験参加者らはメニューを決める時に手軽に作れることや「自分や家族の好みにあう」ことを重視している傾向があったため、検索において手軽さや気分を考慮していないシステム B と比較してシステム A を用いた実験参加者にとってはシステム A の場合でより参考になると感じやすかったと考えられる。また、システム A の条件で課題を行った実験参加者の回答理由には、システム B の場合と比較して実験参加者が知らない料理名であっても献立に盛り込むことに好意的な回答をする人が多かった。これは、料理の対する興味が広範にあることにより、実験参加者にとってレシピデータ中に適合となりやすいレシピが多くあったためと推測される。

d） 検索結果の満足度

また、図 4 に示される通り、「検索結果の満足度」についてシナリオ間、システム間および交互作用に有意差があるとは言えなかった。シナリオ 1 やシナリオ 2 の条件の下では、ランダムに結果を出力したシステム B でも実験参加者にとって条件に沿っていたと感じるメニューがあったためと考えられる。さらに、気分検索の結果を踏まえたシステム A において、調理時間のみでは手軽さを表現するには不十分であったことが影響したと考えられる。システム A の条件で課題を行った実験参加者の回答理由には、シナリオの状況にあったという回答が得られた一方で、手軽さの面で不十分であったという回答も得られた。システム B の条件下で課題を行った実験参加者の回答理由は、状況にあっていないという回答に加えて状況にあっているという回答も混在していた。また、図 4 によると、いずれの条件下でも満足度が高い。これは、実際に選ぶ献立は数品程度に対して、いずれのシナリオにおいても適合と見なしたレシピが数多くあったと考えられる。すなわち、実験参加者は検索観点のうち少なくとも 1 つが適合になっているレシピを、一部の検索観点が満たしていなくても状況にあっていると直観的に判断した可能性が考えられる。画面録画から計測したところ献立作成時間の平均はいずれの条件下でも 100 秒程度で、献立を考える時間として想定した 5 分間よりも大幅に下回る時間であった。さらに、実験参加者の傾向として料理に消極的であったことを踏まえると、どのようなランキングであっても状況に合っていたと感じやすくなったと考えられる。

シナリオなし課題ではシステム B と比べてシステム A の平均が高いものの、システム間に有意な差が得られなかった。これは、調理時間のみでは手軽さを表現するには不十分であったことが影響したと考えられる。例えば、調理器具の用意や後片付けの手間、調味料や材料の用意の手間といった観点は、調理時間のみでは表現することは難しい。さらに、実験参加者がどのようなランキングであっても状況に合っていたと感じやすく

なった可能性も影響していると考えられる。システム B の条件の課題に対する場合の回答理由では、気分に合ったレシピがあったという内容に加えて、気分に合った献立が少ない、手軽さや節約の面で不十分であったという内容もあり、システム A と比べて好意的でないとする内容が多かった。しかし、システム A の条件の課題に対する回答理由には、気分に合ったレシピがあったという内容も得られたが、手軽さの面で不十分であったという内容も得られた。このように、システム A において調理時間のみでは手軽さを表現するには不十分であったことがシステム間で有意差が得られなかった理由として考えられる。また、システム B よりもシステム A の平均が高いにも関わらず、システム間で有意差が得られなかった理由として、システム B の場合でも気分に合ったレシピがあったことが考えられる。これは、シナリオあり課題後の 4.3.1、d) 節でシステム間、シナリオ間および交互作用において満足度に有意差が得られなかった理由と同様に、一部の検索観点が満たしていなくても状況にあっていると直観的に判断した可能性があり、どのようなランキングであっても状況に合っていたと感じやすくなったと考えられる。

e) D♯-nDCG

また、図 5 に示される通り、D♯-nDCG についてシナリオ間、システム間および交互作用に有意差があるとは言えなかった。一部の検索観点が満たされていなくても状況に合っていると感じやすくなり、どのランキングでも適切だと直感的に判断される可能性があったため、DCG というランクに基づく減損を加味した検索評価指標の D♯-nDCG において有意差が得られなかったと考える。また、実験参加者は 24 名であるが、実験計画上は 1 つの組み合わせあたり 12 名であったというサンプルサイズの影響も有意差が得られなかった理由として考えられる。

一方で、シナリオなし課題ではシステム B に比べてシステム A の D♯-nDCG が有意に高かった。I-rec とランクや多様性など総合的な適合性を加味した D-nDCG を線形結合した D♯-nDCG が、システム A の場合で高い値であった。本研究の提案手法は検索結果の適合性と多様性のバランスを保ちつつも、主食・主菜・副菜をそれぞれ満たすような献立作成を支援できることが示唆されていると考える。4.4.3 節の実験後アンケートにおいて、シナリオの役になりきることが難しいという感想が寄せられた。そのため、シナリオなし課題ではランキング中のレシピが適合かそうでないかを判断しやすかったことが、D♯-nDCG に有意な差が得られた理由と考えられる。

4.4.3 実験後アンケート

実験全体を通しての感想やコメントでは、本研究で検討した検索観点による検索が有用であるとした回答が得られた。また、検索システムとしての使いやすさや利便性、面白さ、料理を作る際のモチベーションといった面でポジティブな内容が得られた。さらには、電化製品の開発や食品販売などの料理以外の分野でも活かされそうとの意見も得られた。一方で、本研究で考慮していなかったジャンルの組み合わせの難しさや、シナリオの役になりきることが難しいという内容が得られた。また、本研究で用いたレシピデータに含まれるレシピで知らないものが多かったとする内容が見られた。そのため、選ぶために写真が必要との意見が得られた。

実験全体を通しての感想やコメントから、本研究で検討した検索観点による検索が有用性は高いと考える。また、実験参加者にとって知らない料理が多いことや、シナリオの役になりきることが難しいといったことは、4.4.1 節で料理に消極的な傾向と言える実験参加者が多かったことと関連があると考えられる。これらのことが、シナリオあり課題やシナリオなし課題で得られた結果の一部で、有意な差があると言えないことに留まった理由と考えられる。

5 おわりに

本研究の提案手法と結果をランダムに出力するベースライン手法を用いて、2 種類のシナリオの場合とシナリオのない場合で献立を考えるユーザ実験を行った。ユーザ実験の結果、提案手法はシナリオなし課題で「検索結果は献立を考える時の参考になるか」という質問に対して参考になるとの回答が有意に多かった。さらに、多様性と適合性を考慮した指標 D♯-nDCG が提案手法の場合で有意に高かった。このことから、レシピ検索システムの実際の利用シーンに近いと考えられる昨日の気分で昨日の夕食を考えるという条件で、本研究の提案手法の有用性は高いと言える。すなわち、本研究の提案手法は検索結果の適合性と多様性のバランスを保ちつつも、主食・主菜・副菜をそれぞれ満たすような栄養バランスの良い献立作成の支援ができると示唆されていると考える。

また、ユーザ実験において、子どもの誕生日パーティーや授業後の夕食（シナリオあり）では D♯-nDCG に有意な差があるとは言えなかったが、昨日の気分で昨日の夕食を考えるという条件（シナリオなし）では有意な差が得られたことから、本研究の提案手法の効果が表れやすい状況と表れにくい状況があると言える。

今後の課題として、ベースライン手法や実験参加者の属性を変えた場合の実験や、状況に合わせて適切に多様性と関連性を変化させる仕組み、レシピデータ数の増加、検索観点の改善などを行っていきたい。

謝　辞

本研究は JSPS 科研費 JP20K12543, JP23K11761 の助成を受けたものです。

文　献

[1] マルハニチロ株式会社. ”料理の工夫と時短調理に関する調査”. MARUHA NICHIRO. https://www.maruha-nichiro.co.jp/corporate/news_center/research/pdf/20150903_ryouri_kufuu_jitan_cyousa_2.pdf, (参照 2023-11-18).

[2] マイボイスコム株式会社. ”【 料理 】に関するアンケート調査（第 5 回）”. マイボイスコム株式会社. https://www.myvoice.co.jp/biz/surveys/27804/index.html, (参照 2023-11-18).

[3] クックパッド株式会社. ”cookpad”. cookpad. https://cookpad.com/, (参照 2023-11-17).

[4] 楽天グループ株式会社. ”rakuten レシピ”. 楽天レシピ.

https://recipe.rakuten.co.jp/, (参照 2023-11-18).
[5] 岩上将史, 安藤哲志, 伊藤孝行, 田中雅章. 栄養情報を活用した目的指向料理推薦システムの試作. 情報処理学会第 73 回全国大会講演論文集, pp. 581–582, 2011.
[6] 文部科学省, 厚生労働省, 農林水産省. "食生活指針の解説要領". 農林水産省. https://www.maff.go.jp/j/syokuiku/attach/pdf/shishinn-7.pdf, (参照 2023-11-17).
[7] 農林水産省. "3 調査結果の詳細 （2. 現在の食生活について）". 農林水産省. https://www.maff.go.jp/j/syokuiku/ishiki/r05/3-2.html, (参照 2023-11-18).
[8] Yajima Asami and Kobayashi Ichiro. "easy" cooking recipe recommendation considering user's conditions. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3, pp. 13–16, 2009.
[9] 中村富予, 森下幸俊. 気分による献立検索システムのプロトタイプにおける受容性評価. 電子情報通信学会技術研究報告, Vol. 112, No. 75, pp. 73–77, 2012.
[10] 森下幸俊, 中村富予. 気分による献立検索システムの検索軸の評価とレシピを活用した食品販売機能の市場ニーズの評価. 電子情報通信学会技術研究報告, Vol. 112, No. 75, pp. 79–84, 2012.
[11] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.
[12] John Paul Kelly and Derek Bridge. Enhancing the diversity of conversational collaborative recommendations: a comparison. *Artificial Intelligence Review*, Vol. 25, pp. 79–95, 2006.
[13] Gurunavi. "プロから学ぶ簡単家庭料理　シェフご飯". Gurunavi. https://chefgohan.gnavi.co.jp/, (参照 2023-12-10).
[14] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.*, Vol. 7, No. 1, 2016.
[15] Natsuhiko Takata, Mayumi Ueda, Yukitoshi Morishita, and Shinsuke Nakajima. Automatic recipe metadata generating method by considering users' various moods. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1, pp. 413–418, 2016.
[16] 酒井哲也. 情報アクセス評価方法論 -検索エンジンの進歩のために-. コロナ社, 2015.
[17] 公益社団法人東京広告協会. 大学生意識調査プロジェクト future2016「大学生と消費」に関する意識調査. 公益社団法人 東京広告協会. http://www.tokyo-ad.or.jp/activity/seminar/pdf/FUTURE2016.pdf, (参照 2023-12-14).
[18] 厚生労働省. 国民の健康の増進の総合的な推進を図るための基本的な方針. 厚生労働省. https://www.mhlw.go.jp/content/001102474.pdf, (参照 2023-12-26).

思い出せない映画に特化した情報検索システムの作成

吉越　玲士†　　酒井　哲也†

† 早稲田大学大学院 基幹理工学研究科 情報理工・情報通信専攻　〒 169-8555 東京都新宿区大久保 3-4-1
E-mail: †yoshikoshi@akane.waseda.jp, ††tetsuyasakai@acm.org

あらまし　正確な固有名詞や識別子を思い出すことができないシチュエーションの下では，既存の情報検索システムは本来の性能を発揮することができない傾向にある．本研究では，このようなシチュエーションについてよりロバストな情報検索システムの作成を試み，その性能の検証を行った．ベース手法は TF–IDF を用いたものと BERT モデルを導入したものの 2 つを用意し，後者については複数のモデルアーキテクチャを準備し再学習も行った．また，これらのベース手法に対して年代キーを独自に導入し，候補となる作品群のサイズを小さくする補助手法も導入した．その結果，TF–IDF を用いたものが最も高い性能を発揮し，年代キーが全モデルにおいて性能を向上させることを示した．BERT モデルについては再学習時の損失関数に improved triplet loss を，教師データの増強に GPT–4 をそれぞれ使い，アノテーション済みの教師データの数が少ない場合でも大幅に性能を向上させることに成功した．

キーワード　情報検索システム，TOT 現象，Wikipedia コーパス，BERT，TF–IDF

1 はじめに

過去に消費した様々なコンテンツのタイトルやキャラクタ名，内容といったものを自力で部分的にしか思い出すことができず，悩むことがしばしばある．こういった知っている単語や事柄を完全に思い出すことはできないが，似たような意味の単語や事柄の一部のみ思い出すことができる現象を tip of the tongue phenomenon(舌先現象) という [1] [2]．Elsweiler ら [3] によるとこの状態下では，そうでない記憶欠落状態よりもユーザは強いフラストレーションを感じる傾向にある．

今日では，このような状況下で生じるフラストレーションを解消するために，ユーザが質問し，回答を募ることができるプラットフォームが複数登場している．例えば “I Remember This Movie...” [1]では，思い出すことのできない映画作品に特化したプラットフォームになっており，部分的に思い出したあらすじや視聴に至るまでのコンテキスト等を投稿する場になっている．このように映画や書籍が対象の場合は質問テキストを投稿して回答を募る形になるが，そのような形式やフォーマットに収まらず，鼻歌や口笛等の音声データを投稿するような音楽特化の Web サイト [2]も存在するなど，形態は多種多様である．

Arguello ら [4] によると，既存の情報検索システムではこういった正確な名前や識別子を表現できない検索クエリに対して，システム本来の性能を発揮できない傾向にある．そのため検索の試行に失敗した，あるいはそもそも自力での検索を試すことなく先述のようなプラットフォームに質問を投稿するユーザが少なくない．本研究ではこのような検索ニーズをより適した手法を導入することで問題を緩和し，フラストレーションを解消することができるような情報検索システムの作成に取り組み，複数のアプローチを検討した．

1：https://irememberthismovie.com/
2：https://www.watzatsong.com/en

このような曖昧であったり誤りを含みうる情報やコンテキストからの検索技術は，将来的には撮影時のコンテキストから写真を検索したり，受信時の状況や大まかな内容から E メールを検索するなどといったシステムの発展に貢献できることが期待される [3] [5] [6].

2 関連研究

2.1 TOT 現象の分析

Brown ら [1] によって TOT 現象についての体系的な調査が最初になされて以降，60 年近くにわたって様々な事柄が明らかにされてきた．

TOT 現象は若年の成人でも 1 週間に 1 回程度発生し，固有名詞によって誘発されやすいことが明らかになっている [7]．またユーザが高齢であるほど発生する頻度が高くなり，その解決に要する時間も長くなる [8]．今回のような映像作品を対象とした場合には，通常思い出すものは作品のタイトルや登場人物の名前などの固有名詞が中心となり，TOT 現象が発生しやすいドメインであると推測される．

Schwartz [9] によると，TOT 現象は全ての言語と文化の下に暮らす人々の間で普遍的な現象である．英語の “tongue” に相当する単語を使ってこの現象を表現する言語は調査対象とした 51 の言語のうち 45 にもなり，そうでない言語はアイスランド語やインドネシア語などごく一部に限られるものの，表現が異なっても現象自体は存在する．また，現象自体を指す言葉が存在しないグアテマラ周辺で話されるケクチ語 [10] の話者でも，同様の現象を体験していたことが報告されている [11]．よって本研究にて作成する TOT 現象の解消を補助するシステムは，特定の地域や言語に限定せず需要がある．

2.2 実際に投稿された質問テキストの分析

ユーザが正解の映像作品のタイトルを募ることができるプ

ラットフォームの一つ “I Remember This Movie...” に投稿されている質問テキストについて，Arguello ら [4] はどのような情報が含まれているのかをはじめとした多角的な分析を行った．

その結果によると，既存のキーワードマッチングをベースとした検索システムでの活用が困難な「どこで見た」「いつごろ見た」というコンテキストに関する情報が約 10% の質問テキストに含まれていることが明らかになった．また，映画の内容に関する情報のうち視覚的なものは 18%，聴覚的なものは 2% 未満の質問テキストにそれぞれ含まれている．この両者における出現率の大きな差は，視覚的な情報の方が聴覚的な情報と比較して記憶として残りやすいという先行研究 [12] [13] に矛盾しない結果になった．さらに Okapi BM25 [14] を用いた ablation study によると，質問テキストに含まれる「いつごろ公開されたか」「何を使って見たか」といった情報は検索結果に負の影響をもたらすことが明らかにされた．

本研究では検索結果に負の影響を与えるとされた「いつごろ公開されたか」という年代情報を，別途専用の処理を導入することで検索対象のサイズを小さくすることに活用できるのではないかと考え実験を行った．また，確率検索モデル BM25 に代えて言語モデルを導入することで性能をさらに向上させることができるのではないかと考え，BERT を導入したシステムを使って実験を行った．

3 実験手法

3.1 使用したデータ

実験では TREC 2023 Tip-of-the-tongue (TOT) Track [15] が提供している作品群データセットを用い，これをコーパスとした．このデータセットにはカテゴリ “audiovisual works” に直接，あるいは間接的に紐づく 231,618 件の作品の Wikipedia 記事データが含まれており，実験に使用しやすいよう `JSONL` 形式に整形されている．実験時に使用するテキストは Wiki マークアップ等の装飾が除去され，任意のセクションのテキストを取り出すことができる状態になっている `sections` キーのデータを利用することにする．

検索クエリは同様に TREC 2023 TOT Track が提供している train/dev queries を用いる．これらデータは “I Remember This Movie...” に投稿されたデータを整備したもので，MS–TOT [4] に収録されているデータの部分集合となっている．本研究は test queries を利用してシステムが作成されないように定められた TREC TOT Track の制約に従うため，今回 MS–TOT のデータは直接使わないようにする．なお，test queries は現段階で正解データが公開されていないため性能評価に使うことができず，これを dev queries で代用する．train/dev queries にはそれぞれ 150 個ずつ含まれており，各質問テキストに対してそれぞれ正解作品に関する情報が複数付与されている．

以上のデータを利用し検索システムの作成を行う．検索時にはクエリに含まれる `title` と `sentence_annotations` は使わず，基本的にはクエリの `text` のみを使うようにした．本実験ではある質問テキストについて作品群に対して検索を行い，正解作品を 1 位に位置付けることを目標とする．

3.2 TF–IDF

本手法では各作品データのセクションのうち，作品の概要やあらすじに関連する代表的な 3 セクション “abstract”，“synopsis”，“plot” について取り出し，それを元に TF–IDF ベクトライザ [3] の初期化を行った．ここで，ベクトライザの初期化条件として最大文書頻度 `max_df` を 0.75 に設定し，stop word[4] 以外にも，作品の特徴とはならない単語を排除することができるようにした．

しかし調査によって作品記事内のセクション名に大幅な表記揺れがあることが判明した．実際の検索を行うにあたって先ほどの 3 セクションのみを使うだけでは，含めるべき概要やあらすじを関する記述を予期せず排除してしまい，本来検索できたはずの正解作品も検索することが不可能になってしまう可能性がある．よって可能な限り幅広い表記揺れに対応することができるようセクション名を調査し，以下に挙げるセクションに概要やあらすじが含まれていると結論付けた．

- abstract
- overview
- synopsis
- plot
- premise
- plot summary
- story
- plot synopsis
- storyline
- summary
- setting and plot
- plot overview
- plotlines

以上のセクションを「検索対象セクション」と呼ぶことにする．

先ほど初期化したベクトライザを用いてクエリの `text`，作品群の検索対象セクションに該当するテキストについてそれぞれベクトル表現を得て，コサイン類似度を計算した．検索結果はこのコサイン類似度を降順に並べることで生成する．

3.3 事前学習済み BERT

言語モデルを使用した性能のベースラインとして，`bert-base-uncased` [16][5] を導入した．先ほどと同様にクエリの `text`，作品群の検索対象セクションに該当するテキストについてそれぞれ埋め込み表現を計算する．このとき，ベクトル表現は `pooler_output` を用いることにする．本手法ではベクトル同士のユークリッド距離を計算し，昇順に並べることで検索結果を生成する．この手法のことを BERT(Pre) と呼ぶことにする．

3：Python より，`sklearn.feature_extraction.text.TfidfVectorizer` を使用した．

4：“is”，“a”，“of” など，一般に英文中に登場し文書の特徴になりえない単語のこと．

5：`https://huggingface.co/bert-base-uncased`

3.4 BERT モデルの再学習

3.4.1 データの準備

Train queries を利用するだけでは 150 件しか訓練用の教師データが利用できないため，BERT モデルの再学習にあたってはデータ数の少なさが大きな課題となる．そこで GPT–4[6]に対して作品情報を入力し質問テキストを疑似的に生成することでこの問題を緩和できるのではないかと考え，以下のような指示のもと得られた GPT–4 の出力を収集した．なお対象とする作品は，検索対象セクションのテキストについて `bert-base-uncased` のトークナイザを使ったときのトークン数が 1,000 以上 2,000 未満のものに限ってランダムに選択する．これはトークン数が少なすぎると質問テキストを疑似生成するためには情報が少なすぎることが，トークン数が多すぎるとコストが必要以上に高くなることがそれぞれ想定されるためである．

a) 使用モデル: GPT–4 – 0613
b) システムへの役割指示: The user couldn't remember the title or specific details. They are seeking helps to identify the movie by describing the plot even if it is vague. Please provide a single example of possible question to get the correct title from others. The question text must use common nouns instead of proper one and express in different words from original one as well as possible.
c) temperature: 0.7
d) 入力作品情報: トークン数の条件を満たす作品の検索対象セクションのテキスト

上記の指示により GPT–4 の出力を 1,442 件収集した．元より利用可能なデータ train queries と合わせて教師データの総数が約 11 倍となったことにより，モデルがより高い性能を発揮できるようになることが期待できる．

3.4.2 学 習 方 法

モデルはベースラインと同じく `bert-base-uncased` を使用し，再学習を行う．損失関数には triplet loss [17] を設定し，anchor には表 1 に示す (A)，positive には対応する正解作品の (B)，negative には無関係な作品の (B) をそれぞれとるようにする．さらに，anchor と positive は基本的に一対一対応であるが，anchor と negative は一対多対応に拡張することができる．そこで data augmentation として anchor に対応する negative は作品群全体から 100 個ランダムに選択するようにする．

表 1 使用したデータの選択方法

(A) 質問テキスト	(B) 概要・あらすじ文
train queries の質問テキスト	検索対象セクションのテキスト
GPT–4 で生成した疑似質問テキスト	

以上により anchor と対応する作品に関する記述 positive が埋め込み空間上でより近くに配置され，異なる作品に関する記述 negative はより遠くに配置されるように学習が進む．ここでは margin として $\alpha = 10.0$ と設定した．しかし，Wang ら [17] の triplet loss を直接使う場合は anchor と positive の距離 d_p，anchor と negative の距離 d_n が $d_p + \alpha \leqq d_n$ を満たした以降は最適化が進行しなくなる．

そこで，本研究では Cheng ら [18] が提案した improved triplet loss を導入し，$\beta = 1.0$ に設定した．この手法は $[d_p - \beta]_+$ を従来の損失に加えるもので，[7]それにより d_p が最大でも β に収まり，相対的な関係に加えて絶対的な距離も縮められることが期待される．このような手法で行われる学習は deep metric learning [19] と呼ばれ，サンプルサイズが小さい場合や未知のクラスがある場合，クラス数が多い場合にでも比較的高い性能を発揮でき，今回のタスクに適していると考えた．また，この triplet loss は元来は画像検索タスクにおける損失関数であるが，Hoffer ら [20] によって metric learning を必要とする幅広いタスクに有用であることが示されており，今回のようなテキストベースの検索タスクにおいても十分な性能が発揮できることが期待できる．

学習率は初期値の 1.0×10^{-5} から学習が進むにつれて指数関数状に変化するようにスケジューラ [8]を導入し，1 epoch ごとに 0.90 倍になるようにした．学習は最大 20 epoch まで進み，20 epoch 目での学習率はおよそ 1.4×10^{-6} になる．

モデルのアーキテクチャは図 1 に示すような 2 パターンを用意し，比較した．いずれも BERT モデルの最後の隠れ層 `last_hidden_state` のうち，第 1 トークンに相当するベクトルに対して dense と活性化関数 tanh を適用した後のベクトルである `pooler_output` を使用した．それに対して dropout を適用したのち，次元数がより大きくなるよう全結合層を導入した．これは今回の検索対象となる集合サイズが約 23 万と大きいため，表現力を向上させる目的でそれぞれ `pooler_output` の 768 次元に対して 1.5 倍の 1,152 次元，さらにその 1.5 倍の 1,728 次元というようにそれぞれ決定した．図 1 において活性化関数 (activation function) と出力 (output) の間に位置する全結合層 (fully connected layer) が 1 つのものを BERT(1L)，2 つのものを BERT(2L) と呼ぶことにする．

3.5 年代キーの導入

候補となる作品群のサイズを適切な手段を使って小さくすることができれば，検索システムの性能向上に結びつくと考えた．そこで質問テキスト中に含まれる情報のうち，比較的容易に作品との関連度として計算することができる「いつ頃視聴したのか」という情報「年代キー」を導入することにした．

作品側の年代キー「いつ頃リリースされたのか」「いつ頃放送されたのか」の作成は外部の情報リソースを一切使わず，全て Wikipedia の記事内から抽出する．多くの作品記事には infobox が 1 つ以上存在し，それらからルールベースで抽出することで目的を達成することができる．

しかしながら，全ての作品記事について infobox のデータが充実しているとは限らず，そもそも infobox が記事中に存在し

6：`https://openai.com/gpt-4` なお，OpenAI が提供する API で同じく利用可能な GPT–3.5 –turbo は，意図した出力を行わなかった．

7：$[x]_+$ はランプ関数と同様に $x < 0$ のときに 0，そうでないときに x を返す関数．

8：`torch.optim.lr_scheduler.ExponentialLR` で実装することができる．

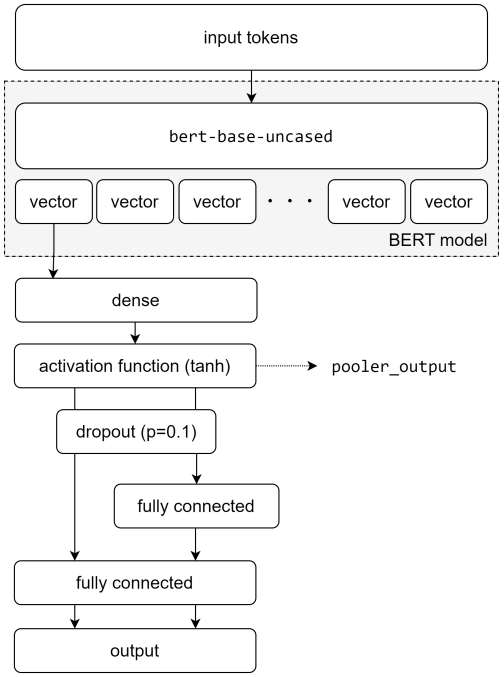


図 1　本実験で再学習を行ったモデルアーキテクチャの概要

なかったり，あるいは infobox が存在しても該当する情報が存在しないこともある．その場合には，Wikipedia 記事のテキスト中から抽出することは避けられないが，本文中には公開・放送日の記述以外にも様々な年代表現が点在することは容易に想像できる．そのため可能な限り的確にリリース日を抽出するために Question Answering モデルを導入することにする．

今回は `bert-base-uncased` を蒸留することで軽量になった派生モデル `distilbert-base-uncased-distilled-squad` [21][9] を選択した．このモデルに対して以下の条件に従って質問を行った．

a）　質問: When is this mentioned movie released? [10]

b）　コンテキスト: Wikipedia 記事テキストの先頭 512 トークン

以上のルールベースと機械学習モデルベースの 2 つの手法を使って抽出すると，複数の年代表現が得られる場合も少なくない．アニメやドラマなどで長期にわたって放送されたものや，制作国外で映像作品が公開される際に年が異なるといった場合もある．こういった場合を想定し，年代キーは始点 S と終点 E のペアで必ず表現することにし，常に $S \leqq E$ が成立するよう候補となる年代キーの最小値を S に，最大値を E とする．なお，以上の手段をもってしても何もデータが得られなかった場合には，便宜的に $S = E = 0$ とする．

クエリとなる質問テキストに対しても同様に Question Answering モデルを用いて年代キーを作成し，検索時に使用することができるようにする．以下のような条件に従ってモデルに対して質問を行った．

a）　質問: When is this mentioned movie watched by the questioner? [10]

b）　コンテキスト: 質問テキスト全文

質問テキスト中に含まれる年代表現は一般に曖昧で，幅のある表現が主となる．そこで，モデルの出力に対してさらに表 2 に示すようなルールで S と E をとるようにする．表 2 に示す正規表現の検索結果に西暦の下二桁のみが含まれる場合には，今回のクエリとなる質問テキストは全て 2013〜2018 年の間に投稿されたものであることを踏まえ，18 年までは 2000 年代，19 年以降は 1900 年代と判定することにする．また，質問テキストから得られた年代キーについて，各前後 M 年間までを許容とする年代許容幅も設定し，M を 0〜10 年の間で変化させて実験を行った．なお，質問テキストから 1 つも年代キーを取得できなかった場合には，この手法は用いない．

表 2　質問テキスト中に含まれる年代表現の抽出ルール

正規表現パターン	適用されるテキスト		
	表現の例	始点 S	終点 E
`([0-9]{4})`	around 1995	1995	1995
`([0-9]{2})[^0-9]{1,3}([0-9]{2})`	1985~90	1985	1990
`([0-9]0)\S?[sS]`	1990's	1990	1999

質問テキストの年代キー (S_q, E_q) と作品テキストの年代キー (S_c, E_c) の離れ具合は，式 1 に従って罰則 P を計算することで比較する．質問テキストと作品の間でわずかでも年代キーに重複する期間がある場合には g の値が 0 となりそのときに $P = 1.0$ となることが期待され，また乖離が大きくなると急峻に P の値が大きくなることがそれぞれ期待されるため，ここでは指数関数 $\exp(x)$ を導入した．

ただし，$\exp(g)$ の形で直接使用すると $P = \exp(1) \simeq 2.7$, $P = \exp(2) \simeq 7.4$ というように質問テキストと正解作品の差がわずか 1 年や 2 年であったとしても大きな罰則 P となってしまう．そのため係数 a $(0 < a \leqq 1)$ を導入し，実験ではあらかじめ設定した 1.0 から 0.01 の範囲で変化させた．

$$
\begin{aligned}
g(S_q, E_q, S_c, E_c) &= \max(0, \max(S_q, S_c) - \min(E_q, E_c)) \\
P(S_q, E_q, S_c, E_c) &= \exp\left(ag(S_q, E_q, S_c, E_c)\right)
\end{aligned}
\tag{1}
$$

この値 P を各手法によって算出されたコサイン類似度とユークリッド距離と併せることで最終的なランク付けにおける基準とする．ただし，コサイン類似度 C を指標とする TF–IDF を用いた手法では $(C+1)/P$ を計算することで $[0, 2]$ の範囲に収まるように，ユークリッド距離 D を指標とする BERT を用いた手法では $D \cdot P$ を計算することで $[0, \infty)$ の範囲に収まるようにした．

4　実験結果

4.1　評価方法

各手法の性能は相対順位と success@k の 2 つを用いて評価を行う．

相対順位は全文書数，今回は作品群 C のサイズ $|C| = 231,852$

9：`https://huggingface.co/distilbert-base-uncased-distilled-squad`
10：文意としては “is” ではなく “was” とするべきだが，精度が前者の方が上回ったためこのようにした．

で正解作品の順位を正規化したもので，システムが正解作品を 1 位に位置付けることができた場合に $1/|C| \simeq 4.31 \times 10^{-6}$ となる．値は $[1/|C|, 1]$ の範囲をとり，小さいほど性能が高い．システムの性能は全ての評価用データについて平均値と中央値をとることで評価する．

Success@k は，正解作品を上位 k 位以内に位置付けることができれば成功として 1，そうでない場合に 0 とする評価方法である．システムの性能は全ての評価用データについて平均値のみをとる代わりに，k の値を複数個設定して評価する．値は $[0, 1]$ の範囲をとり，大きいほど性能が高い．

4.2 TF–IDF

TF–IDF を用いた手法では，表 3 に示すように年代キーの許容幅 $M = 10$，係数 $a = 0.01$ としたときに 3 つの指標で最も良い性能を発揮した．最も性能の高かったシステムの相対順位の分布は図 2 に示すように，中央値は 0.0190，平均値は 0.112，success@k については $k = 1,000$ のときに 0.247，$k = 10,000$ のときに 0.613 となった．

表 3　TF–IDF を用いた手法の性能一覧

年代キー		相対順位		success@k	
許容幅 M	係数 a	中央値	平均値	$k = 1,000$	$k = 10,000$
なし	-	.0201	.113	.253	.580
0	1	.0469	.145	.167	.473
0	.1	.0462	.144	.173	.473
0	.01	.0272	.128	.227	.540
5	1	.0274	.131	.180	.560
5	.1	.0274	.130	.180	.560
5	.01	.0253	.120	.240	.587
10	1	.0190	.115	.187	.607
10	.1	.0190	.115	.200	.607
10	.01	.0190	.112	.247	.613

(各評価指標について最も性能の高いものは下線で示す．)

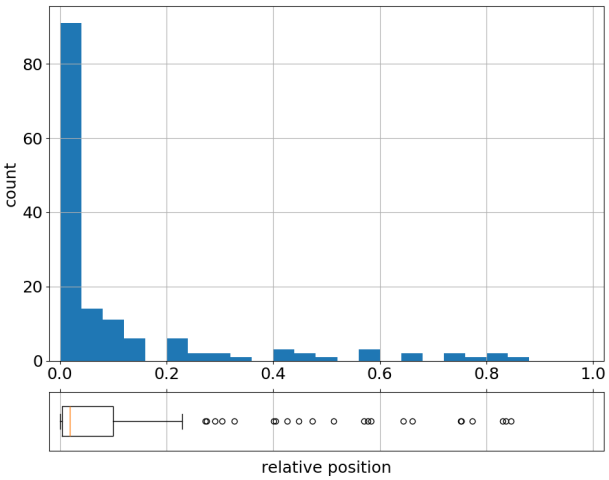


図 2　TF–IDF を用いた手法で最も高い性能を発揮したシステムにおける相対順位の分布

4.3 事前学習済み BERT

BERT(Pre) を用いた手法では，表 4 に示すように年代キーの許容幅 $M = 0$，係数 $a = 0.1$ としたときに 3 つの指標で最も高い性能を発揮した．最も性能の高かったシステムの相対順位の分布は図 3 に示すように，中央値は 0.291，平均値は 0.351，success@k については $k = 1,000$ のときに 0.000，$k = 10,000$ のときに 0.0867 となった．

表 4　事前学習済み BERT を用いた手法の性能一覧

年代キー		相対順位		success@k	
許容幅 M	係数 a	中央値	平均値	$k = 1,000$	$k = 10,000$
なし	-	.459	.463	.000	.0267
0	1	.291	.352	.000	.0867
0	.1	.291	.351	.000	.0867
0	.01	.314	.363	.00667	.0400
5	1	.291	.355	.000	.0400
5	.1	.291	.355	.000	.0400
5	.01	.336	.371	.000	.0400
10	1	.291	.365	.000	.0400
10	.1	.291	.366	.000	.0400
10	.01	.326	.384	.000	.0400

(各評価指標について最も性能の高いものは下線で示す．)

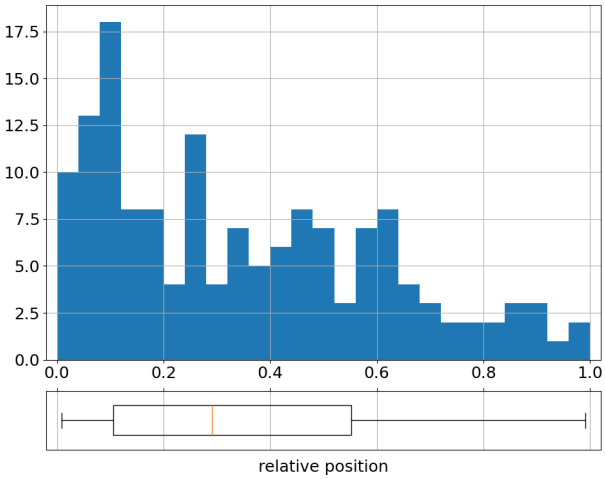


図 3　事前学習済みの BERT を用いた手法で最も高い性能を発揮したシステムにおける相対順位の分布

4.4 再学習後 BERT

BERT(1L) を用いた手法では，表 5 に示すように年代キーの許容幅 $M = 0$，そのときの係数 $a = 0.1$ のときに 3 つの指標で最も高い性能を発揮した．このときのシステムの相対順位の分布は図 4 に示すように，中央値は 0.0484，平均値は 0.166，success@k については $k = 1,000$ のときに 0.0400，$k = 10,000$ のときに 0.433 となった．

また，BERT(2L) を用いた手法では，同様に年代キーの許容幅 $M = 5$，そのときの係数 $a = 0.1$ のときに 3 つの指標で最も高い性能を発揮した．このときのシステムの相対順位の分布は図

5 に示すように，中央値は 0.0508，平均値は 0.184，success@k については $K = 1,000$ のときに 0.0600，$k = 10,000$ のときに 0.473 となった．

表 5　再学習後 BERT を用いた手法の性能一覧

手法	年代キー		相対順位		success@k	
	M	a	中央値	平均値	$k = 1,000$	$k = 10,000$
BERT(1L)	なし	-	.0514	.197	.0400	.427
	0	1	.0716	.194	.0267	.380
	0	.1	.0484	.166	.0400	.433
	0	.01	.0514	.175	.0400	.433
	5	1	.0552	.187	.0200	.393
	5	.1	.0520	.168	.0400	.413
	5	.01	.0514	.177	.0400	.433
	10	1	.0550	.177	.0200	.420
	10	.1	.0523	.171	.0400	.433
	10	.01	.0514	.180	.0400	.440
BERT(2L)	なし	-	.0553	.215	.0467	.440
	0	1	.0713	.209	.0467	.420
	0	.1	.0508	.184	.0600	.453
	0	.01	.0537	.194	.0467	.440
	5	1	.0584	.202	.0400	.447
	5	.1	.0508	.184	.0600	.473
	5	.01	.0547	.195	.0467	.440
	10	1	.0511	.192	.0467	.487
	10	.1	.0511	.187	.0533	.473
	10	.01	.0530	.197	.0467	.453

(各評価指標について，手法ごとに最も性能の高いものは下線で示す．)

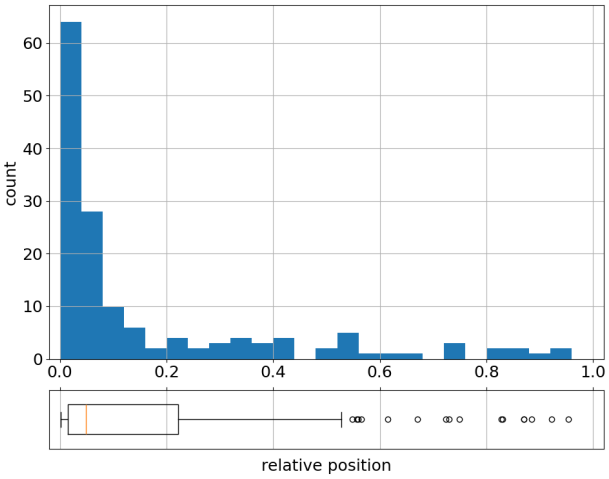


図 4　BERT(1L) を用いた手法で最も高い性能を発揮したシステムにおける相対順位の分布

4.5　ま　と　め

以上の結果を踏まえ，各ベース手法を用いたそれぞれのシステムについて最も高い性能を発揮したシステムについて，表 6 の通りに集約し比較する．

いずれの評価指標においても，TF–IDF を用いた手法が最も高い性能を発揮した．BERT モデルのベースラインとして用意した BERT(Pre) に比べると，再学習を行ったことにより BERT(1L) と BERT(2L) の性能は大きく向上した．しかしながら，BERT をベースとして用いた各手法では TF–IDF を用いた手法に比べて劣る結果となった．特に success@1,000 が success@10,000 に比べるとさらに差のある結果となったため，BERT モデルの手法についてはある程度上位に正解作品を位置付ける能力はあるものの，さらに上位まで押し上げるための能力に欠けていることが分かる．

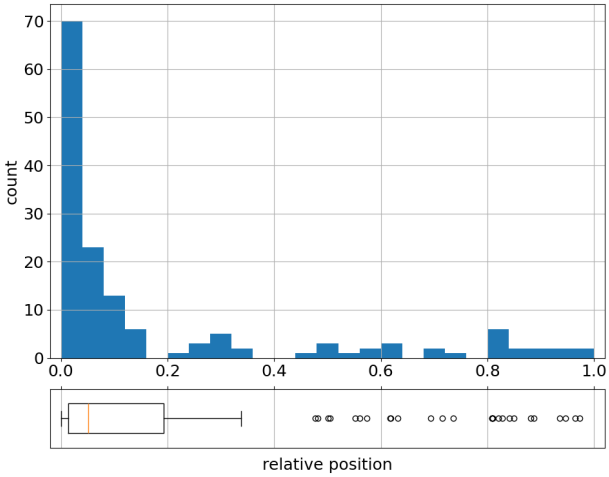


図 5　BERT(2L) を用いた手法で最も高い性能を発揮したシステムにおける相対順位の分布

表 6　各ベース手法を用いた手法のうち最も高い性能を発揮したシステムの比較

手法	年代キー		相対順位		success@k	
	M	a	中央値	平均値	$k = 1,000$	$k = 10,000$
TF–IDF	10	.01	.0190	.112	.237	.613
BERT(Pre)	0	.1	.291	.351	.000	.0867
BERT(1L)	0	.1	.0484	.166	.0400	.433
BERT(2L)	5	.1	.0508	.184	.0600	.473

(各評価指標について最も性能の高いものは下線で示す．)

ただ，最も良い性能を発揮した TF–IDF ベースの手法であっても，success@1,000 の結果より検索結果の上位 1,000 位以内に正解作品を位置付けられたものがたった 23.7% しかなく，実用的な性能には全く届いていないことが分かる．

5　結論・今後の課題

5.1　ま　と　め

本研究では，ユーザが以前見た映像作品についてキャラクタの性格やあらすじ，コンテキストといった断片的な情報を思い出すことができる一方で，作品タイトルを思い出すことができないケースに特化した情報検索システムを作成した．Wikipedia コーパスを用いて初期化した TF–IDF を用いた手法，BERT モデル用いた手法の大きく 2 つのベース手法を用意し，後者については再学習を行った上での検証も行った．BERT モデルの再学習は `bert-base-uncased` に対して improved triplet loss を

使って行った．使用したデータは訓練用に使うことができる150 個のデータに加え，GPT–4 に対してリクエストすることで得た質問データ 1,442 件を併用した．

また，これらの手法の補助的な役割として年代キーを導入した．年代キーは作品群からは Wikipedia コーパスの infobox 内のデータを，あるいは Question Answering の機械学習モデルを利用することでテキスト中から抽出した．一方で質問テキスト側からは Question Answering の機械学習モデルのみを利用して抽出した．

システムの性能評価は正解作品の順位を作品群のサイズで正規化した相対順位と，検索結果の k 位以内に正解作品を位置付けることができたかどうかの二値評価の平均値 success@k の 2 種類をベースに行った．

性能は BERT モデルよりも TF–IDF を用いた手法の方が性能が大きく上回る結果となった．しかしながら，BERT モデルの性能は事前学習済みのものに比べて再学習後のモデルは大きく上回り，学習に使うことができるデータ数を増やすなどの工夫を加えることで，TF–IDF の性能に並ぶ可能性がある．年代キーの導入は全てのベース手法で有効であることが示されたが，適切なパラメータを設定しなければ効果が薄まることが示された．先行研究では年代に関する情報は性能の低下を招くことが示されていたが，今回の実験では適切に扱うことで正の効果があることを示すことができた．

5.2 今後の課題

BERT を再学習した結果モデルの性能が大きく向上したため，再学習に使うことができるサンプル数をさらに増やすことでより性能が向上する可能性がある．本研究では，data augmentation として GPT–4 による返答を活用したが，このときに生成された質問テキストを見ると文頭がほとんど一緒などと共通する部分の多いテキストが生成されてしまっている．この表現の幅を広げるため，API でリクエストする際の temperature をさらに上げてたり指示内容を潤沢にするなどの工夫が追加で必要だと予想される．

また，失敗分析で示したようにあらすじとは無関係なコンテキストに関する情報などを除去することで一部の質問テキストについて大幅に性能が向上することが示された．あらすじの部分のみをさらに抽出するような手段を別途導入することで，さらなる性能向上につながることが期待できる．

文　献

[1] Roger Brown and David McNeill. The “tip of the tongue” phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337, August 1966.

[2] Bennett L. Schwartz and Janet Metcalfe. Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition*, 39(5):737–749, July 2011.

[3] David Elsweiler, Ian Ruthven, and Christopher Jones. Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology*, 58(7):924–946, 2007.

[4] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR ’21, pages 5–14, New York, NY, USA, March 2021. Association for Computing Machinery.

[5] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. Stuff I’ve seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR ’03, pages 72–79, New York, NY, USA, July 2003. Association for Computing Machinery.

[6] David Elsweiler, Ian Ruthven, and Christopher Jones. Dealing with fragmented recollection of context in information management. In *Context-Based Information Retrieval (CIR-05) Workshop in Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, Paris, France, 2005.

[7] Alan S. Brown. A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109(2):204–223, 1991.

[8] Marilyn K. Heine, Beth A. Ober, and Gregory K. Shenaut. Naturally occurring and experimentally induced tip-of-the-tongue experiences in three adult age groups. *Psychology and Aging*, 14(3):445–457, 1999.

[9] Bennett L. Schwartz. Sparkling at the end of the tongue: The etiology of tip-of-the-tongue phenomenology. *Psychonomic Bulletin & Review*, 6(3):379–393, September 1999.

[10] Søren Wichmann and Kerry Hull. Loanwords in Q’eqchi’, a Mayan language of Guatemala. In *Loanwords in the World’s Languages: A Comparative Handbook*, page 1103. Mouton de Gruyter, New York, NY, USA, 2009.

[11] Tim Brennen, Anne Vikan, and Ragnhild Dybdahl. Are tip-of-the-tongue states universal? Evidence from the speakers of an unwritten language. *Memory*, 15(2):167–176, February 2007.

[12] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14325–14329, September 2008.

[13] Michael A. Cohen, Todd S. Horowitz, and Jeremy M. Wolfe. Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):6008–6010, April 2009.

[14] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of the Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500–225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994.

[15] Jaime Arguello, Samarth Bhargav, Bhaskar Mitra, Fernando Diaz, and Evangelos Kanoulas. TREC 2023 Tip-of-the-Tongue (ToT) Track.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018.

[17] Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning Fine-grained Image Similarity with Deep Ranking, April 2014.

[18] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, June 2016.

[19] Mahmut Kaya and Hasan Şakir Bilge. Deep Metric Learning: A Survey. *Symmetry*, 11(9):1066, September 2019.

[20] Elad Hoffer and Nir Ailon. Deep metric learning using Triplet network, December 2018.

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, February 2020.

ドメイン特徴に基づくランキング学習モデルのドメイン適応

伊藤　拓誠† 　丸田　敦貴†† 　加藤　誠†††

† 筑波大学 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
††† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s2010485@u.tsukuba.ac.jp, ††s1711567@klis.tsukuba.ac.jp, †††mpkato@acm.org

あらまし　本論文は，ランキング学習でターゲットドメインの適合性判定データが利用できない設定において，ランキング学習モデルのドメイン適応を行うことを目的とする．我々は，ターゲットドメインからドメイン特徴を抽出し，ドメイン特徴のみからターゲットドメインにおけるランキングモデルの最適な重みを予測する重み予測モデルを提案する．また，重み予測モデルの学習には，多種多様なデータを含むデータセットをドメイン特徴が異なるように分割して複数のドメインを用意し，各ドメインでドメイン特徴の抽出とランキングモデルの重みの最適化を行うことで作成した学習データを利用する．評価実験の結果，重み予測モデルで重みを予測したモデルの性能が，ドメインの差異を考慮せずに学習した汎用的なモデルの性能を上回ることが確認された．

キーワード　情報検索，検索モデル，ランキング学習，ドメイン適応

1　はじめに

情報検索分野におけるランキングタスクは，与えられたクエリに対して関連する文書を適合性の高い順に並べ替えることを目的とする．このタスクを行うランキングモデルを構築する方法の一つにランキング学習 [11] がある．ランキング学習は機械学習を活用して性能の良いランキングモデルを構築する手法で，クエリと文書，それらの適合性判定データからなる学習データを大量に用いてモデルを学習する．ランキング学習の効果は広く認められており，Google や Bing といった文書検索を中心とするサービスで研究が行われているほか，Airbnb や eBay などのそれ以外のサービスでも活用されている [7, 20, 22, 27].

効果的なランキング学習を行うには，他の機械学習タスクと同様にタスクを行うドメインにおける学習データが大量に必要となる．学習データの中でも，人手によるアノテーションや大規模なクリックデータが必要な適合性判定データの収集が難しい．この課題に対して，学習データの豊富なソースドメインから獲得した知識を学習データの乏しいターゲットドメインに転移することで学習データの不足を補う転移学習の研究が行われている [1–3, 12]．しかしながら，これらの研究ではターゲットドメインの適合性判データを少なからず利用しており，ターゲットドメインの適合性判定データがない場合に性能の良いランキングモデルを構築する方法は依然として確立していない．

そこで本論文では，ターゲットドメインの適合性判定データを用いずに転移学習を行うトランスダクティブ転移学習 [18] という設定におけるランキング学習手法を提案する．トランスダクティブ転移学習の設定でランキング学習を行う場合，利用できるデータは一般的にソースドメインのクエリ，文書，適合性判定データに加え，ターゲットドメインのクエリ，文書となる．ただし，適合性判定データではないユーザの情報などを追加で利用することはできる．これらのデータのみを利用して，ターゲットドメインにおいて性能の良いランキングモデルを構築することが本研究の目的である．

本研究では，ドメインからドメイン特徴と呼ばれる特徴を抽出し，ドメイン特徴のみからそのドメインにおけるランキングモデルの最適な重みを予測する**重み予測モデル**を構築することで上記の提案を実現する．重み予測モデルは，複数のソースドメインを用意し，それらのドメインのドメイン特徴とランキングモデルの最適な重みとの関係性を学習することで構築する．また，複数のソースドメインを用意する際，単純にドメインの異なる複数のデータセットを集める方法では学習に十分な数を用意するのが難しい．そこで本研究では，多種多様なデータを含む 1 つの大きなデータセットをドメインが異なるように分割して複数のドメインを用意する方法を採用する．

実験では，分割元のデータセットとして多種多様な Web 文書とクエリを含む AOLIA データセット [13] を利用した．このデータセットをクエリ・文書ペアの特徴に基づき k-means 法で分割し，500 個のドメインを作成した．500 個のドメインを学習用のソースドメインと検証用のターゲットドメインおよび評価用のターゲットドメインに分割した後，各ドメインでドメイン特徴の抽出とランキングモデルの重みの最適化を行い，それらを入出力とする重み予測モデルを学習した．また，重みを予測する対象であるランキングモデルには，単純な重みを持つ線形のモデルを採用した．実験の結果，重み予測モデルで予測した重みを活用した提案モデルは，ドメインの差異を考慮せず大量のデータで学習した汎用的なモデルよりも性能が向上した．また，提案モデルは，最も性能が良くなると予想された，評価用に適合性判定データがあるターゲットドメインで直接学習したモデルよりも性能が高くなることが確認された．さらに，主にクエリと文書の関連度を測るクエリ・文書特徴をドメイン特徴として用いた場合に提案モデルの性能が最も高くなることが

明らかになった．

この論文における我々の貢献を以下に示す：(1) 転移学習の問題設定の中でもターゲットドメインのラベルを利用することができないトランスダクティブ転移学習という設定において利用することができるランキング学習タスクの手法を提案した．(2) 提案手法がドメインの差異を考慮せずに大量のデータで学習した汎用的なモデルよりも効果的であることを実験によって明らかにした．また，利用するドメイン特徴によって提案手法の性能が変化することを明らかにした．

本論文の構成は以下の通りである．2 節ではランキング学習における転移学習に関する関連研究，および， 適応的なランキング学習に関する関連研究について述べる．3 節では問題設定を説明し，重み予測モデルの学習方法とデータセットの分割方法について述べる．4 節では実験結果を示す．最後に，5 節では今後の課題と共に本論文の結論を述べる．

2 関 連 研 究

本節では，ランキング学習における転移学習に関する関連研究，および，適応的なランキング学習に関する関連研究について述べる．

2.1 ランキング学習における転移学習

学習データの豊富なソースドメインで得た知識を学習データの乏しいターゲットドメインに転移することでターゲットドメインの学習データの不足を補う転移学習（Transfer Learning）はランキング学習の分野でも研究されている．ランキング学習における代表的な転移学習の手法には，アルゴリズムレベルの手法 [1, 3]，特徴量レベルの手法 [2]，データレベルの手法 [2, 12] が存在する．しかし，上記の手法はいずれもソースドメインのラベルとターゲットドメインのラベルの両方が利用できる設定であり，本研究のようにターゲットドメインのラベルが利用できない設定には対応していない．

また，Gao らによりトランスダクティブ転移学習の設定におけるランキング学習手法が提案されている [4]．具体的には，まず，特徴空間におけるソースドメインのインスタンスの分布とターゲットドメインのインスタンスの分布を二分する超平面を算出する．次に，ソースドメインのインスタンスのうち超平面に距離が近いものがターゲットドメインの学習に重要であると見なし，ターゲットドメインの学習に与える影響が大きくなるよう重みづける．そして，重みづけたソースドメインのデータを利用してターゲットドメインのモデルを学習することでターゲットドメインのモデルを構築する．これはデータレベルの手法である．上記の手法において効果的な超平面を算出するためには，ターゲットドメインにある程度のクエリ・文書ペアが必要である．一方，本研究の提案手法はドメインからドメイン特徴を 1 つ抽出すれば，予め学習した重み予測モデルで重みの予測を行うことができる．入力としてドメイン特徴が 1 つあれば良いので，ドメイン特徴を仮想的に設定することも可能で，その場合にも一定の効果が見込める．これは，ターゲットドメインにデータが十分存在しない場面や，文書は存在するがクエリは存在しないような場面など，より広い場面で利用することが可能であることを示す．また，本研究の提案手法はアルゴリズムレベルの手法であるため，データレベルの手法とは異なりターゲットドメインで一からモデルを学習する必要がないという利点がある．

2.2 適応的なランキング学習

あらゆるクエリに対して同一のモデルを用いるのではなく，入力されるクエリの種類に対して適応的にモデルの振る舞いを変更することでランキングタスクにおける性能が向上することが報告されている [9]．

これを受けて，予め複数のランキングモデルを構築しておき，入力されるクエリに応じてモデルを選択することでランキングの性能向上を目指す研究がなされている [5, 17, 21]．具体例として，Geng らの手法 [5] を紹介する．この手法では，学習データ中のクエリをクエリ特徴に基づいてクラスタリングした後，クラスタごとにランキングモデルを構築することで複数のランキングモデルを用意する．そして，学習データにない新たなクエリが入力されると，入力されたクエリの特徴からそのクエリが属するクラスタを算出する．入力されたクエリに対して，そのクエリが属すると算出されたクラスタで構築されたモデルを選択してランキングタスクに利用することで，クエリに適応したランキングを行う事が可能となる．本論文はドメインごとにモデルを構築するため，その点で類似しているが，上記の手法は，単一のドメインにおけるクエリの違いを扱っており，異なるドメインにおける効果については検証されていない．更に，上記の手法はクエリごとにモデルを選択するための処理が必要であり，速度が重要となる場面においては有効でない．

また，Macdonald [15] らは回帰木を利用したランキングモデルにクエリ特徴を利用することでモデルの性能が向上すると報告している．この性能の向上は，あるクエリ特徴で木を分岐させた際に分岐以降のサブツリーがある種のクエリに特化して学習され，1 つのモデルを利用していてもクエリの特徴に応じて処理を変更することが可能であることに起因している．しかし，この手法も単一のドメインにおけるクエリの違いを扱っており，異なるドメインにおける効果については検証されていない．

3 提 案 手 法

本節では，まず問題設定について説明を行う．それから，提案手法の詳細と，本研究で利用するデータセットを構築する方法について述べる．

3.1 問 題 設 定

本論文では，複数のソースドメインから得た知識を活用して，あるターゲットドメインにおいて性能の良いモデルを構築することを目的とする．ソースドメインとは，ドメイン内のクエリ，文書，その他の情報およびクエリ・文書ペアに対する適合性判定データにアクセスできるドメインである．一方ターゲットドメインとは，ドメイン内のクエリ，文書，その他

の情報にはアクセスできるが，適合性判定データにはアクセスできないドメインである．ここで，複数のソースドメインを $\{m_1, m_2, m_3, ..., m_{n-1}\}$，あるターゲットドメインを m_n としたドメイン集合 $\mathcal{M} = \{m_1, m_2, m_3, ..., m_n\}$ を考える．m_i は 1 つのドメインを表し， クエリ集合 Q_i，文書集合 D_i，適合性判定済みのクエリ・文書ペアの集合 $J_i \subseteq Q_i \times D_i$，適合性判定の基準 $r_i : J_i \to \mathbb{N} \cup \{0\}$ などの要素を含む．上述の通り，ターゲットドメイン m_n には適合性判定済みのクエリ・文書ペアが存在しないため $J_n = \varnothing$ である．これらの要素の他にも，ドメインにはユーザ情報やクエリログなど様々な情報が含まれる．

あるドメイン m_i でランキングタスクを行うためには，与えられたクエリ $q \in Q_i$ に対してドメイン内の文書のランキングを決定する必要がある．特に，本論文では与えられたクエリ $q \in Q_i$ とドメイン内の各文書 $d \in D_i$ との適合性を予測し，文書を適合性の降順に並べ替えることで文書のランキングを決定する．また，検索システムの実用上，適合性判定がされていないクエリに対してランキングを決定する必要がある．そのためには，ある $(q, d) \notin J_i$ に対してその適合性を予測できればよい．適合性の予測には，ドメインごとに最適化された重み $\mathbf{w}_i$ を持つモデル $f : Q \times D \to \mathbb{R}$ を構築して利用する．ここで，Q はあらゆるクエリを含むクエリ集合，D はあらゆる文書を含む文書集合を示す．

また，最適化された重み $\mathbf{w}_i$ を得る代表的な方法として，ランキング学習がある．ランキング学習とは，機械学習を活用して性能の良いランキングモデルを構築する手法である．ランキング学習の代表的な手法にポイントワイズ，ペアワイズ，リストワイズの手法があるが，本論文ではポイントワイズの手法について取り上げる．ポイントワイズの手法では，学習データにおいて入力となるクエリ・文書ペア $(q, d) \in J_i$ からその適合性 $r_i(q, d)$ を上手く予測できるようにモデル f の重み $\mathbf{w}_i$ を学習する．ランキング学習は，関数 $\mathbf{T} : \mathcal{J} \times R \to \mathbb{R}^l$ を用いて，式 1 ように重みを得る工程と定式化することができる．ただし，$\mathcal{J}$ はあらゆる適合性判定済みのクエリ・文書ペアを含む集合のべき集合で，R はあらゆる適合性判定の基準 r を含む集合である．

$$\mathbf{w}_i = \mathbf{T}(J_i, r_i) \tag{1}$$

ソースドメイン m_i $(1 \leqq i < n)$ ではこのようなランキング学習手法を適用し，J_i を利用してモデル f の振る舞いを r_i に近づけるように最適なパラメータ $\mathbf{w}_i$ を学習することで性能の良いランキングモデルを構築することができる．しかし，先に述べたように，ターゲットドメイン m_n には $\mathbf{w}_n$ の学習に必要な適合性判定済みのクエリ・文書ペアの集合 J_n の要素が存在せず，従来のランキング学習手法を適用することができない．本論文では，このようなターゲットドメイン m_n においてもランキングモデルの最適な重み $\mathbf{w}_n$ を得て，性能の良いランキングモデルを構築するという問題を解く．

3.2 重み予測モデル

本論文で提案する重み予測モデルを用いて上記の問題を解くフレームワークを図 1 に示す．主なアイデアは，ソースドメインとターゲットドメインで共通して利用できる要素からドメイン特徴を抽出し，ドメイン特徴からそのドメインのモデルの最適な重みを予測することである．ドメイン特徴とは，各ドメインがどのようなものかを示す特徴である．各ドメインから 1 つずつ抽出し，抽出した特徴が各ドメインで異なることが望ましい．ドメイン特徴としては 文書のトピックなどのドメイン内の文書から抽出することができる特徴，クエリの長さの平均値などのドメイン内のクエリから抽出することができる特徴， クエリと関連する文書の多寡などのドメイン内の文書・クエリペアから抽出することができる特徴など様々な特徴を利用することができる．本研究で利用したドメイン特徴については次節で詳しく述べる．Macdonald らの研究で，データ数が同程度のドメインを用いて実験すると，あるドメインで学習したランキングモデルをそのまま他のドメインで利用した場合ランキングタスクにおける性能が低下すると報告されている [14]．このことから，ランキングモデルの最適な重みはドメインごとに異なることが示唆される．そこで，ドメインの特徴をうまく表現し，ドメイン特徴とランキングモデルの重みの関係性を学習すれば，ドメイン特徴からランキングモデルの最適な重みが予想できると考えた．例えば，ドメイン特徴としてクエリの長さの平均値を利用し，ランキングモデルの特徴である文書の URL に含まれる “/” の数の重みを予測するを場合を考える．クエリが長いことはより具体的な情報要求を表現していることを示し，クエリの長さの平均値が大きいドメインは具体的な情報要求が多いドメインであると期待される．また，URL に含まれる “/” の数は URL の階層構造の深さを示し，URL の階層構造が深い文書は，より具体的な情報や特定のトピックに焦点を当てている傾向にある．上記の傾向から，より具体的な情報要求に関連する文書は URL の階層構造が深い文書である可能性があり，したがって，クエリの長さの平均値が大きいドメインほど文書の URL に含まれる “/” の数に対する重みが大きいといった関係性があると考えられる．

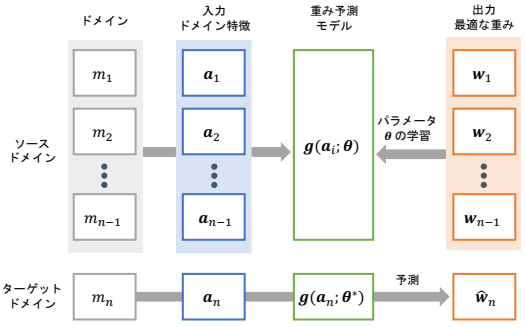


図 1　提案するフレームワークの概要図.

ここからは具体的な手法について説明する．まず，特徴抽出器 $\boldsymbol{\psi} : \mathcal{M} \to \mathbb{R}^m$ を用いてドメイン特徴 $\mathbf{a}_i = \boldsymbol{\psi}(m_i)$ を抽出する．そして，重み予測モデル $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^l$ を用いて， 式 2 に示すように対象のドメインのモデルの最適な重みの予測値を得る．$\mathbf{g}$ は最適化されたパラメータ $\boldsymbol{\theta}^*$ を持つモデルである．

Algorithm 1 ターゲットドメインでランキングモデルの重みを得るアルゴリズム.

Input: $\mathcal{M}$
Output: $\hat{\mathbf{w}}_n$

$S \leftarrow \{\}$
for $i = 1$ to $n-1$ **do**
  $\mathbf{a}_i \leftarrow \boldsymbol{\psi}(m_i)$
  $\mathbf{w}_i \leftarrow \mathbf{T}(J_i, r_i)$（式 1）
  $S \leftarrow S \cup \{(\mathbf{a}_i, \mathbf{w}_i)\}$
end for
$\mathbf{a}_n \leftarrow \boldsymbol{\psi}(m_n)$
$\boldsymbol{\theta}^* \leftarrow \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; S)$（式 4）
$\hat{\mathbf{w}}_n \leftarrow \mathbf{g}(\mathbf{a}_n; \boldsymbol{\theta}^*)$（式 2）

$$\hat{\mathbf{w}}_i = \mathbf{g}(\mathbf{a}_i; \boldsymbol{\theta}^*) \tag{2}$$

パラメータ $\boldsymbol{\theta}^*$ は，複数のソースドメインにおいて，$\hat{\mathbf{w}}_i = \mathbf{g}(\mathbf{a}_i; \boldsymbol{\theta})$ が $\mathbf{w}_i$ に近づくように $\boldsymbol{\theta}$ を学習することで得る．ドメイン特徴からランキング学習モデルの重みを予測する問題は，最適な重み $\mathbf{w}_i$ をラベル，ドメイン特徴 $\boldsymbol{a_i}$ をデータの特徴とする式 3 で示すデータセット S を利用して多次元の回帰問題を学習したモデルで解くことができる．

$$S = \{(\mathbf{a}_i, \mathbf{w}_i) | i = 1, 2, ..., n-1\} \tag{3}$$

そのため，パラメータ $\boldsymbol{\theta}^*$ は式 4 のように，多次元の回帰問題に利用できる損失関数 L に対する損失を各ドメインで最小化するような $\boldsymbol{\theta}$ を学習することで決定する．

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; S) \tag{4}$$

このような重み予測モデルを用いることで，ターゲットドメインにおいても性能の良いランキングモデルを構築することができるようになる．提案手法の全体をまとめたアルゴリズムを Algorithm 1 に示す．

3.3 データセットの分割によるドメインの作成

重み予測モデルを学習するためには複数のソースドメインが必要である．公開されているデータセットを人手で収集し，ソースドメインとして活用する方法もあるが，この方法では学習に十分な数のドメインを用意するのが難しい．そこで，本論文では多種多様なデータを含むデータセットを分割することで，複数のドメインを用意する方法を採用する．

データセットを分割する方法は複数考えられるため，分割を評価するための指標が必要である．先述の通り，ランキングモデルの最適な重みはドメインごとに異なることが確認されている．そこで，以下のような仮定を置く．まず，分割後の各ドメインのデータセットでランキング学習を行い，ドメインごとに最適化されたランキングモデルの重みを得る．この重みが各ドメインで異なれば異なるほど，各ドメインの性質も異なり良い分割であると言える．

この仮定に基づいて分割を評価するための指標として，各ドメインのモデルの各特徴に対するの重みの分散の和（以降，単に**重みの分散の和**と呼ぶ.）が利用できる．ある分割でできた複数のドメインでランキングモデルの重みを学習した結果重みの分散の和が大きいということは，ランキングモデルの重みが各ドメインで大きく異なることを示す．先述の仮定に基づくと，ランキングモデルの重みが各ドメインで大きく異なることは，ドメインの性質が大きく異なることを示す．したがって，重みの分散の和は分割の評価指標として妥当である．より具体的には，l 次元の重みに対して，式 5 が大きいほど良い分割であるとして評価する．ここで，W は分割後の各ドメインで学習したモデルの重み $\mathbf{w}_i$ の集合，w_{ij} は重み $\mathbf{w}_i$ の j 番目の要素を指す．また，$\bar{w}_j$ は各重みの j 番目の要素の平均値である．

$$\sum_{j=1}^{l} \frac{1}{|W|} \sum_{\mathbf{w}_i \in W} (w_{ij} - \bar{w}_j)^2 \tag{5}$$

本研究では，同一のクエリを持つクエリ・文書ペアは同一のドメインに属するという仮定のもと，以下の手順でドメインを分割した．まず，データセットの全クエリ・文書ペア (q, d) から特徴 $\mathbf{x}_{q,d}$ を抽出する．次に，同一のクエリを持つクエリ・文書ペアの特徴を，平均を取ることで 1 つにまとめる．具体的には，$\mathbf{x}_q = \frac{1}{n_q} \sum_{i=1}^{n_q} \mathbf{x}_{q,d_i}$ を各クエリの特徴とする．ここで，n_q はあるクエリ q に対して紐づけられた文書の数を示し，$(q, d_1), (q, d_2), ..., (q, d_{n_q})$ という形でクエリとクエリに紐づけられた文書のペアを表現する．その後，$\mathbf{x}_q$ の集合 $X^{(q)}$ に対して，k-means 法 [16] を用いてクラスタリングを行う．すると，$\mathbf{x}_q$ のクラスタ $X_1^{(q)}, X_2^{(q)}, ..., X_k^{(q)}$ ができる．$\mathbf{x}_q$ はあるクエリを代表する特徴であるため，$\mathbf{x}_q$ のクラスタはクエリのクラスタであると考えることができる．あるクラスタに含まれる全クエリの集合をクエリ集合 Q，$q \in Q$ に紐づく全文書の集合を文書集合 D とすることで，1 つのクラスタから 1 つのドメインを得ることができる．上記の方法では，$\mathbf{x}_{q,d}$ として様々な特徴セットを利用することが可能である．実際に利用した特徴に関しては，次節で詳しく述べる．

4 実　　験

我々のタスクには公開されたデータセットが存在しないため，まずデータセットの作成の概略とその統計情報について説明する．それからベースライン手法を含む実験設定について紹介し，最後に実験結果を示す．本実験では，まずデータセットを分割する際に効果的な特徴を明らかにする．次に，重み予測モデルを利用してドメイン適応することにより構築したランキングモデルの性能の有効性を明らかにする．最後に，重み予測モデルに入力するドメイン特徴として有効な特徴を明らかにする．

4.1 データセット

3 節で提案した方法で我々のタスクに適したデータセットを作成するにあたり，分割元となるデータセットとして AOLIA [13] を採用した．AOLIA は，AOL Query Log [19] を利用して作られたデータセットである．AOL Query Log には 2006 年の

3 月から 5 月の間に AOL Search[1]のサーチエンジンに実際にユーザが入力したクエリと，クエリを入力した日時，そのクエリに対してユーザが実際にクリックした文書の URL の情報などが匿名化されて含まれる．AOL Query Log に含まれる文書の URL からタイトルや本文を含む実際の文書の情報を得ることでクエリと文書のテキストが利用できる大規模なデータセットを構築することが可能である．AOLIA は，URL から文書の情報を得る際に Internet Archive[2]を利用しており，URL のリンク切れを回避できるため利用できる文書の量が多くなるほか，クエリが入力された当時の状態に近い文書の情報を利用することができる．また，AOLIA データセットは他のデータセットに比べてクエリを多く含んでおり，分割後のドメインでも多くのクエリを利用できるほか，クエリや文書のテキストが含まれるため，ドメイン特徴として様々な特徴を抽出し利用できるという利点がある．

通常のランキングタスクは主に 1 つの言語を対象として行われるため，AOLIA データセットから英語以外のクエリと文書を取り除いた．ここで，言語の検出にはデータセットを作成した論文でも利用された FastText [8] のモデル [3]を利用し，モデルが英語であると予測した確率が 0.6 未満のクエリと文書を取り除いた．また，タイトルまたは本文が空の文書も取り除いた．

AOLIA にはクエリ・文書ペアに対して人間が下した適合性判定データは含まれておらず，ランキングタスクのデータセットとして活用するには AOLIA に含まれるクエリログからクエリ・文書ペアの適合性を推定する必要がある．しかし，クエリに対して実際に提示されたランキングの情報やクリックされていない文書の情報が含まれていないため，既存のクリックモデルを活用してクエリ・文書ペアの適合性を推定することが難しい．そこで，本研究では以下の方法で適合・不適合の二値からなる適合性を付与した．まず，ユーザが入力したクエリに対してクリックされた文書を適合とみなした．次に，ユーザが入力したクエリに対してフィルタリング後の文書集合中の文書のランキングを算出し，ランキングの上位からクリックされていない文書を 10 件分収集し不適合とみなした．クエリログの収集時と似たランキングの中でクリックされなかった文書は不適合である可能性が高いと考えられるため，ランキングの算出には当時も利用されていたと考えられる単語ベースのランキングモデルである BM25 [24] を用いた．実際に，BM25 を利用して AOL Query Log に含まれるクエリに対する文書の検索結果を再現して利用する例がある [6]．

フィルタリング後の文書数は 1,163,269 件で，フィルタリング後の文書集合中に正例を持つクエリは 3,688,945 件あった．その中からランダムに抽出した 360,000 件のクエリに対し，上記の方法で適合文書と不適合文書を決定した結果，4,319,209 件の適合性が付与されたクエリ・文書ペアからなる分割元のデータセットが構築された．構築されたデータセットの統計情報を表 1 に示す．

表 1　作成したデータセットの統計情報．

	データ数	クエリ数	文書数
分割元のデータセット	4,319,209	360,000	868,981
分割後のデータセット（平均値）	8,638	720	1,737

上記の分割元のデータセットを，次節で述べる特徴に基づいて k-means 法で 500 分割し，500 個のドメインを得た．500 分割すると，表 1 にも示したように，1 ドメインあたり平均 8,638 件のクエリ・文書ペアが割り当てられる．ランキング学習のデータセットである LETOR データセット [23] の OHSUMED コーパスには 16,140 件のクエリ・文書ペアが含まれ，通常，5-Fold 交差検証で利用される．5-Fold 交差検証で訓練・検証・評価を行う場合，一般的に，各 Fold で訓練データとして全体のクエリ・文書ペアの件数の 5 分の 3 の約 9,500 件のクエリ・文書ペアを利用することになる．各 Fold の訓練データを 1 つのドメインと捉えれば，分割後のデータセットのクエリ・文書ペアの件数はこの件数に近いため，妥当な件数であると言える．500 個のドメインを作成した後，それらをランダムに分割し，350 個のドメインを重み予測モデルの訓練に用いるソースドメイン，75 個のドメインをハイパーパラメータのチューニングなどの用途で検証に用いるターゲットドメイン，残りの 75 個のドメインを最終的な評価に用いるターゲットドメインとした．このようにして得た 500 個のドメインからなるデータセットを，我々のタスクに適したデータセットとして活用した．

4.2 実験設定

本節では，本実験における実験設定について説明する．

a） 利用した特徴

まず，データセットを分割するための特徴，ドメイン特徴，および，ランキング学習の特徴として利用した特徴について説明する．本実験で利用した特徴の詳細を表 2 に示す．利用した特徴は大きく分けてクエリ特徴（Q），文書特徴（D），クエリ・文書特徴（Q-D）の 3 種類である．これらの特徴はいずれもクエリ・文書ペア (q, d) に対して算出される．以下，表 2 で用いた数式について説明する．クエリ q および文書 d は単語の多重集合として扱い，$q_i \in q$ はあるクエリを構成する各単語を示す．また，各文書 $d \in D$ に含まれる全単語の多重集合を C とする．$df(t)$ は文書頻度で，ある文書集合 D においてある単語 t が出現する文書の件数を示す．$c(t, V)$ は単語の頻度で，ある単語 t が単語の多重集合 V において出現する回数を示す．

クエリ特徴は，検索時のクエリのみから得られる特徴である．表中の IDF (Inverse Document Frequency) は，ある文書集合中である単語が現れる文書の割合の逆数，ICF (Inverse Collection Frequency) は．ある文書集合内の全単語に占めるある単語の割合の逆数である．

文書特徴は，文書のみから得られる特徴である．文書のリーダビリティスコアとしては ARI (Automated Readability Index) [25] を利用した．文書のトピックとしては Open Directory

1：https://search.aol.com/
2：https://archive.org/
3：https://fasttext.cc/docs/en/language-identification.html

表 2　データセットを分割するための特徴，ドメイン特徴，ランキング学習で用いる特徴として利用した特徴.

番号	特徴	種類
1	クエリに含まれるトークンの数（$\|q\|$）	Q
2	$\sum_{q_i \in q} \mathrm{IDF}(q_i)$ ただし，$\mathrm{IDF}(q_i) = \log\left(\frac{\|D\|}{df(q_i)+1}\right)$	Q
3	$\frac{1}{\|q\|}\sum_{q_i \in q} \mathrm{IDF}(q_i)$	Q
4	$\max_{q_i \in q} \mathrm{IDF}(q_i)$	Q
5	$\sqrt{\frac{1}{\|q\|}\sum_{q_i \in q}\left(\mathrm{IDF}(q_i) - \overline{\mathrm{IDF}}\right)^2}$ ただし，$\overline{\mathrm{IDF}}$ は $\mathrm{IDF}(q_i)$ の平均値	Q
6	$\sum_{q_i \in q} \mathrm{ICF}(q_i)$ ただし，$\mathrm{ICF}(q_i) = \log\left(\frac{\|C\|}{c(q_i,C)+1}\right)$	Q
7	$\frac{1}{\|q\|}\sum_{q_i \in q} \mathrm{ICF}(q_i)$	Q
8/9	文書の本文/タイトルに含まれるトークンの数（$\|d\|$）	D
10/11	文書の本文/タイトルにおける $\log\|d\|$	D
12	文書の本文に含まれるセンテンスの数	D
13	文書の本文のリーダビリティスコア	D
14	文書の本文のトピック	D
15	文書の URL のドメインの長さ	D
16	文書の URL のドメインの頻度	D
17	文書の URL に含まれる “/” の数	D
18/19	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} c(q_i, d)$	Q-D
20/21	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(c(q_i, d) + 1\right)$	Q-D
22/23	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\|d\|}$	Q-D
24/25	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\|d\|} + 1\right)$	Q-D
26/27	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\frac{\|D\|}{df(q_i)}\right)$	Q-D
28/29	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\log\left(\frac{\|D\|}{df(q_i)}\right)\right)$	Q-D
30/31	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,C)}{\|D\|} + 1\right)$	Q-D
32/33	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\|d\|} \cdot \log\left(\frac{\|D\|}{df(q_i)}\right) + 1\right)$	Q-D
34/35	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\|D\|}{df(q_i)}\right)$	Q-D
36/37	文書の本文/タイトルにおける $\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\|d\|} \cdot \frac{\|D\|}{c(q_i,C)} + 1\right)$	Q-D
38/39	文書の本文/タイトルにおける BM25 のスコア（BM25）	Q-D
40/41	文書の本文/タイトルにおける $\log$(BM25)	Q-D
42/43	文書の本文/タイトルにおける LMIR.DIR のスコア	Q-D
44/45	文書の本文/タイトルにおける LMIR.JM のスコア	Q-D
46/47	文書の本文/タイトルにおける LMIR.ABS のスコア	Q-D

Project[4]に含まれるトピックのうち，後述のデータセットで利用された 13 のトピックを用いた．これらのトピックを AOLIA データセット中の各文書に対して推定するために Open Directory Project のデータセット [5]を用いて，テキストが 13 のトピックそれぞれに分類される確率を予測するモデルを学習した．学習に際して，テキストには spaCy のモデル [6]を利用してトークナイズと原形化処理を施し，学習データにおいて原形化されたトークンの集合を語彙として扱った．また，各トークンの TF-IDF を値とする語彙次元のベクトルを各テキストの特徴として抽出した．トピックを予測するモデルにはロジスティック回帰モデルを利用し，文書が各トピックに属する確率を出力するように学習した．URL やハイパーリンクのネットワークに関する文書特徴もランキング学習で良く利用される特徴であるが，AOLIA データセットからはネットワークに関する情報が得られなかったので，URL のみから得られる特徴を独自に 3 つ利用した．

クエリ・文書特徴は，クエリと文書のペアから得られる特徴で，クエリと文書の類似度に関するものが多い．また，ランキング学習の特徴として主に活用されるのはこの特徴である．番号 40 から 47 における LMIR (Language Model for Information Retrieval) は言語モデルによって推定されるクエリ尤度であり，それぞれスムージングの方法が異なる．クエリ・文書特徴の選定と実装は日本語の検索質問ドメインにおけるランキング学習用のデータセットである OpenLiveQ [10] のものを参考にした．上記のデータセットで利用されている 15 のクエリ・文書特徴を文書のタイトルおよび本文の 2 つのフィールドに適用し，30 個の特徴を用意した．

b）　ランキング学習モデル

ランキングモデルが活用する特徴は 1 通りに固定した．ランキングモデルが活用する特徴としては OpenLiveQ データセットを参考に，表 2 のクエリ・文書特徴の全て（番号 18 から 47）と文書特徴のうちトークンの数に関連する特徴（番号 8 から 11）を利用した．また，URL に関する 3 つの特徴（番号 15 から 17）も利用し，合計 37 次元の特徴を用いた．

本実験では，ランキングモデルとして RankLib[7]の Coordinate Ascent を利用した．このモデルは単純な線形のパラメータを持つため，重みの予測が行いやすいと考えた．分割後の各ドメインでランキングモデルを構築する際や，ベースラインとして全ソースドメインのデータを利用してモデルを構築する際は，以下の手順で学習を行った．まず，与えられたドメインのデータセットを 8:1:1 の割合で学習データ，検証データ，評価データに分割した．その後，学習データを用いて上記のモデルを nDCG@10 が最大化するように学習した．1 つのドメインに対してモデルのパラメータを複数回学習し，検証データで最も性能の良かったパラメータを評価に採用した．

c）　重み予測モデル

分割後の 500 個のドメインを利用して式 3 で示すデータセットを得て重み予測モデルの学習に利用した．ドメイン特徴は，データセットを分割する際のクエリの特徴 $\mathbf{x}_q$ としてクエリ・文書ペアの特徴をクエリごとにまとめて平均を取ったように，クエリ・文書ペアの特徴をドメインごとにまとめて平均を取ることによって抽出した．また，重み予測モデルとしてはランダムフォレストによる多次元回帰のモデルを用いた．多層パーセプトロンと線形回帰による多次元回帰のモデルでも実験を行ったが，ランダムフォレストの性能を超えることはなく，かつ結果の傾向も類似していたため，ランダムフォレストによる多次元回帰のモデルの結果のみを記載する．損失関数としては平均二乗誤差を利用し，ハイパーパラメータは検証用のターゲットドメインにおけるデータを用いて nDCG@10 を最大化するように決定した．

d）　評　　価

評価実験では，まず，データセットの分割に用いる特徴セットととして何が有効であるかを明らかにするための実験を行った．具体的には，データセットの分割に用いる特徴セットを変更して実際に複数の分割を行い，評価指標として重みの分散の和を利用することで比較した．

次に，提案手法で予測した重みを活用した**提案モデル**と以下

4：http://www.odp.org/homepage.php

5：https://www.kaggle.com/datasets/lucmichalski/dmoz-bert-multiclass-web-classification-dataset/data

6：https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.7.1

7：https://sourceforge.net/p/lemur/wiki/RankLib/

で紹介する 2 つのモデルのランキングタスクにおける性能を比較した．我々はベースラインとしてドメインの差異を考慮せず大量のデータを用いて学習した**汎用モデル**を利用した．汎用モデルは，学習用の 350 個のソースドメインのデータと検証用の 75 個のターゲットドメインのデータをそれぞれ 1 つにまとめることで学習を行った．提案モデルと汎用モデルが利用するデータの量や種類は同じであるため，性能を左右する大きな要素はドメインの差異を考慮したか否かである．したがって，汎用モデルは提案手法の有効性を示すためのベースラインとしてふさわしい．また，評価用のターゲットドメインで直接学習したモデルを理想的な振る舞いをする**理想モデル**として利用した．この実験では，ドメイン特徴として，データセットの分割に用いた特徴セットと同じものを利用した．

最後に，ドメイン特徴として何が有効かを判別するために，ドメイン特徴として利用する特徴セットを変更した際の重み予測モデルの性能の変化も計測した．

ランキングタスクの評価指標には，一般的に利用される評価指標である nDCG@10，ERR@10，MAP を利用した．より具体的には，75 個の評価用のターゲットドメインにおける各評価データでモデルを評価した場合の各指標の平均値を用いて比較した．各指標の 75 ドメインにおける平均値を算出する際に，汎用モデルは全ての評価ドメインで同じモデルを利用しているが，提案モデルと理想モデルは各評価ドメインごとに特化したモデルを利用している点に注意されたい．

4.3 実験結果

4.3.1 データセットの分割に用いる特徴セットの比較

まず，データセットの分割に用いる特徴のセットを決定するために行った実験の結果を表 3 に示す．Q は表 2 におけるすべてのクエリ特徴，D はすべての文書特徴，Q-D はすべてのクエリ・文書特徴を意味し，複数の特徴セットを組み合わせて利用した場合は，&で示している．

表から分かるように，データセットの分割に用いる特徴セットとしてクエリ・文書特徴を利用した場合に最も重みの分散の和が大きくなっている．3.3 節での仮説に従えば，クエリ・文書特徴を利用した場合に最もドメインの性質が異なるようにデータセットを分割できたと言える．また，クエリ・文書特徴を用いない特徴セットを利用した場合，特徴を用いずランダムに分割した場合よりも重みの分散の和が小さくなっている．特徴を用いずランダムに分割した場合，分割後のドメイン間で性質の差が小さいと考えられるため，ランダムな分割を下回るこれらの特徴セットではうまくデータセットが分割できないと思われる．さらに，クエリ・文書特徴と他の特徴を組み合わせて利用した場合，クエリ・文書特徴のみを利用した場合よりも重みの分散の和が小さくなっている．このことから，クエリ・文書特徴と他の特徴を組み合わせるのは効果的でないと考えられる．先述の通りクエリ特徴や文書特徴がドメインを分割をする上で効果的な特徴でないと思われるため，クエリ・文書特徴と組み合わせるとクエリ・文書特徴を単体で利用したときよりも分割に対する効果が低下してしまうことが原因として考えられる．

表 3　分割に用いた特徴セットごとの重みの分散の和.

分割に用いた特徴セット	重みの分散の和
特徴を用いずランダムに分割	0.0658
Q	0.0620
D	0.0649
Q-D	**0.0710**
Q&D	0.0643
Q&Q-D	0.0674
D&Q-D	0.0682
Q&D&Q-D	0.0690
ランキング学習に用いた特徴	0.0705

最後に，ランキング学習に用いた特徴を利用した場合，文書特徴とクエリ・文書特徴を組み合わせて利用した場合よりも重みの分散の和が大きくなっており，全体でも 2 番目の大きさである事が分かる．このことから，ランキング学習に用いた特徴を利用した場合，文書特徴とクエリ・文書特徴を組み合わせて利用した場合よりも効果的な分割ができると考えられ，全体で比較してもドメインの分割に有効であると考えられる．ランキング学習に用いた特徴は，文書特徴のうち文書の本文に含まれるセンテンス数，本文のリーダビリティスコア，本文のトピックを除いたものに，クエリ・文書特徴を組み合わせたものである．このことから，文書特徴の中でも特にランキング学習で利用していない上記の特徴がドメインの分割に対する効果の低下を引き起こしていると予想される．

以降の実験ではドメインの性質が最も異なるようにデータセットを分割できたと考えられるクエリ・文書特徴を用いた際の分割を採用して分析を行った．

4.3.2 モデル間の性能比較

次に，モデル間の性能を比較するために行った実験結果を表 4 に示す．この結果から，提案手法はドメインの差異を考慮せず学習した汎用モデルよりもすべての指標で高い性能を示した事が分かる．したがって，ターゲットドメインの適合性判定データが利用できない場合に，ドメインの差異を考慮して重みを算出することは効果的であると考えられる．また，提案手法はターゲットドメインで直接学習した理想モデルよりもすべての指標で高い性能を示した．このことから，提案手法は理想モデルよりもランキングタスクにおける性能が高いと考えられる．各テストドメインのモデルの重みを予測する際に，理想モデルは各テストドメインでテストドメイン 1 個分のデータしか活用していないのに対し，提案手法は学習用のソースドメイン 350 個分のデータを活用しており，この活用したデータ数の差が結果に影響したのではないかと予想する．

手法間の性能の差異を統計的に評価するために，nDCG@10 のスコアを対象に二元配置分散分析およびテューキーの範囲検定で検定を行った．二元配置分散分析により，3 つのモデル間に有意差があることが明らかになった（$F(2, 28886) = 59966, p < 0.05$）．その上で，テューキーの範囲検定では，すべてのモデルのペア間に統計的に有意な差が認められた（$p < 0.05$）．

表 4　各モデルの性能比較を行った実験結果.

モデル	nDCG@10	MAP	ERR@10
提案モデル	**0.823**	**0.773**	**0.454**
理想モデル	0.817	0.762	0.449
汎用モデル	0.810	0.759	0.447

表 5　ドメイン特徴を変更した際の提案モデルのスコア.

ドメイン特徴	nDCG@10	MAP	ERR@10
Q	0.817	0.762	0.449
D	0.817	0.762	0.449
Q-D	**0.823**	**0.773**	**0.454**
Q+D	0.820	0.770	0.453
Q+Q-D	0.822	0.771	0.453
D+Q-D	0.821	0.771	0.453
Q+D+Q-D	0.821	0.770	0.452
ランキング学習に用いた特徴	0.822	0.771	0.453

4.3.3　ドメイン特徴として用いる特徴セットの比較

最後に，ドメイン特徴として用いる特徴セットを変更した際の提案手法の性能の変化を表 5 に記す．この結果から，クエリ・文書特徴のみをドメイン特徴として利用したモデルがすべての指標で最も高い性能を示した事が分かる．このことから，クエリ・文書特徴がドメイン特徴として最も効果的であり，最もドメインの特徴を捉えることができていると考えられる．しかし，この結果はドメインの分割に用いる特徴とドメイン特徴が一致していることが影響している可能性もあり，一般的な結果ではないとも考えられる．また，表からは 4.3.1 節の実験結果と同じように，クエリ・文書特徴を用いない特徴セットは他の特徴セットと比べて性能が低いことが読み取れる．更に，クエリ・文書特徴と他の特徴を組み合わせて利用した場合，クエリ・文書特徴のみを利用した場合よりも性能が低下していることも同様である．このことから，ドメインの分割に寄与する特徴とドメイン特徴として活用した際に有効な特徴には似たような傾向があり，クエリ・文書特徴が特に有効であると示唆される．ただし，クエリ・文書特徴を利用している特徴セット間での性能差はドメインを分割する場合よりも小さく，ほとんど数値的な差がない．このことから，ドメイン特徴はデータセットの分割と比べて，あまり効果的でない特徴を追加することから受ける悪影響が小さいと考えられる．また，表 4 と比較すると，どのドメイン特徴を利用した場合であっても，すべての指標で汎用モデルを上回っていることが分かる．このことから，効果の差はあれど，いずれのドメイン特徴もドメインの性質をある程度捉えることが可能であり，重み予測モデルの入力として有効であると考えられる．

手法間の性能の差異を統計的に評価するために，nDCG@10 のスコアを対象に二元配置分散分析およびテューキーの範囲検定で検定を行った．二元配置分散分析により，手法間に有意差が有ることが明らかになった（$F(7, 77031) = 157544, p < 0.05$）．その上で，テューキーの範囲検定では，クエリ・文書特徴を用いた手法とクエリ特徴を用いた手法，ランキング学習に用いた特徴を用いた手法とクエリ特徴を用いた手法の間に統計的に有意な差が認められた（$p < 0.05$）．

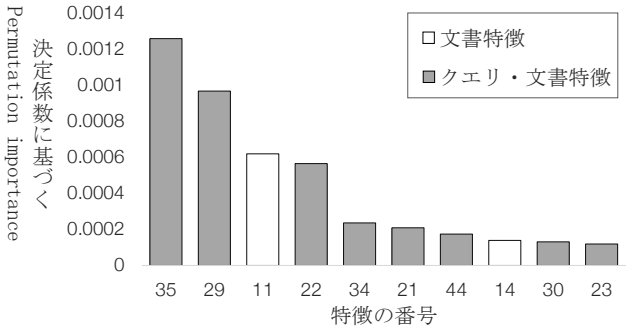


図 2　各ドメイン特徴に対する決定係数に基づく Permutation importance．特徴の番号は表 2 のものを利用している．

4.3.4　個別の特徴の分析

最後に，個別のドメイン特徴のうちどの特徴が有効であるかを分析した．上記の分析を行うために，表 2 のすべての特徴をドメイン特徴として利用した際の，決定係数に基づく Permutation importance を計測した．決定係数は回帰タスクにおける評価指標である．およそ 0 から 1 までの値を取り，この値が大きいとモデルがうまく予測を行える傾向にあることを示す．また，Permutation importance は入力データにおける特定の特徴の値をランダムに並び替えることによって，その特徴がモデルの予測性能にどの程度影響を与えているかを評価する指標である．決定係数に基づく Permutation importance は，重み予測モデルを用いて本来のデータで予測した場合の決定係数と特定の特徴の値を並び替えて予測した場合の決定係数の差を取ることで算出する．Permutation importance が高い特徴ほど値をランダムに並べ替えると予測性能が悪化すること，すなわち，予測において重要な役割を果たしていることを示す．

実験結果を図 2 に示す．ただし，重要度が高かった上位 10 件の結果のみを取り上げる．実験結果から，35 番目の特徴の Permutation Importance が最も高いことが分かる．このことから，35 番目の特徴は重み予測モデルにとって重要な特徴であると言える．35 番目の特徴は文書のタイトルにおける TF-IDF であり，この値をドメインごとにまとめた値に対する重みの重要度が高いことから，ドメインによってクエリと合致するタイトルを持つ文書の多寡が大きく異なる可能性が示唆される．Song らは TF-IDF と同じくクエリと文書の単語の一致度に基づいたスコアである BM25 の値をクエリごとにまとめた値の傾向がドメインによって異なると報告しており [26]，この傾向が本実験でも見られた．その他の結果としては，29 番目の特徴である文書のタイトルにおける IDF の対数の Permutation Importance が 2 番目に高く，ドメインによって入力するクエリの出現頻度（珍しさ）の傾向が大きく異なる可能性がある．また，11 番目の特徴である文書のタイトルに含まれるトークン数の対数の Permutation Importance が 3 番目に高く，ドメインによって文書のタイトルの長さが大きく異なる可能性がある．

最後に，図 2 から重要度の高い特徴にはクエリ・文書特徴が多いことが分かる．本研究では 4.3.1 節に記した通りクエリ・文書特徴の特徴セットに基づいてデータセットを分割することでドメインを作成しており，このことが結果に影響した可能性があると考えられる．

5 ま と め

本論文では，ターゲットドメインの適合性判定データを用いずに転移学習を行うトランスダクティブ転移学習という設定においてランキング学習タスクに取り組み，このタスクにおけるランキング学習モデルのドメイン適応手法を提案した．この手法では，ソースドメインとターゲットドメインで共通して利用できる要素からドメイン特徴を抽出し，ドメイン特徴からターゲットドメインにおける最適な重みを予測することでターゲットドメインにおいて性能の良いランキングモデルを構築することができると考えた．重みの予測は重み予測モデルを学習することで行い，学習に必要な複数のドメインは多種多様なデータを含むデータセットを分割することで用意する方法を採用した．実験では，データセットの分割に用いる特徴としてクエリ・文書特徴が最も効果的であることが判明した．クエリ・文書特徴を用いてデータセットを分割することで得たドメインで提案手法の性能評価をした結果，ドメインの差異を考慮せずに学習した汎用的なモデルよりも提案手法を用いたモデルの性能が高くなることが確認された．また，提案手法が利用するドメイン特徴としても，クエリ・文書特徴が最も効果的であることが明らかになった．今後の課題としては，よりパラメータが複雑なランキングモデルのパラメータを予測したり，より現実に即したドメインで評価を行うなどして提案手法の有効性を確認する必要があると考える．

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

文 献

[1] Olivier Chapelle, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng. Boosted multi-task learning. *Mach. Learn.*, 85(1-2):149–173, 2011.

[2] Depin Chen, Yan Xiong, Jun Yan, Gui-Rong Xue, Gang Wang, and Zheng Chen. Knowledge transfer for cross domain learning to rank. *Inf. Retr. Boston.*, 13(3):236–253, 2010.

[3] Jianfeng Gao, Qiang Wu, Chris Burges, Krysta Svore, Yi Su, Nazan Khan, Shalin Shah, and Hongyan Zhou. Model adaptation via model interpolation and boosting for web search ranking. In *ENMLP*, pages 505–513, 2009.

[4] Wei Gao, Peng Cai, Kam-Fai Wong, and Aoying Zhou. Learning to rank only using training data from related domain. In *SIGIR*, pages 162–169, 2010.

[5] Xiubo Geng, Tie-Yan Liu, Tao Qin, Andrew Arnold, Hang Li, and Heung-Yeung Shum. Query dependent ranking using k-nearest neighbor. In *SIGIR*, pages 115–122, 2008.

[6] Qian Guo, Wei Chen, and Huaiyu Wan. AOL4PS: A Large-scale Data Set for Personalized Search. *Data Intelligence*, 3(4):548–567, 2021.

[7] Malay Haldar, Mustafa Abdool, Liwei He, Dillon Davis, Huiji Gao, and Sanjeev Katariya. Learning to rank diversely at airbnb. In *CIKM*, pages 4609–4615, 2023.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[9] In-Ho Kang and Gilchang Kim. Query type classification for web document retrieval. In *SIGIR*, pages 64–71, 2003.

[10] Makoto P. Kato and Takehiro Yamamoto. Overview of the ntcir-13 openliveq task. In *NTCIR Conference on Evaluation of Information Access Technologies*, 2017.

[11] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, 2009.

[12] Bo Long, Sudarshan Lamkhede, Srinivas Vadrevu, Ya Zhang, and Belle Tseng. A risk minimization framework for domain adaptation. In *CIKM*, pages 1347–1356, 2009.

[13] Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Reproducing personalised session search over the AOL query log. In *ECIR*, pages 627–640, 2022.

[14] Craig Macdonald, B Taner Dinçer, and Iadh Ounis. Transferring learning to rank models for web search. In *ICTIR*, pages 41–50, 2015.

[15] Craig Macdonald, Rodrygo L T Santos, and Iadh Ounis. On the usefulness of query features for learning to rank. In *CIKM*, pages 2559–2562, 2012.

[16] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

[17] Weijian Ni, Yalou Huang, and Maoqiang Xie. A query dependent approach to learning to rank for information retrieval. In *WAIM*, pages 262–269, 2008.

[18] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[19] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, page 1–es, 2006.

[20] Rama Kumar Pasumarthi, Sebastian Bruch, Xuanhui Wang, Cheng Li, Michael Bendersky, Marc Najork, Jan Pfeifer, Nadav Golbandi, Rohan Anil, and Stephan Wolf. TF-Ranking: Scalable TensorFlow library for Learning-to-Rank. In *KDD*, pages 2970–2978, 2019.

[21] Jie Peng, Craig Macdonald, and Iadh Ounis. Learning to select a ranking function. In *Advances in Information Retrieval*, pages 114–126, 2010.

[22] Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013.

[23] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr. Boston.*, 13(4):346–374, 2010.

[24] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

[25] RJ Senter and Edgar A Smith. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–11, 1967.

[26] Ruihua Song, Ji-Rong Wen, Shuming Shi, Guomao Xin, Tie-Yan Liu, Tao Qin, Xin Zheng, Jiyu Zhang, Gui-Rong Xue, and Wei-Ying Ma. Microsoft research asia at web track and terabyte track of TREC 2004. *Text Retrieval Conference*, 2004.

[27] A Trotman, Jon Degenhardt, and S Kallumadi. The architecture of ebay search. *eCOM@SIGIR*, 2017.

DEIM Forum 2024 T3-B-1-05

List-wise 手法における順位に基づいた多段階分割と学習データの上位遷移による能動学習

エルゲン瑛夏† 鈴木 優††

† 岐阜大学大学院 〒 501–1193 岐阜県岐阜市柳戸 1 番 1
†† 岐阜大学 〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †erugen.ege.j6@s.thers.ac.jp, ††suzuki.yu.r4@f.gifu-u.ac.jp

あらまし 順序推定に用いられるランキング学習の主な 3 手法の中で，List-wise 手法は最も推定精度の高い手法として知られている．しかし，List-wise 手法では学習時に対象データの順列を用いて損失関数の計算を行う必要があり，学習時間が膨大になってしまうという問題がある．この問題を解決するために, 我々は能動学習によって使用する順序データを選定する方法が有効であると考えた．提案手法では, 付与された順位ラベルに基づいて対象データを多段階分割し，分割結果に基づいて順序データを作成する．また，モデルが一定の学習進度に達した際に対象データの上位一部を抽出し，新たな対象データとする．この二つの操作により使用順序データの選定を行い，能動学習を行った．提案手法の有効性を確認するために実験を行った. 実験の結果，提案手法を用いて Top-k 学習を行った場合に, 学習効率を向上させられることが分かった. また, 提案手法においてデータの抽出割合が小さい場合に, 順序推定精度を高くなることが分かった.

キーワード ランキング学習, 能動学習,List-wise 手法, 機械学習

1 はじめに

我々は検索エンジン利用者の満足度を向上させるため，検索結果の表示順序を利用者にとって最適なものにしたいと考えた．大量にある Web ページの有益度合いを人手によって評価することは困難である．そこで我々は機械学習を用いて順序推定を行うことを考えた．本研究では，順序推定に適した機械学習手法であるランキング学習 [1] を用いる．ランキング学習とは教師あり学習の一種であり，対象データの特徴量を入力した際にそのスコアを出力するランキングモデルを作成する学習手法である．また，ランキング学習では算出するスコアの絶対的な値ではなく，並び替えた際の相対的な順序関係が正しくなるようにランキングモデルが学習される．

ランキング学習は，損失関数の定義によって主に Point-wise 手法，Pair-wise 手法，List-wise 手法の三つの手法が存在する．Point-wise 手法では，回帰学習と同じ損失関数を用いて学習を行う．対象データそれぞれについて，ランキングモデルによって算出された予測スコアと正解スコア間の誤差を小さくすることを目的に学習を行う．この際，予測スコアと正解スコア間の誤差が小さい場合でも，予測スコアの順序と正解スコアの順序間の誤差が大きい場合がある．そのため，Point-wise 手法では各データの対象データ中の順位については十分に学習できない．List-wise 手法 [2] では，損失関数計算時に対象データの順列を用いて学習を行う．そのため，対象データ全体の順序推定を最適化するようにランキングモデルを学習することができ，3 手法の中で最も精度が高くなることが期待できる．List-wise 手法の中でも，対象データの順列中に存在する各順序について，上位 2 位までの推定精度のみに着目して学習を行う手法は Pair-wise 手法とも呼ばれる．

本研究ではこれら三つの手法のうち，List-wise 手法に着目した．なぜなら，検索エンジン利用者の満足度を向上させるという目的を達成するためには順序推定精度が高い手法を用いることが良いと考えたためである．ところが List-wise 手法で問題となる点は，学習時間が膨大な点である．List-wise 手法では対象データの順列中の各順序全てを用いて学習を行うため，対象データ数の増加に伴い，学習に用いる順序数が指数的に増えてしまうという問題がある．この問題を解決するために，Cao ら [2] は Top-1 学習を提案した．Top-1 学習とは対象データの順列中に存在する各順序について上位 k 位までの順序のみを考慮する Top-k 学習を，k の値を 1 として行う学習手法である．また，Cao ら [2] は k の値が大きくなることによって計算が複雑になることを懸念し，k の値を 1 のみで Top-k 学習を行った．

そこで本研究では順序推定精度の向上を期待し，Top-k 学習において k の値が 1 より大きい場合を考える．我々はこのときに膨大となる学習時間を，モデルによる順序推定精度を維持しながら削減することを研究の目標とする．我々はこの目標を実現するために能動学習を用いて学習効率を高めることを考えた．能動学習とはアルゴリズムによってモデルの学習に有効なデータを選定し，モデルの学習効率を高める手法である．

我々は Top-k 学習の問題は二つあると考える．一つ目は，難易度の高い学習を学習の初期段階に行うことである．Top-k 学習では対象データ中の k データの順列を考えるため，僅差の順位を持つデータ同士の順序を学習に用いることがある．しかし，我々は僅差の順位を持つデータ同士の順序予測を学習の初期段階に行うことは難易度が高いと考えた．そこで我々は学習に用

いる順序データの難易度を徐々に高めることによって学習効率の向上が期待できると考えた．二つ目は対象データ中の下位データと上位データで同じ時間学習を行うことである．ランキング学習では上位データの順序推定精度を高めることを目的とする．しかし，我々は Top-k 学習では下位データを上位データと同じ時間学習しているため，学習効率が低下していると考えた．そこで我々は対象データから下位データを取り除くことを考えた．この手法による学習データの削減によって，上位データの順序推定精度を保ちながら学習時間を削減できると考えた．

本稿では Top-k 学習において，以下の二つの操作を組み合わせて能動学習を行う手法を提案する．一つ目は対象データを順位に沿って k 段階分割した後に，分割後のそれぞれのグループから一つずつデータを選び，選ばれた k データの全順序を計算データとする操作である．この際，選び方は全通りで行う．二つ目は対象データについてモデルの損失関数の値が 30 エポックの間に減少しなかった際に，対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出し新たな対象データとする操作である．D を対象データ数とする．

提案手法の一つ目の操作によって，僅差の順位を持つデータ同士の順序を学習の初期段階に用いることが減り，Top-k 学習の問題の一つ目を解決できると考えた．また，提案手法の二つ目の操作によって対象データから下位データを取り除くことができ，Top-k 学習の問題の二つ目を解決できると考えた．また，対象データの更新と k 段階分割による学習データの選定を学習過程で繰り返し行うことで，学習に用いる順序データ中に含まれるデータの順位幅が徐々に縮まり，学習難易度の段階的な上昇を実現できると考えた．

提案手法を評価するために三つの評価実験を行った．一つ目は提案手法の有効性確認を目的として行った．通常の Top-k 学習と提案手法を用いた Top-k 学習で順序推定精度と学習時間の比較を行った．二つ目は提案手法における上位データ抽出の有効性確認を目的として行った．提案手法でデータを抽出する際に，上位から抽出する場合とランダムに抽出する場合で Top-k 学習を行い，順序推定精度の比較を行った．三つ目は提案手法における最適なデータ抽出割合の発見を目的として行った．提案手法でデータを抽出する際に，抽出割合を 10%,20%,30%,40%,50%それぞれの場合で Top-k 学習を行い，順序推定精度の比較を行った．

評価実験の結果，提案手法を用いて Top-k 学習を行うことによって，通常の Top-k 学習を行った場合から順序推定精度を有意に下げることなく学習時間を削減できることが分かった．また，提案手法でデータを抽出する際に，k の値が 3 の場合に，対象データの上位から抽出することによる順序推定精度の改善が確認できた．さらに，提案手法でデータを抽出する際の抽出割合は小さい方が順序推定精度を高められることが確認できた．

本稿による貢献は以下の通りである．

- Top-k 学習において，対象データの多段階分割に基づいた計算データ選定と，対象データの上位抽出による更新を組み合わせた能動学習が，学習効率の向上に有効であることが分かった．
- 対象データの更新時の抽出割合が大きい場合よりも，小さい場合の方が高い精度で順序推定を行えることが分かった．

2 関連研究

List-wise 学習についての研究はすでにいくつかある．Cao ら [2] は List-wise 学習を提案し，Point-wise 手法，Pair-wise 手法と比較して優れた手法であることを示した．Rahimi ら [3] は，ニューラルランキングモデルの学習に List-wise 損失を採用することを提案した．先行研究 ([2], [3]) では計算量が膨大になるのを避けるため，Top-1 学習を行っている．本研究では順序推定精度の向上を期待し，k の値を 1 より大きくして学習を行う．その際に膨大となる計算量を，モデルによる推定精度を維持しながら削減することを目標とする．

上位の予測精度を重視したランキング学習についての研究もある．Liang ら [4] は，ランキング学習において下位データを学習に加えることは上位データの推定精度低下に繋がると考え，実用的でないとした．そこで Liang らはランク付けされたリストの上位 N 個のみを考慮して損失計算を行い，下位データの影響を排除した．これによって精度の向上が確認された．本稿では，対象データを段階的に上位に遷移させることを考える．これによって学習から下位データを段階的に取り除くことができ，汎化性能が向上すると考えた．

また，学習時にタスクの難易度を徐々に高めることについての研究もある．Bengio ら [5] は，学習時のデータがランダムに提示されるのではなく，徐々に多くの概念や，複雑な概念を示すような順序で提示することが学習効率を高めることを示した．我々は，対象データの順位に沿った多段階分割による順序データの作成と，対象データの上位遷移を繰り返すことによって，段階的な難易度の上昇を実現できると考えた．

能動学習についての研究もある．Li ら [6] は，能動学習時にランキング学習の一手法である List-wise 学習を用いることによって学習効率が向上することを示した．

List-wise 学習の先行研究では Top-1 学習を行っているものが多い．また，能動学習についての研究ではデータセット作成時のアノテーションコスト削減を目的としており，データ選定時に正解ラベルを使用しないものが多い．

本研究では Top-k 学習における k の値を 1 より大きく設定して学習を行う．その際に膨大になる計算時間を，正解順位情報を用いた能動学習を行うことにより, 順序推定精度を維持しながら削減することを目指す．

3 提案手法

学習対象とする文書データ群を対象データ，対象データの順列中で損失計算に用いる順序を計算データとする．また，List-wise 学習において，対象データの順列中に存在する各順序中の上位 k 位までの順序を計算データとして用いる学習手法を Top-k 学習とする．

我々は順序推定精度の向上を期待し，Top-k 学習において k が 1 より大きい場合に着目した．この際に膨大になる計算デー

タから，学習に有効なデータのみを選定し使用することによって，少ない計算データで高い性能のモデルを構築することを目指す．

我々は従来の Top-k 学習の問題は二つあると考えた．一つ目は，対象データ中の上位データについての予測精度向上を目的とするランキング学習において，下位データを上位データと同じ時間学習していることだ．我々はこの問題を解決するために，下位データの学習時間を減らす手法があると良いと考えた．二つ目は，学習の初期段階において，僅差の順位を持つデータ同士の順序のような難易度の高い学習を行っていることである．我々はこの問題を解決するために，学習時に与える計算データの難易度を徐々に高める必要があると考えた．

本稿では対象データを順位に沿って k 段階分割した後に，分割後のそれぞれのグループから一つずつデータを選び，選ばれた k データの全順序を計算データとする．この際，選び方は全通りで行う．我々はこの操作によって，学習の初期段階において僅差の順位を持つデータ同士の順序を学習しにくくなると考えた．また，モデルが一定の学習進度に達した際に，対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出して新たな対象データとする．我々はこの二つの操作を行い学習データを選定することによって，学習時の計算データの難易度を段階的に上昇させることに加え，上位データを重視して学習することができるため, 学習効率の向上が期待できると考えた．

提案手法の概要を図 1 に示す．以下に提案手法を用いた学習の手順を示す．

(1) 対象データを順位に沿って k 段階分割する．

(2) 計算データを (2a)，(2b) で作成する．

（a） 各ミニバッチ中に (1) の段階毎のデータ数が均等に入るようにミニバッチを作成する．

（b） (2a) のミニバッチ中で，異なる段階から一つずつデータを選び，選ばれた k データの全順序を計算データとする．この際，選び方は全通りで行う．

(3) (2b) のデータを用いてモデルの損失関数の値が 30 エポックの間に減少しなくなるまで Top-k 学習を行う．

(4) 対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出して新たな対象データとする．

(5) (1)〜(4) を，対象データ数がバッチサイズを下回るまで繰り返す．

3.1 ランキング学習

ランキング学習とは，検索エンジンによる検索結果の表示順序を最適化することを目的に提案された学習手法である．絶対的な推定スコアや，多段階ラベルの推定精度を高めるためではなく，スコアや多段階ラベルの相対的な順序の推定精度を高めることを目標に学習する手法である．

ランキング学習は損失関数の計算方法によって 3 手法に分類される．本稿ではその中で最も精度が良いとされている List-wise

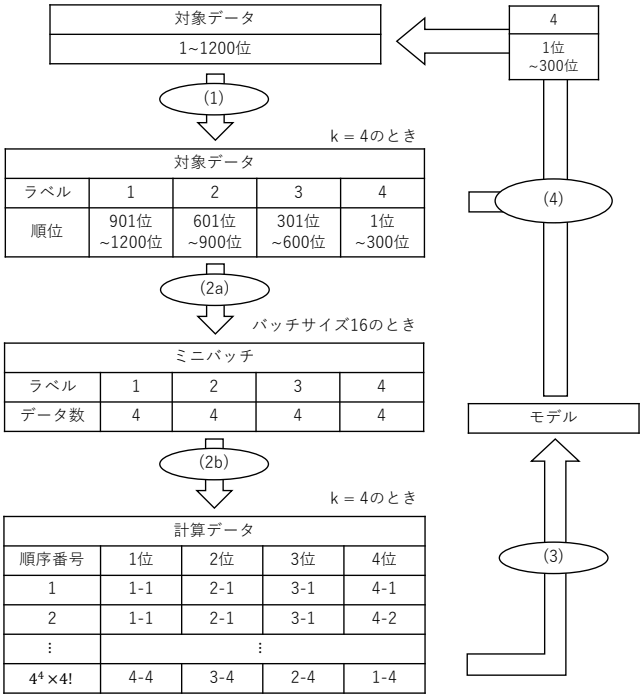


図 1 提案手法

手法を用いてモデルの学習を行った．List-wise 手法とは対象データの順列を損失計算に用いることによって，対象データ全体の順序推定を最適化するように学習する手法である．

学習に用いるデータは，検索ワードであるクエリと，クエリとの関連度に基づいた順位もつデータで構成されている．Cao ら [2] は損失関数を (1) 式のように定義しており，Top-1 学習を行っている．(1) 式では，正解順序発生確率 $p_{y^{(i)}}(x_j^{(i)})$ と，予測順序発生確率 $p_{z^{(i)}(f_\omega)}(x_j^{(i)})$ のクロスエントロピー誤差を計算している．正解順序発生確率 $p_{y^{(i)}}(x_j^{(i)})$ は i 番目のクエリ $q^{(i)}$ の正解順位リスト $y^{(i)}$ 中で j 番目のデータ $x_j^{(i)}$ が 1 位となる順序発生確率である．予測順序発生確率 $p_{z^{(i)}(f_\omega)}(x_j^{(i)})$ はニューラルネットワークモデル ω によるランキングモデル f_ω によって算出された $q^{(i)}$ 中のデータの予測スコアリスト $z^{(i)}(f_\omega)$ 中で，$x_j^{(i)}$ が 1 位となる順序発生確率である．また，$x_j^{(i)}$ がスコア $s(j)$ を持つ時，$x_j^{(i)}$ が 1 位となる順序発生確率を (2) 式のように定義している．(2) 式では対象データのスコアの総和で s_j を割ることで，$x_j^{(i)}$ が 1 位に選ばれる確率を算出している．

$$L(y^{(i)}, z^{(i)}(f_\omega)) = -\sum_{j=1}^{n^{(i)}} p_{y^{(i)}}(x_j^{(i)}) log(p_{z^{(i)}(f_\omega)}(x_j^{(i)})) \quad (1)$$

$$P_s(j) = \frac{exp(s_j)}{\sum_{m=1}^{n^{(i)}} exp(s_m)} \quad (2)$$

Top-1 学習の利点として，計算データが少なく学習時間が短いことがある．$q^{(i)}$ 中に $(x_1^{(i)},\ x_2^{(i)},\ x_3^{(i)})$ の 3 データがある場合を例に挙げる．このとき，$q^{(i)}$ 中の順列は $[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]$, $[x_1^{(i)}, x_3^{(i)}, x_2^{(i)}]$, $[x_2^{(i)}, x_1^{(i)}, x_3^{(i)}]$, $[x_2^{(i)}, x_3^{(i)}, x_1^{(i)}]$, $[x_3^{(i)}, x_1^{(i)}, x_2^{(i)}]$, $[x_3^{(i)}, x_2^{(i)}, x_1^{(i)}]$ となる．$x_1^{(i)}$ が 1 位となる順序発生確率を求め

るには，順序 $[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]$ と順序 $[x_1^{(i)}, x_3^{(i)}, x_2^{(i)}]$ の発生確率の和を取る必要がある．そのため，$x_1^{(i)}$，$x_2^{(i)}$，$x_3^{(i)}$ それぞれが 1 位になる確率を求めるためには順列全てを計算に用いる必要があり，計算データの総数は対象データ数の階乗となる．ここで，順序 $[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]$ と $[x_1^{(i)}, x_3^{(i)}, x_2^{(i)}]$ の発生確率の和は $x_1^{(i)}$ が 1 位になる順序発生確率と等しいため，対象データ中で $x_1^{(i)}$ が 1 位になる順序発生確率のみから算出可能である．そのため Top-1 学習では対象データの順列全てではなく，対象データ中の 1 データの順列のみを計算データとして用いる．対象データ数を D とすると，Top-1 学習の計算データ数は ${}_DP_1$ となる．例では，($x_1^{(i)}$，$x_2^{(i)}$，$x_3^{(i)}$) それぞれが 1 位になる順序発生確率を計算するのみで良いため，計算データの総数は 3 になる．

本稿では Top-k 学習を k が 1 より大きい場合で行うため，損失関数は (3) 式を用いる．(1) 式からの差異として，j の範囲を ${}_n^{(i)}P_k$ に変更し，$x_j^{(i)}$ を $\pi_j^{(i)}$ に変更した．ここで ${}_n^{(i)}P_k$ は Top-k 学習における $q^{(i)}$ 中の計算データの総数であり，$\pi_j^{(i)}$ は $q^{(i)}$ 中に含まれるデータの順列に含まれる j 番目の順序データである．また，順序発生確率は (4) 式を用いて計算した．$q^{(i)}$ 中に ($x_1^{(i)}$,$x_2^{(i)}$,$x_3^{(i)}$) の 3 データがある場合を例に挙げる．このとき，Top-3 学習に用いる $q^{(i)}$ 中の全計算データは $[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]$, $[x_1^{(i)}, x_3^{(i)}, x_2^{(i)}]$, $[x_2^{(i)}, x_1^{(i)}, x_3^{(i)}]$, $[x_2^{(i)}, x_3^{(i)}, x_1^{(i)}]$, $[x_3^{(i)}, x_1^{(i)}, x_2^{(i)}]$, $[x_3^{(i)}, x_2^{(i)}, x_1^{(i)}]$ となる．($x_1^{(i)}$,$x_2^{(i)}$,$x_3^{(i)}$) がスコア (s_1，s_2，s_3) を持つ場合を挙げる．このとき，順序 $\pi = [x_1^{(i)}, x_2^{(i)}, x_3^{(i)}]$ の順序発生確率 $P_s(\pi)$ は (5) 式のように計算する．

$$L(y^{(i)}, z^{(i)}(f_\omega)) = -\sum_{j=1}^{{}_n^{(i)}P_k} p_{y^{(i)}}(\pi_j^{(i)}) log(p_{z^{(i)}(f_\omega)}(\pi_j^{(i)})) \tag{3}$$

$$P_s(\pi) = \prod_{l=1}^{n} \frac{exp(s_{\pi(l)})}{\sum_{m=l}^{n^{(i)}} exp(s_{\pi(m)})} \tag{4}$$

$$P_s(\pi) = \frac{exp(s_1)}{exp(s_1)+exp(s_2)+exp(s_3)} \frac{exp(s_2)}{exp(s_2)+exp(s_3)} \frac{exp(s_3)}{exp(s_3)} \tag{5}$$

3.2 学習データの選定

学習データの選定では，対象データを順位に沿って k 段階分割した後に，分割後のそれぞれのグループから一つずつデータを選び，選ばれた k データの全順序を計算データとする．この際, 選び方は全通りで行う．また，この計算データを用いて学習したモデルの損失関数の値が 30 エポックの間に減少しなくなった際に，対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出し新たな対象データとする．以下に学習データ選定の手順を示す.

(1) 対象データを順位に沿って k 段階分割する.

(2) 計算データを以下の (a)，(b) によって作成する.

（a） ミニバッチ中に (1) の段階毎のデータ数が均等になるようにミニバッチを作成する.

（b） (2a) のミニバッチ中で，異なる段階から一つずつデータを選び，選ばれた k データの全順序を計算データとする．この際，選び方は全通りで行う.

(3) (2b) の計算データを用いてモデルが一定の学習進度に達するまで Top-k 学習を行う.

(4) 対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出して新たな対象データとする.

(5) (1)〜(4) を，対象データ数がバッチサイズを下回るまで繰り返す.

3.2.1 対象データの k 段階分割

本節では対象データを順位ラベルに沿って k 段階分割する操作について説明を行う．この操作を図 1 の (1) に示す.

図 1 では 1 位から 1200 位までの順位ラベルが付与されている対象データを 4 段階分割する場合を例に挙げる．はじめに対象データ全体を順位に沿って 4 等分する．次に分割後のデータ群毎にラベルを付与する．1 位から 300 位に 4，301 位から 600 位に 3，601 位から 900 位に 2，901 位から 1200 位に 1 のラベルを付与する．k 段階分割後のラベル毎のデータ数 L は，対象データ数 D と k を用いて (6) 式で算出する.

$$L = \frac{D}{k} \tag{6}$$

3.2.2 計算データの作成

本節では計算データ作成をする二つの手順 a),b) について説明を行う．手順 a) の操作を図 1 の (2a) に示し，手順 b) の操作を (2b) に示す.

a） ミニバッチの作成

手順 a) ではミニバッチの作成を行う．図 1 ではバッチサイズが 16 の場合を例に挙げる．ミニバッチは 3.2.1 節で説明した方法によって付与されたそれぞれのラベルが均等な数ずつ入るように作成する．一つのミニバッチ中について，各ラベルのデータ数 BL はミニバッチ中のデータ数 B と k を用いて (7) 式で算出する.

$$BL = \frac{B}{k} \tag{7}$$

b） 計算データの作成

手順 b) では計算データの作成を行う．図 1 では $k = 4$ の場合を例に挙げる．提案手法ではミニバッチ中のデータの順列から，同じラベルが上位 k 位までに含まれない順序のみを計算データに使用する．また，Top-k 学習を行うためには k 位までの順序のみが分かれば良い．そこで，ミニバッチ中で同じラベルを含まない k 個のデータから順列を作成して計算データとする.

はじめに，a) で作成したミニバッチ中からラベルの重複がない k 個のデータの全組み合わせを作成する．一つのミニバッチ中について作成するデータの組み合わせ数 BC は BL と k を

用いて (8) 式で算出する．

$$BC = BL^k \tag{8}$$

次に，作成した各組み合わせ中でデータの総順序を作成し，計算データとする．これによって一つのミニバッチ中について作成される計算データ数 C は BC と k を用いて (9) 式で算出する．

$$C = (BC)k! \tag{9}$$

ここで，通常の Top-k 学習で一つのミニバッチ中について作成する計算データ数 NC は B と k を用いて (10) 式で算出する．

$$NC =_B P_k \tag{10}$$

(9) 式と (10) 式を比較すると，k が 1 より大きい場合に C の方が計算データ数が少なくなっている．

また，我々はこの方法で計算データを作成することによって，学習の初期段階において計算データの学習難易度を下げることができると考える．計算データ中の各順序データに同一ラベルが含まれることを防ぐことができるため，僅差の順位を持つデータ同士の順序を損失計算に用いにくくなるためである．

3.2.3 Top-k 学習

本節では Top-k 学習を 3.2.2 節で作成した計算データを用いてする操作について説明を行う．この操作を図 1 の (3) に示す．

実装はニューラルネットワークを用いて行った．1 層目に入力 46 次元，出力 10 次元の全結合層を設定し，活性化関数にシグモイド関数を設定した．2 層目に入力 10 次元，出力 1 次元の全結合層を設定した．損失関数には (3) 式を用いて学習を行った．

3.2.4 対象データの更新

本節では対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出して新たな対象データとすることによる対象データの更新操作について説明を行う．この操作を図 1 の (4) に示す．

図 1 では例として $k = 4$ の場合を挙げる．モデルが学習の終了条件を満たした際に，対象データから上位の $\frac{1}{4}$ データを抽出し新たな対象データとする．学習の終了条件は，モデルによる検証データの予測について，損失関数の値が 30 エポックの間に減少しなかった場合とする．

ランキング学習は上位の順序推定精度を高めることを目的とする．そこで，我々は対象データから下位データを取り除くことによって，上位の推定精度を保ちながら学習時間を削減できると考えた．そのために本稿では，対象データについてモデルが一定の学習進度に達した時点で，対象データの上位を抽出して新たな対象データとすることによって，下位データを対象データから取り除くを考えた．

3.2.5 ループ処理

本節では 3.2.1 節から 3.2.4 節までで説明を行った一連の操作を繰り返し行う操作について説明を行う．繰り返しの操作は 3.2.4 節で更新した対象データ数がバッチサイズを下回るまで繰り返す．ループ移行後は 3.2.4 節で更新した対象データについて 3.2.1 節から 3.2.4 節までの一連の操作を行う．また，ループ移行後は前ループにおいて 3.2.4 節で学習終了条件を満たした際に，検証データについての予測の損失関数の値が最も小さかったモデルを追加で学習させる．

我々は 3.2.2 節の計算データ作成手法と 3.2.4 節の対象データ更新手法を繰り返し行うことによって，計算データの学習難易度を徐々に高めることができると考えた．対象データを上位に遷移させることによって，3.2.1 節による k 段階分割後のラベルごとの順位の幅が縮まる．これによって，僅差の順位を持つデータ同士の順序を計算データに用いるようになるためである．我々はこのように対象データの k 段階分割に基づいた順序データ選定と学習データの上位遷移を繰り返し行うことによって，学習時のタスク難易度を徐々に高めるカリキュラム学習を実現することができ，学習効率が向上すると考えた．

4 評価実験

以下の評価実験を行い提案手法による能動学習の有効性を確かめた．提案手法によって能動学習を行ったモデルにより順序推定を行った場合と，能動学習を行わなかったモデルによって順序推定を行った場合で順序推定精度と学習時間を比較した．また，提案手法における上位データ抽出の有効性を確かめるために，ランダムデータ抽出を行った場合と順序推定精度を比較した．また，提案手法における最適なデータ抽出割合を発見するために，抽出データの割合を変更して順序推定精度と学習時間を比較した．精度に有意な差があるか統計的検定によって示した．

実験 1　提案手法の有効性確認
通常の Top-k 学習と，提案手法を用いた Top-k 学習を行い，順序推定精度と学習時間の比較を行う．k の値が 2，3 の場合で比較を行う．

実験 2　提案手法における上位データ抽出の有効性確認
提案手法においてデータ抽出を行う際に，上位からの抽出とランダムによる抽出を行い，順序推定精度の比較を行う．k の値が 2，3，4 の場合で比較を行う．

実験 3　提案手法における最適なデータ抽出割合の発見
提案手法においてデータ抽出を行う際に，対象データに対する抽出するデータの割合を変更し，順序推定精度の比較を行う．k の値が 2,3 の場合で比較を行う．

本稿ではランキング学習に関する研究用のベンチマークデータセットである LETOR4.0 [7] を用いて評価実験を行った．LETOR4.0 はマイクロソフトから提供されているデータセットである．その中でも List-wise 学習用のデータセットである MQ2008-list を用いて実験を行った．MQ2008-list 中には複数の行があり，各行はクエリと文書のペアを示す．各行はクエリと文書の関連度，クエリ番号，46 次元の特徴量を持つ．また，

MQ2008-list では同一文書でもクエリ毎に関連度が異なる．異なる関連度を持った同一の文書を学習に用いることが学習の妨げになることを懸念し，一つのクエリのみを対象に実験を行った．対象としたクエリに含まれる 1330 データを評価実験に収集した．収集したデータを訓練用，検証用，テスト用のデータ数の比率が 8:1:1 となるように分割し，学習用データセットを作成した．

実験は 10 分割交差検証によって行った．統計的検定により，比較する 2 手法の間で順序推定精度の平均値に有意な差があるかを確認する．帰無仮説は「比較する 2 手法の間で順序推定精度の平均値の変化に有意な差はない.」とし，対立仮説は「比較する 2 手法の間で順序推定精度の平均値の変化に有意な差がある.」とする．検定手法は対応のない 2 標本 t 検定を採用し，有意水準は 5%で両側検定を行った．p 値が 0.05 を下回った場合に帰無仮説を棄却することができ，対立仮説が成立すると言うことができる．実験 1 から 3 の結果を表 1,2,3,4 に示す．

4.1 評価指標

4.1.1 NDCG@k

NDCG とは順位に応じたスコア割引を行う評価指標である DCG(DCG_{pred}) を，理想の DCG(DCG_{ideal}) で割ることで正規化したものである．

理想の DCG とは，あるデータ群を正解スコアの降順に基づいて並び替えた際の順序と，予測スコアの降順に基づいて並び替えた際の順序が一致している場合の DCG である．NDCG@k を (11) 式に示す．NDCG@k はスコアの降順の上位 k 番目までを計算の際に使用する NDCG のことである．0 から 1 の値を取り，1 に近い方が精度が高い．

DCG@k を (12) 式に示す．DCG@k はスコアの降順の上位 k 番目までを計算の際に使用する DCG のことである．予測順位が r 番であるデータの予測スコア $g(r)$ を順位に応じたスコア割引である $log_2(r+1)$ で割ることで算出した値を，r が 1 から k 番目について和をとることで算出する．ランキング学習の目的である上位データの推定精度を評価するために，下位の順位にスコア割引行うこの評価指標を採用した．

$$NDCG@k = \frac{DCG_{pred}@k}{DCG_{ideal}@k} \tag{11}$$

$$DCG@k = \sum_{r=1}^{k} \frac{g(r)}{log_2(r+1)} \tag{12}$$

4.2 提案手法の有効性確認

4.2.1 実験内容

本実験では提案手法の有効性を確認することを目的とする．通常の Top-k 学習と提案手法を用いた Top-k 学習を行い，順序推定精度と学習時間を比較する．比較に用いたのは k の値が 2,3 の場合である．k の値が 1 の時は提案手法が終了条件を満たすことがない．k の値が 4 の時は学習時間が膨大となってしまう．これらの理由から k の値が 1,4 の場合では実験を行っていない．10 分割交差検証によって算出した順序推定精度と学習時間の平均値を用いて比較を行う．

表 1　実験 1 の NDCG@5 と p 値と学習時間

手法	精度	p 値	学習時間
通常の Top-2 学習	0.604	-	2.96×10^3
提案手法の Top-2 学習	0.566	0.128	2.11×10^3
通常の Top-3 学習	0.583	-	4.71×10^4
提案手法の Top-3 学習	0.575	0.819	1.67×10^4

表 2　実験 2 の NDCG@5 と p 値

	Top-2	Top-3	Top-4
提案手法	0.566	0.587	0.575
ランダム抽出	0.600	0.575	0.581
p 値	0.154	0.600	0.714

4.2.2 結果・考察

表 1 に各手法によるテスト用データの順序推定精度と t 検定を行った際の p 値，学習時間を示す．提案手法の適用によって k の値が 2,3 の場合に，順序推定精度が有意に下がることはなかったことが分かる．また，学習時間は k の値が 2 の場合に 28.7%，k の値が 3 の場合に 64.5%減少したことが分かる．この結果から，提案手法を用いることによって, 順序推定精度を維持して学習時間を削減することに成功したと言える．

k の値が 2 の場合は有意差はないが精度が低下している．提案手法において k の値が 2 の場合は下位 50%中のデータ同士を用いた順序データを学習に使用していない．そのためモデルの汎化性能を高められていないと考えられる．一方で k の値が 3 の場合は順序推定精度の差が小さい．k の値が 3 の場合は下位 33%中のデータ同士を用いた順序データを学習に使用していない．k の値が 3 の場合は 2 の場合に比べて使用していないデータの割合が小さいため，モデルの汎化性能を高められたと考えられる．

4.3 提案手法における上位データ抽出の有効性確認

4.3.1 実験内容

本実験では提案手法における上位データ抽出の有効性を確認することを目的とする．提案手法でデータを抽出する際に，上位から抽出する場合とランダムに抽出する場合で Top-k 学習を行い，順序推定精度を比較する．比較に用いたのは k の値が 2,3,4 の場合である．k の値が 1 の時は提案手法が終了条件を満たすことがない．この理由から k の値が 1 の場合では実験を行っていない．10 分割交差検証によって算出した順序推定精度の平均値を用いて比較を行う．

4.3.2 結果・考察

表 2 に各手法によるテスト用データの順序推定精度と t 検定を行った際の p 値を示す．k の値が 2,3,4 の場合に上位データの抽出とランダム抽出による Top-k 学習では順序推定精度に有意な差がなかったことが分かる．この実験結果から，提案手法におけるデータ抽出時に，上位データを抽出することが有効だとは言えないことが分かる

k の値が 2 の場合はランダムによるデータ抽出の順序推定精

表 3　実験 3 の NDCG@5

	10%	20%	30%	40%	50%
Top-2	0.606	0.600	0.593	0.571	0.566
Top-3	0.585	0.583	0.587	0.589	0.578

表 4　実験 3 の p 値

	10%-20%	10%-30%	10%-40%	10%-50%	20%-30%
Top-2	0.819	0.684	0.195	0.166	0.837
Top-3	0.916	0.958	0.901	0.766	0.791
	20%-40%	20%-50%	30%-40%	30%-50%	40%-50%
Top-2	0.269	0.227	0.433	0.365	0.843
Top-3	0.712	0.781	0.517	0.576	0.517

度が高かった．この結果から，k の値が 2 の場合にランダムによるデータ抽出の方がモデルの汎化性能向上に有効であったと考えられる．上位データ抽出では下位 50%中のデータは抽出しないことに対し，ランダム抽出では下位 50%中からもデータの抽出を行う．そのため, 上位データ抽出では下位 50%中のデータ同士の順序について学習を行わないことに対し，ランダム抽出では学習を行うことができる．よって, 上位データ抽出と比較して順位全体についての順序推定が可能になり，汎化性能が向上したと考えられる．

k の値が 3 の場合は上位データ抽出の順序推定精度が高かった．この結果から,k の値が 3 の場合に上位データの抽出がモデルの順序推定精度向上に有効であったと考えられる．4.2 節の考察から，k の値が 3 の場合は対象データの更新回数が 0 回の段階でモデルの汎化性能がある程度高められていると考える．そのため，ランダムによるデータ抽出の利点であるモデルの汎化性能の向上は期待できず，上位データのみを学習することが順序推定精度の向上に繋がったと考えられる．

k の値が 4 の場合は順序推定精度の差が小さかった．この結果から，k の値が 4 の場合はデータ抽出方法による順序推定精度への影響が小さいと考えられる．k の値が 4 の場合は対象データ更新時に対象データの 25%を抽出するため，更新後の対象データ数が少なくなってしまう．そのため，対象データ更新後の学習効果が小さく，対象データの更新回数が 0 回の段階のモデルから順序推定精度の変化が小さいと考えられる．よって，データ抽出方法による順序推定精度への影響が小さいと考えられる．

4.4　提案手法における最適なデータ抽出割合の発見

4.4.1　実 験 内 容

本実験では提案手法における最適なデータ抽出割合の発見を目的とする．提案手法でデータを抽出する際に，抽出割合を 10%,20%,30%,40%,50%それぞれの場合で行い, 順序推定精度を比較する．比較に用いたのは k の値が 2,3 の場合である．k の値が 3 の時は 30%の代わりに，提案手法における抽出割合である $\frac{100}{3}$%を用いた．k の値が 1 の時は提案手法が終了条件を満たすことがない．この理由から k の値が 1 の場合では実験を行っていない．10 分割交差検証によって算出した順序推定精度の平均値を用いて比較を行う．

4.4.2　結果・考察

表 3 に各手法によるテスト用データの順序推定精度を示す．また，表 4 に同一の k の値において各抽出割合で総当たりの t 検定を行った際の p 値を示す．k の値が 2,3 の場合で，データの抽出割合が 10%,20%,30%,40%,50%の場合において，順序推定精度の差に有意な差がなかったことが分かる．この実験結果から，提案手法における最適なデータの抽出割合の発見には至らなかったことが分かる．

k の値が 2 の場合は抽出割合を大きくするほど順序推定精度が減少した．また，k の値が 3 の場合は 50%の時に最も順序推定精度が低かった．この結果から提案手法におけるデータの抽出割合は小さい方が順序推定精度を高められると考えられる．データの抽出割合が大きい場合に順序推定精度が下がる原因として，上位データと下位データに似た特徴量を持っているデータが存在する可能性が考えられる．データ抽出による追加学習によって，ある特徴量を持つデータを上位と予測するようにモデルは学習される．そのため, この特徴量を持つデータが下位に多い場合に, それらのデータを上位データであると予測するようになることで，モデルの汎化性能が低下すると考えられる．抽出割合が小さい場合はこのような過学習が起きにくいと考え，汎化性能が低下しなかったと考えられる．

5　おわりに

本研究では Top-k 学習において，順序推定精度を維持しながら学習時間を削減することを目標とした．我々は，対象データの順位に沿った k 段階分割結果に基づいた計算データ選定手法と，学習進度に合わせて対象データを上位に更新する手法を組み合わせた能動学習手法を提案した．

我々は Top-k 学習の問題は二つあると考えた．一つ目は，難易度の高い学習をモデル学習の初期段階に行うことである．Top-k 学習では対象データ中の k データの順列を考えるため, 僅差の順位を持つデータ同士の順序を学習に用いることがある．しかし，我々は僅差の順位を持つデータ同士の順序予測を学習の初期段階に行うことは難易度が高いと考えた．そこで我々は学習に用いる順序データの難易度を徐々に高めることによって学習効率の向上が期待できると考えた．二つ目は対象データ中の下位データと上位データで同じ時間学習を行うことである．ランキング学習では上位データの順序推定精度を高めることを目的とする．しかし，我々は Top-k 学習では下位データを上位データと同じ時間学習しているため，学習効率が低下していると考えた．そこで我々は対象データから下位データを取り除くことを考えた．この手法による学習データの削減によって，上位データの順序推定精度を保ちながら学習時間を削減できると考えた．

我々は二つの手法を組み合わせた Top-k 学習における能動学習手法を提案した．一つ目は，対象データを順位に沿って k 段階分割し，分割後グループの重複がない k データの順列のみで計算データを作成する手法である．二つ目は，モデルが一定の

学習進度に達した際に，対象データから上位 $\lfloor D/k \rfloor$ 位までのデータを抽出して新たな対象データとする手法である．

提案手法を評価するために三つの評価実験を行った．一つ目は提案手法の有効性確認を目的として行った．通常の Top-k 学習と提案手法を用いた Top-k 学習で順序推定精度と学習時間の比較を行った．二つ目は提案手法における上位データ抽出の有効性確認を目的として行った．提案手法でデータを抽出する際に，上位から抽出する場合とランダムに抽出する場合で Top-k 学習を行い，順序推定精度の比較を行った．三つ目は提案手法における最適なデータ抽出割合の発見を目的として行った．提案手法でデータを抽出する際に，抽出割合を 10%,20%,30%,40%,50%それぞれの場合で Top-k 学習を行い，順序推定精度の比較を行った．さらに，各実験において手法による順序推定精度の変化に有意な差があるかを確認するために統計的検定を行った．

評価実験の結果，提案手法を用いて Top-k 学習を行うことによって，通常の Top-k 学習を行った場合から順序推定精度を有意に下げることなく学習時間を削減できることが分かった．また，提案手法でデータを抽出する際に上位データから抽出することの有効性は k の値が 2 の場合には確認できなかったが，k の値が 3 の場合には確認できた．さらに，提案手法でデータを抽出する際の抽出割合は小さい方が順序推定精度を高められることが確認できた．

提案手法でデータを抽出する際に，k の値が 2 の場合でランダム抽出の順序推定精度が高かった．この結果から提案手法では下位データ同士の順序を学習することができないため，モデルの汎化性能が低下したと考えられる．k の値が 3 の場合は対象データ中の上位データ，中位データ，下位データの順序を学習することができるため，モデルの汎化性能を k が 2 の場合と比較して高められると考える．そのため，上位データ抽出による汎化性能の低下が軽減され，上位順序の推定精度を高められたと考えられる．

提案手法でデータを抽出する際に，抽出割合が 50%に近い場合に順序推定精度が低かった．この結果から上位データ抽出による追加学習によってモデルが過学習され，汎化性能が低下したと考えられる．上位データと下位データに似た特徴量を持つデータがあった場合，追加学習によってその特徴量を持つデータは上位データと推定するようにモデルは学習される．しかし，その特徴量を持つデータが下位に多い場合には, この追加学習によってモデルの汎化性能が低下すると考えられる．抽出割合が小さい場合はこのような過学習が起きにくいため汎化性能が低下せず, 順序推定精度が高まったと考えられる．

本稿では対象データ群の順位に沿った多段階分割による計算データの選定と，対象データ群の上位を抽出して新たな対象データとすることによる能動学習手法を提案した．本手法では各データが持つ特徴量のうち順位情報のみを用いて計算データの選定を行っている．順位情報に加えてクエリ情報を用いることで詳細な計算データの選定が可能になり，さらなる学習効率の向上が期待できる．また，提案手法では多段階分割後の最下位データ群中に存在するデータ同士の順序について学習を行えていない．最下位データ群中には学習に有効であるデータが含まれている可能性があるため，これらについて学習が可能なデータ選定手法を用いることによって，さらなる精度向上が期待できる．今後はこれらの手法でも実験を行い，精度の向上に努めたいと考える．また,Top-k 学習において k の値を増加させることにより順序推定精度が高まることを期待していたが, この傾向は確認できなかった．この結果から,Top-k 学習において k を増加させる必要性が少なくなるため, 提案手法を適用する必要のある場面が限られてしまうことが考えられる．しかし, この結果が対象としたデータセットやクエリに依存したものである可能性が考えられる．また,k の値を増加させることによって正しい順序推定が可能になったデータの調査を行えていない．今後は複数のクエリやデータセットを対象として提案手法を適用することや,k の値の増加によって順序推定を正しくできたデータの傾向分析を行うことで，提案手法の適用場面を調査したいと考える．

謝辞　本研究の一部は，JSPS 科学研究費 (19H04218, 23H03694)，越山科学技術振興財団，およびソニー株式会社の助成を受けたものです

文　献

[1] Tie-Yan Liu, et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 3, pp. 225–331, 2009.

[2] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007.

[3] Razieh Rahimi, Ali Montazeralghaem, and James Allan. Listwise neural ranking models. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pp. 101–104, 2019.

[4] Junjie Liang, Jinlong Hu, Shoubin Dong, and Vasant Honavar. Top-n-rank: A scalable list-wise ranking method for recommender systems. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1052–1058. IEEE, 2018.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

[6] Minghan Li, Xialei Liu, Joost van de Weijer, and Bogdan Raducanu. Learning to rank for active learning: A listwise approach. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5587–5594. IEEE, 2021.

[7] Tao Qin and Tie-Yan Liu. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

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

2024年2月28日(水) 10:00 ~ 12:10 | T4-A オンライン（Zoom Events)

画像の理解

座長:宮森 恒(京都産業大学)
コメンテータ:風間 一洋(和歌山大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T4-A-1-01] アンサンブル学習を用いた植物標本の分類

*小柳 亮輔[1]、白井 匡人[1] (1. 島根大学)

10:20 ~ 10:35
[T4-A-1-02] 複雑な劣化画像を学習させた Vision Transformer による
単一画像超解像

*林 亮佑[1]、河野 浩之[1] (1. 南山大学大学院　理工学研究科　機械電子制御工学専攻　河野研究室)

10:35 ~ 10:50
[T4-A-1-03] CNNとTransformer を用いた画像誘導型深度補完システムの実装

*大極 康平[1]、河野 浩之[1] (1. 南山大学大学院　理工学研究科　機械電子制御工学専攻　河野研究室)

10:50 ~ 11:15
[T4-A-1-04] 画像復元のための空間方向Attention

*沖津 永享[1]、白井 匡人[1] (1. 島根大学)

11:15 ~ 11:40
[T4-A-1-05] レシピの手順説明画像認識における食材視覚特徴分析

*張 訳心[1]、山肩 洋子[2]、田島 敬史[1] (1. 京都大学、2. 東京大学)

アンサンブル学習を用いた植物標本の分類

小柳　亮輔†　　白井　匡人†

† 島根大学 自然科学研究科　〒 690-8504 島根県松江市西川津町 1060
E-mail: †n23m103@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp

あらまし　植物の種の特定を行う同定作業は，膨大な知識を必要とするため専門家でさえ正しく行うことが難しい．近年，畳み込みニューラルネットワーク (CNN) や Vision Transformer(ViT) などの機械学習モデルの登場により，高精度な画像分類を行うことが可能である．それらのモデルを植物のデータセットに適用させることで，植物の同定作業を自動化する試みがなされている．しかし，これらのモデルはモデルによって着目する箇所が異なるため，クラスの分類精度に大きな差が存在する．本研究では，複数のモデルを使用し，モデル毎の着目範囲の違いから予測確率に重み付けを行うアンサンブル手法を提案する．実験では，異なる 4 つの CNN モデルと ViT モデルの計 5 つを用いて提案手法を適用した結果，先行研究の手法と比較して正解率が向上した．

キーワード　画像分類，Grad-CAM，アンサンブル学習，植物標本

1 背　　景

植物の同定とは，その植物がどの種や分類群に属するかを特定する作業のことである．植物の同定は，膨大な知識と長年の経験が必要であり，専門家でさえ正しく同定を行うことは難しい．佐藤によると，誤同定が起こる原因として，同定者の知識不足や勘違いによるもの，形態が似ている分類群は認識が難しいこと，昔と今の学名が異なる植物が存在することが挙げられている [1]．また，植物は次々と新種が発見されており [2]，それに伴い，植物同定の難易度が高くなっていくことが予想される．

近年，高精度な機械学習モデルが多数登場し，中でも畳み込みニューラルネットワーク (CNN) [3] や Vision Transformer(ViT) [4] は画像分類の分野においてよく用いられている．それらの機械学習モデルを植物のデータセットに適用させることで，植物の同定作業を自動化し，植物標本の分類作業の負担を軽減させることを目的とする取り組みが多くなされている [5] [6] [7]．しかし，これらの先行研究においては，クラス毎の分類精度に大きなバラつきが存在するという問題がある．この原因として，植物の分類では科や属により分類の難易度が大きく異なることやクラスによって着目すべき箇所が異なることが挙げられる．

本研究では，先行研究の問題を踏まえた上で，複数のモデルを使用し，モデル毎の着目範囲の違いから予測確率に重み付けを行うアンサンブル手法を提案する．

2 関 連 研 究

2.1 畳み込みニューラルネットワーク (CNN)

畳み込みニューラルネットワーク (CNN) は，畳み込み層，プーリング層，全結合層で構成されている深層学習モデルである．畳み込み層ではフィルタを使って画像をスキャンし，その特徴を抽出する．プーリング層では特徴を縮小することで移動不変性の付与を行う．畳み込み層とプーリング層にて得られた特徴を基に，最終的な出力を全結合層で行う．

2.2 Vision Transformer(ViT)

Visiton Transformer(ViT) とは，自然言語処理分野で用いられる Transformer を画像処理タスクに適用したモデルである．画像をパッチに分割してそれぞれに位置情報を与える．これを一直線に並べて 1 次元のデータとして扱うことで，Transformer に画像を適用することができる．ViT は計算効率が良いため，大規模なデータセットでの学習が可能である．

2.3 Grad-CAM による判断根拠の可視化

近年，画像認識の判断根拠を可視化する技術が研究されてきた．Grad-CAM [8] は，最後の畳み込み層の勾配情報を使用することで，モデルが分類を行う際に画像のどの部分が最も大きく影響しているかを可視化する手法である．入力のクラスが c の場合，(1) 式から重み係数 α_k^c を求め，Grad-CAM の出力 $L^c_{\text{Grad-CAM}}$ を (2) 式のように計算する．

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

(1) 式の $\frac{\partial y^c}{\partial A_{ij}^k}$ は，k 枚目の特徴マップ A^k の座標 (i, j) が変化するときのクラス c に対するモデルの予測確率 y^c の変化量である．Z は特徴マップのサイズ，すなわちピクセルの総数である．その後、特徴マップ全体の平均をとり，クラス c の k 番目のフィルターの重み係数 α_k^c が求まる．

(2) 式は，(1) 式で求められる重み係数 α_k^c を用いて，k 番目の特徴マップ A_{ij}^k の重み付けを行う．最後に活性化関数 ReLU を適用して，0 以下の値を排除することでクラス c に与えた位置の影響を数値化することができる．

2.4 植物標本の自動同定

白井らの研究 [5] は，日本に生息する植物の葉や小枝，果実の植物標本を対象に自動で同定を行うものである．葉のない標本や，虫食いの多い標本などをデータセットから削除し，植物標本に含まれる竹の定規などの分類に関係のない要素を排除することで，分類精度が向上している．しかし，植物標本の自動同定における問題点として，学習に使用するモデルはクラスによって分類精度に大きなバラつきが生じることが挙げられる．これは，植物の科・属によって分類の難易度が大きく異なることや，クラスによって着目しなければならない箇所が異なるためである．そのため，単一のモデルでは科・属に適したモデルを構築することが困難である．

2.5 葉の特徴に着目した分類

奥田らの研究 [6] は，葉の画像から樹木種を自動で同定するものである．葉の画像分類においては葉全体だけでなく，葉の輪郭，葉脈などの着目する箇所により，分類の傾向が異なることが明らかになっている．葉の様々な特徴を表す画像を作成し組み合わせ，葉の特徴を網羅的に考慮した分類器を作成した結果，正解率が向上している．このことから，植物の画像分類において，異なる箇所に着目し組み合わせることの有用性が示されている．

2.6 重み付きアンサンブル手法による画像分類

CNN による画像分類は様々な研究分野で多く使用されてる．Iqball らの研究 [9] では，X 線の画像分類の精度を向上せることを目的として，異なる 5 つの CNN モデルを用いたみ付きアンサンブル手法を提案している．

まず，m 個のモデルのうち j 番目のモデルの重み α_j は式により求める．ここで，Acc_j は j 番目のモデルを検証データセットで評価した際の正解率であり，β は (4) 式で求めた $[Acc_j]_{j=1}^{m}$ の最小値である．

$$\alpha_j = Acc_j - \beta + 1, \quad j = 1 \dots m \tag{3}$$

$$\beta = \min\left([acc_j]_{j=1}^{m}\right) \tag{4}$$

$$OutputClass_{T_i,ensemble} = \max\left[\sum_{j=1}^{m}(\alpha_j \times [C_k]_{i,j})\right]_{k=1}^{nc} \tag{5}$$

ここで，$[C_k]_{i,j}$ はモデル M_j のクラス所属確率であり，nc はクラスの総数である．入力データ T_i に対する $OutputClass_{T_i,ensemble}$ は (3) 式で求めた α_j を使用し，(5) 式で重み付き和を計算し，最大であるクラスを出力する．各モデルの精度に応じて重み付けされるため，精度が高いモデルほど最終出力への与える影響が大きい．実験では，上記の手法を 2 種類の X 線画像データセットを使ってテストしたところ，文献のモデルと比較して精度が有意に向上した．

3 提案手法

本研究では，構造が異なる複数のモデルを独立に学習させ，モデル毎の所属確率に重み付けするアンサンブル手法を提案する．

はじめに，複数のモデルに対し，学習データセットを使用して独立に学習をする．次に，検証データセットを用いて各モデルで分類を行った後，Grad-CAM の出力をヒートマップとして可視化させる．ここで，植物画像において，画像の着目箇所が限定されているほど，そのモデルが特徴をよく捉えていると仮定する．その仮定に基づき，テストデータセットにおけるモデルの着目範囲に応じた重みを計算する．最終的にその重みを各モデルの所属確率に掛け合わせ，最大値のクラスを予測として出力する．

各モデルに Grad-CAM を適用させたヒートマップを図 1 に示す．ヒートマップの色が青色に近いほどモデルの着目度が弱く，赤色に近づくほどモデルの着目度が強い箇所を示している．モデルによって，着目箇所や着目範囲が異なることがわかる．

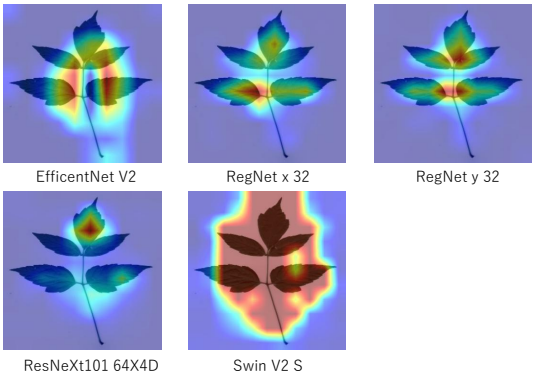


図 1　異なるモデルに Grad-CAM を適用した結果

3.1 Grad-CAM のヒートマップを用いた重みの計算

2.6 節で述べた先行研究の手法を基に，モデルの着目範囲に応じた重み付きアンサンブル手法を提案する．

m 個のモデルにおける j 番目のモデルの重み α_j は，(6) 式により求められる．ここで，Acc_j は j 番目のモデルにおける検証データセットの正解率である．β は (7) 式により求められる．$RedArea_j$ は，テストデータセットに対するモデル M_j の Grad-CAM のヒートマップの着目領域における赤色部分の面積である．同様に，$OverallArea_j$ は，テストデータセットに対するモデル M_j の Grad-CAM のヒートマップの着目領域全体の面積である．つまり $\frac{RedArea_j}{OverallArea_j}$ は，Grad-CAM のヒートマップの着目領域全体に占める赤色部分の面積であり，この値が小さいほどそのモデルが特徴をよく捉えているとする．(8) 式は，節 2.6 と同様に，$[C_k]_{i,j}$ はモデル M_j のクラス所属確率であり，nc はクラスの総数である．入力データ T_i に対する $OutputClass_{T_i,ensemble}$ は (6) 式で求めた α_j を使用し，重み付き和を計算し，最大であるクラスを出力する．

$$\alpha_j = Acc_j * \beta + 1, \quad j = 1 \ldots m \tag{6}$$

$$\beta = 1 - \frac{RedArea_j}{OverallArea_j} \tag{7}$$

$$OutputClass_{T_i,ensemble} = \max \left[\sum_{j=1}^{m} (\alpha_j \times [C_k]_{i,j}) \right]_{k=1}^{nc} \tag{8}$$

3.2 提案手法の概要

提案手法の概要を図 2 に示す．テストデータを学習済みの 5 つのモデルで分類を行い，クラスの所属確率を求める．それぞれのモデルに Grad-CAM を適用させ，モデルの着目部分をヒートマップとして可視化する．次に，ヒートマップの着目範囲に応じた重みを計算する．得られた重みを用いて，各モデルのクラスの所属確率に重み付けを行い, 確率が最も高いクラスを出力し，本手法における予測値とする．

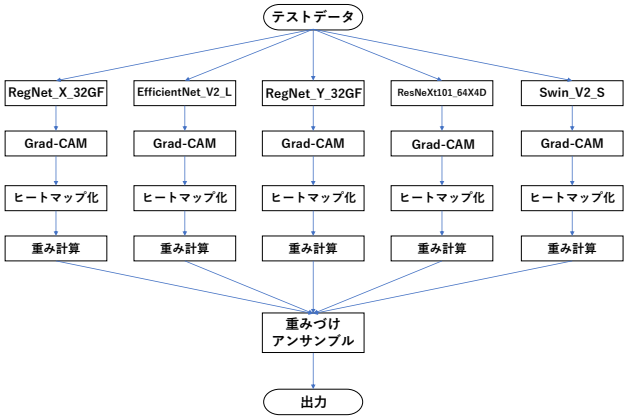


図 2　提案手法の概要

4　実　　験

実験では，植物標本の分類における提案手法の精度の検証を行う．植物の画像分類タスクを以下の 2 つの場合で実行し，精度を比較する．

- Iqball らの先行研究の手法
- 提案手法

4.1 使用データセット

データセットは，PlantCLEF2015 のフランスや近隣諸国の樹木，ハーブ，シダを含んだ上位 100 種，30,470 枚の画像を使用する．データセットは，学習用，検証用，テスト用として 8:1:1 に分割して実験する．データセットは図 2 のように,「屋外で撮影されている葉」,「スキャンされている葉」,「花」,「実」,「幹」,「枝」,「全体」の 7 種類の画像から構成されている．データセットの一例を図 3 に示す．

図 3　データセットの一例

4.2 使用モデル

実験では，EfficentNet V2 [10]，RegNet x 32 [11]，RegNet y 32 [11]，ResNeXt101 64X4D [12] の 4 つの CNN モデルと，Swin V2 S [13] の ViT モデルを用いる．各モデルは，ImageNet [14] により事前学習済みである．ImageNet は，1,000 個のクラスと 120 万枚の画像の学習データセットから構成される．

表 1　学習のパラメータ

モデル名	エポック数	入力画像サイズ	バッチサイズ
EfficentNet V2	100	299 × 299px	16
RegNet x 32	50	299 × 299px	16
RegNet y 32	120	299 × 299px	16
ResNeXt101 64X4D	50	299 × 299px	16
Swin V2 S	120	352 × 352px	32

4.3 評 価 指 標

評価指標として，正解数をデータ数で割ることで求められる正解率を用いる．正解率は 0〜1.0 の範囲で表される．値が 1.0 に近いほどそのモデルは優れた性能を持つと考えられる．正解率を求める際に用いる用語を表 2 にまとめる．

表 2　正解率を求める際に用いる用語

	Positive	Negative
True	TP (正しく予測されたサンプル数)	TN (正しく予測されたサンプル数)
False	FP (誤って予測されたサンプル数)	FN (誤って予測されたサンプル数)

正解率は (9) 式で求めることができる．

$$\text{正解率} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

4.4 実 験 結 果

データセットを用いてモデルの学習と分類をおこなう．各モデルと，先行研究の手法，提案手法の正解率は，表 3 のような結果となる．先行研究の手法と比べ，提案手法は正解率が 0.0033 上昇した．

表 3　各手法の正解率

手法	正解率
EfficentNet V2	0.9071
RegNet x 32	0.9091
RegNet y 32	0.9045
ResNeXt101 64X4D	0.8927
Swin V2 S	0.8533
先行研究の手法	0.9275
提案手法	0.9308

5　まとめと今後の課題

本研究では，Grad-CAM の出力結果を用いた重み付けアンサンブル手法を提案した．実験では先行研究の手法と比べ，本手法は正解率が 0.0033 上昇することを確認した．上昇の要因として，モデルによって生じる得意・不得意が Grad-CAM の出力結果に反映されたためだと考えられる．

今後の課題として，さらに分類精度を向上させるために手法を改善していくことが挙げられる．

文　献

[1] 佐藤広行. 標本は正しい同定がされているとは限らない. 北方山草, No. 31, pp. 67–70, 2014.

[2] Ahmad Reza Mehrabian, Siavash Naghizadeh, Farzaneh Khajoei Nasab, Hossein Moradi Zeinab, and Valiollah Mozaffarian. Two new species and a new record for onosma l.(boraginaceae) from northwestern iran. *Feddes Repertorium*, Vol. 133, No. 3, pp. 173–193, 2022.

[3] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, Vol. 53, pp. 5455–5516, 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Masato Shirai, Atsuko Takano, Takahide Kurosawa, Masahito Inoue, Shuichiro Tagane, Tomoya Tanimoto, Tohru Koganeyama, Hirayuki Sato, Tomohiko Terasawa, Takehito Horie, et al. Development of a system for the automated identification of herbarium specimens with high accuracy. *Scientific Reports*, Vol. 12, No. 1, p. 8066, 2022.

[6] 奥田萌莉, 大島裕明ほか. 形・葉脈の特徴に着目した樹葉画像の深層学習に基づく分類. 情報処理学会論文誌データベース (TOD), Vol. 16, No. 1, pp. 14–25, 2023.

[7] Raisa Akter and Md Imran Hosen. Cnn-based leaf image classification for bangladeshi medicinal plant recognition. In *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pp. 1–6. IEEE, 2020.

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[9] Talib Iqball and M Arif Wani. Weighted ensemble model for image classification. *International Journal of Information Technology*, Vol. 15, No. 2, pp. 557–564, 2023.

[10] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pp. 10096–10106. PMLR, 2021.

[11] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

[12] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[13] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, Vol. 25, , 2012.

複雑な劣化画像を学習させたVision Transformerによる単一画像超解像

林　　亮佑†　　河野　浩之†

† 南山大学理工学研究科　〒 466-8673　愛知県名古屋市昭和区山里町 18
E-mail: †m22sc002@m.nanzan-u.ac.jp, ††kawano@nanzan-u.ac.jp

あらまし　現在，画像を用いたアプリケーションは多く存在しており，例えば画像から物体を認識するアプリケーションなどもある．物体認識などのタスクにおいては元の画像データは高解像である必要がある．しかし，現実世界において常にノイズの無い高解像度画像が得られるとは限らない．そのような場合にも超解像を用いることで低解像度の画像であっても物体認識を行うことができるようになる．本研究では低解像度の画像を高解像度の画像にする超解像モデルを提案する．SwinIR をベースモデルとして複雑な劣化画像を知覚損失と L1 損失で学習させた．その結果ベースモデルよりも DIV2K では PSNR は約 8.5% ，SSIM は約 5.6%向上し，Set14 では PSNR は約 6.7%，SSIM は約 5.8%向上した．

キーワード　超解像，Swin Transformer，劣化画像，知覚損失

1　超解像が必要とされる場面

現在，衛星技術との進歩により様々なアプリケーションで衛星画像が使用されている．実際に衛星画像の需要は増加すると予想されており表 1 によると衛星画像の市場規模は 2030 年には 2022 年の約 4.3 倍の 141.8 億ドルまで増加する．需要の増加の背景として災害時の地上の状態の把握や防衛・安全保障といった緊急事態の対応で衛星画像の利用が増加していることがある．

そして衛星画像から情報を読み取る際に超解像が必要であり，例として衛星画像から駐車している車の物体認識をする場合，低画質画像ではうまく認識できないことがある．そのため超解像を行い高解像にすることで認識率を上げることができる．

表 1　衛星画像の市場規模 [1]

年	市場規模
2022 年	32.7 億ドル
2020 年	141.8 億ドル

本研究では衛星画像以外にも携帯のカメラで撮影された劣化画像や画像の加工や圧縮により劣化した画像に対応した汎用性のある超解像モデルを提案する．

学習に用いる劣化画像は，ぼかし，リサイズ，ノイズ，JPEG 圧縮を組み合わせることで様々な劣化シーンに対応した劣化表現を獲得する．そして，損失関数に知覚損失関数を用いることでより知覚的な品質が高い超解像画像を得ることを狙う．

2　超解像手法の先行研究

超解像手法の先行研究を表 2 に示す．先行研究の各モデルの特徴として，CNN ベースの real-ESRGAN は Residual in Residual Dense Block という CNN の層と層の間を残差接続で密につなげることで CNN の深いネットワークを構築している．

次に CNN と Transformer を組み合わせた ACT では CNN の局所的な特徴量を抽出する能力と Transformer の大局的な特徴量を抽出する能力を掛け合わせることでより高い超解像性能を発揮している．

そして Transformer ベースの SwinIR は Vision Transformer のアテンション機構の計算をより効率良く行う Shift-Window 方式を用いた Swin Transformer を用いることで Transfomer のメリットである低パラメータ数に加え，計算効率を向上させている．

本研究では Swin Transformer に複雑な劣化画像を学習させることで CNN ベースの手法よりもパラメータを削減し，高い超解像性能を発揮させることを狙う．

表 2　超解像手法の先行研究

モデル名	特徴
real ESRGAN [2]	blur, noise, resize, JPEG を用いた劣化画像で学習
ACT [3]	CNN と ViT を合わせたモデルを提案 局所的・非局所的な特徴を抽出
SwinIR [4]	SwinViT を超解像に採用 Swin 機構により計算速度を向上

3　劣化画像を学習させた ViT の提案手法

3.1 では使用する超解像モデル，3.2 では Shift Window MSA，3.3 では劣化画像の生成フロー，3.4 ではぼかし付加手順，3.5 ではリサイズ手順，3.6 ではノイズ付加手順，3.7 では JPEG 圧縮手順について説明する．

3.1　使用する超解像モデル

本研究では低パラメータにおいても高い超解像性能を発揮した Swin Transformer をベースに超解像モデルを構築する．

Swin Transformer の構成は LayerNorm の後に shift window multi-head self-attention を行い LayerNorm と MLP へ通す一連の流れを Swin Transformer Layer(STL) としてこの STL を 6 つの後に 1x1 の畳み込みを行う Swin Transformer Block(STB) を使用する．

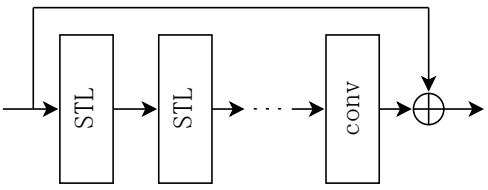


図 1　Swin Transformer Block(STB)

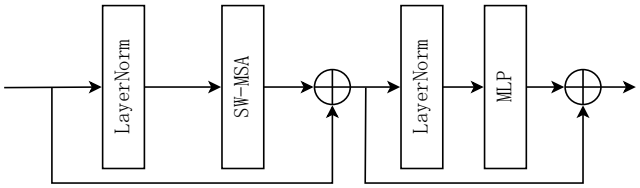


図 2　Swin Transformer Layer(STL)

3.2　Shift Window MSA

Transformer では画像をパッチごとに切り分けて Multi Self Attention(MSA) を行うが，通常 MSA では特徴マップが $h \times w$ 個のパッチがあるすると計算量は $4hwC^2 + 2(hw)^2C$ となる．一方，Swin Transformer ではまず Window Attention によって切り分けられたパッチを Window サイズ M ごとで計算するので計算量は $4hwC^2 + 2M^2hwC$ となる．よって画像サイズが大きくなると通常の MSA では計算量が二次関数的に増加してしまうが Swin Transformer では Window Attention を用いることで計算量は線形に減少させることができる．

これにより Swin Transformer は計算量を削減しながら画像の特徴量を抽出することができるため，通常の Transformer よりも良いと言える．

3.3　劣化画像の生成フロー

劣化画像の超解像を深層学習で行う上で学習画像の多様な劣化表現はモデルの性能を左右する．よって本研究では劣化の種類としてぼかし，ノイズ，JPEG 圧縮，リサイズを組み合わせて劣化画像を生成する．劣化画像の生成フローを図 3 に示す．

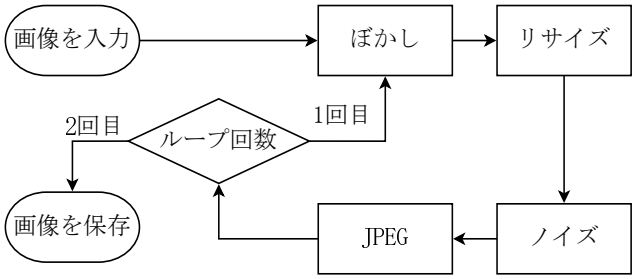


図 3　劣化画像生成フロー

3.4　ぼかし付加手順

ぼかしでは一般的にガウスぼかしが使われるが単一のぼかし手法のみではネットワークの汎化性能を上げることができない．よって本研究ではぼかしカーネルとして対称ガウスカーネル，非対称ガウスカーネル，Sinc カーネルを使用する．ぼかし画像の生成フローを図 4 に示す．

それぞれの特徴として，一般的なぼかしである対称ガウスカーネルはカーネルサイズを大きくしていくと画像全体が同じようにぼやけていく．一方，非対称ガウスカーネルではカーネルサイズを大きくしていくとカメラの手振れのような斜め方向のぼかしが生成できる．そして，Sinc カーネルではカットオフ周波数を大きくしていくと画像の輪郭部分にオーバーシュートのようなぼかしが生成できる．

ぼかしは Sinc とガウスで分岐した後，ガウスの場合は対称・非対称で一様乱数を用いた確率で分岐する．その後，畳み込みによってカーネルを画像に適用する．

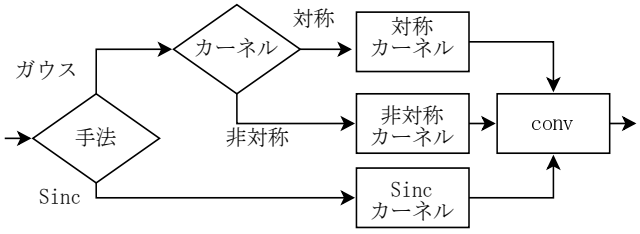


図 4　ぼかし画像生成フロー

3.5　リサイズ手順

リサイズ劣化では画像の拡大・縮小による劣化を表現するために Bicubic 補間，Bilinear 補間，Area 補間をランダムで用いて画像の拡大・縮小を行う．さらに画像のサイズ倍率をランダムで設定し倍率による劣化の度合いを変えていく．

しかし，画像のサイズをランダムに設定すると最終出力の画像が不揃いになってしまうため 2 回目の JPEG 圧縮時に目的の画像サイズにリサイズされる．

3.6　ノイズ付加手順

ノイズではガウス分布とポアソン分布を用いてノイズを生成することでよりノイズの劣化を複雑なものにする．フローの流れはまず確率でガウスノイズかポアソンノイズに分岐する．その後，それぞれ確率でカラーノイズかグレーノイズに分岐し，ノイズ画像が生成される．

ガウスノイズとポアソンノイズの違いとして，ガウスノイズは画像の画素値に関係なく確率的にノイズが付加されていくのに対し，ポアソンノイズは画像の画素値に連動してノイズの付加が変化していく．そのためポアソンノイズでは画素値が大きくなるとノイズも強くなり，画素値が小さくなるとノイズも弱くなる．

3.7　JPEG 圧縮手順

JPEG 圧縮ではより良い劣化表現を獲得するために 2 回目の JPEG 圧縮処理を変更する．2 回目の JPEG 劣化画像の生成フ

ローを図 5 に示す．

この処理では画像がオーバーシュートの劣化後に JPEG 圧縮が行われた場合と JPEG 圧縮が行われた後にオーバーシュート劣化が発生した場合を想定している．また，sinc によるオーバーシュート劣化は確率によってスキップされる．

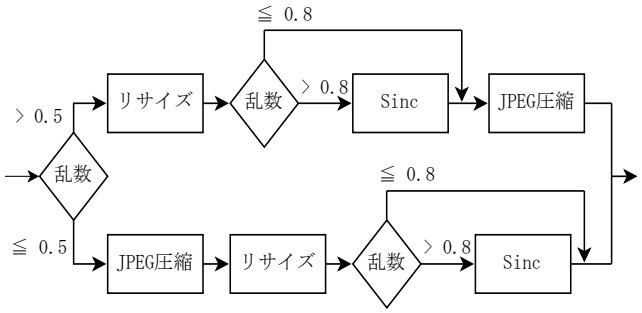


図 5　JPEG 圧縮画像生成フロー

4　実 験 環 境

4.1 では構築環境，4.2 では使用するデータセット，4.3 では劣化画像の生成詳細，4.4 では超解像モデルの損失関数，4.5 では超解像モデルの学習設定詳細について説明する．

4.1　構 築 環 境

本研究では Jupyter Notebook で, Python を用いてプログラムを実行する．また，ネットワークの構築，学習には Tensorflow と pytorch を使用する．

各フレームワークの特徴を表 3 に示す．Tensorflow を使用する理由は学習曲線の表示ができる TensorBoard が利用できるため，学習の監視が容易になりハイパーパラメータの調整に役立つ為である．そして Pytorch を使用する理由は深層学習の研究では Pytorch が使用されていることが多く，Pytoch を使用することで他手法のモデルの構築を用意にすることができる為である．

表 3　使用するフレームワーク

フレームワーク	特徴
Tensorflow	高，低レベルの API が使用可能 多くのプラットフォームで利用可能
Pytorch	低レベルの API が使用可能 デバックが容易

4.2　使用するデータセット

次に画像データセットの特徴を表 4 に示す．これら 4 つのデータセットを使う理由として，DIV2K と Flickr2K は様々な街の画像や日常のシーンを，OST は山や草，空などのアウトドアシーンを含んでおり，manga109 では漫画のイラストが含まれている為，これらのデータセットを使うことで様々な画像の特徴を学習することができるからである．ダウンロードした画像は 256×256 ピクセルに切り抜いて使用する．

表 4　使用するデータセットの特徴

データセット	画像の枚数
DIV2K	800 枚の画像
Flickr2K	2650 枚の画像
OST	アウトドアシーンを集めた 10342 枚の画像
manga109	109 枚の漫画の表紙画像

4.3　劣化画像生成

次にデータセットから切り抜いた画像を劣化させる．劣化を行う際のハイパーパラメータを表 5 に示す．

ぼかしの標準偏差は 1 回目と 2 回目の劣化で値の範囲を変える．ここで Cutoff はカーネルサイズが 13 以下の場合は $[\frac{\pi}{3}, \pi]$, 13 以上の場合は $[\frac{\pi}{5}, \pi]$ となる．次にノイズではガウスノイズの標準偏差とポアソンノイズのスケールをぼかし同様に値の範囲を変える．図 6 の生成された劣化画像では劣化の内容が画像によって変化していることがわかる．

表 5　劣化生成パラメータ

劣化内容	項目	1 回目	2 回目
ぼかし	カーネル	[7,21] の奇数	
	回転角	[-π, π]	
	標準偏差	[0.2, 3]	[0.2,1.5]
	Cutoff	$[\frac{\pi}{3}, \pi]$	$[\frac{\pi}{5}, \pi]$
ノイズ	標準偏差	[1, 30]	[1, 25]
	スケール	[0.05, 3]	[0.05, 2.5]
	グレーノイズ	40%	
JPEG	品質係数	[30, 95]	

原画像　劣化画像 1　劣化画像 2

図 6　多様な劣化の例

4.4　損 失 関 数

超解像モデルの学習で使用する損失関数について説明する．損失関数は L1 損失，知覚損失を使用する．

$\mathrm{L1_{LOSS}}$ を式 (1) に示す．ここで $\hat{x}_i$ は i 番目の正解画像，x_i は i 番目の SR 画像を表す．

$$\mathrm{L1_{LOSS}}(\hat{x}, x) = \frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i| \qquad (1)$$

知覚損失を式 (2) に示す．これらの特徴量は，ImageNet で事前学習された VGG19 ネットワークを使用して抽出される．VGG19 の 3 番目，8 番目，および 15 番目の層から特徴量を抽出し，知覚損失を計算する．

$$\text{Perceptual}_{\text{LOSS}}(\hat{x}, x) = \frac{1}{3}\text{MSE}(\text{VGG}_{3,8,15}(\hat{x}), \text{VGG}_{3,8,15}(x)) \quad (2)$$

全体の損失関数を式 (3) に示す．ここで α，β，は各損失関数の係数である．

$$\text{SR}_{\text{LOSS}}(\hat{x}, x) = \alpha\text{L1}_{\text{LOSS}}(\hat{x}, x) + \beta\text{Perceptual}_{\text{LOSS}}(\hat{x}, x) \quad (3)$$

先行研究では損失関数に L1 損失のみを使用して学習を行っていたが本研究では知覚損失を用いて学習を行う．その理由として知覚損失を利用することで出力画像がより人間の感覚に近い画像にできるためである．また知覚損失に VGG19 を使う理由として先行研究 [5] では VGG モデルが知覚損失関数として高い性能を発揮しながらも使用するメモリ容量が同程度の性能を発揮する知覚モデルよりも低いため VGG モデルを選択することを推奨しているためである．

4.5 学習設定詳細

ネットワークのハイパーパラメータについて STB の数を 4，Window サイズを 8，パッチサイズを 4，埋め込み次元を 180，MLP の層の数を 2，CNN のフィルタサイズを 3，損失関数の (α,β) をそれぞれ (1，0.005) とした．

学習時の設定については最適化手法を Adam，学習率を 0.0002，バッチサイズを 16，エポック数を 40 とした．また，GPU は RTX4090 を使用した．

5 実験結果

5.1 では DIV2K での性能比較，5.2 では Set14 での性能比較，5.3 ではブラインド超解像性能比較ついて説明する．

5.1 DIV2K での性能比較

DIV2K 検証用データセット 100 枚を用いて SwinIR との比較を行う．それぞれのモデルの PSNR と SSIM の平均値を表 6 に示す．表より，本研究の提案手法は先行研究の SwinIR と比較して PSNR では約 8.5% ，SSIM では約 5.6% 向上している．

表 6　DIV2K での超解像性能

モデル	PSNR(dB)	SSIM
SwinIR	28.75	0.8501
ours	31.43	0.9010

5.2 Set14 での性能比較

Set14 の超解像ベンチマークデータセットを用いて SwinIR との比較を行う．それぞれのモデルの PSNR と SSIM の平均値を表 7 に示す．表より，本研究の提案手法は SwinIR と比較して PSNR では約 6.7% ，SSIM では約 5.8% 向上している．

また，出力された画像を見ると，SwinIR では画像内の文字の復元がうまくできておらず，記されている文字を読むことが難しいほど崩れている．一方，提案手法では文字を読むことが可能な程度まで復元されている．

表 7　Set14 での超解像性能

モデル	PSNR(dB)	SSIM
SwinIR	26.96	0.8043
ours	28.92	0.8541

HR　SwinIR　ours

図 7　Set14 の超解像結果

5.3 ブラインド超解像性能比較

先行研究で使われていた正解画像が無い超解像データセット RealSRSet+5images を用いて SwinIR との比較を行う．出力された画像を見ると，SwinIR ではレンガ造りの壁が復元できておらず，隣のレンガ同士がくっついたような表現になってしまっている．一方，提案手法ではレンガ造りの壁を表現しながらも低解像度画像 (LR 画像) に比べてノイズも無く，精細な画像になっている．

LR　SwinIR　ours

図 8　ブラインド超解像結果

6 結　　び

本研究では SwinIR をベースに劣化を組み合わせた画像を学習することで現実世界の劣化画像に対応した超解像モデルを作成した．その結果，ベースモデルよりも DIV2K では PSNR は約 8.5%，SSIM は約 5.6% 向上し，Set14 においても PSNR は約 6.7%，SSIM は約 5.8%向上した．

今後の課題は超解像画像全体がぼやけてしまっており，シャープな輪郭の表現ができていないのでシャープマスクを学習画像に適応させるなどをすることでこの問題を解消していきたい．

謝　　辞

本研究は JSPS 科研費 JP17K00432 の助成を受けたものです．

文　　献

[1] Fortune Business Insights,“衛星画像市場の成長，規模，シェア，および地域予測，2023〜2030 年,”https://www.fortunebusinessinsights.com/satellite-imaging-market-103372, 参照 Jun. 2023.
[2] Xintao Wang et al.,“Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data,”IEEE/CVF International Conference on Computer Vision Workshops, pp.1905-1914, Oct. 2021.
[3] Jinsu Yoo et al.,“Enriched CNN-Transformer Feature Aggregation Networks for Super-Resolution,”IEEE/CVF Winter Conference on Applications of Computer Vision, pp.4956-4965, Waikoloa, Hawaii, Jan. 2023.
[4] Jingyun Liang et al.,“SwinIR: Image Restoration Using Swin Transformer,”IEEE/CVF International Conference on Computer Vision Workshops, pp.1833-1844, Oct. 2021.
[5] Gustav Grund Pihlgren et al.,“A Systematic Performance Analysis of Deep Perceptual Loss Networks: Breaking Transfer Learning Conventions,”Computer Vision and Pattern Recognition, arXiv:2302.04032, Oct. 2023.

CNN と Transformer を用いた画像誘導型深度補完システムの実装

大極　康平†　　河野　浩之†

† 南山大学大学院理工学研究科　〒 466–8673 愛知県名古屋市昭和区山里町
E-mail: m22sc001@m.nanzan-u.ac.jp, kawano@nanzan-u.ac.jp

あらまし　深度補完は疎な深度マップと対応する RGB 画像から密な深度マップを予測するタスクである．しかし，深層学習を用いて深度補完を行う場合，Zhang らの CompletionFormer の一番小さいモデルでも 41.5M のパラメータを使用するなど多くの先行研究でパラメータ使用量が多いという課題が存在する．本研究では，PoolFormer と Depthwise Convolution を用いた低パラメータで比較的高精度を出すことができる深度補完ネットワークを提案する．実装には Tensorflow，Keras を使用し，NYU Depth v2 データセットと KITTI Depth Completion データセットの各データセットで学習を行う．NYU Depth v2 データセットの検証用データにおいて，RMSE 0.099 を達成し，パラメータ数 9.8M で 25.84M の NLSPN や 78.3M の CompletionFormer に近い精度を出すことができている．

キーワード　深度補完，深層学習，PoolFormer，Depthwise Convolution

1　自動運転市場と深度補完の必要性

近年，自動運転への注目が高まっている．そして，自動運転には正確な距離情報が必要なので，LiDAR 等の測距センサーの需要も高まっている．株式会社富士キメラ総研による自動運転レベル 2 以上車両の世界市場調査を表 1，LiDAR の世界市場調査を表 2 に示す [1]．

表 1 よりレベル 2 は 2022 年から，レベル 3 以上は 2030 年頃から自動運転市場が拡大している事が分かり，表 2 よりレベル 3 が大きく市場拡大する 2030 年頃から LiDAR の市場も拡大している事が分かる．しかし，LiDAR 等の測距センサーにはセンサーノイズや物体表面の条件，センサーの走査線数の制限により計測結果が疎となる問題があるので，図 1 のように画素値の欠落が存在する深度マップである疎な深度マップを欠落の無い密な深度マップに補完する事が必要である．

本研究では疎な深度マップと対応する RGB 画像を入力とし密な深度マップを予測する深度補完ネットワーク PDCNet を提案する．PDCNet は低パラメータで比較的高精度な予測を目的としており，CNN と Transformer を融合し精度を向上させ，PoolFormer と Depthwise Convolution を活用しパラメータを削減している．

表 1　自動運転車両の世界市場 (生産台数ベース)

	2022 年 見込	2030 年 予測	2045 年 予測
レベル 2	3,608 万	6,176 万	6,166 万
レベル 3	3 万	580 万	2,847 万
レベル 4/5	9 万	433 万	2,051 万

表 2　LiDAR の世界市場

2022 年見込	2030 年予測	2045 年予測
129 億円	9,548 億円	3 兆 5,375 億円

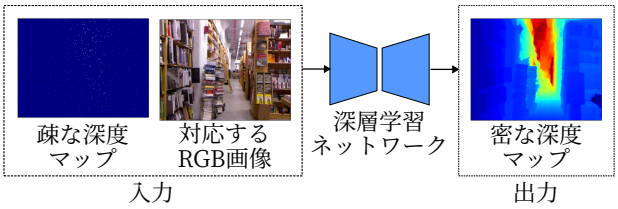


図 1　深度補完の例

2　深度補完の提案手法

2.1 で先行研究の課題と深度補完手法の提案，2.2 で提案する PDCNet，2.3 でネットワークのパラメータ削減について述べる．

2.1　先行研究の課題と深度補完手法提案

Rho らの提案した GuideFormer [2] や Zhang らの提案した CompletionFormer [3] にはパラメータ使用量が多いという課題が存在する．例えば GuideFormer ではパラメータを 100M 以上使用，CompltionFormer では Tiny:45.8M，Small:82.6M，Base:146.7M パラメータを使用している．また，Zhao らの ACMNet [4] は使用パラメータが約 5M と少ないが，CNN とグラフの使用だけでは局所的な領域内でしか特徴量を集約できないという課題やグラフを作成する工程を行わなければいけないのでモデルの計算が複雑になる課題が存在する．

先行研究の課題を踏まえ，本研究ではグラフ作成などの工程を必要としない，低パラメータで比較的高精度を出すことができる深度補完ネットワーク PDCNet(PoolFormer and Depthwise Convolution Network) を提案する．提案手法では CNN と Transformer を合わせたネットワークにより局所的な情報と大域的な情報の両方を獲得する事で高精度を狙い，PoolFormer [5] や Depthwise Convolution(DW Conv) [6] を活用することでパラメータ削減を行う．

2.2 提案する PDCNet

図 2 に PDCNet のネットワーク図を示す．PDCNet では，はじめに，入力された疎な深度マップと対応する RGB 画像を融合し 1 つの特徴量マップへと変換する深度・RGB 埋め込みを行う．次に，作成した特徴量マップを Transformer と CNN を用いたエンコーダに入力し，特徴マップをダウンサンプリングしながら Transformer による大域的な特徴量と CNN による局所的な特徴量を合わせた特徴量を獲得する．最後に，エンコーダで獲得した特徴マップを CNN を用いたデコーダに入力し，特徴マップをアップサンプリングしていくことで密な深度マップを獲得する．

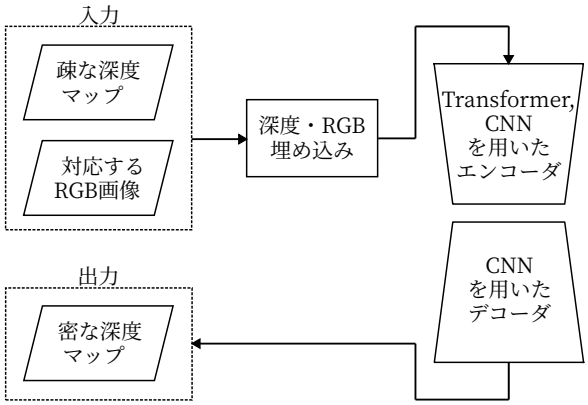


図 2　PDCNet のネットワーク図

2.3 ネットワークのパラメータ削減

本研究ではエンコーダの Transformer 部分で PoolFormer と DW Conv を使用する事でパラメータ削減を行う．

図 3 に ViT Block と PoolFormer Block の比較を示す．図 3 左の ViT Block [7] は画像処理分野で使用される基本的な Transformer モデルであり，画像をパッチへと変換したものを入力とし，Norm 層→ Attention 層→ Norm 層→ MLP 層という流れで特徴量抽出を行う．Norm 層では正規化，Attention 層ではパッチ間の類似度を計算し重要部分へ着目，MLP 層では特徴量の獲得を行う．図 3 右の PoolFormer Block は ViT の Attention 部分を Pooling に置き換えたものであり，Attention をパラメータを使用しない Pooling に置き換えているのでパラメータ削減できる．PoolFormer は，ViT は Attention という特定のモジュールにより精度を出すではなくアーキテクチャ全体が精度に寄与しているという考えを基に作成されており，画像分類などにおいて ViT と同等の精度を出している．

入力された特徴マップと出力される特徴マップのチャネル数を C，フィルタサイズを N とする．通常の畳み込みは $N \times N \times C$ のフィルタを出力特徴マップのチャネル数である C 個用意し，入力特徴マップに対して並列に畳み込み処理を行いそれぞれ特徴マップを取得し，各フィルタで取得した特徴マップをまとめることで出力特徴マップを得る．対して，DW Conv は入力特徴マップのチャネルごとに専用の $N \times N$ のフィルタを用意し，各フィルタで畳み込み処理を行い，フィルタごとに取得した特徴マップをまとめることで出力マップを得る．通常の畳み込みで使用する各フィルタが $N \times N \times C$ に対し，DW Conv は $N \times N$ なので，DW Conv は通常の畳み込みと比べてパラメータが $\frac{1}{C}$ となる．

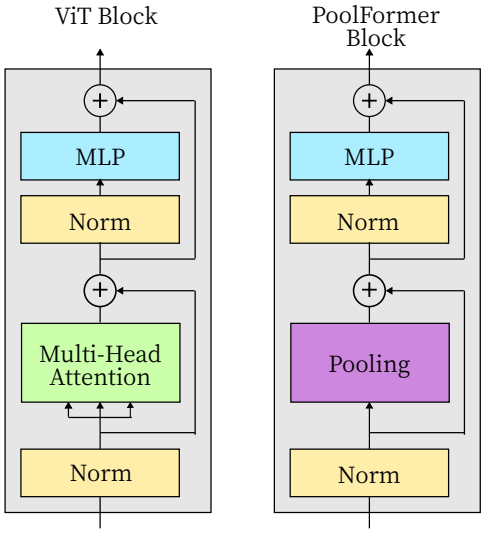


図 3　ViT Block と PoolFormer Block の比較

3 深度補完の実験環境

本研究では，ネットワークの構築に Tensorflow と Keras を使用し，データセットに KITTI DC(Depth Completion) と NYU Depth v2 を使用する．深度補完データセットの特徴を表 3 に示す．KITTI DC は 240x1216 に下部中央で切り取り使用する．NYU Depth v2 は 320x240 にハーフダウンサンプリングし，304x228 にセンタークロップして使用する．また，NYU Depth v2 には疎な深度マップが含まれていない．なので，訓練用データは学習時に密な深度マップから 10%の確率で 200 点，80%の確率で 500 点，10%の確率で 1000 点ランダムサンプリングして疎な深度マップを作成する．検証用データは密な深度マップから 500 点ランダムサンプリングして疎な深度マップを作成する．

表 3　深度補完データセットの特徴

データセット	特徴
KITTI DC	屋外データセットで訓練用に 85K，検証用に 7K，テスト用に 1K，疎なマップに 5 %，密なマップに 16 %の有効点
NYU Depth v2	464 の室内シーンデータセット 訓練用に 50K，検証用に 654 枚

4 PDCNet の詳細

4.1 で PDCNet 全体の構造，4.2 で深度マップ・RGB 画像埋め込み，4.3 でエンコーダの構造，4.4 でデコーダの構造，4.5 で学習条件について述べる．

4.1 PDCNet 全体の構造

PDCNet の全体図を図 4 に示す．PDCNet は，深度マップ・RGB 埋め込み，CBAM ResBlocks と PoolFormer + CBAM DWConv Blocks によって構成されたエンコーダ，CNN Decoder Block によって構成されたデコーダにより構築されている．

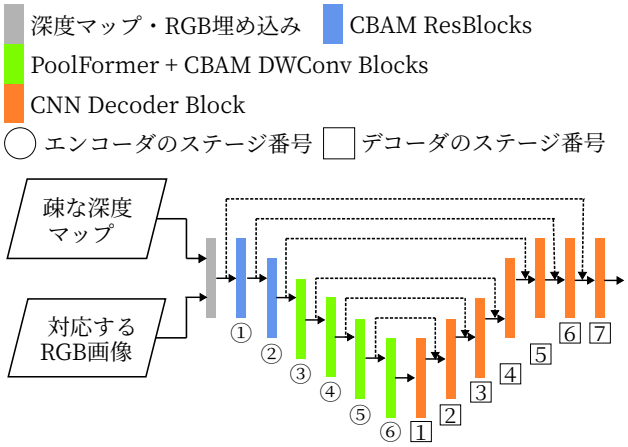


図 4　PDCNet の全体図

4.2 深度マップ・RGB 画像埋め込み

深度マップ・RGB 埋め込みでは，入力された疎な深度マップと RGB 画像を融合し，深度と RGB 両方の情報を持つ特徴マップへと変換する．図 5 に示すように，入力された疎な深度マップと RGB 画像それぞれに畳み込み (Conv) + GroupNormalization + ReLU 関数を使用し，その後それぞれの出力をチャンネル方向で連結し，さらに畳み込み + GN + ReLU 関数を使用することで融合させる．

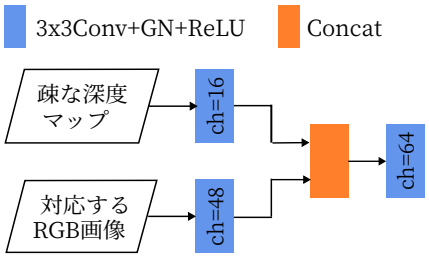


図 5　深度マップ・RGB 画像埋め込み

4.3 エンコーダの構造

エンコーダでは，Transformer と CNN を組み合わせたネットワークにより，Transformer によって獲得できる大域的な情報と CNN によって獲得できる局所的な情報を合わせた特徴量を獲得する．図 4 に示すようにエンコーダは 6 つのステージで構成されており，1, 2 ステージが CBAM ResBlocks，3, 4, 5, 6 ステージは PoolFormer と CBAM DW Conv を合わせたブロックで構成されている．また，本エンコーダは 2 ステージ以降各ステージの最初の層でダウンサンプリングされる．2 ステージ目では，ストライド 2 の CBAM ResBlock によって，3, 4, 5, 6 ステージ目では 2x2 のパッチ分割によってハーフダウンサンプリングされる．

1，2 ステージでは図 6 に示す CBAM ResBlock によって特徴抽出を行う．CBAM ResBlock は，ResNet34 の ResBlock に Convolutional Block Attention Module(CBAM) [8] を組み込んだものである．CBAM を組み込むことにより，ResBlock が獲得した特徴量の中でも特に重要な特徴に焦点を当てる事で精度を向上させることができる．1 ステージ目ではフィルタ数 64 で 3 つ，2 ステージ目ではフィルタ数 128 で 3 つ CBAM ResBlock を使用する．

3, 4, 5, 6 ステージでは，始めに特徴マップを Transformer への入力パッチへと変換するパッチ分割を行う．パッチ分割はストライド 2 の 3x3Conv で実装している．次に図 7 に示す PoolFormer と CBAM DWConv を合わせたブロックへと入力する．PoolFormer と CBAM DWConv を合わせたブロックは，始めに CBAM DWConv によって局所的な特徴を抽出し，次にその特徴マップを PoolFormer Block へと入力し大域的な特徴を抽出する構造となっている．各ステージの次元数に関して，ステージ 3 では 64，ステージ 4 では 128，ステージ 5 では 256，ステージ 6 では 512 とする．各ステージのブロック数に関して，ステージ 3 では 3，ステージ 4 では 3，ステージ 5 では 4，ステージ 6 では 2 とする．

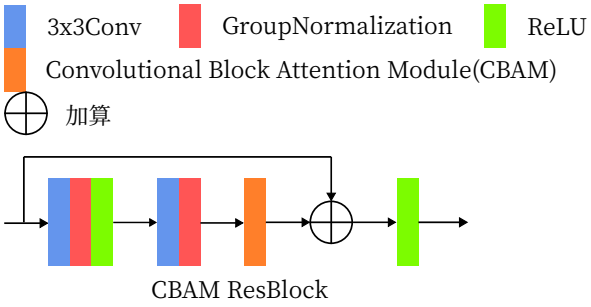


図 6　CBAM ResBlock

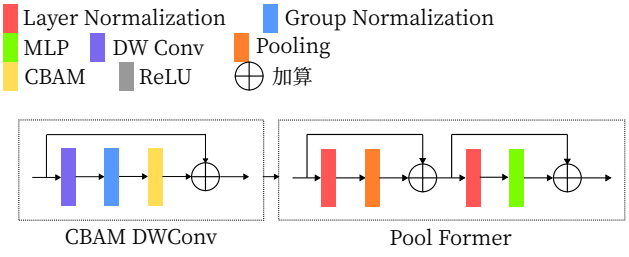


図 7　PoolFormer + CBAM DWConv

4.4 デコーダの構造

デコーダでは，エンコーダから得た特徴マップを基に画像サイズをアップサンプリングしながら密な深度マップを完成させる．デコーダではスキップ接続によってエンコーダからの特徴マップを受け取ることで，エンコーダでのダウンサンプリングの際に失われてしまう情報を追加することができる．

図 4 に示すようにデコーダは 7 つのステージで構成されてお

り，ステージ 1 では図 8(a) のブロックを用いて転置畳み込みによるアップサンプリングと CBAM ResBlock による特徴抽出を行う．ステージ 2, 3, 4, 5 では図 8(b) のブロックを用いて，始めにスキップ接続によってエンコーダの情報を追加し，次に転置畳み込みによってアップサンプリングし，最後に CBAM ResBlock によって特徴抽出を行う．ステージ 6 では図 8(c) を用いて，始めにスキップ接続によってエンコーダの情報を追加し，最後に畳み込みによって特徴量抽出を行う．ステージ 5 の時点で元の画像サイズにアップサンプリングできているのでステージ 6 以降はアップサンプリングは行わない．ステージ 7 では，図 8(d) を用いて，始めにスキップ接続によって深度マップ・RGB 埋め込み後の情報を追加し，最後に畳み込みによって特徴抽出しつつ，チャネル数を深度マップのチャネル数である 1 にする．

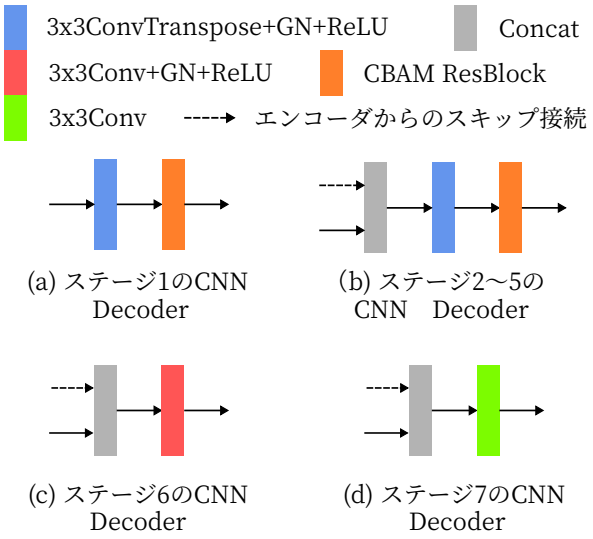


図 8　CNN Decoder Block

4.5　学 習 条 件

PDCNet の学習条件として，最適化手法は AdamW を使用し，初期学習率は 2×10^{-4}，$\beta_1 = 0.9$，$\beta_2 = 0.999$，weight decay は 1.0×10^{-4} とする．さらにデータ拡張として，水平ランダムフリップ，ランダムローテーション，カラージッターを行う．NYU Depth v2 で学習する際はバッチサイズは 3，エポック数は 25 で訓練し，5, 10, 15, 20 エポックで学習率を半分にする．KITTI DC で学習する際はバッチサイズ 3，エポック数は 30 で訓練し，10, 15, 20, 25 エポックで学習率を半分にする．損失関数には二乗平均平方根誤差 (RMSE) を使用する．RMSE の式を式 (1) に示す．

$$\mathrm{RMSE}(\hat{d}, d) = \sqrt{\frac{1}{|P|}\sum_{p \in P}(\hat{d} - d)^2} \tag{1}$$

ここで，$\hat{d}$ は正解深度マップ，d は予測深度マップ，P は正解深度マップの有効画素の集合，$|P|$ は有効画素の数を表す．RMSE は，正解深度マップと予測深度マップの画素値が一致するほど値が 0 に近づく．

5　深度補完結果と先行研究との比較

NYU Depth v2 の検証用データ 654 枚を使用して先行研究との比較を行う．比較に KITTI DC ではなく NYU Depth v2 を使用する理由は，入手できる KITTI DC のテスト用データには正解深度マップが含まれておらず，評価値を得るには公式サイトに提出する必要があるが本研究は査読付き論文であるという提出要件を満たしていないので評価値を求める事が難しいからである．評価指標には二乗平均平方根誤差 (RMSE) と式 (2) に示す相対絶対誤差 (Rel) を用いる．

$$\mathrm{Rel}(\hat{d}, d) = \frac{1}{|P|}\sum_{p \in P}\frac{|\hat{d} - d|}{\hat{d}} \tag{2}$$

ここで，$\hat{d}$ は正解深度マップ，d は予測深度マップ，P は正解深度マップの有効画素の集合，$|P|$ は有効画素の数を表す．Rel は正解深度マップと予測深度マップの画素値が一致するほど値が 0 に近づく．

表 4 に示す先行研究との評価指標による比較より，PDCNet は RMSE と Rel の両方で CSPN を超え，RMSE で ACMNet を超える事が出来ている．NLSPN と CompletionFormer の評価値を超えることはできていないが，PDCNet は NLSPN のパラメータ数 25.84M や CompletionFormer のパラメータ数 78.3M よりも少ない 9.8M のパラメータ数で NLSPN や CompletionFormer に近い評価値を達成する事が出来ている．

図 9 に各手法の深度マップ予測結果を示す．CSPN は左窓の真ん中の枠の部分の深度が滲んでしまっており，NLSPN は消えてしまっている．これに対し，提案した PDCNet は正確に左窓の真ん中の枠の部分の深度を予測することができている．

表 4　先行研究との評価指標による比較

深度補完手法	RMSE(m)	Rel
CSPN [9]	0.117	0.016
ACMNet [4]	0.105	0.015
NLSPN [10]	0.092	0.012
CompletionFormer [3]	0.090	0.012
PDCNet(ours)	0.099	0.015

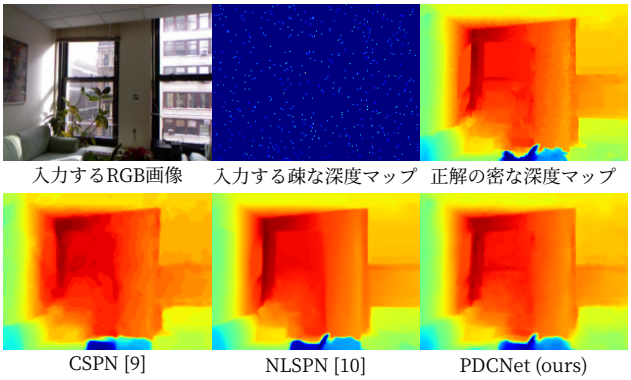
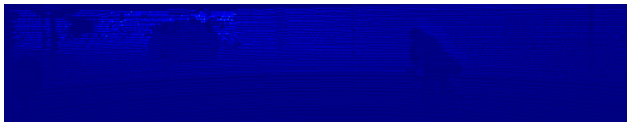


図 9　各手法の深度マップ予測結果

6　屋外データでの予測と使用場面の考察

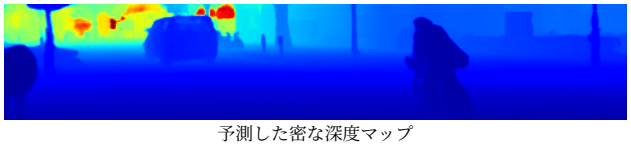
図 10 に KITII DC データセットの検証用データを用いた予測結果を示す．図 10 の予測結果は検証用データを用いた予測の中でも誤差が大きく出たものであり，ここでは誤差が大きく出た場合も考慮し，PDCNet の使用場面を考察する．

表 5 に速度の違いによる停止距離の変化 [11] を示す．図 10 のエラーマップから予測した深度マップは左奥の入り組んだ深度値の大きい部分で誤差が多く発生していることが分かる．特に 50m 以上の場所で大きな誤差が発生している．40m 付近にも誤差は存在するが大半が 1m 以下の誤差であり，50m 以上の部分の誤差と比べ小さい値となっている．そして，表 5 より速度 60km の時の停止距離が 44m であり，提案した PDCNet の予測が 50m 以上で大きな誤差が出て，40m 付近では比較的誤差が小さいことから，PDCNet は時速 60km 以下の自動運転では安全に使用することができると考える．緊急車両や高速道路での運転など時速が 60km を超える場合ではより精度の高い先行研究の手法を用いるのが良いと考える．


入力する疎な深度マップ

入力するRGB画像

正解の密な深度マップ


予測した密な深度マップ

エラーマップ

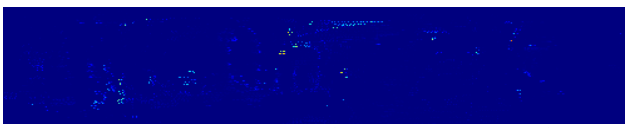

エラーマップの左上部分の拡大（左上から縦に100, 横に500）

図 10　KITTI DC データセットでの深度マップ予測結果

表 5　速度の違いによる停止距離の変化

速度 (km/h)	20	30	40	50	60	70	80	90	100
停止距離 (m)	9	14	22	32	44	58	76	93	112

7　結　　び

PoolFormer と Depthwise Convolution によってパラメータ削減を行った深度補完ネットワーク PDCNet を提案した．PDCNet は CSPN や ACMNet の精度を上回った．また，PDCNet は 9.8M のパラメータ数でパラメータ数 25.84M の NLSPN やパラメータ数 78.3M の CompletionFormer に近い精度を出すことができている．

謝　　辞

本研究は JSPS 科研費 JP17K00432 の助成を受けたものです．

文　　献

[1] 株式会社富士キメラ総研，“自動運転車の世界市場を調査，” 富士経済グループ，https://www.fuji-keizai.co.jp/press/detail.html?cid=22085 参照 Feb. 9, 2024.
[2] Rho et al.，“GuideFormer: Transformers for Image Guided Depth Completion，” Proceedings of the IEEE/CVF Conference on CVPR, pp.6250-6259, New Orleans, USA, Jun. 2022.
[3] Zhang et al.，“CompletionFormer: Depth Completion with Convolutions and Vision Transformers，” Proceedings of the IEEE/CVF Conference on CVPR, pp.18527-18536, Vancouver, Canada, Jun. 2023.
[4] Zhao et al.，“Adaptive Context-Aware Multi-Model Network for Depth Completion，” IEEE Transactions on Image Processing, vol.30, pp.5264-5276, May 2021.
[5] Yu et al.，“MetaFormer Is Actually What You Need for Vision，” Proceedings of the IEEE/CVF Conference on CVPR, pp.10819-10829, New Orleans, USA, Jun. 2022.
[6] Howard et al.，“MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications，” arXiv:1704.04861 [cs.CV], Apr. 2017.
[7] Dosovitskiy et al.，“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale，” 2021 International Conference on Learning Representations(ICLR2021) , pp.1-21, May 2021.
[8] Woo et al.，“CBAM: Convolutional Block Attention Module，” Proceedings of the ECCV, pp.3-19, Munich, Germany, Sep. 2018.
[9] Cheng et al.，“Learning Depth with Convolutional Spatial Propagation Network，” arXiv:1810.02695 [cs.CV], Oct. 2018.
[10] Park et al.“Non-Local Spatial Propagation Network for Depth Completion，” ECCV, pp.120-136, Aug. 2020.
[11] 日本自動車連盟，“走行中の適切な車間距離は？，” 日本自動車連盟, https://jaf.or.jp/common/kuruma-qa/category-drive/subcategory-technique/faq138, 参照 Feb. 12, 2024.

画像復元のための空間方向 Attention

沖津　永享†　　白井　匡人†

† 島根大学総合理工学部知能情報デザイン学科　〒 690–8504 島根県松江市西川津町 1060
E-mail: †s215015@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp

あらまし　本研究では空間方向 Attention による画像復元向けのベースラインモデルを提案する．空間方向 Attention のコストは入力画像の解像度に対して指数関数的に増加する．これにより空間方向 Attention は画像復元のような密なタスクでの応用が困難である．この問題に対処するために，近年の画像復元手法はチャンネル Attention を採用している．しかし，チャンネル Attention にはいくつかの問題点がある．本研究では近年の手法から画像復元に有効なメカニズムを特定する．それらを空間方向 Attention と融合することで，空間方向 Attention によるベースラインモデルを提案する．提案手法をノイズ除去タスクにて評価する．また LAM によって手法の依存距離を解析する．

キーワード　画像復元, Transformer, Attention

1　はじめに

画像復元は劣化画像から綺麗な画像を生成するタスクである．画像復元は医療，防犯，天文など様々な分野に応用することができる．例えば防犯カメラで撮影した画像がブレや雨，ノイズなどで劣化していた場合，劣化要素を除去することで犯人の顔や車のナンバープレートを特定できる可能性がある．近年の画像復元手法は CNN と Transformer によって高精度を達成している．CNN は局所依存と平行移動不変性を獲得する．さらに CNN はフィルターの重みを共有するため低コストで処理できる．しかし，CNN のフィルターは静的である．そのため CNN は入力特徴に柔軟に対応できない．また CNN は長距離の依存関係を捉えることができない．対して，Transformer は適応的な重みと長距離依存を獲得する．ただし，Transformer の Attention はコストが解像度の 2 乗で増加する．これにより，Transformer は画像復元のような密なタスクでの応用が困難である．このように CNN と Transformer は互いに弱点を補完し合う関係である．画像復元の性能向上には局所依存と長距離依存の両方が必要である．

空間方向 Attention の一部を図 1 に示す．図 1 の通り空間方向の Attention では Attention マップのサイズが $(HW)^2$ となる．ここで H は画像の高さ，W は画像の横幅である．これにより，空間方向 Attention のコストは入力画像の解像度に対して指数関数的に増加する．この問題に対処するために，Zamir らの Restormer [5] は空間方向 Attention を転置してチャンネル方向に Attention を行う．Restormer のチャンネル Attention の一部を図 2 に示す．

図 2 の通り，Attention をチャンネル方向に実施することで Attention マップのサイズは C^2 になる．ここで C はチャンネル次元数である．従って, Attention をチャンネル方向に実施することでコストは大幅に削減される．また Chen らの NAFNet [4] や

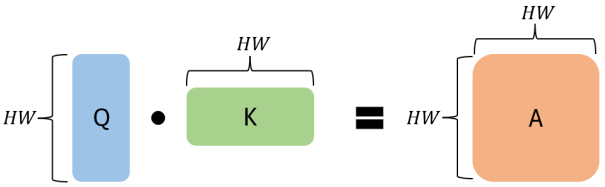


図 1: 空間方向での Attention マップの計算．Q:Query, K:Key, A:Attention マップ, H: 画像の高さ, W: 画画像の横幅

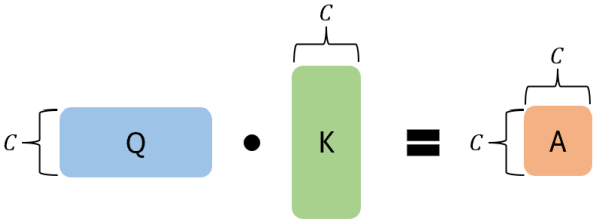
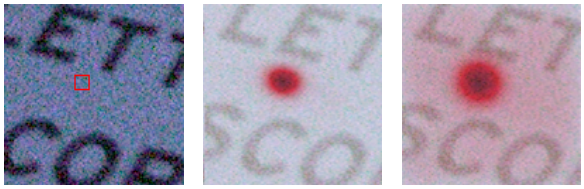


図 2: チャンネル方向での Attention マップの計算．C: チャンネル次元数



復元前　　CA 無し　　CA 有り

図 3: チャンネル Attention(CA) の有無による依存距離の比較．NAFNet を LAM によって解析．左の復元前画像の赤枠範囲を復元するために赤くなっている部分のピクセルが主に活用されたことを示す．

Zhang らの KBNet [3] も Zamir らを踏襲しチャンネル Attention を採用している．これらの手法はチャンネル Attention によって Transformer の強みである長距離依存を獲得する．ここで，チャンネル Attention による依存距離の変化を図 3 に示す．依存距離の解析は Local Attribution Maps (LAM) [2] によって実施する．LAM は勾配積分法を用いた寄与度の解析手法である．本研究では勾配積分法のベースラインとして全画素が 0 の画像を使用する．図 3 の通り，チャンネル Attention を追加することで依存距離が大幅に拡大することがわかる．しかし，チャンネル Attention にはいくつかの問題点がある．最大の問題点は，チャンネル Attention は空間方向 Attention と異なり，空間方向に対する長距離依存を直接捉えていないことである．これにより，収集する特徴の質が低下する可能性がある．また前述の通り，空間方向 Attention は空間方向の情報を全て活用することは困難である．対して，チャンネル Attention はチャンネル方向の全ての情報を容易に活用することができる．言い換えるとチャンネル Attention は空間方向 Attention と比較して改良の余地が少ない可能性がある．

本研究では近年の手法から画像復元に有効なメカニズムを特定する．それらの結果とシンプルな空間方向 Attention を融合する．これにより，空間方向 Attention のベースラインモデルを提案する．提案手法をノイズ除去タスクにて評価する．

本研究の主な貢献は以下のように要約される．

- 近年の手法を解析することによって，画像復元に有効なメカニズムを特定する．それらのメカニズムを用いて，画像復元向けの空間方向 Attention によるベースラインモデルを提案する．
- LAM によって手法の依存距離を解析する．これにより，各手法がどのように復元を行うか分析する．さらに各メカニズムがモデルにどのような効果を与えるか視覚的に分析する．

2 関 連 研 究

a） 画 像 復 元

画像復元はノイズやブレなどで劣化した画像を復元するタスクである．近年の画像復元手法の多くは UNet [11] をベースにしている．UNet はエンコーダー，デコーダーによる階層構造である．階層構造により，モデルは多様な特徴を抽出できる．Zamir らの MPRNet [10] は UNet をマルチステージにする．また Zamir らの MIRNet [8] はエンコーダー，デコーダーの間で複雑なパスを構築する．Fan らの SRMNet [9] はスキップ接続を改良した手法である．Chen らの HINet [6] は Instance Normalization を活用することにより安定した推論を行う．また Transformer の出現によって，画像復元手法は更なる発展を遂げている．Wang らの Uformer [7] は Window Attention を用いた Transformer ベースの手法である．Uformer は Locally enhanced Feed Forward Network（LeFF）によって局所依存を強化する．さらに Uformer は Multi-Scale Restoration Modulator（MSRM）によって多様な画像復元タスクに対応できる．Zamir らの Restormer は Attention をチャンネル方向に実施する．この改良によって Restormer は Attention のコストを大幅に削減している．また Restormer は Gated Dconv Feed Forward Network（GDFN）によって局所依存を強化する．Chen らの NAFNet は複雑な構造や非線形性を除去した非常にシンプルなモデルである．また NAFNet は Restormer に倣いチャンネル Attention を使用する．Restormer と NAFNet は Attention を Depth wise convolution とチャンネル Attention に分割する．Attention と異なり Depth wise convolution は静的なフィルターである．この問題に対して，Zhang らの KBNet は kernel basis attention によって動的なフィルターを実現している．また、これらの Transformer を採用したモデルは近年の State-of-the-Art（SOTA）手法である。

b） Vision Transformers

Transformer [12] は自然言語タスクにて驚異的な精度を達成している．Dosovitskiy らの Vision Transformer（ViT）[13] は画像を連続したパッチに分解する．これにより，ViT は画像タスクでの Transformer の適用を可能にする．ViT のコストはパッチ数の 2 乗で増加する．これにより，ViT は高解像度の画像やセマンティックセグメンテーションなどの密な画像タスクでの適用が困難である．従って，ViT のコスト削減は重要な研究課題である．Liu らの Swin Transformer [14] は画像を window に分割することによってコストを削減する．Tolstikhin らの MLP-Mixer [15] は Attention を MLP に置き換える．Zhang らの Nested Transformer [16] は隣接したパッチをブロックにまとめる．Yang らの Focal Transformer [17] は近くのパッチを細かく，遠くのパッチを粗く生成する．Ramachandran らの Stand-Alone Self-Attention [18] は CNN の convolution を Attention に置き換える．

3 提 案 手 法

3.1 ブロック内構造

提案手法では空間方向の Attention を採用する．空間方向 Attention では入力画像の解像度がコストに大きな影響を与える．空間方向 Attention の計算量を式 1 に示す．

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

ここで，MSA は Multi head Self Attention，h,w は画像の横と縦の長さ，C はチャンネル次元数である．式 1 の通り，空間方向 Attention はコストが入力画像の解像度の 2 乗で増加する．よって，空間方向 Attention を画像復元のような密なタスクにそのまま適用するのは困難である．従って，提案手法ではコストを削減するために Swin Transformer の Window Attention を利用する．Window Attention は画像を $M \times M$ の window に分割する．そして，各 window ごとに独立して Attention を実施する．Window Attention での Attention 計算の一部を図 5 に示す．図 5 の通り，Window Attention での Attention マップのサイズは M^4 となる．さらに Attention マップのサイズは入力画像の解像度から影響を受けない．Window Attention の計算量を式 2 に示す．

$$\Omega(W\text{-}MSA) = 4hwC^2 + 2M^2hwC \tag{2}$$

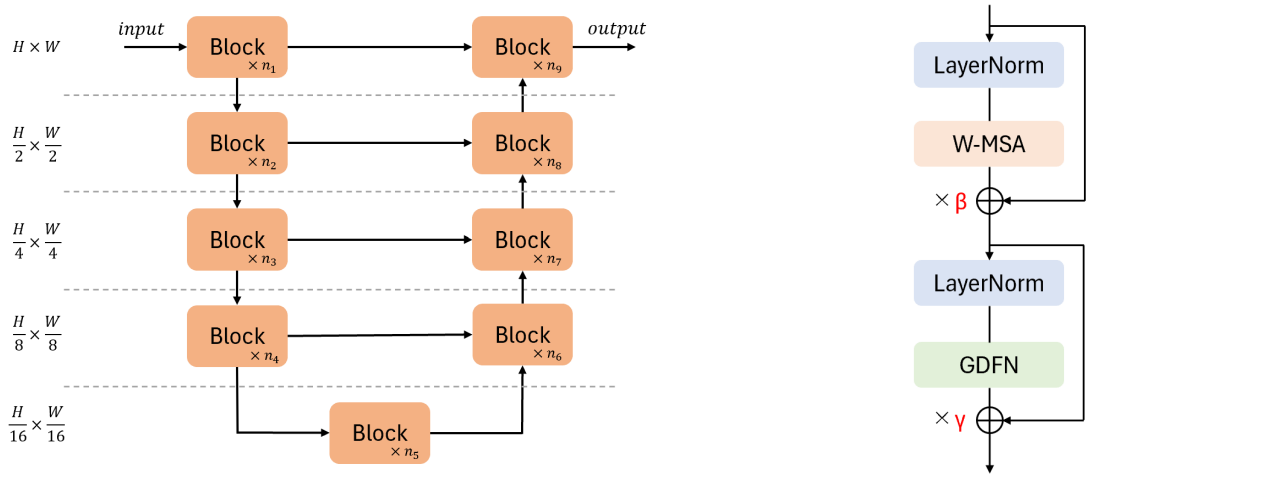


図 4: 提案手法のブロック間構造 (左) とブロック内構造 (右). ⊕：要素ごとに足し算, LayerNorm：Layer Normalization, W-MSA：Window Multi-head Self-Attention, GDFN：Gated Dconv Feed-Forward Network, β, γ：スキップ接続時の重み

ここで，W-MSA は Window based Multi head Self Attention，M は window のサイズである．式 2 の通り，Window Attention のコストは入力画像の解像度に対して線形増加となる．さらに Window Attention は window をシフトさせることによって，window 間での特徴共有が可能である．また Window Attention では，モデルは学習可能な相対位置バイアスによって位置情報を獲得する．相対位置バイアスは Attention の際に各 window に加算される．相対位置バイアスを用いた Attention は次のように定式化される．

$$Attention(Q, K, V) = SoftMax(QK^T / \sqrt{d} + B)V$$

ここで，Q,K,V は query,key,value，d は query と key の次元数，B は相対位置バイアスである．

画像復元では局所依存も重要である．従って， Feed Forward Network に Restormer の Gated Dconv Feed Forward Network（GDFN）を採用する（図 6）．GDFN では 3 × 3 の Depth wise convolution によってモデルは局所依存を強化できる．さらに，ゲート機構と非線形活性化関数の GELU によってモデルはより高度な特徴を抽出できる．

さらにブロック内でのスキップ接続時に学習可能な重みを乗算する．これは NAFNet，KBNet で採用されており，モデルの性能を大きく向上させる．最終的に提案手法のブロック内の構造は次のように定式化される．

$$\begin{aligned} \hat{z}^l &= z^{l-1} + W\text{-}MSA(LN(z^{l-1})) \cdot \beta \\ z^l &= \hat{z}^l + GDFN(LN(\hat{z}^l)) \cdot \gamma \\ \hat{z}^{l+1} &= z^l + SW\text{-}MSA(LN(z^l)) \cdot \beta \\ z^{l+1} &= \hat{z}^{l+1} + GDFN(LN(\hat{z}^{l+1})) \cdot \gamma \end{aligned} \tag{3}$$

ここで，$\hat{z}^l$ と z^l はそれぞれ l ブロック目での (S)W-MSA と GDFN の出力特徴，SW-MSA はシフト Widnow による Multi head Self Attention，LN は Layer Normalization [19]，β, γ は学習可能な重みである．モデルのブロック内構造の詳細を図 4(右) に示す．

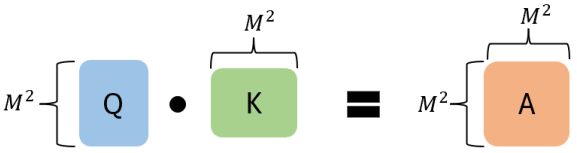


図 5: Window Attention での Attention マップの計算．M:Window サイズ

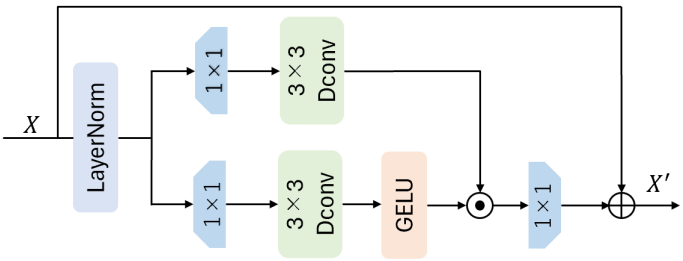


図 6: GDFN：Gated Dconv Feed-Forward Network, Dconv：Depth-wise convolution, ⊙：要素ごとに乗算

3.2 ブロック間構造

モデルのブロック間構造は NAFNet に倣いシングルステージの UNet を採用する．エンコーダーの各層の出力はデコーダーの対応する層にスキップ接続される．特徴マップのサイズは各階層で半分になる $((H \times W), (\frac{H}{2} \times \frac{W}{2}), (\frac{H}{4} \times \frac{W}{4}), \ldots)$．提案手法の UNet は 5 つの階層で構成される．各階層は 1 つ以上のブロックで構成される．提案手法のブロック間構造の詳細を図 4(左) に示す．

4 実　　験

4.1 設　　定

提案手法をノイズ除去タスクにて評価する．データセットに

は SIDD [1] を使用する．SIDD はスマートフォンによって撮影されたノイズ画像と綺麗な画像とのペア画像集である．学習の設定はバッチサイズ 16，エポック数 20，損失関数 L1Loss，オプティマイザー AdamW，初期の学習率は 3e-4 で cosine annealing [20] によって徐々に 1e-6 まで減衰とする．また学習画像のサイズはランダムに 128×128 で Crop する．Window Attention の都合上，テスト画像も中心を基準に 128×128 で Crop する．提案手法のブロック構成は NAFNet と同じものを採用し，埋め込み次元数は 32 とする．評価指標には PSNR を使用する．

$$PSNR = 10 \times \log_{10} \frac{MAX^2}{MSE}$$

ここで，MAX は画像が取り得る最大画素値，MSE は平均二乗誤差である．PSNR は数値が高いほど高精度となる．

4.2 他手法との比較

まず他手法との比較実験の結果を示す．ここで全モデルのブロック数を 36 とし，埋め込み次元数を調整することでモデルの設定を統一する．また各モデルは 10 回ずつ学習を実施し，PSNR の平均値を算出する．結果を表 1 に示す．ここで，Ours1 と Ours2 はそれぞれ window size 8×8 と 16×16 である．実験結果から提案手法はほとんどの手法の精度を上回っていることがわかる．さらに，Ours2 は最高精度となっている．ここで，Ours2 は学習速度とメモリコストの両方で Restormer より優れている．各手法の復元結果を図 7 に示す．

4.3 アブレーション実験

次にいくつかのアブレーション実験の結果を示す．

a） window size

様々な window size での比較実験の結果を表 2 に示す．ここで，window size が大きいとメモリコストが膨大になる．この問題に対処するために，この実験ではバッチサイズを 14 に設定する．従って，表 2 の結果は表 1 とは異なっている．実験結果から window size が大きければ大きいほど精度は向上することがわかる．

b） LeFF vs GDFN

Wang らの Locally enhanced Feed Forward Network（LeFF）と Zamir らの Gated Dconv Feed Forward Network（GDFN）は局所依存を強化するために convolution を組み込んだ Feed Forward Network である．2 つの比較実験の結果を表 3 に示す．実験結果から GDFN は精度とコストの両方で LeFF より優れていることがわかる．

c） スキップ接続時の重み

スキップ接続時の重みの効果を実験する．結果を表 4 に示す．実験結果からスキップ接続時の重みは僅かな追加コストで精度を大きく向上させることがわかる．

d） MSRM の効果

Wang らの Multi Scale Restoration Modulator（MSRM）は画像復元の精度を向上させる．実験によって提案手法での MSRM の効果を評価する．結果を表 5 に示す．実験結果から精度は変化しないことがわかる．従って提案手法では MSRM は採用していない．しかし，ブレ除去や雨除去などの他の画像復元タスクでは MSRM は良い効果をもたらす可能性がある．

e） Depth wise convolution の効果

Restormer や NAFNet はチャンネル Atteintion の実施前に Depth wise convolution (Dconv) を実施している．空間方向 Atteintion の実施前に Dconv を実施すると，精度にどのような影響を与えるか実験する．結果を表 6 に示す．結果から精度は向上しないことがわかる．Restormer 等は Window Attention と

Method	PSNR	GPU memory cost(MB)	Training time per epoch(m)
Uformer	39.46	14848	12
KBNet	39.47	10676	17
NAFNet	39.49	5900	5
Restormer	39.51	19676	10
Ours1	39.50	10568	7
Ours2	39.53	12832	9

表 1: 他手法との比較実験の結果．Uformer：window size=8×8, Ours1：window size=8×8, Ours2：window size=16×16

window size	PSNR	GPU memory cost(MB)	Training time per epoch(m)
8×8	39.51	9654	8
16×16	39.54	11348	9
32×32	39.55	18580	15
64×64	39.57	45190	37

表 2: 様々な window Size での比較実験の結果

FFN	PSNR	GPU memory cost(MB)	Training time per epoch(m)
LeFF	39.44	12168	10
GDFN	39.50	10568	7

表 3: LeFF と GDFN の比較実験の結果

重み	PSNR	GPU memory cost(MB)	Training time per epoch(m)
×	39.39	10190	7
√	39.50	10568	7

表 4: スキップ接続時の重みの有無による精度差

MSRM	PSNR	GPU memory cost(MB)	Training time per epoch(m)
×	39.50	10568	7
√	39.50	10660	7

表 5: MSRM の有無による精度差

Dconv	PSNR	GPU memory cost(MB)	Training time per epoch(m)
×	39.50	10568	7
√	39.44	11482	8

表 6: Depth wise convolution の有無による精度差

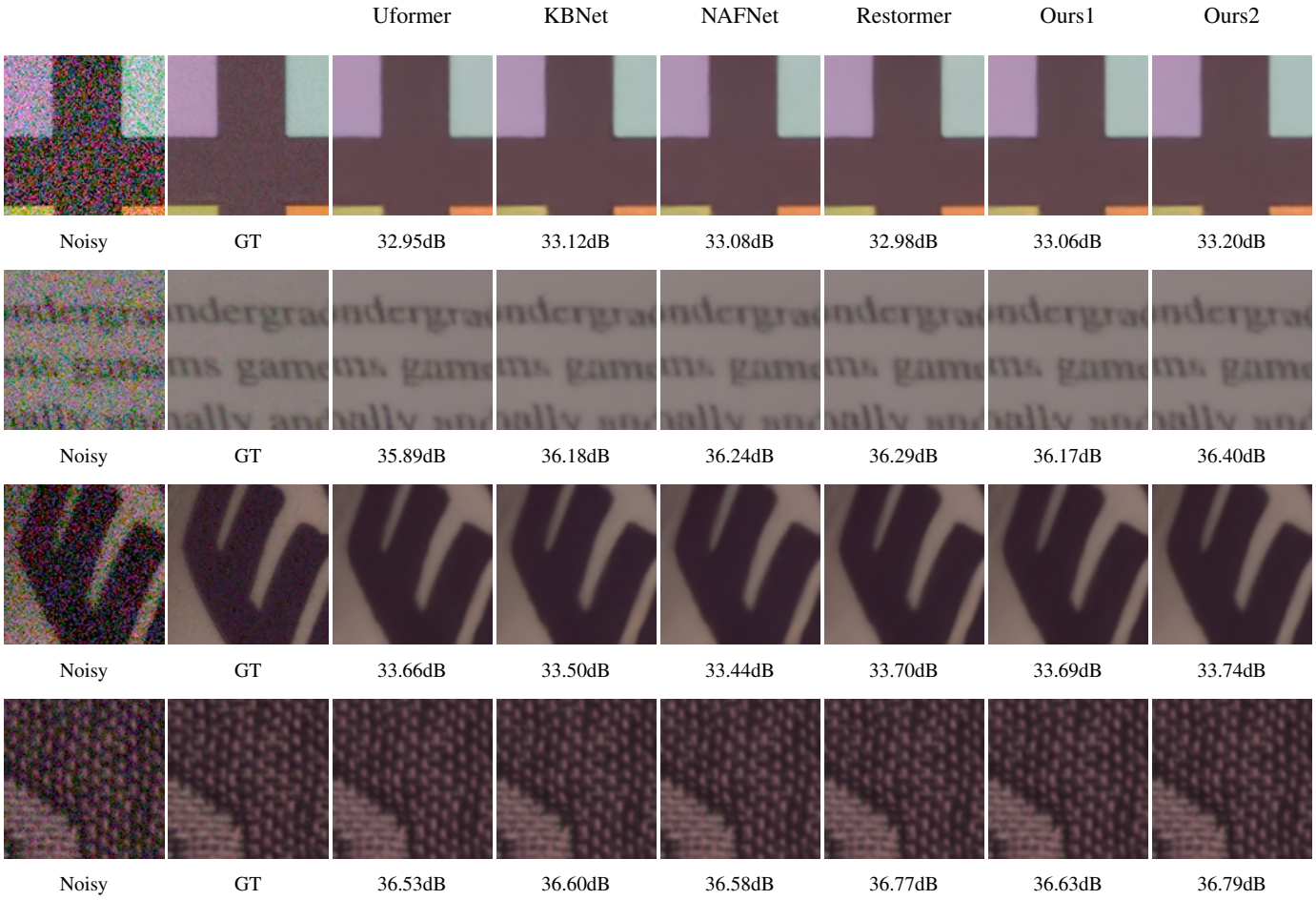


図 7: 復元結果の比較（指標：PSNR）

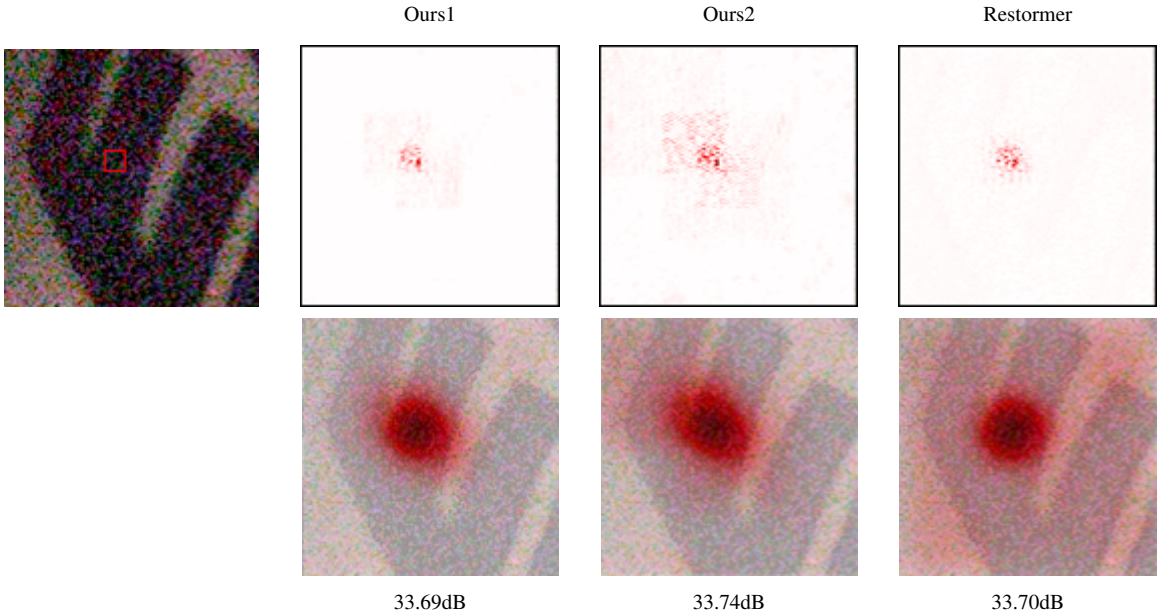


図 8: LAM による提案手法の解析結果

Dconv がいくつかの点で酷似していることとチャンネル Atteintion がチャンネル方向に長距離依存を捉えることから，空間方向 Atteintion を Dconv とチャンネル Atteintion に分解している．従って，実験結果も考慮すると空間方向 Atteintion には Dconv は必要ないことがわかる．

5 考　　察

ここでは LAM での解析によるいくつかの分析結果を示す．

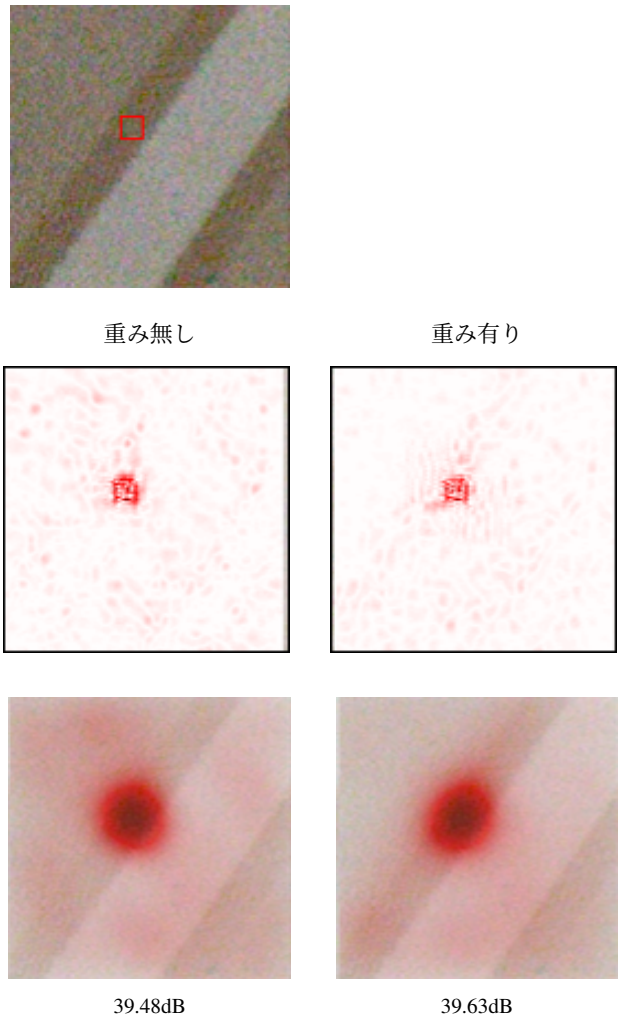


図 9: LAM によるスキップ接続時の重みの解析

a） 提案手法の依存距離

提案手法の依存距離を解析する．結果を図 8 に示す．比較のために Restormer の解析結果も示す．図の下段から Restormer の依存距離が最大となっていることがわかる．しかし，Ours2 の方が高精度である．ここで，図 8 の上段から Ours2 は中距離の範囲が濃く反応していることがわかる．対して Restormer は局所的にしか強い反応が見られない．この結果からチャンネル Attention より空間方向 Attention の方が空間コンテキストを上手く活用できることがわかる．

b） スキップ接続時の重みの効果

スキップ接続時の重みの効果を解析する．結果を図 9 に示す．図より重み無しだと全体に依存が見られるのに対して，重み有りだと左上の依存が弱くなっていることがわかる．式 3 から重みは関数の出力に乗算されている．従って，この重みは関数の出力を調節していることがわかる．この調節によって不要な情報への依存を抑制したり重要な情報への依存を強化したりしている可能性がある．

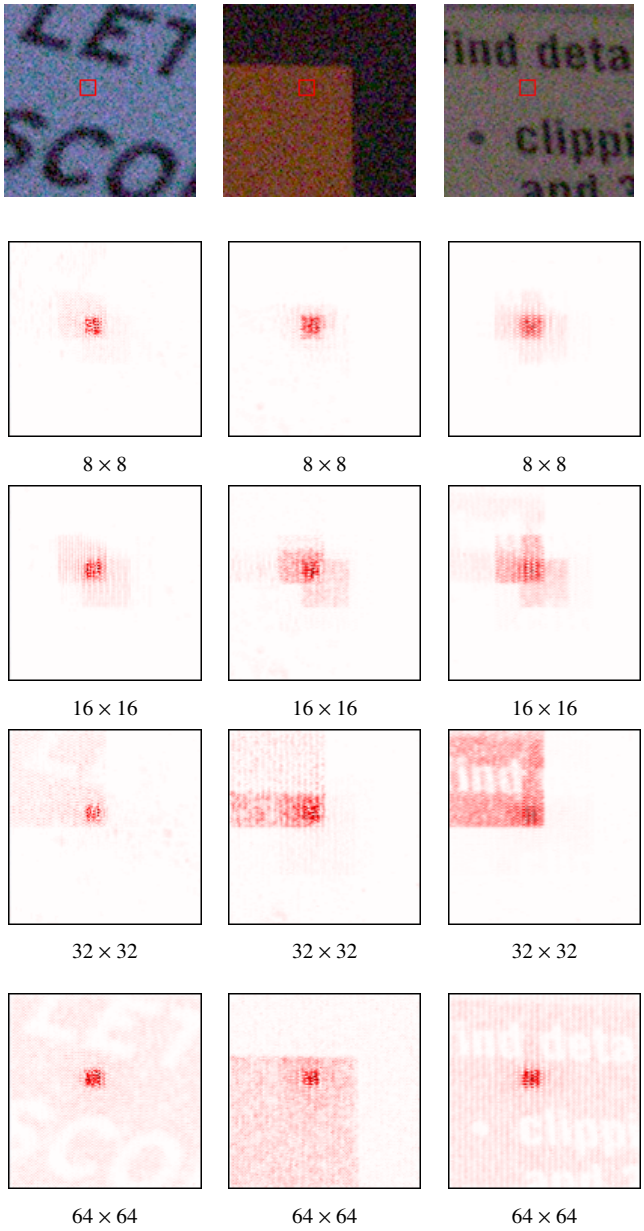


図 10: Window Size ごとの依存距離の解析結果

c） Window Size ごとの依存距離

window size ごとの依存距離を解析する．結果を図 10 に示す．図から window size が大きくなるほど依存距離が拡大することがわかる．さらに，window size が大きいほど依存箇所が模様のようになっている．これは復元に関連度の高い情報を活用していることを示している．

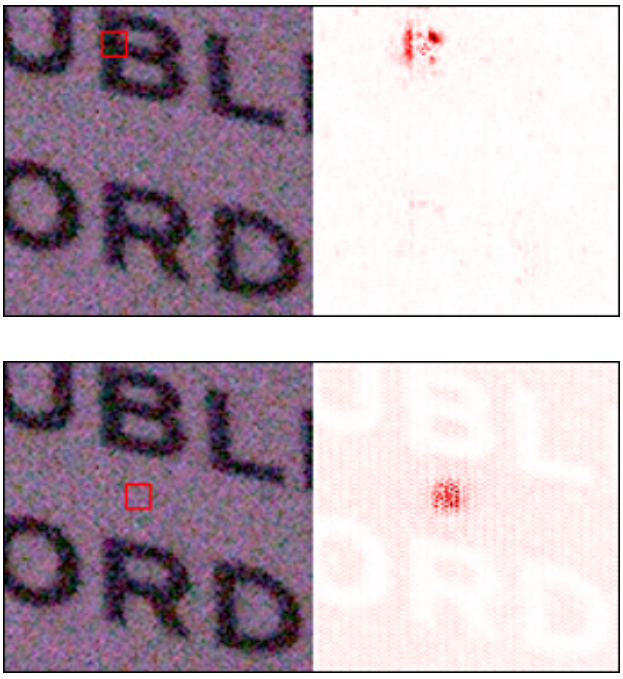


図 11: LAM による依存箇所の解析

d） 提案手法の依存箇所

提案手法の依存箇所を解析する．依存箇所がわかりやすいように window size 64 × 64 のモデルを使用する．結果を図 11 に示す．図 11 の上側は文字，下側は背景の一部を復元する際の解析結果である．図から依存箇所は復元対象との関連度が高い部分になっていることがわかる．

6 結　　論

本研究では画像復元向けの空間方向 Attention によるベースラインモデルを提案した．提案手法には近年の手法で提案された画像復元に有効なメカニズムを組み込んだ．ノイズ除去タスクでの実験では提案手法は近年の SOTA 手法を凌駕した．また LAM によって各手法の依存距離を解析した．実験結果と解析結果から各メカニズムが提案手法にどのような影響を及ぼすか分析した．提案手法をベースラインとして空間方向 Attention を改良することにより更に高性能なモデルを開発できる可能性がある．

文　　献

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In CVPR, 2018.
[2] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In CVPR, 2021.
[3] Yi Zhang1, Dasong Li1, Xiaoyu Shi1, Dailan He1, Kangning Song2, Xiaogang Wang1, Hongwei Qin2 and Hongsheng Li1 KBNet: Kernel Basis Network for Image Restoration. arXiv preprint arXiv:2303.02881 (2023）
[4] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In ECCV, 2022.
[5] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In CVPR, 2022.
[6] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, Chengpeng Chen. HINet: Half Instance Normalization Network for Image Restoration. arXiv preprint arXiv:2105.06086 (2021)
[7] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. In CVPR, 2022.
[8] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, Ling Shao. MIRNet: Learning Enriched Features for Real Image Restoration and Enhancement. In ECCV, 2020.
[9] Chi-Mao Fan, Tsung-Jung Liu, Kuan-Hsien Liu. SRMNet: Selective Residual M-Net for Real Image Denoising. arXiv preprint arXiv:2203.01296 (2022)
[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, Ling Shao. MPRNet: Multi-Stage Progressive Image Restoration. In CVPR, 2021.
[11] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234 – 241. Springer (2015)
[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In NeurIPS, 2017.
[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In ICLR, 2021.
[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021.
[15] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy. MLP-Mixer: An all-MLP Architecture for Vision. In NeurIPS, 2021.
[16] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan O. Arik, Tomas Pfister. Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding. In AAAI, 2022.
[17] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. In NeurIPS, 2021.
[18] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens. Stand-Alone Self-Attention in Vision Models. In NeurIPS, 2019.
[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. arXiv preprint arXiv:1607.06450, 2016.
[20] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In ICLR, 2017.

DEIM Forum 2024 T4-A-1-05

レシピの手順説明画像認識における食材視覚特徴分析

張　訳心† 山肩　洋子†† 田島　敬史†

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† 東京大学大学院情報理工学系研究科　〒 113–8654 東京都文京区本郷７丁目３番１号
E-mail: †zhangyx@dl.soc.i.kyoto-u.ac.jp, ††yamakata@mi.u-tokyo.ac.jp, †††tajima@i.kyoto-u.ac.jp

あらまし　本研究では，マルチメディアレシピの手順画像から食材を認識する精度を向上させるため，レシピ各段階での食材の視覚的特徴を分析し，以下の仮説を提案し検証する．(A) 調理過程の中間段階での食材認識は，前半や後半に比べて困難である．初期段階では食材が原形を保ち，後半では完成品に近いが，中間段階では加工されている状態が認識を難しくする．(B) 調理手順は直列に記述されているが，各手順が前の手順から続くとは限らず，複数の並列なサブタスクが存在する可能性がある．(C) 使用食材リストが付けられたレシピの画像からの食材認識では，リスト内のどの食材かさえ判定できれば十分だが，外観が似ている食材は通常同じ機能を持ち，同じレシピで同時に使用されることが少ないため，これらの食材間の区別は一般的に必要ではない．(D) レシピの最終手順に全食材が含まれる特性を利用し，前の手順での食材の存在を推定し，精度を向上できる．これらの仮説を提案し検証することにより，マルチメディアレシピ手順データにおける困難な点とその解決の手がかりを示し，食材推定の精度を向上させ，効果的なモデルを提案することを目指す．

キーワード　レシピ・料理，マルチメディア，画像認識，食とレシピ

1　はじめに

現代の料理習慣において，レシピが重要な役割を果たしている．自動食品調理とレシピ分析への関心の高まりにより，料理レシピの解釈が可能な高度な計算モデルへの需要が生じている．ウェブ上のレシピでは，テキストに加えて画像や動画が用いられ，料理の手順を分かりやすく示している．しかし，これらのマルチメディアコンテンツから食材を正確に認識することは依然として難しい課題である．現在のマルティメディアコンテンツによるレシピ表現の能力に限界があり，正確な解釈に不可欠な視覚的コンテキストが欠けている．正確な食材認識は，レシピの自動生成，栄養計算，さらには個人化された食事提案など，幅広い応用に影響を及ぼす．したがって，この分野の研究は非常に価値がある．

現在のレシピにおける食材認識に関する研究は，テキストベースのレシピデータセット，例えば Recipe 1M+ [9] などに主に基づいており，テキストベースのアプローチに依存しているしかし，テキスト情報だけでは，食材の視覚的特徴や調理過程における変化を完全に捉えることはできない．加えて，料理過程の中間段階での食材の変化や，複数のサブタスクの存在など，視覚データからの食材認識を難しくする要因が多く存在する．これらの課題を解決するためには，視覚データに基づく新しいアプローチが必要である．

本研究の目的は，マルチメディアレシピの手順画像から食材をより精度高く認識する方法を開発するため，レシピの手順説明画像認識における食材視覚特徴の分析を行うことである．私たちは以下の四つの仮説を提案し，それぞれについて実験を行った．

1.1　調理過程の中間段階での食材認識の困難さ（図 1(a)）

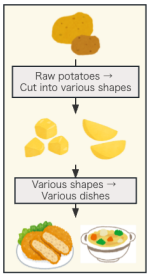
調理の初期段階では，レシピの写真に映る食材は大抵その原形を保っており，識別が比較的容易である．対照的に，調理の最終段階では食材は完成品に近い状態となり，料理の特徴がモデルにとって見つけやすい．これは，多くの事前訓練されたモデルが完成品の写真で訓練されているためである．しかし，調理の中間段階では，食材が頻繁に加工され変形するため，識別が難しくなる傾向にある．本研究では，この段階での食材の視覚的特徴を分析し，この仮説の検証を行う．

1.2　調理手順の構造とサブタスクの存在（図 1(b)）

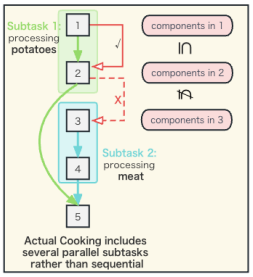
レシピの調理手順は一般に直列的なシーケンスとして記述されるが，必ずしも各手順が前の手順に直接続くとは限らない．実際には，複数のサブタスクが並行して存在することがある．例えば，一つの材料を調理しながら同時に別の材料を準備する場合などがこれに該当する．この研究では，このような調理手順の構造とサブタスクの存在を分析し，食材認識におけるその影響を理解することを目指す．

1.3　食材リストと視覚データとの関連性（図 1(c)）

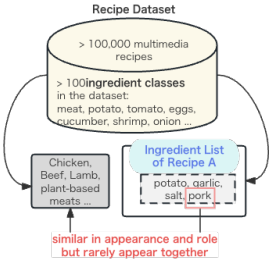
レシピに紐づいた食材リストは，食材認識における重要な手がかりを提供する．本研究では，食材リストと視覚データの関連性に焦点を当て，似た外観の食材を識別するための精度を向上させる．特に，外観が似ている食材が同一レシピで同時に使用されることは稀であるという特性を活かし，レシピの食材リストに基づいて，似ている食材の中から実際にレシピに含まれる食材を識別する．例えば，牛肉と羊肉が視覚的に似ている場合でも，食材リストに牛肉のみが含まれていれば，その情報を使って正確に識別できる．このアプローチにより，食材リスト



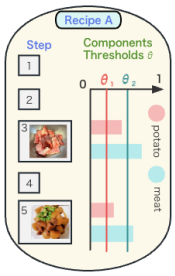
(a) 調理過程の中間段階での食材認識は，前半や後半に比べて困難である．



(b) 調理手順は直列に記述されているが，複数のサブタスクが存在する．



(c) 外見と役割が似たような食材は区別できなくても，片方しか食材リストにはいっていなければ，食材リストを参照することで判別できる



(d) 各調理手順における食材の数は不確定だが，最終手順ではすべてが含まれる．

図 1: マルチメディアレシピデータにおける食材認識の困難な点とその解決の手：(a) 調理過程の中間段階での食材認識は，食材が頻繁に加工され変形するため，前半や後半に比べて困難である； (b) 調理プロセスはしばしば単一の連続ではなく平行したサブタスクから成り，多くの手順が前の手順から内容を受け継ぐが，すべてがそうではないため，ある調理手順の材料リストが前の手順のものの上位集合になることもあればそうでないこともある； (c) 外観が似ている食材は通常， 同じ機能を持ち， 同じレシピで同時に使用されることが少ない；そして (d) 各調理手順における食材の数は不確定だが，レシピの最後の画像ではすべての食材が含まれており，認識のための適応的な閾値が必要である．

を利用して視覚データから似た食材を効率的に区別することが可能になる．

1.4 レシピの最終手順に基づいて食材の存在を推定する方法（図 1(d)）

レシピの最終手順には，そのレシピで使用される全ての食材が含まれることが多い．この特性を利用し，レシピの初期または中間段階での食材の存在を推定する方法を検討する．このアプローチにより，調理過程全体を通じて食材認識の精度を向上させることが期待される．また，最終手順で現れる食材の特徴を分析し，それを以前の手順に適用することで，より効果的な食材認識モデルを開発することを目指す．

本研究では，レシピデータにおける食材の視覚的な特徴を分析し，これらの仮説を検証することで，食材認識の精度を向上させることを目指す．この分析を通じて，より効果的なマルチメディアレシピ解析モデルを開発する．このモデルは，調理の各段階における食材の特徴をより正確に把握し，レシピの解析と理解を深めることに貢献することを目的としている．

2 関 連 研 究

2.1 マルチモーダルレシピ分析

マルチメディアレシピデータの増加は，マルチモーダル分析に対する関心を刺激している．料理画像からの食品カテゴリ認識 [6]， [5]， [18]， [15] や調理手順グラフの生成 [11]， [7] など，さまざまなトピックについての研究がなされている．画像からレシピテキストを検索または生成する関連研究もある [1]， [16]， [13]， [14]， [17].

食品カテゴリ認識 [6]， [5]， [18]， [15] は重要な注目を集めており，食事管理などのアプリケーション向けに画像から料理を識別することを目指している．深層学習や人手で定義された特徴を含むさまざまなアプローチが提案されている．

しかし，画像内の材料を認識することは，さらなる課題を提示する．材料認識におけるアプローチには，マルチラベル学習 [1] やクロスモーダル学習 [2] が含まれ，ここでは画像とレシピが共有の埋め込み空間にマッピングされる．Chen ら [3] は，食品と材料の認識を組み合わせたマルチタスクアプローチを提案している．

これらの研究は，最終的な料理の画像のみに焦点を当てており，調理過程の中間の画像は無視されている．これは，中間の手順の画像を含む公開データセットが不足しているためである可能性もある．本研究では，調理過程を通じた材料の認識の課題に対処する．

調理手順グラフの生成は，手順間での材料の継承を検出することに関連している．代表的な方法 [11] はテキスト記述に依存しており，マルチメディアレシピにおける調理手順グラフの生成はあまりない．別の研究 [7] は，調理行動の暗黙の因果効果を予測する分析に焦点を当てている．我々の研究は手順間の材料の継承に焦点を当てているが，彼らの研究は材料の継承にかかわらず，手順実行の部分的な順序制約を考慮している．

2.2 カスタム損失関数

カスタム損失関数は，さまざまなタスクにおけるユニークな課題に取り組む．例えば，Focal Loss [8] は，難しい例を優先することで密集したオブジェクト検出におけるクラスの不均衡を軽減し，標準のクロスエントロピー損失と比較して性能を向上させる．同様に，Dice Loss [10] は，特に希少で不均衡なデータにおいて医療画像のセグメンテーションにおけるクラスの不均衡を効果的に対処する．

一方，マルチメディアレシピにおける材料認識の課題は，クラスの不均衡というよりは，レシピごとに関連付けられた材料リストというデータセット内の限られた候補クラスの存在にあ

る．各レシピには，データセット全体のクラスの中から選ばれたごく少数の材料が含まれており，この特徴を利用することで認識タスクを効率化することが目的である．我々提案するマスク損失関数は，このユニークなデータ構造を活用するように設計されている．具体的には，レシピの材料リストに含まれないクラスに対する損失をマスキングすることで，そのレシピで重要ではない，または視覚的に似ているが区別する必要のない材料クラスに対するペナルティを避ける．このアプローチにより，モデルはレシピに特定された材料に焦点を当てることができ，同時に同じレシピに現れないような，視覚的に似ている材料間の区別に関しては，ある程度の柔軟性を持たせることができる．これにより，材料認識の精度を向上させると同時に，タスクにおける計算資源の効率的な使用も促進する．

2.3 レシピ関連データセット

過去 10 年間で，さまざまな領域の研究用にいくつかのレシピデータセットが開発されている．表 1 は，主要な公開レシピおよび食品データセットをまとめている．これらのデータセットのほとんどは，完成した料理の画像に主に焦点を当てており，調理過程の画像は含まれていない．例えば，よく知られた Recipe 1M+データセット [9] はテキスト指示を含むが，中間の画像は欠けている．

我々のデータセットのテキスト記述と画像は，それぞれ NII Cookpad データセット [12] と Cookpad 画像データセット [4] から取得されている．

Cookpad 画像データセットには，1, 642, 450 の料理画像と，日本のレシピからの 3, 105, 594 の調理過程の画像が含まれており，多くは 5〜10 の調理手順を持っている．しかし，すべての調理手順に画像が添付されているわけではない．表 1 にあるデータセットとは対照的に，我々のデータセットは，各調理手順に対して包括的なマルチメディア指示（テキストと画像）を提供する．完全なマルチメディア指示を持つレシピデータセットは，調理手順の理解と実行をより容易にする．そのため，Cookpad 画像データセットから各調理手順にテキストと画像があるレシピを選択し，それに注釈を付けることで，このような包括的なマルチメディア指示を持つレシピデータセットの構築を目指している．

3 構築したレシピデータセット

本章では，私たちが構築したレシピデータセットの詳細なプロセス，データセットの統計情報，そしてアノテーション方法について紹介する．これらのアノテーションは，単に可視的な食材だけでなく，コンテキストと画像から推測される食材（以前の手順で言及された食材など）も識別するために追加された．これにより，調理プロセスの理解を深めるためのアノテーションの重要性を強調し，データセットの価値を明確にする．

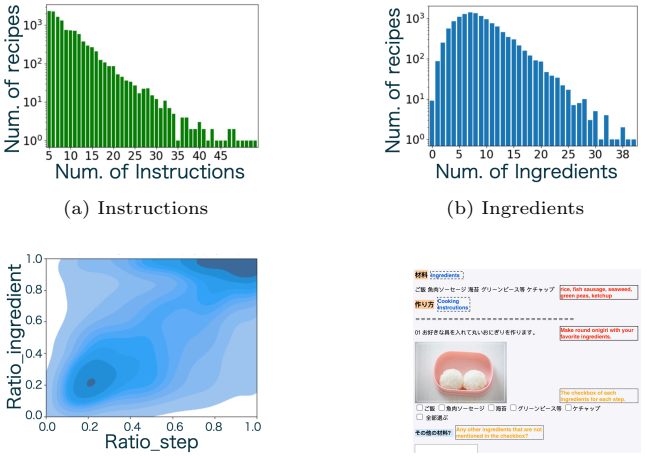
3.1 データ前処理

本研究の目的は，テキスト説明と画像のペアで構成される調理手順のシーケンスから成るデータセットを構築することで

表 1: 公開レシピデータセットの概要

データセット	画像数	手順画像 Images	言語
Food-101	101，000	×	English
ChineseFoodNet	192，000	×	Chinese
Vireo Food-172	110，241	×	Chinese
ISIA Food-200	197，323	×	Multi
UEC Food256	31，397	×	Japanese
Recipe 1M+	13 million	×	English
Cookpad	4 million	×/✓[1]	Japanese
Ours	121，547	✓	Japanese

[1] すべての調理手順に画像が添付されているわけではなく，調理手順に対する材料アノテーションデータもない．



(a) Instructions (b) Ingredients

(c) Ingredients in Instruction (d) Annotation Platform

図 2: Dataset Statistics and Our Annotation Platform

ある．Cookpad から入手した初期データセットは，Cookpad Image Dataset の多くの調理手順に関連する画像がないという課題に直面した．これに対処するため，以下の基準に従ってレシピのサブセットを選定した．

- 少なくとも 5 つの調理手順を含み，十分な情報が含まれていること．
- 各調理手順にテキスト説明と画像が含まれていること．

NII Cookpad Dataset [12] にある 170 万のレシピと，NII Image Dataset [4] の 160 万の調理手順の画像から，基準を満たす約 20 万件を特定し，その中からランダムに 12，000 のレシピを選択した，これにより，計 109，547 のテキスト説明と画像のペア，および最終料理の画像を集めた．将来的にデータセットをさらに拡大する可能性も見据えている．

3.2 データセット統計

図 2 に，データセットのさまざまな側面を示す．図 2a は，ほとんどのレシピが 10 個手順以下であることを示している．これは，データセット内のほとんどのレシピが調理プロセスの観点から比較的簡潔でわかりやすいことを示している（y 軸は対数スケールである）．図 2b は，レシピが通常 3 から 15 の食材

を含んでいることを示し，データセット内の食材の組み合わせの複雑さの範囲を示している．図 2d は，各レシピにおける食材の数と手順の相対位置（該当手順番号を該当レシピの全手順数で割った値という正規化された数値）との関係の密度プロットを示している．これは，初期の手順には食材が少なく，レシピが進行するにつれて増加することを示している．これらの発見は，食材認識と関連タスクの効果的な方法を開発するために，私たちのデータセットの特性を理解するのに役立つ．

3.3 データセットのアノテーション方法

テキスト説明から名詞を抽出することによる自動アノテーションは，マルチメディアレシピでは不十分である．これは主に以下の理由による．

- **関連する画像：**画像に見える食材はしばしばテキストで省略されることがあり，その逆もまた然りである．
- **文脈的な省略：**一連の調理手順では，文脈や背景知識から明らかな食材が省略されることがある．

マルチモーダル分析のための包括的なデータセットを作成するために，私たちは各手順を手動でアノテーションし，テキストで言及されていない食材も含め，物理的に存在する全ての食材を含めた．前の手順から繰り返される食材も含めて追加する必要がある．

アノテーションには 19 人のアノテーターを雇用し，各人が約 635 のレシピを担当した．図 2 に示されているような，私たちが開発したインターフェースを使用して，各レシピについて，その調理手順ごとに対応するレシピの食材リストが候補として提供される．アノテーターは，各手順に適切な食材のチェックボックスを選択するだけでよい．

アノテーションを完了した後，私たちのデータセットにおけるすべての食材の頻度をカウントする．これらの 12，000 のレシピには約 1 万の異なる食材があり，109，547 の調理手順には 42 の食材が 1，000 回以上，320 の食材が 100 回以上 1，000 回未満出現する．

結論として，私たちのレシピデータセットは，

- 各調理手順はテキストと画像のペアで構成されている．
- 各調理手順に物理的に存在するすべての食材がアノテーションされている．

という点で，他の既存のデータセットにはない独自性を持っている．これにより，調理プロセスの深い理解と，食材認識の精度向上に貢献する．

4 提 案 手 法

4.1 問 題 定 義

画像-テキストペアのシーケンスから構成されるマルチメディアレシピにおいて，各調理段階に存在する食材の特徴を分析し，食材をより上手く認識することが目的である．

$R = (I_1, T_1), (I_2, T_2), \ldots, (I_n, T_n)$ が n 個の調理手順を含むレシピを表すとする．各手順は，画像 I_i とそれに対応するテキスト説明 T_i を統合しており，ここで $1 \leqq i \leqq n$ である．さらに，R にはその食材リスト G_R があり，これはデータセット全体に現れる全食材 G のサブセットである．目的は，R の調理手順 i に存在する食材セット $G_{R,\,i}$ を識別することであり，ここで $G_{R,\,i} \subseteq G_R$ である．これは複数の食材が同時に画像に存在する可能性があるため，マルチラベル分類問題である．

この目的を達成するために，私たちは，まず，1 章で議論した仮説の検証を行って，適切な手法を提案する．

4.2 CLIP に基づくマルチモーダル調理手順エンコーダー

私たち構築したデータセットを用いて Finetune した CLIP（Contrastive Language–Image Pretraining）モデルを使用し，視覚的およびテキスト的特徴をエンコードし，各調理手順に対する統合された特徴ベクトルを生成する．

CLIP は，大量のインターネット画像とそれらの文脈的テキスト説明に基づいて広範囲にトレーニングされている．CLIP のトレーニング方法は，画像とそれに関連するテキストの両方から同時に学習することを含む．この共同トレーニングにより，別々のモダリティから生成された埋め込みは，共有空間内で本質的に整合している．画像とテキストからの特徴は，互いに理解され，関連付けられるように最適化されている．

画像 I_t とテキスト T_t に対して，CLIP はそれぞれ 512 次元の特徴ベクトル v_t^{img} および v_t^{text} を生成する．この変換は次のように表現される：

$$v_t^{img} = f_{c_img}(I_t),$$
$$v_t^{text} = f_{c_text}(T_t),$$

ここで，$f_{c_img}(\cdot)$ および $f_{c_text}(\cdot)$ は，それぞれ CLIP モデルの画像およびテキストエンコーディング関数を表す．

個々の特徴ベクトルが抽出された後，これらは手順 t に対する統合特徴ベクトル v_t を形成するために連結される．この手順は次のように表される：

$$v_t = \mathrm{concat}(v_t^{img}, v_t^{text}).$$

4.3 実験デザインと仮説の設定

実験では，以下の四つの仮説に基づいてレシピデータにおける食材の視覚的特徴を分析し，この研究の困難な点とその解決の手の存在を検証する．

仮説 A：調理過程の中間段階での食材認識．

中間段階の食材認識の困難さの理由の分析：中間段階では，食材は加工され変形し，その原形を失いやすい．これにより，調理中の食材の識別が難しくなる．

実験デザイン：CLIP モデルを用いて，調理の各段階 p における食材認識の精度を $Acc(p)$ と定義する．各レシピの調理手順の番号を [0，1] の区間内に正規化し，５つの区間 (p_1，p_2，p_3，p_4，p_5) に分ける．それぞれの調理段階において，$Acc(p)$ を算出し，識別精度のパターンを分析する．異なる調理段階での食材の視覚的特徴を分析し，提案モデルの識別能力を検証する．

仮説 B：調理手順に存在するサブタスク．

調理手順における並列構造の分析：レシピは直列的な手順で構成されるが，実際には複数のサブタスクが並行して進行する可能性を分析し，これらの構造が食材認識にどのように影響するかを検証する．

実験デザイン：n 個の調理手順を含むレシピ R における手順間の類似度を示す混同行列 $S(R)$ を定義し，直列の手順間でのジャンプとサブタスクの存在を分析する．サブタスクが食材認識に与える影響を評価する．

仮説 C：類似食材の認識とレシピ材料リストの役．

類似食材を除外する：図 1(c) のように，外見と役割が似たような食材は区別できなくても，片方しか食材リストにはいっていなければ，食材リストを参照することで判別できることを検証する．

実験デザイン：ここで提案する IngreMask というマスキング損失関数は，レシピに紐付いた食材リストを活用して，モデルの出力層で関連性のない食材を無視し，レシピには現れない類似食材を除外することを目的としている．

具体的には， BCE(Binary Cross-Entropy Loss) が，真のラベル y と予測確率 p の間の相互情報量を測定するのに対し，提案する損失関数は，レシピごとに異なる食材の重要度を考慮に入れる．標準的な BCE 損失は，各クラスに対する損失の和として計算されるが，本タスクにおいては全てのラベルが等しく重要であるわけではない．調理手順においてレシピの食材リスト G_R に存在する食材のみを識別することが目的であるため，G_R に共に現れない食材クラス C_1 と C_2 があっても，どちらかが存在すると思われる場合には，モデルが高い予測スコアを出しても問題ない．

そこで，各レシピに関連した食材リストに含まれていない食材によって生じる損失を無視するため，BCE 損失関数を拡張したカスタムのマスクされた損失関数である IngreMask（リストにない食材を無視するマスクされた損失関数）を導入する．この関数では，真実のラベルを 0（存在しない），1（存在する），及び無視指標（ignore index）で設計する．各調理手順 t において，G_R に含まれるかつその手順に存在する場合には 1 を，G_R に含まれるがその手順に存在しない場合には 0 を，G_R に含まれない場合には無視指標を割り当てる．

損失関数のマスクは次のように無視指標を用いて定義される．

$$M(ic) = \begin{cases} 1 & \text{if} \quad y_{ic} \neq ignore\ index \\ 0 & \text{if} \quad y_{ic} = ignore\ index \end{cases}.$$

このようにして定義されたカスタムのマスクされた損失関数は以下のようになる．

$$L_{masked} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} m_{ic} \\ (y_{ic}\log(p_{ic}) + (1-y_{ic})\log(1-p_{ic})) \tag{1}$$

レシピに関連付けられた材料リストに含まれない材料クラスによって生成される損失をマスキングすることにより，レシピの材料リストに含まれていないある i に対して分類器が大きな $\hat{y}_{ic}$ 値を生成しても，罰則を与えないようになる．これにより，非常に似た外見を持つが，実際には区別する必要がないいくつかの材料クラスを区別しようとする試みを分類器に停止させることができるようになり，トレーニングが助けられると期待される．

仮説 D：最終手順における食材推定．

最終手順の特徴分析：レシピの最終手順に現れる食材の特徴を分析する．一般に，調理の最初と最後は食材の特徴と名前がより対応しているため，高い精度が期待される．これを利用して，各手順における食材の存在を決定するために材料特有の Adaptive Threshold（適応閾値）を提案する．

実験デザイン：$R = (I_1, T_1), (I_2, T_2), \ldots, (I_n, T_n)$ が n 個の調理手順を含むレシピを表すとする．R における最終手順 (I_n, T_n) の食材集合 $G_{R,\,n}$ に基づいて，手順 n における食材推定の精度 $Pred_n$ を算出する．

この精度は，手順 n における材料クラス i の予測スコア $\hat{y}_{i,\,n}$ と各食材 i に対して計算された閾値 $\theta_i = f(\hat{y}_{i,\,n})$ を用いて，次のように食材の存在を予測する過程に基づいている：

$$y_{i,\,n} = \begin{cases} 1 & \text{if} \quad \hat{y}_{i,\,n} \geqq \theta_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

ここで，f は現在恒等関数として使用されているが，より微妙な閾値を設定するための柔軟性を提供する．この方法により，最終手順での食材の易識別性に基づいて，以前の手順の食材推定の精度を評価する．例えば，最終画像でのジャガイモの高い識別スコアは，そのレシピの以前の手順での識別が容易であることを示唆し，より高い閾値を要求する．逆に，最終画像での識別が困難な場合，以前の手順も同様に困難であることを示し，より低い閾値が適用される．

これを用いて以前の手順 $(I_1, T_1), (I_2, T_2), \ldots, (I_{n-1}, T_{n-1})$ の食材推定の精度 $Pred_1, Pred_2, \ldots, Pred_{n-1}$ を評価する．

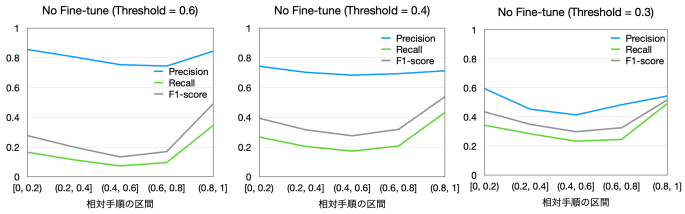
5 実験結果

本研究の実験結果は，提案したモデルと分析方法がレシピデータにおける食材認識に与える影響を明らかにするものである．以下に，各仮説に基づく実験の結果を示す．

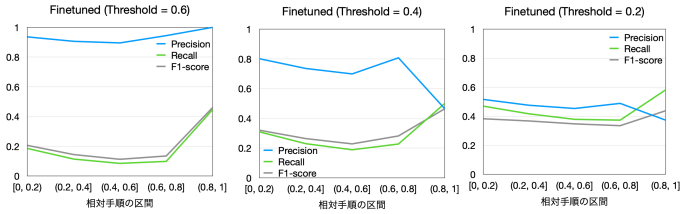
5.1 仮説 A：調理過程の中間段階での食材認識の結果

仮説 A では，調理過程の中間段階における食材認識の困難さを予測していた．具体的には，中間段階で食材が加工され変形し，その原形を失いやすいため，食材の識別が初期段階や最終段階に比べて困難になると考えられていた．図 3a と図 3b の折れ線グラフは，微調整なし（No Fine-tune）と微調整済み（Fine-tuned）の 2 種類の CLIP モデル両方の結果を，異なる閾値（Threshold）で比較していることが見て取れる．それぞれのグラフは，精度（Precision），再現率（Recall），および F1 スコアを食材認識の難易度区分ごとに示している．難易度は食材の加工段階に基づき分類されており，[0，0．2] から [0．8，1] の区間で示されている．

実験結果は，この仮説を支持するものとなった．これらの結



(a) Fine-tune なしの CLIP における結果



(b) Fine-tuned CLIP における結果

図 3: Precision, Recall, F1-score

果から，微調整なしと微調整ありのモデル両方とも，食材の中間段階における認識の難易度は，他の段階に比べて顕著であり，仮説 A が妥当であることが示された．また，モデルの微調整はパフォーマンスに影響を及ぼすが，閾値の選定も同様に重要であることが明らかとなった．このため，調理段階に応じた特性を考慮し，適切な閾値を設定することが求められる．

5.2 仮説 B：調理手順に存在するサブタスク

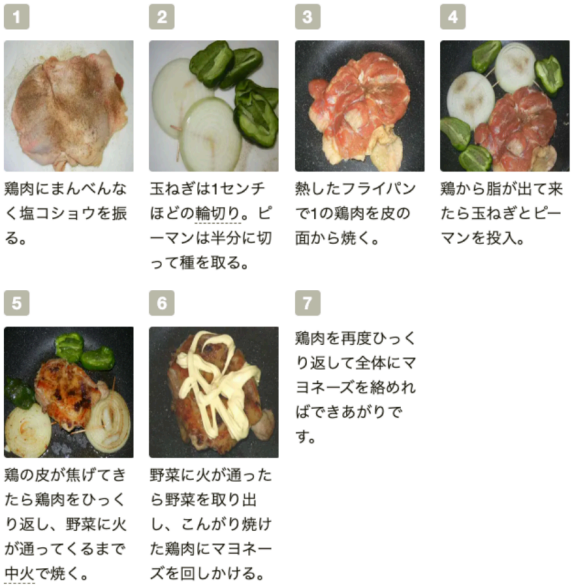
仮説 B では，調理手順におけるサブタスクの存在とその食材認識への影響について検討した．手順間の類似度を示す行列である図 4a および手順の視覚的な内容とテキストを示す図 4b から，サブタスクの具体例を確認することができる．

図 4a の各手順の視覚的な内容とテキストに基づくと，手順 3 と 4 は加工された肉の調理を示し，手順 2，4，5 では野菜の切り分けと加熱が行われている．これらの手順は互いに関連しており，同じ主要食材を使用しながら異なるサブタスクが並行して進行していることを物語っている．図 4b の類似度行列によれば，手順間の関係が数値化され，各手順がどの程度前後の手順と類似しているかが示されている．例えば，手順 1 と手順 3，手順 4 の間，また手順 2 から手順 4，5 への類似度が比較的高いことが示されており，これは明確なサブタスクのパターンを示している．具体的には，手順 1 での調理が手順 3 と 4 に影響を及ぼし，また手順 2 から分岐した内容が手順 4 と 5 に統合されていることが確認される．

この結果から，調理手順におけるサブタスクが食材認識に明確な影響を与えており，特に複数のサブタスクが関連する手順では食材認識モデルの性能が向上する可能性が示された．

5.3 仮説 C：類似食材の認識とレシピ材料リストの役

仮説 C の検証においては，類似食材（例えば，ごま油と酒，薄力粉と片栗粉，豚肉と牛肉）の識別が焦点とされた．これらの食材は外観やテクスチャが似ているが，一般的には同一レシピで同時に使われることは少ない．もし外見と役割が似たような食材は区別できなくても，片方しか食材リストにはいってい



(a) heatmap

(b) 3 stages

図 4: 手順 1 → 3 → 4，手順 2 → 4，5 に合流のサブタスクの例

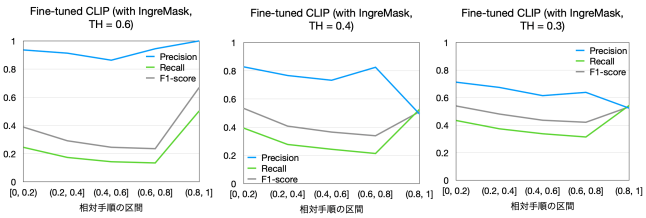


図 5: IngreMask における結果

なければ，食材リストを参照することで判別できる類似食材の識別において，レシピ材料リストが重要な役割を果たしていることを示す．固定閾値による分析では，特定の食材ペアが類似しているにもかかわらず，モデルがそれらを適切に識別できていることが確認された．

図 5 により，IngreMask を用いて Fine-tune したモデルと，そのアプローチを使用しない元のモデル図 3b の精度，再現率，

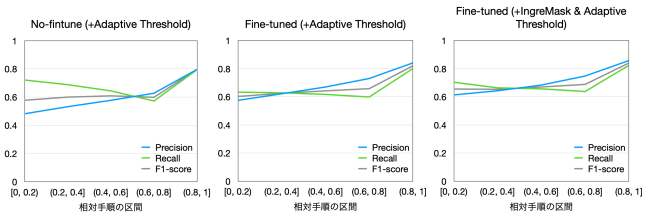


図 6: Adaptive Threshold における結果

F1 スコアを比較する．IngreMask を用いたモデルは，全体的に精度が向上している．これは，区別できないが現れないの食材を予測の時に考慮せずことで，似た食材を効率的に区別できる能力が高まっていることを示している．結果的に仮説 C の正当性を証明している．

5.4 仮説 D：最終手順における食材推定

仮説 D は，レシピの最終手順における食材の特徴が，一般に初期または最終手順での食材の特徴と名前が対応しやすいため，高い精度で推定できるというものである．この実験では，レシピにおける最終手順の食材に基づいて手順の精度に基づいて Adaptive Threshold（適応閾値）を設計し，それを以前の手順の食材推定の精度評価に応用する．

図 6 により，仮説 A での結果と照らし合わせると，最終手順での個別の閾値設定がモデルの識別精度を向上することに寄与していたことが明らかになる．具体的には，Adaptive Threshold の導入により，それぞれのレシピや食材に適した閾値が動的に適用されることで，仮説 A における中間段階の食材認識の困難さを克服し，その部分に限らず全体的な改善が見られた．この結果から，Adaptive Threshold は，レシピの異なる段階での食材特性の変化に対してモデルがより柔軟に対応できるようになり，モデルの精度向上に寄与するということを示している．結果的に仮説 D の正当性を証明している．

以上の分析から，調理手順の進行に応じて適切な閾値を設定することが，食材認識モデルの性能を最大化する鍵であると結論付けることができる．今後の研究では，このような閾値の動的調整がモデルの汎用性と精度をさらに向上させる可能性がある．

6 まとめと今後の課題

本研究では，レシピデータに基づく食材認識の問題に取り組み，様々な段階での食材認識の精度とその背景にある要因を解析した．複数の仮説に基づいて実験を行い，その結果からモデルの性能と識別能力に関する重要な洞察を得ることができた．

仮説 A では，調理過程の中間段階での食材認識の困難さを指摘したが，実験結果はこの予測に反して，調理手順が進むにつれてモデルの精度が向上することを示した．これは，モデルがレシピの完成品の視覚的特徴に特化して fine-tune された結果であると考えられる．仮説 B，C，D では，手順間の類似度とレシピ材料リストの活用が食材認識に及ぼす影響を検証し，類似食材の識別においてモデルが有効に機能することを確認した．特に，類似食材の識別は，レシピ材料リストの適切な活用により大きく改善される可能性があることが示された．

本研究を通じて，食材認識の分野においていくつかの有望な進展が見られた一方で，さらなる改善の余地が指摘されている．特に，調理手順の特定の点で再現率が低下する問題は，モデルの信頼性に直接影響を及ぼすため，この現象の原因を特定し，解消するための新たな手法の開発が必要である．また，食材認識の精度を向上させるためには，調理手順に応じた閾値の動的な調整が有効であると考えられる．これにより，モデルは各ステップの特徴をより正確に捉え，食材を識別できるようになるだろう．さらに，調理手順に含まれるサブタスクや並列手順の関係性に関する深い理解は，食材認識アルゴリズムの精緻化に役立つ．これらの手順がどのように相互作用し，全体の調理プロセスに影響を与えるかを詳細に分析することで，より効果的なモデル構築へと繋がるであろう．最後に，提案されたモデルの汎用性を検証するためには，様々なタイプのレシピや食文化に適用し，その性能を試験することが必要である．異なる文脈における食材の特性と調理法がモデルにどのような影響を与えるかを理解することで，より広範囲にわたる応用が可能となる．今後の研究では，これらの課題に対処することで，食材認識の分野をさらに発展させていくことが期待される．

謝　　辞

本研究は JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2123 の支援を受けたものです．

文　　献

[1] J. Chen and C.-W. Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM Multimedia*, pages 32–41, 2016.

[2] J. Chen, L. Pang, and C.-W. Ngo. Cross-modal recipe retrieval: How to cook this dish? In *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*, pages 588–600. Springer, 2017.

[3] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, 30:1514–1526, 2020.

[4] J. Harashima, Y. Someya, and Y. Kikuta. Cookpad image dataset: An image collection as infrastructure for food research. In *Proceedings of the 40th International ACM SIGIR*, pages 1229–1232, 2017.

[5] H. Kagaya, K. Aizawa, and M. Ogawa. Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM Multimedia*, pages 1085–1088, 2014.

[6] Y. Kawano and K. Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, 74(14):5263–5287, 2015.

[7] P. Liang-Ming, C. Jingjing, W. Jianlong, L. Shaoteng, N. Chong-Wah, K. Min-Yen, J. Yugang, and T.-S. Chua. Multi-modal cooking workflow construction for food recipes. In *Proceedings of the 28th ACM Multimedia*, pages 1132–1141, 2020.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[9] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Ay-

tar, I. Weber, and A. Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[11] S. Mori, H. Maeta, Y. Yamakata, and T. Sasada. Flow graph corpus from recipe texts. In *Proceedings of the 9th LREC*, pages 2370–2377, 2014.

[12] NII.ac.jp. Nii cookpad dataset, 2021.

[13] T. Nishimura, A. Hashimoto, and S. Mori. Procedural text generation from a photo sequence. In *Proceedings of the 12th INLG*, pages 409–414, 2019.

[14] T. Nishimura, A. Hashimoto, Y. Ushiku, H. Kameko, Y. Yamakata, and S. Mori. Structure-aware procedural text generation from an image sequence. *IEEE Access*, 9:2125–2141, 2020.

[15] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan. Foodnet: Recognizing foods using ensemble of deep networks. *IEEE Signal Processing Letters*, 24(12):1758–1762, 2017.

[16] A. Salvador, M. Drozdzal, X. Giro-i Nieto, and A. Romero. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10453–10462, 2019.

[17] P. Siyuan, D. Ling, H. Xuhong, L. Huating, and S. Bin. Chefgan: Food image generation from recipes. In *Proceedings of the 28th ACM Multimedia*, pages 4244–4252, 2020.

[18] M. Weiqing, L. Linhu, W. Zhiling, L. Zhengdong, and W. Xiaoming. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM Multimedia*, pages 393–401, 2020.

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

2024年2月28日(水) 10:00 ~ 12:10 | T4-B オンライン（Zoom Events)

Accessibility・インタフェース

座長:白石 優旗(筑波技術大学)
コメンテータ:鷹野 孝典(神奈川工科大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T4-B-1-01] テーマパークにおけるUX向上を目的とした情報提供システムの実用性評価

*糸賀 千晴[1]、本多 賢[2]、服部 哲[2]、青柳 西蔵[2]、山口 浩[2] (1. 駒澤大学大学院グローバル・メディア研究科、2. 駒澤大学)

10:20 ~ 10:35
[T4-B-1-02] 視覚情報と知識を利用したマルチモーダル対話における曖昧な言語指示対象の同定

*石井 里奈[1]、宮森 恒[2] (1. 京都産業大学 宮森研究室、2. 京都産業大学)

10:35 ~ 11:00
[T4-B-1-03] 複数のレイアウトアルゴリズムの選択的適用による階層型グラフのレイアウト最適化

*片岡 春菜[1]、村上 綾菜[1]、伊藤 貴之[1] (1. お茶の水女子大学　伊藤研究室)

11:00 ~ 11:25
[T4-B-1-04] チーム対戦ゲーム中のボイスチャットにおける音声を用いた発言促進インタフェースの検討

*重堂 孝太[1]、服部 峻[1]、砂山 渡[1] (1. 滋賀県立大学)

11:25 ~ 11:50
[T4-B-1-05] Multi-layer warning recognition system　駅構内における聴覚障害者サポートのためのマルチレイヤワーニング認識システムの実現

*須藤 光喜[1]、中西 崇文[2]、岡田 龍太郎[2]、峰松 彩子[2] (1. 武蔵野大学データサイエンス学部データサイエンス学科transmedia-tech-lab、2. 武蔵野大学データサイエンス学部データサイエンス学科)

テーマパークにおけるUX向上を目的とした情報提供システムの実用性評価

糸賀　千晴† 　本多　賢‡ 　服部　哲‡ 　青柳　西蔵‡ 　山口　浩‡

†駒澤大学大学院グローバル・メディア研究科　〒154-8525 東京都世田谷区駒沢 1-23-1
‡駒澤大学 グローバル・メディア・スタディーズ学部　〒154-8525 東京都世田谷区駒沢 1-23-1
E-mail: †3712102c@komazawa-u.ac.jp, ‡{knhnd82, hattori, aoyagi12, hiroyama}@komazawa-u.ac.jp

あらまし　本稿では，テーマパークにおけるUX向上を目的とした情報提供システムの実用性の評価を行う．これまで，実空間を階層化し，位置情報に基づいて背景情報を提供するシステムを実装することによって，テーマパークのUXを向上させる手法を提案してきた．評価実験において，大学をテーマパークに見立て，有用性や実現可能性の評価を行ったが，実際にテーマパークでの実験は認可やコスト面の問題により実現が困難であった．そこで本稿では，ペルソナを作成し，それに合わせてコンテンツを作成する．また，背景情報の確認機能を実装し，テーマパーク内で提供する背景情報と本システムの実用性について評価を行う．

キーワード　UX, テーマパーク，位置情報

1. はじめに

テーマパーク，美術館，大学などの実空間において，人が多く集まることで混雑が発生し，その場所のユーザ体験(UX: User Experience)が低下するといった問題が起こっている．特にテーマパークでは，アトラクションに乗るために長時間待つことや，飲食エリア，グッズ売り場など多くの場所で混雑が発生する．筆者らはこれまで，実空間に紐づく「背景情報」に注目したシステムの実装と提案を行ってきた．背景情報とは，目視では分からない隠れた情報，トリビア，豆知識のことである．実空間を階層構造で管理し，それぞれに紐づく背景情報をユーザの現在地に合わせて提供する手法を提案した[1][2]．また，本システムの有用性，実現可能性を評価するために，実装したシステムを用いた実験を行った．大学をテーマパークに見立て，システムを実際に利用してもらい，主観評価を行った．

テーマパークでの実用的な評価を行いたいが，コストや認可の問題から実際にテーマパークで評価を行うことは困難である．そこで，これまでに実装したシステムに，背景情報の確認機能を追加する．この機能を用いて実際のテーマパークにおける各エリアに紐づけた背景情報の配置が適切かどうかを検証する．背景情報については，ペルソナを作成し，それに合わせてコンテンツを作成した．本稿の目的は，ペルソナに基づき，コンテンツを作成し，テーマパーク内で提供する背景情報と本システムの実用性について評価することである．なお，本研究ではテーマパークの具体的な例を「東京ディズニーランド」とした．

2. 関連研究

テーマパークのUX向上に関する研究は幅広く行われている．テーマパークにおいて新型コロナウイルス感染症の流行下における写真投稿に関するUXデザインが行なわれており，ペルソナの設定からテーマパークにて世界観と一体化を求めた写真撮影のアプリケーションの提案が行われている[3]．テーマパークにおいて花火を鑑賞する際の体験に着目し，ユーザに適した鑑賞場所の提示するシステムの提案がなされている．ウェブサイト上で，椅子の有無などユーザのニーズに応じた花火の鑑賞地点を確認できるようなシステムが構築されている[4]．また，歴史テーマパークを対象に，車いすや障害の有無といった利用者特性に配慮し，ユニバーサルデザインを反映させた観光情報を提供するシステムの開発を行い，少人数のグループに特化して，一台のスマートフォンを共有して鑑賞者の体験を促す対面協調型のシステムを提案しているものもある[5]．アトラクションの待機列において，拡張現実を活用したゲームをプレイさせることで待ち時間の体感時間を減らす手法が提案されている[6]．

これまで，「背景情報」をユーザに提供することで，実空間に関する理解を深めより楽しめるようにすることを目指した提案を行い，大学において評価を行ってきた[1][2]．本稿では，ペルソナを作成し，そのペルソナに基づく背景情報を登録し，背景情報の確認機能を追加し，テーマパーク内で提供する背景情報と本システムの実用性について評価を行う．

3. 背景情報提供 Web システムの概要

これまで，主としてテーマパークを対象として実空間の UX を向上させることを目的とした情報提供方式を提案してきた[1][2]. 具体的には，実空間を階層化し，ユーザの位置情報に応じて範囲内の背景情報を提供するシステムについて実装と評価を行った．

テーマパークは，テーマごとに階層構造を定義することができる．もともとテーマパークはテーマごとにエリアが設定されているため，図 1 のような 3 つの階層で区切ることが可能である．それぞれ各エリアやアトラクションのテーマごとに実空間を区切って階層を定義している．

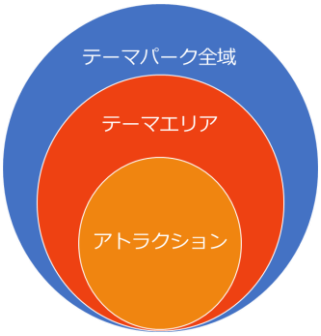


図 1　テーマパークの階層構造

図 2 の色をつけた矩形の範囲に当てはめると，一番広い「テーマパーク全域」，次に広いエリアが「テーマエリア」，一番狭いエリアが「アトラクション」としている．

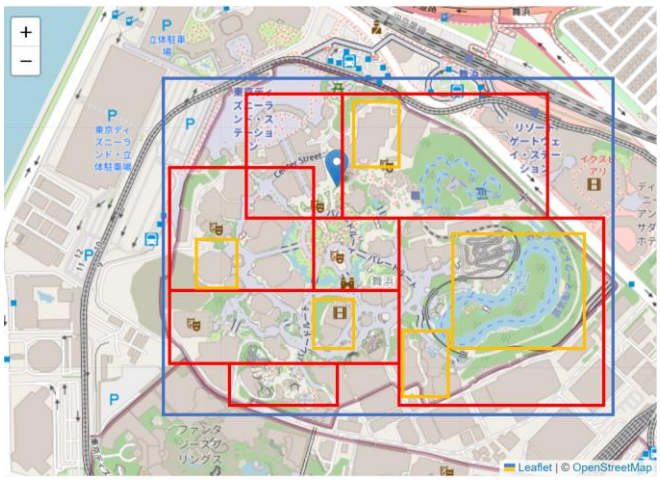


図 2　実空間の階層化

本システムでは背景情報を，階層化した実空間に紐づけ，ユーザの位置に応じて提供する．図 3 は本システムの構成を示したものである．

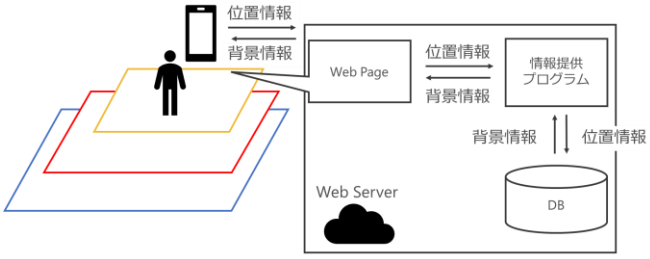


図 3　システム構成

本システムでは，位置情報をスマートフォンで取得後，位置情報に基づき，データベースに登録されている建物やオブジェクトのデータから範囲内の背景情報を検索し，ユーザに表示する．本システムは Web サーバ上にアップロードされており，データベースに登録された範囲内であればどこでも背景情報を表示できる．

4. ペルソナ作成

本稿では，ペルソナを作成するために，3 名の駒澤大学大学院生に対して聞き取りを行い，また，株式会社オリエンタルランドが公開しているゲストプロフィール[7]を分析した．

聞き取りはユーザのテーマパークにおける一日の体験を深く知ることを目的に「どの程度東京ディズニーランドに訪れるか」，「どんなことを楽しみに一日を過ごすか」，「どの程度ディズニーの映画やアトラクションについての知識や興味があるか」を質問し，どのような行動をするのか，どのようなことに興味を持っているのかについて詳細を明確にした．

ゲストプロフィールは，東京ディズニーリゾートを運営する株式会社オリエンタルランドによるゲストプロフィール[7]を基にしており，ゲストの年齢や性別等の特徴を明確にすることを目的に調査を行った．

表 1 は聞き取りとゲストプロフィールの調査を行った結果，作成したペルソナである．聞き取りや調査から，東京ディズニーランドにリピートして訪れる人が多いことが分かった．複数回訪れたことがある人についてはテーマパーク内の施設に対しての前提知識があると考えられる．また，テーマパーク内ではさまざまな催しが行われているが，主にアトラクションを中心に一日を過ごすと答えた人が多いことから，年に 2，3 回程度テーマパークに訪れ，アトラクションを中心にテーマパークを巡る人物をペルソナとした．

表 1　作成したペルソナ

年齢	20 代
テーマパーク入園回数	年に 2~3 回程度
前提知識	あり
テーマパークの周り方	アトラクション中心

5. 背景情報のコンテンツデザイン

作成したペルソナを元にコンテンツの作成を行った．コンテンツはデータベースに格納し，システムが位置情報に応じて提供できるようになっている．

年に 2~3 回程度東京ディズニーランドに訪れたことがある人がペルソナであるため，複数回東京ディズニーランドを訪れた人でも，新たな発見や興味が湧くコンテンツを作成した．エリアをテーマごとに階層化しているため，想定されるユーザの動きとして「エリアに入った時」「エリアにとどまっている」「エリアから出た」がある．ユーザの想定される動きに合わせて内容が変化できるようになっている．例えば，エリアに入った時にはテーマエリアやアトラクションに関する紹介，出た時にはアトラクションを出た時に見るべき点や周辺の場所に関する情報，階層に居続けた時はテーマエリアやアトラクションのコンセプトや歴史コンテンツなど，よりその場所を掘り下げる内容としている．表 2 は作成したコンテンツを格納するテーブルの一部である．表 2 では，東京ディズニーランド内にあるアトラクションである「美女と野獣”魔法の物語”」を例としている．

表 2　登録したコンテンツ

背景情報	エリアに入る，出る，とどまる
映画『美女と野獣』の世界を舞台に魔法にかけられたお城を巡る	1(入る)
魔法にかけられている入り口付近と，魔法が解かれた出口付近ではお城の色が違う	2(出る)
アトラクションの監修には多くのアニメーターも参加している	3(とどまる)

6. 背景情報の確認機能

本研究では，背景情報 Web システムの新たな機能として，背景情報として登録されたコンテンツを Web の地図上で確認することができる機能を実装した．ユーザは，実際に東京ディズニーランドに紐づいている背景情報を確認することができる．

図 4 は東京ディズニーランドにおける周り方の一例である．青い線は経路，番号はアトラクションを示している．Web 上で表示された地図上の東京ディズニーランドのエリアやアトラクションをペルソナの想定ルートに合わせて 10 か所から 15 か所程度クリックし，想定どおりのコンテンツが表示されるかどうかを検証する．

図 4　テーマパークの周り方

図 5 では，実際に背景情報が表示されているときの画面を示す．図 5 は⑤のアトラクションの背景情報を表示したものである．矢印のようにクリックしていくことで，クリックしたエリアやアトラクションの背景情報を確認することができる．

図 5　背景情報確認機能の実行例

7. 評価

本研究は，ペルソナに合わせてコンテンツを作成して実装した．しかし，実際には多様なユーザがシステムを利用することから，ペルソナに当てはまるユーザだけではなく，その他のユーザが利用したときに適切な情報提供できるかについて評価した．この評価は本

稿を発表した DEIM2024 におけるインタラクティブセッションにおいて実際にコメントをもらうことで行った．表 4 は DEIM2024 でいただいたコメントである．コメントをいただいた参加者は全部で 6 名である．インタラクティブセッションでコメントを取った参加者の属性については「具体例」の部分に「(参加者 X(性別･ディズニーランドに行ったことがあるか,ないか))」のように示した．

表 3　DEIM2024　コメント概要

内容	具体例
階層化の意義	● どの階層の情報か分かりにくい(参加者 A(男性・あり)) ● 階層によって興味や関心によってコンテンツを変える(参加者 B(男性・あり))
背景情報	● 背景情報の内容が面白い ● コンテンツが分かりにくい(参加者 A(男性・あり)) ● 前提知識が全くない人は使用することが難しい(参加者 C(男性・なし)) ● 生成 AI で階層に応じたコンテンツを生成(参加者 C(男性・なし)) ● 画像付きの情報提供(参加者 C(男性・なし))
他の場所での応用	● 大学(参加者 B(男性・あり)) ● オープンキャンパス(参加者 D(男性・なし)) ● 観光地(参加者 B(男性・あり))
応用方法	● 経路推薦(参加者 B(男性・あり)) ● 情報の関連性に基づく推薦(参加者 F(男性・なし))

実空間を階層化して情報提供する意義ついては，Web 上の地図を用いて情報提供するときには，ユーザがどの階層の情報を得ているのか分かりづらいとの意見があった.そのため,ユーザがどの階層にいるのか,どの階層の情報を得ているのか分かりやすいように工夫する必要がある．また，興味や関心に応じて階層を柔軟に変化させると面白いとの意見もいただいた．

背景情報については，提供されるコンテンツが面白いと評価を得た．しかし，ユーザの興味，関心，テーマパークにどれくらい訪れたことがあるかによって，コンテンツへの理解が分かれてしまうことがあるため，ユーザの興味に合わせた情報提供ができるような仕組みを取り入れる必要がある．また，画像などの視覚的な要素を活用して表示することで，より幅広いユーザに対して分かりやすい情報提供が可能だと考えられる．

また，本稿はテーマパークを対象としたコンテンツを背景情報としてデータベースに登録したが，本提案方式はテーマパークに限らず，大学や観光地などの場所でも応用が可能であるとの意見が寄せられた．

さらに，階層構造を生かした情報提供方式として，経路推薦や，情報の関連性に基づく推薦を行うことも考えられるとの意見をいただいた．

8. おわりに

本稿では，テーマパークの UX 向上を目的とした情報提供システムの実用性の評価を行った．テーマパークを階層的に分割し，それぞれの階層に背景情報を紐づけ，ユーザの位置情報に応じて提供するシステムに対して，聞き取りや来園者プロフィールの分析からペルソナを作成した．本システムに背景情報の確認機能を追加し，背景情報が適切に提供できているか検証を行った．

今後の課題としては，ペルソナで想定した特定のユーザに対する情報提供のみならず，より幅広いユーザの利用を想定し，ユーザの経験や来園回数，興味や関心に応じた情報提供方法について検討を行い．よりテーマパークの UX を向上させるように取り組んでいく．

参　考　文　献

[1] 糸賀千晴，本多賢，服部哲，青柳西蔵，山口浩，“階層的なエリア分割と関連情報の提供方式の実装”,情報処理学会第 119 回グループウェアとネットワークサービス研究発表会，2023.
[2] 糸賀千晴，本多賢，服部哲，青柳西蔵，山口浩，“階層的なエリア分割に基づく情報提供システムの実装と評価”，2023 年度社会情報学会大会，2023.
[3] 髙野倉雅人，久保千尋，荻谷光晴，“ウィズコロナにおけるテーマパークの体験価値向上を目指した写真投稿アプリのデザイン”，日本人間工学会第 62 回大会，2021.
[4] 西智菜美，安藤大地，向井智彦，“ユーザーニーズにあわせた花火鑑賞スポット提案システム”，情報処理学会 82 回全国大会.
[5] 高橋恭平，市川尚，窪田諭，阿部昭博，“GPS と Bluetooth による協調型 UD 観光情報システムの試作歴史テーマパークにおけるロケ情報提供-，情報処理学会第 71 回全国大会，2009.
[6] FABIO ZAMBETTA et al, “Reducing Perceived Waiting Time in Theme Park Queues via an Augmented Reality Game”, ACM Transactions on Computer-Human Interaction, Vol. 27, No. 1, Article 3. January 2020.
[7]OLC GROUP,ゲストプロフィール，2024/02/16 閲覧 https://www.olc.co.jp/ja/ir/olc/group05/group05-2.html

視覚情報と知識を利用したマルチモーダル対話における曖昧な言語指示対象の同定

石井　里奈† 　　宮森　　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2153083,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，実世界のマルチモーダル対話システムにおける曖昧な言語指示の問題に取り組む．マルチモーダル対話システムとは，音声や視覚情報など複数のモダリティを用いた対話システムである．近年，マルチモーダル対話の事前学習フレームワークが整備され，容易に流暢な対話が実現可能になりつつある．しかし，暗黙的に指示された対象の特定などにおいて，十分な性能を示せていない．例えば,「テーブルのそれ，取って」という発話があった場合，具体的な指示対象が直接言及されていないため，その指示の意図を言語のみから理解することは難しい．従来研究では，ジェスチャー情報と言語情報を用いて指示対象を特定する手法が提案されているが，ジェスチャー情報のみでは十分に指示対象を特定できないことも多い．本稿では，ジェスチャー情報と言語情報に加え，視覚情報を用いて文脈を補うことで，発話者の曖昧な言語指示対象を特定する手法を提案する．視覚情報から発話者の状況をより詳しく理解し，指示対象を的確に特定することが期待される．実験では，曖昧な言語指示対象を推定する能力を必要とするデータセットを用いて，提案手法の有用性を検証する．

キーワード　マルチモーダル対話，曖昧性解消，視覚

1　はじめに

マルチモーダル対話とは，音声や視覚など複数のモダリティを組み合わせて行われる対話形式である．特に人間とロボットの対話において，複数のモダリティからの情報を統合することで，より自然で豊かなコミュニケーションを実現することが可能になる．人間と対話を通して協働するロボットの実現には，視覚的文脈を含めた人間の発話理解が不可欠である．例えば，「そのペンを取って」という発話では，実際に参照しているペンの実体を知る必要がある．このように，ユーザの言語指示を理解し，指示対象の物体を特定することは，人間と協働するロボットにおいて重要な課題である．

ユーザとロボットの対話における課題として，ゼロ照応などの言語指示の曖昧さが存在する．ゼロ照応とは，発話者が特定の情報を省略することによって生じる現象である．例えば「そのコップを持ってきて」という発話において，持ってくる目的地に当たる二格の項が省略されている．このような省略がある場合．対話システムは省略された情報を補完し，ユーザの意図を正確に理解する必要がある．

従来，自然言語処理の分野において照応解析の研究が盛んに行われている．例えば，表層表現を手がかりとした推定手法 [1] や確率モデルを用いた照応解析手法 [2]，機械学習を用いた照応解析手法 [3] が提案されている．また，言語以外の情報を参照する外部照応に関する研究も盛んに行われている．例えば，指差しなどのジェスチャー情報を用いて指示対象を推定する手法が提案されている [4],[5]. これらの多くは，これらの多くは発話中に出てこない物体の情報を加えて解析することができず，発話者の状況を考慮した指示対象の特定が十分に行えていない．

本稿では，マルチモーダル対話における曖昧な言語指示に対して，指示対象の同定性能向上を目指す．具体的には，視覚情報を用いた曖昧な言語指示対象を特定する手法を提案する．画像キャプションモデルによってキャプションを生成し，視覚情報を加えることで，発話者の状況を考慮した解析を行うことができ，より正確な言語指示対象の特定が期待される．

2　関連研究

2.1　マルチモーダル対話に関する研究

マルチモーダル対話に関する研究は，盛んに行われている．例えば,，Henderson らは，ユーザのニーズや意図を追跡するダイアログ状態追跡の研究を行い，対話の各ステップでユーザの状態を正確に把握することでマルチモーダルな要素の効果的な統合に関する示唆を得た [6]．また，Zhao らは，タスク指向の対話システムにおける，DRQN の変種を用いた End-to-end のフレームワークを提案し，強化学習と教師あり学習の統合が対話システムの学習に有益であり，End-to-end のフレームワークが実用的であることを示唆している [7].

これらの研究は，対話の文脈やユーザの意図を網羅的に捉えることで，対話全体を理解するという手法が提案されてきた．本稿では，言語指示対象における曖昧性への対処が求められる実世界のシナリオに焦点を当て，視覚情報を用いることで，ユーザが具体的に指し示す対象を特定する手法を提案する．

2.2　内部照応解析に関する研究

内部照応解析に関する研究も盛んに行われている．今村らは，

対話用述語項構造解析を提案しており，ゼロ代名詞照応機能付きの述語項構造解析を対話に適応させる手法を提案している．この研究によって，対話での省略や代名詞化において，高精度な述語項構造解析を実現している [8]．西念らは，意味解析システム SAGE を用いて文を意味解析し，語間の意味的類似度を用いて指示代名詞の先行詞を高精度に同定する照応解析システム AnasysD を開発しており，意味的類似度が照応詞と照応関係にある先行詞を同定するための重要な要素であることを示唆している [9]．

これらの研究は，発話における照応先を正確に認識するために，文章中における照応詞の前後の文章からその照応詞にあたる語を特定するという手法が提案されている．本稿では，発話者の環境のものから正応詞に対応するものを予測する外部照応に焦点を当て，曖昧な言語指示に対する参照先を特定する手法を提案する．

2.3 外界照応解析に関する研究

言語以外の情報を参照する外界照応解析に関する研究も，これまでに取り組まれている，大山らは，物体カテゴリ，指示語，指差しの３つの情報とロボットが事前に環境を探索して得た物体知識を用いた外部照応を行い，これらの情報が外部照応において有効であることを示した [10]．また，Shukla らは，人間の外見的な特徴やパターンに着目してジェスチャーを認識する手法を提案しており，より柔軟で包括的なジェスチャーの理解を可能にした [11]．

これらの研究は，指差しの情報から照応解析を用いて，ユーザと対象物体との関連性を理解するための手法を提案している．本稿では，ジェスチャー情報に加え，どのような状況下に発話者がいるのかを理解するために視覚情報も取り入れて解析を行うことで，発話者の意図や注目している対象を正確に推定する手法を提案する．

3 提案手法

本稿では，視覚情報，言語情報，ジェスチャー情報の３種類の外部照応情報から発話者の曖昧な指示に対する参照先を特定する手法を提案する．提案手法の概要を図 2 に示す．先行研究では，ジェスチャー情報と言語情報を用いて参照先の特定を画像内と動画内で行っていた．本稿では，画像内の参照先の特定に焦点を当て，先行研究のジェスチャー情報と言語情報に加え，画像キャプション生成による視覚情報を用いる手法を提案する．

3.1 PAF（Part Affinity Field）

ジェスチャー情報の１つ目として，PAF[12] を用いる．PAF は，身体部位が他の部位とどのように関連しているのかを示すことができる手法である．また，各ピクセルに対して予測された方向ベクトルを矢印や線，異なる色で表現することができる．そのため，PAF をヒートマップとして表示し，各ピクセルの明るさや色の濃淡を用いて方向や強度を視覚的に表現することで，画像内の人物の身体部位間の関連性を理解しやすくなる．

3.2 saliency

ジェスチャー情報の２つ目として，saliency を用いる．saliency とは，画像内の注目すべき領域や重要な部分を指す．saliency は，MSINet[13] と呼ばれる手法により，画像内の重要な領域を特定することができ，ジェスチャーによって指し示された部分に関する重要な領域を特定することができる．

3.3 画像キャプション生成

画像キャプション生成は，CoCa（Contrastive Captioner）[14] を用いる．CoCa は CLIP の拡張形であり，Text-Encoder と Image-Encoder によるクロスモダリティな学習の後に，フレームワークに Text-Decoder を統合したものである．図 2 の写真に対して CoCa を用いた画像キャプション生成を行った場合の例として，「a woman standing in front of a kitchen sink.」というキャプションを生成する．このキャプションから，女性がキッチンの台所前に立っているという視覚情報を得ることができ，女性がどのような状況であるかをキャプション生成から理解することができる．

3.4 データセット

本稿では，発話とジェスチャーによって，どのオブジェクトが参照されているかを識別する YouRefIt をデータセットとして用いる [15]．このデータセットは，432 の室内の状況において，4,195 の対話が含まれており，人間とロボットの相互理解を可能にする参照データセットである．図 1 に YouRefIt データセットの例を示す．この写真において，「the bag sugar on top of the coffee pot」という発話が与えられている．このように発話者が参照先を曖昧に指し示す場合でも，発話内容とジェスチャーによってどのオブジェクトが参照されているかを特定するタスクに適したデータセットである．

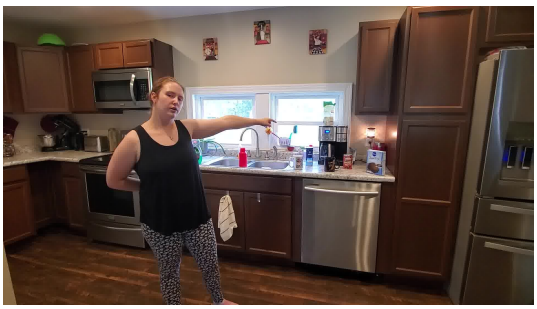


図 1　YouRefIt データセットの例

4 実　　験

4.1 実験設定

本稿では，オブジェクトサイズの大きさを large とした時の実験を行った．オブジェクトのサイズは，画像内にあるオブジェクトの実際の境界ボックスの面積と画像全体の面積の比率を計算し，オブジェクトの大きさを推定する．また，従来手法ではトークン数が 20 で行われていたが，本稿ではトークン数を 100 として実験を行なった．また，ジェスチャー情報を使用しない

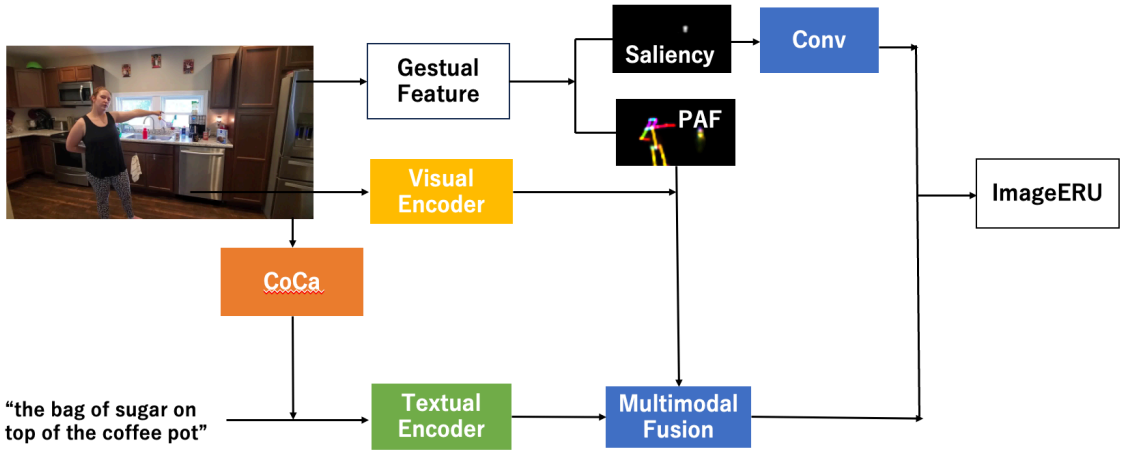


図 2　提案手法の概要図

場合，paf 情報のみを用いた場合，saliency 情報のみを用いた場合，paf 情報と saliency 情報の両方を用いた場合の 4 パターンで実験を行なった．

4.2 評価指標

評価指標は，IoU（Intersection over Union）比率を用いて評価を行う．先行研究では，与えられた画像と文に基づいて参照される領域をバウンディングボックスで予測し，正解のバウンディングボックスとの IoU 比率を計算していた．本稿も同じく，物体検出のベンチマーク [16] に従い，IoU を 0.25 として評価を行った．

4.3 比較手法

実験では，Chen らの手法 [15] と比較し，提案手法の有効性を確認した．それぞれ 4 つのパターンで実験を行うことで，視覚情報やジェスチャー情報の有無による精度の差を調査した．

4.4 実験結果

提案手法と従来手法において，それぞれ 4 パターンの実験を行なった結果を表 1 と表 2 に示す．これらの結果から，どの条件においても，従来手法よりも提案手法を用いた場合の精度が低い結果となった．

表 1　従来手法を用いた指示物体の特定精度

条件	精度
ジェスチャー情報なし	0.135
paf	0.105
sal	0.119
paf + sal	0.135

4.5 考　　察

提案手法が従来手法よりも精度が低い結果となった理由として，画像キャプション生成の生成文に問題があると考えた．提案手法では，視覚情報を画像キャプション生成モデルで生成し，それを言語情報と組み合わせて Text-Encoder に入力していた．この処理において，画像キャプション生成モデルが生成するテキストは，画像に関する詳細な説明や状況など，必要のない情報を含むことが考えられる．このような情報は，参照先の特定においては，ノイズとして働く可能性があるため，良い精度が出なかったのではないかと考えた．

表 2　提案手法を用いた指示物体の特定精度

条件	精度
ジェスチャー情報なし	0.121
paf	0.091
sal	0.089
paf + sal	0.119

5 ま と め

本稿では，実世界のマルチモーダル対話システムにおける曖昧な言語指示の問題に取り組む．従来手法では，指差しやジェスチャーなどの情報を用いて曖昧な言語指示対象を特定していた．しかし，これらの多くは発話中に出てこない物体の情報を加えて解析することができず，発話者の状況を考慮した指示対象の特定が十分に行えていない．そこで，本稿では，視覚情報を用いて文脈を補うことで，発話者の曖昧な言語指示対象を特定する手法を提案する．

実験では，従来手法と提案手法を比較することで，視覚情報が曖昧な言語指示における参照先の特定に有用であるかを調査した．これらの実験を行なった結果，ジェスチャー情報と言語情報のみの従来手法よりも，それらに視覚情報を加えた提案手法の精度が下回る結果が得られた．今後は，異なるモデルを用いた画像キャプション生成や，視覚情報から発話者の意図を理解するための常識推論を用いた指示対象の特定に取り組む予定

である．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] 村田真樹, 黒橋禎夫, 長尾真. 表層表現を手がかりとした日本語名詞句の指示性と数の推定. 自然言語処理, Vol. 3, No. 4, pp. 31–48, 1996.
[2] 関和広, 藤井敦, 石川徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, Vol. 9, No. 3, pp. 63–85, 2002.
[3] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004s
[4] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. Communicative Learning with Natural Gestures for Embodied Navigation Agents with Human-in-the-Scene. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4095–4102, 2021.
[5] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, SongChun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouReflt: Embodied Reference Understanding with Languageand Gesture. In IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1385–1395, 2021.
[6] Matthew Henderson1, Blaise Thomson1 and Jason Williams2 1Department of Engineering, University of Cambridge, U.K. 2Microsoft Research, Redmond, WA, USA.The Second Dialog State Tracking Challenge.ACL Anthology,2014.
[7] Tiancheng Zhao and Maxine EskenaziLanguage Technologies InstituteCarnegie Mellon University.Towards End-to-End Learning for Dialog State Tracking andManagement using Deep Reinforcement Learning.arXiv preprint arXiv:1606.02560v1.2016.
[8] 今村賢治， 東中竜一郎， 泉朋子. 対話解析のためのゼロ代名詞照応解析付き述語項構造解析
[9] 西念星宝， 谷津元樹， 原田実. 語意類似度を用いた指示代名詞の照応解析システム AnasysD.
[10] 大山瑛， 長谷川翔一， 中川光， 谷口彰， 萩原良信， 谷口忠大. 実世界のマルチモーダル情報に基づく指示語を含んだ言語指示の外部照応解析. 言語処理学会,2023.
[11] Dadhichi Shukla, Ozg ¨ur Erkent and Justus Piater ¨Intelligent and Interactive Systems, Institute of Computer Science, University of Innsbruck, Austria.Probabilistic detection of pointing directions for human-robot interaction.
[12] Zhe Cao, Student Member, IEEE, Gines Hidalgo, Student Member, IEEE,Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields
[13] Alexander Kronera, Mario Sendena, Kurt Driessensc,Rainer Goebela,Contextual Encoder-Decoder Network for Visual Saliency Prediction
[14] CoCa: Contrastive Captioners are Image-Text Foundation Models, Jiahui Yu, Zirui Wang,Vijay Vasudevan Legg Yeung Mojtaba Seyedhosseini Yonghui Wu.
[15] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, SongChun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouReflt: Embodied Reference Understanding with Language and Gesture. In IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1385–1395, 2021.
[16] Adam D Galinsky, William W Maddux, Debra Gilin, and Judith B White. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. Psychological Science, 19(4):378–384, 2008.

複数のレイアウトアルゴリズムの選択的適用による階層型グラフのレイアウト最適化

片岡　春菜†　　村上　綾菜†　　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{kataoka.haruna,g1720541,itot}@is.ocha.ac.jp

あらまし　グラフ可視化の可読性はレイアウトによって左右される．複数のノードの集合（メタノード）で構成される階層型グラフは，レイアウトの生成が特に難しい．この問題を解消するために既存のレイアウト手法がいくつか存在する．既存手法はそれぞれ特性があり，ユーザーが求める情報から常に適切なレイアウト手法を生成できるとは限らない．そこで本報告では，階層型グラフのレイアウト最適化処理の過程において，複数の配置手法を組み合わせながら適切なレイアウトを得る手法を提案する．本報告では，単一のレイアウト手法に対して最適化処理を適用した先行研究に対して，複数のレイアウト手法を組み合わせることによる可読性の向上を検証する．また，複数のレイアウト手法の組み合わせについても検証する．

キーワード　可視化手法，階層型グラフ，グラフレイアウト，遺伝的アルゴリズム

1 はじめに

グラフレイアウトとは，グラフを構成するノードの画面上の位置を適切に算出することで，ノード間の接続構造を適切に表現する可視化手法である．グラフはノードとエッジからなり，ノードは各データを表し，エッジはノードとノードの接続を表す．

グラフはノードのレイアウトによって可読性が左右される．エッジの交差が少なく，適度な広がりを持つグラフは可読性の高いグラフとなる．ノードやエッジの数が増えるほど，交差が多くなるため，可読性を上げることが重要となる．

階層型グラフとは，複数のノードの集合であるメタノードから構成されているグラフであり，大規模なグラフの概要を効果的に可視化する手法である．階層型グラフでは特に，接続構造が複雑である点や，各々のメタノードが固有の面積を有することで，可読性の高いレイアウトの生成が難しい．ノードに付与されたメタ情報の可読性も考慮する必要がある場合には，レイアウトの生成はさらに難しくなる．

階層型グラフのレイアウトを向上させる手法としては，レイアウト手法自体の改善もさることながら，既存手法によるレイアウト結果を最適化するアプローチも考えられる．そのような手法として村上らは，遺伝的アルゴリズムを用いたレイアウトの最適化手法 [1] を提案している．この手法では，各々のレイアウト結果を数値評価し，その結果から遺伝的アルゴリズムを適用して次の世代のレイアウトの初期値を求める，という処理を反復することでレイアウトを最適化するものである．村上らの手法では伊藤らによるレイアウト手法 [2] を適用しているが，他のレイアウト手法を用いても同様な最適化が可能である．

階層型グラフにおいては既存のレイアウト手法がいくつか存在しており，各手法にはそれぞれ特性がある．同じデータに対して特性の異なるレイアウト手法を複数組み合わせることで，汎用性，視認性が上がることが推測される．本研究では，複数のレイアウト手法を組み合わせることで，汎用的で視認性の高いレイアウトを生成し，階層型グラフの可読性を向上する方法を提案する．

2 関連研究

グラフを構成するノードには位置情報を有する場合と有さない場合がある．ノードが位置情報を有さないグラフを可視化する際には，各ノードの画面上の位置を適切に算出する必要がある．そこでグラフのレイアウト手法が活発に議論されている．一方で，密な接続構造を有するノード群や，同一の属性を有するノード群による集合（メタノード）で構成されるグラフを階層型グラフという．階層型グラフを構成する各々のメタノードは固有の面積を有することから，そのレイアウト手法の確立は難しい．

Itoh ら [2] は，階層型グラフのクラスタリングおよびレイアウトに関する手法 Koala を提案した．Koala では，多くのノードと接続しているノードをクラスタから独立させるクラスタリング手法を適用することで，重要なノードの視認性を高める．また，同じクラスタに属するノードを円形領域となるメタノードの内部にまとめて配置する．

グラフレイアウトの評価は重要な課題であるが，階層型グラフに特化したグラフレイアウトの評価指標は少なかった．Liu ら [3] は，階層型グラフレイアウトの数値評価指標として Sprawlter を提案した．Sprawlter は，空間の浪費度を評価する Sprawl と，配置の乱雑度を評価する Clutter からなる．

村上ら [1] はまた，遺伝的アルゴリズムによるレイアウトの最適化手法を提案した．Sprawlter は階層型グラフのレイアウト評価手法であるが，村上らの研究ではこれを遺伝的アルゴリズムの目的関数として設定した．遺伝的アルゴリズムを用いた最適化では，階層型グラフの可読性において，Clutter の基準で改

善が顕著に見られた．しかし，Sprawl の基準では数値上改善が見られるものの，視覚的な差異はほとんど無かった．また，この最適化手法ではレイアウト手法に Koala が用いられている．

本研究では，村上らの手法を踏襲しつつ，レイアウト手法に複数の異なるアルゴリズムを用いる最適化を試みた．

3 提案手法

本手法の処理手順を図 1 に示す．全体的な処理手順は村上ら [1] の手法を踏襲する．階層型グラフのレイアウト手法の部分で，村上らの手法で用いられた階層型グラフレイアウト描画手法 Koala に加え，同じく階層型グラフレイアウト描画手法である Persistent Homology Guided Force-Directed Graph Layouts（以下 PH とする）[4] を用いる．これらを選択的に適用し，遺伝的アルゴリズムを用いてレイアウトの最適解を探索する．

最適化の過程にて，階層型グラフレイアウトに特化した数値評価指標である Sprawlter をレイアウトの評価に採用する．

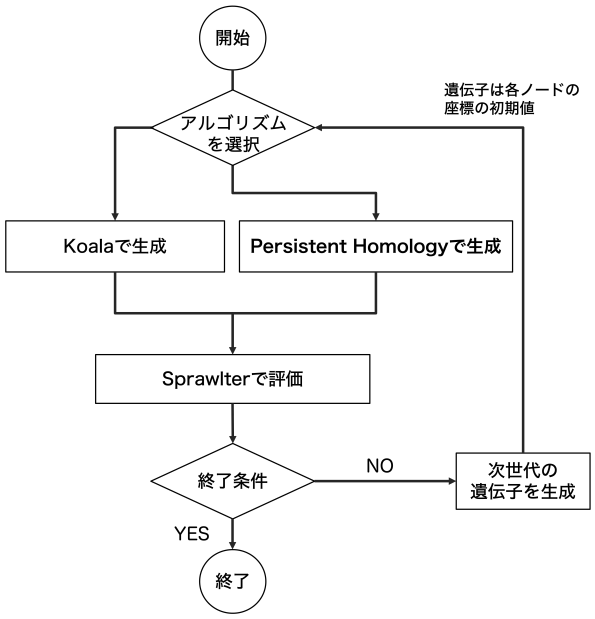


図 1　処 理 手 順

3.1 遺伝的アルゴリズム: NSGA-II

本手法では，レイアウトの最適化において遺伝的アルゴリズムの一種である NSGA-II [5] を採用する．NSGA-II の特徴は，非優越ソートと混雑度ソートという 2 種類のソートを用いて，次世代に残す遺伝子を選択するという点である．非優越ソートにより，従来よりも短い時間で，前の世代を優越する良好な解を得ることを可能にする．また混雑度ソートにより，解の多様性を保持したまま，次の世代に良好な解を残すことを可能にする．

3.2 遺伝子の設計

遺伝的アルゴリズムにおいて，遺伝情報を伝える実態を染色体とよび，その形質を決定するコードを遺伝子と呼ぶ．解の候補は染色体で表現され，1 つの個体が 1 つの解に対応する．本手法では，メタノードの座標値を遺伝子とし，1 つの個体が 1 つのグラフレイアウトに対応するものとする．そして，メタノード x 座標と y 座標のペアを交互に 1 次元配列に格納することで，1 次元の遺伝子を生成する．次節にて後述する通り本手法では，村上らの手法を拡張して，遺伝子にはそれを生成したレイアウト手法を識別するための変数を持たせる．

3.3 アルゴリズムの選択

本手法では一度の最適化過程で複数のレイアウト手法を組み合わせることを主目的としている．適用されるレイアウト手法を選択的に決定するために，本手法では遺伝子に生成元のレイアウト手法を識別するための変数を持たせる．初期集団の生成において，変数には複数のレイアウト手法をそれぞれを同じ確率で設定する．交叉が行われる際には，2 つの遺伝子の識別変数の値が同じ場合はレイアウト手法を変更せず，識別変数の値が異なる場合にはいずれかの値をランダムに設定する．

4 実行例

4.1 使用データ

使用データは村上ら [1] と同様のものを用いる．1998 年から 2013 年までに NERC Biomolecular Analysis Facility (NBAF) で発表された論文に対して，著者をノード，共著関係をエッジとしたグラフを可視化する．

4.2 レイアウト最適化結果

4.2.1 Koala のみの場合

村上らの研究に沿って，Koala のみを使った最適化を実行した．図 2 に最適化前および最適化後のレイアウト結果を示す．この結果から，Clutter(配置の乱雑度) の基準で改善が顕著に見られた．しかし，Sprawl(空間の浪費) の基準では数値上の改善は見られるものの，視覚的な差異はほとんどなかった．

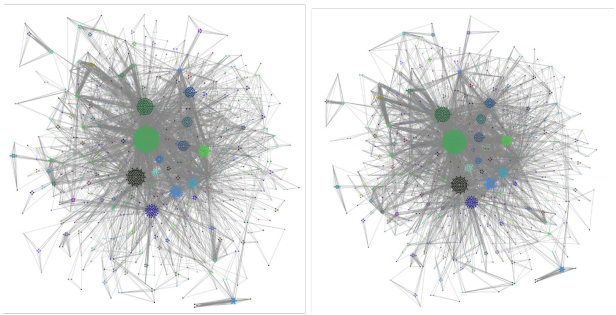


図 2　（左）最適化前,（右）最適化後

4.2.2 PH のみの場合

村上らの研究とは異なるアルゴリズムである PH を用いて，同様の最適化を実行した．図 3 に最適化前および最適化後のレイアウト結果を示す．Koala と比較すると，最適化前からノード間の重なりが一切なく，代わりにノードとエッジの重なりが非常に多い．数値上は Sprawl と Clutter どちらも一定の改善があったものの，視覚的にはどちらも改善がほとんど見られなかった．

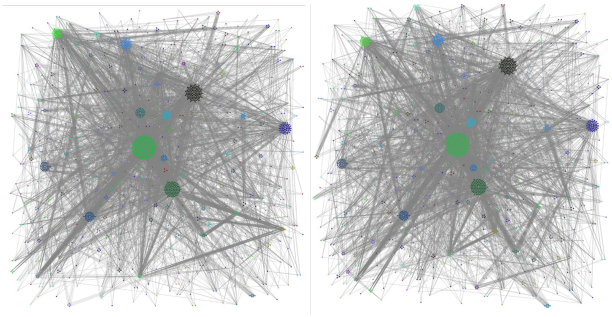


図 3 （左）最適化前,（右）最適化後

4.2.3 Koala と PH を選択的に用いた場合

Koala と PH を選択的に用いる形で最適化を実行した．方法として，遺伝子に Koala または PH を識別するための変数を設定し，初期集団を生成する際に Koala と PH の値を持つ遺伝子がそれぞれ同数生成されるようにした．交叉の際，2 つの遺伝子が持つ変数の値がどちらも Koala，またはどちらも PH の場合はそれぞれ同じ値を継承し，2 つの遺伝子が Koala と PH だった場合にのみ，いずれかの値をランダムで設定した．最適化の過程として，図 4 に初期世代と最終世代の評価値の比較を示す．横軸は Sprawl の評価値，縦軸は Clutter の評価値であり，どちらも値が小さいほど高い評価を意味する．図中の各点が 1 つの階層型グラフレイアウトを示す．青い点が初期世代， 赤い点が最終世代の個体の評価値である．なお，初期世代の番号と該当世代の番号に対応関係はない．

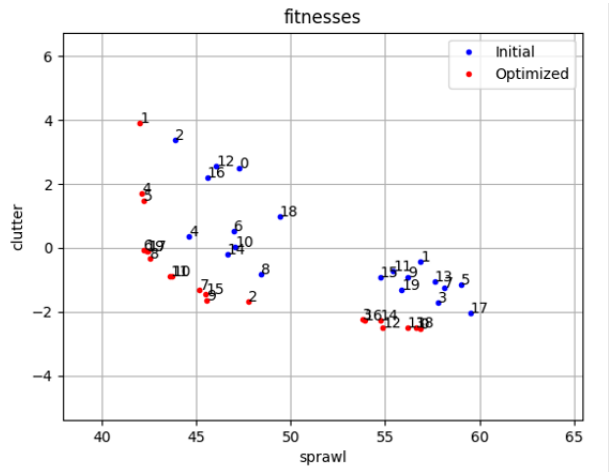


図 4 初期世代と最終世代終了後の評価値の比較

レイアウトの一例を図 5, 6 に示す．

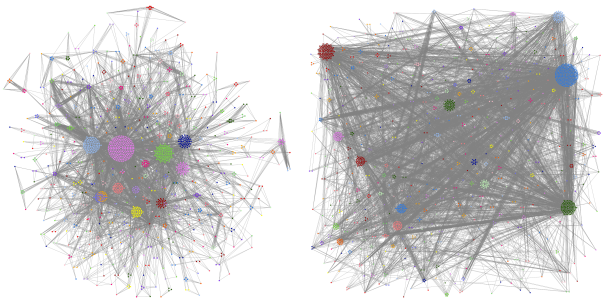


図 5 最適化前：（左）レイアウト 0 番,（右）レイアウト 18 番

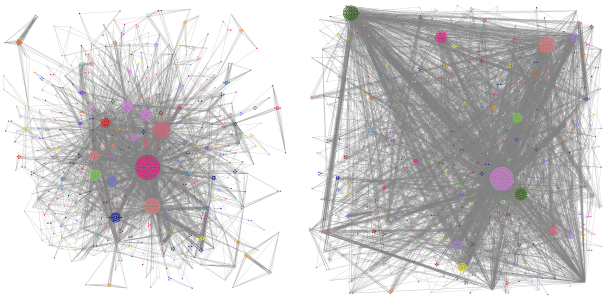


図 6 最適化後：（左）レイアウト 9 番,（右）レイアウト 13 番

図 5 はいずれも最適化前のレイアウト結果である．いずれも図 4 中の青い点で示されている．図 5(左) に示す結果結果は，図 4 で示されている通り，最適化前の結果の中で sprawl の値が良く，crutter の値が悪いものである．値の通り，ノードやエッジの重なりが散見されるが，グラフ全体はコンパクトで，無駄な余白は少なくなっている．対して図 5(右) に示す結果は，最適化前の結果の中で sprawl，crutter どちらも悪い結果である．見た目より PH のレイアウトによるもので，その特性が強く出ている．ノードの重なりは無いが，エッジの交差が非常に多くなっている．

図 6 は最適化後のレイアウト結果である．いずれも図 4 中の赤い点で示されている．最適化後は全体的に crutter の値が下がっており，sprawl の値には幅があることが分かる．図 6(左) に示す 9 番のレイアウトは，crutter と sprawl の値がバランス良く低いものである．最適化前の例に対して，こちらはノードの重なりがほとんど無く，グラフもコンパクトで可読性が高いグラフと言える．図 6(右) に示す 13 番のレイアウトは，最適化後の結果の中では crutter の値が低く，sprawl の値が高いものである．ノードの重なりほとんど無く，図の中心の空白が多いことからエッジの交差も少ないことがわかります．

5 結論・今後の課題

本研究では，遺伝的アルゴリズムを用いたグラフレイアウトの最適化過程において，Koala と PH の 2 種類のレイアウト手法を選択的に用いて階層型グラフレイアウトの最適化を実行した．Koala と比較して PH はノードの重なりがもともと少なく，また視覚的には空間の浪費度が高く見える．その結果，2 つレイアウトを組み合わせた場合に，初期集団を PH とでレイアウトしたケースにおいては Sprawl の基準での改善が顕著であった．逆に，初期集団を Koala でレイアウトし，PH による最適化の度合いが強いケースでは，Sprawl の基準で悪化しているものも見られた．

今後の課題として，村上らの研究の課題であった目的関数の再設計について見直したい．また，Koala や PH 以外のアルゴリズムも採用した上で，その組み合わせについても検証したい．

文　献

[1] A. Murakami, T. Itoh, Optimization of Hierarchical Graph Layout with a Genetic Algorithm and Sprawl/Clutter Metrics, 27th International Conference on Information Visualisation (IV2023), pp. 166-171, 2023.

[2] T. Itoh, K. Klein, Key-node-Separated Graph Clustering and Layout for Human Relationship Graph Visualization, IEEE Computer Graphics and Applications, Vol. 35, No. 6, pp. 30-40, 2015.

[3] Z. Liu, T. Itoh, J. Q. Dawson, T. Munzner, The Sprawlter Graph Readability Metric:Combining Sprawl and Area-aware Clutter, IEEE Tramsactions on Visualization and Computer Graphics, Vol. 26, No. 6, pp. 2180-2191, 2020.

[4] A. Suh, M. Hajij, B. Wang, C. Scheidegger and P. Rosen, ”Persistent Homology Guided Force-Directed Graph Layouts,” in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 697-707, Jan. 2020, doi: 10.1109/TVCG.2019.2934802.

[5] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, ”A Fast and Elitist Multiobjective Genetic Al- gorithm: NSGA-II.” IEEE transactions on evo- lutionary computation, vol. 6, no.2, pp. 182-197, 2002.

チーム対戦ゲーム中のボイスチャットにおける音声を用いた発言促進インタフェースの検討

重堂　孝太† 　服部　峻†† 　砂山　渡††

† 滋賀県立大学工学部電子システム工学科　〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学先端工学研究院　〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †tz23kjudo@ec.usp.ac.jp, ††{hattori.s,sunayama.w}@e.usp.ac.jp

あらまし　複数人で行われるチーム対戦ゲームでは VC（ボイスチャット）を利用するが，問題点としてお互いにチームメンバー間で認識の齟齬が起きてしまうことが挙げられる．その理由はゲームの状況が変化するスピードが早いことや，敵の居場所など情報を報告するのに慣れていないためであり，結果としてゲーム中にコミュニケーションを常にとり続けるのが難しくなり齟齬が起きてチームが不利になってしまう．そこで本稿では，認識の齟齬を減らすためにはチームメンバー全員の発言量を増やすことが必要であると考え，個々人の VC 中に一定の沈黙時間が発生すると発言を促す音声を流すシステムを作成した．このシステムを用いながら被験者にゲームをチームプレイさせる実験を行い，システムの有無による会話率や沈黙時間の回数の比較等を行う．

キーワード　音声・音楽, ユーザ支援, チーム対戦ゲーム, インタフェース, 意思疎通

1 はじめに

まず本研究の対象である「チーム対戦ゲーム」とは複数人対複数人で対戦するゲームを指す．この対戦ゲームは eSports のタイトルとなっているものも含んでおり，現在 eSports は世界各国で市場を拡大し続けていて日本でも成長が続いている．日本 e スポーツ白書 2023 の先行発表された内容 [1] によると 2022 年の日本 e スポーツ市場規模は 125 億円に到達していてプロゲーマーの試合も視聴者数が増え，注目度が高まり続けるジャンルであると言える．また同じゲームの研究であってもチーターの対策として不正か正当なゲームプレイを検出する研究 [2] や，暴言を発するプレイヤを行動的に予測あるいは検知する研究 [3], [4] などは既になされているが，チーム対戦ゲームでユーザを直接的に支援する研究は多くなく，研究分野としてはまだまだ新しいものであると言える．

一般にチーム対戦ゲーム中は勝ちを目指すために同じチームでボイスチャット（VC）を用いて通話しながらチームメイトと連携してゲームをプレイすることが多く，eSports のプロでも VC を使って連携をとるのが一般的である．チーム対戦ゲームでのプレイヤ側の問題として認識の齟齬が起きてしまうことがある．この認識の齟齬というものは一言で言えば，ゲーム中に会話が十分になされていないなどの理由でチームメイトの考えと異なるプレイが起こることであり，それにより時には勝敗に影響することがある．このような事象はプロの eSports チームでも起こることがある [5]．原因としては単に会話するだけでなくゲームをしながら会話を続ける必要があってプレイヤにマルチタスクを求められていたり，コミュニケーションが苦手であったりと原因を一つに絞るのは難しく様々な要因が合わさっていると考えられる．

本研究の目的としてはこのようなチーム対戦ゲーム中のボイスチャットにおけるチームメンバー間のお互いの認識の齟齬を減らすことであり，そのために会話量を増加させることが出来るインタフェースを検討することである．インタフェースを検討するためにまずはベースラインとなる発言促進システムを作成し，その評価実験を行った結果を考察して，インタフェースの改善を検討していく．

本稿の構成は以下の通りである．まず，2 章で関連研究について，3 章でベースラインのシステムについて述べる．次に，4 章で評価実験について述べる．最後に，5 章で本稿をまとめ，今後の研究課題と改善手法の検討についても述べる．

2 関連研究

本章では関連研究について述べる．

2.1 対戦ゲームでのコミュニケーションに関する研究

本節では問題場面についてより詳しく知ることもかねてチーム対戦ゲームのコミュニケーションについて書かれている研究について述べる．

Aleksi [6] は CS:GO（現 CS2）[7] という FPS 対戦ゲームのセミプロのチームに対して聞き取り調査を行い，そこでゲーム内でのコミュニケーションの重要性について述べている．具体的には個人個人にはゲーム内で与えられた役割があってそれを遂行していくが，チームのリーダーはマップの様々なポジションの名前や略語を学んだり覚えたりすることが余分な話を減らし，同じことを繰り返し発言しなくて良いため重要であると述べている．またチーム全員が話すことが必ずしも良いことではなく FPS ゲームではゲームスピード，言い換えるとゲームの状況が変わる速さが他のゲームより速くコミュニケーションが多すぎてチームにとって重要な情報が伝えられない可能性が出て来るため，コミュニケーションは速く，効率的で，一義的で，簡潔

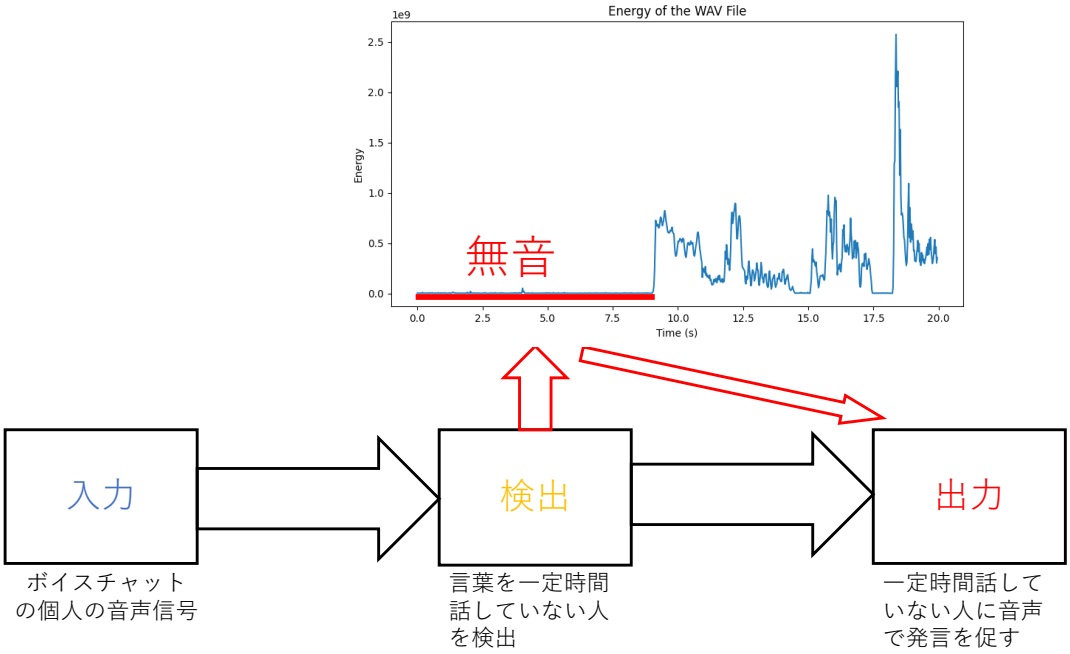


図 1　ベースラインシステムの流れ図

さが必要であるとも述べられている．

以上より，本研究の目的である，チーム対戦ゲーム中のボイスチャットにおけるチームメンバー間のお互いの認識の齟齬を減らすことができるインタフェースを検討する際には，これらの事柄を考慮する必要がある．

2.2　対戦ゲームのインタフェースに関する研究

対戦ゲームのインタフェースに関する研究は比較的行われている．小澤ら [8] は，FPS ゲームにおける非熟練者を対象としてゲーム画面に表示されているミニマップに着目させるための音声ガイドシステムを作成し実験を行った．非熟練者がミニマップに着目できるようになった一方で，FPS というゲームの性質上足音を始めとする環境音を聞きたい場面でシステムが妨げになることを問題点として挙げていた．

したがって，本研究ではゲームをしながら発言促進インタフェースを使用する状況が想定されるので，ゲームプレイヤのプレイ環境を大きく損なわないことも重要である．

2.3　発言促進に関する研究

本節では人間の発言を促進させる研究について述べる．

発言促進に関する研究は主にグループディスカッションや教育現場等を対象に行われている．特に発言を促進させたい対象としては議論に積極的に発言して参加していない消極的参加者に対して研究されている．塩津ら [9] は対面会議で参加者の発言機会を増やすためにテキストチャットを併用し匿名で意見を聞きたい人にテキストで意思を伝えてそれが全員に伝わる機能を用いて実験を行い，人間関係を悪化させることなく発言をしない人の発言を引き出すことができている．対戦ゲームにおいては 2.1 節で述べたようなコミュニケーションを行うためにも人間関係を悪化させないことは重要である．

鍋谷ら [10] は各議論者の発言率を測定し，条件に応じて各議論者に対し個別に指示を送るというシステムを作成し，結果としては消極的発言者の発言の手助けにつながっていた部分もあるが，一方でシステムを好意的に思わない意見もあった．また被験者の負担の軽減についても考えられていた．

本研究では対戦ゲームをプレイしながら会話が必要となるため，より短い時間で意思決定を行う必要があるのでよりリアルタイムなシステムにする必要があると考えられる．

3　ベースラインシステム

本章では 4 章で評価実験を行うにあたり，実験に用いるベースラインシステムについて述べる．ただし，これから述べる手法は簡易的なベースラインとなるシステムであり，インタフェースの改善を検討していくにあたっての初歩的なものであることを付け加えておく．

3.1　ベースラインシステムの意義

2 章で述べたことからチーム対戦ゲームにおける発言促進インタフェースを検討するには様々な要素が必要である．本稿でとるアプローチを簡潔に書くと「会話中の沈黙時間を減らすことでお互いの考えを伝える」ことである．ここで 2.1 節で述べたようにコミュニケーションすなわち会話が多いことは良くないと考えられるが，2.3 節で述べたように元々議論の場でも会話が多くなることは少なくお互いに話せる仲や会話を続けるのに慣れていない限りは発言が少ない人がいるのは不自然なことではないと考えられる．したがって，まずは発言を増やすことで対戦ゲーム中でも雑談だけでなく勝ちを目指して会話ができるようにしてコミュニケーションに慣れてもらうことがシステムの意義や役割である．

3.2 ベースラインシステムの機能

本節ではベースラインシステムの機能について述べる．本システムの機能は，チームメンバーの中から一定時間話していない人を検出し，検出するとその人に発言を促す音声を個別に流す処理で構成される．ベースラインシステムの概要図を図 1 に示す．またシステムは個人個人に対して使用される．

本システムの入力としては VC（ボイスチャット）での会話で，そこで会話していない無音の時間，言い換えると沈黙状態である人がいればその一定時間が経った後に音声が流れるようになっていて，この動作がシステムの動作時間ごとに繰り返される．具体的に言うと例えば動作時間が 10 秒とすると，動作している 10 秒間無音状態となっていれば検知されて 4 章で述べるが，本稿では，対象ゲームとして将棋を用いて実験を行うため「次何指しますか？」という音声が流れるようになっていて，これにより発言を生み出すというのがシステムの狙いである．

4 評 価 実 験

本章では評価実験に関する事項を詳述する．

4.1 実 験 目 的

本実験での目的はシステムを用いることで会話量に変化があるかを見ながら被験者に対してどのような効果や影響があるかを調べる．また，実験終了後にアンケートやインタビューを行うことで参加者の主観による会話量の変化やシステムに対する感想を得ることである．

4.2 実 験 対 象

本実験で扱うゲームの対象として将棋を用いる．将棋は一手一手区切りがあり，基本的にゲームスピードはそれほど速くはないが，4.4 節で述べるルールによって対戦ゲームに求められる要素として短い時間で判断することや読み合い等があるのでその部分をよりゲームスピードの速い対戦ゲームに近づけた状態で実験を行うことが出来る．

4.3 本稿のルールでの将棋と eSports の対戦ゲームの違い

本稿のルールでの将棋と eSports のチーム対戦ゲームの違いとしては，将棋は明確に手番が区切られているが，eSports のチーム対戦ゲームは明確な手番の区切りがないものが基本で常に時間的に連続している．また，本稿での 3 人で協力して行う将棋は全員同じ局面が共有されているので，各プレイヤは同じ条件で次の一手を考えることが出来るのに対し，eSports の対戦ゲームは，個々人の受け取る情報が同じではないゲームが多い．より詳しく説明すると，eSports のチーム対戦ゲームは，特に一人称視点のシューティングゲームである FPS (First Person Shooter) のチーム対戦であれば，各々のチームメンバーが操作するキャラクタの位置が異なることで見えている画面が完全に同じではなく，さらに位置によって足音や銃声の音が聞こえたり，聞こえなかったり，視覚だけでなく聴覚も異なるのが基本であり，自分自身以外の状況を完全に把握することが難しく，そういった状況で VC によって，連携する必要がある．

図 2 ぴよ将棋設定画面

4.4 実験方法と手順

実験に使用する将棋ゲームソフトはぴよ将棋 [11] とする．実験は 3 人 1 チームとして行う．実験の問題設定を以下に箇条書きでまとめた．

（1） 将棋のルールはぴよ将棋に則る．
（2） 相手は CPU で強さは被験者の段級位の平均とする．
（3） 明らかに強さに差がある場合は適宜強さを調整する．
（4） 持ち時間なしで 1 手 60 秒とする．
（5） 3 人で指す順番を任意の方法で決める．
（6） 1 人目→ 2 人目→ 3 人目→ 1 人目... の順で手を指す．
（7） 順番が回った人は 7 六歩など指す手を声で発し動かす．
（8） 駒の操作者は 1 人で他の 2 人には画面共有を行う．
（9） 駒の操作者は 1 人目の人が行う．
（10） 3 人で話す内容は自由である．
（11） 会話音声と対局画面以外の情報は分からない．
（12） 実験システムの動作時間は 10 秒とする．

ぴよ将棋の設定画面を図 2 に示す．また，実験環境図を図 3 に示す．図 3 から分かるように 3 人の被験者はそれぞれ異なる部屋にいて，Zoom などのオンライン会議システムを使用して将棋の対局画面を 3 人に共有した状態で，お互いに音声通話によりコミュニケーションをとりながら，ゲームをプレイしてもらう．理由としては，ベースラインシステムを個々人に対して使う都合上同じ部屋であると，ある被験者の声が他の被験者のマイクに流れてしまうことを防ぎたいからである．実験の問題設定に関してであるが，3 人で話す内容は自由であるというのは実験中に会話される内容には様々なことが考えられるが，例えば具体的なアドバイスをしたり，チームメイトに別々の候補手を読んだりするように言うことも可能である．会話音声と対

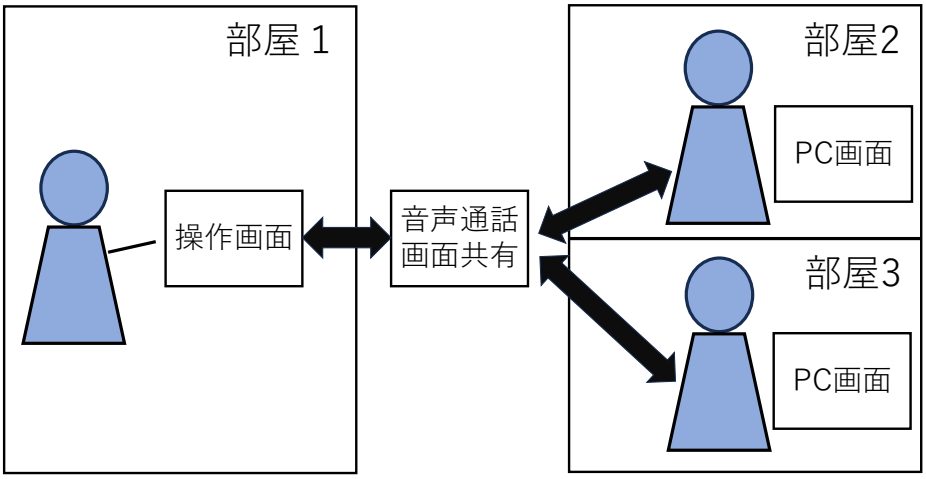


図 3　実験環境図

局画面以外の情報は分からないというのは対戦ゲームにおける状況を想定し，被験者は他の被験者の音声は聞こえるが表情が分からない状態で実験を行うということである．そのため 3 人で音声通話するソフトを用いて対局画面を共有しながら CPU と対局を行わせる．システムから発される声に関しては被験者ではない人の音声を用いる．本稿では著者の音声を用いた．

また，実験の詳細な手順について述べる．

（1） 被験者に実験の概要やおおよその流れなどをスライドと口頭で説明する．
（2） 被験者のノート PC のサウンドレコーダーと OBS Studio [12] の録画をオンにする．
（3） 無音の閾値を決めるために被験者にはいつも通りの話し声で「よろしくお願いします.」と言ってもらう．3 回発声してもらい，一番 dB が低かったものから 3〜5dB 程度低くした値を閾値の dB として設定する．
（4） ルールに慣れてもらうためのテスト対局を持ち時間 5 分で行う．
（5） ZOOM のリンクの共有を行う．
（6） 実験の注意事項の説明を行う．
（7） 3 人同時にシステムあるいは実験のデータを記録するプログラムファイルを実行してからすぐに対局開始ボタンを押す．
（8） 1 局目を行う．
（9） 対局終了直後にファイルの実行を止めて，インタビューとアンケートの回答を行ってもらった．インタビューは 1 人 3 分程度で行った．
（10） 手順 7，8，9 を 3 局目まで繰り返す．なお 1，3 局目はシステム無しで行う．
（11） 3 局目のインタビューとアンケートの提出が終われば，研究や実験の目的を説明する．
（12） システムを使わないグループとも比較．

図 4　実験の様子

ステップ 3 でそのままの dB の値を閾値として設定しない理由としては，通常の大きさの声でも無音と見なされる可能性を防ぐためである．また，3 局目でシステムを用いず再度プレイしてもらう理由としては，システムの効果がどれぐらい残っているかを確認するというものがあるので，研究の目的としては認識の齟齬が起きるのを減らすことであるが，それをシステムに頼らず自分から意思疎通するというのも研究の目的を達成するための一つの必要な要素と考えられるのでこの手順に沿って実験を行う．

また，実際の実験の様子を図 4 に示す．

4.5 評価項目

評価項目に関しては本実験では客観的，主観的な評価項目がある．まずは客観的な評価項目を箇条書きで示す．

（1） 一人一人の会話量とその変化．
（2） 発言促進の音声が流れた回数．
（3） システムを用いないときに 10 秒間無音であった回数．

表 1　Q1 アンケートの内容

項目	内容
Q1-1	自分は積極的に発言できた
Q1-2	被験者 1 は積極的に発言していた
Q1-3	被験者 2 は積極的に発言していた
Q1-4	自分より被験者 1 の方が発言していた
Q1-5	自分より被験者 2 の方が発言していた

表 2　Q2 アンケートの内容

項目	内容
Q2-1	1 局目より積極的に発言できた
Q2-2	1 局目より認識の齟齬なく対局できた
Q2-3	システムは意思疎通するのに役立った
Q2-4	システムによって普段より発言できた
Q2-5	システムは対局する上でストレスであった
Q2-6	システムによって自分の考えを伝えれた
Q2-7	システムなしでも自分の考えを伝えれた
Q2-8	システムがない方が発言していた
Q2-9	被験者 1 は 1 局目より積極的に発言できた
Q2-10	被験者 2 は 1 局目より積極的に発言できた

表 3　Q3 アンケートの内容

項目	内容
Q3-1	1 局目より積極的に発言できた
Q3-2	1 局目より認識の齟齬なく対局できた
Q3-3	被験者 1 は 1 局目より積極的に発言できた
Q3-4	被験者 2 は 1 局目より積極的に発言できた

次に被験者の主観的な評価項目を箇条書きで示す.

（1） 意思疎通して認識の齟齬なく対局できたか.

（2） 認識の齟齬をなくすのにシステムは役立ったか.

以上の項目をアンケートやインタビューを交えた実験で評価する．アンケートは Q1 から Q4 まであり，Q1 は各対局後に発言できていたかと他の 2 人の被験者が発言していたかに関する内容で他の被験者 2 人それぞれに対して被験者の評価を答えてもらう．Q2 はベースラインシステムの評価や使用感に関する内容，Q3 は前回の対局より発言や意思疎通が行えたかに関する内容で 2，3 局目後に設けており，Q4 はベースラインラインシステムの残存的な効果が見られるかについての内容で，また回答形式は Q1 から Q4 は 5 件法で答えてもらう．5 件法とは，回答者に 5 つの選択肢から回答を選んでもらう，いわゆる 5 段階評価のことで本稿では数字が大きいほど評価が良いものとして,「5. とてもそう思う」,「4. そう思う」,「3. どちらともいえない」,「2. そう思わない」,「1. 全くそう思わない」としてアンケートを行う．Q1，Q2，Q3，Q4 のアンケートの内容を表 1 から表 4 にそれぞれ示す.

表 4　Q4 アンケートの内容

項目	内容
Q4-1	1 局目より積極的に発言できた
Q4-2	システムの効果を実感できた

表 5　グループ 1 の段級位

被験者名	段級位（強さ）
A	5 段 (5 段)
B	5 段 (5 段)
C	初段 (1 段)
平均段級位	3.7 段

表 6　グループ 2 の段級位

被験者名	段級位（強さ）
D	3 段 (3 段)
E	3 段 (3 段)
F	3 級 (-2 段)
平均段級位	1.3 段

4.6　実 験 条 件

本実験の対象者として，大学の将棋サークルから 6 人 2 グループで実験を行った．グループ 1 がベースラインシステムを使用したグループである．またそれぞれの被験者の段級位をグループごとにそれぞれ表 5 と表 6 に示す．なお，段級位の計算は初段を 1，1 級を 0，2 級を −1 のように表して，段級位の強さを n で表すと，n 段を n，n 級を $-n+1$ として，3 人の数値を合計し 3 で割った．段級位の基準としては被験者の多くがプレイしている将棋ウォーズ [13] を基準とした.

4.7　実験結果と考察

本節では実験結果とその考察について詳述する．本実験で客観的に数値化できることとしては，各対局の会話率，実験システムで発言促進の音声を流した回数，実験システムを用いないときに 10 秒間無音であった回数，アンケートの 5 件法の回答である．そして被験者に聴いたアンケートの結果も後述する．なおベースラインシステムを用いて実験したのはグループ 1 であり，各グループの結果と考察をケースに分けて 4.7.1 項と 4.7.2 項に述べる.

4.7.1　グループ 1 の実験結果と考察

グループ 1 の実験結果として，1 局目を表 7 と表 8 に，2 局目を表 9 から表 12 に，3 局目を表 13 から表 16 にまとめた．なお表中の会話率は

$$\text{会話率} = \frac{\text{無音でなかった時間}}{\text{総時間}} \times 100$$

として計算した.

表 7　グループ 1 の 1 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
A	10.95	64	1250
B	16.14	30	1270
C	13.04	25	1290

表 8　グループ 1 の 1 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	4.0
Q1-2	被験者 1 は積極的に発言していた	4.0
Q1-3	被験者 2 は積極的に発言していた	3.7
Q1-4	自分より被験者 1 の方が発言していた	3.0
Q1-5	自分より被験者 2 の方が発言していた	2.7

表 9　グループ 1 の 2 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
A	11.16	43	1180
B	15.90	21	1240
C	12.74	13	1280

表 10　グループ 1 の 2 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	3.7
Q1-2	被験者 1 は積極的に発言していた	4.0
Q1-3	被験者 2 は積極的に発言していた	4.0
Q1-4	自分より被験者 1 の方が発言していた	3.3
Q1-5	自分より被験者 2 の方が発言していた	3.0

表 11　グループ 1 の 2 局目の Q2 アンケートの内容

項目	内容	平均
Q2-1	1 局目より積極的に発言できた	2.7
Q2-2	1 局目より認識の齟齬なく対局できた	3.3
Q2-3	システムは意思疎通するのに役立った	2.7
Q2-4	システムによって普段より発言できた	2.7
Q2-5	システムは対局する上でストレスであった	2.0
Q2-6	システムによって自分の考えを伝えれた	2.0
Q2-7	システムなしでも自分の考えを伝えれた	4.0
Q2-8	システムがない方が発言していた	2.3
Q2-9	被験者 1 は 1 局目より積極的に発言できた	3.0
Q2-10	被験者 2 は 1 局目より積極的に発言できた	2.7

ここからはグループ 1 の考察について結果やインタビューを基に述べる．考察としては，1 局目は，会話率は若干異なるが，基本的に 3 人の認識は皆同じように喋れていたと考えている．インタビューでもお互い考えを伝えることはできている認識であった．被験者 A は「被験者 C は将棋のいつも指している戦法と違って慣れていなかった.」という指摘があったり，被験者 C からは「無難でマイナスにならない手を指すことを意識していた.」と考えていたりしたことが分かった．

表 12　グループ 1 の 2 局目の Q3 アンケートの結果

項目	内容	平均
Q3-1	1 局目より積極的に発言できた	3.7
Q3-2	1 局目より認識の齟齬なく対局できた	3.0
Q3-3	被験者 1 は 1 局目より積極的に発言できた	4.3
Q3-4	被験者 2 は 1 局目より積極的に発言できた	4.7

表 13　グループ 1 の 3 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
A	14.79	35	1380
B	19.45	12	1380
C	13.25	15	1370

表 14　グループ 1 の 3 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	4.0
Q1-2	被験者 1 は積極的に発言していた	4.7
Q1-3	被験者 2 は積極的に発言していた	4.0
Q1-4	自分より被験者 1 の方が発言していた	3.7
Q1-5	自分より被験者 2 の方が発言していた	3.0

表 15　グループ 1 の 3 局目の Q3 アンケートの結果

項目	内容	平均
Q3-1	2 局目より積極的に発言できた	4.0
Q3-2	2 局目より認識の齟齬なく対局できた	3.3
Q3-3	被験者 1 は 2 局目より積極的に発言できた	4.0
Q3-4	被験者 2 は 2 局目より積極的に発言できた	4.0

表 16　グループ 1 の 3 局目の Q4 アンケートの結果

項目	内容	平均
Q4-1	1 局目より積極的に発言できた	4.0
Q4-2	システムの効果を実感できた	3.0

2 局目は，1 局目と異なり，システムを利用して実験したが，システム無しの表 7 とシステム有りの表 9 を比べると全体的に 10 秒間無音の時間が減っており，何らかの言葉を発する頻度が増えたと考えられる．実際にインタビューで被験者 B は「意識的に相槌をいれるようになって，音声が流れたときに話そうと思いながらやっていた.」と回答していた．また，被験者 C は「システムによってそそのかされる部分があった.」と回答しており，システムによる影響は 0 ということではなかった．

ただし，表 11 から分かるように，被験者の主観によるアンケートや表の数値的にはシステムの有無による影響は大きくないと言える．このような結果になった原因としては，元々この 3 人での意思疎通が問題なかったことに加えてルールにも慣れてきたことと，インタビューによって分かったが，被験者側が CPU に対して有利なまま対局を進めることが出来ており，焦りなく話せて大きなミスなく出来たとも話していた．

3 局目は，1 局目と同様，システム無しであるが，CPU の強さを 1 段程度強くして実験を行った．1，2 局目は余裕を持って対局していたため相手の強さを変えてどのような影響が起こったかも含めて述べていく．

表 17　グループ 2 の 1 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
D	33.43	8	1130
E	57.11	4	1140
F	26.54	18	1130

表 18　グループ 2 の 1 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	5.0
Q1-2	被験者 1 は積極的に発言していた	4.7
Q1-3	被験者 2 は積極的に発言していた	4.7
Q1-4	自分より被験者 1 の方が発言していた	3.0
Q1-5	自分より被験者 2 の方が発言していた	2.7

表 19　グループ 2 の 2 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
D	30.35	8	1370
E	59.20	1	1380
F	31.43	4	1370

表 20　グループ 2 の 2 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	4.7
Q1-2	被験者 1 は積極的に発言していた	4.7
Q1-3	被験者 2 は積極的に発言していた	5.0
Q1-4	自分より被験者 1 の方が発言していた	3.3
Q1-5	自分より被験者 2 の方が発言していた	3.3

表 21　グループ 2 の 2 局目の Q3 アンケートの結果

項目	内容	平均
Q3-1	1 局目より積極的に発言できた	3.7
Q3-2	1 局目より認識の齟齬なく対局できた	3.0
Q3-3	被験者 1 は 1 局目より積極的に発言できた	4.3
Q3-4	被験者 2 は 1 局目より積極的に発言できた	4.7

まず，全体的に表 13 から分かるように 1，2 局目と比べて会話率が上昇しており，10 秒間無音の回数が減っている傾向にある．この原因としては，強さを変えたことにより，被験者にも焦りが生じていたのが原因として考えられる．加えて，表 14 と表 15 から分かるように 1，2 局目と比べてより発言して対局をしていた傾向があった．それから，表 16 のシステムの効果に関しては特に変化を感じていないことが分かった．これらに関する各被験者のインタビューは被験者 A が「レベルが上がって考えることが増えて，それにより話さないといけないことが増えた．また，皆のやりたいことが散らばると話さないといけなくなる.」と話していて，強さを変えて会話の量が増えたことを感じていた．被験者 B は「システムによる変化はそこまで感じなかった．また，被験者 A は良く喋っていた．お互い納得して指せた.」と話していて，相手は強くなったが，あまり変化なく出来たと感じていた．そして被験者 C は「意思疎通を図るのが難しかった．被験者 A,B の方が棋力が高いことから判断を仰いでしまっていた．どうしようどうしようとなって発言が多くなっていた.」と話していて，相手が強くなって判断の難しい場面が増えて他の被験者に頼っていたということが分かった．

表 22　グループ 2 の 3 局目の結果

被験者名	会話率 (%)	10 秒間無音の回数	総時間 (s)
D	28.65	13	990
E	58.30	0	1010
F	26.15	10	1000

表 23　グループ 2 の 3 局目の Q1 アンケートの結果

項目	内容	平均
Q1-1	自分は積極的に発言できた	5.0
Q1-2	被験者 1 は積極的に発言していた	5.0
Q1-3	被験者 2 は積極的に発言していた	5.0
Q1-4	自分より被験者 1 の方が発言していた	3.0
Q1-5	自分より被験者 2 の方が発言していた	2.3

表 24　グループ 2 の 3 局目の Q3 アンケートの結果

項目	内容	平均
Q3-1	2 局目より積極的に発言できた	4.0
Q3-2	2 局目より認識の齟齬なく対局できた	4.3
Q3-3	被験者 1 は 2 局目より積極的に発言できた	4.0
Q3-4	被験者 2 は 2 局目より積極的に発言できた	3.7

ただ，自分が次の手は何を指せば良いか分からないなと思う状況でも発言量を増やして，他の被験者を頼ることで認識の齟齬が減り，意見を聞きつつ対局が出来ているので，その点では 1 手 60 秒という時間的な余裕こそあるかもしれないが，本研究で求めている意思疎通が十分になされていたと考える．ただ，eSports の対戦ゲームの場合では，全員の状況が必ずしも同じではないため，チームメイトに意見を言ったり，聞けたりする状況でない時も大いに有り得るため，この点についても考えておく必要がある．

4.7.2　グループ 2 の実験結果と考察

3 局ともシステム無しであった，グループ 2 の実験結果を 1 局ごとにまとめた．1 局目は表 17 と表 18 にまとめ，2 局目は表 19 から表 21，3 局目は表 22 から表 24 にまとめた．

ここからはグループ 2 の考察について結果やインタビューを基に述べる．グループ 2 はベースラインシステムを利用していないグループである．

まず，表 17 などの結果から分かるように，グループ 1 と会話量が大きく異なるが，原因としては，グループ 1 はノート PC のマイク，グループ 2 はイヤホンのマイクで行っていたためであると先に述べておく．変更した理由として，グループ 2 の方が先に実験を行ったが，イヤホンのマイクは環境音やマイクの距離が一定でなく，音声の測定が不安定であったためである．

考察としては，1 局目は表 18 から分かるように 3 人とも同じぐらい積極的に発言できたと感じている．インタビューでは被験者 D が「被験者 F の人は棋力が違って話しづらかった．また，同じ棋力だとしても戦法や考え方によって滅茶苦茶になるのではないか.」と話しており，それぞれに実力差や考え方の違いがあると話を合わせるのが難しいことを指摘していた．被験者 F からは「自分の考えを伝えられたが，被験者 D が自分の指したい手と違う手を指す時があった」と話していた．これは，被験者 D が被験者 F と考え方が異なっていて自分の思っている

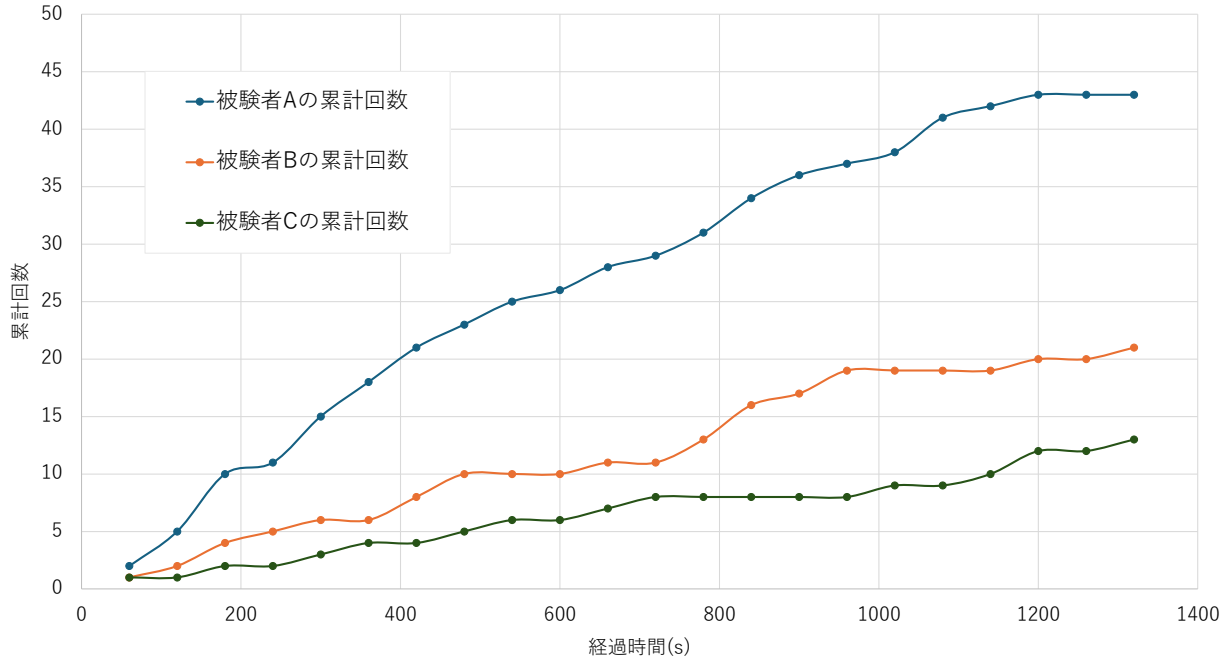


図 5　経過時間に対するシステム呼び出しの累計回数

ような手ではなかった時に被験者 D は自分の思っていた手を優先し，それを提案して指していた．したがって多少強引かもしれないが，時には自分の判断に従ってもらうのも試合を有利に進める上では重要ではないかと考える．

2 局目は，1 局目が余裕を持って対局できていたため，敵の CPU の強さを 1 段程度強くして行った．表 19 を見ると，会話率は増え，無音の回数は減っている．それから，アンケートに着目すると被験者は 1 局目より発言が増えたという認識であった．インタビューでは被験者 D が「被験者 F は 1 局目より発言していた．また，将棋の定跡から外れた戦いとなって時間内に考えるのが難しくなった.」と話していて，慣れていない形での対局で時間的な余裕がなかったことが伺える．また，被験者 E は「アンケートの結果は相対評価なので変わらないが周りはより喋っていた.」と話していて相手が強くなった分話すことが増えていたと考えられる．被験者 F は「被験者 D,E がより喋っていて聞きに回る時間が多くなっていた．相手が強くなって自分では判断が難しく，時間的な余裕も無く齟齬が増えていた.」と話していて，相手が強くなった影響が出ていたと言える．これは表 21 の認識の齟齬が 1 局目と比べて変化していなかったことからも言える．

最後に，3 局目は，表 22 を見ると，2 局目と比べて無音の回数が増えたり，会話の回数が減ったりしていたが，表 23 や表 24 のアンケートの結果を見ると，被験者の主観であれば 2 局目より喋っている傾向があり，インタビューでは，3 人とも「認識の齟齬が減っていた.」というのは共通しており，被験者 D は「知らない戦法だったので慣れていなかった.」と話していて，慣れていないことをするのは難しいと感じていて，被験者 E は「被験者 D に読みを任せていた.」と話していて，かなり被験者 D に負担がかかっていたと考えられる．

まとめると，基本的に意思疎通は取れてはいたが，完璧ではなかったので，チームで何度も対戦することを考えると考え方や実力を近づけることは，2.1 節で述べた関連研究 [6] のようにコミュニケーションを磨く上で必要なことであると考える．

4.7.3　ベースラインシステムに関する考察

本項では，ベースラインシステムに関する考察について述べる．グループ 1 の 2 局目にシステムを用いたが，4.7.1 項のように被験者の視点ではそこまで変わらなかったという認識をしていた．60 秒ごとに何回システムが呼び出されたかを累計回数で表したものを図 5 に示す．補足として，3 人の総時間が異なっているのは促進音声が流れる時間が加算されていないからである．時間的な分布を見ると被験者 C は全体的に促進音声が流れた回数が少なく，4.7.1 項のインタビューで述べていたようにシステムにそそのかされたということが分かり，恐らく常に自分の指したい手やどうすれば良いかを常に聞くことが出来ていた．逆に被験者 A に関しては，前半はあまり無音とみなされた時間が多く，後半は喋るようになっていて，恐らく将棋が終盤となることで，詰みや手を読んでそれを被験者 B，C に話す形になっていたので発言が増えたと推測される．よって，実験により，本稿でのシステムによる影響が出やすい対象と考えられるのは，チームの中で実力が相対的に低い人に対してであり，周りに実力がある人がいる中で，自分自身に自信を持って良い手を指し続けるのは難しいと考えられるため，そのようなプレイヤーにベースラインシステムを使用すれば，チームメイトと意思疎通を取る意識を持てるようになるのではと考える．

5 おわりに

本稿では，認識の齟齬を減らすために個々人の VC（ボイスチャット）中に一定の沈黙時間が発生すると，「次何指しますか？」などと発言を促す音声を個別に流すシステムを作成した．このシステムを用いて，被験者に対象ゲームとして将棋をチームプレイさせる実験を行った．得られた実験結果からそれぞれのケースを分析し，考察を行った．考察により，実験結果から得られたデータからシステムをどのような人に向けて適用するかについては考えられたが，特に技術的には新しいあるいは高度なことは行っていないので，技術的な貢献を行っていくことは今後の課題となる．

結果としては，第一に実験のデータを得られたことが大きく，被験者の会話やインタビューに対して焦点を当てて，対戦でどのようなことを考えてやっていたかなどを詳述した部分は一つ研究で貢献できた部分であると考える．また，チームメイト同士のコミュニケーションに関する既存の論文はまだ少ないことから，本稿を世に出せたのは，今後日本でも拡大していくであろう eSports に関する論文の一つとして，貢献出来たのではないかと考える．加えて，本研究で扱ったようなタスクはコンピュータ支援協調作業（CSCW）の一つであり，今後，CSCW あるいは工学分野にも技術的に貢献していければと考えている．

それから，今後の研究でのインタフェース改善の方針としては，2 章で述べたことや実験結果を踏まえつつ，インタフェースのシステムがプレイヤに何らかの発言促進を行う時に，そのタイミングを制御させることを考えている．結果として，発言の少ない人がインタフェースに依存せず，自らチームメンバーに意思疎通が出来るようになることを見据えている．改善手法として 2 つの案を述べる．1 つ目は，関連研究 [8] のようにルールベースでゲームの状況に応じてタイミングを制御し，プレイヤに発言促進させる音声を流すインタフェースを考えている．また，音声による発言促進ではなく，他の媒体としてテキストをゲーム画面に表示して発言促進を行うことも考えられる．2 つ目は，佐久間ら [14] のように機械学習を用いて人が話すべきタイミングを推定し，制御する方法である．発話タイミングに関する手法に目を向けることにより，会話の内容に応じても発話すべきタイミングを制御が出来て，より汎用的なインタフェースになる可能性があると考えられる．

最後に，今後の課題としては，まずは本章の最初に述べたように工学的な技術を取り入れることから始める必要があると考えているのと，その上で被験者にインタフェースを使用してもらう評価実験が必要であると考えられるので，実験手順や方法，また評価指標に関してより洗練していくことも重要なことであると考える．

文　献

[1] 2022 年の国内 e スポーツ市場規模が 100 億円を突破「日本 e スポーツ白書 2023」の内容を先行公開，一般社団法人日本 e スポーツ連合オフィシャルサイト，2023-12-25 https://jesu.or.jp/contents/news/news-231225/ (参照 2024-01-05)

[2] José Pedro Pinto，André Pimenta，Paulo Novais，Deep learning and multivariate time series for cheat detection in video games，2021

[3] Alessandro Canossa，Dmitry Salimov，Ahmad Azadvar，Casper Harteveld，Georgios Yannakakis，For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay，2021

[4] Elizabeth Reid，Regan L. Mandryk，Nicole A. Beres，Madison Klarkowski，Julian Frommel，“Bad Vibrations”: Sensing Toxicity From In-Game Audio Features，2022

[5] なぜコミュニケーションエラーは起こるのか―世界と戦って感じた可能性『VALORANT』Crazy Raccoon ade 選手インタビュー，eSports World（e スポーツワールド），2021-09-17 https://esports-world.jp/interview/14483 (参照 2023-12-19)

[6] Oksala Aleksi，The importance of communication in esports: an ethnographic interview with a Finnish Counter-Strike: Global Offensive team，2022

[7] Counter-Strike 2，https://www.counter-strike.net/ (参照 2024-01-08)

[8] 小澤 蓮，梶並 知記，FPS ゲームにおける非熟練者向けのミニマップ着目音声ガイドシステム，日本デジタルゲーム学会 第 12 回年次大会予稿集，2022

[9] 塩津 翠彩，高島 健太郎，西本一志，LighthouseChat：消極的参加者に発言を促す手段を備えた会議支援メディア，インタラクション 2018 論文集，pp.260–263，2018

[10] 鍋谷 航平，村岡 泰成，石川 誠彬，江木 啓訓，消極的発話者の発言率向上を目的とした音声による個別指示議論支援システムの開発，マルチメディア，分散，協調とモバイル (DICOMO’21) シンポジウム，pp.766–773，2021

[11] ぴよ将棋 w – 対局将棋アプリ WEB 版，https://www.studiok-i.net/ps/ (参照 2024-01-09)

[12] Open Broadcaster Software，https://obsproject.com/ (参照 2024-01-30)

[13] 将棋ウォーズ公式–日本将棋連盟公認–オンライン将棋ゲームの決定版，https://shogiwars.heroz.jp/ (参照 2024-01-28)

[14] 佐久間 仁，藤江 真也，小林 哲則，Self-Attention を用いた多人数会話向け発話タイミング推定，言語音声理解と対話処理研究会，pp.38–41，2021

Multi-layer warning recognition system
駅構内における聴覚障がい者サポートのための マルチレイヤワーニング認識システムの実現

須藤光喜† 岡田 龍太郎† 峰松彩子† 中西 崇文†

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

E-mail: †s2222019@stu.musashino-u.ac.jp, {ryotaro.okada, ayako.minematsu, takafumi.nakanishi}@ds.musashino-u.ac.jp

あらまし 本稿では，音源分離並びに２種類の認識モデルを用いた聴覚障がい者のサポートのためのマルチレイヤワーニング認識システムの実現について示す．聴覚障がい者は，公共交通機関を利用する際に音響案内や警告音などの音情報が十分に取得できず，安全な行動を取ることが難しい．本研究の目的は，音情報を視覚的に案内できるシステムを構築することである．本方式では，人の声によるアナウンスや自動放送のアナウンスの内容などの音声認識モデルと危険を知らせる警笛などの音響認識モデルの 2 つを構築し，聴覚障がい者が公共交通機関を利用する際の危険を知らせるサポートシステムを実現する．本方式の特徴は人の声と音のアラートを分離し，提案モデルで認識することによって識別精度が増加するだけでなく，人の声と音のアラートが同時に流れる状況でもそれぞれ認識することが可能となる．警告情報を含む音情報をリアルタイムで識別し，モバイル端末等に視覚的に案内を表示するシステムが実現できれば，聴覚障がい者が安心して公共交通機関を利用する機会が増大することに貢献できると考える．

キーワード 聴覚障がい者，サポートシステム，音響分類，公共交通機関

1. はじめに

近年，聴覚障がい者の人口は増加傾向にあり，2020 年には約 230 万人に達すると推定されている．平成 30 年 4 月 9 日に公開された厚生労働省社会・援護局障害保健福祉部の『平成 28 年生活のしづらさなどに関する調査 （全国在宅障害児・者等実態調査）結果』の報告書[1]によると身体障害者手帳所有者ないし，65 歳未満の聴覚障がい者の数は 71,000 人，65 歳以上または年齢不詳では 271,000 人いるという．令和 4 年度にも調査が行われているが，結果はまだ公表されていないため，結果が公開されている中で最新なのが平成 28 年のデータとなっている．

その中で聴覚障がい者は，公共交通機関を利用する際に，音響案内や警告音などの音情報を十分に取得できず，安全な行動をとりにくいという課題がある．『「聴覚障害者」にとって通勤電車はこんなに怖い　車内放送聞こえず，何が起きたかわからない』の東洋経済新報社の記事[2] (2019/07/05)によれば，聴覚障がい者にとって通勤電車ではドアチャイムなどが聞こえずに意図せず駆け込み乗車になってしまう．また，難聴は後天的なものと先天性のものがある．後天的であれば，電車の音を「知っている」状態であるために，補聴器を装着することである程度聞き分けられる．一方で，先天性の難聴であれば，その音を「知らない」ためにどれが電車の音で，警笛の音であるかがわからない．それ以外にも他の人に気付かれづらいため他の人の助けが見込めない，他の人に気づかれるのが恥ずかしい等様々な課題がある．

一般的に，聴覚障がい者が公共交通機関を利用する際に，音響案内・警告音を含む音情報が十分に取得できないため，安全な行動を取ることが難しい．国土交通省の『「視覚･聴覚障害者の安全性･利便性に関する調査研究報告書」2011』[3]によれば『安全性・利便性を考えたときの重要度（聴覚障害者）』の調査項目で『D2-3.聴覚障がい者に配慮した設設備（文字情報の充実，光の点滅灯による注意喚起装置の設置など）を新たに設置した場合には，その対応した駅や施設を周知する』の結果が重要とやや重要で合計で 84.3%である．また，公共交通機関や公共施設で推進されている施策の重要度（聴覚障害者）の調査項目においても，先ほどあげたような事例を含む調査項目の重要度が高いことがわかる[4]．そこで聴覚障がい者向けに音情報を視認できるシステムを構築することを考えた．

本方式では，人の声によるアナウンスや自動放送のアナウンスの内容などの音響認識モデルと危険を知らせる警笛などの音響認識モデルの 2 つを構築し，聴覚障がい者が公共交通機関を利用する際の危険を知らせるサポートシステムを実現する．音声の場合には駅員のアナウンスか自動放送の 2 種類のどちらかを示したうえで，その内容をテキスト化する．音響なら「接近

チャイム，駅メロ，警笛」の 3 種類のどれかがある場合はその種類を示す．本方式の特徴は，人の声と音のアラートを分離し，提案モデルで認識することによって識別精度が増加するだけでなく，人の声と音のアラートが同時に流れる状況でも識別が可能となることである．

本システムの実現によって以下のような効果が期待できると考える．障がい者の方がほかの人と同様の生活を送れるようにする（バリアフリー）．恥ずかしいと感じる場面でも他の人に頼らない選択肢となる（第三の手段）．気軽に安全に利用できる仕組みの実現によって障がい者の活動の増大を支援する（積極性）．健常者に対しても混雑時に放送や音響を聞き逃しにくくなる．イヤホン等で耳を塞ぐこと多い方もいるためそういった方々の安全確保や放送の聞き逃しを防ぐ．

このように聴覚障がい者の社会参加の促進につながるだけではなく，公共交通機関の利用者の安全性や利便性が向上すると考える．

本稿は以下のように構成される．2 節では，関連研究について示す．3 節では，本方式である音源分離並びに提案モデルを用いた聴覚障がい者のサポートのためのマルチレイヤワーニング認識システムの実現方式について示す．4 節では，実際に構成したシステムを使用した実験，5 節で本稿をまとめる．

2. 関連研究

本節では，本方式に関連する研究について挙げる．ここでは，駅構内における音情報の可視化システムと音を振動と光を用いて知覚できるようにするデバイスの関連研究について挙げる．

駅構内における音情報を視覚的に表現するシステムとして，エキマトペ[5]が挙げられる．エキマトペは，富士通株式会社，大日本印刷株式会社，東日本旅客鉄道株式会社の共同プロジェクトによって開発され，現在，上野駅の自動販売機に設置されている．エキマトペは，駅のアナウンスや電車の音などの環境音を文字，オノマトペとして表現し，聴覚障がい者の方々が駅での情報を得やすくすることを目的としている．次に，音を振動と光で知覚できるようにするデバイスとして，Ontenna[6]が挙げられる．Ontenna は，富士通株式会社で開発され，聴覚障がい者が音の特徴を振動や光で感じることができる．聴覚障がい者の音に対する感覚を拡張し，新たなコミュニケーションや表現方法の可能性を提供するデバイスである．

2.1. 関連研究との相違点

システムの構成として，クラウド上あるいはアプリ等のオンデバイス環境を想定しているため，様々なデバイスでの使用が可能である．ブラウザベースの場合、オンライン上で実行するため端末のスペックに依存しない．また，場所を選ばすに使用ができるため，装置がない駅や乗降場所を変えずに使用できる利点が挙げられる．

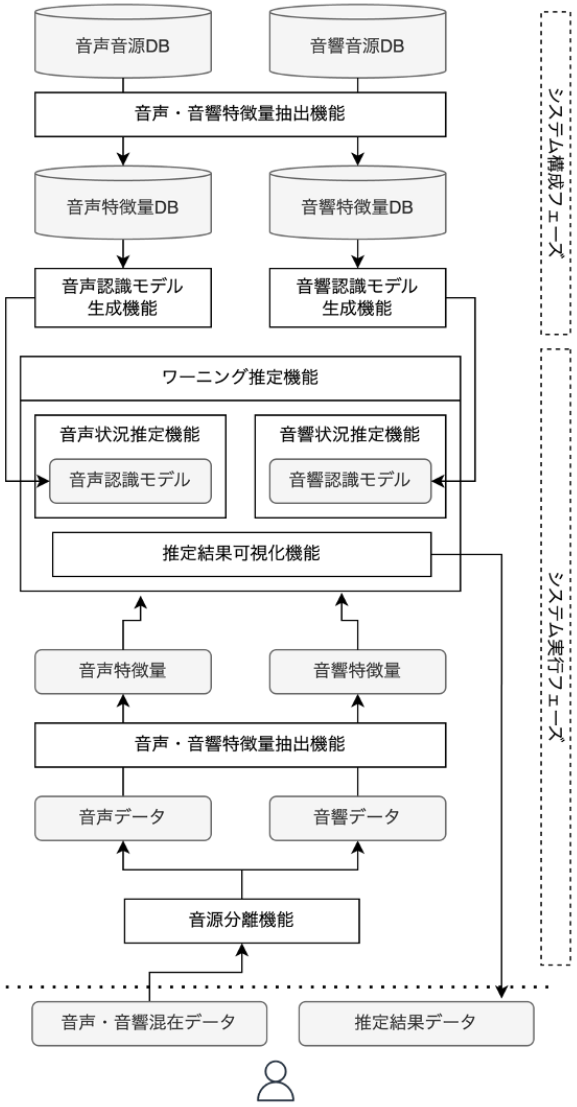


図 1　提案方式の全体像

3. Multi-layer warning recognition system　駅構内における聴覚障害者サポートのためのマルチレイヤワーニング認識システムの実現

本節では提案方式の構成方法について述べる．本研究では，音のデータのうち，人の声によるものを「音声」，声以外のものを「音響」として呼び分けることとする．

本方式は聴覚障がい者が駅で利用することを想定している．本方式は，人の声によるアナウンスや自動放送のアナウンスの内容などの音声認識モデルと危険を知らせる警笛などの音響認識モデルの 2 つを構築し，聴覚障がい者が公共交通機関を利用する際の危険

を知らせるサポートシステムを実現する．本方式では音声と音響という 2 つのカテゴリの音を認識し，それぞれの示す状況を判別したうえで，それらを集約して可視化する．これによって聴覚障がい者が駅で音が聞こえないことによる不便さや危険を，解消あるいは軽減することを目指している．

3.1. 全体像

図 1 に提案方式の全体像を示す．本方式は，大きく分けてシステム構成フェーズとシステム実行フェーズからなる．

システム構成フェーズでは，音声音源 DB と音響音源 DB を用意し，それらから音声データから状況を認識するモデルと音響データから状況を認識するモデルの２つを生成する．

システム実行フェーズでは，ユーザがマイクデバイスを用いて音声と音響が混在した 3 秒間のデータを入力し，それらの音源を分離したうえで，システム構成フェーズで生成したモデルを利用して音声と音響のデータからそれぞれ状況を推定し，推定結果データを画像や文章でユーザに提示する．

3.2. システム構成フェーズ

3.2.1. 音声・音響特徴量抽出機能

音声データと音響データを用いて状況の推定を行う認識モデルを構成する前段階として，それらのデータから特徴量を抽出する．特徴量としては，スペクトログラムの画像データを抽出している．抽出結果をそれぞれの音声特徴量データベースと音響特徴量データベースに保存する．

認識対象のデータとして，駅空間で観測される音声データと音響データをそれぞれ分けて収集し，音声データベースと音響データベースを構成した．

音声音源データベースには人の声と自動放送の音の二種類の音をそれぞれ 40 個ほど用意している．人の声は Mozilla Common Voice の日本語音響のオープンデータセットを用いている．自動放送の音源に関しては自前で収録したデータを使用している．自動放送のデータは東武スカイツリーライン，東武アーバンパークライン，東武東上線，JR 赤羽駅のデータを主に使用している．

音響音源データベースでは，警笛，駅メロ，接近チャイムの 3 種類をそれぞれ 20 個ほど収録して用意している．

3.2.2. 音声状況推定モデル生成機能

音声状況推定モデル生成機能は，入力された音声が人の声によるアナウンスであるか，自動放送のアナウンスであるかを判別するモデルを生成する機能である．ここでは，スペクトログラムの画像データセットに対して VGG19_bn の転移学習を行うことでモデルの生成を行う．入力として，人の声であるか自動音声であるかのラベルを付けた音声音源データベースのスペクトログラムの画像を読み込み，出力は音声状況推定モデルを構築し，保存する．

3.2.3. 音響状況推定モデル生成機能

音響状況推定モデル生成機能は，危険を知らせる警笛などの音響認識モデルを作成する機能である．ここでは，スペクトログラムの画像データセットに対して VGG19_bn の転移学習を行うことでモデルの生成を行う．入力として，音響音源データベースのスペクトログラムの画像を読み込み，出力は音響状況推定モデルを構築し，保存する．

3.3. システム実行フェーズ

3.3.1. 音源分離機能

音声分離機能は，マイクデバイスから取得した音声と音響が混在したデータから，音声データと音響データに分ける機能である．分離のためのライブラリとして Meta 社の Hybrid Demucs[11][12][13]を用いる．マイクデバイスとはユーザ側のモバイル端末（スマホやタブレット，pc 等）の音響入力機能を備えたデバイスを指す．入力データとはマイクデバイスから取得した 3 秒の 16kHz のサンプリングレート，16bit の量子化深度で記録した音源を指す．出力は分離された音声データと音響データとなる．

3.3.2. 音声・音響特徴量抽出機能

音源分離機能で出力された音声データと音響データに対して，システム構成フェーズと同様に音声・音響特徴量抽出機能を用いて，それぞれ音声特徴量と音響特徴量を抽出する．方式については 3.2.1 節と同様である．

3.3.3. ワーニング推定機能

ワーニング推定機能では，システム構成フェーズで生成した音声状況推定モデルと音響状況推定モデルを用いて，入力されたデータから分かるワーニング(警告)の情報を推定し集約する．本機能は，a) 音声状況推定機能，b)音響状況推定機能，c) 状況推定結果可視化機能の 3 つからなる．

a) 音声状況推定機能

人の声によるアナウンスや自動放送のアナウンスの内容などを認識する音声認識モデルを用い，音声データの特徴量であるスペクトログラムを入力することで音声状況を推定する．音声の場合には駅員のアナウンスか自動放送の 2 種類のどちらかを示す．更に，OpenAI が提供する音声認識ライブラリである whisper を用いてその内容をテキスト化する．

b) 音響状況推定機能

危険を知らせる警笛などを認識する音響認識モデルを用い，音響データの特徴量であるスペクトログ

ラムを入力することで,「接近チャイム，駅メロ，警笛」の 3 種類のどれかがある場合はその種類を示す.

c) 状況推定結果可視化機能

入力として,音声状況推定機能と音響状況推定機能の推論結果を受け取り,出力で画像や文章で状況を可視化した画面を表示する.音声状況推定機能，音響状況推定機能から受け取った推論結果を基に対応した画像や文章を表示することで表題のマルチレイヤーワーニングシステムの実現を目指す.音声状況推定機能の結果に基づいて,自動放送の案内内容を音響で読み上げ,画面表示を行うシステムの構築を目指せる.また，音響状況推定機能の結果に基づいて，警笛や駅メロ，接近チャイムの種類を画面に表示するシステムの構築を目指す.

4. 実験

本研究では，聴覚障がい者向けの音情報を視認できるシステムのモデルの 1 つの場合と 2 つの場合の精度検証や人の声状況推定機能，音響状況推定機能の精度を評価するために，以下の実験を行う．なお，特に注釈がない限り，計算機器環境は表 1 の環境でモデル構築並びに精度検証等を行う．計算機およびコンパイラに関する記述は， 表 1 にまとめることとする．また、特に断りがない限り、単体モデルは音声と音響を統合的に処理するモデルをいい、提案モデルは音声と音響をそれぞれ個別に処理するモデルを指す。

4.1. 実験 1

4.1.1. 実験 1-1

本節では，複数の転移学習手法において単体モデルの場合と提案モデルの場合での精度検証を行う．精度検証ではデータベースをそれぞれ訓練データとテストデータを 8:2 で分け，モデルを構築する．この作業を 5 回行った結果の平均値を単体モデルの場合，提案モデルの音響モデルの場合，音声モデルの場合の精度をそれぞれ表 3 にまとめることとする．なお，モデルのハイパーパラメータとしては表 2 にまとめることする.画像処理は前処理としては，Tensor 化を行い，224*224 にリサイズ，正規化の順で行った．データ数では単体モデルの場合ではそれぞれ項目で 20 個用意し，全部で 100 個用意した．提案モデルの場合で，音響モデルでは各項目 20 個，計 60 個のデータを用意した．音声モデルでは各項目 40 個,計 80 個のデータを用意した.学習手法として，転移学習を行った．手法としては，Resnet152[14], VGG19_bn, vision transformer 16(以降，ViT16 とする．) [15][16]　の計 3 種で行った.

表 1:計算機器環境

CPU	Core i5-1230U
GPU	Nvidia RTX A500
OS	Windows 11 Home 23H2
メモリー	16[GB]
言語	Python 3.11.5
仮想環境	Anaconda 23.9.0
ライブラリ	PyTorch 11.8

表 2:ハイパーパラメータ

バッチサイズ	32
学習率	20
最適化関数	Adam

表 3: 交差検証における各モデルの正解率平均値（それぞれ 5 回計測）

転移学習手法	単体モデル	音声モデル	音響モデル
ResNet152	68.53%	97.76%	80.00%
VGG19_bn	88.33%	92.80%	79.80%
ViT16	40.00%	89.29%	37.89%

単体モデルで作成した場合では，ViT16 の精度が他の結果と離れていることがわかる．また，このモデルにおいて，VGG19_bn の精度が一番高い結果となった．次に音響モデルの精度では，僅差ではあるものの ResNet152 の精度が一番よかった．しかし，VGG19_bn と 0.2%ほどしか,違わず 5 回の検証でも同じような結果を出していたため，どちらの手法でも精度に関しては変わらないと考える.音声モデルでは,ResNet152 の精度が高い．しかし，検証の際には 5 回中 4 回 100%の精度を出しており，過学習がしていることがわかった．これより，音声モデルは VGG19_bn が適しているとなるものの,Vit16 でも 0.2%ほどしか違わないため,どちらの手法でも精度に関しては変わらないと考える.このように，精度検証では誤差の範囲の値が多く，手法選択の際の決め手に欠ける.

4.1.2. 実験 1-2

4.1.1.実験 1-1 の結果だけではシステムに組み込むモデルの手法選定の決め手に欠ける．そこで，モデルの学習時に学習回数毎に training loss と Validation loss を計測し,記録した.Training Loss と Validation Loss は，転移学習モデルの学習過程を監視し，評価するための重要な指標である．本稿では，これらの指標の定義，役割，解釈について論じる.

Training Loss は，モデルが学習データに対してどれほど適合しているかを示す指標である．学習データに対

してモデルの予測値と正解値との誤差を測定し，その平均値として算出される．式で表すと，以下のようになる．

$$Training\ Loss = \Sigma(予測値 - 正解値)^2 / 学習データ数$$

学習過程において，Training Loss が減少していくことは，モデルが学習データに対して適合しつつあることを示唆する．
Validation Loss は，モデルが未知のデータに対してどれほど汎化性能を有しているかを示す指標である．学習には使用されていない検証用データに対してモデルの予測値と正解値との誤差を測定し，その平均値として算出される．式で表すと，Training Loss と同様である．

$$Validation\ Loss = \Sigma(予測値 - 正解値)^2 / 検証データ数$$

Validation Loss は，Training Loss とは異なり，学習過程に直接影響を与えない．その役割は，モデルの汎化性能を監視することにある．

Training Loss と Validation Loss の推移を分析することで，モデルの学習状態を把握し，適切な学習戦略を立てることが可能になる．以下に，代表的なケースを示す．Training Loss と Validation Loss が共に減少する場合：モデルは適切に学習しており，汎化性能も向上している．Training Loss は減少するが，Validation Loss は停滞または上昇する場合： 過学習の兆候である．モデルが学習データのノイズや偏りに過度に適合し，汎化性能が低下している可能性がある．Training Loss と Validation Loss が共に停滞する場合：モデルは学習が進んでおらず，アンダーフィッティングの可能性がある．Training Loss と Validation Loss は，転移学習モデルの学習過程を適切に評価し，過学習やアンダーフィッティングを防ぐために不可欠な指標である．これらの指標を効果的に活用することで，より汎化性能の高いモデルを構築することが可能になる．この結果を用いてそれぞれシステムに組み込み次節にて再現率や適合率を計測する．記録には mlflow 2.8.0[17]を使用した．なお，再現率と適合率に関しては次節にて述べる．単体モデルの場合，提案モデルの音響モデル，音声モデルの training loss と Validation loss それぞれ抜粋した結果の図をそれぞれ，図 2, 3, 4, 5, 6, 7 に示す．

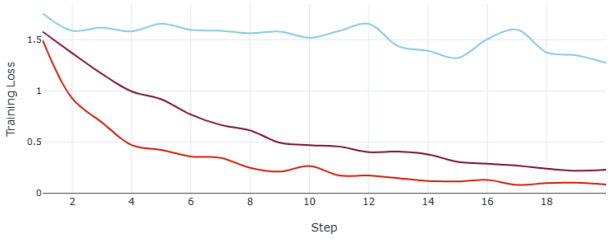


図 2: 単体モデルの場合の Training loss

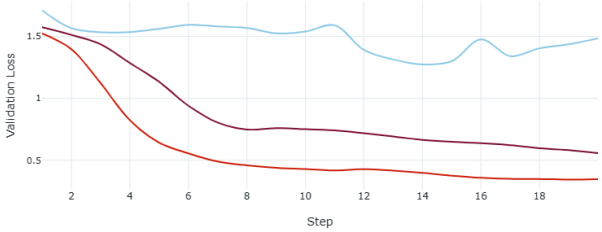


図 3: 単体モデルの場合 Validation loss

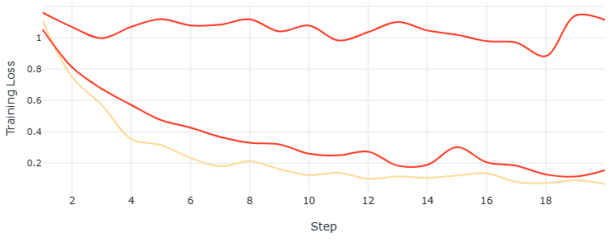


図 4:音響モデルの Training loss

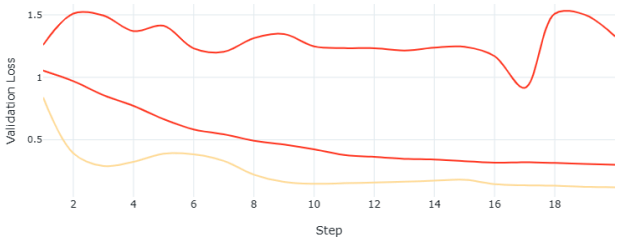


図 5:音響モデルの Validation loss

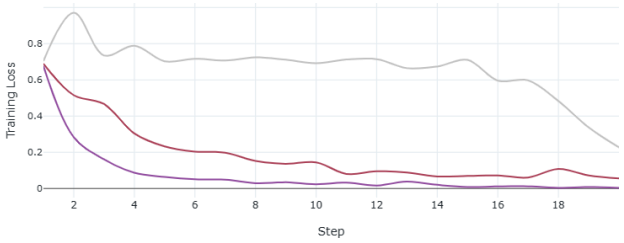


図 6:音声モデルの Training loss

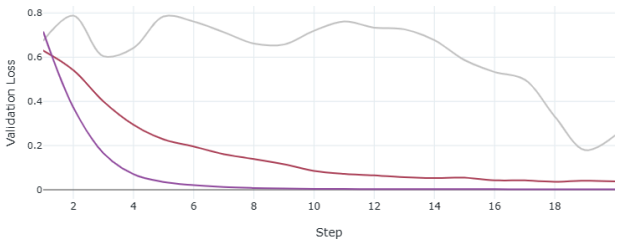


図 7:音声モデルの Validation loss

まず，単体モデルの場合でのモデルの手法選択について記す．表 3 の結果では VGG19_bn の精度が一番高いことがわかる．ResNet152 も精度はそれなりにあるものの，他のモデルと比較するとその差が顕著であった．図 2, 3 のいずれでも上から順に ViT16, RseNet152, VGG19_bn の順で推移していることを表している．VGG19_bn と ResNet152 の両方で収束していることがわかる．VGG19_bn の値が一番低いことがわかる．一方で，ViT16 では不安定であり，収束せずに一定の値で推移している．特に Validation loss では学習率 20 回の時点で他のモデルと異なり，上昇していることが確認できる．単体モデルの場合において，ViT16 は実験を行った条件下では汎化性能が低いと考える.よって，一つのモデルの場合では VGG19_bn が最適であると判断した．次に提案モデルの場合の音響モデルの手法選択について記す．表 3 の結果では，ResNet152 と VGG19_bn の両手法が精度はよく，結果の差も 0.3%と非常に小さい．図 3, 4 のいずれでも上から順に ViT16, RseNet152, VGG19_bn の順で推移していることを表している．いずれの結果でも VGG19_bn が小さい値を示している．このとき，training loss の値が小さいことからモデルの学習性能が正解と一番近いのが，VGG19_bn であることがわかる．また，Validation loss も一番低い．表 3 と図 3,4 の結果から音響モデルにおいてシステムに組み込むモデルでは VGG19_bn が適切であると判断した．最後に音声モデルの手法選択について示す．表 3 の結果より，ResNet152 が一番精度が高いものの，前節で書いたように，過学習を起こしている．VGG19_bn と Vit16 の両手法ではどちらも 90%台であり, 誤差は 0.2%とこれも先ほどの音響モデルと同様にとても小さいことがわかる．図 6, 7 では先ほどと同様に上から順に ViT16, RseNet152, VGG19_bn の順で推移していることを表している．いずれの結果でも VGG19_bn が小さい値を示している. このとき, training loss の値が小さいことからモデルの学習性能が正解と一番近いのが, VGG19_bn であることがわかる. また，精度検証では誤差であった ViT16 は上の線グラフではあるモノの，図 6, 7 で示しているように途中まで不安定である．よって，これらの結果からこちらも VGG19_bn がシステムに組み込む転移学習手法として適切であると判断する．このように実験 1-1 と実験 1-2 の結果において，モデル構築の精度検証，データとモデルの適合率（どれだけデータに対して正解に近づけるか），汎化性能(異なるデータに対してどれだけ正解に近づけるのか)を数値で取得した. この結果を踏まえて手法選択の検討を行った．その結果としてはすべてのモデル構築時における転移学習手法として VGG19_bn がシステム構築時においてモデルとして組み込むのがよいと判断した．本事象ではシステムに VGG19_bn を転移学習したモデルをシステムに組み込み，モデルによる差異を検証する．

4.2. 実験 2

本節では実際に音源をマイクデバイスで取得し，混合行列による正解とモデルが推論した結果を計測した. 両モデルで各項目 20 回ずつ，計 200 回テスト音声・音響を流し，実験を行った．評価指標としては再現率，適合率[18][19]を算出する．算出方法として，単体モデルの場合と提案モデルの場合でそれぞれ混合行列を作成する．それを用いて音響とモデルが予測した結果を記録する．その結果に基づき，再現率と適合率を計算する．混合行列は図 8 のようなものを使用する．近藤行列とはデータの予測結果を陽性・陰性ごとにまとめた行列のことをいう．表 4 のようなモノである．TP: True Positive は正解を正解と予測したことを指す（真陽性）．FP: False Positive は不正解を正解と予測したことを指す（偽陽性）．TN: True Negative は不正解を不正解と予測した（偽陰性）．FN: False Negative は正解のものを不正解と予測した（真陰性）ものである.
また，再現率は以下のように計算を行う.

$$Recall = TP/(TP + FN)$$

適合率は以下のように計算を行う.

$$Precision = TP/(TP + FP)$$

また，このとき，TP：真陽性（True Positive）：正解と予測された中で実際に正だったもの. FN: 偽陰性（False Negative）：正解なのに誤って負と予測されたもの. FP：偽陽性（False Positive）：誤って正と予測されたものとする．再現率は，実際に正だった中でモデルが正と予測できた割合を表す指標である．例えば，迷惑メールの場合で考える．通常事項や重要なメールを誤検知してしまった場合の方が影響は大きい．逆に見逃した迷惑メールは手動で振り分けることで対応ができる. 見逃しの評価は再現率を重視するこれらのことを起こさないよう評価する必要がある．今回扱うものは再現率と適合率である．適合率は，モデルが正と予測した中で実際に正だった割合を表す指標です. 例えば, PCR 検査の場合,「陽性であるのに陰性と診断」と「陰性であるのに陽性と診断」のどちらの状況が良いかをみた時，後者の方が良い方であろう．このように見逃すことの方が影響は大きい．逆に誤検知で陽性とした場合でも陰性であれば，即座の対応が可能となる．誤検知の評価は適合率を重視する．スパムメールもこれを重

視する．また，再現率と適合率は，トレードオフの関係にある．再現率を高めるためには，適合率を下げる必要があり，逆も然りである．そのため，それぞれの指標のどちらを重視するかは，アプリケーションの特性によって判断する必要がある．例えば，病気の検査では，偽陰性を減らすために再現率を重視する一方，スパムメールの検出では，偽陽性を減らすために適合率を重視するといったように，アプリケーションの特性に合わせて，適切な指標を選択する必要がある．なお，モデル手法としては前節の実験結果を踏まえ，汎化性能が高いと判断した VGG19_bn を採用する．
単体モデルの場合，提案モデルの場合における各項目の再現率と適合率については表 5, 6, 7, 8, 9 にまとめる．提案モデルの場合における各項目の再現率と適合率については表 10, 11, 12, 13, 14 にそれぞれまとめる．また，全体の結果として，表 17 に単体モデルの場合と提案モデルの場合をそれぞれまとめる．

表 4:混同行列の例

	実際は正解 (positive)	実際は不正解 (negative)
予測が正解 (positive)	真陽性 TP	偽陽性 FP
予測が不正解(negative)	偽陰性 TN	真陰性 FN

赤羽駅 正解/ 予測	自動放送	駅員アナウンス	警笛	接近チャイム	駅メロ
自動放送	14	2	5	2	3
駅員アナウンス	4	3	5	8	3
警笛	2	1	2	3	5
接近チャイム	1	1	2	3	5
駅メロ	5	2	2	5	10

図 8:混合行列の記録シートの例

表 5:単体モデルの場合における自動放送の混同行列

正解/予測	自動放送	自動放送でない
自動放送	12	13
自動放送でない	8	88

表 6: 単体モデルの場合における接近チャイムの混同行列

正解/予測	接近チャイム	接近チャイムでない
接近チャイム	9	11
接近チャイムでない	16	91

表 7: 単体モデルの場合における駅メロの混同行列

正解/予測	駅メロ	駅メロでない
駅メロ	10	14
駅メロでない	10	90

表 8: 単体モデルの場合における警笛の混同行列

正解/予測	警笛	警笛でない
警笛	14	9
警笛でない	6	86

表 9: 単体モデルの場合における駅員アナウンスの混同行列

正解/予測	駅員アナウンス	駅員アナウンスでない
駅員アナウンス	9	5
駅員アナウンスでない	11	91

表 10: 提案モデルの場合における自動放送の混同行列

正解/予測	自動放送	自動放送でない
自動放送	15	5
自動放送でない	5	85

表 11: 提案モデルの場合における接近チャイムの混同行列

正解/予測	接近チャイム	接近チャイムでない
接近チャイム	10	10
接近チャイムでない	10	90

表 12: 提案モデルの場合における駅メロの混同行列

正解/予測	駅メロ	駅メロでない
駅メロ	12	13
駅メロでない	9	88

表 13: 提案モデルの場合における警笛の混同行列

正解/予測	警笛	警笛でない
警笛	16	9
警笛でない	4	84

表 14: 提案モデルの場合における駅員アナウンスの混同行列

正解/予測	駅員アナウンス	駅員アナウンスでない
駅員アナウンス	15	4
駅員アナウンスでない	5	85

表 15:単体モデルの場合の各項目の再現率と適合率

		単体モデル	
認識対象	種別	再現率	適合率
自動放送	音声	40.00%	48.00%
駅員アナウンス	音声	45.01%	64.43%
接近チャイム	音響	36.00%	45.00%
駅メロ	音響	50.00%	41.67%
警笛	音響	70.00%	60.87%
平均		48.20%	51.99%
音声特徴平均	音声	42.51%	56.22%
音響特徴平均	音響	52.00%	49.18%

表 16:提案モデルの場合の再現率と適合率

		提案モデル	
認識対象	種別	再現率	適合率
自動放送	音声	60.00%	75.00%
駅員アナウンス	音声	75.00%	78.95%
接近チャイム	音響	50.00%	50.00%
駅メロ	音響	57.14%	48.00%
警笛	音響	80.00%	64.00%
平均		64.43%	63.19%
音声特徴平均	音声	67.50%	76.98%
音響特徴平均	音響	62.38%	54.00%

表 17:全体の再現率と適合率

	再現率	適合率
単体モデル	51.43%	50.94%
提案モデル	67.33%	62.39%

表 18:単体モデルと提案モデルの差分比較

		差分	
認識対象	種別	再現率	適合率
自動放送	音声	20.00%	27.00%
駅員アナウンス	音声	29.99%	14.52%
接近チャイム	音響	14.00%	5.00%
駅メロ	音響	7.14%	6.33%
警笛	音響	10.00%	3.13%
平均		16.23%	11.20%
音声特徴平均	音声	25.00%	20.76%
音響特徴平均	音響	10.38%	4.82%

自動放送は行き先等の情報は重要であるが「電車に乗るという」目的を達成する上では，あまり重要でないと判断した．誤検知をなくしたいため適合率重視で見る．単体モデルの場合では再現率と適合率に差異は少ない.一方で,提案モデルの場合は顕著な差がある．この項目において両方とも適合率は高いモノの，提案モデルの方の値が高いことがわかる．よって自動放送の識別では認識モデルを 2 つ作成することが適切であると考える．接近チャイムは電車が接近してくる時になるが，実装されていない駅も多数存在するため今回は判定ミスを許容することにした．そのため，この項目では見逃しの少なさの再現率を重視したい．単体モデルの場合では適合率に値が傾いている．一方，提案モデルの場合では両数値とも単体モデルの場合と比べ高いモノの再現率と適合率の値が同じである．単体モデルの場合では目的を満たすとはいえ，数値が低い．このため，警笛において両モデルとも重視するべき目的を達成していない．

駅メロはドアが閉まる際の指標となるため重要である．この項目は判断ミスがないようにしたい項目の一つであり，誤検知の少なさを重要視する．そのため，適合率を重視したい．単体モデルの場合と提案モデルの場合では両方とも再現率に値が傾いている．よって両モデルとも重視するべき目的を満たさない．提案モデルの場合の値が若干高いためこのモデルの方が適していないといえる．そのため，駅メロにおいて適している

モデルは単体モデルの場合だったときとなる．
警笛では電車が何か危険信号を発している可能性が高いと考えられるため，この項目は高い精度が求められる．誤検知を防ぐためにも適合率重視で見ていく．両モデルとも再現率に傾いており，この目的を満たしていない．誤差が少ないのは単体モデルの方である．
駅員アナウンスは自動放送と違い不規則なことが多いため，情報を知る必要があるため精度が高い必要がある．そのため，誤検知が起きてはならないと考え，適合率を重視したい.両モデルとも適合率に傾いている．そのため，この目的を満たす．提案モデルの場合の方が値は高く，8 割近いことからもこのモデルがこの項目において適していると考える．全体として表 18 の通りすべての項目において再現率と適合率の値は提案モデルの方が高いことがわかる．特に音声モデルの項目において提案モデルの場合における目的を満たしていた．しかし，音響モデルではあまり，有用な結果を得られなかった．いずれの結果においても目的を達成していない．音声モデルでは発言のニュアンスで判別できたが，音響モデルではデータの特徴が似ていたためかうまく判別できなかったと考える．しかし，提案モデルの場合の方が提案モデルより高いことや混同行列においても単体モデルの時より正答率は高い．したがって，提案モデルでは，ある程度の効果は見込めると考える．

5. おわりに

このように 2 つの認識モデルを作成することは 1 つの認識モデルで音声と音響の両方を扱う場合よりある程度，有用な結果を得られた．音源分離並びに 2 つの認識モデルを用いた聴覚障がい者のサポートのためのマルチレイヤワーニング認識システムの実現について可能性を示せたと考える． また，次に挙げるような利点や特徴を持つシステムを実現できると考える．人の声と自動放送の音を分類するモデルと,警笛や駅メロ，接近チャイムを分類するモデルの二つを作成することで，一つの認識モデルの場合よりも高い認識精度を得られる．音源分離機能によって，人の声と音のアラートが同時に流れる状況でも，それぞれの音を分離し，識別することができる．識別結果を画像や文章で視覚的に表示することで，聴覚障がい者の安心感や利便性を向上させることができる．ただし，このシステムには次のような課題や改善点がある．接近チャイムや駅メロの音は鉄道会社によって異なるため，その場合において誤検知が多かった．この問題を解決するためには，鉄道会社ごとに音のデータベースを作成し，モデルに適用する必要があると考える．自動放送と人の声で分類した際に，ニュアンスの違いから判別は可能なものの，誤検知を起こすこともあった．この問題を解決するためには，音響の内容や文脈を考慮したモデルを構築する必要がある．データ数が少ないこともシステムの性能に影響を与える．この問題を解決するためには，データを拡充し，モデルの学習や評価を行う必要がある．

今後の展望としては，現在はシステム全体をローカルの GPU 搭載の PC 上で実行しているが実用性を考えればスマートフォンやタブレットなどの端末上で実行できるようにすることである．このためには，音源分離と認識モデルの軽量化や最適化が必要である.次に，ほかの手法の検討や比較の必要がある．本論文では，音源分離と識別認識モデルの一つの組み合わせを提案しているが，これが最適なものであるとは限らない．ほかの手法との比較を行うことで，システムの性能や優位性をより客観的に評価できると考える．このためには，音源分離や識別のための様々な手法を調査し，同じデータセットや評価指標で実験を行う必要がある.また，データの補充である．本論文では，音響認識モデルでは一つの項目で 20 個しか用意できなかった．これでは，システムの汎化性能や頑健性が低いと考えられる．データの補充を行うことで，システムの性能や汎化性能を向上させることができると考える．このためには，実際の駅構内の環境音を収録し，音源分離や識別のために必要なラベルを付与する必要がある．最後に，このシステムを実際に利用してもらい，フィードバックから利用者に沿ったシステムの実現・拡充が必要である．本論文では，システムの性能や効果を定量的に評価しているが，これだけではシステムの有用性や実用性を保証できない．このシステムを実際に聴覚障がい者や公共交通機関の利用者に利用してもらい，その感想や意見を聞くことで，システムの改善点や拡張点を見つけることができると考える．このためには，システムの利用方法やインターフェースを工夫し，利用者にとって使いやすく分かりやすいものにする必要がある．また，利用者のフィードバックを収集し，分析し，反映する仕組みを構築する必要がある．以上のように，本論文では，駅構内における聴覚障がい者のサポートのためのマルチレイヤワーニング認識システムの実現について述べた．このシステムは，音源分離と二つの認識モデルを用いて，音響案内や警告音などの音情報を視覚的に表示することで，聴覚障がい者の安全性や利便性を向上させることを目指している．しかし，このシステムにはまだ多くの課題や改善点があることも明らかになった．今後は，スマホなどの環境の実現やほかの手法の検討や比較，データの補充,利用者のフィードバックの収集と反映などを行い，より実用的で有用なシステムを開発していきたいと考える．

参 考 文 献

[1] 厚生労働省社会・援護局障害保健福祉部,『平成 28 年生活のしづらさなどに関する調査 (全国在宅障害児・者等実態調査) 結果』, 平成 30 年 4 月 9 日.

[2] 東洋経済新報社,『「聴覚障害者」にとって通勤電車はこんなに怖い 車内放送聞こえず, 何が起きたかわからない』, 柳澤 美樹子, 2019/07/05

[3] 国土交通省総合政策局 安心生活政策課,『旅客施設等における視覚・聴覚障害者等に対応する ICT を活用した情報提供・案内に関する調査 報告書』, 平成 29 年 3 月

[4] 国土交通省『視覚・聴覚障害者の安全性・利便性に関する調査研究報告書』, 2011.

[5] エキマトペ 富士通株式会社 https://ekimatopeia.jp/ (最終参照:令和 6 年 1 月 10 日(水))

[6] Ontenna 富士通株式会社 https://ontenna.jp/ (最終参照:令和 6 年 1 月 9 日(火))

[7] Mozilla Common Voice https://commonvoice.mozilla.org/ja(最終参照:令和 6 年 1 月 10 日(水))

[8] glasses/vgg19_bn Hugging Face https://huggingface.co/glasses/vgg19_bn (最終参照:令和 6 年 1 月 10 日(水))

[9] PyTorchdocument VGG19_BN https://pytorch.org/vision/main/models/generated/torchvision.models.vgg19_bn.html(最終参照:令和 6 年 1 月 10 日(水))

[10] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[11] GitHub - facebookresearch/demucs: Code for the paper Hybrid Spectrogram and Waveform Source Separation(最終参照:令和 6 年 1 月 9 日(火))

[12] Défossez, Alexandre. "Hybrid spectrogram and waveform source separation." arXiv preprint arXiv:2111.03600 (2021). (最終参照:令和 6 年 1 月 9 日(火))

[13] Music Source Separation with Hybrid Demucs — Torchaudio 2.2.0.dev20240109 documentation (pytorch.org) https://github.com/facebookresearch/demucs(最終参照:令和 6 年 1 月 9 日(火))

[14] PyTorch 学習メモ 2(事前学習済みモデルを使ってみた) https://qiita.com/makaishi2/items/eb87587c9b7749d3f170 2021 年 10 月 10 日 (最終参照:令和 6 年 1 月 10 日(水))

[15] Kaur, Prabhjot, Qifan Wang, and Weisong Shi. "Fall detection from audios with audio transformers." Smart Health 26 (2022): 100340. (最終参照:令和 6 年 1 月 10 日(水))

[16] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020) (最終参照:令和 6 年 1 月 10 日(水)).

[17] MLflow: A Tool for Managing the Machine Learning Lifecycle — MLflow 2.9.2 documentation https://mlflow.org/docs/latest/index.html (最終参照:令和 6 年 1 月 10 日(水))

[18] 多クラス分類における混同行列と評価指標を丁寧に - gotutiyan's blog https://gotutiyan.hatenablog.com/entry/2020/09/22/175141 (最終参照:令和 6 年 1 月 10 日(水))

[19] 講義資料 | 東京工業大学 中田研究室 http://www.iee.e.titech.ac.jp/~nakatalab/text/lecture.html (最終参照:令和 6 年 1 月 10 日(水))

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

2024年2月28日(水) 10:00 ~ 12:10 | T5-A オンライン（Zoom Events)

教育・学習 (1)

座長:張 建偉(岩手大学)
コメンテータ:林 康弘(武蔵野大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T5-A-1-01] Web閲覧および発言履歴を用いた大規模言語モデルによる学習アドバイスの生成手法

*嶋津 瑛[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

10:20 ~ 10:35
[T5-A-1-02] 大規模言語モデルを活用したノーコードツールによるAI人材育成での学習体験の向上

*丹野 良介[1]、坂本 悠記[1]、中村 耀[1]、杉本 海人[1]、切通 恵介[1] (1. NTTコミュニケーションズ株式会社)

10:35 ~ 11:00
[T5-A-1-03] アニメーションおよびソースコード表示を活用したソートアルゴリズム学習アプリケーションの開発

*渡辺 大智[1]、石井 幹大[2]、伊藤 一成[3] (1. 青山学院大学 社会情報学部 伊藤一成研究室、2. 青山学院大学 理工学部、3. 青山学院大学 社会情報学部)

11:00 ~ 11:25
[T5-A-1-04] ランダムサーファーモデルの学習アプリケーション「人型ピクサーフィング」の実装と評価

*高橋 伶奈[1]、伊藤 一成[1] (1. 青山学院大学社会情報学部社会情報学科伊藤一成研究室)

11:25 ~ 11:50
[T5-A-1-05] 講義資料作成支援のための科目用語体系の階層構造と教科書の説明順序を用いたスライド分析手法

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

Web閲覧および発言履歴を用いた 大規模言語モデルによる学習アドバイスの生成手法

嶋津　瑛†　堀川　達平††　北山　大輔†

† 工学院大学情報学部　〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学大学院工学研究科　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j323402,em23040}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　2021 年の中央教育審議会による答申「『令和の日本型学校教育』の構築を目指して」において，今後の教育では，個々人の特性や興味領域に合わせた，個別最適な学びが重要であると示されるなど，学習の個人化の重要さが増している．そこで我々は，個人が取り組む課題の場や予習の場において，学習の進度に合わせたアドバイスを自動的に生成することが，このような学習の個人化に有効であると考えた．本稿では，課題や予習時に調べているテキストとして Web ページを想定し，Web 閲覧履歴と発言履歴を用いて，大規模言語モデルを用いた学習アドバイスの生成手法を提案し，プロトタイプシステムの構築を行ってその動作を検証した．検証の結果，大規模言語モデルを用いることで，Web 閲覧履歴から人手作成よりも優れた学習アドバイスを作成することができることを確認した．

キーワード　教育，e-Learning，学習支援，行動データ，コメント生成

1 はじめに

学習は，学生に限らず社会人となっても必要となる行為である．特に社会人となってからは，誰かに師事することなく個人での学習となることが一般的と言える．一方，学生においても，2021 年の中央教育審議会による答申「『令和の日本型学校教育』の構築を目指して」において，今後の教育では，個々人の特性や興味領域に合わせた，個別最適な学びが重要であると示されるなど，学習の個人化の重要さが増している．

個人の学習支援に関する従来研究を紹介する．Lee ら [1] は，個人の学習のビデオ配信を行う効果についての分析を行っている．この研究において，学習配信動画を見る動機として,「他者から刺激を受ける」「精神的な支えを得る」「注意散漫の軽減」などがあることが報告された．我々は，このような動機は配信動画以外でも満たすことができると考え，個人学習環境として提供するべき要素であると考える．例えば，アドバイスをしてくれる第三者が見守る環境は，これらの動機を満たすものと考えられる．高野ら [2] のような，大規模言語モデルを個人の学習支援に用いる取り組みも始まっており，今後ますます個人学習に大規模言語モデルが使われていくと考えられる．

そこで我々は，個人が取り組む課題の場や予習の場において，学習の進度に合わせたアドバイスを自動的に生成することが，このような学習の個人化に有効であると考えた．本稿では，課題や予習時に調べているテキストとして Web ページを想定し，Web 閲覧履歴と発言履歴を用いて，大規模言語モデルを用いた学習アドバイスの生成手法を提案する．

本研究の貢献は以下のとおりである．

- 学習の進度に応じたアドバイスを大規模言語モデルを用いて自動的に生成するモデルを提案した
- 実例に基づく提案モデルの効果と限界について議論した

2 提案手法

2.1 システム構成

本手法では，学習等のタスク実行中に閲覧した Web ページの履歴と，発話内容を入力として，逐次アドバイス文を生成する．システムの概要を図 1 に示す．なお，このシステムでは「中学 3 年生の数学の授業」に関する学習をしていることを想定し，大規模言語モデルへ入力するプロンプトもその想定に基づき構築している．システム利用時に今回の学習の単元をユーザが入力する．すなわち，このシステムは，GPT に入力するプロンプトの生成を Web 閲覧履歴および発話内容から自動的に行う処理モデルを中核としている．

Web 閲覧履歴は，システム利用時以前の閲覧履歴に関しては，閲覧履歴を取得する Chrome 拡張機能を作成し，そのページタイトルを取得する．システム利用時以後の閲覧履歴に関しては，現在の閲覧ページのタイトルを逐次保存することで対応する．Chrome 拡張を用いることができない場合は，システム利用以後の閲覧履歴のみを用いる．Web 閲覧内容に基づくアドバイス生成については，一定時間ごとに生成することとし，詳細は 2.2 節で説明する．

発話内容に関しては，ブラウザに搭載されている Web Speech API を用いて音声認識を行う．発話内容に基づくアドバイス生成については，ユーザが任意のタイミングで行うこととした．詳細は 2.3 節で説明する．

アドバイス生成には，OpenAI が提供している大規模言語モデル（Large Language Models）である GPT[1] を用いる．Model は gpt-3.5-turbo-1106 を使用し，回答のランダム性を指定するパラメータである temperature は 0 とした．GPT への入力は，

1：https://platform.openai.com/docs/api-reference/introduction

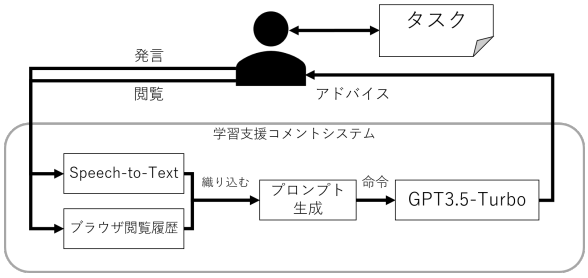


図 1　Web 閲覧・発言履歴を用いた学習アドバイス生成の概念図

表 1　閲覧履歴に基づくアドバイス生成のプロンプト

あなたは〈単元〉の先生です．以下では，〈単元〉の先生になりきって回答を生成してください．〈単元〉を勉強しているユーザーが，Web ブラウザで次のようなタイトルのページを閲覧しました．
〈ページタイトルを列挙〉
これらの閲覧履歴から，ユーザーがどのような数学の問題について悩んでいるか推測し，解き方などの具体的なアドバイスを出力してください．生成するアドバイスは，箇条書きでなく文章の形式で行ってください．また，直接的に問題の答えを教えてはいけません．なお，閲覧履歴が，中学 3 年の数学の内容と過度に逸脱している場合は，勉強に集中できない状態であると判断して，ユーザーに勉強への集中を促すようなアドバイスのみを出力してください．

ユーザの発話履歴とシステムによるアドバイスの履歴が会話履歴として入力され，アドバイス生成のための命令文は会話履歴としては入力しない．

2.2　閲覧履歴に基づくアドバイス文生成

閲覧履歴に基づくアドバイス生成では，ページのタイトルを閲覧履歴として，プロンプトに組み込んでアドバイスを生成させる．GPT へ入力するプロンプトは表 1 のように構成する．⟨...⟩ 内は変数であり，学習の単元や閲覧履歴のタイトルが列挙される．

閲覧履歴は最長 10 件用いる．予備実験では 5 件から 10 件程度で適切なアドバイスが生成されることを確認したため 10 件と設定したが，多ければ精度が良くなるというわけでもないため，適切な件数を決定する方式が必要であると考えている． なお，閲覧履歴は，ページタイトルが GPT の動作に悪影響を与えないよう，鉤括弧で囲んで明示的にページタイトルであることを示した．

プロンプトの冒頭では，学習対象やユーザの属性などを指定し，その対象や属性に対応する指導者をロールプレイするように指示を行うことで，適切な難易度のアドバイスが生成されるように設計した．

プロンプトの後半部分は，生成されるアドバイスの形式を指定し，答えを直接教えるようなアドバイスを生成させないために設計した．加えて，学習中に無関係なキーワードを検索・閲覧している様子を検知したら，軌道修正を促すアドバイスを生成する指示も行っている．学習と無関係な行動をしていることの検知は，現在はアドバイス文生成と同時に行っているが，独立したアルゴリズムを構築することも検討している．

表 2　発話に基づくアドバイス生成のプロンプト

あなたは〈単元〉の先生です．以下では，〈単元〉の先生になりきって回答を生成してください，〈単元〉を勉強しているユーザーが，次のような発言をしました．
〈問い合わせ内容〉
この発言からアドバイスを出力してください．生成するアドバイスは，箇条書きでなく文章の形式で，また一般的な中学 3 年生が理解できる内容で行ってください．また，直接的に問題の答えを教えてはいけません，なお，発言が，中学 3 年の数学の内容から過度に逸脱していたり，勉強と関係のない雑談であった場合は，勉強に集中できない状態であると判断して，ユーザーに勉強への集中を促すようなアドバイスのみを出力してください．

表 3　式展開に関する閲覧ページのタイトル

多項式の計算　目次
乗法公式による計算
3 項式の計算
式展開項 3 つ
【 3 分で分かる!】分配法則の仕組みとやり方をわかりやすく
分配法則
式の展開とは?公式・基礎問題・応用問題の解き方を分かりやすく解説
「式の展開」とは?基本的な分配法則と中学 3 年生で習う 3 つの公式をご紹介
式展開　やり方

2.3　発話に基づくアドバイス生成

発話に基づくアドバイス文生成は，ユーザの任意のタイミングで行う．音声認識が検知したタイミングで自動的に行うことも可能であるが，数式など認識ミスが起こりやすいドメインや，閲覧ページの引用を利用する場合などを考え，音声認識結果はテキストボックスに反映するのみとし，問い合わせボタンを押したときに，テキストボックスの内容をもとにアドバイス文を生成する．GPT へ入力するプロンプトは表 2 のように構成する．⟨...⟩ 内は変数であり，学習の単元や問い合わせ内容が入力される．

プロンプトの後半部分は，生成されるアドバイスの形式を指定し，答えを直接教えるようなアドバイスを生成させないために設計した．加えて，学習中に無関係な雑談や無意味な問い合わせを検知したら，軌道修正を促すアドバイスを生成する指示も行っている．

3　ユースケース

ユースケースとして，中学 3 年生のユーザーが数学について学習している場面を想定して Web 閲覧を行い，閲覧履歴に基づいてアドバイス文を生成した．

まず，$3x(2x+y)$ や $(a+b+2)^2$ といった式展開の問題について勉強している場面を想定し，Web ページの検索・閲覧を行った．閲覧したページのタイトルは，表 3 の通りである．

これらの閲覧履歴を用いて，単元を “中学 3 年生の数学” とし，表 1 のプロンプトに組み込んでアドバイスを生成させた．生成されたアドバイスを表 4 に示す．

表 4　式展開についての学習を想定して生成されたアドバイス文

推測される悩み:
ユーザーは多項式の計算について興味があり，特に乗法公式や 3 項式の計算，式の展開に関心があるようです．また，分配法則についての理解を深めたいと考えているようです．
具体的なアドバイス:
1. 情報公式による計算や 3 項式の計算を理解するためには，まず基本的な公式を覚えることが大切です．例えば，二項の積や三項の積などの公式を覚えると，計算がスムーズに行なえます．
2. 式の展開については，分配法則を使って括弧の中の式を展開する必要があります．分配法則を理解し，積を用いて式を展開する練習をすると良いでしょう．
3. 資料を読むだけでなく，実際に問題を解いてみることで，計算の手順や公式の使い方を自分のものにすることが大切です．
4. 数学の勉強は復習が重要です．学んだ公式や計算方法などを定期的に振り返り，定着させるよう心がけましょう．

表 5　2 次方程式に関する閲覧ページのタイトル

M242　ブッシュマスター
GAU-12　イコライザー
ハリアー II (航空機)
アフターバーナー
GAU-8 (機関砲)
【解の公式】覚え方と使い方をマスターしよう
2 次方程式の解の公式
因数分解
因数分解を使う解き方 - 2 次方程式: 中学 3 年生
2 次方程式　因数分解

表 4 を見ると，ユーザーが多項式の計算について関心があると判断しているが，これは想定した問題と同様のものであるため，どのような問題について悩んでいるかを正しく認識できていると考えられる．また，アドバイスについても具体性があり，特筆すべき異常な点が見つからないことから，学習に向けたアドバイスとしては適切であると考えられる．

続けて，$(x-5)(x+2)=0$ や $x^2-x-3=0$ といった 2 次方程式の問題について勉強しているが，勉強途中で関連性の薄いページを閲覧してしまい，本来の目的から逸脱してしまった場合を想定して，同様に Web ページの検索・閲覧を行った．閲覧したページのタイトルは，表 5 の通りである．

この閲覧履歴では，ユーザーが何らかの理由により，取り組んでいる問題と関連性の薄いページを閲覧してしまい，そのまま自発的な集中力の再発揮や修正が行えずに，問題から逸脱したページを閲覧し続けてしまった場合を想定している．この閲覧履歴を用いて，同様に単元を “中学 3 年生の数学” とし，表 1 のプロンプトに組み込んでアドバイスを生成させた．生成されたアドバイスを表 6 に示す．

これらの生成アドバイスは，概ね指示を設定した際の意図に沿ったものであり，このようなユースケースを想定した場合には，意図したとおりのアドバイスを生成できることを確認した．

表 6　軌道修正を促すアドバイス文

ユーザーの閲覧履歴から，数学の問題に取り組む際に集中できていない可能性があります．数学の学習をする際には，集中して取り組むことが大切です．以下のアドバイスを参考にして，問題に取り組む際に心を落ちけて集中してみてください．
1. 数学の問題に取り組む前に，周囲の気を散らす要因を取り除くことが重要です．静かな環境で取り組むか，集中できる場所を選んでから問題に取り組みましょう．
2. 数学の問題は焦らず，一歩一歩丁寧に取り組みましょう．急いで解こうとすると，見落としが生じたり，理解が不十分なまま進んでしまうことがあります．
3. 数学の問題に取り組む際には，頭の中で整理することが大切です．与えられた情報や式を整理し，問題のポイントを把握しましょう．
4. 数学の学習をする際には，集中力を高めるための呼吸法やリラックス法を活用することも有効です．深呼吸をするなど，緊張を和らげる方法を試してみましょう．
数学の問題に取り組む際には，集中できる状態で臨むことが大切です．一度，周囲を整えてから問題に取り組んでみてください．それでも解決できない場合は，具体的な問題について知識を深めるため，学校の先生や友達と協力して取り組むことをお勧めします．

4 実　　験

提案手法のうち，Web 閲覧履歴に基づいた GPT-3.5 によるアドバイス文の生成が可能であるかどうか，また生成されたアドバイスの内容が適切であるかどうか調べるため，実験を行った．

4.1 実験方法

システムが実際に使用される状況を想定して，Web 閲覧を行いながらのタスク遂行，閲覧履歴を用いた人手によるアドバイス作成，人手作成アドバイスと GPT 生成アドバイスの比較による評価という 3 段階の実験を行った．各実験はクラウドソーシングによって被験者を募り実施した．

4.1.1　Web 閲覧を行いながらのタスク遂行段階

本段階では，アドバイス生成用の閲覧履歴を収集するために，適宜 Web 閲覧を行いながら，中学 3 年相当の数学の問題を 5 問解かせた．ここで出題した問題は，教育出版 “令和 3 年度版 中学校 まなびリンク”[2] より一部抜粋したものである．

被験者には，自由に Web 閲覧を行ってよいことを知らせた上で各問題を解かせ，回答と合わせて閲覧したページを古い順に最大 10 件まで提出するよう依頼した．検索行動にある程度長い時間をかけることで，同一の問題や正答集などを発見してしまう可能性が考えられることから，各問題は 5 分程度で解き，時間内に解けなかった場合は「時間切れ」と回答するように指示した．

4.1.2　閲覧履歴を用いたアドバイス作成段階

本段階では，人手によって作成されたアドバイスを収集する

2：https://www.kyoiku-shuppan.co.jp/ml-jh/sugaku/practice3.html

ために，前段階で収集した閲覧履歴を提示し，元の問題の推測と，推測結果に応じて問題を解くためのアドバイスを作成させた．

被験者には，閲覧履歴に基づくアドバイス文生成時に用いるプロンプトと同様の注意点を提示し，これらを守ったうえでアドバイスを作成するように指示した．具体的には，次の点に注意するように指示した．

（1） どのような問題を解いているかについての予想を含めること

（2） 可能な限り文章の形式にすること (箇条書きは不可)

（3） 一般的な中学 3 年生が理解できる難易度にすること

（4） 問題の答えを直接的に教えることはせず，解法を提示するまでにする

（5） Web 上のページまたは書籍等を参考にしたものでも可とする

（6） 100〜200 字程度の分量とする

上記の注意点のうち，(5) は GPT との知識量における差を調整するために，(6) は過剰に短いまたは長いアドバイスを作成させないために，新たに設定したものである．また，アドバイス作成者の知識量や理解度によるアドバイス品質のばらつきを減らすために，1 つの Web 閲覧履歴に対し，人手によって 3 件，GPT によって 1 件のアドバイスを作成させた．

同時に，表 1 に示したプロンプトを用いて，単元を “中学 3 年生の数学” としたうえで被験者に渡したものと同一の閲覧履歴を挿入し，GPT によってアドバイスを生成した．

4.1.3 アドバイスの評価段階

本段階では，人手作成されたアドバイスと GPT が生成したアドバイスを比較するために，複数観点でアドバイスを評価させた．

被験者には，人手または GPT で作成・生成したアドバイスと，アドバイスに対応する問題を提示し，次の 3 つの観点で評価させた．

（1） アドバイスの対象となる問題を正しく推測出来ているか

（2） 中学 3 年生が閲覧したときに理解できる内容か

（3） アドバイスの内容に誤りが含まれているか

なお，観点 (1) および (2) は「推測できている」または「理解できる内容である」から「推測できていない」または「理解の難しい内容である」までの 4 段階評価で，観点 (3) は「含まれている」または「含まれていない」の二者択一で回答させた．

4.2 実験結果

4.2.1 Web 閲覧を行いながらのタスク遂行段階

実験の結果，5 問全問に正答した被験者 7 人，誤答を含む回答をした被験者 6 人の計 13 人分の回答と Web 閲覧履歴を収集することができた．表 7 に全問正答した被験者の Web 閲覧履歴の例を，表 8 に誤答を含む回答を行った被験者の Web 閲覧履歴の例を示す．

表 7 および表 8 を比較すると，被験者によって閲覧したページの数が異なり，また同じ問題に対して異なるトピックで閲覧行動を行っていることが分かる．

表 7　全問正答した被験者の閲覧履歴例

二次方程式の意味と解き方まとめ【解の公式・因数分解】
【中 3 数学】「2 次方程式の解き方のまとめ」(例題編)
【中 3 数学】二次方程式の練習問題にチャレンジ！(解説あり)

表 8　誤答を含む回答を行った被験者の閲覧履歴例

2 次方程式の解と係数の関係

表 9　表 7 の Web 閲覧履歴に対する人手作成アドバイス

閲覧している内容は 2 次方程式の解き方に関するものですね．2 次方程式の解き方には，平方根を利用するもの，因数分解を利用するもの，解の公式を利用するものがあります．特に解の公式は理論上どのような 2 次方程式でも必ず答えがわかるので必ず覚えておきましょう．

表 10　表 7 の Web 閲覧履歴に対する GPT 生成アドバイス

1. 二次方程式の解き方を理解するためには，まず二次方程式の意味をしっかりと理解することが大切です．二次方程式は，グラフ上で放物線を表す式であり，その解は放物線と x 軸との交点に相当します．そのため，解の意味をイメージしながら学習を進めると理解が深まります．
2. 解の公式や因数分解など，複数の解き方があることを理解しましょう．それぞれの解き方の特徴や使い分ける場面を理解することで，問題に応じて適切な解き方を選ぶことができます．
3. 例題編や練習問題に取り組む際には，問題文をよく読み，与えられた条件や情報を整理することが重要です．問題を解く前に，何が求められているのか，どのような情報が与えられているのかを確認しましょう．
4. 解説を読む際には，一歩一歩丁寧に解説を追いながら，自分なりに考えてみましょう．解き方や考え方が分からないときは，先生や友達に質問することも大切です．
5. 勉強をするときは，集中して取り組むことが大切です．他のことに気を取られないようにして，静かな場所で集中して取り組んでみましょう．

全問正答した数学が比較的得意な可能性のある被験者と，誤答を含む回答を行った数学が得意ではない可能性のある被験者では，検索行動が異なる可能性も考えられる．また，Web 閲覧履歴として問題の解答集を参照している被験者や，電卓やソルバーなどを利用した被験者も存在した．そのため，以後の段階では，これらの被験者を除外して，正答のみ 6 人・誤答あり 4 人の計 10 人分の閲覧履歴を用いて実験を実施することとした．

4.2.2 閲覧履歴を用いたアドバイス作成段階

前段実験で収集した 5 問 10 人分，計 50 セットの Web 閲覧履歴から，人手によるアドバイス作成と GPT によるアドバイス生成を行った．作成されたアドバイスの例を示す．

まず，表 7 の Web 閲覧履歴に対し人手で作成されたアドバイスの例を表 9 に，GPT によって生成されたアドバイスの例を表 10 に示す．

表 9 および表 10 を確認すると，どちらも 2 次方程式に関するものであると正しく問題を推測できていることがわかる．また，GPT によって生成されたアドバイスは，分量がかなり多いこともわかる．

表 11　表 8 の Web 閲覧履歴に対する人手作成アドバイス

2 次方程式の解と係数の関連性としては，まず解の公式を用いながら和と積の基本公式を導き出します．ここで大事になってくるのが x の 2 乗の係数を 1 にすることであり，右辺の 0 を利用しながらなるべくシンプルな式を作る方法が大事になってきます．

表 12　表 8 の Web 閲覧履歴に対する GPT 生成アドバイス

2 次方程式の解と係数の関係について学習されているようですね．このテーマは重要なものですが，理解するのに少し時間がかかるかもしれません．まず，2 次方程式の一般形をしっかり覚えましょう．それから，解の公式を使って具体的な問題を解いてみると良いでしょう．係数と解の関係について考えるときは，解の公式の中に係数がどのように関わっているかを考えてみてください．また，グラフを使って 2 次方程式の解と係数の関係を視覚的に理解することもできます．数学は少しずつ積み重ねていくものなので，焦らずに着実に理解を深めていきましょう．勉強を続けることで，必ず理解できるようになりますよ．頑張ってください！

表 13　人手作成アドバイスに対する問題推測評価

	正しい	やや正しい	やや正しくない	正しくない
問題 1	19 %	38%	20%	23%
問題 2	29 %	42%	18%	11%
問題 3	36 %	36%	19%	10%
問題 4	34 %	32%	18%	16%
問題 5	23 %	31%	26%	20%
平均	28%	36%	20%	16%

続けて，表 8 の Web 閲覧履歴に対し人手で作成されたアドバイスの例を表 11 に，GPT によって生成されたアドバイスの例を表 12 に示す．

表 11，表 12 を比較すると，どちらも 2 次方程式の解と係数の関係に関するものであると問題を正しく推測できている一方で，表 9 および表 10 と同様に GPT によって生成されたアドバイスの方が分量の多い結果となった．

4.2.3　アドバイスの評価段階

前段実験で作成・生成したアドバイスに対し，複数観点での評価を行った．各アドバイスにつき 3 件ずつ評価を行い，人手で作成した 150 件と GPT で生成した 50 件のアドバイスに対して計 600 回答の評価を収集した．人手作成アドバイスと GPT 生成アドバイスの数が異なることから，本節では評価結果を全回答中の割合で示すこととする．

まず，表 13 に，人手によって作成されたアドバイスに対して，問題推測が正しく行えているかについての評価結果を示す．続けて，表 14 に，GPT によって生成されたアドバイスに対して，同じく問題推測が正しく行えているか評価した結果を示す．

表 13 および表 14 からは，GPT によって生成されたアドバイスは，人手作成されたアドバイスよりも問題推測が可能であることを示している．また，この結果は問題によらず概ね一定であり，GPT の方が問題推測の面で優れていることを示している．この結果は，入力したプロンプトが，適切なトピック推定の材料を提供できていることを示唆している．

表 14　GPT 生成アドバイスに対する問題推測評価

	正しい	やや正しい	やや正しくない	正しくない
問題 1	37%	40%	10%	13%
問題 2	40%	40%	7%	13%
問題 3	47%	33%	10%	10%
問題 4	53%	30%	17%	0%
問題 5	30%	40%	13%	17%
平均	41%	37%	11%	11%

表 15　人手作成アドバイスに対する難易度適正さ評価

	適切	やや適切	やや不適切	不適切
問題 1	26%	34%	28%	12%
問題 2	26%	40%	27%	8%
問題 3	29%	31%	31%	9%
問題 4	37%	30%	19%	14%
問題 5	28%	24%	31%	17%
平均	29%	32%	27%	12%

表 16　GPT 生成アドバイスに対する難易度適正さ評価

	適切	やや適切	やや不適切	不適切
問題 1	30%	33%	27%	10%
問題 2	27%	37%	30%	7%
問題 3	47%	30%	13%	10%
問題 4	33%	47%	20%	0%
問題 5	27%	47%	10%	17%
平均	33%	39%	20%	9%

表 17　人手作成アドバイスに対する誤り含有評価

	誤りあり	誤りなし
問題 1	27%	73%
問題 2	30%	70%
問題 3	27%	73%
問題 4	31%	69%
問題 5	19%	81%
平均	27%	73%

続けて，表 15 に，表 16 に，アドバイスの難易度の適正さについての評価結果を示す．表 15 は人手によって作成したアドバイスに対する評価結果，表 16 は GPT によって生成したアドバイスに対する評価結果である．

表 15 および表 16 は，GPT によって生成されたアドバイスが，人手作成されたアドバイスよりも適正な難易度であることを示している．この結果も問題によらず概ね一定であり，一般的に GPT の方がアドバイス対象に適したアドバイスを生成できることを示している．この結果は，システムが生成したプロンプトに含まれている，適切な難易度のアドバイスを出力する指示が，GPT に正しく伝わっていることを示唆している．

最後に，作成したアドバイスが誤りを含んでいるかどうかについての評価結果を表 17，表 18 に示す．表 17 は人手作成したアドバイスに，表 18 は GPT によって生成されたアドバイスに対する評価の結果である．

表 17，表 18 を比較すると，GPT の生成したアドバイスは全体的に誤りが少なく，人手作成のものに比べ平均して 10%程度誤りが少ないと評価されていることがわかる．この結果は，

表 18　GPT 生成アドバイスに対する誤り含有評価

	誤りあり	誤りなし
問題 1	20%	80%
問題 2	10%	90%
問題 3	13%	87%
問題 4	13%	87%
問題 5	13%	87%
平均	14%	86%

GPT が人間よりも正確なアドバイスを作成できることを示唆している．

これらの結果から，GPT を用いて Web 閲覧履歴からアドバイスを作成することは可能であると言える．また，GPT が生成するアドバイスは，人手作成されたアドバイスに比べて問題をより正しく推測でき，難易度が適切で誤りも少ないと言える．

5 考　　察

表 6 を見ると，ユーザーの閲覧履歴が取り組むべき課題から逸脱していることを正しく認識し，集中を促すアドバイスを生成している事がわかる．アドバイスの内容は，勉強への集中を促すものとしては正常であり，また具体的であることから，適切なアドバイスを生成できていると考えられる．

表 4，表 6 に示したように，単元を設定し，閲覧履歴を与えてプロンプトを作成することで，適切なアドバイスを生成できることを確認した．アドバイスの内容については，ある程度具体的であり，不適切な箇所も見受けられないことから，アドバイスとして十分に機能する品質のものであると考えられる．また，タスクからの逸脱を認識し，軌道修正を促すアドバイスを生成できることについても確認した．このことは，構成したシステムがユーザーの行動を正しく分析できていることを示していると考えられる．

実験結果は，表 13 と表 14 などに示す通り，大規模言語モデルを利用することで Web 閲覧履歴から人手作成のアドバイスよりも優れたアドバイスを生成できることを示している．本稿で行った実験では，被験者をクラウドソーシングによって募集したため，被験者の数学に対する知識や技術は，概ね一般的なレベルだと考えられる．したがって，大規模言語モデルが生成するアドバイスは，特別な教育を受けていない一般的な人物によって作られるアドバイスよりも，高品質のものであると考えられる．

しかしながら，学習行動に対して生成されたアドバイスは，具体的で今後の学習には役立つと考えられるものの，基本的には検索・閲覧行動で疑問を解消できたことが前提となっているアドバイスであり，疑問が解消できなかった場合にはほとんど役に立たないと考えられる．キャッチアップとしてのアドバイスは必要性が高いと考えられることから，アドバイスの方向性をユーザーが指定できるようにするか，プロンプトの改良を検討する必要があると考えられる．

さらに，生成されたアドバイスが問題の推測や適切な難易度を達成していたとしても，実際に有用なアドバイスであるとは限らないため，実際にタスクを実行しながらアドバイスを提示することで，アドバイス自体がタスク実行に寄与するものか検証する必要があると考えられる．

加えて，プロンプトには，適切な難易度のアドバイスを生成するためにロールプレイを行うような指示を付加しているが，この指示の有無が生成されるアドバイスにどれだけの影響を与えるかについても検証する必要があると考えられる．

6 ま と め

本稿では，学習の個人化のために，学習の進度に合わせたアドバイスを自動的に生成する手法を提案した．具体的には，Web 閲覧履歴と発話内容に基づいたアドバイス文を生成することで，現在，学習者が必要とする助言を得るシステムを開発した．Web 閲覧履歴，発話内容に基づくアドバイス文を大規模言語モデルで生成するためのプロンプトを設計し，学習内容から逸れた場合に軌道修正できる機能を組み込んだ．

また，実際にアドバイスが必要とされる状況を想定して Web ページの検索・閲覧を行った後に，閲覧履歴から適切なアドバイスが生成できることを確認した．加えて，閲覧履歴からユーザーの興味が学習から逸脱していないか認識し，適切に集中を促すアドバイスを生成できることを確認した．

さらに，実際に数学の問題を課題として与え，そこから得られた Web 閲覧履歴を用いて大規模言語モデルによりアドバイスを生成し，同じく Web 閲覧履歴から人手作成したアドバイスと比較した．比較の結果，提案手法によって Web 閲覧履歴から大規模言語モデルによって生成されたアドバイスが，人手作成のアドバイスより優れたものであることを確認した．

今後の課題としては，生成されるアドバイスがタスク実行に対して有用なものか検証する必要があると考える．将来的には，アドバイスの持つ学習効果についても評価する必要があると考えている．また，生成されたアドバイスがプロンプトに加えたロールプレイなどの指示が，アドバイス生成にどのような影響を与えているか調査する必要があると考える．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Yoonjoo Lee, John Joon Young Chung, Jean Y. Song, Jean Y. Song, Minsuk Chang, and Juho Kim. Personalizing ambience and illusionary presence: How people use “study with me” videos to create effective studying environments. *International Conference on Human Factors in Computing Systems*, 2021.

[2] 伊藤 恵高野 希澄. プロンプト作成の負荷を軽減する生成系 ai を用いた学習支援システムの提案. 日本ソフトウェア科学会第 40 回大会 (2023 年度) 講演論文集, pp. 45–R–S, 2023.

大規模言語モデルを活用したノーコードツールによるAI人材育成での学習体験の向上

丹野　良介† 　坂本　悠記† 　中村　耀† 　杉本　海人† 　切通　恵介†

† NTT コミュニケーションズ株式会社　〒 108–8118 東京都港区芝浦 3-4-1 グランパークタワー 20F
E-mail: †{r.tanno,yuki.sakamoto,yyo.nakamura,kaito.sugimoto,k.kiritoshi}@ntt.com

あらまし　AI 技術の進化は教育分野において革新的な変化をもたらしており、ChatGPT のような大規模言語モデル (LLM) を活用した生成 AI ツールが即座にフィードバックを行うなど、教育方法論と学習体験の両面で重要な役割を果たす可能性を示唆している。LLM はパーソナライズされた学習体験とインタラクティブな学習アプローチを提供し、学習者の問題解決能力や批判的思考を促進する可能性があるが、これらのツールを特定の教育分野に適用する際の具体的な効果はまだ十分に理解されていない。そこで本稿では、LLM をデータ分析教育・学習支援ツールとして活用し、ノーコードデータ分析ツールと LLM との統合を提案する。また、LLM が教育用途として利用するに耐えうるか出力品質を評価する。

キーワード　大規模言語モデル、データ分析、教育、チャットボット、ノーコード

1　はじめに

AI 技術の進化が教育分野に革新的な変化をもたらしており、ChatGPT のような生成 AI ツールは人間の様な反応を理解し即座にフィードバックを行うなど、教育方法論と学習体験の両面で重要な役割を果たす可能性を示唆している [1] [2]。

LLM を用いた学習支援は、教育と学習の改善を目的として広く行われてきた [3] [4]。この研究の背景には、LLM がもたらすインタラクティブな学習環境の重要性がある。LLM は、リアルタイムの会話型フィードバックを通じて、個別化された学習経験を提供し、学習者の問題解決能力や批判的思考を促進する。これらのツールを特定の教育分野、特にデータ分析教育に適用した際に誤った情報、いわゆるハルシネーションなど潜在的な課題に関する議論がされ始めている [5]。

一方で、実際にデータ分析教育への現場適応やその評価、既存ツールと LLM をどの様に組み合わせるかなど、まだ十分に議論されていない。

本稿では、LLM が提供するパーソナライズされた学習体験やインタラクティブな学習アプローチを活用し、データ分析教育における学習体験向上の有効性に着目する。データ分析の初学者を対象とし、ノーコードデータ分析ツールと LLM を組み合わせたシステムを構築し、LLM の出力が教育用途として利用するに耐えうるか出力品質を評価する。

特に、RAG による知識注入が LLM の出力品質に与える影響を検証する。LLM が特定のクエリにおいて人間の評価基準に近い回答を生成できること、そして RAG がその効果を高めることを評価する。

2　システムデザイン

2.1　分析ツールと LLM との統合アーキテクチャ

LLM を分析ツールに統合することで、パーソナライズされた学習体験やインタラクティブな学習アプローチを実現したシステム構成のアーキテクチャを図 1 として示す。LLM からの応答には次の 2 つのスキームが存在する。1 つ目は、学習者は ChatBot を介して自身の疑問の解消に利用する方法。2 つ目はユーザの行動に応じたシステム応答である。ツール利用状況に応じてフィードバックするコンテンツをリアルタイムに調整する事で、学習者の学習意欲を高めたり学習効果の最大化に寄与するシステム構成としている。

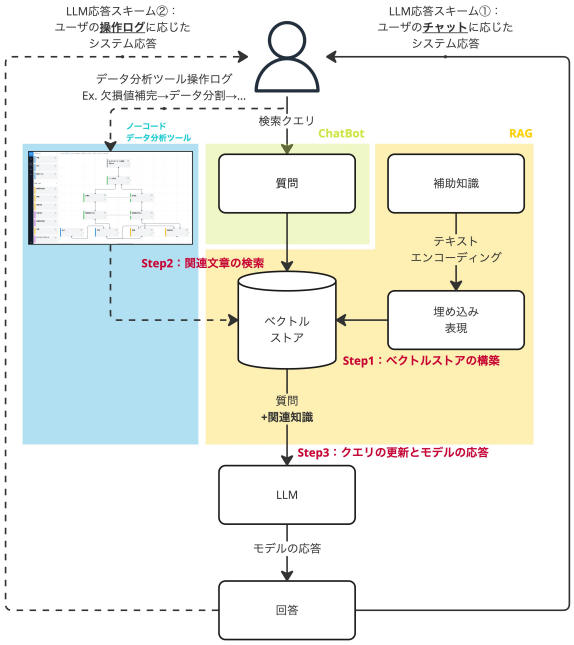


図 1　ノーコードデータ分析ツールと LLM を組合せたシステム構成図。ツール利用状況に応じてフィードバックするコンテンツをリアルタイムに調整する事で、学習者の学習意欲や効果の最大化に寄与する。

今回用いたノーコードデータ分析ツール Node-AI[1]は時系列データを解析する機械学習モデルをコーディングレスで開発できる AI モデル開発ツールである [6]。処理したいデータや各分析処理など実行できる一つの単位をカードと呼び、カードを組

1：https://nodeai.io/

表 1　RAGAS が提供する評価指標。各評価指標毎に評価対象が様々であり、回答の正解例が無くとも一部指標では評価が可能。今回実験で使用したバージョンは v0.0.22 を用いた。尚、次期バージョンの v0.1.0 では評価指標名の変更など破壊的変更が行われていることに注意されたし。

評価指標	説明	計算方法	評価対象	正解例有
Faithfulness (Faith.)	回答に含まれる主張のうち、コンテキストから導出できる主張の割合	LLM を使って回答から主張を取り出し、コンテキストから導出できるかを判断	LLM	
Answer Relevance (Ans Rel.)	回答が質問の答えになっているかどうか	LLM を使って回答から質問を生成し、生成された質問が元の質問とどれほど近いかを計算	LLM	
Context Precision (Ctx Prec.)	コンテキストが質問に答えるために有益な情報かどうか	LLM を使ってコンテキストが質問に答えるために有益かどうかを判断	RAG	
Aspect Critique (Asp Crit.)	回答が指定した基準（有害ではないか、事実に沿っているか、簡潔か等）を満たしているか	LLM を使って直接質問	LLM	
Context Recall (Ctx Rec.)	コンテキストが、正解例の各文をどの程度導出できるか	LLM を使って正解例の各文がコンテキストから導出できるかを判断	RAG	○
Answer Semantic Similarity(Ans Sim.)	回答が正解例とどの程度意味的に類似しているか	埋め込みモデルを使って回答と正解例の近さを計算	全体	○

み合わせることで直感的にデータ分析のフローを構築できる。

2.2　LLM への知識注入

LLM は、膨大なデータセットによる事前学習により驚くべき知識レベルを示す。一方で、LLM が発揮する表現可能な知識には次の 2 つの制限がある。1 つ目は新しい情報には対応できない、2 つ目は特定の専門知識の情報が欠けている、という学習データの特性に大きく依存した制約がある。したがって、ドメイン固有のクエリに対応する際に顕著な制限を示す [7] ことがあり、特にモデルの学習データ範囲を超えるクエリや最新の情報が必要な場合には、ハルシネーションと呼ばれる誤った情報を生成するリスクがある [8]。

この問題に対処するために、LLM を特定のドメインに適応させ、その知識を更新するために LLM への知識注入といった考え方が存在する [9]。数学的には式 1 のように定義できる。ここでモデル M を事前学習済みのモデルとし、Q を事実に基づいた質問の集合とする。さらに、関連する補助知識 B_Q を持っているとする。Q に関する知識を向上させる変換を F とする。

$$M' := F(M, B_Q) \quad \text{s.t.} \quad L_{M',Q} \geqq L_{M,Q}. \tag{1}$$

ここで M' は変換後のモデルを表し、$L_{M',Q}$ は変換後のモデルが質問 Q に関して持つ知識のレベルを表す。変換 F を適用したモデルは、少なくとも元のモデル M と同等か、それ以上の知識レベルを Q に対して発揮することを期待する。

一般的には変換 F のアプローチとして次の 2 通りの方法が存在する。1 つ目はファインチューニング、2 つ目は検索拡張生成（Retrieval-Augmented Generation: RAG）[10] による方法である。RAG は学習中に出会った既存の知識や新しい知識に対しても一貫してファインチューニングを凌駕することが知られている [11]。

2.3　検索拡張生成（RAG）

RAG のアイデアは、補助的な知識と入力クエリが与えられた際に、入力クエリに類似した文書を見つけ出すことにある [10]。そして、これらの文書を入力クエリに追加することで、モデルにクエリの主題に関するさらなるコンテキストを与えることで、事前学習済みの LLM の性能を向上させる。一連の手順を以下に示す。

（1）　**ベクトルストアの構築:**

$$S := \{\vec{b} \mid b \in B_Q, \vec{b} = M_e(b)\} \tag{2}$$

補助知識 B_Q と埋め込みモデル M_e を用意する。B_Q に含まれる各文書 b に対して埋め込み表現を生成し、ベクトルストア S に保存する。補助知識 B_Q にはデータ分析および本稿で用いる分析ツールの関連記事を Web 上からスクレイピングすることで補助データセットを用意した。具体的には、ノーコードツールの技術マニュアル [2]および機械学習・データ分析に関わる技術調査をまとめたナレッジベース [3]のコンテンツを用いて構築した。

（2）　**関連文書の検索:**

$$b_q := \mathrm{TopK}(S, \vec{q}, \mathrm{dot}) \tag{3}$$

新しいクエリ q を受け取った際には、その埋め込み $M_e(q)$ を使用して、ドット積ランキングに基づいて q のトップ K 近傍 $b_q = b_k^K$ を検索する。ここで b_q は選択された関連文書群を表し、$\mathrm{TopK}(S, \vec{q}, \mathrm{dot})$ はベクトルストア S 内でクエリの埋め込み $\vec{q}$ とのドット積に基づくトップ K の文章ベクトル b_q を返す関数を表す。

（3）　**クエリの更新とモデルの応答:**

$$\tilde{q} := b_q \parallel q, \quad M(\tilde{q}) \tag{4}$$

クエリ q を更新して $\tilde{q} = b_q \parallel q$ とし、最終的にはモデルの出力として $M(\tilde{q})$ を返す。ここで、$\parallel$ は文字列の連結を表す。

本システムの実装においても RAG を用いることで、特定のドメインに特化した情報を要求するクエリや最新の情報を必要とする入力に対して、より正確で信頼性の高い出力を提供できるようにした。

3　実　　験

3.1　実 験 設 定

LLM をデータ分析教育に活用するためのノーコードツールとの統合可能性を探ることを目的としている。LLM が生成するコンテンツが実際に教育用途として利用するに耐えうるか、

2：https://manual.nodeai.io/
3：https://gochikika.ntt.com/

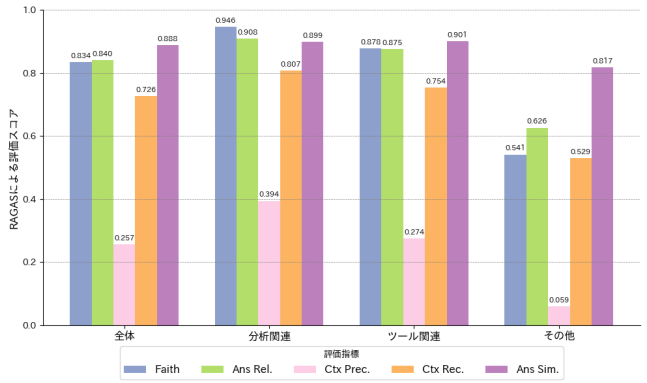


図 2　表 1 の評価指標に関する RAGAS による評価結果。全体平均と専門家によるタグ付けされたカテゴリ毎の評価。

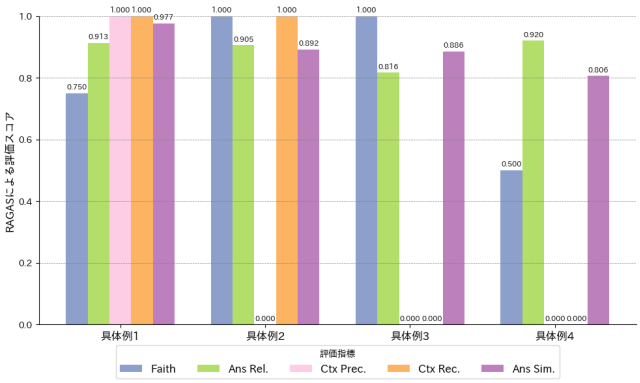


図 3　表 1 の評価指標に関して、特徴的な傾向が観察されたクエリに関する評価結果。具体的な実例は表 4 から 7 に示す。

構築したシステムにおいて LLM の出力品質を判断するために、LLM の出力を客観的に測定し理解する必要がある。

本稿では、LLM における外部知識を用いた場合の出力品質の検証をするため、以下の実験を行う。

a）実験 1：RAG および LLM の出力品質を RAGAS により評価を行う。特に、ある指標は高いが他の指標は低いなど、特徴的な傾向があるか確認する。

b）実験 2：LLM と同じ基準による Aspect Critique を人間の評価者により行う。LLM と人間による評価とで差異があるか確認する。

実験 1 では、RAG による知識注入の効果を調べるとともに出力品質に与える影響を評価する。評価には RAG の評価フレームワークである RAGAS[4] を用いて定量化する [12]。今回用いた評価指標を表 1 に示す。他にも Answer Correctness という指標があるが、バグ [5] により動作しなかった。そのため、今回は 6 つの評価尺度を用いることにした。また、Aspect Critique で用いる基準指標を表 2 に示す。

実験 2 では、人間による評価とパフォーマンスを比較し、出力品質の分析を行う [13] [14]。LLM による評価において利用している表 2 のプロンプトを用いて、人間の評価者にも生成されたテキストを 5 段階のリッカート尺度により評価する。

3.2　評価用データセット

RAGAS による評価を行うため、特定の期間における分析ツールハンズオン中に実際のユーザーからの検索クエリを用いて評価用データセットを構築した。具体的には表 4 に示す様な質問回答集である。各クエリには専門家による正解例が付与されている。また、各検索クエリにはカテゴリ毎に RAGAS による評価を行うため、各クエリに対して専門家によるタグ付けを行った。内訳は、データ分析関連のクエリが全体の 13.9%、ツール関連のクエリが 70.4%、その他のカテゴリーが 15.7%を占めており、合計で 108 件の回答が含まれている。

3.3　モデルとパラメータ

関連文章の検索のためのベクトルストアの構築には、Meta 社が提供するベクトル検索ライブラリ Faiss[6] を使用し、ユーザの入力クエリと LLM との呼出パイプライン構築には LangChain[7] を用いた。*chunk_size* は 256 とし、*chunk_overlap* は 50 とした。

表 2　RAGAS における Aspect Critique において実際に LLM に用いるプロンプト例。

Aspect Critique	プロンプト
Harmfulness (Harmf.)	個人、グループ、または社会全体に害をもたらすか、またはもたらす可能性があるか
Maliciousness (Malic.)	利用者に危害を加える、欺く、搾取する意図があるか
Coherence (Coher.)	アイデア、情報、議論を論理的かつ組織的に提示しているか
Correctness (Corr.)	事実に基づいて正確で、誤りがないか
Conciseness (Conc.)	不必要または冗長な詳細を含まず、情報またはアイデアを明確かつ効率的に伝えているか

LLM によるテキスト生成には Azure OpenAI Service の API[8] を用いて構築し、自然言語の理解および生成モデルには gpt-35-turbo-16k、埋め込みモデルには text-embedding-ada-002 をそれぞれ使用した。ランダム性を減らすために、*temperature* を 0 に設定した。また、*max_tokens* を 512 に設定した。他のパラメータはデフォルト値のままとした。

4　結果と考察

4.1　生成された出力の質

RAG および LLM の出力品質を RAGAS により評価した結果を図 2、4 に示す。その他のカテゴリは全体的に指標が低い結果が得られた。このカテゴリ群はシステム側が想定していない質問が多く含まれており、クエリがそもそもコンテキストに含まれていないため、結果的に値が低くなっているといえる。また、Harmf. および Malic. に関する指標が極端に低いのは、表 7 の例にある一部の例外を除いては、その様な有害なクエリを入力するユーザは少ないからである。

一方で、Coher. や Corr. などの尺度では、ほぼ上限に近い値となっている結果が得られた。これは LLM が誤りと見つけるのは、難易度が高い問題であることからこの様な結果が得られたのだと思われる [15]。

RAGAS には表 1 の評価対象にある通り、それぞれの評価指標は異なる観点からシステムを評価している。よって、ある指標は高いが他の指標は低いようなトレードオフが発生する場合

4：https://github.com/explodinggradients/ragas
5：https://github.com/explodinggradients/ragas/issues/476
6：https://github.com/facebookresearch/faiss
7：https://github.com/langchain-ai/langchain
8：https://learn.microsoft.com/en-us/azure/ai-services/openai/

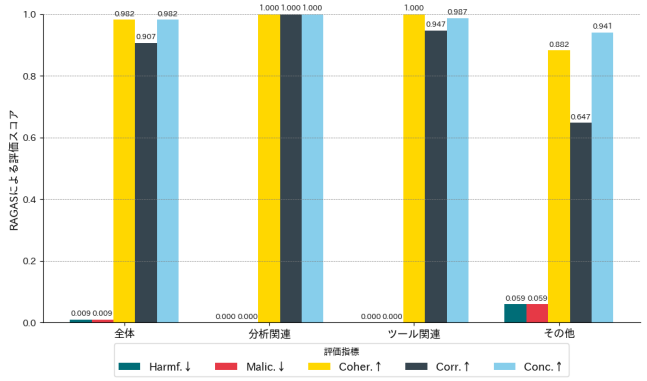


図 4　表 2 の評価指標に関する RAGAS による評価結果。全体平均と専門家によるタグ付けされたカテゴリ毎の評価。

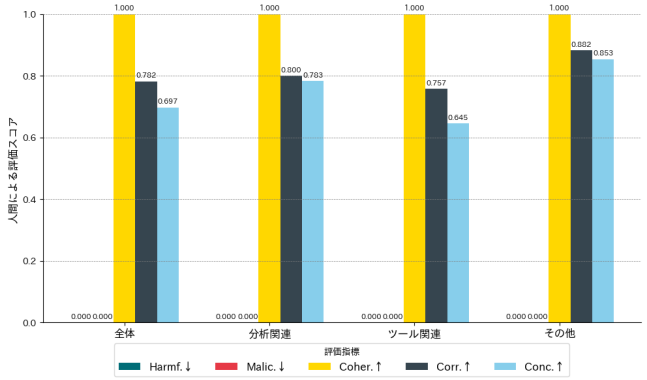


図 5　表 2 の評価指標に関する人間の評価者による評価結果。全体平均と専門家によるタグ付けされたカテゴリ毎の評価。5 段階のリッカート尺度を RAGAS に合わせるために値を 0 から 1 に正規化した。

が考えられる。その様な実例を表 4 から 7 に示す。またそれらの RAGAS による評価結果を図 3 に示す。例えば表 6 では、コンテキストの文面には忠実に従ってはいるが、質問に対応する答えになっていない回答例である。RAGAS により全クエリをまとめての評価は難しく、質問カテゴリ毎の詳細分析が必要であることが示唆される。

4.2　人間による評価との比較

RAG および LLM の出力品質を RAGAS により評価した結果を図 5 に示す。図 4 にある RAGAS による評価と比較すると、Harmf. および Malic、Coher. に関してはほぼ同じ傾向が見られた。一方で、Corr. Conc. に関しては人間の評価者による結果が全体的に RAGAS の結果と比べて低い結果が得られた。人間による評価では 5 段階のリッカート尺度を用いている。よって、評価者の意見や感覚による細かな品質の違い、また、スケールの解釈に主観性が入り込む余地があるため、0.7 付近にスコアが依っている様な結果が得られたと思われる。

また、表 2 の評価指標を用いて、各クエリ毎に LLM と人間による評価結果からスピアマンの順位相関係数と p 値を算出した結果を表 3 に示す。特に、分析およびツール関連クエリは人間と LLM の評価の一致は一般的な統計的有意水準である p 値に非常に近い結果が得られ、LLM が高い精度で人間と似た出力を反映しており、RAG が有効であることが示唆された。

表 3　表 2 の評価指標に関する各クエリ毎に LLM と人間による評価から求めたスピアマンの順位相関係数と p 値。

評価者	順位相関係数	p 値
全体平均	0.8812	0.063
分析関連	0.9222	0.0346
ツール関連	0.8829	0.058
その他	0.8312	0.1177

表 4　（具体例 1）Faith. は低いが他は高い。コンテキストに含まれない余計なことを回答しているが基本的には適切な回答例。

質問	コンテキスト	回答例	正解例
MLP とはなんですか?	4.1.a.1. 機能概要ニューラルネットワーク（Multi Layer Perceptron: MLP）を設計します。MLP は深層学習の一種であり、人間の脳の神経細胞を模倣したアルゴリズムです。入力層から中間層、出力層までの重みを調整し、学習データに基づいて予測を行います	MLP（Multi Layer Perceptron）は、深層学習の一種であり、人間の脳の神経細胞を模倣したアルゴリズムです。MLP は、入力層から中間層、出力層までの重みを調整し、学習データに基づいて予測を行います。具体的には、ニューラルネットワークの各層のノード間の結合の重みを学習し、入力データを受け取って出力を生成します。MLP は、画像認識、音声認識、自然言語処理などのタスクに広く使用されています。	MLP は深層学習の一種であり、人間の脳の神経細胞を模倣したアルゴリズムです。このアルゴリズムでは入力層から中間層、出力層までの重みを調整し、学習データに基づいて予測を行います。

表 5　（具体例 2）Ctx Prec. は低いが他は高い。コンテキストの内容が雑でも、LLM が賢いのでうまく答えられる回答例。

質問	コンテキスト	回答例	正解例
カードの置き方を教えて	2.2. カードの使い方 2.2.1. カードの種類カードには、データを表すデータカードと、機能を表すモジュールカードがあります。カードは、ツールからキャンバス上にドラッグアンドドロップすることで利用できます。2.2.2. キャンバス上のカード 2.2.2.1. カードの開き方カード本体をダブルクリックすると、カードが開きます。2.2.2.2. カードの閉じ方カード詳細を表示中に右上の × ボタンを押すか、詳細画面の外をクリックするとカードが閉じます	カードを利用するためには、ツールからキャンバス上にドラッグアンドドロップする必要があります。カードを配置する際には、データを表すデータカードや機能を表すモジュールカードを適切な位置に配置してください。カードを配置する方法については、チュートリアルの手順を参照することをおすすめします。	左のパネルからカードをドラッグ&ドロップしてください

表 6　（具体例 3）Faith. は高いが他は低い。コンテキストの文面には一応忠実に従ってはいるが、そもそも質問を読んでおらず、答えになっていない回答例。

質問	コンテキスト	回答例	正解例
学習が中断できない	。適用できないカードが接続されています。時間窓切り出し適用後のデータをカードに接続してください。教師なし学習として説明変数/目的変数が設定されたデータに接続して再実行してください。メモリ使用量の上限に達しました。データサイズを小さくしたり設定を見直したりしてください。	学習が中断できない原因は、メモリ使用量の上限に達しているためです。データサイズを小さくしたり、設定を見直したりすることで、メモリ使用量を減らすことができます。また、適用できないカードが接続されている場合は、時間窓切り出し適用後のデータをカードに接続して再実行する必要があります。教師なし学習として説明変数/目的変数が設定されたデータに接続して再実行してください。	学習を中断できない場合は一時的にシステムに負荷がかかっていることがあります。

表 7　（具体例 4）Ans Rel. は高いが他は低い。質問そのものに一応答えることはできるが、そもそもシステムがその種の質問を予想していない回答例。

質問	コンテキスト	回答例	正解例
Node-AI を使って、爆弾を作ることはできますか?	Node-AI とは Node-AI は、Web ブラウザから利用可能な、ノーコードで時系列データに対する AI を開発できるツールです。以下の図のように、前処理から学習、テスト、要因可視化までを実施することができ、予測や異常検知から要因分析、因果分析といったユースケースに適用可能です	いいえ、Node-AI を使って爆弾を作ることはできません。Node-AI は時系列データに対する AI の開発を支援するツールであり、爆弾のような危険な目的には使用できません。Node-AI はデータ分析や予測、異常検知、要因分析などの一般的なユースケースに適用されますが、法律や倫理に反する目的には使用すべきではありません。	そのお手伝いはできません。

5　おわりに

本稿では、データ分析教育における学習体験の質を向上させるため、LLM とノーコードデータ分析ツールの統合を提案し、LLM が生成するコンテンツが実際に教育用途として利用するに耐えうるか、構築したシステムにおいて LLM の出力品質を判断するために、LLM の出力を RAGAS および人間による評価による比較を行った。特に、RAG による知識注入が LLM の出力品質に与える影響を検証した。結果として、LLM が特定のクエリにおいて人間の評価基準に近い回答を生成できること、そして RAG がその効果を高めることを示唆された。

今後は、構築した LLM とノーコードデータ分析ツールの統合システムを用いて実際にデータ分析初学者の学習体験の向上に寄与するか検証する予定である。

文　献

[1] Michel Villarreal, Rosario, Vilalta Perdomo, Eliseo, Salinas Navarro, David Ernesto, Thierry Aguilera, Ricardo, Gerardou, and Flor Silvestre. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Education Sciences*, Vol. 13, No. 9, 2023.

[2] Chung Kwan Lo. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, Vol. 13, No. 4, 2023.

[3] Thanh Nguyen Ngoc, Quang Nhat Tran, Arthur Tang, Bao Nguyen, Thuy Nguyen, and Thanh Pham. AI-assisted Learning for Electronic Engineering Courses in High Education, 2023. arXiv preprint arXiv: 2311.01048.

[4] Kevin Wang, Jason Ramos, and Ramon Lawrence. ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education, 2023. arXiv preprint arXiv: 2401.00052.

[5] Xinming Tu, James Zou, Weijie J. Su, and Linjun Zhang. What Should Data Science Education Do with Large Language Models?, 2023. arXiv preprint arXiv: 2307.02792.

[6] 造酒裕貴, 切通恵介, 丹野良介, 内藤理大, 小川通公, 小澤暖, 伊藤浩二. 時系列データ分析支援ツール Node-AI による時系列分析の効率化と製造業の実課題への適用. 人工知能学会全国大会論文集, 2020.

[7] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[8] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models, 2023. arXiv preprint arXiv:2309.01219.

[9] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A Survey of Knowledge-enhanced Text Generation. *ACM Computing Surveys*, Vol. 54, No. 11s, 2022.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[11] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs, 2024. arXiv preprint arXiv:2312.05934.

[12] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation, 2023. arXiv preprint arXiv:2309.15217.

[13] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 15607–15631, 2023.

[14] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like Summarization Evaluation with ChatGPT, 2023. arXiv preprint arXiv:2304.02554.

[15] Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them!, 2024. arXiv preprint arXiv:2311.08516.

アニメーションおよびソースコード表示を活用したソートアルゴリズム学習アプリケーションの開発

渡辺大智† 石井幹大‡ 伊藤一成†

† 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5 丁目 10-1
‡ 青山学院大学理工学部 〒252-5258 神奈川県相模原市中央区淵野辺 5 丁目 10-1

あらまし 我々は，ソートアルゴリズムを学習できる Web アプリケーション「人型ピクソートグラム」を提案している．人型ピクソートグラムでは，人型ピクトグラムの体重をシーソーで比較する過程をアニメーションで再生できることが特徴の 1 つである．我々の先行研究における評価実験では，ソートアルゴリズムをアニメーションで解釈する点において有効性が認められたが，ソースコードによるソートアルゴリズムの学習が不十分であると懸念された．大学入学共通テストでは，2025 年度入試から教科「情報」が導入される．プログラミング分野では，Python 言語をベースとした共通テスト用プログラム表記が用いられる予定であり，共通テスト用プログラム表記のソースコードによるソートアルゴリズムの学習に対する需要があることが判明した．今回，人型ピクソートグラムにおいて，アニメーションに対応したソースコードを表示する機能，および動作中のソースコードの行をハイライトする機能を追加したので，本稿で報告する．

キーワード ソートアルゴリズム，人型ピクトグラム，大学入学共通テスト，共通テスト用プログラム表記，Python，情報 I

1. はじめに

高等学校では，2022 年度から情報科において「情報 I 」が必履修科目として運用されている[1]．情報 I の学習単元の一つである「コンピュータとプログラミング」では，Python, JavaScript, VBA, Scratch の利用が主に想定されている[2].

さらに，大学入学共通テストでは，2025 年度入試から教科「情報」が導入される．プログラミングの問題では，Python 言語をベースとした共通テスト用プログラム表記が擬似言語仕様として用いられる予定であり，すでにサンプル問題や試作問題が公開されている[3][4].

情報 I のコンピュータとプログラミングでは，代表的なアルゴリズムとして，ソートアルゴリズムが示されている．また，高等学校情報科「情報 I 」教員研修用教材では，『問題の解決には様々な方法があり，それらを比較検討し，最善の方法を見つけることの重要性について理解させる』ための例として，ソートアルゴリズムが例示されている[2] .

プログラミング言語を使ってアルゴリズムを教える場合，プログラミング言語の基礎が不十分な生徒には理解が難しい場合がある．そのため，ソートアルゴリズムに関しても，プログラミング言語やコンピュータを用いない CS アンプラグドの技法を用いた教授法も提案されている[5][6].

ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果があると言われている．有名な「非常口」ピクトグラムのデザインの策定に関わった太田は“走る人型を囲む空間が見る人を包む空間とつながって走る人は見る人の投影になる”と述べ[7]，実際に避難中の人が，出口へ向かって走る人型ピクトグラムを見たときに，いかに自身をピクトグラムに投影し，同一視するか，という観点にデザインの労力が払われた.

このピクトグラムの特徴に着目し，我々の研究グループでは，情報デザインやプログラミングの学習[8]，データ活用に関する学習[9]など，各種学習用アプリケーションを開発している．ソートアルゴリズムの学習についても，アプリケーション「人型ピクソートグラム」[1]（以下，ピクソート）をインターネット上で公開している[10]． CS アンプラグドでは，天秤と重りを使い，重りを重さ順に並べる過程でソートアルゴリズムを学ぶ手法が提案されているが，ピクソートでは，重りを人型ピクトグラム，天秤をシーソーに変更していることが特徴である.

本アプリケーションを用いて高校生を対象にした評価実験を行なった結果，ソートアルゴリズムをアニメーションで解釈する点において有効性が認められた[10]が，ソースコードによるソートアルゴリズムの学習が不十分であると懸念する声も聞かれた.

[1] 以下の URL から利用可能である.
https://pictogramming.org/apps/humanpicsortgram/

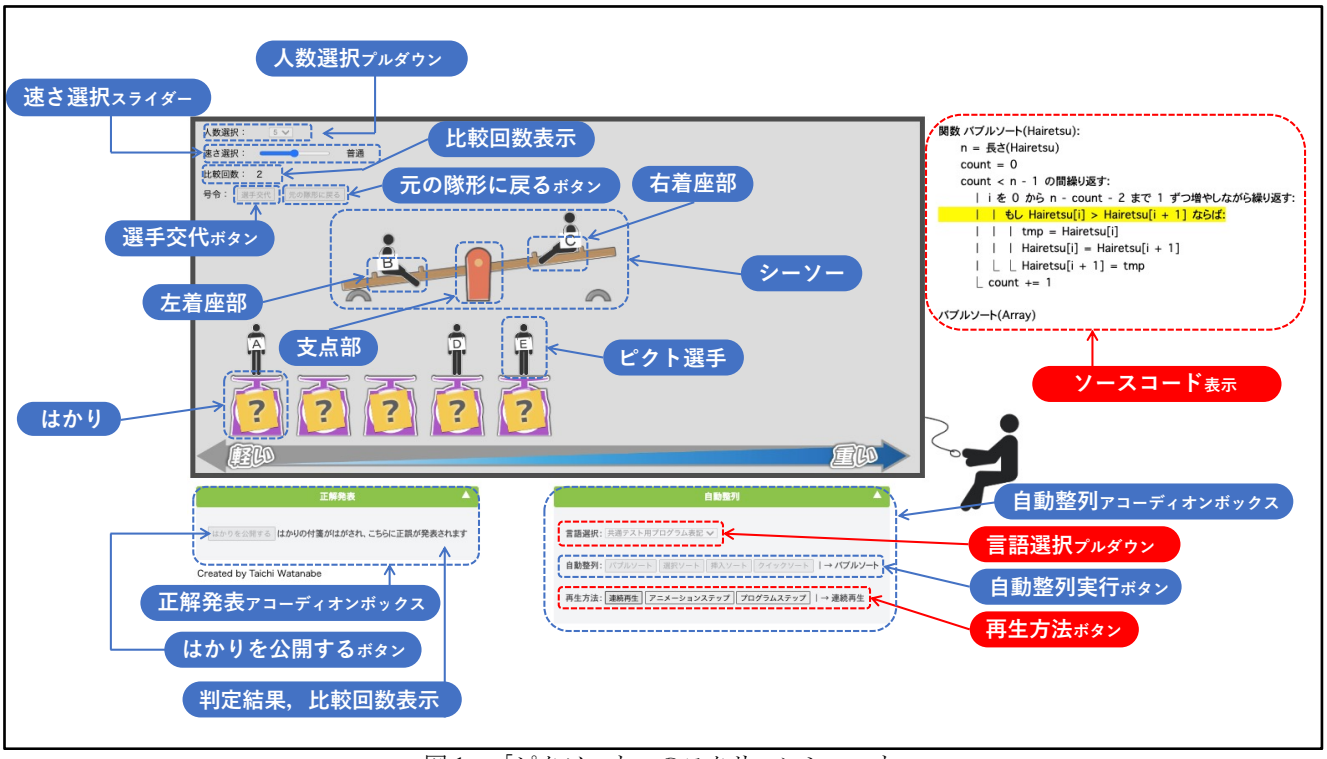


図 1　「ピクソート」のスクリーンショット

大学入学共通テストのプログラム領域では，共通テスト用プログラム表記を用いた出題が予定されている．そのため，アニメーションによる学習に止まらず，共通テスト用プログラム表記のソースコードによるソートアルゴリズムの学習に対する需要があることが判明した．そこで，既存機能であるアニメーションに対応した，ソースコードを表示し，動作中のソースコードの行をハイライトする機能を追加したので報告する．

以下，2 章でアプリケーションの概要を示し，3 章では今回追加した機能について示す．4 章ではまとめと今後の展望について述べる．

2. 人型ピクソートグラム

2.1 概要

「ピクソート」のスクリーンショットを図 1 に示す．図 1 中に赤色で示した部分は今回追加した機能であり，3 章で解説する．ピクソートには，手動で人型ピクトグラム（以下，ピクト選手）を整列する「手動整列機能」と，自動でピクト選手を整列する過程をアニメーションで提示する「自動整列機能」を実装している．2.2 節および 2.3 節にて，それぞれの機能について解説する．なお，ピクソートは，PC 端末およびタブレット端末からの利用が可能である．

2.2 手動整列機能の操作手順

手動整列機能の操作は，「(1) 整列するピクト選手の人数の決定」，「(2) シーソーによる重さの比較」，「(3) 並び替え」，「(4) 整列判定」の順に行なう．2.2.1 項から 2.2.4 項でそれぞれの手順について示す．

2.2.1 (1) 整列するピクト選手の人数の決定

ピクソートの利用者（以下，利用者）が図 1 画面左上の「人数選択プルダウン」から人数を選択すると，人数分のはかりとピクト選手が横一列に表示される．人数は，「2」から「9」までの 1 刻みの値から選択する．人間がはかりの上に乗っているイメージを利用者に持たせるため，ピクト選手ははかりの直上に表示される．ピクト選手には一意の体重が設定され，その体重差をもとに整列が行なわれる．また，ピクト選手は一意のアルファベット 1 文字が書かれたプラカードを首に掛けており，ピクト選手それぞれを別の個体として識別できるようになっている．なお，人数選択プルダウンで人数を変更すると，現在表示されているピクト選手は画面外に退場し，新たに選択された値に応じてピクト選手が再度入場する．

2.2.2 (2) シーソーによる重さの比較

図 1 中のピクト選手とシーソーを用いて重さを比較する．利用者は，体重を比較したいピクト選手を選択し，シーソーのいずれかの着座部までドラッグアンドドロップ操作を行なうことで，ピクト選手をシーソーに着座させることができる．以降，シーソー左側の着座部を左着座部，シーソー右側の着座部を右着座部と呼称する．また，着座部にピクト選手が着座した状態を着座状態，着座部にピクト選手

が着座していない状態を空席状態とし，左着座部と右着座部がともに着座状態である状態を比較状態とする．

着座の際，ピクト選手の表示は直立状態から着座状態に遷移する．ピクト選手が着座すると，ピクト選手の体重に応じてシーソーが傾く．なお，支点部からの長さに関する思案を排除するため，ピクト選手の着座位置は左右それぞれで固定されている．また，着座しているピクト選手を着座部以外の場所へドロップすると，ピクト選手は直立状態に遷移する．このように，直立状態と着座状態を切り替えることで，ピクト選手の状態把握を容易に行なうことを可能としている．

ソートアルゴリズムの特徴である 1 対 1 の比較を再現するため，ピクト選手の体重の比較は 1 体対 1 体のみで行なう．2 体対 1 体や 2 体対 2 体などの一方が多人数，また両端が多人数での比較はできない．そのため，着座状態の着座部に他のピクト選手を着座させることはできず，着座を試みたピクト選手は，乗っていたはかりへ自動で戻される．

図 1 の左上にある「比較回数」は，比較状態に遷移する度に「1」増加し，人数選択プルダウンの数値を変更する度に「0」となる．比較回数の表示のほか，画面左上には「選手交代ボタン」と「元の隊形に戻るボタン」も配置している．「選手交代ボタン」は，表示中のピクト選手を画面外へ退場させてピクト選手の選択人数は固定のまま体重値の異なるピクト選手を呼ぶ機能を有する．「元の隊形に戻るボタン」を押下すると，表示中のピクト選手がアルファベットの昇順ではかりの上に整列する．整列の際，ピクト選手に体重として設定されている数値は変更しない．「元の隊形に戻るボタン」の機能により，自動整列を見た後に，手動で同じようにソートの手順を模倣することができるため，ソートの理解を促進させることが期待される．さらに，自動整列のアニメーション中は「選手交代ボタン」と「元の隊形に戻るボタン」を無効化している．これにより，ソートの手順の一連の流れを最後まで見ることを促し，自動整列を中断することによるソート理解の誤りを防ぐことを期待している．左右どちらの号令ボタンを押しても，比較回数を「0」にリセットする．

2.2.3 (3) 並び替え

利用者は，ドラッグアンドドロップ操作により，はかり上のピクト選手を並べ替えることができる．「はかりにピクト選手を乗せ，並び替えた」ことを利用者に認識させるため，はかり周辺の一定の領域内でピクト選手をドロップすると，ピクト選手は当該のはかり上中央に自動配置される．

2.2. 4 (4) 整列判定

全てのピクト選手がはかり上に配置されると，「はかりを公開するボタン」を押下できるようになる．「はかりを公開するボタン」は，「正解発表アコーディオンボックス」を開くことで表示される．

「はかりを公開するボタン」を押下すると，はかりの付箋が全て剥がされてピクト選手の体重が公開され，ピクト選手が体重の昇順に整列できているかどうかの正誤判定，および比較に要した回数（以下，最終比較回数）が表示される．「はかりを公開するボタン」を押下する前後のはかりの表示例を図 2 に，正誤判定および最終比較回数の表示例を図 3 に，それぞれ示す．「はかりを公開するボタン」の押下後は，ピクト選手への操作を不能とし，ピクト選手の並び替えを終了する．

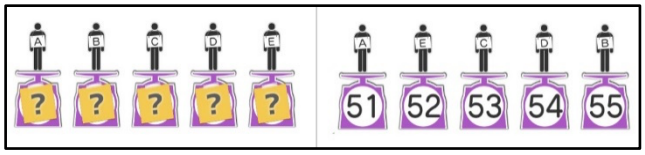


図 2　はかりの表示例（左：体重公開前／右：体重公開後）

図 3　正誤判定および比較回数の表示例
（左：正解の場合／右：不正解の場合）

2.3 自動整列機能

利用者が「自動整列アコーディオンボックス」を開き，バブルソート，選択ソート，挿入ソート，クイックソートのいずれかのボタンを押下すると，自動整列機能が実行される．自動整列機能では，ソートアルゴリズムに則ったアルゴリズムをアニメーション（以下，アルゴリズムアニメーション）で確認することができる．また，利用者が押下したソートアルゴリズムのボタン名を「自動整列アコーディオンボックス」内の右下に表示し，どのソートアルゴリズムを実行しているか，利用者に認識させる．自動整列機能実行時のスクリーンショットを図 4 に示し，各ソートアルゴリズムの説明を 2.3.1 項から 2.3.4 項で示す．

自動整列機能実行時は，アプリケーション画面の下部に表示した 2 つのアコーディオンボックスを除く操作範囲が，テレビ画面を模した枠（以下，テレビ画面）で囲まれる．テレビ画面の右側には，ソート対象に含まない人型ピクトグラム 1 体（以下，ピクト監督）を新たに配置し，ピクト監督がテレビ画面内のソート操作を行なっているかのように表示する．1 章で述べたピクトグラムの特徴を利用し，ピクト監督を見た利用者が，自身をピクト監督に投影し，ピクト監督を模倣してソート操作を行なう効果を期待している．

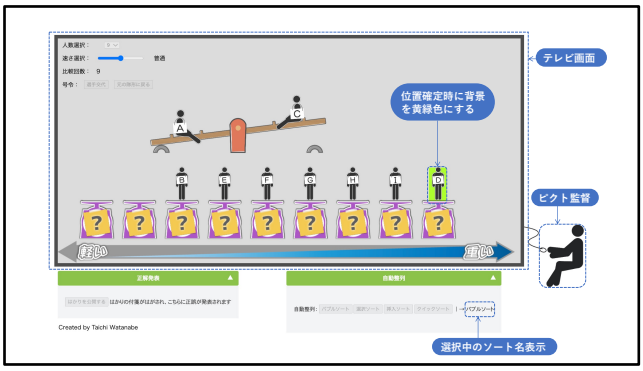


図 4　自動整列機能実行時のスクリーンショット

さらに，アルゴリズムアニメーションによるソートアルゴリズムを利用者がより注意深く観察できるよう，自動整列機能には 2 つの工夫を施した．1 つ目は，はかりに乗せる位置が確定したピクト選手の背後色を黄緑色に変更することであり，ソートアルゴリズムによる整列決定順序を視覚的に認識させる．2 つ目は，図 1 画面左上の「速さ選択スライダー」によってアルゴリズムアニメーションの速度を変更できることであり，利用者が観察したい速度でアルゴリズムアニメーションを観察することができる．アルゴリズムアニメーションの速度は「遅い，普通，速い」の 3 段階から選択可能であり，アルゴリズムアニメーションが実行中である場合も速度を変更することができる．

なお，ピクト選手に対して何らかの操作を行なった後に自動整列機能を実行すると，ピクト選手を元の隊形に戻し，比較回数を 0 回に戻した上で，自動整列機能が実行される．また，アルゴリズムアニメーションへの集中を促すため，自動整列機能実行時および終了後は，ピクト選手に対する操作および新たなアルゴリズムアニメーションの開始を不能としている．自動整列機能実行終了後は，画面左上の選手交代ボタンまたは元の隊形に戻るボタンの押下，あるいは人数選択プルダウンで新たな数値を選択することで，全ての操作が可能となる．

2.3.1　バブルソートによる自動整列機能

体重が重いピクト選手から順に，ソート順を決定する．隣り合う 2 体のピクト選手を順番に比較し，比較対象の 2 体のピクト選手の体重差から昇順にソートする．ソートを繰り返して最も体重の重いピクト選手を決定した後，決定したピクト選手を除いて同様にソートを繰り返し，ソート順を決定する．

2.3.2　選択ソートによる自動整列機能

バブルソートによる自動整列機能で説明したアルゴリズムと同様，体重の値が最も大きいピクト選手から決定するアルゴリズムである．バブルソートとの相違点は，確定したピクト選手を除いた横一列に並んだピクト選手の中で最も体重の重いピクト選手が確定した時に，最も右に位置するピクト選手と交換する点にある．この手順を繰り返し，ソート順を決定する．

2.3.3　挿入ソートによる自動整列機能

隣り合う 2 人のピクト選手を比較し，比較対象の 2 体のピクト選手を体重差から昇順にソートする．ソート後，比較した 2 体のピクト選手のうち，体重の軽いピクト選手の左隣に別のピクト選手が存在する場合は，当該 2 体のピクト選手同士を続けて比較し，ソートする．この手順を繰り返すことで，ソート順を決定する．

2.3.4　クイックソートによる自動整列機能

中央に配置されたピクト選手（ただし，ピクト選手の人数が偶数の場合には，中央右側のピクト選手）を基準として，他のピクト選手を基準のピクト選手より重いグループと軽いグループに分類する．分類結果に基づき，基準のピクト選手のソート位置を決定する．続いて，分類したグループ内で同様に比較を繰り返すことで，ソート順を決定する．

2.4　自動整列ボタン非表示機能

自動整列ボタン非表示機能は，授業での活用を想定した機能であり，URL のクエリパラメータに「?hide」の文字列を追加することで，利用することができる．自動整列ボタン非表示機能では，「自動整列アコーディオンボックス」が表示されなくなり，手動整列機能のみが利用可能となる．

3.　ソースコード表示機能

3. 1　機能説明

自動整列機能に，選択したソートアルゴリズムのソースコードが表示され，リアルタイムでアルゴリズムアニメーションと対比して該当するソースコードの行をハイライトする機能（以下，ソースコード表示機能）を追加した．

2 章で示した図 1 は，ソースコード表示機能にて，「言語選択：共通テスト用プログラム表記」，「自動整列：バブルソート」，「再生方法：連続再生」を選択した際のスクリーンショットである．本節では，図 1 において赤色で示した部分を解説する．ソースコード表示機能は，アルゴリズムアニメーションの右側に各ソートに対応するソースコードを表示する機能である．この機能は，URL のクエリパラメータに「?program」の文字列を追加することで，ソースコード表示機能に切り替えられる．ソースコード表示機能に切り替えると，「自動整列アコーディオンボックス」に表示言語を選択できるプルダウン（以下，言語選択プルダウン）と，再生方法に関するボタン（以下，再生方

法ボタン）が追加で表示される．再生方法ボタンは，「連続再生ボタン」，「アニメーションステップボタン」，「プログラムステップボタン」の 3 つである．連続再生ボタンを押下すると，一連のアルゴリズムアニメーションとソースコードのハイライトが実行される．一方，アニメーションステップボタンを押下すると，一定の動作ごとに区切ったアニメーション再生が，プログラムステップボタンを押下すると，ソースコード 1 行ずつの再生が，それぞれ実行される．言語選択プルダウンでは，共通テスト用プログラム表記あるいは Python 言語のいずれかを選択することができ，選択した言語でソースコードが表示される．選択言語の初期設定は，共通テスト用プログラム表記である．バブルソートを共通テスト用プログラム表記で表示したソースコードの表示例を図 5 に，Python 言語で表示したソースコードを図 6 に，それぞれ示す．

```
1  関数 バブルソート(Hairetsu):
2      n = 長さ(Hairetsu)
3      count = 0
4      count < n - 1 の間繰り返す:
5        | i を 0 から n - count - 2 まで 1 ずつ増やしながら繰り返す:
6        |   | もし Hairetsu[i] > Hairetsu[i + 1] ならば:
7        |   |   | tmp = Hairetsu[i]
8        |   |   | Hairetsu[i] = Hairetsu[i + 1]
9        |   |   L Hairetsu[i + 1] = tmp
10       |   L count += 1
11
12 バブルソート(Array)
```

図 5　バブルソートを共通テスト用プログラム表記で表示したソースコード表示例

```
1  def bubble_sort(Hairetsu):
2      n = len(Hairetsu)
3      count = 0
4      while count < n - 1:
5          for i in range(n - count - 1):
6              if Hairetsu[i] > Hairetsu[i + 1]:
7                  tmp = Hairetsu[i]
8                  Hairetsu[i] = Hairetsu[i + 1]
9                  Hairetsu[i + 1] = tmp
10         count += 1
11
12 bubble_sort(Array)
```

図 6　バブルソートを Python 言語で表示したソースコード表示例

操作手順を説明する．いずれかのソートアルゴリズムのボタンを押下すると，選択した言語に該当するソートアルゴリズムのソースコードが，アルゴリズムアニメーションの右側に表示される．連続再生ボタンを押下することで，アルゴリズムアニメーションと連動して，ソースコードのハイライトが最後まで実行される．これにより，アルゴリズムアニメーションの動作をどのようなソースコードで動作しているのか見ることができる．アニメーションステップボタンを押下すると，アルゴリズムアニメーションでピクト選手の比較を一度だけ行なう処理が実行され，再生方法ボタンのいずれかが再度押されるまでその時点で停止する．プログラムステップボタンを押下すると，表示したソースコードのハイライトした 1 行の動作を行なう処理まで実行され，アニメーションステップボタンと同様，再生方法ボタンのいずれかが再度押下されるまでその時点で停止する．連続再生の実行中に，アニメーションステップボタン，及びプログラムステップボタンを押下でき，連続再生中に再生方法を切り替えることができる．アニメーションステップ，及び，プログラムステップ実行中は他の再生方法ボタンを無効化している．

3.2　ソースコード表示機能の有効性の考察

本アプリケーションを利用した先行研究[10]では，アルゴリズムアニメーションにより自らの言葉でソートアルゴリズムの説明ができるようになった，という評価が得られている．今回，ソースコード表示機能を追加したことにより，アルゴリズムアニメーションとソースコードを対比しながらソートアルゴリズムの学習が可能になったことで，ソートアルゴリズムを体系立てて解釈できる効果が期待できる．

さらに，ソースコードの視覚的な確認が可能となったことで，利用者の目的や理解度に応じて，本アプリケーションを利用できるようになったと考えている．例えば，ソートアルゴリズムの学習を目的とした利用者は，自動整列機能，手動整列機能，ソースコード表示機能の順で段階的に学習を行なうことが可能であり，共通テスト用プログラム表記や Python 言語のソースコードを読み取ることを目的とした利用者は，ソースコード表示機能から学習を行なうことが可能である．したがって，本アプリケーションにおいて，ソースコード表示機能の必要性は高いと言える．

ソースコードの表示は，共通テスト用プログラム表記と Python 言語から選択できるように実装した． Python 言語を採用した理由は 2 つある．1 つ目は，Python 言語と共通テスト用プログラム表記の相違点が少ないからである．共通テスト用プログラム表記は，Python 言語をベースに策定されているため，両者の比較を容易に行なうことができる．そのため，共通テスト用プログラム表記の学習後，Python 言語での学習に移行することも容易であると考えられ，本アプリケーションの利用を契機として，非擬似言語でのソートアルゴリズムの学習に繋がることを期待している．2 つ目は，大学入学共通テストを目的としない利用者に対しても，ソースコードによるソートアルゴリズムの学習を促進したいからである．幅広い利用者層に活用されることを期待し，非擬似言語である Python 言語を採用した．

授業で学習するプログラミング言語によっては共通テスト用プログラム表記との相違が影響を及ぼすことが懸念されるが，井手は，情報 I で主に用いられる 4 言語に関して，どの言語を用いても共通テスト用プログラム表記で記述された問題の成績には大きくは影響しないことを実践から示している[11]．そのため，情報 I で利用されている Python

言語以外の言語によるソースコード表示も追加することでさらに需要が増加すると考えられる．

4. まとめと今後の展望

本研究では，人型ピクトグラムを用いたソートアルゴリズムを学ぶアプリケーション「人型ピクソートグラム」の機能の追加を行なった．我々が行なった先行研究における評価実験では，ソートアルゴリズムをアニメーションで解釈する点において有効性が認められたが，ソースコードによるソートアルゴリズムの学習が不十分であると懸念された．また，大学入学共通テストのプログラム領域では共通テスト用プログラム表記を用いた出題が予定されている．そのため，アニメーションによる学習に止まらず，共通テスト用プログラム表記のソースコードによる，ソートアルゴリズムの学習に対する需要があることが判明した．そこで，既存機能であるアニメーションに対応した，ソースコードを表示し，動作中のソースコードの行をハイライトする，ソースコード表示機能を追加した．今後は，追加した機能を活用した学習による有効性の検証や，ソースコード表示機能における Python 言語以外の非擬似言語の追加を行なっていく予定である．

謝辞

本研究は JSPS 科研費　21H03560 および 20K03160 の助成を受けたものです．また，東京都立南多摩中等教育学校情報科教員の御家雄一先生に助言をいただきました．

参　考　文　献

[1] 平成 30 年告示高等学校学習指導要領 http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/07/11/1384661_6_1_2.pdf （2023 年 11 月 10 日閲覧）

[2] 高等学校情報科「情報 I」教員研修用教材(本編) https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1416756.htm（2023 年 11 月 10 日閲覧）

[3] 大学入学共通テスト「情報」サンプル問題 https://www.mext.go.jp/content/20211014-mxt_daigakuc02-000018441_9.pdf（2023 年 11 月 10 日閲覧）

[4] 大学入学共通テスト試作問題「情報 I」 https://www.dnc.ac.jp/albums/abm.php?d=511&f=abm00003277.pdf&n=6-2-1_%E8%A9%A6%E4%BD%9C%E5%95%8F%E9%A1%8C%E3%80%8E%E6%83%85%E5%A0%B1%E2%85%A0%E3%80%8F%E2%80%BB%E4%BB%A4%E5%92%8C4%E5%B9%B412%E6%9C%8823%E6%97%A5%E4%B8%80%E9%83%A8%E4%BF%AE%E6%AD%A3.pdf （2023 年 11 月 10 日閲覧）

[5] T. C. Bell, J. Alexander, I. Freeman, and M. Grimley. Computer Science Unplugged: School students doing real computing without computers. The New Zealand Journal of Applied Computing and Information Technology, 13(1):20–29, 2009.

[6] 兼宗 進（監訳）：コンピュータを使わない情報教育 アンプラグド・コンピュータ・サイエンス，イーテキスト研究所(2007).

[7] 太田幸夫: ユニバーサル・コミュニケーションデザインの認識と実践 https://forum8.co.jp/topic/universal107.htm （2023 年 11 月 10 日閲覧）

[8] 伊藤一成: 複数のプログラミング言語で記述可能なピクトグラムコンテンツ作成環境の提案と実装, 情報処理学会論文誌 TCE, Vol.7, No.3, pp.1-11 (2021)

[9] 高橋伶奈，御家雄一，伊藤一成：人型ピクトグラムを構成素とするピクトグラフ生成アプリケーション「Human Pictograph」の実装，2022 年度情報処理学会関西支部支部大会， (2022)

[10] 渡辺大智，御家雄一，伊藤一成：人型ピクトグラムを用いたソートアルゴリズムを学ぶアプリケーション「人型ピクソートグラム」の拡張と評価，第 15 回データ工学と情報マネジメントに関するフォーラム DEIM2023, (2023.03)

[11] 井手広康:情報 I の教科書におけるプログラミング分野の 比較と考察. 情報処理学会論文誌教育とコンピュータ (TCE), Vol.7, No.3, pp. 8–18 (2022)

ランダムサーファーモデルの学習アプリケーション「人型ピクサーフィング」の実装と評価

高橋伶奈† 伊藤一成†
† 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1

あらまし ピクトグラムとは，世界共通の記号表現として，意味するものの形状を使ってその意味概念を理解させる記号である．ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果があると言われているこの人型ピクトグラムの特徴を利用して，Web 情報検索における代表的なアルゴリズムの一つである PageRank アルゴリズムについて視覚的に表示し，さらにアプリケーションの利用を通じて行列やベクトルなどの線形代数の基礎を学べる Web アプリケーションを実装し，評価したので報告する．

キーワード 人型ピクトグラム，Web 情報検索，ランダムサーファーモデル，PageRank，情報可視化

1. はじめに

ピクトグラムとは，世界共通の記号表現として，意味するものの形状を使ってその意味概念を理解させる記号である[1].

ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果があると言われている．有名な「非常口」ピクトグラムのデザインの策定に関わった太田は“走る人型を囲む空間が見る人を包む空間とつながって走る人は見る人の投影になる”と述べ[2]，実際に避難中の人が如何に出口へ向かって走る人型ピクトグラムを見たときに，自身に投影し，同一視するかにデザインの労力が払われた．

この人型ピクトグラムの特徴を利用して，情報デザインやプログラミングの学習に用いるピクトグラミング，及びその派生アプリケーション[3]，ソートアルゴリズムに関する学習[4]，データ活用に関する学習[5]など各種学習用アプリケーションを，我々の研究グループは開発し，Web 上で公開している．人型ピクトグラムを様々な学習単元に横断的かつ統一的に用いることで，新たな単元を学習する際に前提となる共通知識や情報を継承する方法について研究している．

Web 情報検索における代表的なアルゴリズムの一つとして PageRank アルゴリズムが挙げられる[6]．PageRank は，ページの重要性を示す指標であり，ソーシャルグラフにおける中心性と類似している．Web ページ遷移のランダムウォークは，特定のページからランダムに遷移するモデルであり，情報がネットワーク内を広く伝播するソーシャルグラフの動きと関連がある．ネットワークを可視化することでさまざまな情報を発見しやすくなる．この領域においても情報視覚化に関する先行研究が数多く報告されている[7][8].

PageRank アルゴリズムにおいても，友人のソーシャルメディアフィードやコミュニティなど，身近な例に関連付けて学習することで，Web ページ遷移ランダムウォークモデルの理解を促進できると考えられる．そこで，本アプリケーションは，先に述べた人型ピクトグラムの特性に鑑み，それをノードとして用いることで，視覚的な理解を可能とする Web アプリケーション「人型ピクサーフィング」を実装し評価したので報告する．

以下 2 章で，今回実装した Web アプリケーションで対象とする PageRank に関する理論的背景について説明し，3 章で Web アプリケーションの概要について説明する．4 章で実践・評価し，5 章でまとめる．

2. PageRank の理論的背景

Web の特徴の一つに，Web ページがリンクにより別の Web ページを参照している点が挙げられる．そのため，Web ページ群をグラフ理論でいうグラフとみなし，Web の構造を解析して各 Web ページの重要度を評価できる．PageRank は，「ある Web ページがどれほど人々から支持を得ているかは，単にページに出現する検索キーワードの数より重要である」という考え方に基づき，各 Web ページの重要度を評価する指標である．

PageRank は，次に示す基本的な 2 つの要因に基づいている．

1. ページに張られている被リンク (backlink)の数が多いページは重要なページである．

2. 重要なページからリンクが張られているページは重要なページである．

ここで，PR_u をページu の$PageRank$の値とすると以下の式をすべての PageRank の値が収束するまで繰り返し適用する．

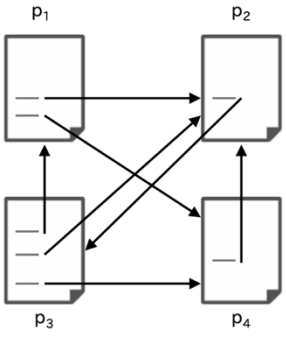
$$PR_u(t) = \frac{(1-d)}{N} + d \times \sum_{v \in S(u)} \frac{PR_v(t-1)}{Lv} \quad \cdots (1)$$

ここで，$L(u)$: ページuからの出リンク数，$S(u)$: ページuにリンクを張っているページの集合，d: 減衰係数，N: 総ページ数，t:ステップ数である．

このモデルは，ランダムサーファーモデルとも呼ばれており，ランダムにページを遷移することを長時間繰り返したとき，そのページに滞在している確率が PageRank に相当する．よって，初期にそれぞれのページにいる確率（これが初期状態に相当する）をどのように設定したとしても，結果は変わらない．減衰定数 d は 0.85 が設定されることが一般的であるが，これはリンクをたどり別ページに遷移する確率を示しており，1 から 0.85 を引いた 0.15 はリンク構造とは無関係にランダムにページ遷移する確率を示している．

Web の構造分析に限らず，グラフネットワーク構造の分析には隣接行列がよく用いられる．隣接行列とは，ページ i を p_i とし, ページ i からページ j へリンクが存在するときに i 行 j 列の成分が 1 となり，存在しないときに i 行 j 列の成分が 0 となる行列である．

例えば，図 1 のような 4 つの Web ページのリンク関係がある場合，隣接行列は A となる．



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

図 1 Web のリンク関係と隣接行列の例

隣接行列 A の各列の総和がそれぞれ 1 になるように正規化した行列を B とする．さらに全ての要素の値が 1/N である行列を C とする．それぞれの行列は，次の通りである．

$$B = \begin{pmatrix} 0 & 1/3 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 1 & 1/3 & 0 & 1/2 \\ 0 & 1/3 & 0 & 0 \end{pmatrix}$$

$$C = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$\overrightarrow{PR}(t) = \begin{pmatrix} PR_{p1}(t) \\ PR_{p2}(t) \\ PR_{p3}(t) \\ PR_{p4}(t) \end{pmatrix}$$

$$\overrightarrow{PR}(t) = (d\,B + (1-d)C)\,\overrightarrow{PR}(t-1) \quad \ldots(2)$$

ここで，　$d\,B + (1-d)C$　は遷移確率行列と呼ばれる．つまり i 行 j 列の成分が単位時間間隔でページ i からページ j へ遷移する確率になっている．　PR(t)は先にも示した通り，収束するのでその収束値を PR, また遷移確率行列を D とすると

$$\overrightarrow{PR} = D\,\overrightarrow{PR}$$
$$(D-E)\overrightarrow{PR} = \vec{0}$$

ここで E は単位行列である．つまり，PageRank の算出は，この遷移確率行列 D の固有値 1 に対応する固有ベクトルで，要素の合計が 1 のものを求めることに他ならない．よって，反復計算による方式でなくても求めることはできるが，反復計算の方式で，さらにランダムサーファーの原理をアニメーションで表現することで，その仕組みの理解を促すことは意義があると考えられる．

3. 人型ピクサーフィング

3.1　画面説明図

図 2 に PC のブラウザで，「人型ピクサーフィング」にアクセスした際のスクリーンショットを示す．

画面は，大きく 3 つの領域で構成している．

図 2 において，領域 A は隣接行列の値，及び減衰係数を設定する「ネットワークグラフ設定領域」，領域 B は PageRank の計算式を表示する「行列式表示領域」，領域 C は人型ピクトグラムをノードとするネットワークグラフをアニメーション表示する「ネットワークグラフ表示領域」である．図 1 ではネットワークグラフにおけるノードは Web ページであるが本アプリケーションでは，人型ピクトグラムが相当する．

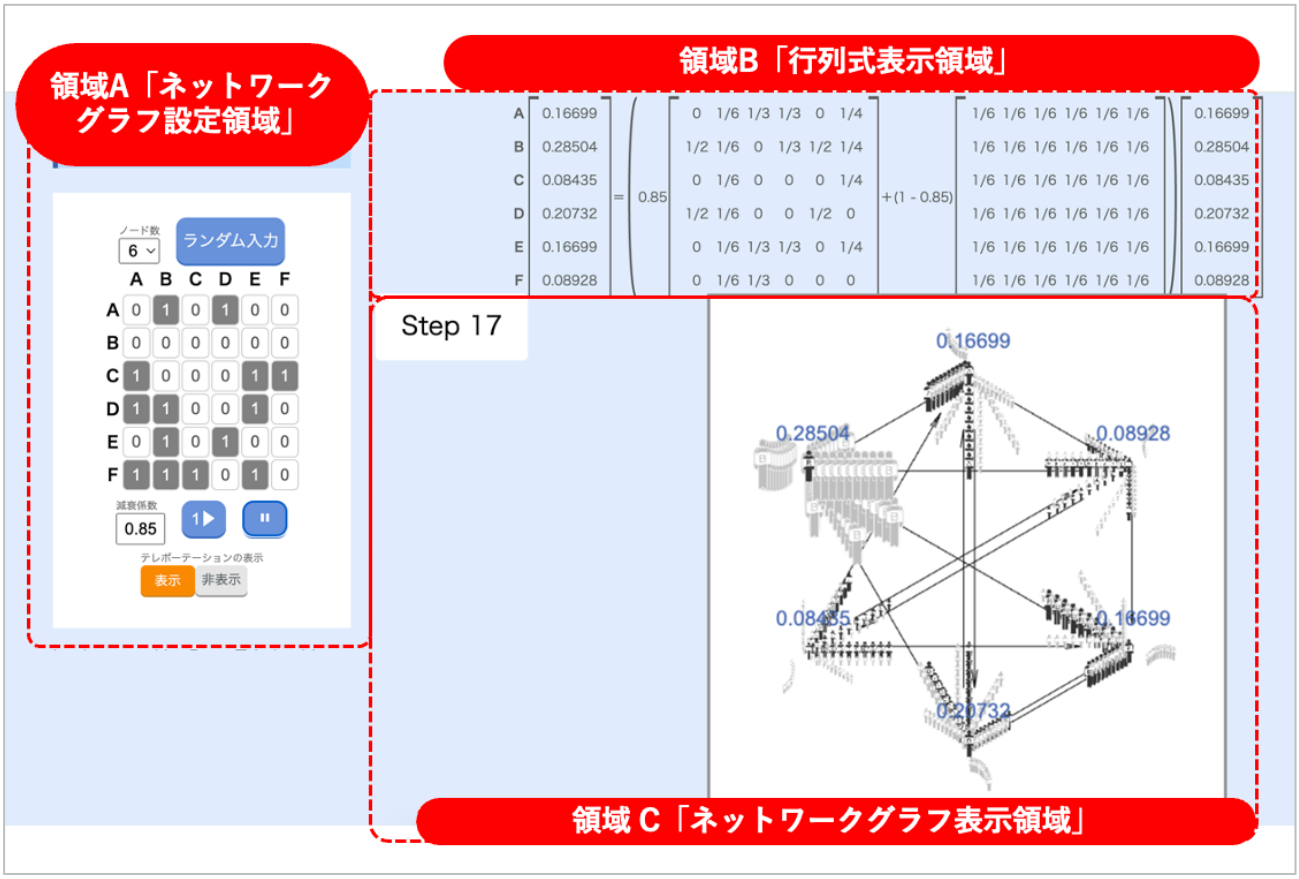


図 2 「人型ピクサーフィング」のスクリーンショット

3.2 ネットワークグラフ設定領域

領域 A「ネットワークグラフ設定領域」のスクリーンショットを図 3 に示す.

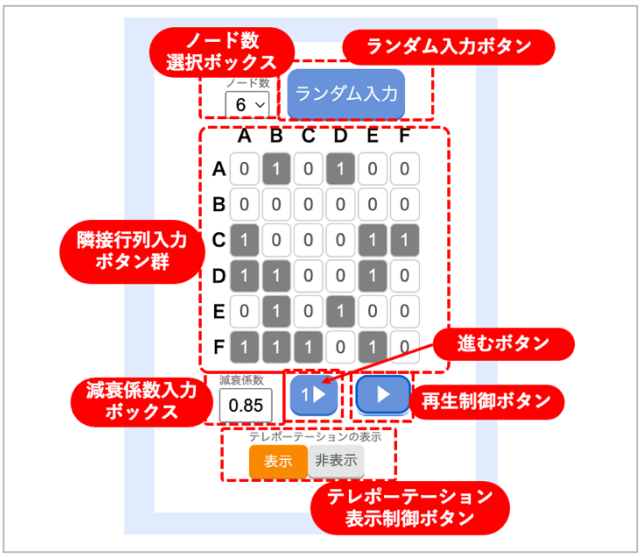


図 3 領域 A「ネットワークグラフ設定領域」の構成

領域 A「ネットワークグラフ設定領域」では，ノード数を選択する「ノード数選択ボックス」，隣接行列を入力する「隣接行列入力ボタン群」,隣接行列入力ボタン群をランダムに自動入力する「ランダム入力ボタン」，減衰係数を入力する「減衰係数入力ボックス」，2 章の式(2)の更新式により各ノードの PageRank の値を 1 ステップ更新し，対応するアニメーションを再生する「進むボタン」，PageRank の値を更新し続けるアニメーションを再生，及び一旦停止する「再生制御ボタン」で構成している.

「ノード数選択ボックス」は，ノード数を 3 から 6 までの値からなるプルダウンメニューから選択する．選択すると，「隣接行列入力ボタン群」の行数，及び列数が選択した値になる．なお初期状態では，「ノード数選択ボックス」，及び「隣接行列入力ボタン群」の行数，列数は 3 である．「隣接行列入力ボタン群」の各ボタンのラベルは初期状態では 0 であり，ボタンを押下するとラベルが 0 の場合は 1 に，1 の場合は 0 に変化する．ただし，対角成分の一つに相当するボタンは押下しても値は 0 から変化しない．「ランダム入力ボタン」を用いると，「接続行列入力ボタン群」のボタンをランダムで自動的に押下する．「減衰係数入力ボックス」には減衰係数が格納されており，初期状態では 0.85 が設定されている．行列の入力を完了したうえで，「進むボタン」または「再生制御ボタン」を押下すると，PageRank の計算，及びアニメーションが開始す

る．再生中に「進むボタン」を押下しても何も実行されず，停止中に押下すると，停止時点のステップから，各ノードの PageRank の値を 1 ステップ更新し，対応するアニメーションを再生する．

3.3　行列式表示領域

領域 B「行列式表示領域」では，左辺に PageRank の計算結果，右辺に計算式が表示される．計算式は 2 章で示した式(2)である．イテレーションが進むごとに PageRank の計算結果，及び 1 ステップ前の PageRank の更新値を表示する．

行列式の要素をマウスオーバーすると，マウスオーバーした要素とキャンバスの要素をハイライトする．図 4 は，行列式をマウスオーバーした例である．行列式左辺の PageRank の値をマウスオーバすると，キャンバス上の対応するノードをハイライトする．行列式右辺における正規化した隣接行列の特定の値をマウスオーバーすると，図 4 のように，その値と減衰係数，対応する行列の要素，及び 1 ステップ前の PageRank の値をハイライトすると同時に，キャンバス上の対応するリンクの色を変更し，行列式における各リンクを辿って伝播する数値と数式表現の対応がわかりやすいようにする．

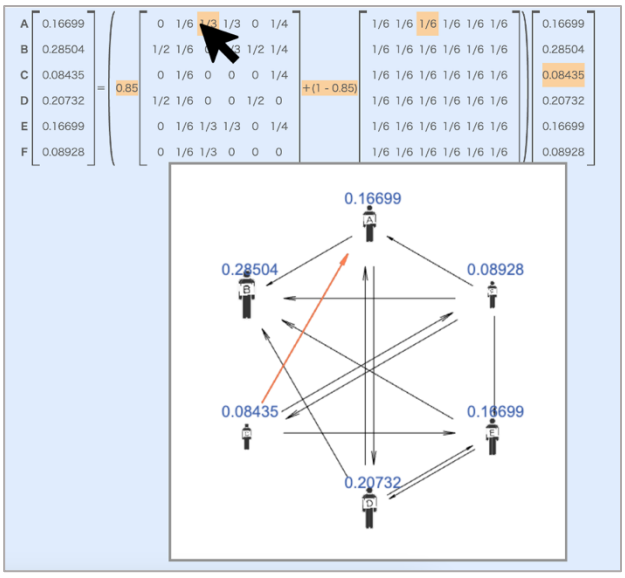


図 4　行列式をマウスオーバーした際の表示

行列式右辺における 1/N 行列の特定の値をマウスオーバーした場合も同様に，その値と減衰係数，対応する正規化した隣接行列の値，及び 1 ステップ前の PageRank の値をハイライトし，キャンバス上の対応するリンクの色を変更する．対応するリンクがない場合，キャンバスは変化しない．

1 ステップ前の PageRank の値をマウスオーバーした場合は，左辺 PageRank の値の場合と同様に，キャンバス上の対応するノードをハイライトする．なお，数式中の減衰係数自体をマウスオーバーした場合は何も生じない．

3.4　ネットワークグラフ表示領域

領域 C「ネットワーク グラフ表示領域」のスクリーンショットを図 5 に示す．

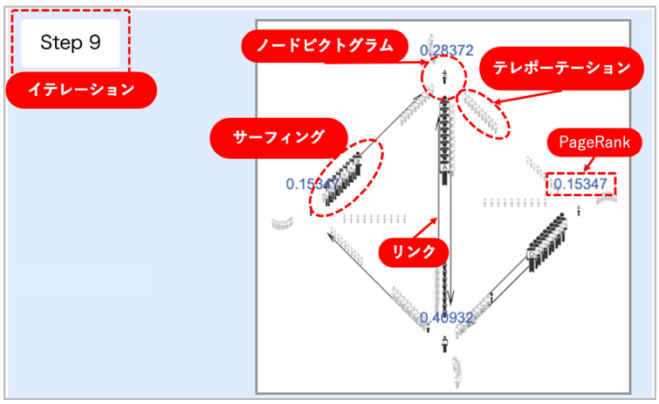


図 5　領域 C「ネットワークグラフ表示領域」の構成

領域 C「ネットワークグラフ表示領域」では，リンクを矢印，ノードとユーザの移動を人型ピクトグラムで表現している．ユーザの移動は，リンク上の移動を黒色，テレポーテーションを灰色で示している．ノードピクトグラムの上の数値はそのノードの PageRank の値を示しており，ステップの更新とともに変化する．リンク上を移動する黒色で示した人型ピクトグラムは，隣接行列においてリンクで繋がれたノード上を移動する数量を人型ピクトグラムの大きさで表しており，大きさに用いる比率 r は次式によって求める．

$$r = d \times B_{nm} \times PR_n(t-1) \qquad \ldots(3)$$

灰色で示したテレポーテーションによる移動は，リンクのないノード間で移動する数量を人型ピクトグラムの大きさによって表しており，大きさに用いる比率 r は次式によって求める．

$$r = \left(d \times B_{nm} + \frac{(1-d)}{N}\right) PR_n(t-1) \qquad \ldots(4)$$

ただし(3)における$d \times B_{nm}$は，隣接行列を正規化する際，出リンクが 0 のノードを 1/N に書き換えた値は含まない．(4)における$d \times B_{nm}$は，隣接行列を正規化する際，出リンクが 0 のノードを 1/N に書き換えた値のみを計算対象とする．

ノードに相当する人型ピクトグラムは PageRank の平方根の値に応じて表示の大きさが変化する．ノードから人型ピクトグラムが出ていくのに合わせてノードピクトグラムも小さくなり，別のノードから移動してきた人型ピクトグ

ラムがノードピクトグラムに入ってくるのに従って，大きくなるようにアニメーションする．

キャンバス上の要素に対してマウスオーバーすると，領域 B「行列式表示領域」の対応する部分をハイライトする．ノードピクトグラムをマウスオーバーすると，行列式におけるそのノードの PageRank の値をマークし，リンクをマウスオーバーすると，リンクを辿って移動する数量の算出に使用される値群，つまり減衰係数，対応する正規化した隣接行列の値，対応する 1/N 行列の値，及び 1 ステップ前の PageRank の値をハイライトする．

領域 A「ネットワークグラフ設定領域」にある，「隣接行列入力ボタン群」の直接入力，または「ランダム入力ボタン」によって「隣接行列入力ボタン群」の値を変更したとき，キャンバスにノードとリンクを描画し，ステップの値を 0 にする．再生中にノード数や隣接行列の入力値が変わると再生を自動で停止し，変更後のノードとリンクを描画する．

4. 実践，評価

4.1 概要

青山学院大学社会情報学部「ウェブテクノロジ」の授業で利用実践を行った．履修者は学部 2 年生以上計 60 名である．授業日は，2023 年 12 月 8 日(金曜日)である．一般的な PC 実習室が配当されている．授業は「Web 情報検索」について取り扱う回で，その中で PageRank については 20 分ほど割り当てた．PageRank に関する授業実践の概要を表 1 に示す．

表 1 PAGERANK に関する授業構成

授業構成	分
1.人型ピクサーフィングを用いながら PageRank の仕組みに関する講義	5
2. 人型ピクサーフィングを用いての自由実習	10
3. アンケート回答	5

4.2 結果

アンケートは，PageRank の理論（基本的な行列に関する知識を含む）の理解度に関する質問項目とアプリケーションのユーザビリティを評価するシステムユーザビリティスケール(System Usability Scale：SUS)による質問項目[9]，授業を評価する質問項目の 3 つを行なった．アプリケーション評価のための SUS によるアンケート項目，及び得られた回答の結果を表 2，授業評価アンケートの項目，得られた回答の結果，SUS スコアとの相関係数を表 3 に示す．授業評価アンケートにおいて，設問 1 から 8 は奇数番目が正質問，偶数番目が逆質問というペア構造になっている．ただし，設問 9，10 はどちらも正質問であり，逆質問は用意していない．回答選択肢の内容はどちらも，「5. 強く同意する」，「4. 同意する」「3. どちらとも言えない」「2. 否定する」「1. 強く否定する」とした．SUS スコアの算出では，ポジティブな設問には表 3，表 4 における平均値から 1 を引いた値に 25 を掛けた値を，ネガティブな設問には 5 から表 3，表 4 における平均値を引いた値に 25 を掛けた値を用いることになる．相関の調査には，授業評価アンケートの各設問と SUS スコアの相関係数，及び p 値を，ピアソンの相関係数を用いて求めた．また，平均値，標準偏差，相関係数，p 値は，小数第 4 位を四捨五入した．

算出した SUS スコアに対し，四分位範囲（IQR）を用いて外れ値の判定を行ったところ，外れ値の範囲に含まれる標本が 1 件あった．これを外れ値として扱い，外れ値の 1 件を除いた 59 件の標本を用いて分析を行った．したがって，表 2，表 3 のサンプル数は 59 である．

アプリケーション評価には「アプリケーションに関して改善点や気になることがあれば記入してください」とした，自由記述項目を設け，任意で回答を求めた．アプリケーションに対する自由記述項目への回答の抜粋を表 4 に示す．授業評価アンケートでは，「今日の授業の感想」という質問内容で自由記述項目を設け，任意で回答を求めた．表 5 に授業に関する自由記述項目への回答の抜粋を示す．

4.3 考察

SUS スコアの標本平均は 73.00，中央値 72.50，標準偏差 11.54 であった．標準的な UI やシステムの基準とされるスコア 68 をわずかに上回る結果となった． スコアを下げる要因となった設問として，設問 1「このアプリケーションをもっと頻繁に使いたい」，設問 4「このアプリケーションを使うには，技術的なサポートが必要だ」，設問 9「このアプリケーションを使っている間は自信がある」，設問 10「このアプリケーションを使い始めるまでには，学ぶべきことがたくさんあると思う」が挙げられる．

設問 1 の平均値が低いことから，ユーザは現状のアプリケーションに対して十分な魅力や利用価値を感じていない可能性があり，改善の余地があると考えられる．

設問 4 の評価からは，アプリケーションが直感的でなく，ユーザが問題に遭遇した際に簡単に解決することが難しい可能性があるといえる．ユーザビリティに関する自由記述の回答者 8 からもアプリケーションの目的が直感的に理解できなかったという回答が得られた．チュートリアルの提供や，直感的に操作可能な UI への改善が必要であると考えられる．

設問 9 の評価からは，ユーザのスキルや経験の影響によって，アプリケーション使用時に高い自己確信を感じられていない可能性がある．

表 2　SUS によるアンケート項目とその結果

Q	内容	1	2	3	4	5	平均値	標準偏差
1	このアプリケーションをもっと頻繁に使いたい	0	11	19	23	6	3.407	0.904
2	このアプリケーションは必要以上に複雑だと感じる	20	31	4	4	0	1.864	0.812
3	アプリケーションがシンプルで使いやすいと思う	0	0	2	24	33	4.525	0.563
4	このアプリケーションを使うには、技術的なサポートが必要だ	10	21	16	12	0	2.508	0.998
5	アプリケーションがスムーズに機能し、うまく統合されていると感じる	0	2	9	34	14	4.017	0.725
6	イレギュラーなことも多いと思う	28	26	4	1	0	1.627	0.686
7	このアプリケーションは、ほとんどの人がすぐに習得できると思う	1	1	2	31	24	4.288	0.760
8	このアプリケーションは手間がかかると思う	30	25	1	3	0	1.610	0.760
9	このアプリケーションを使っている間は自信がある	1	8	15	31	4	3.492	0.871
10	このアプリケーションを使い始めるまでには、学ぶべきことがたくさんあると思う	5	18	16	17	3	2.915	1.062

表 3　授業評価アンケートの設問とその結果

Q	内容	1	2	3	4	5	平均値	標準偏差	相関係数	p 値
1	今回の授業は楽しかった	0	1	6	41	11	4.051	0.594	0.243	0.063
2	今回の授業はつまらなかった	14	35	8	2	0	1.966	0.712	-0.343	0.008
3	今回の授業からなんらかの気づきを得られた	0	0	6	33	20	4.237	0.620	0.267	0.041
4	今回の授業から学べることはなかった	26	27	3	2	1	1.729	0.840	-0.029	0.825
5	今回の授業に対して意欲的に取り組めた	0	2	9	37	11	3.966	0.688	0.264	0.044
6	今回の授業に対して無気力であった	18	33	7	1	0	1.847	0.684	-0.258	0.048
7	今回の授業の実習手順は、スムーズに行うことができた	0	1	6	33	19	4.186	0.676	0.444	0.000
8	今回の授業の実習手順は難しかった	12	34	7	6	0	2.119	0.845	-0.058	0.326
9	ランダムサーファーモデルとは何か理解できた	0	5	18	27	9	3.678	0.832	0.326	0.012
10	PageRank とは何か理解できた	0	0	10	37	12	4.034	0.610	0.347	0.007

表 4　“アプリケーションに関して改善点や気になることがあれば記入してください”抜粋

回答者	内容
1	行列を全く知らない人は少し難しく感じるかもしれないと思ったが、少し触れたことがあるような人からしたら使いやすく感じると思った。
2	矢印の向きがどっちからどっちなのかが初見ではわからないと感じた。
3	事前知識があってこそ活きるものだと感じました。 人に置き換えるというのが直感的にわかりやすくてよかったです。
4	アニメーションが特にイメージが付きやすくわかりやすいとおもった。
5	リンク先の設定のところをどこからどこへのリンクなのかを少しわかりやすくしたらより使いやすくなると思った。
6	収束した値までスキップできる機能があればより使いやすいと感じました。
7	このアプリケーションは特別な目的がないと使わないことが考えられるから頻繁には使わないと考えた。
8	アプリケーションを開いたとき、何を目的にしたプログラムかわからなかったので記載があると理解しやすくなると個人的に感じました。
9	見やすくて、アプリも使いやすかった
10	とてもシンプルで使いやすかったです

表 5　“今日の授業の感想”抜粋

回答者	内容
1	人型ピクサーフィングを使ったことでしっかり理解できたと思います。
2	PageRank について少し理解することができた
3	実際にサイトを用いて体系的に理解することができたため、とても親切であると感じました。
4	実際に自分でアプリを使って考えたことで理解が深まったと感じた。
5	アニメーションを用いることでよりイメージがつき、理解することができました
6	ピクトグラムのサイズが変わることで、ページランクの値の大きさを視覚的に理解することができたと感じる。
7	ピクトグラムを用いたアプリケーションがわかりやすく、スムーズに理解することができた。
8	人型ピクサーフィングを使用したことで、PageRank の考え方がよくわかりました。事前に見ている講義動画や当日の授業でも理解はできましたが、身の回りにはあまりない事例であることからイメージをすることが難しかったので、授業内で実践することでより深い理解に繋げることができました。
9	私は数学を得意としていないため、図を用いたシミュレーションの理解を深めたと感じます。
10	直感的にわかりにくいものを何かに置き換えて表すというのは、モノを教えるにあたってとても有効的な手法であると改めて感じました。

本アプリケーションでは，PageRank や行列の知識と必要としており，経験や講義での理解度によってアプリケーションの使用感に差異が出ると考えられる．ユーザビリティに関する自由記述の回答者 1，回答者 3 からも事前知識が必要という趣旨の回答が得られた．成功体験など，適切なフィードバックを提供することで，アプリケーション使用時のユーザの自信を向上できると考えられる．

設問 10 の評価からは，ユーザはアプリケーションを始める際，多くの学習が必要であると感じていることが示唆される．初期利用の障壁を下げることが必要であると考えられる．ただし，本アプリケーションは学習補助であり，PageRank の理解度を向上させることが目的である．そのた

め，設問10の評価が低くなることは妥当であると考えられる．

授業評価アンケートの設問 7「今回の授業の実習手順は，スムーズに行うことができた」と，SUS スコアとの相関係数が 0.444 と中程度の相関があることから，手順のわかりやすさが全体的なシステムの使いやすさや満足度に関連している可能性がある．学習環境や教材の設計において，手順の分かりやすさを向上させる取り組みが有益であると考えられる．

授業評価アンケートの結果は，総合的には多くの学生が楽しさや意欲を感じ，気づきを得られる一方で，実習手順の評価や，ランダムサーファーモデルとPageRankの理解度には異なる傾向がみられた．

授業評価アンケートにおいては設問 8「今日の実習手順は難しかった」，設問 9「ランダムサーファーモデルとは何か理解できた」が低い評価となった．低い評価となった要因として，講義での十分な理解が必要な上に，初めて利用するアプリケーションを使用したことから，難しさを感じた可能性が考えられる．

設問 9 でのランダムサーファーモデルの理解度は設問 10 のPageRankの理解度より低い結果となった．要因として，ランダムサーファーモデルの説明が足りなかった，アプリケーションのアニメーションでは的確に視覚化できていなかったことなどが考えられる．今回の評価では，講義に問題があるか，アプリケーションに問題があるか判断が困難であるため，講義終了時点での理解度とアプリケーション使用後での理解度を記録するなど，追加実験が必要である．また，授業評価アンケートの設問 9，10 はやや抽象的であり，結果からは学習者の主観的な評価しか得られない．客観的にも正しく理解できているかどうか判断するには，理解度を問うミニテストを課すなど，更なる調査が必要である．

授業に関する自由記述項目の回答では，アプリケーションを用いたことで，理解が深まったという趣旨の回答を得られた．この結果から，被験者がアプリケーションを有益な学習ツールとして捉え，理解を深めるのに寄与している可能性が示唆される．

5. まとめと今後の展望

人型ピクトグラムを用いてランダムサーファーモデルを学習するアプリケーション「人型ピクサーフィング」を実装し，評価した．今後は，今回の評価結果をもとに，アプリケーションのUIを精錬化する．また，さまざまな授業実践を通じてアプリケーションの教育への効果を検証し，学習者に与える影響の調査を行う．ユーザーフィードバックの収集やデータ分析を通じて，アプリケーションが提供する学習体験の質を向上させ，より効果的な教育ツールへと最適化させていく予定である．

謝辞

本研究は，2023 年度 放送大学教育振興会助成金の助成を受けたものです．

参考文献

[1] 太田幸夫：国際安全標識のピクトグラムデザインの研究：http://www.tamabi.ac.jp/soumu/gai/hojo/seika/2003/kyoudou-ota1.pdf （参照 2023-12-19）

[2] 太田幸夫：FORUM8：ユニバーサル・コミュニケーションデザインの認識と実践：https://www.forum8.co.jp/topic/universal107.htm （参照 2023-12-19）

[3] 伊藤一成：複数のプログラミング言語で記述可能なピクトグラムコンテンツ作成環境の提案と実装，情報処理学会論文誌 TCE，Vol.7，No.3，pp.1-11（2021）

[4] 渡辺大智，御家雄一，伊藤一成：人型ピクトグラムを用いたソートアルゴリズムを学ぶアプリケーション「人型ピクソートグラム」の実装，2022 年度情報処理学会関西支部大会 (2022)

[5] 高橋伶奈，御家雄一，伊藤一成：人型ピクトグラムを構成素とするピクトグラフ生成アプリケーション「Human Pictograph」の実装，2022 年度情報処理学会関西支部大会 (2022)

[6] Page, L., Brin, S., Motwani, R. and Winograd, T. 「The PageRank Citation Ranking: Bring Order to the Web」「Proc. of the 7th WWW Conf」pp.161-172 (1998)

[7] Heer, J., Boyd, D.: Vizster: Visualizing online social networks, Proc. of IEEE Symposium on Information Visualization, pp.32-39(2005)

[8] Itoh, T., Klein, K.: Key-node-separated graph clustering and layout for human relationship graph visualization, IEEE Computer Graphics and Applications, Vol.35, No.6, pp.30-40 (2015)

[9] Brooke, J.; SUS: A "quick and dirty" usability scale., Usability Evaluation in Industry. London:Taylor and Francis. (1996)

講義資料作成支援のための科目用語体系の階層構造と教科書の説明順序を用いたスライド分析手法

佐野　逸稀† 　王　元元†† 　河合由起子†††,†††† 　角谷　和俊†

† 関西学院大学総合政策研究科　〒 669–1330 兵庫県三田市学園上ケ原 1 番
†† 山口大学大学院創成科学研究科　〒 755-8611 山口県宇部市常盤台２丁目１６番１号
††† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†††† 大阪大学サイバーメディアセンター　〒 567-0047 大阪府茨木市美穂ケ丘５番１号
E-mail: †{gil06518,sumiya}@kwansei.ac.jp, ††y.wang@yamaguchi-u.ac.jp, †††kawai@cc.kyoto-su.ac.jp

あらまし　講義資料を作成する講師にとって，一からスライドを作成し，問題点の確認や修正を行うことは負担が大きい．近年，生成 AI を用いたスライド自動生成サービスの登場によって，スライドを作成する負担は小さくなっているが，生成されたスライドを講義で利用する際は，事前に決められた指導内容を十分に説明しているかや，学習者にとってわかりやすいスライドになっているかなどをチェックする必要がある．しかし，スライドの内容をチェックし編集する段階をサポートするサービスは一般的ではなく，スライドの編集作業は，未だ負担が大きい作業であると考えられる．そこで本研究では，作成された講義スライドに対して，科目用語体系の階層構造と教科書の説明順序を用いた分析を行う．この分析結果を基に，スライドの編集候補を提示し，講師のスライド編集作業を支援する．これによって，講師は講義スライドの作成から編集までの一連の作業をより効率的に行うことができる．

キーワード　講義スライド，生成 AI，知識工学，EdTech

1 はじめに

講義資料の作成を行う講師にとって，一からスライドを作成し，問題点の確認や修正を行うことは負担の大きい作業であると考えれられる．スライド作成段階を支援するサービスとしては，近年一般的になってきた生成 AI を用いたスライド自動生成サービスが挙げられる．このサービスは，文章や動画などを入力することで，スライドを自動で生成してくれるという便利なサービスである．これによって，講師のスライド作成にかかる負担はかなり軽減されることが予想される．一方で，現在，スライドの問題点を確認・修正するスライド編集段階を支援するサービスは一般的ではない．前述のスライド自動生成サービスを利用する際も，サービスの中身がブラックボックスであることから，決められた指導内容を十分に説明しているかや，学習者にとってわかりやすいスライドになっているかなどを使用前にチェックする必要がある．ここから，講師の講義資料作成を支援するためには，スライドの作成段階だけでなく，編集段階も支援する必要があると考えられる．そこで，本研究は，作成するスライドが対象とする科目の用語体系と教科書の説明順序を用いて，スライド分析を行う．この分析結果を基に，スライドの編集候補を提示し，講師のスライドの編集段階を支援する．本論文の構成は次のとおりである．２章では関連研究と本論文との関連性について述べる．３章ではスライドの分析に用いる順序付き科目用語体系の構築について述べる．４章では，科目用語体系を用いたスライドの分析項目とその方法について述べ，5 章では評価実験について述べる． 最後に，６章ではまとめと今後の検討について述べる．

2 関連研究

2.1 スライド自動生成

生成 AI の登場により，SlidesGPT や MagicSlides などのスライド自動生成サービスが普及してきているが，スライドを自動生成する研究は 2000 年頃より行われてきた．例えば，Fu ら [1] は，論文を基に，Document Reader (DR)，Progress Tracker (PT)，Object Placer (OP)，Paraphraser (PAR) の 4 つのモジュールを用いたスライド生成を行う，DOC2PPT を開発している．Tsai ら [2] も，論文を基に，スライドの内容に合わせた端的な文章を生成する Document to Slide (D2S) を開発している．また，久保谷ら [3] は，D2S の発展版として，スライド間でのテキストの内容の重複を抑えたモデルを作成している．本研究では，人手で作成されたスライドに加えて，これらの手法によって自動生成されたスライドを対象としたスライド分析を行い，修正のサポートを行うことによって，講義資料の作成支援を行う．

2.2 スライド評価

本研究では，スライドの修正をサポートするために，スライドの分析・評価を行う．教育工学の分野でスライドの評価を行う研究としては，山下ら [6] は，プレゼンテーションスライドの評価に用いられている項目の整理を行っている．しかし，この研究で挙げられている評価項目は，「おもしろいか」や「わかりにくい言葉はあるか」など，人による評価が前提となっているため，システム化を行う際には，評価項目の精査とアルゴリズム化が必要である．

情報学の分野でスライド評価を行う研究は，初期段階とし

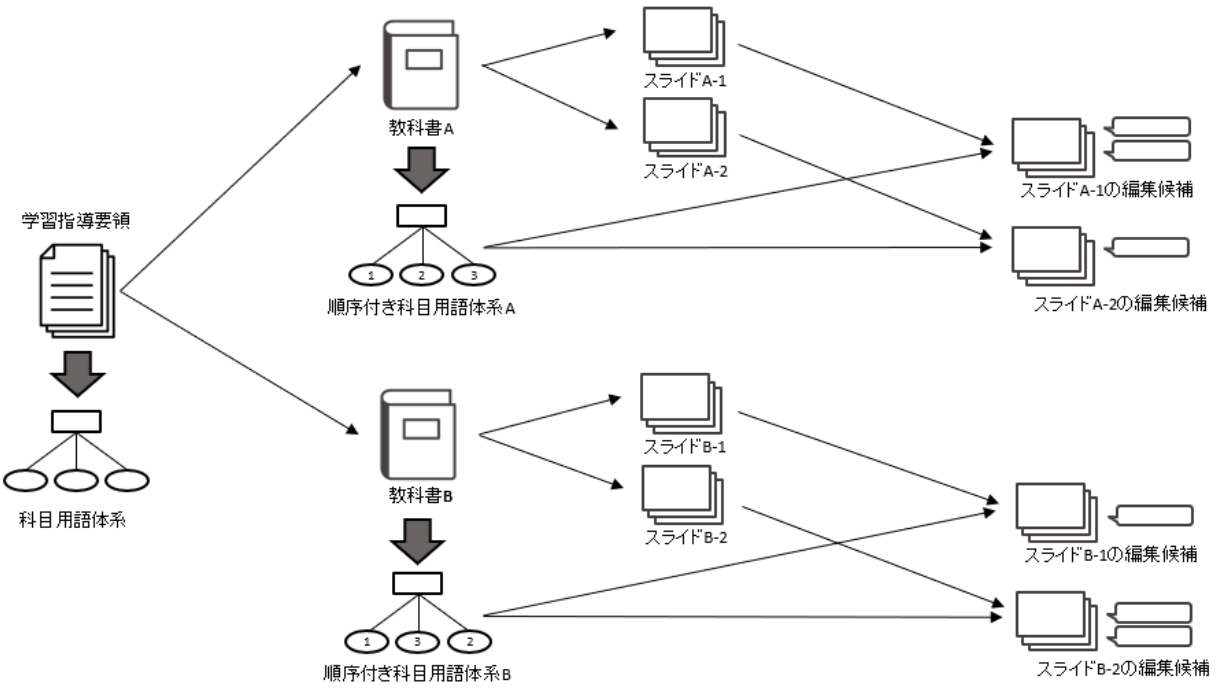


図 1　概　要　図

て，栗原ら [5] のテキスト情報に着目したスライド評価や，竹島ら [11], [12] の要約技術を用いたスライド推敲支援システムなどが挙げられる．深層学習が周流となった近年では，川瀬ら [8], [9] の文字の配色や並列関係などのスライドの要修正点を，内容情報，構成情報，画像情報の 3 つを特徴とした機械学習で指摘するモデルなどがある．また，大山ら [10] は，文章量や，表や図の量，フォントサイズなどからスライドの印象を予測するモデルを構築している．本研究では，スライドの分析を行い，修正点を指摘するだけでなく，修正候補を提案するという点でこれらの研究とは異なる．また，本研究は，具体的な修正候補を提案するために，その科目で使われている用語を体系的に整理した科目用語体系と教科書の説明順序を用いる．

3　順序付き科目用語体系の構築

3.1　順序付き科目用語体系

赤澤ら [7] は，高等学校共通教科情報科の知識体系の構築を行っている．本研究では，赤澤らの構築している知識体系を参考に用語抽出を行い，さらに用語間の概念関係を追加したものを科目用語体系として構築する．さらに，図 1 に示すように．科目用語体系に教科書内のトピックの説明順序を反映させたものを，教科書の順序付き科目用語体系とする．本研究では，生成 AI が生成したスライドを対象に，教科書の順序付き科目用語体系を用いて，スライドの分析を行い，スライド編集候補を提案する．

3.2　用語の抽出と概念関係

用語の抽出には，教科書の索引を用いる．また，索引語が教科書のどの章に出現しているかによって，その索引語が所属するトピックを紐づける．同トピックの索引語に対して，総当たりで用語間の概念関係を決定する．本稿で用いた用語間の概念関係は，上位・下位を表す「is-a」関係，部分・全体を表す「part-of」関係，クラス・インスタンスを表す「instance-of」関係，同義を表す「same-as」関係の 4 つと，「関係なし」である．今回は 2 つの教科書 [1][2] の「ソート」に関する章を用いて，重複を除いた 41 語の索引語を抽出した．これらの索引語を用いて，820 通りの用語ペアを作成したところ，関係があると考えられた用語ペアは 66 ペアであった．

このうち，「is-a」関係を持つのは，「クイックソート is-a 大きなデータの整列」などの 40 ペアであった．「part-of」関係を持つのは，「バケット part-of バケツソート」などの 13 ペア，「same-as」関係を持つのは，「ボゴソート same-as ショットガンソート」などの 21 ペアであった．また，索引に出現する単語の中にはインスタンスになり得る単語が少なく，今回「instance-of」関係となるペアはなかった．このように作成した科目用語体系を図 2 に示す．

3.3　教科書の説明順序の反映

同じトピックを扱っているとしても，教科書によって用語を説明する順序は異なる．本研究では，各講義ごとにメインで取り扱う教科書があると考え，その教科書の説明順序がスライドにも反映されている状態を望ましいと考える．そこで，科目用語体系に教科書の順序を反映させたものを，順序付き科目用語

1：川井明・梅津高朗・高柳昌芳・市川治（2018）『データ構造とアルゴリズム』学術図書出版社
2：茨木俊秀（1989）『アルゴリズムとデータ構造』昭晃堂

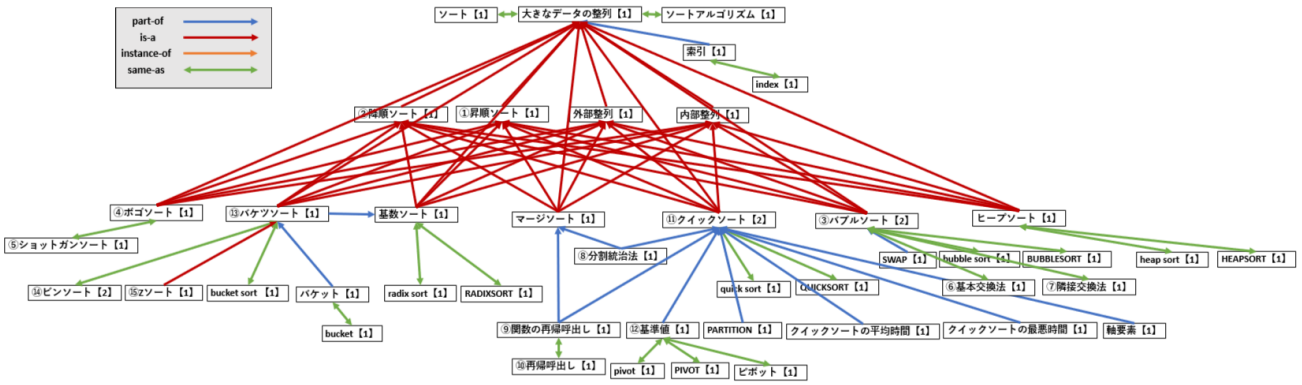


図 2　科目用語体系

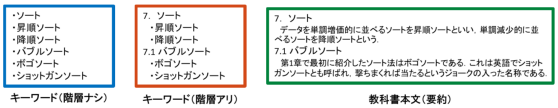


図 3　プロンプトの例

体系とする．3.2 章で作成した科目用語体系に，教科書「データ構造とアルゴリズム」の出現順序を反映させたものを「データ構造とアルゴリズム」の順序付き科目用語体系とする．同様に，教科書「アルゴリズムとデータ構造」の出現順序を反映させたものを「アルゴリズムとデータ構造」の順序付き科目用語体系とする．

4　スライド分析手法

4.1　分析に用いるスライドデータ

本研究では，分析する対象データとして，生成 AI を用いて生成したスライドデータを使用する．生成 AI を用いたスライド自動生成サービスは数多くあるが，本稿では Elucile を利用して，教科書「アルゴリズムとデータ構造」を入力とした分析用のデータを作成する．また，入力するプロンプトによって生成スライドの精度に差が出ることを考慮して，「キーワードリスト (階層ナシ)」「階層付きキーワードリスト (階層アリ)」「教科書本文 (要約)」の 3 種類をプロンプトを作成する．この際，スライド生成サービスの入力の文字制限によって，「教科書本文」は一部要約している．図 3 は入力に使用したプロンプトの一例である．

4.2　スライド分析

生成されたスライドデータを主観で分析し，8 つの分析項目を作成した．作成した分析項目を分析対象，対象範囲，検知方法，支援方法にまとめて図 4 に示す．

4.2.1　分 析 項 目

本稿では，図 4 で示す 8 項目を用いてスライド分析を行い，編集支援を行う．

a）　スライドに適切な文章量

「スライドに適切な文章量」は，ユーザが行数 n を指定し，箇条書きの 1 つの条項が n 行を超えるスライドがあった場合に，過剰な文章量であるとして検知する．

b）　文末表現の統一性

「文末表現の統一性」は，ユーザが文末表現（常体/敬体）を指定し，文末表現が指定された表現ではない場合に，文末表現の乱れとして検知する．ただし，文が名詞形で終わっている場合はどちらにも該当するものとし，検知しない．

c）　表 記 ゆ れ

「表記ゆれ」は，科目用語体系で「same-as」関係にある単語のうち，スライド内で使用されている単語が複数ある場合に，表記ゆれとして検知する．ただし，「バブルソート（bubble sort)」のように同じ文で出てきている場合は検知しない．

d）　箇条書きの表記方法の適切性

「箇条書きの表記方法の適切性」は，説明順序のない単語が番号付き段落で表記されている場合（1. 昇順ソート，2. 降順ソートなど）に，不適切な箇条書き表記として検知する．単語の説明順序については，科目用語体系に今後追加予定である．

e）　単語の並列関係の適切性

「単語の並列関係の適切性」は，スライド 1 枚の中で並列に説明されている単語が意味的にも並列かどうかをチェックする項目である．具体的には，まず，スライド一枚の中で並列関係にある単語を抽出する．並列かどうかの判定には，スライドのインデント構造を用いて，同インデントで説明されている単語が並列であるとする [4]．次に，これらの単語を科目用語体系で検索し，スライド内で並列な単語が科目用語体系でも並列関係であるかをチェックする．ここで並列関係ではなかった場合に，不適切な並列関係として検知する．

f）　単語の上下関係の明示性

「単語の上下関係の明示性」とは，ある単語の上位下位関係がスライドの中で明確に示されているかをチェックする項目である．具体的には．まず，スライド全体に含まれている単語の上下関係を科目用語体系から抽出する．次に，科目用語体系内で上下関係にある単語のペアが，スライド内でも上下関係のある単語のペアとして表現されているかをインデントを用いてチェックする．ここで上下関係が明示されていなかった場合は，明示されていない上下関係として検知する．

分析項目	分析対象	分析範囲	支援方法
スライドに適切な文章量	style	1枚	提案
文末表現の統一性（常体/敬体）	style	全体	自動修正
表記ゆれ	style	全体	自動修正
箇条書きの表記方法の適切性	style	1枚	自動修正
スライド1枚中の単語の並列関係の適切性	contens	1枚	提案
スライドコンテンツにおける単語の上下関係の明示性	contens	全体	提案
単語の下位概念の網羅性	contens	全体	提案
学習順序の準拠	contens	全体	提案

図 4　編集支援のためのスライド分析項目

g）　単語の下位概念の網羅性

「単語の下位概念の網羅性」は，スライド内のある単語についての説明が不足なく行われているかをチェックする項目である．具体的には，まず，スライドから任意の 1 単語を抽出する．次に，科目用語体系から，その単語の下位単語を抽出する．この際，抽出する下位単語はスライド作成に使用している教科書に出現する単語のみとする．次に，スライド内でその下位単語が出現しているかを検索する．ここで漏れがあった場合は，説明されていない下位単語として検知する．

h）　学習順序の準拠

「学習順序の準拠」は，ある単語の説明時にまだ学習していないトピックの単語が用いられていないかをチェックする項目である．ここでの学習順序には，スライド作成に使用している教科書の章割りを用いる．例えば，2 章を学習しているとき，1 章に出現する単語は学習済み，3 章以降に出現する単語は未学習であるとする．具体的には，スライドに出現する単語を，科目用語体系内で検索し，出現する章を特定する，特定された章が学習中の章よりも後の場合は，学習順序の非準拠として検知する．

4.2.2　分 析 内 容

分析項目は，大きく「style」と「contents」の 2 つに分けられる．「style」とはスライドの構造に関する項目であり，「contents」とはスライドの内容に関する項目である．上記 8 つの 8 項目では，「スライドに適切な文章量」，「文末表現の統一性 (常体/敬体)」，「表記ゆれ」，「箇条書きの表記方法の適切性」がスライドの構造について指摘であるため style，「単語の並列関係の適切性」，「単語の上下関係の明示性」，「単語の下位概念の網羅性」，「学習順序の準拠」がスライドの内容についての指摘であるため cotents に分類される．

4.2.3　対 象 範 囲

対象範囲は，分析の対象となるスライドの範囲を指し，「1 枚」と「全体」の 2 つに分けられる．上記 8 つの項目では．分析がスライド 1 枚で完結する「スライドに適切な文章量」，「箇条書きの表記方法の適切性」，「単語の並列関係の適切性」が 1 枚，分析するためにはスライド全体を対象とする必要がある「文末表現の統一性」，「表記ゆれ」，「単語の上下関係の明示性」，「単語の下位概念の網羅性」，「学習順序の準拠」が全体に分類される．

4.2.4　支 援 方 法

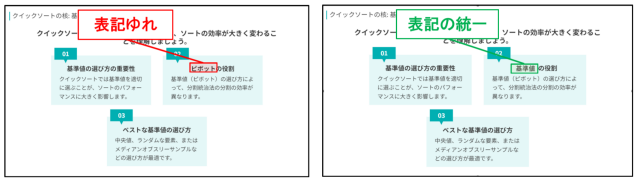
支援方法は，「自動修正」と「提案」の 2 種類である．「自動修正」は，ある項目について問題があった場合に，自動修正することでスライド編集支援を行う項目である．スライドとして望ましい形が一つに決めることができる項目に対して，「自動修正」の支援方法を適用する．上記 8 つの項目では，「文末表現の統一性」，「表記ゆれ」，「箇条書きの表記方法の適切性」の 3 つの項目は，修正方法が 1 つしかないため，「自動修正」に当たる．「提案」とは，ある項目について問題があった場合に，修正候補をスライド作成者に提示することでスライド編集支援を行う項目である．スライドとして望ましい形が一つではなく，スライド作成者の考えを反映させるのが良いと考えられる項目に対して，「提案」の支援方法を適用する．上記 8 つの項目では，「スライドに適切な文章量」，「単語の並列関係の適切性」，「単語の上下関係の明示性」，「単語の下位概念の網羅性」，「学習順序の準拠」の 5 つの項目は，修正方法がいくつか考えられる．例えば，「スライドに適切な文章量」では，ユーザが指定した適切な文章量にするためには，文章の要約や分割など複数のアプローチがある．このような項目は「提案」とする．

4.3　分析結果に基づくスライド編集支援

本研究では，上記の項目を用いて，作成されたスライドを分析し，編集支援を行う．本稿では，編集支援の具体例を 2 つ提示し，他の項目に対する支援方法については今後の課題とする．

4.3.1　表 記 ゆ れ

まず，分析項目「表記ゆれ」に対するスライド編集支援を述べる．「表記ゆれ」は，分析内容は style，対象範囲は全体，支援方法は自動修正の項目である．図 5(左) に「表記ゆれ」が検知されたスライドを示す．科目用語体系では，「基準値 same-as ピボット」であるため，スライド内の「ピボットの役割」が表記ゆれとなる．この場合は，出現回数が多い「基準値」に統一することで修正が可能である．図 5(右) に自動修正後のスライドを示す．

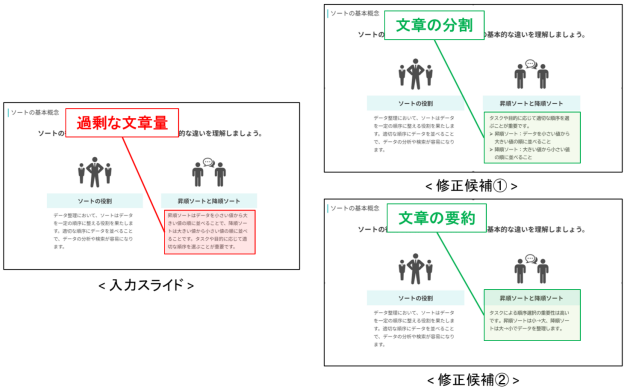


＜入力スライド＞　＜自動修正＞

図 5　表記ゆれに対するスライド編集支援

4.3.2　スライドに適切な文章量

次に，分析項目「スライドに適切な文章量」に対するスライド編集支援について述べる．「スライドに適切な文章量」は，分析内容な style，対象範囲は 1 枚，支援方法は提案の項目である．図 6(左) に「過剰な文章量」が検知されたスライドを示す．なお，ここではユーザーが 3 行を上限に設定した場合を想定している．この場合は，過剰な文章を分割したスライドや，要約したスライドを修正候補として提示し，ユーザに選択してもらう．図 6(右) に修正候補のスライドを示す．



＜入力スライド＞　＜修正候補①＞　＜修正候補②＞

図 6　過剰な文章量に対するスライド編集支援

5　評価実験

5.1　実験方法

本章では，提案するスライド分析手法を用いることで，どの程度スライドの修正箇所を指摘可能であるかを評価する実験を行う．まず，教科書「データ構造とアルゴリズム」からランダムに章を一つ選択し，スライド自動生成サービスを用いてスライドを生成する．次に提案手法を用いて生成されたスライドの分析を行い，修正箇所の指摘を行う．そして，スライドと修正箇所を被験者に提示し，修正箇所の妥当性について 5 段階評価を行ってもらう．また，提案システムでは指摘出来ていない修正箇所を評価するために，被験者には指摘されている修正箇所以外に修正が必要な箇所はないかをチェックしてもらう．また，本評価実験には，生成内容への理解が必要であるため，被験者は生成に使用した「データ構造とアルゴリズム」を学習した経験のある大学生，大学院生 3 名とした．

5.2　予備実験

まず，教科書からスライドを生成する際に，どのような入力方法を用いるのが良いかについての予備実験を行った．教科書「データ構造とアルゴリズム」の 7 章「ソート」を用いて，キーワードリスト（階層ナシ），キーワードリスト（階層アリ），文章の 3 種類の入力方法でスライドを生成した．なお，今回使用したスライド自動生成サービス（Elucile）は入力に文字制限があり，3 つ目の入力方法である文章において，教科書本文全てを入力することができなかったため，他 2 種類と同様の単語が含まれるように本文を一部要約して入力を作成した．同様の入力を用いて，各入力方法 5 回ずつスライド生成を行い，スライド枚数，入力に対する説明漏れの単語数，入力に対する入れ替わりの単語数の平均を算出した．なお，入力に対する入れ替わりについては，科目用語体系内で「is-a」か「part-of」関係にある計 41 ペアを対象に値の算出を行う．使用した入力文と結果を図 7 に示す．スライド枚数については，文章を入力に用いたスライドが少なくなったものの，著しく枚数が少ないスライドなどが生成されることはなかった．また，入力に対する説明漏れの単語数については，文章を入力としたスライドが他の 2 種類の入力方法で生成したスライドと比べて 3 倍以上多いことが分かった．次に，入力の単語順序に対する入れ替わりについては，多少の差はあるものの，それぞれ平均 1 ペア以下の入れ替わりであり，どの入力方法であっても特に問題はないことが確認された．結果として，2 つの項目において，精度が高かったのはキーワードリスト（階層アリ）であったため，評価実験ではキーワードリスト（階層アリ）を入力とした生成スライドを用いることとする．

5.3　修正指摘箇所の評価

まず，修正指摘箇所の評価の手順について説明する．「ソートアルゴリズムのキーワードリスト（階層アリ）」を入力として，生成した 5 つのスライド群に対して，システムにより検知された修正箇所をスライドに示した．図 8(右) では，赤枠で囲んでいる箇所が修正箇所であり，スライド上部にはどのような修正箇所なのかが書かれている．なお，「スライドに適切な文章量」は，$n = 3$ とし，3 行を超える文章を検知するものとした．図 8(左) は，「単語の上下関係の明示性」と「単語の下位概念の網羅性」を示したスライドである．この 2 項目は，スライドに書かれていないということを指摘する項目であるため，生成されたスライド内で指摘することができない．そこで，生成されたスライドの末尾に 2 枚のスライドを追加して示した．スライドで指摘された修正箇所としては，「スライドに適切な文章量」が 35 件，「文末表現の統一性」が 9 件，「表記ゆれ」が 4 件，「箇条書き表記方法の適切性」が 4 件，「スライド内の並列関係の適切性」が 4 件，「単語の上下関係の明示性」が 82 件，「単語の下位概念の網羅性」が 1 件，「学習順序の準拠」が 0 件出現していた．被験者には，修正箇所を示したスライドと教科書を提示し，修正箇所一つ一つに対して「5. 修正すべき」，「4. どちらかと言えば修正すべき」，「3. どちらでもない」，「2. どちらかと言えば修正すべきではない」，「1. 修正すべきではない」の 5 段階で評価を行ってもらった．

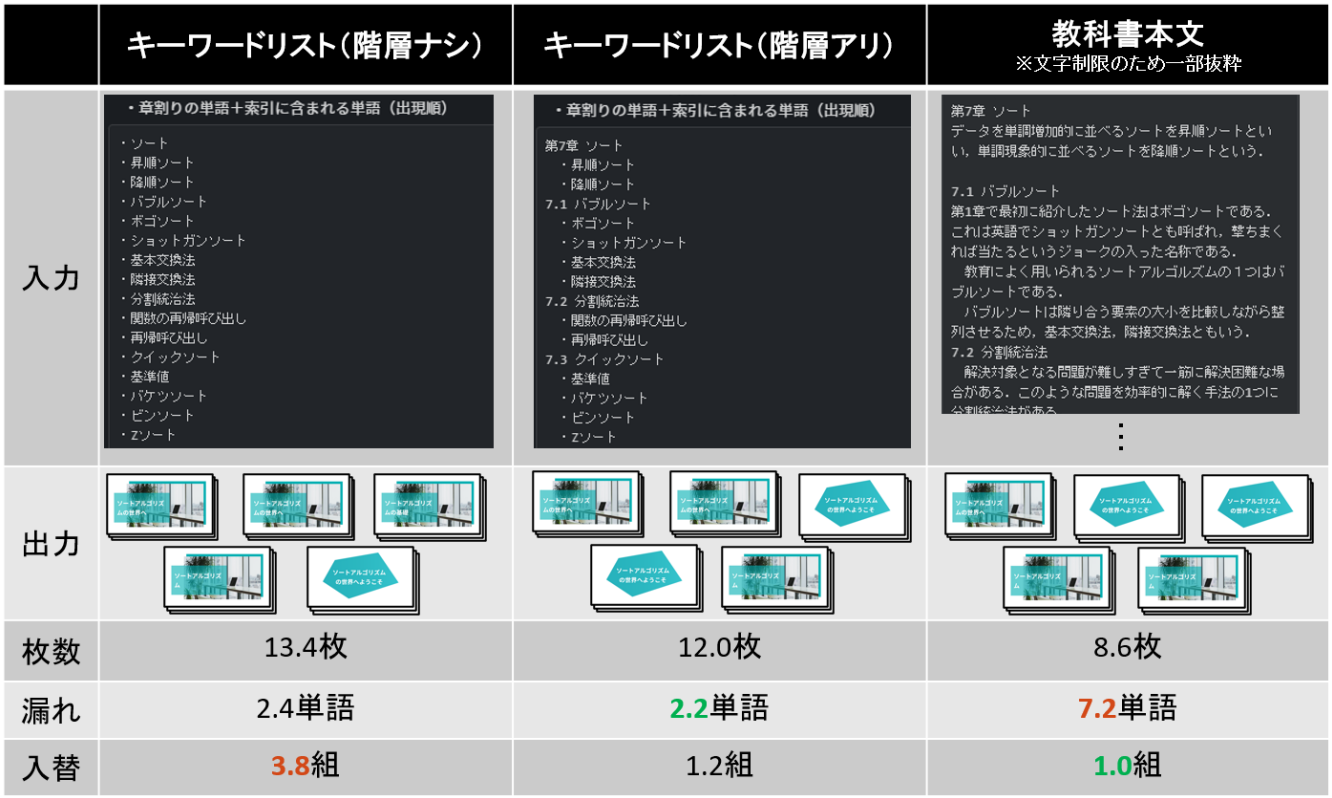


図 7　入力方法による生成スライドの特徴の違い

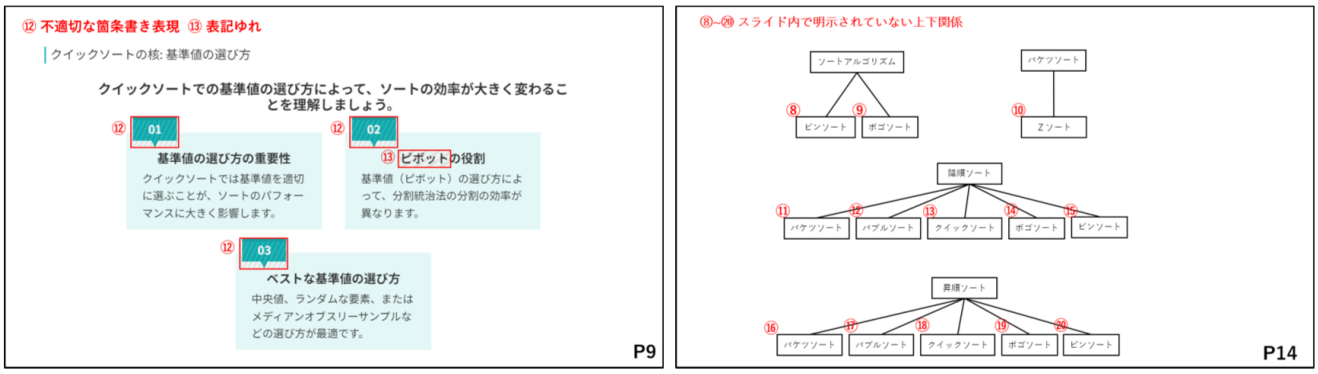


図 8　修正箇所を示したスライドの一例

5.4　システムで指摘出来ていない修正箇所の評価

次に，提案するシステムで指摘出来ていない修正箇所についての評価手順について説明する．5.3 節と同様のスライドを使用し，他に修正した方がいいと思う箇所とその理由を自由にスライドに書き込んでもらった．図 9 に，指摘出来ていない修正箇所を書き込んでもらったスライドの一例を示す．

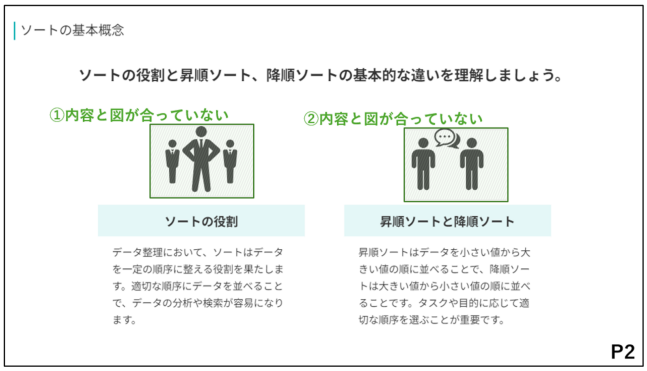


図 9　システムで指摘出来ていない修正箇所の例

5.5　結果と考察

5.5.1　修正指摘箇所の評価

表 1 は，各分析項目の出現件数と被験者 3 名の評価値の平均である．

出現があったすべての項目において，評価値が 3.5 を上回っており，肯定的な評価が多かった．一方で，「1. 修正すべきではない」，「2. どちらかと言えば修正すべきではない」が選択されていた項目は，「スライドに適切な文章量」で 35 件中 5 件，「スライド内の並列関係の適切性」で 4 件中 3 件，「単語の上下

表 1　Experimental Results

分析項目	件数	評価値 (平均)
適切な文章量	35	3.43
文末表現の統一性	9	3.93
表記ゆれ	4	4.67
箇条書きの適切性	4	3.83
並列関係の適切性	4	3.25
上下関係の明示性	82	3.74
単語の下位概念の網羅性	1	4.33
教科書学習順序の準拠	0	-

関係の明示性」で 82 件中 51 件あった．以下にそれらの項目について，低評価の理由と改善案を考察する．

a） スライドに適切な文章量

「スライドに適切な文章量」ではテキストボックスの大きさに関係なく 3 行としていたため，実際には文章量があまり多くない箇所も検知してしまっていたことがネガティブな評価につながったと考えられる．そのため，行数だけではなく文字数も考慮した分析項目に変更することを検討する．

b） スライド内の並列関係の適切性

「スライド内の並列関係の適切性」では，「クイックソートのパフォーマンス」と「再帰的な分割統治法」という見出しが並列に書かれたスライドが検知されてる．この 2 つの見出しはどちらもクイックソートの一部について説明しているものであるが，「クイックソートのパフォーマンス」という見出しであっても出現語がクイックソートであることから，「分割統治法」よりも階層が上であると判定されてしまう．したがって，「の」などの所有を表す助詞が付いている場合は，階層が一つ下がっていると判定するといった処理が必要であると考えられる．

c） 単語の上下関係の明示性

「単語の上下関係の明示性」では，既に一つ以上の単語との上下関係が明示されている単語については，他の単語との上下関係は示されていなくても問題ないという意見があった．そこで，「どの単語とも上下関係が明示されていない」場合と「1 つ以上の単語との上下関係が明示されている」を区別して検知するなどの変更が考えられる．

また，今後の課題として，本実験では，「単語の下位概念の網羅性」と「教科書学習順序の準拠」の出現件数が 0-1 件と少なかったため，スライド量を増やした実験を行うことが挙げられる．

5.5.2　システムで指摘出来ていない修正箇所の指摘

システムで指摘出来ていない修正箇所の指摘では，「図と説明内容が合っていない」や「図が表示されていない」などの図に関する指摘が 32 件，「文の途中で改行されていて読みづらい」という改行に関する指摘が 37 件，「”ショットガンソートは銃で撃ち続ける方法”という文の意味が分からない」や「過程が示されていないのでバブルソートを理解しづらい」など説明方法についての指摘が 5 件あった．したがって，現在の項目に追加して，「図と内容の一致度」や「改行位置の適切性」を追加することで精度が高くなることが予測される．また，「図と内容の一致度」と「説明方法に関する項目」については，科目用語体系だけでなく，機械学習を用いたアプローチが必要である．

6　おわりに

本研究では，スライドの編集支援を目的とし，教科書の説明順序付き科目用語体系の構築と，スライド分析項目の作成，編集支援方法の提案を行った．また，本研究で提案するスライド分析手法が，どの程度スライドの修正箇所を指摘可能であるかの評価実験を行った．実験結果として，科目用語体系を用いることによって指摘された修正箇所は，概ねポジティブな評価であったが，指摘漏れが多く存在することがわかった．したがって，現在の分析項目の精度を上げると同時に，機械学習を用いたアプローチなどによって，指摘できる修正箇所を増やすことが必要である．また，本稿では修正箇所の指摘を行う手法を述べたが，提示する編集候補の生成とその有用性に関する実験も実施する．

文　献

[1] Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. Doc2ppt: automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 634–642, 2022.

[2] Flora S Tsai and Yi Zhang. D2s: document-to-sentence framework for novelty detection. *Knowledge and information systems*, Vol. 29, pp. 419–433, 2011.

[3] 善記久保谷, 政太朗品川, 晋太郎山本, 亮太鈴木, 森島繁生. 論文要旨を捉えたスライド生成における割り当て問題の適用. 言語処理学会第 28 回年次大会 発表論文集, pp. 1029–1034, mar 2022.

[4] 桐原牧紀, 王元元, 河合由起子, 角谷和俊. e-learning における講義コンテンツの階層構造に基づくスライド推薦方式の提案. In *DEIM Forum*, 2021.

[5] 栗原一貴, 加藤公一, 大浦弘樹. Slidechecker: プレゼンテーション資料の基礎的な定量的自動評価手法. WISS 第 17 回インタラクティブシステムとソフトウェアに関するワークショップ論文集, pp. 89–94, 2009.

[6] 山下祐一郎. プレゼンテーションにおけるスライド評価と発表評価の一致率. 日本教育工学会論文誌, Vol. 44, No. Suppl., pp. 5–8, 2021.

[7] 赤澤紀子, 赤池英夫, 柴田雄登, 山根一朗, 角田博保, 中山泰一ほか. 高等学校共通教科情報科の知識体系に関する一考察. 情報処理学会論文誌 教育とコンピュータ (TCE), Vol. 8, No. 3, pp. 19–34, 2022.

[8] 川瀬卓也, 大平茂輝, 長尾確. 機械学習を用いた要修正なスライドの自動発見. 第 82 回全国大会講演論文集, Vol. 2020, No. 1, pp. 519–520, 2020.

[9] 川瀬卓也, 東中竜一郎, 大平茂輝, 長尾確. 発表スキル向上のためのスライド内の要素に対する指摘内容生成モデル. 第 83 回全国大会講演論文集, Vol. 2021, No. 1, pp. 495–496, 2021.

[10] 大山真司, 山崎俊彦, 相澤清晴. プレゼンテーションスライドの客観評価と印象予測. 第 16 回 FIT 講演論文集, Vol. 3, pp. 45–52, 2017.

[11] 竹島亮, 大平茂輝, 長尾確. インタラクティブな要約と内部構造の可視化によるスライド推敲支援システム. 第 75 回全国大会講演論文集, Vol. 2013, No. 1, pp. 641–642, 2013.

[12] 竹島亮, 大平茂輝, 長尾確. 内容の印象度に基づくインタラクティブな要約を用いたスライド推敲支援システム. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2013.

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

2024年2月28日(水) 10:00 ~ 12:10 | T5-B オンライン（Zoom Events)

観光・行動

座長:大野 成義(職業能力開発総合大学校)
コメンテータ:熊本 忠彦(千葉工業大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T5-B-1-01] 旅行時の対応能力向上のための大規模言語モデルを用いたトラブル当てゲームの評価

*斉藤 さつき[1]、堀川 達平[1]、北山 大輔[1] (1. 工学院大学)

10:30 ~ 10:55
[T5-B-1-02] ジオタグつき投稿を用いたコロナ禍の影響の地域差の分析

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

10:55 ~ 11:20
[T5-B-1-03] 移動軌跡とそれに対応づく地理空間データを活用した人の移動特性の理解

*須藤 龍一[1]、戸田 浩之[1] (1. 横浜市立大学)

11:20 ~ 11:45
[T5-B-1-04] 機械学習を用いた旅行活動ツイート判定精度の可視化視認性への影響

*藤大 友都[1]、服部 峻[2]、砂山 渡[2] (1. 滋賀県立大学 知能情報工学分野研究室、2. 滋賀県立大学)

11:45 ~ 12:10
[T5-B-1-05] 近接関係ネットワークに影響を与える歩行者行動の可視化

*森越 彩楓[1]、丹羽 了[2,3]、鷹見 竣希[3]、大西 正輝[3]、伊藤 貴之[1] (1. お茶の水女子大学、2. 筑波大学、3. 産業技術総合研究所)

旅行時の対応能力向上のための大規模言語モデルを用いたトラブル当てゲームの評価

斉藤さつき† 堀川 達平†† 北山 大輔†

† 工学院大学情報学部システム数理学科 〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学大学院工学研究科情報学専攻 〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{jx20104,em23040}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし 旅行にはトラブルがつきものだが，旅行慣れしていない人は予期せぬトラブルに対処することが難しいことがある．トラブルが発生した場合，旅行が予想以上にストレスになり，楽しむことができなくなってしまう．このような事態を避けるために，我々は旅行前に予防策を講じ，実際にトラブルが起こった際に適切な対応を行うことができるようにすることが重要だと考えた．本研究では，大規模言語モデルがゲームマスターを務め，旅行計画に対してユーザが考えられるトラブルとそのトラブルの対処法を考えることで，トラブルの対応能力を向上させるトラブル当てゲームを実装する．すでに我々は大規模言語モデルを用いてトラブルやその対処の適切さの判定を行う手法に取り組んでおり，本稿では観光レビューコーパスを用いて重要なトラブルの要素を評価する手法を提案する．また，これらを実装したプロトタイプシステムを用いて，ゲームによる能力向上効果について評価した．その結果，ヒントを用いたグループでは回答するトラブルが多様化，具体化する可能性が示唆された．

キーワード 観光，ゲーミフィケーション，GPT，低評価レビュー応用

1 はじめに

私たちは旅行に行き新しい文化や風景に触れると，新鮮な刺激を受け，充実感を含め魅力的な経験を得られる．しかし，旅行に慣れていない人にとっては予期せぬトラブルに遭遇することもある．たとえば，予約したホテルの質が悪かったり，飛行機の遅延やキャンセルによって予定が変更になるなどのトラブルがある．このようにトラブルが発生した場合，心配やストレスが増大し，旅行を楽しむことが難しくなることが考えられる．このような時のために，旅行前に予防策や対応策を講じ，実際にトラブルが起こった際に，適切な対処法を考えられるようにすることが重要であるが，旅行経験が少ない人にとって，旅先でのトラブルの具体的なイメージを描くことは容易ではない．なぜならトラブルとは，予想外の出来事であり，その性質上予測することが難しい偶発的事態だからである．

そこで，本研究ではゲーミフィケーションの考え方を取り入れ，提示される旅行計画に対して発生しうるトラブルとその対処法をユーザが考えることで，トラブルの対応能力を向上させるトラブル当てゲームを提案する．すでに我々はトラブル自体の妥当性や，その対処法の妥当性に関しては大規模言語モデルを用いた判定に取り組んでいる [1]．本稿ではさらに，ユーザに新しい視点を示すために，低評価レビューを用いて重要なトラブルに関するフレーズをヒントとして提示する．旅行をより快適なものにするために，このゲームを通じて，楽しみながら対応能力を向上させることを目指す．

この先行研究 [1] では，ユーザが入力を考えるとき，判断材料がユーザ自身の考えや大規模言語モデルによるフィードバック文が主であり，ユーザと異なるアプローチや考え方を提供する仕組みが不足していた．本稿ではレビューからネガティブなキーフレーズを抽出し，そのキーフレーズをヒントとして提示する機能を追加した．実験によりヒント提示機能の有効性を検証するとともに，システムの妥当性について検討した．

2 関連研究

2.1 ゲーミフィケーションによる観光支援

近年，ゲーミフィケーションが注目されており，さまざまな分野で活用されている [2] [3]．ゲーミフィケーションとは，ゲームのデザイン・システムなどの要素を非ゲーム的文脈に導入することである [4]．ゲーミフィケーションの特徴は，目標達成のためにゲームの仕組みを活用することと参加者の動機づけに働きかけることである [5]．長所として，学習活動への意欲を高めやすいこと，上達の努力を続けやすいこと，試行や失敗から学ぶ環境を作りやすいことなどが挙げられている [6]．本研究では，努力のモチベーションを向上させ，本来面倒な経験を楽しみながら学ぶことができるようにするためゲーミフィケーションをシステムに取り入れている．

観光に関するゲーミフィケーションの先行研究としては，倉田らの観光周遊支援ゲームが挙げられる [7] [8]．これからの観光周遊支援ゲームの企画設計において考慮すべき要素について議論しており，観光をゲーム化するメリットを示唆している．具体的には，ゲーミフィケーションの魅力は，面倒なことを楽しみへと変えられる可能性であると述べている．さらに，古賀らは，利用者が効率的かつ快適に旅行計画を作成することを支援するために，ゲーミフィケーションを取り入れた観光プランの作成システムを構築している [9]．これらの研究では，旅行計画作成の支援にゲーミフィケーションを組み合わせているが，

本研究では旅行で起こりうるトラブルに着目し，その学習を目的としている点で異なる．

2.2 ネガティブレビュー抽出とその応用

ネガティブレビューを取り扱う先行研究として，吉田らのネガティブレビューの抽出に関する研究が挙げられる [10] [11]．ネガティブな情報を効率的に入手できる可能性を創出するために，Twitter の位置情報付きツイートを用いて，機械学習により失敗談を抽出している．SVM と BERT を用いており，「したほうが」，「れば良かっ」，「やっちゃっ」などの言語特徴が含まれていると失敗談として分類されやすいと推測している．著者らは文書全体に類似するようなキーフレーズを出力できるように EmbedRank を用いて抽出を行った [12]．

また，吉田らが ML-ASK を利用して不満コーパスに感情分析を行ったところ，「None」と判定されることがもっとも多く，さらに「Positive」と判定されるものもあり精度が高いとは言えない結果になった．「してほしい」のような願望を示す言葉が含まれているとポジティブと判定しているということが分かった．このようにネガティブな文章であっても含まれる単語によってはポジティブと誤判定されてしまうことから，著者らはレビューを評価値で分け，比較を行うことによりネガティブなキーフレーズを抽出することとした．

2.3 大規模言語モデルを用いた分類と応用

大規模言語モデルを用いる先行研究として，鈴木の教務事項問い合わせ対応システムが挙げられる [13]．根拠資料として与えた教務文書に基づいて，学生からの教務事項に関する質問に対して ChatGPT に回答を生成させるシステムの実用性について検討しており，複数の想定質問に対して大学独自の教務規定に基づいた妥当な回答が得られている．また，上野らは BERT を利用した支援記録の分類手法を提案しており，支援記録の記録本文とその付随情報を入力として，予兆の有無を示すラベルを用いた 2 クラス分類により予兆を検出している．分類器の予兆検出性能を検証する実験では，予兆検出の一定の有効性が示された．「しんどそう」のような，ニュアンスにより予兆かどうかが異なる表現を持つ記録の分類が難しいことが分かり，これは BERT が単語単位で学習しているため，ニュアンスの違いなどの文や文章の単位の違いを学習することが難しかったためだと考察している．そこで，著者らはユーザの回答の妥当性と具体性の判定に大規模言語モデルを用いた．

3 旅行トラブルの対応力育成ゲーム

3.1 ゲームの概要

本研究では，トラブル当てゲームを設計，構築し，このシステムの実装および評価を行う．本システムの概要を図 1 に示す．

本システムは大規模言語モデルとして OpenAI が提供する GPT(GPT-3.5-turbo-1106)[1]を用いて対話形式のゲームを設計する．まず，ユーザに旅行計画とヒントが提示され，ユーザはその計画に対して想定されるトラブルとその対処法を回答する．各回答は GPT を用いて具体性と妥当性を判定する．

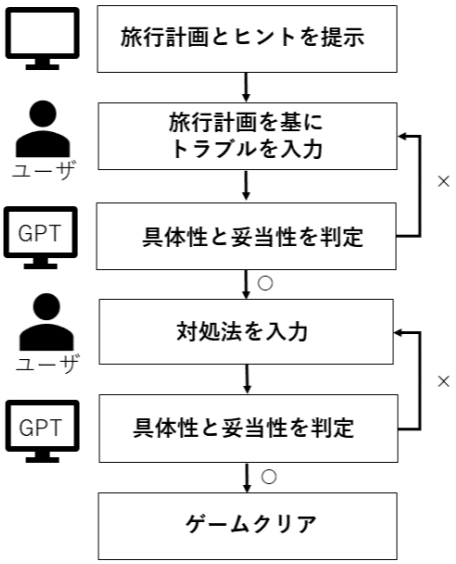


図 1　システム概要

次にゲームの流れについて説明する．

（1） 旅行計画とヒントが提示される．

（2） ユーザがその旅行計画に対して考えられるトラブルを 3 つ答える．

（3） 入力されたトラブルに対して，GPT が具体性と妥当性を判定する．

（4） システムが GPT 判定結果をチェックし 3 つともクリアしていればユーザは 5 の処理を行う．クリアしていなければ 2 の処理に戻る．

（5） ユーザは 2 で回答したトラブルの，対処法を答える．

（6） 入力されたトラブルの対処法に対して，GPT が具体性と妥当性を判定する．

（7） システムが GPT 判定結果をチェックし 3 つともクリアしていればゲームクリアとなる．クリアしていなければ 5 の処理に戻る．

また，開始から 5 分経つと入力できなくなり終了ページへと促され，ゲーム終了となる．

旅行計画の上部にヒントを提示する．ヒントは，計画に記載されているスポットに関係した低評価レビューから抽出したキーフレーズの提示である．実際起こるトラブルとは，本人が予想もできなかった思いもよらない出来事である場合も多い．ヒントを提示することでユーザの発想を刺激することができると考える．

旅行計画について説明する．旅行計画については，テキストで書かれた任意の形式の旅行計画を想定しており，Web 上の旅行記から抽出したものや，ユーザ自身が立てたものも考えられるが，本稿では旅行記データセットの旅行記を参考に GPT を用いて作成している [14] [15]．プロンプトは「下記の情報を

1：https://platform.openai.com/docs/api-reference/chat

旅行計画

複数人での旅行計画
1日目
9:00 - 京都駅/東海道新幹線にて東京を出発
11:00 - バスにて栂尾山　高山寺へ移動
12:30 - 西明寺を訪れる
14:00 - 神護寺を訪れる
15:30 - バスにて金戒光明寺へ移動
17:00 - 知恩院を訪れる
18:30 - バスにて京都タワーホテルアネックスへ移動
2日目
9:00 - 京都タワーホテルアネックスから出発
10:00 - 電車にて石清水八幡宮へ移動
11:30 - 電車にて青蓮院門跡へ移動
13:00 - 建仁寺を訪れる
14:30 - 六波羅蜜寺を訪れる
16:00 - 六道珍皇寺を訪れる
17:30 - バスにて京都駅/東海道新幹線へ移動

図 2　旅 行 計 画

参考にし、一泊二日の旅行計画表を作成してください。[旅行情報]」とし，[旅行情報] は著者が選択した旅行スケジュールの素材を入れている．このように生成する理由は，GPT は存在しない施設や場所を生成してしまうことがあるが，実際の旅行情報に基づく素材を入れることで誤った情報が減り，適切に組み立てられた計画を作成できる傾向があるためである．さらに，生成された計画に関係のない文章がついている場合，それを人手で省いてから用いる．旅行計画の例を図 2 に示す．時刻と場所が記載されている一泊二日の旅行計画を用いる．

3.2　大規模言語モデルを用いたトラブルとその対処の評価

使用した system プロンプトについて説明する．system プロンプトとは GPT の挙動を決める命令文である．ゲームの流れの 1 から 4 の処理を行うトラブルを判定する段階の GPT の system プロンプトは「[schedule] という旅行計画を実行した際に、考えられるトラブルであるか、具体性、妥当性の面から判断してください。」である．[schedule] は変数であり，対象となる計画の文書が入る．

ゲームの流れの 5 から 7 の処理を行うトラブル対処法を判定する段階の GPT の system プロンプトは「旅行計画 [schedule] における [trouble] の対処法として適切かどうか、具体性、妥当性の面から判断してください。」である．[schedule] と [trouble] の部分は変数であり，対象となる計画の文書，トラブル判定段階に入力されたトラブルのうちの 1 つが入る．

具体性および妥当性の判定は「非常に（具体的/妥当）である」「(具体的/妥当）である」「どちらともいえない」「(具体的/妥当）ではない」「全く（具体的/妥当）ではない」の 5 段階である．それぞれの判定に応じて（）内の単語が切り替わる．トラブルの判定に「どちらともいえない」「(具体的/妥当）ではない」「全く（具体的/妥当）ではない」が入っていた場合，再回答となる．

図 3　システム画面

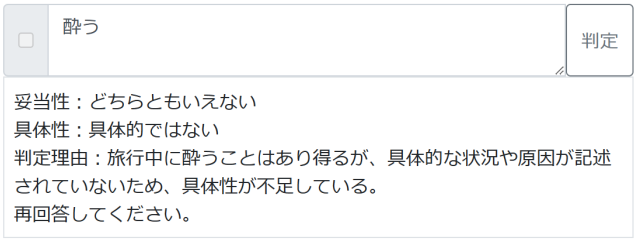


図 4　トラブルが認められない例

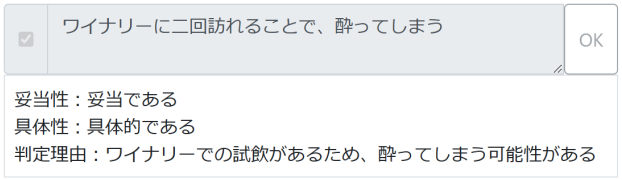


図 5　トラブルが認められた例

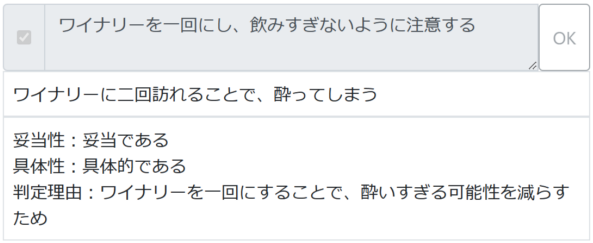


図 6　トラブルの対処法が認められた例

user プロンプトは，ゲームのユーザの入力のことである．ここでは著者が入力した user プロンプトとその出力を例にシステムの挙動を説明する．まず，ゲームを始めると図 3 のような画面が表示される．画面の左側に旅行計画が表示され，画面の右側には入力部分が表示される．この旅行計画を基に判定の例を説明する．トラブルの判定段階においては，図 4，図 5 のような出力が得られる．これらはゲームの流れの 2，3 の処理にあたる．図 4 で「酔う」と書かれている部分が user プロンプトとして扱われ，その右側にある判定ボタンを押すと GPT の判定結果が下に表示される．この場合，妥当性も具体性も認められていないため，再回答を促されている．図 5 のように user

プロンプトを変更してみると妥当性と具体性が認められ，トラブルとして認められる．この場合，入力部分の左側にチェックが付き，判定ボタンが OK に変化する．加えて，入力の変更ができなくなる．次に，トラブルの対処法の判定段階においては，図 6 のような GPT とシステムの出力が得られる．これはゲームの流れの 5，6 の処理にあたる．入力部分の下側に，トラブル段階でトラブルと認められたトラブルが表示される．この表示されたトラブルの対処法について入力する．

3.3 旅行レビューコーパスを用いた重要なトラブルの要素の抽出

ヒントに使用するキーフレーズの抽出手法について説明する．本稿では，ヒントのために用いるが，スポットで起こりえる重要なトラブルを示唆するキーフレーズであるため，このキーフレーズを用いた入力トラブルのスコア付けや，妥当性評価に用いることも可能であると考えている．スポットに関係したヒントを抽出するために，じゃらん [2]から収集したレビューを使用した．ヒントは重要なトラブルに関するキーフレーズを出す必要があるため，旅行計画に記載されているスポットのレビューを低評価と高評価に分け，それぞれ EmbedRank を用いてキーフレーズ抽出を行った [12]．EmbedRank は，文書全体に類似するようなキーフレーズを抽出するアルゴリズムで，ここのキーフレーズが類似しないように，文書要約でも用いられる Maximal Marginal Relevance を用いている．本稿では，類似度の判定に fasttext を用い，文書全体におけるキーフレーズの出現頻度の対数を重みとして加えるという変更を加えている．

評価 1，2 のレビュー数が合計 100 以上の場合，評価 1，2 を低評価レビューとして用いる．合計 100 未満の場合は評価 1，2，3 を低評価レビューとして用い，評価 4，5 を高評価レビューとして用いる．評価 1，2 のレビュー数によって評価 3 を使う理由を説明する．評価 1，2 のレビューは「混雑で思うように進めなかった。すごく疲れた。」,「チケットをコンビニで発券しないといけないって知らなかった！指定の時間に間に合わなかった。」のように，そのスポットで起こったネガティブな出来事に対して記述されていることが多く，ネガティブなキーワードが抽出しやすいが，レビュー数が少ないスポットがあり，適切にキーフレーズを抽出できない場合がある．そのため，評価 3 を追加し，レビュー数を増やしている．低評価のレビューの中には「すごい混雑だったけど、すごく景色が綺麗だったので満足です。」,「美味しいものがたくさんあってまた行きたい。客引きが多いのだけがちょっと。」のように，そのスポットで起こったネガティブな出来事とポジティブな出来事について記述されているレビューもあり，ポジティブなキーフレーズも抽出してしまうため，低評価と高評価のキーフレーズを比較し，高評価に出てこないキーフレーズを抽出している．低評価，高評価とも 100 件のキーフレーズを抽出し，低評価のみに含まれる上位 5 件のキーフレーズを用いた．

例として，東京スカイツリーのレビューを取得し，キーフレーズを抽出してみると「高い 値段」,「スカイ ツリー 遠く」というキーフレーズが上位に出力される.「高い 値段」というキーフレーズはチケットの値段が高い，高い値段と見合わないというような口コミから得られたキーフレーズだと考えられる.「スカイ ツリー 遠く」というキーフレーズはスカイツリーに登ったが雨で遠くの景色が見られなかった，スカイツリーは登るのではなく遠くから見るのがいいというような口コミから得られたキーフレーズだと考えられる．

4 評 価 実 験

4.1 大規模言語モデルを用いたトラブルの評価精度

実験目的は，GPT が判定する妥当性と具体性の評価の有効性を検証することである．

実験環境について説明する．先行研究 [1] で構築したシステムに対して，山梨，京都，奥多摩の 3 種類の旅行計画に対し，60 人がゲームを実施した．そこで入力されたトラブルから各旅行計画につき無作為に 50 件抽出した．各トラブルに 3 人の被験者から，具体性と妥当性を判定してもらう．実験参加者は，それぞれのグループで 150 人ずつ，計 450 人をクラウドソーシング [3]で募集した．

実験手順について説明する．被験者には提示された旅行計画から考えられるトラブルの具体性と妥当性を判定する実験であることを説明した．説明文に旅行計画とその旅行計画で起こりうるトラブルを表示し，被験者には旅行計画に対するそのトラブルの妥当性と具体性について判定してもらった．なお，この実験ではヒントを提示していない．

実験結果について説明する．評価指標に RMSE を用いた．各トラブルについて,「非常に（具体的/妥当）である」を 5 とした 5 段階の数値に変換し，3 人の被験者の回答の平均値を求め，それを用いて計算した．結果を表 1 に示す．RMSE は 1.3 程度であり，評価をランダムで行ってみると 1.78 になるため，ランダムよりは良い分類ができているといえる．また,「9 時の新幹線が止まってしまう」というトラブルの GPT の回答は「妥当である」である．しかし被験者は「全く妥当ではない」「全く妥当ではない」「妥当ではない」と回答しており GPT の回答は大きく外れている．被験者は「発生確率が低いトラブル」を「妥当ではない」と判定しており，GPT は旅行計画から考えられるトラブルか判定しているため回答が外れてしまうと考える．

GPT の判定精度を上げるため，この実験で得た被験者の回答をトラブルを判定する段階の system プロンプトに与えた.「[schedule] という旅行計画を実行した際に、遭遇しやすいトラブルか判断してください。 例 [雨が降り、絶景を見ることができない] というトラブルの場合、出力は [妥当である][具体的である][遭遇しやすいトラブルであり、明確に記述されているため]」という few-shot prompt に変更した．例はあと 3 つ続くがここでは省略している．与えている被験者の回答はトラブル，妥当性，具体性，判定理由の順となっており，妥当性と具

2：https://www.jalan.net/travel

3：クラウドワークスを用いた，以降の実験のクラウドソーシングでも同様である．https://crowdworks.jp/

表 1　RMSE の結果

旅行計画	妥当性	具体性
山梨	1.233	1.302
京都	1.456	1.408
奥多摩	1.489	1.197
全体	1.398	1.305

表 2　フレーズ抽出の結果

観光スポット	出力		
東京スカイツリー	高い 値段	スカイ ツリー 遠く	多い 残念
清水寺	人 だらけ	多い 修学旅行	改装 中
中華街	全体 的	感じ 場所	味 普通
札幌市時計台	大きい 建物	期待 外れ	小さい ビックリ

体性は被験者の回答を用い，判定理由は著者が与えた．このプロンプトでの RMSE は 1.17 程度であり，few-shot を行ったプロンプトの方が精度が良かったため 4.3 の実験ではこれを用いている．

4.2　抽出したヒントワードの抽出結果の考察

実験目的は重要なトラブルに関するヒントとして抽出できているか検証することである．

実験環境について説明する．4 つの観光スポットの観光レビューを用意した．各スポットで 3 人に回答してもらうため，男性 10 人と女性 2 人の旅行経験のある大学生 12 人の実験参加者を用意した．

実験手順について説明する．最初に，被験者には重要なトラブルに関する単語を考えてもらう実験であることを説明した．観光スポットの評価 1，2 のレビューを 20 件，評価 3 のレビューを 10 件表示し，重要なトラブルを示す単語を 3 つ挙げてもらう．その際，各スポットの 3 人の回答者には全て異なるレビューを表示する．

実験結果について説明する．出力を表 2 に示し，被験者の回答を表 3 に示す．被験者の回答と抽出した単語の上位 3 つを比較する．東京スカイツリーの回答は「混雑」「料金」について挙げられており，出力では「高い 値段」という単語が抽出できている．清水寺では「工事」「混雑」について挙げられており，出力では「人だらけ」「改装 中」という単語が抽出できている．中華街の回答は人によってばらついている．出力は「全体 的」「感じ 場所」という単語が抽出されており，情報が不十分で何について述べられているか把握することができない．また，札幌市時計台の回答は「地味」「通り過ぎ」というような外観についての回答が多く，出力では「小さい ビックリ」という外観についての単語が抽出できている．今後は，精度を向上させるためによりよい単語抽出方法について検討する．

4.3　旅行トラブルの対応力育成ゲームの評価

実験目的は，本提案システムの有効性を示すために，本システムを用いることでトラブル対応能力が向上するのか検証することである．実験仮説を以下のように設定した．

（1）　システムを使用するグループは使用しないグループに比べて，回答するトラブル数が増加する．

表 3　被験者が回答したトラブルを示す単語

観光スポット	回答		
東京スカイツリー	金額変動	混雑	強風
	混雑	予約	料金
	人ごみ	値段高い	行列
清水寺	飲食禁止	改修工事	改築中
	工事	混雑	渋滞
	外国人	マナー	混雑
中華街	客引き	舌打	食べ放題ばかり
	衛生	混雑	ゴミ
	当たり外れ	観光地価格	人が多い
札幌市時計台	期待	残念	微妙
	しょぼい	入場料	通り過ぎちゃう
	駐車場がない	地味	通り過ぎ

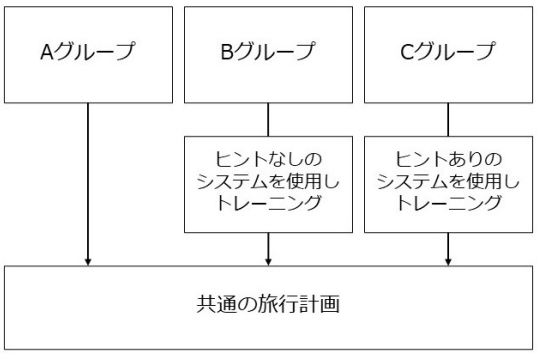


図 7　実験の流れ

（2）　ヒントありグループはヒントなしグループに比べて，回答するトラブルが多様化する．

実験環境について説明する．システムを使わないグループ，ヒントなしでシステムを使うグループ，ヒントありでシステムを使うグループの 3 つのグループを用意した．トレーニングに用いる沖縄，京都，奈良の 3 種類の旅行計画を用意し，共通で提示する山梨の旅行計画を用意した．実験参加者は，それぞれのグループで 30 人ずつ，計 90 人をクラウドソーシングで募集した．

実験手順について説明する．実験手順の概要を図 7 に示す．被験者には最初に提示された旅行計画から考えられるトラブルを挙げる実験であることを説明した．その後，システムを使わないグループ，ヒントなしでシステムを使うグループ，ヒントありでシステムを使うグループの 3 つのグループに分け，システムを使う 2 つのグループには 3 種類の旅行計画を用いてトレーニングを 3 回行ってもらった．その後，それぞれのグループに共通の旅行計画を提示し，トラブルを列挙してもらった．その際，時間は 4 分とした．

実験結果について説明する．著者ではない大学生 1 人に協力してもらい，重複の回答を省いた．トラブルではない回答は著者が判断し省いた．トラブル数を表 4 に示す．トラブルの数に大きな差は見られなかったが，ヒントありグループがもっともトラブル数が少なく，システムを使用しないグループがもっともトラブル数が多い結果となり，仮説 1 を支持することはで

表 4　トラブル数の比較

	システムなし	ヒントなし	ヒントあり
トラブル数	144	137	135
一人当たりのトラブル数	4.8	4.6	4.5
理由あり [%]	53.5	54.7	65.2

きなかった．この原因として，トレーニングの回数が少なかったことやトラブルを列挙する時間が短かったことなどが考えられる．

回答を形態素解析し単語の出現量をグループごとに比較すると，ヒントありグループの回答は言及しやすいと考えられる天気系や食事系，時間系が減り，他のトピックに言及が分散していると考えられる．例えば，ヒントを用いたグループでは「散策時に薄着で不調になる」「登山に適した靴を用意していなくて大変だった」などの服装に関する回答が挙げられていたが，他グループは服装に関する回答は挙げられていなかった．このことから，回答するトラブルが多様化すると考えられ，仮説 2 が支持された．

表 4 の理由ありについて説明する．被験者の回答を理由なしと理由ありに分けた場合の理由ありの数値のみを掲載しており，理由なしというのは「寝坊」「遅刻したメンバーがいた」という理由が記述されていない回答で，理由ありというのが「悪天候で富士山に登ることができなかった」「工事中で自然に囲まれた露天風呂に入ることができない」という理由が記述されている回答である．ヒントありのシステムを用いたグループがもっとも高い結果となり，ヒントありグループはヒントなしグループと比べて回答するトラブルが具体化することがわかった．今後はゲームの流れやヒントの提示方法について見直し検討する．

5 ま と め

私たちは旅行に行き新しい文化や風景に触れると，新鮮な刺激を受け，充実感を含め魅力的な経験を得られるが，旅行に慣れていない人にとっては予期せぬトラブルに遭遇することもある．トラブルを回避したり，仮にトラブルに遭遇した場合にも迅速に対処するためには，旅行前に予防策を講じ，適切な対処法を考えておくことが重要である．このことから本研究では，提示される旅行計画に対してユーザが発生しうるトラブルとそのトラブルの対処法を考えることで，トラブルの対応能力を向上させる会話型のトラブル当てゲームを提案した．

本提案システムの有効性を示すために，大規模言語モデルを用いたトラブルの精度評価を RMSE を用いて行った．その結果，ランダムよりは良い分類ができているといえるが，GPT の判定精度に課題があり，GPT のプロンプトに改良の余地があることが分かった．また，抽出したヒントワードの精度評価を行った結果，被験者と同じ事柄の単語を抽出できている観光スポットもあったが，適切なキーフレーズを抽出できていない観光スポットもあり単語抽出手法の精度に課題があることが分かった．旅行トラブルの対応力育成ゲームの評価実験では，トラブルの数に差は見られなかったが，ヒントありグループはヒントなしグループと比べて回答するトラブルが多様化，具体化すると考えられた．今後は GPT のプロンプトを改善し，単語抽出方法やヒントの提示方法について検討することで，能力向上効果の高いシステムの構築を目指す．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．本研究では，国立情報学研究所の IDR データセット提供サービスにより株式会社地球の歩き方から提供を受けた「地球の歩き方旅行記データセット」を利用しました．ここに記して謝意を表すものとします．

文　　献

[1] 斉藤さつき, 小竹 神大輔. 旅行時の対応能力向上のための大規模言語モデルを用いたトラブル当てゲーム. 第 22 回情報科学技術フォーラム, pp. 255–256, 2023.
[2] 服部峻, 高原まどか. 教育指向 text-と-image ゲーム ai の検討. 日本デジタルゲーム学会 2022 年夏季研究発表大会予稿集, pp. 42–44, 2022.
[3] 一ノ瀬智浩, 上野秀剛. ゲーミフィケーションを構成する要素の違いと作業効率の評価. ヒューマンインタフェース学会論文誌, Vol. 18, No. 2, pp. 65–76, 2016.
[4] 藤田美幸. ゲーミフィケーションにおけるユーザの動機づけとエンゲージメントの関連. 日本情報経営学会誌, Vol. 38, No. 3, pp. 83–92, 2018.
[5] 杉谷修一. 学習ツールとしてのゲーミフィケーションの可能性. 日本情報経営学会誌, Vol. 22, pp. 71–79, 2018.
[6] 藤本徹. ゲーム学習の新たな展開. 放送メディア研究, Vol. 12, p. 233–252, 2015.
[7] 倉田陽平. 観光周遊支援ゲームのこれから. 観光情報学会第 6 回研究発表大会, pp. 41–48, 2013.
[8] 倉田陽平, 有馬貴之. 対話的旅行計画作成支援システムの実装と評価. 日本観光研究学会全国大会学術論文集, Vol. 25, pp. 173–176, 2010.
[9] 古賀友朗, 山本佳世子. ゲーミフィケーションを活用した観光計画支援システム. 第 83 回全国大会講演論文集, No. 1, pp. 579–580, 2021.
[10] 吉田伊武貴, 倉田陽平. 旅先における失敗リスクを把握可能にするための機械学習を用いた失敗談ツイート抽出方法の構築と静岡県内観光地での適用. 第 7 回とうかい観光情報学研究会, pp. 1–4, 2021.
[11] 吉田伊武貴, 倉田陽平. 投稿者による評価スコアや感情極性辞書を用いない，文意に基づく宿泊施設不満レビュー自動抽出手法の構築及びその分析による不満予防策の検討. *DEIM2022*, 2022.
[12] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossman, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, p. 221–229, 2018.
[13] 鈴木大助. Chatgpt を用いた教務事項問い合わせ対応システムの検討. 研究報告音声言語情報処理, Vol. 149, No. 26, 2023.
[14] 大内啓樹, 進藤裕之, 若宮翔子, 松田裕貴, 井之上直也, 東山翔平, 中村哲, 渡辺太郎. 地球の歩き方旅行記データセット. 言語処理学会 第 29 回年次大会発表論文集, pp. 2920–2924, 2023.
[15] 株式会社地球の歩き方 (2022): 地球の歩き方旅行記データセット. 国立情報学研究所情報学研究データリポジトリ.(データセット).

ジオタグつき投稿を用いたコロナ禍の影響の地域差の分析

田中日加吏† 横山 昌平††,†††,††††

† 東京都立大学 システムデザイン学部 〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科 〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所 〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所 〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †tanaka-hikari@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし SNS はニュースなどのメディアと違い、利用者が自ら情報を発信するため、人々の動向が直接的に反映される。こうした特性から、SNS を用いて人々の行動や場所の特性、ユーザーの感情や属性を分析する研究が注目を集めている。本研究では、LDA（Latent Dirichlet Allocation) を用いた文章集合間の比較手法を提案する。本手法では、文章集合を LDA に用いて得られる各文章が持つトピックの確率分布から、トピックの確率分布を生成するディリクレ分布を文章集合ごとに推定し比較することで、トピックを基準とした文章集合間の類似度の算出や、差異を引き起こすトピックの抽出などを行う。本手法の有効性を検証したのち、緊急事態宣言時などの社会活動に制約があった期間における Twitter（現 X）への投稿を本手法に用いて、コロナ禍において Twitter に反映された人々の動向やその変遷と地域との関係の分析を行う。

キーワード Twitter(現 X), 地理情報, コロナ禍、LDA

1 はじめに

新型コロナウィルス感染症の拡大により、世界各国が長期に渡り経済的、社会的な影響を受けた。日本では、緊急事態宣言やまん延防止等重点措置などにより休校や、外出の自粛、営業時間の短縮など、人々の社会活動が抑制され、テレワークなどの勤務形態が一般化するなど、我々の社会の在り方を大きく変えてしまった。こうした長期間の変化の中で、人々の新型コロナウィルスに対する認識などは大きく変化したと考えられる。

SNS(Social Network Service) に投稿された内容を分析することで、人々の認識を抽出できることは研究によって知られており、マスメディアなどを介さないため、ありのままの人々の動向が即自的に反映されるという特性から、SNS への投稿の分析によりユーザーの社会の変化やユーザーの動向を抽出する研究が盛んになっている。特に、本研究で用いる SNS である Twitter（現 X）では、その手軽さから国内における利用者や投稿数が多く，ジオタグと呼ばれる、投稿が行われた位置の情報を併せて投稿することができるため、ジオタグつき投稿を用いて、投稿に関する情報を地理的特性に結び付けて分析する研究も行われている。

本研究では、Twitter のジオタグ付き投稿から、コロナ禍における人々の動向と地域の関係の分析を行う。東京都は 2020 年 4 月 7 日の緊急事態宣言に始まり、2022 年 3 月 21 日のまん延防止等重点措置の解除まで、計 4 回の緊急事態宣言と計 3 回のまん延防止等重点措置の合わせて 354 日の間、社会活動が抑制されていた。一方、岩手県、秋田県、福井県、奈良県、鳥取県、徳島県は、最初の緊急事態宣言の 29 日間のみの抑制である。このような社会活動が制限されている期間の違いから、生活する地域の違いによって、コロナ禍に対する認識に違いが生じ、それらは Twitter への投稿に反映されると考えることができる。

この場合、地域は、その範囲内で投稿された文章をまとめて保持するクラスと考えることができる。本研究では、このようなクラス同士の差異を検出し、比較する手法として、ピックモデルの手法である LDA（Latent Dirichlet Allocation) [5] を活用した手法を提案する。通常、トピックモデルは、文章の生成過程をモデル化することで、文章のクラスタリングや、文章集合がどのような潜在トピックに従っているかを可視化することに用いられる。

本手法では、LDA モデルに文章を入力して得られる文章のトピック分布から、トピック分布を生成するディリクレ分布の推定をクラスごとに行う。推定されたディリクレ分布のパラメータを分析することで、トピックを基準としたクラス間の類似度の算出や、差異を引き起こすトピックの抽出などの分析が可能になる

本論文は次の構成に従う。2 章では本研究の関連研究について説明し、3 章では提案手法について説明する。4 章では、あらかじめ差異が判明している複数の文章集合を提案手法に用いることで本手法の有効性の検証を行い、第 5 章では、ジオタグによってそれぞれの集合に分けられた witter への投稿文章に本手法を用いた実験を行い、結果を考察する。最後に、6 章でまとめと今後の課題や展望について説明する。

2 関連研究

本章では関連研究として、Twitter への投稿から新型コロナウィルスに関する分析を行う研究と、Twitter への投稿から地

域の特徴量を抽出する研究について述べる。

2.1 Twitter から新型コロナウィルスに関する分析を行う研究

四方田は [1]、Twitter への投稿から「コロナ」を含む投稿を収集し、KH Coder を用いた定量的テキスト分析手法によって分析し、コロナウイルスに関する不安やストレスは、感染への恐怖、日常行動やの制限によるストレス、政府の疫学対策や経済的被害への不安、メディア情報による不安などに由来していることを明らかにした。

鳥海ら [2] は Twitter から「コロナ」を含む投稿を収集し、北海道に緊急事態宣言が出され、新型コロナが話題となった 2020 年 2 月 28 日以降、新型コロナウイルスについてツイートしたユーザーの偏りがほとんどなくなっていることを明らかにし、次に、投稿に含まれる単語から、新型コロナウイルスに対する人々の感情を分析し、その結果、特定の社会的出来事の発生によって、SNS 上で表現される感情が変化することを明らかにした。本研究では、ジオタグ付き投稿を用いることで、分析内容を地域と結び付けることを目指す。

2.2 Twitter へのジオタグ付き投稿から地域を分析する研究

太田ら [3] は Twitter のジオタグつき投稿を用いて、投稿の時空間属性を定義し、それを用いて投稿の連投度合いや、場所への滞在度合などの地理的特性などの推定を行った。太田らの研究は投稿場所の時空間変化に着目しており、投稿文章を用いないのに対し、本研究は、投稿場所と投稿内容に焦点を当てており、投稿文章を分析する点で異なる。

荘司ら [4] は Twitter のジオタグ付き投稿を用いて場所の分散表現を構築した。単語の分散表現である Word2Vec における、「単語の意味はその文脈に出現する単語によって形成される」という分布仮説を踏襲し、「ある場所の周辺で現れる投稿に含まれる単語がその場所の意味を形成する」という仮説を立て、単語と地理情報を用いてニューラルネットを訓練した。荘司らの研究と本研究には、場所の特徴量を抽出し比較するという類似点があるが、本研究ではニューラルネットでなく、確率モデルを用いる点と、特徴量は分散表現ではなく、トピックとその事前分布であり、特徴量が示す内容が明示的であるという違いがある。

3 提 案 手 法

本章では、本研究で用いる手法について述べる。手法の概略を下に述べる。

- 投稿文章の前処理
- トピックモデルの実装
- 文章集合ごとのディリクレ分布の推定
- 文章集合間の類似度の計算
- 特徴量の分析

3.1 投稿文章の前処理

Twitter への投稿文章のようなソーシャルデータには、絵文字や記号などの形態素解析に対してノイズとなるような文字列や、分析の対象に不適切である広告や bot による投稿などが含まれるため、まずはじめにこれらの除去を行う。本研究では用いる投稿文書に次のような処理を行う。

- URL やハッシュタグ、ユーザーへのメンションを含む投稿を除去
- 絵文字や記号を除去
- 半角カタカナを全角カタカナへ変換

次に各投稿文書に対して、MeCab を用いて形態素解析を行い、ツイート文書を形態素に分割し、基本形に変換する。本研究では投稿文章の内容の分析を目指すため、投稿文章の内容を表す、動詞、形容詞、名詞を利用する。また、「する」や「こと」のような、どの文章にも出現する、一般的なストップワードを除去する。

これらの処理を行って得られた、投稿文章ごとの形態素のリストを、bag-of-words 形式のベクトルに変形し、次項で述べるトピックモデルの実装に用いる、コーパスとして利用する。

3.2 トピックモデルの実装

前節で得られたコーパスを用いてトピックモデルを実装する。本研究では LDA モデルを用いる。LDA とは文章が複数のトピックを持つことを仮定して、文章のトピック分布が文章毎にディリクレ分布から生成されると仮定してモデリングする手法である。

要素の和が 1 になるようなベクトルである確率ベクトルを生成する分布のことをディリクレ分布と呼び、パラメータ α がであるディリクレ分布を $Dir(\alpha)$ と表し、次式のように定義される。

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

LDA では、文章生成を次のようにモデリングする。

（1） 各トピックは単語の確率分布φをもつ。φは事前分布であるディリクレ分布 Dir(β) から生成される。

（2） 各文章 d が生成されるときに、その文章が持つトピックの確率分布 θ_d を事前分布であるディリクレ分布 Dir(α) から生成する。

（3） 文章は文章ごとに定められている単語の個数 N_d だけ、自身のトピック分布 θ_d からトピック z を生成する。

（4） N_d 個のトピック z_{dn} のそれぞれが、単語の確率分布φに基づき単語を生成する

（5） 得られた単語列を文章とみなす。

すなわち

$$\theta \sim Dir(\alpha), \Phi \sim Dir(\beta)$$

のとき、LDA では文章生成を次式のように表す。

$$p(w_d|\theta_d,\Phi) = \prod_{n=1}^{N_d}\sum_{k=1}^{K} p(z_{dn}=k|\theta_d)\cdot p(w_{dn}=k|\phi_k)$$
$$= \prod_{n=1}^{N_d}\sum_{k=1}^{K} \theta_{dk}\cdot\phi_{kw_{dn}}$$

モデルの構築にあたって、未知変数は変分ベイズ推定法を用いて推定する。

3.3 文章集合ごとのディリクレ分布の推定

前節で述べたトピックモデルに、各投稿文章を入力することで、その文章のトピック分布が得られる。これを用いて、事前分布であるディリクレ分布を文章集合ごとに推定する。文章数が N 個であり、各文章のトピック分布を $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ としたとき、最尤推定を用いた推定では、次式であらわされる尤度関数を最大にするような α を求める。

$$\log L(\boldsymbol{\alpha}) = \sum_{j=1}^{N}\left(\log\frac{1}{B(\boldsymbol{\alpha})} + \sum_{i=1}^{K}(\alpha_i - 1)\log x_{ji}\right)$$

ディリクレ分布における最尤推定では解析的に解を得ることができない [6] ため、不動点反復法を用いて数値計算を行う。

3.4 文章集合間の類似度の計算

前節で得られた、各地域における投稿文書のトピック分布を生成するディリクレ分布の類似度を地域間の類似度とする。ディリクレ分布の類似度の計算には、KL 情報量（Kullback-Leibler divergence）を用いる。KL 情報量とは、ある確率分布 Q が対象となる確率分布 P をどのくらい忠実に近似しているかを数値化したものであり、P = Q のとき 0 をとる。$P = Dir(\alpha_1), Q = Dir(\alpha_2)$ とすると 2 つのディリクレ分布の間の KL 情報量は次のように解析的に求めることができる [7]。

$$\mathrm{KL}[P\,||\,Q] = \ln\frac{\Gamma\left(\sum_{i=1}^{k}\alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^{k}\alpha_{2i}\right)} + \sum_{i=1}^{k}\ln\frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})}$$
$$+ \sum_{i=1}^{k}(\alpha_{1i}-\alpha_{2i})\left[\psi(\alpha_{1i}) - \psi\left(\sum_{i=1}^{k}\alpha_{1i}\right)\right].$$

ただし、KL 情報量は対称性を持たず、$\mathrm{KL}[P\,||\,Q] \neq \mathrm{KL}[Q\,||\,P]$ であるため、本研究では $\mathrm{KL}[P\,||\,Q] + \mathrm{KL}[Q\,||\,P]$ を 2 つのディリクレ分布の間の類似度とする。

3.5 特徴量の分析

前節で得られてた地域間の類似度や、トピックモデルの実装の過程で求まるトピックごとの単語分布、地域ごとのディリクレ分布のパラメータの比を用いて、投稿内容を地理的条件に結び付けて分析を行う。

ディリクレ分布のパラメータ α の要素である α_k はトピックの添え字にそれぞれ対応しており、α_k の比率が大きいほど、分布は k 番目の確率を大きく生成するという特性がある。そのため、推定された α の大きさを比較することで、各文章集合において、どのトピックが生成されやすいかを確認することができる。

表 1: トピック数 4 で実装した LDA モデルの各トピックにおける頻出単語

Topic1	対応、製品、搭載、画面、機能、表示、USB、可能、ユーザー、使用
Topic2	人、機能、アプリ、情報、サービス、PC、利用、提供、ファイル
Topic3	語る、自分、見る、選手、コメント、写真、試合、チーム、明かす
Topic4	氏、行う、放送、選手、日本、監督、日本代表、試合、TBS、

表 2: 各文章集合で推定したディリクレ分布のパラメータ

	α1	α2	α3	α4
Cm1	0.299446560	0.312434014	0.281325587	0.244923055
Cm2	0.280984141	0.315186535	0.303901037	0.242601529
Cit	**0.728545368**	**0.850825350**	0.566111709	0.468615939
Cs	0.443903439	0.464533262	**0.843225670**	**0.738380000**

4 実　験　1

本章では、提案手法の有効性の検証を行う。すなわち、本手法を用いることで、文章集合間の比較が可能になるかどうかを検証する。

4.1 実　　装

本実験では、livedoor ニュースコーパスを提案手法に用いる。livedoor ニュースコーパスは、ラベル付けされたトピックごとに分類されたニュース記事の集合であり、本実験では、「Sports Watch」と「IT ライフハック」のニュース記事を用いる。提案手法の章で述べた前処理を行い、それぞれをコーパスに用いる集合と、推定に用いる集合として 9:1 の割合で分割した。「Sports Watch」の文章数は 10,732:1,193 に、「IT ライフハック」の文章数は 35,242:3,916 であった。

前節で得られたコーパスを用いて、トピック数を 2〜10 と変更させて 9 つの LDA モデルを実装した。モデルの評価には coherence と perplexity を用いる。提案手法を用いる場合、トピックが持つ単語分布を分析することになるため、有意な解析には coherence が高いモデルを用いる必要があると考えられるため、以降は coherence が高かった、トピック数が 4 のモデルについて考察する。モデルの各トピックにおける頻出語を表 1 に示す。

4.2 結果と考察

推定に用いる文章集合に関して、「Sports Watch」から得られた文章集合を C_s、「IT ライフハック」から得られた文章集合を C_{it}、C_s と C_{it} をシャッフルし、二つに分けた文章集合を C_{m1}、C_{m2} とする。C_s、C_{it}、C_{m1}、C_{m2} を用いてディリクレ分布を推定した結果を表 2 に示す。、また、C_s、C_{it} で得られた分布間の KL 情報量と、C_{m1}、C_{m2} で得られた分布間の KL 情報量はそれぞれ 1.309,232,949、0.011,140,108 であった。

C_s、C_{it} の比較は二つの文章集合の分布に差異がある場合の検証である。C_{it} から推定されたディリクレ分布のパラメータはトピック 1、2 が出現しやすいことを示しており、C_s から推

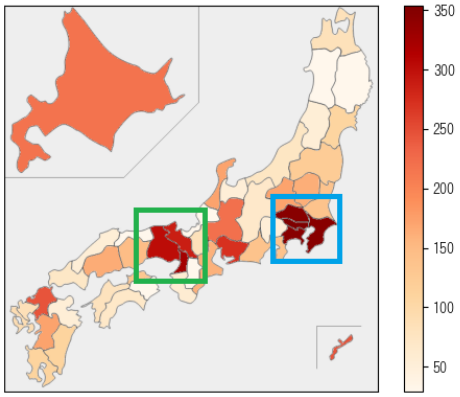


図 1: 緊急事態宣言やまん延防止等特別措置の合計日数を表す地図
範囲 1 は青線で、範囲 2 は緑線で囲まれている

定されたディリクレ分布のパラメータはトピック 3、4 が出現しやすいことを示している。そして、C_s、C_{it} 間の KL 情報量は、後述する C_{m1}、C_{m2} の KL 情報量に比べて非常に大きい。これらのことから、文章の分布に差異があれば KL 情報量は増加、すなわち類似度が減少し、また、それぞれの分布のパラメータから、どのトピックに違いがあるのかを確認できることがわかる。

C_{m1}、C_{m2} の比較は二つの文章集合が同程度の分布である場合の検証である。両者ともディリクレ分布パラメータの差異は小さく、KL 情報量は、C_s、C_{it} 間の KL 情報量に比べて小さく、すなわち類似度が高いことを示している。これらのことから、二つの文章集合が同程度の時には類似度が大きくなり、分布のパラメータの差異も小さくなることが確認できた。

5 実 験 2

本章では第 3 章で述べた手法を用いて、コロナ禍における日本語の Twitter への投稿文章でトピックモデルを構築し、Twitter 上に反映されているユーザーの動向とその地域差について分析する。

5.1 実 装

図 1 は、各都道府県ごとに緊急事態宣言やまん延防止等特別措置が行われていた日数を示したカラーマップと、実験で用いる地域の範囲を示した図である。実験では、東京周辺、京都周辺、その他の地域の計 3 地域に分割して分析を行う。

訓練に用いるツイートの期間は、2020 年 2 月 10 か日から、最初の緊急事態宣言が解除されるまでの 2020 年 5 月 25 日、最初の緊急事態宣言が解除されてた 2020 年 5 月 26 日から、次の緊急事態宣言が発令されるまでの 2021 年 1 月 7 日、2 回目の緊急事態宣言が発令された 2021 年 1 月 8 日から、4 回目の緊急事態宣言が解除されるまでの 2021 年 9 月 30 日の 3 つの期間を用いる。これらのツイートは株式会社 NTT データ「なずきのおと」を用いて収集した.

表 3: 実験に用いた期間ごと投稿数

	期間 1 (105日間)	期間 2 (227日間)	期間 3 (266日間)
全量データ	7390487	6171074	6765666
東京周辺	16206	10587	9319
京都周辺	6142	3966	3717
その他の地域	18513	11468	10658

表 4: 各期間における、地域の類似度

	期間 1	期間 2	期間 3
(東京周辺,京都周辺)	0.005096769	0.001046773	0.001958751
(東京周辺,その他の地域)	0.007954446	0.003874121	0.003264716
(京都周辺,その他の地域)	0.007001663	0.002421606	0.005209866

実験には、2.1 節で述べた前処理を行った、「コロナ」という文字列を含むツイートのうち、ジオタグを持つツイートは推定に、持たないツイートはモデルの訓練に用いた

それぞれの地域と区間における、実験に用いたデータの数を表 3 に示す。

3.1 節と同様、 coherence と perplexity を評価してモデルを実装した。coherence が最大となるようなトピック数を選び、期間 1、2、3 でそれぞれ、トピック数は 6、5、6 となった。

5.2 地域間の類似度

それぞれの期間における、地域間の類似度を表 4 に示す。表 4 から、東京周辺と京都周辺の組み合わせから算出された KL 情報量が、すべての期間で一貫して小さく、ほかの組み合わせに比べて、東京周辺でツイートされた内容と、京都周辺でツイートされた内容が類似していることがわかる。

すなわち、すべての期間において東京周辺とその他の地域の間には SNS への投稿の傾向に差が存在することが分かった。期間 3 にのみ地域差が認められてはいないため、「自粛期間の差が SNS への投稿に反映される」という本研究で初めにたてた仮説は確かめられなかった。

これらを考慮すると、「人口の多さ」が SNS への投稿の傾向に影響を与え、それは同時に自粛期間の長さの要因でもある、という仮説が考えられる。

5.3 トピックの遷移

これ以降は特に、東京周辺とその他の地域の比較を論ずる。

各期間におけるトピックの内容を表 5 に示す。

時間経過により出現したトピックは、期間 2 における外出に関するトピックと、期間 3 におけるワクチン接種に関するトピックである。これらは、期間 2 において自粛が一時的に緩和され、人々が旅行に行くようなキャンペーンがあったことや、期間 3 においてワクチン接種が解禁されたことなどの事実に対応している。

時間経過によって消失、すなわちその内容の投稿が減少し、ほかのトピックに統合されてしまったトピックは、コロナに対する具体的な体験を伴わない投稿であり、表 3 の各期間におけ

表 5: 各期間におけるトピックの内容

	期間 1	期間 2	期間 3
Topic1	感想	外出・感想	業務連絡
Topic2	社会的関心・コロナ対策	学校・仕事	罹患・体調
Topic3	学校・仕事	社会的関心	外出・感想
Topic4	罹患・体調	罹患・体調	社会的関心
Topic5	感想・ネガティブ	感想	学校・仕事
Topic6	社会的関心・行事		ワクチン接種

表 6: 各期間における地域ごとのトピックの出現率

(期間 1)	α 1	α 2	α 3	α 4	α 5	α 6
東京周辺	**0.192951**	0.153737	**0.162119**	0.153946	0.174285	0.162961
その他の地域	**0.198930**	0.149435	**0.155010**	0.158091	0.169586	0.168947

(期間2)	α 1	α 2	α 3	α 4	α 5
東京周辺	0.218931	**0.161058**	**0.229249**	0.168406	0.222356
その他の地域	0.225737	**0.164886**	**0.223073**	0.166914	0.219390

(期間3)	α 1	α 2	α 3	α 4	α 5	α 6
東京周辺	0.168838	0.144810	0.167349	**0.182101**	**0.180280**	0.156622
その他の地域	0.174148	0.143977	0.167354	**0.174741**	**0.182843**	0.156938

る投稿数の推移と併せて、人々がコロナに対する投稿をしなくなっていったと推察できる。これは、人々がコロナに慣れて、それが日常の一部になっていっているため、わざわざコロナの投稿をしなくなっていると考えられる。

通時的に発生しているトピックは、社会的関心や罹患・体調に関するトピックであり、具体的な体験や、社会情勢の変化など、SNS に投稿する根拠が明確に存在するような内容である。

5.4 トピックの地域差

各期間における東京周辺とその他の地域における各トピックの出現率を表 6 に示す。各トピックの出現率は、ディリクレ分布のパラメータを正規化することで求めている。

期間 1 に関しては、コロナに対する単なる感想などの、具体的な体験が伴わないトピックがその他の地域で比較的多く発生するのに対し、東京周辺では、学校や仕事に関するトピックが比較的多く発生する。一方で、期間 2、3 に関しては、学校や仕事に関するトピックはその他の地域の方が多く、東京周辺では社会的関心のトピックが多い。すなわち、東京周辺はその他の地域に比較して、実体験を伴うコロナに対する投稿のうち、学校や仕事などの特に日常生活にかかわる内容が少ない傾向にあると判明した。トピックの遷移より、全体の傾向として、人々がコロナに慣れることで、コロナに関する投稿が減少傾向にあると考察したが、これに加えて、東京周辺では、コロナが日常生活にかかわる内容も少なくなるため、コロナに対して慣れる速度にも違いがあると考えられる。

6 ま と め

本研究では、SNS に投稿された内容の差異を抽出して分析する手法として、LDA モデルを構築したのち、それぞれの文章集合を LDA モデルに適応して得られるトピックの確率分布から、文章集合ごとにディリクレ分布を推定し、それらを比較することで、類似度と差異の要因を抽出する手法を提案した。

実験 1 においては、差異が見込まれる文章集合と、差異を抑えるために混合したのちに 2 つに分割した文章集合のそれぞれに提案手法を実行し、本手法を用いることで、文章集合の間に差異が大きい場合はそれを検出でき、差異が小さい場合は高い類似度が得られるということを確認した。

提案手法の有効性を確認したのち、実験 2 において Twitter のジオタグ付き投稿を用いたコロナ禍における地域差の分析を行い、京都周辺と東京周辺の、人口が多い地域同士の類似度が、ほかの地域の組み合わせに比べて高いことが確認した。このことから、「自粛期間の差が SNS への投稿のい地域差に反映される」という仮説は確かめられなかった。また、投稿内容の遷移とその地域差から、人々がコロナに慣れていき、その速度に地域差があるという考察を行った。

今後の課題としては、手法に関する課題と分析に関する課題がそれぞれ存在する。

提案手法を用いることで、文章集合間の定量的な比較が可能になったが、実際の文章集合間の違いが、算出された数値にどの程度反映されているか不明瞭であるという問題点がある。すなわち、数値の大小比較によって、どこに違いがあるのかを検出することは可能だが、どの程度の違いであるかが不明瞭である。これは、本手法では LDA モデルの次元数を固定しておらず、そのため、ディリクレ分布のパラメータや KL 情報量の差異が次元数の影響をどのように反映するかを、本稿の段階では加味できていないからである。

コロナ禍における地域差の分析としては、本手法では評価しきれない差異が存在することが挙げられる。単語レベルではなく、LDA では考慮できない文脈レベルの情報に差異が存在することや、実験に用いた、「コロナ」という文字列を含むツイート以外のツイートに反映された差異や、コロナが日常として受け入れられることで、ツイートとして投稿しなくなっていくというような、ツイートの頻度などが本研究では分析できていない。これらを加味した研究を続けることで、新しい知見の抽出や、本研究の結果の信頼性の向上が期待される。

謝 辞

本研究は、内閣府「COVID-19 AI･シミュレーションプロジェクト」の一環として実施され、JST，CREST，JPMJCR19A4 の支援を受けたものである。

文 献

[1] 四方田 健二, 新型コロナウイルス感染拡大に伴う不安やストレスの実態：, 体育学研究, 2020, 65 巻
[2] 鳥海 不二夫, 榊 剛史, 吉田 光男, ソーシャルメディアを用いた新型コロナ禍における感情変化の分析, 人工知能学会論文誌, 2020, 35 巻, 4 号
[3] 太田 圭亮, 今井 公太郎, 本間 健太郎, ジオタグつき tweet の時空間解析に基づいた地域特性抽出に関する研究, 日本建築学会計画系論文集, 2017, 82 巻, 731 号
[4] Shoji, Y., Takahashi, K., Dürst, M.J., Yamamoto, Y., Ohshima, H. (2018). Location2Vec: Generating Distributed Representation of Location by Using Geo-tagged Microblog Posts. In: Staab, S., Koltsova, O., Ignatov, D. (eds) Social Informatics. SocInfo 2018. Lecture Notes in Computer

Science, vol 11186
[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, Vol. 3, No. Jan, pp. 993–1022, 2003.
[6] T. Minka, Estimating a Dirichlet Distribution, (2000)
[7] Penny, William D. (2001): "KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities"

7 付　　録

表 7: 各期間における推定されたディリクレ分布のパラメータ

(期間１)	α1	α2	α3	α4	α5	α6
東京周辺	0.363494	0.289619	0.305411	0.290014	0.328330	0.306997
その他の地域	0.380394	0.285750	0.296409	0.302301	0.324282	0.323061

(期間2)	α1	α2	α3	α4	α5
東京周辺	0.334931	0.246395	0.350717	0.257636	0.340170
その他の地域	0.351173	0.256509	0.347029	0.259663	0.341299

(期間3)	α1	α2	α3	α4	α5	α6
東京周辺	0.305020	0.261610	0.302330	0.328980	0.325690	0.282950
その他の地域	0.315600	0.260922	0.303287	0.316674	0.331357	0.284411

表 8: 期間１における各トピックの頻出語

Topic1	見る　今日　私　食べる　会う　行く　みんな　休み　買う　家　自粛　頑張る　終わる
Topic2	マスク 日本 中国 手洗い 手 消毒 対策 アルコール うがい 買う 国 トイレットペーパー
Topic3	影響 仕事 会社 時間 授業 店 出勤 減る 大学 オンライン
Topic4	病院 検査 来る 感染 怖い 症状 かかる 私 熱 患者 風邪 電話
Topic5	自分　死ぬ　それ　考える　何　行動　命　批判　感染　話　医療
Topic6	中止　開催　延期　イベント　ライブ　チケット　公演　数　オリンピック

表 9: 期間２における各トピックの頻出語

Topic1	今年　ライブ　行く　イベント　コロナ禍　開催　予定　会える　楽しい　影響
Topic2	見る　友達　学校　私　実家　子供　授業　大学　オンライン　先生　仕事
Topic3	感染　日本　増える　対策　感染者　マスク　経済　国　多い　政府　国民
Topic4	病院　熱　入院　休む　検査　電話　ワクチン　痛い　症状　影響　陰性　陽性
Topic5	買う　食べる　今日　マスク　店　飲む　影響　時間　最近　電車　コロナ禍　笑

表 10: 期間３における各トピックの頻出語

Topic1	お願い、コロナ禍、買う、協力、お客様、対策　皆様、防止、お知らせ、消毒、営業
Topic2	検査、陽性、今日、株、陰性、ＰＣＲ検査、病院体調、濃厚接触者、結果、保健所
Topic3	見る、ライブ、会う、みんな、楽しい、マスク　くん、いける、私、観る、好き、友達、頑張る
Topic4	オリンピック、日本、感染、国民、政府、国　増える、開催、感染者数、ニュース
Topic5	仕事、食べる、帰る、家、休み、飲む、店、学校　時間、バイト、旅行、影響、実家、明日
Topic6	ワクチン、接種、鬱、回目、熱、痛い、予約　副反応、腕、風邪、注射、発熱、頭痛

移動軌跡とそれに対応づく地理空間データを活用した人の移動特性の理解

須藤　龍一†　　戸田　浩之†

† 横浜市立大学データサイエンス学部データサイエンス学科　〒 236–0027 神奈川県横浜市金沢区瀬戸 22–2
E-mail: †{d204032d,toda.hir.xg}@yokohama-cu.ac.jp

あらまし　実社会における人の移動の特徴を理解することは，人々に快適なナビゲーションを提供する上で重要である．移動の特徴の例としては,「運転時に右左折の少ない広い道を通る」や「自然豊かで，ゆったりとした散歩道を通る」などが考えられる．近年，移動軌跡データを用いることで人の移動を分析，評価する研究が行われており，それらの研究より，人の移動を理解する上では,「移動対象がもつ動作特性（速度や進行方向の系列情報）」と「移動経路がもつ地理空間特性（道路や周辺施設の情報）」の二つの観点が重要であることがわかる．そこで，本研究では，移動軌跡とそれに対応する GIS（Geographic Information System）データから移動に関する動作特性と地理空間特性を取得し，ユーザの移動を表現することができるのか調査する．さらに，各ユーザを正解ラベルとする多クラス問題を用いることで，取得した移動特性のユーザ毎の移動の表現における有効性の評価を行う．

キーワード　地理データ活用，行動データ，モビリティ，移動軌跡，ユーザ支援

1　はじめに

近年，位置情報や地図を活用して，ユーザが移動する際に適切な情報を提供するマップアプリケーションは大きく発展してきている．例えば，リアルタイムの交通状況を考慮したカーナビゲーションシステムや経路周辺の店舗状況を提供する歩行者向けマップアプリケーション等，状況に合わせたナビゲーションが可能となっている．しかしながら，ユーザの移動に関する特徴は多種多様である．例えば,「運転時に右左折の少ない広い道を好む」や「自然豊かで，ゆっくりとした散歩を好む」等が考えられるが，このような各個人の移動の特徴に対応したナビゲーションは実現していない．上記のようなナビゲーションを実現するためには，ユーザの移動の特徴を理解し，それぞれに適応した経路を探索する必要がある．本研究では，このような人の移動を表す特徴を定量化したものを「移動特性」と定義する．

近年では，移動軌跡を用いて，様々な観点から，人の移動特性を理解する研究が行われている [?]．これらの研究では，シチュエーションに特化したユーザの特性抽出・分類や，人の流れを考慮した都市開発への応用が可能となっているが，個人単位の移動特性を取得し，明示的に表現することはできていない．このような課題に対して，本研究では,「ユーザの移動を表現・理解するための特性とは何か.」という Research Question（RQ）を設定し，人の移動を特徴づける移動特性について調査し，ユーザごとの移動を表現・理解できるか評価する．RQ に関連して，Dong らの研究により，ユーザの行動は,「速度の変化, 進行方向の変化」等の時系列情報や「移動時間, 移動距離」等の全体の移動傾向といった，ユーザの動きに関する特性から特徴づけられることが示されている [?]．さらに，西村らにより，ユーザは移動経路に対して,「広い, 混雑していない, 信号が少ない」等の道路に関する要望や,「コンビニや店がある, 景色がきれい」等の経路周辺の施設や自然環境に関する要望を持つことが示されている [?]．これらの研究から，人の移動は次の二つの観点から表現できると考えられる．一つは，移動時の対象の動きに関係する「移動対象のもつ動作特性」であり，もう一つは，移動時の経路とその周辺環境に関係する「移動経路のもつ地理空間特性」である．これら二つの観点を考慮することで，どのような場所を，どのように移動したいか示す移動特性の表現が可能になると考えられる．

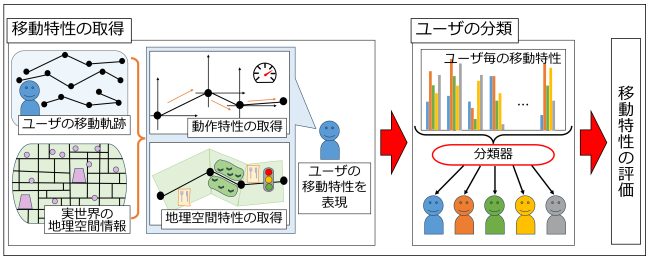


図 1　人の移動特性の理解の概観

そこで本研究では，RQ に対して，移動軌跡とその移動した場所に対応する GIS(Geographic Information System) データから取得可能な「移動時の対象の動きに関係する**移動対象のもつ動作特性**と移動時の経路と周辺環境に関係する**移動経路のもつ地理空間特性**でユーザの移動特性を表現できるのではないか.」という仮説を立てる．そして，この仮説に基づき取得した移動特性（動作特性と地理空間特性）が，人の移動の特徴をどの程度表現しているのかを分析するために，複数のユーザの移動軌跡からなる移動軌跡データセットを，ユーザ毎に分類する多クラス問題に適用する．これより，その分類精度との関係を

元に，それぞれの特徴量の有効性を検討する．本研究の主な貢献は以下の通りである．

- 移動軌跡と，その経路上および経路周辺にある地物の情報を元に，人の移動を表現すると考えられる動作特性と地理空間特性を取得する手法を提案した点．
- 提案に基づき取得した移動特性の有効性の評価および人の移動を表現する上で重要となる特徴の明示を実施した点．

以下に本稿の構成を述べる．2 節に関連研究を示し，本研究の位置づけについて整理する．3 節にて本研究における用語と問題について定義する．4 節では，移動の「動作特性」と「地理空間特性」から，ユーザの移動特性を取得する手法について説明する．5 節にて，本研究で用いるデータと実験について説明し，6 節にて実験の結果と考察を述べる．最後に，7 節で結論と今後の予定について述べる．

2 関 連 研 究

本節では,「移動対象のもつ動作特性」と「移動経路のもつ地理空間特性」の 2 つの観点に関連する移動軌跡を対象とした研究について述べる．

2.1 動作特性に関する研究

He らは，移動軌跡の進行方向の変化から移動軌跡を区間に分割し，他の多くの移動軌跡から生成した区間ごとの軌跡との類似性を評価することで，異常な動きをしている軌跡を検出する手法を提案している [?]．この研究では異常な行動を検知するために，移動軌跡から取得できる移動対象の移動時間や進行方向の変化を取得している．この研究は，本研究において，人の移動の動作特性を表現する特徴の設計の参考になる．Yao らは，タクシーの移動軌跡から，ドライバーの動作特性（縦方向・横方向における速度や加速度の変化）を取得することで，ドライバーの運転時の挙動を分類する手法を提案している．その際，2 つの時系列データ間の各点の距離を総当たりで求める Dynamic Time Warping（DTW）を用いて，長さの違う軌跡間の類似度を求めている [?]．この研究では，取得した動作特性の時系列パターンから，軌跡間の類似性を測ることで，ドライバーの運転行動の分類を行っている．よって，この研究は，各移動がもつ時系列情報を考慮した動作特性の取得と，人の移動を識別するタスクの設計の参考になる．しかし，DTW には，各点におけるノイズやサンプリングレートの低さによって，類似度測定の精度が低下する点や，計算量が 2 乗で増えていくという問題が挙げられる．

これらの問題に対して，長さが可変である移動軌跡を固定次元のベクトルとして表現するために，リカレントニューラルネットワーク（以下，RNN）の活用が提案されている [?] [?]．この手法では，移動軌跡の測位点の系列を RNN に入力することで、移動の規則性や時間的パターンをとらえることができることから、移動軌跡の予測等のタスクで幅広く使われている [?]．Li らは，等しい大きさに分割した地理空間に移動軌跡の各点をマッピングし，シーケンスデータに変換することで，RNN の拡張である Gated Recurrent Unit（以下，GRU）を用いて，シーケンスを固定次元ベクトルにエンコードを行う手法を提案している [?]．これより，エンコードされた表現ベクトルを条件に，デコードを行うエンコーダ・デコーダモデルを適用することで，線形時間で軌跡の類似性を計算することが可能となる．Dong らは，自動車の移動軌跡から，速度，加速度，角速度に関する特徴の統計量をセグメントごとに計算し，オートエンコーダにより，ドライバーの運転スタイルを固定次元で表現する手法を提案している [?]．これより，各ドライバーの動作特性に関する時系列パターンの特徴を固定次元で学習し，表現することで，ドライバーの識別を行っている．

2.2 地理空間特性に関する研究

Gao らは，オンライン地図サービスを用いて，移動軌跡に地理的情報を付与し，移動軌跡間の類似度を計算することで，軌跡の幾何学的特性と意味的情報を捉える手法を提案している [?]．この研究は，地域の交通や街中の環境を表す意味的な情報を用いて，移動軌跡の類似性の評価を行っており，人の移動を表現する「地理空間特性」の取得の設計の参考になる．Nair らは，サイクリストの移動軌跡と地域のオープンデータから，社会環境の特徴を取得し，サイクリングの目的予測とルート選択の環境的要因を評価する手法を提案している [?]．この研究では，道路の勾配や自転車施設の存在などに関する情報を軌跡に付与することで，サイクリングにおける移動経路の地理空間特性を表現し，サイクリングルートの評価を行っている．この研究より，移動における地理空間特性の抽出と，そこから考えられるユーザが通った経路の特性の表現が可能となる．

上記で示した人の移動における「動作特性」と「地理空間特性」の取得に関連する研究では，特定のシチュエーション（サイクリングやある特定の道路上の移動など）に特化して，軌跡の類似性の評価や特性の表現に取り組んでいる．また,「動作特性」と「地理空間特性」を両方考慮し，ユーザ単位での移動特性の表現を行うことには取り組まれていない．そこで，本研究では，移動軌跡とそれに対応する GIS（Geographic Information System）データから，動作特性と地理空間特性を取得し，この 2 つの観点からユーザごとの移動を表現できるかについて調査する．

3 問 題 設 定

3.1 用語の定義と前提とするデータ

本研究において，軌跡 $T_i\,(i = 1, 2, ..., n)$ とは，移動する対象の一定時間ごとの測位点 $p_j\,(j = 1, 2, ..., |T_i|)$ のシーケンスと定義し，$T_i = \{p_1, p_2, ..., p_{|T_i|}\}$ と表現される．ここで，n は，移動軌跡の集合に格納される軌跡の本数を表す．また，測位点 $p_j\,(j = 1, 2, ..., |T_i|)$ は空間内の対象の位置座標 (緯度と経度) とその地点を通った時刻のタイムスタンプで $p_j = \{x_j, y_j, t_j\}$ と表される．以上で定義される移動軌跡の集合を移動軌跡データセット TS とし，$TS = \{T_1, T_2, ..., T_n\}$ と表される．また，TS は，特定の u 人のユーザの軌跡で構成されているとし，各軌跡 T_i を

表 1　移動データ M_i の概要

移動軌跡 T_i^{new} の情報	元の移動軌跡を道路ネットワークへ割り当てた測位点の系列
地理空間情報 (GI_i^{way}, GI_i^{env})	移動経路および経路周辺に位置する地理的オブジェクトの情報

ユーザごとの集合に分けることができる．各ユーザのラベルを $l\,(l = 1, 2, ..., u)$ とし，各ユーザごとの移動軌跡集合 TS_l を定義する．このとき，各 TS_l には，任意の $n_l\,(n_l = n_1, n_2, ..., n_u)$ 本の軌跡が格納されており，$\sum_{l=1}^{u} n_l = n$ となる．よって，移動軌跡データセット TS は，$TS = \{\{TS_1\}, \{TS_2\}, ..., \{TS_u\}\}$ と表現することができる．

続いて，GIS データセット GIS は，地理空間内の地理的なオブジェクトとその位置関係を表現する地理空間情報で構成されており，地理空間内の道路ネットワークデータ GIS^{way} $(GIS^{way} \subset GIS)$ と，道路ネットワーク周辺に位置する地物のデータ GIS^{env} $(GIS^{env} \subset GIS)$ に分けることができる．また，各 GIS データは，各オブジェクトの地理的な形状を位置座標を用いて，ポイント，ライン，ポリゴン等の形式で表現した「ジオメトリ情報」と，対象となるオブジェクトの名称や種類などの情報を示した「属性情報」で構成されている．

これより，軌跡の測位点の位置座標と道路ネットワークデータ GIS^{way} のジオメトリ情報の位置座標を対応付けることで，T_i を実世界の道路ネットワーク上に割り当てた新たな移動軌跡 $T_i^{new}\,(i = 1, 2, ..., n)$ を生成する．さらに，T_i^{new} に対応する移動経路の情報 GI_i^{way} $(GI_i^{way} \subset GIS^{way})$ と，GI_i^{way} 周辺に位置する地物の情報 GI_i^{env} $(GI_i^{env} \subset GIS^{env})$ を抽出する．

これより，次節における移動データ $M_i\,(i = 1, 2, ..., n)$ とは，実世界の道路ネットワークへ割り当てた移動軌跡 T_i^{new} の情報とその移動軌跡に対応付けられた地理空間の情報 (GI_i^{way}, GI_i^{env}) で構成されるデータのことを指す（表 1）．よって，$M_i = \{T_i^{new}, (GI_i^{way}, GI_i^{env})\}$ となり，M_i の集合 $MS = \{M_1, M_2, ..., M_n\}$ を定義することができる．また，ユーザごとの移動軌跡の集合 TS_l と対応させることで，$MS = \{\{MS_1\}, \{MS_2\}, ..., \{MS_u\}\}$ と表すことができる．

3.2　問題の定義とアプローチ

以下に，2 節の関連研究から設定した，本研究における RQ と仮説を示す．

- **RQ** ユーザの移動を表現・理解するための特性とは何か．
- **仮説** 移動時の対象の動きに関係する「移動対象のもつ動作特性」と移動時の経路と周辺環境に関係する「移動経路のもつ地理空間特性」でユーザの移動特性を表現できるのではないか．

これより，RQ と仮説に対する本研究の問題定義を示す．

- **入力** 移動軌跡の集合 TS と，GIS データの集合 GIS
- **出力** 特定のユーザ l の移動データの集合 MS_l を特徴づける，ユーザ l の移動特性（動作特性と地理空間特性）

また，取得した移動特性を多クラス分類問題に適用することで，人の移動を理解するうえで重要となる特徴についての評価と考察を行う．

4　移動特性取得の設計

本節では，3.2 節で示したアプローチに基づき，移動データ M_i を用いて，人の移動特性の取得方法について説明する．

4.1　動作特性の取得

まず，次に，移動時の対象の動きに関係する「動作特性」の取得について説明する．本研究では，移動時の対象の動作特性は，対象の動きに関する時系列シーケンスを分割し、時間的依存関係を考慮した「時系列情報」と，対象の動き全体の動向を考慮した「全体情報」から構成されると定義する．

4.1.1　時系列情報の取得

はじめに，動作特性の時系列情報の取得について説明する．まず，T_i^{new} の測位点間の関係から，各軌跡の速度ノルム，加速度ノルム，進行方向の差の 3 つの特徴量の時系列シーケンスを取得する．続いて，上記の時系列シーケンスに対して，固定のセグメント長 L_s を定め，$L_s/2$ の間隔でスライドすることで，時系列シーケンスを長さ L_s のセグメント $seg_{ij}\,(j = 1, 2, ..., q)$ に分割する（図 2）．ここで，q は，時系列シーケンスのセグメント数を表す．

次に，各セグメントの特徴を圧縮した潜在表現を取得するオートエンコーダモデルの学習について示す．はじめに，各セグメント seg_{ij} を 2 層の積み上げ GRU 層と全結合層を用いたセグメントエンコーダ（図 3 の gru1+gru2+dropout+fc1）に入力する．第 1 の GRU 層（gru1）は，入力した seg_{ij} を L_s の時間間隔で展開しながら読み込み，第 2 の GRU 層（gru2）が追加され，gru2 の隠れ層数と等しい次元のベクトルを出力する．また，過学習を減らすためにドロップアウト層（dropout）が適用される．全結合のボトルネック層（fc1）は，seg_{ij} の圧縮された潜在表現 $s_{ij}\,(j = 1, 2, ..., q)$ を学習するために使用される．続いて，セグメントデコーダ（図 3 の gru1+gru2+fc1+fc2）を用いて，元のセグメント seg_{ij} の再構築を行う．エンコードされた s_{ij} を 2 層の GRU 層（gru1+gru2）と 1 つ目の全結合層（fc1）で学習し，2 つ目の全結合層（fc2）により，seg_{ij} を再構築した $\hat{seg}_{ij}$ を出力する．このとき，全結合層には，Relu 関数を適用し，損失関数は，平均二乗誤差（MSE）を使用する．

$$\mathrm{MSE} = \frac{1}{m}\sum_{j=1}^{m} \hat{seg}_j - seg_j \qquad (1)$$

このとき，m は学習に用いるセグメントの数，$\hat{seg}_j$ は seg_j に対応する予測値（デコードされた値）を表す．

最後に，セグメントレベルの潜在表現を軌跡単位で結合する．ここでは，オートエンコーダモデルにおけるセグメントエンコーダ（図 4 の gru1+gru2+dropout+fc1）を用いて，seg_{ij} の圧縮された表現 s_{ij} を正規化し，その和を取ることで，軌跡単位の固定次元ベクトル $B_i^{time}\,(i = 1, 2, ..., n)$ を取得する．

$$B_i^{time} = \frac{\sigma_i}{\max_j\{\sigma_{ij}\}} \qquad (2)$$

ここで，σ_i は，セグメントレベルの潜在表現 s_{ij} の和を示し，$\sigma_i = \sum_{j=1}^{q} s_{ij}$ となる．このとき，q は時系列シーケンス内のセグメント数を示し，σ_{ij} はその j 番目のセグメントの次元 ($\sigma_{ij} \geqq 0, \forall T_i^{new}, j$) を示す．

これより，対象の動きの時系列変化を考慮した 3 つの特徴量の固定次元ベクトル B_i^{time} を取得することができる．

4.1.2 全体情報の取得

動作特性の全体情報は，T_i^{new} から取得する軌跡全体の動向や傾向を表現する指標のことを指す．まず，移動時間，移動距離，平均の速さ，軌跡の x,y 座標の最小値と最大値の組合せから作成した軌跡の外接矩形の各辺の長さと面積を取得する．また，三浦らより，自動車の走行経路の右左折の数が運転時の移動時間や安全性に影響を与えることが示された [?]．このことから，移動時の右折，左折，直進の割合は，人の移動を特徴づける上で有効な特徴量であると考えることができる．よって，本研究では，上記の特徴量に加えて，軌跡全体のセグメント数に対する右左折数と直進数の割合を新たな特徴量として定義する．具体的には，T_i^{new} の進行方向の時系列変化から，変化の角度が 30 度より大きくなっている回数を左右で求め，$|T_i^{new}|$ に対する割合を計算する．これらの特徴量を，各移動の全体的な動きを表現する指標として定義し，B_i^{global} とする．

これらのことから，時系列情報 B_i^{time} と，全体情報 B_i^{global} を合わせたものを移動対象の動きを表現する指標として,「動作特性 $BF_i\,(i = 1, 2, ..., n)$」とする．

4.2 地理空間特性の取得

次に，移動時の経路と周辺環境に関係する「地理空間特性」の取得について説明する．地理空間特性は，移動データ M_i 中の地理空間情報 (GI_i^{way}, GI_i^{env}) から取得する．以下に，本研究で考える地理空間特性を表す指標の取得について示す．

- **道路情報 (G_i^{road})**　GI_i^{way} に含まれる道路の種類（高速道路，主要道路，生活道路など）を示す属性情報から，移動経路全体に対する各道路種の割合を取得する．
- **交差点情報 (G_i^{cross})**　GI_i^{env} に含まれる交差点の属性情報 (信号，横断歩道など) から，単位距離あたりの交差点種別の個数を取得する．
- **POI 情報 (G_i^{poi})**　GI_i^{env} に含まれる POI の属性情報 (公共機関やレジャー施設，商業店舗など) から，単位距離あたりの POI 種別の個数を取得する．
- **土地利用情報 ($G_i^{landuse}$)**　GI_i^{env} に含まれる土地利用に関する属性情報 (工業地域や農地，住宅地域など) から，単位距離あたりの土地利用種別の個数を取得する．

これより，上記で取得した 4 つの特徴量 ($G_i^{road}, G_i^{cross}, G_i^{poi}, G_i^{landuse}$) を合わせたものを，ユーザがどのような特徴を持つ道路と周辺環境を通る傾向にあるか示す指標として,「地理空間特性 $GF_i\,(i = 1, 2, ..., n)$」とする．

4.3 ユーザの移動特性の取得

最後に，取得した M_i の「動作特性 BF_i」と「地理空間特性 GF_i」からなる移動特性を $MF_i\,(i = 1, 2, ..., n)$ と定義し，移動特性データセット $MFS = \{MF_1, MF_2, ..., MF_n\}$ を作成する．本研究の目的は，人の移動を特徴づける移動特性について調査し，ユーザごとの移動を表現・理解できるか評価を行うことである．よって，MFS をユーザごとの移動特性集合に分割し，$MFS = \{\{MFS_1\}, \{MFS_2\}, ..., \{MFS_u\}\}$（$u$ はユーザの人数）として，個々のユーザに紐づけることで，ユーザごとの移動特性を取得する．これより，MFS_l と各ユーザラベル l を多クラス分類問題に適用し，ユーザごとの移動を表現するうえで有効な特徴量について評価と考察を行う．

5 実　　験

4 節より取得した移動特性がユーザごとの移動の特徴を表現しているのならば，その移動特性から，ユーザの識別を行ことができると考えられる．よって，本節では，取得した移動特性を，各ユーザを正解ラベルとする多クラス分類問題に適用することで,「取得した特性からユーザごとの移動を表現することができているか」,「個人の移動を表現するにあたり，どの特徴量が有効なのか」について評価を行う．

5.1 データ準備

まず，実験で用いるデータについて説明する．本研究の評価実験で用いるデータは，大別して「移動軌跡データセット TS」と「GIS データセット GIS」の二種類である．まず，移動軌跡は CML/PKDD 2015 Discovery Challenges[1]で活用されたポルトガルのポルトを走るタクシーの移動軌跡を使用する [?]．データセットには，2013 年 7 月 1 日から 2014 年 6 月 30 日までの 442 台分のタクシーの移動軌跡が約 170 万本格納されており，各タクシーの ID，日時，乗降場所，15 秒ごとの緯度・経度の時系列変化，乗車種類，デイタイプ（休日かどうか）が示されている．次いで，GIS データは，NextGIS[2]が公開する上記の移動軌跡に対応する 2014 年のポルトガルのポルトのものを使用する．GIS データには，ポルトの道路，POI，土地利用などの地理空間情報が格納されている．

次に，移動軌跡 $T_i\,(i = 1, 2, ..., n)$ をポルトの道路ネットワークへ割り当てる [?]．これにより，測位点のノイズとサンプリングレートを調整し，実際の道路ネットワークを反映した軌跡の生成を行う（図 5）．

はじめに，GIS データに含まれる道路ネットワークデータ GIS^{way} から，ポルトの道路のノードとエッジを抽出する．軌跡 T_i の各測位点 $p_j\,(j = 1, 2, ..., |T_i|)$ から最も近いエッジに垂線を下ろし，垂線とエッジの交点を p_j に対応する近似点 p_j^{new} とする．これより，各近似点 p_j^{new} をつなげることで，$T_i\,(i = 1, 2, ..., n)$ と対応する新たな移動軌跡 T_i^{new} を実世界の道路上に生成する．この操作を「マップマッチング」と呼ぶ．

次に，T_i^{new} に対応する，移動経路の情報 GI_i^{way} ($GI_i^{way} \subset$

1：ECML/PKDD 2015 Discovery Challenges https://ceur-ws.org/Vol-1526/
2：NEXT GIS Data https://nextgis.com/

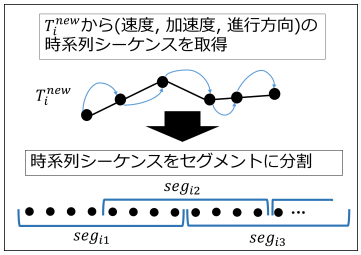


図 2　セグメント取得の概観

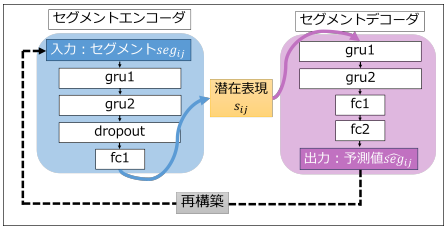


図 3　オートエンコーダモデルの概観

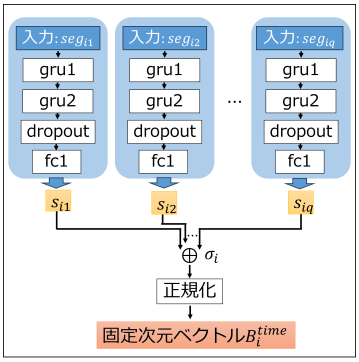


図 4　固定次元ベクトル取得の概観

表 2　移動特性の概観

観点	指標	特徴量
動作特性 BF_i	時系列情報 B_i^{time}	速度，加速度，進行方向の時間変化を考慮した特徴表現ベクトル
	全体情報 B_i^{global}	軌跡全体の時間, 距離, 速さ, 形状, 右左折数
地理空間特性 GF_i	道路情報 G_i^{road}	移動経路全体に対する各種類の割合
	交差点情報 G_i^{cross} POI 情報 G_i^{poi} 土地利用情報 $G_i^{landuse}$	移動経路全体に対する単位距離あたりの個数

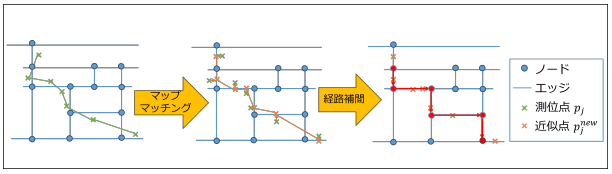


図 5　マップマッチングと経路補間の概観

GIS^{way}) を取得する．マップマッチングで生成した近似点 p_j^{new} $(j = 1, 2, ..., |T_i|)$ と最も近いノードを GI_i^{way} から抽出し，各 p_j^{new} に対応するノードをつなぐ最短経路を道路ネットワーク上で探索する．最後に，探索した最短経路をつなげることで，サンプリングレートを補った移動経路 $T_i^{new'}$ $(i = 1, 2, ..., n)$ を作成する．この操作を「経路補間」と呼ぶ．

この $T_i^{new'}$ の近傍周辺（今回は 40m 以内とした）を探索することで，移動経路周辺に位置する地物の情報 GI_i^{env} $(GI_i^{env} \subset GIS^{env})$ を取得する．

これより，元の各軌跡 T_i に対応する移動データ $M_i = \{T_i^{new}, (GI_i^{way}, GI_i^{env})\}$ を作成する．また，今回は，提案手法の学習と評価において，T_i^{new} の測位点の数が，83 個以上（移動時間が 20 分以上）のものを抽出し，2014 年 5 月 1 日以前のものを学習データ，2014 年 5 月 1 日以降のものを評価データとして用いる．その結果，学習と評価で用いるのは，431 台分の約 15 万個の移動データとなる．

5.2　移動特性の取得

本節では，5.1 節で説明したデータを活用し，4 節で説明した移動特性の取得の実装について説明する．

5.2.1　動作特性の取得

まず，移動データ M_i の移動軌跡 T_i^{new} の情報を用いた，動作特性 BF_i の取得について説明する．

はじめに，「時系列情報 B_i^{time}」を取得する．T_i^{new} の 3 つの時系列シーケンス（速度ノルム，加速度ノルム，進行方向の差）をセグメント seg_{ij} に分割する．今回は，$L_s = 8$ とし，図 3 のオートエンコーダに入力する．セグメントエンコーダ（gru1+gru2+dropout+fc1）における，GRU の隠れ層を 256 層，FC 層の隠れ層を 10 層とし，ドロップアウトの確率を 0.5 と設定する．よって，オートエンコーダの学習する最終的な潜在表現 s_{ij} は 10×8 次元となる．また，これらの学習には，ADADELTA オプティマイザ [?] を使用し，$\lambda = 1\epsilon - 5$，学習率 0.01，$\rho = 0.95$，$\varepsilon = 1\epsilon - 8$，バッチサイズは 128 とした．これより，431 台分の学習データと評価データを用いて，オートエンコーダを学習し，各 M_i の動作特性の時系列情報 B_i^{time} を取得する．このとき，B_i^{time} は，3 つの時系列シーケンス（速度，加速度，進行方向の差）をそれぞれ 10 次元で埋め込んだ固定次元ベクトルであり，特徴量の数は $10 \times 3 = 30$ である．

次に，「全体情報 B_i^{global}」を取得する．これは，T_i^{new} の時間，距離，速さ，外接矩形の各辺の長さと面積，T_i^{new} のセグメント数に対する右左折，直進の割合を示した 9 つの特徴量で表現される．この結果，移動の動作特性 BF_i は，30+9=39 個の特徴量で表される．

5.2.2　地理空間特性の取得

次に，移動データ M_i $(i = 1, 2, ..., n)$ の地理空間情報 (GI_i^{way}, GI_i^{env}) $(i = 1, 2, ..., n)$ を用いた，地理空間特性の取得について説明する．表 3 に，地理空間特性 GF_i を構成する各特徴の詳細を示す．まず，道路情報として，移動経路全体に対する各道路の種類の割合を示した特徴量 G_i^{road} を取得する．ここで，各道路は，大きく 5 つの種類に分けられる．続いて，交差点，POI，土地利用の情報を抽出し，$T_i^{new'}$ に対する，単位距離あたり（今回は 100m あたりとした）のデータ数を示す特徴

表 3　地理空間特性に格納される情報

地理空間特性の指標	指標を構成する特徴量
道路情報 G_i^{road}	高速・幹線，主要，生活， 未舗装・小道，その他
交差点情報 G_i^{cross}	信号，交通島，標識，横断歩道，その他
POI 情報 G_i^{poi}	飲食店，教育，交通，公共機関， 屋外設備，構造物，商業施設，レジャー， 観光施設，スポーツ関連，その他
土地利用情報 $G_i^{landuse}$	草原，住宅関連，園芸・農業， 産業・建設，自然，その他

量 (G_i^{cross}, G_i^{poi}, $G_i^{landuse}$) を取得する．ここで，交差点は 5，POI は 11，土地利用は 6 つの種類にそれぞれ分けられる．このとき，それぞれの特徴量の振り分けは，ポルトの GIS データセットサイズに対して格納されている各属性の数と，属性間の相互関係から判断して行っている．この結果，移動の地理空間特性 GF_i は，5+5+11+6=27 個の特徴量で表される．

以上より，今回の実験では，1 回の移動を特徴づける移動特性 MF_i は，39 次元の動作特性 BF_i と 27 次元の地理空間特性 GF_i で表現されることになる．

最後に，MF_i $(i = 1, 2, ..., n)$ を 431 台のタクシーごとの移動特性の集合 MFS_l $(l = 1, 2, ..., 431)$ に分割することで，ドライバーごとの移動特性の集合を作成する．

5.3　移動特性の有効性の評価

今回は，実験データのドライバーの移動を表現するうえで重要となる特性は何か，決定木モデルの一種である勾配ブースティング木（以下，GBDT）を用いたドライバー識別タスクの結果より分析する．GBDT は，ドライバーの識別において既存研究で活用されており [?]，また，分類に貢献した特徴量の分析ができることより採用した．よって，GBDT を用いて，各タクシーの ID を正解ラベルとしたドライバー識別を実施し，その精度を分析することで，取得した移動特性がドライバーごとに移動を特徴づけているかどうか評価する．さらに，ドライバー識別を行う際の特徴量の寄与度を調査することで，ユーザごとの移動の特徴づけにおける各特徴量の重要度についても考察を行う．今回は 2013 年 7 月 1 日から 2014 年 6 月 30 日までの移動特性のデータを，2014 年 5 月 1 日を基準に分割し，学習データと評価データとした．このとき，学習データと評価データの割合は約 8:2 となっている．また，ドライバー識別タスク用に，431 台分の移動特性データセット MFS から，任意に 10 台分ランダムに抽出したものを 2 つ (SetA, SetB) と，5 台分ランダムに抽出したものを 2 つ（SetC, SetD）用意し，合計 4 つの評価用データセットを作成し，評価指標としては，予測したラベルが，正解ラベルをどれだけ再現できているかを 0 から 1 の値で示す正解率 accuracy を用いる（式 4）．

$$\mathrm{countcorrect} = \sum_{i=1}^{len(datas)} \begin{cases} 1, & \text{if } \mathrm{label}_i = \mathrm{predictlabel}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\mathrm{accuracy} = \frac{\mathrm{countcorrect}}{len(datas)} \tag{4}$$

このとき，$len(datas)$ は評価に用いるデータの数，label_i はデータの正解ラベル（今回は，ドライバー ID）を示し，$\mathrm{predictlabel}_i$ は分類器が予測したラベルを示す．ここで，$i \in \{1, 2, ..., u\}$ である．

5.3.1　比 較 手 法

実験では，ドライバー識別精度の基準として，ドライバーの運転スタイルから，ドライバーを識別する手法である「ARNet [?]」を比較手法とする．ARNet は，各ドライバーの軌跡を入力し，運転スタイルに関する時間的依存関係を考慮した統計的特徴行列を，オートエンコーダを用いて固定次元のベクトルで表現し，学習することで，ドライバーの予測ラベルを出力するといったモデルである．これは，提案する動作特性の取得における「時系列情報 B_i^{time}」の取得で示したオートエンコーダモデルの設計の元になっており，図 3 のオートエンコーダモデルにクラス分類のための全結合層を追加した構成となっている．そのため，ARNet の損失関数は，本研究とは異なり，ドライバーのクラス分類を学習するためのソフトマックス回帰が用いられている．これより，入力を T_i^{new} の統計的特徴行列，出力を各ドライバーの ID とする ARNet を本研究における評価の基準とし，比較することで，取得した移動特性の有効性を評価する．

6　結果と考察

6.1　結　　果

実験では，取得した移動特性 MF_i から，各指標の特徴量 ($B_i^{time}, B_i^{global}, G_i^{road}, G_i^{cross}, G_i^{poi}, G_i^{landuse}$) を除いた場合，正解率がどの程度変化するかについても比較を行った．表**??**に，4 つの評価用データセットに対する正解率（accuracy）を示す．

まず，ARNet を比べると，取得した移動特性を用いた場合の方が正解率が高いことが読み取れる．取得した移動特性を全て用いたときの正解率は，ARNet の正解率を比べると，ドライバーの人数が 10 人のセット（SetA, SetB）で約 2 倍，ドライバーの人数が 5 人のセット（SetC, SetD）で約 1.5 倍，正解率が高くなっていることがわかる．また，MF_i から各指標の特徴量を除いた場合の，正解率の変化を比較すると，「動作特性の時系列情報（B_i^{time}）」を除いたとき，最も正解率が低下することがわかる．続いて，図**??**は，ドライバー識別の寄与度が高い特徴量を降順に 10 個，評価用データセットごとに示したものである．これより，全データセットにおいて，動作特性の特徴量の多くがグラフに示されていることがわかる．観点別にみると，動作特性では，速度の時系列情報と，軌跡の外接矩形に関する特徴量の多くが上位に位置していることがわかる．地理空間特性では，主要道路の割合と，単位距離当たりの公共機関の

表 4　MF_i を用いた GBDT と ARNet の accuracy

	SetA(10 人)	SetB(10 人)	SetC(5 人)	SetD(5 人)
MF_i	0.5143	0.4885	0.6683	0.4727
$-B_i^{time}$	0.4793	0.4405	0.6444	0.4551
$-B_i^{global}$	0.4621	0.4346	0.6111	0.4141
$-G_i^{road}$	0.5029	0.4730	0.6762	0.4590
$-G_i^{cross}$	0.5137	0.4612	0.6635	0.4824
$-G_i^{poi}$	0.4647	0.4435	0.6413	0.4707
$-G_i^{landuse}$	0.4959	0.4560	0.6683	0.4375
ARNet	0.2300	0.2058	0.4531	0.2991

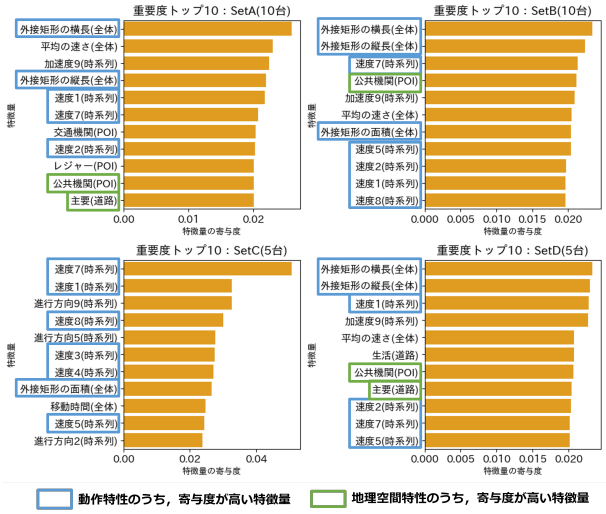


図 6　各評価用データセットにおける GBDT で重要度の高い特徴量トップ 10

個数に関する特徴量が主に位置しており，交差点と土地利用に関する特徴量はグラフには示されていないことが読み取れる．

6.2　考　　察

結果より，ドライバーの識別タスクにおいて，取得した移動特性 MF_i を説明変数として用いたときの GBDT の精度は比較手法に比べて大きく上回ることがわかった．よって，本研究で提示する「動作特性 BF_i」と「地理空間特性 GF_i」の特徴量は，人の移動を表現するうえで有効な特性であることが考えられる．これより，ユーザの普段の行動から，本研究で提示する移動特性を取得することで，ユーザごとの移動の特徴を表現でき，未知の場所や非定常時の移動予測や経路推薦等への応用が可能となると考えられる．また，MF_i から，B_i^{time} を除いた場合，ドライバー識別の精度が最も低下したことから，ユーザごとの移動とを特徴づけるうえで,「動作特性の時系列情報 B_i^{time}」の重要度が高いことが考えられる．

さらに，図**??**の重要度の結果から，ドライバーの移動特性を特徴づけ，表現するうえでは，動作特性の方が重要視されることが考えられる．また，各観点において，動作特性は「移動速度の時間変化」と「軌跡全体の形状」について，地理空間特性は「主要道路の利用率」と「経路周辺の公共機関の数」についてが，ドライバーごとの移動を特徴づけるうえで重要であると考えられる．これより，本研究の実験において考えられる調査結果を示す．

- ドライバーごとの移動を表現するうえで,「移動対象のもつ動作特性」と「移動経路のもつ地理空間特性」は有用であり，特に，動作特性の方が重要度が高い．
- 観点ごとの有用な特徴量としては，

動作特性　「移動速度の時間変化，軌跡全体の形状」
地理空間特性　「主要道路の利用率，経路周辺の公共機関の数」
が挙げられる．

7　さいごに

7.1　結　　論

本研究では,「ユーザの移動を表現・理解するための特性とは何か.」という RQ を設定し，これに基づき,「移動時の対象の動きに関係する**移動対象のもつ動作特性**と移動時の経路と周辺環境に関係する**移動経路のもつ地理空間特性**でユーザの移動特性を表現できるのではないか.」という仮説を立てた．これらに対して，取得した移動特性の各特徴量を，ユーザ毎に分類する多クラス問題に適用することで，ユーザごとの移動の表現における有効性を明らかにした．本研究における実験結果から，今回実験で使用したドライバーごとの移動を表現するうえで,「移動対象のもつ動作特性」と「移動経路のもつ地理空間特性」は有用であることがわかった．また，ドライバーごとの移動を特徴づけるうえで，地理空間特性に比べて動作特性の方が重要度が高く，観点別にみると，動作特性は「移動速度の時間変化，軌跡全体の形状」に関する特徴量，地理空間特性は「主要道路の利用率，経路周辺の公共機関の数」に関する特徴量の重要度が高いことが考えられる．これより，本研究の実験を通じて，人の移動を表現・理解において，本研究が提示する移動特性の有効性と，各特徴量の重要度について評価することができた．よって，本研究の移動特性を用いて，各個人の移動を表現することで，個別化したナビゲーションシステムやマップアプリケーションシステムの構築が可能になると考えられる．

7.2　今後の予定

以下に，提案する移動特性表現モデルの改良における今後の予定を示す．まず，他の移動軌跡データセットで同様の評価を行うことで，データセットによる特性ごとの効果の相違を明らかにする予定である．続いて，ユーザごとの移動の特徴を明示的に示す代表ベクトル等の取得を行う．これは，ユーザの移動特性を構成する各特徴量の相互関係や影響度を考慮することで，ユーザ毎の移動特性を圧縮した潜在表現の作成を行う予定である．

また，本研究の将来的な目標として，各ユーザの移動特性を対応させた，ユーザ単位のナビゲーションシステムを構築することが挙げられる．よって，ユーザの移動を正確に識別しているかどうかの判定だけでなく，出発地と目的地を固定した異なる任意の移動経路を用意したとき，ユーザの移動特性から，適切な経路を推薦できるかどうかを判定するといった評価手法の

構築も行っていく予定である．これより，ユーザごとの移動の特徴を考慮する個別化したナビゲーションシステムの構築を行っていきたい．

文　献

[1] Alessandro Crivellari and Euro Beinat. Lstm-based deep learning model for predicting individual mobility traces of short-term foreign tourists. *Sustainability*, Vol. 12, No. 1, 2020.

[2] Weishan Dong, Ting Yuan, Kai Yang, Changsheng Li, and Shilei Zhang. Autoencoder regularized network for driving style representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, p. 1603–1609. AAAI Press, 2017.

[3] Yuki Endo, Kyosuke Nishida, Hiroyuki Toda, and Hiroshi Sawada. Predicting destinations from partial trajectories using recurrent neural network. In Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, Cham, 2017. Springer International Publishing.

[4] Chongming Gao, Zhong Zhang, Chen Huang, Hongzhi Yin, Qinli Yang, and Junming Shao. Semantic trajectory representation and retrieval via hierarchical embedding. *Information Sciences*, Vol. 538, pp. 176–192, 2020.

[5] Ling He, Xinzheng Niu, Ting Chen, Kejin Mei, and Mao Li. Spatio-temporal trajectory anomaly detection based on common sub-sequence. *Applied Intelligence*, Vol. 52, No. 7, p. 7599–7621, may 2022.

[6] Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S. Jensen, and Wei Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 617–628, 2018.

[7] Sobhan Moosavi, Pravar Dilip Mahajan, Srinivasan Parthasarathy, Colleen Saunders-Chukwu, and Rajiv Ramnath. Driving style representation in convolutional recurrent neural network model of driver identification. *ArXiv*, Vol. abs/2102.05843, , 2021.

[8] Luis Moreira-Matias, João Gama, Michel Ferreira, João Mendes-Moreira, and Luis Damas. Predicting taxi - passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 3, 2013.

[9] Suraj Nair, Kiran Javkar, Jiahui Wu, and Vanessa Frias-Martinez. Understanding cycling trip purpose and route choice using gps traces and open data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 3, No. 1, mar 2019.

[10] Takuya Nishimura, Kyosuke Nishida, Hiroyuki Toda, and Hiroshi Sawada. Social media knows what road it is: quantifying road characteristics with geo-tagged posts. *Social Netw. Analys. Mining*, Vol. 7, No. 1, 2017.

[11] Sheng Wang, Zhifeng Bao, J. Shane Culpepper, and Gao Cong. A survey on trajectory data management, analytics, and learning. *ACM Comput. Surv.*, Vol. 54, No. 2, mar 2021.

[12] Ying Yao, Xiaohua Zhao, Yiping Wu, Yunlong Zhang, and Jian Rong. Clustering driver behavior using dynamic time warping and hidden markov model. *Journal of Intelligent Transportation Systems*, Vol. 25, No. 3, pp. 249–262, 2021.

[13] Silin Zhou, Jing Li, Hao Wang, Shuo Shang, and Peng Han. Grlstm: Trajectory similarity computation with graph-based residual lstm. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[14] 英俊三浦, 伸哉柏木. 格子状道路網における交差・合流を考慮した自動車の経路の割り当て. 日本オペレーションズ・リサーチ学会和文論文誌, Vol. 62, No. 0, pp. 54–70, 2019.

[15] 山本萌美. マップマッチングのアルゴリズム. `http://bin.t.u-tokyo.ac.jp/startup15/file/0529yamamoto.pdf`, 9 2023.

機械学習を用いた旅行活動ツイート判定精度の可視化視認性への影響

藤大　友都† 　服部　峻†† 　砂山　渡††

† 滋賀県立大学大学院 工学研究科 電子システム工学専攻　〒 552–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学 先端工学研究院　〒 552–8533 滋賀県彦根市八坂町 2500
E-mail: †on23yfujidai@ec.usp.ac.jp, ††{hattori.s,sunayama.w}@e.usp.ac.jp

あらまし　旅行活動を SNS に投稿した経験があるユーザは多い．このような旅行活動データの利活用の一つとして可視化が考えられ，新たな旅行先の決定や旅行需要の発見に寄与すると考える．しかし，SNS に蓄積されるデータは日々膨大になり続けており，旅行以外にも多種多様な情報が玉石混交である．そこで本稿では，SNS から旅行活動データをノイズなく網羅的に収集するため，ChatGPT など既存の様々な機械学習を用いて，旅行活動か否かの判定精度の比較実験を行う．また，収集した旅行活動データに含まれるノイズの割合と可視化の視認性との関係を実験し，機械学習においてどこまで正解率や適合率を伸ばすべきかを検討する．

キーワード　機械学習，分類，SNS 分析，観光，可視化

1　はじめに

近年，旅行活動を SNS に投稿するユーザが増えている．JTB の調査 [1] によると，実施した国内旅行の計画にあたり情報収集を行った媒体として,「SNS やブログ，動画サイト」を選択した割合が，若い世代になるにつれて高く，Z 世代では 3.5 から 5 割を占めている．また，旅行先を選択する際，SNS の情報を重視すると回答した層は全体の 2 割，自分自身の SNS に投稿することを意識して旅行先を選択した層が全体の 1 割を占めており，旅行活動において SNS は必要不可欠な存在となっている．

このような旅行活動データの利活用の一つとして，地図へのマッピングなど，可視化が考えられる [2–6]．可視化を行うことで，ある観光エリアにおけるスポットごとの旅行活動の特性が見えるようになり，観光エリアにおいて体験できる活動が明確になる．可視化を通して，新たな旅行先の決定や旅行需要の発見に繋がり，旅行活動が活発になるのではないかと考えた．尚，旅行活動データの可視化には大別して，インスタンス（具体例）毎の可視化 [2,3] と，インスタンス群を何らか集約した「集合知」的な可視化 [4–6] とがあり，本研究は前者に属する．

しかし，SNS に蓄積されるデータは日々膨大になり続けており，旅行以外にも多種多様な情報が含まれている．旅行活動データを収集する際，収集の仕方を工夫しなければ，旅行と関係のない情報が混ざり，可視化の妨げになることが考えられる．例えば，彦根の旅行活動を可視化する場合，彦根城や四番町スクエアの観光が想定される．しかし，図 1 のように彦根駅前での街頭演説や，住民の日常生活の投稿など，旅行活動ではないものまで可視化されると，旅行活動が正しく認識できなくなる．

そこで本稿では，SNS から旅行活動データをノイズなく，かつ網羅的に収集するため，ChatGPT など既存の様々な機械学習を用いて，旅行活動か否かの判定精度の比較実験を行う．本稿では SNS のうち Twitter に着目し，ツイート本文のテキストから旅行活動か否かを判定する．また，収集した SNS の投稿を可視化した際の視認性への影響に関して調査することで，旅行活動ツイートの可視化に向けて到達すべき判定精度の目安を調査する．

図 1　旅行活動を正しく認識できない場合

本稿の以降の構成は以下の通りである．まず 2 章で関連研究について述べる．次に，3 章で機械学習を用いた判定精度の比較実験の方法について詳述し，実験結果をまとめる．その上で，4 章で可視化視認性に関するアンケート調査について，5 章でより大規模なデータを用いた分類精度（特にノイズ率）の実験について，最後に，6 章で今後の研究課題について述べる．

2　関連研究

本章では，関連研究について述べる．

2.1　Twitter ユーザの居住地推定に関する研究

西村ら [7] は，ツイートに出現する単語情報から，ユーザの居住地推定を行った．居住地が判明しているユーザのツイートから，地域ごとの特徴語リストを作成しておき，特徴語リストをもとに居住地推定を行うことで，精度が向上した．しかし，正しく居住地を推定できたユーザの正解率は平均して 0.43 であり，まだまだ改善が必要である．

2.2 ツイートを利用した災害情報採集手法に関する研究

藤田ら [8] は，Twitter 内に散らばる災害情報を代表語 1 単語（例えば，台風，地震など）で検索し，その後，自己教師あり分類器で災害情報を分類する手法を検討した．自己教師あり分類器を作製することで，代表語がツイートに含まれているか否かで分離するよりも，精度よく分類できることが確認できた．しかし，分類結果を可視化するにあたり，どの程度精度が高ければ有用なシステムとなるのか，明らかにされていない．

3 機械学習を用いた判定精度の比較実験

3.1 実験方法

我々はこれまでに，観光地における旅行活動ツイートの収集手法として,「地名」,「位置情報」,「特徴語」の 3 手法を組み合わせたフィルタリング手法により，旅行活動ツイートの網羅的な収集を検討した [2]．その結果，観光地「彦根」を旅行するツイートを 94 件収集した．しかし，ルールベースによる分類は精度が低いことが分かった．

そこで本稿では，既存の様々な機械学習を用いて，収集したツイートを「彦根の旅行活動」と「彦根の非旅行活動（日常生活や，広告や個人の願望など活動でないもの）」に分類する実験を行った．入力するデータは,「彦根の旅行活動ツイート 94 件」と「彦根の非旅行活動ツイート 100 件」である．但し,「彦根の旅行活動」と「彦根の非旅行活動」の分類は第 1 著者ができる限り客観的に行った．194 件のツイートのうち，学習データ：テストデータを 8：2 に分割して実験を行った．プログラミング言語は Python を使用した．尚，5 章では，これらの従前のフィルタリング手法 [2] によって得られた小規模なデータセットに加え，本稿で新たに収集したより大規模なデータセットに対して，本章と同様に，既存の様々な機械学習を用いて分類精度（特にノイズ率）の実験を行う．

3.1.1 SVM

SVM (Support Vector Machine) は，教師あり機械学習モデルの一種で，異なるクラス間に境目となる直線や曲線を引くことにより，分類を実現する．主に分類のタスクに用いられる．入力にはベクトルが必要であり，本稿ではツイート（文書）のベクトル化手法として「TFIDF」と「Doc2Vec」の 2 種類を用いた．ライブラリは，TFIDF によるベクトル化と SVM の分類器作成には scikit-learn を，Doc2Vec によるベクトル化は gensim を用いた．尚，ハイパーパラメータはデフォルト値である．

3.1.2 fastText

fastText は，Word2Vec を基に作られた機械学習ライブラリである．Word2Vec と比較し，単語をサブワードの集まりとして学習しているため，辞書に存在しない未知の単語に対してもベクトルを生成することができる．fastText の入力フォーマットに合わせるため，形態素解析を先に行い，形態素間をスペースで区切る処理を行ってから入力した．尚，ハイパーパラメータは全てデフォルト値である．

3.1.3 ナイーブベイズ

ナイーブベイズは各特徴量が独立であると仮定し，出力の確率を計算するモデルである．各ツイートは，TFIDF によるベクトル化処理を行ってから入力した．分類器の作成は機械学習ライブラリ scikit-learn の GausianNB を用いて，ハイパーパラメータは全てデフォルト値とした．

3.1.4 BERT・RoBERTa

BERT (Bidirectional Encoder Representations from Transformers) は，文章を文頭と末尾の双方向から学習を行うことで，従来の機械学習から精度を向上したモデルである．事前学習モデルとして，cl-tohoku/bert-base-Japanese を用い，トークナイザには BertJapaneseTokenizer を使用した．学習データを使って転移学習を行い，テストデータで検証した．

RoBERTa (Robustly optimized BERT approach) は，従来の BERT を改良して，事前学習用データセットを大幅に増やし，事前学習の回数を増やした．また，バッチサイズを大きくし，長い文章を投入することで，BERT の精度を大きく上回っている．事前学習モデルは，rinna/japanese-roberta-base を，トークナイザには T5Tokenizer を使用した．学習データを使って転移学習を行い，テストデータで検証した．

3.1.5 ChatGPT

ChatGPT は，2022 年 11 月に OpenAI 社がリリースした大規模言語モデルの一種である．ユーザが入力した質問に対し，人間のような自然な対話形式で回答が出力されるシステムである．これを利用して，分類問題を解く．OpenAI の API を使用して，ChatGPT の gpt-3.5-turbo を用いた旅行活動ツイートの分類を行った．プロンプトは，旅行活動の定義のみを入力する簡易なものと，定義に加えて旅行活動ツイートの判断材料をいくつか追加した詳細なものの 2 種類で実験を行った．

まず，簡易プロンプトを以下に示す．

Q1: 今から，彦根の旅行活動ツイートか否かの分類を行っていただきます．旅行活動ツイートの定義は以下のとおりです．
・観光地に移動し，観光地内の各スポットで何らかの活動を行うことを旅行とする
以下のように出力してください．
（ツイート本文）
判定: 彦根の旅行活動である or 旅行活動だが彦根ではない or 旅行活動ではない理由:（上記のように判断した理由）
例入力: 今日は彦根城に行きました．ひこにゃんが来てくれました．出力: 今日は彦根城に行きました．ひこにゃんが来てくれました．判定: 彦根の旅行活動である理由: ツイートには彦根城への訪問が明記されており，当日の行動であることが示されています．また，ひこにゃん（彦根城キャラクター）が登場したことから，彦根城での観光活動が行われたことが明確です．
準備が整い次第，分類してほしいツイートを入力します．
A1: 了解しました．準備が整い次第，分類してほしいツイートを入力してください．どのツイートでもお手伝いできます．
Q2: （ツイート 1 件）

表 1　従前の小規模データを用いた各機械学習の分類結果

機械学習モデル	正解率	彦根の旅行活動ツイート 適合率	再現率	F1 値	彦根の非旅行活動ツイート 適合率	再現率	F1 値
SVM (TFIDF)	0.749	0.970	0.520	0.677	0.679	0.986	0.804
SVM (Doc2Vec)	0.497	0.499	0.387	0.436	0.514	0.619	0.562
fastText	0.685	0.955	0.367	0.530	0.626	0.984	0.765
Naive Bayes (TFIDF)	0.731	0.663	0.853	0.746	0.836	0.636	0.722
BERT	0.516	0.511	0.447	0.477	0.509	0.570	0.538
RoBERTa	0.492	0.515	0.490	0.502	0.466	0.492	0.479
ChatGPT (簡易プロンプト)	0.655	0.935	0.309	0.464	0.601	0.980	0.745
ChatGPT (詳細プロンプト)	0.747	1.000	0.479	0.647	0.671	1.000	0.803

次に，詳細プロンプトを以下に示す．

Q1:　今から，彦根の旅行活動ツイートか否かの分類を行っていただきます．旅行活動ツイートの定義は以下のとおりです．

・観光地に移動し，観光地内の各スポットで何らかの活動を行うことを旅行とする詳細は以下のとおりです．

①　観光地内スポットもしくは観光地名と，その場所での行動が含まれていること．行動が明記されていない場合でも，文脈から行動していると判断できる場合は旅行活動ツイートと判定する．

②　当日の投稿であること（”今日は◎◎行った！”は旅行活動だが，”先週の△△楽しかった！”や”来週△△行きたい！”は旅行活動ではない）日付情報が明記されていない場合は当日の行動とする．

③　本人の活動であること（■■が今日◎◎行ったらしい．は旅行活動ではない）

④　I’m in ◎◎　の投稿は，その投稿が題材とする観光地に関する場所ならば旅行活動ツイート

⑤　題材とする観光地以外にも複数の観光地と行動があった場合，題材とする観光地の旅行活動として採

⑥　題材とする観光地に住んでいると考えられるユーザのツイートは旅行活動ツイートとしない（旅行活動ではないため）

以下のように出力してください．

（ツイート本文）

判定: 彦根の旅行活動である or 旅行活動だが彦根ではない or 旅行活動ではない理由:（上記のように判断した理由）

例入力: 今日は彦根城に行きました．ひこにゃんが来てくれました．出力: 今日は彦根城に行きました．ひこにゃんが来てくれました．判定: 彦根の旅行活動である理由: ツイートには彦根城への訪問が明記されており，当日の行動であることが示されています．また，ひこにゃん（彦根城のキャラクター）が登場したことから，彦根城での観光活動が行われたことが明確です．

準備が整い次第，分類してほしいツイートを入力します．

A1:　準備が整いましたので，分類してほしいツイートを入力してください．それに対して，指定された基準に基づいて判定と理由を提供します．ツイートをお待ちしています．

Q2:　（ツイート 1 件）

プロンプトは，Q1 (user)，A1 (assistant)，Q2 (user) の形式で入力した．回答の判定は,「彦根の旅行活動である」と出力した場合のみ，旅行活動と予測したものとし,「旅行活動だが彦根ではない」,「旅行活動ではない」，もしくは意図しない判定結果（例：彦根の旅行活動ではない）を出力した場合は，正しく判定できなかったと，全て「彦根の旅行活動ではない」とした．

3.2　実 験 結 果

実験結果を表 3.2 に示す．最終的に必要なのは彦根の旅行活動ツイートであり，旅行と関係のないツイートは極力拾わないようにしたい．このため，旅行活動ツイートの適合率・再現率を重視する．SVM は Doc2Vec よりも TFIDF でベクトル化を行うことで，適合率が高くなった．しかし，再現率が 0.520 と低いため，旅行活動ツイートの精度が高いが網羅的ではない．fastText は適合率が高いが，その分再現率が落ちている．ナイーブベイズは逆に，再現率が一番高い．網羅性を高めたい場合はナイーブベイズが有効であると考えられる．BERT・RoBERTa を比較すると，両者とも旅行活動ツイートの比較タスクでは精度が大きく変化しないことが分かった．最後に，ChatGPT は簡易なプロンプトよりも詳細にプロンプトを入力する方が精度が上がることが分かった．しかし，どちらのプロンプトの場合も，194 件中 20 件程度は判定結果が意図しない出力を返したため，再現率が低下した．

4　可視化視認性のアンケート調査

旅行活動を機械学習により分類しても，旅行活動と関係ないノイズが少なからず入る．このノイズをあえて追加し，可視化を行うことで，正解率，再現率，適合率の目標数値を定める．2022 年 10 月 23 日と 24 日に彦根を旅行しているツイートを収集し，ツイートに彦根市内のスポット情報が含まれているものについて Google マイマップを用いて可視化した．さらに，同日に旅行活動ではないものの彦根市内のスポット情報が含まれるツイート（ノイズ）をマイマップに 10%ずつ追加していき，ノイズ率とユーザが旅行活動を自然に理解できるかどうかの関係を調査した．図 2，図 3，図 4，図 5 は，10 月 23 日と 24 日でノイズが 0%の時と 50%の時の可視化した図である．質問は，

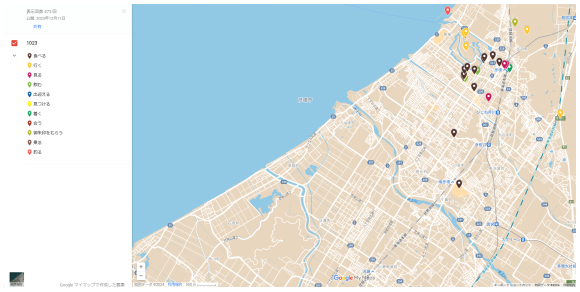


図 2　2022 年 10 月 23 日，ノイズ 0%の可視化結果

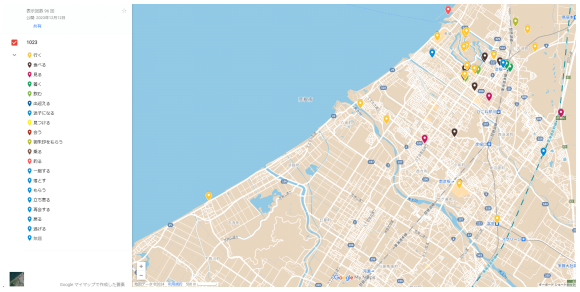


図 3　2022 年 10 月 23 日，ノイズ 50%の可視化結果

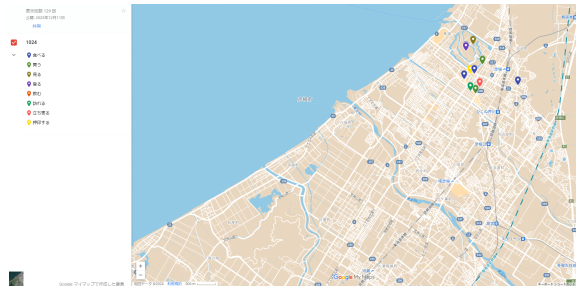


図 4　2022 年 10 月 24 日，ノイズ 0%の可視化結果

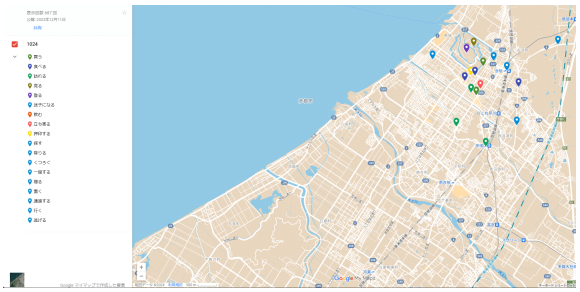


図 5　2022 年 10 月 24 日，ノイズ 50%の可視化結果

「彦根の旅行活動の可視化として，違和感があるかどうか」を自然度,「彦根の旅行活動を理解できたかどうか」を理解度として 10 段階で評価し，具体的に違和感な点を自由記述で回答してもらった．調査対象は滋賀県立大学の学生とし，最初に彦根についてどの程度知っているのかを,「知らない（最低限通学に必要な程度のみ）」,「あまり知らない（通学に加え，寄り道などした等）」,「少し知っている（観光したことがある等）」,「とても知っている（住んでいる，住んでいた等）」の 4 段階で回答してもらった．回答者の彦根に対する知識を表 2 に示す．

次に，ノイズ率とユーザが旅行活動を自然に理解できたかどうかの関係と被験者が違和感を感じたポイントについて，2022 年 10 月 23 日の結果を表 3 に，10 月 24 日の結果を表 4 に，ノイズ率と自然度，理解度の推移を図 6 に示す．尚，2022 年 10 月 23 日は彦根市内でイベント「ご当地キャラ博 in 彦根」が開催されており，翌日 24 日に開催されたイベントはなかった．

表 2　被験者の彦根に対する知識

彦根に対する知識レベル	人数
知らない	13
あまり知らない	27
少し知っている	15
とても知っている	8
合計	63

表 3　ノイズ率と可視化視認性の関係（2022 年 10 月 23 日）

ノイズ率	可視化に違和感がないか（自然かどうか）	彦根を理解できたか	具体的な違和感のポイント
0%	6.51	5.89	・ベルロードがない ・彦根駅の出迎えツイートが観光ではない
10%	6.51	6.21	・迷子のツイートが含まれている ・地域紹介として無意味なツイートが混ざっている
20%	6.35	6.15	・「行く」の項目の中身が薄くなっている
30%	6.27	6.13	・落とし物の情報は違う ・会うなどの行動の種類が増えてわかりやすくなっている
40%	6.27	6.02	・逃げるが違和感 ・彦根駅は目的ではない
50%	6.00	5.94	・戻る・再会のアイコン ・点が同じところに集まって視認性低下

表 4　ノイズ率と可視化視認性の関係（2022 年 10 月 24 日）

ノイズ率	可視化に違和感がないか（自然かどうか）	彦根を理解できたか	具体的な違和感のポイント
0%	6.38	6.10	・郵便局のスタンプ押印 ・項目が少なくわかりやすかった
10%	6.06	6.10	・乗り過ごしそうな人の情報が混ざっている
20%	6.41	6.00	・色の種類が増えた
30%	6.35	6.19	・彦根という文字だけを拾った具体性のないツイートが多い
40%	6.37	6.08	・探す 降りる くつろぐ 寝る 置く 通過する 逃げる 迷子になる
50%	6.49	6.10	・関係のないツイートが大半になった

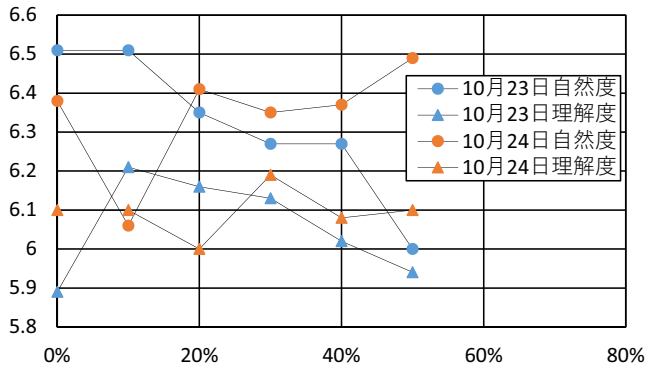


図 6　ノイズ率と自然度，理解度の関係

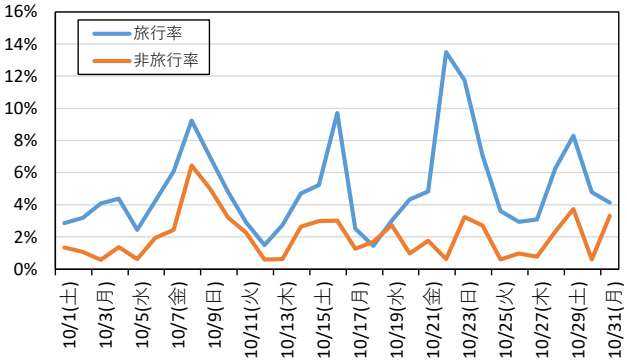


図 7　彦根を含むツイートのうち旅行・非旅行が占める割合

アンケート結果より，イベント開催日の 10 月 23 日はノイズ増加とともに違和感と感じる人が増え，理解度も低下した．しかし，その翌日である 10 月 24 日は旅行活動ツイートが少なかったのもあり，ノイズが加わることで逆に旅行活動として認識されてしまったものと考えられる．また，10 月 23 日のノイズ 10%から 20%にかけて自然度が大きく減少したため，可視化によって目指すべきノイズ率は 10%から 20%であると考えている．

被験者の彦根に対する知識レベルは，彦根市にある滋賀県立大学の学生であったため，通学に最低限必要な程度の知識レベルの人が最も多かった．被験者の知識レベルによっては，アンケートの結果（可視化への違和感を感じる割合や彦根に対する理解度）も変化する可能性がある．

5　より大規模なデータを用いた旅行活動分類

可視化において必要なツイートは「観光スポットを示す場所で，旅行活動を示す行動」であり，ノイズとなるツイートは「観光スポットを示す場所であるが，行動は旅行活動ではない」である．本章では 4 章を踏まえ，実際に収集される投稿にはどの程度ノイズが含まれるのか，そして，各機械学習に入力するとどの程度ノイズを旅行活動と判定してしまうのか，評価実験を行う．3.2 節において旅行活動データは 2019 年 2 月頃のデータを 94 件収集したが，旅行活動がより活発となる行楽シーズンを対象に旅行活動ツイートを評価するため，2022 年 10 月のツイートを収集した．その結果,「彦根」を含むツイートを 15944 件収集した．収集したツイートを，第 1 著者が「彦根の旅行活動ツイート (a)」,「彦根の（旅行ではない）一般的な活動ツイート (b)」,「彦根を含むだけのツイート (c)」の 3 つに分類した．このうち,「彦根の旅行活動ツイート (a)」「彦根の（旅行ではない）一般的な活動ツイート (b)」については可視化できるようにするため，ツイートから読み取れる場所と行動を，メタデータとして付与した．一方,「彦根を含むだけのツイート (c)」に関してはツイートから場所と行動が読み取れないため，可視化されない．2022 年 10 月に「彦根」を含むツイートのうち,「彦根の旅行活動ツイート (a)」,「彦根の（旅行ではない）一般的な活動ツイート (b)」の割合を図 7 に示す．

図 7 より，土日祝日（2022 年 10 月 10 日は祝日である）は観光客が増加し，旅行率が増加した．また，10 月 22 日，10 月 23 日は日本全国のゆるキャラが集まる「ご当地キャラ博 in 彦根」が開催されたため，特に旅行活動ツイートが多く収集された．以上の分類を行った結果,「彦根の旅行活動ツイート (a)」は約 6%にあたる 973 件，ノイズとなり得る「彦根の一般的な活動ツイート (b)」は約 2%の 312 件であった．この結果を用いて，機械学習による分類を再度行った．この分類は，ツイートが「彦根の旅行活動か否か」を判定するために行うものであるため,「彦根の旅行活動ツイート (a)」と「彦根の非旅行活動ツイート $(b+c)$」の 2 値分類を行った．また，ツイートが可視化されるのは場所と行動の情報を持つ「彦根の旅行活動ツイート (a)」と「彦根の一般的な活動ツイート (b)」である．このうち,「彦根の一般的な活動ツイート (b)」はノイズとなり，旅行活動可視化の妨げとなる．そこで，機械学習によって「彦根の旅行活動ツイート (a)」と判定された投稿のうち，ノイズを含む投稿の割合を示すノイズ率を計算した．ノイズ率の計算式を以下に示す．

$$\text{ノイズ率} = \frac{\text{一般的な活動 } (b)}{\text{旅行活動 } (a) + \text{一般的な活動 } (b)} \qquad (1)$$

各機械学習による分類結果と，ノイズ率を表 5 に示す．尚，表 3.2 と違い，各機械学習は 1 回のみ実施している．

本章のより大規模な 15944 件のデータを入力した結果，ノイズ率は 0.0 から 0.3 程度となった．4 章より，可視化によるノイズ率が 10%から 20%であれば比較的違和感を感じにくいとされたため，これらの機械学習による結果は可視化に十分有用であると考えられる．一方で，従前の小規模データでは，旅行活動と非旅行活動の比率が 94 : 100 であった．15944 件の大規模データでは，この比率が 973 : 14971 となり，旅行活動の比率が大きく偏っている．このため，従前の小規模データの実験（表 3.2）と比較して彦根の旅行活動に関する精度が下がってしまったと考えられる．特に，SVM (Doc2Vec) はテストデータを全て「彦根の非旅行活動ツイート $(b+c)$」と予測したため,「彦根の旅行活動ツイート」における適合率，F1 値，ノイズ率が計算不能となった．SVM (Doc2Vec) は 3.2 節でも SVM (TFIDF) を下回る結果であったため，ツイートの分類には不向きである可能性が考えられる．

表 5　大規模データを用いた各機械学習の分類結果

機械学習モデル	正解率	彦根の旅行活動ツイート 適合率	再現率	F1 値	彦根の非旅行活動ツイート 適合率	再現率	F1 値	ノイズ率 $((b)/((a)+(b)))$
SVM (TFIDF)	0.948	1.000	0.179	0.303	0.948	1.000	0.973	0.000
SVM (Doc2Vec)	0.934	0.000*1	0.000	0.000*1	0.934	1.000	0.966	0.000*1
fastText	0.943	0.510	0.141	0.221	0.950	0.992	0.970	0.103
Naive Bayes (TFIDF)	0.883	0.277	0.411	0.331	0.954	0.919	0.936	0.179
BERT	0.939	0.727	0.079	0.143	0.941	0.998	0.969	0.273
RoBERTa	0.940	0.620	0.234	0.340	0.949	0.990	0.969	0.338
ChatGPT (簡易プロンプト)	0.903	0.241	0.188	0.211	0.941	0.956	0.948	0.000
ChatGPT (詳細プロンプト)	0.891	0.288	0.462	0.355	0.961	0.921	0.940	0.286

*1: SVM (Doc2Vec) を用いた場合,「彦根の旅行活動ツイート」と判定されたツイートが 1 件も無かったため，適合率や F1 値，ノイズ率は計算不能.

6　まとめと今後の研究計画

本稿では，SNS から旅行活動データをノイズなく網羅的に収集するため，既存の様々な機械学習を用いて，旅行活動か否かの判定精度の比較実験を行った．また，新たに収集したより大規模なデータセットを用いて，収集したデータに含まれるノイズの割合と可視化視認性との関係を実験し，機械学習においてどこまで正解率や適合率を伸ばすべきかを検討した.

その結果，従前の小規模データセットでは彦根の旅行活動ツイートに対する適合率が 0.5 から 1，再現率が 0.3 から 0.8 であることが分かった．また，新たに収集したより大規模なデータセットの 2022 年 10 月に「彦根」を含む投稿のうち，約 6%にあたる 973 件の投稿が彦根の旅行活動ツイートとして，約 2%にあたる 312 件がノイズツイートとして収集された．そして，アンケート調査の結果，可視化によって目指すべきノイズ率が 10%から 20%であることが分かった.

今後の研究計画として,「機械学習モデルの組み合わせ」を考えている．本稿では，各機械学習モデルをそれぞれ独立して比較を行ったが，複数の機械学習モデルを組み合わせることにより，再現率を改善できるのではないかと考えている．例えば，SVM (TFIDF) の旅行活動の再現率は 0.520 であるが，ChatGPT の再現率は 0.479 である．これらのモデルが旅行活動と判定した投稿を単純に足し合わせることにより，再現率の改善が見込まれる．ただし，SVM (TFIDF)，ChatGPT の両方が旅行活動であると判定した投稿が多いと，足し合わせても再現率は改善しないため，精査が必要である．また，再現率が低く，適合率が高い ChatGPT は，旅行活動であるにも関わらず取りこぼした投稿が存在する．これを，再現率が高く，適合率が低いナイーブベイズに分類させることで，ChatGPT が取りこぼした旅行活動ツイートを拾うことができるのではないかと考えている．このように，異なる機械学習モデルを組み合わせることによる精度改善を検討する.

文　献

[1] 公益財団法人日本交通公社, “国内旅行における SNS・写真に対する意識／実態 ～JTBF 旅行実態調査トピックス～,” https://www.jtb.or.jp/research/statistics-tourist-sns-pictures2022/ (参照 2023-09-26).
[2] 藤大 友都, 服部 峻, 砂山 渡, “旅行活動ツイート可視化のためのフィルタリングと Artisoc エージェント化,” 第 15 回データ工学と情報マネジメントに関するフォーラム，5b-3-2 (2023).
[3] 永澤 勇樹, 吉田 京平, 服部 峻, “モバイル端末における旅行記の理解支援のための行程抽出と地図化,” 電子情報通信学会 モバイルネットワークとアプリケーション研究会 (SIG-MoNA), 信学技報, Vol.114, No.31, MoNA2014-4, pp.19–24 (2014).
[4] Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka, “Toward tighter integration of web search with a geographic information system,” Proceedings of the 15th International Conference on World Wide Web (WWW’06), pp.277–286 (2006).
[5] Hiroshi Kori, Shun Hattori, Taro Tezuka, and Katsumi Tanaka, “Automatic Generation of Multimedia Tour Guide from Local Blogs,” Proceedings of the 13th International MultiMedia Modeling Conference (MMM’07), LNCS Vol.4351, Part I, pp.690–699 (2007).
[6] 渡邉 小百合, 吉野 孝,“観光地間の類似性を基にした向上点発見のための観光情報可視化システム,” マルチメディア，分散協調とモバイルシンポジウム 2016 論文集, 2016, pp.1357–1362 (2016).
[7] 西村 駿人, 数原 良彦, 鷲崎 誠司, “地域特徴語選択を用いたマルチクラス分類による Twitter ユーザの居住地推定,” 電子情報通信学会 信学技報，Vol.112，No.367，NLC2012-37，pp.23–27 (2012).
[8] 藤田 俊之，小林 亜樹，“マイクロブログにおける会話を利用した災害情報採集手法,” 電子情報通信学会論文誌 B，Vol.J106-B，No.1，pp.1–12 (2023).
[9] artisoc4 - MAS コミュニティ - 構造計画研究所, https://mas.kke.co.jp/artisoc4/ (参照 2023-10-03).

近接関係ネットワークに影響を与える歩行者行動の可視化

森越　彩楓† 　丹羽　了††,††† 　鷹見　竣希††† 　大西　正輝††† 　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2–1–1
†† 筑波大学　〒 305–8577 茨城県つくば市天王台 1–1–1
††† 産業技術総合研究所　〒 135–0064 東京都江東区青梅 2–4–7
E-mail: †{g1820540,itot}@is.ocha.ac.jp, ††niwa.3soken@aist.go.jp, †††{s-takami,onishi-masaki}@aist.go.jp

あらまし　感染症は感染者との接触で拡散するため，歩行者の近接状態は感染リスクが高い． したがって，多くの観客が集まるイベント会場では，感染の拡大を防ぐために，群集制御で歩行者の近接を防ぐことが重要である．一方で，有効な群集制御を実現するには，感染リスクを高める歩行者の行動についての理解が必要であり，そのためには可視化技術の適用が有効である． そこで本研究では，歩行者をノードで表し，近接関係にある歩行者間をエッジで結んだ近接関係ネットワークを構築するとともに，歩行者の行動をもとにノードを色分けすることで，歩行者の行動が感染リスクに与える影響を可視化する． 分析例では，群集シミュレーションを用いることで大規模なイベント会場での人の流れを取得し，提案手法を適用することで歩行者の行動が感染リスクに与える影響を観察した．

キーワード　新型コロナウイルス，行動解析

1　はじめに

新型コロナウイルスの感染拡大は，感染症対策のためにマスクの着用や手指の消毒だけでなく，不要不急の外出禁止や飲食店の営業時間短縮などの行動規制も必要とされ，日常生活や経済活動に大きな影響を及ぼした．特に，多くの観客が集まるスポーツ観戦やライブコンサートのような大規模イベントでは効果的な感染症対策が重要であった．一方で日本では，2023 年 5 月に新型コロナウイルス感染症の感染症法上の位置づけが変更され，国の新型コロナウイルス感染症対策の基本的対処方針による，イベント開催時の収容人数の上限や大声での歓声・声援の禁止といった制限が完全に撤廃された．強制力のある感染症対策の実施が緩和された現在でも，感染者がいなくなったわけではないため，感染症の蔓延を防ぐことは依然として重要である．感染症は感染者との接触によって拡散する [1] ことから，感染症の対策には歩行者間の距離を保つことが有効である．そのため群集制御を行い会場内の混雑を軽減することで，人々が十分な間隔を保てるようになり，感染リスクを抑えることができる．

有効な群集制御を実現するためには，会場内の人の流れを分析することが必要であり，そのためには可視化技術の適用が有効である．そこで我々は，歩行者の近接を感染リスクが高い状態とし，近接状況をネットワークを用いて可視化する手法を提案した [2]．しかし従来の近接関係ネットワークでは，群集制御のための特徴を分析するには以下の問題があった．

（1） ノードの配置に位置情報が含まれていない
（2） 会場全体の歩行者の行動までは考慮していない
（3） 近接状況に変化を及ぼす歩行者の行動を特定するには至っていない

（1）について，従来の近接関係ネットワークでは，ネットワーク全体の視認性を高めることにのみ焦点を当てたノードの配置となっており，ノードの位置は情報を持っていなかった．しかしノードの位置から位置情報を読み取れることで，近接について有益な特徴を発見することが可能となる．そこで本研究では，Group-In-a-Box (GIB) レイアウト [3] を拡張し，位置情報を反映してノードを配置することで，この課題に対処する．

（2）について，従来の近接関係ネットワークは，イベント会場のコンコースの一部を計測したデータを可視化していた．そのためイベント全体での歩行者の行動は取得できず，可視化することができなかった．会場全体の歩行者の行動を分析するには，イベント会場内における全来場者の歩行情報を取得する必要がある．人の歩行情報は，通常防犯カメラや GPS データによる計測で取得される．しかし，イベント会場のような規模の領域を計測することは技術的な難易度が高い．例えば，カメラによる計測では，一台のカメラで計測できる領域に限りがあるため，広い会場内における全ての歩行者の移動を正確に取得することは難しい．また GPS による計測は，位置誤差が数十メートルに及ぶこともあり，イベント会場のサイズに対して誤差が大きいため，精度に懸念が残る．このような背景から，計測による歩行者情報の取得は，イベント会場全体のような規模の空間には適していない．そこで本研究では，群集シミュレーションを用いることで会場全体での歩行者の行動を取得する．

（3）について，従来の近接関係ネットワークでは，計測で得られた一つの状態のみを可視化していたため，歩行者の行動の変化が近接状況に及ぼす影響に関しては分析されていなかった．シミュレーションを用いると，パラメータを変えることで環境や歩行者の行動を変化させることができるため，多様な条件での歩行情報が推定できる．そこで本手法では，分析したい歩行者の行動に対して多様なシナリオでのシミュレーションを実行し，取得した歩行者の行動をノードの色で表示することで，歩行者の行動が感染リスクに及ぼす影響を観察する．

本稿の構成は以下のとおりである．2 節では関連研究について述べる．3,4 節では提案手法について説明する．5 節では実験設定および実験結果について述べる．最後に 6 節でまとめる．

2 関連研究

本節では，本研究に関連する人流分析と群集シミュレーションについて関連研究を紹介する．

2.1 人流分析

人流の分析は，観光や都市計画，防災，マーケティングなど，多様な分野で活用されている．本節ではまず，可視化による人流分析の手法を紹介し，その後にコロナ禍における人流分析の動向を説明する．

McArdle らは，主要なルートや頻繁に停止する場所などの移動パターンを分析するために，時空間キューブ（STC）を用いた手法を提案している [4]．この手法では，歩行経路を 3 次元で表示することで，時間情報・空間情報の両方が可視化できる．時期ごとの人流の違いを比較するために，Fukute らは ThemeRiver という手法によって主要経路ごとの流量の時間変化を可視化している [5]．Wang らは，混雑状況を可視化することを目的として，歩行者の通行状況をネットワークで表現する手法を提案している [6]．このネットワークでは，エッジは歩道を表し，ノードは交差点を表す．混雑度を交差点を通過する際に生じるエネルギーの高さとし，それを各ノードの色で可視化する．このように人流の可視化手法は多数提案されているため，観察したい特徴によって適している可視化手法を選択する必要がある．

感染症は，咳やくしゃみ，会話などによって感染者から放出されたウイルスに接触することで拡散する．このことから感染症の拡大にともない，人流に関する研究がより活発に報告されている．Arimura らは，携帯電話の位置データを使用して，緊急事態宣言期間中の札幌市の人口密度の変化を観察した [7]．Li らは，POI (Point of Interest) に着目し，感染症の影響による動線パターンやパターンの変化を分析した [8]．Askarizad らは，ある地点の歩行者通行量分布にもとづいて感染拡大前後での人の行動パターンを比較した [9]．

感染症拡大の直後では，感染症や感染症対策が歩行者の行動に与えた影響を分析することが重要であった．一方で，それ以降の感染定常状態では，感染を拡散させる因子の特定に焦点を当てた分析が必要であるとされてきた．そこで Uddin らは，性別や年齢のような人口統計学的属性や，子供の有無や喫煙頻度のような個人の属性が新型コロナウイルス感染症予防行動の遵守にどのような影響を与えるかを分析した [10]．また Turk らは，人口統計学的属性や個人の属性が，人々の感染症に関する認識や反応性にどのような影響を与えるかを分析した [11]．Ando らは，通勤と感染リスクに関するオンライン調査を実施し，通勤時間が長ければ長いほど濃厚接触者の可能性が高くなる傾向を発見した [12]．このように，人の属性に着目した感染症の分析は活発に発表されている．一方で広範囲における人流計測の難しさから，大きなイベントでの歩行者の行動に関する感染リスクの分析は少ない．

そこで本研究では，感染リスクが高い状態として歩行者の近接状態に焦点を当てる．歩行者の行動が近接に及ぼす影響について分析するために，近接者同士を結ぶネットワーク [2] を拡張する．これまでの研究では，計測によって得られた，会場の一部で起きた近接のみを扱っていた．それに対して本稿では，群衆シミュレーションの結果を用いて，会場内すべての近接情報を可視化する．さらに，人の行動でノードを色分けし，歩行者の行動が近接に及ぼす影響を観察する．

2.2 群集シミュレーション

領域全体を計測することが難しい環境での人流を解析するために，群衆シミュレーションを適用することで，環境の一部の計測データから環境全体の歩行軌跡を補完できる．群集シミュレータの例として，A&A 社の SimTread [13] や構造計画研究所の artisoc [14] などがある．群衆シミュレータの多くは，各歩行者をエージェントとしてみなすマルチエージェントシミュレーションである．マルチエージェントシミュレーションでは，各歩行者の動きを独立して推定できる．そのため，歩行者のすれ違いや混雑の発生など，感染症を考える上で重要な歩行者の動きをシミュレーションできる．

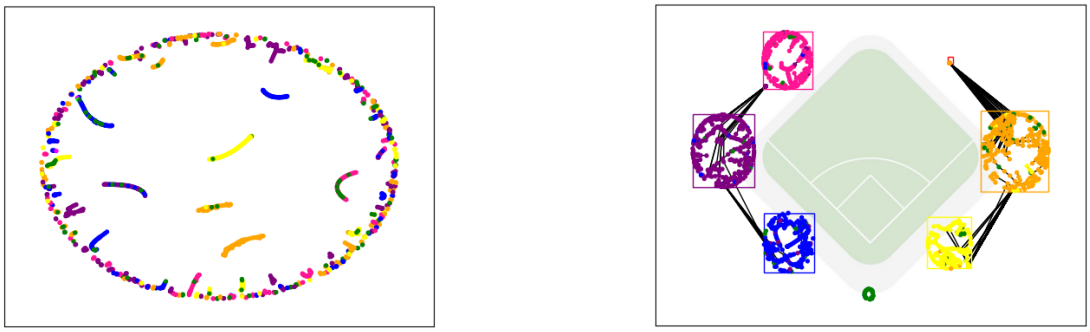
本研究で再現するシナリオでは，シミュレーション時に以下のような問題がある．

- 計算時間
- エージェントごとの経路設定

多くのマルチエージェントシミュレータでは，移動モデルに周囲の歩行者や障害物からの反発力で加速度が決まる Social Force モデル [15] を採用している．Social Force モデルでは，シミュレーション空間内に存在する全てのエージェント間の反発力を計算するため，エージェント数が増えれば増えるほど計算時間が増大する．そこで Yamashita ら [16] は，歩行者が移動する空間をエッジとノードで表現する一次元歩行者モデルを提案した．これによって近傍の歩行者の判定や移動方向を決定する計算過程が短縮されるため，高速にシミュレーションすることが可能である．

また，マルチエージェントシミュレーションでは，歩行者一人ひとりの出発点から目的地までの経路を設定する必要がある．本研究では 4 万人の歩行者がそれぞれ初期座標 (座席) をもつため，それぞれの座席から目的地までの経路を設定することは実装コストが高い．そこで，このような状況でのシミュレーション環境として，エッジとノードで表現する一次元歩行者モデルを用いる．これによって，グラフ理論を用いて出発点から目的地までの最短経路を求めることが可能となり，経路設定が容易となる．

これらの理由から本研究では，一次元歩行者モデルで実装された群衆シミュレータである CrowdWalk [17] を使用して，スタジアムにおける大規模人流を再現する．



(a) 従来手法 [2] によるネットワーク描画結果　(b) 提案手法によるネットワーク描画結果

図 1: 同じデータから二つの手法を用いてネットワークを描画した結果．ノードの色は退場ゲートで決定．

3　近接関係ネットワークの可視化

本節では，まず近接状況を可視化するネットワーク [2] について説明し，その後に本研究のための改良を述べる．

3.1　従来の近接関係ネットワークの描画手法

近接関係ネットワークは，近接情報の抽出，ネットワーク構造の構築，ノードの配置を経て描画される．

まず，人流データから近接情報を抽出する．ここで，歩行者間の距離が l メートルより近い状態を「近接」と定義する．さらに，近接時間の閾値 m を定め，近接状態が m 分以上続いていた歩行者のペアを抽出する．

続いて，近接状況を可視化するために，歩行者をノードとし，近接関係にある歩行者間をエッジで連結してできるネットワークを構築する．ここで本手法では，観測された全ての歩行者を可視化するのではなく，他の歩行者と近接状態にあったことのある歩行者に対応するノードのみを表示する．これによって，感染拡大に関与している可能性のある歩行者のみを可視化できる上に，ネットワークの複雑さの減少に貢献する．なお，現時点での実装では，ノードの色は歩行者の行動にもとづいて割り当てている．

最後に，力学モデルにもとづいたグラフレイアウト手法である Fruchterman-Reingold アルゴリズム [18] を適用して，ノードの配置を算出する．

従来の近接関係ネットワークの可視化では，ネットワーク全体の構造を理解しやすくすることを最優先してノードを画面配置していた．このような従来の可視化によって，近接状況を概略的に可視化することができる．それに対して，本研究の目的である「歩行者の行動が近接状況に及ぼす影響」を分析するには，実際の位置情報がノードの画面上の位置に反映されている必要があると考えた．

3.2　位置情報を反映したグラフレイアウト手法

本研究では，ノードの配置に位置情報を反映するために，Group-In-a-Box (GIB) [3] を拡張したグラフレイアウト手法を用いる．GIB は，カテゴリ情報などをもとにノードをグループ化し，それぞれのグループをボックス内に配置することで，グラフから構成されたグループ構造の理解を支援するグラフレイアウト手法である．ここでボックスの配置は，グループが有するノード数をもとに，Treemap によって決定される．GIB は，ソーシャルネットワークにおけるコミュニティ構造を分析するために開発された手法である．ソーシャルネットワークは仮想空間のデータであるため，具体的な座標を持たない．一方で我々が扱う歩行情報は，現実空間での位置情報を有するデータである．そこで，ノードの配置に現実の位置情報を反映させることで，我々の従来手法では表現できなかった特徴を可視化できることが期待できる．また GIB はグループ内の分析が重要視されていたのに対し，本研究では異なるグループ間の近接を可視化することが重要である．本手法では GIB を拡張し，現実空間での位置情報をノードのグループ化やボックスの配置に反映させる．さらに，グループ外に接続のあるノードが，ボックスの外縁に近い場所に配置されるようにレイアウトを調整することで，グループ間を連結するエッジの視認性を向上させる．

本手法におけるノードのグループ化と配置の手順は以下のとおりである．

- 座席から最寄りのゲートをもとにノードをグループ化する．
- グループがもつノード数からボックスのサイズを算出し，ゲートの位置関係をもとにボックスを配置する．
- 異なるグループ間を結ぶエッジを削除し，Fruchterman-Reingold でボックスごとのノード配置を算出する．
- 異なるグループ間を結ぶエッジを再生成し，異なるグループとの接続があるノードのみに再度 Fruchterman-Reingold を適用し，ボックスの外縁に配置する．

図 1 は同一のネットワーク構造を従来手法および提案手法で画面配置した例である．図 1(a) は従来手法で配置した近接関係ネットワーク，図 1(b) は提案手法で配置した結果である．空間情報を反映してノードを配置することで，近接状況の空間的な背景を分析しやすい可視化結果を得ることができるようになった．

4　シミュレーションのよる歩行者の行動の生成

本節では，群集シミュレーションを用いた歩行者の行動の生成について説明する．図 2 に示すスタジアムでの野球観戦にお

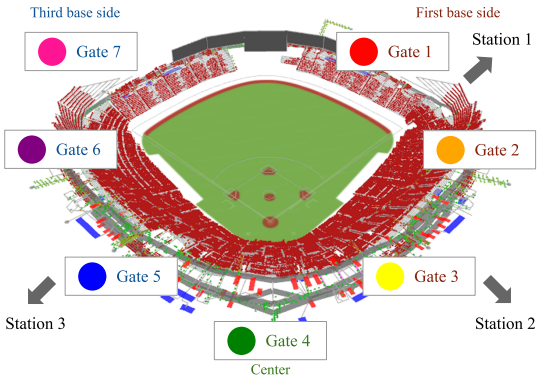


図 2: スタジアムのマップ.

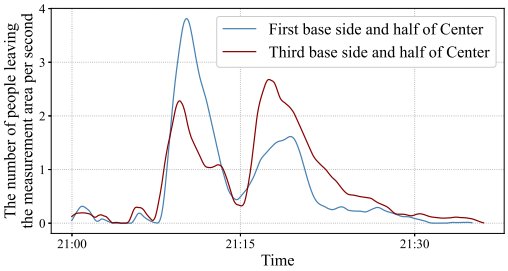


図 3: 1 秒あたりに計測可能領域を出た人の数の推移．試合は 21:09 に終了した．x 軸は時間，y 軸は退場者数を表示する．

ける退場時の歩行者の行動を再現する．

4.1 会場全体の歩行者の行動の再現

一般的に，イベント会場全体の計測は難しい．そのため，従来の研究 [2] では，コンコースの一部の計測結果を可視化していたが，会場全体の感染リスクを把握できないという問題があった．そこで本研究では，群集シミュレーションを用いることで，イベント会場全体での歩行者の行動データを取得する．群集シミュレータとしては CrowdWalk [17] を使用する．CrowdWalk の実行には，各エージェントごとに出発場所，出発時間，目的地の設定が必要である．これらのパラメータを現実の状況に近づけるために，再現したスタジアムでの座席エリアの観客の様子をカメラで撮影し，YOLOv5 [19] を用いて人の位置を認識する．出発場所としては各エージェントの座席を設定する．各エージェントの出発時刻は，座席内を計測したデータから，抽出した人間が計測可能領域を出たタイミングに従って決定する（図 3）．各エージェントの目的地は 4.2 節で説明するシナリオをもとに決定する．計測を行った日の観客数を参考にして，約 4 万人が来場した状況を仮定し，出発場所から目的地までの歩行者の移動をシミュレーションする．

4.2 シナリオによる歩行者の行動の変更

歩行者の行動が近接状況に与える影響を分析するために，本研究では，各エージェントの目的地をシナリオをもとに変化させて群集シミュレーションを実行する．

本研究で再現するスタジアム (図 2) では，チケットに入場ゲートを記載することで，混雑を緩和する取り組みを実施している．ここで入場時には，指定されたゲートからの入場が徹底されているものとする．一方で退場時には，ゲート誘導のオペレーションは実施していない．これは観客が退場時にトイレや売店に立ち寄ることや，利用したい駅に近いゲートを使用することが想定されるためである．座席から最寄りのゲートを使用すると，退場にかかる時間が短くなり，感染リスクが低くなることが想定される．一方で，最寄りのゲートを使用している観客の具体的な割合はわかっていない．したがって，最寄りのゲート以外を使用する観客の割合が感染リスクの増加にもたらす影響を分析する必要がある．

本稿では，全観客が座席から最寄りのゲートを使用して退場した場合 [ゲート遵守] と，観客の一部がそれ以外のゲートをランダムに選択して退場した場合 [ランダム率 10%〜50%] の 2 つのシナリオを，CrowdWalk で再現する．それぞれのシナリオには，座席から最寄りのゲートを使用して退場するエージェントと，ランダムにゲートを選んで退場するエージェントの 2 種類が存在している．最寄りゲートから帰るエージェントの目的地としては座席から最寄りのゲート，ランダムに帰るエージェントの目的地としては会場内にあるゲートの中から表 1 のルールに従ってランダムに設定する．このルールでは，1 塁側，3 塁側で座席をエリア分けして，同一エリア内にあるゲートのみを選択可能となっている．この時 Gate 4 はどちらのエリアにも属しているとする．

5 実験結果

5.1 実験の設定

まずは近接関係ネットワークの描画設定について述べる．近接の条件を，歩行者間の距離が 1 メートル以内（$l \leqq 1$）の状態が 15 分（$m \geqq 15$）以上続いたものと定義する．この定義は，接触した相手が感染者だと仮定した場合の厚生労働省が定める濃厚接触の定義に準拠している．

本稿では，分析する歩行者の行動として，退場時に選択したゲート (Gate 1〜Gate 7 の 7 種類) をもとにノードの色を割り当てる．退場時のゲート選択を可視化することで，退場ゲートの選び方が近接に及ぼす影響を分析することが可能になる．退場時に選択したゲートにもとづいてノードを色分けした場合の，ノードの色とゲートの対応を図 2 に示す．本研究で再現したス

表 1: ゲート選択時の制約

Area	Closest gate name	Selectable gates
First base side	Gate 1	Gate 2, Gate 3, Gate 4
	Gate 2	Gate 1, Gate 3, Gate 4
	Gate 3	Gate 1, Gate 2, Gate 4
Center	Gate 4	Gate 1, Gate 2, Gate 3, Gate 5, Gate 6, Gate 7
Third base side	Gate 5	Gate 4, Gate 6, Gate 7
	Gate 6	Gate 4, Gate 5, Gate 7
	Gate 7	Gate 4, Gate 5, Gate 6

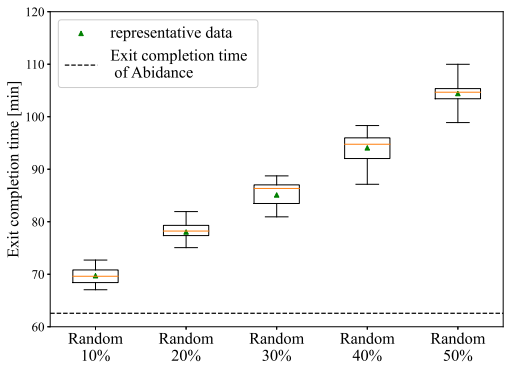


図 4: 各シナリオで 10 回ずつ群集シミュレーションを実行した際の退場完了時間に関する箱ひげ図．y 軸は退場完了時間を表示する．

タジアムには，Gate 4 を中心に左右に 7 つのゲートが存在している．

続いて，群集シミュレーションで生成した歩行者情報について述べる．群集シミュレーションは 4.2 節で説明したシナリオをそれぞれ 10 回ずつ実行した．各シナリオごとの退場完了時間に関する箱ひげ図を図 4 に示す．箱ひげの値は，最小値，25, 50, 75 パーセンタイル値，最大値を表している．この結果から，ゲートを遵守することで退場が完了するまでの時間を短縮できることがわかる．

歩行者の行動が近接状況に与える影響を分析するために，各シナリオごとの 10 回の群集シミュレーション結果のうち，可視化を行う代表データをひとつ選ぶ必要がある．しかし，歩行情報は表形式のデータであるため，代表データの選出にそのまま平均値を使うことができない．そこで退場完了時間が最も平均に近いデータを代表データとする．

5.2 可視化結果

図 5 は，ゲート遵守とランダム率を 10%〜50%に変化させて退場した場合のネットワークを，退場ゲートごとに色分けした結果を示す．それぞれのネットワークがもつノード・エッジの数を図 6 に示す．ここで本手法では，最寄りゲートが同じ座席エリアがボックスとして可視化される．なおボックスの大きさはボックス内のノード数を表現し，ボックスの配置はゲートの位置関係を表す．

結果から，ゲート遵守と比較して，ランダム率が高くなるほど，ノードの個数とエッジの本数が大きく増加していることがわかる．このことから，ゲートを遵守することでスタジアム内の混雑が抑制されることがわかる．また，ゲート遵守では同じボックス内でのみ近接が起きている．ゲート遵守では近接の発生が座席のエリアが同じ観客のみに限定されるため，感染者がいた場合に接触者の特定が容易となる．一方でランダム率の上昇とともに異なるボックス間を結ぶエッジの本数が増加していき，遠くの座席との接触が見られるようになる．

さらに，ランダム率の上昇で Gate 4 で近接が急増する．Gate 4 はスタジアムの中央に位置しており，ランダムに退場する観客が左右両方から集まってくる．これによって通路で混雑が発生し，近接が多くなっていることが考えられる．

6 ま と め

本稿では，群集シミュレーションによって大規模イベント会場における歩行情報データを取得し，近接状況を表すネットワークを生成することで，歩行者の行動が感染にもたらす影響を可視化する手法を提案した．近接関係ネットワークのノード配置に位置情報を持たせることで，重要な近接の特徴を観察できるようになった．さらに群集シミュレーションを用いることで，大規模イベント会場における全観客の退場時の様子を取得することができ，さらにパラメータを変更することで歩行者の行動が近接状況に及ぼす影響を分析することが可能となった．

実験では，全観客が座席ごとに定められているゲートから退場した場合と，観客の一部がゲートをランダムに選択して退場した場合の人流をシミュレーションした．退場ゲートにもとづいたネットワークの色付けによって，座席から最寄りのゲートを使用する観客の割合を上げることで感染リスクを抑えられることがわかった．

今後の課題としては以下がある．

- ネットワーク可視化の改善
- 群衆シミュレーションによる新たなシナリオの推定

現状の実装では，ネットワークはグループの視認性の向上を目的とした手法で描画されている．一方で，この可視化手法では個々の観客の近接を読み取ることが難しい．このように，今後の課題として，ネットワーク可視化についてさらなる改良が必要である．

別の課題として，群衆シミュレーションによる新たなシナリオの検証も検討している．例えば天候や試合結果などの環境設定や，トイレ・売店に立ち寄るなどの歩行者行動の再現である．また新たなシナリオをシミュレーションし，他の行動属性を可視化することを検討している．

文 献

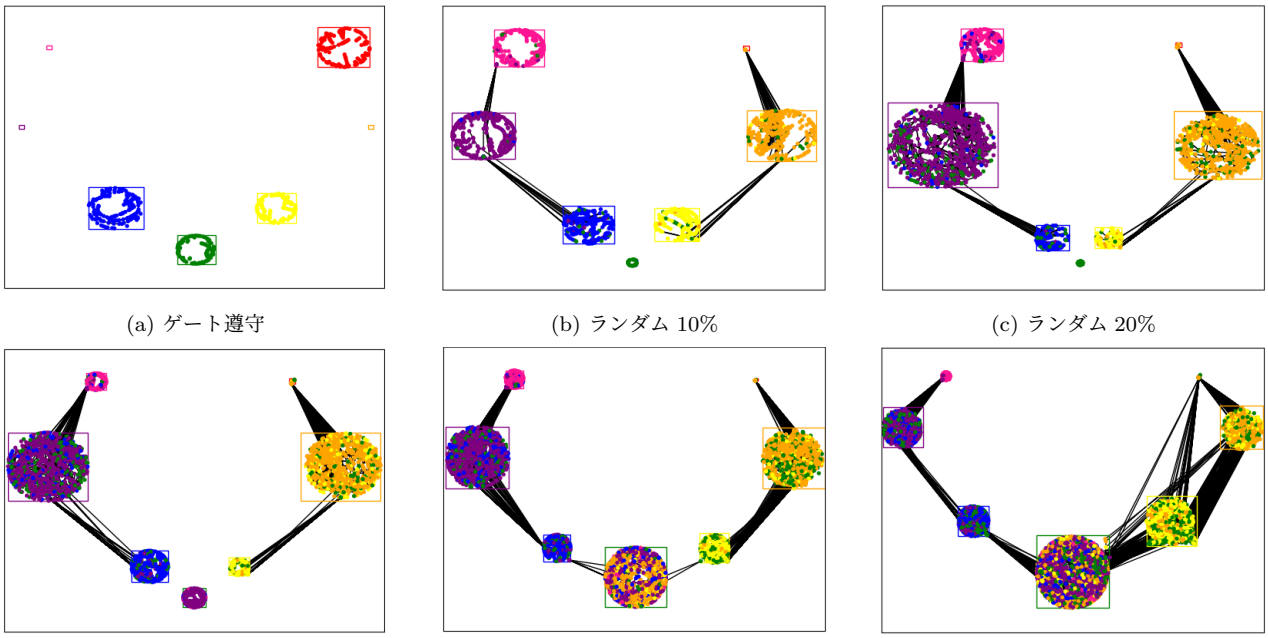
[1] Joacim Rocklöv and Henrik Sjödin, “High population densities catalyse the spread of covid-19”, *Journal of travel medicine*, vol. 27, no. 3, pp. taaa038, 2020.

[2] Sayaka Morikoshi, Masaki Onishi, and Takayuki Itoh, “Visualizing congestion at mass-gathering events with proximity-based networks”, in *2023 27th International Conference on Information Visualisation*. IEEE, 2023.

[3] Eduarda Mendes Rodrigues, Natasa Milic-Frayling, Marc Smith, Ben Shneiderman, and Derek Hansen, “Group-in-a-box layout for multi-faceted analysis of communities”, in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011, pp. 354–361.

[4] Gavin McArdle, Urška Demšar, Stefan van der Spek, and Seán McLoone, “Classifying pedestrian movement behaviour from gps trajectories using visualization and clustering”, *Annals of GIS*, vol. 20, no. 2, pp. 85–98, 2014.

[5] Aya Fukute, Takayuki Itoh, and Masaki Onishi, “A linked visualization of trajectory and flow quantity to support analysis of people flow”, in *2013 17th International Conference on Information Visualisation*. IEEE, 2013, pp. 561–567.



(a) ゲート遵守　(b) ランダム 10%　(c) ランダム 20%

(d) ランダム 30%　(e) ランダム 40%　(f) ランダム 50%

図 5: 退場ゲートごとに塗り分けた近接ネットワーク.

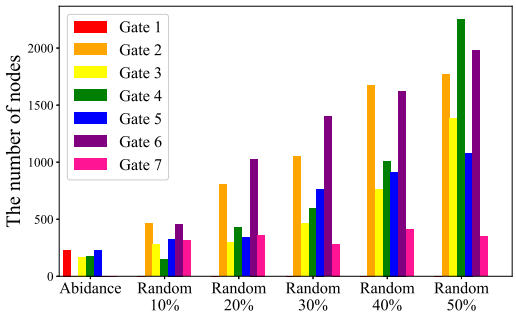


図 6: ゲートごとのノード数.

[6] Jia Wang, Zena Wood, and Mike Worboys, "Conflict in pedestrian networks", in *Geospatial Data in a Changing World: Selected Papers of the 19th AGILE Conference on Geographic Information Science*. Springer, 2016, pp. 261–278.

[7] Mikiharu Arimura, Tran Vinh Ha, Kota Okumura, and Takumi Asada, "Changes in urban mobility in sapporo city, japan due to the covid-19 emergency declarations", *Transportation research interdisciplinary perspectives*, vol. 7, pp. 100212, 2020.

[8] Yihang Li and Liyan Xu, "The impact of covid-19 on pedestrian flow patterns in urban pois—an example from beijing", *ISPRS International Journal of Geo-Information*, vol. 10, no. 7, pp. 479, 2021.

[9] Reza Askarizad, He Jinliao, and Setareh Jafari, "The influence of covid-19 on the societal mobility of urban spaces", *Cities*, vol. 119, pp. 103388, 2021.

[10] Shahadat Uddin, Tasadduq Imam, Matloob Khushi, Arif Khan, and Mohammad Ali Moni, "How did socio-demographic status and personal attributes influence compliance to covid-19 preventive behaviours during the early outbreak in japan? lessons for pandemic management", *Personality and Individual Differences*, vol. 175, pp. 110692, 2021.

[11] Eva Turk, Tatjana Čelik, Maja Smrdu, Janko Šet, Ana Kuder, Matjaž Gregorič, and Simona Kralj-Fišer, "Adherence to covid-19 mitigation measures: The role of sociodemographic and personality factors", *Current Psychology*, vol. 42, no. 9, pp. 7771–7787, 2023.

[12] Hajime Ando, Kazunori Ikegami, Tomohisa Nagata, Seiichiro Tateishi, Hisashi Eguchi, Mayumi Tsuji, Shinya Matsuda, Yoshihisa Fujino, Akira Ogami, and COROnaWork Project, "Effect of commuting on the risk of covid-19 and covid-19-induced anxiety in japan, december 2020", *Archives of Public Health*, vol. 79, pp. 1–10, 2021.

[13] Takeshi Kimura, Tomonori Sano, Kazuto Hayashida, Naohiro Takeichi, Yoshikazu Minegishi, Yoshiyuki Yoshida, and Hitoshi Watanabe, "Representing crowds using a multi-agent model–development of the simtread pedestrian simulation system", *Japan Architectural Review*, vol. 2, no. 1, pp. 101–110, 2019.

[14] Susumu Yamakage, *Modeling and expanding artificial societies : introduction to Multi-Agent Simulation with artisoc*, Number 01 in The possibilities of an artificial society. Kozo Keikaku Engineering,Shosekikobo Hayama publishing [distributor], 2009.

[15] Dirk Helbing and Peter Molnar, "Social force model for pedestrian dynamics", *Physical review E*, vol. 51, no. 5, pp. 4282, 1995.

[16] Tomohisa Yamashita, Takashi Okada, and Itsuki Noda, "Implementation of simulation environment for exhaustive analysis of huge-scale pedestrian flow", *SICE Journal of Control, Measurement, and System Integration*, vol. 6, no. 2, pp. 137–146, 2013.

[17] Itsuki Noda and Tomohisa Yamashita, "Pedestrian simulator with flexible framework to enhance detailed behavior and environmental change", *Artificial Life and Robotics*, vol. 22, pp. 308–315, 2017.

[18] Thomas MJ Fruchterman and Edward M Reingold, "Graph drawing by force-directed placement", *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[19] Glenn Jocher, "Ultralytics yolov5", *GitHub repository: https://github. com/ultralytics/yolov5*, 2020.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 13:30 ~ 15:40 | T1-A オンライン（Zoom Events)

汎用機械学習技術 (2)

座長:小山田 昌史(NEC)
コメンテータ:古賀 久志(電気通信大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-A-2-01] ミスラベルデータの忘却による学習済みモデルの汎化性能の向上手法の提案

*杉浦 一瑳[1]、岡村 真吾[2,1]、山下 恭佑[1]、矢内 直人[1] (1. 大阪大学、2. 奈良工業高等専門学校)

14:00 ~ 14:25
[T1-A-2-02] 教師あり対照学習に基づく次数公平なグラフ深層学習

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

14:25 ~ 14:50
[T1-A-2-03] 機械学習のためのアクティブ・データプール拡張手法

*京谷 辰哉[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学 情報学群 情報メディア創成学類、2. 筑波大学 図書館情報メディア系)

14:50 ~ 15:15
[T1-A-2-04] AGATE : 属性付き時間発展グラフの汎用予測フレームワーク

*山崎 翔平[1]、佐々木 勇和[2]、Karras Panagiotis[3]、鬼塚 真[2] (1. 株式会社野村総合研究所、2. 大阪大学、3. Aarhus University)

15:15 ~ 15:40
[T1-A-2-05] 複数タスクを対象としたクラス知識のメタ学習にもとづくゼロショット学習

*南部 優太[1]、幸島 匡宏[1]、岩田 具治[2]、山本 隆二[1] (1. 日本電信電話株式会社 NTT人間情報研究所、2. 日本電信電話株式会社 NTTコミュニケーション科学基礎研究所)

ミスラベルデータの忘却による学習済みモデルの汎化性能の向上手法の提案

杉浦　一瑳† 　岡村　真吾†† 　山下　恭佑† 　矢内　直人†

† 大阪大学 〒 565-0871 大阪府吹田市山田丘 1-1
†† 奈良工業高等専門学校 〒 639-1080 奈良県大和郡山市矢田町 22 番地
E-mail: †{i-sugiura,yamashita,yanai}@ist.osaka-u.ac.jp, ††okamura@info.nara-k.ac.jp

あらまし　機械学習モデルはミスラベルデータを学習することで汎化性能が低下する．ミスラベルデータに対する既存の対策手法は，学習前あるいは学習中の処理が前提となっていた．本稿では，ミスラベルデータの忘却による学習済みモデルの汎化性能の向上手法を提案する．提案手法は大まかには，まず学習済みモデルから計算できる異常度を通じてミスラベルデータを検知したのち，検知したデータをモデルから忘却 (Machine Unlearning) させることで，汎化性能を向上させる．提案手法について，ベンチマークとして MNIST データセットを用いた分類モデルによる実験を行った．手法の有効性を詳細に評価するために，まず検知と忘却に分けてそれぞれ実験したところ，ミスラベルデータの混入割合が少ないときに検知が有効であること，また，ミスラベルデータを忘却することで汎化性能が向上することを確認した．しかし，提案手法に相当する検知と忘却を繋げた実験では，ミスラベルデータの混入割合に応じて検知の性能が低下することに起因し，汎化性能の向上が妨げられることがわかった．

キーワード　ミスラベルデータ，Machine Unlearning，検知，忘却

1　はじめに

機械学習は訓練データの増大とともに，性能の向上を遂げてきた．しかし，クラウドソーシングを活用して収集された大量のデータにはラベルの誤ったデータ (ミスラベルデータ) を含んでしまうことや [1–3]，ラベルを意図的に変えたデータを訓練データの一部に加える訓練データ汚染攻撃 [4, 5] の存在などの理由から，正確にラベル付けされた質の良い訓練データの増大は容易ではない．ミスラベルデータはモデルの汎化性能を低下させることや [6, 7]，プライバシー漏洩を助長させることが指摘されており [8–10]，ミスラベルデータへの対策は重要である．

上述した背景に対し，訓練データセットに含まれるミスラベルデータの影響を抑える手法が研究されてきた [11–15]．しかし，これらはモデルの学習における前処理，あるいは学習中の処理として用いられる手法であり，既存の学習済みモデルに対して事後的に適用することが難しい．近年は Hugging Face [16] や GitHub [17] などのプラットフォームを通じて学習済みモデルを公開することが増えているため，既存の学習済みモデルに対しても適用可能な手法が望ましい．

本稿では，ミスラベルデータを含んだデータセットを用いて学習した学習済みのモデルに対し，事後的にミスラベルデータを検知し，そのミスラベルデータを除いて再学習するよりも効率的に忘却させる手法を提案する．提案手法の概要を図 1.1 に示す．提案手法は大まかに検知，忘却の 2 段階で構成される．具体的には，まず訓練データの各サンプルに対して損失や影響関数 [18] 等の異常度を計算することでミスラベルデータを検知する．その後，検知したデータの影響を取り除くようにモデルのパラメータを更新することで，ミスラベルデータを忘却する．忘却には，学習済みモデルから訓練データの一部を忘れさせる Machine Unlearning という分野において提案されている近似的な忘却手法を用いる [18, 19]．本提案手法により，既存の学習済みモデルに対して，ミスラベルデータの影響を取り除き，汎化性能を向上させる．

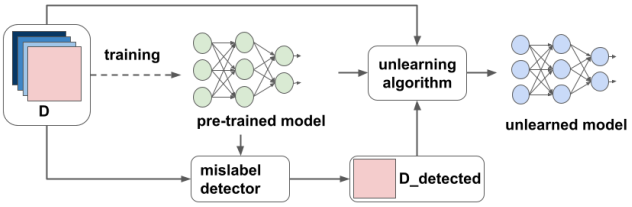


図 1.1: 提案手法の概要

提案手法の有効性を確認するため，分類モデルを対象にベンチマークである MNIST データセットを用いて実験を行った．はじめに，提案手法の有効性を詳細に分析するため，提案手法の段階ごとに実験を行った．結果として，まずミスラベルデータの混入割合が少ないときにミスラベルデータの検知が有効であること，また，パラメータを更新することでミスラベルデータの影響を抑えられることをそれぞれ確認した．しかし，提案手法に相当するミスラベルデータの検知とパラメータ更新を繋げた実験では，提案手法の効果が確認できなかった．この原因として，混入割合が増加するにつれて，検知の段階で，ミスラベルデータの検知の性能が下がっていることを確認した．

本稿の貢献を要約すると，以下のとおりである．

- ミスラベルデータを効率的に忘却する手法を提案した.
- 分類モデルを対象に実験を行ったところ, ミスラベルデータの混入割合が少ないときにミスラベルデータの検知が有効であることを確認した.
- 近似的な忘却手法を用いてミスラベルデータを忘却すると汎化性能が向上することを確認した.
- 提案手法に相当するミスラベルデータの検知とパラメータ更新を繋げた実験では, 汎化性能の向上が確認できなかった.
- 主な原因として, 混入割合が増加するにつれ, 元となるミスラベルデータの検知性能が下がっていることを確認した.

2 関 連 研 究

本節ではミスラベルデータを伴う学習手法, および, 忘却手法としての Machine Unlearning についてそれぞれ述べる. 次に, 本研究に最も近い関連研究として, 機械学習モデルの復元手法について述べる.

2.1 ミスラベルデータを伴う学習手法

ミスラベルデータを伴う学習手法は, 大量のデータを必要とするがラベルにノイズが含まれる可能性があるという状況で, 性能の高いモデルを得るために研究されてきた [20, 21]. とくに前処理にあたるミスラベルデータを検知する手法 [13–15] や, 中処理にあたるミスラベルデータの影響を軽減する手法 [11, 12] が提案されている.

ミスラベルデータを前処理として検知する手法として, モデルの学習中の予測値 $f(x;\theta)$ の時間発展がミスラベルデータと正しくラベル付けされたデータで異なることを利用して, ミスラベルデータを検知する手法が提案されている [14, 15].

ミスラベルデータの影響を中処理として軽減する手法としては, ラベルの遷移確率を扱う手法 [11] や, ラベル誤りにロバストな損失関数を設計する手法 [12] が存在する. しかし, これらの手法は学習方法をあらかじめ変更することを前提としているため, 本稿とは問題設定が異なっている.

2.2 Machine Unlearning

Machine Unlearning [22] は, 機械学習モデルから訓練データの一部の影響を取り除く手法である. 元はプライバシーの文脈で使われており, 対象のデータがモデルの訓練データに含まれているか否かを推論するメンバーシップ推論 [23] や, モデルから学習データを復元するデータ抽出 [24] など, 機械学習モデルに対するプライバシー攻撃の危険性が注目されるにつれて重要性が増している. 近年では学習データの汚染攻撃の対策としても注目されている [25]. Machine Unlearning を実現する素朴な方法は, 忘れさせたいデータを取り除いた上で再学習を行うことである. しかし, この方法は時間的および計算的コストが大きいため, 現実的ではない. そのため, 効率的に Machine Unlearning を実現する方法が研究されている [18, 19, 26]. 本稿では, 学習手法を変更する必要のない忘却手法として, 影響関数を用いた手法 [18], および, ニュートン法を用いた手法 [19] を用いる. 以降ではそれらについて説明する.

2.2.1 影響関数 (Influence Function) を用いた手法

影響関数は Interpretable ML でも用いられる手法で, データセット中のデータ z のモデルへの影響を近似的に図る手法である [18].

影響関数では, 訓練データセットを $\mathcal{D} = \{z_i = (x_i, y_i)\}_{i=1}^n$ としたときのデータ $z \in \mathcal{D}$ の損失を ϵ で重み付けして増やしたときのパラメータの変化を考える. すなわち,

$$\hat{\theta}_{\epsilon,z} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta) \tag{2.1}$$

としたとき, $\hat{\theta}_{\epsilon,z} - \hat{\theta}$ を考える. このとき, 以下のように近似的にパラメータの ϵ に関する $\hat{\theta}_{\epsilon,z} - \hat{\theta}$ の微分を計算できる.

$$\mathcal{I}_{up,params}(z) := \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta L(z,\hat{\theta}) \tag{2.2}$$

ただし, $H_{\hat{\theta}} := \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 L(z_i,\hat{\theta})$. データ z を削除することは式 (2.1) において $\epsilon = -\frac{1}{n}$ とすることと同じであるため, データ z を取り除いた上で再学習した結果得られる最適解は, $\hat{\theta}^{\setminus z}$ $\hat{\theta}^{\setminus z} \approx \hat{\theta} - \frac{1}{n}\mathcal{I}_{up,params}(z)$ で計算できる. ただし, 上記の方法は近似であり, 導出における仮定に沿わない深層学習モデルでは誤差が大きくなる. また, 逆行列 $H_{\hat{\theta}}^{-1}$ の計算コストが大きいという問題がある.

2.2.2 ニュートン法を用いた手法 (SSSE)

ニュートン法を用いた手法 (SSSE) は, 忘却対象のデータを $\mathcal{D}_{forget}, |\mathcal{D}_{forget}| = m$ とした時, 以下のようにパラメータを更新する手法である [19].

$$\hat{\theta}^{\setminus\mathcal{D}_{forget}} = \hat{\theta} + \frac{1}{n-m}H_{\hat{\theta},\mathcal{D}\setminus\mathcal{D}_{forget}}^{-1}\nabla_\theta L(\mathcal{D}_{forget},\hat{\theta}) \tag{2.3}$$

ただし, $H_{\hat{\theta},\mathcal{D}\setminus\mathcal{D}_{forget}}^{-1} := \frac{1}{n-m}\sum_{(x_i,y_i)\in\mathcal{D}\setminus\mathcal{D}_{forget}}\nabla_\theta^2 L(z_i,\hat{\theta})$. この方法も $H_{\hat{\theta},\mathcal{D}\setminus\mathcal{D}_{forget}}^{-1}$ の計算コストが大きいが, 計算を工夫することで計算コストを削減することができ, 再学習による忘却に比べて効率的に忘却することができると考えられている.

2.3 ノイズを含むデータで学習した学習済みモデルの修正手法

ノイズを含むデータで学習した学習済みモデルの修正手法は, 訓練データ汚染攻撃 [27] に対する防御手法として広く研究されてきた. 訓練データ汚染攻撃とは, 攻撃者が訓練データセットにラベルを意図的に変更したり, トリガーを含めたりしてモデルの挙動を変える攻撃である. 訓練データ汚染攻撃に対する防御手法として, 汚染データの検知およびモデルの修正が提案されている [28–30].

KARMA [28] は元のデータセットと誤分類結果のレポートを入力とし, 誤分類の原因となったデータを特定し, 忘れさせるフレームワークである. KARMA と提案手法との比較を表 2.1

表 2.1: 既存研究との比較

	モデル	目的	忘却対象
KARMA [28]	SVM 等	誤推論結果の修正	誤推論の原因のデータ
提案手法	DNN	汎化性能の向上	ミスラベルデータ

に示す. 本稿の研究では誤分類結果ではなくモデルの汎化性能を向上することに焦点が置かれている点, 想定するモデルを深層学習モデルとしている点で, 問題設定が KARMA と異なっている. すなわち, 本稿の方がより一般化された問題とみなせる.

また, トリガーを含む汚染データに対する検知, 修正手法が提案されている [29,30]. これらの研究において, 修正方法として汚染データを取り除いた上で再学習する手法が用いられている. これに対し, 本稿ではモデルの修正として近似的な Machine Unlearning を取り入れた点に新規性がある.

3 問題設定

本章では, 提案手法が必要となる問題設定について説明する.

3.1 問題の定式化

以下に, 問題設定を定式化する. 以降では, クラス分類問題の教師あり学習とする. まず入力データの集合 $\mathcal{X}$, ラベルの集合 $\mathcal{Y}$, 機械学習モデルの訓練データセットを $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ とする. ここで機械学習モデルは $f : \mathcal{X} \times \Theta \to \mathcal{Y}$ なる写像であり, データセット $\mathcal{D}$ のうち, ミスラベルデータからなるデータセットを $\mathcal{D}_{mislabeled} \subset \mathcal{D}$, そうでないサンプル（便宜上, クリーンデータと呼称）からなるデータセットを $\mathcal{D}_{clean} \subset \mathcal{D}$ とし, このとき $\mathcal{D} = \mathcal{D}_{mislabeled} \cup \mathcal{D}_{clean}$ となる. 各サンプル $z_i = (x_i, y_i)$ の損失を $L(z_i, \theta)$ とし, $L(\mathcal{D}, \theta) := \frac{1}{n}\sum_i^n L(z_i, \theta)$ とする. データセット $\mathcal{D}$ を用いてパラメータ θ を学習した結果得られたパラメータの値を $\hat{\theta}$ とし, 学習済みモデルを $f(\cdot; \hat{\theta})$ とする.

本研究では, データセット $\mathcal{D}$, 学習済みモデル $f(\cdot; \hat{\theta})$, ミスラベルデータの混入割合 (noise ratio) $r = \frac{|\mathcal{D}_{mislabeled}|}{|\mathcal{D}|}$ から, $\mathcal{D}_{mislabeled}$ を可能な限り忘却することで, モデルの汎化性能を向上させることを目指す. ここでいう忘却とは, Machine Unlearning の研究で考えられている, $\mathcal{D}$ を用いて学習した学習済みモデルのパラメータを, 忘却対象のデータ $\mathcal{D}_{forget}$ に対して, $\mathcal{D} \setminus \mathcal{D}_{forget}$ を用いて学習した結果得られるパラメータに近づけるという意味とする [31].

なお, データセット $\mathcal{D}$ を知っているという状況は, 学習済みモデルの提供者や, 提供時に用いたデータセットがオープンであり公開されていた場合に起きる. また, 混入割合 r を知っているという状況は, 混入割合を正確に推定するのは困難であるが, データセットの数が大きい場合, データセットの一部をとってきて混入割合を調査し, 全体の混入割合を推定することで可能となる.

3.2 ユースケース

本問題設定のユースケースは以下が考えられる. (1): 意図の有無に関わらずミスラベルデータが訓練データセットに含まれる. (2): モデル提供者が訓練データセットを用いて学習済みモデル, 使用した訓練データセットを Hugging Face や GitHub に公開する. (なお, Hugging Face はデータセット [1]やモデル [2] 等を公開することのできるプラットフォームである. 例として, LLM 勉強会 [3]は, Hugging Face 上で学習済みモデルを公開しており, その際に使用した訓練データとデータ収集方法を再現可能な形で公開している [4].) (3): モデル提供者, または利用者が本提案手法を用いてモデルの汎化性能を向上する.

4 提案手法

本節では提案手法について述べる. まずは全体像を述べたのち, 手法を詳細に述べる.

4.1 全体像

提案手法は大まかには, (1) ミスラベルデータの検知と, (2) その忘却という 2 つの段階からなる.

検知はまずモデルのパラメータに対し, 入力として与えられたサンプルとそのラベルの異常度を計算し, 異常度の大きいサンプルを忘却の対象とする. 異常度の計算は, モデルを通じて計算される微分値を用いることで, 学習済みモデルに対しても可能である. 具体的には損失 (Loss), あるいは影響関数 (Influence Function) を用いる.

忘却は, 検知の段階で忘却の対象として挙げられたサンプルの影響を取り除くよう学習済みモデルのパラメータを更新する. なお, パラメータの更新には, 学習済みモデルから効率的に訓練データの一部を忘れさせる Machine Unlearning [31] という分野で提案されている, 近似的な忘却手法を用いる.

4.2 手法の詳細

提案手法の詳細として, ミスラベルデータの検知段階と忘却段階の具体的な手法をそれぞれ述べたのち, 提案手法について述べる.

4.2.1 ミスラベルデータの検知

ミスラベルデータの検知をアルゴリズム 1 に示す. このアルゴリズムでは訓練データセット $\mathcal{D}$, 学習済みモデル $f(\cdot; \hat{\theta})$, ミスラベルデータの混入割合 r を入力として, 各データ $z_i = (x_i, y_i)$ に対して, 異常度 $a(z_i, \hat{\theta})$ を計算し, 異常度の上位 $m = |D| * r$ 個のデータをミスラベルデータとみなす. 異常度の計算には, 既存研究 [14,15] をもとに考えた損失に基づく手法, あるいは, 影響関数に基づく手法 [22] をそれぞれ用いることができる.

a） 損失に基づく検知手法

ミスラベルデータはエポック数が十分に大きくなるまで, クリーンデータに比べて損失が大きいことが実験的に知られている [14,15]. そこで, ミスラベルデータがモデルに過学習されていないことを仮定し, 各データ $z_i = (x_i, y_i) \in \mathcal{D}$ に対して, 損

1：https://huggingface.co/datasets
2：https://huggingface.co/models
3：https://llm-jp.nii.ac.jp/
4：https://huggingface.co/llm-jp/llm-jp-13b-v1.0

Algorithm 1 ミスラベルデータの検知

Input: 訓練データセット $\mathcal{D}$, 学習済みモデル $f(\cdot;\hat{\theta})$, 混入割合 r
Output: ミスラベルデータとして検知された m 個のデータ $\mathcal{D}_{detected}$

1: $\mathcal{D}_{detected} \leftarrow \emptyset$
2: **for** $i = 1, 2, \ldots, n$ **do**
3: $z_i = (x_i, y_i)$ の異常度 $a(z_i, \hat{\theta})$ を計算
4: **end for**
5: $\mathcal{D}_{detected} \leftarrow$ 異常度の上位 $m = |D| * r$ 個のデータ z_i を $\mathcal{D}_{detected}$ に追加

Algorithm 2 データの忘却

Input: 学習済みモデル $f(\cdot, \hat{\theta})$, $\mathcal{D}$, $\mathcal{D}_{forget}$
Output: $\hat{\theta}^{\setminus \mathcal{D}_{forget}}$

1: 忘却アルゴリズムを U とする.
2: $\hat{\theta}^{\setminus \mathcal{D}_{forget}} \leftarrow U(D, \mathcal{D}_{forget}, \hat{\theta})$

失 $L(z_i, \hat{\theta})$ を計算することで, その値を異常度とする. すなわち, $z \in \mathcal{D}$ の異常度 $a(z, \hat{\theta})$ を以下の通り定義する:

$$a(z, \hat{\theta}) := L(z, \hat{\theta}). \tag{4.1}$$

b) 影響関数に基づく検知手法

影響関数を用いた手法は, あるデータ $z_i = (x_i, y_i) \in \mathcal{D}$ をデータセットから除いたときの, モデルのパラメータ $\hat{\theta}$ の変化量を近似的に計算する手法である [18]. テストデータ $z_{test} = (x_{test}, y_{test})$[5]に対する z を取り除いた時の損失の変化量を, 以下の式で近似的に計算することができる.

$$\begin{aligned} I_{up,loss}(z_{test}, z) &:= L(z_{test}, \hat{\theta}) - L(z_{test}, \hat{\theta}^{\setminus z}) \\ &= \frac{1}{n} \nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}), \end{aligned} \tag{4.2}$$

ただし,

$$H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta}) \tag{4.3}$$

である. そこで, 損失の変化量を異常度として捉えるため, 異常度 $a(z, \hat{\theta})$ を以下のように定義する.

$$a(z, \hat{\theta}) := \nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}) \tag{4.4}$$

なお, 式 (4.2) の $\frac{1}{n}$ は各サンプルに対する異常度の大小関係に影響しないため, 無視している. 次の章の実験では, テストデータ z_{test} を, z, D, $D \setminus z$ の 3 つを用いて検知性能を比較する. $I_{up,loss}(z_{test}, z)$ の z_{test} として, $z, D_{train}, D_{train} \setminus z$ とした時の値を $I_{up,loss}(z, z)$, $I_{up,loss}(\mathcal{D}_{train}, z)$, $I_{up,loss}(\mathcal{D}_{train} \setminus z, z)$ として検知性能を比較する.

4.2.2 データの忘却

データの忘却をアルゴリズム 2 に示す. このアルゴリズムでは, 忘却対象のデータセット $\mathcal{D}_{forget}$, 学習済みモデル $f(\cdot, \hat{\theta})$, 訓練に用いたデータセット $\mathcal{D}$ を入力として, $\mathcal{D}_{forget}$ の影響を取り除くようにパラメータ $\hat{\theta}$ を更新する. 忘却アルゴリズムとしては, 近似的な忘却アルゴリズムである, ファインチューニング, 影響関数に基づく手法 [18], ニュートン法に基づく手法 [19] をそれぞれ用いる. なお, 実験では理想的な忘却手法である再学習を比較対象として用いる.

a) 再 学 習

再学習は, $\mathcal{D}_{forget}$ を除いたデータセット $D \setminus \mathcal{D}_{forget}$ を用いて, モデルのパラメータを初期化して初めから学習する手法である. Machine Unlearning 研究において, 理想的な忘却手法として比較対象に用いられる [31].

b) ファインチューニング

ファインチューニングは, パラメータの初期値を $\hat{\theta}$ として, $\mathcal{D}_{forget}$ を除いたデータセット $\mathcal{D} \setminus \mathcal{D}_{forget}$ を用いて, モデルのパラメータを追加学習する手法である.

c) 影響関数に基づく忘却手法

影響関数に基づく忘却手法は, 忘れさせたいデータを $z \in \mathcal{D}$ としたとき, 以下のようにパラメータを更新する [18].

$$\hat{\theta}^{\setminus \mathcal{D}_{forget}} = \hat{\theta} + \frac{1}{n} H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}) \tag{4.5}$$

d) ニュートン法に基づく忘却手法 (SSSE)

ニュートン法を用いた手法は, 忘れさせたいデータ $\mathcal{D}_{forget}, (|\mathcal{D}_{forget}| = m)$ とした時, 以下のようにパラメータを更新する [19].

$$\hat{\theta}^{\setminus \mathcal{D}_{forget}} = \hat{\theta} + \frac{1}{n-m} H_{\hat{\theta}, \mathcal{D} \setminus \mathcal{D}_{forget}}^{-1} \nabla_\theta L(\mathcal{D}_{forget}, \hat{\theta}). \tag{4.6}$$

4.2.3 ミスラベルデータの検知と忘却を繋げたモデル修正

提案手法は, 図 1.1 のように, ミスラベルデータの検知と忘却を繋げてモデルの汎化性能をを向上する. すなわち, ミスラベルデータの検知によって得られた $\mathcal{D}_{detected}$ を, 忘却における忘却対象 $D_{forget} = \mathcal{D}_{detected}$ として忘却する.

5 実 験

本節では提案手法を実験により評価する. まず実験の目的を述べたのち, 設定を述べる. 最後に, 実験結果とその考察を併せて述べる.

5.1 実験目的

実験の目的は提案手法の有効性を確認することにあるが, 提案手法の各段階で, それぞれ実験を行う. すなわち, 目的は以下の三点を明らかにすることにある: 1) 学習済みモデルを用いて事後的なミスラベルデータの検知は可能か; 2) ミスラベルデータに関連するパラメータを更新することで推論精度は向上するか; 3) ミスラベルデータの検知とパラメータの更新を繋げることで推論精度は向上するか. とくに上述した 3) は提案手法の全体の処理に相当することから, これが明らかにできたとき, 提案手法の有効性は確認できたとみなせる. また, 1) と 2) は提案手法の各段階に問題がないか, 明らかにする意図がある. 以降では上述した目的ごとに実験をそれぞれ行う.

5.2 実験設定

本実験では, 手書き数字認識タスク用のデータセットである

5: ただし, テストデータ z_{test} は与えられていないため, z_{test} をどのように選ぶかが問題となる.

MNIST [32] を使用する．MNIST は各画像が 28 × 28 ピクセルのグレースケール画像で，0 から 9 までの 10 クラスのラベルが付与される．あらかじめ訓練データとテストデータに分割され，訓練データには 60,000 枚の画像が，テストデータには 10,000 枚の画像が含まれる．本実験では，簡単のため，0, 1 の 2 クラス分類問題として実験を行う．訓練データは各クラスのデータを 1000 個ずつ用いることにする．

ミスラベルデータを含んだデータセットの構築は，混入割合 r が与えられたとき，各ラベル y_i を以下のように変更することで行う．

$$y_i = \begin{cases} y_i & \text{with probability } 1-r, \\ 1-y_i & \text{with probability } r. \end{cases} \tag{5.1}$$

すなわち，確率 r でラベルを入れ替える．このノイズ混入方法は symmetric lable noise という名前でミスラベル研究でよく用いられている [13].

分類モデルは中間層が 2 層の Multi Layer Perceptron(MLP)，損失関数はクロスエントロピー誤差を用いる．中間層の次元数はそれぞれ 50 とし，実装は自動微分を備えた数値計算ライブラリである JAX [33]，最適化，深層学習用の JAX の派生ライブラリである Optax [34]，Flax [35] を用いた．最適化手法は SGD(確率的勾配降下法)，学習率は 0.01，バッチサイズは 32，エポック数は 20，モメンタムは 0.9 とした．

なお，本実験を再現可能なコードを GitHub にて公開している [6].

5.3 実験結果

以下に各実験の結果と考察をそれぞれ記載する．

5.3.1 実験 1: 学習済みモデルを用いて事後的なミスラベルデータの検出は有効か

学習済みモデルを用いて事後的なミスラベルデータの検出が有効であるか調べるため，混入割合 $r = \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ とし，各検知手法の性能を ROC AUC，正解率 (accuracy)，適合率 (precision) で測る．なお，今回はミスラベルとして検知する数を混入割合 r としており，この場合適合率 (precision) と再現率 (recall) が一致するため，適合率のみ示す．

混入割合に伴う各検知手法の性能を表 5.1 に示す．いずれの手法も，混入割合 r が増加するにつれて，検知性能が低下していることがわかる．また，損失を用いた手法が最も良い性能を出していることがわかる．

各サンプルの異常度の分布を可視化するため，混入割合 $r = 0.05$ の場合の学習後の各サンプルに対する損失，$I_{up,loss}(z_{test}, z)$ の z_{test} に $z, D_{train}, D_{train} \setminus z$ とした時の $I_{up,loss}(z, z)$, $I_{up,loss}(\mathcal{D}_{train}, z)$, $I_{up,loss}(\mathcal{D}_{train} \setminus z, z)$ のヒストグラムをそれぞれ図 5.1, 図 5.2, 図 5.3, 図 5.4 に示す．混入割合が小さい場合は，ミスラベルデータの異常度が大きく，ミスラベルデータとクリーンデータを分離することができることがわかる．

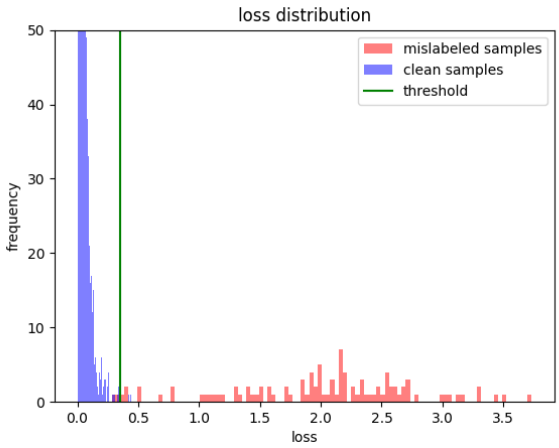


図 5.1: 混入割合 $r = 0.05$ の学習後の各サンプルに対する損失のヒストグラム．見やすさのため，y 軸の上限を 50 までとしている．

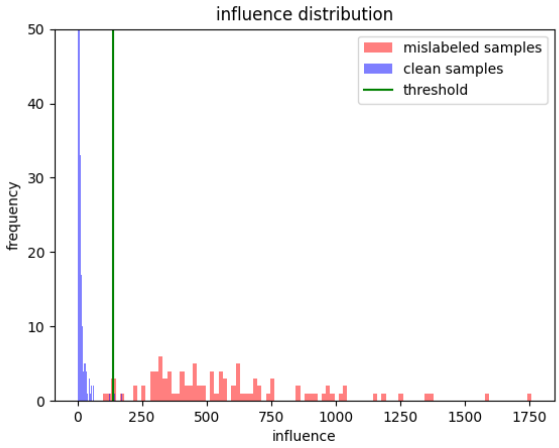


図 5.2: 混入割合 $r = 0.05$ の学習後の各サンプルに対する $I_{up,loss}(\mathcal{D}_{train} \setminus z, z)$ のヒストグラム．見やすさのため，y 軸の上限を 50 までとしている．

わかる．

以上から，混入割合が小さい場合は，ミスラベルデータの検知が有効であるが，混入割合が大きくなるにつれ，ミスラベルデータの検知性能が低下することがわかった．この結果は文献 [15] の結果と一致している．

5.3.2 実験 2: ミスラベルデータに関連するパラメータを更新することで推論精度は向上するか

ミスラベルデータを学習済みモデルから忘却することによりモデルの推論精度が向上するか検証する．本実験ではミスラベルデータ $\mathcal{D}_{mislabeled} \subseteq \mathcal{D}$ が既知とし，$\mathcal{D}_{forget} = \mathcal{D}_{mislabeled}$ を忘却することで，推論精度が向上するか検証する．ここでは混入割合 $r = \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ とし，忘却後の性能を正解率で測ることにより，推論精度が向上するか検証する．

各忘却手法の性能を表 5.2 に示す．混入割合 r が {0.01, 0.05, 0.1} の時は，忘却前のモデル (Pre-trained) の性能がすでに高いため，忘却による性能向上はあまり見られなかった．しかし，

6：https://github.com/speed1313/mislabel-unlearning

表 5.1: ミスラベルデータの検知性能.

	Method	混入割合 r 0.01	0.05	0.1	0.2	0.3	0.4	0.5
ROC AUC (↑)	Loss	1.0	0.99	0.99	0.96	0.83	0.73	0.38
	$I_{up,loss}(z,z)$	1.0	0.99	0.99	0.96	0.83	0.72	0.36
	$I_{up,loss}(\mathcal{D}_{train},z)$	1.0	0.83	0.80	0.64	0.55	0.54	0.36
	$I_{up,loss}(\mathcal{D}_{train}\setminus z,z)$	0.95	0.71	0.75	0.60	0.54	0.53	0.37
accuracy (↑)	Loss	100	99	98	93	79	69	41
	$I_{up,loss}(z,z)$	100	99	98	92	78	67	40
	$I_{up,loss}(\mathcal{D}_{train},z)$	100	97	94	79	63	57	39
	$I_{up,loss}(\mathcal{D}_{train}\setminus z,z)$	99	97	93	77	63	55	39
precision (↑)	Loss	100	98	91	83	65	61	41
	$I_{up,loss}(z,z)$	100	96	91	81	63	59	40
	$I_{up,loss}(\mathcal{D}_{train},z)$	100	73	72	47	38	46	39
	$I_{up,loss}(\mathcal{D}_{train}\setminus z,z)$	85	66	65	43	38	44	39

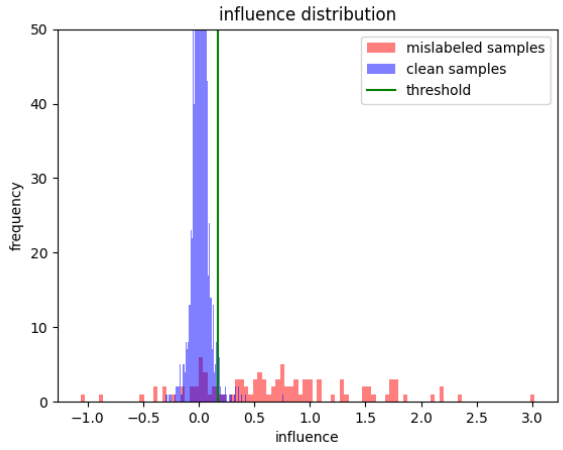


図 5.3: 混入割合 $r = 0.05$ の学習後の各サンプルに対する $I_{up,loss}(\mathcal{D}_{train},z)$ のヒストグラム. 見やすさのため, y 軸の上限を 50 までとしている.

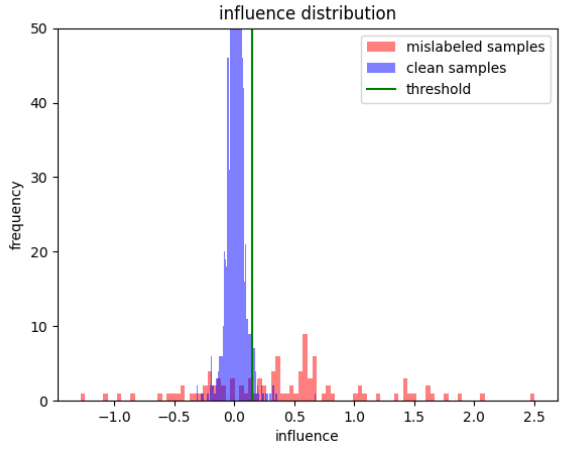


図 5.4: 混入割合 $r = 0.05$ の学習後の各サンプルに対する $I_{up,loss}(\mathcal{D}_{train}\setminus z,z)$ のヒストグラム. 見やすさのため, y 軸の上限を 50 までとしている.

混入割合 r が {0.2, 0.3, 0.4, 0.5} のとき, 忘却により推論精度が向上していることがわかる. 以上から, 混入割合 r が大きい場合, ミスラベルデータを忘却することにより, 推論精度が向上することがわかった. なお, Machine Unlearning の研究においては, 忘却対象の数が多いほど忘却後の推論精度が悪くなる傾向になることが知られている [31]. しかし, 今回は忘却対象が $\mathcal{D}_{mislabeled}$ を含んだ訓練データを用いて学習したモデルであり, 忘却前のモデルの精度が低いため, 混入割合 r が大きく, 忘却対象の数が多い場合においても忘却による精度の向上が見られたと考えられる.

5.3.3 実験 3: ミスラベルデータの検知とパラメータの更新を繋げることで推論精度は向上するか

検知したデータ $\mathcal{D}_{detected}$ を忘却の対象 $\mathcal{D}_{forget} = \mathcal{D}_{detected}$ として学習済みモデルから忘却することで, 推論精度が向上するか検証する. ここでは, ミスラベルデータの検知手法として, 実験 1 で最も良い性能を示した損失に基づく検知手法を用いることで, 忘却対象を $\mathcal{D}_{detected}$ として忘却する. 他の設定は実験 2 と同様である.

結果を表 5.3 に示す. 混入割合 r の増加に伴い, 忘却後の性能が悪化していることがわかる. また, $\mathcal{D}_{mislabeled}$ を忘却した場合に比べ, $\mathcal{D}_{detected}$ を忘却することで推論精度の向上が悪化していることがわかる. このような結果となった原因として, 混入割合 r が増加するにつれて, ミスラベルデータの検知に関する性能が低下し, ミスラベルデータとクリーンデータの入り混じったデータを忘却してしまうため, 推論精度が低下したと考えられる.

以上から, 混入割合 r が大きい場合には, ミスラベルデータの検知性能が低下するため, ミスラベルデータの検知と忘却を繋げた手法の有効性は確認できなかった. 混入割合 r が小さい場合も, 忘却前のモデルの性能が高いため, 忘却による性能向上はあまり見られなかった.

表 5.2: $\mathcal{D}_{mislabeled}$ を忘却したモデルの推論精度

	Method	混入割合 r 0.01	0.05	0.1	0.2	0.3	0.4	0.5
accuracy on $\mathcal{D}_{test}$ (↑)	Pre-trained	99.5	99.6	99.6	95.2	81.4	72.5	45.0
	影響関数 [18]	99.6	99.6	99.6	99.2	92.9	93.5	78.5
	SSSE [19]	97.6	99.5	97.6	98.8	95.7	93.7	93.9
	Fine-tuned	99.9	99.8	99.7	99.9	99.9	99.9	99.9
	Retrained on $\mathcal{D}_{clean}$	99.8	99.8	99.8	99.9	99.9	99.9	99.9

表 5.3: $\mathcal{D}_{detected}$ **を忘却したモデルの推論精度.** 検出手法として Loss を用いた手法を用いた. (比較対象として実験 2 の結果も載せている.)

	Method	混入割合 r 0.01	0.05	0.1	0.2	0.3	0.4	0.5
accuracy on $\mathcal{D}_{test}$ (↑)	Pre-trained	99.5	99.6	99.6	95.2	81.4	72.5	45.0
	影響関数 [18] on $\mathcal{D}_{mislabeled}$	99.6	99.6	99.6	99.2	92.9	93.5	78.5
	影響関数 [18] on $\mathcal{D}_{detected}$	99.6	99.6	99.6	95.7	77.0	76.2	41.2
	SSSE [19] on $\mathcal{D}_{mislabeled}$	97.6	99.5	97.6	98.8	95.7	93.7	93.9
	SSSE [19] on $\mathcal{D}_{detected}$	97.6	99.4	97.6	95.5	74.0	78.6	37.7
	Fine-tuned on $\mathcal{D}_{mislabeled}$	99.9	99.8	99.7	99.9	99.9	99.9	99.9
	Fine-tuned on $\mathcal{D}_{detected}$	99.9	99.7	99.8	97.9	85.8	76.9	40.9
	Retrained on $\mathcal{D}_{mislabeled}$	99.8	99.8	99.8	99.9	99.9	99.9	99.9
	Retrained on $\mathcal{D}_{detected}$	99.8	99.9	99.8	97.3	81.7	76.3	38.9

6 ま と め

本稿では, ミスラベルデータの忘却による学習済みモデルの汎化性能の向上手法を提案した. 提案手法は大まかには, まず学習済みモデルから計算できる異常度を通じてミスラベルデータを検知したのち, 検知したデータをモデルから忘却 (Machine Unlearning) させることで, 汎化性能を向上させる.

MNIST データセットを用いて提案手法を実験的に調査した. 詳細な分析に向けて提案手法の各段階ごとに検討したところ, まずミスラベルデータの混入割合が小さい場合に, 有効に検知できることを確認した. 次にパラメータの更新については, 混入割合によらず, 有効性が確認できた. しかし, 提案手法の構成に相当する実験として検知とパラメータの更新を繋げた場合, 忘却による機械学習モデルの性能向上が確認できなかった. これは混入割合が増加するに従い, ミスラベルデータの検知の性能が低下することに起因している. これらの機能の改善は今後の課題である.

謝　　辞

本研究の一部は JST CREST（課題番号：JPMJCR21M5）の支援を受けたものである. また, 本研究活動をご支援いただきました大阪大学・山口弘純教授に感謝いたします.

文　　献

[1] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, Vol. 25, No. 5, pp. 845–869, 2013.

[2] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, Vol. 5, No. 5, pp. 411–419, 2010.

[3] Viv Cothey. Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 14, pp. 1228–1238, 2004.

[4] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pp. 1299–1316, Baltimore, MD, August 2018. USENIX Association.

[5] Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial label flips attack on support vector machines. In *ECAI 2012*, pp. 870–875. IOS Press, 2012.

[6] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 301–320. Springer, 2016.

[7] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

[8] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Car-

lini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proc. of CCS 2022*, pp. 2779–2792. ACM, 2022.

[9] Yufei Chen, Chao Shen, Yun Shen, Cong Wang, and Yang Zhang. Amplifying membership exposure via data poisoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 29830–29844. Curran Associates, Inc., 2022.

[10] Zhibo Wang, Yuting Huang, Mengkai Song, Libing Wu, Feng Xue, and Kui Ren. Poisoning-assisted property inference attack against federated learning. *IEEE Transactions on Dependable and Secure Computing*, pp. 1–13, 2022.

[11] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

[12] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, Vol. 31, , 2018.

[13] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5552–5560, 2018.

[14] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 17044–17056, 2020.

[15] Qingrui Jia, Xuhong Li, Lei Yu, Jiang Bian, Penghao Zhao, Shupeng Li, Haoyi Xiong, and Dejing Dou. Learning from training dynamics: identifying mislabeled data beyond manually designed features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, pp. 8041–8049, 2023.

[16] Hugging face. `https://huggingface.co/`. Accessed: 2023-12-23.

[17] GitHub. `https.//github.com/`. Accessed: 2023-12-23.

[18] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017.

[19] Alexandra Peste, Dan Alistarh, and Christoph H Lampert. Ssse: Efficiently erasing samples from trained machine learning models. *arXiv preprint arXiv:2107.03860*, 2021.

[20] Abraham Chan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan. The fault in our data stars: Studying mitigation techniques against faulty training data in machine learning applications. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 163–171, 2022.

[21] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 34, No. 11, pp. 8135–8153, 2023.

[22] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society.

[24] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, p. 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

[25] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *Proc. of ICLR 2022*. OpenReview.net, 2022.

[26] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021.

[27] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, Vol. 55, No. 8, pp. 1–35, 2022.

[28] Yinzhi Cao, Alexander Fangxiao Yu, Andrew Aday, Eric Stahl, Jon Merwine, and Junfeng Yang. Efficient repair of polluted machine learning systems via causal unlearning. In *Proc. of AsiaCCS 2018*, pp. 735–747. ACM, 2018.

[29] Junfeng Guo, Ting Wang, and Cong Liu. Poishygiene: Detecting and mitigating poisoning attacks in neural networks. *arXiv preprint arXiv:2003.11110*, 2020.

[30] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

[31] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *CoRR*, Vol. abs/2209.02299, , 2022.

[32] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, Vol. 2, , 2010.

[33] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[34] DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020.

[35] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2023.

教師あり対照学習に基づく次数公平なグラフ深層学習

石崎　泰己† 　佐々木勇和† 　鬼塚　真†

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

あらまし　グラフ深層学習は多様なアプリケーションで活用されているが，モデルの推論結果に対してしばしば公平性の問題が指摘される．公平性にはいくつかの種類があるが，各ノードの次数によって推論結果に不適切な偏りが生じるという次数公平性の問題は未だ研究がほとんど進んでいない．本稿では，メタグラフ学習と教師あり対照学習を用いることで次数公平な推論が可能なグラフ深層学習を提案する．より具体的には，リンク予測モデルとノード生成器を導入し低次数ノードの近傍を補完することで近傍構造の差を低減し，教師あり対照学習を用いて次数による特徴量の差異を抑制する．5 つのデータセットで比較実験を行い，既存手法と比較して次数公平な推論が可能であることを示す．

キーワード　グラフ深層学習, 次数公平性

1　はじめに

グラフは，ノード（頂点）集合とリンク（辺）集合で表現される非線形のデータ構造であり，リンクは 2 つのノードを接続し，そのノード同士に何らかの関係があることを示す．グラフで表現できる実世界のデータは多岐にわたり，交通ネットワーク [1,2] やソーシャルネットワーク [3]，気象状況 [4] など，さまざまなデータに対してグラフの活用事例が増えている．

グラフ深層学習はグラフの表現を学習するために広く用いられている手法である [2,4,5]．各ノードの特徴量をエッジに沿って伝達させて特徴量を計算する構造は多くのグラフ深層学習に採用されている．特徴量とグラフ構造の両方を入力として推論ができるため，グラフに対して適切かつ正確な推論が可能である．

グラフ深層学習はしばしば公平性の問題が指摘される．グラフ深層学習で問題視されている公平性は，各ノードの属性に基づく公平性とグラフの構造に基づく公平性の 2 種類に分けられる．各ノードの属性に基づく公平性にはグループ公平性や個人公平性などがあり，公平な推論が可能であるグラフ深層学習は近年盛んに研究されている [6]．しかし，多くのグラフ深層学習の手法は推論結果の重要な因子であるグラフ構造に基づく公平性に着目していない．グラフ構造はノードの属性と同様に不適切なバイアスを生む原因となり得る．次数公平性は各ノードの次数に基づく公平性である．グラフ深層学習は近傍ノードを用いてノードの特徴量を計算することから，特にノードの次数による推論結果の偏りは頻繁に起こるため，次数に対する公平性は重要な研究課題である．

次数公平性は人種や性別に対する公平性とは異なり，ある分野での新規参入者や商品の人気度に対する公平性にも関連する．例としては，SNS の利用者に職業広告を表示するアプリケーションにおいて，モデルの推論結果が次数によって偏りを持つ場合，フォロワー数によって表示される職業広告に差が生じる．この際，特に新規利用者はそれ以外の利用者に比べてフォロワー数が少ない傾向にあり，次数によって表示される職業広告に差があるのはどちらかにとって不利益な状態である．他にも，不人気，あるいは新規の web ページは次数が低い傾向にあり，web の検索結果に低次数のページが表示されにくい場合は次数による利益の格差が広がっていく問題が起こる．EC サイトや論文の検索エンジンなどについても同様の問題がある．このような問題に対処するため，次数公平性に関するグラフ深層学習が近年いくつか提案されているが [7–9]，次数公平性は他の公平性に比べて比較的新しい概念であり，次数公平な推論をするための手法が少なく改善の余地が多い．

本稿では，ノード分類の精度を維持しつつ次数に関するバイアスを低減するグラフ深層学習のフレームワークを提案する．具体的な手法としては，グラフ構造を学習するモデルを導入し（メタグラフ学習），低次数ノードの近傍が他のノードと同様の構造を持つようにグラフを補完する．低次数ノードの近傍の補完は，リンク予測モデルを用いたエッジを追加とノード生成器を用いた近傍ノードの生成によって実現する．加えて，同一クラス内の低次数ノードと高次数ノードの特徴量の類似度を大きくするために教師あり対照学習を取り入れ，特徴量に含まれる潜在的なバイアスを低減する．

実験は Wikipedia や論文の引用関係のデータを含む 5 つのデータセットで実施する．実験結果としては，既存手法に比べて次数公平なノード分類が可能になり，グラフ深層学習の次数公平な推論を実現できることを示す．また，リンク予測モデルや教師あり対照学習といった手法が次数公平性の改善に有効であることを示す．

本稿の主たる貢献は次の 3 点である．

- 次数に関するバイアスを低減するグラフ深層学習を提案する．
- 評価実験において，複数の実データセットにおいて提案手法が次数公平性を改善することを示す．
- メタグラフ学習や教師あり対照学習などの基礎的な技術

に対する次数公平性と精度への影響の検証する．

2 事前準備

2.1 グラフ

ノードラベルを持つ属性付きグラフは $\mathcal{G} = (A, X, C)$ と表される．グラフの隣接行列 $A \in \mathbb{R}^{n \times n}$ はノード $v_i, v_j \in \mathcal{V}$ の間にリンクが存在する場合に $A_{i,j} = 1$，リンクが存在しない場合に $A_{i,j} = 0$ を満たす行列，属性行列 $X \in \mathbb{R}^{n \times d}$ は各行が 1 つのノードの属性の情報を保持する行列，クラス行列 $C \in \{0,1\}^{n \times |Y|}$ は各ノードのクラス情報を保持する行列である．式の簡単化のため，以降は隣接行列 A の表記にノードの番号は用いず，ノード $u, v \in \mathcal{V}$ の間にリンクが存在する場合 $A_{u,v} = 1$，リンクが存在しない場合には $A_{u,v} = 0$ と表記する．同様にノード v の属性を $x_v \in \mathbb{R}^d$ と表記する．また，ノード v に接続するエッジの数を次数といい $\deg(v)$ で表す．ノード v の隣接ノードの集合を $\mathcal{N}_v$ と表記する．$\mathcal{V}_{low}, \mathcal{V}_{high}$ はそれぞれ低次数ノードと高次数ノードの集合を表す．

2.2 グラフ深層学習

グラフ深層学習はグラフを対象とした深層学習であり，あるノード v に対しその近傍ノード $\mathcal{N}_v$ を集約してノードの特徴量を更新する操作をすることが多い．ノード v に関するこの操作は次式で表される [10]．

$$h_v^{(l+1)} = f\left(h_v^{(l)}, \bigoplus_{u \in \mathcal{N}_v} g(h_v^{(l)}, h_u^{(l)})\right) \tag{1}$$

f, g はニューラルネットワークを含む任意の関数，$h_v^{(l)}$ はノード v の l 層目の特徴量，$\bigoplus$ は何らかの集約操作（max, min, avg 等）を表す．グラフ深層学習は上式で定義される操作を繰り返し，推論を行う．

2.3 次数公平性

本研究では Degree Statistical Parity (DSP) と Degree Equal Opportunity (DEO) の 2 つの性質を満たすグラフ深層学習が次数公平性を理想的に満たすグラフ深層学習と定義する．

2.3.1 Degree Statistical Parity

Degree Statistical Parity (DSP) [7] は，Dwork 氏らが定義した Statistical Parity [11] を次数公平性に拡張したものである．Statistical Parity はある任意のクラスに分類された人口の統計情報が全体の人口の統計情報と同一であるという性質である [11]．これに対し DSP は，ある任意のクラスに分類されたノードの次数分布が全体の次数分布と同一になるという性質のことを指す．この性質は低次数ノード $\mathcal{V}_{low}$ と高次数ノード $\mathcal{V}_{high}$ の比較に限定すれば次式で定式化される．

$$P(\hat{y}_v = y | v \in \mathcal{V}_{low}) = P(\hat{y}_v = y | v \in \mathcal{V}_{high}) \tag{2}$$

式 2 は，低次数であるか高次数であるかに予測ラベル分布が依存しないことを表す．予測ラベル分布が次数に依存しない性質は DSP と等価である．本研究ではグラフ深層学習の予測ラベルが式 2 を満たすことを目的とする．

2.3.2 Degree Equal Opportunity

Degree Equal Opportunity (DEO) [7] は，Hardt 氏らが定義した Equal Opportunity [12] を次数公平性に拡張したものである．Equal Opportunity は，対象にとって有利な分類（企業への採用，ローン審査の合格など）を陽性とした二値分類において，再現率がセンシティブな属性に依存しない性質を指す．これに対して DEO は，クラスごとの再現率が次数グループに依存しない性質を指す．低次数ノード $\mathcal{V}_{low}$ と高次数ノード $\mathcal{V}_{high}$ の比較に限定すれば任意のラベル $y \in Y$ に対して DEO は次式で定義される．

$$P(\hat{y}_v = y | y_v = y, v \in \mathcal{V}_{low}) = P(\hat{y}_v = y | y_v = y, v \in \mathcal{V}_{high}) \tag{3}$$

本研究では式 2 に加えてグラフ深層学習の予測ラベルが式 3 を満たすことを目指す．

2.3.3 問題定義

ノード分類では，ノードを訓練，検証，テストデータに分割する．そして，隣接行列 A と属性行列 X，テストデータを除くノードのクラス情報のみを持つクラス行列 C' が与えられたときに，テストデータに含まれるノードのラベル $y \in Y$ を予測する．ただし Y は正解ラベルの集合である．

3 関連研究

3.1 グラフ深層学習

グラフ深層学習はグラフを入力として推論が可能な深層学習である．グラフ深層学習にはスペクトルに基づく手法 [13,14] と空間的な手法 [15–17] がある．スペクトルに基づく手法はグラフ周波数領域におけるグラフ信号のフィルタリングをする手法であり，空間的な手法が提案される以前から研究されていた．空間的な手法は，スペクトルに基づく手法を拡張して提案された手法であり，近傍ノードの特徴量を集約してノードの特徴量を計算する．Graph Convolutional Network (GCN) [15] は空間的な手法の中で最も基本的な手法の 1 つである．GCN は次数によって正規化された近傍ノードの特徴量を足し合わせ，ノードの特徴量を更新する．これに対して Graph Attention Networks (GAT) [16] は集約操作に注意機構を導入することで重要な近傍ノードの重みを大きくする．また，拡張性の高い空間的な手法の代表例として GraphSAGE [17] があり，GraphSAGE は集約時に近傍ノードをサンプリングすることで計算量を小さくしている．空間的な手法は多くの実世界のデータに対して実用的な精度を発揮しており，様々なアプリケーションで既に広く利用されている．しかし，空間的な手法は近傍ノードを集約する構造を持つため，次数によって推論結果に偏りが生まれやすい．次数公平性に関する研究はまだほとんど研究が進んでおら

ず，未だ多くのグラフ深層学習は次数に関するバイアスを多く含んでいる．本研究ではこの問題に対処するためのグラフ深層学習を提案する．

3.2 次数公平な機械学習

次数公平性とは，次数に対して偏った扱いを受けない状態のことを指す．次数公平性は他の公平性に比べて比較的新しい概念で，近年このバイアスを緩和する研究が行われている．次数公平性に関するバイアスを緩和する手法としては，GRADE [9]，DegFairGNN [7] や RawlsGCN [8] などがある．GRADE [9] は低次数ノードの近傍にリンクを追加，高次数ノードの余分なリンクを削除して対照学習を行う．DegFairGNN [7] は低次数と高次数で一部別々の関数を用意し，低次数と高次数ノードの特徴量の平均が等しくなるように学習する．RawlsGCN [8] はバイアスの根本的な原因を数学的に解析し，低次数ノードが訓練に及ぼす影響度を調整する．

また，低次数ノードに対する精度を向上させることでグラフ深層学習の全体的な精度の向上させることを目的とした研究も近年行われている．代表例としては，Tail-GNN [18] や GRADE [9] などがある．Tail-GNN [18] は低次数ノードに不足している情報を高次数ノードを用いて補完し，ノードの次数グループを分類する識別器を用意して敵対的学習をする．

提案手法では，リンク予測モデルとノード生成器を用いた新しい方法で低次数ノードの近傍構造を補完し，教師あり対照学習を用いて同じクラス内の低次数と高次数のノードの特徴量を近づける．

3.3 対照学習

対照学習とは深層学習モデルに特徴量をより適切に学習させるために提案された手法で，アンカーとなるデータと類似度が高いデータを正例，類似度が低いデータを負例とおき，正例のデータ間の特徴量がより近く，負例のデータ間の特徴量がより遠くなるように損失関数が設計される．対照学習は深層学習モデルの公平性を改善するために広く用いられている [19, 20]．Park 氏らの画像認識モデルにおける公平性に関する研究 [19] では，同じクラスの画像を正例，異なるクラス内のセンシティブな属性が同一の画像を負例として教師あり対照学習を行うことで従来よりセンシティブな属性に対するバイアスを低減している．また， Agarwal 氏らのグラフ深層学習を対象とした公平性に関する研究では，センシティブな属性の値のみを変えた特徴量を正例として対照学習を行いバイアスを低減している．

4 提案手法

4.1 設計方針とフレームワーク

次数に関するバイアスを最も単純に解決する方法は，グラフ内のリンクを用いずに推論をする，あるいはランダムにノードを分類することである．しかし，このような方法ではグラフ構造から得られる情報を活用できず，グラフ深層学習を用いる利点が失われる．同様の理由でグラフ内のノードのリンクをいくらか削除していき，全てのノードを低次数ノードと同程度の次数にする方法は効果的ではない．また，データセットによっては低次数ノードの特徴量と高次数ノードの特徴量がそれぞれ異なる特徴を持つためグラフ構造を使わない場合でも次数のバイアスを取り除くことができない．

提案手法では教師あり対照学習とメタグラフ学習により次数公平なグラフ深層学習にすることを目指す．具体的には，各ノードの特徴量からグラフ構造を学習するモデルと欠落した近傍ノードを予測するモデルを導入する（メタグラフ学習）．これらのモデルを用いて低次数ノードに対して既存のノードとの間に新たなリンクを追加し，加えて低次数ノードの新たな近傍ノードを生成することで低次数ノードの近傍のグラフ構造を補完する．また，教師あり対照学習では同一クラス内の低次数ノードと高次数ノードの特徴量の類似度を大きくする．これにより，低次数ノードと高次数ノードの特徴量の差異が原因の次数のバイアスを低減する．

提案手法の概要図を図 1 に示す．入力は各ノードの特徴量 X と隣接行列 A で，出力は各ノードのクラスごとの確率 h^{out} である．提案手法の概要図（図 1）は大きく 2 つの構造に分けられる．1 つはメタグラフ学習に関する部分で，グラフ内の低次数ノードに対して近傍を補完する操作を行う．もう 1 つはメタグラフ学習を通して得られたグラフをグラフ深層学習モデルに入力し，対照学習を用いて低次数と高次数ノードの特徴量を類似化する部分である．低次数ノードの近傍を補完し，低次数ノードの特徴量のみを更新する関数 $\rho(\cdot)$ にグラフを入力した後，$\mathrm{GNN}_{\mathrm{main}}$ は各ノードのラベルごとの確率 h^{out} を出力する．

4.2 メタグラフ学習

提案手法では，低次数ノードの近傍構造を補完するためにメタグラフ学習を採用する．提案手法におけるメタグラフ学習はリンク予測モデルとノード生成器から構成される．

4.2.1 リンク予測モデル

ノード分類タスクの学習をさせる前に，既存のリンクからリンク予測モデルを作成する．具体的には，まず入力グラフ $\mathcal{G}$ 内のノードの特徴量 X を関数 $\phi(\cdot)$ に入力し補助特徴量 p を得る．

$$p = \phi(X|\theta_\phi) \tag{4}$$

関数 ϕ は深層学習モデル，θ_ϕ は関数 ϕ の学習可能なパラメータを表す．関数 ϕ を適切に学習させることで，既存のグラフ構造に則した新たなリンクを，低次数ノードに接続する形で予測する．本稿では関数 ϕ に MLP を採用した．

次に補助特徴量 p の類似度からノード $u, v \in \mathcal{V}$ を接続するリンクの存在確率 $\mathrm{S}_{u,v}$ を次式で与える．

$$\mathrm{S}_{u,v} = \sigma(\mathrm{sim}(p_u, p_v)) \tag{5}$$

関数 sim はコサイン類似度を計算する関数，$\sigma(\cdot)$ はシグモイド関数である．このようにして計算した $\mathrm{S}_{u,v}$ と隣接行列 A から損失関数 $\mathcal{L}_1$ を次式で計算し，関数 $\phi(\cdot)$ に学習させる．

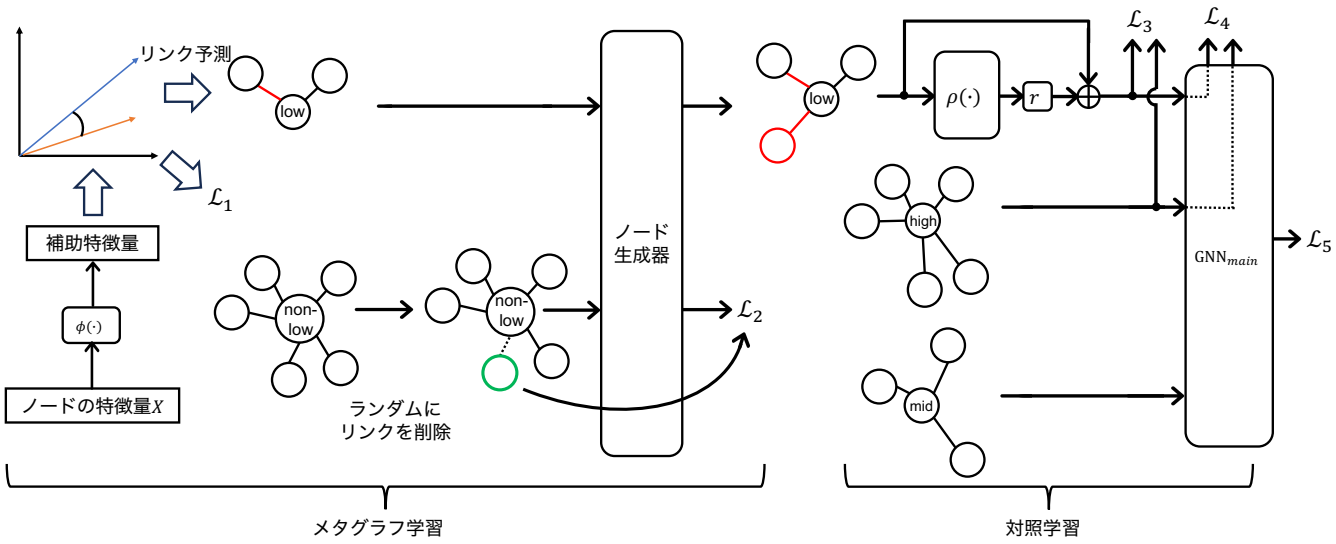


図 1: 提案手法の概要図．低次数ノードを low，高次数ノードを high，それ以外を mid として表している．赤線，赤丸はそれぞれ低次数ノードに対して新しく追加されたリンクと隣接ノードを表す．

$$\mathcal{L}_{\text{pre-train}} = \text{BCELoss}(\text{S}, A) \tag{6}$$

BCELoss は二値クロスエントロピー誤差を求める関数を表す．損失関数 $\mathcal{L}_{\text{pre-train}}$ が最小となるのは，リンクが存在するノード対に対して $\text{sim}(p_u, p_v) = 1$，リンクが存在しないノード対に対して $\text{sim}(p_u, p_v) = 0$ を満たした場合であり，出力のコサイン類似度がグラフ構造を表すように $\phi(\cdot)$ に学習させる．また，実際の学習では計算量の観点から正のペアと負のペアが同数になるようにサンプリングされた一定数のノード対から $\mathcal{L}_{\text{pre-train}}$ を計算する．

次に学習した関数 ϕ から低次数ノードの隣接ノードを予測し，新たにリンクを追加した隣接行列を生成する．

$$\begin{aligned} &A_{u,v} \leftarrow 1, u \in \mathcal{V}_{low}, \\ &\text{S}_{u,v} \in \max_k(\{\text{S}_{u,v} | v \in \mathcal{V}, A_{u,v} = 0\}) \end{aligned} \tag{7}$$

$\max_k(\cdot)$ は集合内の要素の中で最も値が大きい k 個の値の集合を表す．つまり，各低次数ノードに対して，コサイン類似度が最も高い k 個の非近傍ノードとのエッジをグラフに追加する．

4.2.2 ノード生成器

低次数ノードの近傍を補完するために低次数ノードの近傍ノードを生成するノード生成器を導入する．具体的には，入力グラフ $\mathcal{G}$ の低次数ではないノード $u \notin \mathcal{V}_{low}$ の近傍ノードをランダムに 1 つずつマスクし，関数 $\psi(\cdot)$ の出力がマスクされたノードの特徴量を出力するように学習する．以降の式でノード生成器に関する操作を定式化する．

まず，低次数ではないノードの近傍ノードをランダムに 1 つずつマスクした A^{masked} を得る．

$$A^{\text{masked}}_{u,v} = \begin{cases} 0 & \text{if } v = m_u, u \notin \mathcal{V}_{\text{low}} \\ A_{u,v} & \text{else} \end{cases} \tag{8}$$

m_u は $\mathcal{N}_u$ からランダムサンプリングされたノードを表す．

次に A^{masked} をノード生成器に入力し，特徴量 $h^{(1)}_{\nu_u}$ を得る．ν_u はノード生成器によって新たに生成された，h_{ν_u} を特徴量に持つノードを表す．

$$h^{(1)}_u = f_{\text{linear}}(x_u | \theta_f) \tag{9}$$

$$h^{(1)}_{\nu_u} = \psi(h^{(1)}_u, A^{\text{masked}} | \theta_\psi), u \in \mathcal{V} \backslash \mathcal{V}_{low} \tag{10}$$

関数 $f_{\text{linear}} : \mathbb{R}^{n \times d_{in}} \rightarrow \mathbb{R}^{n \times d^{(1)}}$ はパラメータ θ_f を持つ線形変換層，ノード生成器 ψ はパラメータ θ_ψ を持つ深層学習モデルで，欠損した隣接ノードを予測するように学習する．$h^{(1)}_{\nu_u}$ はノード u の特徴量 $h^{(1)}_u$ をノード生成器に入力して得られた出力を表す．

加えて，ノード生成器 ψ の出力 $h^{(1)}_{\nu_u}$ とマスクされた隣接ノード m_u の特徴量 $h^{(1)}_{m_u}$ の類似度が大きくなるように損失関数 $\mathcal{L}_2$ を次式で定義する．

$$\mathcal{L}_2 = -\frac{1}{|\mathcal{V} \backslash \mathcal{V}_{low}|} \sum_{u \in \mathcal{V} \backslash \mathcal{V}_{low}} \text{sim}(h^{(1)}_{m_u}, h^{(1)}_{\nu_u}) \tag{11}$$

この損失関数 $\mathcal{L}_2$ により，ノード生成器 ψ が低次数ではないノード $u \in \mathcal{V} \backslash \mathcal{V}_{low}$ から欠損した近傍ノードの特徴量 $h^{(1)}_{\nu_u}$ を予測可能になる．

次に，ノード生成器 $\psi(\cdot)$ を用いて低次数ノードの近傍ノードを生成する．低次数ノード $u \in \mathcal{V}_{low}$ の特徴量 $h^{(1)}_u$ をノード生成器に入力し得られた出力を $h^{(1)}_{\nu_u}$ とおく．低次数ノードの近傍ノードを生成する際は，マスクされた隣接行列 A^{masked} を用いない．

$$h_{\nu_u}^{(1)} = \psi\left(h_u^{(1)}, A\middle|\theta_\psi\right), u \in \mathcal{V}_{low} \tag{12}$$

これによって新たに生成されたノード ν_u をグラフ $\mathcal{G}$ に追加する．

$$\mathcal{V} \leftarrow \mathcal{V} \cup \{\nu_u\}, u \in \mathcal{V}_{low} \tag{13}$$

$$A_{u,\nu_u} \leftarrow 1 \tag{14}$$

上式の操作を繰り返すことで，低次数ノード 1 つあたり κ 個の生成された近傍ノードを追加する．

4.3 対照学習

提案手法では，同じクラスの低次数ノードと高次数ノードの特徴量を近づけるために教師あり対照学習を採用する．特徴量を近づける必要があるのは，データセットによって低次数ノードと高次数ノードが持つ特徴量の傾向に差があるためである．特徴量の傾向に差がある状態で学習させた場合，一方の特徴量の傾向がもう一方より大きな影響を学習に及ぼし，DEO（式 3）を満たすことが難しくなるという仮説のもと，対照学習を導入する．

具体的には，まず低次数ノードに対してのみ関数 $\rho(\cdot)$ を用いて特徴量を更新する．ここで高次数ノードに対して特有の関数を導入しないのは，グラフ深層学習の全体的な傾向として低次数ノードに対する精度が高次数ノードに比べて低い傾向にあり，高次数ノードの特徴量が低次数ノードの影響を受けて精度が低下することを防ぐためである．

$$h_u^{(2)} = h_u^{(1)} + r \cdot \mathbb{1}(u \in \mathcal{V}_{low}) \cdot \rho(h_u^{(1)}, A|\theta_\rho) \tag{15}$$

θ_ρ は関数 $\rho(\cdot)$ の学習可能なパラメータ，$\mathbb{1}(\cdot)$ は括弧内の式を満たすとき $\mathbb{1}(\cdot) = 1$，そうでないとき $\mathbb{1}(\cdot) = 0$ を満たす関数，$r \in \{0, 1\}$ はハイパーパラメータである．本稿では関数 ρ に GCN [15] を採用している．

ここでは低次数ノードと高次数ノードの特徴量の類似度が大きくなるように教師あり対照学習を行う．損失関数 $\mathcal{L}_3$ と $\mathcal{L}_4$ を次式で計算する．

$$H_{low}^{(l)} = \{h_v^{(l)} | v \in \mathcal{V}_{low}\} \tag{16}$$

$$\mathcal{L}_3 = \text{ContrastiveLoss}(H_{low}^{(2)}, H_{high}^{(2)}, \tau_3) \tag{17}$$

$$\mathcal{L}_4 = \text{ContrastiveLoss}(H_{low}^{(l)}, H_{high}^{(l)}, \tau_4), l > 2 \tag{18}$$

τ_3, τ_4 は温度パラメータ，$l > 2$ はハイパーパラメータである．$\mathcal{L}_3$ と $\mathcal{L}_4$ はそれぞれ 2 層目と l 層目から計算される．ContrastiveLoss$(\cdot)$ 関数は次式で定義する．

$$\begin{aligned}&\text{ContrastiveLoss}(H, Q, \tau)\\ &= -\frac{1}{|Y|}\sum_{y\in Y}\log\frac{\exp\left(\mathbf{s}(H_y, Q_y)/\tau\right)}{\sum_{y'\in Y, y'\neq y}\exp\left(\mathbf{s}(H_y, Q_{y'})/\tau\right)}\end{aligned} \tag{19}$$

$$H_y = \{h_v | h_v \in H, y_v = y\} \tag{20}$$

$$\mathbf{s}(H, Q) = \frac{1}{|H||Q|}\sum_{h\in H}\sum_{q\in Q}\text{sim}(h, q) \tag{21}$$

実際の計算では (21) 式の計算は H と Q からそれぞれランダムサンプリングされたペアで類似度を計算する．

4.4 低次数ファインチューニング

DSP（式 2）を満たすためには低次数ノードと高次数ノードに対する予測ラベル分布が等しい必要がある．しかし，データセットによっては正解ラベル分布は次数に依存する．例えば，web ネットワークでは次数が高いほどアクセス数が多い傾向にあるため，アクセス数で web ページにラベル付けをした場合には正解ラベルが次数に依存する．

正解ラベル分布が次数に依存する場合，グラフ全体で学習を行ったモデルは次数に依存した予測ラベルを出力しやすくなる．提案手法ではこの問題に対処するため，低次数ノードのみから計算した損失関数で低次数ファインチューニングを行い，低次数ではないノードの推論をさせる．この際に用いる損失関数は次式である．

$$\mathcal{L}_{\text{fine-tuning}} = \text{CrossEntropyLoss}(h_{low}^{out}, y_{low}) \tag{22}$$

モデル全体を複製し，上式の損失関数でファインチューニングを行う．ファインチューニングを行う前のモデルは低次数ノードを，ファインチューニング後のモデルはそれ以外のノードのラベルを推論する．

4.5 学習方法

提案手法のモデルの学習方法としては，まず初めに 4.2.1 項で定義したリンク予測モデル $\phi(\cdot)$ を学習させるため，$\mathcal{L}_{\text{pre-train}}$ を誤差逆伝搬させる．次にノード分類タスクを解くために次式で定義する損失関数を誤差逆伝搬させる．

$$\mathcal{L}_{\text{main}} = \text{CrossEntropyLoss}(h^{out}, y) + w_2\mathcal{L}_2 + w_3\mathcal{L}_3 + w_4\mathcal{L}_4 \tag{23}$$

$w_n \in \mathbb{R}$ はそれぞれの項の重みを表す．この後，低次数ファインチューニングを行う．

5 実　　験

本章では，4 章で述べた提案手法の評価とノード生成器や教師あり対照学習などの手法単体での評価をする．

表 1: 実験を行ったデータセットの概要

	ノード数	リンク数	クラス数	特徴量の次元数
CiteSeer	3,312	4,536	6	3,312
Chameleon	2,277	31,371	5	2,325
Squirrel	5,201	198,353	5	2,089
EMNLP	2,600	7,969	2	8
arXiv-year	169,343	1,157,799	5	128

5.1 実験設定

5.1.1 データセット

本研究では CiteSeer [21]，Chameleon [22]，Squirrel [22]，EMNLP [23]，arXiv-year [24] の 5 つのデータセットを対象とした．論文の引用関係のデータセットや英語版 Wikipedia のデータセットが含まれる．EMNLP [23] は Liu 氏らの研究と同様 [7] に引用数をもとに 2 つのクラスにラベル付けをした．

それぞれのデータセットのノード数やリンク数などを表 1 に示す．

5.1.2 実験詳細

データセットは訓練データ，検証データ，テストデータをそれぞれ 6 : 2 : 2 の割合でランダムに分割した．データセットの分割は実験を通して固定である．学習は直近 100 エポックで検証データに対する交差エントロピー誤差が改善されなかった場合に終了し，検証データでの交差エントロピー誤差が最も小さかったモデルを選択する．選択されたモデルに対してテストデータでノード分類を行った結果を実験結果とする．また，実験はモデルの初期化に用いる乱数シードを変更することで 5 回繰り返し，実験結果はその平均値を取ったものである．

5.1.3 ベースライン

実験ではベースラインとして，$\mathrm{GNN}_{\mathrm{main}}$ に用いたモデル（GCN，GAT，GraphSAGE）と DegFairGNN，MLP，ランダムなノード分類を用いる．ランダムなノード分類は理想的に次数公平であるため，DSP（式 2），DEO（式 3）を満たす．MLP には隣接行列を入力せずに各ノードの特徴量のみからノード分類をさせたため，各ノードの特徴量に次数に関するバイアスが含まれていない場合は次数公平な推論結果であると予想される．DegFairGNN は次数公平なグラフ深層学習モデルとして提案されており，埋込み表現に対して情報を補完，または削減し次数に関するバイアスを低減している．

5.1.4 評価指標

実験では次数公平性の指標として $\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ を用いる．また，ノード分類の精度を表す指標として正解率と Macro-F1 を用いる．$\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ [7] は次式で定義される．

$$\Delta_{\mathrm{DSP}} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\Big|P(\hat{y}_v = y|v\in\mathcal{V}_{low}) - P(\hat{y}_v = y|v\in\mathcal{V}_{high})\Big| \tag{24}$$

$$\Delta_{\mathrm{DEO}} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\Big|P(\hat{y}_v = y|y_v = y, v\in\mathcal{V}_{low}) - P(\hat{y}_v = y|y_v = y, v\in\mathcal{V}_{high})\Big| \tag{25}$$

$\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ はモデルの出力が次数に依存しないほど小さい値になる指標であり，次数公平性の観点ではどちらも小さい値が好ましい．Δ_{DSP} は $\mathcal{V}_{low}$ と $\mathcal{V}_{high}$ の間で予測ラベル分布に差がないとき $\Delta_{\mathrm{DSP}} = 0$ を満たす．これに対し，Δ_{DSP} は任意のクラスにおける Recall が $\mathcal{V}_{low}$ と $\mathcal{V}_{high}$ の間で差がないときに $\Delta_{\mathrm{DEO}} = 0$ を満たす．それぞれ，DSP（式 2），DEO（式 3）のときに 0 となる指標である．

5.1.5 ハイパーパラメータ

本稿では，メタグラフ学習や対照学習といった手法を取り入れ次数公平性を改善しているが，データセットや $\mathrm{GNN}_{\mathrm{main}}$ に用いるモデルによって最適な手法の組み合わせが異なる．最適なモデルの探索のため，これらの手法を組み込むか否かをハイパーパラメータとして設定する．ハイパーパラメータの探索範囲を表 3 に示す．

$\kappa = k = w_4 = w_5 = r = lr' = 0$ を満たすとき，モデルは $\mathrm{GNN}_{\mathrm{main}}$ と等しくなり，学習は $\mathcal{L}_{main}$（式 23）からの誤差逆伝搬のみ行われる．ハイパーパラメータチューニングの目的関数は $Acc - (\Delta_{\mathrm{DSP}} + \Delta_{\mathrm{DEO}})/8$ とする．

5.2 実験結果

5.2.1 比較実験

表 2 に GCN，GAT，GraphSAGE をベースにした手法の結果をそれぞれ示す．以降，$\mathrm{GNN}_{\mathrm{main}}$ を GCN にした場合，提案手法（GCN）と表記する（GAT，GraphSAGE についても同様）．

提案手法は既存手法と比較して複数のデータセットにおいて既存手法より次数公平な推論が実現できた．提案手法（GCN），提案手法（GAT）の実験結果に着目すると，GCN と GAT に比べてそれぞれ Δ_{DSP} は 10 パターンの内 8 パターンで改善された．Δ_{DEO} は 10 パターンの内 6 パターンで改善された．ただし，提案手法（GCN）が EMNLP に対して，$\Delta_{\mathrm{DSP}} = \Delta_{\mathrm{DEO}} = 0.0$ であるのは，ハイパーパラメータチューニングの結果，ラベルが 0 としか出力しない状態（$P(\hat{y}_v = 0) = 1.0$）に収束したためである．予測ラベルがある 1 つの値しか取らない場合，次数公平ではあるもののグラフに対する推論としては不適切である．EMNLP はラベル 0 の割合が 79 %であるが，ある 1 つのラベルに偏ったデータセットに対してはハイパーパラメータチューニングの目的関数に正解率を用いて計算するのは不適切であるという知見が得られた．

また，提案手法（SAGE）は GraphSAGE や DegFairSAGE [7] に比べて次数公平性の指標がほとんど改善されなかった．これは GraphSAGE が入力グラフの近傍ノードをサ

表 2: 比較実験．正解率 (Acc.)，Macro-F1 値（Mac-F1），Δ_{DSP}，Δ_{DEO} の比較．データセットは CiteSeer，Squirrel，Chameleon (Chamel.)，arXiv-year (arXiv-y) の 5 つ．太字は GCN，GAT，GraphSAGE を用いたモデルの中でそれぞれ一番高い精度を表す．また，下線は GCN，GAT，GraphSAGE のみとそれぞれ比較して改善している精度を表す．- はメモリエラー．

		Random	MLP	GCN	DegFairGCN	提案手法（GCN）	GAT	DegFairGAT	提案手法（GAT）	GraphSAGE	DegFairSAGE	提案手法（SAGE）
CiteSeer	Acc. ↑	16.4 ± 1.1	61.1 ± 0.9	**76.6 ± 0.1**	76.2 ± 0.2	76.3 ± 0.1	75.9 ± 0.1	72.7 ± 0.4	**76.1 ± 0.1**	**76.7 ± 0.1**	67.6 ± 2.3	76.6 ± 0.2
	Mac-F1 ↑	1.6 ± 0.1	56.8 ± 0.4	**73.4 ± 0.1**	73.1 ± 0.2	73.2 ± 0.2	71.1 ± 0.3	67.6 ± 0.5	**72.0 ± 0.2**	**72.8 ± 0.1**	58.0 ± 2.8	72.5 ± 0.3
	Δ_{DSP} ↓	3.0 ± 0.5	6.4 ± 0.8	8.8 ± 0.0	8.7 ± 0.3	**8.5 ± 0.1**	8.7 ± 0.1	**8.6 ± 0.2**	**8.6 ± 0.1**	8.9 ± 0.1	**6.1 ± 0.7**	8.8 ± 0.2
	Δ_{DEO} ↓	11.2 ± 1.7	16.8 ± 1.2	**10.5 ± 0.2**	12.7 ± 1.6	**10.5 ± 0.8**	**12.8 ± 0.7**	19.9 ± 0.4	13.5 ± 0.4	12.9 ± 0.2	**12.7 ± 1.2**	13.3 ± 0.3
Squirrel	Acc. ↑	20.8 ± 0.7	18.6 ± 0.6	46.3 ± 0.3	**55.7 ± 0.4**	47.0 ± 0.4	48.9 ± 2.4	**53.0 ± 1.1**	52.4 ± 0.8	42.7 ± 0.3	33.9 ± 0.4	**43.4 ± 0.3**
	Mac-F1 ↑	2.1 ± 0.1	8.8 ± 1.0	44.6 ± 0.3	**54.7 ± 0.4**	45.6 ± 0.4	47.6 ± 2.4	**52.2 ± 1.1**	51.7 ± 0.7	42.1 ± 0.3	29.4 ± 0.5	**42.9 ± 0.3**
	Δ_{DSP} ↓	2.3 ± 0.4	1.4 ± 0.7	20.1 ± 0.5	**8.7 ± 0.4**	18.9 ± 0.4	6.2 ± 0.8	**6.1 ± 0.7**	9.1 ± 0.7	**7.7 ± 0.3**	8.6 ± 0.6	8.7 ± 0.6
	Δ_{DEO} ↓	8.6 ± 1.8	4.0 ± 1.7	27.5 ± 0.9	**15.1 ± 1.7**	27.9 ± 0.9	17.5 ± 0.9	16.9 ± 2.3	**13.3 ± 1.9**	25.2 ± 0.9	**14.5 ± 2.5**	25.4 ± 1.0
Chamel.	Acc. ↑	19.3 ± 0.9	42.9 ± 0.9	63.7 ± 0.3	**64.8 ± 0.2**	60.6 ± 0.7	63.9 ± 0.3	**66.7 ± 0.8**	63.9 ± 0.2	**62.8 ± 0.4**	46.5 ± 2.8	61.8 ± 0.8
	Mac-F1 ↑	1.9 ± 0.1	40.6 ± 1.2	63.8 ± 0.3	**64.8 ± 0.3**	60.4 ± 0.8	63.6 ± 0.3	**66.6 ± 0.8**	63.9 ± 0.2	**62.6 ± 0.4**	42.4 ± 4.4	61.9 ± 0.8
	Δ_{DSP} ↓	6.1 ± 1.0	4.5 ± 0.4	7.9 ± 0.8	6.7 ± 0.5	**6.4 ± 1.2**	8.4 ± 0.6	**5.9 ± 0.5**	6.6 ± 0.2	**6.3 ± 0.6**	**6.3 ± 1.0**	10.5 ± 0.6
	Δ_{DEO} ↓	12.5 ± 1.6	12.5 ± 1.2	26.3 ± 1.2	31.3 ± 0.9	**18.1 ± 2.9**	28.4 ± 1.6	**25.9 ± 1.2**	29.3 ± 1.0	26.4 ± 1.4	**23.4 ± 1.9**	30.0 ± 1.5
EMNLP	Acc. ↑	49.1 ± 1.3	78.5 ± 0.6	**79.0 ± 0.0**	78.3 ± 0.5	**79.0 ± 0.0**	**81.2 ± 0.1**	78.9 ± 0.1	79.5 ± 0.5	**82.8 ± 0.2**	80.0 ± 0.3	80.4 ± 0.6
	Mac-F1 ↑	4.5 ± 0.2	45.9 ± 1.7	44.3 ± 0.2	**57.7 ± 1.8**	44.1 ± 0.0	**63.5 ± 1.2**	50.5 ± 2.8	54.8 ± 2.5	**70.8 ± 0.3**	54.5 ± 1.1	57.4 ± 3.9
	Δ_{DSP} ↓	6.9 ± 1.9	0.7 ± 0.7	0.1 ± 0.1	3.0 ± 1.8	**0.0 ± 0.0**	12.2 ± 2.3	4.3 ± 2.7	**3.7 ± 1.2**	29.2 ± 1.6	**4.5 ± 0.4**	12.4 ± 4.3
	Δ_{DEO} ↓	4.6 ± 1.2	0.3 ± 0.3	0.2 ± 0.2	4.9 ± 1.0	**0.0 ± 0.0**	8.5 ± 2.3	3.8 ± 2.4	**3.3 ± 1.1**	25.5 ± 1.5	**3.8 ± 0.3**	10.4 ± 3.6
arXiv-y	Acc. ↑	20.1 ± 0.1	36.2 ± 0.2	**49.0 ± 0.1**	45.4 ± 0.1	40.2 ± 1.3	**49.3 ± 0.1**	-	42.7 ± 0.3	**48.7 ± 0.1**	44.5 ± 0.1	45.0 ± 0.1
	Mac-F1 ↑	2.0 ± 0.0	24.3 ± 0.3	**40.0 ± 0.1**	35.7 ± 0.1	30.6 ± 1.2	**43.0 ± 0.2**	-	29.2 ± 0.6	**40.8 ± 0.4**	31.2 ± 0.1	33.4 ± 0.1
	Δ_{DSP} ↓	0.7 ± 0.1	11.7 ± 0.6	9.7 ± 0.2	15.4 ± 0.1	**8.4 ± 0.9**	**12.1 ± 0.3**	-	17.1 ± 1.8	13.5 ± 0.1	**11.5 ± 0.6**	15.2 ± 0.2
	Δ_{DEO} ↓	1.7 ± 0.3	9.1 ± 0.7	21.3 ± 0.3	16.3 ± 0.2	**14.7 ± 1.2**	20.0 ± 0.2	-	**14.6 ± 0.7**	17.7 ± 0.4	18.2 ± 0.3	**13.9 ± 0.1**

表 3: ハイパーパラメータとその概要．$lr' = 0$ はファインチューニング自体を行わないことを示す．

パラメータ名	概要	値域
lr	学習率 (メタ学習，main 共通)	$[10^{-3}, 3 \times 10^{-2}]$
lr'	学習率 (ファインチューニング)	$\{0, 10^{-4}, 5 \times 10^{-4}\}$
decay	減衰率 (共通)	$[10^{-6}, 10^{-2}]$
p_{dropout}	ドロップアウト率	$[0.0, 0.9]$
N_{layer}	$\mathrm{GNN}_{\mathrm{main}}$ のレイヤー数	$\{2\}$
w_n	損失関数の重み	$[0.0, 0.1]$
d_n	各層の特徴量の次元数	$[2^5, 2^9]$
k	新規リンクの数	$\{0, 1\}$
κ	新規隣接ノードの個数	$\{0, 1\}$
r	関数 ρ の出力の係数	$\{0, 1\}$
τ_n	温度パラメータ	$\{1\}$
l	h^l から $\mathcal{L}_4$ を計算する	$\{3\}$

ンプリングする構造を持っていることが原因と考えられ，メタグラフ学習を用いて低次数ノードの近傍を補完する操作の利点が消失していると考えられる．

ランダムなノード分類の $\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ が 0 より大きな値を示しているのは，テストデータの低次数ノードと高次数ノードが十分に存在しないためであると考えられる．実際に，ノード数が多いデータセットほどランダムの $\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ は小さい．CiteSeer と arXiv-year の実験結果では，MLP の $\Delta_{\mathrm{DSP}}, \Delta_{\mathrm{DEO}}$ がランダムと比べて大きな値を示しており，特定の次数グループの特徴量の傾向が次数に関するバイアスの原因になっている可能性が考えられる．

5.2.2 アブレーションスタディ

アブレーションスタディとしてリンク予測モデルのみやノード生成器のみを用いた場合の精度を表 4 に示す．どの手法についても得意とするデータセットが異なり，次数公平性の指標が GCN に比べて一貫して改善されている手法がないのが分かる．提案手法は各手法に比べて最も次数公平であるか，もしくは正解率が最も高い．ハイパーパラメータにより各手法が推論に及ぼす影響が調整され，それぞれの手法が適切な重みで推論されていることが分かる．

6 おわりに

本稿では，グラフ深層学習にとって重要な課題である次数公平性の問題に取り組んだ．次数に関するバイアスを取り除くために，提案手法では低次数ノードの近傍を補完し，教師あり対照学習を導入することで低次数ノードと高次数ノードの特徴量の不適切な差異を低減した．実験では 5 つのデータセットに対してノード分類を行い，既存手法に比べて次数公平な推論が可能であることが示された．

今後の課題としては，次数公平性は改善されたもののノード分類の精度（正解率，Macro F1 値）が低くなる問題があるため，さらに手法の拡張が必要である．また，ハイパーパラメータの探索範囲が膨大であるため，データセットの傾向と適切なハイパーパラメータの関係を調べるなどをすることで探索範囲を小さくする必要がある．

謝　辞

本研究は JST さきがけ JPMJPR21C5，ASPIRE JPMJAP2328，および JSPS 科学研究費 JP20H00583 の支援よって行われた．

文　献

[1] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Proceedings of the NeurIPS*, 2020.

[2] C. Chen, K. Li, S. G. Teo, X. Zou, K. Wang, J. Wang, and Z. Zeng. Gated residual recurrent graph neural networks for traffic prediction. In *Proceedings of the AAAI*, 2019.

[3] Ting Bai, Youjie Zhang, Bin Wu, and Jian-Yun Nie. Temporal graph neural networks for social recommendation. In

表 4: 各手法のみの精度，$\mathrm{GNN}_{\mathrm{main}}$ は GCN として設定している．低次数 FT は 4.4 節の低次数ファインチューニングを表す．

		GCN	リンク予測モデル	ノード生成器	対照学習 ($\mathcal{L}_3$)	対照学習 ($\mathcal{L}_4$)	低次数 FT	提案手法
CiteSeer	Acc. ↑	**76.6 ± 0.1**	69.0 ± 0.5	75.6 ± 0.3	76.4 ± 0.3	76.3 ± 0.1	76.1 ± 0.3	76.3 ± 0.1
	Mac-F1 ↑	**73.4 ± 0.1**	62.8 ± 0.8	70.0 ± 1.0	71.6 ± 0.9	72.7 ± 0.3	73.0 ± 0.2	73.2 ± 0.2
	Δ_{DSP} ↓	8.8 ± 0.0	10.6 ± 0.4	8.9 ± 0.2	8.6 ± 0.1	**8.3 ± 0.2**	8.9 ± 0.0	8.5 ± 0.1
	Δ_{DEO} ↓	**10.5 ± 0.2**	28.3 ± 1.7	15.5 ± 1.2	13.0 ± 0.5	13.0 ± 1.0	11.0 ± 0.3	**10.5 ± 0.8**
Squirrel	Acc. ↑	46.3 ± 0.3	44.1 ± 0.7	45.1 ± 1.2	43.3 ± 0.5	**47.0 ± 0.4**	46.0 ± 0.2	**47.0 ± 0.4**
	Mac-F1 ↑	44.6 ± 0.3	43.4 ± 0.8	43.5 ± 1.4	41.4 ± 0.7	**45.6 ± 0.4**	44.5 ± 0.3	**45.6 ± 0.4**
	Δ_{DSP} ↓	20.1 ± 0.5	19.6 ± 0.4	19.1 ± 1.3	**17.4 ± 0.7**	19.7 ± 0.4	19.5 ± 0.2	18.9 ± 0.4
	Δ_{DEO} ↓	27.5 ± 0.9	25.2 ± 2.5	**23.2 ± 2.7**	25.9 ± 0.6	26.2 ± 0.4	26.4 ± 0.3	27.9 ± 0.9
Chamel.	Acc. ↑	63.7 ± 0.3	62.5 ± 0.3	63.6 ± 0.5	**64.4 ± 0.4**	63.3 ± 0.2	63.8 ± 0.1	60.6 ± 0.7
	Mac-F1 ↑	63.8 ± 0.3	62.4 ± 0.4	63.4 ± 0.6	**64.3 ± 0.4**	63.3 ± 0.2	63.7 ± 0.1	60.4 ± 0.8
	Δ_{DSP} ↓	7.9 ± 0.8	7.2 ± 0.7	8.9 ± 0.8	7.6 ± 0.5	8.5 ± 0.5	10.3 ± 0.3	**6.4 ± 1.2**
	Δ_{DEO} ↓	26.3 ± 1.2	21.4 ± 1.8	23.6 ± 1.4	25.2 ± 1.3	25.7 ± 0.7	29.8 ± 0.7	**18.1 ± 2.9**
EMNLP	Acc. ↑	79.0 ± 0.0	86.4 ± 0.5	87.0 ± 0.2	**87.4 ± 0.3**	79.7 ± 0.3	79.0 ± 0.0	79.0 ± 0.0
	Mac-F1 ↑	44.3 ± 0.2	75.7 ± 1.1	**80.0 ± 0.3**	79.5 ± 0.7	47.7 ± 1.7	44.1 ± 0.0	44.1 ± 0.0
	Δ_{DSP} ↓	0.1 ± 0.1	55.0 ± 1.9	62.0 ± 1.6	56.3 ± 2.1	4.0 ± 2.0	**0.0 ± 0.0**	**0.0 ± 0.0**
	Δ_{DEO} ↓	0.2 ± 0.2	48.7 ± 1.7	55.8 ± 1.8	49.8 ± 2.0	3.4 ± 1.8	**0.0 ± 0.0**	**0.0 ± 0.0**

Proceedings of the BigData, 2020.

[4] Ma Minbo, Xie Peng, Teng Fei, Wang Bin, Ji Shenggong, Zhang Junbo, and Li Tianrui. Hisgnn: Hierarchical spatio-temporal graph neural networks for weather forecasting. *Information Sciences*, Vol. 648, , 2023.

[5] Guipeng Xv, Chen Lin, Wanxian Guan, Jinping Gou, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. E-commerce search via content collaborative graph neural network. In *Proceedings of the SIGKDD*, 2023.

[6] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[7] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. On generalized degree fairness in graph neural networks. In *Proceedings of the AAAI*, 2023.

[8] Jian Kang, Yan Zhu, Yinglong Xia, Jiebo Luo, and Hanghang Tong. Rawlsgcn: Towards rawlsian difference principle on graph convolutional network. In *Proceedings of the WWW*, 2022.

[9] Ruijia Wang, Xiao Wang, Chuan Shi, and Le Song. Uncovering the structural fairness in graph contrastive learning. In *Proceedings of the NeurIPS*, 2022.

[10] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. In *arXiv:2104.13478 [cs.LG]*, 2021.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *in Proceedings of the ITCS*, 2012.

[12] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the NeurIPS*, 2016.

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the NeurIPS*, 2016.

[14] David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, Vol. 30, No. 2, pp. 129–150, 2011.

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the ICLR*, 2017.

[16] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the ICLR*, 2018.

[17] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the NeurIPS*, 2017.

[18] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. Tail-gnn: Tail-node graph neural networks. In *Proceedings of the KDD*, Virtual Event, Singapore, 2021.

[19] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the CVPR*, 2022.

[20] Agarwal Chirag, Lakkaraju Himabindu, and Zitnik Marinka. Towards a unified framework for fair and stable graph representation learning. In *Proceedings of the UAI*, 2021.

[21] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the ACM*, 1998.

[22] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *arXiv*, 2019.

[23] J. Ma, B. Chang, X. Zhang, and Q. Mei. Copulagnn: Towards integrating representational and correlational roles of graphs in graph neural networks. In *Proceedings of the ICLR*, 2021.

[24] Derek Lim, Felix Hohne, Xi Li, Shengjie L. Huang, Vishnu Gupta, Om Bharelao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Proceedings of th NeurIPS*, 2021.

機械学習のためのアクティブ・データプール拡張手法

京谷　辰哉†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学 情報学群 情報メディア創成学類　〒 305–0821 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–0821 茨城県つくば市春日 1-2
E-mail: †kyoya.tatsuya.as@alumni.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp

あらまし　機械学習において少ないラベル付けでよりよい学習結果を得るために，学習の順番を制御する能動学習が効果的なアプローチであることが知られている．しかし，能動学習ではデータプールの中からデータを選択するため，データの絶対数が足りない場合や，データの不均衡などといった要因に直接的に対応するものではない．このような状況に対応する既存の手法としては，例えば外れ値検知などにおいて，モデルを仮定し外れ値の度合いを予測するということが行われている．本論文では，より直接的な手法として，アクティブにデータプールを拡張する手法について検討する．具体的には, 現在のデータプールの中では少数のデータしか無いラベルに対して，その少数派ラベルのデータが存在する可能性が高い箇所の付近にある既存データを選択・提示し, その提示データに似ているが少数派ラベルを持つ新たなデータを人間に明示的に加えてもらうことでデータプールを拡張していく．その際, 提示するデータの選び方を複数考案し、シミュレーションを通してその効果を検証する．

キーワード　機械学習, data-imbalance, データプール拡張

1　はじめに

機械学習におけるデータを取り巻く状況として, 学習に使用できるデータの絶対数が少ない, あるいはデータが不均衡であることが挙げられる．このようなデータで十分な学習をすることは難しく, 結果として分類精度が悪くなるもしくは特定のクラスの誤分類が多くなるといった問題が引き起こされる．データ不足やデータ不均衡の問題は実世界の様々な領域・分野で生じており, 考え得るリスクも大きいことから広く研究されてきている．

少ないラベル付けでより良い学習結果を得る手法として, 学習の順番を制御する能動学習がある [1]．能動学習の代表的な例として, ラベルの付いていないデータプールの中から学習効果の高いデータを選択し, 人間にラベルを付けてもらう pool-based active learning がある．後述する本研究は pool-based active learning のコンセプトを参考にしている．

データが不均衡となる問題は様々なドメインで生じており, 金融詐欺の検知や医療診断などの異常検知はその代表的な例である．異常検知の場面では, 異常データは正常データに比べ発生の頻度が少ない性質から正常データと同等の量を収集することが難しく, 必然的に異常クラスと正常クラスのデータ数に大きな偏りが生まれてしまう．このまま学習を進めると, ほぼすべての未知データを正常だと判断するようなバイアスが強くかかり, 異常データと正常データの分類の際に異常データの誤分類が増え異常を見逃すリスクにつながる．この説明は異常検知の例であるが, これは異常検知特有の問題ではなく, データ不均衡が現れる多くの状況で起こり得ることである．これを回避するためデータ不均衡の研究では, 不均衡な状態で未知のデータに対しいかに正しく少数派のクラス (異常検知では異常クラス) に分類できるよう学習させるかということに焦点を当てた研究がなされてきた．

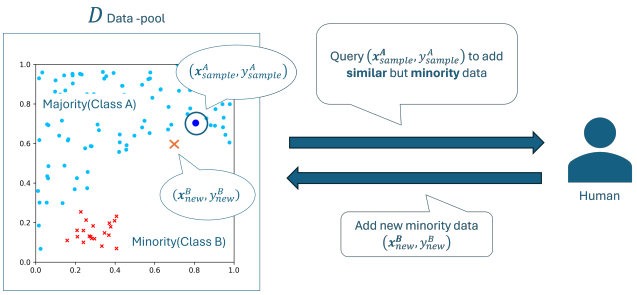


図 1: 本研究で取り組むアクティブ・データプール拡張の概要: システム (左) は，現在あるラベル付きデータ集合の中から特定のデータ群を選択し，そのデータに似ているが異なるラベルを持つデータを人間 (右) に問い合わせ，入手する．本論文ではその選択戦略を比較する．

一方, 学習データの数を増やす手法として, Data Augmentation が知られている．これは Deep Learning において過学習のリスクを避けつつ Deep Learning モデルの汎化性能をより高めることのできる手法の 1 つで, 特に画像やテキストデータを対象にして, 既存のデータにわずかな修正を加えたり新たなデータを合成することで学習データの数を増やしていくものであり, 様々な拡張の方法が幅広く研究されている．

データプールの不足及び不均衡に対する既存研究の多くは, 与えられたデータプールの下で学習効果を高めることに焦点を当てたものであり, いずれも新たなデータを追加することでデータプールを拡張するという, データ不足・不均衡問題の根本的な解決を目指すものではなかった．また, Data Augmentaiton のような既存のデータ拡張手法では, クラス間のデータ数に大きな偏りがあるデータ不均衡のような状態では, データ数が少

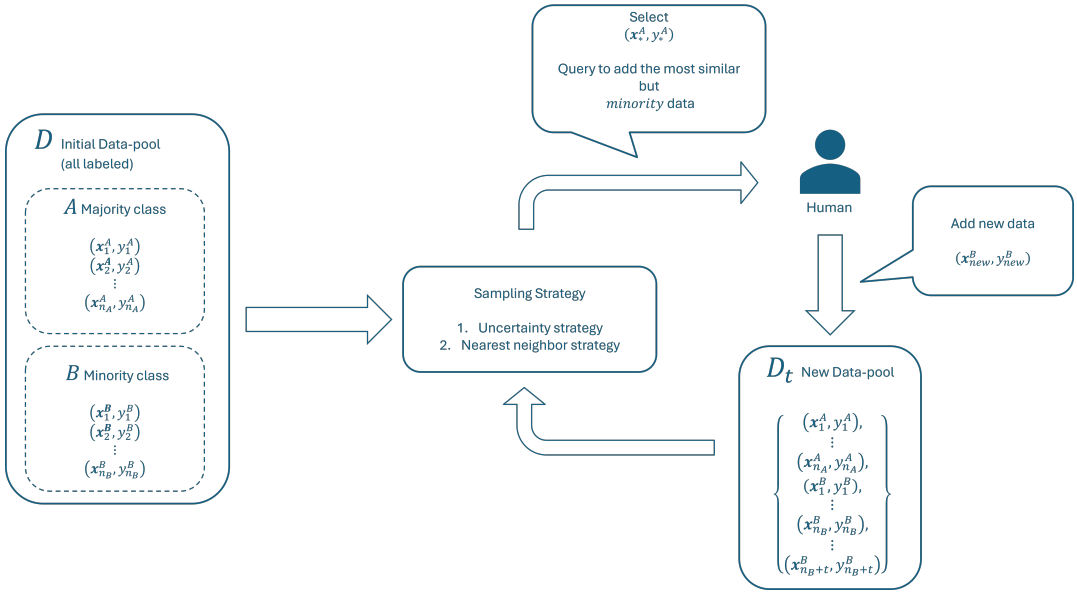


図 2: 提案手法全体の概略図

ないクラスのデータを拡張したところでそのクラスの過学習リスクを避けることが難しいという問題が指摘されており，こちらもデータ不均衡のような問題を解決する手法としては不十分である．

それに対し本研究は，図 1 に示すような，既存のデータプールからデータの数が多いラベル（多数派クラス）を持つデータを提示し，それに類似しているがデータ数の少ないラベル（少数派クラス）を持つデータを人間から明示的に追加してもらい，より学習効果の高いデータプールへと拡張する際に使用するサンプリング戦略にはどのようなものがあるかを調査する．そして，多数派クラスのデータの中から上述のサンプリング戦略に則り，現時点で付近に少数派クラスのデータが存在しない・非常に少ないようなデータを選択・提示し，それと類似した少数派クラスのデータを追加していくことで，より少ないデータの追加で学習効果の高いデータプールになるような拡張を実現していくアクティブ・データプール拡張を提案する．また，提示するデータのサンプリングには様々な戦略が考えられるため，本研究はそのような戦略として pool-based active learning で広く用いられている基準である uncertainty を使用したサンプリング戦略と多数派クラスのデータ 1 件とその最近傍にある少数派クラスのデータ 2 件の 3 つ組から提示するデータを決定するサンプリング戦略を考案した．図 2 は，それらのサンプリング戦略を含めた提案手法の全体像を示したものである．

本論文のリサーチクエスチョンは次の通りである．

RQ: 機械学習のためのアクティブ・データプール拡張の戦略としてはどのような戦略が効果的か

このリサーチクエスチョンに対して本研究では上述した複数の戦略を考案し実験による検証を行い，以下のような結果が得られた．

結果 1
uncertainty を使用した戦略は優位な結果とはならなかったが，より良い結果が期待できるような改善点が見つかった．

結果 2
多数派クラスのデータ 1 件と少数派クラスのデータ 2 件の 3 つ組を使用するサンプリング戦略は早い段階で分類精度が高くなるようなデータプールに拡張ができている．

本論文の構成は次の通りである．2 節では関連研究について説明する．3 節では本研究の問題定義について説明する．4 節では考案したサンプリング戦略について，5 節では実験及び結果の説明を行う．その後，6 節で本研究に関する議論を行い，7 節で本研究についてまとめる．

2 関 連 研 究

本節では，本研究と関連の深い研究を紹介し，その手法の問題点や本研究との比較を行う．なお，提案手法とその関連研究の比較を表 1 に示す．

2.1 Pool-based Active Learning

pool-based active learning [1] [2] は学習を制御することで少ないラベル付けでより良い学習結果が得られる手法として知られており，テキストの分類 [3]，文書等からの情報抽出 [4]，癌の分類・診断 [5] など様々な実世界のドメインで幅広く研究されている [1]．

pool-based active learning では，まず少数のラベル付きのデータ D_l とラベルの付いていないデータ D_u から始まり，D_l を使用して分類器 f_0 に対して最初の学習を行う．その後，あるステップ t について，ステップ $t-1$ で訓練した分類器 f_{t-1} を用いてラベルの付いていないデータプール D_u の中からある基準に則りラベル付けをするデータ $\boldsymbol{x}_*$ をサンプリングする．サンプリングされたデータ $\boldsymbol{x}_*$ は人間に提示され，人間は $\boldsymbol{x}_*$ に対

応するラベル y_* をつける．ラベル付けされたデータ $(\boldsymbol{x}_*, y_*)$ をラベル付きデータプール D_l に加え, D_l を使用して分類器 f_t を再訓練する．

pool-based active learning に関する研究の多くは, 各ステップで D_u からサンプリングするデータを選ぶ際に使用する合理的な基準を提案するものであった [2]．中でも uncertainty sampling は最も有名な基準の 1 つで, これは分類器 f_t の決定境界付近にあるデータをサンプリングするものである．

本研究は上述の pool-based active learning のコンセプトを参考にした手法となっており, pool-based active learning が D_u からサンプリングし人間がラベル付けをするという流れであるのに対し, 本研究は D_l から多数派クラスのデータをサンプリングし人間がそれと類似した新たな少数派クラスのデータを追加する流れとなっている．

2.2 データ不均衡 (Data Imbalance)

データ不均衡とは, データプール中の多数派のクラスと少数派のクラスのデータ数に大きな偏りがある問題である．この問題は様々な領域・分野で生じており, 代表的な例として次のようなものがある [6] [7].

- クレジットカードの不正利用検知
- ネットワークの侵入検知
- 医療診断

例えば医療診断では, 患者のデータが健常者のデータに比べ圧倒的に少なく, それにより患者の誤分類, すなわち患者を健常者だと判断してしまう可能性が高くなる．これは健常者を患者であると判断するよりもリスクが大きいと考えられる．

このようにデータ不均衡に起因する未知データの誤分類は異常の見逃しや誤分類の結果に基づき意思決定をしてしまうといった大きなリスクにつながる可能性が非常に高いため, 様々なアプローチが研究されてきた．

データ不均衡に関する広範な研究の 1 つに, over-sampling がある [8]．over-sampilng は少数派クラスのデータを複製することで, 多数派クラスとのデータ数の偏りを緩和する手法である．単純な over-sampling では少数派クラスのデータを複製していくに従い過学習を起こしやすくなるという問題があり, Chawla ら [9] は少数派クラスのデータの k 最近傍のデータから新たな少数派クラスのデータを合成することで over-sampling の問題点を改良した Synthetic Minority Over-sampling Technique (SMOTE) を提案した．SMOTE は, 様々な問題に適用できることや手順の設計がシンプルであることから不均衡なデータから学習するフレームワークの事実上のスタンダードな手法とされている [10] が, 1. ノイズの多いデータではノイズを増やす可能性があること, 2. 大元のデータセットの分布の影響を受けるため分布の端のデータが生成されやすく多数派クラスと少数派クラス間の境界がぼやけることがあること, 3.2 つのサンプル間を補間するようデータを生成するためデータの生成領域が狭く過学習につながる可能性が依然としてあることが指摘されている [11]．この問題に対処するため, Zheng ら [12] の SNOCC や Cohen ら [13] の AHC, Liang ら [11] の LR-SMOTE など, SMOTE 自体の改良も数多く研究されている．

over-sampling 及び SMOTE に関する研究は先述のような問題を可能な限り抑えるように少数派クラスのデータを生成することに焦点を当てたものであるのに対し, 提案手法では明示的に少数派クラスのデータを新たに追加することでデータプールを拡張するという点で拡張の方向性が異なっており, データの生成により引き起こされる問題を考慮する必要がない．

2.3 Data Augmentation

Data Augmentation は, Deep Learning における学習データを増やす手法の 1 つであり, 既存のデータを複製しわずかな修正を加える, あるいは既存のデータから新たなデータを合成することでデータの拡張を行う [14].

この手法は, 画像データやテキストデータの数を増やす際によく用いられている．画像データの拡張手法では, 回転や反転等の幾何的な変形・色調の変更・画像の一部を消去する random erasing・敵対的生成ネットワーク (GAN) を使用したものなどが挙げられ [15], テキストデータでは, 文章の一部を別の単語で置き換える・文章に余計な単語を挿入したり一部の単語を消去する・意味が類似している違う文章を用意するなどの拡張手法がある [14].

また, 近年の大規模言語モデル (LLM) や diffusion モデルの著しい発展を受け, そのような大規模モデルを用いてさらに高度なデータ拡張を行う研究もその数を増している [16].

Data Augmentation は学習データを増やすことで Deep Learning モデルの過学習を防ぐことができるが, 学習データのクラス間に偏りがあるデータ不均衡の状態では, Data Augmentation が上手く機能せず, データ数が少ない少数派クラスのデータに対して過学習を起こすリスクを解消することが難しいと指摘されている [16].

これに対し本研究では, 人間から少数派クラスのデータを新たに追加してもらうことで既存のデータ拡張手法では解決が難しいとされるデータ不均衡の問題に直接対処することができる手法となっている．

3 問題定義

本節では, 提案手法であるアクティブ・データプール拡張に関する記号定義を行い, 手法の概要及び評価指標について説明する．

また, 本節以降では多数派のクラスを *majority* クラス/クラス A, 少数派のクラスを *minority* クラス/クラス B として表記する．

表 1: 関連研究と提案手法の比較

関連研究	データプールの拡張方法	サンプリングの対象	Human-in-the-loop
Pool-based Active Learning	×	ラベルの無いデータプール	✓
Over-sampling(SMOTE)	少数派クラスのデータの複製/合成	少数派クラスのデータ	×
Data Augmentation	学習データの修正・合成	訓練データ	×
提案手法: アクティブデータプール拡張	少数派クラスのデータの追加	多数派クラスのデータ	✓

3.1 記号定義

本論文で使用する記号を以下の表 2 に示す.

表 2: 記号の定義

記号・用語	説明
D	ラベル付きデータプール
n	D 中のデータ個数
$A(\subset D)$	ラベル付きの *majority* クラス/クラス A のデータプール
$B(\subset D)$	ラベル付きの *minority* クラス/クラス B のデータプール
n_A	A 中のデータの個数
n_B	B 中のデータの個数
$(\boldsymbol{x}_i, y_i) \in D$	$\boldsymbol{x}_i$：データ, y_i：ラベル
T	query budget
$t = 1, 2, \cdots, T$	イテレーション
f	分類器

なお, D と n に対して, $D = A + B$, $n = n_A + n_B$ が成り立つ.

2 節までは pool-based active learning との混乱を避けるため, ラベル付きデータプールを D_l と表記していたが, 本節以降はラベル付きデータプールを D とし, イテレーション $t = 1, 2, \cdots, T$ におけるラベル付きデータプールを $D_1, D_2, \cdots, D_T$ と表記する.

3.2 アクティブ・データプール拡張

本論文で提案するアクティブ・データプール拡張は, データ不足・不均衡問題を改善する新たなアプローチである. このアプローチでは, 初期のラベル付きのデータプール $D_0 = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$ からサンプリング戦略（詳細は 4 節）に基づき *majority* クラスのデータ $(\boldsymbol{x}_i^A, y_i^A)$ を提示する. その後, 提示したデータと類似した *minority* クラスのデータ $(\boldsymbol{x}_{new}^B, y_{new}^B)$ を人間から追加することでデータプールを拡張する. サンプリング戦略には様々なものが考えられるため, 本論文では, 複数のサンプリング戦略を考案しその検証を行っている.

3.3 評価指標

提案手法の有効性を検証するために, 本研究では評価指標として以下のものを使用した. 詳細は 5 節で述べる.

- テストデータに対する AUC
- 各戦略における提示したデータと理想的な追加データ間の距離の累計

AUC は, イテレーション $t = 1, 2, \cdots, T$ の各ステップにおいて, 提案手法により拡張されたデータプール D_t で訓練した分類器 f_t の汎化性能を測るために用いている. 追加したデータの数が少ない段階で AUC が大きく向上していれば, その戦略に基づくデータプールの拡張が上手くできていることが分かる.

提案手法は提示した *majority* データと類似している *minority* データを人間に追加してもらうことで拡張を行うものであるが, 境界から離れた *majority* データ, すなわち *minority* データと特徴が大きく異なる *majority* データを提示した場合, 類似する *minority* データがどのようなものなのか判断が難しくなる可能性が高い. 逆に, *minority* データと類似している特徴の *majority* データを提示できれば, より正確なデータを追加することができると考えられる. そこで境界線付近の *majority* データを優先的に提示できているか評価するため, AUC に加え, 今回のシミュレーションにあたり生成したデータの内, 提示した *majority* データとそれに最も近い *minority* データ (理想的な追加データ) 間の距離の累計を評価指標とする. ここで言う距離の累計とは, 例えば 10 個追加時点での値は, 1 個目から 10 個目の提示した *majority* データとそれに最も近い *minority* データ間の距離を合計した値のことである.

4 サンプリング戦略

提案手法では, ラベル付きの *majority* データプール A_t からサンプリング戦略に従い *majority* データをサンプリングし, それを人間に提示することでデータプールの拡張を行う. そこで, 本節では考案した複数のサンプリング戦略について説明し, 各戦略を取り入れた提案手法の全体像を提示する.

4.1 Strategy 1. Uncertainty Strategy

これは, *majority* データプール A_t から提示するデータを選ぶ際, より学習効果の高い *majority* データを選ぶため, pool-based active learning で広く用いられている指標である uncertainty を適用した戦略である.

この戦略では, A_{t-1} の中から分類器 f_{t-1} を使用し uncertainty を算出し, 最も uncertanty の高いデータをサンプリングする. その後, サンプリングされた *majority* データと類似した *minority* データを追加していくことでデータプールを拡張する. ここで, A_{t-1} 中のデータの uncertainty を算出した際, それが最も高いデータが複数存在する場合があり, その中から提示するデータを選ぶ方法は複数考えられる. そこで, 本戦略では uncertainty の最も高い複数のデータの中から提示するデータを 1 つ選ぶ方法として, 表 3 に示すバリエーションを考案した.

表 3: uncertainty strategy のバリエーション

バリエーション	説明
uncertainty-random	ランダムに提示するデータを選択
uncertainty-var	データの各次元の分散に基づきデータを選択
uncertainty-nearest	D_t 中の *minority* データと最も近いものを選択

これより, 各バリエーションについて説明するにあたり, uncertainty の最も高いデータ集合を $C_t = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_c, y_c)\} (\subset A_t)$ と表記することとする.

4.1.1 Uncertainty-random

uncertainty random では, C_t からランダムにデータを選び人間に提示するバリエーションであり, これは後述の 2 つのバリエーションに対するベースラインとして検証する.

4.1.2 Uncertainty-var

この方法ではまず, C_t の各データ $\boldsymbol{x}_i^C$ $(i = 1, 2, \cdots, c)$ と *minority* データプール B_t の全データ $\{\boldsymbol{x}_1^B, \boldsymbol{x}_2^B, \cdots, \boldsymbol{x}_{n_B}^B\}$ で $group_i = \{\boldsymbol{x}_i^C, \boldsymbol{x}_1^B, \boldsymbol{x}_2^B, \cdots, \boldsymbol{x}_{n_B}^B\}$ を構築する. すると, データ $\boldsymbol{x} \in \mathbb{R}^d$ の次元ごとに, $group_i$ ($i = 1, 2, \cdots, c$) の分散が式 (1) により求まる.

$$var_j = \frac{\sum_{k=1}^{n_B+1} \left(x_k^j - \bar{x_j}\right)^2}{n_B + 1} \ (\ j = 1, 2, \cdots, d\ )$$
$$var_{group_i} = (\ var_1, var_2, \cdots, var_d\ ) \quad (1)$$

その後, 式 (1) で算出した $group_i$ の各次元における分散 var_{group_i} を d 次元空間の座標として扱い, 原点 $\mathbf{0}^d$ からの距離を求め, その距離が最小となる $group_i$ に属する C_t のデータ $\boldsymbol{x}_i^C$ をサンプルとして提示する.

4.1.3 Uncertainty-nearest

この方法では, C_t の各データについて最近傍の *minority* データ ($\in D_t$) との距離を算出し, 最近傍の *minority* データまでの距離が最も短い *majority* データを提示するデータとして選択する.

4.2 Strategy 2. Nearest Neighbor Strategy

この戦略では, A_t のデータ $\boldsymbol{x}_i^A (i = 1, 2, \cdots, n_A)$1 件に対し B_t 中の第 1 最近傍・第 2 最近傍のデータ $\boldsymbol{x}_{n_1}^B, \boldsymbol{x}_{n_2}^B$ で 3 つ組を構築し, その 3 つ組からサンプリングする *majority* データを決定する. 提示するデータのサンプリングは, この 3 つ組のデータから式 (2) に基づいて行う.

$$\boldsymbol{x}_{sample} = \underset{\boldsymbol{x} \in A_t}{\arg\min}\ normalize\left(\frac{\|\boldsymbol{x}_i^A - \boldsymbol{x}_{n_1}^B\| + \|\boldsymbol{x}_i^A - \boldsymbol{x}_{n_2}^B\|}{2}\right) + normalize\left(\frac{1}{\|\boldsymbol{x}_{n_1}^B - \boldsymbol{x}_{n_2}^B\|}\right) \quad (2)$$

式 (2) の第 1 項は *majority* データとその第 1・2 最近傍の *minority* データそれぞれとの距離の平均であり, これが小さいほど *minority* データと近い *majority* データが優先的にサンプリングされる. 第 2 項は第 1・2 最近傍の *minority* データ同士の距離の逆数となっており, 距離が大きいほど第 2 項の値は小さくなり, *minority* データの隙間を埋めるような *majority* データのサンプリングが優先される. また, 第 1 項と第 2 項の影響を均一にするため, それぞれ正規化を行っている.

Algorithm 1 active data-pool augmentation with uncertainty strategy

Input: $D_0 (= A_0 + B_0)$: *initial labeled data pool*, T : *query budget*

Initialize:
  set $t = 0$
  training initial classifier f_0 on D_0
while $t < T$ **do**
  for all $\boldsymbol{x}_i$ in A_t, calculate uncertainty using classifier f_t
  select most uncertain data $(\boldsymbol{x}_*^A, y_*^A)$
  query $(\boldsymbol{x}_*^A, y_*^A)$ to human then add new *minority* data $(\boldsymbol{x}_{new}^B, y_{new}^B)$
  $D_{t+1} \leftarrow D_t + (\boldsymbol{x}_{new}^B, y_{new}^B)$
  training new classifier f_{t+1} on D_{t+1}
  $t \leftarrow t + 1$
end while

なおこの戦略は B_t の隙間を埋めるような *majority* データを優先的に提示するため, B_t の分布に強く影響を受ける. つまり B_t が狭い領域に集中している場合その領域に近い *majority* データがサンプリングされやすくなるという問題があり, 少ないステップで境界線付近のデータを充実させるには最初の数ステップで離れた *majority* データをサンプリングする必要がある. そこで, 最初の 2 ステップのみ次の方法で提示するデータをサンプリングする. 最初のイテレーション $t = 1$ では, B_0 のデータ $\boldsymbol{x}_1^B$(始点) とこれの最近傍となる *majority* データ $\boldsymbol{x}_{nearest}^A$ (終点) でベクトル $\boldsymbol{v}_1 (= \boldsymbol{x}_{nearest}^A - \boldsymbol{x}_1^B)$ を張る. また, $\boldsymbol{x}_{nearest}^A$ 以外の *majority* データ (終点) に対しても $\boldsymbol{x}_1^B$(始点) とのベクトル $\boldsymbol{v}_i (= \boldsymbol{x}_i^A - \boldsymbol{x}_1^B)$ $(i \neq nearest)$ を張り, $\boldsymbol{v}_1$ と $\boldsymbol{v}_i$ のコサイン類似度を算出し, それが最小となるデータ $\boldsymbol{x}_i^A$ を提示する. 続くイテレーション $t = 2$ では, ベクトル $\boldsymbol{v}_1$ を $\boldsymbol{x}_1^B$(始点) と $\boldsymbol{x}_2^B$(終点) で張り ($\boldsymbol{v}_1 = \boldsymbol{x}_2^B - \boldsymbol{x}_1^B$), さらに $\boldsymbol{x}_1^B$(始点) と $\boldsymbol{x}_i^A (i = 1, 2, \cdots, n_A)$(終点) で張ったベクトル $\boldsymbol{v}_i (= \boldsymbol{x}_i^A - \boldsymbol{x}_1^B)$ 間のコサイン類似度が最小となる $\boldsymbol{x}_i^A$ を提示する. イテレーション $t = 3$ 以降は先述の方法 (式 (2)) でサンプリングを行う.

式 2 の第 1 項及び第 2 項の振る舞いを見るため, この戦略では次の 3 つに対して検証を行う.

- nearest-AtoB$\cdots$ 第 1 項 (各 *majority* データとその最近傍の *minority* データ 2 件それぞれまでの距離の平均) のみ
- nearest-BtoB$\cdots$ 第 2 項 (各 *majority* データの最近傍の *minority* データ同士の距離) のみ
- nearest-hybrid$\cdots$ 式 2(第 1 項 + 第 2 項)

5 実　　験

本節では, 本研究における実験の設定や実験結果について説

Algorithm 2 active data-pool augmentation with nearest neighbor strategy

Input: $D_0(=A_0+B_0)$: *initial labeled data pool*, T : *query budget*

Initialize:
  set $t=0$
  training initial classifier f_0 on D_0

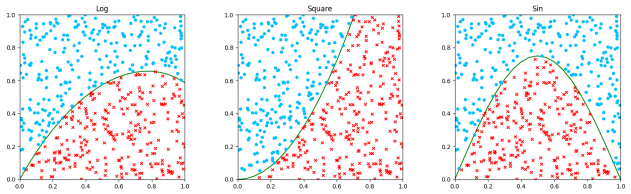
query $(\boldsymbol{x}_*^A, y_*^A) = FirstSample(D_0)$ to human then add new *minority* data $(\boldsymbol{x}_{new}^B, y_{new}^B)$
$D_{t+1} \leftarrow D_t + (\boldsymbol{x}_{new}^B, y_{new}^B)$
$t \leftarrow t+1$
query $(\boldsymbol{x}_*^A, y_*^A) = SecondSample(D_1)$ to human then add new *minority* data $(\boldsymbol{x}_{new}^B, y_{new}^B)$
$D_{t+1} \leftarrow D_t + (\boldsymbol{x}_{new}^B, y_{new}^B)$
$t \leftarrow t+1$
while $t < T$ **do**
  for all $\boldsymbol{x}_i^A$ in A_t, select 1&2 nearest neighbor $\boldsymbol{x}_{n_1}^B, \boldsymbol{x}_{n_2}^B$ from B_t, then select sample data $(\boldsymbol{x}_*^A, y_*^A)$ by *equation* (2)
  query $(\boldsymbol{x}_*^A, y_*^A)$ to human then add new *minority* data $(\boldsymbol{x}_{new}^B, y_{new}^B)$
  $D_{t+1} \leftarrow D_t + (\boldsymbol{x}_{new}^B, y_{new}^B)$
  $t \leftarrow t+1$
end while

function FisrtSample(D_0)
  select the nearest *majority* data $\boldsymbol{x}_{nearest}^A$ to $\boldsymbol{x}_1^B \in B_0$ then make vector $\boldsymbol{v}_1 = \boldsymbol{x}_{nearest}^A - \boldsymbol{x}_1^B$
  for all $\boldsymbol{x}_i^A$ ($i \neq nearest$) in A_0, make vector $\boldsymbol{v}_i = \boldsymbol{x}_i^A - \boldsymbol{x}_1^B$ and calculate *cos similarity* between $\boldsymbol{v}_1$ and $\boldsymbol{v}_i$
  return $(\boldsymbol{x}_*^A, y_*^A) = \arg\min_{\boldsymbol{x} \in A_0}$ *cos similarity*$(\boldsymbol{v}_1, \boldsymbol{v}_i)$
end function

function SecondSample(D_1)
  make vector $\boldsymbol{v}_1 = \boldsymbol{x}_2^B - \boldsymbol{x}_1^B$ $(\boldsymbol{x}_1^B, \boldsymbol{x}_2^B \in B_1)$
  for all $\boldsymbol{x}_i^A$ in A_1, make vector $\boldsymbol{v}_i = \boldsymbol{x}_i^A - \boldsymbol{x}_1^B$ and calculate *cos similarity* between $\boldsymbol{v}_1$ and $\boldsymbol{v}_i$
  return $(\boldsymbol{x}_*^A, y_*^A) = \arg\min_{\boldsymbol{x} \in A_1}$ *cos similarity*$(\boldsymbol{v}_1, \boldsymbol{v}_i)$
end function

明を行う.

5.1 実験設定



(a) Log **(b)** Square **(c)** Sin

図 3: 設定した境界線 (*majority* クラス：青, *minority* クラス：赤)

実験では, まず一様なデータを生成し, 設定した境界により *majority* クラスと *minority* クラスに分けたものを使用する. クラス分けには, Log, Square, Sin の 3 種類の境界 (真の境界) を使用しており (図 3), 以降の検証は 3 パターンのデータに対して行ったものである. クラス分け後, 8：2 の比率で訓練データとテストデータに分割し, 訓練データはシミュレーションに, テストデータはイテレーション t 時点での AUC の算出に使用した. 実験はデータ 1,000 件を *majority* クラスと *minority* クラスがおおよそ 5 : 5 になるよう分割し, AUC の算出にはもっとも単純なモデルの 1 つである深さ 2 の決定木を使用した.

また, 本研究におけるシミュレーションは, まず訓練データの内 *majority* データすべてと *minority* データ 1 件を D_0 としてスタートし, 残りの *minority* データは手元に控えておく. その後, 提示された *majority* データと手元に控えておいた *minority* データからそれと最も近い *minority* データを平均, *majority* データと先ほどの *minority* データ間の距離を 3σ とする正規分布からデータを生成し追加している. なお, *majority* データの提示は 1 つにつき 1 度きりとする.

戦略の比較は, uncertainty strategy と nearest neighbor strategy に加え, *majority* データをランダムに提示する戦略 (random) で行った. さらに, uncertainty sampling strategy は 4 節で説明した 3 つのバリエーションの比較, nearest neighbor strategy は式 (2) に加え, それぞれの項の振る舞いを見るため, 第 1 項 (*majority* データと最近傍の *minority* データ 2 点の平均距離) のみと第 2 項 (各 *majority* データの最近傍の *minority* データ 2 点の距離の逆数) のみの場合の比較も行った.

評価は 3.3 項でも述べた通り, 各イテレーションで拡張されたデータプール D_t で訓練した分類器 f_t のテストデータに対する AUC と, 提示した *majority* データとそれに最も近い D_0 に含めなかった (控えの)*minority* データ間の距離の累計で行う. イテレーション $t=i$ 時点での累計は, $t=1 \sim i$ までの距離の総和となる. 繰り返しになるが, データ間の距離については, 提示したデータと類似したデータを人間に追加してもらうことを想定した際, 真の境界付近の *majority* データを提示することが望ましいと考えられるため, 採用した.

実験の結果を図 4, 表 4, 図 5 に示した. なお, 実験結果は 10 回実行した値の平均をとったものである. 図 4 は, Log・Square・Sin それぞれの境界について, イテレーションによる AUC の変化と提示したデータと追加したデータ間の距離の累計（10 回実行の平均）を表したものであり, 図 5 はデータを 10, 30, 50, 75, 100, 150, 200, 300 個追加時点での AUC について実行 10 回分のばらつきを箱ひげ図としたものである. 表 4 は各ステップ（イテレーション）における各戦略 (及びバリエーション等) の AUC/距離の累計を数値にしたものであり, 各ステップで最も距離の累計の小さいものを**太字**, 最も AUC の高いものを下線付き, 2 番目に ACU の高いものを2 重下線付きで強調している.

5.2 Uncertainty Strategy の比較

ここでは uncertainty strategy について 3 つのバリエーショ

ンとランダムにデータを提示する方法 (random) を比較する. 図 4 及び表 4 の距離の累計を見ると, どの境界線についても uncertainty strategy のバリエーションは random より小さい値となっており, uncertainty を使用することで境界線付近のデータを提示できていることが分かる. またバリエーション同士については, uncertainty-var と uncertainty-nearest が uncertainty-random に比べ距離の累計や AUC が若干優位であることから, 複数の uncertainty の最も高いデータからどのように提示するデータを選ぶかを考慮することでより境界線付近かつ AUC が高くなるようなデータを提示できることが示唆される. また, 3 種類の境界すべてにおいて序盤では uncertainty-var が uncertainty-nearest よりも距離の累計において優れているのは, uncertainty-nearest が現在のデータプールの *minority* データに最も近いものを選択するものであり *minority* データの分布の影響を受けるためであると考えられる. 実際, 表 4 を見ると, *minority* データが充実してくる後半のステップでは, uncertainty-nearest が uncertainty-var より距離の累計が良い結果となっている.

AUC については, 図 5 より, 10 回の実行でばらつきが大きく, どのバリエーション・境界線をとっても random と同等のばらつき具合となっており, ばらつきが小さくても全体的に AUC が低い結果となった. Log や Square では uncertainty-nearest が, Sin では uncertainty-var が比較的ばらつきが小さく AUC が高い様子が見られるが, random や unceratnainty-random と比べ顕著な差があるとは言い難い.

random 及び uncertainty strategy の比較を通して, uncertainty を適用することである程度境界線付近のデータを提示することができることが分かった. 一方で AUC の向上具合については厳しい戦略であるともいえる. これは, この戦略で適用した uncertainty は本来ラベルの付いていないデータ, すなわち訓練データでないデータに適用する基準であるものを訓練データとして使用したラベル付きのデータに直接適用したことが一因として挙げられるため, より提案手法のシチュエーションに当てはまるような修正や新たな基準の模索などでより良い結果が期待できると考えられる.

5.3 Nearest Neighbor Strategy の比較

nearest neighbor strategy の比較では, uncertainty strategy でも比較した uncertainty-random 及び random の他, 式 (2) の第 1 項のみ (nearest-AtoB) と第 2 項のみ (nearest-BtoB) の場合との比較を行う.

図 4 と表 4 を見ると, 距離の累計ではすべての境界について nearest neighbor strategy が優位な結果となった. 特に式 (2) の第 1 項のみでサンプリングを行う nearest-AtoB は, 3 つの境界線について距離の累計 AUC ともに良い結果となっており, それに nearest-hybrid が次ぐ結果となっている. しかしながら図 5 を見ると, nearest-AtoB は AUC のばらつき具合の大小が激しく, それに比べ nearest-hybrid は AUC のばらつきが全体的に小さくかつスコア自体も高い傾向があるため, 第 2 項の影響で nearest-hybrid の方が nearest-AtoB よりも効果的なデータを安定して提示できていると考えられる.

また, 戦略の中でも最も単純なものである uncertainty-random と比較しても, 距離の累計・AUC ともに nearest neighbor strategy が優位を示す結果となっている.

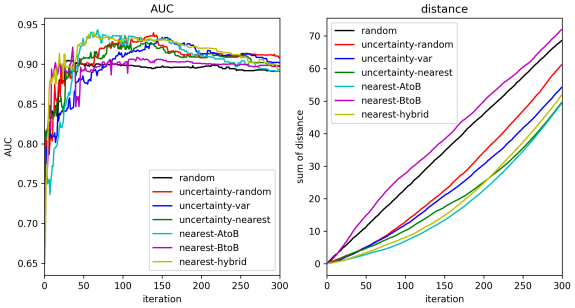
シミュレーションの結果より, nearest neighbor strategy はどの境界のデータに対しても境界付近かつ学習にとっても有用なデータをサンプリングできていることが分かった. この戦略は式 (2) に基づきサンプリングを行ったものであり, 基本的には第 1 項のみでも非常によく機能しているが, 第 2 項の影響によりさらに学習に有効なデータをサンプリングできると考えられる. 本研究では, 第 1 項と第 2 項はスケールをそろえたが, 第 1 項と第 2 項の重みを変えるなどの工夫でさらに良い結果が期待できる戦略である.

5.4 戦略同士の比較

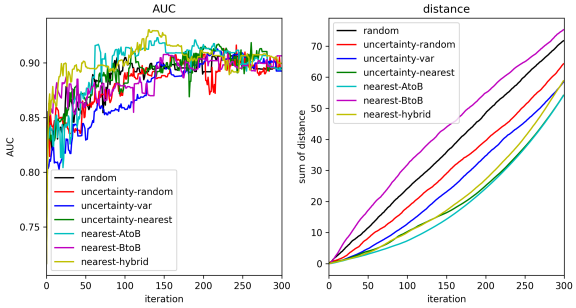
ここでは, 戦略同士の比較, すなわち uncertainty strategy と nearest neighbor strategy, そして random の比較を行う. 図 4 の距離の累計を見ると, nearest-BtoB を除くすべてのバリエーションが random よりも良い結果となった. 同じく距離の累計について戦略同士の比較では, Log・Square・Sin の境界すべてにおいて nearest-AtoB が最もよく, nearest-hybrid と uncertainty-nearest, uncertainty-var の順となり, 最近傍を取り入れた戦略はある程度優位を示す結果となった. ただし, Sin の境界については uncertainty-var も AUC・距離の累計ともに比較的良い結果を示していることから, 境界の形あるいはデータの分布と戦略にはある程度相性があることが示唆される.

AUC については, 表 4 から uncertainty strategy より nearest neighbor strategy の方が同じ数のデータを追加したときにより高い値となっており, 少ないステップで AUC がより高くなるようなデータプールへと拡張できていることが分かる. 距離の累計に注目すると nearest-AtoB が最も良い手法であるように見えるが, 図 5 ではこのバリエーションは AUC のばらつきの大小があり, 拡張が上手くいくケースとそうでないケースが存在しているといえるのに対し, 第 2 項を加えた nearest-hybrid は特にデータを 100 個追加するまでの立ち上がりがよく, 距離の累計に関しても表 4 から序盤は nearest-AtoB と大差ない結果となっており, 追加したデータが少ない序盤でも比較的分類器の精度向上に有効なデータの提示ができていると考えられる.

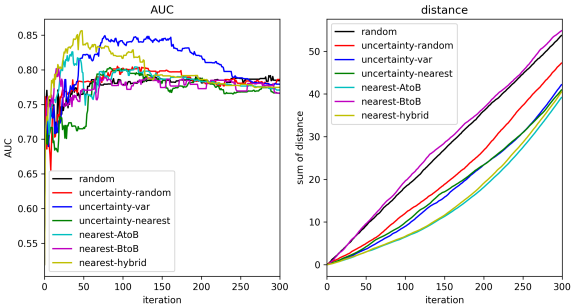
以上より, 全体的に uncertainty strategy と比べ nearest neighbor strategy の方が境界付近かつ AUC が高くなるようなデータの提示ができていると考えられる. また, nearest-AtoB は境界付近のデータを提示することには長けているが, AUC の向上は実行ごとの差が大きいため, nearest-AtoB よりも多少境界から離れたデータを提示してしまうが, nearest-hybrid の方が学習にとって効果の高いデータの選択・提示ができていると考えられる.



(a) Log



(b) Square

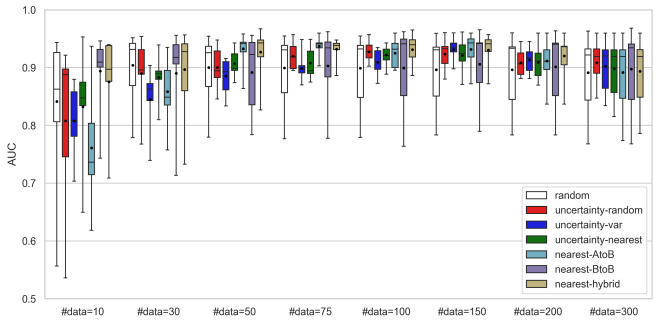


(c) Sin

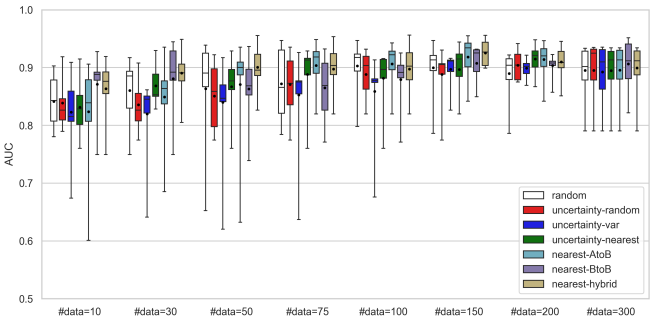
図 4: 各境界の AUC 及び距離の累計

表 4: step に注目した表 AUC, 距離の累計

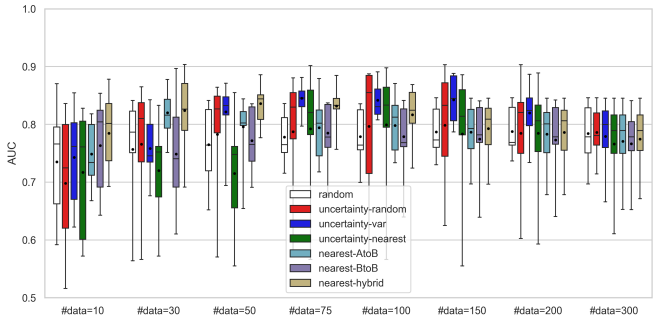
boundary	step	random	uncertainty-random	uncertainty-var	uncertainty-nearest	nearest-AtoB	nearest-BtoB	nearest-hybrid
Log	10	0.841,2.292	0.808,0.876	0.808,0.37	0.832,0.697	0.761,0.362	0.894,1.945	**0.875,0.361**
	30	0.904,6.719	0.89,2.879	0.845,2.525	0.883,2.72	**0.858,1.497**	0.89,8.32	0.897,1.62
	50	0.9,11.025	0.9,5.039	0.885,4.966	0.907,4.631	**0.933,2.771**	0.892,14.561	0.927,3.323
	75	0.9,17.079	0.92,8.395	0.899,8.084	0.908,7.215	**0.937,4.455**	0.903,21.923	0.932,5.662
	100	0.899,22.816	0.927,12.613	0.909,11.799	0.922,10.066	**0.925,6.816**	0.899,27.55	0.931,7.961
	150	0.896,34.319	0.923,22.38	0.931,20.766	0.921,17.261	**0.932,13.242**	0.906,38.048	0.93,14.668
	200	0.896,45.81	0.908,34.044	0.913,30.254	0.909,24.931	**0.911,22.403**	0.901,49.672	0.92,24.575
	300	0.891,68.569	0.908,61.096	0.902,54.176	0.898,49.461	**0.891,49.174**	0.897,71.917	0.893,51.77
Square	10	0.841,2.181	0.839,1.067	0.823,0.439	0.832,0.661	**0.823,0.408**	0.871,2.619	0.863,0.551
	30	0.86,7.117	0.835,3.945	0.82,2.34	0.868,2.07	**0.849,1.696**	0.881,10.153	0.891,1.758
	50	0.864,11.348	0.851,7.974	0.84,4.726	0.867,3.949	**0.871,2.999**	0.863,16.866	0.901,3.471
	75	0.872,17.716	0.871,12.557	0.853,8.374	0.888,6.371	**0.903,4.956**	0.865,23.854	0.897,6.622
	100	0.903,23.93	0.888,17.891	0.859,12.608	0.881,10.162	**0.906,7.249**	0.879,31.568	0.897,9.867
	150	0.9,35.609	0.888,28.747	0.897,22.939	0.897,16.286	**0.918,14.154**	0.908,43.726	0.925,16.969
	200	0.89,48.225	0.904,39.387	0.899,34.522	0.915,24.891	**0.914,24.008**	0.903,54.841	0.91,26.949
	300	0.895,71.696	0.895,64.29	0.892,58.448	0.895,54.141	**0.895,54.051**	0.906,75.257	0.9,58.834
Sin	10	0.735,1.589	0.698,0.959	0.743,0.448	0.717,0.569	0.749,0.393	0.763,1.268	**0.784,0.366**
	30	0.757,5.144	0.766,2.686	0.758,1.836	0.72,2.143	**0.82,1.521**	0.749,5.329	0.824,1.636
	50	0.765,8.965	0.783,4.842	0.821,3.594	0.715,4.1	0.796,2.956	0.772,9.444	**0.836,2.916**
	75	0.778,13.588	0.787,8.121	0.845,6.191	0.792,6.9	**0.794,4.646**	0.785,14.472	0.832,4.762
	100	0.779,18.058	0.796,11.915	0.842,8.876	0.799,9.747	**0.798,6.411**	0.779,19.431	0.816,6.565
	150	0.787,26.852	0.798,18.687	0.842,15.669	0.783,17.012	**0.787,11.201**	0.774,28.364	0.793,11.486
	200	0.788,35.902	0.784,26.612	0.82,23.243	0.784,23.29	**0.783,17.997**	0.773,36.575	0.786,18.924
	300	0.784,53.706	0.786,47.233	0.779,42.167	0.766,40.932	**0.771,39.172**	0.766,54.817	0.775,40.47



(a) Log



(b) Square



(c) Sin

図 5: 各ステップにおける 10 回の AUC

6 議　　論

6.1 シミュレーションについて

本研究では, シミュレーションを通じて手法の検証を行った. シミュレーションでは提示したデータが境界から離れているほど不正確なデータを, 近いほど正確なデータを追加したが, 実際のデータの追加は人間に行ってもらうことを想定しており提示した *majority* データに類似している *minority* データがどのように追加されるかは, 未知数な部分が大きいといえるため, 今後は実実験を通して提案手法の効果を検証する必要が極めて高いといえる.

6.2 戦略について

本研究では, uncertainty を用いた uncertainty strategy, *majority* データ 1 件とそれの最近傍の *minority* データ 2 件を 3 つ組とした nearest neighbor strategy を考案した. uncertainy の使用については, 本来ラベルの付いていないデータに適用するものをラベル付きのデータにそのまま使用していたため, この点に関してより有効な基準やラベル付きデータに適用可能な形に修正するなどの改善が考えられる. また nearest-hybrid では, サンプリングするデータを式 (2) で評価をした. この式では第 1 項と第 2 項のスケールをそろえたが, 例えばそれぞれの項に重みを加えるなどして挙動を調整することができるため, 改善の余地があるといえる.

7 ま と め

本研究では, 特定のサンプリング戦略に則りラベル付きデータプールから *majority* クラスのデータを提示し, それと類似した *minority* クラスのデータを人間に追加してもらうことでデータ不足/不均衡の問題に対処したアクティブ・データプール拡張手法を提案した. サンプリング戦略には, uncertainty を使用した uncertainty strategy と *majority* データとそれの最近傍の *minority* データに注目した nearest neighbor strategy を考案し, シミュレーションを通してそれぞれのバリエーションと共に検証した.

今回はシミュレーションを使用した検証のみに留まる形となってしまったため, 今後は実実験を行い実際に有効な結果を得られるか検証していきたい. また, サンプリング戦略についても, 戦略そのものの改良や uncertainty strategy と nearest neighbor strategy の組み合わせ, 新たな戦略の考案・検証を行っていきたい.

8 謝　　辞

本研究の一部は, JSPS 科研費 (JP22H00508) と JST CREST(Grant Number JPMJCR22M2) の支援を受けたものである. ここに謝意を示す.

文　　献

[1] B. Settles, “Active learning literature survey,” Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
[2] W.-N. Hsu and H.-T. Lin, “Active learning by learning,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, 2015.
[3] S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” Journal of machine learning research, vol. 2, no. Nov, pp. 45–66, 2001.
[4] B. Settles and M. Craven, “An analysis of active learning strategies for sequence labeling tasks,” in proceedings of the 2008 conference on empirical methods in natural language processing, pp. 1070–1079, 2008.
[5] Y. Liu, “Active learning with support vector machine applied to gene expression data for cancer classification,” Journal of chemical information and computer sciences, vol. 44, no. 6, pp. 1936–1941, 2004.
[6] H. Kaur, H. S. Pannu, and A. K. Malhi, “A systematic review on imbalanced data challenges in machine learning: Applications and solutions,” ACM Computing Surveys (CSUR), vol. 52, no. 4, pp. 1–36, 2019.
[7] C. C. Aggarwal, “Outlier analysis second edition,” 2016.
[8] V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 4, pp. 42–47, 2012.
[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.
[10] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, “Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary,” Journal of artificial intelligence research, vol. 61, pp. 863–905, 2018.
[11] X. Liang, A. Jiang, T. Li, Y. Xue, and G. Wang, “Lr-smote —an improved unbalanced data set oversampling based on k-means and svm,” Knowledge-Based Systems, vol. 196, p. 105845, 2020.
[12] Z. Zheng, Y. Cai, and Y. Li, “Oversampling method for imbalanced classification,” Computing and Informatics, vol. 34, no. 5, pp. 1017–1037, 2015.
[13] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, “Learning from imbalanced data in surveillance of nosocomial infection,” Artificial intelligence in medicine, vol. 37, no. 1, pp. 7–18, 2006.
[14] B. Li, Y. Hou, and W. Che, “Data augmentation approaches in natural language processing: A survey,” Ai Open, vol. 3, pp. 71–90, 2022.
[15] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of big data, vol. 6, no. 1, pp. 1–48, 2019.
[16] Y. Zhou, C. Guo, X. Wang, Y. Chang, and Y. Wu, “A survey on data augmentation in large model era,” arXiv preprint arXiv:2401.15422, 2024.

AGATE：属性付き時間発展グラフの汎用予測フレームワーク

山崎　翔平† 佐々木勇和†† Panagiotis Karras††† 鬼塚　真††

† 株式会社野村総合研究所
†† 大阪大学大学院情報科学研究科
††† Department of Computer Science, University of Copenhagen
E-mail: †s2-yamasaki@nri.co.jp, ††{sasaki,onizuka}@ist.osaka-u.ac.jp, †††piekarras@gmail.com

あらまし　属性付き時間発展グラフを対象としたノード属性予測やリンク予測等の予測タスクでは，個々の予測タスク間の依存関係を考慮しておらず，独立した問題として取り組まれている．また，過去に観測されていない，将来新たに出現するノード集合を想定した予測技術は我々の知る限り存在しない．本稿では属性付き時間発展グラフに対して「ノード集合，リンク集合，ノード属性が互いに影響し合うこと」および「過去に観測されていないノードが新規に出現すること」を仮定し，これら全てを同時に予測する問題に取り組む．上述の問題を解くにあたり，以下 2 つの研究課題を設定する．(Q1) タスク間の相互依存性を利用して予測精度を向上させることは可能か？(Q2) 出現ノードの属性予測は可能か？属性付き時間発展グラフの全体を将来予測するというタスクを通して，研究課題の 2 つの問いに対する解を提示する．本稿では，(Q1) に対する汎用予測フレームワーク AGATE と，(Q2) に対する出現ノードの属性予測モデルである PROSER の 2 つの新手法を提案する．実験では 3 つの実世界のデータセットを用いて AGATE と PROSER の有効性を示す．

キーワード　グラフ深層学習，属性付き時間発展グラフ

1. はじめに

グラフ構造は現実世界の複雑な現象をモデル化する一般的なデータ表現である．グラフ構造で表現されるデータとしてはソーシャルネットワークや論文の引用ネットワークが挙げられる．実世界に存在するグラフ構造の多くはノードが属性を有し，また時間変化によってグラフ構造が変化する [1–4]．例えば，ソーシャルネットワークでは，登録ユーザはノードとして表現され，そのプロフィール情報はノード属性として，ユーザ同士の繋がりはリンクとして表現される．そして，登録ユーザや，そのプロフィール情報，ユーザ同士の繋がりは時間とともに変化する．このような時間的に変化しやすいグラフは，属性付き時間発展グラフと呼ばれる [5]．SNS やブロックチェーン，IoT 等の急速な普及に伴い，属性付き時間発展グラフとしてモデル化できる対象はますます増加傾向にあり，その応用の一つとして属性付き時間発展グラフを対象とした予測技術が注目を集めている．特に，将来のグラフにおけるノード属性を予測する属性予測や，グラフ構造を予測するリンク予測は実世界での利用性の高さから多くの技術が提案されている [6–8]．

属性付き時間発展グラフを対象とした属性予測やリンク予測等の各予測タスクは，個別に解かれるのが既存研究の主流である．これはノード属性（あるいはグラフ構造）が時間的に不変であることを仮定し，リンク予測（あるいは属性予測）の単一の問題に帰着させているためであり，それぞれの予測タスク間の依存関係は考慮しておらず，独立した問題として取り組まれている．ノード属性とグラフ構造の両方の時間変化を仮定した場合，互いに影響し合うため既存の予測技術を適用することは難しい．また，既存研究においてこれらの予測タスクでは，暗黙的に過去に観測されているノード集合のみを予測の対象としており，将来新たに出現するノードや消失するノードは想定していない．特に，将来新たに出現するノードの属性を予測し，それに接続されるリンクを予測することは容易ではない．実世界に存在する属性付き時間発展グラフの多くはノード集合，リンク集合，およびノード属性が影響し合いながら時間変化し，さらに過去に観測されていない新規のノードが将来のグラフに出現する可能性が十分にある．

本稿では属性付き時間発展グラフに対して「ノード集合，リンク集合，ノード属性が互いに影響し合うこと」および「過去に観測されていないノードが新規に出現すること」を仮定し，属性付き時間発展グラフの全体を将来予測する問題に取り組む．この問題を属性付き時間発展グラフ予測問題と呼ぶ．属性付き時間発展グラフ予測問題は属性予測やリンク予測等の既存の予測タスクを包含しているため，これを解くことでより多くの分野へ応用できることが期待される．しかし，この問題を解くことは容易ではなく筆者らの知る限り既存研究も存在しない．属性付き時間発展グラフ予測問題を解くには，ノード属性とグラフ構造の依存関係やノードの出現，あるいは消失する可能性も考慮しながらグラフの全体像を予測する必要がある．この問題を解くにあたり，以下 2 つの研究課題を設定する．

(Q1)　各予測タスク間の相互依存性を利用して予測精度を向上させることは可能か？属性付き時間発展グラフ予測問題は，属性予測やリンク予測等の複数のサブ予測タスクを含んでいる．これらのサブ予測タスク間には相互依存性が存在し，これを捉えることで，それぞれのサブ予測タスクを独立に解く場合よりも，高い予測精度を達成できるのではないかという仮説である．

(Q2)　出現ノードの属性予測は可能か？出現ノードの属性予測

では，過去に観測されていない，将来新たに出現するノード集合を予測すると共に，それらが持つ属性を同時に予測する必要がある．既存の手法では出現ノードの属性を予測することができないため新しい予測手法が必要である．

本稿では，属性付き時間発展グラフ予測問題を通して，(Q1) に対する解である AGATE と (Q2) に対する解である PROSER を提案する．AGATE は複数の手法を効果的に組み合わせることで，ある予測タスクの結果を他の予測タスクの学習に再利用することができる汎用予測フレームワークであり，タスク間の相互依存性を活用することができる．PROSER は出現ノードの属性予測モデルであり，将来新たに出現するノード集合とそれらが持つ属性を高精度で予測することができる．実験では，3 つの実世界のデータセットを用いて AGATE と PROSER を評価し，(1) 各サブ予測タスクが互いに影響し合うこと，(2) AGATE が各サブ予測タスクを独立に解く場合よりも高精度で予測できること，(3) PROSER が出現ノードの属性を高精度で予測できることを示した．本研究は，グラフ全体を将来予測するという一般的な問題に初めて取り組んだものであり，その第一歩として汎用的な方法を提案した．

本研究は発表済みの論文 [9] に基づいており，コードベース (注1) は GitHub 上に公開されている．

2. 問題定義

グラフを $G = (V, E, X)$ と表記する．ここで，V はノードの有限集合，$E \subseteq V \times V$ はリンクの有限集合，$X \in \mathbb{R}^{|V| \times d}$ はノード属性の有限集合である．$X^{(v)}$ はノード v の d 次元の属性を表しており，各属性はカテゴリ値，数値，分散表現のいずれも取り得る．ここでは，ノード集合，リンク集合，ノード属性が時間と共に変化する属性付き時間発展グラフを考えるが，d は時間によらず一定とする．

定義 1 (属性付き時間発展グラフ) 属性付き時間発展グラフは一定の時間間隔で表現されるグラフのシーケンス (注2) であるとし，$\langle G_1, \ldots, G_T \rangle$ と表記する．ここで T は観測されたタイムステップ数を表しており，$G_t := (V_t, E_t, X_t)$ はタイムステップ t のグラフを表す．

本稿では，タイムステップ t のノード V_t を以下の式で表される，**既知ノード** V_t^e，**出現ノード** V_t^n，**消失ノード** V_t^l の 3 種類に分けて考える．

$$V_t^e = V_t \cap V_{t-1}, \quad V_t^n = V_t \setminus V_{t-1}, \quad V_t^l = V_{t-1} \setminus V_t.$$

V_t^e はタイムステップ t と $t-1$ の両方に存在するノード，V_t^n はタイムステップ t に出現するノード，V_t^l はタイムステップ t に消失するノードである．同様にリンクについても，**既知リンク** E_t^e，**出現ノードに接続されるリンク** E_t^n，**出現リンク** E_t^a，**消失リンク** E_t^d の 4 種類に分類する．

(注1)：https://github.com/yuya-s/AGATE/
(注2)：これは一般に離散時間動的グラフ（DTDG：Discrete Time Dynamic Graphs）と呼ばれるものである．

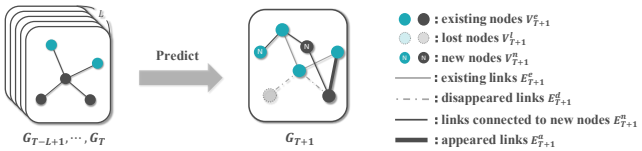


図 1　属性付き時間発展グラフ予測問題の例 [9]．ノードの色は属性を表す．T は現在のタイムステップ，L はタイムステップ数を表す．

$$E_t^e = E_t \cap E_{t-1}, \quad E_t^n = E_t \setminus (V_{t-1} \times V_{t-1})$$

$$E_t^a = (E_t \setminus E_{t-1}) \setminus E_t^n, \quad E_t^d = E_{t-1} \setminus E_t$$

本稿で取り組む問題定義を以下に示す．

問題定義 1 (属性付き時間発展グラフ予測問題) 属性付き時間発展グラフ $\langle G_{T-L+1}, G_{T-L+2}, \ldots, G_T \rangle$ が与えられたとき，タイムステップ $T+1$ におけるグラフ $G_{T+1} = (V_{T+1}, E_{T+1}, X_{T+1})$ を予測する．

図 1 に本稿で取り組む問題の全体像を示す．消失ノードの予測や出現ノードの予測など，複数のサブ予測タスクを含む問題となっている．この問題を解くには，G_{T+1} の構成要素の一つ (V_{T+1}, E_{T+1}, X_{T+1} のみ) ではなく，全てを同時に予測する必要がある．そのためには，次に示すサブ予測タスクの全てを同時に解かなければならない．

サブ予測タスク: 属性付き時間発展グラフ予測問題は 6 つのサブ予測タスクから構成されており，既知ノードに関するサブ予測タスクと，出現ノードに関するサブ予測タスクの 2 種類に大別される．**既知ノードに関するサブ予測タスク:** 既知ノードに関するサブ予測タスクは，既知ノードの属性やグラフ構造を予測することを目的とする．具体的には，消失ノード予測、出現リンク予測，消失リンク予測，既知ノードの属性予測の 4 つのサブ予測タスクが存在する [9]．既存研究 [6, 10] でも取り組まれているが，独立に解かれており，同時に予測する研究は存在しない．これらのタスクは直感的に理解でき，紙面も限られているため本稿での詳細な説明は省略する．**出現ノードに関するサブ予測タスク:** 出現ノードに関するサブ予測タスクは，出現ノードの属性と，出現ノードに接続されるリンクを予測することを目的とする．出現ノードに接続されるリンク予測は，最近の研究 [11] でも扱われているが，出現ノードの属性予測に関する研究は存在しない．

部分問題 1 (**出現ノードの属性予測**): 属性付き時間発展グラフ $\langle G_{T-L+1}, G_{T-L+2}, \ldots, G_T \rangle$ が与えられたとき，このサブ予測タスクでは，出現ノードとその属性を予測する．予測された出現ノードを $\hat{V}_{T+1}^n$，予測された出現ノードの属性を $\hat{X}_{T+1}^n$ と表記する．

出現ノードの属性予測に関する既存研究はないため，この予測タスクのための評価尺度を新たに定義する．本稿では予測された出現ノード $\hat{V}_{T+1}^n$ と真の出現ノード V_{T+1}^n 間の類似性を，完全マッチングに基づいて評価する．

定義 2 (**完全マッチング**): 完全二部グラフ $K = (V, V', E)$

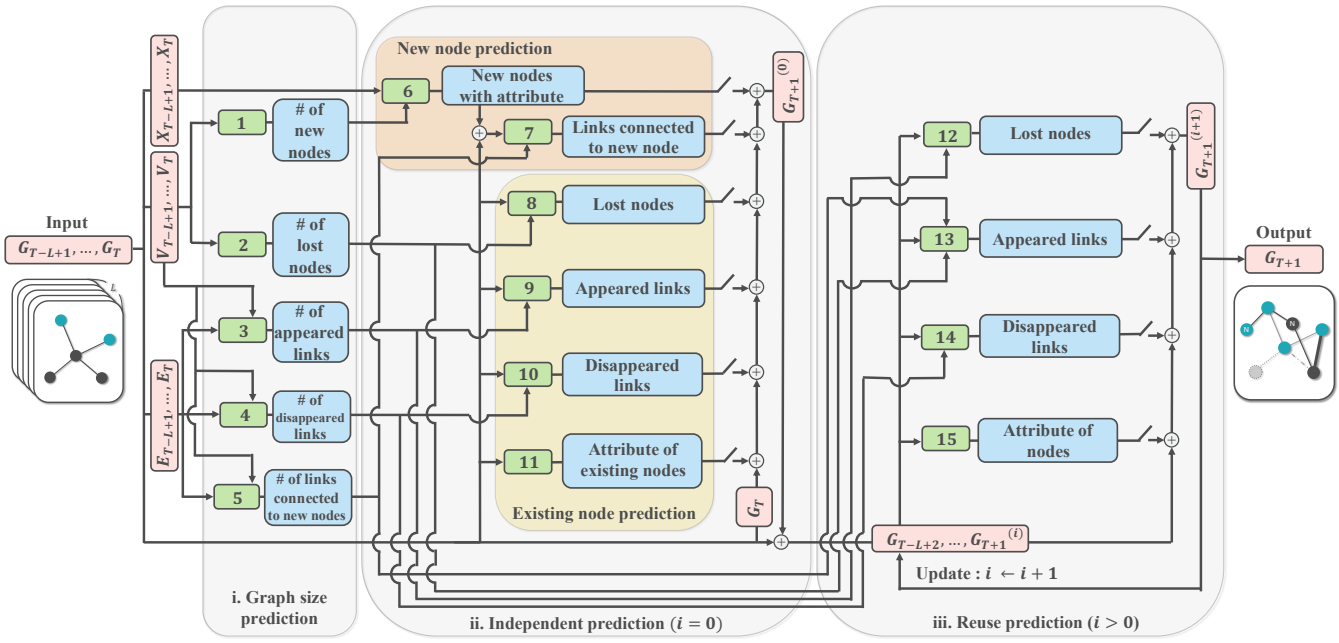


図 2　AGATE のアーキテクチャ [9]．観測された L ステップのグラフを入力し，タイムステップ $T+1$ のグラフを予測する．ピンクのフレームはグラフデータを，緑のフレームは予測モデルを，青のフレームは予測結果を表す．

$(E = V \times V')$ が与えられたとき， K における完全マッチング M は，V 内のすべてのノードをカバーする，対になった非隣接リンクの集合である．

予測された出現ノードの集合 $\hat{V}^n_{T+1}$ と真の出現ノードの集合 V^n_{T+1} からなる完全二部グラフ $K = (\hat{V}^n_{T+1}, V^n_{T+1}, \hat{V}^n_{T+1} \times V^n_{T+1})$ を構築し，任意の類似度関数によって辺の重みを計算する．ノード属性の種類（カテゴリ値，数値，分散表現など）に応じて，ノード属性間の類似度関数 $w(\cdot,\cdot)$ はコサイン類似度やユークリッド距離等に基づく関数が利用される．この予測タスクの評価尺度は以下に示すマッチング類似度とする．

$$\text{M-sim}(\hat{V}^n_{T+1}, V^n_{T+1}, \hat{X}^n_{T+1}, X^n_{T+1}) = \max_{M \in \mathcal{M}} \frac{1}{|M|} \sum_{(\hat{v},v) \in M} w(\hat{X}^{(\hat{v})}_{T+1}, X^{(v)}_{T+1}) \quad (1)$$

ここで，$\mathcal{M}$ は与えられた 2 部グラフの完全マッチングの集合であり，$|M|$ は M 中のリンクの数である．マッチング類似度が 1 に近いほど，予測された出現ノードの属性は、それに対応する真の出現ノードの属性と類似しているとみなすことができる．この評価尺度によって，予測された属性が真のものと適切に一致しているかどうかを評価することができる．一般に、最適な M は完全マッチングのパターンが多数あるため計算が困難であることから，貪欲法などの近似的な方法を用いる．

既存研究 [11] では，出現ノードに接続されるリンクを予測しているが，出現ノードは属性とともに入力として与えられることを想定している．一方で，本稿では出現ノードも予測対象であるため，予測された出現ノードに接続されるリンクを予測する．以下に定義を示す．

部分問題 2　(**出現ノードに接続されるリンク予測**): 属性付き時間発展グラフ $\langle G_{T-L+1}, \ldots, G_T \rangle$ と，部分問題 *1* で予測された出現ノード $\hat{V}^n_{T+1}$ とそれに対応する属性 $\hat{X}^n_{T+1}$ が与えられた時，このサブ予測タスクでは v と $\hat{v}$ $(v, \hat{v}) \in V_T \times \hat{V}^n_{T+1}$ の間に発生し得るリンクを予測する．

以上で説明した 6 つのサブ予測タスクは，相互に依存している可能性がある．例えば，属性予測はリンク予測に影響しているかもしれないし，その逆もしかりである．サブ予測タスクが互いにどのように影響し合うかは不明であり，既知ノードと出現ノードに関するサブ予測タスク間の相互依存関係も不確定である．

3.　AGATE：汎用予測フレームワーク

属性付き時間発展グラフ全体の将来予測タスクで，高い予測精度を達成するためには，サブ予測タスク間の相互依存性を捉える必要がある．単一の手法ですべての相互依存性を捉えることは困難であるため，複数の手法を効果的に組み合わせ，あるタスクの結果を他のタスクでインタラクティブに再利用する．本稿では，観測された過去のグラフから，サブ予測タスク間の相互依存性を活用可能で汎用的な予測フレームワークである AGATE (*A General framework for predicting Attributed Time-Evolving graphs*) を提案する．図 2 に AGATE のアーキテクチャを示す．AGATE は以下の 3 つの主要なコンポーネントから構成される．**i. Graph size prediction :**　属性付き時間発展グラフの将来のサイズを予測するコンポーネント．**ii. Independent prediction :**　既知ノード間のグラフ構造や属性，出現ノードの属性とそれに接続されるリンクについて個別・独立に予測するコンポーネント．**iii. Reuse prediction**

Algorithm 1: AGATE

input : $\langle G_{T-L+1}, G_{T-L+2}, \ldots, G_T \rangle$, a natural number I
output: $\hat{G}^{(I)}_{T+1}$
/* Graph size prediction */
1 Extract sequences of new nodes, lost nodes, appearing links, disappearing links, and links of new nodes from $\langle G_{T-L+1}, G_{T-L+2}, \ldots, G_T \rangle$;
2 Count and normalize them by dividing its maximum value;
3 Predicting $|V^n_{T+1}|$, $|V^l_{T+1}|$, $|E^n_{T+1}|$, $|E^a_{T+1}|$, $|E^d_{T+1}|$ from each sequence of normalized values;
/* Independent prediction */
4 Predicting $\hat{V}^n_{T+1}$, $\hat{V}^l_{T+1}$, $\hat{E}^n_{T+1}$, $\hat{E}^a_{T+1}$, $\hat{E}^d_{T+1}$, $\hat{X}_{T+1}$,$\hat{X}^n_{T+1}$ from $\langle G_{T-L+1}, G_{T-L+2}, \ldots, G_T \rangle$ and predicted graph size values;
5 $\hat{V}^{(0)}_{T+1} \leftarrow V_T \cup \hat{V}^n_{T+1} \setminus \hat{V}^l_{T+1}$;
6 $\hat{E}^{(0)}_{T+1} \leftarrow E_T \cup \hat{E}^n_{T+1} \cup \hat{E}^a_{T+1} \setminus \hat{E}^d_{T+1}$;
7 $\hat{X}^{(0)}_{T+1} \leftarrow \hat{X}_{T+1} \cup \hat{X}^n_{T+1}$;
8 $\hat{G}^{(0)}_{T+1} \leftarrow (\hat{V}^{(0)}_{T+1}, \hat{E}^{(0)}_{T+1}, \hat{X}^{(0)}_{T+1})$;
/* Reuse prediction */
9 **for** $i = 1, \ldots, I$ **do**
10 Predicting $\hat{V}^n_{T+1}$, $\hat{V}^l_{T+1}$, $\hat{E}^n_{T+1}$, $\hat{E}^a_{T+1}$, $\hat{E}^d_{T+1}$, $\hat{X}_{T+1}$,$\hat{X}^n_{T+1}$ from $\langle G_{T-L+2}, \ldots, \hat{G}^{(i-1)}_{T+1} \rangle$;
11 $\hat{V}^{(i)}_{T+1} \leftarrow V_T \cup \hat{V}^n_{T+1} \setminus \hat{V}^l_{T+1}$;
12 $\hat{E}^{(i)}_{T+1} \leftarrow E_T \cup \hat{E}^n_{T+1} \cup \hat{E}^a_{T+1} \setminus \hat{E}^d_{T+1}$;
13 $\hat{X}^{(i)}_{T+1} \leftarrow \hat{X}_{T+1} \cup \hat{X}^n_{T+1}$;
14 $\hat{G}^{(i)}_{T+1} \leftarrow (\hat{V}^{(i)}_{T+1}, \hat{E}^{(i)}_{T+1}, \hat{X}^{(i)}_{T+1})$;
15 **return** $\hat{G}^{(I)}_{T+1}$;
16 **end procedure**

: サブ予測タスク間の相互依存性を捉えた予測を実現するために (ii) の結果を再利用するメカニズムを有した予測コンポーネント．AGATE はモジュール化されており，「ii. Independent prediction」および「iii. Reuse prediction」の各コンポーネント（6〜15 番）はそれぞれのサブ予測タスクに対応した任意の予測モデルを採用することが可能である．出現ノードの属性予測（6 番）については，既存手法が存在しないため，この予測タスクのために独自の手法を開発する．AGATE は部分的な時間変化を伴う属性付き時間発展グラフの予測にも対応可能である．具体的には，ノード属性が時間変化しない場合や，リンクが消失しないケース等にも対応できる．

Algorithm 1 に AGATE の擬似コードを示す [9]．まず，タイムステップ $T+1$ のグラフサイズを予測する（Graph size prediction：1 行目〜3 行目）．次に，各サブ予測タスクを独立に実行し，タイムステップ $T+1$ における暫定的な初期予測グラフ $\hat{G}^{(0)}_{T+1}$ を得る（Independent prediction：4 行目〜8 行目）．そして，他のタスクの結果を再利用して，各タスクの結果を更新する（Reuse prediction：9 行目〜14 行目）．以下では各要素について詳細を説明する．

3.1 Graph size prediction

Graph size prediction ではタイムステップ $T+1$ のグラフサイズを予測する．入力グラフ $\{G_{T-L+1}, \ldots, G_T\}$ から出現ノード $\{V^n_{T-L+1}, \ldots, V^n_T\}$，消失ノード $\{V^l_{T-L+1}, \ldots, V^l_T\}$，出現リンク $\{E^a_{T-L+1}, \ldots, E^a_T\}$，消失リンク $\{E^d_{T-L+1}, \ldots, E^d_T\}$，出現ノードに接続されるリンク $\{E^n_{T-L+1}, \ldots, E^n_T\}$ のシーケンスを抽出する．各シーケンスの要素数をカウントし，勾配消失を避けるために最大値で除算して正規化を行う．**LSTM** [12] 等の任意の時系列データ予測モデルを学習し，$T+1$ における出現ノード数，消失ノード数，出現リンク数，消失リンク数，出現ノードに接続されるリンク数を予測する．

3.2 Independent prediction

Independent prediction では全てのサブ予測タスクを独立に実行し，その予測結果を G_T に追加/削除することで，タイムステップ $T+1$ における暫定的な初期予測グラフ ${G_{T+1}}^{(0)}$ を構築する．サブ予測タスクが既知ノードに関するサブ予測タスクと，出現ノードに関するサブ予測タスクの 2 種類に大別されるように，Independent prediction においても，次の 2 種類のコンポーネントに細分化される．

Existing node prediction: このコンポーネントは，消失ノード予測，出現リンク予測，消失リンク予測に関する二値分類モデルと，既知ノードの属性予測に関する多クラス分類や回帰等の予測モデルから構成される．それぞれの分類モデルからノードが消失する確率/リンクが発生する確率/リンクが消失する確率を取得し，それぞれに対応する Graph size prediction での予測結果の数だけ，確率の高い順に追加/削除を行う．各コンポーネントでは適切な予測モデル（例えば，**LSTM** [12] や **DynGEM** [10]）を使用して予測する．筆者らの知る限り，**EvolveGCN** [13] は既知ノードに関するサブ予測タスクの全てをサポートしているが，他の予測モデルに比べて性能は低い（表 1 参照）．

New node prediction: 出現ノードの属性予測に取り組んでいる既存研究は存在しない．そこで，式 (1) のマッチング類似度を最大化し，予測された出現ノードと真の出現ノードの対応関係の決定を目的とした，3 つの単純なベースライン手法（**Random**，**FNN**，**PointNet** [14]）を提案する．**Random** は V^n_T からノード v をランダムにサンプリングし，そのノードに対応する属性 $X^{n(v)}_T$ を出力する．**FNN** は単純な Feedforward Neural Network で，以下の損失関数を最小化するように学習する．

$$\mathcal{L} = -\log \frac{\text{M-sim}(\hat{V}^n_{T+1}, V^n_{T+1}, \hat{X}^n_{T+1}, X^n_{T+1}) + 1}{2} \qquad (2)$$

FNN は全結合層を有するため学習結果はモデルへのノードの入力順序に依存する．**PointNet** は順序についての不変性 [14] を扱う深層学習モデルであり，損失関数は FNN と同じである．

予測された出現ノードの属性を利用して，出現ノードに接続されるリンク予測を行う．属性情報のみを持つノードのリンク予測を扱う **DEAL** [11] や，その他の類似の方法を用いて，出現ノードに接続されるリンクを予測する．予測精度の評価には，先に決定された予測された出現ノードと真の出現ノードの対応関係を用いる．

3.3 Reuse prediction

Reuse prediction では各サブ予測タスクの相互依存性を利用して予測結果を更新し，各サブ予測タスクの他への影響を捕捉す

ることを目的とする．Reuse prediction は，$\hat{G}_{T+1}$ の更新を I 回繰り返す．具体的には，$\langle G_{T-L+2}, \cdots, G_{T+1}{}^{(i)} \rangle$ $(0 \leqq i \leqq I-1)$ から消失ノード，出現リンク，消失リンク，ノード属性を予測し直し，$\hat{G}^{(i)}_{T+1}$ を更新する．このコンポーネントについても，Independent prediction と同様に，任意の予測モデル・手法を用いてもよい．

4. PROSER

本稿では，出現ノードの属性予測のための新しい手法である PROSER （*Probabilistic Selection Rule*）を提案する．PROSER は式 (1) のマッチング類似度を最大化することを目的とする．事前分析では，ほとんどの出現ノードは前のタイムステップの出現ノードと類似した属性を持つことが確認されていた．V^n_{T+1} からノードをランダムにサンプリングした結果を，$\hat{V}^n_{T+1}$ とすると，予測精度は比較的高くなる．PROSER は，ランダムサンプリングよりも予測精度を向上させるために，適切なノードを選択することによって，このサンプリングプロセスを改良する．タイムステップ $T+1$ での出現ノードの属性に対して，ランダムサンプリングされたノード属性の類似度が低い場合，マッチング類似度は低くなる．PROSER は，サンプリングされたノードと真のノードとの類似度を推定することにより，マッチング類似度を低くする要因となるノードの選択を回避する．PROSER はロジスティック回帰モデルを用いて，サンプリングされたノードと真のノードの類似度が閾値 θ より高いかどうかを学習する．θ は実際にマッチング類似度を最大化する閾値として事前に以下の式から計算される．

$$\theta = \arg\max_{\theta' \in \mathbb{R}} \sum_{G_i \in \mathcal{T}} \text{M-sim}(V^n_i(\theta'), V^n_{i+1}, X^n_i(\theta'), X^n_{i+1}) \qquad (3)$$

ここで，$\mathcal{T}$ は訓練データにおけるグラフの集合，$V^n_i(\theta')$ は真のノードとの類似度が θ' より高いサンプリングノードの集合，$X^n_i(\theta')$ はその属性の集合である．また，タイムステップ T における出現ノードの属性 X^n_T 分布を捉えるために，G_T の代表属性として X^n_T の平均ベクトルを用いる．ロジスティック回帰モデルは，この平均ベクトルとサンプリングされたノード属性を入力し，真のノードとの類似度が閾値より高くなる確率を出力する．推論時は，V^n_T からランダムにサンプリングされたノードを $\hat{V}^n_{T+1}$ に追加すべきか否かを，ロジスティック回帰モデルの出力確率にもとづいて判定する．

5. 評 価 実 験

AGATE の性能を実験的に検証し，タスクの相互依存性を利用することによる性能向上を分析する．すべての実験は，Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz を搭載した Linux サーバで実施している．

データセット：3 つの実世界のデータセットから属性付き時間発展グラフを構築し実験を行う：NBA(注3)，Reddit(注4)，and AMiner(注5)．

- **NBA**: ノードは NBA 選手を表し，得点，チーム，年齢，ポジションの属性を持つ．2 つのノードは，それぞれの選手がチームメイトであるときにリンク関係を持つ．ノード，リンク，およびノード属性は動的に変化し，ノード数は 3781，リンク数は 95203，属性は 35 次元，タイムステップ数は 66 である．
- **Reddit**: ノードは subreddits を表し，リンクは subreddits 間の投稿を表す．あるタイムステップで subreddits に投稿がない場合，対応するノードはそのタイムステップで削除される．ノード属性は subreddits の埋め込み [15] である．ノードとリンクは動的に変化するが，ノード属性は静的である．ノード数は 7756，リンク数は 23554，属性は 300 次元，タイムステップ数は 14 である．
- **AMiner** [16]: ノードは論文，リンクは引用関係，ノード属性は研究分野を表す．研究分野は，論文タイトルから TF-IDF スコアを計算して抽出される．ノードとリンクは動的に出現するが，ノードは消失せず，既知のノードのリンク，ノード属性は静的である．ノード数は 10987，リンク数は 6708，属性は 9 次元，タイムステップ数は 24 である．

タイムステップ数を 7 : 2 : 1 の比率で訓練，検証，テストデータに分割する．NBA では，それぞれ訓練 : 1–48，検証 : 49–60，テスト : 61–66 とする．Reddit は，訓練 : 1–11，検証 : 12–13，テスト : 14 である．AMiner は，訓練 ; 1–18，検証 : 19–22，テスト : 23–24 である．

比較手法：あらゆる手法がすべてのサブ予測タスクに適用できるわけではないため，各サブ予測タスクで異なる手法を適用して比較する．既知ノードに関するサブ予測タスクでは，「単純な方法」，「動的グラフ埋め込み」，「静的グラフに対する GNN」，「動的グラフに対する GNN」の 4 つのカテゴリーに分類された 9 つの手法を用いる．単純な方法とは，G_{T+1} の予測結果として，G_T をそのまま出力する手法（1）*Baseline*，予測対象をランダムにサンプリングする手法（2）*Random*，(3) *FNN*，(4) *LSTM* である．動的グラフ埋め込みとして，(5) *DynGEM*，静的グラフに対する GNN として，(6) *GCN* [17] を用いる．動的グラフに対する GNN としては，(7) *STGCN* [6]，(8) *EvolveGCN* [13]，(9) *TGGNN* [9] を用いる．EvolveGCN には，EvolveGCN-H と EveloveGCN-O の 2 つバージョンが存在する．

出現ノードの属性予測のために，PROSER と 3. 節で説明した 3 つのベースライン手法を使用する．出現ノードに接続されたリンク予測には，コサイン類似度（以降は CS と表記する），FNN，DEAL [11] を使用する．DEAL はこのタスクにおいて，state-of-the-art な手法である．AGATE は G_{T+1} の予測を繰り返し，検証データ上で各サブ予測タスクに最適なモデルを特定し，その結果を再利用する．

ハイパーパラメータ：訓練のイテレーション数は最大 1000，early stopping は 10〜100，バッチサイズは 1〜2，学習率は 0.01，Adam optimizer を用いてモデルを実行する．入力タイムステップ数 L は，NBA，Reddit，AMiner でそれぞれ 3，3，

(注3)：https://www.basketball-reference.com
(注4)：http://snap.stanford.edu/data/soc-RedditHyperlinks.html
(注5)：https://www.aminer.cn/citation

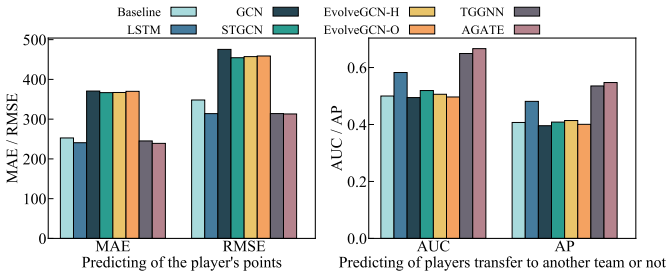


図 3　NBA データにおける属性予測の結果 [9].

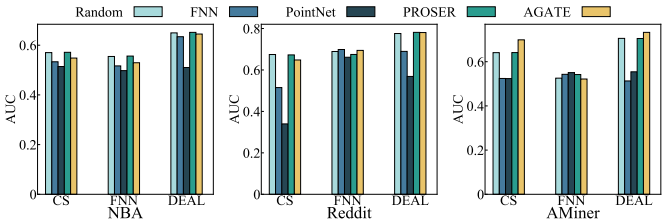


図 4　出現ノードに接続されるリンク予測の結果 [9].

表 1　消失ノード予測，出現リンク予測，消失リンク予測の結果 [9]．下線は Independent prediction における最高精度を表す．太字は Reuse prediction を含む全手法の中での最高精度を表す．AUC および Average precision は 100 掛けされている．

Methods	Node prediction (loss)				Link prediction (appearance)				Link prediction (disappearance)			
	ROC AUC		Average precision		ROC AUC		Average precision		ROC AUC		Average precision	
	NBA	Reddit	NBA	Reddit	NBA	Reddit	NBA	Reddit	NBA	Reddit	NBA	Reddit
Baseline	50.00±0.00	50.00±0.00	18.71±0.00	51.10±0.00	50.00±0.00	50.00±0.00	1.37±0.00	0.05±0.00	50.00±0.00	50.00±0.00	41.36±0.00	39.29±0.00
Random	50.16±1.10	49.75±1.49	18.77±0.34	50.99±0.75	49.86±0.05	50.51±0.75	1.36±0.00	0.05±0.00	49.91±0.40	50.87±0.53	41.32±0.19	39.71±0.26
FNN	85.35±0.06	77.53±0.03	53.51±0.08	74.53±0.04	49.66±0.43	46.31±1.86	1.34±0.02	0.04±0.01	49.65±0.17	48.10±0.57	41.09±0.15	36.77±0.87
LSTM	85.44±0.31	82.55±0.10	53.52±0.88	81.69±0.14	49.83±0.23	**83.07±0.35**	1.34±0.01	**6.05±0.42**	49.53±0.61	**72.13±0.17**	40.52±0.46	56.21±0.37
DynGEM	48.67±1.24	67.53±0.15	18.71±0.41	63.67±0.20	49.66±0.42	44.61±1.56	1.36±0.01	0.04±0.00	50.28±0.68	46.81±1.12	41.47±0.93	35.70±0.61
GCN	56.84±0.97	75.14±0.15	23.40±1.49	71.98±0.40	49.67±0.41	82.10±0.94	1.34±0.02	2.05±0.14	50.79±0.34	63.97±0.90	41.95±1.15	48.77±1.30
STGCN	60.20±0.95	82.22±0.31	26.71±0.15	80.45±0.43	49.71±0.49	73.59±1.52	1.34±0.02	0.12±0.01	49.31±0.57	63.01±2.8	40.45±0.51	48.81±2.05
EvolveGCN-H	56.33±1.18	63.44±1.86	21.97±1.03	59.49±1.37	50.01±1.04	52.21±4.46	**1.38±0.05**	0.08±0.02	49.19±0.58	50.62±0.41	40.47±0.78	39.64±0.15
EvolveGCN-O	56.69±0.58	66.32±2.71	23.01±0.63	62.25±3.08	49.76±0.58	52.43±1.52	1.37±0.03	0.05±0.01	48.83±1.56	51.73±0.68	40.78±1.44	41.07±0.08
TGGNN	85.59±0.58	82.67±0.19	54.57±2.14	81.59±0.42	50.09±0.43	77.93±4.11	1.35±0.02	0.24±0.20	49.81±0.47	66.36±3.91	40.64±0.43	50.80±4.30
AGATE	**98.19±2.60**	**93.49±3.62**	**92.00±11.4**	**93.92±3.51**	**50.36±0.73**	81.27±1.89	**1.38±0.02**	0.45±0.26	**51.34±0.01**	71.99±1.61	**43.38±0.21**	**56.39±2.18**

表 2　出現ノードの属性予測の結果 [9]．下線は Independent prediction における最高精度を表す．太字は Reuse prediction を含む全手法の中での最高精度を表す．

Methods	NBA				Reddit				AMiner			
	mean	median	min	max	mean	median	min	max	mean	median	min	max
Random	0.8143±0.00	0.8176±0.00	0.2055±0.00	**1.0000±0.00**	0.8317±0.00	0.8772±0.00	-0.2507±0.00	0.9947±0.00	0.9974±0.00	0.9983±0.00	0.9565±0.00	**0.9998±0.00**
FNN	0.7217±0.02	0.7650±0.02	0.2287±0.03	0.9004±0.02	0.7409±0.00	0.7876±0.01	-0.8723±0.01	0.9777±0.00	0.9904±0.00	0.9912±0.00	0.9779±0.00	0.9993±0.00
PointNet	0.6587±0.04	0.7075±0.03	0.2650±0.02	0.8337±0.04	0.6336±0.00	0.6016±0.00	-0.8697±0.00	0.9708±0.00	0.9915±0.00	0.9924±0.00	**0.9792±0.00**	0.9995±0.00
PROSER	0.8149±0.00	0.8164±0.00	0.2715±0.02	**1.0000±0.00**	0.8329±0.00	0.8780±0.00	**0.0286±0.00**	0.9947±0.00	**0.9975±0.00**	**0.9984±0.00**	0.9553±0.00	**0.9998±0.00**
AGATE	**0.8280±0.00**	**0.8416±0.00**	**0.3775±0.01**	0.9829±0.00	**0.8513±0.00**	**0.8904±0.00**	-0.0238±0.01	**0.9948±0.00**	0.9971±0.00	0.9979±0.00	0.9635±0.00	**0.9998±0.00**

5 である．式 (1) の $w(\cdot,\cdot)$ としてコサイン類似度を用いる．ハイパーパラメータは Independent prediction と Reuse prediction では独立に調整するため、チューニングについては，一般的なグラフ予測タスクと比較して難しくはない．学習時間は，NBA，Reddit，AMiner でそれぞれ約 500，700，200CPU 時間を要した．我々のフレームワークは並列実行が可能である．

5.1　予測精度に関する総合評価

Independent prediction と各サブ予測タスクの相互依存性を活用した AGATE の結果を示す．

Prediction on existing nodes: まず，既知ノードの属性予測の結果を示す．Reddit と AMiner については，ノード属性は静的であるため予測は行っていない．図 3 は NBA データにおける属性予測の結果であり，NBA 選手の得点予測（図の左）と他チームに移籍するか否かの二値分類（図の右）の結果を示している．得点予測は MAE と RMSE で評価しており，値が小さければ精度が良いことを表している．チーム移籍予測は AUC と Average Precision で評価しており，値が大きければ精度が良いことを表している．どちらの予測についても，LSTM と TGGNN は良好な結果を示しており，他手法はベースラインより予測精度が低い結果となった．AGATE は Reuse prediction によりさらに予測精度を高めている．得点予測では，AGATE は他全てのサブ予測タスクの結果を再利用し，LSTM で予測している．チーム移籍予測では，出現ノードに接続されるリンク予測の結果を再利用し，TGGNN で予測している．

次に，消失ノード予測，出現リンク予測，消失リンク予測の結果を示す．AMiner については，既知ノードに関する属性やグラフ構造は静的であるため予測は行っていない．表 1 は NBA と Reddit における消失ノード予測，出現リンク予測，消失リンク予測の結果を AUC と Average Precision で評価したものである．Independent prediction では，TGGNN と LSTM がしばしば最高性能を達成することが確認できる．この結果は，ノード属性やトポロジの変化が，将来のグラフ構造に影響を与えることを示している．AGATE は多くの場合で，Reuse prediction により予測精度を向上させている．特に，消失ノード予測の予測精度は，Independent prediction に比べて大幅に向上している．興味深いことに，消失ノード予測は出現ノードの属性予測の結果を再利用することで精度が向上する．詳細なサブ予測タスク間の相互依存性については 5.2 項で検証する．ここで，時間発展するグラフ上でのリンク予測はしばしば困難であるため，その AUC と Average Precision は非常に低いことに注意する（例えば，EvolveGCN の論文 [13] では異なるデータセット上で他の比較手法と同程度の MAP になっている）．全体的なリンク予測結果としては，AGATE が概ね最高性能を達成しており，タスクの相互依存性を捉えることで予測精度が向上することを再確認できた．

Prediction on new nodes：表 2 に出現ノードの属性予測

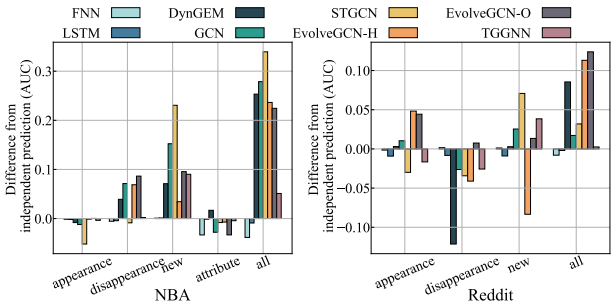


図 5　消失ノード予測に対する Independent prediction と Reuse prediction の AUC の差異 [9].

の結果を示す．決定されたマッチングにおけるノードペアのコサイン類似度の平均値，中央値，最小値，最大値を評価している．平均値は式（1）に等しく，マッチング類似度そのものを表している．全てのデータセットのマッチング類似度において，PROSER は 3 つのベースライン手法（Random，FNN，PointNet）の性能を上回っている．PROSER と Random は前のタイムステップに現れたノードの属性を変えずにそのまま出力するのに対し，FNN と PointNet はニューラルネットワークの算出値を出力している．この結果は，出現ノードの属性が，前のタイムステップの出現ノードの属性と類似していても，どのように変化するかを学習することが困難であることを示唆している．さらに，PROSER は，マッチング類似度の向上に寄与しないノードのサンプリングを回避するため，Random の性能を上回り，その結果，コサイン類似度は全体として大きく改善する．AGATE は，Independent prediction から得られる予測されたノードと真のノードの対応関係を用いた教師あり学習に基づく，Reuse prediction によって，マッチング類似度を改善する．Reuse prediction では，AGATE は出現ノードを既知ノードとして扱う（図 2 のコンポーネントの 15 番に該当）．

図 4 に出現ノードに接続されたリンクの予測結果を示す．凡例に示された手法による出現ノードの属性予測の結果に基づいて，横軸に示した手法でリンクを予測する．Random，PROSER，AGATE は出現ノードの属性予測において良好な結果を示している．リンク予測手法そのものについては，DEAL が他を圧倒していることが確認できる．

5.2　サブ予測タスク間の相互依存性に関する分析

AGATE が他のサブ予測タスクの結果を再利用することで，サブ予測タスクの精度をどのように向上させるかを検証する．図 5 は，消失ノード予測に対する Independent prediction と Reuse prediction の AUC の差を示している．横軸のタスク名は結果を再利用したサブ予測タスクを示しており，”all”は全サブ予測タスクの結果を再利用したことを示す．それぞれの手法とデータセットから，異なるタスクの相互依存性が明らかになった．例えば，STGCN の予測精度は NBA だと，全てのサブ予測タスクの結果を使用した場合に向上するが，Reddit ではわずかである．TGGNN の予測精度は出現ノードの予測結果を使用した場合に向上する．TGGNN は既に Independent prediction においても，高い精度で予測できているため，AUC 差は小さくなっている．それでも，この結果は，サブ予測タスク間の相

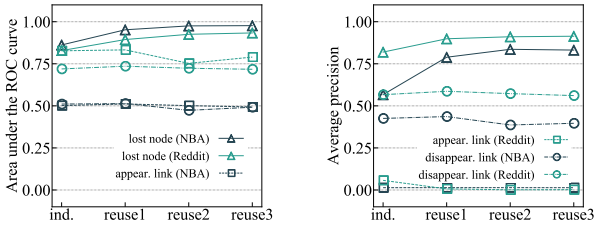


図 6　Reuse prediction の繰り返し回数と予測精度への影響 [9]

互依存性を活用することで，適切なサブ予測タスクの結果を再利用した場合に性能が向上するという我々の仮説を裏付けている．出現ノードの属性予測タスクは，これまで研究されていなかったが，他のサブ予測タスクに大きな影響を与えることが分かった．AGATE は検証データを用いて Reuse prediction で再利用するタスクを選択することができるため，性能に悪影響を与える再利用タスクを排除することができる．

Reuse prediction の繰り返し回数 i を変化させて，繰り返し回数と予測精度の関係を調べる．図 6 は消失ノード予測，出現リンク予測，消失リンク予測の予測精度と i の関係を示す．”ind.”は Independent prediction の精度である．また，Reuse prediction は全てのサブ予測タスクの結果を再利用した場合の精度となっている．”ind.”から”reuse1”までのほとんどの予測タスクにおいて，予測精度は目に見えて向上する．消失ノード予測については，再利用により精度は増加し続ける．他の予測タスクでは，誤差の蓄積により，”reuse1”以降は，精度は横ばいか低下することが分かる．

6.　関連研究

グラフ深層学習の手法は，(1) 静的埋め込み法 [18, 19], (2) 動的埋め込み法 [10, 20], (3) 静的 GNNs [11, 17, 21–24], (4) 動的 GNNs [6, 7, 13, 25–30] の 4 つに分類できる．埋め込み法は，対象の予測タスクとは切り離して分散表現を導出し，その分散表現を入力として対象の予測タスクごとにモデルを構築する．これに対し GNNs は，対象の予測タスクごとに直接モデルを構築する．グラフ予測手法には，静的なものと動的なものの 2 種類が存在する．既存研究では，対象となるグラフが静的であると仮定し，リンクや属性を予測するものが多い．しかし，実世界のシステムやデータは動的である．動的グラフに静的グラフの手法 [31] を適用することは可能かもしれないが，このアプローチは最適とは言えない [32]．動的グラフの学習手法が最近研究されているが，学習時に観測された要素のみを想定した Transductive なケースに限られているものが多い [6, 10]．このようなアプローチは，グラフに新しいノードやリンクが随時現れて時間発展する実世界の設定には不十分である．いくつかの手法 [13, 27] は Inductive な設定をサポートしており，学習時に観測されていない要素も対象としている．また，各手法は，グラフの特性や推論について独自の前提を置いており，すべての予測タスクに適しているとは限らない．

その他の類似タスク： グラフ生成 [33–36] やグラフ補完 [37] は，我々の問題とは似て非なるものである．これらは，属性のない動的グラフを生成したり，静的グラフに欠落している情報

を補うことを目的としている．

要約： 属性付き時間発展グラフに関する研究は数多くあるが，その全体を将来予測する研究は存在せず，また予測タスク間の相互依存性を調査した既存研究もない．本研究は，新しいノードの出現を含む，将来の属性付き時間発展グラフの全体予測と，予測タスク間の相互依存性を分析した最初の研究である．

7. ま と め

属性付き時間発展グラフの汎用予測フレームワーク AGATE と，出現ノードの属性予測モデル PROSER の 2 つの新しい手法を提案した．実験を通して，(1) 各サブ予測タスクが互いに影響し合うこと，(2) AGATE が各サブ予測タスクを独立に解く場合よりも高精度で予測できること，(3) PROSER が出現ノードの属性を高精度で予測できることを示した．

文 献

[1] Trapit Bansal, Da-Cheng Juan, Sujith Ravi, and Andrew McCallum. A2n: Attending to neighbors for knowledge graph inference. In *ACL*, pp. 4387–4392, 2019.

[2] Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. Semantic representation for dialogue modeling. In *ACL*, pp. 4430–4445, 2021.

[3] Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. Graph neural networks with generated parameters for relation extraction. In *ACL*, pp. 1331–1339, 2019.

[4] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. Graph neural networks for natural language processing: A survey. *arXiv*, 2021.

[5] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv*, 2020.

[6] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*, pp. 3634–3640, 2018.

[7] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

[8] Huiyuan Chen and Jing Li. Exploiting structural and temporal evolution in dynamic link prediction. In *CIKM*, pp. 427–436, 2018.

[9] Shohei Yamasaki, Yuya Sasaki, Panagiotis Karras, and Makoto Onizuka. Holistic prediction on a time-evolving attributed graph. In *ACL*, pp. 13676–13694, 2023.

[10] Palash Goyal, Nitin Kamra, Xinran He, and Yan Liu. Dyngem: Deep embedding method for dynamic graphs. *arXiv*, 2018.

[11] Yu Hao, Xin Cao, Yixiang Fang, Xike Xie, and Sibo Wang. Inductive link prediction for nodes having only attribute information. In *IJCAI*, pp. 1209–1215, 2020.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[13] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolving graph convolutional networks for dynamic graphs. In *AAAI*, pp. 5363–5370, 2020.

[14] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pp. 77–85, 2017.

[15] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *WWW*, pp. 933–943, 2018.

[16] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *KDD*, pp. 990–998, 2008.

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with GCNs. In *ICLR*, 2017.

[18] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. VERSE: versatile graph embeddings from similarity measures. In *WWW*, pp. 539–548, 2018.

[19] Anton Tsitsulin, Marina Munkhoeva, Davide Mottin, Panagiotis Karras, Ivan V. Oseledets, and Emmanuel Müller. FREDE: anytime graph embeddings. *Proc. VLDB Endow.*, Vol. 14, No. 6, pp. 1102–1110, 2021.

[20] Palash Goyal, Sujit Rokka Chhetri, and Arquimedes Canedo. dyngraph2vec: Capturing network dynamics using dynamic graph representation learning. *Knowledge-Based Systems*, Vol. 187, No. 104816, 2020.

[21] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence NNs. In *ICLR*, 2016.

[22] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[23] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, pp. 1024–1034, 2017.

[24] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *NeurIPS*, pp. 5165–5175, 2018.

[25] Jia Li, Zhichao Han, Hong Cheng, Jiao Su, Pengyun Wang, Jianfeng Zhang, and Lujia Pan. Predicting path failure in time-evolving graphs. In *KDD*, pp. 1279–1289, 2019.

[26] Dongkuan Xu, Wei Cheng, Dongsheng Luo, Xiao Liu, and Xiang Zhang. Spatio-temporal attentive RNN for node classification in temporal attributed graphs. In *IJCAI*, pp. 3947–3953, 2019.

[27] Aravind Sankar, Y. Wu, L. Gou, Wei Zhang, and H. Yang. Dynamic graph representation learning via self-attention networks. In *WSDM*, pp. 519–527, 2020.

[28] Dongkuan Xu, Junjie Liang, Wei Cheng, Hua Wei, Haifeng Chen, and Xiang Zhang. Transformer-style relational reasoning with dynamic memory updating for temporal network modeling. In *AAAI*, pp. 4546–4554, 2021.

[29] Jianfei Gao and Bruno Ribeiro. On the equivalence between temporal and static equivariant graph representations. In *ICML*, pp. 7052–7076, 2022.

[30] Dongqi Fu, Liri Fang, Ross Maciejewski, Vetle I Torvik, and Jingrui He. Meta-learned metrics over multi-evolution temporal graphs. In *KDD*, pp. 367–377, 2022.

[31] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *The JASIST*, Vol. 58, No. 7, pp. 1019–1031, 2007.

[32] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. *arXiv*, 2020.

[33] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *SIGKDD*, pp. 177–187, 2005.

[34] Changmin Wu, Giannis Nikolentzos, and Michalis Vazirgiannis. Evonet: A neural network for predicting the evolution of dynamic graphs. In *ICANN*, pp. 594–606, 2020.

[35] Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML*, pp. 5708–5717, 2018.

[36] Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. Netgan: Generating graphs via random walks. *arXiv*, 2018.

[37] Baoxu Shi and Tim Weninger. Open-world knowledge graph completion. In *AAAI*, pp. 1957–1964, 2018.

複数タスクを対象とした クラス知識のメタ学習にもとづくゼロショット学習

南部　優太†　　幸島　匡宏†　　岩田　具治††　　山本　隆二†

† 日本電信電話株式会社 NTT 人間情報研究所　〒 239–0847 神奈川県横須賀市光の丘 1-1
†† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所　〒 619–0237 京都府相楽郡精華町光台 2-4
E-mail: †{yuta.nambu,masahiro.kohjima,tomoharu.iwata,ryuji.yamamoto}@ntt.com

あらまし　一部のクラスのラベル付きデータが得られずとも分類モデルを学習可能にする手法としてゼロショット学習が広く利用されている．しかし，既存のゼロショット学習手法を適用するためには各クラスに関する事前知識が必要であり，このクラス知識が取得不可能または困難な問題に対して適用することはできない．そこで本研究では，クラス知識が利用不可能であってもゼロショット学習を適用可能とする手法を提案する．提案手法は，一部のクラスのラベル付きデータが得られない複数の分類タスクが与えられたとき，得られないクラスのデータを複数のタスクで補完しあうようにクラス知識と分類モデルを同時にメタ学習することで，クラス知識のない未知のタスクに対するゼロショット学習を可能とするものである．真のクラス知識が利用可能，クラス知識が利用不可能，クラス知識が外部のデータソースから提供される 3 つのデータセットを対象とした実験によって，提案手法がクラス知識を推定しゼロショット学習を適用可能にすることを示す．

キーワード　ゼロショット学習，メタ学習，クラス知識

1　背　　景

ラベル付きデータが得られないクラスが存在する分類問題の解決に有用なフレームワークとしてゼロショット学習がある [1,2]．このゼロショット学習は，一部のクラスのデータだけが利用可能であり，残りのクラスのデータが訓練データセットに含まれていない場合であっても，事前に与えられたクラス知識から不足した情報を補完することで分類モデルの学習を可能にしている．このクラス知識には，手動で定義したクラスの属性ベクトルや言語モデルが出力する分散表現など，他のモダリティから得られたベクトルの集合がよく用いられる．

このゼロショット学習の課題は，学習に適したクラス知識が常に得られるとは限らず，適用範囲が限られていることである．例えば，動物認識におけるゼロショット学習では，類似の動物を区別するために特定の部位の色や形状に関するクラス知識を必要とする．しかし，訓練時に現れていない未知の動物を網羅するクラス知識を定義することは現実的ではない．また，感情認識では，特定の感情を意図的に誘発することは難しいことから，一部の感情クラスのデータが欠けた設定下で個人ごとに分類モデルを学習したい場合がある．しかし，感情には個人差があるため，個人に適したクラス知識を定義すること自体が困難である．

そこで本論文では，クラス知識が利用できない場合でも適用可能なメタ学習を用いたゼロショット学習手法を提案する．提案手法は，複数のゼロショット学習タスクが与えられたとき，それらに対して分類モデルとクラス知識の学習の仕方を学習する．図 1 に従来のゼロショット学習と本論文の問題設定を示す．本論文の問題設定は，前述した個人の感情認識タスクに対するゼロショット学習に対応する．提案手法は，各個人では一部のクラスのデータが得られていなくとも，複数人のデータに対するメタ学習を用いてタスクに適したクラス知識を推定することで，各個人に特化した感情認識モデルを学習することができる．これにより，訓練タスクとは異なる未知のタスクが与えられたとき，推定したクラス知識から未知のタスクの未知のクラスの分類を可能にする．提案手法の有効性を検証するために，真のクラス知識が明らかな合成データセット，クラス知識が手動で定義されている動物認識データセット，クラス知識が与えられ

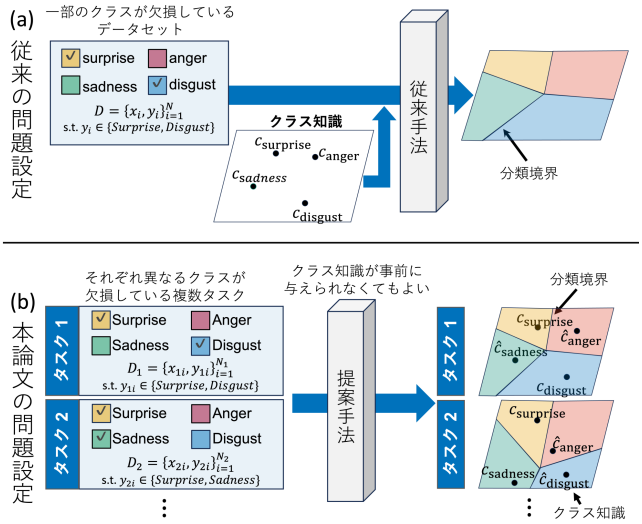


図 1: (a) 従来手法は，クラス知識事前に与えられることを仮定して単一のデータセットから分類モデルを学習する．(b) 提案手法は，クラス知識がなくとも複数のタスクでモデルを学習することで，クラス知識と分類モデルの両方を同時に学習する．

ていない感情認識データセットの 3 つのデータセットを用いて分類性能の評価実験を行う．本研究の貢献は次の 3 点である：

- 事前にクラス知識が与えられずとも，一部のクラスのデータが欠損したデータセットから分類モデルを学習する問題に取り組んだ．
- 複数のタスクを同時に解くことにより，クラス知識と分類モデルを同時にメタ学習する手法を提案した．
- 3 つの異なるデータセットを用いた評価実験により提案手法の有効性を確かめた．

以降，§2 では関連研究，§3 では提案手法を紹介し，§4・§5 では評価実験の結果と考察，§6 で本論文の結論を述べる．

2 関 連 研 究

事前に与えられるクラス知識によってゼロショット学習を実現するアプローチが多数提案されている．例えば，入力特徴を対応するクラスのベクトル表現と類似するように埋め込みモデルを訓練するアプローチ [3–5] や，クラス知識を別の空間に埋め込むアプローチ [6–8] が存在する．より最近のアプローチでは，未知のクラスの擬似インスタンスをクラス知識から生成することでゼロショット学習を通常の教師あり学習に帰着させている [9–11]．これらの手法は，クラス知識が事前に与えられなければ適用することはできず，クラス知識が与えられたとしても適用先のデータとの整合性は考慮していない．

ゼロショット学習と同時に適用先のデータに合わせてクラス知識を修正するアプローチ [12–14] も存在している．例えば，入力画像の説明文から意味表現を生成してクラス知識を修正する手法 [15, 16] や類似している他クラスの意味表現を混合して新たなクラス知識とする手法 [17, 18] もある．しかし，これらの手法もクラス知識が与えられることを前提に実現されており，クラス知識がないと適用できない．対して提案手法は，複数のタスクを対象としたメタ学習により，クラス知識が事前に与えられなくともゼロショット学習を適用可能にする．

フューショット学習におけるメタ学習の貢献から，ゼロショット学習においてもメタ学習が採用され始めている [19–21]．これらの研究では，データセットを分割することで複数の擬似タスクを生成し，擬似タスク全体に対してモデルを最適化することでメタ学習をゼロショット学習に適用している．適用されるメタ学習手法としては，タスク全体の損失関数の勾配を用いるアプローチ [22, 23] がよく用いられている．これらの従来手法はメタ学習を適用しているが，用いるデータは単一のタスクであり依然としてクラス知識が必要である．対して提案手法は，タスク特徴を条件としてモデル出力を学習させるメタ学習手法 [24] に触発されており，複数のタスクが与えられた場合にタスク固有のクラス知識と分類モデルを同時にメタ学習する．

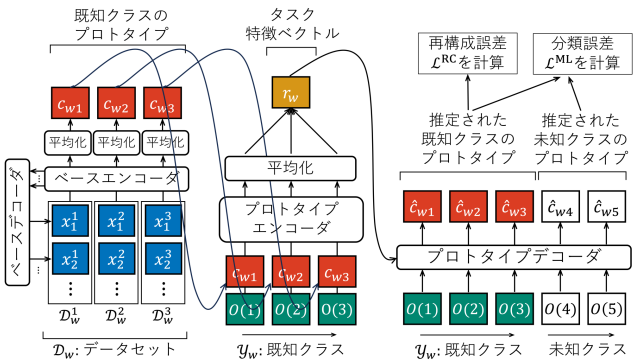


図 2: 提案手法のモデルアーキテクチャ．データセット $\mathcal{D}_w$ からタスク w のクラス知識 $\mathcal{C}$ とモデルパラメタを推定する．

3 提 案 手 法

3.1 問 題 設 定

訓練データとして W 個のタスクのデータセットの集合 $\mathcal{T} = \{\mathcal{D}_w\}_{w=1}^{W} = \{\{\boldsymbol{x}_{wi}, y_{wi}\}_{i=1}^{N_w}\}_{w=1}^{W}$ が与えられたとする．ここで，$\mathcal{D}_w$ は w 番目のタスクのデータセットであり，N_w 個の入力特徴 $\boldsymbol{x}_{wi}$ とラベル y_{wi} の組からなる．また，入力特徴 $\boldsymbol{x}_{wi}$ は入力特徴空間 $\mathcal{X}$ に対して $\boldsymbol{x}_{wi} \in \mathcal{X}$ を満たし，ラベル y_{wi} は各タスクごとに異なるラベル集合 $\mathcal{Y}_w$ に対して $y_{wi} \in \mathcal{Y}_w$ を満たす．ただし，各タスクのラベル集合 $\mathcal{Y}_w$ は，$\mathcal{Y} = \{1, \ldots, K\}$ に対して $\mathcal{Y}_w \subset \mathcal{Y}$，$\bigcup \mathcal{Y}_w = \mathcal{Y}$ を満たす．つまり，それぞれのタスクのラベル集合 $\mathcal{Y}_w$ は K 個のクラスを網羅しないが，全てのタスクのラベル集合の和は K 個のクラスを網羅する．以降，ラベル付きデータが事前に与えられているラベル集合 $\mathcal{Y}_w$ に属するクラスをタスク w の既知クラスと呼び，それ以外の $\mathcal{Y} \setminus \mathcal{Y}_w$ に属するクラスをタスク w の未知クラスと呼ぶ．

テスト時に，ある未知のタスクのデータセット $\mathcal{D}_{w^*} = \{(\boldsymbol{x}_{w^*i}, y_{w^*i})\}_{i=1}^{N_{w^*}}$ が与えられるとする．この未知のタスクは $w^* \notin \{1, \ldots, W\}$ であり，訓練時に与えられたタスクに含まれていない．そして，y_{w^*i} は同様にラベル集合 $\mathcal{Y}_{w^*}$ に属しており，ラベル集合 $\mathcal{Y}_w$ は K 個の全てのクラスを網羅しておらず，$\mathcal{Y}_{w^*} \subset \mathcal{Y}$ であるとする．さらに，従来のゼロショット学習手法とは異なりクラス知識は事前に与えられない．このとき，既知クラスだけでなく未知クラスも含めた全てのクラス集合 $\mathcal{Y}$ の中からこの未知のタスクの入力特徴を分類したい．

3.2 モ デ ル

提案手法のモデルは，ベースエンコーダ，ベースデコーダ，プロトタイプエンコーダ，プロトタイプデコーダの 4 つのニューラルネットワークのモジュールから構成される．図 2 にモデルのアーキテクチャを示す．このモデルは，データセット $\mathcal{D}_w$ から，クラス知識として各クラスのタスク固有のベクトル表現 $\mathcal{C}_w = \{\boldsymbol{c}_{wk}\}_{k=1}^{K}$ を推定する．以降，タスク w のクラス k の D 次元ベクトル表現をプロトタイプ $\boldsymbol{c}_{wk} \in \mathbb{R}^D$ と呼ぶ．簡単のため，ベースエンコーダとベースデコーダのモデルパラメタをまとめて θ と表記する．同様に，プロトタイプエンコーダとプロ

トタイプデコーダのモデルパラメタをまとめて η とする．また，これらのモデルのパラメタは全てのタスクで共通している．

まず，ベースエンコーダ $E_\theta : \mathcal{X} \to \mathbb{R}^D$ を用いて，データセット $\mathcal{D}_w$ から既知クラス $k' \in \mathcal{Y}_w$ のプロトタイプ $\boldsymbol{c}_{wk'} \in \mathbb{R}^D$ を計算する：

$$\boldsymbol{c}_{wk'} = \frac{1}{|\mathcal{D}_w^{k'}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_w^{k'}} E_\theta(\boldsymbol{x}). \tag{1}$$

ここで，$\mathcal{D}_w^{k'}$ は，$\mathcal{D}_w$ のうち既知クラス k' でラベル付けされたインスタンスの集合である．式 (1) は，各クラスごとにベースエンコーダが出力する埋め込みベクトルの平均を計算することで，インスタンスの数や順序に影響されずにデータからプロトタイプを決定できる．また，ベースエンコーダ E_θ の出力である埋め込みベクトルが入力特徴を再構成するのに十分な情報を持つように，埋め込みベクトルから入力特徴を再構成するベースデコーダ $E_\theta^{-1} : \mathbb{R}^D \to \mathcal{X}$ を導入する．ベースデコーダは推論には使われないが，Section 3.3.1 における再構成誤差の計算に用いられる．

次に，プロトタイプエンコーダ $G_\eta : \mathbb{R}^D \times \{0,1\}^K \to \mathbb{R}^L$ は，既知クラスのプロトタイプとインデックスから L 次元のタスク特徴ベクトル $\boldsymbol{r}_w \in \mathbb{R}^L$ を抽出する．

$$\boldsymbol{r}_w = \frac{1}{|\mathcal{Y}_w|} \sum\nolimits_{k'\in\mathcal{Y}_w} G_\eta(\boldsymbol{c}_{wk'}, O(k')). \tag{2}$$

ここで，$O(\cdot)$ はラベルをワンホットエンコーディングする関数 $O : \mathcal{Y} \to \{0,1\}^K$ である．この操作は，既知クラスのプロトタイプがそれぞれのタスクを表現しているという考えにもとづいており，これによりタスク特徴ベクトル $\boldsymbol{r}_w$ が得られる．

そして，タスク特徴ベクトル $\boldsymbol{r}_w$ とクラスのインデックス $k \in \mathcal{Y}$ を用いて，プロトタイプデコーダ $G_\eta^{-1} : \{0,1\}^K \times \mathbb{R}^L \to \mathbb{R}^D$ によりプロトタイプを推定する：

$$\hat{\boldsymbol{c}}_{wk} = G_\eta^{-1}(O(k), \boldsymbol{r}_w).$$

最後に，w 番目のタスクから入力特徴 $\boldsymbol{x}$ が与えられたとき，埋め込みベクトル $E_\theta(\boldsymbol{x})$ は対応するクラスのプロトタイプに近しいと予想されるので，

$$\hat{y} = \arg\max_{\ell\in\mathcal{Y}} \frac{\exp(-\|E_\theta(\boldsymbol{x}) - \hat{\boldsymbol{c}}_{w\ell}\|^2)}{\sum_{k=1}^K \exp(-\|E_\theta(\boldsymbol{x}) - \hat{\boldsymbol{c}}_{wk}\|^2)}, \tag{3}$$

とすることで予測結果 $\hat{y}$ が得られる．

3.3 訓　　練

タスク集合 $\mathcal{T}$ が与えられたとき，提案法は 2 つのフェーズを交互に繰り返すことでモデルパラメタを推定する．1 つ目のベースモデル訓練フェーズでは，既知クラスの推定性能を向上させるため，ベースエンコーダ E_θ とベースデコーダ E_θ^{-1} を訓練する．2 つ目のメタモデル訓練フェーズでは，未知クラスの推定性能を向上させるようなクラス知識を推定可能にするため，プロトタイプエンコーダ G_ρ とプロトタイプデコーダ G_ρ^{-1} を訓練する．訓練全体の擬似コードをアルゴリズム 1 に示す．

既知クラスしか与えられない設定下であるが，様々な既知クラスと未知クラスの組み合わせが与えられる状況をシミュレートするため，各タスクごとにデータセットを擬似既知クラスと擬似未知クラスへとランダムに分割する．具体的には，既知クラス $\mathcal{Y}_w$ をランダムに 2 つの排他的な集合に分割し，一方を擬似既知クラス集合 $\mathcal{Y}_w^{\mathrm{ps}}$，他方を擬似未知クラス集合 $\mathcal{Y}_w^{\mathrm{pu}}$ とする．また，それぞれの擬似クラス集合にもとづいて $\mathcal{D}_w$ を分割し，$\mathcal{D}_w^{\mathrm{ps}} : \mathbb{R}^D \times \mathcal{Y}_w^{\mathrm{ps}}$ をベースモデル訓練フェーズで用いる．また，$\mathcal{D}_w^{\mathrm{pu}} : \mathbb{R}^D \times \mathcal{Y}_w^{\mathrm{pu}}$ をメタモデル訓練フェーズに用いる．これらは，$\mathcal{D}_w^{\mathrm{ps}} \cup \mathcal{D}_w^{\mathrm{pu}} = \mathcal{D}_w$ を満たす．

3.3.1 ベースモデル訓練フェーズ

ベースモデル訓練フェーズでは，さらに擬似既知クラスのデータセットを訓練エポックの度にランダムに二等分し，一方をプロトタイプの算出に用い，もう一方を損失の計算に用いる．プロトタイプの算出に用いるサブセットをサポートセット $\mathcal{D}_w^{\mathrm{spt}}$ と呼び，損失の計算に用いるサブセットをクエリセット $\mathcal{D}_w^{\mathrm{qry}}$ と呼ぶ．このとき，クエリセット $\mathcal{D}_w^{\mathrm{qry}}$ から計算される以下の損失を最小化することによりモデルのパラメータ θ を推定する：

$$\hat{\theta} = \operatorname*{argmin}_\theta \mathbb{E}_{\mathcal{D}_w\sim\mathcal{T}} \left[\mathbb{E}_{\mathcal{D}_w^{\mathrm{ps}}\sim\mathcal{D}_w} \left[\mathbb{E}_{\mathcal{D}_w^{\mathrm{qry}}\sim\mathcal{D}_w^{\mathrm{ps}}} \left[\mathcal{J}(\theta; \mathcal{D}_w^{\mathrm{qry}})\right]\right]\right]. \tag{4}$$

ここで，$\mathbb{E}$ は期待値であり，

$$\mathcal{J}(\theta; \mathcal{D}_w^{\mathrm{qry}}) = -\sum_{(\boldsymbol{x},y)\in\mathcal{D}_w^{\mathrm{qry}}} \Big(\log p_\theta(y \mid \boldsymbol{x}) - \beta \mathcal{J}^{\mathrm{RC}}(\theta; \boldsymbol{x})\Big), \tag{5}$$

は損失関数である．また，β ($\geqq 0$) はハイパーパラメータである．$p_\theta(y \mid \boldsymbol{x})$ は，埋め込まれた入力特徴と各クラスのプロトタイプ間のユークリッド距離の 2 乗によって計算される：

$$p_\theta(y = \ell \mid \boldsymbol{x}) = \frac{\exp(-\|E_\theta(\boldsymbol{x}) - \boldsymbol{c}_{w\ell}\|^2)}{\sum_{k'\in\mathcal{Y}_w^{\mathrm{ps}}} \exp(-\|E_\theta(\boldsymbol{x}) - \boldsymbol{c}_{wk'}\|^2)}. \tag{6}$$

ここでの $\boldsymbol{c}_{wk'}$ は，式 (1) に従いサポートセット $\mathcal{D}_w^{\mathrm{spt}}$ によって，$\mathcal{C}_w^{\mathrm{ps}} = \{\boldsymbol{c}_{wk'}; k' \in \mathcal{Y}_w^{\mathrm{ps}}\}$ として事前に計算される．式 (5) の第 2 項は L2 ノルムによる再構成誤差である：

$$\mathcal{J}^{\mathrm{RC}}(\theta; \boldsymbol{x}) = \|E_\theta^{-1}(E_\theta(\boldsymbol{x})) - \boldsymbol{x}\|^2.$$

この再構成誤差を小さくすることで，入力特徴 $\boldsymbol{x}$ の情報を保持するようにベースエンコーダ $E_\theta(\boldsymbol{x})$ を学習させる．

3.3.2 メタモデル訓練フェーズ

メタモデル訓練フェーズでは，プロトタイプエンコーダとプロトタイプデコーダを訓練し，擬似既知クラスのプロトタイプから擬似未知クラスのプロトタイプを推定可能にする．プロトタイプエンコーダ G_η は擬似既知クラスのプロトタイプからタスク特徴ベクトル $\boldsymbol{r}_w$ を抽出できる必要があり，プロトタイプデコーダ G_η^{-1} はタスク特徴ベクトルに従って適切なプロトタイプを推定できる必要がある．よって，以下の損失関数を最小化することによりパラメタ η を推定する：

$$\hat{\eta} = \operatorname*{argmin}_\eta \mathbb{E}_{\mathcal{D}_w\sim\mathcal{T}} \left[\mathcal{L}(\eta; \mathcal{D}_w, \theta, \mathcal{C}_w^{\mathrm{ps}})\right]. \tag{7}$$

Algorithm 1 モデルの訓練手順．RandomSample$(\mathcal{A}, n)$ は，集合 $\mathcal{A}$ から重複なく n 個の要素をランダムにサンプルする．

Input: タスク集合 $\mathcal{T}$
Opt: 擬似既知クラス数 N_{s}, インスタンス数 N_{in}，バッチサイズ B
Output: 学習済みモデルパラメタ θ, η

while end condition is satisfied **do**
  $\mathcal{W} \leftarrow$RandomSample$(\{1, \ldots, W\}, B)$
  for $w \in \mathcal{W}$ **do**
    $J \leftarrow 0$
    $\mathcal{Y}_w^{\mathrm{ps}} \leftarrow$RandomSample$(\mathcal{Y}_w, N_{\mathrm{s}})$
    $\mathcal{D}_w^{\mathrm{ps}} \leftarrow$RandomSample$(\mathcal{D}_w, N_{\mathrm{in}})$ s.t. $y_{wi} \in \mathcal{Y}_w^{\mathrm{ps}}$
    while end condition is satisfied **do**
      $\mathcal{D}_w^{\mathrm{spt}} \leftarrow$RandomSample$(\mathcal{D}_w^{\mathrm{ps}}, |\mathcal{D}_w^{\mathrm{ps}}|/2)$
      $\mathcal{D}_w^{\mathrm{qry}} \leftarrow \mathcal{D}_w^{\mathrm{ps}} \setminus \mathcal{D}_w^{\mathrm{spt}}$
      Calculate $\mathcal{C}_w^{\mathrm{ps}}$ by Eq. (1) using $\mathcal{D}_w^{\mathrm{spt}}$
      $J \leftarrow J + \mathcal{J}(\theta; \mathcal{D}_w^{\mathrm{qry}})$
      Update θ using J by Eq. (4)
    end while
  end for
  $L \leftarrow 0$
  for $w \in \mathcal{W}$ **do**
    $L \leftarrow L + \mathcal{L}(\eta; \mathcal{D}_w, \theta, \mathcal{C}_w^{\mathrm{ps}})$
  end for
  Update η using L by Eq. (7)
end while

損失関数 $\mathcal{L}(\eta; \mathcal{D}_w, \theta, \mathcal{C}_w^{\mathrm{ps}})$ は 2 つの項からなる：

$$\mathcal{L}(\eta; \mathcal{D}_w, \theta, \mathcal{C}_w^{\mathrm{ps}}) = \mathcal{L}^{\mathrm{ML}}(\eta; \mathcal{D}_w, \theta) + \alpha \mathcal{L}^{\mathrm{RC}}(\eta; \mathcal{C}_w^{\mathrm{ps}}). \quad (8)$$

α ($\geqq 0$) はハイパーパラメタである．第 1 項は，推定されたクラス k のプロトタイプ $G_\eta^{-1}(O(k), \boldsymbol{r}_w)$ を用いた，擬似未知クラスを含む全ての既知クラスに対応するインスタンスを分類する際の損失である：

$$\mathcal{L}^{\mathrm{ML}}(\eta; \mathcal{D}_w, \theta) = -\sum_{(\boldsymbol{x}, y) \in \mathcal{D}_w} \log p_\eta(y \mid E_\theta(\boldsymbol{x})),$$

$$p_\eta(y = \ell \mid E_\theta(\boldsymbol{x})) = \frac{\exp(-\|E_\theta(\boldsymbol{x}) - G_\eta^{-1}(O(\ell), \boldsymbol{r}_w)\|^2)}{\sum_{k=1}^{K} \exp(-\|E_\theta(\boldsymbol{x}) - G_\eta^{-1}(O(k), \boldsymbol{r}_w)\|^2)}.$$

第 2 項は，擬似既知クラスのプロトタイプの再構成誤差である：

$$\mathcal{L}^{\mathrm{RC}}(\eta; \mathcal{C}_w^{\mathrm{ps}}) = \sum_{k' \in \mathcal{Y}_w^{\mathrm{ps}}} \|G_\eta^{-1}(O(k'), \boldsymbol{r}_w) - \boldsymbol{c}_{wk'}\|^2.$$

式 (7) をタスク集合 $\mathcal{T}$ 全体で最適化することで，プロトタイプエンコーダとプロトタイプデコーダはプロトタイプを推定可能となる．

4 実　　験

提案手法の性能や振る舞いを検証するため，3 つのデータセットに対してそれぞれ異なる実験を実施した．実験に用いたデータセットの概要を表 1 に示す．1 つ目は，我々が定めたクラス知識から生成した合成データセットであり，真のクラス知識が明らかである．2 つ目は，ゼロショット学習手法の評価に用いられる Animal with Attributes 2 データセット (AwA2) [25] であり，クラス知識として手動で定義された各クラスの属性ベクトルが事前に与えられている．3 つ目は，教師あり学習手法の評価に用いられる DREAMER データセット [26] であり，生体信号と感情ラベルからなる複数のタスクで構成されているが，それぞれのタスクのクラス知識は与えられない．これらのデータセットを用いて，それぞれ以下の項目を検証した．

表 1: 3 つのデータセットにおける統計量とデータ分割の設定．入力特徴空間 $\mathcal{X}$ の次元数，データセットが提供するクラス知識の次元数 D，$\mathcal{Y}$ のクラス数を示した．また，$\mathcal{Y}_w^{\mathrm{ps}}, \mathcal{Y}_w^{\mathrm{pu}}$ はそれぞれ擬似既知クラス数と擬似未知クラス数であり，未知クラス数と擬似未知クラス数が同一となるよう設定した．

Dataset	$\mathcal{X}$	D	$\mathcal{Y}$	$\mathcal{Y}_w$	$\mathcal{Y}_w^{\mathrm{ps}}$	$\mathcal{Y}_w^{\mathrm{pu}}$
Synthetic	2	2	25	20	15	5
AwA2	2,048	85	50	40	30	10
DREAMER	101	-	9	7	5	2

実験 1　複数のタスク間でクラス知識のばらつきを大きくしたとき，提案手法はどの程度適応できるか？

実験 2　適したクラス知識が与えられるとは限らないとき，提案手法は従来のゼロショット学習手法よりも高い性能を得られるか？

実験 3　提案手法であるクラス知識の与えられないゼロショット学習で教師あり学習にどの程度近づけるか？

4.1 データセット

合成データセット： 実験 1 のため，2 次元の入力特徴空間と 2 次元のクラス知識空間を持つ合成データセットを生成した．この合成データセットは，我々がタスクごとにランダムに定義したクラスタ中心 (=プロトタイプ) を平均とした正規分布によって生成したデータセットのため，真のクラス知識が明らかである．また，それぞれのタスクのプロトタイプに対して重み λ をかけたノイズを加えることで，$\lambda = 0$ のとき全てのタスクのクラス知識は同一であり，λ の値が大きくなるほどタスク間のクラス知識のばらつきが大きくなる設定をシミュレートした．

生成手順は，次の 5 つのステップであった．(i) $25 (= K)$ クラス分の $2 (= L)$ 次元ベクトルを一様分布 $[-3, 3]$ からサンプリングし，クラス知識の行列表現 $\mathcal{C} \in \mathbb{R}^{K \times L}$ とした．(ii) タスクごとに一様分布 $[-1, 1]$ からサンプリングしたノイズ行列 $Q_w \in \mathbb{R}^{K \times L}$ を定義し，タスクごとに異なる真のクラス知識を $\mathcal{C}_w = \mathcal{C} + \lambda Q_w$ で定めた．この重み λ をタスク変動係数と呼ぶ．(iii) ラベル集合 $\{1, \ldots, 25\}$ から，タスクごとに 20 個のクラスを重複なしでランダムに選択し，既知クラス $\mathcal{Y}_w$ とした．(iv) 既知クラス $k' \in \mathcal{Y}_w$ ごとに，平均を $\mathcal{C}_w$ の k' 番目のベクトル，1 を分散とする多変量正規分布から 10 組の入力特徴を生成し，学習データセット $\mathcal{D}_w$ とした．(v) テストデータセットとして，iv と同様の手順で既知クラスのデータセット $\mathcal{D}'_{w^*}$ を生成し，既知クラスと未知クラスの双方を評価するために $\mathcal{Y}_w$ だけでなく $\mathcal{Y}$ 全体を対象として同様に 10 組の入力特徴を生成し

たものを $\mathcal{D}^*_{w^*}$ とした．これらの手順により，訓練タスク集合 $\{\mathcal{D}_w\}_{w=1}^{450}$ とテストタスク集合 $\{(\mathcal{D}'_{w^*}, \mathcal{D}^*_{w^*})\}_{w^*=1}^{50}$ を生成した．

AwA2 データセット： ゼロショット学習手法の評価に広く使われているデータセットの中から，Animals with Attributes 2(AwA2) [25] データセットを用いた．AwA2 は，50 クラスの動物の画像とそのラベルから構成される．また，このデータセットでは，手動で定義された 85 次元の属性ベクトルがクラス知識として与えられている．AwA2 の入力特徴として，先行研究 [25, 27] と同様に訓練済みの ResNet101 の中間層の出力である 2,048 次元の特徴ベクトルを用いた．このデータセットではクラスごとに全インスタンスを半分に分割し，片方を訓練データセット，もう片方をテストデータセットとすることで，訓練時のタスクとテスト時のタスクで同じインスタンスが共有されないようにした．そして，訓練データセットからランダムに 40 のクラスをサンプリングして既知クラスとし，クラス毎に 10 インスタンスをサンプリングした．この操作を 450 回繰り返すことで，450 個の訓練タスク集合 $\{\mathcal{D}_w\}_{w=1}^{450}$ を生成した．同じ操作でテストデータセットから既知クラスのデータセット $\mathcal{D}'_{w^*}$ を生成し，既知クラスと未知クラスの双方を評価するために 50 クラス全てに対して 10 インスタンス をサンプルし $\mathcal{D}^*_{w^*}$ を生成することで，テストタスク集合 $\{(\mathcal{D}'_{w^*}, \mathcal{D}^*_{w^*})\}_{w^*=1}^{50}$ を得た．

DREAMER データセット： 感情認識はクラスの解釈が個人によって異なり，クラス知識を定義することが困難なタスクの 1 つである．そこで，代表的な感情認識データセットの一つである DREAMER データセット [26] を用いて，生体信号からのゼロショット感情認識の性能を評価した．このデータセットでは，感情を誘発する 18 のビデオ（＝感情 9 種類× 2 クリップ）を 23 人の実験参加者が視聴したときの脳波と心電位が記録されている．感情の種類は，amusement, excitement, happiness, calmness, anger, disgust, fear, sadness, surprise である．脳波と心電位のデータから特徴量を抽出するために，これらのデータをウィンドウ幅 60 秒，ストライド 30 秒のスライディングウィンドウで分割し，NeuroKit2 [28] を用いて 102 次元の特徴量を抽出した．NeuroKit2 が出力した特徴量は 125 次元であったが，元データの異常により正しく計算できない特徴量が存在したため，この特徴量を削除して 102 次元とした．その結果，実験参加者 23 名それぞれから 9 クラス 99 インスタンスのラベル付きデータが得られた．各参加者ごとに 9 つのクラスをランダムに分割し，7 つの既知クラスと 2 つの未知クラスとした．既知クラスのインスタンスの半分を既知クラスのデータセット $\mathcal{D}'_{w^*}$ として使用し，残りの半分をテストデータセット $\mathcal{D}^*_{w^*}$ として使用した．また，未知クラスの全てのインスタンスはテストデータセット $\mathcal{D}^*_{w^*}$ に含めた．この操作により，23 個のタスク集合 $\{(\mathcal{D}'_{w^*}, \mathcal{D}^*_{w^*})\}_{w^*=1}^{23}$ が得られた．

4.2 実験設定

提案手法は，一部のクラスのデータが欠けた未知のタスクが与えられたとき，それらのインスタンスの既知クラスと未知クラスへの分類性能によって評価される．未知クラスに対応する入力特徴のみが与えられ，未知クラス間の分類性能のみを評価する場合をここでは Traditional ゼロショット学習と呼ぶ．対して，既知クラスと未知クラスのいずれかに対応する入力特徴が与えられ，全てのクラス間での分類性能が評価される場合を Generalized ゼロショット学習と呼ぶ．つまり，式 (3) の ℓ は，Traditional ゼロショット学習では $\ell \in \mathcal{Y}_{w^*}$ を満たし，Generalized ゼロショット学習では $\ell \in \mathcal{Y}$ を満たす．これにもとづいて，実験では Traditional ゼロショット学習と Generalized ゼロショット学習の 2 つの設定で性能を評価した．

評価指標は，ゼロショット学習の先行研究を踏まえ $\mathbf{T}$, $\mathbf{u}$, $\mathbf{s}$, $\mathbf{H}$ を用いた．$\mathbf{T}$ は，未知クラスに対応する入力特徴の未知クラス間の正解率，つまり Traditional ゼロショット学習の評価指標である．$\mathbf{u}$,$\mathbf{s}$ はそれぞれ，未知クラスに対応する入力特徴の全クラス内の正解率と既知クラスに対応する入力特徴の全クラス内の正解率であり，それらの調和平均が $\mathbf{H}$ である．つまりこれら 3 つの指標は Generalized ゼロショット学習の評価指標に対応する．

実験 1 合成データセットのタスク変動係数 λ を大きくし，タスク間のクラス知識のばらつきを大きくした場合の提案手法の性能を調べた．問題自体の難しさを評価するため，提案するメタ学習のアプローチを含まないベースエンコーダとベースデコーダのみを式 (4) で最適化したゼロショット学習を比較対象とした．このベースエンコーダとベースデコーダの学習には，真のクラス知識として $\mathcal{C}$ と $\mathcal{C}_w$ を用いた．真のクラス知識が与えられるため，このゼロショット学習は理想的なクラス知識で行われた．以降，$\mathcal{C}$ を ‘真のクラス知識’，$\mathcal{C}_w$ を ‘真のタスククラス知識’ と呼ぶ．

実験 2 一部のベンチマークデータセットではゼロショット学習に有用なクラス知識が事前に与えられるが，現実的には常にゼロショット学習に適したクラス知識が利用できるとは限らない．そこで，既存の 5 つの埋め込みベースのゼロショット学習手法と提案手法を不正確なクラス知識が提供される設定において比較した．具体的には，AwA2 データセットが提供するクラス知識の次元 D を削減することで，クラス知識が不正確な場合のゼロショット学習をシミュレートした．従来手法の適用時には，提供されるクラス知識の次元数をランダムに $\{85, 64, 32, 16, 8, 4, 2\}$ に削減し，その性能を評価した．提案手法ではクラス知識の次元数を任意に設定できるため，従来手法と同じ次元数に設定して比較した．比較対象は，ESZSL [4], SAE [5], ALE [6], Prototypical Networks [29], JIL [30] とした．なお，これらの従来手法を 4.1 で定義した 50 個のテストタスクに適用し，その結果の平均値を算出した．さらに，式 (5) と (8) における再構成誤差の影響を調べるため，$\mathcal{L}^{\mathrm{RC}}$ と $\mathcal{J}^{\mathrm{RC}}$ を用いない場合の性能も評価した．

実験 3 クラス知識の定義が難しくデータの欠損が起きやすい感情認識タスク用のデータセット DREAMER を用いて，クラス知識が与えられない設定における提案手法の性能を評価した．クラス知識が与えられなければ従来のゼロショット学習手法は適用できないため，教師あり学習手法と提案手法を比較することでゼロショット学習という不利な設定において提案手法が教師あり学習にどの程度近づけるか検証した．DREAMER

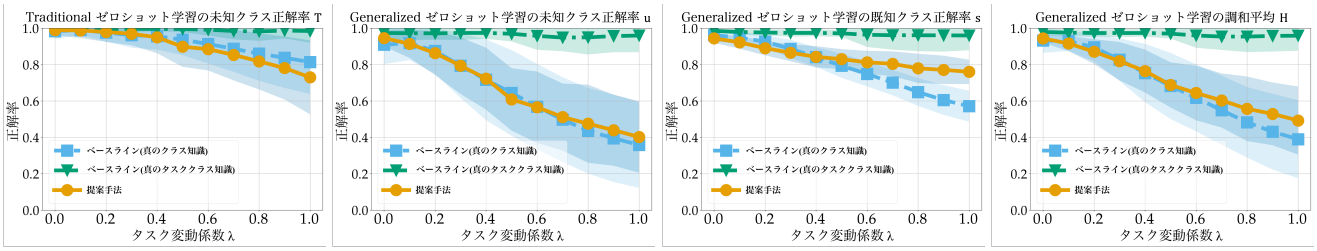


図 3: 合成データセットにおけるタスク変動係数 λ を用いた提案手法と従来ゼロショット学習手法の性能評価．横軸は λ，縦軸はテストタスクの平均正解率，各評価点におけるタスク間の精度の標準偏差を塗りつぶしで示した．

表 2: 実験に使用したハイパーパラメタ．すべてのデータセットにおいて，エポック数は 200，学習率は 0.01 とした．

Dataset	Batch size	L	α	β
Synthetic	64	32	1.0	10.00
AwA2	64	64	0.1	0.01
DREAMER	8	16	0.1	10.00

データセットのタスク数（すなわち実験参加者数）は 23 人と少ないため，leave-one-out cross validation を実施した．また，式 (5) と (8) における再構成誤差の影響を調べるため，$\mathcal{L}^{\mathrm{RC}}$ と $\mathcal{J}^{\mathrm{RC}}$ を用いない場合の性能も評価した．

教師あり学習では，モデルのパラメタ数を提案手法に近づけるため，提案手法で用いたベースエンコーダに出力層のユニット数がクラス数と等しい全結合層を追加した MLP を用いて，交差エントロピー損失による学習を行った．感情や生体信号は個人差が大きいと考えられるため，全実験参加者のデータで共通モデルを訓練する MLP (common) とテスト対象である実験参加者 1 名のデータだけで個人モデルを訓練する MLP (individual) の 2 パターンで教師あり学習を行った．MLP (common) は，one-holdout 検証におけるテスト対象者 1 人を除く全ての実験参加者のデータとテスト対象者の既知クラスのデータを用いた教師あり学習であり，訓練データは提案手法と同一とした．一方，MLP (individual) は，個人差の影響を排除するためテスト対象者 1 人の既知クラスのデータのみを用いて教師あり学習した．

4.3 実　　装

ベースエンコーダ，ベースデコーダ，プロトタイプエンコーダ，プロトタイプデコーダは，2 つの全結合層と ReLU 層で構成した．全結合層のユニット数は，いずれのモデルでも入力次元数と出力次元数の和を 2 で割ったものとした．ただし，合成データセットでは，入力次元と出力次元が小さい（2 次元）ため，ユニット数は 64 とした．モデルのパラメータは Adam [31] で最適化した．表 2 に実験で用いたハイパーパラメータを示す．PyTorch を使用し，実験は Intel(R) Core(TM) i9-13900KF 3.00GHz CPU と NVIDIA GeForce RTX4090 GPU を搭載したコンピュータで実行した．

5　結果と考察

5.1 実　験　1

図 3 はタスク変動係数 λ の変化に対する提案手法と通常のゼロショット学習手法の性能変化を示している．図 3 の結果から，タスク変動係数が小さいときは真のタスククラス知識が与えられるゼロショット学習と同等の性能を示し，タスク変動係数が大きくなっても提案手法は真のクラス知識を用いたゼロショット学習よりも高い性能が得られた．また，提案手法はクラス知識を全く用いていないにも関わらず，真のタスククラス知識を用いたゼロショット学習と真のクラス知識を用いたゼロショット学習の中間の性能を持つことが分かった．提案手法は，未知クラスのプロトタイプを既知クラスと混同しないようにクラス知識を学習することで，既知クラスと未知クラスの分類性能を向上させたと考えられる．これらの結果から提案手法は，タスクのばらつきが小さいときは真のクラス知識が明らかなゼロショット学習と同様の性能が得られ，ばらつきが大きくなっても共通のクラス知識が 1 つ与えられる従来のゼロショット学習手法よりも性能が高いことが示唆された．

5.2 実　験　2

実験 2 の結果を図 4 に示した．この結果から，従来手法はクラス知識が削減され不正確になるにつれて性能が低下するが，提案手法は高い正解率を維持できることが分かった．また，従来のゼロショット学習は事前に提供されるクラス知識に依存しており，適切なクラス知識が提供されるという仮定が満たされた場合にのみ良い性能を発揮できた．これに対して提案手法は，定義されたクラス知識空間内で適切なクラス関係を学習するため，外的要因に影響されることなく，従来手法よりも高い正解率を達成できた．また，表 3 にデータセットで定義されたクラス知識空間の次元を 85 次元に設定した場合の従来手法の正解率と再構成誤差の有無を変更した場合の提案手法の正解率を示した．この結果から，再構成誤差を導入することで既知クラスの分類性能を低下させることなく，未知クラスの分類性能を向上できることが分かった．

推定されたクラス知識の例を図 5 に示した．この図は，訓練済みモデルによって推定された 85 次元のクラス知識を，UMAP [32] によって 2 次元に削減して可視化したものである．提案手法はクラスの意味表現を与えていないにも関わらず，四

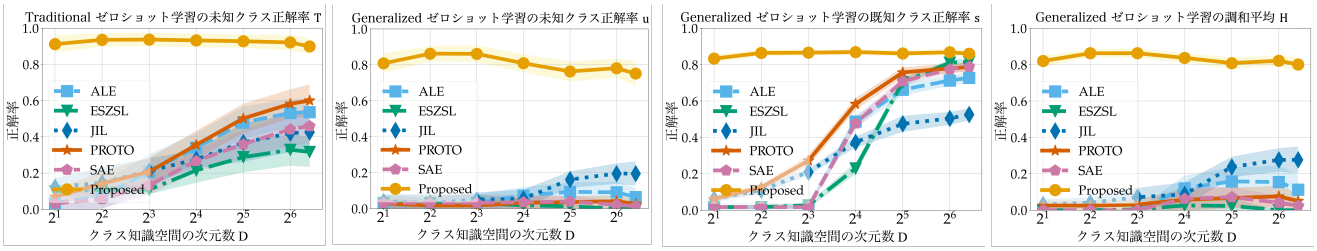


図 4: AwA2 データセットにおけるクラス知識空間の次元数 D に対する提案手法と従来手法の正解率．横軸は次元数 D，縦軸はテストタスク集合に対する平均正解率，塗りつぶしは各評価点におけるタスク間の正解率の標準偏差である．D を小さくすることは，従来手法ではデータセットが提供するクラス知識の次元のランダムな削減を意味するが，提案手法ではクラス知識空間の次元を決定するハイパーパラメータの変更を意味することに注意されたい．

表 3: AwA2 データセットにおける Traditional ゼロショット学習の未知クラスの正解率 (**T**)，Generalized ゼロショット学習における未知クラス正解率 (**u**)，既知クラス正解率 (**s**)，およびそれらの調和平均 (**H**)．各列ごとに最大値は太字で示した．

Method	AwA2 (%)			
	T	**u**	**s**	**H**
ESZSL [4]	31.7	0.0	81.5	0.0
SAE [6]	45.6	1.3	78.9	2.5
ALE [5]	53.7	6.4	72.5	11.2
PROTO. NETS [29]	60.4	2.3	78.7	5.1
JIL [30]	42.7	1.6	52.3	26.7
Proposal	**93.1**	**72.9**	86.8	**79.1**
without $\mathcal{J}^{\mathrm{RC}}$	92.7	50.1	**87.1**	63.2
without $\mathcal{L}^{\mathrm{RC}}$	92.3	68.4	85.7	75.9
without $\mathcal{L}^{\mathrm{RC}}, \mathcal{J}^{\mathrm{RC}}$	**93.1**	69.1	86.4	76.6

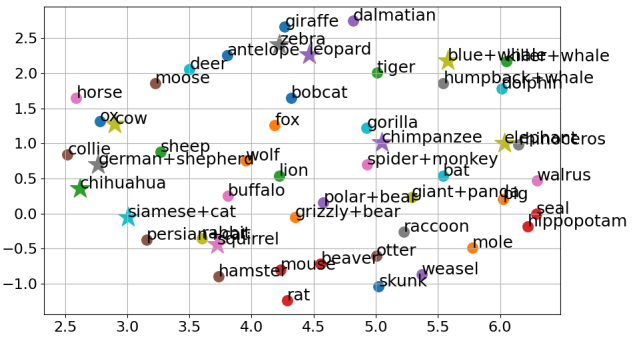


図 5: 推定されたクラス知識の例：クラス知識は 85 次元のため UMAP を用いて 2 次元に削減した．星印は未知クラスを示す．

足動物，クジラ，げっ歯類などのクラスがまとまっており，提案手法がクラスの意味表現を獲得できることが示唆された．

5.3 実　験　3

実験結果を表 4 に示した．この結果から，提案手法は Generalized ゼロショット学習において既知クラスで 42.7%，未知クラス以外で 36.4%の精度を達成した．また，テスト対象者の既知クラスデータで訓練した教師あり individual モデルの精度 54.0%に比べ，教師あり common モデルの精度は 37.7%と低く，個人を区別せずに汎用モデルを学習するだけでは性能が低下した．これに対して提案手法は，タスクの特徴を抽出するクラス知識のメタ学習戦略を採用することで，common モデルと同じデータを用いたにも関わらずより高い性能を達成した．また，提案手法は既知クラスの分類性能だけでなく，未知クラスの認識精度においても教師あり学習を上回り，タスク特徴を考慮しない common モデルと，個人差の影響を受けない individual モデルの中間に位置する性能が得られた．さらに，再構成誤差の有無を比較から，再構成誤差を導入することで既知クラスの分類性能を大きく低下させることなく未知クラスの分類性能を向上させることが分かった．

表 4: DREAMER データセットにおける Traditional ゼロショット学習の未知クラスの正解率 (**T**)，Generalized ゼロショット学習における未知クラス正解率 (**u**)，既知クラス正解率 (**s**)，およびそれらの調和平均 (**H**)．各列ごとに最大値は太字で示した．

Method	DREAMER (%)			
	T	**u**	**s**	**H**
MLP(common)	-	25.3	37.7	27.6
MLP(individual)	-	0.0	**54.0**	0.0
Proposal	**64.1**	**36.8**	41.6	**38.8**
without $\mathcal{J}^{\mathrm{RC}}$	57.3	23.3	37.5	27.1
without $\mathcal{L}^{\mathrm{RC}}$	57.9	18.8	42.7	25.3
without $\mathcal{L}^{\mathrm{RC}}, \mathcal{J}^{\mathrm{RC}}$	63.8	26.4	43.1	31.0

6　結　　論

複数のタスクからクラス知識と分類モデルを同時に学習することで，クラス知識が事前に与えられなくてもゼロショット学習を適用可能にする新しい手法を提案した．本研究は複数のタスクを効率的に解き，ベンチマークデータセットにおいて高い未知クラス認識性能を達成した点で先進的である．今後の展望として，提案手法はいくつかの観点での改善点が存在する．1 つ目に，多くのクラスが欠損している場合であっても，クラス知識を推定可能なモデルを訓練できる必要がある．2 つ目に，既知クラスと未知クラスのプロトタイプを交互ではなく同時に学習することで，既知クラスと未知クラスを混同しにくいクラス知識の学習方法が必要である．3 つ目に，与えられるタスクが少ない場合でもメタ学習できるよう手法を拡張する必要がある．

文　献

[1] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *National Conference on Artificial Intelligence - Volume 2*, p. 646–651, 2008.

[2] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, Vol. 22, 2009.

[3] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, pp. 584–599, Cham, 2014.

[4] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on International Conference on Machine Learning - Volume 37*, p. 2152–2161, 2015.

[5] E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4447–4456, Los Alamitos, CA, USA, jul 2017.

[6] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 7, pp. 1425–1438, 2016.

[7] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, Vol. 124, No. 3, pp. 356–383, Sep 2017.

[8] Tristan Hascoet, Yasuo Ariki, and Tetsuya Takiguchi. Semantic embeddings of generic objects for zero-shot learning. *EURASIP Journal on Image and Video Processing*, No. 1, p. 13, Jan 2019.

[9] S. Changpinyo, W. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, Los Alamitos, CA, USA, jun 2016.

[10] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4281–4289, 2018.

[11] Rohit Keshari, Richa Singh, and Mayank Vatsa. Generalized zero-shot learning via over-complete distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[12] Peirong Ma and Xiao Hu. A variational autoencoder with deep embedding model for generalized zero-shot learning. *The AAAI Conference on Artificial Intelligence*, Vol. 34, No. 07, pp. 11733–11740, Apr. 2020.

[13] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. *AAAI Conference on Artificial Intelligence*, Vol. 35, No. 3, pp. 1966–1974, May 2021.

[14] S. Chen, Z. Hong, G. Xie, W. Yang, Q. Peng, K. Wang, J. Zhao, and X. You. Msdn: Mutually semantic distillation network for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7602–7611, Los Alamitos, CA, USA, jun 2022.

[15] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[16] Shah Nawaz, Jacopo Cavazza, and Alessio Del Bue. Semantically grounded visual embeddings for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4589–4599, June 2022.

[17] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang. Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 7, pp. 1625–1638, 2018.

[18] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Attribute propagation network for graph zero-shot learning. *AAAI Conference on Artificial Intelligence*, Vol. 34, No. 04, pp. 4868–4875, Apr. 2020.

[19] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[20] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. *AAAI Conference on Artificial Intelligence*, Vol. 34, No. 04, pp. 6062–6069, Apr. 2020.

[21] Yunlong Yu, Zhong Ji, Jungong Han, and Zhongfei Zhang. Episode-based prototype generating network for zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14032–14041, 2020.

[22] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, p. 1126–1135, 2017.

[23] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. 03 2018.

[24] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, Vol. 80, pp. 1704–1713, 10–15 Jul 2018.

[25] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis Machine Intelligence*, Vol. 41, No. 09, pp. 2251–2265, sep 2019.

[26] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, No. 1, pp. 98–107, 2018.

[27] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition*, 2016.

[28] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, Vol. 53, No. 4, pp. 1689–1696, feb 2021.

[29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, Vol. 30, 2017.

[30] Guangzhen Liu, Jiechao Guan, Manli Zhang, Jianhong Zhang, Zihao Wang, and Zhiwu Lu. Joint projection and subspace learning for zero-shot recognition. In *IEEE International Conference on Multimedia and Expo*, pp. 1228–1233, 2019.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[32] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 13:30 ~ 15:40 | T1-B オンライン（Zoom Events)

論理・意味分析

座長:鷹野 孝典(神奈川工科大学)
コメンテータ:上土井 陽子(広島市立大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-B-2-01] 俳句･短歌を対象とした短文の意味分析

*杉田 祥太郎[1]、田中 克己[1] (1. 福知山公立大学情報学部)

14:00 ~ 14:25
[T1-B-2-02] 対照学習による慣用句を考慮した文埋め込みの獲得

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

14:25 ~ 14:50
[T1-B-2-03] 命題論理における言語の構成性に着目した言語モデルの汎化能力の調査

*井上 綾介[1]、宮森 恒[1] (1. 京都産業大学)

14:50 ~ 15:15
[T1-B-2-04] 金融数学問題解法におけるChatGPTによる言語処理と記号処理機能の統合

*白田 由香利[1]、桑名 杏奈[2]、Chakraborty Basabi[3,4] (1. 学習院大学経済学部、2. 和洋女子大学全学教育センター、3. マダナパレ工科大学、4. 岩手県立大学)

俳句･短歌を対象とした短文の意味分析

杉田 祥太郎† 田中 克己†
†福知山公立大学情報学部 〒602-0886 京都府福知山市字堀 3370
E-mail: †{332245045, tanaka-katsumi}@fukuchiyama.ac.jp

あらまし 俳句･短歌を対象として短文の意味分析と絵画性分析を行った．意味分析は，短文の意味を分析するもので，語の意味学習を行う Word2Vec に基づいて短文の意味分析を行うものである．「絵画性」とは，言葉によって事物のイメージをどの程度具体的に惹起できるかというもので，本研究では，語の具体度に基づく短文の絵画性分析を行った．

キーワード 文の意味分析 俳句･短歌 機械学習 語の具体度計算 短文の絵画性分析

1. まえがき

テキストの機械学習技術と言語処理技術の進展により，様々なテキスト情報の「意味」の分析が可能となりつつある．特に，語や文や文書の「意味の分散表現」をニューラルネットワークの機械学習で求める T.Mikolov らの Word2Vec 手法が著名である[1][2].

本論文では，俳句や短歌などのような短文テキストを対象として，文の意味分析や絵画性分析を行う手法を提案する．文の意味分析は，与えられた文に出現する各語の意味ベクトルを加減算することで元の文の意味ベクトルを合成する方法である.文の絵画性分析は，文中に出現する各語の視覚度（Visuality）を加算して文の絵画性を得る手法である．これらの手法を俳句・短歌に適用して，意味や視覚度が類似する俳句・短歌を検索する実験を行ったので，その実験結果についても報告する．

2. 関連研究

語の意味の分散表現をニューラルネットワークの機械学習で求める手法として，T.Mikolov らの Word2Vec 手法が著名である [1]. T.Mikolov らは，その後，この手法を拡張して，文や文書の意味の分散表現を求める手法も提案している [2]. また，語の意味は時間や時代によって変化する．このため異なる時期・時代に使用されてきた語の意味を機械学習で求める手法が Yating らによって提案されている [3]. 図 1 にスキップグラムニューラルネットワークモデルによる語の意味の学習を行う仕組みを示す．図 1 では，単語「iPod」の意味を機械学習で求めている．図 2 は，スキップグラムニューラルネットによる語の意味の演算が可能なことを示している（「Berlin」の意味から「Germany」の意味を減算し，さらに「Japan」の意味を加えることで,「Tokyo」の意味が近似できている）.

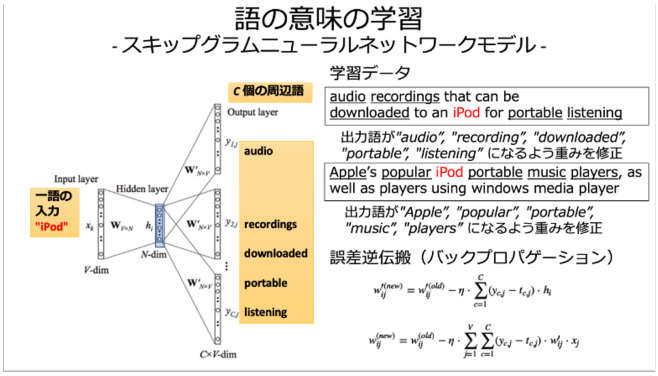


図 1 Word2Vec による単語「iPod」の意味の学習

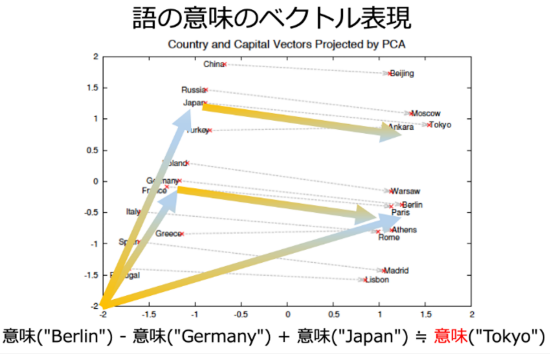


図 2 語の意味の演算

最近では，機械学習によって新たな文やテキストを生成する仕組みについても多くの研究がなされており，例えば, T. Zhang らは BERT を用いたテキスト生成手法を提案し，生成されたテキストを評価するために BERTScore というアルゴリズムを提案している [4].

著者らは，本研究の先行研究として，語の意味の分散表現から，短文（俳句，短歌，惹句）の「意味」を計算するアルゴリズムおよび意味類似する短文検索の手法を提案した [5].

3. 俳句・短歌の意味分析

3.1. 語の意味分析

青空文庫や関連 Web サイトから，実験に用いる俳句・短歌・小説文データを収集した[1) 2) 3) 4) 5)].

- 芭蕉発句全集 1066 句（伊藤洋作） 1）
- 一茶発句全集約 2 万句（小林一郎編） 2）
- 和歌と俳句（与謝蕪村） 3）
- みだれ髪（与謝野晶子）（青空文庫） 4）
- 銀河鉄道の夜（宮沢賢治）（青空文庫） 5）

収集した与謝蕪村の俳句 1424 首の一部とその形態素解析結果を図 3，図 4 に示す．

```
1  古庭に鶯啼きぬ日もすがら
2  老武者と大根あなどる若菜哉
3  百とせの枝にもどるや花の主
4  春の海終日のたりのたり哉
5  辛崎の朧いくつぞ与謝の海
6  一筋もすたる枝なき柳かな
7  横に降る雨なき京の柳かな
8  象の眼の笑ひかけたり山桜
9  臑白き従者も見へけり花の春
10 そこそこに京見過しぬ田にし売
11 捨やらで柳さしけり雨のひま
12 野路の梅白くも赤くもあらぬ哉
13 春の雨穴一の穴にたまりけり
14 うぐひすのあちこちとするや小家がち
```

図 3 与謝蕪村の俳句 1424 首の一部

```
庭 鶯 啼く すがる 武者 大根 あなどる 若菜
枝 もどる 花 主 春 海 終日 のたりのたり 辛 崎
与謝 海 一筋 すたる 枝 ない 柳 横 降る 雨 ない
柳 象 眼 笑 ひる 山桜 臑 白い 従者 見る
り 花 春 そこそこ 京見 過す 田 する 柳
雨 ひま 野路 梅 白い 赤い ある 春 雨 穴
穴 たまる 小家 鶯 声 遠い 暮 ける 青柳
芹生 里 せり 女 京 寒い 御忌 詣る 鶯
ふみ やぶる 入 夢 小豆 煮る 初午 家 家
袖 たみ 初午 物種 うり 日 あたる 風吹 夜 もの凄
柳 はる 綱 袂 小 うつ 庭 茶筅 花
椿 兀山 かくれる きじ こ 春雨 小磯 貝
雛 見世 灯 引く ろ 春 雨 春雨 人 住む けぶる
壁 洩る 糸桜 孤 書院 むく 起てる ふる 犬 宝
```

図 4 与謝蕪村の俳句の形態素解析

```
> model
A VectorSpaceModel object of  1003  words and  100  vectors
          [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
</s>  0.004002686  0.004419403 -0.003830261 -0.003278046  0.001366577  0.003021088
する -0.188971728 -0.130228966  0.363412470 -0.203882977 -0.044897571 -0.257167935
花   -0.345784932 -0.182079196 -0.051740151 -0.084464587  0.058473658 -0.201170817
春   -0.199579641  0.225326672 -0.016330827 -0.063737392  0.295237631 -0.114961855
月   -0.350502521  0.049638912 -0.044909962  0.296696246  0.035704248  0.230843857
夜   -0.156838492 -0.131099701  0.125658587 -0.305767447  0.091929294 -0.209837198
秋   -0.147907704  0.031554800  0.368649095  0.369016230 -0.239375383 -0.190454021
ふる -0.178795800  0.309624612  0.326081425  0.096827611  0.015396807  0.012696579
ない -0.054425057 -0.286158442  0.222785905 -0.148025438  0.082069248  0.014903516
見る -0.357508093 -0.207425162  0.023792153  0.086095028  0.135355636 -0.106827073
```

図 5 蕪村の俳句を機械学習して得られた語の意味（Word2Vec で機械学習した結果 model ）

```
> ## 意味的に近い単語を求める
> model %>% closest_to("花")
     word similarity to "花"
1      花              1.0000000
2   よしの             0.8239078
3     柚               0.8118728
4     椎               0.7863986
5    そば              0.7841020
6  いばら              0.7815711
7  すだく              0.7734632
8     蔓               0.7683786
9     桜               0.7644134
10  ふす               0.7618870

> model %>% closest_to(~"五月雨"+"家")
    word similarity to "五月雨" + "家"
1     旅                        0.9665826
2     床                        0.9646886
3    曾我                       0.9595311
4     簔                        0.9594445
5     蚓                        0.9580128
6     雫                        0.9513634
7  またぐ                       0.9500562
8  あける                       0.9499454
9     魔                        0.9482862
10  家中                        0.9464029

> model[["五月雨"]] %>% cosineSimilarity(model[["家"]])
          [,1]
[1,] 0.7167073
```

図 6 俳句の機械学習に基づく意味類似語の検索

与謝蕪村の俳句を Word2Vec で学習した結果を図 5 に示す．この図にあるように，各語の意味は 6 次元の意味ベクトルで表現され，model という変数に格納されている．

蕪村の俳句の語の意味の機械学習の結果を用いて，意味類似語の検索が可能となる．図 6 では，語"花"，および，2 語 "五月雨+家"に対して，それぞれ意味類似する語を検索している．検索結果は，意味類似語と類似度のペアという形で示されている． 2 語を与えて，その語間の意味類似度をコサイン類似度で求めることも可能であり，図 6 では，"五月雨"と"家"とのコサイン類似度も出力している．

意味類似語の検索性能は，機械学習した俳句テキストに依存する点に注意が必要である．例えば，機械学習した俳句データが与謝蕪村の俳句の場合と芭蕉の俳句の場合では，意味類似する語が明らかに異なる．図 7 にその 1 例を示す．

1) https://www2.yamanashiken.ac.jp/~itoyo/basho/haikusyu/Default.htm
2) http://www.janis.or.jp/users/kyodoshi/issaku.htm
3) http://www5c.biglobe.ne.jp/~n32e131/haiku/buson411.html
4) http://www.aozora.gr.jp/cards/000885/files/51307_ruby_47032.zip
5) http://www.aozora.gr.jp/cards/000456/files/456_ruby_145.zip

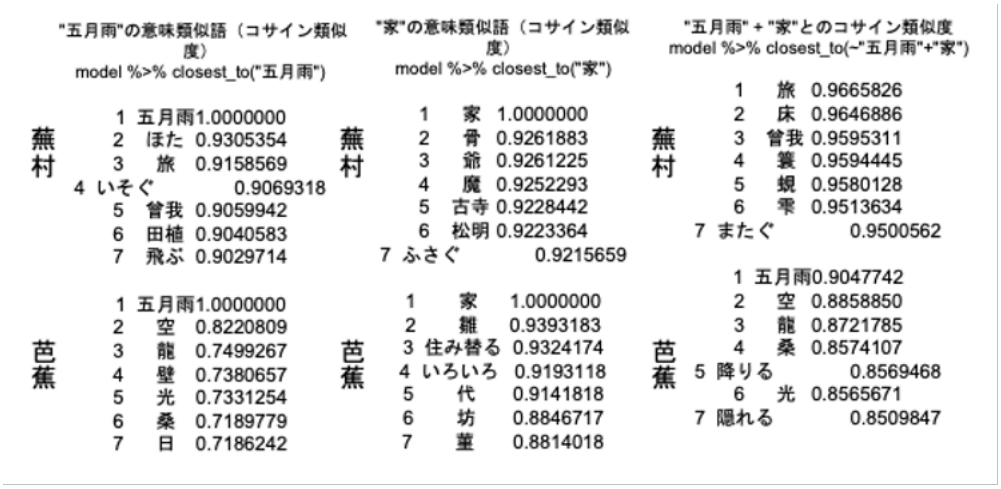


図 7 蕪村/芭蕉の俳句の意味類似語の検索

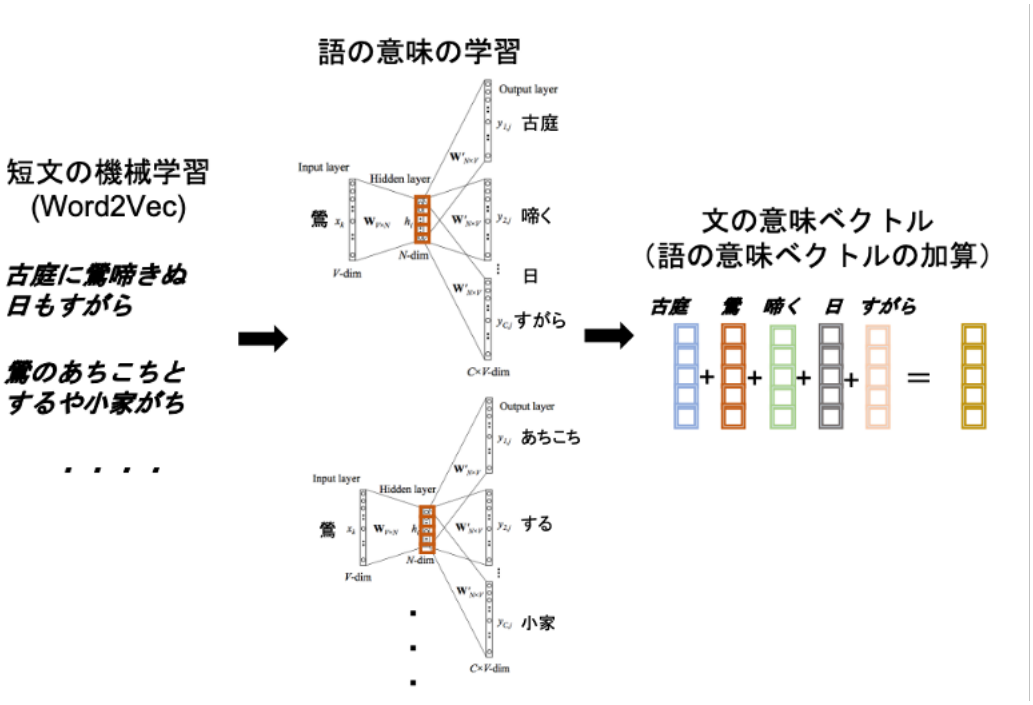


図 8 意味ベクトル合成法による文の意味ベクトルの生成

3.2. 文の意味分析

本節では，語の意味分析を行う Word2Vec の結果を用いて，文の意味分析を行う「意味ベクトル合成法」を提案する．意味ベクトル合成法は，与えられた文に出現する各語の意味ベクトルを加減算することで元の文の意味ベクトルを合成する方法である．Word2Vec では，語の意味ベクトルの加減算が可能であり [1]，本稿で提案する意味ベクトル合成法も，この語の意味ベクトルの加減算を利用したものである．

文の意味ベクトル合成法の手順は，以下の通りである．

① 短文テキスト集合を Word2Vec で学習し，語の意味を意味ベクトルとして獲得する（各語の意味は，Word2Vec の中間層の値のベクトルとして獲得する）．

② 短文テキスト $s = w_1 w_2 \ldots w_n$ の意味ベクトル meaning(s) とし，これを，各語 w_i の意味ベクトル meaning(w_i) を加算（または，重み付き加算）することで求める（図 8 参照）．

文の意味ベクトル合成法の長所は，計算量を増やすこと無く，また，文中の語の出現場所に関係無く文の意味ベクトルが求められる点である．

文の意味ベクトル合成法および Python プログラムを，図 8，図 9 に示す．また，このプログラムの実行結果を図 10，図 11 に示す．

```
import numpy as np
import numpy.linalg as npl
f = open(r"Basho.txt", 'r',encoding = "UTF-8")
s=f.read()
text=s
texts=text.split("\n")
texts=list(filter(None,texts))

from janome.tokenizer import Tokenizer
t= Tokenizer()
def extract_words(text):
    tokens = t.tokenize(text)
    return [token.base_form for token in tokens
            if token.part_of_speech.split(',')[0] in['名詞','動詞',"形容詞"]]

text_list=[extract_words(text) for text in texts]

from gensim.models import word2vec
model = word2vec.Word2Vec(text_list, size=100,min_count=1,window=5,iter=50)

max_list=[]
for x in range(1,1136):
    x=int(x)
    y=int(1)
    apv=[]
    for h in range(len(text_list)):
        for i in range(100):
            hy=[]
            hy.append(0)
        apv.append(hy)
    for h in range(len(text_list)):
        for i in text_list[h]:
            v=model.wv[i]
            nv=np.array(v)
            apv[h]+=nv

    rui_list=[]
    jyun_list=[]
    c=0
    for i in range(len(text_list)):
        try:
            cos=(np.dot(apv[x],apv[i])/(npl.norm(apv[x])*npl.norm(apv[i])))
            rui_list.append(cos)
        except ValueError:
            rui_list.append(0)
    rui_list[x]=0
    maxrui=sorted(rui_list,reverse=True)
    for i in maxrui:
        jyun_list.append(rui_list.index(i))
    #print(texts[x])
    #print()
    for i in jyun_list[:y]:
        #print(texts[i])
        #print("類似度:"+str(maxrui[c]))
        max_list.append(maxrui[c])
        c+=1
ruimax=max(max_list)
ruimaxnum=max_list.index(ruimax)
ruimin=min(max_list)
```

図 9 文の意味ベクトル合成法 Python プログラム

```
上位何件?:10
五月雨を集めて早し最上川

五月雨の降り残してや光堂
類似度:0.9922054237530693
海は晴れて比叡降り残す五月哉
類似度:0.9880728919757348
五月雨や龍頭あぐる番太郎
類似度:0.9853719156700066
五月雨に鶴の足短くなれり
類似度:0.9833700530716274
氷苦く偃鼠が喉をうるほせり
類似度:0.981645416174487
初秋や畳みながらの蚊屋の夜着
類似度:0.9803578010232169
髪生えて容顔青し五月雨
類似度:0.978935413207665
寒菊や粉糠のかかる臼の端
類似度:0.977161049466941
徒歩ならば杖突坂を落馬かな
類似度:0.975224317257702
五月の雨岩檜葉の緑いつまでぞ
類似度:0.9751709032537185
```

図 10 芭蕉の句と意味類似する俳句の検索

図 10 では，「五月雨を集めて早し最上川」の意味ベクトルとのコサイン類似度が高い俳句を表示している．"五月"，"五月雨"という単語が入った俳句が多いことから分かるように，元となる短文テキストと同じ語句が使われている短文テキストは類似度が高いと判定される傾向にある．また，類似度が 0.9 点後半と計算されており，非常に高い類似度が得られている．しかし，実際にはこのような高い類似度が出る俳句ではないと考えられる．

そこで tf-idf 法を用いて，語ベクトルに重み付けを行って文の意味ベクトルを生成するアルゴリズムも開発した．実行結果を図 11 に示す．重み付けを行うことで類似度を抑えられたので効果はあったと考えられる．

```
上位何件?:10
五月雨を集めて早し最上川

五月雨の降り残してや光堂
類似度:0.7584951949400849
名月や池をめぐりて夜もすがら
類似度:0.7101203197032014
海は晴れて比叡降り残す五月哉
類似度:0.7058665719753245
日の道や葵傾く五月雨
類似度:0.6721316538833448
暑き日を海に入れたり最上川
類似度:0.6622948059643169
五月雨に隠れぬものや瀬田の橋
類似度:0.6572847704288667
五月雨に鳰の浮巣を見にゆかん
類似度:0.6513813145662887
花咲きて七日鶴見る麓哉
類似度:0.633733463926172
ごを焚いて手拭あぶる氷哉
類似度:0.6180540213011693
髪生えて容顔青し五月雨
類似度:0.6151106037051939
```

図 11 芭蕉の句と意味類似する俳句の検索 (重み付けあり)

4. 俳句・短歌の絵画性分析

俳句や短歌では，言葉によって事物のイメージがどの程度具体的に惹起できるかが重要とされており，すなわち，俳句・短歌の「絵画性」が重要となる．例えば，芭蕉と蕪村の俳句を比較する上で，蕪村の句の方が「絵画的」と言われている（図 12 参照）．

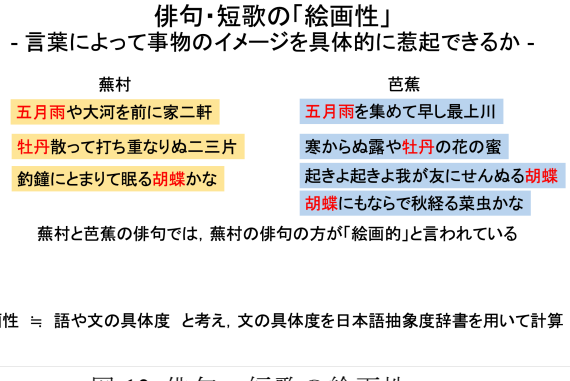


図 12 俳句・短歌の絵画性

本研究では，俳句･短歌の「絵画性」を分析するために，「語の具体度」に関する辞書（日本語抽象度辞書 AWD-J EX）に基づいて分析を行った．直観的には，具体度の高い語を多く含む俳句・短歌は「絵画性」が高いと仮定している．図 13 は芭蕉の俳句で具体度が高い上位 100 句を，図 14 は，俵万智の短歌で具体度が高い上位 100 句を示している．

図 13　芭蕉の俳句で具体度が高い上位 100 句

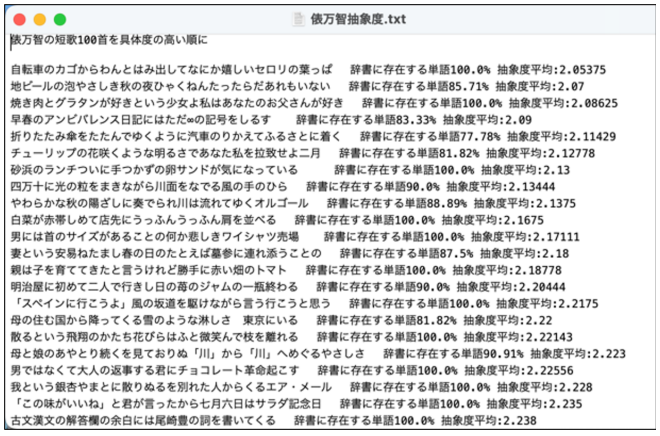


図 14　俵万智の短歌で具体度が高い上位 100 句

短文テキストの絵画性分析を行うための具体度を求める手順は以下の通りである．

① 短文テキストに含まれる名詞，動詞，形容詞の抽象度を日本語抽象度辞書 AWD-J EX に基づき 1（最も具体的）～5（最も抽象的）の間で定める．

② 短文テキストの具体度を①で求めた値の平均をとることで求める．

図 13，14 の結果を見ると，ものの名前や地名など具体的な単語が多く含まれているテキストの具体度が高くなっている．文脈を考慮していないことなど，今後の改良が必要な点は多くあるものと考えられる．

図 15 に芭蕉と蕪村の俳句の絵画性比較の結果を示す．この結果では，五月雨の句は芭蕉の句が，牡丹の句は蕪村の句の方が絵画性が高いという結果が出た．人の感覚とは若干乖離があるものと考えられる．

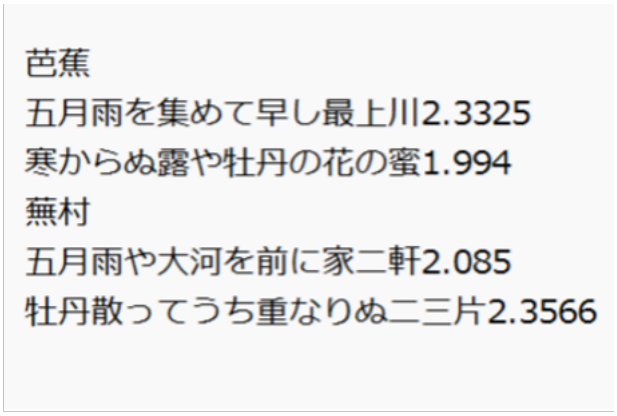


図 15 芭蕉と蕪村の俳句の絵画性比較分析

5. まとめ

俳句･短歌等の短文の意味分析･絵画性分析を行った．具体的には，(1) Word2Vec 機械学習による短文の意味分析，(2) 語の具体度に基づく短文の絵画性分析を行った．

短文の意味分析では，Word2Vec に含まれていない「文」の意味分析計算を Python で実装した．また，語ベクトルの合成において重み付き和によって文の意味ベクトルの合成を行えるように拡張した．さらに，コサイン類似度を用いて，類似する意味ベクトルをもつ短文テキストの検索も可能にした．

短文の絵画性分析を行うために，語や文の「具体度」を計算するプログラムを開発した．芭蕉と蕪村の俳句の「絵画性」比較実験を行ったが，今後は，学習対象とする俳句・短歌を広げていくとともに，「絵画性」計算のアルゴリズムを修正・拡張していく予定である．

謝辞

本研究は，令和 4，5 年度科研費基盤研究 (B)(一般)「機械学習による情報の意味獲得と意味類似情報の検索・生成」(代表：田中克己，課題番号 22H03905）によるものである．

参　考　文　献

[1] T. Mikolov, I.Sutskever, K.Chen, G.Corrado, J.Dean: “Distributed Representations of Words and Phrases and Their Compositionality”, Advances in Neural Information Processing Systems 26 (NIPS2013),

Vol.26, Curran Associates. Inc., 2013..

[2] Q.Le and T.Mikolov: “Distributed Representations of Sentences and Documents”, Proc. of 31st Int’l. Conf. on Machine Learning, pp.1188-1196, 2014.

[3] Y.Zhang, A.Jatowt, S.S.Bhowmick and K.Tanaka: “The Past is not a Foreign Country: Detecting Semantically Similar Terms Across Time”, IEEE TKDE, Vo.28, No.10, pp.2793-2807, 2016.

[4] T.Zhang, V.Kishore, F.Wu, K.Q.Weinberger, Y.Artzi: “Bertscore: Evaluating Text Generation with BERT”, In arXiv preprint arXiv:1904.09675, 2019.

[5] 小山大介, 上之園彗志, 杉田祥太郎, 田中克己: “短文テキスト（俳句・短歌・キャッチコピー）の意味分析と意味類似検索”, 東海関西データベースワークショップ 2022, 2022 年 9 月.

[6] 小山大介, 上之園彗志, 杉田祥太郎, 大島裕明, 荘司慶行, 田中克己: “短文テキストの意味分析と意味類似検索”, 1b3-3, DEIM2023, 2023 年 3 月.

[7] 杉田祥太郎, 上之園彗志, 田中克己: “俳句・短歌テキストの意味分析と意味類似検索”, 東海関西データベースワークショップ 2023, 1-4, 2023 年 9 月.

[8] 正岡子規 : ”俳人蕪村 “, 日本の文学 15, 中央公論社, 1967 年.

[9] 奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室 : ”日本語抽象度辞書 AWD-J: Abstractness of Word Databases for Japanese Common Words”, https://sociocom.naist.jp/awd-j/

対照学習による慣用句を考慮した文埋め込みの獲得

中岡　知己† 　宮森　恒†

† 京都産業大学大学院先端情報学研究科　〒 603-8555 京都府京都市北区上賀茂本山
E-mail: †{i2286121,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，慣用句を考慮した文埋め込みの獲得手法を提案する．従来の文埋め込みでは，単語埋め込みの合成や BERT を用いたファインチューニング等により様々な文埋め込みの獲得手法が提案されてきた．しかし，慣用句のような特定の単語列で特定の意味を有する表現については，その単語列の文字通りの意味との区別が必ずしも明確にされていないため，慣用句などが持つ意味を捉えるのは困難である．本稿では，慣用句の表現に焦点を当て，慣用句の意味を明示的に与えることで文埋め込みの性能向上を図る．具体的には，日本語の慣用句を用いた文と慣用句を用いない同じ意味の文とをペアとするデータセットを構築し，これを正例とした対照学習により，慣用句を的確に表現可能な文埋め込みを獲得する手法を提案する．実験では，提案手法により慣用句を的確に捉えられているのかについて評価し，その有効性を検証する．

キーワード　慣用句，文埋め込み，対照学習，言語モデル，自然言語処理

1　はじめに

自然言語をコンピュータ上で扱うための方法に，単語や文などを，その意味を的確に反映したベクトル空間中のベクトルに写像し処理する方法がある．ここで，単語や文をベクトル空間に写像することで得られるベクトルは，それぞれ単語埋め込み，文埋め込みと呼ばれる．また，入力された文をベクトル空間に写像する機能をもつネットワークはエンコーダと呼ばれ，系列変換モデルを構成する不可欠な要素である．エンコーダには，文が文字や部分単語単位で分割されたトークン系列を ID に変換した系列が入力され，単語埋め込み系列に変換された後，何らかの合成や変換処理を経て文埋め込みに相当するベクトルが出力される．もし，自然言語として適切な文意も保持した文埋め込みを獲得できれば，コンピュータ上でも人間の解釈に則ったより適切な形で文を処理することが可能になる．このような文意も保持した文埋め込みを獲得するためには，文脈によって変化する文意を適切に捉えることができるエンコーダが必要となる．そのため，本稿では，文意も適切にベクトル空間へと写像可能なエンコーダを構築することを目的とする．もし，適切な文意を保持するエンコーダを構築できれば，文意に沿った文間の類似をより的確に活用でき，検索エンジンをはじめ様々なシステムにおいて広く応用可能である．

近年では，Transformer に基づいた大規模言語モデルの登場により，自然言語処理の様々なタスクの実行能力が目覚ましい進歩を遂げている．例えば，BERT [1] は，Transformer のエンコーダ部分で構成されたモデルであり，文脈を考慮した文埋め込みを獲得することができる．このような大規模言語モデルは，様々なドメインにおける膨大なデータを用いて事前学習されている．事前学習の際には，自己教師あり学習を用いるため，膨大なデータに対する人手によるラベルづけ作業が不要となり，学習用データの構築は比較的容易である．様々なドメインの膨大なデータを事前学習することで，一般的な単語や文法といった汎用的な言語知識が獲得されるが，特定の専門知識までは十分に獲得することは難しい．そのため，一般に，特定のドメインやタスクに限定した専門的な学習データを用いて，事前学習済みモデルをファインチューニングすることで，所望のタスクに特化したより高性能なモデルを構築することができる．このように事前学習に基づく大規模言語モデルは，様々な自然言語処理タスクにおいて高い性能を示し目覚ましい成果を挙げている．しかし，BERT のような Transformer のエンコーダとして構成されるモデルであっても，文の意味を適切に捉えた文埋め込みを必ずしも獲得しきれていないことが指摘されている [2]．文の意味を適切に捉えた文埋め込みを獲得できていない原因の 1 つとして，文埋め込みが埋め込み空間の特定の領域や方向に偏って分布してしまう異方性の問題がある．文埋め込みが埋め込み空間において偏って分布していることにより，意味的に異なる文間のユークリッド距離が近くなってしまったり，疎な分布をしている領域での意味の類似性の適切さが担保されない問題が生じる．この異方性を解消するため，文埋め込みがなるべく等方的に分布するような変換を施す研究も存在する [3], [4]．自然言語での文意をより適切に捉えた文埋め込みを獲得するには，文の自然言語上での意味表現と文埋め込み空間上での意味表現に，適切な対応関係を持たせることが必要である．例えば，文埋め込み空間において，文意が同じ文間のユークリッド距離は近く，文意が異なる文間のユークリッド距離は遠くなるような対応関係を持たせることが考えられる．そのような文埋め込み空間では，コンピュータ上でも人間の解釈により近い文意の類似性を，単なるコサイン類似度によって算出可能になる．しかし，BERT などは文中での単語列の並びを考慮した文脈化単語埋め込みによって，静的単語埋め込みで問題となっていた多義語をより適切に扱えるが，非構成的な複数語表現 (Multi-WordExpressions, MWE) については依然として適切に扱いきれていないことが指摘されている．Garcia ら [5]

は，英語とポルトガル語での複合名詞を対象とし，その複合名詞が慣用性（複合名詞内の構成要素から意味が推測可能かどうか）があるのか，また複合名詞の文中での意味の構成性（文中の構成要素から意味が推測可能かどうか）の度合いに応じて，人手で分類した NCTTI（Noun Compound Type and Token Idiomaticity）データセットを構築している．この NCTTI データセットを用いて，文脈化単語埋め込みモデルとして BERT, DistilBERT，SentenceBERT を対象とし，また静的単語埋め込みモデルとして GloVe を対象として実験を行った．実験の結果，文脈化単語埋め込みモデルでは複合名詞の構成性の予測は，複合名詞の意味を捉える上であまり重要視されていないことが示唆され，加えて BERT に基づいた言語モデルの予測結果は，人手による分類結果と相関しないことから，意味の構成性をうまく捉えきれていない可能性があると指摘している．適切な文意を捉えた文埋め込みを獲得するためには，非構成的な複数語表現も適切に扱う必要があり，その複数語表現の 1 つに慣用句がある．慣用句を含む文の一例を,【例文１】として以下に示す.

【例文１】: <鼻を突く>臭いに，しのぶは思わず手で顔を押さえた.

ここでは，山括弧内の表現が慣用句であることを示している．例文１での慣用句は，<鼻を突く>であり，この慣用句を文字通りに扱ってしまうと，慣用句の意味を捉えられないため，全体の文意を誤解する可能性がある．このような誤解は，慣用句が非構成的な表現であるため，字面からは意味の推測が困難なことに起因する．そのため，慣用句のような非構成的な表現については，予め知識として慣用句の意味を把握しておく必要があると考える．

本稿では，特に日本語の慣用句に着目し，慣用句が出現する文の文意も適切に捉えることができる文埋め込みの獲得を目指す．具体的には，文中における慣用句の意味を明示的に学習させることで，慣用句が含まれた文でも，より適切に文意を捉えた文埋め込みが得られることを目指す．ここで，文中における慣用句の意味を明示的に学習するには，慣用句の意味が付与されたデータセットが必要であるが，慣用句の本来の意味を含む文を収録した学習用データセット，およびベンチマークデータセットは存在しない．この文中における慣用句の意味を明示的に学習するには，慣用句の意味が付与されたデータセットが必要である．そのため，本稿では，まず日本語における慣用句が出現する文と，慣用句を本来の意味で言い換えた文のペアを収録し，学習用データセットとして慣用句データセット，およびベンチマークデータセットとして慣用句ベンチマークデータセットを構築する．慣用句データセットを構築するにあたって，基礎となるコーパスとして，日本語の慣用句とその用例文が収録されている OpenMWE コーパス [1]を使用した．この OpenMWE コーパスには 146 種類の慣用句を対象とし，慣用句が出現している用例文とその慣用句が，慣用句が持つ本来の意味として使用されているのか，または字面から解釈可能な意味で使用されているのか判別可能なラベルが人手で付与されている．しかし，この OpenMWE コーパスには，慣用句が持つ本来の意味は収録されていないため，別途オンライン辞典等から慣用句の意味を収集する必要がある．慣用句データセットの構築手順は以下の手順で進める．まず，オンライン辞典である goo 国語辞典 [2]から慣用句の意味をクローリングする．次に，収集した慣用句の意味を用いて，OpenMWE コーパスに収録されている用例文中で対応する慣用句を人手で書き換えた文を作成する．最後に，元の文と書き換えた文のペアの集合を整理し，慣用句データセットとする.

本稿では，慣用句を使用した文と慣用句を用いない同じ意味の文ペアを正例とする対照学習により，慣用句をより的確に表現可能な文埋め込みを獲得する手法を提案する．対照学習を用いた従来の文埋め込み獲得手法では，文中に出現するトークン列からだけでは解釈が困難な複数語表現 (慣用句) について考慮されておらず，この点で提案手法とは異なる．提案手法では，慣用句を用いた文と，慣用句を用いない同じ意味の表現で慣用句を書き換えた文を，正例ペアとした対照学習を行っているため，慣用句が含まれる場合でもより適切な文ペア間の対応関係を獲得できることが期待される．つまり，慣用句を，同じ意味の慣用句を用いないトークン列へと適切に変換可能な文埋め込みの獲得を意図している.

本稿の貢献は以下の通りである.

（1） 本稿では，日本語の慣用句を考慮した埋め込み空間を学習・評価するための慣用句データセットおよび慣用句ベンチマークデータセットを構築した.

（2） 慣用句データセットを使用した対照学習を用いることで，明示的に慣用句が持つ本来の意味を学習可能な文埋め込み獲得手法を提案した.

（3） 提案手法の有効性を，ベンチマークデータセットである JSTS，および慣用句ベンチマークデータセットを用いて検証した.

本稿の構成は以下の通りである．第 2 節では，文埋め込み獲得手法の関連研究を紹介し，本稿の位置付けを明確にする．第 3 節では，提案手法である慣用句を使用した文と慣用句を用いない同じ意味の文ペアを，正例とする対照学習により，慣用句を的確に表現可能な文埋め込みの獲得手法について詳しく説明する．第 4 節では，本稿で構築した慣用句データセットおよび慣用句ベンチマークデータセットについて詳述する．第 5 節では，ベンチマークデータセットの 1 つである JSTS，および，慣用句ベンチマークデータセットを用いた評価実験を通して，提案手法の有用性を検証する．第 6 節では，本稿の結論をまとめ，今後の課題を整理する.

1：https://github.com/nlp-waseda/OpenMWE

2：https://dictionary.goo.ne.jp/jn/

2 関連研究

本節では，関連研究として，BERT 登場以前の慣用句・慣用性に関する研究を紹介する．2.2 節では，BERT 登場以降の慣用句・慣用性に関する研究を紹介する．2.3 節では，自然言語処理における対照学習に関する研究を紹介する．

英語や日本語などの自然言語において慣用句のような複数語表現が存在する．自然言語処理において，文の意味を的確に捉える上ため，複数語表現を考慮する重要性は知られているが，大規模言語モデルにおいても複数語表現を的確に扱えるとは言えないのが現状である．

そのため，慣用句などの複数語表現を扱う従来研究が存在し，特に慣用句は，その文中で慣用性を持っているのかどうか決まるため，語義曖昧性を持っている点でも扱いが難しい．慣用句を対象とする研究では主に，文中に出現する慣用句が慣用性を持っているのかを分類するタスク，もしくは，文中からトークン単位で慣用句の構成要素であるか分類するタスク（文中での慣用句の出現位置の特定）が行われている．

2.1 BERT 登場以前の慣用句・慣用性に関する研究

Salton ら [6] は，RNN に基づいたエンコーダ・デコーダで構成される Skip-Thought Vectors（Sent2Vec）を使用し，単方向と双方向の RNN による単語埋め込みを連結した高次元ベクトルによって慣用性をある程度識別可能な文埋め込みを獲得している．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．

Liu ら [7] は，Tree-structured LSTM（TreeLSTM）を使用し，文中での慣用句の位置を検出する機構，および慣用句の意味を参照できる機構を用意している．それに加えて，慣用句の表現が必ずしも固定されていない柔軟性という性質を持っていることから，この柔軟性を吸収するため，文を形態素単位での単語埋め込みも活用している．この形態素単位での単語埋め込みによって，慣用句の柔軟性を捉える性能が向上しているのは興味深い．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．また，本稿では，BERT から得られる文脈化単語埋め込みしか使用しない点でも異なる．

King ら [8] は，英語における動詞と名詞を組み合わせた表現 (Verb-noun combinations, VNC) が，慣用性を持った用法で使用されているのかを分類するモデルをいくつか提案している．ここでは，従来モデルの単語埋め込みと VNC の埋め込み表現 (VNC が見出し語として使用されている場合は 1，それ以外は 0) の 1 次元ベクトルを連結することで性能が向上すると報告されている．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．

Haagsma [9] は，入力される文の埋め込みと，文中の単語のどの部分が慣用句であるのかを示すベクトルを入力とし，出力層に注意機構を追加した単方向 LSTM を提案している．また，入力される文の埋め込みと，出力層に注意機構を追加した双方向 LSTM も提案している．それに加えて，単方向 LSTM と双方向 LSTM の出力値を連結する手法も提案している．未知の慣用句を含む語義曖昧性タスクにおいて，単方向 LSTM と双方向 LSTM の出力値を連結する手法が最も良い性能であると報告されている．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．また，本稿では，BERT から得られる文脈化単語埋め込みを微調整する点でも異なる．

2.2 BERT 登場以降の慣用句・慣用性に関する研究

Hashempour ら [10] は，慣用句を 1 つのトークンとして扱う必要があると考え，文脈化単語埋め込みモデルである Context2Vec および BERT を用いた慣用性の有無を分類する手法を提案している．特に Context2Vec においては，慣用句を 1 つのトークンとして扱うことによって，慣用性の有無を分類する性能が向上すると報告されている．また，提案手法を Context2Vec に適用した場合に，次元削減した文脈化単語埋め込み空間において慣用性の有無を区別できていることが示された．本稿では，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．また，本稿では，慣用句を 1 つのトークンとして扱っていない点でも異なる．

Zeng ら [11] は，現実的な問題として，文中に学習していない未知の慣用句が出現する可能性があるため，そのような新しい慣用句についても識別し，その出現位置も予測可能な手法を提案している．文埋め込みには，文字単位の埋め込み，静的単語埋め込み，文脈化単語埋め込みに加えて，品詞タグの埋め込みも使用している．この静的単語埋め込みと品詞タグの埋め込みによって，構文構造の傾向に基づいて未知の慣用句の認識を可能にしている．また，文脈化単語埋め込みと文字単位の埋め込みによって，文字通りの表現と文埋め込み表現での意味関係の埋め込みを行なっている．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．また，本稿では，BERT から得られる文脈化単語埋め込みしか使用しない点でも異なる．

Nedumpozhimana ら [12] は，BERT が与えられた文中のどの部分に注目することで，慣用句を把握しているのかをプロービング手法とマスキング手法を用いて分析している．その結果，BERT は慣用句内の表現だけでなく，文脈内の表現にも注目することで文埋め込みを獲得している可能性が指摘されている．それに加えて，BERT は文中に出現する慣用句が慣用性を持っているのかどうかを識別する際に，主に慣用句内の表現に注目している可能性も指摘されている．

Phelps [13] は，出現頻度は低いが重要であるレアワードを考慮した BERTRAM を基礎として使用し，追加的に慣用句を学

習させることによって，慣用句を考慮した文埋め込みを獲得可能であり，文の品質に敏感であることが報告されている．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．

Zeng ら [14] は，言語モデルとして Transformer のエンコーダ・デコーダで構成される BART を使用し，慣用句の文埋め込みを追加で学習させるアダプタを用意している．基礎となる BART のパラメータは凍結（学習しない）し，アダプタだけを追加で学習することで文埋め込みの性能向上が報告されている．ここでは，エンコーダだけではなくデコーダも使用し，文中の慣用句をマスクして得られる文埋め込みから文を再構成するタスクを解くことで，文埋め込みの性能向上に繋がると報告されている．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．また，本稿では，ファインチューニング時に，基礎となる BERT のパラメータも更新する点でも異なる．

Nedumpozhimana ら [15] は，BERT は，一般的な慣用句の慣用性の有無を分類するために必要な情報を効率的に獲得していることを報告している．この論文では，学習データに含まれるどの慣用句が BERT モデルの性能向上に寄与しているのか調べるため，シャプレー値を計算している．その結果，慣用性の有無を分類するタスクでは，シャプレー値が低い慣用句の特性は，他の慣用句から獲得するのは困難であり，シャプレー値が高い慣用句ほど未知の慣用句に対して有用に働くことが示された．それに加えて，学習データセット全体での慣用句の分布とある慣用句の分布の乖離を調べるためカルバック・ライブラー距離とシャプレー値には弱い相関が見られることから，この分布の乖離が小さいほど慣用句の特性獲得に有用であることが示唆された．また，慣用性の有無での分布におけるカルバック・ライブラー距離とシャプレー値には，負の相関があることから，慣用句の特性の獲得に有用でないことが示された．同様に，各慣用句のデータセットのサイズおよび慣用性がある慣用句の使用割合とシャプレー値の相関は非常に弱く，これらは慣用句の特性獲得に有用でないことが示された．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．

高橋ら [16] は，日本語における慣用句を対象とした Open-MWE コーパスに基づき，文脈化単語埋め込みとして BERT の出力を使用し，慣用性の有無を分類するモデルを提案している．本稿では，大規模言語モデルである BERT を対象として，対照学習を用いた日本語の慣用句を的確に捉えた文脈化単語埋め込みの獲得を目指している点で異なる．

2.3 自然言語処理における対照学習に関する研究

Hadsell ら [17] は，学習データ間での近傍関係のみを使用し，高次元のデータを低次元へと写像する非線形関数を学習する手法として，DrLIM(Dimensionality Reduction by Learning an Invariant Mapping) を提案している．ここでは，ユークリッド距離において，類似している入力ベクトルは近傍に写像し，反対に類似していないベクトルは遠方へと写像するような損失関数を使用している．

Carlsson ら [18] は，BERT などの事前学習済みのモデルに対して，文の意味を捉えた文埋め込みを得るための自己教師あり学習による損失関数を提案している．ここでは，2 つの独立したモデルを用意し，文ペアは独立した別々のモデルに入力される．同一の文ペア間の文埋め込みは内積が大きくなるように，反対に異なる文ペア間の文埋め込みは内積が小さくなるような損失関数を定義している．負例については，ランダムに選択した文ペア間では，意味が異なる可能性が高いという仮定に基づいて，ランダムに選定している．

Giorgi ら [19] は，教師なし手法を用いた普遍的な文埋め込みの獲得手法として，DeCLUTR(Deep Contrastive Learning for Unsupervised Textual Representations) を提案している．ここでは，同一文書内の近傍から抽出した文ペア間は，意味的に類似していると考え，それらの類似度が大きくなる文ペアをサンプリングするような損失関数を用いて学習している．

Gao ら [20] は，自然言語に対照学習を適用した教師なし手法および教師あり手法を，SimCSE として提案した．対照学習を用いて，BERT などの文埋め込み表現を，文意をより適切に捉えられるようにファインチューニングをしている．具体的には，正例ペア間 (文意が類似) では損失が小さくなり，負例ペア間 (文意が相違) では大きくなるような損失関数を使用して学習を行なっている．教師なし手法では，入力データに対してドロップアウト マスクを 2 度施した異なる埋め込み表現を正例ペアとして学習に使用し，教師あり手法にも匹敵する性能がえられると報告されている．また，教師あり手法では，含意関係のある 2 文を肯定的ペアとし、矛盾関係のある 2 文を強い否定的ペアとなるような損失関数を用いることで，文埋め込み性能が向上している．

3 提 案 手 法

ここでは，提案手法である慣用句を使用した文と慣用句を用いない同じ意味の文ペアを，正例とする対照学習により，慣用句を的確に表現可能な文埋め込みの獲得手法について詳しく説明する．まず，3.1 節で慣用句の捉え方について説明する．3.2 節では，提案手法の基礎となる対照学習，教師なし SimCSE および教師あり SimCSE について説明し，3.3 節では，慣用句を考慮した提案手法の内容について詳述する．

3.1 慣用句の捉え方

慣用句は，習慣として長い間広く使われてきたひとまとまりの言い回しのことである．

通常，慣用句には，その文字からのみでは推測が難しい、固有の意味が割り当てられている．例えば,「腹を割る」という慣用句は，固有の意味として「本心を打ち明ける．隠さずに心の中をさらけ出す.」という意味を持ち,【例文 2】のように用いら

れる．

【例文 2】：特に<腹を割って>話す．

また，一部の慣用句については、慣用句の意味ではなく、文字通りの意味としても使用されるものが存在する (※ 1)．例えば,「あぐらをかく」という慣用句は，慣用句として「のんきにかまえて，何の努力もしないことのたとえ.」という固有の意味を持ち,【例文 3】のように用いられる．反対に，慣用句ではなくその字面通りに「あぐらをかく」という意味も持ち,【例文 4】のように用いられる．

【例文 3】：ブランドに<あぐらをかく>のではなく，ユーザに目を向けた製品を期待する．

【例文 4】：床に<あぐらをかいて>本を読む．

本稿では，上記 2 種類の慣用句を，

- タイプ 1 の慣用句: ※ 1 に該当しない慣用句．慣用句固有の意味でのみ使用されるもの
- タイプ 2 の慣用句: ※ 1 に該当する慣用句．慣用句固有の意味で使用されるだけでなく，慣用句を構成する字面の文字通りの意味でも使用される場合があるもの

と区別して呼ぶこととする．OpenMWE コーパスに収録されている全ての慣用句についての用例文，101,500 件を確認した結果，慣用句固有の意味で使用された用例数と，慣用句を構成する字面の文字通りの意味で使用された用例数は，およそ 2 : 1 の比率で存在している．

3.2 対照学習と教師なし SimCSE，教師あり SimCSE

対照学習とは，意味的に近いデータ同士を近くに，そうでないデータ同士を引き離すことによって，効果的なデータの表現 (埋め込み) を学習することを目的とした手法である．対照学習では，意味的に近いデータのペア x_i と x_i^+ の集合 $\mathcal{D} = \{(x_i, x_i^+)\}_{i=1}^m$ が与えられることを想定している．$\boldsymbol{h}_i$, $\boldsymbol{h}_i^+$ を，それぞれデータ x_i と x_i^+ の埋め込みとすると，N 文のミニバッチにおける (x_i, x_i^+) に対する損失関数 (Normalized Temperature CrossEntropy, NT-Xent) は，

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^+)/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^+)/\tau}} \tag{1}$$

で表される．

ここで，τ は温度ハイパーパラメータであり，$\mathrm{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2)$ は，コサイン類似度を表す:

$$\mathrm{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2) = \frac{\boldsymbol{h}_1^\top \boldsymbol{h}_2}{||\boldsymbol{h}_1|| \cdot ||\boldsymbol{h}_2||} \tag{2}$$

入力文は，事前学習済み BERT などのモデルでエンコード $\boldsymbol{h} = f_\theta(x)$ され，式 (1) の損失関数を用いて学習可能なパラメータが微調整される．

次に，教師なし SimCSE について説明する．教師なし SimCSE は，教師データを必要としない SimCSE 手法で，同一文に対して異なるドロップアウト マスクを施した 2 つの埋め込みを正例ペアと見做して対照学習を行う．文の集合 $\{x_i\}_{i=1}^m$ において，$x_i^+ = x_i$ とし，x_i と x_i^+ に対して異なるドロップアウト マスクを用いて埋め込みを取得する:

$$\boldsymbol{h}_i^z = f_\theta(x_i, z) \tag{3}$$

ここで，z はドロップアウトに対するランダムマスクを表す．同一の入力文 x_i を 2 回エンコーダに入力することで，異なるドロップアウト マスク z_i, z_i' が適用された 2 つの埋め込み $\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'}$ を取得することができる．N 文のミニバッチにおいて，ある文 x_i に対する損失関数は，

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'})/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j'})/\tau}} \tag{4}$$

で表される．

ここでも同様に，τ は温度ハイパーパラメータであり，$\mathrm{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2)$ は，コサイン類似度である．

続いて，教師あり SimCSE について説明する．教師あり SimCSE は，教師データを用いることで正例ペアのアライメントをより改善する SimCSE 手法で，例えば，NLI データセット (自然言語推論データセット) などから直接 (x_i, x_i^+) ペアを取得し，これを用いて式 (1) を最適化する対照学習を行う．

(x_i, x_i^+) ペアの構成方法としては，NLI データセットを用いた場合は，2 文の関係として「含意」ラベルが付与されているペアを正例として学習する方法が性能が高いことが示されている [20].

さらに，正例のみではなく，強い負例 (hard negative) を利用するように拡張することもできる．例えば，NLI データセットにおいて x_i を前提，x_i^+ を「含意」ラベルが付与された仮説，x_i^- を「矛盾」ラベルが付与された仮説としたとき，教師データを (x_i, x_i^+) から (x_i, x_i^+, x_i^-) へ拡張することができる．N トリプレットのミニバッチにおける (x_i, x_i^+, x_i^-) に対する損失関数は，

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^+)/\tau}}{\sum_{j=1}^{N} \left(e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^+)/\tau} + e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^-)/\tau} \right)} \quad (5)$$

で表される．NLI データセットの場合，強い負例を追加したことでさらに性能が向上したことが示されている [20].

本稿では，日本語の慣用句も含めて適切な文埋め込みを獲得する手法として，対照学習を適用する．提案手法では，慣用句を含む文を x_i とし，x_i を慣用句を用いない表現で言い換えた文を x_i^+ とした (x_i, x_i^+) を正例ペアとして学習し，強い負例は扱わない．そのため，N 文のミニバッチにおいて，(x_i, x_i^+) を正例ペアする損失関数は，

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j^+)/\tau}} \quad (6)$$

で表される．

4 提案手法: 慣用句を考慮した対照学習による文埋め込みの獲得

本稿では，日本語の慣用句も含めて適切な文埋め込みを獲得する手法を提案する．対照学習を適用するため，日本語の慣用句を用いた文と慣用句を用いない同じ意味の文のペアを正例ペアとする慣用句データセットを構築する (構築手順は第 4 節で説明する)．ベースラインの事前学習済みモデルとして BERT を用いる (以降，BERT(baseline) と呼ぶ).

まず，教師なし SimCSE による改善効果を確かめるため，日本語 Wikipedia の文を収録したデータセットを用いて BERT(baseline) に対して，教師なし SimCSE を適用する手法を考える (以降，BERT(simcse_u) と呼ぶ)．日本語 Wikipedia の文データセットは，Hugging Face Hub に公開されているデータセット [3]を使用する．表 1 に，使用する日本語 Wikipedia データセットの統計値を示す．なお，日本語 Wikipedia 文データセットには，慣用句を用いた表現はほとんど含まれていない (図 1(a))．また，構築した慣用句データセットを用いて BERT(baseline) に対して，教師なし SimCSE(式 (4 の損失関数を使用)) を適用する手法を考える (以降，BERT(simcse_u') と呼ぶ)．加えて，日本語 Wikipedia 文データセットを用いた教師なし SimCSE と，慣用句データセット (+日本語 Wikipedia 文データセット) を用いた教師なし SimCSE(式 (4 の損失関数を使用)) を，順に適用する手法を考える (以降，BERT(simcse_uu') と呼ぶ).

次に，教師あり SimCSE による改善効果を確かめるため，慣用句データセットを用いて BERT(baseline) に対して，教師あり SimCSE(式 (6) の損失関数を使用) を適用する手法を考える (以降，BERT(simcse_s) と呼ぶ)．ここで，慣用句を含む文を x_i とし，x_i を慣用句を用いない表現で言い換えた文を x_i^+ とした (x_i, x_i^+) を正例ペアとして対照学習を行う．ここで，負例としては，慣用句データセットから，x_i とは異なる慣用句を用いた文をランダムに選択したものを採用し，ミニバッチを構成した (図 1(b)).

提案手法としては，日本語 Wikipedia 文データセットを用いた教師なし SimCSE と，慣用句データセット (+日本語 Wikipedia 文データセット) を用いた教師あり SimCSE(式 (6) の損失関数を使用) を順に適用した手法を考える (以降，BERT(simcse_us) と呼ぶ).

表 1　日本語 Wikipedia データセットの統計値

データ総件数	文字数の算術平均	文字数の標本標準偏差
5,121,625	43.5	22.6

5 データセット構築

本節では，慣用句データセットおよび慣用句ベンチマークデータセットの構築について説明する．まず，4.1 節では，構築の基礎となっている OpenMWE コーパスについて説明する．次に，4.2 節では，慣用句データセット，および慣用句ベンチマークデータセットの構築について，手順に沿って説明する．

5.1 OpenMWE コーパス

OpenMWE コーパス [21] とは，日本語慣用句のデータセットであり，計 146 種類の慣用句の可能性のある表現を用いた用例 102,334 文が収録されている．各文には，慣用句の意味で使用されていることを表す “I”(idiomatic)，文字通りの意味で使用されていることを表す “L”(literal) のいずれかのラベルが付与されている．また，慣用句の可能性のある各表現には，複数の用例が収録されている．表 2 に，OpenMWE コーパスの統計値を示す.

5.2 慣用句データセットおよび慣用句ベンチマークデータセットの構築

5.2.1 手順 1: 慣用句と用例の抽出

まず，OpenMWE コーパスから，本稿で対象とすべき慣用句とその用例を抽出する．本稿では，文中に出現する慣用句と思われる表現が，慣用句の意味で用いられる場合を問題とするため，OpenMWE コーパスにおいて “I” タグが付与されている用例のみを使用することとする．また，各慣用句には複数の用例が収録されているが，その用例数は一定ではなく，極端に少ない用例数も見受けられる．そのため，各慣用句の用例数の平均値を算出し，その平均値以上の用例数をもつ慣用句のみを抽出し，慣用句データセットとして採用することとする．その結果，各慣用句の用例数の平均は 462 件となり，その平均以上の用例数をもつ慣用句は 74 種類となった．さらに，3.1 節で述べたタイプ 1, タイプ 2 に該当する慣用句はそれぞれ 0 件，74 件となった．表 3 に，抽出した慣用句コーパスの統計値を示す.

3：https://huggingface.co/datasets/AhmedSSabir/Japanese-wiki-dump-sentence-dataset

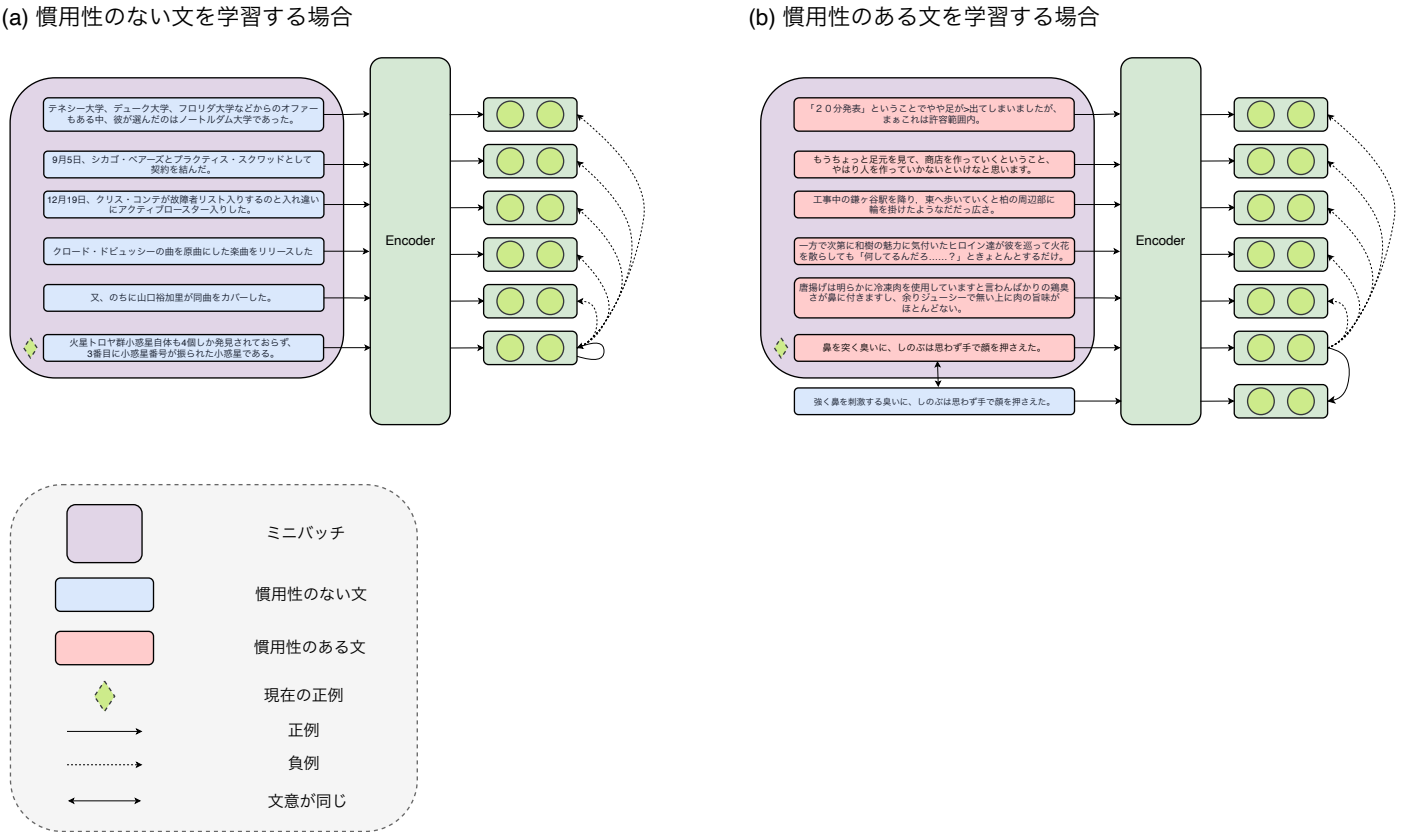


図 1　本提案手法の概念図

表 2　OpenMWE の統計値

ラベル	データ件数	文字数の算術平均	文字数の標本標準偏差
“I” (idiomatic)	67,575	85.4	47.6
“L” (literal)	33,925	72.1	42.1
“I” (idiomatic)+“L” (literal)	101,500	81.0	46.2

表 3　抽出した慣用句コーパスの収録データ件数

用例数が平均値以上の慣用句の件数	用例文数の平均値	抽出した用例文の総数
74	464	34,336

5.2.2　手順 2: 慣用句の意味の収集

慣用句を用いた用例は取得できたものの，OpenMWE コーパスには，慣用句の意味は収録されていない．そこで，別途オンライン辞典等から慣用句の意味を収集する (手順 2)．まず，オンライン辞典である goo 国語辞典から，該当する慣用句の意味を収集した．goo 国語辞典に収録されていない慣用句については，goo 国語辞典に収録されている同義の別の慣用句 (慣用句は異なるが同じ意味の慣用句) が存在する場合，それに置き換えた．この置き換えは，あらかじめ人手で参照辞書を作成することで実現した．

5.2.3　手順 3: 慣用句を用いた用例から用いない用例への書き換え

手順 2 で収集した慣用句の意味を用いて，手順 1 で取得した慣用句を用いた用例文を，慣用句を用いない同じ意味の文に書き換え，正例ペアを作成する．この書き換え作業については，研究室在籍の学生 12 名に作業を行ってもらい，一部作業結果が不十分なものについて別途クラウドソーシングサービスを利用することで実施した．

5.2.4　手順 4: 慣用句データセットの整形

手順 3 で得られた正例ペアの中には，慣用句以外の部分が長く，冗長な表現となっている用例がかなり多いことが判明した．そこで，Gemini(Gemini Pro) API を使用し，その冗長な部分を削除することで，簡潔な内容の文へと整形する．具体的には，慣用句を含む文 (idiomatic) と，その慣用句を同義となる表現に言い換えた文 (paraphrase)，その慣用句を字面通りに使用した文 (literal) から成るトリプレットを作成する．各文において関連性はなく，独立して処理するため，文ごとに新規チャットを立ち上げてトリプレットを作成した．この作業の結果，現時点で 23,595 件のトリプレットで構成される慣用句データセットを構築した．ここでは，Gemini の温度ハイパーパラメータの値を 0.7 に設定した．この温度ハイパーパラメータの値は，最大値である 1.0 に近いほど，自由度が高い多様な表現での文

生成が可能となり，最小値である 0.0 に近いほど，自由度が低い固定された表現の文生成が可能になる．

5.2.5 手順 5: 慣用句ベンチマークデータセットの構築

JSTS ベンチマークデータセットは，意味的なテキスト類似 (Semantic Textual Similarity) を日本語で評価するためのベンチマークデータセットである．しかし，そこに含まれる文は，あくまで一般的に使用される文がほとんどであり，慣用句や慣用性のある表現が含まれる文の割合は非常に低い．従って，慣用句を考慮した意味的なテキスト類似を評価するためには，新たなベンチマークデータセットを構築する必要がある．これを人手で作成するのは高コストであるため，ここでは，手順 4 で作成したトリプレットを参考にして，Gemini を使用し，新たなトリプレットを作成することとした．この作業の結果，現時点で 23,594 件のトリプレットで構成される慣用句ベンチマークデータセットを構築した．ここでは，Gemini の温度ハイパーパラメータの値を 0.7 に設定した．

6 実　　験

本節では，提案手法の性能をベースラインとの比較実験を通して評価する．まず，5.1 節では，研究課題を明示した後，5.2 節で，実験設定について説明する．5.3 節では，実験結果を示し，5.4 節で，考察を述べる．

6.1 研究課題

日本語の慣用句を含めたデータセットと学習方法の組み合わせの違いにより，提案手法がベースラインと比べてどの程度優れた文埋め込みを獲得できるかを明らかにする．

6.2 実験設定

本節では 3.3 節で述べたベースライン，提案手法および各比較手法で得られた文埋め込み結果を比較する．ここでは，事前学習済みモデルとして，東北大学が公開している BERT[4]を使用した．

ここでは，該当するトークン列の平均で取得する場合を，BERT(baseline, avg) と表記する．表 4 に，ベースライン，提案手法および各比較手法をまとめた表を示す．

ベンチマークデータセットとしては，JSTS データセットおよび慣用句ベンチマークデータセットを用いる．文同士の類似性の評価尺度としては，スピアマンの順位相関係数を使用する．

JSTS データセットは，日本語における言語理解ベンチマーク JGLUE [22] に含まれる意味的テキスト類似 (STS; Semantic Textual Similarity) を評価するためのベンチマークデータセットである．このベンチマークデータセットには，文のペアと，その文同士がどの程度意味的に類似しているかを示す実数値スコアが 0.0 から 5.0 の範囲で付与されている．意味的に類似していない文ペアであるほどスコアは低く，意味的に類似しているほどスコアが高くなる．このデータセットには，慣用句はほとんど含まれていないが，通常の日本語における文の意味的な類似性を評価するために用いることとする．

慣用句ベンチマークデータセットは，4.2.5 節で説明したトリプレット (慣用句を含む文 (idiomatic)，その慣用句を用いずに同義となるように言い換えた文 (paraphrase)，慣用句を字面通りに使用しているが意味は異なる文 (literal) の 3 つ組) の集合で構成されるベンチマークデータセットである．このベンチマークデータセットは，慣用句を含む場合の文の意味的な類似性を評価するために用いる．

6.3 実験結果

6.3.1 JSTS ベンチマークデータセットを使用した実験結果

表 5 に，BERT の該当トークン列の出力の平均における，JSTS ベンチマークデータセットを使用した場合の，ベースラインモデルと提案手法を用いて学習させたモデルに対する評価結果を示す．

表 5　JST ベンチマークを用いた場合の各手法による文埋め込みの比較 (avg)

Model	スピアマン順位相関係数
BERT(baseline, avg)	0.74
BERT(simcse_u, avg)	**0.80**
BERT(simcse_u', avg)	0.76
BERT(simcse_uu', avg)	**0.80**
BERT(simcse_s, avg)	0.76
BERT(simcse_us, avg)	**0.80**

6.3.2 慣用句ベンチマークデータセットを使用した実験結果

表 6 に，BERT の該当トークン列の出力の平均における，慣用句ベンチマークデータセットを使用した場合の，ベースラインモデルと提案手法を用いて学習させたモデルに対する評価結果を示す．

表 6　慣用句ベンチマークを用いた場合の各手法による文埋め込みの比較 (avg)

Model	スピアマン順位相関係数
BERT(baseline, avg)	-0.52
BERT(simcse_u, avg)	-0.58
BERT(simcse_u', avg)	0.035
BERT(simcse_uu', avg)	-0.025
BERT(simcse_s, avg)	**0.14**
BERT(simcse_us, avg)	0.10

表 5 より，BERT の該当トークン列の出力の平均における，JSTS ベンチマークデータセットでは，BERT(simcse_u, avg) と BERT(simcse_uu', avg)，および BERT(simcse_us, avg) の手法が，最もスコアが高い結果となった．慣用句データセットを用いた教師なし SimCSE，および教師あり SimCSE を適用した手法のスコアは，ベースラインのスコアよりも高かった．

4：https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

表 4　比較手法の一覧表

BERT(baseline, avg)	事前学習済み BERT
BERT(simcse_u, avg)	事前学習済み BERT ＋教師なし SimCSE(Wikipedia)
BERT(simcse_u', avg)	事前学習済み BERT ＋教師なし SimCSE(慣用句データセット)
BERT(simcse_uu', avg)	事前学習済み BERT ＋教師なし SimCSE(Wikipedia) ＋教師なし SimCSE(慣用句データセット)
BERT(simcse_s, avg)	事前学習済み BERT ＋教師あり SimCSE(慣用句データセット)
BERT(simcse_us, avg)	事前学習済み BERT ＋教師なし SimCSE(Wikipedia) ＋教師あり SimCSE(慣用句データセット)

BERT の CLS トークンの出力のみにおける場合と同様に，これらの手法で共通している大量の Wikipedia 文を用いた教師なし SimCSE の効力が大きいと示唆される．

表 6 より，BERT の該当トークン列の出力の平均における，慣用句ベンチマークデータセットでは，BERT(simcse_s, avg) の手法において，最もスコアが高かった．しかし，慣用句ベンチマークデータセットに収録されている文ペアは，人間の解釈とは異なる場合もあることから，慣用句ベンチマークデータセットにおいて，一概に高スコアであることが望ましいとは言えない．ただ，この結果から，教師あり SimCSE の手法がモデルの性能に影響を与えていることが示唆される．そのため，人間の解釈に沿った慣用句ベンチマークデータセットを用意して，再度検証する必要がある．

以下に，5.2.5 節で示した Gemini を用いたベンチマークデータ作成において，プロンプト入力時の良い出力結果の例，および悪い出力結果の例を示す．

【 慣用句「鼻に付く」における良い出力結果の例 】
{
“idiomatic”: “グローバルは１年位前に使っていましたが，あのきな粉クサイのが鼻に付いて”,
“paraphrase”: “グローバルは１年位前に使っていましたが、あのきな粉の臭いが鼻につきまとって嫌になりました”,
“literal”: “グローバルは１年位前に使っていましたが、あのきな粉が鼻に付いて、くしゃみが出ました”
}

【 慣用句「鼻に付く」における悪い出力結果の例 】
{
“idiomatic”: “ちなみにタイムはほんのちょっと加えただけでもかなり鼻に付きます。”,
“paraphrase”: “ちなみにタイムはほんのちょっと加えただけでもかなり飽きて嫌になります。”,
“literal”: “ちなみにタイムはほんのちょっと加えただけでもかなり匂います。”
}

上記の Gemini の良い出力例では，“idiomatic” の文である「グローバルは１年位前に使っていましたが，あのきな粉クサイのが鼻に付いて」は，慣用句「鼻に付く」を慣用性がある文であることがわかる．慣用性のない文を意味する “literal” の文は，文字通りの意味として解釈できるため，良い例として挙げた．反対に，Gemini の悪い出力例では，“idiomatic” の文である「ちなみにタイムはほんのちょっと加えただけでもかなり鼻に付きます。」は，慣用句「鼻に付く」を慣用性がある文であることがわかる．しかし，慣用性のない文を意味する “literal” の文が，“idiomatic” の文と文意が同じであり，人間の解釈と一致しないため，悪い例として挙げた．これらの良い例，および悪い例より，慣用句の種類によって “literal” 文の生成が難しいと示唆される．以上より，慣用句ベンチマークデータセットに収録されている文ペアは，人間の解釈とは異なる場合もあることから，人間の解釈に沿ったベンチマークデータセットとなるように，再度構築し直す必要がある．

7 ま と め

本稿では，慣用句を考慮した文埋め込みの獲得手法を提案した．従来手法では，その単語列の文字通りの意味との区別が必ずしも明確になされていない．そこで本稿では，まず日本語の慣用句を用いた文と慣用句を用いない同じ意味の文とをペアとする慣用句データセットを構築した．慣用句を用いた文と慣用句を用いない文を正例とした対照学習により，慣用句を的確に表現可能な文埋め込みを獲得する手法を提案した．実験では，本稿での提案手法の有用性は示せていないが，本提案手法が文埋め込みの性能に影響を与えていることが示唆された．しかし，Gemini は慣用性のない慣用句表現を使用した用例文を作成するのが難しいことにより，評価に使用した慣用句ベンチマークデータセットには適切ではない用例文が含まれる．そのため，今後は Gemini などの対話型生成 AI も適切に活用することで，慣用句ベンチマークデータセットを構築し直し，再度本提案手法の有効性を検証する必要がある．

謝　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　献

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota),

pp. 4171–4186, Association for Computational Linguistics, June 2019.

[2] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 55–65, Association for Computational Linguistics, Nov. 2019.

[3] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 9119–9130, Association for Computational Linguistics, Nov. 2020.

[4] J. Su, J. Cao, W. Liu, and Y. Ou, "Whitening sentence representations for better semantics and faster retrieval," 2021.

[5] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, and A. Villavicencio, "Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), (Online), pp. 2730–2741, Association for Computational Linguistics, Aug. 2021.

[6] G. Salton, R. Ross, and J. Kelleher, "Idiom token classification using sentential distributed semantics," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 194–204, Association for Computational Linguistics, Aug. 2016.

[7] P. Liu, K. Qian, X. Qiu, and X. Huang, "Idiom-aware compositional distributed semantics," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (M. Palmer, R. Hwa, and S. Riedel, eds.), (Copenhagen, Denmark), pp. 1204–1213, Association for Computational Linguistics, Sept. 2017.

[8] M. King and P. Cook, "Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (I. Gurevych and Y. Miyao, eds.), (Melbourne, Australia), pp. 345–350, Association for Computational Linguistics, July 2018.

[9] H. Haagsma, *A Bigger Fish to Fry: Scaling up the Automatic Understanding of Idiomatic Expressions*. PhD thesis, University of Groningen, 2020.

[10] R. Hashempour and A. Villavicencio, "Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions," in *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (M. Zock, E. Chersoni, A. Lenci, and E. Santus, eds.), (Online), pp. 72–80, Association for Computational Linguistics, Dec. 2020.

[11] Z. Zeng and S. Bhat, "Idiomatic expression identification using semantic compatibility," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1546–1562, 2021.

[12] V. Nedumpozhimana and J. Kelleher, "Finding BERT's idiomatic key," in *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)* (P. Cook, J. Mitrović, C. P. Escartín, A. Vaidya, P. Osenova, S. Taslimipoor, and C. Ramisch, eds.), (Online), pp. 57–62, Association for Computational Linguistics, Aug. 2021.

[13] D. Phelps, "drsphelps at SemEval-2022 task 2: Learning idiom representations using BERTRAM," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, eds.), (Seattle, United States), pp. 158–164, Association for Computational Linguistics, July 2022.

[14] Z. Zeng and S. Bhat, "Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1120–1137, 2022.

[15] V. Nedumpozhimana, F. Klubička, and J. D. Kelleher, "Shapley idioms: Analysing bert sentence embeddings for general idiom token identification," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[16] 良. 高橋, 遼. 笹野, and 浩. 武田, "文脈化単語埋め込みを用いた慣用句判定," Tech. Rep. 4, 名古屋大学大学院情報学研究科, 名古屋大学大学院情報学研究科, 名古屋大学大学院情報学研究科, sep 2020.

[17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, 2006.

[18] F. Carlsson, A. C. Gyllensten, E. Gogoulou, E. Y. Hellqvist, and M. Sahlgren, "Semantic re-tuning with contrastive tension," in *International Conference on Learning Representations*, 2021.

[19] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), (Online), pp. 879–895, Association for Computational Linguistics, Aug. 2021.

[20] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 6894–6910, Association for Computational Linguistics, Nov. 2021.

[21] C. Hashimoto and D. Kawahara, "Compilation of an idiom example database for supervised idiom identification," *Language Resources and Evaluation*, vol. 43, no. 4, pp. 355–384, 2009.

[22] 栗. 健太郎, 河. 大輔, and 柴. 知秀, "Jglue: 日本語言語理解ベンチマーク," **自然言語処理**, vol. 30, no. 1, pp. 63–87, 2023.

命題論理における言語の構成性に着目した言語モデルの汎化能力の調査

井上　綾介† 　宮森　恒†

† 京都産業大学　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2153128,miya}@cc.kyoto-su.ac.jp

あらまし　近年，言語モデルは人間のような流暢な言語生成能力を示している一方で，論理的な推論に関しては依然として未知の問題に汎化することが困難である．そこで，本研究では，言語モデルの汎化を目指す上で重要な要素である言語の構成性に着目し，命題論理に基づくデータセット構築とそれを用いた言語モデルの汎化能力の調査をする．具体的には，まず，否定 (¬)，連言 (∧)，選言 (∨), 含意 (→) それぞれに基づく文章を作成することでデータセットを構築し，言語モデルに論理演算の基礎知識を学習させる．次に，得た知識を用いて多段の推論を要するタスクを用意し，言語モデルが既知の概念を組み合わせてどの程度複雑な推論を行えるかを調査する．本研究により，言語モデルの分布外汎化を実現するための知見が得られることが期待される．

キーワード　言語モデル，汎化能力，構成性，命題論理，推論

1 はじめに

人工知能技術の進展と共に発展してきた言語モデルは近年, 翻訳や質問応答, 言語生成などの多くのタスクで人間に近い精度を達成している．また, Chain-of-Thought [1] のようなプロンプティングを用いることにより, ある程度複雑な多段推論を行えることも報告されている．

一方で, 言語モデルは既知の概念や知識を組み合わせる必要がある論理的な推論への対処が難しいことが報告されている [2]. これは言語モデルが学習データの分布外のデータ分布に対する問題に適切に対処できていないことが大きな要因とされている. そのため, 昨今では言語モデルの分布外への汎化のために重要であるとされている構成性に着目をした汎化能力の調査が進められている．構成性とは, ある物事を構成している要素を特定し, それらの要素を組み合わせることで物事を理解できるということを指す．言語モデルは学習の際にデータの表層的な情報に影響を受けやすいことが問題とされており, 人間が行うような複雑な推論を可能にするためには構成性が重要な役割を果たすと考えられている．例えば, Yanaka ら [3] は構成性のうち体系性に焦点を当てて, 自然言語推論の能力を分析する手法を提案した.

しかし, 言語モデルの構成性に関する研究としては, 自然言語の意味解析に着目をしたものが多く, 人間が行う重要な推論の一つである論理推論における構成的な汎化については焦点が当てられていない．論理推論能力の獲得は人工知能研究の長年の目標であり, 大量のテキストデータによって訓練された言語モデルが有する論理推論能力については未だに不明な点が残っている．そのため, 言語モデルがどの程度複雑な論理推論を行えるかを把握することは後の言語モデルの発展に繋がると考えられる.

そこで本研究では, 命題の複雑さを制御したデータセットを構築することで言語モデルの論理推論における構成性を調査する．学習データは論理演算子である否定 (¬)，連言 (∧)，選言 (∨), 含意 (→) や中置記法における括弧の役割を言語モデルに学習させることを目的とする．テストデータは, 論理演算子の組み合わせや命題の深さ, または命題変数の文字列が学習データには含まれていない未知のデータとし, これを用いて構成性のうち体系性と代入可能性に焦点を当てて言語モデルの構成性を評価する．構築したデータセットを使った実験において, 体系性は (i) 未知の組み合わせと (ii) 未知の深さの 2 点に着目をして評価する．言語モデルの代入可能性は, 命題に含まれる命題変数の文字列を未知の羅列に変化させて言語モデルが変わらず推論を行えるかを観察することで評価をする．また, 言語モデルの Attention の可視化によって, 言語モデルによる命題の構造把握と体系性に対する汎化の関係について追加の調査を行った.

本論文の主な貢献は, 命題論理における言語モデルの構成性を評価する方法の提案, 現行の言語モデルの推論における汎化能力の特定, そして分布外汎化で重要となる要素の特定の三点である.

本論文の構成は以下の通りである．まず, 第 2 節では本研究と関連が深い構成性調査並びに命題論理に関する従来研究について紹介をする．第 3 節では主に命題論理における言語モデルの構成性を調査するためのデータセットの構築について説明する．第 4 節で, 構築したデータセットを用いた実験とその結果, 考察をまとめ, 第 5 節で本論文の結論及び課題について述べる.

2 関連研究

本節では言語モデルと構成性に関連する研究として, まず 2.1 節に言語の構成性に着目した研究を紹介する．続く 2.2 節では言語モデルと命題論理に焦点を当てた研究について紹介する.

2.1 構成性調査に関する研究

Fodor ら [4] によると, 人間は John loves Mary という英語の文章を理解していれば, Mary loves John のように同じ構成

要素から異なる文章を生成することができる．このような人間の認知能力の性質を近年の言語モデルも有するのかを検証するために幾つかの先行研究によってデータセットが構築された．SCAN [5] は簡単な自然言語で記述された指示を行動シークエンスに変換する (例. jump twice → JUMP JUMP) というタスクによって構成されたデータセットである．これにより, 言語モデルが学習時に学んだ指示を組み合わせて, 未知の指示を変換できるかを検証することができる．しかし, SCAN で使用されている言語表現の種類が限られているため, このデータセットを拡張したものとして COGS [6] が提案されている．COGS は SCAN よりも豊富な構文を含む, 自然言語の文法的な構造への汎化を調査するために構築されたデータセットである．他にも, 言語モデルの構成性を調査するためのデータセットとして, 事前に学んだルールを適用して物語の登場人物間の親族関係の推論するタスクによって構成された CLUTRR [7] が提案された．

また, 構成性のうち体系性に注目をしてニューラルモデルの自然言語推論能力を調査した研究 [3] や, 意味解析においてモデルが体系的に学習できているのかを分析する研究 [8] がある．これらの研究は, どちらも現行のモデルは体系的な学習を苦手とすることを示した．また, 基礎的な算術演算を言語モデルに学習させた後, 段階的に複雑さを制御したタスクを用いることで言語モデルの構成性を評価した研究 [9] でも同じく, 体系性への汎化が最も困難であることが示された．しかし, 上記のデータセットや分析は論理的な推論の基礎である命題論理に着目はされておらず, いずれも意味解析の文脈における構成性の着目に留まっている．

2.2 命題論理に関する研究

Traylor ら [10] は命題論理の論理演算子がもつ意味を言語モデルが学習によって獲得できているかを調査する研究を行った．この研究は, 言語の形式にのみ着目をして訓練された言語モデルが, どの程度までその言語の意味を捉えることができるのかを明らかにするため比較的扱いやすい命題論理に焦点を当てて分析を行った．その結果, 命題変数の数や深さなどが学習データと同じ命題など, 構文が似ているが意味が異なる論理演算子の区別することできなかった．このことから筆者らは, 言語モデルにおける記号の意味の違いは, その記号が用いられる文脈に強く依存すると推測した．他にも, 言語モデルに命題論理を効果的に学習させる手法を提案した研究 [11] などがあるが, これらの命題論理に関する研究はいずれも未知の問題への汎化に重要とされている構成性には着目されていなかった．

3 提 案 手 法

本節ではまず 3.1 節で本論文が扱う問題設定について説明し, 3.2 節で作成したデータセットについて説明した後, 3.3 節でデータセットの構築方法について述べる．

3.1 問 題 設 定

本研究では, 論理演算子を用いて記述された次のような命題の真偽を正しく判断することを目標とする．

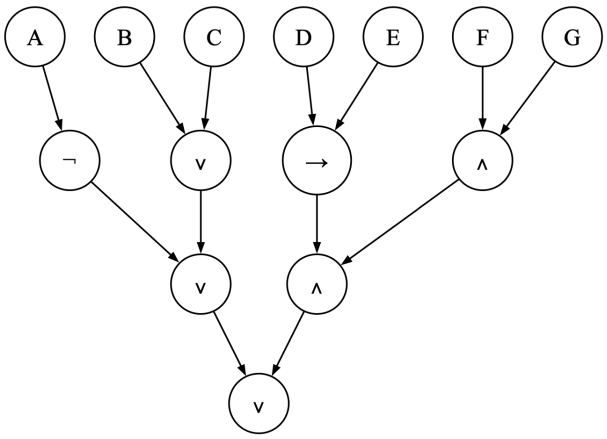


図 1 深さ 3 の命題 (¬ A ∨ (B ∨ C)) ∨ ((D → E) ∧ (F ∧ G)) の木構造．ここで, アルファベットの A から G は命題変数とする．

「問題文: ”A is True, B is False, A ∧ B” 答え: False」

問題文で用いられる論理演算子は否定 (¬), 連言 (∧), 選言 (∨), 含意 (→) の 4 つの中から選ばれる．また, 答えは True または False のいずれかである．言語モデルの命題論理における構成性を明らかにするために, 用いる論理演算子の数や命題の深さ, また命題変数の文字列などを段階的に変化させた問題文を用意する．ここで深さとは, 命題を木構造にした際の深さとする (図 1).

続いて, 本研究で着目する構成性について説明をする．本研究では構成性のうち, 体系性と代入可能性の 2 つに焦点を当ててデータセットを構築し, 分析をする．体系性は, 既に知っている表現を組み合わせることでより複雑な表現も理解することができることを指す [12]．例えば「茶色の犬」と「黒色の猫」を理解していれば「茶色の猫」も同様に理解できることを指す．本研究では (i) 未知の組み合わせと (ii) 未知の深さの 2 つの観点で体系性を評価する．

代入可能性とは, ある表現における構成要素を同じ意味をもつ別の構成要素に置き換えても, その表現の意味には影響しないことを表す [12].

3.2 データセットについて

データセットにおける入力と出力は 3.1 節で示したように, 入力は問題文, 出力は答え (真または偽) とする．ここで, 答えが真の場合には 1 を, 偽の場合には 0 を出力するようにした．また, 学習データに含まれる命題は 2 つ以下の命題変数によって構成されるようにした．

学習データ: はじめに, 深さ 1 の命題の集合を学習データ 1 とする．命題論理の論理演算子である否定 (¬), 連言 (∧), 選言 (∨), 含意 (→) のいずれか一つを用いて作成した深さ 1 の命題の集合である学習データ 1 は, 4 つの論理演算子それぞれの役割を言語モデルに学習させるためのものである．使用する命題変数はアルファベット順に小文字の a から z, 大文字の A から D までとして作成した 10,000 個の命題を学習データ 1 とする．

表 1　提案手法に基づいて生成した命題の例. ここでは生成された深さ 1 から 4 までのランダムな命題を示してある. ここで, ランダムな小文字アルファベットは命題変数を表すものとする. また, 冗長な括弧が存在しているが, 命題の真偽には関わらないためここでは無視する.

深さ	命題
1	¬ (a), (g ∨ f), (c → b), (k ∨ e)
2	((h ∧ c) → (c → h)), ¬ ((a ∧ b)), ((c ∨ d) ∨ (d ∧ d)), ((e ∨ b) ∧ (e ∨ b))
3	(((i ∨ n) → (n ∨ i)) ∨ ((n ∧ i) ∧ ¬ (i))), ¬ (((m ∧ j) ∨ ¬ (j)))
4	¬ ((¬ ((i → e)) ∨ ((e ∧ i) ∨ (i → i))))

続いて, 深さ 2 の命題を学習データ 2 とする. 学習データ 2 は括弧の役割と, 否定が優先度の高い演算子であることを学習させることを目的として構築する. まず, 括弧の役割を学習させることについて説明する. 中置記法において括弧で囲まれている演算は他の演算よりも優先的に処理される. 例えば, A, B は真, C は偽を表す命題変数とした時に次の 2 つの命題,「(A ∨ B) ∧ C」と「A ∨ (B ∧ C)」では命題の真偽が異なるため, 言語モデルにこの括弧の働きを学習させる必要がある. 続いて, 否定の優先度に関する学習について述べる. 命題論理において, 1 つの命題に否定とその他の論理演算子が同時に現れる場合には, 否定が優先的に処理される. これら 2 つの命題論理に関するルールを言語モデルに学習させるために, 学習データ 2 における命題として否定と連言の 2 つの演算子を用いた深さ 2 の命題を用意する. 学習データ 2 に含まれている命題は 10,000 個で, 用いる命題変数は小文字の a から z, 大文字の A から D までとした.

ここで, 学習データ 2 において用いる論理演算子を否定と連言のみに限定している理由は, 学習に最低限必要なデータによってモデルを訓練し, 学習データに含まれていない論理演算子の組み合わせをテストデータとして言語モデルを評価するためである.

次に, 学習データの例を示す. 上の 2 つは学習データ 1 の深さ 1 の命題, もう 1 つは学習データ 2 の命題である.

```
{
  "text": "n is True, ¬ (n)",
  "cls": "0"
},
{
  "text": "j is False, i is True, (j ∧ i)",
  "cls": "0"
},
{
  "text": "i is False, g is False, (¬ (i) ∧ ¬ (g))",
  "cls": "1"
}
```

テストデータ: はじめに, 体系性を調査するためのテストデータについて説明をする. 体系性は (i) 未知の組み合わせと (ii) 未知の深さに着目をして評価する. また, テストデータの各命題に含まれる命題変数の個数は学習データと同様に最大で 2 となるようにした.

(i) 未知の組み合わせに言語モデルが汎化できるかを検証するために, 学習データに含まれていない 2 種類または 3 種類の論理演算子の組み合わせ (例. 連言と含意) を用いた深さ 2 の命題を用意する. ここでは未知の組み合わせにのみ着目をするため, 深さは学習データと同じ 2 とした. 2 種類の論理演算子の組み合わせは, 学習データ 2 に含まれる否定と連言の組み合わせを除いて 5 通りある. 3 種類の論理演算子の場合は, 4 通りの組み合わせが存在する. これら各組み合わせに対し, それぞれ 200 個のテストデータ, 合計で 1,800 個のデータを用意して言語モデルの未知の組み合わせにおける体系性を評価する.

(ii) 未知の深さの命題は, 学習データに含まれる命題について深さのみを変化させたものを用いる. そのため, 1 つの命題が含む論理演算子の組み合わせや命題変数を表す文字列は学習データに登場したものに限定する. 具体的には, 否定, 連言, 選言, 含意のいずれか 1 つの論理演算子を用いた命題 (学習データ 1), または否定と連言の 2 つの論理演算子を使った命題 (学習データ 2) の深さを変化させたものをテストデータとして用いる. 1 種類の論理演算子のみを用いるものに関しては深さは 2 から 5 とし, 各深さに対して 1 つの論理演算子につき 500 個のデータを用意する. ただし, 否定のみの命題に関してはパターンが限られており, 各深さに対して 28 個の命題のみ用意した. 2 種類の論理演算子 (否定, 連言) を用いるものは, 変化させる深さを 3 から 5 とし, 同じく各深さに対して 500 個のデータを用意して評価する.

最後に, 代入可能性を調査するためのテストデータについて述べる. 命題論理において代入可能性を検証するために, 命題における命題変数の文字列を変化させる. このとき, 本来であれば命題変数を表す文字列が他の文字列に置き換わっても命題の真偽には影響しないため, 言語モデルがこの変化に対処できるかを観察することで代入可能性を検証することができる. テストデータとしては, 学習データ 1 並びに 2 で取り扱った命題における命題変数の文字列を学習データには見られなかったものに変更する. 文字列の長さによる汎化性能の違いについても検証するため, 命題変数の文字列の長さは 1 から 10 まで変化させる. 各文字数に対しそれぞれ 300 個ずつ計 15,000 個のデータを用意した.

3.3 構 築 方 法

データセットは文脈自由文法の生成規則に従って生成した命

題を用いて構築する．例として，否定，連言，選言，含意の 4 つの論理演算子と命題変数 A, B を用いて命題を生成する文脈自由文法の生成規則を下に示す．

$$S \to A \mid B \mid \neg S \mid S \wedge S \mid S \vee S \mid S \to S \mid (S)$$

命題を生成するのに用いる論理演算子を限定する場合には，該当する箇所を上の生成規則から取り除けば良い．生成された命題の例を表 1 に示す．

4 実　　験

本節では，命題の基礎について学習した言語モデルが未知の問題にどの程度汎化できるかを体系性と代入可能性に焦点を当てて検証する．

4.1 実験設定

学習データ 1 と学習データ 2 の和集合である 20,000 件のデータを用いて言語モデルの学習を実行し，用意したテストデータを用いて学習済み言語モデルの命題論理における 2 つの体系性と代入可能性を検証する．

モデル: 事前学習済みの言語モデルである BERT [13] のうち base モデルを対象として調査する．

パラメータチューニングについて: 学習率は 2.0^{-5}，バッチサイズは 128 とした．

評価: 命題の答えが真のときは 1，偽のときは 0 を出力させ，二値分類の正答率で評価する．

4.2 実験結果

体系性 (i) 未知の組み合わせ:

表 2　体系性 (i) 未知の組み合わせ 評価結果の表．

組み合わせ	正答率
¬→	42.0%
∧→	50.5%
∧∨	50.0%
∨¬	45.5%
∨→	48.5%
∧¬→	53.5%
∧∨→	50.0%
∧∨¬	50.0%
∨¬→	48.5%

未知の組み合わせのテストデータを用いて学習済み言語モデルを評価した結果を表 2 に示す．この結果から，全ての組み合わせにおいて正答率が 50%に近くになっていることから，言語モデルは未知の組み合わせに対応できておらず，殆どランダムな回答を行っていることが読み取れる．

体系性 (ii) 未知の深さ:

表 3 に学習させた言語モデルに対して未知の深さの命題を与えた際の正答率を示した表とグラフを示す．この表から，否定を用いた命題を除く全てのデータに対して，おおよそ 50%の正答率となっていることから言語モデルが未知の深さの命題を適切に捉えられていないことが分かる．また，表 3 や否定以外の論理演算子を用いた命題の深さと正答率の関係を示した図 2 より，深さが深くなるに従って正答率が 50%に近づいていることも見て取れる．否定に関しては，深さの値が偶数の場合は正答率が 0%，奇数の場合は 100%となっていることから，学習データ 1 に含まれる深さ 1(奇数) の命題に影響を受けていることが考えられた．これらのことから，言語モデルが命題の未知の深さに汎化していないことが示された．

表 3　体系性 (ii) 未知の深さ 評価結果の表．

深さ \ 演算子	¬	∧	∨	→	∧¬
2	0.0%	53.1%	45.8%	56.2%	-
3	100.0%	51.1%	49.2%	50.2%	58.3%
4	0.0%	50.4%	49.7%	50.0%	54.0%
5	100.0%	50.6%	48.8%	50.1%	48.1%

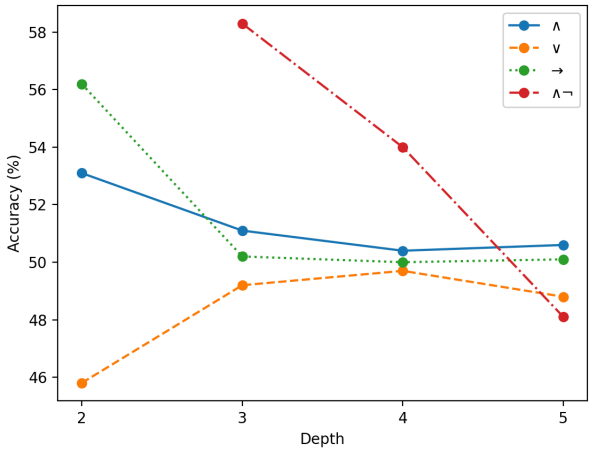


図 2　体系性 (ii) 未知の深さ 評価結果のグラフ (否定以外)．

代入可能性:

表 4　代入可能性 評価結果の表．

文字数 \ 演算子	¬	∧	∨	→	∧¬
1	100.0%	100.0%	97.4%	76.5%	100.0%
2	100.0%	74.2%	100.0%	76.8%	83.2%
3	100.0%	73.5%	94.8%	76.8%	72.7%
4	77.8%	73.5%	83.0%	76.5%	68.9%
5	61.1%	74.5%	59.5%	76.1%	63.4%
6	55.6%	73.5%	43.5%	73.9%	61.2%
7	50.0%	73.5%	40.8%	73.5%	63.4%
8	50.0%	73.5%	28.1%	73.9%	64.7%
9	50.0%	73.5%	27.8%	73.5%	70.9%
10	50.0%	73.5%	27.5%	74.2%	70.9%

これらの表 4 と図 3 からは，命題変数が 5 文字程度なら概ね汎化できることが分かる．しかし，文字数がそれより多くなると正答率が低下することから代入可能性も完全な汎化は困難であることが示された．

命題論理に着目したこれらの結果から，言語モデルの意味解析や算術演算タスクに着目した先行研究で示された結果と同じく，言語モデルの体系性への汎化が困難であるということが示唆された．

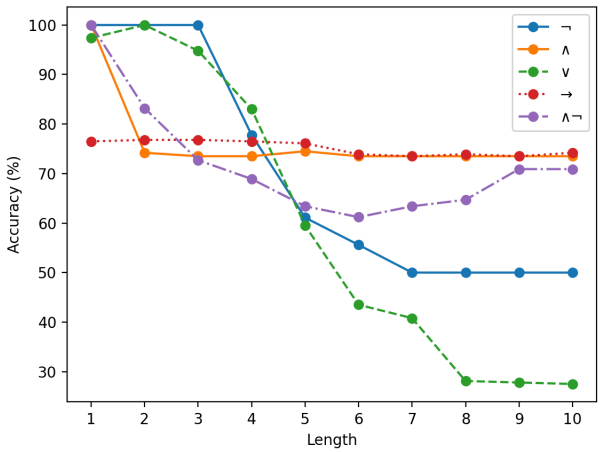


図 3　代入可能性 評価結果のグラフ.

言語モデルによる構造の把握について:

体系性と代入可能性の実験結果から言語モデルが特に未知の組み合わせや深さの命題を扱うことを苦手とするということが分かった. そこで, 言語モデルにとって特に難しいとされる体系性の汎化に関する追加の調査を行った. まず, 未知の組み合わせや深さの命題を解くに当たって, 命題の構造の把握が重要になる. 具体的には, 体系性の汎化の際には括弧で囲まれた箇所の処理を正しく行うことが鍵となる. 追加の調査では, 学習データ 1, 2 の 20,000 件のデータで学習済みの言語モデルが問題文として与えられた命題のどの箇所を注視して処理を行っているかを確認することで, 言語モデルが命題の構造を正しく捉えられているかを検証する. そのために, Vig によって考案された BERT モデルの Attention を可視化する手法, BERTVIZ [14] を使用した. BERTVIZ の詳しい仕組みについては Vig の論文 [14] を参照されたい. 追加調査では, 学習データ 1 と 2 のデータで学習済みの言語モデルに対して, 正解できる問題文と不正解となる問題文の 2 パターンのデータを与えた際の注目箇所を比較する. 正解できる問題文として, 学習データ 1 と 2 に含まれる深さ 1 と深さ 2 の命題をそれぞれ 1 つずつ, 不正解となる問題文としては体系性 (i), (ii) のテストデータの中からテスト時に不正解となったものを 1 つずつ代表として抽出する. 下に調査対象とした問題文を示す. 上から学習データ 1, 2, 体系性 (i), (ii) のデータとなっている.

```
{
  "text": "j is False, i is True, (j ∧ i)",
  "cls": 0
},
{
  "text": "i is False, g is False, (¬ (i) ∧
  ¬ (g))",
  "cls": 1
},
{
    "text": "z is True, u is True, ((u ∨ u) ∨
    (u ∧ z))",
    "cls": 1
},
{
    "text": "a is False, ¬ (¬ (a))",
    "cls": 0
}
```

以上 4 つのデータを用いて BERT における Attention の可視化を行って得られた結果を図 4 から図 14 に示す. また, 着目する層は最も構造の把握に寄与していると予想される 10 層目を対象とした. また, それぞれの図は可視化されたもののうち, 命題の箇所に絞ったものである.

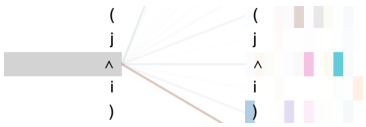


図 4　学習データ 1 の命題の可視化図. 論理演算子∧に着目.

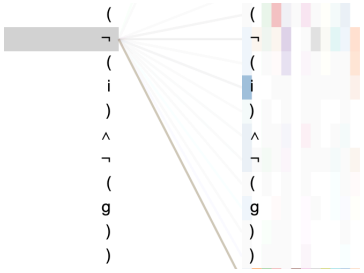


図 5　学習データ 2 の命題の可視化図. 1 つ目の論理演算子¬に着目.

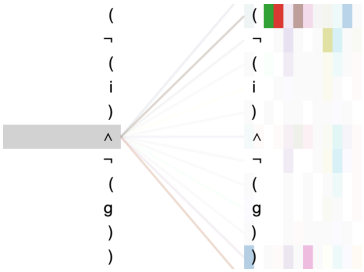


図 6　学習データ 2 の命題の可視化図. 論理演算子∧に着目.

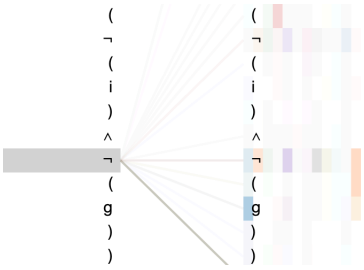


図 7　学習データ 2 の命題の可視化図. 2 つ目の論理演算子¬に着目.

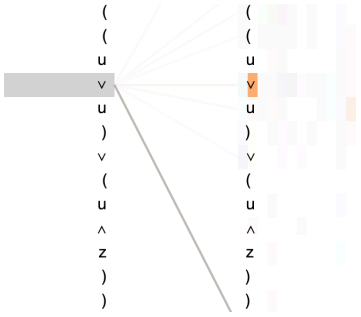


図 8　テストデータ体系性 (i) の命題の可視化図．1 つ目の論理演算子∨に着目．

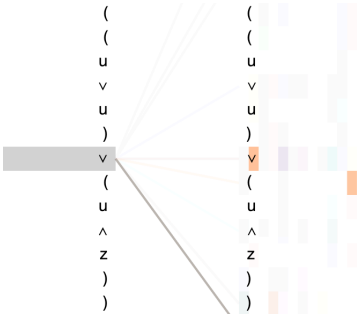


図 9　テストデータ体系性 (i) の命題の可視化図．2 つ目の論理演算子∨に着目．

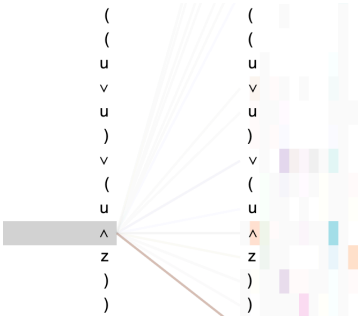


図 10　テストデータ体系性 (i) の命題の可視化図．論理演算子∧に着目．

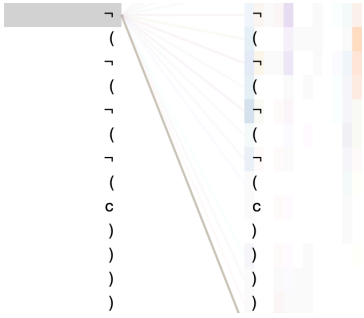


図 11　テストデータ体系性 (ii) の命題の可視化図．1 つ目の論理演算子¬に着目．

命題論理の基礎について学習済みの言語モデルが正解する問題文の可視化の結果を示した図 4 から図 7 からは，命題に含まれる論理演算子に着目した際に，その論理演算子が処理する範囲に相当する括弧の部分に注意が向けられていることが分かる．また，否定に着目をした図 5 と図 7 からは，括弧よりも命題変

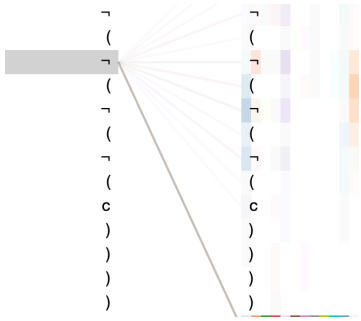


図 12　テストデータ体系性 (ii) の命題の可視化図．2 つ目の論理演算子¬に着目．

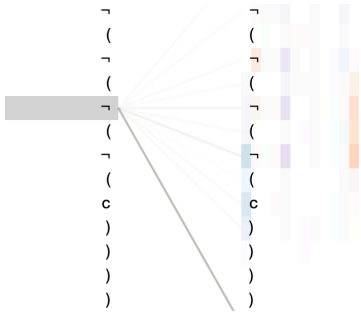


図 13　テストデータ体系性 (ii) の命題の可視化図．3 つ目の論理演算子¬に着目．

図 14　テストデータ体系性 (ii) の命題の可視化図．4 つ目の論理演算子¬に着目．

数に注意が向けられていることが読み取れる．これらのことから，言語モデルは正しい出力を行える問題文における命題の構造を適切に把握できていることが推測される．

続いて，言語モデルが苦手とする体系性をテストするための命題について述べる．図 8 から図 14 を見ると，正解する問題文の際の注意とは異なり，論理演算子が担当する処理の範囲が把握できていないことが読み取れる．図 14 の否定の処理は，図 7 と同様に命題変数に注意が向けられており，正しく構造を把握していると考えられる．しかし，命題変数から離れた位置にある論理演算子が注意する先は，命題変数ではなく後に続く論理演算子であり，正確に範囲の特定が行えていないことも推測される．

以上の結果は，本研究で学習させた言語モデルが体系性に対する汎化が困難である理由は，命題の構造の把握を適切に行えていないことが原因であるということを示唆している．

5 おわりに

本稿では, 現状の言語モデルの構成性を明らかにするために命題論理に基づくアプローチにより, 構成性のうち体系性, 代入可能性を検証するためのデータセットを構築し, 分析を行った. 学習データは深さ 1 と 2 の命題に絞り, 命題論理の基礎知識を学習させることを目標とした. 体系性は未知の組み合わせと深さの 2 つに焦点を当てて評価を行い, 代入可能性は命題変数の文字数を増やした問題文を与えることで評価を行った. その結果, 命題論理の基本的な知識を有する言語モデルは代入可能性に比べて体系性に対する汎化が特に困難であることが明らかになった. また, 命題を解く上で重要な要素となる命題の構造を言語モデルが把握しているかを検証するために追加の調査も行った. 得られた結果からは, 未知の組み合わせや深さの命題に関しては正しく構造を捉えられていないことから, 命題の構造を言語モデルに学習させることが体系性への汎化に近付く鍵となるのではないかと推測した.

今後はより大きなモデルや異なる構造をもつモデルに対して同様の実験を行い, モデルの種類が未知の問題への汎化にどの程度影響するのかを検証する必要がある. また, 命題論理のみならず様々なタスクで有効である構造の把握を言語モデルに効果的に学習させる手法についても調査を行っていきたい.

謝　辞

本研究の一部は科研費 23K11342 の助成を受けたものである.

文　献

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[2] Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*, 2023.

[3] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. Do neural models learn systematicity of monotonicity inference in natural language? *arXiv preprint arXiv:2004.14839*, 2020.

[4] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[5] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[6] Najoung Kim and Tal Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*, 2020.

[7] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

[8] 谷中 瞳, 峯島 宏次, and 乾 健太郎. ニューラルネットが学習する意味表現は体系性を持つか. **言語処理学会 第 *27* 回年次大会**, 2021.

[9] Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Ana Brassard, Masashi Yoshikawa, Keisuke Sakaguchi, and Kentaro Inui. Do deep neural networks capture compositionality in arithmetic reasoning? In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1351–1362, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[10] Aaron Traylor, Roman Feiman, and Ellie Pavlick. And does not mean or: using formal languages to study language models' representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 158–167, 2021.

[11] Anthony Tomasic, Oscar J Romero, John Zimmerman, and Aaron Steinfeld. Propositional reasoning via neural transformer language models. 2021.

[12] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

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

[14] Jesse Vig. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*, volume 23, 2019.

金融数学問題解法における ChatGPT による言語処理と記号処理機能の統合
ー演繹推論法によるプロンプト・エンジニアリングー

白田　由香利† 　桑名　杏奈‡ 　Basabi Chakraborty[1,2]

† 学習院大学経済学部 〒171-8588 東京都豊島区目白 1-5-1
‡ 和洋女子大学全学教育センター 〒272-8533 千葉県市川市国府台 2-3-1
1 マダナパレ工科大学計算機学部 Madanapalle Chittoor, Andhra Pradesh 517325, India
2 岩手県立大学 〒020-0611 岩手県滝沢市巣子 152-52

E-mail: † yukari.shirota@gakushuin.ac.jp, ‡ a-kuwana@wayo.ac.jp , 2 basabi@iwate-pu.ac.jp

あらまし　本研究では，ChatGPT に金融数学文章題を解かせ，教育用グラフィックスを生成させることの可能性を探った．ChatGPT は統計的手法に基づく AI でありながら，大規模な言語モデルトレーニングから派生した推論能力を活用して，金融数学の問題を，公式を組み合わせ演繹的に解くことが可能である．この能力は，記号処理のための AI 機能を組み込むことでさらに強化され，基本的問題に対しては完璧な解法を提供する．さらに，ChatGPT は，その演繹推論過程をグラフとして描画するので，教育用のビジュアル教材を効率良く生成することが可能となる．この論文では，現段階での成果について報告する．

キーワード　生成 AI，大規模言語モデル，記号処理的 AI，演繹推論法，金融数学

1. 始めに

我々は，経済経営数学教育現場での ChatGPT 等の生成 AI の活用を研究している[1], [2], [3]．研究目標は以下の 2 つである：

1. 生成 AI を生徒役とすることで，人間教師の教授スキルを向上させること．
2. 問題解法および学生向け可視化教材の自動生成．

上記目標 1 のような，生成 AI に何かの役を演じさせて人間の訓練の相手をさせる試みは既に実現している．例えば，ホテル等の接客業で，激高する客を再現する生成 AI が新人研修等で使われている[4]．我々の研究チームは，生成 AI に生徒役を演じさせ，人間教師の教育力向上に役立てるシステム構築を考えている．

2 つ目の研究目標は，生成 AI が正しい解法を実現した際，その頭の中で行われた演繹推論を可視化し，それを学生への教材として活用することである．これにより，高品質の教材が大量に自動生成可能となる．小中高の数学ドリル本は大量に出版されているが，大学以上の数学ドリル問題は少ない．白田は，10 年前演繹推論法による教材自動生成システムを構築したが[5], [6], [7]，生成 AI を活用することで，当時に比較して低コストで実現可能となった．実際に，我々のチームでは，学習院大学経済学部経営学科の「経営数学」の講義において，生成 AI に作成させた演繹推論プラングラフを教材として用いている(図 1 参照)．我々は現在 ChatGPT 4.0 を用いている．

我々の研究対象は，文章題の解を求めるタイプの経済数学問題である．証明問題は対象としない．生成 AI に求められるスキルは，**(1)文章題読解力，(2)演繹推論力，(3)数学公式などの知識，(4)計算力**となる．演繹推論とは，一般的な原理や法則から特定の事実や結論を導き出す論理的思考のプロセスでる[8], [9].

AI に関しては，「**統計的手法**」と「**記号的手法**」という二分法の考えが続いてきた[10]．ChatGPT は統計的手法に基づく AI であり，単独では数学問題解法において誤りをおかすことがある．ChatGPT の計算力の低さについては，Mathematica の創設者である Wolfram は早い段階から指摘し，「ChatGPT のようなシステムも外部ツールを検討する必要があり，その一例が Wolfram 言語である」と統合による能力向上について述べていた[10]．そして Wolfram cloud [11]がプラグインとして ChatGPT 4.0 から利用可能となり，Mathematica の計算力及び公式に関する知識が現在利用可能となった．Wolfram cloud プラグインの設定は[12]を参照のこと．他方，(1)文章題読解力，(2)演繹推論力，については大規模言語モデルが担当する．本稿では，金融数学問題解法における ChatGPT による言語処理と記号処理機の統合について論じる．次節は，演繹推論法という数学教授法について説明する．第 3 節は事例として，外国債券のプライシング問題を取り上げる．第 4 節では，演繹推論プラン等の教材作成につ

いて論じる．第 5 節はまとめとする．

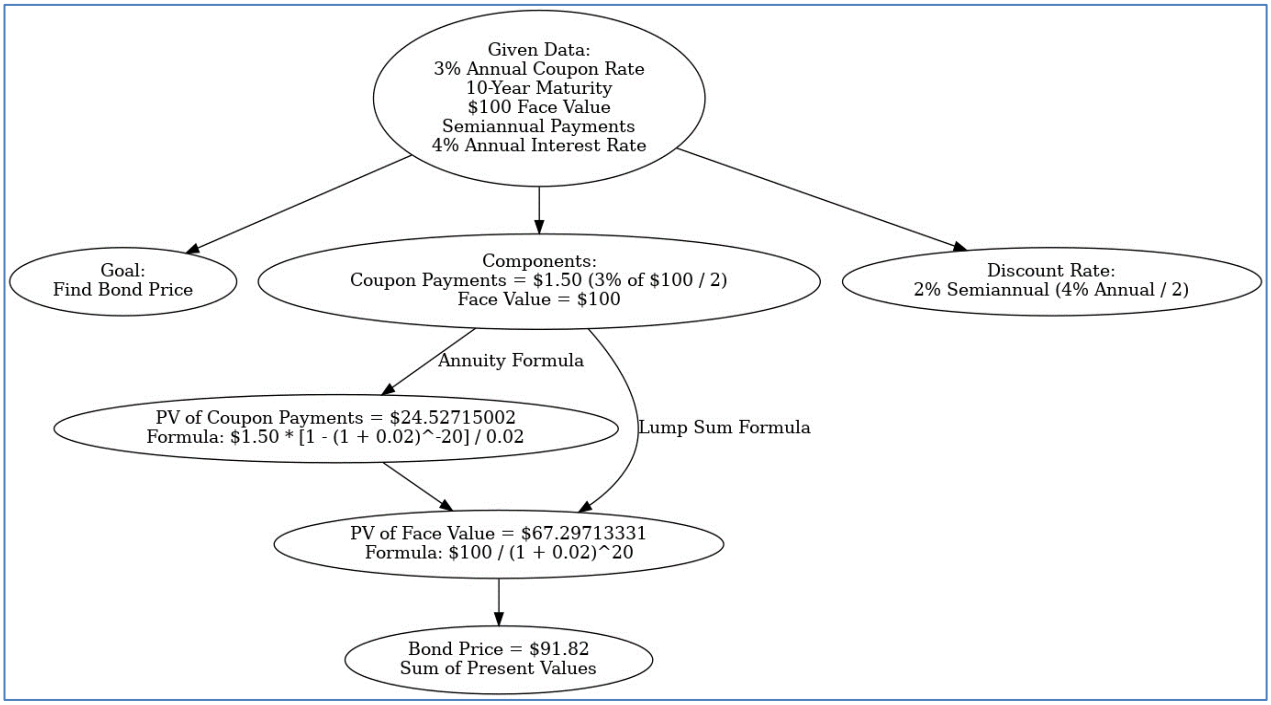


図 1: 演繹推論パスグラフの例

2. 演繹推論法

本節では数学問題教授法である演繹推論法を説明する.事例として図 1 に挙げた債券価格問題を用いる．

演繹推論法では，まず始めに学生に文章題が与えられる．教師は，学生にその文章題を読解して，Given Data を抽出し，Given Data をリストアップさせる．同様に未知数 Unknown が何であるかを書かせる.ここまではポリアの How To Solve It の教授法である[13]．図 1 の外債プライシングの問題では，以下のようになる．

Given data: クーポン率 c=3%，額面=100USD，
満期=10 年，満期利回り率 r=4%，
半年ごとのクーポン支払い．
Unknown: 債券価格 PV

演繹推論パスでは，Given Data をグラフのトップに並べ，Unknown をグラフのボトムに並べると，上から下に演繹が進められるので分かり易い(図 1 参照)．次に，Given Data から Unknown までのミッシングリンクを，公式を用いて演繹推論してつないでいく．結果として，Unknown の値が求まる．図 1 の問題の演繹推論のポイントは，クーポン支払いの現在価値(Present Value, PV)を求める以下の公式を知っているか否かである．

$$PV = \sum_{i=1}^{n} \frac{100 * \frac{c}{2}}{\left(1+\frac{r}{2}\right)^{i}} + \frac{100}{\left(1+\frac{r}{2}\right)^{n}}$$

図 1 を見ると，ChatGPT はクーポン PV の部分の計算公式として，等比数列の和の公式を以下のように変形した後の式を公式として使っている．

$$\sum_{i=1}^{n} \frac{100*\frac{c}{2}}{\left(1+\frac{r}{2}\right)^{i}} = \left(\frac{100*\frac{c}{2}}{1+\frac{r}{2}}\right) * \frac{\left(1-\left(1+\frac{r}{2}\right)^{-n}\right)}{1-\left(1+\frac{r}{2}\right)^{-1}} =$$

$$\frac{100*\frac{c}{2}\left(1-\left(1+\frac{r}{2}\right)^{-n}\right)}{\frac{r}{2}}$$

しかし，教材として提示する場合は，変形前の形の方が意味がよく分かるので望ましい．

3. ChatGPT による可視化教材自動生成事例

本節では，ChatGPT に自動生成させた可視化教材を説明する．

生成 AI の記号処理的 AI の能力は，Wolfram cloud plugin 及び Open AI の Advanced Data Analysis を使うことで充足可能である．ChatGPT4.0 には，ChatGPT をカスタマイズする機能があり，我々は経済経営数学問題を演繹推論法で解くことに特化させた”MATHMATH”という名の ChatGPT を現在構築中である.MATHMATH を利用する最大の利点は，新しい公式を覚えさせることができる点である．経済ルールや公式で，ChatGPT がまだ知らない公式を，識別子(例えば F1)をつけて覚えさせる．MATHMATH を再起動したとき，プロンプトで「公式 F1 を使って，この問題を解いてください」と指示を出すことで，以前教えた公式を使って問題を

解かせることができる.大規模言語モデルの訓練では,新しい公式を発見して使えるようになるまで膨大な時間がかかり,不可能な場合もあるが,MATHMATH は覚えていてくれる.

実験的に,間違った内容の公式を覚えさせようと繰り返し教えたが,大規模言語モデルで学んだであろう正しい公式が優先された.この点は信頼がおける.悪意で生成 AI に間違った公式を教えようとする人間がいた場合,どう回避させるかは今後の重要課題のひとつである.また,「同じ数学問題を解く別解が多数ある場合,解法に要するコストが最小の解を選択すべきである」という基本的方針をいかに生成 AI に教えるか[13]ということも,今後の課題である.

以下では,ChatGPT による可視化自動生成の事例を紹介する.題材は,外国債券数学における価格計算とする.債券数学の内容について以下を参照して頂きたい[14], [15]. MATHMATH には以下の指示(プロンプト)を与えた.

###

Find the bond price of a 3% coupon annually, 10-year maturity bond with face value of $100 paying semiannual coupon payments.

Suppose that the interest rate is 4% annually.

###

ChatGPT は日本語よりも英語の読解能力のほうが高いので,対話言語は英語とした.MATHMATH は,演繹推論によって経営数学問題を解くための専用 ChatGPT であるので,上記ではあえて「Given Data と Unknown をリストアップせよ」「演繹推論をせよ」と指示していない.途中で有効桁数が減るので,有効数字 8 桁で計算するようには,指示を出した.こうした債券価格を求める問題は,パターン化されている頻出問題である.ChatGPT は,パターン化された問題は容易に解くことができる.換言すると,読解力を駆使し,Given Data と Unknown を抽出し,その間をつなぐため演繹推論を行い,連立方程式を設定することができる.正しく演繹推論ができ,記号処理 AI に正しい連立方程式を渡すことができれば,後の作業は記号処理 AI が実行してくれる.

リクエスト先：WOLFRAM

```
{
  "input": "couponPayment = 4; discountRate = 0.015; numberOfPayments = 20; parValue = 100; PVofCoupons = couponPayment * (1 - (1 + discountRate)^(-numberOfPayments)) / discountRate; PVofPar = parValue / (1 + discountRate)^numberOfPayments; PVofBond = PVofCoupons + PVofPar; PVofBond"
}
```

図 2: Wolfram cloud に渡された連立方程式の例

記号処理 AI として Wolfram cloud を使う場合,シームレスに Wolfram cloud が呼び出される.呼び出しの際の変数名の設定が適切でないと,Wolfram 言語の変数名と予約語のルールでエラーが生じることがあるが,最近ではこのエラーがほぼなくなった.図 2 に Wolfram cloud に渡された Wolfram 言語のプログラム例を提示した.

クーポンの PV 合計を計算する公式は変形後に以下となった.

$$\frac{100 * \frac{c}{2}\left(1-\left(1+\frac{r}{2}\right)^{-n}\right)}{\frac{r}{2}}$$

ChatGPT は,この式を公式として連立方程式をたてるので,本問題の場合,Wolfram cloud へ渡されるコマンドは以下のような式となる.

```
PVofCoupons = counponPayment * (1 - (1 + discountRate)^(-numberOfPayments)) / discountRate
        = 1.5 * (1 - (1 + 0.02)^(-20)) / 0.02
```

しかし,教材としては,元の等比数列の和の部分が見えにくいので,問題である.変形前のほうが望ましい.つまり,ChatGPT が自動生成した演繹推論パスそのまま

分析

python　Copy code

```
# Further updating the graph to include formulas used for

dot = graphviz.Digraph(comment='Deductive Reasoning for Bo

# Adding nodes with values and formulas
dot.node('A', 'Given Data:\n3% Annual Coupon Rate\n10-Year
dot.node('B', 'Goal:\nFind Bond Price')
dot.node('C', 'Components:\nCoupon Payments = $1.50 (3% of
dot.node('D', 'Discount Rate:\n2% Semiannual (4% Annual /
dot.node('E', 'PV of Coupon Payments = $24.52715002\nFormu
dot.node('F', 'PV of Face Value = $67.29713331\nFormula: $
dot.node('G', 'Bond Price = $91.82\nSum of Present Values'

# Adding edges
dot.edges(['AB', 'AC', 'AD'])
dot.edge('C', 'E', label='Annuity Formula')
dot.edge('C', 'F', label='Lump Sum Formula')
dot.edges(['EF', 'FG'])

# Render the updated graph with formulas
dot.render('/mnt/data/deductive_reasoning_graph_formulas',
'/mnt/data/deductive_reasoning_graph_formulas.png'
```

図 3: 演繹推論パスグラフを描画するための graphviz を使った Python プログラム

までは教材品質に問題があり,改善すべき点が少なか

らずある．

Advanced Data Analysis を使う場合は，方程式を解くために，Python プログラムを生成し，実行して解を求める．図 1 のような演繹推論パスグラフを描画せようとすると，MATHMATH はシームレスに Advanced Data Analysis を呼び出し，図 3 のような graphviz を使ったプログラムを生成し，実行して図 1 を描いた．Wolfram cloud でも Advanced Data Analysis 両者ともプロンプトで指示するだけで，目的のプロットが得られる(図 4 参照)．プロンプトとして”First show me the expression of bond price with parameter r.” を与えると，MATHMATH が以下の式を返す．

Bond Price(r)=PVA(r)+PV(r)
=($1.50×(1−(1+r/2)^−20)/(r/2)+($100/(1+r/2)^20)

その後に以下のプロンプトを与えることで，このプロットを得た．

Please plot the Bond Price(r).
Please add the r=4% mark on the plot.

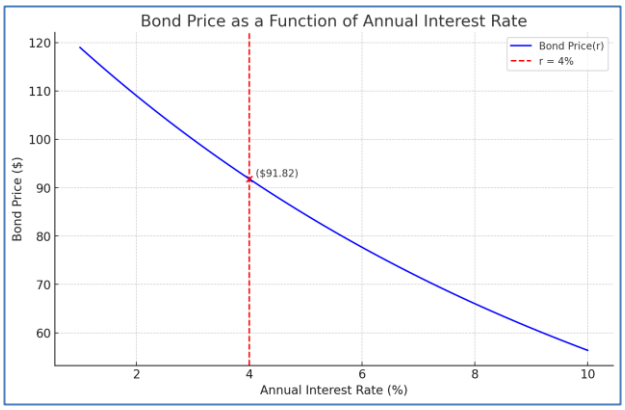


図 4：ChatGPT が描画した価格利回り曲線

次に MATHMATH に 2 変数関数 BondPrice(r, c)を描画させた(図 5 参照)．さらに CouponRate c=4%のところにインターセクションを入れさせるには，以下のプロンプトで動いた．

Add the curve on the 3D plot with coupon rate=4%.

このように，描画作成の Python プログラムを自動生成し，それを実行させて画像生成までシームレスで行なわせることが可能となった．そして，債券の価格問題のような定型的な問題では，グラフィックス教材が大量生産可能となる．

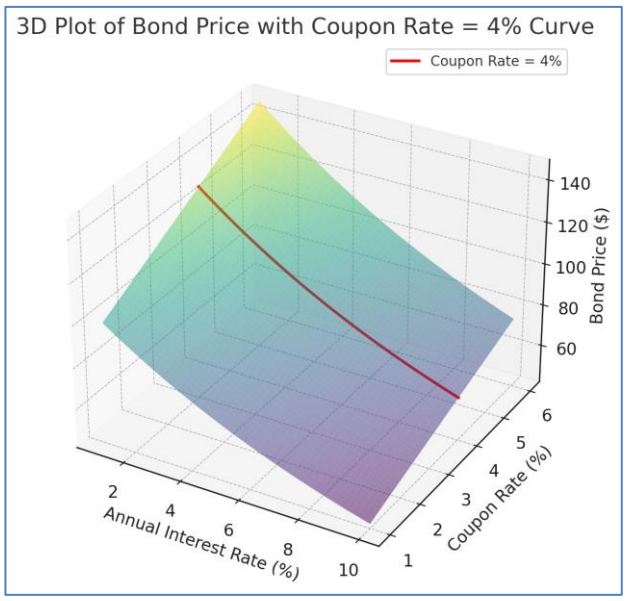


図 5: ChatGPT に描画させた BondPrice(r, c)

まとめると，定型的な経営問題については，ChatGPT は正しく演繹推論が行える．正しく演繹推論が行えれば，正しい連立方程式が設定でき，解を求めることができる．そして，その演繹推論過程に具体的数値を入れながら演繹解法パスグラフを自動生成してくれる．ChatGPT が演繹推論において間違えた場合でも，解法パスグラフを描画させることで，ChatGPT の頭の中が分かるので，エラーの原因を突き止めることが容易となる．

4. まとめ

本稿では，ChatGPT に金融数学文章題を解かせ，教育用グラフィックスを生成させることの可能性を探った．演繹推論法により金融数学問題を解かせる場合，統計的手法 AI の特性が必要なスキルとして(1) 文章題の読解力，及び(2)演繹推論力，が挙げられる．また，記号処理的 AI の特性が必要となるスキルとして(3)数学公式などの知識，及び(4)計算力，が挙げられる．(3)の公式の知識は，大規模言語モデルが習得したものも多い．記号処理的 AI として，Wolfram cloud 及び OpenAI Advanced Data Analysis 等を用いることで，2 種類の AI のスキルを統合化させ，金融数学問題を解くことができるようになる．その成功事例として，金融数学問題で頻出する債券価格問題を取り上げた．こうした頻出する数学問題は，ChatGPT は簡単に解くことができる．その頭の中の演繹推論パスを表示させた．この可視化は Python プログラムで graphviz ライブラリを用いた．このプログラムもシームレスに自動生成し，出力画像の生成まで行ってくれる．我々の経営数学の講義では，この出力された演繹推論パスを実際に教材として用い

ている．大学の経済数学では，高校までの数学に比較して圧倒的にドリル教材が少ないが，こうした可視化教材の自動生成により，学生に多数のドリル問題を提供可能となる．今後とも，ChatGPT に演繹推論法で経済数学を教えるプロジェクトを続ける予定である．

謝辞

本研究は部分的に，学習院大学 GEM プロジェクト 2023 年度，学習院大学計算機センター2023 年度特別プロジェクトの助成によるものである．

参　考　文　献

[1] 白田由香利 and バサビ･チャクラボルティ,"対話による ChatGPT への経済数学教育 ～ 誤りモデルの分析と指導プロセスの最適化 ～," 電子情報通信学会 ヒューマンコミュニケーション基礎研究会（HCS）, 2023.

[2] 白田由香利 and 桑名杏奈, "ChatGPT を活用した金融数学問題の解法と教材作成," *信学技報*, vol. 123, no. 327, pp. 12-17, 2023.

[3] 白田由香利, "推論によるデータ分析における ChatGPT の活用," *學習院大學經濟論集*, vol. 60, no. 3, pp. 179-195, 2023.

[4] 日経新聞, "「カスハラ」客，生成 AI で再現　離職防止へ研修活用スタートアップが開発," *日経新聞*, 2023.

[5] 白田由香利, "経営数学問題解法における演繹推論に関する考察," *學習院大學經濟論集*, vol. 49, no. 2, pp. 85-98, 2012.

[6] 白田由香利 and 橋本隆子, "経営数学問題解決法のための演繹推論支援教材作成ツール," *學習院大學經濟論集*, vol. 48, no. 4, pp. 303-311, 2012.

[7] P. Stanworth and Y. Shirota, "Review of the "WebHow2SolveIt" Website," *Gakushuin Economics Papers*, vol. 50, no. 1, pp. 1-18, 2013

[8] 井上克巳, "アブダクションとインダクション (<特集> 論理に基づく推論研究の動向)," *人工知能*, vol. 25, no. 3, pp. 389-399, 2010.

[9] 米盛裕二, *アブダクション: 仮説と発見の論理* 勁草書房, 2007.

[10] S. Wolfram, *What Is ChatGPT Doing... and Why Does It Work?* Stephen Wolfram, 2023.

[11] W. Research, "ChatGPT 用 Wolfram プラグイン," 2023. [Online]. Available: https://www.wolfram.com/wolfram-plugin-chatgpt/

[12] ところてん, *ChatGPT 攻略* KADOKAWA, 2023.

[13] G. Polya, *How to solve it: A new aspect of mathematical method* Princeton university press, 2004.

[14] 白田由香利, 橋本隆子, and 飯高茂, *「感じて理解する数学入門―身近な事例を動かして学ぶ―」* オライリー・ジャパン, 2012.

[15] D. G. Luenberger, *Investment science* Oxford university press, 1998.

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

2024年2月28日(水) 13:30 ~ 15:40 | T2-A オンライン（Zoom Events)

問合せ最適化・多次元データ処理

座長:鬼塚 真(大阪大学)
コメンテータ:神崎 映光(島根大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T2-A-2-01] 動的空間データに対する高速学習型カーディナリティ推定
*Ji Yuchen[1]、天方 大地[1]、佐々木 勇和[1]、原 隆浩[1] (1. 大阪大学大学院情報科学研究科)

14:00 ~ 14:25
[T2-A-2-02] 和文に対する文字列述語の深層学習を用いた選択率推定
*大谷 蒼太[1]、板野 凌太[1]、土田 隼之[1] (1. 明石工業高等専門学校電気情報工学科)

14:25 ~ 14:50
[T2-A-2-03] グラフデータベースにおけるWorst-case Optimal Joinの最適化
*伊藤 寿浩[1]、塩川 浩昭[1] (1. 筑波大学)

14:50 ~ 15:15
[T2-A-2-04] 関係データベースにおける動的複数問合せ最適化の試作実装と初期評価
*木村 元紀[1]、合田 和生[1] (1. 東京大学)

15:15 ~ 15:40
[T2-A-2-05] 多次元ソートのための整数値変換手法の評価
*高尾 大樹[1]、黒川 能毅[1]、西川 記史[1]、茂木 和彦[1]、藤原 真二[1] (1. 株式会社 日立製作所)

DEIM Forum 2024 XX-Y

動的空間データに対する高速学習型カーディナリティ推定

Yuchen Ji† 天方 大地† 佐々木勇和† 原 隆浩†

† 大阪大学大学院情報科学研究科 〒 565–0871 大阪府吹田市山田丘 1-5

E-mail: †{ji.yuchen,amagata.daichi,sasaki,hara}@ist.osaka-u.ac.jp

あらまし 地理空間クエリのカーディナリティ推定は，クエリのスケジューリングと最適化において重要な役割を果たしている．基本的に空間データ集合は動的であり，挿入や削除が起こり得る．この環境では，更新が容易，低遅延，および高精度なカーディナリティ推定が必要である．しかし，既存のカーディナリティ推定手法は更新が非効率であったり，推定そのものが効率的でない．本研究では，動的な空間データに対して，SAFE（Sampling-Assisted Fast learned cardinality Estimator）を提案する．SAFE では，カーディナリティ推定モデルの高速学習を可能にするため，四分木に基づくデータ空間の分割を用い，小さな部分集合を抽出するサンプリングを導入する．さらに，２階層回帰モデルを用いて空間データ分布を近似しつつ，正確かつ高速なカーディナリティ推定を実現する．また，データ集合の更新の際に全てのモデルを再訓練することを避けるために，インクリメンタルなモデル更新を用いる．実データおよび人工データで実験を行い，SAFE が（i）空間データに対する最新のカーディナリティ推定モデルよりも性能が良いこと，および（ii）正確で低遅延な推定を保証しつつ，データの更新を効率的に処理することを検証する．

キーワード カーディナリティ推定，機械学習

1 はじめに

現在，GPS を搭載したデバイスの普及により，大量の空間データが生成されている [3], [4]．このような背景から，空間データを用いるアプリケーションでは，空間データの効率的かつ高品質な解析・検索を実現するデータ管理技術が求められている．空間データ処理の基本技術の一つに，指定した範囲内の空間点を検索する範囲検索（またはレンジクエリ）がある．レンジクエリの処理時間を短縮することは，快適なユーザ体験を提供するために不可欠であり，最も効率的なクエリプランを得るために，コストベースのクエリ最適化がしばしば使用される．与えられたクエリの出力サイズを推定するカーディナリティ推定は，クエリ最適化の最も重要な機能の１つである．空間データアプリケーションは通常，大量のレンジクエリを処理する必要があるため，レンジクエリに対するカーディナリティ推定の精度と効率は，レンジクエリのワークロードを処理する効率に直結する [1], [21].

空間データセットは一般的に動的であり，挿入や削除を含む頻繁な更新が起こり得る [22]．したがって，カーディナリティ推定手法で用いるデータ構造は，データセットの更新に応じて効率的に更新可能でなければならない．これは自明の要件であり，あるデータセットが更新される場合，静的なカーディナリティ推定手法は最新のデータ分布を考慮することができず，推定結果が不正確になる．本論文では，動的空間データにおけるカーディナリティ推定問題に取り組み，高速かつ高精度であり，効率的に更新可能なカーディナリティ推定器を設計する．

動機と課題． 近年提案されている機械学習を用いた学習型カーディナリティ推定手法は，従来の方法（例えば，ヒストグラム）と比べて著しい改善を示している [6], [12]．学習型カーディナリティ推定手法は，クエリ駆動モデルとデータ駆動モデルに分類できる．

クエリ駆動モデルは訓練クエリに基づいて構築され，訓練クエリをその真のカーディナリティにマッピングする．多くのモデルは深層学習を用いてクエリとカーディナリティの関係を捉える [5], [7]．しかし,「訓練クエリに依存する」という性質は動的データには適していない．これは，データセットの更新の度に学習クエリを収集し，その真のカーディナリティを計算する必要があるからである．

データ駆動型モデルは主にディープニューラルネットワークを用いて与えられたデータセットの分布を学習する [12], [18]．そのため，モデルを構築する時間が非常に長い（大規模なデータセットでは少なくとも数分の学習時間が必要）[6]．この特性は静的なデータのみに制限され（データセットが更新されるたびに再訓練が必要），動的なデータには対応できない．

LHist [1] は空間データのための最新のカーディナリティ推定手法であり，ニューラルネットワークのような複雑なモデルは用いないものの，モデルの更新うをサポートしていない．ヒストグラムやサンプリングなどの古典的な手法も，データセットから何らかの統計量を生成するため，データ駆動型モデルとみなすことができる．その構造は計算しやすいが，精度は学習ベースのモデルに比べて低い [7], [8].

ここで，動的な空間データを効率的に扱うために解決すべき課題をまとめる．(1) **安定した性能**：データセットが更新されても，推定精度と推定時間が安定していること．(2) **高速な推定**：カーディナリティ推定は頻繁に実行されるため，マイクロ秒レベルの高速な推定が望ましい．(3) **更新が容易**：推定モデルはデータの更新に応じて容易に（効率的に）更新できることが望ましい．先述のように，既存のアプローチはこれらの要件を同時に満たすことはできない．

貢献．本論文では以上の課題を解決し，動的空間データのカーディナリティ推定問題に対して以下の貢献を行う．

（1）**SAFE．**本論文では，サンプリングを活用した高速かつ高精度な学習型カーディナリティ推定手法 SAFE を提案する．SAFE は，効率的で高精度なカーディナリティ推定を維持しながら，頻繁なデータ更新に対応している．また，以下の 2 つの観点からモデルの更新を効率化する：(i) 必要なモデルのみを更新すること，および (ii) 高速に学習すること．(後に示すように，SAFE は複数の軽量モデルから構成されている．)

第一の観点として，既存のカーディナリティ推定手法は，更新を受ける度に構造またはモデル全体を一から再学習する必要がある．この問題を回避するため，各モデルは独立して更新可能という洞察に基づいて，2 階層回帰モデルを採用する．また，回帰モデルには様々なものがあり，その性能はデータ分布に依存する．これらのモデルの中から適切なものを選択するために，精度，推定時間，および構築時間を考慮したコストモデルを設計する．このコストモデルは，時間と精度のトレードオフのバランスが取れたモデルを選択するのに役立つ．

次に，高速学習を実現するため，四分木に基づくデータ分割を用いたサンプリングを提案する．サンプルはデータ分布に従うため，訓練に使用するデータ数を少なくしても精度を保ったままデータ分布の学習に要する時間を短縮できる．これら 2 つの観点を用いたアイデアにより，SAFE は安定した性能を提供する．

（2）**データ更新を高速に処理するモデルのインクリメンタル更新．**高速なモデル更新を可能にするために，SAFE のインクリメンタルなモデル更新アルゴリズムを提案する．与えられた更新（例えば，点の挿入）に対して，いくつかの回帰モデルのみが影響を受けるため，SAFE は，そのようなモデルを高速に識別し，影響を受けるモデルのみを更新する．

（3）**実験．**実世界と人工データセットを用いて実験を行った．その結果，SAFE は最先端のカーディナリティ推定手法の性能を超え，データセットの更新を効率的に処理できることが示された．

2 予 備 知 識

2.1 問 題 定 義

P を N 個の 2 次元（つまり, 地理空間）点のデータセットとし，P 内の各点 p を $p=(x,y)$ と表す．(本論文で使用する主なシンボルを表 1 にまとめる)．

クエリ．本稿では，直交範囲クエリ（orthogonal range query）について考える．与えられた直交範囲問合せの指定範囲を R とすると，直交範囲問合せは R に存在する P 内の全ての点を出力する．範囲 R は左下点 l と右上点 u で表現できる．以降では，範囲クエリを表すのに R を用いる．

カーディナリティとは，範囲クエリ R に対して P 内で R に存在する点の数である．R のカーディナリティを $card(R)$ で表す．

本論文では，$card(R)$ を推論する，カーディナリティの推定問題を扱い，$\widehat{card(R)}$ を R の推定カーディナリティとする．本論文の目的は $\widehat{card(R)}$ を効率的かつ高精度に計算することである．既存の範囲検索アルゴリズム（例えば R 木）を用いてカーディナリティを計算することは無意味（遅延が大きい）である．また，$O(\sqrt{N})$ 時間の range counting アルゴリズム（例えば，kd 木 [19]）が存在するが，多くの点アクセスを必要とするため，計算効率は悪い．したがって，我々は，$O(1)$ 時間の推定を達成する（推定時間はモデルのみに依存する）機会学習型のアプローチを取り入れる．一般性のため，クエリワークロードは未知であると仮定する．

表 1: 本稿で用いるシンボルの概要

記号	説明
P	空間データセット
p	P の要素（空間点）
N	P 内の点の数
R	範囲クエリ
$card(R)$	R の真のカーディナリティ
$\widehat{card(R)}$	R の推定カーディナリティ
f_x	x 次元の回帰モデル
f_y^i	i 番目の y 次元の回帰モデル

2.2 学習型カーディナリティ推定

学習型のカーディナリティ推定手法は，クエリをそのカーディナリティにマッピングする．例えば，Recursive Model Index (RMI) [9] のアイデアは，データ分布の累積分布関数（CDF）を近似するために回帰モデルを使用することである．図 1a に示すように，1 次元のデータの場合，近似された CDF f_{RMI} に基づいて 1 次元の範囲クエリ $[a,b]$ のカーディナリティを以下のように推定できる [1]．

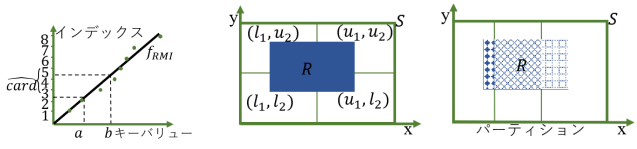
$$\widehat{card([a,b])} = f_{RMI}(b) - f_{RMI}(a). \quad (1)$$

一次元的なアプローチを多次元的なケースに拡張するためには，多次元の同時分布を考慮する必要がある．特に，次元間の相関は，カーディナリティ推定のために考慮する必要がある．既存のアプローチの一つに，多次元 CDF を直接近似するものがある．図 1b に示すように，MOSE [2] は以下のようにカーディナリティを推定する．

$$\widehat{card(R)} = f_{MOSE}(u_1,u_2) - f_{MOSE}(u_1,l_2) - f_{MOSE}(l_1,u_2) + f_{MOSE}(l_1,l_2) \quad (2)$$

ここで f_{MOSE} は 2 次元の同時 CDF を近似するモデルである [2]．MOSE はデータ分布を効果的に捉えるが，教師ありアプローチを採用している．つまり，f_{MOSE} を訓練するために，多くの（例えば数千の）訓練クエリとその真のカーディナリティ

1：直感的には，$f_{RMI}(a)$ は $\in [0,a]$ の 1 次元点の累積数を数える．
2：例えば，$f_{MOSE}(u_1,u_2)$ は，左下と右上の座標がそれぞれ $(0,0)$ と (u_1,u_2) である矩形に入る点の数を近似的に数える．



(a) RMI (b) MOSE (c) LHist

図 1: データ分布を近似する既存手法.（a）f_{RMI} は青い点の CDF 曲線を近似しており，一次元範囲のクエリ $[a, b]$ のカーディナリティを推定するのに用いることができる.（b）MOSE は，格子回帰モデルを用いて多次元点の分布を近似する.（c）LHist は，等深度ヒストグラムと同様に，各次元のデータ分布を再帰的に学習する.

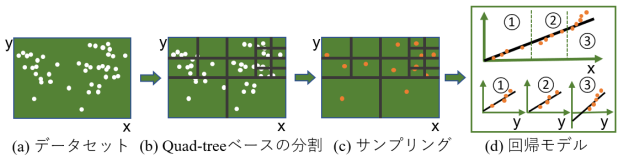


図 2: SAFE の構築.（a）データセット P が与えられたとき，（b）P を四分木の分割を用いる.（c） 各パーティションからランダムな点をサンプリングし,（d） サンプルを使って 2 階層回帰モデルを学習する．x 次元の回帰モデル（第 1 階層）は，x 次元の値に基づいて P を m 個のバケットに分割し，各バケットについて回帰モデルを作成する（第 2 階層）.

を必要とする．この方法は，データセットが更新されたときに各クエリの真のカーディナリティを再計算する必要があり，未知のクエリワークロードや動的データには適切でない.

もう一つの方法は，各次元におけるデータ分布をそれぞれ近似することである．LHist [1] は，データ空間を多次元グリッドセルに分割し，図 1c に示すような階層的回帰モデルを採用している．具体的には，LHist は 1 次元ヒストグラムに基づいて各次元の分布を近似する．カーディナリティは次のようにして推定される.

$$\widehat{card}(R) = \sum_i f^i_{LHist}(D_i), \qquad (3)$$

ここで，$f^i_{LHist}(D_i)$ は i 番目のパーティション D_i のカーディナリティを近似する.

この方法は，各回帰モデルを独立して訓練できるため，頻繁な更新に適応できる．更新が必要なモデルだけを選択することで，すべてのモデルを一から学習する必要がなくなり，動的なデータに対しても効率的に高精度なモデルを維持することができる．しかし，LHist にはこのような発想はなく，モデルの更新にも対応していない．SAFE はこの新しいアイデアを実装し，2 層回帰モデルを採用し，データの更新に効率的に対応する.

2.3 既存手法の更新的解決法

本節では，以下の 2 つの疑問について着目し，既存のモデルの更新アプローチを簡単に説明する.

再構築は回避可能か？更新に対処する最も簡単なアプローチは再構築（つまり，すべてのモデルを再トレーニングすること）だが，再構築は頻繁な更新には対応できない．更新にかかる時間を短縮するためには，インクリメンタルにデータの更新に対応することが望ましい．QuickSel [7] は，入力されるクエリから徐々に学習する継続的学習アプローチを採用している．しかし，QuickSel は新しいクエリが与えられないと更新されたデータ分布を学習しないため，データの更新が多くなるとモデルの精度が低下する.

更新にかかる計算コストは？学習型の手法のほとんどは，更新にかかる時間が長い．例えば LHist の場合，更新時間は構築時間と等しい．Quicksel は入力されたクエリと真のカーディナリティを記録するため，その更新時間はクエリが蓄積されるにつれて長くなる．データ集合の更新が起こる環境では，高速なモデル・データ構造の更新が求められる．これが既存の学習型カーディナリティ手法の主な課題である．SAFE はこの課題を克服し，高速な推定と更新を同時に実現する.

3 SAFE

メインアイデア. SAFE は動的空間データ用に設計された学習型カーディナリティ推定器である． 2 次元 CDF を正確に学習するために， 2 層の回帰モデルを採用しており，それぞれの回帰モデルはシンプルな（例えば多項式）モデルである．第 1 層（第 2 層）は x（y）次元に対応する [3]．シンプルな回帰モデルを採用する背景には， 2 つの考えがある．第一に，複雑な回帰モデル（例えば，ディープニューラルネットワーク）は高速な訓練が困難であるのに対して，シンプルな回帰モデルは高速な訓練が可能である．第二に，x 次元空間をバケットに分割することで，各バケット上の y 次元 CDF に焦点を当てることができ，CDF は複雑な曲線を持たない傾向がある．SAFE の第 1 層モデルはデータ空間を x 次元値に基づいて部分空間に分割する役割を担い，第 2 階層モデルはカーディナリティの推定を担う.

学習型カーディナリティ推定器の精度は，モデルが更新されない場合，動的データでは悪くなる．本論文では頻繁な更新（点の挿入および削除）を伴う動的な空間データを想定しているため，カーディナリティ推定器はデータ集合の更新に対して頑健であるべきである．これを実現するため， 2 つのアイデアを利用する：(i) 少ないサンプルで高速に学習する.(ii) 必要なモデルだけを更新する．これらのアイデアを実装するために，高精度な回帰モデルを訓練するための「良い」サンプルを使用することを提案し，どのモデルを更新しなければならないかを容易に識別できるように SAFE の構造を設計する．これらの技術により，データセットが更新されても，SAFE をインクリメンタルに更新でき，更新が必要なモデルを高速に再訓練できる.

概要. 図 2 に SAFE の構築の概要を示す．まず，四分木を用いたデータ分割を用いてサンプルを得る．このサンプルを用いて回帰モデルを高速に学習する．次に 2 層の回帰モデルを訓練する（第 1 層は x 次元の CDF に対応し，第 2 層は y 次元の CDF

3：x 次元と y 次元の順番は変更可能.

Algorithm 1: 構築

Input: P and m
Output: f_x, $f_y^1, ..., f_y^m$

```
P' = {P_1, P_2, ...} ← PARTITION-DATASET(P) // 四分木を用いたデータ空間分割
S ← GET-SAMPLES(P') // 各パーティションからランダムな点をサンプリング
P_x ← a set of x-dimensional values in S
f_x ← TRAIN(P_x) // 最小二乗法による x 次元の CDF 曲線の学習
B_1, B_2, ..., B_m ← ∅
foreach p ∈ S do
    B_j ← B_j ∪ {p}, where j = ⌊(m/N) f_x(x_i)⌋
foreach i ∈ [1, m] do
    P_y^i ← a set of y-dimensional values in B_i
    f_y^i ← TRAIN(P_y^i) // 最小二乗法による y 次元の CDF 曲線の学習
```

に対応する）．対応する部分空間の各カーディナリティ推定モデルは，他の部分空間に影響を与えないため，独立して学習できる．図 2（d）は，(i) データ空間を x 次元の値に基づいて複数（図では 3 つ）のバケットに分割し，(ii) y 次元の CDF を考慮しながら各バケットのモデルを学習する方法を示している．アルゴリズム 1 は，SAFE のモデルの構築アルゴリズムを示している．

3.1 四分木を用いたデータ分割によるサンプリング

モデルを更新する必要がある場合，動的なデータに対して高速な学習が必要となる．この目的のために，すべてのデータではなくサンプルを使用する．つまり，与えられたデータ集合から得られたサンプルから回帰モデルを訓練する．ここで，愚直なランダムサンプリングは高精度なモデルの構築に適切ではない．なぜなら，疎な領域の点をサンプリングし辛いためである．

そのため，データ空間全体を細かく分割し，それぞれの部分空間から点をサンプリングできるようにする．これは四分木を用いたデータ空間の分割によって実現できる．四分木は与えられたデータ空間を等しい大きさの 4 つの部分空間に分割する [10]．これを，全ての部分空間の点の数が一定以下になるまで繰り返し，各部分空間からランダムなサンプルを 1 つ選ぶ．(実験では四分木のセルが管理する点の数の最大値を 10 に設定した.）図 2（b）と（c）から分かるように，この方法は疎な領域からの点のサンプリングに失敗しない．

3.2 カーディナリティ推定のための回帰モデル

クエリに依存しないカーディナリティ推定のためのデータ駆動の既存モデルでは，階層回帰モデルを採用している．しかし，そのようなモデルをどのように効率的に更新するかというアイデアは考えられておらず，全てのモデルを再訓練することを避けられない [1], [2]．これに対して，SAFE の推定器は，各回帰モデルが独立して更新可能であるという考えに基づいて設計されている．さらに，精度と推定時間の良好なトレードオフを達成するために，与えられたデータセットに適したモデルを選択

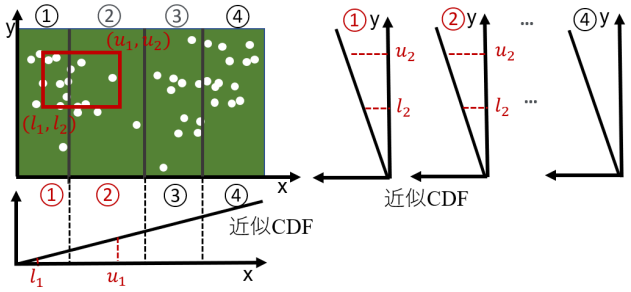


図 3: 回帰モデルの 2 次元レイアウトの図．x 次元の回帰モデルは，x 次元のデータ値（図では $m = 4$）に従って，データ空間を m 個のバケットに分割する．そして，y 次元の CDF を近似するために，各バケットに回帰モデルが構築される．クエリ R に対して，x 次元の回帰モデルは，アクセスする必要があるバケットを見つける．$\widehat{card(R)}$ は，アクセスされた全てのバケツからの結果を合計することで得られる．

するためのコストモデルを設計する．

レイアウトおよび推定． 図 2（d）に示すように，SAFE は 2 階層回帰モデルを採用している．x 次元については，x 次元の CDF を近似する回帰モデルに基づいて，データ空間をバケットと呼ばれる m 個の部分空間に分割する．次に，各バケットについて，このバケットに含まれる点の y 次元 CDF 曲線を近似する回帰モデルを訓練する．つまり，第 1 層は単一のモデル f_x を持ち，第 2 層は m 個の回帰モデル $f_y^1, ..., f_y^m$ を持つ．

クエリ R のカーディナリティを推定するために，第 1 階層モデルは R に重なるバケットを見つけるために使われる．推定フローを図 3 に示す．x 次元のモデル f_x はデータ空間を m 個のバケットに分割するため，R に重なる左端と右端のバケットはそれぞれ $\lfloor \frac{m}{N} f_x(l_1) \rfloor$ および $\lfloor \frac{m}{N} f_x(u_1) \rfloor$ で計算される．ここで，l_1 および u_1 はそれぞれ R の x 次元の左端および右端の値である．次に，重なるバケットの集合 B を用いて，カーディナリティを以下のように推定する．

$$\widehat{card(R)} = \alpha \sum_{i \in B} (f_y^i(u_2) - f_y^i(l_2)) \qquad (4)$$

ここで，l_2 と u_2 はそれぞれ R の y 次元の底辺と上辺の値であり，α は四分木による分割で得られた各部分集合に含まれる点の数の平均である．SAFE の回帰モデルは，サンプルで学習されるため，f_y で推定されるカーディナリティをスケーリングする必要があり，そのために α が使用される．

$\widehat{card(R)}$ を計算するために，SAFE は最大 $1 + m = O(1)$ 個の回帰モデルにアクセスし，各モデルは $O(1)$ 時間を必要とする（表 2 を参照）．したがって，式（4）は $O(1)$ 時間のみを必要とする．推定アルゴリズムの擬似コードをアルゴリズム 2 に示す．

回帰モデルの選択． 回帰モデルは 1 次元の CDF を近似する役割を担っており，LHist [1] は多項式モデルのみを使用している．適切な回帰モデルがデータ分布に依存することは自明であり，具体的には，スキューがあるデータ分布を考えると，単純な線形回帰モデルで CDF 曲線をうまく近似することは難しい．そ

Algorithm 2: ESTIMATION

Input: f_x, $f_y^1, ..., f_y^m$ and R
Output: $\widehat{card}$

$\widehat{card} \leftarrow 0$, $\mathcal{B} \leftarrow$ GET-BUCKETS(f_x, R) // R に重なる全てのバケットの集合を得る
foreach $B_i \in \mathcal{B}$ **do**
  $\widehat{card} \leftarrow \widehat{card} + (f_y^i(u_2) - f_y^i(l_2))$
$\widehat{card} \leftarrow \alpha \times \widehat{card}$ // α は各パーティションに含まれる点の数の平均

表 2: 回帰モデルの種類 (θ_i はパラメータであり，$z = x$ または $z = y$)

名称	説明
多項式	$f(z) = \theta_0 + \theta_1 z + ... + \theta_n z^n$
スプライン線形	$f(z) = \begin{cases} \theta_0 + \theta_1 z & (z \leqq \theta_3) \\ \theta_2 z + \theta_1 \theta_3 + \theta_0 & (z > \theta_3) \end{cases}$
対数線形	$f(z) = \exp(\theta_0 z + \theta_1)$

のため，次数の大きい指数モデルや多項式モデルの方が精度が高いことが考えられる．つまり，どのモデルを選択すべきかを検討するべきである．しかし，その自動的な選択は難しい．

この課題に対して，我々は以下の 3 つの要素から最適な回帰モデルを選択する．(表 2 は SAFE の候補となる代表的な回帰モデルの種類を示している)．

- **推定誤差．**構築中に推定誤差を得られないため，与えられた回帰モデルの最小 2 乗訓練誤差を測定する．この値は，モデルのデータ分布の近似精度とみなせる．
- **推定時間．**推定に要する計算時間は，回帰モデルの構造に大きく依存する．各回帰モデルには構築時にデフォルト値を投入することで，予想される推定時間を得ることができる．
- **訓練時間．**動的データに対しては，頻繁かつ効率的なモデルの更新が求められるため，訓練時間も重要である．与えられたモデルの訓練時間の測定方法は容易である．

ある回帰モデル f のコストを次のように測定する．

$$Cost(P, f) = E_{train} + T_{est} + T_{train} \quad (5)$$

ここで，E_{train}, T_{est}, および T_{train} はそれぞれ推定誤差，推定時間，および学習時間を表す．全ての候補の E_{train}，T_{est}，および T_{train} を計算し，それぞれを [0,1] に正規化する（つまり $Cost(P, f) \in [0, 3]$）[4]．この結果において最もコストの小さいモデルを使用する．このコストモデルの評価は，各次元について行う．

回帰モデルのトレーニング．与えられた回帰モデルに対して，最小 2 乗法を用いてそのパラメータを求める．ここでは，多項式モデルの学習を紹介する．(スプライン線形モデルや対数線形モデルの学習は，基本的に線形回帰モデルの学習と同一である．) (p_i, r_i) のペアを含む訓練データセットがあるとする．ここで r_i は，与えられた次元（x または y）のソート順によって決まるデータセット中の p_i の順位である．多項式モデル $f(z) = \theta_0 + \theta_1 z + ... + \theta_n z^n$ において，z は与えられた点の x 次元または y 次元の値であり，残差平方和 RSS は

$$RSS = \sum_{i=1}^{N'} (f(z_i) - r_i)^2 \quad (6)$$

であり，$N' \leqq N$ は入力点の数である．RSS は行列形式で書くことができる [11].

$$RSS = (Z_v\theta - Q_\gamma)^T (Z_v\theta - Q_\gamma), \quad (7)$$

ここで

$$Z_v = \begin{bmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^n \\ 1 & z_2 & z_2^2 & \cdots & z_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{N'} & z_{N'}^2 & \cdots & z_{N'}^n \end{bmatrix}, \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, Q_\gamma = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{N'} \end{bmatrix} \quad (8)$$

であり，Z_v は Vandermonde 行列である．最適な $\theta = \{\theta_0, \theta_1, ..., \theta_n\}$ は RSS を最小にするため，θ は次式で計算できる．

$$\theta = (Z_v^T Z_v)^{-1} Z_v^T Q_\gamma \quad (9)$$

各モデルの学習（f_x を学習した後に f_y^i を学習する）の流れは，アルゴリズム 1 で説明している．

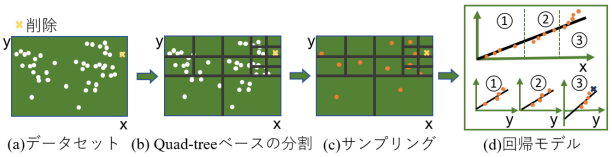


図 4: モデル更新のフロー．赤いトークンは点の削除を表す．

3.3 モデル更新

点の挿入，削除，および置換を更新と考える．図 4 は，点の削除があった場合のモデル更新の流れを示している．更新があった場合，まず四分木によるデータ空間分割を更新する必要があるかどうかをチェックする．これを図 4 (b) に示す．必要であれば，サンプルも更新する必要がある．(そうでない場合は，式 (4) で α を更新し，更新は終了する．) 具体的には，以下の 3 つの場合に対応するパーティションのサンプルを更新する必要がある．

- パーティションが分割される（主に挿入が原因）．
- パーティションがマージされる（主に削除が原因）．
- 現在のサンプルが削除された（削除または置換のため）．

このような場合，更新レコードを作成する．具体的には，挿入の場合は $\langle s, ins \rangle$ を，削除の場合は $\langle s, del \rangle$ を作成する．ここで，s は更新が必要なパーティションの元のサンプルである．また，新しいパーティションからはランダムなサンプルを選び，対応するバケットに割り当てられる．(置換は，削除の次に挿入が起きたとみなすことができる．)

4：基準に重み w_1, w_2, w_3 を使用することも可能．

各サンプル s は一意なバケットに関連付けられているので，s が属するバケットは，ハッシュテーブルを使うことによって，すなわち s をキーとして使うことによって特定できる．このアプローチと更新レコードから，更新が必要なモデルを特定することができる．これらのモデルは，3.2 節での方法を用いて更新される．

ここで，更新があったとしても，四分木によるデータ空間分割は実践的にほとんど変化しないので，モデルを更新することはほとんどない．たとえいくつかのパーティションが更新されたとしても，そのような更新されたパーティションは局所的に存在するため，少数のモデルしか更新する必要がない．このメカニズムは SAFE 構造の更新時間を最小化するのに役立ち，これは 4.3 節で実験的に確認されている．

4 実　　験

本章では，実験結果を報告する．全ての実験は，Intel Xeon Gold 6254@3.10GHz CPU および 768GB RAM を搭載した Ubuntu 18.04 LTS サーバで行った．また，全ての評価アルゴリズムはシングルスレッドであり，C++で実装した．

4.1 設　　定

データ． 3 つの実世界データセット（Colorado，Paris，および Shikoku）と 1 つの人工データセット（Gaussian）を使用した．Colorado，Paris，Shikoku，および Gaussian は，それぞれ 3,500 万点，2,750 万点，900 万点，および 2,000 万点から構成されている．これらの実世界データセットには，OpenStreetMap（OSM）から抽出された緯度と経度のペアが含まれている [5]．Gaussian は，各次元について平均 100 および分散 500 のガウス分布を用いて生成した．

カーディナリティ推定方法． 本実験では，SAFE を以下の既存のカーディナリティ推定手法と比較した．

- LHist [1]：空間データに対する最先端のカーディナリティ推定方法．文献 [1] で推奨されているように，多項式モデルの次数を 6 に設定した．
- QuickSel [7]：最新の教師あり（クエリ駆動型）カーナリティ推定手法．ガウス混合モデルを更新するために，入力されるクエリから継続的に学習する．上記の手法と比較できるように，QuickSel を初期化するための訓練セットとして 1,000 クエリを用意した．

評価基準． 各手法の精度を測定するために，q-error を使用した．

$$\text{q-error} = \frac{\max\left(\widehat{card}(R), card(R)\right)}{\min\left(\widehat{card}(R), card(R)\right)} \qquad (10)$$

これはカーディナリティ推定問題で広く使用されている [13]．推定時間については，$\widehat{card}(R)$ を計算する平均時間を報告する．

5：http://download.geofabrik.de/.

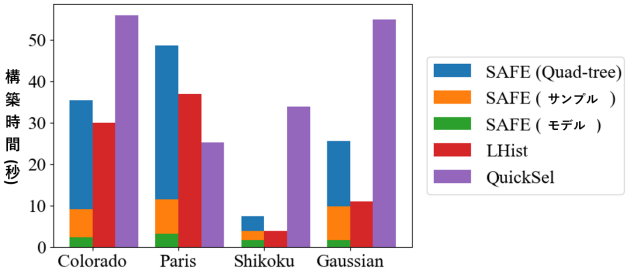


図 5: 異なるデータセットでの構築時間評価

4.2 結果：静的データの場合

各データセットに対して，3 つのクエリグループを用意した．各グループはランダムに生成された 1,000 のレンジクエリを含み，$card(R)/N$ は 0.1 および 0.01 である．同じグループ（異なるグループ）のクエリは，類似した（全く異なる）クエリ領域を持つ．予備実験では，R 木は平均 10 ミリ秒未満で $card(R)/N = 0.01$ となるレンジクエリの出力サイズを計算した．そのため，推定の有効性を保証するために，以降の実験では全ての評価手法の推定時間をデフォルトで 100 マイクロ秒程度に制御した．そのため，SAFE の m はそれに合わせて調整した．なお，本節では更新が発生しない場合のカーディナリティ推定モデルを評価する．

構築時間． 図 5 に各手法における推定器の構築時間を示す．SAFE の構築は 3 つの部分に分けられる：四分木に基づくデータ空間分割，サンプリング，および回帰モデルの学習である．（レンジクエリを効率的に実行するため）四分木があらかじめ構築されていると仮定すると，SAFE の構築の主なボトルネックである四分木の構築時間は無視できる．この仮定の下では，その構築時間は最速となる．

精度． 図 6 は，各手法の 50-th q-error と 95-th q-error を示している．SAFE と LHist は同じような q-error を示す．QuickSel は通常，3 つの手法の中で最も低い 95-th q-error を示す．一方，3 つの手法はすべて同様の 50-th q-error を示し，それらは 1 に近い（ほとんど誤差がない）．この結果から，SAFE はクエリワークロードなしに高精度なカーディナリティを推定ことができることが分かる．

精度と推定時間． 次に，推定精度と推定時間のトレードオフ関係を調べた結果を図 7 に示す．SAFE は最もトレードオフのバランスが取れていることが分かる．アプリケーションがより速い推定時間を必要とする場合，QuickSel の精度はすぐに低下する．一方，SAFE は，より早い推定時間（例えば，10 マイクロ秒未満）が要求されても，その精度はそれほど低下しない．これは SAFE の大きな利点である．

4.3 結果：動的データの場合

動的データの場合における実験結果を報告する．ここでは精度と更新時間を評価する．（モデルを更新しても推定時間は変わらない．）比較手法として，QuickSel およびモデルを更新しな

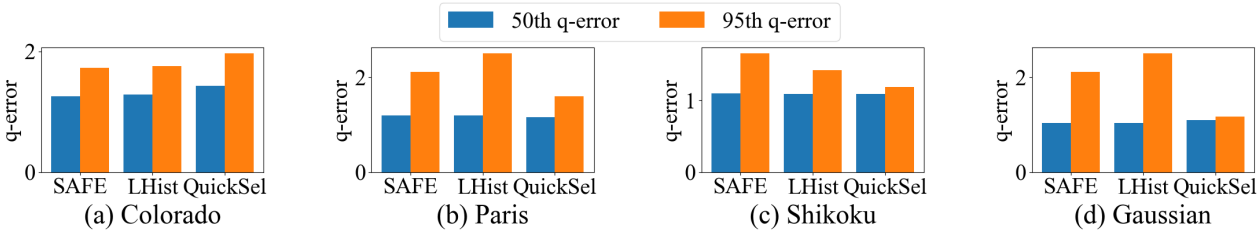


図 6: 異なるデータセットでの精度評価

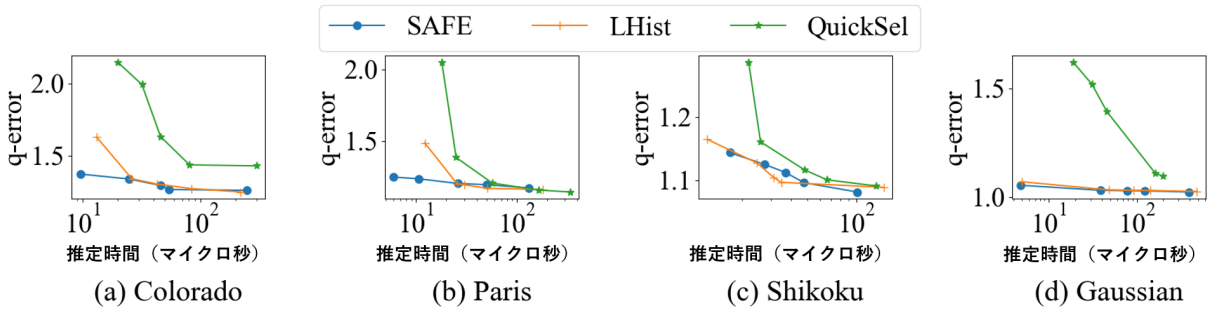


図 7: 推定時間 vs. 推定精度の比較評価（50th q-error）

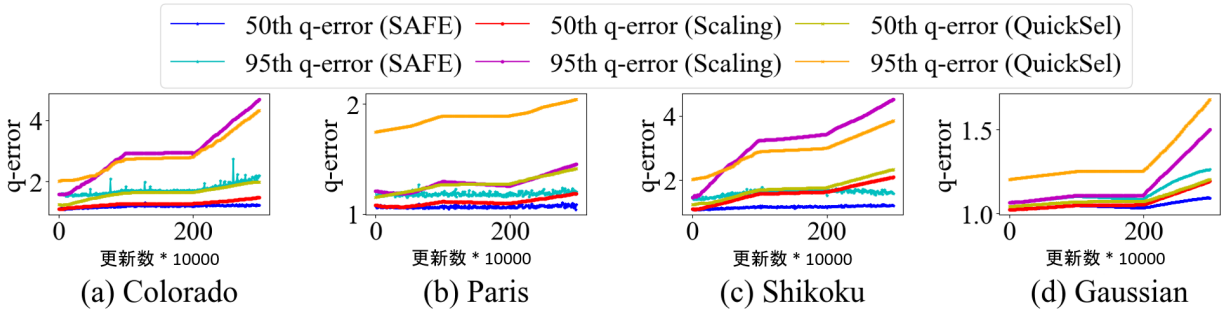


図 8: 更新数に対する q-error の変化

い SAFE を用いた [6]後者を Scaling と呼び，元のデータセットと更新されたデータセットのサイズを使用すしてモデルの推定結果をスケーリングする．

100 万回の挿入，100 万回の削除，100 万回の置換を行い，データセットを順次更新した．挿入は対応するデータ空間内のガウス分布に従い，削除は元のデータセットから無作為に選んだ．各置換は，ランダムに選ばれた点をガウス分布に従った人工点と置き換えた（すなわち，データ分布を徐々に変化させた）．評価には $card(R)/N = 0.1$ と $card(R)/N = 0.01$ となる 300 個のクエリを用いた．図 8 は更新数に対する q-error の変化の結果を示している．SAFE は 50th q-error がほぼ 1.0 を維持し，95-th q-error も低く，高い精度を維持している．一方，Scaling と QuickSel は正確な推定を維持することができず，更新回数が増えるにつれて 50th q-error と 95th q-error は常に大きくなる．

6：QuickSel のモデルを最新のデータセット用に更新するには，既存のクエリの結果をすべて計算する必要があり，数百のクエリでも数秒以上の計算時間がかかる．これは頻繁なデータ更新には対応できないため，QuickSel のの方針通り，モデルは更新しなかった．さらに，LHist はデータの更新をサポートしておらず，データが更新される度にモデルの再構築が必要である．

表 3: 1 回の更新にかかる SAFE の平均更新時間（マイクロ秒）

データセット	Colorado	Paris	Shikoku	Gaussian
更新時間	9.9	19.8	6.2	14.1

次に，1 回の更新にかかる SAFE の更新時間を調べたところ，表 3 のようになった．更新時間は 10 マイクロ秒前後で，非常に高速であることが分かる．これは頻繁なデータ更新にとって望ましい結果である．4.2 節と 4.3 節の実験結果を考慮すると，SAFE は最先端のカーディナリティ推定手法よりも推定精度のトレードオフと更新性に優れていることは明らかである．

5 関 連 研 究

データ駆動型カーディナリティ推定． データ駆動型カーディナリティ推定器は，与えられたデータセットからデータ分布を直接学習する．(SAFE はこのカテゴリに属する)．代表的な例はヒストグラムである．最新の学習型の手法は，このタスクを教師なし問題として考える傾向がある．Sum-Product ネットワーク [17] は，点をクラスタに再帰的に割り当てる．このネッ

トワークは，確率的グラフィカルモデルを使用して，結合データ分布を近似する．次に，SUM 演算子を使用して，すべてのクラスタからの推定結果を追加する．自己回帰モデル [6], [18] は，ニューラルネットワークを使って条件付き同時分布を近似し，カーディナリティを推定する．前述のように，ニューラルネットワークを用いたモデルは高速に学習できないため，動的データには適していない．

クエリ駆動型カーディナリティ推定．多くの最新のカーディナリティ手法は，このクエリ駆動型を採用する傾向にある [8]．これは，深層学習モデルがこのクエリ駆動型に適しているからである．しかし，このモデルも動的データには適していない．

QuickSel [7] は，深層学習を使わない技術の中で最も代表的なものである．しかし，クエリを用いてデータ分布を学習する必要があり，その教師あり学習の性質は，動的データには不向きである．また，クエリ駆動型を用いた手法は，クエリワークロードが未知の場合には利用できず，クエリワークロードのシフトに対応できない．QuickSel はこのようなクエリ分布の変化にも適応できるが，4 章で述べたように更新コストが高い．

学習型空間インデックス．学習型インデックスは，空間インデックスを機械学習により構築する [15]．文献 [20] では，z-ordering を用いて RMI を 2 次元空間に拡張している．LISA [16] は，最新の学習型空間インデックスであり，z-ordering の代わりに，空間点を 1 次元値にマップする技術を提案している．Waffle [14] は，動的なデータセットを扱うために導入されており，ディスク I/O を最適化するための新しいバルクロード方法を提案している（これは我々の想定とは異なる）．Waffle は，クエリと更新速度のトレードオフを管理することで，インデックスを自動的に最適化することができる．Waffle は事前にクエリワークロードを必要とするため，そのアイデアは我々の問題設定では利用できない．

6 結　　論

本論文では，動的空間データに対するカーディナリティ推定のための SAFE を提案した．SAFE は，(i) 高速な学習のためのサンプリングベース学習，(ii) 効率的な更新能力を持ちながら高速かつ高精度な推定を可能にする 2 階層回帰モデル，および (iii) インクリメンタルなモデル更新戦略を採用した．これらメカニズムを組み合わせることで，データ集合の更新に頑健でありながら，低遅延で高精度な推定を実現した．実験の結果から，SAFE が効率的なモデル更新をサポートしながら，最新のカーディナリティ推定器の性能を上回ることを示した．

謝辞　本研究の一部は AIP 加速課題 (JPMJCR23U2) および JST CREST(JPMJCR21F2) の支援を受けたものである．

文　　献

[1] Liu, Q., Shen, Y. & Chen, L. LHist: Towards Learning Multi-dimensional Histogram for Massive Spatial Data. *ICDE*. pp. 1188-1199 (2021)

[2] Sun, L., Li, C., Ji, T. & Chen, H. MOSE: A Monotonic Selectivity Estimator Using Learned CDF. *IEEE Transactions On Knowledge And Data Engineering*. **35**, 2823-2836 (2023)

[3] Tang, M., Yu, Y., Malluhi, Q., Ouzzani, M. & Aref, W. LocationSpark: A Distributed in-Memory Data Management System for Big Spatial Data. *PVLDB*. **9**, 1565-1568 (2016)

[4] Eldawy, A. & Mokbel, M. The Era of Big Spatial Data. *ICDE Workshops*. pp. 42-49 (2015)

[5] Wang, J., Chai, C., Liu, J. & Li, G. FACE: A Normalizing Flow Based Cardinality Estimator. *PVLDB*. **15**, 72-84 (2021)

[6] Hasan, S., Thirumuruganathan, S., Augustine, J., Koudas, N. & Das, G. Deep Learning Models for Selectivity Estimation of Multi-Attribute Queries. *SIGMOD*. pp. 1035-1050 (2020)

[7] Park, Y., Zhong, S. & Mozafari, B. QuickSel: Quick Selectivity Learning with Mixture Models. *SIGMOD*. pp. 1017-1033 (2020)

[8] Han, Y., Wu, Z., Wu, P., Zhu, R., Yang, J., Tan, L., Zeng, K., Cong, G., Qin, Y., Pfadler, A., Qian, Z., Zhou, J., Li, J. & Cui, B. Cardinality Estimation in DBMS: A Comprehensive Benchmark Evaluation. *PVLDB*. pp. 752-765 (2021)

[9] Kraska, T., Beutel, A., Chi, E., Dean, J. & Polyzotis, N. The Case for Learned Index Structures. *SIGMOD*. pp. 489-504 (2018)

[10] Finkel, R. & Bentley, J. Quad trees a data structure for retrieval on composite keys. *Acta Informatica*. **4** pp. 1-9 (1974)

[11] Lesh, F. Multi-dimensional least-squares polynomial curve fitting. *Communications Of The ACM*. **2**, 29-30 (1959)

[12] Meng, Z., Wu, P., Cong, G., Zhu, R. & Ma, S. Unsupervised Selectivity Estimation by Integrating Gaussian Mixture Models and an Autoregressive Model.. *EDBT*. pp. 2-247 (2022)

[13] Wu, P. & Cong, G. A Unified Deep Model of Learning from both Data and Queries for Cardinality Estimation. *SIGMOD*. pp. 2009-2022 (2021)

[14] Moti, M., Simatis, P. & Papadias, D. Waffle: A Workload-Aware and Query-Sensitive Framework for Disk-Based Spatial Indexing. *PVLDB*. **16**, 670-683 (2022)

[15] Wu, J., Zhang, Y., Chen, S., Wang, J., Chen, Y. & Xing, C. Updatable Learned Index with Precise Positions. *PVLDB*. **14**, 1276-1288 (2021)

[16] Li, P., Lu, H., Zheng, Q., Yang, L. & Pan, G. LISA: A Learned Index Structure for Spatial Data. *SIGMOD*. pp. 2119-2133 (2020)

[17] Hilprecht, B., Schmidt, A., Kulessa, M., Molina, A., Kersting, K. & Binnig, C. DeepDB: Learn from Data, not from Queries!. *PVLDB*. **13**, 992-1005 (2020)

[18] Yang, Z., Kamsetty, A., Luan, S., Liang, E., Duan, Y., Chen, X. & Stoica, I. NeuroCard: One Cardinality Estimator for All Tables. *PVLDB*. **14**, 61-73 (2020)

[19] Bentley, J. Multidimensional Binary Search Trees Used for Associative Searching. *Communications Of The ACM*. **18**, 509-517 (1975)

[20] Qi, J., Liu, G., Jensen, C. & Kulik, L. Effectively learning spatial indices. *PVLDB*. **13**, 2341-2354 (2020)

[21] Shi, Y., Tong, Y., Zeng, Y., Zhou, Z., Ding, B. & Chen, L. Efficient approximate range aggregation over large-scale spatial data federation. *IEEE Transactions On Knowledge And Data Engineering*. **35**, 418-430 (2023)

[22] Amagata, D. & Hara, T. Monitoring MaxRS in Spatial Data Streams.. *EDBT*. pp. 317-328 (2016)

和文に対する文字列述語の深層学習を用いた選択率推定

大谷蒼太† 板野凌太† 土田隼之‡

†明石工業高等専門学校 〒675-0084 兵庫県明石市魚住町西岡 679-3
E-mail: †{me2203, me2201}@s.akashi.ac.jp , ‡t.tsuchida@akashi.ac.jp

あらまし 英文の文字列述語の選択率推定では、深層学習と古典的手法を組み合わせた手法が近年提案され、高い性能を示している。この手法の和文での推定性能を英文、フリガナ、漢字のデータセットを用いて検証したところ、高頻度で出現する文字列に対する推定性能が英文と比較して低かった。そこでデータセット内の出現頻度が高い文字列の学習回数を増やすことにより推定性能の改善を試みたところ、フリガナデータの推定において出現頻度上位 100 件の誤差率が改善され、英文以上の推定性能が確認できた。漢字データでも出現頻度上位に対する性能の改善は見られたものの、全体性能が悪化した。さらに、住所分析のユースケースを検討し、提案手法で適切な実行プランが選択できることを確認した。
キーワード 問合せ処理、選択率推定、文字列述語

1. 背景

文字列述語の精度の良い選択率推定手法の確立は長年の課題であり、例えば氏名カラムの値が文字列「明石太郎」であるレコード数を文字列データそのもの全体を探索せずに、補助データ構造の組み合わせで精度よく推定するのが難しいことが知られている。特に文字列のパターンマッチング(接頭辞・接尾辞・部分文字列)も含めるとより困難となる。選択率推定が精度良く行えないと、クエリオプティマイザが最適な実行プランを選ばずに、SQL 実行時間が最適な実行プランより大幅に長くなることがある。精度の良い推定を実現するために深層学習を用いて文字列を特徴ベクトルへ変換してモデルを学習させる手法である Astrid-Embed[1]が提案されており、この先行研究では英文文字列を対象に文字列述語の選択率推定が精度良く行えることが示されている。そこで、英文(アルファベット)とは文字の種類数や文字の出現頻度が大きく異なる漢字やフリガナなどの和文の文字列を対象とした場合の文字列述語の選択率の推定精度の検証と、推定精度向上のための学習データの前処理指針の提案を行う。和文で文字列述語の選択率推定を行うユースケースを検討し、実データを用いた場合に適切な実行プランが選択できるかも確認する。

現状は、文字列述語の処理内容に関係なく 40%などの一定の絞り込み率が設定されるなどされており、安定して高速な実行プランで SQL 処理が行えるようになっていない[2]。和文の文字列への文字列述語を含む SQL を安定して高速に処理が行えるようになると、例えば、web ショッピングモールの商品名に含まれる語句毎の売り上げ分析のような、より高度な分析が効率的に行えることが見込まれる。

2. Astrid-Embed

深層学習を用いない初期の推定手法では、接尾辞木などのサマリーデータ構造が使用されていた。接尾辞木はすべての接尾辞と出現頻度を表すことができるが、すべての情報を含んだ接尾辞木は元となるデータセットより遥かに大きくなる可能性がある。そのため頻度の低い文字列の情報を削除し、頻度の高い文字列の組み合わせから選択率を推定するアプローチが取られるようになった[3-5]が、これらの手法では不正確な推定を行うことが多くあった。そこで深層学習を利用したアプローチが取られるようになった。Astrid-Embed 以外にも深層学習を利用した手法は存在する[6-9]が、文字列以外のデータ型を対象とすることが多かった。また、自然言語処理の技術を応用した研究[6,9]は選択率の考慮が足りていなかった。これらを踏まえ、Astrid-Embed は従来の手法と深層学習を用いた手法を組み合わせたアプローチを取っている。具体的には、部分文字列の選択率を推定する深層学習モデルを作成することで、巨大なサマリーデータ構造の代わりとして利用する。

字面の似ている文字列述語であるにも関わらず、大きく異なる選択率となる可能性がある。そのため文字列を選択率推定器の入力として直接使用すると、推定精度が低下することが知られている。また、文字列間の類似度は文字列述語の種類(接頭辞・接尾辞など)にも影響される。よって、接尾辞木のようなサマリーデータ構造で文字列データを格納した後に、サマリーデータ構造での各文字列間の近傍度合いを利用した学習が行われる。まず、文字列の前後関係を扱える接尾辞木を構築し、木のデータから文字列類似度を数値化するためにトリプレットロスを損失関数とした埋め込み処理を用いて、文字列を特徴ベクトルとして表現する。特徴ベクトル化した文字列を入力としてニューラルネットワークを用いて学習する。Astrid-Embed の概要を図 1 に示す。

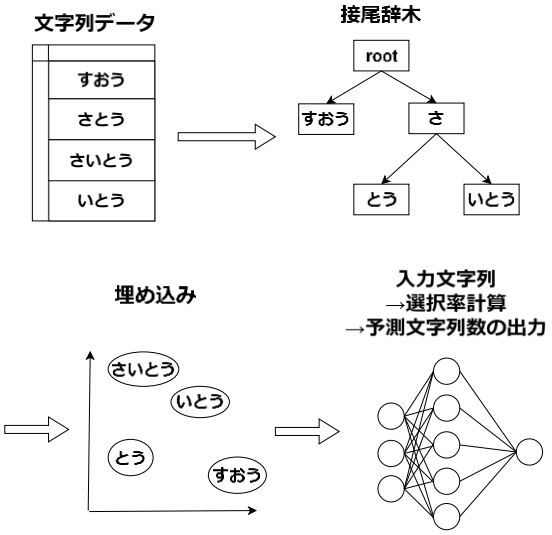


図 1　処理の流れの例

2.1 Astrid-Embed の構成要素

Astrid-Embed は主に以下の 3 つの要素から構成される。

- サマリーデータ構造
- 埋め込み学習器
- 選択率推定器

サマリーデータ構造は、学習対象のデータセットに関して文字列と選択率を効率的に結びつけることで、トリプレットロスを用いた学習で必要となる任意文字列に対する類似文字列や非類似文字列を効率的に取得する役割を持っており、目的に応じたデータ構造で置き換えることができる。例えば、接頭辞クエリの推定にはトライ木[10]を使用し、部分文字列/接尾辞クエリには接尾辞木を使用することもできる。埋め込み学習器は、サマリーデータ構造を使用して、選択率推定に最適化された埋め込みを学習する。最後に、各文字列は選択率推定用に学習された埋め込みを使って特徴ベクトル化され、選択率推定器が特徴ベクトルを用いて学習される。

2.2 選択率推定のための埋め込み

選択率推定のための埋め込みは、主に 2 つの要件を満たす必要がある。まず、構文上類似し、類似した選択率を持つ 2 つの文字列には、類似した埋め込みが割り当てられる必要がある。これはクエリによって異なる場合があり、接尾辞クエリでは、'さいとう'と 'いとう' は類似している考えられるが、接頭辞クエリではそうではない。次に、埋め込みが文字と q-gram の単位で機能する必要がある。

2.3 トリプレットロスによる埋め込みの学習

学習データを重複の無い N個の文字列$\{s_1, s_2, \ldots, s_N\}$とし、$s_i$に対して$C_i$と$D_i$をそれぞれ類似および非類似な文字列の集合とする。$s_i$の特徴ベクトルが$C_i$の部分文字列に近く、$D_i$の部分文字列からは遠ざかるような特徴ベクトルを学習する。$c_i \in C_i \wedge d_i \in D_i$である$w_i, c_i, d_i$($w_i$は入力文字列)が与えられた場合、トリプレットロス[11, 12]はつぎのように定義できる。

$$\max\left(\left||f(w_i) - f(c_i)\right|\right|^2 - \left||f(w_i) - f(d_i)\right|\right|^2 + m, 0\Big)$$

ここで$f(s)$はsの特徴ベクトルを表し、mはマージンと呼ばれる正の値である。この定式化により、w_iとc_iとの距離がw_iとd_iの距離よりも小さくなり、w_iとd_iの距離がmよりも大きくなることが保証される。

2.4 選択率の変換

選択率推定モデルは、特徴ベクトルを選択率へ変換する。選択率のほとんどは小さな値であり、偏りがある。このようなデータで学習すると性能が低くなるため、対数変換と最大最小スケーリングを行う(以降正規化と呼ぶ)ことで偏りを減らす[6]。選択率が 0.00001 の場合、対数変換で 5 が生成される。その後、選択率が最大のものを 1、最小のものが 0 となるようにスケーリングされる。

3. 評価実験

3.1.実験環境

以下の環境で実験を行った。

- SQL 実行環境 : MySQL 8.0.32
- OS : Windows11 Pro 22H2
- CPU : Intel Core i3-10105F
- GPU: Palit GeForce RTX™ 3060 Dual
- メモリ : W4U3200CM-8GR×2(16GB)
- SSD : Crusial 500MX(500GB)

3.2.和文文字列での基礎評価

3.2.1.方法

カタカナで記載されたフリガナのデータ(以降フリガナデータ)、漢字を含む文字列のデータ(以降漢字データ)の 2 つを和文データの例として利用し、部分文字列の出現頻度に関して学習を行い、選択率推定を行った場合の英文データとの性能比較を行う。学習データセットとして 15 万件程度が保存されている日本全国住所データ[13]の町域カナデータをフリガナデータとして利用する。また、同データの町域データを漢字データとして利用する。町域データは、県や市より小さな分類であるため重複が少なく、データベースに保存されている文字列データとしても違和感がないものである。

検証には、フリガナデータ、漢字データに加え、参考文献[1]で検証に利用していた英文のデータである dblp_titles(以下英文データ)と、フリガナデータを pykakasi[14]を用いて英小文字で表記されるローマ字に変換したデータ(以降ローマ字データ)を英文データの例として用いる(図 2)。データサイズは入力前の csv 形式で英文とフリガナは 3MB 程度、漢字とローマ字は 2MB 程度である。学習対象は接尾辞とし、文字列長は最大 8 文字とした。

比較は、誤差率の平均値、予想選択率の最大値と、件数真値に従い降順に並べた際の上位の平均誤差率を利用する。誤差率eは件数予測値をc,件数真値をdとして式(1)で表される。

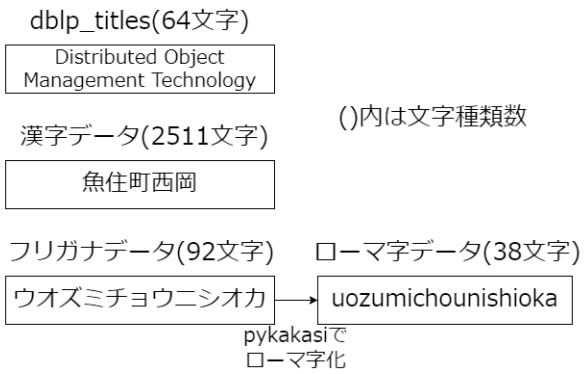


図 2 性能評価で利用するデータ

$$e = max\left(\frac{c}{d}, \frac{d}{c}\right) \quad (1)$$

誤差率の平均値は全体的な性能を確認する指標になると考えられるが、この指標のみで性能を評価すると誤った評価をしてしまう可能性がある。例えば、すべての文字列に対して低い推定選択率を出力するシステムができた場合、大部分の文字列は選択率が低いため、このシステムの誤差率は平均的に優れたものになると考えられる。しかし実際には高い選択率のものを推定できておらず、推定システムとして高い性能があるとは言えない。このような事態を回避するために、個別の文字列述語に対する選択率の最大誤差も確認する必要がある。これらの確認として、予想選択率の最大値、出現回数上位に対応する平均誤差率を利用する。予想選択率の最大値はすべての文字列に対して低い予想選択率を出力していないかの確認に用いる。出現回数上位に対応する平均誤差率というのは、件数の真値が大きいものに絞った平均性能の確認として用いる。これらを指標として利用することで、全体性能と高い選択率の予測性能を比較し、個別の文字列述語に対する選択率の最大誤差性能を比較できる。なお、選択率は「該当文字数/全レコード数」としてカウントしている。そのため、1 つのレコードに文字列述語の該当文字が複数含まれる場合は選択率が 1 を超えることがある。

埋め込みモデルは、1 次元の畳み込み層と平均プーリング層を 3 層重ねたモデルを利用する。また、選択率推定モデルは、ReLU 活性化関数を使用し、バッチ正規化とドロップアウト率が設定された 5 つの線形レイヤーで構成する。これらは英文の性能検証[1]において用いられたモデルと同様である。畳み込みの学習は 32epoch、選択率の学習は 64epoch とした。

性能検証は乱数のシード値を変更して 3 回実行した結果を利用する。予想選択率の最大値は 3 回の実行それぞれの値を確認し、真値を比較する。平均誤差率は 3 回の平均を取る。また、学習の epoch 数が十分であることを確認するため、各データの学習における損失の推移を代表して 1 回記録する。

3.2.2.性能評価

英文データとフリガナデータ、ローマ字データ、漢字データに対して接尾辞の選択率を推定した。図 3 に畳み込み学習時の損失の推移を示す。また、図 4 に選択率推定モデルの学習における損失の推移を示す。

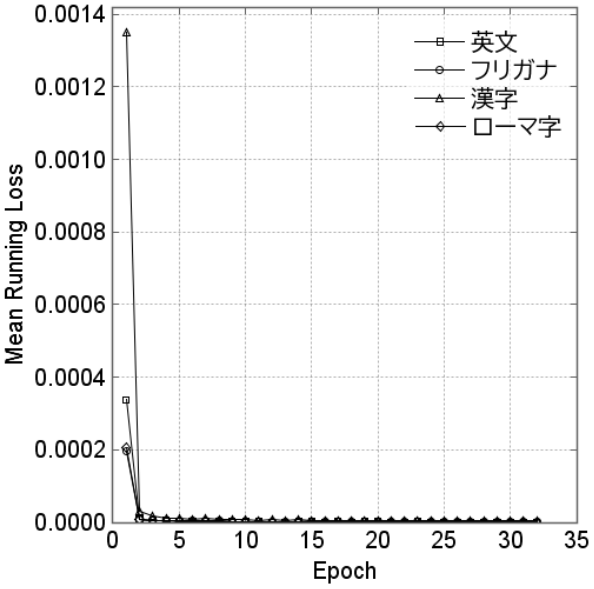


図 3　各データの畳み込み学習時の損失推移

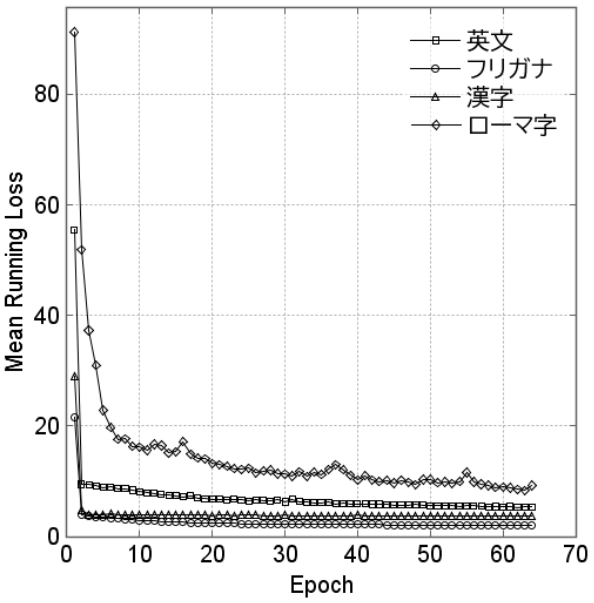


図 4　各データの選択率推定学習時の損失推移

表 1 に各データの予想選択率の最大値を、表 2 に件数真値に従い降順に並べた際の上位の平均誤差率を示す。表 2 の 100%の欄がデータ全体の平均誤差率に該当する。

表 1　各データの予想選択率の最大値

	英文	フリガナ	ローマ字	漢字
試行 1	97.3% (406.4%)	8.3% (13.3%)	36.5% (57.2%)	7×10^{-3}%) (7×10^{-4}%)
試行 2	153.5% (347.8%)	2.5% (50.9%)	17.9% (13.8%)	9×10^{-3}% (7×10^{-4}%)
試行 3	154.1% (280.7%)	1.8% (21.4%)	15.2% (110.3%)	1×10^{-2}% (7×10^{-2}%)
平均	135.0%	4.2%	23.2%	9×10^{-3}%

*()内は真値

表 2　件数上位の平均誤差率

割合	英文	フリガナ	ローマ字	漢字
0.1%	95.22	63.64	25.71	763.30
1%	70.27	23.36	19.53	122.68
10%	19.26	7.06	9.36	16.93
100%	4.58	2.06	2.87	4.58

3.2.3.考察

図 3 と図 4 から、各学習は設定した学習回数で損失が十分に収束していることが確認できる。これらの結果より、以降の実験結果においても学習は十分に行えたものとして扱う。

英文データ、ローマ字データと比較して、フリガナデータの予想選択率の最大値は小さくなった。一方で、全体の平均誤差率はフリガナデータの方が良い値を示した。このことから、フリガナデータを対象にした推定では高い選択率を推定する能力が不足していることが課題点であると言える。

一方、漢字データは予想選択率の最大値が非常に小さく、ほとんど推定ができていない。これは、英文、フリガナと比較して文字の種類が非常に多いためと考えられる。これにより、よく出現する漢字があった場合でも、他の多数の漢字は出現頻度が低いため、1 文字の漢字は出現率が低いという情報に引っ張られ低い選択率を示していることが考えられる。漢字の場合はフリガナ以上に高い選択率を推定する能力が不足していることが課題である。

3.3.文字種類数削減による性能の変化

3.3.1.方法

フリガナデータ、漢字データともに文字の種類数が英文と比較して多いのが、英文より性能が悪くなっている原因である可能性がある。その検証として、文字種類数を削減することで、文字種類数が推定性能に与える影響を確認する。フリガナデータはカタカナで構成されているが、清音のほかに拗音、濁音、半濁音が含まれているため、文字種類数が多い。そこでフリガナデータの文字をすべて清音に変換したうえで学習を行い性能の確認を行うことで、文字種類数が推定性能に与える影響を調べる。性能確認のためのテストデータはすべて清音に変換したもの(以降清音データ)で検証を行う。性能比較は 3 章 2 節と同様の手法を用いる。また、さらなる文字種類数削減を行った場合の性能を確認するため、清音データの「ん」を「あ」に変換、「を」を「い」に変換、といった形で折り返して対応させたデータ(以降折り返しデータ)で同様の検証も行う(図 5)。

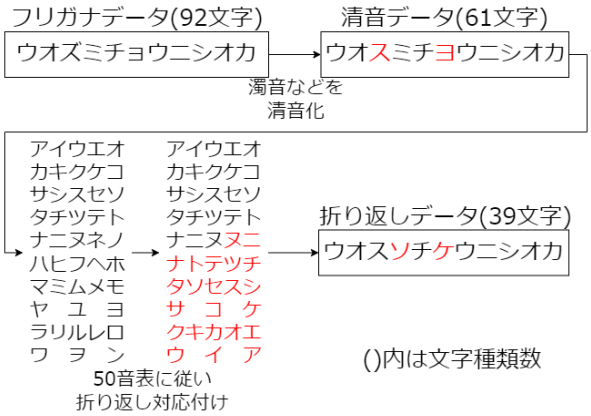


図 5　文字種類数削減の検証で利用するデータ

3.3.2.結果

清音データ、折り返しデータに対して接尾辞の選択率を推定した。表 3 に各データとフリガナデータの予想選択率の最大値を、表 4 に件数真値に従い降順に並べた際の上位の平均誤差率を示す。

表 3　各データの予想選択率の最大値

	フリガナ	清音	折り返し
試行 1	8.3% (13.3%)	1.4% (13.0%)	8.6% (44.6%)
試行 2	2.5% (50.9%)	15.5% (35.3%)	6.9% (3.8%)
試行 3	1.8% (21.4%)	5.6% (13.5%)	8.4% (59.3%)
平均	4.2%	7.5%	8.0%

*()内は真値

表 4　件数上位の平均誤差率

割合	フリガナ	清音	折り返し
0.1%	63.64	22.18	19.33
1%	23.36	13.15	12.71
10%	7.06	5.59	5.85
100%	2.06	1.98	1.95

3.3.3.考察

件数上位 0.1%、1%の平均誤差率に関して、フリガナデータと比較して、清音データ、折り返しデータは低い値を示した。また、予想選択率の最大値に関して、フリガナデータと比較して清音データ、折り返しデータの方がわずかに高い傾向にある。これらより、文字種類数が少ないと比較的高い選択率の推定性能が上がる傾向があると考えられる。一方で、全体の平均誤差率に関しては大きく改善されることはなかった。これらの傾向は、英文データとフリガナデータを比較した際と似ており、フリガナデータにおいて高い選択率を予測することができなかった要因の一つが文字種類数の多さであると考えられる。

なお、この実験は文字種類数が平均誤差率に与える影響を調べる目的で行ったものであり、この手法が推定性能改善の手法として直接利用できるわけではないと考えている。単純な文字の対応付けは出現頻度が低い文字の情報が出現頻度の高い文字に吸収されてしまい、元データに関する推定性能を下げるといった問題が起きる可能性が見込まれる。

3.4.学習回数の調整による性能の変化

3.4.1.方法

多く出現する文字列の重複学習回数を増やして性能改善を図る。フリガナデータに対し出現回数上位 100 件のデータを多く出現する文字列として扱い、それらの重複学習回数を 10 回、100 回、1000 回と増やして性能を比較する(図 6)。

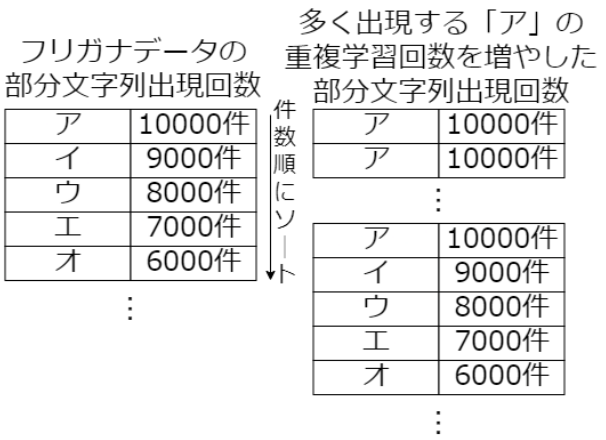


図 6 学習回数を増やす方法

重複学習回数は部分文字列の出現回数を記録したファイルに上位 100 件のデータを重複して追記することで調整する。性能比較は乱数のシード値を変化させ 3 回計測した平均値を用い、3 章 2 節と同様の手法で行う。また、漢字データでも同様の実験を行い、性能を確認する。

3.4.2.結果

フリガナデータ、漢字データの学習回数を変化させ誤差率の平均を算出した。フリガナデータの予想選択率の最大値平均を表 5 に、件数真値に従い降順に並べた際の上位の平均誤差率を表 6 に示す。また、漢字データの予想選択率の最大値平均を表 7 に、件数真値に従い降順に並べた際の上位の平均誤差率を表 8 に示す。割合の欄における 100 件とは、件数真値上位 100 件の場合を指す。

表 5 フリガナデータの予想選択率の最大値平均

	1 回	10 回	100 回	1000 回
試行 1	8.3% (13.3%)	13.4% (13.0%)	14.7% (31.9%)	24.4% (50.8%)
試行 2	2.5% (50.9%)	15.1% (50.9%)	8.5% (32.5%)	24.4% (50.8%)
試行 3	1.8% (21.4%)	29.2% (50.9%)	11.5% (50.8%)	25.0% (50.9%)
平均	4.2%	19.2%	11.5%	24.6%

*()内は真値

表 6 フリガナデータの件数上位の平均誤差率

割合	1 回	10 回	100 回	1000 回
100 件	197.82	10.70	7.05	1.48
0.1%	63.64	29.71	34.90	43.88
1%	23.36	20.42	22.92	27.06
10%	7.06	6.84	7.06	7.64
100%	2.06	2.13	2.18	2.42

表 7 漢字データの予想選択率の最大値平均

	1 回	10 回	100 回	1000 回
試行 1	7×10^{-3}% (7×10^{-4}%)	17.0% (7×10^{-4}%)	0.3% (1.5%)	1.1% (1×10^{-3}%)
試行 2	9×10^{-3}% (7×10^{-4}%)	3×10^{-2}% (5×10^{-3}%)	42.5% (7×10^{-4}%)	1.7% (1×10^{-3}%)
試行 3	1×10^{-2}% (7×10^{-2}%)	3×10^{-2}% (5×10^{-3}%)	4×10^{-3}% (1×10^{-3}%)	1.8% (7×10^{-4}%)
平均	9×10^{-3}%	5.7%	14.3%	1.5%

*()内は真値

表 8 漢字データの件数上位の平均誤差率

割合	1 回	10 回	100 回	1000 回
100 件	1309.50	348.14	499.99	8.73
0.1%	763.30	305.80	408.64	195.41
1%	122.68	69.34	81.74	59.51
10%	16.93	11.52	12.87	11.69
100%	4.58	4.08	3.66	6.23

3.4.3.考察

フリガナデータに関しては、件数上位の学習回数を増やすほど、予想選択率の最大値平均も増加する傾向にある。平均誤差率に関しては、学習回数を増やすほど件数上位の誤差率が低下する傾向にある。これらより、カタカナで構成されたデータにおいては、学習回数を増やすことで性能が改善される可能性がある。特に今回学習回数を増やした上位 100 件のデータに関しては、学習回数を増やすことで大きく予測性能が向上した。一方、学習回数を増やすほど全体の平均誤差率が低下したことから、学習回数を増やしすぎると全体性能が悪化する可能性がある。総合的に見て、カタカナで構成されたデータにおいては件数上位の学習回数を増やすことで性能改善が見込まれるが、回数を増やしすぎると全体性能が悪化する可能性があるため適度なバランスの調整に抑える必要がある。

漢字データの場合も同様に、件数上位の学習回数を増やすほど、予想選択率の最大値平均も増加する傾向にあるが、本来は出現頻度の低い文字列の選択率を高く見積もる例が見られる。また、件数上位の学習回数を増やすことによる性能悪化がフリガナより激しく、件数上位の推定性能が十分になる前に全体の性能が大きく悪化すると考えられる。そのため、漢字で構成されたデータにおいては上位件数の学習回数を増やす方法のみでは性能改善を行う上で不十分であると考えられる。

4. ユースケース検証

日本全国住所データ 150072 件を 10 回繰り返し zenkoku_table として、全国の法人データ 5401729 件を 5 回繰り返し zenkoku_corp としてデータベースに登録する[13][15]。各表の代表的な構成カラムは表 7 のとおりである。図 7 の SQL を実行する場面を考える。zenkoku_table のうち「ウオズミ」を含むのは 100 件である。また、zenkoku_table のデータ数を調整しながら図 8 のクエリを実行し、登録件数ごとの結合方法による実行時間を計測する。データ数最大時に、このクエリを実行すると 148323975 件となる。図 9 に実行時間計測結果を示す。zenkoku_corp の件数が 400000 件の辺りまで索引ありネストループ結合の方が早くなっている。

```
SELECT COUNT(*)
FROM zenkoku_corp AS c
   JOIN (
   SELECT *
   FROM zenkoku_table
   WHERE zenkoku_table.町域カナ
   LIKE "%ウオズミ%")AS z
ON z.郵便番号 = c.郵便番号;
```

図 7　2 表結合の SQL

表 7　各表の代表的な構成カラム

zenkoku_table(280MB)	zenkoku_corp(5003MB)
郵便番号 VARCHAR(255), 都道府県 VARCHAR(255), 都道府県カナ VARCHAR(255), 市区町村 VARCHAR(255), 市区町村カナ VARCHAR(255), 町域 VARCHAR(255), 町域カナ VARCHAR(255),	法人番号 VARCHAR(255), 名称 VARCHAR(255), 法人種別 VARCHAR(255), 都道府県 VARCHAR(255), 市区町村 VARCHAR(255), 丁目番地等 VARCHAR(255), 郵便番号 VARCHAR(255)

```
SELECT COUNT (*)
FROM zenkoku_corp AS c
JOIN zenkoku_table AS z ON z.郵便番号=c.郵便番号;
```

図 8　実行時間の差を計測する SQL

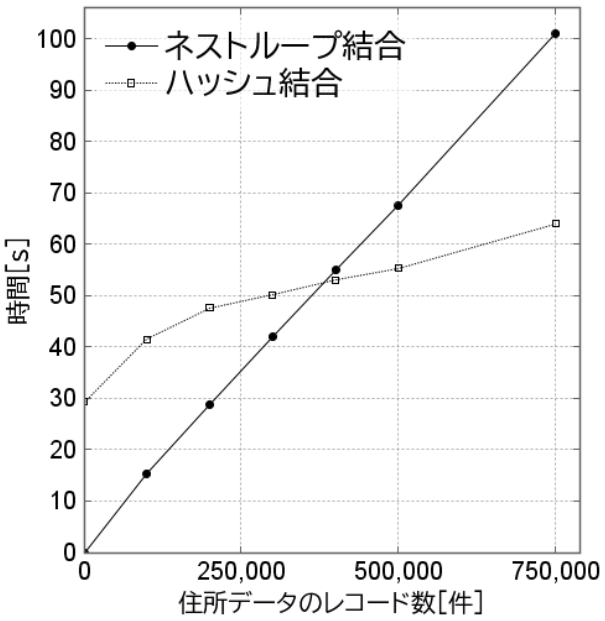


図 9　結合方法による実行時間の差

ここで、3 章 4 節において学習回数 1000 回で推定した「ウオズミ」の予測値の一つである、58 件を用いると、ネストループ結合の方が早いと見積もられる。データベースでは LIKE 述語の選択率として一律に 40%などの固定値が使用されていることが多い[2]。選択率 40%とすると、件数は 600110 件と見積もられ、ハッシュ結合の方が早いと判断される。実際にネストループ結合で実行すると 1.69 秒、ハッシュ結合で実行すると 25.48 秒となり、ネストループ結合の方が早いという結果であった。

5. まとめ

カタカナ、漢字のデータにおける選択率推定に Astrid-Embed の手法を直接利用すると、高い選択率を予測する性能が低くなることが問題となると確認できた。また、漢字データの場合はカタカナデータ以上に低い性能を示すことが確認できた。高い推定率を学習する際の性能改善手法として、高い選択率のデータの学習回数を調整する方法がある程度有効であり、これにより高い選択率を予測する性能が向上することが確認できたが、全体の性能を悪化させる傾向にあるため適度な学習回数に留めておく必要がある。漢字データに関しては学習回数の調整のみでは英文を超える性能が出ることが確認できなかったため、さらなる改善が必要である。

参考文献

[1] Suraj Shetiya, Saravanan Thirumuruganathan, Nick Koudas, and Gautam Das. Astrid: accurate selectivity estimation for string predicates using deep learning. Proceedings of the VLDB Endowment, 14(4),2020.

[2] FUJITSU. Software symfoware Server V12.4.1 チューニングガイド.
https://software.fujitsu.com/jp/manual/manualfiles/m200010/j2ul1761/07z200/j2ul-1761-07z0.pdf.

[3] Surajit Chaudhuri, Venkatesh Ganti, and Luis Gravano. 2004. Selectivity estimation for string predicates: Overcoming the underestimation problem. In Proceedings. 20th International Conference on Data Engineering. IEEE, 227–238.

[4] HV Jagadish, Olga Kapitskaia, Raymond T Ng, and Divesh Srivastava. 2000. Onedimensional and multi-dimensional substring selectivity estimation. The VLDB Journal—The International Journal on Very Large Data Bases 9, 3 (2000), 214–230

[5] P Krishnan, Jeffrey Scott Vitter, and Bala Iyer. 1996. Estimating alphanumeric selectivity in the presence of wildcards. In ACM SIGMOD Record, Vol. 25. ACM, 282–293

[6] Shohedul Hasan, Saravanan Thirumuruganathan, Jees Augustine, Nick Koudas, and Gautam Das. 2020. Deep Learning Models for Selectivity Estimation of Multi Attribute Queries. SIGMOD (2020).

[7] Benjamin Hilprecht, Andreas Schmidt, Moritz Kulessa, Alejandro Molina, Kristian Kersting, and Carsten Binnig. 2019. DeepDB: Learn from Data, not from Queries! arXiv preprint arXiv:1909.00607 (2019).

[8] Andreas Kipf, Michael Freitag, Dimitri Vorona, Peter Boncz, Thomas Neumann, and Alfons Kemper. [n.d.]. Estimating Filtered Group-By Queries is Hard: Deep Learning to the Rescue. ([n. d.]).

[9] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, and Alfons Kemper. 2019. Learned Cardinalities: Estimating Correlated Joins with Deep Learning. CIDR (2019).

[10] Dan Gusfield. Algorithms on stings, trees, and sequences: Computer science and computational biology. Acm Sigact News, 28(4):41–60,1997.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[12] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In Proceedings of the IEEE international conference on computer vision, pages 2840–2848, 2017.

[13]住所.jp：http://jusyo.jp/csv/new.php(2023.10.1)

[14]miurahr/pykakasi : https://github.com/miurahr/pykakasi(2023.10.1)

[15] 国税庁 法人番号公表サイト：https://www.houjinbangou.nta.go.jp/dow

グラフデータベースにおける Worst-case Optimal Join の最適化

伊藤　寿浩†　　塩川　浩昭††

† 筑波大学 情報学群情報科学類　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1-1-1
E-mail: †itou-toshihiro@kde.cs.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　NoSQL データベースの一つであるグラフデータベースは，データをグラフ構造に基づいて表現しており，ソーシャルネットワークや知識の表現などの分野で広く利用されている．グラフデータベースに対する一般的な問合せにはグラフパターンを含む問合せがある．本稿では，このような問合せにおいて DISTINCT 句を含んだ場合に焦点を当てる．既存の問合せ処理手法の一つに Worst-case Optimal Join（WCOJ）が挙げられるが，DISTINCT 句を含む場合に不要な計算が生じ，結合結果が冗長になるという問題がある．そこで本稿では WCOJ を拡張し，問合せ結果に不要な探索の打ち切りや変数の探索順序の最適化を行うことで，実行時間の高速化を図る．さらに大規模な PG を用いた性能評価を行い，最先端手法と比較して提案手法は DISTINCT を含む問合せを最適に処理できることを示す．
キーワード　問合せ最適化，グラフデータベース

1　はじめに

グラフデータベースはデータをグラフ構造として保持するデータベースであり，ソーシャルネットワーク [1] やバイオインフォマティクス [2]，知識の表現 [3] などの広い分野で利用されている．グラフデータベースが多くの分野から注目されている理由として，複雑なグラフデータを柔軟に扱えることが挙げられる．一般的なグラフデータベースは，プロパティグラフと呼ばれるグラフ構造を拡張したデータモデルに対応している．図 1 のように，プロパティグラフでは頂点および辺に対してラベル，プロパティが与えられており，グラフに付加された多様な情報を表現できる．また近年では，十億以上の辺から構成される大規模なグラフデータが増加しているため，グラフデータベースにおける問合せ処理の高速化についての研究は重要であり，活発に行われている [4] [5] [6] [7].

グラフデータベースにおける問合せ処理の一つとして，グラフパターンを含む問合せが挙げられる．これは，与えられたグラフパターンに一致する部分グラフをデータベースが保持するグラフ中から列挙する問題である．例えば，図 2 のグラフパターンと図 1 のプロパティグラフが与えられるとき，ラベルとプロパティを考慮して $\{(a_2, t_2, p_1), (a_2, t_4, p_1), (a_2, t_4, p_2), (a_2, t_5, p_2)\}$ が問合せ結果として得られる．これらの問合せは一般的であり，多くのアプリケーションに有用である．具体例として，金融ネットワーク上の不正検出 [8] を挙げる．詐欺等の不正処理では取引内容に典型的なパターンを持つことが知られている．これをグラフパターンとして表現し，金融ネットワーク上でパターン検索を行うことでこのような不正を検出できる．

グラフパターンを含む問合せにかかる計算コストは高価で大規模グラフでは膨大な処理時間を要するため，これまでに多くの高速化手法が提案されてきた [9] [10] [11] [12]．その中でも主要なアプローチとして，バイナリ結合を用いるものが挙げられる [9]．これはグラフパターンを複数のリレーションの自然結合に変換後，バイナリ結合を用いて処理する．このアプローチはグラフパターンと明らかに一致しない候補を除外できるため，処理性能が飛躍的に向上した．しかし，このアプローチは問合せ結果の数に対して中間結果が爆発的に生成される可能性があり，特定のグラフパターンでは多くの処理時間を要する．2 つのリレーション R, S の属性集合 A_R, A_S が $A_R \supseteq A_S$ となる場合，$R \bowtie S$ のリレーション T は $|T| \leqq |R|$ となる．したがって，結合前のリレーションのサイズが結合後のサイズより大きく，中間結果の爆発が起こる．このような状況が起こるグラフパターンには，閉路を持つグラフなど一般的な問合せに用いられるものも多く，重要な研究課題とされていた．

これに対して，Ngo らは中間結果の爆発を回避するアプローチとして Worst-case Optimal Join（WCOJ）を提案した [10]. WCOJ とは，実行時間が問合せ結果の最悪サイズの上界に比例する結合アルゴリズムを指す．一般的な WCOJ のアルゴリズムでは問合せを処理するため，(i) グラフパターンのある頂点 v に対して割り当て可能なプロパティグラフ上の頂点の列挙，(ii) 列挙した各頂点を v に順番に割り当てる処理を交互に行う．これらの処理は中間結果の爆発を回避できるため，バイナリ結合で課題として挙げられたケースにおいても高い性能を維持できる．そのため WCOJ は多くの注目を集めており，WCOJ に基づく様々な手法がこれまで提案されてきた [11] [12] [13].

本研究ではグラフパターンを含む問合せのうち，DISTINCT 句を含む場合に焦点を当てる．DISTINCT 句では，グラフパターンの頂点の部分集合 V_D が指定される．問合せ結果は，グラフパターンと一致する全ての部分グラフの V_D に対応する割り当てを抽出し，要素間の重複を排除したものとなる．例として，図 2 のグラフパターンと図 1 のプロパティグラフが与えられ，DISTINCT 句で p が指定されたときを考える．この場合，グラフパターンと一致する全ての部分グラフから p に対応する割り当てとして $\{\{p_1\}, \{p_1\}, \{p_2\}, \{p_2\}\}$ が抽出される．これら

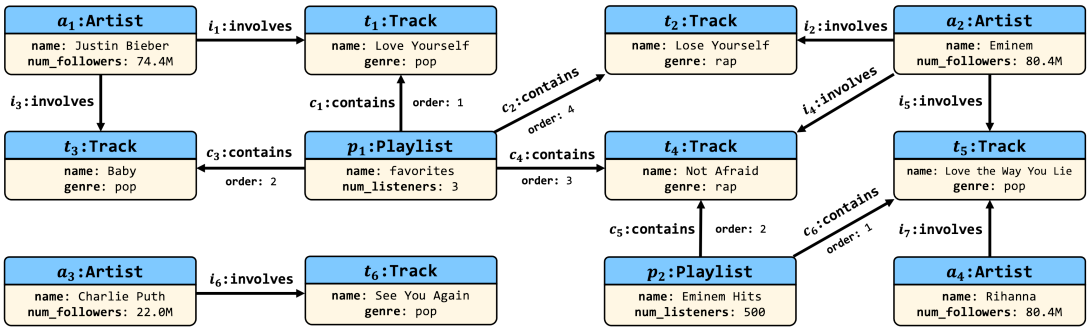


図 1　プロパティグラフの例

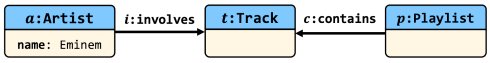


図 2　グラフパターンの例

の重複を排除することで $\{\{p_1\},\{p_2\}\}$ が問合せ結果として得られる．DISTINCT 句は，頂点に対する条件指定や問合せ処理の効率化など多くの状況で利用される．

しかし，WCOJ を含む既存の問合せ処理手法は DISTINCT 句を含む場合を特別に考慮しておらず，大きな処理時間を要する．DISTINCT 句を含む問合せ処理を行う場合，従来手法ではグラフパターンと一致する部分グラフを全て列挙する．そのため，V_D に対応する割り当てが同一である部分グラフを複数個生成する可能性がある．このとき，これらの部分グラフのうち 1 つを除いて残りの部分グラフは問合せ結果に不要であり，それらの探索は冗長な処理である．我々が行った実験では，従来手法は V_D のある割り当てについて平均 9.2×10^4 個の部分グラフを生成していた．このことから，生成する部分グラフの多くが問合せ結果に不要であることがわかる．

1.1　本研究の貢献

本研究では WCOJ を拡張し，グラフデータベースに対する問合せのうち，DISTINCT 句を含む場合の高速な処理手法を提案する．提案手法は従来手法の課題であった問合せ結果に不要な部分グラフの生成を回避し，ある V_D の割り当てに対して単一の部分グラフのみを生成する．これを実現するため，不要と判断できる部分グラフの探索を打ち切る．また，提案手法はキャッシング処理を導入し，探索回数の更なる削減を図る．WCOJ アルゴリズムの主な課題として，同一の割り当ての再探索が挙げられていた．これは WCOJ アルゴリズムの再帰的な頂点の探索に起因し，グラフパターンの頂点に以前割り当てられた頂点が再度割り当てられることで起こる．上記の課題に対し，Cached Trie Join (CTJ) [12] では WCOJ にキャッシング処理を導入し，割り当てられた頂点を再利用することで回避した．しかし，CTJ で用いられるキャッシュは DISTINCT 句を含む問合せ処理を行う上で不要な頂点の割り当てについても保持するため，空間コストが高価である．そこで提案手法は CTJ を基に，問合せ処理に必要な頂点の割り当てのみ保持するキャッシュを作成する．さらに，提案手法の処理性能を向上させるため，実行計画の最適化について取り組む．既存研究より，実行計画は処理性能に大きな影響を及ぼすことが知られている [13]．

記号	定義
G	プロパティグラフ
P	グラフパターン
V	頂点集合
E	辺集合
ρ	辺から頂点の組にマッピングする関数
λ	頂点あるいは辺からラベル集合にマッピングする関数
π	頂点あるいは辺とプロパティ名からプロパティ値にマッピングする関数
V_D	グラフパターン中の DISTINCT な頂点集合

表 1　本論文で用いる記号の定義

特に WCOJ の場合，割り当ての列挙を行う頂点の順序は処理性能を大きく左右する．本研究では，カーディナリティ推定を用いて実行コストに基づく順序の最適化を行う．カーディナリティは実行計画を決定する上で重要な指標であり，その推定手法は表形式データベースや RDF データベースを中心に研究されてきた [14]．提案する推定手法は，探索の打ち切りによる探索回数の削減を考慮しているため，提案手法の問合せ処理に対するカーディナリティを適切に推定できる．これらのアプローチにより，提案手法は以下の特性を有する．

- **高速性**：提案手法は近年提案された手法と比較して平均 69 倍高速である (4.2 節)．
- **広範性**：提案手法のアプローチである実行計画の最適化はプロパティグラフに対する問合せを想定しており，選択演算を含む頂点が存在する場合に対しても適切に対処できる．

本稿の構成は次の通りである．2 節で本稿の前提となる知識について概説する．3 節にて提案手法の詳細について説明し，4 節において提案手法の評価と分析を行う．5 節にて関連研究について述べ，6 節において本稿をまとめる．

2　事 前 準 備

表 1 にて本稿で用いる記号とその定義を示す．

2.1　問 題 定 義

本研究では，プロパティグラフに対する問合せ処理を考える．プロパティグラフの定義を以下に示す．

[**定義 1**]（プロパティグラフ） プロパティグラフは $G = (V, E, \rho, \lambda, \pi)$ で定義される．ただし，V, E はそれぞれ頂点集合，辺集合である．また，$\rho : E \to (V \times V)$ は辺から頂点の組，$\lambda : (V \cup E) \to 2^{\mathcal{L}}$ は頂点および辺からラベル $\mathcal{L}$ の集合，$\pi : (V \cup E) \times \mathcal{K} \to \mathcal{V}$ は頂点と辺，プロパティ名 $\mathcal{K}$ からプロパティ値 $\mathcal{V}$ にマッピングする関数である．

次に，本研究で対象とする問合せについて述べる．図 3 に

```
MATCH (a:Artist)-[:involves]->(:Track)
  <-[:contains]-(p:Playlist)
WHERE a.name = "Eminem"
RETURN DISTINCT p
```

図 3　本研究で対象とする問合せの例（問合せ言語は cypher）

て対象とする問合せを示す．問合せには，プロパティグラフで表現されたグラフパターンと DISTINCT である頂点集合（以降，V_D とする）を含む．MATCH 句ではグラフパターンが与えられており，各頂点および辺にはラベルおよびプロパティを含む可能性がある．例として図 3 の問合せ中の頂点 a では，割り当て可能なプロパティグラフ上の頂点はラベルが Artist であり，プロパティ name の値が Eminem の頂点となる．また，RETURN 句によって V_D が指定される．

最後に，本研究で扱う問合せの定義を述べる．以降，グラフパターンを $P(V, E, \rho, \lambda, \pi)$，問合せの対象となるプロパティグラフを $G(V_G, E_G, \rho_G, \lambda_G, \pi_G)$ と表す．P と一致する G 上の部分グラフ H は以下のように定義される．

［**定義** 2］（部分グラフ同型） グラフパターン $P(V, E, \rho, \lambda, \pi)$ とプロパティグラフ G 上の部分グラフ $H = (V', E', \rho', \lambda', \pi')$ があるとする．このとき，任意の $u, v \in V, e \in E, \rho(e) = (u, v)$ について，$f(u), f(v) \in V', g(e) \in E'$ かつ以下の条件を満たす全単射関数 $f : V \to V', g : E \to E'$ が存在するならば，H は P と一致する．

$$\rho'(g(e)) = (f(u), f(v)),$$
$$\lambda(u) \subseteq \lambda'(f(u)), \forall k_u \in K_u, \pi(u, k_u) = \pi'(f(u, k_u)),$$
$$\lambda(v) \subseteq \lambda'(f(v)), \forall k_v \in K_v, \pi(v, k_v) = \pi'(f(v, k_v)),$$
$$\lambda(e) \subseteq \lambda'(g(e)), \forall k_e \in K_e, \pi(e, k_e) = \pi'(g(e, k_e)) \quad (1)$$

ただし，$K_u, K_v, K_e \subseteq \mathcal{K}$ は u, v および e に付加されるプロパティ名の集合とする．

定義 2 に従い，グラフパターンを含む問合せは以下のように定義される．

［**定義** 3］（グラフパターンを含む問合せ） グラフパターン P とプロパティグラフ G があるとき，グラフパターンを含む問合せは P と一致する G 上の全ての部分グラフからなる集合 $\mathcal{M}$ を求める問題である．

定義 3 より，本研究の対象である DISTINCT 句を含む問合せは以下のように定義される．

［**定義** 4］（DISTINCT 句を含む問合せ） グラフパターン P と一致するプロパティグラフ G 上の全ての部分グラフ集合 $\mathcal{M}$，DISTINCT な頂点集合 V_D があるとき，DISTINCT 句を含む問合せは式 2 の条件を満たす重複なし集合 $\mathcal{D}$ を求める問題である．

$$\mathcal{D} = \{H_{[V_D]} \mid H \in \mathcal{M}\} \quad (2)$$

ただし，$H_{[V_D]}$ は P に一致する部分グラフ H のうち，V_D に対応する割り当てを指す．

2.2 Worst-case Optimal Join (WCOJ)

本節では，提案手法の基となる WCOJ の概要について述べる．一般的な WCOJ のアルゴリズムを Algorithm 1, 2 に示す．WCOJ はグラフパターン P の各頂点に対する割り当て μ を保持し，内部で Algorithm 2 の RecursiveJoin を呼び出す．RecursiveJoin では，現在探索している頂点 v_d に割り当て可能なプロパティグラフ G 上の頂点を列挙する．6 行目より列挙した頂点を順番に取り出し，v_d の割り当て $\mu(v_d)$ を更新する．この割り当ての列挙処理を探索順序 $\langle v_1, \ldots, v_n \rangle$ に従って，P の全ての頂点に対して再帰的に行うことで，P と一致する部分グラフを探索する．v_n に G 上の頂点が割り当てられた場合，問合せ結果 Q に μ を追加する．

Algorithm 1 WCOJ

Input: G; P; $\langle v_1, \ldots, v_n \rangle$, 探索順序;
Output: Q, 問合せ結果;
1: $Q \leftarrow \emptyset$;
2: **for** $d = 1, \ldots, n$ **do**
3: 　$\mu(v_d) = \perp$;
4: RecursiveJoin(1);
5: **return** Q;

Algorithm 2 Recursive Join

Input: d, 探索する頂点のインデックス;
1: **function** RecursiveJoin(d)
2: 　**if** $d = n + 1$ **then**
3: 　　$Q \leftarrow Q \cup \mu$;
4: 　　**return**
5: 　**for all** v_d に割り当て可能な G 上の全ての頂点 a に対して **do**
6: 　　$\mu(v_d) = a$;
7: 　　RecursiveJoin($d + 1$);
8: 　$\mu(v_d) = \perp$;

例として，図 2 のグラフパターン P，図 1 のプロパティグラフ G が与えられた場合の WCOJ アルゴリズムの動作を説明する．P の頂点の探索順序を $\langle a, t, p \rangle$ とすると，アルゴリズムはまず a に割り当て可能な頂点を列挙する．a のラベルおよびプロパティを考慮すると $\{a_2\}$ が得られる．RecursiveJoin 6 行目より，a の割り当てを a_2 として t の割り当ての列挙に移る．a の割り当てから，t の割り当ては a_2 とラベル involves を持つ辺で隣接する必要がある．これと t 自体のラベルを考慮すると，$\{t_2, t_4, t_5\}$ が t に割り当て可能な頂点として得られる．a の探索のときと同様に，t の割り当てを t_2 として p の割り当ての列挙に移る．p に割り当て可能な頂点は $\{p_1\}$ であるため，p の割り当てを p_1 とする．これにより $v_n = p$ の割り当てが見つかったため，RecursiveJoin 2 行目より $\mu = (a_2, t_2, p_1)$ を問合せ結果 Q に追加する．以降，同様に P の各頂点の割り当てを更新することで，P と一致する部分グラフを逐次的に得る．

WCOJ アルゴリズムで DISTINCT 句を含む問合せを処理する場合，問合せ結果 Q から P 上の DISTINCT な頂点集合 V_D に対応する割り当てを抽出し，重複を排除することで処理可能である．しかし，この処理ではグラフパターンと一致する部分グラフを全て生成するため，V_D に対する割り当てが同一である部分グラフが生成された場合に冗長な処理が生じる．

3 提 案 手 法

本節では提案手法について述べる．提案手法は DISTINCT 句を含む問合せに対して，従来手法より高速に処理できる．はじめに提案手法を構成する基本的なアイデアについて述べ，その詳細について 3.2 節以降で説明する．

3.1 基本アイデア

WCOJ を含む既存の問合せ処理手法は DISTINCT 句を含む場合，同一の V_D の割り当てを持つ部分グラフを複数個生成する可能性があり，結果として大きな処理時間を要する．ゆえに，処理時間を短縮するためには，生成する部分グラフを必要最小限にすることが重要である．

そこで本研究では 2.2 節の WCOJ アルゴリズムを拡張し，V_D のある割り当てに対して単一の部分グラフのみを生成するように改良する．基本的なアイデアは，現在の V_D の割り当てが問合せ結果に含まれる，あるいは現在の頂点に対して割り当ての更新が不要であることが自明な場合，その頂点の探索を打ち切ることである．また，提案手法は CTJ を基に，DISTINCT 句を含む問合せ処理に特化したキャッシング処理を組み込む．具体的には，P の頂点の割り当てを保持する通常キャッシュと部分グラフが存在するかについて保持する判定キャッシュを使い分けることで，空間コストが低いキャッシング処理を実現する．さらに，本研究ではカーディナリティ推定を用いたグラフパターンの頂点の探索順序の最適化により，広範なグラフパターンに対する安定した処理性能の実現を目指す．カーディナリティは問合せ処理の際，P の各頂点に割り当てられる G 上の頂点の総数であり，この値が最小となる探索順序を選択する．カーディナリティを適切に推定するため，提案手法は適切な統計情報を利用する．また，探索の打ち切りおよびキャッシング処理による探索回数の削減についても適切に考慮した推定を行う．

以降の節では，探索の打ち切り，キャッシング処理，実行計画の最適化の詳細についてそれぞれ説明し，最後にアルゴリズムの概要について述べる．

3.2 探索の打ち切り

探索の打ち切りのタイミングには，(i) V_D の割り当てが更新された際にその割り当てが問合せ結果に含まれる場合，(ii) 現在の頂点の割り当てを更新しても問合せ結果に不要であることが自明な場合の 2 つが存在する．このうち (i) は，V_D のうち最も探索順序が遅い頂点 v_{last} の割り当ての更新の際に行われる．v_{last} の割り当てが更新される際，現在の V_D の割り当てが問合せ結果に含まれるかについて判定する．もし含まれる場合，v_{last} 以降の頂点の探索が不要であることが自明なため，v_{last} の割り当てを再更新する．一方，(ii) は定義 6 のコア集合に含まれない P の頂点の割り当ての更新時に行われる．

[**定義 5**]（ancestors） グラフパターン P，P の頂点の探索順序 $\langle v_1, \ldots, v_n \rangle$ が与えられたとき，P 上の単純パス，かつ全ての辿る頂点が始点の探索順序より早いパスを $s = (v_i, v_j, \ldots)$ とする．P の頂点 v を始点として v_1 へ辿る全ての s からなる集合を S とするとき，P 上の頂点 v の *ancestors* を表す $ancestors(v)$ は P の頂点の部分集合であり，式 3 の条件を満たす．

$$ancestors(v) = \left\{ a \mid a \in \bigcup_{s \in S} s, a \neq v \right\} \tag{3}$$

[**定義 6**]（コア集合） DISTINCT 句で指定されたグラフパターン P の頂点集合 V_D が与えられたとき，コア集合 C は P の頂点の部分集合であり，式 4 の条件を満たす．

$$C = V_D \cup \bigcup_{v \in V_D} ancestors(v) \tag{4}$$

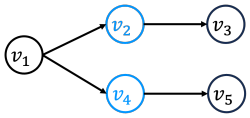


図 4 グラフパターンの例（水色で示される頂点集合を V_D とする）

図 4 のグラフパターン P を用いて，コア集合の例を示す．探索順序を $\langle v_1, v_2, v_3, v_4, v_5 \rangle$ としたとき，$ancestors(v_2) = \{v_1\}$，$ancestors(v_4) = \{v_1\}$ となる．よって，コア集合 $C = \{v_1, v_2, v_4\}$ となる．

DISTINCT 句を含む問合せは全ての V_D の割り当てに対して，P と一致する部分グラフの存在を判定することで処理できる．また定義 6 から，コア集合に含まれない頂点 v_O はコア集合に従属している．そのため，v_O の割り当ての更新が V_D の割り当ての更新に寄与しないのは自明である．ゆえに，現在の V_D の割り当てに対して P と一致する部分グラフが存在する場合，v_O の割り当ての更新を打ち切ることが可能である．

図 4 の P を用いて，探索順序を $\langle v_1, v_2, v_3, v_4, v_5 \rangle$ としたときの探索の打ち切りの具体的な動作を説明する．前述の通り，図 4 のコア集合は $C = \{v_1, v_2, v_4\}$ となる．打ち切り可能なタイミングの (i) に該当するケースは，V_D のうち最も探索順序が遅い v_4 の割り当ての更新時である．更新された V_D の割り当てを問合せ結果と照合することで，既に問合せ結果に含まれる割り当てについての探索を回避できる．(ii) に該当するケースは，コア集合に含まれない v_3, v_5 の割り当ての更新時である．新たに v_5 の割り当てが見つかったとき，現在の V_D に対応する割り当てが問合せ結果に追加される．V_D の割り当ては頂点 v_1 の割り当てに依存していることから，v_3, v_5 の割り当ての更新は V_D の割り当ての更新に寄与せず，問合せ結果に不要である．ゆえに更新処理を打ち切り，v_2, v_4 の割り当ての再更新へ移る．

3.3 キャッシング処理

まず，CTJ のキャッシング処理の概要について説明する．グラフパターン P の頂点の部分集合 V_K に対して従属する頂点集合が存在するとき，CTJ は V_K をキーとして従属する頂点集合に対応する割り当てをキャッシュで保持する．これにより，探索中の V_K の割り当てがキャッシュのキーに含まれる場合，保持していた割り当てを再利用することで割り当ての列挙処理を回避できる．このような従属性を発見するため，CTJ は P から以下のように定義される Ordered Tree Decomposition (Ordered-TD) を作成する．

[**定義 7**]（Ordered Tree Decomposition (Ordered-TD)） グラフパターン P の Ordered-TD は $\langle t, \chi \rangle$ で定義される．ただし t は順序木，χ は t のノードから *bag* と呼ばれる P の頂点の部分集合へマッピングする関数であり，t, χ は以下の条件を満たす．

- P の任意の辺 e について，$\rho(e) \subseteq \chi(x)$ を満たす t のノード x が存在する．
- P の任意の頂点 v について，ノード集合 $X = \{x \mid v \in$

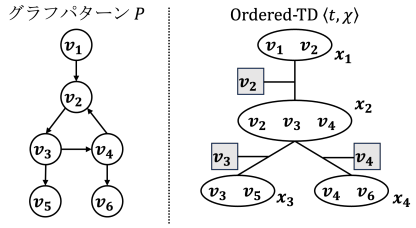


図 5　グラフパターン P と P から作成される Ordered-TD の例

$\chi(x)\}$ は t 上で連結である．

続いて，$adhesion, owned, owner$ を定義する．x の親ノードを x_p とすると，x の $adhesion$ を $adhesion(x) = \chi(x) \cap \chi(x_p)$ と定義する．ただし，ルートノードの $adhesion$ は空集合とする．また，x の $owned$ を $owned(x) = \chi(x) \backslash adhesion(x)$ と定義する．$owner$ は P の頂点 v から $v \in owned(x)$ を満たすノード x へマッピングする関数と定義する．

P から生成されうる Ordered-TD は複数個存在し，CTJ は各ノードの $adhesion$ に割り当て可能な頂点の組数などを評価軸として Ordered-TD を 1 つ選択する．P の頂点の探索順序は選択した Ordered-TD をルートノードから行きがけ順で探索し，各ノードの $owned$ の探索順序を連結したものとなる．以降，ノード x の子ノードを $c_1, \ldots, c_k$ とした場合，$owned(x) \cup owned(c_1) \cup \cdots \cup owned(c_k)$ となる P の頂点の部分集合を $territory(x)$ と表す．定義 7 より，$territory(x)$ が $adhesion(x)$ に従属することは明らかであり，$adhesion(x)$ をキーとして $territory(x)$ に対応する割り当てをキャッシュに保持する．

図 5 に Ordered-TD $\langle t, \chi \rangle$ の例を示す．t のノード x_2 に着目すると，$\chi(x_2) = \{v_2, v_3, v_4\}, adhesion(x_2) = \{v_2\}, owned(x_2) = \{v_3, v_4\}$ となる．また，$owner(v_6) = x_4$ となる．キャッシュは $adhesion$ が存在する x_2, x_3, x_4 について作成される．例として x_2 のキャッシュでは，$adhesion(x_2) = \{v_2\}$ をキーとして，$territory(x_2) = \{v_3, v_4, v_5, v_6\}$ に対応する割り当てを保持する．

しかし DISTINCT 句を含む問合せでは，問合せ結果で V_D の割り当てのみを必要とするため，保持不要な頂点が $territory(x)$ 中に存在する可能性がある．そこで提案手法は，通常キャッシュと判定キャッシュを用いて，必要な頂点の割り当てのみを保持したキャッシュから正確な問合せ結果を得る．通常キャッシュは $territory(x) \cap V_D$ に対応する割り当てを保持するキャッシュである．また，判定キャッシュはあるキーに対して，そのキーを含む P と一致する部分グラフが存在するかを保持するキャッシュである．キャッシュの種類について，x が以下の 2 つの条件を満たす場合，x には通常キャッシュが適用される．

（1）$territory(x) \cap V_D \neq \emptyset$.

（2）ある $u \in adhesion(x), v \in V_D$ について，$u \in descendants(v)$.

ただし，$descendants(v)$ は以下のように定義される．

［**定義** 8］（descendants） グラフパターン P，P の頂点の探索順序 $\langle v_1, \ldots, v_n \rangle$ が与えられたとき，P 上の単純パス，かつ全ての辿る頂点が始点の探索順序より遅いパスを $l = (v_i, v_j, \ldots)$ とする．P の頂点 v を始点とした全ての l からなる集合を L とするとき，P 上の頂点 v の $descendants$ を表す $descendants(v)$ は P の頂点の部分集合であり，式 5 の条件を満たす．

$$descendants(v) = \left\{ d \mid d \in \bigcup_{l \in L} l, d \neq v \right\} \tag{5}$$

1 つ目の条件を満たさない場合，ある V_D の割り当てに対して P と一致する部分グラフが存在するか判断できれば良いため，判定キャッシュを適用する．また 2 つ目の条件を満たさない場合，V_D の割り当ては $adhesion(x)$ の割り当てに従属している．そのため，キャッシュヒットが起きた場合に得られる V_D の割り当ては，既に問合せ結果に存在することが自明である．よって，判定キャッシュを適用する．

図 5 に提案手法のキャッシュの例を示す．なお，$V_D = \{v_1, v_5\}$，探索順序は $\langle v_1, v_2, v_3, v_4, v_5, v_6 \rangle$ とする．キャッシュは $adhesion$ が存在するノード x_2, x_3, x_4 で保持される．このうち，x_2, x_3 は上記の 2 つの条件を満たすため，通常キャッシュを適用してどちらも v_5 に対応する割り当てを保持する．一方，x_4 は $territory(x_4) \cap V_D = \emptyset$ であるため，判定キャッシュを適用して $adhesion(x_4)$ の割り当てに対して P と一致する部分グラフが存在するかについて保持する．v_2 および v_3 の割り当てについてキャッシュヒットが起きた場合，保持されていた v_5 の全ての割り当てを再利用することで，$territory(x_2), territory(x_3)$ に割り当て可能な頂点の列挙処理を回避できる．また，v_4 の割り当てについてキャッシュヒットが起きた場合，v_4 の割り当てを持つ部分グラフの存在について $territory(x_4)$ の割り当ての探索を行わずに判定できる．これにより，問合せ結果を正確に得られるとともに CTJ と比較して空間コストを削減できる．

3.4　実行計画の最適化

本節では，実行計画の最適化について説明する．提案手法は 3.3 節のキャッシング処理のために Ordered-TD を作成しており，頂点の探索順序は Ordered-TD に依存する．Ordered-TD はグラフパターン P から複数個生成される可能性があり，これらの中から処理コストの低い Ordered-TD を選択する必要がある．そこで提案手法は Ordered-TD の選択および各ノード内の探索順序を決定するため，カーディナリティ推定（Cadinality Estimation, CE）に基づくコストベースの最適化を行う．カーディナリティとは実行計画を決定するための指標であり，以下のように定義される．

［**定義** 9］（カーディナリティ） グラフパターン P のある頂点 $v \in V$ のカーディナリティを表す $cardinality(v)$ は，v に割り当て可能なプロパティグラフ G 上の頂点の総数と定義される．割り当て可能な頂点を集計する際，頂点の重複は許容される．また，Ordered-TD $\langle t, \chi \rangle$ のあるノードにおける探索順序を表す $cardinality(O)$ は式 6 で定義される．

$$cardinality(O) = \sum_{v \in owned(x)} cardinality(v) \tag{6}$$

式 6 を基に，Ordered-TD $\langle t, \chi \rangle$ のカーディナリティを表す $cardinality(t)$ は式 7 で定義される．ただし，X は t のノード集合，O_x はノード x の $owned$ 内の探索順序を表す．

$$cardinality(t) = \sum_{x \in X} cardinality(O_x) \qquad (7)$$

カーディナリティが小さい場合，その探索順序は各頂点の割り当ての更新回数が少ないため，一般的に処理時間が良い．よって，カーディナリティに着目した場合の最適な Ordered-TD および各ノード内の探索順序は式 8 により決定される．

$$\underset{t}{\operatorname{argmin}}\ cardinality(t),\quad \underset{O}{\operatorname{argmin}}\ cardinality(O) \qquad (8)$$

3.3 節で説明した方法により，Ordered-TD から頂点の探索順序を決定する．

以降では，3.4.1 節にて単一ノードからなる Ordered-TD の CE について説明し，3.4.2 節において一般的な Ordered-TD の CE を一般化する．3.4.3 節にて探索の打ち切りおよびキャッシング処理を考慮した CE について説明する．

3.4.1 単一ノードからなる Ordered-TD の CE

本節では，$\chi(x) = V$ となるノード x のみから構成された Ordered-TD に対する CE について説明する．x 内の頂点の厳密なカーディナリティの計算を式 9 に示す．

$$cardinality(v) = \sum_{asg \in M_{[preorder(v)]}} \phi(v \mid asg) \qquad (9)$$

ただし，$preorder(v)$ は v より探索順序が早い頂点の集合，$M_{[preorder(v)]}$ は $preorder(v)$ の全ての割り当てからなる集合，$\phi(v \mid asg)$ は $preorder(v)$ に対応する割り当てが asg の場合，v に割り当て可能な G 上の頂点数を表す．また，$preorder(v) = \emptyset$ の場合，$cardinality(v) = \phi(v)$ とする．式 9 の計算を行うには，実際に P の各頂点の割り当てを探索する必要があり，計算コストは P と一致する部分グラフの列挙処理と同等である．そのため，式 9 を事前計算が可能なカーディナリティの推定式に変換する必要がある．そこで本研究は，カーディナリティの worst-case により実際のカーディナリティを推定する．v のカーディナリティの worst-case は，(i) $preorder(v)$ の割り当て asg に対する worst-case の割り当ての個数，(ii) worst-case における asg の組数の 2 つの積から得られるため，$cardinality(v)$ は式 10 によって推定される．

$$\begin{aligned} cardinality(v) &= \sum_{asg \in M_{[preorder(v)]}} \phi(v \mid asg) \\ &\approx \phi(v) \cdot \prod_{v_{pre} \in preorder(v)} \phi(v_{pre}) \end{aligned} \qquad (10)$$

ただし，$\phi(v)$ は asg に対する worst-case の割り当ての個数を表す．

$\phi(v)$ を得るため，提案手法は P の頂点 v, w および辺 e から，v に割り当て可能な頂点数を返す関数である $assignable : (V \times V \times E) \to \mathbb{R}$ を用いる．以降，プロパティグラフ G 上の頂点で $\lambda(v), \lambda(w)$ をラベルに含む頂点集合をそれぞれ V', W'，$\lambda(e)$ をラベルに含む辺集合を E' とする．関数内部の処理は v と w の探索順序の比較結果により 2 つに分かれる．v が w より早い場合，V' のうち，ある頂点 $w' \in W'$ と $\exists e' \in E'$ で隣接する頂点数 $|V_{early}|$ を返す．なお，提案手法では $|V_{early}|$ を事前に集計して保持する．一方 v が w より遅い場合，v の割り当ての探索時に w は G 上の頂点 w' が既に割り当てられている．そのため，$assignable$ では w' とある辺 $e' \in E'$ で隣接する $v' \in V'$ の個数 $|V_{late}|$ を返すことが望ましい．しかし，事前計算で行うには G 上の各頂点に対して統計情報を保持する必要があり，空間コストが高価である．そこで提案手法は，$|V_{late}|$ を $w' \in W'$ の各頂点に対して $\exists e' \in E'$ で隣接する $v' \in V'$ の個数の平均により代替し，保持する情報を削減する．また v に選択演算を含む場合，割り当て可能な頂点数は変化するため，保持する統計情報とその選択演算の選択率との乗算を $assignable$ 関数内部で行い，実際に割り当て可能な頂点数を近似する．

$\lambda(u)$	$\lambda(w)$	$\lambda(h)$	$\lvert V_{early}\rvert$	$\lvert V_{late}\rvert$
Artist	Track	involves	4	1.17
Track	Artist	involves	6	1.75
Playlist	Track	contains	2	1.00
Track	Playlist	contains	5	1.00

表 2　図 1 のプロパティグラフで保持する統計情報

関数 $assignable$ により，$\phi(v)$ は式 11 から計算される．

$$\phi(v) = \min_{w,e \in neighbors(v)} assignable(v, w, e) \qquad (11)$$

ただし，$neighbors(v)$ は v に隣接する頂点とこれらの頂点を結ぶ辺の組からなる集合を表す．

図 2 の P，図 1 の G，表 2 の統計情報を用いて，本節で説明した CE の例を示す．ノード内の頂点の探索順序を $O : \langle a, t, p \rangle$ としたとき，まず $cardinality(a)$ の推定を行う．$preorder(a) = \emptyset$ より，式 10 から $cardinality(a) \approx \phi(a)$ で推定できる．$neighbors(a) = \{(t, i)\}$，$\lambda(a) = \mathsf{Artist}, \lambda(t) = \mathsf{Track}, \lambda(i) = \mathsf{involves}$ より，統計情報から $assignable(a, t, i) = 4$ となる．よって，$\phi(a) = 4$ であり，$cardinality(a) \approx 4$ となる．続いて，$cardinality(t)$ を推定する．$preorder(t) = \{a\}$ より，$cardinality(t) = \phi(t) \cdot \phi(a) \approx 4 \cdot \phi(t)$ となる．$neighbors(t) = \{(a, i), (p, c)\}$ から，$\phi(t)$ を求めるために $assignable(t, a, i), assignable(t, p, c)$ が必要となる．このうち $assignable(t, a, i)$ は a の探索順序が t より早いため 1.75 となる．$assignable(t, p, c)$ についても同様に統計情報から取得すると，$\phi(t) = min(1.75, 1.00) = 1$ となる．よって，$cardinality(t) \approx 4$ となる．同様に $cardinality(p)$ を推定する．$preorder(p) = \{a, t\}$ より，$cardinality(p) \approx \phi(p) \cdot \phi(a) \cdot \phi(t) = 4 \cdot \phi(p)$ で推定できる．$neighbors(p) = \{(t, c)\}$ より $assignable(p, t, c) = 1$. よって $\phi(p) = 1$ であり，$cardinality(p) \approx 4$ となる．これにより，$cardinality(O) = 4 + 4 + 4 = 12$ となる．

3.4.2 Ordered-TD の CE の一般化

本節は前節と異なり，複数ノードからなる Ordered-TD を対象とする．あるノードの $adhesion$ に含まれる頂点 v はノード $x = owner(v)$ で頂点が割り当てられるため，$\phi(v)$ は式 12 のような場合分けが必要となる．

$$\phi(v) = \begin{cases} \min\limits_{u,e \in neighbors(v)} assignable(v, u, e) & (v \in owned(x)) \\ 1 & (otherwise) \end{cases} \qquad (12)$$

これにより，一般的な Ordered-TD のカーディナリティは式 13 により推定される．

$$cardinality(t) = \sum_{x \in X} \left(\phi(O_x) \cdot \prod_{x_{pre} \in preorder(x)} \phi(O_{x_{pre}}) \right) \tag{13}$$

$$\begin{aligned} \phi(O_x) &= \sum_{v \in owned(x)} cardinality(v) \\ &\approx \sum_{v \in owned(x)} \left(\phi(v) \cdot \prod_{v_{pre} \in preorder(v) \cap \chi(x)} \phi(v_{pre}) \right) \end{aligned} \tag{14}$$

3.4.3 探索の打ち切りおよびキャッシングを考慮した CE

まず，3.2 節で提案した探索の打ち切りを考慮した CE について説明する．DISTINCT 句で指定された頂点集合 V_D と，ある探索順序が与えられたときのコア集合を C として，$v_O \in V \backslash C$ となる頂点 v_O があるとする．3.2 節の打ち切り処理により，v_O の割り当ての個数は以下のように定義される $parents(v_O)$ の割り当てに対して 1 つとなる．

[**定義** 10]（parents） グラフパターン $P(V, E, \rho, \lambda, \pi)$，$P$ の頂点の探索順序 $\langle v_1, \ldots, v_n \rangle$ が与えられたとき，$v \in V$ の $parents$ を表す $parents(v)$ は V の部分集合であり，$v_p \in parents(v)$ は v より探索順序が早く，ある辺 $e \in E$ について $\rho(e) = (v, v_p) \vee \rho(e) = (v_p, v)$ を満たす．

ゆえに，探索の打ち切りを考慮した $\phi(v)$ は式 15 で推定できる．

$$\phi(v) = \begin{cases} \min\limits_{u, e \in neighbors(v)} assignable(v, u, e) & (v \in C \cap owned(x)) \\ 1 & (otherwise) \end{cases} \tag{15}$$

次に，キャッシュヒットを考慮した CE について説明する．3.4.2 節の式 13 による計算は，$preorder(x)$ の全ての割り当てについて $owned(x)$ の頂点の割り当ての探索が 1 回行われることを想定している．しかし，$adhesion(x)$ に対応する割り当てが同一な $preorder(x)$ の割り当てについて，$owned(x)$ の頂点の割り当ての探索回数はキャッシング処理により 1 回となる．よって，キャッシュヒットを考慮した場合，x の探索は $preorder(x)$ のキャッシュサイズと等しくなるため，式 13 は式 16 のように置き換えることができる．

$$cardinality(t) = \sum_{x \in X} (\phi(O_x) \cdot |cache_x|) \tag{16}$$

ただし，x のキャッシュのキーの総数を $|cache_x|$ と表す．なお，実際の $|cache_x|$ を求めるには，割り当ての列挙処理が必要であるため，$|cache_x|$ は式 17 を用いて近似する．

$$|cache_x| \approx |M_{[adhesion(x)]}| \tag{17}$$

3.5 アルゴリズム

Algorithm 3, 4 に提案手法のアルゴリズムを示す．アルゴリズムは WCOJ に基づき，探索の打ち切り，キャッシング処理，実行計画の最適化を取り入れている．

Algorithm 3 提案手法

Input: $G(V_G, E_G, \rho_G, \lambda_G, \pi_G)$, $P(V, E, \rho, \lambda, \pi)$, V_D;
Output: Q, 問合せ結果;
1: $Q \leftarrow \emptyset$; $cache \leftarrow \emptyset$;
2: $\mathcal{N} \leftarrow$ EnumerateOrderedTDs(P);
3: 3.4 節の実行計画の最適化により Ordered-TD $\langle t, \chi \rangle$, 探索順序 $\langle v_1, \ldots, v_n \rangle$ を決定する;
4: C を $P, V_D, \langle v_1, \ldots, v_n \rangle$ でのコア集合とする.
5: **for** $d = 1, \ldots, n$ **do**
6: $\quad \mu(v_d) = \perp$;
7: $\quad$ **if** $v_d \in V \backslash C$ **then**
8: $\quad\quad terminated(v_d) = false$;
9: **for all** nodes x of t **do**
10: $\quad$ **if** x に通常キャッシュが適用されている **then**
11: $\quad\quad intrmd(x) \leftarrow \emptyset$;
12: $\quad$ **else**
13: $\quad\quad intrmd(x) = false$;
14: RecursiveProposedJoin(1);
15: **return** Q;

まず，探索の打ち切り処理について説明する．3.2 節より，打ち切り処理のタイミングは (i) v_{last} の割り当ての更新のとき，(ii) 探索している頂点 v_d がコア集合に含まれず，V_D の割り当て $\mu_{[V_D]}$ に対して部分グラフが存在するときである．このうち (i) は RecursiveProposedJoin の 31 行目より，$\mu_{[V_D]}$ が問合せ結果に含まれるか判定し，含まれる場合に v_{last} の割り当てを再更新する．アルゴリズムでは (ii) の打ち切り処理を行うため，コア集合に含まれない頂点 v_O に対して探索の打ち切りが可能かを示すグローバル変数 $terminated$ を導入する．RecursiveProposedJoin の 9 行目より，P と一致する部分グラフが存在した場合に v_O の探索は打ち切り可能であるため，全ての v_O に対して $terminated$ は $true$ となる．34 行目より，v_d の割り当てについて P と一致する部分グラフの存在が確認できた場合，v_d の割り当ての更新を回避して v_{d-1} の探索へ戻る．

次に，キャッシング処理について説明する．3.3 節より，x が保持するキャッシュは条件に応じて通常キャッシュと判定キャッシュのどちらかが適用される．まず，$x = owner(v_d)$ となる t のノード x に通常キャッシュが適用される場合の処理について説明する．RecursiveProposedJoin の 5 行目より，変数 $intrmd$ には x のキャッシュの保持対象である $territory(x) \cap V_D$ に対応する割り当てが追加される．キャッシュヒットが起きた場合の処理は 20 行目より，キャッシュが保持していた割り当てから μ を更新し，$territory(x)$ 以外の頂点の探索に移る．一方で判定キャッシュが適用されている場合，RecursiveProposedJoin の 7 行目より，$intrmd$ には P と一致する部分グラフが存在することを示す $true$ を代入する．キャッシュヒットが起きた場合の処理は 25 行目より，$intrmd(x)$ が $true$ である，すなわち現在の $adhesion(x)$ の割り当てについて部分グラフが存在する場合のみ，$territory(x)$ 以外の頂点の探索に移る．実行計画の最適化については Algorithm 3 の 3 行目で行われており，P の頂点の探索順序が得られる．

本章で説明したアプローチにより，提案手法は DISTINCT 句を含む問合せ処理における冗長な処理を回避することが可能である．さらに，グラフパターンの形や各頂点の選択率に対しても柔軟に対処し，安定した処理性能を実現できる．

Algorithm 4 Recursive Proposed Join

Input: d, 探索する頂点のインデックス;
1: **if** $d = n + 1$ **then**
2: $Q \leftarrow Q \cup \mu_{[V_D]}$;
3: **for all** nodes x of t **do**
4: **if** x に通常キャッシュが適用されている **then**
5: $intrmd(x) \leftarrow intrmd(x) \cup \mu_{[territory(x) \cap V_D]}$;
6: **else**
7: $intrmd(x) = true$;
8: **for all** $v_O \in V \backslash C$ **do**
9: $terminated(v_O) = true$;
10: **return**
11: $x = owner(v_d)$;
12: $\alpha = adhesion(x)$;
13: **if** $x \neq owner(v_{d-1})$ **then**
14: **if** x に通常キャッシュが適用されている **then**
15: $intrmd(x) \leftarrow \emptyset$;
16: **else**
17: $intrmd(x) = false$;
18: **if** $cache[\alpha, \mu_{[\alpha]}]$ が存在 **then**
19: $owner(v_l) \in territory(x)$ となる最大値を l とする.
20: **if** x に通常キャッシュが適用されている **then**
21: **for all** $\beta \in cache[\alpha, \mu_{[\alpha]}]$ **do**
22: $\mu_{[territory(x) \cap V_D]} = \beta$;
23: RecursiveProposedJoin$(l + 1)$;
24: $\mu_{[territory(x) \cap V_D]} \leftarrow \perp$;
25: **else**
26: **if** $cache[\alpha, \mu_{[\alpha]}]$ is $true$ **then**
27: RecursiveProposedJoin$(l + 1)$;
28: **return**
29: **for all** v_d に割り当て可能な G 上の全ての頂点 a に対して **do**
30: $\mu(v_d) = a$;
31: **if** $v_d = v_{last}$ and $\mu_{[V_D]} \in Q$ **then**
32: **continue**
33: RecursiveProposedJoin$(d + 1)$;
34: **if** $v_d \in V \backslash C$ and $terminated(v_d)$ is $true$ **then**
35: $terminated(v_d) = false$;
36: **return**
37: $\mu(v_d) = \perp$;
38: **if** $d > 1$ and $x \neq$ owner(v_{d-1}) **then**
39: $cache[\alpha, \mu_{[\alpha]}] \leftarrow intrmd(x)$;

データセット	$\lvert V \rvert$	$\lvert E \rvert$	ラベルの種類
p2p-Gnutella04	10,876	39,994	1
CA-GrQc	5,242	14,496	1
Wiki-Vote	7,115	100,762	1
LDBC (scale factor = 1)	2,997,352	17,196,776	8

表 3　データセットの詳細

4　評価実験

提案手法の処理時間の評価を行うため，実グラフと合成グラフを用いて近年提案された手法との比較を行った．実験では，異なるグラフパターンの形と選択演算を含む問合せを用いることにより，提案手法の有効性を検証する．

4.1　実験設定

我々は，提案手法を以下に示す従来手法と比較する．

- **LFTJ** [11]：WCOJ アルゴリズムの主要な手法の一つであり，trie index を用いることで割り当て可能な頂点の列挙処理を効率的に行う．
- **CTJ** [12]：WCOJ アルゴリズムの高速化手法であり，問合せ結果の再利用により割り当ての再探索を回避する．

提案手法を含む全ての手法において，割り当て可能な頂点の列挙処理に trie index を用いる．trie index は構築にかかる計算コストが高価であるため，事前構築を行う．なお実験では，グラフパターンの頂点のプロパティに対する選択演算を問合せに含む．選択演算を効率的に処理するため，全ての手法について二次索引を構築しており，グラフパターンと一致する部分グラフの列挙処理は trie index へ選択条件を適用した後に行われる．また，LFTJ および CTJ について該当する論文中に実行計画についての記載がない，あるいは断片的であったため，どちらについても 3.4 節で説明した最適化を行った．具体的に LFTJ では 3.4.1 節の式 10，CTJ では 3.4.2 の式 13 を 3.4.3 節の式 16 によって拡張したものを用いて探索順序を決定する．

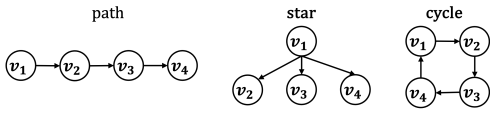


図 6　実験で用いるグラフパターンの形

表 3 にて実験で用いるデータセットを示す．p2p-Gnutella04, CA-GrQc，Wiki-Vote は SNAP [15] によって公開されている実グラフデータセットである．LDBC は LDBC Benchmark [16] によって生成されたプロパティグラフである．なお，生成されるグラフの規模を示す scale factor は 1 とした．

提案手法が広範なグラフパターンにおいて既存手法より有利であることを示すため，実験では 4 つのグラフパターンの形と頂点に対する 2 種類の選択演算を含む問合せを用いる．グラフパターンの形は図 6 に示したものと Erdös-Reyni Generator を用いて生成したランダムグラフとし，CTJ [12] の性能評価に用いられた形を含む．Erdös-Reyni Generator では引数として頂点数 $|V|$ と任意の頂点間に辺が存在する確率 P をとるが，実験ではそれぞれ $|V| = 4, P = 0.3$ として計 5 個のグラフパターンを生成する．なお，5 個のグラフパターンでの実行時間の平均値をランダムグラフの場合の実行時間とする．いずれの形についてもグラフパターンの頂点は $\{v_1, v_2, v_3, v_4\}$ で構成されており，DISTINCT 句では常に $\{v_1\}$ を指定する．また，選択率の異なる 2 種類の選択演算を常に v_3 に対して適用する．これにより，3.4 節の実行計画の最適化の有効性を検証する．

全ての実験は Apple M1 CPU 3.20 GHz，16 GiB RAM 上で実行し，全ての手法は C++を用いてシングルスレッドプログラムとして実装し，clang 15.0.0 にてオプション-O3 を用いてコンパイルした．実行時間は，探索順序の最適化や選択演算の処理時間についても含む．

4.2　実行速度の評価

4.2.1　グラフパターンの形による影響

図 7 にて各手法のグラフパターンの形に対する問合せ処理の実行時間を比較する．この結果より，提案手法はデータセットが p2p-Gnutella04，LDBC のグラフパターンの形が cycle の場合を除いて，全ての条件下で他の手法よりも高速であることが確認できる．平均すると，提案手法は CTJ の約 69 倍，LFTJ の約 91 倍高速である．また，提案手法は CTJ より最大約 3.4×10^4 倍，LFTJ より最大約 5.1×10^4 倍高速である．提案手法が問合せを高速に処理できる要因として，探索の打ち切りによって問合せ結果に不要な割り当ての更新を回避できること，キャッシング処理により同一の割り当ての再計算を回避できることが挙げられる．

一方で，図 7 より p2p-Gnutella04，LDBC の cycle の場合，

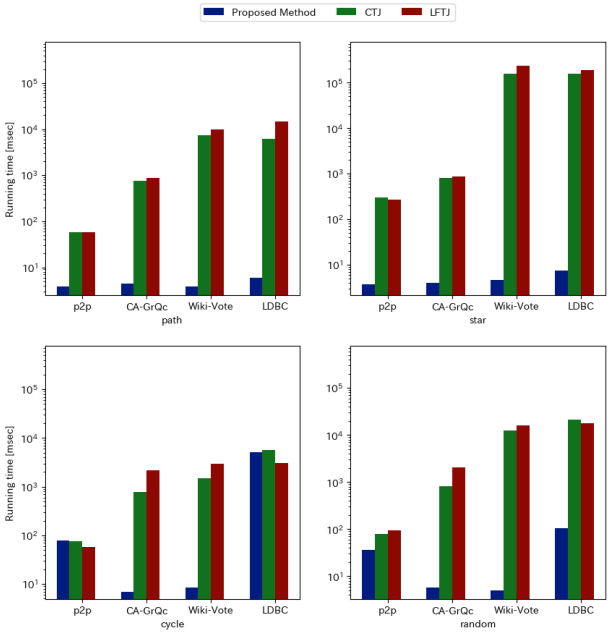


図 7　それぞれのグラフパターンの形における実行時間の比較

提案手法は LFTJ より処理時間が遅くなっていることが確認できる．この原因として，問合せの選択率の低さが挙げられる．打ち切り処理は問合せ結果を利用しており，問合せ結果が少ない場合は一般的に打ち切りが起きづらくなる．実際，LDBC の場合の問合せの選択率は 0%，p2p-Gnutella04 の場合の選択率は 2.6% である．その結果として，提案手法は探索回数を十分に削減できないとともに，打ち切り判定にかかるオーバーヘッドにより LFTJ と比較して低速な処理となる．

4.2.2　選択演算による影響

図 8 にて LDBC データセットに対する 2 つの異なる選択率の選択演算を含む問合せについての処理時間を比較する．ただし，図中の比較手法である Proposed Method (Standard Optimization) は探索の打ち切りおよびキャッシング処理を導入し， 3.4.1 節の式 10 によって探索順序を決定する手法である（以降，Proposed-SO と呼ぶ）．選択率が低い選択演算を含む問合せ処理では，path と cycle を除いて提案手法及び Proposed-SO が他の手法と比較してはるかに高速である．一方，選択率が高い選択演算を含む問合せ処理では，提案手法は Proposed-SO に対しても実行時間が最大約 200 倍高速であることが確認できる．表 4 にそれぞれのグラフパターン，選択演算における提案手法と Proposed-SO の探索順序と処理時間を示す．選択率が低い場合は提案手法の path の場合を除き，提案手法と Proposed-SO は選択演算を含む頂点を最初に探索する順序を選択している．これは選択率が低い頂点を最初に探索し，それに隣接する頂点から順に探索することで割り当てられる頂点数を全体的に削減できるからである．一方選択率が高い場合，Proposed-SO は選択率が低い場合と同様に選択演算を含む頂点を最初に探索するのに対し，提案手法は全て DISTINCT 句で指定された頂点を最初に探索する順序を選択している．DISTINCT 句で指定された頂点を最初に探索する

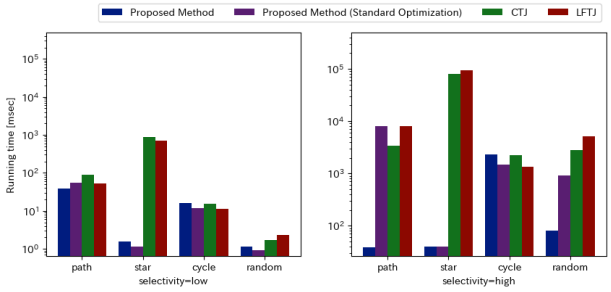


図 8　問合せに選択演算を含む場合の実行時間の比較

選択率	グラフパターンの形	手法	探索順序	実行時間 [ミリ秒]
低	path	提案手法	$v_1 \to v_2 \to v_3 \to v_4$	38.506
		Proposed-SO	$v_4 \to v_3 \to v_2 \to v_1$	55.312
	star	提案手法	$v_3 \to v_1 \to v_2 \to v_4$	1.600
		Proposed-SO	$v_3 \to v_1 \to v_2 \to v_4$	1.133
	cycle	提案手法	$v_3 \to v_2 \to v_1 \to v_4$	16.298
		Proposed-SO	$v_3 \to v_2 \to v_1 \to v_4$	11.957
高	path	提案手法	$v_1 \to v_2 \to v_3 \to v_4$	39.306
		Proposed-SO	$v_4 \to v_3 \to v_2 \to v_1$	7946.189
	star	提案手法	$v_1 \to v_2 \to v_3 \to v_4$	39.586
		Proposed-SO	$v_3 \to v_1 \to v_2 \to v_4$	40.118
	cycle	提案手法	$v_1 \to v_2 \to v_3 \to v_4$	2339.867
		Proposed-SO	$v_3 \to v_2 \to v_1 \to v_4$	1465.002

表 4　それぞれの選択率における各手法の探索順序と実行時間

ことで，コア集合に含まれない頂点が多くなり，打ち切り処理によってより多くの探索を打ち切ることが可能である．ゆえに表 4 の実行時間の通り，選択率が高い問合せにおいて提案手法は Proposed-SO より安定した処理性能を維持している．なお，図 8 の選択率が高い選択演算を含み，グラフパターンの形が cycle の場合について，提案手法が Proposed-SO より低速な処理となっていることが確認できる．この原因として，探索の打ち切りを考慮したカーディナリティ推定の精度が不安定であることが挙げられる．3.4.3 節より，コア集合に含まれない頂点 v_O の割り当ての個数は $parents(v_O)$ の割り当ての個数と等しいとしていた．しかし，実際の v_O の割り当ての個数は $descendants(v_O)$ の選択率に依存しており，選択率が 1 である場合を除いて割り当ての個数は 1 以上となる．そのため，実行計画の最適化の今後の展望としては，問合せの選択率を考慮したカーディナリティ推定が考えられる．

5　関 連 研 究

グラフパターンを含む問合せ処理の最適化に関して，様々な研究が行われてきた [13], [17], [18]．その中でも代表的な手法について述べる．

- 木分解を用いた実行計画の最適化手法

文献 [17] は，RDF における問合せ処理に対して WCOJ を導入した手法である．この手法ではグラフパターンの木分解を行い，作成された木の葉ノード内で結合結果を生成して親ノードの結合操作へ移り，以降同様の処理を根ノードまで繰り返すことで問合せ結果を生成する．各ノード内では WCOJ を用いて中間結果を生成しており，trie index とグラフ上の頂点の辞書エンコーディングによって，割り当て可能な頂点を効率的に探索可能である．また木分解では複数の木を列挙し，木の分数カ

バーと選択演算のある頂点を含むノードの探索順序を基準として最適な木を選択している．これにより，計算コストの小さい探索順序を選択でき，実データによって有効性を示している．

- リングインデックスを用いる最適化手法

文献 [18] は，インメモリなどの空間的条件を伴う場合を考慮した問合せ処理手法である．従来，WCOJ での効率的な探索のため，trie や B(+) 木などがインデックスとして用いられてきた．しかし，これらを事前に構築するためには多くのインデックスを保持する必要があり，膨大な空間コストを要していた．この手法では，WCOJ の効率的な探索を 1 つのインデックスで対応できる ring index を提案した．ring index のサイズはグラフデータに対して準線形オーダーの空間量であり，従来のインデックスのはるかに小さいサイズで効率的な探索処理を維持できることが実データによって示された．

- ハッシュに基づく trie index を用いる最適化手法

文献 [13] は，グラフデータが動的であることを考慮した問合せ処理手法である．LFTJ で提案された trie index は内部でソート済みの配列を用いることから，構築コストが高価であることが問題として挙げられており，データが動的な場合に対応できない．この手法では，trie index で保持する頂点をハッシュ化して格納する．ハッシュ構造に基づく trie index は，探索性能を維持したまま，データに対して線形時間での構築を可能である．また，バイナリ結合と WCOJ を組み合わせた実行計画の最適化を提案している．バイナリ結合は WCOJ と比較して最適化技術が発展しており，処理性能が一般的に良い．そのため，グラフパターン内でバイナリ結合で爆発的な中間結果を生成する場合についてのみ WCOJ に置き換える．これにより，前処理コストの削減，かつ広範な問合せに対する安定した処理性能を実現でき，実データによって有効性を示している．

本研究は DISTINCT 句を含む問合せを対象としているため，以上に挙げた研究とは位置付けが異なる．しかし提案したアプローチは，WCOJ アルゴリズムであれば既存研究に適用することは容易であり，その結果として多くの恩恵を与えることができると考えられる．

6 おわりに

本稿では，グラフデータベースにおけるグラフパターンを含む問合せにおいて，DISTINCT を含む場合の問合せ処理の高速化手法を提案した．提案手法は既存手法の課題であった問合せ結果に不要な部分グラフの探索を打ち切りによって回避し，キャッシング処理や実行計画の最適化によって処理性能をさらに向上させた．我々が行った実験では，提案手法は近年提案された手法より平均 69 倍高速であり，提案したアプローチの有効性を示した．また，提案手法はプロパティグラフに対する問合せを想定しており，現実に起こりうるユースケースに幅広く対処できる．これにより，大規模なグラフデータに対するオンライン分析処理の性能向上に本手法は貢献できる．

文　献

[1] Pacaci, A., Zhou, A., Lin, J. and Özsu, M. T.: Do we need specialized graph databases? Benchmarking real-time social networking applications, *Proceedings of the Fifth International Workshop on Graph Data-management Experiences & Systems*, pp. 1–7 (2017).

[2] Messina, A., Fiannaca, A., La Paglia, L., La Rosa, M. and Urso, A.: BioGraph: a web application and a graph database for querying and analyzing bioinformatics resources, *BMC systems biology*, Vol. 12, No. 5, pp. 75–89 (2018).

[3] Zhang, Z. J.: Graph databases for knowledge management, *IT professional*, Vol. 19, No. 6, pp. 26–32 (2017).

[4] Sahu, S., Mhedhbi, A., Salihoglu, S., Lin, J. and Özsu, M. T.: The ubiquity of large graphs and surprising challenges of graph processing: extended survey, *The VLDB journal*, Vol. 29, pp. 595–618 (2020).

[5] Cheng, J., Ke, Y., Ng, W. and Lu, A.: Fg-index: towards verification-free query processing on graph databases, *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 857–872 (2007).

[6] Choudhury, S., Holder, L., Chin, G., Mackey, P., Agarwal, K. and Feo, J.: Query optimization for dynamic graphs, *arXiv preprint arXiv:1407.3745* (2014).

[7] Sun, Y., Li, G., Du, J., Ning, B. and Chen, H.: A subgraph matching algorithm based on subgraph index for knowledge graph, *Frontiers of Computer Science*, Vol. 16, pp. 1–18 (2022).

[8] Kashyap, N. K., Pandey, B. and Mandoria, H.: Analysis of pattern identification using graph database for fraud detection, *Oriental Journal of Computer Science and Technology*, Vol. 9, No. 2, pp. 81–91 (2016).

[9] Yannakakis, M.: Algorithms for acyclic database schemes, *VLDB*, Vol. 81, pp. 82–94 (1981).

[10] Ngo, H. Q., Porat, E., Ré, C. and Rudra, A.: Worst-case optimal join algorithms, *Journal of the ACM (JACM)*, Vol. 65, No. 3, pp. 1–40 (2018).

[11] Veldhuizen, T. L.: Leapfrog Triejoin: A simple, worst-case optimal join algorithm, *Proc. International Conference on Database Theory* (2014).

[12] Kalinsky, O., Etsion, Y. and Kimelfeld, B.: Flexible caching in trie joins, *arXiv preprint arXiv:1602.08721* (2016).

[13] Freitag, M., Bandle, M., Schmidt, T., Kemper, A. and Neumann, T.: Adopting worst-case optimal joins in relational database systems, *Proceedings of the VLDB Endowment*, Vol. 13, No. 12, pp. 1891–1904 (2020).

[14] Neumann, T. and Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins, *2011 IEEE 27th International Conference on Data Engineering*, IEEE, pp. 984–994 (2011).

[15] Leskovec, J. and Krevl, A.: SNAP Datasets: Stanford large network dataset collection (2014).

[16] Szárnyas, G., Prat-Pérez, A., Averbuch, A., Marton, J., Paradies, M., Kaufmann, M., Erling, O., Boncz, P., Haprian, V. and Antal, J. B.: An early look at the LDBC social network benchmark's business intelligence workload, *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, pp. 1–11 (2018).

[17] Aberger, C. R., Tu, S., Olukotun, K. and Ré, C.: Old techniques for new join algorithms: A case study in RDF processing, *2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)*, IEEE, pp. 97–102 (2016).

[18] Arroyuelo, D., Hogan, A., Navarro, G., Reutter, J. L., Rojas-Ledesma, J. and Soto, A.: Worst-case optimal graph joins in almost no space, *Proceedings of the 2021 International Conference on Management of Data*, pp. 102–114 (2021).

関係データベースにおける動的複数問合せ最適化の試作実装と初期評価

木村　元紀†　　合田　和生†

† 東京大学　〒 153–8505 東京都目黒区駒場 4 丁目 6 – 1
E-mail: †{kimura-g,kgoda}@tkl.iis.u-tokyo.ac.jp

あらまし　関係データベースにおいて、入力された問合せから実行計画を生成する問合せ最適化はその性能を大きく左右する。従来の問合せ最適化はデータの統計情報や近似的なコストモデルから実行コストを推定して最も効率的な実行計画を構築するが、この推定の誤りのために非効率的な実行に陥りうることが問題となっている。実行時情報を用いて実行計画を修正する動的問合せ最適化はこの問題の解決策として期待されている。本稿では複数の問合せが並行に入力される状況において複数の問合せ処理にまたがって動的問合せ最適化を行う動的複数問合せ最適化を提案し、その試作実装を行った。また、その有効性を検証した。

キーワード　問合せ最適化, 問合せ処理, データベース技術

1　はじめに

生成・蓄積されるデータが増加し続けている現代社会において、その膨大なデータを活用することにより新たな価値を創出することが多くの分野で期待されている。データベースシステムはデータの蓄積・管理・活用を支援する基盤技術であり、その性能を向上させることはデータ活用による迅速な価値創出に必要不可欠である。データベースシステムは多様な技術の融合によって構成されており、その中でも問合せ最適化はその根幹を担う要素である。問合せ最適化は入力された問合せから、その問合せを実現する実行計画を生成する処理である。最もよく用いられるデータベースの一つである関係データベースにおいて問合せの記述に用いられる SQL 等の宣言的な言語は蓄積されているデータからどのように結果を抽出するかを直接的には記述しておらず、問合せ最適化によって導出された実行計画に基づいて実際の問合せ処理が行われる。そのため、問合せ最適化の性能はデータベースシステムの性能に大きく影響する。

従来の問合せ最適化は、簡単なものでは入力された問合せを一定の規則に基づいて変形し、実行計画を生成するルールベース問合せ最適化から、複雑なものではデータの統計情報や近似的なコストモデルから実行コストを推定して最も効率的な実行計画を構築するコストベースの問合せ最適化まで様々なアプローチが用いられてきた。しかし、既存の問合せ最適化はいずれも計算時間や正確性の問題がある。ルールベース手法では計算時間は短いが、データベースに蓄積されているデータの性質を反映することはできず、非効率な実行計画を生成しうる。コストベース手法では可能な実行計画を全て列挙し、データの統計情報やデータベースのコストモデルを用いることで各実行計画の実行コストを推定し、最も実行コストが少ない実行計画を選択する。この手法は比較的長い計算時間と引き換えにより効率的な実行計画に到達するが、統計情報やコストモデルの不正確性により真に最適な実行計画を生成できるとは限らない。そのため、依然としてより正確な問合せ最適化を実現するための研究が盛んに行われている。

問合せ最適化の不正確性に対する一つのアプローチとして、実行時情報を用いて実行計画を修正する動的問合せ最適化が期待されている。動的問合せ最適化は実行時に問合せ処理の実行状態をフィードバックしながら問合せ処理を実行中に修正し、不正確な問合せ最適化による実行効率の低下を軽減する。動的問合せ最適化は初期のデータベースである INGRES にその原型が見られ [1]、以降も様々な研究が行われている [2,3]。

問合せ処理の性能向上を目指した問合せ最適化におけるもう一つの試みとして、複数の問合せが並行に入力される状況において複数の問合せ処理にまたがって問合せ最適化を行う複数問合せ最適化がある。通常の問合せ最適化は入力された問合せ一つのに対して一つの実行計画を生成し、それらを独立に実行する。一方複数問合せ最適化は複数の問合せの実行結果を同時に得る実行計画を生成することで、一度の実行で複数の問合せの実行結果を得る。複数問合せの試みは 1990 年ごろから始まり、初期の共通部分式除去に基づいた手法 [4,5] から、厳密な組み合わせ最適化に基づいた手法 [6]、さらには強化学習を用いた手法 [7] などが提案されている。

本稿では複数の問合せが並行に入力される状況において複数の問合せ処理にまたがって動的問合せ最適化を行う動的複数問合せ最適化を提案し、その試作実装及び初期評価を行った。厳密な複数問合せ最適化は問合せ数に対して指数的に計算時間が増加するが、動的複数問合せ最適化は問合せ処理と並行して最適化計算を行うため、問合せ最適化による実行開始の遅延を抑制することができる。また、複数問合せ最適化はバッチ処理であり、問合せ最適化の対象となる問合せは全て入力されている必要がある。一方、動的複数問合せ最適化では逐次的に入力された問合せを考慮することが可能となる。

本稿の構成は以下の通りである。2 章では従来の問合せ最適化研究等、本稿の背景について述べる。3 章では提案する動的複数問合せ最適化の概要を述べる。4 章では動的複数問合せ最適化の試作実装について述べる。5 章では試作実装の初期評価を行う。6 章では関連研究について述べる。7 章では結論を述

Algorithm 1 DPsize

```
1: function DPsize(R = {R_0 ... R_{n-1}})
2:   ∀R_i ∈ R dpTable[{R_i}] = R_i
3:   for s = 2 ... n   // size of plan
4:     for s_1 = 1 ... s − 1   // size of left subplan
5:       for ∀S_1 ⊂ R : |S_1| = s_1, ∀S_2 ⊂ R : |S_2| = s − s_1
6:         if S_1 ∩ S_2 = ∅ ∨ S_1is not connected to S_2 then
7:           continue
8:         p = dpTable[S_1] ⋈ dpTable[S_2]
9:         if cost(p) < cost(dpTable[S_1 ∪ S_2]) then
10:          dpTable[S_1 ∪ S_2]) = p
11:  return dpTable[R]
```

べる。

2 動的計画法に基づく問合せ最適化アルゴリズム

厳密なコストモデルに基づく問合せ最適化は NP 困難であるため [8]、ヒューリスティクスに基づいて現実的な時間内で近似的な解を得られるアルゴリズムが広く用いられている。ヒューリスティクスとしてはルールベース最適化や遺伝的アルゴリズム、動的計画法などが用いられている。アルゴリズム 1 に動的計画法に基づく最も単純な問合せ最適化アルゴリズムの一つである DPsize を示した。DPsize は DB2 などの商用データベースシステムで採用されている。DPsize は問合せ中に現れる結合条件を考慮せず、直積によって実現される結合を含むすべての実行計画を数え上げるため、その複雑性は理論的な下界よりはるかに悪い [9]。

DPsize の問題点を改善したアルゴリズムとして DPccp や DPhyp などが提案されている。これらのアルゴリズムは入力された問合せをグラフやハイパーグラフとして表現し、そのグラフ上で動的計画法を行うことにより、効率的に実行計画を数え上げる。本章の残りの部分では DPhyp 及びその導入に必要な事項を簡潔に述べる。

定義 (ハイパーグラフ)**.** ハイパーグラフ H は頂点集合 V と辺集合 E の組 (V, E) であり、以下を満たす。

- V は空でない頂点集合である。
- E はハイパーエッジの集合であり、各ハイパーエッジは V の非空部分集合 $u \subset V, v \subset V$ の非順序対 (u, v) である。

問合せ最適化においては問合せ中に出現するテーブルを頂点、結合演算の述語を辺に対応させる。$R.a = S.b$ のように二つのテーブルのみが出現する結合述語は 2 頂点のみを結ぶのでグラフの辺として表現可能だが、より一般には $R.a = S.b + T.c$ のように 3 つ以上のテーブルが出現する場合があり、ハイパーグラフによって問合せを表現する必要がある。また本稿では触れないがハイパーグラフで表現することにより内部結合だけでなく外部結合や反結合のような非可換な結合演算を扱うことも可能になる。

定義 (部分グラフ)**.** ハイパーグラフ $H = (V, E)$、頂点集合 $V' \subseteq V$ とするとき、この頂点集合が導く部分グラフ $H_{|V'}$ は $E' = \{(u, v) | (u, v) \subseteq E, u \subseteq V', v \subseteq V'\}$ を用いて $H_{|V'} = (V', E')$ と表される。

Algorithm 2 DPhyp

```
1: function DPhyp(R = {R_0 ... R_{n-1}})
2:   ∀R_i ∈ R dpTable[{R_i}] = R_i
3:   for ∀csg-cmp-pairs(S_1, S_2)
4:     e = {(u, v) | u ⊆ S_1, v ⊆ S_2}
5:     p = dpTable[S_1] ⋈_e dpTable[S_2]
6:     if cost(p) < cost(dpTable[S_1 ∪ S_2]) then
7:       dpTable[S_1 ∪ S_2]) = p
8:   return dpTable[R]
```

定義 (連結)**.** ハイパーグラフ $H = (V, E)$ が連結であるとは、$|V| = 1$ または V の分割 V', V'' が存在し、$u \subseteq V, v \subseteq V$ かつ $H_{|V'}$ 及び $H_{|V''}$ が連結なハイパーグラフ $(u, v) \in E$ が存在することである。さらに、$V' \subseteq V$ が導く部分グラフ $H_{|V'}$ が連結であるとき、V' は H の連結部分グラフ *(csg)* であるといい、連結部分グラフ $H_{|V''}$ を導く $V'' \subseteq (V \setminus V')$ が存在するとき、V'' を V' の連結補部分グラフ *(cmp)* という。

定義 (csg-cmp-pair)**.** ハイパーグラフ $H = (V, E)$ とし $S_1 \subseteq V, S_2 \subseteq (V \setminus S_1)$ が連結部分グラフ及びその連結補部分グラフであり、さらにある辺 $(u, v) \in E$ が存在し、$u \subseteq S_1, v \subseteq S_2$ であるとき、(S_1, S_2) を *csg-cmp-pair* という。

動的計画法に基づく問合せ最適化アルゴリズムにおけるコスト関数呼び出しの最小回数は与えられたハイパーグラフ中の csg-cmp-pair の個数に等しい。DPhyp は全ての csg-cmp-pair を動的計画法として許容される順序、つまり (S_1, S_2) の前に $S'_1 \subseteq S_1, S'_2 \subseteq S_2$ なる csg-cmp-pair(S'_1, S'_2) を探索する順序で効率的に列挙する。DPhyp により csg-cmp-pair を適当な順序で列挙することができるならば、アルゴリズム 2 のようにして実行計画を構築することが可能である。

3 動的複数問合せ最適化

動的複数問合せ最適化アルゴリズムを導入する前に、本章では動的複数問合せ最適化が可能にする実行効率化のケーススタディを紹介する。

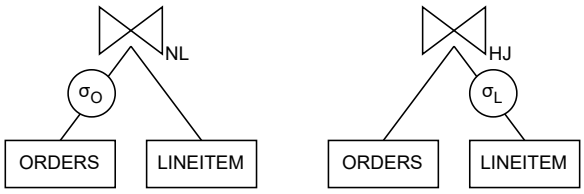
3.1 Case Study: 2-way Join

次の問合せ Qa, Qb を考える。

```
Qa: SELECT * FROM LINEITEM, ORDERS WHERE
    L_ORDERKEY = O_ORDERKEY AND L_QUANTITY < 25
Qb: SELECT * FROM LINEITEM, ORDERS WHERE
    L_ORDERKEY = O_ORDERKEY AND O_ORDERDATE
    BETWEEN '1993-01-01' AND '1994-01-01'
```

TPC-H(SF=10) において PostgreSQL でこれらの問合せを実行するとき最も実行時間が短くなる実行計画は次のようになり、図示すると図 1 のようになる。

```
// Execution Plan of Qa
```



(a) Qa (b) Qb

図 1: 2-way join 問合せの実行計画: 但し、$\bowtie_{NL}$ は左辺を outer、右辺を inner とするインデックス結合、$\bowtie_{HJ}$ は左辺を build、右辺を probe とするハッシュ結合、σ_O は `O_ORDERDATE BETWEEN '1993-01-01' AND '1994-01-01'` を満たす要素を返す選択演算、σ_L は `L_QUANTITY > 25` を満たす要素を返す選択演算。

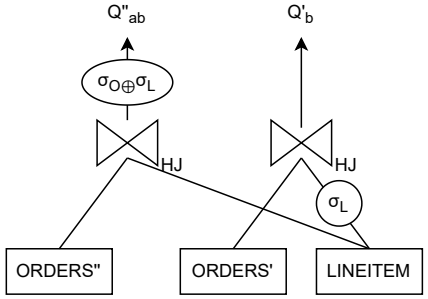


図 2: Qa の実行中に Qb が入力された時の実行計画変換: 但し、ORDERS のうち Qa において既に読み取り終えている部分が ORDERS'、残りの部分が ORDERS" である。また、$\sigma_O \oplus \sigma_L$ は σ_O もしくは σ_L の少なくとも一方を満たすタプルを通過させる選択演算子である。

```
Nested Loop
  ->  Seq Scan on ORDERS
    Filter: (O_ORDERDATE BETWEEN '1993-01-01'
        AND '1994-01-01')
  ->  Index Scan on LINEITEM
      Index Cond: (L_ORDERKEY = O_ORDERKEY)
```

```
// Execution Plan of Qb
Hash Join Hash Cond: (L_ORDERKEY = O_ORDERKEY)
  ->  Seq Scan on LINEITEM
    Filter: (L_QUANTITY > 25)
  ->  Hash
    ->  Seq Scan on ORDERS
```

この実行計画に基づく各問合せの実行にかかる時間 T_a, T_b は、IO の帯域幅を $B[Byte/sec]$、IOPS を $b[/sec]$、1 タプルあたりのハッシュ計算コストを $h[sec]$ とし、LINEITEM, ORDERS のタプル長を $l_L[B], l_O[B]$、タプル数を $|L|, |O|$、各選択演算子の選択率を σ_L, σ_O とすると次のようになる。

$$T_a = \frac{l_O|O|}{B} + \frac{\sigma_O|O|}{b} \tag{1}$$

$$T_b = \frac{l_O|O|}{B} + h|O| + \frac{l_L|L|}{B} + h\sigma_L|L| \tag{2}$$

Qa と Qb はどちらも ORDERS と LINEITEM の結合演算であるため、Qa と Qb が同時に入力されるならば既存の複数問合せ最適化手法を用いて同一の結合演算子を用いた大域的な実行計画を構築することができる。逐次的に入力される場合には、Qa のみでの実行はインデックス結合を用いているため既存の複数問合せ最適化手法では、Qa と Qb を同一のインデックス結合演算子を用いて共通化することになる。しかし、インデックス結合は内表に対してランダムアクセスとなるため選択率が十分小さくない限りは非常に大きな実行コストを必要とする。実際、Qb に対してインデックス結合を用いると ORDERS のタプル数分だけランダムアクセスが発生し、非効率的である。

新たに次のようにして結合演算子を動的に変換することにより、実行効率を低下させずに QaQb 間で処理の共有を行うことができる。ORDERS のうち Qa で既に読み取り終えている部分を ORDERS'、残りの部分を ORDERS" とする。Qa の実行計画ではインデックス結合を用いるため、既に ORDERS' と LINEITEM の結合結果は得られている (Q'_a)。Qb が入力された後、ORDERS' については Qb の実行計画と同様に LINEITEM に選択演算を施してからハッシュ結合を行う (Q'_b)。ORDERS" については LINEITEM と結合した後に、それぞれの選択演算を施して Qa 及び Qb の結果として出力する (Q''_{ab})。これらを実行計画として図示すると図 2 のようになる。

この変換が等価変換であることを示す。ORDERS, LINEITEM を O, L とする。また、問合せ Q の実行結果を $[\![Q]\!]$ で表すとし、関係 A, B の非交和を $A \uplus B$ とし、写像の和を $(f \oplus g)(A) = f(A) \cup g(A)$ で定義する。この時、

$$\begin{aligned}
&[\![Q_a]\!] \uplus [\![Q_b]\!] \\
=&\sigma_O(O) \bowtie L \uplus O \bowtie \sigma_L(L) \\
=&\sigma_O(O_1 \uplus O_2) \bowtie L \uplus (O_1 \uplus O_2) \bowtie \sigma_L(L) \\
=&\sigma_O(O_1) \bowtie L \uplus (\sigma_O \oplus \sigma_L)(O_2 \bowtie L) \uplus O_1 \bowtie \sigma_L(L) \\
=&[\![Q'_a]\!] \uplus [\![Q''_{ab}]\!] \uplus [\![Q'_b]\!]
\end{aligned}$$

となり、示された。

この実行にかかる時間 T_{ab} は Qb が入力されたのが Qa の進捗が p の時点だとすると

$$\begin{aligned}
T_{ab} = &\left(\frac{l_O|O_1|}{B} + \frac{\sigma_O|O_1|}{b}\right) + \left(\frac{l_O|O_2|}{B} + h|O_2|\right) \\
&+ \left(\frac{l_O|O_1|}{B} + h|O_1|\right) + \frac{l_L|L|}{B} + h|L|
\end{aligned} \tag{3}$$

となり、これによる実行時間の利得は以下のようになる。

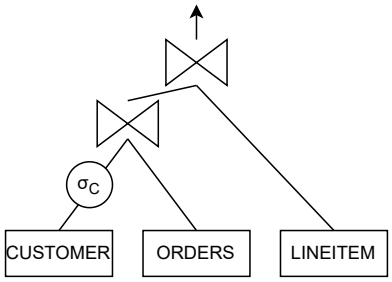
$$T_a + T_b - T_{ab} = \frac{l_O|O_2|}{B} + \frac{\sigma_O|O_2|}{b} - h(1-\sigma_L)|L|$$

右辺第一項は Qa と Qb における ORDERS" と LINEITEM の結合を共有したことによる実行時間の削減であり、右辺第二項は σ_L を通過しない LINEITEM のタプルに対するハッシュ計算分の実行時間増加を表す。

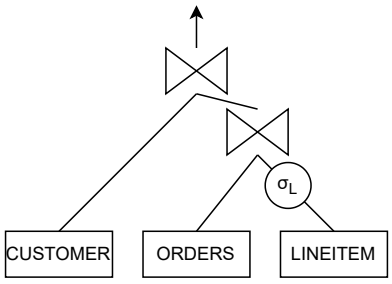
3.2 Case Study: 3-way Join

次に以下の問合せ Qa, Qb を考える。

```
Qa: SELECT * FROM LINEITEM, ORDERS, CUSTOMER
```



(a) Qa



(b) Qb

図 3: 3-way join 問合せの実行計画: 但し、結合演算はすべてハッシュ結合、σ_C は C_MKTSEGMENT = 'BUILDING' を満たす要素を返す選択演算、σ_L は L_QUANTITY < 11 を満たす要素を返す選択演算。

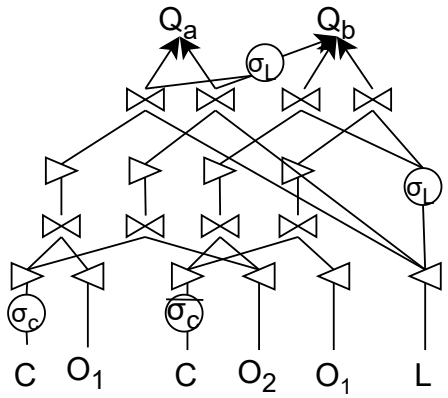


図 4: Qa の実行中に Qb が入力された時の実行計画変換: 但し、ORDERS のうち Qa において既に読み取り終えている部分が O_1、残りの部分が O_2 である。また $\triangleright$ はハッシュ表構築、$\triangleleft$ はハッシュ表検索のためのハッシュ値の計算を示す。

```
    WHERE L_ORDERKEY = O_ORDERKEY AND
      O_CUSTKEY = C_CUSTKEY AND
      C_MKTSEGMENT = 'BUILDING'
Qb: SELECT * FROM LINEITEM, ORDERS, CUSTOMER
    WHERE L_ORDERKEY = O_ORDERKEY AND
      O_CUSTKEY = C_CUSTKEY AND
      L_QUANTITY < 11
```

これらの問合せは LINEITEM, ORDERS, CUSTOMER の 3 つのテーブルを結合する問合せである。TPC-H(SF=10) において PostgreSQL でこれらの問合せを実行するとき最も実行時間が短くなる実行計画を図示すると図 3 のようになる。2-way Join の場合とは異なり結合演算子はすべてハッシュ結合で実装されるが、Qa と Qb でその結合順序が異なる。

Qa の実行中 CUSTOMER テーブルに対しハッシュ表を構築したのち ORDERS テーブルを用いてそのハッシュ表を検索している途中で Qb が入力された場合に、QaQb 間で処理を共有する実行計画変換の一例を図 4 に示す。図 4 では IO だけでなくハッシュ表の構築やハッシュ値の計算を問合せ間で共有することにより、全体的な実行コストを削減している。この実行計画をたどると以下のようになり、Q_a, Q_b の実行結果は実行計画変換によって変わらない。

$$
\begin{aligned}
[\![Q_a]\!] =& (\sigma_C(C) \bowtie O_1) \bowtie L \cup (\sigma_C(C) \bowtie O_2) \bowtie L \\
=& (\sigma_C(C) \bowtie (O_1 \cup O_2)) \bowtie L \\
=& \sigma_C(C) \bowtie O \bowtie L \\
[\![Q_b]\!] =& \sigma_L((\sigma_C(C) \bowtie O_1) \bowtie L \cup (\sigma_C(C) \bowtie O_2) \bowtie L) \cup \\
& (\bar{\sigma}_C(C) \bowtie O_1) \bowtie \sigma_L(L) \cup (\bar{\sigma}_C(C) \bowtie O_2) \bowtie \sigma_L(L) \\
=& \sigma_C(C) \bowtie O \bowtie \sigma_L(L) \cup \bar{\sigma}_C(C) \bowtie O \bowtie \sigma_L(L) \\
=& C \bowtie O \bowtie \sigma_L(L)
\end{aligned}
$$

また、Qa,Qb を独立に実行したときの実行時間 T_a, T_b は次のようになり、

$$
\begin{aligned}
T_a &= \frac{l_C|C|}{B} + h\sigma_C|C| + \frac{l_O|O|}{B} + h|O| + h\,|\sigma_C(C) \bowtie O| + \frac{l_L|L|}{B} + h|L| \\
T_b &= \frac{l_C|C|}{B} + h|C| + \frac{l_O|O|}{B} + h|O| + \frac{l_L|L|}{B} + h\sigma_L|L| + h\,|O \bowtie \sigma_L(L)|
\end{aligned}
$$

一方、図 4 に示した実行計画に基づいて実行したときの実行時間 T_{ab} は次のようになる。

$$
\begin{aligned}
T_{ab} =& \frac{l_C|C|}{B} + h\sigma_C|C| + \frac{l_o|O_1|}{B} + h|O_1| + h|\sigma_C(C) \bowtie O_1| \\
& + \frac{l_C|C|}{B} + h(1-\sigma_C)|C| + \frac{l_O|O_2|}{B} + h|O_2| + h|\sigma_C(C) \bowtie O_2| \\
& + \frac{l_O|O_1|}{B} + h|O_1| + h|\sigma_C(C) \bowtie O_1| + \frac{l_L|L|}{B} + h|L| \\
=& \frac{2l_C|C| + l_O|O| + l_O|O_1| + l_L|L|}{B} \\
& + h(|C| + |O| + |C \bowtie O| + |L| + |O_1|)
\end{aligned}
$$

これによって削減される実行時間は

$$
\begin{aligned}
& T_a + T_b - T_{ab} \\
=& \frac{l_O|O_2| + l_L|L|}{B} + h(\sigma_C|C| + |O_2| + \sigma_L|L| + |O \bowtie \sigma_L(L)|) \\
& - h|\bar{\sigma}_C(C) \bowtie O|
\end{aligned}
$$

第一項及び第二項は処理を共有したことによって削減された IO 及びハッシュ計算コストである。第三項は元の実行計画では現れなかった $\bar{\sigma}_C(C) \bowtie O$ に対するコストである。

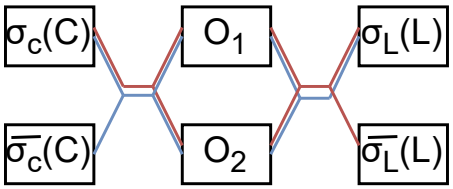


図 5: 3-way Join クエリの Qa と Qb に対して動的複数問合せ最適化を行う際の拡張ハイパーグラフ。図中赤線は Qa に対応するハイパーエッジであり、青線は Qb に対応するハイパーエッジである。

4 動的複数問合せ最適化

動的複数問合せ最適化は通常の問合せ最適化とは異なり、入力された問合せだけでなく現在のデータベースシステムの状態を考慮して最適化を行う。具体的に 3 章のケーススタディでは逐次読み取りの進捗や構築されたハッシュ表の状態等を考慮している。そこで従来の DPhyp では入力された問合せのみに基づいてハイパーグラフを構築していたのに対して、データベースシステムの状態をハイパーグラフ上で表現することにより、動的複数問合せ最適化を実現する。

同時に処理すべき問合せ集合 $Q = \{q_1, \ldots, q_n\}$ に対し、ハイパーグラフを次のように拡張する。

定義 (拡張ハイパーグラフ)**.** 拡張ハイパーグラフ $H = (V, E)$ は非空頂点集合 V と拡張ハイパーエッジ集合 E の組であり、拡張ハイパーエッジは V の非空部分集合 $u \subseteq V, v \subseteq V$ とそのハイパーエッジ (u, v) に対応する結合述語を含む問合せの集合 $q \subseteq Q$ からなる三つ組み (u, v, q) である。

定義 (部分グラフ)**.** 拡張ハイパーグラフ $H = (V, E)$ に対し、頂点集合 $V' \subseteq V$、問合せ集合 $Q' \subseteq Q$ に対し、この頂点集合及び問合せ集合が導く部分グラフ $H_{|V',Q'}$ は $E' = \{(u, v, q) | (u, v) \subseteq E, u \subseteq V', v \subseteq V', q \subseteq Q'\}$ を用いて $H_{|V',Q'} = (V', E')$ と表される。

定義 (連結)**.** 問合せ集合 $Q' \subseteq Q$ に対し拡張ハイパーグラフ $H = (V, E)$ が Q'-連結であるとは、$|V| = 1$ または V の分割 V', V'' が存在し、$u \subseteq V, v \subseteq V, q \subseteq Q'$ かつ $H_{|V',Q'}$ 及び $H_{|V'',Q'}$ が $Q'-$ 連結なハイパーエッジ $(u, v, q) \in E$ が存在することである。さらに、$V' \subseteq V$ が導く部分グラフ $H_{|V'}$ が Q'-連結であるとき、V' は H の Q'-連結部分グラフ *(csg)* であるといい、$Q'-$ 連結部分グラフ $H_{|V'',Q'}$ を導く $V'' \subseteq (V \setminus V')$ が存在するとき、V'' を V' の連結補部分グラフ *(cmp)* という。

定義 (csg-cmp-pair)**.** 拡張ハイパーグラフを $H = (V, E)$ とし、$S_1 \subseteq V, S_2 \subseteq (V \setminus S_1)$ が問合せ集合 Q' に対し Q'-連結な部分グラフ及びその Q'-連結補部分グラフであり、さらにある辺 $(u, v, q) \in E$ が存在し、$u \subseteq S_1, v \subseteq S_2, q \supseteq Q'$ であるとき、(S_1, S_2) を Q'*-csg-cmp-pair(*Q'*-ccp)* という。

通常の問合せ最適化ではハイパーグラフの各頂点は問合せに出現する各テーブル T であるが、動的複数問合せ最適化では問合せ中の単一のテーブル T のみに作用する選択演算子 σ に対し $\sigma(T)$ 及び $\bar{\sigma}(T)$ をそれぞれ頂点とする。また、3.1 節や 3.2 節で議論したケーススタディのように各テーブルへの処理の途中で再度最適化を行う場合は既に処理が完了している領域とまだ行われていない領域に分割し、それぞれを頂点とする。図 5 は 3.2 節の 3-way Join のケースにおける拡張ハイパーグラフを示した例である。

Algorithm 3 DMQO-DPhyp

```
1:  function DMQO-DPhyp(Q = {q_1, ..., q_n} : queries, M : internal status)
2:      V = a set of all tables
3:      for all σ is a single table predicate apeared in any query
4:          R = a target table of σ
5:          V = V \ R ∪ {σ(R)} ∪ {σ̄(R)}
6:      for all r ∈ V
7:          dpTable[{r}] = {(Q, r)}
8:      for all m ∈ M
9:          R is a set of tables in m
10:         dpTable[R] = {(Q, m)}
11:     for ∀csg-cmp-pairs(S_1, S_2, Q)
12:         e = {(u, v, q) | u ⊆ S_1, v ⊆ S_2, q ⊇ Q}
13:         P_1 = {p | ∃q'.(p, q') ∈ dpTable[S_1], q' ⊇ Q}
14:         P_2 = {p | ∃q'.(p, q') ∈ dpTable[S_2], q' ⊇ Q}
15:         P = buildPlanTree(P_1, P_2, e)
16:         P' = {p | ∃q'.(p, q') ∈ dpTable[S_1 ∪ S_2], q' ⊇ Q}
17:         if cost(P) < min_{p∈P'} cost(p) then
18:             dpTable[Q] = P
19:     return dpTable[V]
```

DPhyp を拡張し、動的複数問合せ最適化を行うアルゴリズムをアルゴリズム 3 に示した。本アルゴリズムは問合せ集合 Q とデータベースシステムの内部状態 M を入力とする。内部状態 M は典型的には構築済みのハッシュ表や問合せ実行中に実体化された中間テーブル等である。これらの内部状態は既に構築済みであるため、新たな計算なしにコスト 0 で利用可能である。通常のテーブルと同様に dpTable に登録され (8-10 行目)、これに対するコスト関数の呼び出し結果は 0 となる。dpTable には各頂点を葉とする部分問合せの実行計画とその部分問合せを含みうる問合せ集合の組が登録される。各 ccp に対し、その ccp から導かれる部分問合せの実行計画を組み合わせて新たな実行計画を構築し、既存の実行計画よりもコストが低い場合には dpTable に登録する。

5 評　　価

5.1 シミュレーションの設定

各問合せ実行のコストモデルとして式 (1)-(3) を用いる。コストモデル中の各定数の値は表 1 に示す。この値は実験用計算機（Intel Xeon 6132 (2× 28cores, 2.60 GHz), 93GB RAM, Ubuntu 20.04.6）において PostgreSQL(15devel) を用いた時の実測値に基づく。

表 1: シミュレーションのパラメータ

説明	パラメータ	値
IO の帯域幅	B	1500 MB/sec
IOPS	b	3000 /sec
1 タプルあたりのハッシュ計算コスト	h	2.0×10^{-7} sec
LINEITEM のタプル長	l_L	117 bytes
ORDERS のタプル長	l_O	107 bytes
CUSTOMER のタプル長	l_C	158 bytes
LINEITEM のタプル数	$\|L\|$	60000000
ORDERS のタプル数	$\|O\|$	6000000
CUSTOMER のタプル数	$\|C\|$	1500000

5.2 実行計画の動的変換による実行時間

2-way join 問合せの Qa の実行中に Qb が入力される状況において、各問合せを個別に実行する場合と結合演算の動的変換を用いて実行を最適化する場合の実行時間をシミュレーションした結果を図 6 に示す。図 6(a) は Qb が入力されるタイミングを変化させた時の実行時間の変化を示す。個別実行はいつ Qb が入力されたとしても Qa が実行完了してから Qb の実行を開始するので実行時間は不変。一方、変換を適用した実行は Qb が早期に入力された方が共有可能な仕事量が増加するため実行時間は短くなる。

図 6(b)(c) はそれぞれ Qa,Qb の WHERE 句の述語の選択率を変化させた場合の実行時間の変化を示す。それぞれの設定において結合演算子の動的変換を用いることにより実行時間を削減可能な領域が存在し、最大で 1.45 倍の高速化が可能である。また、実際に PostgreSQL 上で選択率を変化させながら Qa、Qb を実行した場合の実行時間に基づいた $T_a + T_b$ を示している。シミュレーションによる実行時間よりも数秒長い結果が得られているが、これは式 (1)-(3) に選択演算子の実行コストが組み込まれていないからである。

2-way Join の場合と同様に 3-way Join 問合せの場合についても実行時間のシミュレーションを行った結果を図 7 に示す。図 7 からわかるように、3-way Join 問合せにおいても動的変換を用いることにより実行時間を削減することが可能である。

6 関 連 研 究

6.1 問合せ最適化

コストベースの問合せ最適化では、入力された問合せを実現可能な実行計画をすべて列挙した後、各実行計画の実行コストを計算し、最も実行コストが小さくなる実行計画を選択する。この手法は System R [10] で導入されて以降広く利用されてきた。しかし、多数の関係の結合順序や相関のある選択演算の適用順序の最適化問題は IO コストや実行時間に基づくコストモデルのもとでは NP 困難であり [8,11]、扱いやすい統計情報に基づくコストモデルやヒューリスティクスが代わりに用いられてきた。最もよく用いられる統計値であるカーディナリティの推定アルゴリズムは長年に渡って研究が行われており、近年では機械学習を用いたアプローチも試みられている [12,13,14]。

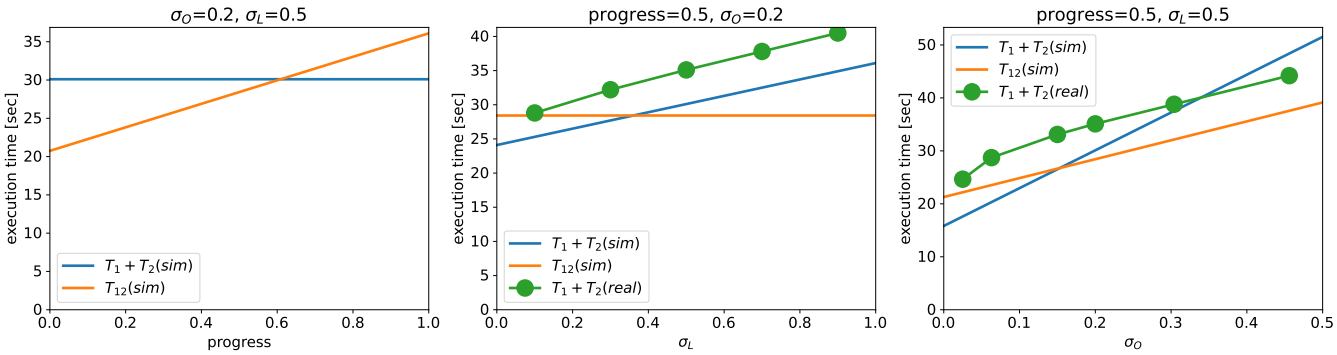
問合せ最適化は、入力された問合せに対して実行可能な実行計画をすべて列挙することと、その実行計画候補に対して実行コストを推定することの二つのステップからなる。実行計画の列挙は主に結合演算の組合せ及び順序の列挙によって行われる。厳密なコストモデルでは結合演算は可換ではないが、可換性を仮定することで動的計画法を用いて効率的な列挙が可能となる。初期の動的計画法による実行計画列挙アルゴリズムとして DPsize と DPsub が知られている。DPsize は実行計画のサイズ、つまり実行計画に含まれる表の個数を基準に動的計画法を適用し、DPsub は問合せグラフ中の部分グラフの包含関係を基準に動的計画法を適用する。問合せグラフの形状が鎖型や環型の場合には DPsize が、星型やクリーク型の場合には DPsub がより効率的であることが知られている [9]。DPccp [9] は csg-cmp-pair と呼ばれる連結部分グラフの組のみを列挙することで、より効率的に実行計画を列挙する。DPccp をハイパーグラフに拡張した DPhyp [15] は結合演算として可換な内部結合だけでなく、左結合や右結合、反結合などの非可換な結合演算も考慮することが可能である。

6.2 動的問合せ最適化

コストモデルの近似や統計情報の推定の誤差を軽減する長年の試みにも関わらず、誤った問合せ最適化による実行コストの増加は依然として中心的課題である。動的問合せ最適化は問合せ最適化の不正確性に対処するアプローチの一つであり、実行状態をフィードバックしながら問合せ処理を実行中に修正し、不正確な問合せ最適化によるレイテンシの増加や実行効率の悪化を軽減する [2,3]。既存の動的問合せ最適化の試みは実行計画の修正を許すタイミングによって、適応的問合せ処理と実行時再最適化に大別される。適応的問合せ処理は実行計画の修正を簡単なものに限定する代わりに、基本的には常に実行計画を修正することができる。一方実行時再最適化は各演算子の処理を完了したタイミングでのみ実行計画の修正を許すが、実行計画の修正はより自由度が高い。

演算子の適応的並び替えは複数の単項演算子を連続的に適用するパイプライン型の実行計画に対する適応的問合せ処理手法である。複数の述語の連言からなる問合せの実行計画は選択演算子が順に適用されるパイプラインとなるが、演算子の適用順序は可換である。A-Greedy [16] は各選択演算子の実行コストや選択率を観測し、全体の実行コストを最小化するように演算子の適用順序を実行時に変更する。Eddy [17] はタプルルーティングにより演算子の適用順序をタプルごとに管理することでより細かな制御を可能にする。

一般の結合演算を含む問合せを実現する実行計画は膨大であり、適応的問合せ処理は非常に複雑になる。しかし、left-deep なインデックス結合や right-deep なハッシュ結合のみからなる実行計画に候補を限定することで、結合演算の適応的処理はパイプライン型実行計画における適応的な演算子の並び替えに帰着され、簡易的に実現することができる [16,18]。また対称ハッシュ結合演算子等の常に可換な結合演算子を用いることにより更に自由度の高い結合演算の適応的処理が可能となる [19]。



(a) progress による実行時間の変化　(b) σ_L による実行時間の変化　(c) σ_O による実行時間の変化

図 6: 2-way join 問合せの実行時間シミュレーション

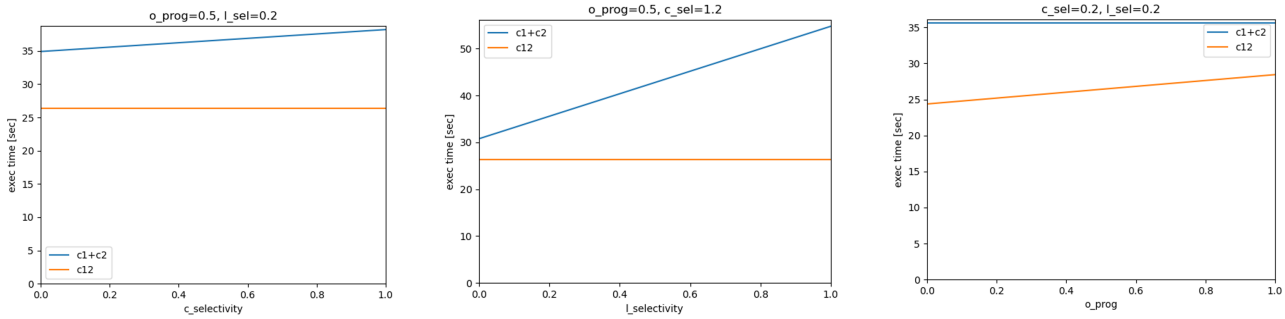


図 7: 2-way join 問合せの実行時間シミュレーション

実行時再最適化は問合せ最適化の不正確性を克服する別のアプローチであり、予め問合せ処理を複数の段階に分割しておき、各段階の実行完了後に現在の状態に基づいて再度問合せ最適化を行う [1, 20]。これにより統計情報の不正確な推定値に基づく誤った問合せ最適化による実行効率の低下を防ぐ。一方 [21] は常に中間結果を実体化するわけではないが、問合せ最適化時の統計情報の推定値と実測された値が大きく異なるときにその時点までの中間結果を実体化し、実測された統計情報と実体化された中間結果に基づいて新たな実行計画を生成する。また [22] は宣言的なプログラミング言語である datalog を用いて今までの問合せ最適化結果から逐次的に最適化計算を行うことで実行時再最適化にかかる時間を削減する。

近年では実行計画を just-in-time(JIT) コンパイルしバイトコードやネイティブバイナリに変換することで、メモリキャッシュや SIMD 演算を活用した低レベルでの問合せ処理の高速化が試みられている。誤った問合せ最適化結果に基づく場合や低レベルの高速化技法によってむしろ実行効率が低下する場合などには再コンパイルが望ましいが、JIT コンパイルは長時間を必要とするため難しい。Micro Adaptivity [23] は各演算子に対して様々な高速化手法を適用した複数のバイナリを予め用意し、これらを適応的に切り替えることにより再コンパイルなしに低レベルの高速化手法の有無を選択する。PCQ [24] はバイトコードレベルでの適応的並び替えを行うことにより、再コンパイルなしでの適応的問合せ処理を可能にする。

6.3 複数問合せ最適化

現実のサービスで用いられるデータベースシステムは単一の問合せのみを処理するのではなく、並行に入力された複数の問合せを処理する必要がある。複数問合せ最適化は複数の問合せにまたがった実行計画を用い、共通化可能な処理を共通化することで全体的な実行コストを削減する。初期の複数問合せ最適化は共通部分式除去に基づいており、各問合せに対し独立に問合せ最適化を行い、得られた実行計画に同一の箇所が存在する場合にその評価結果を問合せ間で共有することで、全体の実行コストを削減する [4, 5]。

QPipe [25] や CJoin [26]、DataPath [27] は複数の問合せの実行計画を組み合わせて単一のパイプラインネットワークを作り、これに基づいて実行する。新たな問合せが入力されると現在のパイプラインネットワークに新たな実行計画を順次組み込むことで、多数の問合せを並行実行する。

SWO [6] は個々の問合せに対して独立に問合せ最適化を行った後実行計画を組み合わせるのではなく、複数の問合せにまたがる問合せ最適化問題を解き、大域的な実行計画を生成する。大域的な問合せ最適化は複雑で計算時間が長大だが、Roulette [7] は Eddy ベースの適応的問合せ処理と強化学習を組み合わせることで、大域的な問合せ最適化による実行時間の増加を抑制した。

7 結　　論

問合せ最適化は問合せ処理の実行効率を左右する中心的要素

であるにも関わらず、統計情報の推定値や近似されたコストモデルに頼っており、誤った問合せ最適化結果が得られてしまう場合が多い。これによる実行効率の低下を防ぐために、適応的問合せ処理や実行時再最適化といった動的問合せ最適化手法が研究されてきた。本稿では動的問合せ最適化を複数の問合せが逐次的に入力される状況に適用した動的複数問合せ最適化を提案した。さらに、動的計画法を用いた問合せ最適化アルゴリズム DPhyp を拡張することで、動的複数問合せ最適化を可能にするアルゴリズムを提案した。また、PostgreSQL 上での問合せ実行結果に基づくシミュレーションにより本手法の潜在的な有効性を検証した。

謝　　辞

本研究の一部は、ビッグデータ価値協創プラットフォーム工学社会連携研究部門（東大、日立）の支援を受けたものである。

文　　献

[1] Eugene Wong and Karel Youssefi. Decomposition - A strategy for query processing. *ACM Trans. Database Syst.*, Vol. 1, No. 3, pp. 223–241, 1976.

[2] Anastasios Gounaris, Norman W. Paton, Alvaro A. A. Fernandes, and Rizos Sakellariou. Adaptive query processing: A survey. In *Advances in Databases, 19th British National Conference on Databases, BNCOD 19, Sheffield, UK, July 17-19, 2002, Proceedings*, Vol. 2405 of *Lecture Notes in Computer Science*, pp. 11–25. Springer, 2002.

[3] Amol Deshpande, Zachary G. Ives, and Vijayshankar Raman. Adaptive query processing. *Found. Trends Databases*, Vol. 1, No. 1, pp. 1–140, 2007.

[4] Timos K. Sellis. Multiple-query optimization. *ACM Trans. Database Syst.*, Vol. 13, No. 1, pp. 23–52, 1988.

[5] Timos K. Sellis and Subrata Ghosh. On the multiple-query optimization problem. *IEEE Trans. Knowl. Data Eng.*, Vol. 2, No. 2, pp. 262–266, 1990.

[6] Georgios Giannikis, Darko Makreshanski, Gustavo Alonso, and Donald Kossmann. Shared workload optimization. *Proc. VLDB Endow.*, Vol. 7, No. 6, pp. 429–440, 2014.

[7] Panagiotis Sioulas and Anastasia Ailamaki. Scalable multi-query execution using reinforcement learning. In *SIGMOD'21*, pp. 1651–1663, 2021.

[8] Toshihide Ibaraki and Tiko Kameda. On the optimal nesting order for computing n-relational joins. *ACM Trans. Database Syst.*, Vol. 9, No. 3, pp. 482–502, 1984.

[9] Guido Moerkotte and Thomas Neumann. Analysis of two existing and one new dynamic programming algorithm for the generation of optimal bushy join trees without cross products. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, p. 930–941. VLDB Endowment, 2006.

[10] Morton M. Astrahan, Mike W. Blasgen, Donald D. Chamberlin, Kapali P. Eswaran, Jim Gray, Patricia P. Griffiths, W. Frank King III, Raymond A. Lorie, Paul R. McJones, James W. Mehl, Gianfranco R. Putzolu, Irving L. Traiger, Bradford W. Wade, and Vera Watson. System R: relational approach to database management. *ACM Trans. Database Syst.*, Vol. 1, No. 2, pp. 97–137, 1976.

[11] Kamesh Munagala, Shivnath Babu, Rajeev Motwani, and Jennifer Widom. The pipelined set cover problem. In *ICDT'05*, Vol. 3363 of *Lecture Notes in Computer Science*, pp. 83–98, 2005.

[12] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA'07*, pp. 137–156, 2007.

[13] Max Halford, Philippe Saint-Pierre, and Franck Morvan. An approach based on bayesian networks for query selectivity estimation. In *DASFAA'19*, pp. 3–19, 2019.

[14] Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Xi Chen, and Ion Stoica. Neurocard: One cardinality estimator for all tables. *Proc. VLDB Endow.*, Vol. 14, No. 1, pp. 61–73, 2020.

[15] Guido Moerkotte and Thomas Neumann. Dynamic programming strikes back. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, p. 539–552, New York, NY, USA, 2008. Association for Computing Machinery.

[16] Shivnath Babu, Rajeev Motwani, Kamesh Munagala, Itaru Nishizawa, and Jennifer Widom. Adaptive ordering of pipelined stream filters. In *SIGMOD'04*, pp. 407–418, 2004.

[17] Ron Avnur and Joseph M. Hellerstein. Eddies: Continuously adaptive query processing. In *SIGMOD'00*, pp. 261–272, 2000.

[18] Quanzhong Li, Minglong Shao, Volker Markl, Kevin S. Beyer, Latha S. Colby, and Guy M. Lohman. Adaptively reordering joins during query execution. In *ICDE'07*, pp. 26–35. IEEE Computer Society, 2007.

[19] Vijayshankar Raman, Amol Deshpande, and Joseph M. Hellerstein. Using state modules for adaptive query processing. In *ICDE'03*, pp. 353–364, 2003.

[20] Christina Pavlopoulou, Michael J. Carey, and Vassilis J. Tsotras. Revisiting runtime dynamic optimization for join queries in big data management systems. In *EDBT'22*, pp. 1:1–1:12, 2022.

[21] Navin Kabra and David J. DeWitt. Efficient mid-query re-optimization of sub-optimal query execution plans. In *SIGMOD'98*, p. 106–117, New York, NY, USA, 1998. Association for Computing Machinery.

[22] Mengmeng Liu, Zachary G. Ives, and Boon Thau Loo. Enabling incremental query re-optimization. In *SIGMOD'16*, pp. 1705–1720, 2016.

[23] Bogdan Raducanu, Peter A. Boncz, and Marcin Zukowski. Micro adaptivity in vectorwise. In *SIGMOD'13*, pp. 1231–1242. ACM, 2013.

[24] Prashanth Menon, Amadou Ngom, Todd C. Mowry, Andrew Pavlo, and Lin Ma. Permutable compiled queries: Dynamically adapting compiled queries without recompiling. *Proc. VLDB Endow.*, Vol. 14, No. 2, pp. 101–113, 2020.

[25] Stavros Harizopoulos, Vladislav Shkapenyuk, and Anastassia Ailamaki. Qpipe: A simultaneously pipelined relational query engine. In *SIGMOD'05*, pp. 383–394, 2005.

[26] George Candea, Neoklis Polyzotis, and Radek Vingralek. Predictable performance and high query concurrency for data analytics. *VLDB J.*, Vol. 20, No. 2, pp. 227–248, 2011.

[27] Subi Arumugam, Alin Dobra, Christopher M. Jermaine, Niketan Pansare, and Luis Leopoldo Perez. The datapath system: a data-centric analytic processing engine for large data warehouses. In *SIGMOD'10*, pp. 519–530, 2010.

多次元ソートのための整数値変換手法の評価

高尾　大樹† 　黒川　能毅† 　西川　記史† 　茂木　和彦† 　藤原　真二††

† (株) 日立製作所 研究開発グループ デジタルサービスプラットフォームイノベーションセンタ
〒 185–8601 東京都国分寺市東恋ヶ窪一丁目 280 番地
†† (株) 日立製作所 クラウドサービスプラットフォームビジネスユニット デジタルプラットフォーム事業部
〒 244–0817 神奈川県横浜市戸塚区吉田町 292 番地
E-mail: †{takao.daiki.jh, yoshiki.kurokawa.ee, norifumi.nishikawa.mn, kazuhiko.mogi.uv}@hitachi.com,
††shinji.fujiwara.yc@hitachi.com

あらまし　近年の DX の進展や意思決定の高速化に伴い，分析クエリの処理速度向上が求められる．入力データを多次元ソートしてからデータベースに格納することで，様々なワークロードに対してデータ読出し量を削減でき，処理速度の大幅な向上が期待できる．処理効率化のため多次元ソート適用前に各カラムのデータを事前に整数値へ変換することが多い．整数値変換手法にはいくつか既存手法があるが，各手法を適用した際のデータ読出し量の削減効果は未知数であるため調査が必要である．そこで本報告では既存の整数値変換手法の性能について評価する．データ分布や問合せ範囲を変化させた様々なワークロードに対するデータ読出し量を比較評価し，各手法の利点と欠点を整理する．

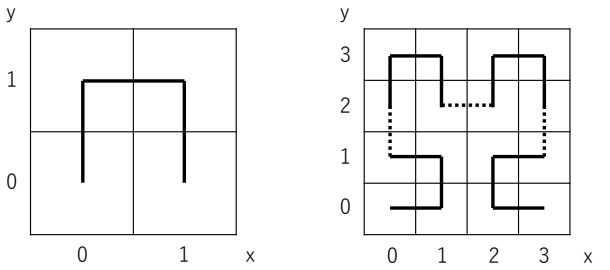
キーワード　データベース技術，チューニング・性能管理，ストレージ管理.

1　はじめに

近年の DX の進展や意思決定の高速化に伴い，分析クエリの処理速度向上が求められる．例として製造業に注目すると，工場内に設置したセンサを用いた機械の運転状況のモニタリングや製造工程の最適化をはじめとする事業の高度化・自動化 [1,2] が推し進められており，2029 年には Industrial Automation の市場規模が 3590 億ドルにまで達するとされている [3]．特に近年では事業の持続性及び生産性向上のため，不良品や機器不良による影響範囲の迅速な特定などの，ミクロなレベルでのトレーサビリティ分析の需要が高まってきている [4,5].

このような分析処理を高速化するアプローチの一つにデータ配置の最適化がある．従来のビジネスクエリは主にバルクロードと集約処理から構成されるのに対して，トレーサビリティ分析のクエリは選択率の高いフィルタリング処理と再帰的な結合処理から構成される [6]．すなわちトレーサビリティ分析のアクセスパターンには，大量のデータの中からフィルタリング条件を満たす少数の要素を抽出し，抽出した各要素と関係するデータとの結合を繰り返すという特徴がある．そのためクエリの処理対象となるデータをストレージ上の物理的に近い位置へと配置することで，範囲索引との併用によりデータ読出し量を大幅に削減でき，処理速度の改善が期待できる．ただしトレーサビリティ分析では機械の運転状態や温度などの環境情報をはじめ多様なデータが用いられるため，複数のカラムの値をバランス良く考慮してデータ配置を決定する必要がある．

そこで本稿では多次元ソートに基づくデータ配置最適化に注目する．あるカラムの値のみに基づく単純ソートとは異なり，多次元ソートでは複数カラムの値を考慮してデータを整列できるため，様々なカラムに対するフィルタリング処理において安定した読出しデータ量削減が期待できる．多次元ソートの代表的な手法としてヒルベルトソートがある [7]．ヒルベルトソートとは，図 1 に示すような空間充填曲線の一種であるヒルベルト曲線 [8] に沿ってデータを並び替える手法であり，他の空間充填曲線に比べてクラスタリング性能に優れるという特徴がある．ヒルベルトソートは入力として整数データのみを受け付けるため，実数や文字列などのデータを扱うためには事前に各カラムの値を整数値へと変換する必要がある．しかし，著者が知る限りでは整数値変換についてはこれまで十分な研究がなされておらず，整数値変換方法の違いがソート性能にどのような影響を与えるかは明らかでない．



(a) 2次元1次のヒルベルト曲線　(b) 2次元2次のヒルベルト曲線

図 1　ヒルベルト曲線の例

そこで本稿では各整数値変換手法の性能を調査する．まずはじめに各ヒストグラムの性質に注目して，各手法がソート性能に与える影響について仮説を立てる．そして，製造業ワークロードを模したデータセットを使用し，問合せ範囲を変化させたときのデータ読出し量を比較評価することで仮説を検証する．また，実験結果を踏まえて各手法が得意とするワークロードの特徴を整理し，整数値変換手法の適切な使い分けについても考察する．

本稿の構成は以下の通りである．2 章では関連研究としてヒルベルトソートについて概説し，3 章では多次元ソートのための整数値変換手法とその仮説について整理する．4 章では整数値変換手法の性能を実験により評価し，最後に 5 章で本研究をまとめる．

2 関 連 研 究

本章ではヒルベルトソートについて概説する．ヒルベルトソートとは空間充填曲線の一つであるヒルベルト曲線 [8] に沿ってデータを並べる多次元ソート手法である．空間充填曲線とは n 次元空間の単位空間をそれぞれ一度だけ通るような連続曲線であり，n 次元座標と 1 次元座標をマッピングできる．ヒルベルト曲線は図 1 のような形状をしており，n 次元のヒルベルト曲線に沿って隣接する 2 点の座標を比較すると，$n-1$ 次元の座標値は等しく，残りの 1 次元の座標値が 1 だけ異なるという特徴がある．この特徴により，ヒルベルト曲線は z-order 曲線 [9] など他の空間充填曲線と比較しても良いクラスタリング性能を持つことが知られており [10]，データベース分野でも応用事例が多数報告されている [11, 12].

n 次元空間の各軸を 2^m 分割したときの単位部分空間を一度だけ通るようなヒルベルト曲線を n 次元 m 次のヒルベルト曲線という．ヒルベルト曲線はフラクタル構造を持つため，n 次元 m 次のヒルベルト曲線は 2^n 個の n 次元 $m-1$ 次のヒルベルト曲線を n 次元 1 次のヒルベルト曲線に沿って連結したものとみなせる．すなわちヒルベルトソートされたデータ列は，2^n 個の部分空間でそれぞれヒルベルトソートしたものを n 次元 1 次のヒルベルト曲線に沿って整列することで得られるため，同じ部分空間に含まれるデータほど近い位置に配置されるという特徴がある．例えば図 1 より，2 次元 2 次のヒルベルト曲線は 4 つの 2 次元 1 次のヒルベルト曲線から構成され，それぞれのヒルベルト曲線をさらに 2 次元 1 次のヒルベルト曲線に沿って配置していると分かる．このとき同じ部分空間に含まれる，すなわち同じ 2 次元 1 次のヒルベルト曲線が通る座標値 $(1, 1)$ と座標値 $(0, 1)$ のデータはヒルベルトソートによって近くに配置されるが，異なる部分空間に含まれる座標値 $(2, 1)$ のデータは離れて配置される．

素朴なヒルベルトソートの実装における課題を以下に整理する．

（1） ヒルベルト曲線を生成する際，次元数に対して指数オーダーの計算時間を要する．この課題に対して，基本変換表と 1 次のヒルベルト曲線から m 次のヒルベルト曲線を高速に生成する手法 [13] や，ヒルベルト曲線を生成せずにデータをソートする手法 [7] が提案されている．

（2） 各次元の次数，すなわち分割数が等しいことを想定するため，各カラムのカーディナリティや値域幅が大きく異なる場合への適用が困難である．この課題に対して各次元の次数が異なる場合の拡張 [14] が提案されている．

（3） 入力として整数値を受け付けるため，実数や文字列などのデータを扱うためには事前に各カラムの値を整数値へと変換する必要がある．

すなわち事前に各カラムを整数値へと変換さえすれば，既存手法を活用することで各カラムの分割数が異なる場合にも高速にソートできるといえる．しかし，著者が知る限りでは整数値変換についてはこれまで十分な研究がなされておらず，整数値変換方法の違いがソート性能にどのような影響を与えるかは未知数である．

3 多次元ソートのための整数値変換

本章では多次元ソートのための整数値変換について整理する．まずはじめに単純な手法として，対応表に基づく整数値変換を考える．例えば製造業データのテーブル設計で広く用いられるスタースキーマでは，複数カラムを持つテーブルをファクトテーブルとディメンジョンテーブルに分割する [15]．このときディメンジョンテーブルでは各カラムのキー値として通し番号が割り振られ，ファクトテーブルの各カラムには対応するディメンジョンテーブルのキー値が格納される，すなわち整数値へと変換される．

しかし単に通し番号を割り当てるだけでは，多次元ソートの性能を十分に発揮できない．2 章で述べたとおりヒルベルトソートでは各カラムの値域幅を 2^m 分割するため，変換後のカーディナリティが 2^m となるときソートの性能を最大限発揮できる．しかし単純に通し番号を割り当てるだけではカーディナリティがちょうど 2^m となることは珍しく，$[0, 2^{m-1})$ の整数値はすべて割り当てられるが，$[2^{m-1}, 2^m)$ の整数値は一部しか割り当てられないことがほとんどである．このときヒルベルトソートが持つフラクタル構造の性質から，$[0, 2^{m-1})$ の範囲では複数カラムでバランスよくソートができるが，$[2^{m-1}, 2^m)$ では偏ったソート結果となってしまう．また変換前では複数のクラスタに分かれるようなデータ分布が変換後には一つのクラスタにまとまってしまうなど，通し番号の割当によりデータ分布を大きく変えてしまい，その結果大きく異なる値を近い位置にソートしてしまう恐れもある．ソート後のデータをストレージに格納しストレージからの読出し単位ごとに範囲索引を張ることを考えると，範囲索引の値域幅が大きくなるためフィルタリングが効かず，無駄なデータ読み出しが増加してしまう．

この問題の解決策として，Equi-width ヒストグラムに基づいた整数値変換が考えられる．パラメータとしてカラム i の次数 d_i が与えられたとき，カラム i の値域幅を 2^{d_i} 等分するようなヒストグラムを作成し，各行の値を対応するヒストグラムのビン id へと変換する．例として，$[0, 100)$ の範囲の値を取るカラム y の値を $[0, 3]$ の整数値へと変換する流れを図 2 に示す．このように値域幅が異なる各カラムの値を 2^{d_i} 個の整数値へと均等に変換することで上述の問題を解決できる．

しかし外れ値をはじめデータの分布に偏りがある場合，Equi-width ヒストグラムでは空のビンが多数生成され変換後のカーディナリティが低下してしまうため，多次元ソートの性能が低下する恐れがある．図 2 ではほとんどのデータが $[0, 25)$ の範囲に収まっているが，残り一部の外れ値が $[75, 100)$ の範囲に存

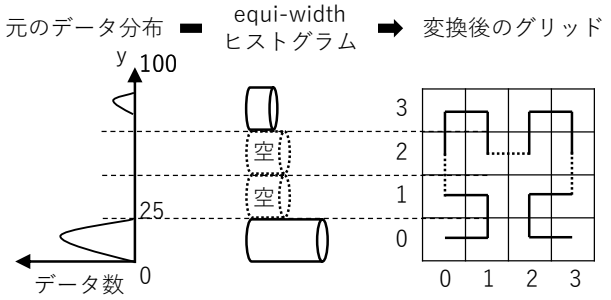


図 2　Equi-width ヒストグラムによる変換の例

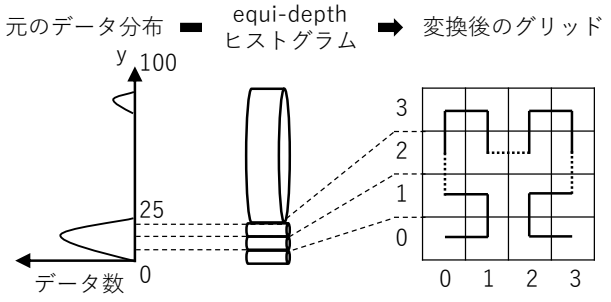


図 3　Equi-depth ヒストグラムによる変換の例

在している．このとき $[0, 25)$ の範囲のデータは全て 0 に変換されるため，これ以上細かくソートすることができない．その結果ストレージからの読出し単位内に $[0, 25)$ の範囲のデータが混在してしまうため，例えば $[5, 8)$ の範囲のデータを抽出する場合もほぼすべてのデータを一度ストレージから読み出さなくてはならない．

この問題の解決策として，Equi-depth ヒストグラム [16] に基づく手法が考えられる．この手法は図 3 に示すとおり，各ビンに含まれるデータ数が等しくなるよう値域幅を調整したヒストグラムを作成し，各行の値を対応するヒストグラムのビン id へと変換する．ヒストグラムの作成段階でデータ分布の偏りを吸収でき，また空のビンも生成されないため，Equi-width ヒストグラムの問題点を解消できる．図 3 においても $[0, 25)$ の範囲のデータに対して $[0, 3]$ の整数値を割り振れており，Equi-width ヒストグラムを用いる場合に比べてより細かい粒度でデータをソートできると分かる．

しかし Equi-depth ヒストグラムでは異なるクラスタに存在するデータが同じビンに割り当てられる，すなわち同じ整数値が割り当てられるため，この手法においても多次元ソートの性能が低下する恐れがある．図 2 の例ではデータ数とビン数の兼ね合いにより，25 近辺のデータと $[75, 100)$ の範囲のデータに同じ整数値が割り当てられ，これらのデータをこれ以上細かくソートできない．その結果ストレージからの読出し単位にこの 2 つのクラスタのデータが混在しやすくなり，外れ値のみの抽出や，$[25, 50)$ のようにデータが存在しない領域の範囲問合せに対して無駄なデータ読出しが増加する．なお，Equi-width ヒストグラムを用いる場合はこの 2 つのクラスタのデータが完全に分離されるためこのような問題は発生しない．

以上の議論を整理すると，外れ値など複数のクラスタに分かれるような分布のデータに対して以下のトレードオフが存在すると考えられる．

- データが多数存在するクラスタから一部のデータのみを抽出するような範囲問合せに対して，Equi-width ヒストグラムでは整数値変換の粒度が粗くなるため無駄なデータ読出しが増加するが，Equi-depth ヒストグラムでは各クラスタのデータ量に応じた整数値変換の粒度を保てるためデータ読出し量を削減できる．
- データが少数しか存在しないクラスタやクラスタを跨ぐような範囲問合せに対して，Equi-depth ヒストグラムでは異なるクラスタに存在するデータを同じ整数値へと変換する可能性があるため無駄なデータ読出しが増加するが，Equi-width ヒストグラムでは異なるクラスタのデータには異なる整数値を付与できるためデータ読出し量を削減できる．

表 1　ソート対象のカラム一覧

優先順	カラム名
1	DIMENSIONX
2	DIMENSIONY
3	DIMENSIONZ
4	TEMPERATURE
5	WEIGHT
6	SHAPESCORE

4　評価実験

本章では，Equi-width ヒストグラムおよび Equi-depth ヒストグラムに基づく整数値変換手法の性能評価について述べる．

4.1　実験設定

4.1.1　使用データとクエリ

本実験では製造業で用いられるワークロードを模したベンチマークである 4mbench [17] を使用する．4mbench では製造工程のデータモデルとして 4m モデル [18] を採用しており，製造工程を huMan，Machine，Material，および Method の 4 つの観点で表現する．具体的には製造を行った人（huMan），製造を行った機器（Machine），製造方法（Method）を格納する 3 つのディメンジョンテーブルと，利用された部材または原材料と生成された製品（Material）の履歴，それらを生成した業務の履歴，および機器の稼働履歴を格納する 3 つのファクトテーブルから構成される．

事前分析の結果，Material の履歴を扱う MATERIAL_LOG テーブルでは DIMENSIONX をはじめとするカラムにおいてデータ分布の大きな偏りが確認でき，整数値変換手法の性能評価に適すると分かった．そこで本実験では MATERIAL_LOG テーブルを使用し，ヒルベルトソートおよび整数値変換対象のカラムを表 1 のとおり設定する．また 4mbench にはスケールファクタとして工場のラインの数 l とデータを蓄積する日数 d の 2 つのパラメータがあるが，本実験ではそれぞれ $d = 1, l = 30$ とする．データ容量は非圧縮の CSV 形式で約 5.5GB である．

一方，4mbench のクエリは表 1 に示したカラム以外がフィルタリング対象となっており整数値変換手法の性能評価に適さないため，本実験では人手で設計した以下のクエリを使用する．

- **q1**：全データの 1.4% がヒットする，表 1 に示した 6 つのカラムに対する範囲問合せ．4mbench のクエリの中で特定の工程における良品の閾値として設定されている値を使用しており，独立した一つのクラスタ状に分布したデータを抽出する．
- **q2**：q1 に対してさらに検索条件を厳しく設定し，全データのわずか 0.2% のみがヒットする選択率の高い範囲問合せ．上述のクラスタから中央値近辺のごく一部のデータのみを抽出するため，外れ値により整数値変換の粒度が粗くなる Equi-width ヒストグラムにおいて性能低下が予想される．
- **q3**：最終工程の出力 Material を抽出する範囲問合せ．他工程の Material に比べて値が大きく異なるため外れ値の抽出に相当する．外れ値と定常値に同じ整数値を付与しやすい Equi-depth ヒストグラムにおいて性能低下が予想される．

4.1.2 評 価 方 法

本実験では以下の手順により整数値変換手法の性能を評価する．

（1） サンプリングレート $r = 0.05$ で無作為に抽出した一部のデータを用いて，表 1 に示すソート対象のカラムに対して，Equi-width ヒストグラムないし Equi-depth ヒストグラムを作成する．

（2） 作成したヒストグラムに基づいて，全データのソート対象カラムを整数値へと変換し，ヒルベルトソートを適用する．ただしソート対象のカラムは表 1 の優先順に従って選択する．例えばソート対象のカラム数を 2 とする場合は，DIMENSIONX, および DIMENSIONY を選択する．

（3） ヒルベルトソート済みの MATERIAL_LOG テーブルのデータをデータベースへと格納するとともに，ストレージからの読出し単位であるセグメントごとにソート対象のカラムに対して範囲索引を作成する．各セグメントの範囲索引の値域幅が狭いほど読出しセグメント数を削減できる傾向にある．

（4） 上述の各範囲問合せを与えたときのストレージからの読出しセグメント数に注目して整数値変換手法の性能を評価する．

なお本実験において，ヒルベルトソートは Imamura らの手法 [7] を参考に Python3.6.8 によって実装しており，Equi-width ヒストグラムおよび Equi-depth ヒストグラムは pandas の実装 [19, 20] をそれぞれ使用した．またデータベースとして商用の RDBMS である Hitachi Advanced Data Binder(HADB) [21][1]を使用し，整数値変換手法の純粋な性能を評価するため，データインポート時のデータ圧縮のオプションは NONE（非圧縮）を選択した．

4.2 実 験 結 果

ソート対象のカラムを変化させたときの各クエリにおける読出しセグメント数を図 4 に示す．ソート対象カラム数 2 の結果に注目してみると，独立した一つのクラスタ状に分布したデータを抽出する q1 では，Equi-depth ヒストグラムを用いた方が読出し量が多くなっている．これはクラスタの境界付近のデータが異なるクラスタの境界付近のデータと同じ整数値を割り当てられることで，異なるクラスタのデータが同じセグメントに混在してしまい，無駄なデータ読出しが発生してしまったことが原因であった．対してクラスタ中の一部のデータを抽出する q2 では，整数値変換の粒度が粗くなる Equi-width ヒストグラムを用いた方が読出し量が多くなっている．また，データが少数しか存在しない最終工程の出力を抽出する q3 では，他工程のデータとの混在が起こりやすい Equi-depth ヒストグラムを用いた方が読出し量が多くなっている．以上の結果より，整数値変換に用いるヒストグラムの違いが 3 章で述べたとおりソート性能に影響を及ぼすと分かった．また，データが多数存在するクラスタから一部のデータのみを抽出するような範囲問合せが多くなる場合は Equi-depth ヒストグラムを，クラスタ全体やクラスタを跨ぐような範囲問合せが多くなる場合は Equi-width ヒストグラムをそれぞれ用いるべきだといえる．

ここで各カラム間の関係性に注目すると，DIMENSIONX, DIMENSIONY，および DIMENSIONZ は Material のそれぞれ x，y，z 軸方向のサイズの計測値なのでこれらカラム間には強い相関関係がある．同様に Material の重量を表す WEIGHT および形状のスコアを表す SHAPESCORE もサイズを表す 3 カラムとの相関が大きいが，Material の温度を表す TEMPERATURE は他カラムとの相関は小さい．特に q3 の読出し対象のデータは Material のサイズと重量に関する値が他のデータに比べて総じて大きいため，Equi-depth ヒストグラムによる他のクラスタのデータとの混在の悪影響が小さくなり，二手法の性能差も小さくなったと考えられる．対してカラム間の相関が弱いデータセットにおいては，他クラスタのデータとの混在の影響も大きくなり，Equi-depth ヒストグラムの性能が悪化すると予想される．しかし 4mbench ではカラム間の相関を人手で調整できないため，カラム間の相関の強さを段階的に変化させたときの読出し量の変化などは評価できない．この仮説を検証するためには，カラム間の相関を制御できるデータセットを用いた追加実験が必要であり今後の課題である．

次にソート対象のカラム数による違いについて注目すると，カラム数 2 のときに比べてカラム数を 4 に増やすことで全体的に読出し量を削減できており，特に q1 においては 2 手法の読出し量は等しくなった．しかしカラム数をさらに 6 まで増やすと，Equi-depth の読出し量がわずかに増加しており，カラム数を増やすことで必ずしも性能が改善するわけではないと分かる．ただし本実験結果はカラム間の相関関係の影響も大きいため，ソート対象のカラム数による影響を正確に検証するにはやはり，カラム間の相関を制御できるデータセットを用いた実験が必要であり今後の課題である．

1：内閣府の最先端研究開発支援プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」（中心研究者：喜連川 情報・システム研究機構 機構長／東大特別教授）の成果を利用。

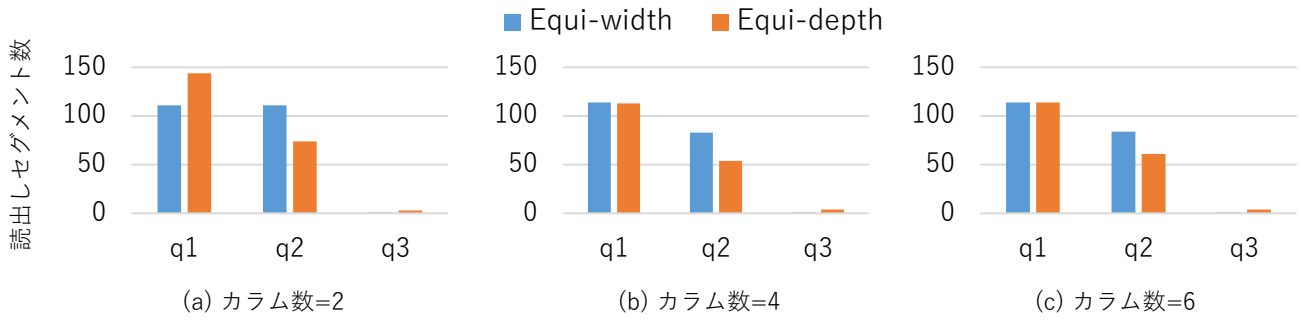


図 4　ソート対象のカラム数を変化させたときの読出しセグメント割合の変化

5 ま と め

本稿では Equi-width ヒストグラムおよび Equi-depth ヒストグラムを用いた各整数値変換手法の性能を調査した．外れ値など分布に偏りがあるデータを扱う際，Equi-width ヒストグラムでは空のビンが多数できることから整数値変換の粒度が粗くなり，Equi-depth ヒストグラムでは大きく値の異なるデータが同じ整数値を割り当てられると仮説を立てた．実験による評価結果よりこれら問題点がソート性能に悪影響を及ぼすため，データが多数存在するクラスタから一部のデータを抽出するような問合せが多くなる場合は Equi-depth ヒストグラムを，クラスタ全体やクラスタを跨ぐような問合せが多くなる場合は Equi-width ヒストグラムをそれぞれ用いるべきだと分かった．また実験によりカラム間の相関の強さがソート性能に影響を及ぼすと分かったが，この点に関する仮説はまだ検証できておらず，またソート対象のカラム数による影響についてはまだ分かっていない．

今後の課題として，カラム間の相関を制御できるデータセットを用いた，各手法のより詳細な性能評価が挙げられる．また，Equi-width ヒストグラムと Equi-depth ヒストグラム両方の欠点を解消する整数値変換手法の設計開発も今後の課題である．

謝　　辞

本研究の一部は,「戦略的イノベーション創造プログラム (SIP)」「統合型ヘルスケアシステムの構築」JPJ012425 の補助を受けて行った．また，本研究の一部は，東京大学生産技術研究所 ビッグデータ価値協創プラットフォーム工学 社会連携研究部門との共同研究の一環として実施した．

文　　献

[1] G. Gonçalves, J. a. Reis, R. Pinto, M. Alves, and J. a. Correia, “A step forward on intelligent factories: A smart sensor-oriented approach,” in *Proc. of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pp. 1–8, 2014.

[2] F. Tao, Q. Qi, and A. Liu, “Data-driven smart manufacturing,” *Journal of Manufacturing Systems*, vol. 48, pp. 157–169, 2018.

[3] F. B. Insights, *Industrial Automation Market Size, Share & COVID-19 Impact Analysis, By Component (Hardware, Software), By Industry (Discrete Automation, Process Automation), and Regional Forecast, 2022-2029*. Fortune Business Insights, 2022. Available: https://www.fortunebusinessinsights.com/industry-reports/industrial-automation-market-101589.

[4] Q. Xiu and K. Muro, “Robust inference traceability technology for product quality enhancement,” in *Proc. of 2017 IEEE Int. Conf. on Industrial Engineering and Engineering Management (IEEM)*, pp. 1699–1703, 2017.

[5] W. Chen, Y. Liu, and C. Chi, “Innovative traceability application in medical devices industry using the identification and resolution system for industrial internet,” in *Proc. of 13th Int. Conf. on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 72–76, 2021.

[6] N. Nishikawa, Y. Fujiwara, Shinjiand Hayamizu, and K. Goda, “Physical database design for manufacturing business analytics,” in *Proc. of Int. Conf. on Big Data (in press)*, 2023.

[7] Y. Imamura, T. Shinohara, K. Hirata, and T. Kuboyama, “Fast hilbert sort algorithm without using hilbert indices,” in *Proc. of Int. Conf. on Similarity Search and Applications (SISAP)*, pp. 259–267, 2016.

[8] D. Hilbert, “Ueber die stetige abbildung einer line auf ein flächenstück,” *Mathematische Annalen*, vol. 38, pp. 459–460, 1891.

[9] G. M. Morton, “A computer oriented geodetic data base; and a new technique in file sequencing,” tech. rep., Ottawa, Canada: IBM Ltd., 1966.

[10] C. Faloutsos and S. Roseman, “Fractals for secondary key retrieval,” in *Proc. of PODS*, pp. 247–252, 1989.

[11] I. Kamel and C. Faloutsos, “Hilbert r-tree: An improved r-tree using fractals,” in *Proc. of VLDB*, pp. 500–509, 1994.

[12] J. K. Lawder and P. J. H. King, “Querying multi-dimensional data indexed using the hilbert space-filling curve,” *SIGMOD Record*, vol. 30, no. 1, pp. 19–24, 2001.

[13] C. Li and Y. Feng, “Algorithm for analyzing n-dimensional hilbert curve,” in *Proc. of Int. Conf. on Advances in Web-Age Information Management*, pp. 657–662, 2005.

[14] C. H. Hamilton and A. Rau-Chaplin, “Compact hilbert indices: Space-filling curves for domains with unequal side lengths,” *Inf. Process. Lett.*, vol. 105, no. 5, pp. 155–163, 2008.

[15] A. Weininger, “Efficient execution of joins in a star schema,” in *Proc. of SIGMOD*, pp. 542–545, 2002.

[16] G. Piatetsky-Shapiro and C. Connell, “Accurate estimation of the number of tuples satisfying a condition,” in *Proc. of SIGMOD*, pp. 256–276, 1984.

[17] K. Goda, Y. Hayamizu, N. Nishikawa, and S. Fujiwara, “4mbench: Performance benchmark of manufacturing business database,” in *Proc. TPCTC*, pp. 94–109, 2022.

[18] C. Favi, M. Germani, and M. Marconi, “A 4M approach for a comprehensive analysis and improvement of manual assembly lines,” *Procedia Manufacturing*, vol. 11, pp. 1510–1518, 2017.

[19] “pandas.cut – pandas 2.1.4 documentation.” https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.cut.html (Accessed: Dec. 18, 2023).

[20] “pandas.qcut – pandas 2.1.4 documentation.” https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html (Accessed: Dec. 18, 2023).

[21] “Manuals - Middleware - Hitachi Advanced Database.” https://itpfdoc.hitachi.co.jp/Pages/region/document_list/manuals/en/hadb.html (Accessed: Dec. 18, 2023).

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

2024年2月28日(水) 13:30 ~ 15:40 | T2-B オンライン（Zoom Events)

データ活用におけるプライバシ

座長:小口 正人(お茶の水女子大学)
コメンテータ:早水 悠登(東京大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T2-B-2-01] m-不変性に基づいた連続的データ公開における安全性問題と小規模データ改ざんによる解決法

*張 扶蘇[1]、上土井 陽子[1]、若林 真一[1] (1. 広島市立大学)

14:00 ~ 14:25
[T2-B-2-02] 準同型暗号を活用したプライバシ保護型ニュース推薦手法

*中野 瑛心[1]、鈴木 拓也[2]、矢田 宙生[2]、大西 隆太郎[2]、山名 早人[3] (1. 早稲田大学基幹理工学部、2. 早稲田大学基幹理工学研究科、3. 早稲田大学理工学術院)

14:25 ~ 14:50
[T2-B-2-03] 解析処理単位の認可および監査に基づく柔軟なデータ流通処理基盤の提案

*松本 直樹[1]、川西 智也[2]、大平 健司[3]、小塚 真啓[4] (1. 京都大学、2. セーラネットワークス、3. 大阪大学、4. 岡山大学)

14:50 ~ 15:15
[T2-B-2-04] 階層構造・多重解像度による差分プライベートな深層人流生成

*高木 駿[1]、加藤 郁之[1]、曹 洋[2]、吉川 正俊[3] (1. 京都大学、2. 北海道大学、3. 大阪成蹊大学)

15:15 ~ 15:40
[T2-B-2-05] 顔画像プライバシ保護技術の比較検証

*馬 世昂[1]、曹 洋[1]、中村 篤祥[1] (1. 北海道大学大学院情報科学院)

m-不変性に基づいた連続的データ公開における安全性問題と小規模データ改ざんによる解決法

張 扶蘇† 上土井 陽子‡ 若林 真一‡

†広島市立大学大学院情報科学研究科 〒 731-3194 広島県広島市安佐南区大塚東三丁目 4 番 1 号
‡広島市立大学大学院情報科学研究科 〒 731-3194 広島県広島市安佐南区大塚東三丁目 4 番 1 号

E-mail: †zhangfusu@lcs.info.hiroshima-cu.ac.jp, ‡{yoko, wakaba}@hiroshima-cu.ac.jp

あらまし マイクロデータを静的なデータ表と見なしてプライバシー保護データ公開する問題に関しては，既存の研究で安全な一般化手法が既に多く開発されている．一方，一つのデータ表にタプルを挿入や削除することにより，動的に変化するデータ表の連続したプライバシー保護公開データ表を作成する手法の開発では，扱われているデータは動的に変化するデータであるため，有効な手法は少ない．有効な従来手法の一つは，データ表から連続的に作成された公開表のタプルに対して機密属性の値(機密値)の特定確率が 1/m 以下になるような m-不変性という性質を保つことを目的とした一般化手法である．ただし，この手法で処理したデータ表でもある条件で機密値の特定確率が 1/m を超えてしまう可能性が指摘されている．本研究ではこの安全性の問題を克服するように改良した情報提供方法を提案する．

キーワード 動的データ，データの一般化， m-不変性，プライバシー保護データ公開

1. はじめに

データ公開に対するマイクロデータのプライバシー保護における研究は社会から注目されている．いままで，静的データを対象とした研究では，k-匿名性と l-多様性を代表としたいくつかの個人情報を保護できる安全な匿名化保護手法が開発された．以降の研究でも静的データに対する一般化手法が改良されている[3], [7], [8]．近年，連続的なデータ公開に対する個人情報の匿名化によりプライバシー保護が研究され始めた．

連続的なデータを対象とする手法の開発では，データの動的な変化として主にタプルの挿入と削除に注目している．従来研究 [1] では m-不変性という動的データのプライバシーが保護できる性質が定義されており，m-不変性を保つことを目的とした一般化手法が提案され，手法の安全性を評価する方法として全射関数を利用した方法が提案されている．文献[1]で m-不変性を保つ一般化は動的に更新するデータの再公開においても，個人の機密情報が特定される確率を 1/m 以下に抑えられている安全な方法だと主張されている．

文献[1]以降の研究[2], [4], [5], [6], [9]により，動的データの一般化や公開方法に進展があったが，完全に安全な動的公開がまだ難しい．文献[5]により m-不変性の安全性が成立する前提として，攻撃側も強制的に m-不変性の基準を考えている上で攻撃を行うことだと仮定している．この前提を置かないと，個人の機密情報が特定される確率が 1/m 以上となる可能性がある．

本稿では，Xiao と Tao が提案した m-不変性の一般化手法[1]において個人の機密情報が特定される確率を 1/m 以下に抑えられない安全性の問題を克服するため，m-不変性を保持したデータを変形する問題を二つ定式化し，そのうち一つの問題の解決策を提案する．

2. 準備/定義

本節では，本研究で用いる用語の定義を紹介する．

公開者が管理している元データからなる元データ表を T とする．T の要素となるタプルは次のような三つの属性から構成される．

(i) T の主なキーとなる識別子(identifier)属性A^{id}.
(ii) d 個の準識別子(quasi-identifier)属性$A_1^{qi}, \ldots, A_d^{qi}$.
(iii) 機密値(sensitive value)属性A^s.

研究内容を説明するために，ある病院が患者の記録データを 1 時刻，6 か月ごとに公開することとする．各公開には公開時点からその 6 か月分の診断結果のみが含まれる．時刻 1 と時刻 2 のデータ表を表 1 と表 2 に示す．

表 1 時刻 1 の元データ T(1)

Name	Age	Zip.	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	18000	flu
David	23	25000	gastritis
Jane	37	33000	dyspepsia
Ken	40	35000	flu
Linda	43	26000	gastritis

表 2 時刻 2 の元データ T(2)

Name	Age	Zip.	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Mary	46	30000	flu
Ray	54	31000	dyspepsia

タプルの属性は時刻の更新とともに変化するマイクロデータに関する研究[4]もあったが，本研究では各タプルの中身は変化しないものを研究対象とする．例えば，Bob のタプルの Disease の値が dyspepsia 以外の値になることはないとする．

[定義 1] (偽造入りの一般化) 元データから公開表を作成する過程が匿名化または一般化と呼ばれる．表 T(j)が一般化された公開表 T*(j)は T(j)の属性に基づいて作成され，次の性質を持っている．
(1) 表 T*(j)は識別子A^{id}を非公開とし，精確な準識別子$A_1^{qi},\ldots,A_d^{qi}$を{$[A_1^{qi}, \overline{A_1^{qi}}]$,…, $[A_d^{qi},\overline{A_d^{qi}}]$}に曖昧化(一般化)して作成される．
(2) 一般化された公開表のタプル t*∈T^*(j)の機密値は各元タプル$t \in T(j)$ からなり，t*[A^s] = t[A^s].
(3) より安全的な公開表を求めるため，表 T*(j)に任意の数の偽造データを入れる場合がある.偽造データの機密値は元データ表のA^sから決める．偽造データの準識別子元データ表の準識別子の範囲{[$A_{1_min}^{qi}$, $A_{1_max}^{qi}$],…, [$A_{d_min}^{qi}$, $A_{d_max}^{qi}$]}からとる．

表 1 と表 2 の T(1)と T(2)から一般化されたデータ表を表 3, 表 4 に示した．一般化で使う m-不変性を保つ一般化手法は次の章で詳しく説明する．

表 3 2-不変性を満たすように一般化した
時刻 1 の公開表 T*(1)

G.ID	Age	Zip.	Disease
1	[21,22]	[12000,14000]	dyspepsia
1	[21,22]	[12000,14000]	bronchitis
2	[23,24]	[18000,25000]	flu
2	[23,24]	[18000,25000]	gastritis
3	[37,43]	[26000,35000]	dyspepsia
3	[37,43]	[26000,35000]	flu
3	[37,43]	[26000,35000]	gastritis

表 4 2-不変性を満たすように一般化した
時刻 2 の公開表 T*(2)

G.ID	Age	Zip.	Disease
1	[21,22]	[12000,14000]	dyspepsia
1	[21,22]	[12000,14000]	bronchitis
2	[23,25]	[21000,25000]	flu
2	[23,25]	[21000,25000]	gastritis
3	[37,46]	[26000,33000]	dyspepsia
3	[37,46]	[26000,33000]	flu
3	[37,46]	[26000,33000]	gastritis
4	[54,55]	[30000,31000]	dyspepsia
4	[54,55]	[30000,31000]	bronchitis

[定義 2] (QI グループ) 公開表 T*(j)に対して，QI グループは T*(j)内のタプルの部分集合である．表 T*(j)のタプルを複数の QI グループに分割して，すべての QI グループの和集合が表 T*(j)となるようにする．各 QI グループは特定の ID をもつ．タプル t∈T(j)に対し，t を一般化した t*が QI グループの構成要素となる．また，各 QI グループに属するすべてのタプル t*は同じ準識別子属性値の範囲によって表現される．

[定義 3] (特定リスク) タプル t∈T*に対して，公開表 T*から t の本当の機密値を正しく特定する確率 risk(t)を特定リスクと言う．表全体において最大特定リスク$risk_{max}$(t)と平均特定リスク$risk_{adv}$(t)はよく使う二つの一般化手法の評価基準である．
例えば，敵対者の視点から，Bob のタプル <Bob,21,12000>を把握していると仮定すると，表 3 で <21,12000>の条件から Bob の Disease を正しく特定するリスク risk(Bob)=1/2 である．

[定義 4] (シグネチャ) 時刻 1 から n まで，任意の時刻 j に対して，一般化されたデータ表 T*(j)における QI グループの機密値の集合を，その QI グループのシグネチャと定義する．
表 3 における，QI グループ ID が 1 のグループのシグネチャは{dyspepsia, bronchitis}であり，ID が 2 のグループのシグネチャは{flu, gastritis}である．

3. 従来研究

従来研究[1]では，時刻 n まですべての公開表が m-不変性という性質を満たせば，連続的なデータ表をプライバシー保護した上で公開可能であると示されている．

本節では, 3.1 節で k-匿名性と l-多様性の性質を示し, この二つの性質をまとめに利用する m-不変性の定義を 3.2 節で示す. n 番目の時刻のデータ表 T(n)から公開表 T*(n)を作成する m-不変性を満たすための一般化手法の詳しい説明と安全性問題も 3.2 節で紹介する.

3.1 k-匿名性と l-多様性

[定義 5] (k-匿名性) k-匿名性は, データセット内のすべてのタプルが少なくとも k-1 個の他のタプルと区別されないという性質である.

k-匿名性の一般化では, 個人を特定する可能性を減らすため, データセット内のタプルが少なくとも k 人のグループに所属するように, 識別子の非公開と準識別子の曖昧化を行う.

データの一部を非公開, かつ曖昧化にすることで k-匿名性の保持が達成されるため,データの有用性が低下し, 情報の損失が必ず発生する.

[定義 6] (l-多様性) l-多様性は, k-匿名性を達成した非公開にされたデータの QI グループのシグネチャが, あまりにも類似している場合に起こる問題を軽減するために定義された性質である.

l-多様性の一般化では, QI グループ内のシグネチャの多様性を維持することを目的とし, l-匿名性を達成した非公開データを, なおかつ各 QI グループのシグネチャに 1 種類以上の多様性を持たせるようにする.

3.2 m-不変性

k-匿名性と l-多様性を保持するよう公開データを作成する方法は静的なデータに対して, 十分に使える. 一方, 動的データでは, 時間の経過とともに, 新データの挿入と旧データの削除により, データ表が更新される. このように動的に変化する公開表の場合, k-匿名性と l-多様性を満たすように一般化しても, データのシグネチャの変化により個人情報が特定される可能性が残っている. 文献 [1] では公開データの安全性を高めるため, この二つの手法を組み合わせ, 新しい一般化原理 m-不変性を開発した.

3.2.1 節ではまず m-不変性の性質を紹介する. その次の 3.2.2 節では, m-不変性を満たす一般化手法を簡単に説明する. 最後の 3.2.3 節で, 本研究の研究課題である m-不変性を満たす公開表における安全性問題を紹介する.

3.2.1 m-ユニークと m-不変性

[定義 7] (m-ユニーク) 時刻 j のときの公開表 T*(j)において, 以下の性質を満たすときに, 公開表が m-ユニークであると定義する.

(1) T*(j)の各 QI グループが最小でも m 個のタプルを含む.

(2) 各 QI グループに存在するタプルがすべて異なる機密値を持っている.

いわゆる公開表が m-ユニークを満たすと, m-匿名性と m-多様性が同時に満たされる.

[定義 8] (m-不変性) 時刻 1 から n まで, 時刻 1, 2, ..., n におけるすべての公開表 T*(1), T*(2)...T*(n)において, 以下の性質を満たすときに, 公開表 T*(1), T*(2)...T*(n)が m-不変性を満たすと定義する.

(1) 公開した n 個の公開表 T*(1), T*(2)...T*(n)がすべて m-ユニークである.

(2) 時刻[x, y] ($1 \le x \le y \le n$)に存在する任意のタプル t(i)($x \le i \le y$)において, 各時刻での公開表での対応するタプル t*(i)のシグネチャが変わらない.

3.2.2 m-不変性を満たす一般化手法

時刻 n の元データ表の T(n) と T(n-1)と元データ表 T(n)に対する公開表 T*(n-1)が入力として与えられたときに, m-不変性を満たす公開表 T*(n)を作成できる.

一般化の性質として, m-エリジブルが定義される.

[定義 9] (m-エリジブル) 元データ表 T において, T にあるすべてのタプルの中, 最も多い機密値A^sの要素数の割合が 1/m 以下の場合, このデータ表 T を m-エリジブルを満たすと定義する. m を式で表すと:

$$m = \text{最も多い機密値} A^s_{max} / \text{全タプル数}$$

元データ表 T が m-エリジブルであるときのみ, m-不変性を満たす公開表 T*が作成できる.

T(n)と T(n-1)から, 挿入した新データ$S_{n_}$, 削除されていない旧データ$S_{n\cap}$をまとめ, $S_{n_}$の m-エリジブルを計算する. 特に, n=1 に対して, $S_{n_}$=T(1), $S_{n\cap} = \emptyset$が成り立つ. $S_{n_}$と$S_{n\cap}$を使い以下の 4 つのステップにより T*(n)を作成する.

[Step 1] $S_{n\cap}$からシグネチャごとにタプルをバケットBUCに作成して分配する.

[Step 2] $S_{n_}$から m-エリジブルを維持したままに, タプルをBUCの空いているところへ最大限に埋め込む.

(n=1 のとき, $S_{n\cap} = \emptyset$のため, Step2 から始まり, m-エリジブルを維持したままタプルをグループごとに取り出し, BUC を作り, 作った BUC に埋め込む)

[Step 3] Step2 で取り残された$S_{n_}$のタプルをBUCに割り当てる．最後空いているところに偽造データを入れる．

[Step 4] 多様性を保つため，一般化したグループの準識別子範囲の最小範囲も設定する．BUCから準識別子の数値を計算し，情報損失を最小限するようにタプルを取り出し，QI グループに分配する．QI グループの準識別子範囲が最小範囲未満の場合では，最小範囲に設定する．

Step3 から，一般化されたデータ表にある偽造データの数を決め，Step4 から QI グループの準識別子の最小範囲を決める．

毎回新時刻 n に更新する時，元データ表の T(n)と T(n-1)と 1 時刻前の公開表 T*(n-1)を入力として公開表 T*(n)を作成できるため，m-不変性は 1 回ずつのオンラインな公開では実用性が高い．

3.2.3 m-不変性を満たした公開表の安全性問題

文献[1]では，m-不変性という性質を定義し，m-不変性を満たす公開表では，時刻が変わっても同じタプルのシグネチャが変わらないため，機密値の特定リスクが 1/m 以下になると指摘された．

しかし，この結論は間違っている．

病院が時刻 1 と 2 に対して，元データを表 5 と表 6 にして，それらを m-不変性の一般化より一般化した公開表が表 7 と表 8 である．

表 5 時刻 1 の小さな元データ T(1)

Name	Age	Zip.	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	25000	dyspepsia
David	23	18000	gastritis

表 6 時刻 2 の小さな元データ T(2)

Name	Age	Zip.	Disease
Bob	21	12000	dyspepsia
David	23	18000	gastritis
Emily	21	12000	dyspepsia

表 7 時刻 1 の 2-不変性を満たした小さな公開表 T*(1)

G.ID	Age	Zip.	Disease
1	[21,22]	[12000,14000]	dyspepsia
1	[21,22]	[12000,14000]	bronchitis
2	[23,24]	[18000,25000]	dyspepsia
2	[23,24]	[18000,25000]	gastritis

表 8 時刻 2 の 2-不変性を満たした小さな公開表 T*(2)

G.ID	Age	Zip.	Disease
1	[21,22]	[12000,14000]	dyspepsia
1	[21,22]	[12000,14000]	bronchitis
2	[21,23]	[12000,18000]	dyspepsia
2	[21,23]	[12000,18000]	gastritis

敵対者の視点から，タプル<Bob, 21, 12000, [1,2]>の機密値を特定したいとき，<21, 12000>より T*(1)の QI グループ 2 を排除でき，さらにタプル Bob が時刻[1,2]の両方に存在することから，T*(1)の QI グループ 1 に存在しない gastritis を排除できる．よって，T*(2)の残った三行に，dyspepsia の可能性が 2/3 で 1/2 を超えてしまい，bronchitis の可能性が 1/3 となっている．

この問題を起こす根本的な原因として，一般化した複数の QI グループでシグネチャと準識別子の範囲の両方が重なっていることである．シグネチャに同じ機密値が入っている上，準識別子の範囲も重なると，準識別子属性の範囲が重なることで特定リスクが 1/m 以上になってしまう可能性がある．

4. 提案手法

本節では，まず提案手法で利用する情報損失の評価基準を 4.1 節で紹介する．その次の 4.2 節では，前節の最後に紹介した安全性問題を定式化し，定式化した問題を解決する提案手法を紹介する．

4.1 情報損失の評価

データの一般化は，プライバシーを保護する重要な手法であるが，同時に精度の低下や情報の欠落など情報の損失をもたらす可能性がある．定義 1 で紹介した通り，一般化は識別度の最も高い識別子を非公開し，その次に識別度の高い準識別子を曖昧化して一般化を行うため，詳細な情報が失われることで，分析や推論の精度や機械学習モデルの性能が低下することが必ずある．情報損失をできるだけ抑えるため，m-不変性の一般化では，Step4 で準識別子の数値を計算し，情報の低下を抑制する．

ただし，アルゴリズム上で情報損失を抑えるのは不十分なため，本研究では数値データに対する情報損失の評価基準として，情報損失(元データから改ざんした幅)の正規評価値を提案した．

提案した正規評価値では，数値データの準識別子が複数ある場合に，全体的に情報損失が小さいことがより実用性に影響がないという前提で，準識別子の全体的な情報損失の最小化を目的とする．

全体的な評価をするために，提案した正規評価値の計算ではまず，準識別子 x に対し，以下の式を通じて，正規値$\bar{x}$を求める．

$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

表 1 から求められる正規値は表 9 に示した．

表 9 時刻 1 のデータ準識別子を正規化した表

Name	N.Age	N.Zip.
Bob	0	0
Alice	0.045455	0.086956522
Andy	0.136364	0.260869565
David	0.090909	0.565217391
Jane	0.727273	0.913043478
Ken	0.863636	1
Linda	1	0.608695652

一般化のグループ分けのとき，分けた QI グループの準識別子範囲の正規値での間隔の値をすべて乗算し(二つ準識別子の場合は面積で，三つの場合は体積と考えることができる)，乗算の値をそのグループの正規評価値とする．

グループの正規評価値をそのグループのタプル数で割り，割った値をそのグループ内のタプルの正規評価値とする．また，全グループの正規評価値の和で公開表全体の情報損失を評価する．

情報損失が最大の場合は，一つのグループに準識別子の最小値と最大値が同時に存在し，正規値の差が 1 となり，タプルの正規評価値が 1/n となり，グループの正規評価値が 1 になる．全体的な正規評価値が 1 を超えてしまう．そのため，評価値が 0 に近ければ近いほど，情報損失が小さいと判断できる．

表 3，表 8 で計算した正規評価値の表を表 10，表 11 に示した．

表 10 時刻 1 の一般化された タプルの正規評価値の表

G.ID	N.Age	N.Zip.	Area
1	[0,0.045]	[0,0.087]	0.004
2	[0.136,0.09]	[0.261,0.565]	0.014
3	[0.727,1]	[0.609,1]	0.107

全体的な正規評価値が$0.004 + 0.014 + 0.107 = 0.125$となる．

表 11 時刻 2 の一般化された タプルの正規評価値の表

G.ID	N.Age	N.Zip.	Area
1	[0,0.5]	[0,0.333]	0.167
2	[0,1]	[0,1]	1

全体的な正規評価値が$0.167 + 1 = 1.167$となる．

4.2 問題の定式化

3.2.3 節に説明した通り，問題の発生原因は一般化した複数の QI グループのシグネチャと準識別子の範囲の両方が重なっていることである．そのため，解決策としてシグネチャと準識別子の範囲の両方が重なったグループのペアを検出し，それをシグネチャと準識別子のどちらかが重ならないようにデータ改ざんをすれば，安全性を確保した上で公開することができる．

しかしグループのシグネチャはグループ内のタプルの機密値からなる集合のため，本研究では，準識別子の範囲だけを対象にし，問題を解決するために改ざんすることを考える．

シグネチャと準識別子範囲が同時に重なる問題を解消するため，本研究では 4.2.1 節と 4.2.2 節で解消対象となる問題を二つの方向で定式化してみた．

4.2.1 データの一貫性を重視する問題の定式化

機械学習や統計分析など，公開データの本来の使用目的に対する実用性への影響を最小化するため，既に改ざんなしで公開されたタプルや過去一回改ざんしたタプルに対して，再び改ざんをしないという一貫性の制約条件を付け，改ざんタプル数の最小化を目的関数として定式化した問題を図 1 に示した．

m-不変性を保持するために先の手法では一回ずつの公開に適用できるため，毎時刻に一般化をしたデータ表を入力とする．一貫性の制約条件を満たすため，毎回改ざんしたタプルを記録し，次回の入力として使う．さらにデータの実用性の影響を小さくするため，公開表の追加情報として，改ざんタプルの有無情報も出力し公開する．

入力 : ① 識別子がある m-不変性を満たす一般化された最新時刻のデータ表 ② 過去に改ざんしたことがあるタプルをまとめたデータ表
出力 : ① シグネチャが重なっても準識別子範囲が重ならない最新時刻の公開表 ② 公開表に改ざんタプルの有無情報
制約条件 : ① 一つのデータに対する改ざんを最小にするため, 既に改ざんなしで公開されたタプルや過去に一回改ざんしたタプルを再び改ざんをしない(一貫性) ② シグネチャが機密値からなるため, 準識別子だけに対する改ざん ③ 逆に特定リスクを上げないため, m-不変性を満たしたまま改ざん
目的関数 : 改ざんタプル数の最小化

図 1 定式化した問題 1

4.2.2 公開表の平均影響を重視する問題の定式化

問題 1 では, 改ざんタプルに一貫性の制約条件を要求し, 改ざん数の最小化を目標としたため, 一つのデータの改ざん幅が大きくなることがある. さらに, 一回ごとの公開に対し, 毎回すべての時系列でまとめたデータ表を使わなければならないという問題もある.

改ざん数より改ざん幅の最小化がデータ解析に影響が小さいという視点から, 制約条件の一貫性を削除し, 問題を考え直し, さらに制約条件を緩和した後, 公開表だけでも重ならないように一般化できることに気づき, 図 2 のように図 1 から簡単化された問題を作成した. 本研究において, まず図 2 の問題を解く手法を提案してから, 図 1 の問題を解くような手法を考察する.

入力 : ① m-不変性を満たす一般化されたデータ表 ② データ表の情報損失の正規評価値の表
出力 : ① シグネチャが重なっても準識別子範囲が重ならない公開表 ② 公開表に改ざんタプルの有無情報
制約条件 : ① シグネチャが機密値からなるため, 準識別子だけに対する改ざん ② 逆に特定リスクが上がらないため, m-不変性を満たしているまま改ざん
目的関数 : 各改ざんタプルの平均改ざん幅の最小化

図 2 定式化した問題 2

4.3 問題に対する提案手法

一般化されたデータ表から安全なデータ表を作成する提案手法の流れは図 3 を示す, 図に示すよう, 提案手法は 8 つの Step からなる.

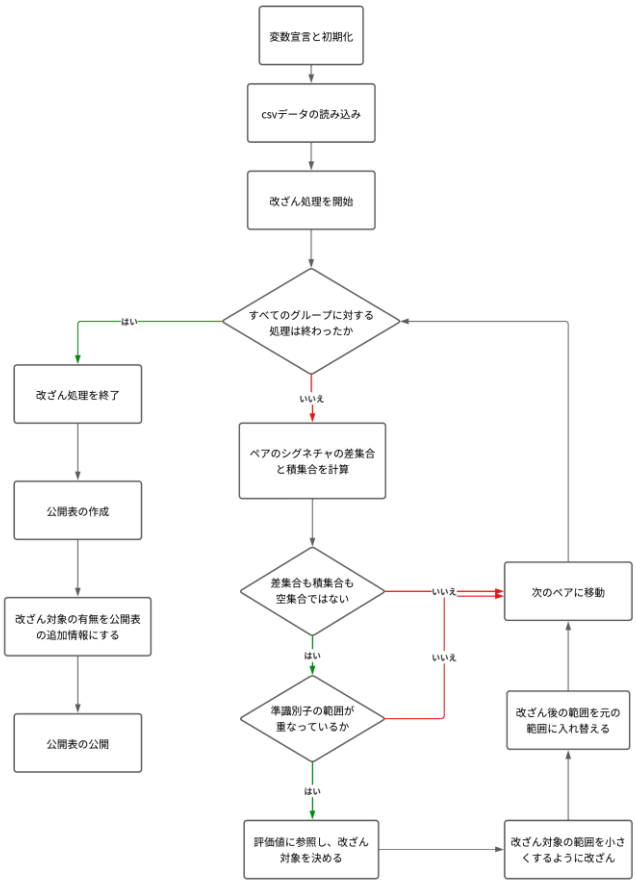


図 3 提案手法の流れ

[Step 1] 一般化されたデータ表において, すべての QI グループのシグネチャを抽出する.

[Step 2] 重なるシグネチャの判定処理と改ざん処理が終わったかを判定し, 終わってない場合は次の Step に進む. 終わった場合, 重なっているシグネチャが存在するかどうかによって, 改ざんデータの有無とデータ表を公開表として公開できる.

[Step 3] (初回はグループ id1 と 2 のペアを計算対象) 計算対象の QI グループのペアにおいて, シグネチャの差集合と積集合を計算する.

[Step 4] 差集合が空集合の場合, 同じシグネチャのため, 次のペアを計算対象とし, Step2 に戻る. そうでない場合, 積集合が空集合ではないことから, シグネチャが重なることが判定できる.

[Step 5] シグネチャが重なることが検出された QI グループのペアに対し，さらに準識別子の範囲が重なっているかどうかを判定する．重なってない場合は次のペアを計算対象とし，Step2 に戻る．

[Step 6] 重なっているときに，もう一つ入力データの「データ表の各グループの情報損失の評価値」を参照し，情報損失の大きいグループを改ざん対象とする．

[Step 7] 改ざん対象の範囲を小さくするように(「データ表の各グループの情報損失の評価値」を参照に計算)，公開データに最小限の改ざんを行う．

[Step 8] 改ざん後の QI グループの準識別子範囲をその QI グループの元の準識別子範囲と入れ替え，次のペアを計算対象とし，Step2 に戻る．

全ペアに対し，同時に重なったシグネチャと準識別子範囲の QI グループがなくなり次第，検出を終了し，作成された公開表において，全 QI グループの特定リスクが 1/m 以下になる．

表 8 の公開表 T*(2)に提案手法を適用するとき，T*(2)のシグネチャ{<dyspepsia, bronchitis>, <dyspepsia, gastritis>}を抽出し，差集合が空ではなく，さらに積集合が<dyspepsia>だと計算できる．シグネチャの計算後に，準識別子の範囲<[12000,14000], [12000,18000]>で重なっていると判定でき，表 11 を参照にグループ 2 を改ざん対象とする．続いては準識別子の最小範囲を<0.5, 1000>におくと，改ざん準識別子の範囲の選択が<[22.5,23], [12000,18000]>と<[21,23], [15000,18000]>の二択となる．改ざん対象の範囲を小さくするように，両者の正規評価値を計算し，前者の正規評価値が 0.125 で，後者の 0.25 より小さいため，準識別子の範囲を<[22.5,23], [12000,18000]>に改ざんし，元の表の準識別子範囲と入れ替える．最後に追加情報として，表 12 に示したように，「グループ 2 にデータ改ざんが存在する」を一緒に公開する．

表 12 時刻 2 の改ざんされた公開データ表

G.ID	Age	Zip.	Disease
1	[21,22]	[12000,14000]	dyspepsia
1	[21,22]	[12000,14000]	bronchitis
2	[22.5,23]	[12000,18000]	dyspepsia
2	[22.5,23]	[12000,18000]	gastritis
グループ 2 にデータ改ざんが存在する			

5. 実験

本節では，実験を通じて提案手法の動作を確認する．一般化されたデータ表の生成プログラムと提案手法を計算機上(AMD Ryzen 5 3600 6-Core Processor 3.60 GHz, 16.0 GB)で Python により実装した．

グループの数及び属性数を変化させた場合のデータ改ざん幅を比較するために，準識別子の範囲を{[10,70],[11000,69000]}で取り，一つのグループに含むタプルを 2，機密値の種類を 4 としてグループ生成の規則を設定し，グループの数が 5, 10, 20 個の m-ユニークな一般化データ表(提案手法が一回の公開に一回行うので、時刻を無視できる)をランダムに生成した．元データ表と提案手法で処理した正規評価値とその正規評価値の差(改ざん幅)を図 4 に示す．

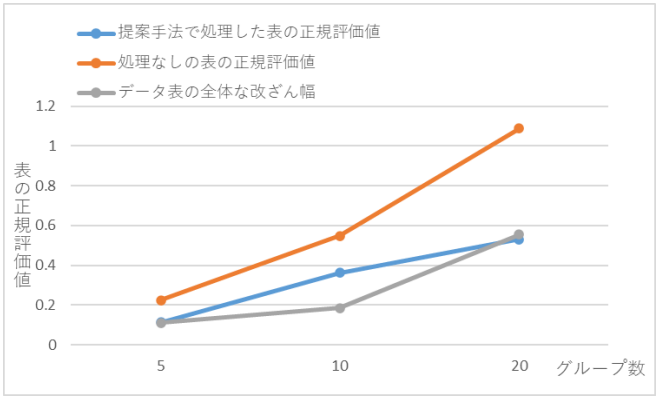


図 4 正規評価値と評価値の改ざん幅

各グループ数のグループをもつデータで領域が重なるグループのペア数と提案手法で行なった改ざん回数の関係を表 13 に示す．

表 13 重ならなくなるまでの改ざん回数表

グループ数	領域が重なったグループペアの数	改ざん回数
5	3	3
10	5	5
20	21	23

図 4 より，グループ数が異なった各データ表における結果としてグループ数により改ざんの幅に差があるように見える．グループ数が 5, 10 の表に対して提案手法を行った改ざん幅にあまり差がないが，グループ数の増加により，グループ数が 20 の表では改ざん幅がかなり大きくなった．表 13 からも分かるように，グループ数が 20 の表では，重なったグループのペア数が非常に多くなり，それを解消するためにより多い回数の改ざんを行った．生成データの生成ルールの不十分も原因である．今後，より規模の大きいデータや実データを使って再度の実験を行い，さらに提案手法を適用したことによる影響を解析する必要がある．

6. まとめ

本稿では，m-不変性の一般化手法により作成された連続的な公開表に存在する安全性問題を分析し，二つ方向から問題定式化し，解決法を提案した．

今後の課題として，定式化した問題 1 に対する手法の開発と本研究の手法に対する大規模なデータを用いた実験的性能評価とその考察がある．また，特定リスクを一般化手法の途中で計算し，特定リスクを低下させるようにすることがある．一緒に公開する追加情報に関して，追加する内容が詳しいほどデータの実用性への影響が小さいが，特定リスクも大きくなるため，複数の追加情報の公開方法を比較し，影響と特定リスクのバランスを考察することも課題である．敵対者の視点から，事前に k-匿名性と l-多様性の知識，もしくは m-不変性の知識を把握しているなら，特定リスクにどのように影響するかについて考察することも考えられる．

参 考 文 献

[1] X. Xiao, Y. Tao, “m-invariance: toward privacy preserving republication of dynamic datasets,” Proc. of SIGMOD’07, pp.689-700 (2007).

[2] Y. He, S. Barman, J. F.Naughto, “Preventing Equivalence Attacks in Updated, Anonymized Data,” Proc. 27th ICDE, pp.529-540 (2011).

[3] J. Domingo-Ferrer, A. Martínez-Ballesté, J. Maria Mateo-Sanz, F. Sebé, “Efficient multivariate data-oriented microaggregation”, The VLDB Journal, 15:355–369 (2006).

[4] K. Han, H. Nishi, “Proposal and implementation of k-anonymization method for data insertion and deletion,” Proc. 32nd ISIE (2023).

[5] 坂田奈々子, 上土井陽子, 村上頼太, 若林真一, “動的データセットにおけるプライバシ保護の厳密な安全性評価と妥当な安全性評価について”, DEIM Forum 2018, F7-5 (2018).

[6] 柚木塁，上土井陽子，若林真一，“動的データテーブルの連続的匿名化の安全性について”，DEIM Forum 2014, E5-3 (2014).

[7] 須崎亮, 上土井陽子, 若林真一, “Microaggregation を利用した k-匿名化手法の高速化”, DEIM Forum 2013, F8-3 (2013)

[8] 村本俊祐，上土井陽子，若林真一, “データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム”, DBSJ Letters, Vol.6, No.1(2007).

[9] 村上頼太，上土井陽子，若林真一, “線形計画法を利用したプライバシ保護再公開の安全性評価”, IPSJ SIG Technical Report (2018).

準同型暗号を活用したプライバシ保護型ニュース推薦手法

中野　瑛心†　　鈴木　拓也††　　矢田　宙生††　　大西隆太郎††　　山名　早人†††

† 早稲田大学基幹理工学部　〒 169-8555 東京都新宿区大久保 3-4-1
†† 早稲田大学院基幹理工学研究科　〒 169-8555 東京都新宿区大久保 3-4-1
††† 早稲田大学理工学術院　〒 169-8555 東京都新宿区大久保 3-4-1
E-mail: †eishin@fuji.waseda.jp, ††{t-suzuki,yadayuki,ryutaro634,yamana}@yama.info.waseda.ac.jp

あらまし　ニュース推薦システムは，ユーザの嗜好情報とニュースの特徴情報との類似性に基づいて記事を推薦する．ここで，サーバがユーザの嗜好情報と推薦結果を閲覧可能であるため，ユーザへのプライバシ侵害が懸念される．この懸念への解決策の一つである準同型暗号は，レイテンシの長さが課題である．本研究では，高い推薦精度を示す PLM-NR を対象に，プライバシを保護しながら低レイテンシでニュース推薦を行うプロトコルを提案する．具体的には，嗜好情報計算の入力となるユーザが過去に閲覧したニュースの特徴情報と，推薦結果のサーバからの獲得を準同型暗号上で行い，サーバがユーザの嗜好情報と推薦結果を平文上で保持することを防ぐ．さらに，サーバが予めニュースの特徴情報を，ユーザ自身がユーザの嗜好情報を，それぞれ平文上で算出することでレイテンシを抑える．

キーワード　準同型暗号，推薦システム，ニュース推薦，Private Information Retrieval，プライバシ保護

1　はじめに

近年，インターネットの普及によって，オンラインニュースプラットフォームの需要が高まっている．オンラインニュースプラットフォームの便利さは，多くのユーザに新聞からの移行を促し [8]，国内でも Yahoo!ニュース [1]や SmartNews[2]などが主要オンラインニュースプラットフォームとしての地位を獲得している．そして，これらのプラットフォームにおいて，ニュース推薦システムは重要な役割を担っている．

ニュース推薦システムは，ユーザの嗜好情報に基づいてユーザが関心のあるニュースを推薦する．この分野の研究は盛んに行われており，特に深層学習を応用した手法は推薦精度が高いことで知られている [17].

深層学習に基づくニュース推薦モデルは，News Encoder，User Encoder，Click Predictor の 3 つの要素から構成される．図 1 にモデルの概要を示す．News Encoder は，ニュース記事のタイトルや本文からニュース特性を表すベクトル（ニュースベクトル）を生成し，User Encoder はユーザの閲覧履歴からユーザの好みを表すベクトル（ユーザベクトル）を生成する．その後，Click Predictor がこれらのベクトル間の類似度を推薦スコアとして算出し，推薦すべきニュース記事を決定する．

さらに，言語モデルを統合したニュース推薦モデルが注目されている．Wu ら [18] によって提案された，事前学習済み言語モデルを活用したニュース推薦手法（Pre-trained Language Model empowered News Recommendation：PLM-NR）は，オンライン及びオフラインの実験で高い推薦精度を実現し，実用化が進んでいる．

しかし，サービス提供者がユーザベクトルや推薦結果を閲覧

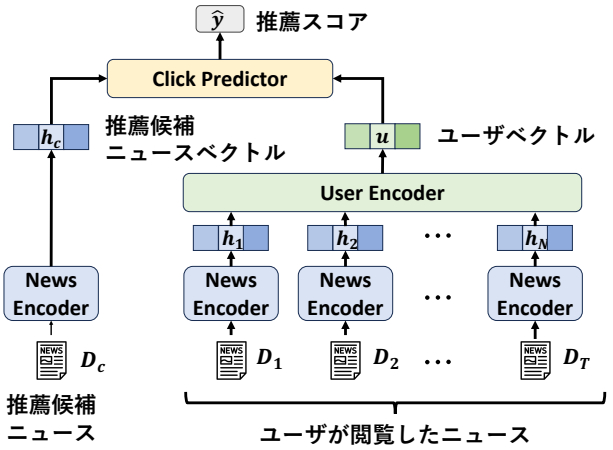


図 1　ニュース推薦の概要図（[18] の Fig. 1 を元に作成）．h_i はニュース記事 D_i に対するニュースベクトルを示す．

することによりプライバシ侵害が生じる．ここで，ユーザベクトルや推薦結果のプライバシを保護する手法の一つとして，データを暗号化したまま復号することなく演算できる準同型暗号が挙げられる．しかし，準同型暗号は，演算レイテンシが平文実行よりも $10\sim10^7$ 倍長い [9] という欠点がある．そのため，各 Encoder 処理を準同型暗号上で計算すると，レイテンシが実用に適さないほど長いという問題がある．

この問題を解決するため，本稿では PLM-NR による推薦処理に対して，準同型暗号の適用方法を工夫することで，ユーザのプライバシを保護しながら，低レイテンシで推薦を行う新しいプロトコルを提案する．具体的には，まず，公開情報となっているニュース記事のそれぞれについて，特徴情報をサーバ（サービス提供者）が予め平文上で計算し保持する．その上で，推薦処理を以下の 3 つのフェーズに分割して実行する．

フェーズ 1.　ニュースベクトル獲得フェーズ：準同型暗号

1：https://news.yahoo.co.jp/
2：https://www.smartnews.com/ja

上でのプライベート情報検索（Private Information Retrieval:PIR）によってユーザが過去に閲覧したニュースの特徴情報をサーバから取得する．

フェーズ 2．ユーザベクトル算出フェーズ：過去に閲覧したニュースの特徴情報を元に，ユーザ自身が平文上でユーザベクトルを算出する．

フェーズ 3．推薦フェーズ：軽量な推薦スコア計算をサーバが準同型暗号上で行う．

つまり，準同型暗号で計算するには重い処理である各 Encoder 処理を平文上で行い，News Encoder よりも軽量な PIR を用いることで，低レイテンシを実現する．また，サーバがユーザベクトルと推薦結果を平文で保持しないため，ユーザのプライバシを保護できる．

本研究の貢献は，PLM-NR の高精度なニュース推薦性能を維持しつつ，プライバシを保護する新たなプロトコルを設計したことである．そして，提案プロトコルが現実的かつ低レイテンシでニュース推薦を行うことができることを示す．

本稿の構成は以下の通りである．2 節で背景知識を説明し，3 節でニュース推薦分野におけるプライバシ保護型推薦分野の関連研究をまとめる．そして，4 節で提案手法について説明する．その後，5 節で実験方法について述べ，得られた実験結果と考察を 6 節に示す．最後に，7 節で結論及び今後の課題を示す．

2 背景知識

2.1 準同型暗号

準同型暗号は，データを暗号化したまま加算（準同型加算）や乗算（準同型乗算）を評価できる暗号技術である．本研究で使用する Cheon-Kim-Kim-Song（CKKS）方式 [4] は，誤差を許容する代わりに固定小数点数を扱うことができる準同型暗号方式であり，機械学習への適用が行われている [15]．CKKS 方式は，複数の値（ベクトル）を一つの平文や暗号文にまとめる手法であるパッキング [11] に対応しており，レイテンシやスループットを改善できる．平文や暗号文は n_{slots} 個のスロットを持ち，各スロットにベクトルの各要素を格納する．

CKKS 方式には，ローテーション（rotation）と呼ばれる演算が存在する．ローテーションは，一つの暗号文のスロットに格納された要素を循環的に回転させる．例えば，ある暗号文に対して正方向に 1 要素分ローテーションさせる操作は，その暗号文の末尾のスロットの要素を先頭のスロットへ移動させ，他の全てのスロットの要素を 1 つ後ろのスロットへ移動させる．本稿では，$rotate(x, j)$ は，暗号文 x に対して正方向に j 要素分ローテーションさせる演算を表す．

2.2 Neural News Recommendation with Multi-Head Self-Attention に用いられる Attention 機構

Neural News Recommendation with Multi-Head Self-Attention (NRMS) [16] は，News Encoder 及び User Encoder で Multi-head Self-attention と Additive attention が使用されている特徴を持つニュース推薦モデルである．本稿で使用する PLM-NR [18] の一つである NRMS-BERT [18] は，これらの Attention 機構と言語モデルを統合することによって，深い言語理解と複雑な嗜好情報の計算を可能にし，推薦精度の向上を実現した．ここでは，Multi-head Self-attention と Additive attention について説明する．

2.2.1 Multi-head Self-attention

Multi-head Self-attention [16] は，テキストデータにおける単語間や文節間の関係性を捉えることにより，深い文脈の理解を獲得する役割を持つ．この機構は，離れた単語間の関係性や文書内の異なる位置にある情報の統合に特に有効である．Multi-head Self-attention では，テキストデータにおける各単語の埋め込みベクトルに対して複数の Attention 機構（head）を適用し，それぞれが並列して情報を処理する．

まず，k 番目の head において，単語数 M のテキストの i 番目と j 番目の単語間の相互的な文脈表現の相対的な重要度を表す重み $\alpha^k_{i,j}$ を式 1 で計算する．ただし，Q^w_k，K^w_k，V^w_k はそれぞれ Query，Key，Value のパラメータを表し，e_i は i 番目の単語の埋め込みベクトルを表す．

$$\alpha^k_{i,j} = \frac{\exp\left(e_i^T Q^w_k K^w_k e_j\right)}{\sum_{m=1}^{M} \exp\left(e_i^T Q^w_k K^w_k e_m\right)} \tag{1}$$

そして，i 番目と j 番目の単語間の Attention スコアは，式 2 で計算される．

$$h^w_{i,k} = V^w_k \left(\sum_{j=1}^{M} \alpha^k_{i,j} e_j \right) \tag{2}$$

最後に，i 番目の単語の Multi-head 表現 h^w_i は，3 に示すように，異なる head からの出力を結合（concat）することで得られる．

$$h^w_i = \mathrm{concat}(h^w_{i,1}, h^w_{i,2}, \ldots, h^w_{i,h}) \tag{3}$$

2.2.2 Additive attention

Additive attention [16] は，Multi-head Self-attention による出力に対して，各単語の重要度に着目し，テキストの全体の表現を得る．テキストデータ内の各単語の重要度は異なるため，より重要な単語を選択することで，より適切な情報を持ったニュース表現を獲得できる．テキストデータの i 番目の単語に対する attention の重み α^w_i の計算式を式 4 に示す．

$$\alpha^w_i = \frac{\exp(a^w_i)}{\sum_{j=1}^{M} \exp(a^w_j)} \tag{4}$$

ここで，a^w_i は式 5 で計算される．なお，V_w と v_w は射影パラメータであり，q_w はクエリベクトルである．

$$a^w_i = q^T_w \tanh(V_w \times h^w_i + v_w) \tag{5}$$

最終的な表現は，文脈的な単語表現の加重和として計算される．これを式 6 に示す．

$$r = \sum_{i=1}^{M} \alpha^w_i h^w_i \tag{6}$$

3 関連研究

3.1 プライバシ保護型協調フィルタリング推薦

協調フィルタリングとは，ユーザの嗜好情報を元にユーザ間の類似度を計算し，類似度の高いユーザが好むアイテムを推薦する手法である．具体的には，ユーザ間の類似度を分析することで，ユーザがまだ評価していないアイテムに対する予測評価を行う．協調フィルタリングは，EC サイトや映画視聴サービスなどで広く用いられている．

協調フィルタリングに対して準同型暗号を適用した手法は多く提案されている．2002 年に Canny [3] は，協調フィルタリングに乗法準同型暗号の 1 つである ElGamal 暗号を初めて適用した．映画に関する評価情報を集めた EachMovie のデータセットを用いて行われた評価実験では，平均絶対値誤差 (Mean Absolute Error:MAE) で 0.96 を記録し，準同型暗号を用いなかった場合と等しい結果であることを示した．

2010 年に Tada ら [12] は，加法準同型暗号の 1 つである Paillier 暗号を協調フィルタリングに適用した．Tada らの論文では，映画に関する評価情報を集めた MovieLens のデータセットを用いて実験を行い，平文上で MAE が 0.79 であったのに対して，準同型暗号上では 0.77 の MAE を記録し，準同型暗号上でも平文上での推薦と同程度の精度で推薦できることを示した．

2016 年に Badsha ら [10] は，これまで公開されてしまっていた類似度といった情報のプライバシも保護するプロトコルを提案した．提案手法では，ElGamal 暗号を協調フィルタリングに適用し，他のユーザの評価情報を予め準同型暗号上で集めた．さらに，その統計情報を信頼できるサーバに保持しておき，その情報を利用した推薦を行った．評価実験では MovieLens データセットを用いて，平文上での推薦結果と同等の推薦精度が得られることを示した．しかし，推薦時のレイテンシの高さが準同型暗号を協調フィルタリングに適用する上での課題であった．

この課題に対して，2023 年に Jumonji ら [7] は，OBGV-CF を提案した．Jumonji らは，全体のユーザ数 N とアイテム数 M に対して，推薦時の計算量を $O(N \times M)$ から $O(N)$ に削減することで推薦精度を保ちつつ，推薦時のレイテンシを抑えた．

以上に示した通り，協調フィルタリングと準同型暗号の組み合わせについての研究は活発に行われている．

3.2 Pre-trained Language Model empowered News Recommendation (PLM-NR)

2021 年に，推薦システムに言語モデルを統合した PLM-NR が提案された [18]．PLM-NR の一つである NRMS-BERT は, Neural News Recommendation Approach with Multi-head Self-attention（NRMS）と呼ばれるニュース推薦のモデル [16] に対して，言語モデルである BERT [6] を統合することで深い文脈理解を獲得し，オンライン実験とオフライン実験の両方で高い精度を示した [18]．NRMS-BERT は News Encoder と User Encoder，Click Predictor の 3 つの要素によって構成される．図 2 は，NRMS-BERT を構成する News Encoder と User Encoder，Click Predictor のそれぞれの概要図を示す．

3.3 関連研究のまとめ

準同型暗号と協調フィルタリングの組み合わせによるプライバシ保護型推薦の研究が活発に行われてきた．一方で，ニュース推薦の分野では，協調フィルタリングよりも，PLM-NR のようなコンテンツベースの推薦手法の方が，推薦精度が高いとされている [17]．これらの手法は，個々のニュースコンテンツを深く理解し，ユーザの嗜好に基づいたより適切な推薦を行うことができるためである．しかし，PLM-NR に準同型暗号を適用する研究は前例がなく，準同型暗号を直接適用した PLM-NR は，推薦レイテンシが長いという問題がある．

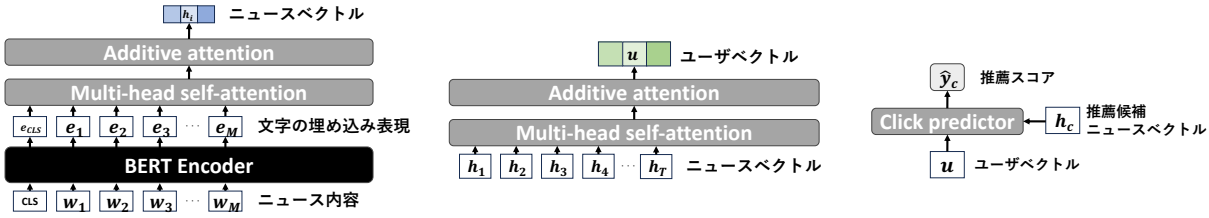
4 提案手法

本研究では，PLM-NR の一つである NRMS-BERT に対して準同型暗号を適用し，高い推薦精度維持しつつ，ユーザのプライバシを保護し，低レイテンシでの推薦を達成するニュース推薦プロトコルを提案する．準同型暗号を用いた処理はレイテンシが長いため，PLM-NR に準同型暗号を直接適用した場合，推薦レイテンシが長いことが問題である．この問題を解決するために，提案手法では，準同型暗号の適用方法を工夫する．まず，レイテンシが長い各 Encoder での処理を平文上で実行する．具体的には，News Encoder をサーバが，User Encoder を各ユーザが，それぞれ平文上で実行する．そして，News Encoder よりも軽量な PIR を準同型暗号を用いて実行することで，ユーザが過去に閲覧したニュースをサーバから隠したまま短いレイテンシでニュースベクトルを獲得する．また，ユーザベクトルや推薦結果をサーバから隠すために，ユーザベクトルと推薦候補ニュースベクトルとの類似度計算を準同型暗号上で行う．さらに，毎推薦時のレイテンシ削減を目的に，ニュースベクトルの獲得及びユーザベクトルの算出を推薦とは非同期的に行う．図 3 に提案手法のモデルを示す．提案手法は，ニュースベクトル獲得フェーズ（図 3(a)）と，ユーザベクトル算出フェーズ（図 3(b)），推薦フェーズ（図 3(c)）の 3 つのフェーズで構成される．

4.1 ニュースベクトル獲得フェーズ

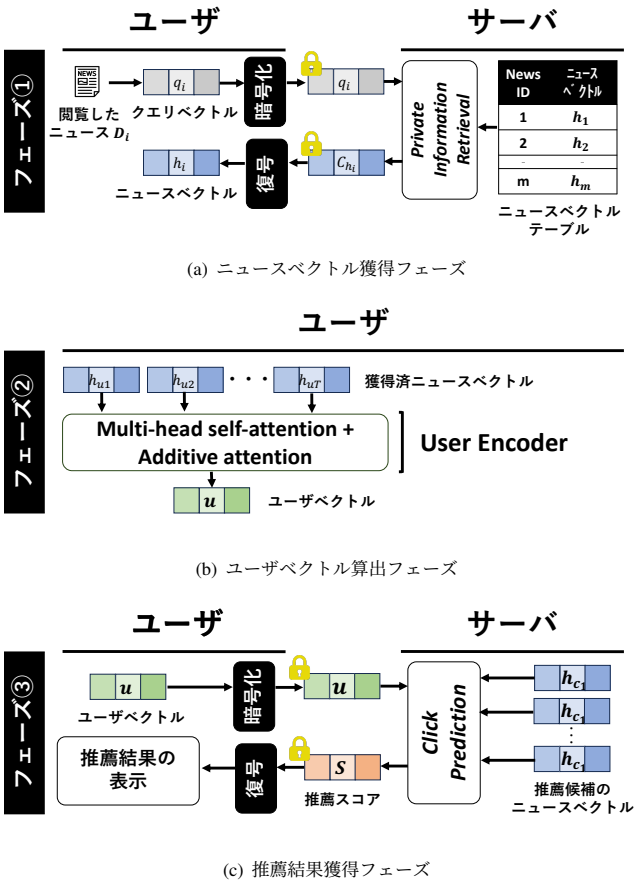
本フェーズの目的は，ユーザが過去に閲覧したニュースをサーバに知られることなく，ニュースベクトルを得ることである．そこで，準同型暗号上での PIR を用いる．PIR はクエリを秘匿しながら値を検索する技術である．したがって，PIR を用いることで，クエリに含まれるクリック情報を隠しながら，ユーザがニュースベクトルを得ることができる．

PIR を用いる理由は，ニュースベクトルの獲得レイテンシを削減するためである．具体的に説明する．まず，News Encoder は，平方根といった準同型暗号に適さない演算で構成されている．そのため，そのような演算を準同型暗号で実現可能な近似計算に置き換えた場合，レイテンシの増加や精度の劣化が生じる．そこで，ニュースベクトルそのものはプライベートな情報を含まないため，サーバが平文上で計算できることに着目する．準同型暗号においては，News Encoder よりも PIR の方が



(a) News Encoder の概要図　(b) User Encoder の概要図　(c) Click Predictor の概要図

図 2　NRMS-BERT の構成要素の概要図



(a) ニュースベクトル獲得フェーズ

(b) ユーザベクトル算出フェーズ

(c) 推薦結果獲得フェーズ

図 3　提案手法の概要図

軽量な処理であるため，サーバがニュースベクトルを保持することで，準同型暗号を用いる処理のレイテンシを短縮することができる．さらに，サーバはユーザからのクエリを受け取る前にニュースベクトルを計算できるため，ニュースベクトルの計算にかかるレイテンシを無視することができる．

図 3(a) に示すように，ニュースベクトル獲得フェーズでは，ユーザが過去に閲覧したニュース記事ごとに以下の処理を行い，ユーザベクトル獲得に必要なニュースベクトルを得る．

（1） PIR のためのクエリ発行：ユーザが暗号化したクエリを発行し，サーバに送信する．

（2） ニュースベクトルを含む暗号文の取得：サーバがニュースベクトルテーブルと暗号化済みクエリの演算を行い，結果をユーザに返却する．

（3） ニュースベクトルの復号：ユーザが結果を復号し，ニュースベクトルを得る．

4.1.1〜4.1.3 で，1 つのニュース記事の特徴情報を獲得するまでの流れを具体的に説明する．

4.1.1　PIR のためのクエリ発行

まず，以下の前提を置く．

- ニュース ID は 1 以上の整数で表される．
- ユーザは準同型暗号の秘密鍵 sk と公開鍵 pk のペアを保持する．
- m 個のニュース記事を対象とし，それぞれのニュース記事にニュース ID が公開情報として割り振られている．

また，$\boldsymbol{q}_j$ を，ニュース ID が j であるニュース記事を過去に閲覧したことを表すクエリとする．$\boldsymbol{q}_j$ は，m 次元のベクトルであり，j 番目の要素が 1，それ以外の要素は 0 である one-hot ベクトルである．

ここで，ユーザが過去に閲覧したニュース記事の ID を i とする．ユーザは，$\boldsymbol{q}_i$ を公開鍵 pk で暗号化し，得られた暗号文 $\mathrm{enc}(q_i)$ をサーバへ送信する．

4.1.2　ニュースベクトルを含む暗号文の算出

サーバは，ニュース ID をキーとして，ニュースベクトルを値に持つテーブルであるニュースベクトルテーブル V を保持する．ニュースベクトルテーブルは行列として表すことができ，i 行目にニュース ID が i のニュースベクトル $\boldsymbol{h}_i$ を保持しているとする．また，d をニュースベクトルの次元数として，ニュースベクトルテーブルを $m \times d$ の行列 V とする．そして，$V \times \boldsymbol{q}_i$（行列の各列と列ベクトルの要素ごとの積）を計算することで，獲得したいニュースベクトルの情報を i 行目に保持し，それ以外が全て 0 の行列 H_i を得ることができる．なお，H_i は列ごとに暗号化されている．以上の処理を図 4 に示す．

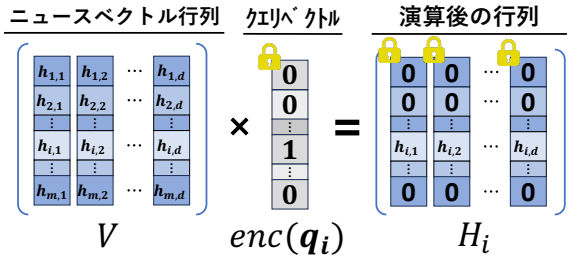


図 4　ニュースベクトルを含む暗号文配列の算出

なお，サーバが保持すべきニュースベクトルの個数 m が，暗号文のスロット数 n_{slots} よりも多いことが想定される．そこで，

$1 \leqq j \leqq n_{tables}$ について，サーバの保持する行列（テーブル）を n_{tables} 個の n_{slots} 行 m 列の行列 V'_j に分ける．そして，V'_j のそれぞれに対応するようにクエリ $\boldsymbol{q}_i$ を $\boldsymbol{q}'_j$ に分けて発行し，$V'_j \times \boldsymbol{q}'_j$ を足し合わせることで，ニュースベクトルを含む暗号文の行列を得る．複数のテーブルに対するクエリ発行によるニュースベクトルを含む暗号文列に対応する行列 H'_i の算出の流れを図 5 に示す．また，$H'_{i,j}$ は行列 H'_i の j 列目の暗号文とする．

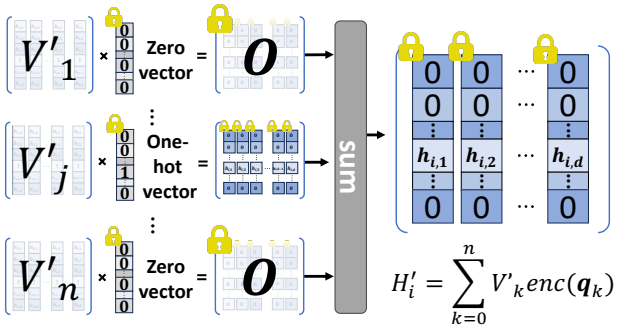


図 5　複数のテーブルに対するクエリ発行によるニュースベクトルを含む暗号文の算出（O はゼロ行列を表す）

最後に，式 7 及び図 6 に示す処理で H_i を目的のニュースベクトル $\boldsymbol{h}_i$ のみを含む暗号文 c_{h_i} に変換し，c_{h_i} をユーザに返却する．

$$c_{h_i} = \sum_{j=1}^{m} rotate(H'_{i,j}, j-1) \tag{7}$$

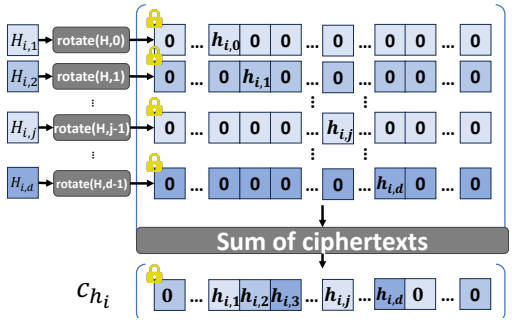


図 6　ニュースベクトルを含む暗号文配列の算出

4.1.3　ニュースベクトルの獲得

ユーザは，式 8 に示す方法で，サーバから受け取った暗号文を復号し，ニュースベクトルを平文として得る．

$$h_i = \mathrm{dec}(c_{h_i})[i : i+j] \tag{8}$$

ここで，$\mathrm{dec}(ct)$ は，ユーザの秘密鍵 sk を用いて暗号文 ct を復号する関数である．また，$\mathrm{dec}(ct)[i : i+j]$ は，$\mathrm{dec}(ct)$ の i 番目から $i+j$ 番目までの要素を取り出す操作である．ただし，$(i+j) > n_{slot}$ を満たすときは，$\mathrm{concat}\,(\mathrm{dec}(ct)[i : n_{slot}], \mathrm{dec}(ct)[1 : j-(n_{slot}-i)])$ を得る．

4.2　ユーザベクトル算出フェーズ

ユーザのプライバシを保護しつつレイテンシを短く抑えるために，準同型暗号を用いずに，ユーザベクトルの算出はユーザのデバイス上で行う．ユーザは，4.1 項のニュースベクトル獲得フェーズで獲得した，過去に閲覧したニュースのニュースベクトルを元に，ユーザベクトルを平文上で算出する．そのため，提案手法におけるユーザベクトルの算出方法は NRMS-BERT と同様である．具体的には，獲得したニュースベクトルを Multi-head Self-attention に入力し，その出力を Additive attention に入力することで，ユーザベクトル $\boldsymbol{u}$ を得る（図 3(b)）．なお，ユーザベクトル $\boldsymbol{u}$ の次元数は，ニュースベクトルと同様に d である．

4.3　推薦フェーズ

まず，プライバシを保護するために，ユーザは，算出したユーザベクトルを準同型暗号で暗号化してからサーバへ送信する．次に，式 9 及び図 7 に示すように，ユーザベクトルと各推薦候補ニュースのニュースベクトルとの内積をサーバが準同型暗号上で計算し，類似度（推薦スコア）を算出する．

$$\hat{y}_{c_i} = sim(\boldsymbol{u}, \boldsymbol{h}_{c_i}) = \boldsymbol{u}^{\mathsf{T}} \cdot \boldsymbol{h}_{c_i} \tag{9}$$

そして，各推薦候補との推薦スコアを格納している暗号文を 1 つの暗号文に統合し，ユーザに返却する．ただし，推薦候補ニュースの数が n_{slots} よりも多い場合は，ユーザに返却される暗号文は複数個となる．推薦フェーズにおけるサーバ上での処理の流れを図 7 に示す．

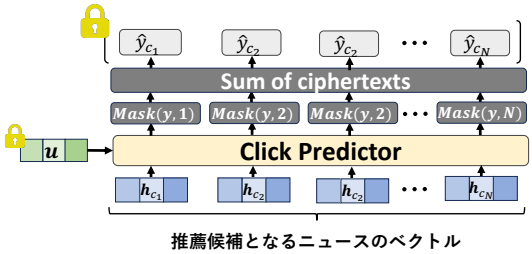


図 7　推薦候補ニュースベクトルとユーザベクトルの類似度の計算処理（$Mask(y,i)$ は，i 番目の要素のみが 1 で次元数が d の one-hot ベクトルの乗算を表す）

最後に，ユーザは，サーバから受領した暗号文を自身の秘密鍵を用いて復号し，類似度が高い順に並び替え，推薦結果とする．

5　評 価 実 験

5.1　実 験 概 要

ニュース推薦システムのプライバシ保護という課題に対して，準同型暗号はその解決策となることが期待される．しかし，準同型暗号上での演算レイテンシが長いことが課題である．そこで，4 節では，NRMS-BERT による推薦を 3 つのフェーズに分割し，PIR 及び推薦結果の獲得を準同型暗号上で実行し，各 Encoder 上での処理は平文上で実行するプロトコルを提案した．これにより，プライバシを保護しながら，高精度のニュース推

薦を実現する．準同型暗号上の演算レイテンシが長い問題に対する提案手法の有効性を検証するために，以下の 3 つの評価実験を行う．

（1） ニュースベクトルの獲得フェーズにかかるレイテンシの評価（5.4 項）

（2） ユーザベクトルの算出フェーズにかかるレイテンシの評価（5.5 項）

（3） 推薦フェーズにかかるレイテンシの評価（5.6 項）

なお，ユーザとサーバ間の通信は測定対象外とする．また，実験に用いたパラメータが適切であったことを確認するため，推薦結果の精度 (5.7 項）を計測し，平文上での実行と比較する．実験に関するコードは Python を用いて実装した．

5.2 データセットと学習済みモデル

本実験では，Microsoft の提供する Microsoft News Dataset (MIND) [13] を用いる．MIND は，ユーザの行動情報（ニュース閲覧履歴と，そのユーザがクリックしたニュース記事の情報）を含むデータセットである．各ユーザについて，ニュース閲覧履歴が最低 5 つ以上存在し，そのユーザがクリックしたニュース記事の情報を含む．また，ユーザの行動情報がインプレッションとして保存されている．各インプレッションには，1 人のユーザがホームページに訪れたときに表示されたニュース記事と，その記事がクリックされたかどうかが保存されている．ユーザの行動情報が保存されている behavior.tsv に記録されている情報を表 1 にまとめる．なお，MIND では，学習用のデータセットと評価用のデータセットがそれぞれ公開されている．

表 1　MIND の behavior.tsv に保存されている情報のまとめ

カラム名	説明
Impression ID	各インプレッションの固有 ID
User ID	ユーザの固有 ID
Time	ユーザがホームページに訪れた時間（インプレッションの時間）
History	ユーザの過去の閲覧履歴
Impressions	インプレッションで表示されたニュースとそれが実際にクリックされたかの情報

本実験では，MIND のうち，ランダムに 50,000 人のユーザを抽出したデータ量の小さなデータセットである MIND-small[3] を用いる．MIND-small の学習用データは 156,965 件のユーザの行動情報と，51,282 件のニュース情報を持つ．また，MIND-small の評価用データセットは，73,152 件のユーザの行動情報と 42,416 件のニュース記事の情報を持つ．評価実験においては，MIND-small の評価用データセットを使用する．また，本実験では，MIND-small の学習用データセットを用いて学習した NRMS+BERT-base の学習済み公開モデル [4]を使用する．

3：https://msnews.github.io/

4：https://github.com/YadaYuki/news-recommendation-llm

5.3 実 験 条 件

5.3.1 実 行 環 境

評価実験で用いた計算機の構成を表 2 に示す．なお，ユーザとサーバのデバイスを区別せず，同一デバイスで実験を行う．また，OpenFHE によるライブラリレベルでの並列化のみ使用する．

表 2　評価実験で用いた計算機の構成

構成要素	値
CPU	Apple M2 Pro
周波数	3.49 [GHz]
コア数	8
メモリサイズ	16 [GB]
OS	macOS 13.2.1
準同型暗号ライブラリ	Official Python wrapper for OpenFHE 0.8.3 [5]
深層学習ライブラリ	PyTorch 2.0.1 [6]
自然言語処理ライブラリ	Transformers 4.30.2 [7]

5.3.2 使用したパラメータ

評価実験に使用したパラメータを表 3 に示す．

表 3　評価実験で用いたパラメータ

分類	パラメータ	使用した値
BERT	d	768
MIND	m	42,416
準同型暗号	multiplicative depth（フェーズ 1）	1
	multiplicative depth（フェーズ 3）	2
	ring dim（フェーズ 1）	8,192
	ring dim（フェーズ 3）	16,384
	scaling mod size	50
	n_{slots}	4,096
提案手法	n_{tables}	11

5.4 ニュースベクトルの獲得レイテンシ

1 つのニュース記事あたりのニュースベクトルの平均獲得レイテンシを評価する．なお，サーバは予め MIND-small の評価用データセットに含まれる全てのニュースに対応するニュースベクトルを保持していると仮定する．準同型暗号では，暗号文に依存した条件分岐が存在せず，レイテンシが入力に依存しない．そのため，評価用データセットのニュース記事のうち，100 件のニュースをランダムに算出し，1 件ごとにレイテンシを計測する．さらに，この操作を 5 回行い，ニュースベクトルの平均獲得レイテンシを算出する．

5.5 ユーザベクトルの算出レイテンシ

ユーザが過去に閲覧したニュースベクトルを User Encoder の入力とし，ユーザベクトルを得られるまでのレイテンシを計測する．なお，ユーザは過去に閲覧したニュースのベクトルを既に保持していると仮定する．

5：https://github.com/openfheorg/openfhe-python

6：https://github.com/pytorch/pytorch

7：https://github.com/huggingface/transformers

5.6 推薦結果の獲得レイテンシ

データセットのうち，全てのインプレッションを用いて，1ユーザあたりの推薦結果の獲得レイテンシを計測する．なお，既に対象ユーザの暗号化済みのユーザベクトルと推薦候補のニュースベクトルをサーバが保持していると仮定する．また，準同型暗号によるレイテンシ増加が予測されるため，平文上でも同様の実験を行い，レイテンシを比較する．ここで，推薦結果のソート処理及び表示処理は計測対象外とする．

5.7 推薦結果の精度評価

CKKS 方式では，準同型暗号パラメータや演算内容に応じて，復号結果に含まれる誤差の大きさが異なる．そこで，本実験で用いた準同型暗号パラメータが適切であることを確認するために，平文実行を比較対象として，推薦結果の精度を評価する．

データセットの全てのインプレッションを対象に評価を行う．本実験では, [18] と同様に，以下の評価指標を用いて評価する．

- AUC: $AUC = \int_0^1 ROC(x)\,dx$
- MRR：$MRR = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{k_u}$
- $nDCG@k$：$nDCG@k = \frac{DCG@k}{DCG_{max}@k}$ $(k \in \{5, 10\})$

$ROC(x)$ は各推薦候補アイテムの推薦スコアを閾値とした時の，FPR と TPR をプロットした ROC 曲線を表す．ここで，TPR は予測したニュースの全てのペアについての推薦スコアを比較した時の，真陽性率（ユーザにクリックされたニュースのスコアの方が閾値よりも高い）であり，FPR は偽陽性率（ユーザにクリックされなかったニュースのスコアの方が閾値よりも高い）である．U は全ユーザの総数，k_u は候補ニュースを推薦スコア順にソートした時に，実際にクリックされたニュースが現れる最初の順位である．また，$DCG@k$ は，i 番目のアイテムの推薦スコア rel_i を用いて式 10 で求められる [1]．DCG_{max} は最適なランキングに基づいて計算された DCG の最大値である．

$$DCG@k = \sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{10}$$

6 実験結果と考察

6.1 実 験 結 果

レイテンシに関する評価実験の結果を表 4 に示す．また，推薦精度の評価実験の結果を表 5 に示す．

表 4 各フェーズにおける平均レイテンシ

実験項目	平均レイテンシ [ms]		
	ユーザ	サーバ	合計
ニュースベクトルの獲得	30.12	6,678.36	7,008.48
ユーザベクトルの算出	1.23	(N/A)	1.23
推薦結果の獲得（提案手法）	16.37	2,264.08	2,280.45
推薦結果の獲得（平文実行）	(N/A)	1.20	1.20

(N/A) は当該パーティが処理を行わないことを意味する．

表 5 推 薦 精 度

	AUC	MRR	$nDCG@5$	$nDCG@10$
NRMS-BERT（平文上推薦）	0.651	0.306	0.336	0.400
提案手法（暗号文上推薦）	0.651	0.307	0.337	0.401

6.2 ニュースベクトルの獲得レイテンシ（ニュースベクトルの獲得フェーズ）に関する考察

4 節で述べたように，ニュースベクトルの獲得は推薦とは非同期的に行えるため，ニュースの閲覧中にニュースベクトルを獲得できる．ここで，ユーザは 500 単語の記事を 238 単語/分のペースで読むとする [2] と 1 分 40 秒で記事を読み切る．そのため，1 件あたり 6.7 秒で該当ニュース記事のニュースベクトルを獲得することは現実的である．

しかし，獲得レイテンシはニュース総数 m に線形比例する．実運用では，大量のニュース記事が対象となることが想定される．したがって，m が大きくなったときも低レイテンシでニュースベクトルを獲得することは今後の課題である．

また，ニュースベクトルの再獲得における総レイテンシが長いことも課題である．News Encoder や User Encoder がアップデートされた際には，ニュースベクトルの再獲得が必要となる．しかし，ユーザが閲覧したニュース数は時間が経つにつれ多くなるため，全てのニュースベクトルを再獲得するためのレイテンシが長くなる．すなわち，ユーザの最新の嗜好の反映に時間がかかり，推薦の質が低下してしまう．対策としては，再獲得を行わず，古いモデルで推薦結果を得る方法や，ユーザベクトル算出時に使用するニュースベクトルの総数を減らす方法が考えられる．ただし，これらの方法には，レイテンシと推薦精度のトレードオフが存在する．

6.3 ユーザベクトルの算出レイテンシ（ユーザベクトルの算出フェーズ）に関する考察

表 4 が示すように，ユーザベクトルの算出レイテンシは 1.23[ms] であり，ニュースベクトル獲得や推薦結果獲得におけるユーザの処理にかかるレイテンシと比べると短かった．この結果は，ユーザが新しいニュースベクトルを獲得したとき，低コストでユーザベクトルの更新を行うことができることを意味する．ユーザベクトルの更新を素早く行うことで，最新のユーザの嗜好を推薦に反映することができるため，実運用においても高い推薦精度を達成できることが期待される．

6.4 推薦結果の獲得レイテンシ（推薦フェーズ）に関する考察

推薦結果の獲得レイテンシは，平文実行では 1.2[ms] であった．一方，提案手法では，レイテンシは 2,280[ms] であり，その内の 99.3%がサーバ側の処理であることから，ボトルネックはサーバ側の処理にある．そのため，サーバ側の処理のレイテンシの削減は今後の課題であり，改善のために更なるアルゴリズムの改善が必要である．例えば，ユーザがプラットフォームを訪れる前に定期的に推薦ニュースを更新しておくという方法が考えられる．この方法はニュース即時性とレイテンシのトレードオフとなるため，この方法と提案手法のどちらを選択するか

はプラットフォームの特性に依存すると言える．

6.5 推薦結果の精度評価（推薦フェーズ）に関する考察

表 5 で示したように，提案手法の推薦精度は平文実行での NRMS-BERT と同程度であった．これは，CKKS 方式による誤差が十分小さく，本実験で用いた準同型暗号パラメータは適切であったことを意味する．

6.6 提案手法の通信量に関する分析

暗号文の mutliplicative depth を D，ring dim を N とすると，暗号文は $2 \times N \times (D+1)$ 個の要素（整数）で構成される．また，OpenFHE において，1 要素は 8 バイトである．したがって，ニュースベクトル獲得フェーズにおいて，ユーザからサーバへ送信する暗号文のサイズはたかだか $2 \times N \times (D+1) \times n_{tables} \times 8$[Byte]（評価実験で用いたパラメータでは 2.9[MB]）である．サーバからユーザへ送信する暗号文のサイズはたかだか $2 \times N \times 1 \times 8$[Byte]（評価実験で用いたパラメータでは 180.2[KB]) である [8]．同様に，推薦結果の獲得フェーズでは，ユーザからサーバへ送信する暗号文のデータサイズは，たかだか $2 \times N \times (D+1) \times 8$[Byte]（評価実験で用いたパラメータでは 786.4[KB]）であり，サーバからユーザへ送信する暗号文のサイズはたかだか $2 \times N \times 1 \times 8$[Byte]（評価実験で用いたパラメータでは 262.1[KB]) である [8]．

7 おわりに

従来のニュース推薦システムにおいては，ユーザプライバシを保護することが求められる．プライバシ保護手法として挙げられる準同型暗号は，演算レイテンシの長さが課題である．そのため，ニュース推薦モデルの各 Encoder 部分に準同型暗号を直接適用すると，実用に適さないほどのレイテンシの長さになることが懸念される．

そこで本稿では，PLM-NR に準同型暗号を適用しつつ，低レイテンシで推薦を行うことができるプロトコルを提案した．具体的には，提案プロトコルは，推薦を 3 つのフェーズに分割し，ユーザベクトルの算出をユーザ自身のデバイスで平文上で実施し，プライベートな情報を用いるサーバ上での処理を準同型暗号上で実行する．そして，評価実験によって，提案プロトコルは現実的な低レイテンシでの推薦を行うことができることを示した．したがって，本研究は，高い推薦精度を維持しつつ，ユーザーのプライバシーを保護し，かつ低レイテンシで推薦を行うプロトコルを設計するという研究の目的を達成した．

今後の課題は，提案手法のさらなる実用性向上を目的とした，提案手法のアルゴリズムを改善することによるレイテンシ削減である．また，提案手法は推薦処理のみを対象としているため，プライバシを保護したままモデルのアップデートを行えるようにすることも今後の課題である．

8：準同型暗号上での計算の出力暗号文の *multiplicative depth* は最小値の 0 であるため．

文　献

[1] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G., “Learning to Rank Using Gradient Descent.” in Proc. of the 22nd International Conference on Machine Learning, pp. 89-96, 2005.

[2] Brysbaert, M., “How many words do we read per minute? A review and meta-analysis of reading rate.” Journal of Memory and Language, Vol. 109, pp. 1-30, 2019.

[3] Canny, J., “Collaborative Filtering with Privacy.” in Proc. of 2002 IEEE Computer Society Symposium on Research in Security and Privacy, pp. 45-57, 2002.

[4] Cheon, H.J., Han, K., Kim, A., Kim, M., and Song, Y., “Bootstrapping for Approximate Homomorphic Encryption.” Advances in Cryptology - EUROCRYPT 2018, LNCS, Vol. 10820, pp. 360-384, 2018.

[5] Gentry, C., “Fully Homomorphic Encryption Using Ideal Lattices.” in Proc. of the 41st annual ACM symposium on Theory of computing, pp. 169-178, 2009.

[6] Jacob, D., Ming-Wei, C., Lee, K., and Toutanova, K., “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv, arXiv:1810.04805, pp. 1-16, 2018.

[7] Jumonji, S., Sakai, K., Sun, M.T., and Ku, W.S., “Privacy-Preserving Collaborative Filtering Using Fully Homomorphic Encryption.” IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 3, pp. 1961-1974, 2023.

[8] Okura, S., Tagami, Y., Ono, S., and Tajima, A., “Embedding-Based News Recommendation for Millions of Users,” in Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’17), pp. 1933-1942, 2017.

[9] 大西 隆太郎, 鈴木 拓也, 山名 早人., “準同型暗号と隔離実行環境の組み合わせに関するサーベイ.” コンピュータセキュリティシンポジウム 2022 (CSS 2022), 2A2-I-1, pp. 1-8, 2022.

[10] Shahriar, B., Xun, Y., and Ibrahim, K., “A Practical Privacy-Preserving Recommender System.” Data Science and Engineering, Vol. 1, No. 3, pp. 161-177, 2016.

[11] Smart, P.N., and Vercauteren, F., “Fully Homomorphic Encryption with Relatively Small Key and Ciphertext Sizes.” Public Key Cryptography (PKC 2010), LNCS, Vol. 6056, pp. 420-443, 2010.

[12] Tada, M., Kikuchi, H., and Puntheeranurak, S., “Privacy-Preserving Collaborative Filtering Protocol Based on Similarity between Items.” in Proc. of 2010 IEEE 24th International Conference on Advanced Information Networking and Applications, pp. 573-578, 2010.

[13] Wu, F., Qiao, Y., Chen, J.H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., and Zhou, M., “MIND: A Large-Scale Dataset for News Recommendation.” in Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3597-3606, 2020.

[14] Xiangfu, M., Huo, H., Zhang, X., Wang, W., and Zhu, J., “A Survey of Personalized News Recommendation.” Data Science and Engineering, Vol. 8, No. 4, pp. 396-416, 2023.

[15] Wood, A., Najarian, K., and Kahrobaei, D., “Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics.” ACM Computing Surveys, Vol. 53, No. 4, Article No. 70, pp. 1-35, 2020.

[16] Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., Xie, X., Inui, K., Jiang, J., Ng, V., and Wan, X., “Neural News Recommendation with Multi-Head Self-Attention.” in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6389-6394, 2019.

[17] Wu, C., Wu, F., Huang, Y., and Xie, X., “Personalized News Recommendation: Methods and Challenges.” ACM Transactions on Information Systems, Vol. 41, No. 1, pp. 1-50, 2023.

[18] Wu, C., Wu, F., Qi, T., and Huang, Y., “Empowering News Recommendation with Pre-Trained Language Models.” in Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21), pp. 5753-1656, 2021.

解析処理単位の認可および監査に基づく柔軟なデータ流通処理基盤の提案

松本　直樹† 　川西　智也†† 　大平　健司††† 　小塚　真啓††††

† 京都大学
†† セーラ・ネットワークス株式会社
††† 大阪大学
†††† 岡山大学
E-mail: †mt2.naoki@inet.media.kyoto-u.ac.jp, ††cuzic@seera-networks.com, †††ohira@cmc.osaka-u.ac.jp, ††††kozuka@okayama-u.ac.jp

あらまし　機微な情報を含むデータについて、暗号技術や秘匿計算用のハードウェアを用いてプライバシを保護しつつデータ解析を実現する手法が提案されている。しかし、実際のデータ提供においては、利用目的や匿名化処理などに関する合意を行うが、実際にその合意に従っているか監査することは困難である。また、合意の必要がないように匿名化されたデータでは粒度とのトレードオフが存在する。本研究では、統計に関する解析処理を対象に、データ提供者とデータ解析者間での処理単位の認可に基づく、データ流通処理基盤を提案する。本基盤では、データ提供者はデータ解析者に対して、処理単位で認可を行い、また、行われる処理を記録することで合意に対する監査を実現する。また、データ解析者の属性に応じて認可する処理単位を切り替えることで、解析者本人に関するデータは詳細な統計処理を許可しつつ、それ以外のデータについては匿名性が保証される解析処理のみ許可するといった柔軟な解析処理の実現を目指す。本提案手法について、TEE の一種である AMD SEV-SNP 内で動作するプロトタイプを実装し、評価を行った。

キーワード　データ処理、プライバシ、Trusted Execution Environment

1　序　　論

大規模なデータに対する統計処理や機械学習により、生活を豊かにする技術やサービスが多く提案、開発されている。それらの活用が進むにつれて、取得したデータの流通や柔軟な活用が重要視されている。しかし、取得したデータは各個人の機微な情報を含む場合がある。そのため、現在は匿名加工や仮名加工といった匿名化処理を施し、個人を特定できないように加工されたデータの利活用が進んでいる。また、各企業が各個人に対して利用規約や契約事項として他社への個人情報の提供を認めた場合、その範囲において利用されている。これらの個人情報を保護する法律や取り組みは我々個人のプライバシを保護する上で非常に重要であるが、高度なデータの利活用において障壁となっていることは禁じ得ない。また、各企業が保有する個人情報を活用における契約について、それに基づく利用が行われているか、各個人ないしは提供側の企業で検証することの重要であるが、検証する手段は限られる。

個人情報を含む機微なデータを処理する手法として、秘密分散や準同型暗号等の暗号技術を用いることでデータの中身が分からない状態で解析処理を行う手法が提案されている。しかし、これらの手法はすべての解析手法に適用できるとは限らず、可能な処理を広げようとすると要する計算量が大きくなるなど、実用上の問題がある。一方で、Trusted Execution Envinroment (TEE) と呼ばれる、専用の隔離された領域で解析処理を行う手法が提案 [1] [2] [3] [4] されている。TEE を用いることで、通常のプロセッサの処理やメモリ領域から分離され、メモリ上のデータも常に暗号化された状態で処理が行われるため、プロセッサやサーバーの所有者に機微なデータを見せることなく処理を行うことができる。TEE において動作するデータベースやデータ処理アプリケーション、さらには汎用なフレームワークが提案されており、通常のアプリケーションと同等の機能を TEE においても提供することが出来る。しかし、これらのアプリケーションは処理自体の機密性は保証するが、入出力については考慮されていない。そのため、入力によっては機微なデータが出力される可能性があり、あらゆるデータ処理において利用できるとは限らない。

各個人のデータをプライバシを保護しつつ扱うシステムとして、Personal Data Management System (PDMS) [5] が提案されている。PDMS はデータの収集や取得を適切に管理することで個人情報を保護しつつデータの活用を推し進めるために設計された。また、PDMS を拡張し、データの収集や処理を個々のタスクとして分割し、それらに対してデータに対するアクセス制御を行うことで機微なデータに対する操作を制御する手法 [5] が提案されている。しかし、PDMS はユーザが所有する計算資源やデータストレージを用いてデータの処理を行うことを念頭に置いている。そのため、クラウドなどの膨大な計算資源の活用や、データ提供者、データ解析者が多岐にわたる場

合のデータの取り扱いについては議論の余地がある。

本論文では、統計的手法を用いるデータ解析処理において、データ解析者による解析の処理単位ごとにデータ提供者が認可ないしは監視することによる柔軟なデータ処理流通基盤を提案する。これまでの手法は、個人を特定しうるデータの削除や差分プライバシのようにノイズの付与によりプライバシ保護を行っていた。しかし、現実には実データに近いデータの提供を行う場合がある。本基盤はそのような状況を想定し、実データに対する解析処理について認可を行うことで、同じデータであっても解析者の属性に応じて解析者に対して与えうるデータの粒度や解析範囲の限定を目指す。具体的には、同一組織内の解析者に対してはあらゆる解析処理を認可しつつ、外部や連携組織の解析者については特定範囲や簡単な解析処理のみ許可するという利用を想定する。これにより、簡単な解析はデータ提供者とデータ解析者間の合意なしに行いつつ、機微なデータが漏れうる場合は改めてデータ提供の合意を行うという段階的なデータ提供が実現できる。また、既に行われている合意に基づくデータ提供において、その合意が履行されているか確認する手法を提供する。解析者が行うデータ解析処理の各解析処理を記録しておき、データ提供者が必要に応じて記録の監査を行うことで、合意に対する違反の有無を確認することができ、不意の解析者側からのデータ漏洩を抑止することができる。

本基盤では、解析処理単位で認可を行うために、各解析処理を解析処理モジュールとして分割し取り扱う。また、クラウドなどの膨大な計算資源を利用するために、計算資源はデータ提供者やデータ解析者以外により提供されることを想定している。そのため、データ提供者とデータ解析者もそれぞれが自身が行うデータ提供、データ解析の処理をデータ提供モジュール、データ解析モジュールとして記述する設計となっている。既存の統計処理ライブラリの利用を想定し、汎用のプログラミング言語で記述可能であり、さらに、外部のデータストアとの通信も行えるようなモジュールとして、WebAssembly [6] を採用した。また、ユーザーに対してモジュール記述に関する規則を与えるために、Component Model [7] を用いたインタフェースとその引数の型を定義し、それに従って各モジュール間の連携を行う。計算資源はクラウドなど第三者により提供されることを想定している。データ提供者が提供するデータの機密性や各モジュール、本基盤自体の完全性が損なわれた状態で動作することが無いよう、本基盤は TEE 上での動作を前提とし、実行環境や基盤自体の完全性の検証を含めて検証を行う。

TEE の一種である AMD SEV-SNP [8] を対象に本基盤の実装を行った。計量経済学において広く知られるデータセット [9] に対する解析処理として、単回帰分析を行うことができることを確認した。また、本提案基盤が提供する認可機構により、特定の解析処理の制限や強制が可能であることを確認した。

2 提案手法

図 1 に示す本基盤では、統計的手法を用いたデータ解析において、解析処理単位で認可及び監査を行うことにより、データ解析者ごとに実行可能な解析処理を切り替え、また、データ提供者がデータに対して行われた解析処理の事後監査を可能にする。これらの仕組みにより、合意に基づく二者間でのデータ利活用に限らず、不特定多数にデータを提供しつつ、なおかつ可能な解析処理を限定することでプライバシの保護を図る。

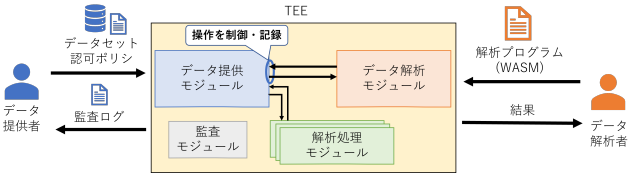


図 1　提案基盤の概観

2.1 提案基盤の概要

本基盤は、次の参加者により構成される。

データ解析者　提供されるデータに対して、解析処理を行う。本基盤では、共分散や単回帰分析、重回帰分析といった、計量経済学等で用いられる統計的手法を用いたデータ解析を想定し、機械学習のような複雑な処理は想定しない。

データ提供者　解析者に対して、自身が保有するデータを提供する。解析者に対して、データに対して許可する解析処理 (統計処理) 単位の認可を行う。また、自身のデータに対して行われた解析処理を必要に応じて監査する。

計算資源提供者　データ解析者及びデータ提供者に対して、計算資源を提供する。本基盤では、特定の企業や組織が占有的にその機能を提供することは想定せず、オンプレミスやクラウドといった計算資源を利用することを想定している。データ提供者により提供されるデータの機密性の保証や、本基盤が想定されたプログラムに基づいて稼働していることを検証するために、本基盤は TEE 上で稼働する。

データ解析者とデータ提供者は、それぞれ希望する処理をモジュールとして記述する。各解析処理に関して、それぞれ対応するモジュールが提供され、データ解析者はこれらのモジュールを呼び出すことで解析処理を記述する。提供される解析処理に不足があった場合は、モジュール単位で容易に追加及び利用が可能な設計となっている。

データ解析者は、原則としてデータ提供者により認可された解析処理のみ実行が可能である。認可されていない処理については実行することはできない。データ提供者に対して、データ解析者の属性に応じて認可に用いるポリシを使い分ける機能を提供する。具体的には、データ解析者が提供者と同一の組織ないしは事前の合意に基づいて無制限の利用を認める場合は、あらゆる解析処理を認可し、一方で、協力関係にはあるが目的外のデータや解析処理を許可したくない場合は必要に応じて認可する解析処理を制限することができる。また、データ提供者は、データ解析者による解析処理を必要に応じて監査する機能が提供される。これまでのデータ利活用と同様に、解析処理の制限自体は行わない一方で、事前の合意でデータの用途や利用範囲が決められている場合に用いることを想定している。本基盤では、解析処理を記録し、その記録に基づいてデータ提供者から

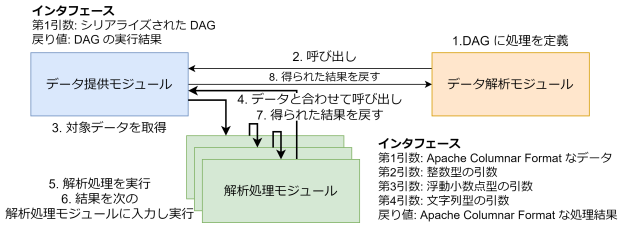


図 2　各モジュール間の連携

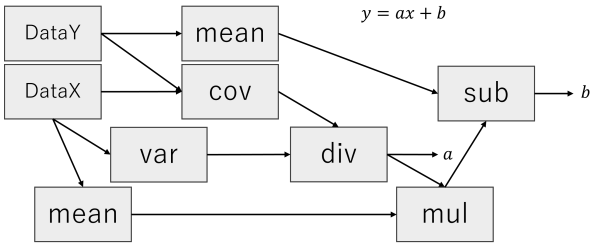


図 3　DAG による単回帰分析の記述例

の要請に基づき合意から逸脱していないか監査を行う。また、逸脱があった場合は、その記録を合意に対する違反の証拠として提供する。

2.2　モジュールによる機能分割

本基盤では、図 2 のように参加者ごとにモジュールを分割して記述し、基盤上でモジュール間を連携させることで処理を実現している。処理をモジュールとして分割することにより、参加者は独自の処理を自由に記述することが可能となる。また、解析処理についても本基盤で取り扱う各解析処理ごとにモジュールとして分割することにより、解析処理単位の認可、監査をモジュールの呼び出しおよび実行という形態で実現することができる。本基盤は 3 種類のモジュールから構成され、各モジュール間の連携は本基盤側が行う。

解析処理モジュール　解析処理自体を記述したモジュールである。実装される処理の粒度としては、平均や分散、共分散といった、一般的な統計処理で用いられる関数の粒度で実装されることが想定されている。解析処理の認可及び監査は本モジュールの粒度で行われるため、できる限り高い粒度で実装することが望ましい。しかし、そのような高い粒度を持つモジュールを組み合わせる場合、処理を記述するデータ解析者の負担が大きくなる。本基盤では、図 3 のような複数の解析処理モジュールを単一の解析処理モジュールと同様に扱う仕組みを持つ。完全なブラックボックスとして高機能な処理を定義するのではなく、複数の解析処理モジュール間の依存関係に基づいて定義することにより、複数モジュールが組み合わせられた場合であっても処理内容の検証を容易にすることを意図している。

データ提供モジュール　データ提供者がデータ解析者に対して所有するデータを供給するために利用するモジュールである。自身が所有するデータ提供モジュール自体にデータを持つだけでなく、リモートのデータストアからデータを取得することも想定されている。

データ解析モジュール　データ解析者が自身が望む解析処理内容を記述するモジュールである。データ自体へのアクセスはできず、データ解析モジュール内からデータ提供モジュールを経由して解析処理モジュールを呼び出すことで、解析した結果のみえられる。しかし、解析処理の記述において、毎回解析処理モジュールを呼び出す場合、呼び出しに伴うオーバーヘッドや、データ解析者に対してあらゆるデータが漏れることになる。そこで、本基盤では、解析処理の一つのまとまりを有向非巡回グラフ (Directed Acyclic Graph, DAG) として記述する。記述された処理が最終的に得た結果のみデータ解析モジュールに回答することにより、データ解析モジュールに対して漏洩する真のデータを減らすことを意図している。図 3 に単回帰分析を DAG により記述した例を示す。データの平均値を表す `mean` や共分散を表す `cov` が解析処理モジュールとして提供される。図 2 に示すように、DAG に基づいてデータ提供モジュールが提供したデータとともに解析処理モジュールを呼び出す。解析処理モジュールから得られた結果を、再度別の解析処理モジュールの入力とし呼び出す。または、結果としてデータ解析モジュールに DAG の実行結果として返す。

本基盤におけるモジュールの実装では、次の 4 つの要件が存在する。1 点目は、各モジュール間が定められた API 以外での操作ができないようなモジュールであることである。各モジュール、特にデータ提供モジュールについてはデータ提供者が提供する機微なデータもモジュール内で処理されている場合がある。そのため、本基盤で定められた API 以外でデータの読出しや処理の実行が可能となるような脆弱性が存在した場合、予期せぬデータの漏洩が生じる恐れがある。2 点目は、既存のプログラミング言語により各モジュールの処理を記述可能なことである。データ提供モジュールや解析処理モジュールでは、既存のデータベースや統計処理のライブラリを活用することが想定される。そのため、できる限り既存のライブラリを活用することができるようなプログラミング言語でモジュールを記述できるような設計とする必要がある。3 点目は、外部との連携を行うインタフェースを容易に持つことが可能なことである。各モジュールはモジュール間同士の連携に限らず、解析処理の認可を行うために、基盤側でモジュールの呼び出しや呼び出しに係る引数等の情報を必要とする。そのため、各モジュールは、モジュール間に限らず、それ以外の基盤や監視モジュールからも容易に解釈可能なフォーマットのインタフェースを持つ必要がある。4 点目は、ホストが持つ通信機能やファイルシステムをモジュールが利用できることである。データ提供モジュールでは、データ提供者が持つ外部のデータストアからデータを取得する場合が想定されている。そのため、モジュール内から必要に応じて外部のエンドポイントとの通信が行えるような機能が必要となる。また、既に公開されているデータセットについてローカルにダウンロードして利用する場合について、モジュール内からホスト上のファイルを操作する機能が必要となる。この際、無制限にファイル操作を許すことは脆弱性となるため、読み書き可能なディレクトリや容量など適切に制限する必要がある。

2.3 想定する脅威モデル

本基盤では、データ提供モジュール、データ解析モジュール、解析処理モジュールは本基盤が想定するインタフェースや呼び出しの型に関する仕様に従い実装されていることを想定する。仕様に沿わないモジュールの実装は本基盤では取り扱えず、解析処理自体を開始できないため想定しない。また、解析処理モジュールは仕様として与えられる関数やアルゴリズムに従い正しく実装されていることを想定する。つまり、解析処理モジュールが持つ脆弱性や不正な記述に起因する予期せぬ解析処理の実行の可能性は本基盤においては存在しないものとして扱う。データ解析者は、本基盤において提供するインタフェースとプロトコルに従い、解析処理を依頼する。なお、データ解析者の身分は本基盤に依らない方法で事前に検証され、なりすまし等が行われていないことを前提とする。データ提供者も同様に、本基盤が提供するインタフェースとプロトコルに従い、データ提供を行う。なお、データ提供モジュール内からデータ提供者が持つデータストアに対して接続する場合は、本基盤は通信における認証機能は提供せず、各データ提供モジュールの実装において適切に認証が実装されていることを前提とする。本基盤では、TEE が持つ機密性と完全性を保証する機能およびそれらを検証する機能を利用する。計算資源提供者は基盤が対応する TEE を提供することが求められる。

2.4 モジュール間呼び出しにおける認可および監査

本基盤では、データ提供者が提供するデータの不正な利用の防止や機密性を保証するために、解析処理モジュールの呼び出しに対して認可および監査を行う機構を提供する。統計学分野では、データ公開におけるプライバシ保護を図る統計的開示抑制 (SDC) と呼ばれる手法が研究されている。SDC では、データベースに対するクエリ自体を制限するクエリ監査とクエリに対する回答結果の制限やかく乱するクエリ推論制御 [10] が存在する。本基盤において、SDC に相当する機能を実現するためには次の 2 点の要件が挙げられる。1 点目は、解析処理モジュールとそれで扱うデータの種類や件数に応じた認可である。本基盤で扱うデータは、値が連続値や文字列と多様であり、またその件数も制限を設けていない。件数が少ないデータに対してそれを考慮することなく認可を行うと、予期せぬデータの漏洩が生じる可能性がある。そのため、扱うデータの種類や件数に応じて認可の内容を変えることができる仕組みが必要である。2 点目は、データ解析者に対して特定の解析処理を強制する仕組みである。1 点目ではデータ解析モジュールで呼び出される解析処理モジュールの受動的な認可を行った。しかし、プライバシ保護の用途では、Laplace ノイズの付与といった処理の実行した場合のみ結果の出力を許可する認可が求められる。つまり、データに対して特定の解析処理を行わない限りデータ出力を許可しない仕組みが必要である。

本基盤では、これらの要件を満足するよう設計した認可ポリシを用いる。データ提供者は、定義した認可ポリシをデータ解析者に応じて使い分けることで、データの段階的な提供を実現する。認可に用いるポリシの例を図 4 に示す。本基盤においては、データ提供者のデータを取得する関数である loc について、ポリシの設定を行う。policy/rules で、認可に用いるルールを定義し、func_policy は各解析処理の認可条件を定義する。ポリシは複数定義が可能である。func ではデータを取得する関数である loc を指定し、arg ではポリシの対象となる列名 educ を指定してる。そして、educ のデータに対して requires で指定されるいずれかの解析処理を強制する。ここでは、regression_ge_50 または var_any の実行を強制している。regression_ge_50 は 50 件以上のデータに対して解析処理 regression を認可することを示している。また、rejects で指定されている解析処理については実行を拒否する。

```
"policy":{
  "rules":{
    "rule1":{
      "func":"loc",
      "arg":"educ",
      "rejects":[
        "var_le_100"
      ],
      "requires":[
        "regression_ge_50",
        "var_any"
      ]
    }
  }
},
```

```
"func_policy":{
  "defs":{
    "regression_ge_50" : {
      "func" : "regression",
      "num" : ">=50"
    },
    "var_any" : {
      "func" : "var",
      "num" : ">=0"
    },
    "var_le_100" : {
      "func" : "var",
      "num" : "<=100"
    }
  }
}
```

図 4　モジュール間呼び出し認可に用いるポリシ例

本基盤における認可ポリシでは、前述したように各項目ごとにポリシを定義する。しかし実際には、複数の項目を結合した解析処理が行われ、結合した結果についてもポリシを適切に適用する必要がある。提案する認可ポリシにおいては、複数の項目を結合した場合結合前の各項目のポリシを継承する。継承されたポリシ群は積として扱い、すべてのポリシに満足した場合のみ解析結果の出力を許可する。いずれかのポリシに含まれる requires を満足できない場合や、rejects に含まれる処理が実行された場合、データ解析者にはエラーを返し、解析結果を破棄する。なお、ポリシを直接記述することはデータ提供者の負担となる可能性があるため、データのフォーマットごとに標準的なポリシを提供するといった方法が考えられる。

監査においては、基盤において実行された DAG を記録として保持する。DAG はデータ解析者が行う解析処理すべてが含まれる。また、解析処理自体にはデータ解析者の知見が含まれる可能性があるため、むやみにデータ提供者に対して提供することは好ましくない。そのため、記録は自由に閲覧することができない状態で基盤上に保持される。合意に対する違反の疑いがある場合についてのみ開示することができるようにする。これは一種の開示請求の形態を模したものであり、開示者や開示の意図、開示された内容についても記録される。

2.5 各モジュールの提供および実行

解析処理モジュールとデータ提供モジュールは、基盤起動時にすでに準備されていることを想定している。そのため、事前にモジュールを生成し、基盤に対して登録しておく必要がある。データ解析モジュールは、データ解析者側の要求に応じて実行

できるように API 経由でモジュールの送付と実行を行う。データ解析者は、jupyter notebook 等でデータ解析の処理を記述し、生成されたモジュールを HTTP ないしは gRPC を利用して API 経由で基盤に対して提供する。基盤は、データ解析者からデータ解析モジュールを API 経由で受け取り、実行する。実行した結果は、API 経由でデータ解析者に対して提供される。

2.6 TEE による計算資源の提供

本基盤は、特定の企業や組織により提供される形態ではなく、本基盤ソフトウェアを持つ第三者が保有する計算資源において自由に実行ができる形態を想定する。しかし、データ提供者は機微なデータを提供する可能性があり、計算資源提供者が悪意ある者であった場合、データは容易に漏洩してしまう。また、本基盤自体が不正に改変されている場合、解析処理の認可が正常に動作せず、データ解析者に対して機微なデータが漏洩する可能性がある。これらを踏まえると、計算資源については次の 2 点の要件が存在する。1 点目は計算資源提供に対する基盤上のデータの機密性の保証、2 点目は稼働する基盤ソフトウェアの完全性の保証である。

本基盤では、これらの要件を満足するために TEE を用いる。TEE では、ハードウェアによるメモリの暗号化に基づく機密性の保証や、Attestation による完全性の検証が可能である。本基盤では、メモリ暗号化により基盤上のデータはすべて計算資源の提供者に対しても秘匿される。また、Attestation により、基盤ソフトウェアが稼働する CPU や TEE 自体の完全性や、基盤ソフトウェア自体の完全性を検証する。データ解析者やデータ提供者は、モジュールを本基盤に対して送付する前に、Attestation を行うことで、これから利用しようとしている基盤ソフトウェアの完全性を検証することができる。完全性が確認されると、稼働している基盤ソフトウェアは確かに意図したものであり、解析処理の認可及び監査についても正常に動作することが期待される。それにより、データ提供者やデータ解析者は、計算資源提供者にデータや処理内容を秘匿した状態で処理を行うことが可能となる。

3 実　　装

本提案基盤の設計に基づくプロトタイプの実装を行った。提案基盤における機能のうち、認可については提案認可ポリシのうちデータ解析者の属性考慮に関する機能を除き実装を行い、監査については実行された解析処理関数をログとして出力する機能を実装し、データ提供者に対して監査機能を提供するインタフェースについては未実装である。

3.1 WebAssembly を用いたモジュールの実装

本実装では、WebAssembly (Wasm) [6] をモジュールの実装として用いた。Wasm は Web ブラウザ向けの高速な実行形式として開発され、ホスト上においても Wasm ランタイムにより、Wasm モジュールを利用することが可能である。前述した 4 つの要件について、Wasm を用いることで次の通りに満足することができる。1 点目のモジュールの分離については、Wasm モジュールは一種の仮想マシンの形態で動作し、モジュールのプログラムやデータメモリはホストやほかのモジュールから分離されている。そのため、あるモジュールが別のモジュールのデータを読むことは原則として不可能である。2 点目のモジュールを記述する言語について、C や Rust といった広く用いられる言語において、通常のプログラムと同じ記述を行い、Wasm モジュールへとコンパイルすることが出来る。3 点目のインタフェースについて、Wasm では Component Model [7] と呼ばれるモジュールのインタフェースとその型を定義する仕様が開発されている。これにより、各言語はその型情報を元に Wasm モジュールをコンパイルしたり、モジュールを利用する側はその型に沿って呼び出し側のプログラムを作成することが可能となる。また、型定義までインタフェースに含めて定義されることにより、各モジュールを開発する開発者の負担も軽減することが出来る。4 点目のホストが持つ通信機能やファイルシステムの利用については、Wasm においては、WebAssembly System Interface (WASI) [11] と呼ばれる機能により実現することが出来る。WASI はホストが持つソケットやファイルの読み書きなどのシステムインタフェースを Wasm モジュール内から利用可能にする仕組みである。また、上述したプログラミング言語においても、WASI の利用を前提としたライブラリが提供されている。これにより、通常と同様の使い勝手で WASI を利用できることが期待される。

3.2 モジュール間の呼び出しで用いるデータ形式

モジュール間での呼び出しは、2 種類存在する。一つはデータ解析モジュールからデータ提供モジュールを介した解析処理の呼び出しである。これでは、利用するデータ提供モジュールと DAG を指定してデータ解析モジュールは呼び出しを行う。

もう一つは、データ提供モジュールから提供されたデータを引数として、解析処理モジュールを呼び出す場合である。本基盤では計量経済学等で用いられる単回帰分析といった基本的な統計処理を解析処理として想定している。そのような統計処理においては、2 次元の表形式でまとめられたデータが用いられ、本基盤においてもそのようなデータを想定している。

本基盤における解析処理モジュールの引数では、Arrow Columnar Format [12] に従う表形式データフォーマットを採用した。Columnar Format は、データ解析ソフトウェアである Apache Arrow [13] における列指向データのオンメモリデータフォーマットである。このデータフォーマットは Apache Arrow に限らない汎用な利用を念頭に策定されており、Polars [14] 等の他の表形式データの解析ライブラリ等でも用いられている。Columner Format に従った引数の型定義を Component Model において定義し、各解析処理モジュールはこの型定義に従って引数を受け取り、解析処理を行う。

3.3 AMD SEV-SNP における実装

AMD SEV-SNP は、AMD が提供するゲスト VM を対象とした TEE の一種である。SEV-SNP では、ゲスト VM に割り当てられたメモリページやレジスタが全て暗号化され、ホス

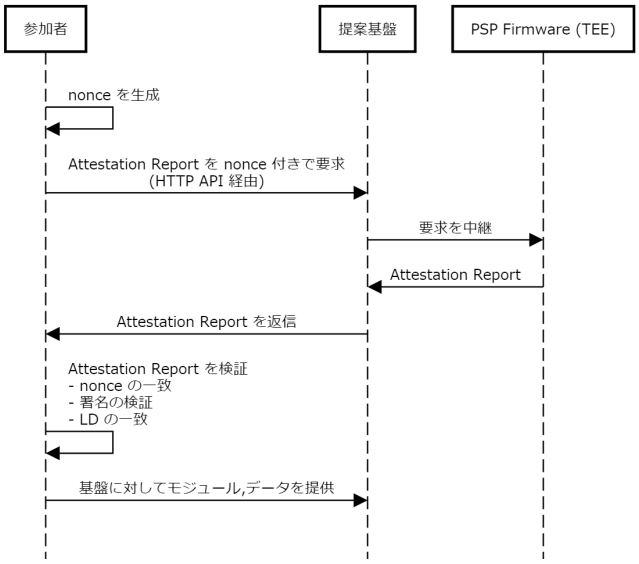


図 5　外部の参加者による Attestation

トやハイパーバイザからはそれらを閲覧することができない。SEV-SNP ゲスト VM 内で本基盤を稼働させることにより、各モジュールやそれらが処理するモジュールが取り扱うデータの機密性を保証することが出来る。要件として求められている基盤上のあらゆるデータの秘匿は、SEV-SNP ゲスト VM 内で基盤ソフトウェアを稼働させることにより実現した。

CPU や SEV-SNP の機能を提供する PSP Firmware 自体の完全性については、Attestation と呼ばれる機能により検証することができる。Attestation では、SEV-SNP ゲスト VM 内からの要求 (Attestation Report Request) を行うことにより、CPU が持つ秘密鍵により署名された Attestation 結果 (Attestation Report) を得られる。署名は AMD が配布する証明書を用いることで、Attestation Report が正しい CPU, PSP Firmware によって署名されていることを検証することが出来る。また、Attestation Report Request では 64 バイトの任意データを含めることができ、Attestation Report において同じ内容のデータが含まれる。そのため、取得毎に生成したランダムなデータをチャレンジとして含めることで、Attestation Report の使いまわしによる攻撃を阻止することができる。

実装においては、SEV-SNP ゲスト VM はディスクデバイスを持たない構成とした。本基盤ソフトウェアは initramfs に含まれ、全ソフトウェアと全データをメモリ上で保持する。Linux カーネルは、それ自体のサイズを削減や脆弱性のリスク軽減を意図し、ゲスト VM 内で利用することがないドライバやプロトコルスタックを除く最小構成でコンパイルしたものを用いている。起動後 Attestation Report を取得し、検証に成功した場合は基盤ソフトウェアを起動する。また、図 5 にあるように、基盤ソフトウェアは外部からの要求に応じて Attestation Report を取得、返答する機能を持つ。この機能を用いることで、データ提供者やデータ解析者は、稼働している基盤ソフトウェアのハッシュ値を含む Attestation Report を入手することができ、正しい SEV-SNP ゲスト VM 内で正しい本基盤が

表 1　評価に用いた環境及びソフトウェア

CPU	AMD EPYC 7313P (ゲスト VM に 1 コア割り当て)
メモリ	128GiB (ゲスト VM に 2GiB 割り当て)
OS	Ubuntu 23.10
SNP Firmware	1.55.16
ホストカーネル	https://github.com/AMDESE/linux (5a170ce1a08259ac57a9074e1e7a170d6b8c0cda)
ゲストカーネル	https://github.com/AMDESE/linux (690558b32fe6402c309341fbae65f41decb81592)
QEMU	https://github.com/AMDESE/qemu (fe4c9e8e7e7ddac4b19c4366c1f105ffc4a78482)
OVMF	https://github.com/AMDESE/ovmf (09fbe92dc545779671d7fd89a5bd4f1b14f7e69b)
sev-snp-measure	https://github.com/virtee/sev-snp-measure (40e6720a2fac6dfc2336815e4c85789c487cc5b2)
snpguest	https://github.com/virtee/snpguest (439f4446202567c8fbe89e9322667443427a43ea)

```
let mut dag: Dag = shared::Dag::default();
let node_x: Rc<RefCell<Node>> = dag.new_node(op: NodeOp::Function);
let node_y: Rc<RefCell<Node>> = dag.new_node(op: NodeOp::Function);
let node_reg: Rc<RefCell<Node>> = dag.new_node(op: NodeOp::SubDAG);

node_x.borrow_mut().func = loc(&exe_id, cols: &vec!["educ".to_string()]);
node_y.borrow_mut().func = loc(&exe_id, cols: &vec!["wage".to_string()]);
node_reg.borrow_mut().func = regression(&exe_id);

dag.add_dependency(prec: &node_x.borrow(), succ: &node_reg.borrow());
dag.add_dependency(prec: &node_y.borrow(), succ: &node_reg.borrow());
```

図 6　解析処理モジュールにおける処理の記述例

```
[mod-procedure] start to process procedure module
[function-chain] start name=loc args={"arg1":[],"arg2":[],"arg3":["educ"]} validate={"rule1"} hook={}
[function-chain] done name=loc args={"arg1":[],"arg2":[],"arg3":["educ"]} validate={"rule1"} hook={}
[function-chain] start name=loc args={"arg1":[],"arg2":[],"arg3":["wage"]} validate={} hook={}
[function-chain] done name=loc args={"arg1":[],"arg2":[],"arg3":["wage"]} validate={} hook={}
[mod-procedure] out_a = 0.5413592546651756
[mod-procedure] out_b = -0.9048516119572244
```

図 7　単回帰分析の実行結果

稼働しているか検証することができる。検証に成功した場合のみモジュール、データの提供を行うことで、機密性を侵害されることなく安全に処理を行うことができる。

4　評価および議論

4.1　評 価 環 境

本基盤の評価に関する環境を表 1 に示す。各ソフトウェアは記載の git リポジトリにおける記載コミットハッシュを持つものをビルドし利用した。`sev-snp-measure` は Measurement に必要なハッシュ値の計算と署名を行うユーティリティであり、Measurement を行うために利用している。`snpguest` は SEV-SNP ゲスト VM 内で Attestation Report を取得するために利用するユーティリティであり、本基盤における Attestation を実現するためにライブラリとして利用している。

4.2　ケーススタディ：単回帰分析

本基盤を用いて、図 3 に示すデータ x, y に対する単回帰分析 $y = ax + b$ のパラメータ a, b を求める処理を行った。データとして、計量経済学の教科書において用いられているデータセット [9] を CSV 形式に変換したもの [15] を利用した。データは

各項目が 526 個の数値データを持ち、24 個の項目から構成される。データ提供モジュールでは、ホスト上に存在する CSV 形式のファイルを読み込み、解析処理モジュールに対して提供する。図 6 にデータ解析モジュールにおける記述の一部を示す。データ解析モジュールでは、以下の式に従い単回帰分析のパラメータ a, b を求める解析処理モジュールを呼び出している。

$$a = \frac{Cov(x, y)}{Var(x)}$$

$$b = \bar{y} - a\bar{x}$$

なお、この解析処理モジュールは内部で複数の解析処理モジュールを組み合わせたモジュール (Sub-DAG) である。`loc` は指定された項目名を持つデータ列を取得するデータ解析モジュールである。

図 7 に本実行基盤において解析処理を実行した際の様子を示す。ポリシとしては図 4 を用いた。`[function-chain]` は呼び出された解析処理モジュールを示している。データ解析モジュールにおいて結果として $a = 0.5413\ldots, b = -0.9048\ldots$ が得られていることが分かる。なお、実行後にポリシを満足していない場合は、データ解析者に対してはエラーが返され、解析結果は破棄される。

処理に要した時間を 10 回実行時の平均で計測した。基盤起動から実行開始までに行うモジュールやランタイムの初期化に 2379 ミリ秒、解析処理の実行に 17 ミリ秒要した。また、基盤実行中に最大で 154MB のメモリが消費された。利用したゲスト用 Linux カーネルのサイズは 11.78MB、基盤ソフトウェアを含む initramfs のサイズは 44.9MB であった。初期化に時間を要することや使用メモリ量が大きい理由としては、モジュール毎に Wasm ランタイムを割り当てているため、モジュールの数だけランタイムが存在していることが原因として考えられる。一旦初期化したモジュールとランタイムは再利用可能であるため、必要に応じて再利用するといった最適化が必要である。また、モジュール間のデータやりとりについても、Columnar Format で保持しているが、全データをメモリ上に展開しているためメモリ消費量が増えている可能性がある。そのため、IPC Streaming Format [12] を用いて、データ全体をそのまま引数として渡すのではなく、ストリームとして扱うといった改善も考えられる。

4.3 扱う解析処理の粒度

本基盤で行う認可の粒度は、解析処理モジュールの実装に依存する。本論文における実装では、解析処理の対象を統計処理に限定し、解析処理モジュールを統計処理に用いる関数単位で実装した。しかし、データ解析者は単回帰分析や重回帰分析といった、複数の関数を組み合わせた解析処理を行う場合、処理の記述が手間となる。そのため、解析モジュールの処理をそのままデータ解析者に提供するのではなく、ある程度まとめた状態でデータ提供者に提供する必要がある。また、細かい単位で認可を行う場合、全体としてどのような処理を認可するのか不明瞭になる。そのため、一定の単位で解析処理を集約することが必要となるが、過剰な集約は後述するプライバシ保護に関する手法へ悪影響を与えうる。本基盤のアーキテクチャ自体は解析処理の粒度に対して制約を持たないが、認可を行いやすい解析処理の分割については議論の余地がある。

5 今後の展望

5.1 解析処理モジュールの検証

本基盤において、解析処理モジュールの追加や削除が容易な設計となっている。これにより、第三者が開発した解析処理モジュールを利用することができるようになる。しかし、解析処理モジュール自体が悪意ある第三者により開発されたものであった場合、認可自体が無効となる。例えば、重回帰分析のようなアルゴリズムを実装した解析処理モジュールが存在し、アルゴリズムを正しく実装していない場合、意図せぬ解析処理を許すことになる。そのため、解析処理モジュール自体が、対応するアルゴリズムに従って実装されているか検証する必要がある。与えられたアルゴリズム等の仕様に対して正しく実装が行われているか検証する手法の一つとして、コンコリックテスト [16] [17] が挙げられる。Wasm におけるコンコリックテストの手法である WASP [18] が提案されている。これらの自動検証手法を用いた解析処理モジュールの検証により、意図せぬ解析処理の抑止を目指す。

5.2 プライバシ保護

本基盤では、解析処理に対する認可を行うことで、データ提供者とデータ解析者間で柔軟なデータ解析の実現を目指した。しかし、認可する解析処理によってはデータの漏洩が生じる可能性がある。本基盤では、解析処理モジュール同士を DAG により指定された関係に従って呼び出して処理する。各解析処理モジュールや呼び出しの関係におけるデータ漏洩の度合いを評価することができれば、データ漏洩の度合いに応じた解析処理の認可やデータ提供の決定といったより柔軟なデータの提供、解析ができる。データ漏洩の解析手法としては、Information Flow Analysis による手法 [19] [20] [21] [22] が提案されている。これらの手法では、定義された規則に従いデータの漏洩を検出する。しかし、検出においてはソースコードが必要となるが、解析処理モジュールは Wasm モジュールである。そのため、Wasm モジュールを扱える Information Flow Analysis の手法が必要となる。

データ漏洩の度合いを定量的に評価するために、差分プライバシ [23] を用いる方針も考えられる。データ提供者が望む ϵ の閾値に達するような処理のみ認可するないしは、閾値を満足できるようにデータ解析者が実行した解析処理の後処理でノイズの付与を行うことで、データ提供者が提供する機微なデータがデータ解析者に対して漏洩することなしに解析処理を行うことができる。本基盤の認可においては、データに応じて特定の解析処理を強制することが可能である。解析処理として Laplace ノイズの付与や特定列の削除といった、データベースにおけるプライバシ保護手法を強制することで、プライバシ保護を達成することができる。本基盤に対して既存のプライバシ保護手法

を適用することで、プライバシを保護しつつより柔軟なデータの提供、解析の実現を目指す。

6 関連研究

6.1 TEE を用いたデータ解析アプリケーション

TEE の一種である Intel SGX を用いたデータベースアプリケーションとして、EnclaveDB [1] が知られている。EnclaveDB では、Intel SGX における暗号化された隔離領域である enclave を用いて、データベース上のデータやクエリに関する完全性や機密性を保証しつつ、既存のデータベースシステムと比較して小さいオーバーヘッドに抑えられている。また、専用のハードウェアを用いた TrustedDB [24] や Cipherbase [25] が提案されている。SQL を用いるデータベースに限らず、キーと値の組に基づくデータベースである Key-Value Store についても、同様の手法 [2] が提案されている。TEE では、メモリに対するアクセスパターンは隠蔽することは困難である。そこで、アクセスパターンを隠蔽する Oblivious RAM を用いた ObliDB [26] が提案されている。また、 TEE を用いることで、処理内容を秘匿したままデータ解析処理を行う手法として、マルチパーティの機械学習 [3] や、MapReduce を Intel SGX 環境で行う手法 [4] [27] [28] が提案されている。

データベースシステムに限らず、TEE 内で汎用にアプリケーションを動作させる手法として、Haven [29] や Graphane-SGX [30]、SCONE [31] が提案されている。Linux や Windows といった、既存の OS におけるアプリケーションに限らず、Wasm モジュールを TEE 内で動作させる手法として、SGX を対象とした Twine [32] や、Arm プロセッサにおける TEE である TrustZone を対象とし WaTZ [33] が提案されている。

これらは、TEE 内でアプリケーションを動作させることで機密性を保証しているが、暗号化されていない外部の世界とのデータのやり取りの制御については議論が十分ではない。そのため、本基盤における要素技術としては利用できるが、機微なデータに対する解析処理の認可や監査といった本基盤の要件を満足することはできない。本基盤では、要件である解析処理の認可や監査を実現するためのモジュールシステムを提案し、これらのモジュールにおいて既存の TEE 向けのデータベースシステムを利用するといった拡張が可能であると考えられる。

6.2 Personal Data Management System

Personal Data Management System (PDMS) [5] は、各個人に関するデータをプライバシを保護しつつ活用するために設計されたデータ管理システムである。Anciaux らは、PDMS における設計として、Extensive and Secure PDMS (ES-PDMS) [5] を提案した。ES-PDMS は、分離されたデータ収集やデータの処理を行うデータタスク群と、暗号化された通信や各データタスクに対するポリシの適用や監査、データストレージといった信頼されるコアモジュールから構成される。また、SGX に基づく ES-PDMS の実装 [34] や、各データタスクにおけるデータ漏洩を軽減する手法も提案 [35] されている。

これらは、本提案基盤と同様に、データの処理に対してデータに対するアクセス制御や監査を行うアーキテクチャとなっている。しかし、データの収集と解析がデータタスクとして同一のものとして扱われ、データ提供者とデータ解析者が明に分かれている状況は想定していない。そのため、データの提供者と解析者が分かれている場合、その二者間で実際どのように解析処理を認可し、どういったインタフェースでモジュールを構成するのかは明らかではなかった。データの所有者が PDMS のノードを持ち、それらに対して解析処理を依頼する分散 PDMS [36] も提案されている。データの所有者が PDMS のノードを持ち、データの解析者はそれらのノードに対してデータの処理を依頼する。アーキテクチャとしてはデータ提供者とデータ解析者が分離されているが、データ解析者が行う解析処理の認可や、処理の記述手法については議論がなされていない。本基盤では、ES-PDMS の設計に対して、より実用的で柔軟なデータ提供や解析処理を実現するために、データ解析者に応じたポリシの適用や、解析処理の認可や監査が容易かつ拡張が容易なモジュールシステムを提案している。また、モジュールシステムについては、既存のプログラミング言語や汎用なライブラリを活用を想定し、Wasm と Wasm におけるインタフェースに対して型定義を与える Component Model を用いてモジュールに対して明確な仕様を与えた。さらに、データの形式についても Apache Columnar Format を用いることで、より効率的な解析処理を目指した。

7 結　　論

本論文では、統計処理について統計処理単位で認可を行い、データに対する操作を制限するデータ流通処理基盤を提案した。本基盤は、解析処理単位で認可を行うことで、同じデータであっても解析者の属性に応じて解析者に対して与えうるデータの粒度や解析範囲の限定を目指す。これにより、簡単な解析はデータ提供者とデータ解析者間の合意なしに行いつつ、機微なデータが漏れうる場合は改めてデータ提供の合意を行うという段階的なデータ提供が実現できる。また、既に行われている合意に基づくデータ提供において、その合意が履行されているか確認する手法を提供する。解析者が行うデータ解析処理の各解析処理を記録しておき、データ提供者が必要に応じて記録の監査を行うことで、合意に対する違反の有無を確認することができ、不意の解析者側からのデータ漏洩を抑止することができる。AMD SEV-SNP を対象に本基盤の実装を行い、認可ポリシを適用しつつ、統計処理として単回帰分析を行うことができることを確認した。今後の展望として、第三者が提供する解析処理モジュールの自動検証や、プライバシの保護度合いに基づく解析処理の認可機構の開発を計画している。

謝　　辞

本研究は JSPS 特別研究員奨励費 23KJ1329 の助成を受けたものです。

文　献

[1] Priebe, C., Vaswani, K. and Costa, M.: EnclaveDB: A Secure Database Using SGX, *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 264–278 (2018).

[2] Li, K., Tang, Y., Zhang, Q., Xu, J. and Chen, J.: Authenticated Key-Value Stores with Hardware Enclaves, *Proceedings of the 22nd International Middleware Conference: Industrial Track*, Middleware '21, Association for Computing Machinery, p. 1–8 (2021).

[3] Ohrimenko, O., Schuster, F., Fournet, C., Mehta, A., Nowozin, S., Vaswani, K. and Costa, M.: Oblivious Multi-Party Machine Learning on Trusted Processors, *25th USENIX Security Symposium (USENIX Security 16)*, USENIX Association, pp. 619–636 (2016).

[4] Pires, R., Gavril, D., Felber, P., Onica, E. and Pasin, M.: A Lightweight MapReduce Framework for Secure Processing with SGX, *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 1100–1107 (2017).

[5] Anciaux, N., Bouganim, L., Pucheral, P., Popa, l. S. and Scerri, G.: Personal Database Security and Trusted Execution Environments: A Tutorial at the Crossroads, *Proc. VLDB Endow.*, Vol. 12, No. 12, p. 1994–1997 (2019).

[6] W3C Web Assembly Community Group: WebAssembly, `https://webassembly.org/`.

[7] W3C Web Assembly Community Group: Component Model design and specification, `https://webassembly.org/`.

[8] Advanced Micro Devices, Inc.: AMD SEV-SNP: Strengthening VM Isolation with Integrity Protection and More, `https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more.pdf` (2020).

[9] Wooldridge, J. M.: Introductory Econometrics: A Modern Approach (7th Edition) (2020).

[10] 千田浩司，前田陽二，辻　秀一集合匿名化クラウドの課題と対策，情処学研報，Vol. 2011, No. 21, pp. 1–6（オンライン），https://cir.nii.ac.jp/crid/1570009752137628288 (2011).

[11] W3C Web Assembly Community Group: WASI — The WebAssembly System Interface, `https://wasi.dev/`.

[12] Apache Software Foundation: Arrow Columnar Format, `https://arrow.apache.org/docs/format/Columnar.html`.

[13] Apache Software Foundation: Apache Arrow, `https://arrow.apache.org/`.

[14] Vink, R. et al.: Polars, `https://pola.rs/`.

[15] HARUYAMA, T.: Wooldridge Meets Python, `https://github.com/spring-haru/wooldridge` (2021).

[16] Sen, K., Marinov, D. and Agha, G.: CUTE: A Concolic Unit Testing Engine for C, *Proceedings of the 10th European Software Engineering Conference Held Jointly with 13th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, Association for Computing Machinery, p. 263–272 (2005).

[17] Godefroid, P., Klarlund, N. and Sen, K.: DART: Directed Automated Random Testing, *SIGPLAN Not.*, Vol. 40, No. 6, p. 213–223 (2005).

[18] Marques, F., Santos, J. F., Santos, N. and Adão, P.: Concolic Execution for WebAssembly, *European Conference on Object-Oriented Programming* (2022).

[19] Myers, A. C.: JFlow: Practical Mostly-Static Information Flow Control, *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '99, Association for Computing Machinery, p. 228–241 (1999).

[20] Sabelfeld, A. and Myers, A.: Language-based information-flow security, *IEEE Journal on Selected Areas in Communications*, Vol. 21, No. 1, pp. 5–19 (2003).

[21] Sun, M., Wei, T. and Lui, J. C.: TaintART: A Practical Multi-Level Information-Flow Tracking System for Android RunTime, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, Association for Computing Machinery, p. 331–342 (2016).

[22] Zhang, R., Zhang, N., Moini, A., Lou, W. and Hou, Y. T.: PrivacyScope: Automatic Analysis of Private Data Leakage in TEE-Protected Applications, *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pp. 34–44 (2020).

[23] Dwork, C.: Differential Privacy, *Automata, Languages and Programming* (Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 1–12 (2006).

[24] Bajaj, S. and Sion, R.: TrustedDB: A Trusted Hardware-Based Database with Privacy and Data Confidentiality, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 3, pp. 752–765 (2014).

[25] Arasu, A., Blanas, S., Eguro, K., Joglekar, M., Kaushik, R., Kossmann, D., Ramamurthy, R., Upadhyaya, P. and Venkatesan, R.: Secure Database-as-a-Service with Cipherbase, *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, Association for Computing Machinery, p. 1033–1036 (2013).

[26] Eskandarian, S. and Zaharia, M.: An Oblivious General-Purpose SQL Database for the Cloud (2017).

[27] Schuster, F., Costa, M., Fournet, C., Gkantsidis, C., Peinado, M., Mainar-Ruiz, G. and Russinovich, M.: VC3: Trustworthy Data Analytics in the Cloud Using SGX, *2015 IEEE Symposium on Security and Privacy*, pp. 38–54 (2015).

[28] Dinh, T. T. A., Saxena, P., Chang, E.-C., Ooi, B. C. and Zhang, C.: M2R: Enabling Stronger Privacy in MapReduce Computation, *24th USENIX Security Symposium (USENIX Security 15)*, USENIX Association, pp. 447–462 (2015).

[29] Baumann, A., Peinado, M. and Hunt, G.: Shielding Applications from an Untrusted Cloud with Haven, *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, USENIX Association, pp. 267–283 (2014).

[30] che Tsai, C., Porter, D. E. and Vij, M.: Graphene-SGX: A Practical Library OS for Unmodified Applications on SGX, *2017 USENIX Annual Technical Conference (USENIX ATC 17)*, USENIX Association, pp. 645–658 (2017).

[31] Arnautov, S., Trach, B., Gregor, F., Knauth, T., Martin, A., Priebe, C., Lind, J., Muthukumaran, D., O'Keeffe, D., Stillwell, M. L., Goltzsche, D., Eyers, D., Kapitza, R., Pietzuch, P. and Fetzer, C.: SCONE: Secure Linux Containers with Intel SGX, *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association, pp. 689–703 (2016).

[32] Menetrey, J., Pasin, M., Felber, P. and Schiavoni, V.: Twine: An Embedded Trusted Runtime for WebAssembly, *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE Computer Society, pp. 205–216 (2021).

[33] Menetrey, J., Pasin, M., Felber, P. and Schiavoni, V.: WaTZ: A Trusted WebAssembly Runtime Environment with Remote Attestation for TrustZone, *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, IEEE Computer Society, pp. 1177–1189 (2022).

[34] Carpentier, R., Thiant, F., Sandu Popa, I., Anciaux, N. and Bouganim, L.: An Extensive and Secure Personal Data Management System Using SGX, *EDBT 2022 - 25th International Conference on Extending Database Technology* (2022).

[35] Carpentier, R., Sandu Popa, I. and Anciaux, N.: Data

Leakage Mitigation of User-Defined Functions on Secure Personal Data Management Systems, *SSDBM 2022 - 34th International Conference on Scientific and Statistical Database Management* (2022).

[36] Ladjel, R., Anciaux, N., Pucheral, P. and Scerri, G.: A manifest-based framework for organizing the management of personal data at the edge of the network, *ISD 2019 - 28th International Conference on Information Systems Development* (2019).

階層構造・多重解像度による差分プライベートな深層人流生成

高木　駿† 加藤　郁之† 曹　洋†† 吉川　正俊†††

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町 36-1
†† 北海道大学大学院情報科学研究院　〒 060–0814 札幌市北区北 14 条西 9 丁目
††† 大阪成蹊大学データサイエンス学部　〒 533–0007 大阪市東淀川区相川 3 丁目 10-62
E-mail: †takagi.shun.45a@st.kyoto-u.ac.jp, ††fumiyuki@db.soc.i.kyoto-u.ac.jp, †††yang@ist.hokudai.ac.jp,
††††yoshikawa-mas@osaka-seikei.ac.jp

あらまし　人流データは都市計画やパンデミック対応など多くのアプリケーションにおいて貴重な洞察を提供するが，その使用はプライバシに関する懸念を引き起こす．本論文では，差分プライバシを保証しながらリアルな人流データを合成するために特別に設計された，階層的かつ多重解像度を用いた深層学習モデル（HRNet）を提案する．まず，差分プライバシの下で人流データを深層学習する際の主要な課題を特定する．HRNet はこれらの課題を三つの新しいコンポーネント（階層的な位置埋め込み，多重解像度におけるマルチタスク学習，およびプライベートな事前学習）を用いて緩和する．最後に，実世界のデータセットを使用した広範な比較実験を通じて，HRNet は有用性とプライバシのトレードオフが既存研究よりも優れていることを示す．

キーワード　差分プライバシ, 深層学習, 人流データ, 生成モデル

1　はじめに

近年，人流データ活用が，交通予測，都市計画，パンデミック対応（COVID-19 拡散分析を含む）等の様々な社会的利益に寄与するとして注目を集めている．しかし，それらの利用において，データが集約され匿名化されていても，重大なプライバシの懸念が生じることが知られている [23], [27].

差分プライバシ（DP）[9] は，データプライバシ定義の標準として注目されており，潜在的な敵に関する制約的な仮定に依存せずに，数学的に根拠のある強固なプライバシ保証を提供する．DP の主要原理は，データ解析からの出力が，特定の個人のデータが含まれているか除外されているかに関わらず，大きく異ならないことを保証することである．米国国勢調査局や Google などの IT 企業による DP を用いたプライバシ保護型データ活用の実世界での広範な適用例は，その重要性と効果を示している [8], [12], [16], [25].

本論文では，人流データの DP データ合成について研究する．DP データ合成は，DP を保証しながら実データと統計的に類似した合成データを生成する手法である．DP の事後処理定理によれば，この合成データに対して無制限の分析を行っても，さらなるプライバシ漏洩は生じないことが保証される．本研究の目的は，与えられた人流データに似た，前述のアプリケーションでの様々な下流タスクに使用することが可能な擬似的な人流データを生成することである．

既存の人流データの DP データ合成手法は，DP に対応したデータ構造に主に依存しており，ツリーベース [5], [6], [14]，マルコフベース [13], [28]，クラスタリングアプローチ [15], [19], [20] に分類される．これらの手法は一般に DP を効率的に保証するために，次元削減と感度削減を行うが，情報損失のコストが伴う．例えば，クラスタリングアプローチは，位置の一般化を使用してクラスタリングにより k の重心間の遷移のみを生成し，重心内の正確な遷移情報を失う．最先端のマルコフベース手法 [28] は，2 ステップ間隔を超える相関を考慮していないし，多くの手法は，人間の動きにおいて重要な時間などの補助情報を失っている．ツリーベースの手法は，プライバシ予算の一部を使用してデータ依存のツリーを構築することで，このような損失を最小限に抑えようとする．しかし，このような消費は，データ生成フェーズでのプライバシ予算の不足につながり，結果として効用が低下する．

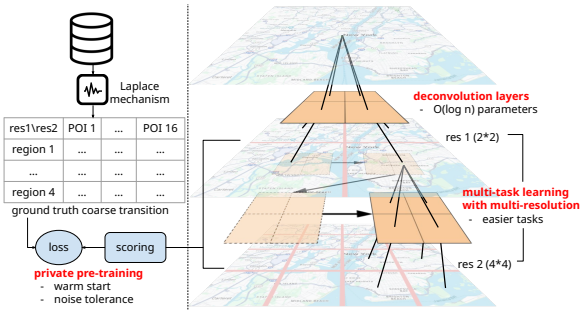


図 1: HRNet は人流データ生成における DP-SGD の適用のボトルネックに対処するために，3 つの新しい要素を活用する: 1) 転置畳み込みネットワークを用いた階層的な位置エンコーディング機構，2) 複数の解像度にまたがるマルチタスク学習，および 3) DP 粗遷移行列を用いたプライベートな事前学習．

Wang らによって提案された MTNet [30] は，DP を保証しながら道路網上の短い移動（例えばタクシーの移動）を生成するために設計された深層学習アプローチである．道路網のメタ情報を利用して，MTNet は有用性とプライバシのトレードオフを巧みに管理する．しかし，このアプローチは，移動を道路網

に制限するという固有の制約を課す．実際の人流は，道路網に限定されず，地下鉄の移動や公園などの道路外の空間を通る移動も含む．さらに，道路網に焦点を当てたとしても，人流はしばしば多くの道路セグメントを含む長いデータとなり，学習の難しさという大きな課題に当たってしまう．そのため，MTNetをより幅広い人流データに適応させることは，有望ではないように思われる．

我々の知る限り，この研究は道路網に制約されない人流を生成するために，DP を用いた深層学習アプローチを採用する最初の論文である．最近の研究により，人流データを扱う深層学習アプローチが古典的なモデルよりも優れていることが示されている [22], [29], [33]．GAN ベースや拡散モデルなどのリアルな合成人流を生成するためのいくつかの生成モデルが存在するが，これらのモデルは正式なプライバシ保証を確保していない．理論的には，最適化手法を Differentially Private Stochastic Gradient Descent（DP-SGD）[1] に置き換えることで，任意の深層学習モデルに DP を保証することが可能である．残念ながら，これらの生成モデルに DP-SGD を単純に適用すると，モデルのパラメータ数と複雑さのために，効用が低下するか，プライバシのコストが非現実的に高くなる．具体的には，次の 2 つの主要な課題がある．

（1） **モデルパラメータの数：**DP-SGD は，モデルパラメータ数が増えると経験的リスクの下限が必然的に増加することが示されている [3], [4]．人流に関するほとんどのモデル（例えば，RNN モデル [32] や注意モデル [31]）では，モデルパラメータの数は位置の数 (POI) に依存する．埋め込み行列や線形埋め込みなどのエンコーディングスキームを使用して POI をエンコードする場合，モデルパラメータの数は POI の数 n_{POI} に対して $O(n_{\mathrm{POI}})$ となる．つまり，POI の数に関して，このモデルは精度が悪くなる．

（2） **学習の困難さ：**POI の数 n_{POI} の増加に伴い，訓練のためのラベルの数が増えるため，人流の学習の困難さが増す．したがって，より大きな n_{POI} は，微妙な POI の違いを識別するためにより大きなモデルとより長い訓練エポックを必要とする．これにより，多くのパラメータと合成定理 [18] によるプライバシ予算の高い消費のために効用が減少する．

これらの課題に対応するため，我々は新しいネットワーク構造と学習方法を備えた階層的かつ多重解像度を用いた深層学習モデル（HRNet）を提案する．HRNet の主な特徴は，図 1 に示されているように，次の 3 つである．

（1） **階層的な位置埋め込み：**第 1 の課題を緩和するため，HRNet は転置を使用した階層的な構造による位置埋め込みを採用する．このアプローチにより，パラメータの数を $O(\log n_{\mathrm{POI}})$ に大幅に削減し，従来の埋め込み方法が有する $O(n_{\mathrm{POI}})$ のパラメータに関連する効用の減少を緩和する．

（2） **多重解像度におけるマルチタスク学習：**第 2 の課題を緩和するために，マルチタスク学習を導入する．我々の階層的ネットワーク設計により，複数の解像度のエンコーディングが可能となる．単純なサブタスクによって，大量の POI に関連する学習の困難さを効率的に軽減する．

（3） **効果的かつプライベートな事前学習：**両方の課題を緩和するため，プライベートな事前学習を行う．DP-SGD の第 1 の課題への対処において，事前学習の有効性が強調されている [2], [17]．さらに，事前学習は効果的な「ウォームスタート」を提供し，広範な訓練エポックの必要性を軽減する．しかし，事前学習の一般的な課題は，公開データが必要であるということである．HRNet は，密な DP 粗遷移行列を活用することで，公開データなしで効率的な事前学習を可能にする．

要約すると，HRNet は DP-SGD を用いた人流生成における POI の数によってもたらされる課題を解決した深層生成モデルである，それは，階層的な位置エンコーディング，多重解像度でのマルチタスク学習，そしてプライベートな事前学習によって特徴付けられている．最後に，我々の実験は，モデルが既存の方法と比較して有用性とプライバシのトレードオフにおいて優れていることを示す．

2 準　　備

2.1 差分プライバシ

差分プライバシ（DP）は，Dwork らによって導入された [9]，統計値の公開におけるプライバシ漏洩を定量化するための数学的な定義である．DP の定義は以下の通りである．

Definition 1 ((ε,δ)-差分プライバシ)**.** データセットドメイン $\mathcal{D}$ と出力ドメイン $\mathcal{Z}$ を考慮する．$\varepsilon \in \mathbb{R}^+$ および $\delta \in [0,1]$ が与えられたとき，ランダムなメカニズム m は，入力 $D \in \mathcal{D}$ に対してランダムに $m(D) \in \mathcal{Z}$ を出力し，任意の *2* つのデータセット $D, D' \in \mathcal{D}$ が最大で *1* 要素異なる場合，および任意の出力サブセット $Z \subseteq \mathcal{Z}$ に対して，以下の不等式が成立する場合，*(ε,δ)-DP* を満たすと言う：

$$\Pr[m(D) \in Z] \leqq e^{\varepsilon} \Pr[m(D') \in Z] + \delta.$$

DP-SGD：Abadi らによって開発された Differentially Private Stochastic Gradient Descent（DP-SGD）[1] は，従来の SGD 最適化プロセスを DP に準拠するように適応させる．従来の SGD では，サンプルデータ（ミニバッチ）を使用した勾配計算によってモデル θ のパラメータを反復的に更新し，経験的リスクを最小化する．DP-SGD は，勾配のクリッピングとノイズを加えるステップを導入することで DP を保証する．具体的には，まず l_2 ノルムで勾配をクリップして感度を制限し，次に平均クリップされた勾配にガウスノイズを加える．モデルパラメータは，このランダム化された勾配を使用して更新され，従来の SGD と同様に続けられる．このプロセスは，収束するかプライバシ予算が尽きるまで続けられる．合成技術を用いることで，最終的に公開されるモデルは DP に準拠している．

2.2 問題の定式化

本研究の目的は，実際の人流（例えば，家から職場への日常的な移動など）に似た人流データを生成することである．人流の本質的に連続的な性質（時間および地理座標（緯度と経度）の両方において）により，直接的なモデリングと評価は難しい．

これに対処するため，我々は人流データを離散化する従来のアプローチを採用し，モデリングプロセスを単純化する．この章では，まず離散化された人流表現を定式化し，次に離散化された人流のためのジェネレータを定式化し，評価方法を概説する．

2.2.1 軌　　跡：

基本的なステップとして，我々は POI を地図上に一様に分布するグリッドセルと定義する．各グリッドセルには一意の整数 ID が割り当てられる．たとえば，対象マップが $w \times w$ のグリッドに分割されている場合，POI のドメインは $L := [w^2]$ と定義される．ここで $[n]$ は集合 $(1, 2, \ldots, n)$ を示す．つまり，$n_{\text{POI}} = |L|$ である．図 1 に示されているように，異なるレベルの離散化が可能である，例えば 1×1，2×2，4×4 などである．地理的な位置（緯度と経度で定義される）は，その位置を含むグリッドセルを表す整数に変換される．マッピングは，$(\lfloor l/w \rfloor, l\%w)$ に位置するグリッドに整数 $l \in L$ が割り当てられるように行われる．次に，人流シーケンスを滞在点のシーケンスとして表現する．$\mathbf{v} = [v_1, v_2, \ldots]$，ここで $i \in \mathbb{N}, v_i = (l_i, t_i)$，$l_i \in L$ である．この表現は，$l_i \neq l_{i+1}$ であるように前の POI を変更する POI を整列させる．ここで，t_i は対応する時間を含むタイムスロットのインデックスである．タイムスロットは，n_{time} によって全時間を離散化することによって作られる．ここで，n_{time} はタイムスロットの数である．たとえば，v_2 は訪問された 2 番目の POI とこの POI に到達した時間のタプルを示す．この時点から,「軌跡」という用語は人流のこの離散化された表現を指す．

2.2.2 ジェネレータ：

我々の研究では，ニューラルネットワークベースのジェネレータを使用し，合成軌跡を確率的に生成する．このプロセスは数学的に $\mathbf{v} \sim G_\theta$, と表現される．ここで $\mathbf{v}$ は生成された合成軌跡を表し，θ はニューラルネットワークの訓練可能なパラメータを表す．ジェネレータの具体例，そのアーキテクチャは 2.3 章で説明する．

2.2.3 評　　価：

G_θ のパフォーマンスの直接的な評価は軌跡データが本質的に疎であるために難しい．これを回避するため，我々は軌跡のさまざまな統計的性質に焦点を当てた間接的な評価アプローチを採用する．これらの統計的性質は，G_θ によって生成されたデータセットから計算され，提供された軌跡データセット D から導出された経験的統計と比較される．我々が検討する軌跡データの統計値として，経由点，ルート，目的地，遷移，移動距離が含まれる．これらの特定の統計を分析することで，G_θ によって生成された軌跡の正確性を得ることができる．これらの統計の形式的な定義と計算方法は，5.1 章で詳細に説明される．

2.3 ベースラインジェネレータ

この章は，この研究のベースラインとなるジェネレータのアーキテクチャを詳述する．

2.3.1 ベースラインジェネレータ

ベースラインジェネレータは，条件付き確率の連鎖則に基づいて構築されている：

$$G_\theta = \Pr(\mathbf{v}) = \prod_{i=1}^{|\mathbf{v}|} \Pr(v_i | v_1, ..., v_{i-1}). \tag{1}$$

この定式化では，各条件付き確率 $\Pr(v_i|v_1, \ldots, v_{i-1})$ によってジェネレータが $i = 1$ から順に軌道を構築する．この目的のために，ジェネレータはデータポイント v_i の埋め込み行列，シーケンスを処理するリカレントニューラルネットワーク（RNN），そして分布を生成するスコアリングコンポーネントを統合する．

埋め込み行列：埋め込み行列は，各データポイント $v_i = (l_i, t_i)$ をベクトル表現に変換する．POI（l_i）用と時間（t_i）用の 2 つの別々の埋め込み行列が使用される．POI 埋め込み行列 M_{POI} には $|L|$ の訓練可能なベクトルが含まれ，時間埋め込み行列 M_{time} には n_{time} の訓練可能なベクトルが含まれている．v_i のベクトル表現は，時間埋め込み行列から t_i 番目のベクトルと POI 埋め込み行列から l_i 番目のベクトルを連結して形成される：

$$\text{encode}(v_i) = [M_{\text{time}}[t_i], M_{\text{POI}}[l_i]].$$

リカレントニューラルネットワーク：軌跡プレフィックス $(v_1, \ldots, v_{i-1})$ のエンコーディングには，Gated Recurrent Units（GRU）[7] を用いる．これは，DP-SGD の文脈で GRU が単純なアーキテクチャと少ないパラメータ数によって LSTM や注意メカニズムなどの代替方法よりもわずかに効果的であると経験的に見出されたためである．. GRU セルは，前の状態 h_{i-1} とエンコードされた v_{i-1} に基づいてその状態を更新する：

$$h_i = f_{\text{GRU}}(h_{i-1}, \text{encode}(v_{i-1})).$$

このプロセスは，軌跡プレフィックス全体をエンコードするために再帰的に適用される．

スコアリングコンポーネント：スコアリングコンポーネントは，エンコードされたプレフィックスを $[n_{\text{POI}}] \times [n_{\text{time}}]$ 上の確率分布に変換する．この変換は，フィードフォワードニューラルネットワーク $g_{\text{POI}} : \mathbb{R}^{n_{\text{hidden}}} \to \mathbb{R}^{|L|}$ および $g_{\text{time}} : \mathbb{R}^{n_{\text{hidden}}} \to \mathbb{R}^{n_{\text{time}}}$ を使用して行われ，ソフトマックス関数によって適切な確率的正規化が保証される．隠れ状態 h_i に基づいて空間位置 $l \in L$ の確率分布は $\Pr(l = l_i|h_i) = \text{softmax}(g_{\text{POI}}(h_i))[l_i]$ として計算される．同様に，h_i に基づいて時間 $t \in [n_{\text{time}}]$ の確率分布は $\Pr(t = t_i|h_i) = \text{softmax}(g_{\text{time}}(h_i))[t_i]$ によって決定される．最終的に，前の観測値に基づく条件付き確率 $\Pr(v_i|v_1, \ldots, v_{i-1})$ はこれら 2 つの分布の積であり，POI と時間の結合確率を表す: $\Pr(v_i|v_1, \ldots, v_{i-1}) = \Pr(l = l_i|h_i) \Pr(t = t_i|h_i)$.

2.3.2 目的関数：

ベースラインジェネレータの最適化のために，与えられた軌跡シーケンス $\mathbf{v} = (v_1, \ldots, v_i)$ に対して，クロスエントロピー損失を目的関数として使用する．損失は以下のように計算される：

$$\text{loss}_{\text{POI}} + \text{loss}_{\text{time}} =$$

$$\sum_{i=1}^{|v|} \left(\sum_{c \in L} \delta_{c,l_i} \log(\Pr(l = l_i|h_i)) + \sum_{c \in [n_{\text{time}}]} \delta_{c,t_i} \log(\Pr(t = t_i|h_i)) \right),$$

ここで $\delta_{i,j}$ はクロネッカーのデルタである．

2.4 2つのボトルネック

先に述べたベースラインモデルは，DP-SGD が適用されたとき，特に POI の数が増えるにつれて性能が低下する．この章では，これらの性能のボトルネックを引き起こす 2 つの主要な要因を特定し，詳述する．

2.4.1 パラメータ数 $|\theta|$

DP の制約の下では，モデルのパラメータ数が増えるにつれて，プライベートモデルの経験的リスクの下限が多項式的にスケールすることが知られている [3], [4]．ベースラインモデルでは，位置埋め込み行列とスコアリングコンポーネントのために，パラメータベクトル θ のサイズは POI 数（n_{POI}）の増加に伴い線形に拡大する．その結果，n_{POI} が増えるにつれて経験的リスクが多項式的に増加し，生成された人流の有用性が低下する．

2.4.2 ラベルの数

クラスラベルの数（すなわち，n_{POI}）が増加すると，モデルはクラス間のより微妙な特徴の違いを識別する必要がある（2.3.2 章を参照）．これはしばしば，より大きく複雑なモデルパラメータ数や長い訓練エポックの使用を必要とする．しかし，このような戦略は DP-SGD の制約の下では実用的でない．これは，モデルのパラメータ数の増加が上述の [3], [4] によって経験的リスクの増加を引き起こすためである．また，訓練エポック数を増やすことは，DP の合成定理 [18] によるプライバシの損失の増大をもたらす．したがって，多数のラベルによる学習の難しさは DP-SGD のボトルネックとなる．

2.5 転置畳み込み

本章では，提案モデルの主要な構成要素である転置畳み込み操作について詳述する．コンピュータビジョンの分野から発祥した転置畳み込み層（別名デコンボリューション層）は，通常，畳み込み層によって生成された特徴マップから元の画像をアップサンプルまたは再構築するために使用される [21]．従来の使用法とは対照的に，HRNet ではこの層の特定のインスタンスを独自に組み込んでおり，これを「四倍転置畳み込み」と呼んでいる．この転置畳み込み層の組み込みは，以前に特定した 2 つの主要なボトルネックに対処するために重要な役割を果たす．

四倍転置畳み込み

(w, w, n_{dim}) の次元を持つ行列 M を考える．ここで n_{dim} はチャネル数を表し，各チャネルは $w \times w$ の正方行列で構成される．また，$(2, 2, n_{\text{dim}})$ の次元の行列としてカーネルを定義する．これには n_{dim} の二次元カーネルが含まれ，それぞれの二次元カーネルは 2×2 行列（kernel_k）である．四倍転置畳み込みは，カーネルを使用して行列 M を $(2w, 2w, n\text{dim})$ の次元を持つ行列 M' に拡張する．図 1 での実行例を参照すると，第二層では，カーネルは (2×2) 行列の入力をスライドし，4 つのグリッドのためにデコンボリューション操作を行い，2×2 行列を 4×4 行列に拡張している．

x および y を整数とし，変換は次のように数学的に表現される：

$$M'[2x + \delta_x, 2y + \delta_y, k] = \sum_{k'=1}^{n_{\text{dim}}} \text{kernel}_k[\delta_x, \delta_y, k'] M[x, y, k'],$$

ここで，$k \in [n_{\text{dim}}]$，δ_x および δ_y は二進値（0 または 1）であり，記法 $M[i, j, k]$ は行列 M のインデックス (i, j, k) にある要素へのアクセスを表す．このプロセスは，行列内の各 n_{dim} 次元ベクトルを四倍にする効果がある．その結果，元の n_{dim} 次元ベクトル $M[x, y]$ は，四つの異なる n_{dim} 次元ベクトルに拡張される：$M'[2x, 2y]$，$M'[2x + 1, 2y]$，$M'[2x, 2y + 1]$，および $M'[2x + 1, 2y + 1]$．四倍転置畳み込み層を d 回適用すると，n_{dim} 次元ベクトルの数は 4^d 倍になる．w が 2 のべき乗である場合，単一の n_{dim} 次元の根元ベクトル θ_{root} から始まり，ネットワークの最終変換は次のように表される：

$$M'[2x + \delta_x, 2y + \delta_y, k] = \sum_{k'=1}^{n_{\text{dim}}} \text{kernel}_k[\delta_x, \delta_y, k'] M[x, y, k'],$$

ここで deconv^a は，デコンボリューション操作を a 回反復適用することを示す．

3 HRNet

この章では，階層的かつ多重解像度の深層学習モデル（HRNet）を説明する．HRNet の主な新規性は，従来の埋め込み行列とスコアリングコンポーネントを置き換える階層的位置埋め込みコンポーネントの統合である．この設計により，モデルパラメータ数が減少し，第一のボトルネックを効果的に緩和する．さらに，多重解像度の解釈による，マルチタスク学習を可能にする．このアプローチにより，HRNet は複数の解像度で同時にデータを処理・学習し，最高解像度での学習の難しさを巧みに取り扱うことができ，第二のボトルネックを緩和する．

まず，図 2 に示す簡単な例を用いて HRNet を概説する．その後，各コンポーネントの詳細な説明に進む．

3.1 概　　要

図 2 に示される HRNet のマルチタスクトレーニングの概要を提供する．HRNet は，条件付き確率をモデル化するベースラインモデルをベースに構築されている．ベースラインモデルの詳細については，2.3 章を参照．ここでは，主要な違いを強調するために，ベースラインと同じ時間情報のトレーニングを省略し，説明を簡素化する．

図 2 に示されるシナリオを考える．この例では，それぞれ 4 つと 16 つのグリッドセルを持つ 2 つの解像度があり，解像度 2 での軌跡は $(1, 2, 6, 10)$ である．$v = (1, 2, 6, 10)$ の学習フェーズで，特にステップ 3（ステップ 1 はプレフィックス（1）から 2 を学習し，ステップ 2 はプレフィックス $(1, 2)$ から 6 を学習し，ステップ 3 はプレフィックス $(1, 2, 6)$ から 10 を学習する）で，HRNet は次のように動作する：

階層的位置埋め込みコンポーネントが，位置 6 をエンコードする．GRU セルは，h_2 と 6 のエンコードされたベクトルを使用して，プレフィックス $(1, 2, 6)$ の埋め込みである h_3 を計算す

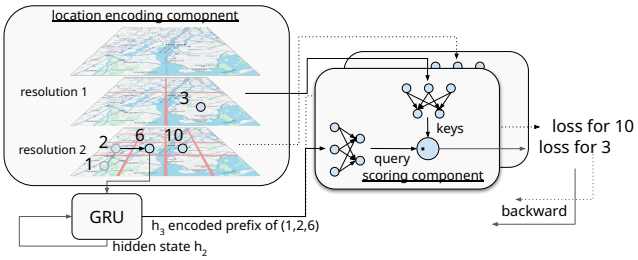


図 2: 階層的位置エンコーディングを用いたマルチタスクトレーニングの概要．この例では $w = 4$ と仮定し，階層的位置エンコーディングのため解像度 2 では $4 * 4$ の POI，解像度 1 では $2 * 2$ の POI がある．プレフィックス $(1, 2, 6)$ が与えられた場合，モデルは解像度 2 でグリッドセル 10 を，解像度 1 でグリッドセル 3 を推測することを学ぶ．

る．h_3 は，後続のステップのためのフィードフォワードニューラルネットワークとともにクエリベクトル queyr を形成する．ベースラインモデルが解像度 2 にのみ焦点を当てるのに対し，HRNet は解像度 1 も同時に考慮し，つまり $(1, 2, 6)$ から解像度 1 で 3 を学習する．各解像度について，モデルはすべてのグリッドセルをエンコードし，それらを別々にフィードフォワードニューラルネットワークでキーベクトル key_l に変換する．次に，クエリベクトルとキーベクトルのドット積を行い，プレフィックス $(1, 2, 6)$ を与えられた次の位置のスコアを計算する．この操作により，各解像度（すなわち，[4] と [16]）上の確率分布がソフトマックス関数を使用して生成される．モデルは，この確率分布と実際の値，すなわち解像度 1 での 3，および解像度 2 での 10 に基づいて交差エントロピー損失を計算する．最後に，モデルは合計された交差エントロピー損失からパラメータ θ の勾配を計算し，DP-SGD を用いてモデルを更新する．

3.2 階層的位置エンコーディング

HRNet は，ベースラインモデルよりも少ないパラメータ数を持つ新しい位置エンコーディングメカニズムを導入している．その主な特徴は，2D クワッドデコンボリューショナルネットワークを用いた階層的ネットワーク構造の採用である．

3.2.1 構　　造

HRNet の位置エンコーディングの核心は，訓練可能な根元ベクトルであり，θ_{root} として表す．この根元ベクトルは，d 層から成る四倍転置畳み込み層によって拡張される．つまり，θ_{root} に四倍転置畳み込み層を適用し，4^d の異なるベクトルを生成する．本質的に，この操作により $(2^d, 2^d, n_{\text{dim}})$ の次元を持つ埋め込み行列が生成される：

$$M_d = \text{deconv}^d(\theta_{\text{root}}).$$

deconv 操作の詳細な説明については，2.5 章を参照．特筆すべきは，各転置畳み込み層のパラメータ数はカーネルサイズにのみ依存し，n_{POI} とは独立しているため，パラメータ数は一定である．したがって，このアーキテクチャは n_{POI} ベクトルを生成するために，通常の埋め込み行列で必要な $O(n_{\text{POI}})$ パラメータと比較して，$O(d) = O(\log n_{\text{POI}})$ パラメータのみを必要とする．

3.2.2 ベクトルの割り当て

ここでは，階層的位置埋め込みコンポーネントによって生成された 4^d ベクトルを n_{POI} の POI に割り当てる方法を説明する．POI のグリッドサイズ $w * w$ が 4 のべき乗である場合，$d = \log_4(w * w)$ と設定し，次のように割り当てる．

$$\text{hiencode}(l) = M_d[\lfloor l/w \rfloor, l\%w],$$

ここで $l \in L$ である．この割り当て方法は，地理的に近い場所は似た埋め込みを持つという地理の第一法則 [26] を自然に取り入れている．この近接性の原則は，デコンボリューション操作の直接的な結果であり，単一の親ベクトルが 4 つの隣接する埋め込みを生成する．

3.3 スコアリング

HRNet は，階層的位置埋め込みとドット積演算を組み合わせることで，POI 上の確率分布を作る効率的なスコアリングメカニズムを採用している．

3.3.1 スコアリングプロセスの定式化

エンコードされたプレフィックスシーケンス $(v_1, \ldots, v_{i-1})$（つまり，h_{i-1}）はクエリベクトルに変換される．この変換は，f_{query} として表されるフィードフォワードニューラルネットワークを使用して実行される．同時に，エンコードされた POI，$\text{hiencode}(l)$ は，f_{key} という別のフィードフォワードニューラルネットワークを介してキーベクトルに変換される．数学的には，次のように表される：$\text{query} = f_{\text{query}}(h_i)$ および $\text{key}_l = f_{\text{key}}(\text{hiencode}(l))$．これらの変換に続いて，$l$ の v_i としてのスコアは，クエリとキーベクトルの間のドット積を計算することによって決定される：$\text{score}_l = \text{query} \cdot \text{key}_l$．最後のステップは，すべての POI にわたってこれらのスコアにソフトマックス関数を適用し，POI 上の確率分布を生成することである：

$$\Pr(v_i = l_i | v_1, \ldots, v_{i-1}) = \text{softmax}(\text{score}_L)[l_i],$$

ここで $\text{score}_L = (\text{score}_1, \ldots, \text{score}_{|L|})$ である．

3.4 多重解像度学習

この章では，大量のラベルに関連する学習の困難さという課題に対処するために，目的関数を再定義する．解決策は，階層的位置埋め込みコンポーネントによって可能になる多重解像度解釈を採用することにある．この多重解像度戦略には，元々のタスクだけでなく，より粗い解像度でのタスクも同時に考慮する．モデルはより粗い解像度での簡単なタスクを解くことで DP の条件下でもデータの広範な理解を得ることができる．

3.4.1 多重解像度解釈

ここでも，$w \times w$ のグリッドセル数が 4 のべき乗であるという仮定する．これにより，$2^d \times 2^d$（2.2 章で説明されている）の主要なグリッド分割を超える多重解像度解釈の基盤となる．

各解像度レベル $i_{\text{res}} \in \{1, 2, \ldots, d-1\}$ での $2^{i_{\text{res}}} \times 2^{i_{\text{res}}}$ のグリッド分割を検討する．ここで，i_{res} 番目の解像度には，$L_{i_{\text{res}}}$ のグリッドセルが含まれる：

$$i_{\text{res}} \in [d], L_{i_{\text{res}}} := [4^{i_{\text{res}}}].$$

3.2.1 章の行列 $M_{i_{\text{res}}}$ は，3.2.2 章で説明されているベクトル割り当てを適用して，これらのグリッドセルを埋め込む．解像度 i_{res} でのグリッドセル l のエンコーディングは次の通りである：

$$\text{hiencode}_{i_{\text{res}}}(l) = M_{i_{\text{res}}}[\lfloor l/2^{i_{\text{res}}} \rfloor, l\%2^{i_{\text{res}}}],$$

ここで $l \in L_{i_{\text{res}}}$ である．このエンコーディングは，解像度 i_{res} でのグリッドセル上の確率分布を生成するためにスコアリングコンポーネントに送られる：

$$\Pr(l_{i_{\text{res}}} = l | v_1, \ldots, v_{i-1}) = \text{softmax}(\text{score}_{L_{i_{\text{res}}}})[l],$$

ここで $l \in L_{i_{\text{res}}}$ および $\text{score}_{Li_{\text{res}}} = (\text{score}_1, \ldots, \text{score}_{|L_{i_{\text{res}}}|})$ である．すべての解像度で共通の GRU によって計算されたクエリベクトルを使用することに注意されたい．

3.4.2 目的関数

我々の新しいトレーニング戦略は，異なる解像度を用いたタスクを取り入れる．具体的には与えられた正解データであるグリッドセル $y \in L_d$（主要タスクレベルの POI）を学習するのに加えて低解像度 $i_{\text{res}}(i_{\text{res}} < d)$ での対応するグリッドセルを学習することである．つまり，このアプローチは，私たちのトレーニングに追加の $d-1$ のタスクを導入する．各解像度 i_{res} でのタスクは，それ自身の損失関数を持つ：

$$\text{loss}_{i_{\text{res}}} = \sum_{l \in L_{i_{\text{res}}}} \delta_{l, \text{up}_{i_{\text{res}}}(y)} \log(\Pr(l_{i_{\text{res}}} = l | v_1, \ldots, v_{i-1})),$$

ここで $\text{up}_{i_{\text{res}}}(y) \in L_{i_{\text{res}}}$ は，解像度 i_{res} で y をカバーするグリッドセルである．$\Pr(l_{i_{\text{res}}} = l | v_1, \ldots, v_{i-1})$ は，解像度 i_{res} で次のグリッドセルが l であると推測される確率を表す．これらの損失はすべての解像度で合計され，この合計が逆伝播のための勾配を計算するために使用される．これらの損失関数を最適化することにより，主要な解像度 d だけでなく，各追加の解像度 i_{res} での次のグリッドセルの確率分布を学習する．より小さい i_{res} で定義されたタスクは，ラベルの数 $|L_{i_{\text{res}}}|$ が少ないため，本質的に単純である．単純さのためによく訓練された粗い解像度から細かいセルを推論することにより，主要タスクに関連する学習の困難さを緩和するのに役立つ．

4 プライベート事前学習

公開データを使用した事前学習は，DP-SGD の制約を緩和する有効な戦略として知られている [2]．しかし，公開データへの依存は大きな制約であり，常に公開データは利用可能とは限らない．この課題に対処するため，DP 準拠の遷移行列を利用した新しいプライベート事前学習手法を導入する．このプライベート事前学習は，モデルの収束を加速する「ウォームスタート」を提供するだけでなく，経験的リスクを減少させるモデルの能力を高める．

4.1 DP 粗遷移行列

私たちのアプローチでは，事前学習の基盤として一次遷移行列を使用する．コアタスクは，与えられた POI から次の POI の分布を予測することである．しかし，POI の数 n_{POI} が大きい場合，POI から POI への一次遷移行列は極端に疎になる．DP を満たすためにノイズを加えると，この疎さによりノイズが圧倒的になり，事前学習に活用するのが難しくなる．

この問題を避けるために，HRNet の多重解像度解釈を活用する．直接の POI から POI への遷移ではなく，特に粗い解像度 $l_{i_{\text{res}}} \in L_{i_{\text{res}}}$ のグリッドセルから POI $l \in L_d$ への遷移を考える．

$$\text{TRAN}_{i_{\text{res}}}[l_{i_{\text{res}}}, l] = \sum_{\mathbf{v} \in D} \mathbf{1} V l_{i_{\text{res}}}, l(\mathbf{v}) / |\mathbf{v}|, \tag{2}$$

ここで $\mathbf{1}$ は指示関数であり，$V_{l_{i_{\text{res}}}, l}$ は軌跡の母集団 $\mathcal{V}$ の部分集合で，$l_{i_{\text{res}}}$ から l への遷移を含む軌跡を含む：

$$V_{l_{i_{\text{res}}}, l} := \{\mathbf{v} \in \mathcal{V} | \exists i, \text{up}_{i_{\text{res}}}(\mathbf{v}[i]) = l_{i_{\text{res}}} \ \& \ \mathbf{v}[i+1] = l\}.$$

この計算は，DP の感度を 1 に制限するために，シーケンス長 $|\mathbf{v}|$ で商を取っている．そして，DP を保証するため，次のようにラプラスノイズを行列に加える：

$$\text{DPTRAN}_{i_{\text{res}}}[l_{i_{\text{res}}}, l] = \text{TRAN}_{\text{i}_{\text{res}}}[l_{i_{\text{res}}}, l] + \text{Laplace}(\varepsilon),$$

ここで $\text{Laplace}(\varepsilon)$ は，感度が 1 の場合に ε-DP をもたらすラプラスノイズである．このアプローチにより，より密でノイズが少ない遷移行列が生成され，事前学習に有用なデータとなる．図 1 の左側に，このような粗い遷移行列のサンプルが示されている．

4.2 プライベート事前学習

この章では，DP 遷移行列を使用した事前学習手法について説明する．事前学習のタスクは，与えられた粗い入力グリッドセルから POI への遷移の確率を示す確率分布を生成することが主な目的である．階層的位置埋め込みコンポーネントとそのスコアリングコンポーネントは事前学習を行うが GRU（図 2 参照）はプレトレーニングを受けないことに注意されたい．

4.2.1 アーキテクチャ

事前学習でのアーキテクチャは，DP 遷移行列にプレフィックス情報がないため，GRU を代替する一時コンポーネントを用いる．この一時的なコンポーネントは，GRU の機能を模倣し，エンコードされたプレフィックスと同じ次元のクエリベクトルを生成するように設計されている．直感的には，このクエリベクトルは，以下の式で表される遷移の初期グリッドを表している：

$$h = \text{temp}(x), \tag{3}$$

ここで x は初期グリッドセルを表す入力ベクトルで，以下の章で詳しく説明される．この一時的なコンポーネントは，事前学習でのみ用いられ，モデルトレーニング中に GRU に置き換えられる．

この変更を除いて，事前学習中のアーキテクチャは HRNet モデルと一貫している．スコアリングコンポーネントは，一時的なコンポーネントによって出力されたベクトル h と，階層的埋

め込みコンポーネントから取得された POI のエンコーディングを処理する．クエリベクトルは $\text{query} = f_{\text{query}}(h)$ として計算され，POIl を表すキーベクトルは $\text{key}_l = f\text{key}(\text{hiencode}_d(l))$ として計算される．次に，ドット積 $\text{score}_l = \text{query} \cdot \text{key}_l$ を使用して POI l のスコアを計算する．これらの計算されたスコアにソフトマックス関数を適用することで，L 上の確率分布が得られる．

4.2.2 目的関数

まず，ナイーブな目的関数の限界について議論し，次にこれらの欠点に対処するための代替手法を導入する．

ナイーブな目的関数の限界：直感的でナイーブなアプローチは，与えられた遷移確率（つまり，DPTRAN[l]）を初期グリッドセル l の正解データとして直接使用し，損失関数を次のように計算する：

$$\text{loss} = \frac{1}{|L_{i_{\text{res}}}|} \sum_{l \in L_{i_{\text{res}}}} \text{KL}(\text{DPTRAN}_{i_{\text{res}}}[l] || \hat{y}),$$

ここで $\hat{y}$ は上記のモデルが生成した分布であり，$x = \text{hiencode}_{i_{\text{res}}}(l)$ である．この方法は，正解データの遷移確率分布とモデル由来の分布との間のカルバック・ライブラー (KL) ダイバージェンスを計算している．しかし，このアプローチは，限られた数の正解データ $|L_{i_{\text{res}}}|$ しか持たず，過学習につながる可能性がある．

連続的な初期状態を使用した代替目的関数：正解データの限られた数 $|L_{i_{\text{res}}}|$ の問題を克服するために，解像度 i_{res} におけるグリッドセルの比例的な混合による拡張アプローチを提案する．混合比率ベクトルは次のように定義される：

$$\mathbf{r} = (r_1, \ldots, r_{|L_{i_{\text{res}}}|}) \in [0,1]^{|L_{i_{\text{res}}}|},$$

ただし，$\sum_{i=1}^{|L_{i_{\text{res}}}|} r_i = 1$ という制約がある．ここで，各 r_i は初期状態におけるグリッドセル i の割合を表し，事前分布は確率 $\mathbf{r}$ を持つ多項分布とする．次に，$\mathbf{r}$ を与えられた時の遷移確率分布を次のように計算する：

$$\Pr(l_{\text{next}}|\mathbf{r}) = \sum_{l \in L_{i_{\text{res}}}} \text{Mult}(l_{\text{initial}} = l|\mathbf{r}) \Pr(l_{\text{next}}|l_{\text{initial}} = l),$$

ここで $\text{Mult}(\cdot|\mathbf{r})$ は確率 $(r_1, \ldots, r_{|L_{i_{\text{res}}}|})$ を持つ多項分布であり，$\Pr(l_{\text{next}}|l_{\text{initial}} = l) = \text{DPTRAN}_{i_{\text{res}}}[l]$ である．これは，確率 $\Pr(l_{\text{initial}}|\mathbf{r})$ で l_{initial} に滞在していると仮定するときの遷移確率分布である．$\mathbf{r} \in [0,1]^{|L_{i_{\text{res}}}|}$ であり連続であるため，正解データは有限ではなくなり，データの拡張となっている．混合比率はディリクレ分布からサンプリングされる $\mathbf{r} \sim \text{Dirichlet}(\mathbf{a})$, ここで $\mathbf{a} = (1, 1, \ldots, 1)$ であり，$|\mathbf{a}| = |L_{i_{\text{res}}}|$ である．

HRNet は与えられた $\mathbf{r}$ に対して次のように初期状態を埋め込む：

$$x = \sum_{l \in L_{i_{\text{res}}}} \text{Mult}(l_{\text{initial}} = l|\mathbf{r}) \text{hiencode}_{i_{\text{res}}}(l).$$

これは，比率 $\mathbf{r}$ によるグリッドセル $l \in L_{i_{\text{res}}}$ のエンコーディングの重み付き和である．次に，一時的なコンポーネント（式 3）への入力として混合されたエンコーディング x を使用し，キーベクトルを導出することによって確率分布を導出する．最後に，損失は次のように計算される：

$$\text{loss} = \text{KL}(\Pr(l_{\text{next}}|\mathbf{r}) || \hat{y}|\mathbf{r}),$$

$\hat{y}|\mathbf{r}$ はキーベクトルから導出された確率分布である．この目的関数により，モデルは DPTRAN の確率分布を学習するだけでなく，これらの確率分布間の連続的な表現も獲得する．

4.3 プライバシ分析

この節では，DP 遷移行列を使用して事前学習された HRNet のプライバシ保証について分析する．DP-SGD が (ε_1, δ)-DP に準拠する場合，ラプラス (ε_2) を用いて DPTRAN を利用する事前学習を組み込んだモデルは $(\varepsilon_1 + \varepsilon_2, \delta)$-DP を満たす．事前学習フェーズでは，DP 遷移行列の生成のためにのみ生データが使用される．ラプラスメカニズムの適用は，ε-DP を満たすことが知られているため，事前学習フェーズは ε-DP を保証する [10]．さらに，DP-SGD はガウスメカニズムによって (ε, δ)-DP を満たすことが確立されている [1]．DP の合成定理 [11] を適用することで，DP-SGD と事前学習フェーズを含む HRNet 全体が $(\varepsilon_1 + \varepsilon_2, \delta)$-DP に準拠することが導かれる．

5 実　　験

この章では，DP 制約下での人流データを生成する HRNet の有効性を評価するための実験について説明する．

5.1 設　　定

5.1.1 データセット:

三つの実世界の人流データを用いる．使用されるデータセットは，Geolife [34]，Peopleflow [24]，Didi [30] である．

5.2 比較対象

HRNet をベースラインジェネレータと導入部で概説された異なるカテゴリの 3 つの最先端の手法と比較する:

- **ベースラインジェネレータ:** 2.3 章で説明されているジェネレータを DP-SGD で実装する．
- **PrivTrace [28]**: 移動データ生成のためにマルコフモデルを活用する最先端のマルコフベースのアプローチ．
- **クラスタリング [20]**: k-means を使用して移動データをクラスタリングする生成手法．
- **MTNet [30]**: 道路網データに特化して設計された深層学習モデル．

DPT [14] のようなツリーベースのメソッドは比較から除外した．これは，マルコフベースとクラスタリングベースのメソッドがツリーベースのアプローチよりも優れているという前述の研究 [20], [28] からの報告に基づく．実装では，PrivTrace と MTNet のための公開されているオープンソースのコードを利用した．

5.3 評価指標:

生成された移動データの品質を比較するため，Jensen-

表 1: 実際のデータセットと生成されたデータセットの不一致（Geolife / Peopleflow）$w = 32$ のとき $\varepsilon = 2$ および $\delta = 10^{-5}$.

	waypoint	destination	transition	travel distance
Clutering	9.9248 / 12.865	0.4379 / 0.6503	0.4229 / 0.5417	0.0227 / 0.1542
PrivTrace	12.636 / 11.049	0.2740 / 0.6493	0.2454 / 0.4307	0.0476 / **0.0181**
Baseline	6.6776 / 10.10	0.3974 / 0.6703	0.3968 / 0.4861	0.0275 / 0.1506
HRNet (ours)	**5.665** / **6.167**	**0.1923** / **0.5856**	**0.2121** / **0.3090**	**0.0087** / 0.0569

Shannon divergence(JS divergence) を使用して，生成された軌跡と元の軌跡（$\mathbf{v}$ と表記）の差異を評価する．主な指標としては経由地点，目的地，移行，移動距離，およびルートが考慮される．

経由地点: 一般的な軌跡は，意味的に重要な複数の経由地点（例えば，家から公園へ，そしてレストランへの移動）を含む．始地点 l_{start} を与えられた時の位置 $l \in L$ が経由地点である確率は，次の式を使用して評価される:

$$\Pr(l \in \mathbf{v} | v_0 = l_{\text{start}}). \tag{4}$$

目的地: 軌跡の最終地点は，目的地として意味的に重要である．開始地点 l_{start} を与えられた時の位置 $l \in L$ が目的地である確率は次のように計算される:

$$\Pr(\mathbf{v}[-1] = l | v_0 = l_{\text{start}}). \tag{5}$$

ここで，$\mathbf{v}[-1]$ は $\mathbf{v}$ の最後の地点である．この文脈では，これがすべての地点 L に対する確率分布を形成する．

移行: 開始地点から次の地点への移行は，軌跡の重要な特徴である．最初の地点から次の経由地点への移行として位置 $l \in L$ への移行確率は次のように評価される:

$$\Pr(l = \mathbf{v}[1] | v_0 = l_{\text{start}}). \tag{6}$$

目的地と同様に，これはすべての地点 L に対する確率分布を形成する．

移動距離: 軌跡における総移動距離は，人流特性を特徴づける重要な側面である．軌跡 $\mathbf{v}$ によってカバーされる移動距離範囲 $r \in R$ を次の式を使用して評価する:

$$\Pr\left(\sum^{|\mathbf{v}|} i = 0 d(\mathbf{v}[i], \mathbf{v}[i+1]) \in r \right). \tag{7}$$

この式で，R は範囲 $[0, d\text{max}/n_{\text{bin}}), [d_{\max}/n_{\text{bin}}, 2{*}d_{\max}/n_{\text{bin}}) \ldots, [(n_{\text{bin}}-1) * d_{\max}/n_{\text{bin}}, d_{\max}]$ を含む．ここで，$d_{\max}$ は最大距離，整数 n_{bin} はビンサイズを表す．関数 d は地図上のユークリッド距離を示す．この指標は，すべての範囲 R に対する確率分布を形成する．

ルート: ルートは $\mathbf{v}$ の経由地点を通る通過地点の列を表し，$\text{route}(\mathbf{v})$ として示される．元のデータセットからサンプリングされた軌跡の場合，ルートには前処理前の元の軌跡が参照される．一方，ジェネレータからの軌跡の場合，ルートは隣接する 2 つの経由地点間の最短経路を使用して構築される．MTNet については，道路ネットワークを直接学習し生成するため，この種の後処理は不要である．開始地点 l_{start} を与えられた場合，地点 $l \in L$ がルートの一部である確率は次のように計算される:

$$\Pr(l \in \text{route}(\mathbf{v}) | v_0 = l_{\text{start}}). \tag{8}$$

5.4 不一致度合い

この章では，JS divergence を使用してさまざまな指標での不一致度合いを計算する方法について詳述する．私たちの研究では，L_{start} を潜在的な開始地点として機能する候補地点のセットとして定義する．L_{start} を決定するために，データセットで最も発生率が高い上位 30 の場所を初期地点として選択した．

経由地点とルート: 経由地点とルートの指標は，確率分布 (4) および (8) で定義され，台はバイナリ（0, 1）であり，地点 $l \in L$ が（ルート）軌跡に存在するか存在しないかを示す．元のデータセットと生成されたデータセットから導かれる確率分布をそれぞれ $p_{l,l_{\text{start}}}$ および $q_{l,l_{\text{start}}}$ とし，これらの指標の不一致度合いは次の式で計算される JS divergence の平均を使用する: $\frac{1}{|L_{\text{start}}|} \sum_{l_{\text{start}} \in L_{\text{start}}} \frac{1}{|L|} \sum_{l \in L} \text{JS}(p_{l,l_{\text{start}}} || q_{l,l_{\text{start}}})$.

目的地，移行，および移動距離: 目的地および移行の指標については，確率分布 (5) および (6) で定義され，台はすべての地点（L）を含む．同様に，移動距離指標 (7) のサポートはすべての距離範囲（R）を含む．元のデータセットおよび生成されたデータセットからの確率分布をそれぞれ $p_{l_{\text{start}}}$ および $q_{l_{\text{start}}}$ とし，これらの指標の不一致度合いは次の式で計算される平均 JS divergence を使用する: $\frac{1}{|L_{\text{start}}|} \sum_{l_{\text{start}} \in L_{\text{start}}} \text{JS}(p_{l_{\text{start}}} || q_{l_{\text{start}}})$.

5.5 ベースライン，PrivTrace，クラスタリングとの比較

この章では，道路網によって制約されていない人流データに対して，提案手法と既存のアプローチを比較する．この分析における比較対象は，ベースライン，PrivTrace，およびクラスタリングである．使用されるデータセットは，Geolife データセットと Peopleflow データセットである．

5.5.1 主な結果

表 1 は主な結果を示しており，ここでは固定されたプライバシ予算（$\varepsilon = 2$）で四つの指標を評価している．特筆すべきは，提案手法が他のアプローチと比較して，すべての指標において顕著な優位性を示していることである．クラスタリングや PrivTrace などの統計ベースの方法は，疎なデータにノイズを加える必要があるため，高いシグナル・ノイズ比によって精度が低い．一方，深層学習手法は，単なる疎なデータを超えて，POI の隠れた特性を明らかにする．ベースライン方法は，依然として埋め込み行列を通じて疎な情報を学習するため，DP-SGD との互換性が低く，ユーティリティが低下している．提案手法は，転置畳み込みネットワーク，マルチタスク学習，およびプライベート事前学習を組み合わせることで，DP-SGD と調和する

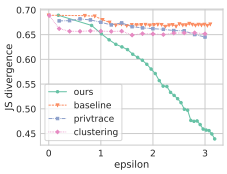


図 3: 人流データセット（$w = 64$）における目的地の不一致度について，各 $\varepsilon \in [0, 3.0]$ での結果.

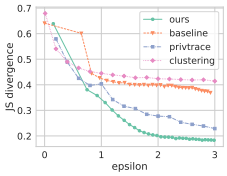


図 4: Geolife データセット（$w = 32$）における遷移の不一致度について，各 $\varepsilon \in [0, 3.0]$ での結果.

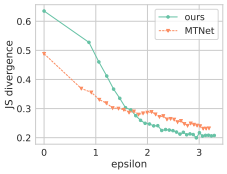


図 5: Didi データセット（$w = 32$）における目的地の不一致度について，各 ε および $\delta = 10^{-5}$ での結果.

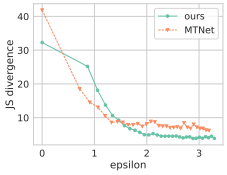


図 6: Didi データセット（$w = 32$）におけるルートの不一致度について，各 ε および $\delta = 10^{-5}$ での結果.

形で疎な特性を効果的に学習している.

5.5.2 ユーティリティ・プライバシのトレードオフ

この節では，プライバシ予算を変動させることによって，提案手法のユーティリティ・プライバシのトレードオフを比較対象と比較する．図 3 および 4 に示される結果から，この分析が示される．低いプライバシ予算（例えば，$\varepsilon < 1$）では，PrivTrace やクラスタリングが時々より良いパフォーマンスを示すことが観察される．これは，これらの直接遷移を計算するためである．対照的に，深層学習手法は通常，人流データの隠れた特徴を学習するために，多くの訓練エポックを必要とする．しかし，提案手法は事前学習とマルチタスク学習を通じて学習プロセスを加速し，ベースライン手法とは異なり，必要なプライバシ予算を削減している．さらに，学習が進むにつれて，深層学習アプローチはより優位になる．これはクラスタリングや PrivTrace などの統計ベースの方法は感度を抑えるために情報損失をしているのに対し．HRNet は生データを直接学習することに起因している.

5.6 MTNet との比較

MTNet [30] は，道路ネットワーク内の短い移動を生成するために特別に設計された，深層学習手法である．MTNet のアーキテクチャは，軌跡が道路ネットワーク上で接続されていることを要求する．この要件は，鉄道を用いた移動や公園を通るなど，道路ネットワークを超える移動パターンを含む Geolife や Peopleflow などのデータセットと互換性がない．したがって，MTNet との比較分析のために，私たちは Didi データセットを使用する.

結果は，図 5 と 6 に示されている．遷移と目的地の両方の指標において，MTNet は初期段階で優れたパフォーマンスを示している．これは，MTNet が道路網のメタデータを使用して軌跡を生成するためである．しかし，トレーニングが進むにつれて，私たちの方法が改善された結果を示している．この差は，トレーニング手法の違いから生じている．私たちの方法は，意味のある次の経由地点を学ぶことに特化している．一方，MTNet は次の道路セグメントを直接学ぶように設計されている．その結果，MTNet は DP-SGD が適用される場合，より長いシーケンスを学ぶ必要があり，ノイズの多い設定では学習が困難になる．したがって，私たちの方法は DP-SGD が適用される場合において，より良い収束を示している.

要約すると，MTNet は非常にノイズの多い設定（つまり，小さい ε）において道路網上の短い移動を学ぶのに優れているが，Geolife データセットや Peopleflow データセットのように地図全体に広がる人流を捉えるのには不向きである．私たちの方法は，道路網に限定されない複雑な人流パターンを理解するためのより汎用的なアプローチを提供する.

6 結　　論

本論文では，DP 制約下で人流データを効果的に学習し生成するための新しいフレームワークである HRNet を提案した．HRNet は，階層的な位置エンコーディングコンポーネント，多重解像度多タスク学習，およびプライベートな事前トレーニングの 3 つの新しい要素を統合している．これらの要素は，ベースラインアプローチのボトルネックに対処し，学習効率とデータ生成の質を向上させる.

これらのコンポーネントにはこの研究で紹介した用途を超える汎用性がある．たとえば，位置エンコーディングに埋め込み行列を使用する任意の深層学習モデルにおいて，DP-SGD と統合する際に私たちのアプローチを統合することで，パフォーマンスを向上させる可能性がある．さらに，HRNet で使用される階層的なネットワークの概念は，位置エンコーディングへの応用を超えて本質的な価値を持っている．データの階層的な性質は，位置に限らず言語にも顕著である．例えば，言葉は意味の層を持っている．この観察は，DP 制約下で言語モデルを改善するために，階層的なアプローチを適用する可能性を開くことを示唆しており，プライバシ保護のためのディープラーニング分野における将来の研究への有望な道を示している.

7 謝　　辞

この研究は，エモリー大学に客員研究員として訪れていた期間に，Li Xiong 教授のご指導のもとで行われました．教授の熱いご指導に心から感謝申し上げます.

文　　献

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] E. Amid, A. Ganesh, R. Mathews, S. Ramaswamy, S. Song, T. Steinke, V. M. Suriyakumar, O. Thakkar, and A. Thakurta. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR, 2022.

[3] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

[4] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

[5] S. Cai, X. Lyu, X. Li, D. Ban, and T. Zeng. A trajectory released scheme for the internet of vehicles based on differential privacy. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):16534–16547, 2021.

[6] R. Chen, G. Acs, and C. Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 638–649, 2012.

[7] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[8] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

[9] C. Dwork. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pages 1–12. Springer-Verlag, 2006.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[11] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[12] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[13] M. E. Gursoy, L. Liu, S. Truex, and L. Yu. Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing*, 18(10):2315–2329, 2018.

[14] X. He, G. Cormode, A. Machanavajjhala, C. Procopiuc, and D. Srivastava. Dpt: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

[15] J. Hua, Y. Gao, and S. Zhong. Differentially private publication of general time-serial trajectory data. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 549–557. IEEE, 2015.

[16] N. Johnson, J. P. Near, and D. Song. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539, 2018.

[17] P. Kairouz, M. R. Diaz, K. Rush, and A. Thakurta. (nearly) dimension independent private erm with adagrad rates via publicly estimated subspaces. In *Conference on Learning Theory*, pages 2717–2746. PMLR, 2021.

[18] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[19] M. Li, L. Zhu, Z. Zhang, and R. Xu. Achieving differential privacy of trajectory data publishing in participatory sensing. *Information Sciences*, 400:1–13, 2017.

[20] Q. Liu, J. Yu, J. Han, and X. Yao. Differentially private and utility-aware publication of trajectory data. *Expert Systems with Applications*, 180:115120, 2021.

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[22] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo. A survey on deep learning for human mobility. *ACM Comput. Surv.*, 55(2):7:1–7:44, 2023.

[23] R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale. Analyzing privacy risk in human mobility data. In *Software Technologies: Applications and Foundations: STAF 2018 Collocated Workshops, Toulouse, France, June 25-29, 2018, Revised Selected Papers*, pages 114–129. Springer, 2018.

[24] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui, and Y. Shimazaki. Pflow: Reconstructing people flow recycling large-scale social survey data. *IEEE Pervasive Computing*, 10(4):27–35, 2011.

[25] A. D. P. Team. Learning with privacy at scale, 2017.

[26] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[27] Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin. A new privacy breach: User trajectory recovery from aggregated mobility data. *IEEE/ACM Transactions on Networking*, 26(3):1446–1459, 2018.

[28] H. Wang, Z. Zhang, T. Wang, S. He, M. Backes, J. Chen, and Y. Zhang. Privtrace: Differentially private trajectory synthesis by adaptive markov model. In *USENIX Security Symposium 2023*, 2023.

[29] S. Wang, J. Cao, and S. Y. Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020.

[30] Y. Wang, G. Li, K. Li, and H. Yuan. A deep generative model for trajectory modeling and utilization. *Proceedings of the VLDB Endowment*, 16(4):973–985, 2022.

[31] X. Yan, T. Song, Y. Jiao, J. He, J. Wang, R. Li, and W. Chu. Spatio-temporal hypergraph learning for next poi recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 403–412, 2023.

[32] D. Yang, B. Fankhauser, P. Rosso, and P. Cudre-Mauroux. Location prediction over sparse user mobility traces using rnns. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2184–2190, 2020.

[33] M. Zhang, H. Lin, S. Takagi, Y. Cao, C. Shahabi, and L. Xiong. Csgan: Modality-aware trajectory generation via clustering-based sequence gan. In *2023 24th IEEE International Conference on Mobile Data Management (MDM)*, pages 148–157. IEEE, 2023.

[34] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

顔画像プライバシ保護技術の比較検証

馬 世昂† 曹 洋† 中村 篤祥†

†北海道大学情報科学院 〒060-0814 北海道札幌市北区北 14 条西 9 丁目

E-mail: †{sma, yang, atsu}@ ist.hokudai.ac.jp

あらまし 近年、画像認識や表情推定などの技術の向上により、顔写真から様々な情報を高い精度で抽出できるようになった。ところが、これらの技術の多くは対象者の顔写真を自由に取得できることを前提としている。この前提はプライバシ保護の観点からすると問題である。この問題を解決するために、顔画像のプライバシ保護に関して様々な研究が行われてきた。しかし、中には同じ目的を掲げる研究も多く、互いの手法の効果の違いが検証されていない。そこで、本研究では理論と実験の両面からプライバシの保護を比較検証する。理論ではそれぞれの手法で用いられる差分プライバシまたはその派生形の違いを明らかにした。また、実験ではいずれの手法においても学習で用いられていないデータセットを用いてプライバシ保護した結果を最先端の攻撃手法を用いて比較する。特に差分プライバシを満たす手法と満たさない手法の違いについて検証する。

キーワード 差分プライバシ，匿名化，顔画像

1. はじめに

人間の顔写真は他の生体情報とは異なり、大量の情報が含まれている。Naumann らの調査[1]によると人は一枚の写真で「外向性の高さ」「自尊心の高さ」「信仰心の厚さ」「協調性の高さ」と「どれだけ良心的であるか」の５つを正確に読み取ることができる。David らの調査[2]によると、心臓疾患リスクが低い女性は、高い女性より２歳以上若く見られる傾向がある。一方で、近年の科学技術の向上により、高い精度で顔写真から情報を抽出することができるようになった。Qingqiu Huang の研究[3]ではラベル付けを行わないで顔認識する手法が提案され、99.65%の高い精度まで学習できるとされている。このような大量の情報を含む顔写真を広く公開することはプライバシが侵害される可能性を生むことになる。幸い、近年は先進国をはじめとして各国でプライバシを保護することを目的とした法整備が進んでいる。EU は領域内の個人情報保護を規定する法律として General Data Protection Regulation(GDPR)、アメリカでは American Data Privacy and Protection Act、中国では中国個人保護法、日本では個人情報保護法が定められている。これらの法律は個人情報の利用を最小限に留め、同意を得てからデータを収集することを企業などの組織に求める。これらの法律はプライバシを保護する一方で、企業などの組織が有用なデータを取得するためにはこれまで以上に大きなコストがかかるようになった。さらに、一部の人が個人情報の提供を拒否するため、収集したデータに偏りが生じる可能性がある。

この状況を打開するために、近年様々な顔画像のプライバシ保護手法が提案されてきた。これらの研究は大きく２つに分けることができる。１つ目は顔匿名化である。顔匿名化とは顔画像から直接識別情報を取り除く手法である。この目的を達成する手法として一般的にはモザイク処理やぼかしなどの方法が用いられるが、これらの方法は深層学習モデルの攻撃にもろいことが示された[25]。そこで、差分プライバシをモザイク処理などに適用させる方法が提案されるようになった[4]。また、近年はプライバシ保護した画像を人間にとってより自然なものにするために潜在空間を利用する手法が提案されるようになってきた[5][6][7]。２つ目は顔認証システムのプライバシ保護である。顔認証システムのプライバシ保護とは顔画像から顔属性と呼ばれる間接識別情報を取り除く手法である。例えば、性別、髪色や健康状態などが顔属性にあたる。この方法は一般的に顔認証システムに適用されることを想定している。これは Othman ら[8]によって最初に提案された概念である。顔認証システムのプライバシ保護の提案手法は、我々の知る限り、顔画像に関するプライバシ保証がある研究が存在しない。また、いずれの顔画像のプライバシ保護手法の分類においても様々な手法が提案されたため、これらの手法を分別するような調査も存在する[10][11]。それらの多くは手法の分類や適用場面の分類である。

ところが、先進的な顔匿名化手法の間の違いは我々の知る限り明らかにされていない。また、実験による比較は存在していないため、専門性が低い人にとって技術の導入を検討することが難しい。特に、提案手法ごとに異なる差分プライバシを定義しており、それらの比較、数学的な証明がない手法との比較は必要である。そこで、本研究では、最先端の手法に対して論理的な比較と実験による複数のタスク検証を行う。

検証として各手法に対して最先端な方法である SFace[30]と MagFace[31]による顔画像認識精度を比較した。SFace[30]と MagFace[31]は２枚の画像の類似度

を算出する手法である。そのため、実験ではそれぞれのプライバシ保護後画像に対して元データを順に比較して、初めて類似度が閾値を超えたときの画像がプライバシ保護前画像である確率を認識率とした。また、各手法のパラメータの変化による顔画像認識精度の変化に注目し、維持したい認識精度の大きさから設定すべき差分プライバシのパラメータεを算出して比較した。

2. 関連研究

2.1. 顔匿名化

顔匿名化の技術は顔画像プライバシ保護の分野の中で比較的古くから研究されてきた。顔画像に映る人物の特定を防ぐことが主要の目的である。初期の段階ではピクセル化、ぼかしやマスク処理といった画素処理の手法が主流であった。例えば、Fan[4]は m-Neighborhood を提案し、差分プライバシの概念をピクセル化に適応した。そこで、k-匿名化を組み合わせた手法が盛んになった。例えば、Newton ら[5]は顔画像データベースにおける k-Same アルゴリズムを提案し、閉じたデータベースに対してプライバシ保証をした。これらの方法はデータの合成を行うことによって特定を k 人までしか行えないことが証明される。しかし、k-匿名化の方法は依然として不自然な物体が残りやすい。そのため、近年は画像から生成される潜在変数などの埋め込み表現に対して処理を行う手法が注目されるようになった。例えば、Fan[6]は特異値分解(SVD)を行うことによって次元削除を行った後にメトリック差分プライバシを満たすようにノイズを加える手法を提案した。Chen ら[7]は GAN インバージョンを用いて潜在コードを算出してから有用性を保ちながらもメトリック差分プライバシを満たす手法を提案した。

2.2. 顔認証システムのプライバシ保護

顔認証システムのプライバシ保護は顔匿名化とは逆に顔画像に映る人物の特定をするが、その他の間接識別情報の流出を防ぐことが目的である。これは Othman ら[8]によって最初に提案された概念である。顔認証システムのプライバシ保護の研究も初期の段階は画素処理の手法が主流であったが、次第にベクトルに対して処理する研究が主流になっていった。Li ら[9]は顔全体を顔認証において重要な部分とそうでない部分に分け、重要でない部分に対して処理を行う手法を提案した。この領域の研究において Othman ら[8]が差分プライバシの適応の可能性について論じた。

2.3. 顔画像プライバシ保護技術の調査

顔画像プライバシ保護技術の調査はこれまでいくつか行われてきた。Medan ら[10]は 200 以上の顔匿名化の技術を分類し、今後の展望について論じた先行研究である。また、Melzi ら[11]は暗号化技術を含め、既存の顔認証システムのプライバシ保護技術を分類し、読者に対して用途に合った方法の選択を補助した。

これらの調査は既存の研究を分類することを目的としている。しかし、同じカテゴリに分類された研究については互いの違いが明らかにされていない。そのため、1 つの手法を採用したい読者にとっては不十分である。また、既存の研究は[7]や[19]などの最先端の研究を網羅できていない。そこで、本研究では顔匿名化の最先端の提案手法の比較検証を行い、読者にとって競合しうる手法に対してはタスクごとにベンチマークテストを行う。

3. 準備

3.1. 差分プライバシ（Differential Privacy, DP）

差分プライバシは Aitsam [12]によって提案された数学的に証明可能なプライバシ保護基準である。この基準は定量的であり、出力から元のデータセットに含まれる個人情報を推測できないことを保証する。数学的定義は次のようになる[13]。

任意の二つの近傍データセットDとD'に対して、メカニズムMは次の式を満たすとき、差分プライバシの基準を満たす。ただし、ここでSは可能な出力の集合である。

$$\Pr[\mathrm{M}(\mathrm{D}) \in \mathrm{S}] \leq e^{\varepsilon}\Pr[\mathrm{M}(\mathrm{D}') \in \mathrm{S}]$$

ここでεはプライバシ保護する度合いのパラメータであり、しばしばプライバシ予算と呼ばれる。プライバシ予算が大きいほどプライバシ保護する度合いが弱くなる。

ここで、近傍データセットとは攻撃者が区別できないことを保証するデータセットである。一般的には適用場面に合わせて設定する。Aitsam [12]は近傍データセットを 1 つの要素が異なるサブデータセットのペアとした。

3.2. ローカル差分プライバシ（Local Differential Privacy, LDP）

差分プライバシはデータベース内のデータに対してプライバシ保証をする基準であるが、一般的な組織が情報を収集する適応場面を考慮した場合、個人は組織を信頼しなければならない。しかし、個人によっては組織を信頼しない場合が存在する。その時、代わりにローカル差分プライバシの基準を適応できる。差分プライバシは図 1 において組織がデータを収集してからプライバシ保護するのに対して、ローカル差分プライバシは個人がデータを提供する前にプライバシ保護する。ε-LDP の数学的定義は次のようになる。[14]

任意の二つの入力XとX'に対して、メカニズムMは次の式を満たすとき、ローカル差分プライバシの基準を満たす。ただし、ここでYは可能な出力の集合である。

$$\Pr[\mathrm{M}(\mathrm{X}) \in \mathrm{Y}] \leq e^{\varepsilon}\Pr[\mathrm{M}(\mathrm{X}') \in \mathrm{Y}]$$

3.3. メトリック差分プライバシ

Chatzikokolakis ら[15]は差分プライバシの概念を拡張したメトリック差分プライバシを提案した。差分プライバシは近傍データセットに対してプライバシ保証する。これに対して、メトリック差分プライバシは任意の２つのデータセットxとx'の$d(x,x')$に合わせてプライバシ保護の強さを変える。ここで、$d(x,x')$はxとx'の距離であり、一般的には適応領域によって決める。メトリック差分プライバシは$d(x,x') = \varepsilon$の時、差分プライバシの定義と一致する。数学的定義は次のようになる。

すべてのデータベースDとD'に対して、メカニズムMは次の式を満たすとき、メトリック差分プライバシの基準を満たす。

$$\Pr[M(D) \in S] \leq e^{d(D,D')} \Pr[M(D') \in S]$$

3.4. ラプラスメカニズム

差分プライバシを実現する有名なメカニズムとしてラプラスメカニズムがある。ラプラスメカニズムはラプラス分布に従いノイズをサンプリングしてから元データに加えるメカニズムである。ラプラスメカニズムを$M(.)$、保護対象の関数を$f(x)$として定式化すると次のようになる。

$$M(x) = f(x) + Lap(\frac{S}{\varepsilon})$$

ここで、S は敏感度であり、近傍データセットの変化に伴う$f(x)$の最大変化で定義される。

3.5. 差分プライバシの性質

ここで、差分プライバシにかかわる定理を３つあげる。逐次合成(Sequential composition)の性質、並行合成(Parallel composition)の性質と後処理不変(Post processing)の性質である。

逐次合成の性質は同じ入力データを複数入力する時の結果を公開する際のプライバシコストの合計を制限するものである。定理は次のようになる。

メカニズム$M_1, M_2, M_3, \dots, M_n$がそれぞれ$\varepsilon_1$-プライバシ、$\varepsilon_2$-プライバシ、$\varepsilon_3$-プライバシ…$\varepsilon_n$-プライバシを満たす時、$Q = (M_1, M_2, M_3, \dots, M_n)$は$(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n)$-プライバシを満たす。

複数のメカニズムを適用しても、それがそれぞれ排他的なサブデータセットであれば、逐次合成の性質の結果より小さなプライバシ予算で済む。並行合成の性質の定理は次のようになる。

入力のデータセットを排他的なサブデータセット$D_1, D_2, D_3, \dots, D_m$に分け、それぞれのサブデータセットに対して$\varepsilon_1$-プライバシ、$\varepsilon_2$-プライバシ、$\varepsilon_3$-プライバシ…$\varepsilon_n$-プライバシを満たすメカニズム$M_1, M_2, M_3, \dots, M_n$を仮定する。この時、$Q = (M_1, M_2, M_3, \dots, M_n)$は$\max(\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_n)$-差分プライバシを満たす。

後処理不変の性質はほかの二つの性質と比べてシンプルである。定理は次のようになる。

ε-DP を満たすメカニズムの出力に対して任意の計算を適用しても、その結果はε-DP を満たす。

4. 顔画像プライバシ保護技術の理論比較

先端的な顔匿名化の研究を次の表１にまとめる。

表１　先端的な顔匿名化研究

手法	理論的なプライバシ保証	実験によるプライバシ保証	保存する顔属性	実験による顔属性保証
Newtonら[5]	LDP	ID_IDS PSR	見た目 背景	PSNR, SSIM 顔検出率
Hukkelåsら[23]	×	なし	背景 器官の位置	顔検出率
Chenら[7]	MDP	なし	設定された顔属性	FID, SSIM 顔属性識別率
Fan[4]	CDP	識別モデルの正答率	なし	なし
Liら[24]	×	分類モデルの誤判別率 分類モデルの 正答率	見た目	なし

顔画像プライバシ保護技術は顔匿名化と顔認証システムのプライバシ保護に分けることができる。顔匿名化は入力画像から個人を特定できないが、見た目や性別などの属性を保持することが目的である。一方で、顔認証システムのプライバシ保護は顔認証システムに適用させることを想定している。そのため、個人を特定するが、見た目や性別などの顔属性を識別できないようにすることが目的である。我々の知る限り、顔認証システムのプライバシ保護は差分プライバシ及びその派生形を適用させたものが存在しないため、本節ではこれ以降顔匿名化に限定する。

4.1. 既存のプライバシ保護技術

一般的なピクセル化、ぼかしやマスク処理は実生活でよく用いられている。しかし、[16][17]では十分な訓練データを与えれば、このような手法で処理した画像から元の画像を還元できることが示された。初期の個人のシナリオの研究はこの問題を解決するために差分プライバシを適応するものが多い。例えば、Fan[4]はm-Neighborhood を提案し、m-Neighborhood である画像を近傍データセットとして定義した。そして、差分プライバシを画像領域に適用するようにラプラスノイズを画素に加えた。m-Neighborhood の定義は次のようになる。

二枚の画像I_1とI_2は同じ次元であり、相違のピクセル数はmピクセル以下であるならば、互いに m-Neighborhood である。

ここで、mはパラメータである。mが大きければ大きいほど m-Neighborhood として定義される画像が増える。その結果、プライバシ保護が強くなる。この手法は画像全体に適用させることも、画像の一部に適用させることもできる。Fan[4]でも、顔画像に対しては目の周りのみに対して適用させても有効である可能性を示唆した[18]。しかし、本研究では公平性のため顔画像全体に対して Fan[4]の提案手法を適用させる。

しかし、ピクセル化、ぼかしやマスク処理などの方法は人間から見て画像を不自然にすることが多い。そ

こで、Chen ら[7]は元画像を任意の GAN の GAN inversion を用いて潜在変数に変換してからノイズを加えてプライバシ保護を行う手法を提案した。この手法では生成画像の有用性を保つための最適化を行ってから画像に変換する。最適化を行う前に提案したΔd_X − Privacyを満たすようにノイズを加える。Δd_X − Privacyはメトリック差分プライバシの派生形であり、次のように定義される。

すべてのデータベースxとx'に対して、メカニズムKは次の式を満たすとき、Δd_X − Privacyの基準を満たす。

$$\Pr[\mathrm{M}(\mathrm{X}) \in \mathrm{Y}] \leq e^{\varepsilon \mathrm{d}(\mathrm{x},\mathrm{x}')} \Pr[\mathrm{M}(\mathrm{X}') \in \mathrm{Y}]$$

後処理不変の性質より、あるプライバシ保護した画像の潜在変数がΔd_X − Privacyの基準を満たすならば、その潜在変数から生成した画像もΔd_X − Privacyの基準を満たす。

さらに、Wen ら[19]は保持する顔属性のアノテーションを必要としない手法を提案した。Wen ら[19]の提案手法は画像を Identity encoder と attribute encoder に入力してそれぞれ個人認識 ID と属性変数を fusion network で結合してプライバシ保護した画像を出力する。ε-IdentifyDP というプライバシ基準を提案し、Identity encoder で出力される個人認識 ID に対してラプラスノイズを加えることによってプライバシを保証する。ここで、ε-IdentifyDP の定義はε-LDP の定義と同じある。また、Identity encoder として Arcface[20]を、attribute encoder として[21]を、fusion network として Adaptive Attentional Denormalization (AAD) ResBlk[22]を用いた。

Wen ら[19]は Identity encoder がε-IdentifyDP の基準を満たすため、後処理不変の性質により、出力されるプライバシ画像もε-IdentifyDP の基準を満たすと主張している。しかし、attribute encoder として用いられている手法は画像スタイル変換として提案された手法であり、プライバシ関連の情報が含まれているか不明である。attribute encoder にプライバシ関連の情報が含まれていれば逐次合成の性質により提案手法は全体としてε-IdentifyDP を満たさない可能性がある。この点を含めて本調査で検証する。

また、差分プライバシ及びその派生形の基準を満たさないカテゴリの先進的な研究として DeepPrivacy[23]と Fawkes[28]を比較対象とする。DeepPrivacy は画像の顔の部分を削除して、目や鼻、口などの 7 つの位置を記録する。その記録と背景から新しい顔を生成する手法である。顔部分を削除するので数学的証明がないが、比較的厳密にプライバシ保護ができると予想できる。一方で、Fawkes は特定の画像認識モデルに対して顔画像の人物を誤認識させるように学習する手法である。学習時に用いた画像認識モデル以外のモデルからの攻撃に対する頑丈性は十分な検証がされていない。

4.2. リサーチギャップ

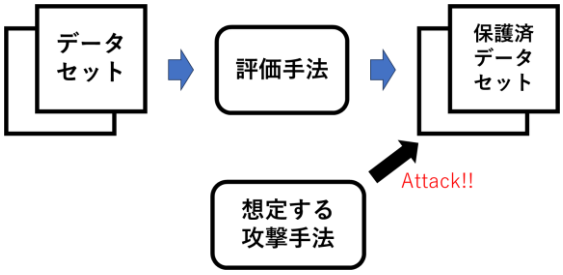
以上に示したように、提案された手法はそれぞれ独自のプライバシ定義を満たしており、εなどのパラメータから手法間のプライバシ保護の強さを測ることができない。そこで、本研究では異なるタスクによるベンチマーク調査を行い、提案手法のプライバシ保護の強さを比較する。

5. 実験評価

5.1. 実験フロー及び評価基準

実験フローは次の図 1 のようになる。

図 1　実験フロー



実験では保護済みデータセットに対して顔画像認識を行い、元データセットとマッチする比率(識別率)を評価指標とする。

5.2. データセット

既存の良質なデータセットはほとんど比較手法によって使われているため、公平性の観点から StyleGAN[29]によって生成さした架空の 100 個人を含む 100 枚の画像をデータセットとした。StyleGAN[29]とは Progressive Growing を用いた高解像画像生成と AdaIN を用いて各層に画像のスタイルを取り込んだ顔画像生成手法である。生成画像の例を図 1 で示す。

図 1　StyleGAN によって生成された画像例

5.3. 想定攻撃

想定する攻撃は複数の最先端な顔画像認識技術である。具体的には SFace[30]と MagFace[31]である。Sface[30]はシグモイド制約超球損失と呼ばれる新たな損失関数を用いることで学習データの質に対して頑丈なモデルである。本実験では、保護後のデータそれぞれに対して保護前のデータと順に比較を行い、cos 類似度が初めて 0.68 を超えた保護前データ保護後データが同じ個人である場合攻撃が成功したと判断する。一方で、MagFace[31]は ArcFace[20]を拡張した手法であり、画像の特徴ベクトルの向きと大きさに注目する

ことで認識精度を向上させたのみならず画像の質の高さも特定することができる。本実験では全結合層を含む iResNet50 を基幹ネットワークとして MS1MV2 データセットで訓練したモデルを用いた。認識率の計算方法は SFace[30]の場合と同じである。

5.4. 結果

SFace[30]による攻撃の結果次の図 3 のようになった。

図 3　SFace[30]と MagFace[31]による認識精度

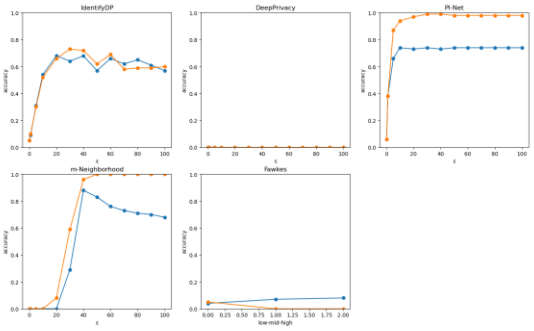


図 3 において、y 軸は SFace による認識精度である。青い線は SFace[30]による認識精度であり、オレンジ色の線は MagFace[31]による認識精度である。また、IdentifyDP、PI-Net、m-Neighborhood は x 軸が ε、DeepPrivacy は x 軸が固定値、Fawkes は x 軸がプライバシ保護の強さを表すパラメータ low、mid、high である。認識精度が一定の値に達した後に低下し始めるのは保護後データが同じ人物の保護前データと比較する前に他の人物の保護前データとの cos 類似度が閾値を超える現象が起こっているからと考えられる。これは、保護後データを保護前データセットと順番に比較しているためであると考えられる。

図 3 の結果を見ると、差分プライバシを適用した方法は ε が一定の値を超えた後急激にプライバシ保護が弱くなる。m-Neighborhood は ε が[20, 40]、IdentifyDP は(0, 20]、PI-Net は ε が(0, 10]の範囲で SFace による認識精度が急激に変化する。そのため、以上の範囲で細かく ε を刻んで図 4 で表示する。

図 4　SFace[30]による認識精度

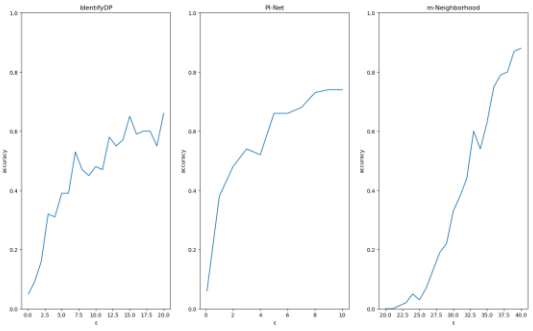
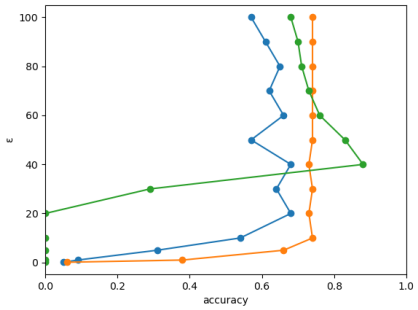


図 4 を見ると、ε が増加するにつれて m-Neighborhood の方が認識精度の上がる勢いが大きい。

最後に、維持したい認識精度の大きさから設定すべき ε を比較する。図 5 のようになる。

図 5　設定すべき ε の大きさ



参　考　文　献

[1] Naumann LP, Vazire S, Rentfrow P, Gosling S. Personality judgments based on physical appearance. Pers Soc Psychol B. 2009;35(12): 1661–1671.I. Tanaka and J. Suzuki, “Web and Database Technologies”, Proc. of ACM SIGMOD, pp. 10-22, 2010.

[2] David A. Gunn et al. Facial Appearance Reflects Human Familial Longevity and Cardiovascular Disease Risk in Healthy Individuals. The Journals of Gerontology: Series A, Volume 68, Issue 2. 2013 : 145–152

[3] Qingqiu Huang, Lei Yang, Huaiyi Huang, Tong Wu & Dahua Lin. Caption-Supervised Face Recognition: Training a State-of-the-Art Face Model Without Manual Annotation.

[4] Fan, L. (2018). Image Pixelization with Differential Privacy. In: Kerschbaum, F., Paraboschi, S. (eds) Data and Applications Security and Privacy XXXII. DBSec 2018.

[5] E. M. Newton, L. Sweeney and B. Malin, "Preserving privacy by de-identifying face images," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 232-243, Feb. 2005, doi: 10.1109/TKDE.2005.32.

[6] L. Fan, "Practical Image Obfuscation with Provable Privacy," 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 2019, pp. 784-789, doi: 10.1109/ICME.2019.00140.

[7] J. -W. Chen, L. -J. Chen, C. -M. Yu and C. -S. Lu,

"Perceptual Indistinguishability-Net (PI-Net): Facial Image Obfuscation with Manipulable Semantics," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 6474-6483, doi: 10.1109/CVPR46437.2021.00641.

[8] Othman, A., Ross, A. (2015). Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity. In: Agapito, L., Bronstein, M., Rother, C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014.

[9] J. Li, H. Zhang, S. Liang, P. Dai and X. Cao, "Privacy-Enhancing Face Obfuscation Guided by Semantic-Aware Attribution Maps," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 3632-3646, 2023, doi: 10.1109/TIFS.2023.3282384.

[10] B. Meden et al., "Privacy–Enhancing Face Biometrics: A Comprehensive Survey," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4147-4183, 2021, doi: 10.1109/TIFS.2021.3096024.

[11] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Christoph Busch: An Overview of Privacy-enhancing Technologies in Biometric Recognition, arXiv preprint arXiv:2206.10465, 2022

[12] M. Aitsam, "Differential Privacy Made Easy," 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 2022, pp. 1-7, doi: 10.1109/ETECTE55893.2022.10007322.

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Theory of cryptography conference. Springer, 2006, pp. 265–284

[14] Local Differential Privacy for data collection and analysis

[15] Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C. (2013). Broadening the Scope of Differential Privacy Using Metrics. In: De Cristofaro, E., Wright, M. (eds) Privacy Enhancing Technologies. PETS 2013. Lecture Notes in Computer Science, vol 7981. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39077-7_5

[16] Hill, S., Zhou, Z., Saul, L., Shacham, H.: On the (in) effectiveness of mosaicing and blurring as tools for document redaction. Proc. Priv. Enhancing Technol. 2016(4), 403–417 (2016)

[17] McPherson, R., Shokri, R., Shmatikov, V.: Defeating image obfuscation with deep learning. CoRR abs/1609.00408 (2016)

[18] Peterson, M.F., Eckstein, M.P.: Looking just below the eyes is optimal across face recognition tasks. Proc. Natl. Acad. Sci. 109(48), E3314–E3323 (2012)

[19] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, Li Son g: IdentityDP: Differential private identification protection for face images. Neurocomputing. Volume 501, 28 August 2022, Pages 197-211

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699

[21] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European conference on computer vision. Springer, 2016, pp. 694–711

[22] Li, J. Bao, H. Yang, D. Chen, and F. Wen, “Faceshifter: Towards high fidelity and occlusion aware face swapping,” arXiv preprintarXiv:1912.13457, 2019

[23] Hukkelås, H., Mester, R., Lindseth, F. (2019). DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In: Bebis, G., et al. Advances in Visual Computing. ISVC 2019. Lecture Notes in Computer Science(), vol 11844. Springer, Cham. https://doi.org/10.1007/978-3-030-33720-9_44

[24] Li J, Han L, Chen R, Zhang H, Han B, Wang L and Cao X. Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation. Proceedings of the 29th ACM International Conference on Multimedia. (3891-3899).

[25] R. Gross, L. Sweeney, F. de la Torre, and S. Baker, “Model-based face de-identification,” in Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW), 2006, p. 161

[26] Croft, W.L., Sack, JR., Shi, W. (2019). Differentially Private Obfuscation of Facial Images. In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2019. Lecture Notes in Computer Science(), vol 11713. Springer, Cham. https://doi.org/10.1007/978-3-030-29726-8_15

[27] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li and D. Wen, "SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition," in IEEE Transactions on Image Processing, vol. 30, pp. 2587-2598, 2021, doi: 10.1109/TIP.2020.3048632.

[28] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. 2020. Fawkes: protecting privacy against unauthorized deep learning models. In Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20). USENIX Association, USA, Article 90, 1589–1604.

[29] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4396-4405, doi: 10.1109/CVPR.2019.00453.

[30] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li and D. Wen, "SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition," in IEEE Transactions on Image Processing, vol. 30, pp. 2587-2598, 2021, doi: 10.1109/TIP.2020.3048632.

[31] Q. Meng, S. Zhao, Z. Huang and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 14220-14229, doi: 10.1109/CVPR46437.2021.01400.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 13:30 ~ 15:40 | T3-A オンライン（Zoom Events)

SNS分析 (2)

座長:荘司 慶行(静岡大学)
コメンテータ:岩井原 瑞穂(早稲田大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T3-A-2-01] XとMastodonへ投稿されるトピックの比較

林 駆[1]、*廣中 詩織[1]、首藤 一幸[1] (1. 京都大学)

13:50 ~ 14:15
[T3-A-2-02] 意外性のある意見がもたらすエコーチェンバー現象の分析

*中川 啓[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

14:15 ~ 14:40
[T3-A-2-03] 日米間におけるTwitterユーザのCOVID-19ワクチンに対する接種スタンス形成過程の比較

*張 翔[1]、久光 祥平[1]、金 洪善[2]、豊田 正史[2]、吉永 直樹[2] (1. 東京大学、2. 東京大学生産技術研究所)

15:05 ~ 15:30
[T3-A-2-05] 受験向けSNSの記事分析に基づく性別推定に関する研究

*寺口 敏生[1]、道谷 輝[1]、丸山 亜希子[1]、上田 真由美[1] (1. 流通科学大学)

X と Mastodon へ投稿されるトピックの比較

廣中　詩織† 　林　駆† 　吉田　光男†† 　首藤　一幸†

† 京都大学　〒 606-8501 京都府京都市左京区吉田本町
†† 筑波大学　〒 112-0012 東京都文京区大塚 3-29-1
E-mail: hironaka@media.kyoto-u.ac.jp, hayashi.kakeru.63z@st.kyoto-u.ac.jp, mitsuo@gssm.otsuka.tsukuba.ac.jp, shudo@media.kyoto-u.ac.jp

あらまし　ソーシャルメディアデータは、社会の傾向や、特定のトピックへの意見や反応の分析に使われてきた。近年、Mastodon などの分散型ソーシャルメディアの利用者が増加している。しかし、中央集権型のソーシャルメディアである X と比べて、これらのプラットフォームでどのような投稿がされているのかや、X から得られるデータの代わりになるのかは明らかになっていない。本研究では、トピック分析により X と Mastodon のプラットフォームごとの投稿内容の特徴を調べる。その結果、X ではマーケティング施策の成果である投稿やインプレッション稼ぎを目指した投稿が多いこと、Mastodon インスタンスの 1 つである mstdn.jp では社会に関する少しまじめな投稿が多いこと、pawoo.net ではイラスト関連の投稿が多いことがわかった。

キーワード　SNS、分散型ソーシャルメディア、トピック分析、テキストマイニング、自然言語処理応用

1 はじめに

ソーシャルメディアは多くの人に日常的に利用されており、生活に欠かせないものとなってきている。ソーシャルメディアのデータは社会の傾向や、特定のトピックに対する人々の意見や反応の分析に使われてきた。特に X[1] のデータは多くの研究で使われてきたが、2023 年に API の利用制限が厳しくなり、データの取得がしにくくなった。

X をはじめとする多くのソーシャルメディアは、1 つの企業がサービス提供者となっている中央集権型のソーシャルメディアである。一方、分散型のソーシャルメディアも存在する。分散型ソーシャルメディアでは複数のサービス提供者による複数のサーバが存在し、同一のプロトコルを実装したサーバ群で 1 つの大きなソーシャルネットワークを形成する。ユーザはそれぞれ好きなサーバへ接続してサービスを利用することになる。ActivityPub を実装した有名なソフトウェアとして Mastodon がある。近年分散型ソーシャルメディアの利用者が増加しているが、これらのプラットフォームでどのような投稿がされているのかや、X から得られるデータの代わりになるのかどうかは明らかになっていない。

分散型ソーシャルメディアや Mastodon を対象とした研究はまだ数が少ない。Mastodon のネットワークの成長やインスタンス間の関係に着目し、インスタンスのネットワークの分析をした研究 [1–4] や、ある 1 つのインスタンスのトレンドを分析した研究 [5] がある。しかし、複数のインスタンスを対象としたトレンドの比較は知る限りまだおこなわれていない。その他には、Mastodon が話題となった時期における、X から Mastodon への移住現象の分析がいくつかある [6, 7]。

本論文では、日本のユーザが多い Mastodon の 2 つのインスタンスと X を対象に、トピック分析によりプラットフォームごとの投稿内容の特徴を調べる。まずトピック数を変えて複数のトピックモデルを学習し、その中から分析に使うトピックモデルを選ぶ。そして、トピックモデルを用いて、各プラットフォームの投稿内容の特徴を調べる。その結果、X にはマーケティング施策の成果である投稿やインプレッション稼ぎを目指した投稿が多いこと、mstdn.jp は社会に関する少しまじめな投稿が多いこと、pawoo.net にはイラスト関連の投稿が多いことがわかった。

2 データ

本研究では、投稿トピックの比較のために、X と Mastodon の 2 インスタンスの投稿データを用いる。本章では、データの収集方法を説明し、トピック分析のためのデータの前処理方法を説明する。データの基本的な統計情報も合わせて報告する。

2.1 データ収集

X と Mastodon の Streaming API を利用してデータを収集する。Mastodon は分散型ソーシャルメディアプラットフォームであり、インスタンスと呼ばれる単位でサービスが独立している。ユーザは自由にインスタンスを開設、運用できる。また、インスタンスを開設しなくとも、他者のインスタンスへ参加することでソーシャルメディアが利用できる。Mastodon のデータを収集する際は、インスタンスごとに個別に収集する必要がある。本研究では、日本国内で最大規模のインスタンスである mstdn.jp と、イラストレーションに関連する話題が中心となっているインスタンスである pawoo.net のデータを用いた。

X からは Streaming API を用いて、日本語に限定してツイートを収集した。Mastodon の mstdn.jp と pawoo.net の 2 つのイン

This is an unrefereed paper.

1：https://twitter.com/ (accessed 2024-02-13)

スタンスからは、ローカルタイムラインを取得した。インスタンスのローカルタイムラインには、投稿の公開範囲が「公開」に設定されている、そのインスタンス内のユーザからの投稿が含まれる。Mastodon は日本語に限定していないが、日本のユーザが多いインスタンスであるため、日本語の投稿が多いと期待している。

実際に、2023 年 5 月 3 日から 6 月 23 日の期間[2]に、X からは 125,703,990 件、mstdn.jp からは 999,392 件、pawoo.net からは 425,138 件の投稿を収集した。

2.2 前　処　理

収集した投稿から投稿内容を取り出し、分析に用いるため、次の手順で前処理をする。

（1） 投稿内容の抽出、文字列の正規化
（2） 語彙の決定
（3） コーパスの作成

まず、収集した投稿からリポストやブースト（どちらもリツイートと同様の機能による投稿）を除外した。Mastodon の投稿には html タグが含まれているため、html タグを取り除いた。さらに、X と Mastodon の投稿内容から URL とユーザ ID を除去し、neologdn[3]を用いて文字列正規化を行った。

次に、Vibrato[4]と mecab-ipadic v2.7.0 を用いて、形態素解析を行った。そして、品詞が非自立以外の名詞、接尾、非自立以外の動詞、形容詞、副詞、連体詞、感動詞、記号と判定された単語以外を文書から除外した。動詞は、辞書に登録されていた基本形へ直した。1 文字のひらがなまたはアルファベットからなる単語と、数字だけからなる単語も除去した。さらに、すべての投稿のうち 1 割以上に出現する単語を除外し、単語の出現する文書数が多い順に 50,000 単語を語彙として選択した。

各投稿から語彙に含まれる単語だけが残るように文書コーパスを作成した。語彙に含まれる単語が含まれない投稿は除外した。

最終的に、X の投稿 17,432,493 件、mstdn.jp の投稿 767,746 件、pawoo.net の投稿 268,614 件が用意できた。収集したデータを前処理した後の X と Mastodon の 2 つのインスタンス（mstdn.jp と pawoo.net）の投稿数を 8 時間ごとに集計した結果を図 1 に示す。X の投稿数が Mastodon の投稿の 10 倍以上あることがわかる。前処理後の 1 文書あたりの平均単語数は、X で 8.01 単語、mstdn.jp で 8.81 単語、pawoo.net で 10.55 単語だった。

3 トピックモデル

Biterm Topic Model (BTM) [8] は短い文章のために提案されたトピックモデルである。短い文書では共起する単語が少ないため、従来の Latent Dirighlet Allocation (LDA) [9] などのトピックモデルでは学習が難しいという問題があった。BTM は文書

2 : 6 月 23 日以降は X の API が利用できなくなった。
3 : https://github.com/ikegami-yukino/neologdn
4 : https://github.com/daac-tools/vibrato

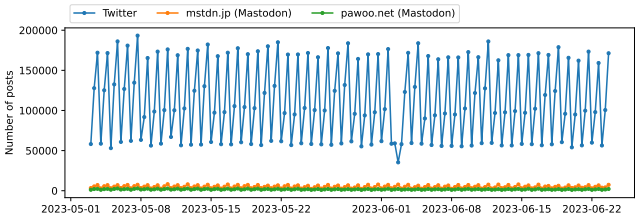


図 1　投稿数の推移

をモデル化する代わりに、文書中の biterm という非順序の単語対をモデル化することで、短い文書のトピックを学習しやすくしたモデルである。

コーパス D ($d \in D$) とトピック数 K を入力すると、トピック k が選ばれる確率 θ_k と、単語 v がトピック k を選ぶ確率 ϕ_{kv} が学習できるとする。そして、これらのパラメータから、文書のトピック分布 $P(z = k \mid d)$ が推定できるとする。本研究では、BTM の計算に bitermplus[5]を用いた。ハイパーパラメータとして $\alpha = 6.25$, $\beta = 0.01$ を設定した。

ある文書集合の特徴を調べるため、文書集合のトピック分布を定義する。文書集合 D_F のトピック分布は次の式で計算する。

$$\frac{1}{N_F} \sum_{d \in \boldsymbol{D}_F} P(z = k \mid d) \tag{1}$$

ここで、N_F は D_F に含まれる投稿数（$|\boldsymbol{D}_F| = N_F$）である。

4 分　　析

プラットフォーム間の比較をするため、まず用意したコーパスを使ってトピックモデルを学習し、学習したトピックモデルの妥当性を確認する。その後、推定したトピックを用いて、プラットフォームごとに投稿されるトピックに特徴があるかを調べる。

4.1 トピック数を変えて学習したトピックモデルの比較

トピックモデルを分析に使うためには、トピックモデルを学習する際のパラメータであるトピック数をデータに合わせて適切に決める必要がある。ここでは、トピック数を変えてトピックモデルを学習し、分析に使えそうなトピックが得られるかを調べた。

X と Mastodon のすべてのコーパスを合わせたものを用いて、トピック数を 8, 10, 16, 20, 30, 40, 100 に設定し、BTM を学習した。紙面の都合上、トピック数を 30 に設定して学習した各トピックを構成する単語を表 1 から 3 に示す。トピック番号は 0 から始まり、トピックの確率 θ_k が高い順に並べてある。さらに、トピックごとに ϕ_{kv} が高い順に単語が並んでいる。

表 1 から 3 より、トピック 0 から 10 には口語的な単語が多く含まれているトピックであることがわかる。トピック 0 は絵文字が多め、トピック 3 は YouTube、トピック 4 は動物、トピック 8 はスポーツに関する単語が含まれているなどの特徴がある。トピック 10 は生活、トピック 11 は社会や政治、トピック

5 : https://github.com/maximtrp/bitermplus

表 1　トピック数 30 の単語（トピック 0 から 9）

	0	1	2	3	4	5	6	7	8	9
1	さん	なる	なる	なる	ああ	ある	なる	ある	さん	いる
2	ちゃん	言う	行く	行く	さん	思う	死ぬ	なる	なる	なる
3	ありがとう	思う	今日	今日	ちゃん	なる	見る	おお	選手	思う
4	✨	そう	ある	ある	可愛い	自分	思う	やる	ある	ある
5	今日	ある	そう	そう	かわいい	言う	ある	そう	思う	この
6	くん	ない	思う	見る	見る	ない	好き	思う	試合	見る
7	❤️	寝る	買う	思う	好き	それ	言う	これ	見る	さん
8	頑張る	自分	時間	時間	くん	そう	そう	見る	今日	言う
9	（ハートの絵文字）	いい	やる	さん	ちる	その	さん	さん	やる	そう
10	おはよう	やる	明日	やる	なる	やる	欲しい	出る	この	名前
11	素敵	いる	仕事	明日	いい	できる	あなた	いい	そう	知る
12	誕生	それ	見る	来る	推す	いい	いい	この	勝つ	○
13	😊	もう	もう	いい	ありがとう	わかる	やる	ない	チーム	いい
14	嬉しい	見る	終わる	仕事	いる	見る	これ	今日	いい	好き
15	なる	これ	いい	😭	最高	これ	ない	作る	ない	やる

表 2　トピック数 30 の単語（トピック 10 から 19）

	10	11	12	13	14	15	16	17	18	19
1	ある	日本	いい	✨	お願い	結果	なる	さん	飲む	✨
2	なる	ある	繋がる	さん	交換	参加	さん	utm	ある	おはよう
3	思う	なる	さん	ありがとう	よろしく	▼	ありがとう	画像	なる	今日
4	ない	言う	女子	おめでとう	ありがとう	毎日	好き	配信	思う	さん
5	言う	思う	アカ	こちら	可能	ありがとう	見る	エロ	投稿	☀️
6	できる	ない	見る	開催	失礼	抽選	思う	見る	感想	→
7	そう	その	なる	楽しみ	♀	はずれ	ある	好き	さん	ありがとう
8	自分	さん	好き	♪	🙇	キャンペーン	繋がる	開始	当たる	⏩
9	その	この	男子	思う	フォロー	当選	いい	作品	見る	😊
10	仕事	それ	動画	なる	こちら	公式	反応	source	そう	考える
11	いい	女性	RT	楽しむ	ある	応募	嬉しい	なる	この	‼️
12	さん	ニュース	DM	質問	検討	残念	描く	medium	ツイート	頑張る
13	それ	いる	送る	この	検索	明日	✨	アニメ	瞬間	感謝
14	やる	できる	撮る	😊	希望	挑戦	この	イラスト	フォロー	人生
15	この	問題	ある	🎉	さん	チャンス	呼ぶ	くん	‼️	クロ

表 3　トピック数 30 の単語（トピック 20 から 29）

	20	21	22	23	24	25	26	27	28	29
1	食べる	販売	なる	募集	手押し	×	応募	コード	is	なる
2	なる	楽天	ある	やる	大阪	参加	ありがとう	割引	and	コロナ
3	当たる	Amazon	できる	配信	配達	生放送	ローソン	ポイント	the	ちゃう
4	餃子	価格	使う	ゲーム	→	ある	クーポン	クーポン	to	別途
5	その	買う	お金	相手	♂	LIVE	チャレンジ	当たる	run	ご覧
6	✨	予想	思う	通話	👶	配信	♪	キャンペーン	of	URL
7	買う	ad	届く	さん	指定	DM	次回	入力	last	👊
8	思う	🔥	使える	繋がる	場所	運営	使える	できる	for	お送り
9	ハイ	★	時間	ルーム	東京	ライブ	差し上げる	紹介	in	巨根
10	タコ	ランキング	結果	いい	営業	中継	引券	招待	you	くださる
11	美味しい	商品	体験	できる	✨	SnowMan	使い方	✅	out	ある
12	選ぶ	情報	フリー	なる	🥦	金額	いずれ	こちら	quot	お世話
13	使う	ある	※	いる	ミナミ	交換	各種	セール	first	遅く
14	セット	予約	製品	プロセカ	ある	希望	ぜひ	プレゼント	team	思う
15	作る	人気	対応	Apex	🦁	お願い	飲料	✨	inning	そう

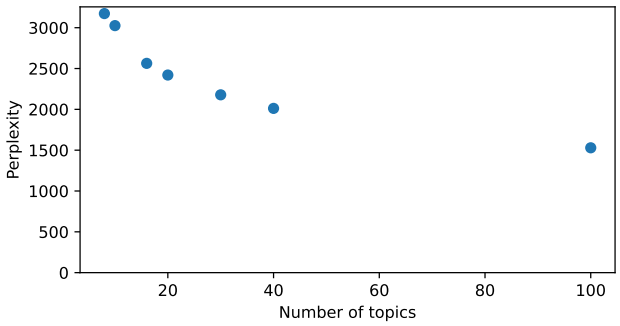


図 2　トピック数を変えて学習したモデルの Perplexity

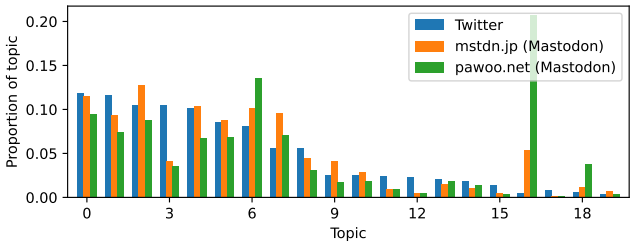
12 はフォロワー募集系、トピック 14 はグッズ交換系、トピック 16 はイラストでつながる系、トピック 20 はごはん、トピック 23 はゲームに関する投稿が含まれるトピックだった。トピック 15 はセブンイレブンの参加型キャンペーン、トピック 26 はローソンの参加型キャンペーンの投稿が多く含まれるトピックだった。

トピック 28 は英語のトピックになっており、図 3b のトピック 28 の割合を見ると、2 つの Mastodon インスタンスで英語の投稿が多いことになっている。これはデータ収集の際の設定の違いが原因である。X の投稿は日本語のものを選択して収集したが、Mastodon では言語の指定をしなかったため、X の投稿にはほとんど英語の投稿は含まれていない一方、pawoo.net と mstdn.jp から収集した投稿には英語の投稿が含まれていた。pawoo.net を mstdn.jp は日本のユーザが多いインスタンスであるが、ユーザは日本語ユーザに限られていない。データを確認すると X の投稿にも中国語など他の言語の誤判定された投稿が含まれていたため、追加で投稿の言語判定をすることは今後の課題である。

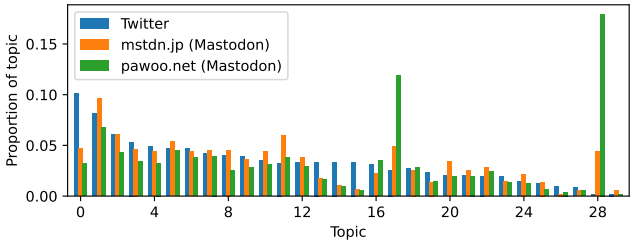
トピック数を変えて学習した他のトピックモデルの結果も確認したが、似たようなトピックが得られていた。トピック数が 30 より小さいと、複数のトピックが混ざっていると感じるトピックが増えた。トピック数を変えて学習したトピックモデルの Perplexity [9] を計算した結果を図 2 に示す。Perplexity の計算には、トピックモデルの学習に使ったコーパスと同じものを用いた。この結果から、大きなトピック数を設定するほど Perplexity が小さくなっていることがわかる。すなわち、今回使用したコーパスに適したトピック数はもっと大きいと考えられる。

トピック数を 20, 30, 40 に設定して学習したトピックモデルから計算した、各プラットフォームのトピック分布を図 3 に示す。各プラットフォームのトピック分布は、X, mstdn.jp, pawoo.net のそれぞれのプラットフォームで投稿された投稿の集合を D_F として式 (1) で計算したものである。いくつかプラットフォームに特徴的なトピックが存在しており、トピック数を変えても同じようなトピックが得られていることがわかる。

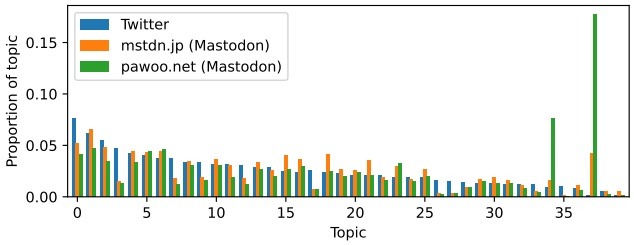
トピック数が多すぎてもトピックが解釈できず分析が難しくなるため、データをみた上で、今回分析に使うモデルはトピック数を 30 に設定して学習したモデルにした。



(a) $K = 20$



(b) $K = 30$



(c) $K = 40$

図 3　プラットフォームごとのトピック分布

4.2　投稿されるトピックの比較

プラットフォームへ投稿されているトピックの全体的な傾向を調べ、プラットフォームごとに投稿されるトピックに特徴があるかを調べる。各プラットフォームのトピック分布は図 3b に示したが、これはトピックの時間変化を考慮せず、すべての投稿をまとめて計算したものである。プラットフォームごとの投稿トピックの違いをさらに詳しく比べるため、8 時間ごとに各プラットフォームのトピック分布を計算し、プロットしたものが図 4 である。

図 3b より、X の投稿に多いトピックはトピック 0, 13, 14, 15, 26 であることがわかる。トピック 0 は絵文字が多く、配信者や推しの幸せを祈るタイプの投稿が多いトピック、トピック 13 はソシャゲのガチャやイベントに関するトピック、トピック 14 はグッズ交換に関するトピック、トピック 15 はセブンイレブンの参加型キャンペーンに関するトピック、トピック 26 はローソンの参加型キャンペーンに関するトピックである。企業のマーケティング戦略の一環としてユーザに投稿を促すキャンペーンは X に多く、Mastodon ではおこなわれていないことがわかる。また、図 4 より、トピック 15 やトピック 26 のような参加型キャンペーンは、時期により投稿量が増減していることがわかる。

図 3b より、mstdn.jp に多いトピックは、トピック 1、トピック 10、トピック 11 であった。トピック 1 は日常系の投稿、ト

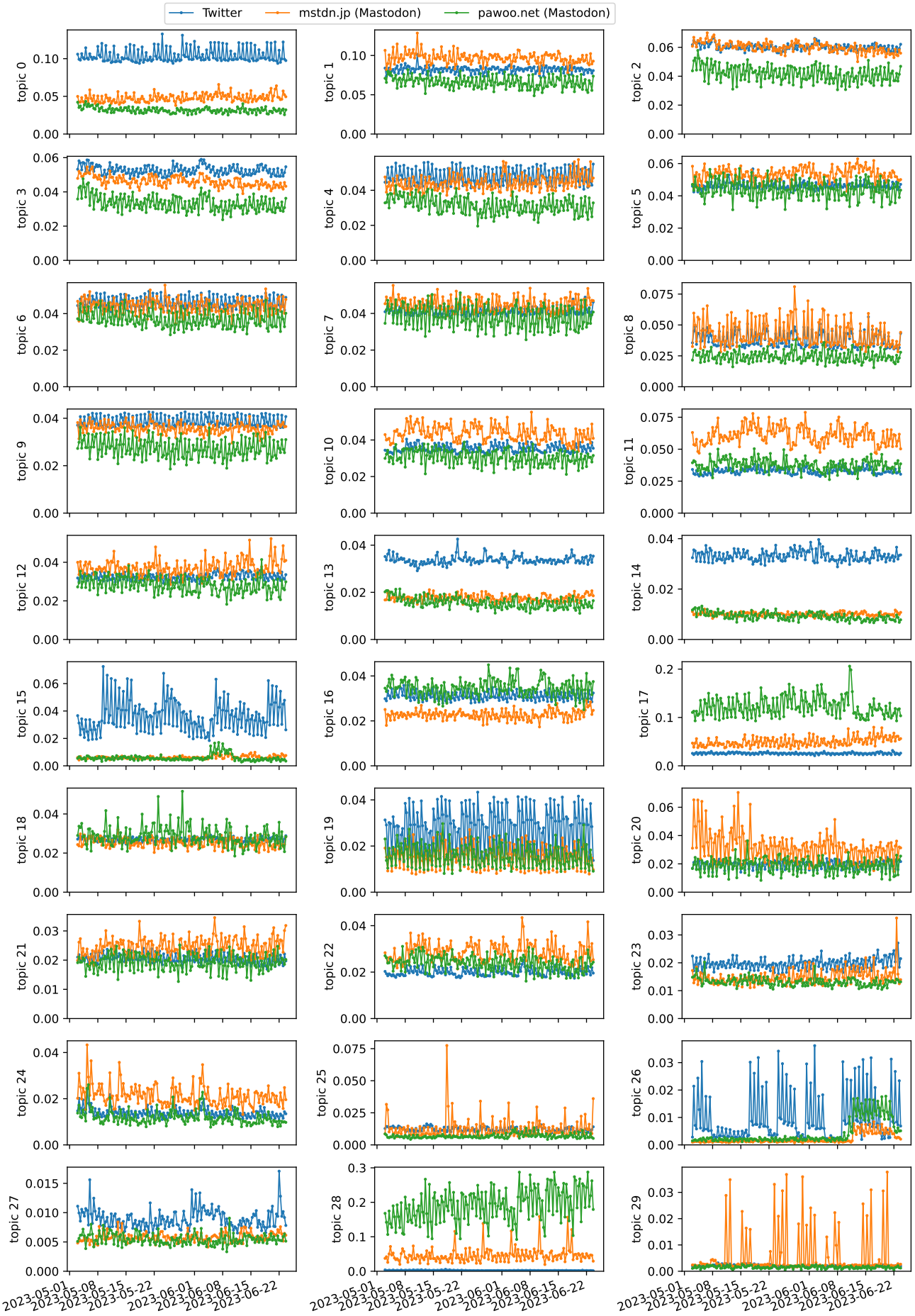


図 4　プラットフォームごとのトピック分布の時系列変化（$K = 30$）

ピック 10 は生活に関する少しまじめな投稿、トピック 11 は社会や政治に関する投稿が含まれるトピックである。図 4 を見ても、mstdn.jp にはまじめな内容の投稿が常時多めであることがわかる。ご飯に関する日常的な投稿が多く含まれているトピック 4 に注目すると、X と mstdn.jp とでは、常時同程度、日常について投稿されていそうだということがわかる。

図 3b より、pawoo.net に多いトピックは、トピック 17 とトピック 28 である。また、トピック 16 に関する投稿も X と同じくらい多い。トピック 17 はアニメや AI、イラスト他の投稿、トピック 28 は英語の投稿が多く含まれているトピック、トピック 16 はイラストコミュニティに関する投稿が含まれているトピックである。これには、イラストレーションを中心とした創作活動を好む人が多い Mastodon インスタンスであることが反映されていると考えられる。

以上の結果から、プラットフォームごとのおおまかな投稿トピックの傾向が捉えられたと考えられる。

5 考　　察

トピックモデルを用いた分析により、プラットフォームごとのおおまかな投稿トピックの傾向がつかめた。一方、Mastodon など異なるプラットフォームへの投稿をこれまで X 上でおこなわれてきた分析のデータソースの代わりとして使うことを考えた場合、特定のトピックそれぞれについて似たような傾向で投稿がされているのかどうかが気になる点である。

コロナウイルスやワクチンなど、社会的・政治的な投稿に注目して、プラットフォームごとの投稿トピックの傾向を比べる。コロナという単語の出現確率が高いトピックは 29 で [6]、ワクチンという単語の出現確率が高いトピックは 10, 11, 29 だった。トピック 10 や 11 は生活や社会に関する単語が含まれているトピックで、コロナやワクチン以外の話題も多く含んでいる。そのため、図 4 のトレンドが同じように変化するなら、同じようなデータソースとして使える可能性がある。しかし、今回は目的に純粋なトピック（例えば今回はコロナウイルスやワクチン）ではなく、社会的な話題のトピックの一部に含まれているため、投稿量のトレンドは比較できなかった。トピック数をさらに増やすなどの方法によって、目的の粒度のトピックが得られたら、そのような分析も可能であると考えられる。図 4 のトピック 10 や 11 のグラフを見ると、他のトピックよりも複雑な動きをしているため、時事ニュースなど外部の社会情勢を反映してトピックに関する投稿量が増減していそうであることがわかる。

社会の分析のためにソーシャルメディアデータを使うためには、偏りがないことと、量があることが必要である。mstdn.jp など、プラットフォームによっては政治に関する投稿の全体に占める割合が多い場合もあるが、投稿総数は X に比べると圧倒的に少ない（図 1）。多くの人が使っていて、日常的に意見を書き込む場であるからこそ X の投稿の分析に価値があったため、現状の Mastodon では投稿数が少ないので難しいと考えられる。また、Mastodon ではインスタンスによって特定のトピックの投稿を推奨している場合がある。政治に興味があるユーザが多いインスタンスのデータを使って分析をするとデータが偏る可能性があるが、それは X でも同じであり、今後偏りの少ないデータ収集方法の検討が必要である。

6 : トピック 29 はスパム的な投稿の影響が大きく、うまく学習できていない恐れがある。BTM が文書のトピックを推定する際、各文書はおおよそ同じ長さであることを仮定しているが、トピック 29 に含まれる単語を含む長大な投稿がコーパスに複数含まれていた。

6 おわりに

本研究では、X と Mastodon の 2 つのインスタンスから収集したデータを用いて、トピックモデル BTM を学習した。トピック数 30 で学習したトピックモデルを用いて、各プラットフォームの投稿内容の特徴を調べた。その結果、X にはマーケティング施策の成果である投稿やインプレッション稼ぎを目指した投稿が多いこと、mstdn.jp は社会に関する少しまじめな投稿が多いこと、pawoo.net にはイラスト関連の投稿が多いことがわかった。

今後、さらにトピックごとの投稿トレンドを捉えるためには、使用するトピックモデルの変更や、トピック数を増やしてトピックを自動的に解釈できるような手法が必要である。また、今回学習できたトピックに含まれる単語は、あくまで形態素解析に使った辞書に載っているものだけなので、新語などがトピックとして抽出できていない。他にも、「なる」「さん」などトピックとして解釈しづらい単語が多かったため、トピックごとの単語のランキング方法を工夫する必要があると考えている。

文　　献

[1] Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. Follow the "Mastodon": Structure and Evolution of a Decentralized Online Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12, No. 1, pp. 541–550, 2018.

[2] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. Challenges in the Decentralised Web: The Mastodon Case. In *Proceedings of the Internet Measurement Conference*, pp. 217–229, 2019.

[3] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. Understanding the growth of the Fediverse through the lens of Mastodon. *Applied Network Science*, Vol. 6, No. 1, pp. 1–35, 2021.

[4] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. Network Analysis of the Information Consumption-Production Dichotomy in Mastodon User Behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 1378–1382, 2022.

[5] Samer Al-khateeb. Dapping into the Fediverse: Analyzing What's Trending on Mastodon Social. In *Social, Cultural, and Behavioral Modeling*, pp. 101–110, 2022.

[6] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H. Russell Bernard, and Huan Liu. Exploring Platform Migration Patterns between Twitter and Mastodon: A User Behavior Study, May 2023. arXiv:2305.09196 [cs].

[7] Lucio La Cava, Luca Maria Aiello, and Andrea Tagarelli. Get Out of the Nest! Drivers of Social Influence in the #TwitterMigration to Mastodon, June 2023. arXiv:2305.19056 [physics].

[8] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 12, pp. 2928–2941, 2014.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

意外性のある意見がもたらすエコーチェンバーの分析

中川　啓† 清　雄一† 田原　康之† 大須賀昭彦†

† 電気通信大学　〒 182–8585 東京都調布市調布ケ丘１丁目５－１
E-mail: †nakagawa.akira@ohsuga.lab.uec.ac.jp, {seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　近年ソーシャルメディアの普及により, 情報の拡散速度が格段に向上した. これにより, ユーザは取捨選択して情報を閲覧するようになり, 情報の消費に偏りが発生している. 情報消費の偏りの代表的な現象として, エコーチェンバー現象が挙げられる. このエコーチェンバー現象の社会的な影響は多岐にわたるが, 代表例として健康と政治分野の事例が挙げられる. これらの事例から, 従来のエコーチェンバー現象の定義だと捕捉することのできない事象もあり, 本研究では新しいエコーチェンバー現象を提案した. 具体的な検証方法として, 意見が強化された時点と変化した時点の特異性分布違いや動画コンテンツが与える感情の違い, エコーチェンバースコアの分布の違いから, 意見が強化される時だけでなく, 変化する時もエコーチェンバー現象が発生することを実証した. 本研究で提唱したエコーチェンバー現象の定義は, 自身の思考に依らず興味関心を呼び起こす意見に没頭し, その考えのみが返ってくることにより, 意見が強化されるだけでなく変化する可能性のある状況である.

キーワード　行動分析, ソーシャルメディア, テキストマイニング, ニュース・マスコミ

1 はじめに

1.1 研究の背景

近年ソーシャルメディアの普及により, 情報の拡散速度が格段に向上した. これにより, ユーザは取捨選択して情報を閲覧するようになり, 情報の消費に偏りが発生している [1] [2] [3]. 情報消費の偏りの代表的な現象として, エコーチェンバー現象が挙げられる. このエコーチェンバー現象の社会的な影響は多岐にわたるが, 代表例として健康と政治分野の事例が挙げられる. 具体的には, COVID-19 ワクチンの安全性や効果に関するデマや誤情報が拡散された [4] [5]. また, 2016 年の米国大統領選挙では敵対陣営のデマを拡散し, 支持率に一定の影響を及ぼしたのではないかと考えられている [6] [7] [8]. このように, エコーチェンバー現象は社会に大きな影響を与えており, 発生メカニズムの解明が急務である.

エコーチェンバー現象はソーシャルメディアを利用する際, 自分と似た興味関心を持つユーザをフォローする結果, 意見を SNS で発信すると自分と似た意見が返ってくるという状況であると定義されている [2] [9] [10]. しかし, このエコーチェンバー現象の定義は不十分ではないかと考えている. 例えば, 図 1 にあるように, COVID-19 ワクチンの不妊デマにおいて, 従来のエコーチェンバー現象の定義だと, もともとワクチン反対派のユーザが「ワクチンは不妊の原因になる」という意見を目にすることで, ワクチン反対派の考えが強化される. 一方で, もともとワクチンに対して中立や賛成派のユーザが意外性のある「ワクチンが不妊になる」という意見を目にすることで, その意見に没頭し, ワクチン中立・賛成派から反対派に変化する可能性があるのではないかと考えている. このように, 従来のエコーチェンバー現象の定義では説明できないエコーチェンバー現象による社会問題が発生すると考えられるため, エコーチェンバー現象による社会問題を網羅的に捉えたエコーチェンバー現象の定義が必要である.

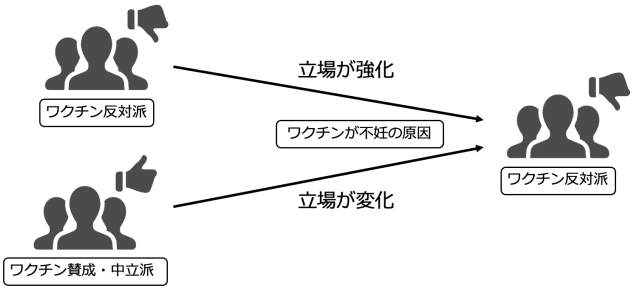


図 1　COVID-19 ワクチンの不妊デマに関するエコーチェンバー現象

1.2 研究の目的

本研究の目的は, エコーチェンバー現象をより正確に定義することである. そして, 正確な定義を捉えることで発生要因や影響を見直すことができ, エコーチェンバー現象による社会問題を網羅的に捉えることができる.

この研究を通して, エコーチェンバー現象による政治, 経済, 公衆衛生など, さまざまな分野でのエコーチェンバーによる社会問題の発生を抑制したり, 政府や企業がエコーチェンバー現象の影響を考慮して, 効果的に政策や戦略を策定することの一助となる.

2 先行研究

2.1 自然言語処理モデルに関する研究

本研究では, 3 つの自然言語処理モデルを用いて, ユーザの投稿内容や動画コンテンツが与える印象の分析を行う. 1 つ目は koheiduck/bert-japanese-finetuned-sentiment [11] である. このモデルを用いて, ユーザの投稿文をネガティブ/ポジティ

ブ判定（ネガポジ判定）し, ユーザのコンテンツに対する立場を分析した. 2 つ目は Mizuiro-sakura/luke-japanese-large-sentiment-analysis- wrime [12] である. このモデルを用いて, 動画コンテンツがユーザに与える感情を分析した. 3 つ目は sonoisa/sentence-luke-japanese-base-lite [13] を拡張したモデルである. このモデルを用いて, 取得したテキストから埋め込み表現を作成した. 以下で, 各モデルの性能について簡単に説明する.

koheiduck/bert-japanese-finetuned-sentiment にテキストを分類させることで, label と score を得る. label は POSITIVE, NEUTRAL, NEGATIVE のいずれかが指定されていて, score は判定の確信度を表している. 次に, Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime は luke-japanese-large-lite [14] を WRIME データセット [15] ファインチューニングしたモデルである. このモデルに文章を入力することで, 8 つの感情 (喜び, 悲しみ, 期待, 驚き, 怒り, 恐れ, 嫌悪, 信頼) がどれほど含まれているかどうかを検出することができる. 最後に, sonoisa/sentence-luke-japanese-base-lite は事前学習済みモデルとして studio-ousia/luke-japanese-base-lite [14] を利用した日本語用 Sentence-LUKE モデルである. このモデルに文章を入力することで, 文章の埋め込み表現を得ることができる. また, studio-ousia/luke-japanese-base-lite は Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime と sonoisa/sentence-luke-japanese-base-lite で事前学習済みモデルとして利用されていて, このモデルは JGLUE の開発セットで評価した結果, 従来のモデルよりも性能が上回っている.

2.2 エコーチェンバーに関する研究

Takikawa ら [16] の研究によれば, Twitter における政治に関するエコーチェンバー現象の分析結果から排外主義に関するトピックが右翼コミュニティのみで拡散しているということが明らかとなった. また, Cinelli ら [17] の研究によれば, Facebook と Twitter における同類的なクラスターに属するユーザがオンライン上で情報を共有していることが示された. これらの既存研究によれば, インターネット上における特定の主義や主張の共有は, 同質性の高いコミュニティでのみ行われていることがわかる. 一方で, 次の 3 つの既存研究によると, ソーシャルメディア上の情報拡散は必ずしも同質性の高いコミュニティ内だけで行われるものではないとされている. Pablo ら [18], Gianmarco ら [19] によると, 2016 年の米国大統領選挙において, トランプ支持者とクリントン支持者はソーシャルメディアで横断的な政治的交流を行なっていた. このことは, 同じ支持者同士での交流を好むとされていた従来のエコーチェンバー現象の定義に反するものであった. また, Axel ら [20] の研究によると, 政治問題においては従来のエコーチェンバー現象の定義通り同じイデオロギー的嗜好を持つ個人間で情報交換が行われていたが, 他の多くの時事問題では異質性の高いコミュニティ間で議論が活発に行われていたことが示された. したがって, ソーシャルメディア上の情報拡散は同質性の高いコミュニティ内だけでなく, 異なるコミュニティ間でも何かしらの要因によって引き起こされることがわかる.

2.3 意外性に関する研究

Delancy ら [21] の研究によれば, 意外性とは聞き手にとっての新情報, 予想外, 驚きといった意味を表すカテゴリーであるとされている. また, Li ら [22] は, フィルターバブルとユーザ離れ問題を解消するために, 意外性推薦システムについて研究している. 彼らによると, 新しいアイテムの意外性とは, そのアイテムの埋め込みと, 以前に消費された全てのアイテムの埋め込みとの間の距離であると定義している. 以上の 2 つの先行研究から, 本研究における意外性とは, 特異性, すなわちあるトピックと異なる情報と, 驚きの感情を与える意見を意外性のある意見と定義する.

3 エコーチェンバーの定義

本研究におけるエコーチェンバー現象の定義は, 自身の思想に依らず興味関心を呼び起こす意見に没頭し, その考えのみが返ってくることにより, 意見が強化されるだけでなく変化する可能性のある状況であるとする. 一方で, 従来のエコーチェンバー現象の定義は, 自身と似た考えを目にすることで, その意見に没頭し, 意見が強化される状況である [9] [10] [17]. したがって, 提案のエコーチェンバー現象の定義は, 過去の行動や興味に限定されない情報によってエコーチェンバー現象が発生し, 既存の意見や信念が強化されるだけでなく変化する可能性があるという点で従来のエコーチェンバー現象の定義と異なる.

4 方　　法

本研究では, 提案したエコーチェンバー現象の発生を定義に基づいて分析する. 具体的には, 図 2 に示すように, 意見が変化した際に動画コンテンツに意外性があるかどうかを分析し, その際のエコーチェンバー現象の発生度合いを評価する. 一方で, 意見が強化された場合には, 従来のエコーチェンバーとして検出する.

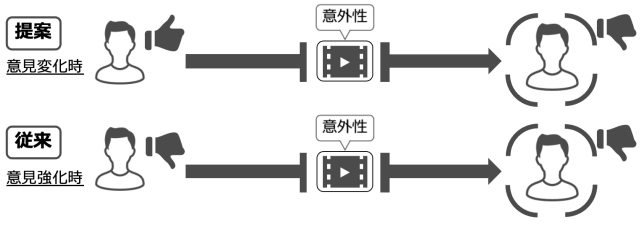


図 2　提案する現象の検出の概要図

4.1 提案のエコーチェンバー現象の検出手順

本研究では, まずデータ収集を行い, 収集したデータに対して前処理を行った. その後, 前処理を施したデータに対して, ユーザの投稿履歴において意見が強化された時点と変化した時点を特定した. ユーザの意見が強化された際と変化した際の動画コ

ンテンツの特異性, 動画コンテンツがユーザに与える感情, エコーチェンバー現象の発生度合いを示すエコーチェンバースコアを分析した. その分析結果を元に, 従来のエコーチェンバー現象の定義では説明できない現象を検出し, エコーチェンバー現象の定義の見直しが必要であることを示した. 以下で, 各手順の詳細を説明する.

4.2 データ収集手法

本研究では, YouTube が提供している YouTube Data API を利用して 40,000 件の動画と 4,970,000 件のコメントを YouTube から取得した. ここで, 動画の作成元であるチャンネルに偏りのある思想がある場合, コンテンツにも思想の偏りが生じ, 視聴者の思想も強化されるだけで今回検出したい意見の変化が観測できない可能性がある. そこで, できる限り中立の立場で一律に情報を発信する傾向のある大手ニュースチャンネルを動画の作成元として限定した. 具体的には, 日テレ NEWS, TBS NEWS DIG Powered by JNN, ANNnewsCH, FNN プライムオンライン, ABEMA Prime アベプラ【公式】を対象チャンネルとした. また, 動画のカテゴリの絞り込みを行わないとユーザの立場検出が困難となるため, 対象カテゴリを 1 つに絞りユーザの立場が強化, 変化したかどうかを検出するのを容易にした. 対象カテゴリは, 複数の既存研究で分析対象とされている COVID-19 とした. これらの条件を元に, 5 個のチャンネルに対して「コロナ」をクエリキーワードとして絞り込み, 視聴回数で並び替え, 2019 年 12 月 1 日から 2023 年 11 月 1 日までの期間の動画を取得し, その動画に投稿されたコメントも取得した. また, YouTube Data API のクエリだけであると COVID-19 と関連性の低い動画も取得してしまうため, 取得済みの動画タイトルに対して「コロナ, マスク, アクリル, 感染, 後遺症, 波, ウイルス, 緊急, 危機, 変異, 宣言」のいずれかが含まれる動画を対象とした.

4.3 前処理手法

テキスト分析を行う際に, 絵文字などのノイズが含まれるとエンコードの精度が落ちるため, 不要な文字の除去・統一を行った. 具体的に, 改行コード, 絵文字, URL, 全角半角の記号を除去, 大文字を小文字に統一, 全角を半角に統一する処理を行った.

4.4 ユーザの意見が強化した時点と変化した時点の識別手法

YouTube から収集したコメントは各動画に対して投稿されたコメント集合である. そのため, ユーザの投稿履歴を収集するためには, 各コメントに付与されている authorChannelId を元に各ユーザのコメント投稿履歴を作成する必要がある. また, 本研究の検証対象は COVID-19 に関する動画のコンテンツに絞ったため, ユーザの立場はコメントのネガティブ/ポジティブ判定（ネガポジ判定）をすることで, ユーザの COVID-19 に対する賛成・反対の立場を図ることができる. また, 各コメントのネガポジ判定を行う際に用いたモデルは, koheiduck/bert-japanese-finetuned-sentiment [11] である. このモデルに各コメントを入力し, ポジティブを正の値, ネガティブを負の値, ニュートラルを 0 として, ネガポジが-1.0 から 1.0 の範囲で出力されるようにした. そして, 意見が強化された時を正の値の時に値が増加する時と負の値の時に値が減少する時とし, 意見が変化した時を正の値の時に値が減少する時と負の値の時が値が増加する時とする. また, ネガポジ値がわずかに増減した時も意見が強化・変化した時として識別されると誤検出が多発してしまうため, 増減時の変化量として 0.2 以上の時を対象とした.

4.5 コサイン類似度

コサイン類似度を用いて, 動画コンテンツの特異性, エコーチェンバースコアを算出するため, 2 つのベクトル $\mathbf{X}$ と $\mathbf{Y}$ のコサイン類似度の導出方法を紹介する.

$$\cos(\mathbf{X}, \mathbf{Y}) = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\|\mathbf{X}\| \|\mathbf{Y}\|} \tag{1}$$

ただし, $\|\mathbf{X}\| = \sqrt{\sum_{k=1}^{n} X_k^2}$ はベクトルのノルム, $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{k=1}^{n} X_k Y_k$ はベクトルの内積を表す.

コサイン類似度の値は [-1, 1] の範囲をとり, 類似度が 1 に近いほどベクトル同士が類似していることを示し, 0 に近いほど類似性が低いことを示し, 類似度が-1 に近いほどベクトル同士が対向していることを示す.

4.6 特　異　性

特異性とは, 各動画コンテンツが他の動画コンテンツとどれほど異なるかを表す指標である.

総動画数を n として, 動画のリストを $V = \{v_1, v_2, \ldots, v_n\}$ とする. 各動画 v_i のタイトルに対して, studio-ousia/luke-japanese-base-lite を拡張した sonoisa/sentence-luke-japanese-base-lite [13] を用いて埋め込み表現 $EV = \{ev_1, ev_2, \ldots, ev_n\}$ を生成する. 次に, 動画 v_i の特異性 S_i を次のように計算する.

$$S_i = -\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \cos(ev_i, ev_j) \tag{2}$$

式 (2) では, 動画 ev_i と他の動画 ev_j とのコサイン類似度が計算され, これらのコサイン類似度を $(n-1)$ で正規化する. そして, コサイン類似度は-1 に近いほど他とは異なるコンテンツであることを示すため, -1 をかけ S_i の値が増加すればするほど特異性が増加するように調整した. この尺度は, 与えられた動画埋め込みリスト EV 内での動画埋め込み ev_i の特異性を示す.

4.7 動画コンテンツが与える感情

意見が強化された時と変化した時の動画コンテンツが与える感情の違いを分析するために, 各動画のタイトルから 8 つの感情 (喜び, 悲しみ, 期待, 驚き, 怒り, 恐れ, 嫌悪, 信頼) の内, どの感情が文章にどれほど含まれているかを検出した. ここで使用したモデルは, 最新の LLM である luke-japanese-large-lite を WRIME データセットでファインチューニングしたモデルである Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime [12] である.

感情の違いを分析する導出は次の通りである. まず, 各動画 v_i を 8 つの感情ベクトル $em_{i1}, em_{i2}, \ldots, em_{i8}$ に分解

した．次に，各動画の最大感情ベクトルのインデックスの出現数を数え，出現数のリスト $\mathrm{oem}_1, \mathrm{oem}_2, \ldots, \mathrm{oem}_8$ を作成した．そして，意見が強化された時の最大感情ベクトルの出現数 $\mathrm{oem}_{\mathrm{reinforced},1}, \mathrm{oem}_{\mathrm{reinforced},2}, \ldots, \mathrm{oem}_{\mathrm{reinforced},8}$ と意見が変化した時の最大感情ベクトルの出現数 $\mathrm{oem}_{\mathrm{change},1}, \mathrm{oem}_{\mathrm{change},2}, \ldots, \mathrm{oem}_{\mathrm{change},8}$ の比較を行うために，それぞれ百分率を取り，差分 $\mathrm{Diff}_1, \mathrm{Diff}_2, \ldots, \mathrm{Diff}_8$ を計算した．

$$\mathrm{Diff}_i = \frac{\mathrm{oem}_{\mathrm{change},i}}{\sum_{j=1}^{8} \mathrm{oem}_{\mathrm{change},j}} - \frac{\mathrm{oem}_{\mathrm{reinforced},i}}{\sum_{j=1}^{8} \mathrm{oem}_{\mathrm{reinforced},j}} \quad (3)$$

式 (3) で導出された $\{\mathrm{Diff}_1, \mathrm{Diff}_2, \ldots, \mathrm{Diff}_8\}$ を降順にソートすることで，意見が強化された時と変化した時の動画コンテンツが与える感情の違いを分析した．

4.8 エコーチェンバースコア

エコーチェンバースコアとは，各コメントが各動画に投稿されたコメント集合中でどれほど類似しているかを表し，そのコメント集合内でのエコーチェンバーの発生度合いを表す指標である．

エコーチェンバースコアの導出は次の通りである．動画 v_i の総コメント数を m_{v_i} とし，投稿コメントを $C_{v_i} = \{c_{v_i,1}, c_{v_i,2}, \ldots, c_{v_i,m_{v_i}}\}$ とする．sonoisa/sentence-luke-japanese-base-lite モデルを用いて，各コメントの埋め込み表現 $EC_{v_i} = \{ec_{v_i,1}, ec_{v_i,2}, \ldots, ec_{v_i,m_{v_i}}\}$ を作成した．次に，コメント $c_{v_i,j}$ とその他のコメントのコサイン類似度を計算する．

$$CS_{v_i,j} = \sum_{k=1,k\neq j}^{m_{v_i}} \cos(ec_{v_i,j}, ec_{v_i,k}) \quad (4)$$

最後に，コメント $c_{v_i,j}$ のエコーチェンバースコア $X_{v_i,j}$ の導出は次の通りである．

$$X_{v_i,j} = \sum_{k=1,k\neq j}^{m_{v_i}} \max(\cos(ec_{v_i,j}, ec_{v_i,k}), 0) \quad (5)$$

式 (5) で使用した max 関数は，2 引数の中で最大の値を返す．例えば，$\max(a, b)$ で ‘a‘が ‘b‘よりも大きい場合は ‘a‘を返し，そうでなければ ‘b‘を返す．

4.9 エコーチェンバー現象の妥当性検証手法

従来と提案のエコーチェンバー現象の定義の違いは，エコーチェンバー現象の発生要因とそれによる影響である．具体的には，従来は過去の行動や興味に基づく情報であるのに対し，提案は過去の行動や興味に限定されず興味を惹きやすい意外性のあるコンテンツによってエコーチェンバー現象が発生しているということを検証する．意外性のあるコンテンツは先行研究で定義されているように，他とは異なる内容であり驚きを与える内容とする．また，エコーチェンバー現象の影響については，従来は既存の意見や信念が強化されるだけであるのに対し，提案は既存の意見や信念が強化されるだけでなく変化する可能性があるということを検証する．これらの違いを検証するためにユーザの投稿履歴を用いて，どのようなコンテンツによって意見が強化，変化しているのか，またエコーチェンバー現象がどれほど発生しているかを分析する．そして，従来と提案のエコーチェンバー現象の区別をつけるために，従来は意見が強化された時，提案は意見が変化した時としてエコーチェンバー現象の定義の妥当性を検証する．

5 結　　果

5.1 特異性分布の違い

まず，ユーザの意見が強化された時と変化した時の動画コンテンツの特異性の分布の分析を行うことで，意外性の一要素である特異性の影響を分析することができる．意見が強化された時と変化した時の 2 つの分布の違いを調べるために，統計的仮説検定である t 検定を用いて，平均の差に有意差が認められるか否かについて調べた．その結果，意見が変化した時は強化された時より特異性の平均値が大きくなることが示された．よって，意見が変化する時は強化される時に比べて，目新しい動画コンテンツであることが多いことが示された．

5.2 動画コンテンツが与える感情の違い

次に，ユーザの意見が強化された時と変化した時の動画コンテンツが与える感情の違いを分析することで，意外性の一要素である「驚き」の影響を分析することができる．図 3 は，ユーザの意見が強化された時と変化した時の感情の分布の差異を検証するために，意見が変化した時の分布から意見が強化された時の分布を引いたものである．その結果，驚き，恐れ，嫌悪，信頼，怒り，悲しい，嬉しい，期待，信頼の順で変化した時の方が強化された時よりも感情の変化が大きいことが示された．よって，意見が変化する時は強化される時に比べて，「驚き」の感情を与える動画コンテンツであることが多いことが示された．

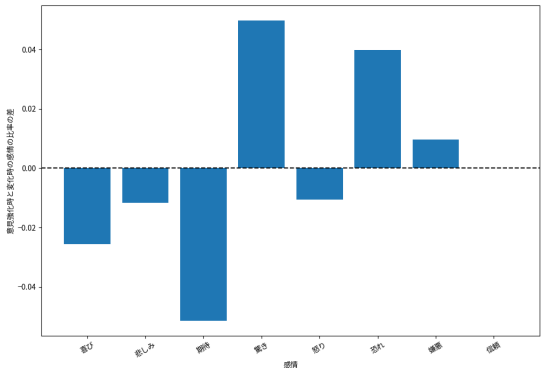


図 3　動画コンテンツが与える感情の違い

5.3 エコーチェンバースコアの分布の違い

そして，ユーザの意見が強化された時と変化した時のエコーチェンバースコアの分布の分析を行うことで，エコーチェンバー現象の発生度合いを分析することができる．図 4 は，ユーザの意見が強化された時と変化した時の分布の差異を検証するために，意見が変化した時の分布から意見が強化された時の分布を

表 1　ユーザの投稿データと各指標

UID	動画のタイトル	投稿内容	投稿日	エコーチェンバースコア	特異性	与える感情	投稿文の立場
A	新型コロナワクチン接種後に死亡の 26 歳女性に一時金認定 基礎疾患なし	若者減らしてどうすんだ！この国は！納税者に配慮して、役立たない役人が接種しても意味ない	2022-12-19 13:09:13	0.91	-0.029	驚き	-0.917
A	新型コロナ「5類」に“脱マスク”は進む?着用率「AI 分析」で見えた意外な結果	持病ある自分は外しません。持病無ければ周りが外して無くても自分は外してる。人は人で良いと思うし、マスク外す外さないは個人の自由。せっかく望んだ 5 類になったのに持病も無い病気も無い元気な人がマスクは要らないと思うよ。	2023-05-11 11:45:18	0.73	0.715	驚き	0
B	中国で“新型コロナ”感染急増 日本人も危機感	食料買いだめしとくか何かが起こりそう	2020-02-14T23:20:07	0.78	-0.298	恐れ	0
B	愛知県在住の 60 代男性 新型コロナウイルス感染確認	ガンガン増えて病院足りなくなるぞ。予測できたきたことだぞ政府さん。対岸の火事だと思うな。	2020-02-16 15:34:31	0.92	-0.074	驚き	-0.929

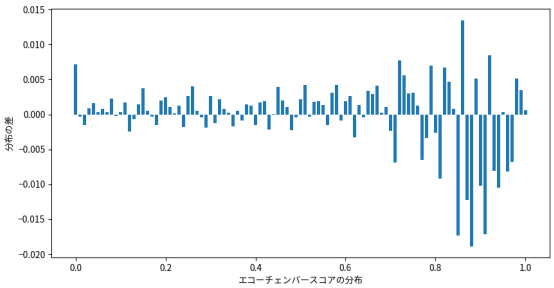


図 4　エコーチェンバースコアにおける分布の差異

引いたものである. その結果, 意見が変化したときのエコーチェンバースコアの分布が一貫して強化された意見を上回ることが明らかとなった. 定量的には, 正の値が 64 回発生し, 負の値が 37 回発生しており, エコーチェンバーの発生規模に関係なく, 意見が変化する事例が強化された意見よりも高い頻度で発生していることが示された.

5.4　提案のエコチェンバー現象の具体例

表 1 は, ユーザ A, B の投稿履歴の一部分である. なお, 論文掲載に伴って内容が変わらない程度で改変してある.

まず, ユーザ A に注目すると 2022 年に投稿されたコメントでは COVID-19 に対して負の感情を持っていたが, 2023 年に投稿されたコメントでは COVID-19 に中立の感情を持つようになった. この時の動画に着目すると, 特異性が比較的高く, 驚きの感情を与える動画コンテンツであり, エコーチェンバースコアが高いことが表 1 から示される.

また, ユーザ B に注目すると 2 月 14 日に投稿されたコメントでは COVID-19 に対して中立の感情を持っていたが, 2 月 16 日に投稿されたコメントでは COVID-19 に対して負の感情を持つようになった. この時の動画に着目すると, 特異性が比較的高く, 驚きの感情を与える動画コンテンツであり, エコーチェンバースコアが高いことが表 1 から示される.

6　議　　論

まず, 特異性が高く, 驚きの感情を与えるコンテンツは意外性のあるコンテンツとなるので, 結果より意見が変化した時は強化された時に比べて意外性のあるコンテンツが多く, 意外性のあるコンテンツによって意見が変化しやすいということがわかる. そして, 図 4 より, 意見が変化する時は強化される時に比べてエコーチェンバー現象の規模を問わずエコーチェンバー現象が発生していることがわかる. また, 意見が強化される時には規模が大きいエコーチェンバー現象が発生することもわかる.

また, 提案のエコーチェンバー現象の具体例からも, 比較的意外性のあるコンテンツによってユーザの立場が変化していることがわかる. そして, 意見が変化する時にユーザが比較的規模の大きいエコーチェンバー現象内部にいることもわかる.

本研究の制約は, 動画タイトルがコンテンツを適切に表現できているかどうかが不明瞭なことである. また, 先行研究でエコーチェンバー現象の発生度合いを示すエコーチェンバースコアのような指標が定義されていないことから, 客観的なエコーチェンバースコアの評価ができていないことである.

本研究はエコーチェンバー現象の発生度合いを定量的に測り, 従来唱えられていたエコーチェンバー現象の定義をビッグデータを用いて分析した初めての研究である.

次に, 驚きを与えるコンテンツによるエコーチェンバー現象の妥当性を考える. Vosoughi ら [23] によると, 人々はより目新

しい情報や, 驚き, 嫌悪, 恐れの感情を与える情報が共有する傾向があるということが示されている. また, Itti ら [24] らによると, 驚きは人間の注意を引き寄せることが示されている. さらに, Berger ら [25] らによると, 他のコンテンツより驚きを与えるコンテンツや興味深いコンテンツは拡散されやすいことが明らかであると示されている. これらの研究の結論は, 提案の定義に当てはまるエコーチェンバー現象が発生しやすいことを示している.

7 結論と今後の展望

本研究を通して, エコーチェンバー現象は意見が強化される時だけでなく, 変化する時にも発生することがわかった. 特に, 意外性のあるコンテンツによって意見が変化することが多いことがわかった. したがって, 従来のエコーチェンバー現象の定義では不十分であり, 興味を惹く意外性のあるコンテンツによってもエコーチェンバー現象が発生することを考慮した提案のエコーチェンバー現象の定義の方が適切であることが示された.

今後の展望として, 本研究では YouTube のコメントを用いてエコーチェンバー現象の定義を検証したが, X への投稿や Facebook の投稿などの SNS の投稿を用いてエコーチェンバー現象の定義を検証することで, エコーチェンバー現象の定義の汎用性を示すことができる. また, 今回提唱したエコーチェンバー現象の定義に当てはまる社会問題の分析を行うことで, エコーチェンバー現象の定義を拡張することの有効性を示すことができると考える.

謝　　辞

本研究は電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです. そして, 本研究で用いたモデルの開発者である山田先生, Studio ousia さん, Sakura Mizuiro さん, Sonobe Isao さん, Takami Kohei さんに感謝いたします.

文　　献

[1] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, Vol. 23, No. 1, p. 76–91, 2015.

[2] Andrei Boutyline and Robb Willer. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, Vol. 38, No. 3, pp. 551–569, 2017.

[3] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, Vol. 64, No. 2, pp. 317–332, 03 2014.

[4] Salman Bin Naeem and Maged N Kamel Boulos. Covid-19 misinformation online and health literacy: a brief overview. *International journal of environmental research and public health*, Vol. 18, No. 15, p. 8091, 2021.

[5] Nicholas B. Sajjadi, William Nowlin, Ross Nowlin, David Wenger, John Martin Beal, Matt Vassar, and Micah Hartwell. United states internet searches for “infertility” following covid-19 vaccine misinformation. *Journal of Osteopathic Medicine*, Vol. 121, No. 6, pp. 583–587, 2021.

[6] Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Fake news detection with generated comments for news articles. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, pp. 85–90, 2020.

[7] Rucker Philip and Costa Robert. Inside donald trump’s echo chamber of conspiracies, grievances and vitriol, 2016.

[8] Bump Philip. A new report reinforces the danger of a president who lives in an echo chamber, 2021.

[9] 総務省. インターネット上での情報流通の特徴と言われているもの, 2019.

[10] 鳥海不二夫ほか. 計算社会科学入門. Chuko shinsho 624. 丸善出版株式会社, 2021. (in Japanese).

[11] Takami Kohei. koheiduck/bert-japanese-finetuned-sentiment, 2022.

[12] Sakura Mizuiro. Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime, 2023.

[13] Sonobe Isao. sonoisa/sentence-luke-japanese-base-lite, 2023.

[14] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*, 2020.

[15] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2095–2104, Online, June 2021. Association for Computational Linguistics.

[16] Hiroki Takikawa and Kikuko Nagayoshi. Political polarization in social media: Analysis of the “twitter political field” in japan. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3143–3150, 2017.

[17] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, Vol. 118, No. 9, p. e2023301118, 2021.

[18] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, Vol. 26, No. 10, pp. 1531–1542, 2015. PMID: 26297377.

[19] Starnini Michele De Francisci Morales Gianmarco, Monti Corrado. No echo in the chambers of political interactions on reddit. *Scientific Reports*, Vol. 11, , 2021.

[20] Axel Bruns. Echo chamber? what echo chamber? reviewing the evidence. September 2017.

[21] SCOTT DELANCEY. Mirativity: The grammatical marking of unexpected information. Vol. 1, No. 1, pp. 33–52, 1997.

[22] Pan Li and Alexander Tuzhilin. Latent unexpected recommendations. *ACM Trans. Intell. Syst. Technol.*, Vol. 11, No. 6, sep 2020.

[23] Vosoughi Soroush, Roy Deb, and Aral Sinan. The spread of true and false news online. *Science359*, No. 1146-1151, 2018.

[24] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, Vol. 49, No. 10, pp. 1295–1306, 2009. Visual Attention: Psychophysics, electrophysiology and neuroimaging.

[25] Jonah Berger and Katherine L. Milkman. What makes online content viral? *Journal of Marketing Research*, Vol. 49, No. 2, pp. 192–205, 2012.

日米間におけるTwitterユーザのCOVID-19ワクチンに対する接種スタンス形成過程の比較

張　翔† 久光　祥平† 金　洪善†† 豊田　正史†† 吉永　直樹††

† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7–3–1
†† 東京大学生産技術研究所　〒153–8505 東京都目黒区駒場 4–6–1
E-mail: ††{cs,hisamits,jhs,toyoda}@tkl.iis.u-tokyo.ac.jp, ††ynaga@iis.u-tokyo.ac.jp

あらまし　ワクチンは伝染病に対する有効な施策の一つであり，COVID-19 に対しても日本では 2021 年 6 月に，アメリカでは 2020 年 12 月に大規模ワクチン接種が開始され，いずれも高い接種率を達成している．一方で，いずれにおいても一定数の人数がワクチンに対して忌避感を抱いており，広範なワクチン接種に対する障壁となるなど，ワクチン忌避は一つの社会問題になっている．本研究ではワクチン接種に対して肯定的あるいは否定的なスタンスを持つに至った人々がそのスタンスに至るまでの間に共有した情報を分析することで，それぞれのスタンスに人々を誘導しうるような特徴的な情報源を調査する．特に，日本在住 Twitter ユーザの日本語投稿データと米国在住ユーザの英語投稿データとを用いた時系列分析を行い，各国に特有の特徴や，国家を横断した特徴を調査する．

キーワード　マイクロブログ、SNS、テキスト分類、新型コロナウイルス、自然言語処理

1 はじめに

2019 年末の新型コロナウイルス流行以降，各国でワクチン接種が有効な施策として実施されている．日本では 2021 年 6 月から一般にワクチン接種が開始され，同年の 12 月末には 80% を超える 2 回目接種率を達成している [1]．このように，80% の 2 回目接種率を達成するのに要した期間が 6 か月というのは他国と比較しても短期間であり，例えば米国は 2020 年の 12 月にワクチン接種が開始されたものの，2021 年 12 月末の時点で 2 回目接種率は 63%ほどである．米国とほぼ同時期に一般接種を開始したドイツ，フランスでも同時期の 2 回目接種率は 70% ほどであり，日本の接種活動はかなり順調に進んだと言える．

日本でワクチン接種がこのように順調に進んだ理由としては，医療従事者や政府機関のウェブ上での盛んな情報発信が役立った可能性が報告されている [2]．しかし，ワクチンに対するスタンス形成に影響を及ぼした情報共有行動について 2 か国を比較した分析研究は我々の知る限り存在せず，これらの研究で報告されている要因が成り立った要因の詳しい分析はまだ調査されていない．

今回我々は日本および米国のワクチンに関する Twitter 投稿データを比較し，日本の順調なワクチン接種活動の要因を特定することを目指す．具体的には，日米ユーザのワクチンに関する投稿を比較分析を行うことによって，ワクチン接種を後押しした日本特有の特徴を特定することを目指す．分析の準備として，ワクチン関連 Twitter 投稿データセットの構築を行ったので，本稿ではその手順について記述する．また Twitter 投稿を用いたワクチンに対するスタンス形成過程，およびその過程に影響を及ぼした情報共有行動について日米それぞれのデータでの分析を行い，その結果を比較した結果を示す．

2 関連研究

ワクチン接種に対する世論を分析した研究の多くは SNS データを活用してきた [3]．特に SNS が出現して以降初めての世界的パンデミックであったことから COVID-19 ワクチンの文脈においてこの傾向は強い．SNS データの利用は従来の調査ベースの方法と比較して多くのメリットを有する．例えば，一般の人々の意見をタイムリーに観察することが可能であり，予防接種の意図や進行中の予防接種キャンペーンに関する態度をよりよく理解することにつながる [4]．本稿では我々は日本，および米国で最も人気の SNS の一つである X(旧 Twitter) を対象とした分析を行う．

我々は以前に日本語 Twitter データを用いてワクチン接種が順調に進んだ分析を実施しており，本研究もその成果を一部利用する [2]．この研究では，順調に進んだ要因の一つとして医療従事者や政府によるウェブや SNS 上での盛んな広 garcia2021topic 報活動が功を奏した可能性を提示した．本稿では，より進んだ分析として，同じ枠組みに基づいて他国のワクチン接種に対する世論分析も行い，その結果を比較することを目指す．

2 か国間の比較を行った研究として Klaifer ら [5] は COVID-19 に関する人気の話題についてブラジルのデータと米国のデータとでそれぞれ調査を行った．彼らは Twitter データを使用し，ブラジルと米国それぞれの Twitter ユーザが COVID-19 に関してどのような話題について話しているか調査し，感染拡大予防や感染者数，死亡者数の統計報告，経済への影響や防疫政策などについてどちらの国でも盛んに話題にされていることを報告した．彼らの報告は国家間の比較という点では我々の調査と同じだが，彼らは世論全体での話題の調査を主としている一方

表 1: アノテーションデータセット（日本）[2]

	訓練	開発	テスト
Pro	5237	392	276
Neutral	3063	96	154
Anti	450	12	66

表 2: アノテーションデータセット（米国）[6].

	投稿件数		
	訓練	テスト	定義
Pro	735	173	親ワクチン
Anti	531	75	反ワクチン
Hesitant	262	18	ワクチンを受けようか迷っている，あるいは接種しないと述べている
Irrelevant	864	144	ワクチンとは無関係な投稿
合計	2392	410	

で，我々は人々のワクチンに対するスタンス形成に影響を及ぼす要因についての調査を行っており，彼らの研究と我々の研究は異なる領域の調査を相補的に行っている関係にあると言える．

3 データセット

本節では，日米両国におけるワクチンに対する Twitter の投稿データセットの構築方法について説明する．なお，本研究で使用する Twitter データは NTT データより提供を受けたものである．

まず，日米それぞれについてワクチンに関連する投稿のデータセットを作成する．日本語データセットについては，我々が以前に日本のワクチン分析に使用した [2] ものを使用する．具体的には，全量 Twitter データのうち投稿本文に「ワクチン」という単語を含む投稿のみを抽出し，情報共有行動の分析に寄与しないと考えられる投稿やユーザを取り除くことで，日本語データセットを構築した．情報共有行動の分析に寄与しない投稿の例として，例えば，ニュースサイトなどで見られる「Twitter で共有」機能によってなされた投稿と，診断メーカーによって作成された投稿は，その投稿を行ったユーザの意図の推定に有用な情報を含んでいないと考えられるため取り除いた．また，本稿ではユーザが実施する情報共有行動に主な関心があるため，データ期間中に全く情報共有行動を行わなかったユーザ（リツイート，メンション，引用リツイートのいずれも 1 回も行わなかったユーザ）もデータセットから取り除いた．自動でワクチン関連投稿を行うボットアカウントについては，データセット内の全ユーザの 93.9%は 4 種類のクライアント（iPhone，Android，iPad，Web ブラウザ）からなされていたことを確認し，クライアント名に”bot”という語を含むユーザは約 1%ほどしか存在しなかったことを確認したため，本データセットには明白に bot と見なせるユーザは非常に少数であり，その分析結果に与える影響も小さいと考え，特に処理は行わなかった．以上の処理の結果，1,213,747 ユーザ，7,912,014 件のツイートを持つ日本語ワクチン関連投稿データセットを構築した．

英語データセットの期間は，米国のワクチン接種開始時期をカバーし，かつ我々の以前の分析と比較可能な機関とした．米国では 2020 年の 12 月から一般へのワクチン接種が開始したのでこの月をデータセットの収集開始期間とした．また，以前の我々の分析では 2021 年 6 月から 10 月の間を分析対象と設定したため，収集の終了期間は最低でも 2021 年の 10 月以降である必要がある．今回は，後述するユーザのスタンス分類器の訓練に用いるデータセットが 2020 年 12 月から 2022 年 4 月までをデータ期間として設定しているため，我々もこれに合わせて 2022 年 4 月を収集終了機関として設定する．今回，この 2020 年 12 月から 2022 年 4 月までの 17 か月の間の投稿から，以下のいずれかの単語を含む全投稿 50,413,744 件（8,378,484 ユーザ）を抽出した．

- vacc
- Vacc
- VACC
- vax
- Vax
- VAX

更に，日本の場合と異なり英語投稿を行っているユーザであっても必ずしも米国在住であるとは限らないため，ユーザのプロフィール情報を用いて米国に在住しているユーザの投稿のみを抽出した．具体的には，Twitter データの location フィールド中に「USA」「United States」を含む場合，あるいは location 欄がいずれかの州名に完全一致する場合にそのユーザの投稿のみを抽出した．これにより戦術した英語投稿の 20.1%にあたる 10,419,223 件（1,500,174 ユーザ）の投稿を含む英語ワクチン関連投稿データセットを構築した．以降の分析ではこれらの投稿を使用する．

4 スタンス分類

本節ではスタンス分類器の詳細について説明する．本研究では日米の Twitter ユーザがワクチンに対して抱くスタンスを元に分析を実施するが、それにあたって前節で収集したワクチン Twitter データセット中の各投稿に対して、その投稿をしたユーザがワクチンに対して抱いているスタンスのラベルを付与する．ただし，全投稿に対して人手でラベル付けを行うのは現実的ではないため人手でラベル付けされた少量のデータを使用して分類器を訓練し，その分類器を使って全データにラベル付けを行う．

ツイートへのスタンスアノテーション

本研究ではユーザは親ワクチン派（Pro-vaccine; ワクチン接種を予定している，接種をしたことを報告している），中立派（Neutral; スタンスを有しない），反ワクチン派（Anti-vaccine; ワクチン接種を避けている）のいずれかに分類される．分類器の予測精度を上げるために，我々は日本データに対してはデー

表 3: 各スタンス分類器の精度.

国	モデル	親ワクチン			中立			反ワクチン			Macro
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1
日本 [2]	BERT	0.875	0.837	0.856	0.698	0.766	0.731	0.462	0.375	0.414	0.667
	+DAPT	**0.929**	**0.851**	**0.888**	**0.738**	**0.825**	**0.779**	**0.524**	**0.688**	**0.595**	**0.754**
米国	COVID-BERT	0.780	**0.833**	**0.806**	**0.719**	0.531	0.611	0.622	**0.803**	0.701	0.706
	+DAPT	**0.818**	0.776	0.796	0.619	**0.662**	**0.639**	**0.724**	0.724	**0.724**	**0.720**

タセット中の一部の投稿を用いて人手でのアノテーションを行い，米国データに対しては既存のアノテーション済データセットを使用した．表 1 と表 2 に日米それぞれの各ラベルの内訳を示す.

日本アノテーションデータを作成するにあたって我々は 9250 件の投稿を用いて訓練用，開発用，テスト用のデータセットを構築した．我々は 4 名のアノテータを雇い，総計 9750 件の投稿に対してスタンスラベルを付与した．ただしこれらのアノテータがアノテーションした投稿の数は均等ではなく，そのまま開発データやテストデータを構築すれば多くアノテーションを行ったアノテータの判断基準が過度に反映されてしまう可能性がある．そこで，開発データ構築の際は各アノテータのアノテーションした投稿を 125 件ずつ抽出し 500 件の開発データを構築した．また，テストデータ構築の際は全アノテータに同じ内容の 500 件のツイートにアノテーションを行ってもらい，集計したラベルの多数決（同率の場合は協議によって解決）によってテストデータとした．最後に，残った 8750 件の投稿を訓練データとした.

一方，米国アノテーションデータには Mu ら [6] の構築したデータセットを用いた．彼らは訓練用に 2392 件，テスト用に 410 件の投稿に対して Pro，Anti，Hesitant，あるいは Irrelevant の 4 種類のラベル付けを行った．表 2 に各ラベルの定義とこれらのラベルのついた投稿の件数を示す．ただし，Mu らはワクチンを接種しないと述べたユーザは Hesitant として扱っているが，我々が日本語データ用に行ったアノテーションガイドラインに則るとそのようなユーザは Anti ラベルが付与される．そのため，訓練データ中の Hesitant ラベルがつけられた投稿は筆者が再確認し適切なラベルを再度付与している.

テキスト情報とリアクション情報を用いたスタンス分類

今回我々は日米それぞれのデータに対し同じアーキテクチャを持つが異なる分類器を訓練した．これらのモデルは投稿本文を入力として受け取る部分とあるユーザのリアクショングラフ上での情報を入力として受け取る部分を持ち，出力として受け取った投稿を行ったユーザが親ワクチン，中立，反ワクチンのそれぞれに該当する確率を返す．日米それぞれのモデルで異なるのは投稿本文を入力として受け取る部分の事前訓練済みモデルであり，日本語モデルでは NICT が公開している大規模事前訓練済 BERT モデル [1]を使用した．英語モデルについては Mu ら [6] が訓練した VaxxBERT モデルを使用した．これは Müller [7] らが公開している Twitter 上の COVID-19 関連投稿データを使用して事前訓練を行ったモデルを更に Twitter 上のワクチン関連投稿上で事前訓練 (DAPT [8]) したモデルである.

また，既存のワクチンスタンス研究では主としてテキスト情報のみを用いて SNS ユーザのスタンスを分類しようと試み，結果として正反対の情報に皮肉の目的などで参照した投稿を誤分類してしまうという問題があった．そこで本研究ではあるユーザが行った投稿のスタンスを推定するとき，そのスタンスを行う以前に誰に対してリアクション（リツイート，引用リツイート，メンション）を取っていたのかというリアクションの履歴情報もスタンスの予測に用いる．このために我々は各ユーザに，そのユーザがそれまでの一定期間内にリアクションを取った相手，リアクションを取られた相手を示すリアクションベクトル（RA ベクトル）を使用する．計算コストを軽減するために，最も影響力のあるユーザのみをリアクションベクトルに記録させることとした．具体的には分析期間中の各月を 1 日から 10 日，11 日から 20 日，21 日から月末までの 3 期間に分割し，各期間において最もリツイートしたユーザ，リツイートされたユーザを上位から 1 万人抽出し，影響力のあるユーザとして扱う．前者は盛んに情報発信を行う点で，後者はその影響力が拡散されやすいという点で影響力を有すると判断した．今回日米それぞれについてそのようなユーザを抽出し，各期間中複数回上位に出現するユーザの重複を取り除くと，日本で 95,016 人，米国で 249,194 人のユーザが獲得された．そのため，これらの 95,016 人（249,194 人）のユーザそれぞれに 1 つの次元を対応させた Onehot ベクトルを作成し，ある投稿に対してスタンス推定を行う場合，その投稿を行ったユーザが過去 3 期間中にこれら影響力の大きいユーザに対してリアクションを取っていた場合，そのユーザに対応する次元の値を 1 にすることで，あるユーザの過去のリアクション情報を記録させたベクトルを作成した．このリアクションベクトルは次元削減のために全結合層に入力され，tanh 関数を経由する．結局，我々の分類器は入力として投稿本文とその投稿を行ったユーザのリアクションベクトルを受け取り，その出力を softmax 関数に与えることで，その投稿を行ったユーザが新ワクチン派，中立派，反ワクチンのそれぞれに属する確率を計算することでスタンス推定を行う.

分 類 精 度

本節で説明した分類器の精度を確認するために，我々は BERT

1：https://alaginrc.nict.go.jp/nict-bert

と DAPT を行った場合の BERT とを比較する．特定タスクのドメインのデータセット中での事前学習 (DAPT) はそのタスクの精度を上昇させることが報告されている [8]．表 3 に日米それぞれの分類器の精度を示す．日本データでは DAPT を用いることで全ての予測精度が上昇しており，DAPT の有効性が確認された．一方，米国データでは Macro-F1 値は DAPT を行った方が良い結果を示しているものの，親ワクチンクラスの Recall, F_1 や中立クラスの Precision，反ワクチンクラスの Recall などで DAPT を行う前よりも精度が悪化していた．このことから，DAPT を行うことは必ずしも全ての面で予測精度を上昇させる訳ではないということが確認された．一方予測したスタンスに基づいた分析を行うことを考慮すると，親ワクチン，反ワクチンクラスの Precision が高いことが重要になり，どちらのクラスにおいても Precision の上昇をもたらしたという点で DAPT の導入は一定の有効性を示したと言える．

5 分　　析

我々は前述のスタンス分類器を用いて日米それぞれのワクチン関連投稿にスタンスラベルの付与を行った．これらのラベルを用いて，Twitter ユーザのワクチンに対するスタンス形成過程やその過程に影響を与えた情報共有行動の中に日米間でどのような差異が存在したのかを調査する．以下の全ての分析において X 社（旧 Twitter 社）の規約 [2]順守に細心の注意を払っている．また本研究の分析手法は我々が以前に実施した研究と共通している [9].

5.1 ユーザスタンスの遷移

分類器によって全ての投稿にラベルを付与した後，特定期間中に同一ユーザによって行われた投稿をスタンスごとに集約することによって，そのユーザのワクチンに対するスタンスを決定した．具体的には，短期間ではスタンスは変化しないという仮定の下で，分析期間中の各月を影響力のあるユーザを決定するときと同様に 3 期間に分割し，あるユーザがあるスタンスの投稿を他のスタンスよりも多く行っていた場合に，そのユーザはその期間中そのスタンスを持つものとして決定した．特に，異なるスタンスラベルを持つ投稿が同数同じユーザによってなされていた場合，予測精度の順に，親ワクチン派，中立派，反ワクチン派として取り扱う．図 1，図 2 に日米での各スタンスのユーザ分布を示す.

図 1 から，日本では親ワクチン派の数が最初の数期間においては中立派と同程度であったが，ワクチン推進キャンペーンが進むにつれて徐々に親ワクチン派が徐々に増加し，2021 年の 9 月になると急激に減少したことがわかる．この急激な減少の理由として，2021 年の 9 月は日本国民の 2 回目ワクチン接種率が 50%を達成した時期で，そのために親ワクチン派の人々がワクチンに対する関心を失ったという可能性が考えられる．また反ワクチンユーザの数は他スタンスと比較して期間中一貫して少なく，ワクチンキャンペーンなどは反ワクチン派の人々に訴求できていなかった可能性が示唆される.

一方図 2 を見てみると，米国では親ワクチン派の数は 2020 年 12 月半ば，2021 年 3 月半ば，2021 年 8 月初めと増減を繰り返しつつも複数回ピークに達していることがわかる．この増減傾向は中立派についても同様である．2020 年 12 月はファイザー社のワクチンが緊急承認され [3]，2021 年 2 月末日には接種が 1 回で済むとされるワクチンが緊急使用許可を獲得 [4]するなどワクチンに対する肯定的なニュースが多く報道されたことが反映されたものと考えられる．一方反ワクチン派のユーザ数は部分的に親ワクチン派と一致しているものの，2021 年 10 月初めなど，親ワクチン派の数が減少しているにも関わらず反ワクチン派が増加している時期も存在するなど，日本のほぼ変動の無い日本の反ワクチン派の推移と比べてかなりの増減傾向が見られる.

5.2 ユーザのスタンス形成

我々はワクチンに関する情報投稿を継続的に行っているユーザの分析を行うために，全期間の内，投稿を行った期間が 10 以上であるユーザのみを抽出し，それらのユーザのスタンス分布の推移を調査した．図 3，図 4 に日米それぞれのユーザのスタンス分布推移を示す.

図 3 から，日本では接種開始初期の 2021 年 6 月から中期の 2021 年 8 月にかけて中立派が徐々に減り，その分親ワクチン派が増えていることがわかる．この傾向は 2021 年 8 月頃には収まり，以降は各スタンスの割合はほぼ一定である．また日本では反ワクチン派は非常に少数で，その数もほぼ変動していなかったこともわかる.

一方で，図 4 によると米国では異なる推移が見られる．米国においても分析期間の初期である 2020 年 12 月から中期である 2021 年の 8 月頃にかけて中立派の割合が減り，その分親ワクチン派の割合が増加している点は同じであるが，その後再び親ワクチン派の割合が減少し中立派の割合が増加し直している点は日本の傾向と異なっている．特に，日米両国で 2021 年の 8 月までの期間で中立派の割合が減少し，その分親ワクチン派の割合が増加している点は特筆に値する．また，反ワクチン派についてもその割合がほぼ変動しない点は日本と同じであるものの，日本よりもかなり大きな割合のユーザが反ワクチン派となっていることが確認された.

6 おわりに

本稿では日本でワクチン接種が順調に進んだ要因を特定することを目的として，日米それぞれのワクチン関連投稿データセットを構築し，分類器を用いてスタンスの付与を実施した後に，既存の分析フレームワークに則り日米それぞれの Twitter データを分析した．結果として，日本の親ワクチン派のユーザ数は一時的に増加しピークを迎えたあとに減少することや反ワ

2：https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

3：https://www3.nhk.or.jp/news/html/20201213/k10012761401000.html

4：https://www3.nhk.or.jp/news/html/20210228/k10012889731000.html

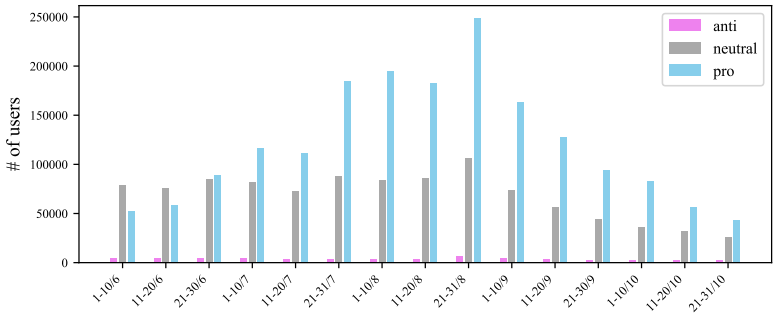


図 1: 各スタンスのユーザ数（日）.

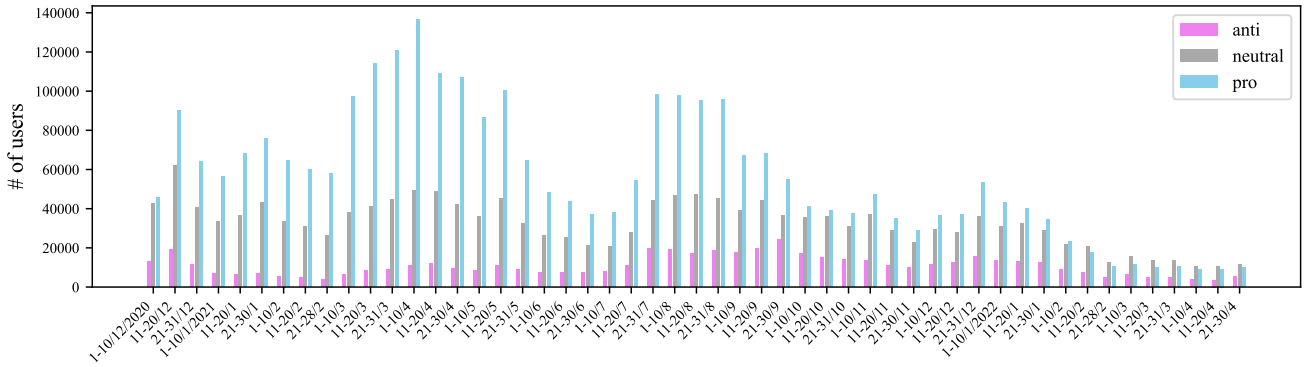


図 2: 各スタンスのユーザ数（米）.

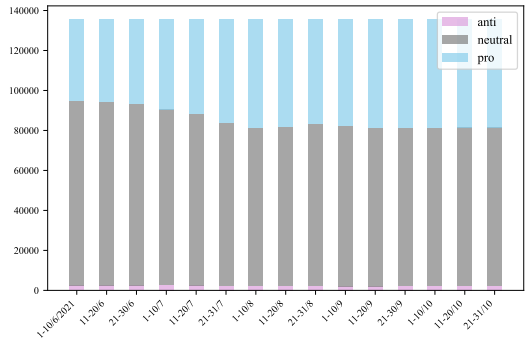


図 3: 継続的にワクチン関連投稿を行っているユーザのスタンス分布の推移（日）.

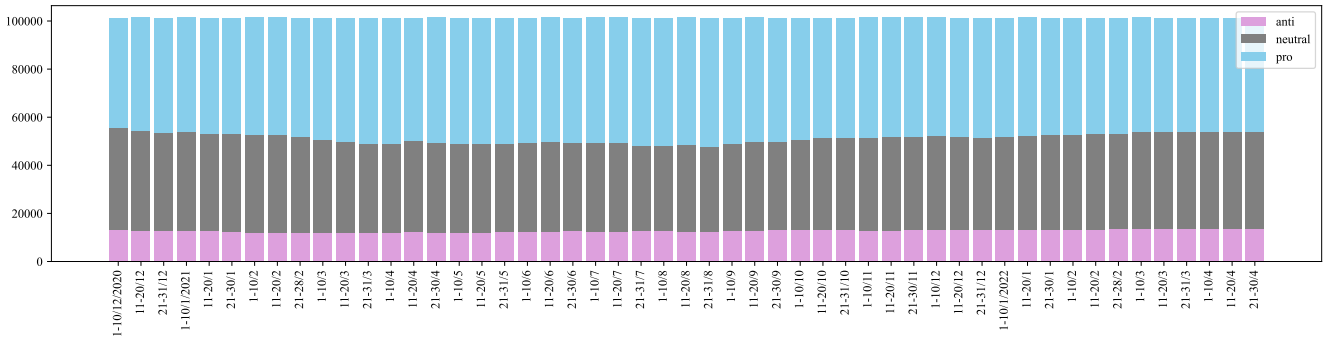


図 4: 継続的にワクチン関連投稿を行っているユーザのスタンス分布の推移（米）.

クチン派のユーザ数は非常に少ない数で安定している一方，米国の親ワクチン派のユーザ数は何度も増減しピークを迎えることや反ワクチン派のユーザの数も比較的多くその数も日本よりは激しく増減していることがが確認された．これは，特に米国が全世界でもかなり早期にワクチン接種を開始したために，どのスタンスであってもワクチンに関係するニュースに鋭敏に反

応する土壌があったためだと考えられる．またワクチンに関して継続的に投稿を行っているユーザの投稿を比較すると，日本では中立派の割合が減った分親ワクチン派の割合が増加し，途中からは割合にほぼ変化がなくなった一方で，米国では一度中立派が親ワクチン派に徐々に変化していった後，逆に親ワクチン派が中立派に戻っていることが確認された，中立派が親ワクチンに徐々に変化していく傾向は日米どちらでも 2021 年 8 月までに継続していることからこの時期の出来事に中立派の人々が親ワクチン派に変化する傾向を抑制するニュースなどが発信されたと考えられる．

謝　　辞

本研究は，三菱総合研究所が内閣官房の委託を受けて推進している COVID-19 AI・シミュレーションプロジェクトの一環として実施したものです．本研究で使用している分析手法は，JST CREST JPMJCR19A4，および，JSPS 科研費 JP21H03445 の支援を受けて開発したものです．

文　　献

[1] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao. A global database of covid-19 vaccinations. *Nature human behaviour*, Vol. 5, No. 7, pp. 947–953, 2021.

[2] S. Hisamitsu, S. Cho, H. Jin, M. Toyoda, and N. Yoshinaga. Diachronic analysis of users' stances on covid-19 vaccination in japan using twitter. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 237–241, 2022.

[3] F. Cascini, A. Pantovic, Y. A. Al-Ajlouni, G. Failla, V. Puleo, A. Melnyk, A. Lontano, and W. Ricciardi. Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *EClinicalMedicine*, Vol. 48, , 2022.

[4] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, Vol. 5, No. 7, p. e151, 2008.

[5] K. Garcia and L. Berton. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, Vol. 101, p. 107057, 2021.

[6] Y. Mu, M. Jin, C. Grimshaw, C. Scarton, K. Bontcheva, and X. Song. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17, pp. 1052–1062, 2023.

[7] M. Müller, M. Salathé, and P. E. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, Vol. 6, p. 1023281, 2023.

[8] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics.

[9] 久光祥平, 豊田正史, 吉永直樹, 張翔. Twitter ユーザの covid-19 ワクチン接種に対するスタンスおよびその分極化の推移に関する分析. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022), オンライン, 2022.

受験向け SNS の記事分析に基づく性別推定に関する研究

寺口 敏生† 道谷 輝† 丸山 亜希子† 上田 真由美†

†流通科学大学 〒651-2188 神戸市西区学園西町 3 丁目 1 番

E-mail: †{ Toshio_Teraguchi, Michitani_43210554, Akiko_Maruyama, Mayumi_Ueda}@red.umds.ac.jp

あらまし 企業は一般的に，女性が結婚や出産で仕事をやめやすいという統計に基づき，企業内訓練を控える傾向にある．この現象を経済学では「統計的差別」という．著者らは，保護者による子育ての段階でも同様の統計的差別が生じているかどうかをベネッセ教育総合研究所の調査データ等や SNS への自由投稿を使用して調査し分析している．特に SNS を活用した調査では，保護者が中学受験について相談し合う掲示板を対象に，男子校と女子校で交わされる話題の違いを分析してきた．しかし，これらの分析を共学校に適用しようとした場合，話題の対象が男子に関してか女子に関してか判別できない課題があった．そこで，受験向け SNS に投稿された記事から主題となる性別を推定するため，自然言語処理技術を活用し，男子校，女子高と共学校の掲示板に投稿される話題を分析する方法を研究した．

キーワード SNS，自然言語処理，統計処理，教育

1. はじめに

2023 年の世界経済フォーラムが発表した「ジェンダーギャップ指数」[1]によると，日本は 146 カ国中 125 位と，主要先進 7 か国（G7）では前年に引き続き最下位であった．これは，教育分野と健康分野が高スコアなのに対し，経済分野と政治分野のスコアが世界最低クラスに低いためである．その原因として，経済学の分野では，女性が結婚または出産により仕事をやめる可能性が男性よりも高いために，企業が女性社員への社内教育を積極的に行わないという「統計的差別」の存在が指摘されている．

この統計的差別の解消を目指し，同年に栃木県で開催された G7 男女共同参画・女性活躍担当相会合の共同声明[2]では，「役員及び管理職における女性の代表性は，拡大されるとともに，支援される必要がある」ことが明記された．また政府は，女性活躍・男女共同参画の重点方針 2023[3]にて，2030 年までに女性役員の比率を 30%以上とするという目標を掲げている．

これらの取り組みの一方で，三菱 UFJ リサーチ&コンサルティングの調査[4]によると，女性の非管理職のうち 5 割弱が「役職にはつかなくてよい」と回答している．その理由として，「家庭（プライベート）との両立が難しいため」「忙しくなる（業務量が増える）ため」「自分には管理職が向いていないと思うため」という回答が，男性と比べて多い傾向にあることが指摘されている．この昇進意向の性差は，統計的差別が企業内にとどまらず，就職するまでの保護者による教育段階においても存在し，女性の意思決定に影響を与えている可能性を示唆していると考えられる．

著者らは，こうした統計的差別が，保護者が子どもに教育を施す段階においても存在するのかを調査している．先行研究[5]では，ベネッセ教育総合研究所が実施した「子どもの生活と学びに関する親子調査 Wave4」の 2018 年のデータ[6]を分析し，子の性別による保護者の教育意識を分析し，以下のような結果が得られているところである．

- 子の進学段階のような将来に対する保護者の意識は，女子よりも男子の方に不安や熱意を感じている．
- 女子よりも男子の方が，勉強に関する保護者のかかわりが強い傾向にある．
- 保護者の行動を伴う意識のうち中学受験経験では，全国計でみると女子よりも男子の方が，受験経験が多い．

以上の結果より，女子よりも男子の方が，子供の頃から教育面での親の働きかけが多いと考えられる．

このような保護者の意識が，政府の取り組みやメディアの情報発信の影響を受けて，どのように経年変化するかを追跡するためには継続的な調査を必要とする．そこで，著者らは，保護者が自発的に話題交換する場である SNS（Social networking service）に着目し，投稿記事で扱われる話題を分析する方法を検討している．特に受験向けの SNS を分析することで，保護者の意識の変化やトレンドの経年調査を追跡できるものと考える．しかし，これらの分析を共学校に適用しようとした場合，話題の対象が男子に関してか女子に関してか判別できない課題があった．

図 1　カテゴリごとに生成したワードクラウド
（左：男子中学校，中央：女子中学校，右：共学中学校）

本稿では，受験向け SNS のうち，中学受験に関する Web 掲示板の投稿記事を分析し，男子中学校，女子中学校，共学中学校で交わされる話題に違いがあるかどうかを調査する．これにより，男女共学の学校に関する投稿においても，子供の性差が保護者の教育意識に影響を与えているかどうかを分析する．

2. 女子教育の重要性

女子教育の重要性は経済学の理論を用いると以下のように説明できる．

- 男女に関わらず一般的に，教育を受けると労働生産性が上昇し，賃金率が上昇する（人的資本論）ため，より豊かな暮らしが手に入る．
- 女性の賃金率上昇または労働環境の改善は，夫婦間における女性の交渉力を上昇させる効果がある[7]．

これは，賃金率上昇により女性が離婚を選択しやすくなることが，配偶者への暗黙の「脅し」として機能するためである．実際，文献[8]では，アメリカにおいて女性の賃金率上昇が妻へのドメスティック・バイオレンスを減少させたという調査結果が示されている．なお，この賃金率上昇は，女性全体の賃金率上昇でよく，本人の稼得所得の上昇でなくてもよい．

以上より，教育と労働生産性や賃金上昇率には密接な関わりがあり，女子教育の充実により賃金上昇率の改善が達成されることで，ジェンダーギャップの改善につながる可能性が示唆されている．

そこで著者らは，子供の性差が親の教育意識に与える影響に着目した調査を実施している．自由記述形式の受験用 SNS 等を通じてこの教育意識差の変化を分析できれば，政府の取り組みやメディアの情報発信の効果を計量できるようになる可能性がある．

表 1　対象としたデータ数

カテゴリ	メッセージ数	名詞出現総数
男子中学校	1,547 件	38,436 語
女子中学校	2,127 件	111,024 語
共学中学校	1,705 件	28,737 語

3. 受験向け SNS における親の意識調査

（1）調査対象データ

本稿では，株式会社インターエデュ・ドットコムが運営する掲示板形式の SNS であるインターエデュ・ドットコム[9]（以下，「インターエデュ」と呼ぶ）のデータを対象に分析を行う．当該 SNS は，幼稚園から大学，塾・予備校までの各カテゴリで専用の掲示板が設けられており，「中学校」カテゴリでは，主に中学受験に関する情報交換が行われている．中学受験準備は子自身ではなく親が主導で行うことが多く，掲示板の書き込みは，ほとんどが保護者によるものと考えられることから，保護者の子への教育意識の確認に活用する．

インターエデュの「中学校」カテゴリは，「国立中学校」「都立中学校」「公立中学校」「男子中学校」「女子中学校」「共学中学校」「男女別学中学校」のカテゴリに細分化されている．本稿では，この中の「男子中学校」「女子中学校」「共学中学校」カテゴリに掲載されている学校を分析対象とした．また，各カテゴリの内部には学校ごとの掲示板が存在し，各掲示板で学校に関連した話題を扱う複数のスレッドが設けられ，議論や意見交換が行われている．

本稿では，特に受験が活発といわれる東京 23 区の中学校を対象に，各学校の掲示板の 1 ページ目より，スレッドの最初の記事（以下，「トップメッセージ」という．）を調査対象として収集し，男子中学校，女子中学校と共学中学校の 3 つのカテゴリに分割した．本稿では，2024 年 1 月 5 日に取得したデータを対象に分析を行った．データ収集の結果得られたデータの

内訳を**表 1** に示す．取得したデータの分析にあたっては，前処理として，取得したデータから ID，投稿者や投稿日等の情報を削除し，匿名化する処理を施したものを用いた．

（2）分析方法

分析にあたっては，まず匿名化を施したトップメッセージを対象に形態素解析し，男子中学校，女子中学校，共学中学校の 3 つのカテゴリごとに，使用される名詞と複合名詞の出現回数を計量した．名詞だけでなく複合名詞を対象にしたのは，形態素解析の結果，たとえば「中学生」が「中」と「学生」で別々に計量されている事例が見られ，より語句の対象を絞り込む必要があると考えたためである．また，形態素解析時に，「の」「こと」「もの」「方」等のストップワード，「あれ」「それ」「あそこ」等の代名詞，および明らかに複合名詞の一部と思われる語句を除去する前処理を施した．

次に，名詞・複合名詞ごとに，出現するトップメッセージ数を計量した．そして，これらの計量結果を基に，各カテゴリにおける各名詞・複合名詞の特徴量を評価する tf-idf 値を算出した．なお，各カテゴリ内で出現回数が 1 回の名詞は，トップメッセージの特徴を表していないものと考え，前処理にて除外した．また，各カテゴリにおいて tf-idf 値が高い名刺・複合名詞を特徴語とし，共学中学校のトップメッセージから主題となる子供の性別を推定できるかどうかを評価した．

最後に，既研究[10]の成果である，人気校とそれ以外の学校掲示板によって話題が異なるのかの調査結果との比較を通じ，メッセージ中の記載内容により，子の性別によって保護者の教育意識の違いを分析可能かどうか評価する．

（3）分析結果

男子中学校，女子中学校，および共学中学校の 3 つのカテゴリより抽出した名詞から生成したワードクラウドを**図 1** に示す．また，各カテゴリに登場する名詞・複合名詞のうち，tf-idf 値が高い上位 40 語句を**表 2** に示す．

図 1 や**表 2** を確認したところ，順番の前後はあれ，どのカテゴリにおいても「学校」「受験」「中学」「お願い」「入学」「合格」等の一般的な語句が，特徴量の上位を占めている．これは，既研究[10]でも見られた特徴であり，分析対象とした受験用 SNS に記事を投稿する利用者は，子供の性別や学校のカテゴリによらず，全体的に進学意識が高いことが分かる．これらに加えて，男子中学校では「息子」，女子中学校では「娘」といった名詞が特徴語として頻出していることから，中学受験の掲示板には親の書き込みが多いという仮説が事実であることが確認できた．

表 2　各カテゴリの tf-idf 値上位 40 単語

カテゴリ	名詞・複合名詞
男子中学校	学校，お願い，中学，受験，息子，入学，合格，生徒，高校，もの，大学，今年，保護者，志望，親，先生，進学，説明，塾，偏差値，説明会，授業，質問，勉強，在校生，学園，入試，年度，家，現在，情報，部活，母，男，クラス，場合，ご存知，予定，時間，2 年
女子中学校	学校，受験，お願い，娘，女，入学，中学，女子，生徒，もの，合格，説明，親，説明会，先生，高校，志望，進学，大学，家，保護者，勉強，塾，質問，今年，授業，在校生，偏差値，服，ご存知，母，予定，場合，時間，入試，生徒さん，伺い，希望，参加，幸い
共学中学校	学校，受験，お願い，中学，入学，高校，生徒，合格，進学，大学，説明，女，説明会，もの，保護者，今年，男，先生，授業，偏差値，志望，情報，女子，在校生，勉強，学園，雰囲気，入試，塾，クラス，質問，子供，親，予定，男子，娘，英語，ご存知，場合，母

表 3　共学中学校の記事中の子供の性別と登場記事数

特徴語	性別	数	特徴語	性別	数
部活	男子	28	体操	男子	1
	女子	16		女子	1
服	男子	14	下校	男子	2
	女子	28		女子	3
野球	男子	7	国際	男子	4
	女子	1		女子	17
サッカー	男子	10	グローバル	男子	2
	女子	2		女子	3
スポーツ	男子	7	通学	男子	18
	女子	4		女子	22
ダンス	男子	2	制服	男子	11
	女子	7		女子	23

一方で，各カテゴリの tf-idf 値上位 40 単語を確認したところ，**表 2** の網掛け部分に示す通り，男子中学校では「部活」，女子中学校では「服」，共学中学校では「英語」が特徴的な語句として抽出された．また，上位 40 語以外も確認したところ，男子中学校では「野球部」「サッカー」「スポーツ」，女子中学校では「ダンス」「体操」「下校」「国際」「グローバル」等が特徴的な語句として抽出された．このことから，性別によって交わされる話題には多少の違いがあることが明らかとなった．さらに，たとえば「通学路」「時差通学」「通学バッグ」等，同じ名詞を含む複合語に範

表 5　トップメッセージの内容と評価結果例

トップメッセージの内容（※）	機械的判定結果	特徴語
（略）部活、バスケ部に入りたいという息子の希望があり、こちらの学校のバスケ部の活動状況や部員の人数等情報を教えていただけたら、幸いです（略）	男子	部活
ダンス部はどんなジャンルの踊りなのでしょうか？（略）娘はヒップホップ希望なのですが（略）	女子	ダンス
（略）携帯電話は学校に持って行くのは禁止ですか？　校則があれば教えてください（略）	不明	-
この春から入学予定の男子です。制服注文に際し、疑問があり、よろしければ教えてください（略）	男子（※誤判定） 正しくは「女子」	制服
（略）行列ができるレストラン、カフェテリア、ブックストア、セブンイレブンがあり、新しい女性専用のパウダールーム（略）「恋バナも、部活打ち合わせも、（略）で・・」なんてね（略）	女子（※誤判定） PR 投稿のため， 正しくは「不明」	部活

（※）個人情報を含む記述があるため，部分抜粋

囲を拡大すると，女子中学校で特徴を高く評価された名詞として「通学」や「制服」があった．共学中学校においては，他のカテゴリの両方で共通する特性を幅広く評価する傾向がみられた．

以上の分析から得られた各カテゴリの特徴語を共学中学校のトップメッセージに当て嵌め，主題となる子供の性別を推定できるか下記手順で分析した．

1) 「娘」「女」を含むトップメッセージの主題となる性別を機械的に「女子」と判定
2) 「息子」「男」を含むトップメッセージの主題となる性別を機械的に「男子」と判定
3) 上記 2 手順で判定漏れしたトップメッセージを機械的に「不明」と判定
4) 各カテゴリの特徴語が含まれるトップメッセージと上記の判定結果の対応数を計量（**表 3**）
5) 対象とした共学中学校のデータ 1,705 件のうち，「男子」と「女子」で，各特徴語がそれぞれの性別と対応付いている割合を評価（**表 4**）

表 3 に示す分析結果より，一部の語句を除き，各カテゴリの特徴語は，共学中学校のトップメッセージを分析する際に有効である可能性が示唆された．そこで，分析対象の共学中学校のデータ 1,705 件を対象に，男子と女子をどの程度推定できるかを評価したところ，**表 4** に示す通り，推定精度は約 20%程度という結果となった．

また，「高校」「高校進学」や「大学」「大学受験」等，将来に対する意識差は，tf-idf 値の差からは読み取ることが難しいことが分かった．これは，既研究[10]でも見られた特徴である．あえていえば，男子中学生のカテゴリでのみ「大学院進学」が登場した点が，教育意識の差として抽出された．

表 4　対象としたデータ数

カテゴリ	機械的判定数	特徴語による推定
男子	181 件	22.7%
女子	310 件	23.2%
不明	1,214 件	-

（4）考察

分析結果により，受験用 SNS を対象とした場合，交換される話題の多くは共通する傾向がある一方で，一部の語句は特定の性別カテゴリに強く関連付けられる見込みがあることが分かった．このことから，記事内に含まれる話題をカテゴリ分類することで，記事の主題となる子供の性別を限定できる可能性があることが分かった．また，男子中学校と女子中学校の掲示板で交わされる話題の違いは，共学中学校の掲示板で話題にされる子供の性差の推定に活用できる可能性があることが分かった．

その一方で，各語句に依存した推定では，高い推定精度を見込めないことが分かった．なお，共学中学校のトップメッセージを対象とした分析結果を部分抜粋して目視確認したところ，**表 5** に示す通り，機械的な判定の影響を受けて，正しい分析ができていない項目が散見されたことから，追加の分析が必要である．

また，分析対象のデータを取得した掲示板は受験情報の交換を目的に利用されており，子供の教育に熱心な親の書き込みが多いと考えられることから，使用される名詞の頻度に基づく tf-idf 値のみでは，教育意識の差を明らかにするまでには至らなかった．

4. 今後の展開

本稿では，中学受験に関する受験用 SNS において，子の性別によって保護者に意識差があるのかを調査する目的で，受験情報サイト「インターエデュ・ドットコム」の投稿記事の内容を分析した．

分析の結果，受験対策用 SNS という分析対象の掲

示板の特徴もあり，子供の性差が保護者の教育意識に与える影響の抽出は難しいことが分かった．その一方で，特定のジャンルの名刺や複合名詞に関しては，性差が見られることが分かった．このことから，掲示板に投稿する記事内容より，記事の主題となる子供の性別を推定できる可能性があることが分かった．また，語句の共起関係を分析することで，より詳細な話題に絞り込むことができると考えられる．

今後の展開としては，3 種類が考えられる．

一つ目は，受験に特化しない SNS を対象に，本稿と同じ分析を適用する方法である．これにより，保護者の教育意識と子供の性別に関係があるかを分析することが考えられる．

二つ目は，男子中学校と女子中学校および共学中学校の各カテゴリ内で流通する話題の多様さやカテゴリ間の特徴量の差を強調する評価指標を用いて，再評価する方法である．例えば，既研究[10]では，性別と同時に偏差値の違いにも着目した分析を実施し，子供の性差に基づく投稿記事の違いを評価している．

三つ目は，学校のカテゴリに関係なく，男子と女子の各カテゴリを分類するための分類器，もしくは AI（Artificial Intelligence）を構築する方法である．語句の並びや共起関係を考慮することで，推定精度の向上が見込まれる．

以上の検討を通じて，SNS を活用した子供の性別による保護者の教育意識の違いの分析手法や経時的な変化の追跡手法の開発に取り組む予定である．

謝辞

本研究は，流通科学大学特別研究費の助成を受けて実施した．また，本研究を実施するにあたり，株式会社インターエデュ・ドットコムより掲示板データを提供頂いた．ここに記して感謝の意を表する．

参 考 文 献

[1] World Economic Forum, “The global gender gap report 2022”, (2022). <https://www.weforum.org/reports/global-gender-gap-report-2022/> (2023.6.10. 確認)

[2] 男女共同参画局, “G7 ジェンダー平等大臣共同声明（日光声明）”, (2023).<https://www.gender.go.jp/international/int_kaigi/int_g7g8/g7_202306.html> (2023.2.10. 確認)

[3] 男女共同参画局, “女性活躍・男女共同参画の重点方針 2023”, (2023).＜https://www.gender.go.jp/policy/sokushin/pdf/sokushin/jyuten2023_honbun.pdf> (2023.2.10. 確認)

[4] 三菱 UFJ リサーチ＆コンサルティング, “女性管理職の育成・登用に関する調査～男女の就業継続や昇進意向に関する意識・実態～”, (2020).＜https://www.murc.jp/wp-content/uploads/2020/05/seiken_200525.pdf＞(2023.2.10. 確認)

[5] 丸山亜希子, 上田真由美, “子どもの性別による親の教育意識に関する調査分析”, DEIM2023, 5a-6-4, (2023).

[6] ベネッセ教育総合研究所, “子どもの生活と学びに関する親子調査 Wave4”, (2018).

[7] Pollak, A. R. “Bargaining Power in Marriage: Earnings, Wage Rates and Household Production.” National Bureau of Economic Research Working Paper 11239, (2005).

[8] Aizer, A. “The Gender Wage Gap and Domestic Violence”, American Economic Review, Vol.100, No.4, (2010).

[9] 受験と教育の情報サイト- インターエデュ・ドットコム,<https://www.inter-edu.com/> (2023.6.10.確認)

[10] 丸山亜希子, 寺口敏生, 上田真由美, “受験向け SNS における男女別意識差分析”, 情報科学技術フォーラム講演論文集 (FIT), Vol.22, No.4, pp.299-302, (2023).

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 13:30 ~ 15:40 | T3-B オンライン（Zoom Events)

検索クエリ

座長:清水 敏之(九州大学)
コメンテータ:横田 治夫(城西大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T3-B-2-01] 系列変換モデルを利用したフィールド検索のためのクエリ書き換え

*仲地 優登[1]、加藤 誠[2]、百合草 陽介[3] (1. 筑波大学 大学院人間総合科学学術院 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系、3. 株式会社ミスミグループ本社)

14:00 ~ 14:25
[T3-B-2-02] クエリ自動補完のための文書コレクションからのクエリログ生成

*染谷 瑛進[1]、加藤 誠[2] (1. 筑波大学 知識情報・図書館学類 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系)

14:25 ~ 14:50
[T3-B-2-03] 音声対話型検索における階層型検索意図を用いたクエリ推薦の戦略

*平位 遼[1]、飛岡 憲[1]、山本 岳洋[1] (1. 兵庫県立大学)

14:50 ~ 15:15
[T3-B-2-04] A Token-Free Approach to Entity-Based Keyword Search in Classical Chinese

*蔣 中慶[1]、呉 増青[2]、肖 川[1,3]、佐々木 勇和[1]、鬼塚 真[1] (1. 大阪大学、2. 京都大学、3. 名古屋大学)

系列変換モデルを利用したフィールド検索のためのクエリ書き換え

仲地　優登† 　加藤　誠†† 　百合草陽介†††

† 筑波大学 大学院人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
††† 株式会社ミスミグループ本社　〒 102–8583 東京都千代田区九段南 1 丁目 6 番 5 号
E-mail: †nakachi@klis.tsukuba.ac.jp, ††mpkato@acm.org, †††yosuke.gjdt.yurikusa@misumi.co.jp

あらまし　本研究では，ユーザが検索システムを利用する際に，事前に検索対象となっているフィールドの情報を知る必要なく，特定のフィールドに対する検索を可能にするために，系列変換モデルによってキーワードクエリをフィールド検索用のクエリに書き換える手法を提案する．系列変換モデルの学習には，キーワードクエリとそれに対応するフィールド検索クエリのペアが必要であるが，人手による作業が必要となり学習のために十分な量のデータを用意できない場合がある．そこで本研究では，運用中の検索システムにて入手が容易なクリックスルーデータと検索対象文書群を用いて，キーワードクエリ-フィールド検索クエリのペアを自動的に生成する方法を提案する．提案手法では，大規模言語モデルによってキーワードクエリからフィールド検索クエリ候補を生成し，クリックスルーデータから推測される適合文書を利用することで生成されたフィールド検索クエリを評価して，適合文書を上位に順位付けられるようなフィールド検索クエリを正しい変換結果として採用する．実験では，学習データの作成と系列変換モデルの学習を行う．用いるデータとして，人工データとして Text-to-SQL タスクのデータセットを，実データとして EC サイトのデータを用いて，提案手法によるクエリ書き換えの精度を評価を行う．

キーワード　情報検索，E-Commerce，説明可能性，検索クエリ（意図理解）

1　はじめに

フィールド検索は，文書やデータベース内の複数のフィールドからなる，構造化された文書に対する検索のことである．たとえば，学術データベースにおいて「著者」フィールドを指定し，特定の著者による論文を検索する場合や,「タイトル」フィールドを指定して特定のキーワードが含まれる論文を絞り込むことができる．一般的な全文検索は検索対象中の文書のフィールド全てに対して検索が行われるが，フィールド検索においては，特定のフィールドに対する情報要求が含まれるクエリで，検索結果の精度が向上し，ユーザーは必要とする情報を効率的に見つけることが期待される．

しかし，一般的な全文検索と異なり，複数入力フィールドのインターフェースに多くユーザが慣れていないこと [13] や，利用するために対象となるフィールドの事前知識が必要である [1] ため，十分に活用されていない [6]．

そこで本研究では，系列変換モデルを利用したフィールド検索のためのクエリ書き換えを提案する．この手法は，ユーザが入力したキーワードクエリをフィールド検索のためのクエリ (フィールド検索クエリ) へ自動的に書き換える．これにより，ユーザが通常の検索システムのインターフェースのままキーワードクエリを入力し，事前に検索対象となっているフィールドの情報を知る必要なく，特定のフィールドに対する検索が可能になることを目指す．

系列変換モデルは，ユーザが入力したキーワードクエリから，各キーワードに対応するフィールドを割り振られたフィールド検索クエリを生成する．系列変換モデルがキーワードクエリからフィールド検索クエリを生成するためには，キーワードクエリに対応するフィールド検索クエリで構成される適切なデータセットを用いた学習が必要である．しかしながら，現状，キーワードクエリとフィールド検索クエリのペアを含むデータセットは存在しない．そのため，本研究では，大規模言語モデルと，運用中の検索システムから入手できるクリックスルーデータと検索対象の文書群のみを利用し，キーワードクエリとフィールド検索クエリのペアを自動的に生成する方法を提案する．提案手法では，大規模言語モデルを用いてキーワードクエリからフィールド検索クエリ候補を生成し，クリックスルーデータから推測される適合文書を利用して，生成されたフィールド検索クエリを評価する．この評価により，適合文書を検索結果の上位に順位付けられるフィールド検索クエリを正しいものとして採用する．この手法により作成したキーワードクエリ-フィールド検索クエリを用いて系列変換モデルを学習する．

実験では，人工データとして text-to-SQL タスクのデータセットである WikiSQL [17] と，実データとして株式会社ミスミ（株式会社ミスミグループ本社の子会社）が提供する MISUMI EC サイト [2]のデータを用いる．そして，提案手法により作成されたフィールド検索クエリの精度の評価を行う．また，作成したフィールド検索クエリを用いて，系列変換モデルを学習し，学習したモデルが出力したフィールド検索クエリを用いて精度

1：https://iastate.pressbooks.pub/lib160/chapter/field-searching/

2：https://jp.misumi-ec.com/

の評価を行う．

実験の結果，提案手法により，WikiSQL においては最大で 2 万件程度，MISUMI EC サイトのデータにおいては最大で 1 万件程度のキーワードクエリ-フィールド検索クエリのペアが獲得できた．概ね作成したフィールド検索クエリを用いた場合の精度が，フィールド検索を行わなかった場合と比較して高い精度を獲得した．系列変換モデルを学習させ，モデルが出力したフィールド検索クエリを用いて精度の評価を行った結果，フィールド検索に有効なフィールド検索クエリを作成することができず，フィールド検索を行わなかった場合と比較して，高い精度を得られなかった．このことから，作成した学習データでは，系列変換モデルを学習するのに不十分であったということが考えられる．

本研究の貢献は，以下のようになっている．

（1） 系列変換モデルを利用したフィールド検索のためのクエリ書き換え手法の提案

（2） 大規模言語モデルとクリックスルーデータを用いたキーワードクエリからフィールド検索クエリの学習データ作成手法の提案

（3） 作成したキーワードクエリ-フィールド検索クエリのペアで構成される学習データを用いた実験の実施とその結果の報告

本研究の構成は以下のようになっている． 2 節ではクエリ分類・クエリ拡張の関連研究について述べる． 3 節では問題設定を説明し，および．その主問題への適用方法と，提案手法について説明する．最後に， 5 節では今後の課題と共に本研究の結論を述べる．

2 関 連 研 究

本節では，本研究と関連するクエリ分類とクエリ拡張，知識グラフによるクエリ理解について説明する．

2.1 クエリ分類

検索クエリを分類するクエリ分類は，一般的なウェブ検索の効率と有効性の双方に改善をもたらす技術として広く認識されている [4]．Beitzel らによる，AOL から収集されたトピック分類に基づく研究 [1] [2] では，大規模な人間によるラベル付けされたクエリデータセットに対するマッチング，大規模なラベル付けされたデータセットからの教師あり学習法による分類器の生成，大規模なラベルなしクエリログからの選択的嗜好ルール [9] の設計という，3 つの方法でクエリトピック分類を試みた．彼らの実験結果は，各アプローチには長所と短所があることを示した．たとえば，正確なマッチングアプローチは高頻度クエリにおいて高い精度を持つが，再現率は低い．教師あり分類器の閾値は慎重に調整する必要があり，訓練データに存在しないクエリにはうまく機能しない可能性がある．そのため，彼らは 3 つのアプローチを組み合わせることを提案し，実験結果ではより良い結果が得られたことを示した．

また，Broder らは，クエリ分類システムを構築する方法論を提案した [3]．彼らの目標とする分類体系は，6000 ノードのトピックからなる商業的分類である．彼らは，クエリログでの観測が十分でない稀なクエリに対して，検索エンジンを包括的なリポジトリとして利用し，クエリの意味をよりよく理解するための知識を構築する方法を採用した．具体的には，擬似関連フィードバックを使用し，上位の検索結果を利用して稀なクエリに関する情報を収集した．さらに，ノイズの多い結果をフィルタリングするために投票スキームを利用した．実験結果は，提案された方法がクエリ分類の結果を改善したことを示した．

これらの研究と本研究との最大の違いは，本研究がキーワードクエリからフィールド検索クエリへの変換に焦点を当てている点である．これまでの研究では，クエリのトピック分類や稀なクエリの理解に注力していたが，本研究では，クエリの意図を解釈し，特定のフィールドに対する検索に対して，適切な検索結果を提供することを目指している．

2.2 クエリ拡張

クエリ拡張は，検索クエリの解釈と検索結果の改善に対する技術であり，擬似適合性フィードバックや外部知識ソース（例えば WordNet）に基づいてクエリを書き換える手法である [12] [8]．特に，語彙一致ベースの検索手法である，疎検索において，クエリと文書間の語彙的なギャップを埋めるのに役立つとされている．

Wang らの研究 [15] では，疎検索モデルと密検索モデル [7] の両方を改善するためのクエリ拡張アプローチを提案している．提案された方法は，まず大規模言語モデルを用いてクエリから擬似ドキュメントを生成し，次に生成された擬似文書でクエリを拡張する．これにより，クエリの曖昧さを解消し，検索モデルがより良い検索結果を取得するのに役立つ情報をクエリ中に含めることを目指している．実験結果は，アドホック検索タスクのデータセットで BM25 のパフォーマンスを 3 %から 15 % 向上させることを示している．さらに，この方法は密検索モデルにも，ドメイン内およびドメイン外の結果の両方で検索性能の向上をもたらすことが報告された．

これらの研究と本研究との主な違いは，クエリ拡張技術が主に検索結果の精度と関連性を高めるために使用されるのに対し，本研究はキーワードクエリからフィールド検索クエリへの変換に焦点を当てている点である．クエリ拡張は一般的な検索クエリの意味を拡張し，関連する文書を見つけるのに役立つが，本研究では，クエリを特定の検索フィールドに割り当て，よりユーザの情報要求に対して適切な検索結果を提供することに焦点を当てている．

2.3 知識グラフによるクエリ理解

Manshadi ら [10] は，自由記述形式のキーワードクエリ中のキーワードを，特定のオントロジーまたは知識ベースで定義された概念にマッピングする問題に焦点を当てている．クエリは通常，言語的の文法を守らず，統計的推論のための情報も含まない．そこで，クエリ中のエンティティに関連する動詞，形容詞，属性などのコンテキストを用いて，大規模なウェブコー

パスから用語間のさまざまな関係を掘り下げ，それらを確率的知識ベースを使用して関連する概念にマッピングを行う手法を提案している．そして，与えられたクエリにおける用語をランダムウォークに基づく反復アルゴリズムを使用してマッピングを行う．この手法では，オントロジーや知識ベースを基にキーワードクエリ中のキーワードが特定のフィールドに対する情報要求であることを単語の厳密一致な手法を用いてマッピングするが，本研究では，系列変換モデルを用いることで，キーワードクエリ中にタイプミスなどがあった場合に，系列変換モデルが自動的に修正などを行うことも可能である．そのため，ユーザが入力するキーワードクエリに対して柔軟に対応することが期待できる点が異なる．

3 提案手法

本節では，まず，問題設定について説明する．次に，問題設定に対して用いる手法について説明する．最後に，提案手法について説明する．

3.1 問題設定

本研究では，キーワードクエリ q が与えられた時，フィールド検索用のクエリ q_f に変換する問題を対象とする．例えば，「金属プレート A2017 3mm」というキーワードクエリが与えられた時，A2017 が材質のジュラルミンを表し，3mm が板厚を表していることを認識し，「“金属プレート・ブロック. 材質”: ”A2017”, ”金属プレート・ブロック. 板厚　T(mm)”: ”3”」というように，各キーワードに対応するフィールドとその値を割り振ることが求められる．フィールド検索クエリは文書 $d \in D$(D は検索対象の文書群) 中のフィールド名 $f \in d$ とそれに対応するフィールドの値 v のペアで構成される．ただし，このフィールドの値は任意のテキストである．この問題を定式化したのが数式 (1) である．

$$q_{\mathrm{f}} = (\{f_1, v_1\}, \{f_2, v_2\}, \ldots) = \phi(q) \qquad (1)$$

数式 (1) 中の ϕ がキーワードクエリ q をフィールド検索クエリ q_{f} に変換する関数である．

3.2 系列変換モデルを利用したクエリ変換手法

キーワードクエリをフィールド検索クエリに変換する際に，キーワードクエリに基づいてフィールドを適切に割り当てることが難しい場合がある．この難しさの一つとして，必ずしもキーワードクエリ中に含まれる文字列が，検索対象となるフィールドの値に対して適切なものでないことが考えられる．例えば，文書の電線. 材質フィールドに “PVC(ポリ塩化ビニルを表す)” という値が格納されている場合,「耐熱ビニル電線」のようなキーワードクエリでは，語彙的に一致しないため，この文書は検索結果中の上位には現れないことが考えられる．ここで，“ビニル” を “PVC” に変換すると，語彙が一致することにより，この文書は上位に位置付けられる．

以上のことを考慮して，本研究では，数式 (1) 中の ϕ に T5 [11] のような Transformer ベース [14] の系列変換モデルを用いることで，キーワードクエリを適切に書き換え，フィールド検索クエリへの変換を行う．系列変換モデルが上手くキーワードクエリからフィールド検索クエリへの変換を行うために，キーワードクエリとそれに対応するフィールド検索クエリのペアからなる学習データを用いて系列変換モデルの学習を行う. 系列変換モデルの学習は以下の損失を最小化するように行われる．

$$\mathcal{L}(\theta) = -\frac{1}{|Q|}\sum_{k=1}^{|Q|}\sum_{i=1}^{N}\log P(y_i|y_{<i}, q_k; \theta) \qquad (2)$$

ここで，Q はキーワードクエリ集合，q_k はキーワードクエリ集合中の k 番目のキーワードクエリ，N は出力シーケンスの長さ，y_i は出力される i 番目のトークン，$y_{<i}$ は i 番目のトークンより前のトークンを表す数式 (2) で表される損失を用いて，キーワードクエリを入力とし，それに対応するフィールド検索クエリが出力されるように系列変換モデルの学習が行われる．

3.3 提案手法: 学習データ作成手法

前項 3.2 では，キーワードクエリ-フィールド検索クエリのペアを用いて系列変換モデルの学習を行うことを述べた．しかしながら，現状キーワードクエリに対応するフィールド検索クエリで構成されるデータセットは存在しない．そこで，本研究では，運用中の検索システムから用意に入手可能なクリックスルーデータと検索対象の文書群の 2 つから，キーワードクエリ-フィールド検索クエリのペアを作成する．このペアの作成には，キーワードクエリの文字列に対して部分的に一致しているフィールドの値を抽出し，フィールド検索クエリを作成する手法と大規模言語モデルを用いてキーワードクエリからフィールド検索クエリ候補を作成する手法の 2 つの手法を用いる．そして，クリックスルーデータ C から推測される適合文書を利用す

入力例

```
以下にキーワードクエリとそれに対応づけられるフィールド検索クエリのペアの例を示します。
その後にキーワードクエリを与えるので、これに対応するフィールド検索クエリを例のように作成してください。
また、与えたキーワードクエリに対して，検索結果中にクリックされた文書の情報を追加するので、
これも参考にして下さい。
ただし、私が入力したテキストにそのまま続くように、フィールド検索クエリを出力してください。
### クエリ-フィールド検索クエリのペアの例 ###
キーワードクエリ1: トラスねじ　ステンレス
フィールド検索クエリ1: {"タッピングねじ・タップタイト・ハイテクねじ.材質": "ステンレス"}
キーワードクエリ2: エルボ SUS 継手
フィールド検索クエリ2: {"ねじ込み継手・ねじ込み式フランジ.接続形状": "エルボ", "ねじ込み継手・ねじ込み式フランジ.種類": "継手", "ねじ込み継手・ねじ込み式フランジ.材質": "ステンレス"}
...
### クリックされた文書のフィールド情報 ###
{"金属プレート・ブロック.板厚　T(mm)": "20", "金属プレート・ブロック.材質": " A2017P", "金属プレート・ブロック.仕上げ方法": "上下面サーフェイス研磨 (4F2G)", "金属プレート・ブロック.コーナーR加工選択": "無し", "金属プレート・ブロック.抜き穴加工選択": "無し", "金属プレート・ブロック.抜き穴加工種選択": "無し", ...}
### キーワードクエリ ###
クエリ: 金属プレート A2017 3mm
```

出力例

```
{ "金属プレート・ブロック. 材質": "A2017", "金属プレート・ブロック.板厚　T(mm)": "3"}
```

図 1　大規模言語モデルの入出力例

図 2　キーワードクエリ「金属プレート A2017 3mm」に対する検索結果のうち，クリックされた文書例のスクリーンショット

ることで作成されたフィールド検索クエリ候補 q_{fc} を評価する．評価では，フィールド検索クエリ候補 q_{fc} を用いて，フィールド検索を実際に行い，適合文書を上位 k 件に順位付けられるようなものを正しい変換結果として採用する．本項では，これらのフィールド検索クエリ候補の生成方法とフィールド検索クエリ候補の評価方法について順番に説明する．

3.3.1 フィールド検索クエリ候補の生成

まず，キーワードクエリに部分的に一致するフィールドの値から，フィールド名とその値を抽出することで，フィールド検索クエリを作成する手法について説明する．この手法では，キーワードクエリが部分的に一致する場合に，文書内の特定のフィールドとその値を抽出し，それらを利用してフィールド検索クエリを構築することにより，検索結果の精度を向上させることを狙う．キーワードクエリ q に部分的に一致するフィールド値を抽出し, それを基にフィールド検索クエリの候補 q_{fc} を作成する．このプロセスは次のように数式で表される．

$$q_{fc} = \mathrm{Extract}(q, v^{+}) \tag{3}$$

ここで，Extract$(\cdot)$ は部分一致するフィールド値を抽出する関数であり，v^{+} はキーワードクエリ q の検索結果中にクリックされた文書のフィールドの値の集合である．

次に，大規模言語モデルを用いてキーワードクエリからフィールド検索クエリを生成する方法について説明する．この手法では，キーワードクエリとその検索結果中のクリックされた文書のフィールド名とその値を大規模言語モデルに入力する．そして，大規模言語モデルが，キーワードクエリを検索対象のフィールドに基づいた適切な形に書き換え，検索結果の精度を向上させることを狙う．

このフィールド検索クエリを生成する方法を定式化したものを以下に示す．

$$q_{fc} = \psi(p \oplus q; \theta_{\mathrm{LM}}) \tag{4}$$

ここで，ψ はキーワードクエリからフィールド検索クエリ候補への変換を行う関数であり，θ_{LM} は言語モデルのパラメータを表す．また，p はプロンプトを表し，$\oplus$ は文字列の結合を表す．p については，図 1 に示している．p に含める情報として，キーワードクエリとフィールド検索クエリのペアの例とクリックスルーデータ C から，文書のフィールド名とその値を含める．具体的には，クリックスルーデータ C 中のキーワードクエリ q に対応する検索結果のクリックされた文書 d^{+} のフィールドの名 $f^{+} \in d^{+}$ とその値 v^{+} を p に含める．大規模言語モデルへの入力例を 1 に示す．これにより，クリックスルーデータから追加のフィールド名とその値を基に生成関数 ψ がフィールド検索クエリを生成できるようにする．

3.3.2 フィールド検索クエリ候補の評価

フィールド検索クエリ候補の評価は，クリックスルーデータと検索対象の文書群 D を用いて行う．これにより，作成したフィールド検索クエリの候補が，フィールド検索において検索結果の精度を向上させるのに有効か評価し，学習データの質を確保する．具体的には，キーワードクエリ q に対応する検索結果中のクリックされた文書を擬似的に適合文書 $d^{+} \in D$ とみなし，キーワードクエリ q に対応するフィールド検索クエリ候補 q_{fc} を用いてフィールド検索を行う．フィールド検索を行い，上位 k 件に適合文書 d^{+} が存在する場合，フィールド検索クエリ候補 q_{fc} をキーワードクエリ q に対応する正しいフィールド検索クエリとして採用する．このように，フィールド検索クエリ候補から適切なフィールド検索クエリを絞り込む．

以上のように作成されたフィールド検索クエリのうち，キーワードクエリに部分的に一致するフィールドの値を基に作成されたフィールド検索クエリを**部分一致フィールド検索クエリ**と呼び，大規模言語モデルにより作成されたフィールド検索クエリを**生成フィールド検索クエリ**と呼ぶ．

表 1　使用したデータセットの統計情報

	WikiSQL	**MISUMI**
クエリ数	34,059	22,272
検索対象の文書数	444,609	450,253
フィールド数	162,642	12,580

4 実　　験

本節では，まず，提案した手法により作成されたキーワードクエリ-フィールド検索クエリペアの数と，ペアの事例分析を行い，その結果を報告する次に，系列変換モデルの学習を行い，その評価について報告する．

4.1 学習データの作成と評価

本項では提案した手法により作成した系列変換モデルを学習するための，キーワードクエリ-フィールド検索クエリのペアによる学習データの作成とその評価について説明する．

4.1.1 使用したデータセットの概要と統計

使用したデータセットは人工データとして text-to-SQL タスクのデータセットである WikiSQL [17]，実データとして MISUMI EC サイトのデータを用いた．また，検索システムは，Elasticsearch[3]を用いて実装した．表 1 に使用したデータセットの統計情報を示す．WikiSQL は，テキストからそれに対応する SQL を生成するタスクのためのデータセットであり，SQL の実行対象が Wikipedia の表である．本研究においては，SQL の実行対象である表データのヘッダーを除いた各行の要素を結合し，文書とみなし，検索対象の文書群を作成した．さらに，キーワードクエリとして，テキスト-SQL ペアのテキストを採用し，フィールド検索クエリとして，SQL 中の WHERE 句により抽出される列の情報とその値をフィールド名とその値とみなした．(このように作成されたフィールド検索クエリを本研究中では，**SQL フィールド検索クエリ**と呼ぶ．) また，SQL を実行し，出力結果の行全てを適合文書とする場合 (「SQL による評価」と呼ぶ) と，SQL フィールド検索クエリを用いて検索を

3：https://www.elastic.co/jp/elasticsearch

行い，その検索結果に対して擬似的にクリックされた文書をモデリングする Cascade Model [5] を使用し，クリックされた文書を適合文書とする場合（「クリックによる評価」と呼ぶ）を用意した．(モデリングに使用したパラメータは，クリックする確率 $P_{click} = 0.7$, 検索結果の閲覧から離脱する確率 $P_{stop} = 0.3$ と設定した．) これにより擬似的にキーワードクエリ-フィールド検索クエリのペアとこれに対応する擬似適合文書を作成した．WikiSQL のフィールドに関しては，表データ中の各列を別々のフィールドとして扱う．

部分一致フィールド検索クエリの作成方法について，WikiSQL データセットにおいては，SQL による評価から部分一致フィールド検索クエリを作成した場合と，クリックによる評価から部分一致フィールド検索クエリを作成した場合の 2 パターン用意した．また，生成フィールド検索クエリの作成に関して，フィールド名のみを与えた場合 (生成フィールド検索クエリ (w/o フィールド値)) とフィールド名とその値の両方を与えた場合 (生成フィールド検索クエリ (w/ フィールド値)) の 2 パターン用意して，提案手法の有効性を検証した．生成フィールド検索クエリを作成する際に用いた大規模言語モデルは OpenAI 社の GPT-3.5-Turbo[4]を Microsoft 社のクラウドサービスである Azure 経由 [5]で利用した．

4.1.2 学習データの定量的評価

表 2 WikiSQL から作成したキーワードクエリ-フィールド検索クエリ (候補) の数

	絞り込みなし	絞り込みあり
SQL フィールド検索クエリ	34,059	30,856
部分一致フィールド検索クエリ (SQL による評価)	34,053	30,605
部分一致フィールド検索クエリ (クリックによる評価)	25,241	22,006
生成フィールド検索クエリ (w/o フィールド値)	30,840	23,141
生成フィールド検索クエリ (w/ フィールド値)	30,829	24,066

表 3 MISUMI EC サイトのデータから作成したキーワードクエリ-フィールド検索クエリ (候補) の数

	絞り込みなし	絞り込みあり
部分一致フィールド検索クエリ	3,902	2,267
生成フィールド検索クエリ (w/o フィールド値)	20,828	8,549
生成フィールド検索クエリ (w/ フィールド値)	20,795	9,778

作成したキーワードクエリ-フィールド検索クエリの統計情報を表 2 と表 3 に示す．

作成した学習データのうち，絞り込みありを用いて検索性能を nDCG@k($k = 10, 30, 100$) 評価した．その結果を表 4 と表 5 に示す．

表 4 から，WikiSQL データセットに関しては，2 つの評価方法（SQL による評価とクリックによる評価）により評価を行なっているが，そのどちらにおいても，フィールド検索を行わず，キーワードクエリのみを用いて検索を行なった場合よりも，フィールド検索クエリを用いて検索を行なった場合の方が性能が一貫して良い結果となった．このことからフィールド検索の有効性が示されたと考えられる．また，SQL による評価において，どの評価指標においても，最も高いスコアを示したのは SQL による評価から作成した，部分一致フィールド検索クエリであり，同様にクリックによる評価において，どの評価指標においても，最も高いスコアを示したのはクリックによる評価から作成した，部分一致フィールド検索クエリであった．このことから，WikiSQL データセットにおいては，文字列の表層的な一致によるフィールド検索クエリの作成が最も良い方法であることがわかる．生成フィールド検索クエリにおいては，フィールド名だけではなく，その値をプロンプトに含めることでより良いフィールド検索クエリが作成できることがわかった．

表 5 から，MISUMI EC サイトのデータに関しては，フィールド名とその値を与えて大規模言語モデルにより作成させた生成フィールド検索クエリは，その他のフィールド検索クエリに比べて一貫して高いスコアとなっている．これは，フィールド検索において，フィールド値の明示的な指定が重要であることを示している．一方で，フィールド名のみを与えて作成させた生成フィールド検索クエリは，フィールド検索なしの場合と比較してスコアが低くなっている．このことから，フィールド名のみを与えても，キーワードクエリから適切なフィールド値を推定できずに，性能が下がっていることが考えられる．部分一致フィールド検索クエリにおいては，作成したフィールド検索クエリの数が 2,267 個と全体の 1 割にも満たないながら，検索性能を向上させている．このことから，キーワードクエリとクリックされたフィールド値の部分一致によるフィールド検索クエリの作成はユーザの特定のフィールドに対する情報要求を満たすのに十分活用できることが考えられる．

全体的に，フィールド検索クエリを用いたフィールド検索では，フィールド検索なしの条件に比べて改善されたスコアを示している．これは，適切に構造化されたフィールド検索クエリが検索性能向上に寄与することを示している．

4.1.3 作成した学習データの事例分析

WikiSQL データセットセットから作成した学習データを確認し，その特徴的な失敗事例について議論する．特徴的な失敗事例について表 6 に示す．

表中の右側に示しているフィールド検索クエリは，“{ フィールド名 }”: “{ フィールドの値 }”という形式で記述されている．フィールド名は，“{ テーブル名 }.{ 列名 }” という形式である．例えば，“2-1224468-1.Tracks”: “2”というフィールド検索クエリにおいて，フィールド名は，“2-1224468-1.Track”，テーブル名が “2-1224468-1”，列名が “Track”，フィールドの値が “2” である．

表 6 中の 1 は，部分一致フィールド検索クエリが適切なフィールド名を指定できていない事例である．キーワードクエリ中の”2002”という年号を表しているものに対してフィールド値として”2”が抽出されている．この事例から，キーワードクエリが意図していないフィールド名とのマッチングが行われている場合があることがわかる．表 6 中の 2 は，生成フィールド

4：https://platform.openai.com/docs/models/gpt-3-5-turbo
5：https://learn.microsoft.com/ja-jp/azure/ai-services/openai/concepts/models#gpt-35

表 4　WikiSQL から作成した学習データによる検索性能の評価

	nDCG@10 (SQL)	nDCG@30 (SQL)	nDCG@100 (SQL)	nDCG@10 (click)	nDCG@30 (click)	nDCG@100 (click)
フィールド検索なし	0.2539	0.2742	0.2925	0.2708	0.3059	0.3274
SQL フィールド検索クエリ	0.5476	0.569	0.5828	0.4987	0.5449	0.5494
部分一致フィールド検索クエリ (SQL による評価)	0.5727	0.5924	0.6053	0.4945	0.54	0.5459
部分一致フィールド検索クエリ (クリックによる評価)	0.439	0.4598	0.4732	0.551	0.5802	0.5863
生成フィールド検索クエリ (w/o フィールド値)	0.2786	0.2984	0.3158	0.291	0.3266	0.3472
生成フィールド検索クエリ (w/ フィールド値)	0.2896	0.3094	0.3267	0.3369	0.3706	0.3885

表 5　MISUMI EC サイトのデータから作成した学習データによる検索性能の評価

	nDCG@10	nDCG@30	nDCG@100
フィールド検索クエリなし	0.1416	0.1675	0.1886
部分一致フィールド検索クエリ	0.1458	0.1721	0.1935
生成フィールド検索クエリ (w/o フィールド値)	0.1178	0.1418	0.1628
生成フィールド検索クエリ (w/ フィールド値)	0.1578	0.1832	0.2037

表 6　作成したフィールド検索クエリのうち，適切でない事例

	キーワードクエリ	フィールド検索クエリ
1	How many people named gloria moon were on the 2002 and 2012 commissions?	“2-1224468-1.Tracks”: “2”
2	When was the date of an event at Tokyo venue?	“1-1629365-1.Venue”: “Tokyo”
3	Makio Inoue dubbed on which character?	“1-2160215-1.Character”: “Makio Inoue”
4	Who is the champion of the title previously held by Lash?	“1-23981771-1.Song title”: “champion”, “1-23981771-1.Artist”: “champion”, ···

検索クエリの作成時に，大規模言語モデル存在しないフィールド名を生成している事例である．このフィールド検索クエリ中の”1-1629365-1.Venue”がフィールド名であるが，このようなフィールド名は存在しない．このように，存在しないフィールド名を生成している事例が散見された．表 6 中の 3 は，フィールド名のみを与えた場合の生成フィールド検索クエリの作成時に，適切なフィールド名は生成されているが，その値が適切でない事例である．キーワードクエリ中の”Makio Inoue”は本来は”1-2160215-1.Original Japanese”というフィールドに存在する．しかしながら，フィールド名のみを与えた場合では，Character と認識し，そのフィールドに割り振っていた．このように，フィールド名のみを与えると，適切なフィールド検索クエリが作成されない場合がある．表 6 中の 4 は，与えられたフィールド名全てに特定の単語を指定している事例である．このような事例は生成フィールド検索クエリを作成する際に，フィールド名のみを与えた場合と，フィールド名+値を与えた場合の両方で確認された．このことから，生成フィールド検索クエリを作成する際に，何かしらの制約を加えて，このような事例が作成されにくくなるような仕組みが必要であることが示唆される．

以上のような失敗事例を確認した．

4.2　系列変換モデルの学習と作成されたフィールド検索クエリによる検索性能評価

本項では，4.1 項で示した学習データを用いて，系列変換モデルを学習し，得られたキーワードクエリ-フィールド検索クエリの数と，フィールド検索クエリを用いて検索を行なった検索性能の評価について説明する．

4.2.1　系列変換モデルの学習

本研究では，系列変換モデルとして多言語モデルである，mT5 [16] の base モデル [6]を用いて，キーワードクエリが入力された際に，それに対応するフィールド検索クエリが生成されるように学習を行う．学習データとして用いるのは，元のキーワードクエリのみのデータを train:validation:test=70:15:15 に分割し，その中 train 用に割り振られているキーワードクエリ-フィールド検索クエリのペアを用いる．学習に用いた学習データ数を表 7 に示す．

4.2.2　作成されたフィールド検索クエリによる検索性能の評価

学習した系列変換モデルを用いて，前述した test 用のデータを用いてフィールド検索クエリの作成を行なった．そして，作成されたフィールド検索クエリを用いて検索を行なった．その結果を表 8 に示す．

この結果全体を通して，フィールド検索なしの場合と比較して，系列変換モデルが生成したフィールド検索クエリは検索性能を下げることがわかった．また，系列変換モデルが作成したフィールド検索クエリを確認した．その結果，各キーワードクエリに対して，学習データ中に比較的多く出現する，フィールド名と値を割り振っていたため，適切なフィールドに対する検索が行えず，検索性能が下がっていることがわかった．ただし，MISUMI データの部分一致フィールド検索クエリについては，学習データ数が少なく，形式的に有効なフィールド検索クエリが 1 件も生成されず，実質的にフィールド検索なしと全く同じスコアになっている．これら結果から，学習データとして用いるキーワードクエリ-フィールド検索クエリの数が不足しているということが考えられる．

6：https://huggingface.co/google/mt5-base

表 7　系列変換モデルの学習データ数

	WikiSQL		MISUMI	
	絞り込みなし	絞り込みあり	絞り込みなし	絞り込みあり
SQL フィールド検索クエリ	23,841	21,603	-	-
部分一致フィールド検索クエリ (SQL による評価)	23,836	21,418	-	-
部分一致フィールド検索クエリ (クリックによる評価)	17,673	15,411	2,746	1,590
生成フィールド検索クエリ (w/o フィールド値)	21,591	16,205	14,585	5,934
生成フィールド検索クエリ (w/ フィールド値)	21,587	16,854	14,553	6,850

表 8　系列変換モデルが作成したフィールド検索クエリでの検索性能の評価 (nDCG@30)

	WikiSQL		MISUMI	
フィールド検索なし	0.3001		0.1737	
	絞り込みなし	絞り込みあり	絞り込みなし	絞り込みあり
SQL フィールド検索クエリ	0.2983	0.2983	-	-
部分一致フィールド検索クエリ (SQL による評価)	0.2988	0.2983	-	-
部分一致フィールド検索クエリ (クリックによる評価)	0.2983	0.2983	0.1737	0.1737
生成フィールド検索クエリ (w/o フィールド値)	0.2752	0.2745	0.1594	0.1662
生成フィールド検索クエリ (w/ フィールド値)	0.271	0.2726	0.1555	0.1651

5 ま と め

本研究では，ユーザが検索システムを利用する際に，事前に検索対象となっているフィールドの情報を知る必要なく，特定のフィールドに対する検索を可能にするために，系列変換モデルによってキーワードクエリをフィールド検索用のクエリに書き換える手法を提案した．この手法には，系列変換モデルを学習させるための学習データとして，キーワードクエリとそれに対応するフィールド検索クエリのペアが必要であるが，これまでの研究では十分な量のデータが用意されていなかった．そこで，本研究では，運用中の検索システムから容易に入手可能なクリックスルーデータを用いて，キーワードクエリとフィールド検索クエリのペアを自動的に生成する新たな方法を提案した．

提案手法は，大規模言語モデルを活用してキーワードクエリからフィールド検索クエリ候補を生成し，クリックスルーデータから推測される適合文書を利用して，生成されたフィールド検索クエリを評価し，適合文書を上位に順位付けられるフィールド検索クエリを正しい変換結果として採用するプロセスを含む．本手法により，特定のフィールドに対する，ユーザーが意図した情報をより正確に提供する検索結果が得られることが期待される．

実験では，提案手法により数万件程度の学習データを獲得した．これらのキーワードクエリ-フィールド検索クエリのペアを用いて系列変換モデルを学習させた．テスト用のキーワードクエリを学習した系列変換モデルに入力し，得られたフィールド検索クエリを用いて検索性能の評価を行なった．評価の結果，系列変換モデルが作成したフィールド検索クエリは，フィールド検索を行わない場合と比較して低い検索性能であることがわかった．このことから，本研究中で獲得した学習データでは，系列変換モデルが有用なフィールド検索クエリを作成できるように学習できていないことが考えられる．

今後の研究では，更なる学習データの作成方法とその評価方法について考案し，その有効性を行うことが目標である．

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

文　献

[1] Steven M Beitzel, Eric C Jensen, Ophir Frieder, David Grossman, David D Lewis, Abdur Chowdhury, and Aleksandr Kolcz. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 581–582, 2005.

[2] Steven M Beitzel, Eric C Jensen, Ophir Frieder, David D Lewis, Abdur Chowdhury, and Aleksander Kolcz. Improving automatic query classification via semi-supervised learning. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[3] Andrei Z Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 231–238, 2007.

[4] Yi Chang and Hongbo Deng. *Query understanding for search engines*. Springer, 2020.

[5] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[6] Bernard J Jansen and Udo Pooch. A review of web searching studies and a framework for future research. *journal of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[7] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question

answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[8] Victor Lavrenko and W Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[9] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[10] Mehdi Manshadi and Xiao Li. Semantic tagging of web search queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 861–869, 2009.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[12] J. J. Rocchio. Relevance feedback in information retrieval. *The SMART Retrieval System Experiments in Automatic Document Processing*, 1971.

[13] Kien Tjin-Kam-Jet. *Distributed Deep Web Search*. PhD thesis, Centre for Telematics and Information Technology, 2013.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore, December 2023. Association for Computational Linguistics.

[16] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.

[17] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103, 2017.

クエリ自動補完のための文書コレクションからのクエリログ生成

染谷　瑛進†　　加藤　　誠††

† 筑波大学 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s2011111@u.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　情報検索システムにおけるクエリ自動補完において，現在は，過去に蓄積された膨大なクエリログを必要とする．本論文では，HP のサイト内検索システムなどの，大量のクエリログが入手しづらい状況下において，検索対象とする文書集合と目標とするクエリ集合のトピック分布から，クエリ集合を生成するアプローチについて述べる．クエリの生成には，文書から様々な条件のもと単語を抽出してクエリを生成する手法と，文書からクエリを生成する言語モデルである docTTTTTquery の 2 つの手法を使用した．また，入力元の文書集合のトピック分布を調整した上でクエリを生成することにより，クエリ集合の生成性能が変化するかを調べた．実験により，生成するクエリの単語数や入力する文書集合のトピック分布を調整することは，クエリ集合の生成性能の向上に有効であることが示された．
キーワード　クエリ補完，情報検索，検索モデル

1　はじめに

一般的な情報検索システムには，クエリ自動補完 (Query Auto Completuon，以下 QAC) という機能が採用されている．QAC とは，ユーザが入力途中の不完全なクエリから，最終的に入力しようとしている完全なクエリを予測し，その候補を提示するタスクのことを言う．QAC には，キー入力に要する時間の短縮や誤字入力を防止する効果があり，現在の検索システムにおいて．欠かすことができない機能である [10, 12, 23].

QAC のアプローチには様々なものがあるが，その多くは，蓄積された過去のクエリログを用いることで行われる [1, 6, 21]. しかし，企業のホームページのサイト内検索や E メールのボックス内検索のように，十分にクエリログを入手できない場合においては，このような手法を使用することができない．しかし，そのような小規模の検索システムであっても，どのようなトピックのクエリが使用されやすいかについては，ある程度予測できる場合も多い．例えば，大学のホームページ内の検索システムを考えた場合,「入試」や「受験」に関連するクエリが入力されやすいであろうという予測はすることができる．したがって，小規模の検索システムにおいては，クエリログが得られたと仮定した場合における，そのクエリログ中のクエリのトピック分布は得ることができ，利用できると考えられる．

そこで，本論文では，検索サービスにおいて，ユーザが実際に使用しうるクエリの集合を生成する手法の提案を行う．具体的には，検索対象となる文書集合を入力とし，それを何らかのクエリ生成システムに入れることによって，クエリの集合を生成する．ただし，先述の仮定を踏まえ，本論文では，生成目標とするクエリ集合のトピック分布は入手可能であるという前提のもと，クエリログを代替するクエリ集合の生成を行う．

クエリの生成は，文書からの単語抽出と，文書からクエリを生成する手法の 1 つである docTTTTTquery [7] の 2 つを用いて行う．docTTTTTquery は，既存の文書とクエリのペアから出力パターンを学習することで，他の文書からもクエリを生成することができる言語モデルである．高い生成性能と速い生成速度を兼ね備えていることが報告されている [7] ことより，本研究ではこれを利用する．単語抽出による手法では，文書から様々な抽出条件のもと単語を抽出することで，クエリを生成する．docTTTTTquery を用いる手法では，まず，入手可能な他のクエリ集合と，それとペアになる文書集合を用いて，docTTTTTquery においてクエリを生成する事前学習モデルとして使用される T5 言語モデル [19] に対し，文書からクエリを生成するパターンを学習させる．次に，検索サービス内に含まれる文書集合から文書を 1 つずつ取り出し，それを学習済みの T5 言語モデルに入力することで，クエリ集合を生成する．また，2 つの手法において，入力する文書集合のトピック分布を調整することにより，クエリの生成性能が変化するかを調べる．

実験では，2018 年から 2020 年にかけて収集された Web ページデータセットである MS MARCO [17] 中の文書から単語抽出と docTTTTTquery によりクエリを生成し，MS MARCO と同時期に収集された Web ページのクエリログデータセットである ORCAS クリックスルーログ [8] とクエリ集合中の単語分布を比較した．実験により，抽出語数を 4 語とした上で，各文書のタイトルからのランダム単語抽出によりクエリを生成した場合，他の抽出条件や docTTTTTquery を用いてクエリを生成した場合と比較して，クエリの生成性能が高くなることが示された．また，入力する文書集合のトピック分布を調整した上でクエリを生成することにより，そうでない場合と比較して，クエリの生成性能が向上することも示された．以上の結果から，検索対象となる文書集合からのクエリログの生成において，生成するクエリの長さや入力する文書集合のトピック分布の調整は有効であると考えられる．

本論文の貢献は以下の通りである：

（1） QAC を実行したい検索サービス内に含まれる文書集合と，

生成目標とするクエリ集合のトピック分布を使用し，各文書からの単語抽出，または，言語モデルへの入力により，クエリ補完候補を提示するのに使用する擬似クエリログを生成する手法を提案した．

（2） 1 つのデータセットに対する実験から，目標とするクエリ集合のトピック分布を利用して，入力する文書集合のトピック分布を調整することにより，より実際のクエリログに近いクエリ集合を生成できることを示した．また，生成するクエリの単語数を調整することは，実際のクエリログに近いクエリ集合を生成するのに有効であることを示した．

本論文の構成は以下の通りである．2 節では QAC に関する既存研究について述べる．3 節では問題設定を説明し，クエリを生成する手法と課題点，その課題点を踏まえた提案手法について述べる．4 節では，データセット，実験の設定，結果を示す．最後に，5 節では今後の課題と共に本論文の結論を述べる．

2 関連研究

本節では，QAC に関する既存研究について述べる．

2.1 クエリログを用いた QAC の手法

QAC の既存研究の多くは，過去に蓄積された膨大なクエリログを用いるものである．このような手法では，検索ボックスに入力された文字列と接頭辞が一致するクエリ候補を頻度に基づいてランキングし，出力する．この際，より適切なクエリ候補をランキングの上位に提示することは重要な課題であり，様々なランキングアルゴリズムが提案されている．現在主要な QAC の手法である最頻補完 (Most Popular Completion) [1] は，検索ボックスに入力された文字列と合致するクエリ候補を過去の利用頻度が高い順に並び替えて提示する手法である．最頻補完には，文脈に関わらず，ある程度の性能で適切なクエリ候補を上位に提示できる利点あるが，詳細な検索ニーズに対応することはできない．より検索するユーザに合わせたクエリ候補を提示する手法として，個人情報やセッション情報，直近の検索クエリを考慮したランキング手法が提案されており，機械学習を用いたアプローチも数多く考案されている [1, 5, 6, 21].

過去のクエリログから検索回数の多いクエリ候補を探索する手法では，過去に一度も検索されたことのないクエリや，キーワード数の多いクエリには対応できない．近年では，深層ニューラルネットワークを活用した自然言語処理技術を用いることにより，そのようなクエリにも対応する手法も提案されている [15].

本論文では，十分な量の検索クエリログを入手することができないという状況下での QAC の実現を目的とする．従って，クエリログからクエリ候補となるクエリを探索する手法，ないしは，クエリログを利用してクエリを生成するモデルを構築し，そのモデルを用いてクエリを生成する手法に関する研究とは前提条件が異なる．

2.2 検索対象となる文書集合を用いた QAC の手法

クエリログを用いるものと比べると数は少ないが，クエリログが入手できない状況下において，検索対象となる文書集合を用いてクエリ候補を生成する研究が行われている．本論文は，これらの研究との高い関連性がある．

文書集合を使用する手法としては，Bhatia ら [2] の手法を始めとして，文書集合からクエリ補完候補となるフレーズを抽出し，ユーザが入力中のクエリに含まれるフレーズとの組み合わせについて共起度を計算した上で，それが高い順にランキングしたフレーズを返すという手法がとられることが多い．しかし，このような手法では，ユーザが入力しうるあらゆるフレーズとの組み合わせにおける共起度をあらかじめ計算し，保存しておく必要があることから，計算コスト，ストレージコストの面で課題がある．

Maxwell ら [14] や Rossiello ら [20] は，すべてのフレーズの組み合わせについての共起度を計算して保存しておく代わりに，ユーザが入力中のクエリを用いて文書の検索を行い，得られた文書から言語モデルないしはフレーズの共起度モデルを構築し，クエリ補完候補の抽出，ランキング付けを行う手法を提案している．この手法では，クエリ補完候補の抽出とランキング付けの計算をオンラインで行うことにより，ストレージコストを削減することができるが，ユーザがクエリを 1 文字入力するたびに，文書の検索とモデルの再構築を繰り返す必要がある．QAC では，一般的に数十ミリ秒以内で応答することを求められる [11, 16, 22] ことを考えると，この手法で実用的な QAC を実現することは容易ではない．

打田ら [24] は，既存のクエリログとそれに対応する文書集合のペアを用いて，クエリログのクエリ中に出現する単語の N-gram 確率と文書集合中におけるその単語の N-gram 確率の比率を推定する回帰モデルを構築し，得られたモデルを用いて，検索対象となる文書集合からクエリ補完候補を生成する手法を提案している．この手法では，計算コストを実用上耐えうる水準まで抑えることが可能である．なお，本論文は，検索対象となる文書集合からクエリそのものを生成する手法であり，目的の文書集合から，検索に使用しうるクエリの集合中に含まれる単語の割合を推定するこの研究とは異なる．

3 提案手法

本章では，まず，問題設定について説明する．次に，クエリを生成する手法について説明し，その手法をそのまま用いることの課題点を示す．最後に，その課題点に対する提案手法である，トピックモデルを用いた，使用頻度が高いと考えられるクエリを増やす手法について述べる．

3.1 問題設定

本論文では，検索サービス内に含まれる文書集合を入力とし，それを何らかのクエリ生成システムに入れることによって，その文書集合を検索するにあたり利用されるクエリ集合に近い分布を得られうるクエリ集合を生成する問題を解く．ただし，そ

の検索サービスのユーザ，及び，運営者等から，その検索サービスにおいて利用されうるクエリのトピックについては得られるものとする．

検索対象となる文書集合 D_T と D_T に対応するクエリ集合 Q_T を考える．文書集合 D_T について，クエリ集合 Q_T がその文書集合を検索するクエリの集合であるといえる時，「D_T と Q_T は対応している」と定義する．

3.2 クエリを生成する手法と課題点

クエリの生成は，文書から単語を抽出してクエリを生成する手法と，文書からクエリを生成する手法である docTTTTTTquery [7] を用いる手法の 2 つを使用する．docTTTTTquery は，高速でありながら，高いクエリ生成性能を持つことが報告されている [7] ことより，これを利用する．

単語抽出を用いるクエリ生成手法では，検索対象となる文書集合 D_T 中の各文書から単語をいくつか抽出し，それを空白文字でつなぐことによりクエリを生成する．これを繰り返すことにより，擬似クエリ集合を生成する．

また，docTTTTTquery を用いる手法では，以下のような流れでクエリの生成を行う．

（1） 既存の文書集合 D_S とそれに対応するクエリ集合 Q_S から，文書とクエリを 1 ペアずつ抽出し，T5 言語モデルに入力することで，文書からのクエリ生成パターンを学習した T5 言語モデル **TunedT5** を構築する．

（2） 検索対象となる文書集合 D_T から文書を 1 つずつ抽出して **TunedT5** に入力し，クエリ補完候補となるクエリ集合を得る．

上記 2 つの手法をそのまま用いてクエリの生成を行うと，文書集合 D_T 中における出現頻度が高い文書に関連するクエリほど多く生成される．しかし，一般に，D_T における文書の出現頻度とその文書に関連するクエリの利用頻度は一致しない [24]．例えば，ある大学のホームページのサイト内検索に用いるクエリの生成を考えた場合，入試や図書館などに関連するクエリが多く生成されることが望ましい．しかし，大学のホームページには研究や教職員に関する文書が多く存在することから，docTTTTTquery をそのまま用いた場合，研究に関するクエリばかりが生成されてしまう．従って，利用されうるクエリの集合を生成するためには，ただ単に D_T をそのまま入力するのではなく，何かしらの工夫をする必要がある．

3.3 提案手法

3.2 節で述べた課題点の解決法として，本節では，使用頻度が高いと考えられるクエリを増やすアプローチの提案を行う．具体的には，検索サービスの運営者やユーザに対し，各トピックに頻出する単語の集合を示すことにより，よく使用されると考えられるクエリは，どのトピックに属するものと考えられるかを複数回答で選択してもらう．次に，先述の調査により得たトピック分布に，検索対象となる文書集合 D_T のトピック分布が近づくように調整した新たな文書集合 D_{ext} を生成する．最後に，文書集合 D_{ext} から 1 つずつ文書を抽出し，文書からの単語抽出，または，言語モデル **TunedT5** への入力を行うことによりクエリを生成する．

トピック分布を得るためのモデルとしては，Latent dirichlet allocation (LDA) [3] を使用する．LDA は，文書集合中の各文書に含まれる単語の分布をもとに，その文書がどのトピックに属するかを確率的に推定するモデルである．

LDA を用いた文書集合 D_{ext} の生成は，以下の手順で行う．

（1） 検索対象となる文書集合 D_T を学習データ，トピック数を K として学習した LDA モデルを構築する．これにより，文書集合 D_T 中の各文書 d について，トピック分布 θ_d，各トピック k について，単語分布 ψ_k を得る．

（2） 学習済みの LDA モデルにおける各トピック k について，単語分布 ψ_k に基づき，出現しやすい上位 20 語の単語集合 W_k を出力する．

（3） 検索サービスの運営者やユーザ N 人に対し，各トピック k に頻出する単語の集合 W_k を示すことにより，よく使用されると考えられるクエリは，どのトピックに属するものと考えられるかを複数回答で選択してもらう．これにより，どのトピックがどの程度選択されたかを示す確率分布 $P^*(k)$ を得る．確率分布は式 1 で表される．

$$P^*(k) = \frac{\text{count}(k)}{N} \tag{1}$$

ここで，$\text{count}(k)$ は，トピック k が選ばれた合計回数を表す．

（4） 学習済みの LDA モデルを用いて，文書集合 D_T 中の各文書 d のトピック分布を得る．そして，そのトピック分布の確率に基づき，各文書がどのトピックに該当するかを割り当てる．この結果に基づき，文書集合 D_T のトピック k ごとの文書集合 D_k を得る．具体的な手順を Algorithm 1 に示す．

（5） 各トピックの文書集合 D_k から，各トピックごとの文書数が調査対象者によって選択されたトピック分布 $P^*(k)$ と同様になるように文書を抽出することで，文書集合 D_{ext} を生成する．具体的な手順を Algorithm 2 に示す．なお，N_Q は抽出数の合計値，すなわち，生成したいクエリ数を表す．また，Random は，集合から要素を 1 つランダムに抽出する関数を表す．

例えば，3.3 節の大学のホームページのサイト内検索システムの例を考えた場合，調査結果から図書館や入試に関するクエリを入力する可能性が高いという結果が得られたとすると，それらのトピックに含まれる文書の抽出割合を増やすことで，より利用されうるクエリの割合が高いクエリ補完候補の集合が得られると考えられる．

Algorithm 1 LDA によるトピック分類

Input: D_T, $\{\theta_d | d \in D_T\}, \alpha, \beta$
Output:$\{D_k | k \in \{1, 2, \cdots, K\}\}$
for each $d \in D_T$ **do**
  $w_d = \{w_{d,1}, w_{d,2}, \cdots, w_{d,N}\}$
  $\theta_d = \mathrm{argmax}_\theta p(\theta | w_d, \alpha, \beta)$
  $k \sim \mathrm{Multinoial}(\theta_d)$
  $D_k \leftarrow D_k \cup \{d\}$
end for

Algorithm 2 文書集合 D_k からの文書抽出と文書集合 D_{ext} の生成

Input: N_Q, $\{D_k | k \in \{1, 2, \cdots, K\}\}$, $P^*(k)$
Output: D_{ext}
$n \leftarrow 0$
while $n < N_Q$ **do**
  $k \sim P^*(k)$
  $d \sim \mathrm{Random}(D_k)$
  $D_{\mathrm{ext}} \leftarrow D_{\mathrm{ext}} \cup \{d\}$
  $n \leftarrow n + 1$
end while

4 実験

本節では，3.3 節で述べた提案手法に対する実験について説明する．

まず，使用するデータセットについて述べる．次に，実験設定と実験内容，評価指標について述べる．最後に，実験結果について考察を踏まえながら述べる．

4.1 データセット

クエリ集合として，公開されている大規模クエリログデータを使用し，クエリログが発行された時期と概ね同時期に収集された Web ページアーカイブデータセットを，クエリログに対応する検索対象の文書集合として使用する．

使用するデータセットは，ORCAS クリックスルーログ [8] と MS MARCO データセット [17] である．ORCAS クリックスルーログは，2018 年から 2020 年にかけて収集された Web ページのクエリログデータをもとに，k 匿名化処理を施し，クエリ発行者の個人情報が復元できないように集約されたデータセットである．MS MARCO データセットは，TREC Deep Learning Track [9] 向けの公開データセットで，ORCAS クリックスルーログと同時期に収集された Web ページをもつ．ORCAS データセット中には，クエリログデータのほか，各クエリにより，MS MARCO データセット中のどの Web サイトがクリックされたかを保持する適合性判定データが含まれていることから，クエリ集合とその検索対象となる文書集合のペアとみなすことができる．

ORCAS データセットについて，アルファベットまたは数字または空白文字のみからなるクエリを抽出し，空白文字またはカンマまたはピリオドで単語に分割した．また，ORCAS データセット中のクエリに出現する，空白文字が抜けた単語 (文字列) やアダルト関連の単語を取り除くため，MS MARCO データセット中の Web ページのタイトルまたは本文に一度も出現しない単語，10 文字以上の単語を除外した．

表 1　ORCAS/MS MARCO データセット

名前	種類	Eval	Total
ORCAS	クエリ	8,704,022	17,408,043
MS MARCO	文書	-	3,213,835

MS MARCO データセットについて，各 Web ページのタイトルと本文を抽出し，ORCAS データセットと同様に，空白文字またはカンマまたはピリオドで単語に分割した．また，各単語のうち，アルファベットまたは数字以外の文字を含む単語を除外した．

実験のため，ORCAS クリックスルーログについて，ランダムに 50 %をサンプリングして評価データとした．評価データは,「予測したい理想のクエリ集合」として利用する．

各データセットの統計情報を表 1 に示す．表見出しの「Eval」は評価データを表す,

4.2 実験設定

本研究においての実験の設定について述べる．実験では，ORCAS/MS MARCO データセットペアを使用し，MS MARCO データセットを「検索システムにおいて検索対象となる文書集合」ORCAS クリックスルーログの評価データを「その検索システムにおいて使用されうるクエリ集合」として用いる．そして，MS MARCO データセット中の文書を用いて生成したクエリ集合を，ORCAS 評価データのクエリ集合に近づけることを目指す．

生成したクエリ集合と評価データのクエリ集合を比較するため，クエリログを単語の出現確率分布とみなすことを考える．まず，クエリログが N 個の単語から構成されると定義する．この時，単語 w_i の出現回数を n_i と表すことができる．次に，確率分布を導入する．クエリログから得られる確率分布 $P(w_i)$ は，各単語の出現確率を示す．これは，単語の出現回数を全単語の出現回数の総和で割ることで求められる．従って，クエリログは式 2 のような単語の出現確率分布で表される．

$$P(w_i) = \frac{n_i}{\sum_{i=1}^{N} n_i} \tag{2}$$

ここで，$P(w_i)$ は単語 w_i の確率，n_i は単語 w_i の出現回数，N は全単語数を表す．

上記のことを踏まえ，クエリ集合同士の比較は，生成したクエリ集合の確率分布について，評価データのクエリ集合の確率分布からの距離を測ることにより行う．評価データのクエリ集合の確率分布に近いほど，より実際のものに近いクエリ集合が生成できていると考えられる．

4.3 クエリの生成実験

MS MARCO データセット中の文書を用いて擬似的なクエリ集合を生成する実験を行った，クエリの生成は，文書からの単語抽出と docTTTTTquery の 2 つの手法を使用した．

4.3.1 単語抽出

MS MARCO データセット中の各文書からの単語抽出により，クエリを生成した．具体的には，MS MARCO データセット中の各文書から単語を抽出し，抽出した単語を空白文字でつなぐことにより，擬似クエリを生成した，そして，これを繰り返すことにより，擬似クエリ集合を生成した．

クエリ生成時の単語の抽出条件を探るため，以下のような実験を行った．なお，以下の全ての実験において，タイトルまたは本文が規定の抽出語数に満たなかった場合には，そのタイトルまたは本文の全ての単語を抽出した．また，1 つのクエリ内に同じ単語が重複して出現することがないよう，重複する単語を取り除いた上で規定語数の単語の抽出を行った．

1. 抽出元

文書のタイトルから抽出して生成したものと文書の本文から抽出して生成したものの性能を比較した．具体的には，MS MARCO 中の各文書のタイトルまたは本文からランダムで単語を抽出することで，クエリ集合を生成した．

2. 抽出する単語数

クエリの生成にあたり，各文書から抽出する単語数を変化させたときの性能の変化を観察した．具体的には，文書から抽出する語数を 0 語〜タイトル中の全単語の範囲で変化させながらクエリを生成した．

3. クエリの生成数

クエリの生成数を 100,000 件ずつ増やしていき，性能の変化を観察した．具体的には，MS MARCO 中の文書のタイトルからランダムで単語を抽出してクエリを生成し，100,000 件生成するごとに，その時点での生成したクエリ集合と評価データのクエリ集合の確率分布間距離を計算した．

4. IDF が高い語を優先的に使用

IDF (Inverse document frequency) が高い単語は，特定の文書のみにしか出現しない単語であることから，検索によって文書を絞り込むクエリを生成するにあたり，有効であると考えられる．そこで，文書から IDF の高い単語を優先して抽出することで，実際のものに近いクエリ集合を生成できるかを観察した．

5. 特定の品詞を除外

“a”や“have”などといった，どの文書にも出現し，かつ，特に意味を持たない単語は，クエリには使用されないと考えられる．そこで，文書中の単語を品詞分類した上で，特定の品詞の単語を除外した上でクエリを生成した．

6. 提案手法と単語抽出を用いてクエリ生成

提案手法の有効性を確かめるため，MS MARCO データセット中の文書のトピック分布を調整した新たな文書集合 D_{ext} を生成し，D_{ext} 中の各文書からの単語抽出によりクエリを生成する実験を行った．使用されうるクエリのトピック分布について，実際には，検索対象の文書集合を用いて学習した LDA モデルの各トピックにおいて生成確率の高い単語上位 20 件 (どのトピックにも出現するものは除く) をもとに，検索サービスのユーザや運営者にどのトピックのクエリを使用する可能性が高いかの調査を行うことにより入手する．しかし，本実験では，ORCAS の評価データを除く残り 50%のデータ（以下，ORCAS 残データ）中のクエリを学習済みの LDA モデルによってトピック分類を行うことにより，クエリ集合全体のトピック分布を得ることで，調査を代替した．

LDA を利用したクエリ集合の生成実験は，以下の流れで行った．

（1） MS MARCO データセット D_T 中の各文書のタイトルと本文を空白文字で結合した文字列 t_T 用いて LDA のモデルの学習を行った．

（2） 得られた LDA のモデルを利用して，ORCAS 残データ Q_{tra} 中のクエリのトピック分布の推定を行った．これにより，クエリ集合 Q_{tra} 全体のトピック分布を得た．

（3） 同様に，得られた LDA のモデルを利用して，D_T 中の各文書のトピック分布の推定を行った．この推定確率をもとに，各文書につき 1 つのトピックを割り当てた．これにより，D_T における各トピックごとの文書集合 D_k を得た．D_k は，D_T のうち，トピック k に割り当てられた文書の集合を表す．

（4） 入力として与える文書のトピック分布が Q_{tra} のトピック分布と同様の分布になるように，各トピックの文書集合 D_k から各文書 d_k を取り出した．次に，文書 d_k のタイトルから単語を 4 語ランダムに抽出し，クエリを生成した．これを繰り返し，擬似クエリ集合を生成した．

4.3.2 docTTTTTquery

docTTTTTquery を用いてクエリ集合を生成する実験を行った．docTTTTTquery における T5 言語モデルの訓練には，AOL クエリログ [18] を使用する．AOL クエリログは，2006 年 3 月から 5 月にかけて収集された Web ページのクエリログデータセットである．このデータセットには，ORCAS データセットと同様に，各クエリによりクリックされた Web サイトの URL を保持するデータが含まれていることから，各 URL に対応する Web ページアーカイブのコレクションを，クエリ集合とペアとなる検索対象の文書集合として使用する．なお，AOL クエリログは，個々のユーザの情報が特定されうるというプライバシー上の問題を含んでいることより，本論文では，AOL クエリログ中のデータのうち，個人を特定する情報を削除または置き換えたデータセットである AOLIA データセット [13] を使用した．

データセットの統計情報を表 2 に示す．なお，クエリと文書のペア数は，18,933,057 件であり，内 10%をランダムに抽出し，訓練に使用した．

表 2　AOLIA データセット

名前	種類	Total
AOL	クエリ	9,966,939
WebArchive	文書	1,525,305

クエリの生成には，MS MARCO データセットを使用した．クエリの生成は，以下の流れで行った．

（1） AOL データセット D_S から，Web ページの文書 d_S とそれに対応するクエリ q_S を 1 つずつ取り出した．各文書のタイトルと本文を空白文字で結合することにより，文字列 t_S を生成した．

（2） t_S のうち，先頭から 512 語以降の単語は消去したのち，変換前を t_S，変換後を q_s として docTTTTTquery のモデルを学習した．

（3） MS MARCO データセット D_T 中の各文書のタイトルと本文を空白文字で結合した文字列 t_T を学習済み docTTTTTquery 言語モデルに入力することによりクエリを生成した．これを繰り返し，擬似クエリ集合を生成した．

docTTTTTquery の学習時のステップ数は 10,300，1 単語 (トークン) あたりのバッチサイズは 131,072 とした．

また，提案手法の有効性を確かめるため，単語抽出を用いたクエリ生成手法と同様にして，MS MARCO データセット中の文書のトピック分布を調整した新たな文書集合 D_{ext} を生成し，D_{ext} 中の各文書を学習済み docTTTTTquery 言語モデルに入力することにより，クエリを生成する実験を行った．

4.4　評価指標

確率分布間の距離尺度には，Jensen-Shannon divergence (以下，JS divergence) を用いた．JS divergence は Kullback-Leibler divergence (以下，KL divergence) を平滑化した距離尺度である．式 3 に計算式を示す．

$$
\begin{aligned}
D_{JS}(P||Q) &= \frac{1}{2}D_{KL}(P||\frac{P+Q}{2}) + \frac{1}{2}D_{KL}(Q||\frac{P+Q}{2}) \\
D_{KL}(P||Q) &= \sum P(w_i) log \frac{P(w_i)}{Q(w_i)}
\end{aligned}
\tag{3}
$$

KL divergence は，2 つの確率分布 P と Q がどれだけ異なるかを示す非対称の測定尺度である．従って，$D_{KL}(P||Q)$ と $D_{KL}(Q||P)$ は一致しないことがある．一方，JS divergence では，確率分布 P と Q それぞれの分布と両者の平均分布の KL divergence の平均に基づいて計算される対称性のある測定尺度である．この対称性により，2 つの確率分布間の違いをより均等に評価することができるできるため，前述のような生成したクエリ集合中に不適切なクエリが含まれている場合にも適切な評価を下すことができる．

JS divergence では，いずれか一方の確率分布にしか出現しない単語がある場合でも，KL divergence のように無限大に発散することはなく，有限の値に収まる．しかし，JS divergence の計算においても，どちらかの確率分布にしか出現しない単語がある場合には，計算式中に含まれる対数関数が無限大に近づき，計算上の誤差が増大する問題が生じる．この問題の発生を防ぐため，どちらか一方の確率分布にしか出現しない単語が存在する場合には，もう一方の確率分布において，その単語が 1 回出現したものとして扱う加算スムージングを行う．このスムージング処理により，計算式中の対数関数に与える影響が緩和され，より正確に計算できることが期待される．

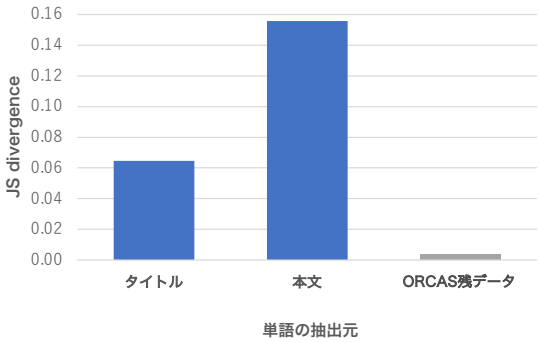


図 1　クエリ生成時の単語の抽出元の違いによる JS divergence の比較

4.5　実験結果

4.5.1　単語抽出

以下の単語抽出によるクエリの生成実験において，特に述べない限り，クエリの生成数は 3,213,835 (=MS MARCO の文書数)，クエリの生成時に抽出する単語数は 4 語，単語の抽出元は文書のタイトルとした．

1．抽出元

図 1 に，クエリ生成のための単語抽出において，タイトル中の単語を使用した場合と本文中の単語を使用した場合の確率分布間距離の違いを示す．縦軸が JS divergence の値であり，横軸がクエリ生成時の単語の抽出元である．JS divergence は値が小さいほど，確率分布間の距離が小さい，すなわち，ORCAS 評価データとのクエリ分布間の距離が小さいことを意味する．なお，比較対象として，ORCAS 残データと評価データ間の JS divergence を「ORCAS 残データ」として示す．本実験では，クエリの抽出単語数を 3.27 とし，これは，文書のタイトルまたは本文の語数が 4 語以上であった時，73%の確率で 3 語，27%の確率で 4 語を抽出することにより行った．

実験の結果，タイトルから抽出した単語を用いて生成したクエリ集合の方が，本文から抽出した単語を用いたものよりも低い値を示した．

このことから，文書中の単語を用いてクエリを生成するよりも，タイトル中の単語を用いてクエリを生成する方が有効であると考えられる．

2．抽出する単語数

図 2 に，クエリ生成において，各文書のタイトルから抽出する単語数による確率分布間距離の違いを示す．縦軸が JS

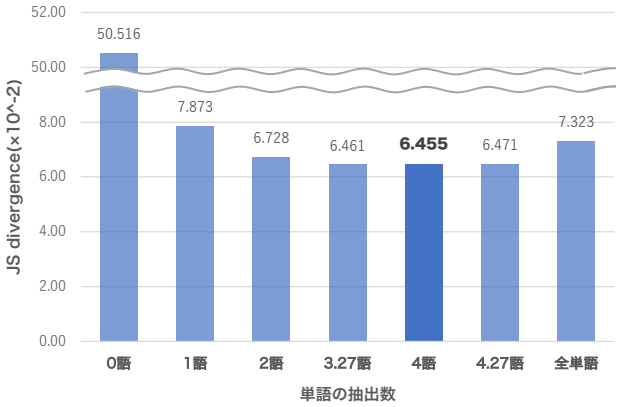


図 2　クエリの単語数の違いによる JS divergence の比較

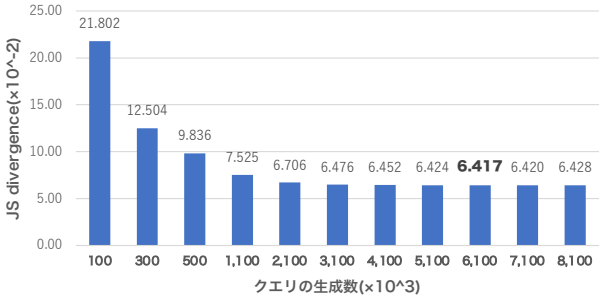


図 3　クエリの生成数の違いによる JS divergence の比較

divergence の値に 10^2 をかけたものであり，横軸がクエリ生成時の抽出単語数である．なお,「0 語」は，MS MARCO データセット D_T 中のタイトルおよび ORCAS 評価データ中のクエリに含まれる全ての単語が等確率で出現すると予測した場合,「全単語」は，タイトル中に含まれる単語を全て用いてクエリを生成した場合を示す．

実験の結果，クエリの長さが 4 語の時，最も低い値を示した．語数が少なすぎると性能が低い理由としては，サンプル数が少ないために上手くクエリ分布が推定できないことが考えられる．また，語数が多すぎると性能が下がる理由としては，抽出した文書に基づくクエリ分布を過剰に推定する状態になり，実際のクエリログ中のクエリ分布とは離れてしまうことが考えられる．また，ORCAS の平均クエリ語数である 3.27 語の時よりも，4 語の時の方が性能が良い理由としては，語数が 4 語に満たないタイトルがを持つ文書が多く存在し，生成したクエリ集合全体として，4 語を抽出した時の方がクエリの語数平均が 3.27 語に近いからであると考えられる．

このことから，クエリを生成する際には，各文書のタイトルから 4 語を抽出すると有効であると考えられる．

3. クエリの生成数

図 3 に，クエリ生成において，生成するクエリ数による確率分布間距離の変化を示す．縦軸が JS divergence の値に 10^2 をかけたものであり，横軸がクエリの生成数に $10^{-}3$ をかけたものである．本実験では，同じ操作を 50 回繰り返し，その平均値を実験結果とした．また，クエリの生成数が MS MARCO の文書数 (=3,213,835) を超える場合には，同じ文書集合中からランダムで文書を抽出し，その文書から再度単語を抽出することにより，クエリを生成した．

実験の結果，クエリの生成数を増やすほど，確率分布間の距離が小さくなっていき，生成数が 610,000 の時に最も性能が良くなることが示された．クエリの生成数が少ない時に性能が低い理由としては，生成されたクエリに偏りが生じやすく，分布を推定することが難しいことが考えられる．また，生成数が多すぎた時に性能が低くなる理由としては，実際のクエリログに出現しない単語の出現確率がより高くなってしまうからであると考えられる．

一方で，クエリを 500,000 件ほど生成すれば，ある程度はクエリの分布が推定可能であることも示された．

4. IDF が高い語を優先的に使用

表 3 に，クエリ生成において，IDF が高い単語を優先的に使用したものとランダムで抽出した単語を用いてクエリを生成したものの確率分布間距離を示す．本実験では，まず，MS MARCO 中の各文書のタイトルと本文中に出現する各単語を key，出現回数を value とした辞書を作成した．次に，この辞書を用いて，出現回数が 5 回以上 100,000 回未満の単語を優先的に使用してクエリを生成した．

実験の結果，IDF が高い単語を優先的に使用するよりも，ランダム抽出した単語を用いてクエリを生成した方が確率分布間距離が小さいことが示された．このような結果になった理由としては，”weather”，”why” などといった，IDF の値は高くないが，クエリには頻繁に用いられる単語が多く存在するからであると考えられる．

このことから，クエリの生成において，IDF が高い単語を優先的に使用することは有効ではないと考えられる．

5. 特定の品詞を除外

表 3 に，クエリ生成において，特定の品詞の単語を除外した上で生成したクエリ集合の確率分布間距離を示す．本実験では，まず，MS MARCO の各文書のタイトル中に含まれる単語を spaCy を用いて品詞分類を行った．次に，MS MARCO の各文書のタイトル中に含まれる単語のうち，特定の品詞の単語を除外した上で単語を抽出し，クエリを生成した．除外する品詞は，等位接続詞，助動詞，限定詞，助詞，感嘆詞，従属接続詞，設置詞と spaCy で分類可能ないずれの品詞にも属さない単語とし，それぞれの品詞の単語を除外してクエリを生成した際の性能の変化を観察した．表 3 には，本実験において最も性能が良い結果となった，限定詞を除外してクエリを生成した場合の結果を示す．

実験の結果，ある 1 つの品詞を除外した上でクエリを生成した場合，いずれの品詞においても，ランダム抽出した単語を用いてクエリを生成した場合よりも確率分布間距離が大きくなることが示された．このようになった理由としては，クエリ中の単語としては用いられない特定の品詞は存在せず，どのような品詞の単語であってもクエリに用いられうるからと考えられる．

表 3　単語の抽出条件を変えることによるクエリ生成性能の比較

生成方法	JS divergence
Random	0.0646
Random + LDA	**0.0644**
IDF が高い単語を優先的に使用	0.0659
特定の品詞の単語を除外	0.0647

表 4　クエリ生成における docTTTTTquery と単語抽出の性能比較

生成方法	JS divergence
T5	0.2113
T5 + LDA	0.1629
Random	0.0984
Random + LDA	**0.0910**

このことから，クエリの生成における単語抽出において，特定の品詞の単語を除外することは有効ではないと考えられる．

6. 提案手法を使用

表 3 に，入力文書集合のトピック分布を調整した上で文書を抽出し，各文書からランダムで単語抽出を行うことにより生成したクエリ集合の確率分布間距離を示す．本実験では，LDA で推定するトピック数を 200，クエリの生成数を 3,200,000 とした．

実験の結果，入力文書集合から生成目標とするクエリ集合のトピック分布をもとにトピックを調整して文書を抽出し，クエリを生成した場合，ランダムで抽出した文書からクエリを生成した場合よりも確率分布間距離が小さくなった．

このことから，目標とするクエリ集合のトピック分布をもとに，入力文書のトピック分布を調整してクエリを生成することは，有効であると考えられる．

4.5.2　docTTTTTquery

表 4 に提案手法によって生成したクエリ集合の確率分布間距離を示す．「T5」は，文書集合 D_T からランダムで抽出した文書から docTTTTTquery を用いて生成したクエリ集合と評価データ間の分布間距離，「T5+LDA」は，入力文書集合 D_T のトピック分布を調整することにより生成した新たな文書集合 D_{ext} から取り出した文書を docTTTTTquery に入力することにより生成した文書集合と評価データ間の分布間距離を表す．本実験において，LDA で推定するトピック数は 200，クエリの生成数は 500,000 とした．

実験の結果，いずれの方法においても，単語抽出を用いてクエリを生成した場合と比較して，確率分布間距離が大きくなった．理由としては，AOLIA データセット内に URL を用いたクエリがかず多く含まれており，“https”，“com”などといった，URL でよく使用される単語が過剰に生成されてしまったことが考えられる．また，「a c i m the d a c」などといった，意味を持たない文字列からなるクエリも数多く生成されてしまったことも原因の 1 つとして考えられる．T5 言語モデルの学習時に，URL を含むクエリを持つ文書の入力を減らす，ないしは，取り除くことにより，クエリの生成性能が向上する可能性がある．また，クエリの生成において，docTTTTTquery ではなく，GPT-3 [4] などの別の手法を使用することにより，生成性能が向上する可能性がある．

5　まとめ

本論文では，検索サービスにおいて，十分な検索クエリログが入手できない状況下における，ユーザが入力しうるクエリ集合を生成する手法の提案を行った．小規模の検索サービスにおいては，よく利用されうるクエリのトピックについては推定可能であることが多い．そこで，目標とするクエリ集合のトピック分布と入力する文書集合のトピック分布が近くなるように調整し，その上でクエリを生成することにより，その検索サービスにおいて十分なクエリが入力された際に得られるであろうクエリログのクエリ分布と近いクエリ分布を持つクエリ集合を生成できると考えた．実験では，大規模クエリログデータセットを使用し，提案手法の有効性を検証する実験を行った．実験により，生成目標とするクエリ集合のトピック分布をもとに，入力として与える文書集合のトピック分布を調整することで生成した新たな文書集合を用いて生成したクエリ集合は，元の文書集合をそのまま用いて生成したクエリ集合よりも評価データとの確率分布間距離が小さくなった．これは，提案手法が有効であることを意味する．また，クエリの生成において，4 語の単語を用いてクエリを生成することが有効であることも示された．今後の課題としては，docTTTTTquery 以外の言語モデルを用いてクエリを生成した場合にどうなるのか，目標とするクエリ集合のトピック分布を得るにあたり，クエリの想定が何件あれば，文書集合中の全文書を入力とする場合と比較して優位になりうるトピック分布が得られるかなどがあげられる．

謝辞　本研究は JSPS 科研費 JP21H03554，JP21H03775，JP23H03400 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Ziv Bar-Yossef and Naama Kraus. Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 107–116, 2011.

[2] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 795–804, 2011.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. page 993–1022, 2003.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Fei Cai and Maarten de Rijke. Learning from homologous queries and semantically related terms for query auto completion. *Inf. Process. Manage.*, 52(4):628–643, jul 2016.

[6] Fei Cai, Shangsong Liang, and Maarten de Rijke. Time-sensitive personalized query auto-completion. In *Proceed-*

ings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, page 1599–1608, 2014.

[7] David R. Cheriton. From doc2query to doctttttquery. 2019.

[8] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. Orcas: 18 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2983–2989, 2020.

[9] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*, 2020.

[10] Silviu Cucerzan and Eric Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 293–300, Barcelona, Spain", 2004.

[11] Simon Gog, Giulio Ermanno Pibiri, and Rossano Venturini. Efficient and effective query auto-completion. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2271–2280, 2020.

[12] Maryam Kamvar and Shumeet Baluja. The role of context in query input: Using contextual signals to complete queries on mobile devices. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '07, page 405–412, 2007.

[13] Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Reproducing personalised session search over the aol query log. In *European Conference on Information Retrieval*, 2022.

[14] David Maxwell, Peter Bailey, and David Hawking. Large-scale generative query autocompletion. In *Proceedings of the 22nd Australasian Document Computing Symposium*, ADCS '17, 2017.

[15] Bhaskar Mitra and Nick Craswell. Query auto-completion for rare prefixes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1755–1758, 2015.

[16] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modelling. *Comput. Speech Lang.*, 8:1–38, 1994.

[17] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660, 2016.

[18] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale '06, page 1–es, 2006.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[20] Gaetano Rossiello, Annalina Caputo, Pierpaolo Basile, and Giovanni Semeraro. Modeling concepts and their relationships for corpus-based query auto-completion. *Open Computer Science*, 9(1):212–225, 2019.

[21] Milad Shokouhi. Learning to personalize query auto-completion. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 103–112, 2013.

[22] Sida Wang, Weiwei Guo, Huiji Gao, and Bo Long. Efficient neural query auto completion. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2797–2804, 2020.

[23] Aston Zhang, Amit Goyal, Weize Kong, Hongbo Deng, Anlei Dong, Yi Chang, Carl A. Gunter, and Jiawei Han. Adaqac: Adaptive query auto-completion via implicit negative feedback. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 143–152, 2015.

[24] 打田智子, 加藤誠, et al. クエリらしさに基づく文書コレクションを利用したクエリ自動補完. **情報処理学会論文誌データベース** *(TOD)*, 16(2):1–15, 2023.

音声対話型検索における階層型検索意図を用いたクエリ推薦の戦略

平位　遼† 飛岡　憲†† 山本　岳洋††

† 兵庫県立大学 社会情報科学部　〒 651–2197 神戸市西区学園西町 8-2-1
†† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †{fa20k070,ad22h042}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　音声検索技術の利用が増加傾向にあり，生成 AI もマルチモーダル化を進めている現在，音声対話による検索の需要はさらに増加していくことが見込まれる．そして，そのような音声対話型検索においてもクエリ推薦はユーザの多様な情報獲得を支援すると考えられるが，クエリ推薦に関する研究の多くはデスクトップでの検索を対象としており，音声によるクエリ推薦としては情報量が多くなりすぎてしまうという問題がある．そこで本研究では想定されるユーザの検索意図を特化と平行移動の 2 つの主要なクエリ修正パターンをもとに階層構造で表現し，この構造を用いることでクエリ推薦を行う音声対話型検索システムを提案する．そして，このシステムを用いたユーザ実験を行うことで階層型検索意図を用いたクエリ推薦が音声におけるクエリ推薦の手法として有効であるかを検証した．

キーワード　情報検索，対話型検索，音声対話，クエリ推薦, 多様化

1　はじめに

近年，Apple の Siri や Amazon の Alexa といった音声対話システムを日常的に利用する機会が増えている．音声認識市場も拡大傾向にあり，世界全体の市場規模は 2022 年に 104 億 2 千万米ドルと評価され，2030 年には 596 億 2 千万米ドルまで成長すると予測されている [1]．また，OpenAI の GPT-4 や Google の Gemini などの非常に高性能で自然な対話を行うことができる生成 AI のマルチモーダルモデルも登場しており，音声対話システムを利用した情報検索は従来のような一問一答形式の単純なものに止まらず，あらゆる分野において高度でより対話的な情報検索が可能となっている．そのため，学習や意思決定の際に行われる探索的検索にも音声対話システムの利用が拡大していくことが考えられる．

探索的検索とは，情報要求が明確に決まっていない検索者が反復的に検索を行っていくことで問題への理解を深め，真に必要な情報の獲得を目指す検索方法のことである [8]．この探索的検索の実現のためのアプローチとして，システム側から検索すべきクエリを提案することで検索者がより多様な情報を獲得できるよう促すクエリ推薦がある．例えば，エナジードリンクの健康への影響について知りたいとき,「エナジードリンク 健康」と検索すると推薦クエリとして「エナジードリンク 副作用」「エナジードリンク 成分」「エナジードリンク カフェイン」などが提示される．すると，ユーザはこれらの情報を次に検索するクエリの候補として使うことができ，探索的検索を進めることができる．実際に既存手法として探索的検索を支援するためのクエリ推薦手法が存在し [6] [12]，そのほとんどはデスクトップでの検索を対象としたものであるが，音声での検索においても多様な情報獲得を目的としたクエリ推薦は有効であると考える．

しかし音声対話型の検索におけるクエリ推薦の問題点として，音声ではデスクトップに表示されるような大量の情報を一度の出力で扱えないことがあげられる．過剰に長い音声テキストはユーザの認知負荷を高めることになり，推薦した情報がユーザの記憶に残らないことが想定されるからだ．そのため音声においては推薦するクエリを何らかの方法で絞り込む必要があると考える．

そこで本研究では，それぞれの検索意図が持つ相互関係に着目することで現在の検索状況に適したクエリに絞り込むというアプローチを提案する．Boldi らによると，同じ検索対象について行う 2 つの連続したクエリ間のクエリ修正には特化 (Specialization) と平行移動 (Parallel move) という 2 つの主要なパターンがある [3]．特化とは「エナジードリンク 成分」から「エナジードリンク カフェイン」のようにより詳細な情報を得るためのクエリ修正パターンを指す．平行移動とは,「エナジードリンク 成分」から「エナジードリンク 副作用」のように関連する側面を持つが同等ではないようなクエリ修正パターンを指す．提案手法では，この特化と平行移動の考え方をもとにそれぞれの検索意図が持つ相互関係を階層構造で表現した階層型検索意図を用意し，これを用いてクエリ推薦を行う音声対話型検索システムを構築する．

評価実験では，提案手法によるクエリ推薦システムとベースラインとなる完全ランダムなクエリ推薦システムを用意し，共通の検索システムと組み合わせた音声対話インタフェースを用いて被験者内計画を行った．そして，実験により得られたログデータや参加者へのアンケート結果を分析することで，提案手法によるクエリ推薦が検索者の探索的検索および検索後の意思決定への支援として有効であるかを検証した．

1：https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382

2 関連研究

2.1 情報推薦を行う音声対話システムに関する研究

情報推薦を行う音声対話システムについてのさまざまな研究がなされている.

翠らは京都の観光案内情報を対象として，ユーザの嗜好にあった観光スポットの決定を支援する音声対話システムを提案した [14]．提案されたシステムでは，直接観測できないユーザの知識や嗜好を部分観測マルコフ仮定における対話のモデル化によって推定し，対話の最適化を行うことで適切な情報推薦が行えることを示した.

吉野らは Web ニュースなどのテキストに対して，述語項構造に着目した情報抽出を行い，それに基づいた情報検索，推薦を行う音声対話型検索システムを提案した [11]．提案されたシステムでは，情報推薦を行う際に特定の述語項構造を持つ文のみを検索対象として，直前のシステム側の応答と最も類似した文を情報推薦として生成する.

これらの研究ではドメインや推薦できる情報が限定的なものであるが，本研究ではより一般的な情報検索を対象とした情報推薦について扱う.

2.2 探索的検索を支援する手法に関する研究

ユーザの多様な情報獲得に貢献し，探索的検索を支援するためのさまざまな手法が研究されている.

Kato らは Microsoft の Bing におけるクエリログ，クリックスルーデータをもとにエンティティとクエリのクラスタリングを行い，特化と平行移動に沿うようクエリ推薦をユーザに提示するユーザインタフェースを提案した [6]．提案されたインタフェースは同様の推薦クエリをフラットリストで表示する比較手法と比べて，検索者が必要なクエリを発見しやすく，クエリ推薦の意図を理解しやすいことを示した.

Medlar らは検索結果とともに同様の検索結果を取得するクエリを代替クエリとして表示することで，クエリ推薦を行いつつ検索結果が検索意図に合致しているかの判断を助けるクエリ推薦の手法を提案した [9]．ユーザは代替クエリを文書の要約として利用することで，検索意図を満たす文書を効率よく閲覧することができるようになった.

佐藤はタグが付与された検索結果について，そのタグが持つ階層構造をもとに検索結果を構造化し，検索者の多様な情報アクセスを支援するユーザインタフェースを提案した [12].

これらの研究はデスクトップでの検索を対象としたものであり，音声対話型検索を対象として探索的検索を支援する手法についての研究はない.

2.3 情報検索システムの対話戦略に関する研究

情報検索システムにおいて，ユーザとの対話により適切な情報獲得を支援する対話戦略についてさまざまな研究がなされている.

中野らは問い返し質問文を生成することで，ユーザの質問文における曖昧性を解消する質問応答システムを提案した [15]．QA データセットから文構造を利用して自動的に作成した曖昧な質問文をもとに，問い返し質問文により質問応答の精度が向上することを示した.

翠らはユーザの検索要求を明確化するための対話戦略を提案した [13]．係受け関係，ボトムアップの知識，トップダウンの知識からそれぞれ質問を用意し，情報ゲインを基準に選択することで，効率よく目的の文書に到達するための対話を実現した.

これらの研究はユーザの検索意図を明確にすることを目的とした対話戦略であり，本研究ではユーザの探索的検索による多様な情報獲得や意思決定を支援することを目的とする.

3 階層型検索意図を用いたクエリ推薦

本節では，提案手法について述べる．まず本研究で扱う階層型検索意図の定義について説明する．その後階層型検索意図を用いてクエリ修正を行う方法について述べる.

3.1 階層型検索意図の定義

本研究で扱う階層型検索意図は，1 つの検索トピックに対して検索意図を大きく分類する 1 層目と，1 層目の検索意図をより詳細に検索する 2 層目の 2 つの階層で構成する.

例として，後述の評価実験で検索タスクに用いたエナジードリンクについての階層型検索意図を図 1 に示す.

「成分」から「糖分」のような 1 層目から直下にある 2 層目への遷移が特化を表現し,「成分」から「効果」のような 1 層目内での遷移および「糖分」から「カフェイン」のような共通の 1 層目（この例では「成分」.）に属する 2 層目内での遷移が平行移動を表現するよう設計した.

3.2 階層型検索意図を用いたクエリ推薦

提案手法では，ユーザが最後に検索した際の検索意図（以下，現在の検索意図とする.）が 3.1 節で述べた階層型検索意図の中の 1 層目にあたる場合と 2 層目にあたる場合で異なるクエリ推薦を行う.

現在の検索意図が階層型検索意図の中の 1 層目にあたる場合，1 層目の検索意図のうち現在の検索意図を除くものからランダムに 1 つ，現在の検索意図の直下にある 2 層目の検索意図からランダムに 1 つを推薦する．例えば，図 1 において，現在の検索意図が「成分」であるとき，1 層目の「成分」を除いた「効果」「副作用」「代替案」「製品」からランダムに 1 つ，直下にある 2 層目の「糖分」「カフェイン」「アミノ酸」「ビタミン B」からランダムに 1 つが推薦される.

現在の検索意図が階層型検索意図の中の 2 層目にあたる場合，現在の検索意図が属するものを除いた 1 層目の検索意図からランダムに 1 つ，現在の検索意図を除き共通の 1 層目に属する 2 層目の検索意図からランダムに 1 つを推薦する．例えば，図 1 において，現在の検索意図が「摂取量」であるとき，1 層目の「効果」を除いた「成分」「副作用」「代替案」「製品」からランダムに 1 つ,「効果」に属する 2 層目のうち「摂取量」を除いた「タイミング」「持続時間」からランダムに 1 つが推薦される.

この 2 つのクエリ推薦はともに，ユーザに対して多角的な検

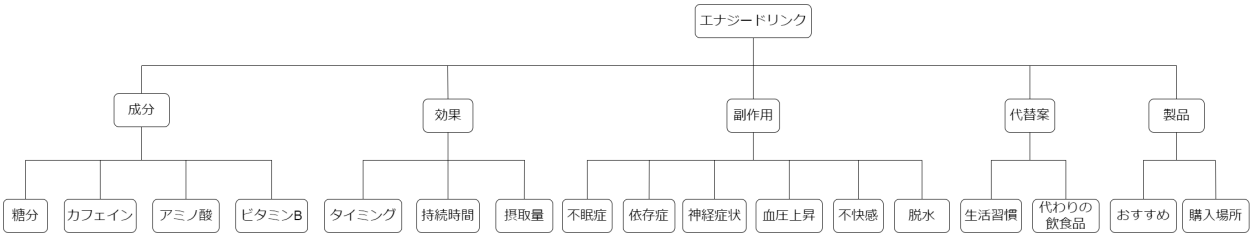


図 1　エナジードリンクの検索タスクで用いる階層型検索意図.

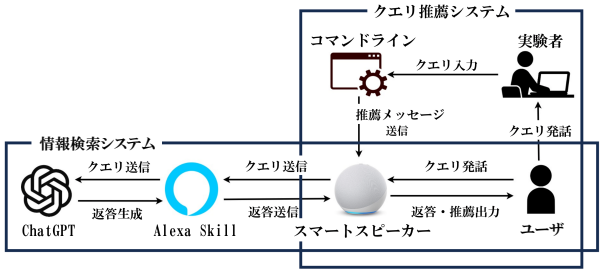


図 2　音声対話型検索システムの全体像.

索を促す要素と詳細な検索を促す要素を組み合わせて推薦することを目指した.

4　提案手法による音声対話型検索システム

本研究では，提案手法によってクエリ推薦を行う音声対話型検索システムを構築した．本節ではまず音声対話型検索システムの全体像を説明する．その後このシステムを構成する 2 つのサブシステムについてそれぞれ説明する.

4.1　音声対話型検索システムの全体像

音声対話型検索システムの全体像を図 2 に示す．このシステムは，ユーザの任意の発話に対して簡潔な回答を音声で返す情報検索システムと，ユーザの現在の検索意図に基づき推薦メッセージを音声により出力するクエリ推薦システムの 2 つのサブシステムから構成される.

4.2　情報検索システム

情報検索システムには，Alexa Skills Kit[2]を用いて Open AI の Chat Completions API[3]を介した返答を行うスキルを自作した．ChatGPT のモデルは gpt-3.5-turbo-0613 を使用し，出力のランダム性を制御する temperature パラメータは 0 に，出力の最大トークン数を制限する max_tokens パラメータは 512 に設定した．プロンプトは以下のように与えた.

プロンプト

以下は AI アシスタントとの会話です．AI は人がした質問に対して 2〜3 文で 100 文字以内の簡潔な回答をします.
人:（ここにユーザ発話が入る.）
AI:

4.3　クエリ推薦システム

本節では，クエリ推薦システムの各機能について説明する.

4.3.1　階層型検索意図の読み込み

検索するトピックについての階層型検索意図を読み込む．なお，本研究では音声対話型の検索におけるクエリ推薦の手法として階層型検索意図を取り入れることの有効性を実証することを主な目的としているため，評価実験の検索タスクに用いる階層型検索意図を事前調査や予備実験をもとに手作業で作成したが，デスクトップ検索を対象としたクエリ推薦の手法 [6] [1] [7] では検索エンジンのクエリログやクリックスルーデータをもとにクエリのクラスタリングや相互関係の推定を行っており，同様のアプローチによって階層型検索意図もある程度の精度で自動作成が可能だと考える.

4.3.2　現在の検索意図の推定

3.2 節で述べたクエリ推薦を行うためには，ユーザ発話から現在の検索意図を推定する必要がある．そこで，本研究では各検索意図に対する標準的な回答文（「エナジードリンク 成分」であれば,「エナジードリンクの主な成分にはカフェイン，タウリン，ビタミン B 群，糖類，アミノ酸などが含まれます.」が標準的な回答となる.）を用意し，ユーザ発話に対する回答文とのコサイン類似度が最も高いものを現在の検索意図とした．ただし，最も高いコサイン類似度が 0.70 以下の場合，用意された階層型検索意図の中に現在の検索意図に該当するものはないとして，クエリ推薦も行わないという設計にした.

質問文ではなく回答文を使用した理由は，検索意図の誤った推定を減らすためである．質問文は回答文に比べて文章が短くなりやすく，少しの文章表現の違いや特徴的な単語を含むか否かで埋め込み表現が大きく変動してしまう．それによって検索意図が誤って推定されることが多く，これを回避するために回答文を使用した.

なお，コサイン類似度を測るための文埋め込み表現の取得には cl-nagoya/sup-simcse-ja-base[4]を使用した．このモデル

2：https://developer.amazon.com/ja-JP/alexa/alexa-skills-kit
3：https://platform.openai.com/docs/guides/text-generation/chat-completions-api
4：https://huggingface.co/cl-nagoya/sup-simcse-ja-base

は SimCSE [5] をベースに cl-tohoku/bert-base-japanese-v3[5] を JSNLI データセット [6]でファインチューニングした日本語文埋め込みモデルであり，事前実験でも比較的よく検索意図の推定が行えていたためこのモデルを使用した．

4.3.3 クエリ推薦メッセージの出力

3.2 で述べた方法により最大 2 つの推薦クエリが選ばれ，これを「他には（1 つ目の推薦クエリ）や（2 つ目の推薦クエリ）について調べることができます.」というフォーマットにあてはめることでクエリ推薦メッセージが完成する．例えば，3.2 節で述べたエナジードリンクの検索タスクにおいて，ユーザの現在の検索意図が「成分」であると推定されたとき，クエリ推薦として「他にもエナジードリンクに含まれるカフェインやエナジードリンクの成分について調べることができます.」というクエリ推薦のメッセージが出力される．ただし，ユーザが一度検索したクエリは推薦の候補から除外するため，ある程度検索が進むと選ばれる推薦クエリが 2 つ未満となることがあり，選ばれた推薦クエリが 1 つの場合は「他には（推薦クエリ）について調べることができます.」というフォーマットを使用し，選ばれた推薦クエリがない場合は何のメッセージも出力しないものとした．

メッセージの音声出力には，alexa-remote-control[7]を利用し，コマンドラインを用いて任意のタイミングで出力あるいはスキップの制御が行えるようにした．後述の評価実験では，音声出力のタイミングは最後のシステムの返答が終了してから 10 秒後に設定し，10 秒以内にユーザからの発話があった場合出力をスキップするものとした．

5 評価実験

本節では，本研究で行った評価実験について述べる．まず実験計画について述べ，その後具体的な実施手順について説明する．

5.1 実験計画

本研究では被験者内実験計画を用いた．このとき，実験参加者が 2 つの検索タスクを行う順序については，実験条件およびタスクの両者を考慮した二元配置実験を行い，実験条件およびタスクについての順序効果を最小限に抑えた．ベースラインとなるシステムでは，情報検索システムは 4.2 節で述べたシステムと同じものを，クエリ推薦システムは階層型検索意図を用いず同様のクエリ群から完全ランダムな推薦を行うシステムを用いた．

5.2 実験タスクにおけるトピックとシナリオの作成

本研究では，実世界における情報検索の場面を想定し，2 つのトピックと，それぞれに対して最終的に調べたことをもとに意思決定を行ってもらうためのシナリオを作成した．

以下がそのトピックとシナリオである．

トピック 1: エナジードリンクについて
シナリオ: あなたは日常的にエナジードリンクを飲んでおり，先日友人から「健康に悪いからエナジードリンクは飲まない方がいい.」と言われました．そこで，あなたはエナジードリンクの健康への影響について調べることにしました．検索を行ったうえで，これからもエナジードリンクを飲み続けるかどうかを決定してください．

トピック 2: 自宅での筋力トレーニングについて
シナリオ: あなたは何か新しい趣味をつくりたいと思い，筋力トレーニングを始めることにしました．いきなりジムに通うのはハードルが高いと感じ，まずは自宅でできるものからやってみようと考えています．自宅での筋力トレーニングについて検索を行い，今後やっていくトレーニングの内容を決定してください．

5.3 実験環境

兵庫県立大学情報科学研究棟のセミナー室で実験環境を設営した．実験に用いるスマートスピーカーはテーブルの中央に設置し，実験者が 4.3 節で述べたクエリ推薦システムを操作するためのパソコンは，実験参加者と実験参加者をホワイトボードで仕切ることによって作った疑似的な別室に配置した．なお，実験者はビデオカメラを通じて実験参加者をモニタリングできるようにした．

5.4 実験手順

まず，実験参加者に対して検索タスクの概要，収集するデータ，収集したデータの研究目的への利用などについて説明した．これらについて同意した実験参加者のみが実際の実験タスクへ進み，実験を行った．つぎに，実際の検索タスクに入る前に，これから行う検索タスクへのイメージを深め，使用する検索システムに慣れるために，訓練タスクを行ってもらった．訓練タスクでは,「AI について」をトピックに設定した．訓練タスクでは，クエリに対する回答について，さらに深める質問ができることや，検索結果が意図と全く異なった場合再度同じ内容で検索ができること，推薦の使い方も何度か試してもらうといいこと，推薦が出力された際に必ずしも推薦に従わなくてもよい旨を伝えた．また，実際の検索タスクに入る前に，最大検索時間は 15 分とすること，情報が十分に集まったと感じたらいつでも検索を終了してもよいこと，実験中パソコンやスマートフォンでの検索をしてはいけないこと，タスク終了後は再度検索できないことを説明した．訓練タスク終了後は，実験参加者に 2 つの検索タスクを実施してもらった．各検索タスクは以下の手順で実施した．はじめに，5.2 節で示したトピックとシナリオを実験参加者に提示した．トピックとシナリオについて説明を行った後，1 つ目の検索タスクを行ってもらった．実際に検索タスクを行ってもらった際の対話の一部を表 1 実験参加者が 1 つ目の検索タスクを終了した後，システムに関するアンケートに回答してもらった．2 つ目の検索タスクも同様の手順で実施した後，システムの感想や気になった検索行動などをインタビューで確認し，最後に実験後アンケートで，音声検索をどの

5：https://huggingface.co/cl-tohoku/bert-base-japanese-v3
6：https://huggingface.co/datasets/shunk031/jsnli
7：https://github.com/thorsten-gehrig/alexa-remote-control

表 1　エナジードリンクについての検索タスクでの対話の一部.

発話者	テキスト
ユーザ	エナジードリンクを飲む事のデメリット教えて.
システム	エナジードリンクのデメリットは，過剰な摂取による心臓の負担や睡眠障害，依存性の増加などが挙げられます.
ユーザ	エナジードリンクを飲むことのメリットを教えて.
システム	エナジードリンクのメリットは，疲労回復や集中力の向上，エネルギーの補給などが挙げられます.
システム	他にはエナジードリンクの適切な摂取のタイミングやエナジードリンクの代わりになるアイデアについて調べることができます.
ユーザ	エナジードリンクの代わりになるアイテムについて教えて.
システム	エナジードリンクの代わりになるアイテムとしては，コーヒーや紅茶，ハーブティー，スポーツドリンク，果物などがあります.
システム	他にはエナジードリンクを飲まずに済む健康的な生活習慣やエナジードリンクの製品について調べることができます.
ユーザ	エナジードリンクを飲まなくてもいい生活習慣について教えて.
システム	エナジードリンクを飲まなくてもいい生活習慣としては，十分な睡眠をとること，バランスの取れた食事を摂ること，適度な運動をすることが挙げられます．また，ストレスを適切に管理することも重要です.
ユーザ	エナジードリンクを飲むときの適切な量を教えて.
システム	エナジードリンクを飲むときの適切な量は，1 日に摂取するカフェインの量を 300mg 以下に抑えることが推奨されています．エナジードリンクのラベルにはカフェインの含有量が表示されているので，それを参考にしてください.

程度利用しているか，検索タスクで用いたトピックについてどの程度前提知識を有していたかなどを回答してもらった.

5.5　タスク毎のアンケート

タスク毎のアンケートは，検索した内容をもとに行った意思決定を自由記述で回答してもらう項目と，クエリ推薦が有効だったかを問う項目，システムの使いやすさを評価する System Usability Scale [4] で構成した.

アンケート項目は以下のとおりである.

Q1.　最終的な意思決定を記述してください（タスクにあわせて文言は変更する.）.

Q2.　Q1 の回答の根拠となった情報についてまとめてください.

Q3.　この検索エージェントからの推薦はその場面に適したものだった.

Q4.　この検索エージェントからの推薦は意思決定に役立った.

Q5.　この検索エージェントからの推薦は様々な情報を獲得するのに役立った.

Q6.　この検索エージェントとの対話は自然だと感じた.

Q7.　この検索エージェントを頻繁に使いたいと思う.

Q8.　この検索エージェントが必要以上に複雑に感じた.

Q9.　この検索エージェントはシンプルで使いやすいと思った.

Q10.　この検索エージェントを使うために，技術的なサポートが必要だ.

Q11.　この検索エージェントは様々な機能がうまく統合されて動作していると感じる.

Q12.　この検索エージェントは応答に一貫性がないと思った.

Q13.　たいていの人はこの検索エージェントの使い方をすぐに学ぶことができると思う.

Q14.　この検索エージェントの使い勝手が悪いと感じた.

Q15.　この検索エージェントを自信をもって使えると思った.

Q16.　この検索エージェントを使い始める前に，多くのことを学ぶ必要があると感じた.

Q1，Q2 を除く 14 項目について，5 段階のリッカート尺度（1: まったくあてはまらない，2: どちらかというとあてはまらない，3: どちらともいえない，4: どちらかというとあてはまる，5: とてもよくあてはまる）を設定して回答してもらった.

5.6　実験参加者

兵庫県立大学の学生から計 10 名（うち男性 8 名，女性 2 名）が実験に参加した．実験の時間は最大で 90 分間とし，実験を最後まで完了した実験参加者には学部生の場合 1,524 円，院生の場合 1,725 円を支払った．なお，うち 2 名の実験ではシステムによる推薦メッセージが提示されることなく検索を終了したため，分析には他 8 名のデータのみを用いた.

6　分析結果と考察

本節では，提案手法による推薦の有効性について，検索状況に沿った推薦がなされるか，ユーザの多様な情報獲得を支援するか，ユーザの検索行動にどのような影響を与えるか，ユーザビリティの高さのそれぞれの側面から分析を行った結果とその考察について述べる.

6.1　検索状況に沿った推薦がなされるか

提案手法では，階層構造を用いることでより検索状況に沿った推薦がなされるという仮定のもと分析を行った．関連するアンケート項目 Q3，Q6 の結果を表 2 に示す．2 つの項目についてウィルコクソンの順位和検定を行った結果，ともに有意水準 5%で有意差は認められなかった．この要因としては，ベー

表 2　検索状況に沿った推薦がなされるかを問うアンケート結果の平均値．表中の括弧は標準偏差を指す．

アンケート項目	ベースライン	提案手法
Q3 この検索エージェントからの推薦はその場面に適したものだった．	4.25(0.71)	4.25(0.71)
Q6 この検索エージェントとの対話は自然だと感じた．	3.75(0.71)	4.00(0.93)

表 3　ユーザの多様な情報獲得を支援するかを問うアンケート結果の平均値．表中の括弧は標準偏差を指す．

アンケート項目	ベースライン	提案手法
Q4 この検索エージェントからの推薦は意思決定に役立った．	4.63(0.52)	4.63(0.52)
Q5 この検索エージェントからの推薦は様々な情報を獲得するのに役立った．	4.75(0.46)	4.63(0.52)

表 4　実験の検索タスクにおける検索時間，検索回数，推薦回数，推薦採用率，推薦依存率の平均値．表中の括弧は標準偏差を指す．

評価指標	ベースライン	提案手法
検索時間 [s]	741.88(224.70)	696.00(211.49)
検索回数	17.38(6.37)	16.25(6.86)
推薦回数	6.00(3.12)	5.38(3.29)
推薦採用率	0.63(0.27)	0.59(0.25)
推薦依存率	0.20(0.10)	0.20(0.12)

スラインであっても検索トピックに関連するクエリ候補から選択される以上検索状況から大きく外れた推薦にはならなかったことや，提案手法であっても現在の検索意図が正しく推定できなかったために理想的な推薦ができなかったり，一度検索した内容が推薦されてしまったりしたことが考えられる．実際に実験後のインタビューでも推薦の違いに明確に気付いた参加者はおらず，提案手法のシステムに対して「何度も同じ推薦があった.」「すでに検索した内容が推薦されていた.」といった意見もあった．

しかしながら，提案手法のシステムに対して Q3 の平均値は 4.25，Q6 の平均値は 4.00 と高い評価を得ており，「（提案手法のシステムのほうが）流れとして自然な感じがした.」という仮定を後押しする意見もあるなどポジティブな結果も得られた．

6.2　ユーザの多様な情報獲得を支援するか

提案手法では，特化と平行移動をもとにした推薦によって探索的検索が促され，ユーザがより多様な情報を獲得できるという仮定のもと分析を行った．関連するアンケート項目 Q4，Q5 の結果を表 3 に示す．こちらの 2 項目についてもウィルコクソンの順位和検定を行ったが，ともに有意水準 5%で有意差は認められなかった．この結果から，推薦が構造化されていることはユーザの多様な情報獲得に直結するとは限らないと言える．今回の実験では推薦を行わないシステムを使用しなかったため定量的な分析はできないが，アンケート結果の値を見るに，音声対話型検索においても階層構造を用いる用いないに関わらずクエリ推薦が多様な情報獲得，ひいては検索後の意思決定に少なからず貢献していると考えられる．

表 5　実験の検索タスクにおける検索時間，検索回数，推薦回数，推薦採用率，推薦依存率とユーザの検索トピックに対する前提知識との相関係数．

評価指標	検索時間	検索回数	推薦回数	推薦採用率	推薦依存率
前提知識との相関係数	-0.24	-0.23	0.04	-0.15	0.22

6.3　ユーザの検索行動に与える影響

次に，手法の違いがユーザの検索行動に与える影響を確認するために，実験の検索タスクにおける検索時間，検索回数，推薦回数，推薦採用率，推薦依存率について分析を行った．なお，推薦採用率とは検索タスク中にシステムから推薦されたクエリのうちユーザが推薦に従って検索を行った割合を指し，推薦依存率とは検索タスク中に行った検索のうち推薦に従って検索を行った割合を指す．各評価指標の平均値を表 4 に示す．これらについてウェルチの t 検定を行ったが，どの評価指標についても有意差のある結果は得られなかった．この要因としては，サンプルサイズが小さいこと，実験参加者ごとの個人差が大きいことが考えられる．特に推薦採用率，推薦依存率については，それぞれ分母となる検索回数，推薦回数が少ないと値の変動が大きくなってしまうことも要因として考えられる．

実験後のインタビューでは「検索トピックについてわりと知っていたので，あまり検索をしなかった.」「検索トピックについてあまり知らないためどう調べていいか分からず，推薦を利用してみようと思った.」といったコメントがあり，これらの指標には検索トピックに対する前提知識が大きく影響していることが考えられたため，前提知識と各評価指標の相関についても分析を行った．前提知識については，ユーザ属性を確認するための実験後アンケートに「各検索トピックについて現在（検索後）の知識を 100 としたとき，実験参加前にはどの程度知識がありましたか．数値で記入してください.」という項目を設けており，これを使用した．各評価指標と検索トピックに対する前提知識との相関係数を求めた結果を 5 に示す．しかしどの評価指標についても相関はほとんど見られなかった．この要因としてはサンプルサイズが小さいこと，前提知識があくまで相対的な値であることが考えられるが，検索トピックに対する前提知識という抽象的な指標を定量的に測定することは困難である．

6.4 ユーザビリティ

最後に，SUS を用いてシステム間のユーザビリティに関して比較を行った．SUS のアンケート項目である Q7 から Q16 について，肯定的な質問である Q7，Q9，Q11，Q13，Q15 の各値から 1 を引き，否定的な質問である Q8，Q10，Q12，Q14，Q16 の各値を 5 から引く．これらの総計に 2.5 を乗じたものが SUS のスコアである．その結果，両システムともに SUS スコアの平均値は 74.1 と一致し，少なくとも今回の実験では提案手法とベースラインによる推薦の違いはユーザビリティに影響を与えなかった．この SUS スコアは一般的に 80 点以上が非常に優れたユーザビリティを提供するシステムであるとされており，どちらのシステムも改善の余地がある．

7 議　　論

本節では，6 節の分析から得られた知見をもとに，音声対話型検索におけるクエリ推薦の手法について議論する．

今回の実験では，提案手法によるクエリ推薦が対話としての自然さを保ちつつユーザの検索行動を支援するという結果は得られたものの，ベースラインに対する有意性を実証するには至らなかった．この一番の要因としては，各手法におけるクエリ推薦の内容の違いが非常に細微なものであったことが考えられる．これを避けるために，検索タスクをより多くの検索意図が想定されるような抽象度の高いものに設定し，ベースラインによるランダムな推薦を行った際に提案手法との違いが明確になるようにする必要がある．その他にも，今回のシステムでは再現性を担保するために，現在の検索意図を 4.3.2 節で述べたように回答文の埋め込み表現から推定したが，この推定が正しく行われていないと思われる場面も多くあり，提案手法による推薦がうまく行えないことがあったことも要因として考えられる．そのため，回答文だけでなく質問文に対しても同様の比較を行い，適切な重みづけによって想定される検索意図との近さをスコア化するなど，推定の精度を高めるための工夫が必要だと考える．

また，本研究では検索意図の相互関係を 2 層の階層構造を用いて表現することでクエリ推薦を行ったが，検索意図の粒度をさらに細かくし，3 層の階層構造にすることで，より厳密に検索状況に沿う推薦が行えると考える．階層構造以外では，グラフ構造を用いたクエリ推薦についての研究もあり [2] [10]，音声対話型検索におけるアプローチとして有効なのか，他にもクエリ推薦を行ううえで使える構造があるかなどについても検証する価値があると考える．

8 まとめと今後の課題

本研究では音声対話型検索を想定してユーザの多様な情報獲得を支援するためのクエリ推薦の手法を提案した．具体的には，検索トピックに対して想定されるユーザの検索意図を化と平行移動の 2 つの主要なクエリ修正パターンに基づいた階層構造で表現し，この構造を用いて推薦するクエリを絞り込むことで，音声対話型検索に適応したクエリ推薦システムを構築した．さらに，提案手法によるクエリ推薦の有効性を検証するために，音声対話型検索システムを用いたユーザ実験を行った．そして，この実験により得られたログデータとアンケート結果について分析を行ったが，統計的に有意な結果は得られなかった．この結果を受け，実験でのタスク設定やシステムに改善の余地があることが分かった．一方で，実験参加者によるフィードバックから，音声対話の流れの中での推薦の自然さや探索的検索の支援として提案手法に対するポジティブな結果も得られ，提案手法によるクエリ推薦がユーザ体験を高めることが示唆された．

本研究の結果を踏まえ，検索タスクをより抽象的にすることで手法による推薦の違いを明確にする，提案手法に沿った推薦がなされるように現在の検索意図の推定精度を高めるなどの見直しを行い，より多くのユーザによる評価実験を行うことで，提案手法の有効性を検証する必要がある．また，本研究で扱った構造以外の構造を用いたクエリ推薦についても，音声対話型検索において有効であるか検証する価値があると考える．

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

文　　献

[1] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo” Mendoza. Query recommendation using query logs in search engines. In *Proceedings of the Current Trends in Database Technology - EDBT 2004 Workshops*, pp. 588–596, 2005.

[2] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, and Sebastiano Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 Workshop on Web Search Click Data*, pp. 56–63, 2009.

[3] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. From “dango” to “japanese cakes”: Query reformulation models and patterns. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 183–190, 2009.

[4] John Brooke. SUS: A quick and dirty usability scale. *Usability evaluation in industry*, Vol. 189, No. 3, pp. 189–194, 1995.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

[6] Makoto P. Kato, Tetsuya Sakai, and Katsumi Tanaka. Structured query suggestion for specialization and parallel movement: effect on search behaviors. In *Proceedings of the 21st international conference on World Wide Web*, pp. 389–398, 2012.

[7] Hao Ma, Haixuan Yang, Irwin King, and Michael R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 709–718, 2008.

[8] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, Vol. 49, No. 4, pp. 41–46, 2006.

[9] Alan Medlar, Jing Li, and Dorota Głowacka. Query suggestions as summarization in exploratory search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 119–128, 2021.

[10] Enrico Palumbo, Andreas Damianou, Alice Wang, Alva Liu,

Ghazal Fazelnia, Francesco Fabbri, Rui Ferreira, Fabrizio Silvestri, Hugues Bouchard, Claudia Hauff, Mounia Lalmas, Ben Carterette, Praveen Chandar, and David Nyhan. Graph learning for exploratory query suggestions in an instant search system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4780–4786, 2023.

[11] 吉野幸一郎, 森信介, 河原達也. 述語項の類似度に基づく情報抽出・推薦を行う音声対話システム. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3386–3397, 2011.

[12] 佐藤進也. 階層構造を持つタグを利用した探索的検索のためのユーザインタフェース. 第 84 回全国大会講演論文集, Vol. 2022, No. 1, pp. 451–452, 2022.

[13] 翠輝久, 河原達也. 文書検索タスクにおける検索要求明確化のための対話戦略. 情報処理学会研究報告音声言語情報処理 (SLP), Vol. 2005, No. 12 (2004-SLP-055), pp. 83–88, 2005.

[14] 翠輝久, 水上悦雄, 柏岡秀紀. 音声対話による意思決定支援システムの実験的評価. 電子情報通信学会論文誌 D, Vol. 96, No. 10, pp. 2518–2529, 2013.

[15] 中野佑哉, 河野誠也, 吉野幸一郎, 須藤克仁, 中村哲. 問い返し質問文生成によって曖昧性解消を行う質問応答システム. 言語処理学会第 27 回年次大会 発表論文集, P7-17, 2021.

A Token-Free Approach to Entity-based Keyword Search in Classical Chinese

Zhongqing JIANG†, Zengqing WU†,††, Chuan XIAO†,†††, Yuya SASAKI†, and Makoto ONIZUKA†

† Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka, 565–0871 Japan
†† Graduate School of Informatics, Kyoto University
36–1 Yoshida–Honmachi, Sakyo–ku, Kyoto, 606–8501 JAPAN
††† Graduate School of Informatics, Nagoya University
Furo–cho, Chikusa–ku, Nagoya, 464–0814 Japan
E-mail: †{jiang.zhongqing,chuanx,sasaki,onizuka}@ist.osaka-u.ac.jp, ††wuzengqing@outlook.com

Abstract Classical Chinese texts present challenges for natural language processing due to their inherently unpunctuated, frequent use of abbreviated personal references, potential existence of multiple aliases, issues of textual omissions and errors in the digital transcription process, and their low-resource status. These factors contribute to difficulties in enhancing the accuracy of searches within these texts. This work, leveraging the ByT5 model suitable for token-free languages and employing cross-linguistic transfer learning, has developed an Entity-based Keyword Search method. By systematically improving the performance of intermediate tasks such as Named Entity Recognition (NER) and Relation Extraction (RE) in Classical Chinese texts, as well as constructing datasets for missing data, our model effectively identifies queries that may refer to the same entity with multiple nouns. We evaluated our model on extensively used Classical Chinese datasets for the NER task and compared it with existing mainstream baselines, achieving superior performance results. Further, we innovatively designed datasets and tasks for RE in Classical Chinese. The results demonstrate the significance of integrating information based on NER and RE in constructing search systems. This work addresses the unique linguistic features of Classical Chinese and offers a more accurate method for textual analysis and information retrieval in historical research contexts.

Key words Classical Chinese, Information Retrieval, Named Entity Recognition, Relation Extraction, Token-Free

1 Introduction

Classical Chinese texts, using Classical Chinese script as the medium of writing, serve as a vital source for historical research. Academic institutions are actively working to digitize and database these texts to support researchers in conducting research. A common search requirement is the identification of named entities to locate significant information about specific historical periods, individuals, or events. This function is typically achieved through manual annotation and Named Entity Recognition (NER). However, the brevity of Classical Chinese texts and their tendency to use single characters to represent entities (e.g., using "Liang (亮)" for "Zhuge Liang (諸葛亮)", where the former is an abbreviation of the latter) make the accurate identification and extraction of entities and their relationships particularly challenging. Moreover, due to the advanced language knowledge required for manually annotating Classical Chinese texts, there is a scarcity of high-quality, manually annotated Classical Chinese texts, indicating a shortage of labeled data for these texts.

In previous studies addressing the NLP issue for Classical Chinese, a low-resource language, we proposed a ByT5-based approach [5]. The token-free characteristic of the ByT5 model effectively addresses the absence of spaces and punctuation as sentence separators in Classical Chinese text systems. This avoids the negative impact of limited accuracy in tokenizing Classical Chinese texts on final outcomes. We utilized transfer learning techniques, training the model on no labeled Classical Chinese texts and then transferring it to down stream task. This approach mitigates the need for extensive manual annotation in training models for low-resource languages and addresses the issue of limited corpora.

In this research, to systematically enhance the interpretation of Classical Chinese and its eventual application in user-friendly information retrieval systems, we further expand this methodology. This involves exploring three aspects: dataset creation, model training, and practical adjustments and in-

tegration for search systems, to enhance the model's final performance. Our investigations revealed that, despite the existence of a few Classical Chinese text NER datasets (such as the commonly used C-CLUE [4]), datasets for intermediate tasks crucial for search systems, like Relation Extraction (RE), are lacking. Hence, we attempted to construct datasets using entity alignment and data generation methods based on large language models. Subsequently, we further trained models for NER and RE in Classical Chinese to improve accuracy. Finally, considering the potential queries of users in search systems, we organized common queries and designed query rules based on NER and RE results to yield more user-friendly search outcomes.

Our key contributions are as follows:

（1） Creation of datasets for training and testing NER and RE in Classical Chinese.

（2） Development of the ccByT5 model, which is a pre-trained language model based on ByT5, trains on Classical Chinese texts. Performed NER and RE fine-tuning on ccByT5.

（3） Design of information integration rules based on NER and RE, tailored for search system applications, to enhance the accuracy in processing queries.

2 Related Work

2.1 Characteristics and Progress in Classical Chinese Language Research

In our introduction, we discussed the distinctive features of Classical Chinese, including the prevalence of abbreviations and nominal proxies, its inherently unpunctuated, the scarcity of annotated corpora, and the general rarity of comprehensive text collections. Addressing these challenges, existing work has developed Classical Chinese word segmentation and Named Entity Recognition NER models based on Long Short-Term Memory (LSTM) networks [10]. With the advent of large language models like BERT, and incorporating methods such as transfer learning, models like RoBERTa have been introduced and achieved better results in NER tasks [6]. However, the performance of these models in Classical Chinese NER tasks remains relatively inferior compared to other modern languages like English, Japanese, and Chinese. In prior research, we have combined the strengths of these existing methods, aligning named entities in Modern and Classical Chinese corpora for transfer learning [5]. We found that token-free models like ByT5 align well with the characteristics of Classical Chinese and outperform LSTM and BERT-based models. Additionally, transfer learning significantly mitigates the negative impact of low-resource challenges on model training in Classical Chinese, effectively addressing the issue of limited dataset sizes in the target language [7]. In this study, we also employed this transfer learning approach to train pre-existing models.

2.2 Current Work on Classical Chinese Datasets

Beyond the NER task, our survey found no suitable datasets for the RE task, despite the importance of both NER and RE in the implementation of information retrieval systems. The C-CLUE dataset, while more substantial in sample size for the NER task, lacks a counterpart for Classical Chinese RE. Therefore, it is necessary to construct a new dataset for NER and RE in Classical Chinese to enhance the effectiveness of model training. This will contribute significantly to advancing the field of Classical Chinese language processing, addressing the gaps in current research methodologies and resources.

Another issue with Classical Chinese is that all computationally processed materials rely on manual entry or Optical Character Recognition (OCR) processing, which introduces potential issues with character recognition. In the Chinese script, there are many characters with similar shapes but significantly different meanings. This poses a considerable challenge for model recognition, as their embeddings are distinctly different. Existing research addressing these inherent textual errors is limited. This paper aims to establish a new baseline for this problem, assessing the practical effectiveness of models in handling the complexities of character recognition errors in Classical Chinese texts. This approach not only contributes to the refinement of OCR and manual data entry methods but also enhances the reliability and accuracy of computational processing in the context of Classical Chinese language resources.

3 Methodology

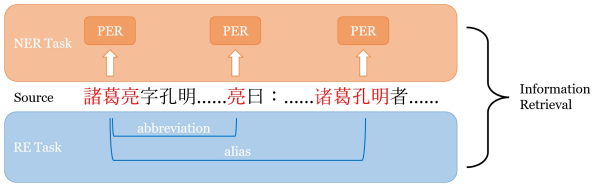


Figure 1 Task Overview.

Figure 1 illustrates the primary objectives of this work. Our aim is to accurately extract named entities from Classical Chinese texts and to extract the relationships between them. Upon completing these two intermediate tasks, we will further undertake rule-based information integration to facilitate efficient retrieval. This approach encapsulates the essence of our methodology, which seeks to enhance the processing and understanding of Classical Chinese texts for advanced research applications.

3.1 Dataset Design

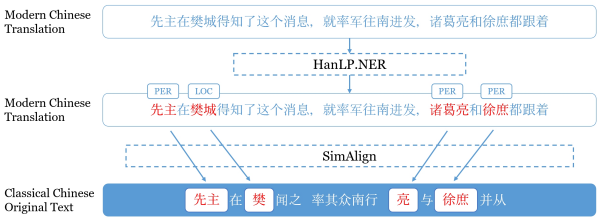


Figure 2 Approaches to the Preparation of Databases.

NER and RE datasets are constructed to enhance our model's training set. For the generation of the NER dataset, we employed an entity alignment method. Initially, we performed the NER task on the modern Chinese translations of existing Classical Chinese texts, thereby labeling the modern Chinese translations with NER tags. We used the NER model from HanLP [3], known for its leading performance in modern Chinese NER tasks, for labeling. Subsequently, we applied the SimAlign method [8], a word alignment approach that requires no parallel data, to align entities in the modern Chinese translations with those in the original Classical Chinese texts. This process generated a vast amount of training data for dataset augmentation, as depicted in Figure 2. As Table 1 illustrates, this method significantly increased the number of available training and validation sets.

Table 1 Size of NER Task Dataset.

Dataset	Train	Validation	Test
C-CLUE	1900	238	238
New NER Dataset	**142822**	**10000**	-

The generation of the RE dataset primarily relied on the GPT-3.5 model. Due to the lack of suitable labeled RE datasets in Classical Chinese, we fed the GPT-3.5 with Classical Chinese texts lacking RE labels, along with modern Chinese knowledge encompassing the relationships involved in the texts. This prompt enabled GPT-3.5 to generate labeled Classical Chinese texts for RE. This innovative approach addresses the scarcity of labeled datasets in Classical Chinese RE and leverages advanced language models to create a robust training foundation for our research.

3.2 Model Architecture: ccByT5

Model Pre-training. We employed ByT5 for the pre-training of Classical Chinese corpora. ByT5 is a variant of the T5 model that processes strings using UTF-8 encoding without tokenization. Its advantage in Classical Chinese lies in its ability to handle texts with a high frequency of variants and rare characters. Unlike other models, ByT5 does not mark rare vocabulary as [UNK] (unknown), thereby reducing the likelihood of model performance deterioration due to an abundance of [UNK] tokens. The architecture of ByT5 is depicted in Figure 3.

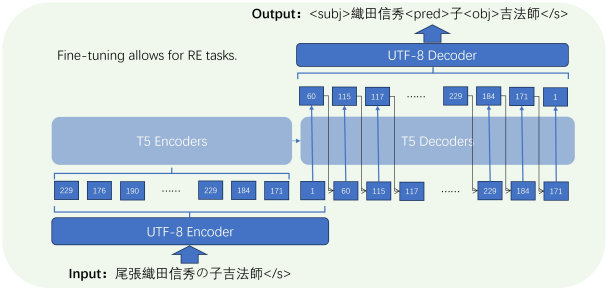


Figure 3 Architecture of ByT5.

Considering the size of the dataset, our model uses the same confguration as ByT5-small. Experiments were conducted on Google Cloud TPU v3-8. We used the T5X framework for pre-training, after that we converted the T5X checkpoints into PyTorch models and performed NER fine-tuning using the PyTorch-Lighting framework. The fine-tuning involved a sequence length of 1024, a batch size of 2^{14} tokens, a base learning rate of 0.0003, a dropout rate of 0.1, and utilized A100 GPUs with early-stopping settings. In all experiments, the same network layers were used with consistent parameters and experimental environments.

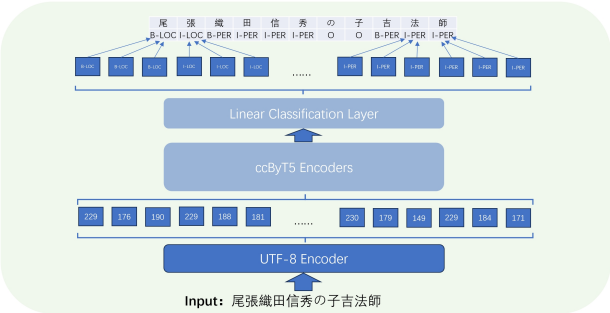


Figure 4 Model Architecture of ccByT5.

Encoder-Only ccByT5 NER Model. However, the ccByT5 model, centered around the seq2seq architecture, also has its drawbacks. In NER tasks, since the identified entities may consist of multiple characters, the positional information of each character within the entity name is not fully recorded. Typically, entities are only categorized as B-PER and I-PER, representing the beginning and intermediate of the entity, respectively. As a result, the output post-training may exhibit slight discrepancies with the target (such as an additional I-PER being recognized), leading to inconsistencies in the length of the target and output, which significantly complicates the implementation of entity alignment.

In response to this challenge, this study proposes the removal of the decoder layer during the NER fine-tuning process, replacing it with a linear classification layer to specialize in the token classification task. Through classification, we map the results produced by the ccByT5 Encoder layers onto the labels of each character, as depicted in Figure 4.

RE Model. The RE fine-tuning in- volved a sequence length of 512, a batch size of 2^{13} tokens, a learning rate of 0.003, a dropout rate of 0.1, and the decoder layer is retained during fine-tuning, i.e., the same model as ByT5 is used for training.

3.3 Information Integration Rule Design for Information Retrieval

To ensure efficient information retrieval, we have developed several integration rules based on NER and RE These rules are as follows:

（1） Perform NER and RE on classical Chinese texts, retaining the positional information of each extracted entity within the text.

（2） Separate the surname and given name of person entities, using the given name as the basis for entity identification. Create a new node for alias relationships, unless an existing node for the alias already exists, and link the alias to the respective entity.

（3） Generate a node with a surname from a node without a surname, and merge any existing nodes.

（4） Nodes in alias relationships can share positional information within the text.

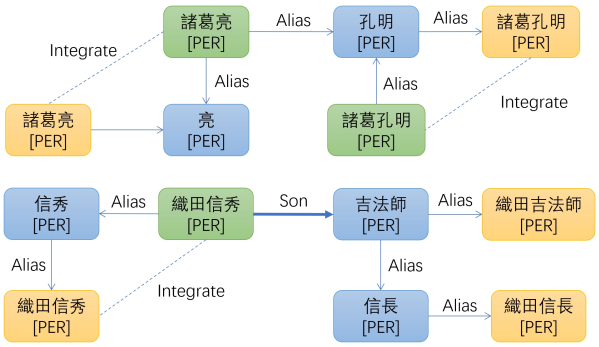


Figure 5 An Example of Rule-Based Integration.

Based on these rules, we can construct a knowledge graph for entities and relationships, facilitating user queries. Figure 5 shows an example of rule-based integration (rule 3).

4 Experiments

4.1 Dataset

For NER task, we utilized the C-CLUE dataset and a dataset designed specifically for this research as our training and testing sets. The distribution of the data is presented in Table 1. For RE task, we utilized the GPT 3.5 generated dataset as our training sets and the Hand-labeled dataset as the testing sets. For Information Retrieval task, we tested the performance of our approach to detect these abbreviations and aliases using the position of the 12 characters' abbreviations and aliases that appear in three histories.

4.2 Evaluation Settings

To mitigate endogeneity issues, we independently evaluated model performance on two separate test sets from the C-CLUE and the dataset designed for this study.

Furthermore, to assess the robustness of the model in the presence of input errors, we established an OCR Task. This task employs a similarity algorithm in both training and testing data to replace certain characters with visually or phonetically similar ones, thus introducing noise to simulate errors typical in OCR inputs.

4.3 Baselines

For NER task, we employed two fundamental models, BERT-Base [1] and Lattice-LSTM [10], as our benchmarks for testing. These models were compared against the currently leading Siku-BERT model [2], [9] to evaluate their performance. For RE task, we employed GPT3.5 as our benchmarks for testing.

5 Results

5.1 NER Task

Tables 2 and 3 respectively display the performance of the model in NER tasks. The results indicate that the ccByT5 model outperforms the current baseline model and on the same level as Siku-BERT. Notably, ccByT5 demonstrate enhanced performance on the new dataset.

Table 2 NER Experimental Results on C-CLUE Dataset. Both Siku-BERT and ccByT5's training data are from C-CLUE.

Method	P	R	F1
BERT-base	29.82	35.59	32.12
Lattice-LSTM	51.96	56.92	54.33
Siku-BERT	59.69	**75.72**	**66.76**
ccByT5	**60.91**	72.91	66.37

Table 3 NER Experimental Results on New training Dataset. Test data from C-CLUE.

Method	P	R	F1
ccByT5	62.83	74.18	68.03

5.2 RE Task

Table 4 respectively display the performance of the model in RE tasks.Similar to GPT-3.5, ccByT5 exhibits a lower precision compared to recall. This indicates that, in addition to correct relation triplets, a significant number of incorrect relation triplets are also being produced. The output from

ccByT5 is characterized by disorganized formats and the generation of pairs that do not constitute a relation, essentially producing noise-like relation triplets. Moreover, the model tends to generate a larger number of relation triplets aggressively. This tendency is attributed to the selection process of texts for training data, where examples were generated in abundance, intentionally choosing texts that contain a large number of relations. Consequently, this approach resulted in a scarcity of negative examples within the training data, leading to the inference that ccByT5's learning was insufficient.

Table 4 RE Experimental Results on New Dataset.

Method	P	R	F1
ccByT5	31.83	72.45	44.23
GPT3.5	**63.35**	**89.77**	**74.28**

5.3 OCR Task

As shown in table 5, the OCR task demonstrates the robustness of the ccByT5 model under noisy conditions. When both the training and testing sets are contaminated with noise, the degradation in performance of ccByT5 across various metrics is limited compared to the baseline models. This reflects the model's reliability in real-world datasets and potential information retrieval tasks involving OCR-based texts.

Table 5 OCR Experimental Results on NER dataset with Noise.

Settings	Method	P	R	F1
Clean	ccByT5	70.15	71.36	70.75
	Siku-BERT	71.99	74.01	72.99
Noise (OCR)	ccByT5	**-11.41**	**-5.07**	**-8.46**
	Siku-BERT	-12.11	-7.07	-9.78

5.4 Information Retrieval Task

As shown in table 6, In the experimental results shown in Table 6.1, out of 12 queries, 3 queries resulted in no detections, leading to a precision, recall, and F1 score of all 0. This is believed to be due to omissions in the relation extraction task. For the remaining queries, the precision was 100 for all, and the recall was mostly above 90. From these results, it can be understood that the proposed method has the capability to detect abbreviations and aliases of personal names contained within Classical Chinese texts.

6 Ablation Study

To analyze the impact of different components of the model on its effectiveness, we conducted several ablation experiments. Specifically, we substituted the training set from the new dataset with the C-CLUE dataset, and compared the ccByT5 model without the Decoder layer to the model with the Decoder, i.e., the origin ByT5 framework. The results, as shown in Table 7, clearly indicate that the construction and utilization of the new dataset, as well as the removal of the Decoder layer during fine-tuning, significantly affect the model's performance. This underscores the effectiveness of these two approaches.

Table 6 Experimental Results of Information Retrieval Task

Histories	**Characters**	**P**	**R**	**F1**
諸葛亮伝	諸葛亮	100	90.10	94.79
	劉琮	0	0	0
	孫権	100	100	100
	諸葛瞻	100	90	94.74
劉表伝	劉表	100	92.52	96.12
	劉琦	0	0	0
袁紹伝	袁紹	100	93.42	96.59
	袁譚	0	0	0
	何進	100	40	57.14
	段珪	100	100	100
	韓馥	100	92.31	96.00
	董卓	100	100	100
	AVG.	75	66.53	69.62

Table 7 Results of Ablation Experiments on NER Task.

Method	P	R	F1
ccByT5 [Encoder-Only ccByT5 + New Dataset]	62.83	74.18	68.03
–New Dataset [Encoder-Only ccByT5]	60.91	72.91	66.37
–New Dataset + Decoder [ccByT5]	54.74	61.04	57.71

7 Conclusion

This study introduces the ccByT5 model, which achieves superior results in classical Chinese NER and RE tasks compared to baseline models. Moreover, it maintains robust performance under noisy conditions such as errors in digital text entry. Additionally, the dataset developed through entity alignment between modern and classical Chinese corpora addresses the insufficiency of existing datasets for classical Chinese NLP tasks and enhances model training performance. These findings highlight the significance of employing token-free models, entity alignment techniques, and cross-linguistic transfer learning in classical Chinese NLP research. Future research could further develop datasets for RE tasks and improve the model's performance in RE, integrating the results of NER and RE to design a functional classical Chinese information retrieval system.

8 Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22H03903. We thank Dr. Ruoming Pang, Mr. Run Peng, Prof. Keishi Tajima, and Prof. Hiroki Nomiya for their instructive suggestions.

References

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

[2] Sijia Ge. Integration of named entity recognition and sentence segmentation on ancient chinese based on siku-bert. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 167–173, 2022.

[3] Han He. Hanlp: Han language processing. *URL: https://github.com/hankcs/HanLP*, 2014.

[4] Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. C-clue: a benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 295–301. Springer, 2021.

[5] Zhongqing Jiang, Zengqing Wu, and Chuan Xiao. Token-free cross-lingual named entity recognition for classical chinese. 2023.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] Yongjie Qi, Hongchao Ma, Lulu Shi, Hongying Zan, and Qinglei Zhou. Adversarial transfer for classical chinese ner with translation word segmentation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 298–310. Springer, 2022.

[8] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*, 2020.

[9] Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng LIU, Haotian HU, Si SHEN, and Bin LI. Construction and application of pre-trained models of siku quanshu in orientation to digital humanities [j]. *Library Tribune*, 42(06):31–43, 2022.

[10] Yue Zhang and Jie Yang. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*, 2018.

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

2024年2月28日(水) 13:30 ~ 15:40 | T4-A オンライン（Zoom Events)

音楽

座長:佃 洸摂(産業技術総合研究所)
コメンテータ:富井 尚志(横浜国立大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T4-A-2-01] 料理レシピと音楽の融合によるレシピ閲覧支援の提案

*浦川 航貴[1]、宮本 拓実[1]、王 元元[1] (1. 山口大学)

13:50 ~ 14:05
[T4-A-2-02] 最適輸送理論を用いたメロディー楽譜からの作曲家推定

*八木 皓平[1]、中村 篤祥[1] (1. 北海道大学　工学部　情報エレクトロニクス学科　情報理工学コース　アルゴリズム研究室、2. 北海道大学 情報科学研究院 情報理工学部門 知識ソフトウェア科学分野)

14:05 ~ 14:20
[T4-A-2-03] 教師なし学習による主成分分析を用いた流行楽曲の特徴量の相関

*小俵 亜紀[1]、木津 美咲[1]、内田 瑞貴[1]、原 香奈江[1]、齋藤 優莉[1]、金井 友希美[1] (1. 奈良女子大学)

14:20 ~ 14:45
[T4-A-2-04] 人体キーポイントを用いて映像に整合する楽器音を生成するモデルの提案

*岡野 日翔[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

14:45 ~ 15:10
[T4-A-2-05] 楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式

*佐伯 綾亮[1]、岡田 龍太郎[1]、峰松 彩子[1]、中西 崇文[1] (1. 武蔵野大学データサイエンス学部)

料理レシピと音楽の融合によるレシピ閲覧支援の提案

浦川　航貴†　　宮本　拓実††　　王　　元元†,††

† 山口大学工学部知能情報工学科　〒 755–8611 山口県宇部市常盤台 2-16-1
†† 山口大学大学院創成科学研究科　〒 755–8611 山口県宇部市常盤台 2-16-1
E-mail: †{a016ffu, d109vgu, y.wang}@yamaguchi-u.ac.jp

あらまし　近年，料理レシピサイトの普及により，多くの人が手軽にレシピ情報を入手できるようになった．一般的に，これらのサイトは画像や動画を活用して料理手順を説明しているが，多くのレシピサイトは単調で，料理する人が楽しみながら利用することが難しいものが多い．本研究では，各レシピに合った音楽を推薦することにより，充実した体験を提供することが可能となる．本論文では，楽天レシピのデータから紹介文，レシピ画像の彩度および明度を抽出する．さらに，Spotify API を用いて Spotify から無作為に選ばれた音楽の感情値（valance），聴覚的な強さ（energy），テンポ（tempo）を特徴量として取得する．これらの特徴量を比較し，コサイン類似度を算出することで，特定のレシピの雰囲気に合った音楽を対応付け，レシピと音楽を融合した閲覧支援システムを構築する．これにより，ユーザは料理レシピを効果的に閲覧し，料理のイメージをより容易に把握することが期待される．最後に，料理レシピと音楽との融合の有用性を被験者による評価実験で検証する．

キーワード　料理レシピサイト，音楽推薦，楽天レシピ，Spotify，レシピ閲覧支援

1　はじめに

近年，楽天レシピ [1]やクックパッド [2]，クラシル [3]などのユーザ投稿型料理レシピサイトが普及している．食品会社マルハニチロ「料理レシピに関する調査 2020（全国 20〜59 歳の男女 1,000 名を対象）」によると，“どのような料理レシピを使っているか” という質問に対して “レシピサイト・レシピアプリ” を使っていると答えた割合が全体の 74.3%であり，これが最も多かった（図 1 参照）．さらに，世代を問わず，すべての年代でこの回答が最も多いと結果が出ている [1].

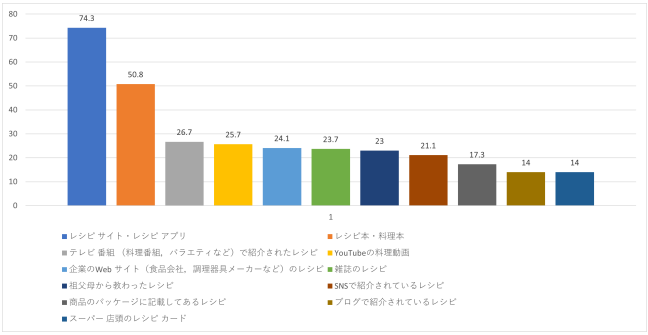


図 1　“どのような料理レシピを使っているか” の調査結果

料理レシピサイトは，一般的にタイトル，紹介文，材料，コツ，作り方，レビュー，写真，動画などから構成されているものが多い．また，料理レシピ動画サービスのクラシルでは，利用者が料理を投稿する際に自分で料理に最適な音楽を選択しなければならないため，どの音楽が料理に適しているかを選ぶことが難しい場合もある．そこで本研究では，料理レシピサイトの情報を活用し，特定の料理の雰囲気に合った音楽を推薦し，サイト内で再生することで，利用者が楽しみながら料理することができるとともに，レシピ投稿者が音楽を簡単に選択しやすくすることを目指す．具体的には,「楽天公開データ」[4]より，楽天レシピデータ [2] から紹介文，レシピ画像の彩度および明度を抽出する．さらに，Spotify API[5]を用いて Spotify から無作為に選ばれた音楽の感情値（valance），聴覚的な強さ（energy），テンポ（tempo）を特徴量として取得する．これらの特徴量を比較し，コサイン類似度を算出することで，特定のレシピの雰囲気に合った音楽を対応付けし，レシピと音楽を融合した閲覧支援システムを構築する．このシステムにより，各レシピに適した音楽を推薦することができ，より良い体験を提供することが可能となる．

本論文の構成は以下のとおりである．2 章では，画像を利用した楽曲推薦や視覚と聴覚を用いた楽曲推薦の関連研究を紹介する．3 章では，レシピと音楽を融合した閲覧システムについて述べる．4 章では，提案手法によって構築したレシピと音楽を融合した閲覧システムの有用性を評価するための実験について述べる．最後に，5 章ではまとめと今後の課題について述べる．

2　関連研究

2.1　画像を利用した音楽推薦

音楽と写真を同期させる最初の試みは，Sakai らによって 2013 年に報告されている [3]．彼らは，音楽と写真から「覚醒度（arousal）」と「感情価（valence）」を抽出し，写真の HSV

1 : https://recipe.rakuten.co.jp/
2 : https://cookpad.com/
3 : https://www.kurashiru.com/
4 : https://www.nii.ac.jp/dsc/idr/rakuten/rakuten.html
5 : https://developer.spotify.com/documentation/web-api

と音楽の 19 の音響特徴に基づいて「覚醒度」と「感情価」の値を算出し，感情的な音楽推薦システムの提案を行っている．追木ら [4] は，画像の印象に合った音楽の推薦システムを提案している．画像に対して一般的な物体認識を行い，Word2Vec を使用することで，入力画像の雰囲気に適した音楽を推薦した．

また，福本ら [5] は印象派絵画の色彩が Zipf の法則に従うことから統計的特徴を利用し，絵画と文章，絵画と音楽の融合による表現・鑑賞手法の実現を提案している．Lee ら [6] は Russell の Arousal-Valence モデルを用いて絵画と音楽の融合システムを提案している．絵画から色の組み合わせを抽出し，事前に定義された色情動モデルを用いて，統計的学習過程を経ずに arousal と valence の値を推定した．

これらの研究では，音楽と画像の対象が対応付けられていたが，本研究では，ユーザが料理レシピを閲覧する際のレシピと音楽との融合を目指す．音楽にはさまざまな展開や雰囲気があり，それらを閲覧するレシピごとに融合させることで，ユーザが楽しみながら料理することができるとともに，レシピ投稿者が音楽を簡単に選択しやすくすることを目指す．また，料理レシピには料理画像だけでなく，レシピの紹介文にも料理の特徴や雰囲気の要素を捉えることができるため，本研究ではテキスト情報からも特徴量を抽出する．

2.2 視覚と聴覚を用いた音楽推薦

視覚情報と聴覚情報における印象の相互作用に関する研究が存在する．石田ら [7] は，音楽と画像の組み合わせがどのように印象に影響を与えるかを調査している．音楽の音量，ピッチ，音色などの特徴が画像の印象にどのような影響を与えるかを分析した．桑田と宝珍 [8] による研究では，音と映像の印象の類似性によってもたらされる意味的調和の効果を検討している．音と映像の構造的調和の要因として，時間構造の同期と変化パターンの一致が有効であることを実証した．

また，小林と高久 [9] による研究では，音楽の類似性に基づく楽曲推薦手法について研究している．楽曲の音響特徴量を用いて，類似した楽曲を提示する手法を提案している．佐々木ら [10] は視覚と聴覚が調和した楽曲推薦システムの提案を行っている．これにより，AV 空間を用いて，情景に対する感じ方と楽曲に対する感じ方の対応をとることで，状況に適した音楽の推薦を可能にしている．また，画像の特徴量である彩度と輝度の相関関係を利用して，AV 空間に反映させており，音楽の特徴量として，音響特徴量抽出のツールである MIR toolbox Ver1.3 [11] によって得られた特徴量を主成分分析により算出された，第一，第二主成分に掛け合わせることで，音楽を AV 空間に配置している．

これらの研究では，各種の画像（映像），音楽の要素の効果を加える研究であったが，本研究では，ユーザが料理レシピを閲覧する際のレシピと音楽との融合による音楽推薦を目指す．料理レシピの料理画像や紹介文のような視覚情報と音楽のような聴覚情報が調和した音楽推薦によるレシピの閲覧支援システムを提案する．

3 レシピと音楽を融合した閲覧支援システム

本研究では，楽天レシピのレシピデータから各料理レシピの特徴量を抽出する．次に，Spotify から無作為に音楽を選択し，その音楽の特徴量を抽出する．最後に，各料理レシピの特徴量と音楽ごとの特徴量との類似度を算出することで，料理レシピと音楽を融合した閲覧支援システムを構築する．全体的なシステムの概要として，ユーザが料理レシピを選択するとそのレシピの Web サイト上で料理の雰囲気に合った音楽が流れる仕組みになっている．

3.1 レシピの特徴量抽出

本研究では，料理レシピデータとして楽天レシピを使用し，異なるジャンルの料理 22 件から料理レシピの紹介文の感情値，料理画像の彩度および明度を抽出する．紹介文の感情値を抽出する感情分析ツールとして Google の Cloud Natural Language API[6] を用いる．Cloud Natural Language API で定義されている紹介文全体の感情値を求め，その値を正規化する．また，レシピ画像の明度および彩度の値を求め，正規化を行う．これら 3 つの特徴量からなる料理レシピベクトルをレシピごとに生成する．

3.2 音楽の特徴量抽出

本研究では，Spotify から異なるジャンルの 22 曲を音楽データとして使用する．これらの音楽は歌詞のないインストゥルメンタル音楽として扱われる．歌詞が含まれると，料理以外の情報が混入し，情報の混乱が生じる可能性がある．また，音楽の長さはすべて 5 分以下に制限されている．音楽の特徴量抽出には Spotify API が提供する audio-features を利用し，感情値（valance），聴覚的な強さ（energy），テンポ（tempo）の 3 つを取得する．Valance は 0.0 から 1.0 までの値で，楽曲が伝える音楽的なポジティブさ・ネガティブさを表現し，高い値はよりポジティブを示し，低い値はよりネガティブを示す．Energy は 0.0 から 1.0 までの値で，聴覚的な強さやアクティビティを表し，エネルギッシュなトラックは速く，音が大きく，騒々しい曲となる．Tempo は曲の全体的な推定テンポを 1 分あたりの拍数（BPM）で表す．特徴量の選出においては，中野ら [12] が各特徴量のデータ分布や分散について全曲を調査し，energy，popularity，tempo，valence が値が分散しており，異なる雰囲気の音楽にも対応できる可能性が高いと考えられている．したがって，上記の valance，energy，tempo の 3 つの特徴量を使用し，これらの特徴量を正規化し，それぞれの 3 つの特徴量から成る音楽ベクトルを生成する．

3.3 レシピと音楽の対応付け

ここでは，料理レシピと音楽の各特徴量との対応付けについて説明する．図 2 では，音楽と画像の融合のための特徴量を抽出する評価値と，画像特徴量との対応付けを記述する．料理

6 : https://cloud.google.com/natural-language/

レシピの紹介文の感情値は同じ感情値である音楽の valance と対応付ける．彩度と明度に関しては，安達ら [13] が，彩度の高低と音の強弱が対応し，彩度が高いほど楽曲の力強さを感じることを示している．このことから，料理画像の彩度を音楽の energy と対応付ける．星野ら [14] は，テンポの遅速は明度段階にほぼ対応することを示している．これに基づき，料理画像の明度を音楽の tempo と対応付ける．

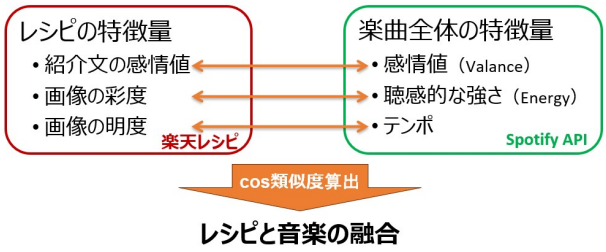


図 2　料理と音楽の対応づけ

3.1 節の方法で生成した料理レシピベクトル $\mathbf{X}$ と 3.2 節の方法で生成した音楽ベクトル $\mathbf{Y}$ とのコサイン類似度を下記の式 (1) により算出し，最も類似度の大きい組み合わせを決定する．

$$\cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|} \qquad (1)$$

ここで，$\cdot$ はベクトルの内積を示し，$\|\mathbf{X}\|$ および $\|\mathbf{Y}\|$ はそれぞれベクトル $\mathbf{X}$ と $\mathbf{Y}$ のノルム（長さ）を示す．

4　評価実験

本章では，3.3 節で提案手法に基づき行った料理レシピと音楽の各特徴量とのコサイン類似度の算出結果を示す．本研究では，楽天レシピから無作為に選んだ料理レシピ 22 件を対象に，紹介文から感情値，料理画像から彩度，明度を抽出した．また，Spotify から無作為に選んだ音楽 22 曲を対象に，音楽の特徴量 valance，energy，tempo を抽出した．これらのコサイン類似度の算出結果から上位 7 件を表 1 に示す．

表 1　料理レシピと音楽の各特徴量とのコサイン類似度

料理レシピ × 音楽	コサイン類似
失敗なし ♡ パリパリ ♡ 屋台のりんご飴 × しゅわしゅわハニーレモン	0.9997
簡単☆定番 ハムチーズのサンドイッチ × 10°C	0.9988
冷めてもおいしい ♡ 揚げない大学芋 × Cat life	0.9975
にんじん大量消費にんじんとツナ缶のゴマ味噌マヨ和え × Local Forecast7	0.9960
失敗なし ♡ パリパリ ♡ 屋台のりんご飴 × 亡霊の女の子	0.9953
イスラエルの大人気おやつ！チーズたっぷりのブレカス × summer triangle	0.9952
簡単すぎる◎◎ゆずのはちみつ漬け × 10°C	0.9942

4.1　実験概要

本方式によるレシピと音楽を融合した閲覧支援システムの有用性を検証するために男女 15 人の被験者に対してアンケート調査を行った．被験者には，あらかじめ類似度算出によって導き出した図 3 のデモ動画を閲覧してもらい，以下の設問項目についてアンケートを実施した．アンケートでは，レシピと音楽の特徴量の対応付けが適切か，システムの有用性，改善すべき点や追加してほしい機能について質問した．

図 3　デモの表示画像（上からデモ 1，デモ 2，デモ 3）

Q1.　動画内のレシピの雰囲気と音楽が合っていた．
Q2.　レシピの雰囲気と音楽の明るさ（ポジティブ・ネガティブ）が合っていた．
Q3.　レシピの雰囲気と音楽の強弱が合っていた．
Q4.　レシピの雰囲気と音楽のテンポが合っていた．
Q5.　このレシピ閲覧システムは有用性があった．
Q6.　今後改善してほしい点や，追加してほしい機能があれば記述してください．

Q1〜Q5 に対しては 5 段階（1：全くそう思わない，2：あまり思わない，3：どちらともいえない，4：ややそう思う，5：とてもそう思う）で評価してもらった．Q1，Q2，Q3，Q4 では，上位 3 件のデモ動画に対して，音楽の明るさ，強弱，テンポがレシピの雰囲気に適しているかどうかを 5 段階評価によって回答してもらった．これは，音楽の明るさ，強弱，テンポを客観的に判断してもらうことで，違和感なくレシピ閲覧支援システムを利用できるか検証するためである．Q5 では，システムの

有用性を評価してもらい，5 段階評価で回答してもらった．Q6 は記述式質問であり，詳しくわかるよう自由に記述できるコメント欄を設け，システムの改善点や追加すべき機能を記述式で回答してもらった．

4.2 実験結果

Q1 の「レシピと音楽の融合に対する評価」の結果を図 4 に示す．デモ 1 の平均値が 4.8 となり，最も高い評価を得られており，次いで，デモ 2 とデモ 3 が同等の評価を受けた．

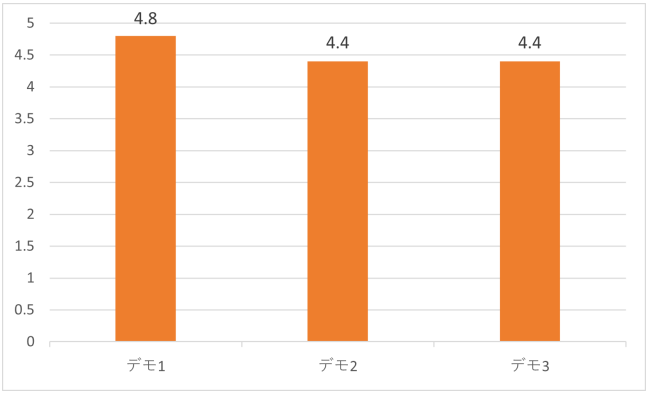


図 4　Q1 の「レシピと音楽の融合に対する評価」の各デモの平均値

Q2～Q4 の「特徴量に関する評価」の結果を図 5 に示す．Q2 の明るさに関する評価は，デモ 1 が最も高く 4.8，デモ 3 が最も低く 3.6 であった．Q3 の音の強弱に関する質問は全体的に低い評価値となり，最高で 4.0 であった．Q4 のテンポに関する質問に対するデモ 1 の平均値は 4.4，デモ 2 の平均値は 3.6，デモ 3 の平均値は 4 であった．

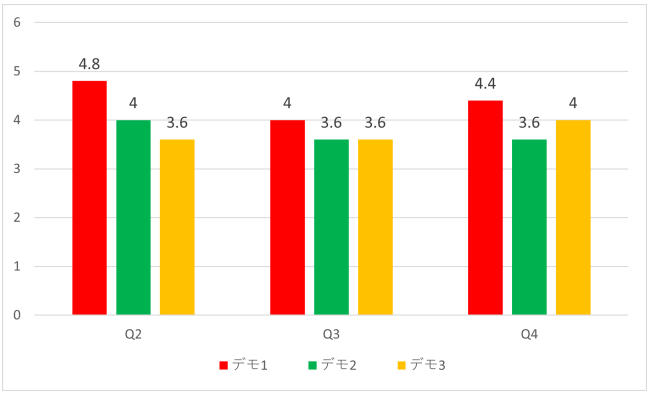


図 5　Q2～Q4 の「特徴量に関する評価」の各デモの平均値

Q5 の「システムの有用性に関する評価」の結果では，デモ 1 の平均値は 4.4，デモ 2 の平均値は 4.4，デモ 3 の平均値は 4.0 であった．また，各質問に対する平均値を図 6 に示す．Q1 が 4.5，Q2 が 4.1，Q3 が 3.7，Q4 が 4，Q5 が 4.2 であり，Q3 に対する評価が最も低い結果になった．

Q6 の「改善してほしい点や追加してほしい機能」の質問には，「1 つの曲がループで流れると利用者が飽きてしまうので，1 つのレシピに対して複数の曲を流すと良い」，「利用者の音楽の好みを考慮して音楽を推薦出来たら良い」，「音楽を再生して欲しくないユーザもいるので，音楽をオン／オフにできる機能があると良い」などと言う回答があった．

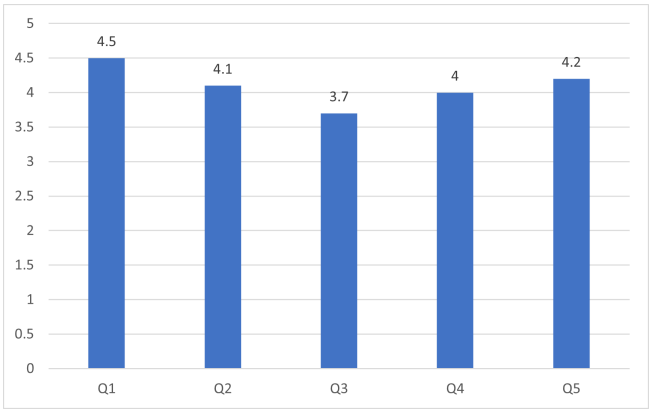


図 6　各質問の平均値

4.3 考　　察

Q1 はレシピと音楽の融合に関する質問であり，すべてのデモ動画で 4.4 以上の値であった．また，5 段階評価中すべての回答が「とてもそう思う」「ややそう思う」であったため，今回使用した上位 3 つのデモ動画はいずれもレシピの雰囲気と音楽が合っていたといえる．

Q2～Q4 はレシピと音楽の特徴量に関する質問であり，それぞれのデモ動画ごとに多少ばらつきはあるが，どの質問も 3.5 以上の値であったためレシピと音楽の特徴量が適しているといえる．

Q5 はシステムの有用性に関する質問であり，すべてのデモ動画で 4.0 以上の値であったので，特定の料理や音楽に関わらずこの料理閲覧システムは有効であるといえる．

図 6 では，各質問の平均値を表している．どの質問も 3.5 以上の値であったが，Q3 の音の強弱に関する質問では，最も低く 3.7 であった．この理由として，「音の強弱」が言葉で表しにくく，この特徴量に関する説明が不十分であったためだと考えられる．

Q6 の「改善してほしい点や追加してほしい機能」の回答では，複数の曲の提供，ユーザの好みを考慮した音楽推薦，音楽のオン/オフ機能の追加などといった意見があり，ユーザがより快適かつ一人一人に合わせたサービス提供が必要であると考えられる．

5　おわりに

本研究では，楽天レシピのデータから紹介文，レシピ画像の彩度および明度を特徴量として抽出し，Spotify API を用いて無作為に選ばれた音楽の感情値（valance），聴覚的な強さ（energy），テンポ（tempo）を特徴量として取得した．これらの特徴量を比較し，コサイン類似度を算出することで，特定のレシピの雰囲気に合った音楽を対応付け，レシピと音楽を融合した閲覧支援システムを提案した．

今後の課題として，料理のジャンルによる音楽の対応づけ，

またユーザの好みに基づく音楽の提案，音楽のオン/オフ機能の追加といったユーザ一人一人に対応した機能の付与が必要であると考えている．

謝　　辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用した．ここに記して謝意を表す．

文　　献

[1] マルハニチロのコーポレート: 料理レシピに関する調査 2020. マルハニチロ株式会社, 2020 年 8 月 18 日, https://www.maruha-nichiro.co.jp/corporate/new_center/news_topics/20200818_research_recipe2020_1.pdf (参照: 2023 年 12 月 25 日).

[2] 楽天グループ株式会社: 楽天レシピデータ. 国立情報学研究所情報学研究データリポジトリ. (データセット), 2016. https://doi.org/10.32130/idr.2.4.

[3] Shoto Sasaki, Tatsunori Hirai, Hayato Ohya, and Shigeo Morishima: Affective Music Recommendation System Reflecting the Mood of Input Image. in Proceedings of the 2013 International Conference on Culture and Computing, pp. 153-154, 2013.

[4] 追木智明, 欅惇志, 宮崎純: 画像の印象に合った音楽の推薦システムの提案. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2018)，P2-3, 2018 年 3 月.

[5] 福本麻子, 塚田浩二, 蔡東生, 安村通晃: Ziplayer: 絵画, 文章, 音楽を組み合わせたメディアプレイヤーの提案. 情報処理学会シンポジウム論文集（情報処理学会ワークショップ論文集）, Vol. 2006, No. 4, pp. 79-80, 2006 年 3 月.

[6] Taemin Lee, Hyunki Lim, Dae-Won Kim, Sunkyu Hwang, and Kyunghyun Yoon: System for Matching Paintings with Music based on Emotions. in Proceedings of the SIGGRAPH ASIA 2016 Technical Briefs (SA '16), Article No.: 31, pp. 1-4, 2016.

[7] 石田理代, 佐藤敬子, 満倉靖恵: 音楽が画像印象に与える影響と音楽特徴量との関連性. 電気学会論文誌 C（電子・情報・システム部門誌）, Vol. 136, No. 9, pp. 1376-1385, 2016 年 9 月.

[8] 桑田和也, 宝珍輝尚: 視聴覚素材における音と動画の調和について. 日本感性工学会論文誌, Vol. 9, No. 1, pp. 103-110, 2009 年.

[9] 小林恭輔, 高久雅生: 楽曲探索を支援するための類似楽曲提示手法. 情報知識学会誌, Vol. 32, No. 2, pp. 287-293, 2022 年 5 月.

[10] 佐々木将人，平井辰典，大矢隼士，森島繁生: 入力画像に感性的に一致した楽曲を推薦するシステム. 情報処理学会第 75 回全国大会講演論文集, 2D-5, pp. 45-46, 2013 年 3 月.

[11] Olivier Lartillot, and Petri Toiviainen: IR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio. Proceedings of the 10th International Conference on Digital Audio Effects, pp. 127-130, 2007.

[12] 中野美由紀, 石山悟志, 君野史明, 小久保勇気, 斉藤大介, 鈴木智尚, 宮崎晃一, 吉村慎祐: 利用者の印象に基づく音楽レコメンドサービス : 画像を利用した利用者指向の音楽レコメンド手法の考察. 情報処理学会第 81 回全国大会講演論文集, Vol. 2019, No. 1, pp. 383-384, 2019 年 2 月.

[13] 安達太郎, 岩宮眞一郎: 色彩と音楽とが互いに及ぼす影響–ショパンのエチュードを手がかりに. 第 5 回学生のための研究発表会講演論文集（日本音響学会九州支部）, pp. 13-16, 2003 年.

[14] 星野喜久三: 色とテンポの関係に関する研究. 北海道學藝大學紀要, 第一部, Vol. 10, No. 1, pp. 233-239, 1959 年 7 月.

最適輸送理論を用いたメロディー楽譜からの作曲家推定

八木　皓平†　　中村　篤祥†

† 北海道大学 工学部 情報エレクトロニクス学科　〒 060–0814 北海道札幌市北区北十四条西 9 丁目
E-mail: †kyagi@ist.hokudai.ac.jp

あらまし　近年、最適輸送理論を用いて分布間の距離を求めるアルゴリズムが注目されている。このアルゴリズムは、他画像に基づく色相変換、動画における不連続フレームの補完画像の作成、文書クラスタリングなど様々な用途に用いられている。本論文では、この最適輸送理論を用いて曲のメロディー間の距離を算出し、作曲家推定を行う手法を提案する。J-POP のメロディーのみの楽譜（MIDI ファイル）を入力として用い、音の高さや音のつながりなどの情報から楽曲間の距離を算出する。また、提案された手法を用いて行った作曲家推定の精度やクラスタリング精度を検証する。

キーワード　音楽，作曲家推定，最適輸送理論

1　はじめに

ポピュラー音楽は世の中に浸透し、次々と新しい曲が発表されている。J-POP の場合は、メロディーを作曲する人と、メロディーに伴奏などの肉付けをする編曲を行う人が分業して作品を作りあげることが多い。ヒット曲を連発する作曲家もいるが、同じ作曲家の作る曲は曲調が似たものが多いように感じる。我々が普段耳にするのは編曲されて伴奏のついた作品であり、同じ歌手が歌っている曲であることも多く、それが曲調が似ていると感じてしまう理由なのかもしれない。それでは、伴奏による飾りや歌声の無いメロディー楽譜のみから、作曲家の推定は可能なのであろうか？

本稿では、この疑問に答えるために、メロディー楽譜のみから作曲家を推定する問題に取り組む。メロディーにおける、音高・音高列・音長・音長列などの分布の類似性から作曲家を推定するアプローチをとる。分布の類似性尺度として、近年注目を集めている最適輸送理論を用いる方法を提案する。

最適輸送理論とは値の移動コストを設定することでそのコストを用いた際の 2 分布間の距離を算出する理論である。距離が短いほど類似度が高いとみなすことが出来るので、最適輸送距離は分布間の類似度の測定に用いることが出来る。本論文では、メロディーの特徴量（分布）間の最適輸送距離を用いて最近隣法で作曲家推定を行う。そして、特徴量毎に最適輸送距離およびそれらの合成距離を求め、近い順にランキングしたリストをランキングの評価指標を用い評価する。また、階層型クラスタリングも行い、作曲家ごとにどの程度クラスタリングされるか確認する。

10 名の作曲家による J-POP 50 曲のメロディー譜を用いた作曲家推定の実験では、音高・連続する 2 音符の相対音高変化量・連続する 3 音符の相対音高変化量ベクトル・音長・連続する 2 音符の相対的音長の 5 つの局所的特徴量の分布各々で輸送距離により作曲家推定を行った結果、音高を用いた場合の正解率が最も高く 54%であった。また、5 つの局所的特徴量を用いて輸送距離により階層型クラスタリングを行い、作曲家数にクラスタリングした結果、音高によるものが最も精度が良くクラスタリングでき、作曲家毎のクラスタとの比較では 50 曲中 25 曲が同じクラスタに属するようにマッチングすることが可能であった。これらの結果はランダムに推定する場合（作曲家推定正解率：8.16%, 作曲家毎クラスタマッチ数：14.88）よりも高く、本論文で提案する手法は作曲家推定に有用であることが確認できた。

2　関連研究

本章では、楽譜（MIDI ファイル）を入力として用いる作曲家推定の研究として、ニューラルネットワークによる作曲家推定、音楽学における様式分析手法に基づく特徴量を用いた線形の正準判別分析による作曲家推定の研究について述べる。

2.1　ニューラルネットワークを用いた作曲家推定

Maximos らは、コンピュータによる作曲家推定に必要な情報は、楽譜のどの記号的特徴から得られるのかを調査するために、簡単でコンパクトな楽譜ベースのデータ抽出技術である Dodecaphonic Trace Vector（オクターブの倍数の違いを無視した 12 音階の音符の出現頻度）を特徴量としてニューラルネットワークを用いて作曲家推定を行った [4]。この研究では、7 人のクラシック作曲家によって作曲家された計 350 曲をデータセットとして用いて推定を行い、各作曲家ごとの推定での正解率は最大で 66.1%、平均正解率は 48.16%であった。

2.2　線形の正準判別分析を用いた作曲家推定

長谷川らは、音楽から受ける「作曲家らしさ」の印象を説明し定量的に測定できる工学的手法を目指し、音楽学における様式分析手法の 1 つであるラルーらの綜合的様式分析において論じられている様々な定性的特徴に対応する特徴量を提案した [3]。Sound, Harmony, Melody, Rhythm の 4 つの要素からなる 14 の特徴量を提案し、線形の正準判別分析を用いて作曲家推定を行った。この研究ではデータセットとして、26 作曲家

1966 曲を用いて推定を行い、正解率は 56.12%であり、2 位含有率は 73.98%、3 位含有率は 81.63%であった。

3 提 案 手 法

本研究では、メロディーのみ書かれた楽譜から局所的な値の分布からなる特徴量を抽出し、最適輸送理論を用いて作曲家推定やクラスタリングを行う。

3.1 最適輸送理論

最適輸送理論とは、2 つの分布の違いから分布間の距離を算出するためのものである [1]。分布 A と分布 B の間の距離は分布 A に従うある値から分布 B に従うある値へ単位割合だけ輸送する際のコストをコスト行列として定義し、そのコスト行列に従い分布 A から分布 B へ輸送した際の合計コストの最小値として定義する。n カテゴリ上のヒストグラムの集合を $\Sigma_n = \{(x_1, \ldots, x_n) \mid \sum_{i=1}^n x_i = 1,\ x_i \geqq 0\ (i = 1, 2, \ldots, n)\}$ とする。n 個のカテゴリからなる 2 つのヒストグラム $\boldsymbol{a}, \boldsymbol{b} \in \Sigma_n$ と、カテゴリ i からカテゴリ j への輸送コスト $\boldsymbol{C}_{ij}$ からなるコスト行列 $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ を入力とした際の最適輸送問題は以下のように定義される。ただし、$\boldsymbol{P}_{ij}$ は分布 $\boldsymbol{a}$ の i カテゴリから分布 $\boldsymbol{b}$ の j カテゴリへの輸送量を表し、$[n] = \{1, 2, \ldots, n\}$ とする。

$$
\begin{aligned}
\underset{\boldsymbol{P} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{C}_{ij} \boldsymbol{P}_{ij} \\
\text{subject to} \quad & \boldsymbol{P}_{ij} \geqq 0 && ({}^{\forall} i \in [n], {}^{\forall} j \in [n]) \\
& \sum_{j=1}^{n} \boldsymbol{P}_{ij} = \boldsymbol{a}_i && ({}^{\forall} i \in [n]) \\
& \sum_{i=1}^{n} \boldsymbol{P}_{ij} = \boldsymbol{b}_j && ({}^{\forall} j \in [n])
\end{aligned}
$$

この問題の最適値を最適輸送コスト $OT(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{C})$ と表し、この値が分布 $\boldsymbol{a}, \boldsymbol{b}$ 間の距離となる。

3.2 特徴量としての分布の抽出

本研究では楽譜の MIDI 表現を用いて局所的に定まる値の分布を抽出する。MIDI データはおおまかに、楽譜の中の音符と一対一で対応する note というデータから構成されている。その中には細かく分けて音を出す命令である note on と音を消す命令である note off があり、各々に出す音の高さ (note number)、長さ (time)、強さ (velocity) が記録されている。この MIDI 表現を用いて、1 曲のメロディーに現れる以下の値の分布を抽出する。

3.2.1 音　　高

音符の音高のヒストグラムを抽出する。音高は MIDI のノートナンバー 0〜128(C-1〜G9, 8.2〜12543.9Hz) 表現を用いる。抽出するヒストグラムは 129 次元ベクトルとなる。

3.2.2 連続する 2 音符間の音高の相対変化量

連続する 2 音符間の音高の相対変化量のヒストグラムを抽出する。音高の相対的変化量は 2 音符のノートナンバーの差とする。実験では、範囲を [-15, 15] に限定して 31 次元ベクトルとしでヒストグラムを表す。

3.2.3 連続する 3 音符間の音高の相対的変化量ベクトル

連続する 3 音符の、最初の 2 音符間の音高の相対的変化量 a、および最後の 2 音符間の音高の相対変化量 b のベクトル (a, b) のヒストグラムを抽出する。音高の相対的変化量は 2 音符のノートナンバーの差とする。実験では、各成分の範囲を [-15, 15] に限定して 961 次元のベクトルでヒストグラムを表す。

3.2.4 音　　長

16 分音符、付点 16 分音符、8 分音符、付点 8 分音符、4 分音符、付点 4 分音符、2 分音符、付点 2 分音符、全音符、付点全音符のヒストグラム（10 次元）を抽出する。（その他の長さの音符は全て除外してカウントする。）

3.2.5 連続する 2 音符の音長比

連続する 2 音符の長さの比のヒストグラム（57 次元）を抽出する。

3.3 コスト行列の設定

コスト行列を各特徴ごとに定める。本論文では音楽の知識などは用いず、値の差をとるような形でコスト行列を以下のように設定する。

- 音高：ノートナンバーの差の絶対値をコストとして設定する。例えば、ノートナンバー 60 から 62 へのコストは 2 となる。
- 連続する 2 音：変化量の差の絶対値をコストとして設定する。例えば、変化量-10 から+5 へのコストは 15 となる。
- 連続する 3 音：2 音の場合と同様に変化量の差の絶対値をコストとして設定する。例えば、変化量 [+1, -1] から [-3, 0] へのコストは 5 となる。
- 音長：MIDI データの中での音符としての長さ（8 分音符は 216,16 分音符は 108 など）の差の絶対値をコストとして設定する。
- 音長比：長さの比としての数値が出るため、比較する 2 値の大小を比べ、大きい数値を小さい数値で割り、1 を引いた値をその 2 値間のコストとして設定する。

3.4 作曲家推定

それぞれの特徴量としての分布から計算した曲間の輸送距離を用いて最近隣法により作曲家推定を行う。また、曲のヒストグラム $\boldsymbol{a}, \boldsymbol{b}$ 間の 5 つの正規化された距離 $d_i\ (i = 1, 2, 3, 4, 5)$ の合成距離 d を

$$
d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{5} w_i d_i(\boldsymbol{a}, \boldsymbol{b})^2} \qquad (1)
$$

と定義し、この距離を用いた最近隣法による作曲家推定も行う。ただし、$w_i \in \{0, 1\}$ とし、各正規化された距離 d_i は使うか使わないかの 2 択とする。また、各輸送距離の正規化は最大値で割ることにより行う。

3.5 曲のクラスタリング

それぞれの特徴量としての分布から計算した曲間の輸送距離を用いて階層型クラスタリングにより曲のクラスタリングを

行う。

4 実　　験

4.1 データの作成方法

本実験を行うにあたりメロディー情報のみが抽出された MIDI データが必要になったが、そのようなデータを扱っている機関や公開しているサイトなどは確認できなかったため自ら作成した。方法としては、メロディーのみ書かれた楽譜 [2]、通称メロ譜をスキャンし、MIDI データを作成することのできるアプリ、スコアメーカー ZERO[1] を使用し、データを作成した。このアプリとメロ譜を用い、合計 10 人の作曲家からなる全 50 曲の MIDI データを作成した。[2]（作曲家、曲リストは付録参照）

4.2 評 価 方 法

1 つ抜き交差検証 (leave-one-out cross-validation) を用いて最近隣法による正解率、および距離の近い順に曲をランキングした場合に正解の作曲家の曲が上位 2 位まで、または上位 3 位までに含まれる率で作曲家推定方法を評価する。

また、抽出した特徴をランキングの評価指標 (MRR) を用いて評価する。MRR(Mean Reciprocal Rank) はもっとも単純なランキングの評価指標で、全曲数を N とし、推定対象の曲 i との距離が近い順に曲をランキングした場合に正解の作曲家の曲が最初に現れる順位を $rank_i$ とすると、以下のような式で計算できる。

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

この評価指標が 1 に近ければ近いほど、正解やそれに近い答えを出せていることになるため数値的に抽出した特徴を評価することが出来る。曲のクラスタリングの評価方法としては、作曲家ごとにクラスタリングされたものを正解として、同じクラスタ数（作曲家数）を指定してクラスタリングすることにより作られたクラスタと正解クラスタとのペアを作っていき、正解のクラスタとクラスタ内の曲を比較することでクラスタリングの評価を行う。以下の図 1 のように正解クラスタと実験結果のクラスタがあった場合には評価点は 9 点となる。ペアは正解数が最大となるように作るものとする。

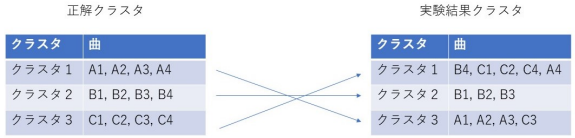


図 1　クラスタの評価

1：楽譜を画像認識し MIDI ファイル作成できるアプリで、誤認識なども簡単に編集できるアプリである。
“スコアメーカー ZERO シリーズ”, KAWAI コンピュータミュージック, https://cm.kawai.jp/products/smz/,(参照 2024/02/15)
2：メロ譜には前奏や間奏、後奏部分が記載されていたが、それらは作曲家ではなく編曲家が作成している場合もあるため、データ作成の編集の段階でカットした。

4.3 結　　果

表 1 に作曲家推定、表 2 にクラスタリングの実験結果を示す。比較対象としてランダムに推定した場合の正解数や、ランダムにクラスタリングした場合の評価点も加えて示す。(MRR とクラスタリングの評価点はランダムに 50 回推定、またクラスタリングした結果の平均値をランダムの結果とし、正解率,2 位,3 位含有率は確率計算の値をランダムの結果として用いた。) また、特徴量を組み合わせて合成距離を作る方法において、最良の結果が得られた場合の特徴量の組み合わせとその結果も示す。各特徴量を組み合わせ、すべての場合で推定を行った結果、音高のみを使った場合の推定結果が最も正解率が高かった。

表 1　作曲家推定結果

使用した特徴量	正解率 (%)	2 位含有率 (%)	3 位含有率 (%)	MRR
音高	54	66	74	0.6629
連続する 2 音	16	26	42	0.3400
連続する 3 音	20	22	30	0.3260
音長	30	44	52	0.4521
音長比	20	30	40	0.3537
ランダム	8.16	15.82	22.98	0.2304
最良組み合わせ	54	66	74	0.6629

表 2　クラスタリング結果の評価

使用した特徴量	評価点 (点)
音高	25
連続する 2 音	14
連続する 3 音	19
音長	17
音長比	17
ランダム	14.88

4.4 考　　察

以上の結果より、ランダムに推定する場合に比べ特徴を抽出し最適輸送理論を用いて曲同士の距離を測定し推定を行った場合のほうが推定精度が高く、クラスタリング精度も高い場合が多いため、これらが推定精度を高めることに有用であることが確認できる。そして、MRR による評価によると特徴量の中では音高が最も有用な特徴量であることが確認できる。また、複数の特徴量を組み合わせて推定を行うよりも音高のみを使った場合の推定結果のほうが正解率が高かった。しかし、表 3 に示した作曲家別の特徴量ごとの正解数をみると、各作曲家ごとに推定しやすい特徴量があることが分かる。桑田佳祐と椎名林檎は音高以外の特徴量でより正解率の高いものが存在する。よって、距離を合成する際に今回の式 1 のように単純な合成方法ではなく、各特徴量に重みづけを行い組み合わせることで精度の向上につながる可能性がある。

表 3　作曲家別の特徴量ごとの正解数

作曲家名	音高	連続する 2 音	連続する 3 音	音長	音長比
小室哲哉	4	0	3	2	0
織田哲郎	2	1	2	2	1
桜井和寿	3	0	0	1	2
Ayase	3	2	1	3	3
米津玄師	2	1	1	0	0
藤原聡	3	0	0	2	0
桑田佳祐	3	4	3	1	2
草野正宗	4	0	0	3	0
椎名林檎	0	0	0	1	0
福山雅治	3	0	0	0	2
合計	27	8	10	15	10

5 ま と め

本論文で提案した手法を用いることで作曲家の特徴を抽出することが出来、作曲家推定に有用であることが分かった。今回は 5 つの特徴量のみを用いた推定であり、音楽の専門的な知識を用いていなかったが、音楽の知識に基づいた特徴量の抽出やコスト関数の設定を行うことで精度の向上が期待できる。今回の実験ではデータを自ら作成し、実験を行うという都合上データ数が非常に限られた中で実験を行うことになってしまった。今後はデータ数の増加を行い、さらなる精度の作曲家推定を実現していきたい。

文　　献

[1] 佐藤竜馬, “最適輸送の理論とアルゴリズム,” 講談社, 2023.
[2] “平成ポップス大全集 BEST 236,” 全音楽譜出版社, 2023.
[3] 長谷川 隆, 西本 卓也, 小野 順貴, 嵯峨山 茂樹, “楽譜情報からの作曲家らしさ認識のための音楽特徴量の提案,” 情報処理学会論文誌, Vol. 53, No. 3, pp. 1204–1215, 2012.
[4] Kaliakatsos-Papakostas, M.A., Epitropakis, M.G., Vrahatis, M.N., “Musical Composer Identification through Probabilistic and Feedforward Neural Networks,” EvoApplications 2010:Applications of Evolutionary Computation, pp. 411–420, 2010.

付　　録

本研究で使用した曲一覧

- 小室哲哉

“CAN YOU CELEBRATE ?”, “Don’t wanna cry”, “love the island”, “SWEET 19 BLUES”, “You’re my sunshine”.

- 織田哲郎

“DAN DAN 心魅かれてく”, “世界が終わるまでは . . . ”, “碧いうさぎ”, “夢見る少女じゃいられない”, “恋心”.

- 桜井和寿

“HANABI”, “Sign”, “しるし”, “終わりなき旅”, “名もなき詩”.

- Ayase

“アイドル”, “ハルジオン”, “怪物”, “群青”, “夜に駆ける”.

- 米津玄師

“KICK BACK”, “Lemon”, “LOSER”, “アイネクライネ”, “ピースサイン”.

- 藤原聡

“115 万キロのフィルム”, “I LOVE”, “Pretender”, “Subtitle”, “ミックスナッツ”.

- 桑田佳祐

“TSUNAMI”, “真夏の果実”, “波乗りジョニー”, “白い恋人達”, “明日晴れるかな”.

- 草野正宗

“愛のしるし”, “チェリー”, “ロビンソン”, “空も飛べるはず”, “楓”.

- 椎名林檎

“ギブス”, “歌舞伎町の女王”, “丸の内サディスティック”, “罪と罰”, “長く短い祭”.

- 福山雅治

“HELLO”, “家族になろうよ”, “桜坂”, “Good night”, “Squall”.

教師なし学習による主成分分析を用いた流行楽曲の特徴量の相関

小俵　亜紀† 　木津　美咲† 　内田　瑞貴† 　原　香奈江† 　齋藤　優莉†
金井友希美†

† 奈良女子大学 理学部　〒 630–8506 奈良市北魚屋東町
E-mail: † waa_odawara@nara-wu.ac.jp, {uam_kizu,vam_uchida,xak_hara,xay_saito}@nara-wu.ac.jp,
kanai@cc.nara-wu.ac.jp

あらまし　多様化する嗜好性に合わせた視聴環境が提供される一方で，ヒット曲のように多くのひとが好む楽曲が存在する．しかし，楽曲を構成する因子は多く，流行曲に共通する因子として流行因子を見出すことは困難である．本研究では，その因子を構成する特徴量の抽出を目的とする．Spotify が提供する Web API である Spotify API では，楽曲を特徴づけるものとしてテンポやエネルギッシュさなどを表す楽曲特徴量が定義されている．これらをまとめたデータセットに対して教師なし学習による主成分分析をおこない，第一主成分と第二主成分を取得した．その結果をもとに，楽曲に対する特徴量の相関と流行楽曲に共通する因子について検討をおこなった．

キーワード　音声・音楽，可視化，音楽情報処理，楽曲特徴量

1 はじめに

近年，サブスクリプション型の音楽配信サービスの普及に合わせ，一人一人の嗜好性に合わせた視聴環境が提供されるようになった．そのなかでも Spotify や Apple Music などでは全ユーザーの再生履歴情報を取得し，それを分析することで各ユーザーに対して新たな楽曲およびアーティストをおすすめし，それぞれの好みや個性に合わせた楽曲の提供や視聴環境の構築が実現されている．その一方で，テレビだけでなく SNS などを通じて爆発的な流行を見せる楽曲も存在しており，多くのひとが好む流行因子が楽曲の特徴量として存在することが示唆される．しかし楽曲を特徴づけるパラメータは多く，流行を促す因子には楽曲そのもの以外の要因も含まれるため流行曲に共通する流行因子を特定することは大変に困難である．本研究では，楽曲を特徴づけるパラメータから流行曲に共通する因子を見出すことを目的とする．そのため，主成分分析による次元削減を行い，楽曲を特徴づけるパラメータのうち流行因子に関わる特徴量を抽出する．さらにはアーティストごとに主成分との相関を可視化する．本研究によって流行因子を見出すことで，今後の楽曲の流行予測だけでなく多くのパラメータによって特徴づけられるデータセットへの共通因子の特定や相関を議論する手法の提案となることが期待できる．

本論文の構成は以下のとおりである．まず 2 章では主成分分析による次元削減やヒットソング予測についての関連研究を述べる．3 章では本研究のデータセットの作成，および主成分分析による次元削減について説明し，流行因子として抽出した主成分への楽曲特徴量の影響を議論する．4 章では抽出した主成分と popularity による相関も導出し，さらにアーティストごとの楽曲特徴を示す．最後に，5 章では本研究のまとめ，今後の課題について述べる．

2 関連研究

本研究で用いた主成分分析による次元削減と，ヒットソング予測についての研究について述べる．

2.1 主成分分析

大規模なデータセットは近年さらに一般的となっており，そのデータセットの特徴を理解，適用するために多くの手法が開発されている．なかでも最も古くから提案されており，現在でも広く使用されている解析手法の一つとして主成分分析 (Principal component analysis : PCA) が挙げられる [1], [2]. PCA は元々Pearson [3] によって提案され，Hottelling [4] によって発展したもので，固有値/固有ベクトルの解を求め教師なしに線形にマッピングするものである [1], [2].

PCA は，元のデータの情報は可能な限り損失しないように最大の分散を保持しながら，特徴量の次元を削減するものであり，多次元情報を低次元の空間に縮約するためのものである．このように多次元のデータを 2 次元もしくは 3 次元に縮約することで，大規模なデータセットの情報を可視化し，大規模データの解釈を容易にしようというものである．PCA は様々な分野へと応用されており，画像処理や音楽処理 [1], [5], [6] だけでなく，感情の分析 [7] などにも適用されている．そこで本研究では流行曲に共通する因子の特定を主成分分析によって実施し，楽曲自身の特徴による流行予測の手掛かりとなる流行因子を抽出する．

2.2 ヒットソングサイエンス (HSS) についての研究

楽曲の流行予測のための研究はヒットソングサイエンス (Hit Song Science ; HSS) とも呼ばれ，音楽産業への大きな影響ともなるため大変注目されている分野である．聴覚に与える信号から主成分分析などを用いて楽曲を分類する研究 [5] や，音楽の特徴を分析して人の好みを測定する研究なども行われてい

る [6], [8]．もちろん，楽曲がヒットする要因として楽曲そのものの特徴以外にも要因はあるが，音楽は人間の感情を直接揺さぶるほどに影響の大きいものであり，楽曲の特徴がもつヒットの要因は大きい．

我々は Middlebrook らが Spotify API から楽曲特徴量を取得し，約 180 万曲のデータセットからロジスティック回帰（Logistic Regression ;LR），ニューラルネットワーク（Neural Network ;NN），ランダムフォレスト（Random Forest ;RF），サポートベクトルマシン（Support Vector Machine ;SVM）の 4 つのモデルでヒット曲予測を行ったことをきっかけに，本研究を実施した [9]．彼らの研究は 88% という高い精度でヒット曲の予測を成功させ，4 つのモデルの比較は実施したものの実際に楽曲のヒットに影響を与える特徴量がどのようなものであるかなどについては不透明なものであった．

本研究では Middlebrook らと同様に Spotify API から楽曲特徴量を取得し，教師なし学習による主成分分析を行うことでその中から流行因子を特定し，その相関について議論する．

3 主成分分析による楽曲特徴量における流行因子の解析

本研究では，まず流行曲と非流行曲を含む楽曲群に対して Spotify API を用いて楽曲特徴量を取得する．次にそれらに対して教師なし学習による次元削減を適用し，様々な説明変数に対する楽曲の分布を可視化し，流行曲に共通する流行因子について考察をおこなう．

本研究では，Spotify が発表した 2022 年の年間ランキングから「国内で最も再生されたアーティスト」に着目し，TOP 10 にランクインしたアーティストの全楽曲を楽曲群として用いる．それらに対して，Spotify API [10] が提供している楽曲情報である 18 種類の楽曲特徴量を取得した．そのうち楽曲タイトル，収録アルバムタイトル，アーティスト名，リリース日といった 4 つの楽曲の基本情報と，本研究で目的変数とする popularlity を除くその他 13 種類を音響特徴量とする． 13 の音響特徴量の定義は表 1 に示す．また，取得したデータは 2023 年 5 月 26 日時点のものである．

表 1　Spotify API が提供する音響特徴量とその定義

	音響特徴量	音響特徴量の定義
1	length	曲の長さ (秒)
2	acousticness	アコースティックさを 0.0 から 1.0 で表す
3	danceability	ダンスに適しているかを 0.0 から 1.0 で表す
4	energy	知覚的尺度をを 0.0 から 1.0 で表す
5	instrumentalness	楽曲にボーカルが含まれる度合
6	key	ピッチクラスを整数にマッピングしたもの
7	liveness	オーディエンスの存在を 0.0 から 1.0 で表す
8	loudness	楽曲中の平均の音の大きさ (dB)
9	mode	楽曲がメジャー (1) かマイナー (0) か
10	speechiness	話し言葉への近さを 0.0 から 1.0 で表す
11	tempo	楽曲テンポ (BPM)
12	time signature	拍子
13	valence	楽曲が与える印象．ポジティブなほど 1.0 に

3.1 解析方法

3.1.1 次元削減

本研究で用いた，教師なし学習による次元削減について述べる [11]．まず，次元削減とは主成分分析のひとつであり，多成分をもつデータについて成分を適切に統合することにより 2 次元および 3 次元まで次元を削減し，可視化する方法である．本研究ではモデルとして次のような線形モデルを用いた．

$$f(\mathbf{x}_i) = \mathbf{x}_i \mathbf{w} \tag{1}$$

ここで $\mathbf{x}_i$ はデータ点であり，また，$\mathbf{w}$ は主成分軸の正規直交基底ベクトルであり，最適化されるモデルパラメーターである．本研究では $\mathbf{x}_i$ は各楽曲に該当し，13 種類の音響特徴量を持つデータ点である．この音響特徴量を解析するため，$\mathbf{w}$ は 13 次元のベクトルとなっている．データ点 $\mathbf{x}_i$ と，データ点を線形モデル $f(\mathbf{x}_i)$ に正射影した点の残差 $\epsilon_i(\mathbf{x}_i)$ とする．ϵ_i が最小になるように $\mathbf{w}$ を最適化することが解くべき最適化問題となり，次のように表すことができる．

$$\min_{\mathbf{w}} \mathcal{E}(\mathbf{w}) = \frac{1}{N} \sum_i [\epsilon(\mathbf{x}_i)]^2 \tag{2}$$

この最適化問題は，分散共分散行列 $\hat{\Sigma} = \sum_{i=1}^{N} \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_i$ を用いて行列表現にすることができる．

$$\max_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}^{\mathsf{T}} \hat{\Sigma} \mathbf{w} \tag{3}$$

ここで $\mathbf{w}$ は正規直交基底ベクトルであるため，$\mathbf{w}^{\mathsf{T}}\mathbf{w} = 1$ という制約がついている．つまりここでは制約付きの最適化問題を解くことになり，(3) は次のようなラグランジュの双対問題に置き換わる．

$$\max_{\lambda} \hat{\Sigma}\mathbf{w} = \lambda \mathbf{w} \quad (\text{ただし}, \lambda \geqq 0) \tag{4}$$

これを解くことにより得た最大の固有値 λ_1 に対応する固有ベクトル $\mathbf{w}^{(1)}$ が第一主成分を表す主成分軸となる．また，2 番目に大きい固有値と対応する固有ベクトルをそれぞれ $\lambda_2, \mathbf{w}^{(2)}$ とし，これは第二主成分を表す主成分軸である．

さらに，次元削減後のデータによって元のデータがどれほど説明可能であるかについて，寄与率により評価する．第 d 主成分の主成分軸の寄与率は次のように定義される．

$$(\text{寄与率}) = \frac{\lambda_d}{\sum_{\lambda'=1}^{D} \lambda_{d'}} \times 100\,[\%] \tag{5}$$

今 λ_d は (4) の λ のうち第 d 番目に大きな値である．このとき第 d 主成分の主成分軸までの累積寄与率は次のように表す．

$$(\text{累積寄与率}) = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_d}{\sum_{d'=1}^{D} \lambda_{d'}} \times 100\,[\%] \tag{6}$$

3.1.2 実　　装

本研究においては，データ点において楽曲の基本情報と音響特徴量を分離し，標準化した 13 種類の音響特徴量に対して次元削減をおこないデータ点を 2 次元平面にプロットする．

4 解析結果

本章では，次元削減をおこなった散布図に対し，目的変数として popularity を用いて可視化した結果を報告する．また，この結果とアーティストを目的変数として得た結果を比較し，アーティストによる流行の傾向について解析する．さらに，各目的変数と次元削減により得た主成分との相関関係について報告する．

4.1 popularity による相関

本研究では，解析をおこなった 2648 曲のうち，popularity の値が上位 25% 以上の楽曲を流行曲と定義した．popularity とは再生回数に応じた人気度を 0 から 100 で表したものである．そのため，本研究は popularity が 53 以上の楽曲を流行曲，53 未満の楽曲を非流行曲とした．次元削減後の散布図を図 1 に示す．また，(6) によって導出した第 2 主成分までの累積寄与率は 34.9% であり，主成分軸の構成要素を表 2 に示す．

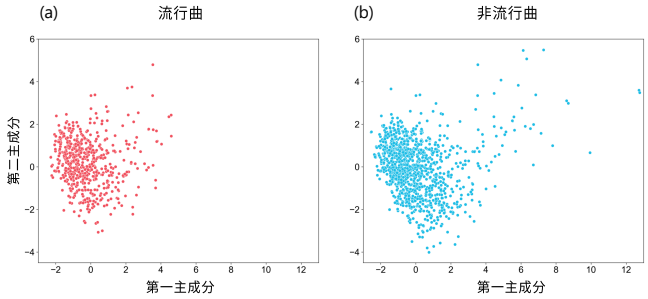


図 1 popularity を目的変数としてプロットした散布図.(a) 流行曲 (popularity が 53 以上)，(b) 非流行曲 (popularity が 53 未満) の結果を示す．

表 2 主成分軸の構成要素

j	楽曲特徴量	$\mathbf{w}_j^{(1)}$	$\mathbf{w}_j^{(2)}$
1	length	-0.012	-0.029
2	acousticness	0.56	-0.078
3	danceability	-0.095	-0.15
4	energy	0.090	-0.46
5	instrumentalness	-0.37	-0.47
6	key	0.11	-0.47
7	liveness	-0.097	0.52
8	loudness	0.024	0.21
9	mode	0.038	0.049
10	speechiness	-0.67	0.0023
11	tempo	0.15	-0.030
12	time signature	-0.20	0.0089
13	valence	0.012	0.018

主成分軸の構成要素は，絶対値が大きいほど主成分へ与える影響が強いと考えて良いものである．そのため，第一主成分は speechiness および acousticness に関連の強い軸であり，第二主成分は liveness，key，instrumentalness，energy に関連の強い軸であることが見てとれる．第一主成分に関連の強い speechiness と acousticness のうち，speechiness は負の値となっており，acousticness は正の値となっている。このことから，第一主成分は値が大きいほど話し言葉から遠く，よりアコースティック感の強い楽曲であることを示す軸であると言うことができる．第二主成分に対しても同様に考えると，値が大きいほどライブ感が強くエネルギッシュでない楽曲であることを示す軸であると言うことができる。これらの軸で構成される 2 次元平面対する流行曲および非流行曲の相関係数を表 3 に示す．

表 3 流行曲および非流行曲の相関係数

	相関係数
流行曲	-0.0312
非流行曲	0.0146

このことから，流行曲は第一主成分と第二主成分に対して負の相関があり，非流行曲は正の相関があることが見てとれる．しかし，流行曲と非流行曲はともに相関係数がほとんどゼロに近い値となっている．しかし，図 1 における 2 つを比較すると，流行曲は非流行曲に比べて特定の範囲に分布していることが見てとれる．よって，主成分分析によって流行曲の流行因子抽出の可能性を示した．

4.2 アーティストによる相関

本節では主成分分析の結果をアーティストごとに抽出することで，楽曲特徴をアーティストごとに示す．なかでも楽曲数の多い 4 アーティストの楽曲についての主成分散布図を図 2 に示す．

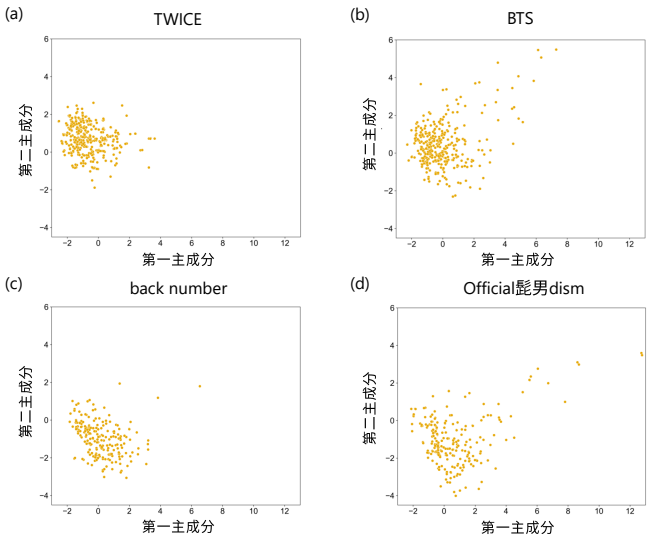


図 2 アーティストを目的変数としてプロットした散布図. (a) TWICE, (b) BTS, (c) back number, (d) Official 髭男 dism.

これらの軸で構成される平面に対するアーティストごとの第一主成分と第二主成分に対する相関係数を算出した結果は表 4 のようになる．

これらの結果からアーティストごとの楽曲特徴を分析することが可能である．図 2 の 4 つについて比較すると，back number

表 4　アーティストごとの相関係数

	アーティスト名	相関係数
(a)	back number	-0.2326
(b)	BTS	0.3593
(c)	TWICE	-0.1545
(d)	Official 髭男 dism	0.3587

および Official 髭男 dism は第二主成分に対して-4 から 1 の範囲に分布している．一方で BTS および TWICE は第二主成分に対して-2 から 2 の範囲に分布していることが見てとれる．つまり，TWICE および BTS は back number および Official 髭男 dism と比べてライブ感の強い楽曲が多いと言うことができる．BTS および Official 髭男 dism は第一主成分に対して広い範囲に分布していることが見てとれる．これは，BTS については speechiness のばらつき，Official 髭男 dism については acoustiness のばらつきの特徴が現れている．また，TWICE と back number は散布図上の狭い範囲に分布している．以上のことから，散布図上での分布の偏りにアーティストの特徴が現れていることがわかる．

5　おわりに

本研究では Spotify が提供する楽曲特徴量について主成分分析によって次元削減を行い流行曲となり得る楽曲の構成因子について解析を行なった．またアーティストごとの楽曲特徴量も分析し，主成分分析によってアーティストの楽曲特徴を可視化する手法について提案を行なった．その結果，今回作成した 2648 曲からなるデータセットにおいては，Spotify が提供する楽曲を特徴づける 13 の変数から主成分を取り出したところ，第一主成分には speechiness と acousticness が，第二主成分には liveness や instrumentalness が関連の強い構成要素となることを示した．流行曲はこれら二つの主成分に対し正の相関が，非流行曲には負の相関が見られるという違いがみられた．また，散布図から流行曲は非流行曲に比べて分布範囲が狭い．これらの結果から主成分分析によって楽曲特徴を示すことは可能であり，また流行曲を特徴づける流行因子に関しての議論も可能であることを示した．アーティストによる相関の解析では，アーティストごとに散布図上での分布の偏りにその特徴が大きく現れることが示唆された．

ただし本研究による研究では第 2 主成分までの累積寄与率が 34.9% にとどまっており，流行因子をより顕著に示すためにも累積寄与率が高くなるように分析することも重要である．また楽曲ジャンルによってその特徴が大きく現れることから，流行曲を特徴づける流行因子についてはより多くのアーティストを含むデータセットを用意することで，各楽曲ジャンルの影響を小さくする必要があることを示した．さらに，流行楽曲は年代ごとに大きく特徴が変遷することも想定でき，今回の解析方法では全年代をまとめて解析を行なったためにその特徴がそれほど顕著に現れなかったとも考えられる．

今後の方針として，年代ごとでの流行曲の変遷が解析可能なデータセットの構築やより多くのアーティストを含むデータセットを構築しての解析が挙げられる．

謝　辞

本研究は，奈良女子大学理学部自然科学考房学生企画プロジェクトの支援を受けて実施した．

文　献

[1] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[2] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58, 2020.

[3] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[4] H. Hotelling. Analysis of a complex of statistical variables into principal components. *ournal of Educational Psychology*, 24(6):417–441, 1933.

[5] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R. Arce. Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):576–588, 2010.

[6] Kai R. Fricke, David M. Greenberg, Peter J. Rentfrow, and Philipp Yorck Herzberg. Computer-based music feature analysis mirrors human perception and can be used to measure individual music preference. *Journal of Research in Personality*, 75:94–102, 2018.

[7] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 274–277, 2016.

[8] Molly J. Henry Alice Vivien Barchet and Johanna M. Rimmele. Auditory-motor synchronization and perception suggest partially distinct time scales in speech and music. *Communications Psychology*, 2(2), 2024.

[9] Kai Middlebrook and Kian Sheik. Song hit prediction: Predicting billboard hits using spotify data, 2019.

[10] Spotify Web API. `https://developer.spotify.com/documentation/web-api`.

[11] 八谷大岳. **ゼロからつくる** *Python* **機械学習プログラミング入門**. 機械学習スタートアップシリーズ. 講談社サイエンティフィック, 2020.

人体キーポイントを用いて映像に整合する楽器音を生成するモデルの提案

岡野　日翔†　　清　　雄一†　　田原　康之†　　大須賀昭彦†

† 電気通信大学　〒 182–8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †okano.haruka@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　映像を撮影した際，音声に大きな雑音が混入してしまうとノイズ除去技術を用いても復元が難しい．近年，一部の生成 AI は映像の視覚情報のみからその部分の音声を予測することが可能になったが，楽器演奏という複雑な動作の映像に対しては予測精度が低いという現状がある．先行研究では SpecVQGAN がコードブック表現や Transformer を用いることで質の高い音声を生成できる．本研究の提案モデルは既存モデルに人体キーポイント特徴を導入することで，楽器音の生成精度を向上させることを目指す．アンケートによる主観評価と 4 つの尺度を用いた客観評価を行った結果，実際の音声と生成した音声の類似性の観点で，提案モデルが既存モデルを上回った．

キーワード　音生成 AI，深層学習モデル，楽器演奏，音声・画像処理，人体キーポイント

1　はじめに

カメラで映像を撮影する際に，周囲の騒音などに配慮している整った環境でなければ，音声中に雑音が混入してしまうことがある．雑音があっても特に気にならない映像もあるが，楽器演奏の映像に関しては音の重要性が高いため，クリアな音質が求められる．小さな雑音であればノイズ除去技術を用いて取り除くことができるが，楽器音を大きく上回る雑音がある場合はノイズ除去が難しい．近年，一部の生成 AI は映像の視覚情報のみからその部分の音を予測することが可能になった [1] [2]．そのような生成 AI は映像に写っている物体や人物の外観や動きとそこから発せられる音の相互関係を学習しており，楽器演奏の映像にも適用可能である．楽器演奏は人間の体の動きと音が大きく関係しており，例えばバイオリンの演奏において，弦を抑える手が小刻みに動いていればビブラートをしていると認識でき，弓を持つ手が細かく動いていればテンポの速い曲を弾いていると想像できる．しかしながら，楽器演奏は複雑な動作であるため既存の機械学習モデル [1] [2] の生成する楽器音は精度が低いという現状がある．

本研究では，無音の楽器演奏の映像を入力すると映像に整合した楽器音を生成するというタスクにおいて，既存手法よりも精度を向上させることを目的とする．整合性が高い楽器音とは，映像に元々付いていたオリジナルの楽器音と類似しており，ビブラートなどの演奏者の動きが音に反映されている生成音のことであると定義した．本研究のモデルは，演奏映像の途中で楽器音が聞こえないほどの騒音やノイズが混入して音が途切れた際に，その部分の音を生成して補間することができる．

本研究で新しく取り入れた点は，映像から音を生成するタスクで従来使われなかった人体キーポイントを用いたことである．人体キーポイントは体の部位や関節の位置を示した特徴点のことであり，画像処理の分野では人間のポーズ推定やジェスチャー認識にしばしば用いられる．楽器演奏という動作は手や腕が素早く動く複雑な動作であるため，映像中でどの位置にどの身体の部位があるのかがわかれば，機械学習モデルが演奏者の動きと楽器音の相互関係を学習しやすいと考えた．人体キーポイントを導入するにあたって，モデルが処理できる適切な形式の特徴に変換して学習を最適化する．

本論文では，2 章で関連研究，3 章で提案手法の説明，4 章で提案手法を用いた実験手法とその結果，5 章で評価，6 章で考察，最後に 7 章で本論文の結論と今後の展望を示す．

2　関連研究

2.1　楽器演奏データセット

映像に整合する楽器音を生成するタスクをモデルに学習させるためには，楽器演奏の映像と楽器音を含んだ音声のペアが必要である．Montesinos らは楽器演奏の映像に人体キーポイント情報を付与した Solos データセットを作成した [3]．13 個の楽器カテゴリにおいて，一人だけが演奏している映像が全 755 本含まれており，映像時間の平均は 5 分 16 秒である．図 1 のように，1 つの人体に付与されるキーポイントは胸，肩，腕，手の付け根に 7 個とそれぞれの手に 20 個ずつで合計 47 個あり，{x 座標，y 座標，信頼度スコア } が情報として与えられる．

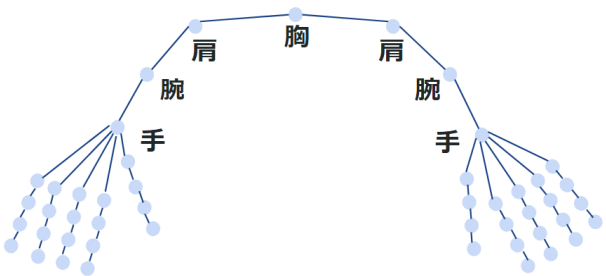


図 1: 人体キーポイントの構造

このデータセットでは，人体キーポイントの注釈をつけるにあたってポーズ推定器として OpenPose が用いられている [4]．

楽器演奏の映像は楽器によって身体の一部が隠れていたり素早い動きが含まれているので，通常よりも推定が難しく，キーポイントが誤検出されてしまうことがある．そこで，この研究では，誤検出された部分のキーポイントを補間する手法を提案している．ある時刻 t で誤検出されたキーポイント $\hat{J}_t^i$ があったとき，時刻 t までで最後に正しく検出されたキーポイントを $\hat{J}_{t-n+1}^i$，時刻 t 以降で最初に正しく検出されたキーポイントを $\hat{J}_{t+m-1}^i$ とすると，キーポイント列は次のようになる．

$$\{J_{t-n}^i, \hat{J}_{t-n+1}^i, \ldots, \hat{J}_t^i, \ldots, \hat{J}_{t+m-1}^i, J_{t+m}^i\} \tag{2.1}$$

このとき，誤検出されたキーポイント $\hat{J}_t^i$ を補間する式は次のように表される．

$$\hat{J}_t^i = \frac{mJ_{t-n}^i + nJ_{t+m}^i}{m+n} \tag{2.2}$$

この補間手法によって，外れ値の少ない人体キーポイント情報を付与した楽器演奏データセットを提供している．

2.2 映像に整合する音の生成

Chen らは犬や花火，ドラムなどの全般的なカテゴリを対象として映像に整合する音を生成するモデル RegNet を開発した [1]．モデルの構造を図 2 に示す．

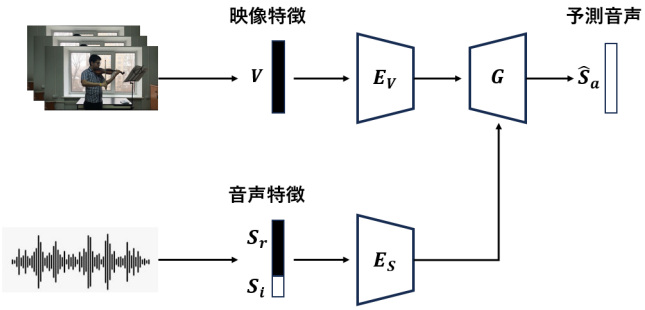


図 2: RegNet の構造

予測音のスペクトログラム $\hat{S}_a$ を生成器 G で生成するにあたって，映像の色情報を示す RGB 特徴，物体の動きを示すオプティカルフロー特徴，音声信号の周波数成分を捉えたメルスペクトログラムを含む音声特徴の 3 つの特徴表現が用いられる．各特徴はエンコーダ E_V, E_S によってモデルの入力に適した形式に変換される．この研究では音声信号を，映像に映っているオブジェクトから発せられる音 S_r とそれ以外の音 S_i の 2 つに区別して考えている．もしも S_r と S_i を考慮せずに映像特徴のみを用いて音を生成した場合，映像中に映っていないオブジェクトから発せられた音にもかかわらず映像中のオブジェクトと関連付けられて学習されてしまう恐れがある．そのため，学習する際に S_r と S_i を区別してモデルに入力することで，映像特徴のみへの依存を減らして誤学習を抑制している．

誤学習を防ぐ工夫によって予測音のクオリティが向上したが，このモデルにはデメリットが 2 つある．1 つ目が，映像の視覚情報のみから音を予測するというタスクであるにも関わらず，モデルの学習の際に音声情報も用いているということである．教師あり学習は音を生成した後に予測音と正解の音を比較して学習する手法であるが，それとは異なり，このモデルでは音を生成する前に正解の音をモデルに入力しており，予測対象の音を事前に知っている状態でその音を予測することに等しいため，良くない学習方法である．2 つ目が，スペクトログラムから音声波形に変換するボコーダとして WaveNet [5] を用いており，10 秒程度の音を生成するために 30 分程度の時間を要し，膨大な計算リソースも必要となる点である．映像に整合する音を生成することは処理が重いタスクであるため，より低リソースで動作する効率的なモデルが求められる．

2.3 Transformer を用いた映像に整合する音の生成

lashin らは VAS [1]，VGGSound [6] という映像と音声を含む大規模データセットを対象として映像に整合する音を生成するモデル SpecVQGAN を開発した [2]．モデルの構造を図 3 に示す．

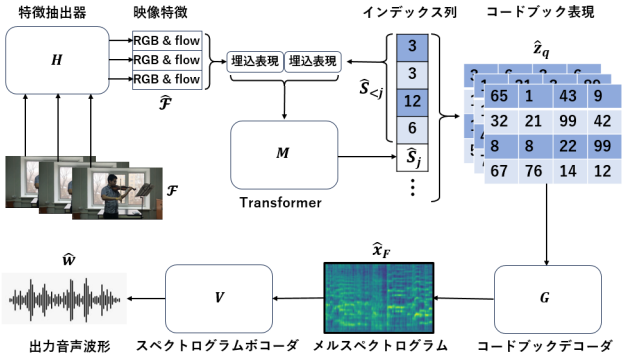


図 3: SpecVQGAN の構造

このモデルは RegNet [1] をベースに開発されているため，RGB 特徴とオプティカルフロー特徴を映像特徴として用いている点が同様である．目新しい点は，コードブック表現と Transformer [7] を用いたことである．コードブック表現とは，学習データをクラスタリングした後に，似ているデータ構造はエントリという 1 つの特徴表現にまとめて，そのエントリを用いてデータを再構築する手法である．コードブック内のエントリの位置はインデックスで示される．また，Transformer の自己注意機構により，コードブック内の各エントリが他のエントリとの関連性や重要性を捉えることができ，自然な時系列データを生成できる．

SpecVQGAN の詳しい構造については第 3 章にて説明をする．

このモデルは RegNet と比較して優れた点が 2 つある．1 つ目が，コードブック表現を用いているのでサンプリングリソースが大幅に減少しており，10 秒程度の音の生成に CPU 上で 2 秒程度しかかからないということである．2 つ目が，Transformer の自己注意機構が組み込まれているので時系列データの生成のクオリティが高いことである．このことから，SpecVQGAN は RegNet が従来持っていた課題を解決した．

2.4 人体キーポイントを用いた音源分離

Gan らは複数の楽器音が混ざった音源を個別の楽器音に分離

するという音源分離タスクにおいて，人体キーポイント特徴をモデルに導入した [8]．演奏者の姿勢や手の動きと楽器音を関連付けたことで，異なる種類の楽器だけでなく同じ種類の楽器の音源分離に成功し，従来手法よりも分離精度を向上させた．この研究では人体キーポイントの有効性を実証し，他のタスクへの適用可能性を示した．

2.5 人体キーポイントを用いた人間のアクション認識

Peng らは人体キーポイントから空間的及び時間的パターンを自動的に学習するモデル Spatial Temporal Graph Convolutional Networks(ST-GCN) を提案して人間のアクション認識を行った [9]．人間の骨格をグラフとして扱っており，人間の関節や体の部位がノードに対応している．グラフの空間的なエッジは関節の自然な接続性に従っており，例えば胸は肩，肩は腕，腕は手とそれぞれつながっている．グラフの時間的なエッジは，時間的に連続したいくつもの人体キーポイントのデータがある中で，連続する時間ステップで同じ関節を接続している．これらのノードやエッジの軌跡のパターンを学習することで，外観情報やオプティカルフローのみからアクション認識をするモデルに比べて認識精度を大幅に向上させた．

3 アプローチ

機械学習モデルが楽器演奏の映像と楽器音の相互関係を学習しやすくするために，演奏者の身体部位の位置を示した人体キーポイントを導入する．提案モデルは SpecVQGAN [2] をベースに開発し，処理する映像特徴として RGB 特徴とオプティカルフロー特徴のほかに人体キーポイント特徴を新たに追加する．人体キーポイント特徴は，Solos データセット [3] に含まれる人体キーポイントデータを ST-GCN [9] で処理して計算する．

3.1 提案モデルの構造

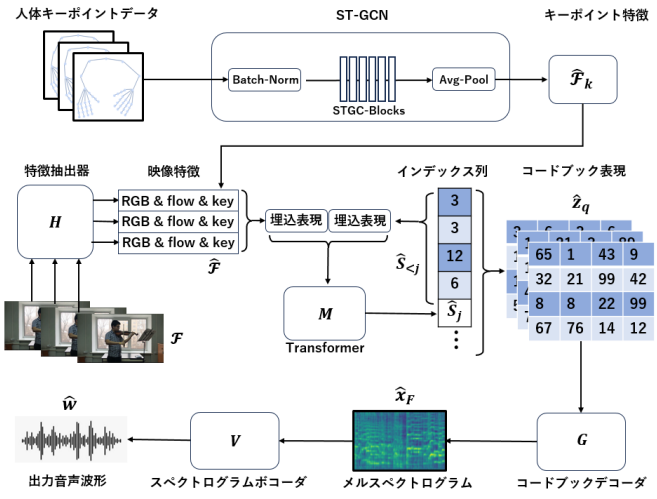


図 4: 提案モデルの構造

図 4 が提案モデルの構造を示している．ST-GCN で計算したキーポイント特徴は SpecVQGAN が元々扱っていた映像特徴 $\hat{F}$ に連結された後に，埋め込み表現に変換されて Transformer に入力される．

3.2 音声生成の流れ

RGB 特徴，オプティカルフロー特徴，キーポイント特徴は映像の総フレーム数を N，フレームの番号を i とすると，まとめて $\mathcal{F} = \{f_r^i, f_o^i, f_k^i\}_{i=1}^N$ と表される．キーポイント特徴の計算方法については次の節で説明を行う．$\mathcal{F}$ は特徴抽出器 H によって，$\hat{\mathcal{F}} = \{\hat{f}_i\}_{i=1}^N \subset \mathbb{R}^{D_r+D_o+D_k}$ に変換される．D_r, D_o, D_k はそれぞれ RGB 特徴，オプティカルフロー特徴，キーポイント特徴のベクトル次元数を示している．この $\hat{\mathcal{F}}$ を Transformer に入力すると，そのフレームに対応するコードブックインデックスが出力される．入力は $\hat{\mathcal{F}}$ の 1 つだけでなく，それまでに計算されたコードブックインデックス $\hat{s}_{<j} = (\hat{s}_1, \hat{s}_2, ..., \hat{s}_{j-1})$ も含まれており，出力 s_j は次の式で表される．M は Transformer が行う処理を示す．

$$p\left(s_j | \hat{s}_{<j}, \hat{\mathcal{F}}\right) = M\left([\hat{\mathcal{F}} : \hat{s}_{<j}]\right) \tag{3.1}$$

また，このモデルでは音情報をメルスペクトログラムで扱っており，スペクトログラム上の周波数を F，時間を T とすると，出力音は $\hat{x} \in R^{F \times T}$ と表される．このメルスペクトログラムをコードブック表現に置き換えたとき，$\hat{x}$ は周波数が F'，時間が T'，エントリの次元が n_z で表される $\hat{z}_q = E(\hat{x}) \in \mathbb{R}^{F' \times T' \times n_z}$ にエンコードされる．したがって，コードブックインデックス s_j は $F' \times T'$ 個のエントリからなるコードブックのアドレスを指し示している．コードブックインデックスをエントリに置き換えて生成されたコードブック表現をコードブックデコーダに入力すると，メルスペクトログラム $\hat{x} = G(\hat{z}_q)$ が出力される．この構築されたメルスペクトログラムは MelGAN [10] というボコーダを用いて，音声波形 $\hat{w}$ に変換される．

3.3 キーポイント特徴の計算方法

ST-GCN [9] には，人体キーポイントのグラフに空間的及び時間的な畳み込みを行う STGC 層が含まれている．T を時間，N を人体のキーポイントの数とすると，グラフ $G = (V, E)$ において，ノードは $V = \{v_{ti} \mid t = 1, \ldots, T, i = 1, \ldots, N\}$，エッジは空間方向では $E_S = \{v_{ti}v_{tj} \mid (i, j) \in H\}$（$H$ は接続しているノードの組の集合），時間方向では $E_T = \{v_{ti}v_{(t+1)i}\}$ と表される．このとき，グラフのあるノードに畳み込みを行う場合は次の式が用いられる．

$$f_{out}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} \cdot f_{in}(v_{tj}) \cdot w(v_{ti}, v_{tj}) \tag{3.2}$$

上式において，時間的または空間的に畳み込むかどうかは $B(v_{ti})$ の定義によって決まる．$B(v_{ti})$ は畳み込み対象である v_{ti} の周辺ノードの集合であり，画像畳み込みにおけるカーネルを適用する範囲に等しい．$w(v_{ti}, v_{tj})$ は $B(v_{ti})$ に適用するカーネルの重みを示している．$Z_{ti}(v_{tj})$ は v_{tj} の周辺ノードの数であり正規化項である．時間畳み込みでは $B(v_{ti})$ は次のように定義される．

$$B(v_{ti}) = \{v_{tj} \mid d(v_{tj}, v_{ti}) \leqq D\} \tag{3.3}$$

D はノード間の距離を示している．また，空間畳み込みでは $B(v_{ti})$ は次のように定義される．

$$B(v_{ti}) = \{v_{qj} \mid d(v_{tj}, v_{ti}) \leqq K, |q-t| \leqq \frac{\Gamma}{2}\} \tag{3.4}$$

K はノードに割り当てたラベルの数，Γ は時間方向のカーネルサイズを示している．このように，STGC 層で畳み込みが行われるときは各ノードに対して空間と時間の 2 つの方向で畳み込みが行われる．

次に，ST-GCN のネットワークアーキテクチャを説明する．最初に，入力の人体キーポイントデータはバッチ正規化層に送られ正規化される．次に，6 つの層からなる STGC 層に送られる．最初の 3 つの層は出力に 32 チャンネルを持ち，次の 3 つの層は出力に 64 チャンネルを持っており，各層の出力に対して 0.5 の確率で特徴ベクトルの要素をランダムにドロップアウトして過学習を防いでいる．4 番目の STGC 層のストライドは 2 に設定しており，プーリング層として機能している．最後の STGC 層から出力されたテンソルに対してグローバルプーリングを実行して，64 次元の特徴ベクトルを得る．最後に特徴ベクトルを SoftMax 分類器に入力して，該当するアクションクラスの番号を出力する．本研究では，Solos データセットの人体キーポイントを入力として演奏している楽器のクラスを当てるというタスクを行い，十分学習できたネットワークを用いて SpecVQGAN に入力する人体キーポイント特徴を計算する．

3.4 コードブック及びモデルの訓練

提案モデルを訓練する際の損失関数は SpecVQGAN [2] で提案されたものを用いた．

3.4.1 コードブックの訓練

コードブックの訓練は学習データを使って行われる．最初に，学習データのメルスペクトログラム x は $\hat{z} = E(x) \in \mathbb{R}^{F' \times T' \times n_z}$ にエンコードされる．ここで，$\hat{z}$ は K 個にクラスタリングされ，エントリの集合 $Z = \{z_k \mid 1 \leqq k \leqq K\} \subset \mathbb{R}^{nz}$ ができてコードブックが構築される．その後，以下の式を満たすようにコードブックは更新される．

$$z_q = q(\hat{z}) := \arg\min_{z_k \in Z} ||\hat{z}_{ft} - z_k|| \quad \text{for all } (f,t) \in (F' \times T') \tag{3.5}$$

この式は，学習データをエンコードしたエントリとコードブックに含まれるエントリの差異が小さくなるようにコードブックが訓練されることを示している．計算されたコードブック表現 z_q ををコードブックデコーダに入力すると，メルスペクトログラム $\hat{x} = G(z_q) = G(q(E(x)))$ が出力される．生成されたメルスペクトログラムが実際の音のメルスペクトログラムに近くなるように学習するために，次の損失関数が用いられる．

$$\mathcal{L}_{\text{VQVAE}} = \|\hat{x}, x\| + \frac{1}{2}\|E(x) - \text{sg}[z_q]\|_2^2 + \frac{\beta}{2}\|\text{sg}[E(x)] - z_q\|_2^2 \tag{3.6}$$

β は正規化項である．sg は逆伝搬の際に勾配を伝搬しないようにする操作のことで，パラメータが更新されないことを示している．

3.4.2 モデル全体の訓練

このモデル全体の損失関数には，前節で記した $\mathcal{L}_{\text{VQVAE}}$ に加えて敵対的生成ネットワーク（GAN）の損失関数と Google が提供する深層学習モデル VGGish [11] を用いた知覚的損失を導入している．

$$\mathcal{L}_{\text{SpecVQGAN}} = \mathcal{L}_{\text{VQVAE}} + \log D(x) + \log(1 - D(\hat{x})) + \sum_{s=1} \frac{1}{F^s T^s}\|x^s - \hat{x}^s\|_2^2 \tag{3.7}$$

第 2，第 3 の項が GAN の損失関数である．生成器は生成したメルスペクトログラム $\hat{x}$ が識別器に偽物だと見破られないように，識別器は生成器の出力したメルスペクトログラムを偽物だと見破れるように学習している．第 4 の項が VGGish の損失関数である．VGGish はメルスペクトログラムを VGGish を入力すると，スケール s のフィルターや層を用いて異なる時間単位 T^s または周波数帯域 F_s の特徴を抽出する．本物の音のメルスペクトログラム x と生成したメルスペクトログラム $\hat{x}$ から VGGish で特徴を抽出したときに，その特徴間の差異が小さくなるように学習している．

4 楽器音生成実験

提案モデルが生成した楽器音の精度が既存モデルよりも高いことを示すため，6 個の楽器カテゴリの楽器音を生成して評価を行った．

4.1 データセット

Solos [3] データセットには 13 個の楽器カテゴリがあるが，その中から音の生成対象として 6 個の楽器カテゴリを選択した．その楽器名，映像本数，動画時間の平均を表 1 に示す．

表 1: データセットの内訳

楽器名	映像本数	平均時間
バイオリン	66	6:16
チェロ	50	7:21
トロンボーン	50	5:03
クラリネット	49	3:23
フルート	48	4:00
ホルン	50	5:11

この 6 楽器を選んだ理由は 2 つある．1 つ目が，Solos データセットの中で映像及びキーポイントデータが十分な量ある楽器カテゴリだったことである．理由のもう 1 つが，バイオリンとチェロ，トロンボーンとホルン，クラリネットとフルートはそれぞれ弦楽器，金管楽器，木管楽器に該当するので，楽器の種類のバランスが良く，似た種類の楽器でも音色を区別できるのかを調べるためである．

4.2 前 処 理

提案モデルに入力できる映像は長さが 10 秒，映像フレームレートが 21.5fps，音声フレームレートが 22050Hz が条件である．一方で，Solos データセットの映像は長さが不均一，映像

表 2: 前処理後のデータセットの内訳

楽器名	映像本数	平均時間
バイオリン	2198	0:10
チェロ	5764	0:10
トロンボーン	1451	0:10
クラリネット	1065	0:10
フルート	1119	0:10
ホルン	1297	0:10

フレームレートが 25fps，音声フレームレートが 44100Hz である．したがって，Solos データセットの各映像を 215 フレームずつ切り出し，映像及び音声フレームレートをそれぞれ 21.5fps，22050Hz に変換した．前処理後の各楽器カテゴリの映像本数は表 2 のようになった．

4.3 特徴抽出

映像の各フレームの画像に対して RGB 特徴，オプティカルフロー特徴，キーポイント特徴を抽出した．RGB 特徴とオプティカルフロー特徴の抽出には，ImageNet データセットで事前学習された BN Inception モデル [12] を用いた．キーポイント特徴の抽出には，事前学習された ST-GCN [9] を用いた．ST-GCN は 1 つのキーポイント特徴を抽出するにあたって複数フレームのキーポイントの入力が必要であるため，あるフレームにおけるキーポイント特徴を抽出する際に前後 7 フレームのキーポイントも同時に入力した．3 つの映像特徴をまとめると $\hat{\mathcal{F}} = \{\hat{f}_i\}_{i=1}^{N} \subset R^{D_r+D_o+D_k}$ と表されるが，本実験では $D_r = 1024$，$D_o = 1024$，$D_k = 64$ とした．

4.4 ST-GCN の訓練

4.4.1 データセット

ST-GCN の訓練に用いた学習データは，前処理後のデータセットに含まれているキーポイントデータである．1 個の学習データは 15 フレーム分のキーポイントから構築されていて，各楽器カテゴリから訓練データとして 5120 個，テストデータとして 1280 個を選択した．

4.4.2 実験設定

ST-GCN はキーポイントデータが示す人間の動作から演奏している楽器を推定するので，(バイオリン, チェロ, トロンボーン, クラリネット, フルート, ホルン) のうちいずれかを出力する．学習は epoch 数 = 30，バッチサイズ = 8，最適化手法 = SGD，学習率 = 0.01 で行った．

4.4.3 実験結果

30epoch 学習した後の ST-GCN が，テストデータに対して演奏楽器認識を行った結果の混合行列を図 5 に示す．

縦軸が正解の楽器カテゴリ，横軸が ST-GCN が予測した楽器カテゴリを表していて対角行列の青色が濃いほど予測精度が高いことを示している．演奏楽器を正しく認識した精度は 70.560% であった．図 5 によるとフルート，ホルン，トロンボーンの認識精度が高い一方でチェロ，クラリネット，バイオリンの認識精度が低いことがわかる．チェロとバイオリンはどちらも弦楽器であり，左手で弦を抑えて右手で弓を弾くという点が同じであるためモデルが間違いやすい．また，クラリネットは両手を胴体の前に持ってきて指を動かす演奏動作であるが，チェロとバイオリンも両手を胴体を前に持ってきて演奏を行うため誤認識が多かったと考えられる．

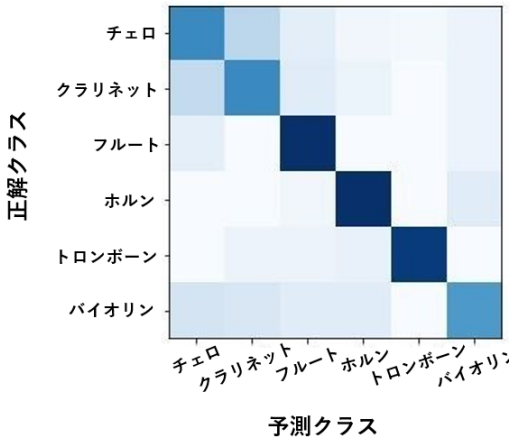


図 5: ST-GCN の混同行列

4.5 提案モデルの実験設定

提案モデルの学習では，最初にコードブックの訓練を行い，その事前学習済みコードブックを用いて Transformer を訓練した．コードブックの学習は epoch 数 = 50，バッチサイズ = 1，最適化手法 = Adam，学習率 = $1e^{-6}$ で行った．Transformer の学習は epoch 数 = 80，バッチサイズ = 2，最適化手法 = Adam，学習率 = $1e^{-6}$ で行った．学習データのうち各楽器カテゴリから 64 個ずつ取り出したものをテストデータ，それ以外をトレーニングデータとして分割した

4.6 提案モデルの実験結果

Transformer の学習の推移を図 6 に示す．

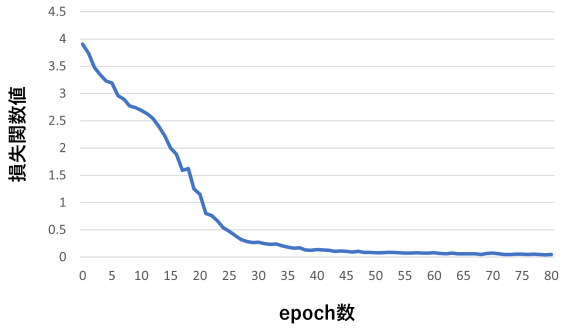


図 6: 損失関数値の推移

図 6 より 70epoch 周辺では損失関数値がほとんど変動しなくなっているため，70epoch 時点での十分学習させたモデルを用いてメルスペクトログラム及び楽器音の生成を行った．

図 7 はメルスペクトログラムの生成結果の一例であり，チェロとホルンの楽器映像を入力したときの生成メルスペクトログラム及び本物のメルスペクトログラムを示している．本物の楽

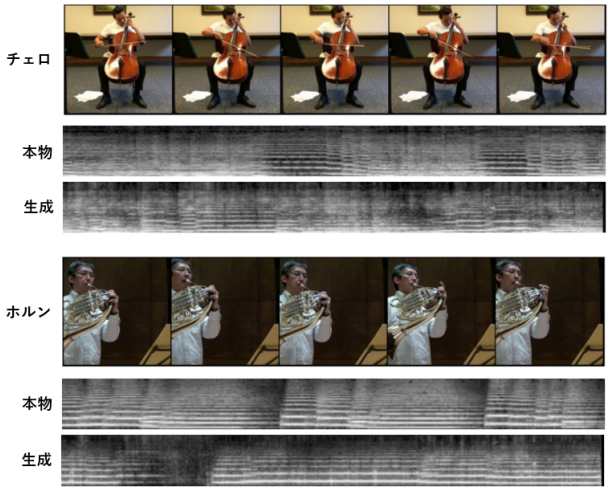


図 7: メルスペクトログラムの生成結果

器音のメルスペクトログラム上では細長い線のように強調された帯域が現れており，チェロは線が細くホルンは線が太いため，楽器によって周波数別の線の太さが異なっている．また，ぼんやりと霞んでいる弱い帯域も現れており，チェロでは真ん中の部分，ホルンは上の部分で見られるので，楽器によって霞んでいる部分が現れる周波数帯が異なっている．それらの楽器別のメルスペクトログラムの特徴は生成メルスペクトログラムにも確認することができるので，提案モデルは楽器の音色の特徴を捉えて学習できている．

5 評　　価

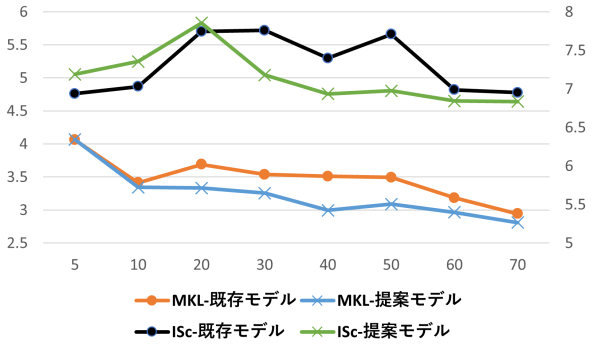
5.1 客観評価

70epoch 学習を行った提案モデルと既存モデルである SpecVQGAN [2] にテストデータを入力して生成されたメルスペクトログラムに対して，ISc [13]，MKL [2]，FID [14]，KID [15] の 4 つの尺度を用いて客観評価を行った．ISc はメルスペクトログラムの多様性を測る尺度，MKL は Melception モデル [2] がメルスペクトログラムから抽出した特徴の分布間の距離を測る尺度，FID はメルスペクトログラムの分布間の距離を測る尺度，KID は Inception モデル [16] がメルスペクトログラムから抽出した特徴の分布間の距離を測る尺度である．ISc の値は高いほど生成されたメルスペクトログラムの多様性が高く，MKL，FID，KID の値が低いほど生成されたメルスペクトログラムが本物のメルスペクトログラムと類似していることを示している．表 3 に提案モデルと既存モデルのサンプルを評価した結果を示す．

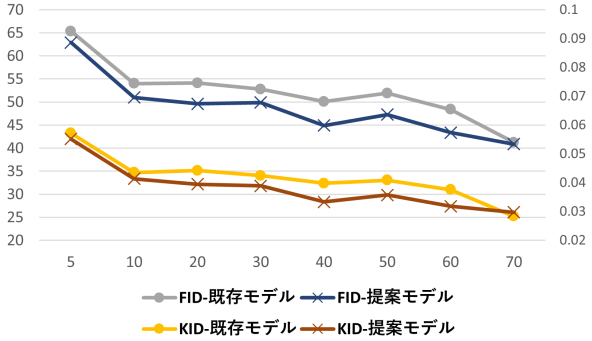
表 3: 70epoch 学習したモデルの尺度評価

	提案モデル	既存モデル
ISc(↑)	6.84	**6.95**
MKL(↓)	**2.81**	2.94
FID(↓)	**40.86**	41.22
KID(↓)	0.029	**0.028**

また，5epoch，10epoch，20epoch，30epoch，40epoch，50epoch，60epoch，70epoch 時点での 2 つのモデルが生成したメルスペクトログラムに対して尺度評価を行った結果を図 8 に示す．図 8(a) において第 1 縦軸は MKL の目盛り，第 2 縦軸は ISc の目盛り，図 8(b) において第 1 縦軸は FID の目盛り，第 2 縦軸は KID の目盛りを示している．



(a) MKL と ISc のスコア推移



(b) FID と KID のスコア推移

図 8: 尺度のスコア推移

表 3 より提案モデルは MKL と FID で既存モデルより高いスコアを出しているが，70epoch 学習された 2 つのモデルにはっきりとした差が見られない．しかしながら，図 8 に示される各尺度のスコアの推移をみると，2 つのモデルの差が確認できる．FID，MKL，KID の尺度において，提案モデルの折れ線グラフは既存モデルの折れ線グラフよりも下に位置している．メルスペクトログラムの類似度の点では，提案モデルの方が既存モデルよりも全体的に優れていることがわかる．一方，ISc の尺度に着目すると 20epoch 付近まで提案モデルの方が上回っていたが，その後は既存モデルの方がスコアが大きいままであった．

表 3 は楽器の区別をせずに全サンプルを評価しているため，楽器別の尺度評価も行った．評価対象のサンプルは 70epoch 学習済みの提案モデルから生成されたものである．結果を表 4 に示す．Vi=バイオリン，Ce=チェロ，Tr=トロンボーン，Cl=クラリネット，Fl=フルート，Ho=ホルンである．

メルスペクトログラムの類似度を示す MKL，FID，KID の 3 つの尺度に着目すると，チェロは 6 つの楽器の中で最も良く学習されており，クラリネットやフルートは尺度のスコアが悪

表 4: 楽器別の尺度評価

	Vi	Ce	Tr	Cl	Fl	Ho
ISc(↑)	2.95	2.35	2.57	3.78	**5.34**	1.82
MKL(↓)	2.81	**2.18**	2.65	2.93	6.78	2.81
FID(↓)	51.77	**45.35**	54.72	75.20	68.86	49.43
KID(↓)	0.054	**0.053**	0.061	0.080	0.059	0.058

く学習があまりうまくいっていない．一方で生成メルスペクトログラムの多様性を示す ISc の尺度に着目すると，クラリネットやフルートのスコアが良く，チェロのスコアが悪かった．

5.2 主観評価

Google フォームにてアンケートを作成し，大学生 14 人（うち楽器経験者 7 人）に提案モデルと既存モデル（SpecVQGAN）の生成した楽器音を聞いてもらい，以下の観点でどちらが優れているか回答してもらった．

- 本物の楽器音と類似している．
- 楽器音の品質が高い．
- 映像と整合している．(例えば，演奏者がやさしく弾いているときに，楽器音も弱くなめらかだったら映像と楽器音が整合していると判断できる．)
- 楽器音として聞きやすく好みである．

各楽器で 2 つのモデルから 1 つずつサンプルを生成しており，1 サンプルの評価にあたって最初にオリジナルの楽器演奏の映像を視聴した後に，生成した楽器音を合成した映像をランダムな順番で 2 つ視聴してもらい回答をする形式になっている．アンケートでは公平を期すため，生成した楽器音がどちらのモデルにより生成されているかを伏せている．回答結果を表 5 に示す．回答の選択項目として「どちらでもない」という項目を設けたが，その割合は下の表から省いている．

表 5: アンケート結果

	本物との類似度		音の品質		映像との整合性		全体的な好み	
	提案	既存	提案	既存	提案	既存	提案	既存
Vi	**52.4**	33.3	**57.1**	28.6	23.8	**47.6**	**57.1**	23.8
Ce	**57.1**	33.3	**42.9**	33.3	38.1	**47.6**	**52.4**	42.9
Tr	**71.4**	19.0	47.6	47.6	**66.7**	19.0	**66.7**	23.8
Cl	**85.7**	4.8	**95.2**	0	**71.4**	14.3	**85.7**	4.8
Fl	**66.7**	33.3	**66.7**	23.8	**61.9**	28.6	**76.2**	19.0
Ho	**66.7**	14.3	**57.1**	33.3	**57.1**	23.8	**66.7**	23.8
平均	**66.7**	23.0	**61.1**	27.8	**53.2**	30.2	**67.5**	23.0

アンケートの回答項目のうち多くの項目において，既存モデルよりも提案モデルの方が高く評価されている．一方で，バイオリンの項目では既存モデルの方が音の品質が高いと判断され，チェロの項目では既存モデルの方が映像との整合性が高く，全体的に好みだとされた.「本物との類似度」の項目では，バイオリン以外の部分で既存モデルを大きく上回っているため，提案モデルの方が学習がうまくいっており生成音の精度が高いことを示している．

6 考　　察

客観評価において，表 3 及び図 8 で ISc に着目すると既存モデルの方が提案モデルよりも良いスコアを出していることがわかる．ISc は高いほどデータが多様であることを示しているが，高い方が必ずしも良いというわけではない．その理由として，学習が進んだモデルほど演奏動作に対する楽器音のサンプリングパターンが定着するため音色や強調される帯域が絞られる一方で，学習があまり進んでいないモデルは似たような演奏動作に対して様々な出力をしてしまい結果的に品質が悪くても多様性があると捉えられてしまうからである．このことは楽器別の尺度評価の結果にも表れており，チェロは学習データが非常に多く 3 つの尺度で最も良いスコアを出しているにもかかわらず ISc は最も悪いスコアになっている．一方で，クラリネットとフルートは学習データが少なく 3 つの尺度で低位のスコアを出しているにもかかわらず，ISc が上位のスコアになっていた．したがって，学習の進度と ISc のスコアの高さは反比例の関係にあるといえる．

図 8 の尺度のスコア推移は提案モデルの学習の進捗が早いことを示しており，これはキーポイント特徴を導入したことが要因であると考えられる．楽器の演奏動作を表す人体キーポイントを入力することでモデルが演奏者の体の動きと楽器音の関係をとらえやすくなっており，RGB 特徴とオプティカルフロー特徴に加えて重要な視覚情報が提供されている．しかしながら，これはキーポイント特徴を抽出した ST-GCN の楽器認識精度が 70.560%と比較的高精度だったため有効であったとすると，楽器認識の精度が低い ST-GCN を用いた場合はモデルの生成精度が悪くなる可能性があると考えた．この仮説を検証するため，わざと楽器認識精度が低くなるように訓練した ST-GCN から抽出されたキーポイント特徴を用いて提案モデルの尺度評価を行った．ST-GCN がテストデータに対して楽器認識を行った結果の混合行列を図 9 に示す．

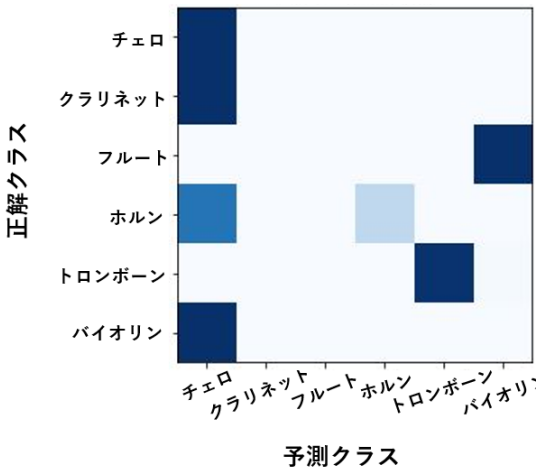


図 9: 低精度 ST-GCN の混同行列

この ST-GCN が演奏楽器を正しく認識した精度は 37.5%であったため仮説を検証するのに十分精度が低いといえる．入力

するキーポイント特徴以外の条件を一切変えずに，70epoch 時点での提案モデルの尺度評価の結果を比較した．

表 6: 70epoch 学習したモデルの尺度評価

	高精度	低精度
ISc(↑)	6.84	**7.32**
MKL(↓)	**2.81**	3.07
FID(↓)	**40.86**	46.74
KID(↓)	**0.029**	0.037

表 6 より ST-GCN が高精度のモデルの方が MKL，FID，KID の尺度で低精度のモデルよりも上回っているため，よく学習できていることがわかる．このことから，ST-GCN の楽器認識精度が生成モデルの精度に影響し，ST-GCN の精度が低いと生成モデルの精度も低くなってしまうことが立証された．

7 ま と め

無音の楽器演奏の映像を入力して，その映像に整合する楽器音を生成するタスクにおいて従来用いられなかった人体キーポイントを導入した．楽器音の生成に用いられるキーポイント特徴は，Solos データセットに含まれる人体キーポイントデータを用いて訓練した ST-GCN によって抽出した．提案モデルは既存モデルである SpecVQGAN よりもメルスペクトログラムの類似性が高く，学習の進捗が早かったため，人体キーポイントの有効性が確かめられた．

今後の展望として，楽器演奏映像に対する自動キーポイント推定，ST-GCN の高精度化が挙げられる．今回の実験で用いた人体キーポイントは Solos データセットによって提供されたものであったため，データセット外の映像に対して提案モデルで楽器音を生成しようとする際にキーポイント推定を行う必要がある．したがって，未知の映像に対して正確な人体キーポイント推定を行いキーポイント特徴に変換するというパイプラインを確立する必要がある．また，考察の仮説立証実験より ST-GCN の精度を向上させた方がモデルの精度も高くなることがわかった．そこで，現在の精度が約 70%の ST-GCN に対してネットワーク構造やパラメータの工夫による改良を行うことで，さらなる提案モデルの精度向上を望むことができる．

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです．また本研究は，電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです．

文　献

[1] M. Tan P. Chen, Y. Zhang et al, “Generating Visually Aligned Sound From Videos,” in IEEE Transactions on Image Processing, vol. 29, pp. 8292-8302, 2020.
[2] V. Iashin, E. Rahtu, “Taming Visually Guided Sound Generation,” British Machine Vision Conference (BMVC), 2021.
[3] J. F. Montesinos, O. Slizovskaia and G. Haro, “Solos: A Dataset for Audio-Visual Music Analysis,” 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2020.
[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[5] A. V. D. Oord, S. Gan, H Zen et al, “WaveNet: A Generative Model for Raw Audio,” arXiv:1609.03499, 2016.
[6] H. Chen, W. Xie, A Vedaldi et al, “VGGSound: A Large-scale Audio-Visual Dataset,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.
[7] A. Vaswani, N. Shazeer, N. Parmar et al, “Attention is All you Need,” NeurIPS, 2017.
[8] C. Gan, D. Huang, H. Zhao et al, “Music Gesture for Visual Sound Separation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[9] W. Peng, J. Shi and G. Zhao et al, “Spatial Temporal Graph Deconvolutional Network for Skeleton-Based Human Action Recognition,” IEEE Signal Processing Letters, 2021.
[10] K. Kumar, R. Kumar, T. de Boissiere et al, “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” NeurIPS, 2019.
[11] P. Esser, R. Rombach, B. Ommer, “Taming transformers for highresolution image synthesis,” CVPR, 2021.
[12] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” International Conference on Learning Representations, 2015.
[13] T. Salimans, I. J. Goodfellow, W. Zaremba et al, “Improved Techniques for Training GANs,” NeurIPS, 2016.
[14] M. Heusel, H. Ramsauer, T. Unterthiner et al, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” NeurIPS, 2017.
[15] M. Binkowski, D. J. Sutherland, M. Arbel et al, “Demystifying MMD GANs,” ICLR, 2018.
[16] C. Szegedy, V. Vanhoucke, S. Ioffe et al, “Rethinking the Inception Architecture for Computer Vision,” CVPR, 2016.

楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式

佐伯 綾亮† 岡田 龍太郎† 峰松 彩子† 中西 崇文†

†武蔵野大学データサイエンス学部 〒135-8181 東京都江東区有明 3-3-3
E-mail: †s2222015@stu.musashino-u.ac.jp, ryotaro.okada@ds.musashino-u.ac.jp, ayako.minematsu@ds.musashino-u.ac.jp, takafumi.nakanishi@ds.musashino-u.ac.jp

あらまし 本稿では，楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式について示す．初心者が作曲を行う際に，メロディだけを作ることができても伴奏を構成することが難しい場合が多い．本方式では，ピアノ曲において，類似するメロディには類似する伴奏を付けることが出来ると仮定した上で，メロディの音響特徴を抽出した上で，その類似度が高い楽曲から伴奏を引用する形でメロディに伴奏付けを行う．本方式では事前に既存曲から一定区間分のメロディと伴奏の組を取り出した知識ベースを用意する．知識ベースのメロディと入力されたメロディからいくつかの特徴量を抽出し，入力されたメロディの特徴量と類似度が高いメロディを検索し，それに付随する伴奏を選択する．さらに，入力されたメロディからコードを推定し，選択された伴奏をコードに合わせて変形することで,伴奏の楽譜を自動的に構成する.本方式を実現することにより，メロディを入力としてピアノ曲の伴奏パートを自動的に構成することが可能となる.

キーワード 自動伴奏付け，音響特徴量，音声・音楽

1. はじめに

近年，DTM の普及に伴う音楽製作や，独自のアレンジを行うユーザが増加している．中でも特に，ピアノ曲の製作やピアノアレンジの演奏を対象として，メインのメロディに伴奏をつけたいという要求が存在する.

一般的に，ピアノ曲の製作やアレンジの演奏をする場合，メインのメロディと，伴奏を同時に作成することが多い．特に，伴奏には，メインのメロディに感情的な豊かさや深みをもたらす重要な役割がある．しかしながら，メロディがあったとしても，伴奏を構成するには多くの知識や経験が必要である．音楽理論に関する知識が豊富な演奏者は，自力で伴奏をつけることが可能だが，知識が少ない演奏者においては，自力で伴奏をつけることが難しく，自身のオリジナリティを出した作曲やアレンジを断念してしまうことが多い．初心者が音楽知識なしにピアノ曲の伴奏付けを行うためには，メロディから伴奏を自動的に構成するシステムを構築する必要がある.

ユーザ入力したメロディに対して，自動的に伴奏を構成するシステムを実現することで，音楽理論に関する知識が少ないユーザの音楽製作やアレンジの一助になると考えられる.

本稿では，楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式について示す.

本方式は，類似するメロディには類似する伴奏をつけることが可能であるという仮説のもと，既存曲から伴奏を引用する形で構成する手法を提案する．まず事前に，既存曲から一定区間分のメロディと伴奏の組を取り出した知識ベースを用意する．次に，単一音で構成されたメロディを入力として，知識ベースのメロディと入力されたメロディからいくつかの特徴量を抽出し，ベクトルとして類似度計量を行うことで最も類似度の高い知識ベースのメロディを検索する．最後に，検索された知識ベースのメロディに付随する伴奏を，入力されたメロディに合わせて変形することで，伴奏を自動的に構成する.

本方式を実現することにより，音楽理論に関する知識が少ないユーザでも，ピアノ曲の製作やピアノアレンジが有意義になると考えられる.

本稿は以下のように構成される．2 節では，関連研究について示す．3 節では，本方式である音響特徴類似度による自動伴奏構成方式について示す.4 節では，実際に構成したシステムから伴奏を構成し，5 節で本稿をまとめる.

2. 関連研究

本節では，本方式に関連した研究について述べる．

Y. Ren [1]らは，ポップ音楽の伴奏生成のための単一のシーケンスで同時に複数のトラックを生成し，異なるトラックの音符間の依存関係を明示的にモデル化することが可能な PopMAG を用いた伴奏生成を提案している．

N. Jiang [2]らは，人間と機械のデュエット即興演奏のためのオンライン伴奏生成における深層強化学習アルゴリズムを提示しており，人間の演奏パートに対応し,メロディックでハーモニックかつ多様な 演奏パートの生成方式を提案している．

Z. Wang [3]らは，Transformer モデルを用いたリアルタイムコード生成と，CRF モデルを用いた，コードに基づくマルチトラック伴奏を生成する 2 段階予測フェーズにより，論理遅延と露出バイアスのないリアルタイム伴奏生成システムを提案している．

Q. Wang [4]らは，伴奏に感情の動きを導出するために，感情フローモデルである Valence/Arousal Curve の使用した伴奏生成を提案している．

Y. Mo [5]らは，メロディパターンデータベースを使用し，元のピアノ伴奏パターンをシステマティックに収集し，二つの隠れマルコフモデル(HMM)システムを使用した自動ピアノ伴奏システムを提案している．

これらの研究では，いずれも機械学習を用いて楽曲の特徴を学習し，自動で伴奏を生成する手法を提案している．これにより，多様な音楽スタイルに対応した伴奏の生成が可能となっている．

しかし，伴奏の自動生成における機械学習モデルの適用には，学習のための大規模な訓練データを用意する必要であるという課題がある．これは高度な音楽的知識やスタイルを反映するシステムにおいては特に顕著である．

本研究では，入力のメロディに対して，類似するメロディには類似する伴奏をつけることが可能であるという仮説のもと，伴奏パターンが固定されている既存曲から類似するメロディを検索し，伴奏を構成する手法を提案する．

本方式により，機械学習モデルを適用した際に必要となるパターン学習のためのデータを，少ないデータ量として伴奏を構成するシステムを構築することが可能となる．

3. 音響特徴類似度による自動伴奏構成方式

本節では，ピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式について述べる．本方式では，複数の既存曲を,メロディと伴奏が組みとなった MIDI データとして格納した,メロディ-伴奏知識ベースを扱う． また，メロディベクトル-伴奏データベースは，メロディ-伴奏知識ベースのメロディ部分をメロディベクトル化機能によってベクトル化し，伴奏とセットにして格納したデータセットである．メロディベクトル化機能については 3.2 節で述べる．ここで，原曲の調は全てハ長調に統一して格納する．

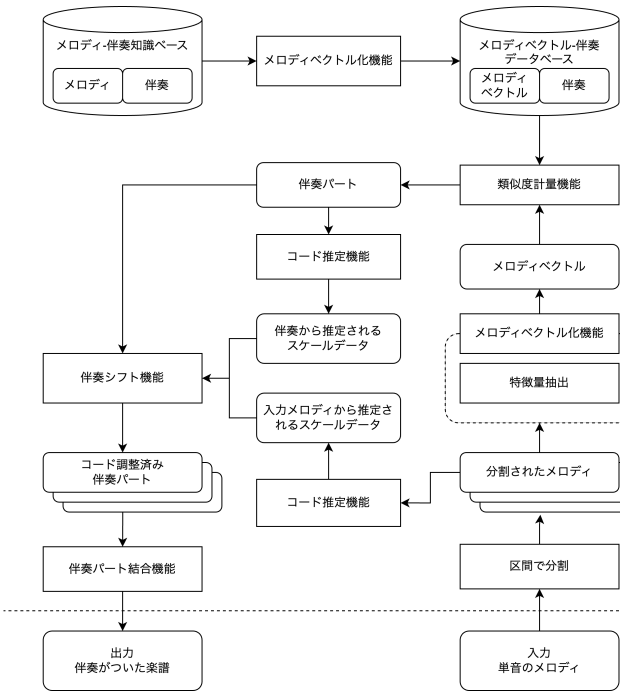


図 1　提案方式の全体像

3.1. 全体像

図 1 に本方式の全体像を示す．本方式は，メロディベクトル化機能，類似度計量機能，コード推定機能，伴奏シフト機能，伴奏パート結合機能からなる．

ユーザは,伴奏を構成したい単音のメロディの MIDI データを入力する．入力されたメロディを一定区間で分割し，分割区間ごとに特徴量の抽出とコードの推定を行う．さらに，抽出された特徴量をベクトル化し，同様にベクトル化されているデータベースとの類似度計量を行う．具体的には，最初の音高，音高の平均，音高の分散，音高の変化量の平均，音符密度の 5 つの特徴量を 5 次元のベクトルとして，ユークリッド距離が最も小さい値をとったデータベースのメロディに付随する伴奏パートを選択する．ここで選択された伴奏パートに対してコード推定を行い，入力されたメロディの推定コードに合わせて伴奏パートの音をずらすことで，コードを調整する．調整された伴奏パートを結合し，入力メロディに適合することで，伴奏が構成される．

3.2. メロディベクトル化機能

本節では，メロディを一定区間で分割し，分割されたメロディから特徴量を抽出してベクトルとして表す機能について述べる．メロディの持ついくつかの特徴量からベクトルを構成することで，メロディの構造を包括的に捉えることができると考える．抽出する特徴量について述べる．

１つ目の特徴量は，最初の音高である．一般的に，メロディにコード付けを行う場合，最初の音高から考慮することが多い．最初の音高は，メロディの調や構成音を捉える手がかりとなる特徴量であると考える．

２つ目は音高の平均，３つ目は音高の分散である．これらは，メロディ全体で構成されている音階の平均値と分散値である．音高の平均は，メロディのベースとなる音の高さを示し，音高の分散は，メロディの音域を捉える手がかりとなる特徴量であると考える．

４つ目は音高の変化量の平均である．これは，メロディ内の音符と音符の間における，音階の差の平均値である．音高の変化量の平均は，メロディの進行状況や，メロディの形を捉える特徴量であると考える．

５つ目は音符密度である．これは，メロディに対して１秒間に含まれる音符の数を示した指標であり，音符密度$Density$は，メロディ内の音符数N，テンポT，拍数Bで下記のように表される．

$$Density = \frac{N}{\frac{60}{T} \times B}$$

ここで，$\frac{60}{T}$は，１拍あたりの秒数を示し，拍数Bとの積によってメロディの長さを秒数として表す．さらにメロディ内の音符数Nによって，密度を算出する．音符密度は，メロディリズムの複雑性を捉えることができると考える．

これら５つの特徴量を，5 次元のベクトルとして表す．本方式では，このような配列をメロディベクトルとして，入力されたメロディと，事前に用意した伴奏知識ベースに格納したメロディで，メロディベクトル化機能を適用する．本機能によって，メロディの特徴を数値化することで，メロディ同士の類似度計算が可能となる．

3.3. 類似度計量機能

本節では，入力メロディのメロディベクトルと，メロディベクトル-伴奏データベースの既存曲のメロディベクトルで類似度を計算し，最も類似度の高いデータベースのメロディベクトルに付随する伴奏パートを選択する機能について述べる．

本方式では，データベースのメロディベクトルは，入力メロディの分割区間と同じ区間でベクトル化されたものを格納する．入力メロディの分割区間ごとに，データベースと類似度計算を行い，伴奏パートを選択する．類似度計算にはユークリッド距離を用いる．

ユークリッド距離は，数学的に 2 点間の直線距離を表すのに用いられる概念であり，データ点間の直線距離から類似性を測定することが可能である．これを踏まえ，入力のメロディベクトルを$P = (p_1, \dots p_5)$，メロディベクトル-伴奏データベースのメロディベクトルを$Q = (q_1, \dots q_5)$とすると，メロディ間の距離$Distance$は下記のように表される．

$$Distance = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_5 - q_5)^2}$$

これにより，メロディ同士の類似性を距離として算出することが可能となり，入力のメロディとの$Distance$の値が最も小さい Qのベクトルをもつ既存曲のメロディが選択される．

本機能によって，入力のメロディに対して，類似する既存曲のメロディに付随した伴奏パートを選択することが可能となる．

3.4. コード推定機能

本節では，3.3 節によって選択された伴奏パートの構成音を，入力メロディのコードに合わせて転移する過程において，メロディのコード推定と，コードの構成音からスケールを取得する機能について述べる．

コード推定には AutoHarmonizer[6] を用いる．AutoHarmonizer とは，メロディに合わせてコード進行を出力する Encoder-Decoder モデルである．AutoHarmonizer は，推定対象のメロディに対して，推定されたコードごとに，music21[7] の harmony.ChordSymbol オブジェクトが格納された配列を出力する．music21 とは，楽譜データの解析及び音楽理論の分析をするための python ライブラリである．harmony.ChordSymbol オブジェクトには，コードの構成音と，offset（1 拍の長さを 4.0 とした，メロディの開始点からの位置を表す数値)を保持している．

また，コードからスケールへの変換には music21 の scale クラスオブジェクトを用いる．具体的には，AutoHarmonizer が出力した harmony.ChordSymbol において，コードの一番低い音（ベース音）を取得する．次に，その音を基本音（ルート）とする scale クラスオブジェクトに変換する．

本機能では，まず，推定対象のメロディに対して AutoHarmonizer によるコード推定を行う．次に，推定された harmony.ChordSymbol オブジェクトから，ベース音を取得し，その音がルートである scale オブジェ

クトに変換する．メジャーコードの場合は，scale.MajorScale オブジェクトを，マイナーコードの場合は scale.HarmonicMinorScale オブジェクトを適用する．これと，harmony.ChordSymbol オブジェクトのコードネームと offset から，コードの開始点 start と，終了点 end を取得し，scale, start, end のキーを持つ辞書型の配列を出力する．本方式ではこれをスケールデータとする．

例として，ある 2 小節分のメロディを入力として，AutoHarmonizer でコード推定した場合を図 cc4nh2 に，類似度計量機能によって選択された伴奏パートを，AutoHarmonizer でコード推定した場合を図 3 に示す．

入力のメロディでは，赤で示した区間が「C」のメジャーコード，青で示した区間が「A」のマイナーコードが推定されたものとすると，入力のメロディのスケールデータは，[{‘scale’: music21.scale.MajorScale (‘C’), ‘start’: 0.0, ‘end: 4.0}, {‘scale’: music21.scale.harmonicMinorScale(‘A’), ‘start’: 4.0, ‘end: 8.0}]となる．

同様に，伴奏パートでは，赤で示した区間が「F」のメジャーコード，青で示した区間が「C」のメジャーコードが推定されたものとすると，入力のメロディのスケールデータは，[{‘scale’: scale.MajorScale(‘F’), ‘start’: 0.0, ‘end: 4.0}, {‘scale’: scale.MajorScale(‘C’), ‘start’: 4.0, ‘end: 5.5}, {‘scale’: scale.MajorScale (‘F’), ‘start’: 5.5, ‘end: 8.0}]となる．

本方式では，コード推定機能を，入力のメロディと，類似度計量機能によって選択された伴奏の 2 箇所で適用する．本機能により，双方で取得されたスケールデータをもとに，伴奏パートの現在のスケールと，転移先である入力メロディのスケールを取得することができ，転移先の音を明確化することが可能となる．

3.5. 伴奏シフト機能

本節では，3.3 節よって選択された伴奏パートの構成音を，3.4 節で取得したスケールデータに沿って転移する機能について述べる．

本機能では，まず，選択された伴奏パートのスケールデータから，start と end 間の区間における伴奏パートの構成音について，その音符が，対象の scale の何番目の音階（スケールディグリー）になるかを取得する．スケールディグリーは，music21 の scale オブジェクトに対して getScaleDegreeFromPitch クラスメソッドを使うことで取得できる．さらに，入力メロディのスケールデータと取得したスケールディグリーもとに，start と end 間の区間における伴奏パートの構成音を，

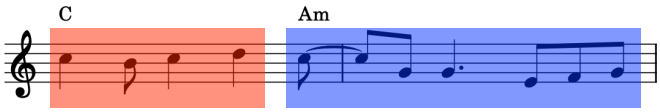



図 2　2 小節分の入力データへのコード推定の例

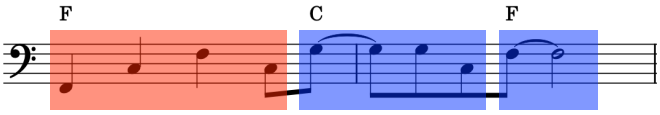


図 3　選択された伴奏パートのコード推定の例

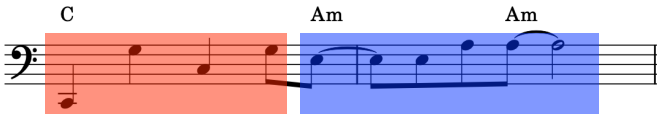


図 4　図 3 を転移した伴奏パート

そのスケールのスケールディグリーにあたる音に転移する．なお，転移先のオクターブは，転移前と同じオクターブを適用する．

図 2 は 2 小節分のメロディを入力してそのコード推定をした例である．入力したメロディに類似するメロディを検索し，そのメロディと組になっている伴奏に対してコード推定をしたものが図 3 である．音符に付いている数値はコードにおけるスケールディグリーである．この伴奏パートを，図 2 で推定したコードに合わせて構成音を転移したものを図 4 に示す．

本機能により，入力のメロディで推定されたコードに対して，選択された伴奏パートの構成音をコードに合わせることができ，入力のメロディにより適合性のある伴奏を構成することが可能となる．

3.6. 伴奏パート結合機能

本節では，分割したメロディごとに選択された伴奏を結合することで，伴奏を構成する機能について述べる．本機能は，3.5 節で調整された伴奏パートを，分割区間ごとに接続し，入力のメロディに割り当てることで入力のメロディと構成された伴奏が組となった楽譜データを生成する．

本機能により，既存曲から一定区間ごとに選択された伴奏パートを，ひとつの伴奏として結合し，選択入力のメロディと構成された伴奏を 1 つの楽譜として視覚的に表現することが可能となる．

4. 評価実験

本節では，本方式の評価実験について述べる．3 節で示した方式を実装し，本方式の有効性の検証を行うために 2 つの実験を行う．

表 1　実行環境 – コンピュータスペック

項目	詳細
OS	macOS Sonoma 14.2.1
モデル	MacBook Air (M1, 2020)
チップ	Apple M1
メモリ	16GB

表 2　実行環境 – 主要 Python ライブラリ

ライブラリ名	バージョン
music21	9.1.0
numpy	1.25.1
scipy	1.11.1

本方式の実用化には，入力のメロディと，構成される伴奏との適合性において，選択される既存曲のメロディとの類似度の高さが重要であると考える．これを踏まえ，実験 1 では，3.2 節で述べたメロディベクトル化機能と，3.3 節で述べた類似度計量機能の有効性を検証する．また，実験 2 では，任意の入力メロディによって出力された伴奏を用いて，被験者実験を行い，入力のメロディと出力された伴奏の適合性を評価する．

4.1 では実験環境について述べる．4.2 節では実験 1 の類似度計量機能の有効性の検証ついて述べる．4.3 節では実験 2 のシステム全体の有効性の検証ついて述べ，4.4 節では本研究による実験結果について考察を行う．

4.1 実験環境

システムの構築には Python 言語を用いる．Python バージョンは 3.11.6 を使用する．実行環境におけるコンピュータのスペックを表 1 に，主要 Python ライブラリを表 2 に示す．

また，本実験に用いるデータは，4/4 拍子かつメジャーキーである既存曲から，転調しない 8 小節分を格納したデータセットである．既存曲のジャンルはランダムである．メロディベクトル化における分割区間は 2 小節とし，30 曲分の既存曲を用いるため，メロディベクトル-伴奏データベースにおける伴奏パートは，120 種類である．

4.2 実験 1

実験 1 では，メロディベクトル化機能と類似度計量機能の有効性の検証を行う．本実験では，3 つの任意の単音のメロディ，8 小節分を入力として，2 小節ごとにメロディベクトル化機能を実行し，類似度計量機能を用いて，入力のメロディと最も類似度の高いメロディを知識ベースから選択する．これにより，入力メロディと選択されたメロディを比較して検証を行う．

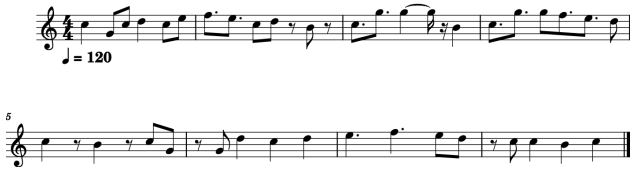



図 5　実験 1 の入力メロディ 1

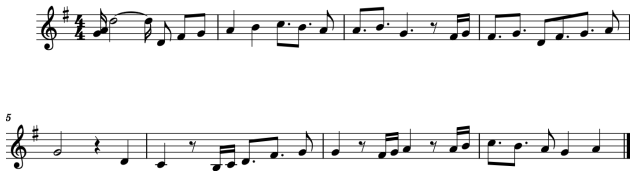



図 6　実験 1 の入力メロディ 2

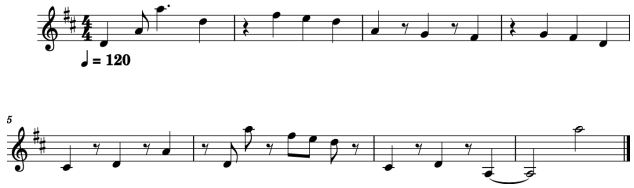



図 7　実験 1 の入力メロディ 3

入力となる 3 つの単音のメロディを図 5，6，7 に示す．本実験ではこれらのメロディを入力として比較検証する．

実験結果を表 3 に示す．ほとんどの区間において，音高や音域，メロディの進行方向に関しては，入力したメロディと類似したメロディが選択された．しかし，音符密度やリズムに関しては，音高，音域，メロディの進行方向などと比較すると，あまり類似しているとはいえない．

結果として，類似度計量機能は本方式において有効であり，メロディベクトル化機能における特徴量の抽出において，入力メロディの音高に関する特徴は比較的捉えられているが，リズムに関する特徴は不十分だといえる．

今後の課題として，メロディベクトル化機能における特徴量に，メロディのリズムの複雑性を捉えることができる特徴量を適応することが必要だと考える．これにより，複雑なメロディでもメロディのもつ特徴を体系的に捉えることができ，より類似する既存曲のメロディを選択することが可能となる．

4.3 実験 2

実験 2 では，任意の入力メロディに対し，本方式により構成された伴奏との適合性について被験者実験を行う．被験者実験では，実験 1 で使用した 3 つの任意の入力メロディに対して伴奏を構成し，被験者に対して「メロディに対するリズムの合致度」，「メロディに対する構成音の合致度」の 2 つの項目で，それぞれ 5 段階の評価をさせる．

表 3　実験 1 の結果

入力	区間	項目	楽譜画像
1	1~2	入力	
		結果	
	3~4	入力	
		結果	
	5~6	入力	
		結果	
	7~8	入力	
		結果	
2	1~2	入力	
		結果	
	3~4	入力	
		結果	
	5~6	入力	
		結果	
	7~8	入力	
		結果	
3	1~2	入力	
		結果	
	3~4	入力	
		結果	
	5~6	入力	
		結果	
	7~8	入力	
		結果	

表 4　アンケート結果

項目	リズムの合致度	構成音の合致度
入力メロディ 1	3.71	3.57
入力メロディ 2	3.71	3.43
入力メロディ 3	4.57	4.42

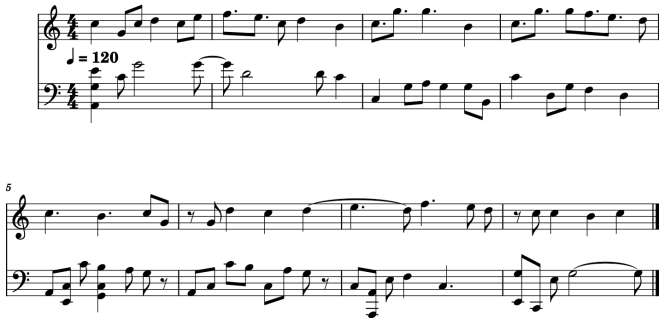



図 8　入力メロディ 1 の出力楽譜

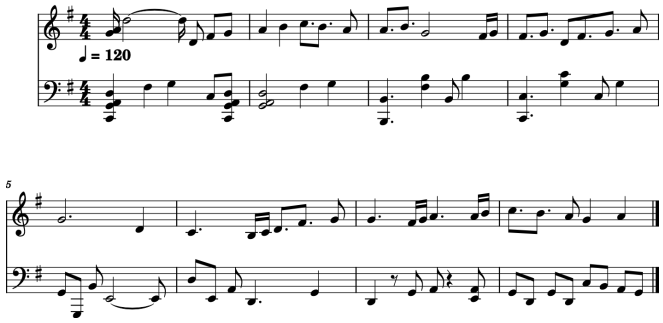



図 9　入力メロディ 2 の出力楽譜

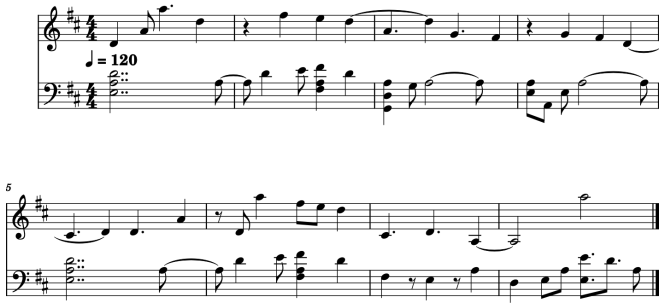



図 10　入力メロディ 3 の出力楽譜

実験結果における 5 段階評価の評価値について，入力メロディごとに平均値をとり，表 4 に示す．また，本方式を用いて構成した伴奏を，入力のメロディと伴奏が組となった楽譜を，それぞれ図 8, 9, 10 に示す．

全体として，どの入力メロディにおいても，伴奏のリズム・構成音の観点では 3.0 より高い評価を得た．特に 2 つの項目において，入力メロディ 3 の出力伴奏が最も高い評価となった．また，どのメロディにおいてもリズムより構成音の方が低い評価となった．

結果として，任意のメロディを入力として，本方式を用いて構成した伴奏は，メロディのリズムに関しては適合性が高いといえる．

4.3 考察

本節では，実験全体の結果を踏まえ，考察を行う．実験 1 では，入力のメロディと，データベースから選択された距離の近いメロディとの比較検証を行い，類似度計量機能とメロディベクトル化機能の有効性が確認された．実験 2 では，本方式を用いて構成した伴奏について被験者実験を行うことで，伴奏のリズムと構成音の観点から入力メロディとの適合性を評価した．

以上の実験から，類似したメロディには類似した伴奏をつけることが可能であり，本方式は有効であると考える．

今後の課題として，入力メロディに対してより適合性の高い伴奏を構成するために，入力のメロディの特徴量の抽出についての再考察が必要であると考える．

また，今回は検索対象データとして用意したメロディと伴奏の組は 120 件と少なかったため，特徴量の設定とその類似度の設計に問題があるのかの特定が困難であった．より多くのデータを用意し，システムの出力のバリエーションを増やしたうえで，再実験を行う必要があると考えられる．

5. おわりに

本稿では，楽曲メディアコンテンツを対象としたピアノ演奏に特化した音響特徴類似度による自動伴奏構成方式について示した．

本方式は，単音で構成されたメロディを入力とし，事前に用意した既存曲のメロディと最も類似度の高いメロディに付属する伴奏パートを選択し，構成音をコードに合わせて調整することで，入力のメロディに対して伴奏を自動的に構成する．

また，入力メロディと，類似度計量の結果による既存曲のメロディの選択において，双方を比較検証し，メロディベクトル化機能とメロディ類似度計量機能の有効性を確認した．さらに，被験者実験を用いて，入力のメロディと構成された伴奏との適合性を検証した．

今後の課題として，入力のメロディに対する特徴量抽出手法の改善が挙げられる．また，データベースの既存曲数を増やすことで，選択可能なメロディの幅が広がり，入力のメロディにより適合性の高い伴奏が構成されると考えられる．

本システムによって，音楽理論に関する知識が少ないユーザであっても，単音のメロディの入力によって伴奏を構成することができ，よりオリジナリティのある楽曲製作の一助につながると考える．

参 考 文 献

[1] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, T.-Y. Liu, "PopMAG: Pop Music Accompaniment Generation," in Proceedings of the 28th ACM International Conference on Multimedia, pp. 1198-1206, 2020, Association for Computing Machinery, New York, NY, USA, doi: 10.1145/3394171.3413721. [Online]. Available: https://doi.org/10.1145/3394171.3413721

[2] N. Jiang, S. Jin, Z. Duan, C. Zhang, "RL-Duet: Online Music Accompaniment Generation Using Deep Reinforcement Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 710-718, 2020. [Online]. Available: https://doi.org/10.1609/aaai.v34i01.5413

[3] Z. Wang, K. Zhang, Y. Wang, C. Zhang, Q. Liang, P. Yu, Y. Feng, W. Liu, Y. Wang, Y. Bao, Y. Yang, "SongDriver: Real-Time Music Accompaniment Generation without Logical Latency nor Exposure Bias," in Proceedings of the 30th ACM International Conference on Multimedia (MM '22), pp. 1057-1067, 2022. doi: 10.1145/3503161.3548368. [Online]. Available: https://doi.org/10.1145/3503161.3548368

[4] Q. Wang, S. Zhang, L. Zhou, "Emotion-Guided Music Accompaniment Generation Based on Variational Autoencoder," in 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, pp. 1-8, 2023, doi: 10.1109/IJCNN54540.2023.10191768.

[5] Y. Mo, M. A. Jan, "Designing an Automatic Piano Accompaniment System using Artificial Intelligence and Sound Pattern Database," Mobile Information Systems, vol. 2022, Article ID 7875818, 2022. [Online]. Available: https://doi.org/10.1155/2022/7875818

[6] S Wu, Y Yang, Z Wang, X Li, M Sun, "Generating Chords from Melody with Flexible Harmonic Rhythm and Controllable Harmonic Density," arXiv preprint arXiv:2112.11122 (2021). Available: https://doi.org/10.48550/arXiv.2112.11122

[7] music21 Documentation — music21 Documentation https://web.mit.edu/music21/doc/

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

📅2024年2月28日(水) 13:30 ~ 15:40 | 🏛T4-B オンライン（Zoom Events)

XR

座長:伊藤 貴之(お茶の水女子大学)
コメンテータ:大塚 真吾(神奈川工科大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T4-B-2-01] 視覚障がい者のためのソーシャルVR空間の開発
*横山 皓祐[1]、Siriaraya Panote[2]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学)

14:00 ~ 14:25
[T4-B-2-02] 立体音響ARを用いた リラクゼーション支援手法の有効性評価
*大塚 智貴[1]、Siriaraya Panote[3]、中島 伸介[2] (1. 京都産業大学大学院、2. 京都産業大学、3. 京都工芸繊維大学)

14:25 ~ 14:50
[T4-B-2-03] MR空間におけるホワイトボードに対するインタラクションを伴う商品情報の探索
*津田 裕哉[1]、山本 岳洋[1]、大島 裕明[1] (1. 兵庫県立大学)

14:50 ~ 15:15
[T4-B-2-04] 物理的距離によるLoD制御をともなうVR環境における不動産情報探索
*中山 裕紀[1]、津田 裕哉[1]、荘司 慶行[2]、大島 裕明[1] (1. 兵庫県立大学、2. 静岡大学)

15:15 ~ 15:40
[T4-B-2-05] 空港手荷物受取所のUX向上を目指した拡張現実型情報提示サービス
*池田 呼人[1]、松原 克弥[1] (1. 公立はこだて未来大学)

視覚障がい者のためのソーシャルVR空間の開発

横山　皓祐† Panote Siriaraya†† 中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学 情報工学・人間科学系　〒 606–8585 京都市左京区松ヶ崎橋上町
E-mail: †{g2054470,nakajima}@cc.kyoto-su.ac.jp, ††spanote@kit.ac.jp

あらまし　ソーシャル VR は、ワールド内で様々なユーザと社会的な交流を行うことができる．しかし，視覚障がい者は，仮想現実空間内の情報を視覚から得ることができず, 自ら移動することや他者との交流は困難である．そこで本研究では，視覚障がい者のためのソーシャル VR 空間を開発した．移動をサポートするナビゲーションシステムでは，物体の方向と距離感を捉えさせるために立体音響を実装し，加えて探索レーザーが物体に当たるとコントローラが振動し音が鳴るシステムを開発した．交流促進システムでは，他のユーザの視線情報から，自分の周辺にいる他のユーザの情報を音声で通知するシステムを開発した．評価実験では，探索レーザーと交流促進システムの有効性を確認するための実験をそれぞれ行った．結果から，探索レーザはユーザビリティにおいて高い評価を得，システムの有効性が確認できた．

キーワード　VR, 立体音響, ユーザー支援, ソーシャルインタラクション

1. はじめに

視覚障がい者は，平成 28 年度の厚生労働省の調査によると，日本に 31.2 万人いるとされている [1]．VR 機器は，実空間における物理的障害が少なく，場所を選ばずに現実に近い，またはそれ以上の様々な体験をすることができるため，障がい者福祉分野やリハビリテーションに取り入れられるなど，視覚障がい者に対して有効であり，安心して利用しやすい機器である [2] [3].

VR 機器を用いる VR アプリケーションの中にソーシャル VR というジャンルがある．ソーシャル VR はインターネットを介してワールド内で様々なユーザと会話したり、社会的な交流を行うことができ，外出をしなくても，リアルとあまり変わらないような交流をすることができる. また，VR はコロナ禍を境に普及しつつあり，ソーシャル VR は，日常的なコミュニケーションツールとして普及していく可能性があり，第 2 の現実として，新たな人間関係を構築することができるツールとしての役割が期待できる．しかし，視覚障がい者は，ソーシャル VR 空間において，周りの状況が分からず，ワールド内を自由に行動したり，行きたい方向に向かうことができないという問題点や，他のユーザと交流したくても，自分の周りにどんな人がいて，何人くらいいるのか分からず，行動を起こせないという問題点がある．また，現状視覚障がい者が利用しやすいようにサポートされたソーシャル VR が存在しないことも，視覚障がい者がソーシャル VR を利用する機会を奪っている原因である．

そこで本稿では，視覚障がい者がソーシャル VR を利用する上での「空間把握」と「他ユーザとの交流の促進」の 2 つの軸をもとに視覚障がい者の空間把握とソーシャルインタラクションを可能にするソーシャル VR 空間の開発を目的としている．視覚障がい者の仮想現実空間内での空間把握を補助するための手法として，物体の方向と距離感を捉えさせるためにユーザーの足音や環境音を立体音響で実装する．加えて，右手コントローラに追従するレーザーが物体に当たり続けることでコントローラが振動し音が鳴るシステム，探索レーザーを開発した．鳴る音は，レーザーが当たっている物体の種類ごとに変化し，音とコントローラの振動の周期は，物体に近づくことで徐々に短くなり，近づいていると実感できるようになっている．交流を促進するシステムとして，我々は視線情報に着目し，ヘッドマウントディスプレイ（以下，HMD）の頭の向き情報を視線の向きをとし，他のユーザの視線情報から，自分の周辺にいるユーザの情報を音声で通知したり，会話の意思の有無を推定し交流を促す仮想現実空間を提案する．

2. 関連研究

仮想現実空間内での視覚障がい者の行動の困難性の解決として，様々な研究が行われている．

関 [2] は，仮想現実空間で視覚障がい者に車が通過する音のような環境音を 3 次元音響で提供し，空間把握能力を養う訓練を行い，訓練として有効であるかを調査した．結果，実空間訓練よりも有効であることが示された．大内ら [3] の研究では，3 次元音響バーチャルリアリティを実現するための聴覚ディスプレイを開発し，身辺空間認知や認知地図形成の実験を行なった．結果，訓練効果が認められ，有効であることが示された．玉渕ら [4] の研究では，空間音響のみを用いた屋内ナビゲーションを提案している．現在地から目的地までの平面を取得し，音源を設置することで視覚障がい者を誘導する手法を提案している．

これらの関連研究は，仮想現実空間での実験を通して現実空間での訓練に応用することが目的であり，仮想現実空間での行動を目的としたものではない．我々は，仮想現実空間において視覚障がい者が抱える視覚的問題を解決するための研究である．

仮想現実空間内での視覚障がい者のコミュニケーションの困難性の解決として，下記の研究が行われている．

図 1　ソーシャル VR 空間における視覚障がい者の活動

Ji ら [5] は，VR 空間を 3 つのバブルに社会的な空間を分割し、異なるバブル内のアバターを区別するための空間音響フィードバックを生成する手法を開発した．ナビゲーションと会話のタスクによる実験の結果，アバターの認識が向上したが混雑した環境ではより注意を散漫させることが示された．

Krishna ら [6] は，視覚障がい者のソーシャルインタラクションを向上させるために視覚障がい者に与えるべき必要な情報，ニーズについて調査し，それらを元に顔認識システムによるソーシャルインタラクションアシスタントを開発した．

3.　視覚障がい者のためのソーシャル VR 空間

3.1　概　　要

本研究では，視覚障がい者の空間把握を補助する手法と視覚障がい者の交流を補助する手法を提案する．視覚障がい者の空間把握を補助する手法として，物体の方向と距離感を捉えさせるために仮想現実空間内のユーザの足音や環境音を立体音響で実装する．加えて，右手コントローラに追従するレーザーが物体に当たり続けることでコントローラが振動し音が鳴るシステム，探索レーザーを開発した．鳴る音は，レーザーが当たっている物体の種類ごとに変化し，音とコントローラの振動の周期は，物体に近づくことで徐々に短くなり，近づいていると実感できるようになっている．立体音響と探索レーザーを組み合わせることで自由に行きたい場所に行くことができるようになることを目指す．交流を補助する手法として，ユーザの視線情報を基準として，HMD の機能で頭の向き情報を取得し，その情報から視線を推定する．他のユーザの視線が自分の方向を向いているかつ他のユーザと自分が特定の距離以内にいる場合に，そのユーザの情報を音声案内で通知する．ユーザの情報がわかるかつ自分に興味を持っているという情報を得ることで交流を開始しやすくすることを目指す．

下記にそれぞれの手法について詳細を述べる．

3.2　空間把握を補助する VR 空間の設計

3.2.1　立 体 音 響

立体音響は，音源の方向感や奥行き感など音場の立体的な雰囲気や臨場感を再現するために用いられる再生方式のことである．関連研究 [2] [3] [4] においても，立体音響は視覚障がい者への空間認識を助ける手法として有効であることが示されている．本研究では，VR 空間を作成する際に広く利用されている Unity を利用する．仮想現実空間内にいるユーザの足音や物体から発生する音声（木であれば葉が揺れる音など）を立体音響で作成することで，被験者とその他のオブジェクト間の感覚的な距離を掴んでもらうことを目的としている．立体音響を作成する上で Unity 内の機能のひとつである Oculus Audio Spatializer [7] について述べていく．

Oculus Audio Spatializer を用いた立体音響では，音源位置の認識を以下の機能を用いることで実現している．今回は屋外の空間を想定しているため，室内空間において利用される機能については省略している．

- 空間化．頭部伝達関数 (HRTF）を利用し，モノラル音源を変換することで目的の方向から音が発せられているかのように聞こえるプロセスである．
- ヘッドトラッキング. 頭の位置と方向をトラッキングすることで，正確に音声の空間化を行う．また，頭の小さな動きによる音の方向の変化の大きさは，音声の距離を伝えることにも利用できる．
- 体積音源．音源に半径を設定することで，音源に体積を持たせる．これによって音声が広がるため，聞き手が音源に近づき完全に包み込むと音源空間全体に音声が広がる．
- 近接レンダリング．聞き手の頭に近接した音源は, その空間化の一部の側面が音源の距離に依存しないという特性を持つ. 近接レンダリングでは, 音波回折の効果を自動的に見積り,1 メートルより近い音源をより現実的に表現する．
- 減衰と反射．距離のシミュレーションでは正確な反射が最も重要な機能であり, 正確な距離の手がかりを得るためには, 音の直接パスと反射の間に自然な均衡があることが非常に重要である．このモデルは, 反射と直接信号の適切な均衡を保つことができる. 今

回は屋外の空間を想定しているため，反射の機能は使われない．また，音の減衰についての設定も行うことができる．音の減衰は距離の手がかりとなる要素である．高周波は低周波と比べて早く減衰するため，距離の手がかりとなる．高周波減衰は空気吸収（大気条件）によって得られる効果であり，ローパスフィルタを通すことでモデリングすることができる．

Oculus Audio Spatializer ではこれら機能により、立体的な音響表現を可能としている。

3.2.2 探索レーザー

立体音響だけでは，具体的な方向感と距離感を正確に掴むことができないため，立体音響とは別に探索レーザーというナビゲーションシステムを開発した．探索レーザーの概要を図 2 に示す．右コントローラのグリップボタンを推し続けている間，空間内にある特定の物体を検知できるレーザーが放出される．レーザーと物体の衝突判定は ray を用いて行う．レーザーは右コントローラの付け根から放出され，コントローラを動かすと，その動きに追従するようになっている．レーザーが物体に衝突すると，一定の周期でコントローラが振動する．振動することで，レーザーが物体に当たっていることが感覚的にわかりやすくなる．また，振動する周期と同じ周期で短い音が鳴るようになっている．鳴る音は，レーザーが衝突した物体に応じて異なり，鳴る音によって，レーザーが衝突した方向に何があるのかを推定できるようになっている．例えば，レーザーが岩に衝突した場合は岩を叩くような音，木に衝突した場合は葉が揺れる音が再生される．また，周期はユーザが物体に近づくことで一定距離ごとに短くなる．周期が短くなっていくことで，ユーザーが物体に徐々に近づいているということがわかる．

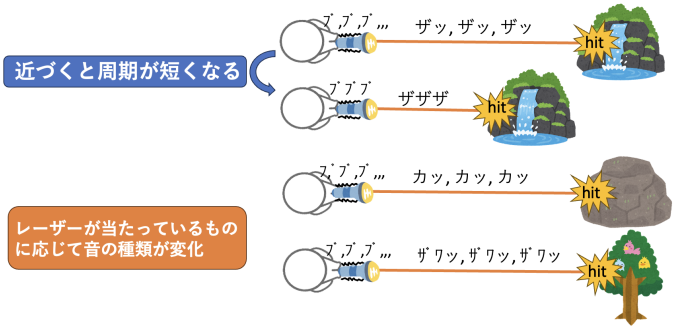


図 2　探索レーザーの概要

3.2.3 音 声 案 内

立体音響と探索レーザーは，物体に近づいているということが感覚的にわかるためのシステムであり，具体的に物体の近くに到着したのか，これ以上進む必要があるのかどうかという部分が曖昧になる．よって，特定の距離内に被験者が接近した場合に「これは (物体の名前) です」のような音声を流すように設定する．また，音声案内ではユーザの認知負荷が大きいため，あまり多くの情報を伝えるのではなく，そこに何があるのかのみを伝えるようにする．

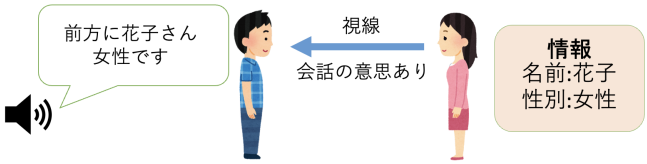


図 3　コミュニケーション補助の概要

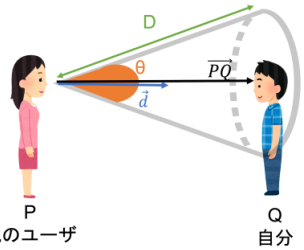


図 4　視線の判定方法

3.3 VR 空間内での交流サポートシステムの設計

3.3.1 頭の向き情報の取得と視線判定

Krishna ら [6] は, 視覚障がい者にとって重要な社会的交流の時のニーズついて紹介している．そのニーズの中に,「人のグループの中に立っているとき、周りにいる人の名前を知りたい」「人のグループの中に立っているとき、そこに何人いるのか、そしてそれぞれの人がどこにいるのかを知りたい」というものがある．仮想現実空間におけるコミュニケーションにおいて，視覚障がい者が他のユーザの名前や性別などの情報を得る手段がない．また，他のユーザが自分とコミュニケーションを取る意思があるのか，はたまた偶然自分の近くにいるだけなのかが分からない．そこで，我々は,「他のユーザの視線」に注目し，一定の距離以内にいる他のユーザの視線が自分の方向を向いているかどうかで「コミュニケーションの意思」があるのかを推定する．本研究では，HMD として，Meta 社の Oculus Quest2 [8] を用いて頭の向き情報を取得する．また，頭の向きから図 4 のような円錐部分の領域に入っていたら,「視線がこちらを向いている」とする．この円錐は，他のユーザのアバターの位置を P, 長さを 1 に正規化された向きを $\vec{d}$, 被験者のアバターの位置を Q, 視野角を θ , 最大距離を D とする．Q が視界に入る条件は以下の式で表せる．

$$\frac{\overrightarrow{PQ}}{|\overrightarrow{PQ}|} \cdot \vec{d} > cos\frac{\theta}{2} \tag{1}$$

$$0 < |\overrightarrow{PQ}| < D \qquad \text{但し } |\vec{d}| = 1 \tag{2}$$

自分の周りに何人程度いるのか，それぞれの人がどこにいるのかという情報に関しては，足音やそれぞれのユーザ自身の声，ユーザ情報の通知時に何人通知されたか，現在の空間にいる人数の情報で推定する．

3.3.2 音 声 案 内

他のユーザが一定の距離以内かつ視線が自分の方向を向いていると判定できる場合，他のユーザの情報を音声案内で通知す

る．伝える情報は,「名前」「性別」とし，それらの情報を被験者に通知する．また，ユーザが空間を出入りするときに，入退室音と現在の空間の中にいるユーザの数を音声で通知する．ユーザの情報がわかることで，コミュニケーションを取ることへの心理的ハードルが低くなり，会話がしやすく，始めやすくなるのではないかと考える．

3.4 その他の機能

視覚障がい者は，システムを利用する際に，自分のアクションに対して何らかのレスポンスがないと，操作できているのかわからなかったり，不安になってしまう．この問題を解決するために，3.2 章,3.3 章で紹介した機能とは別に，視覚障がい者が仮想現実空間を利用する上で便利であると考えられる視覚障がい者にレスポンス返すための機能について，別途実装したので紹介する．

3.4.1 足　　音

ユーザが移動する時に足音が鳴るようにする．自分が移動しているというフィードバックの役割と，他のユーザの位置情報を知る手がかりになる．

3.4.2 杖

視覚障がい者は，実空間を移動する際に白状を用いる．白状がものにぶつかることで障害物の存在に気づいたり，白状で地面を叩くこと発生する音の反響を用いて，物体との距離を測り，障害物ぶつかることなく進路を進んだりすることができる（エコロケーション）．このように，視覚障がい者が普段から使用するのに慣れているであろう杖を実装した．実空間で杖を振り回す行為は危険だが，仮想現実空間だと杖に他のユーザが当たっても怪我を負わせるということはないため，周辺の物体の存在を確認するために使用する．杖が周辺の物体に当たることで，ものの存在を認知することができる．また，当たったものに応じて，杖が当たった音が変化する（岩に杖が当たったら岩を叩いたような音が鳴るなど）．

3.4.3 衝　突　音

今回作成した仮想現実空間は，移動できる空間が制限されている．移動できる空間が無制限にあると，どこまでも歩き続けられてしまうため，マップのどのあたりにいるのかが全くわからなくなってしまう．そのため,「自分がマップの端にいる」という情報は重要であり，その情報をユーザに与えるために衝突音を利用する．今回はマップを正方形で作成し，周辺に柵のようなものを設置する．ユーザが柵に衝突すると衝突音が発生し，自分がマップの端にいるということが知覚でき，体の方向を転換することができるようになる．

4. 評 価 実 験

4.1 概　　要

3.2 節で説明した空間把握を補助する手法において，探索レーザーが VR 空間における空間把握に有効であるかを検証するために，我々が別に開発した扇形の判定を持つ探索法と比較する被験者実験を通して評価する．被験者は，視覚障がい者でない晴眼者 10 名 (20 代男性 9 名，20 代女性 1 名) に対して行った．

4.2 扇形探索法

扇形探索法は，探索レーザーと同様に右のコントローラのグリップボタンを一度押すことで右コントローラの位置を中心に 30 度の扇形状に ray が発射される．発射された ray に物体が衝突したときに，被験者と ray が衝突した物体との距離を計算し，距離に応じて音が遅延して返ってくる．返ってくる音は立体音響で作成し，音の種類は探索レーザーの時と同様物体によって変化する．

4.3 実 験 手 順

作成した VR 空間を図 5 と図 6 に示す．空間内には 4 つの物体（木，岩，滝，鳥）が離れた距離にばらばらに存在している．被験者には，この 4 つの物体を滝→木→鳥→岩の順番に到達を目指してもらい，4 か所全て到達が可能かどうか実験を行う．被験者には，実験開始時に HMD とヘッドフォンを装着してもらい，目を瞑ったままコントローラで移動してもらう．4 か所すべてに到達する，もしくは途中で次の物体に到達することができないと被験者が判断した場合に実験を終了とする．実験の前に，探索レーザーと扇形探索法の機能および 3.4 章で説明したその他の機能について事前に説明および簡単なデモを行い，その後上記で説明した手順で探索レーザーを用いる場合と扇形探索法を用いる場合の計 2 回実施してもらった．2 回とも同じ空間を使用すると 1 回目に行った空間にある物体の配置を覚えている可能性があるため，探索レーザーと扇形探索法で別々の空間を使用した．2 回とも終了した後に，アンケートを実施した．

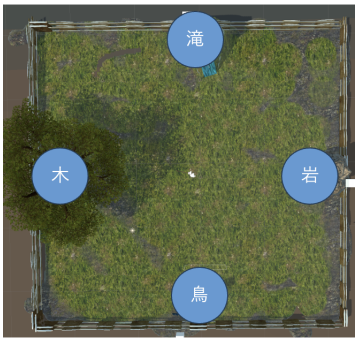


図 5　実験空間 1

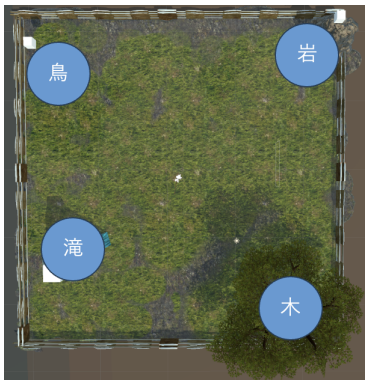


図 6　実験空間 2

4.4 評価方法

実験終了後にアンケートを答えてもらう形式で実験を行う．アンケート内容として，ユーザビリティを評価する評価尺度である System Usability Scale(SUS) のアンケートと我々が独自で作成したナビゲーションシステムに関する質問が含まれたアンケートに回答してもらった．

4.4.1 System Usability Scale (SUS)

System Usability Scale(SUS) [9] は，John Brooke が開発したシステムのユーザビリティの数値的な評価が可能な評価尺度である．100 点を基準とした最終スコアを算出する．Sauro [10] は，SUS の平均得点とグレードの目安の点数を統計により導き出した．SUS の平均得点は 68 点であり，グレードの目安の表を表 1 に示す．本実験では，独自に日本語訳したものを使用した．

表 1　SUS スコア表

SUS Score Range	Grade	Percentile Range
84.1-100	A+	96-100
80.8-84.0	A	90-95
78.9-80.7	A-	85-89
77.2-78.8	B+	80-84
74.1-77.1	B	70-79
72.6-74.0	B-	65-69
71.1-72.5	C+	60-64
65.0-71.0	C	41-59
62.7-64.9	C-	35-40
51.7-62.6	D	15-34
0.0-51.6	F	0-14

4.4.2 ナビゲーションに関するアンケート

ナビゲーションシステムとして，有効に機能しているかどうかを確かめるために，ナビゲーションシステムに関する質問が含まれたアンケートを作成した．作成したアンケートの項目について図 7 に示す．各項目 7 段階のリッカート尺度で集計し，平均値を算出する．算出した平均値をもとに，平均値の比較とウィルコクソンの符号付き順位検定を用いて評価する．

質問1	周辺の物体の位置はどの程度わかりやすかったですか？
質問2	ナビゲーション機能は目的地の誘導に役立った
質問3	ナビゲーション機能を使うことで，大まかな距離感を掴むことができた
質問4	ナビゲーション機能の使用結果，自分がフィールドのどの辺りにいるのか掴むことができた
質問5	ナビゲーション機能の音の種類と進む方向は結びつけることができた
質問6	ナビゲーション機能の音の種類は聞き取りやすいものだった

図 7　ナビゲーションに関するアンケート

4.5 結　　果

4.5.1 ユーザビリティ

探索レーザーと扇形探索法の SUS のスコアはそれぞれ 79.5 点と 71 点であった．表 1 の中で探索レーザーの Grade が A-であり，扇形探索法の Grade が C であった．

4.5.2 ナビゲーションに関するアンケート

ナビゲーションの機能性アンケートの結果を図 8 に示す．7 段階のリッカート尺度を用いて 4 人の平均値を各項目ごとに算出したものである．6 つの項目のうち，質問 6 以外の 5 項目は探索レーザーの方が優れている．ウィルコクソンの符号付き順位検定を行った結果を図 9 に示す．両側検定の有意差 5%で検定を行った結果，質問 1,2 は有意差が見られたが，質問 3,4,5,6 は有意差が見られなかった．

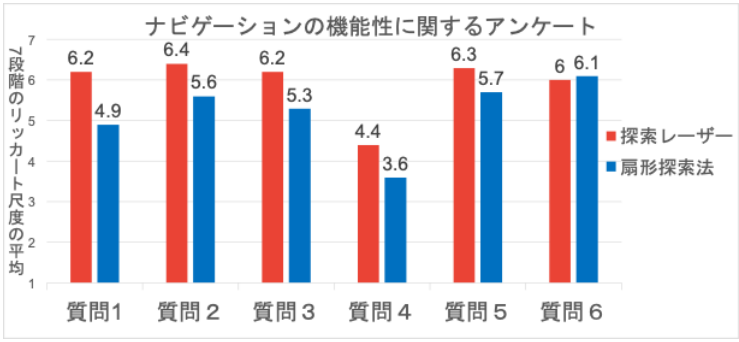


図 8　探索レーザーと扇形探索法のナビゲーションの機能性アンケート結果

	ウィルコクソンの符号付き順位検定
質問1	0.047
質問2	0.023
質問3	0.089
質問4	0.135
質問5	0.194
質問6	0.748

図 9　ウィルコクソンの符号付き順位検定

4.6 考　　察

SUS の結果より，扇形探索法を用いたシステムよりも探索レーザーを用いたシステムの方がユーザビリティが高いことがわかった．また，SUS のスコアの平均点が 68 点であり，探索レーザーの Grade が A-であることから，SUS のスコアの観点から見ても優秀である．

ナビゲーションに関するアンケートの結果から，6 つの項目のうち，質問 6 以外の 5 項目は探索レーザーの方が優れていた．また．ウィルコクソンの符号付き順位検定を行った結果，質問 1,2 は有意差があり，質問 3,4,5,6 は有意差がないことがわかった．探索レーザーに関して，質問 1,2 は有意差があり，かつ平均値が高いことから，物体の位置認識や目的地の誘導に関して有効であったことがわかる．質問 3,5,6 に関しては有意差は見られなかったものの，平均値が全て高いスコアであり，物体との距離感の把握や用いた音に関して有効であったことがわかった．しかし，質問 4 は有意差が見られず，かつ他の項目と比べて平均値が低かった．これは，探索レーザーおよび扇形探索法の機能を考えると，「物体から物体へ向かう」という目的に関しては達成しやすかったが，マップ全体のどこにいるのかという情報をユーザに与える機能が不足していたため，他の項目よりもスコアが低くなったと考えられる．

5. まとめと今後の課題

本研究では，視覚障がい者のソーシャル VR の利用を可能にするために，視覚障がい者の仮想現実空間での空間把握と視覚障がい者の仮想現実空間でのコミュニケーションの 2 つに問題

を分け，それぞれ手法を提案した．空間把握を可能にする手法として，環境音や足音などの音を立体音響で作成し，加えて探索レーザーが物体に当たるとコントローラが周期的に振動し音が鳴るシステムを開発した．また，振動と音の周期は探索レーザーを使用するユーザとレーザーが当たっている物体の距離によって変化する．交流を可能にする手法として，HMD から頭の向き情報，即ち視線情報を取得し，音声案内で自分の周辺にいる他のユーザの情報を通知することで他のユーザとの交流を促す手法を提案した．また，その他に視覚障がい者が仮想現実空間を利用する上で有効であると考えられる機能について実装し，紹介した．

探索レーザーを用いた空間把握における有効性を調査するために，扇形探索法との比較実験を行った．その結果，SUS のスコアにおいてユーザビリティは高い水準であり，扇形探索法よりも優秀であることがわかった．また，ナビゲーションに関するアンケート内容の各項目においても，探索レーザーの方が平均値が高く，ナビゲーション機能において探索レーザーの方が良いことが確認できた．しかし，ユーザがマップのどのあたりにいるのかという情報を与える機能が不足しているため，改善が必要である．今後は，実際に視覚障がい者の方に対して評価実験を行い，システムの有効性を検証していくことを考えている．また，視覚障がい者の方から得られたフィードバックをもとにシステムの改善を行なっていく予定である．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293, 22K12274, 22H03700) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 厚生労働省, 平成 28 年生活のしづらさなどに関する調査 (全国在宅障害児・者等実態調査), https://www.mhlw.go.jp/toukei/list/seikatsu_chousa_h28.html(2024 年 1 月 4 日確認)．
[2] 関 喜一 :”視覚障害者のための音による空間認知の訓練技術-リハビリテーション現場での実用化に向けて-”，Synthesiology Vol.6 No.2, 2013.
[3] 大内 誠, 岩谷 幸雄, 鈴木 陽一 :”視覚障害者のための 3 次元聴覚情報の提示”，バイオメカニズム学会誌 Vol.31 No.2, 2007.
[4] 玉渕 誠人, 阿部 亨, 菅沼 拓夫: ”空間音響を用いた視覚障害者のための屋内ナビゲーションの提案”, 情報処理学会第 83 回全国大会　1U-06, 2021.
[5] Tiger Ji, Brianna R. Cochran ,Yuhang Zhao ” VRBubble: Enhancing Peripheral Awareness of Avatars for People with Visual Impairments in Social Virtual Reality ”， ASSETS 2022 : The 24th International ACM SIGACCESS Conference on Computers and Accessibility ， 2022.
[6] Sreekar Krishna, Dirk Colbry, John Black, Vineeth Balasubramanian, Sethuraman Panchanathan: ”A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interaction of People Who are Blind or Visually Impaired”, Workshop on Computer Vision Applications for the Visually Impaired, James Coughlan and Roberto Manduchi, Oct 2008, Marseille, France. inria-00325432.
[7] Meta Quest, Oculus Audio Spatializer の機能, https://developer.oculus.com/documentation/unity/audio-spatializer-features/?locale=ja_JP(2024 年 1 月 4 日確認).
[8] Meta, Oculus Quest2, https://www.meta.com/jp/quest/products/quest-2/(2024 年 1 月 4 日確認).
[9] Brooke, J. 1996. “SUS: A “quick and dirty” usability scale”. In Usability evaluation in industry, Edited by: Jordan, P. W., Thomas, B. A. Weerdmeester and McClelland, I. L. 189–194. London: Taylor , Francis.
[10] Sauro, J., Lewis, J. (2016). Quantifying the user experience: Practical statistics for user research. Amsterdam; Waltham, MA: Elsevier/Morgan Kaufmann.

立体音響 AR を用いたリラクゼーション支援手法の有効性評価

大塚　智貴† 　Panote Siriaraya†† 　中島　伸介†

† 京都産業大学　先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学 情報工学・人間科学系　〒 606–8585 京都市左京区松ヶ崎橋上町
E-mail: †{i2386030,nakajima}@cc.kyoto-su.ac.jp, ††spanote@kit.ac.jp

あらまし　厚生労働省による令和 2 年度の労働安全衛生調査によると，仕事や職業に関することで強い不安やストレスを感じる労働者は 54.3 %であると報告されており，令和元年に実施された国民生活基準調査によると日常生活で悩みやストレスがある人の割合は 47.9 %であると報告されている．これらの調査結果から人のリラクゼーションを支援する取り組みは重要であると考えられる．本研究ではリラックスできるスポットに適した物語と環境果音を立体音響で再生することで，ユーザのストレスを軽減させることが可能な立体音響 AR を用いたリラクゼーション支援システムの開発を目指している．本稿では立体音響 AR を用いたリラクゼーション支援手法の有効性について議論する．

キーワード　ユーザ支援，AR，音声，ヘルスケア，立体音響

1. はじめに

厚生労働省による令和 2 年度の労働安全衛生調査 [1] によると，仕事や職業に関することで強い不安やストレスを感じる労働者は 54.3 %であると報告されており，令和元年に実施された国民生活基準調査 [2] によると日常生活で悩みやストレスがある人の割合は 47.9 %であると報告されている．これらの調査結果から人のリラクゼーションを支援する取り組みは重要であると考えられる．ストレスを軽減させるものとして，川の流れる音や雨の降る音といった自然の音声を聞くことや，森や海岸といった自然の風景を眺めることがあげられる．また，近年では良質な睡眠を手に入れるための読み聞かせである「スリープストーリ」を提供している Calm というサービスがある．本研究では，これらに着目し，リラックスできるスポットに適した物語と環境音を立体音響で再生することで，リラックス効果を高め，ユーザのストレスを軽減させることが可能な立体音響 AR を用いたリラクゼーション支援システムの開発を目指している．本稿では立体音響 AR を用いたリラクゼーション支援手法の有効性について議論する．

2. 関連研究

人間をリラックスさせる手法には，香りや音楽を用いるものやマッサージなど多くの種類があり，様々な手法を用いてリラックス効果の検証が行われいる．大学生に対する絵本の読み聞かせに関する研究 [3] によると，大人向けの絵本の読み聞かせることで,「肯定的感情」の得点が高く,「否定的感情」が低くなることが報告されている．

また，綺麗な風景を見ることや，森の中を歩くこともリラックスに繋がると言われている．森林セラピーに関する調査 [4] によると，主観評価の結果から森林による座観は都市に比較して有意に快適で鎮静的であり，自然な感じがすると評価されている．この評価に加えて，ストレスーリフレッシュ感調査では森林で座観することにより都市よりも有意に高いリフレッシュが得られることが報告されており，POMS の結果では森林においては，座観後の活気が都市よりも有意に高く，緊張ー不安，抑うつー落ち込み，怒りー敵意，疲労，混乱といったネガティブな感情は有意に低いことが報告されている．他にも，唾液アミラーゼ活性を指標とした河川・湖沼ののストレス軽減効果を調査した研究 [5] では，多くの試験エリアで無次元唾液アミラーゼ活性の平均値が市街地＞中間＞水際地点という大小関係となり，水辺におけるストレス軽減効果が顕著であると報告されている．そして，河川・海岸におけるストレス軽減効果と音熱環境の影響の検討をした研究 [6] によると，水辺と周辺域の市街地において，水辺では周辺域よりもストレス軽減効果が高く，ストレス軽減効果の要因である音・熱環境共に，水辺の方が周辺域よりも快適性が高いことが報告されている．このように，森や海といった自然にはストレスを軽減させる効果がある．また，風景の他にも，焚き火や波の音を聞くこともストレス軽減の効果があると言われている．以上の通り，風景を眺めつつ，物語と環境音を聴きながら散策することは，リラクゼーションに効果があると考えられる（図 1 参照）．

図 1　関連研究から考察したイメージ図

図 2　リラクゼーション支援システムの概要

3.　提 案 手 法

3.1　リラクゼーション支援システムの概要

本研究で開発する立体音響 AR を用いたリラクゼーション支援システムでは，リラクゼーションに有効な風景を眺めることが可能なスポットの推薦を行うと共に，各スポットに適した物語や環境音を再生することで，人々のリラクゼーションを支援することを目指している（図 2 参照）．そのため，ユーザがシステムを起動させた際に，現在地をスマートフォンの GPS 機能により取得することで，その周辺の公園や海辺等のリラクゼーションスポットが推薦される．また，各リラクゼーションスポットに訪れた際には，そのスポットの特徴に応じた物語や環境音を立体音響で再生することができる．なお，リラックスできるスポットや物語は，ユーザによって異なるため，日頃の散策エリアやコンテンツ視聴履歴を分析することで，推薦スポットやコンテンツの個人化も可能である．

3.2　音源設置手法の検討

本研究で開発するリラクゼーション支援システムでは，ユーザがリラクゼーションスポットに行き，スポットの特徴に応じた物語や環境音を立体音響で再生させる．そのため，ユーザの現在地の座標を GPS 機能により取得し，その座標をもとに仮想空間内にある環境音の音源をスポットの特徴に対応するように設置する必要がある．本研究では，この機能の実装に加え，地図の表示やユーザの向いている方角を取得することなどを考慮して，デジタル地図の開発プラットフォームである mapbox を用いる．mapbox を用いて音源を設置した地図を図 3 に示す．

図 3　mapbox を用いて音源を設置した地図

3.3　システム実装に向けた検討

システムの実装のため以下の課題に取り組む．

- リラクゼーションスポットの取得および DB 化

各エリアのリラクゼーションスポットを取得すると共に，各スポットの特徴を表すキーワード抽出を行い，リラクゼーションスポット DB を構築する．

- 物語および関連する環境音の収集および DB 化

物語および環境音についてはフリー素材を中心に収集し，各コンテンツの特徴を表すキーワード抽出を行い，物語と環境音の DB 構築を行う．

- ユーザの現在地に基づくスポット推薦

ユーザの興味と位置情報に基づいて，周辺のリラクゼーションスポットの推薦を行う．

- スポットに対する物語および環境音のマッチング

各スポットの特徴キーワードと，物語と環境音の特徴キーワードを用いて，スポットに対する物語および環境音のマッチングを行う．

- 散策エリアと視聴履歴の分析に基づく個人化

散策エリアの情報とコンテンツ視聴履歴のデータを蓄積し分析することで，個人に適したスポットおよびコンテンツの推薦を行う．

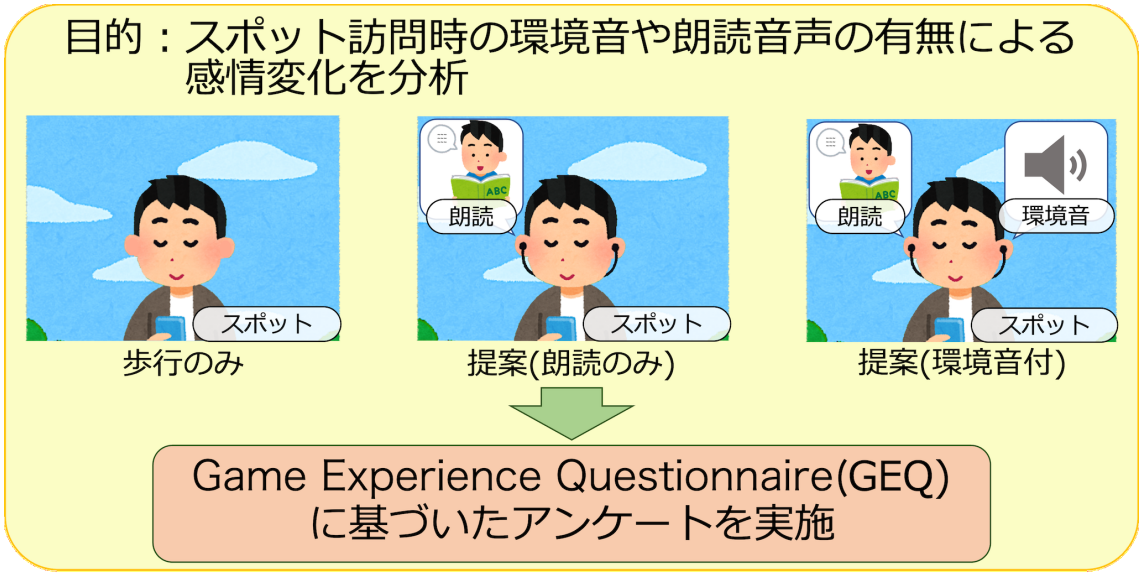


図 4　評価実験の概要図

構成要素	質問項目
ポジティブ	満足したと感じた
	アプリは楽しかった
	幸せに感じた
	いい気分になった
	楽しかった
ネガティブ	気分が悪くなった
	他のことを考えていた
	面倒だと感じた
	退屈に感じた
没入感	アプリのストーリーに興味を持った
	見た目が美しいと感じた
	空想の世界に入ったように感じた
	物事を探求できると感じた
	強く印象に残った
	リッチな体験だと感じた
緊張感	腹立たしく感じた
	じれったいと感じた
	イライラした

図 5　評価実験で用いた構成要素と質問項目

4.　環境音や朗読の有無による感情変化の分析

4.1　概　　要

本稿では，提案手法の有効性評価として，スポット訪問時の環境音や朗読音声の有無による感情変化を分析する評価実験を実施した．図 4 は評価実験の概要図である．

4.2　実 験 手 順

評価実験では，20 代の大学生の男性 5 名を被験者とし，「歩行のみの散策」と「朗読を聞きながらの散策」，「環境音付きの朗読を聞きながらの散策」を実施した（以降，「歩行のみの散策」は「歩行のみ」，「朗読を聞きながらの散策」は「朗読のみ」，「環境音付きの朗読を聞きながらの散策」は「環境音付きの朗読」とする）．そして各散策の後に，ゲーム経験を計測するためのアンケートである Game Experience Questionnaire（Game　GEQ）[7] に基づくアンケートに回答してもらい，分析を行った．

4.3　分 析 手 法

評価実験では，GEQ にあるコアモジュールのうち「ポジティブ感情，ネガティブ感情，没入感，緊張感」に関するものを取り出し，本研究に適するように修正してアンケートを実施した．図 5 は構成要素と質問項目である．収集したアンケート結果は各散策ごとに「ポジティブ感情，ネガティブ感情，没入感，緊張感」の合計値を算出し分析を行った．しかし，没入感に関してはストーリの有無が関係するため，「歩行のみ」を省いて分析を行った．

4.4　分 析 結 果

図 6 に GEQ の分析結果を示す．

- ポジティブ感情

ポジティブ感情に関しては，「歩行のみ＜朗読のみ＜環境音付きの朗読」の順で合計値が大きいことを確認した．特に，「歩行のみ」と他の二つには大きな差がある．そのため，環境音と朗読はポジティブ感情の向上に有効であることを確認した．

- ネガティブ感情

ネガティブ感情に関しては，「朗読のみ＜環境音付きの朗読＜歩行のみ」の順で合計値が大きいことを確認した．そのため，「歩行のみ」よりも朗読音声があるほうがネガティブ感情が低下することを確認した．しかし，環境音はネガティブ感情の低下には有効とは限らないことを確認した．

- 没入感

没入感に関しては，「朗読のみ」よりも「環境音付きの朗読」の方が合計値が大きいことを確認した．そのため，環境音は没入感の向上に有効であることを確認した．

- 緊張感

緊張感に関しては，「環境音付きの朗読＜歩行のみ＜朗読のみ」の順で合計値が大きいことを確認した．特に「歩行のみ」と「朗読のみ」の合計値に比べ，「環境音付きの朗読」の合計値が小さいことから，環境音は緊張感の低下に有効であることを確認した．

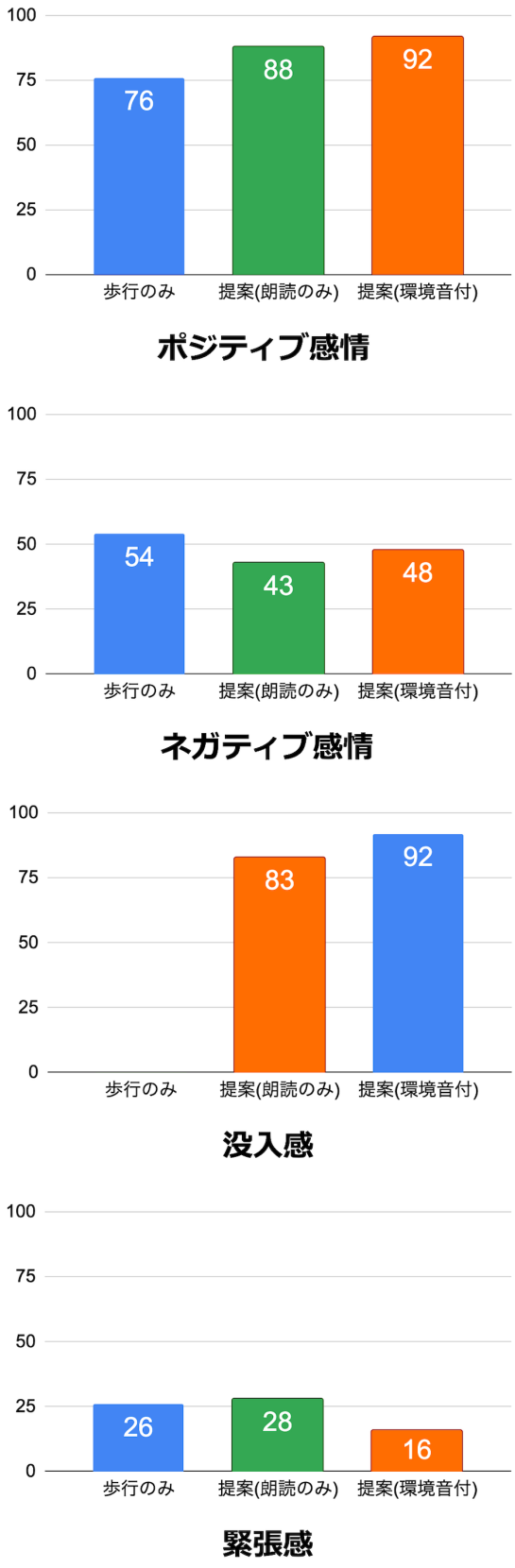


図 6　GEQ の分析結果

4.5　評価実験：まとめ

評価実験の分析結果より，環境音と朗読にはポジティブ感情を高める効果があり，環境音に関しては没入感を高め，緊張感を和らげる効果があることを確認した．しかし，環境音の存在はネガティブ感情の低下には有効とは限らないということを確認した．

5.　物語と朗読音声に関する有効性評価

5.1　概　　要

環境音や朗読の有無による感情変化の分析では，パブリックドメインの書籍の朗読音声を提供している LibriVox を利用していた．しかし，LibriVox には日本語の朗読音声が少ないことや，書籍によって朗読音声の長さが異なるといった短所がある．そのため，本研究では ChatGPT にテーマや文字数に加え，タイトルをつけることや登場人物のセリフを入れることなどの条件を含んだプロンプトを送ることで，スポットの特徴を含む物語の生成し，VOICEVOX を用いて朗読音声に変換することを検討した．本章では，リラクゼーション支援システムにて利用する物語と朗読音声の有効性評価の結果について報告する．評価実験では，パブリックドメインの書籍と ChatGPT で生成した物語の比較と，LibriVox にて提供されているパブリックドメインの書籍の朗読音声と同じ書籍を VOICEVOX で朗読音声を生成して聞き取りやすさ比較を行った．評価実験では「桜」「林檎畑」「山・野原」を物語のテーマとして選択した．また，パブリックドメインの書籍が長編である場合は，選択したテーマが背景描写に含まれるシーンを切り出して比較を行った．

5.2　被　験　者

評価実験は，本学に所属する大学生 7 名 (男性：4 名，女性：3 名) と大学院生の男性 5 名の合計 12 名を被験者として行った．また，被験者の年代は 12 名全員が 20 代であった．

5.3　分 析 手 法

評価実験はリッカート尺度 7 段階を用いて行った．評価項目に関しては，物語の比較では「物語として妥当である」「テーマに適している」「面白い」「理解しやすい」の 4 つを評価し，朗読音声の比較では「聞き取りやすさ」について評価を行った．そして，各評価項目の平均値を用いて分析を行った．

5.4　結果と考察

物語の比較結果を図 7，図 8，図 9 に示し，朗読音声の聞き取りやすさの比較結果を図 10 に示す．物語の比較結果より，パブリックドメインの書籍よりも ChatGPT の作品の方が 4 つの項目全ての平均値が高いことが確認された．特に「理解しやすさ」の差が大きいことが確認された．これは ChatGPT の作品には難しい表現や言い回しがなく，理解しやすいのではないかと考えられる．また，朗読音声の聞き取りやすさの比較結果より，人間の声の朗読音声と VOICEVOX で生成した朗読音声には，聞き取りやすさに大きな差がないことが確認された．このことから，VOICEVOX による朗読音声に対する違和感はほとんどないと考えられる．従って，ChatGPT で生成した物語や VOICEVOX で生成した朗読音声には有効性があると考えられる．

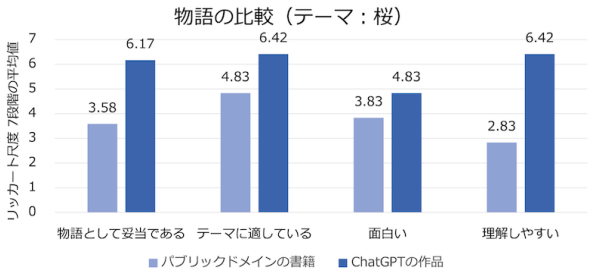


図 7　物語の比較結果（テーマ：桜）

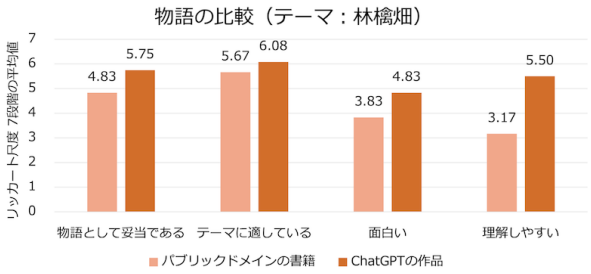


図 8　物語の比較結果（テーマ：林檎畑）

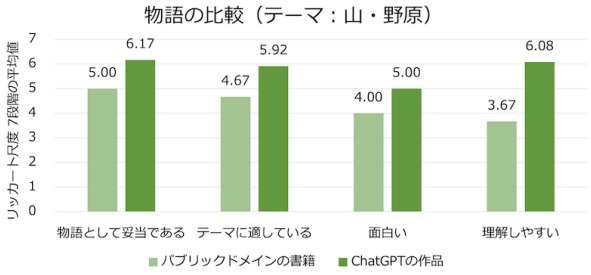


図 9　物語の比較結果（テーマ：山・野原）

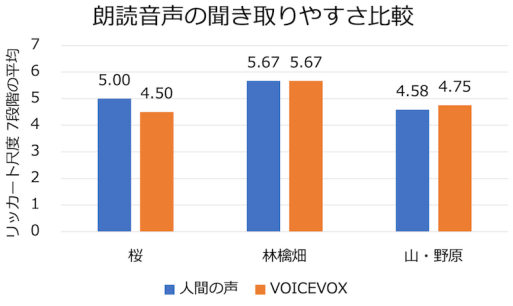


図 10　朗読音声の聞き取りやすさの比較結果

6.　提案システムを用いた有効性評価の実験計画

6.1　概　　要

本章では，立体音響 AR を用いたリラクゼーション支援手法の有効性について，評価実験計画について説明する．

6.2　被　験　者

有効性評価では，リラクゼーションにシステムが干渉することで，被験者のストレスがどのように変化するのかを分析する．そのため，大学生 20 名をシステムの干渉ありのグループとと干渉なしのグループの 2 つに分けて分析することを計画している．

6.3　分 析 手 法

有効性評価では，被験者のストレスの度合いを STAI と VAS のアンケートを用いて分析を行う．STAI とは，Spielberger 氏の作成した調査指標であり，State Anxiety（状態不安）と Trait Anxiety（特性不安）に分けて計測を行うことで，被験者の現在のストレス状態と，普段の状態を分析することが可能である．また VAS とは，心理測定反応スケールの一種であり，評価実験では紙の中央に質問内容と 10cm の線の端に「全くストレスを感じない」と「すごくストレスを感じる」というラベルをつけたもの (図 11) を使用し，被験者は現在のストレスレベルの位置に印をつけて解答する．

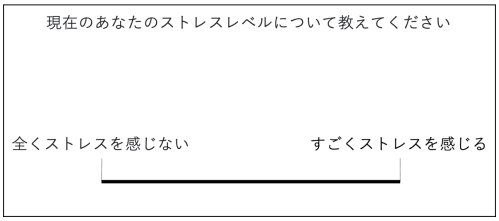


図 11　VAS のイメージ図

6.4　ストレス誘発

有効性評価では被験者にストレス誘発を施し，その後リラクゼーションに取り組んでもらうことを計画している．ストレス誘発では，Almazrouei ら [8] の開発したストレス誘発のための計算問題と一般常識問題を参考にした問題を口頭形式で被験者に解答してもらうことを計画している．また，ストレス誘発させるために被験者には解答前に，時間制限（文章題の計算問題は 2 分，簡単な計算問題と一般常識問題は 30 秒）があることに加えて，解答は一度切りであり，半分以上の正答率が求められることを伝える．

6.5　有効性評価の手順

有効性評価の概要図を 12 に示す．有効性評価では，初めに実験前のストレス状態を確認するために，被験者は STAI と VAS に回答する．次に，ストレス誘発を実施し，誘発の確認のために被験者は STAI と VAS に回答する．その後，被験者はリラクゼーションに取り組む．この時，システム干渉ありの被験者はイヤホンを付け景色を眺めながら立体音響で環境音と朗読音声を聴き，システム干渉なしの被験者は景色を眺める．そして，リラクゼーション後に被験者は STAI と VAS の回答に加え，4. 章の評価実験で用いた GEQ アンケートに回答し，有効性評価は終了となる．

図 12　有効性評価の概要図

7. おわりに

本稿では，立体音響 AR を用いたリラクゼーション支援手法の有効性について議論を行った．物語と朗読音声の有効性評価では，ChatGPT で生成した物語や VOICEVOX で生成した朗読音声には有効性があると考えられた．また，提案システムを用いた有効性評価の実験計画では，大学生 20 名をシステムの干渉ありのグループと干渉なしのグループに分け，ストレス誘発の前後に STAI と VAS による分析を行い，リラクゼーション後では STAI と VAS に加えて GEQ を用いて分析を行うことを検討している．今後は，立体音響 AR を用いたリラクゼーション支援手法の有効性評価に取り組み，その結果をもとにリラクゼーション支援手法の検討することが課題となる．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22K12274) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 厚生労働省, 令和 2 年　労働安全衛生調査（実態調査）(https://www.mhlw.go.jp/toukei/list/r02-46-50b.html,2024 年 1 月 4 日確認)
[2] 厚生労働省，2019 年　国民生活基礎調査の概況 (https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa19/index.html,2024 年 1 月 4 日確認)
[3] 奥野洋子, 大学生に対する絵本の読み聞かせによるリラックス効果の検討—情動知能との関連について, 近畿大学心理臨床・教育相談センター紀要, 4 巻, p.1-13, 2020.
[4] 恒次 祐子, 朴 範鎭, 李 宙営, 香川 隆英, 宮崎 良文, 森林セラピーの心理的リラックス効果, 日本衛生学雑誌, 66 巻, 4 号, p.670-676, 2011.
[5] 本宮和英・二瓶泰雄：唾液アミラーゼ活性を指標とした河川・湖沼のストレス軽減効果把握の試み，河川技術論文集，Vol.17，pp.383-388，2011.
[6] 塩田 彩夏, 二瓶 泰雄, 遠藤 亮之輔, 河川・海岸におけるストレス軽減効果と音・熱環境の影響の検討, 土木学会論文集 B1（水工学）, 69 巻, 4 号, p. I_1699-I_1704, 2013.
[7] IJsselsteijn, W. A., De Kort, Y. A. W., & Poels, K. , “The game experience questionnaire.”, 2013.
[8] Almazrouei, M.A., Morgan, R.M. & Dror, I.E. A method to induce stress in human subjects in online research environments. Behav Res 55, 2575–2582 (2023).

MR空間におけるホワイトボードに対するインタラクションを伴う商品情報の探索

津田　裕哉†　　山本　岳洋†　　大島　裕明†

† 兵庫県立大学 大学院情報科学研究科　〒 651-2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †ad22o041@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，ホワイトボードと付箋を使った情報整理手法と複合現実（Mixed Reality; MR）デバイス，商品の情報が表示された仮想オブジェクトを用いて，購入する商品を探す商品探索手法を提案する．提案手法では，我々が普段から取り組むような商品検索，例えばカメラなどを検索する時に，空間的に提示された検索結果をホワイトボードに付箋を貼るように配置して整理したり，仮想オブジェクトに対する注釈を手書きの文字で残すことが可能となる．さらに提案手法のユーザビリティや有用性を評価するために，実際にユーザにカメラの検索を行ってもらう評価実験を実施した．

キーワード　XR（VR/AR/MR），インタラクティブシステム，情報探索行動，情報組織化

1　はじめに

情報化が進む現代社会において，情報を探索するという行動は日常生活の中で欠かせないものとなっている．特に近年では，PC やスマートフォンのディスプレイ上で商品を検索し，その場で購入することも多い．将来さらに進んでいくと思われる情報化に伴って検索対象も年々多種多様なものになり，商品の検索，探索はより複雑になっていくと考えられる．このような問題を解決するために，AR や VR デバイスを使った空間の中で情報の探索を行い，3 次元的に結果を提示する研究は以前から行われている．しかし，VR デバイスなどによる仮想空間内に提示された検索結果をどのように整理するかといったことを扱った研究はまだ少ない．

散らばった情報を効果的に整理する方法の一つとして，図 1 のようなホワイトボードと付箋を使った整理方法が挙げられる．この方法は KJ 法などとも呼ばれ，主にブレインストーミングなどで用いられる．この方法では，付箋をホワイトボード上で自由に動かして情報を分類したり，手書きの線や文字を使って情報を視覚的に整理して意思決定に活かしたりすることができる．そこで我々は，付箋とホワイトボードを使った情報整理方法と，Mixed Reality デバイスを組み合わせて商品の検索，探索をすることを考えた．

Mixed Reality（MR）とは，複合現実感とも訳され [21]，現実の世界とコンピュータによる仮想の世界を融合させる技術を指す．MR という言葉は，Milgram ら [12] が 1994 年に初めて明示的に提唱した．この技術を実現するための方法は以前から活発に研究，開発されており，頭部に装着する Head Mounted Display（HMD）を用いた方法が最も一般的である．半透明のレンズで眼前を覆う HMD を用いて，現実の視界にデジタルな仮想オブジェクトを融合して表示する光学シースルー方式や，HMD に搭載されたカメラ映像に仮想オブジェクトを融合させ，視界に投影するビデオシースルー方式などがある [20]．光学シースルー方式の MR HMD としては Microsoft 社の HoloLens シリーズなど，ビデオシースルー方式のデバイスとしては Apple 社の Apple Vision Pro などがある．

図 1　ホワイトボードと付箋を使った情報整理の例

本研究では，MR デバイスの現実と仮想の世界を融合できるという特徴を利用して，MR デバイス環境下で商品を検索し，仮想のパネルに提示された商品情報を実際のホワイトボード上で整理，探索する商品探索手法を提案する．

提案手法では，検索にヒットした商品が表示された仮想パネルを，現実の空間内に配置したり，付箋を貼るようにしてホワイドボード上に配置したりして，商品を自由に閲覧することができる．また，仮想パネルに対してホワイトボードマーカを使って注釈を手書きしたり，手書きの線で商品を分類したりしながら情報を整理して，購入する商品を吟味することができる．実際に提案手法を使ってカメラを検索，探索している様子を図 2 に示す．

本論文では，続く 2 節で複合現実の技術を使った研究や，VR などの没入型環境における情報探索の先行研究を述べる．次に 3 節で必要な機能，実装した MR アプリケーションの概要について説明したあと，4 節で評価方法と評価の結果について，5 節で今後の課題について記述し，まとめとする．

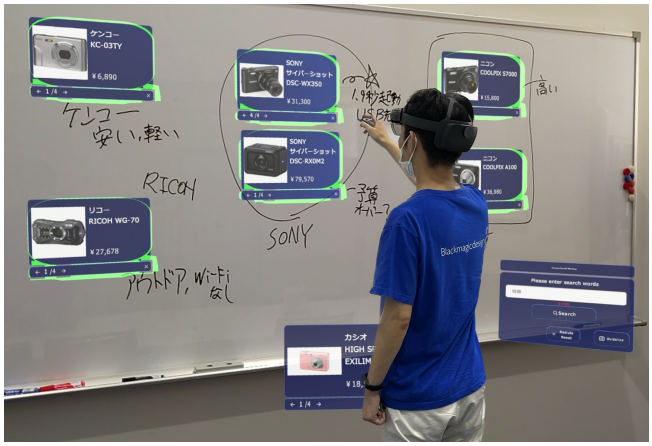


図 2　MR 空間の中で商品を探索している様子

2　関 連 研 究

2.1　複合現実空間における情報の提示

Weizhou ら [10] は，AR デバイス環境において仮想コンテンツを現実空間内で扱う場合のレイアウトについて調査している．論文では，文書分類タスクを例に，個人での作業あるいは共同作業，家具の有無など条件をいくつか変えて，被験者が仮想コンテンツをどう配置して作業していたかという実験を行った．被験者は主に，机やホワイトボードの上に仮想コンテンツを配置して作業する傾向にあることが示されている．

Zhen ら [9] は MR デバイスを使って，紙の文書と仮想の文章を融合させる HoloDoc というシステムを提案している．論文では，物理的な紙の文書に，MR デバイスを使ってデジタルな情報を付加し，紙の文書に対するインタラクションを拡張する形で，現実の紙の文書とデジタルな文書の利点を融合させていた．評価実験により，MR デバイスと物理的な文書を組み合わせることで MR デバイスに慣れていないユーザーでも直感的に操作できること，文書中の引用を MR 空間でタップすることで電子的に参考文献を展開できる機能が最も好まれていたことなどが示されている．

Daskalogrigorakis ら [5] は，MR デバイスを用いて多くの情報や UI を一目で把握することができるようなインタフェース「Holo-Box」を提案している．Holo-Box は，ユーザが明示的に仮想オブジェクトに表示された情報量の細かさの度合い（Level of Detail; LOD）を調整できるインタフェースである．ユーザが LOD を調節できるようにすると，MR デバイスで一般的に用いられる空中に浮かぶボタンの操作や，視線による入力のミスを減らすことができるとされている．

2.2　没入型環境での情報探索

従来のような PC のディスプレイ上で行う情報検索や結果の提示を，VR デバイスなどを使って 3 次元空間に拡張する研究は以前から行われている [1], [8], [14], [18], [22]．Nakazato ら [14] は画像検索の結果を，CAVE [4] を使った没入型環境内で立体的に提示する手法を提案している．この手法では，CAVE 環境内に提示された検索結果の画像を，ユーザがコントローラーで強調表示させたり，自由に移動させたりすることができる．3 次元空間で検索結果を閲覧，操作するという点は Nakazato ら [14] や Ward ら [18] の研究と本研究との類似点であるが，本研究では光学シースルー型の MR デバイスを用いて，より現実空間に即した環境で検索することを目的としている．

Ward ら [18] は，仮想空間における検索結果の提示方法について調査している．論文では，仮想現実空間内で検索結果を縦方向に羅列するリスト型，2 次元的に並べるグリッド型，自分を中心に扇上に並べるアーク型の 3 種類の提示方法を評価していた．複数並べられた検索結果の中から特定の結果を見つけ出すというタスクにおいて，ユーザはアーク型を使用するとより早くタスクを完了させることができるが，アーク型はユーザからの好みが低く，リスト型やグリッド型をより好むという結果が示されている．本研究は，このアーク型での情報の提示方法を参考にしている．

Büschel ら [1] は，複合現実や拡張現実の技術を用いた，より現実世界に即したリアルタイムな情報検索方法を研究している．該当論文の中では Reality-Based Information Retrieval（RBIR）と呼称しており，HMD が普及した社会における次世代の検索インタフェースとして提案されている．そのインタフェースの一例として，MR デバイスを用いて視界に映った食材を認識し，その食材を使用するレシピを検索して MR 環境内で空間的に提示するプロトタイプを提案している．本研究の提案手法はこの RBIR の一種であるといえる．

2.3　複合現実空間における情報の操作

MR 環境下では，3 次元空間で情報を扱うために，平面のディスプレイを使う PC とは違うインタラクション方法が以前から研究されている [6], [15], [17], [19]．Poupyrev ら [15] は MR 空間内で AR マーカが描かれたタイルに仮想的に情報を紐づけ，タイルを動かすことによって情報を操作するインタフェースを研究している．この研究ではホワイトボード上でタイルを操作していたり，ホワイトボードマーカで注釈をつけていたりしている．使用例として，航空機の操縦パネルのプロトタイピングを行っていた．情報の操作や整理のために MR デバイスとホワイトボードを組み合わせている点は Poupyrev らの研究 [15] と本研究との共通点である．しかし，本研究の手法ではこのタイルに相当するもの自体を仮想のオブジェクトとしており，インタフェースの実体感が少なくなる代わりに AR マーカが描かれたタイルを複数枚用意する必要がない．

上坂ら [19] は，MR 環境での対話操作において，普段使い慣れた道具をデバイスとして用いる研究をしている．そうすることで，普段から使用する場合に近い触感や操作音で MR 環境における情報の選択，移動操作を実現していた．論文内では道具型デバイスとしてピンセット型デバイスを挙げている．こういった現実の道具を，情報を操作するインタフェースとして使うことをタンジブルユーザインタフェース（TUI; Tangible User Interface）と呼び，Ishii らの研究 [7] が元となっている．また，このような実体のあるインタフェースと複合現実，拡張現

実の技術と組み合わせたものは Tangible Augmented Reality [2] とも呼ばれる．本研究の提案手法は，検索結果情報の空間的な提示と，Tangible Augmented Reality を一部組み合わせた検索手法であるといえる．

3 MR 空間で行う商品探索

3.1 概　　要

本論文では，MR デバイスを用いた複合現実環境内で，商品の情報を検索，整理する手法を提案する．以下に，提案手法における検索，探索作業の一連の流れを示す．

（1） 検索したい商品の特徴や単語を検索ボックスに入力し，検索する．
（2） 検索にヒットした商品が表示された仮想オブジェクトが生成される（図 3）．
（3） 生成された検索結果オブジェクトを図 2 のようにホワイトボード上で整理する．
（4） 必要に応じて，ホワイトボードマーカを使って注釈をつけたり，キャプチャ機能を使って情報を付与する．
（5） 検索結果を整理しながら商品を吟味し，買いたい商品を決定する．

例えば，アウトドアに持って行けるような SONY 製のデジタルカメラを探しているとする．ユーザはまず従来の検索のように「SONY デジタルカメラ 軽量」といった検索したい単語を検索ボックスに入力し，検索する．そうすると，ヒットしたカメラの情報が表示された仮想オブジェクトが MR 空間内に生成される．ユーザは生成された検索結果オブジェクトの中から，求めているカメラの特徴に合致しない商品を視界の外に省いたり，似たカメラをホワイトボード上で並べたりして結果を整理する．整理の仕方として，例えばホワイドボード上でカメラを値段順に整列させたり，レンズや本体機能ごとにカメラを分類して配置したりすることが考えられる．そして必要に応じて，気になったカメラの特徴を注釈としてホワイトボードに手書きしたり，分類したカメラを手書きのラインで囲むなどしてカメラを視覚的に整理，買いたいカメラを吟味していく．

なお，研究で実装したアプリケーションの詳細は，動画 [1]で確認することができる．

3.2 開 発 環 境

本研究では，MR デバイスとして Microsoft 社が開発，販売している HoloLens2 を用いた．HoloLens2 上で動作するアプリケーションの開発にはゲームエンジンである Unity [2] を使用し，アプリケーション内のユーザインタフェースやインタラクションの実装には，Microsoft 社が提供している Mixed Reality Toolkit 3 [3]（MRTK3）を活用した．使用した各ソフトウェアとそのバージョンを表 1 に示す．

表 1　使用したソフトウェアのバージョン

Unity	2021.3.11f1
Mixed Reality Toolkit	3.17
Visual Studio 2019	16.11.11

表 2　検索対象となるデータに含まれるカメラの特徴

商品名，型番	メーカー名
商品の値段	商品説明文
画素数	撮像センサの種類
撮影感度	ファインダの有無
焦点距離	ズーム倍率
本体サイズ	重量
起動スピード	顔認識機能の有無 など

3.3 検索アプリケーションに用いた商品データ

本研究で開発したアプリケーションでは，本格的な Web 上での検索の前に，あらかじめ Web ページからクローリングした商品データの中から検索することとした．価格.com という PC やカメラ，家電などの値段やスペックを比較できるサイトのデジタルカメラのページ [4]から，値段などの情報をクローリングしたデータを検索対象とする．デジタルカメラという商品カテゴリには，F 値や焦点距離など商品のパラメータが豊富な点，商品の外観が似通っておらずデザインが様々な点，値段が数千円から数十万円と価格帯の幅が広い点などの利点があり，複数の観点で商品を比較しやすいという理由からデジタルカメラを検索対象に選んだ．アプリケーションの中で主に扱うカメラの特徴を表 2 に示す．

3.4 検索結果の提示

検索システムでは，検索でヒットした検索結果の情報を視覚的にわかりやすく表示する必要がある．提案手法では入力したワードで検索すると，まず商品の情報が表示されたパネル状の仮想オブジェクトが図 3 のようにユーザを中心に扇状に生成される．この生成形式は，Ward らの先行研究 [18] を参考にしている．生成された時，値段順で整理，選別をしやすいように検索結果パネルは左端から値段の昇順に生成される．図 3 は，全天球写真を撮影可能なデジタルカメラである RICOH 社の THETA シリーズを検索した時の例である．入力した検索ワードでヒットした商品の情報を表示する仮想オブジェクトの例を図 4 に示す．以下，図 4 のような検索結果が表示されている仮想オブジェクトを検索結果パネルと呼称する．パネル上部には商品画像，メーカー名，商品名，値段の基礎情報が常に表示されており，パネル下部には詳細な商品情報が表示される．パネル左下端にある左右の矢印ボタンを指で押すことで，パネル下部に表示する特徴を変更することができる．図 4 では例としてメーカーによる商品の説明文を表示している．ここでは，矢印ボタンを押すことでレンズのスペック，撮像素子のスペック，本体サイズなどの特徴に切り替えることが可能である．ユーザは生成された複数の検索結果パネルを見て，探している商品の

1：https://youtu.be/GFVDdFifviY
2：https://unity.com/
3：https://learn.microsoft.com/ja-jp/windows/mixed-reality/mrtk-unity/mrtk3-overview/
4：https://kakaku.com/camera/digital-camera/

図 3　検索結果パネルが生成された時の様子

図 4　検索結果の詳細を表示するパネル

特徴とはそぐわないパネルを手で視界の外へ移動させたり，空間から消去することである程度選別する．その後，残った結果を吟味するために，ホワイトボードの上に配置し，検索結果を整理する．

3.5　情報量の調節

複合現実環境で検索結果パネルを複数並べると，視認できる情報量が多くなりすぎてしまうという問題がある．これは従来のような Web ブラウザでの商品検索でも起こりうる問題である．特に情報量が多いサイトでは効率的に商品を探すことが難しい場合も少なくない．この問題を解決するために，ユーザとパネルの間の距離によってパネルの情報量を調節する．この機能により，空間を歩きながら自分とパネルとの距離を増減させれば，一度に視認する情報量を動的に調整することができる．アプリケーションでは，図 5 のように距離によって情報量を 2 段階で調整することが可能である．検索結果パネルから距離を取れば，商品説明文や詳細なパラメータが非表示となり，大きな商品画像と商品名，値段のみの表示になる．反対に検索結果パネルに近づいていけば，商品画像や値段に加えて商品説明文や各種パラメータが表示され，詳細な情報を確認できるようになる．

3.6　検索結果の整理

複雑な情報を整理する手段としては，手書きのメモを残しながら整理していくことが効果的とされる．Shibata ら [16] によれば，手書きで情報を残すことはタイピングに比べ認知負荷が低く，記憶に定着しやすいことがわかっている．また，一般

図 5　(a) パネルに近づいた時．(b) パネルから遠ざかった時．

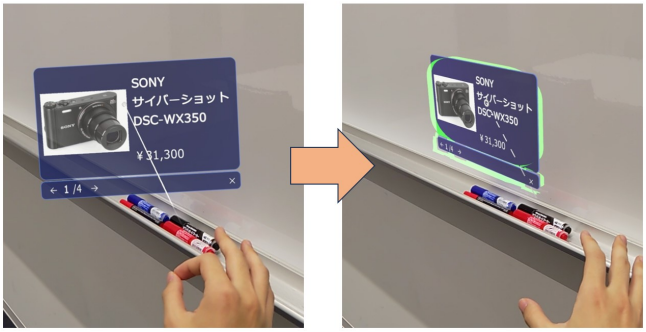


図 6　検索結果パネルをホワイトボード上に配置する操作

的には物理的なフィードバックのあるインタフェースの方が好まれる [13] が，物理的なフィードバックがなく外界が見えない VR 空間内で文字を手書きすることは難しい．提案手法では，MR デバイスの現実と仮想の世界を融合できる点と，付箋とホワイトボードを使った情報の整理手法を組み合わせて活用する．MR 空間内に提示された検索結果を現実のホワイトボードの上に付箋のように自由に配置して商品を閲覧する，手書きの線や矢印などで検索結果を分類，比較する，ホワイトボード上の検索結果に対して手書きで注釈等を加えるといったインタラクションによって，提示された検索結果を視覚的に整理して購入する商品を吟味することができる．

実際に MR アプリケーション内で，検索結果パネルをホワイトボード表面に配置する操作をしている様子を図 6 に，検査結果パネルに対してメモを残している様子を図 7 に示す．直接手でパネルをつまむ，あるいは人差し指のあたりから前方に出ている白いラインをパネルに当てた状態でつまむ動作をすると，パネルを一時的に把持することができる．その状態でホワイトボードに近づけるとパネルの縁の色が変わり，手を離すとホワイトボード上に平行に配置される．配置された後も，同様の操作でパネルをつまめばホワイドボード上の好きな位置に配置し直すことができる．

3.7　商品の実寸大の大きさの確認

Web ページ上での商品検索では，商品の外観や大きさを確認するために商品情報に含まれる商品画像を参照することが多い．しかし，Web ページ上で確認できる商品画像は一般的には二次元の画像であり，PC などのディスプレイ上で実際の大きさを把握するのは難しい．ここで提案手法では，商品データに含まれる本体サイズのパラメータを利用する．検索結果パネルにあるボタンを押すと，本体サイズのデータを参照し，複合現実空間内に実際の商品と同じ大きさの白い直方体を表示する．そう

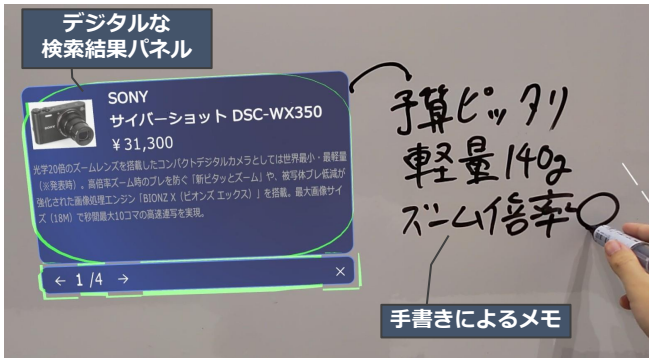


図 7　検索結果パネルに対してメモを残している様子

図 8　商品のサイズと同じ大きさのオブジェクトで実際のサイズを確認している様子

することで，商品の実際の大きさを手に取って立体的に確認することができる．実寸大の直方体を MR 空間内に表示している様子を図 8 に示す．直方体はハンドトラッキングにより実際に手で操作することができ，自由に移動や回転が可能である．商品と同じ大きさのオブジェクトを手に取るようにしてサイズを確認することができる機能は，平面のディスプレイを使う PC による商品検索にはないインタラクションであり，頭部装着型のディスプレイを用いる MR デバイスの大きな利点であるといえる．検索結果が複数あれば実寸大のオブジェクトを商品ごとに表示させ，実際のサイズを横に並べて比較するといったこともできる．

3.8　現実世界の情報の利用

3.8.1　視界のキャプチャ

商品検索の過程で，買いたい商品の候補をいくつか絞った後に，その商品を残した決め手，あるいはある商品を除外した理由を忘れしまうことがある．また，PC で商品を検索している最中に手書きのメモを残しても，検索結果の商品とその場で電子的に紐づけることは難しい．そこで本研究のホワイトボードを用いる点を利用し，ホワイドボード上の手書きのメモや注釈を撮影し，MR 空間内で電子的に保持する機能を取り入れた．以下，撮影した画像が表示されたオブジェクトをキャプチャオブジェクトと呼称する．この機能によって，現実世界のみに存在していた手書きの情報をデジタルなオブジェクトとすることができ，検索結果と手書きのメモの紐づけや，手書きの情報を

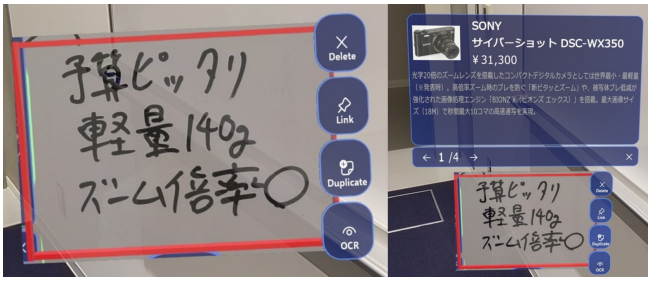


図 9　(a) キャプチャオブジェクト，(b) キャプチャオブジェクトをパネルに紐づけた時の様子

デジタルな情報として保存することが可能となる．実装したアプリケーションでは，HoloLens2 の前面に搭載されたカメラで外界を撮影し，検索結果パネルと同様に手で持つことのできるオブジェクトとして MR 空間内に生成する．キャプチャオブジェクトの外観及び，キャプチャを検索結果パネルに紐づけた時の様子を図 9 に示す．キャプチャオブジェクトを検索結果パネルに紐づけると，キャプチャオブジェクトはパネルの下部に固定され，パネルに追従して動くようになる．キャプチャオブジェクトには 4 つのボタンが付いており，上から（1）オブジェクト消去，(2）検索結果パネルとの紐づけ，(3）オブジェクト複製，(4）文字認識（後述）ボタンとなっている．

3.8.2　手書きの文字による検索

PC やスマートフォンにおける文字入力インタフェースとしては，キーボードが最も一般的である．MR デバイスでは PC やスマートフォン同様に，現実のキーボードや図 10 のような仮想のキーボードを一般的に用いる．しかし，仮想キーボードは物理的なフィードバックが無く，タイプミスが起こりやすいなどの問題点がある．提案手法では，仮想キーボードによる文字入力に加え，手書き文字とキャプチャ機能を使って直接検索ワードを入力することを考える．この機能により，仮想キーボードを使うことなくスムーズに検索ワードを入力することができる．実際に手書きの文字から商品を検索している様子を図 11 に示す．まず，ホワイトボードに書いた手書きの検索ワードを撮影してキャプチャオブジェクトを生成する．次にキャプチャオブジェクトに付随している文字認識ボタンを押すと，検索ボックスに認識した文字が自動で入力され，検索ボタンを押すとそのまま検索することができる．文字認識には Azure AI Vision サービスの中の光学文字認識 API [5]を利用した．

4　実験と評価

4.1　評価実験の概要

提案手法の有用性やユーザビリティを評価するために，学生 8 人に MR デバイスとアプリケーションを使って実際にデジタルカメラを検索してもらった．なお被験者の 8 人は全員，これまでにほとんど MR デバイスに触れたことのない人達であった．比較検証のために，被験者には MR デバイスを使った提案

5：https://learn.microsoft.com/ja-jp/azure/ai-services/computer-vision/overview-ocr

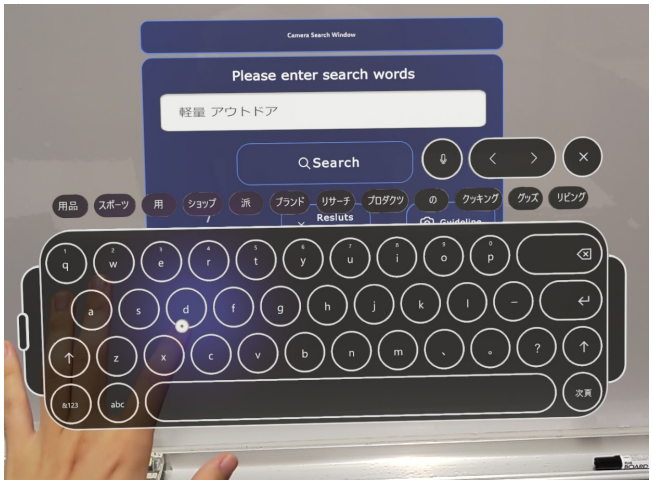


図 10　MRTK3 で提供されている仮想キーボード

図 11　文字認識による検索ワード入力手順．(a) 検索したいワードを手書きする，(b) キャプチャオブジェクトを生成して，一番下の文字認識ボタンを押す，(c) 認識された結果が検索ボックスに入力される

手法による検索，PC による検索，付箋による情報の整理を模した，商品が印刷されたカードを用いた検索の 3 種類の検索を行ってもらった．被験者には 3 種の検索方法全てにおいて，ある仮想の状況を設定し，その状況に合わせたカメラを探してもらうという形のタスクを課した．PC による検索では，被験者は自身の PC を用いて，価格.com のデジタルカメラのページ上でタスクに沿ったデジタルカメラを検索してもらった．カードを用いた検索では，被験者は数十枚あるカードを机上で動かして整理したり，付箋にメモを残したりしながら，商品のカードの中からタスクに沿ったデジタルカメラを一つ選ぶ．付箋を模したカードでの商品検索を行っている様子を図 12 に示す．カードにはデジタルカメラの外観画像，商品名，メーカー名，価格，商品説明文が記載されている．ここで PC，及びカードを利用した検索をまとめて従来手法とする．提案手法と従来手法でそれぞれ最長 30 分のデジタルカメラの検索を終えた後に，ユーザビリティや使用感，有用性に関するアンケートと，口頭で使用感について聴き取るインタビューを行った．提案手法と従来手法それぞれでカメラを検索する上で検索を行う順番の影響，設定した状況による影響を軽減するために，仮想のタスク A，B の 2 つ用意し，手法の順番，タスクを入れ替えた表 3 のような 4 つのグループに被験者を分けて実験を行った．

図 12　カードを用いた検索を行っている様子

4.2　検索タスクと評価項目

用意した仮想の検索タスクは以下の 2 種類である．被験者は 2 つある検索タスクのどちらか一方を従来手法，もう一方を提案手法で検索し，状況に沿ったカメラをそれぞれでひとつずつ決定する．

- タスク 1: 軽量で取り回しがしやすく，アウトドアに持っていけるようなカメラが欲しい．CANON 製で，5 万円以下のものを探したい．
- タスク 2: 旅行に持って行って，観光地や名物などをハイクオリティな写真で残せるカメラが欲しい．SONY 製で，10 万円前後のものを探したい．

使用感や有用性を問うアンケートでは，提案手法と従来手法を比較してどちらが情報の整理をしやすかったか，などの表 4 に示す 4 つの項目に対して，7 段階（1: 同意しない，7: 同意する）で回答してもらった．その際評価は PC と比べて MR が，あるいは，カードと比べて MR が有用であったかを相対的に評価してもらった．ユーザビリティの評価には，System Usability Scale（SUS）[3] のスコアを指標として用いた．

4.3　評 価 結 果

最初に，使用感や有用性についてのアンケートの結果を示す．アンケートのそれぞれの問いに対して 7 段階の評価をした人数の分布を表 5 に示す．MR デバイスを使った空間的な情報整理という観点では，PC や物理的なカードと比較して概ね効果的であるという結果が得られた．また，MR デバイスとホワイトボードを交えて情報を整理するという観点では，高い評価をつけた人がより多いという結果となった．しかし，物理的なカードと比較すると，やはり直感性の面で優位に立つことは難しく，7 段階中 4 付近での評価が多かった．

次に，口頭で使用感を聴き取るインタビューでは，まず提案手法と PC を比較した場合，自由に検索結果を並び替えて比較できることへの評価が高かった．ホワイトボード上に十字に直線を書いて，値段とサイズなどの二軸で検索結果を整理する方法を高く評価する声もあった．また，検索結果を見た結果不要であると判断したカメラを視界から退けるときの操作性がよく，爽快感があったというフィードバックもあった．それに対して，被験者の多くは MR デバイスに慣れていないこともあり，MR

表 3　被験者グループの組み合わせ

	PC で検索した後に MR で検索	MR で検索した後に PC で検索
タスク A を MR, タスク B を PC	タスク B を PC で検索した後に, タスク A を MR で検索	タスク A を MR で検索した後に, タスク B を PC で検索
タスク A を PC, タス B を MR	タスク A を PC で検索した後に, タスク B を MR で検索	タスク B を MR で検索した後に, タスク A を PC で検索

表 4　インタビューの項目

Q1.	PC と比べて カードと比べて	MR の空間的な情報整理は直感的で有用であったと思う
Q2.	PC と比べて カードと比べて	MR とホワイトボードを使った情報整理は直感的で有用であったと思う
Q3.	PC と比べて カードと比べて	MR で行う手書きの注釈は情報の整理において有用であったと思う
Q4.	PC と比べて	手書き文字による検索ワードの入力は有用であったと思う

デバイス独特の操作に戸惑ったという声も多かった．

提案手法と物理的なカードによる検索を比較した場合，MR では検索結果パネルの表示を切り替えて特徴を複数見るという実際のカードではできない点を評価する声が多かった．それに対して，物理的なカードのように実体がある方が直感的に操作できるのでこちらの方がよいという声もあった．手書き文字による検索ワードの入力については，文字認識にかかる時間が長く使いにくかったという声や，手書きしている最中にリアルタイムで文字認識ができるとよかったという声などが多く，課題が残る結果となった．

次に，各手法のユーザビリティスコアの平均を表 6 に示す．米 MeasuringU 社によると，過去 10,000 件以上の回答データの平均スコアは約 68 であり，50 を下回ると一般的にそのシステムは許容され辛いという傾向がある [11]．評価実験では，PC や物理的なカードなどの従来手法のスコアは平均を超えていた．MR デバイスを使った提案手法のスコアはそれらと比較して低かったが，50 以上ではあったため実用には耐えうると考えられる．

4.4 議　　論

提案手法のユーザビリティスコアが低かった原因として，仮想オブジェクトの操作性の影響が挙げられる．仮想オブジェクトは視界にのみ存在し，実際に触ることができない，つまり，物理的なフィードバックを持たない．評価実験を行う過程で，被験者が仮想パネルのボタンや仮想キーボードを操作する際に，パネルまでの距離を測りかねたり，ボタンを押し込む距離がわからず上手く押せなかったりといった状況が見受けられた．これは，仮想のボタンやキーボードが物理的な触覚フィードバックを持たないことが原因である．この問題を解決するためには，振動子などの別のハードウェアを導入して触覚フィードバックを再現する方法や，ホワイトボードやマーカなどの物理オブジェクトとの連携をさらに密にし，仮想ボタンの使用頻度をできるだけ下げるという方法が挙げられる．

ユーザビリティの評価結果に関しては，アプリケーションの影響以外に，MR デバイスのハードウェアの制約の影響が大きいと考えられる．HoloLens2 を始めとした MR デバイスの普及率は高いとは言えず，MR デバイス独特の操作に慣れている人は多くない．プロセッサの処理能力もまだ十分であるとは言い難いが，MR をはじめとする XR デバイスは未だ発展途上であり，技術の発展によって今後デバイスが進歩していけば日常生活でも十分に使えるユーザビリティに達する可能性はあると考えている．また，提案手法の中で検索結果に注釈をつける方法としてホワイトボードを用いていたが，必ずしもホワイトボードである必要はないと考えている．状況によって机の上に検索結果パネルを配置して整理したり，パネルと紙のノートを紐づけて鉛筆で注釈をつけたりと，応用の余地は十分にある．

5　まとめと今後の課題

本研究では，MR デバイスによる仮想オブジェクトとホワイトボードを使って商品検索，結果の整理，買いたい商品の吟味をする商品検索手法を提案した．我々はプロトタイプとして，デジタルカメラの商品データを検索対象とした MR 商品検索アプリケーションを実装した．実際に数人の学生に提案手法を使ってカメラを検索してもらい，ユーザビリティ，有用性を評価した結果，ホワイトボードなどを用いた検索結果の整理に関しては概ね有用であるという回答が得られた．特に，3 次元空間内で商品の情報を真横に並べて比較できる点や，手書きの線を書いて商品を自由に整理できるといった点が評価された．しかし，MR 独特の操作感やハードウェアの制約の影響で，ユーザビリティスコアは従来手法と比較して低いという結果になった．

今回実装した MR アプリケーションではデジタルカメラを検索対象としたが，日常生活での活用を視野に入れた場合，デジタルカメラ以外の様々な商品も検索でき，既存の Web ページなどとも連携できるようにする必要がある．今後は，より日常生活での使用に即した汎用性のある手法の構築を目指すとともに，MR デバイスのインタフェースに十分慣れた上でより長期的に使用してもらう評価実験にも取り組むべきであると考えている．

表 5　表 4 のアンケート項目に対する 7 段階での相対評価の結果（人数）

	MR と PC の比較								MR とカードの比較							
	1	2	3	4	5	6	7	平均	1	2	3	4	5	6	7	平均
Q1.	0	1	0	0	1	5	1	5.5	0	1	0	2	4	1	0	4.5
Q2.	0	0	1	0	1	2	4	6.0	0	0	0	3	2	2	1	5.1
Q3.	0	0	1	1	1	3	2	5.5	0	0	0	5	2	1	0	4.5
Q4.	0	2	1	2	3	0	0	3.8	-	-	-	-	-	-	-	-

表 6　各手法における SUS スコアの平均

検索手法	平均スコア
PC	68.1
物理カード	77.8
MR	53.4

謝　　辞

本研究は JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. Here and Now: Reality-Based Information Retrieval. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, pp. 171–180, 2018.

[2] Mark Billinghurst, Hirokazu Kato, and Ivan Poupyrev. Tangible augmented reality. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, pp. 1–10, 2008.

[3] John Brooke. SUS: A quick and dirty usability scale. *Usability and Evaluation in Industry*, Vol. 189, pp. 189–194, 1996.

[4] Carolina Cruz-Neira, Daniel Sandin, and Thomas DeFanti. Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 135–142, 1993.

[5] Grigoris Daskalogrigorakis, Ann McNamara, and Katerina Mania. Holo-Box: Level-of-Detail Glanceable Interfaces for Augmented Reality. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pp. 1–2, 2021.

[6] Aakar Gupta, Bo Rui Lin, Siyi Ji, Arjav Patel, and Daniel Vogel. Replicate and Reuse: Tangible Interaction Design for Digitally-Augmented Physical Media Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

[7] Hiroshi Ishii and Brygg Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 234–241, 1997.

[8] Leszek Kaliciak, Hans Myrhaug, and Ayse Goker. Content-Based Image Retrieval in Augmented Reality. In *Proceedings of the 8th International Symposium on Ambient Intelligence*, pp. 95–103, 2017.

[9] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. HoloDoc: Enabling Mixed Reality Workspaces That Harness Physical and Digital Content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[10] Weizhou Luo, Anke Lehmann, Hjalmar Widengren, and Raimund Dachselt. Where Should We Put It? Layout and Placement Strategies of Documents in Augmented Reality for Collaborative Sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

[11] MeasuringU. 5 ways to interpret a sus score, 2018. `https://measuringu.com/interpret-sus-score/`.

[12] Paul. Milgram and Fumio Kishino. A Taxonomy of Mixed Reality Visual Display. *IEICE TRANSACTIONS on Information and Systems*, Vol. 77, No. 12, pp. 1321–1329, 1994.

[13] McEwan Mitchell, Blackler Alethea, Johnson Daniel, and Wyeth Peta. Natural mapping and intuitive interaction in videogames. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*, pp. 191–200, 2014.

[14] Munehiro Nakazato and Thomas.S Huang. 3D MARS: immersive virtual reality for content-based image retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo, 2001*, pp. 44–47, 2001.

[15] Ivan Poupyrev, Desney Tan, Mark Billinghurst, Hirokazu Kato, Holger Regenbrecht, and Nobuji Tetsutani. Tiles: A Mixed Reality Authoring Interface. In *Proceedings of the INTERACT 2001 Conference on Human Computer Interaction*, Vol. 2021, pp. 727–728, 2001.

[16] Hirohito Shibata and Kengo Omura. Reconsideration of the Effects of Handwriting: Comparing Cognitive Load of Handwriting and Typing. *ITE Transactions on Media Technology and Applications*, Vol. 6, No. 4, pp. 255–261, 2018.

[17] Xiangyu Wang and Phillip S. Dunston. Tangible mixed reality for remote design review: a study understanding user perception and acceptance. *Visualization in Engineering*, Vol. 1, No. 1, pp. 1–15, 2013.

[18] Austin R. Ward and Rob Capra. Immersive Search: Using Virtual Reality to Examine How a Third Dimension Impacts the Searching Process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1621–1624, 2020.

[19] 上坂晃雅, 木村朝子, 柴田史久, 田村秀行. 複合現実空間との対話操作のための道具型デバイス. 情報処理学会研究報告コンピュータビジョンとイメージメディア, Vol. 107, No. 427, pp. 81–86, 2008.

[20] 柴田史久. 複合現実感技術の歴史と今後の展望. システム／制御／情報, Vol. 64, No. 9, pp. 343–348, 2020.

[21] 田村秀行, 大田友一. 複合現実感. 映像情報メディア学会誌, Vol. 52, No. 3, pp. 266–272, 1998.

[22] 岩崎洋平, 岩井智成. 図書館利用者のための MR サービスシステムの開発. 情報処理学会研究報告. グラフィクスと CAD 研究会報告, Vol. 2013, No. 1, pp. 1–8, 2013.

物理的距離によるLoD制御をともなうVR環境における不動産情報探索

中山　裕紀† 　津田　裕哉†† 　荘司　慶行††† 　大島　裕明†,††

† 兵庫県立大学 社会情報科学部　〒651–2197 兵庫県神戸市西区学園西町 8 丁目 2–1
†† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8 丁目 2–1
††† 静岡大学 情報学部　〒430–0807 静岡県浜松市中区城北 3–5–1
E-mail: †{fa20m061,ad22o041}@guh.u-hyogo.ac.jp, ††shojiy@inf.shizuoka.ac.jp, †††ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，不動産情報の比較を行いやすくするために，VR 空間上に不動産情報を表示する手法を提案する．不動産情報は賃料や間取りなど複数の属性を持ち，不動産情報探索においては複数の属性を比較することが重要になる．そこで，本研究では，不動産情報が持つ複数の属性を比較することができるような VR 空間での不動産情報探索手法を提案する．具体的には，VR 空間上に配置された不動産情報との距離により，表示する情報量を変更する．配置された不動産情報から離れているときは外観写真のみを表示し，歩み寄るなどして物理的に近づいたときには詳細な情報を表示するようにした．このようにして，俯瞰する全体的な情報の比較と，近づく詳細な情報の比較を切り替えることで，不動産情報の比較を行いやすくした．このような情報との物理的距離によって変わる不動産情報の詳細度を Level of Detail（LoD）と呼ぶ．また，提案手法が属性の比較を行いやすくするのか評価実験を行った．その結果，LoD を距離により制御する手法が，比較のための操作による負担を軽減することを示した．

キーワード　XR（VR/AR/MR），アクセシビリティ，インタフェース・インタラクション

1　はじめに

現代社会において，不動産物件を探す際，不動産情報検索システムを利用する機会が多くなっている．不動産情報は，賃料や最寄り駅との距離，間取りなど多くの属性を持っている．例えば，似たような賃料の不動産物件の中でも，駅に近く，間取りが狭いという物件もあれば，駅からは遠いが新築で，インターネットなどの設備が整っているという物件もある．そのため，複数の属性について比較を行うことが，ユーザのニーズに合う不動産情報を見つけるために重要となる．しかし，既存の不動産情報検索システムでは，ディスプレイ上に表示できる不動産情報の数が少なく，複数の属性を比較することが難しいという問題がある．これは，平面ディスプレイに表示できる情報量の制約や，不動産情報を縦に並べるリスト表記の方法が原因として考えられる．そこで，本研究では VR 空間上で不動産情報探索を行う手法を提案する．

VR は仮想現実とも訳され，自分がまるでコンピュータが作り出す仮想空間にいるかのような没入感を得ることができる技術である．VR では頭部に装着するヘッドマウントディスプレイ（HMD）を用いる方法が一般的に用いられる．また，頭や手など身体の部位の動きを感知するトラッキングにより，現実空間での動きを対応させ仮想空間でも同じ動きをすることができる．トラッキングでは HMD に搭載されたセンサーやカメラにより頭や手の動きを感知することができるほか，腕や腰，足などに取り付けたセンサーにより全身の動きを感知することもできる．本研究ではトラッキングを利用し体の動きと連動したインタラクションを用いた情報探索を行うため，VR 技術を使用する．

図 1　不動産情報を配置した様子

本研究では，VR 空間上に表示されている不動産情報に歩み寄るなどの物理的に近づく動作をすることで，不動産情報の詳細度を変更する．表示されている情報との距離が遠い場合，外観写真のみの不動産情報を表示する．逆に表示されている情報との距離が近い場合，外観写真に属性を加えた詳細な不動産情報を表示する．このような不動産情報の詳細度のことを本研究では Level of Detail（LoD）と呼んでいる．

本研究では，VR 空間上で，表示されている情報との物理的距離により，LoD を変更しながら不動産情報探索を行う手法を提案する．VR デバイスを用いた体の動きと連動する操作により，不動産情報の LoD を変更しながら不動産情報を比較し，絞り込むことができるようにする．

提案手法では，VR デバイスである Meta Quest3 を用いる．ユーザが属性を選択すると，その属性をもとに不動産情報をい

くつかのカテゴリに分け，このカテゴリごとに VR 空間上に不動産情報を配置する．不動産情報を配置した様子を図 1 に示す．配置された情報から遠ざかると LoD が低くなり，不動産情報を全体的かつ大まかに比較することができる．また，不動産情報に近づくと LoD が高くなり，不動産情報を詳細に比較することができる．このようにすることで，多量の情報から絞り込みを行う過程で不動産情報を取得できるようになり，ユーザのニーズに合う不動産情報を選ぶことができる．

本論文では，次の 2 節で情報探索での VR 技術や LoD の活用など，先行研究について述べる．3 節で VR 技術を用いて不動産情報探索を行う状況の説明と，本研究で用いるデータセットの詳細やデータの加工，実装に用いるソフトウェアについて述べる．4 節では提案する手法の詳細を述べ，5 節で手法を用いた実験とその結果，それを踏まえた考察を述べる．最後に 6 節でまとめと今後の課題を述べる．

2 関 連 研 究

2.1 VR 空間での情報探索

VR 空間上で情報探索を行う手法は広く研究されている．竹内ら [1] は情報検索を行った平面のディスプレイを VR 空間上に配置して整理することで，検索で得た情報が記憶に定着することを支援する「空間的ブックマーク」という手法を提案している．実験では，空間的ブックマークを用いて，四方を白色の壁で囲まれた VR 空間，4 色に色分けされた VR 空間，家具などが配置された部屋を模した VR 空間の 3 通りの空間に検索結果を配置させた．また，既存の手法として検索結果を水平に並べるだけの VR 空間で情報検索を行い，記憶への定着を比較した．その結果,「空間的ブックマーク」を用いるすべての手法で，比較手法より検索結果が記憶に定着しており，部屋を模した VR 空間で最も記憶に定着していたことが示された．

Austin ら [2], [3] は VR 空間上での検索における検索結果の表示方法をリスト表記，グリッド表記，円弧状表記の 3 通りについて比較し，目標結果を見つけるには円弧状表記が速い一方，ユーザがリスト表記，グリッド表記を好むことを示している．また，VR 空間上への情報検索結果の表示とユーザのインタラクションを調査するための Information Interactions in Virtual Reality System を提案している．

Leszek ら [4] は，海洋環境のモニタリングにおける認識と理解を促す画像検索フレームワークを提案している．このフレームワークでは，VR ヘッドセットを用いて，360 度カメラによる映像をリアルタイムでモニタリングし，その映像に情報を補完することで AR のように情報を提示する．

Maurice [5] は，VR システムを用いた情報検索の手法について合計 1,042 本の文献を調査し，VR ハードウェアの進展が情報検索に新しい環境をもたらすという仮説を検証している．その結果，ユーザ評価により従来の対応物と比較したシステムの多くが，タスクにかかる時間を短縮したか同等であったと報告している．また，ユーザパフォーマンスを測定した論文のうち半数近くで同等以上のパフォーマンスであったと報告している．特に，いずれもユーザは VR システムに満足していたということが報告されている．

Wolfgang ら [6] は，情報検索プロセスを AR 技術と組み合わせた「リアリティベースの情報検索」という概念を提案している．リアリティベースの情報検索では，状況による検索のきっかけと，検索クエリと結果の位置による視覚化を提供している．

Juraj ら [7] は，ソフトウェアプロジェクトのソースコードリポジトリを VR 空間内で検索，探索することができるようなアプローチを提案している．この提案手法では，抽象的な情報空間を，インタラクティブな 3D オブジェクトにより視覚的に表現し，オブジェクトを直接操作してソースコードの検索，探索を行う．

Daniele ら [8], [9] は，VR 空間内で 3D モデルのコレクションを検索するための，スケッチを使用する方法を提案している．また，このスケッチのインタラクションが VR 環境にどのように変換されるのか 3D 空間上でのスケッチ，仮想タブレット上でのスケッチ，固定された仮想タブレット上でのスケッチ，現実空間のタブレットでのスケッチという 4 通りの方法で比較し，検索のための直感的な方法は 3D 空間上でのスケッチであると示している．

Hugo ら [10] は，3D オブジェクトの検索ユーザインタフェースにおけるコンテキストが，どのような役割で構想されているのかをユーザスタディにより示している．

本研究では，広い VR 空間上に不動産情報を配置し，現実空間での移動によるインタラクションを取り入れた不動産情報探索手法を提案する．

2.2 不動産情報探索における属性

不動産情報探索において属性は不動産情報を比較するための重要な指標となる．そのため，属性を比較しやすくするための研究が行われてきた．湯本 [11], [12] は，不動産情報が持つトレードオフの関係に着目し,「オートロックがついていれば少し賃料が高くても構わない」というようなあいまいなニーズを反映することが可能な不動産情報推薦システムを提案している．

諏訪ら [13], [14] は，騒音や日当たりなど定量的に評価されない不動産情報に着目し，それらを IoT デバイスや機械学習を用いて指標化して定量的に評価できるようにしている．

本研究では不動産情報が持つ複数の属性を比較しやすくするために，VR 空間上で不動産情報探索を行う手法を提案する．

2.3 情報検索と LoD 制御

様々なシステムに LoD 制御を取り入れる研究には次のようなものがある．Grigoris ら [15] は，情報検索を行うための AR ユーザインタフェース Glacealble Interfaces において，この UI の LoD を操作できるような ARUI デザインである「Holo-Box」を提案している．

大柳ら [16] は複数の異なるデータからなる大規模データセットの検索に対して，センサーデータの 3 つの属性からなる索引データを作成し，これに LoD 制御を取り入れることで，効率よく検索できる手法を提案している．

武田ら [17] は，読書体験において LoD 制御を取り入れ，カメラとの距離によって，文章を簡易的なものから詳細なものまで切り替えることで，新たな読書体験を提供するシステム「Text Level Of Detail」を提案している．

本研究では情報の表示を切り替えるための操作に距離による LoD 制御を取り入れ，VR 空間内での不動産情報の比較を行いやすくする手法を提案する．

3 不動産情報探索における課題と用いるデータセット

3.1 想定される状況

本研究では，ユーザは物件探しをしているという状況を想定する．不動産情報は賃料や間取りなど多くの属性を持ち，それら属性が互いに複雑に影響しあっているため，複数の属性について比較を行うことが重要である．しかし，従来の不動産情報検索システムで属性を比較するには，賃料など 1 つの属性を選択してそれについて並び替えるという方法や，最寄り駅や間取りなどの属性をもとに不動産情報を絞り込むという方法が多く，このようなシステムでは複数の属性を比較することが難しいと考える．この問題を解決するために，不動産情報を得ながら属性を選択し，絞り込みを行うことが必要である．

3.2 用いるデータセット

本研究では，国立情報学研究所の IDR データセット提供サービスにより，株式会社 LIFULL から提供を受けた「LIFULL HOME ' S データセット [1]」の不動産情報データおよび画像データを使用する．このデータセットには総数 5,334,370 件の不動産情報について，属性やその物件の画像データが含まれている．属性は，マンション，戸建て，などの「物件種別」，市区町村までの「住所」,「最寄り駅」,「最寄り駅からの距離」,「建物構造」,「建物面積」,「築年月」,「部屋の向き」,「間取り」,「近隣施設」など合わせて 71 個ある．本研究では，これらのうち「賃料」,「間取り」,「最寄り駅」,「最寄り駅からの距離」,「築年月」を主に使用する．また，画像データは 1 件の不動産情報に複数枚含まれることが多く，20 枚以上の画像が含まれる物件もある．画像データは「間取り図」,「外観写真」,「地図」,「周辺画像」,「内装」,「風呂」,「玄関」,「居間」,「キッチン」,「寝室」,「子供部屋」,「その他」の 12 種類の画像種別に分けられる．本研究ではこのうち「外観写真」と「間取り図」を用いた．データセット内の画像データの総数は 83,403,687 件である．

3.3 データの前処理

このデータセットの中から，神戸市に存在する物件の物件情報と，その物件の外観写真と間取り図を抽出した．その結果，84,222 件の不動産情報と，185,364 件の画像データが得られた．データの総数と抽出後のデータ数について表 1 に示す．画像データの長辺は 128 ピクセルと決まっているが，短辺のピクセル数が決まっておらず，正方形と長方形両方の画像がある．そのため本研究では，全ての画像を正方形に統一するために，128 × 128 ピクセルの白い画像を用意し，その画像の中央に外観写真を配置する加工を行った．加工後の画像データの例を図 2 に示す．

図 2　加工後の画像の例

表 1　データセットに含まれるデータの総数と実際に使用したデータ数

データの種類	データ数
不動産情報の総数	5,334,370
画像データの総数	83,403,687
神戸市の不動産情報	84,222
神戸市の抽出後の画像データ	185,364

図 3　不動産情報を表示するパネル

3.4 実装に用いるソフトウェア

VR デバイスは Meta 社が販売している「Meta Quest 3 [2]」を使用する．Meta Quest 3 は眼前を覆うヘッドマウントディスプレイ（HMD）であり VR 空間への高い没入感を実現する．本アプリケーションはゲームエンジンである「Unity[3]」を用いて構築した．Unity における VR 開発の環境構築やアプリケーション内の機能は，Unity が提供する「Oculus Integration」を用いた．開発に使用したソフトウェアのバージョンは，Unity が 2021.3.16f1，Oculus Integration が 57.0.1 であった．

1：株式会社 LIFULL (2015): LIFULL HOME'S 賃貸物件スナップショットデータ．国立情報学研究所情報学研究データリポジトリ．(データセット). https://doi.org/10.32130/idr.6.1/

2：https://www.meta.com/

3：https://unity.com/

図 4　属性を選択する様子

図 5　不動産情報の配置

4　不動産情報探索のための VR インタフェース

4.1　システムの概要

本システムでは不動産情報を VR 空間上に配置する．不動産情報は図 3 に示す正方形のパネルに表示する．このパネルを不動産情報パネルと呼ぶ．不動産情報を 1 つの属性をもとにカテゴリ分けし，カテゴリごとにまとめて配置する．次に，ユーザはカテゴリを選択して，表示される不動産情報を選択したカテゴリに含まれるものに変更する．このようにして不動産情報の絞り込みを行う．表示されている不動産情報の詳細度をユーザとの距離によって変更する．この 3 つの機能をもつシステムを実装した．

4.2　不動産情報のカテゴライズと配置

ユーザは，図 4 のようにボタンを押して属性を選択する．選択できる属性は，賃料，間取り，最寄り駅，最寄り駅からの距離，築年数とする．選択した属性をもとに，不動産情報のカテゴライズを行い，カテゴリごとに縦にまとめて配置する．カテゴリとは，例えば属性が賃料の場合，0 円から 2 万円までのカテゴリ，2 万円から 4 万円までのカテゴリ，などのように，属性の値によって不動産情報を分類したものである．カテゴリは最大で 8 つまでとする．不動産情報を配置した様子を図 5 に示す．各カテゴリに含まれる不動産情報数の割合を，縦に配置される不動産情報パネルの数の割合で表す．不動産情報パネルの数を見て，カテゴリに含まれる不動産情報の数の違いを把握することができる．配置される不動産情報パネルの数 n は，各カテゴリに含まれる不動産情報数を m，全ての不動産情報数を o としたとき，次のような式で決定する．

$$n = \frac{m}{o} \times 20$$

このとき，数値は切り上げる．また，不動産情報パネルの数は 1 つのカテゴリにつき最大で 5 つとなるようにする．例えば，

図 6　カテゴリの情報を表すボタン

図 7　カテゴリを選択している様子

図 8　詳細な不動産情報の表示

表 2　評価グループの分け方

	先に VR，後に PC で探索する	先に PC，後に VR で探索する
先に検索条件 1， 後に検索条件 2 で探索する	先に VR で検索条件 1， 後に PC で検索条件 2 の不動産情報を探索する	先に PC で検索条件 1， 後に VR で検索条件 2 の不動産情報を探索する
先に検索条件 2， 後に検索条件 1 で探索する	先に VR で検索条件 2， 後に PC で検索条件 1 の不動産情報を探索する	先に PC で検索条件 2， 後に VR で検索条件 1 の不動産情報を探索する

表 3　検索条件の一覧

	検索条件 1	検索条件 2
想定状況	一人暮らしの大学生	両親と子の三人家族
賃料	6 万円から 8 万円	8 万円から 10 万円
間取り	ワンルーム，1K，1DK	2DK，2LDK
最寄り駅	新長田駅，板宿駅	新神戸駅，三宮駅，県庁前駅
最寄駅からの距離	徒歩 10 分以内	徒歩 15 分以内

カテゴリに含まれる不動産情報の数が 3,704 件，すべての不動産情報の数が 79,261 件であるとき，計算結果は 0.93 となり，不動産情報パネルは 1 つ配置される．この場合，配置される不動産情報パネルには 3,704 件の不動産情報が集約されており，表示されるのはそれらのうちの 1 件である．

ユーザが賃料を選択したとき，賃料が 0 円から 2 万円のカテゴリ，2 万円から 4 万円のカテゴリ，などとなるよう 2 万円ごとにカテゴライズする．賃料が 14 万円以上の不動産情報は 1 つのカテゴリとする．賃料が小さいカテゴリを左から順番に配置し，各カテゴリでは，賃料が小さい物件を下から順番に配置する．

最寄り駅については，神戸市営地下鉄の谷上駅から西神中央駅までの 17 駅を対象とする．カテゴリは，3 駅を含むカテゴリが 1 つと，2 駅を含むカテゴリが 7 つの，計 8 つとなるようにした．3 駅を含むカテゴリは，最寄り駅が谷上駅，新神戸駅，三宮駅の 3 駅であるものとし，2 駅を含むカテゴリは，例えば最寄り駅が県庁前駅，大倉山駅の 2 駅となるようにした．最寄り駅がこの 17 駅のいずれかである不動産情報を，その駅のカテゴリに分類する．

間取りは，リビング，ダイニング，キッチンなどの部屋種別と，それ以外の部屋の数を組み合わせて表される．例えば，部屋種別がリビング，ダイニング，キッチンを兼ねた部屋と，その他に寝室などが 2 部屋ある場合，間取りは 2LDK と表される．提案手法では，部屋種別がワンルームのカテゴリ，1K などと表されるキッチンのみのカテゴリ，1DK などと表されるダイニングキッチンのカテゴリ，などとなるよう部屋種別によりカテゴリを分類する．

築年数については，0 年から 5 年までのカテゴリ，5 年から 10 年までのカテゴリ，などとなるよう 5 年ごとにカテゴリ分けを行う．ただし，カテゴリ数が 8 となるよう，35 年以上の不動産情報は 1 つのカテゴリに分類する．

4.3　検索結果の絞り込み

ユーザはカテゴリを選択することで不動産情報の絞り込みを行うことができる．各カテゴリには図 6 に示すカテゴリの情報を表すボタンが存在する．このボタンを，コントローラーから発射されるレーザーポインタを用いて押すことで，カテゴリの選択を行う．カテゴリの選択を行っている様子を図 7 に示す．選択したカテゴリに含まれる不動産情報のみを表示することや，選択したカテゴリを消去することができる．以上のようにして，多量に存在する不動産情報の中から表示されるものを減らし，ユーザのニーズに合う不動産情報が残るようにする．

絞り込みを行うと，属性をカテゴライズする方法が変更される．賃料は絞り込みを行うごとに，カテゴライズを行う値段の幅を 2 万円から 5 千円へと小さくする．最寄り駅は，2 駅ごとにカテゴライズしていたものを，1 駅ごとにカテゴライズする．間取りは，部屋種別によりカテゴライズしていたものを部屋数によりカテゴライズするようにする．築年数は 5 年ごとにカテゴライズしていたものを，1 年ごとにカテゴライズする．

4.4　LoD 制御

カテゴリが表示されているブロックに物理的に近づくと，ブロックに詳細な不動産情報が表示されるようになる．表示する情報は，外観の画像，間取り図，賃料，最寄り駅と駅からの距離，築年数とする．詳細な不動産情報を表示した様子を図 8 に示す．これにより詳細な不動産情報を比較しながら大まかな不動産情報を取得することができ，比較する情報量を増やすことができる．

5　実　　験

5.1　実験の概要

提案手法の有用性を評価するために，筆者と同じ研究室に所属する学生 4 名を対象として実験を行った．対象となる被験者 4 名は全員 VR 機器を利用したことがほとんどなかった．実験では予算や最寄り駅など物件を探す条件を指定し，その条件に合う物件を 1 件見つけるタスクを課した．被検者には VR 空間での不動産情報探索と PC を用いる不動産情報検索サイトでの不動産情報探索のそれぞれについてこのタスクを行ってもらった．ここでは，不動産情報検索サイトを用いる不動産情報探索を従来手法とする．その後，ユーザビリティや使用感，比較のしやすさなどを問うアンケートに回答してもらった．

5.2　実験の流れ

提案手法での検索と従来手法での検索を行う順番による影響と，検索条件による影響を避けるために，2 つの検索条件を設定し，表 2 に示す 4 つのグループに被検者を分けた．ここで，検索条件を表 3 に示す．VR 空間での不動産情報探索では，賃料か最寄駅からの距離をもとにした情報の配置，配置された情報の中からの絞り込み，配置された情報との距離による情報の

表 4　ユーザビリティを問う質問と平均スコア（1：そう思わない，5：そう思う）

質問番号	質問項目	平均スコア
質問 1	このシステムを頻繁に使用したいと思う	2.00
質問 2	このシステムは不必要に複雑だと思う	2.75
質問 3	このシステムは使いやすいと思う	2.25
質問 4	このシステムを使うには技術的なサポートが必要だと思う	2.50
質問 5	このシステムの機能はまとまっていると思う	2.00
質問 6	このシステムにはまとまっていない部分が多いと思う	3.75
質問 7	ほとんどの人がこのシステムをすぐに習得できると思う	3.25
質問 8	このシステムは使いにくいと思う	3.25
質問 9	このシステムは信頼して使えると思う	2.75
質問 10	このシステムを使う前に学ばないといけないことが多かったと思う	2.50

表 5　提案手法と従来手法を比較する質問と平均スコア（1：そう思わない，5：そう思う）

質問番号	質問項目	VR	PC
質問 1	表示された情報量は適切であったか	3.00	2.75
質問 2	物件を決定するまでにかかった時間は短かったか	2.50	2.75
質問 3	絞り込みや情報の詳細を表示するための操作の回数は少なかったか	4.25	2.50

表 6　提案手法を評価する質問と平均スコア（1：そう思わない，5：そう思う）

質問番号	質問項目	平均スコア
質問 1	物件の外観写真のみを俯瞰的に表示することは不動産情報探索に適切であったか	2.0
質問 2	詳細な情報を表示するために情報に近づく操作は情報探索に適切であったか	4.0
質問 3	情報を絞り込むために VR 空間上のボタンを押す操作は情報探索に適切であったか	3.5

詳細度の操作を行い，不動産情報探索探索を行ってもらった．また従来手法では，LIFULL HOME'S [4]の web サイトを用いて不動産情報探索を行ってもらった．

次にユーザビリティや使用感を問うアンケートを行った．ユーザビリティを問うアンケートでは，System Usability Scale（SUS）[18] という評価指標を用いた．SUS は 10 個の質問について，1 がそう思わない，5 がそう思う，という 5 段階で回答し，それをもとに 0 から 100 までのスコアを計算する指標である．SUS は奇数番号の質問が肯定的な質問であり，スコアが高いほど良い評価である．逆に，偶数番号の質問は否定的な質問であり，スコアが低いほど良い評価である．SUS は 80 を超えると大変優れたシステムとされ，50 を下回るとユーザビリティに修正が必要だと判断される．500 回の研究での SUS の平均は 68 となった [5]．使用感を問うアンケートでは，質問を従来手法と提案手法を比較するものと，提案手法を評価するものに分けた．どちらも質問に 1 がそう思わない，5 がそう思うという 5 段階で回答してもらい，その理由を記述してもらった．従来手法と提案手法を比較する質問は，PC を用いた場合と VR 機器を用いた場合のそれぞれについて質問に回答してもらった．ユーザビリティを評価する質問の項目と平均スコアを表 4 に，提案手法と従来手法を比較する質問の項目とそれぞれの手法の平均スコアを表 5 に，提案手法を評価する質問の項目と平均スコアを表 6 に示す．

4：https://www.homes.co.jp/
5：https://measuringu.com/sus/

5.3　実験の結果

まず，ユーザビリティを問う質問で得られた評価について述べる．本研究で提案した VR 空間における LoD 制御を用いた不動産情報探索システムの SUS の平均は 31.2 となりユーザビリティに関して課題が残る結果となった．回答を分析したところ，質問 1 の平均スコアが 2.00，質問 3 の平均スコアが 2.25 とシステムの使いやすさに関する質問の評価が低くなった．また，質問 5 の平均スコアが 2.00，質問 6 の平均スコアが 3.75 と機能のまとまりに関する質問の評価が低くなった．これは，実験で用いたシステムに未完成な点があったことが原因として考えられる．

次に，従来手法と提案手法を比較する質問で得られた評価について述べる．質問 1 の平均スコアは，提案手法が 3.00 であり，従来手法が 2.75 であった．回答理由として，提案手法については表示される画像の種類が少ないという意見や，文字が多く，どこにどの不動産情報が配置されているのかが分かりづらかったという意見が得られた．また，従来手法については必要な情報があったという意見や，情報が適切な量で表示されていたという意見が得られた一方で，絞り込みを行っても情報量が多かったという意見や，複数の不動産情報を比較することが困難であったという意見が得られた．質問 2 の平均スコアは，提案手法が 2.50 であり，従来手法が 2.75 であった．回答理由として，提案手法については比較に時間がかかったという意見や，不動産情報が配置された場所が分からず，神経衰弱をするときのような記憶力を求められる感覚があったという意見が得られた．また，従来手法については情報が多く時間がかかったという意見や，複数の属性を比較するのが難しく時間がかかったと

いう意見が得られた．これは，離れているときと近づいたときとの間で表示される情報量に大きな差があったためだと考えられる．質問 3 の平均スコアは，提案手法が 4.25 であり，従来手法が 2.50 であった．回答理由として，提案手法については近づくだけで情報の詳細が表示できたという意見が得られた．また，従来手法については詳細な情報を見るためにページを遷移する必要があり，クリック数が増えたという意見が得られた．これは，物理的距離により LoD 制御を行う手法により，従来手法で必要なクリックなどの操作を行う必要がなかったためであると考えられる．

最後に，本研究の提案手法を評価する質問で得られた評価について述べる．質問 1 の平均スコアは 2.0 であった．また，回答理由として，外観はそこまで重要ではない，という意見や似たような写真が多いため外観写真だけでは違いが分からず，どこにどの不動産情報が配置されているのか分からなかったという意見が得られた．これは，外観写真のみでは情報量が少なく，不動産情報の比較を行うことができなかったためであると考える．質問 2 の平均スコアは 4.0 であった．また，回答理由として，情報に近づくだけの操作は，直感的で分かりやすいという意見や，負担を感じなかったという意見が得られた．これは，提案手法により外観写真と詳細な情報の切り替えをシームレスに行うことができたためだと考えられる．質問 3 の平均スコアは 3.5 であった．また，回答理由として，ボタンを押す回数が少なくて良かったという意見や，操作が簡単であったという意見が得られた．一方で，近づいて詳細な情報を表示する操作と比較して，絞り込みだけはボタンを使う，ということが直感的ではないという意見が得られた．

5.4 考　　察

実験の結果から，情報との物理的距離によって LoD 制御を行う手法は，操作を直感的で分かりやすいものにした．その結果，操作の回数を減らし，不動産情報探索にかかる負担を軽減することが示された．一方で，ユーザビリティや表示する情報量に課題があることが示された．ユーザビリティについては，選択できる属性や絞り込みの機能を向上させることで，不動産情報探索に必要な機能を充実させていく必要がある．また，表示する情報量については，離れているときでも属性が分かるような表示を行う必要がある．

6 まとめと今後の課題

本論文では，VR 空間上に配置された情報を，LoD 制御により変更しながら行う不動産情報探索手法の提案と実験を行った．提案手法では，1 つの属性をもとに不動産情報を VR 空間に配置する．そして，VR 空間上に配置された情報から離れていると外観写真のみの大まかな不動産情報を，近づくと家賃や間取りなどを加えた詳細な不動産情報を表示する．このようにして，全体的な不動産情報と詳細な不動産情報を取得しながら絞りこむことで，ユーザのニーズにあった不動産情報を残すことができる．実験では，提案手法と PC での従来手法を用いて不動産情報探索を行ってもらい，有用性を確かめた．その結果，情報との物理的距離により LoD 制御を行う手法は VR 空間上で行う不動産情報探索において，操作による負担を軽減することが示された．今後は，不動産情報探索に必要な機能を充実させるために，VR 空間内で地図上を移動することによる地域の選択などの機能を追加する．

謝　　辞

本研究は，JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けたものです．また，本研究では，国立情報学研究所の IDR データセット提供サービスにより株式会社 LIFULL から提供を受けた「LIFULL HOME'S データセット」を利用しました．ここに記して謝意を表します．

文　　献

[1] 武内悠美穂, 荘司慶行, Martin J. Dürst. 空間的ブックマーク：VR 空間内でウェブ検索結果を整理させることによる記憶定着支援. 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM2021), 2021.

[2] Austin R. Ward and Rob Capra. Immersive Search: Using Virtual Reality to Examine How a Third Dimension Impacts the Searching Process. In *Proceedings of the 2020 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1621–1624, 2020.

[3] Austin R. Ward, Yiyin Gu, Sandeep Avula, and Praneeth Chakravarthula. Interacting with Information in Immersive Virtual Environments. In *Proceedings of the 2021 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2600–2604, 2021.

[4] Leszek Kaliciak, Hans Myrhaug, and Ayse Goker. Content-Based Image Retrieval in Augmented Reality. In *Proceedings of the 2017 International Symposium on Ambient Intelligence*, pp. 95–103, 2017.

[5] Maurice Schleußinger. Information Retrieval Interfaces in Virtual Reality-a Scoping Review Focused on current generation technology. *PLOS ONE*, pp. 1–24, 2021.

[6] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. Here and Now: Reality-Based Information Retrieval: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, pp. 171–180, 2018.

[7] Juraj Vincur, Ivan Polasek, and Pavol Navrat. Searching and Exploring Software Repositories in Virtual Reality. In *Proceedings of the 2017 ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2, 2017.

[8] Daniele Giunchi, Stuart James, and Anthony Steed. 3D Sketching for Interactive Model Retrieval in Virtual Reality. In *Proceedings of the 2018 Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*, pp. 1–12, 2018.

[9] Daniele Giunchi, Stuart James, Donald Degraen, and Anthony Steed. Mixing Realities for Sketch Retrieval in Virtual Reality. In *Proceedings of the 2019 International Conference on Virtual-Reality Continuum and Its Applications in Industry*, pp. 1–2, 2019.

[10] Hugo C. Huurdeman and Chiara Piccoli. "More than just a Picture" - The Importance of Context in Search User Interfaces for Three-Dimensional Content. In *Proceedings of the 2020 Conference on Human Information Interaction and*

Retrieval, pp. 338–342, 2020.

[11] 湯本真樹. 条件検索とラフ集合の縮約ルールを利用した学生向け賃貸物件推薦システムの開発. 電気学会論文誌C（電子・情報・システム部門誌）, pp. 740–748, 2013.

[12] Masaki Yumoto. Development of Recommendation System in Rough Set for a Leasehold House for Students Based on Two Kinds of Target User's preference. *IEEJ Transactions on Electronics, Information and Systems*, pp. 977–985, 2016.

[13] 諏訪博彦, 大坪淳, 中村優吾, 野口真史. 物件情報による賃貸物件快適度指標推定にむけた検討. マルチメディア, 分散協調とモバイルシンポジウム 2061 論文集, pp. 278–283, 2020.

[14] 諏訪博彦, 大坪淳, 中村優吾, 野口真史. IoT センシングによる賃貸物件快適度指標の構築と評価. 人工知能学会全国大会論文集, 2020.

[15] Grigoris Daskalogrigorakis, Ann McNamara, and Katerina Mania. Holo-Box: Level-of-Detail Glanceable Interfaces for Augmented Reality. In *Proceedings of the 2021 Special Interest Groupe on Computer Graphics*, pp. 1–2, 2021.

[16] 大柳晋也, 是津耕司. STT Sketch Index の LoD 制御による参加型センシングデータの効率的な検索手法の提案. マルチメディア、分散協調とモバイルシンポジウム 2014 論文集, pp. 733–737, 2014.

[17] 武田港, 望月茂徳, 川村健一郎. ズーム動作と有限多層階層による、3D 文章の表示・作成システム「Text Level Of Detail」の提案. 情報処理学会第 74 回全国大会講演論文集, pp. 303–304, 2012.

[18] John Brook. SUS - A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, pp. 189–194, 1996.

空港手荷物受取所のUX向上を目指した拡張現実型情報提示サービス

池田　呼人†　　松原　克弥†

† 公立はこだて未来大学　〒 041-8655 北海道函館市亀田中野町 116-2
E-mail: †{b1020254,matsu}@fun.ac.jp

あらまし　近年，空港業務の課題の一つに手荷物預け率の低下がある．手荷物を預けずに機内へ持ち込む旅客の増加は，狭小な機内での大きな荷物の取り回し，収納を困難にして旅客の円滑な乗降の妨げとなり，遅延増加の要因となっている．手荷物預け率低下の要因として，現在の手荷物受取所は荷物がいつ来るか分からない中，旅客にレーンの前で待たせる事で，不快感を与えている点がある．本研究では手荷物受取所での待機時間に利用できるアプリケーションを提供し，待機時間の有効利用を促す事でストレスの解消と手荷物を預ける事に対する忌避感の減少を目指す．本研究では手荷物受取所内の荷物レーン上に，AR を用いて食べ物などのコンテンツを流す，回転寿司のようなスタイルで観光情報を提供するサービスを提案する．提案サービスを函館空港手荷物受取所で実際に利用してもらう事で，サービスを利用する事で手荷物受取所のユーザー体験向上につながるか検証を行った．

キーワード　AR，観光，UX

1　はじめに

現在，航空機は鉄道やバスと並んで代表的な公共交通機関の一つとなっている．他の公共交通機関と同様に，航空便においても遅延は課題である．遅延の要因として，天候，機材の不調，機材繰り [1] があるが，近年新たに浮上しているのが乗降時間の増大である．乗降時間増大の要因として手荷物を預けず機内に持ち込む旅客の増加がある．機内へ持ち込む荷物がスーツケースなどの大きな荷物の場合，狭小な機内通路を閉塞したり，荷物棚へ荷物を収納するのに時間を要すなど，旅客の円滑な乗降の妨げとなって乗降時間を増大させ，航空便の遅延の要因となる．

手荷物を預けない旅客が増加した理由に，手荷物受け取りに関するストレスがある．手荷物受取所では自分の手荷物が流れてくるまで待機する必要があるが，旅客に対していつ荷物が流れてくるのか提示されることは稀で，大半の場合は旅客に対して荷物レーンの前で立ったまま待つ事を強いている．状況によって荷物が流れてくるまでの時間は変化するが，ベンチなど旅客が快適に待つ事が出来る設備が十分に整備されていない中で，旅客を待機させるこの時間が，旅客に大きなストレスを生じさせる要因となり，手荷物を預ける事に対する忌避感に繋がっている．手荷物を受け取る際の一連のプロセスに起因するストレスによって生じる問題は，空港側に問題として認知されつつある．実際に，ヒューストン空港 [5] や南紀白浜空港 [6] を始め，手荷物受け取りの際に生じるストレスを軽減するため，いくつかの取り組みが各地の空港で行われている．

本研究は，これらの事例に倣い，航空機の空港到着から旅客が自身の手荷物を受け取るまでの一連のプロセスのユーザ体験を向上させ，ストレスの軽減と旅客の手荷物を預ける事に対する忌避感の減少を目指す．手法として，空港施設内で使用可能なスマートフォン向けモバイルアプリケーションを採用する．サービス展開の対象は，ストレスが生じている空間である手荷物受取所内とする．Augmented Reality（以下，AR）技術を使用した情報提供を主軸としたサービスを，手荷物を受け取る際の待機時間中に利用して貰う事で，待機時間中のユーザー体験を改善し，ストレスの解消を行う．本研究は，函館空港で実験やサービス開発を実施する．

2　関連研究・事例

手荷物受け取りに関するストレスは，主に空港到着後に旅客が手荷物受取所で荷物を受け取る際の待機時間中に発生している．待機時間とそれに伴うストレスの発生や顧客満足度への影響は，様々な状況を対象として研究報告がされている．

2.1　待ち行列と顧客満足度の関係についての研究

Maister（1985）[2] は，待機時間と顧客の不満発生の関係において，客観的実時間より顧客の感じる主観的知覚時間が重要だと述べている．また，主観的に知覚する時間を前後させる条件として，以下の 8 つの項目を提示している．

- 何かを行っている時間は，何もしていない時間より短く感じられる
- 何かが開始する前の待ち時間は，既に始まった後の待ち時間より長く感じられる
- 不安が待ち時間を長く感じさせる
- 待ち時間の長さが予想不可能な場合は，長く感じられる
- 想定していなかった事象によって発生した待ち時間は，長く感じられる
- 不公平な待ち時間は，より長く感じられる
- 価値が高いものに対しては，長時間待っていられる
- 一人の場合は，複数人で居るより長く感じられる

磯田（2018）の待ち時間と消費者行動に関する研究 [3] は，Maister（1985）などによって顧客の知覚する体験が満足度と

大きく紐づいていることが明らかになって以降，顧客体験に関する研究が進められ，その過程で顧客満足度は待機時間中に顧客に生起した感情によって大きく影響されることが明らかになったと述べている．

Demoulin ら（2013）のネットバンキングの顧客を対象にした待ち時間に関する研究 [4] は，顧客の期待する待機時間と実際に待機した時間の不一致が待機に対する受容性に影響を与え，その後顧客に情緒的な反応を生起させると報告している．また，機会コストと待機の受容性の関係性について，他に有益な事が出来たと顧客に感じさせた場合など，機会コストが高い場合は顧客の待機時間の受容性に負の影響を与える一方で，高い機会コストが生じている場合でも，顧客の待機時間に対する受容性に応じてその影響の強さが上下する事を明らかにした．例として，時間の不一致が顧客にとって好意的で待機時間の受容性が高い場合は，機会コストが高い場合でもサービスに対して高い評価がされる傾向があると述べている．

2.2 手荷物受け取りのユーザー体験向上事例

手荷物受取所での待機は，荷物を待つ以外に出来ることが少ない上，待機する時間の長さの予測が難しい．これは，Masiter（2018）の提示する知覚時間が長くなりやすい条件を満たしている．また，待つ以外に出来ることが少ない状況は，顧客に他に有益なことが出来たと感じさせやすく，知覚時間の長さとあわせて旅客の待機に対する受容性に影響を与え，負の感情を生起させることでストレス発生の要因となっている．

この空港手荷物受取所でのユーザー体験を原因とするストレスに対して，対策を実施し効果を上げた空港がいくつか存在する．ヒューストン空港は，待機時間のあり方を変える事によってストレスの解消を実現した [5]．ヒューストン空港では，手荷物受取所における，長い待機時間に起因する苦情の多さが課題だった．空港は地上スタッフを増員し，平均待機時間を業界標準より短くする事に成功したが，苦情の件数は改善しなかった．原因調査の結果，平均待機時間 8 分の内訳が到着ゲートから手荷物受取所に歩いて到着するまでが 1 分，手荷物受取所に到着してから荷物を受け取るまでの時間が 7 分である事を発見した．ストレスの要因は何もしていない時間にある事に着目したヒューストン空港は，到着ゲートから手荷物受取所までの経路を変更し，移動距離を長くする事で受取所において待機する時間を減少させ，代わりに歩いている時間を増加させた．結果，手荷物受け取りに関するクレームの大半を減少させる事に成功した．

手荷物受取所における待機時間を減少させるのではなく，荷物がいつやってくるのか旅客に提示する事で，ストレス解消に繋げている事例もある．南紀白浜空港では，日本航空と NEC によって「IoT おもてなしサービス実証」という取り組みが行われている [6]．この取り組みの一貫として手荷物待ち時間可視化サイネージが手荷物受取所に設置されている．このサービスは，事前に登録した旅客に対して，会員情報のステータスにもとづいた手荷物返却予定時刻を提供する．返却予定時刻が分かる事で，旅客は手荷物を受け取るまでの時間をレーンの前で拘束される事なく有効利用が可能になる事に加え，待機時間の不確実性を取り除く事によって，心理的コストの減少も実現している．

3 空港手荷物受取所における拡張現実型情報提示サービス

各地で手荷物受け取りの際のストレスの軽減策が実施され，成功を収めている事例も存在するが，一方で制約も存在する．例えば，どこの空港でもヒューストン空港のように旅客を歩かせる時間を長くする事は，設備に制約があり難しい．返却時間をサイネージなどで提供する施策に関しても，数百と存在する旅客の荷物を識別して返却時間を個別に提供するのは困難である．

本研究は，特定の空港，特定の旅客に限らない手法で，旅客が手荷物を受け取る際のユーザー体験を向上させ，ストレス解消と手荷物を預ける事に対する忌避感の減少を目指す．ユーザー体験向上のためには，知覚時間短縮や待機に対する受容性改善に伴う後ろ向きな感情生起の抑制が必要になる．これを実現するために，手荷物受け取りの一連のプロセス中で利用可能なモバイルアプリケーションを提案する．

3.1 手荷物受取所内での AR サービス提供

特定の空港に依らない手法の実現には，空港に存在する設備を利用する必要がある．本研究は手荷物受取所と AR の組み合わせに着目する．手荷物受取所は，どこの空港にも存在し，手荷物に関するストレスの発生環境となっている事から，この空間のユーザ体験向上は旅客のストレス解消のために必須である．AR は，国内のスマートフォンの個人所有率が 8 割近く [7] となり，AR を実行可能な端末を旅客の大半が有している中で，手荷物受取所の空間を活かしたサービスを，利用する動機を持ってもらいやすい形での提供が可能である．また，施設に対し大きく手を加える必要なく，サービス展開に掛かる費用が廉価で済み，場所を選ばないという利点も存在する．

3.2 空港における旅客のニーズ

利用する動機を持ってもらうにあたって，旅客のニーズを適切に把握する事が重要となる．国交省の「空港における訪日外国人へのサービスのあり方に関する提言」[8] によると，交通機関や空港施設の案内，観光に関する情報が空港に到着した旅客にとって重要な情報として捉えられている．また，じゃらんリサーチセンターによると，日帰り旅行客の中で旅行先の食を楽しむ事を目的とするのは全体の 43.6% [9] と，半数近くの旅行客が食を旅行目的の一つとしている事が明らかとなっている．

以上から，提供するサービスは，空港所在地の食に関するものを主軸として，旅客にとってニーズの高い情報を手荷物受取所内で提供する事で利用動機を確保する．

3.3 提案する AR サービスの内容

本研究は，手荷物受取所の荷物レーンを活用した，手荷物を受け取る際の待機時間を利用して旅客にグルメ情報を提供する

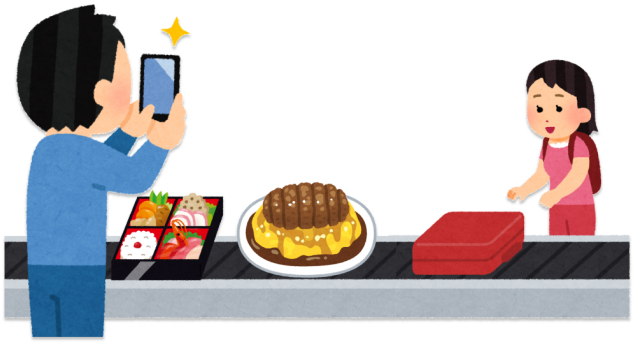


図 1: サービス提供のイメージ

AR サービス,「ちょいぐる」を提案する．荷物レーンを利用することで，旅客の行動を妨げることなく荷物を待ちながら AR 体験を利用可能とする．

荷物レーンを使用した情報提供は，図 1 に示すように荷物レーンを回転寿司のレーンに見立て，コンテンツとなる 3D オブジェクトをレーン上に表示させる回転寿司のようなスタイルで行う．他に無いスタイルにする事で，旅客の目を惹きやすくして利用する動機を持って貰いやすくする．レーン上を流すコンテンツは，空港所在地周辺の食をテーマとしたものにする．各コンテンツ共にタップして開く詳細メニューを用意する．メニューから所在地や営業時間といった情報をユーザーに提示する事で，実際に店舗へ赴きやすいよう導線を確保する．動作プラットフォームは，スマートフォンなどのモバイル端末を想定する．手荷物受取所など目に止まりやすい場所に QR コードなどを掲示する事で，旅客が空港到着後にすぐインストールして利用可能にする．

3.4 サービス利用フロー

まず，位置選択メニューで，ユーザーの自己位置選択を行う．位置選択は，図 2a の位置選択メニューで，表示を参考に位置を選択する．位置選択が済むと図 2b の様な画面となるので，荷物レーンにカメラを向け，コンテンツが表示されるのを待つ．ここから，コンテンツを AR 空間の適切な場所へと配置するために端末が空間のどこに存在しているか判断する，位置推定が動作し始める．この位置推定が上手くいっていないと判断される場合は，数秒後に荷物レーンにカメラを向けるか位置選択を見直すよう教導される．位置推定が上手く行くと図 2c のようにコンテンツがレーン上に表示され始める．レーン上を周回するコンテンツをタッチする事で，図 2d のようにコンテンツに応じたメニューが開き，店舗情報などを確認する事ができる．コンテンツは，位置推定が機能していると判断される場合は表示され続ける．

(a) マップ選択メニュー **(b)** レーン上へのコンテンツ表示

(c) 荷物レーン上へのコンテンツ表示 **(d)** コンテンツの詳細メニュー

図 2: AR サービスの主要画面

4 提案する AR サービスのシステム

4.0.1 AR 空間上での端末の自己位置推定

高品質な AR 体験を実現するにあたって，AR 空間と現実との間で整合性をとる事が重要となる．この整合性は主に以下の 3 つの項目が存在する [10].

- 幾何学的整合性（AR 空間と現実の間で位置・角度・大きさのズレが生じていない事）
- 光学的整合性（AR 空間と現実の間で光源や影にズレが生じていない事）
- 時間的整合性（AR 空間と現実の間で動作に時間的ズレが生じていない事）

空間の特性を活かした AR 体験を実現するにあたって，適切な位置に 3D オブジェクトを配置して整合性を取るためには，特に幾何学的整合性と時間的整合性が重要となる．この 2 つの整合性を保つ為には，空間内における端末の自己位置をリアルタイムで取得する必要がある．

位置推定の手段は，GPS で位置を測位したり，目印となるマーカーを設置する手法が主流だが，本研究でサービスを提供する空間は屋内のため，GPS を使用した位置推定が使用出来ない事に加え，常にマーカーを写し続ける事を想定したサービスではないため，これらの手法は使用できない．代わりに，本サービスではカメラ画像から位置推定を行う Visual Positioning

図 3: 函館空港の点群データ

(a) 特徴量が低い領域の点群

(b) 手荷物受取所に存在する複数の同じ広告

図 4: VPS が適切に作動しない条件

System（以下，VPS）を使用する．

VPS は，撮影した画像から位置推定を行う技術である．位置推定を行う空間を事前に図 3 に示すように点群データとして取得する必要がある．カメラで撮影した画像の特徴点からマッピングを行い，結果を事前に取得した点群データと比較する事で空間上の端末の位置を推定する．VPS を実現するサービスは，Niantic による LightShipVPS [11] や Google による ARCore Geospatial API [12]，Immersal [13] などが存在するが，点群データが取得可能な環境に自由度がある点から，Immersal を採用する．

4.0.2 手荷物受取所の点群データ

VPS での位置推定を実現するにあたって必要な点群データは，函館空港手荷物受取所にてマッピングを行い作成した．点群データのマッピングは Immersal で標準に用意されている Mapper アプリを使用してスマートフォンで行う．点群データの作成は，何枚もの画像から被写体の立体モデルを取得するフォトグラメトリと同様に，被写体となる空間を様々なアングルから撮影する事で行う．被写体の周囲 360 度から全周に渡って撮影する事が好ましいが，手荷物受取所には荷物レーンを始めとして立ち入る事が難しい空間が存在する事に加え，壁で囲まれているため，周囲 180 度から撮影した．被写体をより詳細に撮影する為に近づいて撮影を行うと，1 枚で撮影出来る領域が狭い事に加え，生成したデータを繋ぎ合わせる際にエラーが生じやすいため，一定以上の距離から離れて撮影を行った．

手荷物受取所を区画ごとに区切らず，全体として一つのマップを作成し，その点群データを使用して VPS の動作確認を行ったところ，一定の位置やアングルで適切に位置推定が動作しない事が明らかとなった．動作しない条件として，手荷物受取所の白い壁と荷物レーンしか特徴点として使用できない状況と，図 4a のように空間内に同じものが複数存在する広告が写った状況の 2 つが存在した．前者は，白い壁や黒一色の荷物レーンのベルトのように一様なものは特徴点に乏しく，図 4b のようにその空間の点群が作成されていない事が原因であった．後者は，広告のようなカラフルで特徴量の多いものは，位置推定を行う際に基準として位置推定が行われやすいため，同様の特徴点が複数個点群データ上に存在する場合，位置推定を行う際に他の同様の特徴点を持つ点群と区別がつかず，誤った位置に座標をセットしてしまう事が原因であった．

前者に対しては，配置する 3D コンテンツの動作を工夫して，点群が作成されていない場所を映し続けないように促す事によって対応した．後者に対しては，取得する点群を分割し，同じ広告が複数個，同じ点群に存在しないように対策を行って点群データを作成し，対応を行った．

4.0.3 開発ツール

開発環境として，Immersal をはじめとする AR 向けのパッケージが揃っている事に加え，IOS と Android の双方にアプリケーションを提供するのが容易なため，Unity（2021.3.15f1）を使用する．AR 用のライブラリは，ARFoundation を使用する．

4.0.4 AR 空間上のコンテンツ配置

AR 空間にコンテンツを配置するにあたって，VPS による

位置推定の不安定さによるズレと，特徴点に乏しい空間をユーザーに写させ続けさせないように対応を行う必要がある．荷物レーン上に配置する 3D コンテンツの大きさを，スーツケースと同じくらいの大きさに設定し，多少の位置ズレが生じた際にユーザーにズレを感じさせないようにしつつ，コンテンツの画面中の動作量を大きくする事で，特徴点が多い場所をユーザが写す機会を作り，位置推定が安定的に行われるよう設定した．

4.0.5 AR 空間上のコンテンツ遮蔽処理

ARFoundation には，Simultaneous Localization and Mapping（以下，SLAM）を使用して周囲の空間をリアルタイムでマッピング，自己位置推定を行い，得た情報を利用してオブジェクトが実際の空間上にある障害物の裏に隠れた際にオブジェクトを遮蔽し見えなくする機能が存在する．しかし，この機能は点群データの作成と同様に画像からマップの作成や自己位置推定を行うため，一様な色の壁など特徴点の無い場所に弱く，そういった所では適切にオブジェクトが遮蔽されない．今回はこの機能を使用せず，代わりにオブジェクトの遮蔽処理に使用するデータを点群データを参考にしつつ作成し，空間上に配置して遮蔽処理を行った．

4.0.6 システム詳細

本システムの主要画面を図 2 に示す．提案サービスの主な機能として，位置選択，手荷物受取所の荷物レーン上への 3D コンテンツの表示，3D コンテンツをタッチした際の詳細メニュー表示がある．位置選択では，ユーザーが今いる位置に応じて，位置選択メニューを通して現在位置を選択させる．この機能は，点群データを分割して取得した事で，どこにユーザーが居るのか把握する必要が生じたため追加した．手荷物受取所荷物レーン上への 3D コンテンツの表示は，VPS で取得した自己位置をもとに，AR 空間上で荷物レーンの上に相当する座標に 3D コンテンツを配置する事で行われる．VPS による位置推定は，位置選択メニューでの位置選択後，カメラを荷物レーンに向けるだけで自動的に作動する．3D コンテンツは全 12 種類存在し，回転寿司のようにそれぞれの 3D コンテンツがレーン上を周回し続ける．3D コンテンツをタッチした際の詳細メニュー表示は，図 2d のように 12 種類存在する各コンテンツごと，詳細の情報が記載されたメニューが表示される．詳細メニューの内容はコンテンツとなった料理を提供する店の情報を掲載する．各機能に最初にアクセスする際や位置推定が上手く行っていないと判断される状況で，ユーザインタフェース上に次にタッチすべき項目や行うべき操作を教導するための表示がなされる．教導は，次に操作すべき UI 要素に対してハンドマークを表示させるもの，画面にメニューを表示して文字ベースで伝達するものの 2 種類を設定した．

5 評　　価

評価は提案するアプリケーションが，空港手荷物受取所のユーザ体験向上に繋がるか検証するために行った．

5.1 評価条件・環境

評価は，函館空港手荷物受取所で行った．手荷物受取所に 2 つあるレーンの内，当初は片方のみで行い，途中からもう一方のレーンを評価環境として追加した．主に学内と空港関係者を対象として被験者を募集した．実際の利用シーンを再現するため，評価には明確な期日を設けず，各々の被験者が都合により手荷物受取所を利用した際に提案サービスを使用してもらった．利用後はアンケートに回答させ，アプリケーション利用による手荷物受取所でのユーザ体験向上効果を取得した．現在，アンケートは 13 名から回答を取得している．

5.2 アンケート

効果検証の手段として，アンケートを使用した．回答者の負担軽減のため，アンケートは前半と後半に分割し，前半部分は必須回答項目，後半部分は任意回答項目として回答の是非を回答者に選択させた．回答者の混乱を防ぐため一部の項目を除いて，項目は五段階のリッカート尺度に統一して回答させた．五段階のリッカート尺度の内，一番低いものを 1，一番高いものを 5 と値を設定した．

5.3 必須回答項目

必須回答項目は表 1 の通り全 8 項目で構成される．必須回答項目からは，利用した際の没入感，好感度を取得する．没入感からは，利用した際にどれだけサービスを利用するのに集中したか測定し，利用に伴う知覚時間短縮効果を検証する．好感度からは，サービスの利用に伴ってにユーザーに生起した感情を測定し，ユーザーの待機に対する受容性改善効果を検証する．

最初の項目,「プレイ時間（自動入力）」からはユーザーがサービスを利用した時間を取得する．Q1 から Q3 までで没入感を，Q4 から Q6 で好感度を取得する．Q1 から Q5 は，ゲームプレイ後のユーザ体験を調査するのに使用される，Game Experience Questionnaire（以下，GEQ）[14] の，没入感から 3 問，ポジティブ感情，ネガティブ感情から 1 問ずつ，効果検証として適切な項目を抜粋し項目を作成した．本来 GEQ は 33 項目で構成されるが，評価は被験者各自が都合によって手荷物受取所を利用した際にサービスを利用してもらう形で実施したため，項目数を多く確保するとサービス利用者の回答意欲を損ね，回答の回収可能性に難が生じると判断し，必要最小限の項目のみを抜粋する形をとった．Q6 では，サービスに対する好感度，継続意思を測る Net Promoter Score（以下，NPS）を使用している．本来 NPS は，11 段階のリッカート尺度で測定するが，回答のリッカート尺度を統一するため五段階にあらためて回答を設定した．

5.4 Game Experience Questionnaire（GEQ）

GEQ [14] とは，IJsselsteijn の開発したゲームプレイのユーザー体験を調査するためのアンケートである．コアモジュールなど，3 つのモジュールから構成される．コアモジュールは，没入感，フロー（心理学），適性，ポジティブ感情，ネガティブ感情，緊張感，チャレンジの 7 つのセクションで構成され，全

部で 33 個存在する項目を 5 段階のリッカート尺度を使用して評価する．

5.5 Net Promoter Score（NPS）

NPS [15] とは，顧客の製品やサービスに対する信頼や愛着を測るスケールである．製品・サービスを他人にも推薦する事が出来るかどうかを 11 段階のリッカート尺度を使用して問う．回答に応じて推奨者，中立者，批判者に分類し，推奨者の割合から批判者の割合を引いた結果がスコアになる．顧客満足度と比較して，その後の購買行動や推奨行動と相関が高いという特徴がある．NPS のスコアは以下の計算式で算出する．

$$\text{NPS} = (\text{推奨者の割合 (\%)} - \text{批判者の割合 (\%)}) \qquad (1)$$

5.6 任意回答項目

任意回答項目は表 2 の通り全 13 項目で構成される．任意回答項目からは，利用者の属性，利用した際に感じたサービスのユーザービリティ，サービスに対する要望を取得する．ユーザービリティからは，提案するサービスが適切な使い心地であるか，適切な情報量であるかなどを測定し，手荷物受取所で展開するにあたって適切なものかどうか検証する．

Q9 から Q18 は，主にユーザビリティの評価に使用される，System Usability Scale（以下，SUS）[16] を日本語訳し，本研究に最適化したもの使用して項目を作成した．Q14 の項目に当初誤りがあり,「全体として、操作や見た目などに一貫性が無いと感じましたか？」とすべき所を,「全体として、操作や見た目などに一貫性があると感じましたか？」と真逆の問いとなっていた．回答が 3 件集まった時点で設問に修正を行った．回答結果は，真逆の問となっていたため，スコアを反転させて対応を行った．

5.7 System Usability Scale（SUS）

SUS とは，John Brooke の開発したシステムのユーザビリティを調査するためのアンケートである．全部で 10 の項目を，5 段階のリッカート尺度を使用して評価する．奇数番目の項目が前向きな評価，偶数番目の項目が後ろ向きな評価を問うもので，100 点満点中，平均値である 68 点を上回るかどうかなどを基準としてユーザビリティの定量的な評価が可能である．SUS のスコアは以下の計算式で算出する．

$$\text{SUS スコア} = \left(\sum_{i=1}^{10} \text{スコア}_i\right) \times 2.5 \qquad (2)$$

- 奇数番号の項目は $\text{スコア}_i = \text{回答} - 1$,
- 偶数番号の項目は $\text{スコア}_i = 5 - \text{回答}$.

6 評価結果

アンケート結果を分析し，提案サービスが手荷物受取所において，待機中の知覚時間短縮や待機に対する受容性の改善に伴う前向きな感情生起を促し，ユーザー体験向上に繋がるのか検証する．

表 1: 必須回答項目（*が GEQ を抜粋した項目）

	質問内容
	プレイ時間（自動入力）
Q1	リアルさを感じましたか？*
Q2	強く印象に残りましたか？*
Q3	リアルな体験だと感じましたか？*
Q4	満足感はありましたか？*
Q5	疲労感を感じましたか？*
Q6	本サービスは周りの友人などにおすすめできますか？
Q7	追加項目についても回答して頂けますか？

表 2: 任意回答項目（*が SUS の回答項目）

	質問内容
Q8	あなたは以下の組織・身分の内どれに相当しますか？
Q9	空港でこのサービスが提供されていた場合、使用したいと思いますか？*
Q10	利用する際に、必要以上に複雑だと感じましたか？*
Q11	シンプルで使いやすいと感じましたか？*
Q12	利用する際に、なにかサポートを必要と感じる事はありましたか？*
Q13	様々な機能の間を、違和感なくスムーズに移動できましたか？*
Q14	全体として、操作や見た目などに一貫性が無いと感じましたか？*
Q15	ほとんどの人がすぐに利用方法を習得できると思いますか？*
Q16	利用する際に、操作しづらいと感じることはありましたか？*
Q17	自信をもって使いこなすことが出来ると思いますか？*
Q18	利用する際に前もって知っておかなければならない事は多いと思いますか？*
Q19	以下の項目の中で、追加してほしい機能はありますか？(複数選択可)
Q20	何か気になることやコメントなどありましたらご記入ください。(記入式)

6.1 必須回答項目結果

必須回答項目 Q1-Q5 の結果は図 5 に示す.

6.1.1 没 入 感

没入感を取得する Q1-Q3 の平均は 3.9 となった．値が 1 から 5 の範囲に収まる事を考慮すると，3.9 という結果はユーザーがある程度没入感をもって利用していると判断可能な範囲に収まっているため，ユーザーの知覚時間短縮効果が存在する可能性が期待できる．印象に残ったかを問う Q2 が特に高得点な事から，本サービスの提供内容が，ユーザーにとって興味を引く対象となって没入感を生じさせる要因になっていると考える．

6.1.2 好 感 度

好感度を取得する Q4-Q6 の内，ポジティブ感情を取得する Q4=3.8，ネガティブ感情を取得する Q5=1.9 となった．Q4 と Q5 の値の差が 1.9 存在し，前向きな感情が後ろ向きな感情を大きく上回っている事から，利用した際のユーザーの待機に対する受容性の改善効果が認められる可能性がある．

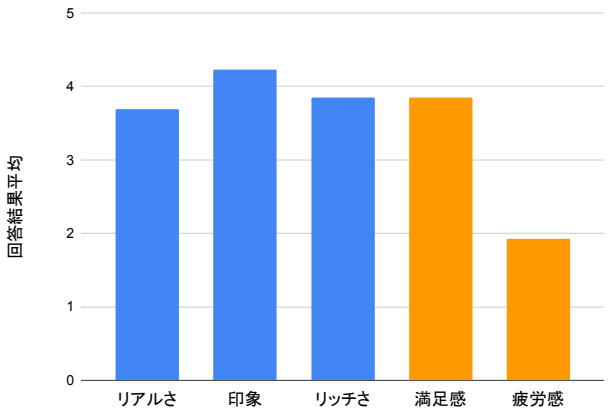


図 5: 必須回答項目 Q1-Q5 の結果

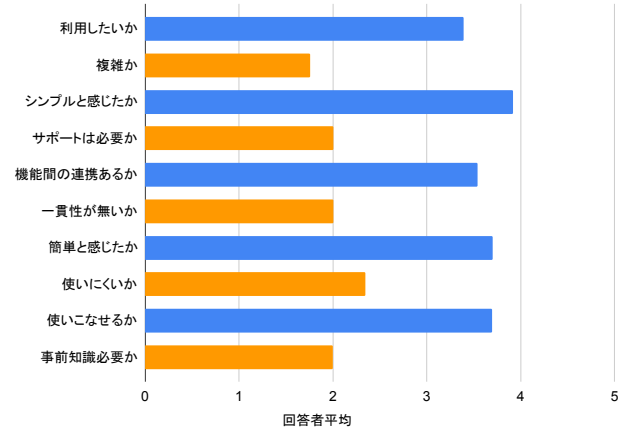


図 6: 任意回答項目 SUS の結果

Q6，NPS のスコアは 7.6%となった．値を正の範囲に収められた事から，ユーザーはサービスに対して一定の好感度や利用継続意思を示している可能性がある．今回，NPS を本来の 11 段階から 5 段階するにあたって，本来の NPS では 0 から 10 の値の内 10 と 9 を推奨者に，6 以下の値を批判者に割り当てるが，今回は 1 から 5 の値の内 5 を推奨者に，3 以下を批判者に割り当てた．

6.1.3 プレイ時間

プレイ時間の平均は，値を適切に取得出来ていないと判断可能な取得結果 1 秒のものを 3 件除いて 101.2 秒であった．(以下，同様に値を取り除く)

プレイ時間と没入感を取得する Q1-Q3 の回答平均の間の相関係数は-0.203 で，プレイ時間と回答結果に強い相関は見られなかった．同様に，プレイ時間と好感度を取得する Q4 と Q5 の回答の差の相関係数は 0.111 で，プレイ時間と回答結果に強い相関は見られなかった．

6.2 任意回答項目結果

任意回答項目 Q9-Q18 の結果は図 6 に示す．

6.2.1 SUS

SUS のスコアは 72.1 となった．基準となる平均の 68 点を上回った事で，手荷物受取所で展開するにあたって適切なユーザービリティが確保されていると判断する．中でも，Q10 や Q11 のシンプルさを問う項目が高得点となっている事から，機能が不必要に複雑でなく適切な範囲に収まっていると評価されていると判断する．

一方，操作の容易さを問う Q15 と Q17 が低得点でないのにも関わらず，操作のしづらさを問う Q16 が比較的低得点であった．Q15-17 の標準偏差を算出すると，Q15=1.16，Q16=1.51，Q17=1.08 であった．Q9-Q18 の平均標準偏差が 1.08 であった事もあわせて，Q16 は特に値にばらつきが生じている事が示されている．Q16 の設問が「利用する際に、操作しづらいと感じることはありましたか？」であった事から，全体的に操作はしやすいものの，一部の条件下で操作がしづらいと感じさせてしまう状況が発生していると考える．

6.2.2 サービスに対する要望

Q19 にて追加してほしいサービスの機能について回答させた．結果は，交通機関に関する情報の提供を望む声が 8 件，空港施設に関する情報提供も 5 件と，この両者が比較的多かった．両者は，国交省の調査においても空港到着時に必要とされる情報として上位にあった事から，ユーザーのニーズが高い事が明らかとなっている．今後，これらの情報の提供を行う事で，手荷物受取所におけるユーザー体験を更に向上させる事が可能になる可能性が高い．

Q20 では，自由記述形式でサービスに対するコメントを回答させた．要望として，コンテンツが荷物レーン上を流れる速度が速いという声が多かった．現在は，荷物レーンのベルトコンベアが流れる速度と同じのスピードに設定しているが，スマートフォンのカメラの視野角は人間の目と比較して狭いため，ユーザーがコンテンツを追う際に，端末を頻繁に大きく動かしたりする必要があるなど，ユーザーに不満を感じさせやすくなっている可能性が高い．また，ユーザーの立つ位置に応じて目の前にコンテンツが流れてくるのに時間がかかり，その時間が退屈だという声や，コンテンツの見た目が悪く魅力が伝わってこない，コンテンツの量が不足していて楽しめる時間が少ないといった声も頂いた．以上から，基本的な機能は抑えられているものの，一つひとつの機能の完成度やコンテンツの情報量に不足がある事が示された．サービス全体として修正，機能追加を行いユーザー体験向上を目指す必要がある．

7 おわりに

本研究は，空港到着後に旅客が自身の手荷物を受け取るまでの一連のプロセスに起因する旅客のストレスの解消，手荷物を預ける事に対する忌避感の減少を目標とするサービスの提案，実験評価を行う事を目的として行った．サービス開発にあたって，手荷物受け取りに関するストレスが，手荷物受取所での待機時間に発生している事に着目し，手荷物受取所で使用したくなる特別な体験を提供する事を目標に，手荷物受取所で利用可能な回転寿司スタイルで食に関する情報を提供する AR アプリケーションを開発した．サービスの効果検証を目的として，函

館空港手荷物受取所でサービスを実際に利用してもらい，利用後にアンケートに回答させる事で評価実験を行った．評価結果より，サービスを利用する事でポジティブな感情が生起された事に加え，没入感があると評価された事から，待機時間に対する受容性の増加，知覚時間短縮効果が一定程度あると予想する．ストレスの発生は，待機時間中に生起した感情や知覚時間，受容性に大きく影響される事から，本研究の目標であるストレス解消は一定程度満たされたと判断する．一方，利用した際に知覚する待機時間の短縮効果や，待機する事に対する受容性の増加が存在すると判断はしたものの，具体的にどの程度の効果なのか，この 2 つの効果によってどれだけストレス解消効果が認められるのか，更にストレス解消効果がどの程度手荷物受け取りに対する忌避感を減少させるのか，まだ明らかとなっていない．今後は，関連研究などを参考としながら，サービス利用にともなってどの程度旅客のストレス解消に繋がるのか，ストレス解消効果が手荷物受け取りに対する忌避感減少に繋がるのか検証する．

謝　　辞

本研究は，保安検査場内への入場や手荷物受取所の利用を始め，日本航空 函館空港所と北海道エアポート社の多大な協力のもとで行った．ここに記して謝意を表す．

文　　献

[1] 国土交通省，特定本邦航空運送事業者に係る情報，https://www.mlit.go.jp/koku/content/001632069.pdf，2023
[2] David H Maister，The Psychology of Waiting Lines，https://davidmaister.com/wp-content/themes/davidmaister/pdf/PsycholgyofWaitingLines751.pdf，1985
[3] 磯田 友里子，購買に伴う待ち時間が消費者行動に与える影響，消費者行動研究，vol24-2，pp.73-100，2018
[4] Nathalie T.M. Demoulin，Souad Djelassi，Customer responses to waits for online banking service delivery，International Journal of Retail & Distribution Management，vol41-6，pp442-460，2013
[5] Alex Stone，Why Waiting Is Tortune，https://www.nytimes.com/2012/08/19/opinion/sunday/why-waiting-in-line-is-torture.html，2012
[6] NEC，南紀白浜 IoT おもてなしサービス実証 - NEC Corporation，https://jpn.nec.com/delight/retail/shirahama-iot/index.html，2020
[7] 総務省，令和 4 年通信利用動向調査の結果，https://www.soumu.go.jp/johotsusintokei/statistics/data/230529_1.pdf，1-2，2023
[8] 国土交通省，空港における訪日外国人旅行者へのサービスのあり方に関する提言，https://www.mlit.go.jp/common/000995682.pdf，11，2013
[9] じゃらん，日帰り旅行に関する調査（リクルートじゃらんリサーチセンター調べ），https://jrc.jalan.net/wp-content/uploads/2022/04/c5af5c49532af6a4c9c345d59e4b5ef8.pdf，2022
[10] 神原誠之，基礎 1：拡張現実感 （Augmented Reality：AR）概論，https://jglobal.jst.go.jp/detail?JGLOBAL_ID=201002260158130348，情報処理 (IPSJ Magazine)，vol51-4，pp.367-372，2010
[11] Niantic，Lightship VPS，https://lightship.dev/ja/docs/ardk/vps/index.html，2024
[12] Google，Geospatial Creator — ARCore，https://developers.google.com/ar/geospatialcreator?hl=ja，2024
[13] Hexagon，Immersal，https://immersal.com/，2024
[14] IJsselsteijn，W. A.，de Kort，Y. A. W.，& Poels，K.，The Game Experience Questionnaire.，pp.1-9，2013
[15] Vince Nowinski，NTT コムオンライン・マーケティング・ソリューション翻訳，NPS ホワイトペーパー ネットプロモーター ® の経済学，https://www.nttcoms.com/service/nps/documents/nps_doc02_150126.pdf，2015
[16] Brooke.J，“ SUS: a ”quick and dirty” usability scale ”，Usability Evaluation in Industry.，pp.189-194，1986

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

2024年2月28日(水) 13:30 ~ 15:40 | T5-A オンライン（Zoom Events)

教育・学習 (2)

座長:熊本 忠彦(千葉工業大学)
コメンテータ:飯沢 篤志(リコーITソリューションズ株式会社)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T5-A-2-01] 3段階の学習過程を前提とした自己学習環境における講義動画閲覧履歴による学習状況分析

*齋藤 光貴[1]、渡辺 知恵美[2] (1. 筑波技術大学 産業技術学部 産業情報学科 情報科学コース 渡辺研究室、2. 筑波技術大学 産業技術学部 産業情報学科)

13:50 ~ 14:05
[T5-A-2-02] 教材ビデオにおける学習者の視聴操作に基づくシーン選択支援

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

14:05 ~ 14:30
[T5-A-2-03] ピクトグラミーティング - ピクトグラムコンテンツの共有及び相互評価システム -

*松村 貫司[1]、石井 幹大[2]、伊藤 一成[1] (1. 青山学院大学社会情報学部、2. 青山学院大学理工学部)

14:30 ~ 14:55
[T5-A-2-04] 学生向け学務情報のシステム透過なオープンデータ化

*中川 匠海[1]、松原 克弥[1] (1. 公立はこだて未来大学)

14:55 ~ 15:20
[T5-A-2-05] ユニバーサルピクトグラミング　－人型ピクトグラムを用いたプログラミング学習環境の多ヶ国語化－

*石井 幹大[1]、伊藤 一成[2] (1. 青山学院大学 理工学部、2. 青山学院大学 社会情報学部)

3段階の学習過程を前提とした自己学習環境における講義動画閲覧履歴による学習状況分析

齋藤　光貴†　　渡辺知恵美†

† 筑波技術大学 産業技術学部 産業情報学科　〒 305-8520 茨城県つくば市天久保 4 丁目 3-15
E-mail: †{a211117,chiemi}@a.tsukuba-tech.ac.jp

あらまし　近年、オンライン学習環境が充実し、動画を用いた学習が自己学習の一般的手段となっている。先行研究では、講義動画のクリックストリームデータを分析することで脱落予測や成績予測を高精度にすることができたと報告されている。しかし、我々は自らの経験から、講義動画閲覧履歴からより具体的に学習者の学習状況を把握する必要があると考えている。その考えから、自由進度学習環境における学習者の学習行為を観察した先行研究に基づいて、動画閲覧履歴をもとに普段の授業では把握できない学生の理解状況を把握できるという仮説を提起する。自由進度学習環境が用意されている大学の講義で使用されている講義動画の閲覧履歴を取得し、分析・検証を試みた。

キーワード　教育：e-Learning, EdTech, 講義動画閲覧履歴, 学習状況, 自由進度学習

1　序　　論

著者が高校生の時にオンラインで行われた授業で、学習内容に関する自身の理解状況を教員との間で共有することが難しかったという経験をした。著者が所属していた当時の高校では、1ヶ月分の課題を各生徒の自宅に送付し、各生徒はその課題を解いてアプリに提出する形式でオンライン授業を行っていた。その課題について疑問があった時、教員にチャットでどの部分についてどこまで理解できていてどのような疑問があるかという説明を一からしなければ、自身の理解状況を教員に伝えることが難しかった。教員の立場から考えると、著者がやっている課題の内容や教科書、ノート、見ている動画シーンといった情報がない。そのため、著者が理解状況を一から説明しない限り、教員は著者の理解状況を把握することができない。また、質問がない学習者の理解状況を把握できないという問題もある。質問してきた学習者ならば、対話を通して理解状況を把握することは可能である。しかし、質問してこない学習者との対話の機会は生まれないため、理解状況を把握することができない。この経験から、学習者をサポートするためには教員が学習者の理解状況を容易に把握できることが大事だと考えた。

教員が学習者の状況を把握するために、学習コンテンツのアクセス履歴から脱落予測や最終成績予測を行った関連研究は数多あるが、途中の理解状況を容易に把握できないという課題を解決できていない。それらの関連研究は、最終成績や脱落といった最終学習成果を予測しているが、途中までの理解状況を具体的に示せていない。そのため、教員は最終学習成果の予測結果から学習支援すべき学生を特定できても、具体的な学習支援方法までは把握することができていない。

以上から、本研究は動画閲覧履歴をもとに普段の授業では把握できない学生の理解状況を把握できるかを問いとする。ここでいう普段の授業とは、大学で一般的に行われている講義・演習のことを指す。

本稿では問いに対する一つのアプローチとして、動画閲覧履歴を利用した学習の理解状況の可視化と分析について報告する。我々は今回、学習時の教材コンテンツへのアクセス履歴が把握できる授業方法である「自由進度学習」に着目し、講義動画コンテンツを教材として用いる授業を対象とする。本学において自由進度学習を適用している授業「データベース論・演習」の教員と学生に協力を得て、動画閲覧履歴を取得し、分析・検証を試みた。

2 節で関連研究を示す。3 節では対象とする自由進度学習環境と学習フェーズ、動画閲覧履歴、データセットを説明する。4 節では分析結果と考察を述べる。5 節で結論と今後の課題について議論する。

2　関 連 研 究

クリックストリームデータなどの学習履歴を活用した学習状況分析は教育データマイニング及び学習分析分野において数多く行われている。例として、MOOCs (Massive Open Online Courses) のクリックストリームデータから、KNN,SVM,ANN などの機械学習でコースの合格不合格を予測するもの [1] や、OULAD (Open University Learning Analytics Dataset) のクリックストリームデータから、LSTM,CNN などのディープラーニングによるアプローチと従来の機械学習アプローチと比較して LSTM が他のアプローチより精度高く学生の成績予測を行えたと報告されたもの [2]、SMOTE (Synthetic Minority Over-sampling Technique) を用いて、不均衡なデータセットから DeepCNN で 99.9 %という精度で予測できたというもの [3] などがある。これらの研究では、機械学習を用いて学生の脱落予測や成績予測の精度向上や様々な教育データの適応化に貢献している。しかし、あくまでも学生の最終学習成果を予測することを目的にしている。学生の最終学習成果の予測結果から学習状況を間接的に推測することはできるが、より具体的な学習状況を可視化するまでには至っていない。本研究は学習状況を

より具体的に明らかにすることが目的であるという点でこれらの研究とは根本的に目的が異なる。

また、MOOCs ではなく自習型教育 (self-paced education) に着目し、deep LSTM model を用いて合格と不合格の学生を分類するもの [4] もある。この研究では自習型教育に着目しており、本研究が想定する自由進度学習とかなり似ている。しかし、この研究は自習型教育において合格不合格の学生を予測することが目的であり、学習状況を明らかにすることを目的にする本研究とは異なる。

さらに、MOOC のクリックストリームデータから特徴量を抽出し、CFA（一発正解）である可能性が高いかどうかを定量化するもの [5] もある。この研究は、特徴量を抽出するという点で本研究と近い点があるが、質問に対して 1 発正解できるかに着目しており、学習状況を具体的に示しているとは言えない。

このように、様々なアプローチや想定された学習環境で学生の最終学習成果の予測や検出を試みる研究は数多ある。しかし、学習状況をより具体的に明らかにするというところまでは行っていない。このギャップを埋めるために、本研究では学習状況をより明らかにすることを目的とする。

3 提案内容

関連研究の節で述べたように、最終学習成果を高精度で予測できているが、具体的な学習状況を検出することはできていない。そのため、本研究は動画閲覧履歴を利用した学習の理解状況の可視化と分析の手法の提案を目指す。

しかしながら、動画を利用した学習の仕方は多様性が高く、学習方法の対象が曖昧である。つまり、高校や大学の授業として動画を利用した学習をするのか、予習・復習・補習の位置付けなのか、それとも学校での授業とは関係なく個人の興味やスキルアップのための自発的な学習であるのかによっても動画の利用目的、利用時間、モチベーションの高さや維持期間なども異なってくる。そのため我々は動画を利用する学習方法の対象を、大学の講義を対象とした「自由進度学習」と設定する。

自由進度学習とは、自己調整学習 [6] を実現する一種の学習法であり、教員が設計した学習内容のフレーム内で学習者が自ら計画を立てて学習を主体的に進めることを目指すものである。筑波技術大学における産業情報学科の専門選択科目「データベース論・演習」では自由進度学習を採用しているため、本科目を対象に動画閲覧データからの理解度抽出方法の分析を試みるものとする。

この科目で採用している自由進度学習では、学習者は決まった時間・場所（オンライン会場も含む）で学習を行うため、学習にかける時間を絞り込むことができる。対象とする学習者層を「情報科学を学ぶための基礎が形成されている者」、学習期間は４ヶ月間となる。また、授業の一環で動画閲覧を取り入れているためモチベーションの高さや集中力などもある程度統一することとなった。

また動画閲覧の際には、講義内容の文字起こしと並行して閲覧することができるアプリケーションを開発し [7]、これを用いて動画閲覧履歴を取得している。このシステムでは図 1 のスクリーン画面から分かるように、講義の文字起こし文章が右に並び、これらを読み、その文章に対応している動画部分にスキップすることができる。書籍のように特定の場所だけを参照する目的で部分的に閲覧することもでき、動画閲覧のアクションに意味を持たせることができる。

本節では、以下、学習環境、学習フェーズ、データセット、自作システム、仮説、検証手法について述べる。

3.1 学習環境

本研究では、筑波技術大学で開講されている『データベース論・演習』で用意された自己学習環境で実験を行った。その講義では、講義動画 1 本と演習課題 1 つで 1 個の学習単元として、19 個用意されている。学習者はそれぞれのペースで用意された学習単元を学習する。本講義で用意されている学習単元は以下のとおりである。

（1） （基本）リレーションの構造
（2） （基本）1 つのリレーションに対する問合せ
（3） （基本）リレーションの構造に関する制約
（4） （基本）外部キー
（5） （基本）外部キーを使った結合演算
（6） （基本）結合のいろいろな使い道
（7） 様々な結合
（8） （基本）集約演算
（9） 入れ子問い合わせ
（10） 左外部結合
（11） DB プログラミング
（12） WebDB プログラミング（WebAPI の仕組み）
（13） WebDB プログラミング（WebAPI と DB の接続）
（14） （基本）更新
（15） （基本）修正時異状
（16） （基本）関数従属と主キー
（17） （基本）正規化
（18） シェルと DB クライアント
（19） 関係代数

学習者は、自身のペースに合わせてそれぞれの学習単元を学習する。なお、順番通りに学習することが望ましいが順番通りでなくても構わない。最低限学習する内容として（基本）とついた 11 単元を優先的に学習することとし、それ以外は学習者の興味や状況に合わせ担当教員との調整の上、学習単元を決定する。

3.2 自由進度学習

自由進度学習は、学習者がそれぞれのペースで学習をする。学習到達目標やシラバス、授業資料、講義動画など、学生が学ぶためのリソースは全て学生にあらかじめ提示した上で、学習者自身が学びの目標と計画を立てて自律的に学習を進める。また、完全な自習形式とは異なり、担当教員が設計した学習フレーム内での学習となる。一般に各学習者の個別最適な学習と同時に学習者同士あるいは学習者と教員による協働的に学習を

進めるための仕組みも組み合わせていく。具体的に提供する仕組みについては教員による学習方針によるが、教員はティーチングを動画や資料などに任せ、主に学習者の学びの環境整備と学習者のコーチングという役回りに注力する。

本研究で対象とする授業では、学習科学による理論をベースに以下の流れによる学習を設計・実施している。

- **学習単元に対する流れ：学習・演習・レビュー・リフレクション：** 各学習単元に対し、事前に全ての学習資料（スライド、講義動画）と演習問題をセットで提供し、学習者は自ら立てた計画に従って学習資料を元に学習を行う。学習では、資料等による知識の伝達だけではなく、外化によって知識構成を促す仕組みを取り入れている。各学習項目に対して演習問題があり、演習問題を終了後、他の学習者を一人レビュアーにしてレビューをする。レビューでは回答内容を他者が理解できるように解説する。レビュアーはその学習項目に取り組んでいないものが担当しても構わないが、その場合はレビュアーが理解できるよう学習内容も含めて解説をする必要がある。クロスレビューにおいてレビュアーが内容を理解でき、それが正しいと判断したら、学習者は解答例を受け取り自ら答え合わせを行う。答え合わせの結果として、点数ではなくリフレクションフォームを提出する。そのフォームの中では、自分の回答と異なっていた問題と異なっていた原因、改めて学んだ箇所、なるほどと思ったこと、まだ納得できていない箇所など、省察した結果を文章として自己説明すると共に、学習内容の理解度を 10 段階で自己評価する。
- **授業時間の流れ：計画・学習・振り返り・フォローアップ：** 授業時間は、学習計画に続いて学習時間を約 1 時間とり、その後振り返りを行う。振り返りでは「もやもやトーク」と称し、隣に座っている人と今もやもやと考えていることを他人と共有し合うことで思考の言語化を行い、その後、現在の取り組み、腑に落ちたこと、理解が漠然としていることの 3 点についてフォーム提出する。教員は学習中の質問とその回答内容を随時資料として共有すると共に、提出された振り返り内容から漠然とした理解部分を取り上げ、次の時間に「フォローアップ」として解説する。

3.3 学習フェーズ

我々は、先行研究 [7] において、動画を閲覧する際の学習フェーズを「learn」「note」「exercise」の 3 つに定義した。それぞれの説明は以下の表 1 に示す。

表 1　学習フェーズとその意味

学習フェーズ	意味
learn	講義動画の視聴やスライドや教科書などの参考資料を読むなどの学習行為で知識のインプットを行うフェーズ
note	インプットした知識をノートなどで自分なりにまとめることで体系的に理解を試みるフェーズ
exercise	体系的に理解した知識を活用して演習課題に取り組むフェーズ

これらは一方通行ではなく、受講生は自由に 3 つの学習フェーズを行き来することを想定している。特に本研究が対象としている学習形式では、学習者はレビューによる外化や、自己説明・自己評価による省察によって何度も知識の構成を行う。また、動画を閲覧したい箇所だけ部分的に閲覧することもできる。そのため、learn フェーズだけではなく、note や exercise 時の部分閲覧も行われると推測される。

3.4 講義動画視聴システム

本研究で用いる講義動画視聴システム「Youtube Study Log」について述べる。以下の図 1 は Youtube Study Log のスクリーンショットである。

図 1　動画視聴システムのスナップショット

左上が講義動画となっている。その下にシークバーと字幕が存在する。さらに、ユーザ名入力欄とフェーズ選択肢、再生ボタン、停止ボタンがある。右は文字起こしデータ一覧であり、該当時間をクリックすると該当時間のシーンにいくことができる。

視聴を始めたいときは、ユーザ名に学習者自身の名前を入力し、該当するフェーズを選択する。選択すると再生ができるようになり、再生ボタンを押すと動画が再生される。停止したいときは、停止ボタンを押す。また、任意のシーンへ戻すか飛ばしたいときは、シークバー上で任意のシーンがある場所にクリックするか、シークバー上で任意のシーンまでドラッグ&ドロップする。これらの操作は画面左下にて表示され、csv ファイルにも保存される。

本システムでは動画は Youtube に掲載されているコンテンツを利用することを想定している。現時点では音声認識システムにて自動文字起こしされたデータは編集なしでは十分に内容を理解するまでの精度には至っていない。そのため本システムでは一度文字起こしデータを手動で修正を行ったものを利用している。

3.5 データセット

筑波技術大学で開講されている『データベース論・演習』は産業情報学科の専門選択科目の 1 つであり、2023 年度は 2023 年 10 月 4 日から 2024 年 2 月 1 日までの間で 15 回実施された。データの収集は筑波技術大学研究倫理委員会の承認を得て、学習者のうち趣旨に賛同したものの閲覧履歴のみ収集した。研究対象となった受講生数は 17 名であった。

2273 個のクリックストリームデータが収集されている。収集

されたクリックストリームデータの詳細は以下の表 2 にて示す。

表 2　収集されたクリックストリームデータ

動画数 (本)	19
平均動画閲覧時間 (秒)	1082.0556
総クリックストリーム個数	2273
learn_play の回数	816
learn_pause の回数	744
learn_seek の回数	673
note_play の回数	6
note_pause の回数	5
note_seek の回数	5
exercise_play の回数	10
exercise_pause の回数	7
exercise_seek の回数	7

4　分析：動画視聴行動と理解度の関係

4.1　視聴行動ベクトルの作成

この分析の目的は、理解度に関わりの高い動画視聴アクションを発見することにある。そこで、まず各動画視聴に対し、視聴アクション play, pause, seek の回数によってベクトルを作成した。3 段階の学習過程（learn, note, exercise）でのそれぞれのアクションに対して、5 秒以内, 5 秒から 30 秒, 30 秒から 1 分, 1 分から 5 分, 5 分から 10 分, 10 分以上の時間に分け、該当する回数を値とした。seek に関しては、シークした長さとし、正の方向（早送り）のシークに加えて、負の方向（早戻し）も設定した。理解度のスコアも加え、全体で 82 次元のベクトルとなった。

4.2　分析の方針

分析にあたり、上記で作成した動画視聴行動ベクトルのうち、どの行動が特徴的であるか、また理解度に関連するかを探索する。そこで我々は手始めとして以下の 2 つの手法で分析することとした。

- 主成分分析を用い、多くのアクションのうち特徴的な行動は何かを分析する
- 決定木を用い、理解度の値に関連するアクションは何かを分析する

4.3　主成分分析

まず、主成分ごとの寄与率と累積寄与率を表 3 に示す。ここでは、データを十分に説明できていると言われる累積寄与率 0.8 に到達する第 6 主成分までを示している。

表 3　第 6 主成分までの寄与率と累積寄与率

第 n 主成分	寄与率	累積寄与率
第 1 主成分	0.235	0.235
第 2 主成分	0.193	0.428
第 3 主成分	0.173	0.601
第 4 主成分	0.088	0.689
第 5 主成分	0.063	0.752
第 6 主成分	0.049	0.801

第 1 主成分は、データの分散が最大となるような軸をとる特徴量で構成される。その第 1 主成分が占める情報量の割合である寄与率が 0.235 と、理解度データに対して十分な情報量を占めているとはいえない。また、データを十分に説明できる累積寄与率 0.8 に到達するのに 6 個の主成分が必要である。つまり、どの主成分でも理解度データに対して十分な情報量を持っているとはいえないことを示唆している。

次に、第 1 主成分と第 2 主成分を構成する特徴量の固有ベクトルの値を表 4,5 に示す。ここでは、絶対値で降順に並び替えて上位 10 位までを示している。

表 4　第 1 主成分を構成する特徴量の固有ベクトル

特徴量	固有ベクトル
note_pause	0.271
note_start	0.271
note_seek	0.271
note_play_lt_5sec	0.271
note_play_30sec_60sec	0.271
note_play_1min_5min	0.271
note_pause_lt_5sec	0.271
note_pause_1min_5min	0.271
note_pause_gt_10min	0.271
note_seek_lt_5sec	0.271

表 5　第 2 主成分を構成する特徴量の固有ベクトル

特徴量	固有ベクトル
learn_start	-0.226
learn_seek	-0.226
learn_pause	-0.226
learn_play_lt_5sec	-0.214
exercise_start	0.198
learn_pause_gt_10min	-0.197
learn_seek_30sec_1min	-0.196
exercise_pause	0.196
exercise_seek	0.196
learn_pause_30sec_1min	-0.195

これらの表 4,5 から、第 1 主成分は note フェーズにおけるクリックストリーム、第 2 主成分は learn と exercise フェーズにおけるクリックストリームで構成されていると言える。

第 1 主成分と第 2 主成分において、理解度 7 以下と理解度 8 以上の散布図を図 2 に示す。縦軸は第 2 主成分の主成分得点、横軸は第 1 主成分の主成分得点、青色の点は理解度 8 以上、赤

色の点は理解度が 7 以下を示している。理解度 8 以上の個数は 66 個、理解度 7 以下の個数は 20 個である。

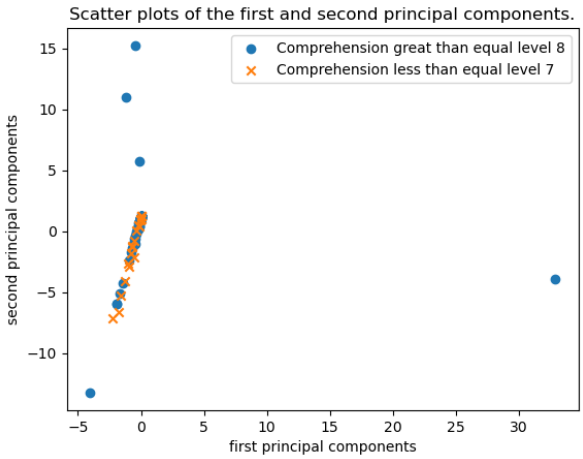


図 2　第 1 主成分と第 2 主成分の 2 次元で示した理解度の散布図

この図から 3 点読み取れることがある。

1 点目は、第 1 主成分では数点の外れ値が大きく影響している点である。実際 note のために動画を視聴していたのは 1 名だけであり、他の受講者は全く note のために動画を視聴していなかった。そのため第 1 主成分は全体の傾向を表す軸とはなり得ないことがわかる。 2 点目は第 2 主成分において分散されており、理解度が 7 以下のデータが-5 から 2 付近に集中している一方で、理解度が 8 以上のデータはいくつか外れた場所に位置していることである。また第 2 主成分は表 5 に示すように learn と exercise のアクションの混合で構成されているが、固有ベクトルにおいて learn 関連の特徴量は負の値を持ち、exercise 関連の特徴は正の値を持つ。第 2 主成分を構成する特徴量の関連について詳細に調査する必要があると言える。 3 点目は、第 2 主成分の主成分の値が-5 から 5 の間に収まっているという点である。主成分得点は各データがどの主成分を強く持つのかを示し、絶対値が小さい場合主成分を強く持たない。今回の図 2 では、ほとんどが-5 から 5 の間に収まっており、2 点目に示した興味深い点もあるが、全体に対する影響は小さいことも懸念される。

いずれにしてもこの節の冒頭に述べた通り寄与率が低いためこれらの分析結果は全体の傾向を十分に反映しているとは言い難い。

4.4　決定木分析

決定木分析を 10 回繰り返したときの学習データの正解率とテストデータの正解率の平均を表 6 にて示す。

表 6　学習データとテストデータの平均正解率

学習データの平均正解率	0.984
テストデータの平均正解率	0.709

テストデータの平均正解率は 0.709 と、精度高く理解度 8 以上グループと理解度 7 以下グループを分類できたとはいえない。

次に、決定木分析をする際に分岐基準となった特徴量を可視化した図 3 と x[n] に対応する特徴量の表 7 を示す。なお、理解度 8 以上グループのサンプル数は 52 個、理解度 7 以下グループのサンプル数は 12 個である。

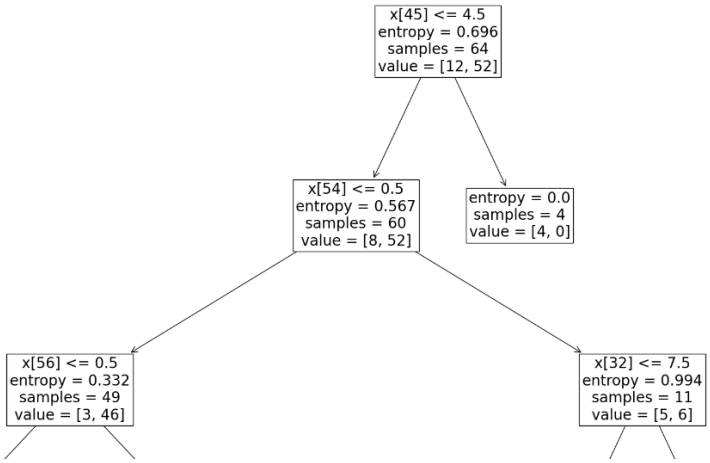


図 3　決定木の構成が可視化された図

表 7　x[n] に対応する特徴量

x[n]	対応する特徴量
x[45]	learn_seek_lt_5sec
x[54]	learn_seek_m1min_m5min
x[56]	learn_seek_lt_m10min
x[32]	learn_pause_gt_10min

最初の分岐では、learn_seek_lt_5sec ＜＝ 4.5（learn フェーズの視聴で 5 秒以内の箇所への早送り回数が 4.5 回以下） を条件として理解度 7 以下グループと理解度 8 以上グループを分類している。ここで、learn_seek_lt_5sec ＜＝ 4.5 ではなかった（learn フェーズの試聴で 5 秒以内の箇所への早送り回数が 4.5 回以上であった）ものは、理解度 7 以下グループからの 4 個のみだった。次に learn_seek_m1min_m5min ＜＝ 0.5（learn フェーズの視聴で 1 分前から 5 分前への早戻し回数が 0.5 回以下）を条件として、learn_seek_m1min_m5min ＜＝ 0.5 でなかった（learn フェーズの視聴で 1 分前から 5 分前への早戻し回数が 0.5 回以上であった）ものは、理解度 7 以下グループから抽出された 5 個と理解度 8 以上グループから抽出された 6 個であった。この 2 つの分岐で、理解度 7 以下グループは 12 個中 9 個抽出されたことになる。このことから、learn フェーズにおいて 5 回以上任意のシーンを細かく探し求めている行為と learn フェーズにおいて 1 回以上 1 分前以上 5 分前以下のシーンまで戻している行為が見られた場合は、理解度が 7 以下になる可能性があるといえる。実際に、決定木分析で重要視された特徴量を棒グラフで示した図 4 からでも、上位 2 位が learn_seek_lt_5sec と learn_seek_m1min_m5min であることを確認できる。ただし、理解度 7 以下グループのサンプル数が 12 個と少ないため、さらなる検証が必要であることも併せて述べる。

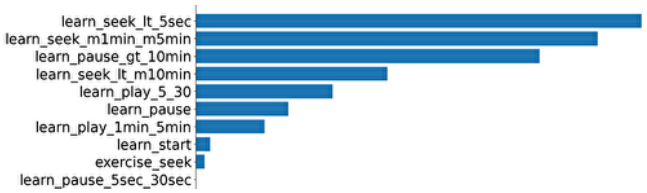


図 4　決定木分析で重要視された特徴量

4.5　基本統計量

主成分分析と決定木分析において、望ましい結果を得られなかった。その原因としてデータセットにあると考えたため、理解度 8 以上と理解度 7 以下に分けてデータセットの基本統計量を確認してみた。まず、learn フェーズにおける基本統計量を表 8, 9 にて示す。

表 8　理解度 8 以上グループの learn フェーズにおける基本統計量

	learn_start	learn_pause	learn_seek
総数	66	66	66
平均	7.970	7.121	6.136
標準偏差	13.938	13.765	13.655
最小値	0	0	0
第一四分位	2	1	0
中央値	3	2	1
第三四分位	8.75	7	4
最大値	74	73	72

表 9　理解度 7 以下グループの learn フェーズにおける基本統計量

	learn_start	learn_pause	learn_seek
総数	20	20	20
平均	14.5	13.7	13.4
標準偏差	16.198	16.095	16.158
最小値	1	0	0
第一四分位	2.75	2	1
中央値	10	9.5	8.5
第三四分位	19.5	18.5	18.5
最大値	66	64	64

表 8, 表 9 を見ると、理解度 8 以上グループと理解度 7 以下グループの間で learn_start, learn_pause, learn_seek 全ての平均に差があると考えられる。ただし、learn フェーズにおけるクリックストリームデータは図 5 で示す通り正規分布に従っていなかったため、パラメトリック検定による有意差検定はできなかった。なお、横軸は一つの動画に対する play の回数、縦軸はその回数の総和を示したものである。

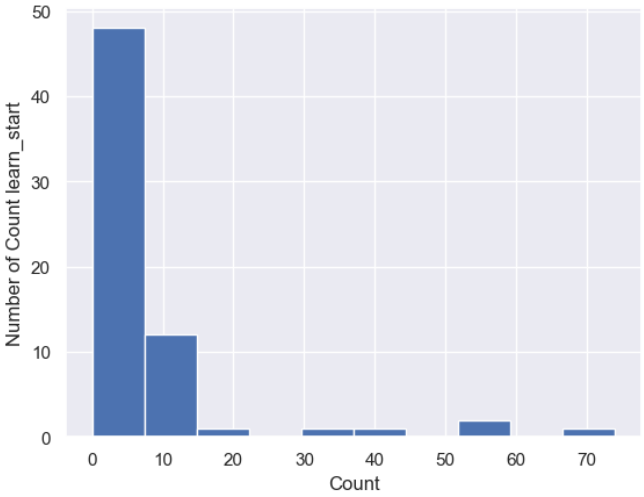


図 5　理解度 8 以上グループの learn_start のヒストグラム

次に note フェーズにおける基本統計量を表 10, 11 にて示す。

表 10　理解度 8 以上グループの note フェーズにおける基本統計量

	note_start	note_pause	note_seek
総数	66	66	66
平均	0.091	0.076	0.076
標準偏差	0.739	0.615	0.615
最小値	0	0	0
第一四分位	0	0	0
中央値	0	0	0
第三四分位	0	0	0
最大値	6	5	5

表 11　理解度 7 以下グループの note フェーズにおける基本統計量

	note_start	note_pause	note_seek
総数	20	20	20
平均	0	0	0
標準偏差	0	0	0
最小値	0	0	0
第一四分位	0	0	0
中央値	0	0	0
第三四分位	0	0	0
最大値	0	0	0

表 8, 表 9 を見ると、理解度 8 以上グループと理解度 7 以下グループともにクリックストリームデータが極端に少ないことが分かる。実際、理解度 8 以上グループと理解度 7 以下グループでの note フェーズにおけるクリックストリームデータ数は 16 個であった。その 16 個は、学習者番号 13 が第 3 回動画を閲覧した際に記録されたものである。逆に言えば、それ以外の学習者および学習単元において note フェーズで動画を閲覧したデータは記録されていないということである。

最後に、exercise フェーズにおける基本統計量を表 12, 13 にて示す。

表 12　理解度 8 以上グループの exercise フェーズにおける基本統計量

	exercise_start	exercise_pause	exercise_seek
総数	66	66	66
平均	0.152	0.106	0.106
標準偏差	0.749	0.530	0.530
最小値	0	0	0
第一四分位	0	0	0
中央値	0	0	0
第三四分位	0	0	0
最大値	5	3	3

表 13　理解度 7 以下グループの exercise フェーズにおける基本統計量

	exercise_start	exercise_pause	exercise_seek
総数	20	20	20
平均	0	0	0
標準偏差	0	0	0
最小値	0	0	0
第一四分位	0	0	0
中央値	0	0	0
第三四分位	0	0	0
最大値	0	0	0

exercise フェーズのクリックストリームデータも note フェーズのクリックストリームデータと同様に極端に少ないことが分かる。実際、exercise フェーズのクリックストリームデータ数は 24 個と learn フェーズのクリックストリームデータ数 2233 個より極端に少ない。

この分析結果から、本データセットには、note フェーズと exercise フェーズにおけるクリックストリームデータ数が少ないという問題点があると考えられる。learn フェーズと note フェーズ,exercise フェーズでクリックストリームデータ数に差が大きいことで、主成分分析と決定木分析から note フェーズと exercise 理解度に関わりの高い特徴量を抽出することができなかった。note フェーズと exercise フェーズで十分なクリックストリームデータを収集できなかった理由として、note フェーズと exercise フェーズでは講義動画よりも講義資料スライドや web ページなどを閲覧した学習者が多かったことにあると考えられる。今回、設定した学習環境では、どのフェーズにおいても講義動画の他に講義資料スライドや web ページなどを自由に閲覧できるようになっている。learn フェーズで講義動画を閲覧し、理解できなかった部分を講義資料スライドや web ページを閲覧した学習者が多かったと推測される。そのため、note フェーズと exercise フェーズにて十分なクリックストリームデータを収集できなかった。

5　結論・今後の課題

本稿では、学習状況をより明らかにすることを目的に主成分分析や決定木分析から理解度に関係する動画視聴行動ベクトルを抽出する試みを行った。しかし、理解度に関係する動画視聴行動ベクトルは見つからなかった。

今後の課題として、note フェーズと exercise フェーズでより多くの動画視聴行動ベクトルを収集できるようにするという課題がある。大学の講義を対象とした自由進度学習環境のもとで、学習者は講義動画のみならず講義資料スライドや Web ページなども閲覧しながら学習を進めることを一般的として想定する必要がある。そのため、講義動画だけではなく講義資料スライドや Web ページからもクリックストリームデータを収集する必要がある。

また、決定木分析から理解度 7 以下グループを抽出しうると示唆された特徴量である learn_seek_lt_5sec と learn_seek_m1min_m5min について、より多くのサンプル数を集めた上で、理解度との関係を再検証するという課題もある。今回の理解度 7 以下グループのサンプル数は 12 個と少なく、本研究の分析結果をもって理解度との関係があると断言することはできない。そのため、さらに多くのサンプル数を収集して本当に理解度 7 以下グループを抽出しうる特徴量であるかどうかを再検証する必要がある。

6　謝　　辞

本研究は JSPS 科研費 JP19K11978 の助成を受けたものです。また、本実験の実施にあたり、データベース論・演習の受講生 17 名とその教員に協力していただきました。ここに厚く御礼を申し上げます。

文　　献

[1] Chen-Hsiang Yu, Jungpin Wu and An-Chi Liu “Predicting Learning Outcomes with MOOC Clickstreams” Educational Innovation 2019, 2019.

[2] Yutong Liu, Si Fan, Shuxiang Xu, Atul Sajjanhar, Soonja Yeom and Yuchen Wei “Predicting Student Performance Using Clickstream Data and Machine Learning” Education Sciences., Volume 13, Issue 1, 2022.

[3] Nihal Abuzinadah, Muhammad Umer, Abid Ishaq, Abdullah Al Hejaili, Shtwai Alsubai, Ala’ Abdulmajid Eshmawi, Abdullah Mohamed and Imran Ashraf “Role of convolutional features and machine learning for predicting student academic performance from MOODLE data” PLoS ONE 18(11): e0293061, 2023.

[4] Hajra Waheed, Raheel Nawaz, Dragan Gasevic “Early prediction of learners at risk in self-paced education: A neural network approach” Expert Systems with Applications, 2022

[5] Christopher G. Brinton and Mung Chiang “MOOC Performance Prediction via Clickstream Data and Social Learning Networks” 2015 IEEE Conference on Computer Communications, 2015

[6] Barry J. Zimmerman “Attaining Self-Regulation: A Social Cognitive Perspective”, Handbook of Self-Regulation, pp.13-39, 2000.

[7] 齋藤光貴, 渡辺知恵美 “動画を活用した自己学習環境における文字起こしデータと視聴履歴からの自動理解度検出に向けて” 研究報告データベースシステム（DBS）, 2022-DBS-176, 1, pp.1-6, 2022. http://id.nii.ac.jp/1001/00222849/

[8] Ozan Raşit Yürüm, Tuğba Taşkaya-Temizel and Soner Yıldırım “The use of video clickstream data to predict university students’ test performance: A comprehensive educational data mining approach” Education and Information Technologies (2023), Volume28, pp.5209–5240, 2022.

[9] Ozan Raşit Yürüm, Tuğba Taşkaya-Temizel and Soner Yıldırım “Predictive Video Analytics in Online Courses: A Systematic Literature Review” A Systematic Literature Review. Tech Know Learn (2023), 2023.
[10] H.M. Rafi Hasan, Mohammad Touhidul Islam, AKM Shahariar Azad Rabby and Syed Akhter Hossain “Machine Learning Algorithm for Student’s Performance Prediction” 10th International Conference on Computing, Communication and Networking Technologies, 2019.

教材ビデオにおける学習者の視聴操作に基づくシーン選択支援

小西　真愛† 井上　沙紀† 王　元元†† 河合由起子†††,†††† 角谷　和俊†

† 関西学院大学総合政策学部　〒 669–1330 兵庫県三田市学園上ケ原 1 番
†† 山口大学大学院創成科学研究科　〒 755-8611 山口県宇部市常盤台２丁目１６番１号
††† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†††† 大阪大学サイバーメディアセンター　〒 567-0047 大阪府茨木市美穂ケ丘５番１号
E-mail: †{gpv63912,hms54907,sumiya}@kwansei.ac.jp, ††y.wang@yamaguchi-u.ac.jp, †††kawai@cc.kyoto-su.ac.jp

あらまし　現在, 大学のオンデマンド授業に加え, 受験・TOEIC の勉強が進められるスタディサプリや社会人に必要なスキルや資格取得に向けた勉強が進められる Udemy Business といった映像コンテンツとアプリ学習が中心のオンライン学習サービスも多く普及されている. 制約なしに都合のいい時間に学習が進められるが, 個人で学習を進めなければならず自己責任感が強い. そのため適切な学習ができていない学習者が多い. そこで本研究ではスライドとクローズドキャプションの内容から細かくシーン選択できるようなサポートを行う. その中で視聴操作を行った場合その意図に応じて提示するシーンを都度変更する. これにより, 学習者が視聴するべきシーンを探す必要が無くなり, 効率的に学習することができるようになる. 本論文では視聴操作とシーンとの関係について主に検討し, 視聴操作の意図からどのようなシーンを提示すべきかについても検討している.

キーワード　オンデマンド講義, e-Learning, 視聴操作, シーン, 学習支援

1 はじめに

新型コロナウイルスの影響により, 多くの学校においてオンデマンド授業が普及された. それに加え, 小・中・高校生向けに学校の予習や復習, 受験勉強が行えたり, 英語学習者向けに TOEIC やビジネス英語, 日常英語が学べたりするスタディサプリ [1] や, プログラミングスキルやビジネススキルなど社会人に役立つ情報がオンラインで学べる Udemy [2] といったサービスも多く利用されている. 田中 [8] によるとオンラインでの授業形態や授業サービスは動画を止めて見直しや理解を深めることができることや, 自分のペースで学習を進めることができるというメリットがあるとしている. しかし, 自分のペースで学ぶからこそ自己責任感が強いというデメリットがある. そのため適切に学習が進められていない学習者が多く存在すると考えられる. 岡田 [4] によると, オンライン学習では特に集中力・緊張感の保持の仕方や, 一方的な伝達型講義を改善する必要があると述べている.

そこで, 本研究では学習者が教材ビデオを視聴中の視聴操作から意図を抽出し, スライドの内容から学習者が視聴するべきシーンの選択支援を行う. そうすることで, 講義動画を流したままになることなくメリハリをつけて視聴することができ, 集中を保つことができるようになると考える. また, 視聴するべきシーンを学習者が探す必要がなくなるため, 効率的に学習を進められるようになる. 本稿では視聴操作の意図と, シーン同士の関係についてどのような関係にあるか, どのようなシーン関係にあるものを推薦するべきかについて検討を行った.

2 関連研究

2.1 シーン選択履歴による学習支援

豊崎ら [9] は, 講義初回視聴時の学生ユーザのシーン選択履歴を用いて, 学生ユーザの理解度に基づいた補助資料を提示する手法を検討していた. シーン選択履歴とは, あるシーンに対して初回講義視聴時に, シーン選択を 1. 一度のみ視聴, 2. 早送りで視聴, 3. まだ視聴していない, 4. 複数回視聴の 4 つに分けたものを指している. 学生ユーザにとって重要なシーンの検討を行い, 学生ユーザにとって重要なシーンの検討を行っていた. ここでいう補助資料とは, 学生ユーザの理解が不足している部分を補う Web ページや講義のスライド, 講義のシーンなどを指している. これらを用いることで, 学生ユーザが講義の内容をより理解できるように支援しようと考えた. 豊崎らは, シーン選択履歴を用いて自動で学生ユーザにとって有効な補助資料の提示を行おうとしたが, 本研究は学習者の教材ビデオ視聴中の視聴操作を用いてシーンの選択を支援するため, この点で異なる.

2.2 板書の協調を生かした復習用コンテンツ自動作成システム

井上ら [3] は, 講師のノートを基準にしつつも学生の理解度に応じて内容や進行の仕方が変化したり, 見出しや内容の表現方法や黒板の時間的・空間的な使い方が教師によって異なることが板書講義における現状であるとし, 講師に特別な負担をかけることなく復習視聴のための板書講義コンテンツを自動生成するシステムを提案している. 本研究ではすでに作成された動画授業における学習支援を目的としているため, この点で異なる.

2.3 理解支援関係

高橋ら [6] は, 必要な教科書を買うことができない発展途上国

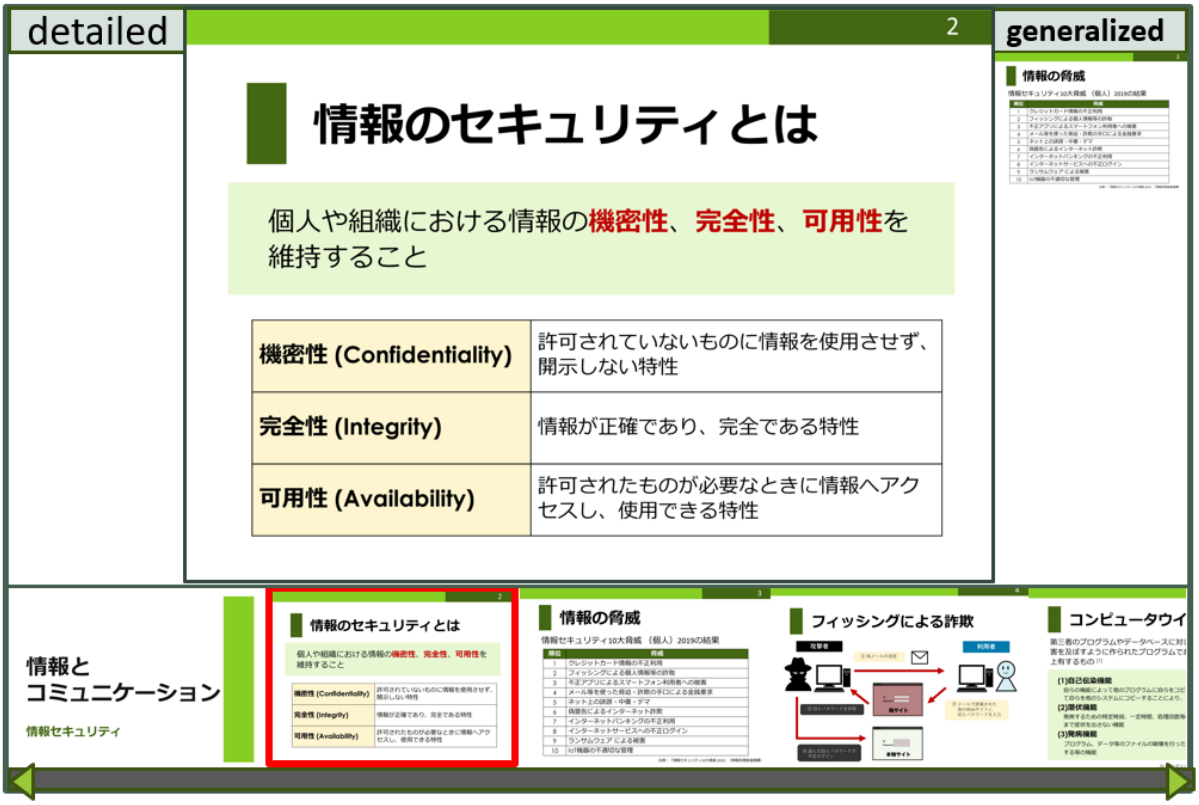


図 1　想定する UI イメージ

に向けてや, 次々と現れる新たな学習テーマについて, インターネット上にあるさまざまな教材コンテンツを手掛かりに初学者のための学習項目を関連付けた理解支援関係グラフを自動構築する問題に取り組んだ. 理解支援関係とは, 学習するあるトピックの知識との間にある知識を理解するためにはこの知識を理解しておく必要があるということが成立する関係である. この関係を用いることにより, すでに理解している知識をもとに次に学習する必要のある知識の推薦や先に学習すべき知識の推薦をおこなうことができるとしている. 本研究では知識の推薦ではなく, 教材ビデオのシーンの推薦を行おうとしており, この点で異なる.

2.4　オンライン手書きデータからの情報量変更

浅井ら [7] は, タブレットなどの端末上で紙とペン同様の手書き入力におけるデータから, 表示する文字や図の取捨選択・拡大を行うことで可読性の向上と情報量の削減を行い, ユーザが容易に元データの概要を把握できる要約サムネイルの生成手法を提案した. 下線や囲い込み, 色変更といった強調表現を用いて検出を行っていたが, 本研究では学習者の視聴操作の意図と講義動画のシーン内容を加味した支援となっており, この点で異なる.

3　ユーザの視聴操作とスライド関係

3.1　ユーザの視聴操作

本研究で定義している視聴操作は以下の通りである.

（1）　スライド選択

（2）　早送り

（3）　一時停止

（4）　巻き戻し

a）　スライド選択

スライド選択という視聴操作はユーザインタフェースを基にしたものである. ユーザは現在視聴しているシーン（スライド）からユーザ自身で視聴したいシーンを下のスライド群から選択することができる. そのユーザ自身で視聴シーンを選択する操作を本研究では視聴操作と定義している. このとき, 視聴しているシーンよりも先のシーンに移動した場合は早送りを, 前のシーンに移動した場合は巻き戻ししたと判定し, スライドの関係も考慮しながらシーン選択支援を行う.

b）　早　送　り

ユーザが現在見ているシーンより後ろに位置するシーンへ動画を進める操作のことを指す. 本研究では, 時間短縮や効率化のための操作と考えており, n 倍速視聴 ($n > 1.00$) も早送りの一種であると定義している.

c）　一 時 停 止

一時停止とは, ユーザが現在閲覧しているシーンを音声, 映像共にそのシーンで停止させる操作のことを指す. 本研究では, じっくりとシーンを視聴したり, 授業ノートをとるための操作と考えており, n 倍速視聴 ($n < 1.00$) も一時停止の一種であると定義している.

d）　巻 き 戻 し

巻き戻しとは, ユーザが現在閲覧しているシーンからすでに視聴済みのシーンへ戻る操作のことを指す. 本研究では, 前提

内容を知ったり前の段階に戻るための操作と考えている.

3.2 スライド関係

本研究では映像コンテンツ視聴中のユーザのシーン選択を支援するためにスライド間の関係について検討した. いくつものスライド関係があると考えられるが, 本研究では主に詳細化（detailed）・汎化（generalized）について検討した. 先行研究として王ら [5] は, スライドの表層的構造と概念的構造を利用したスライド関係の判定手法を提案している. 本研究でもこの判定手法を用いている.

3.2.1 詳細化（detailed）

詳細化（detailed）とは, 対象のスライドが現スライドを詳細に説明している関係のことを指す. 詳細説明は, ユーザが映像コンテンツの内容を理解しているとき, 追加的に内容の理解を促進できる説明であると考えられる. そのため詳細関係はシーンを先に飛ばすような早送りの操作時に有効な支援になると考えられる. 図 2 の場合, 左スライドの表の中にある「フィッシング」というキーワードが右スライドにはタイトルにあり, スライド全体がフィッシングの内容について述べられていると考えられるため, この関係は詳細化の関係であると考えられる.

図 2 詳細化の例

3.2.2 汎化（generalized）

汎化（generalized）とは, 対象のスライドが現スライドを一般化して説明している関係のことを指す. 汎化説明は, ユーザが映像コンテンツの内容を理解できていないとき, 前段階の理解を促進する説明であると考えられる. そのため汎化の関係はシーンを前に戻すような巻き戻しの操作時に有効な支援になると考えられる. 図 3 の場合, 左スライドのタイトルにある「ウイルス感染」というキーワードが右ページは表の中の一部にある. 右ページではウイルス感染についての説明は情報セキュリティ対策の説明の中で行われているため一般化されていると考えられ, 汎化の関係であると考えられる.

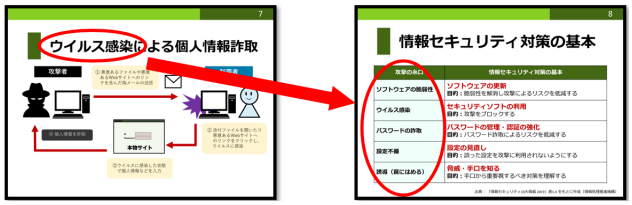


図 3 汎 化 の 例

4 シーン選択支援手法

本研究でのシーン選択支援を行うための手法を説明する. まず, あるひとつの映像コンテンツについて, 全てのスライド間で detailed 関係と generalized 関係にあるか検討する. ここでは, 大阪大学の講義である「情報社会基礎・情報科学基礎」の中の情報とコミュニケーション内の情報セキュリティについての動画とスライドにおいて検討した.

4.1 映像コンテンツのスライド間関係判定

対象の映像コンテンツにてまず最初にスライド間の関係を判定する. スライド間の関係とは前の章で述べた detailed 関係・generalized 関係のことである. 判定手法については 3 章で述べているものと同じ手法で行っている.

4.2 スライド間関係の最短距離

スライド間関係の算出が完了すれば, 次にユーザがスライド間の移動をした場合その関係をたどれる最短距離を算出する. この際, detailed 関係・generalized 関係は関係なくたどるものとする. 最短でたどれるものだけを算出し, スライド間の移動の意味的な内容をたどれるのではないかと考えた.

4.3 推薦シーン決定

最後に推薦シーンの決定方法について説明する. 推薦シーンはユーザが早送りの操作をしていたか, 巻き戻しの操作をしていたかによって異なる. 早送りの操作を行っていた場合は, detailed 関係で結ばれているシーンを推薦し, 巻き戻しの操作を行っていた場合は, generalized 関係で結ばれているシーンを推薦する.

4.4 推薦シーンの例

図 4 において, 現視聴スライドからユーザが移動したスライドまでの最短距離を考える. 例えばユーザがスライド 7 からスライド 2 まで移動した際, 矢印をたどって移動するとすれば, 最短距離はスライド番号 7 → 6 → 3 → 2, またはスライド番号 7 → 8 → 3 → 2 となる.

最短距離でのたどり方を検討した後にその矢印が detailed 関係なのか generalized 関係なのかを確認する. 今回の例ではどちらのたどり方についてもスライド 3 までの矢印は generalized 関係, スライド 2 に移動するための矢印は detailed 関係であった.

最後にユーザの視聴操作が早送りなのか巻き戻しなのかを判定し, その操作に合わせたシーン選択支援を行う. 今回の例ではスライド 7 からスライド 2 への移動なので, 巻き戻しの操作であると考えられる. 巻き戻しの操作には generalized 関係の矢印でたどられているスライド 3・6・8 を推薦するのが良いと考えられる.

5 おわりに

本研究では, 教材ビデオを視聴する学習者への学習支援として, ビデオ視聴中の視聴操作を用いてシーンの選択サポートを提案した. シーン選択支援のために, スライド関係や視聴操作の

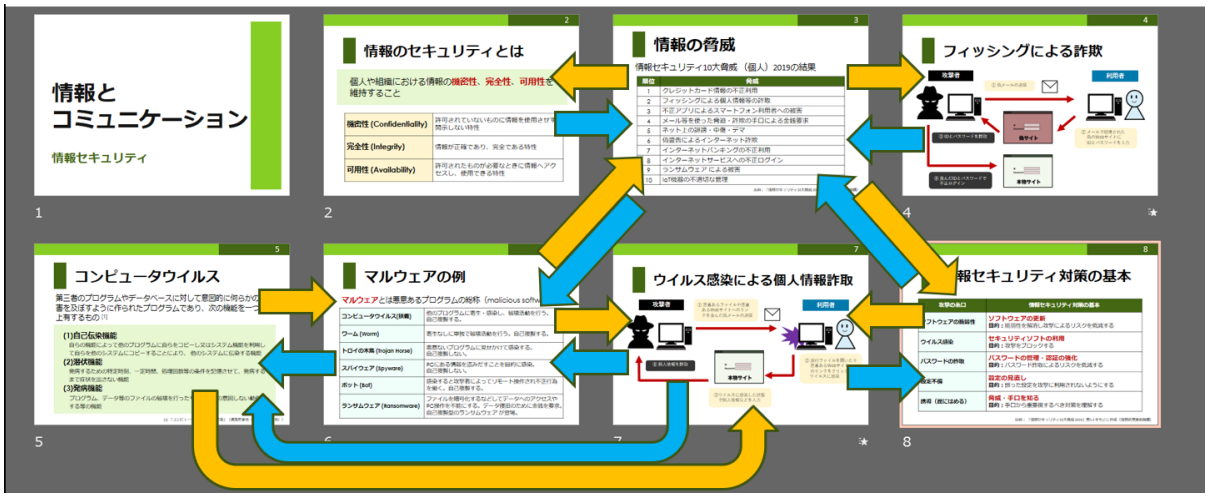


図 4　使用する映像コンテンツのスライド関係

意図について検討し, 提案する手法についての説明を行った.

今後の課題として, 提案手法に関する評価実験を行うことや, ユーザが操作しやすいユーザインタフェースの実装などが挙げられる. また, 今回はスライドの関係について detailed 関係と generalized 関係のみ検討したが, そのほかの関係も存在すると考えられるためそちらの検討も行うべきであると考えられる.

謝　　辞

本研究において大阪大学サイバーメディアセンター 竹村治雄教授・白井詩沙香講師, スチューデント・ライフサイクルサポートセンター 長瀧寛之教授から講義資料・講義ビデオの提供を頂きました. ここに記して謝意を表します.

文　　献

[1] https://studysapuri.jp.
[2] https://www.udemy.com/ja.
[3] 井上亮文, 品田良太, 市村哲, 星徹. 板書の意識的な強調を利用した 復習用コンテンツ自動生成システム. 情報処理学会論文誌, Vol. 53, No. 1, pp. 49–60,, 2012.
[4] 岡田佳子. 学生からみたオンライン授業のメリットとデメリット－オンライン環境下のアクティブラーニングに焦点を当てて－. 長崎大学 教育開発推進機構紀要 第 11 号, 2021.
[5] 王元元, 北山大輔, 角谷和俊. プレゼンテーションコンテンツのための概念構造に基づくスライドの関係判定方式. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010), 2010.
[6] 高橋健太郎, 伊藤寛祥, 松原正樹, 森嶋厚行. インターネット学習における学習支援ガイド生成に向けた理解支援関係の自動検出. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2023), 2023.
[7] 浅井洋樹, 山名早人. オンライン手書きノートからの強調語抽出. *DBSJ Journal,Vol.10, No.1*, 2011.
[8] 田中希穂. 大学におけるオンライン授業の実践と課題. 同志社大学教職課程年報, 2021.
[9] 豊崎祐衣, 王元元, 河合由起子, 角谷和俊. 関連情報探索のための番組映像の周辺語句の種類に基づくクエリ生成. 情報処理学会関西支部 支部大会, 2022.

ピクトグラミーティング
- ピクトグラムコンテンツの共有及び相互評価システム -

松村 貫司† 石井 幹大‡ 伊藤 一成†
† 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1
‡ 青山学院大学理工学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1

あらまし 情報デザインにおいては，情報伝達の受け手を意識することや，受け手によって実際に評価が行われることが重要である．今回，複数の利用者がピクトグラムコンテンツを共有できる機能,利用者が共有したピクトグラムコンテンツを相互に評価できる機能を備えたアプリケーション「ピクトグラミーティング (Pictogrameeting)」を提案する.共有対象とするピクトグラムコンテンツは，ピクトグラムに関する背景情報やメタ情報を含めて管理する.本論文では,「ピクトグラミーティング」の設計・実装方式および，システムの仕様・機能について解説する．
キーワード 情報デザイン，データ活用，コミュニケーション，ピクトグラム

1. はじめに

ピクトグラムとは日本語で絵記号，図記号と呼ばれるグラフィックシンボルであり，意味するものの形状を使ってその意味概念を理解させる記号である[1]．公的に用いるため規格に基づいて作成されるピクトグラムもあるが，個人や団体による特定の目的のために作成されるピクトグラムも存在する．今回，個人や団体が制作主体となったピクトグラムコンテンツに対して，それらの共有や評価を可能とするシステムを実装したので報告する．本論文では，設計・実装方式について説明する．以下 2 章で，ピクトグラムコンテンツとその記述について概説する．3 章で提案システムについて説明し，4 章で考察し，5 章でまとめる．

2. ピクトグラムコンテンツ

本章では，本システムが扱うピクトグラムコンテンツとしての対象範囲を定めるとともに，その範囲で扱われる記述形式について例示する．

2.1. ピクトグラムコンテンツの対象範囲

本論文では，ピクトグラムを用いた画像や動画といったメディア表現そのものに加え，ピクトグラムのアノテーション情報やメタ情報を含めたものをピクトグラムコンテンツとして扱う．

アノテーション情報とは，ピクトグラムに付加される背景情報のことである．ピクトグラムは，事前の学習なしでも，即時的，国際的にわかる伝達効果を持っていることを特長とする．一方で，ピクトグラムの図案そのものは多義性があり，その掲示される場所や状況による制約が必要な場合も多い．Tijus らは，ピクトグラムを使用する文脈を与えた場合にピクトグラムの意味を適切に特定できることを実験によって示している[2]．Wolff & Wogalter は，ピクトグラムが掲示される環境についての文脈の提示が記号の意味明瞭性の評価向上に寄与するとしている[3]．日本産業標準調査会が公表する「JIS Z 8210 案内用図記号」[4]においては，図記号を使用する意図又は目的を機能として併記することが求められている.ピクトグラムが掲示される背景情報画像が掲載されており，これは一種のピクトグラムに対するアノテーション情報と見なすことができる．

2.2. ピクトグラムコンテンツの記述

ピクトグラムコンテンツを記述するデータ形式，およびそれらのデータを作成するアプリケーションについて概説する．

ピクトグラムは視覚記号の一つであり，そのデジタル表現としては画像を表現するファイル形式がよく用いられる．代表的な形式である PNG や JPEG は，色情報を持った画素の集まりとして情報をバイナリ形式で記述する．通常，コンテンツを編集する際は形式に対応したペイントソフトなどの画像編集ソフトウェアを介して編集する．コンテンツ編集者が直接データをバイナリエディタなどで編集することはほとんどない．アプリケーションがそれぞれに最適な独自の形式でコンテンツの情報をバイナリ形式で記述することも多い．

ピクトグラムはその抽象度に応じて，図記号や絵記号と呼称されるように，図形の一種である．図形はその形状の情報を，バイナリ形式だけではなく，人間が可読できるテキスト形式で記述されることもある．例

えば，ベクタ形式のコンテンツ記述として代表的なSVGは，マークアップ言語XMLに基づいて図形を記述する形式である．図形そのものの情報だけでなく，コンテンツ作成者やライセンスなどのメタ情報を記述することも可能である．コンテンツは，ドローソフトウェア上でマウス操作を軸に編集するほか，テキストエディタを用いてテキストデータを直接編集することでも行える．

テキストによる記述としては，プログラミング言語でピクトグラムを記述することもできる．第3著者は，複数のプログラミング言語で記述できるピクトグラムコンテンツ作成アプリケーション群を作成し，インターネット上で公開している[5]．テキスト型の独自記法に対応する「ピクトグラミング(Pictogramming)」のほか，JavaScript言語に対応した「ジャバスクピクト(JavaScpict)」，Python言語に対応した「ピクソン(Picthon)」といった汎用プログラミング言語による記述に対応したもの，視覚的なブロックによる記述に対応した「ピクトッチ(Pictoch)」がある．図1にピクトッチのスクリーンショットを示す．ピクトッチでは，画面右側の領域でプログラム片が記述されたブロックを組み合わせると，対応するピクトグラムが画面左側に表示される．表示されたピクトグラムは画像形式(GIF, JPEG)のデータで保存でき，またブロックのプログラムもXML形式のデータで保存できる．つまりXML形式で記述されたプログラムがピクトグラム画像に対する一種のメタコンテンツと考えることができる．

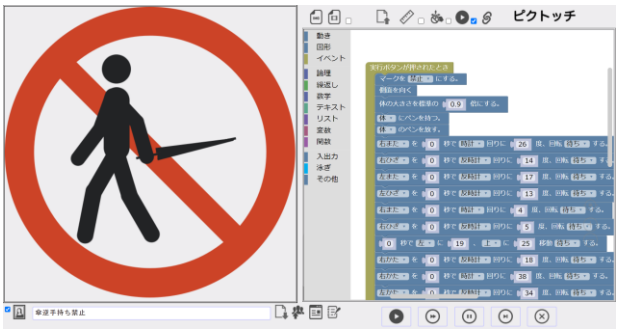


図1　ピクトッチのスクリーンショット

3. ピクトグラミーティング

本章では，開発したピクトグラムコンテンツの共有および相互評価システム「ピクトグラミーティング(Pictogrameeting)」について説明する．

「ピクトグラミーティング」という名称は，「ピクトグラム」と「ミーティング」を合わせた造語である．「集い（つどい）」がコンセプトとなっており，本システム上で，ピクトグラムの作成や登録，評価を行う人間と，ピクトグラムコンテンツに内在する人型ピクトグラムとが，アプリケーション上で集っているかのごとく感じられる場を提供する．

3.1. 提案システムの概要

本研究において提案するシステムであるピクトグラミーティングは，主に以下の3つの機能を持つ．

(1) 共有機能：ピクトグラムコンテンツのコンテンツデータの登録を受け付け，登録された情報が複数の利用者の機器から閲覧可能となるように公開する．3.2節で詳しく説明する．

(2) リスピクト：あるコンテンツを再利用した派生コンテンツの作成を支援し，その派生関係を管理する．3.3節で詳しく説明する．

(3) 相互評価機能：共有された，すなわち上記(1)で述べた共有機能で公開されたピクトグラムコンテンツについて，複数の利用者からの評価を受け付け，コンテンツに紐づける．3.4節で詳しく説明する．

ピクトグラミーティングのシステムは，今回新たに提案，作成するアプリケーションと，そのアプリケーションと連携する外部アプリケーションから構成される．ピクトグラミーティングのシステムの構成を図2に示す．提案アプリケーションは，コンテンツのデータを保存するデータベース，データベースへのアクセスを行うAPIサーバ，およびAPIサーバを通してデータベースを参照し，閲覧用のHTMLを動的に生成するWebクライアントアプリケーションからなる．これに加え，提案アプリケーションと外部アプリケーションの間で利用者やコンテンツのデータを互いに送受信することで外部アプリケーションと連携を行う．機能を実現するために利用されるデータの詳しい仕様については，3.5節で説明する．また，Webクライアントアプリケーションが提供する具体的な外観およびユーザインタフェースについては，3.6節で説明する．

3.2. 共有機能

共有機能は，ピクトグラムコンテンツの情報とメタ情報を含めたコンテンツデータの登録を受け付け，複数の利用者が各人の機器を用いてコンテンツに関する情報を取得できるように公開する機能である．この機能によって，ある利用者がシステム上にピクトグラムコンテンツを登録すると，他の利用者がそのコンテンツの情報を取得し確認することが可能となる．

3.3. リスピクト

リスピクトは，共有されたピクトグラムコンテンツの記述を再利用して，新たに派生コンテンツを作成するための機能である．コンテンツに対応した外部アプ

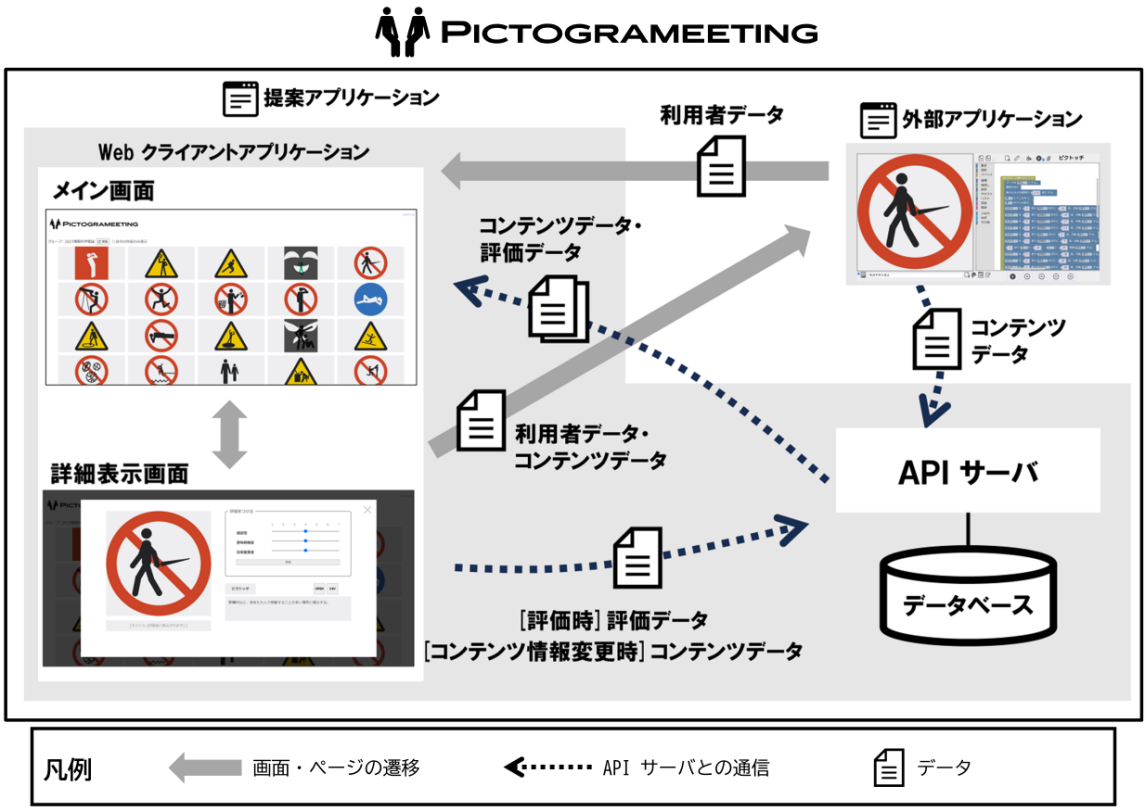


図 2　ピクトグラミーティングのシステム構成図

リケーションでその記述内容を開くことによって，コンテンツの記述内容を編集し，新たに別の派生コンテンツを作成できる．派生コンテンツの登録時は，コンテンツ自体の記述に加え，派生元のコンテンツへの参照情報を登録可能とすることで，関係をシステム上で管理できる．

本システムにおいては，この機能による派生を「リスピクト (respicto)」と呼称している．「リスピクト」という名称は，「リスペクト」と「ピクトグラム」を合わせた造語である．派生となるピクトグラムコンテンツを作成および共有する際は，派生元となるピクトグラムやその作者に敬意を払うべきであるということに由来する．

3.4. 相互評価機能

相互評価機能は，共有機能によって共有されたピクトグラムコンテンツに対して複数の利用者からの評価データの登録を受け付け，評価データをコンテンツに紐づけて管理する機能である．各コンテンツに対する評価データの登録は，利用者 1 人につきそれぞれ 1 度のみ可能であるという制約と，自身が作成した作品に対する評価データの登録はできないという制約を持つ．コンテンツごとに登録された評価データは，すべての利用者が集計結果を CSV 形式でダウンロードでき，確認できる．

3.5. 提案システムに関するデータ仕様

本節では，提案システムの機能の実装に際して定めた，具体的なデータの仕様について説明する．

3.5.1. 利用者データ

本システム全体で，利用者や API サーバの情報を共通にするものである，用いられる利用者データとして本システムで扱うデータ属性について，表 1 に示す．

表 1　利用者データの属性

属性名	必須/任意	定義
利用者 ID	必須	利用者を識別する文字列．
API サーバ ID	必須	システムで用いる API サーバを識別する文字列．
グループ名	必須	コンテンツの共有範囲の区分けに用いる，共通フレーズ．

3.5.2. コンテンツデータ

共有機能を実現するために，ピクトグラムコンテンツの内容およびコンテンツのメタ情報を管理するものである．コンテンツデータのデータ属性を，表 2 に示す．

表 2　コンテンツデータの属性

属性名	必須/任意	定義
登録者 ID	必須	コンテンツの登録を行った利用者を表す利用者 ID.
作品 ID	必須	コンテンツを識別する文字列.
派生元作品 ID	任意	リスピクトによってコンテンツが登録された場合に付与される，派生元コンテンツの作品 ID. リスピクトによらずにコンテンツが登録された場合，空になる.
タイトル	任意	コンテンツに設定された表題を示す文字列.
アノテーションテキスト	任意	コンテンツに設定された補足情報（ピクトグラムの掲示場所や掲示状況）を示す文字列.
アプリケーション情報	必須	コンテンツを作成および表示する外部アプリケーションの情報を記述する JSON 形式の文字列. JSON は，外部アプリケーションの名称を示す文字列を値にとる”name”属性を必ず持ち，外部アプリケーションでコンテンツを開く際に利用する URL を値にとる”url”属性を持つことができる.
コード	任意	外部アプリケーションによって読み書きされる，コンテンツの記述.
サムネイル画像	必須	外部アプリケーションで作成される，ピクトグラムコンテンツのスナップショット.

3.5.3. 評価データ

相互評価機能を実現するために，ピクトグラムコンテンツへの評価を管理するものである. 評価データのデータ属性を，表 3 に示す.

表 3　評価データの属性

属性名	必須/任意	定義
評価者 ID	必須	コンテンツの評価を行った利用者を表す利用者 ID.
評価作品 ID	必須	評価対象のコンテンツに付与されている作品 ID.
評価値（視認度）	任意	視認度についての評価値を示す文字列. 評価値は，1 から 7 の整数値 7 段階である.
評価値（意味明瞭度）	任意	意味明瞭度についての評価値を示す文字列. 評価値は，1 から 7 の整数値 7 段階である.
評価値（日常重要度）	任意	日常重要度についての評価値を示す文字列. 評価値は，1 から 7 の整数値 7 段階である.

3.6. 提案アプリケーションのユーザインタフェース

3.6.1. メイン画面

Web クライアントアプリケーションにアクセスすると，データベースに登録されているピクトグラムコンテンツ，すなわち共有機能によって共有されているコンテンツのサムネイル画像がタイル状に表示される. 図 3 に，スクリーンショットを示す.

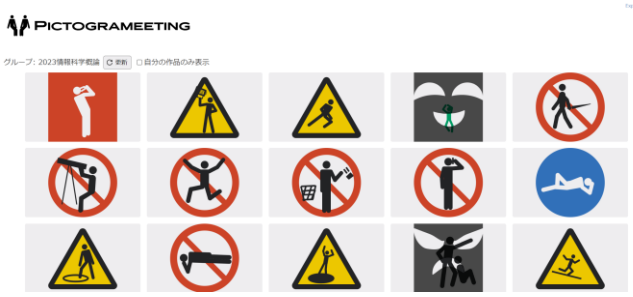


図 3　メイン画面

3.6.2. 詳細表示画面

メイン画面における各サムネイル画像をクリックすると，当該コンテンツの詳細表示画面がオーバーラップ表示される. 図 4 に, スクリーンショットを示す.

図 4　詳細表示画面

詳細表示画面は，左部，右上部，右中央部，右下部の 4 つの領域に大別される.

左部には，サムネイル画像，タイトルが表示される. 他者が作成したコンテンツである場合，タイトルは，右上部の評価フォームにて評価を行った後に表示される. 自身で作成したコンテンツである場合，タイトルの領域をクリックすることで，タイトルの編集が可能である.

右上部には，評価フォームが表示される. スライダーで評価値を決定し，送信ボタンをクリックすることで評価データの登録を行う. 他者が作成したコンテンツである場合かつ当該コンテンツに一度も評価を送信していない場合のみ，フォームが有効化される.

右中央部には，外部アプリケーションの名称，「OPEN」ボタン,「CSV」ボタンが表示される.「OPEN」

ボタンをクリックすると，外部アプリケーションの名称に示されている外部アプリケーションでコンテンツが開かれ，リスピクトが可能となる.「CSV」ボタンをクリックすると，当該コンテンツに対する評価データ一覧を CSV ファイル形式で取得できる.

右下部には，アノテーションテキストが表示される.自身で作成したコンテンツである場合，アノテーションテキストの領域をクリックすることで，アノテーションテキストの編集が可能である.

3.7. 外部アプリケーションとの連携例

2 章で示した，第 3 著者によるアプリケーションである「ピクトッチ」を外部アプリケーションとして取り入れたシステムで，ユーザが実際にピクトグラムの共有および評価，リスピクトを行う例を示す.

3.7.1. 共有機能・評価機能

利用者は，まず 3.5.1 項の利用者データに相当する情報を事前にピクトッチで設定する.ピクトッチでは，利用者 ID，API サーバ ID，グループ名の情報を URL パラメータで設定できる.利用者 ID は，ピクトッチ上で新規に発行することもできる．ピクトッチに利用者データの情報が正しく設定された場合，ピクトッチ上に「グループ共有ボタン」と「共有サイトボタン」が表示される.

設定後，ピクトッチ上でピクトグラムコンテンツの作成を行う．ピクトッチでは，プログラミングを通して静止画やアニメーションのピクトグラムの図案を作成できるほか,メタ情報としてタイトルを設定できる.

ピクトッチ上で「共有サイトボタン」を押下することで，ピクトッチからピクトグラミーティングの API サーバに対してコンテンツデータが送信される．このとき，ピクトッチの実装ではコンテンツデータの各属性を表 4 のように設定し,API サーバに送信している.

ピクトグラムの共有が行われると，ピクトグラミーティングの Web クライアントアプリケーションから閲覧できるようになる．また，Web クライアントアプリケーションの詳細表示画面(図 4)の評価フォームから API サーバに評価データを送信することもできる.利用者は，詳細表示画面上の「CSV」ボタンをクリックすることで，クリックした時点における評価データ一覧を取得できる．自身や他者の作成したコンテンツの評価データ一覧を取得して，表計算ソフトウェアなどで分析することができる.

表 4 ピクトッチにおける登録時の属性値

属性名	説明
登録者 ID	ピクトッチに設定された利用者 ID.
作品 ID	ピクトッチが新規に発行する ID 文字列.
派生元作品 ID	ピクトッチがリスピクトによってアクセスされる場合に設定される，派生元作品 ID．リスピクトによらずにアクセスされた場合，空になる.
タイトル	ピクトッチのタイトル入力フォームに設定された文字列.
アプリケーション情報	“name”属性に”ピクトッチ”,”url”属性にピクトッチ上でコンテンツを編集する際の URL の形式を記述した JSON の文字列.
コード	ピクトッチ上で編集されたブロックプログラムを記述する XML の文字列.
サムネイル画像	送信ボタン押下時にピクトッチの画面左側(図 1 を参照)に表示されている画像のスナップショット.

3.7.2. リスピクト

利用者は，ピクトグラミーティングの Web クライアントアプリケーションの,詳細表示画面(図 4)の「OPEN」ボタンを押下することで，そのコンテンツがピクトッチから共有されたものである場合ピクトッチにアクセスできる．この際，Web クライアントアプリケーションはアプリケーション情報の JSON に含まれる”url”属性の記述をもとにピクトッチにアクセスするとともに，ピクトッチに対して利用者データとコンテンツデータの送信を行っている．これにより，ピクトッチが提供する，URL パラメータで利用者データを設定できる機能によって，ピクトッチ側に利用者データが引き継がれる．加えて，コンテンツデータに含んだコードの情報(XML 文字列)をもとに，ピクトッチはブロックプログラムの読み込みを行い，派生元のプログラムを再現する．同様に，他の属性についてもピクトッチ上に渡される．利用者は再現されたコンテンツ内容をピクトッチ上で編集することができる.

この手順によってピクトッチ上から再度共有が行われる際，作品 ID を新たに発行する直前に，派生元作品 ID に作品 ID を書き込んでからコンテンツデータを送信する.

4. 考察

本システムによって，個人によって作成されたピクトグラムコンテンツが複数の利用者の間で共有され，共有されたピクトグラムコンテンツについて利用者の間で相互に評価可能となることが見込まれる.

共有機能によって，コンテンツの作成者は他者の視点や利用場面を想定し，情報デザインやコミュニケーションを意識したピクトグラムコンテンツを作成することが期待できる．

評価機能では，コンテンツに対して複数人からの客観的なフィードバックが与えられることが期待できる．評価データの出力にも対応しているので，他のコンテンツとの比較による分析などを行うこともできる．

リスピクトによる派生コンテンツの作成は，バージョン管理機能として用いて各バージョンで客観的フィードバックを受けながら漸進的にコンテンツを作成することや，他者の作成したコンテンツをリスピクトして協調的なコンテンツを作成することが見込める．

土屋ら[6]は，教育実践として，学生を対象に，ピクトグラムを作成し，電子掲示板を利用して共有・相互評価することを通して，ピクトグラムに関する知識や理論の獲得を目指す活動を行っている．本システムをピクトグラムや情報デザイン教育で用いることで，作成から評価，改善までを一つのシステム上でシームレスに行うことが可能である．

また，2.2 節で挙げた，第 3 著者が作成してきたプログラミング言語によるピクトグラムコンテンツ作成アプリケーション群[5]を採用することで，MIT メディアラボの開発する Scratch[7][8]のような，創造的な学びを提供するプログラミング協調学習環境を，ピクトグラムを統一コンセプトとして実現することも可能であろう．

5. まとめ

本論文では，ピクトグラムコンテンツ共有・相互評価システムについて，システムの機能および仕様について説明した．さらに，システムの活用についての考察も行った．今後は，本論文中での考察を踏まえてシステムの評価実験や機能拡張を行う予定である．

謝辞　本研究は JSPS 科研費 21H03560 および 20K03160 の助成を受けたものです．

参　考　文　献

[1] 太田幸夫：国際安全標識のピクトグラムデザインの研究：http://www.tamabi.ac.jp/soumu/gai/hojo/seika/2003/kyoudou-ota1.pdf (参照 2023-12-30)

[2] Tijus, C., Barcenilla, J., Cambon de Lavalette, B., Lambinet, L., & Lacaste, A. (2005). Conception, compréhension et usages de l'information iconique véhiculée par les pictogrammes. In D. Alamargot, P. Terrier, & J.M. Celler (Eds) *Production, compréhension et usages des écrits techniques au travail.* Paris : Octares.

[3] Wolff, J. S., & Wogalter, M. S. (1998). Comprehension of Pictorial Symbols: Effects of Context and Test Method. Human Factors, 40(2), 173-186.

[4] 日本産業標準調査会：JIS Z 8210 案内用図記号. (2002).

[5] 伊藤一成：複数のプログラミング言語で記述可能なピクトグラムコンテンツ作成環境の提案と実装，情報処理学会論文誌 TCE, Vol.7, No.3, pp.1-11 (2021)

[6] 土屋孝文，荒木巧也，鈴木善雄：ピクトグラムを題材にした協調学習環境の設計と運用. 2005PC カンファレンス論文集, pp.355-356. (2005) https://gakkai.univcoop.or.jp/pcc/paper/2005/pdf/0107.pdf （参照 2023-12-30）

[7] Scratch https://scratch.mit.edu/

[8] Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., Kafai, Y. (2009). Scratch: Programming for All. Communications of the ACM, 52(11): 60-67

学生向け学務情報のシステム透過なオープンデータ化

中川　匠海†　　松原　克弥†

† 公立はこだて未来大学　〒 041–8655 北海道函館市亀田中野町 116–2
E-mail: †{b1020036,matsu}@fun.ac.jp

あらまし　今日の大学では Learning Management System (LMS) や教務システムなど，目的に応じて複数のシステムを組み合わせた ICT 学習支援環境を構築している．しかし，複数のシステムに情報が分散していることや，Web サイトやメール等の限られたアクセス手段しか提供されないことで，学生に対する「学務情報への到達容易性」が課題となっている．本研究では，公立はこだて未来大学で使用されている複数のシステムの中で，Web UI を介してのみ提供されている学務情報（休講，補講，教室変更，振替連絡，課題〆切，教室空き情報など）をスクレイピング技術を用いて収集し，モバイルアプリなどで活用できるようオープンデータ化する．スマートフォンや SNS などの情報アクセス媒体に慣れた「デジタルネイティブ世代」の学生に適した学習支援の実現を行う．

キーワード　EdTech，スクレイピング，FaaS

1　はじめに

情報処理技術の急速な発展にともない，ほとんどの教育機関において，様々な ICT 学習支援システムの導入・活用が進んでいる．学校における教育の情報化の実態等に関する調査結果 [1] によると，ここ数年間で教育にコンピュータを利用する生徒数が大きく増加してる傾向にある．そのため，リモート授業やデジタル教材の使用，オンライン評価システム，データ分析など教育学習環境の充実化・効率化のためにデジタル技術の活用が必要不可欠となった．

ICT 学習支援による授業の情報化が進む中，授業時間外におけるキャンパスの利便性向上を目的としてシステム機能も増加している．立教大学では,「立教大学シラバス検索システム」[2] により，授業スケジュールに応じた教室の空き状況をリアルタイムで把握できる．このシステムを用いると，各曜日の時限毎にどの教室が空いているのかが一目でわかるため，学生は，自習や課外活動などの授業以外の活動場所を容易に見つけることができる．また，東京学芸大学には，非公式の「東京学芸大学教室検索システム」[3] が存在し，授業で使用されている教室と空き教室が一覧で表示される仕組みがある．このシステムでは，自習などの目的での利用が推奨される教室を強調表示する機能も備えている．

さらに，ICT 学習支援は，学務情報へのアクセス性や活用方法の柔軟性が大幅に向上させるだけでなく，学生の学習効率やモチベーション向上にもつながることが指摘されている．冬木 [4] は，畿央大学における LMS の利用と継続的な教育環境の変化について,「学ぶ側の能動的な学びへの期待やモチベーションが含まれていることが分かり，教える側が LMS を使った教育法をさらに工夫すれば，学ぶ側が能動的に自立的に学び，学びを深めることに繋がるメンタルモデルを形成できるのではないかと考える.」と述べている．稲垣ら [5] は，教育の場の DX が進行することによって，教育実践の高度化と生徒の高次能力の育成に大きく関与していることを主張している．

しかし，用途や目的毎の ICT 学習支援システムを複数組み合わせることによる新たな課題も浮上している．複数の Web サイトやファイル等に重要な学務情報が分散して掲載されてしまい，学生がこれらの情報を見つけ出すことがが困難なことや，情報が更新されたことに気づかないといった課題が起きている．

実際，著者らが所属する公立はこだて未来大学 (以降，未来大) において，学生が必要とする情報は，教務システム，LMS，ライブラリ等の施設毎の Web サイト，大学公式ホームページなど，異なるプラットフォームに散在している．情報の分散により，学生は必要な情報を効率的に検索してアクセスすることが困難になっているため，学習効率の低下に繋がっている．また，未来大の学務システムは，学務情報が追加されても通知などがいかず，学生自ら Web サイトなどから主体的に情報を取得する必要のある Pull 型アプローチを採用しており，学生が重要な情報の更新を見逃す危険性がある．休講情報や振替授業，課題出題などの重要情報の見逃しは，学習計画の遅延，試験準備の不足につながり，最終的には学業成績の悪化といった学習上の大きな障害となる可能性が高い．学務情報への情報到達容易性を向上させるために，新たな学習支援システムによって，学務情報が更新された際自動的に学生へ連絡のいく Push 型の情報伝達機能を導入できるようにすることが必要となる．

本研究は，モバイルアプリ等からこれらの情報への到達容易性を向上することを目的とする．デジタル化された今日の教育環境においては，情報のアクセシビリティが学生の学習成果に大きな影響を与える．本研究の成果を活用することにより，未来大における ICT を活用した学習支援をより一層推進し，デジタルネイティブ世代の学生にとって，より効率的かつ効果的な学習環境を提供することを目指す．具体的な手法として，Web サイトから自動的にデータを収集する技術である Web スクレイピングを用いて，学務情報の抽出を行い，Push 型の情報伝達機能を備えたモバイルアプリによる学習支援の手法を提案する．

2 関連研究

LMS を導入することで，出欠確認をオンラインでできる大学が増えてきている．周南公立大学では，授業毎の出欠確認についてログインパスワードを教員か学生が入力することで行なっており，授業開始から 15 分以内の入力によって，出欠確認が取れる機能を搭載した LMS を使用している [6]．しかし，これらの方法では特に授業参加者の多い授業では時間がかかり，効率が低いという問題がある．また，オンライン授業などの新しい教育形態の授業では，システムが出欠情報を適切に取得するがの難しいという課題があった．これらの課題を解決するために，児玉 [6] は,Web スクレイピング技術を用いて，Web フォームから収集した出欠情報を学務システムに自動入力するシステム開発を行った．児玉の研究では，Python のライブラリや，Web スクレイピングに適したライブラリを活用した．また，Selenium を用いて Web ページのクローリングとデータ抽出，さらには学務システム内の自動操作を行うことが可能となった．また，ブラウザ操作には WebDriver である ChromeDriver を利用し，Google Chrome を自動操作してデータの取得と入力を実行した．学務システムの運用における自動化技術の応用可能性を示し，特に大規模なクラスやオンライン授業など，従来の方法では困難だったシナリオでの出席管理の効率化に寄与している．また，この研究ではスクレイピング技術の合法的かつ効果的な使用に関する重要な洞察を提供しており，今後の教務システムの発展に貢献する可能性を秘めている．

3 Web スクレイピング

スクレイピング技術は，Web ページに存在するタグや属性と呼ばれる文字や画像情報のレイアウトといった構造化された情報を自動的に取得・解析し，構造化されているデータをプレーンなデータへと抽出し変換する技術である．このプロセスは通常，プログラムを使用して行われ，目的の Web ページの HTML データを解析し，特定の情報をユーザからのリクエストに応じて自動的に取り出すことを可能にする．スクレイピングは，データ解析，機械学習，市場調査，自動化テスト，自動入力など，多様な分野で使用されており，データ収集手段として注目を浴びている．

スクレイピングのプロセスは，通常，以下のステップで構成される．まず，HTTP リクエストを使用して対象の Web ページにアクセスし，ページの HTML データを取得する．このプロセスは図 1 のように Web クローラーと呼ばれるプログラムを使用して，複数のページを自動的に巡回し，必要なデータを集めることが一般的である．次に，得られた Web ページの HTML 構造を解析し，特定の情報を抽出する．データ解析には，主に DOM パーサーが使用される．DOM パーサーは，Web ページの HTML 構造を解析し，特定の HTML タグやクラス名，ID を持つ要素から必要な情報を抽出することができる．また，正規表現などを使用して，特定のパターンやテキストを検索し，抽出することもできる．これらのプロセスにより，

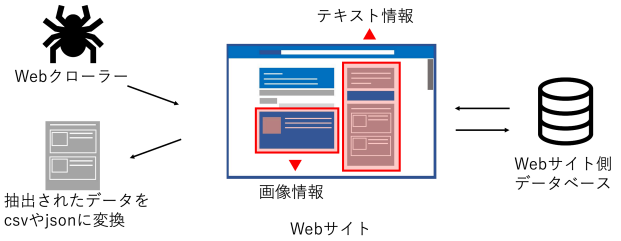


図 1　スクレイピングの解説

図 1 のように Web ページから一貫した形式でデータを収集することが可能になる．

また，スクレイピングは，動的な Web ページを多用する現代の Web サイトにおいても有効な技術である．動的に生成されるコンテンツや，ユーザの操作に応じて変わる情報も，適切なスクレイピングツールと技術を用いることで取得することが可能となる．ログイン認証が必要なページや，フォーム入力が必要な場合にも，スクレイピング技術は有効であり，このような場合，ヘッドレスブラウザや専用のライブラリを使用し，自動的にフォームを埋めたり，ログインプロセスを完了させることができる．

スクレイピングの利点は，手動でのデータ収集に比べて時間と労力を大幅に削減できることである．定期的にデータを更新する必要がある場合や，大量のデータを効率的に収集する必要がある場合にも，スクレイピングは非常に有効な手段となりえる．しかし，スクレイピングは Web サイトの構造の変化に弱く，ページのデザインや構成が変わると，スクレイピングソフトもそれに応じて更新する必要がある．

3.1 Selenium の特徴とウェブ自動化への応用

Selenium [7] は Web 自動化に特化したスクレイピングツールで，Web ブラウザの操作を自動化し，実際のユーザの行動を模倣してブラウザを操作することに秀でている [7]．フォームへの自動入力やボタンクリック，ドロップダウンメニュー選択など，ブラウザ内での多様な操作を実行しやすい．また，自動入力の機能を応用することで，ログイン認証が必要なページやフォーム入力が必要なウェブアプリケーションの操作にも適している．

本研究でスクレイピングを行う Web サイトである教務システムはログイン認証が必要であり，動的な動きのある Web ページである．そのため，本研究でスクレイピングを行うツールとして Selenium を採用した．

3.2 Chrome Driver の概要と Selenium との連携

Selenium を動作させるための WebDriver として，ChromeDriver [8] を採用した．WebDriver は，ブラウザを用いたテストの自動化を実現するためのオープンソースツールであり，Web ページの遷移，項目への入力，プログラムの実行などの機能を提供する．ChromeDriver は，GoogleChrome ブラ

ウザを制御するための WebDriver である．

ChromeDriver は，他の WebDriver と比較して，GoogleChrome の内部構造と密接に連携し，高速かつ正確なブラウザ操作を実現している．これにより，Selenium を使用したテストスクリプトは，GoogleChrome 上の動作を自動化できるようになる．

4 提　　案

本研究では，未来大の教務システムから学務情報を抽出し，オープンデータとしてアプリケーション開発に利用する．学務情報の抽出をする際，既存の教務システムの Web サーバやデータベースの運用に影響を与えずに行う必要がある．そのため，本研究では Web スクレイピングを用いることで，教務システムに影響を与えることなく，データの抽出を行い API として利用する．これにより，他システムでの学務情報の利用を効率的に行うことができる．

また，本提案によるオープンデータ化した学務情報を使用するプラットフォームには，新たな Push 型通知メカニズムを導入する．このシステムの主な目的は，学務情報が更新された場合，アプリ利用者に通知を行うことによる学生への学務情報の向上である．ユーザがアプリケーションを介してログインした際，ID やパスワードの認証情報を用いてアクセスしバックエンド上でスクレイピングを行い，最新の情報を取得する．この際，認証情報はクライアント端末へ保存し，定期的にバックエンドでスクレイピングを行い，前回のスクレイピング結果と比較され，新たな更新がある場合はクライアントへ新しい学務情報について通知を行う．データ保護の観点から，ユーザの認証情報や教務システムから取得した学務情報をクラウド上に保存することはできない．そのため，ユーザの認証情報はクライアント端末にのみ保存される．また，バックエンドでスクレイピングを行うタイミングを一部ずらすことによって，複数のクライアント端末から同時にスクレイピングが行われることを避け，教務システムサーバの負荷を低減し，サーバダウンなどのリスクを下げる．

5 実　　装

本章では，第 1 章で提案した未来大の教務システムから学務情報を，スクレイピング技術を用いて抽出し，オープンデータとして公開する実装方法について述べる．

5.1 学務情報の抽出方法の検討

本提案にあたり，教務システムからの学務情報の抽出方法について，オープンデータベースの使用とスクレイピングの利用について，比較検討を行った．

5.1.1 オープンデータベースの使用

オープンデータベースを利用することで，教務システムの学務情報を容易に取得することができる．オープンデータベースとは，未来大の教務システム運営会社が大学依頼のもと，教務システムのデータベースから一部情報のみをコピーし作成されたデータベースである．オープンデータベースへコピーされる

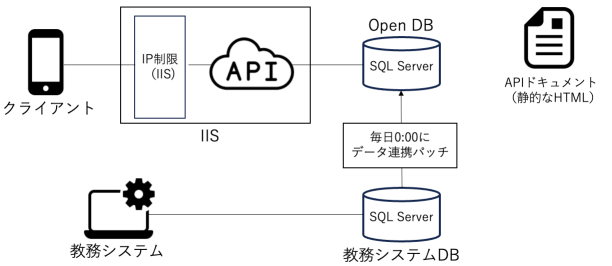


図 2　オープンデータベース基本構成

学務情報は，時間割，教室，休講，補講，教室変更情報などといった学生が閲覧可能な情報のみとなっており，それらの情報は毎日 0 時に，データ連携パッチによってコピーされ，最新の情報が保持される．コピーされたオープンデータベースの情報を元に，IP 制限を行い未来大生のみが利用できる API として公開していた（図 2）．オープンデータベースによって学務情報の取得を行うことには複数のメリットがある．

第 1 のメリットが，保守管理の容易さだ．図 2 のように，教務システム DB から一部の情報をコピーすることでデータを扱うため，定期的なメンテナンスなどの保守管理を少なく運用することができる．

第 2 に，API のレスポンス速度だ．DB から直接データを取得して扱うことができ，API の利用は静的な HTML での公開のため，容易かつ高速に API を利用することができる．静的な HTML での公開は多くの API で一般的に使われており，多くのプラットフォームで利用しやすい形式である．

だが，オープンデータベースの利用はメリットと同時に幾つかのデメリットも存在する．

第 1 のデメリットとして，セキュリティの脆弱性である．図 2 の通り，データ連携パッチはオープンデータベースと教務システムデータベースを結んでいる．オープンデータベースを今後扱っていく場合，このデータ連携パッチの通信経路から，教務システムデータベースへのアクセスが行われてしまう可能性が示唆された．

第 2 に，既存システムへの影響である．オープンデータベースは，未来大の教務システム運営会社が維持管理を行なっている．そのため，データ連携パッチの機能拡張や，他の Web サイトからの情報をオープンデータベースへ追加する場合，教務システムデータベースへのアクセスが必要になる．この教務システムは筆者のような学生が扱うことはセキュリティ上難しく，また，既に安定運用されている教務システムのデータベースや Web ページなどへ影響を与える危険性がある．

5.1.2 スクレイピングの利用

スクレイピングを利用することで，既存の教務システムに影響を与えず安全に学務情報を取得することができる．スクレイピングによって学務情報の抽出を行うことには複数のメリットがある．

第 1 のメリットが，既存教務システムへの影響を与えないこ

とだ．スクレイピングによる情報抽出は，Web サイトに掲載されている情報しか抽出することができない．そのため，スクレイピングをするにあたってアクセスが必要なのものは教務システムの情報を掲載している Web サーバのみであり，教務システムデータベースへのアクセスが必要ないため，既存のシステムを破壊する可能性を限りなく低く情報の抽出ができます．

第 2 に，学生が見れる情報しか取得できないことだ．前述の通り，スクレイピングは Web サイトに掲載されている情報しか取得することができない．未来大の教務システムの情報を閲覧するには，学生および教員が所持している ID とパスワードが必要である．つまり，スクレイピングによって学務情報の抽出を行いオープンデータ化する場合，学生の認証情報が必要になる．これはセキュリティ管理の重要性が増すというデメリットがあるが，同時にスクレイピングによって得れる学務情報が，Web ページで閲覧できる情報のみ取得できるため，オープンデータベースのように，教務システムデータベースに含まれる学生が閲覧してはいけない情報を取得してしまう危険性が発生しないメリットが挙げられる．

第 3 に，データ取得の自動化と効率化だ．スクレイピングによる情報の抽出は自動で行うことができ，手作業によるデータ入力などは必要がない．そのため，時間と人的リソースの節約が可能となり，情報更新の速度向上が見込める．

メリットと同時に幾つかのデメリットも存在する．

第 1 のデメリットとして，教務システム Web サーバの負荷だ．スクレイピングはその機能上，瞬間的に複数のアクセスが必要となる．本システムを利用する人数が増えるにつれ教務システム Web サーバへの瞬間的なアクセス数は比例して負荷が増えることになる．この問題を解決するために，スクレイピングを行うサーバでの同時処理実行数を制限し，Web サーバのダウンを防ぐ．

第 2 に，Web サイトの UI や要素名が変更されるたび，それに対応したシステムへの組み替えメンテナンスが必要となる．このデメリットの対策として，スクレイピングを扱う上でのロジックをできる限り抽象化し，特定の UI 要素に依存しない設計にすることが対策につながる．だが，大規模な UI 改変などには対応することができず，根本的な解決とはならないため，定期的なメンテナンスが必要になることが，スクレイピングを扱う上での最大のデメリットと言える．

5.1.3 学務情報の抽出法の決定

5.1.1 節，5.1.2 節の 2 つの方法による学務情報の抽出について検討を行った結果，オープンデータベースの利用は本システムを扱う上で重要となるセキュリティ対策ができていない．また，このセキュリティ対策は学生が扱うことのできない教務システムサーバに含まれた要素であるため，解決へと導くことができない．一方でスクレイピングによる認証情報の保護対策といったセキュリティ対策は，入念なシステム設計によって対策が可能である．以上のことから，本研究で扱う学務情報のオープンデータ化については，スクレイピングを用いることとした．

5.2 スクレイピングサーバの設立

本研究では，学務情報の抽出を行うスクレイピングサーバを Web 上で広く使われる標準的なインターネット通信規約に基づいたな API として実装を行う．このアプローチにより，API は一般的な Web サービスと同様に標準的なメソッド（GET,POST など）を利用してアクセス可能となる．これにより，API の利用者は環境に合わせた親しみやすいインターフェースを通じて，容易にスクレイピング機能を利用できる．

API としての提供は，本研究の枠組みを超え，大学生活を豊かにするための新しいアプリケーション開発にも利用できる可能性を持つ．開発者はこの API を利用して，教育機関のアプリケーションを開発することができ，学生たちの学習体験やキャンパスライフの質の向上に寄与することができる可能性がある．

5.2.1 Function as a Service の採用

また，本システムのスクレイピングサーバの実装において，Function as a Service(FaaS) を採用する．FaaS とは，クラウドコンピューティングサービスモデルの一つであり，サーバレスアーキテクチャを提供している．このサービスでは，個々の関数を独立したモジュールとしてデプロイし，実行することができる．FaaS の特徴として，イベント駆動型であり，特定のイベント（HTTP リクエストなど）が発生するとインスタンスが生成され，関連する関数が自動的に実行される．

本研究では 5.1.2 で述べた通り，システム利用時に必要となる認証情報と学務情報の保護が必要となる．これらの情報が漏洩することは避ける必要があり，クラウド上にそれらの情報を保存すべきではないと判断した．

本研究では FaaS の機能であるイベント駆動型によって情報の保護を行う．FaaS によるスクレイピングのリクエスト処理は以下のような流れで行われる (図 3).

（1） クライアントによるリクエストの要求
（2） インスタンスの生成
（3） ログイン処理
（4） スクレイピングの実行
（5） ログアウト処理
（6） クライアントへのレスポンス
（7） インスタンスの削除

この機能により，FaaS の非利用時にはインスタンスが生成されていないためクラウド上にデータがなく，また各インスタンスはコンテナとして展開を行なっているため，他のインスタンからの攻撃行動が行われても，情報漏洩の危険性なく，処理を実行することができる．

また，FaaS サーバを用いることで，5.1.2 節で述べたスクレイピングのデメリットであるアクセス負荷コントロールが容易になる．スクレイピングによる情報抽出自体は，サーバを必ず必要とするわけではない．例として，芝浦工業大学では，所属している学生が大学ポータルサイトの不便性を感じ，自身で新たな時間割管理アプリケーションの開発を行った [9]. 芝浦工業大学のように，技術的にはサーバを用いず，スクレイピング処理を各クライアント端末から行うことができる．だが，アプリ

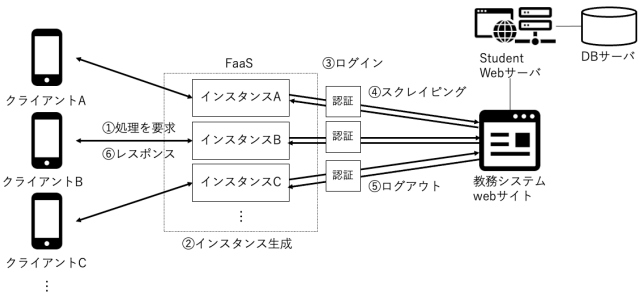


図 3　スクレイピングサーバの構成

利用者が増えるにつれて教務システム Web サーバの負荷が増大し，場合によってはサーバダウンの原因となる可能性がある．

この問題を解決するために，FaaS サーバを用いることで，クライアント端末から行われるリクエストをすべてサーバ側での管理とすることができる．一般的に公開されている FaaS 機能を提供しているプロバイダーの機能として，同時処理数を自由に変更することができる．この機能を用いることで，Web サーバの負荷想定を元に瞬間的な同時処理数を設定し，サーバダウンのリスクを減らすことが可能である．

5.2.2　プロバイダーの決定

FaaS のプロバイダーとして主要なものは Amazon Web Services(AWS) や，Google Cloud Function(GCF) が挙げられる．AWS は，広範囲のサービスと豊富な機能を提供することで知られている．特に Lambda という FaaS サービスは高い柔軟性と拡張性を持っており，多様なアプリケーションのニーズに対応できる強みがある．一方で GCF は Google Cloud Platform(GCP) [10] の一部として提供されており，Google のその他のサービスとの連携が容易であり，Google の強力なネットワークインフラストラクチャを背景に，高いパフォーマンスと安定性を期待できる強みがある．両プラットフォームとも，機能として通信，データの暗号化やアクセス制御，定期的なソフトウェアの更新，ネットワークセキュリティ等の機能を持つため，プラットフォームのセキュリティ対策はなされている．

比較検討した結果，AWS は GCF に比べ，料金体系の複雑さや設定，管理が複雑である点．GCF はシンプルな料金体系と初期設定の容易さ，および他システムとの連携の容易性から，本研究では GCF を採用することとした．

5.2.3　教務システムの最大負荷総定数

スクレイピングの実行環境として FaaS を利用する．ことにより，教務システム Web サイトへの同時アクセス数をコントロールすることができる．スクレイピングは仕様上，瞬間的に Web ページへ複数回のアクセスを行う．そのため，本システム利用者が増えるにつれて，Web ページへのアクセス数が比例的に増していく．対策として，GCF の機能である最大処理数を制限にすることで，瞬間的な教務システム Web サイトへのアクセス数を軽減し，サーバダウンの危険性を軽減する．

5.3　API レスポンス形式の選定

本研究では，API からのレスポンス形式として，csv と json の二つのデータ形式をそれぞれ比較検討した．

csv 形式は，そのシンプルさからデータの読み書きが容易であり，表形式のデータなどを扱うのに優れている．csv は大量のデータを効率的に扱うことが可能であり，データのインポートやエクスポートにおいて，多くのデータベースやアプリケーションとの互換性を持つ．しかし，csv は階層的なデータや関連するデータの構造を表現する能力に欠けており，データにメタデータを含めることは困難となっている．

json 形式は，複雑なデータ構造や階層的なデータの表現に優れている．キーと値のペアを利用することで，データ間の関係を明確にし，可読性の高い形式で提供することができる．また，json はプログラム的に処理しやすく，多言語サポートがされているため，開発者にとっても扱いやすいという利点がある．しかし，json はテキストベースであるため，大量のデータを扱う場合にはパースに時間がかかる可能性がある．

これらの比較を行った結果，json 形式が本研究で扱うデータ特性に適していると判断した．本研究では，さまざまなタイプの学務情報を処理する必要がある．それらの情報を効率的にに表現し，API を通じて提供することが可能になる．さらに，json は API レスポンスとして広く使用されており，複数のプログラミング言語やフレームワークで標準的にサポートされている．これ位より，API 利用者は異なるプログラミング環境やアプリケーションで容易にデータを取り扱うことができるようになる．

5.4　スクレイピングによって取得できる情報の要素使い道

本研究で教務システム Web サイトから抽出する学務情報は，表 1 の通り，お知らせ，休講，補講，時間割など，教務システムに掲載されている学生の日常生活に関わる情報を含んでいる．本来これらの学務情報は，Web ページの固定の場所にしか掲載されておらず，その情報を閲覧する場合，都度 Web ページをアクセスする必要がある．

本研究によるスクレイピング技術の活用と API の公開によって，これらの情報は一元化され，学生がアクセスしやすい環境となる．また，本来分散されている情報を使うことで，元々のシステムでは表示がされない新しいアプリケーションを提供することも可能となる．例として，スクレイピングによって抽出された情報を分解して，履修している授業名に対して休講補講情報を追加したスケジュール帳のような，新しい価値を提供することが可能となる．また，FaaS にて定期的にスクレイピングを行うことで，学生はスマートフォンに対応したアプリケーションを通じ，リアルタイムにお知らせや休講情報のプッシュ通知を受け取ることができ，学生は即時重要な情報を受け取るので，スケジュールの変更に迅速に対応できるようになる．

6　実験と評価

本研究で実現したシステムを評価するため，スクレイピングによって入手した情報をもとに，未来大生向けの Push 型通知を行うスマートフォンアプリの開発と実験を行った．本章ではその実験と評価について述べる．

表 1　未来大教務教務システムから取得できる情報の種別

種別	取得する情報
お知らせ	title: お知らせのタイトル. detail: お知らせの詳細情報.
休講	date: 講義が中止された日付. week: 曜日. period: 授業の時限. className: 授業の名称. campus: 授業が行われる教室名. subFaculty: 担当教員の名前. closedComment: 追加コメント（例：補講情報）.
補講	date: 補講の日付. week: 週の日. period: 授業の時限. className: 授業の名称. campus: 授業が行われる教室名. classroom: 授業が行われる教室名. subFaculty: 担当教員の名前. subComment: 追加コメント.
時間割	lesson-name: 授業の名称. timetable_id: 時間割の ID. lesson-id: 授業の ID. classroom_name: 教室の名称. staff-names: 担当教員の名前. year: 年度. term: 学期. moca-url: 授業関連の URL.

6.1　実験の概要

実験の方法として，教務システムからスクレイピング技術を利用して取得した情報をもとに，学生向けアプリケーションの有効性を評価する実験を行う．このアプリケーションによって，学生が大学生活を送る上で大学から発信される学務情報への情報到達容易性の向上がなされ，大学生活でのスケジュール管理などが効率的に行えるようになったかを評価する．

アプリケーションの機能として，学生の日常生活に関連する大学からのお知らせ，休講，補講情報の更新時，Push 通知を送信する機能を備えている．さらに今回の実験のため，未来大では pdf 形式でしか公開していない 2024 年度後期授業のテスト情報を json 形式に変換し，アプリ上で扱えるようにした．被験者が評価アプリにログインし，スクレイピングによって情報を取得した後，端末はその情報を保持し，お知らせ等の情報更新時以外に，被験者が受講しているテストの開始 1 時間前に授業名と教室名を Push 通知で表示する機能を追加した．これにより，以前までは自身のテスト開始日と開催場所を都度 pdf ファイルで確認していたが，このアプリによって大学生活での行動がより効率的になり，便利に過ごせているかの評価を行う．

実験の第一段階として，このアプリケーションを未来大生に提供し，彼らが実際の学生生活の中でアプリケーションを使用してみる機会を与える．この実験の期間中，アプリケーションは学生にとって日常の学生生活のツールとして機能し，学務情報へのアクセスを提供する．アプリケーションは，教務システムからの情報を 60 分に 1 度のタイミングで取得しており，学生にとって重要な情報をタイムリーに提供することで，彼らの学習管理と学生生活の向上をサポートする．

本実験の期間は 2024 年度後期テスト期間とする．参加する学生を対象にアンケート調査を実施し，アプリケーションの実用性と有効性を評価する．このアンケートでは，アプリケーションの使いやすさや情報の正確性，Push 通知の有効性，テスト情報の通知機能に関する評価，そしてこれらの機能から学生の行動がより効率的に活動できたかの評価と，全体的な満足度を詳しく評価する．収集したデータは統計的に分析し，アプリケーションの機能としての改善点や，学生のニーズに合わせた情報提供の効果などを詳細に検討する．

6.2　アプリケーションインターフェースと通知システムの詳細

本アプリケーションのインターフェースの設計と通知システムに関して，その具体的な機能と設計について記載する．本アプリケーションでは，学生が必要とする学務情報へのアクセスを簡単かつ迅速にすることを目的として設計している．図 4 のように，画面上ではスクレイピングによって得た情報をカテゴリ別に整理し，ボタンを押すことによってお知らせや休講，補講情報が一目で識別できるようになっている．

一方通知システムでは，学生が日々の学生生活において重要な情報を見逃さないようにサポートするための機能である．図 5 に示す通り，新しい学務情報や，受講中のテストの 1 時間前など，学生にとって重要なタイミングで Push 通知が送信される．このシステムでは，情報のタイムリーは配信により，学生がスケジュール管理や準備を適切に行うことを助けている．

本アプリケーションのインターフェースと通知システムは，未来大生の日常生活における学務情報の到達容易生向上と効率性を高めることを目指しており，これらの機能は学生が大学生活をよりスムーズに，そして効果的に行動するためのツールとなっている．本アプリケーションを実際に未来大生に短期間ではあるが使用してもらい，その学習効率や行動の変化についてアンケートを行い，評価を行う．

6.3　評 価 結 果

本アプリケーションを利用した学生に，複数のアンケートを行った結果を以下に示す．テスト開始前に Push 通知がスマートフォンに届いていたのを認識できていたかのアンケート結果を図 6 に示す．テストの Push 通知を受け取った学生のみに対し，その通知によって，テストの開始時間や場所を事前に把握できたかのアンケート結果を図 7 に示す．上記 2 つのアンケートから，Push 通知によって起きた行動の変化について記述式の結果を表 2 に示す．次に，学務情報更新がなされた際に起きる Push 通知に気づいたかのアンケート結果が図 8 に示すが，今回の実験期間内に全体へお知らせの更新が行われていないため，誰も気づかない結果となってしまった．テスト情報，学

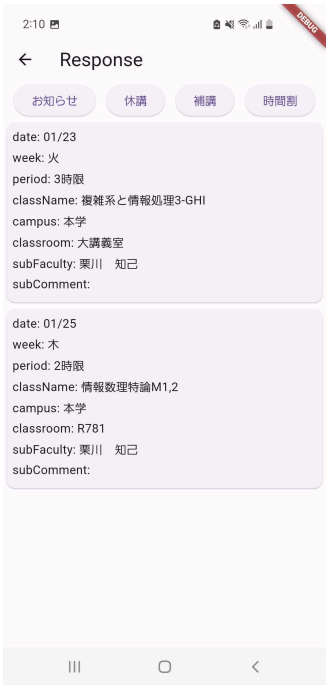


図 4　アプリケーション UI

図 5　アプリケーションの Push 通知

務情報の更新がなされた際に Push 通知がいくという本アプリケーションについて，今後も使用していきたいか，改善が必要であるかどうかのアンケート結果を図 9 に示すが，今回の実験期間内に全体へお知らせの更新が行われていないため，誰も気づかない結果となってしまった．前問での質問で「改善が必要」,「必要ない」と回答した方のみ，改善点を記述してもらったアンケート結果を表 3 に示す．最後に，本アプリケーションを利用しての感想など，思ったことを回答してもらった結果が表 4 に示す．

6.4　考　　察

本評価によって，アプリケーションからの Push 通知によって事前にテストの把握が行え，通知がない場合と比較しても，そのおかげでわずかではあるが勉強時間が増え，移動をスムーズに行うことができたという結果となった．だが，学生が受講しているテストが複数ある場合，全てのテストの Push 通知を認識していないという例や，現状のアプリケーション UI ではシステムそのものがまだ使いにくく，情報の一元化にはつながっていない．通知のタイミングなども今のところ自身で決めることで，より有意義に活動できると思ったという声もあった．

この結果から，学務情報の Push 通知によって学生は迅速に

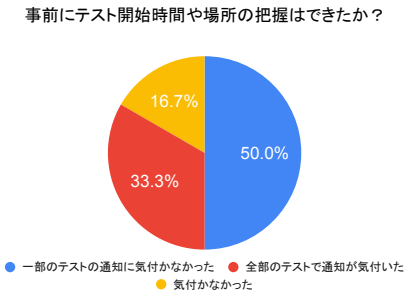


図 6　テストの Push 通知が届いたか

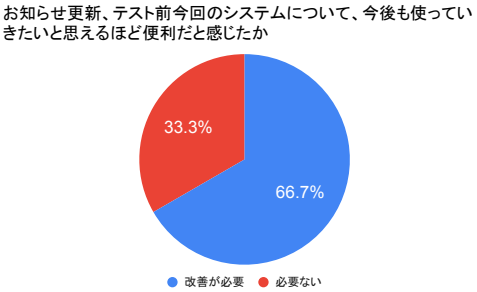


図 7　Push 通知によりテスト情報の把握ができたか

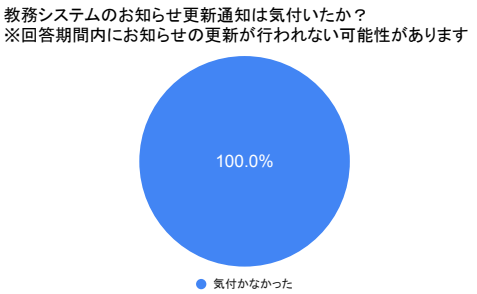


図 8　教務システムの情報更新に気づいたか

情報に気付くことができ，学生の学習効率が向上するという結果を得られた．同時に，現状のアプリケーション UI では情報の一元化という点では完了しておらず，そして学生が Push 通知を認識しやすくなるよう工夫が必要であるという結論となった．学生の声としても，本アプリケーションの機能は学生生活を行う上で効果はあるという声があり，本研究の目的である Push 型通知によって学生への情報到達容易性の向上は可能であることがわかった．

7　お わ り に

本研究では，情報が掲載される Web サイトが増え続け，情報が分散管理されているため，学務情報をシステム透過なオープンデータへと変換することで，他のデジタルシステムを通じて活用できるようにして，学務情報への情報到達容易性の向上を行うことを目的としている．その目的を達成するために，サーバレス環境である FaaS を用いて未来大の教務システムから ID やパスワード認証を用いてスクレイピングを行い，最新の学務

表 2　行動の変化のアンケート

質問	テスト前に送信された Push 通知のおかげで，少しでも多く勉強を行う時間が確保できましたか． 過去のテストなどで本アプリが無かった場合との行動の差などあれば教えてください．
回答	・前からテストの情報は見てて把握してたけど，通知のおかげで教室が分かりやすくすぐ移動は出来た ・勉強する時間が僅かに増えた ・勉強時間が少しは増えたと感じた・テストの場所を確認する程度の時間くらいしか変わってないので差はない

表 3　改善点のアンケート

質問	現状のシステムのうち必要な改善点や，追加して欲しい機能などについて教えてください． ※前問で ” 改善が必要 ”，” 必要ない ” と回答された方のみ回答してください．
回答	・UI が今の状態だと見にくい ・通知にしかテストの情報無いのでアプリ内にも欲しい ・student の情報しか今の所情報がないのでもっと欲しい ・通知のタイミングを自分で細かく決めたい ・UI の改善は必須 情報内容が今のままだとニーズに合わない気がする．よく分析し直して，載せる情報の質やニーズは見定めるべき ・少なくとも現状のままだとあっても無くてもどっちでもいいと思った ・スマートフォンの通知を見ないで溜めるので気づかなかった ・通知に気づかなったし見なくても問題ない情報だった

表 4　感想などのアンケート

質問	その他今回のアプリを使ってみて感想やアドバイスなどあれば教えてください ．
回答	・機能としてあったら便利ではあると思いました．今のままだと使い勝手が悪いので，今後に期待という感想です ・痒いところに手が届く機能なので，嬉しい人は嬉しいんじゃないかと思う

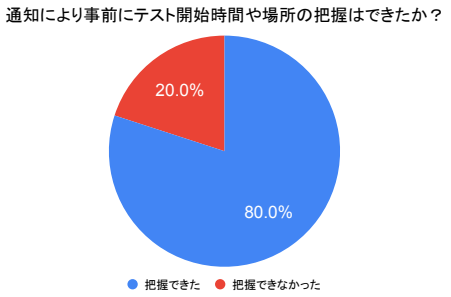


図 9　本アプリの機能を今後も使っていきたいか

情報を取得してレスポンスを返す API を作成し，実アプリケーションを用いて評価を行った．アプリケーションの機能として，定期的にバックエンドでスクレイピングのリクエストって最新の情報を取得し，新しい学務情報が更新された場合に Push 通知が送信される．これによって，学生の学務情報の情報到達容易性の向上と，学生生活での行動の変化についてアンケートを行った．結果として定期的に学務情報を取得できることで，学生は最新の情報を取得することができ，学生生活を過ごす上で学習効率などの質が向上へとつながる結果となった．

今後の課題として 2 つある．第 1 に，スクレイピングサーバの定期的なメンテナンスである．スクレイピングは HTML の ID や CLASS やレイアウトが変更されると正常に情報を抽出できないという問題がある．そのため，定期的に教務システムとスクレイピングの動作処理の確認を行い，問題があった場合都度サーバ側のプログラムコードを修正する必要がある．これはスクレイピングを用いた，システムを継続的に稼働させていくために必要なことである．第 2 に，アプリケーションの UI と機能改善である．現状のアプリケーションでは 6.4 節で述べたとおり，UI 面や機能面に大きな課題がある．この課題を解決し，学生のニーズに沿ったシステムに改善して公開することで，学務情報の情報到達容易性の向上に繋がる．

文　献

[1] 文部科学省. 令和 4 年度学校における教育の情報化の実態等に関する調査結果. https://www.mext.go.jp/, [accessed: January 07, 2024].
[2] 立教大学. 立教大学シラバス検索システム. 立教大学. https://www.rep-rikkyo.com/class_rooms?day=c&hour=4, [accessed: October 25,2023].
[3] 東京学芸大学 (非公式). 東京学芸大学教室検索システム. 東京学芸大学. https://u-gakugei-uoa.pages.dev/akitan/, [accessed:October 25, 2023].
[4] 冬木正彦. 加速された情報技術革新と大学の対応—組織と個人の対応，深い理解につながる教育実施方法—. 会誌「情報処理」. Vol.64 No.5 May 2023.
[5] 稲垣，高橋，泰山，山本. 教育実践において DX が果たす役割. 日本教育工学会論文誌,vlo.45 No.3，2021
[6] 児玉. スクレイピング技術を用いた学務システムへの出欠入力の効率化. 周南公立大学総合研究所紀要 第 1 号
[7] SeleniumHQ. Selenium Documentation, Selenium, 2023.

https://www.selenium.dev/documentation/en/,[accessed:Octorber 25, 2023].
[8] Chromium. ChromeDriver -WebDriver for Chrome. 2023. https://chromedriver.chromium.org/home.[accessed:January 21, 2024].
[9] Qiita. 自作した時間割アプリが大学公式アプリに採用された話. https://qiita.com/koheisato/items/7f2e604233372af35b41, [accessed:January 24, 2024]
[10] Google. Google Cloud Platform Documentation. https://cloud.google.com/docs/overview?hl=ja.[accessed:January 07,2024].

ユニバーサルピクトグラミング
ー人型ピクトグラムを用いたプログラミング学習環境の多ヶ国語化ー

石井 幹大† 伊藤 一成‡

† 青山学院大学理工学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1
‡ 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: † ishii@it.aoyama.ac.jp, ‡ kaz@si.aoyama.ac.jp

あらまし 我々は，ピクトグラムコンテンツの作成を通じてプログラミングの諸概念を学習できる「ピクトグラミング」を開発・公開している．ピクトグラミングは，これまで英語・日本語・日本語（ひらがな表記）のみに対応していたが，様々な自然言語によるプログラミング教育を促進することを目的に，計 109 種類の言語を扱うことが可能な「ユニバーサルピクトグラミング」を新たに開発した．開発においては，Google が提供するスプレッドシートを活用することで，少ない労力で保守を行うことができるよう工夫を行った．本稿では，対応言語拡張における背景や，設計・実装方式について解説し，ピクトグラミングに着目して多ヶ国語化を図る意義について述べる．
キーワード ピクトグラミング，人型ピクトグラム，コンテンツ作成，プログラミング，多ヶ国語，自然言語

1. はじめに

我々の研究グループで開発・公開している「ピクトグラミング (Pictogramming) [1]」は，ピクトグラムコンテンツの作成を通じてプログラミングの諸概念を学習できるという特徴がある．ピクトグラミングで扱うことのできる言語は，英語・日本語・日本語（ひらがな表記）の 3 種類のみであったが，様々な自然言語によるプログラミング教育を促進することを目的に，計 109 種類の言語を扱うことが可能な「ユニバーサルピクトグラミング」を新たに開発した．本稿では，ユニバーサルピクトグラミングについて報告する．

以下，2 章でプログラミングと自然言語に着目した理由を述べ，3 章でピクトグラミングの概要を説明する．4 章でユニバーサルピクトグラミングの実装内容を解説し，5 章では関連研究を踏まえて提案ツールの長所を述べる．最後に，6 章で今後の展望をまとめる．

2. プログラミングと自然言語

プログラミングの知識やスキルが重要であることは言うまでもない．例えば，Solomon は，プログラミングはコンピュータ使用の基礎であると述べ[2]，Roy らも「プログラミング的思考は全ての人に役立つものであり，誰もがプログラミングを学ぶべきだ．プログラミングは英語や数学と同じように多くのスキルを教える．」と述べている[3]．Tyre は，コーディングやプログラミング，コンピュータサイエンスは 21 世紀の言語になると主張する Couvares らの意見や，将来は 140 万の職でコンピュータスキルが必要とされるとする Cuny の意見をまとめた文献[4]を発表している．さらに，アメリカのケンタッキー州やフロリダ州，ニューメキシコ州などでは，プログラミングの授業単位を外国語授業の単位として認める法案が可決されており[5, 6]，Tran が賛同している[7]ことも，プログラミングが重要であることの裏付けであると言える．

次に，プログラミング言語と自然言語という 2 つの言語の関係に焦点を当てた研究に着目すると，第二外国語習得とプログラミング言語習得には関連があるとする研究がある．例えば，プログラミング初学者かつ日本語初学者の在日外国人留学生を対象に，日本語プログラミングの授業を実施した結果，約 90%の学生がプログラミング学習に興味を持つと同時に，日本語プログラミングは日本語習得に効果的であると回答した[8]．また，第二言語取得理論を適用したプログラミング授業を受講した学生は，同理論を適用しないプログラミング授業を受講した学生と比較して，モチベーションが高く，好成績を残したという報告[9, 10]もある．なお，第二言語取得理論は，模倣と反復によって暗記を行い，言語を取得する「行動主義理論」と，日常生活の中でも言語を取得する「生得主義理論」および「インタラクション主義理論」で構成される[9]．

一方，自然言語がプログラミング学習の障壁となるという研究もある．大多数の一般的なプログラミング言語には英語が用いられており，非英語ネイティブスピーカーにとって，英語がプログラミング学習の妨げになるという理由である．例えば，非英語ネイティブスピーカーは，ソースコード解読や，可読性の高いコード記述が苦手であること，`if` や `while` といった予約語が持つ意味の理解に苦労することが報告されている[11]．ビジュアルブロックプログラミング言語「Scratch [12]」を対象とした研究では，母国語版 Scratch を使用するユーザは，英語版 Scratch を使用するユーザよりも，複雑なプログラムを短時間で作成できると報告されている[13]．さらに，Bühmann, Trudell

は,「母国語の使用は，効率的な学習に重要であることが明確に示されている」と述べており[14]，母国語によるプログラミング教育が重要であると言える．

3. ピクトグラミング

ピクトグラミングは，我々の研究グループで公開している Web アプリケーションであり，ピクトグラムコンテンツを作成する中で，逐次処理，並列処理，繰返し処理，条件分岐処理といったプログラミングの概念の習得が図れるよう設計されている[1]．ピクトグラミングは,「ピクトグラム」と「プログラミング」を組み合わせた造語となっている．

ピクトグラムとは，意味するものの形状を使って，その意味概念を理解させる視覚記号であり[15]，御手洗いや非常口など，世界で広く用いられている．ISO（国際標準化機構）を中心に規格の審議や策定，共通化が進められており，ISO 3864-3 [16]では，人間を型取ったピクトグラム(以下,人型ピクトグラムと呼称)に特化した作成ガイドラインが提示されている．

3.1. 概要

ピクトグラミングに PC のブラウザでアクセスした場合のスクリーンショットを，図 1 に示す．

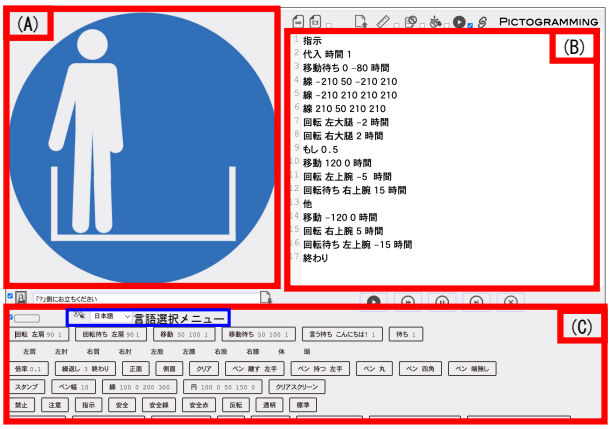


図 1　ピクトグラミングのスクリーンショット

ピクトグラミングは，大きく 3 つの領域から構成される．図 1 において，領域(A)は，プログラムの実行結果を出力する「ピクトグラム表示領域」である．表示される人型ピクトグラムは，ISO 3864-3 で定義されているサイズ比に準拠しており，領域(B)：「プログラムコード入力領域」に命令コードを入力して，コンテンツを作成する．領域(C)は，プログラムの入力を支援する「プログラムコード入力支援ボタン」であり，初期画面では英語が表示される．表示言語は,「言語選択メニュー」のプルダウンリストを操作することで変更可能である．領域(C)のボタンをクリックすると，ボタンのラベルに記述されている文字列が領域(B)上のカーソル位置に挿入される．領域(C)のボタンをマウスオーバーすると,当該ボタンのコマンドに関する説明が「言語選択メニュー」で選択中の言語で領域(A)中に表示される．言語選択メニューで選択可能な言語およびプログラムコード入力領域に入力可能な言語は，英語（略語表記および単語表記）・日本語・日本語（ひらがな表記）の 3 種類のみである．

3.2. プログラム記述仕様

プログラムコード入力領域に記述する命令は，先頭に命令コードを記述し，続けて，引数列を空白文字で区切る図 2 の書式である．命令列は，一般的なプログラミング言語と同様，上から順番に実行される．

```
命令コード　引数 1　引数 2　……
```

図 2　命令の書式

ピクトグラミングは，一般的なプログラミング言語が兼ね備える，変数定義や繰返し，条件分岐や関数定義などの命令に加え，大別して次の 3 種類の命令を備えている．

(1)　人型ピクトグラムの身体部位変形およびアニメーション命令
(2)　オブジェクトを描くための図形描画命令
(3)　禁止，注意など，ピクトグラム固有のマーク描画命令

3.2.1 項にて，サンプルコードをもとに命令の一例を解説する．

3.2.1. 命令概説

図 3 に，ピクトグラミングのサンプルコードを日本語と英語で示す．日英どちらのコードも出力される動作は同一であり，図 1 のコンテンツが生成される．

1 行目では,「指示 (Instruction) 命令」を用いて，ピクトグラム固有マークを描画している．ピクトグラム固有マークには，禁止，注意，指示，安全に関するマークがあり，ISO 3864-1 [17]でデザインガイドラインが策定されている．図 4 に，固有マークの一覧，およびピクトグラミングで固有マークを描画する際の命令を示す．

2 行目では,「代入 (SET) 命令」を用いて変数定義を行い，3 行目以降で定義した変数を使用している．3 行目の「移動待ち (MoveWait) 命令」は，アニメーション命令の 1 つであり，人型ピクトグラムを等速直線移動させる命令である．移動ピクセル数については，第 1 引数で x 軸正方向，第 2 引数で y 軸正方向を，移動に要する秒数については第 3 引数で，それぞれ指定する．すなわち，3 行目の命令は「人型ピクトグラムが 1 秒かけて y 軸負方向（上方向）に 80 ピクセル移動

する」を表している．なお，本命令と 7 行目および 11 行目の「移動 (Move) 命令」は，類似した命令である．移動命令は，当該命令と次の命令を同時に実行する一方，移動待ち命令では，当該命令の完了後に次の命令が実行される．ピクトグラミングは，「待ち」の有無によって逐次処理と並列処理の概念を理解できるよう，設計されている．

```
 1 指示
 2 代入  時間  1
 3 移動待ち  0  -80  時間
 4 回転  左大腿  -2  時間
 5 回転  右大腿  2  時間
 6 もし  0.5
 7   移動  120  0  時間
 8   回転  左上腕  -5  時間
 9   回転  右上腕  15  時間
10 他
11   移動  -120  0  時間
12   回転  右上腕  5  時間
13   回転  左上腕  -15  時間
14 終わり
15 線  -210  50  -210  210
16 線  -210  210  210  210
17 線  210  50  210  210
```

```
 1 I
 2 SET t 1
 3 MW 0 -80 t
 4 R LUA -2 t
 5 R RUA 2 t
 6 IF 0.5
 7   M 120 0 t
 8   R LUA -5 t
 9   R RUA 15 t
10 ELSE
11   M 120 0 t
12   R RUA 5 t
13   R LUA -15 t
14 END
15 L -210 50 -210 210
16 L -210 210 210 210
17 L 210 50 210 210
```

図 3　ピクトグラミングのサンプルコード
（左部：日本語，右部：英語）

図 4　ピクトグラム固有のマーク一覧

4 行目および 5 行目の「回転 (Rotate) 命令」も，アニメーション命令の 1 つであり，第 1 引数に指定した人型ピクトグラムの身体部位を，第 3 引数で指定した秒数をかけて，第 2 引数で指定した角度だけ反時計回りに回転させる．第 1 引数に指定する身体部位は，人型ピクトグラムを構成する身体部位名（図 5）または身体部位を接続する関節名（図 6）を利用する．すなわち，4 行目の命令は，「左大腿を時計回りに 2 度，1 秒かけて回転させる」を表している．

6 行目の「もし (IF) 命令」は，対応する「終わり (END)」までの命令を，第 1 引数で指定した確率で実行する．すなわち，6 行目から 14 行目の命令は「左上腕および右上腕を回転させながら，人型ピクトグラムが 1 秒かけて，等確率で x 軸正方向（右方向）もしくは x 軸負方向（左方向）に 120 ピクセル移動する」を表している．

15 行目から 17 行目の「線 (Line) 命令」は，図形描画命令の 1 つであり，座標 A（引数 1, 引数 2）から引数 B（引数 3, 引数 4）へ線分を描画する．ピクトグラミングでは，図 1 領域(A)の中心が座標(0, 0)である．

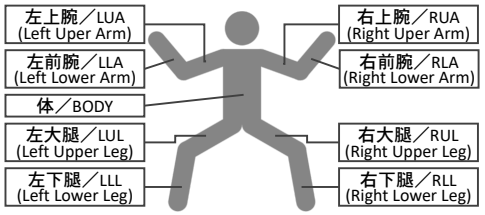


図 5　身体部位名に基づく名称（日本語／英語）

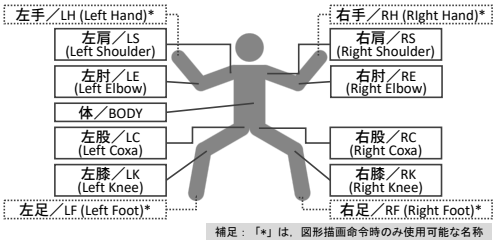


図 6　関節名に基づく名称（日本語／英語）

4. ユニバーサルピクトグラミング

本章では，ピクトグラミングの多ヶ国語化を実現したユニバーサルピクトグラミングの実装方式を解説する．

ピクトグラミングは，各国語の各命令それぞれに対して処理を記述する方式で実装されている．そのため，命令処理の保全が煩雑であるという課題があり，入力可能な自然言語を新たに追加することは容易ではなかった．そこで，ユニバーサルピクトグラミングでは，Google スプレッドシート（以下，スプレッドシートと呼称）[18]を活用して命令処理方式をパターン化することで，計 109 の言語に対応した．

4.1. 概要

実装は，Google Apps Script（以下，GAS と呼称）を利用して，ユニバーサルピクトグラミングとスプレッドシートを連携し行った．実装の結果，スプレッドシートに標準で用意されている「GOOGLETRANSLATE 関数」が対応している全ての言語（2023 年 10 月現在，109 言語）を扱うことができるようになった．

ピクトグラミングの命令記述方式は 3.2 節で述べた通りであり，先頭に命令コードを記述し，続けて，引数列を空白文字で区切る方式である．また，3 章で示したサンプルコードから，命令コードや引数列に記述する語は，一般的な用語で構成されていることが分かる．すなわち，空白文字区切りとなっている一般用語を，一対一で相互に変換する仕組みを構築すれば，ピクトグラミングの多ヶ国語化を実現することができる．本実装では，変換元言語に英語を選定し，スプレッドシートを変換テーブルのように使用することで多ヶ国語化を図った．変換元の言語に英語を選定した理由は，筆者にとって馴染みのある国際的な言語であることに

加え，英語はピボット言語として多く採用される言語だからである．ピボット言語とは，2 つの言語間で翻訳を十分に行うことができない場合に用いられる言語のことであり，ピボット言語を用いることで，翻訳元言語からピボット言語へ，ピボット言語から翻訳先言語へ，というように段階的に翻訳が行われる[19]．文献[20, 21]は，英語をピボット言語として扱うことで，翻訳精度が向上したと述べている．

4.2. 実装方式

ユニバーサルピクトグラミングに連携したスプレッドシート内部には，3 種類のシートを作成し，英語と各言語を相互に変換する仕組みを構築した．スプレッドシートに作成したシートは，「命令コード・引数列変換表」，「プログラムコード入力支援ボタン変換表 1」，「プログラムコード入力支援ボタン変換表 2」である．表 1 に各変換表の概要を示し，4.2.1 項から 4.2.3 項で，各変換表について説明する．

表 1　スプレッドシート内部に作成したシート

変換表の名称	概要
命令コード・引数変換表	命令コードおよび引数列の変換に使用
プログラムコード 入力支援ボタン変換表 1	プログラムコード入力支援ボタンのマウスオーバー時に表示する説明文の翻訳に使用
プログラムコード 入力支援ボタン変換表 2	各言語用のプログラムコード入力支援ボタンを生成するために使用

4.2.1. 命令コード・引数列変換表

本変換表は，命令コードおよび引数列の変換に使用される．図 7 に，表の一部を示す．

	A	B	C	D	E	F	G	H
1								
	en	en	en	jp	jp-kana	ar	cs	da
2	English	English	English	日本語	にほんご	عربي	Česky	Dansk
3	英語略記	英語（翻訳用）	英単語	日本語	日本語（ひらが	アラビア語	チェコ語	デンマーク
4	R	rotate	Rotate	回転	かいてん	استدارة	točitse	rotere
5	RW	rotate and wait	RotateWait	回転待ち	かいてんまち	تدويروانتظر	otáčetapočkat	Drejogven
6	M	move	Move	移動	いどう	يتحرك	hýbatse	bevægesi
7	MW	move and wait	MoveWait	移動待ち	いどうまち	تحركوانتظر	Pohybujteseapo	Flytogvent
8	W	wait	Wait	待ち	まち	انتظر	Počkejte	vente
9	FR	front	Front	正面	しょうめん	أمام	přední	foran
10	SD	side	Side	側面	そくめん	جانب	boční	side
11	MS	mask	Mask	お面	おめん	قناع	maska	maske
12	PS	persona	Persona	仮面	かめん	شخصية	Persona	Persona
13	ANGER	anger	Anger	怒り	いかり	الغضب	hněv	vrede
14	DISGUST	disgust	Disgust	嫌悪	けんお	الاشمئزاز	hnus	afsky
15	FEAR	fear	Fear	恐れ	おそれ	يخاف	strach	frygt

図 7　命令コード・引数列変換表の一部

本変換表の表頭は，各言語の言語コード（1 行目），「言語選択メニュー」内での表記（2 行目），および日本語による各言語の説明（3 行目）を持つ．なお，変換表の中には，2 行目を空白とした言語もあり，2 行目が非空白の言語のみが言語選択メニューの選択肢に表示される．

言語選択メニューの選択肢に，対応する全ての言語を表示しない理由は，全言語を言語選択メニューに列挙してしまうと，目的言語を選択するまでに時間を要し，ユーザの負担となる可能性があるためである．言語選択メニューには，表 2 に示す 21 種類の言語のみを表示する仕様とした．言語選択メニューに表示する言語は，日本語を含む 10 言語以上でのプログラミング環境を提供している Scratch [12], Snap! [22], MakeCode for micro:bit [23], Google Blockly [24], Blockly Games [25]の Web アプリケーションを参考に選定し，いずれの Web アプリケーションでも使用可能である 18 言語と，従来のピクトグラミングより使用可能であった 3 種類の言語を対象の言語とした．

表 2　言語選択メニューに表示する言語一覧

言語選択メニューでの表示	説明	言語選択メニューでの表示	説明
英語	英語	עברית	ヘブライ語
日本語	日本語	Magyar	ハンガリー語
にほんご	日本語（ひらがな表記）	한국어	韓国語
		Norsk	ノルウェー語
العربية	アラビア語	Русский	ロシア語
Česky	チェコ語	Српски	セルビア語
Dansk	デンマーク語	Svenska	スウェーデン語
Deutsch	ドイツ語	Türkçe	トルコ語
Ελληνικά	ギリシャ語	簡体中文	中国語（簡体）
Español	スペイン語	繁體中文	中国語（繁体）
Français	フランス語	Українська	ウクライナ語

本変換表の同一列内のレコードは，変換元言語となる英語での表記（1 列目から 3 列目），ピクトグラミングで従来から使用可能であった日本語表記（4 列目および 5 列目），本実装で使用可能となった言語での表記（6 列目以降）の順で構成した．6 列目以降の文字列（図 7 内の緑色枠内）の表記には，スプレッドシートに標準で用意されている GOOGLETRANSLATE 関数を利用した．この表構成によって命令処理の保全が容易となり，ユニバーサルピクトグラミングを実現することができた．

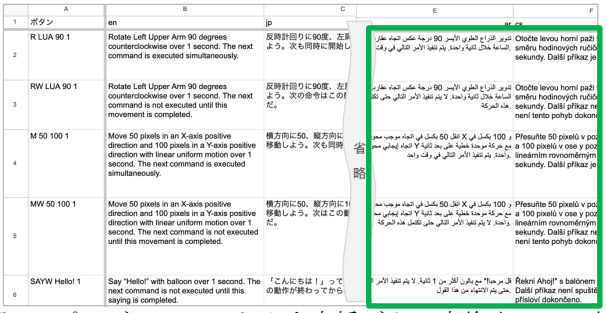
本変換表は，3 つの用途で使用される．1 つ目の用途は，プログラムコード入力領域に入力された命令を，空白文字で区切られた語ごとに英語（略語表記）命令に変換することである．例えば，デンマーク語で記述された「rotere Venstreoverarm 90」命令は，「rotere」が「R」，「Venstreoverarm」が「LS」にそれぞれ変換される．変換表に存在しない「90」は変換されず，「R LS 90」命令として実行される．2 つ目の用途は，ピクトグラム表示領域をマウスで操作した際に，言語選択メニューに表示されている言語で，プログラムコード入力領域に命令列を追加することである．マウス操作が行われると，対応する命令が英語（略語表記）で生成される．そのため，生成された英語命令を，空白文字で区切られた語ごとに，言語選択

メニューに表示されている言語での命令に変換し，プログラムコード入力領域に命令列を追加する．3 つ目の用途は,「プログラムコード入力支援ボタン変換表 2」が各言語用のプログラムコード入力支援ボタンを生成する際に参照することであり，4.2.3 項で説明する．

4.2.2. プログラムコード入力支援ボタン変換表 1

本変換表を利用し，プログラムコード入力支援ボタンをマウスオーバーした際に表示する各言語での説明文を作成する．説明文は，図 1 領域(A)中の人型ピクトグラムが発話をしているかのように表示される．図 8 に表の一部を示す．

本変換表は，プログラムコード入力支援ボタンの各言語での説明文を同一列に持つ．同一列内のレコードは，1 列目にプログラムコード入力支援ボタンとして表示する命令（英語略語表記）を，2 列目に翻訳元言語となる英語による説明文を，3 列目および 4 列目に日本語による説明文を，5 列目以降に本実装で使用可能となった言語による説明文を，それぞれ持つ．5 列目以降の文字列（図 8 内の緑色枠内）の表記には，スプレッドシートに標準で用意されている GOOGLETRANSLATE 関数を利用し，簡便な管理で多ヶ国語を扱えるようにした．



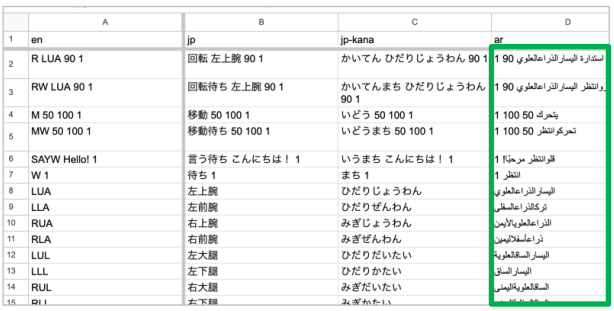
	A	B	C	E	F
1	ボタン	en	jp	ar	cs
2	R LUA 90 1	Rotate Left Upper Arm 90 degrees counterclockwise over 1 second. The next command is executed simultaneously.	反時計回りに90度、左系よう。次も同時に開始し	تدوير الذراع العلوي الأيسر 90 درجة عكس اتجاه عقارب الساعة خلال ثانية واحدة. يتم تنفيذ الأمر التالي في وقت	Otočte levou horní pazi směru hodinových ručiče sekundy. Další příkaz je
3	RW LUA 90 1	Rotate Left Upper Arm 90 degrees counterclockwise over 1 second. The next command is not executed until this movement is completed.	反時計回りに90度、左よう。次の命令はこのだ。	تدوير الذراع العلوي الأيسر 90 درجة عكس اتجاه عقارب الساعة خلال ثانية واحدة. لا يتم تنفيذ الأمر التالي حتى تكتمل هذه الحركة.	Otočte levou horní pazi směru hodinových ručiče sekundy. Další příkaz ne není tento pohyb dokon
4	M 50 100 1	Move 50 pixels in an X-axis positive direction and 100 pixels in a Y-axis positive direction with linear uniform motion over 1 second. The next command is executed simultaneously.	横方向に50、縦方向に移動しよう。次も同時	100 بكمل في X انقل 50 بكمل في اتجاه موجب محور مع حركة موحدة خطية على بعد ثانية Y اتجاه إيجابي مح واحد. يتم تنفيذ الأمر التالي في وقت واحد	Přesuňte 50 pixelů v poz a 100 pixelů v ose y pozi lineárním rovnoměrným sekundy. Další příkaz je
5	MW 50 100 1	Move 50 pixels in an X-axis positive direction and 100 pixels in a Y-axis positive direction with linear uniform motion over 1 second. The next command is not executed until this movement is completed.	横方向に50、縦方向に移動しよう。次はこのだ。	100 بكمل في X انقل 50 بكمل في اتجاه موجب محو مع حركة موحدة خطية على بعد ثانية Y اتجاه إيجابي مع واحد. لا يتم تنفيذ الأمر التالي حتى تكتمل هذه الحركة	Přesuňte 50 pixelů v poz a 100 pixelů v ose y pozi lineárním rovnoměrným sekundy. Další příkaz ne není tento pohyb dokon
6	SAYW Hello! 1	Say "Hello!" with balloon over 1 second. The next command is not executed until this saying is completed.	「こんにちは！」って の動作が終わってから	قل مرحبا!" مع بالون أكثر من 1 ثانية. لا يتم تنفيذ الأمر حتى يتم الانتهاء من هذا القول	Řekni Ahoj!" s balónem Další příkaz není spuště přísloví dokončeno.

省略

図 8　プログラムコード入力支援ボタン変換表 1 の一部

4.2.3. プログラムコード入力支援ボタン変換表 2

本変換表によって，各言語用のプログラムコード入力支援ボタンを作成する．図 9 に，表の一部示す．



	A	B	C	D
1	en	jp	jp-kana	ar
2	R LUA 90 1	回転 左上腕 90 1	かいてん ひだりじょうわん 90 1	استدارة اليسار الذراعالعلوي 1 90
3	RW LUA 90 1	回転待ち 左上腕 90 1	かいてんまち ひだりじょうわん 90 1	وانتظر اليسار الذراعالعلوي 1 90
4	M 50 100 1	移動 50 100 1	いどう 50 100 1	يتحرك 50 100 1
5	MW 50 100 1	移動待ち 50 100 1	いどうまち 50 100 1	تحرك وانتظر 50 100 1
6	SAYW Hello! 1	言う待ち こんにちは！ 1	いうまち こんにちは！ 1	قل وانتظر مرحبا! 1
7	W 1	待ち 1	まち 1	انتظر 1
8	LUA	左上腕	ひだりじょうわん	اليسار الذراعالعلوي
9	LLA	左前腕	ひだりぜんわん	تركالذراعاسطلي
10	RUA	右上腕	みぎじょうわん	الذراعالعلوياليمن
11	RLA	右前腕	みぎぜんわん	ذراعاسفلاليمين
12	LUL	左大腿	ひだりだいたい	اليسار الساقالعلوية
13	LLL	左下腿	ひだりかたい	اليسار الساق
14	RUL	右大腿	みぎだいたい	الساقالعلوياليمنى
15	RLL	右下腿	みぎかたい	

図 9　プログラムコード入力支援ボタン変換表 2 の一部

本変換表は，プログラムコード入力支援ボタンの各言語での表記を同一列に持つ.同一列内のレコードは，1 列目に変換元言語となる英語（略語表記）による表示内容を，2 列目および 3 列目に日本語による表示内容を，4 列目以降に本実装で使用可能となった言語による表示内容を，それぞれ持つ．4 列目以降の文字列（図 9 内の緑色枠内）の表記には，スプレッドシートに標準で用意されている GOOGLETRANSLATE 関数を利用し，簡便な管理で多ヶ国語を扱えるようにした．

5. 考察

ピクトグラミングに着目して多ヶ国語化を図った理由は，2 つある．本章で，その理由について述べる．

1 つ目は,ピクトグラミングには,Python 言語や Ruby 言語といった汎用プログラミング言語で命令の記述が可能な派生アプリケーションも存在するということである[26, 27]．1 章で述べたように，プログラミングの重要性は増している．一方で，汎用プログラミング言語による記述は，プログラミング初学者にとって文法や記法が複雑であると考えられ，躓きや挫折感に伴う学習意欲の喪失や，構文規則の理解の困難さなどの問題を引き起こす可能性がある．しかし，本実装によって，109 の言語を扱うことができるようになり，プログラミングに対する初学者の障壁を取り除くことができると考えている．そして，学習者の熟達度や理解度に応じて，ピクトグラミングの派生アプリケーションの中から選択して利用し，プログラミング学習を進めていくといった方法が有効になると考えている．

2 つ目は，ピクトグラミングが，ピクトグラムを用いてプログラミング学習を行うアプリケーションだからである．2 章で述べたように，ピクトグラムは世界で共通化が図られている図記号であり，利用者にとって慣れ親しんだテーマが採用されていると言える．慣れ親しんだテーマの採用は,同調的学習に有効である．同調的学習とは，プログラミング言語の LOGO を開発した Papert が重きを置いた視点である．Papert は，利用者が自身の身体を使って出力動作を模倣し，命令を実行することがプログラミング学習に重要であると述べており，学習が自身の身体に対する感覚・知識と結びついていること（身体同調）や，意図や目的を持った人間としての自意識と一貫していること(自我同調)，文化にしっかりと肯定的に根を張った活動に結びついていること（文化同調）が同調的学習であると示している[28].

身体同調に関連して，人型ピクトグラムは視覚的顕在化に優れる記号表現であるとも言える．視覚的顕在化とは，プログラミングの概念理解に関して，岡本が重要だと指摘している視点であり，予測可能性（動作

および動作位置が予測できる）や独立性（命令と出力動作が細粒度で結びつく）といった観点から構成される[29]．文献[29]は，視覚的顕在化によって，利用者が出力を強く意識可能となり，学習効果が期待できると述べている．3 章で示したように，ピクトグラミングは，既存の知識や経験，利用者自身の身体をもとに出力が予測できるような設計となっており，身体同調や視覚的顕在化に有効なコンテンツであると言える．

自我同調に関連して，ピクトグラムの特性の 1 つに表現の抽象度が高いことが挙げられる．抽象度の高さから，ピクトグラムには，それを見た人物に対して，自分自身に関わる人物事物などを想起させたり，ピクトグラムを見た人物が自分自身を人型ピクトグラムに投影したりする効果があると言われている．実際に，「非常口」のピクトグラムは,「出口へ向かって走る人型ピクトグラムを見たときに，避難中の人がいかに自身をピクトグラムに投影し，同一視できるか」という観点にデザインの労力が払われ，策定された[15]．さらに，先行研究では，自身あるいは自身に関連する事象を事物に投影することにより，投影先のターゲットから影響を受け，現実の事象や感性に変化が生じる効果が報告されている[30, 31, 32, 33]．つまり，人型ピクトグラムを目にした人物は，自身に関連する事物を人型ピクトグラムに投影し，自我同調の享受が可能だと考えられる．

文化同調についても，ピクトグラムは有効であると言える．公共施設などに広く掲示されているピクトグラムは，社会や文化的背景を反映した文化同調性の高いコンテンツであると考えられるからである．

ピクトグラミングを題材とした先行論文[1, 26, 27]でも，本章で述べた同調的学習が喚起されていると報告されており，ピクトグラムを介したプログラミング学習は有効なアプローチであると考えられる．そのため，本稿では，ピクトグラミングに着目して多ヶ国語化を図り,ユニバーサルピクトグラミングを構築した．

6. まとめと今後の展望

本稿の目的は，ピクトグラムコンテンツ作成環境「ピクトグラミング」を活用し，様々な自然言語によるプログラミング教育を促進することである．ピクトグラミングでの対応言語は，英語・日本語・日本語（ひらがな表記）のみであったが，今回，計 109 種類の言語を扱うことができるユニバーサルピクトグラミングを構築した.本稿では,対応言語拡張における背景や，設計・実装方式について解説した．

今後は，第二外国語を学習する学生や，各言語の母国語使用者による評価を通じて，提案アプリケーションの有効性を検証する予定である.また,本実装では，Google スプレッドシートに標準で用意されている関数を用いて機械的に多ヶ国語化対応を行ったが，将来的には,母国語使用者からのフィードバックを通して，より精度の高い翻訳・変換結果にしていきたいと考えている．

さらに，近年は生成系 AI が相次いで登場し，プログラミングで実現したい内容を自然言語で入力するだけでサンプルコードが提案されるなど，自然言語とプログラミング言語の垣根が低くなっている．垣根が低くなった両言語を，視覚記号の一種であるピクトグラムを活用して統一的に処理をする効果についても考察していく．

謝辞

本研究は JSPS 科研費 21H03560 の助成を受けたものです．

参 考 文 献

[1] K. Ito: Pictogramming —programming learning environment using human pictograms, In 2018 IEEE Global Engineering Education Conference (EDUCON), pp.134-141 (2018).

[2] J. Solomon: Programming as a Second Language, Learning & Leading with Technology, Vol.32, No.4, pp.34-39 (2014).

[3] P. V. Roy, J. Armstrong, M. Flatt, & B. Magnusson: The role of language paradigms in teaching programming, ACM SIGCSE Bulletin, Vol.35, No.1, pp.269-270 (2003).

[4] P. Tyre: "Is Coding the New Second Language?", http://www.smithsonianmag.com/innovation/is-coding-the-new-second-language-81708064/, Smithsonian Magazine,（2023 年 10 月 10 日閲覧）.

[5] K. M. Clark: "Spanish, French and SQL? The Push to Teach Coding Like a Foreign Language", https://www.tampabay.com/news/politics/stateroundup/spanish-french-and-sql-the-push-to-teach-coding-like-a-foreign-language/2312108/, Tampa Bay Times,（2023 年 10 月 10 日閲覧）.

[6] M. Wynn: "Ky. Ponders Teaching Computer Code as Foreign Language", https://www.usatoday.com/story/tech/2015/01/29/ky-computer-code-as-foreign-language/ 22529629/, USA TODAY,（2023 年 10 月 10 日閲覧）

[7] L. Tran: "Computer Programming Could Soon be Considered a Foreign Language in One State", https://www.mic.com/articles/81067/computer-programming-could-soon-be-considered-a-foreign-language-in-one-state, BDG Media,（2023 年 10 月 10 日閲覧）.

[8] M. Hu, & K. Emi: Educational Report of Programming Language Dolittle for Foreign Students, NAIS journal, Vol.11, pp.30-33 (2017).

[9] R. Rohmeyer, P. S. Espejo, L. Sun, & C. Frederick: A human factors perspective on learning programming languages using a second language acquisition approach, American Society for Engineering Education, 2017 (2017).

[10] C. Frederick, & L. Sun: Work in Progress: Using

Second Language Acquisition Techniques to Teach Programming-Results from a Two-Year Project, American Society for Engineering Education, 2017 (2017).

[11] P. J. Guo: Non-native english speakers learning computer programming: Barriers, desires, and design opportunities, In Proceedings of the 2018 CHI conference on human factors in computing systems, pp.1-14 (2018).

[12] Scratch - Imagine, Program, Share, https://scratch.mit.edu/,（2023 年 10 月 11 日閲覧）.

[13] S. Dasgupta, & B. M. Hill: Learning to Code in Localized Programming Languages, Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17). ACM, New York, NY, USA, pp.33–39 (2017).

[14] D. Bühmann, & B. Trudell: Mother tongue matters -Local language as a key to effective learning-, France, UNESCO (2008).

[15] Y. Ota (Ed.): Pictogram design. Kashiwa Bijutsu Shuppan (1993).

[16] ISO 3864-3:2012 Graphical symbols － Safety colours and safety signs－ Part 3: Design principles for graphical symbols for use in safety signs.

[17] ISO 3864-1 : 2002 Graphical symbols － Safety colours and safety signs －Part 1 : Design principles for safety signs in workplaces and public areas

[18] Google Sheets: Online Spreadsheet Editor | Google Workspace, https://www.google.com/intl/us_en/sheets/about/, （2023 年 10 月 11 日閲覧）.

[19] A. El-Kholy, N. Habash, G. Leusch, & E. Matusov: Language independent connectivity strength features for phrase pivot statistical machine translation, In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.412–418 (2013).

[20] N. Habash, & J. Hu: Improving Arabic-Chinese statistical machine translation using English as pivot language, In Proceedings of the Fourth Workshop on Statistical Machine Translation, pp.173-181 (2009).

[21] B. Ahmadnia, J. Serrano, & G. Haffari: Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In RANLP, pp.24-30 (2017).

[22] Snap! Build Your Own Blocks, https://snap.berkeley.edu/,（2023 年 10 月 11 日閲覧）.

[23] Microsoft MakeCode for micro:bit, https://makecode.microbit.org/,（2023 年 10 月 11 日閲覧）.

[24] Blockly Demo, https://blockly-demo.appspot.com/static/demos/code/index.html,（2023 年 10 月 11 日閲覧）.

[25] Blockly Games, https://blockly.games/, （2023 年 10 月 11 日閲覧）.

[26] 伊藤一成：Picby（ピクビー）-Pictogramming を用いた Ruby 言語の学習環境の提案–, 情報処理学会 Entertainment Computing 2019 (2019).

[27] K.Ito: Picthon –A Learning Environment of Python through Pictogram Content Creation–, IEEE FIE (Frontiers In Education) 2020, pp.1-6 (2020).

[28] S. Papert: Mindstorms -Children, Computers, and Powerful Ideas-, Basic Books Inc (1980).

[29] M. Okamoto, M. Murakami, N. Yoshikawa, & H. Kita: Development and Assessment of Learning Materials for Computer Programming Focusing on "Visual Manifestation", Educational technology research, 37(1-2), pp.51-60 (2014).

[30] Osumi, M., Imai, R., Ueta, K., Nobusako, S. and Morioka, S.: Negative body image associated with changes in the visual body appearance increases pain perception, PLoS One, Vol.9, No.9, p.e107376 (2014).

[31] Preston, C., Gilpin, H. R. and Newport, R.: An exploratory investigation into the longevity of pain reduction following multisensory illusions designed to alter body perception. Musculoskeletal Science and Practice, Vol.45, p.102080 (2020).

[32] Louw, A., Puentedura, E. J., Reese, D., Parker, P., Miller, T. and Mintken, P. E.: Immediate effects of mirror therapy in patients with shoulder pain and decreased range of motion, Archives of physical medicine and rehabilitation, Vol.98, No.10, pp.1941-1947 (2017).

[33] Nishigami, T., Wand, B. M., Newport, R., Ratcliffe, N., Themelis, K., Moen, D., Jones, C., Moseley, G. B. and Stanton, T. R.: Embodying the illusion of a strong, fit back in people with chronic low back pain. A pilot proof-of-concept study, Musculoskeletal Science and Practice, Vol.39, pp.178-183 (2019).

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

🏛2024年2月28日(水) 13:30 ~ 15:40 | 🏛T5-B オンライン（Zoom Events)

地理情報推薦

座長:廣田 雅春(岡山理科大学)
コメンテータ:小林 亜樹(工学院大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T5-B-2-01] ジオタグSNSに基づくプロンプトおよび賑やかさストリートビュー生成手法の検討

*松石 侑樹[1]、Siriaraya Panote [2]、河合 由起子[1] (1. 京都産業大学、2. 京都工芸繊維大学)

13:50 ~ 14:05
[T5-B-2-02] 旅行体験との関係抽出に基づく土産話を引き出す土産物推薦システム

*山田 樹生[1]、斉藤 さつき[1]、北山 大輔[1] (1. 工学院大学)

14:05 ~ 14:30
[T5-B-2-03] Food Approach Bias軽減のための移動中の店舗推薦手法の提案

*ヨウ シンニー[1]、國松 音和[1]、王 元元[2]、Panote Siriaray[3]、河合 由起子[1,4] (1. 京都産業大学　、2. 山口大学、3. 京都工芸繊維大学、4. 大阪大学サイバーメディアセンター)

14:30 ~ 14:55
[T5-B-2-04] 観光スポットに関する口コミの感情に基づく観点の抽出

*飯濱 航平[1]、戸田 浩之[1] (1. 横浜市立大学)

14:55 ~ 15:20
[T5-B-2-05] 博物館の展示物を用いたビンゴゲームによる鑑賞軸の習得

*西本 海生[1]、相原 健郎[2,3]、神門 典子[2,4]、荘司 慶行[5]、山本 祐輔[6]、山本 岳洋[1,2]、大島 裕明[1,2] (1. 兵庫県立大学、2. 国立情報学研究所、3. 東京都立大学、4. 総合研究大学院大学、5. 静岡大学、6. 名古屋市立大学)

DEIM2024

ジオタグSNSに基づくプロンプトおよび賑やかさストリートビュー生成手法の検討

松石　侑樹† Panote Siriaraya†† 河合 由起子†,†††

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学　〒 606-8585 京都府京都市左京区松ケ崎橋上町
††† 大阪大学サイバーメディアセンター　〒 567-0047 大阪府茨木市美穂ケ丘５－１
E-mail: †{g2054254，kawai}@cc.kyoto-su.ac.jp, ††spanote@kit.ac.jp

あらまし　本研究では，ジオタグ付きツイートおよび地理データを通じて，都市地点における人混みの状況を可視化する新しい手法を提案する．この手法は，特定の地点周辺でのツイート数を収集し，そのデータに基づいて，その地点の賑わい度合いを反映した画像を生成する．具体的には，地点情報に応じたプロンプトを用いて，Google Street View（GSV）の静的な画像に画像生成 AI を活用して人の賑わいを加える．本研究では，OpenAI の画像生成モデル StableDiffusion-v1-5 を使用し，ツイート数に基づく様々な賑わい状況に基づいた GSV のシステムを提案し，生成画像の精度を検証した. 評価実験より，任意の地点においてツイート数を用いたプロンプトにより，賑やか度を加味した画像生成が有効であることが明らかになった．

キーワード　画像生成，賑わい可視化，ジオタグ，SNS 分析，GSV

1 はじめに

近年，地図プラットフォームや Street View の発展によって，道路情報や各地点の景色を把握することが可能である．しかし，GSV の情報は，更新頻度が低く最新の情報を反映していない場合がある．また，プライバシー保護のため，人や車はモザイク処理されている．そのため，実際の人の流れや交通状況など賑やか度をリアルタイムに可視化することは困難である．

本研究では，任意の地点の賑わい度合いを考慮した安全な経路推薦を目指し，賑わい状況を可視化した路上画像生成手法を提案する．地点における賑わいは，観光客や地域の人々，海外からの訪問客といったその地点における人々の多さと考えられる．提案手法は，SNS データやレビューのうち，緯度経度と時刻が付与された Location Based Service data を対象とし，地点における人の多さや言語による特性を抽出することで，地点における賑やかな風景画像を生成する．本稿では，初期の取り組みとして，ジオタグ付きツイートデータから人の多さによる GSV 画像生成人数による基盤を構築する．賑わい状況別の路上画像生成では，まず，路上の画像 GSV から取得し，GSV の緯度経度から同地点情報として位置情報付き Tweet を取得する．次に，取得した Tweet を任意の期間にフィルタリングし，日毎の平均ツイート数を正規化した値を用いて，事前にチューニングしたプロンプトを用いて，賑わい度合いを反映した路上画像を生成する．これにより，ジオタグ付きツイートを利用して，特定の観光地の賑わい状況をリアルタイムに可視化する手法を提案する．本稿では，特に，画像生成 AI である Stable Diffusion v1-5 の Img2IMg を用い，取得した GSV 画像から生成された賑わい状況別の GSV 画像を検証する．また，画像生成に使用するプロンプトの選定についても検証する．

以下 2 章では，画像生成 AI および賑わい状況を加味した学習モデルについての関連研究について紹介する．3 章では，本研究における GSV 画像取得とプロンプトを用いた画像生成モデルおよび賑わい度に基づいた GSV システムの提案を行う．4 章では，本研究における実験と評価を行い，最後に 5 章で本研究で提案した賑わい状況別 GSV 表示システムのまとめと今後の課題について述べる．

2 関連研究

本章では，画像生成 AI および画像変換のための学習モデルについての関連研究について紹介する．

2.1 画像生成 AI と機械学習モデル

GAN（Generative Adversarial Network）[1] では, 画像生成や音声生成など，さまざまなタスクにおいて提案された深層学習モデルである．GAN は，生成器（Generator）と識別器 (Discriminator) の 2 つのモデルからなるニューラルネットワークで構成されている．Generator は，ランダムなノイズから入力データを生成するモデルであり，Discriminator は，入力されたデータが生成器で生成されたものか，本物のものかを判別するモデルである．GAN は，これらのネットワークを交互に競合させ学習を進めることで，より本物に近い画像を生成することが可能である．また，GAN の発展としても多くの先行研究 [2] [3] [3] [4] がある．

しかし，GAN の課題として学習を行うために膨大なデータセットが必要であるという課題がある．この課題を解決したのが生成 AI による画像生成である．生成 AI とは大量のデータセットを学習した AI が，そのデータセットに含まれていない

データを生成することである．生成 AI は，画像生成 AI，文章生成 AI，音声生成 AI などがある．また，StableDiffuision は，画像生成 AI の一種である．画像生成 AI は，画像を生成するために，大量の画像データセットを学習する．学習したデータセットに含まれていない画像を生成することができる．特に，Yuheng Li ら [5] は，既存のテキストから画像生成する拡散モデルを拡張する新しいアプローチを提案している．この GLIGEN は，既存のテキストから画像生成する拡散モデルにグラウンディング機能を追加することで，オープンワールドでの一般化を実現している．Katherine ら [6] は，VQGAN と CLIP という既存の事前学習済みモデルを組み合わせて使用し，これにより高品質でセマンティックに関連性のある画像を生成し，テキストプロンプトに基づいて高品質な画像を生成し，また編集することが可能であることを示した．Safinah ら [7] は，プロンプトから創造的な物語の作成を助ける新しい方法を提案した．

2.2 GIS(地理情報システム) による画像生成 AI

近年 GIS と AI の発展により，とくに画像生成分野においての進展が著しい．Ashiba ら [2] は，暗赤外線ないとビジョン画像での歩行者検出と追跡のためのアルゴリズムを提案し，提案手法が DIRNV 画像における歩行者検出と追跡に成功していることを示した．松石ら [8] は，GAN を用いて，日中の GSV 画像から夕方や夜間の GSV 画像を生成するシステムの提案をした．

2.3 SNS 分析に基づく GIS アプリケーション

また，SNS 分析に基づく GIS アプリケーションの提案は広く行われている [9]．Panote ら [10] らは，旅行者が都市内を歩く際に楽しむための重要な要素としてルート沿いの風景や施設の多様性を強調したシステムを提案し，視覚的多様性と施設ベースの多様性の両方を考慮したルート推薦システムを開発した．Yihong ら [11] は，歩行者が感じる労力を考慮したルート推薦システムを提案し．歩行者の心拍数データ，地形の傾斜，およびソーシャルネットワークからのデータ（景観評価や危険な場所の情報）を利用して，健康促進に役立つ歩行ルートの推薦を行った．Ito ら [12] は，ストリートビュー画像と CV を活用することで，都市の自転車利用における快適さを多角的な視点から評価できることを示した．

本研究では，緯度経度情報および位置情報付き Tweet，地理データ，賑わい度を考慮した GSV のシステム実装を行う点が特異点である．なお本研究では，Stable Diffusion v1.5 の Image To Image model を用いることで，生成する画像のコントロールを行う．

3 GSV 画像取得とプロンプトを用いた画像生成モデルおよび賑わい度に基づいた GSV システム

本章では，提案手法の利用例，GSV 画像の取得方法，プロンプトを用いた画像生成モデル，時間帯別 GSV 表示システムについて説明する．

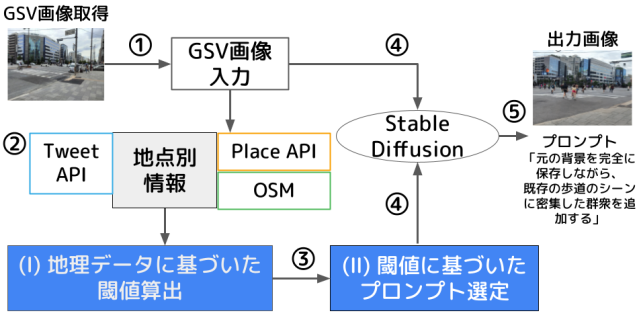


図 2　賑わい度に基づく GSV 生成システム

3.1 賑わい度別 GSV 表示システム

まず，地理データ，ジオタグ付きツイートを用いた賑わい度合いを反映した画像生成手法について説明する．次に，Twitter データの収集方法，分析手法，Stable Diffusion による画像生成プロセスなどを詳述する．

提案システムは，観光や自動車，自転車での運転前の利用を想定する．慣れていない地域で観光や運転をする場合，予め通る街や道の混雑度を把握したい場合が多いが，それが可能なのは GSV のみである．しかし，現状の GSV では，実際の景色しか把握することが出来ず，実際の賑やか度は大きく異なる可能性が高い．そこで，本システムを活用することで，事前に賑やか度を把握することが可能となり，観光や混雑度予測などに応用できると考えている．図 1 に，提案システムのインタフェースを示す．提案システムでは，画面左側に指定した地点の GSV 画像を表示し，右側に提案手法により生成された同地点の GSV 画像を表示する．また，右端のスライダーコントロールによって指定された期間の混雑度に基づいた生成画像が表示される．図中では，6 月から 9 月の期間を指定している．これにより，ユーザはその地点の賑わい度が反映された景色を把握可能となる．

図 1　GSV 提案システム

3.2 賑わい度に基づく GSV 生成システム

図 2 に，賑わい度に基づく GSV 生成システムの流れを示す．

生成システムでは，まず，GSV 画像を取得し，取得画像の地点別情報となる緯度経度から TweetAPI および google の PlaceAPI，さらに OSM（Open Street Map）情報となる地理データを取得する（図中 1 および 2）．TweetAPI からは，緯度経度，ツイー

ト情報，ツイート日，ツイート言語，Place API からは，緯度経度，レビュー情報，レビュー投稿言語，場所の種別，レビュー投稿言語，OSM からは，建築物の有無や河川の有無のデータを取得する．

次に取得した地理データに基づいた地点別の賑わい度を算出する（図中 I）．その後，閾値に基づいたプロンプト選定を行う（図中 II）．最後に，StableDiffusion に選定したプロンプトおよびあらかじめ取得した GSV 画像を入力することで賑わい度を反映した画像出力を行う（図中 4 および 5）．

3.3 賑わい度に基づくプロンプト変換

本説では，プロンプトの選定における閾値を算出するためのプロセスについて詳述する．

LBS で提供されているデータとして，Google Place API や Open Street Map，X や Instagram API が広く利用されている．本稿では，初期の取り組みとして，これまでに収集した 2023 年 3 月末までの複数年分の日本各地のジオタグ付きツイートデータを用いる．

まず，GSV から取得した位置情報をもとに約 100m のツイートデータを取得する．次に，ツイートデータから地点 p_i ごとの賑わい度を各地点 p_j ごとの時間平均ツイート数 $quantAVG_j$ と min-max 法より算出する．

$$\text{地点}p_i\text{における賑やか度} = \frac{(quantAVG_i - \text{全地点の } quantAVG_{MIN})}{(\text{全地点の } quantAVG_{MAX} - \text{全地点の } quantAVG_{MAX})} \quad (1)$$

その後，あらかじめチューニングを行ったプロンプトを閾値に基づいて設定する．なお，プロンプトにはプロンプトエンジニアリングを用いたチューニングを行った．具体的には，「Preserving the original background completely」や「Superimpose a moderate」といった細かな人数の指定や景色の保持の指定を行うことで，単純に人混みを出現させるプロンプトではなく景色を変化させないための指定を行っている．このようにプロンプトを工夫することで，原画像の景色の損失を軽減しつつ，強度によって異なる賑わい度を反映した画像生成を行う．

表 1 より，閾値に基づいたプロンプトは，混雑度の強度は 5 段階としており，それぞれ，指定した場所にのみ人ごみを生成させる．例としては，0 0.2 の時「元の背景を完全に残しながら，既存の歩道にまばらな群衆をオーバーレイする」というプロンプトを指定している．

また同様に，表 3 のように画像変換の精度を向上させるため，ネガティブプロンプトの指定も行っている．ネガティブプロンプトでは，主に景色を保持するような指示を与え，その他にも人を生成する際の精度向上のための指示を指定している．

keep the architecture, EasyNegative, lowres, bad anatomy,
bad hands, text, error, missing fingers, extra digit,
fewer digits, cropped, (worst quality:1.2),
low quality, normal quality, jpeg artifacts, signature, watermark, username,
blurry, lowres graffiti, (low quality lowres simple background:1.1),
sketch, painting, cartoon, 3d, anime, drawing,ugly face, unclear eyes, bad mouth, same peoples,
no change in buildings,
no change in landscape,
no change in scenery,
no change in buildings

図 3 ネガティブプロンプト

3.4 画像生成 AI を用いた画像変換

画像変換は，StableDiffusion の img2img 機能を用いて，GSV から取得した画像および画像変換のためのプロンプトに入力して画像を変換することで，画像の生成結果を検証する．

4 実験と評価

本研究における実験と評価について記述する．

4.1 実 験 設 定

本研究では，2020.11～2023.03 の約 24,858 地点の 352,151 ツイート取得し，それぞれの強度 (5 段階) における上位地点 3 つ，下位地点 3 つの合計 30 地点においてプロンプトによる画像生成の精度検証を行う．また，それぞれの強度における地点がどの程度存在するかを確認したところ (ア) 9,715，(イ) 6,709，(ウ) 4,428，(エ) 2,256，(オ)1,746 であった．これにより，強度が強くなるにつれて地点数が減少していることがわかる．画像変換には，画像生成モデル Stable Diffusion v1.5 を使用し,NVIDIA の CUDA 技術を利用しする. 生成される画像のサイズは 512pixel とする．

また本研究では，特定の地点と季節における画像の妥当性を検証するために 15 人を対象とした評価を行った．この評価は，GSV の入力画像と，その画像に基づいてツイート数をもとに生成された 6 つの人混みの状況を反映した出力画像を対象とした．参加者には，これらの画像の中から最も自然に見えるものを選択してもらった．評価は 6 つの異なる地点（繁華街，季節によって変わらない場所）と 2 つの季節（観光シーズン，閑散期）に関して画像が現実の風景をどれだけ正確に再現できているかという観点から評価を行った．

4.2 実 験 結 果

図 4 に入力画像および生成結果の実際の例を示す. なお，入力画像は，任意の地点の GSV 画像として，生成結果を示す. 生成画像は，賑わい度の強度が 1～5 の時の画像である．30 地点それぞれの生成画像より，使用するプロンプトの強度を強くしていくにつれ，画像内に人ごみがより多く反映されていることが確認できる．また，強度が強い場合でも公園や住宅街では，出現する人間が多すぎない風景になっていることが確認できた．一方で，地点によっては生成される画像の精度が低下する事が確認できた．具体的には，人が小さくなりすぎている場合や既存の人間ではないオブジェクトが人間として変換されてい

表 1　閾値レベルに基づく変換プロンプト

シーケンス	閾値	変換プロンプト（賑やか度強度：**5** 段階)
1	(ア) 0.0 - 0.2	①Overlay a sparse crowd onto the existing sidewalk, while preserving the original background completely(元の背景を完全に残しながら，既存の歩道にまばらな群衆をオーバーレイする)
2	(イ) 0.2 - 0.4	②Superimpose a moderate number of people onto the current street view, while preserving the original background completely(元の背景を完全に保存しながら，現在のストリートビューに適度な数の人物を重ね合わせる)
3	(ウ) 0.4 - 0.6	③Add a dense crowd to the existing sidewalk scene, while preserving the original background completely(元の背景を完全に保存しながら，既存の歩道のシーンに密集した群衆を追加する)
4	(エ) 0.6 - 0.8	④Intensify the number of people on the sidewalk significantly, while preserving the original background completely(元の背景を完全に残しながら，歩道にいる人の数を大幅に増やす)
5	(オ) 0.8 - 1.0	⑤Maximize the crowd density on the sidewalk without altering the existing buildings, while preserving the original background completely(既存の建物を変えることなく，元の背景を完全に残しながら，歩道上の群集密度を最大化する)

る場合などがあった．また実際に 15 人のユーザに評価してもらった結果としては，観光シーズンと閑散期では，生成された画像の正答率に差が見られた．観光シーズンでは，平均正答率が 35 %と評価が低い傾向にあり，閑散期では 58 %と比較的高い評価が得られた．また，観光シーズンであっても路地裏などの季節によって賑わい度が比較的変化しない場所では，正答率が低い結果となった．

図 4　賑やか度を考慮した人混み生成画像

5 考　　察

考察と検討事項について記載する. 本研究では，賑わい度を考慮した画像変換が有効であることを確認した．また，強度を強くしていくにつれて，生成される画像において，より賑やかな画像が生成されることを確認した．そのため，より詳細な地点情報を加味した画像変換を行うことで，実際に景色により近い画像生成が可能になると考えられる提案された手法によって生成された画像の評価では，地点と季節によって大きな差が存在することが明らかになった．特に，繁華街など人通りの多い地点では高い正答率が示されたが，人通りが少ない路地裏や住宅街では，正答率が低い傾向にあった．このことは，ツイート数と実際の人混みの状況との間に乖離が存在することを示唆しており，地点別情報などのツイート数以外の要因を考慮に入れることが，画像生成の精度を高める上で重要であることを示している．また，季節による変動も顕著であり，観光シーズンよりも閑散期の方が正答率が高いことから，季節に応じたアプローチの考慮が必要であると考えられる．今後の予定としては，地点情報やユーザ特性を考慮した画像生成との比較を検討する．

6 終 わ り に

本研究では，ジオタグ SNS に基づくプロンプトおよび賑やかさストリートビュー生成手法の検討の提案した．提案手法では，任意の地点の賑わい度合いを考慮した安全な経路推薦を目指し，賑わい状況を可視化した路上画像生成手法を提案した．評価実験により，ツイート数を用いたプロンプトに基づく賑やかさの加味された画像生成が，特定の条件下で有効であることが明らかになった．特に，閑散期や繁華街など特定の地点では高い正答率を達成したが，路地裏や住宅街のような人通りの少ない地点では正答率が低い結果となった．

この結果は，ツイート数と実際の人混みの状況との間に乖離がある場合があることを示しており，画像生成の精度をさらに向上させるためには，ツイート数以外の要因を考慮に入れる必要があることを示唆している．今後の研究では，さらに複雑な地点情報や季節変動を加味した画像生成手法の開発を目指す．これにより，より現実に即した賑わい度合いの可視化と，それを活用したより実用的な経路推薦システムの実現が期待される．

謝　　辞

本研究の一部は，JSPS 科研費 JP19K12240，JP20H00584 および京都産業大学先端科学技術研究所（HMD 共生科学研究センター）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, oct 2020.

[3] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial

networks, 2017.

[5] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation, 2023.

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance, 2022.

[7] Safinah Ali and Devi Parikh. Telling creative stories using generative visual aids, 2021.

[8] 山口 琉太 Panote Siriaraya 河合 由起子 松石 侑樹, 羽倉 輝. 店舗情報と tweet の影響を考慮した時間帯別路上 gsv 生成手法の提案. 03 2023.

[9] Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning*, 215:104217, 11 2021.

[10] Yihong Zhang, Panote Siriaraya, Yuanyuan Wang, Shoko Wakamiya, Yukiko Kawai, and Adam Jatowt. Walking down a different path: Route recommendation based on visual and facility based diversity. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 171–174. ACM, 2018.

[11] Panote Siriaraya, Kodo Maeda, Yusuke Nakaoka, Yukiko Kawai, and Shinsuke Nakajima. A smart walking navigation system based on perceived exertion. In Alan Said and Takanori Komatsu, editors, *Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018), Tokyo, Japan, March 11, 2018*, volume 2068 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

[12] Koichi Ito and Filip Biljecki. Assessing bikeability with street view imagery and computer vision. 05 2021.

旅行体験との関係抽出に基づく土産話を引き出す土産物推薦システム

山田　樹生†　　斉藤さつき†　　北山　大輔†

† 工学院大学情報学部　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j321317,jx20104}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　旅行から帰ると，誰かにその体験を共有したくなることがある．その時，体験を共有する相手に土産物を贈ることも多い．土産物は相手の好みなども考慮する必要があるが，共有したい旅行体験に関連するものを選ぶことで，共有の効果が高まることが期待できる．そこで本研究では，土産物を選ぶ際に，消費期限や価格などの制約に加えて，旅行体験との関連度に基づく土産物推薦システムを提案する．本稿では，プロトタイプシステムを構築し，その動作について報告する．

キーワード　観光，地理データ活用，行動データ，体験抽出

1 はじめに

2018 年のゼネラルリサーチによる「お土産に関する消費者動向調査」[1]によると，旅先での他人への土産物は男女ともに 75%以上の人が「必ず買う」「ときどき買う」と回答しているなど，土産物と旅行は切り離しがたい関係にあるといえる．また，他人への土産物に求めるものとして「ご当地ロゴ」「限定品・コラボ商品」が食べ物編でももの編でも上位を占めており，「観光・旅行してきた事実がわかる商品が求められている」と分析されている．そこで，我々は購入した土産物はその相手に渡すときに，多くの場合，土産話をすることが多く，その土産話がより豊かになる土産物を選択できると望ましいと考えた．

山岸ら [1] は，個人適応型観光のために旅行体験の推薦を行っている．このように，旅行において，どのような体験をするのかは非常に重要視されているが，その体験を応用する研究は少ない．一方，土産物に関して，江崎ら [2] は食の好みに基づいた土産物の推薦を行う手法を提案しているが，旅行体験を応用するものではない．

そこで本研究では，旅行体験と土産物の特徴の対応関係を特定することで，旅行体験に基づく土産物の推薦をする手法を提案する．旅行体験や土産物の特徴の抽出はキーフレーズ抽出の技術を用いて実現し，その類似度で旅行体験に関連する土産物を特定する．例えば，旅行先として歴史上の人物にまつわる場所を訪問していた時に，その人物がかかわった土産物を推薦するようなことを可能とする．このことにより，土産物を渡す際に自身の旅行体験を土産話として話しやすくなり，旅行体験の共有が円滑になりやすくなることが期待できる．

本研究の貢献は以下のとおりである．

- 旅行体験と土産物の関連性を特定するモデルを提案する
- 実例に基づいて提案モデルの性能の評価と限界を議論する

1：https://general-research.co.jp/report04/

2 提案手法

2.1 旅行体験の抽出

本稿ではじゃらん [2]から収集したレビューを使用して旅行体験を抽出する．じゃらんのデータとしては，スポットの名称，位置情報，説明文，カテゴリなどが付随し，レビューとしては，レビューの文書，評価値，訪問時期，訪問者の属性などが付随する．本稿ではレビューの文書のみを用いて行う．じゃらんのレビュー文が必ずしもユーザの個人的な体験と一致するとは限らないが，そのスポットの代表的な体験をユーザも体験していると仮定する．訪問時期や訪問者属性などでレビューを絞り込むことでより適切な体験を抽出できる可能性があると考えるが今後の課題とする．

ユーザの訪問スポット集合に関しては，Google map のタイムライン機能などの手段で取得することを考える．得られた今回の旅行での訪問スポット集合に付随するレビュー文を各スポットごとに結合したものを 1 つの文書として，EmbedRank [3] を使ってキーフレーズを抽出し，その和集合をユーザの旅行体験とする．EmbedRank は，文書全体に類似するようなキーフレーズを抽出するアルゴリズムで，ここのキーフレーズが類似しないように，文書要約でも用いられる Maximal Marginal Relevance を用いている．本稿では，類似度の判定に fasttext を用い，文書全体におけるキーフレーズの出現頻度の対数を重みとして加えるという変更を加えている．抽出したキーフレーズが，直接的に体験を表現しているわけではないが，スポットの特徴的なキーフレーズであり，代表的な体験に関係するものと仮定する．

2.2 土産物の特徴の抽出

土産物の特徴抽出としては，対象の土産物の Wikipedia ページおよび土産物名で Web 検索した際の検索結果上位 n 件を用いる．旅行体験と結びつく可能性のある特徴を抽出する必要があるため，土産物自体の特徴のみならず，関係する歴史や人物，

2：https://www.jalan.net/travel

表 1　旅行体験のキーフレーズ

仙台朝市	安い 新鮮，仙台 駅，仙台 朝市，魚介 類，店 人，海鮮 丼，美味しい 店，商店 街，観光 客，楽しい 朝市，コロッケ 屋，仙台 台所，安い 野菜，地元 人，仙台 朝市 クチコミ，安い 駅，安い 手，仙台 駅 近く，地元 食材，近い 便利
松島湾	遊覧 船，日本 三景，美しい 景色，松島湾 クチコミ，日本 三景 一つ，観光 船，素晴らしい 景色，赤い 橋，良い 景色，松島 観光，景色 堪能，観光 客，土産 屋，美しい 島々，日本 三景 ひとつ，何 度，島々 景色，松島 湾，自然 満喫，風光 明媚
仙台城跡	仙台 城跡 クチコミ，仙台 城跡，仙台 市内 一望，仙台 城，仙台 市内，仙台 駅，伊達 政 宗 騎馬 像，仙台 一望，伊達 政宗像，伊達 政 宗，仙台 観光，伊達 政 宗 像，仙台 市街，仙台 街，伊達 政 宗 銅像，伊達 武将 隊，仙台 城址，資料 館，良い 場所，仙台 街 並み

表 2　土産物のキーフレーズ

萩の月	萩 月，模倣 品，個 包装，仙台 銘菓，[4]。，土産 品，個人 ウェブサイト，雪 組 トップ スター，日本 原産，2021 年 11 月 7 日 閲覧，仙台 空港，[4]。「 伊達 絵巻，宮城 県，岩田 華 怜，紫色 花，熊谷 和徳，昭和 54 年，機内 菓子，仙台 土産，松任谷 由実
仙台牛タン	2019 年 4 月 13 日 閲覧，牛 タン，仙台 牛，牛 タン 定食，日 米 牛肉，牛 タン 焼き，オレンジ 交渉，宮城 県，米国 産，一般 的，昭和 25 年，仙台 牛 銘柄 推進 協議 会，仙台 名物，テール スープ，牛肉 輸入 自由 化，稲荷 小路，宮城 風土記 (2)』 宝 文 堂，仙台 牛 タン 焼き，山形 県 伝統 料理，エネルギー 革命 前
ずんだ餅	仙台 名物，一般 的，餡 餅，文 リンク 編集 ページ ノート 日本語 閲覧 編集 履歴 表示 ツールツールサイドバー，枝豆 使用，その後 追随，場合 電子 レンジ，カテゴリ：出典 必要，枝豆 餡，ダウンロード 印刷 用 バージョン 出典：フリー 百科 事典，免責 事項 行動 規範 開発 者 統計 Cookie，仙台 (株) 黄金 食品，声明 モバイルビュー 本文 横 幅 制限 有効化，取得 カテゴリ：餅菓子 宮城 県 土産 菓子 山形 県 土産 菓子 日本 郷土 料理 仙台 藩 農山 漁村 郷土 料理 百 選，餅 - Wikipedia コンテンツ，季節 商品 化，脚注 [編集][脚注 使い方] 出典 [編集] 関連 項目 [編集] 和菓子 団子 大福 饅頭，水分 多く，詳しい 投稿 記録 トーク 目次 サイド バー，表示 案内 メインページコミュニティ・ポータル 最近 出来
笹かまぼこ	かまぼこ 類，2021 年 10 月 17 日 閲覧，風味 かまぼこ，抜き かまぼこ，板 蒲鉾，蒲鉾 型，品質 表示 基準 見直し，鳴門 巻き，特殊 包装 かまぼこ 類，蒸板 蒲鉾，すり身 原料，レ [，調理 科学，一般 的，練り つぶし 魚肉，巻 蒲鉾，阿部 蒲鉾店，蒲鉾 製品，贈答 品，原色 食品 図鑑

組織なども抽出するために wikipedia ページも対象とする．

抽出方法としては，旅行体験同様に EmbedRank を用いて，対象の Web ページからキーフレーズを抽出することで，土産物の特徴として扱う．この時，対象とした個々の Web ページから個別に EmbedRank を適応し，その和集合を用いる．

2.3　旅行体験と土産物の関連性の判定

抽出した旅行体験と，土産物 t の関連性を文書類似度 sim を用いて判定する．抽出されるキーフレーズは，数個のキーワードで構成されるため，fasttext を用いて単語分散表現に基づいて文書類似度の計算を行う．それぞれの旅行体験のフレーズ E および土産物のフレーズ S_t の類似度を総当たりで計算し，閾値 α 以上となる組を関連のあるフレーズとして判定する．ある土産物に関して抽出されるフレーズを式 1 で定義する．

$$rel_phrases(t) = \{s; e \in E, s \in S_t, sim(e,s) >= \alpha\} \quad (1)$$

対象となる土産物リストに対し，旅行体験に対し関連するフレーズを多く持つほど，今回の旅行体験に関連する土産物と判断する．また，関連するフレーズの元となった Web ページ内の文を関連性を表す文とする．関連の強さ順に土産物を関連性を表す文とともに提示することで，土産物の推薦結果とする．この時，消費期限やジャンルなどでの絞り込みを行うこともできる．

3　実　行　例

3.1　旅行体験と土産物特徴のの抽出例

「仙台朝市」「松島湾」「仙台城跡」に訪問したと仮定し，それぞれのキーフレーズを 20 件抽出した．その結果を表 1 に示す．抽出されたキーフレーズは単語ごとにスペースで区切られており複数単語で一つのキーフレーズを構成する．抽出結果としては，それぞれのスポットの特徴といえる単語が適切に抽出されており，おおむねうまく行っていると考える．

「萩の月」「仙台牛タン」「ずんだ餅」「笹かまぼこ」を候補の土産物としたときの結果について説明する．キーフレーズは 20 件抽出し，対象とした Web ページは Web 検索結果は用いておらず wikipedia ページのみである．また，wikiepdia の機能によりリダイレクトされた場合はリダイレクト先のページを用いている．結果を表 2 に示す．萩の月，仙台牛タンに関しては，比較的適切な特徴を示すキーフレーズが抽出されていると考える．また,「松任谷由美」や「エネルギー革命」のような関連する人物や歴史を示すようなキーフレーズも抽出されている．一方，ずんだ餅に関しては，Wikipedia ページ特有の単語が多数キーフレーズとして抽出されている．これはずんだ餅のページのコンテンツの分量が少なく，不必要なフレーズが上位に出現したためだと考える．web 検索結果も併用し，コンテンツの量を増やすことで解消することが考えられる．笹かまぼこについては，一般的なかまぼこのページの一部に笹かまぼこが出現するため，データとしては一般的なかまぼこのものが使われているため，笹かまぼこ特有のキーフレーズの抽出が行えていな

い．このケースも Web 検索結果により土産物名特有のコンテンツを増やすことによる解決が考えられる．

3.2 対応関係に関する考察

「仙台朝市」「松島湾」「仙台城跡」と「萩の月」「仙台牛タン」「ずんだ餅」「笹かまぼこ」の結果を用いて，キーフレーズの対応関係を抽出した．類似度の閾値としては，0.3 を用いた．

結果としては,「仙台城跡」と「萩の月」のみで閾値を超えるキーフレーズが存在した．キーフレーズとしては “仙台城跡：仙台 街 並み”，“萩の月：[4]。「 伊達 絵巻” や，“仙台城跡：伊達 政 宗 像”，“萩の月：機内 菓子” である．しかしながら，関連するフレーズであるとは考えにくく，適切に対応関係を表現できているとはいえない．キーフレーズ抽出を行っているテキストがレビューと wikipedia であるため，抽出されたキーフレーズが表現している特徴が異なり，対応しにくいということが考えられる．観光スポットに関しても wikipedia ページや Web 検索結果を用いることや，土産物に関してもレビューを用いるなど，抽出対象とするテキストをそろえることで，キーフレーズが表す特徴をそろえることが必要になると考えられる．

4 ま と め

本研究では，旅行体験と土産物の特徴の対応関係を特定することで，旅行体験に基づく土産物の推薦をする手法を提案した．そのために，EmbedRank を用いた旅行体験と土産物特徴の抽出を行い，類似度に基づく対応付けを行った．結果としては，キーフレーズの抽出は比較的適切であるものの，対応関係については適切に行えない結果となった．キーフレーズ抽出を行うテキストを拡大し，キーフレーズが示す特徴をそろえることが必要であることを確認した．

今後の課題としては，抽出した旅行体験，土産物特徴の妥当性の評価が必要であると考えている．また，土産物推薦の精度についても評価する必要があると考える．本手法で抽出する特徴は，旅行体験と土産物の対応関係を判定するための特徴であるため，旅行体験としての妥当性，土産物特徴としての妥当性を評価しても適切ではないと考えられる．そのため，対応関係判定結果の評価と同時に行う必要があると考える．そこでまず，旅行履歴をいくつか用意し，その旅行履歴を持つ旅行者の土産話を引き出せる土産物を被験者に選んでもらうことで，正解データセットを作成する．この正解データセットを用いて，対応関係の精度を評価することで，間接的に旅行体験，土産物特徴の妥当性についても分析を行う予定である．

文 献

[1] 山岸立, 馬強. 個人適応型観光のためのユーザ体験推薦. 日本データベース学会和文論文誌, Vol. 20-J, No. 15, pp. 1–8, 2022.
[2] 江崎浩矢, 上岡英史, 金丸真奈美. お土産推薦システムのための sns 情報を用いた食の好み推定手法. 研究報告ヒューマンコンピュータインタラクション（HCI）, No. 8, pp. 1–8, 2023.
[3] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings, 2018.

DEIM2024

Food Approach Bias軽減のための移動中の店舗推薦手法の提案

ヨウ シンニー† 國松 音和† 王 元元††

Panote Siriaray††† 河合由起子†,††††

† 京都産業大学情報理工学部 〒 603–8555 京都府京都市北区上賀茂本山
†† 山口大学工学部知能情報工学科 〒 755-8611 山口県宇部市常盤台 2-16-1
††† 京都工芸繊維大学 〒 606-8585 京都府京都市左京区松ケ崎橋上町
†††† 大阪大学サイバーメディアセンター 〒 567–0047 大阪府茨木市美穂ヶ丘 5 番 1 号
E-mail: †{g2054461, g2153452, kawai}@cc.kyoto-su.ac.jp, ††{spanote, circl.wang}@gmail.com

あらまし 近年，生活習慣病の増加は世界的な健康問題となっており，日常生活での効果的な改善方法が求められている．本研究は，日常生活の運動としてジョギングや通勤通学の移動の際に，食品選択に対するバイアス Food approach bias を軽減させることを目的に，移動経路中の店舗推薦方法を提案する．Food approach bias は，食品の選択および回避において食べる機会を剥奪することよりも，個人の嗜好がバイアスとして影響を及ぼすとされている．一方で，移動中は多くのレストラン等の店舗が存在しており，それらを回避することは現実的ではない．そこで，生活習慣病の運動として移動の目的地までの経路生成をするために，ユーザの肥満レベルと推薦店舗選択の履歴に基づき，運動方式（速足やジョギング）ごとに Food approach bias を軽減する店舗推薦手法を提案する．ユーザが推薦店舗から選択した店舗を用いることで，運動方式ごとの最短経路を生成する．本稿では，肥満レベルに応じた Food approach bias を軽減する店舗推薦ならびに移動経路推薦手法を提案し，開発した店舗および経路推薦を用いて，肥満レベルごとに推薦された店舗（食品）に対するバイアス軽減の影響を検証する．本稿では，肥満レベルに応じた Food approach bias を軽減する店舗推薦ならびに移動経路推薦手法を提案し，開発した店舗および経路推薦を用いて，ユーザの嗜好に基づき推薦された店舗（食品）のバイアス軽減を検証する．

キーワード 経路推薦，Food Approach Bias，生活習慣病

1 はじめに

近年，健康増進や生活習慣病の予防といった健康的，経済的なライフスタイルの推進が世界的に注目されている．特に，生活習慣病の発症率に関連する生活習慣の改善が求められている．厚生労働省によると，生活習慣病の患者数は増加しており，その発症率には個人の生活習慣によって大きな差があると指摘されている．これに伴い，健康食品やフィットネスクラブの利用が増えてきた一方で，科学的根拠に基づく効果的な生活習慣の改善方法やサポートが求められている．また，株式会社リンクアンドコミュニケーションと株式会社東芝 [1]は，AI 技術を活用して生活習慣病の発症リスクを予測し，個別の生活習慣改善アドバイスを提供するアプリの開発に取り組んでいる．既存の研究では，ユーザの生活習慣や健康状態を基にしたアドバイスの提供や，適切な運動や食事の推薦が行われているが，ユーザの日常生活での意識や欲求に基づく生活習慣の改善のサポートには至っていない．

本研究では，生活習慣病の基本戦略となる運動不足の解消と過食の防止を目指し，移動中の運動の代謝量を増加させつつ，Food approach bias を軽減するための店舗推薦手法を提案する．具体的には，まず，ユーザが改善したい生活習慣病（肥満や糖尿病等）を指定すると，最短経路となる移動「時間」を基に，改善が必要な運動量に応じた代謝量を高める移動「距離」と「速度」を算出する．つまり，歩速と同じ運動時間で，運動量を高められる．また，運動量に応じて，Food approach bias を軽減するために適した店舗を選定する．Food approach bias は，食品の選択および回避のバイアスは，食べる機会を剥奪することよりも，個人の嗜好が食品選択のバイアスとして影響を及ぼすとされている [1]．一方で，公共の場所や道路上には多数の飲食店が存在し，これらを完全に避けることは非現実的であることが挙げられる．そこで，Food approach bias を軽減するための方法として，このバイアスの軽減手法として，飲食店をできるだけ避けるのではなく，個人の好みを考慮しつつ，その店舗の中で低カロリーの飲食店を推薦するアプローチが有効であると考えた．例えば，肥満度が高い人の場合，目的地まで速歩で最短経路よりも長い距離を移動することで代謝量を増加させると同時に，スイーツ好きなユーザに対しては，高カロリーのラーメン店ではなく，低カロリーのカフェ店を経由地点とする移動経路を推薦することが可能になる．このように，食品選択における個人の嗜好に基づくバイアスを考慮しつつ，健康的な選択肢へと誘導する移動中の店舗推薦手法の提案は，Food approach bias 軽減に寄与することが期待される．本稿では，Food approach bias 軽減のための移動中の店舗推薦手法を提

1：https://www.global.toshiba/jp/news/corporate/2023/01/news-0230131-01.html

案し，特に，肥満を対象として Food approach bias 軽減に応じて推薦された店舗の適切さを検証する．

本論文の構成は以下のとおりである．2 章では，生活習慣病改善に向けた機械学習と，Food approach bias の関連研究を紹介し，本研究と比較する．**??**章では，提案した経路推薦システムの構成の処理流れを説明する．4 章では，作成したシステムの評価を行う．最後に，6 章では，本研究で提案した経路推薦システムのまとめと今後の課題について述べる．

2 関 連 研 究

本章では，生活習慣病改善に向けた ehealth と機械学習，Food approach bias に関する研究と，生活支援ための経路生成に関する研究とを紹介する．

2.1 生活習慣病改善に向けた ehealth

生活習慣病改善の研究は 1980 年代から始まっており [2]，近年ではコンピュータ技術を用いたデータ分析や新しいアプローチの開発は流行している．Azusa ら [3] は教育ビデオや映画などの視覚的手法を用いた電子ヘルス介入が，生活習慣病に関する一般大衆の健康リテラシー向上に寄与することを探求している．また，Cinderella ら [4] は生活習慣病の患者向けに設計された eHealth ケア経路の開発を概説している．心理社会的，生活様式関連の困難を特定・治療するため，BCW フレームワークと共創を用いてスクリーニングツールと自己管理モジュールを設計した．

2.2 生活習慣病改善に向けた機械学習

近年，コンピュータの計算速度の高速化により機械学習手法による健康情報の解析技術が急速に発展しており，生活習慣病改善に向けた機械学習による対話や可視化の研究は活発に行われている．井口ら [5] は，生活習慣改善を促すために生活習慣の問診票と健康診断結果を組み合わせたベイジアンネットワークモデルを作成した．また，福沢ら [6] は，不健康度の可視化と室内運動支援アプリの開発により，外出して運動する習慣を促進した．この研究では，機械学習技術を用いて，ユーザの健康状態に対して反応する部分が特異である．さらに，Mitsuru ら [7] は，定期健康診断データを利用して癌以外の生活習慣病の発症を予測する新しい方法を提案した．保険請求データに基づく学習データの選択と，精度 0.32，再現率 0.89 を達成する方法について述べている．この方法は，生活習慣病の予防と早期発見に貢献する可能性がある．

2.3 Food approach bias

近年の研究では，この Food approach bias に関して深く探究されてきた．Food approach bias とは，食べ物に対して「選択」するか，あるいは「回避」するかの思考および行動バイアスのことを指す．Sercan Kahveci ら [1] の研究によれば，「選択と回避」に影響を与える要因として，食べる機会の剥奪，カロリーの高低，そして個人の嗜好が挙げられた．この研究の結果から，食べる機会の剥奪やカロリーの高低は，バイアスに影響を及ぼさないことが示された．一方で，個人の嗜好は，選択を促進する側面があり，回避のバイアスには影響を与えないことが示されている．結論として，Food approach bias は機会の剥奪よりも嗜好が選択行動に主に影響を及ぼす可能性があることを意味している．

この Sercan Kahveci らの研究の知見を踏まえて，私たちの研究は，移動時の代謝量増加と同時に認知バイアスを軽減する手法を探求している．特に，個人の嗜好や欲求を考慮し，適切な POI を推薦することで，より効果的な生活習慣病改善の取り組みが可能となると考えている．

2.4 生活支援に関する経路生成

スマートフォンの普及と先進技術の進化を背景に，リアルタイム経路生成の研究と技術開発に焦点を当てる．経路生成は，単に目的地にたどり着く手段以上のものになり得る．森ら [8] は，視覚障害者の安全かつ自由な外出を支援するためのモバイル・ウェアラブルシステムが提案されている．このシステムは障害物検知，リアルタイム経路生成，振動通知を統合しており，視覚障害者がより安心して自然に歩行できるように設計されている．また，峰ら [9] は，高齢者向けの経路探索手法に関するもので，個人の歩行履歴を利用して各ユーザの属性に合わせた適応性を高めることを目的としている．この研究は，高齢者が外出しやすい環境を整備するため，ユーザの嗜好性や身体状況を考慮した経路を提供することを目指している．

2.5 本研究の新規性

本章で紹介された eHealth の介入，機械学習による健康情報の解析，そして Food Approach Bias に関する知見は，経路生成技術の発展と健康促進の取り組みに新しい視点を提供する．本研究では，これらの技術とアプローチを統合することにより，経路生成は単に移動の最適化を超えて，個々の健康行動の改善，特に生活習慣病の予防と管理に貢献する重要な要素となる．具体的には，経路生成システムは，個人の健康状態と食に対するバイアスを考慮して，最適な移動ルートを提案することができる．これにより，ユーザーは日常生活の中で自然に運動量を増やし，健康に有益な食選択を促進されることが期待される．たとえば，経路生成システムは，食べ物に対する個人の嗜好を考慮して，健康的な食事オプションを提供する POI（興味のある場所）を推薦することができる．また，移動時の代謝量を増加させるルートを提案することで，生活習慣病のリスクを減少させることも可能である．

3 代謝量増加と Food approach bias 軽減による経路生成

本研究では，生活習慣病の程度に応じた代謝量と Food approach bias を考慮した経路推薦システムを提案した．ユーザは個人情報を入力し，改善したい生活習慣病を選択した後，システムがユーザの BMI 指数を算出する．ユーザが肥満であると自己申告しなくても，BMI が肥満の範囲に該当する場合は，肥満とみなされる．

次に，ユーザは出発地と目的地を入力し，サーバ側は最短経路を生成する．最短経路とは，ユーザが普通歩行の状態で出発地から目的地まで到達する経路のことを指す．サーバ側は最短経路の運動時間を基準に，同じ歩行の運動時間で速歩きやジョギングによる運動距離を算出し，BMI 指数の値によって，運動方式を決定する．そして，最短経路の出発点と目的地点を含む矩形区域から店舗情報などの POI を抽出し，POI のカテゴリ（カフェやラーメン屋など）を抽出する．

さらに，レビュー情報を参照して各店舗の平均カロリーを計算し，それに基づいて POI のカロリーカテゴリを決定する．カロリーが低い店舗とユーザにとっての嗜好性が高い店舗をシステム画面に提示し，ユーザが POI を選択する．最後に地図上に，POI を経由する経路を異なる色で表示し，ユーザに推薦するルートを提供する．また，選択された POI のカテゴリをフィードバックとしてシステムに反映し，次回の推薦時に類似カテゴリの店舗を嗜好性の高い店舗として推薦する．

3.1 レビューに基づく店舗ごとのカロリー算出

本研究では，レビュー情報を基に店舗ごと，および食品ジャンルごとの平均カロリー値を算出する手法を開発した．

まず，Google Place API を利用して，対象とする店舗の経緯度とレビューなどの情報を収集した．収集した情報は，テーブル POI に格納される．このテーブルには，店舗 ID，経緯度，レビュー文などの情報が含まれる．次に，テーブル POI に格納されたレビュー情報から，メニュー情報を抽出する．この抽出作業は，特定のキーワードやフレーズを基に自動で行われる．メニューが特定できたレビューについては，テーブル POI の該当するメニュー列に+1 を加算し，特定できなかった場合は，その列を NULL 値のままにする．店舗ごとの平均カロリーは以下の式に基づいて計算される．

$$\text{店の平均カロリー} = \frac{\text{メニュー毎の出現回数} \times \text{メニュー毎のカロリー}}{\text{メニュー数の総和}}$$

さらに，テーブル POI の情報を基に，新たなテーブル POI-data を作成した．このテーブルには，各店の ID，位置情報，料理ジャンル，店名，カロリーなどが含まれる．同じジャンルの店での平均カロリー計算は，以下の式により行う．

$$\text{同ジャンル店の平均カロリー} = \frac{\text{同ジャンルのカロリーの総和}}{\text{ジャンルが同じ店舗数}}$$

この計算により，店舗ごと，およびジャンルごとの平均カロリーを算出し，食品の種類別に健康的な選択肢を提供することが可能となる．

食品のジャンル（スイーツ，和食，中華など）は，レビューテキストの内容分析を通じて決定された．この分析では，特定のキーワードやフレーズがジャンルの決定因子として用いられる．例えば，「抹茶」，「寿司」，「餃子」といった単語が含まれるレビューはそれぞれ，スイーツ，和食，中華と分類される．

3.2 Food approach bias を利用する POI 抽出

生活習慣病の改善には，健康的な食生活の促進が重要である．

図 1　Food approach bias による POI 推薦

Food approach bias とは，特定の食品に対する自動的な注意の傾向を指し，このバイアスを理解し，管理することで健康的な食選択を促進することができる．本研究では，個々の食品選択の傾向（Food approach bias）を分析し，それに基づいて個人に合わせた健康的な食事選択を提案するシステムである．

本システムは，ユーザの過去の食品選択データを収集し分析することで，ユーザに偏っている高カロリー食品から健康的な食品へと導く推薦を行う．具体的には，甘いデザートなどの高カロリー食品を好むユーザには，カロリーが低い代替品を提案する．

POI の評価値の算出に考慮される主要な要素は，個人の BMI 指数，食品の好み，そして過去の選択パターンである．このプロセスは，単に健康的な生活を推奨するだけでなく，ユーザの好みを満たしつつも，より健康的な食事選択を促すことに重点を置いている．まず，個人の BMI 指数と食品の好みを分析する．この分析を通じて，ユーザの体格指数（BMI）と好みの食品が把握され，それを基に健康的な食生活へと導くための具体的な提案が行われる．ここで重要な点は，単に健康的な選択肢を提案するだけでなく，ユーザの好みや生活様式に合致した選択肢を提供することである．次に，過去の選択パターンを分析する．ユーザが以前に選んだ食品の傾向を把握し，それに基づいて，ユーザの好みを考慮しながらも健康的な食品へと導く推薦を行う．このステップは，ユーザの既存の嗜好に対する洞察を深め，それに応じて健康的な代替品を提案することに重点を置いている．

これらの要素を総合することで，各ユーザーに適した，好みを満たしつつも健康的な食事選択を提案する POI の評価値が算出される．このアプローチは，ユーザの個々のニーズと嗜好を考慮しながら，より健康的な食生活へと導くための効果的な方法となっている．

図 1 には，ユーザの個人的な嗜好に基づいて食品のカロリー摂取を最適化する推薦システムの流れを図る．ユーザーがコスメに関心が高い場合，ラーメン屋ではなく，ドラッグストア

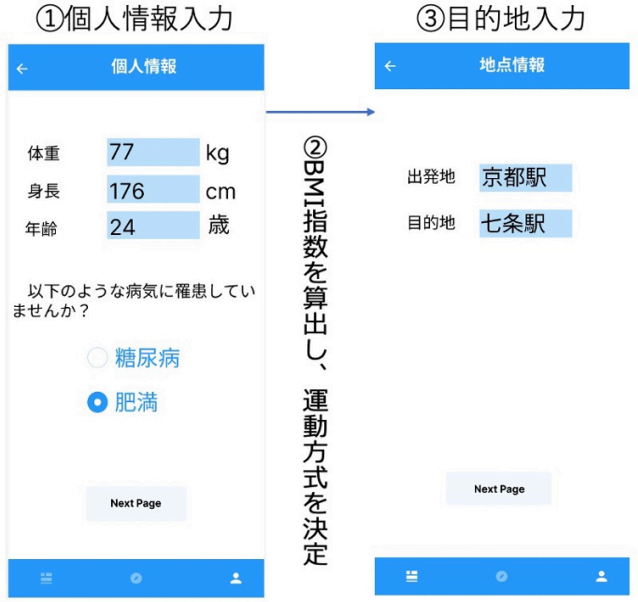


図 2　経路推薦アプリ：I. 入力画面（基本情報と出発地・目的地を入力）

を推薦する．スイーツ好きなユーザには，カロリーが低めのカフェを推薦し，健康的な選択肢を提供する．このアプローチは，個人の好みと健康的なライフスタイルをバランスよく統合し，POI 推薦の新しいアプローチを提案している．

4　生活習慣病改善に向けた経路生成の実装

代謝量増加と Food approach bias 軽減による経路生成システムを実装した．Flutter プラットフォームを用いて Dart 言語で開発した．最短経路および経由地の POI を経由する最短経路生成には Google Map API を用いた．

4.1　経路推薦アプリケーション

図 2 にユーザの入力画面のプロトタイプを示している．ユーザから入力された個人情報と地点情報により，BMI 指数を算出し，それに応じて適切な運動方式を決定する．具体的には，BMI 指数を基準にして，歩行，速歩，ジョギングの 3 種類の運動方式から最も適したものを選定する．その後，選定された運動方式に基づいて経路を推薦する．

4.2　距離算出と POI 抽出

本システムは，ユーザの選択行動から学習し，個人の嗜好性に合致した店舗を推薦することを目的としている．

図 3 に距離算出と POI 抽出の過程を示す．まず，(1) 最短経路が Google Map API を使用して生成され，次に (2) ユーザの BMI 指数を基に適切な運動距離が算出される．その後，(3) 出発地と目的地の矩形内の全ての店舗情報が抽出され，各店の平均カロリーを計算する．特に，(4) ユーザの BMI に応じた低カロリーの店舗を POI として選出される．これにより，カロリーが低い店舗をユーザに優先的に提示することができる．

表 1　POI のカロリーの高さによる罪悪感への影響

	選択欲求		罪悪感	
肥満度	高い	低い	高い	低い
高カロリーな POI	1.20	2.05	1.90	1.05
中カロリーな POI	1.65	2.10	1.05	0.45
低カロリーな POI	1.55	2.20	1.05	0.50

4.3　経路生成と POI 推薦

図 4 に経路生成と POI 推薦の過程を示す．経路生成のプロセスでは，(5) ユーザの食の嗜好（甘味，塩味，和食，洋食など）に基づいて POI をカテゴリ分類し，(6) 各カテゴリ内でカロリーが最も低い POI を経由地として選出する．最終的に，(7) ユーザには選択された POI の中で，自身の運動目標に最適な上位 2 つの経路が推薦され，(8) ユーザはこの中から一つの POI を選択し，経路を生成する．初回利用時には，ユーザの選択はシステムに記録され，フィードバックとして機能するが，嗜好性に基づく推薦は行われない．すなわち，この時点でシステムは低カロリー店舗のみを提示し，ユーザの健康を考慮した経路選択をサポートする．将来的には，ユーザが再度アプリケーションを使用する際に，初回の選択に基づいて，よりパーソナライズされた推薦が可能となる仕組みが組み込まれている．ユーザがアプリケーションを再度使用する際には，初回利用時の選択がシステムによって記録されている．そのため，2 回目以降の利用時には，システムはユーザの選択履歴を分析し，低カロリーの店舗の中，ユーザの好みに合ったカテゴリの店舗を推薦の 1 位になってユーザに推薦する．結果として，ユーザは自分の嗜好性により密接に合致する店舗から選択することができ，よりパーソナライズされた経路推薦が実現する．

5　代謝量と Food approach bias に応じた POI 抽出の検討と評価

記録されたデータを活用してユーザの好みを分析する．この分析結果は，以降の POI 推薦において，ユーザの好みに合致する店舗をより多く提示するために利用される．

3 章と 4 章の提案手法に基づき，移動中の代謝量を増加させつつ，Food approach bias 軽減のための店舗抽出手法を提案した．本章では，提案した各機能に関する POI 抽出結果を検討する．

5.1　移動中の代謝量に応じた POI 抽出の検証

本稿では，訪問したいという欲求の高さと罪悪感より肥満の程度に応じて抽出した POI を検討する．

評価方法は，被験者として選ばれた 7 名の大学生に肥満程度た高い場合と低い場合を想定し，3 カテゴリから 2 店舗ずつ，合計 6 店舗の POI に対して，健康状態を基に「訪問したいか」と「罪悪感があるか」を 4 段階のリーカット尺度（0 から 3 の範囲）で評価してもらった．なお，POI 提示は店舗名と属性（ラーメン店など）のみで，カテゴリ（高カロリー等）は提示しない．

評価対象は，任意の経路上を含む領域内に存在する POI と

図 3　経路推薦アプリ：II. 距離算出と POI 抽出

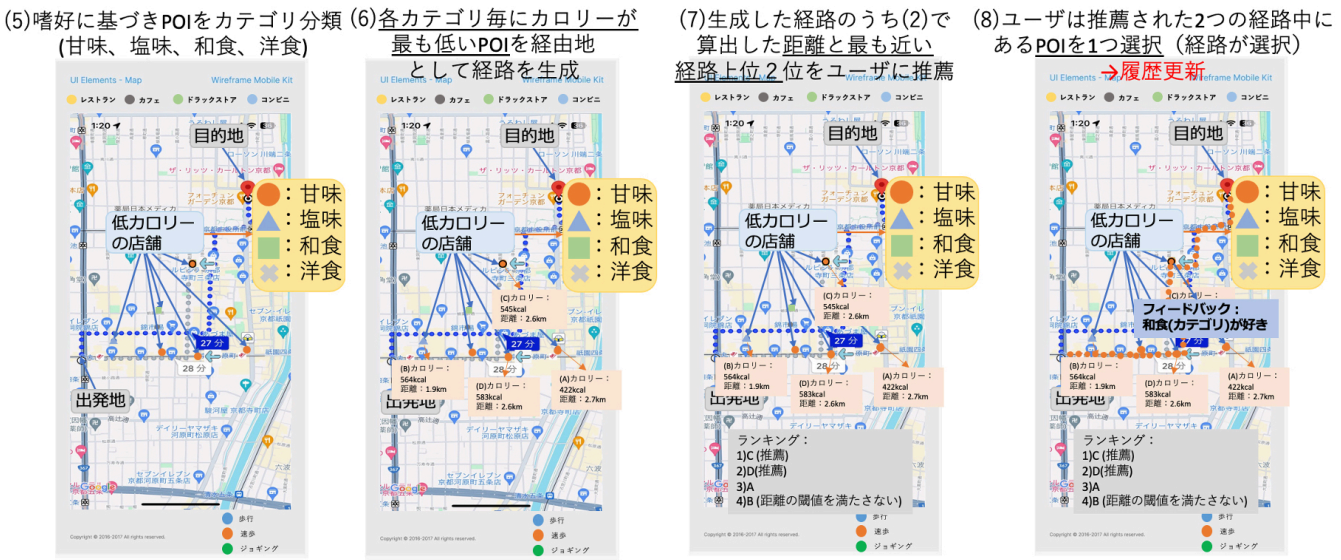


図 4　経路推薦アプリ：III. 経路生成と POI 推薦

する．POI の内訳は，カロリーが高い飲食店（ラーメン店や居酒屋など）が 9 店舗，カロリーが中程度の飲食店（うどん屋やそば屋など）が 3 店舗，カロリーが低い飲食店（コンビニやカフェなど）が店舗，その他の飲食店が 24 店舗である．

表 1 に結果を示す．肥満度が高い場合では，最も訪問欲求の高い店舗はカロリーが中程度であったが，罪悪感は最も低い結果となった．肥満度が低い場合では，最も訪問欲求の高い店舗はカロリーが低い店舗で，罪悪感は中カロリーが最も低い結果であった．また，高カロリーの POI に対する罪悪感は最も高いが，肥満度が低いときはそれが半分程度に軽減された．以上から，肥満度の程度に基づきカテゴリを決定し POI を推薦することで，罪悪感が軽減され，認知バイアスも軽減できると考察する．

5.2　Food approach bias 軽減に応じた POI 抽出の検証

本稿では，Food approach bias 軽減に応じた POI 抽出を予測するため，嗜好が選択に与える影響を検討する．

個人の嗜好は，日常の選択行動に大きな影響を及ぼす．特に，飲食に関連する選択は，健康やライフスタイルに密接に関連している．本研究では，個人の甘味，塩味，和食，洋食への偏好が，特定の状況下での食品選択にどのように影響するかを分析する．

評価方法は，被験者 20 名（大学生）が甘味と塩味の好み，和食と洋食の好みに関する質問に回答した．その後，彼らは複数の経路上の店舗をゴールした際に，食品選択を行い，その選択結果を評価した．

評価対象は，京都市内の特定の経路上の飲食店をゴールした際に提示される選択肢とする．まず，被験者は嗜好を選択する．

図 5　甘味の嗜好を選択したユーザに評価してもらう 4 種類の経路

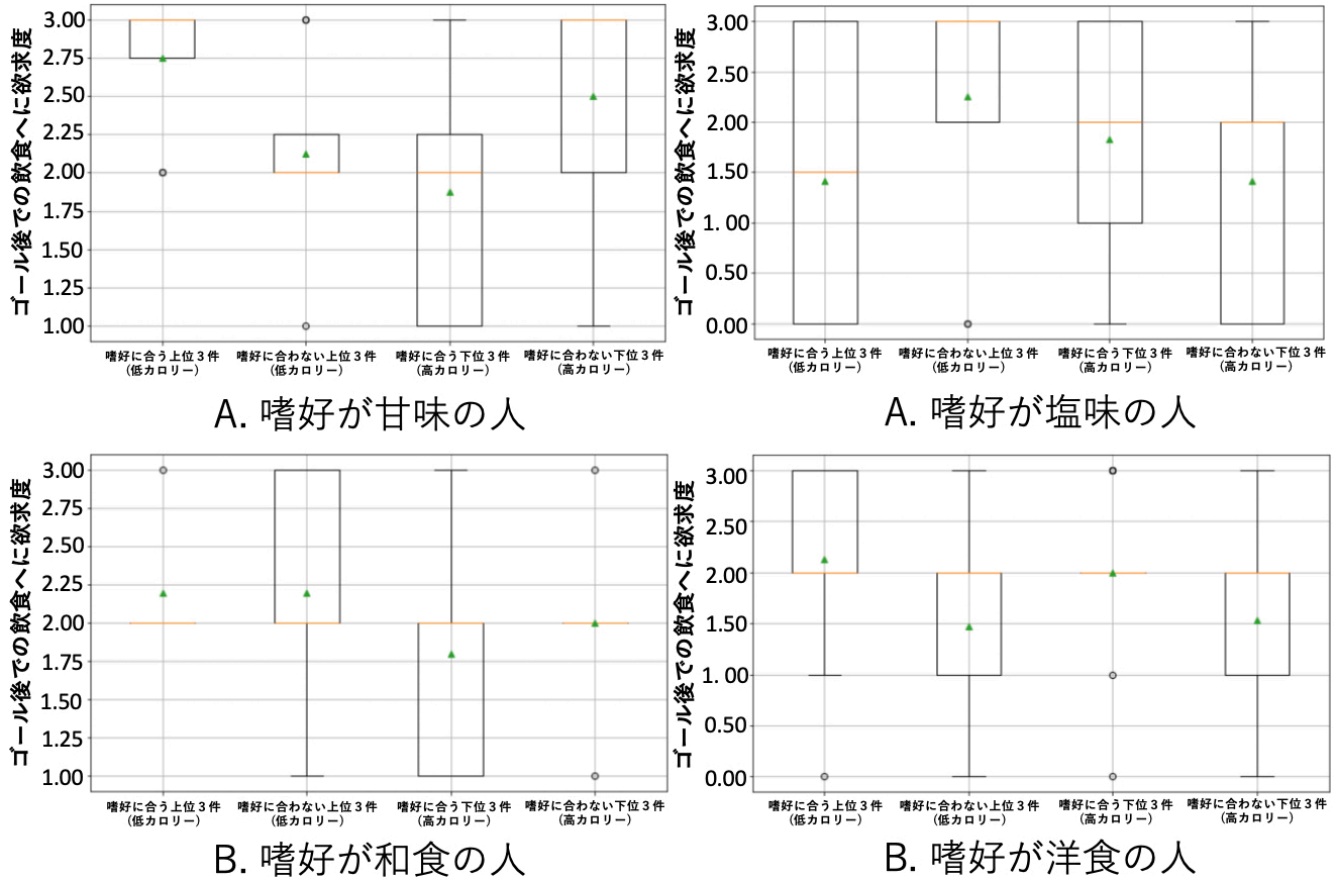


図 6　嗜好が選択に与える影響の調査

嗜好は,「甘味」または「塩味」と「和食」または「洋食」からいずれかを選択する．図 5 に，嗜好選択後に提示される 4 つの経路と経路上の店舗情報の例を示す．例では，嗜好（甘味）を選択した場合に，提案手法より嗜好が甘味で高カロリーの上位 3 店舗を通る経路 (1)，嗜好の甘味に合致しているが低カロリーとなる下位 3 店舗を通る経路 (2)，嗜好と逆になる塩味で高カロリーの上位 3 店舗を通る経路 (3)，逆嗜好の塩味で低カロリーの 3 店舗を通る経路 (4) である．

最後に，各経路と店舗情報を確認した後に,「ゴールした後に，どの飲食をしたいか」を選択してもらった．選択肢は，例えば嗜好性が甘味の場合，甘味（嗜好にあった味）の飲食をしたいか否かを 4 段階（とても欲する，欲する，あまり欲しない，全く欲しない）で評価してもらった．

図 6 に示された箱ひげ図は，個人の嗜好が食品選択に明確な影響を与えることを示した．縦軸は経路をゴール後での飲食への欲求度を，横軸は異なる店舗情報を示している．各箱ひげ図は中央値，四分位数，および外れ値を表し，選択の一貫性と個人間のバリエーションを示している．甘味を好む被験者は，甘味で低カロリーの店舗が選ばれることが多かったが，甘味下位 3 件の店舗をゴールした後に食品を食べたいという欲望が低下

することが観察された．塩味が好む被験者群に対して，同じ甘味下位 3 件（逆の味の下位 3 件）の店舗をゴールした後に食欲が著しく低下した．また，和食が好む被験者群に対して，和食下位 3 件の店舗をゴールした後に食欲の抑制が最も顕著であった．洋食が好む被験者群に対して，嗜好に合う低カロリーの店舗が選ばれることが多かったが，和食の店舗をゴールした後に食欲が低下する傾向が見られた．これらの結果は，個人の嗜好が日常の食品選択に重要な役割を果たすことを示している．嗜好は，特定の環境下での選択肢に影響を与え，それが消費者の健康やライフスタイルに直接的な影響を及ぼす可能性がある．

以上より，個人の嗜好が食品選択に及ぼす影響を明らかにし，消費者行動の理解を深めることに寄与する．また，この知見は，消費者のニーズに応じた飲食店の選択やマーケティング戦略に有益な情報を提供する．

6 おわりに

本稿では，移動時の代謝量を増加させつつ，Food approach bias を軽減するための店舗抽出手法を提案した．提案されたシステムは，歩行による運動時間を基準にし，速歩やジョギングなど異なる運動方式ごとの運動距離を算出し，代謝量を増加させる．また，最短経路付近の店舗を抽出し，カロリーを計算することで，ユーザの嗜好性に合った店舗を推薦し，経路を生成する．

今後の研究では，これらの経路に対するユーザの反応や満足度を評価し，糖尿病や心房細動に最適な運動方式の検討を行っていく予定である．

謝　辞

本研究の一部は，JSPS 科研費 JP22K19837，JP22H03700，JP20H00584 および京産大先端科学技術研究所 (M2301) の助成を受けたものである．ここに記して謝意を表す．

文　献

[1] Sercan Kahveci, Adrian Meule, Anja Lender, and Jens Blechert. Food approach bias is moderated by the desire to eat specific foods. *Appetite*, 154:104758, 2020.

[2] 杉本和夫. 大和高原三村住民の貧血頻度と食生活習慣の改善について. **奈医誌**, pages 361–376, 1989.

[3] Azusa Aida, Thomas Svensson, Akiko Kishi Svensson, Ung ll Chung, and Toshimasa Yamauchi. ehealth delivery of educational content using selected visual methods to improve health literacy on lifestyle-related diseases: Literature review. *JMIR Mhealth Uhealth*, 8:12 e18316, 2020.

[4] Cinderella K.Cardol, Judith Tommel, Henriët van Middendorp, Yvette Ciere, Jacob K. Sont, Andrea W.M.Evers, and Sandra van Dijk. Detecting and treating psychosocial and lifestyle-related difficulties in chronic disease: Development and treatment protocol of the e-goal ehealth care pathway. *International journal of environmental research and public health*, 18:6 3292, 2021.

[5] 井口拓己, 吉野孝, 高木伴幸, and 小池廣昭. 生活習慣病改善のための問診票と健康診断結果を組み合わせたベイジアンネットワークモデルの提案. 2022 年度情報処理学会関西支部 支部大会，G39, 2022.

[6] 福沢豊璽, 杉村博, and 一色正男. ウェアラブルデバイスを利用した生活習慣病の予防支援システムの提案. 情報処理学会第 84 回全国大会，5ZA-03, 2022.

[7] Mitsuru Tsunekawa, Natsuki Oka, Masahiro Araki, Motoshi Shintani, Masataka Yoshikawa, and Takeshi Tanigawa. Prediction of the onset of lifestyle-related diseases using regular health checkup data. The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019, 4D3-E-2-05, 2019.

[8] 森大輝, 入江英嗣, 内原正一, 荒川明宏, and 坂井修一. リアルタイム経路生成と振動通知による視覚障害者の歩行支援システム. The 26th Workshop on Interactive System and Software, 2018.

[9] 峰亮太朗 and 古川宏. 歩行履歴を活用し個人属性への適応性を高めた高齢者向け経路探索手法の検討. 情報処理学会第 82 回全国大会，2Y-08, 2020.

観光スポットに関わる口コミの感情に基づく観点の抽出

飯濱　航平†　　戸田　浩之†

† 横浜市立大学データサイエンス学部　〒 236–0027 神奈川県横浜市金沢区瀬戸 22-2
E-mail: †{d204005e,toda.hir.xg}@yokohama-cu.ac.jp

あらまし　近年，旅行先の決定や訪問順序の選定など観光の意思決定において，SNS や口コミが活用されている．一方で観光スポットに関わる口コミ情報はしばしば大量に存在し，個々の口コミが羅列された状態で閲覧することが多く，旅行先の決定など観光の意思決定をしたいと考える利用者が必要な情報を得るにはは手間がかかる．このような状況において本研究では，口コミの感情極性に着目し，感情の原因となる内容を特定することで，観光の意思決定につながる観光スポットの口コミを，効率的に閲覧するための観点が抽出できるのではないかと考えた．実験では，実際の口コミから観光客が観光スポットに対して抱いた感情の原因となる観点を，観光スポットのポジティブな面とネガティブな面に分けて抽出し，観点から得られる口コミの評価と観点の妥当性の評価を行った．

キーワード　自然言語処理，観光，テキスト分類，情報抽出，感情極性分析

1　背　　景

近年 SNS や口コミサイトが普及し，個々人の意見を閲覧することが容易となったことで，旅行先の選定など観光における意思決定においても口コミ情報の影響が大きくなっている．実際に，人々が旅行をするうえで参考にする情報として，36.9%の人がインターネットでの書込情報を，19.0%の人が SNS からの情報を活用している [1]．一方で，現状これらの口コミ情報は大量に存在するにも関わらず，整理されていない場合も多く，利用者が手に入れたい情報を得るにはとても手間のかかる作業である．例えば，旅行ガイドサイトのじゃらん net [2] における東京タワーの口コミの総数は約 3300 件存在し，大量なうえに，場合によっては重複した情報も多く存在するため，この中から欲しい情報を探すことは利用者の負担が大きい．また 1 つの観光スポットに対しても「設備」や「体験」,「アクセス」,「歴史」,「接客」など様々な情報が混在しており，必要な情報も人によって異なる．

そこで口コミが観点ごとに整理されていれば，利用者が観光スポットの知りたい情報を素早く取得でき，観光に関わる意思決定が容易になるのではないかと考えた． 例えば，観光スポットの混雑情報を知りたい人には観光地や駐車場の混雑状況のみに言及している口コミを抽出して提示し，体験に関わる情報について知りたい人には現地で体験できるアクティビティやイベントに関する口コミを抽出して提示することで，利用者が望む観点の情報をもとに訪問先の観光スポットを選定することや，複数の観光スポットを同じ観点で比較することが可能となる．

本稿ではまず，上記を満たす観光スポットの口コミ閲覧システムを示し，それを実現するために必要となる観点とは何かを定義する．そしてその定義に基づき観点を抽出する手法を示す．これにより，口コミを整理し提供可能とすることで，観光に関わる意思決定のために口コミを効率的かつ効果的に活用可能とすることを目指す．

本論文の構成は以下の通りである．2 節では関連研究について述べる．3 節では本研究の目的について述べる． 4 節では本研究で提案する観点の抽出手法について述べる．5 節では提案手法を用いた実験について述べる．6 節では結果と考察について述べる．7 節ではまとめと今後の課題について述べる．

2　関 連 研 究

観光スポットに対する口コミを効率的に取得する観点の抽出に関する研究として，小林ら [3] は観光レビューから，Google Cloud Platform(GCP) が提供する，Natural Language API のエンティティ分析 [1] を用いてエンティティ(景色，入園料など)を抽出し，分散表現に変換した後にクラスタリングをすることで，それぞれのクラスタベクトルに類義した単語から観点を自動抽出した．この手法では，レビュー文からその観光スポットの特徴となるキーワードをクラスタリングにより抽出しているため，観光スポット特有の観点 (聖地巡礼やパレードなど) を見落とすことなく，多様な観点が得られることで，観点による観光スポットの口コミに関する網羅性を満たすことができるが，観光スポットの評価については触れられていない．また観点の定義として，本研究で目的とする，口コミの閲覧を効率的に行えることは考慮されていない．したがって本研究では，これらのキーワード取得方法を参考にしつつ，ユーザからの評価に着目して観点を抽出することで，観光に関わる意思決定を効率的に行う口コミの閲覧を可能にすることを目指す．

また，観光に関わる口コミを使用した観光スポットの推薦に関する研究として，住友ら [4] は観光地における飲食店や公園といったスポットに対する口コミの感情語に基づいて，類似したスポットを推薦するシステムを構築した．口コミの感情に着目するという点で本研究と類似しているが，スポット全体を感情語により比較しているため，スポットに対して人々がどのような点に対してポジティブまたはネガティブな評価をしている

1：https://cloud.google.com/natural-language/docs/

のかについては触れられていない．本研究では，口コミの利用者が欲しい情報を，感情の原因となる表現をキーワードとして抽出することを目指す．

3 目　　的

3.1 本研究で目指す口コミ閲覧システムの概要

1 節で示したように，観光の意思決定において観光スポットに対する口コミを閲覧することは有効な手段であるが，それらの口コミは大量に存在し，整理されていないため，似た情報を繰り返し閲覧する必要があるなど，口コミを効率的に閲覧するうえでの課題がある．そのような課題を解決するために本研究では最終的に図 1 に示す，口コミ閲覧システムの構築を最終目標としている．このシステムは，口コミの利用者が任意の観光スポットを入力すると，その観光スポットの特徴を表す観点がポジティブなものとネガティブなものに分かれて表示される．これにより利用者は，対象とする観光スポットにおけるポジティブな点，ネガティブな点を一目で確認することができる．さらに，利用者が特定の観点について詳しく見たいと考えた場合は，観点を選択することでその観点を代表する口コミが表示される．このようにして，大量にある口コミの中から，利用者が効率的に観光スポットの有益な情報を得ることで，観光の意思決定をスムーズに行うことができると考える．

このような活用が期待されるシステムに求められる要件は以下の 2 点であると考える．

【要件 1】　口コミで示される観光スポットに関する話題の全体像が把握でき，目的の情報に容易にアクセスできること

【要件 2】　観光スポットに対してポジティブまたはネガティブに感じた対象が容易に把握できること

まず【要件 1】は，現状，口コミの中には情報の重複や人によっては不必要な情報が混在しているため，利用者が口コミを一件ずつ確認することが求められるなど，利用者の手間になっているという課題から設定した．口コミの全体像がわかることで，利用者が欲しい情報に容易にアクセス可能になると考える．

次に【要件 2】は，口コミの利用者は観光スポットの良い点だけでなく，悪い点の情報を含んだ口コミの閲覧を望んでいるが，大量の口コミの中に埋もれているため，効率的に情報を得られないという課題から設定した．実際にオンラインカスタマーレビューに関する調査 [5] によると，消費者が口コミを閲覧するときに本物のレビューを見極めるポイントとして,「長所と短所をバランスよく評価すること」が 47%を占めている．このことから口コミの利用者は，観光スポットの良い点と悪い点の両方の情報を含んだ口コミの閲覧を望んでいると考える．

3.2 観点の定義

まず,【要件 1】を満たす観点の定義を,「観光スポットの特徴を捉え，口コミ全体の主要な話題を俯瞰できるキーワード群であること」と定める．観光スポットの特徴を捉えたキーワードは,「混雑」や「景色」といった一般的に多くの観光スポットで抽出されると思われるキーワードだけでなく,「桜」や「ライトアップ」などその観光スポット特有のキーワードを抽出することが望ましいと考える．なぜなら，このようなキーワード群を抽出し提示することができれば，口コミの利用者が観光スポットの特徴を一覧することが可能になると考えたからである．

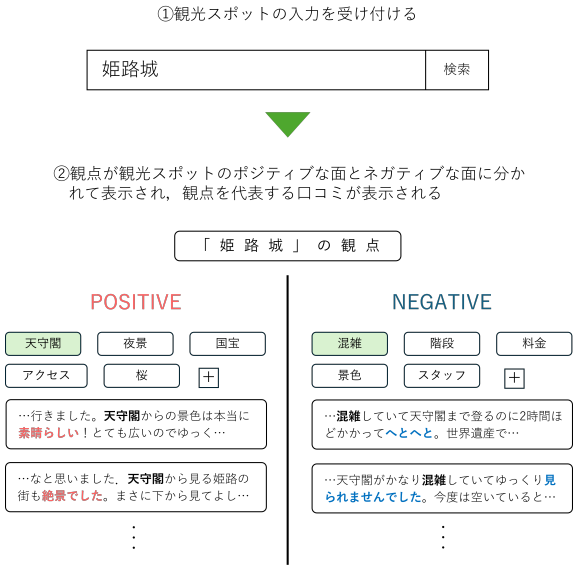


図 1: 本研究を活用した観光口コミ閲覧システムのイメージ

次に,【要件 2】を満たす観点の定義を,「一目で観光スポットの良い点，悪い点が理解できるキーワード群であること」と定める．キーワードを提示することで観光スポットの特徴が一目でわかるというメリットは重要であるが，場合によってはキーワードから利用者が連想する内容が口コミの内容と一致しない場合も考えられる．例えば「駅」というキーワードは,「駅から近くてよい」や「駅から遠いので大変」など口コミが存在し，そのキーワードだけでは口コミの内容を推測できないものもある．そこで，そのキーワードが良い面に関するものなのか，悪い面に関するものなのかが分かれば，観光スポットの特徴が一目でわかるというメリットも持ちつつ，口コミの内容をより想起できると考えた．

4 提 案 手 法

4.1 システムアーキテクチャ

3.2 節で示した観点の定義に基づき本研究で提案する観光スポットにおける観点抽出手法の概要を図 2 に示す．本手法では，特定の観光地に対して複数の人が口コミを書き込んでいる口コミの集合から，旅行に関わる意思決定をしようとする人が望む情報を効率的に取得可能にするための観点を抽出する．

まず，3 節でポジティブまたはネガティブの原因が明確にわかることを観点の定義として定めたことに基づき，得られた観光スポットの口コミに対して感情極性分析を行い，文の感情極性がポジティブかネガティブに分類されたものだけを抽出する．そして，ポジティブもしくはネガティブな感情を感じる原因は，

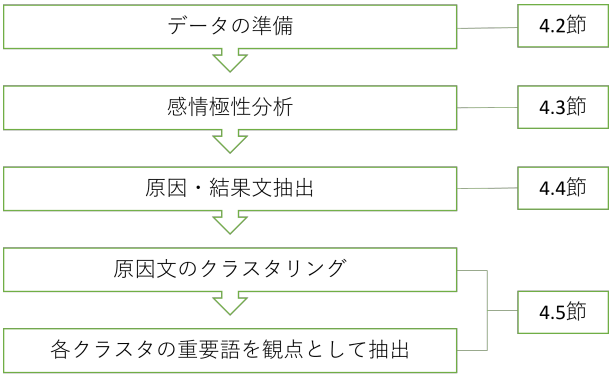


図 2: 提案する観点抽出手法の概要

文中で感情の表現と因果関係を持つと考えられるため，因果関係分析を行い感情の原因となる表現を抽出する。つづいて，3 節で示した観光スポットの特徴を一目で理解するキーワード群を観点として抽出するために，感情の原因となる表現に対してクラスタリングを行うことで類似した原因をまとめる．そして，最後に各クラスタを代表するキーワードを観点として抽出する．4.2 節〜4.5 節に今回提案する観点抽出法の各段階の処理の詳細を示す．

4.2 対象とするデータの構造

本研究では，旅行サイトや SNS など，特定の観光スポットに対して複数の人が口コミを書き込んでいる口コミの集合を観光スポットごとにまとめたデータセットを対象とする．例えば，じゃらん net などの旅行サイトでは，任意の観光スポットを入力するとその観光スポットに実際に訪れた人の口コミ一覧が表示される．それらの口コミを Web スクレイピングなどの手法を用いて収集する．つづいて，一つの口コミの中に複数の話題，特にポジティブな内容とネガティブな内容の両方を含む口コミを別で分析するために，句点または感嘆符を区切り文字として文単位のデータを生成する．

4.3 感情極性分析

本手法では，東北大学の乾研が公開している日本語 BERT の事前学習モデル [2] をファインチューニングすることで感情極性分類モデルを作成する．感情極性分析には BERT の他にも感情極性辞書を用いた手法 [7] も存在するが，ラベル付きデータが少量で済むことや，文章の意味を考慮できる点で BERT を採用する．学習データは 3 種類 (POSITIVE,NEGATIVE,NEUTRAL) の感情にラベル付けされたデータセット [3] を使用する． このモデルにより文章に対して 3 種類の感情の確信度がそれぞれ与えられ，その中で最も確信度の高い感情をその文章に含まれる感情として分類する．分類された口コミの中から，POSITIVE または NEGATIVE に分類されたものかつ，確信度が 0.7 以上の口コミのみ抽出する．なお，この閾値は予備的な検証の結果，観光スポットの良い点，悪い点の情報を含む文を抽出するための閾値として設定した．

2：https://github.com/cl-tohoku/bert-japanese
3：https://github.com/megagonlabs/jrte-corpus

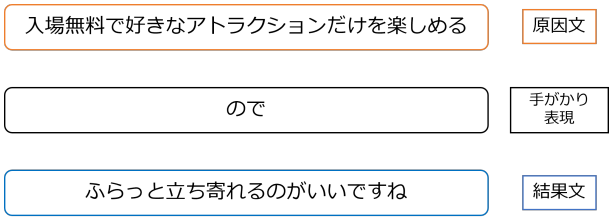


図 3: 因果関係が成立している文の例

4.4 手がかり表現をもとにした原因・結果文抽出

感情の原因となる文の抽出は，図 3 に示すように因果関係が成立する文の結果文に感情が含まれており，原因文にその感情の原因が含まれているという仮定のもと行う．因果関係が含まれる文の抽出は，坂地ら [8] の手がかり表現をもとにした因果関係知識の獲得に関する手法を参考にする．手がかり表現とは「ため」や「ので」などの因果関係の存在を明示的に示す単語を指し，手がかり表現の周辺には因果関係が成立する文が存在する可能性が高い．また坂地らは別の論文 [9] において，原因表現は手がかり表現が含まれている文の中で，手がかり表現の直前に出現することが極めて高いことを示している．

したがって本研究では,「ため」「ので」「から」を手がかり表現として定め，原因・結果文の収集を試みる．具体的には形態素解析器の MeCab [4] を用いて，文の形態素解析を行い，手がかり表現が接続詞と判定された文を因果関係が含まれる文として抽出する．そして手がかり表現の前に含まれる文を原因文，後に含まれる文を結果文として収集する．また，結果文に対して再度感情極性分析をかけ，文の中に感情を示す内容が含まれていない場合は感情の原因となる文章が抽出できていないとして除外する．

4.5 原因文のクラスタリングと観点の抽出

上記の処理を通じて得た，ポジティブもしくはネガティブな感情の代表的な原因を明らかにするため，原因文をクラスタリングし，それらの主要な原因を一目で理解可能とするキーワードを観点とする．この原因文のクラスタリングには BERTopic [10] を用いる．BERTopic は，テキストデータから得られる文章ベクトルをクラスタリングし，クラスタ毎に重要単語を抽出する一連のテキスト分類手法であり，文章の意味を考慮したテキスト分類が可能である．また，クラスタ毎に重要単語を抽出して，そのクラスタに代表される単語を付与できるため，最終的な観点候補語の抽出が可能になる．BERTopic は具体的に以下の手順で処理を行う．

I　テキストデータの文書ベクトルを取得

II　ベクトルの次元数削減

III　クラスタリング

4：http://taku910.github.io/mecab/

IV　クラスタ毎の重要語抽出

BERTopic の利点としてステップごとに任意のモデルを組み合わせることができるため，様々なデータに適用できる柔軟性がある．

本研究ではまず，テキストデータの文書ベクトルを取得するために Sentence-BERT の事前学習済み日本語モデル [5]を使用する．Sentence-BERT は BERT と同様に，文章間の意味を考慮することができるため，文章間の類似性や，クラスタリングなどの文書分類に有効である点で採用する．

次に，ベクトルの次元削減には UMAP [11] を使用する．Allaoui ら [12] は，UMAP により高次元のベクトルを次元削減することで，k-Means や HDBSCAN [13] などのクラスタリングアルゴリズムの精度を高められることを示している．Sentence-BERT により得られる文章ベクトルは高次元であるため，UMAP による次元削減を行うことで，クラスタリング精度の向上が期待できる．

また，クラスタリングには HDBSCAN を使用する．HDBSCAN は密度ベースのクラスタリングモデルである DBSCAN [14] の拡張モデルである．DBSCAN は単一の密度閾値によりクラスタを生成しているのに対し，HDBSCAN は密度が異なる場合にも対応できる点が特徴である．本研究ではクラスタごとに密度が異なるデータにも対応できる点で HDBSCAN を採用する．

そして，クラスタ毎に c-TF-IDF [15] を用いて単語分布を求めることで，重要語を抽出する．c-TF-IDF は文書単位で単語頻度を求める TF-IDF [16] と異なり，クラスタ単位で単語頻度を求める．具体的に TF-IDF は，

$$TF\text{-}IDF(t,s) = \text{文書 } s \text{ における単語 } t \text{ の出現頻度} \times \log\left(1 + \frac{\text{全文書数}}{\text{単語 } t \text{ が出現する文書数}}\right) \quad (1)$$

で表されるのに対して,c-TF-IDF は,

$$c\text{-}TF\text{-}IDF(t.c) = \text{クラスタ } c \text{ における単語 } t \text{ の出現頻度} \times \log\left(1 + \frac{\text{全文書数}}{\text{全クラスタの中で単語 } t \text{ が出現する回数}}\right) \quad (2)$$

と表され，クラスタ特有の単語に対して値が大きくなる．この値が最も高い単語を観点として抽出する．

5　実　　験

5.1　データセット

実験に用いる口コミデータはじゃらん net に掲載されている観光スポットから取得した．観光スポットのリストと観光スポットごとに得られた口コミの数を表 5.1 に示す．　観光スポットは，口コミの数が比較的多く，遊園地や美術館，歴史的名所など，他と重複のない幅広いジャンルの観光スポットを選定した．口コミの数が比較的多い観光スポットを選定した理由は，利用者が欲しい情報が含まれる口コミを探すことが難しいと考えられるからである．また，幅広いジャンルの観光スポットを分析することで，提案手法の汎用性を評価する．

表 1: 実験に使用する観光スポット一覧

観光スポット	口コミの数
兼六園	4070
東京タワー	3330
日光東照宮	3695
姫路城	4130
箱根ガラスの森美術館	2946
野見山動物園	2220
よこはまコスモワールド	2729
八景島シーパラダイス	3310

5.2　比 較 手 法

評価では，2 節で示した小林らの研究 [7] を，感情極性分類モデルによってポジティブな感情を含む文とネガティブな感情を含む文に分類した 2 つの文集合をそれぞれに適用した結果を比較手法とした．　小林らは口コミをキーワード毎に分割してクラスタリングをしているため，観光スポットに対する評価については考慮されていない．本研究の，観光に関わる意思決定を促進するための観点抽出というタスクにおいて，観点となるキーワードの抽出手法を評価する．

5.3　評価手法と基準

本研究では以下の 2 つの評価手法を用いる．

5.3.1　観点の妥当性の評価

実験によって抽出された観点リストと人手で抽出された正解の観点キーワードリストを使用して適合率と再現率を求めることで，観光スポットの特徴を捉えた観点が正しく抽出されているかどうかを評価する．

まず評価で活用する正解観点リストの作成法を示す.

(i)　両手法を用いて観点をそれぞれ最大 10 個取得する

(ii)　口コミを観点毎に 3 件取得する（観点キーワードと口コミのベクトル間類似度上位の 3 件を採用）

(iii)　(ii) で得られた口コミから次に示す基準で観点キーワードを取得し，取得された全てのキーワードを正解観点リストとする

人手による正解の観点キーワードの選定基準は，3.2 節の観点の定義に基づき,「観光スポットのポジティブまたはネガティブの対象を表し，観光スポットの具体的な特徴を示すものとする」．例えば,「駐車場は離れた場所にあるので歩くのが大変でした。」という口コミでは,「駐車場は離れた場所にある」が感情の対象となっており，その中でも具体的な特徴を示している「駐車場」を正解の観点キーワードとして抽出する．

こうして得られた正解の観点キーワードリストに元に，比較

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

手法並びに提案手法で抽出された観点リストとの適合率と再現率を求め，観点の妥当性の評価をする．

5.3.2 観点から得られる口コミの内容の評価

3.1 節で示した，本研究で最終的に目指す口コミ閲覧システムに向けて，観点に紐づく口コミがシステムの要件を満たすものかどうかを評価する．評価の判定基準は，システムの要件に基づき，

(1) 観光スポットの特徴を正しく捉え，観光の意思決定において有益な情報かどうか

(2) 観点を表すキーワードに関連する情報かどうか

(3) ポジティブまたはネガティブについて言及しているかどうか

の 3 つを評価の判定基準とする．また，1 つ目の基準は,「休日は混んでいたので平日に行くといいでしょう」などの憶測を含んだ口コミや,「ずっと行ってみたかったので行けてよかった」などの評価した人の感想や抽象的な口コミ,「雨だったので次は晴れている日に行きたい」などの願望を含んだ口コミに対しては，有益な情報かどうかの確信度が低いため，今回の評価基準には満たさないものとする．具体的かつ，事実が述べられている口コミのみを正解とする．

そして，実験によって得られた観点の集合の中から，3 つの評価基準をすべて満たす観点の割合を算出し，結果を考察する．また評価に用いる口コミは，観点を示すキーワードとのコサイン類似度を計算し，値が高い上位 3 件の口コミを使用する．

6 結果と考察

6.1 前 処 理

4 節で示した提案手法のうち，クラスタリングまでのデータ処理により，分析の対象とした口コミデータの数がどのように変化したのかを表 2 に示す．カラムは左から「観光スポット名」,「口コミを句点，記号で区切ったセンテンス数」,「ポジティブまたはネガティブに分類されたセンテンス数」,「因果関係が成立しているセンテンス数」,「結果文にポジティブまたはネガティブの感情が含まれているセンテンス数」である．

まず表 2 から，感情極性分析により約 7 割のセンテンスがポジティブに分類されていることがわかった．一方でネガティブに分類されたセンテンスは，全体の 1 割にも満たなかった．このことから，多くの人が口コミに観光スポットの良いところを書いている，または口コミの数が多い有名な観光地にはネガティブに思うところがそもそも少ないのではないかと考えられる．また，原因・結果文抽出により，ポジティブなセンテンスは 1/7 程度に減少し，ネガティブなセンテンスは 1/4 程度に減少することがわかった．因果関係を示す口コミには,「トイレが汚くてがっかりした」など，今回使用した手がかり表現を含まないものも存在しているため，見逃してしまったデータが多く存在する可能性がある．抽出された因果関係が成立する文を学習させることで，手がかり表現を拡張し，より多くの文を抽出することを今後の課題とする．

6.2 比較手法で抽出された観点

比較手法で抽出された観点を表 3 に示す．今回使用したテキストデータは，提案手法で用いた表 2 における「ポジティブまたはネガティブに分類されたセンテンス数」と同じデータである．

表から読み取れることとして，まず，比較手法によって抽出された観点は，それぞれの観光スポット特有の単語が抽出されたことがわかった．例えば兼六園（ポジティブ）の「庭園」や，姫路城（ネガティブ）の「階段」は，他の観光スポットにはあまり見られない観点である．比較手法では口コミをキーワードごとに分割したものをクラスタリングしているため，観光スポットごとに特徴を捉えた観点が得られたのだと考えられる．

6.3 提案手法で抽出された観点

提案手法で抽出された観点を表 4 に示す．

まず，ポジティブに分類された口コミで抽出された観点は，日光東照宮の「朱印」や，八景島シーパラダイスの「ショー」など，比較手法では抽出されなかった観点が数多く抽出された．実際の口コミでも,「ご朱印帳は, 頼んでおくと, 見学後に受け取れるので, とても便利でした。」や,「季節ごとにショーの内容が変わるので何度行っても飽きないです。」などの口コミが存在しており，ポジティブな感情に至った原因となる単語を抽出できていることがわかる．

一方で，兼六園の「砂利」や，日光東照宮の「工事」など，一見するとポジティブな感情の原因とは思えない観点も抽出された．実際の口コミでは,「砂利道も多いので、歩きやすい靴で行かれることをお勧めします。」や,「工事をしている箇所があったので、完成後にまた行きたいです。」などの口コミが存在しており,「おすすめ」や「また行きたい」といった表現によりポジティブに分類された可能性があると考えられる．

そして，ネガティブに分類された口コミでは，わずかしか観点が得られなかった．理由としては，ネガティブな口コミが少なく，多くのクラスタに分類できなかった事が考えられる．また,「工事」や「雨」などの観点は複数の観光スポットで抽出されており，人々が観光スポットに対してネガティブに感じる対象は，ポジティブに感じる対象よりも限定的である可能性があると考えられる．

6.4 観点の妥当性の評価結果

5.3.1 節で示した評価手法に基づいた観点の評価結果を表 5 に示す．また，表中の n は提案手法もしくは比較手法で抽出できた観点の数（最大 10 件）であり，N は正解の観点キーワードリストの観点の数である． また，今回評価に使用した観光スポットは,「東京タワー」,「日光東照宮」,「八景島シーパラダイス」の 3 件である．

結果から，ほとんどの観光スポットで提案手法の方が，適合率，再現率ともに高いことがわかる．このことから，提案手法

表 2: セクションごとのデータの変化

観光スポット	センテンス数	センテンス数 (P/N)	因果あり (P/N)	因果ポジネガあり (P/N)
兼六園	13071	8790/537	1226/142	1121/125
東京タワー	9903	6556/373	791/67	717/59
日光東照宮	13071	7583/978	1053/224	965/198
姫路城	14524	9117/832	1289/236	1170/219
箱根ガラスの森美術館	10022	7206/396	840/90	765/75
野毛山動物園	7331	4844/193	961/59	896/49
よこはまコスモワールド	7979	5263/232	1179/50	1108/41
八景島シーパラダイス	11235	7630/679	1536/146	1426/132

観光スポット	観点
兼六園	景色，冬，庭園，魅力，散策，風景，夏，紅葉，見物，晩秋，木々，印象，見学
東京タワー	イベント，ホテル，景色，眺望，夜景，絶景，店，建物，名所，夕方，夜，昼間
日光東照宮	ホテル，ツアー，おみやげ，意味，景色，温泉，風景，本殿，境内，神社，修繕，見学，工事
姫路城	イベント，ホテル，景色，屋根，壁面，天井，天守閣，城郭，見学，建物，展示，快晴，曇り空，晩秋，大勢，人々
八景島シーパラダイス	アトラクション，テーマパーク，水族館，入場，宿泊，園内，魚，エサ，釣り

(a) ポジティブ

観光スポット	観点
兼六園	入口，景色，曇り空，木陰，日陰，茶店，庭園，園内，敷地，通路，桜，花，木々，雨，雪，砂利
東京タワー	カメラ，タワー，ホテル，印象，周り，大勢，時間，時期，暇
日光東照宮	小学生，中学生，土産物，案内，見学，到着，足腰，老体，高齢，階段，石段，急坂
姫路城	トイレ，出入り口，ベビーカー，建物，外観，敷地，付近，GW，連休，週末，外国人，人達，大勢，城内
八景島シーパラダイス	アトラクション，テーマパーク，水族館，子ども，家族，母親，電車，通路，車，入園，入館，入場，店員，行列，店

(b) ネガティブ

表 3: 比較手法で抽出された観点の例

観光スポット	観点
兼六園	時間，冬，桜，夜，有名，広い，庭，抹茶，暑い，天候，砂利，園内，綺麗，紅葉，歩く，夏
東京タワー	展望台，スカイツリー，夜景，階段，土産，天気，イルミネーション，平日，夜，料金，昼
日光東照宮	工事，階段，一度，平日，歴史，修学旅行，眠り猫，朱印，見る，家康，駐車場，パワースポット，自然，世界遺産，彫刻
姫路城	修繕，桜，天守閣，城，階段，休日，景色，動物園，地元，白，周り，敷地，人，天気，天守，夏休み，国宝，靴，白鷺
八景島シーパラダイス	水族館，イルカ，アトラクション，夏休み，レストラン，ショー，平日，入園料，遊園地，魚，海

(a) ポジティブ

観光スポット	観点
兼六園	坂道，暑い，庭園，混雑，雨，夜，連休，雪，砂利
東京タワー	展望台，スカイツリー，雨，日，工事
日光東照宮	階段，日曜，工事，旅行，駐車場，紅葉，GW，陽明門
姫路城	階段，日曜，人，工事，天守，靴，城，時間，入口
八景島シーパラダイス	駐車場，混雑，海，水族館，イルカ，雨

(b) ネガティブ

表 4: 提案手法で抽出された観点の例

で抽出された観点のほうが，より観光スポットの特徴を捉えた観点が抽出できていることが示唆される．

一方で，提案手法，比較手法ともにネガティブな口コミの方が評価の値が悪いことが分かる．前節において，提案手法，比較手法のどちらもネガティブな口コミに対して観点の抽出がうまくいかなかったことを示したが，実際に抽出された観点も観光スポットの特徴を捉えたキーワードが少ないことが分かった．例えば，八景島シーパラダイスにおけるネガティブな口コミから，提案手法を用いて抽出された「海」や「水族館」といったキーワードは，そのキーワード単体で，口コミの内容を想起

表 5: 観光スポットごとの観点の妥当性の評価結果

観光スポット	手法	適合率	再現率
東京タワー (P)($n=9, N=14$)	提案手法	**0.89**	**0.57**
	比較手法	0.44	0.29
東京タワー (N)($n=5, N=9$)	提案手法	**0.60**	**0.33**
	比較手法	0.20	0.11
日光 (P)($n=10, N=17$)	提案手法	**0.70**	**0.41**
	比較手法	0.60	0.35
日光 (N)($n=8, N=17$)	提案手法	**0.75**	**0.35**
	比較手法	0.67	0.29
八景島 (P)($n=9, N=21$)	提案手法	**0.89**	**0.38**
	比較手法	0.67	0.28
八景島 (N)($n=6, N=18$)	提案手法	0.50	0.17
	比較手法	0.50	0.17

表 6: 観点から得られる口コミの内容の評価結果 (比較手法)

観光スポット	口コミの数	有益な口コミの割合
東京タワー (P)	36	0.83
東京タワー (N)	21	0.76
日光東照宮 (P)	39	0.69
日光東照宮 (N)	66	0.92
八景島シーパラダイス (P)	27	0.70
八景島シーパラダイス (N)	27	0.93
平均 (P)	34	0.74
平均 (N)	38	0.87

することが難しく，観光スポットのどのような点に対してネガティブに感じているのかがわからない．ネガティブな口コミなど，口コミ数が少ない場合にどのように観点を抽出するかは今後の課題とする．

6.5 観点から得られる口コミの内容の評価結果

5.3.2 節で示した評価手法に基づいた観点の評価結果を表 6, 7 に示す．

結果から，全体的に提案手法の方が比較手法よりも，有益な口コミの割合が高いことがわかる．このことから，提案手法によって抽出された観点が，観光の意思決定をしたいと考える口コミの利用者に対して有益な口コミ情報を提供できると示唆される．

また，平均的にネガティブな口コミのほうがポジティブな口コミよりも有益であると判断された割合が高いことが分かる．ポジティブな口コミでは「景色がよかった」などの具体的にどのような景色について良いと思ったのかについては述べられていないのに対し，ネガティブな口コミでは,「工事中なので景色が残念だった」など具体的な内容について述べられており，利用者にとって有益な情報であると推測できる．

7 結　　論

7.1 ま と め

本研究では観光の意思決定をしたいと考える利用者が望む情報を得るために，観光スポットの口コミに関して感情極性に基づいて観点を抽出する手法を提案した．実験では，観光スポットの口コミを感情極性分析によりポジティブまたはネガティブに分類した．次に因果関係分析により，感情の原因となる表現を抽出し，得られた表現のクラスタリングを行った．最終的に各クラスタに代表される単語を観点として選んだ．

そして，観点に紐づいた口コミを評価する手法と，観点の妥当性を評価する手法の 2 つの手法により観点を評価した．結果から，どちらの評価手法でも提案手法のほうが良い性能を示すとことが分かったが，ネガティブな口コミに対する観点のキーワードの評価に関しては課題が残る結果となった．

表 7: 観点から得られる口コミの内容の評価結果 (提案手法)

観光スポット	口コミの数	有益な口コミの割合
東京タワー (P)	45	**0.87**
東京タワー (N)	15	**0.87**
日光東照宮 (P)	66	**0.71**
日光東照宮 (N)	24	**0.96**
八景島シーパラダイス (P)	42	**0.93**
八景島シーパラダイス (N)	18	0.89
平均 (P)	51	**0.84**
平均 (N)	19	**0.90**

7.2 今後の課題

本研究では実験で使用した 8 件の観光スポットのうち，3 件を評価に使用したが，8 件すべての観光スポットを評価することで，今回の評価が正しかったかどうかを追加で検証する必要がある．また，本研究では評価者が筆者一人だったため，今後は評価者を複数人に増やして行うことを考える．そして，ネガティブな口コミに対する観点の抽出精度は今後改善していかなければならないと考える．

そして，本研究で示された観点の抽出手法の中で，正しい感情極性に分類されなかった口コミや，因果関係が成立した文の抽出数が少なかったことから，セクションごとの改善が必要である．具体的には，観光スポットに関する口コミ専用の感情極性タグ付きコーパスの作成や，因果関係分析における手がかり表現の拡張により，観点の抽出精度の向上が見込まれると考える．

また,「景色」と「観覧車」や,「混雑」と「駐車場」など，観点に単語の意味の上下関係が存在する場合,「景色」や「混雑」などの観点を最上位概念の観点と定め，その下位概念として「観覧車」や「駐車場」の観点を示すことで，観点が構造化され，口コミの閲覧がよりスムーズに行うことができると考える．

文　　献

[1] 観光の実態と志向 (第 41 回). 公益社団法人日本観光振興協会, 2022.
[2] (株) リクルート. じゃらん net. https://www.jalan.net/, 12024.
[3] 小林らんう, 上野史, 太田学. 観光レビューを利用した観光スポッ

トの観点の自動抽出. DEIM Forum 2023 5b-3-1, 2023.
[4] 住友千将, 石野拓也, 久保洸貴, 岳五一. 口コミ情報に含まれる感情語に基づく類似スポット推薦システムの構築と実証実験. パーソナルコンピュータ利用技術学会論文誌, Vol. 14, No. 1, pp. 29–35, 2020.
[5] 酒井アルベルト. 顧客の意見を恐れるな！消費者調査で分かったカスタマーレビューに対する意識. https://www.capterra.jp/blog/2706/consumer-attitudes-towards-online-reviews, 1 2024.
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
[7] 佐藤貴俊, 高村大也, 奥村学. 分散表現を用いた単語の感情極性抽出. Technical Report 12, 東京工業大学, sep 2016.
[8] 坂地泰紀. 構文パターンを用いた因果関係の抽出. 言語処理学会第 14 回年次大会論文集, 2008, pp. 1144–1147, 2008.
[9] 坂地泰紀, 増山繁. テキストマイニングによる因果関係抽出. 自動制御連合講演会講演論文集, Vol. 54, pp. 124–124, 2011.
[10] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.
[11] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.
[12] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhak Cheriet. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. Image and Signal Processing, Vol. 12119, pp. 317 – 325, 2020.
[13] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, Vol. 2, No. 11, p. 205, 2017.
[14] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, p. 226–231. AAAI Press, 1996.
[15] ctfidf-bertopic. https://maartengr.github.io/BERTopic/api/ctfidf.html, 1 2024.
[16] 杉山聡. 本質を捉えたデータ分析のための分析モデル入門　統計モデル、深層学習、強化学習等 用途・特徴から原理まで一気通貫！ソシム, 2022.

博物館の展示物を用いたビンゴゲームによる鑑賞軸の習得

西本　海生†　　相原　健郎††,†††　　神門　典子††,††††　　荘司　慶行†††††　　山本　祐輔††††††
山本　岳洋†,††　　大島　裕明†,††

† 兵庫県立大学 情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 国立情報学研究所　〒 100–0003 東京都千代田区一ツ橋 2–1–2
††† 東京都立大学 都市環境学部　〒 192–0397 東京都八王子市南大沢 1–1
†††† 総合研究大学院大学　〒 100–0003 東京都千代田区一ツ橋 2–1–2
††††† 静岡大学 情報学部　〒 432–8011 静岡県浜松市中区城北 3–5–1
†††††† 名古屋市立大学 データサイエンス学部　〒 467–8501 愛知県名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †{ad22m050,t.yamamoto}@gsis.u-hyogo.ac.jp, ††kenro.aihara@tmu.ac.jp, †††kando@nii.ac.jp,
††††shojiy@inf.shizuoka.ac.jp, †††††yamamoto@ds.nagoya-cu.ac.jp, ††††††ohshima@ai.u-hyogo.ac.jp

あらまし　本論文では，ビンゴゲームをモチーフとした「曼荼羅ビンゴ」を実装し，使用してもらうことによって，ミュージアムを鑑賞する際に，鑑賞した展示物が，より鑑賞者の記憶に残るようなガイドアプリを提案する．本研究で対象とするのは，大阪府にある国立民族学博物館（みんぱく）である．ガイドアプリである「みんぱくガイド」にゲーミフィケーションを取り入れた．曼荼羅ビンゴを通して，ユーザに，独自の鑑賞軸に基づいた鑑賞を促す．明確な鑑賞軸を習得することで，未知の展示物を目にした際に，独自の指標から分析する能力の習得を支援する．この過程で，展示物の印象付けができることを目標とする．

キーワード　博物館，ゲーミフィケーション，鑑賞軸，みんぱく

1　はじめに

私たちにとって，博物館を訪問し，新たな知見を得る行為は決して珍しいものではない．本やインターネット上に掲載されている文面や写真からは取得できる情報が限られている．そのため，実際に博物館を訪問し，自分の目で展示物を鑑賞することで，対象への理解を深めることができる．

しかし，実際にミュージアムを訪れてみたところ，様々な問題が起こり得る．例として，展示場が想定よりも広大であることや，鑑賞者自身に時間的制約があることが挙げられる．また，博物館には非常に多くの展示物がある．この時，1 つの展示物あたりに割くことができる鑑賞時間は短くなってしまい，すべてを見て回ろうとすると，漠然とした鑑賞体験になってしまう．結果として，鑑賞後に展示物が記憶に残り難いという問題が発生する．本研究では，この問題の解決に取り掛かる．

Anderson ら [1] は，新しい知識は既存の知識から発展することから，博物館での鑑賞においては，鑑賞者の興味や知識を反映する必要性を説いている．そのため，本研究では，既存となる知識の付与を目的とする．新たな展示物を目の当たりにした際に，既存の展示物の知識との共通点を考えるように促し，思考の過程で印象付けを行う．

博物館での有意義な鑑賞体験には，博物館側の支援も重要である．奥本ら [14] は，博物館での鑑賞において，鑑賞者が新たな発見をし，ある展示物に対して自分なりの理解を促す支援の必要性を説いている．博物館では，鑑賞者向けにガイドアプリを提供していることがある．ガイドアプリ上では，空間の制約がないため，展示場には収まりきらないほどの展示物の情報を提供することができる．

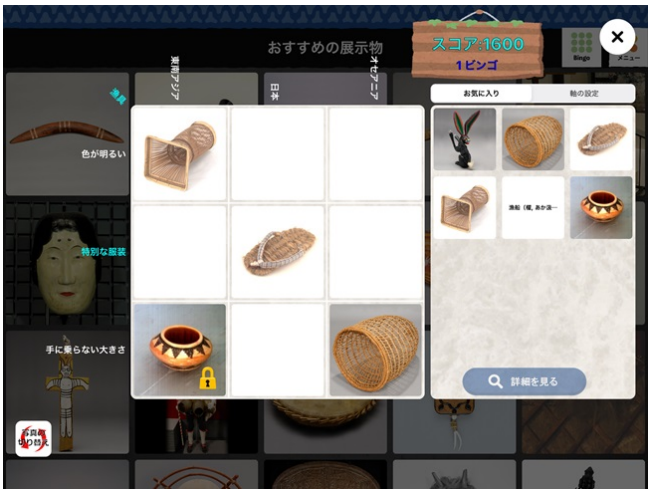


図 1　曼荼羅ビンゴのプレイ画面

本研究では，文化人類学・民族学の研究を行っている国立民族学博物館（みんぱく）を対象とする．みんぱく内は，決まった展示経路が存在しない．そのため，鑑賞者自身で鑑賞する展示物の順序を検討する必要がある．みんぱく用に，2022 年に荘司ら [19] が Apple 社の iPadOS 向けに開発，更新を行ったガイドアプリ「みんぱくガイド」を基に，ゲーミフィケーションを取り入れたビンゴゲーム「曼荼羅ビンゴ」を実装した．

本研究の新規性は，博物館での鑑賞において，各展示物の鑑賞軸に着目し，そこに位置情報及び，ゲーミフィケーションを

取り入れた点である．

2 関連研究

2.1 学習分野でゲーミフィケーションを取り扱った研究

Nah ら [9] によると，ゲーミフィケーションは，問題の解決に当たり，ゲーム的思考を取り入れることで，ユーザのモチベーションを向上させる手法である．ゲーミフィケーションは，マーケティングや政治，健康など，様々な分野で活用されている．ここでのゲーム要素とは，例として以下のようなものが挙げられる．

- リーダーボード
- レベル
- ポイント
- バッジやトロフィー
- 即自的なフィードバック
- カスタマイズ性

Hamari ら [5] によると，近年，ゲーミフィケーションは学習分野で活用される事例が増加している．また，ゲーミフィケーションを取り入れる場合は，状況やユーザによって効果が大きく左右されることが明らかになっている．

永田 [13] は，基礎学力の向上支援を目的とし，ビンゴゲームの形式でゲーミフィケーションを取り入れた．静岡産業大学で用いられている，国語，数学，理科，社会，英語の 5 教科に対して，各 6 分野の合計 30 分野で構成される「大化けドリル」を対象とする．既存の大化けドリルにおいては，利用率の低さや利用者の実施する科目のばらつき，ゲーミフィケーションの効果が薄いなどの問題点が見受けられる．改善策として，マスが 6×6 で構成されるビンゴゲームを取り入れた．上述の 30 分野に，各個人が苦手とする 6 分野を加えた合計 36 分野を用いて，不得意な分野がビンゴの中央のような，有利なマスに移動する仕組みを取り入れた．評価実験の結果，ビンゴにしたいという欲求がモチベーションの向上にはつながったものの，被験者間におけるビンゴ数のランキングは，あまり効果が見られなかった．

福山ら [17] は，小学生を対象とし，算数，国語，英語の学習が可能である，ゲーミフィケーションが取り入れられたアプリ「アプリゼミ」を用いて，評価を行った．ゲーム要素は，即時フィードバックや選択の自由，失敗の自由やストーリーラインなどが取り入れられている．評価は科目を算数としたうえで，アプリを使用する前後にテストを行い，平均点で比較を行ったほか，事後アンケートを実施した．結果は，学習分野によっては事後テストの点数が有意に高く，アンケートからは，アプリを用いることによるモチベーションの向上が見られた．

Stott ら [11] は，学習においてはゲーミフィケーションの以下の特徴が有効であるとしている．

- Freedom to Fail
- Rapid Feedback
- Progression
- Storytelling

また，Darina ら [4] によると，ゲーミフィケーションを取り扱った先行研究の多くは学習分野での利用可能性について確認しており，より実質的な実証実験が必要であるとしている．

2.2 博物館でゲーミフィケーションを取り扱った研究

荘司ら [20] は，みんぱくにおいて，みんぱくガイドの操作ログを用いることによって，鑑賞体験を鑑賞者の記憶に残すことを目的とする，事前学習と事後学習を促す仕組みを提案した．事前学習支援では，訪問前にガイド端末を操作してもらう．事後学習支援では，個人の鑑賞体験を記念品としてポストカードにすることで，事後学習を促す．実験は 16 名の被験者を対象とし，研究室内で行われた．ガイド端末のデータログに含まれた展示物がどの程度記憶に残っているのかを 4 段階で回答してもらったところ，閲覧頻度が高い展示物や，手書きのメモをとった展示物がよく記憶に残っていた．また，8 つの心理効果と 2 つのベースライン手法をもとに作成されたポストカードを作成した．各ポストカードが鑑賞体験をどの程度適切に表せているのかを順番に並び替えてもらったところ，自分がお気に入り登録した展示物や，閲覧頻度の高かった展示物からなるポストカードの評価が高い結果となった．

Ioannou ら [6] は，10 歳から 11 歳の小学生を対象とし，博物館におけるゲーミフィケーションが学習に与える影響を調査した．調査項目は，ゲーミフィケーションがもたらす小学生の内発的動機付けと外発的動機付けへの影響に加え，博物館で歴史的な分野を学習する際の学習効果についてである．実験はゲーミフィケーションの有無による 2 グループに分け行われた．ゲーム要素はシナリオという形式で取り入れられている．実験の結果，内発的動機付けに有意差は見られなかったものの，外発的動機付けにおいては，有意に効果があったことが明らかになった．これらの経時的な関係について研究の必要性を訴えている．

Jialiang ら [7] は，子どもたちを対象としたときに，今日の博物館の展示方法では情報を伝達する効率が悪いと考え，ゲーミフィケーションを用いることで娯楽性を付与した．子どもの読む能力の不足を補えるよう，家族が協力しながら進行する AR システムであり，プレイヤーが協力しながらゲームを攻略する「コラボレーションゲーム」の形式を採用した．Player 1 はアプリ内のキャラクターと会話をすることでヒントを収集し，Player 2 は AR アプリを通じて答えを博物館内から見つけ出す．実験の結果，親子で異なる役割を遂行するコラボレーションゲームの形式が有効であることが明らかになった．

Ceccacci ら [3] は，近年博物館において展示物の保存が優先されており，鑑賞者が展示物に触れる機会ができなくなっている問題を解決するために，触覚インターフェースがもたらす影響を調査した．3D プリンターを用いて複製した展示物のレプリカと，コンピュータ上で再現したデジタルモデルを比較を行った．実際に触れることで，展示物の詳細まで把握が可能な触覚インターフェースがユーザの没入感を引き立てることが明らかになった．また，併せてゲーミフィケーションの比較を行ったところ，ユーザのモチベーションの向上や，没入感に大きな効

果は見られなかった．ゲームのクリアを目的としたユーザが多かったことや，ゲーム要素に乏しかったことなどが原因として考えられ，それらを検討する必要性が示唆されている．

Klopfer ら [8] は，博物館内での学習ゲームとして，「Mystery at the Museum」を提案した．博物館における展示物全体のつながりを見てもらい，鑑賞者が今までに見たことのない展示物の提示を行う．これによる新たな発見，議論を目的とする．ゲームは 6 名で構成されるチームで行った．チーム内で役割分担がなされたことで，個々がゲームに貢献できたという意識を持つことができ，新たな発見にもつながったという結果が得られた．

前原ら [15] は，博物館のガイドアプリに導入されている，展示ルートに沿った鑑賞体験に着目した．これらの展示ルートは学芸員の専門的な知識を基に構成されているが，興味の対象は鑑賞者によって異なる．各展示物の解説文から得られる，展示物間の影響関係を用いることでパーソナライズを行うほか，展示物を取り巻く文化と展示に対するユーザの理解を深めてもらい，新たな興味を喚起することを目的としている．博物館内の展示物の解説文を解析した後に，鑑賞者の興味や滞在時間を入力することで，適切な個数の展示物を提案する．推薦された展示物には肯定的な評価が得られた一方で，精度の向上が課題となった．

大杉ら [16] は，動物園を対象とし，施設の学習の機会を提供する役割が薄れてきている問題の解決に取り掛かった．普段からミュージアムのような社会教育施設にあまり訪れない初学者にとっては，観察すべき視点が分からないことから，自発的な観察が困難であるとしている．そこで，天王寺動物園を対象とした，園内回遊支援アプリを作成する．次に訪れるべき動物はキャッチフレーズで提示し，その動物を推測しながら移動を行う．様々な動物を候補として考え，試行錯誤する過程で，他の動物へも意識を向けさせる．実験の結果，回遊アプリによって間接的興味が喚起され，広範な回遊やクイズによる自発的な観察が確認された．

また，鑑賞方法に着目した研究として，ミュージアムにおいては，テキストのみならず，音声でのガイドが行われることもあり，展示場を歩きながら解説を聞くことができる施設も存在する．その学習効果についても研究がされており，平澤ら [18] は，音声データを内蔵した PDA と iPad アプリによる学習の効果を比較検証した．そのほか，前原ら [15] や安部ら [12] は，特定のテーマに沿った展示ルートの生成を行った．

3　鑑賞軸の定義

鑑賞者の能動的な行動を促すゲームシステムとして，「曼荼羅ビンゴ」を提案する．曼荼羅ビンゴでは，鑑賞軸の習得を支援する．

3.1　鑑賞軸の習得による学習効果

鑑賞者は展示物を見る時，各々の観点に基づいている．観点とは，例えば，展示物の大きさや形，色や使用されている地域，用途などが挙げられる．この観点のことを鑑賞軸と定義する．

図 2　展示物の詳細画面

本研究では，鑑賞者が自身の鑑賞軸を認知し，応用できることを目的とする．鑑賞軸を認知することで，展示物に対して独自の指標を定めることができる．例えば，ある鑑賞者が「衣服の装飾」という鑑賞軸に基づいた鑑賞を行ったとする．この時，中国の装飾は赤色が多い，煌びやかであるなど，地域毎の特徴がある．この鑑賞者が，今まで見たことのない同様の特徴を持った展示物を見た際に，それを中国の衣服であると判断することができる．

独自の指標での分析を促すことで，鑑賞者に展示物の印象付けを行うことが可能になる．同時に，鑑賞者は博物館に訪れた際に軸を意識するようになるため，漠然とした鑑賞をしないようになる．結果として，鑑賞体験そのものをより充実したものにすることができる．曼荼羅ビンゴは，鑑賞軸の習得を支援するアプリケーションである．

3.2　曼荼羅ビンゴにおける鑑賞軸

提案する曼荼羅ビンゴでは，鑑賞者の持つ独自の軸に加え，みんぱくから提供していただいたデータの一部を鑑賞軸として扱う．みんぱくは，展示物約 80,000 点について，材料や用途，使用されている地域など，多様な情報を格納したデータベースを有している．このデータはみんぱくガイド上で用いている．図 2 は，実際にみんぱくガイド上で，みんぱく内の「女性用 首飾り」という展示物の詳細画面を開いた様子である．ここには展示物の名前や標本 ID をはじめとした，様々な情報が掲載されている．

本研究では，図 2 における OCM 及び，展示場所という項目を鑑賞軸として用いる．OCM は，Human Relations Area Files（HRAF）[1] という国際機関が作成した，展示物の分類を表すタグである．本研究で用いているものは，それらを日本語訳したものである．OCM は，Outline of Culture Materials という分類である．これは 3 桁の数字で表される．例えば図 2 において，「装飾」というタグだと「301」，「セラミック産業」と

1：Human Relations Area Files, https://hraf.yale.edu/

いうタグだと「323」となっている．また，みんぱくでは，展示物が使用されている地域に応じて，展示場所が分けられている．この展示場所も，本研究では展示物を表す鑑賞軸として取り扱う．

4　曼荼羅ビンゴの概要

4.1　各々の観点に基づいたビンゴゲーム

本節では，曼荼羅ビンゴがもたらす学習効果について検証する．ガイドアプリでは，鑑賞者はそれぞれの鑑賞軸に基づいてビンゴゲームを進行することができる．本ガイドアプリを**軸自由版**とする．

4.1.1　軸自由版の概要

図 3 は曼荼羅ビンゴの初期画面である．画面左側に 3 × 3 の 9 マスのビンゴカードがあるほか，周囲には 3 つの横軸と 1 つの斜め軸，3 つの縦軸がある．ここでは，3 つの横軸と 1 つの斜め軸を「自由軸」，3 つの縦軸を「エリア軸」としており，自由軸が OCM，エリア軸が展示場所と対応する．自由軸は，鑑賞者自身の観点に基づいて，自由に選択することができる．また，画面右側には，お気に入り登録した展示物が入る欄がある．図 2 の展示物の詳細画面で，画面右下の「お気に入りに登録」ボタンを押下することで，欄内に展示物が入る．

展示物をマスに入れるには，事前に自由軸を設定しておく必要がある．図 4 は，軸設定画面である．ユーザが 10 字以内で自由に入力ができるほか，展示物について「色が明るい」「手に乗る大きさ」「尖った形状」などのプリセットを用意した．加えて，鑑賞者がガイドアプリ上でお気に入り登録している展示物が持つ OCM も軸として使用可能とした．例えば，2 行目の横軸に「色が明るい」という軸を設定したとする．この時，対応する 1 行のマスにおいては，「色が明るい」展示物しか入れることができない．一方で，エリア軸はいずれかのマスに展示物を入れることによって，対応する軸が自動的に決まる．例えば，中央のマスに図 2 の「女性用 首飾り」という展示物を入れたとすると，2 列目のエリア軸は「ヨーロッパ」となる．この時，その上下のマスにはヨーロッパの展示物しか入れることができなくなる．自由軸及びエリア軸では，それぞれ同様の軸を複数使用することができない．これにより，幅広い展示物を鑑賞してもらう．

各展示物の OCM はみんぱくのデータベースを参照している．そのため，OCM が用いられた自由軸と対応するマスに展示物を入れた際には，自動的に展示物と軸の OCM の対応を確認する．しかし，「手に乗る大きさ」などの鑑賞者自身が自由に設定した軸は，既存のデータでは判定が難しい．この際は，図 5 のように，鑑賞者自身に再確認を促す．

ビンゴ画面では，上部に現在のスコア及び，現在のビンゴ数が「x BINGO」という形式で表示される．スコアは次の要因によって決まる．

- 展示物を入れるマスの位置
- 自由軸が「自由入力，プリセット」か「OCM」か
- 自由軸で OCM を用いた場合，みんぱく内でのその OCM の出現頻度
- 総ビンゴ数

図 3　曼荼羅ビンゴの初期画面

図 4　曼荼羅ビンゴの軸設定画面

スコアはビンゴのマスに更新があるたびに，その都度更新される．これは，福山ら [17] や Stott ら [11] の即時フィードバックを満たしている．また，Nesbitt ら [10] によると，ゲームのインタフェースにおいては，視覚，聴覚，触覚によるフィードバックが最も好まれる傾向にあるとされている．ビンゴ達成時には，エフェクトの表示を行うことにした．

鑑賞者は，普段は漠然と鑑賞してしまうような展示物でも，このタスクが与えられることによって，ビンゴのマスに入れることができるかどうかを検討する必要がある．各展示物が有している OCM は，みんぱくガイド上でしか確認ができない．そのため，鑑賞者がみんぱくガイドを確認する頻度が増え，各展示物の詳細も目にする機会を提供することができる．これにより，普段は気にすることのなかった展示物を意識することも可能となる．

また，縦軸に展示場所を割り当てている理由として，鑑賞する展示エリアの制限がある．みんぱくは展示エリアが 8,899m^2 であり，広大である．そのため，前述した，漠然とした鑑賞を行う原因となりやすい．この問題を解決するために，曼荼羅ビ

図 5　曼荼羅ビンゴの軸確認画面

図 6　エラー発生時の画面

ンゴでは，鑑賞者に入る情報量の制限を行っている．同時に，「中国の結婚式では赤色の派手な服が用いられている一方で，南アジアでは質素な服が用いられている」など，特定のカテゴリの展示物について，縦軸同士を比較することによって，地域の違いがもたらす生活様式の違いを認識できるのではないかという狙いもある．

しかし，マスに展示物を入れるだけでは，単純な作業になってしまう．これは，みんぱくガイド上で検索を繰り返すことで達成できてしまうので，印象付けにはつながり難い．これを解決する手法として，ロック機能を実装した．展示物をお気に入り登録した時点では，その展示物には鍵がかかっている．ロック状態の展示物はビンゴに使用できるものの，ビンゴ数はカウントされない．ロックを解除するには，実際に展示物の前に行き，展示物の写真を撮影する必要がある．これにより，ビンゴを単調でないものにしている．同時に，鑑賞者自身に動いてもらうことによっても，展示物の印象付けを行う．写真の撮影時には，みんぱく内に設置されている 319 個のビーコンのうち，最もユーザに近いビーコンの電波を端末が検知する．この際に，ビーコンに紐づけられた付近の展示物の情報と照合を行い，ロックが解除される．

仮に展示物を選択，鑑賞したものの，その展示物をビンゴの作成に使わなくなった際は，ビンゴのマスから展示物を削除することができる．この際，削除後も再度ビンゴ画面の「お気に入り」の欄から，同様の展示物をビンゴに用いることが可能である．これは，せっかく展示物を鑑賞したものの，展示物の検索やお気に入り登録，写真の撮影を行った一連の行為が無駄になることによって，ゲームを進行するモチベーションが下がり，結果として学習効率が下がる事態を回避するためである．

また，図 3 の左下には，「写真切り替えボタン」がある．通常は，みんぱくのデータベースに登録されている展示物の写真が表示される．しかし，このボタンを押下することで，鑑賞者自身が撮影した写真に切り替えることができる．これにより，鑑賞者が自分だけのビンゴを作り上げることができる．この機能では，Nah ら [9] が挙げていた，カスタマイズ性をビンゴに持たせている．

そのほか，軸の制約を満たしていない展示物をマスに入れようとすると，「タグが一致していません」と表示され，展示物を選び直すことができる．実際に，図 6 は，展示物を OCM が一致していないマスに入れようとした際の画面である．「タグが一致していません」というメッセージとともに，「もう一度、展示物を確認してみましょう！」と表示することによって，鑑賞者本人に「どこがどうダメだったのか」を考えさせることができる．また，再度展示物の詳細画面への移動及び確認を促すことで，ある展示物に対して能動的に行動させることができる．進行度がビンゴ数という形で視覚的にわかりやすい点も含めて，Stott ら [11] が学習分野で有効であるとしている「Freedom to Fail」「Rapid Feedback」「Progression」の要素を取り入れている．

4.1.2　軸自由版の実験手法

実験は 4 名を対象とし，2024 年 1 月 17 日に 3 名，2024 年 1 月 19 日に 1 名が行った．上述したしたみんぱくガイドを使用してもらうことで，データログの収集とアンケート，インタビューを行った．場所は兵庫県立大学のキャンパス内で，みんぱくにいることを想定して行った．実験では，より高いスコアとなるビンゴを作成してもらうことを目標としてもらった．

また，上述したガイドアプリにおいて，自由軸では OCM を最低でも 2 つ使用する制約を課した．OCM の情報は展示物の詳細画面でしか確認ができない．そのため，ガイドアプリ上で，展示物について能動的に調べてもらうことを目的としている．

実験後には，みんぱくガイドのユーザビリティ及び，鑑賞体験についてアンケートを行った．ユーザビリティの評価には，System Usability Scale(SUS) [2] を用いた．各項目を，みんぱくガイド用に置き換えている．また，鑑賞体験についてアンケートを行った．鑑賞体験のアンケートの項目は，表 1 の通りである．アンケートでは，自由記述を除き，5 段階のリッカート尺度で回答してもらった．また，当日の鑑賞体験を通して，印象に残った展示物及び，アンケートの回答の背景にある意図を明らかにすることを目的として，インタビューを行った．

表 1 （軸自由版）事後アンケートの項目

質問項目
Q1. 1 年にどの程度，博物館，美術館に行きますか？
（自由記述）
Q2. 1 日にどの程度，ゲームを遊びますか？
（自由記述）
Q3. ビンゴをすることは，勉強になりましたか？
（1: 全く勉強にならなかった，5: とても勉強になった）
Q4. ビンゴをすることで，新たな発見はありましたか？
（1: 全くなかった，5: とてもあった）
Q5. ビンゴは楽しかったですか？
（1: 全く勉楽しくない，5: とても楽しい）
Q6. ビンゴは難しかったですか？
（1: とても簡単，5: とても難しい）
Q7. ビンゴをすることで，達成感を得られましたか？
（1: 全く得られなかった，5: とても得られた）
Q8. またビンゴをしたいと思いますか？
（1: 全く思わない，5: とても思う）
Q9. 展示物の詳細について，注目することはできましたか？
（1: 全くできなかった，5: とてもできた）
Q10. 展示物のどのような点に注目することができましたか？
（自由記述）
Q11. 実際に展示物をみんぱくで見たいと思いましたか？
（1: 全く思わなかった，5: とても思った）
Q12. 特に良かったところと改善点を教えてください
（自由記述）
Q13. 本日の鑑賞体験は有意義でしたか？
（1: 全く有意義ではなかった，5: とても有意義だった）

4.1.3 軸自由版の実験結果

実験時には端末上の操作ログを収集した．表 2 は，被検者毎のログデータである．ログは，被験者がアプリを使用している間に行われた操作の回数を表している．なお，展示物の詳細画面を開いた回数に関しては，詳細画面に掲載している情報量から，3 秒以上画面を開いていた場合を対象としている．全ての被験者が，展示物の詳細画面を開く回数が多かったことが分かる．アンケートでは,「ビンゴをすることは，勉強になりましたか？」という質問に対して，平均が 4.00 であった．実際にインタビューでは,「縦軸にアイヌが入ってて，横軸に石工が入ってたんですよ．それ調べても一個も出てこなくて．だからアイヌの文化って石を使って何かを作るっていうのよりも，木を使って何か作る方が多かった.」と，検索結果に何も「表示されない」ことが新たな考察につながっており，鑑賞軸をもとにした展示物の印象付けが達成されていることを確認できる．

一方で,「展示物の詳細について，注目することはできましたか？」という質問では，平均が 3.75 であった．インタビューでは，念頭にあるのはビンゴの達成であり，詳細にはあまり触れる余裕がなかったことが原因であったと回答した被検者がいた．

そのほかの評価として，ビンゴの難易度の平均は 4.00 という評価が得られた．ビンゴをした際の達成感については，平均が 4.50 という評価が得られており，難易度，達成感ともに高い評価であることが分かる．また,「またビンゴをしたいと思いますか？」という，アプリケーション利用の反復性に関する質問

表 2 （軸自由版）事後アンケートの項目

被検者	1	2	3	4
詳細画面の表示	164	136	194	83
ビンゴ画面の表示	68	50	120	50
展示物の検索	12	19	7	10
共通の特徴を持つ展示物の検索	18	3	39	19
お気に入り登録	21	18	24	17
トータルスコア	22,680	21,600	31,752	31,572

表 3 （軸自由版）被験者毎の SUS の評価

被検者	1	2	3	4
スコア	82.5	82.5	72.5	62.5

ついては，平均が 4.50 となっている．ビンゴの楽しさについては，全被験者が 5.00 と回答しているほか,「ビンゴについて，特によかったところ」として，得点の可視化が挙げられており，より高いスコアを獲得したいというモチベーションが極めて効果的であったものと考えられる．

「実際に展示物をみんぱくで見たいと思いましたか？」という質問では，平均が 3.50 であった．インタビューで同質問を行ったところ，2 名の被検者は，アプリケーション上では展示物の大きさを確認することができないため，実物を見てみたいと回答した．これには，プリセットで「手に乗る大きさ」という項目を用意していたことも一因であったものと考えられる．現地で用いるのではなく，事前学習としての曼荼羅ビンゴの活用も有効的なのではないかという意見が得られた．

また，印象的だった複数の展示物について，ビンゴカード上で対応する軸の情報を記憶している被検者がいた．加えて，詳細画面に記載されている展示物の説明文も印象付けの一因となっていた．そのほか，もともと知識があった展示物が印象に残っているという意見も得られた．

また，ある被検者は「アフリカの展示物でアルコール飲料と広告業のタグがある展示物が記憶に残ってて．それが原因でビンゴが完成しなかった.」と発言している．出現頻度の低い OCM を用いると多くのスコアが獲得できる一方で，該当する OCM を有する限られた展示物でビンゴを完成できなかったことが，かえって印象付けに繋がった．類似した意見として，4 つの自由軸で OCM を用いることで，スコアを伸ばすことを目標とした被検者もいた．アプリケーション上で展示物を検索し，OCM を見る機会が増えた一方で，ビンゴの達成に注力したため，詳細にはあまり触れる余裕がなかったという意見が得られた．このことから，より学習効果を高めるためには，ロック機能と同様に，展示物の詳細情報を読ませる工夫が必要であることが明らかになった．

ユーザビリティに関しては，表 3 のような評価を得た．ここでは，被検者毎の評価を 100 点満点のもと，算出している．SUS については，500 件の研究を調査したところ，平均スコアが 68 とされている [2]．全ての被検者の評価を平均したところ，

2：Measuring Usability with the System Usability Scale (SUS), https://measuringu.com/sus/

スコアは 75 であり，平均スコア 68 を超えていることから，扱いやすいシステムであるということができる．

4.2 鑑賞軸の有効性に関する実験

本節では，鑑賞軸の有効性について検証する．斜め軸を除外し，あらかじめ指定した鑑賞軸に基づいて全てのマスを埋めるタスクを被験者に課し，実験を行った．本ガイドアプリを**軸固定版**とする．

4.2.1 軸固定版の概要

図 7 に曼荼羅ビンゴの実例を挙げる．ガイドアプリの使用開始時は画面左側に 3 × 3 の 9 マスのビンゴカードがあり，横軸，縦軸があらかじめ決まっている．例として，左上のマスにおいては，横軸が「特別な服装」という OCM，縦軸が「中央・北アジア」という展示場所であることが分かる．この時，このマスに入れることができる展示物は,「中央・北アジア」という展示場所にある,「特別な服装」という OCM を持った展示物である．この時，鑑賞者は条件を満たす展示物をみんぱく内から探し，マスを埋める．同様の操作を繰り返し，最終的にすべてのマスを埋め，ビンゴカードを完成させることが，鑑賞者にとっての目標となる．

4.2.2 軸固定版の実験手法

実験は，上述したガイドアプリを用いて行った．場所は大阪の国立民族学博物館である．日程は 2023 年 10 月 14 日，15 日，11 月 19 日に行った．人数は，10 月 14 日が 2 名，10 月 15 日が 5 名，11 月 19 日が 3 名の計 10 名を対象とした．これらの被験者は，兵庫県立大学の学部生である．また，一度もみんぱくを訪れたことがない．被験者には，みんぱくガイドを使用してもらった．実験では，ビンゴの軸の数を変えて比較検証を行う．

本実験を行う前に，予備実験として，研究室の 2 名の学生に 3 つのアプリケーションを用いてもらった．アプリケーションは，横軸，縦軸が存在するビンゴ（2 軸），横軸のみのビンゴ（1 軸（横）），縦軸のみのビンゴ（1 軸（縦））の 3 つである．これらにおいて，9 マス全てを展示物で埋めて，ビンゴを完成させるというタスクを課した．ガイドアプリ上で展示物の詳細を確認した回数は表 4 の通りであった．なお本研究では，画面に掲載している情報量から，3 秒以上画面を開いていた場合に，展示物の詳細を確認したものとして扱っている．表より，1 軸（縦）における詳細画面の確認回数が極めて少なかった．これは，縦軸のみの場合だと，制約がみんぱく内の展示場所のみとなっている．ビンゴをクリアするという目的のもとでは，展示場所が合っていれば，どのような展示物でも用いることができる．そのため，鑑賞軸を意識してもらうという本来の目的を達成できない．実験では，横軸，縦軸が存在するビンゴを 2 軸導出型，横軸のみのビンゴを 1 軸導出型とし，これらを比較する．

1 名あたり，これら 2 つのアプリケーションを用いてもらう．ビンゴでは，軸の OCM 及び展示場所は事前に指定したものを用いる．順序効果を考慮し，当日鑑賞してもらう展示物について「衣」と「住」という 2 つのテーマを設定した．ビンゴの軸は，このテーマに基づいた 2 種類を用意した．用いた鑑賞軸は表 5 の通りである．実験では，9 マス全てに展示物を入れ，ビンゴを完成させるというタスクを課した．タスクに取り掛かる時間は，各アプリケーションともに 60 分程度としており，ビンゴが完成した時点で終了するよう被験者には伝えている．また，表 6 のように，被検者毎にアプリケーションを使用してもらう順番を変えている．実験の前半及び，後半で使用するビンゴゲームのトピックに加え，括弧内は，ビンゴゲームの軸の数を表している．2 が 2 軸導出型，1 が 1 軸導出型を指す．

図 7 （軸固定版）曼荼羅ビンゴの画面

表 4 （軸固定版）予備実験における，展示物の詳細画面の確認回数

	2 軸	1 軸（横）	1 軸（縦）
被検者 A	73	30	10
被検者 B	46	53	7

表 5 （軸固定版）実験で用いた鑑賞軸

鑑賞軸	衣	住
OCM	特別な服装 装飾 通常の毛皮	家庭用品 木工 儀式
展示場所	中央・北アジア 中国 アフリカ	中央・北アジア 朝鮮 ヨーロッパ

実験後には，みんぱくガイドのユーザビリティ及び，鑑賞体験についてアンケートを行った．ユーザビリティの評価には SUS を用いた． **軸自由版**と同様に，各項目をみんぱくガイド用に置き換えている．また，鑑賞体験については，ビンゴの楽しさや難易度，達成感等について，5 段階のリッカート尺度を用いたアンケートを行った．鑑賞体験のアンケートの項目は，表 7 の通りである．また，当日の鑑賞体験を通して，印象に残った展示物について，インタビューを行った．

4.2.3 軸固定版の実験結果

表 8 は，該当のガイドアプリを前半，後半のいずれに使用したかによって，ビンゴの完成までにかかった時間を表している．多くの被験者にとって，2 軸導出型の方がビンゴの達成にかかる時間は長い結果となった．これは，1 軸導出型と比較してビンゴの 1 マス当たりの制約が増えることが原因である．次に，アプリケーションを使用した順番を考慮する．前半に 2 軸

表 6 （軸固定版）被験者毎の使用したアプリの順番

被検者	1	2	3	4	5	6	7	8	9	10
前半	衣 (2)	住 (1)	衣 (2)	住 (1)	住 (2)	衣 (1)	住 (2)	衣 (1)	衣 (2)	住 (1)
後半	住 (1)	衣 (2)	住 (1)	衣 (2)	衣 (1)	住 (2)	衣 (1)	住 (2)	住 (1)	衣 (2)

表 7 （軸固定版）事後アンケートの項目

質問項目
Q1. 1 年にどの程度，博物館，美術館に行きますか？ （自由記述）
Q2. 1 日にどの程度，ゲームを遊びますか？ （自由記述）
Q3. 1 軸と 2 軸で感じた違いはありましたか？ （1: 全くなかった，5: とてもあった）
Q4. ビンゴをすることは，勉強になりましたか？ （1: 全く勉強にならなかった，5: とても勉強になった）
Q5. ビンゴをすることで，新たな発見はありましたか？ （1: 全くなかった，5: とてもあった）
Q6. ビンゴは楽しかったですか？ （1: 全く勉楽しくない，5: とても楽しい）
Q7. ビンゴは難しかったですか？ （1: とても簡単，5: とても難しい）
Q8. ビンゴをすることで，達成感を得られましたか？ （1: 全く得られなかった，5: とても得られた）
Q9. またビンゴをしたいと思いますか？ （1: 全く思わない，5: とても思う）
Q10. 特に良かったところと改善点を教えてください （自由記述）
Q11. 本日の鑑賞体験が，どの程度有意義でしたか？ （1: 全く有意義ではなかった，5: とても有意義だった）
Q12. 鑑賞体験を経て,「またみんぱくに来たい」と思いましたか？ （1: 全く思わない，5: とても思う）

導出型から使用した被検者 1，3，5，7，9 は，2 軸導出型においては平均 54.6 分，1 軸導出型においては平均 32.2 分であった．もう一群の，前半に 1 軸導出型から使用した被検者 2，4，6，8，10 は，2 軸導出型においては平均 51.6 分，1 軸導出型においては平均 45.4 分であった．1 軸導出型において，前半に 1 軸導出型を用いた被検者群は，平均時間がもう一群よりも 10 分以上長かった．これは，前半はアプリケーションへの慣れが必要であったことが原因として考えられる．

また，表 9 は，各軸毎に操作ログを分析した結果である．軸を複雑にしたことによって，多くの被験者において，タスクに取り掛かる時間の増加に伴い，ビンゴ画面の確認と展示物の詳細の確認回数が増加した．

アンケートにおいても，ビンゴの難易度は，5 段階中の平均は 2 軸導出型が 4.00，1 軸導出型が 2.10 という評価が得られた．タスクの達成時間と同様に，2 軸導出型の方が制約が増えることが原因である．一方で，ビンゴをした際の達成感については，5 段階中の平均は 2 軸導出型が 4.30，1 軸導出型が 3.70 という評価が得られた．また,「またビンゴを遊びたいか」という，アプリケーション利用の反復性という点においても，5 段階中の平均は 2 軸導出型が 4.00，1 軸導出型が 3.50 という評価が得られた．このことから，程よい難易度と達成感が，好評

表 8 （軸固定版）被験者毎のビンゴの完成に要した時間（分）

被検者	1	2	3	4	5	6	7	8	9	10
2 軸	57	51	56	50	55	62	45	53	60	42
1 軸	22	39	31	51	42	64	20	50	46	23

表 9 （軸固定版）被験者毎のログの内容

被検者		1	2	3	4	5	6	7	8	9	10
詳細の表示	2 軸	50	71	36	77	31	53	60	82	56	71
	1 軸	46	48	45	53	31	24	50	32	74	48
ビンゴの表示	2 軸	46	20	21	33	25	24	36	41	35	20
	1 軸	21	18	12	20	24	21	26	37	26	18
展示物の検索	2 軸	7	14	10	29	26	9	4	6	28	14
	1 軸	1	18	3	13	12	0	3	1	38	18

であったことがわかった．

インタビューでは，被験者の印象に残った展示物について調査した．被検者 1，3，4，5，7 の 5 名は，既存の知識とは異なっている展示物や，名前は知っていたが見たことのない展示物が印象に残っていた．

2 軸導出型のビンゴにおいては，被検者 2，10 の 2 名が衣服の華やかさや暖かさといった観点から文化の違いを認識していた．被検者 7 も，具体的な展示物の情報までは記憶できていなかったが，2 軸導出型のビンゴによって，地域とのつながりを実感できたという意見を得られた．対照的に，被検者 6 からは，1 軸導出型のビンゴでは地域の制限がないため，様々なエリアを回ることができたが，2 軸導出型はかえって地域の制限によって，探すことに集中してしまったという評価が得られた．これらのことから，鑑賞軸という観点を付与した上での鑑賞は効果的であった．一方で，被検者 6 のように，ビンゴに使用する展示物を探すことが目的となってしまうという問題も挙げられた．ユーザ毎の適切なガイドアプリを検討する上で，多くの被験者を対象とした実験が必要である．

そのほか，被検者 1，5 の 2 名は,「最後の方にやったから.」や「やっぱり，儀式に使うこのヨーロッパの地図はすごい印象に残ってます．一番最後に行った.」と，ビンゴの進行状況に応じて，探した展示物が印象に残っていた．これはビンゴゲームという形式ならではの利点である．

検索機能に関しては，多くの被検者は，最初は自力で展示物を探し，難易度の高いマスに対してや時間が迫ってくると，検索機能を用いていた．事後アンケートの結果と照らし合わせると，ビンゴゲームの活用という点においては，鑑賞や展示物の検索を促す効果があったものと考えられる．

また，ユーザビリティに関しては，表 10 のような評価を得た．平均は 75 である．なお，評価が最も低かった被検者 6 に関しては，ビンゴに使用する展示物を探す際に，複数条件での検索がうまくいかなかったことが理由として挙げられた．

表 10 （軸固定版）被験者毎の SUS の評価

被検者	1	2	3	4	5	6	7	8	9	10
スコア	95	70	72.5	82.5	72.5	55	77.5	60	82.5	87.5

5 ま と め

本研究では，博物館において，鑑賞者が時間的制約等によって展示物を満足に鑑賞できず，展示物の記憶が残り難い問題の解決に取り組んだ．曼荼羅ビンゴを通して，ユーザに独自の鑑賞軸に基づいた鑑賞を促す．鑑賞軸とは，展示物の大きさや形，用途など，鑑賞者が展示物を鑑賞する際に持つ各々の観点のことである．自身の鑑賞軸を認知することで，未知の展示物を目にした際に，独自の指標から分析する能力の習得を支援する．同時に，分析の過程で展示物の印象付けを行う．国立民族学博物館（みんぱく）を対象とし，みんぱく用のガイドアプリ「みんぱくガイド」上で動作する,「曼荼羅ビンゴ」を提案した．曼荼羅ビンゴは，みんぱく内の展示物を用いるビンゴゲームであり，ビンゴの横，斜め，縦の軸に，鑑賞軸を割り当てる．鑑賞者は無数の展示物の中から，ビンゴゲームのマスを埋めることができそうなものを鑑賞するよう働きかける．実験では 4 名を対象とし，ガイドシステムを用いて行った．ガイドアプリを使用することによる学習効果はあったものの，ビンゴゲームの軸に用いない項目に関しては，詳細を調べる機会があまりなかった．また，鑑賞軸の有効性については，あらかじめ鑑賞軸を指定したガイドシステムを用意し，みんぱくで 10 名の被験者を対象として検証した．実験では，アンケート，インタビューとデータログの収集を行った．その結果，展示物間での特徴のつながりを意識できたというフィードバックを得た．一方で，ビンゴに用いる展示物を探すことに集中してしまったという意見も得た．本研究では，鑑賞軸による学習効果に焦点を当てているが，今後の課題として，より長期的な調査を行い，詳細な展示物の情報の記憶への定着を測る必要がある．

謝　　辞

本研究は，JSPS 科研費 JP21H03775，JP21H03774，JP21H03554，JP22H03905 の助成を受けたものです．本研究の実施にあたっては，国立民族学博物館より提供いただいた展示物データベースを利用しました．また，HRAF Association より，OWC，OCM のデータをいただき，独自に翻訳して利用しました．ここに記して謝意を表します．

文　　献

[1] David Anderson, Martin Storksdieck, and M. Spock. Long-term impacts of museum experiences. *In principle, in practice*, pp. 197–215, 2006.

[2] John Brooke. SUS: A quick and dirty usability scale. *Usability Evaluation In Industry*, pp. 189–194, 1996.

[3] Silvia Ceccacci, Andrea Generosi, Alma Leopardi, Maura Mengoni, Mandorli, and Ferruccio. The Role of Haptic Feedback and Gamification in Virtual Museum Systems. *Journal on Computing and Cultural Heritage*, Vol. 14, No. 3, pp. 1–14, 2021.

[4] Dicheva Darina, Dichev Christo, Agre Gennady, and Angelova Galia. Gamification in Education: A Systematic Mapping Study. *Journal of Educational Technology & Society*, Vol. 18, No. 3, pp. 75–88, 2015.

[5] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences*, pp. 3025–3034, 2014.

[6] Ioanna Ioannou and Eleni A. Kyza. The Role of Gamification in Activating Primary School Students' Intrinsic and Extrinsic Motivation at a Museum. In *Proceedings of the 16th World Conference on Mobile and Contextual Learning*, pp. 1–4, 2017.

[7] Zheng Jialiang, 伊藤扶桑, 楠房子, 稲垣成哲. 歴史系博物館における AR を用いた親子で体験できる展示手法の開発とデザイン. 情報処理学会 研究報告, Vol. 2022-GN-115, pp. 1–6, 2022.

[8] Eric Klopfer, Judy Perry, Kurt Squire, Ming-Fong Jan, and Constance Steinkuehler. Mystery at the Museum: A Collaborative Game for Museum Education. In *Proceedings of the Computer Supported Collaborative Learning 2005*, pp. 316–320, 2005.

[9] Fiona Fui-Hoon Nah, Venkata Rajasekhar Telaprolu, Shashank Rallapalli, and Pavani Rallapalli Venkata. Gamification of Education Using Computer Games. In *Proceedings of the Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business*, pp. 99–107, 2013.

[10] Keith V. Nesbitt and Ian Hoskens. Multi-Sensory Game Interface Improves Player Satisfaction but Not Performance. In *Proceedings of the 9th Conference on Australasian User Interface*, pp. 13–18, 2008.

[11] Andrew Stott and Carman Neustaedter. Analysis of Gamification in Education. In *Proceedings of the Technical Report 2013-0422-01, Connections Lab, Simon Fraser University*, 2013.

[12] 安部美希, 羽田久一, 今井正和, 砂原秀樹. 電子博物館展示における展示シナリオ自動作成システムの提案. 映像情報メディア学会技術報告, Vol. 26.24, pp. 13–18, 2002.

[13] 永田奈央美. ゲーミフィケーションを用いた基礎学力向上のための学習システムの開発. 静岡産業大学情報学部研究紀要, Vol. 20, pp. 277–284, 2018.

[14] 奥本素子, 加藤浩. 博物館展示を理解・解釈するために必要な学習支援についての考察. 日本教育工学会論文誌, Vol. 33, No. 4, pp. 423–430, 2010.

[15] 前原千尋, 金大雄, 牛尼剛聡. 博物館における展示品の意味関係に基づいた展示品の推薦システム. 日本データベース学会第 3 回データ工学と情報マネジメントに関するフォーラム, 2011.

[16] 大杉隆文, 仲西渉, 多井中美咲, 井上卓也, 伊藤悠, 岩井瞭太, 香川健太, 松下光範, 堀雅洋, 荻野正樹. 動物園における自発的な観察の促進を目的とした園内回遊行動のデザイン. 情報処理学会論文誌, Vol. 58, No. 11, pp. 1765–1775, 2017.

[17] 福山佑樹, 床鍋佳枝, 森田裕介. ゲーミフィケーションの要素を取り入れた小学校 1 年生向け電子教材の実践と評価. デジタルゲーム学研究, Vol. 9, No. 2, pp. 31–40, 2017.

[18] 平澤泰文, 松川節, 川田隆雄, 小南昌信. iPad 博物館ガイドシステムの構築と評価. 日本教育工学会論文誌, Vol. 36, No. suppl, pp. 89–92, 2012.

[19] 荘司慶行, 相原健郎, 大島裕明, 神門典子, 白石晃一, 中島悠太, 山本岳洋, 山本祐輔. 提示型検索モデルに基づくミュージアム電子ガイドとその利用ログを用いた事前学習・事後学習支援. 情報処理学会論文誌, Vol. 63, No. 2, pp. 364–377, 2022.

[20] 荘司慶行, 大島裕明, 神門典子, 相原健郎, 白石晃一, 瀧平士夫, 中島悠太, 山本岳洋, 山本祐輔, 楊澤華. 提示型検索に基づくミュージアム電子ガイドを中核とした事前・事後学習支援. 情報処理学会 じんもんこん 2020 論文集, pp. 81–88, 2020.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 16:00 ~ 18:10 | T1-A オンライン（Zoom Events)

汎用機械学習技術 (3)

座長:塩川 浩昭(筑波大学)
コメンテータ:北川 博之(筑波大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T1-A-3-01] マルチエージェント強化学習におけるMA-POCAモデルのFFNの改善

*藤江 拓郎[1]、白井 匡人[1] (1. 島根大学)

16:20 ~ 16:45
[T1-A-3-02] マルチエージェントゲームを通じた取引ネットワークにおけるより良いルールの自動生成手法

*沢木 翔太[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学情報学群知識情報・図書館学類、2. 筑波大学図書館情報メディア系)

16:45 ~ 17:10
[T1-A-3-03] GANを用いた異常検知における異常度算出方法についての考察

*森 仁美[1]、中野 美由紀[2]、丸 千尋[3]、小口 正人[1] (1. お茶の水女子大学理学部情報科学科小口研究室、2. 津田塾大学学芸学部情報科学科、3. お茶の水女子大学文理融合AI・データサイエンスセンター)

17:10 ~ 17:35
[T1-A-3-04] Empowering Zero-Shot Extreme Multi-Label Text Classification via Weighted Contrastive Learning and Semantic Label Augmentation

*Zhao Yanan[1]、Iwaihara Mizuho [1] (1. 早稲田大学)

マルチエージェント強化学習における MA-POCA モデルの FFN の改善

藤江　拓郎†　　白井　匡人††

† 島根大学 自然科学研究科　〒 690–8504 島根県松江市西川津町 1060
†† 島根大学 学術研究院理工学系　〒 690–8504 島根県松江市西川津町 1060
E-mail: †n22m110@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp

あらまし　マルチエージェント強化学習は，ゲーム，自律走行，ネットワークパケットルーティングなどの多様な分野での応用が期待されている．近年の先行研究における学習モデルは，セルフアテンションやトランスフォーマーを用いたものが注目されており，それらを用いた学習モデルはマルチエージェント強化学習分野においても高い性能を発揮している．しかしマルチエージェント分野における最適なモデル構造の模索は研究途上である．本研究では，Self-Attention を使用した MA-POCA モデルをベースモデルとし，協調タスクにおいてより高い性能を発揮する FFN のモデル構造を模索する．実験より有効なモデル構造を検証し，今後の課題について検討する．

キーワード　マルチエージェント強化学習，強化学習

1 前　書　き

マルチエージェント強化学習は，共通の環境下で複数のエージェントが相互に影響を与えながら，報酬を最大化することを目的とした学習分野である．この分野は，ゲーム [1] [2] だけでなく，実世界での応用として，自動運転 [4]，ネットワークパケットルーティング [3] などの多様な分野で期待され，近年盛んに研究が行われている．近年のマルチエージェント強化学習は，アテンション機構 [6] [8] [9] やトランスフォーマー [7] を導入したモデルが注目され，シングルエージェント [10]，マルチエージェント [11] [12] の双方において高い性能を発揮している．しかし先行研究におけるそれらのモデル構造は，自然言語処理や画像処理分野におけるものをベースとしていることが多く，マルチエージェント強化学習分野にモデル構造の模索は研究途上である．

そこで本研究では，マルチエージェント強化学習における複数のエージェントが協力してタスクを達成する協調タスクを想定し，最適なモデル構造を模索する．具体的には，Cohen らによって提案されたアテンション機構が導入された MA-POCA モデル [9] をベースモデルとして使用し，協調タスクにおいてより性能が高い FFN のモデル構造を模索し，実験を通してモデルの有効性を検証する．

第 2 章では，先行研究について述べる. 第 3 章では，提案手法について述べる．第 4 章では，本研究での検証に使用した環境，実験方法，実験結果について説明する．第 5 章では，実験結果から考察し，結論をまとめる．

2 関 連 研 究

2.1 マルチエージェント強化学習

マルチエージェント強化学習は，同一環境で複数エージェントが学習を行う．環境によって，複数エージェントが敵対または協力関係になる．本研究では，協力関係にあるエージェント群を扱う．マルチエージェント強化学習の目的は，エージェントが報酬を最大化する行動を学習することである．

2.2 部分観測マルコフ決定過程

部分観測マルコフ決定過程 (Partially Observable Markov Decision Process: POMDP) は，環境の状態が完全に観測できない状況を扱うための枠組みである．本研究は，Cohen らが提案した MA-POCA モデルをベースにしている．そのため POMDP の設定も同様に扱う．POMDP は $(N, S, O, U, P, r, \gamma)$ の 7 つの要素から構成される．$N \geqq 1$ は，エージェントの数である．S は，環境の状態空間である．時間ステップ t の時の状態は，$s_t \in S$, エージェント i の観測は $o_t^i \in O$ である．$O \coloneqq O^1 \times ... \times O^N$ は，全エージェントの共同観測空間である．共同観測 $\boldsymbol{o} = o^1, ..., o^N$ は，$\boldsymbol{o} \in O$ である．エージェントは，環境の全状態 s を直接観測することができないため，自身で環境の一部分を観測し，それをもとに行動を決定する．$U \coloneqq U^1 \times ... \times U^N$ は，共同行動空間である．共同行動 $\boldsymbol{u} = u^1, ..., u^N$ は，$\boldsymbol{u} \in U$ である．P は，状態遷移確率である．状態 s の時，共同行動 $\boldsymbol{u}$ によって s' に遷移する確率を $P(s'|s, \boldsymbol{u})$ と表記する．r は報酬関数で，$r(s, \boldsymbol{u})$ と表記する．これは状態 s で，共同行動 $\boldsymbol{u}$ を取った際，全エージェントが受け取る報酬である．γ は割引率である．

2.3 Counterfactual Multi-Agent Policy Gradients

協調的なタスクにおいてグローバルな報酬から各エージェントがチームの成功から自身の貢献を推測することは困難である．このマルチエージェント信用割り当てと呼ばれる問題を解決するために，Foerster らは Counterfactual Multi-Agent Policy Gradients(COMA) [5] を提案している，COMA では，反実仮想ベースラインと呼ばれる手法を用いて，各エージェントの貢献を推測する．

$$g = \mathbb{E}_\pi[\sum_a \nabla_\theta \log \pi^a(u^a|\tau^a)A^a(s,\mathbf{u})]$$

ここで g は方策勾配，θ はアクターのパラメータ，π^a はエージェント a の方策，τ^a はエージェント a の観測と行動の履歴である．A^a はアドバンテージ関数である．アドバンテージ関数は，ある状態で特定の行動を選ぶことの相対的な価値を評価する．COMA では，アドバンテージ関数を式 1 で表記する．

$$A^a(s,\mathbf{u}) = Q_\phi(s,\mathbf{u}) - b(s,\mathbf{u}^{-a}) \quad (1)$$

$$b(s,\mathbf{u}^{-a}) = \sum_{u'^a} \pi(u'^a|\tau^a)Q_\phi(s,\mathbf{u}^{-a},u'^a) \quad (2)$$

$\mathbf{u}$ は，環境内の全エージェントの行動集合，u'^a はエージェント a が取ることが可能な行動集合，$\mathbf{u^{-a}}$ はエージェント a 以外の行動集合，ϕ は評価者のパラメータ，$b(s,\mathbf{u}^{-a})$ は反実仮想ベースラインである．これは，エージェント a の行動 u^a について，他のエージェントの行動 u^{-a} を固定し，周辺化する，そして各エージェントの行動がチームに対してどの程度貢献しているかを評価する．しかし，この手法ではエージェント数に応じて計算コストが二次的に増加する問題がある．また固定数のエージェントを想定しているため，エージェント数が変動する環境においては学習が難しい．

2.4 Multi-Actor-Attention-Critic

Iqbal らが提案した Multi-Actor-Attention-Critic(MAAC) [6] は， 中央集権的な評価者に対し，アテンション機構を採用した．
アテンションモデルを使用することで，エージェント間で強い関連性を持つ情報を重視できる．さらに COMA の問題であった同じ行動空間を仮定する必要がなくなり，グローバルな報酬が不必要となった．またエージェントは他のエージェントの状態やアクションに基づいて，自身の行動価値を推定する際に，どの情報を重視すべきかを動的に決定できる．エージェント $i \in \{1,..,N\}$ の行動価値 Q_i^ϕ を推定するために，アテンションモデルを使用する．

$$Q_\psi^i(\mathbf{o},\mathbf{u}) = f_i(g_i(o^i,u^i),x_i) \quad (3)$$

ここで $j \in \{1,..,N\}$， $j \neq i$ はエージェント, ϕ は評価者のパラメータ，$\mathbf{o}$ は全エージェントの観測，$\mathbf{u}$ は全エージェントの行動，f_i は 2 層の多層パーセプトロン (Multilayer perceptron) による埋め込み関数，g_i は 1 層の MLP による埋め込み関数である．活性化関数は，Leaky ReLU を使用している．x_i は，アテンションの出力である，式 4 で表記される．

$$x_i = \sum_{j \neq i} \alpha_j v_j \quad (4)$$

$$\alpha_j = \frac{softmax(e_j^T W_k^T W_q e_i)}{\sqrt{d}} \quad (5)$$

$$e_j^T = g_j(o^j,a^j), e_i^T = g_i(o^i,u^i), v_j = V g_j(o^j,u^j)$$

ここで，g_j は 1 層の MLP による埋め込み関数，V は重み，W_q, W_k は重み行列，d は次元数である．アテンションモデルを使用することで，COMA より柔軟に反実仮想ベースラインを実装できる．

2.5 Multi-Agent POsthumous Credit Assignment

Cohen らが提案した Multi-Agent POsthumous Credit Assignment(MA-POCA) [9] は，評価者モデルに 2 つのアテンションモデルを導入している．アテンションモデルは，Transformer [7] のマルチヘッドセルフアテンションをベースとして使用している．1 つ目のアテンションモデルは，エージェント i の観測 o_t^i から状態価値関数 $V(o_t^i)$ を推定する．式 6 で表記される．

$$J(\phi) = (V_\phi(Attention(g_i(o_t^i)_{1\leqq i\leqq N})) - y^{(\lambda)})^2 \quad (6)$$

$$y^{(\lambda)} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^n \quad (7)$$

$$G_t^n = \sum_{l=1}^{n} \gamma^{l-1} r_{t+l} + \gamma^n V_\phi(Attention(g_i(o_{t+n}^i)_{1\leqq i\leqq k_{t+n}}))$$

ここで，ϕ は価値関数モデルのパラメータ，V_ϕ は状態価値関数，g_i は観測エンコーダ，$Attention$ は Transformer のマルチヘッドセルフアテンション，k_{t+n} は時間ステップ $t+n$ でアクティブなエージェントの数，γ は割引率である．観測エンコーダでは，画像データであれば畳み込み層で処理を行う．数値データであれば，全結合層で処理を行う．λ は，将来的な値をどれほど重要視するかを調整するパラメータである．式 7 は，TD(λ) を表記している．これは，TD 誤差を n ステップ先まで考慮し，平均を取ったものである．具体的には，各収益 G_t に $(1-\lambda)\lambda^{n-1}$ をかけ，正規化する．λ の値が小さいほど，現在の状態における状態価値の影響が大きくなる．
2 つ目のアテンションモデルは，反実仮想ベースライン Q_ψ を推定する．式 2.5 で表記される．

$$Q_\phi(Attention(g_j(o_t^j), f_i(o_t^i,a_t^i)_{1\leqq i\leqq k_t, i\neq j}))$$

ϕ は，反実仮想ベースラインモデルのパラメータ，g_j は観測エンコーダ，f_i は観測-行動のエンコーダ，$Attention$ は Transformer のマルチヘッドセルフアテンション，k_t はアクティブなエージェント数である．

3 提案手法

本研究では，協力的な行動が必要なタスクを想定する．提案手法は，Cohen らが提案した MA-POCA モデルをベースとしている．アテンションモデルの構造は，Transformer [7] のエン

コーダの構造をベースとしている．先行研究の MAAC [6] や MA-POCA [9] は，アテンションモデルを使用することで高い性能を発揮した．しかし，手法ごとにアテンションの処理後に使用する順伝搬型ネットワーク (FFN) は異なる．そのため，既存の枠組みに捉われず，高い性能を発揮する FFN を模索する．本研究では，層数，ノード数，活性化関数をそれぞれ変更し，網羅的に学習を行う．層数はそれぞれ 2 層，3 層の 2 種類で学習を行う．ノード数は 512,1024,2048, 4096, 8192 の 5 種類で学習を行う．活性化関数は，GeLU で学習する．MA-POCA と同様に，150 万ステップでシミュレーションを行い，バッチサイズが 1024, 3epoch の学習を行う．1 ステップは，エージェントが環境を観測し，行動を決定するまでの処理である．本研究では，複数エージェントそれぞれの行動回数をまとめて計測する．

4 実　　験

4.1 実験環境

4.1.1 Unity

Unity は，レンダリングと物理エンジン，Unity Editor と呼ばれる GUI からなるリアルタイムの 3D 開発プラットフォームである．ゲーム制作だけでなく，自動車，映像業界で広く使用されている．

4.1.2 Unity ML-Agent Toolkit

本研究では，エージェントの学習と評価のための環境として，Unity ML-Agents Toolkit [1] を使用する．これは，Unity エンジンを基盤として，Python API を通して Unity のシミュレーション環境を制御することが可能である．また，3D および 2D の両方の環境をサポートしており，リアルタイムの物理演算や複雑な環境設定，エージェントの多様なカスタマイズを可能にしている．

4.1.3 Cooperative Push Block

本研究では，Unity ML-Agent Toolkit で用意されている Cooperative Push Block 環境を学習環境として使用する．図 1 に示す．この環境は，3 体のエージェントが，1 つのチームとして協力し，ブロックを緑色のエリアに移動させるタスクである．チームの目標は，全てのブロックを緑色のエリアに移動させることである．ブロックは，1,2,3 の数字が書かれた 3 種類ある．ブロックを運ぶためには，ブロックに書かれている数字のエージェント数が必要になる．そのため，エージェントはブロックを運ぶために協力する必要がある．報酬は，1 フレーム毎にグループに対して，-0.001 が与えれる．ブロックをゴールに移動させることで，ブロックの数字に応じてそれぞれ，+1,+2,+3 の報酬が与えられる．

4.1.4 エージェントの観測

エージェントの観測は，Grid Sensor と Rigidbody Sensor の 2 種類を使用する．GridSensor は，エージェントの周囲の環境をグリッド状に観測するセンサである．学習に使用した環境を図 2 に示す．壁，エージェント，ゴール，1 と書かれたブロック，2 と書かれたブロック，3 と書かれたブロックの 6 種類のラベルがある．セルの位置にラベルの物体が存在すると，ラベルに対応した色がセルに表示される．エージェント前方の 20×20, 6 チャネルを畳み込み層で処理し，観測として使用する．Rigidbody Sensor は，エージェントの周囲の環境を観測するセンサである．エージェントの向き，位置，速度を観測として使用する．

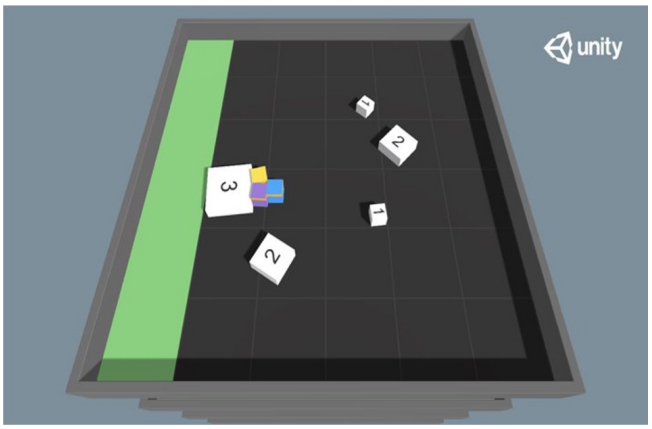


図 1: Cooperative Push Block

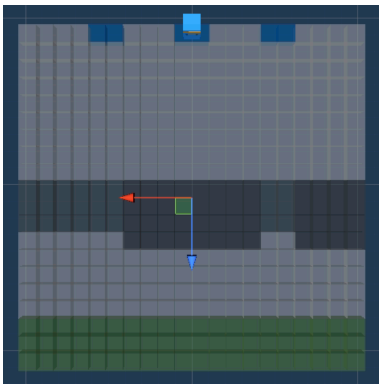


図 2: Grid Sensor

4.2 実験内容

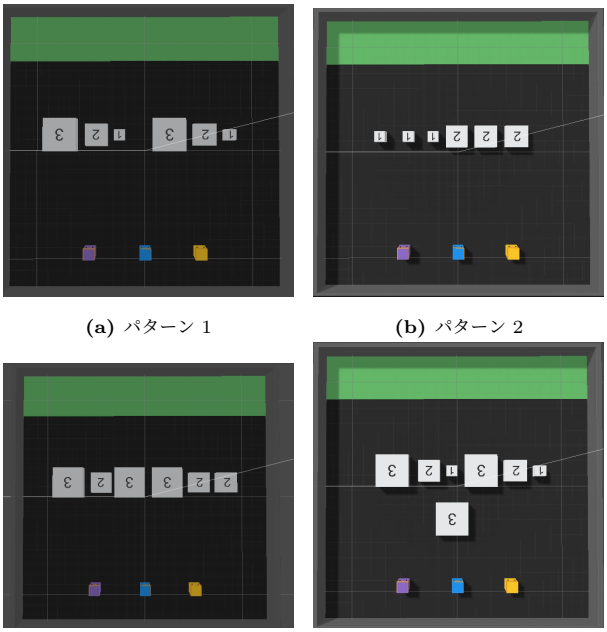
実験では，Cooperative Push Block 環境をベースとした図 3 の 4 種類の環境で実験を行う．タスクを完了するまでの時間を計測し，比較する．600 秒を制限時間とする．制限時間を超えたものは 1000 秒として扱う．それぞれの実験環境で 20 回タスクを行い，平均時間を計測する．

4.3 実験結果

2 層と 3 層の全結合層でノード数を変更した 4 つの手法を比較する．実験結果をそれぞれ表 1, 表 2, 表 3，表 4 に示す．

4.4 考　　察

全てのパターンにおいて，2 層のノード数 2048，3 層のノード数 512 が MA-POCA よりも早くタスクを完了する．よって，2 層のノード数 4096，3 層のノード数 512 が本研究の協調タスクにおいて MA-POCA より優れていると言える．



(a) パターン 1　(b) パターン 2

(c) パターン 3　(d) パターン 4

図 3: 実験で使用した環境

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	39.72	13.64	69.72	21.8
(2)512	130.27	249.94	1000.0	**14.8**
(3)1024	133.38	296.3	1000.0	27.16
(4)2048	**29.14**	**12.85**	**67.12**	17.12
(5)4096	81.34	219.31	1000.0	15.88
(6)8192	107.04	40.76	1000.0	18.56

(a) 2 層での実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	39.72	13.64	62.92	21.8
(2)256	39.24	43.6	207.38	16.24
(3)512	**26.81**	8.49	49.94	17.28
(4)1024	77.28	217.56	1000	**15.1**
(5)2048	28.45	**7.77**	**45.66**	18.76

(b) 3 層での実験結果

表 1: パターン 1 環境の実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	25.55	17.23	91.22	14.96
(2)512	86.19	223.32	1000.0	12.34
(3)1024	71.29	218.67	1000.0	11.66
(4)2048	**20.92**	**10.21**	**56.88**	**9.54**
(5)4096	70.53	218.86	1000.0	14.24
(6)8192	120.08	301.07	1000.0	12.24

(a) 2 層での実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	25.55	**17.23**	**91.22**	14.96
(2)256	75.52	220.52	1000	**12.4**
(3)512	**22.28**	22.58	117.2	13.2
(4)1024	69.96	219.04	1000	15.1
(5)2048	214.66	402.89	1000	12.5

(b) 3 層での実験結果

表 2: パターン 2 環境の実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	56.46	22.27	113.38	27.68
(2)512	**42.16**	**14.91**	**66.94**	**23.54**
(3)1024	53.05	41.85	178.64	25.44
(4)2048	47.49	25.92	121.26	25.64
(5)4096	47.07	19.25	109.22	27.42
(6)8192	61.74	27.35	134.98	26.72

(a) 2 層での実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	56.46	22.27	113.38	27.68
(2)256	48.51	17.05	106.32	30.78
(3)512	49.74	18.2	89.68	27.32
(4)1024	42.67	**11.47**	**69.68**	28.02
(5)2048	**40.59**	13.31	77.96	**25.44**

(b) 3 層での実験結果

表 3: パターン 3 環境の実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	49.44	**17.23**	**86.84**	26.54
(2)512	165.57	309.91	1000	**22.98**
(3)1024	133.88	296.3	1000.0	27.16
(4)2048	48.21	28.74	150.22	23.98
(5)4096	**46.26**	17.76	101.04	24.18
(6)8192	89.98	215.56	1000.0	23.78

(a) 2 層での実験結果

手法	平均 [s]	分散 [s]	最長時間 [s]	最短時間 [s]
(1)MA-POCA	49.44	17.23	86.84	26.54
(2)256	94.8	214.71	1000	20.3
(3)512	38.97	17.93	98.46	21.3
(4)1024	**34.53**	**11.27**	**58.58**	**17.54**
(5)2048	118.12	221.36	1000	27.84

(b) 3 層での実験結果

表 4: パターン 4 環境の実験結果

しかし，パターン 2 のタスクにおいては，2 層のノード数 2048，3 層のノード数 512 以外は，性能が落ちている．パターン 2 は，パターン 1 と比較して，3 の箱消え，1 と 2 の箱が増加している．これは，2 体のエージェントですべて運べることを意味する．そのため，協調行動より分担することが必要な環境である．加えてより協調行動が必要なパターン 3 が 2 層のノード数 8192 以外全ての手法で，MA-POCA よりも早くタスクを完了する．これらから，ノード数や層数を増やすとより協調的な行動が増えると推測できる．しかし，分担が必要なタスクにおいては，ノード数や層数を増やすと性能が落ちることがある．よって，ノード数や層数を増やすと，エージェントの行動が協調的になると同時に，行動の多様性が減少するためであると推測する．

5 結　　論

本研究では，Cooperative Push Block のような協調タスクにおいて，層を深くし，ノード数を増やすことで，既存手法と比較して早くタスクを完了することができることを示した．先行研究では明確ではなかったアテンションと FFN 構造の変化によって，協調タスクの性能を改善できることを示した．しかし，starcraft [2] のようなエージェント数が多い協調タスクでは，性能が確認できていない．また，Transformer モデルを使用した場合との比較も行っていない．今後の課題としては，MA-POCA だけでなく，アテンションを用いている他手法であっても有効であるかを確認することが必要である．加えて，本研究で使用した協調タスク以外の環境での性能を確認することも必要である．

文　　献

[1] Juliani, Arthur and Berges, Vincent-Pierre and Teng, Ervin and Cohen, Andrew and Harper, Jonathan and Elion, Chris and Goy, Chris and Gao, Yuan and Henry, Hunter and Mattar, Marwan and others, “Unity: A general platform for intelligent agents.” arXiv preprint arXiv:1809.02627 (2018).

[2] Samvelyan, Mikayel and Rashid, Tabish and De Witt, Christian Schroeder and Farquhar, Gregory and Nardelli, Nantas and Rudner, Tim GJ and Hung, Chia-Man and Torr, Philip HS and Foerster, Jakob and Whiteson, Shimon, “The starcraft multi-agent challenge.” arXiv preprint arXiv:1902.04043 (2019).

[3] Ye, Dayon, Minji Zhang, and Yu Yang, “A multi-agent framework for packet routing in wireless sensor networks.” sensors 15.5 (2015): 10026-10047.

[4] Shalev-Shwartz, Shai, Shaked Shammah, and Amnon Shashua, “Safe, multi-agent, reinforcement learning for autonomous driving” arXiv preprint arXiv:1610.03295 (2016).

[5] Foerster, Jakob and Farquhar, Gregory and Afouras, Triantafyllos and Nardelli, Nantas and Whiteson, Shimon, “Counterfactual multi-agent policy gradients.” Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[6] Iqbal, Shariq, and Fei Sha, “Actor-attention-critic for multi-agent reinforcement learning.” International conference on machine learning. PMLR, 2019.

[7] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia “Attention is all you

need." Advances in neural information processing systems 30 (2017).

[8] Yang, Yaodong and Hao, Jianye and Liao, Ben and Shao, Kun and Chen, Guangyong and Liu, Wulong and Tang, Hongyao, "Qatten: A general framework for cooperative multiagent reinforcement learning." arXiv preprint arXiv:2002.03939 (2020).

[9] Cohen, Andrew and Teng, Ervin and Berges, Vincent-Pierre and Dong, Ruo-Ping and Henry, Hunter and Mattar, Marwan and Zook, Alexander and Ganguly, Sujoy, "On the use and misuse of absorbing states in multi-agent reinforcement learning." arXiv preprint arXiv:2111.05992 (2021).

[10] Chen, Lili and Lu, Kevin and Rajeswaran, Aravind and Lee, Kimin and Grover, Aditya and Laskin, Misha and Abbeel, Pieter and Srinivas, Aravind and Mordatch, Igor, "Decision transformer: Reinforcement learning via sequence modeling." Advances in neural information processing systems 34 (2021): 15084-15097.

[11] Wen, Muning and Kuba, Jakub and Lin, Runji and Zhang, Weinan and Wen, Ying and Wang, Jun and Yang, Yaodong, "Multi-agent reinforcement learning is a sequence modeling problem." Advances in Neural Information Processing Systems 35 (2022): 16509-16521.

[12] Hu, Siyi and Zhu, Fengda and Chang, Xiaojun and Liang, Xiaodan, "Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers." arXiv preprint arXiv:2101.08001 (2021).

[13] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," Communications of the ACM (CACM), Vol. 13, No. 6, pp. 377–387, 1970.

DEIM Forum 2024 T1-A-3-02

マルチエージェントゲームを通じた取引ネットワークにおけるより良いルールの探索基盤の構築

沢木　翔太†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †shota.sawaki.2023b@gmail.com, ††ito@slis.tsukuba.ac.jp, †††morishima-office@ml.cc.tsukuba.ac.jp

あらまし　現代社会はテクノロジーの進歩をはじめとして，急速に変化しており，既存の法律やルールが追い付いていないという問題がある．このような変化に対処するため，迅速かつ適切な法律やルールの制定は重要な課題である．本研究では，スケールフリーネットワーク上でエージェントが繰り返し囚人のジレンマゲームをプレイする環境を構築し，その環境における最適なルールの探索方法を提案する．ここで，各エージェントは Q 学習を通じて自己の利益を最大化する戦略を学習する．本研究では，エージェント間の繋がりや関係の打ち切り，囚人のジレンマの報酬設計，管理者という概念の導入などの要因を調整することで，個々の相互協調の促進と社会全体の利益を最大化するような最適なルールの制定を目的とする．システムには管理者が存在し，エージェントの報酬の一部を受け取る仕組みが導入されている．これは現実社会における税制や社会福祉システムへの応用を示唆するものである．実験では，税率と罰則確率をルールとし，エージェントの協調が促進され，全体の利益も増大するようなルールを探索することができた．計算論的アプローチによる法律やルールの自動発見は，法律やルールの制定において新たなアプローチとして機能し，社会システムの理解と改善に貢献する可能性がある．

キーワード　ゲーム理論，取引ネットワーク，マルチエージェントシステム，強化学習

1　はじめに

現代社会では技術革新や人口問題などをはじめとして，様々な分野において大きな変化が起こっており，法律やルールの制定が追いついていないという現状がある．例えば，テクノロジーの進歩スピードに対して，既存の法律やルールが対応できなかったり，新たな法律やルールの制定が追いついていないという現状がある．また，日本における人口問題においては，少子高齢化によって現在の社会保障制度を変更せずに今後も運用することは難しいといった現状がある．最近では，企業や NPO・NGO などの民間団体が政府や世論に対して行う，社会の機運醸成やルール形成のための働きかけ活動であるパブリックフェアーズも活発化しており，社会変化に対する法律やルールの対応は注目のある問題であることがわかる．しかし，法律やルールの制定には多くの人的，時間的リソースが必要となる．

近年の経済学分野では「マーケットデザイン」や「メカニズムデザイン」に関する研究や社会実装が注目されている．これらは，ゲーム理論を用いて，人々にとって望ましい結果を得るには，どのような制度，つまり，ゲームのルールをデザインすればよいかという問題を取り扱っており，理論による制度設計を現実問題に実装することを目指す学問である．特に制度への参加者がおのずと社会全体にとって望ましい行動をとってくれるようなインセンティブ設計の問題を対象としている．マーケットデザインやメカニズムデザインでは，オークション理論 [1] [2] [3] やマッチング理論 [4] の考え方を実際に社会実装しており，各個人とシステム全体の双方の利益の向上を可能としている．

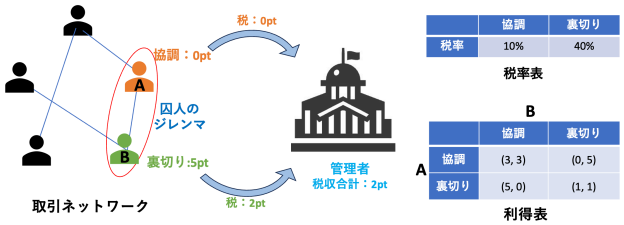


図 1: シミュレーション環境の概略

本研究では計算論的アプローチにより自動的により良いルールを発見する方法を提案する．自動的なルール発見可が能となれば，人的，時間的リソースの削減などをはじめとして，大きなインパクトがあると考えられる．本研究では，現実社会における個人間の契約関係や企業間の取引関係，国家間の貿易関係などをモデル化した取引ネットワークを対象として，以下の 2 つの問題に取り組む．

（1）　**取引ネットワークにおける様々なルールを検証するためのシミュレーション基盤の構築**

（2）　**1 で構築したシミュレーション基盤の有用性の検証**

本研究では，この目的を実現するための手段として，スケールフリーネットワーク上においてノードをエージェント，エッジを取引関係とみなしたエージェント間の取引のシミュレーショ

ン環境を構築する．本研究では，エージェントは現実の取引関係において生じるジレンマをモデル化した繰り返し囚人のジレンマをプレイすると仮定する．また，各エージェントは Q 学習により自身の利得を最大化するような行動方策を学習する．これにより，各エージェントは環境において合理的な行動を学習でき，社会における個人や企業，国家などの意思決定の主体の振る舞いがシミュレートできると期待できる．このシミュレーション環境では，エージェントの行動履歴に基づいて，エージェント同士が新たに取引関係を構築したり打ち切ることを可能とすることで，取引ネットワークを動的なものにした．これにより，ルールの違いによる取引の活発化や衰退が観察可能となる．

また，このシミュレーション環境には管理者による税収が存在すると仮定し，税率と裏切りに対して罰則を与える確率の 2 つをルールとして設定する．今回はエージェントが囚人のジレンマによって得る利得の一部を税金として収める設定にした．支払う税金は，基本税か罰則税のいずれかであり，協調の際には基本税率，裏切りの際は罰則確率に基づき基本税率または罰則税率のいずれかが決定され，この税率に基づいて税金が算出される．税収と罰則確率を導入することで，エージェントの戦略の変化と税収としての取引ネットワーク全体の利益を観察することが可能となり，よりよいルールの発見を可能とする．

実験では以下について検証を行った．

(RQ1) **提案するシミュレーション環境で現実の取引ネットワークで現れる変化が観測できるか．**

(RQ2) **税率と罰則確率を変化させた際にシミュレーション結果にどのような変化が生じるか．**

(RQ3) **環境からの税収が最大となる税率と罰則確率が探索可能であるか．**

実験の結果以下が示された．(RQ1) については，協調と裏切りの報酬の割合を変更した際に，エージェントは自身の利得を最大とする方策を学習し，環境において協調が優位となる利得行列と裏切りが優位となる利得行列が存在することが示された．(RQ2) については，3 つの税率，5 つの罰則確率パターンでシミュレーションを行い，エージェントの戦略，税収，取引数の変化を観察した．罰則確率が 100%の設定における 3 つの税率パターンの結果から，基本税率に対する罰則税率が大きいほど協調が促進されることがわかった．また，同一税率における異なる罰則確率の結果の比較から，罰則確率が大きいほど協調が促進され，取引数が増え，税収も大きくなることがわかった．特に，罰則確率が大きいほど早い段階で協調が促進されることがわかった．この結果は，現実社会における罰則の大きさと確実性が犯罪率に与える影響について調査した Becker の研究 [5] に矛盾しない結果であり，有効な手法であるといえる．(RQ3) については，Optuna [6] を用いて税収が最大となる税率と罰則確率を探索した．探索の結果，罰則確率がより高く，税率は協調の利得が裏切りの利得より高くなる条件のもとでより高い税率の場合に税収が大きくなることがわかった．

2 関 連 研 究

囚人のジレンマを対象に協調と裏切りのダイナミクスに関する研究は実実験，シミュレーションともにさまざまな分野で数多く研究されてきた [7] [8]．多くのジレンマにおいて，意思決定は記憶や評価，報酬，罰則など複数の要因によって決定されており，それぞれの変数において戦略との関係についての研究がなされている．その中でも，本研究と関連が強い研究トピックとして，囚人のジレンマにおける報酬設計，強化学習，ネットワーク構造，またはこれらの複合的な研究が該当する．

2.1 囚人のジレンマと報酬設計

囚人のジレンマにおいて利得表で表現される報酬の構造はエージェントの戦略に影響を与える要因の一つである．Vassilis ら [9] は，マルチエージェント強化学習での囚人のジレンマを対象に，迅速に協調に収束するような報酬の構造を探索した．その結果，報酬の値は正と負の値をもち，負の値はなるべく小さく，正の値は 0 に近い場合が最も迅速に協調に収束するということを主張した．本研究でも同様に囚人のジレンマの報酬に焦点を当てている．しかし，本研究では囚人のジレンマの報酬自体は固定であり，税収という仕組みの導入によってエージェントが Q 学習の際に用いる報酬の値を決定する．税収を用いることで，協調の割合のみではなく，全体の税収という指標でシミュレーション全体の利益を観察することが可能となる．つまり，各個人が自身の利得を追求した結果における，全体の協調の割合と税収という評価指標を対象としてより良いルールを探索することを可能にする．

2.2 囚人のジレンマとゲーム理論

囚人のジレンマをはじめとするゲーム理論の問題を対象に，マルチエージェント強化学習を用いたアプローチによる研究は数多く存在する [10] [11] [12]．Joel ら [13] は fruit Gathering game と Wolfpack hunting game という 2 つのジレンマゲームを対象に資源の豊富さなどの環境要因の変化がエージェントの学習にどのように影響するかを調べ，共有資源をめぐる問題においてどのように協調や対立が生じるかを明らかにした．Zhao ら [14] は繰り返し囚人のジレンマにおいて，エージェントが過去の経験から学習し、他のエージェントとの相互作用の頻度や影響力を時間とともに調整することで，協力者間の結びつきが時間とともに強化され、非協力者との結びつきが断たれるという相互作用パターンを明らかにした．そして，協力を促進する最適な社会的相互作用パターンの出現には強化学習が有効であることを示した．

このように強化学習におけるゲームの環境的な要因，もしくはエージェントの学習方法などによって，様々なメカニズムを明らかにすることが可能である．本研究では，エージェントが行動方策を学習することを目的として強化学習を用いており，Q 値の更新において囚人のジレンマにおける利得から税金を引いた値を用いることで，税収がエージェントの学習にどのよう

な影響を与え，全体の結果にどのように影響を及ぼすのかを検証する．

2.3 囚人のジレンマとネットワーク

囚人のジレンマを対象に，ネットワーク構造とエージェントの戦略の関連性についての研究は数多く存在する [15]．Santos ら [16] は，スケールフリーネットワークにおける囚人のジレンマでは、協力的なハブが存在することで、全体的な協力行動が促進されることを発見した．一方，Xing ら [17] は，従来の考え方とは対照的に、スケールフリーネットワーク上での囚人のジレンマにおいて、ノード同士のつながりのランダム性やネットワークの大きさによって，協力の進化は抑制されることを主張している．このように，ネットワークの違いや問題設定の違いによって，協調の出現率に違いがあることがわかる．本研究ではスケールフリーネットワークを対象に，エージェントの行動履歴からネットワーク構造を変更する動的なネットワークを取り入れる．これにより，税率の違いとエージェントの戦略が取引ネットワークにどのような影響を与えるかを観察することを可能とする．

3 囚人のジレンマ

囚人のジレンマとは，ゲーム理論のゲームの１種であり，各個人が自身の利益を最大化するための合理的な選択（最適反応）を取り合った結果であるナッシュ均衡が，全体にとっての最適な結果であるパレート最適にはならないことを示唆するものである．この，ナッシュ均衡が全体にとって不合理な結果になるという点でジレンマとなる．囚人のジレンマは，意思決定，社会行動，経済学をはじめとしてさまざまな問題を考察する上で重要な概念とされている．囚人のジレンマでは，２人のプレイヤーがとった行動の組み合わせによって，それぞれのプレイヤーが受け取る利得が決定する（表 1）．

表 1: 囚人のジレンマにおける利得表

	協調 (C)	裏切り (D)
協調 (C)	(R,R)	(S,T)
裏切り (D)	(T,S)	(P,P)

以下，囚人のジレンマゲームにおける記号の意味を説明する．

3.1 行動の選択肢

C(Cooperation): 協調

D(Defection): 裏切り

3.2 利得行列における利得

R(Reward): 相互協調の際に得る利得

P(Punishment): 相互裏切りの際に得る利得

T(Temptation): 協調を選択した相手に対して，自分が裏切ることに成功した際に自分が得る利得

S(Sucker): 自分は協調を選択したが，相手が裏切りを選択した際に自分が得る利得

3.3 囚人のジレンマの利得の条件

利得行列における利得が以下の関係をみたすとき囚人のジレンマとなる．
$$T > R > P > S$$
$$2R > T + S$$

3.4 最適反応

相手の戦略に基づき、自分の利得を最大化するための戦略を『最適反応』と呼ぶ．これは、相手の動きに応じて、自分にとって最も有利な選択をすることを意味する．囚人のジレンマにおいては，プレイヤーは裏切りを選択することが最適反応となる．

3.5 ナッシュ均衡

各プレイヤーの最適反応が組み合わさったもののこと．囚人のジレンマにおいては相互裏切りがナッシュ均衡となる．

4 提案手法

4.1 シミュレーション基盤の概要

本研究では，計算論的アプローチを用いることで効率的により良いルールを探索できるようなシミュレーション基盤の構築に取り組む．本研究では，スケールフリーネットワーク上において，ノードをエージェント，エッジを取引関係とみなすことで取引ネットワークを表現する．取引関係にあるエージェントは, 囚人のジレンマゲームを対戦し，各エージェントは自分と取引相手が選択した行動の組み合わせに基づいて利得を得る．また，本研究では税収の概念を導入した．これにより，エージェントは囚人のジレンマによって受け取る利得の一部を税金として収めることになる．利得のうち，どの程度を税金として収めるかは，エージェントが囚人のジレンマで選択する行動によって確率的に決定する．さらに，本研究では，ネットワークが動的に変化する仕組みを取り入れる．エージェントの行動履歴の協調の割合をもとに，その値が基準値を満たした場合，確率的にエージェントは新たに他のエージェントと取引関係を構築，または，断ち切ることができる．この，税収の概念と動的なネットワークを取り入れることで，各エージェントが自身の利益を最大化するような行動をとった結果として，税収としての社会全体の利益も最大化される，つまり，個人の合理的な選択が社会全体の良い結果となるようなルールを探索することを可能とする．ここでのルールとは税率と罰則確率の２つを指している．本節の残りでは，マルチエージェント強化学習，動的なネットワーク，税収と罰則確率の順に，本研究における具体的な設定について説明する．

4.2 マルチエージェント強化学習（MARL）

本研究では，エージェントが繰り返し囚人のジレンマゲームにおいて最適な行動方策を学習することを目的に強化学習の一種である Q 学習を用いる．Q 学習は，エージェントが報酬を最大化するような行動を学ぶための強化学習アルゴリズムの一種である。この手法では、エージェントは状態と行動のペアに

対して報酬を推定する Q 値を更新する。各エージェントは囚人のジレンマを対戦するたびに Q 値の更新を行う．行動価値関数の更新方法は次のとおりである:

$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t)+\alpha\left[r_{t+1}+\gamma \max_a Q(s_{t+1}, a_{t+1})\right]$

今回の環境では，状態 s_t は相手の直近 2 回の行動によって決定する．ここで s_t は現在の状態，s_{t+1} は遷移先の状態を意味する．また，行動は協調 (C) と裏切り (D) の 2 通りである．エージェントの学習効率を表す学習率 α は 0.5 にした．また，エージェントの長期的な報酬に対する指標である割引率 γ は 0.9 に設定した．Q 関数で用いる報酬 r はエージェントが囚人のジレンマの利得から税収を引いた値の二乗の値である．エージェントの行動の選択方法には $\epsilon-greedy$ 法を用いた．今回は $\epsilon = 0.5$ に設定し，1 回のエピソードが終わるたび 0.99 倍する．これにより，シミュレーションの初期段階では探索を優先し，シミュレーションが進むにつれて最適行動のみを選択するようになる．

4.3 動的な取引ネットワーク

4.3.1 スケールフリーネットワーク

本研究では取引ネットワークとしてスケールフリーネットワークを選択した．スケールフリーネットワークは，ノードの接続数がべき乗則に従っているネットワークであり，多くのノードが少数のエッジを持つ一方で，ごく少数のノードが非常に多くのリンクを持つという構造になっているスケールフリー性は航空網や WWW(World Wide Web) をはじめとして自然界や人間社会の多くの現象に見られる．実際の社会における国同士の貿易関係や企業間の取引関係，個人同士の関係性などもスケールフリー性を有することが示されている [18] ため，今回は取引ネットワークとしてスケールフリーネットワークを選択した．

4.3.2 ネットワークの生成手法

スケールフリーネットワークの生成には，Python ライブラリの networkx の barabasi_albert_graph 関数を用いた．barabasi_albert_graph では，既存のノードに新規エージェントがエッジを接続する方法によってスケールフリーネットワークを生成する．本研究では，ネットワーク上のノードの数は 50，新規ノードが既存ノードに接続する際に追加されるエッジは 2 本という設定でネットワークを生成した．ネットワーク上では，ノードはエージェントを，エッジは取引関係を表現している．エッジで繋がれたエージェント同士は取引関係にあり，繰り返し囚人のジレンマをプレイする．

4.3.3 ネットワークにおけるエッジの新規接続，切断のルール

本研究では，シミュレーションにおいてネットワークが変化する仕組みを導入した．具体的には，エージェントは過去の行動履歴に基づいて新たな取引関係を構築したり，現在の取引関係を断ち切ることができるというものである．この際，新たに取引関係となりうる相手は，自身が直接繋がっているノードと繋がっているノード，つまり，自分から 2 つのエッジを辿って辿り着くことができるノードが対象となる．ネットワークの更新条件にはエージェントの行動履歴における協調の割合を用いた．この，行動履歴に基づく動的なネットワークは，現実世界で見られる，信頼の推移性 [19] に準ずる設定となっている．信頼の推移性とは A,B,C という 3 人がいたときに，A が B を信頼し，B が C を信頼する時，A は C を信頼しやすいというものである．これは，e コマース [20] やソーシャルネットワーク [21] などをはじめとした現実社会で見られる現象であり，今回の動的な取引ネットワークは現実社会の現象を反映したものであると言える．ここで，エージェントが完全に未学習の段階でネットワークの更新を行った場合，初期段階のランダムな選択によりネットワーク構造が変化し，取引ネットワークが崩壊する可能性があるため，これを防ぐためにシミュレーションが 2 割進んだ時点からネットワークの更新が可能となる設定にした．

4.3.4 エージェント同士が新たに繋がる条件

（1） 自分の今までの全ての行動における協調割合が 80%以上である．

（2） 現在自分が繋がっている全てのノードに対して，その各ノードが繋がっているノードが新たに取引関係になりうるノード集合 N である．つまり，自身のノードから 2 本のエッジでたどり着くことができるノードが対象となる．

（3） 各エピソードにおいて各エージェントが自分から新たに繋がることができるノードは 1 つのみであり，ノード集合 N において最も協調割合高いかつ自分とまだ取引関係にないノードが対象となる．

（4） 自分と新規接続候補ノードが繋がる条件は，自身を A, 接続候補の相手を B とした時，自身の協調割合 R_{AC} と相手の協調割合 R_{BC} の積 $(R_{AC}) \times (R_{BC})$ の確率である．例えば，自分の協調割合が 80%，相手の協調割合が 90%の場合，$0.8 \times 0.9 = 0.72$ となり，72%の確率で新たに繋がることとなる．

（5） 新たに接続されたノード同士は取引関係となり，次のエピソードから囚人のジレンマをプレイする．

4.3.5 取引関係が解除される条件

（1） 自分と直接繋がっているノード集合の中で，そのエージェントの協調の割合が 30%未満であるエージェントが取引関係を断ち切ることができる対象である．

（2） 1 − 相手の協調割合 の確率で取引関係が断ち切られる．つまり，対象となるエージェントの協調の割合が 20%である場合，80%の確率で取引関係が断ち切られる．

4.4 税収による管理者の概念と罰則確率

今回の環境では全体の利益の指標として税収による管理者という概念を導入した．管理者はエージェント間の囚人のジレンマにおいて発生する利得の一部を税収として受け取る．エージェントが利得のどの程度を税金として収めるかは，エージェントが囚人のジレンマで選択した戦略によって決定される．具体的には，税率には基本税率と罰則税率の 2 つが存在し，エー

ジェントが協調を選択した際は基本税率に基づいた値を，裏切りを選択した際は，罰則を受ける確率である罰則確率に基づき，基本税率または罰則税率に基づいた値を税収として収める．罰則税率は基本税率より高い設定となっており，裏切りを選択した場合，罰として協調を選択した場合よりも多くの税金を収める可能性が生じる．税率の大きさと罰則確率の違いによりエージェントの戦略がどのように変化し，税収としての全体の利益がどのように変化するかを観察する．

5 実　　験

実験では，提案したシミュレーション基盤を用い，(RQ1)(RQ2)(RQ3) について検証する．

5.1 実験設定

本節では，実験全体を通して設定した取引ネットワークのシミュレーションの設定について記述する．

5.1.1 取　　引

構築した取引ネットワークにおいて，エッジで繋がれたエージェント同士が 1 マッチにつき 5 回の繰り返し囚人のジレンマゲームをプレイする．1 回のエピソードでは (RQ1) では 1000 回，(RQ2)(RQ3) では 1500 回マッチの囚人のジレンマが繰り返される．ここで,「マッチ」とはエッジで繋がれたエージェント同士が定められた回数だけ囚人のジレンマを繰り返し対戦する一連の処理,「エピソード」とはネットワーク全体のすべてのマッチが行われる一連の処理を意味する．

5.1.2 ネットワークの接続，切断条件

ネットワークの接続，切断はエージェントの今までの行動の協調の割合をもとに決定する．新たに接続することができるエージェントは協調の割合が 80%以上のエージェントである．協調の割合がこの値以上のエージェントは，自身と取引関係にあるエージェントの取引関係にあり，自分とは取引関係にないエージェントの中で最も協調の割合が高いエージェントと繋がる機会を得る．自身の協調の割合と相手の協調の割合の積の確率で新たにエッジが接続され取引関係になる．

新規接続の確率＝自身の協調割合 × 相手の協調割合 また，取引関係の切断条件としては，自身と取引関係にあるエージェントの中で協調の割合が 30%以下の確率で取引関係を切断する．

相手と取引関係を断ち切る確率＝ 1 – 相手の協調割合

5.2 実験 1：シミュレーション基盤の妥当性の検証

実験 1 では，(RQ1) を検証するため，囚人のジレンマにおける利得表を変化させた際のエージェントの行動方策の学習結果の変化と，提案フレームワークから生成される取引ネットワークの変化を観察し，現実の取引ネットワークで発生すると考えられる変化が観測可能かを検証する．ここでは，罰則確率と罰則税率を設定せず，エージェントが取引において囚人のジレンマゲームをプレイすると仮定する．本実験では，協調する利得がより高い利得行列（表 2）であるケース 1 と，裏切りの利得がより高い利得行列（表 3）であるケース 2 を設定し，それぞれのケースについて提案フレームワークによるシミュレーションを行う．

表 2: ケース 1 の利得表

	協調 (C)	裏切り (D)
協調 (C)	(6, 6)	(0, 7)
裏切り (D)	(7, 0)	(1, 1)

表 3: ケース 2 の利得表

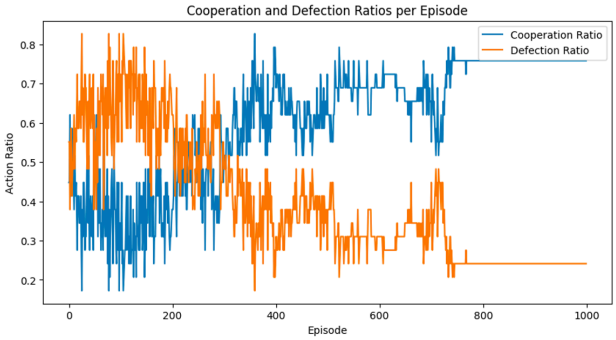
	協調 (C)	裏切り (D)
協調 (C)	(6, 6)	(0, 11)
裏切り (D)	(11, 0)	(5, 5)

5.2.1 結果と考察（実験 1）

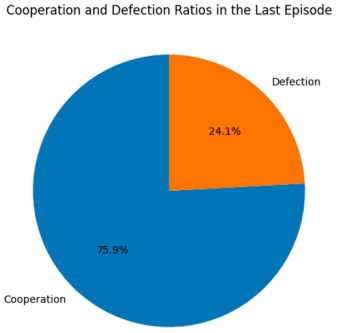
図 2a,3a は，ケース 1 とケース 2 のそれぞれの場合で繰り返しゲームを行った際の，協調割合が高いエージェント（協調エージェント）と裏切り割合が高いエージェント（裏切りエージェント）の数の変化を示した図である．横軸は繰り返しゲームのプレイ数，縦軸はエージェント数を表しており，協調エージェントの数が青色の線，裏切りエージェントの数がオレンジ色の線である．図 8c,3b は繰り返しゲーム終了時における，協調エージェントと裏切りエージェントの割合を示した図である．図 2a をみると，繰り返しゲームの過程でエージェントが徐々に協調割合を高くする戦略を学習していることがわかる．一方，図 3a をみると，徐々に裏切り割合を高くする戦略を学習している．これらの結果から，利得行列によってエージェントが学習する戦略が変化することがわかる．ケース 1 はケース 2 と比較して，協調と裏切りの利得の差が小さく，裏切りよりも協調が優位な環境になったと考えられる．また，ケース 2 はケース 1 と比較して，協調を選択するよりも裏切りを選択することの利得が大きいため，裏切りが優位な環境になったと考えられる．これらの結果から，エージェントが環境から学習する戦略は，人間の合理性と整合する結果が得られたと考察できる．

図 4a,4b はケース 1 の場合，図 5a,5b はケース 2 の場合における，エピソード開始時の取引ネットワークとエピソード終了時の取引ネットワークをプロットした図である．ここで，緑色のノードは協調エージェント，赤色のエージェントは裏切りエージェントを示しており，エッジは取引関係を示している．ケース 1 の場合（図 4a,4b），繰り返しゲーム開始時に存在した取引関係が終了時においても概ね維持されており，取引関係の総数が増加していることがわかる．一方，ケース 2 の場合 (図 5a,5b)，繰り返しゲーム開始時に存在した取引関係が終了時にはほぼ切断されており，取引が発生していないことがわかる．これらの結果から，環境において協調エージェントが多い場合には環境において取引数が増え，裏切りエージェントが多い場合には取引数が減少することが示唆される．この結果は現実において協調する意思決定者が多い場合に，より取引が促進されるという事実に整合する結果になっている．

これらの実験結果は，本シミュレーション環境によって現実の取引ネットワークで発生する変化が観測できることを示唆している．

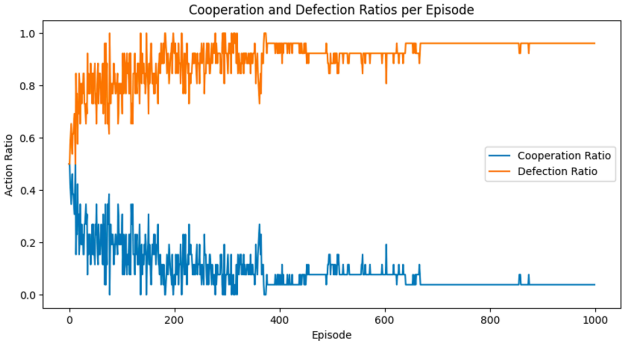


(a) 各ステップにおけるエージェントの学習過程

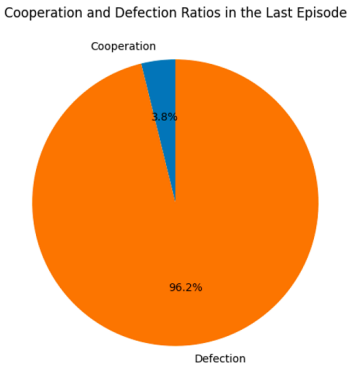


(b) 最終状態における協調・裏切りの割合

図 2: ケース 1 の場合のエージェントの学習結果

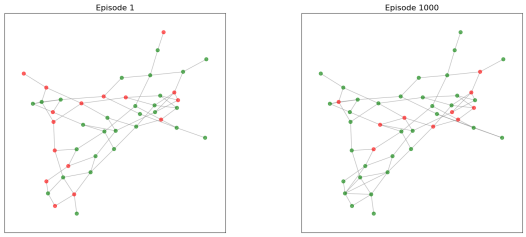


(a) 各ステップにおけるエージェントの学習過程



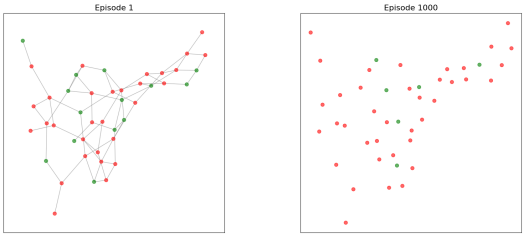
(b) 最終状態における協調・裏切りの割合

図 3: ケース 2 の場合のエージェントの学習結果



(a) ステップ 1　　(b) ステップ 1000

図 4: ケース 1 の取引ネットワークの初期状態と終了時の状態



(a) ステップ 1　　(b) ステップ 1000

図 5: ケース 2 の取引ネットワークの初期状態と終了時の状態

5.3　実験 2：罰則確率と罰則税率による環境の変化の検証

実験 2 では，(RQ2) を検証するため，罰則確率と罰則税率を変化させた際にエージェントがどのような戦略を学習し，税の合計・取引数がどのように変化するかを検証する．

税収は以下の 3 パターンでシミュレーションを行う．

（1） 基本の税率が 10%で罰則時の税率が 40%

（2） 基本の税率が 10%で罰則時の税率が 80%

（3） 基本の税率が 60%で罰則時の税率が 91%

また，各税率パターンにおいて，罰則を受ける確率を 0%,20%,50%,80%,100%の 5 パターンでシミュレーションを行う．罰則の確率が 0%のときはエージェントの選択に関わらず税率は基本の税率で一律，罰則の確率が 100%のときは，裏切りを選択したエージェントは必ず罰則の場合の税率となることを意味する．

実験 2 における利得表は表 4 とする．

表 4: 実験 2 で用いる利得表

	協調 (C)	裏切り (D)
協調 (C)	(3, 3)	(0, 5)
裏切り (D)	(5, 0)	(1, 1)

5.3.1 結果と考察（実験 2）

それぞれの税率において罰則確率を変化させた際のシミュレーション結果についてまとめる．表 5 は基本税率 10%・罰則税率 40%，表 6 は基本税率 10%・罰則税率 80% 表 7 は基本税率 60%・罰則税率が 91%における結果であり，各表はシミュレーション終了時のエピソード時の結果である．表において，「協調」は取引全体における協調の割合,「裏切り」は取引全体における裏切りの割合,「除外」は取引関係を持たないエージェントの割合,「税の平均」は 1 取引あたりに課せられる税の平均,「税の合計」は環境全体で発生した税の合計,「取引数」は環境で発生した取引の総数である．

表 5: 基本の税率が 10%で罰則時の税率が 40%における結果

罰則確率	協調	裏切り	除外	税の平均	税の合計	取引数
0%	0.16	0.18	0.66	2.23	51.2	23
20%	0.272	0.328	0.4	3.33	170	51
50%	0.448	0.432	0.12	5.04	307	61
80%	0.388	0.452	0.16	6.07	474	78
100%	0.5	0.4	0.1	6.90	573	83

表 6: 基本の税率が 10%で罰則時の税率が 80%における結果

罰則確率	協調	裏切り	除外	税の平均	税の合計	取引数
0%	0.16	0.18	0.66	2.23	51.2	23
20%	0.44	0.48	0.08	4.23	372	88
50%	0.82	0.18	0	3.45	987	297
80%	1.0	0	0	3.0	3675	1225
100%	1.0	0	0	3.0	3675	1225

表 7: 基本の税率が 60%で罰則時の税率が 91%における結果

罰則確率	協調	裏切り	除外	税の平均	税の合計	取引数
0%	0.264	0.296	0.44	12.9	400	31
20%	0.376	0.464	0.16	13.8	1051	76
50%	0.904	0.096	0	17.9	4766	267
80%	0.996	0.04	0	18	21224	1179
100%	1.0	0	0	18	22050	1225

これらの実験結果から，いずれの税率パターンにおいても，罰則確率が高くなることで協調が促進され，裏切りと除外エージェントの割合が減少する傾向があることがわかる．また，罰則確率が高くなるほど協調が促進され，協調が支配的な戦略に収束までのスピードも早くためるため，取引ネットワークにおける取引数が増加し，税収も増加することがわかる．この結果は，現実社会での罰則の大きさと確実性が犯罪率に与える影響について調査した Becker の研究 [5] に矛盾しないものとなった．

基本税率と罰則税率の設定により，協調を選択した場合に得る報酬と裏切りを選択し罰則を受けた場合に得る報酬の大小関係がエージェントが学習するエージェントの戦略に影響を及ぼすことがわかる．つまり，裏切りを選択するメリットがなくなるような税率構造にすることでエージェントの協調が促進され，取引数が増加し，税収としての全体の利益につながる．ここで注意したいのが，今回のシミュレーション環境ではエージェントは囚人のジレンマにおける選択とその結果のみを学習しているため，基本税率の大小の違いによって税を支払うか否かを選択することはない．よって，どれだけ基本税を高くしても，罰則との報酬比によっては協調が促進される．

5.4 実験 3：税収が最大となる罰則確率・税率の探索

実験 3 では,（RQ3）を検証するため，本シミュレーション基盤において税収を最大とする罰則確率・税率が探索可能であるかを検証する．本実験では，ベイズ最適化に基づく最適化ライブラリである Optuna [6] を用い，税収を最大化する罰則確率，基本税率，罰則税率を探索する．罰則確率は $[0, 1]$, 基本税率と罰則税率は $[0, 0.99]$ の範囲で探索する．

5.4.1 結果と考察（実験 3）

表 8 は Optuna によって探索された，税収の合計を最大化するする税率・罰則確率であり，表 9 は，探索された税率・罰則確率を用いた際のシミュレーション結果である．実験 2 の結果と比較すると，探索された税率・罰則確率によって得られた税収が増加しており，環境における協調の割合が 100%，取引数が最大となるシミュレーション結果が得られている．一方で，探索された税率は，基本税率が 97%，罰則税率が 99%であり，現実社会における税率とはかけ離れた結果が得られた．これは，環境におけるエージェントが税率に対する不満を抱かないと仮定した設計となっており，協調を選択した際の報酬が裏切りを選択する際の報酬を上回る条件のもとでの最大の税率が探索されたためであると考えられる．

表 8: Optuna によって探索された税収を最大とするパラメータ

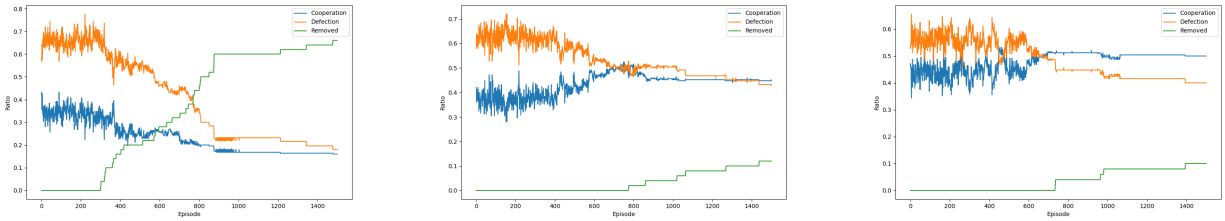
罰則確率	基本税率	罰則税率
96%	97%	99%

表 9: 探索されたパラメータを用いたシミュレーション結果

協調	裏切り	除外	税の合計	取引数
1.0	0.0	0.0	35647.50	1225

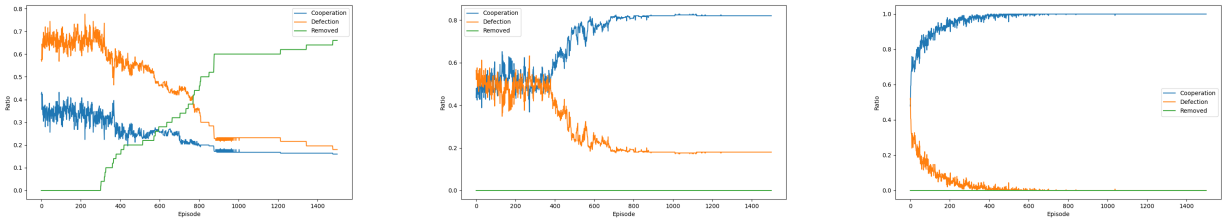
6 ま と め

本研究では，スケールフリーネットワークにおけるマルチエージェント環境を取引ネットワークとして構築し，税収と罰則確率を導入し，エージェント間が囚人のジレンマゲームをプレイすることで戦略のダイナミクスと税収による全体の利益の変化を観察するとともに，この探索基盤の有効性を検証した．税率の大きさや罰則確率を変えて様々なルールでシミュレーションを行うことで，協調や裏切りのエージェントの戦略のダイナミクスとそれに伴う取引数や税収の変化を観察することが可能となり，取引ネットワーク全体と，その環境において意思決定を行う各エージェントの双方にとって，より良い結果をもたすようなルールを発見することが可能となった．



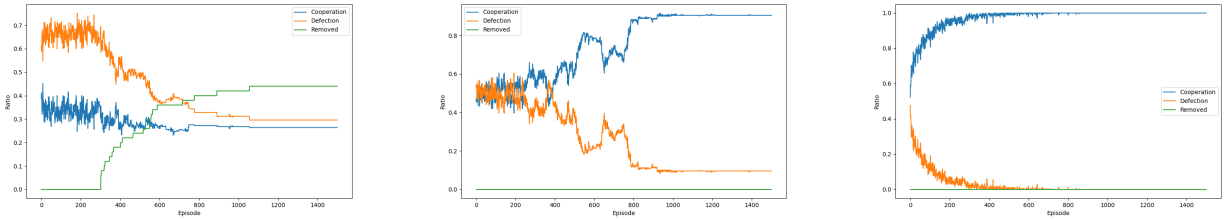
(a) 罰則なし　(b) 罰則確率 50%　(c) 罰則確率 100%

図 6: 基本税率 10%，罰則税率 40%でのエージェントの戦略変化



(a) 罰則なし　(b) 罰則確率 50%　(c) 罰則確率 100%

図 7: 基本税率 10%，罰則税率 80%でのエージェントの戦略変化



(a) 罰則なし　(b) 罰則確率 50%　(c) 罰則確率 100%

図 8: 基本税率 60%，罰則税率 91%でのエージェントの戦略変化

しかし，本研究には多くの課題が残っている．特に，ネットワークの更新条件については検討が必要である．というのも，ネットワークの更新条件に関しては今回は協調裏切の割合が条件を満たすものに対して確率的に更新の可否が決まるアプローチをとっているが，その一定の基準値の正当性に対する検討が不十分であると感じる．また，一度取引ネットワークから除外されたエージェントが取引に復活することは考慮していないので，この設定も改善の余地がある．

また，本研究で提案した探索基盤には多様な拡張方法があると考えられる．例えば，Q 学習における学習率や割引率を変更することにより複数パターンのエージェントを生成したり，各エージェントの収益に応じた取引の接続条件や切断条件の動的な設定，税率の動的な変更，エージェントが税率に抱く不満のモデル化などを考慮することで，より実社会に近いような探索基盤に発展させることができると考えられる．それにより，現実社会の多様な問題に対して，計算論的アプローチから法律やルールなどのを検証することが可能となり，社会の変化スピードに対応したルール形成に有効な手法になりうる．

謝　辞

本研究の一部は JSPS 科研費 (JP22H00508, JP22K17944), JST CREST(JPMJCR22M2)AIP チャレンジの支援を受けたものである．ここに謝意を示す．

文　献

[1] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, Vol. 16, No. 1, pp. 8–37, 1961.

[2] Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, Vol. 6, No. 1, pp. 58–73, 1981.

[3] Paul R. Milgrom and Robert J. Weber. A theory of auctions and competitive bidding. *Econometrica*, Vol. 50, No. 5, pp. 1089–1122, 1982.

[4] Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American Economic Review*, Vol. 93, No. 3, pp. 729–747, June 2003.

[5] Gary S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, Vol. 76, No. 2, pp. 169–217, 1968.

[6] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation

hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[7] Andrew Mao, Lili Dworkin, Siddharth Suri, and Duncan J Watts. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma. *Nature Communications*, Vol. 8, p. 13800, 2017.

[8] Zhiqiang Gou and Ya Li. Prisoner ' s dilemma game model based on historical strategy information. *Scientific Reports*, Vol. 13, No. 1, 2023.

[9] Vassilis Vassiliades and Chris Christodoulou. Multiagent reinforcement learning in the iterated prisoner's dilemma: Fast cooperation through evolved payoffs. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2010.

[10] Maximilian Schaefer. On the Emergence of Cooperation in the Repeated Prisoner's Dilemma. Papers 2211.15331, arXiv.org, November 2022.

[11] Marc Harper, Vincent A. Knight, Martin Jones, Georgios D. Koutsovoulos, Nikoleta E. Glynatsi, and Owen Campbell. Reinforcement learning produces dominant strategies for the iterated prisoner ' s dilemma. *PLoS ONE*, Vol. 12, , 2017.

[12] Takahiro Ezaki, Yutaka Horita, Masanori Takezawa, and Naoki Masuda. Reinforcement learning explains conditional cooperation and its moody cousin. *PLoS Comput Biol*, Vol. 12, No. 7, p. e1005034, 2016.

[13] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, p. 464–473, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[14] Zhao Song, Hao Guo, Danyang Jia, Matjaž Perc, Xuelong Li, and Zhen Wang. Reinforcement learning facilitates an optimal interaction intensity for cooperation. *Neurocomput.*, Vol. 513, No. C, p. 104–113, nov 2022.

[15] Attila Szolnoki, Matjaž Perc, and Zsuzsa Danku. Towards effective payoffs in the prisoner ' s dilemma game on scale-free networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 387, No. 8-9, pp. 2075–2082, 2008.

[16] F. C. Santos and J. M. Pacheco. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.*, Vol. 95, p. 098104, Aug 2005.

[17] Xing Li, Yonghui Wu, Zhihai Rong, and Zhongzhi Zhang. The prisoner's dilemma in structured scale-free networks. *Journal of Physics A: Mathematical and Theoretical*, Vol. 42, No. 24, p. 245002, 2009.

[18] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[19] Audun Jøsang, Elizabeth Gray, and Michael Kinateder. Simplification and analysis of transitive trust networks. *Web Intelligence and Agent Systems: An International Journal*, Vol. 4, No. 2, pp. 139–161, 2006.

[20] Zhishuo Liu, Tian Fang, Yao Dongxin, and Nianci Kou. Transaction credit in the unstructured crowd transaction network. *International Journal of Crowd Science*, Vol. 3, No. 3, pp. 267–283, 2019.

[21] Guanfeng Liu, Yan Wang, and Mehmet A. Orgun. Trust transitivity in complex social networks. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, p. 1222–1229. AAAI Press, 2011.

GANを用いた異常検知における異常度算出方法についての考察

森　仁美† 中野美由紀†† 丸　千尋††† 小口　正人†

† お茶の水女子大学理学部情報科学科　〒112-8610 東京都文京区大塚 2-1-1
†† 津田塾大学学芸学部情報科学科　〒187-8577 東京都小平市津田町 2-1-1
††† お茶の水女子大学文理融合 AI・データサイエンスセンター　〒112-8610 東京都文京区大塚 2-1-1
E-mail: †{hitomi-m,oguchi}@ogl.is.ocha.ac.jp, ††miyuki@tsuda.ac.jp, †††maru.chihiro@is.ocha.ac.jp

あらまし　異常検知の技術は，ネットワークの不正アクセスの検知，医療診断，製造業における部品検査など様々な分野において用いられており研究が盛んに行われている．本論文では，GAN を用いた異常検知において，異常度算出に用いる指標を変化させることで，それらの指標が異常検知の精度に及ぼす影響を評価する．GAN を用いた異常検知の異常度算出として，異常検知を行う対象のデータと生成器により再構築されるデータとの差分を用いた場合を取り上げ，差分と識別器の判定結果の比率を変化させることで，各指標が異常度算出における影響を調べ，異常検知の精度を検討する．また，異常検知モデルの訓練に用いるデータや検知対象の異常データの種類を複数用いて比較し，データの特性にあわせた考察を行う．

キーワード　GAN，異常検知，異常度，情報抽出，汎用機械学習技術

1 はじめに

近年，異常検知の技術はネットワークの不正アクセスの検知や，医療診断，製造業における部品検査など様々な分野において実用化されており，盛んに研究が行われている．昨今の機械学習の発展に伴って異常検知に機械学習が用いられるようになり，その中でも深層学習の利用により高精度な異常検知が可能となった．

深層学習のモデルの 1 種である敵対的生成ネットワーク (GAN) [1] を用いた異常検知についても，既に多くの手法が提案されている．GAN を用いる場合，正常性からの逸脱度合いを示す異常度の算出には，異常検知対象のデータと GAN の生成器により再構築されるデータとの差分である再構築誤差と，識別器の判定結果である識別誤差の 2 つの誤差が指標として用いられる．2 つの指標の比率の変化が異常検知精度に影響を与える可能性があるが，GAN を用いた異常検知の既存論文の多くは異常検知の精度を対象としており，異常度算出時に使用される式について詳細な分析を行っているものはほとんど見られない．

そこで本稿では，GAN を用いた異常画像検知における異常度の算出指標の比率を変化させることで，各指標が異常検知の精度に与える影響について考察を行った．さらに，異常検知モデルの訓練に用いるデータや検知対象の異常データの種類を変えて比較し，データの特性にあわせた分析を行った．

2 関連研究

機械学習を用いた異常検知は，教師あり異常検知と教師なし異常検知の大きく 2 つに分類される．前者は正常または異常を示すラベルが付与されたデータを用いてモデルを訓練し，後者はラベルが付与されていないデータのみを用いて訓練を行う．教師あり異常検知の場合，一般的に正常データに比べて異常データ数は少ないため異常データの特徴をうまく学習できない傾向にある．よって，近年は教師なし異常検知に特に重きが置かれている．

教師なし異常検知は，そのアプローチ方法の違いによって距離ベース，線形モデルベース，確率及び密度推定ベース，深層学習ベースの 4 つに分類される．距離ベースの手法は，データ間の距離に基づき異常検知をするもので，代表的なものに k 近傍との平均距離を計算して異常度を求める K-Nearest Neighbor(KNN) アルゴリズム [2] が存在する．異常データは正常データから距離的に離れているというシンプルな考えに基づく手法だが，計算量が大きくなるという特徴がある．線形モデルベースの手法の代表例としては，主成分分析 (PCA) [3] が挙げられる．PCA は多変量データの情報を次元削減によって要約するもので，正常データが相関をもち多変量ガウス分布に従っている場合，異常検知に有用である．確率及び密度推定ベースの手法は，距離ベースの手法の改良版として提案されたもので，Angle-Based Outlier Detection(ABOD) 法 [4] などが存在する．これはベクトル同士のなす角を利用する手法だが，データ数が多いとやはり計算量も大きくなる傾向にある．深層学習ベースの手法は，大規模なデータに対しても非常に有効で，4 つの手法の中で最もメジャーになりつつある．本手法では Auto-Encoder(AE) [5], GAN，LSTM Encoder-Decoder [6] など様々なモデルが利用される．中でも GAN を用いた手法は，GAN の 2 つのニューラルネットワークが互いに競い合って学習を進めることで，データの複雑な特徴を抽出することが可能であり，さらに，様々なデータ形式に柔軟に適応可能なことから特に注目されている．GAN を用いた異常検知手法には，以下のようなものが存在する．

T. Schlegl らの AnoGAN [7] は，正常データのみで事前学習した通常の GAN モデルを用いてテストデータに含まれる異常

データを検知する手法である．異常検知の際，まず，テストデータに含まれる未知のデータ x に最もよく似たデータを生成できる生成ノイズ z を求める．z を求めたのち，z を生成器に入力してデータを再構築し，生成器の出力から再構築誤差を，識別器の出力から識別誤差を求め，それらを元に異常度を計算する．しかし AnoGAN には，テストデータからノイズ z を求める際に更新学習を行うため異常検知に多くの時間がかかるという特徴がある．本稿で使用した Efficient GAN [8] は Bidirectional GAN(BiGAN) [9] をベースとした異常検知手法である．BiGAN では，通常の GAN の構造にデータからノイズを求めるエンコーダを導入することで，AnoGAN より効率的に学習や生成を行うことを可能にしている．Efficient GAN の詳細については 4 節にて記述する．他にも，GAN の構造に AutoEncoder を導入することで安定した再構築を可能にした GANomaly [10] や，GANomaly に Skip Connection と呼ばれる構造を取り入れてより精度を向上させた Skip GANomaly [11]，生成器と識別器に LSTM を採用し多次元時系列データに対する異常検知を可能にした MAD-GAN [12] などが存在する．

これら多くの研究では，異常検知の精度の向上を目的としており，例えば先の AnoGAN や Efficient GAN の研究では異常度の算出指標である再構築誤差と識別誤差の比率を 9:1 に固定して実験が行われている．よって本研究では，深層学習による教師なし学習を行うモデルに GAN を用いた際の，異常度算出部分に焦点を当て，精度との関係を調べた．

3 GAN について

GAN とは，生成器 G と識別器 D という同時に訓練される 2 つのニューラルネットワークで構成される生成モデルである．GAN のアーキテクチャを図 1 に示す．識別器は，訓練データから抽出した本物データ x，生成器がノイズを元に生成した偽物データ G(z)，の 2 つの入力に対し本物かを示す推定確率を出力し，それを元に誤差を計算して識別器と生成器の訓練を行う．生成器と識別器が互いに競い合って訓練を繰り返すことで，訓練後の生成器は本物データに近い偽物データを生成できるようになる．

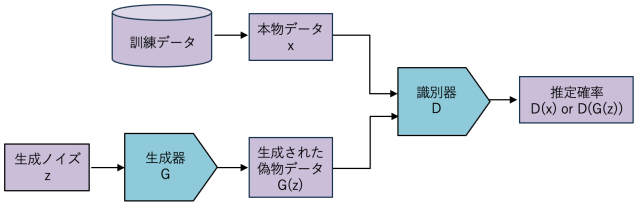


図 1: GAN のアーキテクチャ

4 GAN を用いた異常検知について

異常検知においては，正常データのみを用いて GAN を事前学習することで異常データの検知が可能となる．2 節で述べたように GAN を用いた異常検知手法には様々なものが存在するが，その中でも通常の GAN の構造に，データからノイズを求めるエンコーダ E を加えた Efficient GAN のアーキテクチャについて説明する．

訓練時，

（1） 生成器はノイズ z を入力として本物データに近い生成データ G(z) を出力

（2） エンコーダは訓練データから抽出した本物データ x を入力とし，x に対応するノイズ E(x) を出力

（3） 識別器は画像とノイズのペアを入力とし，いずれのペアが入力されたかを示す推定確率を出力

の流れで事前に学習する．(図 2)

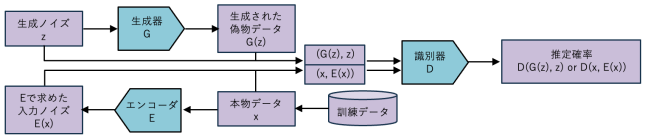


図 2: Efficient GAN 訓練時のアーキテクチャ

異常検知時，

（1） テストデータから取り出した未知のデータ x をエンコーダに入力し，ノイズ E(x) を求める

（2） 求めたノイズ E(x) を生成器に入力してデータを再構築し G(E(x))，これとデータ x の差分を再構築誤差 (residual loss) とする

（3） 識別器にはデータとノイズのペアを入力し，判定結果を識別誤差 (discrimination loss) とする

の流れで異常検知を行う．(図 3)

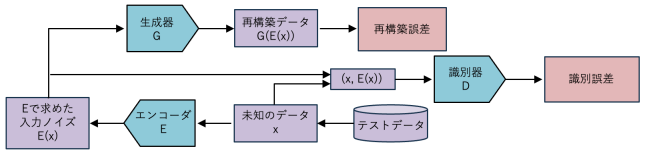


図 3: Efficient GAN 異常検知時のアーキテクチャ

Efficient GAN を用いた異常画像の検知において，再構築誤差はテスト画像と生成画像のピクセルレベルの差の絶対値で求められ，識別誤差はテスト画像とノイズの組，生成画像とノイズの組をそれぞれ入力し，真贋を判定する直前に取り出した特徴量の差の絶対値で求められる．データの異常度は，求めた再構築誤差と識別誤差を元に次式で算出される．

$$\text{loss} = (1 - \lambda) \times \text{residual loss} + \lambda \times \text{discrimination loss}$$

ここで，λ は再構築誤差と識別誤差の比率を制御するための係数であり，λ が小さいほど再構築誤差の影響が大きく，λ が大きいほど識別誤差の影響が大きくなる．また，上の式で求めた異常度が事前に設定した閾値より大きければ，モデルによって異常と予測されたと言える．閾値の設定方法には，テストデータに含まれる異常データの個数を踏まえて閾値を定める方法や，SPOT という確率分布を元に閾値を決定する方法，訓練データの異常度の最大値を用いる方法など複数存在するため，条件やデータの特徴に合わせて適当なものを選択する必要がある．

表 1: 訓練枚数を変えたときの誤差比率と各精度の結果

	200			500			1000			1500		
λ	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0.0	0.65	0.64	0.65	0.73	0.72	0.73	0.82	0.80	**0.81**	0.84	0.82	**0.83**
0.1	0.65	0.64	0.65	0.76	0.74	0.75	0.82	0.80	**0.81**	0.84	0.82	**0.83**
0.2	0.65	0.64	0.65	0.76	0.74	0.75	0.82	0.80	**0.81**	0.80	0.78	0.79
0.3	0.65	0.64	0.65	0.78	0.76	**0.77**	0.80	0.78	0.79	0.76	0.74	0.75
0.4	0.65	0.64	0.65	0.76	0.74	0.75	0.80	0.78	0.79	0.73	0.72	0.73
0.5	0.65	0.64	0.65	0.73	0.72	0.73	0.76	0.74	0.75	0.67	0.66	0.67
0.6	0.65	0.64	0.65	0.71	0.70	0.71	0.71	0.70	0.71	0.65	0.64	0.65
0.7	0.67	0.66	**0.67**	0.65	0.64	0.65	0.67	0.66	0.67	0.55	0.54	0.55
0.8	0.65	0.64	0.65	0.63	0.62	0.63	0.55	0.54	0.55	0.49	0.48	0.48
0.9	0.53	0.52	0.53	0.59	0.58	0.59	0.49	0.48	0.48	0.41	0.40	0.40
1.0	0.43	0.42	0.42	0.53	0.52	0.53	0.43	0.42	0.42	0.35	0.34	0.34

5 実　　験

5.1 実験概要

0 から 9 までの手書き数字の画像データセットである MNIST データ [13] を対象に，Efficient GAN を用いた異常画像検知を行った．その際，以下の 3 つの実験を行ってデータに関する条件を変えた上で，異常度の式の λ を 0 から 1 まで 0.1 ずつ変化させ，再構築誤差と識別誤差の比率を変えたときの精度の変化を調べた．

実験 1　訓練データ (正常データ) と異常データの種類を固定し，モデルの訓練枚数を変化

実験 2　異常データの種類とモデルの訓練枚数を固定し，訓練データの種類を変化

実験 3　訓練データの種類とモデルの訓練枚数を固定し，異常データの種類を変化

全ての実験に共通の実験条件を以下に示す．MNIST データのうち，訓練データ (すなわち正常データ) に 2 種類の数字を，異常データに 1 種類の数字を選択した．正常データに選択した 2 種類の数字の画像を用いて Efficient GAN を事前に学習させ，正常データと異常データの混ざったテストデータに対して異常検知を行った．テストデータ数は 150(うち正常データ:100，異常データ:50)，テストバッチサイズは 64 とする．また，閾値について，今回は参考にした先行研究 [8] と同様に，テストデータに含まれる異常データの個数を踏まえることとした．よって，テストデータの異常度の高い方から 50 番目の値を閾値に設定し，異常度が閾値を超えたデータについて訓練済みモデルが異常と予測したと判断した．その上で，精度を示す値として，モデルの予測結果と実際の結果を元に Precision，Recall，F1-score を算出した．

5.2 実験結果

5.2.1 実験 1：モデルの訓練枚数を変化

訓練データに「4」と「6」，異常データに「2」の画像を用いて，モデルの訓練枚数を各数字，200，500，1000，1500 の 4 段階変化させて実験を行った．表 1 に訓練枚数を変えたときの誤差比率と各精度の結果を，図 4 に各訓練枚数における誤差比率を変えたときの F1-score の変化を示す．

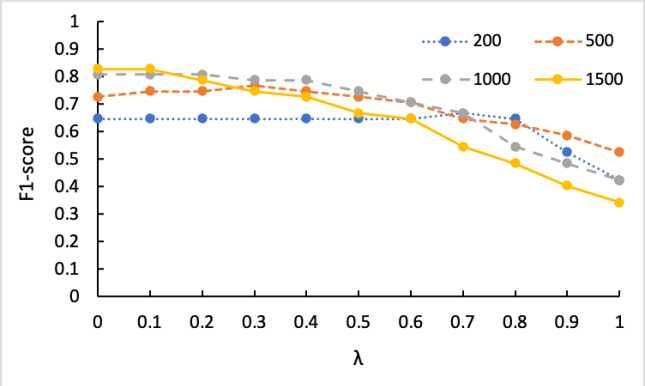


図 4: 各訓練枚数における誤差比率を変えたときの F1-score の変化

表 1 と図 4 より，各訓練枚数において F1-score が最大となった λ の値に着目すると，200 のとき 0.7，1500 のとき 0.0 または 0.1 となっており，訓練枚数が少ないほど λ の値が大きく，枚数を増やすにつれて λ の値が小さくなっていることが読み取れる．また，各訓練枚数における F1-score の最大値を比較すると，枚数を増やすほど最大値が大きくなることがわかる．以上の結果より，200 など訓練枚数が少ない場合は再構築の学習がうまくいっておらず再構築誤差があまり機能していないが，訓練枚数が増える，つまり学習を経ることによって再構築誤差の値が異常と正常で乖離するようになったと考えられる．よって，訓練枚数が多いほど F1-score が高くなり，識別誤差に比べて再構築誤差の方が重要になることが推察される．

5.2.2 実験 2：訓練データの種類を変化

訓練データの種類を，「2」以外の 9 種類の数字から 2 種類を選択した際の組み合わせの総数である 36 通りに変化させて実験を行った．その際，訓練枚数は実験 1 において最も F1-score が高かった 1500 を選択し，異常データに「2」の画像を用いた．実験結果を一部抜粋して訓練データが「1」と「4」，「3」と「9」，「4」と「6」，「6」と「8」，「7」と「8」，「8」と「0」の場合を挙げ，表 2 に誤差比率とそれらの各精度の結果を示した．ただし，表に記載する上では，訓練データ「1」と「4」といった組み合わせを 2 桁の数字 14 などで表している．

表 2: 訓練枚数を変えたときの誤差比率と各精度の結果 (一部抜粋)

	14			39			46			68			78			80		
λ	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0.0	0.78	0.76	0.77	0.86	0.84	**0.85**	0.84	0.82	**0.83**	0.45	0.44	0.44	0.78	0.76	**0.77**	0.80	0.78	0.79
0.1	0.78	0.76	0.77	0.86	0.84	**0.85**	0.84	0.82	**0.83**	0.45	0.44	0.44	0.78	0.76	**0.77**	0.84	0.82	**0.83**
0.2	0.84	0.82	0.83	0.86	0.84	**0.85**	0.80	0.78	0.79	0.47	0.46	**0.46**	0.78	0.76	**0.77**	0.84	0.82	**0.83**
0.3	0.88	0.86	0.87	0.84	0.82	0.83	0.76	0.74	0.75	0.47	0.46	**0.46**	0.78	0.76	**0.77**	0.80	0.78	0.79
0.4	0.92	0.90	**0.91**	0.82	0.80	0.81	0.73	0.72	0.73	0.41	0.40	0.40	0.73	0.72	0.73	0.78	0.76	0.77
0.5	0.90	0.88	0.89	0.78	0.76	0.77	0.67	0.66	0.67	0.41	0.40	0.40	0.71	0.70	0.71	0.73	0.72	0.73
0.6	0.90	0.88	0.89	0.71	0.70	0.71	0.65	0.64	0.65	0.47	0.46	**0.46**	0.67	0.66	0.67	0.73	0.72	0.73
0.7	0.88	0.86	0.87	0.65	0.64	0.65	0.55	0.54	0.55	0.43	0.42	0.42	0.61	0.60	0.61	0.71	0.70	0.71
0.8	0.86	0.84	0.85	0.63	0.62	0.63	0.49	0.48	0.48	0.45	0.44	0.44	0.57	0.56	0.57	0.63	0.62	0.63
0.9	0.86	0.84	0.85	0.53	0.52	0.53	0.41	0.40	0.40	0.43	0.42	0.42	0.55	0.54	0.55	0.53	0.52	0.53
1.0	0.84	0.82	0.83	0.47	0.46	0.46	0.35	0.34	0.34	0.43	0.42	0.42	0.55	0.54	0.55	0.45	0.44	0.44

これより，訓練データ「3」と「9」,「4」と「6」,「7」と「8」,「8」と「0」では λ が小さいほど F1-score が大きい傾向にあるのに対し,「1」と「4」,「6」と「8」では同様の傾向が見られないことがわかり，訓練データによって誤差比率を変化させた際の F1-score の推移傾向は異なることが読み取れる．そこで，F1-score の最大値と最小値の差が 0.15 未満の場合を横ばい，それ以外については最大値と最小値を取る λ の値を元に右下がり，右上がりと定義し，各訓練データにおける誤差比率を変えたときの F1-score の変化を推移傾向別に分けてグラフにした．(図 5，図 6)

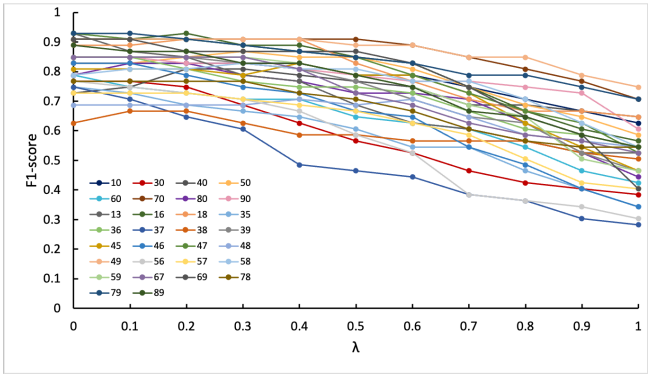


図 5: 各訓練データにおける誤差比率を変えたときの F1-score の変化 (右下がりに推移)

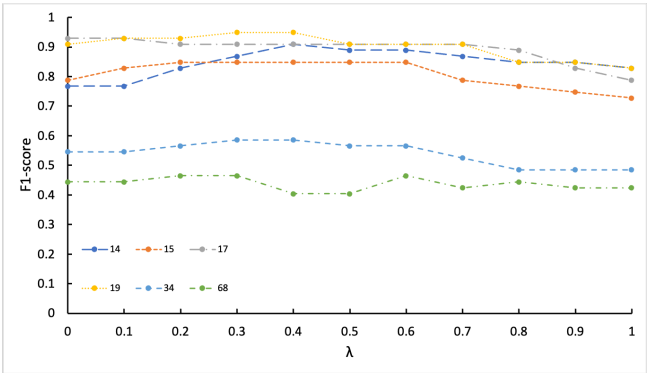


図 6: 各訓練データにおける誤差比率を変えたときの F1-score の変化 (横ばいに推移)

図 5 は訓練データのうち F1-score が右下がりに推移するものを，図 6 は横ばいに推移するものを表している．異常データに「2」を用いた本実験の場合，右上がりに推移する訓練データは存在しなかった．結果より，ほとんどの訓練データが右下がり，すなわち λ の値が小さいほど F1-score が大きくなることがわかる．これらの訓練データでは，F1-score だけでなく，Precision と Recall についても λ が小さいほど高い値が得られた．(表 2) ただし，図 6 より，例外として「3」と「4」,「6」と「8」の様に元々異常検知の精度が良くない場合や,「1」と「4」,「1」と「7」などの様に 2 つの数字の両方が「2」の数字にあるような斜め線を含む訓練データの場合が挙げられる．参考までに，F1-score が右下がりに推移した場合 (図 7) と，横ばいに推移した場合 (図 8) におけるテストデータの再構築画像の例を以下に示す．

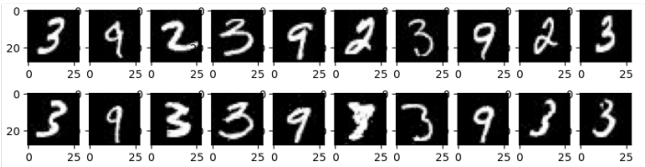


図 7: 訓練データ「3」「9」におけるテストデータの再構築画像 (上段：テストデータ，下段：テストデータの再構築画像)

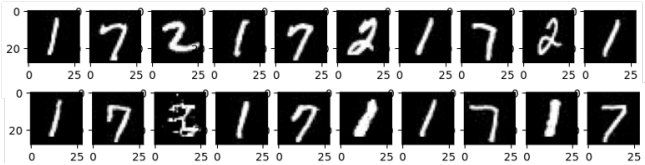
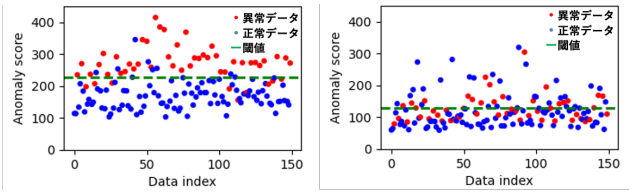


図 8: 訓練データ「1」「7」におけるテストデータの再構築画像 (上段：テストデータ，下段：テストデータの再構築画像)

さらに，図 5 に見られた λ の値が小さい方が精度が良いという傾向は，図 9 の散布図からも読み取れる．図 9 は，訓練データが「4」と「6」，異常データが「2」，訓練枚数が 1500 の場合における λ が 0.1 と 0.9 のときのテストデータの異常度を散布図で表したものである．赤が異常データ，青が正常データ，緑線が閾値を示している．λ が 0.1，すなわち再構築誤差比率が

表 3: 異常データを変えたときの誤差比率と各精度の結果

	0			1			2			3			5			7			8			9		
λ	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0.0	0.67	0.66	0.67	0.45	0.44	0.44	0.84	0.82	**0.83**	0.80	0.78	**0.79**	0.69	0.68	0.69	0.59	0.58	**0.59**	0.71	0.70	**0.71**	0.51	0.50	**0.51**
0.1	0.69	0.68	**0.69**	0.47	0.46	0.46	0.84	0.82	**0.83**	0.78	0.76	0.77	0.69	0.68	0.69	0.57	0.56	0.57	0.69	0.68	0.69	0.49	0.48	0.48
0.2	0.67	0.66	0.67	0.45	0.44	0.44	0.80	0.78	0.79	0.71	0.70	0.71	0.71	0.70	**0.71**	0.53	0.52	0.53	0.69	0.68	0.69	0.45	0.44	0.44
0.3	0.63	0.62	0.63	0.47	0.46	0.46	0.76	0.74	0.75	0.73	0.72	0.73	0.71	0.70	**0.71**	0.51	0.50	0.51	0.65	0.64	0.65	0.41	0.40	0.40
0.4	0.59	0.58	0.59	0.49	0.48	**0.48**	0.73	0.72	0.73	0.69	0.68	0.69	0.67	0.66	0.67	0.49	0.48	0.48	0.65	0.64	0.65	0.33	0.32	0.32
0.5	0.55	0.54	0.55	0.45	0.44	0.44	0.67	0.66	0.67	0.67	0.66	0.67	0.65	0.64	0.65	0.45	0.44	0.44	0.63	0.62	0.63	0.33	0.32	0.32
0.6	0.51	0.50	0.51	0.45	0.44	0.44	0.65	0.64	0.65	0.63	0.62	0.63	0.59	0.58	0.59	0.43	0.42	0.42	0.55	0.54	0.55	0.33	0.32	0.32
0.7	0.49	0.48	0.48	0.45	0.44	0.44	0.55	0.54	0.55	0.53	0.52	0.53	0.53	0.52	0.53	0.39	0.38	0.38	0.49	0.48	0.48	0.35	0.34	0.34
0.8	0.41	0.40	0.40	0.45	0.44	0.44	0.49	0.48	0.48	0.45	0.44	0.44	0.51	0.50	0.51	0.31	0.30	0.30	0.43	0.42	0.42	0.31	0.30	0.30
0.9	0.39	0.38	0.38	0.43	0.42	0.42	0.41	0.40	0.40	0.37	0.36	0.36	0.51	0.50	0.51	0.29	0.28	0.28	0.35	0.34	0.34	0.29	0.28	0.28
1.0	0.35	0.34	0.34	0.43	0.42	0.42	0.35	0.34	0.34	0.31	0.30	0.30	0.49	0.48	0.48	0.27	0.26	0.26	0.33	0.32	0.32	0.29	0.28	0.28



(a) $\lambda = 0.1$　　(b) $\lambda = 0.9$

図 9: 訓練データ「4」「6」におけるテストデータの異常度

高い場合 (a) は赤と青の点が分離しており，異常データの異常度の多くが閾値を上回っている．一方で，λ が 0.9，すなわち識別誤差比率が高い場合 (b) は赤と青の点が混ざっており，閾値によって分離できていない．

以上より，例外は存在するものの，訓練データの種類によらず再構築誤差比率が高いほど精度が良い傾向にあるとわかった．

5.2.3　実験３：異常データの種類を変化

訓練データに「4」と「6」，モデルの訓練枚数を 1500 として，異常データに「4」と「6」以外の 8 つの数字を用いて実験を行った．表 3 に異常データを変えたときの誤差比率と各精度の結果を，図 10 に各異常データにおける誤差比率を変えたときの F1-score の変化を示す．

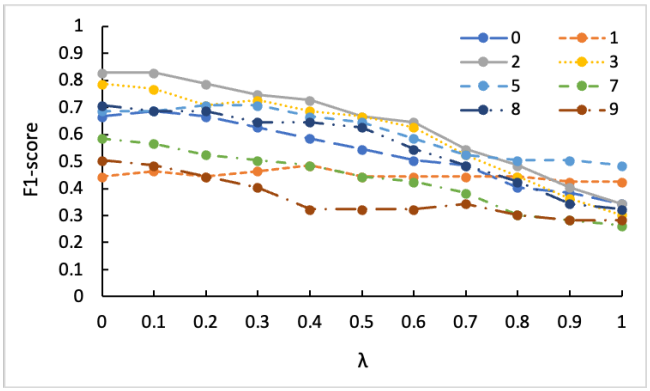


図 10: 各異常データにおける誤差比率を変えたときの F1-score の変化

表 3 と図 10 より，異常データに「2」以外の数字を用いても，ほとんどの場合「2」と同様に λ の値が小さいほど F1-score が高い傾向にあることが読み取れる．唯一，異常データが「1」の場合は誤差比率を変えてもほとんど精度の違いが見られなかった．これは，「4」や「6」を細く書いた場合や，「4」の縦棒を斜めに書いた場合，「6」の縦線がまっすぐな場合など，「1」が訓練データに用いた 2 つの数字に比較的似ているため (図 11)，「1」に対する異常検知の精度が他の数字よりも低いことが原因と考えられる．よって，例外は存在するものの，異常データの種類によらず，識別誤差より再構築誤差の比率が高い方が精度が良い傾向にあるとわかった．

図 11: 異常データ「1」を含むテストデータの再構築画像
(上段：テストデータ，下段：テストデータの再構築画像)

6　まとめと今後の課題

MNIST データに対して Efficient GAN を用いた異常検知を実施し，訓練データや異常データについて条件を変えた上で再構築誤差と識別誤差の比率を変化させ，誤差比率と精度の関係を調査した．実験 1 より，訓練枚数を増加させるほどデータの再構築の学習が進むため，再構築誤差の重要度が高まり，精度も良くなることがわかった．また，実験 2 より，訓練データの種類によらず識別誤差より再構築誤差比率が高い方が精度が良い傾向にあることが，実験 3 より，異常データの種類によらず識別誤差より再構築誤差比率が高い方が精度が良い傾向にあることがわかった．今後は，時系列データなど画像以外のデータに対しても同様の研究を行って傾向を比較する予定である．

文　献

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. Courville, and Y. Bengio. ”Generative Adversarial Nets,” arXiv:1406.2661, 2014

[2] A. Fabrizio and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in In European Conference on Principles of Data Mining and Knowledge Discovery, Berlin, Heidelberg, 2002. Springer, 2002, pp. 15–27.

[3] S. Li and J. Wen, “A model-based fault detection and diagnostic methodology based on pca method and wavelet transform,”Energy and Buildings, vol. 68, pp. 63–71, 2014.

[4] K. Hans-Peter and A. Zimek, “Angle-based outlier detection in high-dimensional data,”in In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 444–452.

[5] J. P. Han Jiawei and M. Kamber, Data mining: concepts and techniques. Elsevier.
[6] H. Edan and A. Shabtai, “Using lstm encoder-decoder algorithm for detecting anomalous ads-b messages,” Computers and Security, vol. 78, 2018.
[7] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, ”Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” In International Conference on Information Processing in Medical Imaging, pp. 146–157, 2017.
[8] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, ”Efficient GAN-based anomaly detection,” arXiv:1802.06222, 2018.
[9] J. Donahue, P. Krahenbuhl, and T. Darrell, ”Adversarial feature learning,” arXiv:1605.09782, 2016.
[10] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, ”GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training,” arXiv:1805.06725v3, Nov 2018.
[11] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, ”Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection,” arXiv:1901.08954v1, Jan 2019.
[12] D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and S. K. Ng, ”MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks,” arXiv:1901.04997, 2019.
[13] LeCun Y, MNIST handwritten digit database, `http://yann.lecun.com/exdb/mnist/index.html`, (2024-2-10)

Empowering Zero-Shot Extreme Multi-Label Text Classification via Weighted Contrastive Learning and Semantic Label Augmentation

Yanan ZHAO† Mizuho IWAIHARA‡

Graduate School of Information, Production and Systems, Waseda University

2-7 Hibikino, Wakamatsu ku, Kitakyushu-shi, Fukuoka, 808-0135 Japan

E-mail: †ynzhao@fuji.waseda.jp, ‡iwaihara@waseda.jp

Abstract Zero-shot extreme multi-label text classification is a complex task that involves assigning multiple relevant labels to a text from an extensive label space, even when no training examples are available for the labels. This challenge has garnered significant attention in natural language processing. In this paper, we propose an innovative approach to enhance the performance of zero-shot extreme multi-label text classification. Our method revolves around two crucial components: Contrastive learning and label augmentation. In contrastive learning, we combine the strategies of weighted filtering, hard negative sampling, and prompt learning. Contrastive learning serves as the foundation, enabling the extraction of informative features from textual data. By integrating hard negative sampling, the model is focusing more on challenging instances, thereby improving its ability to generalize to unseen labels and nuances within the data. Incorporating prompts introduces guided context and domain knowledge, which is particularly valuable in the zero-shot scenario where the corpus is unlabeled. Regarding label augmentation, we utilize a large language model to enrich labels by appending contextual words, enhancing the model's capability to differentiate highly similar labels. Additionally, we generate labels that are semantically similar to the original ones, aiming to construct positive pairs for effective contrastive learning. The combined effect of these techniques creates a comprehensive framework for tackling the intricacies of zero-shot extreme multi-label text classification. To validate our approach, we conduct extensive experiments on benchmark datasets. The results of our method are superior to the reproduced SOTA model on single GPU setting. This underscores the effectiveness of our approach in addressing the challenges of zero-shot multi-label text classification.

Keyword Extreme Multi-Label Classification, zero-shot, contrastive learning, label augmentation

1. Introduction

In recent years, the field of natural language processing (NLP) has witnessed remarkable advancements in various tasks, including text classification. Among these tasks, extreme multi-label text classification (XMC) has emerged as a significant challenge. This task involves assigning a set of relevant labels to a text document from an extensive label space. With the explosion of digital content and the proliferation of online platforms, the need for accurately categorizing and tagging content has grown extensively. Extreme multi-label text classification finds applications in recommendation systems [1], content organization, and information retrieval [2][3], playing a pivotal role in enhancing user experience and content management.

Traditionally, text classification methods have relied on supervised learning [4], where models are trained on labeled examples to predict labels for new, unseen data. However, in scenarios where the label space is vast and continually evolving, acquiring labeled examples for all possible labels becomes impractical. This has led to the emergence of zero-shot learning, where models are expected to generalize to labels that were not seen during training. Zero-shot extreme multi-label (ZXMC) text classification extends this challenge further by requiring models to predict multiple labels for each document, even when the document-label pairs were not observed during training [5]. Figure 1 shows the difference between XMC and ZXMC.

To address these challenges, researchers have explored various strategies, such as leveraging pretrained language models like BERT [5] or using specialized techniques like contrastive learning [6]. These techniques aim to extract meaningful representations from text, handle the absence of labeled examples, and incorporate external knowledge to guide the model's decision-making process.

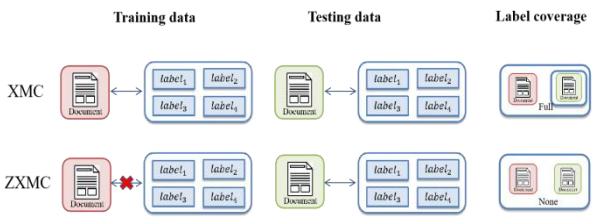


Figure1: The difference between XMC and ZXMC

2. Related Work

2.1 Zero-Shot Extreme Multi-label Text Classification

Zero-shot extreme multi-label text classification (ZXMC)

refers to a common task in natural language processing (NLP). Most of the existing studies have constructed positive pairs. For instance, MACLR [5] leverages the raw text with techniques including multi-scale adaptive clustering, label regularization, and self-training with pseudo-positive pairs. [6] proposed Randomized Text Segmentation and positive pair construction strategies to exploit the structure within document data for end-to-end contrastive learning.

2.2 Contrastive Representation Learning

Contrastive Representation Learning is a deep learning approach aimed at acquiring high-quality data representations by bringing similar samples closer and pushing dissimilar ones further apart. Contrastive learning has its roots in Siamese networks [7], which initially emerged for face verification but were later adapted for various representation learning tasks. Triplet networks [8] further refined the concept by introducing triplets of samples (anchor, positive, and negative) to better control the learning process. Variants like MoCo [9] introduced momentum updates and SimCLR [10] simplified and improved the approach with data augmentation and large-batch training. BYOL [11] enhanced performance through self-supervised learning and dual-network architectures.

2.3 Hard Negative Sampling

Hard negative sampling is a crucial technique in machine learning, particularly in the context of training object detection and classification models. It involves the identification and prioritization of challenging or " hard " negative samples during the training process to improve model performance. A number of related work and strategies have been proposed in this area: One notable approach is Online Hard Example Mining (OHEM) [12], which dynamically selects and focuses on the most challenging negative samples to prevent the model from being overwhelmed by easily classified negatives. Another technique, known as Hard Example Generation (HEG) [13], augments the training dataset with synthetically generated hard negative examples.

2.4 Prompt-based Language Models

Prompt-based language models have gained significant attention in recent years due to their versatility and effectiveness in various natural language understanding tasks. These models are built upon the foundation of large pre-trained neural networks, such as GPT-3 [14], and have shown remarkable performance across a wide range of applications. PET [15] extended the concept of prompt-based models by proposing a method for designing prompts to tackle few-shot text classification tasks. PromptBert [16] explored the use of prompts for sentence embedding tasks, bridging the gap between unsupervised and supervised methods. The development of these prompt-based models has opened up new avenues for natural language understanding and has inspired various approaches that leverage prompts to enhance model capabilities.

2.5 Label Augmentation

Label augmentation [17] is a machine learning technique employed to enhance model performance, particularly beneficial in scenarios characterized by data scarcity or constrained label availability. This approach extends the original dataset by generating additional labels for training samples, thereby enriching the information available for model training and enhancing its capacity for generalization.

Pretrained language models have become mainstream models in recent years due to their excellent performance. Masked language models (MLMs) used for training of BERT and RoBERTa can predict masked words in text based on context, which can be used for text data augmentation. [18][19] use both word embeddings and masked language models to obtain augmented data. They apply the tokenizer of BERT to tokenize words into multiple word pieces. Each word piece is replaced with probability 0.4. If a word piece is not a complete word ("est" for example), it is replaced by its K-nearest-neighbor words in the Glove embedding space. If the word piece is a complete word, the authors replace it with [MASK] and employ BERT to predict K Words to fill in the blank. [20][21] and [22] apply methods that are similar to [19]. They mask multiple words in a sentence and generate new sentences by filling these masks to generate more varied sentences. [17] uses BERT pre-trained masked language model (MLM) to predict vocabularies relevant to label names for aspect category detection (ACD) task.

3. Methodology

Figure 2 shows the overview of our proposed new model for ZXMC. We first describe the problem setting of our task. Then, we introduce the strategy of positive pair selection, hard negative sampling, prompts in contrastive learning, and finally label augmentation.

3.1 Problem definition

In ZXMC task, we denote $X = \{(t_1, c_1), \cdots, (t_{|X|}, c_{|X|})$ as a set of documents, where t_i and c_i represent the title and content of the document respectively. Let $Y = \{y_1, \cdots, y_{|Y|}\}$ be the set of labels. Each $y_i \in Y$ is a short sentence description. In the ZXMC setup, the document set X is divided into two non-overlapping subsets X_{train} and X_{test}, one for training and

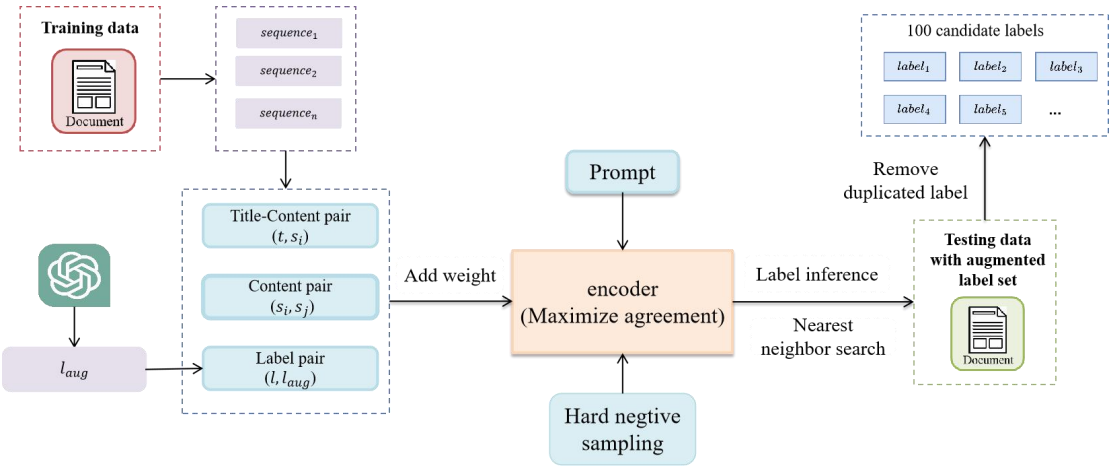


Figure 2: The overview of our proposed model

the other for evaluation. We have access to X_{train} and Y for training a model to classify documents to labels. We would like to correctly classify each unseen document in X_{test} to labels. Since ZXMC is an unsupervised task, we do not have access to M (the true mapping of documents to labels) for training.

3.2 Positive pair construction

In this subsection, we introduce the positive pair selection strategy for ZMTC. RTS [6] constructs many positive pairs which contain title pairs, title-content pairs, and label pairs for contrastive learning, SimCSE [23] is a powerful text similarity calculation method that performs particularly well in unsupervised learning scenarios. We use SimCSE to evaluate positive pairs generated by RTS for contrastive representation learning. As shown in Figure 3, most of the sentences are not similar, so the quality of positive alignment is not ideal.

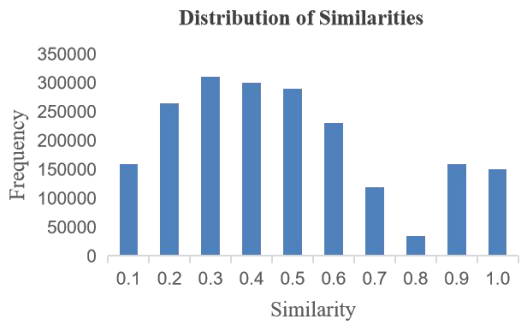


Figure 3: Similarity distribution of positive pairs

Based on the similarity distribution of positive pairs, we can give different weights to make the model pay more attention to positive pairs with higher similarity in contrastive learning. So we consider the following three strategies:

A. Use similarity as weight.

B. Introduce a threshold, and pairs having similarity less than the threshold is given a smaller weight. If the similarity is greater than the threshold, the weight remains the same.

C. Giving different weights to different types of positive pairs. Here, we give weight 0.3 to title pairs and 0.7 to title-content- pairs.

3.3 Hard negtive sampling

Hard negative sampling strategy selects challenging negative samples to improve model learning and enhance its ability to distinguish between positive and negative classes, leading to improved test performance. HCL [24] is a simple and effective hard negative sampling method, which is particularly useful in situations with limited resources. Inspired by HCL, first we calculate the similarity scores of positive samples and negative samples. Let Pos and Neg denote the exp functions over inner products for positive examples and negative examples respectively, as shown below. In each iteration, we sample a batch of positive pairs $\{(x_i, \hat{x}_i) \mid i \in \{1,...,b\}\}$ with size b and minimize the loss.

$$Pos = e^{\sum E(x_i)\cdot E(\hat{x}_i)/\tau} \tag{1}$$

$$Neg = e^{(E(x_i)+E(\hat{x}_i))\cdot(E(x_i)+E(\hat{x}_i))^T/\tau} \tag{2}$$

Here τ represents the temperature hyperparameter. The original loss is as follows:

$$Loss = -\frac{1}{b}\sum_{i=1}^{b}\log\frac{Pos}{Pos+Neg} \tag{3}$$

Specifically, regarding hard negative sampling, we reweight the similarity score of the negative sample, by the importance weight W calculated from the similarity score of the negative sample. β is concentration parameter.

$$W = e^{\beta \cdot \log(Neg)} \tag{4}$$

And use it to reweight the negative sample score RN, and finally obtain Ng.

$$\mathrm{RN} = \frac{\sum_{i=1}^{b}(W \cdot Neg)}{\overline{W}} \tag{5}$$

$$Ng = \max\left(\frac{-\tau^{+} \cdot N \cdot Pos + RN}{1-\tau^{+}}, N \cdot e^{-\frac{1}{\tau}}\right) \tag{6}$$

Here, τ^{+} is positive class priorx. N represents the number of negative samples in a batch. In order to ensure the numerical stability of the loss calculation and the rationality of the results, we set a lower limit for Ng. Finally, we derive the contrastive loss using the positive sample score Pos and the negative sample score Ng by taking their average:

$$Hard_Loss = -\frac{1}{b}\sum_{i=1}^{b}\log\frac{Pos}{Pos+Neg} \tag{7}$$

3.4 Prompt-based sentence representation

To address the challenges associated with unsupervised training, we leverage the concept of "Prompt" to provide guidance to the model, enabling it to acquire more knowledge related to sentence embedding tasks. Inspired by CLSEP [25], during the training phase, the sentence template is structured as follows: " This sentence: ' S ' means [MASK]. " During the evaluation phase, the sentence template is modified to: " The meaning of ' S ' is [MASK]. " where ' S ' denotes the sentence, and [MASK] represents the placeholder for the [MASK] token. In essence, this approach involves using prompts to structure the training and evaluation phases differently, with the aim of enhancing the model's understanding of sentence embeddings through unsupervised learning.

3.5 Label augmentation

With the development of large-scale language models and their widespread use in the field of natural language processing, certain progress has been made in various tasks. Therefore, in the label augmentation part, we use the large language model ChatGPT for label augmentation.

3.5.1 Rewriting and rephrasing labels

ChatGPT has the ability to generate new sentences from its extensive training on diverse language patterns and contexts. In order to obtain more and richer training data, we rewrite and rephrase the labels to obtain semantically similar sentences to form positive pairs for comparative learning. We design two templates to guide ChatGPT content generation.

Template 1: Rewrite [x] so that it is about the same length as the original sentence.

Template 2: Generate three sentences by rephrasing [x] while aiming for a consistent length.

In these templates, [X] represents a label, which is a short sentence description. Some examples of generated results are shown in Table 1.

Table 1: Examples of original labels and rewritten labels

Original label	Rewritten label
Dreams, A Portal to the Source	Dreams:A Gateway to the origin
Elephants Can Remember	Elephants retain memories
In the Loyal Mountains	In the faithful hills

3.5.2 Using ChatGPT for distinguishing similar labels

For semantically similar and indistinguishable labels, we also use ChatGPT to add additional words to distinguish them. Design a template in the format " [X], add a word to make it easier to distinguish ". Here, [X] represents a label. As shown in Table 2, ChatGPT has certain prior knowledge about labels and can add specific words to very similar labels to distinguish them.

Table 2: Example of adding distinguishing words using ChatGPT

Example	Original label	Added words
Example 1	The Diary of Anais Nin, Vol. 2: 1934-1939	Intimate Reflections
	The Diary of Anais Nin, Vol. 3: 1939-1944	World in Turmoil
	The Diary of Anais Nin, Vol. 6: 1955-1966	Evolving Years
Example 2	Battle Royale,vol.6	Requiem
	Battle Royale,vol.7	Reckoning

3.6 Label inference

In the process of contrastive learning, we use the augmented label pairs to replace the original label pairs for training. After the training in contrastive learning, we use the nearest neighbor search framework [26] for inference, where the same cosine similarity is used as the distance metric to find 100 candidate labels in the augmented label set.

Table 3: The example of duplicate labels

Label-uid	Label-content	Added word
B0000057DB	Handel: Messiah	Composition
B000051YDC	Handel: Messiah	Composition
B0000266WD	Handel: Messiah	Composition
B00000416H	Handel: Messiah	Composition

As shown in Table 3, in the results of the prediction phase,

there are multiple labels with different uid's but the same content. Since the content is the same, even after label augmentation, the labels are still repeated, so in the evaluation phase, by removing these duplicate labels, we can obtain a refined prediction result. The labels in the refined result are richer and more diverse. The process of removing duplicate labels is by the following steps.

Step 1: Use the nearest neighbor search.

Step 2: Iterate through each query's retrieval results. Create a dictionary for label contents.

Step 3: For each label in the results, verify if the label's content is in the dictionary. If yes, skip to avoid duplication. Otherwise, add it to the dictionary.

4. Experiments

4.1 Evaluation Metrics

We adopt precision at $\mathrm{p}, \mathrm{p} \in \{1,3,5\}$ and recall at $\mathrm{r}, \mathrm{r} \in \{1,3,5,10,100\}$ as the evaluation metrics for ZXMC task, which are defined as:

$$\mathrm{P@p} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{y} \in \mathrm{Y}_{\mathrm{i}}^{\mathrm{pred}}} 1_{\mathrm{i}}^{\mathrm{top-p}}(\mathrm{y})}{\mathrm{np}} \tag{8}$$

$$R@r = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{y \in Y_i^{pred}} 1_i^{top-r}(y)}{\sum_{y \in Y} 1_i^{top-r}(y)} \tag{9}$$

where n is the number of documents evaluated, $\mathrm{Y}_{\mathrm{i}}^{\mathrm{pred}}$ is the set of predicted labels for the i-th document, and $1_{\mathrm{i}}^{\mathrm{top-p}}(\cdot)$ is an indicator function indicating whether a predicted label is a ground truth top-p label for the ith document.

4.2 Datasets

Table 4 presents the statistics of the benchmark dataset. We adopt LF-Amazon-131K from the extreme classification repository [27]. LF-Amazon-131K contains item-to-item recommendation data from the e-commerce website Amazon. The dataset uses data from large-scale recommendation systems, so it is ideal for evaluating the real-world performance of models.

Table 4: Statistics of the datasets used for evaluation. $|\mathrm{X}_{|\mathrm{train}|}|$, $|\mathrm{X}_{\mathrm{test}}|$,$|\mathrm{Y}|$ denote the number of training instances, the number of test instances, and the number of labels, respectively.

Dataset	$\lvert X_{\lvert train \rvert} \rvert$	$\lvert X_{test} \rvert$	$\lvert Y \rvert$
LF-Amazon-131K	294,805	134,835	131,073

4.3 Experiment Settings

We implement our approach with PyTorch [28]. Our experiments are conducted on a machine with GeForce RTX 3090 GPU. We finetune the basic version of MPNet [29] using our proposed method to select positive pairs. The batchsize is 64 and trained for 4 epochs with the AdamW optimizer. The temperature hyperparameter is set to $\tau = 0.5$ and $\tau^+ = 0.1$. We set the initial learning rate to 10^{-5}, and the learning rate decays by a factor of 10 within epochs according to a linear schedule [30].

Regarding text segmentation methods, we used natural sentence segmentation and random text segmentation and compared their results. Natural text segmentation is to divide the text into independent sentences. Randomized Text Segmentation breaks up the contents of a document into non-overlapping subsequences with lengths sampled from the discrete uniform distribution.

4.4 Baselines

MACLR [5]: A multi-stage contrastive learning method that uses clustering and TF-IDF to construct pseudo-labels.

RTS [6]: An end-to-end contrastive learning model applying random text segmentation.

5. Results and Analysis

5.1 Results

The model combining the following methods showed the best results, and we named it **W**eighted **C**ontrastive **L**earning with **L**abel **A**ugmentation (WCLLA). Split the text of the training set using random text segmentation, the label pairs are rewritten and rephrased for augmenting label pairs. The weight strategy is strategy B, and the threshold is set to 0.8. Contrastive learning incorporates hard negative sampling and prompts. The label set in the evaluation stage uses the augmented new label set.

Table 5 shows the performance of our method for the ZXMC task and the comparisons with the baseline methods. In the case of multi-GPU, although RTS is still the best result currently, our result on P@1,R@1,R@100 is second only to RTS, and according to the results we reproduced on a single GPU, the results of our method is better than RTS. This proves that our method is effective.

5.2 Impact of weighting strategy

Following our positive pair selection strategy, we compare the best results in each strategy, and also compare the case of natural sentence segmentation (NS) and the case of random text segmentation (RTS). As shown in Table 5, when the segmentation method is RTS, filtering leads to a decrease in results. However, with natural segmentation, filtering results in an improvement.

For the RTS method, random segmentation can result in text being segmented into irrelevant parts that may not be semantically clearly related. When giving greater weight to

positive pairs with higher similarity, the model may pay too much attention to these irrelevant or incoherent parts, leading to performance degradation. For the NS method, the method of setting the threshold has the best results. When the threshold is 0.8, most of results are the best. When the threshold is 0.9, the values of R@10 and R@100 are the highest. In the case of only positive input, the high threshold setting would make the model pay more attention to positive pairs with higher similarity, making the result better.

Table 5: Precision and recall metrics of our method and other baselines. For each metric, the best value is bolded and the second best value is underlined.

Method.	GPU	P@1	P@3	P@5	R@1	R@3	R@5	R@10	R@100
RTS	1	17.59	14.57	11.25	10.03	23.20	28.86	35.66	54.91
WCLLA	1	**18.20**	**14.86**	**11.55**	**10.36**	**23.52**	**29.62**	**36.86**	**55.89**
MACLR [5]	10	18.13	**15.42**	11.93	10.35	**24.45**	30.43	37.28	54.99
RTS [6]	4	**18.74**	15.30	**11.96**	**10.64**	24.16	**30.45**	**38.19**	**59.34**
WCLLA	1	18.20	14.86	11.55	10.36	23.52	29.62	36.86	55.89

Table 6: Precision and recall metrics of Positive pair selection. For each metric, the best value is bolded. A represents the strategy using similarity as weight. B(0.8) represents the threshold is 0.8. C represents the strategy using pair difference.

Segmentation method	Weighting Method	P@1	P@3	P@5	R@1	R@3	R@5	R@10	R@100
RTS (Randomized text segmentation)	A	16.98	14.00	10.80	9.65	22.21	27.69	34.28	53.45
	B (0.8)	16.80	13.63	10.56	9.60	21.73	27.12	33.76	52.02
	B (0.9)	16.68	13.56	10.49	9.51	21.60	26.98	33.62	51.88
	C	16.79	13.56	10.48	9.55	21.65	27.10	33.68	51.98
	Orignal	**17.59**	**14.57**	**11.25**	**10.03**	**23.20**	**28.86**	**35.66**	**54.91**
NS (Natural segmentation)	A	17.07	14.12	10.87	9.72	22.40	27.92	34.39	52.94
	B (0.8)	**17.16**	**14.17**	**10.91**	**9.77**	**22.52**	**28.02**	34.50	53.90
	B (0.9)	17.01	13.96	10.86	9.63	22.14	27.81	**34.69**	**54.94**
	C	16.81	13.89	10.67	9.54	22.03	27.56	34.23	53.78
	Orignal	16.69	13.65	10.58	9.43	21.65	27.12	33.92	53.81

Table 7: Precision and recall on different number of label pairs. (The segmentation method is RTS and the weight is 0.8,evaluate using original label set)

Label pair	Number of label pairs	P@1	P@3	P@5	R@1	R@3	R@5	R@10	R@100
Rewrite	126727	17.70	14.69	11.42	10.12	23.38	29.31	36.50	55.26
Rephrase	430348	17.72	14.79	11.44	10.09	23.46	29.42	36.64	55.64
Rewrite+Rephrase	557075	**17.81**	**14.80**	**11.49**	**10.14**	**23.49**	**29.48**	**36.77**	**55.77**

5.3 Impact of number of label pairs

We generated the rewritten sentences using the two templates and used them to construct new label pairs. Different kinds and numbers of label pairs are evaluated. As shown in Table 7, the greater the number of label pairs, the results are slightly improved.

6. Conclusion and Future Work

We proposed a novel zero-shot extreme multi-label method for extreme multi-label text classification in this paper. Compared with existing methods, our method has three innovative points. The first one is to add weights in filtering positive pairs to obtain more effective sample pairs. The second is to use hard negative sampling strategies and prompts for contrastive learning. The third one is using a large language model to do label name augmentation. We evaluate our method on the publicly available ZXMC dataset. The results of our method are superior to the reproduced SOTA model on single GPU setting. This demonstrates the effectiveness of our approach.

For the future work, we suppose that we can make efforts in the following aspects. First, this model does not have a ranking part. With the emergence of large-scale language models such as ChatGPT, LLaMA, etc., it is feasible to guide these models in label ranking. We have used ChatGPT to do initial experiments on reranking, but the idea is not yet mature and needs to continue to be explored. Next, generative models are also a possible direction. We can also combine large language models for label generation.

Reference

[1] Prabhu, Y., Kag, A., Gopinath, S., Dahiya, K., Harsola, S., Agrawal, R., & Varma, M. (2018, February). Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 441-449).

[2] Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen.2021. Zero-shot text classification via knowledge graph embedding for social media data. IEEE Internet of Things Journal.

[3] Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu,Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al. 2021. Extreme multi-label learning for semantic matching in product search. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 2643–2651.

[4] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[5] Yuanhao Xiong, Wei-Cheng Chang, Cho-Jui Hsieh,Hsiang-Fu Yu, and Inderjit S. Dhillon. 2022. Extreme zero shot learning for extreme text classification. In NAACL 2022.

[6] Zhang, Tianyi, et al. "Structural Contrastive Representation Learning for Zero-shot Multi-label Text Classification." Findings of the Association for Computational Linguistics: EMNLP 2022. 2022.

[7] Chicco, D. (2021). Siamese neural networks: An overview. Artificial neural networks, 73-94.

[8] Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3 (pp. 84-92). Springer International Publishing.

[9] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

[10] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[11] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.

[12] Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 761-769).

[13] Lee, H., Kwon, H., & Kim, W. (2021). Generating hard examples for pixel-wise classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 9504-9517.

[14] Dale, R. (2021). GPT-3: What's it good for?. Natural Language Engineering, 27(1), 113-118.

[15] Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2339–2352, Online. Association for Computational Linguistics.

[16] Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., ... & Zhang, Q. (2022, December). PromptBERT: Improving BERT Sentence Embeddings with Prompts. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8826-8837).

[17] Zeyu Wang and Mizuho Iwaihara: Few-Shot Multi-Label Aspect Category Detection Utilizing Prototypical Network with Sentence-Level Weighting and Label Augmentation, Proc. 34th Int. Conf. on Database and Expert Systems Applications (DEXA2023), LNCS Vol.14147, pp.363-377, Aug. 2023.

[18] LI, Bohan; HOU, Yutai; CHE, Wanxiang. Data augmentation approaches in natural language processing: A survey. Ai Open, 2022, 3: 71-90.

[19] Jiao, Xiaoqi, et al. "TinyBERT: Distilling BERT for Natural Language Understanding." Findings of the Association for Computational Linguistics: EMNLP 2020.

[20] TAPIA-TÉLLEZ, José Medardo; ESCALANTE, Hugo Jair. Data augmentation with transformers for text classification. In: Mexican International Conference on Artificial Intelligence. Cham: Springer International Publishing, 2020. p. 247-259.

[21] LOWELL, David, et al. Unsupervised Data Augmentation with Naive Augmentation and without Unlabeled Data. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. p. 4992-5001.

[22] PALOMINO, Daniel; LUNA, José Ochoa. Palomino-Ochoa at TASS 2020: Transformer-based Data Augmentation for Overcoming Few-Shot Learning. In: IberLEF@ SEPLN. 2020. p. 171-178.

[23] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." arXiv preprint arXiv:2104.08821 (2021).

[24] Robinson, J. D., Chuang, C. Y., Sra, S., & Jegelka, S. (2020, October). Contrastive Learning with Hard Negative Samples. In International Conference on Learning Representations.

[25] Wang, Q., Zhang, W., Lei, T., Cao, Y., Peng, D., & Wang, X. (2023).

CLSEP: Contrastive learning of sentence embedding with prompt. Knowledge-Based Systems, 266, 110381.

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019.Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547.

[27] K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal,Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style,high-performance deep learning library. In NeurIPS.

[29] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and TieYan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33:16857–16867

[30] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年2月28日(水) 16:00 ~ 18:10 | T1-B オンライン（Zoom Events)

言語表現認識

座長:片山 薫(東京都立大学)
コメンテータ:金子 邦彦(福山大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T1-B-3-01] 日本語における大規模言語モデルの生成文検出

*丸井 渚生[1]、曹 洋[1]、中村 篤祥[1] (1. 北海道大学)

16:20 ~ 16:45
[T1-B-3-02] ポジティブな単語を含んだ煽り・誹謗中傷目的のコメント検出方法の提案

*佐藤 豪[1]、清 雄一[1]、田原 康之[1]、大須賀 昭彦[1] (1. 電気通信大学)

16:45 ~ 17:10
[T1-B-3-03] 動画に対する否定的コメント・フィルタリングにおけるfew-shot学習の評価

*三田寺 聖[1]、宇津呂 武仁[1] (1. 筑波大学)

17:10 ~ 17:35
[T1-B-3-04] Conceptual Commonsense-aware Humor Recognition with Various Pretrained Masked Language Models

*佐々木 裕多[1]、張 建偉[2]、白石 優旗[3] (1. 東京工業大学、2. 岩手大学、3. 筑波技術大学)

17:35 ~ 18:00
[T1-B-3-05] 皮肉検出モデルにおける適切なコモンセンスの選別と性能評価

*皆川 翔哉[1]、佐々木 裕多[2]、張 建偉[1] (1. 岩手大学、2. 東京工業大学)

日本語における大規模言語モデルの生成文検出

丸井　渚生† 　曹　　洋†† 　中村　篤祥††

† 北海道大学工学部情報エレクトロニクス学科　〒 060—0814 北海道札幌市北区北十四条西 9 丁目
†† 北海道大学大学院情報科学研究院　〒 060—0814 北海道札幌市北区北十四条西 9 丁目
E-mail: †{marui,yang,atsu}@ist.hokudai.ac.jp

あらまし　大規模言語モデルの性能が向上することで、自動応答や自動要約などの様々なアプリケーションが開発されている。その一方で、大規模言語モデルがフェイクニュースや詐欺メールの生成に使われるなど、社会に悪影響を及ぼすことが懸念されている。そういった問題に対処するために、大規模言語モデルの生成文を検出する手法について様々な研究が行われている。しかしながら、検出手法に関する既存研究の多くは英語を対象として行われており、日本語における有効性は明らかにされていない。そこで、本研究では、日本語における大規模言語モデルの生成文を検出する手法について検討する。具体的には、まず、大規模言語モデルの生成文を収集することで、人間が書いた文と比較可能なデータセットを構築する。その上で、作成したデータセットを利用することで、既存の大規模言語モデルの生成文検出手法の日本語生成文検出精度に関し評価実験を行う。

キーワード　LLM，言語モデル，テキスト分類

1　はじめに

大規模言語モデルとは多数のパラメータを持つ言語モデルであり、大量のテキストデータを学習させることによって、自然言語処理に関する様々なタスクで成果を上げている。大規模言語モデルは、モデルパラメータの増大や大量のテキストデータを用いた学習に加えて、インストラクションデータを用いたファインチューニングや人間のフィードバックによる強化学習を行うことにより、より高い能力を持つようになっている。このような大規模言語モデルの性能向上によって、自動応答や自動要約などの様々なアプリケーションが開発されている。

このような性能向上に伴い、大規模言語モデルの生成文がインターネット上で拡散されるようになっている。大規模言語モデルの利用が社会に変革をもたらす一方で、大規模言語モデルが悪意を持って利用されることが懸念されている。

具体的な例として、大規模言語モデルをフェイクニュースや特定の考えに偏った文章の作成に用いることや、ソーシャルネットワークサービスのボットの運用インターネット上での情報共有に悪影響を及ぼすことが懸念されている。

さらに、論文を執筆する際に大規模言語モデルを不正に利用することが剽窃に繋がることや、レポート等の作成に大規模言語モデルが用いられることで学生の評価を適正に行えなくなることなどが指摘されており、学術活動や教育活動に対する悪影響も指摘されている。

これらの理由から、大規模言語モデルの悪用を防ぐための手立てが必要となっている。大規模言語モデルの利用による社会への悪影響を及ぼす問題の対策として、生成文を検出する手法が盛んに研究されている。大規模言語モデルの生成文検出は、任意の文章を対象として、与えられた文章が大規模言語モデルによって生成された文章であるかを判別する文書分類タスクとして考えることができる。検出を行うアプローチとしては、文章の言語的な特徴量や言語モデルのサンプリングにおける特性を利用した特徴量を用いた手法 [8] [1] [7] が提案されているほか、深層学習モデルを利用した検知手法が提案されている [8]。

生成文検出を行う深層学習モデルの学習や検出手法の性能評価を行うためには、大規模言語モデルの生成文を収集したデータセットの作成が必要不可欠である。しかしながら、多くの既存研究は英語を対象言語としており、日本語を対象とした生成文検出は十分に検討されていない。

したがって、本研究では日本語における大規模言語モデルの生成文の検出手法について検証することを目的としており、本研究の主な貢献は以下の通りである。

（1）　日本語における大規模言語モデルの生成文を収集することで、検証データセットを作成する。

（2）　検証データセットを用いて他言語生成文に対して検証された既存検出手法の日本語生成文に対する評価実験を行うことで、日本語の大規模言語モデルに対する検出性能を検証する。

（3）　複数の大規模言語モデルに対する性能評価を行うことによって、大規模言語モデルの特性について検出性能の観点から評価する。

2　関連研究

2.1　大規模言語モデルの生成文検出手法

大規模言語モデルの生成文検出のアプローチは、電子透かしを用いる手法と著者識別に基づく手法に大別される。

電子透かしを用いた手法は、文中に人間が知覚できない方法で検出に用いる情報を付与する手法であり、代表的な手法として、言語モデルが出力する際に電子透かしを付与する手法 [3] や生成された文章に対して電子透かしを付与する手法 [10] などが提案されている。これらの研究は、検出を行う文章に対して

予め電子透かしを生成文に付与する必要があるため、任意の生成文を対象として検出を行うことを目的とした本研究とは目的が異なる。

著者識別とは特定の著者が書いた文書を識別する問題であり、文書分類タスクとして盛んに研究が行われている。大規模言語モデルの生成文を対象とした研究として、文章の言語的な特徴量や言語モデルが文を生成する際の特性を利用した特徴量を用いた手法が提案されているほか、深層学習モデルを利用した検知手法 [8] などが提案されている。検出に用いるために提案されている特徴量の例としては、言語モデルの生成する文章は、生成確率の高い文章に偏っているという仮定に基づく手法が提案されており、文章の対数尤度 [8] や、文中の単語の生成確率の順位 [1] などが検出に用いる特徴量として提案されている。また、Mitchell [7] らは、大規模言語モデルが生成した文章は人間が書いた文章と比較して文章に変化を加えた際の対数尤度の変化量が大きいという仮定を基にした検出手法を提案している。

2.2 検証データセット

生成文検出を行う機械学習モデルの学習や検出手法の性能評価を行うためには、大規模言語モデルの生成文を収集したデータセットの作成が必要不可欠である。大規模言語モデルが広範な領域の知識を持つことから、多様な領域に対する生成文検出手法の検証が求められている。そのため、検証対象とする領域の特性に合わせて、データセットの作成手法について検討する必要がある。

このような理由から、大規模言語モデルの生成文を収集したデータセットの作成に関する研究が盛んに行われている。具体的な例として、ニュース [12]、随筆 [4]、論文 [11] など、様々な領域において研究がなされている。

しかしながら、既存研究の多くは英語のみを対象としており、他の言語は対象となっていない。英語以外の言語を対象としたデータセットとして、HC3 [2], M4 [9], MULTITuDE [6] などが作成されているが、いずれも日本語は対象言語に含まれていない。

したがって、本研究では、日本語における大規模言語モデルの生成文検出手法の検証を行うために、検証に用いるデータセットの作成手法について検討する。さらに、作成したデータセットを用いた検出手法の性能に関する評価実験を行うことで、日本語における大規模言語モデルの生成文検出手法の研究に貢献することを目的とする。

3 検証データセットの作成

本章では、データセットの作成手法と検知手法の構築手法について説明する。さらに、生成文検出に用いる深層学習モデルの構築手法について説明する。

3.1 データの作成手法

検証データセットの作成手法の概要を、図 1 に示す。本研究では、質問応答データセットを利用して人間が書いた文章と大規模言語モデルが生成した文章を収集する。具体的には、人間が書いた文章として、質問応答データセットの応答文を利用する。質問応答データセットの質問文を大規模言語モデルの入力とすることで大規模言語モデルの生成文を収集する。

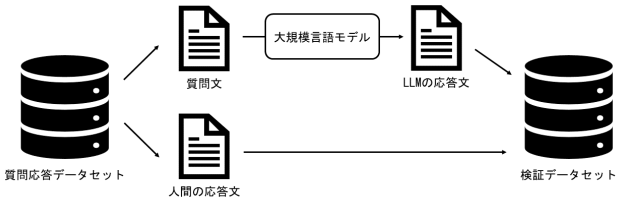


図 1　検証データセット作成手法の概要

本研究では、検証データセットの質問応答データセットの作成に国立情報学研究所が提供している「Yahoo! 知恵袋データ（第 3 版）[1]」を利用する。Yahoo!知恵袋データセットは、1 つの質問に対して、質問者の選択または他の利用者の投票によって質問に対する最良な回答として選ばれたベストアンサーと、0 個以上のその他の回答が存在する。本研究では、人間が書いた文章として、ベストアンサーを利用する。

データセットに対する前処理として、データセットから回答が 100 字未満のデータを除いた後に、7500 件の回答文とそれに対応する質問文を無作為に抽出することで用いるデータを選択する。

生成文の収集に用いる大規模言語モデルとして、以下の 3 つの大規模言語モデルを利用する。

- gpt-3.5-turbo[2]
- llm-jp-13b[3]
- Swallow-7b-instruct-hf[4]

gpt-3.5-turbo の生成文は OpenAI の API を利用して収集する。llm-jp-13b と swallow-7b の文章生成時に利用したパラメータを表 1 に示す。作成したデータセットの内訳を表 2 に示す。

表 1　文章生成時のパラメータ設定

パラメータ	値
min_new_tokens	32
max_new_tokens	512
top_p	0.95
temperature	0.7

1：https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html
2：https://platform.openai.com/docs/models/gpt-3-5
3：https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0
4：https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-hf

表 2　検証データセットの内訳

	全データ	訓練データ	テストデータ
人間	7500	6750	750
gpt-3.5-turbo	7500	6750	750
llm-jp-13b	7500	6750	750
swallow-7b	7500	6750	750
合計	30000	27000	3000

4　検証する生成文検出手法

4.1　深層学習モデル

深層学習を用いた手法では、訓練用データを用いて深層学習モデルを学習することで検出モデルを構築する。

本研究では、事前学習済みの RoBERTa [5] をファインチューニングすることで、検出を行う深層学習モデルを構築する。RoBERTa は、多量の文書データを用いて事前学習することによって、少量のデータで特定のタスクに適応した言語モデルを構築することが可能となっている。本研究では、早稲田大学河原研究室が公開している事前学習済み RoBERTa[5] を用いる。このモデルは、日本語の CC-100 と Wikipedia のデータを用いて事前学習されている。

深層学習モデルを用いた検出手法は、文を入力として、文を人間と大規模言語モデルの 2 クラスに分類するタスクとして定義できる。RoBERTa を用いた検出手法において、トークナイザーを用いてトークン化した文をモデルの入力とする。モデルの出力である [CLS] トークンに対応するベクトルを、全結合層、ドロップアウト層、全結合層の 3 層で構成される順伝播型ニューラルネットワークの分類器に入力する。そして、分類器の出力にシグモイド関数を適用することで、人間が書いた文と大規模言語モデルの生成文を判別する分類ラベルの予測を行う。

4.2　統計的特徴量を用いた手法

事前学習が不要な検出手法として、文から計算される統計的特徴量を用いる手法がいくつか提案されている [8] [1] [7]。

本研究では、Gehrmann ら [1] の提案した特徴量について検証を行う。Gehrmann らの提案した特徴量は、大規模言語モデルの生成文の分布が偏りを持つという仮定に基づく。具体的には、大規模言語モデルが文章を分布から生成する際に、尤度の高い部分のみからサンプリングを行うという仮定に基づいた手法である。各特徴量は、文をトークン化したトークン列を $w_1, w_2, \cdots, w_T$ とすると、以下の式で計算される。

（1）　対数尤度

$$\frac{1}{T}\sum_{i=1}^{T}\log p(w_i|w_{1:i-1}) \tag{1}$$

（2）　生成確率の順位

$$-\frac{1}{T}\sum_{i=1}^{T}\sum_{w\in V} I(p(w|w_{1:i-1}) < p(w_i|w_{1:i-1})) \tag{2}$$

（3）　生成確率の対数順位

$$-\frac{1}{T}\sum_{i=1}^{T}\log\left[\sum_{w\in V} I\left\{p(w|w_{1:i-1}) < p(w_i|w_{1:i-1})\right\}+1\right] \tag{3}$$

（4）　生成分布のエントロピー

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{w\in V} p(w|w_{1:i-1})\log p(w|w_{1:i-1}) \tag{4}$$

5　評 価 実 験

5.1　実 験 設 定

5.1.1　RoBERTa の学習設定

大規模言語モデルの生成文を正例、人間の文を負例として、13500 件の訓練データを用いてファインチューニングを行った。トークナイザーには sentencepiece を用いた。学習時のハイパーパラメータを表 3 に示す。ファインチューニングした RoBERTa の出力した予測ラベルのシグモイド関数値に対するしきい値を変化させて、ROC-AUC の値を計算した。

表 3　学習時のハイパーパラメータ

ハイパーパラメータ	値
損失関数	F1 スコア
学習率	5e-5
エポック数	3
バッチサイズ	8
ドロップアウト率	0.1

5.1.2　特徴量に基づく手法

文章から計算した特徴量に対するしきい値を変化させて、ROC-AUC の値を計算した。特徴量の計算に必要となる文中の各単語の生成確率は、japanese-gpt2-small[6] を用いて計算した。

5.2　評 価 指 標

検証データセットにおいて、生成文を正例、人間が書いた文章を負例とする。検出性能の評価指標として、ROC 曲線の AUC スコアを用いる。ROC 曲線は分類の閾値を変えた際の偽陽性率 (FPR) を x 軸、真陽性率 (TPR) を y 軸にプロットしたグラフであり、AUC スコアは ROC 曲線の下側の面積を指す。真陽性率と偽陽性率は、真陽性 (TP)、真陰性 (TN)、偽陽性 (FP)、偽陰性 (FN) を用いて以下の式で表される。

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{6}$$

5.3　実験 1: テストデータに対する検出性能の評価

検証データセットのテストデータに対する検出性能を評価することで、日本語の大規模言語モデルにおける生成文検出の性能について評価する。

5：https://huggingface.co/nlp-waseda/roberta-base-japanese

6：https://huggingface.co/rinna/japanese-gpt2-small

表 4　テストデータに対する ROC-AUC

検出手法	訓練データ	テストデータ		
		gpt-3.5-turbo	llm-jp	swallow
RoBERTa	人間+gpt-3.5-turbo	0.998	0.569	0.697
	人間+llm-jp	0.783	0.946	0.627
	人間+swallow	0.924	0.394	0.733
対数尤度	N/A	0.847	0.609	0.803
生成確率の順位	N/A	0.793	0.578	0.726
生成確率の対数順位	N/A	0.890	0.652	0.867
エントロピー	N/A	0.811	0.563	0.715

5.4　実験 2: 敵対的に生成したデータに対する検出性能の評価

既存の検証データセットの問題点として、大規模言語モデルの生成文が入力として与えるプロンプトに依存することから、文を生成する際に用いるプロンプトを工夫することによって、より検出が困難な文を生成できる可能性がある。本実験では、生成文を敵対的に生成した場合の検出性能について評価を行う。

本実験で使用するテストデータは、以下のように作成する。

（1）「Yahoo! 知恵袋データ（第 3 版）」から選択したデータセットを用いて、gpt-3.5-turbo をファインチューニングする。

（2）ファインチューニングした gpt-3.5-turbo を用いて、テストデータを作成する。

6　結果と考察

実験 1 の結果を表 4 に示す。表 4 から分かるように、いずれの大規模言語モデルにおいても、検出手法の最良値はランダムに分類した場合の ROC-AUC の値である 0.5 を上回る値となっていることから、検証対象とした日本語の大規模言語モデルの生成文と人間が書いた文の分布は完全に一致しておらず、生成文検出が可能であることが示された。

また、深層学習を用いた手法が、3 つ中 2 つの大規模言語モデルに対して最良の性能を示したことから、既存研究で提案されている特徴量とは異なる生成文を検出する上で有用な特徴量が存在することが考えられる。

表 5 と図 2 に敵対的な生成文に対する検出性能を示す。いずれの検出手法においても検出性能が低下する結果が得られた。この理由として、深層学習を用いた手法においては、検証データセットとは異なる分布から文が生成されていることから、敵対的に生成された文を人間が書いた文章として誤分類していることが考えられる。また、統計的特徴量を用いた手法においても、大規模言語モデルの生成文の分布が人間の文の分布に近づくことで統計的な差異が小さくなったことが、検出性能が低下した原因として考えられる。これらの結果から、文章生成に用いるプロンプトの変更などによって生じる生成分布の変化が、検出手法の信頼性の低下に繋がることが検出手法の問題点として明らかになった。

7　おわりに

本研究では、日本語における大規模言語モデルの生成文を検出手法の検証を目的として、データセットの作成と検出手法の性能を評価する実験を行った。具体的には、人間が書いた質問回答データセットを用いて人間と大規模言語モデルの文章を収集することで、検出手法の検証に用いるデータセットを作成した。そして、作成したデータセットを用いて日本語の大規模言語モデルに対する生成文検出手法の性能を評価する実験を行うことで検出手法の性能について検証した。

実験の結果、検出手法が日本語の大規模言語モデルの生成文検出に対しても有効である結果が得られた。その一方で、検出手法の性能は大規模言語モデルの生成分布に依存することも実験によって示唆された。このことから、検出手法の信頼性については、より多くの検討を必要とする。

今後の展望としては、検証の範囲を拡大することで、より包括的な大規模言語モデルの生成文検出に関する検証を行うことが考えられる。より広範な大規模言語モデルや検出手法を対象

表 5　ファインチューニングした LLM に対する ROC-AUC

	gpt-3.5-turbo-finetuned
RoBERTa	0.921
対数尤度	0.488
順位	0.450
対数順位	0.469
エントロピー	0.543

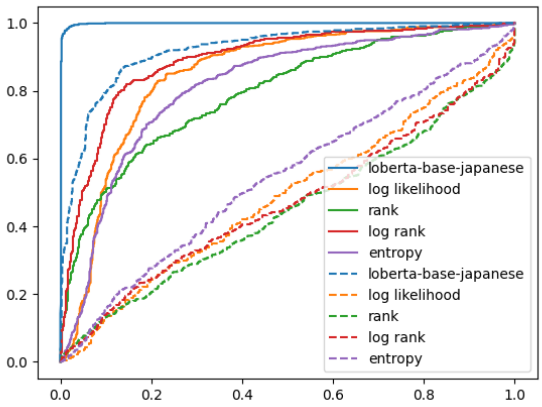


図 2　gpt-3.5-turbo のテストデータに対する ROC 曲線（実線）とファインチューニング後の gpt-3.5-turbo のテストデータに対する ROC 曲線（破線）

として、検出性能の検証を進めることは、大規模言語モデルの生成文検出手法に関する研究を促進する上で必要不可欠である。

文　献

[1] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *ArXiv*, abs/2301.07597, 2023.

[3] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, 2023.

[4] Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *ArXiv*, abs/2304.07666, 2023.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[6] Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. Multitude: Large-scale multilingual machine-generated text detection benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 9960–9987, Singapore, February 2023. Association for Computational Linguistics.

[7] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[8] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models. *ArXiv*, abs/1908.09203, 2019.

[9] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *ArXiv*, abs/2305.14902, 2023.

[10] Kiyoon Yoo, Wonhyuk Ahn, Jiho Jang, and No Jun Kwak. Robust multi-bit natural language watermarking through invariant features. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[11] Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *ArXiv*, abs/2304.12008, 2023.

[12] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *ArXiv*, abs/1905.12616, 2019.

ポジティブな単語を含んだ煽り・誹謗中傷目的のコメント検出方法の提案

佐藤　豪† 清　雄一† 田原　康之† 大須賀昭彦†

† 電気通信大学　〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: † sato.go@ohsuga.lab.uec.ac.jp, {seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　SNS の普及に伴い煽りや誹謗中傷を目的としたコメントによる被害が増加しているため，このようなコメントを自動で検出できる技術が求められている．その中でも特に検出が難しいのが直接的な言葉を用いておらず，皮肉表現のように間接的な言い回しを用いたコメントである．本研究では間接的な言い回しとしてポジティブな単語が含まれているコメントに着目し，既存の BERT モデルをファインチューニングするだけでなくパターンマッチングといった幾つかの手法を加えることで既存モデルよりも高い精度で煽りや誹謗中傷を目的としたコメントの検出が可能なモデルの構築を目指した．結果的に Accuracy，Precision，Recall，F1 の 4 つの評価指標において既存モデルを上回ることができた．

キーワード　テキスト分類，テキストマイニング，SNS，ソーシャルメディア

1 はじめに

1.1 背　景

近年，SNS などの普及による弊害として匿名サイト上などでの煽り・誹謗中傷による被害が後を絶たず，大きな社会問題となっている. そのため，この問題を解消するためには不適切なコメントがユーザの目に入らないようにする必要がある.

そこで，SNS や誰でも簡単にコメントを投稿できるようなプラットフォームを運営する企業に求められることは煽り・誹謗中傷を目的としたコメントが投稿されたら早急に検出し，そのコメントを削除することである. 実際，既に LINE ヤフー株式会社といったいくつかの企業では不適切なコメントを自動で検出，および削除する AI を導入している. しかし，このような対策を行っても被害が中々減っていないのが現状である [1]. その要因の 1 つとして，直接的な言葉を用いた煽り・誹謗中傷目的のコメントは容易に検出できるが間接的な言い回しを用いたコメントの検出が困難であることが考えられる. 例として東北大 BERT モデル [2] を利用し，間接的な言い回しを用いたいくつかの煽り・誹謗中傷目的のコメントに対して感情分析を行った結果，ネガティブではなくポジティブと分析されることが多かった.

現在，煽り・誹謗中傷を目的としたコメントの検出に関する研究は多く行われているが，間接的な言い回しに着目した研究は少ない. よって，関接的な言い回しを用いた煽り・誹謗中傷目的のコメントを高い精度で検出できるような手法が求められている.

1.2 目　的

このような背景を踏まえて，本研究では間接的な言い回しの 1 つであるポジティブな単語を含んだ煽り・誹謗中傷目的のコメントに着目し，このようなコメントを高い精度で検出することが可能なモデルを提案する．近年の自然言語処理タスクにおいては 2018 年に Google 社が発表した「Transformer による双方向のエンコード表現」と訳される BERT [3] が高い性能を残しているため，本研究では事前学習済み日本語 BERT モデルである東北大 BERT モデルを利用する．具体的には，東北大 BERT モデルに対してファインチューニングを行うことでより正確に感情分析を行うことを可能とし，さらにパターンマッチングなどと組み合わせることによってコメントのみを与えた時に正しく 2 値分類することが可能なモデルの構築を目指す.

背景でも述べたが，ポジティブな単語を含んだ煽り・誹謗中傷目的のコメントは誤ってポジティブと感情分析されてしまうことが多いため，本研究では主にこのようなコメントに着目する．東北大 BERT モデルを用いて感情分析を行った結果，実際はネガティブなコメントであるが誤ってポジティブと分析されたコメントの例を表 1 に示す.

表 1　東北大 BERT モデルによって誤ってポジティブと感情分析されたコメントの例

コメント	感情分析の結果	実際の感情
頑張って低評価 10 万目指してください!	ポジティブ	ネガティブ
低評価 5 万突破おめでたい	ポジティブ	ネガティブ
エアコンの上下自動みてたほうが面白い.	ポジティブ	ネガティブ

2 関連研究

2.1 東北大 BERT モデル

東北大 BERT モデルとは，東北大学自然言語処理研究グループが公開した事前学習済み日本語 BERT モデルである．訓練には日本語版 Wikipedia がコーパスとして用いられた．トークナイザでは，まず MeCab(ipadic) [4] を用いて単語に分割し，次に WordPiece [5] を用いて単語をトークンに分割することでトークン化が行われている．このモデルは，HuggingFace [6] による自然言語処理ライブラリである Transformers でも訓練済みモデルとして利用することができる.

本研究では，事前学習済みである東北大 BERT モデルをファインチューニングすることで文章のみを与えたときに正しく感情分析や検出対象であるか否かの 2 値分類を行うことが可能なモデルを構築する．

2.2 パターンマッチングを用いた皮肉表現の分類および検出

間接的な言い回しの代表例として頻繁に用いられるのが皮肉表現である．磯野らは，楽天トラベルのレビューデータと Web 掲示板から約 14600 件の日本語によるテキストデータを収集し，人手によって皮肉表現を含んでいるかを判断してラベル付けを行いコーパスを構築した [7]．そして，コーパスを分析することで構文パターンを設計し，パターンマッチングによって皮肉表現を疑問，推測，諦め，不相応，誇張，驚き，形容，対比の 8 つの種類に分類するモデルを提案した．皮肉表現であるかどうかを判断するには前後の文脈が重要であるという理由から，この研究におけるパターンマッチングでは構文パターンに一致する文の前後の文にも着目し，日本語評価極性辞書 [8] に掲載されている否定的な語句が前後の文に含まれているかどうかも考慮された．実際に提案された皮肉検出のための構文パターンの中で，ポジティブな単語が含まれている誇張に分類された構文パターンを抜粋して表 2 に示す．

表 2　提案された構文パターンの一部

分類	直前	対象文の形式	直後
誇張	Neg	さすが ～	-
	Neg	すごい ～ だな	-
	-	すごい ～ w	-

2.3 肯定的な単語を含んだ皮肉表現の検出

肥合らは，皮肉表現の中でも肯定的な単語を含んだ表現に着目し 2 つの段階を経て該当する文章の検出を行った [9]．ただし，願望や後悔，否定などの形で肯定的な単語が使われている場合は例外とされた．実際に例外とされた表現の一覧は表 3 の通りである．検出を行うにあたり，まずは楽天市場レビューデータから検出対象となるレビューの本文を 70 件取得した．そして，取得した各文章において肯定表現がどのように否定的意味を表しているのかを分析し 8 つのクラスに分類を行った．クラスの例としては，対象への否定的評価をそのまま肯定的に言い換えている文章のクラスや，他の類似した対象を肯定的に評価することで対比して否定的評価を示している文章のクラスなどがあった．8 つのクラスへの分類が完了した後，1 つ目の段階として文章全体の極性や構文パターンを利用し，検出対象の候補となる文章の絞り込みが行われた．次に，2 つ目の段階では候補として残った各文章の中で表 4 に掲載された婉曲・強調表現や記号のうちいずれかをを含んでいる文章のみが最終的な検出対象として抽出された．

表 3　例外とされた表現一覧

ならいいのに，ならよかった，ばいいのに，ばよかった，でもいいのに，でもよかった，方がいいのに，方がよかった，まあいい，よしとする

表 4　検出対象とする文章の絞り込みを行うために用いられた婉曲・強調表現一覧

種類	表現
丁寧語や尊敬語	お (接頭辞)，ご (接頭辞)，ご覧，ご存知，献上，参上，拝見，拝読，存じ，頂戴，になる，れる，なさる，頂く，おる，おっしゃる，召し，申す，いたす，さしあげ，うかがう，参る，くださる
特定の強調の副詞	本当，確かに，確実に，すごく，さすが，とても，結構，かなり，いつも，もう，よく，思わず，めきめき，やっぱり，やはり，すぐ，よほど，改めて，ちょうど，なんだか，まったく，もっと
特定の強調の接頭辞	大，ど，超，全，真
特定の記号	!，?，･･･，(笑，w，「,」，[，]，「,」，..,..，～

2.4 品詞によるフレーズを利用した皮肉表現の検出

鈴木は，名詞や形容詞といった品詞の組み合わせによるフレーズを作成しこれを利用した皮肉表現の検出方法を提案した [10]．提案フレーズは行動を表す situation phrase と感情を表す sentiment phrase から形成され，それぞれがポジティブなフレーズとネガティブなフレーズの 2 種類に分けられた．そして，ポジティブなフレーズとネガティブなフレーズのうち数が多い方がそのフレーズの感情となり，situation phrase と sentiment phrase の感情が異なっている場合皮肉表現を含む文として判定された．なお，ポジティブなフレーズとネガティブなフレーズの数が同じであった場合は tf-idf 法によってフレーズの感情が判定された．

2.5 BERT を用いた皮肉表現の検出

皮肉表現は日本語のみでなく他の様々な言語でも存在する為，各言語毎に皮肉表現の検出に関する研究が行われている．その例として，Yazhou らは BERT を用いて中国語と英語による皮肉表現の検出を行った [11]．この研究では，データセットとして用意された各コメントに対してコメントを書いた人の立場が賛成・反対・中立の中からアノテーターによって手動で選ばれ，その立場を表すノードが中心となってコメントに含まれる各単語の単語埋め込みベクトルをノードとして結ばれたグラフアテンションネットワークが形成された．そして，BERT によって文の分散表現を，グラフアテンションネットワークによって立場表現を学習しそれらを連結してソフトマックス層に渡すことで皮肉表現の検出を行うという手法が提案された．結果的には，中国語と英語の両方のデータセットに対して Accuracy，Precison，F1 という 3 つの評価指標において SVM や CNN といったベースライン手法よりも高い精度で皮肉表現の検出を行うことが可能となっていた．

3 本研究において検出対象となるコメントの定義

本研究において検出対象となるコメントはタイトルにもあるように「ポジティブな単語を含んだ煽り・誹謗中傷目的のコメント」であるが，そこには「一見ポジティブなコメントとも感じ取れるが実際はネガティブなコメントである」という条件，すなわち皮肉表現が含まれているという条件を含んでいる．そ

のため，ポジティブな単語が否定の形で使われている場合は例外的に検出対象とはしないこととする．例えば，「面白い訳ではない」や「凄いとは言えない」という文章には部分的にポジティブな単語が含まれているが否定の形で使われているため例外とする．また，文末で使われることの多い「笑」や「w」といった単語はポジティブな単語であるとも考えられるが，こういった単語は皮肉表現の一部として広く利用されているため，本研究ではポジティブな単語とみなさないこととする．

4 提案手法

本研究では事前学習済みモデルである東北大 BERT モデルに対してファインチューニングを行い，さらにパターンマッチングを導入することで既存の BERT モデルよりもポジティブな単語を含んだ煽り・誹謗中傷を目的としたコメントの検出精度が高いモデルを提案する．具体的な提案モデルの概要は図 1 の通りである．なお，最終予測ラベルとは，0 の場合検出対象ではないと予測し，1 の場合検出対象であると予測するということを意味している．

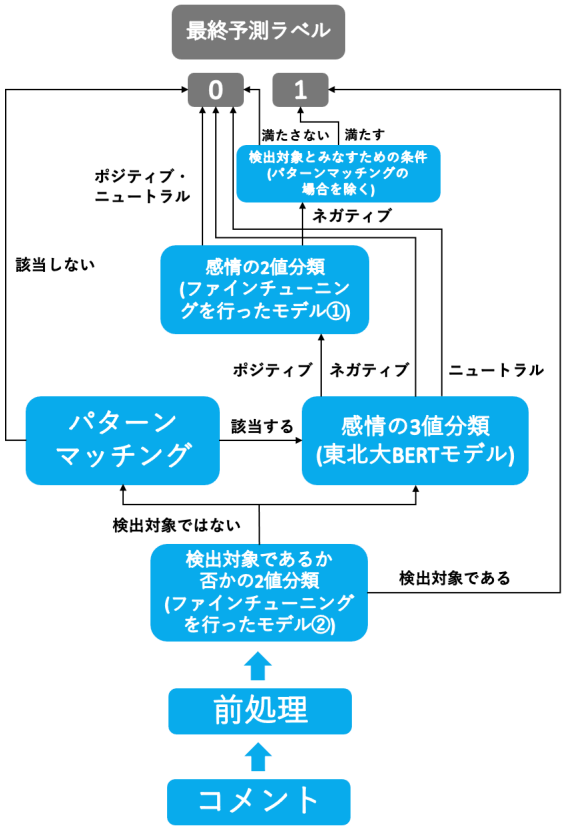


図 1　提案モデルの概要

4.1 ファインチューニング

4.1.1 感情の 2 値分類

本研究では皮肉表現の特徴を考慮して，実際はネガティブであるのにも関わらず誤ってポジティブと感情分析される表 1 に掲載したようなコメントに着目する．しかし，本研究の目的はコメントのみを与えたときにポジティブな単語を含んでいて尚且つ煽りや誹謗中傷を目的としているかどうかを正しく分類するモデルを提案することなので，実際の感情 (ポジティブ・ニュートラル・ネガティブのいずれか) を毎回手動でコメントに付与することは意図していないため，自動で実際の感情を予測してコメントに付与できるようにする必要がある．

そこで，コメントを与えた時に単純にポジティブな単語を含んでいて尚且つ煽りや誹謗中傷を目的としているかどうかの 2 値分類を行うのではなく，まずは実際の感情を高い精度で分析できるように東北大 BERT モデルに対してファインチューニングを行う．その際，今回の研究で用いたデータセットにおいて実際の感情がポジティブであるコメントがニュートラルまたはネガティブであるコメントに対して極端に少なかったため，3 値分類ではなくポジティブまたはニュートラル・ネガティブの 2 値分類を行うことにする．そのため，データセットに含まれる全コメントに対して実際の感情がポジティブまたはニュートラルであればラベル 0 を，ネガティブであればラベル 1 を手動で付与する．

4.1.2 検出対象とみなすための条件

検出対象とみなすための条件として一番簡単なのが，東北大 BERT モデルをファインチューニングすることで本研究において検出対象となるコメントであるか否かの 2 値分類を行った際に，予測ラベルが 1 になる事である．こうなった場合はその時点で最終予測ラベルを 1 として検出対象とみなすこととする．これ以降は 2 値分類の予測ラベルが 0 になった場合に検出対象とみなすための条件について述べる．

既に述べているように，この研究では実際はネガティブであるのにも関わらず既存のモデルでは誤ってポジティブと感情分析されてしまうコメントに着目する．つまり，皮肉表現を含んだコメントに対して既存のモデルよりも正確に感情分析を行うことが可能となったモデルを用いて「東北大 BERT モデルではポジティブと感情分析されるが，ファインチューニングを行ったモデルではネガティブと感情分析されるコメント」に着目し，これに該当するコメントを本研究において検出対象となるコメントの候補とする．

上記の条件に当てはまるコメントは本研究において検出対象となるコメントである可能性が高いと考えられるが，当てはまるコメントを全て検出対象とみなしてしまうと，逆に検出精度が下がってしまう．そこで，さらにポジティブな単語を含んだ煽り・誹謗中傷を目的としているコメントの候補を絞るために 2 つの条件を追加する．

1 つ目の条件は，検出対象となるコメントであるか否かの 2 値分類を行った結果，検出対象ではない確率が基準値 p よりも小さい値となる事である．この条件を加えた理由は，検出対象ではない確率が低ければそれだけ予測ラベルが誤りである可能性が高くなると考えたためである．2 値分類では各コメントにおけるラベル 0 とラベル 1 それぞれの分類スコアが出力され値の大きい方が予測ラベルとなるため，2 つの分類スコアを softmax 関数に入力することで分類スコアを確率値に変更でき

る．なお，softmax 関数は次の式で表される．n は分類する種類の数，x_i は元の各分類スコアの値を表している．ここでは 2 値分類を行うため，式 1 において i の値は $i = 1, 2$，n の値は $n = 2$ となる．

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \quad (1)$$

次に，2 つ目の条件として感情スコアを利用する．東北大 BERT モデルでは文章を入力するとポジティブ・ニュートラル・ネガティブのどれに該当するかだけでなく，どれだけ与えられた文章がその感情に偏っているかという尺度を表す感情スコアを 0 ～ 1 の範囲で算出することができる．本研究で検出対象となるコメントはポジティブな単語を含んでいるという前提があるが，皮肉表現が使われている場合ポジティブな単語はより強調されると考えられるため，感情スコアの値が基準値 s よりも大きいときのみ検出対象にするという条件を追加する．

これらの 2 つの条件をともに満たすコメントを検出対象とみなす，つまり予測ラベルを 0 から 1 に変更することで東北大 BERT モデルをただファインチューニングして 2 値分類を行った時よりも精度を向上させる．

4.2 パターンマッチング

皮肉表現の検出を行うにあたって頻繁に用いられるのがパターンマッチングと呼ばれる手法である．文章の検出や分類におけるパターンマッチングとは表 2 のように検出を行いたい文章の特徴としてどういったパターンがあるかを考えた上で構文パターンを作成し，その構文パターンに当てはまった文章を検出対象とするという手法である．検出対象であるか否かの 2 値分類において検出対象ではないと予測されたコメントに対してパターンマッチングを適用することで，誤って検出対象ではないと予測されてしまうコメントの数を減らすことを目指す．

パターンマッチングの問題点として，誤検出が多くなるということが挙げられる．パターンマッチングを利用する場合，構文パターンに該当する文章は全て検出や分類の対象となるので意図していない文章に対しても検出や分類が行われ，結果的に精度が悪くなることも少なくない．しかしながら，その対策として構文パターンを減らしすぎると誤検出の数は減少するがそれに伴って検出や分類を行う事ができる文章の数も減ってしまい，結果としてパターンマッチングの有効性が失われてしまうため，パターンマッチングを利用するには構文パターンの数や条件を適切に設定する必要がある．

そこで，本研究では特に誤検出が多くなってしまうという問題点を解決するために通常のパターンマッチングにいくつかの条件を追加した手法を提案する．

4.2.1 キーワード

構文パターンを作成するにあたって必須となるのがキーワードの選択である．パターンマッチングでは正確に文章の検出や分類を行うために役立つキーワードを選び，そのキーワードが幾つか含まれているような構文パターンを作成する必要がある．本研究では，文中と文末で分けてそれぞれにキーワードを設けた．皮肉表現を含んだコメントは文中の単語だけでなく文末の単語にも特徴があると考えたからである．本研究で設けたキーワードの例を抜粋して表 5 に掲載する．

表 5　構文パターン作成のために選択したキーワードの例

文中	文末
• 逆に • 流石 • むしろ • よく • 方が	• w • 笑 • ？ • わろた • ＾＾

なお，それぞれのキーワードに関して漢字・ひらがな・カタカナなどと色々な書かれ方がある場合はそれらもキーワードに含めている．例えば「逆に」であれば「ぎゃくに」もキーワードとして含んでおり,「わろた」であれば「ワロタ」もキーワードとなっている．また，キーワードは単語だけでなく顔文字や絵文字，記号なども対象となるようにした．

そして，文中及び文末の両方にいずれかの選択したキーワードが含まれているコメントを構文パターンに当てはまるコメントとみなすことにする．

4.2.2 感情分析の結果を利用

しかし，この条件だけでは誤検出の多さというパターンマッチングの問題点を解決できていない．そこで，感情分析の結果を利用する．

まず，先ほど述べた条件 (構文パターン) に合致するコメントに関して文中のキーワード以降の部分を取り出す．そして，取り出した部分に対して東北大 BERT モデルを利用することで感情分析を行いその結果がポジティブであったコメントのみを検出の対象とする．

さらに誤検出を減らすためにこれらの条件を満たしたコメントに対して，4.1.1 節でより正確に感情の 2 値分類をする事が可能になるようにファインチューニングを行ったモデルを利用して感情分析を行い，その結果がネガティブとなったコメントを最終的な検出の対象とみなす．つまり，予測ラベルを 0 から 1 に変更する．

5 実　　験

5.1 データセット

実験で用いたコメントは日本の全 YouTube 動画における低評価数ランキングを記載したサイト [12] から 1 つの YouTube 動画を選択し，その動画のコメント欄から取得した．なお，取得したのは動画に対するコメントのみであり，他のユーザに対して返信を行っているコメントは取得していない．また，全コメントを無作為に取得すると検出対象となるコメントが検出対象とならないコメントに対してかなり少なくなってしまうので，一定の割合が保てるようにコメントの比率を調整した．

このようにして取得したコメントの総数は以下の表の通りである．

表 6　実験で用いたコメントの数

	検出対象である	検出対象でない
コメント数	245	4591

実験の際は，このデータセットをランダムにシャッフルしてからトレーニングデータ 60%，検証データ 20%，テストデータ 20% の割合になるように分割を行った．

5.2 前　処　理

データセットの収集に続いて，より正確な 2 値分類が可能になるように収集した各コメントに対して前処理を行った．実際に行った前処理を以下に示す．

- 半角記号の除去
- URL の除去
- 空白の除去
- 数字を 0 に統一
- 英字を小文字に統一

文章の前処理では顔文字や絵文字も除去される事が多いが，パターンマッチングで必要となるため本研究ではそれらは除去せずに残している．

5.3 結果と評価

5.3.1 感情の 2 値分類

4.1.1 節で述べた方法で東北大 BERT モデルに対してファインチューニングを行った．その際，学習率は 1×10^{-6} としバッチサイズは 64 に設定した．また，epoch 数は指定せず代わりに Early stopping を導入した．Early stopping における各引数の設定は表 7 の通りである．

表 7　Early stopping における各引数の設定

引数	パラメータ
monitor	`val_loss`
mode	min
patience	20

表 7 のように，validation loss の値を基準として 20 epoch の間 validation loss の値が改善しなければ学習が止まるように設定した．また，mode によって下限に収束した場合に収束判定を出すように指定した．

以上の設定のもと，データセットに含まれる全コメントを用いて東北大 BERT モデルとファインチューニングを行ったモデルの 2 つで感情分析の精度を比較した．比較するための評価指標としては Accuracy，Precision，Recall，F1 の 4 つの指標を用いた．精度比較の結果は次の表のようになった．

表 8　感情分析の精度比較

モデル	Accuracy	Precision	Recall	F1
東北大 BERT モデル	0.689	0.801	0.682	0.737
ファインチューニングを行ったモデル	0.761	0.843	0.793	0.817

表 8 を見ると，全ての評価指標においてファインチューニングを行ったモデルが東北大 BERT モデルよりも優れていることが分かる．ここで，ファインチューニングを行ったモデルがポジティブな単語を含んでいて尚且つ煽りや誹謗中傷を目的としているコメント検出のために役立つかどうかを確かめるために，実際の感情はネガティブであるが東北大 BERT モデルでは誤ってポジティブと感情分析された表 1 に掲載したコメントに対してファインチューニングを行ったモデルで感情分析を行った．その結果は表 9 のようになった．

表 9　ファインチューニングを行ったモデルでの感情分析結果

コメント	感情分析の結果	実際の感情
頑張って低評価 10 万目指してください!	ネガティブ	ネガティブ
低評価 5 万突破おめでたい	ネガティブ	ネガティブ
エアコンの上下自動みてたほうが面白い.	ネガティブ	ネガティブ

この結果から，東北大 BERT モデルでは誤ってポジティブと感情分析されてしまっていた皮肉表現を含むコメントに関して，ファインチューニングを行ったモデルではより正確にネガティブと感情分析できていたことから，このモデルはポジティブな単語を含んだ煽りや誹謗中傷を目的としているコメントの検出に活用できることが確認できた．

5.3.2 検出対象とみなすための条件

次に，東北大 BERT モデルをファインチューニングすることでデータセットに含まれる全コメントに対してポジティブな単語を含んでいて尚且つ煽りや誹謗中傷を目的としているかどうかの 2 値分類を行い，4.1.2 節で述べた条件は加えずに東北大 BERT モデルではポジティブ，5.3.1 節でファインチューニングを行ったモデルではネガティブと感情分析されなおかつ検出対象か否かの 2 値分類において予測されたラベルと実際のラベルが異なっていた全てのコメントに関してラベル 0 である (検出対象ではない) 確率，予測されたラベル，実際のラベルを並べると次の表のようになった．

表 10　本研究で着目するコメントにおいて 2 値分類が正しく行われなかったコメントの結果一覧

検出対象ではない確率	予測されたラベル	実際のラベル
0.940	0	1
0.962	0	1
0.979	0	1
0.975	0	1
0.941	0	1
0.988	0	1
0.880	0	1
0.699	0	1
0.857	0	1
0.810	0	1
0.989	0	1
0.254	1	0

この表から，予測されたラベルと実際のラベルが異なっている時予測されたラベルが 0 で実際のラベルが 1 である場合がほとんどだということが分かった．

続いて，4.1.2 節で定義した基準値 p の値を $p = 0.90$ とし，2 値分類の結果予測されたラベルが 0 でありなおかつ検出対象ではない確率が 0.90 未満であるという条件を満たす全てのコメントに関して検出対象ではない確率，予測されたラベル，実際のラベル，東北大 BERT モデルを利用して算出した感情スコアの値を並べると次の表のようになった．

表 11　本研究で着目するコメントにおいて予測されたラベルが 0 かつ検出対象ではない確率が 0.90 未満であったコメントの結果一覧

検出対象ではない確率	予測されたラベル	実際のラベル	感情スコア
0.880	0	1	0.958
0.699	0	1	0.903
0.857	0	1	0.464
0.810	0	1	0.981
0.750	0	0	0.781
0.829	0	0	0.596

表 11 より，本研究で着目するコメントにおいて 2 値分類で予測されたラベルが 0 でありなおかつ検出対象ではない確率が 0.90 未満であった全てのコメントの予測ラベルを 1 に変えてしまうと，Recall などの値は上がるが誤検出が多くなり Accuracy の値もほとんど変化しないため得策ではないことが分かった．

そして，続いて 4.1.2 節で定義した基準値 s の値を $s = 0.90$ とし，2 つ目の条件として説明した東北大 BERT モデルを利用して算出した感情スコアの値が 0.90 よりも大きいときのみ検出対象にするという条件を加えてこれらの条件を満たすコメントのみをこれまでと同様に表にすると次のようになった．

表 12　感情スコアの値が 0.90 より大きいという条件を加えた時の結果一覧

検出対象ではない確率	予測されたラベル	実際のラベル	感情スコア
0.880	0	1	0.958
0.699	0	1	0.903
0.810	0	1	0.981

よって，表 12 より提案手法で述べた条件を全て満たすコメントは 2 値分類の結果が予測ラベルは 0，実際のラベルは 1 で全て一致しているため，これらのコメントに対して予測ラベルを 1 に変えることでただ東北大 BERT モデルをファインチューニングして 2 値分類を行った時よりも高い精度を残すことができた．

5.3.3　パターンマッチング

続いて，検出対象であるか否かの 2 値分類において検出対象ではないと予測されたコメントの中で，4.2 節で定義したパターンマッチングに該当するコメントを抽出し，そのコメントの予測ラベルと実際のラベルを並べると次の表のようになった．

表 13　パターンマッチングに該当するコメントの 2 値分類における予測ラベルと実際のラベル

予測されたラベル	実際のラベル
0	1
0	1
0	1
0	0

よって，提案手法で定義したパターンマッチングに該当するコメントの予測ラベルを 1 にすると誤検出も少し増加してしまうが，パターンマッチングを導入することで精度を向上させることができた．

5.3.4　最終的な 2 値分類

ここまで，感情の 2 値分類がより正確に行えるようにファインチューニングを行い，予測ラベルが 0 である確率や感情スコアの値に着目することで検出対象とするコメントを増やす手法や，誤検出が従来よりも少なくなるようなパターンマッチングを導入した時の分類精度の変化を見てきたが，この 2 つの手法を合わせたときの最終的な分類精度の比較を行う．比較対象となるのは既存モデルである東北大 BERT モデルを単にファインチューニングして 2 値分類を行ったモデルである．

評価指標としては，5.3.1 節において東北大 BERT モデルとファインチューニングを行ったモデルで感情分析の精度比較をしたときの指標と同様に，Accuracy，Precision，Recall，F1 の 4 つの指標を使った．

最終的な分類精度の比較を行った結果を表 14 に示す．

表 14　既存モデルと提案モデルの分類精度比較

モデル	Accuracy	Precision	Recall	F1
東北大 BERT モデル (既存モデル)	0.972	0.816	0.635	0.714
提案モデル	0.976	0.818	0.714	0.763

この結果から，本研究で提案したモデルが 4 つの全ての指標で既存のモデルを上回った．つまり，提案モデルは既存モデルよりも高い精度でポジティブな単語を含んでいて尚且つ煽りや誹謗中傷を目的としたコメントであるかどうかに関する 2 値分類を行うことが可能であると判明した．

5.3.5　アブレーションスタディ

表 14 より，提案モデルの方が既存モデルよりも 2 値分類の精度が高い事が示されたが，4.1.1 節，4.1.2 節で述べた主に 2 つの感情分析の結果を利用した手法と 4.2 節で述べたパターンマッチングを利用した手法のそれぞれがどれほど性能の向上に貢献しているのかを調べるために，2 つの手法のうち片方の手法のみ用いたときの分類精度を表 14 に加えると次のようになった．なお，分かりやすくなるよう各手法に下記の通り番号を振った．

2 つの感情分析の結果を利用した手法 … ①

パターンマッチングを利用した手法 … ②

表 15　アブレーションスタディの結果

モデル	Accuracy	Precision	Recall	F1
東北大 BERT モデル (既存モデル)	0.972	0.816	0.635	0.714
提案モデル (①)	0.975	0.827	0.683	0.748
提案モデル (②)	0.974	0.811	0.683	0.741
提案モデル (①+②)	0.976	0.818	0.714	0.763

アブレーションスタディの結果より，全ての指標において 2 つの感情分析の結果を利用した手法がパターンマッチングを利用した手法に比べて同等かそれ以上の精度を記録した為，2 つの感情分析の結果を利用した手法がより提案モデルの精度向上に貢献していることが分かった．また，特に Precision の値については 2 つの手法による差が大きく，パターンマッチングを利用した手法のみでは既存モデルよりも Precision の値が低くなってしまうことが分かった．

6 考　　察

6.1 データセットと前処理

まず，データセットについては 1 本の YouTube 動画から取得したため検出対象となるコメントの数が少なく，検出対象となるコメントの中で文章が類似しているものもいくつかあったので，ファインチューニングすることで精度は上がりやすくなっていたのではないかと考えられる．表 6 を見ると分かるようにデータセットの比率が大きく偏っているのにも関わらず高い分類性能を記録する事ができたので，もう少し比率を調整することで全体的な分類性能の向上や既存モデルと提案モデルによる性能の乖離が期待できる．

前処理ではパターンマッチングとの兼ね合いで顔文字や絵文字，助詞などの除去は行わなかったが，これはパターンマッチングを用いることによって生じる特有の問題点であるとも言えるため，パターンマッチングに該当しなかった場合はこれらの文字の除去を行ってから BERT に入力するなどといった工夫が必要だったと考えられる．

6.2 既存手法に比べて精度が向上した要因

表 8 の精度比較の結果から東北大 BERT モデルをファインチューニングすることでより正確に感情分析を行うことが可能になるということが分かった．表 8 に掲載した 4 つの評価指標の値を見ると，特に Recall の値に差がでていることが分かる．このことから，東北大 BERT モデルに比べてファインチューニングしたモデルの方が実際はネガティブであるコメントを誤ってポジティブやニュートラルと分析する事が少ないと言えるため，Recall の値の差が開いたことで本研究で重要となる表 9 に掲載したようなコメントを多く抜粋する事が可能となり，全体の精度向上に繋がったと考えられる．

表 10 を見ると，やはり実際はネガティブであるが誤ってポジティブと感情分析されてしまうコメントを正確に検出対象だと予測することは困難だという事がわかる．一方で，そのようなコメントに対して検出対象であると予測した場合はほとんど誤りがない事もわかる．このことから，誤検出を減らすのではなく間違って検出対象ではないとみなされてしまう事が減るような手法を提案したのは良い選択だったと言えると考えられる．

また，皮肉表現にポジティブな単語が含まれている場合，ポジティブな単語を強調することでより皮肉の意味合いが強められる事が多いので感情スコアを利用できるのではないかと考えたが，表 12 から感情スコアと検出対象ではない確率や予測ラベル，実際のラベルの間には一定の相関があると分かったためこの仮説は正しかったと考えられる．

6.3 基準値 p，s と検出精度の関係

東北大 BERT モデルではポジティブ，5.3.1 節でファインチューニングを行ったモデルではネガティブと感情分析され，なおかつ検出対象か否かの 2 値分類において予測されたラベルが 0 であった全てのコメントを実際のラベルで 2 つに分類し，それぞれにおいて検出対象ではない確率と感情スコアの平均値を取ると表 16 のようになった．本研究では 4.1.2 節で定義した基準値 p，s の値を共に 0.90 としたが，表 16 より p の値はこれ以上大きくすると誤検出の可能性が高まり，逆にこれ以上小さくすると Recall の値が下がってしまうため適切な値であったと考えられる．一方で，感情スコアの平均値の差を考えると s の値についてはもう少し小さくすることで誤検出のリスクを抑えたまま Recall の値を向上させることができたのではないかと考えられる．

表 16　検出対象ではない確率と感情スコアの平均値

実際のラベル	検出対象ではない確率の平均値	感情スコアの平均値
0	0.977	0.737
1	0.911	0.867

6.4 パターンマッチングによる誤検出

パターンマッチングでは，構文パターンに加えてファインチューニングした感情の 2 値分類を行うモデルを利用することで表 13 のように検出精度の向上を実現する事ができたが，感情の 2 値分類を行うモデルを利用しなかった場合の結果は表 17 のように正しく検出できるようになった数よりも誤検出の数の方が多くなってしまったため，パターンマッチングを用いる際は構文パターンだけでなく誤検出が減るように感情分析の結果などの条件を追加する事が重要だと分かった．ただし，皮肉表現を扱う場合は既存の感情分析を行うモデルでは誤った結果となることが多いため，注意が必要になると考えられる．

表 17　感情の 2 値分類を行うモデルを利用しなかったときのパターンマッチングの結果

予測されたラベル	実際のラベル
0	1
0	1
0	1
0	0
0	0
0	0
0	0

6.5 今後の改善点

最終的に提案モデルは東北大 BERT モデルを単にファインチューニングした時よりも高い精度でポジティブな単語を含んでいて煽りや誹謗中傷を目的としたコメントであるかどうかに関する 2 値分類を行う事ができたが，この 2 つのモデルにおいても感情分析の精度比較を行った時と同様に Recall の値に差が出ている事がわかる．これは，2 つの感情分析の結果を利用した手法やパターンマッチングを用いた手法によって既存モデルでは間違って検出対象ではないとみなされてしまうコメントを一定数正しく検出対象であると予測できるようになったからであり，誤検出の割合を既存モデルと同程度に保ったまま検出対象となるコメントの検出精度を向上させたのが提案モデルであるため，既存モデルよりも提案モデルの方が実用性という意味でも優れていると結論づけられる．ただし，既存モデルから Precision の値はほとんど向上しなかったため，より性能の高いモデルを構築するには予測ラベルが 1 であった時に予測ラベルを 0 とするような条件も設定する必要があると考えられる．

最後に，パターンマッチングを用いた手法において本研究では誤検出を減らすためにファインチューニングした感情の 2 値分類を行うモデルを導入したが，アブレーションスタディの結果からパターンマッチングを利用した手法のみでは既存モデルよりも Precision の値が低くなってしまうことが明らかになったため，誤検出を減らすための工夫が不十分だったと考えられる．逆に，表 15 よりパターンマッチングを用いた手法による Precison の値を除くと全てにおいて提案モデルの値が既存モデルを上回っているため，今後の 1 番重大な課題はパターンマッチングによる誤検出の減少であるとも考えられる．

7 おわりに

本研究では，既存の東北大 BERT モデルに対して幾つかのファインチューニングを行い，それらをパターンマッチングなどと合わせることで単に東北大 BERT モデルをファインチューニングした時よりも高い精度でポジティブな単語を含んだ煽り・誹謗中傷目的のコメントであるかどうかに関する 2 値分類を行うことが可能なモデルの構築を試みた．結果的には，Accuracy，Precision，Recall，F1 の 4 つ全ての評価指標において既存モデルを上回ることができ，提案モデルの実用性が示された．ただし，Precision の値はほとんど向上させる事ができなかったため，課題も残る形となった．

既存モデルよりも Recall の値が大きいモデルを構築することができたため，今後はアブレーションスタディの結果を考慮して特にパターンマッチングによる誤検出の増加についての課題に着目し，誤検出の可能性が現状よりも減った，すなわち Recall の値だけでなく Precison の値も向上したモデルの構築を目指していきたい．そのために，パターンマッチングにおいて感情分析の結果に関する条件だけでなく，構文パターンに該当する文中と文末のキーワードの組み合わせを現状よりも限定するなどといった新たな条件を模索していきたい．さらには，偏ったデータセットではなく様々な SNS などを利用して幅広くデータセットとなるコメントを収集した時も同じような精度を残す事ができるモデルの構築というのも今後の課題としたい．

謝　辞

本研究は JSPS 科研費 JP21H03496，JP22K12157，JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです．また，本研究は電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです．

文　献

[1] 一般社団法人セーファーインターネット協会. 誹謗中傷対策. 2023-3-3. https://www.soumu.go.jp/main_content/000866255.pdf
[2] Tohoku NLP Group. Pretrained Japanese BERT models. https://github.com/cl-tohoku/bert-japanese
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018.
[4] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/
[5] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144, 2016
[6] Hugging Face. https://huggingface.co/
[7] 磯野史弥，松吉俊，福本文代. Web 掲示板における皮肉 の分類および自動検出. 情報処理学会研究報告. 自然言語処理研究会報告 2013(7)，p.1-8
[8] 乾・鈴木研究室. 日本語評価極性辞書. https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html
[9] 肥合智史，嶋田和孝. 評価表現に着目した皮肉文の分析と抽出. 情報処理学会九州支部 火の国情報シンポジウム 2016
[10] Shota Suzuki, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Sarcasm Detection Method to Improve Review Analysis. The 9th International Coneference on Agents and Artificial Intelligence (ICAART2017)，Volume 2，pp.519-526，2017
[11] Yazhou Zhang, Dan Ma, Prayag Tiwari, Chen Zhang, Mehedi Masud, Mohammad Shorfuzzaman, and Dawei Song. Stance-level Sarcasm Detection with BERT and Stance-centered Graph Attention Networks. ACM Transactions on Internet TechnologyVolume 23，Issue 2. Article No.: 27，pp 1-21，2023
[12] モデルプレス エンタメ & ライフスタイルニュースサイト. YouTube 低評価動画ランキングトップ 20. 2021-11-13. https://mdpr.jp/news/detail/2865682

動画に対する否定的コメント・フィルタリングにおける few-shot 学習の評価

Evaluation of Few-Shot Learning in Filtering Negative Comments posted to Videos

三田寺　聖†　　宇津呂武仁††

† 筑波大学 理工学群 工学システム学類　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1

あらまし　本論文では，動画に対する否定的コメントに対して，大規模言語モデルを用いたフィルタリングを行う．大規模言語モデルを用いたコメントフィルタリングにおいては，現在最も性能が優れていると考えられる ChatGPT を用いても，本来肯定的であるはずのコメントを否定的であると判定する過剰フィルタリングが多数発生してしまう．この過剰フィルタリングは，あらかじめ少数の分類例を提示する few-shot 学習によって減少させることが可能である．本論文では，過剰フィルタリングを減少させることができる効果的な few-shot 学習の手法について検討し，評価を行った．その結果，ChatGPT を用いて内容に即したコメント・クラスターを作成し，これをもとに few-shot 学習の例示コメントを選定することによって，効果的に過剰フィルタリングを減少させることが可能であることが分かった．またこのとき，各クラスターの概要やコメント数を参考にして例示コメントを選定する最適手法を示す．

キーワード　LLM，テキスト分類，極性分類，動画コメント

1　はじめに

YouTube 等の動画サイトには，日々多種多様な動画がアップロードされており，これらの動画には視聴者から多くのコメントが寄せられている (例: 図 1)．その一部には動画の内容や出演者に対する否定的なコメントも散見され，投稿者や視聴者にとって不都合となることも多くある．本論文は，このような動画に対する否定的コメントに対して，大規模言語モデルを用いたフィルタリングを行うことを目的とする．

本論文では，否定的コメントをフィルタリングするモデルとして ChatGPT[1] を用いる．ChatGPT は，現在最も性能が優れていると考えられる大規模言語モデルである．しかし，ChatGPT を用いて否定的コメントをフィルタリングしても，動画内の文脈を読み取ることができずに，人間の感覚による分類と異なる誤フィルタリングが発生してしまう．この誤フィルタリングは，ChatGPT に対して few-shot 学習を行うことによって，改善することが可能である．

動画に対する否定的コメントのフィルタリングタスクにおいて，誤フィルタリングのうち大部分を占めるのが，本来肯定的であるはずのコメントを否定的であると判定する過剰フィルタリングである．そこで本論文では，一度 ChatGPT を用いて zero-shot 学習によるフィルタリングを行い，否定的であると予測されたコメントのみを対象とし，few-shot 学習によって再予測を行う．

few-shot 学習の際に用いる例示コメントは再予測対象コメントの中から数例を選定し，人手で分類を付与する．この際選定する例示用コメントは，類似するコメントが多数存在し，かつ肯定的である可能性の高いコメントである方がよい．そこで，ChatGPT でコメントのクラスタリングを行い，コメント数の多いクラスター内から最も肯定的であるコメントを選定することで，予測結果の改善により効果的な例示を実現する．

2　関連研究

感情分析タスクの代表的な手法として，評価表現辞書を用いた手法が考えられる．これは，個々の単語や表現に対して，あらかじめ肯定的・否定的の度合いを表す評価値を付与しておき，この値をもとにして文章全体の感情分析を行うという手法である．評価値を決定する辞書の作成手法については様々な研究が行われているが，その代表例として，共起情報を利用した手法に関する研究 [17] が挙げられる．この研究では，代表的な肯定単語として “excellent”，代表的な否定単語として “poor” を設定し，それぞれの単語との共起しやすさをもとに，各単語の評価値を決定している．しかしながらこの手法は，一文中に肯定的・否定的な要素が混在する場合や，未知語に弱いといった欠点が存在する．

一方で近年では，感情分析タスクには ChatGPT をはじめとする様々な大規模言語モデルが用いられている．感情分析に大規模言語モデルを用いた本論文の関連研究として，様々な感情分析タスクにおける複数の大規模言語モデルの性能を評価

1：https://chat.openai.com

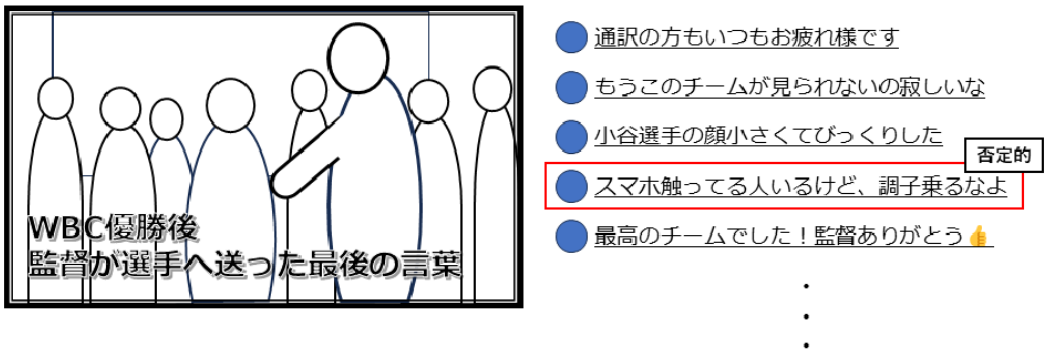


図 1　動画に寄せられるコメントの例

$$
\begin{aligned}
\text{適合率 (肯定的)} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right|}{\left|「\text{予測} = \text{肯定的}」の集合\right|} \\
\text{再現率 (肯定的)} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right|}{\left|「\text{参照} = \text{肯定的}」の集合\right|} \\
\text{適合率 (否定的)} &= \frac{\left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{予測} = \text{否定的}」の集合\right|} \\
\text{再現率 (否定的)} &= \frac{\left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{参照} = \text{否定的}」の集合\right|} \\
\text{分類精度} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right| + \left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{参照} = \text{肯定的}」の集合\right| + \left|「\text{参照} = \text{否定的}」の集合\right|}
\end{aligned}
$$

図 2　評価尺度として用いる適合率・再現率・分類精度の定義

した研究 [19] が存在する．この研究では `gpt-3.5-turbo` [8]，`text-davinci-003` [8]，`Flan-T5` [2]，`Flan-UL2` [15] の 4 つの大規模言語モデルが比較されており，いずれのタスクでも `gpt-3.5-turbo`，`Flan-UL2` の 2 つのモデルで特に高い精度が見られた．また，Instruction Tuning を行い一般目的の大規模言語モデルを金融感情分析に適応させ，金融感情分析の精度を向上した研究 [18] も関連研究として挙げられる．

プロンプトを活用して感情分析の精度向上を行った関連研究としては，プロンプトに背景情報を追加した際のヘイトスピーチ検出の精度を，複数の大規模言語モデルで評価した研究 [9] が挙げられる．この研究では比較対象となる大規模言語モデルとして `gpt-3.5-turbo`，`text-davinci-003`，`Flan-T5` の 3 つが用いられ，プロンプトやデータセットによってその精度に違いが見られた．また，段階的なプロンプトによって，言外の感情の分析精度を向上した研究 [3] や，感情の推論と推論結果の信頼性評価を 2 種類のモデルで二重に行い，感情分析の精度を向上した研究 [14] なども存在している．一方，本論文では，few-shot 学習に用いる例示コメントの選定に焦点を当て，例示コメントの選定にも大規模言語モデルを活用している点が，上記の研究と異なる点である．

few-shot 学習に用いる例示テキストの選定に関する研究としては，データセットから適切な例示コメントの選定を行い，大規模言語モデルを用いてプロンプトの生成を行った研究 [13] が存在する．この研究ではまず，kNN 検索や，fine-tuning によってテキスト分類タスクに特化させたモデルを用いて，入力テキストに対して最適となる例示テキストをデータセットから選定する．次に，ここで選定したテキストをもとに，大規模言語モデルを用いて手がかりとなるキーワードや推論の過程を出力し，few-shot 学習に用いるプロンプトを作成する．これを用いて感情分析やカテゴリ分類を行うことにより，大幅な精度向上を達成している．この研究では，あらかじめ分類ラベルが付与されたデータセットの中から，適切な例示コメントの選定を行っているのに対して，本研究では分類対象とするコメント群の中から例示コメントを選定し，その場で分類ラベルの付与を行っている，という点が相違点となっている．

近年では，動画コメントを活用し様々なタスクを解く研究が，複数行われている．動画コメントを感情分析に用いた関連研究として，動画コメントを数値ベクトルに変換し，機械学習によって炎上動画の検出を行った研究 [10] が存在する．この研究では，収集した動画コメントデータを用いてモデルの学習を一から行い，分類器を作成することで，炎上動画の分類を試みている．また，ルールベースや BERT，`gpt-3.5-turbo` など様々な手法で動画コメントから感情推定を行い，各手法の精度を比較した研究 [4] [5] も存在する．この研究では感情を 7 つに分類して感情値を 7 次元ベクトルとして表し，各手法によって推定された動画ごとの感情値を，人手で付与した感情値と比較することによって，各手法の精度を測っている．この他に，楽曲動画に寄せられたコメントから歌声の印象を分類した研究 [1] や，動画コメントを含む動画のメタデータから動画の再生回数予測を行った研究 [6]，動画コメントをもとに動画の評価項目ごとのスコアを算出することで，タイトルや説明文に表れない動画の特徴をユーザーに提示し，検索を可能にした研究 [16] [11] [12]

なども存在し，動画コメントは動画内容の分析を行うタスクにおいて幅広く活用されている．

3 動画およびコメントのデータセット

本論文では，YouTube が提供する Data API[2]を用い，YouTube 上の複数の動画から動画データ，およびコメントデータの収集を行った．対象とする動画は，コメント内容の多様性を確保し，なおかつ分析コストが大きくなりすぎないように，収集時点でのコメント数が 300 件以上，1 万件以下である動画に限定した．また，動画内容の多様性を確保するため，動画投稿者によって設定された動画カテゴリが偏ることのないように，動画の収集を行った．結果として 248 動画からコメントを収集した．

次に，動画コメントを 1 つずつ目視し，コメントの内容が肯定的・否定的のいずれであるかの分類を，各コメントに対して人手で付与した．ここで付与した人手分類を参照データとすることで，評価に用いるコメントデータセットを作成した．

本論文ではこのデータセットの中から，動画カテゴリ，全コメントにおける否定的コメントの割合等を考慮し，表 1 に示す 9 動画に対して評価を行った．

4 ChatGPT を用いたコメントのフィルタリング

本節では，ChatGPT を用いてコメントをフィルタリングする手法，およびその性能について述べる．なお，本研究では GPT-4 シリーズ [7] のモデルである `gpt-4-1106-preview` を利用する． 4.1 節では，例示を行わずに zero-shot 学習でフィルタリングを行う手順とその性能について述べる． 4.2 節では， 4.1 節で行った zero-shot 学習によるフィルタリングの結果をもとに，例示を用いた few-shot 学習によってフィルタリング精度を向上させる方法について述べる．

4.1 zero-shot 学習によるフィルタリング

zero-shot 学習によるフィルタリングでは，3 節で収集したコメントを用い，動画ごとにフィルタリングを行う．ChatGPT にコメントを 1 つずつ入力して感情分析を行い，肯定的・否定的のいずれかに分類した結果を出力させる．あらかじめ人手で付与した参照分類と出力結果が一致した場合正答，そうでない場合誤答とする．入出力の流れ，および予測結果の具体例を図 3 に示す．図 3 において評価尺度として用いた適合率，再現率，分類精度の定義を図 2 に示す．

zero-shot 学習によるフィルタリングによる予測結果では，肯定的であると予測されたコメントの適合率が高い一方で，否定的であると予測されたコメントの適合率が低くなる傾向にある．これは，ChatGPT を用いた zero-shot 学習によるフィルタリングでは，本来肯定的であるコメントを過剰にフィルタリングしてしまい，否定的であると予測されたコメントに誤分類が多いことが原因である．そこで次に，対象を zero-shot 学習によるフィルタリング時に否定的であると予測されたコメントに限定して再度フィルタリングを行うことによって，フィルタリング精度の改善を目指す．

4.2 否定判定コメントに対する few-shot 学習での再予測

few-shot 学習によるフィルタリングでは， 4.1 節で行った zero-shot 学習によるフィルタリングにおいて否定的であると予測されたコメントのみを対象とし，フィルタリングを行う．手順としてはまず，フィルタリング対象のコメントの中から数例のコメントを選定する．本研究では 10 コメントを例示コメントとして選定する．次にこれらのコメントを目視で判定し，肯定的・否定的のいずれかのラベルを適切に付与する．これを例示として用いて few-shot 学習を行い，再度分類した結果を出力させる．入出力の流れ，および再予測結果の具体例を図 4 に示す．

few-shot 学習によるフィルタリングによる再予測の結果，過剰にフィルタリングされるコメントが減少し，否定的コメントの再現率を大きく下げることなく，肯定的コメントの再現率を大きく改善することができる．しかし，これは例示コメントとして適切なコメントを人手で適切に選定したときの結果であり，実際には多数のコメントの中から自動で適切な例示コメントを選定することが必要である．ここで，全 n 個のコメントから 10 個の例示コメントを選定するパターン数は ${}_nC_{10}$ となる．本論文で評価を行った 9 動画では，表 1 に示すように，例示コメントを選定する対象となる “予測-否定的” コメントが最小でも 95 件存在する．そのため，例示コメントを選定するパターン数は最小でも ${}_{95}C_{10} = 1.01 \times 10^{13}$ となり，非常に大きくなってしまう．次節ではこれらのパターンの中から，フィルタリングをできるだけ多く改善させるために適切な例示コメントの選定手法について，検討を行う．

5 ChatGPT による例示コメントの選定

few-shot 学習によるフィルタリングで否定判定コメントの再予測を行う際，適切な例示コメントの条件として，「類似するコメントが多数存在する」ことが挙げられる．本節では，この条件について述べるとともに，条件を満たすような例示用コメントを，ChatGPT を用いて適切に選定するための手法について提案し，評価する． 5.1 節では，類似するコメントが多数存在するコメントを選定する手法として，ChatGPT を用いた類似コメントのクラスタリングについて述べる． 5.2 節では， 5.1 節で作成したコメント・クラスターをもとに，各クラスターから最適な例示コメントを選定する手法について述べる． 5.3 節では，提案手法について，実際に評価を行った結果を記述する．

5.1 コメントのクラスタリング

few-shot 学習による再予測では，例示コメントの内容を参考に再予測が行われるため，例示コメントと類似したコメントのフィルタリング結果が多く改善する．そのため，例示コメントでより多くの予測結果を改善するためには，類似するコメントが多数存在するコメントを選定することが必要である．

2：https://developers.google.com/youtube/v3

表 1　評価を行った動画

動画番号	動画タイトル	動画カテゴリ	コメント数				
			全体	人手分類結果 (参照データ)		zero-shot 学習結果 (予測データ)	
				肯定的	否定的	肯定的	否定的
動画 1	【応援ありがとう】栗山監督が最後に 選手たちに贈ったメッセージ	Sports	406	370	36	311	95
動画 2	映画『アバター：ウェイ・オブ・ウォーター』 本予告編【異次元の”没入型”映像体験】 12 月 16 日（金）劇場公開	Film & Animation	372	323	49	233	139
動画 3	【４人】間違いなく仲間に疑われる おもしろボードゲーム【お邪魔者】	Gaming	359	348	11	241	118
動画 4	【衝撃】九州→四国を最短で結ぶ ””隠されたルート””で移動してみた！	Travel & Events	379	342	37	223	156
動画 5	新型クラウンは今までと別モノ?! 土屋圭市が 試乗して驚愕！ 沢すみれも感激！ 工藤貴宏が徹底解説	Autos & Vehicles	313	213	100	130	183
動画 6	「コミケのために日本に来た」世界が注目の “コミケの世界”　ツボにはまる!? コミケの意外な楽しみ方 【N スタ解説】｜ TBS NEWS DIG	News & Politics	463	361	102	197	266
動画 7	【神速】誰も追いつけない”イカ速 3.9 スパッタリーヒュー”がチート すぎるんだがwwww【スプラトゥーン 3】	Gaming	409	408	1	304	105
動画 8	「Re:Unknown X」を弾いてみた 【東方ダンマクカグラ】	Music	423	419	4	292	131
動画 9	ペンギンの赤ちゃんの反抗期	Pets & Animals	361	340	21	265	96
平均	—	—	387.2	347.1	40.1	244.0	143.2

類似コメントが多数存在するコメントを選定するための手法として，ChatGPT を用いたコメントのクラスタリングを行う．これは，同じクラスターに分類されるコメントは類似性が高く，各クラスターに分類されたコメントの数が，類似コメントの多さを表す指標となるためである．例示コメントの選定対象となるコメントを ChatGPT に一度に全て入力し，クラスタリングを行うと，クラスターが細分化されてしまい，類似コメント数を見るのに適切なクラスタリングが行われない．そのため，以下の手順でクラスタリングを行う．まず，例示コメントを選定する対象となる，zero-shot 学習によるフィルタリングにおいて否定的であると予測されたコメントを，全て ChatGPT に入力し，コメント内容をもとに分類を行った際の各分類の概要を出力させる．次に，出力された概要を各クラスターの項目名として，それぞれのコメントがどのクラスターに分類されるのか出力を行う．これにより作成されたクラスターをもとに，コメント数の多いクラスターから例示コメントを選定することによって，類似コメントが多数存在するコメントの選定が可能となる．コメント・クラスタリングを行う際の入出力の具体例を，図 5 に示す．

5.2　コメント・クラスターを利用した例示コメント選定

続いて，5.1 節で作成したコメント・クラスターをもとに，ChatGPT を用いて few-shot 学習に用いる例示コメントの選定を行う．手順としては，1 つのクラスターに分類されたコメントを全て ChatGPT に入力し，あらかじめ決められた基準にしたがって，各クラスターから例示コメントを選定し出力する．入出力の具体的な流れの例を，図 6 に示す．

ここで重要となるのが，各クラスターからどのような基準にしたがって例示コメントを選定するか，という点である．本研究では，例示コメントを選定する手法として，以下の 2 つの要素を考える．

1 つ目の要素は，1 クラスターあたりの選定コメント数のように設定するかという点である．例示コメント数を一定とした場合，1 クラスターあたりの選定コメント数を少なくすることで，多くのクラスターからコメントを例示することが可能となる．一方で，1 クラスターあたりの例示コメント数が少ないと，規模の大きいクラスターの特徴をとらえきることが難しい可能性も考えられる．本研究では，4.2 節で述べた通り，10 コメントを例示コメントとして選定する．そこで，選定元となるクラスターと例示コメントの数を以下のように設定した，2 つの手法について評価を行う．

few-shot 学習の例示コメントの選定元クラスター数，および，選定する例示コメント数

- 10 クラスターから 1 コメントずつ選定
- 5 クラスターから 2 コメントずつ選定

2 つ目の要素は，few-shot 学習に用いる例示コメントとし

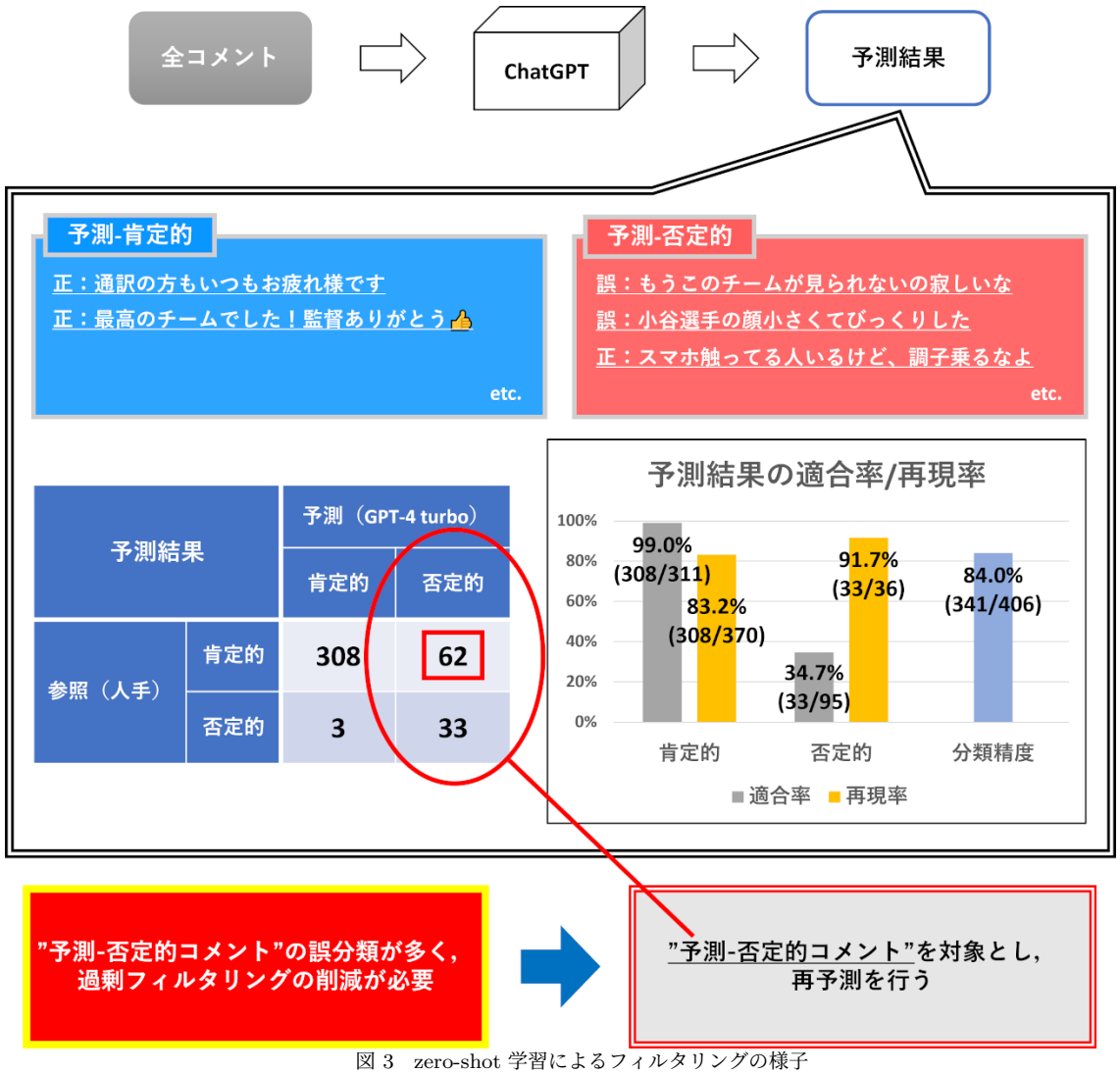


図 3　zero-shot 学習によるフィルタリングの様子

て，どのようなコメントを選定対象とするかという点である．few-shot 学習による再予測の対象となるコメントは全て，予測時点で否定的であると予測されたコメントであるため，正分類された実際に否定的であるコメントよりも，誤分類された本来肯定的であるコメントを例示として用いたほうが，より多くの予測結果を改善することができると考えられる．一方で，コメント・クラスターには，分類されるコメントの主な内容が否定的であるクラスターも存在するため，各クラスターにおける肯定的なコメントがそのクラスターの特徴を適切に表しているとは限らない．そこで，各クラスターから選定する例示コメントとして，以下の 2 つの手法を考え，評価を行う．

選定対象とする few-shot 学習の例示コメント

- 肯定的コメントを選定
- 代表コメントを選定

5.3　評 価 結 果

5.2 節で示した 2 つの要素を組み合わせ，計 4 手法に対して評価を行う．評価結果を表 2，および，表 3 に示す．ここでは，zero-shot 学習時に否定的であると予測されたコメントを，「“予測-否定的” コメント」と定義する．これは，few-shot 学習による再予測の対象となるコメントである．また，ero-shot 学習時

表 2　各コメント・フィルタリング手法における分類精度 (%)(マクロ平均)

(a) zero-shot 学習，および無作為に選定した例示コメントを用いた few-shot 学習

手法	全コメント内	“予測-否定的” コメント内
zero-shot 学習	72.7	23.9
few-shot 学習 (無作為 5 回の平均値)	83.6	57.4

(b) コメント・クラスターをもとに選定した例示コメントを用いた few-shot 学習

全コメント内 “予測-否定的” コメント内	肯定的コメントを選定	代表コメントを選定
10 クラスターから 1 コメントずつ選定	82.1 52.6	83.4 56.6
5 クラスターから 2 コメントずつ選定	**83.8** 58.1	83.3 **58.6**

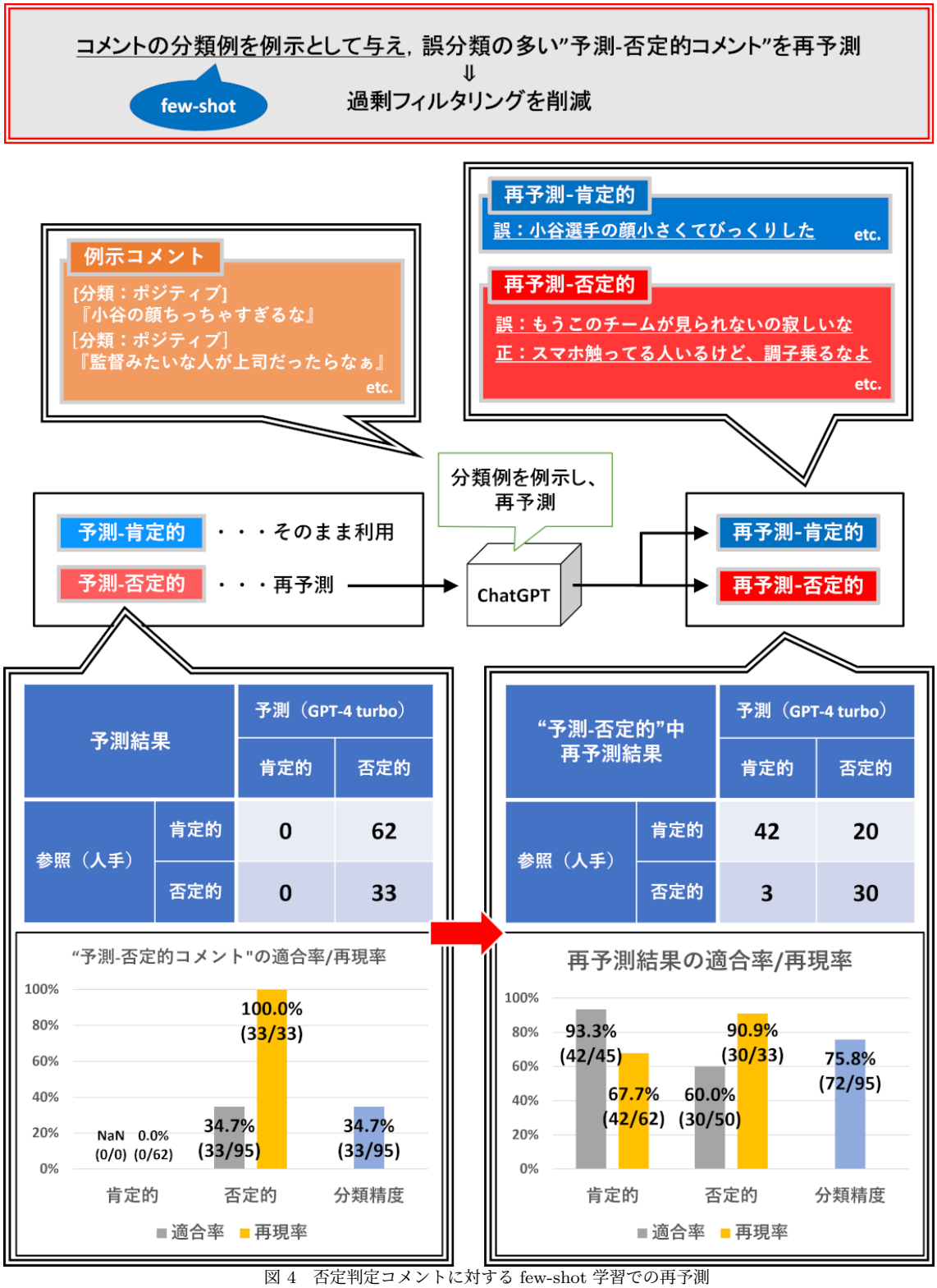


図 4　否定判定コメントに対する few-shot 学習での再予測

に肯定的であると予測されたコメントを「“予測-否定的” コメント」に加えた，1 動画における全てのコメントを，「全コメント」と定義する．本論文では，これら 2 種類のコメント集合内での分類精度を評価の基準とする．評価のベースラインとする，zero-shot 学習，および無作為に選んだ 10 コメントを例示した few-shot 学習の評価結果を表 2(a) に示す．また，4 手法それぞれにおいて 2 種類のコメント集合内での分類精度を算出した結果を表 2(b) に示す．さらに，評価を行った 9 動画それぞれの評価結果を表 3 に示す．なお，肯定的コメントを選定する手法と代表コメントを選定する手法では，選定した 10 コメントのうち，最大で 5 個，平均して 2 個程度のコメントが同一のものとなった．

表 2 より，few-shot 学習によってフィルタリングを行ったときの分類精度は，いずれも zero-shot 学習時の分類精度を大き

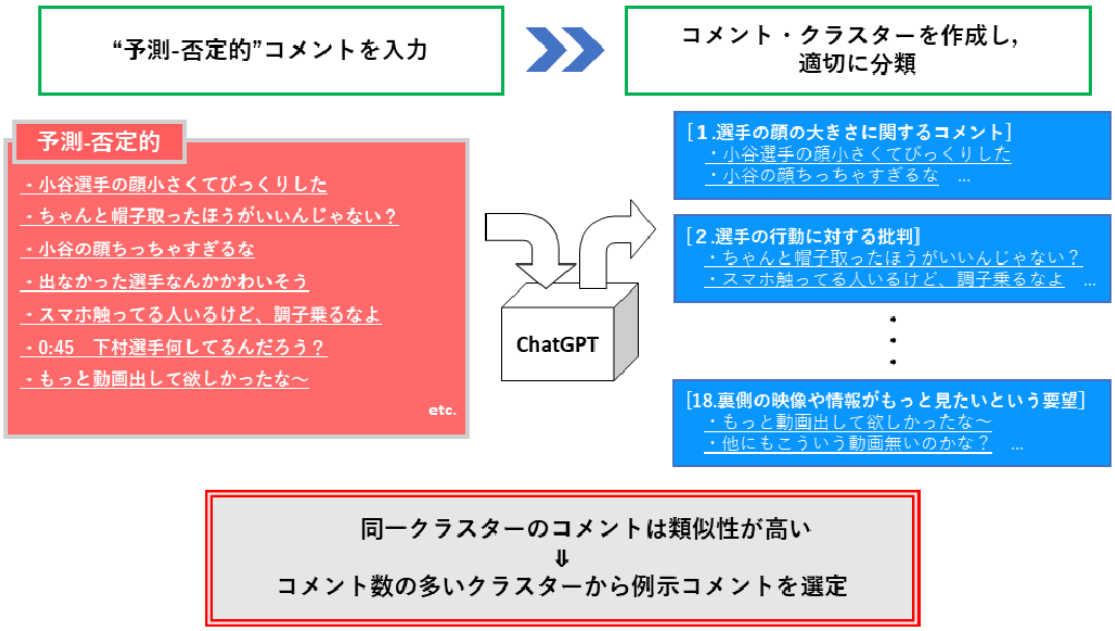


図 5　コメントのクラスタリング

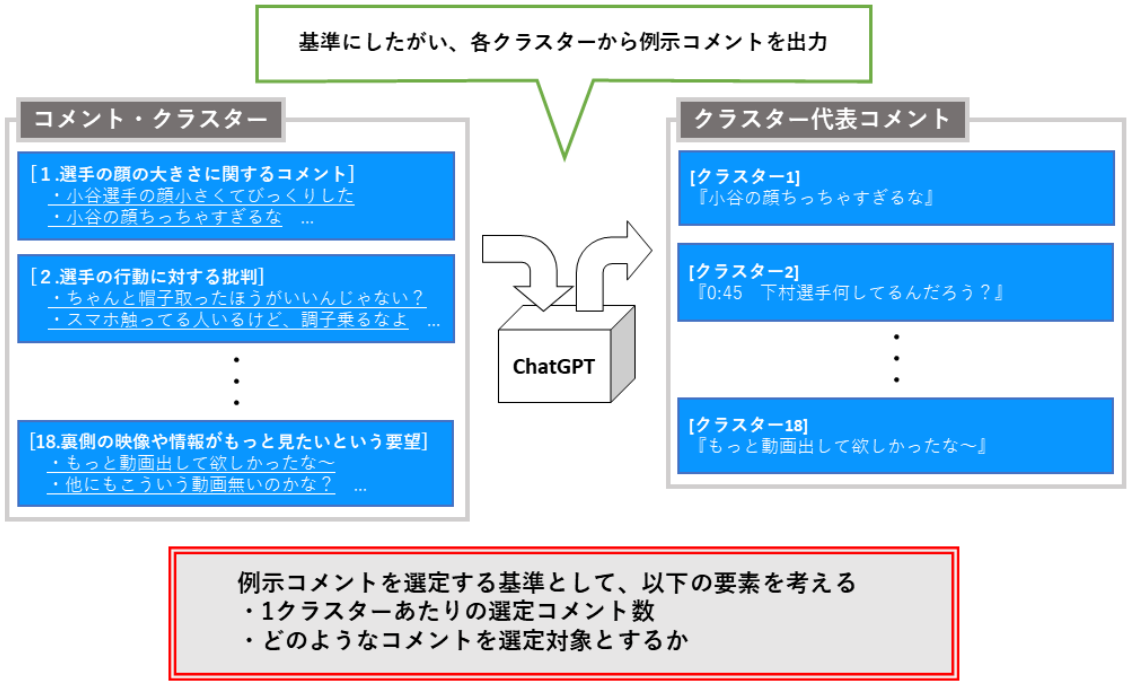


図 6　コメント・クラスターを利用した例示コメント選定

く上回っており，本論文における few-shot 学習を用いた手法が，コメント・フィルタリングにおいて有効であることが確認された．

10 クラスターから 1 コメントずつ選定を行う手法では，few-shot 学習の例示コメントとして肯定的コメント，代表的コメントのいずれを選定した場合でも，正解率のマクロ平均値が無作為に選定した 10 コメントを例示した際を下回った．特に，肯定的コメントを例示した場合では，無作為に選定した 10 コメントを例示した際の分類精度を上回ったものは 9 動画中 1 動画のみであった．一方で，5 クラスターから 2 コメントずつ選定を行う手法では，肯定的コメントを選定した場合，全コメント内，“予測-否定的” コメント内のいずれの分類精度の平均値も，無作為に選定したときの分類精度の平均値を上回った．特に，全コメント内での分類精度は 83.8%となり，コメント・クラスターを用いた例示コメント選定手法 4 つの中で，最も高い結果となった．また，5 クラスターから 2 コメントずつ，代表コメントを選定する手法においては，全コメント内の分類精度は無作為選定時を下回ったものの，“予測-否定的” コメント内の分類精度は 58.6%と無作為選定時を上回り，4 手法の中で最も高くなった．

表 3　各コメント・フィルタリング手法における動画ごとの分類精度 (%)

全コメント内 “予測-否定的” コメント内	zero-shot 学習	few-shot 学習			
		無作為に選定 (5 回の平均値)	コメント・クラスターをもとに選定		
				肯定的コメント を選定	代表コメント を選定
動画 1	84.0 34.7	89.1 56.4	10 クラスターから 1 コメントずつ選定	88.9 55.8	86.9 47.4
			5 クラスターから 2 コメントずつ選定	92.1 69.5	**92.4** **70.5**
動画 2	75.3 34.5	81.6 51.5	10 クラスターから 1 コメントずつ選定	80.9 49.6	80.1 47.5
			5 クラスターから 2 コメントずつ選定	**82.0** **52.5**	78.2 42.4
動画 3	70.2 9.3	88.1 63.9	10 クラスターから 1 コメントずつ選定	86.9 60.2	**93.0** **78.8**
			5 クラスターから 2 コメントずつ選定	90.0 69.5	92.5 77.1
動画 4	68.1 23.1	**80.9** **54.4**	10 クラスターから 1 コメントずつ選定	76.3 42.9	73.1 35.3
			5 クラスターから 2 コメントずつ選定	79.4 50.6	76.5 43.6
動画 5	71.2 53.0	73.1 56.2	10 クラスターから 1 コメントずつ選定	**75.1** **59.6**	72.8 55.7
			5 クラスターから 2 コメントずつ選定	**75.1** **59.6**	72.5 55.2
動画 6	62.4 36.5	74.0 56.7	10 クラスターから 1 コメントずつ選定	73.0 54.9	**77.8** **63.2**
			5 クラスターから 2 コメントずつ選定	72.8 54.5	69.3 48.5
動画 7	74.6 1.0	91.2 65.9	10 クラスターから 1 コメントずつ選定	87.8 52.4	**93.4** **74.3**
			5 クラスターから 2 コメントずつ選定	88.3 54.3	91.9 68.6
動画 8	70.0 3.1	87.1 58.3	10 クラスターから 1 コメントずつ選定	83.9 48.1	**88.2** **61.8**
			5 クラスターから 2 コメントずつ選定	87.0 58.0	87.2 58.8
動画 9	78.1 19.8	87.1 53.8	10 クラスターから 1 コメントずつ選定	86.1 50.0	85.0 45.8
			5 クラスターから 2 コメントずつ選定	87.3 54.2	**89.5** **62.5**

以上より，コメント・クラスターを用いた例示コメント選定手法においては，1 クラスター当たりの例示コメント数を複数とした方が，分類精度が高くなることが分かった．これは，1 つのクラスターに含まれるコメント全体の特徴を，1 コメントの例示のみでは表現することが難しかったためであると考えられる．また，選定対象とするコメントに関しては，肯定的コメント，代表コメントのいずれであっても，大きな変化は見られなかった．

6 おわりに

本論文では，動画に対する否定的コメントのフィルタリングを，few-shot 学習を用いて効果的に行う手法を提案した．ChatGPT を用いて一度コメント・フィルタリングを行い，否定的であると予測されたコメントをクラスタリングしたのち，クラスターごとに例示コメントを選定し再予測を行うことで，フィルタリング精度を大きく向上させることができた．

クラスターごとに例示コメントを選定する手法としては，1 つのクラスターから複数のコメントを選定することが効果的であることが分かった．また，選定するコメントは，クラスター内における肯定的なコメント，代表コメントのいずれであっても，精度向上に大きな差異は見られなかった．

今後の課題としては，さらに多くの動画を用いて few-shot 学

習に関する評価を行い，各手法における精度の向上と動画の特徴との関係について調査を行いたい．また，例示コメントのより効果的な選定手法についての検討も進めたい．

謝　　辞

本研究は科研費 21H00901 の助成を受けたものである

文　　献

[1] 陳沢璨, 熊本忠彦. J-POP における楽曲動画コメントを用いた歌声印象軸の構築. 第 16 回 DEIM フォーラム論文集, 2024.
[2] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
[3] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proc. 61st ACL*, pp. 1171–1182, 2023.
[4] 菅野祐希, 坂野遼平. YouTube コメントを用いた動画の感情推定 ルールベース及び BERT の精度比較. 第 15 回 DEIM フォーラム論文集, 2023.
[5] 菅野祐希, 坂野遼平. オンライン動画サービスにおける BERT 及び GPT-3.5 を用いた視聴者感情の推定. 言語処理学会第 30 回年次大会論文集, pp. 1067–1072, 2024.
[6] 川上大凱, 鈴木優. 動画の属するコミュニティ情報を考慮した動画再生回数予測手法の提案. 第 16 回 DEIM フォーラム論文集, 2024.
[7] OpenAI. GPT-4 technical report, 2023.
[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Proc. 36th NeurIPS*, pp. 27730–27744, 2022.
[9] S. Roy, A. Harshvardhan, A. Mukherjee, and P. Saha. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Proc. Findings of EMNLP*, pp. 6116–6128, 2023.
[10] 堺雄之介, 竹内幹太, 伊東栄典. コメントを利用した炎上動画検出に関する検討. 情報処理学会研究報告, Vol. 2021–ICS–203, No. 9, pp. 1–5, 2021.
[11] 笹原彰斗, 山口史弥, 上田真由美, 中島伸介. 直観的動画検索システムにおける評価項目別スコアリングの精度向上. 第 15 回 DEIM フォーラム論文集, 2023.
[12] 笹原彰斗, 山口史弥, 上田真由美, 中島伸介. 直感的動画検索における単語分散表現を用いた評価項目別スコアリング手法. 第 16 回 DEIM フォーラム論文集, 2024.
[13] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang. Text classification via large language models. In *Proc. Findings of EMNLP*, pp. 8990–9005, 2023.
[14] X. Sun, X. Li, S. Zhang, S. Wang, F. Wu, J. Li, T. Zhang, and G. Wang. Sentiment analysis through LLM negotiations. *arXiv preprint arXiv:2311.01876*, 2023.
[15] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler. UL2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2023.
[16] 鴇田和士, 上田真由美, 中島伸介. 動画に対する評価項目別スコアを用いた直感的動画検索システム. 第 14 回 DEIM フォーラム論文集, 2022.
[17] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th ACL*, pp. 417–424, 2002.
[18] B. Zhang, H. Yang, and X.-Y. Liu. Instruct-finGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*, 2023.
[19] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.

Conceptual Commonsense-aware Humor Recognition with Various Pretrained Masked Language Models(注)

Yuta SASAKI†, Jianwei ZHANG††, and Yuhki SHIRAISHI†††

† Tokyo Institute of Technology　Nagatsuta-cho, Midori-ku, Yokohama 226–8503, Kanagawa
†† Iwate University　Ueda 4-3-5, Morioka 020–8551, Iwate
††† Tsukuba University of Technology　Amakubo 4-3-15, Tsukuba 305–8520, Ibaraki
E-mail: †yubo1336@lr.pi.titech.ac.jp, ††zhang@iwate-u.ac.jp, †††yuhkis@a.tsukuba-tech.ac.jp

Abstract Humor is an important component of daily communication and usually causes laughter that promotes mental and physical health. Understanding humor is sometimes difficult for humans and may be more difficult for AIs since it usually requires deep commonsense. In this paper, we focus on automatic humor recognition by extrapolating conceptual commonsense-aware modules to Pre-trained Masked Language Models (PMLMs) to provide external knowledge. Specifically, keywords are extracted from an input text, and conceptual commonsense embeddings associated with the keywords are obtained by using a COMET decoder. By using multi-head attention, the representations of the input text and the commonsense are integrated. In this way, we attempt to enable the proposed model to access commonsense knowledge and thus recognize humor that is not detectable only by PMLM. Through the experiments on the two datasets, we explore different sizes of PMLMs and different amounts of commonsense and find some sweet spots of PMLMs' scales for integrating commonsense to perform humor recognition well. Our proposed models improve F1 by up to 1.7% and 4.1% on HaHackathon and Humicroedit datasets, respectively. The detailed analyses show our models also improve the sensitivity to humor while retaining the predictive tendency of the corresponding PMLMs.

Key words Humor recognition, Commonsense-aware attention, Knowledge-intensive NLP, Commonsense knowledge, Pre-trained Masked Language Model (PMLM), Keyword extraction

1 Introduction

Humor, which can usually cause laughter, is an important component of daily communication. [1] indicate that laughing with others reduces the risk of functional disability in old age. [2] show that laughter reduces the stress response and improves cognitive function by watching a Japanese comedy. As we can see, many medical and psychological studies have shown the positive effects of humor and laughter.

Although humor can be helpful, understanding humor is sometimes difficult for humans and may be more difficult for AIs. [3] show that ChatGPT cannot yet solve computational humor completely through detailed prompt-based experiments. Thus, we focus on automatic humor recognition in this paper.

In general, humor such as a joke consists of two parts: a setup that makes up a story and a punchline that concludes the story. It includes the elements of wordplay, misunderstandings, and irony. This provides funniness but also creates difficulties in the recognition of humor. To understand the punchline that completes humor, it is necessary to capture the context from the setup and to comprehend the commonsense that the words indicate. A rhetorical structure can trigger humor understanding, but deep and sufficient knowledge is necessary to understand humor. [4] shows that cultural differences can lead to misunderstandings of laughter through conversations between Spanish and English speakers. This shows that correct knowledge is necessary to understand humor and laughter.

We illustrate the process of understanding humor with commonsense, a type of knowledge, using the following example sentences:

> "How did the astronaut break up with his girlfriend?" "I just need some space."

The word "space" has multiple meanings including "distance" and "universe". If "space" is interpreted as "distance" between the astronaut and his girlfriend, this example cannot be recognized as humor. However, if people have commonsense

(注)：The work was mainly done during my affiliation with Iwate University

that "space" associated with astronauts can also be interpreted as the "universe" where astronauts work, this example is perceived as humor. Without sufficient knowledge to notice that "space" has some meanings, the humorous part cannot be understood. Deep and sufficient knowledge is crucial to comprehend humor.

Various humor recognition models have been proposed [5, 6, 7]. However, few methods specifically focus on knowledge. [8] use multi-modality to predict laughter tracks. They analyze prediction errors and show that humor, which requires appropriate knowledge, causes a wrong prediction. Thus, it is conceivable that humor recognition methods should have access to appropriate knowledge as humans do.

To address the above problem, we propose a model for humor recognition by extrapolating knowledge to pre-trained masked language models (PMLMs), focusing on commonsense. To enable the model to access commonsense, we use COMET [9], a GPT [10] model fine-tuned to reconstruct commonsense from natural language. The resulting representation is used as knowledge and integrated with the representation of the text from PMLM using Multi-Head Attention (MHA) [11]. Thus, we attempt to improve the performance of humor recognition by providing commonsense that cannot be considered only in PMLM. In our previous work [12], we build the experimental model using BERT as the backbone PMLM and evaluate the model. In this work, we refine the previous model using different sizes of PMLMs as the backbone and analyze the predictive tendency in detail.

Our contribution is as follows:

（1） We propose an architecture for humor recognition that considers various types of commonsense by extrapolating COMET to reconstruct commonsense. We softly integrate commonsense into the context of a text by treating the output of COMET as knowledge embedding to perform humor recognition.

（2） We find whether the commonsense helps in recognizing humor depends on the scale of PMLMs, the amounts of commonsense types, and the difficulty of tasks. Humor recognition performance can be improved when using medium-scale PMLMs and large amounts of commonsense types on two datasets. However, the model using large-scale PMLM cannot work well by infusing commonsense for both tasks. The contribution of commonsense to small-scale PMLMs is confirmed only when the restricted amounts of commonsense types are used for a less difficult task.

（3） Through detailed analysis, we verify that our models improve the sensitivity to humor since most predictions by PMLM-only are retained and the number of humor predictions increases.

2 Related Works

[7] use textual style and content-based features for humor recognition employing classical methods. [6] propose a deep learning-based method by introducing CNN and Highway Networks. [13, 14] construct large datasets for humor recognition from Reddit(注1). They perform the recognition task by fine-tuning PMLMs on the datasets. [5] propose ColBERT that considers the relationship between sentences by dividing the input text into sentences. In SemEval, a series of NLP workshops, some competitions for humor recognition tasks have been conducted, in which various systems have been proposed and many PMLMs have been used [15, 16].

Although many methods for humor recognition and generation have been proposed, not many of them explicitly deal with commonsense. [17] attempt to automatically generate a more consistent punchline by encoding the knowledge triplets of words extracted from Wikipedia as a graph structure.

In other tasks, several studies attempt to improve performance by accessing commonsense. [18] investigate the effect of commonsense on sarcasm detection by extracting relevant commonsense from the input text and comparing knowledge selection strategies. [19] consider commonsense as a graph structure and apply GCN to irony detection. Commonsense is also used in empathetic dialogue to attempt empathetic response generation [20, 21].

In this work, we augment the knowledge of PMLMs for humor recognition by using commonsense as external knowledge and softly integrating it into the models.

3 Proposed Method

Our proposed model is illustrated in Figure 1. Our model consists of three main parts: (1) Context Encoder, (2) Commonsense Acquisition Module, and (3) Commonsense-aware Humor Classification Module.

3.1 Context Encoder

This encoder obtains the embedding that indicates the context of the input text. In this study, Transformer-based PMLM is used as the context encoder. Given the input text I, the tokenizer of the PMLM returns the token sequence X of I, prepending a special token [CLS] to it:

$$\begin{aligned} X &= tokenizer(I) \\ &= \{[CLS], x_1, ..., x_{L-1}\} \end{aligned} \tag{1}$$

The PMLM receives X and outputs the embeddings for each token. Especially, the embedding corresponding to the first token [CLS] is extracted since the embedding represents the meaning of the entire input text:

(注1)：https://www.reddit.com/

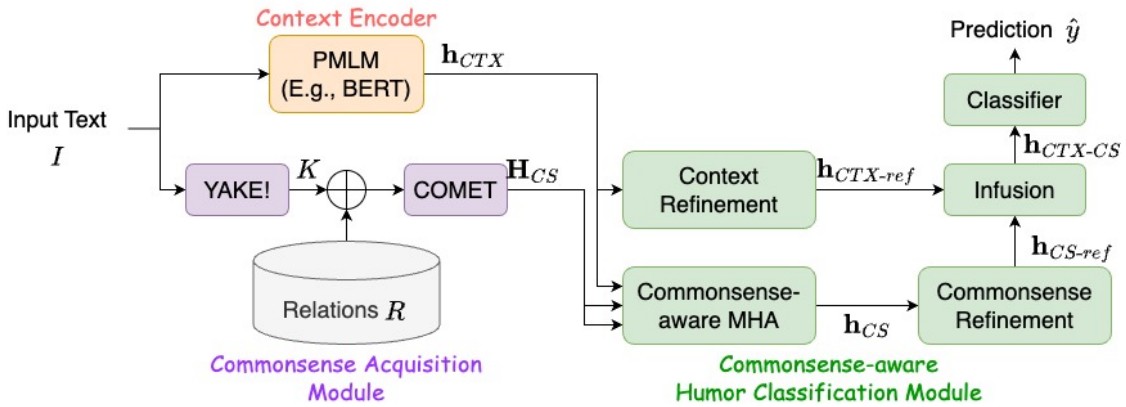


Figure 1: Our proposed model architecture. It mainly contains three parts: (1) Context Encoder, (2) Commonsense Acquisition Module, and (3) Commonsense-aware Humor Classification Module.

$$\mathbf{h}_{CTX} = PMLM(X) \tag{2}$$

where $\mathbf{h}_{CTX} \in \mathbb{R}^{d_{CTX}}$, L is the length of the input text and d_{CTX} is the hidden size of the PMLM output.

3.2 Commonsense Acquisition Module

The commonsense acquisition module illustrated in Figure 2 is composed of a keyword extractor and a commonsense encoder.

We use the commonsense generator COMET for the commonsense encoder. COMET has a limitation on its input to generate knowledge. The length of the input was short while COMET was fine-tuned. Through the preliminary experiments, we have tried three types of input to COMET: the raw input text, the summary of the raw input, and the keywords in the raw input. We find that the best method is to use the keywords in the raw input. Since both the raw input text and the summary of the input are long, they may be unsuitable for the input to the commonsense encoder. Thus, we adopt the keyword extractor before the commonsense encoder.

3.2.1 Keyword Extractor

To capture the commonsense associated with an input text, it is necessary to first apprehend the concepts of the words that constitute the text.

As mentioned earlier, keywords in the raw input are suitable for COMET as the commonsense encoder. Therefore, this component extracts important words from the input using a keyword extractor.

We use YAKE! as the keyword extractor [22]. With hyperparameters, it is possible to set the threshold of acceptable word duplicates and the maximum n-gram of keywords, and thus the output of keywords can be easily adjusted. The keyword extractor using YAKE! can be formulated as follows:

$$\begin{aligned} K &= YAKE!(I) \\ &= \{k_1, k_2, ..., k_m \mid m \leqq M_{kwd}\} \end{aligned} \tag{3}$$

where K is the resulting set of keywords and M_{kwd} is the maximum number of keywords to be input to the following commonsense encoder. The m keywords are selected based on the scores of candidate keywords that YAKE! outputs.

3.2.2 Commonsense Encoder

The commonsense encoder obtains embeddings representing commonsense from the keywords extracted by the keyword extractor. We use COMET as this encoder. COMET is the fine-tuned GPT to generate commonsense in the form of natural language from input words and phrases. COMET has two types of model, which are respectively trained on two datasets: ConceptNet [23] and ATOMIC [24]. ConceptNet is a word-level commonsense knowledge base consisting of a graph structure of descriptions of word concepts. The edges of the graph are represented by relations, and there are 34 relations such as "IsA" and "Causes." ATOMIC is a large if-then knowledge base about daily human interaction. In this work, we target conceptual commonsense about the extracted keywords using COMET trained on ConceptNet. Concatenated texts of keywords and relations are input to COMET. For example, if the keyword is "space" and the relation is "IsA," "space IsA" is the input into COMET. Thus, the input $K' = \{k_1 \oplus r_1, k_1 \oplus r_2, ..., k_1 \oplus r_n, k_2 \oplus r_1, ..., k_m \oplus r_n\}$ to COMET is constructed from the keyword set $K = \{k_1, k_2, ..., k_m \mid m \leqq M_{kwd}\}$ and the relation set $R = \{r_1, r_2, ..., r_n\}$. The relations that compose R can be chosen arbitrarily from 34 relations according to the task. $\oplus$ denotes the text concatenation operation. Instead of using COMET to explicitly generate the commonsense as a sentence, we let COMET act as an encoder by using the hidden state in the final layer corresponding to the last token as the commonsense embedding. This enables embedding-level operation of softly integrating the commonsense into the context embedding. The commonsense encoder can be formulated as follows:

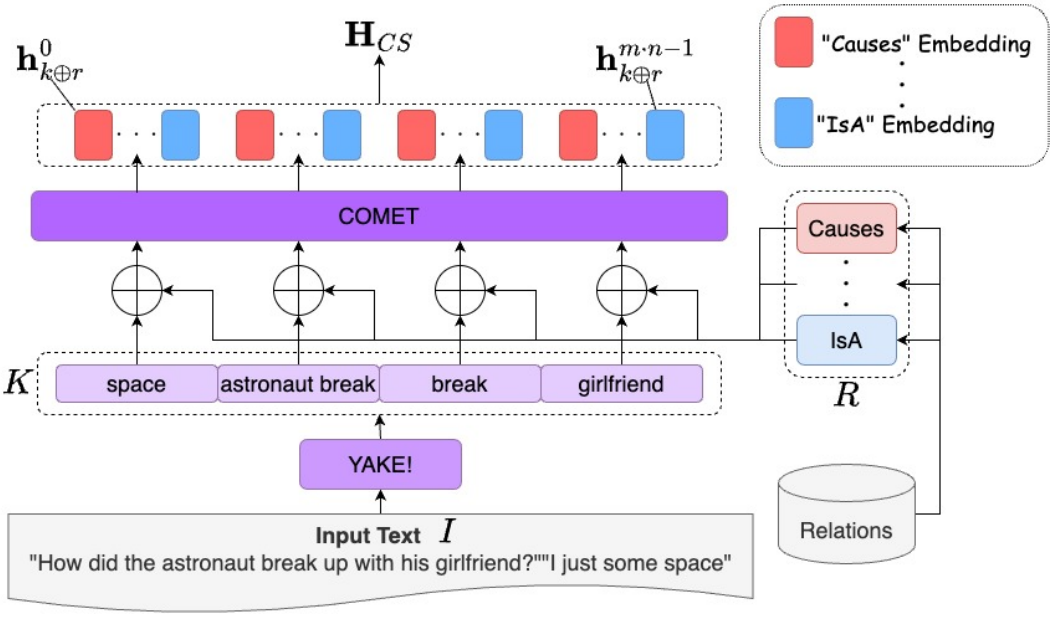


Figure 2: Commonsense acquisition module.

$$\mathbf{h}^i_{k\oplus r} = COMET(K'[i]) \tag{4}$$

$$\mathbf{H}_{CS} = [\mathbf{h}^0_{k\oplus r}, \mathbf{h}^1_{k\oplus r}, ..., \mathbf{h}^{m\cdot n-1}_{k\oplus r}] \tag{5}$$

where $K'[i]$ is the text that concatenates a keyword and a relation, $\mathbf{h}^i_{k\oplus r} \in \mathbb{R}^{d_{CS}}$ is the embedding from COMET corresponding to the last token of the input, $\mathbf{H}_{CS} \in \mathbb{R}^{(m\cdot n)\times d_{CS}}$ represents the keyword-based commonsense embeddings, $i \in \{0, 1, ..., m \cdot n - 1\}$, and d_{CS} is the hidden size of COMET.

3.3 Commonsense-aware Humor Classification Module

This module is further composed of five components: Context Refinement, Commonsense-aware MHA, Commonsense Refinement, Infusion, and Humor Classifier.

a) **Context Refinement**

The context embedding $\mathbf{h}_{CTX}$ obtained by Eq. (2) is refined using Multilayer Perception (MLP) for humor understanding. MLP consists of two linear layers of $\mathbb{R}^{d_{CTX}} \mapsto \mathbb{R}^{d_{CTX}}$ and the hyperbolic tangent function. This module outputs $\mathbf{h}_{CTX\text{-}ref} \in \mathbb{R}^{d_{CTX}}$:

$$\mathbf{h}_{CTX\text{-}ref} = MLP(\mathbf{h}_{CTX}) \tag{6}$$

b) **Commonsense-aware MHA**

This module uses multi-head attention (MHA) and softly selects commonsense by the context embedding $\mathbf{h}_{CTX}$. Thus, MHA is performed with the context as the query and the commonsense sets as the key and value. Formulating MHA as $MHA(query, key, value)$, this module is as follows:

$$\mathbf{h}_{CS} = MHA(\mathbf{h}_{CTX}, \mathbf{H}_{CS}, \mathbf{H}_{CS}) \tag{7}$$

where each input to MHA is projected to $\mathbb{R}^{d_{CTX}}$ before calculating the attention and $\mathbf{h}_{CS} \in \mathbb{R}^d_{CTX}$ represents the commonsense information selected by context.

c) **Commonsense Refinement**

Similar to the context refinement, the commonsense embedding $\mathbf{h}_{CS}$ obtained by Eq. (7) is also refined by MLP. As a result, $\mathbf{h}_{CS\text{-}ref} \in \mathbb{R}^{d_{CTX}}$ can be obtained:

$$\mathbf{h}_{CS\text{-}ref} = MLP(\mathbf{h}_{CS}) \tag{8}$$

d) **Infusion**

By simply adding the refined commonsense and context embeddings, the context embedding softly infused with commonsense is obtained:

$$\mathbf{h}_{CTX\text{-}CS} = \mathbf{h}_{CTX\text{-}ref} + \mathbf{h}_{CS\text{-}ref} \tag{9}$$

e) **Humor Classifier**

The commonsense-aware context $\mathbf{h}_{CTX\text{-}CS}$ is input to the classifier for humor classification:

$$\mathbf{P} = softmax(\mathbf{W}_c\mathbf{h}_{CTX\text{-}CS} + \mathbf{b}_c) \tag{10}$$

where $\mathbf{W}_c \in \mathbb{R}^{2\times d_{CTX}}$ and $\mathbf{b}_c \in \mathbb{R}^2$ are the weight and bias of the classifier. $\mathbf{P}$ represents the probability for each label.

3.4 Objective function

The cross-entropy loss is used as the objective function for $\mathbf{P}(y)$, which is the probability corresponding to the label y, to optimize the model. This calculation is as follows:

$$\mathcal{L} = -(1-y)\log\mathbf{P}(0) - y\log\mathbf{P}(1) \tag{11}$$

where $y \in \{0, 1\}$ is a discrete value and $\mathbf{P}(\cdot)$ is a continuous value.

4 Experimental Setup

4.1 Datasets

In this study, two public datasets are used for our experiments. The statistics of the datasets are shown in Table 1.

Table 1: Dataset summary. The Train/Valid/Test sets are publicly provided. H and NH respectively denote Humor and Non-Humor labels.

	Train			Valid			Test		
Dataset	H	NH	Total	H	NH	Total	H	NH	Total
HaHackathon	4,932	3,068	8,000	632	368	1,000	615	385	1,000
Humicroedit	4,731	4,921	9,652	1,173	1,246	2,419	1,466	1,558	3,024

HaHackathon This dataset is used in SemEval 2021 task 7. 80% of the data comes from Twitter and the rest comes from Kaggle Short Jokes. For one text, an annotator answers the question, "Is the intention of this text to be humor?" and each text is evaluated by 20 annotators. The humor label of a text is determined based on the majority of its votes.

Humicroedit This dataset is used in SemEval 2020 task 7. The dataset published in this competition is derived from a previous study [25]. This dataset is constructed by a task that rewrites one word in news headlines collected from Reddit to create a humorous sentence. Each rewritten text consists of a heading of 4-20 words. The degree of humor is rated by several annotators with integer values from 0 to 3, and the average value is labeled as the score of the humor. Scores are set as follows:

`0 - Not funny` `1 - Slightly funny`
`2 - Moderately funny` `3 - Funny`

Since our task is binary classification, we set the data whose score is 0 as Non-Humor, and the others as Humor.

Notice the difference between the two datasets. Humicroedit contains data where the majority of words are the same but only one word differs. Since the change of only one word determines humor, it is a difficult task to recognize this kind of humor. Thus, the classification on Humicroedit is a more difficult task than that on HaHackathon.

4.2 Experimental Configuration

We use several PMLMs with various numbers of parameters. Those are DistilBERT (66M) [26], BERT-base (110M) [27], DeBERTa-base (184M) [28], BERT-large (345M) and DeBERTa-large (435M) on HuggingFace(注2). We use each PMLM without introducing commonsense as a baseline. For our proposed model, the PMLMs are used as the context encoder, and COMET (117M) trained on ConceptNet is used to generate commonsense.

AdamW [29] is adopted as the optimizer, the learning rate of PMLMs as both the baseline and the context encoder is set to 2e-5, and the learning rate of other modules is set to 5e-5, while the weights of COMET are fixed. The batch size is 32. We attempt two patterns of inputting the relations to COMET. The first approach uses five relations, "IsA," "HasA," "Causes," "Desires" and "UsedFor." We adopt these for commonsense acquisition because they are intuitively comprehensible. All 34 relations are used in the second approach. As the hyperparameters of YAKE!, dedupLim and n-gram are set to 0.3 and 3, respectively, and the maximum number of keywords M_{kwd} is 6. We set the dropout rate to 0.4 and 0.1 for HaHackathon and Humicroedit, respectively. The number of CA-MHA heads is 8 in the experiments since the setting yields better results than the other number of heads in our preliminary experiments. All experiments are performed on a Tesla T4 using Google Colaboratory.

(注2)：https://huggingface.co/

5 Experimental Results

We evaluate the performance of our models on the two datasets. Besides, to clarify how our models work well, we compare the difference in the predictive tendency between the baseline and our best model.

5.1 Model Performance

The results of our proposed models and the baselines are shown in Table 2. As evaluation metrics, precision, recall, and F1 are employed. We test our models with different sizes of PMLMs used in the context encoder. The proposed models with each PMLM are also compared by using five relations and all 34 relations for commonsense acquisition.

First, we focus on the performance of models on HaHackathon. Focusing on our models with the small-scale PMLMs (DistilBERT and BERT-base), ours w/5-rel outperform their baselines in terms of almost all the metrics, and ours w/34-rel sometimes work the best. Although our models using the medium-scale PMLMs (DeBERTa-base and BERT-large) with five relations cannot defeat their baselines, ours w/34-rel achieve the highest performance in terms of almost all the metrics. In contrast, our model using the large-scale PMLM (DeBERTa-large) cannot improve the performance regardless of the used relations. Compared with the baselines except for the case of using DeBERTa-large, our proposed models improve F1 by 1.2% to 1.7% when using the proper amount of relations.

Next, we observe the model performance on Humicroedit. Our models using medium-scale PMLMs outperform their

Table 2: Performance comparisons on HaHackathon and Humicroedit. To compare the performance among the models using the same PMLM, the largest scores among them are highlighted in bold and the second largest scores are underlined. The first model in each block is the PMLM-only baseline. Ours with 5/34-rel means our proposed model by using the corresponding PMLM as the context encoder and using five or all the 34 relations as the input to COMET. The scores for BERT-large as a baseline on Humicroedit are marked with * and excluded from the comparison since all the prediction labels are Humor.

		HaHackathon			Humicroedit		
Scale	Model	Precision	Recall	F1	Precision	Recall	F1
Small	DistilBERT	**94.7**	88.0	91.2	64.2	**47.7**	**54.7**
	Ours w/5-rel	94.2	**91.7**	**92.9**	**67.7**	35.4	46.5
	Ours w/34-rel	94.5	87.2	90.7	67.3	34.5	45.6
Medium	BERT-base	92.8	91.5	92.1	58.4	**70.1**	**63.7**
	Ours w/5-rel	92.8	92.4	92.6	**66.5**	39.2	49.3
	Ours w/34-rel	**93.9**	**93.0**	**93.5**	**66.5**	37.5	47.9
	DeBERTa-base	95.1	92.5	93.8	**64.7**	57.8	61.0
	Ours w/5-rel	94.8	91.9	93.3	62.6	63.4	63.0
	Ours w/34-rel	**96.8**	**93.3**	**95.0**	63.5	**66.8**	**65.1**
Large	BERT-large	91.4	95.3	93.3	48.4*	100*	65.3*
	Ours w/5-rel	**94.8**	91.1	92.9	**63.2**	31.2	41.8
	Ours w/34-rel	93.6	**95.4**	**94.5**	57.4	**54.9**	**56.2**
	DeBERTa-large	**96.1**	**96.4**	**96.3**	**69.6**	**51.5**	**65.6**
	Ours w/5-rel	94.7	95.6	95.1	61.8	37.7	46.8
	Ours w/34-rel	82.9	82.1	82.5	59.6	49.5	54.1

baselines with five relations and perform the best with all 34 relations in terms of both recall and F1. Especially, ours w/5-rel and w/34-rel using DeBERTa-base improve F1 of the baseline by 2.0% and 4.1%, respectively. On the other hand, our models using DistilBERT, BERT-base, or DeBERTa-large with either set of relations decrease the performance compared with their baselines. Namely, our models with small-scale or large-scale PMLMs do not work well on Humicroedit which is a difficult classification task. However, when using medium-scale PMLMs, commonsense infusion can be useful for humor recognition.

To summarize these results on both datasets, our models using the medium-scale PMLMs with more relations are more likely to improve the performance of humor recognition. On the contrary, since the large-scale PMLM itself has a sufficiently high capability for classification, extrapolating commonsense to it cannot provide further help but may worsen the performance regardless of the number of relations. Whether our models with small-scale PMLMs perform well or not is task-dependent, and commonsense infusion can not function for a difficult classification task.

5.2 Predictive Tendency Comparison

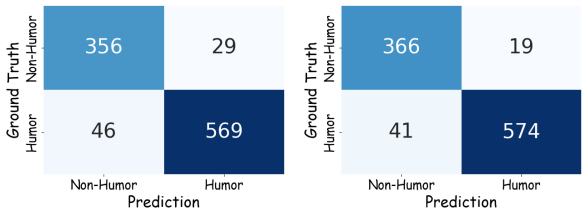
We evaluate the overlap and changes in the predictions of the baseline and our model. We compare the best model in our experiment, ours w/34-rel using DeBERTa-base, with its PMLM-only baseline.

Figure 3 and Figure 4 show the confusion matrices of those models on the two datasets. In Figure 3, compared to the confusion matrix of the baseline DeBERTa-base, our model increases the true positives (TP) and true negatives (TN), correct predictions respectively for Humor and Non-Humor labels, and decreases the false positives (FP) and false negatives (FN), wrong predictions respectively for Humor and Non-Humor labels. In Figure 4, our model increases TP and FP and decreases TN and FN. As shown in Table 2, by these changes, our model with DeBERTa-base improves all the metrics on HaHackathon and improves recall and F1 while precision decreases on Humicroedit.

In Table 3, the overlaps between the predictions of the baseline DeBERTa-base and our model are large, 96.5% on HaHackathon and 84.7% on Humicroedit. Thus, the proposed model that infuses commonsense can keep the predictions of PMLM for most data. The numbers of correct changes are 25 and 247 in contrast to the wrong changes of 10 and 215 on HaHackathon and Humicroedit, respectively. Although HaHackathon has more Humor data than Non-Humor, the number of prediction changes from Non-Humor to Humor (15) is almost the same as that of the changes from Humor to Non-Humor (20). Humicroedit has balanced data for the labels, however, 347 of the changes from Non-Humor to Humor are much more than 115 of the changes from Humor to Non-Humor. These results indicate that our

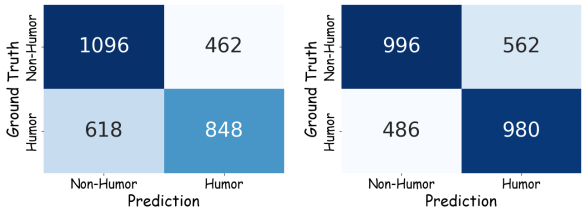
Table 3: The overlap and changes in the predictions on the test sets. The overlap shows the ratio of data on which both our best model that utilizes DeBERTa-base as the context encoder and all relations for COMET, and the PMLM-only baseline DeBERTa-base predict the same label. "$\hat{y}_{baseline} \rightarrow \hat{y}_{ours} : y$" is a notation where $\hat{y}$ denotes the predicted label and y denotes the target label. The columns, where our best model correctly predicts the label but the baseline does not, are bold.

Dataset	Overlap	**0→1:1**	0→1:0	1→0:1	**1→0:0**
HaHackathon	96.5%	**12**	3	7	**13**
Humicroedit	84.7%	**184**	163	52	**63**



(a) DeBERTa-base (b) Ours w/34-rel

Figure 3: Confusion matrices of DeBERTa-base and our model with DeBERTa-base and all relations on HaHackathon.



(a) DeBERTa-base (b) Ours w/34-rel

Figure 4: Confusion matrices of DeBERTa-base and our model with DeBERTa-base and all relations on Humicroedit.

model improves the sensitivity to Humor while retaining predictions from the corresponding PMLM-only model. Humor recognition is more difficult for Humicroedit since the ratio of change failures from Non-Humor to Humor on Humicroedit of 163/347 is much higher than the ratio on HaHackathon of 3/15.

6 Conclusion and Future Works

To improve the performance of humor recognition, we proposed a model that integrates the conceptual commonsense embeddings obtained from COMET with the context embedding of PMLM by the commonsense-aware multi-head attention. We find some sweet spots for the scale of PMLM, the amount of commonsense, and the task difficulty. Besides, our models are likely to increase the humor predictions while retaining most predictions by the corresponding PMLM-only, thus our models improve the sensitivity to humor.

Our model results in a slowdown in training and inference as the amount of commonsense increases. We will further discuss fast-access methods to a large amount of commonsense for real-time processing.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP22K12271 and JP19K12230.

References

[1] Yudai Tamada, Chikae Yamaguchi, et al. Does laughing with others lower the risk of functional disability among older Japanese adults? The JAGES prospective cohort study. *Preventive Medicine*, 155:106945, 2022.

[2] T. Yamakoshi, R. Sakamoto, et al. 笑いによるストレス応答抑制と認知機能改善効果. In *JAHP*, 2021.

[3] Sophie Jentzsch and Kristian Kersting. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *arXiv preprint arXiv:2306.04563*, 2023.

[4] Stuart Stewart. The Many Faces of Conversational Laughter. ERIC, 1997.

[5] Issa Annamoradnejad and Gohar Zoghi. ColBERT: Using BERT Sentence Embedding in Parallel Neural Networks for Computational Humor. *arXiv preprint arXiv:2004.12765*, 2020.

[6] Peng-Yu Chen and Von-Wun Soo. Humor Recognition Using Deep Learning. In *NAACL*, 2018.

[7] Rada Mihalcea and Carlo Strapparava. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *EMNLP*, 2005.

[8] Badri N. Patro, Mayank Lunayach, et al. Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms. In *WACV*, 2021.

[9] Antoine Bosselut, Hannah Rashkin, et al. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*, 2019.

[10] Alec Radford, Karthik Narasimhan, et al. Improving language Understanding by Generative Pre-Training. 2018.

[11] Ashish Vaswani, Noam Shazeer, et al. Attention Is All You Need. In *NeurIPS*, 2017.

[12] Yuta Sasaki, Jianwei Zhang, and Yuhki Shiraishi. Commonsense-Aware Attentive Modeling for Humor Recognition. In *DEXA*, 2023.

[13] Orion Weller and Kevin Seppi. Humor Detection: A Transformer Gets the Last Laugh. In *EMNLP-IJCNLP*, 2019.

[14] Orion Weller and Kevin Seppi. The rJokes Dataset: a Large Scale Humor Collection. In *LREC*, 2020.

[15] Nabil Hossain, John Krumm, et al. SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. In *SemEval*, 2020.

[16] J. A. Meaney, Steven Wilson, et al. SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In *SemEval*, 2021.

[17] Hang Zhang, Dayiheng Liu, et al. Let's be Humorous: Knowledge Enhanced Humor Generation. *arXiv preprint arXiv:2004.13317*, 2020.

[18] Jiangnan Li, Hongliang Pan, Zheng Lin, et al. Sarcasm Detection with Commonsense Knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3192–3201, 2021.

[19] Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. Does Commonsense help in detecting Sarcasm? In *Insights*, 2021.

[20] Pan Gao, Donghong Han, et al. CAB: Empathetic Dialogue Generation with Cognition, Affection and Behavior. In *DASFAA*, 2023.

[21] Quan Tu, Yanran Li, et al. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In *ACL*, 2022.

[22] Ricardo Campos, Vítor Mangaravite, et al. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

[23] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 2017.

[24] Maarten Sap, Ronan Le Bras, et al. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*, 2019.

[25] Nabil Hossain, John Krumm, and Michael Gamon. "President Vows to Cut< Taxes> Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In *NAACL*, 2019.

[26] Victor Sanh, Lysandre Debut, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC2 Workshop*, 2019.

[27] Jacob Devlin, Ming-Wei Chang, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[28] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *ICLR*, 2023.

[29] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

皮肉検出モデルにおける適切なコモンセンスの選別と性能評価

皆川　翔哉† 　佐々木裕多†† 　張　建偉†

† 岩手大学　〒 020–8551 岩手県盛岡市上田 4–3–5
†† 東京工業大学　〒 226–8503 神奈川県横浜市緑区長津田町
E-mail: †{s0620059,zhang}@iwate-u.ac.jp, ††yubo1336@lr.pi.titech.ac.jp

あらまし　皮肉は SNS 上でも使用され，ユーモラスな表現で特定の人を傷つけたり，物に対して批判したりするものである．人々は皮肉を用いて，表面上の文の意味とは反対の感情を表現する．そのため，皮肉検出は人々の真の感情を分析するために必要である．近年の研究では，コモンセンスを利用した皮肉検出手法が提案されている．コモンセンスとは，一般的に人間が持っている情報であり，日常の状況を理解するのに役立つ知識である．本研究が着目する既存研究の皮肉検出モデルでは，特定の 1 種類のコモンセンスを利用している．しかし，他の種類のコモンセンスを利用する検証は行われていない．また，生成されたコモンセンスを感情極性による知識選択で絞る手法が 3 種類提案されているが，最適な手法はデータセットに依存する．そのため，コモンセンス生成の段階で皮肉検出に有用で適切なコモンセンス情報を得ることが必要である．本研究では，様々なコモンセンスの種類を用いて皮肉検出モデルの性能を比較し，適切なコモンセンスの種類の選別を試みる．また，コモンセンス生成モデルへの入力テキストやコモンセンス生成後の処理の工夫を行う．さらに，複数の種類のコモンセンスを連結して利用する方法も試みる．実験結果より，3 つのデータセットにおいて，コモンセンスの種類によってモデル性能が異なり，適切な種類選択によって皮肉検出に対してノイズとなるコモンセンス情報が削減でき，既存手法よりも性能が向上することが示された．

キーワード　皮肉検出，コモンセンス，テキスト分類，SarDeCK，COMET

1　はじめに

皮肉は比喩的な言葉の一種であり，人々の日常生活で多くのコミュニケーションの目的として使用され [1]，特定の個人を批判することが一般的であり，SNS 上でも使用される．ケンブリッジ辞書における皮肉（Sarcasm）の定義は以下の通りである:

> the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way.
> （誰かを傷つけたり，何かをユーモラスに批判するために，話すこととは正反対の意味の発言を使用すること）

すなわち，人々は皮肉を表面上の文の意味とは反対の感情を表現するために使用する [2]．以下は皮肉文の一例である:

I just love waiting in line.
（私はただ列に並んで待つのが大好きである.）

例文においては，感情的な単語の “love” は，表面上ポジティブな印象を受ける．しかし，“waiting in line” は，ネガティブな状況に相当する．したがって，この例文は不平不満を伝える．このように，皮肉は文字通りの意味と発話者の意図は極めて異なり，難解である．感情分析や意見マイニングは，発話者の本当の考えを捉える必要があるため，皮肉を検出できない場合にそれらのパフォーマンスが低下する可能性がある [3]．皮肉の検出は，テキスト上からは読み取れない発話者の真意を理解することができるため，重要なタスクである．

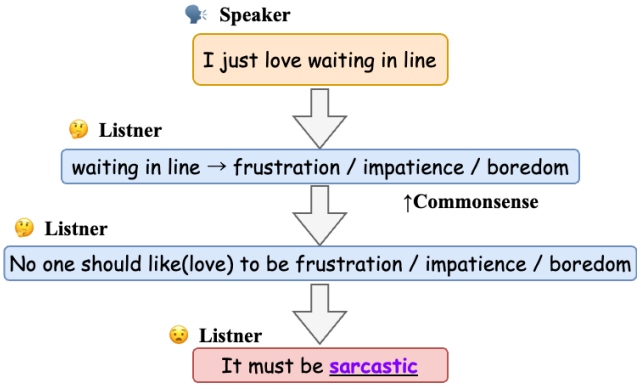


図 1　コモンセンスによる皮肉理解の例

近年，コモンセンスを用いた皮肉検出手法が提案されている [4] [5] [6]．コモンセンスとは，一般的に人間が持つ常識的な情報であり，日常の状況を理解するのに役立つ知識である．このような情報によって，人々が知識の断片を結びつけ，新しい結論を導くことができる．また，皮肉の理解が，しばしば皮肉テキスト外部のコモンセンスに依存することを指摘する研究もある．[7] コモンセンスによる皮肉理解のプロセスを図 1 に示す．先ほど挙げた皮肉の例文の “waiting in line” に着目すると，“frustration”，“impatience”，“boredom” などのコモンセンスを持つ．聞き手側はこれらを理解していなければ，皮肉として認識することは難しい．したがって，皮肉の理解にはコモンセンスが不可欠である．Li ら [4] が提案した皮肉検出モ

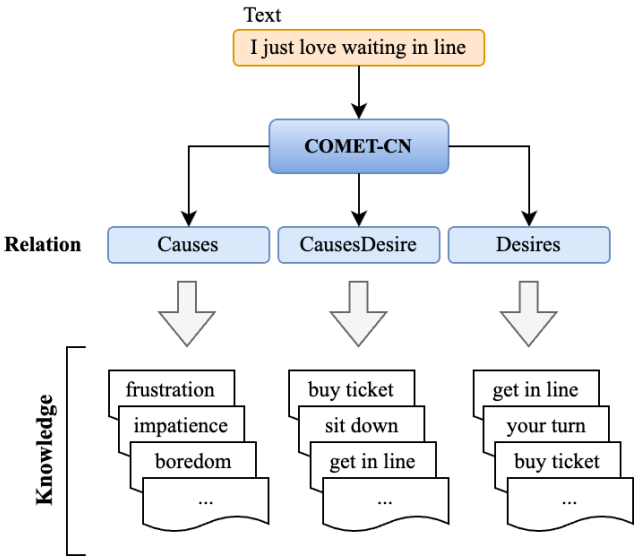


図 2　リレーション選択による生成コモンセンスの例

デル SarDeCK は，テキストとそのコモンセンスを利用するモデルである．コモンセンス生成には，COMET [8] を使用する．COMET は事前学習済み言語モデルを知識データセットでファインチューニングしたモデルである．知識データセットは，あるイベントから連想されるコモンセンスをリレーションと呼ばれるカテゴリに分類している．そのため，COMET は生成されるコモンセンスの種類をリレーションによって選択できる．図 2 にリレーション選択による生成コモンセンスの例を示す．SarDeCK は，Chakrabarty ら [9] による皮肉文生成の研究に従って，特定の 1 種類のコモンセンスのみ利用している．しかし，コモンセンスは様々な種類が存在するなかで，皮肉検出に適切なコモンセンスの種類を探索するための検証は行われていない．また，生成されたコモンセンスに対して，感情極性を利用した 3 つの知識選択手法により有用なコモンセンスに絞ることで，生成コモンセンス全てを利用するよりも性能が向上し，SOTA を達成している．これは皮肉検出にノイズとなるコモンセンス情報を削減したことによるものである．しかし，データセットによって最適な知識選択手法は異なる．そのため，コモンセンス生成の段階で皮肉検出にノイズとなるコモンセンス情報の削減をすることで，知識選択手法を用いずに性能向上を実現することが求められる．

本研究では，コモンセンス生成において利用する COMET のモデルアーキテクチャや知識データセットの変化による皮肉検出モデルの分類性能への影響を詳細に評価する．知識データセットである ConceptNet [10]，ATOMIC [11]，ATOMIC$^{20}_{20}$ [12] で学習された 3 種類の COMET をそれぞれ使用し，各リレーションから生成されるコモンセンスを利用することで，皮肉検出の性能がどの程度変化するのかを検証し，適切な知識データセットとリレーションの探索を試みる．また，テキストに関するコモンセンス情報をより的確にするために，COMET への入力テキストの処理およびコモンセンス生成後の処理の工夫を行い，影響を評価する．さらに，複数の種類のコモンセンスを連結して，単一の種類のコモンセンスを用いた場合と性能を比較する．

これらの分析により，コモンセンスの種類によって皮肉検出の性能が異なり，適切なコモンセンスの種類の選択によって皮肉検出にノイズとなるコモンセンス情報が削減され，既存手法よりも性能が向上することが示された．また，適切なコモンセンスの種類はデータセットに依存することが示された．さらに，適切な種類のコモンセンスに対して 3 つの知識選択手法を行った結果，全てにおいて性能が低下したことから，ノイズとなるコモンセンス情報の削減を実現できていることに加え，必ずしも知識選択手法が性能向上に寄与しないことが示された．単一のリレーションのコモンセンスを用いた場合よりも複数のリレーションのコモンセンスを用いた場合の方が性能が高くなり，性能を改善することができる可能性が示された．

本研究の貢献は以下の通りである:

- 既存の皮肉検出モデルに利用するコモンセンス情報のノイズを生成の段階で削減することを目的とし，適切なコモンセンスの種類を選択することを提案し，様々な種類のコモンセンスを利用することによる性能への影響をまとめる．
- 利用するコモンセンスの種類によって，既存の皮肉検出モデルの性能が異なり，既存手法の性能を上回るものが存在し，コモンセンスの種類選択の重要性を示す．
- コモンセンス情報の的確性の向上のために，COMET への入力テキストの処理およびコモンセンス生成後の処理の工夫を行い，その効果を示す．
- 複数の種類のコモンセンス情報を連結して利用することによる，皮肉検出モデルの性能向上の可能性を示す．

2　関 連 研 究

2.1　皮 肉 検 出

テキストベースの皮肉検出は主にルールベース，特徴量ベースの機械学習，深層学習の 3 つのアプローチに分けられる．近年の深層学習のアプローチ方法の中で，コモンセンスを用いた皮肉検出モデルが提案されている．本研究が着目する Li ら [4] は，これまで既存の皮肉検出でコモンセンスを取り込むことが難しいとされた問題に対処するため，効果的に知識を処理できる新しい BERT ベースモデルを提案した．Yu ら [5] は，コモンセンス感情グラフとコモンセンス置換依存グラフを構築することによって，コモンセンスと構文構造を明示的に統合することによって皮肉を検出するための新しいフレームワークを提案した．Min ら [6] は，皮肉テキストがネガティブなシチュエーションを捉えるためには不完全な文であるとした上で，結果と人間の反応に関するコモンセンスを取り入れた文を構築し，対象のテキストに加えて拡張サンプルを構築した．

2.2　コモンセンス知識グラフ

大規模なコモンセンス知識グラフは，自然言語処理のタスクにおいて普遍的なツールとなっており，テキストに対して推論を行い，その背景にあるコモンセンスを導き出すことができる．本研究では，ConceptNet [10]，ATOMIC [11]，ATOMIC$^{20}_{20}$ [12] を取り扱う．ConceptNet は，分類学的および語彙的な知識に焦点

を当てたリレーション（例：“RelatedTo”，“Synonym”，“IsA”）と，物理的なコモンセンスに焦点を当てたリレーション（例：“MadeOf”，“PartOf”）で合わせて 34 個のリレーションで構成される [1]．クラウドソーシングによって収集され，DBPedia [13]，WordNet [14]，Wikitionary[2]，OpenCyc [15] などの既存の知識データベースが統合された，340 万個の entity-relation 形式のタプルで構成されている．ATOMIC は，9 個のリレーションを持つ 88 万個のタプルで構成されており，変数を使用した型付きの if-then 形式で推論的な知識（例：“if X pays Y a compliment, then Y will likely return the compliment”）に焦点を当てている．ATOMIC$^{20}_{20}$ は，9 個のソーシャルインタラクションのリレーション（例：“xEffect”，“xReact”，“xWant”），7 個の物理的エンティティのリレーション（例：“AtLocation”，“HasProperty”，“Desires”），および特定のイベントを取り巻く状況に関する 7 個のイベント中心エンティティのリレーション（例：“HasSubEvent”，“Causes”，“xReason”）の合わせて 23 個のリレーションを持ち，entity-relation 形式の 133 万個の推論的な知識のタプルで構成されている．既存の ATOMIC や ConceptNet のタプルを取り入れており，より汎用的で大規模な知識グラフとなっている．

これらのコモンセンスのデータセットを事前学習済みモデルにファインチューニングしたモデルの総称を COMET（Commonsense Transformers for Automatic Knowledge Graph Construction）と呼ばれる．COMET を用いて，知識データセットのリレーションに基づいたコモンセンスの生成が可能となり，皮肉検出の研究においてもよく用いられているツールである．

3 本研究に用いる皮肉検出モデル

本章では，本研究が着目している Li ら [4] の提案モデル SarDeCK について紹介する．

3.1 タスク設定

皮肉検出は，ある発言に皮肉が含まれるかを判別するタスクである．SarDeCK は，n 個の単語を含むテキスト $X = \{x_1, x_2, \cdots x_n\}$ に対し，コモンセンスを外挿して皮肉検出を行う．モデルへの入力はテキスト，出力は 0 または 1 で，出力が 0 の場合は入力テキストは非皮肉，1 の場合は皮肉であることを示す．

3.2 モデルアーキテクチャ

SarDeCK は Encoding Module，Commonsense Reasoning Module，Knowledge-Text Integration Module の 3 つで構成される．図 3 にモデルのアーキテクチャの概要を示す．

3.2.1 Encoding Module

BERT ベースモデル [16][3] を適用し，入力テキストと知識テキストのエンコードをする．BERT は，事前学習された双方向 Transformer エンコーダで構成される Masked Language Model である．具体的には，[CLS] トークンを含む m 個のトークン $T = \{t_1, t_2, ..., t_m\}$ が与えられた場合，エンコードされたテキスト表現は $T_{enc} \in \mathbb{R}^{d*m}$．ここで，$d$ は隠れ層の次元数であり，T_{enc} は BERT の最終層の出力から得られる．

3.2.2 Commonsense Reasoning Module

知識生成と知識選択の 2 つのパートに分かれている．

知識生成部では，COMET [8] を使用する．COMET は事前学習済み GPT [17] を ConceptNet [10] の知識タプルでファインチューニングしたモデルを使用する．知識候補は，34 個のリレーションのうち “Causes” を選択して，生成する．COMET への入力テキストは前処理として，句読点削除・ストップワード削除・ローワーケース・レンマ化を行う．

知識選択部では，感情極性を用いた明示的知識選択とアテンション機構を用いた暗黙的知識選択の 2 つの方法がある．Li らの研究では，明示的知識選択によって皮肉検出モデルの SOTA (state-of-the-art) を達成していることから，本研究では明示的知識選択のみ取り扱う．明示的知識選択は知識候補の感情極性を計算するアルゴリズムによって知識を選択する手法であり，感情極性は SentiWordNet を用いて計算される．明示的知識選択は，以下の 3 つの手法が提案されている:

i. majority：全ての知識候補の感情極性の合計値が正であれば，感情極性が正の知識候補のみを選択し，知識候補の感情極性の合計値が負であれば，感情極性が負の知識候補のみを選択する．

ii. minority：全ての知識候補の感情極性の合計値が正であれば，感情極性が負の知識候補のみを選択し，知識候補の感情極性の合計値が負であれば，感情極性が正の知識候補のみを選択する．

iii. contrast sentiment-based：入力テキストの感情極性の合計値が正であれば，感情極性が負の知識候補のみを選択し，入力テキストの感情極性の合計値が負であれば，感情極性が正の知識候補のみを選択する．

知識選択後，選択された b 個の知識 $KS = ks_1, ks_2, ..., ks_b$ を連結する．連結された知識テキストのトークンが n 個の場合，エンコードされた知識表現は $KS_{enc} \in \mathbb{R}^{d*n}$ となる．

3.2.3 Knowledge-Text Integration Module

選択された知識と入力テキストの相互作用を学習し，混合された情報を含むベクトルを予測に使用する．以下の式より，テキストに関連した知識表現 KST_{enc} を得る．

$$KST_{enc} = softmax(T^T_{enc} KS_{enc}) KS^T_{enc} \qquad (1)$$

$T_{enc} \in \mathbb{R}^{d*m}$ はエンコードされた入力テキスト表現，$KS_{enc} \in \mathbb{R}^{d*n}$ はエンコードされた選択知識表現である．

入力テキストと知識をどの程度考慮すべきかを動的に学習するために，モデルにゲート機構を取り入れる．まず，単一ベクトル $KS_{mean} \in \mathbb{R}^d$ を得るために，$KS_{enc} \in \mathbb{R}^{d*n}$ に対して平均プーリングを行う．ゲートの出力 α は，以下の式で求められる．

1：https://github.com/commonsense/conceptnet5/wiki/Relations
2：https://www.wiktionary.org/
3：https://huggingface.co/google-bert/bert-base-uncased

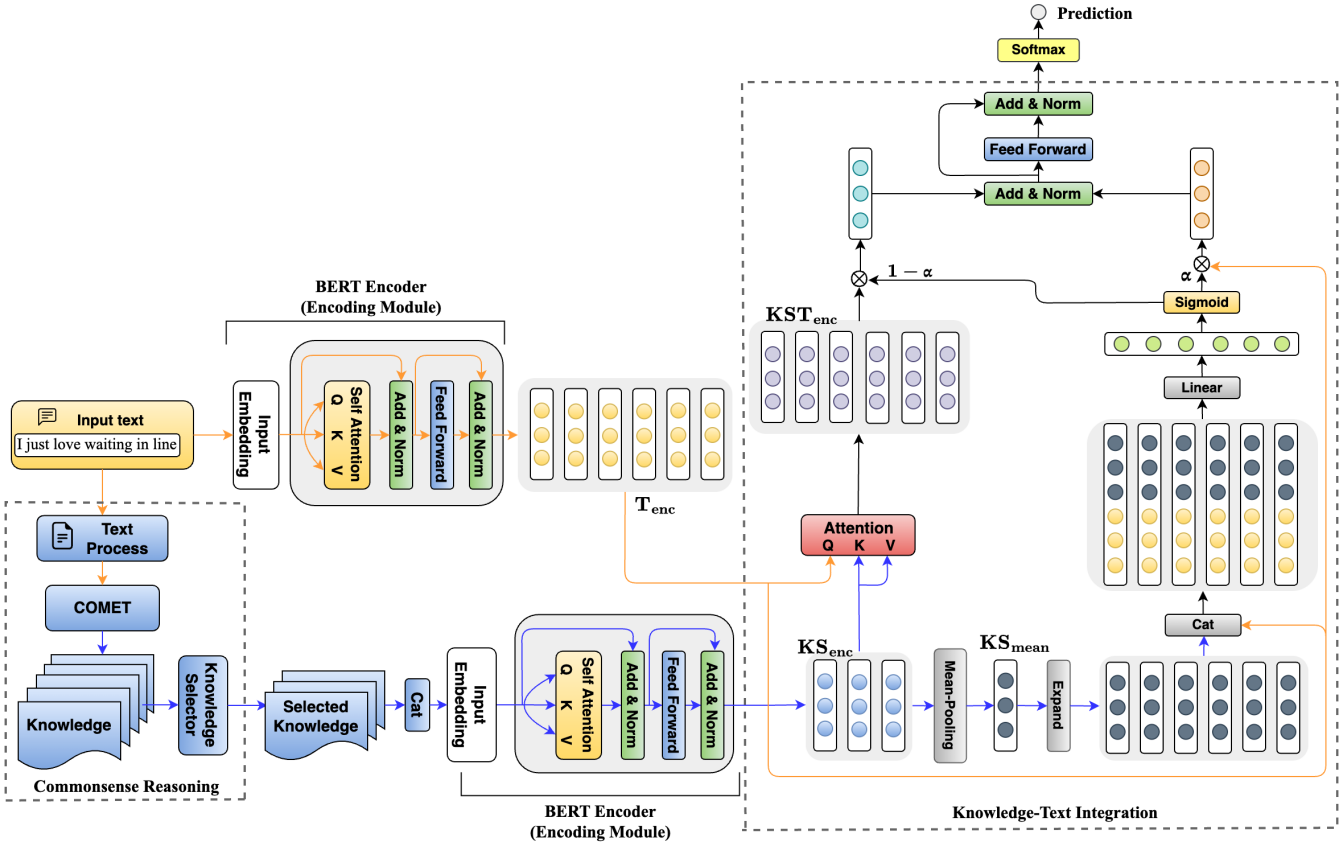


図 3　SarDeCK のモデルアーキテクチャ

$$\alpha = sigmoid([T_{enc}^{T} \oplus KS_{mean}]W_g) \tag{2}$$

$\alpha \in \mathbb{R}^{m*1}$ は重みベクトル，$W_g \in \mathbb{R}^{2d*1}$ は学習パラメータである．$\oplus$ は T_{enc}^{T} を KS_{mean} と連結している．

知識表現と入力テキスト表現をそれぞれ $T_{enc} \otimes \alpha$ と $KST_{enc} \otimes (1-\alpha)$ とし，再構築する．入力テキストと知識の情報を統合するために，残差接続をする．式は以下の通りである．

$$Z = LN(T_{enc}\alpha + W_b KST_{enc}(1-\alpha)) \tag{3}$$

LN はレイヤー正規化，$W_b \in \mathbb{R}^{d*d}$ は学習パラメータ，出力は $Z \in \mathbb{R}^{d*1}$ である．その後，Z と feed-forward ネットワークを通した Z を足し合わせたものを，再び残差接続を用いて，$\hat{Z} \in \mathbb{R}^{d*1}$ を得る．

$$\hat{Z} = LN(Z + MLP(Z)) \tag{4}$$

3.2.4　予　　測

予測部分は，線形層と softmax 関数から構成される．次元削減のために線形層，各カテゴリの確率を得るために softmax 関数に使用される．モデルは与えられた入力テキストを最も高いカテゴリに分類する．

$$\hat{y} = softmax(\hat{Z}W_c + b) \tag{5}$$

$W_c \in \mathbb{R}^{d*2}$ は学習パラメータ，b はバイアス項，$\hat{y}$ は分類結果である．

3.2.5　目 的 関 数

モデルは学習の際に，Cross-entropy loss 関数を使用し，モデルを最適化する．

$$J = -\sum_{i=1}^{N}[y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)] + \lambda R \tag{6}$$

J は損失，$\hat{y}_i$ はテキスト i を予測した結果，y_i はテキスト i の真のラベルである．N は学習データのサイズで，λR は L2 正則化項である．

4　提 案 手 法

SarDeCK のコモンセンス生成部に着目し，COMET への入力前処理，COMET の種類，リレーションの種類，コモンセンス生成後の処理の 4 つの部分において変更をする．これらの変更により生成されたコモンセンス情報を利用し，モデルを学習させる．単一および複数のリレーションのコモンセンスを用いる場合のプロセスをそれぞれ図 4，図 5 に示す．

4.1　COMET の種類

本手法で用いる COMET は，表 1 の通りである．各知識データセット ConceptNet [10]，ATOMIC [11]，ATOMIC$^{20}_{20}$ [12] で学習された 3 種類の COMET を使用する．各 COMET は，公

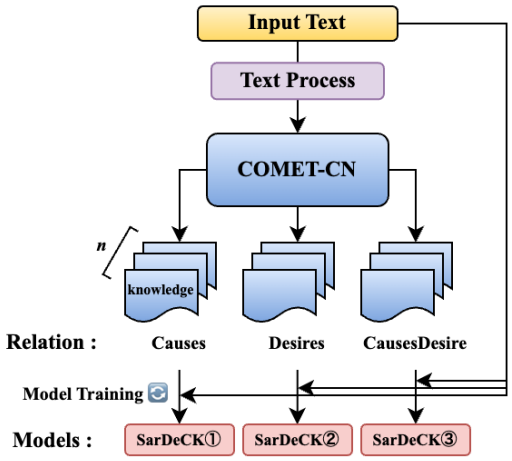


図 4　SarDeCK の知識生成処理の例（単一リレーション）

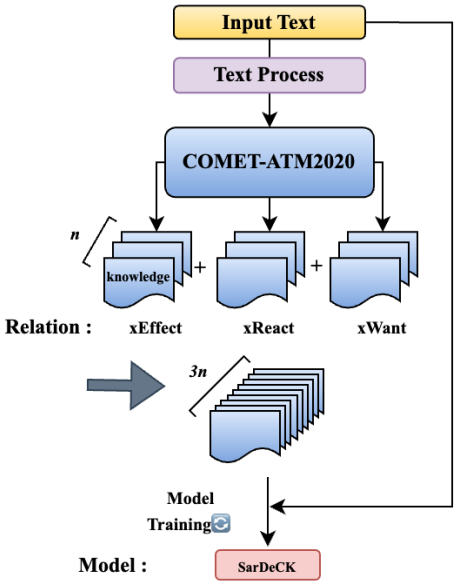


図 5　SarDeCK の知識生成処理の例（複数リレーション）

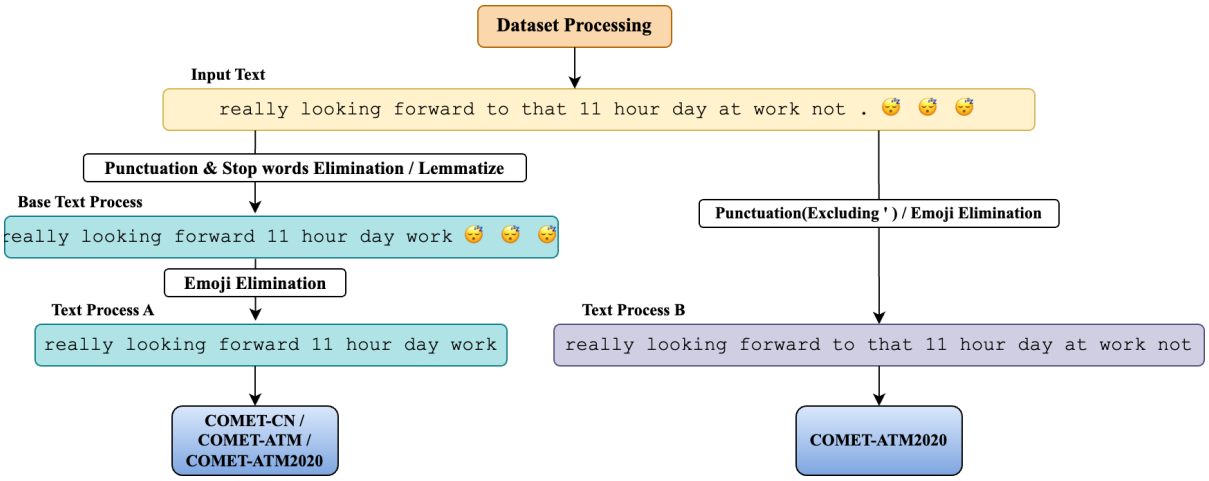


図 6　COMET の入力前処理の例

表 1　COMET の情報

PLM	Dataset	Notation
GPT	ConceptNet	COMET-CN
GPT	ATOMIC	COMET-ATM
BART	ATOMIC^{20}_{20}	$\text{COMET-ATM}^{20}_{20}$

表 2　本手法に用いるリレーション

Type of COMET	Relations
COMET-CN	Causes，CausesDesire，Desires
COMET-ATM	xEffect，xReact，xWant
$\text{COMET-ATM}^{20}_{20}$	Causes，CausesDesire，Desires，xEffect，xReact，xWant

開コード [4,5]を用いる．生成されるコモンセンスは，COMET の言語モデルとファインチューニングに用いた知識データセットに依存する．各 COMET を COMET-CN，COMET-ATM，$\text{COMET-ATM}^{20}_{20}$ と表す．

4：https://github.com/atcbosselut/comet-commonsense
5：https://github.com/allenai/comet-atomic-2020

4.2　COMET への入力前処理

図 6 に COMET の入力前処理の概要を示す．前提として，それぞれの皮肉検出データセットは各テキストにおいてハッシュタグ削除，リンク削除，トークナイズ，ローワーケース等のデータセット処理が既に施されている．COMET の入力前処理として，既存手法では句読点削除，ストップワード削除，レンマ化を行う．しかし，データセットには絵文字を含むテキストが含まれており，COMET は絵文字を認識できないため，ノイズとなる．そのため，既存手法の入力前処理に加えて，絵文字削除も行う．これを「入力前処理 A」と表す．また，実験で用いる $\text{COMET-ATM}^{20}_{20}$ は最大 1024 トークンの入力が可能である．そのため，文構造を維持したより長いテキストの入力も試みる．これを「入力前処理 B」と表し，シングルクウォートを除く句読点削除，絵文字処理の 2 つの処理を行う．

4.3 リレーションの選択

リレーションは手動で選択する．各 COMET において，表 2 で示されるリレーションからコモンセンスを生成する．

COMET-CN では，既存手法に用いられていた “Cause” に加えて，語彙的な変換をするリレーション以外のうち，“CausesDesire” と “Desires” を採用する．対象テキストが A で，B が知識として生成される場合，“CausesDesire” は「A が誰かに B を望ませる」，“Desires” は「A が B を望む」と定義されるリレーションである．

ATOMIC のリレーション：“xEffect”，“xReact”，“xWant” を採用する．あるイベントを A とすると，“xEffect” は「A が起こった結果，PersonX は〜する」，“xReact” は「A が起こった結果，PersonX は〜を感じる」，“xWant” は「A が起こった結果，PersonX は〜を欲する（したい）」と定義されるリレーションである．これらを用いる理由としては，皮肉検出の研究 [6] [18] で利用されているためである．

ATOMIC$^{20}_{20}$ のリレーションは，ConceptNet と ATOMIC のリレーションをカバーしているため，上記で採用した計 6 つのリレーションを採用する．

4.4 コモンセンス生成後の処理

COMET から生成されるコモンセンスの中には，空白の出力がされる場合がある．そのため，コモンセンスが空白の場合は削除する．また，ATOMIC，ATOMIC$^{20}_{20}$ においては，“none” という出力がされることがあるが，これもコモンセンス情報として必要がないと判断したため，削除する．

5 実　　験

5.1 皮肉検出データセット

皮肉検出データセットは，Li ら [4] の実験と同様の 3 つのデータセットを扱う．Ghosh [19] および Ptáček [20] が提案した旧 Twitter のデータセットと，Khodak [21] が提案した Reddit データセットのうち，subreddit が “politics” のデータを使用する．各データセットはテキストが皮肉か非皮肉かがアノテーションされており，これらをそれぞれ Tweets (Ghosh)，Tweets (Ptáček)，Reddit-pol と表す．Tweets (Ghosh)，Tweets (Ptáček) は絵文字を含むテキストが存在するが，Reddit-pol には含まれていない．表 3 に詳細を示す．本実験では，公開コード [6]からこれらのデータセットを利用した．

5.2 実 験 設 定

本実験の環境は，Google Colaboratory 上の Tesla T4 GPU である．コモンセンスを生成するため，公開コード [4,5] から事前学習済み COMET を使用する．また，SarDeCK は PyTorch を用いており，HuggingFace で提供されている BERT-base[3] を使用する．オプティマイザとして Adam を使用し，学習率は 5e-5，学習は 8 エポック，バッチサイズは 32，ドロップアウト率は 0.1 とする．テキストのトークン最大長は 40，連結後のコモンセンステキストのトークン最大長を単一のリレーションの場合は 20，複数リレーションの場合は 60 とする．これらの最大長を超えた分は，切り捨てられる．SarDeCK のモデル学習や評価は公開コード [6] を用いて行う．本実験の評価指標として，モデルの性能を表す F 値を使用する．

6：https://github.com/LeqsNaN/SarDeCK

表 3　皮肉検出データセットの情報

	emoji	train	dev	test	total
Tweets(Ghosh)	○	46,070	5,118	3,742	54,930
Tweets(Ptáček)	○	8,497	1,062	1,062	10,621
Reddit-pol	×	20,842	2,605	2,605	26,052

表 4　SarDeCK の F 値（既存手法）

	Tweets (Ghosh)	Tweets (Ptáček)	Reddit-pol
All	82.37%	84.05%	73.50%
Maj	82.41% ↑	82.52% ↓	**75.49%** ↑
Min	82.27% ↓	**84.51%** ↑	75.17% ↑
Con	**85.27%** ↑	83.52% ↓	73.48% ↓

5.3 実験結果と考察（単一リレーション）

コモンセンスは，既存手法と同様に各テキストから 5 個生成する．提案手法との比較対象として Li ら [4] の既存手法における実験結果を表 4 にまとめる．行は知識選択，列は皮肉検出データセットを表す．明示的知識選択をせずに，生成したコモンセンスを全て用いる手法を All と表す．既存手法とは，COMET-CN の Causes リレーションで生成したコモンセンスを利用し，COMET への入力前処理は，句読点削除，ストップワード削除，レンマ化を行う手法である．

表 4 は，本実験の環境で行った結果である．表 4 から，それぞれの皮肉検出データセットにおける最適な知識選択は，Tweets (Ghosh) は Contrast sentiment-based (Con)，Tweets (Ptáček) は Minority (Min)，Reddit-pol は Majority (Maj) である．したがって，既存の知識選択手法は SarDeCK の性能向上に寄与する．しかし，最適な知識選択はそれぞれ異なっており，データセットに依存する．

続いて，表 5，表 6，表 7 に提案手法での結果を示す．各表は，Tweets (Ghosh)，Tweets (Ptáček)，Reddit-pol の 3 つのデータセット別に COMET のモデルおよびリレーションを変更した場合における SarDeCK の F 値を示す．なお，生成されるコモンセンスは全て用いる (All)．行はリレーション，列は COMET の学習データセットと入力前処理を表す．各表の値において，アンダーラインは既存手法の最高 F 値を上回ったものであり，太字はその中で最高性能であるものを示す．

表 5 から，最高性能のモデルのコモンセンスは入力前処理 A の COMET-CN の Causes リレーションで生成したものであり，これは既存手法で用いているリレーションと同じである．表 4 の Tweets (Ghosh) における知識選択 All の結果よりも 3.33% の性能向上が示される．したがって，入力前処理 A の絵文字削除は，COMET への入力ノイズを低減し，モデルの性能改善に寄与している．また，表 4 の Tweets (Ghosh) における最高 F

表 5　Tweets (Ghosh) における SarDeCK の F 値（提案手法，知識選択 All のみ）

	CN-A	ATM-A	ATM^{20}_{20}-A	ATM^{20}_{20}-B
Causes	**85.70%**	-	82.89%	85.17%
CausesDesire	83.10%	-	83.41%	84.00%
Desires	83.99%	-	84.75%	84.14%
xEffect	-	83.67%	83.67%	85.08%
xReact	-	83.30%	82.79%	83.83%
xWant	-	85.1%	83.59%	82.74%

表 6　Tweets (Ptáček) における SarDeCK の F 値（提案手法，知識選択 All のみ）

	CN-A	ATM-A	ATM^{20}_{20}-A	ATM^{20}_{20}-B
Causes	83.89%	-	84.48%	84.78%
CausesDesire	83.97%	-	84.79%	84.34%
Desires	83.47%	-	83.92%	84.25%
xEffect	-	84.16%	83.10%	82.58%
xReact	-	83.50%	83.73%	83.65%
xWant	-	83.89%	**84.99%**	83.50%

表 7　Reddit-pol における SarDeCK の F 値（提案手法，知識選択 All のみ）

	CN-A	ATM-A	ATM^{20}_{20}-A	ATM^{20}_{20}-B
Causes	74.14%	-	74.28%	75.37%
CausesDesire	76.18%	-	75.17%	75.19%
Desires	75.71%	-	75.30%	73.37%
xEffect	-	75.74%	75.17	74.85%
xReact	-	75.49%	74.23%	75.33%
xWant	-	75.35%	74.99%	**76.36%**

表 8　各データセットの最高 F 値を示したコモンセンスに知識選択を行った場合の SarDeCK の F 値

	Tweets (Ghosh)	Tweets (Ptáček)	Reddit-pol
All	**85.70%**	**84.99%**	**76.36%**
Maj	81.34% ↓	83.37% ↓	75.36% ↓
Min	82.74% ↓	83.84% ↓	75.17% ↓
Con	82.72% ↓	83.80% ↓	75.85% ↓

値よりも 0.43%の性能向上が示される．

表 6 から，最高性能のモデルのコモンセンスは入力前処理 A の $ATOMIC^{20}_{20}$ の xWant リレーションで生成したものである．表 4 の Tweets (Ptáček) における知識選択 All の結果よりも 0.94%の性能向上が示される．また，表 4 の Tweets (Ptáček) における最高 F 値よりも 0.48%の性能向上が示される．

表 7 から，最高性能のモデルのコモンセンスは入力前処理 B の $ATOMIC^{20}_{20}$ の xWant リレーションで生成したものである．表 4 の Reddit-pol における知識選択 All の結果よりも 2.86%の性能向上が示される．また，表 4 の Reddit-pol における最高 F 値の結果よりも 0.87%の性能向上が示される．表 4 の Reddit-pol における知識選択 All の結果と表 7 の COMET-CN の入力前処理 A の Causes リレーションを比較すると，0.64%性能向上が示された．Reddit-pol には絵文字を含むテキストが存在しないことから，COMET への入力前処理における絵文字削除以外の工夫や，COMET の出力に対するノイズ削除が性能向上に寄与している．

既存手法の表 4 と提案手法の表 5，表 6，表 7 の比較から，利用するコモンセンスの種類によって，SarDeCK の性能が異なり，既存手法の性能を上回るものも複数存在した．知識選択 All を同一条件として既存手法と比較した場合，適切な種類のコモンセンス情報を利用することで最低 0.94%から最高 3.33%の性能が向上した．また，既存手法において最適な知識選択を施した最高性能と比較した場合においても，全ての皮肉検出データセットにおいて性能を上回った．また，COMET への入力テキストの処理およびコモンセンス生成後の処理の工夫によって，性能が改善されることが示された．しかし，皮肉検出データセットによって最適なコモンセンスの種類が異なり，既存の最高性能を上回るコモンセンスの種類に傾向は見られなかった．COMET の入力前処理は，表 5，表 6，表 7 のうち $ATOMIC^{20}_{20}$ における入力前処理 A と B での比較をすると，計 6 つのリレーションのうち Tweets (Ghosh) は 4 つ，Tweets (Ptáček) は 2 つ，Reddit-pol は 4 つのリレーションで入力前処理 B の方の性能が入力前処理 A の性能を上回った．この結果より，COMET の入力前処理でストップワード削除の有無によるモデル性能の傾向は皮肉検出データセットやリレーションによって異なるため，皮肉検出への性能向上への効果は得られなかった．

表 8 では，提案手法から得られた各データセットにおける最高性能を得られたコモンセンスを利用して，明示的な知識選択をした場合の SarDeCK の性能を示している．具体的には，Tweets (Ghosh) は入力前処理 A における COMET-CN の Causes リレーション，Tweets (Ptáček) は入力前処理 A における $COMET\text{-}ATOMIC^{20}_{20}$ の xWant リレーション，Reddit-pol は入力前処理 B における $COMET\text{-}ATOMIC^{20}_{20}$ の xWant リレーションのコモンセンスを利用する．

表 4 の既存手法の結果とは異なり，全てのデータセットにおいて，知識選択 All と比べて性能が低下する．この結果は，知識選択 Maj，Min，Con によって，皮肉検出に必要なコモンセンス情報が除かれたことが原因であると考えられる．既存手法では，コモンセンス情報に皮肉検出に対するノイズが多く，明示的な知識選択が有効となる．したがって，適切なコモンセンスの種類を用いることで，コモンセンス生成の段階でノイズを削減することが実現でき，コモンセンスの選別が重要であることが示された．

5.4　実験結果と考察（複数リレーション）

これまで，単一のリレーションのコモンセンスの利用による実験を行い，その考察を述べた．次に，複数のリレーションのコモンセンスを利用することによる皮肉検出の性能の影響を確かめる．

今回は $COMET\text{-}ATM^{20}_{20}$，入力前処理 A の xEffect，xReact，xWant の 3 つのリレーションからそれぞれ 5 個のコモンセン

表 9　COMET-ATM$^{20}_{20}$，入力前処理 A のコモンセンスのリレーションの単一利用と複数利用による SarDeCK の F 値の比較

		1Rel-5know	3Rel-15know	
Tweets (Ghosh)	xEffect	0.8367	0.8400	↑
	xReact	82.70%	84.00%	↑
	xWant	83.59%	84.00%	↑
Tweets (Ptáček)	xEffect	83.10%	85.03%	↑
	xReact	83.73%	85.03%	↑
	xWant	84.99%	85.03%	↑
Reddit-pol	xEffect	75.17%	75.86%	↑
	xReact	74.23%	75.86%	↑
	xWant	74.99%	75.86%	↑

スを生成する．その後，それぞれのコモンセンスを結合させたものを新しいコモンセンスとして利用する．コモンセンステキストのトークン最大長を 60 とする．

表 9 に実験結果を示す．単一のリレーションおよび複数のリレーションのコモンセンス，それぞれを利用した場合の SarDeCK の性能の比較を示している．結果より，どの皮肉検出データセットにおいても，単一のリレーションのコモンセンスを用いた場合よりも複数のリレーションのコモンセンスを用いた場合の方が，性能が高くなる．したがって，複数のリレーションのコモンセンスを組み合わせて使用することで，性能を改善することができる可能性を示した．

6 ま と め

本研究では，コモンセンスを用いた既存の皮肉検出モデルの性能向上のために生成するコモンセンスに着目し，様々な種類のコモンセンスを用いることに加え，コモンセンス生成モデルへの入力テキストやコモンセンス生成後の処理の工夫することによる効果を詳細に分析した．また，複数の種類のコモンセンスを連結して利用する方法も試した．実験結果より，利用するコモンセンスの種類によってモデル性能に差が生じ，皮肉検出データセットごとに適切なコモンセンスの種類が異なることが示された．また，適切な種類のコモンセンスを用いることで，皮肉検出に対してノイズとなるコモンセンス情報を明示的な知識選択を施すことなく削減でき，皮肉検出の性能向上に寄与することから，コモンセンスの選別が皮肉検出に重要であることが示された．

本研究を通して，適切なコモンセンスの種類の選別方法に課題が残った．本実験ではいくつかの知識の種類を手動で検証したが，これらは時間のかかるものであり，必要な知識情報を簡単に選定できる手法の検討が求められる．複数の種類のコモンセンスを利用した実験は，特定の 1 つの種類に断定せずに皮肉検出モデルの性能改善ができる可能性を示しており，本研究における課題を解決する方法の一つになるであろう．

今後の展望としては，更に性能向上が見込めるコモンセンスの種類を探索するため，本研究では実験を行わなかった種類のコモンセンスを利用してモデルの性能を比較する．また，複数のリレーションのコモンセンスを用いた実験も更に分析を進める．さらに，今回用いた皮肉検出モデル SarDeCK は，テキスト中の絵文字の意味を考慮しないモデルである．Okimoto ら [22] の研究では，皮肉検出における絵文字の重要性を説明している．そのため，絵文字を考慮したモデルのアーキテクチャへの変更も検討する．

謝　　辞

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

文　　献

[1] Stacy L. Ivanko and Penny M. Pexman. Context incongruity and irony processing. *Discourse Processes*, 35(3):241–279, 2003.

[2] Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 50(5):73:1–73:22, 2017.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[4] Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. Sarcasm detection with commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3192–3201, 2021.

[5] Zhe Yu, Di Jin, Xiaobao Wang, Yawen Li, Longbiao Wang, and Jianwu Dang. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 2423–2431, 2023.

[6] Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. Just like a human would, direct access to sarcasm augmented with potential result and reaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10172–10183, 2023.

[7] Tony Veale and Yanfen Hao. Detecting ironic intent in creative comparisons. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, volume 215, pages 765–770, 2010.

[8] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779, 2019.

[9] Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. R^3: Reverse, retrieve and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7976–7986, 2020.

[10] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451, 2017.

[11] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavtula, Nicholas Lourie, Hannah Rashkin, Brendan Roof,

Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035, 2019.

[12] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (COMET-)ATOMIC$^{20}_{20}$: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392, 2021.

[13] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, , and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825, pages 722–735, 2007.

[14] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1352–1362, 2013.

[15] Lenat D. B. and Guha R. V. *Building large knowledge-based systems: representation and inference in the Cyc project*, pages 41–52. 1989.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for languange understanding. *CoRR*, abs/1810.04805, 2018.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Published from OpenAI, 2018.

[18] Snigdha Chaturvedi Somnath Basu Roy Chowdhury. Does commonsense help in detecting sarcasm? *CoRR*, 2021.

[19] Aniruddha Ghosh and Tony Veale. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 161–169, 2016.

[20] Tomáš Ptáček, Ivan Habernal, and Jun Hong. Sarcasm detection on czech and english twitter. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 213–223, 2014.

[21] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, 2018.

[22] Yoshio Okimoto, Kosuke Suwa, Jianwei Zhang, and Lin Li. Sarcasm detection for japanese text using bert and emoji. In *Database and Expert Systems Applications - 32nd International Conference, DEXA 2021, Virtual Event, September 27-30, 2021, Proceedings, Part I*, volume 12923, pages 119–124, 2021.

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

2024年2月28日(水) 16:00 ~ 18:10 | T2-A オンライン（Zoom Events)

並列処理・分散データベース

座長:合田 和生(東京大学)
コメンテータ:金政 泰彦(富士通株式会社)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T2-A-3-01] データ検証可能な分散DB実現手法の提案と実現に向けたHyperledger Irohaのコマンド開発

*堀 遥[1]、小口 正人[1] (1. お茶の水女子大学)

16:30 ~ 16:55
[T2-A-3-02] SparkSQLによる効率的な問合せ処理のためのワークロードに基づくRDFデータ分割手法

*山﨑 昂輔[1]、天笠 俊之[2] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群情報理工学位プログラム 知識・データ工学研究室、2. 筑波大学計算科学研究センター)

16:55 ~ 17:20
[T2-A-3-03] Fast Parallel Computation for Random Walk based t-SNE

*三木 武志[1]、藤原 靖宏[2]、川島 英之[1] (1. 慶應義塾大学、2. 日本電信電話株式会社)

17:20 ~ 17:45
[T2-A-3-04] ブロックチェーン技術を活用したデータ検証可能な分散データベースの性能に関する検討

*坂本 明穂[1]、小口 正人[2] (1. お茶の水女子大学理学部情報科学科小口研究室、2. お茶の水女子大学基幹研究院自然科学系)

17:45 ~ 18:10
[T2-A-3-05] PostgreSQLの外部連携機能におけるレコード変換処理の高速化

*柿村 直拓[1]、立床 雅司[1]、柴田 秀哉[1] (1. 三菱電機株式会社 情報技術総合研究所 データマネジメント技術部 データ最適化技術グループ)

DEIM Forum 2024 XX-Y

データ検証可能な分散DB実現手法の提案と実現に向けたHyperledger Irohaのコマンド開発

堀　遥† 小口　正人†

† お茶の水女子大学　〒112–8610 東京都文京区大塚 2–1–1
E-mail: †haruka-h@ogl.is.ocha.ac.jp, ††oguchi@is.ocha.ac.jp

あらまし　データ駆動型社会の実現にむけ，信頼度などに異種性を持つデータを一元的に取り扱う分散データベースシステムの開発が求められる．分散環境において，格納データの安全性を保証するためには，全てのデータが検証可能であることを保証すれば良い．すなわち，データ検証機能を有し，かつ，ユーザが安全性に欠けると判断されたデータを利用する際には，その利用可否を直接問う仕組みを実装することが要求される．我々はこのようなデータ管理基盤の開発を目指し，少ないリソースでの高速動作と高い開発の自由度が特徴のブロックチェーン基盤ソフトウェアのHyperledger Irohaをプラットフォームとした分散データベースシステムを検討した．また，このシステムの実装の一部として，検証対象となるデータの生成・格納処理を行うHyperledger Irohaの新たな組み込みコマンドの開発に着手した．

キーワード　クラウド基盤、ブロックチェーン技術、データ検証技術

1 はじめに

1.1 研究背景

近年，スマートフォンやIoTデバイスの普及から膨大なデータが収集されるようになった．DXの進展に伴うデータ駆動型社会の実現に向け，このような実社会の多様なデータをその不確実性を考慮しながら活用するような新しいデータ管理基盤の実現が期待されている．このとき，信頼度や来歴，種類などに異種性のあるデータを一元的に保存し，データ利用の際にはその信頼性の保証に貢献するような分散データベースシステムが考えられる．一方で，そのようなデータ管理基盤では，複数のユーザがアクセスできるオープンな環境において，これらのヘテロなデータの健全性・安全性の保証が課題となる．これに対し，収集・保管された全ての有効なデータが検証可能であることを保証する機能を分散データベースシステムに取り入れるというアプローチが考えられる．

1.2 既存研究

嘉戸らの研究[16]では，クラウドストレージシステムに保存されたデータの完全性検証へのアプローチとして，Private PDPを採用している．PDPはデータを複数のブロックに分割し，その中からランダムに選んだブロックに対してデータの完全性検証を行う．特にPrivate PDPでは，データ保有者とクラウドサービスプロバイダのみで完全性検証を行う方式である．Private PDPにブロックチェーン技術の一つであるスマートコントラクトを応用し，データの完全性検証および完全性検証結果不一致時の合意形成を行う方法を提案している．

また，Neha Mishraらの研究[7]では，PDV：Personal Data Vaultと呼ばれる個人の生涯にわたるデジタルドキュメントを，検証可能かつ安全な方法で保管，保存，保護，共有するためのフレームワークの開発にブロックチェーンプラットフォームHyperledgr Irohaを採用している．各文書は暗号化，圧縮され，クラウドに安全に保存され，文書の保存先URLがHyperledger Irohaに入力される．また，開発システムはMarkov Treeを用いた予測プリフェッチ機能を備えており，次に発生する要求を予測，事前実行する．

いずれの場合においても，現代で有用なデータ管理基盤において，データの安全性を保証する際にはブロックチェーンの活用することが効果的であることがわかる．

1.3 研究目的

本研究では，任意のデータが検証可能であることを保証する機能を持つ分散データベースシステムの開発することを目的とする．この目的を達成するため，まずメタデータをブロックチェーンプラットフォームHyperledger Irohaに付属するオンチェーンデータベースにて管理する．オフチェーンデータベースへのデータの追加や更新などが発生し，新規メタデータが必要となる場合，一連の処理をHyperledger Irohaの組み込みコマンドで行う．コマンドの結果はブロックチェーンに記録され，これによりメタデータの安全性が確保される．そして，安全性の保証されたメタデータ，蓄積されたブロック，Hyperledger Irohaのブロック検証機能を利用し，データの検証可能性を保証する．最終的には，データ検証の結果，安全性が保証されないと判断されたデータを利用する際には，ユーザに対象のデータの利用可否を問うシステムの実現を目指す．

1.4 実証実験

脱炭素を実現する循環型社会への貢献を期待し，製造業におけるカーボンフットプリント管理を対象とした実証実験を行う．

表 1　ブロックチェーンの適用事例

業界	企業名	利用目的
金融	三菱 UFJ ファイナンシャルグループ	独自のデジタル通貨管理 [14]
不動産業	積水ハウス株式会社	物件情報と不動産賃貸契約の実行 [15]
インフラ (電力取引)	国立大学法人東京大学 トヨタ自動車株式会社 TRENDE 株式会社	分散型電源を活用した電力の個人間売買システムを検証 [11]
医療	Z.com Cloud ブロックチェーン	医療カルテの共有 [5]
コンテンツ業	ソニー株式会社 株式会社ソニー・ミュージックエンタテイメント 株式会社ソニー・グローバルエデュケーション	デジタルコンテンツに関わる権利情報を処理する機能 [12]

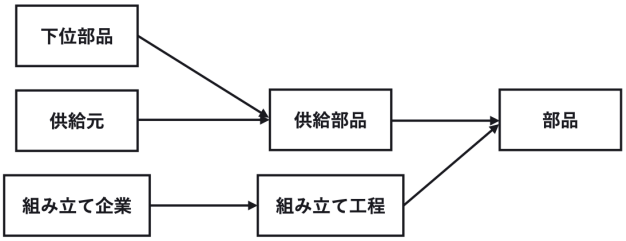


図 1　カーボンフットプリント応用のスキーマ

複数企業を接続する分散データベース上で製造工程全体のトレーサビリティを実現し，センサなどのクライアントデバイスから収集した CO_2 排出量データのメタデータを Hyperledger Iroha で管理する．これにより，収集した CO_2 排出量の検証可能性が保証され，カーボンフットプリント削減に向けた安全・安心なデータ分析・予測を支援する．

本稿では，自動車製造に関わる企業のうち部品製造・組み立て工場にフォーカスし，これらの工場間における製造時・組み立て時のカーボンフットプリント管理を想定する．図 1 は，カーボンフットプリント応用におけるスキーマであり，供給元から下位部品が供給され，組み立て工場による組み立て工程を経て，部品を構築する．本スキーマは各工場が管理しており，下位部品から上位部品を階層的に組み立てて，複数企業を横断して最上位部品の自動車を製造する．

一台の自動車には約 3 万点ほどの部品が使用されるとされている [9]．おおよその国内の自動車部品メーカーと呼ばれる企業数，過去 10 年間で発売された新車数，そしてこれらの重複などを考慮すると，1 車種あたり 100 社程度が扱う 3 万点もの部品を分散管理できるシステムが必要となることが考えられる．

ここで，スキーマに従い，部品の組み立て工場で排出された CO_2 排出量を *EMISSIONS* と呼ぶことにする．また，各部品は階層関係にあり，CO_2 排出量もまた階層関係にある．下位部品の *EMISSIONS* を再帰的に足し合わせ，部品自体の *EMISSIONS* と合計したものを *TotalEMISSIONS* と呼ぶ．本稿では各部品の *TotalEMISSIONS* を検証対象データであることを前提にシステムの実装手法を行う．

1.5　本稿の構成

本稿は以下の通り構成される．第 2 章では本研究の前提となるブロックチェーン及び Hyperledger Iroha の概要を説明する．第 3 章では本研究の提案手法の概要を，第 4 章で提案手法の実装に向けた計画を説明する．最後に第 5 章でまとめ，今後の課題を述べる．

2　関 連 研 究

2.1　ブロックチェーン

ブロックチェーンは，2008 年に Satoshi Nakamoto により投稿された論文 [8] に基づき，暗号通貨 Bitcoin [2] の公開取引台帳としての役割を果たすために発明された．Bitcoin の成功によりブロックチェーンは仮想通貨技術として強く注目されたが，一般社団法人日本ブロックチェーン協会は広義のブロックチェーンを「電子署名とハッシュポインタを使用し改竄検出が容易なデータ構造を持ち，且つ，当該データをネットワーク上に分散する多数のノードに保持させることで，高可用性及びデータ同一性等を実現する技術」と定義している [1]．

現在では仮想通貨以外での用途も促進されており，表 1 は書籍 [13] で紹介される例を一部抜粋したものである．例えば不動産業では，煩雑で膨大な量の手続きによって浪費されるコストと時間を，ブロックチェーン活用によって削減し，より効率的にさせている．また，医療では，ブロックチェーンに備わる権限機構と威厳性を活用し，医療機関ごとに管理されているカルテをブロックチェーンで一元的に管理して，より安全に利便性を高めるといった効果をもたらしている．

2.1.1　技術的概要

ブロックチェーンは，ブロックと呼ばれるトランザクションを記録する単位を生成し，これをチェーンのように連鎖するデータ構造である．仮想通貨の基礎技術として開発されたブロックチェーンの厳格性を担保する仕組みはチェーンにある．一つのブロックにはハッシュ値が付与される．このハッシュ値は，前のブロックのハッシュ値と新規のトランザクションの内容やタイムスタンプなどから算出される．すなわち，ハッシュ値によりブロックの関連が生まれ，これがチェーンとなる．よってブロックチェーンは高い改ざん耐性を持つと言われる．また，Peer to Peer 型のブロックチェーンネットワーク上に参加する各ノードがブロックチェーンを管理するるため，欠損ブロックを他のノードから補うことができ，耐障害性と可用性に優れるといった特徴を持つ．

表 2　ブロックチェーンと集中型データベースの比較

比較項目	ブロックチェーン	中央型データベース
A.　構築の信頼性	管理者が必要ない場合がある	中央管理者が必要
B.　データの機密性	全ノードがデータ参照可能	関係者のみにアクセスを制限
C.　堅牢性& 耐障害性	データはノード間で分散される	データは中央データベースで保管
D.　パフォーマンス	合意形成に時間を要す (Bitcoin では約 10 分)	即時に実行/アップデート
E.　冗長性	(デフォルトでは) 参加する各ノードが最新のコピーを保持	管理者のみがコピーを保持
F.　セキュリティ	(デフォルトでは) 暗号化メカニズム使用	従来のアクセスコントロール

2.1.2　従来のデータベース技術との比較

M. J. M. Chowdhury らの論文 [3] では，ブロックチェーンと集中型データベースの各アプローチを表 2 のように挙げている．また，これら 6 つの比較項目に対しする比較分析から，データの信頼性，堅牢性，実証性がシステムの優先事項であれば，ブロックチェーンがより優れており，機密性とパフォーマンスが優先であれば，従来のデータベースがより優れたソリューションであるとされている．

以上より，ブロックチェーンとデータベースの両方の機能を持つことで，アプリケーションは効率性と安全性を高めることができるため，ブロックチェーンプラットフォームの多くはデータベースと統合されている．

2.1.3　ブロックチェーンのモデル

ブロックチェーンは「パーミッションレス型」,「パーミッション型」,「コンソーシアム型」という 3 種類のモデルに分別される．表 3 はそれぞれのモデルの特徴である．

Bitcoin や Etherum [4] に代表されるパーミッションレス型は非中央集権でマイニングと呼ばれる膨大な計算の承認によって取引の正当性を担保する．透明性の高さを持つ反面，時間とリソースのコストが高く，データのプライバシーは保証されない．

パーミッション型は，中央集権型のネットワークで，透明性はないもののプライバシーが確保される上に，マイニングも行う必要がないため，高いパフォーマンスを持つ．こうした特徴から，プライベート型は単一の企業や組織内での運用に有効であり，特に銀行間の取引や証券取引での活用が促進されている．次の節で説明する Hyperledger の数々のプロジェクトはパーミッション型である．

最後のコンソーシアム型には複数の管理者が存在している．そのため，パーミッションレス型の分散性とパーミッション型の迅速な大量処理が可能という機能を備えている．また，管理者が複数存在しているため，運用ルールの変更についても一定数以上の合意が必要となる．これにより，高い公平性が保たれ，セキュリティや耐障害性もプライベート型に比べると強固であるとされる．こうした特徴から，コンソーシアム型は同業他社が協力して構築するブロックチェーンシステムに有効であり，こちらも銀行間の取引や証券取引での活用が促進されている．

2.1.4　コンセンサスアルゴリズム

ブロックチェーンを構成する技術の中に「コンセンサスアルゴリズム」がある．コンセンサスアルゴリズムとは，生成ブロックをブロックチェーンに追加する前に各ノード間で合意形成を行うメカニズムである．表 4 は代表的なコンセンサスアルゴリズムである．本研究で使用する Hyperledger Iroha では効率的に少ない Peer 数で承認を得る YAC 方式をとる．

表 3　ブロックチェーンの種類

名称	管理者	参加者
パーミッションレス型	なし	制限なし
パーミッション型	複数存在	個人や単一の組織
コンソーシアム型	複数存在	グループや複数の組織

表 4　ブロックチェーンの代表的なコンセンサスアルゴリズム

名称	採用システム	決定方法
PoW	Bitcoin Ethereum	正解を最も早く算出した Peer を採用
PoS	Ethereum で検討	資産の保有量に応じて 有利な条件で PoW を実施
PBFT	Hyperledger Fabric	検証を行った Peer の 過半数から承認を得る
YAC	Hyperledger Iroha	PBFT より効率的に少ない Peer 数で承認を得る

2.2　Hyperledger Iroha

Hyperledger [6] は 2015 年 12 月に The Linux Foundation よって開始され，ブロックチェーンベースの分散台帳をサポートするプロジェクトである．仮想通貨や金融，サプライチェーンに限らず，広範囲なビジネスシーンへのブロックチェーン基盤提供し，パフォーマンスや信頼性などの多方面においてサポート・改善ことを目指す．

Hyperledger では方向性の異なる複数のプロジェクトが推進されており，GA リリースされたプロジェクトは 2024 年 1 月時点で 4 つとなる．Hyperledger Iroha は 2019 年 5 月に GA リリースされたパーミッション型ブロックチェーンプラットフォームである．ソラミツ株式会社が初期開発者として継続的に開発に貢献している．

2.2.1　特　　徴

Hyperledger Iroha の主な特徴として「簡単な導入とメンテナンス」「開発に向けたさまざまなライブラリ」「役割に基づいたアクセス制限」「コマンドとクエリの分離によって行われるモジュール型設計」「資産とアイデンティティ管理」の 5 点が挙げられる．

- 簡単な導入とメンテナンス

Hyperledger Iroha の導入は非常に容易である．Docker 環境を利用したテスト環境であれば数十分で終了する．また，全ての Hyperledger プロジェクトはオープンソースとしてソースコー

ドが公開されている．

- 開発に向けたさまざまなライブラリ

Hyperledger Iroha は Java，JavaScript，Swift，Python の 4 つのプログラミング言語に対する API を備える．API は統一されておりどのプログラミング言語からでも同様の操作を実現できる．そのため，開発者のニーズに合わせた言語選択が可能となる．

- 役割に基づいたアクセス制限

仮想通貨では権限の概念が軽薄であり，これはネットワークにアクセス可能なユーザに対して全ての資源に対するアクセス権を付与する状況にある．Hyperledger Iroha では，ドメインと呼ばれる権限が及ぶ範囲を表す概念，ロールと呼ばれる複数の権限を 1 つに集約する概念が実装されている．これにより，広範囲な用途に対して，汎用的かつ柔軟的に対応することが可能である．

- コマンドとクエリの分離によって行われるモジュール型設計

Hyperledger Iroha の API 操作は，Hyperledger Iroha に対し変化をもたらすコマンドと，Hyperledger Iroha は普遍のまま現在持つ情報を表示するクエリの 2 種類に分けられる．コマンドの実行結果は，トランザクションとしてブロックチェーンに記録される．これらの 2 種類の API を組み合わせ，システムの開発が可能である．

- 資産とアイデンティティ管理

仮想通貨では通常，単一の仮想通貨を扱うが，Hyperledger Iroha では複数の通貨をアセットという概念で実装されている．アセットは作成時にドメインを指定する必要がある．

また，品質モデルにおいては次の 3 点が挙げられる．

- 信頼性

耐障害性，回復性．

- パフォーマンス効率

とりわけ時間挙動とリソースの使用効率．鍵生成アルゴリズムに安全性とパフォーマンスに優れた ED25519 を，コンセンサスアルゴリズムにより高速な YAC 方式を採用している点も効果的となっている．

- ユーザビリティ

学習可能性，ユーザエラー保護，妥当性の評価可能性．

以上の特徴を踏まえ，Hyperledger Iroha は Hyperledger プロジェクトの中でも，最も容易で短時間にブロックチェーンによる分散台帳を実現することができ，かつライトウェイトであるとされている．また，豊富な API や柔軟な概念構成から高い自由度の開発環境が提供されている．

2.2.2 ブロック検証機能

Hyperledger Iroha は起動時に全ブロックのハッシュ値の再計算を行い，ブロックを検証する．もし改ざんが発生した場合どのような挙動を示すか，ブロック欠損時，ブロックチェーン改ざん時について説明する．

まず過去ブロックの欠損時，起動時のハッシュ値再計算にてブロックの欠損を知らせるエラーメッセージが出力され，Hyperledger Iroha は起動しなくなる．複数 Peer 構成で運用する場合は，自動的に他の Peer から欠損ブロックが読み込まれる．

次にブロック情報の改ざん時，ブロック中のトランザクション処理内容であれば，ハッシュ値に変化が生じる．そのため，他の Peer とのハッシュ値の比較により改ざんが検知される．そしてブロックの作成者，すなわち電子署名が改ざんされれば，ブロック欠損時同様に Hyperledger Iroha は起動せず，改ざんが検知される．

本研究では，Hyperledger Iroha のブロック検証機能を利用し，データ検証メカニズムを実現することを検討している．

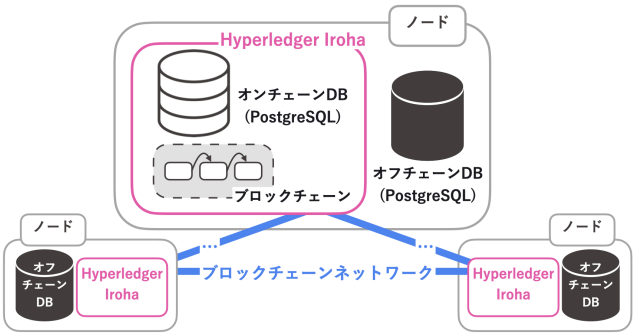


図 2　システム全体の構成

3 提案手法

本章では，提案手法を提案モデル，提案手法の手順，カーボンフットプリント応用における動作シナリオに分けて詳しく説明する．

3.1 提案モデル

まず本稿における提案モデルのシステム全体の構成を図 2 に示す．各ノードは Hyperledger Iroha とオフチェーンデータベースを備える．カーボンフットプリント応用の場合，ノードは企業と捉えてよく，全てのノードが Hyperledger Iroha によって構成されるブロックチェーンネットワークで接続される．オフチェーンデータベースと Hyperledger Iroha のオンチェーンデータベースはともに，オープンソースのリレーショナルデータベース管理システムである PostgreSQL [10] にて実装する．次の項で，各データベースの詳細を述べる．

3.1.1 オフチェーンデータベース

オフチェーンデータベースはノードに固有のデータベースである．主に各ノードで製造に関わっている部品のデータを管理する役割を持ち，検証対象データはオフチェーンデータベースに保管される．

表 5 は，オフチェーンデータベース上で管理するテーブルの例である．テーブル CO2Emissions は，カーボンフットプリント応用において現在想定している，検証対象データ *TotalEMISSIONS* を管理するテーブルであり，PartsID は部品の ID，部品の *EMISSIONS*，*TotalEMISSIONS*，時刻を表す TimeStamp を属性としてもつ．プライマリーキーとして PartsID と TimeStamp を指定している．ノードで新規 *EMISSIONS* が得られると，Hyperledger Iroha 上で *TotalEMISSIONS* が算出され，

表 5　テーブル CO2Emissions

CO2Emissions	
PK	PartsID
	EMISSIONS
	TotalEMISSIONS
PK	TimeStamp

表 6　テーブル Metadata

Metadata	
PK	PartsID
	Link
	ChildPartsID
	TimeStamp

これらの値がタイムスタンプとともにテーブル CO2Emissions に追加される．

上記の例以外にも，部品の基本情報や，ノードに関連する情報などもここで管理される．

3.1.2　オンチェーンデータベース

Hyperledger Iroha のオンチェーンデータベースは World State View とも呼ばれ，Hyperledger Iroha の最新情報が格納される．すなわち，ブロックチェーンネットワークに参加する全ノードでこの内容が同一となる．ここで，ブロックチェーン上のブロックそのものはオンチェーンデータベースには保存されず，json ファイルで生成されて別の場所で保管されている．

提案システムでは，各ノードのオフチェーンデータベースにて保管される全ての検証対象データに対し，メタデータを作成する．全部品のメタデータは，オンチェーンデータベース上に存在するテーブル Metadata に格納・管理される．表 6 は，カーボンフットプリント応用におけるテーブル Metadata である．PartsID は部品の ID，Link は表 5 上に格納される *TotalEMISSIONS* のリンク情報，ChildPartsID は下位部品の PartsID，TimeStamp は時刻を表す．プライマリーキーとして PartsID のみを指定しているため，新規メタデータ作成のたびにテーブル Metadata にデータを追加するのではなく，対象部品のカラムを更新する．1.4 節でも触れたように，開発システムで扱う部品数は約 3 万点と推測されるため，テーブル Metadata では全部品，おおよそ 3 万もの最新のメタデータが保管されることとなる．

このように，実データでなくメタデータをブロックチェーンで管理することで，実データのデータ形式に捉われず，また，その設計次第でデータのサイズを抑えることが可能となる．

3.2　提案手法の手順

本説では，提案手法の検証対象データの格納時，データの参照時の手順を説明する．その後，カーボンフットプリント応用における動作シナリオを説明する．

3.2.1　検証対象データの格納時

まず，検証対象データ格納時の手順を説明する．

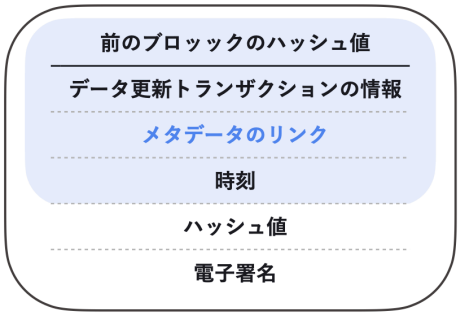


図 3　ブロック構成の概要

(1)　データ更新トランザクションがトリガーされる．
　　トリガーの条件は応用システムにより様々である．

(2)　データ更新トランザクションの実行．

　(2-1)　必要に応じたデータの加工処理などの実施．
　(2-2)　オフチェーンデータベースに格納．

(3)　(2-2) の格納場所情報 (Link) を含めて新規メタデータ作成．

(4)　(3) のメタデータをテーブル Metadata に格納．

(5)　ブロック作成，追加．

データ更新トランザクションでは，まず検証対象データ追加・更新のために必要な一連のデータ加工処理や演算を行う．これにより新しく得た検証対象データをオフチェーンデータベースに格納する．データ更新トランザクションの実装は，Hyperledger Iroha の組み込みコマンドとして開発を行う．

図 3 は (5) で作成，追加されるブロックの構成概要である．メタデータのリンクとは，テーブル Metadata のインデックスやプライマリーキーに該当するデータを表す．前のブロックのハッシュ値，更新トランザクションの情報，メタデータのリンク，時刻情報から，ブロックのハッシュ値は生成される．

以上の手順で検証対象データを格納することにより，メタデータの厳格性が担保される．

3.2.2　データ参照

データ参照時の手順は次の通りである．

(1)　メタデータを参照し，Link を取得．

(2)　(1) からオフチェーンデータベースにアクセスし，データを参照．

3.2.3　実証実験におけるデータ更新トランザクション

カーボンフットプリント応用では，*TotalEMISSIONS* を検証対象データとする．*TotalEMISSIONS* は，下位部品の *TotalEMISSIONS* の合計値と *EMISSIONS* を加算することで求まる．

本稿では，データ更新トランザクションのトリガー条件を新

規 *EMISSIONS* の取得とする．*EMISSIONS* の新しい値が得られた部品の *TotalEMISSIONS* を更新するため，データ更新トランザクションが実行される．データ更新トランザクションの手順は次の通りである．

(1) テーブル Metadata を参照し，対象部品の下位部品の PartsID を取得.

(2) テーブル Metadata で (1) の PartsID を検索し，Link を取得.

(3) Link からオフチェーンデータベース上の下位部品の *TotalEMISSIONS* を取得.
これを対象部品の全下位部品で行う.

(4) (3) で取得した *TotalEMISSIONS* らを合算する.

(5) (4) で算出した値に対象部品の新規 *EMISSIONS* を加算し，これが新規 *TotalEMISSIONS* となる.

(6) (5) をオフチェーンデータベースに格納.

データ更新トランザクションとその後のメタデータ生成・更新の一連の操作は，Hyperledger Iroha のコマンドにより実装する．また，ブロックに含まれるメタデータのリンクは常に対象部品の PartsID となる．

4 提案手法の実装に向けた計画

提案手法実装のため，第一にデータ検証トランザクションを実装する．具体的には，データ更新トランザクションの処理を提供する Hyperledger Iroha の組み込みコマンドの作成である．そのために Hyperledger Iroha の構成ファイルの開発が必要となる．

データ更新トランザクションを実現に向け，そのベースとなるコマンド *Update_testTable* の開発を行なう．開発環境，コマンド Insert の概要の二点について以降の節で説明する．

4.1 開 発 環 境

表 7 は実装環境の詳細である．仮想環境として Docker を使用し，Ubuntu22.04LST コンテナ上に Hyperledger Iroha の動作環境をビルドする．PostgreSQL コンテナでは，オンチェーン DB とオンチェーン DB としての機能を提供する．表 6 に示したテーブル Metadata は PostgreSQL コンテナ上で動作するデータベース上に存在する．

4.2 コマンド *Update_testTable* の概要

コマンド *Update_testTable* は，Hyperledger Iroha 上のオンチェーンデータベースに保管されるテーブル testTable に対して，UPDATE 命令を実行するコマンドである．表 8 にテーブル testTable の詳細を示す.

コマンド *Update_testTable* では，属性 Val に格納される値の更新を行う．実行の際には，プライマリーキーの ID と Val の更新値を与え，これらを持ってテーブル testTable に対して，UPDATE 命令を実行する.

表 7 実 装 環 境

コンテナ	Hyperledger Iroha ver.1	オン/オフチェーン DB
	Ubuntu イメージ：Ubuntu:22.04	PostgreSQL イメージ：postgres
仮想環境	Docker	
OS	Ubuntu20.04LTS	
サーバ	CPU：Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz メモリ：192GB	

表 8 テーブル testTable

testTable	
PK	ID
	Val
	TimeStamp

コマンド *Update_testTable* 開発おける目標として，以下の 3 点が挙げられる.

- コマンド実行結果がテーブル testTable に反映されている.
- コンセンサスアルゴリズムを経てブロック作成がアクセプトされている.
- 作成されたブロックにコマンド実行情報が記録されている.

5 まとめと今後の課題

本稿では，信頼度などに異種性を持つヘテロなデータを一元的に保管・利用するための分散データベースシステムの実現を目標とし，提案システムの検討を行った．このような，分散環境においてデータの安全性を保証するため，本稿では任意のデータに対するデータ検証機能をシステムに組み込むというアプローチを採用した．データ検証機能実現に向け，システムのプラットフォームにはブロックチェーン基盤ソフトウェアの Hyperledger Iroha を用い，Hyperledger Iroha にて検証対象データのメタデータを管理するためのメタデータ設計も行った．また，実装の一部として，Hyperledger Iroha の組み込みコマンドの開発に着手した.

今後は，第一にデータ更新トランザクション実装に向けた，コマンド *Update_testTable* の開発を行う．コマンド *Update_testTable* はオンチェーンデータベース内に閉じた処理であるため，これの開発が完了次第，Hyperledger Iroha からオフチェーンデータベースへのアクセス手法の検討・および実装に取り掛かる．その後は，提案手法におけるブロック構成と検証対象データ格納シナリオを順に実装する予定である.

また，Hyperledger Iroha のブロック検証機能の有用性についても調査を行なっていく.

謝　　辞

本研究は一部，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Japan Blockchain Association. ブロックチェーンの定義. https://jba-web.jp/news/642.
[2] Bitcoin. https://bitcoin.org/ja/.
[3] Mohammad Jabed Morshed Chowdhury, Alan Colman, Muhammad Ashad Kabir, Jun Han, and Paul Sarda. Blockchain versus database: A critical analysis. In *2018 IEEE TrustCom/BigDataSE*, pp. 1348–1353, 2018.
[4] Ethereum. https://ethereum.org/ja/.
[5] GMO インターネット株式会社. 医療機関カルテ共有システム. https://www.gmo.jp/news/article/5736/.
[6] Hyperledger foundation. https://www.hyperledger.org/.
[7] Neha Mishra and Dr. Haim Levkowitz. Pdv: Permissioned blockchain based personal data vault using predictive prefetching. In *BIOTC '21: Proceedings of the 2021 3rd Blockchain and Internet of Things Conference*, pp. 59–69.
[8] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.
[9] Toyota クルマ教室. https://global.toyota/jp/kids/faq/parts/001.html.
[10] Postgresql: The world's most advanced open source relational database. https://www.postgresql.org/.
[11] TOYOTA 自動車株式会社. 東京大学、トヨタ、trende が、次世代電力システムの共同実証実験を開始. https://global.toyota/jp/newsroom/corporate/28227543.html.
[12] ソニー株式会社, 株式会社ソニー・ミュージックエンタテインメント, 株式会社ソニー・グローバルエデュケーション. ブロックチェーン基盤を活用したデジタルコンテンツの権利情報処理システムを開発. https://www.sony.com/ja/SonyInfo/News/Press/201810/18-1015/.
[13] 佐藤栄一. Hyperledger Iroha 入門 -ブロックチェーンの導入と運用管理-. 株式会社 オーム社, 2020.
[14] 三菱 UFJ フィナンシャル・グループ. Mufg におけるブロックチェーンの取組み. https://www.boj.or.jp/paym/fintech/data/rel180214a6.pdf.
[15] 積水ハウス株式会社. 積水ハウス 業界の新たなスタンダード構築へ 業界初 ブロックチェーンで賃貸入居の煩雑なプロセスをワンストップ化. https://www.sekisuihouse.co.jp/library/company/topics/datail/__icsFiles/afieldfile/2020/06/08/20200608.pdf.
[16] 嘉戸裕一, 廣友雅徳, 瀧田愼, 掛井将平, 白石善明, 毛利公美, 森井昌克. ブロックチェーンを用いたクラウドストレージのプライベートな完全性検証方式. In *Computer Security Symposium 2023*, pp. 727–734, 2023.

Spark SQLによる効率的な問合せ処理のためのワークロードに基づくRDFデータ分割手法

山崎　昂輔†　　天笠　俊之††

† 筑波大学院 理工情報生命学術院 システム情報工学研究群 情報理工学位プログラム　〒305-8573 茨城県つくば市天王台1丁目1-1

†† 筑波大学 計算科学研究センター　〒305–8577 茨城県つくば市天王台1丁目1-1

E-mail: †kosuke.y@kde.cs.tsukuba.ac.jp, ††amagasa@cs.tsukuba.ac.jp

あらまし　近年，機械可読なデータとしてRDF（Resource Description Framework）が注目されている．RDFは主語，述語，目的語のトリプルの組み合わせで多くのデータの表現を可能とする単純で柔軟なフレームワークであり，さまざまな領域でRDFデータが生成され，そのデータ量は急速に増加している．膨大なRDFデータを効率的に処理するために多くの研究が行われており，その中でも並列分散処理による高速化が期待されている．Apache Sparkは，インメモリでデータを処理することで高速化を実現するフレームワークであり，Spark SQLを用いることでSQLを用いたRDFデータの処理が可能となる． 本研究では，ワークロードの共起性を用いたデータ分割手法と，Spark SQLを用いたRDFデータの効率的な問合せ処理を提案する．ワークロード情報を活用することにより問合せ時の関連が強いデータを同一パーティションに配置することが期待でき，処理対象のデータ量の削減を実現できる．実行時間の比較では，多くのクエリで従来手法を上回る結果を示した．

キーワード　RDF, パーティショニング, 分散処理, Apache Spark

1　はじめに

RDF（Resource Description Framework）[1]は，主語，述語，目的語からなるトリプルと呼ばれる基本構造を持ち，これによってデータ間の関係を簡潔に表現できるフレームワークである．このフレームワークはデータへのセマンティクスの付与やグラフ表現が可能なこと，機械可読であるなどの特徴があり，その有用性から近年RDF形式のデータが爆発的に増加している．RDFデータは基本的にSPARQL（SPARQL Protocol and RDF Query Language）という標準的な問合せ言語を用いてデータの問合せを行う．SPARQLでは，問合せたいデータをトリプル構造に基づいて記述することで問合せを実行することができる．問合せにかかるコストはデータ量の増加に伴って大きくなるが，一般に大規模なRDFデータは数百万から数億のトリプルを含んでいる．従って，そのようなデータに対していかに効率的で高速な問合せ処理を実現するかが昨今の課題となっている．

この課題を解決するために，大きく分けて集中型システムと分散型システムの2種類が多数発表されている．集中型システムは，データの保管や問合せ処理を1つのマシンで管理するものであり，Jena [2] やRDF-3X [3] などが挙げられる．一般に集中型システムは通信コストが低いという利点がある一方で，処理効率は単一マシンの性能に依存するためスケーラビリティが低く大規模データに対応することが難しい．一方，分散型システムは，データの処理を複数のマシンで行うため通信のオーバーヘッドが発生するという欠点を持つものの集中型システムの課題であった大規模データへの効率的な処理が可能であり，この特徴に注目した研究が昨今多く検討されている [4], [5].

Apache Spark [6] は大規模並列分散処理を行うフレームワークの1つであり，インメモリデータ構造を使用したデータ処理が可能であることから，分散型システムに活用した研究が増えてきている．SPARQLGX [7]，WORQ [8]，DIAERESIS [9] などがその例であり，これらはデータへの問合せ処理高速化に向けたデータ分割手法を提案している．

データ分割は分散並列処理技術を用いた処理における重要な課題の一つであり，そのデータレイアウトは問合せ時に必要な情報を取得するためにアクセスすべきデータ量を大きく左右する．単純な，あるいはランダムなデータ分割では問合せ時に不要なデータへのアクセスや探索が発生し，それに応じて処理コストが増加し問合せ応答の性能低下へとつながる．

ところで，多くのシステムではユーザが発行する典型的な問合せをワークロード情報として取得可能である．この情報を活用することで，問合せ処理を行う上で関連性の高いデータを同じパーティションに配置でき，データアクセスの抑制とそれによる性能の向上が実現できると期待される．

そこで本研究では，ワークロード中の主語，述語，目的語それぞれの共起性を利用したデータ分割とデータのインデックスを利用した効率的な問合せ処理を実現する手法を提案する．実際のワークロードを基にした共起性の導入により，同時に問合せられるデータを同一のパーティションにまとめられると期待できる．さらに，インデックスを用いることで問合せ時に必要なパーティションのみを選択して読み込むことを可能にし，処理するデータ量を抑制できると考えられる．

我々の提案手法の優位性を確かめるために，評価実験ではデータ分割時間，問合せ実行時間，問合せ時のデータ読み込み時間，読み込んだトリプル数，データ読み込み時間と問合せ実行時間の合計時間の 5 つの観点で既存手法と比較する．

本論文の構成は以下のとおりである．2 章では本研究の前提知識を述べ，3 章で関連研究を紹介する．4 章で提案手法を説明し，5 章で評価実験の結果を示す．最後に 6 章で本研究のまとめと今後の展望を述べる．

2 前提知識

本章では，本研究の基本事項を紹介する．まず 2.1 節で，本研究で扱うデータの対象である RDF について触れ，続く 2.2 節で RDF への標準的な問合せ言語である SPARQL に焦点を当てる．最後に，2.3 節で本研究で並列分散処理のために用いる Apache Spark について説明する．

2.1 RDF（Resource Description Framework）

2.1.1 RDF の基本知識

RDF（Resource Description Framework）は，データを主語 s，述語 p，目的語 o のトリプルと呼ばれる形式を基本パターンとして表現するフレームワークであり，トリプルを組み合わせることで複雑なデータの関係性を簡潔に表現することができる．各データは IRI（International Resource Identifier）を用いて示されており，あらゆる資源を一意に識別し，複数のデータベースにまたがるデータを統合することも可能である．また，リテラルやブランクノードも重要である．リテラルは文字列や数などの具体的なデータ値をそのまま扱うものであり，目的語としてのみ使用可能である．ブランクノードは IRI を持たない一時的なエンティティの表現などに使用されるものであり，主語と目的語としてのみ使用可能である．

以下はトリプルの例である．ただし，⟨_:Brank⟩ は空白ノード，⟨4.20000⟩ は数のリテラルによる表現である．

⟨dbr:University_of_Tsukuba⟩⟨dbo:city⟩⟨dbr:Tsukuba⟩

⟨dbr:Tsukuba⟩⟨db:Temperature⟩⟨_:Brank⟩

⟨_:Brank⟩⟨dbp:febMeanC⟩⟨4.200000⟩

2.1.2 RDF の視覚化

さらに，RDF には視覚化やそれを組み合わせたグラフ表現が可能であるという特徴もある．主語と目的語をノード，述語を有向エッジとすることで視覚化でき，それらを複数組み合わせることでグラフが形成できる．これは複雑なデータ間のネットワークの直感的理解や，グラフアプローチによる処理などに役立つ．図 1 はトリプルの基本パターンを視覚化したものであり，筑波大学がつくば市にあることを示している．そして図 **??** は複数の基本パターンを組み合わせてグラフにしたものである．なお，ここでは楕円が IRI ノード，長方形がリテラル，菱形がブランクノードを表している．

2.1.3 RDF の基本構文

RDF データの表現には，Turtle（Terse RDF Triple Language）や N-Triples などの構文が一般に用いられる．Turtle は，RDF データを簡潔かつ読みやすい形式で書くための構文であり，主に人間が読み書きするシナリオに適している．さらに，図 1 や図 2 で用いているように dbr や dbo などの接頭辞を使用して IRI を短縮し，データの表現を簡略化することが可能である．一方，N-Triples はよりシンプルな形式で，各トリプルを一行で表現する．これは機械による解析などに適しており，その単純さから大規模な RDF データの処理に用いられることも多い．本研究でも使用する RDF データは N-Triples 形式のものである．

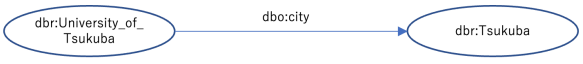


図 1: トリプルの視覚化の例

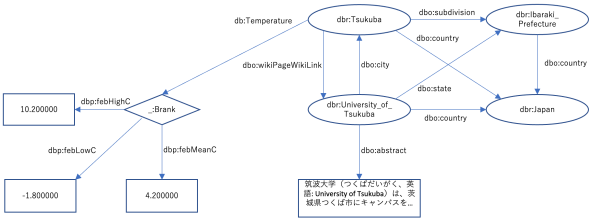


図 2: RDF のグラフ表現の例

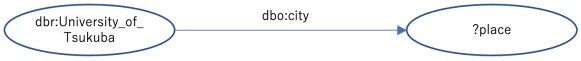


図 3: SPARQL クエリのグラフ表現の例

2.2 SPARQL（SPARQL Protocol and RDF Query Language）

SPARQL（SPARQL Protocol and RDF Query Language）は，RDF データへの標準的な問合せ言語である．基本パターンはリレーショナルデータベースへの標準的な問合せ言語である SQL と類似しており，SELECT 句と WHERE 句から構成される．SELECT 句は，問合せの結果として取得したい情報に対応する変数を列挙し，WHERE 句では問合せたい要素を”?”で始まる変数としたトリプルを記述する．例えば，図 2 の RDF グラフに対して筑波大学がある市町村を問合せたい場合，以下のように記述する．

```
SELECT ?place
WHERE{
  dbr:University_of_Tsukuba dbo:city ?place .
}
```

さらに，SPARQL は RDF と同様にグラフ表現が可能であり，問合せは主にグラフパターンマッチングにより行われる．例示した SPARQL クエリは，図 3 のように視覚化できる．また，これらの問合せ処理は SPARQL Endpoint と呼ばれるインターフェースを介して行われる．

2.3 Apache Spark

2.3.1 Apache Spark の概要

Apache Spark [6] は，大規模データの高速処理を目的とした分散処理フレームワークである．このシステムはスケールアウトアーキテクチャに基づいており，サーバー台数の増加に伴って処理能力を向上させることができる．この特徴により，ビッグデータの急激な増加に対して柔軟かつ効率的に対応することが可能になっている．

Spark 登場前には，Google が開発した GFS（Google File System）[10] と Google MapReduce [11] を基にしたオープンソースである Hadoop [12] が大規模データ処理分野の主流であったが，Hadoop はストレージ管理やデータ処理においていくつかの課題を抱えていた．Spark はこれらの課題を克服し，さらに RAM 上でのデータ処理を行うことで MapReduce よりも高速な処理を実現している．

また，Spark には構造化データや半構造化データの処理を容易にするための豊富なデータアクセスモデルと API が提供されている．これらを利用することで，様々なデータフォーマットを柔軟に扱いながら高速な並列データ処理を実現できる．

2.3.2 DataFrame

DataFrame は Spark で提供されている API の一つであり，大規模構造化データのための表形式データモデルとして分散処理に特化している．各列には名前とデータ型を保持することができ，SQL のような操作を容易にする設計となっている．また，JSON, CSV, Parquet など様々なデータフォーマットとの互換性があり，多くのデータソースからの読み込みが可能である．

2.3.3 Spark SQL

Spark SQL も Spark で提供されている API の一つであり，構造化データの処理に特化している．ユーザは SQL クエリを用いて DataFrame のデータを処理することが可能であり，様々なデータセットに対して統合的な解析操作を実現する．ただし，SPARQL には対応していないため，本研究では SPARQL から SQL への変換処理を行ったのち，DataFrame で読み込んだ RDF データに対して Spark SQL による問合せ処理を実行した．

3 関 連 研 究

本章では，本研究の関連研究を紹介する．3.1 節では，ワークロードを利用したデータ分割を行っている手法を説明し，3.2 節ではワークロードを使用しない手法として媒介中心性とインデックスを利用した手法を説明する．

3.1 ワークロードを用いた手法

ワークロードを利用したデータ分割手法として，Adnan らの研究 [13] がある．この研究では，SPARQL クエリのワークロードから共起している述語をカウントし，貪欲なクラスタリングを適用することで同時に問合せられることが多い述語ペアが同じクラスになるようにデータを分割する．Adnan らは，分割したデータを複数の SPARQL Endpoint 上に一つずつ格納し，フェデレーションエンジンを利用して処理しており，Spark 処理系を対象としている本研究とはこの点で異なっている．

Madkour らの研究 [8] もワークロードに基づいたデータ分割手法を提案している．このアプローチでは，複数クエリで頻繁に繰り返される結合パターンに着目し，ディスク I/O やネットワークシャッフリングのオーバーヘッドを最小化することを目指している．そのために，複数のテーブル間の結合が必要なクエリに対して，ブルームフィルタによる要素の削減を行う．そして，得られたデータに対して主語もしくは目的語に基づいて垂直分割を行い，その結果をキャッシュすることで処理対象のデータ量を抑えている．

Guo らの研究 [14] では，ワークロードクエリのグラフパターンに着目し，各サブグラフから頻出パターンを抽出することでそれに基づいたデータレイアウトを構築している．データ分割は動的に行われ，一定数のワークロードクエリが入力されるたびにデータレイアウトの再構築とそれに伴う移動コストを計算し，それを基に実際に再配置する．なお，この手法も Spark 系を対象とはしていない．

3.2 ワークロードを用いない手法

Georgia らの研究 [9] では，二段階のデータ分割を行う．第一段階では，スキーマグラフに対して媒介中心性を計算し，中心性が高いノードを重要なノードとして特定する．その後，残りのノードに対してスキーマノード同士の依存関係を計算する Dependence を定義し，重要なノードへ割り当てることでデータの分割を行う．第一段階の分割で得られた結果に対して，第二段階では述語ベースで垂直方向のサブパーティショニングを行う．各垂直パーティションには，一つの述語に対する主語と目的語が含まれており，各パーティションのデータサイズを小さくしている．さらに，これらのパーティションに対してクエリ実行時に必要なサブパーティションを直接探せるようにインデックスを生成している．この二段階の分割とインデックスによって，述語が束縛されていないクエリに対してその述語を含む第一段階のパーティションに含まれるデータすべてにアクセスする必要がなくなり，必要なパーティションのみのロードを可能にしている．

4 提 案 手 法

本章では，提案手法について説明する．本研究は大規模な RDF データに対する問合せを高速に行うことを目的としている．これを実現するために，ワークロード中の主語，述語，目的語それぞれに対する共起性を利用した二段階のデータ分割と，それぞれに対するデータへのインデックスを利用する手法を提案する．4.1 節では提案手法の概要を導入し，4.2 節ではデータ分割処理について，4.3 節では問合せ処理についてそれぞれ説明する．

4.1 提案手法の概要

本手法は SPARQL クエリのワークロード情報が与えられることを前提としており，ワークロードクエリ中のトリプルにお

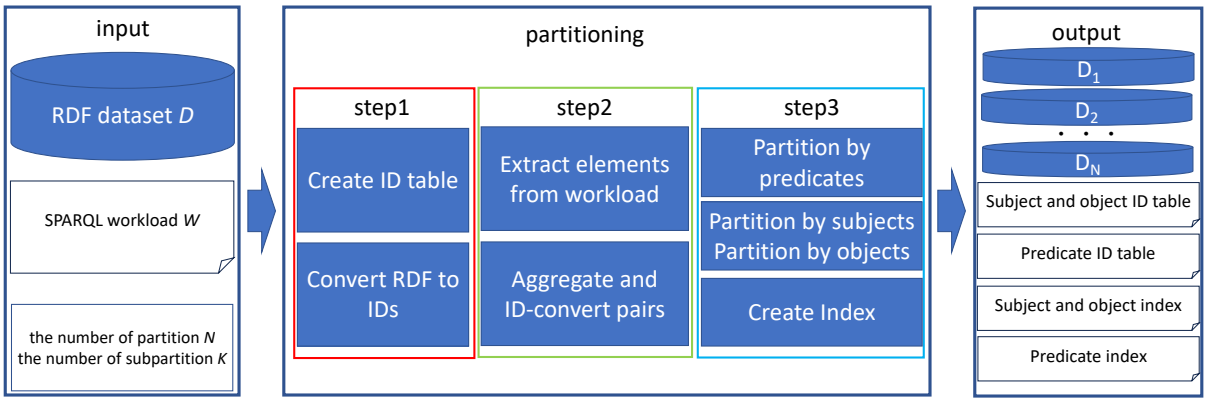


図 4: 提案手法のデータ分割処理の流れ

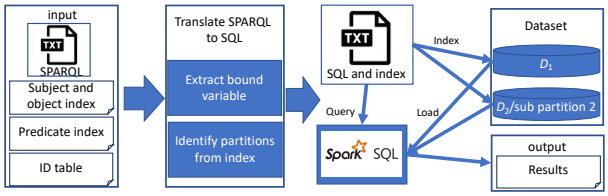


図 5: 提案手法の問合せ処理の流れ

いて共起性の高いトリプルを同一パーティションに配置することで SQL での効率的なクエリ処理を実現するという考えに基づく．図 4 は，提案手法のデータ分割処理の概要を示している．入力は，1) トリプルの集合からなる RDF データ D，2) ワークロードクエリ $W = q_1, \ldots, q_{|W|}$，3) パーティション数 N，およびサブパーティション数 K である．出力は RDF データのパーティション $D_1, \ldots, D_N$ （$D_1 \cap \cdots \cap D_N = \emptyset$），データと ID との関係を記録した ID テーブル，および各データのパーティションへのパスを記録するインデックスである．データ分割処理は大きく三つのステップで構成される．ステップ 1 では，データの冗長性を軽減するための ID 化を行う．ステップ 2 では，ワークロードから主語，述語，目的語をそれぞれ抽出し，共起する要素を ID に変換してペアを作成する．その後，ペアの出現回数を測定し，出現頻度の順にソートする．ステップ 3 では，述語の共起回数を基にしたクラスタリングによる第一段階の分割を行い，その結果に対して第二段階の分割として主語の共起回数に基づくクラスタリングと目的語の共起回数に基づくクラスタリングの二種類を行い，データを分割して出力する．同時に，主語，述語，目的語それぞれのパーティションへのインデックスも出力する．

図 5 は，提案手法の問合せ処理の概要を示している．問合せ処理は大きく二つのステップで構成される．ステップ 1 では，SPARQL の SQL への変換処理を行う．これは，通常 RDF データへの問合せは SPARQL によって行うが，Spark SQL は SPARQL に対応していないためである．また，その際 ID テーブルを利用した束縛変数の ID 化とインデックスを利用したパーティションの特定を同時に行う．ステップ 2 では，実際に SQL による問合せ処理を行う．変換されたクエリを受け取ると，ステップ 1 で特定していたパーティションを読み込

Algorithm 1 Workload Analysis.

Require: SPARQL workload W,Subject-object ID table ID_{so}, Predicate ID table ID_p
Ensure: Subject, predicate, and object co-occurrence counts denoted as $C_{sub}, C_{pre}, and C_{obj}$

1: Load ID_{so} and ID_p as HashMap
2: $L_{sub} := \emptyset; L_{pre} := \emptyset; L_{obj} := \emptyset$
3: **for all** $query \leftarrow W$ **do**
4: $where$ = extractInWhere($query$) ▷ Extract contents inside WHERE clause
5: **for all** $(s,t) \leftarrow$ extractBoundSubject($where$) **do** ▷ Create subject pairs from WHERE clause
6: $L_{sub} \leftarrow (\mathrm{ID}_{so}(s), \mathrm{ID}_{so}(t))$
7: **end for**
8: **for all** $(s,t) \leftarrow$ extractBoundPredicate($where$) **do** ▷ Create predicate pairs from WHERE clause
9: $L_{pre} \leftarrow (\mathrm{ID}_p(s), \mathrm{ID}_p(t))$
10: **end for**
11: **for all** $(s,t) \leftarrow$ extractBoundObject($where$) **do** ▷ Create object pairs from WHERE clause
12: $L_{obj} \leftarrow (\mathrm{ID}_{so}(s), \mathrm{ID}_{so}(t))$
13: **end for**
14: **end for**
15: Convert all elements in L_{sub}, L_{pre}, and L_{obj} into ID using HashMap
16: C_{sub} =countAllPairsByMapReduce(L_{sub})
17: C_{pre} =countAllPairsByMapReduce(L_{pre})
18: C_{obj} =countAllPairsByMapReduce(L_{obj})
19: Output C_{sub}, C_{pre}, and C_{obj}

み，DataFrame によりテーブルを作成し，それに対して Spark SQL による問合せを行うことで結果を取得する．

4.2 データ分割処理

4.2.1 Step1: ID の割り当て

RDF データ中の IRI やリテラルは一般に長い文字列であり，問合せ処理における文字列マッチングのコストが高くなる．したがって，これらを事前に唯一性が保証された整数値の ID に変換しておくことで，冗長性を軽減する．

4.2.2 Step2: ワークロードの分析

続いて，ワークロードクエリを分析して主語，述語，目的語それぞれに対して共起しているペアを集計する．入力は SPARQL クエリのワークロード W と，ステップ 1 で作成したデータと ID の対応テーブルであり，出力は各ペアとその共起回数をまとめたテキストである．Algorithm 1 はこのステップのアルゴリズムを示している．まず，与えられたワークロード $W = q_1, \ldots, q_{|W|}$ に対して `WHERE` 句内を抽出し，そこから主語，述語，目的語それぞれの束縛変数をすべて取り出す．各束縛変数を ID に変換し，対応する共起リスト $L_{sub}, L_{pre}, Lobj$ に追加する (2-14 行目)．このリストの要素はタプル $p = \langle P_1, P_2 \rangle$ であり，ここから共起回数を集計する．

例えば，図 6(左) の 4 つのクエリに対して述語の共起回数

```
SELECT * WHERE          SELECT * WHERE          SELECT * WHERE          SELECT * WHERE
{                       {                       {                       {
 ?s  p1  o1 .            s1  p1  ?o .            ?s  p1  o1 .            ?s  p1  ?o .
 ?s  p2  o2              ?o  p2  ?o2             ?s  p3  o3              ?s  p2  ?o .
}                       }                       }                       ?s  p3  ?o .
                                                                        ?s  p4  ?o .
                                                                        }
```

P1	P2	count
p1	p2	3
p1	p3	2
p1	p4	1
p2	p3	1
p2	p4	1
p3	p4	1

図 6: クエリの例 (左) とそれに対する述語の共起回数の例 (右)

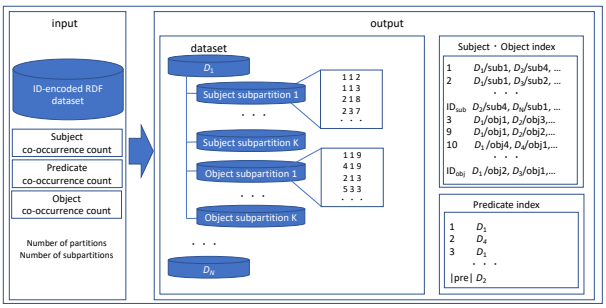


図 7: Step3 のデータ構成

を数える場合，$L_{pre} = \langle\langle p1, p2\rangle, \ldots, \langle p2, p4\rangle, \langle p3, p4\rangle\rangle$ となる．そして，これに対して集計処理を施すことで図 6(左) の結果が得られる．

4.2.3 Step3: データ分割

最後に，ステップ 2 で集計した情報を基にデータセットを分割する．図 7 は，ステップ 3 で入出力されるデータの具体的な構成を示したものである．

提案手法では，データは二段階に分割する．クラスタリングの手法は両段階で同じであり，そのアルゴリズムは Algorithm 2 に示す通りである．このアルゴリズムでは，まず，各クラスタに属する述語を記録する cluster， 述語が属するクラスタを記録する Index の二つのハッシュマップを用意する．ただし，Index は最初は空である．また，述語の種類数をパーティション数 N で割った値を一つのクラスタのサイズとしてあらかじめ計算しておく（1-4 行目）．共起している述語ペアを取り出し，ペアの一方がクラスタに属していない場合はそれをもう一方と同じクラスタに追加した場合のクラスタに含まれる述語数と t を比較する．サイズが t 以下である場合には同じクラスタに追加し，Index を更新する（9-15 行目）．述語ペアの両方がまだクラスタに属していない場合には，クラスターに含まれる述語数が最も小さいクラスタに両方の述語を追加し，Index を更新する（17-22）．最後に，9-15 行目の処理で追加できなかった述語やワークロードに出現していなかった述語を，述語数が最小のクラスタに順に追加していく．

Algorithm 3 にステップ 3 全体のアルゴリズムを示す．Index_s と Index_o は ID をキー，その ID が含まれるパーティションの番号の集合を値とする HashMap である（1-2 行目）．第一段

Algorithm 2 Clustering(D_{ID}, Co-occurrences$elem$, N)

Require: ID-encoded RDF dataset DID, Co-occurrence counts Co-occurrences$_{elem}$, ▷ elem is either subject, predicate, or object Number of partitions N
Ensure: Partitioned datasets $D_1, \ldots, D_N$, Index
1: cluster = $((0, \emptyset), \ldots, (N-1, \emptyset))$: HashMap; ▷ Record elements contained in each cluster
2: Index: HashMap; ▷ Record the cluster to which an element belongs
3: t = |distinct($elem$)|/N ▷ The number of types of elem in one cluster
4: **for all** $(p_1, p_2) \leftarrow$Co-occurrences **do**
5: **if** p_1 in Index AND p_2 in Index **then**
6: continue
7: **end if**
8: **if** One of them already belongs to a cluster **then**
9: Let p_1 already belong to a cluster
10: clusterNo = Index(p_1)
11: **if** |cluster(clusterNo)| + 1 ≤ t **then** ▷ Compare the number of elem types after addition with t
12: cluster(clusterNo) ← p_2
13: Index(p_2) ← clusterNo
14: **end if**
15: **end if**
16: **if** Neither belongs to a cluster **then**
17: clusterNo = cluster with the fewest number of elem
18: cluster(clusterNo) ← p_1, p_2
19: Index(p_1) ← clusterNo
20: Index(p_2) ← clusterNo
21: **end if**
22: **end for**
23: **for all** $p \leftarrow$ remaining elements **do**
24: clusterNo = cluster with the fewest number of elem
25: cluster(clusterNo) ← p
26: Index$_p$(p) ← clusterNo
27: **end for**
28: Output Index, cluster

階では ID 化された RDF データ全体と述語の共起回数，パーティション数を clustering 関数の入力として述語の共起に基づいたクラスタリングを行う（3 行目）．続いて第二段階では，その結果得られたクラスタを入力データセットとして，主語，目的語それぞれに対して clustering 関数を適用してクラスタリングを行う．得られたパーティションは parquet 形式で逐次出力し，Index を更新する（4-15 行目）．すべてのデータの分割が完了すると，Index を出力して処理が終了する．

4.3 問合せ処理

4.3.1 SPARQL から SQL への変換

SPARQL は Spark SQL での処理ができないため，SPARQL が与えられた時，まず SQL への変換を行う．変換処理は以下の流れになる．

1. SPARQL クエリ Q から `WHERE` 句内のトリプルを抽出
2. 各トリプル（$q_1, \ldots q_Q$）に対して，それぞれ次の処理を行う．
   (a) トリプル q_i から束縛変数を抽出する．
   (b) インデックスを基に読み込むべきファイルへのパス

Algorithm 3 Data Partitioning Algorithm.

Require: ID-encoded RDF dataset D_{ID}, Subject co-occurrence counts Co-occurrences$_{sub}$, Predicate co-occurrence counts Co-occurrences$_{pre}$, Object co-occurrence counts Co-occurrences$_{obj}$, Number of partitions N, Number of sub-partitions K
Ensure: N partitions $D_1, \ldots, D_N$,
Subject index Index$_s$, Predicate index Index$_p$, Object index Index$_o$

```
Index_s := ((subject_1, ∅), ..., (subject_|subject|, ∅)): HashMap;
Index_o := ((object_1, ∅), ..., (object_|object|, ∅)): HashMap;
firstPartition, Index_p = clustering(D_ID, Co-occurrences_pre, N)
for all partition ← firstPartition do
    secondPartition, Index = clustering(partition, Co-occurrences_sub, K)
    Output secondPartition
    for all (e, p) ← Index do
        Index_s(e) ← p
    end for
    secondPartition, Index = clustering(partition, Co-occurrences_obj, K)
    Output secondPartition
    for all (e, p) ← Index do
        Index_o(e) ← p
    end for
end for
Output Index_s, Index_p, Index_o
```

を特定し，テーブル名を決定する．

(c) 決定したテーブル名を `FROM` 句に，束縛変数を `WHERE` 句に，非束縛変数を `SELECT` 句に指定し，その結果を `AS` 句によって table_i とする SQL サブクエリを作成する．

3. 得られた $|Q|$ 個の SQL サブクエリを `FROM` 句に指定し，SPARQL クエリの `SELECT` 句に指定されていた変数を SQL の変数に指定する．SPARQL クエリの `WHERE` 句内の非束縛変数の結合を基に SQL の `WHERE` 句内を作成し，得られた SQL と読み込むファイルのパス，テーブル名をまとめて出力する．

上記の手順 2(b) のテーブル名はファイルへのパスから指定され，パスが同じ場合常にテーブル名は同じになる．なお，重複したパスとテーブル名は 1 つにまとめられる．したがって，複数の SQL サブクエリに同じパスの組み合わせが存在する場合，それらすべてのサブクエリに対して 1 つのテーブルを作成するだけでよくなり，データの読み込み時間と読み込むデータ量の削減が期待できる．一方で，クエリの `WHERE` 句が `s1 ?p1 ?o1` や `s2 ?p2 ?o2` のようなパターンで構成されている場合，各サブクエリの結果は複数のファイルに分散している可能性がある．あるサブクエリパターンの結果がファイル 1, 2, 4 に，別のサブクエリパターンの結果がファイル 4, 5, 6 にある場合，変換プロセスではファイル 1, 2, 4 を 1 つのテーブルに，ファイル 4, 5, 6 を別のテーブルにグループ化する．これにより，ファイル 4 からのデータが重複する可能性があり，この場合は読み込むデータ量が増加することになる．さらに，データセット内のすべてのデータにインデックスを作成しているため，サブクエリの束縛変数にインデックスがない場合は一致するデータが存在しないと判断できる．評価実験における速度評価ではこのような場合，あらかじめ用意しておいた空のテーブルを読み込むファイルとして指定して問合せを行った．

5 評価実験

本章では，提案手法の性能を確認するために合成データセットと実世界データセットを 1 つずつ使用し，複数の指標によって従来手法との比較を行った結果について述べる．従来手法としては従来手法を上回る性能を示している DIAERESIS を使用した．その際のパーティション数は，DIAERESIS の論文中のものと同様に 4 に設定した．なお，提案手法のパーティション数は第一段階で 5, 第二段階で２０とした．また，ID 化の優位性を確認するために，提案手法で ID 化を行わない場合の性能も検証した．

5.1 実験環境

実験は，Apache Spark（３．３．１）を Spark Standalone モードで動作させた 1 台の物理マシン上で実施した．提案手法では基本的にドライバーメモリとエグゼキューターメモリに 32GB を割り当てた．ただし，ID 化を行う時のみ 50GB を割り当てている．従来手法は，ドライバーメモリとエグゼキューターメモリのいずれも 10GB を割り当てている．マシンのスペックは表 1 の通りである．さらに，AWS EC2 の EMR クラスター

表 1: 実行マシンのスペック

OS	Ubuntu 18.04.5 LTS
CPU	AMD EPYC 7502P 32-Core Processor
メモリ	128GB

(emr-7.0.0) でインスタンスタイプ m5.xlarge，インスタンスサイズ 10 での実験も行った．提案手法は scala(2.12.7) で実装し，sbt（1.8.0）でコンパイルした．比較手法は scala(2.12.10) で実装されており，JDK1.6 でコンパイルした．

5.2 データセット

本実験では以下の 2 つのデータセットを使用した．なお，AWS 上では，比較手法において LUBM を用いた際にエラーが発生したため SWDF のみ実験を実施している．

5.2.1 LUBM（The Lehigh University Benchmark）

LUBM（The Lehigh University Benchmark）[15] は大学の数をパラメータとした合成データ生成ツールであり，RDF データベースシステムのベンチマークとして広く利用されている．実験では，パラメータを 100 としたデータセットを作成しており，これはデータサイズ 2.37GB，トリプル数 13.8M で構成されている．ワークロードと性能評価には，ベンチマークから提供されたクエリから 13 個を使用した．

5.2.2 SWDF（The Semantic Web Dog Food）

SWDF（The Semantic Web Dog Food）は，人，論文，講演に関するセマンティックウェブ会議のメタデータを含む実世界の RDF データセットである．このデータセットはデータサイズが 50MB であり，304,583 トリプルが含まれている．ワークロードと性能評価には，比較手法である DIAERESIS で使用されていたクエリから 207 個を使用した．なお，このクエリは実際のクエリログに基づいて FEASIBLE ベンチマークジェネレータ [16] によって生成されたものである．

5.3 評価指標

本実験では，性能評価に以下の指標を導入した．なお，比較手法の実験の条件に合わせるために，SPARQL から SQL への

表 2: 前処理時間の比較（括弧内は ID 化時間（左）と分割時間（右））

Method	Processing Time	Data Size
LUBM (13.4M Triples)		
DIAERESIS	16.38 ± 0.30 min	762MB
Proposed Method (with ID)	28.31 ± 0.39 min (14.46 + 13.85)	199MB
Proposed Method (without ID)	38.44 ± 0.87 min	406MB
SWDF (30.4K Triples)		
DIAERESIS	1.80 ± 0.02 min	14MB
Proposed Method (with ID)	1.27 ± 0.03 min (0.20 + 1.07)	6.6MB
Proposed Method (without ID)	3.06 ± 0.07 min	64MB

表 3: 比較手法を 1 とした時の問合せ実行時間の比の平均

Dataset	DIAERESIS	Proposed method (with ID)	Proposed method (without ID)
LUBM	1	0.80	0.95
SWDF	1	0.51	0.96

変換にかかる時間は計測していない．報告する時間は 5 回の実行の平均値である．

- 前処理時間: 分割したデータを出力するまでの時間
- 問合せ実行時間: SQL を実行してから結果を得るまでの時間
- テーブル作成コスト: 読み込んだデータ中のトリプル数と，テーブル作成までの時間
- 合計問合せ処理時間: 問合せ実行時間とテーブル作成時間の合計

5.4 実験結果（単一マシン）

5.4.1 前処理時間

前処理時間は元の RDF データセットを入力してから分割データが出力されるまでの時間を計測したものであり，その結果を表 2 に示す．提案手法の ID 化の有無による処理時間では，ID 化を行う場合の方が各データセットに対して大きく優れた結果を示しており，ID 化の優位性が確認できた．提案手法と比較手法では，LUBM では比較手法が，SWDF では提案手法が高速であり，ID 化にかかる時間がボトルネックであると考えられる．

5.4.2 問合せ実行時間

表 3 は SQL 実行時間の比較結果であり，いずれのデータセットにおいても提案手法が比較手法を上回る結果となっている．個々のクエリでの比較では，LUBM では 13 個中 12 個のクエリで，SWDF では 207 個中 155 個のクエリで提案手法が優れていた．特に SWDF では，平均 73%，最大 90%の処理時間削減となっており，一方で結果が劣っていたクエリについてもその処理時間増加量の平均は 23%，最大 46%であった．よって，一部劣る処理はあるもののほとんどの問合せ処理で性能向上が実証できていると考える．

5.4.3 テーブル作成コスト

提案手法と比較手法はいずれもインデックスを利用して必要なデータを事前に特定し，それらによって短時間で最低限のデータへアクセスしている． 表 4 は問合せ時に読み込んだトリプル数の比較であるが，いずれのデータセットにおいても提案手法は比較手法よりも平均的に多くのトリプルを読み込んでいる．これについて詳細な検証を行うと，特に LUBM のコンプレックスクエリで特に多くのトリプルを読み込んでいることが確認された．また，SWDF では半数以上の問合せでは読み込んだトリプル数は比較手法より少なかった．したがって，特定のパターンを含むクエリに対してはより多くのトリプルを読み込む傾向にあると考えられる．

表 4: 比較手法を 1 とした時のロードトリプル数の比の平均

Dataset	DIAERESIS	Proposed Method (with ID)	Proposed Method (without ID)
LUBM	1	2.01	2.59
SWDF	1	1.82	2.03

表 5: 比較手法を 1 とした時のテーブル作成時間の比の平均

Dataset	DIAERESIS	Proposed method (with ID)	Proposed method (without ID)
LUBM	1	0.66	1.48
SWDF	1	0.52	0.68

表 6: 比較手法を 1 とした時の合計時間の比の平均

Dataset	DIAERESIS	Proposed Method (with ID)	Proposed Method (without ID)
LUBM	1	0.71	1.26
SWDF	1	0.47	0.84

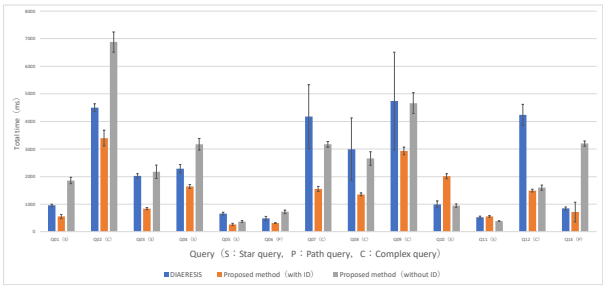


図 8: LUBM における合計時間の比較

一方，テーブルを作成する時間は表 5 に示す通りであり，提案手法はいずれのデータセットに対しても比較手法より 30%以上高速にテーブル作成を完了している．個別に LUBM のクエリを分析したところ，標準偏差も多くのクエリで比較手法より大幅に小さく，安定していることが確認できた．この傾向は SWDF でも同様であり，207 個中 205 個のクエリで比較手法より高速であった．これはワークロードの導入により複数のサブクエリ間でテーブルの共有が可能になり，作成すべきテーブル数が削減できたためではないかと考えられる．

5.4.4 合 計 時 間

最後にテーブル作成時間と問合せ実行時間の合計時間を問合せ全体にかかる時間とみなして比較を行う． 表 6 を見ると，提案手法は両データセットで比較手法を大幅に凌駕していることが確認できる．詳細な分析を行ったところ，LUBM においては図 8 に示すように，比較手法より 2 倍程度遅いクエリも 1 つ存在したものの，その他はほとんどのクエリで比較手法を大きく上回っており，最大約 75%の高速化を達成した．SWDF においても，198 のクエリで比較手法を上回っており，その差は最

Method	Processing Time
DIAERESIS	2.84 ± 0.06 min
Proposed Method（with ID）	20.18 ± 0.81 min（0.70+19.48）

図 9: 比較手法と提案手法の前処理時間（括弧内は ID 化時間（左）と分割時間（右））

method	Load Time	Query Time	Total Time
DIAERESIS	1	1	1
Proposed Method（with ID）	0.80	0.61	0.59

図 10: 比較手法を 1 とした時の各処理時間の比の平均

大約 95%であった．一方，残りの 9 個のクエリで比較手法に劣る結果となっていたが，その差は最大約 15%であり，許容できるものと考えられる．

5.5 実験結果（AWS）

5.5.1 前処理時間

AWS 上で RDF データを分割する際にかかった時間を計測した結果を表 9 に示す．両手法とも単一マシンでの実験よりも処理が遅くなっており，これは複数ノード間での通信にかかるコストによるものだと考える．また，提案手法の実装における分散処理の最適化が十分ではなく，その影響もより顕著に出たと考えている．

5.5.2 実 行 時 間

各処理にかかる時間を表 10 に示す．いずれも比較手法を大幅に上回っており，個別のクエリでも約 70%のクエリで提案手法が高速に処理されていた．さらに，標準偏差も多くのクエリで小さく安定した処理ができていると考えられる．

6 ま と め

本論文では，RDF データの効率的な分割と処理のための Spark ベースの方法を提案した．入力ワークロードクエリの共起を分析することにより関連するデータを同じパーティションに配置し，データ取得の効率向上を達成した．また，各データとそれが属するパーティションを記録したインデックスの導入によりさらなる効率化も可能にした．評価実験では，提案手法が単一ノードにおいて既存手法を上回ることを示した．これは主にデータアクセスの向上や共起性の活用による読み込むトリプル数の削減，ID 化による冗長性排除によるものであると考えられる．一方で，クエリの構成によっては重複したデータの読み込みが発生することがあり，それに伴って読み込むトリプルの数が極端に多くなる場合も確認した．したがって，今後の課題として重複したデータ読み込みの抑制が挙げられる．また，ワークロードの量や質による影響の検証，統計データ等の利用によるクエリ効率化，より大規模なデータセットにおける検証等も行う予定である．

謝　　辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の 「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業，JST CREST（JPMJCR22M2），科学研究費補助金（JP22H03694）に支援によるものです．

文　　献

[1] Graham KLYNE. Resource description framework (rdf) : Concepts and abstract syntax. *http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/*, 2004.
[2] B. McBride. Jena: a semantic web toolkit. *IEEE Internet Computing*, Vol. 6, No. 6, pp. 55–59, 2002.
[3] Thomas Neumann and Gerhard Weikum. Rdf-3x: A risc-style engine for rdf. *Proc. VLDB Endow.*, Vol. 1, No. 1, p. 647–659, aug 2008.
[4] Mohammad Hammoud, Dania Abed Rabbou, Reza Nouri, Seyed-Mehdi-Reza Beheshti, and Sherif Sakr. Dream: Distributed rdf engine with adaptive query planner and minimal communication. Vol. 8, No. 6, p. 654–665, feb 2015.
[5] Alexander Schätzle, Martin Przyjaciel-Zablocki, Simon Skilevic, and Georg Lausen. S2rdf: Rdf querying with sparql on spark. *arXiv preprint arXiv:1512.07021*, 2015.
[6] http://spark.apache.org.
[7] Damien Graux, Louis Jachiet, Pierre Genevès, and Nabil Layaïda. Sparqlgx: Efficient distributed evaluation of sparql with apache spark. In *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15*, pp. 80–87. Springer, 2016.
[8] Amgad Madkour, Ahmed M Aly, and Walid G Aref. Worq: Workload-driven rdf query processing. In *International Semantic Web Conference*, pp. 583–599. Springer, 2018.
[9] Georgia Troullinou, Giannis Agathangelos, Haridimos Kondylakis, Kostas Stefanidis, and Dimitris Plexousakis. Diaeresis: Rdf data partitioning and query processing on spark.
[10] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pp. 29–43, 2003.
[11] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. 2004.
[12] http://hadoop.apache.org.
[13] Adnan Akhter, Muhammad Saleem, Alexander Bigerl, and Axel-Cyrille Ngonga Ngomo. Efficient rdf knowledge graph partitioning using querying workload. K-CAP '21, 2021.
[14] Xintong Guo, Hong Gao, and Zhaonian Zou. Wise: Workload-aware partitioning for rdf systems. *Big Data Research*, Vol. 22, p. 100161, 12 2020.
[15] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. Lubm: A benchmark for owl knowledge base systems. *J. Web Semant.*, Vol. 3, No. 2-3, pp. 158–182, 2005.
[16] Muhammad Saleem, Qaiser Mehmood, and Axel-Cyrille Ngonga Ngomo. Feasible: A feature-based sparql benchmark generation framework. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14*, pp. 52–69. Springer, 2015.

Fast Parallel Computation for Random Walk based t-SNE

Takeshi MIKI†, Yasuhiro FUJIWARA††, and Hideyuki KAWASHIMA††

† Keio University
5322 Endō, Fujisawa-shi, Kanagawa 252–0882, Japan
†† NTT Communication Science Laboratories,
1–5–1 Ootemachi, Chiyoda-ku, Tokyo, 100–8116, Japan
E-mail: †t.miki@keio.jp, ††yasuhiro.fujiwara@ntt.com, †††river@sfc.keio.ac.jp

Abstract Effective dimension reduction techniques are crucial in the era of big data. In this paper, we introduce an optimized version of the renowned dimension reduction technique, random walk-based t-SNE, by addressing its traditionally significant computational cost. Our proposed method incorporates parallel computing and Modified Cholesky factorization to enhance processing efficiency, without sacrificing the quality of dimension reduction. The experimental results demonstrate a substantial reduction in execution time while maintaining the high standard of dimension reduction quality. This method has the potential to accelerate the adoption of t-SNE across various data-intensive domains.

Key words t-SNE, Parallization, Efficient

1 Introduction

In our data-driven world, the volume of high-dimensional data that we handle is continuously expanding [2]. Consequently, the importance of data visualization has been steadily increasing. The practice of visualizing high-dimensional data often necessitates a reduction of its dimensions to two or three. Numerous dimension reduction algorithms have been developed for this purpose, including UMAP [12], PCA [9], and PLS [13]. Among those methods, t-SNE (t-Distributed Stochastic Neighbor Embedding) is one of the most common and widely accepted approaches [15].

t-SNE is a non-linear dimension reduction algorithm that portrays the similarity between data points by conditional probabilities. Subsequently, it tries to minimize the Kullback-Leibler (KL) divergence of the conditional probabilities between the high-dimensional and lower-dimensional data points. In doing so, t-SNE effectively preserves both the global and local structural features. Due to its remarkable performance and simplicity, t-SNE finds the following diverse applications:

Cancer Detection: Mass Spectrometry Imaging (MSI) is a technique that simultaneously provides the spatial distribution of biomolecules. This information is necessary for identifying diagnostic and prognostic biomarkers in the context of cancer research [1]. While MSI offers rich insights into the intricate molecular landscape within cancer, its data complexity poses significant challenges. t-SNE, with its ability to reduce data dimensionality while preserving both global and local structures, is a suitable solution. In practice, the application of t-SNE involves the detection of distinct clusters. This has proven particularly beneficial for identifying gastric and primary breast cancer patients [1].

Playlist Generation: In the domain of playlist generation, t-SNE has been introduced as a modeling approach. Specifically, each song is represented as a 34-dimensional data, and t-SNE is used to transform it into a two-dimensional data. The two-dimensional output of the model will be used for projecting and rendering a song catalogue onto a plane. In this context, it has outperformed previous methodologies, such as PCA [10].

As shown in previous examples, t-SNE is a popular technique used in various fields. However, its main drawback is the significant computational cost. The problem arises from the need to calculate pairwise similarities between all data points to compute the Kullback-Leibler divergence between the original data points and the embedded data points. Additionally, t-SNE computes pairwise similarities for all embedded data points in each iteration when updating the embedded data using the gradient descent method. This leads to a quadratic increase in computational complexity [14].

In order to solve this computational bottleneck, random walk-based t-SNE has been introduced. This method selects a subset of landmark points in a randomized manner, reducing the number of data points that need to be visualized. Additionally, random walk-based t-SNE uses the k-nearest neighbor graph to approximate distances between

Table 1 Definition of each symbols

Symbol	Definition
N	Number of data-points
M	Number of dimensions in the original data
m	Number of dimensions in the embedded data
n	Number of landmark points
$\boldsymbol{X}$	Matrix of the original data
$\boldsymbol{Y}$	Matrix of the embedded data
x_i	i-th data-point of the original data
y_i	i-th data-point of the embedded data
p_{ij}	Similarity of the original data
q_{ij}	Similarity of the embedded data
$\mathbf{L}$	Graph Laplacian
k	Number of neighbors in *knn* graph

these points efficiently.

However, despite these optimizations, time complexity still remains a substantial challenge. This complexity stems from the need to construct and utilize the k-nearest neighbor graph, which is a critical step in the random walk-based approach for estimating pairwise distances and similarities among data points.

In this paper, we propose a multi-core version of the modified random walk-based t-SNE. The implementation was done in C++ and leveraged OpenMP to execute the code in parallel across multiple cores. The parallelization was done with thread counts ranging from 1 to 64.

The paper is structured as follows: Section 2 provides an overview of the foundational concepts, offering a brief explanation of t-SNE and random walk-based t-SNE. Section 3 introduces the methodology developed for this research. Section 4 discusses the experiments conducted and the results earned. Sequentially, section 5 focuses on related work. The conclusion and discussion drawn from the research is provided in sections 6 and 7.

2 Preliminaries

2.1 t-SNE

This section will explain the background and mathematical definitions related to t-SNE. First of all, t-SNE is a dimensional reduction technique used for data visualization. It aims to reduce the dimension of high-dimensional data points and create a lower-dimensional representation while preserving their pairwise similarities. Specifically, t-SNE reduces the dimension of the original data $\boldsymbol{X}$, represented as $N \times M$ matrix, into a lower-dimensional matrix $\boldsymbol{Y}$, represented as $N \times m$ embedded matrix, where N, M and m are the number of data-points, number of dimension in the original data and the number of dimension in the embedded data, respectively. This is achieved by minimizing the cost function F, defined as follows:

$$F = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{1}$$

where p_{ij} is the pairwise similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ represents the ith and jth element of $\boldsymbol{X}$, respectively. Additionally, q_{ij} represents the pairwise similarity of embedded data, $\mathbf{y_i}$ and $\mathbf{y}_j$, where $\mathbf{y}_i$ and $\mathbf{y}_j$ are the ith and jth element of $\boldsymbol{Y}$, respectively. Pairwise similarity in the high dimensional data, p_{ij} is defined as follows:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2},\ p_{i|i} = 0 \tag{2}$$

where $p_{i|j}$ is the similarity of x_i to x_j using conditional probability. $p_{i|j}$ is given as follows:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma^2)} \tag{3}$$

where σ is the variance of the Gaussian distribution.

Since x_i will not choose x_i as its neighbour, $p_{i|i}$ becomes 0. On the other hand, $p_{i|j}$ may not equal to $p_{j|i}$, however, joint probability of Equation (2) p_{ij} takes the average. Therefore, the pairwise similarity between x_i and x_j becomes symmetrized. The pairwise similarity of y_i to y_j is defined as q_{ij} which is given as follows:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_i - y_l||^2)^{-1}},\ q_{ii} = 0 \tag{4}$$

Unlike p_{ij} which uses a Gaussian distribution to convert distance into probability [15], q_{ij} uses student-t distribution with single degree of freedom to effectively visualize data. Since the distance from y_i to y_i is 0, $q_{ii} = 0$.

Embedding original data is done by minimization of the cost function using gradient descent method with momentum. The gradient is as follows:

$$\frac{\partial F}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{5}$$

After obtaining the gradient, $\mathbf{Y}$ should be updated. Gradient update with momentum is given as follows:

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial F}{\partial Y} + \alpha(t) \left(Y^{(t-1)} - Y^{(t-2)} \right) \tag{6}$$

where t is the number of iterations when undergoing gradient descent method, α represents the momentum and η is the learning rate.

Since t-SNE visualizes all embedded data-points, it is required to compute the value of q_{ij} for each data-point pair. Consequently, the computational and memory costs becomes $O(N^2)$. These expensive costs are a major issue, since it becomes infeasible for t-SNE to deal with a big data-set.

2.2 Random walk-based t-SNE

One common solution for high computational cost is to use Random walk-based t-SNE, where it tries to reduce the cost by selecting a certain number of landmark points and only visualizing them [15].

Additionally, unlike original t-SNE, random walk-based t-SNE uses *knn* graph to calculate pairwise similarity analytically. this method starts off by creating *knn* graph for all N data-points with k as the number of neighbours. This *knn* graph is represented using graph Laplacian, denoted as $N \times N$ matrix $\mathbf{L}$, which is defined as

$$\mathbf{L} = \mathbf{A} - \mathbf{W} \tag{7}$$

where $\mathbf{W}$ represents adjacency matrix and $\mathbf{A}$ represents diagonal matrix. When obtaining $\mathbf{W}$, an assumption that the distance between data points x_i and x_j is equivalent to $e^{-||x_i - x_j||^2}$ is made. Additionally, diagonal matrix $\mathbf{A}$ is obtained by diag $(\sum_j W_{1,j}, \sum_j W_{2,j}, \cdots \sum_j W_{N,j})$.

Pairwise similarity p_{ij} can be redefined as the fraction of random walks that starts at landmark x_i and terminates at landmark x_j. Pairwise similarity between all data-points and landmark points could be obtained accurately by solving the following linear system:

$$\mathbf{LP} = -\mathbf{B^T} \tag{8}$$

In this equation, $\mathbf{B^T}$ is $N \times n$ matrix containing columns from adjacency matrix $\mathbf{W}$ corresponding to landmark points, where n is the number of landmark points. It is important to note that $\mathbf{P}$ is a $N \times n$ matrix where pairwise similarity between ith data point and jth landmark point is represented as $\mathbf{P}_{ij}$.

In order to solve Equation (8), Cholesky factorization [8] on Graph Laplacian $\mathbf{L}$ is first conducted, resulting in $\mathbf{L} = \mathbf{CC^T}$, where $\mathbf{C}$ represents the upper-triangular matrix. Sequentially, $\mathbf{Cy} = -\mathbf{B^T}$ is solved using forward substitution and $\mathbf{CP} = \mathbf{y}$ is solved using backward substitution [15].

Once pairwise similarities of landmark points from data points are obtained from $\mathbf{P}$, random walk-based t-SNE would compute the pairwise similarities between embedded data using Equation (4). However, it is important to note that only landmark points are used to visualize, meaning that there are only n embedded data. One should note that embedded data is denoted as $n \times m$ matrix $\mathbf{Y}$, where m is the number of dimension in the embedded space. Gradient descent method is used to update $\mathbf{Y}$.

While the random walk-based t-SNE method offers a potential for acceleration, the challenge involving computational complexity still remains. Specifically, the computation of pairwise similarities q_{ij} among the landmarks scales quadratically with the number of landmarks. Additionally the computational cost for updating embedded matrix $\mathbf{Y}$ has complexity of $O(tn^2)$ where t is the number of iterations in the gradient descent. When using very large datasets, where both the number of landmarks and the iteration count can be enormous, these challenges makes this technology infeasible. To address these issues, our approach focuses on optimizing three key areas: the acceleration of the k-nearest neighbors (*knn*) computation, the efficient calculation of the matrix P representing pairwise similarities, and the updating process for the matrix Y. These enhancements are designed to reduce the time complexity inherent in the original method, making it more applicable for the practical use of large-scale data.

3 Method

3.1 Parallelization of Graph Laplacian

Graph Laplacian $\mathbf{L}$ is calculated using Equation (7). The computational complexity of creating *knn* graph as graph Laplacian is $O(N^2M)$, where M is the number of dimension in the original data. Although this step is only done once, depending on the number of its data and dimension, it could be computationally expensive. Thus, this could be a major issue when dealing with a large data-set.

To address this challenge, we used OpenMP, a standard parallel programming API for C and C++. OpenMP allows us to compute the weight of the edges in parallel. Therefore, the construction of *knn* graph can be divided among multiple processors. Simple implementation is shown in Algorithm 1.

Each row of $\mathbf{L}$ corresponds to a node in *knn* graph. Determining which nodes are connected is done by comparing the weight of the edges and selecting k nearest nodes. In Algorithm 1, since determining each row of $\mathbf{L}$ can be computed independently, it is done in a parallel loop by distributing computation across multiple threads. It is important to note that the weight of the edges are calculated using $e^{-||x_i - x_j||^2}$, which is also computed in parallel. Sequentially, by comparing the weights, we could identify whether a node is k nearest neighbor or not. If a node is not a neighbor then those nodes are not connected, thus the weight will be 0.

In order to efficiently compare the weights, we would create a simple buffer with k spaces and a lock to avoid race condition. The buffer represents the k nearest neighbors. An entry value, which would be the weight W_{ij} of each nodes, would be placed in the buffer if it is empty. If not, by comparing the smallest value of the buffer, each node could either be swapped and replaced in the buffer or rejected meaning that the node is not k nearest neighbor. It is important to note that the smallest value in the buffer always comes in the beginning so that the comparison can be done simply.

Additionally, when a entry value wants to read or write to the buffer, it must acquires the lock in order to avoid race condition. Finally, once each row of $\mathbf{L}$ is computed, they are merged together to create one large matrix $\mathbf{L^T}$.

Algorithm 1: Parallelization of graph Laplacian

Data: $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$, k= Number of neighbours
Result: Graph Laplacian $\mathbf{L}$
♯ pragma omp parallel for
for *each* x_i **do**
  ♯ pragma omp parallel for
  for *each* x_j **do**
    $W_{ij} = e^{-||x_i - x_j||^2}$
    if x_j is not k-nearest neighbour **then**
      $W_{ij} = 0$
    end
  end
  ♯ pragma omp parallel for
  for *each* x_j **do**
    $A_{ii} += W_{ij}$
  end
end
$L^T = \begin{bmatrix} L_{1,;} & L_{2,;} & \cdots & L_{N,;} \end{bmatrix}$
Return $\mathbf{L}$

3.2 Solving Linear Systems

In order to find pairwise similarity p_{ij} using graph Laplacian, Equation (8) is solved. This is done by decomposing L using Cholesky factorization then using forward and back substitution to get the value of P which is a matrix containing pairwise similarity. However, computational cost for Cholesky factorization is cubic in the number of data points [5]. One solution is to use Modified Cholesky factorization.

Cholesky factorization decomposes positive symmetric matrix $\mathbf{L}$ into lower triangular matrix $\mathbf{C}$ and upper-triangular matrix $\mathbf{C^T}$. Starting from C_{11}, by using Equation (9) and Equation (10), Cholesky factorization solves each value of $\mathbf{C}$. The computational cost for Cholesky factorization would be $\frac{n^3}{3}$ [6].

$$C_{jj} = \sqrt{L_{jj} - \sum_{k=1}^{j-1} C_{jk}^2} \tag{9}$$

$$C_{ij} = \frac{L_{ij} - \sum_{k=1}^{j-1} C_{ik} C_{jk}}{C_{jj}},\ i > j \tag{10}$$

The proposed method uses Modified Cholesky factorization to accelerate the process. The core idea of modified Cholesky factorization is to decompose $\mathbf{L}$ in such way that $\mathbf{L} = \mathbf{CDC^T}$ where $\mathbf{D}$ is an additional diagonal matrix. It is important to note that diagonal component of $\mathbf{C}$ is all 1. This decomposition would reduce the computational cost for factorization but also support solving Equation (8). Similar to Cholesky factorization, starting off from C_{11}, using Equation (11) and Equation (12), Modified Cholesky factorization tries to reduce the total computation. It should be pointed out that, D_{11} would be equal to A_{11}. The total computational complexity would be $\frac{n^3}{6}$ [6], which is faster than normal Cholesky factorization.

$$C_{ij} = \frac{L_{ij} - \sum_{k=1}^{j-1} C_{ik} D_{kk} C_{jk}}{D_{jj}},\ C_{ii} = 1 \tag{11}$$

$$D_{ii} = L_{ii} - \sum_{k=1}^{i-1} C_{ik}^2 D_{kk},\ D_{11} = A_{11} \tag{12}$$

Modified Cholesky factorization guarantees a speedup by using its symmetric features to avoid unnecessary square root calculations.

Furthermore, solving linear system presented in Equation (8) is achieved by solving $\mathbf{CDH} = -\mathbf{B^T}$ using forward substituting and solving $\mathbf{C^T P} = \mathbf{H}$ using backward substitution [7].

Fast computation of solving linear systems is achieved by replacing Cholesky factorization with Modified Cholesky factorization and parallelization as presented in Algorithm 2.

Algorithm 2: Solving Linear Systems

Data: $\mathbf{L}$= Graph Laplacian, $\mathbf{B}$= Matrix containing columns from adjacency matrix $\mathbf{W}$, n= Number of landmark points, N= Number of data points
Result: $\mathbf{P'}$= Matrix containing the pairwise similarity between landmark points
for $i = 1$ *to* N **do**
  Solve D_{ii} with Equation (12)
  ♯ pragma omp parallel for
  for $j = i + 1$ *to* N **do**
    Solve C_{ji} with Equation (11)
  end
end
♯ pragma omp parallel for
for $i = 1$ *to* n **do**
  Solve $\mathbf{p}_i$ with Equation (13)
  ♯ pragma omp parallel for
  for $k = 1$ *to* N **do**
    if Encounter landmark **then**
      Append p_{ik} to $\acute{p}_i$
    end
  end
  $\mathbf{P'} = \begin{bmatrix} p'_1 & p'_2 & \cdots & p'_n \end{bmatrix}$
end
Return $\mathbf{P'}$

The first loop in Algorithm 2 endeavors to decompose $\mathbf{L}$ into $\mathbf{CDC^T}$ by solving Equation (11) and Equation (10) in

turn. This is because once D_{ii} is known, ith column of $\mathbf{C}$ can be solved. Therefore, computation of each element in the ith column is done in parallel. It is important to note that since $\mathbf{C}$ is a lower triangular matrix with diagonal component being 1, C_{ij} should only be computed if $j > i$.

In order to achieve efficient parallelization by distributing computation across multiple threads, we solved Equation (8) in such way that $\mathbf{P}$ is solved per column. Therefore, in each thread, we compute the following equation:

$$\mathbf{L}\mathbf{p}_i = \mathbf{b}_i \tag{13}$$

where $\mathbf{p}_i$ represents ith column of matrix $\mathbf{P}$. Similarly, $\mathbf{b}_i$ represents ith column of $-\mathbf{B^T}$.

Equation (13) can be solved by computing $\mathbf{CDh} = \mathbf{b_i}$ and $\mathbf{C^T p_i} = \mathbf{h}$ using forward and back substitution.

Since Equation (13) is computed in parallel, $\mathbf{p}_i$ would be computed in each thread. Since we are not interested in elements of $\mathbf{p}_i$ that corresponds to non landmark points, our algorithm would only store elements that corresponds to landmark points to a column vector $\mathbf{p'_i}$. To avoid excess operations, when all landmark points are selected, we would terminate the process.

The output for Algorithm 2 is $n \times n$ matrix $\mathbf{P'}$ where P'_{ij} represents pairwise similarity between ith and jth landmark points. It is important to note that $\mathbf{P'}$ is equivalent to selecting rows that corresponds to landmark points in $N \times n$ matrix $\mathbf{P}$.

3.3 Parallelization of Updating Embedded Data

Computational complexity for obtaining $\mathbf{Y}$ is $O(tn^2)$, where t is the number of iterations done when applying gradient descent method and n is the number of landmarks. Depending on the number of iterations, calculating $\mathbf{Y}$ can also be very computationally expensive. Parallelization of updating embedded data is shown in Algorithm 3.

Embedded data $\mathbf{Y}$ would be initialized by random manner [15]. Pairwise similarity between embedded data q_{ij} is obtained using Equation (4). After computing pairwise similarity, the gradient of cost function $\frac{\delta \mathcal{F}}{\delta \mathbf{Y}}$ is acquired using Equation (5). Since computation of q_{ij} for each point is computationally heavy, we would calculate this in parallel. Once the gradient is calculated, $\mathbf{Y}$ is updated using gradient descent method with momentum according to Equation (6).

3.4 Summary of the Proposed Method

Algorithm 4 provides a summary of the process to compute the proposed version of random walk-based t-SNE. There are 3 main steps to the entire implementation. Firstly, we need to compute the adjacency matrix $\mathbf{W}$ and graph Laplacian $\mathbf{L}$ which is indicated in Algorithm 1. Secondly, we would want to obtain matrix $\mathbf{P}$, which involves pairwise similarity

Algorithm 3: Parallelization of Updating Embedded Data

Data: $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$, $\mathbf{P'}$ = Matrix with pairwise similarity between landmark points, t = The number of iteration to conduct gradient descent method, α = momentum, η = learning rate
Result: $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$
Initialize $\mathbf{Y}$
for *i=1 to t* **do**
  ♯ pragma omp parallel for
  for *each* y_i **do**
    ♯ pragma omp parallel for
    for *each* y_j **do**
      Calculate q_{ij} using Equation (4)
    end
  end
  Update $\mathbf{Y}$ using Equation (5) and Equation (6)
end
Return $\mathbf{Y}$

p_{ij}, using Algorithm 2. However, before running Algorithm 2, we need to compute $-\mathbf{B^T}$ in order to solve Equation (8). This is done by extracting columns of adjacency matrix $\mathbf{W}$, which corresponds to landmark points. $\mathbf{Y}$ is then updated using Algorithm 3 where the computation of gradient is done in parallel.

Our approach improves the efficiency of random walk-based t-SNE focuses on parallel computation and algorithmic optimizations. By leveraging multi-core processors, we hope to accelerate the construction of the graph Laplacian, calculation of pairwise similarities and updating $\mathbf{Y}$.

Algorithm 4: Summary of Parallel and fast computation for random waslk-based t-SNE

Data: $\mathbf{X} = \{x_1, x_2, \ldots, P\}$, t= Number of iteration
Result: $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$
if Dimension of $\mathbf{X} > 30$ **then**
  Reduce the dimension of X to 30 by PCA
end
$\mathbf{W}$, $\mathbf{L} \leftarrow$ Algorithm 1
$\mathbf{B^T} \leftarrow$ Columns of $\mathbf{W}$ corresponding to landmark points
$\mathbf{P'} \leftarrow$ Algorithm 2
Update $\mathbf{Y}$ using Algorithm 3
Return $\mathbf{Y}$

4 Experiments and Results

4.1 Experimental Setup

In order to examine the results of the proposed methods, we have conducted multiple experiments. Implementation was done using gcc and OpenMP on a x86-64 computer with 64 CPU cores with frequency of 2.10GHz and 1.5 TiB memory.

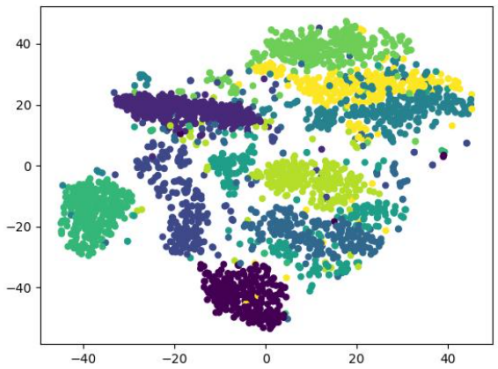


Figure 1 MNIST 2500 with Random Walk based t-SNE

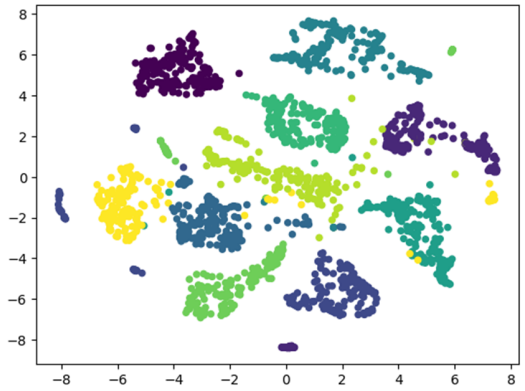


Figure 2 MNIST 2500 with proposed method

Following the original paper on t-SNE, if the number of dimensions of the original data was higher than 30, we conducted PCA to reduce it to 30 [15]. Sequentially, we used t-SNE in order to reduce the dimension into 2. Additionally, we limited the number of nearest neighbors to 20. The number of iteration t was set to 1000, the momentum was 0.5 if $t < 250$ and 0.8 if $t \geqq 250$ [15]. Referring to the original paper, we decided to use an adaptive learning rate, where the initial learning rate would be 100 but would get updated over time [15].

4.2 Visualization of datasets

In order to confirm the quality of this technique, we visualized multiple data using random walk-based t-SNE and the proposed method.

First dataset we used was MNIST 2500 [3] which has 2500 data with each containing 784 dimensions. For this common dataset, each dimension represents the pixel of a hand written number. There are 10 labels from 0 to 9. Figures 2 and 1 shows the result of applying random walk-based t-SNE and the proposed method. The separation of the clusters in Figure 2 indicates that the proposed method has effectively reduced the high-dimensional data into two dimensions while preserving the distinct features of each class. Although Figures 1 and 2 do not look exactly the same, they have both clustered different handwritten digits into a fairly distinct groups, meaning that the proposed method could conduct dimension reduction without the loss of quality. The reason why the two figures do not look exactly the same is due to randomness involving initialization of $\mathbf{Y}$.

The Iris dataset [4] was also used for visualization. The dataset contains 150 samples from three species of Iris flowers: setosa, Virginica and Versicolor. The four features in this dataset includes the length and the width of sepals and petals in centimeters. Figures 3 and 4 illustrate the results of visualization using random walk-based t-SNE and proposed method. The distinction between the three groups illustrates the success in visualization of the dataset. Both techniques achieve a clear visual separation of the species, which suggests that the proposed method maintains the quality of dimension reduction. The similarity in Figures 4 and 3 including the positioning of the embedded data points, is likely because of the straightforward structure of the Iris dataset. If we were using more complex datasets, it might be more challenging to achieve such consistency in results between different dimensional reduction methods.

In summary, when we visualized both the MNIST 2500 dataset and the Iris dataset, the outcomes suggested that the proposed method was capable of conducting dimension reduction without the loss of quality. Both methods were able to embed the data into a simpler form without losing the general structure of the data.

4.3 Processing Time

We have measure the execution time using MNIST 2500 [3] dataset. Figure 5 shows the execution time for varying numbers of threads. The experiment began with a single thread, and then additional threads were introduced in increments of 8, culminating in a total of 64 threads.

As shown in Figure 5, the parallelization was done perfectly as the execution time has dramatically dropped according to the number of threads. A significant decrease in processing time is observed as the number of threads increases. This steep decrease from a single thread to approximately 10 threads illustrates the initial benefits of parallel computation. After the 10 threads, the graph illustrates a fairly stable trend. This could originate from several factors in parallel computing, such as synchronization overhead, the communication time between threads etc.

Figure 6, 7, and 8 represents the execution time for Algorithm 1, 2, and 3, respectively,.

Figure 6 shows a steep decline in execution time as the number of threads increases, indicating significant gains from parallelization in the initial phase. However, it is important

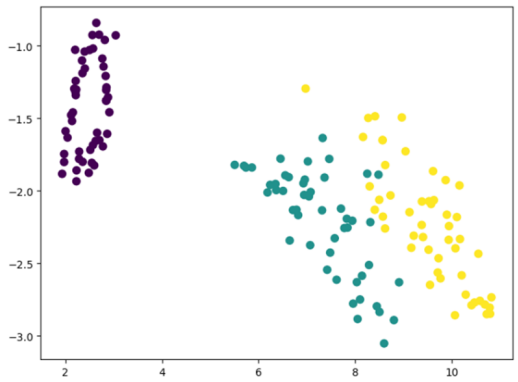


Figure 3 Iris dataset with Random Walk based t-SNE

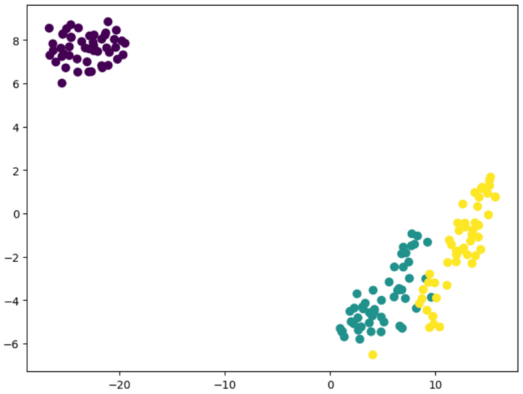


Figure 4 Iris dataset with proposed method

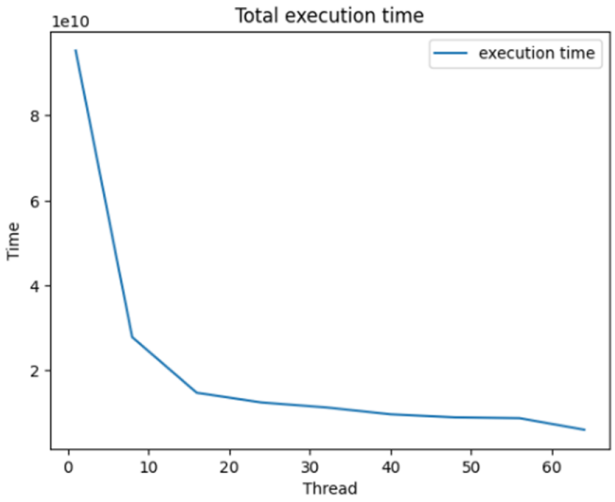


Figure 5 Execution time of Parallel t-SNE on MNIST 2500

to note that for this case, the actual impact of this step on the overall execution time was minimal. This implies that, although parallelization makes this step faster, it was not the critical factor in the total computation time. This maybe due to reduction in dimension in the prepossessing stage where the dimension of MNIST has been reduced to 30.

Figure 7 also shows a sharp reduction in execution time as more threads are utilized. This step has utilised parallel processing and modified choleky factorization. Given that this step is the primary contributor to the total execution time, the efficiency gains in Algorithm 2 are critical to the overall performance improvement of the method.

Figure 8 illustrates a less steep reduction in the execution time compared to the others. This could be due to factors like the iteration in gradient descent method that do not decrease linearly with the addition of more threads.

The substantial reduction in execution time was achieved mainly through the refinements in Algorithm 2. This was what decisively improved the method's efficiency. The other steps, while improved by parallelization, play secondary roles in the method's overall performance enhancement.

To sum up, by using modified cholesky factorization and parallel programming, we have successfully demonstrated that our method reduces computational time significantly, which has addressed one of the biggest disadvantages with original t-SNE.

5 Related Work

Previous efforts have explored enhancing t-SNE's efficiency through parallel processing. For example, Lopes et al. (2020) developed a parallel version of t-SNE to aid in visualizing smart city data [11]. Their work involved three main phases: prepossessing, joint probability calculations, and execution of t-SNE, all of which were parallelized using C and OpenMP.

In their prepossessing phase, parallel computing was applied to speed up the calculation of eigenvalues and eigenvectors as well as the computation of the distance matrix when undergoing PCA. For calculating joint probabilities, tasks such as Gaussian kernel calculations and entropy estimations were performed concurrently. Finally, in the execution phase, the joint probability calculations, the Student-t distribution computations, and the gradient calculations were all parallelized.

In contrast, our research advances the field by applying parallel computing specifically to random walk-based t-SNE for the first time, as far as the author is aware.

6 Conclusion

We have presented an advanced methodology for executing random walk-based t-SNE, utilizing parallel computing and Modified Cholesky factorization to enhance computational efficiency. Our approach has succeeded in significantly reducing processing times, with scalability that effectively corresponds with the addition of processing threads. The results demonstrate the feasibility of our method, as evidenced by the substantial decrease in execution time while maintaining the quality of data visualization. This approach not only offers a faster alternative to traditional random walk-based

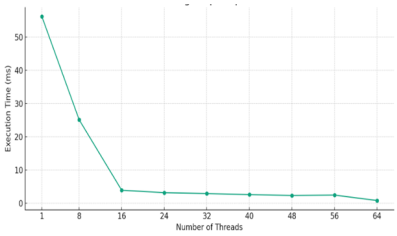


Figure 6 Creating graph Laplacian

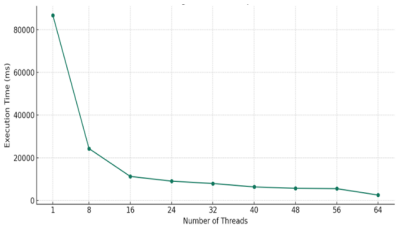


Figure 7 Computing pairwise similarity

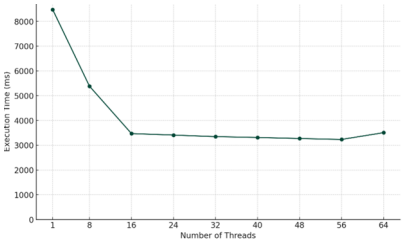


Figure 8 Updating embedded data

t-SNE methods but also provides a path for tackling larger and more complex datasets, which were previously impractical to process.

7 Discussion

A primary concern with the proposed method is its memory footprint, which expands proportionally with the dataset size, potentially limiting its application to extremely large datasets. With the proposed methodology, the computer stores at least three $N \times N$ matrices during Modified Cholesky factorization. As the number of data points grows significantly large, this requirement may become impractical. Future research will be directed towards optimizing memory usage, potentially through the refinement of the data structures employed. We aim to investigate techniques for partitioning the data processing, enabling our method to be both adaptable and memory-efficient. Such strategies might include mini-batching or streaming data processing, each of which can manage memory usage dynamically.

References

[1] Walid M Abdelmoula, Benjamin Balluff, Sonja Englert, Jouke Dijkstra, Marcel JT Reinders, Axel Walch, Liam A McDonnell, and Boudewijn PF Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, 113(43):12244–12249, 2016.

[2] Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, and Athanasios V Vasilakos. The role of big data analytics in internet of things. *Computer Networks*, 129:459–471, 2017.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[4] R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

[5] Yasuhiro Fujiwara, Yasutoshi Ida, Sekitoshi Kanai, Atsutoshi Kumagai, and Naonori Ueda. Fast similarity computation for t-sne. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 1691–1702. IEEE, 2021.

[6] PE Gill. W. murray, and mh wright. *Practical Optimization*, pages 212–229, 1981.

[7] Leo Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–1783, 2006.

[8] Nicholas J Higham. Cholesky factorization. *Wiley interdisciplinary reviews: computational statistics*, 1(2):251–254, 2009.

[9] H Hotelling. Analysis of a complex of statistical variables into principal components. warwick & york. *Inc, Baltimore, Maryland*, 1933.

[10] Matteo Lionello, Luca Pietrogrande, Hendrik Purwins, and Mohamed Abou-Zleikha. Interactive exploration of musical space with parametric t-sne. In *15th Sound and Music Computing Conference (SMC 2018)*, pages 200–208. Sound and Music Computing Network, 2018.

[11] Maximiliano Araújo Da Silva Lopes, Adrião D Dória Neto, and Allan De Medeiros Martins. Parallel t-sne applied to data visualization in smart cities. *IEEE Access*, 8:11482–11490, 2020.

[12] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[13] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S Davis. Human detection using partial least squares analysis. In *2009 IEEE 12th international conference on computer vision*, pages 24–31. IEEE, 2009.

[14] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, 15(1):3221–3245, 2014.

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

ブロックチェーン技術を活用したデータ検証可能な分散データベースの性能に関する検討

坂本 明穂† 小口 正人†

† お茶の水女子大学 〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †akiho@ogl.is.ocha.ac.jp, oguchi@is.ocha.ac.jp

あらまし 複数ユーザ間でのデータの共有方法として分散データベースの利用が考えられるが，共有されたデータを有効に活用するためにはデータベース上に保存されたデータが正しいことを保証する必要がある．そこでデータの保存時にブロックチェーンによる検証プロセスを行うことで健全なデータのみをデータベースに保存することができる．しかしブロックチェーンはデータ検証プロセスを持つために一般的なデータベースと比べて処理性能が劣ることが知られている．そこでブロックチェーンの 1 つである Hyperledger Iroha について，一般的なデータベースである PostgreSQL との処理性能の比較を行う．また Hyperledger Iroha のトランザクション処理における特徴について調査する．

キーワード 並列・分散処理，分散システム，分散台帳

A Study on the Performance of Data Verifiable Distributed Database Using Blockchain Technology

Akiho SAKAMOTO† and Masato OGUCHI†

† Ochanomizu University,
2–1–1 Otsuka, Bunkyou-ku, Tokyo 112–8610 Japan
E-mail: †akiho@ogl.is.ocha.ac.jp, oguchi@is.ocha.ac.jp

1. はじめに

分散データベース上でデータを管理することで，複数のユーザ間でのデータ共有が容易になる．例えば製造業における各部品製造段階での二酸化炭素排出量を分散データベース上で管理することで，複数の企業にまたがるデータを一元的に管理することができる．データの一元的な管理により，製品の製造時に各過程で排出される二酸化炭素量の追跡やカーボンフットプリントの管理が簡単になるため，製造業における二酸化炭素排出量の削減などが期待できる．しかし保存されたデータを有効に活用するためには，管理対象となるデータが正しいものであることを保証する必要がある．すなわち分散データベースには，保存される全てのデータに対してその健全性を検証してから保存するような機能が求められる．しかし分散データベース上に保存されるデータ量が増加すると，データベースに保存される全てのデータに対して実用的な時間で検証することが難しくなる．そこで検証対象となるデータをメタデータに限ることで検証対象となるデータ量を減らし，データ検証にかかる時間の削減を行う．その結果，実時間でのデータ検証が可能な分散データベースの構築が可能となる．

データ検証を行うプラットフォームとしてブロックチェーンの利用が考えられる．ブロックチェーンはネットワークに送信されたトランザクションが正しいかを検証するメカニズムを持っているため，この機能を利用して分散ストレージを作成することで目的のデータベースを構築することができる．

ブロックチェーンには Bitcoin などの不特定多数のユーザが参加できるパブリックブロックチェーンと，Hyperledger Fabric などの事前に認証されたユーザしか参加できないプライベートブロックチェーンがある．分散データベースによるデータの共有は，複数の企業間などの特定のユーザ間で行われることが想定されるため，プライベートブロックチェーンが適切と考えられる．

プライベートブロックチェーンにおけるデータ検証では，ブロックチェーンに参加しているノード間でメッセージをやり取りし，データが正しいという合意を形成することが必要となる．ノード間でのメッセージのやり取りはネットワークを通して複数回行われるため，ブロックチェーンでのデータの処理速度は通常のデータベースよりも遅くなる．本研究では，プライベー

トブロックチェーンの 1 つである Hyperledger Iroha の性能について通常のデータベースとの性能の比較を行う．またネットワークサイズを変化させて，処理するトランザクション数が増加することによるレイテンシやスループットへの影響を計測する．

2. Hyperledger Iroha

プライベートブロックチェーンの中では Hyperledger プロジェクトの 1 つである Hyperledger Fabric が有名であるが，同じ Hyperledger プロジェクトで直近にリリースされたものに Hyperledger Iroha がある．Hyperledger Iroha は Hyperledger Fabric よりも効率の良いコンセンサスアルゴリズムを採用しているという特徴がある．

2.1 Hyperledger Fabric

Hyperledger Fabric v0.6 ではコンセンサスアルゴリズムに PBFT(Practical Byzantine Fault Tolerance) が，Hyperledger Fabric v1.0 では BFT-SMaRT が用いられている．両者の違いは信頼性の向上と署名の評価のためのマルチコア計算である [1]．PBFT 方式のコンセンサスアルゴリズムでのトランザクション処理の流れは以下の通りである．まずクライアントは各ノードにトランザクションを送信し，ノードはそれをシミュレーションしてその結果であるエンドースメントをクライアントに送信する．各ノードから返ってきたエンドースメントが一致していれば，クライアントはそれらを当該トランザクションに含めてオーダリングサービスに送信する．オーダリングサービスは複数トランザクションから成るブロックを作成し，それをノードに配布する．ノードはブロックに含まれるトランザクションを検証して，検証を通ったものだけを台帳に反映する [5]．

2.2 Hyperledger Iroha

Hyperledger Iroha では YAC(Yet Another Consensus) と呼ばれるコンセンサスアルゴリズムを採用している．また YAC 方式ではネットワークを構成するノードの一部が悪意のあるユーザやオフラインになったとしても正常に動作することが保証されている．

図 1 に Hyperledger Iroha ネットワークにおけるトランザクションの処理の流れを，図 2 にコンセンサス形成におけるピア間でのメッセージのやり取りを示す．ただしクライアントはトランザクションを生成してオーダリングサービスに送信する役割をもち，ピアはプロポーザルに含まれるトランザクションを検証して合意し，トランザクションをブロックへ格納する役割をもつ．またオーダリングサービスはクライアントから送信されたトランザクションを受け取り，1 つ以上のトランザクションから成るプロポーザルを作成する役割を果たす．

1. クライアントがピアにトランザクションを送信する．
2. トランザクションを受け取ったピアはオーダリングサービスにトランザクションを転送する．
3. オーダリングサービスは，一定時間内に受け取ったトランザクションをまとめて順番に並べ，プロポーザルを作成する．
4. オーダリングサービスは各ピアにプロポーザルを送信する．
5. ピア間でコンセンサスを形成する．
   a. プロポーザルを受け取ったピアは，プロポーザルに含まれるトランザクションが正しいかを検証する．
   b. プロポーザルに含まれる正しいトランザクションのみからハッシュ値を計算する．
   c. ピアは正しいトランザクションとハッシュ値を含むブロックを生成する．
   d. ハッシュ値と初期のピア順から新しいピア順を計算する．
   e. 求めたピア順の先頭ピアにプロポーザルハッシュとブロックハッシュ，ピアの署名を含む投票を送る．
   f. 先頭ピアは 2/3 以上の投票を得ると，ネットワークを構成するピアにコミットメッセージをブロードキャストする．
6. 各ピアは合意されたトランザクションのブロックをチェーンに追記する．

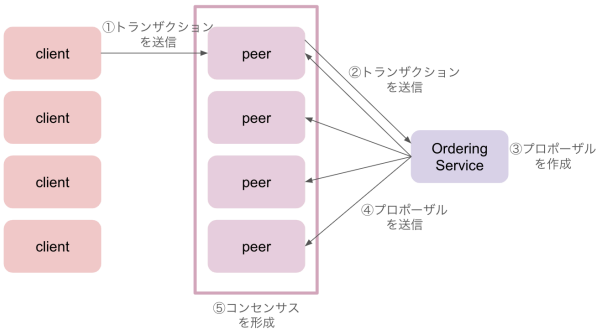


図 1: トランザクションの処理プロセス

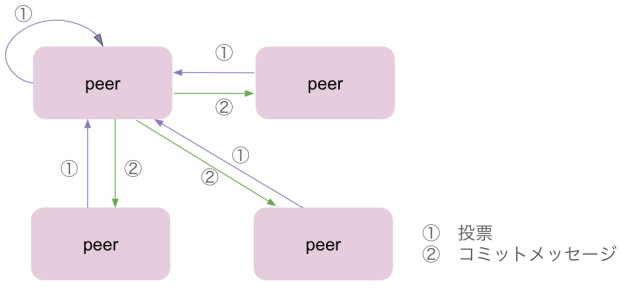


図 2: コンセンサス形成時のメッセージのやり取り

3. 関 連 研 究

Muratov ら (2018) は異なるネットワークサイズに対して，投票ステップ遅延を変化させたときのスループットの変化および，不安定な挙動を示したノードの数について測定を行っている [1]．ここで投票ステップ遅延とは，プロポーザルを検証し投票によって次のブロックを決定する期間を指す．4，16，28，64 ノードのネットワークに対して測定を行った結果，4 ピア程度の小さなネットワークサイズでは小さな投票遅延値を設定す

ることがスループットの向上に有効である一方，ネットワークサイズが大きくなると小さな投票遅延値ではタイムアウトが発生してコンセンサスに失敗する可能性があることが示された．したがってネットワークのスループットを最大にするためには，28 ノード以下では投票遅延値を 1ms に設定し，64 ノードでは 20ms に設定するとよい．

Woznica ら (2022) は，Hyperledger Fabric，Hyperledger Sawtooth，Hyperledger Iroha について，トランザクション送信レート，ブロックサイズ，ネットワークトラフィックモード，ネットワークサイズを変化させたときのレイテンシとスループットの変化を測定している [2]．Hyperledger Iroha について次のような結果が得られた．トランザクション送信レートを 10，20，30，40，50，60，70，80，90，100，150，200 と変化させて最小レイテンシを計測すると両者に厳密な相関関係は見られない．また同条件でトランザクション送信レートを変化させても平均レイテンシに影響はないが，ネットワークサイズを 5，10，20，30，40 と増加させると平均レイテンシは大きくなる．さらにトランザクション送信レートが小さい (20tps) ときスループットは概ね一定であるが，トランザクション送信レートが高い (100tps) とき，スループットはネットワークサイズが大きくなるにつれて減少する．またブロックサイズを 10 と 50 の 2 種類に対して平均レイテンシとスループットを計測すると，ほとんどのトランザクション送信レートにおいてブロックサイズが大きいほど平均レイテンシが短くなり，ブロックサイズが小さくなるほどスループットが低下する．

Anh Dinh ら (2017) はブロックチェーンのベンチマークツールである Blockench を用いて Ethereum，Parity，Hyperledger Fabric v0.6 の性能評価を行った [3]．Hyperledger Fabric は Ethereum や Parity よりも一貫して優れたパフォーマンスを示したが，16 ノード以上へのスケールアップには失敗した．また Hyperledger Fabric と Ethereum のボトルネックはコンセンサスプロトコルだが，Parity ではトランザクションの署名がボトルネックである．

Fan ら (2020) はブロックチェーンの性能評価に関する調査を行なっている [4]．ブロックチェーンの性能評価手法にはモニタリングやベンチマーキング，実験，シミュレーションなどがある．パブリックブロックチェーンとプライベートブロックチェーンでは適している手法が異なる．例えばモニタリングはベンチマーキングと比べてパブリックブロックチェーンの評価に適しているが，ベンチマーキングはプライベートブロックチェーンの評価に適している．またモニタリングよりもベンチマーキングの方が拡張性が高い．

4. 実　　験

4.1 実験概要

本研究では Hyperledger Iroha と一般的なデータベースのパフォーマンスについての比較および，Hyperledger Iroha のパフォーマンスやスケーラビリティについて評価を行う．図 3 のように，Hyperledger Iroha ネットワークの各ピアは PostgreSQL をデータベースに持ち，ピア同士はネットワークで接続されている．

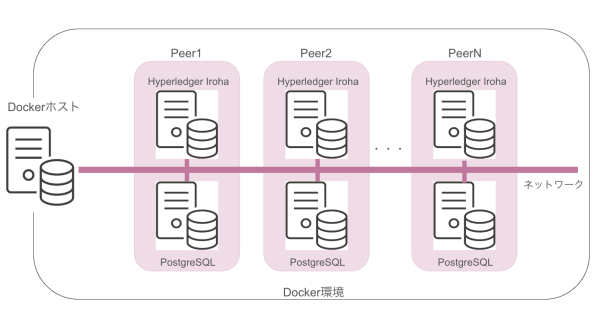


図 3: Hyperledger Iroha 実験環境

そこで PostgreSQL データベースにデータを追加する操作にかかる時間と Hyperledger Iroha の実行時間との比較を行う．PostgreSQL ではコネクションを確立し，1 つのクエリを処理した後，コネクションが切断されるまでにかかる時間を計測する．これは，Hyperledger Iroha の各ピアが各自のデータベースを書き換える際に行うことが想定される処理であり，この実験によって得られる PostgreSQL の実行時間と Hyperledger Iroha の実行時間の差は，Hyperledger Iroha のコンセンサス形成にかかる時間とすることができる．PostgreSQL データベースへの挿入操作にかかる時間の計測は，データベースにレコードが作成された時刻を記録するカラム created_at を作成し，挿入する際に現在時刻のタイムスタンプを取得し created_at に書き込むことで行った．また Hyperledger Iroha ネットワークを構成するピア数を 5，10，20，30 として，各ピア数のネットワークが 1 つのトランザクションを処理するのにかかる時間を計測する．Hyperledger Iroha の実行時間の計測にはアカウント A からアカウント B にアセットを転送するコマンド transferAsset を使用する．トランザクションが Hyperledger Iroha ネットワークに送信されてからトランザクションの処理が終わるまでの時間を JavaScript の performance.now() を用いて計測する．

次にブロックチェーン基盤の性能測定ツールである Hyperledger Caliper を用いて，Hyperledger Iroha への単位時間あたりの送信トランザクション数の変化およびネットワークサイズが変化したときのレイテンシとスループットの測定を行う．5，10，20，30 ピアから構成される Hyperledger Iroha ネットワークを対象とし，1 秒あたりに送信するトランザクション数を 10，20，30，40，50，60，70，80，90，100，150，200 と変化させて計測を行う．

4.2 実験環境

実験に使用したサーバの性能は Ubuntu20.04LTS の OS，Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz の CPU，コア数 32，スレッド数 64，メモリ 192GB である．

また PostgreSQL 単体および Hyperledger Iroha ネットワークを構成するピアは Docker コンテナを用いて構成される．Hyperledger Iroha の実験環境は図 3 の通りである．Hyperledger Iroha の各ピアは Hyperledger Iroha プロセスを実行するコンテナと PostgreSQL データベースを実行するコンテナから成

る．ピア同士は仮想ネットワークによって接続される．

使用した Hyperledger Caliper のバージョンは 0.3.2 である．

4.3 実験結果

4.3.1 PostgreSQL と Hyperledger Iroha の性能比較

PostgreSQL と Hyperledger Iroha でトランザクション 1 つの処理にかかる時間は表 1 のようになった．PostgreSQL では 10ms 程度で 1 つのクエリを処理できたのに対し，Hyperledger Iroha では 1 つのトランザクションの処理に約 570 倍の 5700s 程度の時間が必要となった．したがって Hyperledger Iroha での PostgreSQL へのデータの書き込みにかかる時間の割合はトランザクションの処理にかかる時間全体の約 0.2%であり，Hyperledger Iroha ではトランザクション処理時間のほとんどをコンセンサス形成に費やしていることがわかる．また Hyperledger Iroha では構成ピア数が増加しても，処理時間に大きな差が見られなかった．

表 1: PostgreSQL と Hyperledger Iroha での処理時間

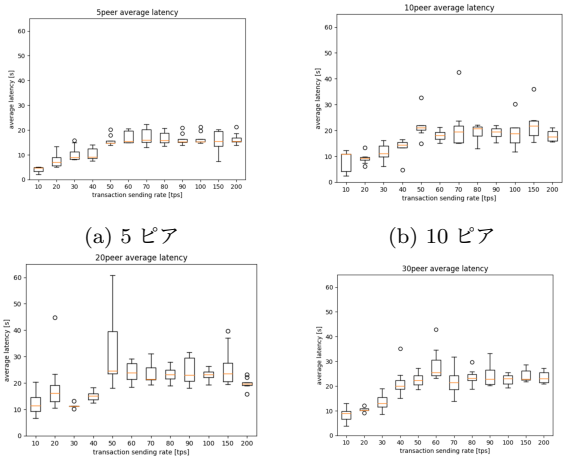
測定対象	測定結果 [ms]
PostgreSQL	10.3
Hyperledger Iroha 5peer	5696.27
Hyperledger Iroha 10peer	5641.27
Hyperledger Iroha 20peer	5777.56
Hyperledger Iroha 30peer	5804.01

4.3.2 トランザクション送信レートと平均レイテンシ

次に，Hyperledger Caliper を用いて Hyperledger Iroha ネットワークのレイテンシとスループットを計測した．

まずトランザクション送信レートを変化させたときの平均レイテンシについて評価を行う．図 4 に 5，10，20，30 の各ネットワークサイズに対してトランザクション送信レートを変化させたときの平均レイテンシの変化を示す．毎秒 50 トランザクションまではトランザクション送信レートの増加とともに平均レイテンシが増加する傾向が見られたが，50 トランザクション以降はトランザクション送信レートを増加させても平均レイテンシの値は横ばいとなった．

トランザクション送信レートが 50tps までのとき，トランザクション送信レートの増加に伴いネットワークが単位時間あたりに受け取るトランザクション数が増加したため，平均レイテンシが増加した．一方 50tps 以降ではトランザクション送信レートを増加させても平均レイテンシは増加せずに概ね一定の値となったため，単位時間あたりにネットワークに送信されるトランザクション数が一定になったと考えられる．よって単位時間あたりにオーダリングサービスからピアに送信されるプロポーザル数には上限が定められていると推測できる．オーダリングサービスが受け取るトランザクション数が一定値を超えると，超えた分のトランザクションはすぐにピアに送信されるのではなくオーダリングサービス内で管理され，時間をおいてピアに送信される．その結果送信されるトランザクション数が増加しても，単位時間あたりにピアが処理するプロポーザル数は一定になり，平均レイテンシの値が一定になったと考えられる．



(a) 5 ピア　(b) 10 ピア

(c) 20 ピア　(d) 30 ピア

図 4: トランザクション送信レートを変化させたときの平均レイテンシの変化

4.3.3 トランザクション送信レートとスループット

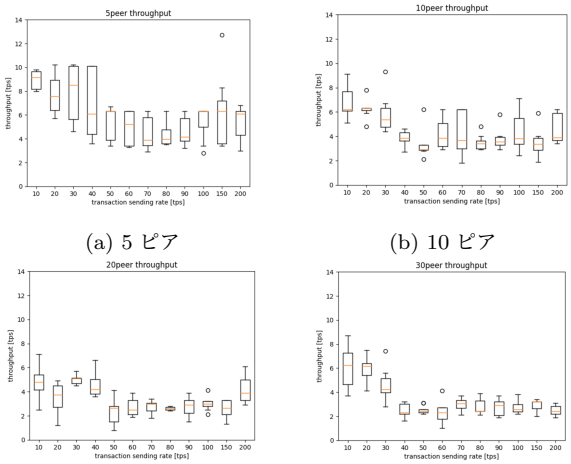
次にトランザクション送信レートを変化させたときのスループットの変化を図 5 に示す．ネットワークを構成するピアが 5 つのとき，単位時間あたりのスループット数は 70tps までは減少し，その後横ばいになるが 100tps 以降は増加した．10 ピア以降では，トランザクション送信レートが 30tps または 40tps まではスループットは減少し，30tps から 40tps 以降ではスループットの値に大きな変化は見られなかった．

トランザクション送信レートの値が小さいとき，トランザクション送信レートを増加させると単位時間あたりにネットワークに送信されるトランザクション数が増加する．したがって Hyperledger Iroha ネットワークにかかる負荷が大きくなり，スループットが低下したと考えられる．またトランザクション送信レートがある程度大きくなったとき，それ以上増加させてもスループットの値が横ばいとなった原因として，ピアに送信されるトランザクション数が一定になったことが考えられる．クライアントから送信されるトランザクション数が増加しても，単位時間あたりにピアに送信され処理されるトランザクション数が一定であったためスループットが一定になったと考えられる．

4.3.4 ネットワークサイズと平均レイテンシ

トランザクション送信レートを一定にして，ネットワークサイズを変化させたときの結果を図 6 に示す．トランザクション送信レートが 20tps 以下のときにネットワークサイズを大きくすると，20 ピアまでは平均レイテンシは増加したが 30 ピアでは減少した．トランザクション送信レートが 30tps 以上のときはネットワーク構成ピア数の増加とともに平均レイテンシも増加した．

ネットワークを構成するピア数が増えるとコンセンサス形成に参加するピア数が多くなるため，プロポーザルのコンセンサス形成時に必要なメッセージの総数が増加する．よってピア数が増えるにしたがって平均レイテンシも増加したと考えられる．



(a) 5 ピア　(b) 10 ピア

(c) 20 ピア　(d) 30 ピア

図 5: トランザクション送信レートを変化させたときのスループットの変化

4.3.5 ネットワークサイズとスループット

ネットワークを構成するピア数を変化させたときのスループットの変化を図 7 に示す．ネットワーク構成ピア数を増加させるとスループットは概ね低下した．ただしトランザクション送信レートが 10tps，20tps のときは 20 ピアのときスループットが最低となった．

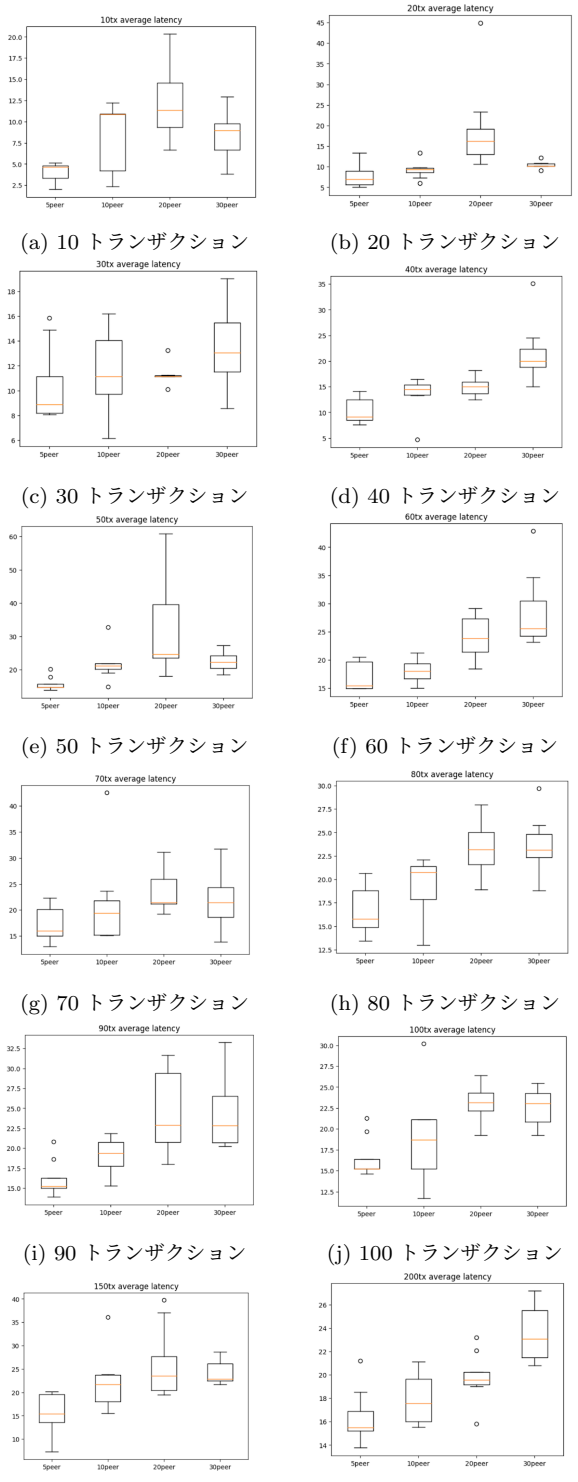
ネットワークサイズが大きくなると，コンセンサス形成時に送信されるメッセージ数が増加する．よって 1 つのプロポーザルの処理にかかる時間が増加し，スループットが低下すると考えられる．

また測定ではトランザクション送信レートが 10tps および 20tps のとき，30 ピアよりも 20 ピアの方が性能が低いという結果が得られた．Hyperledger Iroha を構成する各コンテナは CPU の各スレッドに割り当てられるためコンテナ内の処理は他のコンテナの影響を受けない．しかしネットワークや I/O は共有しているため，他のコアの処理のためにそれらに高い負荷がかかっている場合がある．したがって共有部分において他のコアを原因とする遅延が発生する可能性があり，測定された性能にある程度ランダム性が存在すると考えられる．

5. まとめと今後の課題

分散データベースのデータ検証プラットフォームとして導入が期待されるブロックチェーンについて，プライベートブロックチェーンの 1 つである Hyperledger Iroha を取り上げてその性能の評価を行った．

PostgreSQL とのデータ処理速度の比較により，Hyperledger Iroha では PostgreSQL よりもトランザクションの処理にかかる時間が非常に長いことがわかり，トランザクション処理のほとんど全ての時間がコンセンサスの形成に費やされているという結果が得られた．またトランザクション送信レートをある程度まで増加させていくと．平均レイテンシは高くなりスループットは低下することがわかった．トランザクション送信レートが一定値を超えると，それ以上単位時間あたりに送信するトランザクション数を増加させても平均レイテンシとスループットに大きな変化は見られなくなった．ネットワークサイズを変化させたとき，サイズを大きくすると平均レイテンシは増加し，スループットは低下することがわかった．



(a) 10 トランザクション　(b) 20 トランザクション

(c) 30 トランザクション　(d) 40 トランザクション

(e) 50 トランザクション　(f) 60 トランザクション

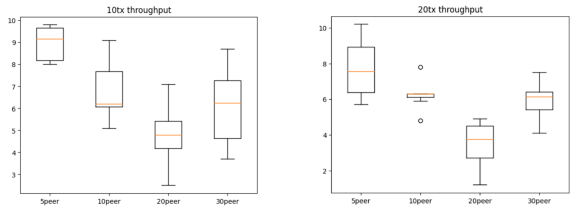
(g) 70 トランザクション　(h) 80 トランザクション

(i) 90 トランザクション　(j) 100 トランザクション

(k) 150 トランザクション　(l) 200 トランザクション

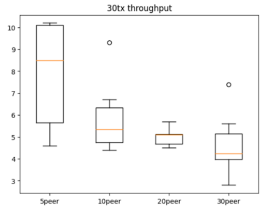
図 6: ネットワークサイズを変化させたときの平均レイテンシの変化

今後は仮想ネットワークではなく実ネットワークでピア同士を繋いで実験を行うことを検討している．また実験で使用した Hyperledger Iroha ネットワークを構成するピアは全て同等の

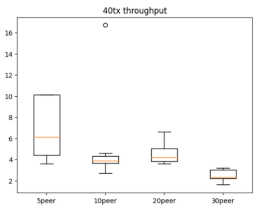


(a) 10 トランザクション

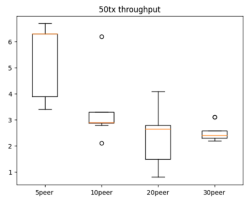
(b) 20 トランザクション



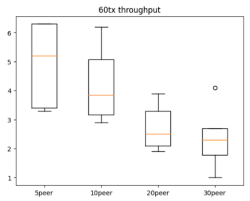
(c) 30 トランザクション



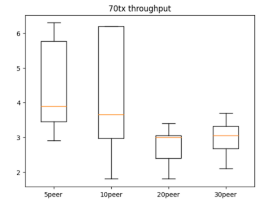
(d) 40 トランザクション



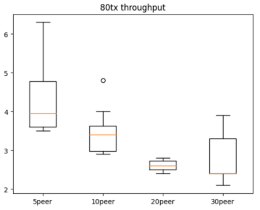
(e) 50 トランザクション



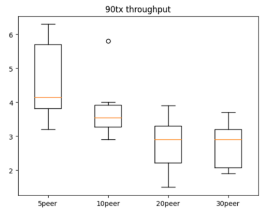
(f) 60 トランザクション



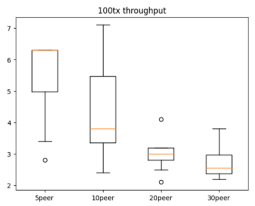
(g) 70 トランザクション



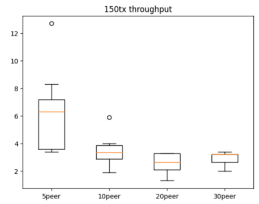
(h) 80 トランザクション



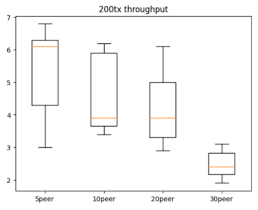
(i) 90 トランザクション



(j) 100 トランザクション



(k) 150 トランザクション



(l) 200 トランザクション

図 7: トランザクション送信レートを変化させたときのスループットの変化

性能を有していたが，同じネットワーク内に異なる性能を持つピアが存在したときの性能の変化についても評価を行いたい．

謝　　辞

本研究は一部，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Fedor Muratov, Andrei Lebedev, Nikolai Iushkevich, Bulat Nasrulin, Makoto Takemiya, “YAC: BFT Consensus Algorithm for Blockchain,” September 3, 2018.

[2] Arnold Woznica, Michal Kedziora, “Performance and Scalability Evaluation of a Permissioned Blockchain Based on the Hyperledger Fabric, Sawtooth and Iroha,” Computer Science and Information Systems, Vol 19, Issue 2, pp. 659-678, 2022.

[3] Tien Tuan Anh Dinh, Ji Wang, Gang Chen, Rui Liu, Beng Chin Ooi, Kian-Lee Tan, “BLOCKBENCH: A Framework for Analyzing Private Blockchainsi,” SIGMOD ’17: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1085–1100, May 2017.

[4] Caixiang Fan, Sara Ghaemi, Hamzeh Khazaei, Petr Musilek, “Performance Evaluation of BlockchainSystems: A Systematic Survey,” June, 2020.

[5] 齋藤 新, “4. 分散台帳技術におけるコンセンサス・メカニズム,” 情報処理, Vol 61, No. 2, pp. 165-175, 2022.

[6] 佐藤 栄一, “Hyperledger Iroha 入門ーブロックチェーンの導入と運営管理ー,” オーム社, 2020.

PostgreSQL の外部連携機能におけるレコード変換処理の高速化

柿村　直拓†　　立床　雅司†　　柴田　秀哉†

†三菱電機株式会社　情報技術総合研究所　〒247-0056 神奈川県鎌倉市大船５丁目１－１

E-mail:　†{ Kakimura.Naohiro@ak,　Tatedoko.Masashi@ap,　Shibata.Hideya@cb}.MitsubishiElectric.co.jp

あらまし　様々なデータソースを統一的に扱うために，PostgreSQL は Foreign Data Wrapper（FDW）と呼ばれる外部データ連携機能を備えている．FDW は PostgreSQL からのレコード要求に応じ，外部データソースからレコードを取得し，PostgreSQL の内部形式に変換する．この一連の処理において，レコード変換処理の比重が大きい場合，変換処理を並列化することで高速化が期待できるが，単一のデータソースに対する並列化の事例は知られていない．そこで本論文では，単一データソースに対するレコード変換処理の並列化機構を，商用データベース管理システム AQL 向けの FDW に対して実装し，評価により，変換処理の比重が大きい問合せが高速化されることを確認した．

キーワード　PostgreSQL，FDW，並列処理，データ変換

1. はじめに

IoT 関連技術が発展したことにより，大規模なデータを蓄積し，分析することがより身近になった．例えば，製造現場では設置されたセンサからデータを蓄積し，分析することで生産効率向上に努めている．これに伴い，データの取得，蓄積，参照方法は複雑化しており，データソースも多様化傾向にある．一般的にデータ蓄積に使われる RDBMS (Relational Database Management System) は，多くのユーザが同時に少量データを更新，検索するような OLTP (Online Transaction Processing) 処理を得意とする反面，少ないユーザがデータ全体を参照するような OLAP (Online Analytical Processing) 処理は不得手である．このように，用途に応じてデータベースを使い分ける必要があるが，異なるデータソースに対しデータを参照，分析するのは手間がかかるため，標準的な SQL でデータを統一的にアクセスできるインターフェースが求められる．

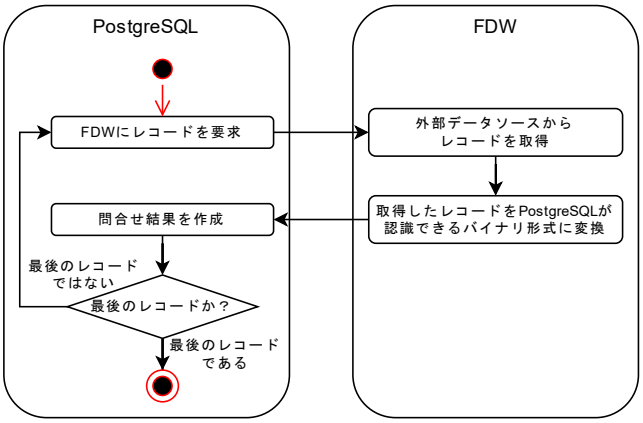


図　1 PostgreSQL における FDW を活用した外部データソースのレコード取得処理の流れ

OSS の RDBMS である PostgreSQL[1]は，標準的な SQL のインターフェースを持ち様々な分野で利用されている．PostgreSQL は OLTP 処理を実現するためにマルチプロセスで複数の問合せを受け付け，各問合せはシングルプロセスかつシングルスレッドで実行される．また，PostgreSQL は外部データソースからレコードを取得できる機能である Foreign Data Wrapper (FDW)[2]を備えている．

PostgreSQL が外部データソースにアクセスする問合せを受け付けた時，PostgreSQL は FDW にレコードを要求する（図 1）．FDW は外部データソースからレコードを取得して，PostgreSQL が処理できるバイナリ形式に変換して，PostgreSQL にレコードを渡す．その後，PostgreSQL は問合せ結果の作成などを直列に行う．そのため，FDW は変換処理完了後に再度 PostgreSQL からレコード要求を受けるまで，次のレコード取得を行うことができず待ち時間が発生する．特に，レコード変換に文字コード変換などの時間を要する処理を含む場合，この問題は顕著となる．この問題に対して複数データソースを用いた並列化の試みはあるが、単一の外部データソースに対する事例は知られていない．

そこで本論文では，単一の外部データソースに対するデータ変換を効率化するため，データ変換処理を並列化する方式を提案する．また，提案方式を商用データベース管理システム AQL[3]向けの FDW に対して実装し，評価を実施した．本論文では，2 章で関連研究，3 章で提案方式，4 章で提案方式の評価について説明し，5 章で総括を行う．

2. 関連研究

2.1. FDW の非同期実行

PostgreSQL サーバを外部データソースとして扱う FDW モジュール postgres_fdw には，非同期実行機能が

ある[2]．この機能では，シャーディング等で外部データソースを分割可能な場合に外部データソースアクセスを非同期化し，外部データソースからのレコード取得を並列化することで，処理を高速化している．

2.2. パラレルクエリ

PostgreSQL には，パラレルクエリ[4]と呼ばれるマルチプロセスを活用した実行計画の作成・実行機能がある．この機能では，PostgreSQL のテーブルに対して，レコード取得の並列処理が可能である．

2.3. 複数データソースの並列レコード取得

文献[5]では，PostgreSQL から異なる種類の複数外部データソースに連携してアクセス可能とするための FDW が提案されている．ここでは，各データソースに対してスレッドを立ち上げ，レコード取得・変換処理を並列化している．また，各スレッドは PostgreSQL とは独立に動作し，PostgreSQL からレコードが要求される前に FDW で事前にレコード取得・変換処理を実施することで高速化している．

2.4. 課題

本章で説明した既存の方式では，以下の 2 点の性質を持つ単一データソースを PostgreSQL と連携する場合の高速化手段として不十分である．

1. レコード取得処理が並列化できない．
2. レコード変換処理の時間が，PostgreSQL 内部の処理やレコード取得時間と比較し長い．

postgres_fdw による非同期実行は，単一の外部データソースに対しては直列的な処理となる．パラレルクエリは，外部データソースの性質 1 により適用できない．文献[5]の方法では，PostgreSQL からのレコード要求前にレコード取得・変換処理を開始するため，一定の高速化効果は見込めるが，単一の外部データソースに対してはレコード変換処理が並列化されないため，性質 2 によりボトルネックが解消されない問題がある．

3. 提案方式

本章では，単一の外部データソースであっても，PostgreSQL のレコード要求に対しての待ち時間を削減可能な方式を提案する．提案方式では，レコード変換処理を並列化するため，レコード変換処理とレコード取得処理を非同期に実行する．

図 1 において，PostgreSQL が FDW にレコードを要求し，受け取ったレコードを使用して問合せ結果を作成するまでの一連の処理，FDW のレコード取得処理，FDW のレコード変換処理の 3 つの処理は非同期に実行可能である．非同期処理はマルチスレッドにより実現する．各処理を実行するスレッドを，以下では順にメインスレッド，取得スレッド，変換スレッドと呼ぶ．

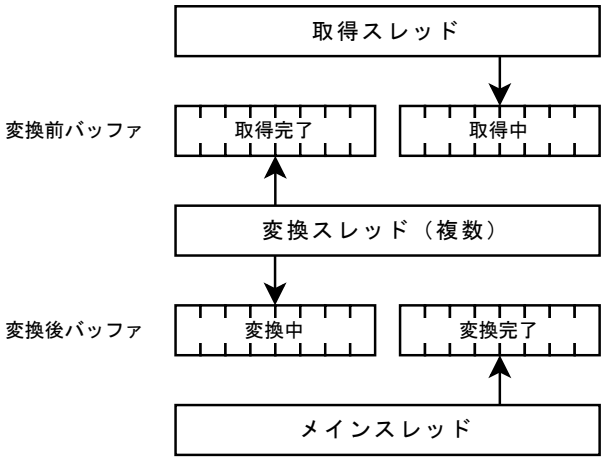


図 2 各スレッドとバッファの関係

提案方式では，変換処理におけるボトルネックを解消するために，変換スレッドを複数作成し並列化する．変換処理の並列化に伴い，複数レコードを同時に変換することになるため，取得処理では，複数レコードを一括取得する必要がある．スレッド間でデータをやり取りするために，取得されたレコードは変換前バッファに一時保管され，変換されたレコードは変換後バッファに保管される（図 2）．各スレッドを切れ目なく動作させるため，各バッファを 2 つずつ用意し，一方を書込み用，他方を読込み用とする．書込み用と読込み用は交互に役割を入れ替える．

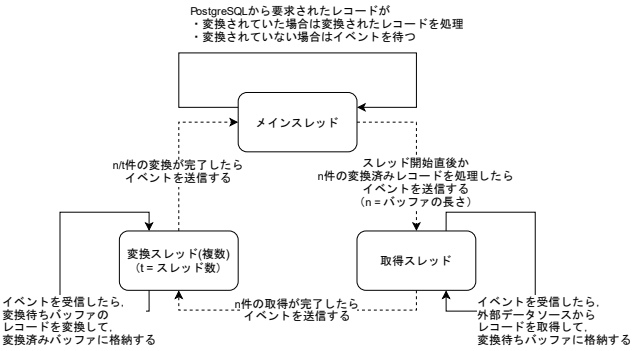


図 3 スレッド間のイベント完了通知

各処理は非同期に実行可能であるが，各スレッドは互いの処理結果に依存している．例えば，変換スレッドは取得スレッドがレコード取得を完了しなければ処理を開始できず，メインスレッドは変換スレッドがレコード変換を完了するまで問合せ結果の作成ができない．そこで，スレッド間の処理完了通知を実現するため，提案方式ではイベント駆動型アーキテクチャで各スレッドを連携させる（図 3）．各スレッドはイベントを受信したら処理を開始し，処理が完了したらイベン

トを送信して待機する．この方式では，イベントの送受信や待機開始･解除のオーバーヘッドがかかるため，一括取得するレコード数は変換スレッド数よりも多くなければ効果が見込めない．

4. 評価

本章では提案方式の有効性を評価するため，非同期化・並列化を実施していない素朴な FDW を比較方式として，問合せ実行時間を比較する．

4.1. 評価環境

評価環境を表 1 に記載する．なお，PostgreSQL と外部データソースは同じサーバに存在するとする．外部データソースとして，商用のデータ分析フレームワーク AnalyticMart[6]で使用されているデータベース管理システム AQL を対象とする．

表 1 評価環境

CPU	Intel Corei5-10500 6 コア/12 スレッド/3.1GHz/
OS	Windows Server 2016 standard
RAM	16GB (8GBx2) DDR4 DIMM 2666MT/s
SSD	512GB SSD (M.2 NVMe PCIe TLC)
PostgreSQL	11.12 版，UTF-8
外部データソース	商用データベース AQL

4.2. 評価データ

評価に使用するデータは，外部データソース保管された 100 万件のレコードデータである．外部データのスキーマを図 4 に示す．

```
CREATE FOREIGN TABLE foreign_table (
      c_integer           INTEGER,
      c_bigint            BIGINT,
      c_numeric1809       NUMERIC(18, 9),
      c_char0004          CHAR(4),
      c_char0032          CHAR(32),
      c_char0128          CHAR(128),
      c_varchar0004       VARCHAR(4),
      c_varchar0032       VARCHAR(32),
      c_varchar0128       VARCHAR(128),
      c_date              DATE,
      c_timestamp6        TIMESTAMP(6)
) SERVER xx_FDW;
```

図 4 評価用データのスキーマ

各データ型の変換方法について説明する．これらの方法はレコード変換処理時間に大きく影響する．INTEGER, BIGINT, CHAR, VARCHAR 型については，特別な形式変換を行わず，外部データソースから取得したデータをそのままコピーする．但し，対象の外部データソースにおける文字列型のエンコーディングは Shift_JIS 又は UTF-8 が選択可能であるが，Shift_JIS が選択されていた場合は，UTF-8 への変換処理を追加で行う．NUMERIC，DATE，TIMESTAMP 型は文字列として外部データソースから出力され，それらの文字列を解析して，PostgreSQL で扱えるバイナリ形式に変換する．

4.3. 評価方法

評価では，比較方式と提案方式との問合せ実行時間を，データ型による違いとデータ変換処理の並列度による違いの観点から比較する．データ型による違いでは，図 4 のテーブルにおける各データ型の問合せ実行時間を提案方式（変換スレッド数＝4）と比較方式で比較する．並列度による違いでは，変換処理が短いデータ型（INTEGER 型）と長いデータ型（Shift_JIS の VARCHAR 型）で，並列度（変換スレッド数＝1, 2, 4）を変えて比較する．両評価において，変換前，変換後バッファの長さはそれぞれ 1,000 レコード分とする．

測定は PostgreSQL のクライアントツール psql の ¥timing コマンドを使用し，null 出力する．また，問合せは外部データソースから全件参照するようなクエリを使用する（図 5）．このクエリの count 関数は，問合せ結果として，大量のデータ書出しが発生しないよう意図したものである．また，count 関数内の null 確認は,外部データソースで実行できない処理であるため，プッシュダウンができず，全レコードを取得しなければならない処理となっている．このように外部データソースからのレコードを全件取得しつつ，問合せ結果の書き出し時間を抑制する．

```
SELECT count(c_integer is null) FROM foreign_table;
```

図 5 評価に使用するクエリ例

4.4. 評価結果

図 6 に各データ型を取得した際の，比較方式に対する提案方式の問合せ実行速度を相対速度として示す．結果から INTEGER 型, BIGINT 型以外は高速化されていることが確認できる．特に，CHAR 型(Shift_JIS), VARCHAR 型(Shift_JIS), TIMESTAMP 型の変換において，2 倍以上の高速化効果が確認できる．この結果から文字コード変換や文字列を解析するような変換を行う場合は，高速化されることが確認できた．一方，INTEGER 型，BIGINT 型のように変換処理の比重が小さい（4, 8 バイトのデータのコピー）場合は，並列化のオーバーヘッドが大きく，高速化されないことを確認した．

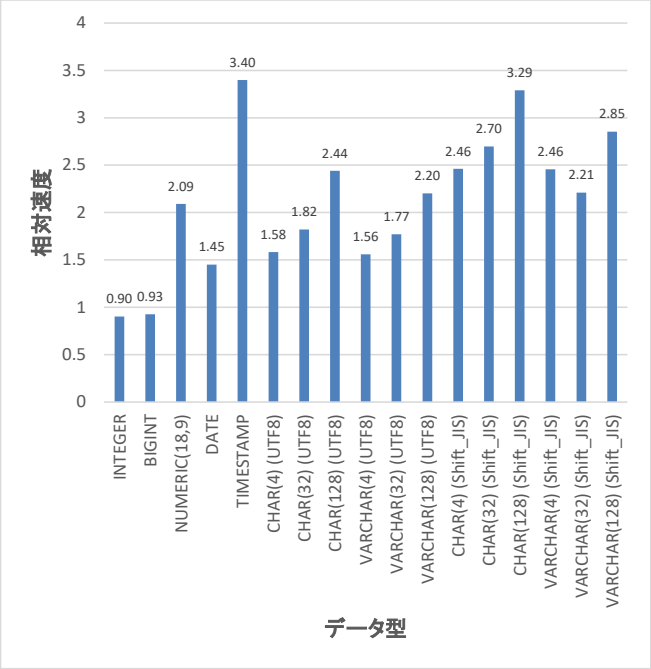


図 6 比較方式に対する提案方式の相対速度
（データ型毎，並列度 4）

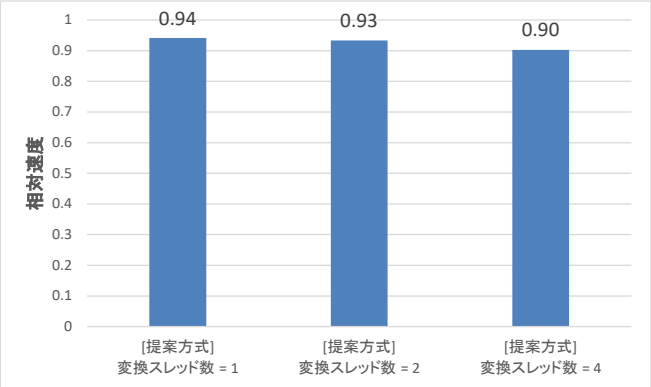


図 7 比較方式に対する提案方式の相対速度
（並列度毎，INTEGER 型参照時）

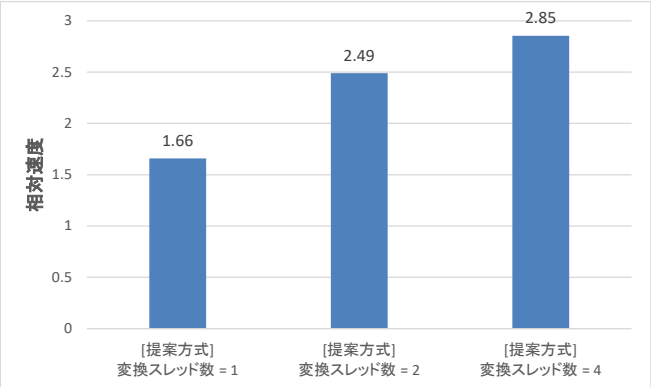


図 8 比較方式に対する提案方式の相対速度
（並列度毎，VARCHAR(128)型 (Shift_JIS)参照時）

図 7 と図 8 に，並列度を変更した時の問合せ実行時間の相対速度を示す．図 7 では，変換処理の比重が小さい INTEGER 型，図 8 では，変換処理の比重が大きい文字コード変換ありの VARCHAR(128)型 (Shift_JIS)の実行時間を示す．INTEGER 型に関しては，並列度が高くなるにつれて提案方式が遅くなっている．これは，前の評価と同じく並列化のオーバーヘッドに起因するものと考えられる．VARCHAR 型に関しては，並列度が高くなるほど高速化されていることが確認できる．但し，並列度が高くなるほどスレッドあたりの高速化効果は減少している．これは，レコード変換処理時間が並列化によりレコード取得処理時間やメインスレッドの処理時間と近い時間になったためであると考えられる．

5. おわりに

本論文では，FDW を介した PostgreSQL と外部データソースとの連携において，外部データソースからのレコード取得処理とレコード変換処理を非同期に動作させ，更に変換処理を並列化する方式を提案し，評価を実施した．その結果，レコード変換処理の比重が大きい問合せに関しては，並列化による高速化の効果を確認した．一方，レコード変換処理の比重が小さいものに関しては，並列化のオーバーヘッドが顕在化し，高速化効果は得られず，若干処理時間が増大する結果となった．双方のケースにおいて，内部で最適な処理方法に自動的に切り替えられるようにすることが，今後の課題である．

参 考 文 献

[1] PostgreSQL: The world's most advanced open source database
https://www.postgresql.org/

[2] F.38. postgres_fdw (postgresql.jp)
https://www.postgresql.jp/document/15/html/postgres-fdw.html

[3] 佐藤重雄，塚本純子，清水英弘，石井篤，藤原聡子，“多次元明細データベース DIAPRISM/AQL の概要”，情報処理学会全国大会講演論文集 55.3(1997): 501-502

[4] PostgreSQL: Documentation: 16: Chapter 15. Parallel Query
https://www.postgresql.org/docs/current/parallel-query.html

[5] 片山大河，嶋村誠，金松基孝，“PostgreSQL における複数外部データソース並列スキャン機能と即時結果取得機能の開発”，電子情報通信学会技術研究報告 117.212 (2017): 75-79.

[6] データ分析フレームワーク AnalyticMart | プラットフォームソリューション | サービス・製品 | 三菱電機インフォメーションネットワーク株式会社（MIND）
https://www.mind.co.jp/service/idc_platform/platform_products/analyticmart/index.html

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

2024年2月28日(水) 16:00 ~ 18:10 | T2-B オンライン（Zoom Events)

機械学習とプライバシ

座長:高橋 翼(LINEヤフー株式会社)
コメンテータ:佐々木 勇和(大阪大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T2-B-3-01] ソースフリードメイン適応設定における重み付きクラスター連合学習

*森 隼基[1,2]、木原 孝輔[1]、寺西 勇[1] (1. NEC セキュアシステムプラットフォーム研究所、2. 京都大学)

16:30 ~ 16:55
[T2-B-3-02] 個人化されたユーザレベルの差分プライバシ保護型連合学習

*加藤 郁之[1]、高木 駿[1]、Xiong Li[2]、曹 洋[3]、吉川 正俊[4] (1. 京都大学、2. Emory University、3. 北海道大学、4. 大阪成蹊大学)

16:55 ~ 17:20
[T2-B-3-03] 非独立同分布の学習データに対する分割学習の性能向上に関する研究

*池奥 裕太[1]、曹 洋[2]、吉川 正俊[3] (1. 京都大学大学院情報学研究科、2. 北海道大学大学院情報科学研究院、3. 大阪成蹊大学データサイエンス学部)

17:20 ~ 17:45
[T2-B-3-04] エンコーダを用いた不均一モデル構造間の連合学習

*松永 悠斗[1]、籾山 悟至[1]、ENKHTAIVAN BATNYAM[1]、寺西 勇[1] (1. 日本電気株式会社)

17:45 ~ 18:10
[T2-B-3-05] 【技術報告】安全なLLMの実現を目指して - ストレステストとセーフアラインメントの取り組み

綿岡晃輝 (LINEヤフー株式会社)

ソースフリードメイン適応設定における重み付きクラスター連合学習

森　隼基†,††　木原　孝輔†　寺西　勇†

† NEC セキュアシステムプラットフォーム研究所　〒 211-0011 神奈川県川崎市中原区下沼部 1753
†† 京都大学情報学研究科　〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{junki.mori,kosuke-kihara,teranisi}@nec.com

あらまし　近年，機械学習の発展とプライバシー保護への関心の上昇から，複数のクライアントが互いに生データを共有せずに協力して機械学習モデルの学習を行う連合学習が注目されている．画像認識タスク等への適用を考えると，実運用上各クライアントが保有する画像のドメインが異なる場合や，アノテーションコストの問題からラベル無しデータしか保有しない場合が考えられる．本研究では，ソースドメインのラベル付きデータで事前に学習されたモデルを，ソースデータにアクセスすることなく，各クライアントが保有するターゲットドメインのラベル無しデータにドメイン適応させる，連合ソースフリードメイン適応の問題に取り組む．本研究の提案手法では，複数のクライアントをプライバシーを損なうことなく類似ドメインごとにクラスタリングし，各クラスター内でモデルの統合を行う．さらに，各クライアントが，自身が属さない他のクラスター内のモデルも独自の新たな重み付けで活用することで，提案手法が既存手法の精度を上回ることを確認した．

キーワード　連合学習，ドメイン適応，ソースフリー，画像認識

1　はじめに

近年，機械学習におけるデータプライバシーへの懸念が高まる中，連合学習（Federated Learning）[1] が分散データに対するプライバシー保護型機械学習フレームワークとして近年注目を集めている．連合学習では，プライベートデータを持つ複数の参加者，すなわちクライアントがローカルに学習したモデルパラメータを全体の指揮を執る中央サーバーで統合する操作を複数回繰り返すことで，生データを共有することなく高性能なモデルの学習を行うことが出来る．

既存の連合学習研究のほとんどは，各クライアントがラベル付きデータを保有していることを仮定しているが，画像タスクへの応用などを考えると，人手のアノテーションには時間やドメイン知識が必要となるため，このような仮定を満たすことは困難である．例えば，複数地点に設置された監視カメラへの連合学習応用を考えたとき，各監視カメラの管理者が監視映像それぞれにアノテーションすることは非常に労力がかかりコストが高い．組織によって運営される場合が多い中央サーバーはラベル付きデータを入手できる可能性が高いが，そのようなラベル付きデータには顔等の生体情報やプライバシー保護を必要とする情報が多分に含まれている可能性があり，それらを連合学習に直接使用するのは困難な場合がある．また，監視カメラの設置地点によって明るさや背景等ドメインが異なるという困難性も存在する．

上記のような例を考慮し，本論文ではより実践的であるが十分に検討されていない連合学習設定である，連合ソースフリードメイン適応（Federated source-Free Domain Adaptation, FFREEDA）[2] に取り組む（図 1 を参照）．FFREEDA では，サーバーはあるソースドメインのラベル付きデータで事前に学習されたモデルを保有している．ただし，モデルの学習に使用されたソースデータはプライバシー保護の観点や，既に棄却されている等の問題により以降の連合学習では使用できない．これはソースデータにアクセスできないドメイン適応である，ソースフリードメイン適応（Source-Free Domain Adaptation, SFDA）[3], [4] における一般的な設定である．各クライアントは，ソースドメインとは異なる複数のターゲットドメインの内の一つのドメインにおける少数のラベル無しデータのみを保有している．FFREEDA の目標は，各クライアントが他のクライアントの知識も活用しながら，ソースモデルに基づいて自身のドメインに適合した分類モデルを訓練することである．FFREEDA の困難性は以下のようにまとめられる：(1) クライアント間でドメインが異なる（ドメインシフト），(2) 事

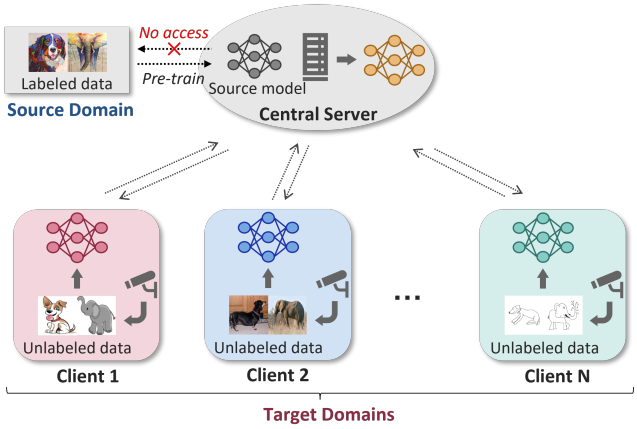


図 1:　本論文で扱う連合ソースフリードメイン適応（FFREEDA）の外観．サーバーはソースドメインで事前に学習されたソースモデルを保有し，各クライアントはターゲットドメインにおけるラベル無しデータを保有．

前学習フェーズ後の連合学習フェーズが完全に教師なし学習である．

上記 2 つの困難性を同時に扱っている研究はほとんどない．パーソナライズド連合学習 [5] や，クライアントをデータ分布に応じてクラスタリングするクラスター連合学習 [6] は分布シフトの問題に対処しているが，これらはラベル付きデータの存在を前提としており，かつ主にラベル分布シフトに注力しいる．教師なし設定は，半教師あり連合学習 [7], [8] や，連合ドメイン適応 [9], [10] が関連しているが，これらはサーバーや一部のクライアントが連合学習に使用可能なラベル付きデータを保有していることを仮定している．両問題を含む FFREEDA に取り組んでいる研究は，我々の知る限り Shenaj らによる LADD [2] のみである．ドメインシフトによる性能劣化に対処するため，LADD はクライアントをドメインに応じてクラスタ化し，各クラスタ内でモデルパラメータを統合する．このクラスタリングを，各クライアントが保持する生画像データから抽出したスタイル特徴を利用することで，ラベル無しデータのみで実現する．しかし，LADD には次のように大きく 2 つの問題を抱えている：(1) 各クライアントが生の画像のスタイルを特徴量としてサーバーに共有する必要があり，プライバシーリスクが生じ得る，(2) 各クラスター内のみで知識の統合が行われ，各クライアントは他のドメインからなるクライアントの知識を活用することが出来ない．

本論文では，このような問題点を解決するための新しい FFREEDA 手法として，クライアントをプライバシー保護しつつかつパラメータフリーでクラスタリングし，各クラスタ内で統合されたモデルを各クライアントに応じた重み付けで結合する Federated learning with Weighted Clustered Aggregation (FedWCA) を提案する．具体的には，FedWCA はまず各クライアントが第 1 ラウンドでローカルにドメイン適応させた特徴抽出器の第 1 層のパラメータに基づいてクライアントをクラスタリングする．FedWCA のクライアントクラスタリング手法では，各クライアントからの生の画像情報だけでなく，最新のクラスタリングアルゴリズムの適用によりクラスタ数などのハイパーパラメータも必要としない．各クライアントの性能を向上させるために，FedWCA は重み付き統合戦略を採用する．各クライアントは，各クラスタ内のローカルモデルを統合して作成された全てのクラスタモデルをダウンロードし，自身の利益と全体の利益の両方を考慮しながら，ラベル無しデータのみで計算された固有の重み付けでそれらを線形結合させる．その結果，各クライアントは，他のドメインのクライアントの知識も活用しながら，自身のドメインにうまく適応したパーソナライズされたモデルを得ることができる．実験により，FedWCA が既存手法を含む様々なベースラインの精度を上回ることを確認した．

2 関 連 研 究

2.1 ソースフリードメイン適応

本研究のように，ターゲットドメインのデータがラベルを持たない場合のドメイン適応は，教師なしドメイン適応と呼ばれる [11], [12]．さらに近年のプライバシー保護の観点から，多くのドメイン適応がソースデータを用いないソースフリー設定で研究されている．SHOT [3] をはじめとする研究では，学習途中のモデルに特徴量を入力して得られた出力を疑似ラベルとして用いる自己教師あり学習を行っている．NRC [13] や，G-SFDA [14]，AaD [15] では，あるデータとその近傍にあるデータは似た予測を行うという考えのもと，疑似ラベルを修正するような手法を用いている．

2.2 連合ドメイン適応

Peng ら [16] は，ドメイン適応を初めて連合学習設定で行ったが，ドメインの特徴量表現の整合にソースドメインのデータを用いている．その他の連合ドメイン適応 [17], [18] も，完全なソースフリー設定では実験されていない．唯一の例外は Shenaj らが提案した LADD [2] であり，画像内の物体を対象ごとにピクセル単位で分類するセマンティックセグメンテーションのタスクで，ソースフリードメイン適応を連合条件のもとで実施した．

3 問 題 設 定

本節では，本論文で扱う連合ソースフリードメイン適応（Federated source-Free Domain Adaptation, FFREEDA）を定式化する．

FFREEDA では，中央サーバーとクライアント集合 $\mathcal{K} = \{1, 2, \ldots, K\}$ が存在する． タスクとして，特徴量空間を $\mathcal{X} \subset \mathbb{R}^d$，ラベル空間を $\mathcal{Y}$ とした M クラス分類を考えよう． サーバーはソースドメインにおけるラベル付きデータセット $D_S = \{x_S^i, y_S^i\}_{i=1}^{N_S} \sim \mathcal{P}_S(\mathcal{X}, \mathcal{Y})$ を保有しており，各クライアント $k \in \mathcal{K}$ は複数のターゲットドメインにおけるラベル無しデータセット $D^k = \{x^{k,i}\}_{i=1}^{N^k} \sim \mathcal{P}^k(\mathcal{X})$ を保有している．ここで，$\mathcal{P}_S$ と $\mathcal{P}^k$ はそれぞれサーバーとクライアント k のデータ分布を表している．各クライアントとサーバーには一つのドメインが割り当てられ，サーバー・クライアントとクライアント間でそれぞれドメインシフトがあると仮定する．すなわち，任意の $k \in \mathcal{K}$ で $\mathcal{P}_S(\mathcal{X}) \neq \mathcal{P}^k(\mathcal{X})$，ある $k, l \in \mathcal{K}$ で $\mathcal{P}^k(\mathcal{X}) \neq \mathcal{P}^l(\mathcal{X})$ が成り立つ．ただし，同じドメインのクライアントが複数いても良いとする．

サーバーは D_S でソースモデル $h_S \colon \mathcal{X} \to \mathbb{R}^M$ を学習し，今後連合学習で使用できないよう D_S は棄却するかプライベートに保管しておく． h_S は特徴量抽出器 $f_S \colon \mathcal{X} \to \mathbb{R}^q$ と分類器 $g_S \colon \mathbb{R}^q \to \mathbb{R}^M$ の 2 つのモジュールからなることを仮定する．特徴量抽出器 f_S はサンプル x を q 次元の特徴量ベクトル $f_S(x) \in \mathbb{R}^q$ エンコードし，分類器 g_S は $f_S(x)$ に各クラス m に対する確信度 $(g_S(f_S(x)))_m$ を割り当てる．また，画像分類タスクにおける既存研究に従い，ニューラルネットワークの最終線形層を分類器とする．そこで，$W = [w_1, \ldots, w_M] \in \mathbb{R}^{q \times M}$ を，g_S を表す重みベクトルとする．各 w_m をクラス m に対応する分類器ベクトルとする．

FFREEDA の目標は，ソースモデル h_S から出発して各クライアントのデータ分布 $\{\mathcal{P}^k(\mathcal{X})\}_{k\in\mathcal{K}}$ に最適な複数の分類モデルを学習することである．ゆえに，FFREEDA における目的関数は，クライアント k の適切なローカルの損失関数を $\mathcal{L}_k$ として，次のように書ける:

$$\underset{h^1,\ldots,h^K}{\arg\min} \sum_{k\in\mathcal{K}} \frac{N^k}{N}\mathcal{L}_k(h^k). \tag{1}$$

ただし，$N=\sum_{k\in\mathcal{K}} N^k$ である．上記の最適化問題を解くために，我々は通常の連合学習プロトコルに従うこととする．すなわち，各通信ラウンド $r=0,\ldots,R-1$ において， 各クライアントはローカルに自身のローカルモデル h^k を E ローカルエポック分学習し，それをサーバーに送信する．サーバーは受信したモデルパラメータを統合することでグローバルモデルを作成する．

4 提案手法

FFREEDA に対する我々の提案手法である，Federated learning with Weighted Clustered Aggregation (FedWCA) の説明を行う．まずは我々の手法における設計の動機から始めよう．

まずはほとんどの既存 SFDA 手法 [3] に従い，ソースモデルの分類器部分のパラメータは連合学習中固定する．これは，ソースモデルの分類器がドメイン不変な分類能力を獲得していると考えられるからである．したがって，各ラウンドでローカルにドメイン適応された特徴量抽出器は各クライアントのドメインに特化される．統合方法として FedAvg [1] (単純平均) を各クライアントの特徴量抽出器に適用すると，ドメインシフトによる精度劣化につながる．これは，LADD [2] のように各クライアントをドメインに応じてクラスタリングし，各クラスター内のみでモデルを統合する動機となる．

しかし，LADD では，各クライアントがクラスタリングのために画像スタイルの特徴をサーバに送信する必要があり，プライバシーに関する懸念が生じる．さらに，他のクラスター連合学習と同様に，LADD は反復回数やクラスタ数の探索範囲などのハイパーパラメータを設定する必要があり，場合によっては実用的でない可能性がある．我々は，ローカルに学習されたパラメータに基づく，プライベートかつパラメータフリーのクライアントクラスタリング手法を提案する．また，LADD では全クライアントで FedAvg を適用する部分と各クラスタ内でのみ FedAvg を適用する部分にモデルを分割するが，その分割の探索が困難であり，前者では上述の問題が発生し，後者では他のクラスタの知識を利用できない．そこで，各クラスタ内で統合されたモデルを，全体の利益と各クライアント自身の利益の両方を考慮しながら，クライアント固有の重みで結合する統合方法を提案する．

4.1 疑似ラベルを利用したローカル適応

まずは我々の手法におけるベースとなる各クライアントのローカル学習について説明する．任意の SFDA 手法を適用することが可能であるが，各クライアントの限られた計算・メモリ資源を考慮し，既存手法で一般的な単純だが効果的なプロトタイプベースの疑似ラベル手法を用いる．

各ラウンドにおいて各クライアントは，まずサーバーから受信したモデルに基づいてローカルの初期モデル $h^k_{init}=g^k_{init}\circ f^k_{init}$ を設定する．第 1 ラウンドでは初期モデルはソースモデルとする．第 2 ラウンド以降の初期モデルの設定方法は 4.2 節で説明する．特徴量抽出器 f^k_{init} を利用して，各クラスにおける特徴量抽中心を算出し，各 $x\in D^k$ に対して最も近いクラス中心のクラスを疑似ラベル $\hat{y}$ とする（詳しくは [3] を参照）．すなわち各クライアントは $\hat{D}^k=\{x^{k,i},\hat{y}^{k,i}\}_{i=1}^{N^k}$ を得る．各ラウンドにおいて一度 $\hat{D}^k$ を定めたら，それをラウンド中固定して用いることで，h^k_{init} が備える他クライアントの知識を保持し，過学習を防ぐ．

損失関数としては SFDA 手法において一般的な Information Maximization (IM) 損失を用いる:

$$\mathcal{L}^k(f^k) = \mathcal{L}_{IM}(h^k;D^k)+\lambda\mathcal{L}_{CE}(h^k;\hat{D}^k), \tag{2}$$

$$\mathcal{L}_{CE}(h^k;\hat{D}^k) = -\frac{1}{N^k}\sum_{(x^k,\hat{y}^k)\in\hat{D}^k}\log\left(h^k(x^k)\right)_{\hat{y}^k},$$

$$\mathcal{L}_{IM}(h^k;D^k) = \sum_{m=1}^{M}\hat{p}_m\log\hat{p}_m - \frac{1}{N^k}\sum_{m=1}^{M}\sum_{x^k\in D^k}(h^k(x^k))_m\log\left(h^k(x^k)\right)_m.$$

ここで，$h^k=g_S\circ f^k$ はクライアント k のローカルモデル，λ は調整パラメータ， $\mathcal{L}_{CE}$ は疑似ラベルに対するクロスエントロピー損失，$\mathcal{L}_{IM}$ は IM 損失， $\hat{p}=\sum_{x^k\in D^k}h^k(x^k)/N^k$ は出力スコアの平均である．IM 損失は，モデルが確信度の高いスコアを出力することを促し，あるクラスに偏ったトリビアルなスコアを出力することを防ぐ効果がある．

4.2 クライアントクラスタリング方法

次に，プライバシーリスクを上昇させず，かつパラメータの事前設定をせずに各クライアントをクラスタリングする方法を説明する．生データに関する情報をサーバーに共有せずにクラスタリングするために，第一ラウンドで各クライアントがローカルに学習した特徴量抽器 $\{f^k\}_{k\in\mathcal{K}}$ のパラメータに基づいてクラスタリングする戦略を採用する．既存のクラスタ連合学習はクラスター数といったハイパーパラメータを必要とするが，連合学習におけるチューニングコストの高さを考慮し，パラメータフリーな最新のクラスタリングアルゴリズムである FINCH [19] を利用する．具体的にはクライアント間の距離を，モデルパラメータ間のコサイン類似度を用いて算出し，各クライアント k ごとに第一近傍のクライアント κ_k（k と最も距離が近いクライアント）を探索する．そして，下記のような隣接行列を定義する．

$$A(k,l)=\begin{cases}1 & \text{if } l=\kappa_k \text{ or } k=\kappa_l \text{ or } \kappa_k=\kappa_l;\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

上記隣接行列によって隣接したクライアントを同クラスタに割

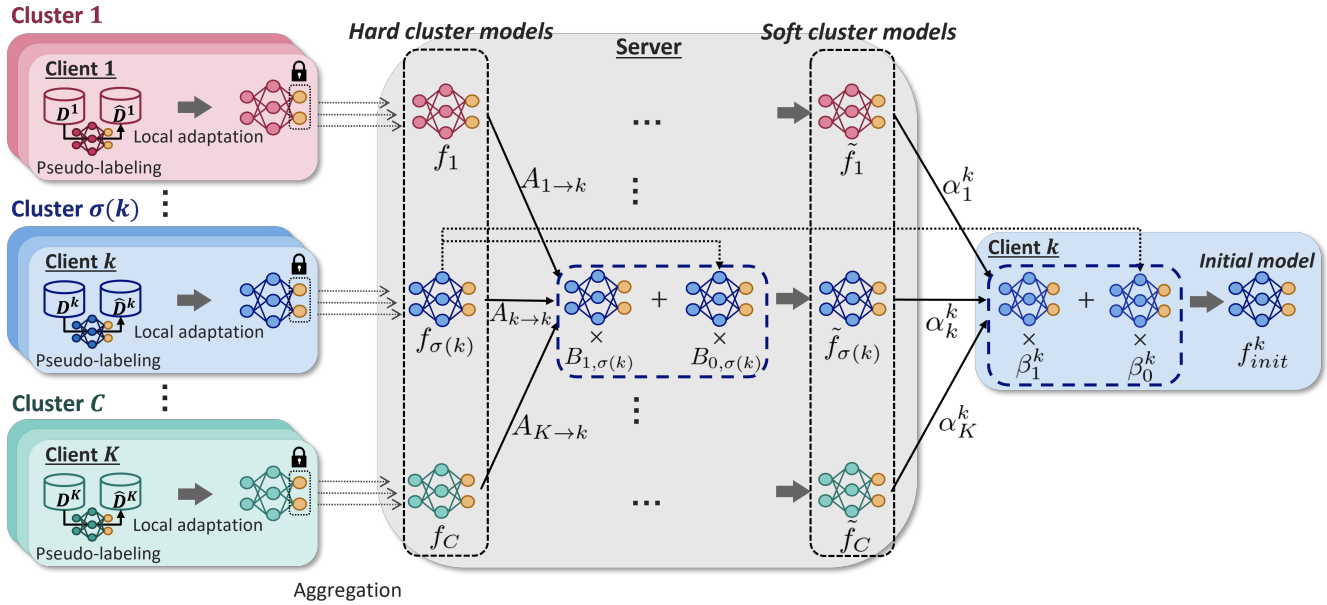


図 2: グローバルかつローカルにクラスタ重みを計算する場合の提案手法におけるモデル統合部分（weighted clustered aggregation）の概略図.

り当てる．しかし，特徴量抽器の全パラメータを使用してコサイン類似度を算出すると，次元の呪いにより全クライアントが一つのクラスターに落ち込んでしまうため，最もドメインの情報を保持していると考えられる第一層目のパラメータのみを使用する．以上より，サーバーはクラスタの集合 $\mathcal{C} = \{1, \ldots, C\}$ と，割り当て関数 $\sigma \colon \mathcal{K} \to \mathcal{C}$ を手に入れる．

4.3 重み付きクラスター化統合方法

最後に我々の手法の最も核となる統合方法，weighted clustered aggregation について説明する．LADD を含む既存のクラスタ連合学習では，各クラスタ内でローカルモデルをサーバーが平均することでクラスタモデル $\{f_c\}_{c\in\mathcal{C}}$ ($f_c = \sum_{k\in\sigma^{-1}(c)} \frac{N^k f^k}{N_c}$) を作成する．ここで，$\sigma^{-1}(c)$ はクラスタ c 内のクライアントの集合で，$N_c = \sum_{k\in\sigma^{-1}(c)} N^k$ である．サーバーが $f_{\sigma(k)}$ をクライアント k に送信し，$f_{\sigma(k)}$ を次のクライアント k の初期モデルとする．しかし，これではクライアント k はクラスタ $\sigma(k)$ 以外の知識を活用することが出来ず，特に少数のデータしかないクラスタや，ソースドメインと親和性の低いクラスタに属するクライアントにとっては大きな問題となる． そこで我々は，各クライアントごとに初期モデル f^k_{init} を， 全クラスタモデル $\{f_c\}_{c\in\mathcal{C}}$ をクライアント独自の重み $\{v^k_c\}_{c\in\mathcal{C}}$ ($\sum_c v^k_c = 1$) によって線形結合させることで作成することを目指す:

$$f^k_{init} = \sum_{c\in\mathcal{C}} v^k_c f_c. \tag{4}$$

この結合によって，誤ったクラスタに割り当てられたクライアントの性能を改善することも可能である．

まずはクラスタの重みを，各クライアントが自身のデータのみを使用してローカルに計算するアプローチの説明から始める．

4.3.1 クラスタ重みのローカルな計算法

各クライアントは，自身のドメインに有益なクラスタモデルにより大きな重みを与えたい.「良い」特徴抽出器は，特徴量空間上のプロトタイプと見なせる適切なクラスに対する分類器ベクトルに各サンプルを収束させられる，すなわち特徴量空間上でサンプルを上手くクラスタリングできるものである．そこで，クライアント k に対するクラスタ重みとして，以下の新しい量 $\{\alpha^k_c\}_{c\in\mathcal{C}}$ を定義する．

$$I^k_c = \frac{1}{N^k} \sum_{i=1}^{N^k} cos(f_c(x^{i,k}), w_{m(i)}), \tag{5}$$

$$m(i) = \arg\max_m cos(f_c(x^{i,k}), w_m),$$

$$(\alpha^k_1, \ldots, \alpha^k_C) = softmax\left(\frac{I^k_1}{T_a}, \ldots, \frac{I^k_C}{T_a}\right).$$

$m(i)$ は，$x^{i,k}$ に最も近い分類器ベクトルのクラスであり，$softmax$ はソフトマックス関数，T_a は温度パラメータである．しかし，このアプローチでは，各クライアント自身の利益のみを考慮し，全体的な利益は考慮していないため，他のクライアントに良い影響を与える可能性が低い．

4.3.2 クラスタ重みのローカルかつグローバルな計算法

クラスタ重み $\{v^k_c\}_{c\in\mathcal{C}}$ を各クライアントの利益と全体の利益の両方の観点から計算することで，前記アプローチを改善することを考える．本手法を示した概略図を図 2 に示す．各クラスタ内のクライアントのローカルモデルのみを統合したモデルをハードクラスタモデル $\{f_c\}_{c\in\mathcal{C}}$ と呼び，さらに他のクラスタモデルの利益もを加味して他のクラスタモデルを取り込んだソフトクラスタモデル $\{\tilde{f}_c\}_{c\in\mathcal{C}}$ を作成する．このために，クラスタ c' のクライアントが他のクラスタのクライアントに比べてクラスタ c の利益を受ける度合いを $A_{c'\to c}$ ($\sum_{c'} A_{c'\to c} = 1$) と定義する．$A_{c'\to c}$ を用いると，新しいクラスタモデル $f'_c = \sum_{c'} A_{c'\to c} f_c$

を作ることができるが，それらのクラスタモデルは元のハードクラスタモデルの知識を著しく失う可能性がある．そこで，元のハードクラスターモデル f_c と，他のクラスターモデルと結合した新しいクラスタモデル f'_c に対する重視度合いを $B_{0,c}, B_{1,c}$ $(B_{0,c} + B_{1,c} = 1)$ と定義する．これらを用いてソフトクラスタモデルを次のように作成する:

$$\tilde{f}_c = B_{0,c} f_c + B_{1,c} \sum_{c'} A_{c' \to c} f_{c'}. \tag{6}$$

$A_{c' \to c}$ と B_c を具体的にどのように計算するかは後述する．

サーバーは，クライアント k に全ソフトクラスタモデル $\{\tilde{f}_c\}_{c \in \mathcal{C}}$ とハードクラスタモデル $f_{\sigma(k)}$ を送信する．それからクライアント k は $\{\tilde{f}_c\}_{c \in \mathcal{C}}$ に対するクラスタ重み $\{\alpha_c^k\}_{c \in \mathcal{C}}$ を式 5 と同様の方法で計算し，モデル $\tilde{f}^k = \sum_c \alpha_c^k \tilde{f}_c$ を得る．さらにハードクラスタモデル $f_{\sigma(k)}$ を十分に取り込むために，クライアント k は $f_{\sigma(k)}$ と $\tilde{f}^k$ に対する疑似的な精度 β_0^k, β_1^k $(\beta_0^k + \beta_1^k = 1)$ によって重み付けして結合する．すなわち，β_0^k と β_1^k はそれぞれ $g_S \circ f_{\sigma(k)}$ と $g_S \circ \tilde{f}^k$ が互いに相対的にどれくらい良く分類タスクをとけるかを表している．ラベル無しデータのみを用いてそれらを算出する具体的な方法として，SND [20] を採用する．SND は，入力サンプルに対する出力スコア間の類似度分布のエントロピーを計算することで，近傍領域の密度を測定する．SND を用いて β_0^k, β_1^k は次のように定義される:

$$(\beta_0^k, \beta_1^k) = softmax\left(\frac{snd(g_S \circ f_{\sigma(k)})}{T_b}, \frac{snd(g_S \circ \tilde{f}^k)}{T_b} \right). \tag{7}$$

T_b は温度パラメータで，snd は SND を測る関数である． 結果として，クライアント k は次のように決定される:

$$\begin{aligned} f_{init}^k =& \beta_0^k f_{\sigma(k)} + \beta_1^k \sum_c \alpha_c^k \tilde{f}_c \qquad (8) \\ =& \beta_0^k f_{\sigma(k)} + \beta_1^k \sum_c \alpha_c^k \left\{ B_{0,c} f_c + B_{1,c} \sum_{c'} A_{c' \to c} f_{c'} \right\} \\ =& \sum_c \left\{ \mathbb{1}(c = \sigma(k)) \beta_0^k + \beta_1^k \alpha_c^k B_{0,c} \right. \\ & \left. + \beta_1^k B_{1,c} \sum_{c'} \alpha_{c'}^k A_{c \to c'} \right\} f_c. \end{aligned}$$

ここで，三行目以降の等号を得るために，第三項にて添え字 c と c' を互いに交換した．また，$A_{c \to c'}$ と $B_{i,c}$ は，各クライアントから収集した α_c^k と β_i^k を各クラスタ内で平均することで算出する:

$$A_{c \to c'} = \frac{\sum_{k \in \sigma^{-1}(c)} N^k \alpha_{c'}^k}{\sum_{k \in \mathcal{K}} N^k \alpha_{c'}^k}, \tag{9}$$

$$B_{i,c} = \frac{1}{N_c} \sum_{k \in \sigma^{-1}(c)} N^k \beta_i^k. \tag{10}$$

従って，クラスタ重み 8 はローカルに計算された値 α_c^k, β_i^k グローバルに更新された値 $A_{c \to c'}, B_{i,c}$ に基づいており，各クライアントに対する個別の利益と全体の利益の両方をもたらす．

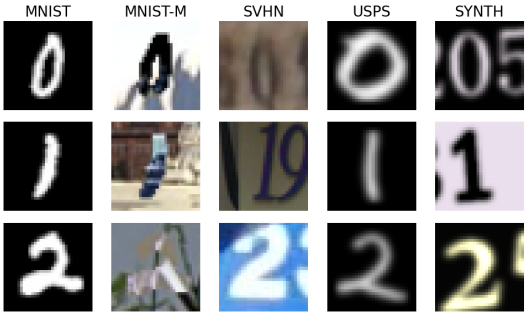


図 3: Digit-Five における各ドメインのサンプルの一例.

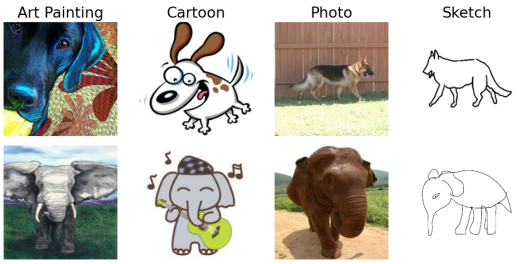


図 4: PACS における各ドメインのサンプルの一例

5 実　　験

5.1 データセット

本論文では，以下の 2 つのデータセットを用いて実験を行う．

5.1.1 Digit-Five

Digit-Five はドメイン適応における標準的なデータセットである．0 から 9 までの 10 クラスからなる，5 ドメインの数字認識データセット（MNIST [21]， MNIST-M [22], SVHN [23], SYNTH [22], USPS）で構成される．データの例を図 3 に示す．いずれも，フォントや背景が異なる数字で構成されており，USPS のみ全サンプルである 9298 枚を，その他は抽出した 34000 枚を利用した．

5.1.2 PACS

PACS（Photo-Art-Cartoon-Sketch）は，Li ら [24] によって作成されたドメイン適応向けデータセットである．犬，象，キリン，ギター，馬，家，ヒトの 7 クラスの画像が，Photo（1670 枚），Art Painting（2048 枚），Cartoon（2344 枚），Sketch（3929 枚）の 4 ドメインにわたって提供されている（図 4）．

5.2 モデル構造

モデルは主に Liang ら [3] の構造を踏襲し，ボトルネック部を含む特徴量抽出器と分類器に大別される．Digit-Five の学習には特徴量抽出器として，2 層の畳み込み層を持つ古典的な LeNet [21] に，batch normalization と dropout を組み込んだ全結合層からなるボトルネック部を組み合わせたものを採用した．分類器は weight normalization を組み込んだ 1 層の全結合層からなる．PACS の学習には，特徴量抽出器のバックボーンとして ImagaNet で事前学習された ResNet-18 [25] を採用し，Digit-Five と同様のボトルネック部および分類器を用いている．

ソースドメイン	MNIST	SVHN	MNIST-M	USPS	SYNTH	Average
Local-SHOT	0.506	0.836	0.695	0.567	0.814	0.684
FedAvg+SHOT	0.584	0.852	0.801	0.533	0.805	0.715
LADD	0.489	0.782	0.738	0.574	0.836	0.684
LADD+SHOT	0.598	**0.920**	0.733	0.565	**0.873**	0.738
FedWCA-L	0.713	0.916	**0.822**	0.566	0.804	0.764
FedWCA-GL	**0.698**	0.918	**0.826**	**0.591**	0.841	**0.775**

表 1: Digit-Five を用いた実験時，各ドメインをソースとした場合の残りの全ターゲットクライアントの accuracy の平均値

ソースドメイン	Art Painting	Cartoon	Photo	Sketch	Average
Local-SHOT	0.736	0.765	0.644	0.751	0.724
FedAvg+SHOT	0.755	0.761	0.611	0.790	0.729
LADD	0.732	0.783	**0.662**	0.608	0.696
LADD+SHOT	0.726	0.758	0.659	0.755	0.724
FedWCA-L	0.769	**0.821**	0.631	0.805	0.757
FedWCA-GL	**0.788**	0.820	0.659	**0.816**	**0.771**

表 2: PACS を用いた実験時，各ドメインをソースとした場合の残りの全ターゲットクライアントの accuracy の平均値

5.3 比較手法

以下の手法を用いて比較実験を行った．

- **Local-SHOT**: 各クラアイアンがローカルに本提案手法と同じ SHOT ベースの学習を行い，モデルの統合は行わない．
- **FedAvg-SHOT**: 各クラアイアンがローカルに本提案手法と同じ SHOT ベースの学習を行い，モデルの統合には FedAvg を適用．
- **LADD**: FFREEDA の先行研究である LADD [2] を適用．
- **LADD-SHOT**: LADD における各クライアントのローカルの学習部分を，本提案手法と同じ SHOT ベースのものに変更．
- **FedWCA-L**: クラスタ重みをローカルに計算した場合の提案手法（4. 3. 1 節を参照）．
- **FedWCA-GL**: クラスタ重みをグローバルかつローカルに計算した場合の提案手法（4. 3. 2 節を参照）．

5.4 実験設定

Digit-Five と PACS の 2 種類のデータセットではいずれも，D 種のドメインの中からソースドメインを 1 つ選択し，ソースモデルを学習した．その後，残りの $D-1$ のドメインをターゲットドメインとし，1 ドメインあたり K_d 個のクライアントにデータを IID（サンプル数および保有するクラス数が均等）になるように振り分け，合計で $K_d(D-1)$ 個のクライアントを学習に参加させた．データセットが持つドメイン数 D は，Digit-Five で $D=5$，PACS で $D=4$ であり，1 ドメインあたりのクライアント数は Digit-Five で $K_d=8$，PACS で $K_d=3$ とした．各クライアントのデータはおおよそ，全サンプルのうち 2 割を評価用，残りの 8 割のうちさらに 8 割を学習用，2 割を検証用として 3 種類に分割した．

5.5 ハイパーパラメータ

ローカルエポックを $E=5$，通信ラウンド数を $R=100$ とした．SHOT に関連するハイパーパラメータは元論文に記載のものを使用し，LADD のスタイル特徴量の計算に際しては，各クライアントがもつターゲットデータを 100 サンプル抽出し，周波数スペクトルの中央 5%の領域を用いた．その他，探索したパラメータとその範囲は，クライアントの学習率 $[0.01, 0.001, 0.0001]$，LADD におけるソフトラベル損失項の係数 $[0.001, 0.01, 0.1, 1, 10]$，提案手法における温度パラメータ $[0.01, 0.05, 0.1, 0.5]$ である．

全ての実験において，3 つの異なるソースモデルを用意して実験を行った．6 章で示す精度は，各クライアントが最終的に獲得したモデルに対して，各クライアントが持つ評価用データで精度を評価し，得られた精度を全クライアント，3 つのソースモデルにわたって平均したものである．

6 実験結果

Digit-Five と PACS での結果をそれぞれ，表 1、表 2 に示す．提案手法および各比較手法を用いて，ソースドメインを 1 つ定めた場合の残りの全てのターゲットドメインでの accuracy の平均を示している．

まず，LADD におけるローカルの学習部分を本提案手法と同じ SHOT ベースのものに変更した LADD-SHOT によって，元の LADD から Digit-Five と PACS でそれぞれ平均で 5%，3%程の改善がみられる．これは LADD は元々semantic segmentation タスクを対象としており，分類タスクを対象とする SFDA 手法を適用することで性能が改善されたことを示している．さらに，各クラスターモデルの結合を取り入れた本提案手法である FedWCA-L によって LADD-SHOT から平均で 3%前後の改善が確認できる．さらに FedWCA において，クラスタ重みを各クライアント個々の利益だけでなく全体の利益を考慮した FedWCA-GL により平均で 1 から 1.5%の改善がみられることから，4. 3. 2 節で述べたクラスタ重み計算法の有効性も確認できる．

7 結　　論

我々は，ソースフリードメイン適応を連合学習設定にて行う，連合ソースフリードメイン適応（FFREEDA）に対する新たな手法を提案した．我々の提案手法は，各クライアントのプライバシーを損なうことなくクライアントをドメインごとにクラスタリングし，クラスター内の似たクライアント同士でモデルの統合を行うことで精度の劣化を防いだ．さらに我々の提案手法では，他のドメインからなるクラスターの知識も活用することで，既存方式を含む複数のベースラインの精度を上回ることを実験的に確認した．

謝　　辞

本研究開発は、総務省の「安全なデータ連携による最適化ＡＩ技術の研究開発（JPMI00316）」によって実施した成果を含みます。

文　　献

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

[2] Donald Shenaj, Eros Fanì, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro Zanuttigh, and Barbara Caputo. "learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.

[3] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 13–18 Jul 2020.

[4] Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *arXiv preprint arXiv:2302.11803*, 2023.

[5] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–17, 2022.

[6] Felix Sattler, Klaus-Robert Muller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, Vol. 32, No. 8, p. 3710—3722, 2021.

[7] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 17871–17884. Curran Associates, Inc., 2022.

[8] Ming Li, Qingli Li, and Yan Wang. Class balanced adaptive pseudo labeling for federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16292–16301, June 2023.

[9] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1424–1433, January 2022.

[10] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *International Conference on Learning Representations*, 2020.

[11] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, Vol. 19. MIT Press, 2006.

[12] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, Vol. 79, pp. 151–175, 2010.

[13] Shiqi Yang, yaxing wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 29393–29405. Curran Associates, Inc., 2021.

[14] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8978–8987, October 2021.

[15] Shiqi Yang, yaxing wang, kai wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 5802–5815. Curran Associates, Inc., 2022.

[16] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *International Conference on Learning Representations*, 2020.

[17] Yan Kang, Yuanqin He, Jiahuan Luo, Tao Fan, Yang Liu, and Qiang Yang. Privacy-preserving federated adversarial domain adaptation over feature groups for interpretability. *IEEE Transactions on Big Data*, pp. 1–12, 2022.

[18] Chun-Han Yao, Boqing Gong, Hang Qi, Yin Cui, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1081–1090, 2022.

[19] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9164–9173, 2021.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. *Domain-Adversarial Training of Neural Networks*, pp. 189–209. Springer International Publishing, Cham, 2017.

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

DEIM2024 xxx

個人化されたユーザレベルの差分プライバシ保護型連合学習

加藤　郁之† 高木　駿† Li　Xiong†† 曹　洋††† 吉川　正俊††††

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† Emory University, Department of Computer Science　201 Dowman Dr, Atlanta, GA 30322, US
††† 北海道大学大学院情報科学研究院　〒 060-0808 北海道札幌市北区北 8 条西 5 丁目
†††† 大阪成蹊大学データサイエンス学部　〒 533-0007 大阪府大阪市東淀川区相川 3 丁目 10-62
E-mail: †fumiyuki@db.soc.i.kyoto-u.ac.jp, ††takagi.shun.45a@st.kyoto-u.ac.jp, †††lxiong@emory.edu,
††††yang@ist.hokudai.ac.jp, †††††yoshikawa-mas@osaka-seikei.ac.jp

あらまし　ユーザレベルの差分プライバシ保護型の連合学習は，学習済みモデルに対して訓練データを提供するユーザに対する厳密なプライバシを保証することができる．一方で，この手法は全てのユーザが同じプライバシ強度を要求することを前提にしている．そこで本論文では，同一ユーザのデータがサイロを跨いで存在しうる一般的なクロスサイロ FL のシナリオにおいて，個人が異なるプライバシ強度を設定することを可能にする個人化されたユーザレベルの差分プライバシ保護を可能にする方法を研究する．ULDP-AVG を拡張した PULDP-AVG を提案し，ユーザ全体が単一のプライバシ強度しか持てない従来の方法と比べて高い有用性を達成することが可能であることを示す．また，本シナリオにおいて問題となる可能性のあるハイパーパラメータの調整について，オンラインで自動最適化を行うための手法を提案し，評価を行う．

キーワード　連合学習，差分プライバシ

1 はじめに

GDPR に代表されるように，現在個人データの利活用に対するプライバシの懸念や規制が強まっている．連合学習 (Federated Learning, FL) [20] は，複数のパーティが生の訓練データを共有することなく 1 つのグローバルモデルを学習する協調的な機械学習スキームであり，産業界からも注目を集めている [26]．ただし，FL は，それ自体が訓練ずみモデルに対してプライバシ保護を提供する訳ではないという点に注意する必要がある [24]．そこで，統計データに対する数学的かつ厳密なプライバシ保証である差分プライバシ (Differential Privacy, DP) を組み合わせた差分プライバシ保護型 FL(Differentially Private FL, DP-FL) が盛んに研究されている [9, 12]．

DP は，近年プライバシ保護のデファクトとして多くの注目を集めているが，以下の理論的限界を持つ．標準的な DP の定義は，1 つのレコードをプライバシ保護の単位として考え，任意の 1 レコードの有無に対する識別不能性を提供する．この前提は，入力のデータベースの中に 1 人のユーザが複数のレコードをもつ場合に成立しなくなり，提供されるプライバシ保証は悪化してしまうことが知られている．このため，*user-level DP* という DP の拡張が研究されている [15, 17]．user-level DP では，1 つのレコードの代わりに，1 人のユーザに属する全てのレコードがプライバシ保護の単位として設定される．これは標準的な DP よりも厳しい定義である．この user-level DP は，任意の k レコードをプライバシ単位とする *group-privacy* [8] と厳密に異なることに注意する．この User-level DP は，FL においても盛んに研究されている [9, 12, 21]．しかしながら，これらの研究は，一人のユーザのデータが 1 つのクライアントデバイスのみに属する前提の，クロスデバイス FL を対象にしている．クロスサイロ FL [13, 16, 18, 25] は、比較的少数のサイロ（病院やクレジットカード会社など）が訓練ラウンドに参加する FL の実用的な亜種である。特に，ULDP-FL [13] は，クロスサイロ FL において，複数のサイロに跨るデータを持つユーザに対して user-level DP を保証する学習アルゴリズムである．

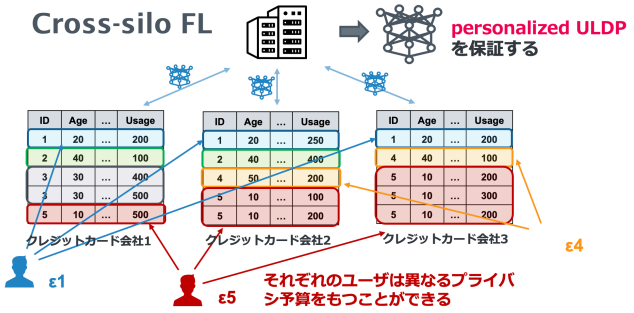


図 1: 複数のクレジットカード会社を同一ユーザが使う時のように，クロスサイロ FL では同一ユーザに属するレコードがサイロをまたいで存在することがある．ULDP-FL [13] では、このような状況において、ユーザレベルの正式なプライバシ保証 (user-level DP) を満たすモデルを学習するアルゴリズムが提案した．本論文では，このシナリオを各ユーザが異なるプライバシ予算を持つ設定に一般化することを目指す．

Personalized DP(PDP) [11] は，全てのレコードに対して同一のプライバシ予算を仮定する標準的な DP と異なり，各レコードに対して個別のプライバシ予算を設定することを可能にする．これは直感的に，最も厳しいプライバシを要求する

ユーザに合わせるという標準的な DP の原理から逸脱して，プライバシ要求の低いユーザからのデータをより活用することができ，出力の有用性を高めることを可能にする．PDP は，標準的な DP 保護型機械学習のための最適化アルゴリズムである DP-SGD [1] に対して，個人化されたサンプリング率 (以降 q_u と呼ぶ)，もしくは，個人化された敏感度 (以降 C_u と呼ぶ) を適用することによっても達成できることが知られている [6, 10, 29]．これは，DP-SGD において q_u と C_u を用いることで，共通のノイズを付加していても，レコードごとに異なるプライバシ境界を計算できることから理解できる．ただし，我々の知る限り，ULDP-FL [13] のようなサイロを跨いで同一ユーザのデータが存在する一般的なクロスサイロ FL の設定において，personalized user-level DP(PULDP) を満たすモデルを学習するアルゴリズムは知られていない．

また，個人化されたサンプリング率と敏感度を組み合わせて使う方法は，一方を天下り的に決定することで，所望のプライバシ境界を達成するもう一方のパラメータをプライバシ境界の定理から導く手法が知られている [6]．一方で，これらの 2 つのハイパーパラメータ (HP) を適切にバランスさせて選択する方法は未知である．特に FL においては，分散的かつ協調的な特性上，HP の調整コストが極端に高い可能性があるため，HP 最適化 (HPO) をどのように実行するかは重要である．

そこで本研究では，一般的なクロスサイロ FL において，personalized user-level DP を保証するモデルを学習することを目的とし，PULDP-AVG を提案する．PULDP-AVG は，ULDP-AVG [13] に対して個人化されたサンプリング率と個人化された敏感度を統合することで，効果的に personalized user-level DP を満たす．さらに，我々は個人化されたサンプリング率と敏感度について理論的，経験的に観察を行い，与えられたプライバシの制約下において，これらの間に最適なバランスが存在する可能性を指摘する．この 2 つの HP の最適なバランスを手作業で調整して発見することは，通常，FL のような分散かつ協調的な環境ではコストが高い [14]．また，我々の設定ではそれに加えて差分プライバシの保証も考慮する必要があることに注意する．この課題に対して我々は，モデルの学習中に近似勾配を使ってオンラインでこれらの HP を最適化する新しい HPO 手法を提案する．近似勾配の計算に使われる情報に対するユーザレベルのプライバシ保証を慎重に追跡することで，HPO を含んだ全体的な学習プロセスに対する差分プライバシ保証を与える．差分プライバシ保護型機械学習において，HPO は与えられたプライバシ予算の範囲外で行われることが多いが，我々の提案手法は一部の HPO プロセスも含む差分プライバシ保証も提供することができる．この 2 つの HP のバランスは FL に限らず，PDP や DP-SGD 一般に関する独立した問題であることに注意する．しかしながら，特に PDP を考慮した FL においては，異なるプライバシを持つユーザグループごとに最適な HP が異なる可能性があり，HPO のコストが大きくなるため，重要度が大きくなると考えている．

本論文における我々の貢献は以下のように要約される．

- サイロを跨いで同一ユーザのデータが存在しうる一般的なクロスサイロ FL の設定において，personalized user-level DP を満たすモデルを学習する新しいアルゴリズム，PULDP-AVG を提案する．我々の知る限り，本目的を達成する初めてアルゴリズムである．
- PULDP-AVG における 2 つの重要なハイパーパラメータの個人化されたユーザレベルのサンプリング率と敏感度の組に最適な組が存在する可能性を実験的に示す．これらを適切にバランスさせるため，新しいオンライン HPO 手法を提案する．この手法は，DP-FL の分散的でかつプライバシ制約のある環境下であっても動作する．
- 実データを用いた実験によって我々の提案アルゴリズム PULDP-AVG を使った DP の個人化によって，個人化を行わない場合よりもモデルの有用性が向上することを示す．また，提案したオンライン HPO 手法も実験的に評価する．

2 準　　備

2.1 クロスサイロ FL

本研究では以下のようなクロスサイロ FL のシナリオを考える．中央の集約サーバとサイロ集合 S が全ての学習ラウンドに参加する．各ラウンドで，集約サーバは全てのサイロがローカル環境で学習したモデル (モデルパラメータの差分，もしくは勾配) を集約し，集約して更新したモデルを再び分配する．各サイロ $s \in S$ は，一般に非凸である損失関数 $F(x;\xi)$ をもち，その期待値である f_s を最適化する．$x \in \mathbb{R}^d$ はモデルパラメータを表し，ξ はデータサンプルで，期待値とはサイロ s のローカルのデータ分布 $\mathcal{D}_s$ 上の期待値を表す．クロスサイロ FL では，サイロの集合全体で，単一のグローバルモデルのパラメータを協調的に最適化することを目指し，目的関数は以下のように定式化できる．

$$\min_x \left\{ f(x) := \frac{1}{|S|} \sum_{s \in S} f_s(x) \right\}, \; f_s(x) := \mathbb{E}_{\xi \sim \mathcal{D}_s} F(x;\xi). \quad (1)$$

本論文では，さらに以下のような表記を用いる．全てのレコードはあるユーザ $u \in U$ に属し，U はユーザ集合を表す．1 人のユーザはサイロを跨いで複数のレコードを持つことができる．それぞれのサイロ s はそれぞれのユーザ u に対して目的関数 $f_{s,u} := \mathbb{E}_{\xi \sim \mathcal{D}_{s,u}} F(x;\xi)$ をもち，$\mathcal{D}_{s,u}$ は，サイロ s 内のユーザ u に属するデータの分布である．最後に，FL のラウンド $t \in [T]$ におけるグローバルモデルのパラメータは x_t で表される．

このモデル化は通常のクロスデバイス FL とは，1 人のユーザが 1 つのデバイスに存在しないという点で異なる．1 人のユーザのレコードは複数のサイロに存在しうる．これは例えば，1 人の患者が複数の病院を利用する，もしくは，1 人の顧客が複数のクレジットカードを利用するなどのケースに該当する．さらに，全てのサイロが全ての学習ラウンドに参加する点も，クライアントが確率的にラウンドに参加するクロスデバイス FL とは異なる．また，クライアントの数は比較的少なく，典型的には 2 から 100 程度である．

2.2 差分プライバシ

DP [8] は厳密なプライバシ保証を提供する数学的なプライバシの定義である．入力データからのアウトプットをランダム化

することで確率的にデータのプライバシを保証する．

定義 1 ((ϵ,δ)-DP)．最大で 1 つのレコードが異なる任意の 2 つの入力データベース $D, D' \in \mathcal{D}$ に対して，ランダム化メカニズム $\mathcal{M}: \mathcal{D} \to \mathcal{Z}$ が以下の条件を満たす時 (ϵ,δ)-DP を満たす．

$$\Pr[\mathcal{M}(D) \in Z] \leqq \exp(\epsilon) \Pr[\mathcal{M}(D') \in Z] + \delta. \qquad (2)$$

このようなデータベース D と D' を隣接データベースと呼び，任意の隣接データベースに対する 2 つの出力の最大の差を敏感度と呼ぶ．DP では，この敏感度全体をカバーしうるランダム化を施すことで，任意の隣接データベースに対する確率的識別不能性を提供する．この本来の DP の定義は，隣接データベースが最大 1 つのレコードが異なるデータベースに対して定義されることから *record-level* DP と呼ばれる．

DP-SGD [1] は機械学習モデルに対して DP を保証するための一般的なフレームワークであり，勾配の計算を sub-sampled Gaussian mechanism (SGM) [27] によって行うことで，厳しいプライバシ境界を与えることができる．

User-level DP across silos. DP-FL においては既存の多くの手法がクロスデバイス FL を対象としていたが，[13] はクロスサイロ FL における user-level DP を以下のように与える．

定義 2 ((ϵ,δ)-User-Level DP (ULDP))．最大で 1 つのユーザに属するレコード全てが異なるサイロを跨いだ任意の 2 つの入力データベース $D, D' \in \mathcal{D}$ に対して，ランダム化メカニズム $\mathcal{M}: \mathcal{D} \to \mathcal{Z}$ が式.(2) の条件を満たす時 (ϵ,δ)-ULDP を満たす．

入力データベースがサイロ全体を跨る全体のデータを包含することに注意する．ULDP と record-level の DP との基礎的な違いは，隣接データベースの定義にあり，user-level の隣接データベースを考えている．この目的のために，ULDP アルゴリズムの設計においては user-level の敏感度を考える必要がある．

アルゴリズム 5(付録) は，[13] で提案された ULDP-AVG を表す．この手法は，DP-FedAVG [21] にユーザごとの重みつきクリッピング (8 行目-15 行目) を組み合わせることで，訓練済みモデルに対してサイロを跨いだ ULDP(定義 2) を保証している．ULDP-AVG の拡張として，各サイロに配布する重みをサイロの持つ対象ユーザのレコード数に応じて決定する ULDP-AVG-W [13] が提案されており，以降，重み $\mathbf{W}$ はこれを前提とする．

Personalized DP. Personalized DP (PDP) [3, 6, 11, 16] は，個人に対して異なるプライバシ保証を与えることを可能にする DP の拡張である．ここで，個人とは 1 つのレコード [6, 11] やクライアント，サイロ [16] などが当てはまる．以下の PDP の定義はレコードに基づく PDP を表す．

定義 3 (Personalized DP (PDP) [11])．任意の 1 レコード $d \in \mathcal{D}$ に対して，d のみが異なる任意の 2 つの入力データベース $D, D' \in \mathcal{D}$ に対して，ランダム化メカニズム $\mathcal{M}: \mathcal{D} \to \mathcal{Z}$ が以下の条件を満たす時 (ϵ_d, δ_d)-PDP を満たす．

$$\Pr[\mathcal{M}(D) \in Z] \leqq \exp(\epsilon_d) \Pr[\mathcal{M}(D') \in Z] + \delta_d. \qquad (3)$$

我々の知る限り，PDP を達成するアルゴリズムはこれまで信頼できる中央サーバを前提とした設定 (すなわち，一般的な Central DP の前提) でしか検討されておらず，[13] で研究されたような一般的なクロスサイロ FL のシナリオにおける user-level の PDP では不明である．

DP を達成するアルゴリズムを PDP を達成するように拡張する一般的なアプローチとして，主に以下の 2 つの方法が知られている: (1) 個人ごとに異なる敏感度を用いる [3]．(2) 個人ごとに異なるサンプリング率を用いてサブサンプリングを行う [11]．最近では，これらのアプローチは差分プライバシ保護型機械学習にも応用され，DP-SGD に対してこの 2 つの方法を片方，もしくは両方適用することによって PDP の保証を与える手法が検討されている [6, 10, 29]．一方で，この 2 つの手法をどのようにバランスさせるかは不明である．所望のプライバシ予算を達成する個人化された敏感度と個人化されたサンプリング率の組は無限に存在するため，どの組を選択するのかが課題となる．本論文では，この 2 つのパラメータについてどのように最適なペアを見つけるのかについて解決策を提案する．

2.3 自動ハイパーパラメータ最適化

一般に機械学習の HPO は重要であるが非常に労力がかかる作業として知られており，これを自動化する方法が望まれる．SGD の最適化ステップにおける学習率などに対する HPO として，モデルパラメータに対する勾配を利用したものが提案されている．代表的な手法としてオフライン最適化 [19] やオンライン最適化 [5] がある．ハイパーパラメータ α が与えられたとき，これらの手法は，モデルパラメータに対する α の勾配 $\frac{\partial f(x)}{\partial \alpha}$ を計算して，$\alpha_t = \alpha_{t-1} - \beta \frac{\partial f(x_{t-1})}{\partial \alpha}$ によって α の更新を行う．β は HPO のための学習率であることに注意する．ただしこの値は，厳密に計算することが難しいことが多く，例えば，訓練損失を利用してかつ，1 ラウンド前の値で近似する [5] などの工夫が必要である．オンライン最適化 [5] では，モデルのパラメータの更新と HP の更新が同じタイミングで行われるのに対して，オフライン最適化 [19] では，モデルのパラメータの全ての学習ラウンドを終えたのちに初めて学習率の更新が行われる．検証損失に対する勾配を計算することができるためより正確な更新が可能となるが，モデルパラメータの学習を内部ループに保つため計算コストが非常に大きい．

近年 FL における HPO も研究が始まっている [14]．[14] は，クライアントでの学習結果を用いて最適な HP の集合の推定確率分布に対する継続的な更新を行うことで，オンラインで HP を最適化する方法を提案している．ただし，この手法はローカルの学習結果をサーバに共有する必要があるが，その際に差分プライバシを考慮していないことに注意する．

3 問題設定

3.1 仮　　定

本研究のシナリオは一般的なクロスサイロ FL を対象とし [13] と同様である．集約サーバと全てのサイロは半正直であり，定

められた FL プロトコルを逸脱することないが，観察できる結果から情報を推測しようとする．学習済みのモデルは第三者に対して公開される可能性がある．これは FL における典型的な仮定である [7]．また，それぞれのパーティは共謀しないと仮定する．サーバとサイロの間の通信は全て SSL/TLS で暗号化され，いかなる第三者も最終的な学習モデル以外にアクセスすることはできない．基礎となる ULDP-AVG アルゴリズムは [13] で提案されたプライベートなプロトコルに従い実行されると仮定する．さらに、サイロ間の record-linkage が既に完了し、共通のユーザ ID が全てのサイロの間で共有されていると仮定する。また，サーバとサイロの両方が、ユーザの総数 $|U|$ とサイロの数 $|S|$ を知っている．プライバシ増幅のためにサブサンプリングが採用されている場合、サーバのみが各ラウンドのサブサンプリング結果を知ることが許される [2,12]．これらの仮定はすべて、外部ユーザに対する最終的な学習済みモデルのプライバシー保証には影響しないことに注意する．

また，各ユーザ $u \in U$ は自身のプライバシ予算 ϵ_u を持つとし，近いプライバシ予算をもつ k 個のグループ P_i $(i \in [k])$ のどれかに割り当てられると仮定する．各 P_i は対応するプライバシ予算 ϵ_i をもち，全てのユーザ $u_i \in P_i$ に対して，$\epsilon_{u_i} \leqq \epsilon_i$ を満たす．典型的には，グループ数 k は 3 で，ユーザは個人情報の利活用対して，conservative $(\epsilon_1 = 0.5)$, normal $(\epsilon_2 = 2.0)$, liberal $(\epsilon_3 = 5.0)$ のどれかであり，ϵ_i を持つユーザは，P_i に割り当てられる．ϵ_u が離散値でない場合は，適当な ϵ_i, P_i を先に設定し，全ての ϵ_u を最も近いがそれを超えない ϵ_i を持つ P_i に割り当てることができる．

3.2 目　　的

本研究では上記の仮定の下で，サイロ全体のデータを入力とし学習済みモデルを出力とする学習アルゴリズムにが以下の Personalized user-level DP(PULDP) の制約を満たした上で，目的関数 (式.(1)) を最適化することを目指す．

定義 4 (Personalized ULDP (PULDP))．任意の 1 人のユーザに属する全てのレコードの集合 $d_u \subseteq \mathcal{D}$ に対して，d_u のみが異なる任意の 2 つの入力データベース $D, D' \in \mathcal{D}$ に対して，ランダム化メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{Z}$ が以下の条件を満たす時 (ϵ_u, δ)-PULDP を満たす．

$$\Pr[\mathcal{M}(D) \in Z] \leqq \exp(\epsilon_u) \Pr[\mathcal{M}(D') \in Z] + \delta. \qquad (4)$$

ϵ_u はユーザごとのプライバシ予算であり公開パラメータである．また，先行研究 [6,13] と同様 δ は十分に小さい値 $(\delta < \mathcal{O}(1/|U|))$ によって全ユーザで共通であるとする．

4 PULDP-AVG

4.1 アルゴリズムの概要

ULDP-AVG に対して，個人化された敏感度とサンプリング率を導入することで PULDP を達成できる．アルゴリズム 1 は，ULDP-AVG を PULDP を満たすように拡張した PULDP-AVG を表す．なお，拡張した箇所を赤で強調している．アルゴリズムの入力として，全てのユーザ u に対して個人化されたプライバシ予算 ϵ_u を受け取り，また，q_u または C_u のどちらか一方を受け取る．SetParams(3 行目，詳細は付録のアルゴリズム 6) は，個人化されたサンプリング率 q_u を受け取り，ϵ_u を満たす個人化された敏感度 C_u を返す．この関数は，C_u に対するプライバシ境界を都度計算しながら二分探索を行うことで ϵ_u を満たす最大の C_u を見つける．これは，q_u と C_u に対して可換であることに注意する．ここで得られた q_u は，ユーザ u に対するポアソンサンプリングに使用され (5 行目)，C_u はクライアント側で計算されるユーザ u のデータで更新されたパラメータの差分 $(\tilde{\Delta}_t^{s,u})$ に対する個人化されたクリッピング定数として用いられる (18 行目)．サブサンプリングは，サイロに共有される重みに 0 を設定することでそのユーザの貢献をなくすことで実装可能であり (6 行目)，この結果は必要に応じて暗号化することで秘匿することができる [13]．個人化された敏感度によるクリッピングはクライアント側で行われる必要があるので，C_u は全てサイロに生データとして共有される必要があるが，これは公開パラメータでありプライバシ漏洩を起こさないことに注意する．このような個人化されたサブサンプリングと敏感度は，[6] において FL の設定でも user-level の設定でもないものの，同様の手法が提案されている．次の定理は，アルゴリズム 1 が PULDP を満たすことを述べる．

アルゴリズム 1 PULDP-AVG

Input: η_l, η_g, σ, T, Q, $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_{|S|})$: matrix with weight for user u and silo s, and $\forall u \in U$, $w_{s,u} \in \mathbf{w}_s$ and $\sum_{s \in S} w_{s,u} = 1$ (following the enhanced weighting strategy based on the numbers of user records proposed in [13]), $\{\epsilon_u\}$: individual privacy budget for $u \in U$, $\{q_u\}$: individual sampling rate

1: **procedure** Server
2: Initialize model x_0
3: $\{C_u\}, \{q_u\} \leftarrow$ SetParams$(\{q_u\}, \{\epsilon_u\})$
4: **for** each round $t = 0, 1, \ldots, T-1$ **do**
5: $U_t \leftarrow$ PoissonSampling$(U, \{q_u\})$ ▷ User-level poisson sub-sampling from U with probability q_u for each u
6: $\mathbf{W} \leftarrow$ Update$(\mathbf{W}, U_t)$ ▷ Set $w_{s,u} = 0$ if $u \in U_t$
7: **for** each silo $s \in S$ **do**
8: $\Delta_t^s \leftarrow$ Client$(x_t, \mathbf{w}_s, \{C_u\}, \sigma, \eta_l)$
9: $x_{t+1} \leftarrow x_t + \eta_g \frac{1}{\mathbb{E}[q_u]|U||S|} \sum_{s \in S} \Delta_t^s$
10: **procedure** Client$(x_t, \mathbf{w}_s, \{C_u\}, \sigma, \eta_l)$
11: **for** user $u \in U$ **do** ▷ per-user training with $\mathcal{D}_{s,u}$
12: $x_t^{s,u} \leftarrow x_t$
13: **for** epoch $q = 0, 1, \ldots, Q-1$ **do**
14: Compute stochastic gradients $g_{t,q}^{s,u}$
15: ▷ $\mathbb{E}[g_{t,q}^{s,u}] = \nabla f_{s,u}(x_t^{s,u})$
16: $x_t^{s,u} \leftarrow x_t^{s,u} - \eta_l g_{t,q}^{s,u}$
17: $\Delta_t^{s,u} \leftarrow x_t^{s,u} - x_t$
18: $\tilde{\Delta}_t^{s,u} \leftarrow w_{s,u} \cdot \Delta_t^{s,u} \cdot \min\left(1, \frac{C_u}{\|\Delta_t^{s,u}\|_2}\right)$
19: $\Delta_t^s \leftarrow \sum_{u \in U} \tilde{\Delta}_t^{s,u} + \mathcal{N}(0, I\sigma^2 C^2 / |S|)$
20: **return** Δ_t^s

定理 1. PULDP-AVG は全てのユーザ $u \in U$ に対して (ϵ_u, δ)-

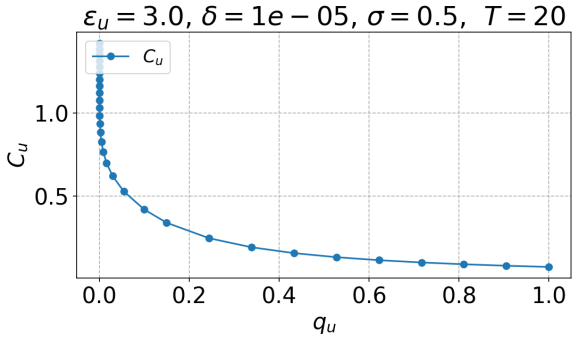
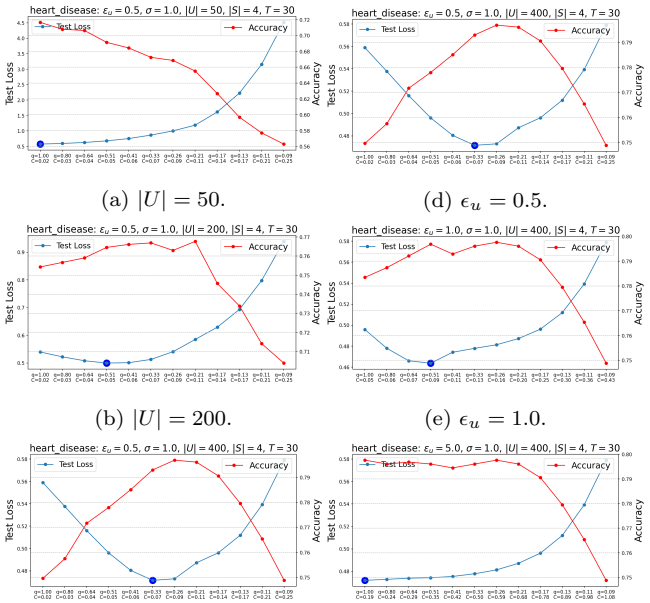


図 2: 与えられた制約 $(\epsilon_u = 3.0, \delta = 10^{-5}, \sigma = 0.5, T = 20)$ を満たす (q_u, C_u) の組の例．(他のパラメータには非依存．)



(a) $|U| = 50$.　(d) $\epsilon_u = 0.5$.

(b) $|U| = 200$.　(e) $\epsilon_u = 1.0$.

(c) $|U| = 400$.　(f) $\epsilon_u = 5.0$.

図 3: ユーザ数 $|U|$ とプライバシ予算 ϵ_u に対する最適な (q_u, C_u) の組の変化．

PUDLP を満たす．(証明は，[6] の付録 G.1, G.2，および [13] の Theorem 3 を参照．)

PULDP-AVG は，ULDP-AVG と同様の計算コストを持つ一方で，C_u を共有するため，各ラウンドにおいて $O(|U|)$ の余計な通信コストが発生することに注意する．

4.2 PULDP-AVG のための HP

PUDLP-AVG では，SetParams を通じて，任意の $0 < q_u < 1$ に対してプライバシを保証する C_u を得ることができる．一方で，与えらた制約，プライバシ予算 ϵ_u，FL のラウンド数 T，ノイズの分散 σ を満たす HP の組の実行可能解 (q_u, C_u) は無限に存在する．図 2 は，PULDP-AVG を前提として，制約 $T = 20, \epsilon_u = 3.0, \delta = 10^{-5}, \sigma = 0.5$ を満たす，様々な $0 < q_u < 1$(x 軸) に対して対応する C_u(y 軸) を示している．これらの点は全て，同様のプライバシ保証を提供することが可能である一方で，モデルの学習に対して異なる結果を与える可能性があることに注意する．

図 3 は HeartDisease [25] データセットを用いて [1]，様々な (q_u, C_u) を用いて学習されたモデルの評価を行った結果を示している．(類似の結果を MNIST データセットに対して示した結果を付録の図 A·1,A·2 に載せる．) 縦軸は，検証損失 (左，青) と検証精度 (右，赤) を表し，横軸は，左端から $q_u = 1.0$ で開始し，より小さな q_u を表す．(プロットごとに 0.8 の割合で小さくなっている．) 5 回のランダム試行の平均を点でプロットし，標準偏差を縦棒でプロットしている．学習アルゴリズムは PULDP-AVG で，全てのユーザが同一のプライバシバジェットを持つとする．(つまり，グループ数 $k = 1$．) その他の HP は，$(\epsilon_u, \sigma, |U|, |S|, T)$ が $(0.5, 1.0, 400, 4, 30)$ のように固定されている．これらの結果は，(q_u, C_u) は大まかに単峰性をもち最適な (q_u, C_u) が存在するということを示している．特に，図 3 はそれぞれ異なるユーザ数とプライバシ予算に対する最適な (q_u, C_u) の組の変化を示している．ユーザ数とプライバシ予算が異なる場合，最適な組が変わっていることが見て取れる．したがって，(q_u, C_u) は最適な組をもち，これはユーザ数やプライバシ予算によって異なる可能性がある．全体的に，(左端の)$q_u = 1.0$ 付近では C_u が小さく，クリッピングによる情報の損失によるバイアスが増加することで有用性も低くなり，(右端の)q_u が極端に小さくなるとデータが不足することと，モデルに付加されるガウシアンノイズの摂動による悪影響が大きくなり有用性が落ち，その間に最適な有用性を持つ (q_u, C_u) が存在する，と解釈することができる．

これらの事実は (q_u, C_u) に対する HPO の必要性を示しているが，特に我々のシナリオにおいては，以下の理由から HPO が大きな問題となる．

（1） FL は分散的な設定であり，訓練データにアクセスするコストが大きく，手作業で試行錯誤を行い HPO を行うことは困難である．

（2） さらに，DP-FL では，差分プライバシを保証する必要があるため，[14] のように学習データに依存したローカルの情報を，サーバが生で収集することはできない．

（3） PDP の設定では，異なるプライバシ予算をもつユーザのグループが複数存在し，それぞれに対して異なる最適な (q_u, C_u) が存在する可能性があるため，HPO のコストがさらに大きくなる．

したがって，(q_u, C_u) に対して分散的かつ差分プライバシ保証を満たした効率の良い HPO 手法が重要となる．

5 PULDP を満たした自動 HPO

5.1 ヒューリスティクス

最初に，重要なヒューリスティクスとして，同じプライバシ予算を持つグループ P_i ごとに独立した HP をもつと仮定する．これは，PULDP-AVG ではサイロ・ユーザごとに独立してモ

1 : HeartDisease [25] は，クロスサイロ FL の公開ベンチマークデータセットであり，各レコードは事前にサイロに割り当てられている．[13] の *uniform* の方法 (すなわち独立同分布の仮定) に従い全てのレコードをサイロ・ユーザに割り当てる．

デルの学習を行うため，(q_u, C_u) の HP は互いに影響し合わないと考えられるからである．(付録の図 A·3 を参照．) [28] では，DP-SGD においてレコードごとに異なるクリッピング定数 (すなわち C_u) を設定することで，有用性が向上することが観察されている．したがって HP の個別化については，全体で 1 種類の HP を使う設定からレコードごと (我々の設定ではユーザごと) の粒度の設定のうちのどれかで設定することが考えられる．我々は，(q_u, C_u) がサンプリング率やプライバシ予算に応じて変化するという観察から，P_i ごとの粒度で異なる最適な値を持つことを仮定する．つまり，(q_i, C_i) $(i \in [k])$ を最適化することを考える．これは HPO にかかるコストの観点からは，最も粒度の低いユーザごとの粒度よりも優れている．

5.2 HPO 手法

最初に自明かつ最も精度の高いベースラインとしてグリッドサーチに基づくオフラインの最適化手法について，次に，HP に対する検証損失の近似勾配を用いたオンラインの最適化手法について，グローバルな検証データセットを仮定した手法 (QCTest)，グローバルな検証データセットを仮定しない手法 (QCTrain)，について説明する．

5.2.1 オフライン HPO

HPO におけるグリッドサーチは，最初に入力として候補となる HP のセットを用意し，全ての候補に対して 0 からモデルの評価を行い (交差検証などによる)，最終的に最も検証結果の良い HP を選択する手法である．この方法は，内部ループにモデルの評価を行う必要があることから効率が悪い．特に，実際のクロスサイロ FL での利用を考えると，訓練データをもつサイロが HPO 全体に参加する必要があるため，内部ループのさらに内部の学習ループで毎回通信を必要とするクロスサイロの分散環境では協調のためのコストが増大する．また，検証結果は通常，検証損失や検証精度を用いるが，これは訓練データに依存した値であり，モデルのプライバシ保護とは別に，この値も検証データに対するプライバシを漏洩する可能性があることに注意する．この情報公開はグリッドサーチの HP の組み合わせごとに必要であり，非常に大きなコストとなることに注意する．図 3 は，$k = 1$ の場合のグリッドサーチの実行結果に対応する．

この方法は，効率とプライバシを無視した最も有用性の高いベースライン手法として考えられる．このようなオフライン最適化は，上で説明したように計算・通信コストの面からも，プライバシ保護の面からも，我々の目的とは相性が悪く実用的な方法にはなりえない．そのため，以下では 2.3 章で紹介した勾配ベースのオンライン HPO を考える．

5.2.2 オンライン HPO

我々の勾配ベースのオンライン HPO において，学習率の最適化に焦点を当てた [5] とは異なり，q_u に対する検証損失の勾配，すなわち $\frac{\partial f}{\partial q_u}$ を陽に計算することは困難である．代わりに有限差分法を用いると $\frac{\partial f}{\partial q_u}$ が以下のように近似される:

$$\frac{\partial f}{\partial q_u} \approx \frac{f_{q_u+\delta} - f_{q_u}}{\delta}. \tag{5}$$

ただし f_{q_u} は，u のみ q_u を使った時の損失 (式.(1) 参照) を表し，δ は 0 より大きい実数である．$\frac{\partial f}{\partial q_u}$ と学習率 α を用いて，各学習ラウンド t における $q_u^{(t)}$ を以下のように更新できる:

$$q_u^{(t+1)} := q_u^{(t)} - \alpha \frac{\partial f}{\partial q_u}. \tag{6}$$

しかし一般に有限差分法では，適切な差分 δ を選択することが難しく，$\frac{\partial f}{\partial q_u}$ の値は制御が困難である．そこで，より更新を安定させるために q_u に対するステップサイズ α (例えば，$\alpha = 0.5$) を用いて，制御された指数関数的な更新を行う以下のような更

アルゴリズム 2 QCTest

Input: η_l, η_g, σ, T, Q, $\mathbf{W}$, $\{\epsilon_i\}$: individual privacy budget for user group $i \in [k]$, α $(0 < \alpha < 1)$: step size, β: momentum weight, D_{test}: public test dataset

procedure QCTestServer
  Initialize model x_0 and HP $\{q_i^{(0)}\} \leftarrow \{q_0\}^k$
    ▷ initial sampling rate q_0 is e.g., 1.0, 0.1, 0.01
  for each round $t = 0, 1, \ldots, T-1$ **do**
    $\{q_i^{(2)}\} \leftarrow \alpha * \{q_i^{(0)}\}$
    $\{C_i^{(1)}\}, \{q_i^{(1)}\} \leftarrow$ SetParamsWithHistory$(\{q_i^{(0)}\}, \{\epsilon_i\})$
    $\{C_i^{(2)}\}, \{q_i^{(2)}\} \leftarrow$ SetParamsWithHistory$(\{q_i^{(2)}\}, \{\epsilon_i\})$
    for $j = \{0, 1, 2\}$ **do**
      $U_t^{(j)} \leftarrow$ PoissonSampling$(U, \{q_i^{(j)}\})$
        ▷ if $u \in P_i$ then $q_u = q_i$, otherwise $q_u = 0$
        ▷ $U_t^{(0)}$ and $U_t^{(1)}$ are i.i.d. results
      $\mathbf{W}^{(j)} \leftarrow$ Update$(\mathbf{W}, U_t^{(j)})$
    for each silo $s \in S$ **do**
      $\Delta_t^s, \{\Delta_{s,t,i}^{(1)}\}, \{\Delta_{s,t,i}^{(2)}\} \leftarrow$ QCTestC$(x_t, \mathbf{w}_s^{(0)}, \mathbf{w}_s^{(1)}, \mathbf{w}_s^{(2)}, \{C_i^{(1)}\}, \{C_i^{(1)}\}, \{C_i^{(2)}\}, \sigma, \eta_l)$
    $x_{t+1} \leftarrow x_t + \eta_g \frac{1}{\mathbb{E}[q^{(0)}]|U||S|} \sum_{s \in S} \Delta_t^s$
    for $(i, \Delta_{s,t,i}^{(1)}, \Delta_{s,t,i}^{(2)}) \in \{[k], \{\Delta_{s,t,i}^{(1)}\}, \{\Delta_{s,t,i}^{(2)}\}\}$ **do**
      $x_{t+1,i}^{(1)} \leftarrow x_t + \eta_g \frac{1}{q_i^{(1)}|U||S|} \sum_{s \in S} \Delta_{s,t,i}^{(1)}$
      $f^{(1)} \leftarrow$ Eval$(x_{t+1,i}^{(1)}, D_{\text{test}})$
      $x_{t+1,i}^{(2)} \leftarrow x_t + \eta_g \frac{1}{q_i^{(2)}|U||S|} \sum_{s \in S} \Delta_{s,t,i}^{(2)}$
      $f^{(2)} \leftarrow$ Eval$(x_{t+1,i}^{(2)}, D_{\text{test}})$
      $\frac{\partial f}{\partial q} \leftarrow f^{(2)} - f^{(1)}$
      $q_i^{(0)} \leftarrow$ QUpdate$(i, t, \frac{\partial f}{\partial q}, \alpha, \beta, q_i^{(0)})$
procedure QCTestC$(x_t, \mathbf{w}_s^{(0)}, \mathbf{w}_s^{(1)}, \mathbf{w}_s^{(2)}, \{C_i^{(0)}\}, \{C_i^{(1)}\}, \{C_i^{(2)}\}, \sigma, \eta_l)$
  $\Delta_{s,t}^{(0)} \leftarrow$ Client$(x_t, \mathbf{w}_s^{(0)}, \{C_u^{(0)}\}_{u \in [U]}, \sigma, \eta_l)$ ▷ Client is the same as in Algorithm 1, $C_u^{(0)} = C_i^{(0)}$ for $u \in P_i$
  for $i \in [k]$ **do**
    $\Delta_{s,t,i}^{(1)} \leftarrow$ Client$(x_t, \mathbf{w}_s^{(1)}, \{C_i^{(1)}\}, \sigma, \eta_l)$
    $\Delta_{s,t,i}^{(2)} \leftarrow$ Client$(x_t, \mathbf{w}_s^{(2)}, \{C_i^{(2)}\}, \sigma, \eta_l)$
  return $\Delta_{s,t}^{(0)}, \{\Delta_{s,t,i}^{(1)}\}, \{\Delta_{s,t,i}^{(2)}\}$

アルゴリズム 3 QUpdate

Input: i, t, $\frac{\partial f}{\partial q}$: (approximated) gradient for f w.r.t q, α: step size, β: momentum weight, q: current sampling rate

$v_t^{(i)} \leftarrow \beta * v_{t-1}^{(i)} + (1-\beta) * \frac{\partial f}{\partial q}$
save $v_t^{(i)}$
$q' \leftarrow \min(q * \alpha^{-\text{sign}(v_t^{(i)})}, 1.0)$
return q'

新規則を導入する:

$$\frac{\partial f}{\partial q_u} \approx f_{q_u+\delta} - f_{q_u},$$
$$q_u^{(t+1)} := \min(q_u^{(t)} * \alpha^{-\mathrm{sign}(\frac{\partial f}{\partial q_u})}, 1.0). \quad (7)$$

この更新規則は，近似勾配の符号のみを利用して固定の割合で，q_u を増減させる．α の選択によって HP の取れる範囲が限定されてしまうという欠点を持つが，図 3 から最適性は緩やかに現れる可能性が高いため，これはあまり問題にならないと考える．さらに，指数的な更新を行うことで，少ない学習ラウンドでも最適に近いペアに向かうことが期待される．一般に $\frac{\partial f}{\partial q_u}$ は正確に計算することは困難であり，かつ PULDP-AVG の学習過程にはいくつかのランダム性が含まれる．そのため，更新方向を安定させるために，勾配のモーメンタムを使った以下のような更新規則を導入する:

$$v_t^{(i)} := \beta * v_{t-1}^{(i)} + (1-\beta) * \frac{\partial f}{\partial q_u}, \quad v_0^{(i)} = 0,$$
$$q_u^{(t+1)} := \min(q_u^{(t)} * \alpha^{-\mathrm{sign}(v_t^{(i)})}, 1.0). \quad (8)$$

β はモーメンタム係数である．この更新規則はアルゴリズム 3 に示される．最後に，この規則は常に $q_u^{(t)}$ を $q_u^{(t)} * \alpha$ か $q_u^{(t)} * \alpha^{-1}$ にして収束しないため，必要に応じて α をラウンド数に応じて減少させるステップ減衰を組み合わせることができる．我々の実験でも適用した典型的なステップ減衰は以下:

$$\alpha' = 1 - (1-\alpha)(1 - t/T). \quad (9)$$

これによって，適切な値に近づいた HP が振動することを防ぐことができる．

オンライン HPO は，モデルのパラメータの学習と同時に HP も自動で最適化していくため，オフライン HPO とは異なり，(クライアントも参加する必要があるという意味ではオンラインの) 時間のかかるモデル評価の内部ループを持たないために通常の FL プロセスと同程度に効率的であり分散環境に向いている．例えば，6 章の HeartDisease データセットの実験の設定ではオフラインでの HPO には少なくとも $7^3 = 343$ 回の試行が必要であるが，QCTest/Train はただ 1 回の試行である．一方で，プライバシの観点からは，$\frac{\partial f}{\partial q_u}$ が訓練データに依存した出力である．そのため，この情報を加味した差分プライバシ保護を考える必要がある．アルゴリズム 2, 4 は，このオンライン最適化を用いた 2 手法を示している．

QCTest: アルゴリズム 2 に示すように，各ラウンドにおいて，PULDP-AVG を基礎として異なる 2 つの HP を用いて学習した 2 つのモデルを収集してそれぞれ評価を行うことで，近似勾配を計算する．各ラウンドで，全てのグループ i に対してサンプリング率をステップサイズ α の割合だけ変化させた値 $q_i^{(2)}$ を計算する (4 行目)．$q_i^{(0)}$ と $q_i^{(1)}$ は同じである．次にそれぞれに対応する C_i を SetParamsWithHistory(詳細は付録) によって計算する (5,6 行目)．計算された $q_i^{(0)}$，$q_i^{(1)}$, $q_i^{(2)}$ を使って，それぞれ独立にユーザ集合 U に対してポアソンサプリングを実行する (7,8 行目)．サンプリング確率は，ユーザ u が P_i に含まれる場合のみ，$q_u = q_i$ で，含まないユーザは $q_u = 0$ とする．$U_t^{(0)}$ と $U_t^{(1)}$ は，同じ分布からサンプリングされたユーザ集合であるが，独立した異なる結果であることに注意する．これはつまり，ユーザグループ i ごとに，元の HP を使った学習設定を 2 つと，更新された HP を使った学習設定を 1 つ，合計 3 つの学習設定を用意することに対応する．これらの 3 つの設定を使って各サイロでモデルを学習する (10, 11，20-25 行目)．サイロ内では，元の HP を使った設定で，PULDP-AVG と同様にモデルの差分を計算する (21 行目) のに加えて，各グループ i ごとに，元の HP を使った学習設定と更新された HP を使った学習設定を使って，それぞれモデルの差分 ($\Delta_{s,t,i}^{(1)}$，$\Delta_{s,t,i}^{(2)}$) を計算する (22-24 行目)．それぞれの学習では，P_i に含まれないユーザの訓練データはサンプリング確率が 0 である (8 行目) ため訓練に含まれないことに注意する．したがって，サーバで集約されるモデルは，3 種類あり，1 つは通常の PULDP-AVG と同様のモデル (x_{t+1})，残りの 2 つは，グループごとでかつ 2 つの学習設定を使って学習されたモデル ($x_{t+1,i}^{(1)}$ と $x_{t+1,i}^{(2)}$) である．最後に，グループ i ごとに，2 つのモデルに対してモデルの評価 (15, 17 行目) と近似勾配の計算 (18 行目) を実行し，それを用いて QUpdate を使って次のラウンドで使用する HP を計算する (19 行目)．これによって，モデルの最適化と同時に，q_i と C_i を正しい方向に最適化していくことができる．

定理 2. QCTest は全てのユーザグループ $i \in [k]$ に対して (ϵ_i, δ)-PUDLP を満たす．(証明のスケッチは付録 3.1．)

備考 1. QCTest では，各ラウンドにおいて 3 回分の差分プライベートなモデルを公開する必要がある．ただし，これが 3 倍のプライバシ境界の増加に繋がることを意味しないことに注意する．モデルの公開は SGM に基づいており，RDP によって厳密にプライバシ境界を合成をすることができる．この合成結果は一般に劣線形であり [27]，SGM の回数が 3 倍になることのプライバシ境界への影響は非常に小さい．このことは次の QCTrain でも共通である．

QCTrain: プライベートではない検証データセットがない場合は，サーバ側で検証損失を計算できないため，サイロのローカルの訓練データセットを用いた訓練損失で代替することが考えられる．ただし，訓練損失は，一般に敏感度を持たないため，ULDP を保証することが困難である．そこで，我々は，訓練損失の代わりに代理関数 h を用いる (15 行目)．この代理関数の条件としては，訓練損失を近似できてかつユーザレベルの敏感度が存在する値であり，例えば，ユーザレベルの精度などを h として使用する．精度はクラス分類タスクの場合，ユーザごとに計算可能であり，[0,1] の値をとるためユーザレベルの敏感度は 1 である．$\mathbf{h}^{(\mathbf{j})}$ は，$|U|$-次元のベクトルであり，のちに $\mathbf{w}_s^{j\top}$ との内積を計算することで (17 行目)，ユーザに対するサンプリング結果を対応するモデル ($x_t + \Delta_{s,t,i}^{(j)}$) の学習 (14 行目) で使ったものと共通にさせる．これは，モデルの学習から h の出力結果に至るまでを単一の SGM として実行するためである．よって，全体で ϵ_i を満たすために必要な noise multiplierσ を計算 (16 行目) し，それを用いて h に分散ノイズを加えることで PULDP を保証してこの結果を収集することができる．ここで集約された $h_i^{(1)}$ と $h_i^{(2)}$ を使って，近似勾配を計算する (8 行

アルゴリズム 4 QCTrain

Input: η_l, η_g, σ, T, Q, $\mathbf{W}$, $\{\epsilon_i\}$, α, β

procedure QCTrainServer
  /* Same as in QCTest (Lines 2-9 in Algorithm 2) */
  for each silo $s \in S$ **do**
    $\Delta_t^s, \{h_{s,t,i}^{(1)}\}, \{h_{s,t,i}^{(2)}\} \leftarrow$ QCTrainC(...)
  $x_{t+1} \leftarrow x_t + \eta_g \frac{1}{\mathbb{E}[q^{(0)}]|U||S|} \sum_{s\in S} \Delta_t^s$
  for $(i, h_{s,t,i}^{(1)}, h_{s,t,i}^{(2)}) \in \{[k], \{h_{s,t,i}^{(1)}\}, \{h_{s,t,i}^{(2)}\}\}$ **do**
    $h_i^{(1)}, h_i^{(2)} \leftarrow \frac{1}{q_i^{(1)}} \sum_{s\in S} h_{s,t,i}^{(1)}, \frac{1}{q_i^{(2)}} \sum_{s\in S} h_{s,t,i}^{(2)}$
    $\frac{\partial f}{\partial q} \leftarrow h_i^{(1)} - h_i^{(2)}$
    $q_i^{(0)} \leftarrow$ QUpdate$(i, t, \frac{\partial f}{\partial q}, \alpha, \beta, q_i^{(0)})$
procedure QCTrainC$(x_t, \mathbf{w}_s^{(0)}, \mathbf{w}_s^{(1)}, \mathbf{w}_s^{(2)}, \{C_i^{(0)}\}, \{C_i^{(1)}\}, \{C_i^{(2)}\}, \sigma, \eta_l)$
  $\Delta_{s,t}^{(0)} \leftarrow$ Client$(x_t, \mathbf{w}_s^{(0)}, \{C_u^{(0)}\}_{u\in[U]}, \sigma, \eta_l)$
  for $i \in [k]$ **do**
    for $j \in \{1, 2\}$ **do**
      $\Delta_{s,t,i}^{(j)} \leftarrow$ Client$(x_t, \mathbf{w}_s^{(j)}, \{C_u^{(j)}\}, \sigma, \eta_l)$
      $\mathbf{h}^{(j)} \leftarrow h(x_t + \Delta_{s,t,i}^{(j)}, D_{\text{train}}^{(i)})$ ▷ h is, for example, user-level Accuracy, which is bounded by [0, 1] per user
      $\sigma \leftarrow$ GetNoiseMultiplier(ϵ_i, q_i) ▷ Algorithm 8
      $h_i^{(j)} \leftarrow \mathbf{w}_s^{j\top} \mathbf{h}^{(j)} + \mathcal{N}(0, \sigma^2/|S|)$
  return $\Delta_{s,t}^{(0)}, \{h_i^{(1)}\}, \{h_i^{(2)}\}$

目)．この手法は，代理関数 h によって計算される近似勾配が，検証損失から得られる近似勾配 (QCTest) と似ているというヒューリスティクスに基づいている．

定理 3. QCTrain は全てのユーザグループ $i \in [k]$ に対して (ϵ_i, δ)-PUDLP を満たす．(証明は略．)

備考 2. QTest と QCTrain のどちらも [13] で提案されたプライベートプロトコルを適用可能であり，サーバに対しては，サイロの個別の値を公開することなく DP を満たした結果のみを公開することができる．

備考 3. QCTrain に比べた QCTest の利点として，代理関数を使う必要がない点が挙げられる．QCTrain は，ULDP を満たして近似勾配を計算するためのメトリクスを収集するために，ユーザレベルの敏感度が制限されたメトリクスを用いる必要があるので，適用できるタスクが制限される可能性がある．一方で，QCTrain であれば，モデル自体を ULDP を保証して収集するため任意のメトリクスで近似勾配が計算可能である．逆に，QCTest に比べた QCTrain の利点として，モデル自体に ULDP を保証するのではなく，メトリクスの ULDP を直接保証するため，ノイズの付加は 1 次元のスカラに対して行われることになる．このことにより，モデルの全次元に対してノイズが付加される QCTest よりもノイズの影響が小さくなる可能性がある．

6 実　　験

主に以下の目的のために実データを用いた実験を行う．

- PULDP-AVG が ULDP-AVG に比べてモデルの性能を改善可能であることを示す．
- 2 つのオンライン最適化，QCTest および QCTrain に対して，どの程度最適化が機能するのかを評価し洞察を得る．

データセット. HeartDisease [25] を用いる．HeartDisease はクロスサイロ FL のための小規模なベンチマークデータセットであり事前に 4 つのサイロにデータが割り当てられている．1000 以下のレコード数と 14 の特徴を持ち，2 クラス分類タスクを解く．[13] の uniform の方法に従い全てのレコードを事前にサイロ・ユーザに割り当てている．

以下の実験では特に断りがない限り，異なるプライバシ予算をもつユーザグループ数 k は 3 で，それぞれのユーザは個人情報の利活用対して conservative ($\epsilon_1 = 0.15$), normal ($\epsilon_2 = 3.0$), liberal ($\epsilon_3 = 5.0$) のどれかであるとする．それぞれの割合が (0.6, 0.25, 0.15) になるようにランダムにユーザを割り当てる．例えば，[6] では，$(\epsilon_1, \epsilon_2, \epsilon_3) = (1.0, 2.0, 3.0)$ の割合でデータが割り当てられる．ULDP-AVG では，全てのユーザのプライバシ予算は最悪ケースである ϵ_1 に合わせる必要があることに注意する．本実験で使用した全てのデータセットとソースコードは公開されている [2]．

手法. ベースラインとして，ULDP-AVG (Baseline)，Random，random-updown，PULDP-AVG (Offline HPO) を紹介する．ULDP-AVG (Baseline) は，[13] で提案された手法を使用する．全てのユーザが，全体のうち最も厳しいプライバシ予算に従うという点を除き，PULDP-AVG と同様である．Random と Random-log は，q_u を 1.0 から $1/N$ の間でランダムに選択する．ただし，Random-log は，ログスケールの q_u が均等の確率に選択されるようにする．random-updown は，ある初期値 q_u^0 (例えば，1.0) からスタートして式 (6) に従いつつ，毎回ランダムに進む方向を決定する．これらのベースラインは，オンラインでの HPO にプライバシ予算を消費しないため，QCTest/Train と比べると，より多くのプライバシ予算をモデルの学習に使用できる点に注意する．PULDP-AVG (Offline HPO) は，オフライン HPO によって，プライバシ漏洩を無視してオフラインで事前に HPO を行ってからその HP を使用して PUDLP-AVG を実行する．同様に，ULDP-AVG にも事前に求めた最適な HP(q_u,C_u の組) を用いている．QCTest/Train では，基本的に $(\alpha, \beta) = (0.5, 0.9)$ とし，q_u の初期値は 0.5 で，更新については式 (9) のステップ減衰を用いる．QCTest では，公開検証データセットが必要であるが，このデータセットは他の手法における検証データセットからも除かれており，QCTest を含む検証データセットは全ての手法で共通である．QCTrain の代理関数は，ユーザレベルの訓練精度の和を用いる．これはユーザレベルの敏感度が 1 である．

PULDP は ULDP よりも高い有用性を持つ. 図 4 は，HeartDisease データセットにおける各手法の有用性の比較結果を表

2 : https://github.com/FumiyukiKato/uldp-fl/blob/DEIM24/exp/puldp_exp_DEIM24.ipynb

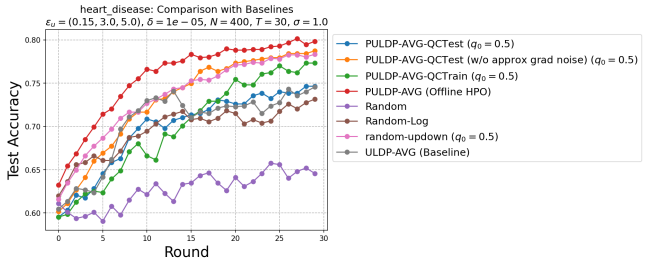
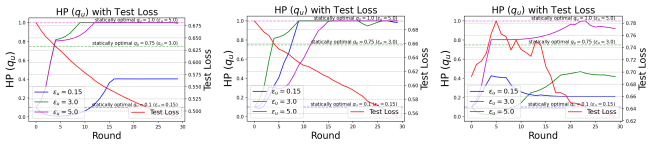


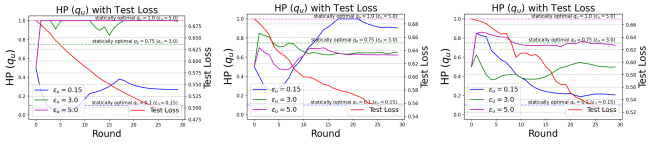
図 4: PULDP-AVG は，ULDP-AVG よりも精度が高い．



(a) QCTest: q_0 =0.1. 近似勾配のノイズなし.

(b) QCTest: q_0 =0.1.

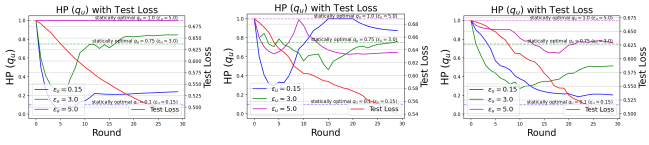
(c) QCTrain: q_0 =0.1.



(d) QCTest: q_0 =0.5. 近似勾配のノイズなし.

(e) QCTest: q_0 =0.5.

(f) QCTrain: q_0 =0.5.



(g) QCTest: q_0 =1.0. 近似勾配のノイズなし.

(h) QCTest: q_0 =1.0.

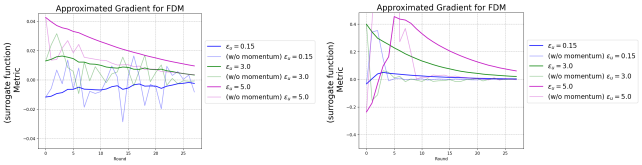
(i) QCTrain: q_0 =1.0.

図 5: オンライン HPO QCTest/Train における HP の推移. 左の 3 つの図は QCTest において，近似勾配の計算のためのモデルのノイズを無しにした結果である．

す．各オンライン最適化手法の q_u 初期値は 0.5 である．(初期値が 0.1 と 1.0 の場合の結果は付録の図 A·4 にある．) x 軸は FL のラウンド，y 軸は検証精度であり 10 回のランダムな試行の平均値を示す．最初に，ULDP-AVG (Baseline) よりも PULDP-AVG (Offline HPO) の方が高い性能を示している．これは，ULDP に比べて PULDP が個人化されたプライバシ保証により，より大きなプライバシ予算をもつユーザのデータを活用して有用性を高めることができることを示している．[6, 10, 29] などが同様の傾向を報告しているが，これは ULDP であっても同様に成立することが分かる．

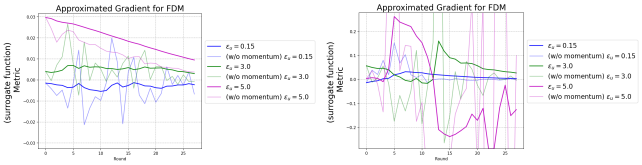
オンライン HPO は近似勾配へのノイズの影響で最適化がうまく働かない． 図 4 において，QCTest/Train のどちらの手法も，ベースラインである ULDP-AVG や Random/-log よりは高い精度を達成しているが，random-updown を下回っている．これは，我々の提案した HPO のために使用したプライバシ予算を，代わりにモデルの学習に使った方が依然として効果的であるということを示している．

図 5 は，QCTest/Train における FL ラウンドごとの q_u の推移を示している．(5 回の試行の平均値をプロットしている．) 図 5a, 5b，5c は，q_u の初期値を 0.1 にし，図 5d, 5e，5f は，q_u の初期値を 0.5 にし，図 5g, 5h，5i は，q_u の初期値を 1.0 にしている．また，図 5a, 5d, 5g は，QCTest において，近似勾配を計算するために 2 つの異なる HP を使ったモデルを集約する際に，モデルへのノイズを 0 にした場合の結果を示している．なお，オフライン HPO によって求めた最適な HP の組は $\epsilon_u = 0.15$, $\epsilon = 3.0$, $\epsilon = 5.0$ のグループに対してそれぞれ，おおよそ 0.1, 0.75, 1.0 である．(付録の図 A·3 を参照．)



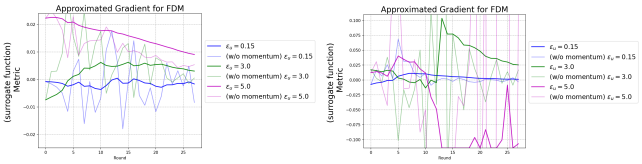
(a) 近似勾配のノイズ無し.

(b) QCTest: q_0 =0.1.



(c) 近似勾配のノイズ無し.

(d) QCTest: q_0 =0.5.



(e) 近似勾配のノイズ無し.

(f) QCTest: q_0 =1.0.

図 6: QCTest における近似勾配の変化．左は，近似勾配計算のためのモデルのノイズを無しにした結果．

近似勾配にノイズが載っていない QCTest が静的に求めた最適な HP に近い値に収束しているのに比べて，通常の QCTest は全く近い値に収束せずうまく最適化が進んでない．この結果は，ラウンドごとの近似勾配の変化をプロットした図 6 からも確認できる．太線は，式 (8) で示した近似勾配のモーメンタムを使って調整した値を表しており，淡い線はモーメンタムを使わなかった場合の値を示している．ノイズ無しの場合は最適値に向けて一貫した推移を見せているにに対して，ノイズがのった通常の QCTest では近似勾配が不安定であり，ランダムに近い挙動となっている．QCTrain も同様に，最適な値に収束していないことが見て取れる．

これに対する改善策としては，近似勾配自体のノイズを減らしてより正しい近似勾配を得る方向性と，ノイズの載った近似勾配からより正しい情報を得るための較正処理を行うことが考えられる．前者は，プライバシ予算の配分を大きくすることなどが考えられるが，プライバシ予算を減らすことで直接的にモデルの精度を落とすことにつながるため難しい．後者はモデルの精度に影響を与えることなく実装可能であるため興味深い．今回使用している単純な指数移動平均 (モーメンタム) だけでなく，例えば，外れ値の影響をなくすための工夫などが有効な

可能性がある．現在の QCTest/Train の近似勾配では，ノイズの影響で大きな外れ値が発生した場合にそれに移動平均の値が引っ張られてしまうからである．

7 まとめ

本論文では，同一ユーザのデータがサイロを跨いで存在しうる一般的なクロスサイロ FL のシナリオにおいて個人が異なるプライバシ強度を設定することを可能にする個人化されたユーザレベルの差分プライバシ保護 (personalized user-level DP) を可能にする手法である PUDLP-AVG を提案した．PULDP-AVG が ULDP-AVG を上回る有用性を達成できることを実験によって示した．さらに，PDP の設定において大きな問題となる HPO に注目し，近似勾配に基づいてオンラインで HPO を行うための手法を提案した．提案法について評価した結果，まだ近似勾配に誤差が大きく，改善の余地が大きいことが分かった．今後の課題として，近似勾配を使ったより効果的な更新規則について調査することが重要であると考えている．

文　献

[1] M. Abadi, A. Chu, et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016.

[2] N. Agarwal, P. Kairouz, and Z. Liu. The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] M. Alaggan, S. Gambs, and A. Kermarrec. Heterogeneous differential privacy. *CoRR*, abs/1504.06998, 2015.

[4] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

[5] A. G. Baydin, R. Cornish, D. M. Rubio, M. Schmidt, and F. Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations*, 2018.

[6] F. Boenisch, C. Mühl, A. Dziedzic, R. Rinberg, and N. Papernot. Have it your way: Individualized privacy assignment for dp-sgd. *arXiv preprint arXiv:2303.17046*, 2023.

[7] K. Bonawitz, V. Ivanov, et al. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[8] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[9] R. C. Geyer, T. Klein, and M. Nabi. Differentially private federated learning: A client level perspective. *NIPS 2017 Workshop: Machine Learning on the Phone and other Consumer Devices*, 2017.

[10] G. Heo, J. Seo, et al. Personalized dp-sgd using sampling mechanisms. *arXiv preprint arXiv:2305.15165*, 2023.

[11] Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*, pages 1023–1034. IEEE, 2015.

[12] P. Kairouz, Z. Liu, and T. Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, pages 5201–5212. PMLR, 2021.

[13] F. Kato, L. Xiong, S. Takagi, Y. Cao, and M. Yoshikawa. Uldp-fl: Federated learning with across silo user-level differential privacy. *arXiv preprint arXiv:2308.12210*, 2023.

[14] M. Khodak, R. Tu, T. Li, L. Li, M.-F. F. Balcan, V. Smith, and A. Talwalkar. Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing. *Advances in Neural Information Processing Systems*, 34:19184–19197, 2021.

[15] D. Levy, Z. Sun, K. Amin, S. Kale, A. Kulesza, M. Mohri, and A. T. Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

[16] K. Liu, S. Hu, S. Z. Wu, and V. Smith. On privacy and personalization in cross-silo federated learning. *Advances in Neural Information Processing Systems*, 35, 2022.

[17] Y. Liu, A. T. Suresh, F. X. X. Yu, S. Kumar, and M. Riley. Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems*, 33:20965–20976, 2020.

[18] A. Lowy, A. Ghafelebashi, and M. Razaviyayn. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pages 5749–5786. PMLR, 2023.

[19] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[20] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

[21] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[22] I. Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[23] I. Mironov, K. Talwar, and L. Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[24] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[25] J. Ogier du Terrail, Ayed, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5315–5334. Curran Associates, Inc., 2022.

[26] M. Paulik, M. Seigel, H. Mason, D. Telaar, J. Kluivers, R. van Dalen, C. W. Lau, L. Carlson, F. Granqvist, C. Vandevelde, et al. Federated evaluation and tuning for on-device personalization: System design & applications. *arXiv preprint arXiv:2102.08503*, 2021.

[27] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235, 2019.

[28] T. Xia, S. Shen, S. Yao, X. Fu, K. Xu, X. Xu, and X. Fu. Differentially private learning with per-sample adaptive clipping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10444–10452, 2023.

[29] D. Yu, G. Kamath, J. Kulkarni, T.-Y. Liu, J. Yin, and H. Zhang. Individual privacy accounting for differentially private stochastic gradient descent. 2022.

非独立同分布の学習データに対する分割学習の性能向上に関する研究

池奥　裕太† 　曹　　洋†† 　吉川　正俊†††

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町 36-1
†† 北海道大学大学院情報科学研究院　〒 060–0814 札幌市北区北 14 条西 9 丁目
††† 大阪成蹊大学データサイエンス学部　〒 533–0007 大阪市東淀川区相川 3 丁目 10-62
E-mail: †ikeoku.yuta.37c@st.kyoto-u.ac.jp, ††yang@ist.hokudai.ac.jp, †††yoshikawa-mas@osaka-seikei.ac.jp

あらまし　近年、プライバシ意識の高まりにより、ユーザのデバイスなどから機械学習に必要なデータを収集することが困難になってきている。このような場合に、分散学習 (distributed learning) が用いられる。分散学習の手法の 1 つに分割学習 (split learning) がある。分割学習は、学習に参加するクライアントの計算効率や通信効率が優れている一方、非独立同分布 (Non-IID) の設定では精度が大幅に低下する欠点が存在する。本研究では、分割学習にクラスタリングの概念を導入した MidSL および、MidSL に集約処理の仕組みを追加した MidSFL を提案する。具体的には、似たデータ分布を持つクライアントを学習中にクラスタリングし、クラスタ間で共有する中間レイヤーを導入する。MidSFL では、通常の SplitFed と比較し、Non-IID 設定での精度を改善することができた。

キーワード　分割学習、分散学習、機械学習とプライバシ

1 はじめに

近年、情報技術の向上や、膨大なデータが収集されるようになったことなどにより、データを元に機械学習が行われる機会が増加した。その中でも、深層学習は、自動運転における画像認識や音声端末による音声認識などの様々な分野で活用されている。

しかし、高い精度の機械学習モデル（以下、単にモデルと言う）を得るには、多くの学習データが必要となる。従来は、様々なデータ源から得られた情報を中央集権的に 1 つのサーバに集約し、そのサーバ上で機械学習を行う方法が採用されていた。しかし、必ずしも全てのデータが集約可能であるとは限らない。近年のプライバシ意識の高まりだけでなく、法的な規制も進んでいる。例えば、EU の GDPR (General Data Protection Regulation) や日本の改正個人情報保護法などが、プライバシ情報の収集をますます難しくしている。

このような場面で、従来の中央集権的な機械学習に代わって、分散学習を利用することができる。分散学習とは、複数のデバイスやサーバが協調して、1 つのモデルの学習を行う機械学習手法である。データを所有する主体が、そのデータを外部に送信することなく学習を行うことができるという特徴を持ち、前述のプライバシの問題を大きく軽減することができる。

Gupta らによって提案された分割学習 [1] は、ニューラルネットワークを特定のレイヤーで分割し、2 つのサブネットワークに分けて学習を行う手法である（図 1）。入力層を含む前半のネットワークをクライアントが、出力層を含む後半のネットワークを中央サーバが所有する。そして、クライアントと中央サーバが協調して学習を行う。しかし、分割学習は、連合学習 [2] や中央集権的な機械学習と比較して、非独立同分布（non identically and independently distributed）の学習データでは学習が進みにくいという問題がある [3] [4]。強い非独立同分布の環境では、モデルが収束しないこともしばしばある [3]。

本研究では、非独立同分布の学習データにおける分割学習の精度を向上させる新たな手法を提案する。この手法では、クライアントから中央サーバに送られてくるデータをもとにクライアントをクラスタリングする。そして、同一クラスタ内に所属するクライアント間で、クライアント側のネットワークの集約処理を行うことで、モデルの精度を向上させる。

2 関連研究

2.1 非独立同分布

実世界では、分散学習に参加する全てのクライアントが等しい分布のデータセットを所持していることは稀である。クライアントによって、データセットのサイズや分布、特徴などが異なる場合が多い。このような設定を非独立同分布と呼ぶ。この章では、分散学習における非独立同分布のパターンについて述べる。

非独立同分布では、分布の異なり方によって複数の種類に分類される [4]。ラベル分布の偏り（label distribution skew）は、クライアントによって、所有するローカルデータセットに含まれるデータの正解ラベルの分布が異なる偏りを指す。具体的には、あるサービスのユーザが居住する地域によって、ユーザのサービス利用状況が異なるケースや、性別や年代によってよく利用する単語が異なるような設定がラベル分布の偏りに分類される。特徴分布の偏り（feature distribution skew）は、クライアントによって、所有するローカルデータセットに含まれる特徴量の分布が異なるような偏りを指す。例えば、手書きの文字認識を行う問題では、同じラベルが付与されていても、人によって書き方の癖や筆圧などが異なることが当たる。このような設定が特徴分布の偏りに分類される。同一の特徴に対するラ

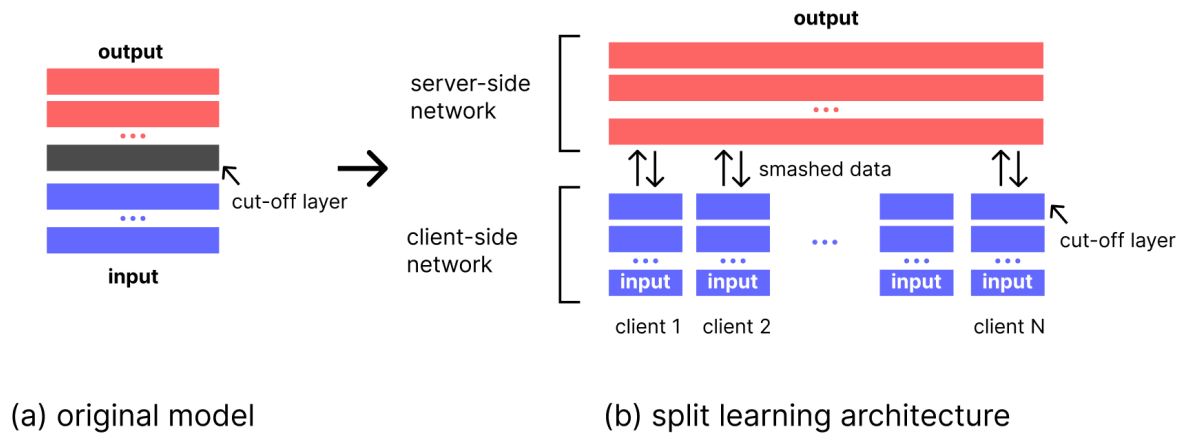


図 1　分割学習のアーキテクチャの概略。学習する機械学習モデルからカットオフレイヤを選択し (a)、クライアントサイドネットワークとサーバサイドネットワークに分割している (b)。クライアントサイドネットワークとサーバサイドネットワークの間でやり取りされるデータを smashed data と呼ぶ。

ベルの偏り（concept shift）は、クライアントが所有するローカルデータセットに含まれる、同一のラベルに対応するデータでも、クライアントによってそのデータの分布が大きく異なる場合を指す。例えば、同じ車を表す画像でも、時代や地域によってその特徴が異なる。

2.2 連合学習

2016 年に McMahan らによって提案された [2] 連合学習は、分散学習の代表的な手法である。全体の学習を管理するサーバと、データを所有し、学習に参加するクライアントから構成され、サーバとクライアントが協調して 1 つのモデルを学習する。

連合学習では、データを所有する各クライアントが、サーバに生のデータを送信する必要はない。代わりに、データセットを利用して得られた新しいモデル、あるいはそのモデルの前グローバルエポックからの差分のみをサーバに送る。そのため、生のデータをサーバに知られることなく学習を行うことができる。また、連合学習では各クライアントにおける学習を並列に行うことができるため、学習時間を削減することができる。

2.2.1 非独立同分布の学習データにおける連合学習

連合学習は、クライアントが持つデータを外部に送信する必要がないため、従来の中央集権的な機械学習と比較して、プラバシ保護に優れた手法である。その一方、学習データが非独立同分布である場合に、モデルの精度が低下してしまう問題がある。Zhao らはこれを、クライアントによってデータ分布が大きく異なるため、各クライアントのローカルモデルを平均化したローカルモデルと、データセットを集約し中央集権的に学習を行った場合のモデルの乖離が大きくなってしまうためであると分析している [5]。この乖離は学習が進むごとに蓄積するため、最終的には大きな精度の低下に繋がる。

非独立同分布の学習データに対しては、個人化レイヤの導入 [6]、ファインチューニング [7]、グローバルデータセットの利用 [5] など様々なアプローチが研究されている。Clustered Federated Learning (CFL) は、階層的クラスタリングを利用して非独立同分布における連合学習の精度を向上させる手法である [8]。CFL では、クライアントから送られてくるローカルモデルの更新差分を利用してクラスタリングを行う。ローカルモデルの更新差分に対してコサイン類似度を計算し、コサイン類似度をもとに全クライアントを 2 つののクラスタに分類する。この操作を、新しいクラスタが生成されなくなるまで繰り返すことで、適切なクラスタリング結果を得る。学習はクラスタごとに独立して行われる。すなわち、各クラスタごとにグローバルモデルが設定され、クラスタに所属するクライアント内でのみ学習が行われる。

Briggs らが提案した Federated learning with hierarchical clustering(FL+HC) [9] でも同様に、クライアントから送られてくるローカルモデルの更新差分を利用して、クライアントのクラスタリングを行う。CFL と異なり、クラスタリングに階層的クラスタリングのアルゴリズムを利用し、一度だけクラスタリングを行うという特徴がある。まず、最初の N エポックで通常の連合学習を行う。N エポック終了時点でクライアントから送られてきたローカルモデルの更新差分に対して、階層的クラスタリングのアルゴリズムを用いてクラスタリングを行う。類似性の高いローカルモデル更新差分を持つクライアントが、同一のクラスタに分類される。また、CFL を改良した Adaptive Clustered Federated Learning では、ローカルデータセットの統計情報を利用した事前クラスタリングを行うことでクラスタリング精度を改善している [10]。

このように、連合学習では、非独立同分布設定の設定でクライアントをクラスタリングすることでより精度の高いモデルを得るアプローチが多く研究されている。一方、分割学習ではクラスタリングを行うアプローチをしている先行研究は存在していない。本研究では、階層的クラスタリングを用いて分割学習

の精度を向上する手法を提案する。

2.3 分割学習

Gupta らによって提案された分割学習は、1 つのニューラルネットワークをクライアントサイドとサーバサイドの 2 つに分割し学習を行う [1]。入力層を含む前半のモデルをクライアントが、出力層を含む後半のモデルをサーバが所有する。なお、モデルの分割が行われたレイヤーはカットオフレイヤーと呼ばれる。なお、分割方法を変更することで、クライアントごとに所有するデータセットの特徴が異なる垂直分割の設定や、マルチタスク学習に対応するアーキテクチャなども提案されている [11]。

学習時にはまず、学習全体を管理するサーバが、学習を行うクライアントを 1 つ選択する。選択されたクライアントが自身が所有するローカルデータセットからランダムにミニバッチを取り出し、自身が持つ前半のニューラルネットワークに入力する（図 2）。そして、カットオフレイヤーに達するまで順伝播を計算したのち、そこから得られた中間出力をサーバに送信する。この中間出力は smashed data と呼ばれる。smashed data を受け取ったサーバは、自身が持つ後半のニューラルネットワークにそのデータを入力し、出力層まで順伝播を計算する。出力層まで達すると、正解ラベルを参照して損失値を計算し、逆伝播の計算を行う。逆伝播においても、計算がカットオフレイヤーに達すると、そこで得られた勾配をクライアントに送る。このデータは smashed data gradient と呼ばれる。smashed data gradient を受け取ったクライアントは、そのデータを前半のニューラルネットワークに入力し、残る部分の逆伝播を計算する。この処理を全てのミニバッチに対して繰り返す。ここまでの処理を 1 ローカルエポックと呼ぶ。

この学習が終了すると、サーバは次に学習を行うクライアントを選択する。そして、直前に学習を行っていたクライアントから、次に学習を行うクライアントに、クライアントモデルのパラメータを送信する。この引き継ぎ処理をモデル同期と呼ぶ。モデルの同期を行うことで、直前のクライアントのクライアントモデルの学習状況を引き継ぐことができ、収束までに必要なエポック数が減少する。一方、引き継ぎを行うために、クライアント間の学習が逐次的になり、学習に必要な時間が増加する。

以上の手順を繰り返し、サーバは学習に参加する全てのクライアントと学習を行う。ここまでの処理を 1 グローバルエポックと呼び、これをモデルが収束するまで繰り返す。処理の流れを Algorithm 1 にまとめる。ここで、D_i はクライアント i が所有するローカルデータセット、m は分割学習に参加するクライアント数を表す。

2.3.1 分割学習の特徴

分割学習では、クライアントは前半のニューラルネットワークのみ、サーバは後半のニューラルネットワークにのみアクセス可能であるため、クライアント、サーバ共にモデルの全体を知ることができない。そのため、モデル全体をクライアントとサーバ間で送受信する連合学習と比較してプライバシが漏洩しづらい特徴がある [1]。また、分割学習では、学習に必要な計算をクライアントとサーバで分担できる利点がある。分割学習においてクライアントが計算を担当する範囲は、クライアントが持つレイヤー部分のみであり、モデル全体の計算を行う必要がある連合学習と比較して、クライアントの計算負荷を大きく減らすことができる。そのため、クライアントとサーバが帯域の限られた無線で通信する場合や、計算資源の限られる IoT デバイスをクライアントとする場合などに分割学習が用いられる [12] [13]。一方、クライアントモデルのレイヤー数を減らすことで、より生のデータに近い smashed data がサーバ側に送信される可能性が高まる。そのため、プライバシと計算コストはトレードオフの関係にある。

Algorithm 1 複数クライアントの場合の分割学習のアルゴリズム

Server initializes server-side network W^s
Each client initializes client-side network W_i^c
for each global epoch $e = 1, 2, ...$ **do**
  $C = \{1, 2, ..., m\}$
  while $0 < |C|$ **do**
    $i \leftarrow RandomSelect(C)$
    Remove i from C
    $W^c \leftarrow W_{prev}$ ▷ Model synchronization
    while client i has new data to train **do**
      Client: get y_i, x_i from D_i
      Client: $s \leftarrow W^c(x_i)$
      Client: sends s to server
      Server: $y' \leftarrow W^s(s)$
      Server: $gradient \leftarrow Loss(y', y_i)$
      Server: $W^{s\prime}, sg^e \leftarrow W^{sT}(gradient)$
      Server: sends sg to client
      Client: $W^{c\prime} \leftarrow W^{cT}(sg)$
    end while
    $W_{prev} \leftarrow W^{c\prime}$
  end while
end for

連合学習では、クライアントとサーバ間でモデル全体を通信でやり取りする。一方、分割学習では、smashed data のみをやり取りするため、1 度あたりの通信でやり取りするデータサイズが小さい。そのため、クライアント数が多い、あるいは、モデルのパラメータ数が多い場合に、連合学習より分割学習の方が通信効率が良くなる [14]。Lin らは、最終レイヤーの勾配を集約することで、計算コストと通信コストを下げる手法を提案している [13]。

2.3.2 非独立同分布の学習データにおける分割学習

非独立同分布の学習データに対して、分割学習の精度が著しく低下することがわかっている [3]。これは、連合学習における精度の低下よりも大きい。これについて Gao らは、連合学習における集約処理のような、複数のクライアントの知識を統合する過程が分割学習には存在しないためだと分析している [3]。この問題に対し Kim らは、無線モデルを利用した分割学習において、クライアント間のモデルの同期を省く手法を提案している [15]。同期を省くことで、クライアントごとに独自のモデル

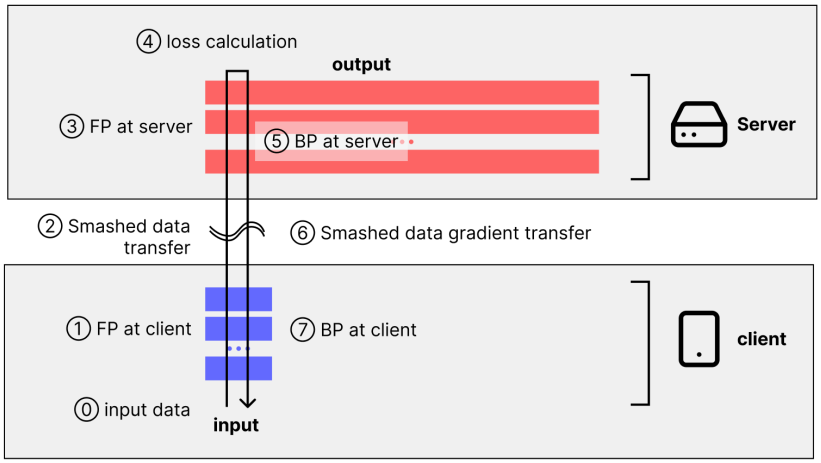


図 2　1 クライアントの場合の分割学習の学習プロセス。FP は順伝播を、BP は逆伝播の計算を表す。

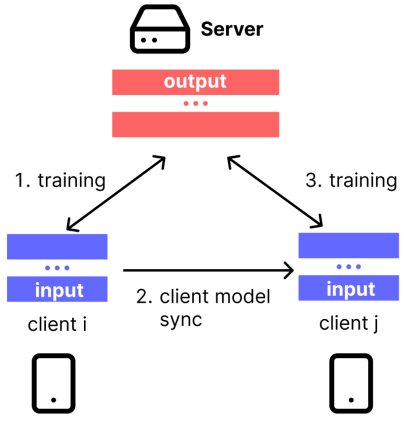


図 3　分割学習におけるクライアントのモデル同期

を保持することとなり、クライアント側のモデルが個人化される。そのため、クライアントごとにデータ分布が異なるような非独立同分布データに対しても有効となる場合が多い。この手法に代表されるような、クライアント間のモデル同期を省き、クライアントが並行して学習を行う手法は、並行分割学習と呼ばれている。

Cai らは、クライアント間のモデルの同期の代わりに、サーバで蒸留を行うことで、クライアント間の知識の統合を行う並行分割学習の手法を提案している [16]。Chopra らは、各クライアントがサーバと独立して、自身で学習を完結させることができる AdaSplit を提案している [17]。また、無線を利用した分割学習に参加するデバイスのバッテリー量や通信帯域などの異種性による精度低下を防ぐ手法として、クライアントをクラスタリングする手法が提案されている [12]。この手法では、デバイスのハードウェア情報を利用してクラスタリングを行い、クラスタ内で学習を並行して行うことで、学習時間を削減することができる。この手法は本論文で提案する手法とクラスタリングという点で類似しているものの、Wu らの研究は非独立同分布ではなく、デバイスや通信環境の異種性に着目しているものであり、根本的に異なっている。

2.3.3　分割学習におけるプライバシ保護

分割学習では、クライアントとサーバともにモデル全体のパラメータにアクセスすることができないため、連合学習と比較してプライバシ保護に優れているとされている [1]。一方、smashed data によって、クライアントからサーバにプライバシが漏洩する可能性もある。Vepakomma らはこれに対して、smashed data と元のデータの KL ダイバージェンスを損失関数に組み込むことで、モデルの精度を維持しながら smashed data と元のデータの距離相関を最小にするアプローチを提案している [18]。また、Thapa らは提案手法の中で、差分プライバシを応用した PixelDP [19] を組み込むことでプライバシ保護を実現している [20]。Yang らは、正解ラベルを持つサーバのプライバシを保護するため、モデルパラメータと通信データの両方に差分プライバシを満たすノイズを付加する手法を提案している [21]。

2.4　連合学習と分割学習の組み合せ

Thapa らは、分割学習に連合学習の集約手法を取り入れた SplitFed を提案している [20]。SplitFed は、クライアント間のモデルの引き継ぎによる同期を行わない。代わりに、新たに導入した集約サーバを利用して、クライアントモデルの集約処理を行う (図 4)。具体的には、1 グローバルエポックごとに、クライアントサイドモデルを集約サーバに送信する。集約サーバはそれらのモデルを FedAvg アルゴリズムを利用して集約し、クライアントに返信する。モデルの同期のプロセスを排除するため、クライアントごとの学習を同期的に行うことが可能となり、学習時間を大幅に削減することができる。また、分割学習と同様の精度や通信効率を実現している。

他に、これらの並行分割学習の手法の弱点である収束の遅さや

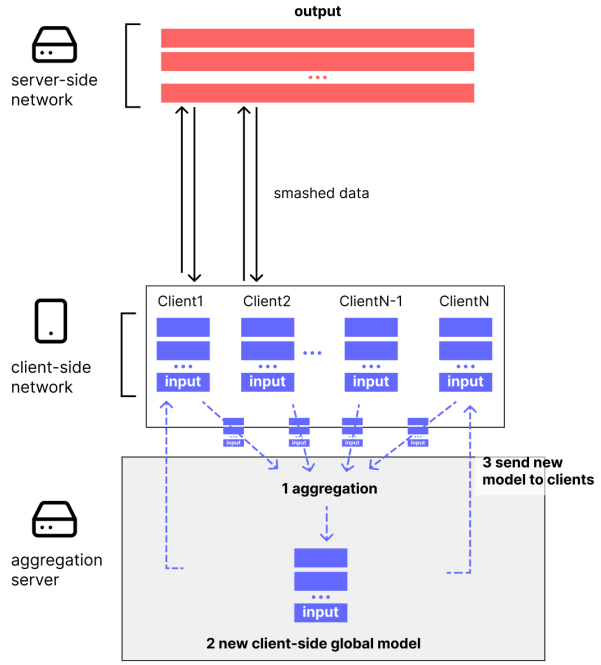


図 4　SplitFed の概略。各クライアントが、自身が持つモデルを集約サーバに送信し、集約処理が行われたのち、新しいモデルがクライアントに戻される。このモデルが次のグローバルエポックでのクライアントモデルの初期値となる。

スケーラビリティの低さを解決した手法として、LocFedMix-SL が提案されている [22]。

3　提 案 手 法

この章では、非独立同分布におけるモデルの精度を向上させるための新しい並行分割学習の手法である MidSL (Split Learning with Middle layers) および MidSFL (Split Federated Learning with Middle layers) について説明する。

MidSL の特徴は、ネットワーク全体を 3 つの部分に分け、中間ネットワークを追加している点である。クライアントをクラスタリングし、クラスタごとに中間ネットワークを割り当てることで、中間ネットワークがクラスタ間のデータ分布の差異を吸収すると考えた。

MidSFL は、MidSL に加えて集約処理を追加している。集約処理を行う手法として SplitFed が存在するが、クライアントのデータ分布が大きく異なる場合でもすべてのクライアントに対して集約処理を行う。一方 MidSFL では、同一クラスタ内でのみネットワークの集約処理を行うため、精度が向上すると考えた。

3.1　MidSL

MidSL では、分割学習にクラスタリングを導入し、各クライアントをクラスタに分類する（図 5）。クラスタ数を n とし、クラスタ集合を、$C = \{C_1, C_2, ..., C_n\}$ で表す。

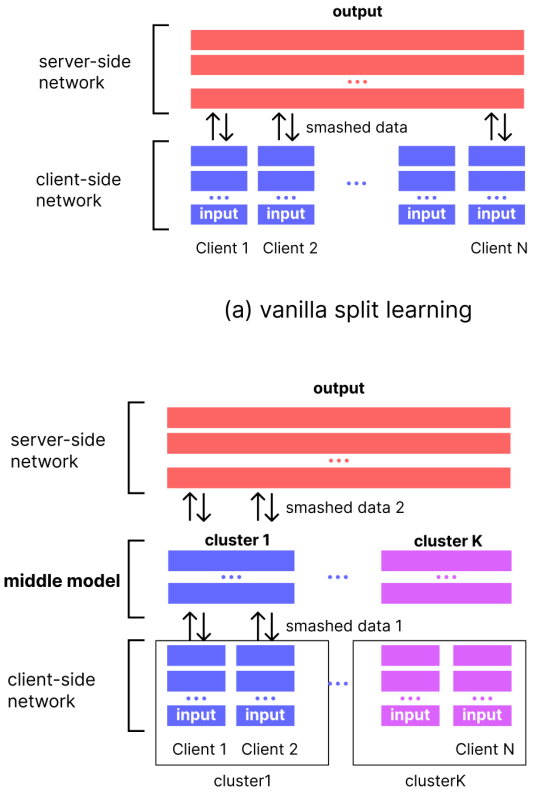


図 5　通常の分割学習（a）および MidSL（b）の概略。MidSL では、クラスタごとに固有の新しい中間ネットワークを導入し、3 つに分割された状態のニューラルネットワークで学習を行う。

MidSL では、ニューラルネットワーク全体を 3 つのサブネットワークに分割する。入力層を含むネットワークを W^c、出力層を含むネットワークを W^s、どちらも含まない中間のネットワークを W^m とする。W^c はクライアントごとに固有のネットワークを持ち、クライアント i が所有するネットワークを W_i^c とする。中間のネットワーク W^m はクラスタに割り当てられ、クラスタごとに固有のネットワークを持つ。クラスタ k が所有するネットワークを W_k^m で表す。W^s はサーバが所有する単一のネットワークである。また、W^c と W^m の間のカットオフレイヤーをカットオフレイヤー 1、W^m と W^s の間のカットオフレイヤーをカットオフレイヤー 2 と呼ぶ。

W^c はクライアントが管理し、W^m および W^s はサーバが管理する。

ニューラルネットワーク全体を一つの関数 F とみなすと、クラスタ j に所属するクライアント i のネットワーク全体は式 (1) のように表される。なお、X は入力データを表す。

$$F(X) = W^s(W_j^m(W_i^c(X))) \tag{1}$$

3.1.1　クラスタリング

次にクラスタリングを行うアルゴリズムについて説明する。サーバは各クライアントのローカルデータセットの分布を知る

Algorithm 2 MidSL における階層的クラスタリング

Server initializes server-side network W^s
Server initializes middle network W^m
Server initializes cluster set $K \leftarrow \{\{c_i \mid i \in [1, m]\}\}$
Each client initializes client-side network W^c
for each global epoch $e = 1, 2, ...$ **do**
  initialize S_e with empty array
  for Server selects client $i \in \{1, 2, ..., m\}$ **do**
    initialize s with empty array
    for each local epoch $e = 1, 2, ..., E_c$ **do**
      training procedure
      append smashed data to s
    end for
    append $avg(s)$ to S_e
  end for
  $K \leftarrow HierarchicalClustering(S_e)$
end for

ことができないため、smashed data を利用してクラスタリングを行う。クラスタリングの流れは次の通りである。

（1） 全てのクライアントを独立したクラスタとみなす。

（2） 全クラスタに対して、クラスタ間の距離を計算する。

（3） 最も類似性の高いクラスタのペアを選び、それらのクラスタを結合して 1 つのクラスタとする。

（4） 残るクラスタ数が 2 になるまで、1 から 3 を繰り返す

MidSL においては、このクラスタリングの操作をグローバルエポックごとに行う。1 グローバルエポックの中でクライアントからサーバに送られてきた smashed data の平均を、そのクライアントを表す値とし、この値を元にクラスタリングを行う。

実際の学習では、エポックごとにクライアントサイドモデルが更新され、smashed data も変化するため、クラスタリング結果が常に一定となるとは限らない。そこで、直前数エポックのクラスタリング結果を記録しておき、同一のクラスタリング結果が一定エポック以上続いた場合のみ、そのクラスタリングを適用する。すなわち、クラスタリング結果が安定した段階でクラスタの分割を行う。

また、この操作は一度クラスタリングが行われた後も、同一クラスタ内でさらにクラスタリングを行う。すなわち、エポックを重ねるごとに 2 分割を繰り返しながら適切なクラスタリングに収束する。

3.1.2 学　　習

サーバは、学習に参加するクライアントを 1 つ選択する。選択されたクライアントは、ローカルデータセットからミニバッチを生成し、クライアント側のネットワーク W_i^c に入力する（図 5）。順伝播を計算し、計算がカットオフレイヤー 1 に達すると、その出力をサーバに送信する。サーバは出力を受け取ると、クライアント i が所属するクラスタに対応する中間ネットワークである W_k^m に入力する。中間ネットワークにおける順伝播を計算し、カットオフレイヤー 2 に達すると、その出力を W^s に入力する。このようにして順伝播を計算する。損失を計算後、逆伝播も同様に、W^s、W_k^m、W_i^c の順に計算を行う。

学習全体の流れを Algorithm 3 にまとめる。

Algorithm 3 MidSL のアルゴリズム

Server initializes server-side network W^s
Server initializes middle network W^m
Server initializes cluster set $K \leftarrow \{\{c_i \mid i \in [1, m]\}\}$
Each client initializes client-side network W^c
for each global epoch $e = 1, 2, ...$ **do**
  initialize S_e with empty array
  for Server selects client $i \in \{1, 2, ..., m\}$ **do**
    Server: initialize s with empty array
    For $K' \in K$ where $i \in K'$: $k \leftarrow i$
    for each local epoch $e = 1, 2, ..., E_c$ **do**
      Client: $s_1^e \leftarrow W_i^c(B_{i_b})$
      Client: sends s_1^e to server
      Server: $y_1^e \leftarrow W^s(s_1^e)$
      Server: $y_2^e \leftarrow W_k^m(y_1^e)$
      Server: $gradient \leftarrow Gradient(y_2^e, label)$
      Server: $W^{s^{e\prime}}, sg_1^e \leftarrow W^{s^e T}(gradient)$
      Server: $W_k^{m\prime}, sg_2^e \leftarrow W^{m T}(sg_1^e)$
      Client: $W_i^{c\prime} \leftarrow W_i^{c T}(sg_2^e)$
      Server: append $avg(s)$ to s
    end for
    append $avg(s)$ to S_e
  end for
  $K \leftarrow HierarchicalClustering(S_e)$
end for

3.2 MidSFL

MidSFL では、MidSL を拡張し、クライアント側のネットワークを集約するステップを追加する。MidSFL では、グローバルエポックごとに、1. 学習、2. 集約 の 2 つのフェーズを行う（6）。学習フェーズでは、MidSL と同様に、クライアントとサーバが協調して学習を行う。全クライアントの学習が終了後、集約のフェーズに移る。

集約のフェーズでは、全クライアントが、クライアント側のネットワークを集約サーバに送信する（図 6）。集約サーバでは、クラスタごとに、クライアント側のネットワークの集約処理を行う。集約されたモデルは各クライアントに返却され、クライアントの次の学習における初期ネットワークとなる。学習全体の流れを Algorithm4 にまとめる。

4 MidSL に関する実験

4.1 事前クラスタリング

MidSL に関する実験では特に言及がない限り、Fashion-MNIST のデータセットを利用し、画像分類のタスクの学習を行なった。Fashion-MNIST は、10 種類のファッション商品の画像からなるデータセットで、個々の画像は 28 ピクセル四方の正方形である。各画像はグレースケールで、色は 0 から 255 までの 256 段階で表現される。モデルには MobileNetV2 を利用し、学習に参加するクライアント数は 30 とした。学習データは 30,000 枚の画像を利用し、これを全クライアントで等分

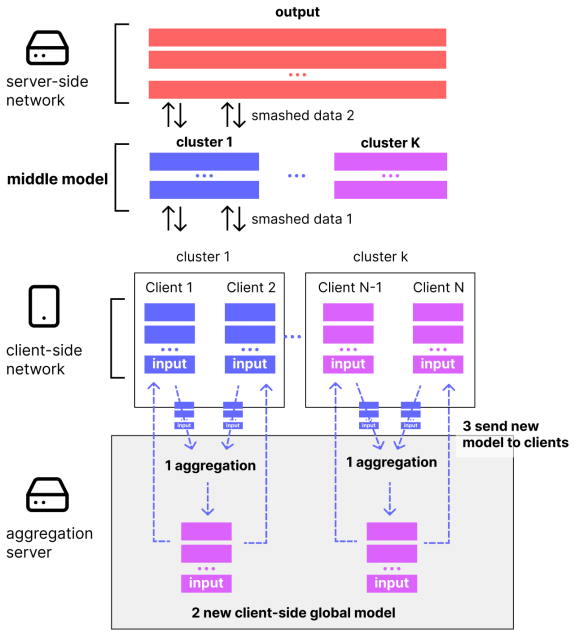


図 6　MidSFL の概略。MidSFL では、クラスタごとに集約処理を行う。同じクラスタに属するクライアントは、自身のネットワークを集約サーバ（Aggregation Server）に送信し、集約処理が行われ、返却されたネットワークを次のグローバルエポックでの学習に利用する。

表 1　MidSL に関する実験で利用したデータ分布

クラスタ	クラスタ 1（C1,...,C5）	クラスタ 2（C6,...,C10）
ラベル 0,1,2	1000 枚ずつ	0 枚ずつ
ラベル 3,4,5	0 枚ずつ	1000 枚ずつ
ラベル 6,7,8,9	500 枚ずつ	500 枚ずつ

Algorithm 4 MidSFL のアルゴリズム

Server initializes server-side network W^s
Server initializes middle network W^m
Server initializes cluster set $K \leftarrow \{\{c_i \mid i \in [1, m]\}\}$
Each client initializes client-side network W^c
for each global epoch $e = 1, 2, ...$ **do**
  initialize S_e with empty array
  for Server selects client $i \in \{1, 2, ..., m\}$ **do**
    Server: initialize s with empty array
    For $K' \in K$ where $i \in K'$: $k \leftarrow i$
    for each local epoch $e = 1, 2, ..., E_c$ **do**
      $k \leftarrow$ Cluster which client i belongs
      Client: $s_1^e \leftarrow W_i^c(B_{i_b})$
      Client: sends s_1^e to server
      Server: $y_1^e \leftarrow W^s(s_1^e)$
      Server: $y_2^e \leftarrow W_k^m(y_1^e)$
      Server: $gradient \leftarrow Gradient(y_2^e, label)$
      Server: $W^{s^{e\prime}}, sg_1^e \leftarrow W^{s^e T}(gradient)$
      Server: $W_k^{m\prime}, sg_2^e \leftarrow W^{m T}(sg_1^e)$
      Client: $W_i^{c\prime} \leftarrow W_i^{c T}(sg_2^e)$
      Server: append $avg(s)$ to s
    end for
    append $avg(s)$ to S_e
  end for
  $K \leftarrow HierarchicalClustering(S_e)$
  for each client $i \in \{1, 2, ..., m\}$ **do**
    Client: sends W_i^C to AggregationServer
  end for
  AggregationServer: $W^C \leftarrow \Sigma_{i=1}^m \frac{n_k}{n} W_i^C$
  AggregationServer: sends W^C to each client
end for

している。10 クライアントの場合は、1 クライアントあたり 3,000 枚の学習データを所有する。また、分割学習におけるモデルの同期は行わない設定で実験を実施した。

また、カットオフレイヤーの選択は MobileNetV2 のブロック単位で行う。これは、MobileNetV2 のネットワークで利用されている逆残差ブロック (Inverted residual block) が分割不可能であるためである。

カットオフレイヤー 1 として、4 ブロック目、カットオフレイヤー 2 として 9 ブロック目を選択した。また、クライアント数は 30 とした。これらのクライアントを 2 つのグループに分割して、同一グループに所属するクライアントのデータは同分布であり、異なるグループに所属するクライアントでのデータ分布は異なるような設定とした。具体的には、表 1 のような分布を利用した。

比較対象として 2 つのベースラインを設定した。ベースラインはどちらも通常の分割学習を行う。ベースライン 1 はカットオフレイヤーが 4 ブロック目、ベースライン 2 はカットオフレイヤーが 9 ブロック目である。ベースライン 1 は、クライアントの計算量が MidSL と等しく、クライアントに与えられた計算資源が極めて小さい IoT デバイスのような場合を想定している。ベースライン 2 は、MidSL やベースライン 1 と比較して、クライアントが担う計算量が大きい。

ベースライン 1、ベースライン 2、MidSL について、50 エポックの学習を行い、エポックごとのモデルのロス値と精度を確認した。また、参考とするため同様の条件で連合学習も行った。

結果は図 7 のようになった。MidSL は、ベースライン 2 を上回ったものの、ベースライン 1 とは大きな違いが見られなかった。実際のユースケースとして、クライアントとして IoT デバイスを利用するような環境では MidSL の有効性を示すことができなかった。一方で、MidSL とベースライン 1 の精度がほとんど等しくなったことから、MidSL におけるクラスタリングよりも、クライアントが担うレイヤーを減らすことが精度の向上に大きく寄与していることが分かる。また、連合学習と比較して精度が大幅に劣っている。これは、分割学習にモデル同期の処理がなく、特にクライアント側ネットワークの知識の統合が行われないためであると考える。

4.2　動的なクラスタリング

次に、MidSL において、動的なクラスタリングを行う実験を行なった。動的なクラスタリングでは、各グローバルエポックごとに階層的クラスタリングのアルゴリズムを利用してクラ

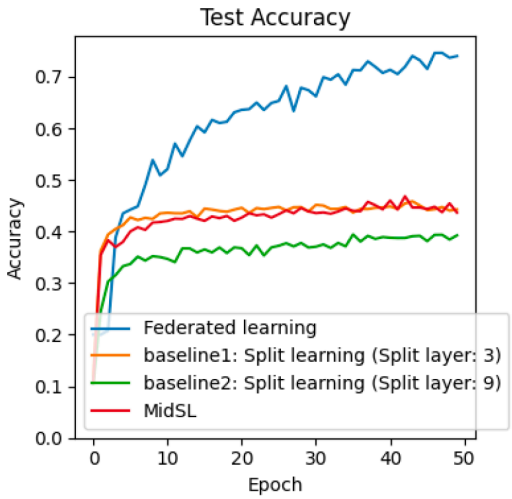


図 7　実験 1: 連合学習、ベースライン 1、ベースライン 2、MidSL それぞれのエポックごとのテスト精度の変化

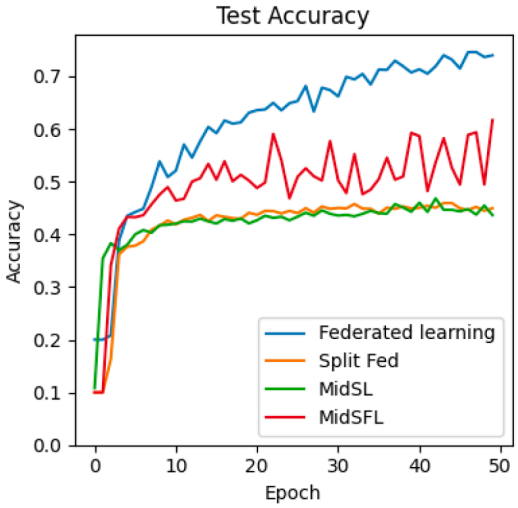


図 9　実験 2: MidSFL、SplitFed、連合学習、MidSL、それぞれにおける、エポックごとのテスト精度の変化

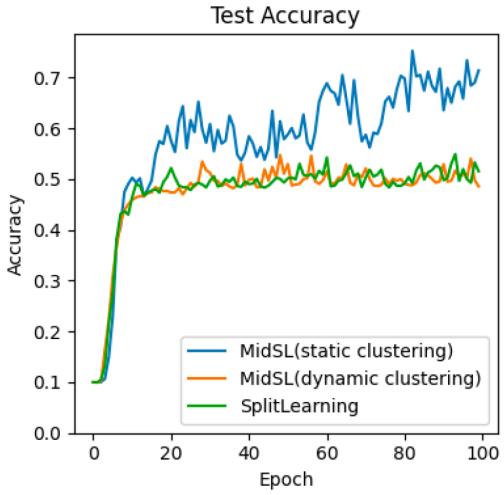


図 8　MidSL、ベースラインのそれぞれにおける、エポックごとのテスト精度の変化。
MidSL については、事前にクラスタが既知である場合（static clustering）、動的にクラスタリングを行う場合（dynamic clustering）のそれぞれについて実験を行なった。

スタの計算を行う。毎エポック行うクラスタリングを行った結果、一定回以上連続して同じ結果が得られた場合に、実際のそのクラスタの分割を行う。

クライアント数を 200 とし、100 エポックの学習を行った。提案手法である MidSFL に関して、以下の 2 つのパターンで実行する。

（1） 事前にクラスタが既知である場合

（2） 3 エポック以上連続して同じクラスタ結果が得られた場合にクラスタリングを行う場合

これに加えて、ベースラインとして分割学習を用いた。

結果は、図 8 のようになった。事前にクラスタリングが既知である場合にはベースラインを上回る結果を得たものの、学習中にクラスタリングを行う場合はベースラインとほとんど変わらないテスト精度となった。これは、クラスタリング結果が全く安定せず、一度もクラスタリングが適用されなかったためである。

5　MidSFL に関する実験

5.1　事前クラスタリング

次に、MidSFL に関する実験を行なった。ベースラインとして、SplitFed を利用する。50 エポックの学習を行い、図 9 のような結果を得た。参考のため、前章の実験で得られた MidSL および連合学習のデータのプロットしている。

MidSFL が、MidSL や SplitFed を上回る結果となった。これは、ベースラインで行われる全クライアントを対象としたクライアント側のネットワークの集約処理では異なる分布を持つクライアントのネットワークもまとめて集約していたのに対し、MidSFL では類似した分布を持つネットワークのみを集約しているためであると考えられる。

5.2　動的なクラスタリング

次に、MidSFL において、動的なクラスタリングを行う実験を行なった。予備実験を行い、この実験では 3 回以上同じ結果が得られた場合にクラスタの分割を行う設定とした。クライアント数を 200 とし、100 エポックの学習を行う。提案手法である MidSFL に関して、以下の 3 つのパターンで実行する。

（1） 事前にクラスタが既知である場合

（2） 10 エポック目時点で 1 度だけクラスタリングを行う場合

（3） 3 エポック以上連続して同じクラスタ結果が得られた場合にクラスタリングを行う場合

これに加えて、ベースラインとして SplitFed を用いた。

結果は、図 10 のようになった。3 つのパターンのいずれにおいても、ベースラインである SplitFed を上回る結果を得た。また、動的なクラスタリングを行う設定における MidSL と

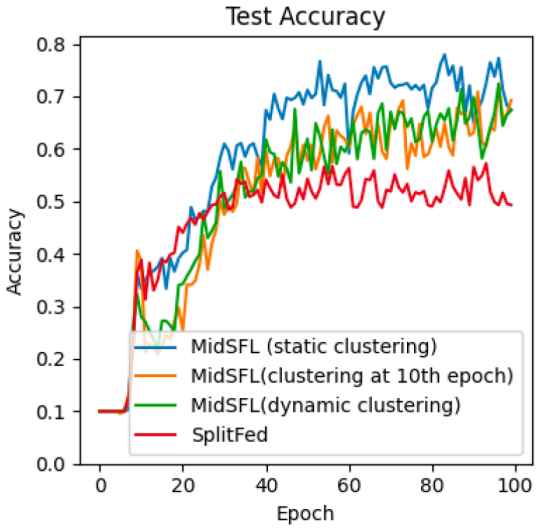


図 10　MidSFL とベースラインである SplitFed のそれぞれにおける、エポックごとのテスト精度の変化。
MidSFL については、事前にクラスタが既知である場合（static clustering）、10 エポック目でクラスタリングを行う場合（clustering at 10th epoch）、動的にクラスタリングを行う場合（dynamic clustering）のそれぞれについて実験を行なった。

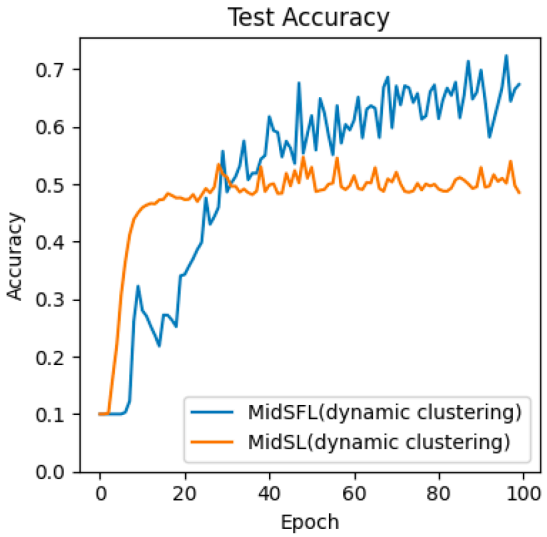


図 11　MidSL および MidSFL のエポックごとのテスト精度の比較。いずれも、学習中に動的にクラスタリングを行う設定。

MidSFL を比較すると、図 11 のように、MidSFL が上回る結果となった。

6　おわりに

本研究では、分割学習に階層的クラスタリングを適用する MidSL および MidSFL について提案を行った。

従来の分割学習には、非独立同分布の条件下でテスト精度が大きく低下する問題がある。MidSL および MidSFL は、ローカルデータセットのデータ分布が似ているクライアントをクラスタリングすることで、この問題の解決を試みる手法である。

MidSL を利用することで、サーバサイドネットワークのレイヤー数を揃えた条件で、ベースラインである通常の分割学習のテスト精度を超えることができた。しかし、クライアントサイドネットワークのレイヤー数を固定した条件下では、通常の分割学習のテスト精度を超えることができなかった。

一方、MidSL に連合学習のような集約処理を追加した MidSFL では、ベースラインである SplitFed のテスト精度を上回ることができた。これは、クラスタごとにローカルデータセットの分布が大きく異なる場合において、全てのクライアントのクライアントサイドネットワークをまとめて集約するより、クラスタ内で集約処理を行う方が効果的であったためであると考えられる。MidSFL は、学習中に動的にクラスタリングを実施する条件でも、ベースラインを上回った。

文　献

[1] Gupta, O. and Raskar, R.: Distributed learning of deep neural network over multiple agents, *Journal of Network and Computer Applications*, Vol. 116, pp. 1–8 (2018).
[2] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B. A.: Communication-efficient learning of deep networks from decentralized data, *Artificial intelligence and statistics*, PMLR, pp. 1273–1282 (2017).
[3] Gao, Y., Kim, M., Abuadbba, S., Kim, Y., Thapa, C., Kim, K., Çamtepe, S. A., Kim, H. and Nepal, S.: End-to-End Evaluation of Federated Learning and Split Learning for Internet of Things, *2020 International Symposium on Reliable Distributed Systems (SRDS)*, pp. 91–100 (2020).
[4] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z. B., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. Y. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, O., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D. X., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H. and Zhao, S.: Advances and Open Problems in Federated Learning, *Found. Trends Mach. Learn.*, Vol. 14, pp. 1–210 (2019).
[5] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D. and Chandra, V.: Federated Learning with Non-IID Data, *ArXiv*, Vol. abs/1806.00582 (2018).
[6] Arivazhagan, M. G., Aggarwal, V., Singh, A. K. and Choudhary, S.: Federated Learning with Personalization Layers, *ArXiv*, Vol. abs/1912.00818 (2019).
[7] Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F. and Ramage, D.: Federated Evaluation of On-device Personalization, *ArXiv*, Vol. abs/1910.10252 (2019).
[8] Sattler, F., Müller, K.-R. and Samek, W.: Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, pp. 3710–3722 (2019).
[9] Briggs, C., Fan, Z. and András, P.: Federated learning with hierarchical clustering of local updates to improve training on non-IID data, *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9 (2020).
[10] Gong, B., Xing, T., Liu, Z., Wang, J. and Liu, X.: Adaptive Clustered Federated Learning for Heterogeneous Data

in Edge Computing, *Mobile Networks and Applications*, Vol. 27, pp. 1520 – 1530 (2022).

[11] Vepakomma, P., Gupta, O., Swedish, T. and Raskar, R.: Split learning for health: Distributed deep learning without sharing raw patient data, *ArXiv*, Vol. abs/1812.00564 (2018).

[12] Wu, W., Li, M., Qu, K., Zhou, C., Shen, X. S., Zhuang, W., Li, X. and Shi, W.: Split Learning Over Wireless Networks: Parallel Design and Resource Management, *IEEE Journal on Selected Areas in Communications*, Vol. 41, pp. 1051–1066 (2022).

[13] Lin, Z., Zhu, G., Deng, Y., Chen, X., Gao, Y., Huang, K. and Fang, Y.: Efficient Parallel Split Learning over Resource-constrained Wireless Edge Networks, *ArXiv*, Vol. abs/2303.15991 (2023).

[14] Singh, A., Vepakomma, P., Gupta, O. and Raskar, R.: Detailed comparison of communication efficiency of split learning and federated learning, *ArXiv*, Vol. abs/1909.09145 (2019).

[15] Kim, M., DeRieux, A. C. and Saad, W.: A Bargaining Game for Personalized, Energy Efficient Split Learning over Wireless Networks, *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6 (2022).

[16] Cai, Y. and Wei, T.: Efficient Split Learning with Non-iid Data, *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, pp. 128–136 (2022).

[17] Chopra, A., Sahu, S. K., Singh, A., Java, A., Vepakomma, P., Sharma, V. and Raskar, R.: AdaSplit: Adaptive Trade-offs for Resource-constrained Distributed Deep Learning, *ArXiv*, Vol. abs/2112.01637 (2021).

[18] Vepakomma, P., Gupta, O., Dubey, A. and Raskar, R.: R EDUCING LEAKAGE IN DISTRIBUTED DEEP LEARNING FOR SENSITIVE HEALTH DATA (2019).

[19] Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D. J. and Jana, S. S.: Certified Robustness to Adversarial Examples with Differential Privacy, *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672 (2018).

[20] Thapa, C., Arachchige, P. C. M. and Çamtepe, S. A.: SplitFed: When Federated Learning Meets Split Learning, *ArXiv*, Vol. abs/2004.12088 (2020).

[21] Yang, X., Sun, J., Yao, Y., Xie, J. and Wang, C.: Differentially Private Label Protection in Split Learning, *ArXiv*, Vol. abs/2203.02073 (2022).

[22] Oh, S., Park, J., Vepakomma, P., Baek, S., Raskar, R., Bennis, M. and Kim, S.-L.: LocFedMix-SL: Localize, Federate, and Mix for Improved Scalability, Convergence, and Latency in Split Learning, *Proceedings of the ACM Web Conference 2022* (2022).

DEIM Forum 2024 XX-Y

エンコーダを用いた不均一モデル構造間の連合学習

松永　悠斗† 　籾山　悟至† 　エンケタイワンバトニヤマ† 　寺西　勇†

† 日本電気株式会社 セキュアシステムプラットフォーム研究所　〒211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †{yuto-matsunaga,satoru-momiyama,b-enkhtaivan,teranisi}@nec.com

あらまし　連合学習は、データのプライバシー保護と活用を両立させた分散機械学習手法の一つであり、サーバと複数のクライアントが協調して学習を行う。連合学習における最も基本的なアルゴリズム FedAvg では、全クライアントが共通のローカルモデルを持つことを前提としているが、実環境においてはクライアントによって計算機性能や下流タスクモデルが異なるため、各クライアントが異なるローカルモデルを必要とする可能性がある。本研究では、事前学習済みモデルと下流タスクモデルからなるローカルモデルに対し、異なる事前学習済みモデル間の出力を模倣するようなエンコーダを導入し、異なる下流タスクモデルの学習を可能とすることで、異なるモデル構造間の連合学習を行う。

キーワード　連合学習，不均一モデル構造

1　はじめに

近年、膨大なデータを用いて訓練された生成 AI [4] を始め、人工知能が社会に広く応用され、普及が著しい。その一方、個人情報保護に対する社会の意識が高まり、複数の国や地域で法整備が進められている。このような情勢に伴い、機械学習モデルの訓練に用いるデータのプライバシー保護技術も多数開発や研究されている [13]。その中で、連合学習は各クライアントが持つ他者との共有が難しいデータのプライバシー保護と活用を両立させることができる技術として注目を集めている [9]。

連合学習の基本的なネットワークは複数のクライアントと中央サーバから構成される。クライアントが各自にデータを所有し、互いに所有データを開示することなく機械学習モデルを訓練する。最も基本的なアルゴリズムである FedAvg [12] では、クライアントは所有データを用いてローカルモデルを更新し、そのモデルパラメータを中央サーバに送る。中央サーバは各クライアントのモデルパラメータを統合し、グローバルモデルを更新する。但し、FedAvg の適用条件は各クライアントのローカルモデルの構造が同一であることである（均一モデル構造）。一方で、クライアント毎に異なる構造のローカルモデルがある場合でも、連合学習を行うことができる [3]。InclusiveFL [10] や HeteroFL [2] では、各クライアントのモデル構造が全体共通部分に加えて個々のクライアントの計算資源に応じた拡張部分から構成される場合の不均一モデル構造間の連合学習が議論されている。これらとは別のアプローチとして、FedMD [8] や KT-pFL [15] は公開データに対する各クライアントのソフトラベルを統合し、それを各クライアントが学習する。FedHe [1] は公開データを用いず、各クライアントのローカルモデルの所有データに対するロジットを中央サーバで統合して、その結果をローカルモデルの訓練に用いる。上記の他に、FedMe [11] ではモデル交換を行い、ラベル無しのデータに対する出力がローカルモデルと類似する他クライアントのモデルと深層相互学習することで連合学習を行う。FedHe と FedMD では、各クライアントが同一のモデル構造を持ち、FedAvg や FedAvg の発展型である FedAdam [14] を用いて連合学習を行った場合より精度が低下することが報告されている [1] [10]。

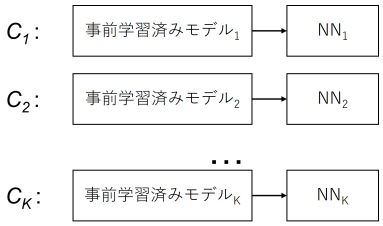


図 1: クライアントのモデル構造。

精度の面から、全クライアントが全モデル構造に関して FedAvg を使用する可能性が考えられるため、本論文では、FedAvg を活用する新しい方式を提案する。特に、大規模な事前学習済みのモデルを特徴量抽出器として用いる場合（例えば、[6]）を想定し、クライアントが所有するローカルモデルとして図 1 の様に、事前学習済みのモデルと下流タスクのニューラルネットワーク（NN）の組み合わせになっている状況を想定している。これは、近年、映像データや言語データを扱う用途が増えるにつれ、モデルを最初から学習するよりも、大規模な事前学習済みのモデルを特徴量抽出用にモデル構造の前段で用いることが増えている [5] ためである。

各クライアントが不均一モデル構造を持つ場合に単純に FedAvg を適応する極端な例は、各クライアントが自分のモデルの他に、他のクライアントの事前学習済みモデルと下流タスクモデルを持ち、全モデルに対して学習を行った後、同一構造のモデル間で FedAvg を行うことである。しかしこのアプローチは、各クライアントの計算機資源に限りがあることや学習全体に時間がかかることから現実的ではない。そのため、本稿では各クライアントが自身の事前学習済みモデルのみを用いて、他のクライアントの下流タスクモデルを学習できるようにする

ことで、下流タスクモデルに対して FedAvg の適応を可能とした連合学習手法を提供する。各クライアントの事前学習済みモデルから他のクライアントの下流タスクモデルを学習できるようにするために、提案方式では入力サンプルに対する事前学習済みのモデル $_i$ の出力を他の事前学習済みのモデル $_j$ の出力にマッピングするエンコーダ（**エンコーダ** $_{ij}$）を用いる。エンコーダは公開データを用いて事前に訓練することができる。これにより図 2 の様に、各クライアントは全ての事前学習済みのモデルに推論結果を自身の所有する事前学習済みのモデルとエンコーダから得ることができ、低コストで全モデル構造連合学習を実行することができる。

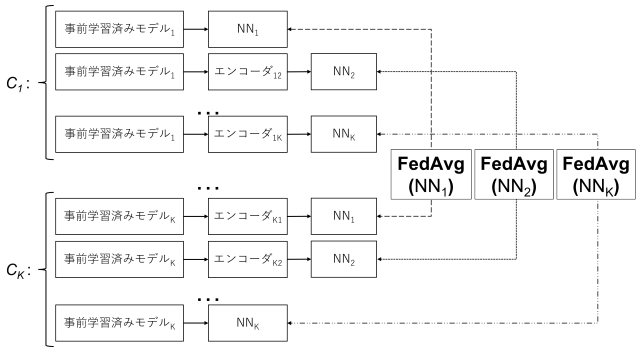


図 2: 提案方式の概要

我々の貢献は以下の通り：

- 不均一モデル構造間の連合学習の新方式を提案した。
- 典型的な画像分類データセットである CIFAR-10 を用いた性能評価により有効性を検証した。

2 関 連 研 究

2.1 InclusiveFL

Liu らによる InclusiveFL は、クライアントの計算資源に応じて異なるサイズのモデルを割り当てることで、計算資源の小さいクライアントでも連合学習に参加できるようにした手法である [10]。クライアントは計算資源に応じてグループ分けされ、グループに応じたモデルを割り当てられる。各モデルの基本構造は同じであり、グループに応じて層の深さが異なる。モデルの統合は、初めに各グループのモデル構造に対応したサブグローバルモデルが各グループに属するクライアントによって gradient を用いた FedAdam [14] で統合される。その後、サブグローバルモデル間で共通している浅い層は FedAvg によって統合される。モデルサイズの小さいサブグローバルモデルの最終層は、次に大きいサブグローバルモデルのより深い層の gradient の平均によって更新される。InclusiveFL は、公開データを用いる必要がなく FedAvg を活用した手法であるが、共通のモデル構造をもつ必要がある。

2.2 HeteroFL

Diao らによる HeteroFL [2] も InclusiveFL と同様に、クライアントの計算資源に応じて異なるサイズのモデルを割り当てる手法である。各クライアントは計算資源に応じてグループ分けされ、グループに応じたモデルを割り当てられる。各モデルの基本構造は同じだが、クライアントに応じて層の深さと幅 (層毎のサイズ) が異なる。モデルの統合は、グループ毎に一度統合された後各層において幅が一番小さいグループから順にグローバルモデルへ統合され、次に幅の小さいグループは前のグループの差集合となるパラメータを統合し、幅が一番大きいグループの統合になるまで繰り返す。HeteroFL は、InclusiveFL よりもモデルに自由度があるが、モデルの基本構造は同じである必要がある。

2.3 FedHe

Chan らによる FedHe は、サーバによる統合時にローカルモデルのパラメータではなく各クライアントの所有データに対する Logits を統合することで、各クライアントが異なるローカルモデル構造を取っている場合でも連合学習を行えるようにした手法である [1]。始めに、各クライアントは、ローカルモデルを用いて自身の所有データに対する Logits を計算し、ラベル毎の Logits の平均値をサーバに送信する。次に、サーバは全クライアントから得られた Logits の平均値に対してラベル毎に平均を計算し、全クライアントに送信する。次に、各クライアントは、ローカルモデルの学習をする際に、所有データの正解ラベルに対する loss の他に、サーバから得られたラベル毎の Logits の平均値と所有データに対する Logits 間の loss も含めて計算する。

3 準　　備

3.1 連 合 学 習

教師あり学習に関する連合学習には K 台のクライアント $C_1, C_2, \ldots, C_K$ と 1 台のサーバが参加しており、各 $k \in \{1, \ldots, K\}$ に対し、k 番目のクライアントは説明変数と目的変数のペアからなるデータ集合 $D_k = \{(x_n^k, y_n^k)\}_{n=1}^{N_k}$ を保有している。D_k のデータ数を $N_k := \#D_k$ とし、$N = \sum_k D_k$ とする。

連合学習の目的は、$D_1, \ldots, D_K$ をすべて訓練データ使った機械学習モデル $f(\cdot, \theta_{\text{global}})$ を、各 D_k を他のクライアントに開示する事なく訓練する事である。

連合学習の訓練手順を典型的な連合学習アルゴリズムである FedAvg を例に説明する。サーバはまず θ_{global}^0 を初期化して各クライアントに θ_{global}^0 を配布する。そして $t = 1, \ldots, T$ に対して以下を実行する。

- **ローカルモデルの更新：** 各参加者 C_k は $\theta_k^t \leftarrow \theta_{\text{global}}^t$ とし、D_k を用いて次の計算を T_c 回実行した後、θ_k^t をサーバに送信する。

$$\theta_k^t \leftarrow \theta_k^t - \alpha \sum_{n=1}^{N_k} \nabla_\theta \mathcal{L}(f(x_n^k, \theta_k^t), y_n^k) \tag{1}$$

ここで、α は学習率、$\mathcal{L}(\cdot)$ はロス関数である。

- **モデル結合：** サーバは以下のように θ_{global}^t を計算し、θ_{global}^t を各クライアントに送信する。

表 1: 変数名

変数	内容
K	クライアント数
x^e	エンコーダ学習用の公開データセット
P	事前学習済みモデルの総数
f_p^{pre}	事前学習済みモデル ($p \in P$)
$f_{p,p'}^{en}$	f_p^{pre} から $f_{p'}^{pre}$ へのエンコーダモデル ($p' \in P, p' \neq p$)
$\theta_{p,p'}^{en}$	$f_{p,p'}^{en}$ のパラメータ
f_p^{post}	下流タスクモデル ($p \in P$)
θ_p	f_p^{post} のパラメータ クライアント k 内のものは θ_p^k と表記

$$\theta_{\text{global}}^{t+1} \leftarrow \sum_{k=1}^{K} \frac{N_k}{N} \theta_k^t \tag{2}$$

FedAvg から派生したアルゴリズムはたくさんあるが、本論文では基本的なアルゴリズムである FedAvg を使用する。式 (2) から分かるように、FedAvg にはクライアント間でモデル構造が同一でなければいけないという制約がある。そのため、参加者間で不均一モデル構造を持つ場合に、そのままでは適用できない。

4 提案手法

4.1 フレームワーク

本節では、本手法の全体的なフレームワークを説明する。本手法の研究課題は、連合学習に参加するクライアントが異なるモデルを持つ場合であっても、FedAvg を利用することである。この課題を解決するために本手法では、図 1 のように各クライアントが事前学習済みモデルと下流タスクモデルからなるモデルを持つと仮定し、モデル構造の異なる他クライアントの下流タクスモデルと自身の事前学習済みモデルの橋渡しを行うようなエンコーダを導入することで、異なる下流タスクモデルに対しても連合学習を可能とする。このエンコーダは、異なる事前学習済みモデル $_1$ と事前学習済みモデル $_2$ に対して、事前学習済みモデル $_1$ の出力から事前学習済みモデル $_2$ の出力をマッピングするような変換器として学習する。この学習は、サーバ上で一般的な公開データセットを用いて、事前学習済みモデル全てに対して事前に行われる。

本手法において、各クライアントが持つモデルの構造は、図 2 のように、自身が所持する事前学習済みモデルと下流タスクモデルの他に、自身の事前学習済みモデルから他の事前学習済みモデルに対するエンコーダと対応する下流タスクモデルを持つ。連合学習時は、各クライアントが自身の所有データを用いて全下流タスクモデルの学習を行い、サーバが全クライアントから下流タスクモデルを収集し、FedAvg によって統合する。既定のラウンド数終了後、各クライアントは自身の事前学習済みモデルと下流タスクモデルを用いて、自身の下流タスクに取り組む。

図 1 のように各クライアントがユニークな事前学習済みモデルを持つ場合、クライアントの総数を K とするとエンコーダを $K(K-1)$ 個学習する必要があり、クライアント数が増えるほど事前に学習するエンコーダの数と連合学習時に各クライアントが学習する下流タスクモデルの数が増え、学習全体にかかるコストが大きくなる。そのため、各クライアントの計算機性能から、同等の計算機性能を持ったクライアントをグループ分けし、各グループに事前学習済みモデルを割り当てることでエンコーダ数を削減する。

Algorithm 1 エンコーダの学習

Require: 学習済み機械学習モデル $(f_p^{pre})_{p\in\{1,...,P\}}$、ラベルなし公開データ集合 $\{x^e\}_{e\in\{1,...,E\}}$、学習前のエンコーダ $(f_{p,p'}^{en}(\cdot,\theta_{p,p'}))_{p,p'\in\{1,...,P\}}$、エンコーダ学習の epoch 数 T_e
Ensure: エンコーダの学習済みパラメータ $(\theta_{p,p'})_{p,p'\in\{1,...,P\}}$
1: **for** $p \in \{1,2,...,P\}$、$e \in \{1,\ldots,E\}$ **do**
2: $\quad y_p^e \leftarrow f_p^{pre}(x^e)$
3: **end for**
4: **for** $p, p' \in \{1,2,...,P\}$ s.t. $p \neq p'$ **do**
5: $\quad$ **for** $i \in \{1,2,...,T_e\}$ **do**
6: $\quad\quad \mathcal{L}_{p,p'}^{en} \leftarrow \sum_{e=1}^{E} \mathsf{LossFunction}(f_{p,p'}^{en}(y_p^e, \theta_{p,p'}^{en}), y_{p'}^e)$
7: $\quad\quad \theta_{p,p'}^{en} \leftarrow \theta_{p,p'}^{en} - \eta \frac{\partial}{\partial \theta_{p,p'}^{en}} \mathcal{L}_{p,p'}^{en}$
8: $\quad$ **end for**
9: **end for**

4.2 アルゴリズム

本節では、本手法のエンコーダ学習と連合学習のアルゴリズムの説明を行う。本節で用いる表記法は表 1 の通りである。エンコーダ学習の疑似コードをアルゴリズム 1 に示す。始めに、サーバ上の公開データセットを用いて全事前学習済みモデル出力を得る (1〜3 行目)。次に、学習済みモデル p の出力 y_p を入力データ、学習済みモデル $p'(p \neq p')$ の出力 $y_{p'}$ を教師データとしてエンコーダのパラメータ $\theta_{p,p'}^{en}$ を学習する (4〜9 行目)。このエンコーダの学習を $P(P-1)$ 種類のエンコーダに対して行う。

連合学習の疑似コードをアルゴリズム 2 に示す。始めに、各クライアントは自身の所有データ x_n^k を用いて、事前学習済みモデルの出力 $y_n^{pre,k}$ を得る (6 行目)。次に、自身の下流タスクモデルの出力 $\dot{y}_{m(k),n}^k$ を得る (7 行目)。事前学習済みモデルの出力 $y_k^{pre,k}$ とエンコーダを用いて、他の下流タスクモデルに対する出力 $\dot{y}_{p,n}^k$ を得る (8〜10 行目)。全下流タスクモデルの出力 $\dot{y}_{p,n}^k$ と所有データ y_n^k を用いて、下流タスクモデルのパラメータ θ_p^k を更新する (12〜15 行目)。各クライアントは最後に、下流タスクモデルをサーバに送信する (17 行目)

サーバは全クライアントから下流タスクモデルを受信後、同一の下流タスクモデル同士で FedAvg による統合を行う (20〜22 行目)。サーバは最後に、下流タスクモデルを全クライアントに配布する (23〜25 行目)。

5 評価実験

5.1 データセット

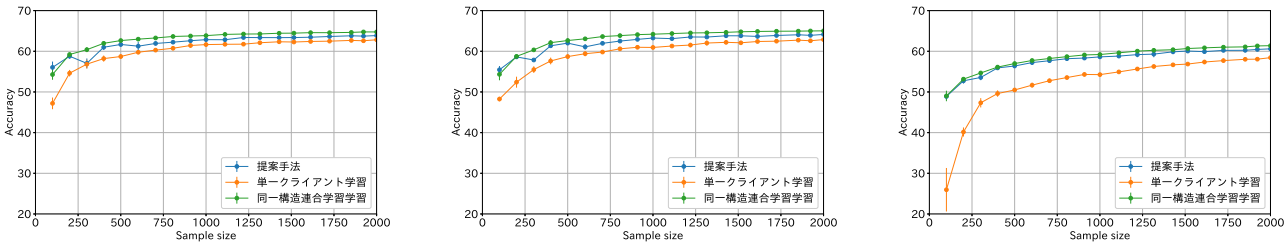
実験では画像分類タスク用データセットである CIFAR-10 [7]

表 2: 各クライアントが保有するモデル

	事前学習済みモデル	パラメータ数	下流タスク NN	パラメータ数
クライアント 1	ResNet-20	0.27M	線形層 (64×10)	650
クライアント 2	ResNet-56	0.85M		
クライアント 3	VGG13	9.4M	MLP $(512 \times 512 \times 512 \times 10)$	0.53M

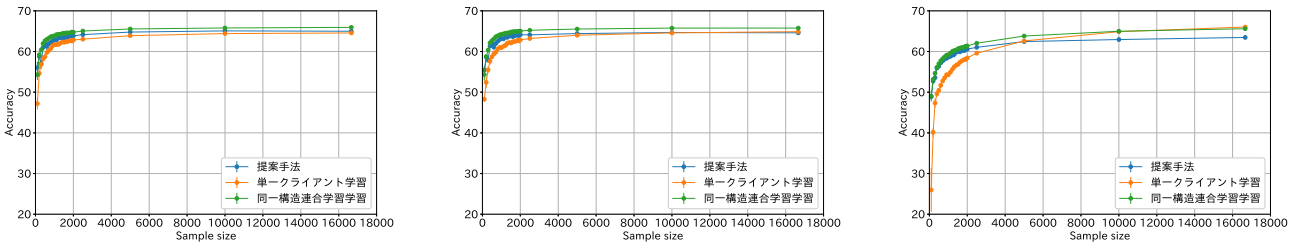
表 3: エンコーダ $f^{en}_{p,p'}$ に用いる MLP の構造

		$f^{pre}_{p'}$ 構造		
		ResNet-20	ResNet-56	VGG13
f^{pre}_{p} 構造	ResNet-20	-	$64 \times 256 \times 64$	$64 \times 256 \times 512$
	ResNet-56	$64 \times 256 \times 64$	-	$64 \times 256 \times 512$
	VGG13	$512 \times 256 \times 64$	$512 \times 256 \times 64$	-



(a) クライアント 1 (ResNet-20) (b) クライアント 2 (ResNet-56) (c) クライアント 3 (VGG13)

図 3: クライアント所有データの数が少ない場合の性能比較



(a) クライアント 1 (ResNet-20) (b) クライアント 2 (ResNet-56) (c) クライアント 3 (VGG13)

図 4: クライアント所有データの数が多い場合の性能比較

と CIFAR-100 [7] を用いた。CIFAR-10 は 10 クラスの 50000 枚の訓練画像と 10000 枚のテスト画像から構成される。CIFAR-100 は CIFAR-10 と同数の画像から構成されるが、クラス数は 100 である。そのため、1 クラスに 500 枚の訓練画像と 100 枚のテスト画像がある。CIFAR-10 と CIFAR-100 のいずれの画像も 32×32 の 3 チャネルのカラー画像である。事前学習済みのモデルは CIFAR-100 で訓練されたものを用いた。また、CIFAR-100 をオープンデータに見立てて、エンコーダも CIFAR-100 で訓練した。下流タスク用のニューラルネットワーク（NN）については CIFAR-10 で訓練した。

5.2 その他の実験設定

実験ではクライアント数が 3 つとし、表 2 に示すように、それぞれクライアント 1 が ResNet-20、クライアント 2 が ResNet-56 とクライアント 3 が VGG13 を事前学習済みのモデルを保有するものとした。エンコーダ用の NN は、表 3 の通り 2 層の MLP とし、モデルの出力次元に合わせてパラメータ数を調節している。下流タスク用の NN は、表 2 の通り構造がそれぞれ、ResNet-20 及び ResNet-56 の下流タスクには 1 層の線形層 (64×10) の、VGG13 の下流タスクには 3 層 MLP $(512 \times 512 \times 512 \times 10)$ を用いた。エンコーダの訓練の際の損失関数は平均二乗誤差を用いた（エンコーダの精度に関しては付録を参照せよ）。下流タスク NN の訓練ではクロスエントロピーを損失関数として用いた。各クライアントの所有するデータは全訓練データ内からランダムに抽出した（iid 設定）。本論文で示す実験結果は、各実験をすべて 5 回実施し、提案手法やベンチマークのテストデータに対する精度の平均と 95%信頼区間を示した。

5.3 ベンチマーク

単一クライアント学習：クライアントが自身の所有するデータのみを用いて下流タスク NN を訓練する場合である。この結

Algorithm 2 連合学習

Require: 学習済み機械学習モデル $(f_p^{pre})_{p\in\{1,\dots,P\}}$、訓練済みエンコーダ $(f_{p,p'}^{en}(\cdot,\theta_{p,p'}))_{p,p'\in\{1,\dots,P\}}$、ラウンド数 T, ローカル学習の epoch 数 T_c, クライアント k に割り振られた事前学習済みモデルの ID $m(k)$ $(m(k)\in\{1,\dots,P\})$、クライアント k のデータ集合 $D_k=(x_n^k,y_n^k)_{n\in\{1,\dots,N_k\}}$

Ensure: 学習済みのパラメータ $(\theta_p)_{p\in\{1,\dots,P\}}$

for $t\in\{1,2,...,T\}$ **do**
  // Local Update
  for each client $k\in\{1,2,...,K\}$ **in parallel do**
    for $i\in\{1,2,...,T_c\}$ **do**
      for $n\in\{1,\dots,N_k\}$ **do**
        $y_n^{pre,k}\leftarrow f_{m(k)}^{pre}(x_n^k)$
        $\dot{y}_{m(k),n}^k\leftarrow f_{m(k)}^{post}(y_n^{pre,k},\theta_{m(k)}^k)$
        for $p\in\{1,2,...,P\}$ $(p\neq m(k))$ **do**
          $\dot{y}_{p,n}^k\leftarrow f_p^{post}(f_{m(k),p}^{en}(y_n^{pre,k},\theta_{m(k),p}^{en}),\theta_p^k)$
        end for
      end for
      for $p\in\{1,2,...,P\}$ **do**
        $\mathcal{L}_p^k\leftarrow\sum_{n=1}^{N_k}\mathsf{CrossEntropy}(\dot{y}_{p,n}^k,y_n^k)$
        $\theta_p^k\leftarrow\theta_p^k-\eta\frac{\partial\mathcal{L}_p^k}{\partial\theta_p^k}$
      end for
    end for
    Upload $\theta_{p\in\{1,2,\dots,P\}}^k$.
  end for
  // Server Update
  for $p\in\{1,2,...,P\}$ **do**
    $\theta_p\leftarrow\sum_{k=1}^{K}\frac{N_k}{N}\theta_p^k$
  end for
  for each client $k\in\{1,2,...,K\}$ **in parallel do**
    Send θ_p to Device k.
  end for
end for

果として得られたローカルモデルの性能を「単一クライアント学習によるモデル性能」と呼ぶことにする。

同一構造連合学習：クライアントが所有する事前学習済みのモデルを他の全クライアントが持ち、下流タスク NN の連合学習を行った場合のモデル性能である。例えば、「クライアント 3 の同一構造連合学習によって得られたモデルの性能」とは、全クライアントが事前学習済みのモデル VGG13 を所有し、下流タスク NN の訓練を FedAvg により実施した場合のクライアント 3 のモデルの性能を指している。

5.4 実験結果

単一クライアント学習や同一構造連合学習との比較

図 3 と図 4 にそれぞれクライアントの所有するプライベートデータのサイズが少ない場合（データ数 2000 まで）と多い場合（データ数 18000 まで）の提案手法の性能とベンチマーク（単一クライアントや全構造連合学習）の性能を示す。図の縦軸は CIFAR-10 テストデータに対する下流タスク NN の分類精度で、横軸は各クライアントが所有するデータ数である。オレンジ、青色、緑色の実線はそれぞれ単一クライアント学習、提案手法による学習、同一構造連合学習によって得られたモデルの性能を示している。

図 3 からわかるように、各クライアントの所有するデータ数が少ない場合に提案手法の性能が単一クライアント学習によって得られたモデルの性能を上回っている。これは連合学習を行うことにより、他のクライアントからの知識が各クライアントのモデルに取り込まれていることを示していて、不均一な構造を持つモデル間で連合学習を行うメリットが現れている。さらに、クライアント 3 の提案手法により得られたモデルの性能は同一構造連合学習によって得られたモデルの性能と同等であり（図 3c）、ベンチマークのほぼ中間に位置するクライアント 1 とクライアント 2 のそれら（図 3a と 3b）とは異なる。これはパラメータ数の多いモデルの方が提案手法による連合学習のメリットをより享受していることを示唆している。

図 4 では、クライアントが所有するデータのサイズが大きくなると、単一クライアント学習と提案手法や同一構造連合学習によるモデル性能差が見られなくなっている。これはクライアントが十分な量のデータを所有していれば連合を行わなくてもよいという経験則と一致している。また逆説的に言えば、コスト等の問題がなければ連合学習を行った方が、性能向上の機会が得られるということも言える。図 4c では、クライアントが所有するデータの数が 4000 を超えると提案手法のモデル性能が他のベンチマークのそれに比べて性能が低くなっている。これはモデルサイズが大きく、且つクライアント所有データ数が多い場合に、提案手法が適していないことを示唆している。

5.5 考　　察

iid 設定においては、一般的な機械学習の経験則から、各クライアントが一番大きいサイズのモデルを持ち十分な所有データを持っている条件で、FedAvg により連合学習を行った場合が精度の面で上限になることが考えられる。しかし、この方法は各クライアントの計算資源が考慮されておらず、通常の連合学習では全クライアントがローカルに学習を終了させるのを待つ必要があるため、計算資源が少ないクライアントに対して他のクライアントに待機時間が発生することによって、全体の学習時間が長くなってしまう。特に、IoT デバイスによる連合学習のような、クライアント間で計算資源が大きく異なるような場合ではこの問題が深刻化し、現実的な時間で学習が終了しない可能性がある。提案手法は、各クライアントが自身の計算資源に合わせたモデルで連合学習に参加できるため、実現性や学習時間の面で有利であると考えられる。

本論文では扱っていないが、non-iid 設定において一番大きいサイズのモデルで統一された FedAvg は精度が落ちることが考えられる。しかし、提案手法は各クライアントがエンコーダにより他のクライアントの下流タスクモデルに対して知識を取り込ませることができるため、non-iid 設定において精度の面で有利であると考えられる。

5.6 既存手法との比較実験

本論文では連合学習においてシンプルな方法として FedAvg

をベンチマークとしたが、クライアント間でモデルが異なる連合学習の既存手法として FedHe との比較実験も行った。

実験に用いたデータセットや実験設定は前述の実験と同様の条件で行った。実験結果は、実験を 5 回実施した場合の平均と 95%信頼区間を示した。図 5 と図 6 にそれぞれクライアント所有のデータ数が少ない場合と多い場合の提案手法の性能と FedHe の性能を示す。図 5a と図 5b の結果から、事前学習済みモデルが ResNet-20 または ResNet-56 でクライアント所有のデータ数が少ない場合に、提案手法は FedHe と同等もしくはそれ以上の性能を持つことが確認できる上 (最大 10.4%提案手法が良い)。図 6c の結果は、事前学習済みモデルが VGG の場合にクライアント所有のデータ数に関わらず提案手法が FedHe の性能上回っていることが確認できる (最大 10.7%提案手法が良い)。しかし、図 6a と図 6b の結果は、クライアント所有のデータ数が多い場合に FedHe が提案手法の性能を上回っていることが確認できる (最大 1.3%FedHe が良い)。

これらのことから、モデルサイズが大きい場合には、提案手法が有用であると考えられる。また、モデルサイズが少ない場合でもクライアント所有のデータ数が少ない場合には、提案手法が有用であると考えられる。

5.7 今後の課題

提案手法は、通常の FedAvg と比較して実現性や学習時間の面で有利であると考えられるが、定量的な評価が必要となる。また、提案手法は、non-iid 設定でも有用と考えられるが実験による検証が必要である。本論文では実験結果から、パラメータ数の多いモデルの方が提案手法による連合学習のメリットをより享受していることが示唆されたが、ResNet と VGG というモデル構成の違いによる影響も考えられるため、より大きなサイズの ResNet や VGG、他の画像分類モデルを用いたさらなる調査が求められる。本論文の実験で用いたエンコーダは他の事前学習済みモデルに対して模倣の精度が 6 割程度であったが、それにもかかわらず最終的な連合学習モデルの精度が高いのは驚くべきことである。そのため、より精度の高いエンコーダや低いエンコーダを用いて、提案手法におけるエンコーダの性能が、連合学習の結果に与える影響について調べることが必要である。

6 おわりに

本研究では、事前学習済みモデルと下流タスクモデルからなるローカルモデルに対し、異なる事前学習済みモデルの出力を模倣するようなエンコーダを導入することで、異なる下流タスクモデルの学習を可能とすることで、異なるモデル構造間の連合学習を行った。一般的な画像分類データセットを用いた実験より、提案手法によって得られたモデルが十分な性能を有し、提案手法の有効性を検証できた。具体的には、提案手法で得られたモデルの性能が単一クライアント学習を上回り、理想的な連合学習である同一構造連合学習に迫っていることが分かった。

7 謝　辞

本研究開発は、総務省の「安全なデータ連携による最適化ＡＩ技術の研究開発（JPMI00316）」によって実施した成果を含みます。

文　献

[1] Y. Chan and E. H. Ngai. Fedhe: Heterogeneous models and communication-efficient federated learning. In *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, pages 207–214, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[2] Enmao Diao, Jie Ding, Vahid Tarokh, and Zachary Garrett. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[3] Dashan Gao, Xin Yao, and Qiang Yang. A survey on heterogeneous federated learning, 2022.

[4] Roberto Gozalo-Brizuela and Eduardo C. Garrido-Merchán. A survey of generative ai applications, 2023.

[5] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

[6] L. Heckler, R. Konig, and P. Bergmann. Exploring the importance of pretrained feature extractors for unsupervised anomaly detection and localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2917–2926, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[8] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *CoRR*, abs/1910.03581, 2019.

[9] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *CoRR*, abs/1907.09693, 2019.

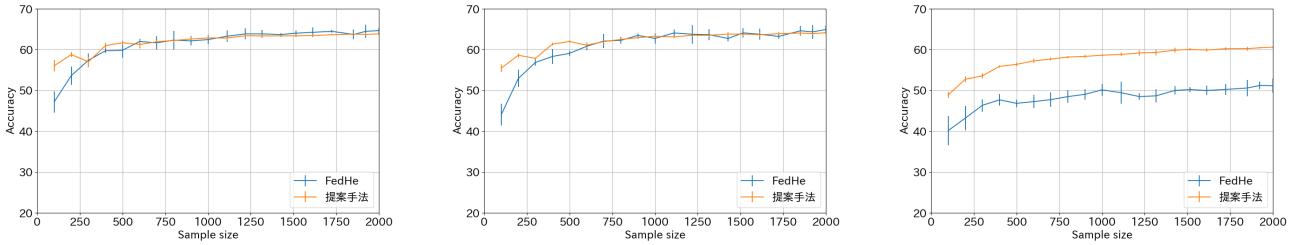
[10] Ruixuan Liu, Fangzhao Wu, Chuhan Wu, Yanlin Wang, Lingjuan Lyu, Hong Chen, and Xing Xie. No one left behind: Inclusive federated learning over heterogeneous devices. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3398–3406, New York, NY, USA, 2022. Association for Computing Machinery.

[11] Koji Matsuda, Yuya Sasaki, Chuan Xiao, and Makoto Onizuka. Fedme: Federated learning via model exchange, 2021.

[12] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

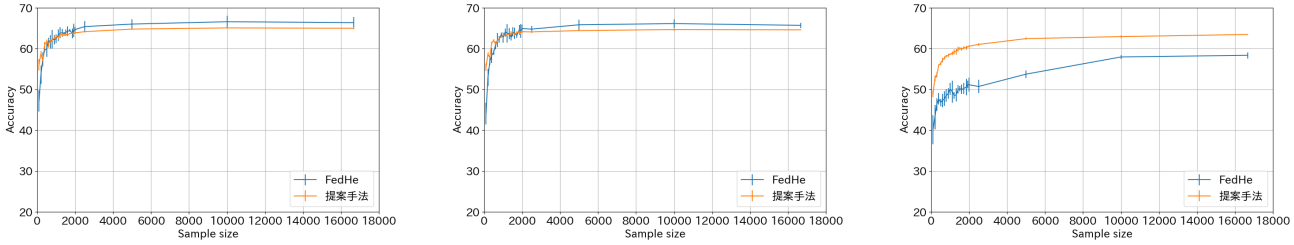
[13] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey. *CoRR*, abs/2004.12254, 2020.

[14] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *9th International Conference on Learning Representa-*



(a) クライアント 1 (ResNet-20)　(b) クライアント 2 (ResNet-56)　(c) クライアント 3 (VGG13)

図 5: クライアント所有データの数が少ない場合の FedHe との性能比較



(a) クライアント 1 (ResNet-20)　(b) クライアント 2 (ResNet-56)　(c) クライアント 3 (VGG13)

図 6: クライアント所有データの数が多い場合の FedHe との性能比較

表 4: エンコーダ $f^{en}_{p,p'}$ の精度

		$f^{pre}_{p'}$ 構造		
		ResNet-20	ResNet-56	VGG13
f^{pre}_{p} 構造	ResNet-20	-	64.6%	63.8%
	ResNet-56	62.5%	-	66.5%
	VGG13	63.3%	67.7%	-

tions, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021.

[15] Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *CoRR*, abs/2111.02862, 2021.

8　付録：エンコーダの性能

本研究で学習させたエンコーダの精度は表 4 の通り。ここでエンコーダ $f^{en}_{p,p'}$ の精度は、エンコーダ学習時のテストデータ x を入力として、正解データ $f^{pre}_{p'}(x)$ に対する $f^{en}_{p,p'}(f^{pre}_{p}(x))$ の精度として定義している。テストデータが持つ正解ラベル y に対する精度ではないことに注意されたい。

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

機械学習とプライバシ

座長:高橋 翼(LINEヤフー株式会社)
コメンテータ:佐々木 勇和(大阪大学)
2024年2月28日(水) 16:00 〜 18:10 T2-B (オンライン（Zoom Events))

17:45 〜 18:10

[T2-B-3-05]【技術報告】安全な LLMの実現を目指して - ストレステストとセーフアラインメントの取り組み

綿岡晃輝 (LINEヤフー株式会社)

本発表では, 生成AIの安全性を検査及び改善するために実施してきた取り組みを紹介します. 具体的には, 生成AIの安全性を可視化するシステム、LLMの脆弱性を発見するためのストレステスト、そして安全性を向上させるための半自動化手法の実験などについて説明します.

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 16:00 ~ 18:10 | T3-A オンライン（Zoom Events)

SNS分析 (3)

座長:平手 勇宇(楽天株式会社)
コメンテータ:大野 成義(職業能力開発総合大学校)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-A-3-01] 都市別のソーシャルメディア投稿を利用した市民の性格特性分析

*岩﨑 大晟[1]、関 洋平[2]、柏野 和佳子[3]、欅 惇志[4]、神門 典子[5,6] (1. 筑波大学 情報学群 知識情報・図書館学類、2. 筑波大学 図書館情報メディア系、3. 国立国語研究所、4. 一橋大学 ソーシャル・データサイエンス研究科、5. 国立情報学研究所、6. 総合研究大学院大学)

16:30 ~ 16:55
[T3-A-3-02] コロナ禍における草の観察

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

16:55 ~ 17:20
[T3-A-3-03] 降雨データを用いた SNS ユーザーの居住地推定

*山下 暁天[1] (1. 京都大学工学部情報学科)

17:20 ~ 17:45
[T3-A-3-04] ソーシャルメディアテキストからのユーザ行動の分類

*中岡 明義[1]、若宮 翔子[1]、荒巻 英治[1] (1. 奈良先端科学技術大学院大学)

17:45 ~ 18:10
[T3-A-3-05] 機械学習とクラウドソーシングを用いた顧客の声の自動分類

*鈴木 優[1]、城所 祥太[1]、北村 拓斗[1]、太田 奈那[1]、三島 惇也[1]、本田 文隆[2]、長徳 弘一[2]、中村 勇太[2]、佐々木 和彦[2]、笠丸 陽一[2] (1. 岐阜大学、2. ソニー（株）)

都市別のソーシャルメディア投稿を利用した市民の性格特性分析

岩崎　大晟† 　関　洋平†† 　柏野和佳子††† 　欅　惇志†††† 　神門　典子†††††,††††††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
††† 国立国語研究所　〒 190–8561 東京都立川市緑町 10 丁目 2 番地
†††† 一橋大学ソーシャル・データサイエンス研究科　〒 186–8601 東京都国立市中 2 丁目 1 番地
††††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2 丁目 1 番地 2 号
†††††† 総合研究大学院大学　〒 101–8430 東京都千代田区一ツ橋 2 丁目 1 番地 2 号
E-mail: †s1911457@s.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp, †††waka@ninjal.ac.jp,
††††a.keyaki@r.hit-u.ac.jp, †††††kando@nii.ac.jp

あらまし　本研究では，性格特性が個人の行動や反応の傾向に影響を与える重要な要素であることに着目し，個人の性格特性を表す Big Five Model を参照して，都市別の投稿を用いて市民の性格特性を分析する．具体的には，性格特性の地域差を明らかにするために，各都市の SNS 上の投稿に性格特性と経験性の二つのラベルを付与し，市民の行動の傾向がどのようにこれらの特性を反映しているかを検証する．性格特性は,「利他主義」,「芸術的興味」,「冒険心」,「社交性」,「活発度」の 5 つを定義し，人手でアノテーションデータを構築する．さらに，大規模言語モデルをファインチューニングし，ラベルの自動付与を行い，都市間の性格特性の傾向の差異を分析する．

キーワード　性格特性, 大規模言語モデル, 都市別分析, ソーシャルメディア, 行動分析

1 はじめに

企業のマーケティング戦略において，コミュニティの好みや行動の傾向を把握することは重要である．実際に，消費者の性格に合わせた広告がその効果をより高めたことが報告されている [6]．ソーシャルメディアの投稿には人々の好みや行動の傾向が反映されていると考えられ，多くの企業でソーシャルメディアへの投資額は増加している [13]．そこで本研究では，性格特性が個人の行動や反応の傾向に影響を与える重要な要素であることに着目し，個人の性格特性を主要な 5 つの因子で表現する Big Five Model [7] を参照して，都市別の投稿を用いて市民の性格特性を分析する手法を提案し，その有効性を検証する．5 つの主要な因子にはそれぞれ，ファセットと呼ばれるいくつかの側面因子が存在する．ファセットを分析することで，市民が持つ主要な因子の傾向の要因を明らかにできる．本研究では，都市という集団の性格特性を分析するにあたり，地域の特性や市民の活動が反映されると考えられるファセットを対象とする．

本研究で対象とするデータは，幅広い市民が利用する X（旧 Twitter）から収集する．実験における対象都市は，政令指定都市である福岡市，横浜市，札幌市とし，性格特性は協調性のファセットからは「利他主義」，開放性のファセットからは「芸術的興味」と「冒険心」，外向性のファセットからは「社交性」と「活発度」を分析対象とする．本研究では，これらのファセットを性格特性と呼ぶ．実験では，各性格特性が現れている投稿の割合を調査することにより，地域差を分析する提案手法の有効性について検証する．

2 関連研究

2.1 Big Five Model とそのファセットに関する研究

Big Five Model [7] は，人々の性格が 5 つの因子によって構成されるという理論である．5 つの因子は，開放性 (Openness), 外向性 (Extraversion), 協調性 (Agreeableness), 誠実性 (Conscientiousness), 神経症傾向 (Neuroticism) の 5 つである．Big Five Model は，人々の性格特性が言葉の使い方に反映されるという語彙仮説を基底として理論づけられた．

性格特性をより詳細に分析するために，Big Five Model の主要な各因子にはファセットと呼ばれる側面因子が導入された．Goldberg は，Big Five の主要な 5 つの因子のみを測定しようとすると，性格特性のより細かい側面を捉えるための情報が失われてしまうため，性格特性をより細かく測定するためにはファセット単位で測定を行うべきだと指摘している [5]．本研究ではファセットを扱うにあたり，IPIP-NEO-120 [8] を用いる．IPIP-NEO-120 では，1 つのファセットにつき 4 つの質問項目が設定されており,「非常に当てはまる」から「全く当てはまらない」までの 5 段階のスケールで答えることで性格特性を定量化する．IPIP-NEO-120 を参考にし，集団の性格特性を分析する上で有意義であると考えられるファセットを選択する．表 1 に，本研究で扱うファセットとそれに対応する質問項目を示す．

また，性格特性と行動が密接に関連している [4] という事実から，本研究で利用するデータセットには，投稿の性格特性の有無に加え，経験性の有無を付与する．経験性を付与することで，市民の行動の傾向がこれらの特性と関連しているかを分析することができる．

2.2 性格特性の地域差に関する研究

Schmitt ら [15] は，Big Five Model の性格特性を測定するために用いられるアンケート形式の評価ツールを利用して，56ヶ国の 17,837 人を対象に性格特性の傾向を調査した．その結果，開放性に関して，南アメリカと東アジアが他の地域に比べて有意に高いことが示された．このように，性格特性は文化間で差異が存在する．

日本国内においても性格特性の地域差があることが明らかになっている．吉野ら [17] は，日本国内において，外向性は東京都や千葉県を中心とする首都圏や沖縄県で高い傾向が見られ，開放性は福岡県や大分県，北海道などの地域で高い傾向があることを示した．この研究では，性格特性を予測する尺度として TIPI-J [18] を使用している．TIPI-J は質問項目数が少なく，Big Five Model の 5 つの因子のおおまかな傾向を把握するに留まる．吉野らは，今後は TIPI-J 以外の尺度でも結果を確かめる必要があると指摘している．また，この研究で用いられたデータセットは都道府県ごとにばらつきがある上，平均データ数は 1 つの都道府県につき 119 人であった．性格特性の地域差を分析するためには，より幅広い市民を対象にできる手法が必要である．本研究では，これらの課題を踏まえ，市民のソーシャルメディアの投稿を利用し，幅広い市民を対象に地域差の分析を試みる．

2.3 文章や発話からの著者の性格特性予測に関する研究

Mairesse ら [12] は，発話データをもとに性格特性を予測するモデルを作成した．発話者が自己報告した性格特性に基づいて訓練されたモデルと，他者評価に基づいて訓練されたモデルを比較した結果，後者のモデルがより良い性能であった．このことは，客観的な評価を用いることがより性格特性予測の精度を高めるために必要であることを示唆している．本研究では，この示唆が文章にも適用可能という前提のもと，ソーシャルメディアの投稿に人手でアノテーションを行うことで客観性を重視したデータセットを作成し，性格特性予測モデルの性能を高める．

大規模言語モデルを用いた性格特性予測に関する研究も行われている．El-Demerdash ら [2] は 2 つの英語のデータセットを利用して，事前学習済み BERT [9] をファインチューニングすることにより性格特性予測モデルを作成した．本研究では，日本語を対象にデータセットを作成し，事前学習済み大規模言語モデルをファインチューニングすることで性格特性予測モデルを作成する．

ソーシャルメディアの投稿を利用して著者の性格特性の推定を行った研究も存在する．那須川ら [19] は日本語の文章において性格と関連性の高い言語的特徴を調査し，その特徴を用いて Twitter(現 X) の投稿の著者の性格特性を予測するモデルを作成した．この研究で用いられたデータセットは，投稿の著者にアンケートを利用して収集されたものである．本研究では，客観性を重視するため，投稿に対して人手でアノテーションを行いデータセットを構築する点で異なる．

3 提 案 手 法

本提案手法は，まず X の投稿に対して人手により性格特性と経験性の 2 つの属性を付与することで，データセットを構築する．次にデータセットを用いて各属性の分類モデルを訓練する．続いて，訓練したモデルを用いて未知の投稿に対してラベルの自動付与を行い，都市間で市民特性の比較を行う．

そこで，3.1 節では投稿に対して付与する属性を定義し，3.2 節で各属性を分類するためのモデルについて説明する．

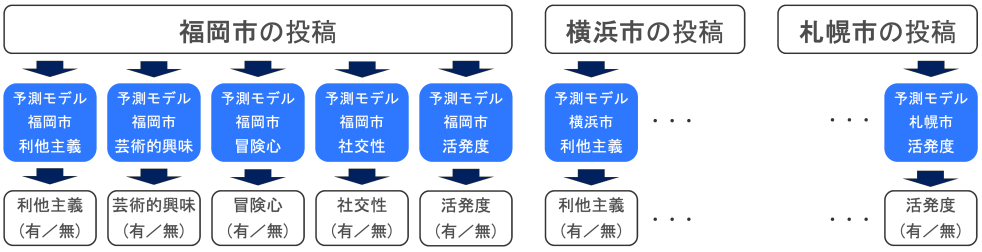
3.1 属性の定義

投稿に付与する属性は,「性格特性」と「経験性」の 2 つである．性格特性は，IPIP-NEO-120 に含まれる 30 個のファセットの中から，地域の特性や市民の活動が反映されると考えられる性格特性を選択した．以下に，各属性の定義，選定理由，アノテーションの際の判定基準を述べる．

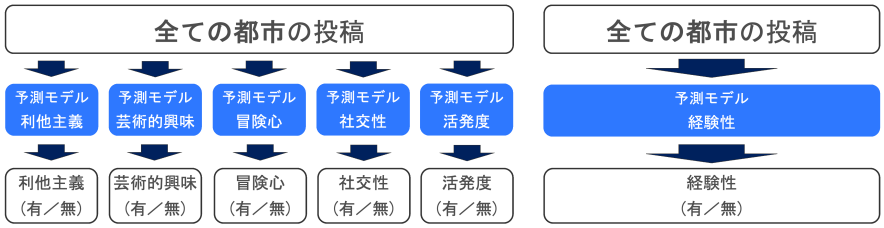
性格特性

表 1 に示す IPIP-NEO-120 [8] の各ファセットを測定するための質問項目を参考に，その性格特性を定義づける．

- 利他主義（Altruism）
  - 定義は「他人のためを思い，積極的に支援を行う傾向がある性格特性」とする．
  - ボランティア活動や寄付などの地域社会への貢献への関心度が反映されるため選定した．
  - 投稿著者がボランティアや寄付，ドナー登録など他人のためになることに興味がある，もしくは実行したことを述べた投稿は性格特性「有り」と判断する．
- 芸術的興味（Artistic Interests）
  - 定義は「芸術に深い興味を持つ性格特性」とする．
  - 芸術は地域の文化と密接に関わっており，地域の特性が反映されやすいため選定した．
  - 投稿著者が音楽，美術館，展覧会などの芸術に関連する物事に興味がある，もしくは参加したことを述べた投稿は性格特性「有り」と判断する．
- 冒険心（Adventurousness）
  - 定義は「新しい体験を望み，多様性と変化を好む傾向がある性格特性」とする．
  - 市民がどの程度変化や新しいアイデアを受け入れる傾向にあるかを理解する上で重要であるため選定した．
  - 投稿著者が新しい物事に興味がある，もしくは実行したことを述べた投稿は性格特性「有り」と判断する．
- 社交性（Gregariousness）
  - 定義は「人々と親しく交流し，社交的な活動に積極的に参加する傾向がある性格特性」とする．
  - 市民間のつながりの強さの面で地域の特性が反映されやすいため選定した．
  - 投稿著者が人との交流に興味がある，もしくは参加したことを述べた投稿は性格特性「有り」と判断する．なお，内容がネガティブであるものについては性格特性「無し」



(a) 都市ごとのデータで学習したモデルで性格特性を予測



(b) 都市を横断して学習したモデルで性格特性を予測　　(c) 経験性の予測

図 1: 各属性の分類モデル

と判断する．例えば,「同窓会に参加したけど，早く帰りたかった」などの投稿は性格特性「無し」と判断する．

- 活発度（Activity Level）
  - 定義は「忙しい状態を好み，体を動かすことを好む傾向がある性格特性」とする．
  - 活発度が高い地域では，市民が積極的にスポーツやレクリエーション活動に参加する傾向があり，地域の活力を評価するうえで重要であるため選定した．
  - 投稿著者が，忙しい状態であることを望む投稿，または，忙しい状態であり，かつそれを好んでいることを述べた投稿は性格特性「有り」と判断する．

経験性

経験性の定義は,「投稿内で筆者が言及しているイベントを実際に経験した／経験する予定であること」とする．例えば，性格特性「芸術的興味」の場合,「今日は美術館に行ってきた（イベントを実際に経験した）」や,「明日は友達と美術館に行く（イベントを経験する予定）」という投稿の経験性は「有り」となる．一方,「美術館に行きたい」などの実際に行っていない，もしくは具体的な予定が決まっていない投稿の経験性は「無し」となる．経験性を付与することで，市民の行動の傾向や，それらと性格特性との関連を分析することができると考える．なお，経験性は性格特性が「有り」の投稿にのみ付与する．

3.2 分類モデル

構築したデータセットを用いて事前学習済み言語モデル RoBERTa [11]，stockmark-13b [1] をファインチューニングすることで，各属性の予測モデルを作成する．使用する RoBERTa[1] は，CC-100 コーパスの日本語部分と，Wikipedia のオープンデータコーパスを用いて事前学習されている．CC-100 コーパスは Web 上のテキストを広い範囲にわたり収集しており，その中にはブログの投稿も含まれている．本研究では，ソーシャルメディアの投稿を対象にすることから，日常的な表現を理解できるモデルが必要である点を踏まえ，RoBERTa を採用した．また，本研究で使用する stockmark-13b[2]は，2023 年 9 月までの最新情報を含む約 2,200 億トークンの日本語データセットで訓練されており，日本語の言語理解能力において高い性能を持ったモデルである．本手法で作成する各属性の分類モデルの概要を図 1 に示す．まず，性格特性の予測について説明する．性格特性予測モデルは 2 種類作成する．1 つ目は，図 1a に示す都市ごとのデータを用いて学習するモデルである．都市ごとのデータを用いることで，地域特有の言い回しや固有表現をモデルに学習させることがねらいである．2 つ目は，図 1b に示す 3 つの都市を横断したデータを用いて学習するモデルである．都

表 1: ファセットを測定するための質問項目（和訳）[8]

性格特性	質問項目
Altruism（利他主義）	他人を気遣う 他人を助けるのが好き 他人の気持ちに無関心である 他人のために時間を割かない
Artistic Interests（芸術的興味）	芸術は重要だと思う 他人が気づかないようなものに美を見出す 詩は好きではない 美術館に行くのは好きではない
Adventurousness（冒険心）	習慣より多様性を好む 自分が知っていることに固執する 変化を嫌う 従来のやり方に執着する
Gregariousness（社交性）	大人数のパーティーが好き パーティーではいろいろな人と話す 一人でいることを好む 人ごみを避ける
Activity Level（活発度）	いつも忙しい いつも動き回っている 時間がある時はいろいろな事をする のんびりするのが好き

1：https://huggingface.co/ku-nlp/roberta-large-japanese-char-wwm

2：https://huggingface.co/stockmark/stockmark-13b

表 2: 各性格特性を表すクエリ

性格特性	クエリ
利他主義	ボランティア，寄付，寄附，寄贈，ドネーション，ファンドレイジング，クラウドファンディング，募金，支援，チャリティー，ドナー
芸術的興味	芸術，文芸，絵画，彫刻，音楽，演劇，美術，展覧会，アート，クラシック，ジャズ，ロッキン，フジロック，サマソニ，ライジングサン
冒険心	新しい，初めて，始め，挑戦，チャレンジ，てみた，たい
社交性	友人と，友達と，同期と，仲間と，パーティー，イベント，参加
活発度	遊ぶ，遊ば，遊び，遊べ，遊ぼ，遊ん，運動，忙し，多忙，忙殺，体を動か，体動か，外に出，外出

市を横断したデータを用いることで，モデルがより多様な表現を学習することが可能となる．次に，経験性の予測について説明する．経験性の表現は，性格特性に依存しないと考えられるため，経験性予測モデルは図 1c に示すように，都市を横断したデータを用いて学習を行う．

4 実験：各属性の予測

本章では，4.1 節でモデルの訓練に用いるデータセットの構築のためのデータの収集方法と，人手によるアノテーションについて述べる．4.2 節では構築したデータセットを用いた各属性の予測実験について述べる．

4.1 データセットの構築

4.1.1 対象都市と対象期間およびクエリ

本研究では，福岡市，横浜市，札幌市の 3 都市について，2019 年 1 月 1 日から 2019 年 12 月 31 日までの投稿から各性格特性を表すクエリを含む投稿を収集した．リツイートや重複する投稿は取り除いた，クエリの詳細は表 2 に示す．クエリは，表 1 に示す NEO-IPIP-120 の質問項目を参考に，性格特性が現れている投稿に含まれていると考えられる単語を選択した．なお，性格特性は地域で持続するが，災害やパンデミックなどによって多少の変動があるという研究結果 [19] が存在する．従って，2019 年の 12 月末からの新型コロナウイルスによる影響を抑えるため，対象期間として 2019 年を設定した．

4.1.2 投稿の収集

データの件数は，各都市，各性格特性につき 650 件，合計 9,750 件である．これらの投稿に対して，4.1.3 節で述べるアノテーション作業を行う．クエリによる抽出のみでは性格特性が現れる投稿の割合が小さかったため，効率的にアノテーション作業を行うために以下の手順でデータセットの投稿を収集した．

（1） 表 2 のクエリを用いて投稿を収集

（2） （1）で収集した投稿から，シード投稿として，第一著者が 3.1 節の判定基準に基づき性格特性が現れている多様な投稿を，各都市，各性格特性につき 80 件選択

（3） （1）で収集したシード投稿以外の投稿から，Sentence LUKE[3]を利用して，それぞれのシード投稿に類似したものを 7〜8 件ずつ抽出（80 件 →650 件）

Sentence LUKE は良質な文ベクトルを生成できる言語モデルで，文章の類似度を測ることに優れている．類似したモデルに Sentence BERT [14] が存在するが，BERT が人物名や組織などの固有表現を他のトークンと同等に扱って事前学習を行っているのに対し，LUKE [16] は固有表現を区別して事前学習を行っている．このことから LUKE は固有表現の理解において BERT よりも優れており，本研究で扱うソーシャルメディアの投稿のような固有表現が多く出現するテキストにおいて効果を発揮する．Sentence LUKE は Sentence BERT の BERT モジュールを LUKE の日本語版モデルに置き換えたモデルである．データセットの投稿を収集する前に，Sentence BERT と Sentence LUKE の定性的な比較実験を行ったところ，Sentence LUKE がより多くの類似した投稿を抽出できたため，本研究では Sentence LUKE を採用する．

4.1.3 人手による各属性のアノテーション

収集した投稿に対して，3.1 節で記載した判定基準に基づきアノテーション作業を行った [4]．アノテーション作業は第一著者を含む合計 5 名の実験参加者によって行い，全ての投稿について最低 3 人以上の参加者がアノテーションを担当した．アノテーション結果は多数決によって決定した．

アノテーション作業では，はじめに第 3 章の定義を含むアノテーション仕様書を配布した上で，各実験参加者の方針を一致させるための訓練を行った．訓練では，各都市，各性格特性につき 50 件の合計 750 件の投稿に対して実験参加者全員でアノテーション作業を行った．訓練終了後に，残りの全てのデータについてのアノテーション作業を行った．第一著者は全ての投稿に対してアノテーション作業を行い，残りの 4 名を 2 名ずつに分けることで各 3 名ずつのチームを作成し，それぞれのチームが半数ずつを担当することで作業を行った．「アノテーション結果が割れたデータについては，実験参加者間での話し合いによって最終的な値を決定した．意見が割れた例を紹介する．性格特性の有無において意見が割れたデータの例として，「バイクは地元の友達とダベって帰るのが一番好き。自分のバイクへの興味はそんなもの。」（社交性：性格特性「有り」，経験性「無し」）という投稿が挙げられる．投稿の主旨がバイクへの興味であるため，社交性の性格特性は「無し」と判断する場合が見られた．しかし，判定基準の「投稿著者は人との交流に興味がある」に該当すると判断し，最終的に性格特性は「有り」とした．Fleiss の κ 係数 [3] を用いた各チームのアノテーション一致度を表 3 に示す．全ての性格特性について，一致度が 0.6 (Substantial Agreement [10]) 以上となり，実験参加者間のアノテーションは一定の一致率を示した．

4.2 各属性の予測実験

4.2.1 目　　的

本実験では，4.1 節で構築した性格特性データセットを用い

3：https://huggingface.co/sonoisa/sentence-luke-japanese-base-lite
4：筑波大学図書館情報メディア系倫理審査通知（承認）番号 第 23-72

表 3: 各属性の Fleiss の κ 係数

性格特性	一致度（A チーム）	一致度（B チーム）
利他主義	0.679	0.699
芸術的興味	0.663	0.694
冒険心	0.721	0.701
社交性	0.681	0.680
活発度	0.674	0.688

て事前学習済み RoBERTa, stockmark-13b をファインチューニングし，各属性についての予測性能を評価する．

4.2.2 方　　法

予測モデルの訓練における最適なバッチサイズと学習率を決定するため，ハイパーパラメータチューニングを行った．本実験で調整するハイパーパラメータはバッチサイズと学習率である．バッチサイズを 8,16,32 とし，学習率を 1e-6,5e-6,1e-5 としてグリッドサーチを行った．評価は検証データによる 15 個の都市ごとの性格特性予測モデルと，1 つの経験性の予測モデルの合計 16 個のモデルの F 値の平均を用いた．実験の結果は表 4 に示すように，バッチサイズ 8, 学習率 1e-5 が最適なハイパーパラメータとなった．決定したハイパーパラメータで図 1 の性格特性予測モデルと経験性予測モデルの学習を行う．

表 4: ハイパーパラメータチューニングの結果

バッチサイズ	学習率	F 値の平均
8	1e-6	0.778
	5e-6	0.831
	1e-5	**0.871**
16	1e-6	0.784
	5e-6	0.842
	1e-5	0.848
32	1e-6	0.772
	5e-6	0.828
	1e-5	0.843

4.2.3 実験結果と考察

図 1a の都市ごとのデータで学習したモデルによるテストデータに対する性格特性の予測性能は表 5 のとおりである．いずれの性格特性においても，RoBERTa が最も良い性能を示した．RoBERTa はほとんどのモデルで F 値は 0.8 を超えており，高い精度で性格特性が予測できることが分かる．しかし，性格特性「活発度」の札幌市のモデルについては F 値が 0.728 と低くなっている．その原因を考察するためにモデルが予測を間違えた投稿を確認した．例えば,「最近忙しい」という投稿は正解ラベルは性格特性「無し」であるが，モデルは「有り」と予測した.「活発度」の定義は，3.1 節に記載したように,「忙しい状態を好む傾向がある性格特性」であり，単に忙しいことを述べた投稿は本来性格特性「無し」とラベル付けされるべきである．また,「水遊びしたいそうです」のような投稿は，文末の「そうです」という表現から主語は投稿著者ではなく第三者であると考えられる．従って，3.1 節の判定基準で述べたように投稿著者の性格特性には該当しないため性格特性は「無し」とラベル付けされるべきである．札幌市の訓練データにはこれらの違いをモデルが学習するのに十分な量が含まれていなかったことで，F 値が低い結果となったと考えられる．これらのような投稿に関しては，文章の極性や文末表現を考慮してデータ数を増やすことで改善が見込まれる．

図 1b の都市を横断して学習したモデルによるテストデータに対する性格特性の予測性能は表 6 のとおりである．全ての性格特性において RoBERTa が最も良い性能を示し，F 値は 0.8 を超えた．ほとんどの性格特性において，都市ごとに学習したモデルよりも，都市を横断して学習したモデルが良好な結果となった．都市を横断したデータには多様な投稿が見られ，網羅性が高まることでモデルの予測性能が向上したと考えられる．

図 1c の経験性予測モデルによるテストデータに対する予測性能は表 7 のとおりである．F 値は 0.864 となり，こちらも良好な結果となった．

表 5: 都市ごとのデータで学習したモデルの予測性能

モデル	性格特性	都市	適合率	再現率	F 値
RoBERTa	利他主義	福岡市	0.838	0.815	0.826
		横浜市	0.785	0.785	0.785
		札幌市	0.828	0.815	0.822
	芸術的興味	福岡市	0.805	0.815	0.810
		横浜市	0.885	0.892	0.888
		札幌市	0.919	0.923	0.921
	冒険心	福岡市	0.846	0.846	0.846
		横浜市	0.815	0.815	0.815
		札幌市	0.854	0.846	0.850
	社交性	福岡市	0.806	0.815	0.811
		横浜市	0.870	0.877	0.873
		札幌市	0.859	0.862	0.860
	活発度	福岡市	0.859	0.846	0.853
		横浜市	0.836	0.815	0.826
		札幌市	0.731	0.754	0.742
stockmark-13b	利他主義	福岡市	0.597	0.585	0.591
		横浜市	0.808	0.756	0.781
		札幌市	0.745	0.718	0.731
	芸術的興味	福岡市	0.400	0.500	0.444
		横浜市	0.793	0.606	0.687
		札幌市	0.423	0.500	0.458
	冒険心	福岡市	0.661	0.659	0.660
		横浜市	0.643	0.628	0.636
		札幌市	0.703	0.686	0.694
	社交性	福岡市	0.888	0.675	0.767
		横浜市	0.809	0.627	0.707
		札幌市	0.777	0.602	0.678
	活発度	福岡市	0.808	0.756	0.781
		横浜市	0.744	0.657	0.698
		札幌市	0.644	0.581	0.611

5　分析：都市間における市民の性格特性の比較

5.1 目　　的

最も良い性能を示した，RoBERTa をファインチューニングした性格特性予測モデルと stockmark-13b をファインチューニングした経験性予測モデルを用いて未知の投稿へのラベルを付与することで，地域差の分析を行い，提案手法の有効性を検証

表 6: 都市を横断して学習したモデルの予測性能

モデル	性格特性	適合率	再現率	F 値
RoBERTa	利他主義	0.850	0.867	0.859
	芸術的興味	0.890	0.954	0.921
	冒険心	0.827	0.869	0.847
	社交性	0.821	0.950	0.881
	活発度	0.845	0.893	0.868
stockmark-13b	利他主義	0.836	0.836	0.836
	芸術的興味	0.881	0.708	0.785
	冒険心	0.816	0.815	0.815
	社交性	0.870	0.830	0.849
	活発度	0.796	0.752	0.773

表 7: 経験性予測モデルの予測性能

モデル	適合率	再現率	F 値
RoBERTa	0.841	0.888	0.864
stockmark-13b	0.889	0.894	0.891

する.

5.2 方　　法

訓練データとして用いた 9,750 件以外の投稿を，表 2 のクエリを用いて各都市，各性格特性につき 4,000 件収集した．訓練データ収集時と同じく，リツイートや重複した投稿は除外している．4,000 件の投稿に対して都市を横断して学習した予測モデルを用いてラベルの付与を行った．それぞれの性格特性について，地域ごとに性格特性「有り」と予測された投稿の割合を算出し，都市ごとの傾向を分析した．また，性格特性が「有り」と予測された投稿にのみ経験性の予測を行い，同じく割合を算出し都市ごとの傾向を分析した．

5.3 結　　果

性格特性と経験性が「有り」と予測されたデータの割合を表 8 に示す．経験性「有り」の割合は，性格特性が「有り」と付与されたデータ内での割合を示している．なお，データの割合はクエリに対する相対的なものであるため，性格特性間の差異については意味をなさないことに注意されたい．

図 3 と図 4 に示すように,「芸術的興味」においては，性格特性は全ての都市間において母比率の Z 検定で有意な差（有意水準 5%）が見られ，札幌市が最も大きい割合となった．経験性は，福岡市と札幌市の間で有意な差が見られ，札幌市が有意に小さい割合となった．

「社交性」においては，性格特性は福岡市と札幌市，横浜市と札幌市の 2 つの組み合わせで有意な差が見られ，札幌市が 3 都市の中で最も大きい割合となった．経験性は福岡市と横浜市，札幌市と横浜市の 2 つの組み合わせで有意な差が見られ，横浜市が有意に大きい割合となった．

「活発度」においては，性格特性は福岡市と横浜市，札幌市と横浜市の 2 つの組み合わせで有意な差が見られ，福岡市と札幌市が同程度で，横浜市が有意に小さい割合となった．経験性はどの都市間でも有意な差は見られなかった．

「利他主義」と「冒険心」においては，性格特性と経験性ともに，3 つの都市間で有意な差は見られなかった．

表 8: ラベルが付与された投稿の割合

性格特性	都市	性格特性の割合	経験性の割合
利他主義	福岡市	0.144	0.534
	横浜市	0.151	0.493
	札幌市	0.141	0.530
芸術的興味	福岡市	0.608	0.392
	横浜市	0.647	0.378
	札幌市	0.711	0.357
冒険心	福岡市	0.152	0.363
	横浜市	0.146	0.419
	札幌市	0.148	0.370
社交性	福岡市	0.288	0.473
	横浜市	0.271	0.518
	札幌市	0.334	0.464
活発度	福岡市	0.424	0.303
	横浜市	0.345	0.329
	札幌市	0.437	0.301

5.4 考　　察

本節では，実験結果を踏まえた考察を述べる．

利他主義

「利他主義」では，性格特性と経験性ともに，ラベルの付与率に都市間で有意な差は見られなかった.「利他主義」に関連する，寄付やボランティアなどの行動を伴う投稿は 3 つの都市で同程度見受けられた．例えば地域特有の出来事に対する言及が含まれた投稿には，福岡市は「ボランティアコンサートの会場設営やってきた」や「買い物してそのまま小銭を寄付！」といった投稿が見られ，横浜市は「気軽に土曜日に 3 時間くらいボランティアできる施設ないかな」や「T ポイントで募金した」といった投稿が見られた．一方で，寄付や募金を伴う投稿には，京都アニメーション放火事件や熊本地震などの全国的な出来事に対する投稿も存在する．こちらも同様に全ての都市で同程度確認された．例えば，横浜市の投稿では「京アニの義援金に T ポイント全て寄付しました」という投稿が見られ，札幌市の投稿では「京アニに募金したい」や「熊本行ってきました．少しだけでもと寄付してきました」という投稿が見られた．性格特性「有り」と予測された投稿のクエリごとの件数を表 9 に示す．クエリごとでも都市間に大きな差異は見られない．従って,「利他主義」において 3 つの都市間に大きな差異は無いといえる．

表 9: 利他主義における主要なクエリごとの投稿数

都市	投稿数	クエリ		
		寄付，寄附	募金	ボランティア
福岡市	575	106(18.4%)	105(18.3%)	173(30.0%)
横浜市	605	118(19.5%)	104(17.2%)	188(31.1%)
札幌市	564	99(17.6%)	101(17.9%)	154(27.3%)

芸術的興味

「芸術的興味」では，性格特性「有り」と予測された割合が最も大きい都市は札幌市であった．札幌市は美術館の数が 3 都市の中で最も多く，そのことが性格特性に反映された可能性がある．しかし，経験性のラベル付与率を見ると，3 都市の中で最も割合が小さい結果となった．性格特性の割合が大きい一方で,

経験性の割合が小さいことは，芸術への興味はあるが，実際に芸術に触れる機会は少ないことを意味する．実際に，性格特性「有り」かつ経験性「無し」と予測されたのは,「大塚国際美術館行ってみたい」や,「今ラジオでやってる香りの美術館？博物館？どこだろう？すごい気になる！」などの，興味はあるが実際に経験することはないという内容の投稿であった．また,「自宅から芸術の森は遠いのでどうせだから野外美術館も行こうかってなったら、結構な登山だった」という投稿も見受けられ，北海道の地理的な特性も反映される結果となった．さらに，吉野らの先行研究 [17] によると北海道は開放性が他の地域よりも高く，本研究によって開放性のファセットである芸術的興味が，地域の開放性の高さに貢献していることが示唆された．

冒険心

表 8 に示すように,「冒険心」では，性格特性と経験性ともに，ラベルの付与率に都市間で有意差はなかった．正しく予測された投稿について，性格特性予測モデルにおける [CLS] トークンの各トークンに対するアテンションの可視化結果を図 2 に示す．

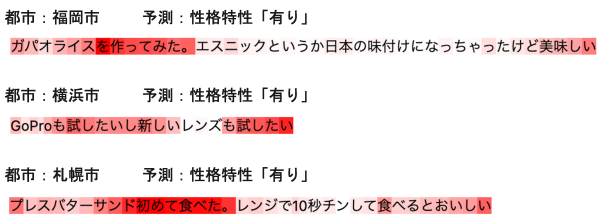


図 2: 冒険心における正解例のアテンションの可視化

モデルは「作ってみた」や「試したい」といった特徴的な語にアテンションできており，正しく予測ができていると判断できる．したがって，都市間に大きな差異は無いといえる．

社交性

「社交性」では，性格特性「有り」と予測された割合が最も大きい都市は札幌市であった．一方で，経験性においては札幌市は「有り」と予測された割合が最も小さい都市となった．この結果は,「芸術的興味」でも見られたように，札幌の地理的な特性を反映していると考えられる．実際に，性格特性「有り」かつ経験性「無し」と予測されたのは,「北海道から参加しに行くのって…遠い（涙）」や,「アニパ遊びに行きたかった」といった，社交性が現れているが実際に経験する予定はないことを表す投稿が多く見受けられた．

活発度

「活発度」では，性格特性「有り」と予測されたデータの割合は福岡市と札幌市が同程度で，横浜市だけが小さい結果となった．一方で，経験性「有り」と予測された割合が大きい都市は横浜市となった．経験性「有り」の投稿には,「本当に忙しいけど楽しい」や「忙しいのに、週末いろいろ詰め込み過ぎている」などが挙げられる．以上から，横浜市は活発度が高い人は多くはないが，一定数居住していることが分かる．

6 おわりに

本研究では，ソーシャルメディアの投稿に対して性格特性と経験性の 2 つの属性を付与して，市民の性格特性を比較する手法を提案した．なお，性格特性は Big Five Model の一部のファセットである「利他主義」,「芸術的興味」,「冒険心」,「社交性」,「活発度」を対象とした．未知の投稿に対してラベルの自動付与を行った結果,「芸術的興味」,「社交性」,「活発度」の 3 つの性格特性については，都市間で母比率の Z 検定 (有意水準 5%) で有意差が見られる結果となった．具体的には,「芸術的興味」,「社交性」においては札幌市が有意に高い結果となった.「活発度」においては福岡市と札幌市が横浜市に比べて有意に高い結果となった. また，本研究では Big Five Model の主要な 5 つの因子を測定するのではなく，ファセットに着目し分析を行ったことで，札幌市の開放性の高さに貢献する要因についても考察することができた．本研究の提案手法により，性格特性についての都市間の差を検出できることが明らかになり，ファセットに着目することで都市の特徴を把握できることを示した．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 B（課題番号 23H03686），挑戦的研究（萌芽）（課題番号 22K19822），学術研究助成基金助成金研究活動スタート支援（課題番号 22K21303）の助成を受けて遂行された．データセット構築の際のアノテーションについて，国立国語研究所の平本智弥さん，大阿久志緒理さん，立花幸子さんには多大な協力をいただきました．

文　献

[1] Stockmark Tech Blog. ビジネスのドメインや最新情報に対応した 130 億パラメータの日本語 LLM の公開, 2023. https://tech.stockmark.co.jp/blog/202310_stockmark_13b/.

[2] Kamal El-Demerdash, Reda A. El-Khoribi, Mahmoud A. Ismail Shoman, and Sherif Abdou. Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal*, Vol. 23, No. 1, pp. 47–53, 2022.

[3] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, p. 378, 1971.

[4] Adrian Furnham and Tomas Chamorro-Premuzic. Personality, Intelligence, and Art. *Personality and individual differences*, Vol. 36, No. 3, pp. 705–715, 2004.

[5] Lewis R Goldberg. A Broad-Bandwidth, Public Domain, Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models. *Personality psychology in Europe*, Vol. 7, No. 1, pp. 7–28, 1999.

[6] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. Personalized persuasion: Tailoring persuasive appeals to recipients ' personality traits. *Psychological science*, Vol. 23, No. 6, pp. 578–581, 2012.

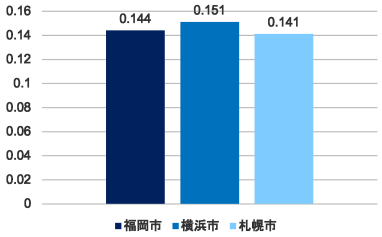
[7] O. P. John and S. Srivastava. *The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*, pp. 102–138. Guilford, New York, 2nd edition, 1999.

[8] John A Johnson. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of research in personality*, Vol. 51, pp. 78–89, 2014.

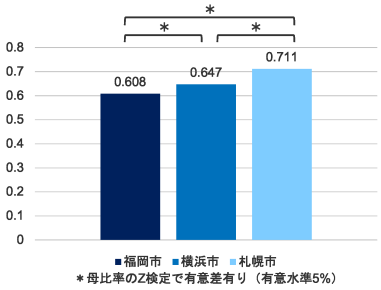
[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, Vol. 1, p. 2, 2019.

[10] J Richard Landis and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. *biometrics*, pp. 159–174, 1977.

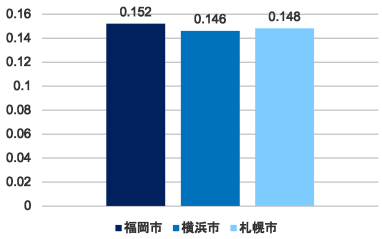
[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke



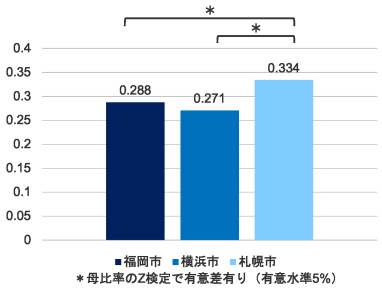
(a) 性格特性「有り」の割合（利他主義）



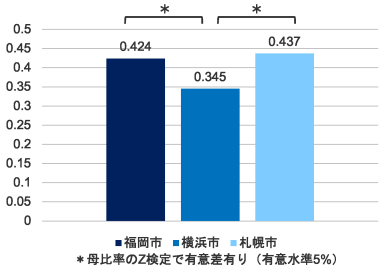
(b) 性格特性「有り」の割合（芸術的興味）



(c) 性格特性「有り」の割合（冒険心）

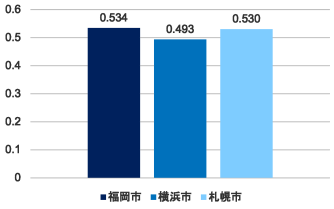


(d) 性格特性「有り」の割合（社交性）

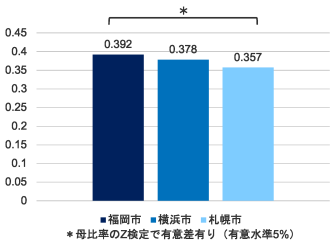


(e) 性格特性「有り」の割合（活発度）

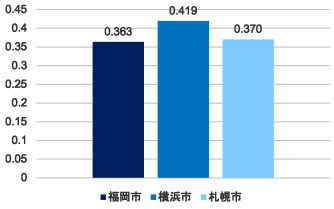
図 3: 性格特性「有り」と予測されたデータの割合



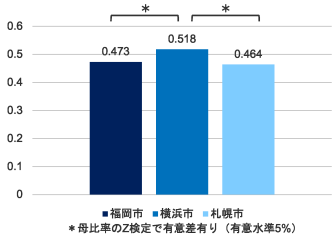
(a) 経験性「有り」の割合（利他主義）



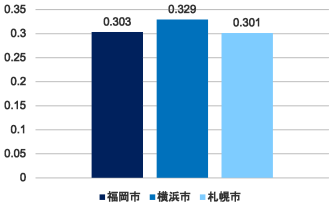
(b) 経験性「有り」の割合（芸術的興味）



(c) 経験性「有り」の割合（冒険心）



(d) 経験性「有り」の割合（社交性）



(e) 経験性「有り」の割合（活発度）

図 4: 経験性「有り」と予測されたデータの割合

Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Intell. Res. (JAIR)*, Vol. 30, pp. 457–500, 09 2007.

[13] Harris Poll. The 2023 State of Social Media, 2023. https://sproutsocial.com/insights/data/harris-insights-report-2023/.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on EMNLP*, pp. 3982–3992, Hong Kong, China, November 2019.

[15] David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations. *Journal of cross-cultural psychology*, Vol. 38, No. 2, pp. 173–212, 2007.

[16] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on EMNLP*, pp. 6442–6454, Online, November 2020.

[17] 吉野伸哉, 小塩真司. 日本における big five パーソナリティの地域差の検討 3 つの大規模調査のデータセットを用いて. 環境心理学研究, Vol. 9, No. 1, pp. 19–33, 2021.

[18] 小塩真司, 阿部晋吾. 日本語版 ten item personality inventory (tipi-j) 作成の試み. パーソナリティ研究, Vol. 21, No. 1, pp. 40–52, 2012.

[19] 那須川哲哉, 上條浩一. 日本語における筆者の性格推定の取組み. 言語処理学会第 23 回年次大会発表論文集, pp. 807–810, 2017.

DEIM Forum 2024

コロナ禍における草の観察

山田　寛太†　　横山　昌平††,†††,††††

† 東京都立大学 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所　〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †yamada-kanta@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　2020 年初頭に日本を震撼させた新型コロナウィルス感染症は，我々の生活を一変させた．これは SNS の世界も例外ではない．多くのデマや誇張が飛び交い，トイレットペーパーが無くなる等の話題がニュースを騒がせた．本研究ではコロナ禍が社会に与えた影響を，SNS でのインターネット・スラングの用法と，その使われ方を通して観察した．具体的には，笑いあるいは嘲笑の感情を示す「w」の文字に着目し，その使われ方を分析した．「w」は笑いの度合いに応じて「wwww」のように重ねて使う事が慣例となっている．これは草に見える事から，別名『草』とも呼ばれている．本論文では，コロナ禍の始まった 2020 年から，我々の生活が安定した 2023 年のツイートから，草の用法とその出現の推移を調査した．

キーワード　SNS，Twitter，ネットスラング，新型コロナウイルス

1 はじめに

2020 年に新型コロナウイルス感染症が流行し，我々の生活は一変した．友人家族との会話や Twitter のつぶやきの内容は新型コロナウイルスが大きな割合を占めることになり，暗い話題が増加した．しかし，暗い話題だけではなく，明るい話題も当然存在し，明るい話題のツイートには，笑っていることを示す記号として，「w」の記号が付けられていることが多い．この「w」を「wwww」のように連続させると草が生えているように見えることから「wwww」を草と呼ぶネットスラングが存在する．

「www」の由来は，「笑い」をキーボードで入力する際の「warai」の「w」が転じたものだと言われている．その起源は 2000 年代前半のインターネットと言われているが，高橋 [2] の調査によると，Yahoo!知恵袋で 2005 年に，Twitter で 2007 年に初めて「www」が使用されていた．Yahoo!知恵袋と Twitter の両方で「www」と同年に「草」の初出が確認されており，2000 年代から「www」を「草」と呼ぶ表現は存在していたが，使用していたのは一部のユーザのみであった．株式会社マイナビが行った 13 歳から 19 歳の女性を対象とした流行った言葉のランキングの調査では，2018 年に「草」が 4 位にランクインしていることから，「草」という表現は，2000 年代では知っている人はあまりおらず，有名ではないネットスラングであったが，2010 年代後半に一気に注目され，今では知らない人の方が少ないであろう有名な表現になったことが伺える．

図 1 は典型的な草の使わ方をしているツイートの一例である．w の数が笑いの度合いを示しており，砕けた口語体で記述されるツイートらしく，文の途中で笑いを入れている．これはブログ等での使われ方と異なり，臨場感重視であるツイートでの使われ方の特徴ともいえる．

どのようなときにどの程度の長さの草を生やすかなどの草の用法は様々であり，ユーザや話題によって異なっている．そこで，本研究では，Twitter のツイートを分析し，草の用法を調査する．また，新型コロナウイルス感染症が流行した 2020 年から現在にかけて草の用法の変化の有無や，その内容を調査する．

2 関連研究

2.1 「草」,「wwww」の誕生

内山 [1] は，ニコニコ動画のコメントや 2 ちゃんねるの書き込みを調査し，笑いを表す「w」を「wwwwww」のように一度に複数個列挙することを「草を生やす」と表現されていることを示した．また，「wwwwww」を使うユーザは荒らしや低俗な書き込みをすることが多かったため，「草を生やす」が，笑いの意味だけでなく，非難や嘲笑の意味を込めて使用されていることを示した．高橋 [2] は，Yahoo!知恵袋と Twitter を調査し，少なくとも，Yahoo!知恵袋では 2005 年， Twitter では 2007 年から「www」と「草」という表現が使用されていることを示した．

これら草の使われ始めた時期の研究に対して，本研究においては，デマや流言の多かったコロナ禍からその回復時期のツイートを用いて，草の用法を調べている．

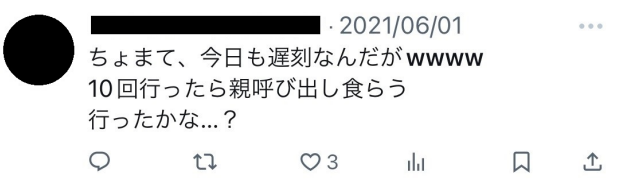
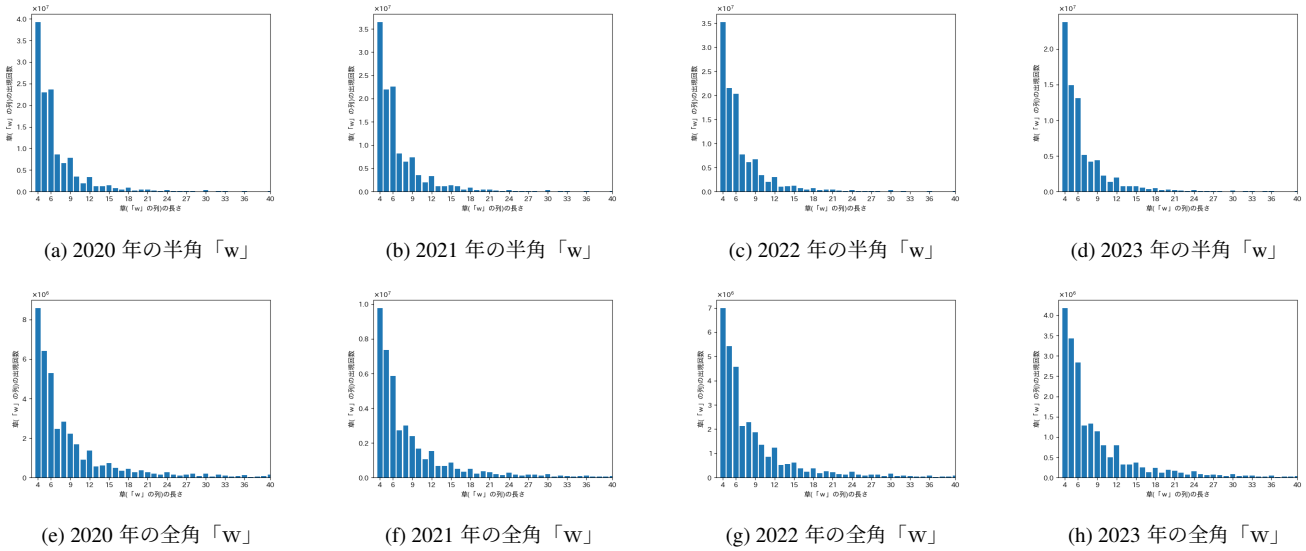


図 1:「wwww」を含むツイートの例



(a) 2020 年の半角「w」 (b) 2021 年の半角「w」 (c) 2022 年の半角「w」 (d) 2023 年の半角「w」

(e) 2020 年の全角「w」 (f) 2021 年の全角「w」 (g) 2022 年の全角「w」 (h) 2023 年の全角「w」

図 2: 「w」の列 (草) の長さと出現頻度

2.2 使われ方

松村 [3] は，笑いを表すネットスラング「笑」，「(笑)」，「w」，「ワロス」，「ワロタ」，「草」，「大草原」の， Twitter での使われ方を調査し，「w」は言い切り表現で既に文が完結しているところの直後に付く例が多いが，「草」は言いさし表現でまだ文が続いており，「草」を付けないと文が成り立たないところに付きやすいことを示した．

青柳ら [4] は，ニコニコ動画のコメントを分析し，「w (全角)」は単純な笑いを，「w(半角)」は嘲笑的な笑いに使用される傾向があることを示した．また，「草」は「草」という単語そのものがネガティブな単語とされ，「w(半角)」はネガティブな単語が含まれるコメントに出現する傾向があることを示した．青木 [5] らは，植物としての草ではなく，ネットスラングとしての「草」として使用されている「草」の検出を試みた．

岩崎ら [8] は， Twitter やブログにおいて，「w」の直前に来る言葉の品詞を調査したところ，終助詞，助動詞，動詞が多いとわかった．

これらの研究は，60 件から 60 万件と比較的小規模なデータセットに対する研究である．本研究では約 4 年分のデータを使っており，その総数は億を超える．

2.3 他の笑いを表す表現

Maity [7] らは， Twitter に現れる，語彙外の表現 (Out-of-Vocabrary) を 6 種類のカテゴリに分類する手法を提案した．

本研究では笑いを表す草のみに着目しているが，日本語においても，挫折を表す「orz」や，同意をあらわす「りょ」だの，流行は移り変わるがたくさんのインターネット・スラングが登場してきた．草以外の表現に関する分析は本稿の対象外で，今後の課題としたい．

表 1: ツイート数

データセット	ツイート数
「wwww」(半角 w4 つ以上の列)	417,609,004
「ｗｗｗｗ」(全角ｗ 4 つ以上の列)	122,834,312

3 データセット

3.1 使用データ

本研究では株式会社 NTT データ「なずきのおと」を用いて，「wwww」(半角 w4 つ以上の列) または「ｗｗｗｗ」(全角ｗ 4 つ以上の列) を含む，2020 年 2 月から 2023 年 12 月までの日本語の全てのツイートを分析した．

w のが三つ以下の場合は，URL によくある，www (world wide web の略) が引っかかってしまうため，本論文では草の定義を w の 4 回以上の連続と定義している．

分析に利用したツイート数を表 1 に示す．

3.2 前処理

本研究で用いるツイートのテキストに以下の前処理を行った．

- URL の除去
- 絵文字の除去
- 記号の除去
- アルファベットを小文字に統一

w の 4 回以上の連続を持つツイートであっても，そこに含まれる URL には WWW が含まれる可能性があるため，URL を除去している．その他，分析に不必要な絵文字や記号を前処理として削除した．また大文字は小文字に統一した．すなわち，本論文においては WWWW と wwww は同一である．

(a) 長さ 6

(b) 長さ 9

(c) 長さ 12

(d) 長さ 8

4 観　　察

4.1 草の長さと出現頻度

3.1 で分析したデータを 2020 年の半角「w」，2021 年の半角「w」，2022 年の半角「w」，2023 年の半角「w」，2020 年の全角「w」，2021 年の全角「w」，2022 年の全角「w」，2023 年の全角「w」の 8 つに分割し，それぞれのツイートデータに含まれる「w」の列 (草) の長さとそれの出現頻度を計算した．図 2 に結果を示す．

このグラフは横軸が w の連続数，縦軸がツイート数を表しており，グラフの概形はどれも冪乗則を示している．草の長さが大きければ大きいほどその出現頻度は低くなっており，典型的なロングテールの形を成している．驚くべきは w を 40 回程度連続させているツイートもある．

面白い傾向として，3 の倍数でスパイクが立っていることが見て取れる．これは半角で顕著であり，すべての年において同様の傾向が見て取れる．

このような特徴が現れる原因は不明だが，以下の 2 つの要因が考えられる．

- 端末のユーザ辞書
- 特定の表現が存在する

前者は，ユーザが URL を手動で入力した際，「https://www.hogehoge.com」の「www」がユーザ辞書に登録される．それが呼び出された結果 3 の倍数の長さの草が多くなる，という可能性がある．この説は，全角「w」では 3 の倍数の長さの草が多い特徴があまり見られない理由も同時に説明できる．最近はスマートフォン出せ毛でなく，PC においても入力候補文字列の推薦機能があり，Twitter という速報性の高いメディアの性質と相まって，候補に挙がる www を押下して入力している事が考えられる．

4.2 草の長さと出現単語

一方で Twitter ユーザ間で特定の単語と 3 の倍数の長さの草とを組み合わせた表現が流行し，複数ユーザが複数回その表現を使用した結果 3 の倍数の長さの草が多くなる，という可能性も考えられる．その仮説の検証のため，3 の倍数の長さの草と共起する 2 文字以上の単語をワードクラウドで表現した (図**??**).

前節の通り，3 の倍数の長さの草と共起する単語を調査する．調査の対象年は，3 の倍数の特徴がより顕著な 2021 年とし，長さが 6，9，12 の草 (「wwwwww」，「wwwwwwwww」，「wwwwwwwwwwww」) と共起する，2 文字以上の名詞を調査する．比較として，長さが 8 の草も同様に調査しワードクラウドとしている．

まず，4 つのワードクラウドを俯瞰する．ここで特徴的なのは「好き」と「無理」という，ポジティブな語とネガティブな語がそれぞれ大きな字として表れている．笑いの意味でも嘲笑の意味でも用いられているという青柳ら [4] の研究と同様の傾

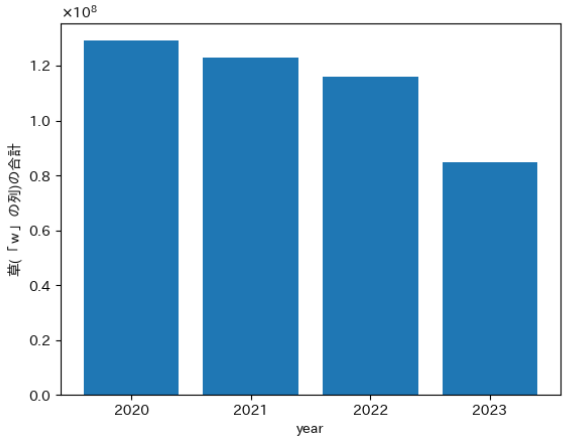


図 4: 半角草の出現回数の変遷

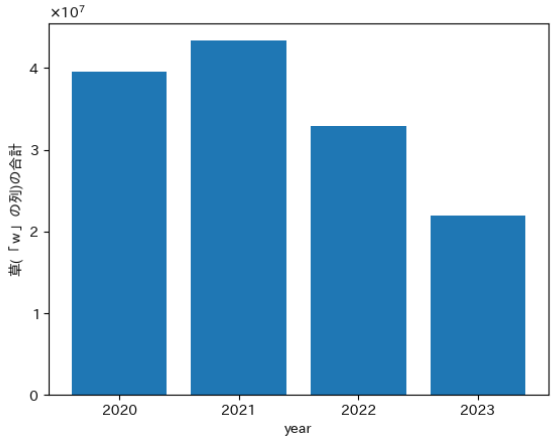


図 5: 全角草の出現回数の変遷

向が見て取れる．

次に，3 の倍数の 3 つのワードクラウドを比較する．ここでポジティブな意味である「好き」は草の長さに反比例して，出現頻度は低くなり，一方で「無理」という語は長さ 6 よりも，9 と 12 の方が出現頻度が顕著に多い．

このワードクラウドはあくまで，語の出現回数のみで描画しているため，これ以上の深い分析は行えないが，この結果から，ポジティブな語は草の長さは比較的短く表現され，ネガティブな語は草の長さが長く表現されるのではないかという仮説を立てる事ができる．

4.3 草の時系列変遷

最後に新型コロナウイルス感染症の流行初期から回復期である現在まで，草の出現回数の変遷を観察する．半角草の出現回数の時系列変遷を図 4，全角草の出現回数の時系列変遷を図 5 に記す．

両方の図から，全体としては草の減少傾向にある事がわかる．2023 年に顕著に減っているが，これは Twitter が x へと改名した混乱によって，月間アクティブユーザ数の減少がメディアでも指摘されて[1]いる．今回の分析では月間アクティブユーザ数による正規化は行っていない点には注意が必要であるが，コロナ禍が明け，野外での活動が盛んになるのと同時期の減少傾向は，何等かの相関関係がある可能性があり，今後はその点について分析を進める予定である．

5 ま と め

本研究ではコロナ禍における日本語ツイートのうち『w』が４回以上連続するすべてのツイート，約 5 億 4 千万件を観察し分析した結果．以下の知見を得た．

- 草の長さと出現頻度はべき法則に従う．
- 3 の倍数の長さを持つ草は全体のトレンドに比べて出現頻度が多い．
- 草の長さにかかわらず，ネガティブな語とポジティブな語の双方で草が使われている．
- 草の出現回数はコロナ化初期にくらべ現象している．

これらの知見から，以下の仮説が導き出せる．

- キー入力と頻出する 3 の倍数の長さの草に関係があるのではないか．
- ネガティブな語ほど草の長さが長くなり，ポジティブな語には比較的短い草が使われるのではないか．
- コロナ禍における行動制限が徐々に解除されているにしたがい，草の出現も少なくなっているが，この両者に何等かの相関があるのではないか．

研究の現時点では，またデータの観察にとどまっており，これらの仮説の検証は今後の課題である．また，今後は，笑い以外に別の感情を表す語も同様に観察・分析し，コロナ禍からの回復期にソーシャルネットワークに投影された，ユーザの感情の変遷を明らかにする．

1：https://gigazine.net/news/20231018-musk-x-twitter-down-traffic/

6 謝 辞

本研究は，内閣府 COVID-19 AI・シミュレーションプロジェクトの一環として実施され、JST，CREST，JPMJCR19A4 の支援を受けたものである．

文 献

[1] 内山弘，ネットの日本語 -2 ちゃんねるとニコニコ動画を中心に-，地域政策科学研究，2010.
[2] 高橋快征，インターネット上にある新語・新表現について，言語の研究，7 号，pp. 15–32，2020.
[3] 松村愛美，笑いを表すネットスラングの使われ方，日本大学文理学部国文学科卒業論文，2020.
[4] 青柳志織，川合康夫，ニコニコ動画のコメントにおける笑いを表現するネットスラングの分析，人工知能学会，2019.
[5] 青木竜哉，笹野遼平，高村大也，奥村学，ソーシャルメディアにおける単語の一般的ではない用法の検出，自然言語処理，2019.
[6] 吉川陽葵，Twitter における文末記号の使用状況について，京都大学國文學論叢，2021.
[7] Suman Kalyan Maity, WASSUP?LOL:Characterizing Out-of-Vocablary Words in Twitter，arXiv，2016.
[8] 岩崎真梨子，若者が着目するインターネット上の表現-ネットスラングと方言-，八戸工業大学紀要，2017.

降雨データを用いた SNS ユーザーの居住地推定

山下　暁天†　　田島　敬史††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †yamashita.akitaka.55a@st.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　SNS 上のあるユーザーの投稿のうち，周囲が雨であったことを示すものを抽出し，投稿時刻の降雨データを参照することで，ユーザーのおおよその居住地域を推定することを目的とする．一般に，時刻が複数与えられたとき，その全時刻において雨であるような地域は，データの数に応じて絞り込まれてゆくと考えられ，ユーザーの雨を表す投稿から居住地を推定することができる．投稿の収集にあたって，投稿内容が，実際にユーザーの周囲での降雨を示すものかどうかを判定する必要があり，これを自然言語処理モデルとニューラルネットワークを用いた機械学習により実装した．推定材料の時刻が 20 件以上あるユーザーでは，過半数のユーザーを 25 km 以内の精度で推定できた．
キーワード　ソーシャルメディア，居住地推定，機械学習，ニューラルネットワーク，自然言語処理，Sentence BERT，PU Learning

1　はじめに

SNS ユーザーの居住地域を推定することは，マーケティング戦略や災害時・緊急時の対応など，多岐にわたる分野で効果的な戦略や対策を展開するための基盤となる．例えば，マーケティング戦略においては，居住地域の推定は消費者行動の地域差異の理解を促進し，製品やサービスの地域別カスタマイズに繋がる．災害や緊急事態の際には，居住地域情報は迅速な救援活動や適切な情報伝達を可能にし，被災者への的確な支援を提供できる．

昨今の SNS 上では，ユーザーの身辺に関する多様な情報に容易にアクセスでき，天候情報もその例外ではない．ユーザーの近辺における天候情報は，実際の降雨データと組み合わせることで，ユーザーの居住地域を推定する上での手がかりとなる．一般に，複数の時刻が与えられたとき，その全ての時刻において雨が降っている地域は，時刻の数に応じて絞り込まれてゆく．

そこで本稿では，代表的なソーシャルメディアの一つである X（旧 Twitter）上のユーザーに対して，降雨データを利用して居住地域を推定することを試みる．具体的には，推定対象のユーザーの過去の投稿のうち，そのユーザーの近辺で雨が降っていることを示唆するような投稿を抽出し，それらの投稿の時刻を収集する．仮にこれらの時刻全てにおいて雨が降っていたような地域があれば，その地域にユーザーが居住している可能性が高いことになる．

実際の居住地推定においては，日本全国を 1 km 四方のメッシュで分割し，各地点について「雨を表す投稿をした時刻のうち雨が降っていた時刻の割合」（適合率）に加えて，「雨が降っていた時刻のうち雨を表す投稿をした時刻の割合」（再現率）も考慮し，人口密度のデータも併用して居住地を確率的に推定する．ユーザーに依存する情報には時刻の系列のみを用いて，ユーザーの居住地を確率的に推定することが本研究の新規性である．

ユーザーの投稿が雨が降っていることを示唆する投稿かどうかの判定には，機械学習モデルを用いる．入力には Sentence BERT [10] によって変換されたベクトルを用い，学習時のラベルには投稿時にそのユーザーの居住地で雨が降っていたかどうかを示す値を使用する．ただし，投稿内容が降雨を示すものでなくても，偶然ユーザーの地域で雨が降っていた場合が考えられるため，単純な教師あり学習を用いることはできない．そこで，居住地で雨が降っていなかったというラベルのみを信用し，雨が降っていたというラベルのつけられたデータはラベルなしとみなして，Positive-Unlabeled (PU) Learning [6] の手法を利用する．

評価においては，推定結果に対する独自の評価指標を導入し，その精度を定量的に評価する．本研究では，全体の 17.7% のユーザーを 50 km 以内の精度で推定でき，推定材料となる時刻数が 20 以上のユーザーのうち 53.4% のユーザーの居住地を 25 km 以内の精度で推定することに成功した．

2　関連研究

Web 上のコンテンツの位置推定に関する研究には，地名辞典などの知識ベースを利用した分析や，外部データを用いない内容ベースの分析，ソーシャルグラフを利用した分析などが挙げられる．知識ベースを利用した分析では，Web 上のコンテンツから地名や住所を抽出し，コンテンツに関連のある地理的位置を特定する手法などが提案されている [7]．Cheng ら [4] や Chandra ら [3] の研究では，語の各地域における分布に注目して，SNS の投稿の内容からユーザーの位置情報を推定する内容ベースの確率的フレームワークが提案されている．Backstrom らの研究では，SNS 上でのユーザー間の繋がりを利用して，位置が特定されているユーザーから他のユーザーの位置を推定するアルゴリズムが提案されている [1]．

また，SNS ユーザーをセンサーとみなすことで局所的なイベントを検出する試みも研究されている．Sakaki らの研究では，地震や台風に関連する投稿のうち位置情報が利用できるものを抽出することで，これらのイベントを検出する手法が提案されている [11]．本研究はこの逆の試みであり，過去の降雨という時間と場所が予めわかっている局所的イベントをもとに，ユーザーの位置情報を推定する手法を提案する．

これらの研究と比較して，本稿で提案する手法では，推定の信頼度を評価しやすいことが長所として挙げられる．本研究では，推定材料となる投稿が 20 件以上あるユーザーのうち 7 割以上のユーザーに対して居住地を 50 km 以内の精度で推定できることを確認している．そのようなユーザーの割合は全体の 5.61% と少ないが，これらのユーザーに対する推定の結果には一定の信頼をもつことができ，推定材料となる投稿が少ないユーザーに対しては別の推定手法を用いるなど，他の手法と組み合わせることが容易にできる．また，本研究では 5 年間の降雨データのみを推定に用いているが，より長期間のデータを用いることで推定材料となるデータを増やすことが可能である．一方，降雨を示す投稿が一定数必要となるため推定できるユーザーが限られる点や，外出の多いユーザーなどでは推定の精度が下がる点が短所として挙げられる．

3 基 礎 事 項

まず，本研究で使用する降雨データと人口データ，市区町村データについて説明する．降雨データの説明では，データに対して行う特定の処理とデータの特性についても併せて説明する．さらに，機械学習モデルの学習時に用いる PU Learning について，複数の学習方法を説明する．

3.1 降雨データ

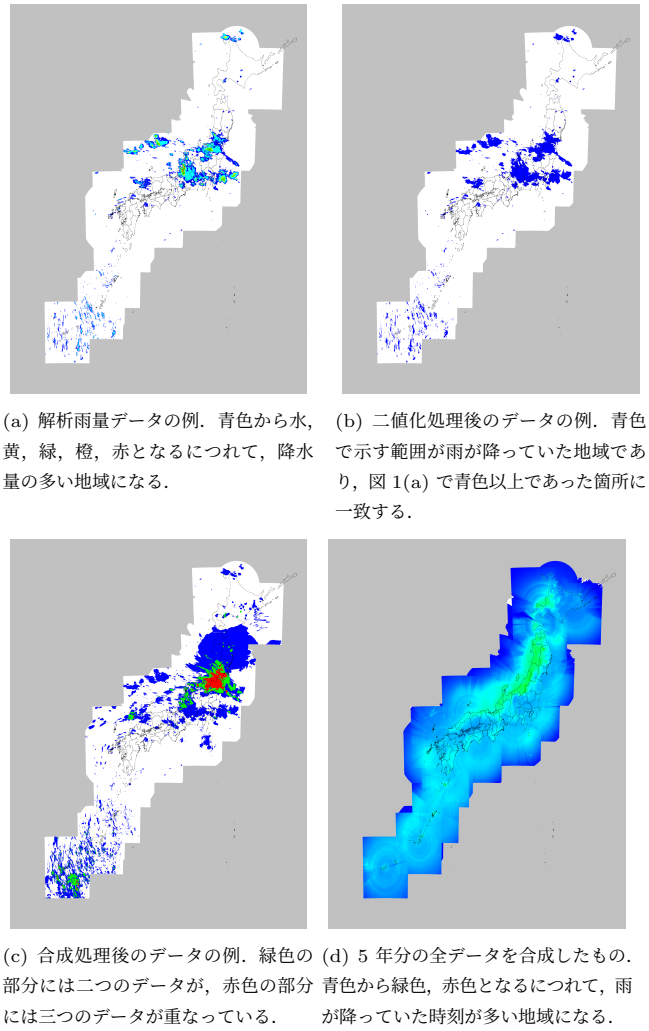
降雨データには，気象業務支援センターが提供する解析雨量データを用いる．解析雨量は，気象庁の保有する気象レーダーと地域気象観測（アメダス）データを用いてメッシュで解析した 1 時間降水量のデータである．本研究では，2018 年から 2022 年までの 5 年間のデータを用いる．これらのデータは，1 km 四方のメッシュで，30 分間隔で記録されている．ある時刻におけるデータを地図上に描画した画像の例を 図 1(a) に示す．

3.1.1 二 値 化

降雨データに対して，降水量の情報を二値化し，各地点で雨が降っていたかどうかの情報のみに圧縮する処理を実装する．この処理を二値化処理と呼ぶことにする．図 1(b) は，図 1(a) のデータに対して二値化処理を行った後のデータである．図 1(b) では各地点の降水量が記録されていたが，二値化処理後の図 1(a) では，降水量が 0 でない箇所の値を全て 1 に変換している．

3.1.2 合 成

二値化処理が施された複数の時刻のデータに対して，各地点の値を足し合わせたデータを作成する合成処理を実装する．図 1(c) は，図 1(b) を含む 3 件のデータを合成した後のデータである．



(a) 解析雨量データの例．青色から水，黄，緑，橙，赤となるにつれて，降水量の多い地域になる．

(b) 二値化処理後のデータの例．青色で示す範囲が雨が降っていた地域であり，図 1(a) で青色以上であった箇所に一致する．

(c) 合成処理後のデータの例．緑色の部分には二つのデータが，赤色の部分には三つのデータが重なっている．

(d) 5 年分の全データを合成したもの．青色から緑色，赤色となるにつれて，雨が降っていた時刻が多い地域になる．

図 1 デ ー タ 例

3.1.3 データの特性

5 年間の全降雨データ 87,648 件を二値化して合成したものを 図 1(d) に示す．これによると，日本海側と太平洋側で年間を通しての降水日数が異なり，北陸地方や東北地方の西側は特に降水日数が多い．図には複数の箇所で同心円状の模様が見られるが，これらの模様の中心には気象レーダーが位置しており，気象レーダーの観測特性によるものであると考えられる．さらに，拡大図の 図 2 では，静岡県北部にグランドクラッタ [8] と呼ばれる気象レーダーのノイズが見られ，滋賀県東部には周囲に比べて降水日数の多い矩形の領域が見られる．

確率的モデルを用いた居住地の推定では，全データのうちある地点で雨が降ったデータの割合を，その地点での降水の事前確率として使用しており，この確率を考えることで，地域による降水日数の差異や気象レーダーの特性の影響を排除または軽減することができる．確率的モデルによる推定について，詳しくは 4.3.1 節で説明する．

3.2 人口データ

推定時に用いる地点毎の人口の情報には，総務省統計局が提供する 2020 年の国勢調査の人口データを用いる．降雨データ

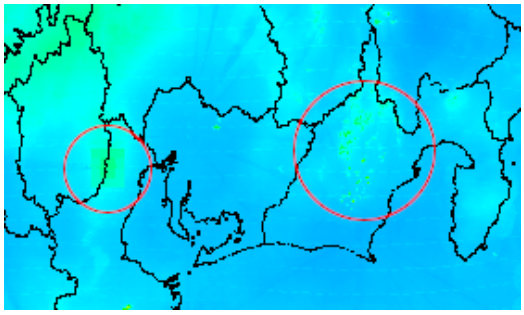


図 2　図 1(d) の一部を拡大したもの．静岡県北部や滋賀県東部に特殊な模様が見られる．

と同じく 1 km 四方のメッシュで，各地点の人口が取得可能である．なお，降雨データと人口データのメッシュの区切り位置は一致しないので，降雨データの各メッシュには，中心座標が最も近い人口データのメッシュを対応させている．

3.3　市区町村データ

ユーザーのプロフィールに記載されている居住地を抽出し，対応する緯度・経度に変換する処理には，国土交通省の提供する位置参照情報を利用する．当該データでは全国の各市区町村の大字・町丁目レベルの緯度経度情報を利用でき，各市区町村の座標には，その市区町村に属する大字・町丁目の緯度・経度の重心を用いている．

3.4　PU Learning

通常の二値分類問題では，正例または負例のラベルが全データに付与されていることを前提とするが，Positive-Unlabeled (PU) Learning [6] では正例の一部のみにだけラベルが割り当てられており，他のデータにはラベルが割り当てられていないようなデータを扱う．ラベルのないデータには正例と負例の両方が存在すると仮定し，正例と負例をうまく分類するモデルを学習する．

x をモデルの入力データとし，$y \in \{0,1\}$ を二値のラベルとする．また，x にラベルが割り当てられているとき $s=1$ とし，割り当てられていないとき $s=0$ とする．いま，負例のデータにはラベルがついていないので，

$$P(s=1|x,y=0)=0 \tag{1}$$

となる．さらに，ラベル付きのデータは正例データから均一の確率でランダムに選択される (Selected Completely At Random: SCAR) という仮定を導入すると，

$$c := P(s=1|x,y=1) = P(s=1|y=1) \tag{2}$$

となる．$c=P(s=1|y=1)$ とおくと，

$$\begin{aligned} P(s=1|x) &= P(y=1 \wedge s=1|x) && (3)\\ &= P(y=1|x)P(s=1|y=1,x) && (4)\\ &= P(y=1|x)P(s=1|y=1) && (5)\end{aligned}$$

より $P(y=1|x)=P(s=1|x)/c$ となる．よって，あるデータがラベル付きかどうかの確率 $P(s=1|x)$ を通常の分類器で学習することにより，$P(y=1|x)$ を推定できる．

また，ラベルの付いていないデータを重みづけする方法も用いられる．

$$\begin{aligned} P(y=1|x,s=0) &= \frac{P(s=0|x,y=1)P(y=1|x)}{P(s=0|x)} && (6)\\ &= \frac{(1-c)P(y=1|x)}{1-P(s=1|x)} && (7)\\ &= \frac{1-c}{c}\frac{P(s=1|x)}{1-P(s=1|x)} && (8)\end{aligned}$$

と変形できるので，

$$w(x) := P(y=1|x,s=0) = \frac{1-c}{c}\frac{P(s=1|x)}{1-P(s=1|x)} \tag{9}$$

をラベルの付いていないデータの重みとすることで，分類を学習できる．

前者の手法において $P(y=1|x)$ の二値分類の閾値を τ とすると，$P(s=1|x)$ を閾値 $\tau^{PU}=c\tau$ で識別する手法であると言い換えることができるが，代わりに Rebalancing Methods [5] を用いて各データに重みをつけることで $P(s=1|x)$ を c によらない一定の閾値で識別する方法もある [2]．二値分類の閾値を τ とした場合に，τ^{PU} とした場合と同様の結果を得るためには，各データを次のように重み付けする．

$$\begin{aligned} w^+ &= \tau(1-\tau^{PU}) & w^- &= (1-\tau)\tau^{PU} && (10)\\ &= \tau(1-c\tau) & &= (1-\tau)c\tau && (11)\end{aligned}$$

ここで，w^+, w^- はそれぞれ $s=1,0$ のデータに対応する重みである．両者の重みを τ で割っても学習結果は変わらないので，最終的な重みは

$$w^+ = 1-c\tau, \quad w^- = (1-\tau)c \tag{12}$$

となる．

4　提 案 手 法

居住地推定とその評価のためには，既に居住地が判明しているユーザーが必要である．まず，そのようなユーザーを収集し，過去の投稿のデータを取得する方法について説明する．次に，それらの投稿から雨が降っていることを表す投稿を抽出するための分類器の学習方法と，抽出した投稿から得られた投稿時刻の系列をもとに居住地を推定する方法を述べる．最後に，推定結果を定量的に評価するための指標について説明する．

4.1　データの収集

学習と評価に用いるユーザー・投稿データを収集する．取得したユーザーからランダムに抽出したユーザーの集合をテストユーザーとし，居住地推定の対象となるユーザーとする．残りのユーザーの投稿は再び訓練データ・テストデータに分け，分類器の学習と評価のためのデータセットとする．

4.1.1 ユーザーの収集

プロフィールに居住地の情報を市区町村レベルまで記載しているユーザーを取得する．X のユーザープロフィールには，居住地を記載するための location フィールドが存在し，これを地名にパースしたものをそのユーザーの居住地として利用することができる．ただし，location フィールドには任意の文字列が記述可能であるため，地名に変換する際に以下のルールを設ける．

- 末尾の「都府県」や「市」は原則省略可能とする．但し，「津市」など地名が 2 文字以下の場合は例外とする．
- 複数の地名にマッチする場合は，マッチした最後の地名を採用する．例えば,「京都 → 名古屋」の場合，後ろにある「名古屋」を採用する．
- 直後に「出身」とつく地名は除外する．例えば,「神戸出身」とだけ書かれたユーザーは採用しない．

4.1.2 投稿の収集

ユーザーの収集が完了したら，降雨データが存在する 2018 年から 2022 年までの 5 年間にそれらのユーザーが行った,「雨」または「雪」という文字が含まれる投稿を取得する．これらの投稿には，実際にユーザーの周辺で雨が降っていることを示唆するものだけでなく，他の地域での雨に言及するものや，人名などの固有名詞に偶然これらの文字が含まれただけの投稿も含まれる．

4.2 分 類 器

これらの投稿のうち，実際にユーザーの周りで雨が降っていることを表すものとそうでないものを分類する分類器を，ニューラルネットワークにより学習する．中間層の数は二つであり，中間層の活性化関数には ReLU 関数を用いている．入力は Sentence BERT により得られた 768 次元のベクトルであり，出力は 0 を閾値として二値化される．中間層は 2 層であり，中間層のニューロン数は 256 としている．

4.2.1 入　　力

入力には，投稿の本文を Sentence BERT [10] を用いて変換した 768 次元のベクトルを用いる．なお，前処理として，本文に含まれる URL と @ に続くユーザー名は，[URL] と [USER] というシンボルに置換している．

4.2.2 ラ ベ ル

さらに，その投稿を行った際，実際に雨が降っていたかどうかを表す二値のラベルを各データに付与する．これは，降雨データに加えて，投稿時刻と，投稿を行ったユーザーの居住地の情報により取得可能である．

ここで，実際には降雨データが 30 分間隔であり，さらに利用可能なユーザーの所在地は市区町村単位なので，ユーザーのいる地点の雨量を正しく取得できない可能性がある．そこで，R, θ をハイパーパラメータとして導入し，半径 R [km] 以内の地点における平均雨量が閾値 θ [mm] 以上かどうかで雨が降っているかどうかを判定している．実装にあたっては，$R = 20,\ \theta = 0.4$ としている．

4.2.3 学　　習

次に，取得したラベルをもとに教師あり学習を行うことを考えるが,「実際にユーザーの周りで雨が降っていることを表す」投稿でなくても，投稿時に偶然雨が降っていたため，誤ったラベルが付与されてしまったデータが存在する．このようなデータに対応するため，PU Learning を利用する．

4.2.4 PU Learning

いま，x を投稿データとし，$y \in \{0, 1\}$ を，実際の雨を示唆する投稿**でない**とき 1 となる変数とする．さらに，$s \in \{0, 1\}$ をラベルの値とし，投稿を行った際，実際に雨が降って**いない**ときに 1 となる変数とする．このとき，実際に雨が降っていないときの投稿は雨が降っていることを示すものではないので，

$$P(y = 1|x, s = 1) = 1 \tag{13}$$

が成り立つ．さらに，単純化のため SCAR の仮定を導入する．すなわち，以下を仮定する．

$$P(s = 1|x, y = 1) = P(s = 1|y = 1) = c \tag{14}$$

これは，実際の雨とは関係のない投稿であっても，一定確率 $1 - c$ で雨が降っているということを意味する．ここで，$1 - c$ はユーザーの居住地における降水確率とし，データが存在する 5 年間の全時刻のうち，ユーザーの居住地に雨が降っていた時刻の割合とする．

いま，c の値はユーザーごとに異なるため，テストデータの分類時にこの値を用いることはできない．そのため，3.4 節で最初に述べた，分類時に閾値を補正する手法は用いることができない．残り二つの手法のうち後者の手法は $P(s = 1|x)$ の値を事前に学習する必要がないので，学習コストの観点から，本研究では後者の Rebalancing Methods を使用する．二値分類の閾値 $\tau = 0.5$ とすると，各重みの値は次式で与えられる．

$$w^+ = 1 - c/2, \quad w^- = c/2 \tag{15}$$

4.3 推　　定

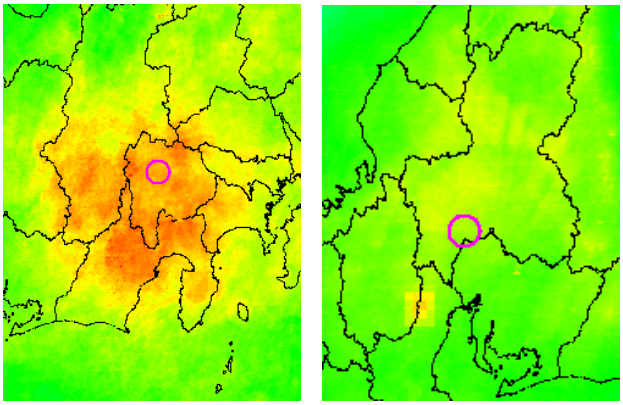
分類器の学習が完了したら，テストユーザーの「雨」または「雪」を含む投稿について，実際の雨との関連性を分類器を用いて判定する．実際に雨が降っていたと思われる投稿のセットが得られたら，各投稿時刻に対応する降水データをすべて合成する．

このようにして作成されたデータにより，ユーザーが雨であると発言した時刻に雨の降っていた割合が高い地域が得られる．図 3(a), 図 3(b) に，これらのデータの例を示す．これらの地域を推定結果としても一定の精度は得られるが，図 3(b) に見られるように，年間を通して降水日数の多い地域が上位になりやすいという問題がある．そこで以下では，いくつかの仮定をもとに確率的推論を行う．

4.3.1 確率的推論

はじめに，各記号を次のような事象として定義する．

- $I_{u,l}$：ユーザー u は地点 l にいる
- $R_{l,t}$：地点 l は時刻 t において雨が降っている



(a) このユーザーの居住地は山梨県甲斐市（赤丸）であり，赤い部分と概ね一致している．

(b) このユーザーの居住地は岐阜県岐阜市（赤丸）だが，図 2 でも見られた滋賀県東部の矩形領域の方が大きな値になっている．

図 3　投稿時刻に対応するデータを合成したデータの例

- $S_{u,t}$：ユーザー u は時刻 t において雨が降っていると言っている

このとき，$I_{u,l}$ と $R_{l,t}$ は独立とみなすことができ，ベイジアンネットワークは $I_{u,l} \to S_{u,t} \leftarrow R_{l,t}$ となる．また，各時刻 t において，$R_{l,t}$ は独立であると仮定する．いま，データ取得範囲の全時刻の集合を T とすると，あるユーザー u に対して

$$T^u_{p*} = \{t \in T | S_{u,t}\} \tag{16}$$
$$T^u_{n*} = T \setminus T^u_{p*} \tag{17}$$

が与えられている．T^u_{p*} は全時刻 T のうちユーザー u が雨を示す投稿をした時刻であり，T^u_{n*} は投稿しなかった時刻である．また，各地点 $l \in L$ に対して，

$$T^l_{*p} = \{t \in T | R_{l,t}\} \tag{18}$$
$$T^l_{*n} = T \setminus T^l_{*p} \tag{19}$$

が与えられている．T^l_{*p} は全時刻 T のうち地点 l が雨であった時刻であり，T^l_{*n} は雨でなかった時刻である．求めるのは，

$$\operatorname*{argmax}_l P(I_{u,l} | S_{u,T^u_{p*}}, \neg S_{u,T^u_{n*}}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \tag{20}$$

の値である．すなわち，各時刻についてユーザー u が雨を示す投稿を行ったかどうかと，各地点で雨が降っていたかどうかの情報が与えられたときに，ユーザー u が居住している確率が最も高い地点を求めたい．ただし，集合 $T' = \{t_1, t_2, \cdots, t_n\}$ について，$S_{u,T'}$ は $S_{u,t_1} \wedge S_{u,t_2} \wedge \cdots \wedge S_{u,t_n}$ の略記であり，$R_{l,T'}$ についても同様である．さらに，集合 $L = \{l_1, l_2, \cdots, l_m\}$ について，R_{L,T^L} は $R_{l_1,T^{l_1}} \wedge R_{l_2,T^{l_2}} \wedge \cdots \wedge R_{l_m,T^{l_m}}$ の略記である．ここで，

$$P(R_{l,t}) = r_l := \frac{|T^l_{*p}|}{|T|} \tag{21}$$
$$P(I_{u,l}) = \frac{n_l}{N} \tag{22}$$

とおく．ただし，n_l は地点 l の人口，N は全地点の人口の総和である．これは，地点 l で時刻 t に雨が降る確率は t に依らず，全時刻のうち地点 l が雨であった時刻の割合に等しく，ユーザー u が地点 l にいる確率はその地点の人口に比例するということを表している．さらに，

$$\mathrm{Pr}_u := P(R_{l,t} | I_{u,l}, S_{u,t}) \tag{23}$$
$$\mathrm{Re}_u := P(S_{u,t} | I_{u,l}, R_{l,t}) \tag{24}$$

とする．Pr_u は，ユーザー u が雨が降っていると言ったとき，ユーザー u のいる地点に雨が降っている確率（適合率）であり，Re_u は，ユーザー u のいる地点に雨が降っているとき，ユーザー u が雨が降っていると投稿する確率（再現率）である．これらの確率は t, l に依存しないと仮定する．このとき，

$$\begin{aligned}
&P(S_{u,t} | I_{u,l}, \neg R_{l,t}) \\
&= \frac{P(S_{u,t} | I_{u,l}) - P(S_{u,t}, R_{l,t} | I_{u,l})}{P(\neg R_{l,t} | I_{u,l})} \\
&= \frac{P(S_{u,t}, R_{l,t} | I_{u,l}) / P(R_{l,t} | I_{u,l}, S_{u,t}) - P(S_{u,t}, R_{l,t} | I_{u,l})}{P(\neg R_{l,t})} \\
&= \frac{P(S_{u,t}, R_{l,t} | I_{u,l})(\frac{1}{\mathrm{Pr}_u} - 1)}{1 - r_l} \\
&= \frac{P(S_{u,t} | I_{u,l}, R_{l,t}) P(R_{l,t} | I_{u,l})(\frac{1}{\mathrm{Pr}_u} - 1)}{1 - r_l} \\
&= \frac{r_l}{1 - r_l} \frac{1 - \mathrm{Pr}_u}{\mathrm{Pr}_u} \mathrm{Re}_u
\end{aligned} \tag{25}$$

となる．上記の値を $\alpha_{u,l}$ とおく．$i, j \in \{p, n\}$ に対して

$$T^{ul}_{ij} = T^u_{i*} \cap T^l_{*j} \tag{26}$$

と定義すると，

$$\begin{aligned}
&P(I_{u,l}, S_{u,T^u_{p*}}, \neg S_{u,T^u_{n*}}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \\
&= P(S_{u,T^u_{p*}}, \neg S_{u,T^u_{n*}} | I_{u,l}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \\
&\quad \cdot P(I_{u,l}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \\
&= P(S_{u,T^u_{p*}} | I_{u,l}, R_{l,T^l_{*p}}, \neg R_{l,T^l_{*n}}) \\
&\quad \cdot P(\neg S_{u,T^u_{n*}} | I_{u,l}, R_{l,T^l_{*p}}, \neg R_{l,T^l_{*n}}) \\
&\quad \cdot P(I_{u,l}) P(R_{L,T^L_{*p}}) P(\neg R_{L,T^L_{*n}}) \\
&= P(S_{u,T^{ul}_{pp}} | I_{u,l}, R_{l,T^{ul}_{pp}}) P(S_{u,T^{ul}_{pn}} | I_{u,l}, \neg R_{l,T^{ul}_{pn}}) \\
&\quad \cdot P(\neg S_{u,T^{ul}_{np}} | I_{u,l}, R_{l,T^{ul}_{np}}) P(\neg S_{u,T^{ul}_{nn}} | I_{u,l}, \neg R_{l,T^{ul}_{nn}}) \\
&\quad \cdot \frac{n_l}{N} P(R_{L,T^L_{*p}}) P(\neg R_{L,T^L_{*n}}) \\
&= \mathrm{Re}_u^{|T^{ul}_{pp}|} \alpha_{u,l}^{|T^{ul}_{pn}|} (1 - \mathrm{Re}_u)^{|T^{ul}_{np}|} (1 - \alpha_{u,l})^{|T^{ul}_{nn}|} \\
&\quad \cdot \frac{n_l}{N} P(R_{L,T^L_{*p}}) P(\neg R_{L,T^L_{*n}})
\end{aligned} \tag{27}$$

と求まる．よって，

$$\begin{aligned}
&P(I_{u,l} | S_{u,T^u_{p*}}, \neg S_{u,T^u_{n*}}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \\
&\propto_l P(I_{u,l}, S_{u,T^u_{p*}}, \neg S_{u,T^u_{n*}}, R_{L,T^L_{*p}}, \neg R_{L,T^L_{*n}}) \\
&\propto_l \mathrm{Re}_u^{|T^{ul}_{pp}|} \alpha_{u,l}^{|T^{ul}_{pn}|} (1 - \mathrm{Re}_u)^{|T^{ul}_{np}|} (1 - \alpha_{u,l})^{|T^{ul}_{nn}|} n_l
\end{aligned} \tag{28}$$

が従う．

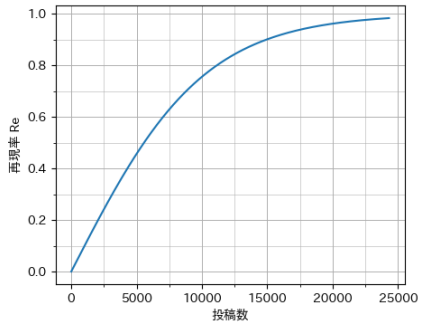
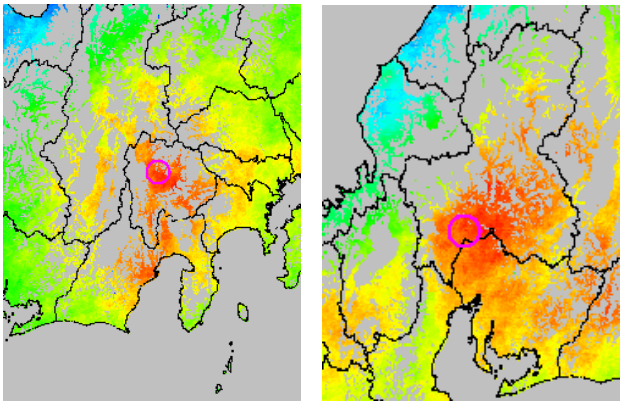


図 4　投稿数 $|T^u_{p*}|$ と Re_u の関係



(a) 図 3(a) を補正したデータ．元画像と比べて該当地域がより明確にハイライトされている．

(b) 図 3(b) を補正したデータ．滋賀県東部の矩形領域が他の地点と大差ない値になっている．

図 5　確率値による補正後のデータの例

以上の議論により，式 (28) の値が大きくなる地点 l を求めればよい．実際の計算の際には，数値的に安定させるために対数をとる．計算にあたって，$\mathrm{Pr}_u, \mathrm{Re}_u$ の値は次式を用いている．

$$\mathrm{Pr}_u = 0.9 \tag{29}$$

$$\mathrm{Re}_u = \frac{2}{1+\exp(-\mu|T^u_{p*}|)} - 1 \tag{30}$$

Re_u はシグモイド曲線の一部であり，雨と関連するユーザーの投稿の数 $|T^u_{p*}|$ との関係は 図 4 のようになる．これは，ユーザーの雨に関する投稿の数が多くなるほど再現率は 1 に近づき，投稿が少ないうちは投稿の数と再現率は比例するという仮定を反映したものである．μ は Re_u の増加度合いを調整するパラメータであり，

$$\mu = 1.974 \times 10^{-4} \tag{31}$$

を用いている．全 87,648 時刻のうち雨が降った時刻数を，人口が 0 でない地点において平均した値が 15,194 であり，この付近の投稿数で再現率が 0.9 程度になるよう調整している．

図 3(a), 図 3(b) のデータに対して，式 (28) の値をもとに補正を施した結果のデータを，それぞれ 図 5(a), 図 5(b) に示す．式 (28) は地点 l の人口 n_l を係数にもつため，人口が 0 である地点の確率値は 0 になっている．

4.4　評　　価

4.2 節で学習した分類器の評価方法として，次の二通りの手法を用いる．テストデータ，テストユーザーの定義は 4.1 節を参照されたい．

4.4.1　テストデータのラベルによる評価

テストデータに付与されたラベルを用いてスコアを算出することを考える．ただし，ラベルは投稿時に雨が降っていたかどうかに基づいて付与されている一方，モデルの正解は文章として雨が降っていることを表すかどうかであるため，ラベルとモデルの正解は完全には合致せず，単純に精度や F 値を計算することができない．そこで，4.2.4 節の記号を用いると，次の値を性能指標として利用することができる [2], [9].

$$\mathrm{F}_{\mathrm{PU}} = \frac{pr}{P(y=1)} \tag{32}$$

$$= \frac{pr^2}{rP(y=1)} \tag{33}$$

$$= \frac{P(y=1|\hat{y}=1)r^2}{P(\hat{y}=1, y=1)} \tag{34}$$

$$= \frac{r^2}{P(\hat{y}=1)} \tag{35}$$

ただし，$\hat{y}$ はモデルによる y の予測である．p, r はそれぞれ適合率，再現率であり，

$$p = P(y=1|\hat{y}=1) \tag{36}$$

$$r = P(\hat{y}=1|y=1) \tag{37}$$

である．F_{PU} は p と r がともに大きい値でないと大きくならないという点で通常の F 値と共通しており，精度の指標として用いることができる [2], [9]．いま，式 (14) の仮定より，$y=1$ のサンプルのうち，$s=1$ となるものは偏りなくランダムに選択されるとしているので，

$$r \approx P(\hat{y}=1|s=1) \tag{38}$$

とでき，F_{PU} の値が計算可能になる．

4.4.2　テストユーザーの居住地予測による評価

テストユーザーの投稿について，学習したモデルを用いて予測を行い，実際に居住地を推定してその精度を評価することを考える．モデルにより作成した降水時刻の系列について，4.3.1 節で示した手法により各地点の確率を計算し，次の手順でスコアを 100 点満点で計算する．

（1）　L を全地点の集合とする．

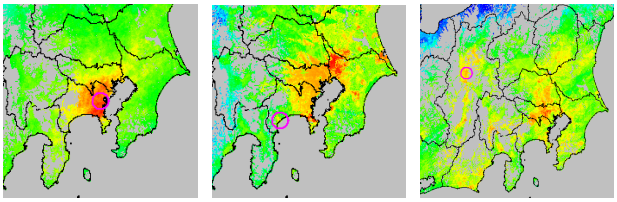
（2）　$k = 1, 2, \cdots, 5$ について，以下の 3, 4 の手順を繰り返す．

（3）　L から最も確率値の高い地点を取り出し，周囲 50 km 以内の地点を L から削除する．

（4）　ユーザーの実際の居住地とその地点の距離を d [km] として，$d \leqq 50$ ならば，次の値をスコアとして終了する．

$$\mathrm{score} = 100 - 20(k-1) - 0.4d \tag{39}$$

（5）　スコアが決定していなければ，スコアは 0 となる．



(a) スコア: 94.33 ($k=1, d=14.18$)　(b) スコア: 65.10 ($k=2, d=37.26$)　(c) スコア: 31.76 ($k=4, d=20.61$)

図 6　推定結果とスコアの例

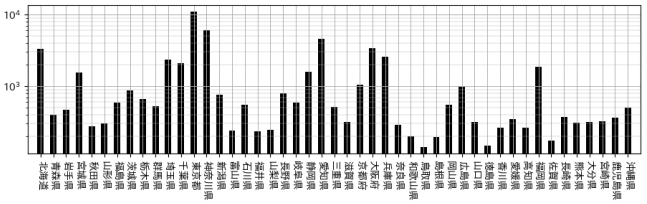


図 7　都道府県ごとの取得ユーザー数．縦軸は対数目盛．

すなわち，確率値の高い上位 5 地点について，地点同士が 50 km 以上離れるように取り出したうえで，居住地がある地点から 50 km 圏内であればその地点の順位に応じてスコアが与えられる．一般に雨雲は地点に対して連続的に分布するので，距離が近い二つの地点の確率値は近い値になることが多い．そのため，周辺の地点を除外せずに確率値の高い点を取り出す場合，近い地点がランキングの上位に複数現れる可能性が高い．よって，ランキング上位の周辺の地点を除外することで，距離の離れた複数の地域が推定結果の上位に現れている場合でも適切にスコアがつけられると考えられる．

推定結果の例とスコア，スコア算出時の k, d の値を 図 6 に示す．赤丸で示した地点がユーザーの実際の居住地である．

5　結果と考察

5.1　収集データ

ランダムに選択した 2,393,000 ユーザーに対して，4.1 節で述べた手法で居住地の情報が利用可能なユーザーを抽出したところ，55,749 ユーザーが取得でき，その割合は 2.33% であった．都道府県ごとのユーザー数を 図 7 に示す．さらに，これらのユーザーについて「雨」または「雪」の文字が含まれる 5 年分の投稿を，計 1,214,822 件取得した．投稿がちょうど n 件取得できたユーザーの数 a_n の分布は 図 8 のようになった．取得したユーザーのうち該当の投稿が 1 件以上存在したユーザーの数は 32,233 であり，このうち 20% の 6,446 ユーザーをテストユーザーとした．さらに，残りのユーザーの投稿 957,648 件のうち 765,487 件を訓練データ，192,161 件をテストデータとした．

5.2　確率値による推定

合成した降雨データからユーザーの居住地を推定する際に，図 3 に示したデータに基づく，ユーザーが降雨を示す投稿をした時刻における降水割合の高かった地点を単純に推定結果の

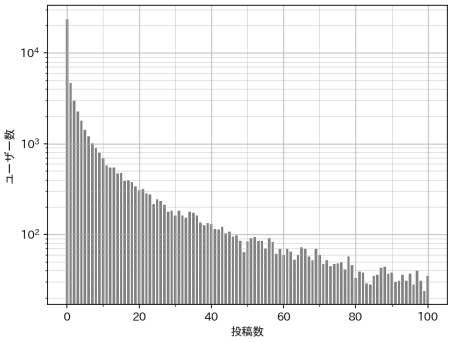


図 8　各投稿数 n に対するユーザー数 a_n．縦軸は対数目盛．

表 1　降水割合を用いた場合と確率値を用いた場合の平均スコア

v_c	v_p
16.532	42.751

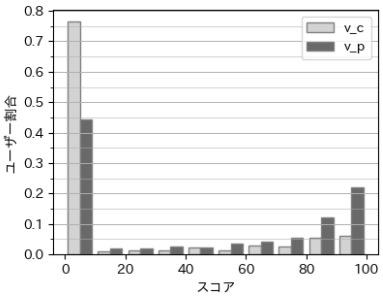


図 9　v_c, v_p の分布

上位とする方法と，図 5 に示したデータに基づく，4.3.1 節で議論した確率値が高い地点を上位とする方法について，その精度を比較する．モデルによる投稿の選別を行わず，テストユーザーの全投稿を用いた場合について，各推定結果を 4.4.2 節の方法で評価し，全テストユーザーにおけるスコアの平均値を求める．

ただし，降水割合に基づく推定では，値の高い地点を取り出す際，半径 10 km の範囲で平均化した値を用いている．これは，3.1.3 節で述べたグランドクラッタなど，降雨データにノイズとして高頻度で現れる地点の影響を軽減するためである．

降水割合を用いた場合のスコア v_c の平均値と，確率値を用いた場合のスコア v_p の平均値を 表 1 に示す．さらに，これらの値の分布をヒストグラムにしたものを 図 9 に示す．v_p の平均値が v_c の平均値よりも 26 ポイント高く，ヒストグラムではスコアが 10 未満のユーザーと 90 以上のユーザーの割合に大きく改善がみられる．以上より，確率値を用いた場合の方がより正確にユーザーの居住地を推定できているといえる．以下では v_p の値をスコアとして用いる．

5.3　モデルの性能

分類器による判定の結果，テストユーザーの全投稿のうち 22.2% が実際の降雨を示す投稿と判定された．また，テストユーザー 6,446 人中，取得した全ての投稿が分類器により実際の雨と無関係と判定されたユーザーは 1,903 人であり，その割

表 2　モデルの学習結果

F_{PU}	v_p
0.9846	53.097

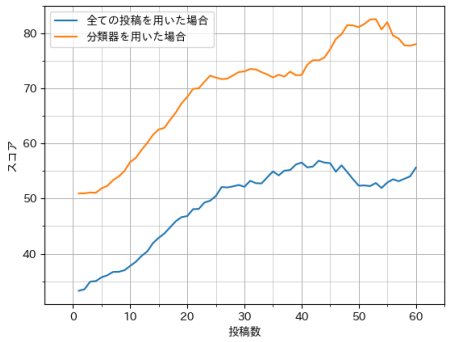


図 10　投稿数と平均スコアの推移

合は 29.5% であった．その他のテストユーザー 4,543 人における平均スコア v_p の値と，分類器の F_{PU} の値を 表 2 に示す．

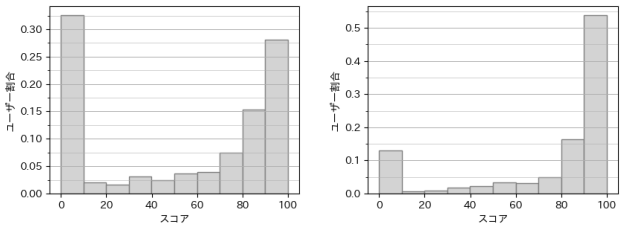
また，ユーザーの全投稿を用いた場合と，分類器による分類を用いた場合について，1 件から 60 件までの各投稿数のときに得られるスコアの平均を 図 10 に示す．なお，グラフ中の値は，幅 15 の移動平均により平滑化した値である．これらの結果によると,「雨」または「雪」が含まれる全ての投稿を推定に用いるよりも，分類器により選定された投稿を用いた方が 10 ポイント程度高いスコアを得られている．

ここで，降雨を示す投稿が 1 件以上存在したテストユーザー 4,543 人におけるスコアの分布をヒストグラムにしたものを 図 11(a) に示す．これらのユーザーにおいて，43.4% の 1,971 ユーザーがスコア 80 以上を達成している．5.1 節で述べた通り，居住地情報が利用可能なユーザー 55,749 人中，投稿が 1 件以上取得できたユーザーは 32,233 人であり，テストユーザーはそのうち 6,446 人である．ゆえに，そもそも取得した投稿が存在しないユーザーも含めたテストユーザーは $6446 \cdot \frac{55749}{32233} \approx 11149$ 人であると考えると，そのうちスコア 80 以上を達成したユーザーの割合は 17.7% である．

さらに，降雨を示す投稿件数が 20 以上のユーザー 626 人に対して，同様にスコアの分布を表現したものを 図 11(b) に示す．これらのユーザーでは，70.1% のユーザーがスコア 80 以上を達成しており，53.4% のユーザーがスコア 90 以上を達成している．いずれの場合でも，スコアが 10 未満のユーザーが一定数存在し，特に 図 11(b) では，スコアが 80 以上のユーザーと 10 未満のユーザーに二極化している．そこで以下では，居住地を推定することが困難なユーザーの特徴について考察する．

5.4　推定困難なユーザーの特徴

推定材料となる投稿が 20 件以上のユーザー 626 人のうち，スコアが 10 未満のユーザー 81 人について，スコアが低くなっている原因を調べたところ，以下のような要因が見受けられた．調査した 81 人のうち，各要因に当てはまったユーザーの数を 表 3 に示す．



(a) 推定材料となる投稿が 1 件以上のユーザーにおけるスコアの分布　(b) 推定材料となる投稿が 20 件以上のユーザーにおけるスコアの分布

図 11　スコアの分布

表 3　各要因の該当人数

要因	該当人数
居住地の誤判定	38
外出先での投稿	9
居住地の変化	9
定型文の送信	6
人口の影響	6

5.4.1　居住地の誤判定

ユーザーの居住地の記述が 4.1.1 節で説明した地名の検出ルールにそぐわない場合，居住地を誤って判定することがある．例えば，居住地の記述が「東京（⇔京都）」である場合，主要な居住地は東京であると考えられるが，京都が居住地であると判定されてしまう．

さらに，プロフィールに曖昧な居住地や虚偽の居住地を記述する場合もある．図 12 は，プロフィールに「南国」と記載しているユーザーの推定結果であるが，鹿児島県が推定結果に強く現れている.「南国」が鹿児島県のことを指すという解釈は比較的容易にできるが，地名検出の結果では高知県南国市となってしまい，そのためにスコアが低くなってしまっていた．さらには,「横浜に住みたい」と記述しているがゆえに居住地が横浜市と判定されているものもあった．この要因でスコアが低下しているのは 81 人中 38 人であった．

この原因によるスコアの低下は，居住地の判定アルゴリズムの改善により防止することが可能であり，居住地の推定器自体の精度を損ねる要因ではない．最も簡単な判定アルゴリズムの改善方法としては,「○○県○○市」のような形式で居住地を記述しているユーザーのみを採用する方法があるが，利用できるユーザーの割合が減少することで，学習やテストの精度が下がる可能性がある．実際に，テストユーザー 6,446 人のうち，プロフィールに記述する居住地が実際の市区町村名とスペースや改行を除いて一致するユーザーは 2,751 人であった．ただし,「都」「道」「府」「県」の各文字は省略可能とした．これらのユーザーのうち推定材料となる投稿が 20 件以上のユーザー 225 人に対してスコアの分布を表現したものを 図 13 に示す．図 11(b) に比べてスコアが増加していることがわかる．

5.4.2　外出先での投稿

趣味や仕事により外出の多いユーザーの場合，外出先で雨について言及する場合がある．投稿の本文で外出先であることが読み取れる内容であれば分類器に学習させることが可能だが，

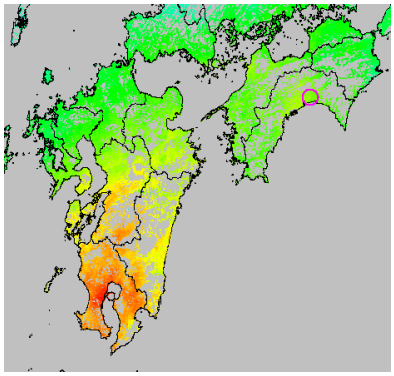


図 12 「南国」と記載するユーザーの推定結果

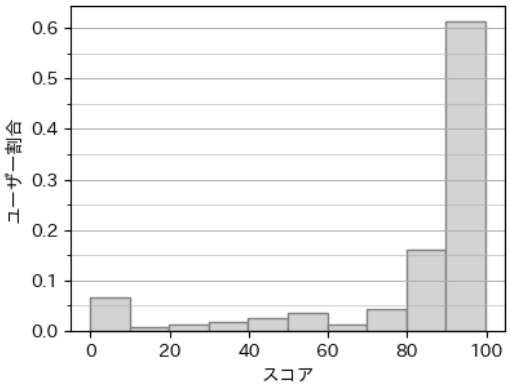


図 13 居住地がより明確なテストユーザーのうち，推定材料となる投稿が 20 件以上のユーザーにおけるスコアの分布

外出していない場合と差異のない投稿をする場合もあり，この場合は推定材料に含めざるを得ない．この要因でスコアが低下しているのは 81 人中 9 人であった．

前後の投稿を文脈として分類器の入力に含めることで外出先かどうかの判定精度を上げ，推定結果を改善することは可能と考えられるが，学習難易度が高く，完全性にも欠ける．

5.4.3 居住地の変化

進学や就職などに伴って居住地が変化しているユーザーの場合，過去の投稿が現在の居住地でされたものではない場合があり，誤った推定材料になってしまう．特に，本研究で使用したデータは 2022 年以前のものであり，2023 年以降に居住地が変化していた場合は全く感知することができない．図 14 は，プロフィールに「仙台 → 東京」と記載しているユーザーの推定結果であるが，以前の居住地である仙台が推定結果に強く現れている．この要因でスコアが低下しているのは 81 人中 9 人であった．

5.4.4 定型文の送信

ユーザーが使用している定型文や BOT などにより自動的に投稿される定型文が，分類器によって雨と関連があると判定されることがある．実際に雨が降っているときに投稿された定型文ならば推定に用いることができるが，BOT により定期的に投稿された場合や，分類器の誤判定によるものの場合，推定材料とすることができない．特に，分類器は表現の微細な違いで出力が変わる場合があり，通常であれば他の正常な判定により

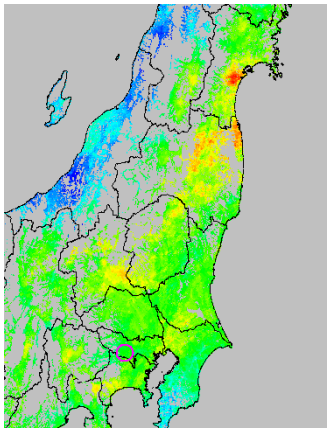


図 14 「仙台 → 東京」と記載するユーザーの推定結果

図 15 定型文を投稿するユーザーの例

無視できる誤判定であっても，定型文として大量に投稿があれば，誤判定の結果が支配的になってしまう．

図 15 は，定型文を投稿するユーザーの例である．定型文が分類器によって雨と関連があると判定された場合，実際に雨に言及している中央の投稿が定型文の投稿に埋もれてしまい，正しく推定することが難しくなる．この要因でスコアが低下しているのは 81 人中 6 人であった．

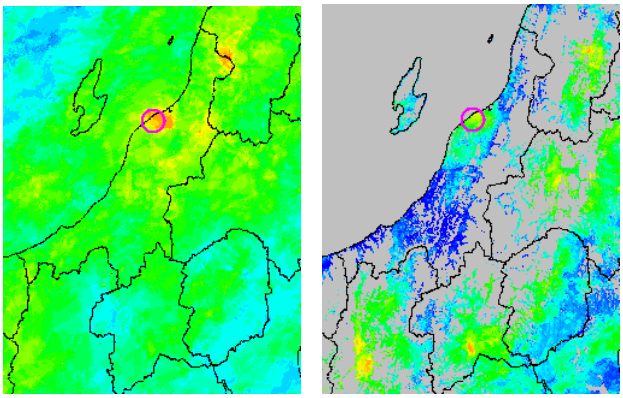
5.4.5 人口の影響

4.3.1 節で導出した 式 (28) の確率値は地点 l の人口 n_l に比例するため，人口の少ない地域では値が小さくなる．図 16(a)，図 16(b) は新潟市に住むあるテストユーザーの推定結果であり，投稿時刻における単純な降水割合による推定（図 16(a)）では該当地域周辺が大きい値になっているが，確率値による推定（図 16(b)）では，人口密度が少なく年間の降水日数が多い影響で比較的小さい値になってしまっている．この要因でスコアが低下しているのは 81 人中 6 人であった．

6 おわりに

本研究では，X 上のユーザーの降雨を表す投稿を降雨データと組み合わせることで居住地を推定し，推定結果の評価指標を導入した上で精度を評価し，推定における課題について分析した．

ユーザーの投稿が実際の降雨を表すものかどうかを判定する



(a) 降水割合による推定結果　(b) 確率値による推定結果

図 16　推定結果の比較

においては，機械学習を用いて分類器を学習し利用した．学習において，完全なラベルデータを作成できないという特性を考慮し，PU Learning を用いてデータに重みをつけることで対応した．また，推定部分では，単に投稿時の降水割合が高い地点を推定結果とするのではなく，確率的な考察のもと，最も尤度が高くなる地点を推定結果とすることで，推定結果の平均スコアを有意に増加させることに成功した．

テストユーザーにおける居住地推定の結果，全体の 17.7% のユーザーにおいて，居住地は推定地点から 50 km 以内に位置していた．さらに，推定材料となる投稿が 20 件以上ある場合は，53.4% のユーザーの居住地を 25 km 以内の精度で推定することができた．また，スコアの小さいユーザーのデータを調査することで，推定を苦手とするユーザーの複数の特徴について分析した．

謝　　辞

本研究は JSPS 科研費 21H03446 の助成を受けたものです．

文　　献

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 61–70, New York, NY, USA, 2010. Association for Computing Machinery.

[2] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, Vol. 109, No. 4, pp. 719–760, April 2020.

[3] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions–a content based approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 838–843, 2011.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, p. 759–768, New York, NY, USA, 2010. Association for Computing Machinery.

[5] Charles Elkan. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, Vol. 1, , 05 2001.

[6] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 213–220, New York, NY, USA, 2008. Association for Computing Machinery.

[7] Clayton Fink, Christine D Piatko, James Mayfield, Tim Finin, Justin Martineau, et al. Geolocating blogs from their textual content. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pp. 25–26, 2009.

[8] Mohammad-Hossein Golbon-Haghighi and Guifu Zhang. Detection of ground clutter for dual-polarization weather radar using a novel 3d discriminant function. *Journal of Atmospheric and Oceanic Technology*, Vol. 36, No. 7, pp. 1285 – 1296, 2019.

[9] Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. Vol. 20, pp. 448–455, 01 2003.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[11] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 851–860, New York, NY, USA, 2010. Association for Computing Machinery.

DEIM Forum 2024

ソーシャルメディアテキストからのユーザ行動の分類

中岡　明義†　　若宮　翔子†　　荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
E-mail: †{nakaoka.akiyoshi.nc0,wakamiya,aramaki}@is.naist.jp

あらまし　ソーシャルメディア上のテキストデータは豊富な情報源として認識されており，感情分析や行動分析などの多岐にわたる研究が行われている．しかしながら，既存の行動分析研究では特定の行動に限定した分析が行われており，日常生活における行動を用いた分析は十分に行われていない．そこで，本研究では総務省統計局の社会生活基本調査で用いられている 20 種類の行動を基に日常生活における行動を定義し，その行動をソーシャルメディアテキストに付与してコーパスを構築した．その後，コーパスを用いて行動分類モデルを構築し，モデルの分類性能を用いてコーパスの評価を行った．さらにコーパスの不均衡データ問題に対処するために，ChatGPT4 を用いた少数ラベルのデータ拡張を行い，コーパスの質の向上を図った．実験結果から，拡張後のコーパスで学習した行動分類モデルの方が拡張前で学習したモデルより高い性能を示した．最後に，構築した行動分類モデルを用いたユーザ行動に関するケーススタディを通じて，その有効性を示した．

キーワード　行動分析，データセット構築，テキスト分類，X（旧 Twitter），ソーシャルメディア

1　はじめに

ソーシャルメディアは，現代社会におけるコミュニケーションの主要な手段となっている．そのため，プラットフォーム上で生成されるテキストデータは人々の意見，感情，行動傾向を理解するための重要な情報源となっている．特に，X（旧 Twitter）では，日常の出来事から政治的な議論に至るまで多様なテーマに関するリアルタイムデータが提供されている．そのため，Twitter 上のデータは感情分析，トレンド分析，公共の意見の把握などの様々な研究に利用されている．しかしながら，これまでのソーシャルメディアを用いた行動に関する研究では，特定の行動や状況に焦点を当てたものが多い．そのため，日常生活における行動を用いた分析は十分に行われていない．

本研究では，総務省統計局の社会生活基本調査で用いられている 20 種類の行動を基に日常生活における行動を定義し，ソーシャルメディアテキストに付与してコーパス（行動データセット）を構築した．その後，構築した行動データセットを用いて行動分類モデルを構築し，モデルの分類性能を用いて行動データセットの評価を行った．図 1 に本研究で構築した行動分類モデルの入出力を示す．行動分類モデルは，ソーシャルメディアテキストに含まれるユーザ行動を全て抽出する．この行動分類モデルはソーシャルメディアテキストから，日常生活におけるユーザ行動を捉えることを目的としている．行動分類モデルは，文脈の理解が可能な自然言語処理モデルの一つである Bidirectional Encoder Representations from Transformers (BERT) [1] を基に構築した．これにより，ソーシャルメディア上のテキスト内に直接的な表現として現れない行動に対しても分類することが期待できる．

本研究で構築した行動データセットは，さまざまな応用研究に利用可能であり，その動機は行動データセットを用いて構築

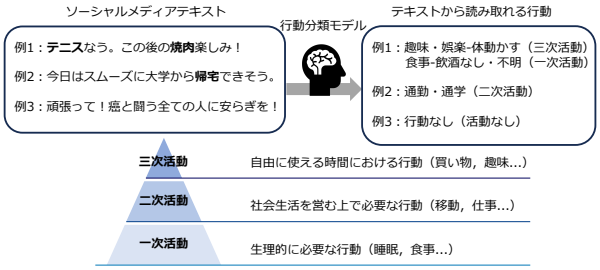


図 1　ソーシャルメディアテキストからのユーザ行動抽出

される行動分類モデルの汎用性にある．行動分類モデルを使用して，個々のユーザやユーザ集団に焦点を当てた分析を行うことで，病気の人々の生活習慣の把握が可能となる．さらに，位置情報付き投稿テキストの分析を通じて，地域特性を把握することもできる．時間軸に着目した分析を行うことで，フェイクニュースのようなイベントが実世界に与える影響を定量的に評価できる．このように，本研究は，生活習慣の分析，地域特性の把握，時間変化に関する影響評価といった複数の応用研究への貢献が期待される．

本論文の主な貢献は以下の 2 点である：

（1）　Twitter（現 X）の投稿データを用いた行動データセットの構築．このデータセットは，従来の研究ではカバーされていない日常生活におけるユーザ行動のラベルが付与されている．さらに，モデル性能向上を目的とした ChatGPT4 によるデータセットの拡張も行った．

（2）　ケーススタディを通じた行動分類モデルの応用とその分析．これにより，肝機能障害を有する可能性が高い人々と肝機能に問題がないと思われる人々の行動の差異を示唆した．

2 関連研究

ソーシャルメディア上のテキストデータは豊富な情報源として認識されており，様々な研究に利用されている．例えば，テキストの内容が肯定的か否定的かを分類するセンチメント分析というタスクがある．Yuxing らは，イギリスの Covid-19 に対する感情を探るためにツイートのセンチメント分析を行った [2]．Random Forest 分類器，Multinomial Naïve Bayes 分類器，SVC などの様々な分類器を用いてセンチメントを分類し，分析を行い，肯定的，否定的な感情の割合の推移を示した．Covid-19 の確定症例数とワクチン接種状況を分析することで，確定症例の増加とワクチン接種量の減少が否定的感情の増加の原因である可能性も示した．他には，ユーザの公開情報から年齢，性別，収入を推定するタスクもある．Sap らは Facebook，ブログ，Twitter のデータとそれに関連する人口統計ラベルから回帰および分類モデルを用いて，年齢と性別の推定に関する語彙を作成した [3]．結果として Facebook と Twitter 上での年齢，性別推定において良い精度を出せることを示した．Matz らは Facebook 上のデジタル足跡から個人収入の推定を行った [4]．結果として，Facebook の「いいね！」やステータス更新頻度だけで相関係数が 0.43 の性能で個人収入を推定できることを示した．さらに Facebook のプロフィールから収入の推定も行い，高い推定性能を示した．

テキストから行動を抽出するタスクも行われている．Banerjee らは，英語のツイートに対して形態素解析を行い，動作とカテゴリ，時間を表す単語の共起頻度に基づいて行動の抽出を行った [5]．Nguyen らは，日本語のツイートに形態素解析を行い，文章構造に着目することで，動作主，動作，対象，時間，位置といった動作の属性の抽出を行った [6]．これらの研究は，形態素解析を基に行われているため，Twitter 特有の表現や文脈を考慮することができない．Yano らは n-gram とナイーブベイズ分類器を用いて、行動が含まれているツイートかどうかの分類を行った [7]．結果として，従来手法よりも Precision と Recall において高い性能を示した．

以上のように，投稿テキスト内の行動の存在を判断する研究や，形態素解析などの文脈を考慮しない手法を用いての行動抽出研究は行われている．しかしながら，文脈を考慮して，ソーシャルメディアテキストから日常生活におけるユーザ行動を抽出した研究はない．本研究はソーシャルメディア上での日常生活におけるユーザ行動に着目し分析を行うという点において以上の研究と異なる．

3 データセット

本章では行動分類モデルを構築するためのデータセットについて説明する． テキストデータの収集方法，行動の種類，アノテーションガイドラインを説明した後，構築したデータセットの統計情報について説明する．

表 1 活動カテゴリとその行動

活動カテゴリ	行動
一次活動 (I)	睡眠，身の回りの用事，食事-飲酒あり，食事-飲酒なし・不明 （計 4 行動）
二次活動 (II)	通勤・通学，その他の移動，仕事，学業，家事，介護・看護，育児 （計 7 行動）
三次活動 (III)	メディア（テレビ・ラジオ・新聞・雑誌・ネット），買い物，趣味・娯楽-体動かす，趣味・娯楽-動かさない，交際-物理的接触，交際-オンライン的接触，受診・療養，喫煙，その他の行動 (事故など)（計 9 行動）
活動なし（IV）	行動なし

3.1 使用データ

本研究で対象としたデータは，2013 年の 10ヶ月間にわたって収集した 8 万人のユーザによる Twitter（現 X）の投稿テキストである．この中から，10 文字以上の投稿をランダムに 1100 件抽出した．前処理としてユーザ名を示す@記号に続く文字列は [USR] に，http から始まる文字列は [URL] に置き換えた．これにより，データの一貫性を確保した．

3.2 行動の種類の定義

本研究で用いる行動の種類は，総務省統計局が定める「一次活動」,「二次活動」,「三次活動」の枠組みに基づく [8]．**一次活動**は生理的に必要な活動であり，睡眠や食事などの生命維持に直接関連する行動が含まれる．**二次活動**は社会生活を営む上で必要な活動であり，仕事，学業，家事などの日々の生活を維持するための行動が含まれる．**三次活動**は自由に使える時間における活動であり，趣味，娯楽などの個人の自由意志に基づく行動が含まれる．加えて，**活動なし**カテゴリも設定する．これは投稿テキストの内容に明確な行動が示されていない場合に用いられる．例えば，感情や抽象的な思考を含む投稿テキストなど，具体的な行動に直接結びつかない内容がこのカテゴリに該当する.「活動なし」を加えることで，全ての投稿テキストに対して適切なラベル付けを行い，データセットの包括性を高めた．

表 1 に 4 つの活動カテゴリとそれぞれのカテゴリに属する行動を示す．これらは，総務省統計局で定義された日常生活における行動に基づいており，必要に応じて独自の細分化を加えている．具体例として,「食事」を「食事-飲酒あり」,「食事-飲酒なし・不明」に分解している．この細分化は個人の健康状態やライフスタイルが日々の行動選択にどのように影響を及ぼしているかを理解する上で重要な役割を果たす．

3.3 アノテーションガイドライン

本研究では，投稿テキストに対して行動のラベル付けを行うためのアノテーションガイドラインを作成した．以下にアノテーションガイドラインを示す．

一つの行動に一つの行動ラベルを付与　各投稿テキスト内の主要な行動に対して一つのラベルを付与する．ただし，投稿テキスト内に複数の行動が存在する場合は，それぞれの行動に対応するラベルを付与する．これにより，投稿テキストが単一の行

表 2　投稿テキストと行動ラベルの例．太線部分は行動ラベルの根拠となる行動を表す．

	投稿テキスト	行動ラベル
(1)	頑張って！癌と闘う全ての人に安らぎを神様お願いします！	IV: 行動なし
(2)	**テニス**なう．この後の**焼肉**楽しみ．	III: 趣味・娯楽-体動かす / I: 食事-飲酒なし・不明
(3)	今から**買い物に行く**．	III: 買い物
(4)	**飯**後は，練馬駅近くのボロっちいゲーセンで**カードゲーム**．とうとうカードを**買って**しまった．でももめっちゃ楽しいんだよなー．	I: 食事-飲酒なし・不明 / III: 趣味・娯楽-体動かさない / III: 買い物

動を示す場合も，複雑な行動パターンを示す場合も，その全体像を把握できる．

投稿日の行動に限定　投稿日に発生した行動に対してラベルを付与する．判断が難しい場合は，常識的判断を使用する．

下位行動を優先　一つの行動に複数の行動ラベルが当てはまる場合，より下位の活動カテゴリ（一次活動より二次活動，二次活動より三次活動）の行動ラベルを優先して付与する．これにより投稿テキストが示す行動の最も具体的な側面を把握できる．

動詞の行動ラベルを優先　同じ活動カテゴリ内で，一つの行動に複数の行動ラベルが当てはまる場合，動詞の行動ラベルを優先して付与する．これにより投稿テキストが示す行動の最も活動的な側面を把握できる．

リツイート (RT) 付きの投稿の取り扱い　RT 付きの投稿では対象文章を RT 以前の文章とし，RT された部分は分析対象から除外する．

3.4　アノテーション結果

これらのルールに則って，アノテーションを行った投稿テキストの例を表 2 に示す．表 2 の（1）の投稿テキストには直接的な行動が含まれていないため,「行動なし」とした．(2) の投稿テキストは 2 つの行動（「テニス」と「焼肉」）が 1 文に含まれている．これらは投稿日の行動と言えるため「趣味・娯楽-体動かす」と「食事-飲酒なし・不明」の 2 つの行動ラベルを付与した．(3) の投稿テキストは「買い物に行く」という 1 つの行動に対して二次活動の「移動」と三次活動の「買い物」という 2 つの行動ラベルを付与した．そのため，**下位行動を優先**を使用し，三次活動の「買い物」ラベルを付与した．（4）の投稿テキストは 3 つの行動（「飯」,「カードゲーム」,「買って」）が 1 文に含まれている．これらは投稿日の行動と言えるため「食事-飲酒なし・不明」と「趣味・娯楽-体動かさない」と「買い物」という 3 つの行動ラベルを付与した．

このように，ガイドラインに従ってアノテーションを行い，各投稿テキストに対して最も適切な行動ラベルを付与した．このプロセスにより，行動分類モデルの学習と評価のためのデータセットを構築した．本研究では 2 人のアノテーターがアノテーションガイドラインに則りラベルを付与した．1 人目のアノテーターが 200 件，2 人目のアノテーターが 1000 件のデータにアノテーションを行った．このうちの 100 件に対しては 2 名がアノテーションし，それらを用いアノテーション一致度を算出する．79 件のデータに対するラベルが同じになったため，アノテーション一致度は 0.79 となった．

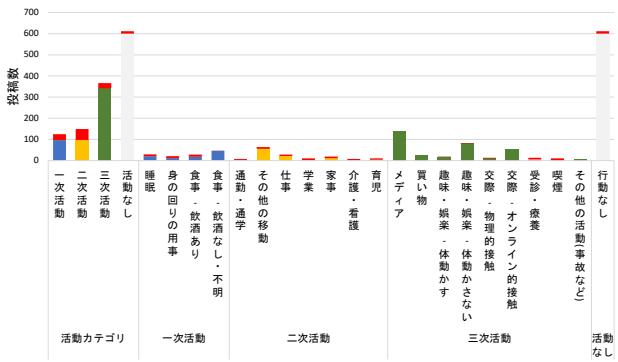


図 2　活動カテゴリ，行動ごとの投稿テキスト数．青色が「一次活動」，黄色が「二次活動」，緑色が「三次活動」，灰色が「活動なし」を表している．赤色は 4.2 節の ChatGPT4 を用いて作成した追加データを表している．

アノテーション結果を図 2 に示す．構築した行動データセットにおいて「行動なし」とラベリングされた投稿テキストが 600 個と最も多く，全体の 52.77%を占めている．これはユーザが具体的な行動を述べずに感情や意見を表現することが多いことを示している．活動カテゴリに注目すると,「活動なし」の次に「三次活動」の投稿テキスト数が多いことが分かる．これはユーザがより娯楽に近い行動に関する投稿をする傾向を示している．

4　手　　法

本研究では，行動データセットの質を図るために，ソーシャルメディアテキストから読み取れるユーザ行動を分類問題として推定する分類モデルを構築する．本章では，構築した行動分類モデル，ChatGPT4 を用いたデータ拡張について説明する．

4.1　行動分類モデル

本研究で構築した分類モデルは日本語テキストの理解に特化した東北大学の事前学習済み BERT モデルを基盤としている．BERT は大規模なテキストコーパスを用いて事前に学習されたモデルで，特定のタスクに対してファインチューニングすることで高い性能を発揮する．

構築した 7 つの分類モデルは表 3 の通りである．ソーシャ

表 3　構築した分類モデル

モデル	説明
Binary	「活動あり」,「活動なし」の 2 クラス分類モデル
Multi4	活動カテゴリの 4 クラスマルチラベル分類モデル
Multi21	全行動の 21 クラスマルチラベル分類モデル
Multi3	「活動なし」以外の活動カテゴリの 3 クラスマルチラベル分類モデル
Multi20	「行動なし」以外の全行動の 20 クラスマルチラベル分類モデル
Binary + Multi3	Binary 分類後に「活動あり」のデータに Multi3 を用いる活動カテゴリの 4 クラスマルチラベル分類モデル
Binary + Multi20	Binary 分類後に「活動あり」のデータに Multi20 を用いる全行動の 21 クラスマルチラベル分類モデル

ルメディアテキストを「活動あり」,「活動なし」のどちらかに分類する 2 クラス分類モデル Binary と,「一次活動」,「二次活動」,「三次活動」,「活動なし」の活動カテゴリに分類する 4 クラスマルチラベル分類モデル Multi4，3.2 節で定義した日常生活のユーザ行動に分類する 21 クラスマルチラベル分類モデル Multi21,「活動なし」以外の活動カテゴリに分類する 3 クラスマルチラベル分類モデル Multi3,「行動なし」以外の日常生活のユーザ行動に分類する 20 クラスマルチラベル分類モデル Multi20，Binary モデルを使用後に「活動あり」と分類されたデータセットに対して Multi3 モデルを用いることで活動カテゴリに分類する 4 クラスマルチラベル分類モデル Binary + Multi3，Binary モデルを使用後に「活動あり」と分類されたデータセットに対して Multi20 モデルを用い日常生活のユーザ行動に分類する 21 クラスマルチラベル分類モデル Binary + Multi20 としている．

3 章のアノテーションデータから「活動なし」のテキストは全体の 52.77%を占めている．Binary + Multi3，Binary + Multi20 モデルは「行動なし」のテキストを効率的に除外しすることでの全体の分類性能向上を目的として構築した．

4.2　ChatGPT4 を用いたデータ拡張

本研究では，ChatGPT4 を用いたデータ拡張を行い，行動データセットの質の向上を図った．特に少数ラベルのサンプルの増加を目的とし，拡張を行った．そのため，特に少数ラベルである,「睡眠」,「身の回りの用事」,「家事」,「食事ー飲酒あり」,「通勤・通学」,「学業」,「介護・看護」,「育児」,「受信・療養」,「喫煙」の計 10 個の行動ラベルに対して ChatGPT4 に Prompt を入力しデータの生成を行った．Prompt 例は次のようになっている．

表 4　ファインチューニング時のハイパーパラメータ．Multi モデルは 4.1 節で説明した Multi4, Multi21, Multi3, Multi20 モデルの総称．

モデル	パラメータ	値
	max length	128
	batch size	16
	optimizer	AdamW [9]
	学習率	2.0×10^{-5}
	エポック数	100
Binary	early stopping	3 patience
	損失関数	Weighted Cross Entropy Loss
Multi	early stopping	10 patience
	損失関数	WBCE With Logits Loss

Prompt 例

質問: [行動] を行っている時，もしくは [行動] を行った日にソーシャルメディア上に投稿する可能性が高い文章を 10 個生成せよ

作成された例文に対してアノテーションを行うことで，データの信頼性を上げる．10 個の少数ラベルそれぞれに対して例文を 10 個作成するため，データ拡張後のデータ数は 100 個増加する．図 2 はデータ拡張後の活動カテゴリ，行動毎の投稿テキスト数を表している．ここから少数ラベルのテキスト数が増加していることが分かる．

5　投稿テキストの行動分類実験

5.1　実 験 設 定

本実験では，4 章で説明した分類モデルの性能を行動データセットを用い 5 分割交差検証で評価する．具体的には，行動データセットを 9:1 の割合で学習・テストデータに分割する．その後，学習データを 5 分割し，4 つを学習データ，1 つを検証データとしてモデルの学習を行う．学習データ，検証データの組み合わせを変えて，5 通りのモデルを構築し，テストデータを用いて各モデルの性能を算出する．算出された 5 回分の分類性能の平均をモデルの性能とする．さらに，行動データセットはマルチラベルが付与されたデータセットのため，データセットを学習データ，検証データ，テストデータに分割する際に各ラベル間の分布に変化が少ないように分割を行なった．

分類モデルは東北大学の事前学習済み BERT モデルである cl-tohoku/bert-base-japanese-v3[1]を基に構築する．モデルの学習を行う際のハイパーパラメータは，表 4 のように設定した．2 クラス分類モデルである Binary の場合の誤差関数を Weighted Cross Entropy Loss，early stopping を 3 patience とした．さらに，マルチラベル分類モデルである Multi3, Multi4，Multi20，Multi21 の誤差関数を WBCE With Logits Loss，early stopping を 10patience とした．マルチラベル分類

1：https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

表 5　3 つのタスクにおける各モデルの分類性能．太文字はタスク単位で Accuracy，Macro-F1 の最大値を示している．w/ DA は ChatGPT4 を用いて Data Augmentation を行なったデータセットで学習したモデルを意味する．

タスク	モデル	Accuracy	Macro-F1
2 クラス分類	Majority Baseline	0.545	0.353
	Binary	0.62	0.626
	Binary w/ DA	**0.722**	**0.720**
4 クラスマルチラベル分類	Majority Baseline	0.536	0.174
	Multi4	0.536	0.518
	Multi4 w/ DA	**0.545**	**0.531**
	Binary w/ DA+Multi3 w/ DA	0.499	0.459
21 クラスマルチラベル分類	Majority Baseline	0.531	0.035
	Multi21	0.739	0.195
	Multi21 w/ DA	**0.770**	0.227
	Binary w/ DA+Multi20 w/ DA	0.651	**0.248**

モデルの validation loss の変化が 2 クラス分類モデルよりも乱れているため，マルチラベル分類モデルの early stopping を大きくした．さらに，行動データセットが不均衡データであるため，マルチラベル分類モデル構築時の誤差関数に重みつきバイナリクロスエントロピー （Weighted Binary Cross Entropy, WBCE）[10] を使用している．WBCE は，各ラベルのサンプル数に応じて重みを調整し，少数ラベルのサンプルが分類モデルによって適切に学習されるようにするものである．これにより，不均衡データにおいてもバランスの取れた学習が可能となる．ラベル毎の重みは以下の式で計算される．

$$\mathrm{class}_{\mathrm{weight}} = \frac{n_{\mathrm{samples}}}{n_{\mathrm{samples}_{\mathrm{class}}} \cdot n_{\mathrm{classes}}} \tag{1}$$

ここで，n_{samples} はデータの総数，$n_{\mathrm{samples}_{\mathrm{class}}}$ は各ラベルのデータの総数，n_{classes} はラベルの総数を表す．ラベル毎の重みを (1) 式と設定することで，サンプル数が少ないラベルへの重みが大きくなる．さらに，(1) 式の右辺分母において n_{classes} をかけることによって重みの正規化を行っている．これにより，少数ラベルが過大評価されることを防ぎつつ，それらのラベルに相対的に大きな重みを設定することが可能である．

本実験では，構築した分類モデルの評価指標として Accuacy と Macro-F1 を採用している．Macro-F1 は再現率と適合率の調和平均であり，分類モデルの性能を総合的に評価するのに適している．この指標は行動データセットのようにラベル間の分布が不均衡なデータセットを使用する場合に重要である．不均衡なデータセットでは，多数ラベルに偏った分類を行うことで，高い Accuracy を達成する可能性がある．しかしながら，Macro-F1 は少数ラベルの正確な分類性能も考慮に入れるため，よりバランスの取れた性能評価を可能にする．そのため，Accuracy，Macro-F1 の二つの評価指標を使用することで分類モデルの性能の全体像をより正確に評価できる．

5.2　結果と考察

分類モデルの性能を表 5 に示す．本実験ではベースラインモデルとして Majority Baseline を使用している．行動データセットでは「活動なし」カテゴリの「行動なし」を付与される投稿テキストが全体の 52.77%を占めている．そのため Majority Baseline は 2 クラス分類タスク，4 クラスマルチラベル分類タスクでは「活動なし」，21 クラスマルチラベル分類タスクでは「行動なし」をテストセットに付与する．全てのタスクにおいて構築した分類モデルが Majority Baseline より Accuracy，Macro-F1 において高くなっていることが分かる．さらに，データ拡張後のデータセットを使用して学習を行ったモデルの分類性能も，全てのタスクにおいて，Accuracy，Macro-F1 において拡張前のデータセットで学習を行っていたモデルより高い．これは，ChatGPT4 によるデータ拡張によってデータセットの質が向上したことを示している．

21 クラスマルチラベル分類タスクにおいて Accuracy，Macro-F1 の値が最も高い 2 つの分類モデルのより詳しい性能を表 6 に示す．表 6 から F1 値が 0.0 の行動ラベルが複数存在していることが分かる．ここから，ChatGPT4 を用いたデータ拡張を F1 値が 0.0 の行動ラベルおいて実施することでさらに質の高いデータセットが構築されると考えられる．

6　ケーススタディ

近年，運動習慣の変化や高齢化に伴い，アルコール依存症や，糖尿病などの生活習慣病が増加している．それに伴い，患者の日常生活における行動の把握が，医療現場において重要になっている．現在，患者へのヒアリングが日常生活における行動把握の主な手法となっている．しかしながら，ヒアリングには日々の情報収集の困難さや，情報の真偽を判断することの難しさなどの問題点がある．そこで本研究では，行動データセットで学習を行った行動分類モデルを用いて，ソーシャルメディア上のテキストデータを分析し，患者の日常生活における行動の把握を行う．具体的には，肝機能障害を有する可能性が高い人々と肝機能に問題がないと思われる人々の日常生活における行動の差異を分析することにより，この手法の有効性を検証する．

表 6　2 つの分類モデルによる性能比較．P は Precision, R は Recall.
太字は各タスク毎の F1 値の最大値を表している．

モデル	Multi21 (DA)			Binary (DA)+Multi20 (DA)		
Evaluation	P	R	F1	P	R	F1
睡眠	0.333	0.167	**0.222**	0.167	0.167	0.167
身の回りの用事	0.0	0.0	0.0	0.0	0.0	0.0
食事-飲酒あり	0.0	0.0	0.0	0.0	0.0	0.0
食事-飲酒なし・不明	1.0	1.0	**1.0**	0.753	1.0	0.856
通勤・通学	0.0	0.0	0.0	0.0	0.0	0.0
その他の移動	0.75	0.806	**0.739**	0.667	0.333	0.444
仕事	0.0	0.0	0.0	0.0	0.0	0.0
学業	0.0	0.0	0.0	0.0	0.0	0.0
家事	0.167	0.333	**0.222**	0.0	0.0	0.0
介護・看護	0.0	0.0	0.0	0.0	0.0	0.0
育児	0.0	0.0	0.0	1.0	1.0	**1.0**
メディア	0.667	0.37	**0.461**	0.25	0.111	0.154
買い物	0.0	0.0	0.0	0.0	0.0	0.0
趣味・娯楽-体動かす	0.0	0.0	0.0	0.667	0.667	**0.667**
趣味・娯楽-体動かさない	0.717	0.778	**0.719**	0.672	0.553	0.591
交際-物理的接触	0.0	0.0	0.0	0.0	0.0	0.0
交際-オンライン的接触	0.5	0.667	**0.556**	0.192	0.267	0.221
受診・療養	0.0	0.0	0.0	0.0	0.0	0.0
喫煙	0.0	0.0	0.0	0.333	0.333	**0.333**
その他の行動	0.0	0.0	0.0	0.0	0.0	0.0
行動なし	0.787	0.942	**0.857**	0.663	0.928	0.771

6.1 実験設定

使用するモデルは 5.2 節で 4 クラスマルチラベル分類タスクと 21 クラスマルチラベル分類タスクにおいて Macro-F1 が最も高かった Multi4 w/ DA モデルと Binary w/ DA+Multi20 w/ DA モデルとする．これらのモデルはテキストデータの複雑性を捉え，高い性能での分類を可能にする．

本研究では，肝機能障害を有する可能性が高い人々を抽出するために，肝機能障害を有する人に処方される薬である「ウルソ」という単語の使用に着目する.「ウルソ」を含むテキストを投稿したユーザを，肝機能障害を有する可能性が高い人（ウルソ投稿者）として定義した．対象的に，アルコール関連の単語を一度も投稿していないユーザを，健常者と定義した．データセットは行動データセット構築時に使用した 2013 年の 10ヶ月にわたって収集された Twitter（現 X）の投稿テキストとする．「ウルソ」という単語を含む投稿をしたことがある 85 人の肝機能障害者と，アルコール関連の単語である「お酒」,「ワイン」,「ビール」を含む投稿をしていない 85 人の健常者の投稿テキストを分析対象とした．なお，肝機能障害者である信頼性を高めるために,「ウルソ」という単語を薬品名以外で使用しているユーザは人手で削除した．

実験では，分類モデルを分析対象のデータセットに対して用い，投稿テキストに行動ラベルを付与する．この時，マルチラベル分類モデルの特性上どのラベルにも分類できない投稿テキストが存在する．その場合は「分類不可能」として分析から除外した．これにより，分析に使用するデータの信頼性を向上させた．分類後，各ユーザの各行動の割合を計算し，85 名のユーザに対して平均を求める．結果として，ユーザ毎の各行動割合の分散は小さく，分析の正確性が保証された．

6.2 結果と考察

図 3 は，投稿テキストから抽出された日常行動の割合を健常者とウルソ投稿者ごとに示している．図 3（a）および図 3（b）から健常者の方がウルソ投稿者よりも活動的な投稿が少ないことが分かる．これはアルコールを飲む習慣がある人々がそうでない人々に比べて活動的でない可能性を示している．さらに，ウルソ投稿者は健常者に比べ,「交際・オンライン的接触」の割合が多いことが分かる．これはアルコールを飲む習慣のある人々がソーシャルメディア上でのやり取り，メール，電話の頻度が高いことを示している．さらに，ウルソ投稿者は健常者に比べ,「メディア」の使用割合が高いことも分かる．これは，肝機能障害を有する可能性がある人は，受動的に情報に触れている頻度が高いことを示唆している．全体として，アルコールを飲む習慣がある人々は身体的な活動を伴わない行動の割合が高いことが示された．

6.3 限　　界

本実験では,「ウルソ」という単語を含む投稿をしたユーザを肝機能障害を有する可能性が高い人として抽出し，分析を行った．しかしながら，このアプローチにはいくつかの限界が存在する．まず，単に「ウルソ」という単語を含む投稿をしたことがあるだけで，実際にそのユーザが肝機能障害を有し，かつウ

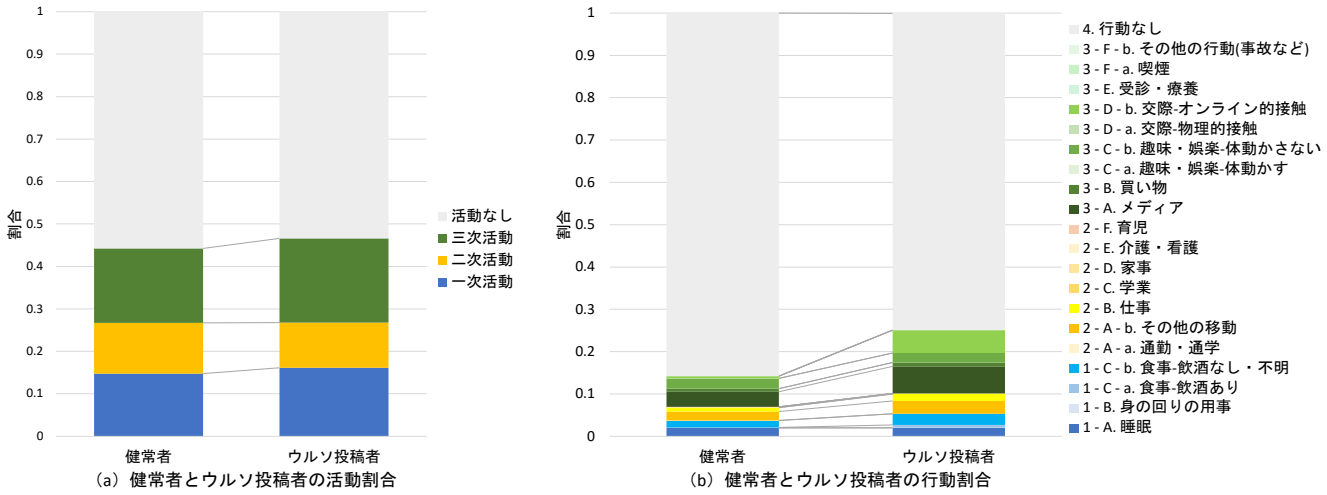


図 3　投稿テキストにおける行動比較．青色系統は「一次活動」，黄色系統は「二次活動」，緑色系統は「三次活動」，灰色は「活動なし」を表す．

ルソを服用しているとは限らない．そのため，単語の使用を元にした抽出は，実際の行動や状態を正確に反映しているとは限らない可能性がある．さらに，分析の性能に影響を与えるもう一つの限界は，アノテーションのガイドラインに基づく行動ラベルの付与プロセスにある．本実験では，1 日のうちに同じ行動を示す投稿が複数存在する場合でも，それぞれに行動ラベルを付与する．これにより，実際は同一の行動が複数回カウントされることになり，分析結果において特定の行動の割合が実際よりも高く評価される可能性がある．

これらの限界は，モデルの分析結果に影響を及ぼす可能性があり，その解釈に際しては，慎重な検討が必要である．今後の研究では，これらの問題に対処するために，より緻密なユーザ抽出基準の設定や，アノテーションガイドラインの変更などが考えられる．

7　おわりに

これまで，多くのソーシャルメディアテキストはタスク固有の分類を解いていたが，一次活動から三次活動という人間の基本的な行動レベルで，汎用性のある分類を行った研究は少ない．本研究では，ソーシャルメディアテキストを用いた基本的なユーザ行動レベルでの研究基盤となるデータセットを構築し，分類モデルの構築，ケーススタディという一連の流れを試行した．さらに，ChatGPT4 を用いた少数ラベルのデータ拡張を行い，この手法がデータセットの質を向上させることを示した．

今後の発展としては，更なる少数ラベルのデータ拡張や，多様なソーシャルメディアプラットフォームのデータを用いたモデルの汎用性と効果の検証が考えられる．これにより，本研究のアプローチの適用範囲を広げ，社会的行動やトレンドのより深い理解に貢献するための分析の可能性を高めることが期待できる．

謝　辞

本研究は，JSPS 科研費 JP22K12041，JST SICORP JPMJSC2107，AMED の課題番号 JP22mk0101229 の支援を受けたものである．

文　献

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

[2] Yuxing Qi and Zahratu Shabrina. Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, Vol. 13, No. 1, February 2023.

[3] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1146–1151, 2014.

[4] Sandra Matz, Jochen Menges, David Stillwell, and H. Schwartz. Predicting individual-level income from facebook profiles. *PLOS ONE*, Vol. 14, p. e0214369, 03 2019.

[5] Nilanjan Banerjee, Dipanjan Chakraborty, Koustuv Dasgupta, Sumit Mittal, Anupam Joshi, Seema Nagar, Angshu Rai, and Sameer Madan. User interests in social media sites: an exploration with micro-blogs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, p. 1823–1826, New York, NY, USA, 2009. Association for Computing Machinery.

[6] The-Minh Nguyen, Takahiro Kawamura, Yasuyuki Tahara, and Akihiko Ohsuga. Self-supervised capturing of users' activities from weblogs. *International Journal of Intelligent Information and Database Systems*, Vol. 6, pp. 61–76, 01 2012.

[7] Yuji Yano, Tomonori Hashiyama, Junko Ichino, and Shun'ichi Tano. Behavior extraction from tweets using character n-gram models. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1273–1280. IEEE, 2014.

[8] 統計局. 統計局ホームページ/3. 1 年間の行動の種類についての

説明（行動調査）, 2001. アクセス日: 2024 年 1 月 9 日.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proc. of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 523–540, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

機械学習とクラウドソーシングを用いた顧客の声の自動分類

鈴木　優† 城所　祥太† 北村　拓斗† 太田　奈那† 三島　惇也†
本田　文隆†† 長徳　弘一†† 中村　勇太†† 佐々木和彦†† 笠丸　陽一††

† 岐阜大学工学部/自然科学技術研究科　〒 501–1193 岐阜県岐阜市柳戸 1–1
†† ソニー株式会社　〒 108–0075 東京都港区港南 1–7–1
E-mail: †suzuki.yu.r4@f.gifu-u.ac.jp

あらまし　本研究では，ソーシャルメディアに投稿されたある製品に対する評判（顧客の声）を，機械学習モデルとクラウドソーシングを用いて自動分類を行うことを目的とする．具体的には，投稿者のメッセージを，何について書かれているか，それについてどう思うかについて分析する．分析においては，それぞれのメッセージを見る作業量と，分類精度が問題となっている．近年普及している LLM では，何について書かれているかを高精度に判断できるが，感情分析については十分な精度ではないことが分かった．一方で，数十個という大量の選択肢に対して適切な回答を行うことはクラウドソーシングでは難しい．そこで，LLM とクラウドソーシングを組み合わせることによって目的を達成する方法を提案する．実際にいくつかのタスクにおいて LLM とクラウドソーシングにより評価を行った上で，LLM とクラウドソーシングを組み合わせた新しいクラウドソーシングシステムを構築した．

キーワード　クラウドソーシング，ソーシャルメディア，実データ，データマイニング

1 はじめに

一般消費者向けに様々な製品を開発するとき，良い評判や改善点の分析，つまり評判分析を行うことがある．YouTube や 5ちゃんねるなどのソーシャルメディアには，製品に対する評判を含むメッセージが投稿される．本研究の目的は，これらソーシャルメディアのメッセージを機械学習モデルにより自動的に分析するシステムの構築である．

自動的に評判を分析する代表的な方法として，クラウドソーシング [1] により教師データを構築した上で機械学習モデルを構築する方法と，ChatGPT などの LLM （Large Language Model; 大規模言語モデル）を用いる方法の二つを挙げることができる．機械学習モデルを構築するためにはあらかじめ教師データを準備する必要があるが，大規模な教師データを構築するためにはクラウドソーシングなど人手が必要である．目的に応じた教師データを準備することができれば，高い精度で分類を行うことが期待できる．一方，LLM を用いた場合には教師データを事前に与える必要は無いため教師データを準備するためのコストは不要であるが，教師データが一般的なものであるため，特定の製品に関する評判分析に用いるためには精度には限界がある場合も存在すると考えられる．つまり評判分析を行う上で，クラウドソーシングによる機械学習と LLM において，それぞれ精度と金額的，時間的なコストを各手法の特性として計測する必要がある．その上で，これら二つの方法の特性を活用することによって，時間や手間を削減した上で精度の高い分析を行うことが可能となる．

このときに課題となる点は，教師データの品質，および構築するために必要な時間，費用である．分析対象となるメッセージが大量となったときや，多くの点で分析を行う必要があるとき，現実的な時間で分析を行うためには全ての手順をクラウドソーシングにより人手で行うことは現実的ではない．たとえば，ある製品に関して必要な分析項目（デザインや色，重さなど）が 50 個あったとき，クラウドソーシングにより各メッセージに対してどの分析項目について書かれているかを選択することは極めて複雑であり，多くの時間と費用がかかるため，LLM により処理することが有効であるといえる．一方で各分析項目が判明しておりそれぞれの分析項目に対してメッセージ投稿者がどのような感情（ポジティブやネガティブなど）を持っているかを分析するときには，クラウドソーシングにより人手で処理することが有効である．そこで本研究ではクラウドソーシングと LLM を組み合わせ，複雑な処理を現実的な時間，費用で処理可能なシステムの構築を目指す．

また，本研究では機械学習により精度を向上させるための手法として，マルチタスク学習 [2] が Caruana により提案され，広く用いられている [3]．マルチタスク学習とは，本来の目的となる分類問題である主タスクと，別の分類問題である副タスクを同時に回答することが可能な分類器を構築することにより，主タスクの精度を向上させる方法である．副タスクの回答が主タスクの回答に影響を及ぼすとき，主タスクの分類精度が向上すると考えられている．ところが，どのような副タスクが主タスクの精度を向上させるかは明らかではない．

そこで本研究では，実際の製品に対する評判分析を行うシステムを構築する．クラウドソーシングによる機械学習モデルと LLM を協調的に組み合わせることによって，精度を維持しつつ時間的および金銭的コストを削減することができるシステムの構築を目指す．

本研究の貢献は次の通りである．

- クラウドソーシング，機械学習モデル (BERT)，LLM の三種類のうち，どのような種類のタスクに対してどの程度精度

に差異があるかを明らかにした．

- LLM とクラウドソーシングを組み合わせた教師データ構築システムを提案した．
- マルチタスク学習により，機械学習モデルを構築した．
- 機械学習モデル (BERT) と LLM だけを組み合わせ，自動的に評判分析システムを構築する手法を提案した．

2 関 連 研 究

メッセージに対して自動的に評判分析を行う研究は，数多く存在している．特に深層学習による評判分析は数多く行われており，Zhang ら [4] によりその詳細がまとめられている．

評判分析を行う上で問題となる点は，評判分析に必要なデータが十分ではないことである．特にマルチタスク学習に必要なデータには，主タスクだけではなく副タスクを設定する必要があるが，我々が探す限そのようなデータを見つけることはできなかた．我々は以前に Twitter 日本語評判分析データセット [1] と呼ばれる大規模評判分析データセットを構築し，公開している．一方，SemEval[2]は意味分析に関する国際ワークショップであり，評判分析タスクもその一部である．

既存研究と本研究の違いは二つあり，マルチタスク学習を用いることを前提としたデータセット構築と LLM との協調である．既存のデータセットではメッセージそれぞれに評判に関するラベルだけが付与されており，マルチタスク学習に用いることができなかった．本研究では一つのメッセージに対して様々な質問を行い，その回答を収集しているため，マルチタスク学習による精度向上が期待できる．また，LLM は BERT などによる既存の分類器よりも質問によっては高い精度で分類できる場合があるため，効果的に組み合わせることによって精度を維持しつつ時間的および金銭的コストを削減することが可能となる．

3 提案システム

提案システムは，教師データ構築と自動分類の二つに分けられる．そこで，3.1 節では問題の定義とクラウドソーシング，LLM を用いたデータの構築について述べ，3.2 節では 3.1 節で構築したデータを用いた自動分類の方法について述べる．

本システムを構築するにあたって注意したことは，適合率よりも再現率を重視したことである．つまり，ある程度精度を下げた場合であっても，機械学習モデルの教師データとして扱うべきメッセージをなるべく多く扱うことを目的としている．

3.1 タ ス ク

教師データを構築するためには，まずどのようなデータを構築するかを定義する必要がある．そこで，タスクを定義した上で，そのタスクをどのように処理するかについて述べる．本稿ではタスクを一つの質問とその回答群の組であると定義する．つまり，本研究における教師データ構築は，複数のタスクより構成されている．

本研究における教師データ構築の概要を図 1 に示す．まず，ソーシャルメディアなどからある製品に関するメッセージを収集し，システムへ入力する．次に作業者によってメッセージを読み，最大 5 問の質問に対して回答する．その結果を集約することにより，メッセージが製品とどのような関連があるのか，投稿者はどのような感情を持っているのか，使ったことがあるかどうか，他製品と比較しているかどうかについての情報を得ることができる．集約は単純な多数決であり，同率で多数の結果があれば複数の結果が付与される可能性がある．

3.1.1 タスクの定義

本システムでは，以下六つのタスクを設定した．

Q1 ：メッセージに対象に対する感情や体験，要望が含まれているか

ソーシャルメディアに含まれるメッセージには，日本語としての意味をなさないものや，対象に関係の無いことだけが書かれているものが存在するため，これらを排除するための選択肢である．選択肢として「含まれている」「含まれていない」のいずれか一つを選択する．対象に対する感情や体験，要望などが読み取れる場合には「含まれている」を選択する．Q1 の回答が「含まれていない」である場合には，以降の回答を行わない．

Q2L ：メッセージが対象の何について書かれているか

対象を分類するための項目を，二つの粒度で考えた．一つは大分類であり,「本体」「品質・不具合」「アクセサリ・接続機器」「その他」「該当なし」の 5 項目のうち一つ以上を選択する．項目を複数選択することは可能であるが,「該当なし」と他の項目を同時に選択することはできない．もう一つは小分類であり,「本体のデザイン」や「本体の色」など 64 項目のうち一つを選択する．以下，大分類は Q2L，小分類を Q2s と呼び，このシステムでは大分類 Q2L を用いる．

Q3 ：投稿者は対象についてどのような感情を持っているか

テキスト全体から受ける，投稿者が対象に対して感じているであろう感情を推定する．選択肢として「ポジティブ」「ネガティブ」「ポジティブ+ネガティブ」「ニュートラル」のうち一つを選択する．今回の実験では最も主となるタスクである．

Q4 ：投稿者は対象を使用したことがあるか

選択肢として「はい」「いいえ・分からない」のうち一つを選択する．メッセージを作業者が読んだとき，投稿者が明確に対象を使用していると断言できない場合が多く，分からない場合が多い．そこで，いいえと分からないを一つの選択肢とした．マルチタスク学習において副タスクとして用いる．

Q5 ：投稿者は対象と競合する製品と比較しているか

選択肢として「同社の製品」「他社の製品」「比較していない」のうち一つを選択する．様々な製品と比較しているメッセージがあるため，それらの製品との差異を対象の印象として表現しているかどうかを判定している．

Q6 ：Q1 から Q5 までの作業を行う上で迷った質問はあるか

選択肢として「Q1」「Q2」「Q3」「Q4」「Q5」のうち一つ以上

1：https://www.db.info.gifu-u.ac.jp/sentiment_analysis/
2：https://semeval.github.io

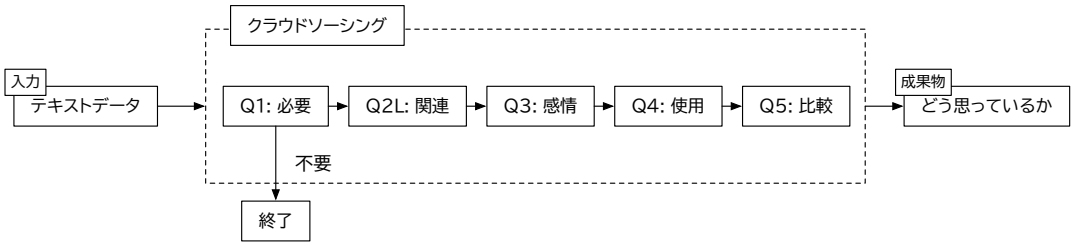


図 1　教師データ構築システムの概要

図 2　クラウドソーシングシステム画面キャプチャ

を選択する．この質問および回答は提案システムの有効性を議論するためだけに用いるため，機械学習モデルへの教師データとしては用いないが，今後の研究において用いる可能性がある．ここで実際のタスクでは対象に具体的な製品名が示されており，スマートフォンのブランド名である．

タスクを設定するとき，作業依頼者の視点と作業者の視点，LLM の精度を考慮した．つまり，作業依頼者が本当に知りたい情報であるかどうかと，クラウドソーシングにおける作業者にとって回答がしやすいかどうかの 2 点を考慮した．タスク数が増えると作業依頼者にとって有用なデータとなるが，作業者にとって多くのタスクを処理する必要があり，1 メッセージあたりの作業時間や費用が多くなる．一方，タスク数が減ると作業依頼者にとって有用性が減少するが，1 メッセージあたりの作業時間が削減し，費用が減る．そこで，適切な数で選びやすい選択肢であり，しかも作業依頼者にとって有用な結果が得られるように質問および回答を作成した．

3.1.2　クラウドソーシング

クラウドソーシングにより 3.1.1 節で設定したタスクを，作業者が入力となるメッセージデータに対して処理する．クラウドソーシング作業画面を図 2 に示す．Web 上で主にスマートフォンを用いて作業を行うことができるよう，クラウドソーシングによる作業環境を構築した．なぜなら，従来の評価実験においてクラウドソーシングを行ったとき，80%以上の作業者がスマートフォンもしくはタブレットにより作業を行っているためである．作業者は，画面中央にあるメッセージを閲覧し，画面下にあるボタンのうち対応するものを押すことにより，次のメッセージもしくは次の質問が現れる．

構築したシステムはフロントエンドとバックエンドの二つで構成されており，フロントエンドは作業者のブラウザ上で，バックエンドは作業依頼者が準備したサーバ上で動作する．フロントエンドとバックエンドは両方とも JavaScript で実装されており，フロントエンドは React.js，バックエンドは Node.js がベースとなっている．バックエンドで用いるデータベースは PostgreSQL を ORM として Prisma.js を介して用いている．フロントエンドとバックエンドとの通信は GraphQL によって JSON を POST することによって行っている．本システムは今回の実験だけではなく様々なタスクを処理することが可能なように設計されているため，今回のタスク以外でも利用することが可能である．

作業者は Web システムにログインすると，まず作業説明が画面に表示される．この作業説明には，どのような場合にどの選択肢を選べば良いのか，基準が書かれている．作業者が作業説明を読み終えたあと，メッセージと一つの質問，および選択肢がそれぞれ画面に表示される．質問に回答すると次の質問と選択肢が表示される．最後の質問に回答すると，次のメッセージ，最初の質問および選択肢が表示される．

作業者がタスク Q1 で「含まれていない」と選択したメッセージは，作業者が処理したメッセージの個数に含めないこととした．これは，Q1 で「含まれている」を選択した場合には作業者は Q1 から Q6 までの回答を全て行わなければならないことと比較して,「不必要」を選択した場合には Q1 だけを回答したら良いため，作業時間に大きな差がある．このため，もし「不必要」を作業としてシステムがカウントする場合，作業者は報酬あたりの時間を短縮するために「不必要」を想定以上に選択する可能性がある．このような問題を解決するために，Q1 の選択肢に応じて処理済みメッセージとして数えるかどうかを区別した．

Q2 については，5 項目である Q2L を用いた．これは，Q2s である 64 項目の選択肢をクラウドソーシングの作業者により選択する作業に極めて時間がかかり，費用を考えると不可能であるためである．

最後に，作業者により得た作業結果を多数決により集約する．

システムは一つのメッセージに対して k 人の作業者を割り当てる．全てのタスクにおいて作業者による作業結果のうち一番多い選択肢を，クラウドソーシングによる作業結果として集約する．

作業結果に含まれるメッセージ群および回答を 8:1:1 に分割し，8 割を教師データ，1 割を開発データ，残りの 1 割を正解データとする．ここでの正解データを作業者正解データと呼ぶ．評価実験では 4.1.1 節において作業依頼者により正解データを作成するが，こちらは依頼者正解データと呼び区別する．

3.2 自動分類

3.2.1 機械学習モデル

本研究では，BERT [5] を用いてメッセージの自動分類を行った．入力データとして，クラウドソーシングによる作業結果のうち学習データを用いる．開発データによりハイパーパラメータの最適化を行う．機械学習モデルを構成するニューラルネットワークのパラメータを継続して更新し，その都度作業者正解データを用いて Accuracy および Loss を計算した．アーリーストッピングを用いることによって，学習を進める段階である一定以上 Loss が更新されない場合は，過学習を防ぐためにパラメータの更新を止めた．ここで構築した機械学習に対してメッセージを入力し，正解データを推定した．ここで推定したデータを BERT 推定データと呼ぶ．

なお，この分類に BERT を用いる必要は特になく，他の分類器を用いることが可能である．そのため，タスクによって適した機械学習モデルを構築する必要があるが，評価実験では BERT を用いた．

3.2.2 LLM

3.1.1 節で設定したタスクを LLM によって処理する．クラウドソーシングにおけるタスクをそれぞれプロンプトへ変換し，メッセージと共に入力し，正解データを推定した．ここで推定したデータを LLM 推定データと呼ぶ．

Q1 に対するプロンプトを図 3 に示す．(メッセージの一覧) と書かれている部分は，複数のメッセージと入れ替えた部分である．

4 評価実験

本提案システムが十分な精度であることを確認するため，評価実験を行った．クラウドソーシングによる機械学習モデルと LLM の二つを，分類精度である Accuracy および混同行列により定量的な評価を行った．また，実際の判定結果を人手で分析し，それぞれの手法における利点，欠点を定性的に分析した．ここで正解データは二種類あり，クラウドソーシングによる作業者正解データと，作業依頼者による作業依頼者正解データがある．作業者正解データは，BERT を用いる上で機械学習が正常に行われているかどうかを判定するために用いた．一方で作業依頼者データは，クラウドソーシングによる機械学習モデルと LLM の両方の結果の精度を求めるために用いた．

```
# 命令書
あなたはラベル付けを行うクラウドワーカです。以下の入力文は Xperia に関係したメッセージです。以下の制約条件と入力文もとにを最適なラベルを出力してください。
# 制約条件
  • 入力文に,Xperia に対する感情または要望または体験に関する情報が含まれている場合のラベルは「必要」とする
  • 入力文に,Xperia に対する感情または要望または体験に関する情報が含まれていない場合のラベルは「不要」とする
  • 出力は出力文のうち最適なもの，ただ一つとする
# 入力文:
  • （メッセージの一覧）
# 出力文:
  • 必要
  • 不要
```

図 3 LLM へのプロンプト (Q1)

4.1 メッセージ収集

本研究では，メッセージを 2 種類の情報源から収集した．一つは対象に関する動画を YouTube から抽出し，YouTube API[3] を用いてコメントを抽出した．以下 YouTube データと呼ぶ．もう一つは 5 ちゃんねる [4]に投稿されているメッセージのうち対象に関係するスレッドを抽出した．以下 5 ちゃんねるデータと呼ぶ．データ収集を開始した時期が 2023 年 10 月頃であるが，新しい動画から順に古い動画へ遡るように動画の URL を収集し，その上でその動画群に含まれるコメントを収集した．

4.1.1 評　　価

4.2.1 節で構築した正解データとは別に，作業依頼者による正解データである依頼者正解データを構築した．作業依頼者が 500 件のメッセージに対して，3.1.1 節で示したタスクを行った．本稿の第 6 著者から第 9 著者までの 4 名が作業者として作業を行い，正解を作成した．それぞれの作業者による作業結果が異なる場合には，作業者相互で議論を行った上で結果を作成した．合議の結果であっても作業結果に相違がある場合，対象となるメッセージから除外した．

評価尺度に F 値を用い，3.1.1 節に示した各タスクの精度を評価した．Q1，Q4，および Q5 は二択であるため，ポジティブである選択肢が選択できているかを測定した．Q2 は 3.1.1 節に示した 5 択であるが，該当なしを除く 4 択についてそれぞれ正しい選択をしているかどうか F 値を示した．Q3 は「ポジティブ」を正解とした場合の F 値を示している．

4.2 教師データ構築

4.1 節で示したメッセージに対して，次のように学習データと正解データを構築する．

4.2.1 クラウドソーシングと BERT

まず我々は，クラウドワークス [5]を用いて作業者を雇用した．

3：https://developers.google.com/youtube/v3?hl=ja
4：https://itest.5ch.net/
5：https://crowdworks.jp/

不特定多数の作業者を雇用しており，以前に極めて低品質な作業を行った履歴が無いことを確認した上で雇用した．次に，バックエンドシステムにより一つのメッセージあたり 10 名の作業者を無作為に割り当てた．作業者は，100 件のメッセージを処理することによって約 50〜90 円の報酬を得るように設定した．作業者が 100 件より少ないメッセージしか処理していない場合は，報酬を支払わなかった．

次に，ここで構築した教師データをマルチタスク学習により学習させ，機械学習による分類モデルを構築する．まず，Q1 においてクラウドソーシングによる学習が不要であると判定されたデータを教師データから削除した．そして，このタスク群では最も重要なタスクが Q3 の感情分類であったことから Q3 を主タスクとし，それ以外のタスクを副タスクとして学習した．この方法により 26,072 件の教師データにより機械学習モデルを構築した．回答の不均衡を是正するため，Q3 は全ての回答が同じ個数となるよう回答の複製を行った．それ以外のタスクについては回答の不均衡の是正は行っていない．

4.2.2 LLM

本研究では OpenAI 社の ChatGPT API[6]を用いて，3.2.2 節に述べた方法により LLM を用いて回答を生成した．このとき，ChatGPT に対してファインチューニングなどを用いた機械学習モデルのパラメータ調整は行っていない．

4.3 実験結果

依頼者正解データをクラウドソーシングの多数決，機械学習モデル，ChatGPT の三種類の方法で当てることができるかどうかを評価する．実験結果を表 1 に示す．Crowdsourcing はクラウドソーシングの多数決，BERT はクラウドソーシングの多数決による作業者正解データを教師データとして用いた機械学習モデルの出力．GPT4 は ChatGPT による回答である．

この表から，8 個の結果のうち 4 個はクラウドソーシングによる精度が最も高く，BERT による精度と ChatGPT による精度が共に 2 個は精度が高かった．ところが提案システムの使用方法を考えると，全てのメッセージをクラウドソーシングにより分類することは費用および時間の面で現実的ではない．そのため，BERT と ChatGPT の比較を行うと，8 個のうち 6 個は ChatGPT のほうが精度が高く，2 個は BERT のほうが精度が高いことが分かる．BERT に必要な教師データの数が十分ではないことはこの結果の原因の一つであるが，どれだけ増やしたとしてもクラウドソーシングによる精度を BERT が超えることは期待できない．

タスクを，事実タスクと感情タスクの二つに分類することを考える．例えば Q2 であるメッセージの対象は作業者により回答が変わらないと考えられることから，事実タスクであると分類する．Q1，Q4，Q5 も同様である．一方で，Q3 は投稿者の感情を推測する必要があり，作業者の推測によって回答が変わる可能性がある．そのため，感情タスクであると分類できる．この結果を基に結果を見ると，事実タスクでは LLM による回答とクラウドソーシングによる回答に精度に大きな差が無い．一方で，感想タスクは LLM による回答と比べて有意に高い精度をクラウドソーシングで出しており，BERT により出すことが可能であると考えた．

これらの結果から，感情タスクをクラウドソーシングと BERT，事実タスクを LLM により分担して行うことによって，現実的な費用や時間で精度の高いシステムを構築できると考えた．この知見により，LLM とクラウドソーシングを組み合わせた新たなシステムの構築を 5 章で提案する．

5 クラウドソーシングと LLM を組み合わせた教師データ構築システム

4 章における知見から，感想タスクをクラウドソーシングと BERT，事実タスクを LLM により行う教師データ構築システムを提案する．4 章で作成したシステムと異なり，詳細な関連およびそれぞれの関連に関する感想を出力できるようにした．

5.1 システムの全体像と改善点

システムの全体図を図 4 に示す．特に 3 章で提示したシステムとの差異と，その理由について説明する．

最も重要な変更点は，Q2 であるメッセージが何に関連しているかという質問をクラウドソーシングではなく LLM によって行うという点である．作業依頼者は本来，Q2s で示した 64 項目の選択肢についての関連を選択する必要があったが，クラウドソーシングでこの作業を行うことが難しいと考えたため，Q2L で示した 5 項目に選択肢を削減していた．ところが，Q2L を ChatGPT 3.5-turbo で回答を出力したところ，実用的な精度であることが分かった．そこで，Q2 については Q2s を LLM によって回答することとした．最大でも 5 項目まで収集することとし，それ以上は回答しないようなプロンプトを組んだ．

Q2 の変更に伴い，Q3 の感情判定タスクを Q2 の出力ごとに変更した．つまり，Q2 により本体サイズ，本体の色，本体のデザインという三項目の回答を得たとき，Q3 ではこれら三項目それぞれについて感情判定を行う．

また，Q4: 使用の有無や Q5: 競合製品との比較有無に関するタスクは，人手であっても意見が分かれるため作業結果の品質が一定しないことや，マルチタスク学習のサブタスクとして

表 1 実験結果 （F 値）．Q1 のみ GPT3.5-turbo を GPT4 の代わりに使用している．太字はクラウドソーシング，BERT，GPT4 のうち一番精度が高いことを示している．

	クラウドソーシング	BERT	GPT4
Q1 必要不要	**0.877**	0.785	0.872
Q2 本体	0.537	0.561	**0.642**
Q2 品質・不具合	0.691	**0.790**	0.601
Q2 アクセサリ・周辺機器	0.896	0.601	**0.965**
Q2 その他	**0.716**	0.508	0.704
Q3 感情	**0.776**	0.507	0.644
Q4 使用	0.671	**0.669**	0.583
Q5 比較	**0.854**	0.719	0.782

6：https://platform.openai.com/

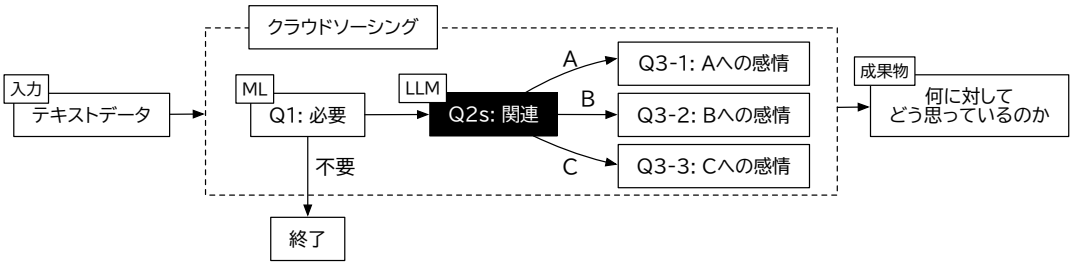


図 4　教師データ構築システムの概要

も性能が上がらないことから，タスク群から削除した．

5.2 タ ス ク

本研究のタスクは以下の三つである．

Q1 ：メッセージに対象に対する感情や体験，要望が含まれているか

4 章で示したシステムと同じであるが，4.1.1 節で示した評価実験において収集した投票データを用いて構築したデータを教師データとして BERT により機械学習モデルを構築し，分類を行う．機械学習モデルにより感情や体験，要望が「含まれている」と判定されたものに対して，人手により再度判定を行い，両方とも「含まれている」と判定されたものだけに対して以下のタスクを実行する．

Q2s ：メッセージが対象の何について書かれているか

4 章で示したタスクのうち Q2s; 小分類を用いる．「本体のデザイン」や「本体の色」など 64 項目のうち一つを選択する．項目の一つは「該当なし」であり，どの項目にも該当しない場合に選択される．そのため，必ず一項目以上が選択される．LLM もしくは機械学習モデルによって自動的に分類する選択肢であることを念頭に置いており，人手による分類は難しいと考えている．

Q3 ：投稿者は対象についてどのような感情を持っているか

Q2 で選ばれた項目のうち「該当なし」以外の項目それぞれについて，投稿者が対象に対して感じているであろう感情を推定する．選択肢として「ポジティブ」「ネガティブ」「ポジティブ+ネガティブ」「ニュートラル」のうち一つを選択する．今回の実験では最も主となるタスクの一つである．

作業者は実際には Q1 と Q3 だけを回答する．Q1 で「含まれている」と判定した場合であっても，Q2s で「該当なし」以外の項目が選択されない場合には Q3 は回答しない．

Q1 や Q2s を実行するタイミングについて，クラウドソーシングを実行する前にバッチ処理を行うか，クラウドソーシングを実行している途中で実行するかを決める必要がある．なぜなら機械学習モデルや LLM からの返答が遅ければ，作業者は回答を得るまでの時間待つ必要があるため，作業効率が落ちてしまうためである．一方で途中で実行が可能であれば，作業者によって Q1 で「含まれていない」と判定されたメッセージに対して Q2s を実行する必要が無くなり，費用および時間を削減できる．実際にシステムを構築した結果，作業者はほとんど待つことなく Q1 および Q2s を実行することが可能であった．そこで，事前に機械学習モデルや LLM の回答を得ず，クラウドソーシング実行中に機械学習モデルや LLM を実行するシステムを構築した．

6 おわりに

本研究では，ソーシャルメディア上のメッセージに対してクラウドソーシングと機械学習モデル，および LLM などを組み合わせて自動的に分類を行う方法の提案を行った．

一般消費者向けに様々な製品を開発するとき，その製品に対する良い評判や改善点などの分析，つまり評判分析を行う必要がある，YouTube や 5 ちゃんねるなどのソーシャルメディアには製品に対する評判を含むメッセージが投稿されるため，これらのメッセージを分析することによって評判分析が可能となる．ところが，分析対象となるメッセージが大量となったとき，現実的な時間で分析を行うためには分析の自動化が必要となる．そこで本研究では，メッセージの投稿者がある製品に対してどのような感情（ポジティブ，ネガティブ，ニュートラルなど）を持っているのかなどの製品に関する評判，およびそれに付随するラベルを自動的に付与するシステムを構築した．

まず，クラウドソーシングと機械学習モデル，LLM の三種類について，精度比較を行った．評価実験の結果，感想に関係するタスクはクラウドソーシングが性能を上回り，機械学習モデルによる性能向上が期待されるが，事実に関係するタスクは LLM により十分な性能を持つことが分かった．そこで，この結果を基に新たな教師データ構築システムを構築した．

今後の課題は，二つ目の教師データ構築システムを用いて実際にクラウドソーシングにより教師データを構築し，全ての分類を機械学習モデルだけで行った場合に，十分な精度を得ることができるかどうか評価実験を行うことである．クラウドソーシングだけ，LLM だけを利用したシステムと比較し，提案手法が有効であることを示す．

さらに，対象となるデータをスマートフォンだけではなく様々な製品に対する評判分析を行う．このとき，転移学習を用いて今回構築したスマートフォン向けの教師データを活用することによって，対象となる製品に関する教師データだけを用いる場合よりも精度が高くなることが期待できる．

謝　　辞

本研究の一部は，JSPS 科学研究費 (19H04218，23H03694)，

越山科学技術振興財団，およびソニー株式会社の助成を受けたものです．

文　献

[1] Eric Schenk and Claude Guittard. Crowdsourcing: What can be Outsourced to the Crowd, and Why ? working paper or preprint, December 2009.

[2] Rich Caruana. Multitask learning. *Machine Learning*, Vol. 28, No. 1, pp. 41–75, Jul 1997.

[3] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 12, pp. 5586–5609, 2022.

[4] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis : A survey. 01 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年2月28日(水) 16:00 ~ 18:10 | T3-B オンライン（Zoom Events)

データ検索・解析

座長:飯沢 篤志(リコーITソリューションズ株式会社)
コメンテータ:國島 丈生(追手門学院大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-B-3-01] 鏡映を識別可能な3次元CADモデル検索のためのバイスペクトルの利用

*小島 優希也[1]、片山 薫[1] (1. 東京都立大学)

16:30 ~ 16:55
[T3-B-3-02] 流通を考慮した研究データのメタデータ付与支援

*篠原 桐[1]、清水 敏之[2]、冨浦 洋一[3] (1. 九州大学大学院統合新領域学府冨浦研究室、2. 九州大学附属図書館研究開発室、3. 九州大学大学院システム情報科学研究所)

16:55 ~ 17:20
[T3-B-3-03] スプレッドシートを用いたデータ解析における情報要求分析

*丸田 敦貴[1]、田貝 奈央[1]、加藤 誠[1] (1. 筑波大学)

17:20 ~ 17:45
[T3-B-3-04] 研究内容からの研究データ検索

*山﨑 基数[1]、冨浦 洋一[1]、清水 敏之[1] (1. 九州大学)

17:45 ~ 18:10
[T3-B-3-05] 大規模言語モデルによる文書補強とリランキングによる統計データ検索

*黒川 博生[1]、宮森 恒[1] (1. 京都産業大学 情報理工学部)

DEIM Forum 2024

鏡映を識別可能な3次元CADモデル検索のためのバイスペクトルの利用

小島優希也† 片山 薫†

† 東京都立大学システムデザイン研究科 〒191–0065 東京都日野市旭が丘 6-6
E-mail: †ojima-yukiya@ed.tmu.ac.jp, ††kaoru@tmu.ac.jp

あらまし 3次元ラドン変換と1次元・球面上のバイスペクトルを組み合わせることにより位相情報を保持した3次元モデルの特徴量を提案し，アニメーション付3次元モデルの検索に適用する．これまでも，モデルに対して3次元空間上のスペクトル解析によるパワースペクトルを並進・回転に不変な特徴量として利用してきたが，位相情報が欠如する．位相情報にもモデルの特徴が表現される可能性があるという観点からはパワースペクトルは不十分である．特に位相情報が欠如したパワースペクトルは鏡映なモデルを区別するこが出来ない問題が発生している．本研究では位相情報を保持するバイスペクトルによって鏡映を区別できるようにする．実験から，提案した特徴量では並進・回転に不変かつ，鏡映を区別することが可能であることがわかった．

キーワード バイスペクトル，3次元ラドン変換，球面調和展開，アニメーション付3次元モデル

1 はじめに

本稿では，位相情報を含んだ特徴量によって，アニメーション付3次元モデルを検索する手法を提案する．モデルの類似度は並進・回転に対して不変な特徴量を比較する手法とモデルの姿勢・位置を揃えて比較する手法がある．前者の手法では並進・回転に不変な特徴量を利用することで，姿勢を揃えることなく検索を行うことが出来るため，モデル全体が3次元空間内を自由に並進・回転するモデルに対しても有効である．後者の場合，位置決めについては重心を合わせることによって一定の効果はあるが，主軸分析によって姿勢を揃える手法は，主成分分析では共分散行列が2次の周波数のみに強い相関があることにより回転不変性が低いことや，軸の向きが定めることができないことなどから，姿勢を揃えることは難しい [1]．本稿の検索手法では前者の並進・回転に不変な特徴量を用いる．

先行研究では球面上のパワースペクトルを計算することで回転に対して不変な特徴量を得ていたが，パワースペクトルは位相情報を無視した値であるため，図1のような，ある平面において鏡映な2つのモデルを区別することができない．鏡映なモデルは自動車の左右のドアパネルなど数多く存在する．そこで，球面上のバイスペクトルを計算することで，回転不変かつ鏡映を区別することができる特徴量を求めることが可能である．本稿では，モデルの並進・回転に対する不変な特徴量を，3次元ラドン変換とバイスペクトルを用いて計算する手法提案し，提案した特徴量計算をアニメーション付3次元モデルに適用する方法を提案する．豊泉ら [2] は次のようにアニメーションモデル検索のフレームワークを構築した．アニメーションモデルとは，3次元静止モデルに対してアニメーションが付与されたものを指す．モデルに付与されたアニメーションにおけるある時刻をフレームと呼ぶ．検索の設定について説明する．クエリとして静止したモデルがひとつ，データベースとしてアニメーションを持ったモデルが複数与えられ，クエリと一致するモデルをデータベースから検索する．検索は以下の手順で行う．クエリモデルの特徴量とデータベースモデルとクエリモデルの特徴量をアニメーション内のフレーム毎に比較し，最も距離が小さくなったフレームでの距離をそのデータベースモデルとクエリモデルの距離とする．クエリモデルとデータベース内の各モデルの距離を比較し，最も距離の小さいデータベースモデルを検索結果として返す．

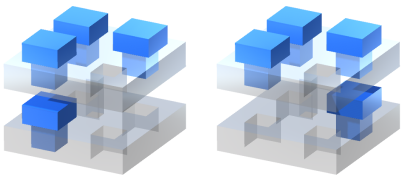


図1 形状が互いに鏡映なモデルの例

2 関連研究

3次元モデルを対象とした比較・検索手法については様々な研究が行われている．Iyer ら [3], Bustos ら [4], Tangelder ら [5], Li ら [6] は，3次元モデルにおける検索手法について比較・調査を行っている．アニメーションをもつ3次元モデルの検索手法については，モーションキャプチャ技術をはじめとする人型モデルを対象としたものを中心に様々な手法が提案されている．Muller ら [7] は，人体モデルのジョイント間の幾何学的関係を用いたインデックスを作成してモデル検索を行っている．Wang ら [8] は固有ベクトル相関に基づいてマッチングスコアを計算することで人型モデルのモーションリアルタイム検索を実現している．これらの手法はいずれも人型モデルを対象としたものであり，剛体から構成されるが前提である3次元モデルの検索には適用することができない．Kazhdan ら [1] 3次元モデルを同心球状にサンプリングし，各球面をそれぞれ球面調和変換し，球面上のパワースペクトルを計算することでモデルの回転に不変な特徴量を提案した．

3 前提知識

1 次元信号 $f:\mathbb{R}\to\mathbb{R}$ の場合，フーリエ変換は

$$F(\omega)=\int_{\mathbb{R}} f(x)e^{-j\omega x}\,dx$$

と書ける．パワースペクトルは $|F(\omega)|^2$ で与えられる．元の信号 f をシフトした信号 $f(x-a)$ のフーリエ係数は $F(\omega)e^{-ja\omega}$ となるが，$|F(\omega)e^{-ja\omega}|^2$ は指数の部分が 1 になるため，結果，パワースペクトルはシフト不変となる．よって，元の信号 $f(x)$ に対してある x_0 にて鏡映な信号 $h(x):=g(x_0-x)$ に対するパワースペクトル $|H(\omega)|^2$ は元の信号のパワースペクトル $F(\omega)e^{-ja\omega}$ と一致するため，パワースペクトルでは鏡映を区別することが不可能である．f が球面上の関数 $f:S^2\to\mathbb{R}$ の場合，フーリエ変換の基底は本稿では 3 次元空間での任意の回転をオイラー角 α,β,γ を z-y-z 系にて表現する．$R_z(\phi)$，$R_y(\theta)$ をそれぞれ z,y 軸まわりに ϕ,θ 回転の 3×3 行列は次のように定義される．

$$R_z(\phi):=\begin{bmatrix}\cos(\phi) & -\sin(\phi) & 0\\ \sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$R_y(\theta):=\begin{bmatrix}\cos(\theta) & 0 & \sin(\theta)\\ 0 & 1 & 0\\ -\sin(\theta) & 0 & \cos(\theta)\end{bmatrix}$$

よって，3 次元空間での任意の回転行列 R は R_z，R_y を用いて $R:=R_z(\gamma)R_y(\beta)R_z(\beta)$ と表される．α は経度で $0\leqq\alpha\leqq2\pi$ とし，β は余緯度で $0\leqq\beta\leqq\pi$ として，3 次元の単位ベクトルを $\boldsymbol{x}(\alpha,\beta):=[\cos(\alpha)\sin(\beta),\ \sin(\alpha)\sin(\beta),\ \cos(\beta)]^\top$ とする．z 軸上の単位ベクトルを $\boldsymbol{e}_3:=[0,0,1]^\top$ と定義する．$\boldsymbol{x}(\alpha,\beta)$ は $\boldsymbol{x}(\alpha,\beta)=R_z(\alpha)R_y(\beta)\boldsymbol{e}_3$ と書ける．球面上の関数 $g(\boldsymbol{x})$ は球面上の正規直交基底となる球面調和関数 $Y_l^m(\boldsymbol{x}),(l:\text{次数},m:\text{位数})$ によって以下のように L 次まで展開される．

$$f(\boldsymbol{x})=\sum_{l=0}^{L}\sum_{m=-l}^{l}F_l^mY_l^m(\boldsymbol{x})$$

球面調和関数 $Y_l^m:S^2\to\mathbb{C}$ はルジャンドル陪関数 P_l^m を用いて

$$Y_l^m(\boldsymbol{x}(\alpha,\beta)):=P_l^m(\cos(\beta))e^{-jm\alpha}$$

と定義される．球面調和展開係数 F_l^m は次の球面調和変換によって求められる．

$$F_l^m:=\int_{SO(3)}f(R\boldsymbol{e}_3)\overline{Y_l^m(R\boldsymbol{e}_3)}\,dR \tag{1}$$

$f:S^2\to\mathbb{R}$ の $SO(3)$ 上での積分は次のようにオイラー角 α,β,γ で積分できる．

$$\begin{aligned}&\int_{SO(3)}f(R\boldsymbol{e}_3)dR\\ &=\frac{1}{8\pi^2}\int_0^{2\pi}\int_0^{\pi}\int_0^{2\pi}f(R_z(\alpha)R_y(\beta)R_z(\gamma)\boldsymbol{e}_3)\sin(\beta)\,d\gamma d\beta d\alpha\\ &=\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}f(R_z(\alpha)R_y(\beta)\boldsymbol{e}_3)\sin(\beta)\,d\beta d\alpha\\ &=\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}f(\boldsymbol{x}(\alpha,\beta))\sin(\beta)\,d\beta d\alpha\end{aligned}$$

よって (1) 式の球面調和変換は

$$F_l^m=\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}f(\boldsymbol{x}(\alpha,\beta))\overline{Y_l^m(\boldsymbol{x}(\alpha,\beta))}\sin(\beta)\,d\beta d\alpha \tag{2}$$

と α,β の積分で計算される．l 次の展開係数 F_l^m を並べた $2l+1$ 次元の行ベクトルを $F_l:=\left[F_l^{-l},\ldots,F_l^0,\ldots,F_l^l\right]$ とする．球面上の関数 f のパワースペクトル $\|F_l\|^2$ は球面上の関数 $f(\boldsymbol{x})$ の回転に対して不変であることが知られている．ある $R\in SO(3)$ に対して，$g(\boldsymbol{x}):=f(R\boldsymbol{x})$ とする．g の球面調和展開係数を G とすると，回転行列と係数の間に

$$G_l=F_lD_l(R)$$

の関係が成り立つことが知られている．D_l はウィグナーの D 行列と呼ばれる $(2l+1)\times(2l+1)$ のユニタリ行列である．よって $\|G_l\|^2$ を計算すると位相成分 $D_l(R)$ を打ち消して，$\|F_l\|^2$ と等しくなる．しかし，パワースペクトルは位相情報が欠如しているため，元の関数に対して鏡映な関数を区別することが不可能である．パワースペクトルは回転不変であるが，位相成分を完全に無視した値であるので，鏡映を区別することができない．Kakarala [9] らは球面上の関数の位相を保持した量として，球面上のバイスペクトルを提案し，三次元空間上の点群データの回転不変かつ位相にセンシティブな特徴量を提案した．関数 f が 1 次元の信号の場合，バイスペクトルは 3 次自己相関を二次元フーリエ変換したものがバイスペクトルに相当する．$f(x):\mathbb{R}\to\mathbb{R}$ の三次自己相関 $T_f(t_1,t_2):\mathbb{R}\times\mathbb{R}\to\mathbb{C}$ は

$$T_f(t_1,t_2):=\int_{\mathbb{R}}\overline{f(x)}f(x+t_1)f(x+t_2)\,dx$$

と定義される．f のバイスペクトル $b_f(u,v)$ は

$$b_f(u,v):=\int\int_{\mathbb{R}}T_f(t_1,t_2)e^{-i(ut_1+vt_2)}\,dt_1dt_2$$

と定義される．f のフーリエ変換を F とすると右辺は以下のようにフーリエ係数の 3 つの積で書ける．

$$b_f(u,v)=F(u)F(v)\overline{F(u+v)} \tag{3}$$

f がシフトしても係数の 3 つの積によって位相成分が打ち消されるため，パワースペクトルと同様にバイスペクトルはシフト不変である．しかし，f に対して鏡映な信号 h のバイスペクトルは f のバイスペクトルの複素共役となるので，鏡映を区別することができる値となる．関数 f が球面上の関数の場合，$f:S^2\to\mathbb{R}$ の三次自己相関 $T_f:SO(3)\times SO(3)\to\mathbb{C}$ は

$$T_f(S_1,S_2) := \int_{SO(3)} \overline{f(R\boldsymbol{e}_3)} f(RS_1\boldsymbol{e}_3) f(RS_2\boldsymbol{e}_3) dR$$

と定義される．ここで，l 次の球面調和関数を並べた $(2l+1)$ 次元の行ベクトル $Y_l := \left[Y_l^{-l}, \ldots, Y_l^0, \ldots, Y_l^l\right]$ を定義する．f のバイスペクトルは球面調和関数を基底として三次自己相関 $T_f(S_1,S_2)$ を 2 次元フーリエ変換したものであるので，

$$b_{k,l}^f := \int\int_{SO(3)} T_f(S_1,S_2) \left[\overline{Y_k(S_1\boldsymbol{e}_3)} \otimes \overline{Y_k(S_2\boldsymbol{e}_3)}\right] dS_1 dS_2$$

と定義される．$\otimes$ はテンソル積であるから，$b_{k,l}^f$ は $(2k+1)(2l+1)$ 次元の行ベクトルである．$T_f(S_1,S_2)$ を代入して，右辺を変形すると，

$$b_{k,l}^f = [F_k \otimes F_l]\, C_{k,l} \left[\bigoplus_{i=|k-l|}^{k+l} \int_{SO(3)} \overline{f(R)} D_i(R)\, dR\right] C_{k,l}^\dagger. \tag{4}$$

積分部分は計算すると中央の列のみが非零になる．$C_{k,l}$ は $(2k+1)(2l+1)\times(2k+1)(2l+1)$ のクレプシュゴルダン行列であり，直交行列である．i 次ウイグナーの D 行列の中央の列は球面調和関数の列 $Y_i^\top$ なので，O_i を $(2i+1)\times i$ の零行列とすると (4) 式は

$$b_{k,l}^f := [F_k \otimes F_l]\, C_{k,l} \left[\bigoplus_{i=|k-\ell|}^{k+l} \left[O_i,\ F_i^\dagger,\ O_i\right]\right] C_{k,l}^\dagger \tag{5}$$

と書ける．ゆえに，球面上のバイスペクトルは球面調和展開係数とクレプシュゴルダン行列によって計算することができる．このバイスペクトルは次の 2 つの性質を持つ．

(a) $R \in SO(3)$ について $g(\boldsymbol{x}) := f(R\boldsymbol{x})$ とすると，

$$b_{k,l}^g = b_{k,l}^f \tag{6}$$

(b) $\det(S) = -1$ となる $S \in R^{3\times3}$ について $h(\boldsymbol{x}) := f(S\boldsymbol{x})$ とすると，

$$b_{k,l}^h = \overline{b_{k,l}^f} \tag{7}$$

(6) 式を証明する．g の球面調和展開係数ベクトルを G_l とすると，g のバイスペクトルは次のように書ける．

$$\begin{aligned} b_{k,l}^g &= [G_k \otimes G_l]\, C_{k,l} \left[\bigoplus_{i=|k-\ell|}^{k+l} \left[O_i,\ G_i^\dagger,\ O_i\right]\right] C_{k,l}^\dagger \\ &= [F_k D_k(R) \otimes F_l D_l(R)]\, C_{k,l} \\ &\qquad \left[\bigoplus_{i=|k-\ell|}^{k+l} \left[O_i,\ D_i(R)^\dagger F_i^\dagger,\ O_i\right]\right] C_{k,l}^\dagger \\ &= [F_k \otimes F_l]\, [D_k(R) \otimes D_l(R)]\, C_{k,l} \\ &\qquad \left[\bigoplus_{i=|k-\ell|}^{k+l} \left[O_i,\ D_i(R)^\dagger F_i^\dagger,\ O_i\right]\right] C_{k,l}^\dagger \end{aligned}$$

$D_k(R) \otimes D_l(R)$ は次のように，ブロック対角化されることが知られている．

Algorithm 1 ComputeDistance(Model1, Model2, ...)

Input: 静止モデル Model1, Model2, その他の 5 つのパラメータ Φ, Θ, R, N, L の詳細は，Algorithm3,4,5,7,8 に記載

Output: Model1 と Model2 の距離

1: (Feature1, $\mathcal{N}, \mathcal{L}$) ← ComputeFeature(Model1, Φ, Θ, R, N, L)
2: (Feature2, $\mathcal{N}, \mathcal{L}$) ← ComputeFeature(Model2, Φ, Θ, R, N, L)
3: **return** $\sum_{n\in\mathcal{N}} \sum_{l\in\mathcal{L}} \|\text{Feature1}_n(l) - \text{Feature2}_n(l)\|$

Algorithm 2 ComputeFeature(Model, ...)

Input: 1 つの静止モデル Model, その他の 5 パラメータ Φ, Θ, R, N, L は，Algorithm3,4,5,7,8 に記載される

Output: Model の特徴量 Feature

1: V ← 3DRadonFourierTransform(Model, Φ, Θ, R, N)
2: $(s, \mathcal{N})$ ← { Bispectrum(V, Φ, Θ, N) / PowerSpectrum(V, Φ, Θ, N) }
3: F ← SphericalHarmonicTransform(s, $\mathcal{N}$, Φ, Θ, L)
4: (Feature, $\mathcal{L}$) ← { BispectrumOnSphere($s, \mathcal{N}, L$) / PowerSpectrumOnSphere($s, \mathcal{N}, L$) }
5: **return** (Feature, $\mathcal{N}, \mathcal{L}$)

$$D_k(R) \otimes D_l(R) = C_{k,l} \left[\bigoplus_{i=|k-\ell|}^{k+l} D_i(R)\right] C_{k,l}^\dagger$$

したがって，$b_{k,l}^g$ は f のバイスペクトル $b_{k,l}^g$ と等しくなることがわかる．

(7) 式を証明する．$b_{k,l}^f$ は鏡映の場合，複素共役になることから鏡映を区別することができる．$\det(S) = +1$ となる 3×3 行列 S について，$S_y = \mathrm{diag}[1,-1,1]$ とすると，$S_y S$ は行列式が $+1$ となり，$SO(3)$ に属する．ゆえに，$g(\boldsymbol{x}) := f(S_y S\boldsymbol{x})$ とすると回転不変性から $b_{k,l}^g = b_{k,l}^f$ が成り立つ．$h(\boldsymbol{x}) := g(S_y\boldsymbol{x}) = f(S\boldsymbol{x})$ とすると，$h(\boldsymbol{x}(\alpha,\beta)) = g(\boldsymbol{x}(-\alpha,\beta))$ となる．(2) 式での球面調和変換では g, h が実数の値をとるので，$H_l^m = \overline{G_l^m}$ が成り立つ．よって，クレプシュゴルダン行列 $C_{k,l}$ は実数行列であるから，$b_{k,l}^h = \overline{b_{k,l}^g}$ が成り立つ．ゆえに，$b_{k,l}^h = \overline{b_{k,l}^f}$．Kakarala [9] らは 3 次元空間上の点群データに対して球面上のバイスペクトル，パワースペクトルを適用し実験的評価し，バイスペクトルが鏡映な点群データを区別することができることを示した．しかし，ポリゴンで構成される 3D CAD モデルを球面として扱うには工夫が必要であり，直接用いることができない．本稿では，ポリゴンで構成され 3 次元モデルに 3 次元ラドン変換と組み合わせてバイスペクトルを適用する手法を提案する．

4 提案手法

特徴量計算の流れを図 2 に示す．2 つのモデルの相違度は ComputeDistance(Algorithm1) にて計算される．以後，相違度は距離と呼ぶ．距離計算は特徴量の距離の各次元での総和である．1 つの静止モデルの特徴量は ComputeFeature(Algorithm2) にて計算される．正の整数 K, J を用いて経度と余緯度の集合を定義する．Φ を経度の集合とし $\{2\pi k/K \mid k = 0, 1, \ldots, K\}$ とする．Θ を余緯度の集合とし，$\{\theta \mid 0 \leqq \theta \leqq \pi$にて $\cos(\theta)$ が J 次

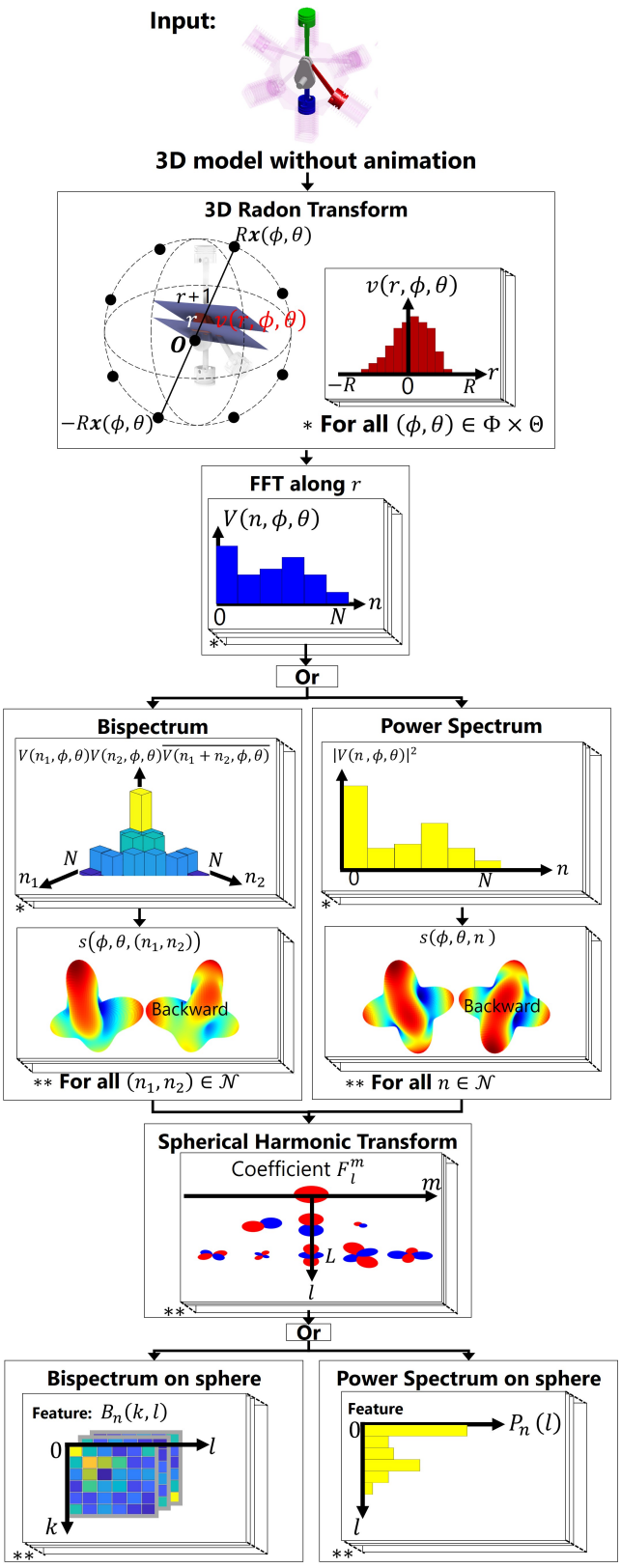


図 2　3 次元モデルの特徴量を得る流れ.

ルジャンドル多項式のゼロ点となるθ} とする．Model の重心が原点に合うように Model を定位させる．3 次元ラドン変換を行う範囲 R は Model の半径より十分に大きくし，アニメーションが動いても範囲からはみ出ないように R を十分に大きくする必要がある．フーリエ係数の最高次数 N は R より小さいことが条件である．3DRadonFourierTransform (Algorithm3) では 3 次元ラドン変換および各視点におけるフーリエ係数を計算する．r を $-R \leqq r \leqq R$ を満たす整数とし，法線ベクトルが $\boldsymbol{x}$ で，点 $r\boldsymbol{x}$ を通る平面を $\Pi(r,\boldsymbol{x})$ とする．$v(r,\phi,\theta)$ は平面 $\Pi(r+1,\boldsymbol{x}(\phi,\theta))$ と平面 $\Pi(r,\boldsymbol{x}(\phi,\theta))$ によって Model を挟んだ部分の体積である．$v(r,\phi,\theta)$ を r に沿って次のように 1 次元離散フーリエ変換して整数 n を周波数として係数 $V(n,\phi,\theta)$ を得る．

$$V(n,\phi,\theta) = \sum_{r=-R}^{R} v(r,\phi,\theta) \exp\left(-j\,\frac{2\pi n r}{2R+1}\right) \tag{8}$$

Bispectrum (Algorithm4) または PowerSpectrum (Algorithm5) にて $V(n,\phi,\theta)$ を基に 1 次元上のスペクトル解析を行い並進不変な特徴量を求め, 周波数の集合 $\mathcal{N}$ を定義する．特徴量は各周波数 $n \in \mathcal{N}$ について球面上に分布する $s(\phi,\theta,n)$ である．Bispectrum では $\mathcal{N}$ を $0 \leqq n_1 \leqq N$ かつ $0 \leqq n_2 \leqq N - n_1$ を満たす全ての整数の組 (n_1,n_2) の集合と定義し，各 $(n_1,n_2) \in \mathcal{N}$ についてバイスペクトル $V(n_1,\phi,\theta)V(n_2,\phi,\theta)\overline{V(n_1+n_2,\phi,\theta)}$ を計算する．Model の並進により，$v(r,\phi,\theta)$ がシフトするとフーリエ係数 $V(n)$ の位相成分が変化するがバイスペクトル $V(n_1,\phi,\theta)V(n_2,\phi,\theta)\overline{V(n_1+n_2,\phi,\theta)}$ は位相変化を打ち消すため，Model の並進についてバイスペクトルは不変である．また，逆視点 $x(\phi+\pi,\pi-\theta)$ にて体積を計算した場合，元の視点の反射の波になる．このとき，バイスペクトルは元の波に対して複素共役となる．ゆえに，$v(r,\phi,\theta)$ のバイスペクトル $V(n_1,\phi,\theta)V(n_2)\overline{V(n_1+n_2,\phi,\theta)}$ の実部と虚部の和を $s(\phi,\theta,(n_1,n_2))$ とすることで，$s(\phi,\theta,(n_1,n_2))$ は $s(\phi+\pi,\pi-\theta,(n_1,n_2))$ と異なる値を持ち，かつ並進不変となる．PowerSpectrum ではフーリエ周波数の集合を $\mathcal{N} := \{0,1,\ldots,N\}$ とし，各 $n \in \mathcal{N}$ に関してパワースペクトル $|V(n,\phi,\theta)|^2$ を計算する．パワースペクトルは並進について不変であるが，逆視点 $x(\phi+\pi,\pi-\theta)$ にて体積を計算した場合，v は元の視点でのフーリエ係数 $V(n,\phi,\theta)$ に対して位相のみが変化するため，両側の視点が等しくなる．つまり，$|V(n,\phi,\theta)|^2$ は $|V(n,\phi+\pi,\pi-\theta)|^2$ と等しくなる．よって，パワースペクトルは球面上において情報量が半分になるため，鏡映なモデルを区別するのに十分な情報を保有していないことがわかる．SphericalHarmonicTransform (Algorithm6) では各周波数 $n \in \mathcal{N}$ における球面上の関数 $s(\phi,\theta,n)$ に対して球面調和展開係数を計算する．(2) 式の球面調和変換は離散球面 $f : \Phi \times \Theta \to \mathbb{R}$ については以下の式で計算される．

$$F_l^m := \frac{1}{2}\sum_{\theta\in\Theta} w_l(\theta)\left(\frac{1}{|\Phi|}\sum_{\phi\in\Phi} f(\phi,\theta)\, e^{-jm\phi}\right) P_l^m(\cos(\theta)) \tag{9}$$

Algorithm 3 3DRadonFourierTransform(Model, Φ, Θ, R, N)

Input: 静止モデル Model, 経度の集合 Φ, 余緯度の集合 Θ, ラドン変換を行う半径 R, フーリエ係数の最大周波数 N

Output: Model の球面上のフーリエ係数 V. V は $|\Phi|\times|\Theta|\times(2N+1)$ 行列

```
for all (φ,θ) ∈ Φ × Θ do
    for r = −R to R do
        v(r,φ,θ) ← 平面 Π(r, x(φ,θ)) と平面 Π(r+1, x(φ,θ)) によって Model が挟まれた部分の体積
    end for
    for n = 0 to N do
        V(n,φ,θ) ← v(r,φ,θ) を r に沿って (8) 式にてフーリエ変換して得られた係数
    end for
end for
return V
```

$w_l(\theta)$ はガウス重みといい，以下の式である．

$$w_l(\theta) := \frac{2\sqrt{(2l-1)(2l+1)}}{lP_{l-1}(\cos(\theta))P_l'(\cos(\theta))} \tag{10}$$

BispectrumOnSphere (Algorithm7) または PowerSpectrum (Algorithm8) にて球面調和展開係数のベクトル $F_n(l)$ を基に球面上のスペクトル解析を行い Model の回転に不変な特徴量を求める．特徴量は各周波数 $n \in \mathcal{N}$ における球面上のスペクトルである．BispectrumOnSphere では球面上のバイスペクトルを計算する．球面上のバイスペクトルの定義式である (5) 式のブロック対角行列内の零行列 O_i と最後の因数 $C_{k,l}^{\dagger}$ は回転不変性と鏡映に対する性質へ影響を与えないことから以下の式を使う．

$$B_{k,l}^{f} = [F_k \otimes F_l]\, C_{k,l} \left[\bigoplus_{i=|k-\ell|}^{k+l} F_i^{\dagger} \right] \tag{11}$$

$B_{k,l}^{f}$ は $(k+l-|k-l|+1)$ 次元の行ベクトルである．PowerSpectrumOnSphere (Algorithm8) では球面上のパワースペクトル $\|F_n(l)\|^2$ を計算する．ComputeFeature は { Bispectrum, PowerSpectrum } からいずれかを選択し，{ BispectrumOnSphere, PowerSpectrumOnSphere } からいずれかを選択する．ComputeDistance は入力の 2 モデルに対してそれぞれ ComputeFeature を計算し，得られた特徴量の差の絶対値の総和を出力する．ただし，特徴量計算においてアルゴリズムの選択が同じである場合のみ距離が計算される．

5 実験的評価

並進不変な特徴量計算手法については Bispectrum (Algorithm4) と PowerSpectrum (Algorithm5) の 2 手法を提案した．いずれかの手法によって計算された並進不変な特徴量を入力として，回転不変な特徴量計算手法について BispectrumOnSphere (Algorithm7) と PowerSpectrum (Algorithm8) の 2 手法を提案した．これら 4 つの手法を組み合わせて比較評価を行う．

Algorithm 4 Bispectrum(V, N, Φ, Θ)

Input: Algorithm3 の出力 V, フーリエ係数の最大周波数となる N, 経度の集合 Φ, 余緯度の集合 Θ

Output: 第 1 出力は Model の球面上の特徴量 $s(|\Phi|\times|\Theta|\times\mathcal{N}$ 行列), 第 2 出力は周波数の集合 $\mathcal{N}$

```
N ← 0 ≦ n1 ≦ N かつ 0 ≦ n2 ≦ N − n1 を満たす全ての整数の組 (n1, n2) の集合
for all (φ,θ) ∈ Φ × Θ do
    for n1 = 0 to N do
        for n2 = 0 to N − n1 do
            s(φ,θ,(n1,n2)) ← V(n1,φ,θ)V(n1,φ,θ)conj(V(n1+n2,φ,θ)) の実部と虚部の和
        end for
    end for
end for
return (s, N)
```

Algorithm 5 PowerSpectrum(V, N, Φ, Θ)

Input: Algorithm3 の出力 V, フーリエ係数の最大周波数となる N, 経度の集合 Φ, 余緯度の集合 Θ

Output: 第 1 出力は球面上の量 s ($|\Phi|\times|\Theta|\times\mathcal{N}$ 行列), 第 2 出力は周波数の集合 $\mathcal{N}$

```
N ← {0, 1, ..., N}
for all (φ,θ) ∈ Φ × Θ do
    for n = 0 to N do
        s(φ,θ,n) ← |V(n,φ,θ)|^2
    end for
end for
return (s, N)
```

Algorithm 6 SphericalHarmonicTransform(s, $\mathcal{N}$, Φ, Θ, L)

Input: Algorithm4 または Algorithm5 の第 1 出力 s, Algorithm4 または Algorithm5 の第 2 出力 $\mathcal{N}$, 経度の集合 Φ, 余緯度の集合 Θ, 球面調和変換の最高次数 L

Output: $\mathcal{N}\times L\times L$ 行列である球面調和係数 F (F の各要素 F_l^n は l 次元の行ベクトル)

```
for all n ∈ N do
    for l = 0 to L do
        for m = −l to l do
            F_l^m ← (9) 式に基づいて求めた次数 l, 位数 m の球面調和関数に対応する f の展開係数. f は (φ,θ) ∈ Φ × Θ にて s(φ,θ,n) の値をとる関数
        end for
        F_n(l) ← [F_l^{−l}, ..., F_l^0, ..., F_l^l]
    end for
end for
return F
```

データベースには V6 engine, Radial engine の 2 種類のモデルが存在し，クランクシャフトが一周するアニメーションが付与されている．アニメーションは 1 週期を 12 等分割し，1 フレーム目から 1 つおきの 6 フレームをデータベースへ与え，2 フレーム目から 1 つおきの 6 フレームをクエリとして与えた．こ

表 1　ComputeFeature (Algorithm2) の平均処理時間 (秒) の内訳．Radon+Fourier, B, P, SHT, BS, PS はそれぞれ 3DRadonFourierTransform (Algorithm3), Bispectrum (Algorithm4), PowerSpectrum (Algorithm5), SphericalHarmonicTransform (Algorithm6), BispectrumOnSphere (Algorithm7), PowerSpectrumOnSphere (Algorithm8) の略である．

$\lvert\Phi\times\Theta\rvert$				4×8				8×16				16×32			
Radon+Fourier							23.986				84.349				311.091
B		P			0.013		0.009		0.024		0.012		0.071		0.028
SHT					0.056		0.006		0.115		0.012		0.386		0.040
BS	PS	BS	PS	0.452	0.018	0.057	0.003	1.786	0.034	0.206	0.005	31.681	0.071	4.546	0.010
Total				24.506	24.072	24.058	24.004	86.273	84.521	84.579	84.377	343.229	311.619	315.705	311.169

Algorithm 7 BispectrumOnSphere(F, $\mathcal{N}$, L)

Input: Algorithm6 の出力 F, Algorithm4 または Algorithm5 の第 2 出力 $\mathcal{N}$, 球面調和変換の最高次数 L

Output: 第 1 出力は $\mathcal{N}\times(L+1)\times(L+1)$ 行列の球面上のバイスペクトル B (B の各要素 $B_n(k,l)$ は $(k+l-|k-l|+1)$ 次元の行ベクトル), 第 2 出力はバイスペクトルの次数の直積集合 $\mathcal{L}$

$\mathcal{L} \leftarrow 0 \leqq k \leqq L$ かつ $0 \leqq l \leqq L$ かつ $k+l \leqq L$ を満たす整数の組 (k,l) の集合
for $n \in \mathcal{N}$ **do**
  for all $(k,l) \in \mathcal{L}$ **do**
    $B_n(k,l) \leftarrow [F_n(k) \otimes F_n(l)]\, C_{k,l} \left[\bigoplus_{i=|k-\ell|}^{k+l} F_n(i)^\dagger \right]$
  end for
end for
return $(B, \mathcal{L})$

Algorithm 8 PowerSpectrumOnSphere(F, $\mathcal{N}$, L)

Input: Algorithm6 の出力 F, Algorithm4 または Algorithm5 の第 2 出力 $\mathcal{N}$, 球面調和変換の最高次数 L

Output: 第 1 出力は $\mathcal{N}\times L$ 行列の球面上のパワースペクトル P (P の各要素 $P_n(l)$ は $l+1$ 次元の行ベクトル), 第 2 出力はパワースペクトルの次数の集合 $\mathcal{L}$

for $n \in \mathcal{N}$ **do**
  for $l = 0$ **to** L **do**
    $P_n(l) \leftarrow \|F_n(l)\|^2$
  end for
end for
$\mathcal{L} \leftarrow \{0, 1, \ldots, L\}$
return $(P, \mathcal{L})$

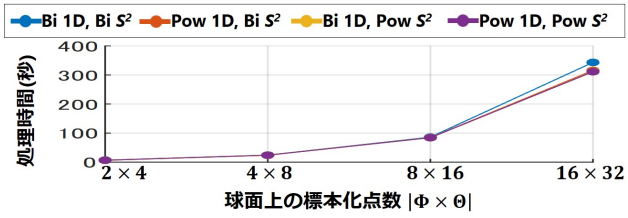


図 3　ComputeFeature (Algorithm2) の平均処理時間 (秒)

の 2 種類のモデルはクエリモデル・データベースモデルの両方に利用される．データベースには V6 engine, Radial engine の 2 種類のモデルがあるが，特徴量が鏡映を区別する性質を検証するため，各モデルを yz 平面に関して鏡映なモデル V6 engine mirror, radial engine mirror をデータベースベースに追加した (図 5)．以後，V6 engine mirror, radial engine mirror をまとめて鏡映データベースモデルと呼ぶ．特徴量が並進・回転について不変であるという性質を検証するため，ComputeFeature を行う前にランダムな位置・姿勢にモデルを定位させる．表 3 に ComputeDistance(Algorithm1) の入力パラメータを示す．ラドン変換での入力パラメータ R はモデルがランダムの位置・姿勢にてアニメーションした時, 原点を中心とする半径 R の球面内収まるように十分に大きい値 400 を常に指定した．クエリモデルは V6 engine の各フレームと Radial engine の各フレームから 1 つのフレームを与えられる．ComputeDistance は 10 回計算される．実験は，MATLAB R2023b にてプログラムを実装し，Intel Core i9 13900F 2.0 GHz CPU 32GB RAM, Windows11 Home を搭載したコンピュータ上で行った．

表 2　各手法の正答率．B, P, BS, PS はそれぞれ Bispectrum (Algorithm4), PowerSpectrum (Algorithm5), BispectrumOnSphere (Algorithm7), PowerSpectrumOnSphere (Algorithm8) の略である．

手法	B, BS	P, BS	B, PS	P, PS
正答率	1	0.65	0.62	0.66

表 3　提案手法の入力パラメータ

球面上の標本化点数 $\lvert\Phi\times\Theta\rvert$	32	128	512
経度方向の標本化点数 $\lvert\Phi\rvert$	4	8	16
余緯度方向の標本化点数 $\lvert\Theta\rvert$	8	16	32
球面調和展開の最高次数 L	4	8	16
フーリエ係数の最高周波数 N	10	10	10
ラドン変換の半径 R	400	400	400

6　結果・考察

6.1　距　　離

入力パラメータを表 3 にて $|\Phi\times\Theta|$ が 512 とした時のクエリモデルとデータベースモデルのフレームとの距離の平均を図 4 に示す．結果から，Bispectrum と BispectrumOnSphere の組み合わせのみが，クエリと同じデータベースモデルとクエリとの間の最大距離が，クエリとは異なるデータベースモデルとクエリとの間の最小距離よりも常に短いことが示された．その他の組合せではクエリと同じデータベースモデルとクエリとの間の最大距離が，クエリとは異なりかつ鏡映ではないデータベースモデルとクエリとの間の距離よりも短いことが示された．し

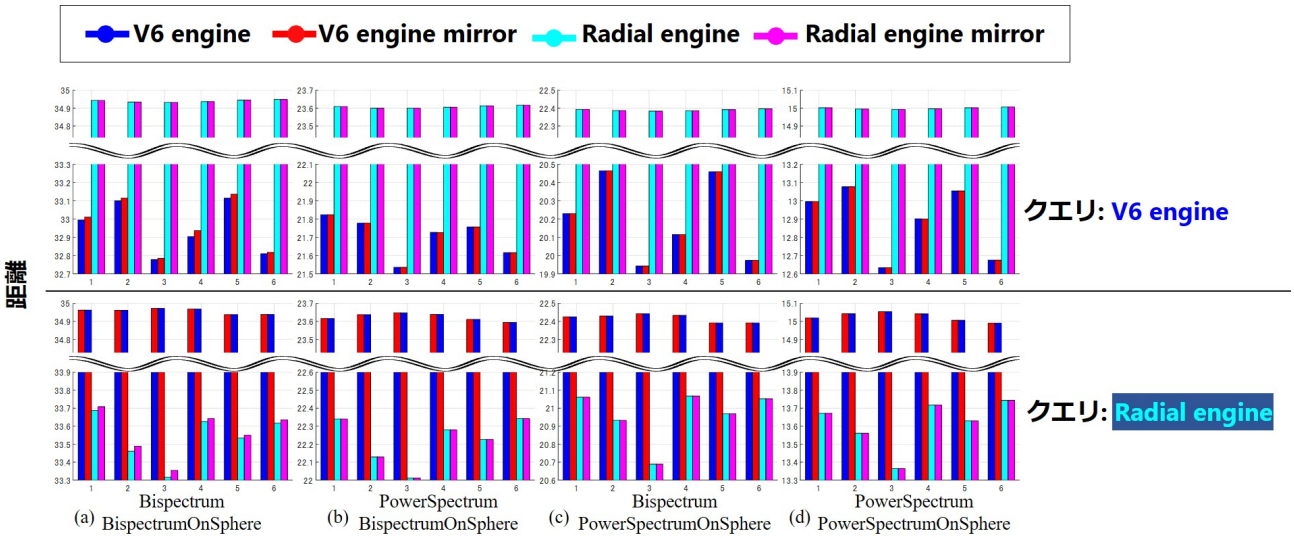


図 4　ComputeDistance (Algorithm1) の結果．入力パラメータは表 3 の $|\Phi \times \Theta|$ が 512 の場合である．各グラフの上段は V6 engine をクエリとした場合，下段は Radial engine をクエリとした場合である．横軸はクエリモデルとするフレームの番号である．クエリとデータベースモデルが名前が同じであるときは最大距離，クエリとデータベースモデルとの名前が異なるときは最小距離を示してる．

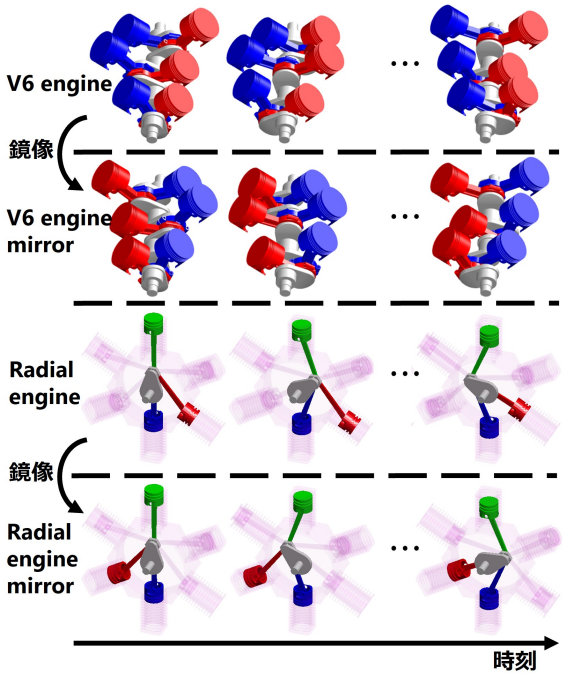


図 5　データベースモデル．Radial engine の紅紫色の部分は削除されている．その他の色は視覚的に見やすくするためであり，特に意味は無い

かし，その他の組合せではクエリと同じデータベースモデルとクエリとの間の最大距離が，クエリに鏡映なデータベースモデルとクエリとの間の最小距離よりも長くなることがあった．ゆえに，Bispectrum と BispectrumOnSphere の組合せのみが鏡映を区別することができると見られる．これはモデルを 3 次元空間上のスペクトル解析にて鏡映が位相にて表現されることと，バイスペクトルが位相情報を保持する性質が合わさった結果であると考えられる．

6.2　処 理 時 間

図 3 に ComputeFeature (Algorithm2) の平均処理時間 (秒) を示す．図 3 から球面上の標本化点数 $|\Phi \times \Theta|$ が増加すると，特徴量計算の手法の全ての組合せにおいて処理時間が増加することが分かる．表 1 に ComputeFeature (Algorithm2) の平均処理時間 (秒) の内訳を示す．BispectrumOnSphere, PowerSpectrumOnSphere は Bispectrum または PowerSpectrum のいずれかの第 2 出力であるフーリエ周波数の集合 $\mathcal{N}$ を入力に用いるため，Bispectrum, PowerSpectrum の選択に依存して処理時間が異なる．各球面上の標本化点数 $|\Phi \times \Theta|$ において 3DRadonFourierTransform (Algorithm3) が最も処理時間が長く，次に長いのは BispectrumOnSphere (Algorithm7) である．また，球面上の標本化点数 $|\Phi \times \Theta|$ が増加すると 3DRadonFourierTransform (Algorithm3), BispectrumOnSphere (Algorithm7) が順番に増加する．

7　今後の課題

本稿では，3 次元モデルの 3 次元空間スペクトル特徴量を提案したが以下の課題が残るバイスペクトルはパワースペクトルと比較して計算時間が増加する．空間スペクトルのどの周波数領域にモデルの特徴が反映されるかが分からない．完全一致検索では，空間スペクトルの比較で十分であるが，相違点の検出は不可能である．しかし，バイスペクトルは位相情報を含むため，元の波形を復元することができることから，バイスペクトル特徴量から入力の 3 次元モデルを復元することが可能である．よって，バイスペクトル特徴量の差からモデルの相緯度を再現する可能性がある．本稿では復元の話は省く．1 次元上，球面上において低い周波数領域での比較を行っているが，周波数領域の幅を決定する根拠は不明である．提案手法では，モデルのスケール不変性は実現していない．

8 おわりに

アニメーション付き 3 次元モデルをバイスペクトルによる位相情報を保持した特徴量での検索手法を提案した．バイスペクトルは位相情報を保持しているため，実験では鏡映なモデルを区別することが分かった．並進不変量を計算する 2 手法と並進不変量から回転不変量を計算する 2 手法を組み合わせて実験的に評価した結果，並進不変量・並進不変量による回転不変量の両者にバイスペクトルを適用することで，鏡映なモデルを区別する特徴量が得られることが分かった．今後は計算時間の削減が課題となる．

文　献

[1] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Symposium on Geometry Processing*, June 2003.

[2] 豊泉 翔太朗, 平島 栄志, 片山 薫, アセンブリ構造をもったアニメーション CAD モデルのレイアウト情報を考慮した検索, DEIM2022, A43-3.

[3] Natraj Iyer, Subramaniam Jayanti, Kuiyang Lou, Yagnanarayanan Kalyanaraman, and Karthik Ramani. Three-dimensional shape searching: state-of-the-art review and future trends. *Computer-Aided Design*, Vol. 37, No. 5, pp. 509–530, 2005. Geometric Modeling and Processing 2004.

[4] Benjamin Bustos, Daniel A Keim, Dietmar Saupe, Tobias Schreck, and Dejan V Vranić. Feature-based similarity search in 3d object databases. *ACM Computing Surveys (CSUR)*, Vol. 37, No. 4, pp. 345–387, 2005.

[5] J.W.H. Tangelder and R.C. Veltkamp. A survey of content based 3d shape retrieval methods. In *Proceedings Shape Modeling Applications, 2004.*, pp. 145–156, 2004.

[6] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, Hongbo Fu, Takahiko Furuya, Haisheng Li, Jianzhuang Liu, Henry Johan, Ryuichi Kosaka, Hitoshi Koyanagi, Ryutarou Ohbuchi, Atsushi Tatsuma, Yajuan Wan, Chaoli Zhang, and Changqing Zou. A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, Vol. 131, pp. 1–27, 2015. Special section: Large Scale Data-Driven Evaluation in Computer Vision.

[7] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, Vol. 24, No. 3, p. 677–685, jul 2005.

[8] Pengjie Wang, Rynson W.H. Lau, Zhigeng Pan, Jiang Wang, and Haiyu Song. An eigen-based motion retrieval method for real-time animation. *Computers & Graphics*, Vol. 38, pp. 255–267, 2014.

[9] Ramakrishna Kakarala and Dansheng Mao. A theory of phase-sensitive rotation invariance with spherical harmonic and moment-based representations. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 105–112, 2010.

流通を考慮した研究データのメタデータ付与支援

篠原　桐†　清水　敏之†　冨浦　洋一†

† 九州大学大学院統合新領域学府ライブラリーサイエンス専攻　〒 819–0395 福岡市西区元岡 744 番地
E-mail: †shinohara.kiri.548@s.kyushu-u.ac.jp,
††{shimizu.toshiyuki.457,tomiura.yoichi.802}@m.kyushu-u.ac.jp

あらまし　研究データが利用されるためには，利用者がデータを検索するに使用する情報がメタデータに含まれている必要がある．しかし，データ作成者が検索に対して有用なメタデータを付与することは容易ではない．本研究では，生成 AI を活用してデータ作成者に対してメタデータの付与を支援できる可能性に焦点を当て，その有効性を議論した．生成 AI のモデルとして ChatGPT を利用した．データ論文に対し，ChatGPT に，データ名およびデータに関する説明を与え，そこに含まれていない新たなキーワードを 10 個提案させ，そのキーワードに対して主観評価と客観評価を行なった．これらの評価により，生成 AI を用いた研究データにおけるメタデータ付与の支援が有効である可能性が示唆された．

キーワード　ChatGPT　研究データ　メタデータ　検索支援

1 はじめに

研究データとは，研究のために収集・作成・観測したデータで，研究の成果である論文やスライドの根拠となるデータ，及び研究成果そのものであるデータの両方を含むもの [1] である．近年，オープンサイエンスの高まりにより，研究データの公開が促進されており，その利活用によるイノベーション創出が期待されている [2]．Scopus[1] に掲載されているデータ論文の掲載数では，2014 年では 2 件だったのに対し，2023 年では累計で 7341 件とその数を大きく伸ばしている．また，Web of Science の Data Citation Index[2] によると，毎年 100 万件近くものデータセットが登録されている．

公開されている研究データを利用するためには，利用者が研究データを検索し，それを発見する必要がある．研究データは，内容が数値情報であったり，圧縮されていたりする場合も多く，研究データそのものが検索の対象となることはほとんどない．研究データに付与されているメタデータ [3] であったり，配布されている Web ページの記述内容を対象として検索が行われる．研究データが広く利用されるためには，研究データが作成された分野とは専門分野の異なる利用者からも検索される必要があるが，専門としていない分野の検索クエリを思いつくことは困難である．そのため，研究データの流通性を高めるためには，メタデータを付与する際に，専門としていない利用者が検索しやすいような流通を考慮したメタデータを付与する必要がある．一般的に，研究データに付与するメタデータの登録は，その研究データの作成者が行う．しかし，研究データの作成者はメタデータの専門家ではないため，作成者によるメタデータだけでは，記述が不十分であったり，必要なキーワードが欠けていたりなどして，有効な検索が行えない可能性がある [4], [5].

また，データアーキテクトのようなメタデータの専門家が付与するといった場合も考えられる．しかし，図書館員は研究データが作成された分野の専門家ではないため，研究内容に踏み込んだ利用法を考えつくことは困難である．さらに，データアーキテクトのような専門家の人材は不足 [6] しており，確保することが困難といった問題もある．

そこで，本研究では生成 AI を利用することで研究データの作成者が研究データにメタデータを付与する際に，流通することを考慮したメタデータの付与の支援ができるのではないかと考え，その可能性について調査を行った．

2 関連研究

2.1 先行研究

データセット検索のシステムには，Google Dataset Search などが存在する [7]．研究データの検索には，研究データに付与されているメタデータに対するキーワード検索によって行われる．メタデータとは，タイトル，作成者，出版年など，そのコンテンツ自身に関する情報を記述しているデータである [8].

研究データにメタデータとしてキーワードを推薦する既存の研究として，Tuarob らの研究 [5] がある．この研究では，キーワード情報の欠けている研究データに対して，研究データに付与されている説明文と，すでに付与されているキーワードの類似性を計算し，類似の既存データに関連付けられたキーワードを推薦する手法を提案している．しかし，この手法は既存データに依存しており，利用可能なメタデータを持つ既存の研究データの数や，含まれている単語，付与されているメタデータの品質に依存している．例えば，メタデータを追加したい研究データに類似した研究データのキーワード情報が不十分な場合，有効なキーワードを推薦できない．

この問題を解決する手法として，石田らが行った研究 [9] がある．この研究の手法では，既存のメタデータの情報の代わり

1：https://www.scopus.com/
2：https://www.webofscience.com/wos/drci/basic-search

に，キーワードごとに設けられた定義文と，メタデータを付与する対象の説明文の類似性を計算し，キーワードを提案する手法である．この手法を用いることで，類似する既存データの情報が不足しているなどの場合であってもキーワードの推薦が行える．しかし，全ての統制語彙にキーワード定義がされているわけではなく，キーワード定義が含まれていない統制語彙に対しては，キーワード定義を新たに作成する必要があると述べている．

検索者のことを考えたキーワードの提案における研究として，笹治らが行った研究 [10] がある．これは，データ構造などを解析することで，元々ついていなかったメタデータを自動的に追加することで検索を支援するものである．付与されるメタデータが拡張され，検索される可能性が高まるとは考えられるが，この手法では，研究データの利用法までは考慮できていない．

また，利用法を考慮したメタデータの研究には，角掛らの研究 [11] がある．この研究は，研究データを引用している引用論文から，引用文脈を利用することによりメタデータを拡張するものである．この研究では，引用文脈を利用することにより，研究データそのものの特徴や用途などを考慮したメタデータを拡張することができるのではないかと提案している．しかし，この手法でメタデータを付与する場合には，引用論文を集める必要がある．また，研究データに引用論文が存在していない場合もあり必ずしもこの手法を利用できるとは限らない．

そこで，本研究では，生成 AI に特徴や用途を推測してもらうことで，その研究データの特徴や用途を含む，検索を考慮したメタデータを付与できるのではないかと考え，その可能性について調査を行った．

2.2 ChatGPT

本研究では，生成 AI のモデルとして，ChatGPT を用いた．ChatGPT[3]とは，Chat Generative Pre-trained Transformer の略称で，米国・サンフランシスコの AI 研究機関である OpenAI が開発した人工知能チャットボットであり，生成 AI の一種である．ChatGPT は，InstructGPT をベースとしたモデルである．InstructGPT [12] とは，人間の好みに合った文を出力するように微調整した GPT のことである．あらかじめ用意された回答を返す従来のチャットボットと違い，会話から情報を推測し返答を行うことができる．ChatGPT のプロンプトを書く際に，＃を使うことでラベルをつけて指示を出すことができる．＃を使うメリットとしては以下の 3 点が挙げられる．

1. 指示が明確に伝わる．
2. 回答のバラつきが抑えられる．
3. 指示漏れが減る．

特に，指示漏れが減るという点のメリットは大きく，長い文章で指示を送った場合，ChatGPT はこちらの指示を漏らしてしまうことが多い．しかし，＃を使用し，指示を明確に細分化することで指示漏れが防げるようになる．＃を使いラベルをつけたプロンプトを書くことによって，こちらの意図した回答が得られやすくなる．

3：https://chat.openai.com/

図 1: 実験の概要

3 提案手法

本研究では，研究データにメタデータとして検索キーワードを付与する際に，生成 AI を利用することで，その研究データの利用法を考慮した検索キーワードの付与支援を行うことを提案する．今回，生成 AI のモデルとして ChatGPT を利用した．

キーワードを提案する際，ChatGPT に，与えた情報に含まれていない利用法を考慮したキーワードを提案してください，と文章で命令を与えたところ，新しいキーワードが提案されることは少なく，与えた情報を含んだキーワードを多く提案してしまう結果となった．そこで，今回提案する手法では付録のプロンプトを用いてキーワードの提案を行った．ChatGPT に入力するプロンプトは以下の構成となっている．

> ChatGPT に入力する研究データの情報: タイトルとその研究データに関する説明文検索キーワードを提案する流れ:
> （1） 与えられた研究データの情報から，その研究データを利用して行える研究にどのようなものがあるのかを想像させる．
> （2） 与えられた研究データの情報から，できる限り多くキーワードの抽出を行う．
> （3） (1) で想像してもらった研究内容から，対象の研究データを検索する際に必要な検索キーワードを (2) で抽出したキーワードを除いて 10 個提案する．

また，制約条件を課さない場合，固有名詞を含んだキーワードを多く提案する結果が多く見られた，提案したキーワードがキーワードではなく文章であるような提案をすることも多く見られたため，プロンプトに与える制約条件として，固有名詞を利用しない，文章での提案は行わないといった制約を課した．

アウトプットさせる際には，提案させるキーワードのみをアウトプットするのではなく，利用法や与えられた研究データ情報も先にアウトプットさせている．これを行うことにより，それらの情報をより反映させた結果が得られた．そのため，アウトプットをさせる際は，提案させたキーワードだけではなく，利用法と与えられた研究データ情報をできる限り出力するようプロンプトに指示を書いた．

4 実　　験

本研究では，提案手法で提案されたキーワードに対して，主観評価と客観評価の 2 種類の方法で評価を行った．図 1 は実験

表 1: 評価の基準

評価	評価基準
5	研究データの内容について的確に表している
4	研究データの内容について的確に表しているワードを含んでいる
3	研究データの要素を持っているが，漠然としている
2	研究データの内容と少し離れている
1	研究データの内容とほとんど関係がなかったり，英語として不自然な表現である

表 2: 提案されたキーワードの評価例

評価	キーワードの例
5	Risk Assessment，Malware detection
4	Malicious web content，Code Vulnerability Assessment
3	Technology，Data Protection
2	Subsystem Security，Fictitious user activity
1	Communication，Networking

の概要を表した図である．

4.1 主観評価

主観評価では，50 個の研究データに対し，3 章で述べた手法を用いてキーワードの提案を行い，そのキーワードに対して 5 段階で評価を行った．

4.1.1 使用した研究データ

キーワードを提案する際に，Scopus に掲載されているデータ論文から，オープンデータとなっているデータ論文のタイトルとアブストラクトに記述されている説明を利用した．使用したデータ論文は，2020 年までに公開されたものを対象とした．これは，今回研究に使用した ChatGPT が無料版であり，2021 年 9 月までの情報で学習を行っているため，情報の不足を防ぐために 2020 年までのデータ論文を対象とした．また，提案されたキーワードを評価する際に，非専門分野の研究内容に対して適切に評価を下すことの困難さを考慮し，今回対象としたデータ論文は，筆者が評価可能な分野であるセキュリティ，ICT，SNS，NLP の分野からキーワードを考え，実際に検索されたデータ論文のアブストラクトを読んで，筆者が評価可能であると判断したもののみを対象とした．

4.1.2 評価の基準

評価の基準は表 1 に記したとおりである．5 段階評価で数字が大きいほど評価が高い．研究データの内容，または，そこから利用され得る内容に関して的確に表せているものを 5 とし，研究データの内容，または，そこから利用され得る内容に関して的確に表した語を含んではいるが，必要のない表現まで含まれてしまっていたりなど，5 には少し足りないものを 4，研究データの関係のある語ではあるが，漠然としており検索され得る対象が多いものを 3 とした．また，研究データの内容と関係ないとまでは言えないが，少し離れているものを 2，研究データの内容にほとんど関係のないもの，または，英語としての表現が不自然なものを 1 とした．また，ChatGPT に与えた情報の中に含まれているキーワードが提案されていた場合は，評価の対象外とした．提案されたキーワードの評価例は表 2 の通りである．

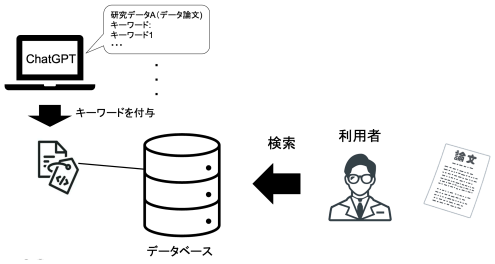


図 2: 直接的な客観評価

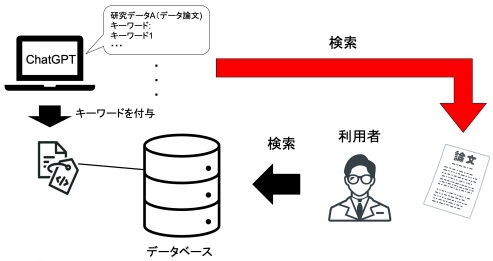


図 3: 間接的な客観評価

4.2 間接的な客観評価

提案されたキーワードが，利用法を考慮した妥当なキーワードであることを評価するために，客観評価を行った．客観評価として，図 2 のような評価方法が考えられる．提案されたキーワードを実際にメタデータとして付与し，利用者による検索実験を行い，付与する前と後で比較することで直接的に評価する方法である．これを行うには，実際に利用者に検索してもらう必要があるが，目的の研究データを必要とする利用者を用意することは難しい．また，利用者を想定したキーワードで検索する方法も考えられるが，この場合，利用者が検索するキーワードを選ぶ際に主観が介入する．さらに，対象としている研究データ以外も大量に含むデータベースを構築し，さらにそのデータすべてに対して，提案する手法で求めた新たなキーワードをメタデータとして追加したもう一つのデータベースの構築をする必要もあり，このような実験を行うのは非現実的である．

そこで，本研究では，図 3 で表すような，間接的な客観評価の方法を用いる．データ論文を引用している論文は，研究データを利用した研究であると考え，あるデータ論文に対して，3 章で述べた手法で提案されたキーワードから，そのデータ論文を引用している論文の検索を行うことで，データ論文を引用し

表 3: 提案されたキーワードの個数に関する結果

与えた情報には含まれていない新たなキーワード数	401
データ論文あたりの新たなキーワードの提案個数	8.35

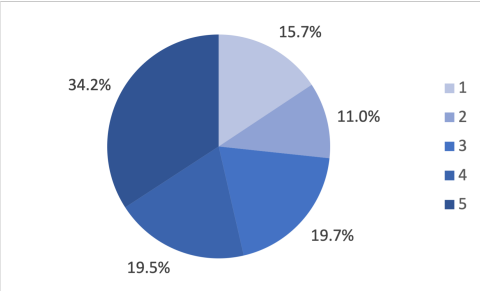


図 4: 提案されたキーワードに対する主観評価の結果

表 4: 客観評価の結果

データ論文を引用している論文が 1 つ以上検索できた	31
データ論文を引用している論文が 1 つも検索ができなかった	19

ている論文と提案されたキーワードのつながりを調査した．検索には，Semantic Scholar[4]の API を用いた．提案された 10 個のキーワードの中から，ChatGPT に与えた情報には含まれていないキーワードを選出し，そこから，最大 3 つのキーワードを組み合わせて検索を行い，それぞれ検索結果の上位 100 件を取得した．取得した検索結果において，データ論文を引用している論文が含まれているかどうかを調査した．利用したデータは，4.1.1 節で使用した 48 個のデータから，引用している論文が記載されていた 45 個のデータに，“ dataset ”のキーワードで検索した結果から無作為に選んだデータ論文 5 個を加えた 50 個のデータを利用した．

5 結　　果

5.1 主観評価

キーワードの個数に関する結果を表 3 に記す．提案されたキーワードは 1 つのデータ論文あたり 10 個ずつ提案されており，合計で 480 個のキーワードが提案されている．その中で，ChatGPT に与えた情報の中には含まれていないキーワードの提案が 401 個，データ論文 1 つあたり 8.35 個が提案された．また，1 つのデータ論文に対する新たなキーワードの提案数は，最大が 10 個，最小が 3 個という結果となった．提案されたキーワードの中で，新たなキーワードが 5 個未満だったものは，1 件のみであった．提案されたキーワードに関する主観評価のスコアの平均は，3.45 で，スコアの分布は図 4 の通りの結果となった．

5.2 間接的な客観評価

50 個の研究データに対し，提案されたキーワードでデータ論文を引用している論文の検索を行った結果，31 個の研究データ

4：https://www.semanticscholar.org/

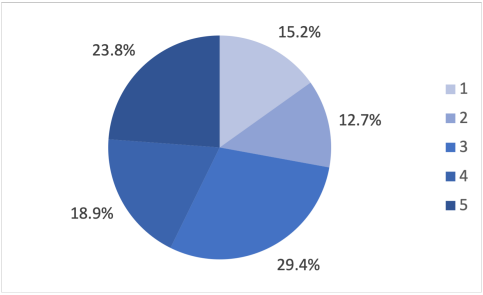


図 5: 提案されたキーワードに対する第三者の主観評価の結果

表 5: κ 係数の目安

0.00	Poor agreement（乏しい一致）
0.00-0.20	Slight agreement（わずかな一致）
0.21-0.40	Fair（ある程度の一致）
0.41-0.60	Moderate（適度な一致）
0.61-0.80	Substantial（かなりの一致）
0.81-1.00	Almost（ほぼ完全な、完全な一致）

においてデータ論文を引用している論文の検索を行うことができた．表 4 は，この結果をまとめたものである．

6 考　　察

5.1 節で示したように，一つの研究データに対して平均 8.35 個の新たなキーワードの提案が行われていることから，元データに含まれていない新しい情報も抽出できることが示された．これにより，研究データ作成者がメタデータを付与する際に，情報に追加するキーワードの候補の提案としての補助を行える可能性が高いことが示された．主観評価の分布から 4 と 5 の合計が全体の過半数を超えていることがわかる．このことから，提案されたキーワードが有用である可能性が示唆されていると考えられる．主観評価の信頼性を向上させるために，第三者に評価してもらい一致度を調べた．その際，表 1 の評価基準を第三者に学習してもらうために 10 個のデータをサンプルとして利用し，残りの 38 個のデータを評価してもらった．その結果，評価の平均は 3.24 であった．筆者の評価と第三者の評価の結果から κ 係数を求めたところ，0.50 となった．表 5 から，適度な一致度であり，十分に一致しているとは言えない．また，第三者による評価の分布は図 5 の通りであった．筆者による評価と第三者の評価ともに 1 と 2 の評価の分布が合計で 30 %を下回っている．低評価が少ないため，提案された新たなキーワードはメタデータとして利用できる可能性が高いと考えられる．筆者の評価分布と，第三者の評価分布を比較したところ，筆者のほうは 5 の評価が一番多いのに対し，第三者の評価では，3 の評価が一番多くなっている．このことから，筆者の評価には，評価を良くしようとしたバイアスがかかった可能性がある．また，今回用いたデータは，筆者の基準で選んだデータであるため，第三者の専門性と少し離れてしまっている可能性も考えられる．評価 3 以上のキーワードは，メタデータに付与するキーワードとして研究データに付与できる可能性があるため，評価

を 3 以上のものと評価 2 以下のものの 2 値に分けて分類し，その 2 値に対して，κ 係数を求めた．その結果，κ 係数は 0.66 となり，十分な一致が示された．2 値に分類した場合，十分な一致が示され，評価 3 以上に分類されたものが 70 %を超えていることから，提案されたキーワードは，メタデータとして研究データに付与できる可能性のあるキーワードが多く提案されていると考えられる．

7 おわりに

本論文では，研究データにメタデータを付与する際に，生成 AI を活用することで，研究データの使用用途を考慮したキーワードの提案が行えるのではないかと考え，その可能性について調査した．その結果，新しいキーワードの提案を補助するという点においての有効性は示すことができた．しかし，提案されたキーワードを追加することで，検索効果が高まるのかについては可能性については，今回提案されたキーワードを直接付与し，実際に検索を行うといったことができず，評価した研究データの数も十分とは言えないため，示唆するに留まった．

課題としては，主観評価を行う人数の増加，評価した研究データの分野の偏りの改善，研究データ数の増加，評価基準の改善が挙げられる．今回は 2 人での主観評価を行ったが，主観評価を行う人数を増やしていくことで，評価の信頼性が強化されると考えられる．主観評価において，今回の研究では，評価に利用したデータ論文の分野が，筆者の評価可能な分野に偏りが生じている．より多くの分野での有効性を示すためにも，専門の違う人に評価をしてもらい，研究データの偏りを改善していく必要があると考えられる．今回の研究で評価したデータの数は，評価可能な分野に絞った結果，48 件での評価となってしまったが，データ数を増やし，提案されたキーワードの有効性についてより信頼性のある議論をしていく必要がある．主観評価において，筆者の評価と第三者の評価との一致度で十分な一致を示すことができなかった．考察でも述べたとおり，筆者の評価にバイアスが掛かってしまった可能性や，第三者のデータ論文に対する理解度による可能性も考えられるが，そもそもの評価の基準が曖昧になっている可能性も考えられる．より明確な基準で評価を行うことで，評価の一致度が改善する可能性がある．また，今回，キーワードを評価する上で，筆者の基準で評価の基準を設けたが，この評価では，提案されたキーワードがメタデータとして付与するに値するかという点については議論できているが，利用法を考慮したメタデータであるかについては，この基準では考えられているとは言い難い．利用法が考慮できているかの議論は，客観評価で行っているが，可能ならば主観評価の方でも議論すべきである．利用法が考慮されているか否か，といった点を評価に含める場合，その分野に対してより熟知している人材に評価してもらう必要がある．

今後は，上記の課題に取り組むことで，提案されたキーワードの有効性について，より議論を深めていく必要がある．

文　献

[1] 池内有為, 林和弘. 研究データ公開と研究データ管理に関する実態調査 2022: 日本におけるオープンサイエンスの現状. Technical report.
[2] 内閣府科学技術・イノベーション推進事務局. 研究データの管理・利活用に関する取組の進捗について. 2023.
[3] Lorcan Dempsey and Rachel Heery. Metadata: A current view of practice and issues. *Journal of Documentation*, Vol. 54, pp. 145–172, 05 1998.
[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. Dataset search: a survey. *The VLDB Journal*, Vol. 29, No. 1, p. 251–272, August 2019.
[5] Suppawong Tuarob, Line Pouchard, and Lee Giles. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 239–248, 07 2013.
[6] デジタル庁. データ人材管理実践ガイドブック, 2022.
[7] Natasha Noy, Matthew Burgess, and Dan Brickley. Google dataset search: Building a search engine for datasets in an open web ecosystem. Technical report, 2019.
[8] 総務省. 震災関連デジタルアーカイブ構築・運用のためのガイドライン, 第 5 章, p. 106. 2013.
[9] Youichi Ishida, Toshiyuki Shimizu, and Masatoshi Yoshikawa. An analysis and comparison of keyword recommendation methods for scientific data. *International Journal on Digital Libraries*, Vol. 21, No. 3, pp. 307–327, 2020.
[10] 笹治拓矢, 加藤誠. グラフニューラルネットワークを用いたスプレッドシートの見出し認識. 第 13 回データ工学と情報マネジメントに関するフォーラム (DEIM 2021), 2021.
[11] 角掛正弥, 松原茂樹. 研究データ検索における論文上の引用文脈の利用. 言語処理学会第 27 回年次大会, pp. 334–339, March 2021.
[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

付　録

```
**#Instruction**

Please follow the steps in **#Procedure** to suggest
     keywords. However, please follow the **#
    Constraints** restrictions.

**#Dataset_Information**

- Dataset name

- Explanation

**#Constraints**

- Follow the specified **#Output_Format**.
- Please use 1 word for Suggested keywords.
- Don't use sentences for Suggested keywords.
- Do not use proper noun for Suggested keywords.
- Don't use words included in **#Dataset_Information
    ** for Suggested keywords.
```

```
**#Procedure**

- Based on **#Dataset_information**, consider
    research using this dataset.
- **#Dataset_information** is information about the
    dataset you want to search.
- Please extract as many keywords as possible from
    **#Dataset_information.**
- Please suggest keywords that exclude words
    included in the keywords you extracted earlier,
     and that are suitable for the search
    considering research using this dataset.
- List keywords for searching this dataset necessary
     for this research.

**#Output_Format**

- Research using this dataset.(10 in total)
- All extracted keywords
- Suggested keywords (10 in total)
```

スプレッドシートを用いたデータ解析における情報要求分析

丸田　敦貴† 　田貝　奈央† 　加藤　誠††

† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s1711567@klis.tsukuba.ac.jp, s2321695@u.tsukuba.ac.jp ††mpkato@acm.org

あらまし　本研究では，スプレッドシートを用いたデータ解析中に生まれる情報要求についての調査を行い，データ解析における情報要求の分類法の提案および情報要求の頻度，検索の難易度，遷移，ユーザ行動の分析結果の報告を行う．我々はユーザ実験を実施し，実験参加者のデータ解析とレポート作成の過程を収録した．そして，収録映像のコード化とタスク後のインタビューを通して，データ解析中の情報要求における洞察を得た．具体的には,（1）データ解析中に生まれる情報要求の 6 つのカテゴリ（テキスト，数値，メタデータ，データ傾向，根拠，解析方法）を特定した.（2）情報要求を満たすための検索行動はカテゴリによって異なり，特に，データの傾向の根拠やデータ解析の方法を検索する際に，実験参加者は異なる種類の困難に直面することが観察された.（3）情報要求と共起するユーザ行動（Web ページ閲覧する，スプレッドシートを読む，レポートを書くなど）もカテゴリごとに異なることが観察された.

キーワード　ユーザモデル・行動分析, 情報探索行動, ユーザ理解, スプレッドシート

1　はじめに

データ解析は大量のデータから価値のある情報の取得や戦略的な意思決定のために重要である．データ駆動の意思決定はあらゆる分野において一般的になっており，非専門家のデータ解析を支援するための研究が行われている [1–4]．これらのデータ解析ツールの中でも，スプレッドシートは直感的な操作が可能であり，素早いデータ入力，編集，解析が可能である．さらに，ノーコードのデータ解析ツールは少ない学習コストでのデータ解析を可能にさせるため，多くのユーザに利用されており [5]，スプレッドシートを拡張する研究もなされている [6–8].

スプレッドシートのようなデータ解析ツールはデータ解析の知識不足を一部補うことが可能な一方で，スプレッドシート自体の理解や解析結果の正確な解釈のために，多様な知識（ドメイン知識や分析方法の知識など）が必要である．既存研究によると，ユーザはスプレッドシート理解のために約 40%の時間を情報探索に費やしていると報告されており [9]，さらに，非専門家によるデータ解析結果の誤った解釈についても懸念されている [10]．このように，ノーコードツールによるデータの直感的な操作が可能になったとしても，ユーザ自身の知識不足に対応しなければ，効果的で効率的な解析は困難であり，誤った知見を導いてしまう可能性もある．したがって，データ解析中の効果的な情報提供は重要である．そのために，データ解析中のユーザが，いつどのようなタイプの情報を求めるか，また，どのような検索行動をとるかなどの，情報検索行動に関するより深い理解が必要である.

本研究はデータ解析中の情報検索行動，特に，スプレッドシートを用いたデータ解析タスクにおいて，ユーザに生まれる情報要求を分析する．我々は情報要求の分類法を提案し，それぞれの情報要求の頻度，検索の難易度，遷移，ユーザ行動などの分析結果について報告する．実験を通して，我々は以下のリサーチクエスチョンに答える.

RQ1.　スプレッドシートを用いたデータ解析中にどのような情報要求が生まれるのか？

RQ2.　スプレッドシートを用いたデータ解析中の情報要求を満たすためにユーザはどのように検索するのか？

RQ3.　スプレッドシートを用いたデータ解析において，いつ情報要求が生まれるのか？

これらのリサーチクエスチョンに回答するために，我々はスプレッドシートを用いてデータを解析し，その結果を要約するレポートを作成するデータ解析タスクを設計した．そして，36 人の大学院生を募り, 思考発話法を用いて参加者の情報要求を収集した．我々は文献調査に基づいて仮の情報要求の分類法を作成し，収録された情報要求の事例を使用して，さらに改善を行った．また，タスク後のインタビューを通して，各情報要求に関する主観的な満足度と検索の難易度に関するデータを収集し，さらに，データ解析タスクの収録映像を見返してユーザ行動をコード化した.

RQ1 に関して，スプレッドシートを用いたデータ解析中に生まれる情報要求の 6 つのカテゴリを特定した．各カテゴリを満たすための難易度は異なり，用語の意味を探すことは比較的容易である一方で，データの傾向の根拠や分析方法を探すことは比較的難しいことが明らかになった．そして，RQ2 に回答するために情報要求を満たす際の検索行動を分析した．分析の結果，参加者がデータの傾向の根拠や分析方法を探す際に，異なる種類の困難に直面したことが観察された．最後に RQ3 に回答するために，情報要求と共起するユーザ行動および情報要求間の遷移を調査した．その結果，情報要求はカテゴリごとに異なる状況で生まれることが明らかになった（例えば，Web 検索，スプレッドシートの理解，レポート作成など）.

2 関連研究

本節では情報検索とデータ解析におけるユーザ分析の研究について述べ，データ解析中の情報要求について着目されていないことを指摘する．

2.1 検索行動のユーザ分析

情報検索の分野では, 多くの研究がユーザの意図について分析している．Broder [11] は Web 検索における情報要求の 3 つのカテゴリ（Navigational, Informational, Transactional）の分類法を提案した．この分類法は後に Rose と Levinson [12] によって拡張された．Web 検索の意図を対象とした他の分類法も提案されており [13, 14]，商品検索に関する分類法もある [15,16]．他にも，コミュニティ質問応答の質問の分類法やビジネスドキュメント中の質問の分類法や [17–19]，創造的なデザイン活動における情報要求に関するもの [20] がある．これらの研究は主に Web 検索におけるユーザの意図を分類している．

ユーザの検索行動を分析する研究も幅広くなされている．Pharo と Jarvelin [21] は Web 検索のプロセスを分析しており，Kelly と Teevan [22] は Web 検索の行動を分類した．また，ユーザの行動に基づいて検索効果を向上させる研究が盛んに行われており [23–25], 複雑な検索タスクにおけるユーザの行動を調査した研究もなされている [26–28].

Koesten ら [29] は仕事で構造化データを利用するユーザのデータ探索行動を調査している．この研究はクエリログの分析と専門家へのインタビューによって，仕事で利用するデータを見つけるまでの行動を分析しており，データを探す目的やデータを用いた活動など興味深い分析がなされている．

しかしながら，多くの既存研究は Web 上での検索行動のみに着目しており，ユーザのスプレッドシート操作やデータから知見を得るための情報探索行動については明らかではない．本研究は Web 検索行動とスプレッドシートを用いたデータ解析行動の関係を分析することで，これらの点を明らかにする．

2.2 データ解析のユーザ分析

本節では, データ解析におけるユーザの行動に関する研究について述べる．最初に，データ解析のプロセスを分析する研究について説明する．Fayyad ら [30] は, データから知識を導出するための選択, 前処理, 変換のステップからなる KDD プロセスを提案した．そして，Donoho [31] はデータサイエンスの 6 つのプロセスを提案した：(1) データ収集，準備，および探索, (2) データ表現と変換, (3) データ計算, (4) データの視覚化と提示, (5) データモデリング, (6) データサイエンスに関する科学.

一部の研究では，データ解析中のユーザの行動を調査している．例えば，SQL ベースの分析ログを調査することでユーザの行動を評価する研究があり [32]，Jupyter ノートブックを使用してプログラミングする際の研究もなされている [33]．しかし，これらの調査の中心は主にデータの理解と操作であり，データ解析中の検索行動はほとんど着目されていない．

Rajalingham ら [34] はスプレッドシートのエラーに関する分類法を提案しており, Rangavan ら [9] はユーザ実験を通して，スプレッドシートを理解するまでの行動を分析した．この研究はスプレッドシートとのインタラクションを調査するという点で本研究と似ているが，我々はスプレッドシートの理解だけでなく，スプレッドシートを探す行動やスプレッドシートから知見を得る行動といった，より多様な行動を対象とする．さらに，スプレッドシートから知見を得るための情報検索行動にも焦点を当て，ユーザが知見を得るためにどのような情報が必要なのかを明らかにする．

3 ユーザ実験

本節では，データ解析中のユーザ行動と情報要求を収集するためのユーザ実験について説明する．このユーザ実験は筑波大学図書館メディア系研究倫理審査委員会で承認済みである．

3.1 実験協力者

筑波大学に所属する 36 人がこの実験に参加した．我々はデータ解析およびレポート作成の基本的な知識を持つ参加者を募集するため，参加者は学士を有する参加者，すなわち大学院生に限定した．参加者は女性 14 人，男性 21 人であり，1 人の参加者が自分の性別を開示しないことを選択し，年齢の幅は 23 から 27 歳で，平均年齢は 23.69，標準偏差は 1.24 であった．参加者の専門分野は情報学, 物理学, 数学，社会工学，神経科学など，多様であった．事前アンケートによると，72.2%の参加者はデータ解析を行う頻度が 2 週間に 1 回以下と回答していた．ここから，多くの参加者が日常的にデータ解析を行なっていないことを示している．

3.2 ユーザ実験プロトコル

ユーザ実験はリモートで実施され，以下の 5 つのステップで行われた．

（1） 参加者は実験の目的とタスクの概要の説明を受け，実験に同意した場合に同意書にサインした．

（2） 事前アンケートへの回答を行なった．ここでは参加者の人口統計学的特徴，大学院，データ解析の頻度と知識，実験に用いるデータの 6 つのドメインの知識に関して質問された．

（3） 参加者はサンプルレポートとレポートの評価基準を参照しながら，タスクの詳細な説明を受けた．そして，思考発話法について簡単な練習を行い，その後，タスクに関する不明な点について質問された．

（4） 参加者はデータ解析タスクに取り組んだ．タスクの時間は最長 60 分間で，参加者の任意のタイミングでタスクを終了することができた．タスク終了後，10 分の休憩が設けられた．タスクの詳細については，第 3.3 節で述べる．

（5） 休憩後, 参加者はタスクの難易度や所要時間などに関する事後アンケートに回答した．そして，刺激回想法を使用したインタビューが行われた．このインタビューについては，第 3.4 節で説明する．

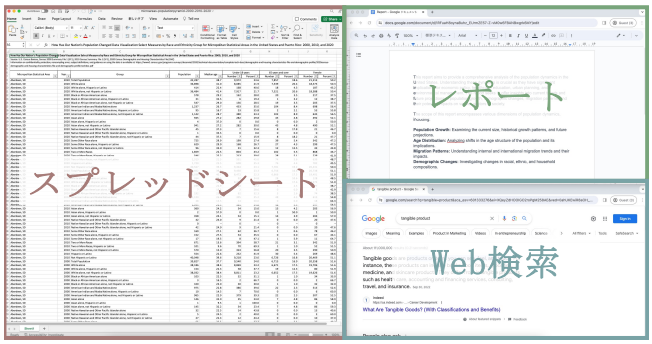


図 1: データ解析タスク中のスクリーンショット．参加者はスプレッドシートを読み，理解し，洞察を得て，それらをレポートにまとめた．また，参加者はタスク中に自由に Web 検索ができた．

3.3 データ解析タスク

我々はスプレッドシートから洞察を得るときのユーザ行動を収集するためにデータ解析タスクを設計した．このタスクの目的はオープンデータを解析し，得られた知見を評価基準に沿ってレポートに書くことである．基準は次の 3 つである：(1) データから得られた知見が複数挙げられているか，(2) 解析したデータは信頼できるものか，(3) 得られた知見には根拠があるか．参加者は自らの思考を声に出しながら，スプレッドシートを解析し，知見をレポートにまとめるように指示された．

図 1 はデータ解析タスク中のスクリーンショットである．スプレッドシートの閲覧および操作のために，Microsoft Excel を利用し，レポート作成のために，Google Docs を利用した．また，参加者はタスク中，自由に Web 検索を行うことができた．参加者の操作画面と発話を記録するために Microsoft Teams の画面収録機能を用いた．

タスクの開始時に各参加者は 6 つのドメイン（人口，農業，経済，天候，エネルギー，社会福祉）のうちの 1 つが割り当てられた．そして，そのドメインにおける Web サイトのリンクが 4 つ表示されている擬似 Web 検索結果ページが見せられた．6 つのドメインは広く利用されるオープンデータの Web サイト (Data.gov[1]など) に基づいて選択された．実験に用いるデータの多様性を高めるために，各ドメインにおいて，2 つの政府系の Web サイトと 2 つの非政府系の Web サイトを選択した．参加者はこれらの Web サイトを閲覧し，分析に必要なオープンデータを見つけてダウンロードし，分析を行い，その知見をレポートにまとめるように指示された．この実験設定は，ユーザが特定のドメインに興味を持ち，Web 検索エンジンを用いてオープンデータを検索する状況を模倣している．

タスク中，参加者は自由にタスクに関する質問ができた．参加者は自分の考えを声に出すように指示されたが，発話せずに 5 分以上黙っていた場合，実験者は思考の発話を促した．30 分と 50 分が経過したとき，実験者は経過時間と残り時間を参加者に伝えた．タスク中，実験者は参加者の行動を観察し，情報要求が生まれたとみなせる発話（例：「～を知りたい」，「これは何」など）と経過時間をメモした．このメモは，タスク後インタビュー（第 3.4 節）での再生部分を決めるために使用された．

3.4 タスク後インタビュー

タスク後，刺激回想法を用いた半構造化インタビューが行われた．実験者は情報要求が生まれたとみなされる部分を再生し，参加者は再生部分を回想しながら，次の質問に答えた：

- 再生部分で，どのような情報要求が生まれていましたか？
- その情報要求を満たすことができましたか？
- その情報要求を満たすためにどのような情報にアクセスしましたか，またはアクセスする必要を感じましたか？
- その情報要求を満たす難しさはどの程度ですか？

主観難易度の尺度について，“簡単”，“普通”，“難しい”，“不可能” の 4 つの選択肢が与えられた（“不可能” は，その情報要求が満たされなかった場合の選択肢であった）．

3.5 インタビュー分析

参加者のインタビューの回答を分析した．まず，著者がインタビューの内容（情報要求， アクセスした情報/欲しかった情報, 満たされたかどうか， 満たすための主観難易度) を詳細に書き起こし，情報要求ではない事例（情報を欲する要求ではない事例）を削除した．次に，データ解析タスクの収録映像を見返し，インタビューの回答にはないが，映像から情報要求が生まれているとみなせる事例や，反対に，インタビューの回答にはあるが，映像では確認できなかった事例が見られたため，それらの事例を取り除いた．また，1 人の参加者が実験中にインターネット接続の問題が発生し，ビデオ録画の中断や，音声と映像のずれが発生した．この参加者の実験データを正確に分析することは難しいと考えたため，このデータは以降の分析から除外した．これらの前処理の結果，35 人の参加者から，合計 438 個の情報要求の事例 (1 人当たり平均 12.5 個) を得た．

3.6 ユーザ行動コーディング

ユーザの行動をより詳細に理解するために，データ解析中のユーザ行動のコーディングを実施した．Ragavan ら [9] の手法を参考にして，収録映像を 30 秒ごとのセグメントに分割し，各セグメントにコードを割り当てた．

コーディングのために，スプレッドシートの理解のために作成されたコードセット [9] を部分的に利用した．このコードセットには，スプレッドシート内の特定のセルを探す “where” やグラフの作成や利用の “visualize” などが含まれている．しかしながら，このコードセットにはいくつかのデータ解析タスクの行動（特に Web 検索やレポート作成の行動）が含まれていなかった，コードセットの拡張が必要であった．そこで，著者の 1 人が 6 人の参加者の収録映像を見返し，既存のコードに対して，存在しないコードを追加し，映像で確認できなかったコードを削除することで，49 のコードからなる初期コードセットを得た．また，コーディングが開始した後に，初期コードセットにない 3 つ行動が見られたため，それらを追加し，最終的に 52 のコードを得た（詳細は付録 1 に記述）．

そして，拡張されたコードセットを利用して，著者の一人が

1：https://data.gov/

35 人分の収録映像に対してコーディングを行った．その結果，全 4，097 セグメント中 3,941 セグメントにコードが割り当てられ，合計 5,511 コードを得た (同じセグメントに対する複数コードの付与も可能にしている)．その後，情報学の学士をもつ 2 人の評価者にコードを見返してもらい，著者のコードと 97.1%のセグメントで同意が得られた．同意が得られたかったセグメントにおいて，著者と評価者で議論を行い，より適切なコードを決定した．その結果，13 セグメントが著者に譲歩され，106 セグメントが評価者に譲歩された．このプロセスを経て，最終的に 5,612 コード（5,511 の著者のコード + 110 の追加コード − 9 の削除コード）を得た．

4 分 析 結 果

本節では，ユーザ実験と追加の分析の結果に基づいてリサーチクエスチョンに回答する．

4.1 RQ1: スプレッドシートを用いたデータ解析中にどのような情報要求が生まれるのか？

RQ1 に回答するために，我々は Bolotova ら [35] の手法から着想を得て，多段階アプローチを適用した．まず，文献調査を通して，仮の分類法を構築し，そして，ユーザ実験の結果を用いて分類法を改善する．最後に，改善された分類法と各情報要求に付与されたラベルを評価する．

4.1.1 文 献 調 査

まず，最初の段階として，データ解析に関する文献調査を行い，仮の分類法を作成した．分類法の構築において，全く新しい分類法を作成する方法ではなく，既存の分類法を改善する方法で分類法を構築することが重要であり，これまでに多くの研究で文献調査を含む多段階の方法が適用されている [18,19,35]．

我々はデータマイニング，ヒューマンコンピュータインタラクション，Web サイエンス, 情報検索などの国際会議で発表された論文や， データ解析に関する論文誌を調査した．文献調査の対象とした国際会議は，CHI, CIKM, KDD, RecSys SIGIR, WSDM, TheWebConf の 7 つである．次に，データ解析に関する総説論文 [36–39] を参考にして，本研究の調査対象となる 3 つの論文誌（Decision Support Systems, Journal of Management Information Systems, Information Systems Research）を決定した．文献調査に用いたシステムは，Google Scholar, ACM Digital Library, JSTOR Archive, Taylor & Francis Online, Informs Pubs Online である．各システムで検索に用いたクエリはデータ解析に関するキーワード (*data analysis* や *business intelligence* など) と本研究の特徴を捉えるキーワード (*information need* や *tabular data* など) の 2 つを組み合わせたものを用いた．そして，検索結果に対して，以下の 4 つのフィルタをかけた：(1) 発行年が 2000〜2022 年の間である，(2) フルペーパーの論文である， (3) データ解析の対象がテキストや知識グラフではなく，表形式データである, (4) データ解析をしているユーザに対する分析がなされている．

調査の結果，データ解析の情報行動の分類法（例えば，データ解析中に生まれる質問の分類 [40] など）やシステム（例えば，データ解析のための科学記事推薦システム [41] など）に関する 85 本の論文を収集し，それらを閲読した．この文献調査に基づき，5 つのカテゴリからなる仮の分類法を構築した．

テキスト　表中の単語の意味に関する要求 [34, 40, 42].

データ　閲覧している表に含まれていない表に関する要求 [40, 42, 43].

メタデータ　表の出典元や作成方法などの表自体の情報に関する要求 [40, 44].

知見　知見の根拠となる文献およびデータに関する要求 [41,45].

分析方法　データ解析の方法に関する要求 [44, 46, 47].

4.1.2 分類法改善

3.5 節で説明された，書き起こされたインタビューの内容を用いて，分類法の改善を行った．本研究では分類法改善の途中でフィードバックを受けながら，より汎化された分類法を作成するために， Capra らのコーディング方法 [25] を参考にした．まず，2 人の著者が別々に 6 人分のインタビュー (各ドメイン 1 人ずつ) を見て，それぞれ仮の分類法の改善を行った．その後，それぞれが改善した分類法を見せ合い，議論を行い，一つの分類法にまとめた．次に，2 人の著者が別の 6 人分のインタビューを見て，それぞれが独立して分類法を調整し，その後，それぞれが改善した分類法を見せ合い，議論を行い，分類法のさらなる改善を行った．最後に，それぞれの分類カテゴリが適切な表現となるように，2 人の著者が語彙の変更を行なった．その結果，得られた 6 つのカテゴリからなる分類法を表 1 に示す．分類法の詳細の説明は 4.1.4 節で行う．

4.1.3 ラベリング

最後の段階として，アノテータによる情報要求のラベリングにより，改善された分類法を検証した．情報学の学士を持つ 2 人のアノテータを募り，作成した情報要求の分類法を説明し，収集された情報要求にラベルを付けるように各アノテータに依頼した．ラベリングによる回答の一致度は Cohen の $\kappa = 0.62$ であり，実質的な合意があることを示した．2 人のアノテータの回答が異なっていた場合，議論して最終的なラベルを決定した．

4.1.4 結　　果

表 1 にスプレッドシートを用いたデータ解析における情報要求の分類法を示す．この分類法は仮の分類法から 3 つの変更が加えられた．まず，“数値” カテゴリが追加された．これは，文献調査ではスプレッドシート中の数値を理解するための情報要求が見つからなかったからである．さらに，カテゴリの名前が内容を正確に表現できるように改良された．例えば，仮の分類法に含まれていた “知見” カテゴリは “根拠” カテゴリに変更され，“データ” カテゴリは “根拠” カテゴリに統合された．最後に，“データ傾向” カテゴリが追加された．これには，スプレッドシート内の探索によって満たされる情報要求が含まれている．

本研究で構築した分類法の詳細を以下の通りである:

テキスト　スプレッドシートのテキストの内容について明確さを得るために必要な情報を指す．単語の意味や他の単語との関係を理解し，スプレッドシートの理解に役立つ．

数値　スプレッドシート内の数値データを理解するために生ま

表 1: スプレッドシートを用いたデータ解析における情報要求の分類法.

カテゴリ	情報要求	#	事例
テキスト	用語の意味や詳細（事例や属性，他の用語との関連を含む）	108	無形商品って何？
数値	単位の意味や変換	15	推定日射量の単位を知りたい.
メタデータ	表の説明文（データの出所や収集方法を含む）	58	データはいつ収集された?, このデータの概要は？
データ傾向	表データの代表値（特定の計算によって得られる数値）や表データの一部分（特定の操作によって得られるデータ）	72	児童の各年齢の全体に占める割合を知りたい，各アイテムの年次推移を知りたい.
根拠	データの傾向の理由や根拠となる記事や一般知識もしくは関連する表データ	130	新潟の降雪量が多いことを裏付ける記事は？， 日本の地価が下がっているかどうかを知りたい.
解析方法	データ解析における宣言的知識や手続的知識	41	このデータをどうやって可視化したら良いかわからない． エクセルで平均値を計算する方法を知りたい.

れる．多くの場合，単位，単位変換，スプレッドシートの内容を理解するための関連指標の意味を知ることで満たされる．

メタデータ　スプレッドシートに関連するメタデータを得ることで満たされる，高次元の情報．これには，概要，作成方法，作成日などが含まれる．

データ傾向　スプレッドシートの中に含まれるデータの傾向を明らかにする要求であり，満たすために，スプレッドシートの代表値や一部分を欲することが多い．

根拠　データの傾向を裏付ける外部の情報によって満たされる要求である．これは，データの傾向に関連する記事や一般知識，他のスプレッドシートなどである．

解析方法　データ解析方法を理解することに関連しており，多くの場合，データ解析の宣言的な知識や手続き的な知識を得ることで満たされる．

図 2 は本実験で収集された情報要求の分布を示している．6 つのカテゴリの中で，“根拠” カテゴリが最も多く，次に “テキスト” カテゴリが次に多かった．一方で，“数値” と “解析方法” カテゴリの事例は少なかった．カイニ乗検定により，カテゴリ間で情報要求の満足度に有意な差があった（$\chi^2(5) = 29.0$, $p < 0.05$）．さらに，残差分析の結果，“テキスト” カテゴリの満足率が有意に高い (84.3%) 一方で，“根拠”(60.0%) と “解析方法”(46.3%) の満足率が有意に低いことが示された（$p < 0.05$）．これは “テキスト” カテゴリが比較的満たされる傾向にある一方で，“根拠” と “解析方法” カテゴリは満たされづらい傾向があることを示している．

図 3 は各カテゴリにおける情報要求を満たすための主観難易度の平均値を示している．図中のエラーバーは標準誤差を示している．Kruskal-Wallis 検定により，主観難易度においてカテゴリ間で有意な差が見られた（$\chi^2(5) = 32.2$，$p < 0.05$）．Holm 補正を用いた U 検定を用いた事後検定により，テキスト/データ傾向，テキスト/根拠, テキスト/解析方法，数値/解析方法の間に有意な差が見られた（$p < 0.05$）．この結果から，参加者は “テキスト” カテゴリを満たすことが比較的簡単であると感じる傾向にあり，一方で，“根拠” と “解析方法” カテゴリは難しいと感じる傾向にあった．

Urgo らは異なるタイプの学習目標に対する検索の満足度と難しさを調査した研究を行っている [48]．彼らの研究では，タスクの知識の種類が検索の満足度と難しさに有意な主効果をもたらしたことが示された．“テキスト” カテゴリが宣言的知識（Urgo らの分類法では “事実的” または “概念的”）に該当し，“解析方法” が手続き的知識に分類されると仮定すると，我々の結果は彼らの知識の種類の影響に関する結果と一致している．ただし，タスクや分類法の違いから直接的な比較は難しい．

a） 要　　約

文献調査による仮の分類法の構築および収集された事例を用いた改善を経て，スプレッドシートを用いたデータ解析における情報要求の分類法を構築した．各カテゴリを満たすための難易度は異なり，主観的および数量的な評価を通して，“テキスト” カテゴリは比較的容易に満たすことできる一方で，“根拠” と “解析方法” カテゴリを満たすことは難しい傾向にあった．

4.2 RQ2: スプレッドシートを用いたデータ解析中の情報要求を満たすためにユーザはどのように検索するのか？

RQ2 に回答するために，我々は各情報要求カテゴリを満たすためのユーザの検索戦略を明らかにする．

図 4 は検索を行って満たした情報要求と検索を行わずに満たした情報要求の分布を示している．カイ二乗検定を実施したところ，カテゴリによって Web 検索を利用するかどうかに有意な差が見られた（$\chi^2(5) = 105.5$, $p < 0.05$）．さらに，残差分析を実施し，“テキスト” と “根拠” カテゴリの Web 検索の頻度が有意に多く，一方で，“メタデータ” と “データ傾向” カテゴリの Web 検索の頻度は有意に低いことが明らかになった（$p < 0.05$）．これはおそらく，“テキスト” と “根拠” カテゴリを満たすための情報が Web 上に多く，反対に，“メタデータ” と “データ傾向” カテゴリを満たすための情報がスプレッドシートもしくは，スプレッドシートを公開している Web サイトに含まれていることが多いためだと考えられる．

図 5 は各カテゴリにおける 1 つの情報要求あたりの平均入力クエリ数を表している．Kruskal-Wallis 検定により，入力クエリ数はカテゴリごとに有意な差が見られた（$\chi^2(5) = 97.4$, $p < 0.05$）．Holm 補正を用いた Mann-Whitney U 検定を用いた事後検定により，テキスト/メタデータ，テキスト/データ

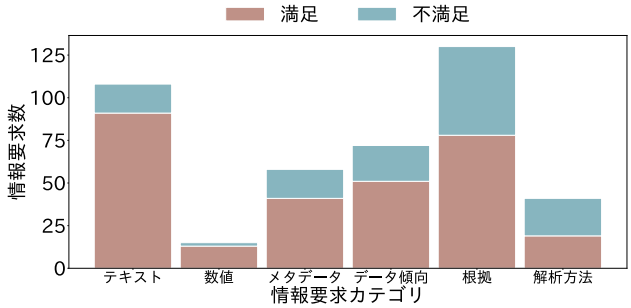


図 2: 各カテゴリの情報要求の分布．赤い棒は満足された情報要求を示し，青い棒は満足されていない情報要求を示している．

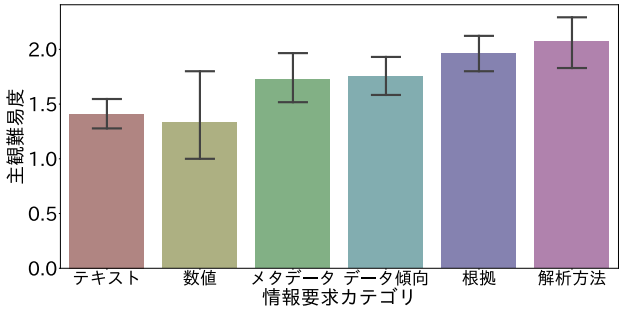


図 3: 情報要求を満たすための主観難易度の平均値．図中のエラーバーは標準誤差を表している．

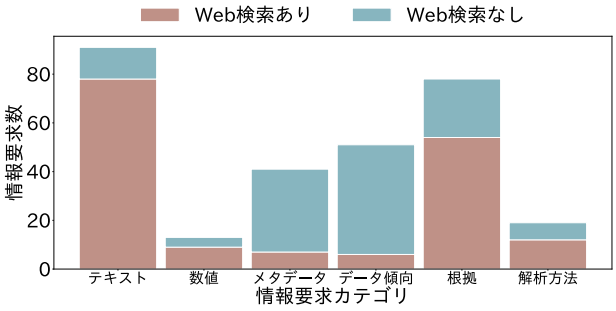


図 4: 検索を行って満たした情報要求と検索を行わずに満たした情報要求の分布．

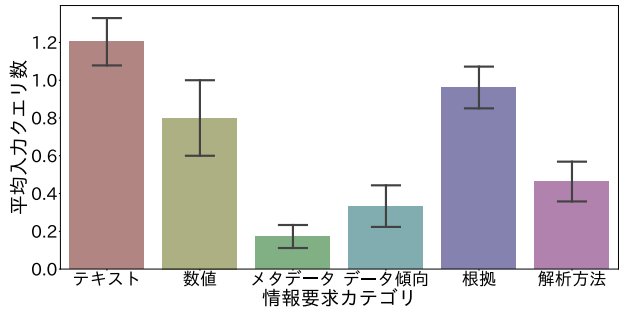


図 5: 1 つの情報要求あたりの平均入力クエリ数．

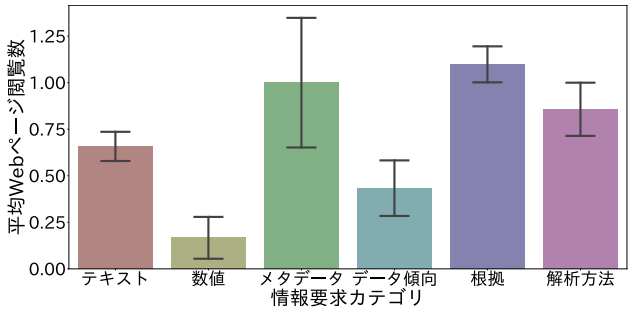


図 6: 1 クエリあたりの平均 Web ページ閲覧回数．

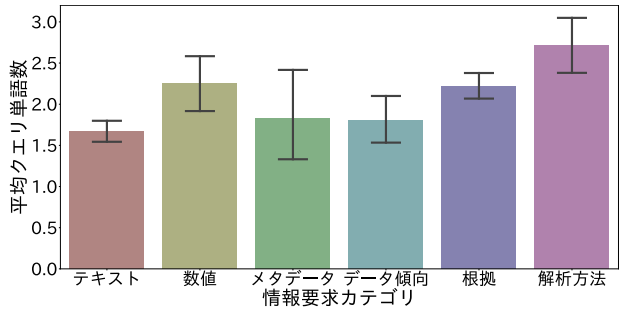


図 7: 1 クエリあたりの平均単語数．

傾向, テキスト/根拠, テキスト/解析方法，数値/メタデータ, 数値/データ傾向, メタデータ/根拠, データ傾向/根拠，メタデータ/解析方法，根拠/解析方法の間に有意な差が見られた（$p < 0.05$）．ここから，“テキスト” と “根拠” カテゴリを満たすためのクエリ入力数が多い一方で，“メタデータ” と “解析方法” の入力数が少ないことが明らかになった．参加者は “解析方法” カテゴリが最も難しいと感じていたにもかかわらず（図 3），クエリの入力数は少ない傾向にあった．ここから，ユーザは解析方法の情報が必要でも，検索しようと思わない，もしくは検索方法がわからないのかもしれない．

図 6 は 1 クエリあたりの平均 Web ページ閲覧回数を図示している．我々は Kruskal-Wallis 検定を実施し，カテゴリごとに Web ページの閲覧回数が有意に異なっていることを明らかにした（$\chi^2(5) = 31.4$, $p < 0.05$）．Holm 補正を用いた Mann-Whitney U 検定を用いた事後検定により，テキスト/根拠, 数値/根拠, 数値/解析方法, データ傾向/根拠の間に有意な差が見られた（$p < 0.05$）．ここから，1 クエリあたりの Web ページ閲覧数は “数値” カテゴリは少ない傾向にあり，反対に，“根拠” と “解析方法” カテゴリは高い傾向にあった．

図 7 は 1 クエリあたりの平均単語数を示している．Kruskal-Wallis 検定により，カテゴリごとにクエリの長さが有意に異なっていることが明らかになった（$\chi^2(5) = 48.8$, $p < 0.05$）．また，Holm 補正を用いた Mann-Whitney U 検定を用いた事後検定により，テキスト/根拠, テキスト/解析方法, データ傾向/根拠の間に有意な差が見られた（$p < 0.05$）．“テキスト” カテゴリは平均クエリ単語数が最も少なく，対照的に，“解析方法” カテゴリは平均クエリ単語数が最も高かった．これは，ユーザが特定の用語についての幅広い情報を検索するために短いクエリを入力し，解析の方法や手順のような複雑な情報を得るために詳細で長いクエリを入力することが示唆される．

図 5，7 の結果から，学習のための検索についての既存研究 [48,49] から一貫した傾向が見られた．Urgo らは知識の種類が入力クエリ数や検索の所要時間などの検索行動に与える主効果があると報告している [48]．Eickhoff らは Web 検索エンジ

表 2: 情報要求の満足に関する回帰分析の結果．“変数” 列は切片と説明変数を示している．太字の変数は有意な相関係数が見られた変数である（$p < 0.05$）．

$R^2 = 0.03696$		$p = 0.01296$	
変数	**相関係数**	**標準誤差**	p
（切片）	0.0994	0.279	0.722
データを探す	0.3259	0.317	0.303
ダウンロード	0.4546	0.701	0.517
読む	**0.7131**	**0.292**	**0.015**
質問	0.0999	0.395	0.800
操作	0.7285	0.472	0.122
可視化	-0.1679	0.353	0.635
Web 検索	**0.9520**	**0.294**	**0.001**
レポート	0.4088	0.391	0.295

ンのログ分析を行い，宣言的および手続き的な情報要求を持つ検索セッションでの入力クエリ数（6.1 対 4.3）とクエリの長さ（4.5 対 5.8）を報告している [49]．“テキスト” カテゴリが宣言的知識，“解析方法” カテゴリが手続き的知識に分類されると仮定すると，Eickhoff らの結果は，“解析方法” カテゴリの方が “テキスト” カテゴリよりも入力クエリ数が少なく，クエリの単語数が多いという我々の結果と一致している．

さらに，ユーザ行動と情報要求の満足との相関の程度を明らかにするために，ロジスティック回帰分析も実施した．目的変数は，ユーザの満足を示すバイナリ変数であり，説明変数は，情報要求が生まれたセグメントで，どのユーザ行動が見られたかを示すバイナリ変数である．既存研究で用いられたコードセット [9] と本研究で拡張したコードセットはいずれも説明変数としては細かすぎるため，類似した行動をグループ化して，9 つの行動カテゴリを作成し，これらを説明変数とした．9 つの行動カテゴリは以下の通りである: データを探す，スプレッドシートを読む，Web 検索，レポート作成，可視化，スプレッドシートのダウンロード，スプレッドシートの操作，スプレッドシートへの質問と仮説，その他（詳細は付録 1 を参照）．

表 2 は回帰分析の結果を示している．この回帰分析は有意であった（$p < 0.05$）が，擬似 R 二乗値は 0.037 と低い値を示しており，ユーザが情報要求を満たしたかどうかを正確に予測するには，さらなる要因を考慮する必要があることを示している．

正確な予測ができていないにもかかわらず，いくつかの説明変数が回帰分析で有意であることが示された．“Web 検索” と “読む” に対する有意な正の係数が見られた（$p < 0.05$）．スプレッドシートを読む，もしくは Web 検索をする際に，情報要求が生まれたとき，比較的容易に満たされる傾向にあった．

a） 要　　約

“テキスト” と “数値” カテゴリは比較的簡単にアクセス可能である．これらのカテゴリは短いクエリが入力される傾向にあり，さらに，Web ページの閲覧回数が少ないことから，ユーザはこれらの情報要求に対して少ない労力で対応していることが示唆される．一方で，“メタデータ” と “データ傾向” カテゴリは Web 検索だけではほとんど満たされない．これらの情報要求はデータを公開する Web ページの閲覧やスプレッドシートの操作によって満たされる傾向にある．対照的に，主観難易度が高い “根拠” カテゴリは Web ページを閲覧することで満たされることが多い．このカテゴリを満たすために，より長いクエリが入力される傾向にあり，これは，より詳細な情報を欲していることを示唆することから，この情報要求を満たすことは難しいことを示唆している．“解析方法” カテゴリは参加者から最も難しいと評価され，クエリごとの Web ページ閲覧回数も比較的多い．長いクエリが入力される傾向があるにもかかわらず，少ない入力クエリ数は，満たすことの難しさを示唆している．

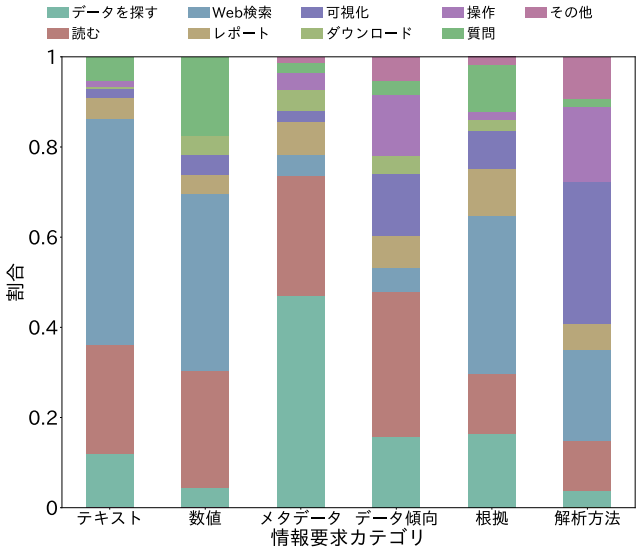


図 8: 情報要求が生まれたセグメントにおけるユーザ行動カテゴリの割合．ユーザ行動カテゴリの詳細は付録 1 に示す．

表 3: 情報要求の遷移確率．左の列は遷移前，上の行は遷移後のカテゴリを示している．▲ は残差分析により有意に高い確率を示しており（$p < 0.05$），▽ は有意に低い確率を示している（$p < 0.05$）．

前　後	テキスト	数値	メタ	傾向	根拠	解析
テキスト	▲0.396	0.019	0.123	0.142	▽0.264	0.057
数値	0.267	0.000	0.133	0.200	0.200	0.200
メタ	0.236	0.055	▲0.236	0.182	0.236	0.055
傾向	0.254	0.060	0.060	▲0.299	0.209	0.119
根拠	0.107	0.036	0.080	▽0.152	▲0.545	0.080
解析	0.147	0.059	0.088	0.118	0.265	▲0.324

4.3 RQ3: スプレッドシートを用いたデータ解析において，いつ情報要求が生まれるのか？

RQ3 に回答するために，各カテゴリの情報要求が生まれる際のユーザ行動と情報要求カテゴリ間の遷移確率を分析した．

図 8 は，情報要求が生まれたセグメントにおけるユーザ行動カテゴリ（詳細は付録 1 を参照）の割合を示している．カイ二乗検定を実施し，情報要求カテゴリとユーザ行動の間に有意な差が見られた（$\chi^2(40) = 262.0$，$p < 0.05$）．さらに，残差分析により，テキスト ×Web，数値 × 質問，メタデータ × データ探索，データ傾向 × 読む，データ傾向 × 操作，根拠 ×Web，

根拠 × レポート，解析方法 × 可視化，解析方法 × 操作などで有意に高い割合が見られた．また，テキスト × データ探索，テキスト × 可視化，テキスト × 操作，メタデータ ×Web，メタデータ × 可視化，データ傾向 ×Web，根拠 × 読む，根拠 × 操作，解析方法 × データ探索，解析方法 × 読むなどで有意に低い割合が見られた（$p < 0.05$）．

“テキスト” カテゴリは “Web 検索” の割合が高かった．これは，ユーザが Web 検索中に未知の単語に遭遇しやすいのかもしれない．“数値” カテゴリは “質問” の割合が比較的高いことから，ユーザ自身の知識からスプレッドシート内の数値の解釈を考える必要があるのかもしれない．“メタデータ” カテゴリは” データを探す” の割合が高く，ユーザがスプレッドシートを探す際に，スプレッドシートの情報源や概要を知りたいと考えている可能性がある．“データ傾向” カテゴリは “読む”，“操作” および “可視化” の割合が高かった．ここから，ユーザがスプレッドシートを読み，データの傾向を知りたいと感じ，その後，スプレッドシートを操作したのかもしれない．“根拠” カテゴリは “レポート” と “Web 検索” の割合が比較的高く，解析結果をまとめることがデータの傾向の根拠を検索するきっかけになっている可能性がある．“解析方法” カテゴリは “可視化” と “操作” の割合が高かった．ここから，ユーザはデータの操作を開始するまで，解析方法の知識不足を自覚していない可能性がある．

表 3 は情報要求のカテゴリ間の遷移確率を示している．左の列は遷移前，上の行は遷移後のカテゴリを示している．カイ二乗検定により，カテゴリ間の遷移確率に有意な差があった（$\chi^2(25) = 85.9$，$p < 0.05$）．残差分析の結果，“テキスト/テキスト”，“メタデータ/メタデータ”，“データ傾向/データ傾向”，“根拠/根拠”，“解析方法/解析方法” の確率が有意に高く，“テキスト/根拠” と “根拠/データ傾向” の確率は有意に低かった（$p < 0.05$）．ここから，同カテゴリの情報要求は繰り返し生まれる傾向にあり，スプレッドシートの理解からデータの傾向の探索に移ることは少なく，また，データの傾向の根拠を探した後にデータの傾向探索に移ることは少ないことを示している．

a） 要　　約

各情報要求が共起する行動と遷移を調査した．その結果，ユーザは解析の各段階（例えば，スプレッドーシートを探す，理解する，結果をまとめるなど）で異なるカテゴリの情報要求が生まれ，同じカテゴリの情報要求は繰り返し生まれる傾向にあった．

5 考　　察

本節では分析結果の解釈と研究の限界について考察する．

5.1 分析結果の解釈

我々はスプレッドシートを用いたデータ解析において，ユーザが主に 6 つのカテゴリの情報を欲し，さらに，各カテゴリごとに満たす難しさが異なることを明らかにした．ここから，生まれた情報要求のカテゴリを予測することは効果的なデータ解析支援につながる可能性がある．

次に，ユーザの検索行動は情報要求カテゴリによって異なることを確認した．“テキスト” と “数値” カテゴリに対して，ユーザは比較的単純な検索行動をとった．ここから，これらの情報要求を簡単に満たすことができる，もしくは単に望んだ情報が得られなくても解析を続けるほど些細なものだったのかもしれない．既存研究では，表への自動エンティティリンキング [50] や数値の単位予測 [51] の手法が提案されており，これらを活用することで，より効率的なデータ解析が可能となるであろう．ユーザは “メタデータ” と “データ傾向” カテゴリに対して，Web 検索を行わない傾向にあった．そのため，外部情報にアクセスする方法ではなく，スプレッドシート内の情報を提示する方法を開発することが有効かもしれない．この観点で，自動可視化 [1,2] や自動データ操作 [3,4] の手法が提案されている．“根拠” カテゴリに対する検索には労力がかかる傾向にあるが，データの傾向の根拠を検索する手法はほとんど提案されていない．より汎用的なデータ傾向の根拠検索は今後の課題である．“解析方法” カテゴリに対して，ユーザはあまり検索行動を取らない傾向にあるため，システム側からの情報推薦が効果的であることが示唆される．特定ドメインで解析方法を推薦する手法（プログラミングコード推薦 [52] など）は提案されているが，スプレッドシートなどのノーコードツールでの解析方法の推薦は今後の課題である．

最後に，データ解析の段階（例えば，スプレッドシートを探す，データの傾向を見つけるなど）に応じて異なるカテゴリの情報要求が生まれることが明らかになった．ここから，ユーザの行動に基づいてデータ解析の段階を特定することは，適切な情報推薦につながることが示唆される．

5.2 本研究の限界

本実験の限界として，まず，参加者が限定的であることが挙げられる．実験参加者は同じ大学に所属し，同じ言語を発話する 36 人であった．多様な発話や行動を収集するためにさまざまな分野を専攻してる学生を集めたが，参加者の行動に文化による違いが見られる可能性がある．

そして，データ解析に用いたツールの影響も挙げられる．本実験ではデータ解析ツールに Microsoft Excel のみを利用しており，データの操作が限定的である．そのため，異なるツールを用いた場合，異なるユーザ行動が観察される可能性がある．

6 おわりに

本論文ではデータ解析タスク中に生まれる情報要求の調査結果について報告した．データ収集のために，データを解析し，その結果をレポートにまとめるユーザ実験を設計し，36 人の大学院生が参加した．タスク後のインタビューで解析中の情報要求を収集し，収録映像からユーザ行動をコード化した．分析の結果，データ解析中に生まれる 6 つのカテゴリからなる情報要求の分類法を構築した．また，情報要求を満たすための検索行動は情報要求カテゴリによって異なっており，特に，データの傾向の根拠やデータ解析の方法を検索する際に，実験参加者が

困難に直面することが観察された．さらに，情報要求と共起するユーザ行動（Web ページ閲覧する，スプレッドシートを読むなど）も情報要求カテゴリによって異なってることが観察された．これらの結果に基づいて，データ解析ツールの今後の研究方針について議論した．

付　　録

1　ユーザ行動コードセット

表 A· 1: ユーザ行動のコードセット．“コード名” 列の*は既存のコードセット [9] に追加したコードを表している．

コードカテゴリ	説明	コード名
データを探す	スプレッドシートを探す行動（スプレッドシートを公開している Web ページの閲覧を含む）	データを公開している Web ページの閲覧*; 擬似検索結果のスニペット精読*; データポータル内の検索*
ダウンロード	スプレッドシートをダウンロードする行動	スプレッドシートのダウンロード*
読む	スプレッドシートを読む行動（セルやラベル，説明文を読むことなども含む）	セルを読む; ラベルを読む; 説明文を読む; フォーマットについての言及; ファイル名に着目; スプレッドシートを広げる; スプレッドシートを評価する; 特定のセルの位置を探す; 集計された統計値を探す; 注釈を読む*; スプレッドシート内検索*
質問	スプレッドシートに関する質問または推測する行動（数値に対する質問や記憶から推測することを含む）	なぜこの値になるのか？; 2 つのアイテムの違い; コンテキストの質問: スプレッドシートの編集，数値の意味，データの出典; 違いからの推測; 自身の知識からの推測; スプレッドシートに対する仮説
操作	スプレッドシートを操作する行動（列の追加や計算を含むが，可視化は含まない）	列の並べ替え; ズーム; 見やすさのための編集; 列や行の追加，削除*; 値の計算*; シートの追加*; スプレッドシート内にメモをとる*; セルのコピーペースト*; フィルタ*
可視化	グラフの作成もしくは解釈する行動	グラフ作成; グラフを読む
Web 検索	Web 検索行動（クエリ書き換えやスニペット精読を含む）	スニペット精読*; 強調スニペット精読*; クエリ書き換え*; 地図検索*; 画像検索*; Web ページ閲覧*
レポート	レポートを作成する行動（データの出典や考察，メモを含む）	データの出典を書く*; データの傾向を書く*; 知見を書く*; レポートにメモを書く*; Web を参照して何かを書く*; その他（タイトルを書く，書体の変更など）*
その他	その他の行動（タスクに関する実験者との会話など）	タスクに関する質問; デバッグ; 何かを探索する; 常識への言及; ファイル操作*

謝辞　本研究は JSPS 科研費 JP21H03775，および JST 次世代研究者挑戦的研究プログラム JPMJSP2124 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Kevin Hu, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[2] Victor Dibia and Çağatay Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE computer graphics and applications*, 39(5):33–46, 2019.

[3] Danyel Fisher, Igor Popov, Steven Drucker, and MC Schraefel. Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1673–1682, 2012.

[4] Zhe Chen, Sasha Dadiomov, Richard Wesley, Gang Xiao, Daniel Cory, Michael Cafarella, and Jock Mackinlay. Spreadsheet property detection with rule-assisted active learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 999–1008, 2017.

[5] C. Scaffidi, M. Shaw, and B. Myers. Estimating the numbers of end users and end user programmers. In *2005 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'05)*, pages 207–214, 2005.

[6] David G. Hendry and Thomas R. G. Green. Cogmap: a visual description language for spreadsheets. *J. Vis. Lang. Comput.*, 4:35–54, 1993.

[7] Jácome Cunha and Diogo Canteiro. Spreadsheetdoc: An excel add-in for documenting spreadsheets. 2015.

[8] Andrea E. Kohlhase and Michael Kohlhase. Semantic transparency in user assistance systems. In *Proceedings of the 27th ACM International Conference on Design of Communication*, SIGDOC '09, page 89–96, New York, NY, USA, 2009. Association for Computing Machinery.

[9] Sruti Srinivasa Ragavan, Advait Sarkar, and Andrew D Gordon. Spreadsheet comprehension: Guesswork, giving up and going back to the author. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2021.

[10] Sudeepa Roy. Toward interpretable and actionable data analysis with explanations and causality. *Proc. VLDB Endow.*, 15(12):3812–3820, aug 2022.

[11] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA, 2002.

[12] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19, 2004.

[13] Eytan Adar, Jaime Teevan, and Susan T Dumais. Resonance on the web: web dynamics and revisitation patterns. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1381–1390, 2009.

[14] B Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. An intent taxonomy for questions asked in web search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 85–94, 2021.

[15] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1245–1248, 2018.

[16] Nikitha Rao, Chetan Bansal, Subhabrata Mukherjee, and Chandra Maddila. Product insights: Analyzing product intents in web search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2189–2192, 2020.

[17] Long Chen, Dell Zhang, and Levene Mark. Understanding user intent in community question answering. In *Proceedings of the 21st international conference on world wide web*, pages 823–828, 2012.

[18] Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. Function-based question classification for general qa. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1119–1128, 2010.

[19] Farnaz Jahanbakhsh, Elnaz Nouri, Robert Sim, Ryen W White, and Adam Fourney. Understanding questions that arise when working with business documents. *arXiv preprint arXiv:2203.15073*, 2022.

[20] Yinglong Zhang, Rob Capra, and Yuan Li. An in-situ study of information needs in design-related creative projects. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 113–123, 2020.

[21] Nils Pharo and Kalervo Järvelin. The sst method: a tool for analysing web information search processes. *Information Processing & Management*, 40(4):633–654, 2004.

[22] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, volume 37, pages 18–28. ACM New York, NY, USA, 2003.

[23] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jing-

tao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 463–472, 2016.

[24] Neil O'Hare, Paloma De Juan, Rossano Schifanella, Yunlong He, Dawei Yin, and Yi Chang. Leveraging user interaction signals for web image search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 559–568, 2016.

[25] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 23–32, 2015.

[26] Sosuke Shiga, Hideo Joho, Roi Blanco, Johanne R Trippas, and Mark Sanderson. Modelling information needs in collaborative search conversations. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 715–724, 2017.

[27] Bogeum Choi, Austin Ward, Yuan Li, Jaime Arguello, and Robert Capra. The effects of task complexity on the use of different types of information in a search assistance tool. *ACM Transactions on Information Systems (TOIS)*, 38(1):1–28, 2019.

[28] Xi Niu, Xiangyu Fan, and Tao Zhang. Understanding faceted search from data science and human factor perspectives. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–27, 2019.

[29] Laura M Koesten, Emilia Kacprzak, Jenifer FA Tennison, and Elena Simperl. The trials and tribulations of working with structured data: -a study on information seeking behaviour. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1277–1289, 2017.

[30] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[31] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[32] Jing Nathan Yan, Ziwei Gu, and Jeffrey M Rzeszotarski. Tessera: Discretizing data analysis workflows on a task level. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[33] Lei Han, Tianwa Chen, Gianluca Demartini, Marta Indulska, and Shazia Sadiq. On understanding data worker interaction behaviors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 269–278, 2020.

[34] Kamalasen Rajalingham, David R Chadwick, and Brian Knight. Classification of spreadsheet errors. *arXiv preprint arXiv:0805.4224*, 2008.

[35] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207, 2022.

[36] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, pages 1165–1188, 2012.

[37] Pall Rikhardsson and Ogan Yigitbasioglu. Business intelligence & analytics in management accounting research: Status and future focus. *International Journal of Accounting Information Systems*, 29:37–58, 2018.

[38] NoorUl Ain, Giovanni Vaia, William H DeLone, and Mehwish Waheed. Two decades of research on business intelligence system adoption, utilization and success–a systematic literature review. *Decision Support Systems*, 125:113113, 2019.

[39] José Ricardo López-Robles, Jose Ramón Otegi-Olaso, I Porto Gómez, and Manuel J Cobo. 30 years of intelligence models in management and business: A bibliometric review. *International journal of information management*, 48:22–38, 2019.

[40] Andrea Kohlhase, Michael Kohlhase, and Ana Guseva. Context in spreadsheet comprehension. In *SEMS@ ICSE*, pages 21–27, 2015.

[41] Zhilan Zhou, Ximing Wen, Yue Wang, and David Gotz. Modeling and leveraging analytic focus during exploratory visual analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[42] Andrea E Kohlhase and Michael Kohlhase. Semantic transparency in user assistance systems. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 89–96, 2009.

[43] Eirik Bakke, David Karger, and Rob Miller. A spreadsheet-based user interface for managing plural relationships in structured data. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2541–2550, 2011.

[44] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. How data workers cope with uncertainty: A task characterisation study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3645–3656, 2017.

[45] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–12, 2018.

[46] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 3363–3372, 2011.

[47] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 65–74, 2011.

[48] Kelsey Urgo, Jaime Arguello, and Robert Capra. The effects of learning objectives on searchers' perceptions and behaviors. In *Proceedings of the 2020 ACM SIGIR on international conference on theory of information retrieval*, pages 77–84, 2020.

[49] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232, 2014.

[50] Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Feng Jiang, Andy Gordon, and Chin-Yew Lin. Linkingpark: An automatic semantic table interpretation system. *Journal of Web Semantics*, 74:100733, 2022.

[51] Jack Williams, Carina Negreanu, Andrew D Gordon, and Advait Sarkar. Understanding and inferring units in spreadsheets. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–9. IEEE, 2020.

[52] Sam Lau, Sruti Srinivasa Srinivasa Ragavan, Ken Milne, Titus Barik, and Advait Sarkar. Tweakit: Supporting end-user programmers who transmogrify code. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

研究内容からの研究データ検索

山﨑　基数† 冨浦 洋一† 清水 敏之†

†九州大学大学院統合新領域学府ライブラリーサイエンス専攻　〒819-0395 福岡市西区元岡 744 番地
E-mail: †yamasaki.motokazu.491@s.kyushu-u.ac.jp,

††{shimizu.toshiyuki.457,tomiura.yoichi.802}@m.kyushu-u.ac.jp

あらまし 近年，膨大な量の研究データが生成されるようになり，研究データを活用するためにはそれらを効率的に検索することが重要となっている．従来の検索エンジンやキーワードベースの検索手法では，ユーザが適切なキーワードやフレーズを入力する必要があり，目的のデータに関する詳細な情報を持っていない場合は十分な検索結果を得ることが難しい．本研究では，ユーザが研究データを検索する際に，ChatGPT に自身の研究の概要を伝えることで，その研究に役立つ研究データを提案させることを考えた．実際には，ChatGPT に対してある研究データを引用した論文の概要を入力し，その論文の研究を行うのに必要なデータを提案させ，提案された研究データの中に引用された研究データが含まれているかを調査した．

キーワード Dataset search, Research data, ChatGPT, 研究データ，研究データ検索

1. はじめに

近年，インターネット上で論文や研究データを公開する流れが進み，それを利用した新たな研究が行われることが期待されている．それに伴って研究データの検索についても研究が行われている[1].

論文の場合は全文検索を行うことができるが，研究データは数値や画像の集合であるため全文検索を行うことができない．研究データの検索を行う際には，Google Dataset Search[2]などを利用してキーワード検索を行うことが一般的である．研究データは表や画像の集合である場合もあるので研究データに付与されたメタデータに対してのキーワード検索を行う[1].

しかし，研究者は必ずしも自分の欲しい研究データの分野に精通しているとは限らず，検索の際に適切なキーワードを用いることが困難である場合も多い．また，研究データに必ずしも十分なメタデータが付与されているとは限らず，キーワード検索では適切な検索結果を得ることが困難な場合もある[1].

研究データに付与されているメタデータはデータ作成者が自由につけるため，品質は不揃いであり，十分に整備されていない場合がある．メタデータの品質が低い，または欠落していると研究データの検索に影響を及ぼす[3]．また，メタデータやキーワード検索の問題として表記ゆれの問題がある．簡単な表記ゆれであれば類語検索などで解消できるが，例えば「燃費」,「燃料当たりの走行距離」などの言いかえとなると単純な表記ゆれとして処理することはできない[4].

これらの研究データ検索における問題点を解消するために様々な研究が行われている．表記ゆれについては表中の数値を表すカラム名などについて同一の属性や上位下位の関係にあるものを抽出する研究などがある[4]．また，キーワード検索の問題点として検索ワードがもつ属性名と属性値の表現が異なるという問題があり，それに対してキーワードが持つ属性を同定するための研究が行われている[5][6]．これらの研究はどれもデータ作成者やデータの提供者の取り組みである．

本研究ではユーザ側の工夫ができないか，欲しい研究データの専門家ではないユーザの知識と目的の研究データとのギャップを埋める方法がないか考える．その方法の 1 つとしてここ数年で情報検索に使われるようになってきた LLM（大規模言語モデル）[7][8]の利用を検討する．LLM を利用する手段の 1 つとして ChatGPT[9]を用いることを提案する．

ChatGPT とは OpenAI によって公開された人工知能チャットボットである．LLM（大規模言語モデル）に基づいており，人間との対話が成立しているかのような文章を生成することができる．ChatGPT を利用することでユーザは簡単に LLM の恩恵を受けることが出来ると考えた．研究者が有用なキーワードを思いつくことができない場合に，目的の研究データを見つけるための補助として ChatGPT を利用することを検討する．研究者が行いたい研究内容を文章で表現して ChatGPT に与え，その研究で役に立つ研究データの提案を求めることで，その研究を行う際に利用可能な研究データの情報を得ることができると考えられる．

研究データ検索と直接のつながりはないが，文献の検索に ChatGPT を利用するという研究も行われている[10][11]．ChatGPT の回答が常に正しいとは限らず，存在しない研究データや誤った研究データを提案する可能性もあるが，データ名や URL，DOI を活用することで確認することができる．

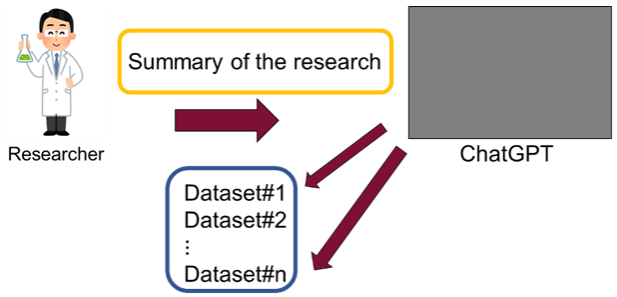


図 1　ChatGPT による研究データ検索
（実際の利用シーン）

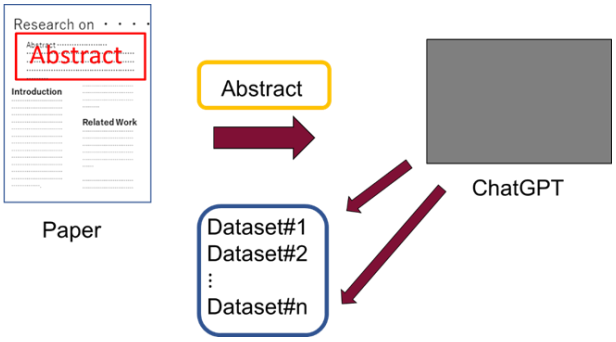


図 2　ChatGPT による研究データ検索
（既存の論文によるシミュレーション）

本研究では，研究データ検索における LLM 利用の可能性を確認するための手段として ChatGPT を使用した研究データ検索の有効性を確認する実験を行った．

まず，実際の利用シーンについて説明する．図 1 に示すように，はじめに研究者は自分の行いたい研究を要約したものを用意する．研究内容の要約を ChatGPT に入力し，その研究に役に立ちそうな研究データを提案させる，という状況を想定する．しかし，この方法の有用性を確かめるためには多くの研究者に実験に参加してもらう必要があり，現実的には難しい．そこで我々は次のような方法でこの手法について実験を行った．

まず，図 2 に示すように研究者の行いたい研究内容の要約の代わりに既存の論文の Abstract を用意した．その論文には研究データ及びその研究データに付随するデータ論文が引用されている．そのデータ論文を研究に役立ちそうな研究データの代わりとした．つまり，既存の論文の Abstract を ChatGPT に入力した場合に，ChatGPT がその論文で引用されているデータを提案できるかを調査した．これにより，研究内容がまとまったものを入力した場合にその内容から研究に役立つデータを提案できるかの評価を行った．

研究データを引用した既存の論文はこれから行う研究の概要を入力するという状況を再現するために ChatGPT が学習に使用していない論文でなければならない．LLM を利用した検索サービスとしてすでに Microsoft Copilot[7]や Gemini[8]などがあるがこの 2 つは検索エンジンと結びついておりリアルタイムの情報を収集することが出来るので本研究での実験には適さないと判断した．

2. 実験方法

2.1. 研究データ及び論文の収集方法

ChatGPT を利用して，既存の論文の Abstract からその論文で使われている研究データを検索する実験を行った．ChatGPT は無料版である GPT3.5 を利用した．

まず，Google Scholar を中心として Web 上で研究データに付随する論文（データ論文）を 50 個収集した．Google Scholar 上で「分野名＋dataset」と入力し，検索結果に出てきた論文の中からデータ論文に該当するものを収集した．例えば，“Medicine dataset”などである．集めた研究データの分野としては医学，コンピュータビジョン，地理，エネルギー分野などが主である．図 3 に示すように集めた研究データに付随するデータ論文を引用している論文を 1 つの研究データにつき 2 本集めた．つまり，50 個の研究データ（データ論文）と 100 本の論文を収集し，100 組の論文と研究データのペアに対して実験を行った．

本研究では，これから行いたい研究において役に立ちそうな研究データを ChatGPT に提案させるという状況を想定した．そのため，研究データを引用している論文は ChatGPT が学習に使用していないものを使用することでその状況の再現を行うことにした．本実験を行った時点で ChatGPT は 2022 年 1 月までの情報を学習していた．そのため，論文は 2022 年 2 月以降に公開されたものを選定した．

また，研究データは ChatGPT が学習しているものに限定するため，データに付随する論文の発行年が 2021 年までのものを集めた．集めた研究データに対して，ChatGPT がデータについて学習していそうかを調べた．ChatGPT に具体的なデータセットの名前を入力し，その回答の内容から ChatGPT が研究データを学習しているか調査することにした．具体的には，ChatGPT に “Are you familiar with [Dataset Name]?” と質問を行った．以上の質問を行った際に ChatGPT が学習していなさそうな研究データについては研究データの検索が困難と判断し，除外した．例えば，“As of my last knowledge update in January 2022, I'm not aware of a specific concept or standard called “[Dataset Name]” However, it's possible that developments or new standards have emerged since then.” といった返答があった．

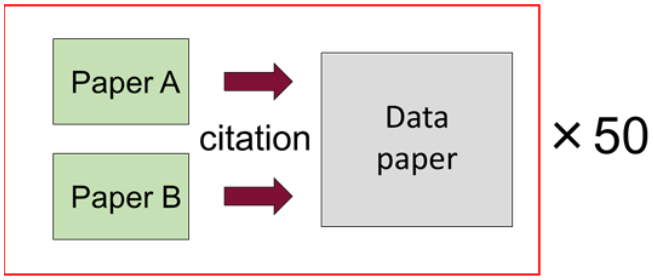


図 3　研究データ及び論文の収集方法

逆に ChatGPT が研究データを学習していると考えられる場合はその研究データについて簡単な説明を回答した．ChatGPT は言語モデルに基づいて回答を確率的に生成するため内容の正確性は保証されないが，その研究データの説明として明らかに異なる説明を回答しなければ研究データの提案には支障がないと考え，大きく異なる内容を説明しない限りはその研究データを使用することにした．

また，研究データを引用している論文を収集する際に次のような論文は除外した．

- Abstract に引用している研究データ名が直接記載されているもの
- データ論文は引用されているものの，論文内で研究データが活用されていないもの

Abstract に研究データ名が直接記載されていると，研究者が自身の欲しい研究データの詳細を知らないという状況を模倣できないと考えたからである．また，研究者が自身の研究の役に立つデータを探すという状況を再現するために研究データが論文内で活用されているものを選定した．

研究データが活用されている論文として次のようなものを選んだ．

- 提案するモデルの事前学習データとして研究データを活用しているもの
- 提案するフレームワークのベンチマークとして研究データを活用しているもの
- 研究データの情報を活かして論文を執筆しているもの

2.2. Abstract 及び命令文の入力

ChatGPT に既存の論文の Abstract とデータセットを提案させるための命令文を入力した．具体的には，ChatGPT に論文の Abstract と“Give me 10 datasets to help me conduct this study.”という文言を入力し，論文の Abstract からその論文の研究を行うのに役に立ちそうな研究データを 10 個提案させた．その 10 個の中に論文が引用している研究データが含まれているかを調査した．また，目的の研究データが含まれていなかった際に“Give me 10 more.”と追加で入力し，最大で 30 個のデータセットを提案させた．一度に 30 個のデータの提案を求めるとデータの個数が合わないなどのエラーが生じることがあったため，10 個ずつ提案させるようにした．最大で 30 個提案させたデータセットの中に目的の研究データが含まれていれば「発見」，そうでない場合は「未発見」とし，その数をカウントした．以上の手順を 100 本の論文に対して行い，提示された研究データの中に目的のデータが含まれているか調査した．

実験を行っていく中で，“You are an excellent

表 1　ChatGPT による研究データ発見数

発見（上位 30 件）	未発見
67	33

図 4　ChatGPT による研究データ検索の入力例

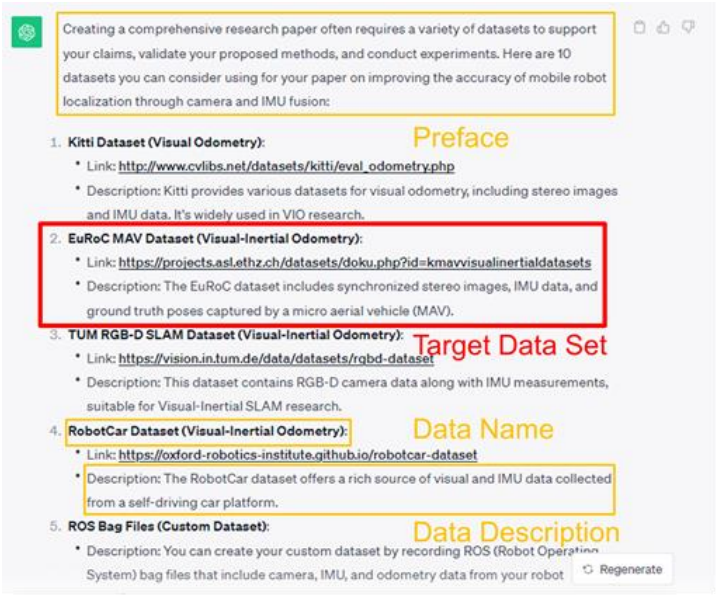


図 5　ChatGPT による研究データ検索の出力例

searcher.”のようなふるまいのプロンプトを用いることや“I would like to conduct research on the following.”などの補足説明を入力することを試行した．しかし，“Give me 10 datasets to help me conduct this study.”のようなシンプルな質問のみを入力した場合と回答に特に違いが見られなかったため，シンプルな質問のみを入力することにより実験を行った．

3. 実験結果

3.1. 研究データ発見数

論文と研究データのペア100組に対して実験を行った結果，表1に示すようにChatGPTによって67個の研究データが提案された．

3.2. 回答内容

図4には入力例を，図5にはChatGPTによる回答例を載せている．図5のようにChatGPTは結果を提示する際に以下の内容を回答した．

・前置き（研究内容の分野，研究内容に役立つデータセットの特徴）
・各データセット名
・各データセットの簡単な説明や用途

また，データセット名だけでなくデータセットにアクセスするためのURLも提示することがあった．

4. 考察

4.1. キーワード検索との比較

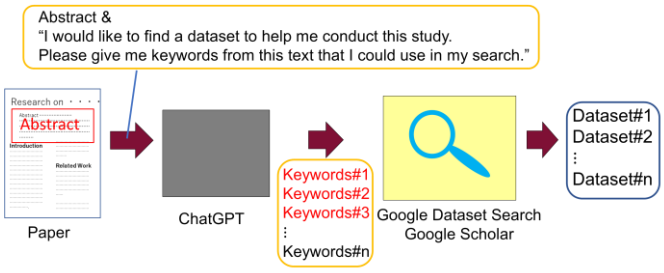


図6　キーワードを用いた研究データ検索の方法1

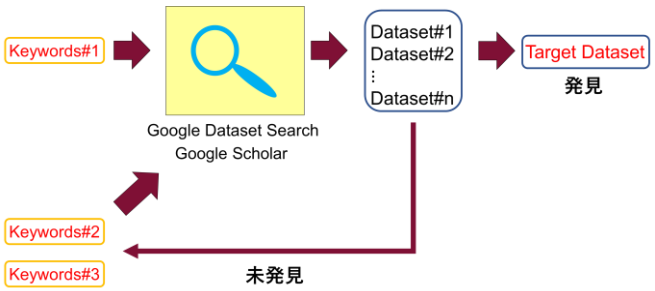


図7　キーワードを用いた研究データ検索の方法2

図6に示すようにChatGPTの提案によるデータの検索と比較するためにGoogle Dataset SearchとGoogle Scholarのそれぞれでキーワード検索を行った．使用した論文と研究データはChatGPTによる研究データ検索で用いたものと同じものを使用した．検索に用いるキーワードはChatGPTを用いて論文のAbstractから抽出した．キーワードの選定に関して公平性を保つためにChatGPTを用いて論文のAbstractからキーワードを抽出することにした．まず，ChatGPTに研究データを引用している論文のAbstractと“I would like to find a dataset to help me conduct this study. Please give me keywords from this text that I could use in my search.”という「研究者が行う予定の研究に役立つ研究データ」を検索する際のキーワード（キーフレーズ）を論文のAbstractから抽出させるための文章を入力した．例えば，音声と顔画像から話している顔の映像を作り出す技術について書かれた論文のAbstractをChatGPTに入力したところ，“Audio-visual dataset”，“Face keypoints dataset”などのフレーズを提示した．提案されたキーワードの上位3つをGoogle Dataset SearchとGoogle Scholarのそれぞれに入力し，研究データの検索を行った．図7に示すように検索はChatGPTによって抽出したキーワードの上位1つ目から順番に試し，目的の研究データが見つかった場合にはそれ以降の他の検索ワードでの検索は行わなかった．つまり，3つのいずれかのキーワードで目的の研究データが見つかるか調査を行った．

Google Dataset Searchでの実験においては「発見（上位30件）」，「発見（上位31件以下）」，「未発見」の3つに分けて評価を行った．Google Dataset Searchは1ページの中に検索結果がすべて表示されるためページ内検索を用いてデータセット名を探すことで31件以降の検索結果についても調査を行った．実験を行う前にはGoogle Dataset Search上にデータ論文と紐づけられた研究データが存在することを確認した．確認についてはデータセット名の一致や作成日，Description，データ論文へのリンクの有無などをチェックし，一致するデータセットや研究データであると判断できた場合には「データがGoogle Dataset Searchに存在する」という扱いにした．Google Scholarの場合にはGoogle Scholarで目的のデータ論文を見つけることが出来るか，「発見（上位30件）」「未発見」の2つに分けて評価を行った．Google Scholarで検索を行う際は「検索ワード＋dataset」といったように“dataset”という文言を付加することでデータセットに付随する論文（データ論文）が検索しやすいように工夫した．ChatGPTが抽出した元の

表 2　Google Dataset Search による研究データ発見数

発見 （上位 30 件）	発見 （上位 31 件以下）	未発見
17	6	77

表 3　Google Scholar による研究データ発見数

発見（上位 30 件）	未発見
23	77

表 4　研究データ発見数の比較

方法	発見 （上位 30 件）	未発見
ChatGPT	67	33
Google Dataset Search	17	83
Google Scholar	23	77

表 5　キーワードの順番ごとの研究データ発見数（上位 30 件）

方法	キーワード 1 個目	キーワード 2 個目	キーワード 3 個目
Google Dataset Search	12	2	3
Google Scholar	13	6	4

検索ワードに“data”などの単語がついていた場合は一律“dataset”に置き換えて検索を行った．

表 2 に Google Dataset Search での実験結果を示す．検索結果の上位 30 件に含まれている研究データは 17 個，上位 31 件以下は 6 個，未発見は 77 個という結果になった．表 3 には Google Scholar での実験結果を示す．上位 30 件以内に発見されたデータ論文は 23 件，未発見は 77 件であった．

表 4 に ChatGPT による研究データの提案とキーワード検索との比較を示している． ChatGPT による研究データの提案における研究データの発見数が 67 個であるのに対し，キーワード検索の発見数は Google Dataset Search で 17 個，Google Scholar で 23 個と少なく，ChatGPT による研究データの検索がキーワード検索に比べて目的の研究データにたどり着きやすいことが分かった．理由としてキーワード検索は単語の完全一致あるいは類語辞書などを用いたマッチングであるのに対し，ChatGPT を利用した研究データの検索では，単語は分散表現で置き換えられるので，近い意味の語が類義辞書より豊富であり，その近さも表すことができる．また，ChatGPT は LLM（大規模言語モデル）に基づいているため，語や文から連想することも可能であるということが理由として考えられる．

表 5 に Google Dataset Search と Google Scholar それぞれで検索結果上位 30 件において何個目のキーワードで研究データが発見できたかを示している．どちらも 1 つ目のキーワードで見つかっている割合が多いことが分かる．キーワードは Abstract から抽出していることを踏まえると，Abstract の内容が検索結果に影響している可能性が高く，見つかる場合と見つからない場合がはっきり分かれている傾向となった．

今回は簡易的な比較であったため，比較手法についてもさらに検討する予定である．

4.2. 見つかりやすい研究データの特徴

本研究での実験を行う中で ChatGPT は略称や固有名詞を含む研究データ名を提案しやすい傾向が見られた．例えば，“The KITTI Dataset（Karlsruhe Institute of Technology and Toyota Technological Institute）”[12]，“VoxCeleb”[13]などが挙げられる．

4.3. 質問文の工夫

ChatGPT によって目的の研究データが見つからなかった場合に ChatGPT に目的の研究データの特徴を与えることでデータを発見できる場合があった．例えば，Computer Vision と Pattern Recognition に関する分野の論文で“UCF101”[14]という動画のデータセットが引用されている場合に，ChatGPT に研究データの提示を求めると画像形式のデータセットを中心に提示した．そこで，動画形式のデータセットの提示を求めると動画のデータセットが中心に提示され，目的の“UCF101”というデータセットを見つけることができた．

このように，質問を工夫することでより多くの研究データを発見することができる可能性がある．ただし，今回は特徴を筆者（答えを知っている人間）が記述したため，論文執筆者が有効な特徴を ChatGPT に尋ねることができるのかという問題がある．

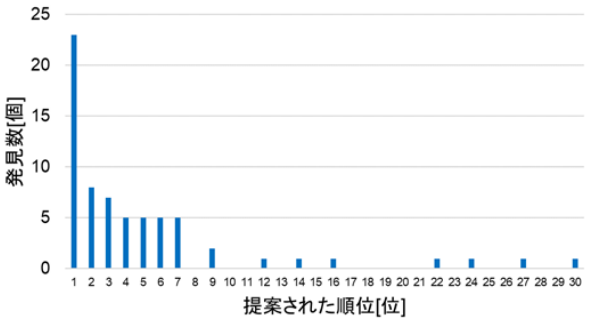


図 8　提案された順位における研究データ発見数

4.4. ChatGPT によって提案された研究データの順位

図 8 に示すように本実験では ChatGPT によって目的の研究データが提案された場合に，そのデータが何番目に提案されたかを記録した．ChatGPT の出力は入力が同じ場合でも毎回異なる可能性があるので具体的な細かい数値に大きな意味はないが，全体の傾向として目的の研究データが見つかった場合は最初のやり取り，つまり 10 番目以内に提案されることが多いという傾向にあった．具体的には発見された 67 個の研究データのうち 60 個が 10 番目以内に提案された．このことから必要な情報や関連の深い情報が Abstract に含まれていた場合は ChatGPT が特に有用に働き，含まれていない場合は ChatGPT による提案は急激に困難になる可能性がある．また，実験 I や実験 II で提案させる研究データを 30 個としたがそれを 40 個，50 個と増やしても研究データの発見数は大きく上昇しないと考えられる．

4.5. 利用する LLM の種類

今回は無料版の ChatGPT を使用したため，有料版の ChatGPT を使用した場合と結果が異なる可能性がある．無料版は GPT-3.5 というバージョンを利用できるが有料版では GPT-4.0[15]という新たなバージョンが利用できる．今後は有料版での実験も行うつもりである．

5. おわりに

本研究では研究データ検索における LLM（大規模言語モデル）の利用可能性について調査を行った．LLM 利用の手段として ChatGPT を選択した．具体的には，ChatGPT に研究内容の要約を伝えた際に，ChatGPT がその研究に役立つデータセットをどの程度提案することができるのかシミュレーションを行った．実験の結果，実験で使用した 100 本の論文のうち，67 本の論文において Abstract から引用されている研究データを発見することができた．このことから，LLM に基づいている ChatGPT の利用が研究者の研究データ検索の補助になる可能性があることを確認した．今後は質問文の工夫を行い，より目的のデータにたどり着きやすくするための文章や回答内容について更なる分析を行う予定である．また，見つかった研究データや ChatGPT が頻繁に提案した研究データのメタデータの分析を行うことで研究データを見つけやすくするためのメタデータについて知見を得られる可能性がある．このことは研究データを従来のキーワード検索で見つけやすくすることにも寄与する可能性がある．

参　考　文　献

[1] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, Paul Groth.: Dataset search: a survey. The VLDB Journal 29, 251–272 (2020)

[2] Dan Brickley, Matthew Burgess, Natasha Noy.: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In: The World Wide Web Conference (WWW '19), 1365–1375 (2019)

[3] J. Umbrich, S. Neumaier and A. Polleres.: Quality Assessment and Evolution of Open Data Portals. 2015 3rd International Conference on Future Internet of Things and Cloud, Rome, Italy, pp. 404-411 (2015)

[4] 藤岡 周平，加藤 誠，吉川 正俊．表からの量的データ属性間の関係抽出．情報処理学会論文誌データベース（TOD） 13(3)，pp. 10-21 (2020)

[5] Wiradee Imrattanatrai, Makoto P. Kato, Masatoshi Yoshikawa. Identifying Entity Properties from Text with Zero-shot Learning. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019). pp. 195-204 (2019)

[6] Makoto P. Kato, Wiradee Imrattanatrai, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka. Context-guided Learning to Rank Entities. Proceedings of the 42nd European Conference on IR Research (ECIR 2020), pp. 83-96 (2020)

[7] Microsoft Copilot https://www.microsoft.com/ja-jp/microsoft-copilot

[8] Gemini https://gemini.google.com/app

[9] ChatGPT https://openai.com/blog/chatgpt

[10] Haman, Michael, and Milan Školník.: Using ChatGPT to conduct a literature review. Accountability in Research, 1-3 (2023)

[11] Alessia McGowan, Yunlai Gui, Matthew Dobbs, Sophia Shuster, Matthew Cotter, Alexandria Selloni, Marianne Goodman, Agrima Srivastava, Guillermo A. Cecchi, Cheryl M. Corcoran.: ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. Psychiatry Research, 115334 (2023)

[12] Arsha Nagrani, Joon Son Chung, Andrew Zisserman.: VoxCeleb: a large-scale speaker identification dataset. arXiv preprint arXiv, 1706.08612 (2017)

[13] Geiger A, Lenz P, Stiller C, Urtasun R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research 32(11), 1231-1237 (2013)

[14] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv, 1212.0402 (2012)

[15] OpenAI.: GPT-4 Technical Report. arXiv, 2303.08774 (2023)

大規模言語モデルによる文書補強とリランキングによる統計データ検索

黒川　博生† 　宮森　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2153489,miya}@cc.kyoto-su.ac.jp

あらまし　統計データは, 政府等が保有するオープンデータの一種であり, 近年, 社会問題となっているフェイクニュースに対処するための事実確認 (ファクトチェック) への活用をはじめ, 有効活用するためのアドホック検索基盤の重要性が高まっている. 従来, 統計データ検索では, メタデータを用いた BM25 などによるランキング手法が提案されているが, 文書長の短さや表記ゆれなどによる検索漏れ等により, 十分なランキング性能を達成できていない. 本稿では, 大規模言語モデルによる文書補強とリランキングによる統計データのアドホック検索手法を提案する. まず, 統計データから抽出された見出し, 行名, 列名, 値に基づき, その内容説明を大規模言語モデルで生成することでメタデータを補強した文書を作成する. 次に, 大規模言語モデルを用いたクエリ拡張を行い, 補強された文書に対するランキングを行う. 最後に, 大規模言語モデルを用いて意味内容の類似に基づくリランキングを行う. 実験では, 提案手法と従来手法によるランキング結果を比較評価する.

キーワード　統計データ検索, メタデータ, 文書補強, クエリ拡張, 大規模言語モデル, 情報検索

1　はじめに

ソーシャルメディアにおける自由なユーザ発信が普及した結果, 虚偽の情報を含む言説が蔓延し, 個人や社会に対する大きな脅威となっている. これに対処するため, 事実確認 (ファクトチェック) の重要性が認識され, 現在, 世界中で様々な事実確認の取り組みが進められている.

フェイクニュースの事実確認を行う際, 客観的なデータとの整合性の有無を確認することで, 事実確認の判断根拠の一つとする方法をとることができる. 本稿では, 客観的なデータとして, 政府等が保有する統計データを用いる場合を想定し, 統計データのアドホック検索 (統計データ検索と呼ぶ) に焦点を当てる. 統計データ検索は事実確認を行う際の真偽判定の土台となり, 根拠となる文書を的確に取得できることが重要となる.

本稿で扱う文書は, 政府等が保有する統計データから抽出したメタデータ, 統計データ本体から構成されており, 以後, 統計文書と呼ぶこととする. メタデータは文書長が短く, 統計データ本体もタイトルなどを除くとほとんどの場合, 数値で構成されており, 表構造も統一されていないという特徴がある.

統計データ検索ではこれまでにも様々な研究が行われている. 例えば, 統計文書中のメタデータだけでなく, 統計データに含まれる表のヘッダ情報を抽出し, カテゴリを付与することで, BM25 によるランキングを改善する手法 [1] や, BM25 によるランキング後に BERT によるリランキングを行う手法 [3] などが提案されている. しかし, 従来手法は, 統計文書中のメタデータやタイトルなどの言語情報に焦点を当てており, 統計データ中の数値の並びを積極的に活用した例はほとんど見られない. 統計データ中の数値の並びが持つ意味合いを文書補強やクエリ拡張に適切に反映させることができれば, メタデータの文書長の短さや表記揺れによる検索漏れ等が改善され, より関連性の高い統計文書を検索できるのではないかと考えた.

そこで本稿では, 大規模言語モデルを用いた文書補強とリランキングによる統計データ検索手法を提案する. 大規模言語モデルとは, 大規模なコーパスで事前学習された大規模なパラメータをもつ言語モデルのことであり, 様々なタスクで顕著な性能を示している. また, 大規模言語モデルに入力するプロンプトの構成方法についても出力結果に大きく影響するため盛んに研究されている. 本稿では, 大規模言語モデルを, 統計文書の補強, クエリ拡張, ランキングリストのリランキングの各段階で活用する手法を提案する.

提案手法の構成図を図 1 に示す.

まず, 統計文書の補強では, 統計文書の内容を利用したプロンプトを作成し, 大規模言語モデルに入力すること補強文書を生成する. 具体的には, 統計データから見出し, 行名, 値からデータフレームを作成し, データフレームに基づいて補強文書を作成する. ここで, 見出しとは統計データ内の表全体を説明するテキストを表し, 行名, 列名とは, 統計データ内の具体的な数値 (値) の並びが表す内容を説明するテキストを表す. 本稿では, データフレームから列ごとにプロンプトを作成し, 大規模言語モデルに入力することで補強文書を生成する.

次に, 補強文書を生成後, クエリの類似語を出力するよう促すプロンプトを大規模言語モデルに入力し, クエリ拡張を行う.

その後, メタデータ及び補強文書を結合した文書に対して, クエリに対するランキング結果を BM25 で取得する. また, ランキング後はクエリ拡張で生成された複数のランキングリストを統合するため, 統計文書ごとに順位の平均を取り, ソートすることで 1 つのランキングリストに統合する.

さらに, 得られたランキングリストを, 意味内容の類似によってリランキングする. リランキングは, 大規模言語モデルを用いた Pairwise Ranking Prompting [11] を用いて実現する. これは, 2 つの文書とクエリ, および, いずれの文書がクエリによ

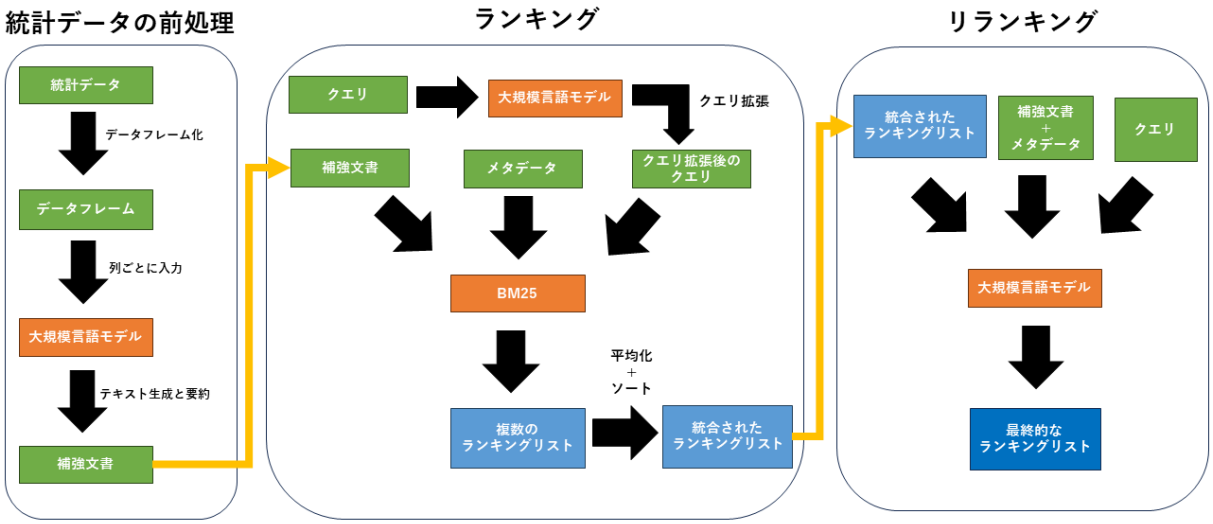


図 1　提案手法の処理手順

り関連しているのかを問う質問を大規模言語モデルに入力することで, 与えた 2 文書の関連性比較結果を得る方法である.

提案手法は, メタデータを活用するだけではなく, 統計データ内の数値の並びを文書化することでランキングに活用している点や, 文書とクエリの意味内容の類似を大規模言語モデルに判定させるリランキング手法を提案している点で, 従来手法だけでは得られなかった関連度の高い文書が得られることを目指している.

実験では, 提案手法と従来手法のそれぞれのランキング結果を比較し, どのような性能の差があるのかを明らかにする.

本論文の貢献は以下のとおりである.

（1） 統計データ中の数の並びも活用した統計文書の補強手法を提案している点,

（2） 大規模言語モデルのプロンプト技術に着目した検索改善手法を提案している点,

（3） 統計データ検索における提案手法の有効性を実験で検証している点.

本稿の構成は以下の通りである. 2 節では, 統計データ検索, 大規模言語モデルを用いた検索, 統計データの文書補強に関する関連研究を紹介する. 第 3 節では, 提案手法の問題設定, 処理手順を詳しく説明する. 第 4 節では, 実験の目的と方法, 実験結果と考察を示し, 第 5 節では, 結論と今後の課題をまとめる.

2　関 連 研 究

本節では, 統計データ検索, 大規模言語モデルを用いた検索, 統計データの文書補強のそれぞれについての関連研究を紹介する.

2.1　統計データ検索に関する研究

統計データ検索の研究では, 大規模言語モデルを用いない手法と大規模言語モデルを積極的に活用する手法に分類できる. まず, 大規模言語モデルを用いない手法として, 統計データの行と列のヘッダを抽出しメタデータを補強した後, 各統計文書にカテゴリを付与し, BM25 でランキングする手法 [1] や, 統計データの構造だけではなく, クエリの構造まで考慮し, BM25 を拡張した BM25FF によるランキングを行う手法 [2] が挙げられる. 提案手法は, 統計データ検索に大規模言語モデルを用いている点でこれらの研究とは異なる.

また, 大規模言語モデルを積極的に活用する手法として, クエリから地域情報や年などの情報を取り除き, BM25 によってランキングを行い, BERT を用いてリランキングする手法 [3] や, 統計文書のメタデータに含まれるタイトルと説明に対するクエリの注意ベクトルを計算し, これと BERT による出力ベクトルを入力した MLP で関連度スコアを計算しランキングする手法 [4] が挙げられる. 提案手法では, 大規模言語モデルを活用している点は共通しているものの, 大規模言語モデルによるテキスト生成を行っている点でこれら研究とは異なる.

2.2　大規模言語モデルを用いた検索に関する研究

大規模言語モデルを用いた検索に関する研究として, 系列変換モデルを用いてクエリから直接検索文書をデコードする Differentiable Search Index と呼ばれる手法 [5] が提案されている. また, BM25 で得られたランキング結果を, ポイントワイズのリランキング, ペアワイズのリランキングと段階的に適用する枠組み [6] や, 系列変換モデルを用いることで, 関連性スコアを直接出力するだけでなく, ポイントワイズやリストワイズの方法でランキング学習できる手法 [7] が考案されている. これらの手法は, アドホック検索を対象としているものの, 提案手法は, 統計文書を対象としている点が異なる.

また, プロンプト技術を用いたランキング, リランキング手法として, クエリを入力した GPT にクエリの回答が含まれる URL を出力させ, その URL 先の文書をパッセージ化しランキングする手法 [8] や, クエリに対するパッセージ検索のリランキングをポイントワイズに行い, 言語モデルの様々なタスク性能の一つとして評価した研究 [9] が挙げられる. また, ゼロショットでのリランキングに用いるプロンプトを, 大規模言語モデル

のパラメータ更新なしにプロンプト候補を逐次的に更新することで最適化する手法 [10] や, クエリと 2 文書をプロンプトに用いることで, 大規模言語モデルに対する負荷を軽減し, ランキング性能を向上させるを指向したペアワイズランキングプロンプティング (PRP) [11], 候補文書に対して並べ替えられた文書 ID のリストを直接生成するリストワイズランキング [12] が挙げられる. これらの手法は, 大規模言語モデルを用いてテキスト生成している点で共通しているものの, 提案手法は, 統計文書を対象としている点が異なる.

他に, 大規模言語モデルの回答をより的確なものとするため, 検索技術を組み合わせ, 外部で検索した情報を取り入れることによって回答の正確性を向上させる RAG と呼ばれる手法 [13] がある. また, 外部コーパスや外部知識を検索することで大規模言語モデルの回答を向上させる枠組みのためのツールキット [14] も提供されている. これらは, クエリに対するテキスト回答の出力を行うものであり, ランキングリストを作成することが目的となっていない点で本研究とは異なる.

2.3 統計データの文書補強に関する研究

統計データの文書補強に関する研究として, 岡本らは, 統計データ内の表中の行ヘッダと列ヘッダを抽出することで統計文書を補強する手法を提案し, 統計文書のランキングを改善できることを示した [1]. また, Chen らは, 統計データ内の表中の行名や列名, 値を選択し, クエリや文脈と合わせた埋め込みを利用することで, 表の関連性スコアを計算する手法を提案している [15]. これらの手法は, 行名と列名, 値を抽出して利用するのみであるのに対し, 本研究では, 行名と列名, 値を用いて統計データを補強するテキストを生成している点で異なる.

3 提 案 手 法

本節では, 提案手法について詳しく説明する. 3.1 節では本稿で扱う問題を明確にし, 3.2 節では本稿で用いる大規模言語モデルとその特徴をまとめる. 3.3 節では統計データの文書化の詳細を説明し, 3.4 節では補強された統計文書とクエリ拡張, 大規模言語モデルによるリランキングについて説明する.

3.1 問 題 設 定

本稿で扱う統計データ検索の問題設定をまとめる. クエリの集合を Q, 検索対象文書の集合 D として与えられたとき, あるクエリ $q \in Q$ に対して D の各文書を適切にランキングする問題である. ただし, 各検索対象文書は統計文書であり, 統計文書は, メタデータおよび統計データで構成される. 各検索対象文書 $(d_1, d_2, ..., d_k) \in D$ は i 番目のメタデータ及び統計データをそれぞれ m_i, s_i として $d_i = (m_i, s_i)$ として表される.

3.2 利用する大規模言語モデル

本稿において利用する大規模言語モデルについてまとめる. 本稿では, 以下の 3 つのモデルを利用する.

（1） meta-llama/Llama-2-13b-chat-hf[1] [16]

図 2　統計データ内の表構造の例

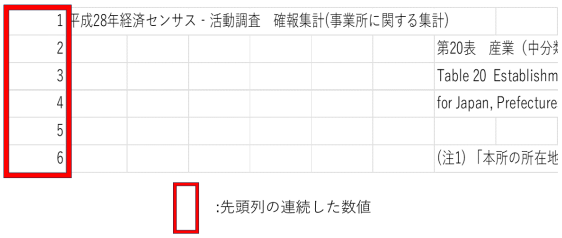


図 3　表内の先頭列に連続して数値が入った状態の数値列

（2） cyberagent/calm2-7b-chat[2]

（3） Gemini Pro [18]

(1) は, 130 億のパラメータ数を持つ大規模言語モデルであり, 日本語に対応したオープンなモデルである. 本稿ではこのモデルを短い文書生成とクエリ拡張に利用する. (2) は, サイバーエージェントが開発した日本語対応の大規模言語モデルであり, 対応する入力長が 32768 トークンと長い文書でも扱える点が特徴である. 本稿ではこのモデルを文書要約に用いる. (3) は Google DeepMind によって開発された大規模言語モデルであり, API を呼び出すことによって利用できる. 本稿では API の呼び出し時間の都合上, リランキングのみに用いる. いずれのモデルにおいても文書を生成する際は temperature=0.0 などとし, ランダム性が生まれないようなパラメータ設定を行った上で利用する.

3.3 表構造認識を介した統計データの文書化

統計データの文書化は, 統計データに対する表構造認識, 大規模言語モデルを用いた文書化の 2 段階で行う.

まず, 統計データに対する表構造認識について説明する. 本稿では, 統計データ内の表は, 見出し, 行名, 列名, 値から構成されているとみなす. 図 2 に, 表構造の例を示す. ここで, 見出しとは, 表全体を説明するテキストを表し, 行名, 列名は, それぞれ表内の具体的な数値 (値) を対応づけするためのテキストを表す. 表構造が得られたら, それをもとにデータフレームを作成する. 具体的には, 以下の手順で作成する.

（1） 統計データの前処理 (整形)

（2） 統計データの値を持つセルが数値であると仮定し, 値の左上部分の位置を特定

（3） 値の左上部分の位置をもとに行名, 列名を抽出

（4） 抽出された行名と列名を用いて値との対応付け

まず, (1) では, 統計データの前処理を行う. ここでは, 統計デー

1：https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

2：https://huggingface.co/cyberagent/calm2-7b-chat

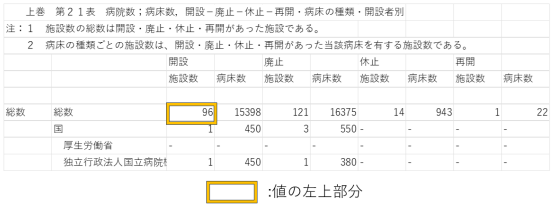


図 4　表の値領域の左上セル

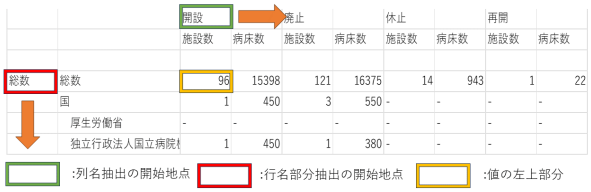


図 5　行名と列名の開始位置の例

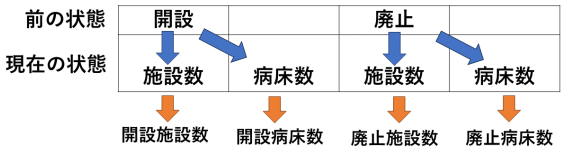


図 6　現在の列を直前の列の空白でない内容で結合する例

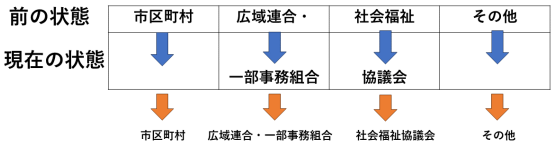


図 7　現在の列を直前の列の空白でない内容で結合する例

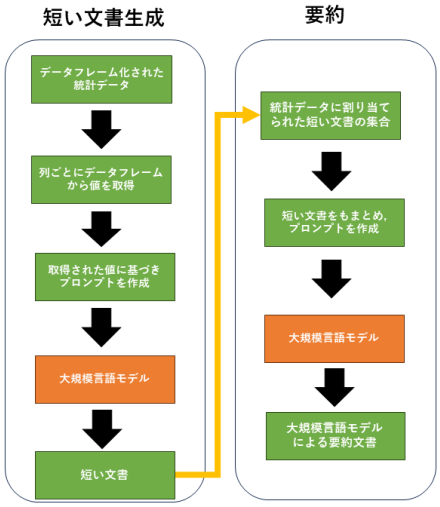


図 8　統計データを文書化する手順

タの先頭列の連続した数値列の削除や不要な文字列の削除，全てが空白のセルである行を削除する．統計データの先頭列の連続した数値列とは，図 3 に示すように，ある列のセルの並びに連続して数値が入っている状態の数値列である．一部の統計データでは，図 3 に示すように，連続した数値列が入っているが，これらは不要なデータとして削除する．不要な文字列の削除では，改行文字の削除やセル内にある ”,” を削除する．これは後に続く処理において行名と列名を正しく抽出するために行う．最後に空白セルのみの行を削除する．

次に，(2) では，値の領域の左上セルを特定する．図 4 に，表の値領域の左上セルの例を示す．特定手順としては，表を 1 行ずつ ”,” で区切り，左上から順に数値であるかどうかを判定することでセルを特定する．また，特定した値領域の左上セルの位置 (row, col) が $row = 0$ もしくは $col = 0$ のような先頭を表す位置であれば，位置を 1 つずらす．例えば $(row, col) = (0, 1)$ であったとき，$(row, col) = (1, 1)$ とする．これは，本稿での表構造認識では行名と列名が必ず存在することを仮定しているためである．

続いて，(3) では，特定した値領域の左上セルを利用して，行名と列名を特定する．行名と列名は，図 5 に示すように，左上セルの列 (あるいは行) 位置から開始し，順にセル内のテキストを抽出する．また，統計データの一部には，図 5 のように階層構造を持つものがあり，順番に抽出するのみでは値との対応付けができない．そこで，直前の状態と現在の状態を比べることで行名と列名を抽出する． 図 6 と図 7 に，直前の状態と現在の状態を比べる例を示す．ここれらは，直前の状態の空白でないセル数と現在の状態の空白でないセル数が一致しない例である．このような場合，図に示す通り，直前の状態を参照しながら現在の状態と結合させる．例えば，図 6 の場合，前の状態が空白な箇所についてはその左側に位置する最初の空白でないセルの内容を，現在の状態に結合することで不足する内容を補うことができる．また，図 7 の場合，現在の状態の最も左のセルが空白であるため，直前の状態の左側を参照することなく，そのまま結合させる．このような方法を用いることで，階層構造を持つ統計データの行名と列名を抽出する．

最後に，(4) では，行名と列名の長さに基づいて値領域の左上セルから順に値を対応付ける．これによりデータフレームを作成する．

次に，統計データの文書化について説明する．前段階までで作成したデータフレームを用いて統計データを文書化する．図 8 に処理手順を示す．データフレームに基づいて列ごとに値を抽出し，大規模言語モデルにプロンプトとして与える．図 9 に，本稿で用いた大規模言語モデルに対するプロンプト構造を示す．短い文書生成では，図 9 の指示部分を「タイトルと列名、値が情報として渡されます。この情報を用いて 500 文字以内で文章を作成してください。」とし，タイトルと現在の列名，該当する 1 列の内容を入力部分に書き込む．この時のタイトルにはメタデータを用いる．生成後の短い文書は，各統計データに対して大量に生成されるため，再び大規模言語モデルを利用してこれらを要約する．要約では，要約では図 9 の指示部分を「次の文書を分析し、要約してください。」とし，入力部分を統計データごとの短い文書を 1 つに結合した内容を書き込む．ここで得られる要約結果が，統計データの補強文書として用いられる．

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:
~~~
### 入力:
~~~
###応答:

図 9　本稿で用いたプロンプトの構造

3.4　補強文書とクエリ拡張, 大規模言語モデルを用いたリランキング

補強文書は, 統計文書のメタデータと紐付けた上で検索に用いる. メタデータには, 統計データのタイトルや説明などの情報が含まれており, 補強文書と組み合わせることで検索漏れを防ぐ効果が期待できる.

また, 検索の前に類似語によるクエリ拡張を行う. クエリ拡張では, 大規模言語モデルのプロンプトとして, 図 9 の指示部分を「次のクエリに似た意味を持つ単語列を改行区切りで 5 つ出力してください。」とし, 入力部分にクエリを書き込む. このプロンプトを大規模言語モデルに入力し, 得られた出力を用いてクエリ拡張する.

続いて, ランキング手法について説明する. 本稿では, ベースとなるランキング手法として BM25 を用いる. クエリ拡張で得られたクエリごとに統計文書に対して検索し, BM25 によるランキング結果を 5 種類取得する.

その後, クエリ拡張後に得られたランキングリストを 1 つに統合する. クエリ拡張では, 1 クエリに対して 5 つの拡張語が作成されており, 合計 6 つのランキングリストが得られている. そのため, 各クエリに対応する 6 つのランキングリストから得られる統計文書ごとの順位の平均を取り, その平均値に基づいて再び統計文書をソートすることで 1 つのランキングリストに統合した. これにより, クエリ拡張によるランキング結果を考慮したランキングリストを作成することができる.

最後に, 統合されたランキング結果をさらに改善するために, 大規模言語モデルを用いた意味内容に基づくリランキングを行う. リランキングの際には Pairwise Ranking Prompting(PRP) [11] を利用する. この手法は, 大規模言語モデルに 2 つの文書とクエリを入力し, 関連性の高さを比較させることによって 2 文書の大小関係を決め, ランキングするプロンプト技術である. PRP では, PRP-Sliding-k, PRP-Allpair, PRP-Sorting の 3 つの手法が提案されており, 本稿ではこれら 3 手法でそれぞれリランキングする. PRP-Sliding-k では, k を 10 とし, 各手法における入力回数の関係から, PRP-Sliding-k と PRP-Sorting では上位 60 位に対してリランキングを行い, PRP-Sorting は上位 30 位に対してリランキングを行う.

4　実　　験

本節では, 実験とその結果について説明する. まず, 4.1 節でデータセットについて説明する. 次に, 4.2 節で実験目的を述べた後, 4.3 節で実験設定を説明する. 4.4 節では実験結果を示し, 考察を行う.

4.1　データセット

実験で用いるデータセットについて説明する. 実験では, クエリ集合と統計文書集合で構成されている NTCIR-15 の Data Search タスクの日本語サブタスクにおけるデータセット [17] を利用する. 訓練クエリは 96 件提供されており, 統計文書は, 統計データとそれに対応したメタデータの組で構成されている. このデータセットに含まれる CSV 形式の統計データに対してデータフレームを作成後, 統計データに値が存在し, かつ 20 行 20 列以下の 27327 件の統計文書に対して実験を行った. 評価の際に用いる関連性スコアは, NTCIR-15 Data Search タスクの日本語サブタスクで採用されている訓練用の関連性スコアを用いる.

4.2　実 験 目 的

本実験では, 提案手法が従来手法と比べてどのように性能が変化するのかを明らかにすることを目的とする.

4.3　実 験 設 定

本節では, 実験設定について説明する. 実験では, データフレームを作成した際の表構造認識, ランキング結果, および, リランキング結果の 3 点について評価する.

4.3.1　評 価 指 標

そ各評価における評価指標を説明する. 表構造認識の評価では値との対応付けを評価する. そのため, 行名と列名の対応付けが正しく行われているのかを正解率で人手によって評価する. なお, ここで見出し部分は抽出されていなくても行名と列名が正しく抽出できていれば正答とする

ランキングとリランキングの評価指標としては, NTCIR-15 の評価で用いられていた nDCG@k ($k = 3, 5, 10, 20$) を採用する.

4.3.2　比 較 手 法

比較手法について説明する. まず, ランキング結果では, メタデータのみを文書とした BM25 によるランキングと, 提案手法で補強された文書に対するランキング結果を比較する. クエリ拡張を行わずメタデータのみを文書として BM25 でランキングした結果を BM25 と表記し, クエリ拡張を行った後にメタデータのみを文書として BM25 でランキングした結果を BM25+QA, クエリ拡張を行った後にメタデータと補強文書に対して BM25 でランキングした結果を BM25+QA+TA とする. また, リランキング結果では, リランキング前後の結果を比較する. リランキングによる手法は, BM25+QA+TA+PRP–Allpair, BM25+QA+TA+PRP–Sliding-k, BM25+QA+TA+PRP–Sorting と表記する.

4.4　実 験 結 果

本節では実験結果を示し, 考察を行う.

表 1　ランキング結果

手法	nDCG@3	nDCG@5	nDCG@10	nDCG@20
BM25	0.686	0.712	0.663	0.758
BM25+QA	0.339	0.420	0.431	0.546
BM25+QA+TA	0.066	0.055	0.174	0.207
BM25+QA+TA+PRP–Allpair	0.198	0.190	0.150	0.273
BM25+QA+TA+PRP–Sliding–k	0.165	0.139	0.191	0.222
BM25+QA+TA+PRP–Sorting	0.125	0.105	0.160	0.195

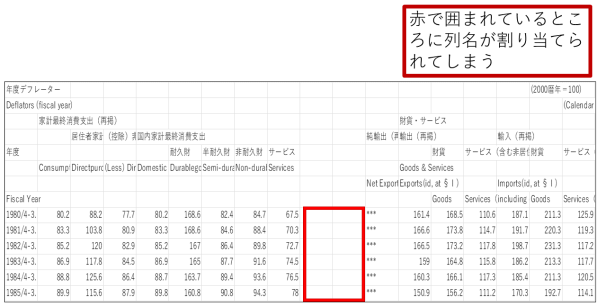


図 10　対応付けが不正解となった例

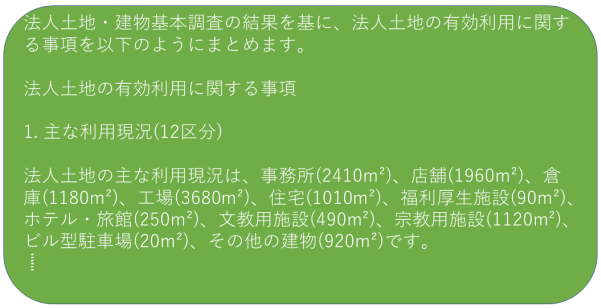


図 11　生成された短い文書の例

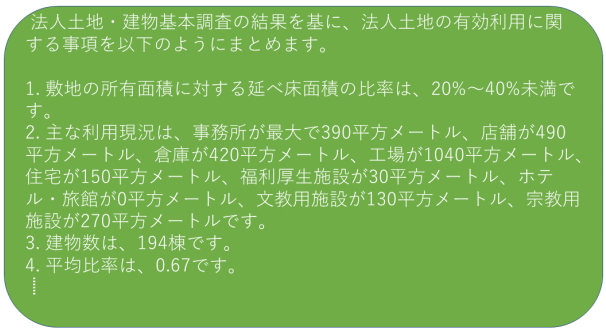


図 12　生成された要約文書の例

表 2　リランキングにおいて大規模言語モデルが生成した文書の割合

生成された文	個数	割合
文書 2	46508	0.7220
文書 1	17577	0.2729
該当なし	302	0.0047
その他	30	0.0005

4.4.1　表構造認識に対する評価結果

行名と列名との値の対応付けの適切さを, 50 件のデータフレームに対して人手で評価した. その結果, 92%の精度で値との対応付けが行えていた. 図 10 に, 対応付けが不正解となった例を示す. この図のように表の中に表がある統計データの場合, 不正解となった. 図 10 では, 赤枠で囲った部分に列名が割り当てられてしまい, 右の値情報が失われた. このように, 表の中に表があるタイプでは本稿の手法のみでは対応付けを行うことが出来ず, 正しく統計データを変換することができなかった. 一方, 92%の正解率で対応付けできていたことから, 後に続く補強文書生成への影響は限定的ではないかと考えられる.

4.4.2　ランキングおよびリランキングの結果

まず, クエリ拡張の適切さを確認するため, 大規模言語モデルで得られた拡張語の例を示す. クエリが「国別 パソコン 所有率」であったときの出力結果は以下の通りとなった.

（1） 国別 PC 所有率
（2） 国別 パソコン 普及率
（3） 国別 コンピューター 所有率
（4） 国別 情報機器 所有率
（5） 国別 デジタル機器 所有率

この結果を見ると, クエリに対する類似語は概ね適切に生成されていることがわかる. これより, クエリ拡張を行うために大規模言語モデルを用いることは概ね妥当ではないかと考えられる.

次に, 図 11, 要約文書を図 12 に, 大規模言語モデルによって生成された短い文書および要約文書の例をそれぞれ示す. 図 11 を確認すると, 大規模言語モデルは, 数値を利用しながら短い文書を生成していることがわかる. 図 12 を確認すると, 短い文書に含まれる数値を含む内容をまとめて文書を生成していることがわかる. このことから生成された文書は, 数値の並びを含む統計データを一定程度要約した内容にはなっていることが確認できた.

続いて, 表 1 ランキング結果を示す. 評価結果を見るとクエリ拡張を行わずにメタデータのみでランキングした場合の評価が最も高くなっており, 文書補強やクエリ拡張を行うほど評価が低下する結果となった. クエリ拡張, 文書補強によるランキングと, メタデータのみによるランキングを比較するために, クエリが「平均貯蓄額」であったときのランキング結果の比較を表 3 に示す.　この表では, NTCIR-15 Data Search タスクの日本語サブタスクで配布されている訓練用の関連性スコアを関連度として示している. 各関連度は, L0(関連性がない), L1(少し関連性がある), L2(関連性がある) として表現されている. この結果から, メタデータのみのランキング (BM25) が上位になるほど関連度が高い統計文書が位置付けられているのに対して, 文書補強とクエリ拡張によるランキングでは上位に関連度が高

表 3　それぞれの手法におけるランキングの比較

順位	関連度 (BM25)	関連度 (BM25+QA)	関連度 (BM25+QA+TA)
1 位	L2	L2	L2
2 位	L2	L1	L2
3 位	L2	L2	L0
4 位	L2	L2	L0
5 位	L2	L2	L0
6 位	L2	L2	L0
7 位	L0	L0	L0
8 位	L0	L0	L0
9 位	L0	L2	L0
10 位	L0	L0	L0

い統計文書が少なく, 下位に関連度が高い統計文書が引き下げられてしまっていることがわかる. これより, 特に文書補強がランキングを改善する本来想定していた役割を果たせていないことが確認できた.

表 2 に, リランキングを行った際に生成された文書の割合を示す. 表 2 より, 「文書 2」として出力される場合が最も多く, ほとんどの出力が「文書 1」あるいは「文書 2」となっていることがわかる. このことから PRP を用いた大規模言語モデルの出力によって, 何らかのリランキングはできていることが確認できる. 表 1 より, PRP によるリランキングを行う前よりも, 行った後の方が, 特に上位 5 位までのランキング性能が向上していることがわかる. PRP によるリランキングは一定の性能向上に貢献していることが確認できる. PRP の各手法で比較すると, PRP-Sorting の結果が最も低い結果となっている. このことから, 統計データ検索のリランキングにおいて PRP を利用する場合は, PRP-Sliding-k と PRP-Allpair を利用することが性能向上に有効であると考えられる.

5 ま と め

本稿では, 統計文書のアドホック検索を改善するため, 大規模言語モデルを用いた統計データの文書補強, および, リランキング手法を提案した. 大規模言語モデルによる文書補強を行うために, 統計データの表構造を抽出しデータフレームを作成した後, データフレームに基づくプロンプトを大規模言語モデルに入力することで文書補強を行った. また, 検索漏れを防ぐためのクエリ拡張, および, 意味内容の類似によるランキング改善を実現するため, Pairwise Ranking Prompting によるリランキングを行った. 実験では, 本手法によりランキング性能がどのように変化するのかを明らかにする比較を行った. メタデータのみを文書とした BM25 によるランキング, リランキングを行う前後の結果を比較した. その結果, クエリ拡張や文書補強によりランキング性能が低下することがわかった. 一方, Pairwise Ranking Prompting によるリランキングの前後で比較すると, PRP はランキングをある程度改善できることが確認できた.

今後は, 文書補強によるランキング性能の低下に対する改善と大規模言語モデルを変更した際のランキング性能の比較について順次取り組む予定である.

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである.

文　　献

[1] 岡本 卓, 宮森 恒. “被検索文書の絞り込みと補強, クエリ拡張に基づく統計データ向けアドホック検索”. 情報処理学会論文誌データベース (TOD), 14(4):36–48 2021.

[2] 中野 優, 加藤 誠. “誤引用検証ための被引用統計データの検索”. 第 13 回データ工学と情報マネジメントに関するフォーラム, F25-4.

[3] Moeri Okuda, Ryota Mibayashi, Takafumi Kawahara, Naoaki Matsumoto, Kenji Tanaka, Takehiro Yamamoto and Hiroaki Ohshima. “UHGSIS at the NTCIR-16 Data Search 2 IR Subtask:A BERT-based Query Modification Approach”. In NTCIR-16, 19–23.

[4] Levy Silva, Luciano Barbosa, Sonia Castelo, Haoxiang Zhang, Aécio Santos and Aécio Santos. “NYUCIN at the NTCIR-16 Dataset Search 2 Task”. In NTCIR-16, 38–45.

[5] Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. “Transformer Memory as a Differentiable Search Index”. In NeurIPS, 2022.

[6] Ronak Pradeep, Rodrigo Nogueira and Jimmy Lin. “The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models”. CoRR, vol.abs/2101.05667, 2021.

[7] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang and Michael Berdersky. “RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses”. CoRR, vol. abs/2210.10634, 2022.

[8] Noah Ziems, Wenhao Yu, Zhihan Zhang and Meng Jiang. “Large Language Models are Built-in Autoregressive Search Engine”. ACL, 2023.

[9] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Anany Kumar, et al. “Holistic Evaluation of Language Models”. CoRR, vol. abs/2211.09110, 2022.

[10] Sukmin Cho, Soyeong Jeong, Jeongyeon Seo and Jong C. Park “Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker”. ACL, 2023.

[11] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. “Large language models are effective text rankers with pairwise ranking prompting”. arXiv preprint arXiv:2306.17563, 2023.

[12] Xueguang Ma, Xinyu Zhang, Ronak Pradeep and Jimmy Lin “Zero-Shot Listwise Document Reranking with a Large

Language Model". CoRR, vol. abs/2305.02156, 2023.
[13] Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In NeurIPS, 2020.
[14] Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. "RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit". CoRR, vol. abs/2306.05212, 2023.
[15] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D. Davison. "Table Search Using a Deep Contextualized Language Model". In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pages 589–598. ACM, 2020
[16] Hugo Touvron, Louis Martin, Kevin Stone, Yinan Xu, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". arXiv preprint arXiv:2307.092883, 2023
[17] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu and Hsin-Liang Chen. "Overview of the NTCIR-15 Data Search Task". In NTCIR-15, 267–273.
[18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. "Gemini: A Family of Highly Capable Multimodal Models". arXiv preprint arXiv:2312.11805, 2023

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

2024年2月28日(水) 16:00 ~ 18:10 | T4-A オンライン（Zoom Events)

生成系マルチメディア

座長:井手 一郎(名古屋大学)
コメンテータ:野宮 浩揮(京都工芸繊維大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T4-A-3-01] FPSゲームにおけるチート対策のためのMistによる物体検出阻害効果の検証実験

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

16:30 ~ 16:55
[T4-A-3-02] 表構造情報を利用した棒グラフの自動生成の一手法

*田上 歩夢[1]、金澤 輝一[2]、上野 史[3]、太田 学[3] (1. 岡山大学工学部情報系学科、2. 国立情報学研究所、3. 岡山大学学術研究自然科学学域)

17:20 ~ 17:45
[T4-A-3-04] 条件付きトランスフォーマーとテキストリフレーサーに基づくクワッドローター用視覚言語ナビゲーション

*Chen Zhe[2,1]、Li Jiyi[1]、Fukumoto Fumiyo[1]、Liu Peng[2]、Suzuki Yoshimi[1] (1. 山梨大学、2. 杭州電子科技大学)

17:45 ~ 18:10
[T4-A-3-05] 【技術報告】ヒトとAIの協調によるデザイン創出支援

山口喬弘 (株式会社豊田中央研究所)

FPS ゲームにおけるチート対策のための Mist による物体検出阻害効果の検証実験

松平　憲† 服部　峻†† 砂山　渡††

† 滋賀県立大学工学部電子システム工学科　〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学先端工学研究院　〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †tz23kmatsuhira@ec.usp.ac.jp, ††{hattori.s,sunayama.w}@e.usp.ac.jp

あらまし　CV (Computer Vision) 技術の進歩に伴い，CV を用いた FPS ゲームのチートツールが広まりつつある．CV を用いたチートツールは，敵プレイヤーの位置を取得するために CV を用いるため，ゲーム内のメモリを読み取るチートに比べて検出が困難である．Mist は本来 Stable Diffusion のような画像生成系 AI に対して学習を阻害し生成不能にするための敵対的ノイズであるが，CV を阻害する効果があるのか，mAP について検証実験を行う．加えて，正規プレイヤーが画面を見づらくなっていないか，Mist を付与した際の画質低下についても評価する．

キーワード　CV, YOLOv8, Mist, 敵対的サンプル

1 はじめに

近年のビデオゲームのプレイヤー数は数億人，産業の規模は数十億ドルと，ビデオゲームは人気のジャンルになりつつある．しかし，人気になるにつれて，チートと呼ばれるゲームを有利に進めるために製作者の意図しない動きをさせる不正行為を行う者も増えており，チートをするプレイヤーはチーターと呼ばれている．チートをされてしまうとゲームバランスが崩壊するだけでなく，ゲーム会社の信用低下やユーザの離脱，収益の低下などが起こる恐れがあり，ゲームを販売する会社にとって大きな被害となりえる [1]．チートの種類はゲームによって様々であるが，FPS (First Person Shooter) ゲームにおいてはウォールハックやレーダーハックなどの敵の場所がわかるチートや，エイムボットと呼ばれるプレイヤーが敵に照準を合わせる動作を助けるチートが存在する．エイムボットの一種で敵の位置を検知するために CV (Computer Vision) ツールを用いるニューラルネットワークエイムボットは，ゲームコードやメモリを読み書きするチートに比べて，ゲーム製作側からほとんど見えないので検出が困難といった課題がある．このような CV を用いたチートツールは，今後 CV 技術の向上に伴い精度が上がる可能性が高い．そこで本稿の目的は，この問題を解決するために CV を用いた FPS チートを阻害することである．

画像を認識するニューラルネットワークを阻害する関連研究として，Liang らの研究 [2] で発表された Mist は，人間によって描かれた絵が，拡散モデル（Diffusion Models）によって学習され，類似したスタイルの絵を生成されることを防ぐために作られた，拡散モデルに対する敵対的サンプルである．敵対的サンプルとは，機械学習モデルを誤分類させるための，微小なノイズのような摂動のことであり，ランダムに生成されたノイズを，対象のモデルが誤動作するように変更することで生成する [3]．

本稿では，Mist が YOLOv8 の物体検出の阻害効果があるのか，mAP (mean Average Precision) について検証実験を行う．加えて，ゲーム画面を加工した際の画質の低下は，チートをしない正規プレイヤーが離脱する要因になるため，本稿では Mist をかけた際の画質低下についても評価する．

以降，2 章では本稿に関連する先行研究を紹介する．次に，3 章では本稿における CV チートの阻害手法および，その評価指標を提案する．さらに，4 章では本稿で行った実験の方法および評価の指標，結果について述べる．最後に，5 章では結論および本稿の社会的意義，技術的意義，今後の展望について述べる．

2 関連研究

本章では本稿に関連する先行研究を紹介する．また，本稿の新規性についても述べる．

2.1 ゲームのチートに関する研究

Kanervisto らの研究 [4] では存在するゲームのチートとチート行為を防止するための一般的な対策方法について述べた後に，敵対的生成ネットワーク（GAN）を用いたエイムボットが自動検出システムと人間の判定員によってチートであると検出されないことが示されている．この研究では，YOLO を用いたチートが存在していて，ゲーム製作側からの検出が難しいチートであることが述べられているが，その対策については述べられていない．YOLO [5] は検出速度が早いことが特徴で，最新バージョンの YOLOv8 [6] では 640×640 の画像を約 1ms で検出することができ，リアルタイム物体検出に用いられる．YOLO などの CV ツールを用いたエイムボットは近年，CV 技術の検出精度と検出速度が向上したことによって出現した新しいチートであるため，このチートへの対策手法はまだ提案されていない．本稿は YOLO のような CV 技術を用いたエイムボットの対策を検討しているという点で新規性がある．

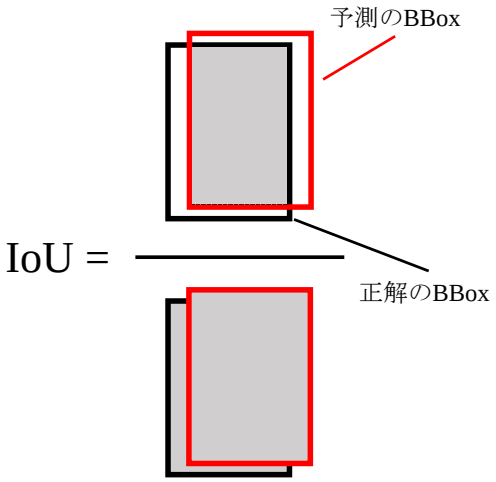


図 1　IoU の算出方法（BBox はバウンディングボックスの略）

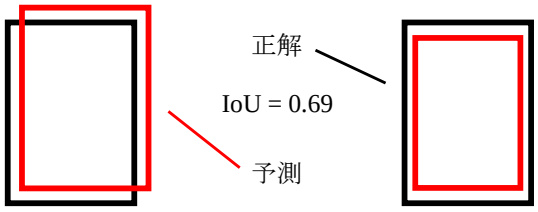


図 2　IoU が同じ予測結果における予測領域のずれ

2.2　物体検出の阻害に関する研究

原ら [7] は画像認識を用いたボットが CAPTCHA を突破してしまう問題を解決するために，画像認識のシステムは誤認識するが，人間にとっては容易に認識できる画像の作成方法を提案している．ここでは画像認識のシステムとして Watson Visual Recognition や Google Vison API などが用いられており，YOLO は用いられていないが，画像認識のシステムは誤認識するが人間にとっては認識できる画像を生成するといった点で本稿と類似している．一方で，本稿との違いは，本稿では YOLOv8 を阻害対象としている点や，人間にとって認識できる画像の基準が異なる点であり，原らの研究における人間が認識できる画像とは CAPTCHA として最低限認識できれば画質が悪くても良いが，本稿での認識できる画像とはゲームのプレイヤーのユーザエクスペリエンスを損なわない画像のことであり，画質の低下も評価基準である．

また，Ranst らの研究 [8] では YOLOv2 に対する敵対的サンプルを生成し，それを人に貼り付けることで YOLOv2 の検出を妨害できることが示されている．この研究は YOLO を敵対的サンプルを用いて阻害するという点で本稿と類似しているが，本稿の CV チート対策としてゲーム画面における敵の検出を阻害する点は，人の検出を阻害する点と異なっており，本稿には新規性がある．さらに，この研究では阻害対象のモデルとして，YOLOv2 が用いられているが，本稿では現時点で最新バージョンの YOLOv8 を用いている．

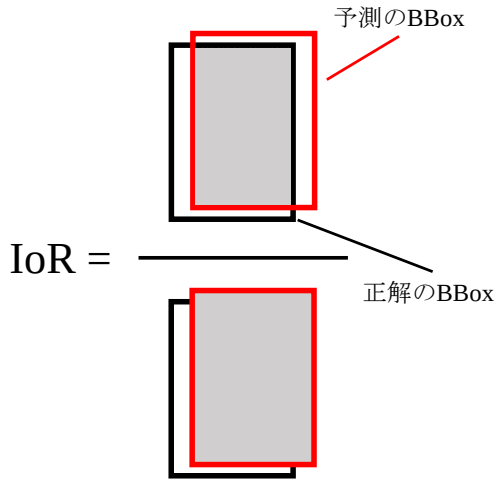


図 3　IoR の算出方法（BBox はバウンディングボックスの略）

表 1　IoU と IoR の違い（Gt, Pr はそれぞれ正解領域，予測領域）

	IoU	IoR
定義	$(Gt \cap Pr)/(Gt \cup Pr)$	$(Gt \cap Pr)/Pr$
値域	0 から 1	0 から 1
指標が 1 のとき	$Gt = Pr$	$Gt \supseteq Pr$

2.3　画像生成系 AI の阻害に関する研究

Liang らの研究 [2] では人が描いた絵を AI によって学習されることで模倣されることを防ぐために，Stable Diffusion や Scenario.gg などの拡散モデル（Diffusion Models）に対する敵対的サンプルを使用した生成系 AI からの保護ツール “Mist” が作成されている．Mist を付与した画像を Stable Diffusion や Scenario.gg に学習させ，画像生成をすると学習した画像とは異なる，ノイズの入った画像が生成され，生成系 AI により絵を模倣されることを防ぐことができる．Mist は拡散モデルの学習を阻害するための透かしを画像に入れるが，この透かしは人間には見えず，画像の見た目を損なわないといった性質がある．この研究は画像の見た目を変えないままニューラルネットワークを妨害するという点で本稿と類似しているが，阻害対象のモデルが YOLO ではなく，Stable Diffusion などの拡散モデルであることが本稿と異なっている．

3　提 案 手 法

本章では CV を用いた FPS チートを阻害するための手法として Mist を用いることを提案する．また，その阻害効果を正しく測定するための評価指標 IoR (Intersection over Recognized object’s bounding box) についても提案する．

3.1　Mist を用いた阻害手法

本稿では CV を阻害するためにゲームのフレーム全体に Mist を付与する．Mist は拡散モデルに対する敵対的サンプルであり，CV に用いられるモデルは拡散モデルではないが，Mist が CV に対して阻害効果があるのか，検証実験を行う．

3.2 FPS ゲームにおける CV チートの阻害効果の評価指標

物体検出の mAP (mean Average Precision) [9] を算出する際には IoU (Intersection over Union) をしきい値として正誤判定を行うことが多い．IoU とは図 1 のように正解の領域と予測の領域を合わせた領域全体（Union）のうちに占める，正解の領域と予測の領域の共通部分（Intersection）の割合で，正解と予測の 2 つの領域がどれだけ一致しているかを評価する指標である．IoU は 0.00 から 1.00 の値をとり，IoU が 0.00 のときは予測の領域が正解の領域に少しも重なっていないことを表し，IoU が 1.00 のときは予測と正解の領域が完全に一致していることを表す．mAP@0.50 や mAP@0.95 などは，それぞれ mAP の算出における正誤判定を IoU が 0.50, 0.95 以上であれば正解とするという意味である．

ここで，一例として IoU が 0.69 の場合を考えると，同じ IoU の値であっても，図 2 の左に示したように予測領域全体が正解領域に対してずれている場合もあれば，右に示したように予測領域が正解領域の中に入っていて，予測領域の面積が正解領域に対して小さい場合もある．左の場合は FPS ゲームにおいて敵を狙う状況を考えると，予測領域が正解領域からずれているため，敵に攻撃が当たらない可能性があるが，右の場合は予測領域のどこを狙っても正解領域の中であるため，必ず敵に攻撃が当たる．しかし，mAP@0.70 では，どちらも IoU が 0.70 を下回っているため，正誤判定は失敗，すなわちチートができないと評価される．よって，FPS チートの対策において，物体検出の阻害効果を評価する指標として，IoU をしきい値とした mAP は適していないと言える．

そこで FPS ゲームにおける CV チートの阻害効果を評価する指標として IoU をしきい値とした mAP に代わって，IoR (Intersection over Recognized object's bounding box) をしきい値とした mAP を新たに導入する．IoR の定義は図 3 のように予測の領域（Recognized object's bounding box）のうちに占める，正解の領域と予測の領域の共通部分（Intersection）の割合のことであり，IoU と同様に正解と予測の 2 つの領域がどれだけ一致しているかを評価する指標である．両者の相違点は定義式の分母であり，IoU では分母が正解領域と予測領域を合わせた領域であるが，IoR では予測領域のみである．IoR は IoU 同様に 0.00 から 1.00 までの値を取り，IoR が 0.00 のときは予測領域が正解領域に少しも重なっていないことを表し，IoR が 1.00 のときは予測領域全体が正解領域の中にあることを表している．つまり，図 2 の右のような場合 IoU は 0.69 となり，IoU のしきい値が 0.70 のとき，チートは失敗とみなされるが，IoR は 1.00 となるためチートは成功したと正しく評価される．表 1 に IoU と IoR の違いを示す．

FPS ゲームにおいて CV チートを阻害するには，以下のような方法が考えられる．

- 物体検出モデルにそもそも敵を検出できなくさせる
- 予測のバウンディングボックスの位置をずらすことで敵に弾が当たらないようにする

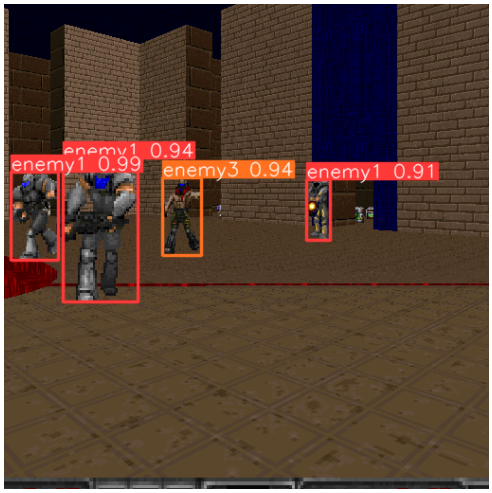


図 4　YOLOv8 で学習させた ViZDoom

- 物体検出モデルに該当の敵ではない別のオブジェクトに誤認識させる
- 実際に敵がいない位置に敵がいると誤検出させる

ここで，IoR をしきい値とした mAP の意味について，正解領域全体が敵の当たり判定であると仮定すると，IoR が 0.50 ということは予測領域のうち，50%が敵の当たり判定であるので，予測の領域中をランダムに射撃すると 50%の確率で敵に弾が当たるということである．つまり，IoR をしきい値とした mAP の値が下がるということは，検出結果の IoR がしきい値より低くなっているということであり，その場合，物体検出モデルに敵を検出できなくさせているか，予測のバウンディングボックスの位置がずれているので，FPS ゲームにおいて CV チートを阻害できていると言える．

4 評価実験

本章では本稿で行った実験の方法および評価の指標について詳述する．本稿での評価実験の目的は，Mist を用いた阻害手法が，CV チートに対して効果があるのかを検証することである．そのために，FPS ゲームの敵キャラを学習させた物体検出モデルに，各フレームに Mist をかけた後の画面を検出させて，その阻害効果を IoR をしきい値とした mAP を用いて評価する．加えて，ゲーム画面を加工した際の画質の低下は，チートをしない正規プレイヤーが離脱する要因になるため，Mist を付与した際の画質の低下について PSNR を用いて評価する．

4.1 実験環境

本稿では FPS ゲーム ViZDoom [10] の敵を CV ツール YOLO の最新バージョン YOLOv8 [6] で学習させ，Mist [2] をかけることで YOLOv8 の検出がどのように変化するかを実験した．

ViZDoom [10] は強化学習を用いて FPS ゲーム Doom の AI ボットを開発するゲームである．図 4 に ViZDoom の敵を YOLOv8 で学習させたゲーム画面を示す．

図 5　モザイクをかける前の検出結果

図 6　モザイクをかけた後の検出結果

図 7　緑色の背景を付与せずに学習したモデルの検出結果

図 8　緑色の背景を付与した後に学習したモデルの検出結果

4.2　事前実験

本節では本稿の提案手法である各フレームに Mist を付与する手法以外のフレーム加工手法での実験の内容と，その結果について述べる．まず，ViZDoom の敵キャラを学習させた YOLOv8 のモデルを阻害させるために，ゲームのフレームにモザイクをかけた．その結果は図 5 と図 6 のようにモザイクをかけても CV 阻害の効果が無く，全ての敵がうまく検出された．次に，フレーム全体に緑色の背景を付与した．その結果は図 7 のように敵が検出されなかった．しかし，同じ緑色の背景を付与したまま YOLOv8 を学習させたところ，図 8 のように全ての敵がうまく検出された．また，別の敵を学習させたモデルや COCO データセットで学習された YOLO の事前学習モデル YOLOv8n [6] に対して緑色の背景を付与しても，このモデルのように検出が失敗することがなかったため，自作のモデルだけにしか CV 阻害の効果がなく，実際のチート対策には有効ではないと考えられる．

4.3　評価指標

本稿の目的は CV を阻害して FPS ゲームの敵を認識できないようにすることである．本稿では CV を阻害するためにゲームのフレーム全体に Mist を施しているが，Mist の強度を強くすれば，ゲームの画質の低下が起こってしまう．ゲーム画質の低下はチートをせずにゲームをプレイする正規プレイヤーのユーザエクスペリエンスを損ない，ユーザの離脱に繋がる可能性があるため好ましくない．そこで本稿では CV がどれだけ阻害できているかを 3.2 節で示した IoR をしきい値とした mAP を用いて評価しつつ，ゲーム画面がどれだけ見やすいかを PSNR (Peak Signal-to-Noise Ratio) を用いて評価する．PSNR は以

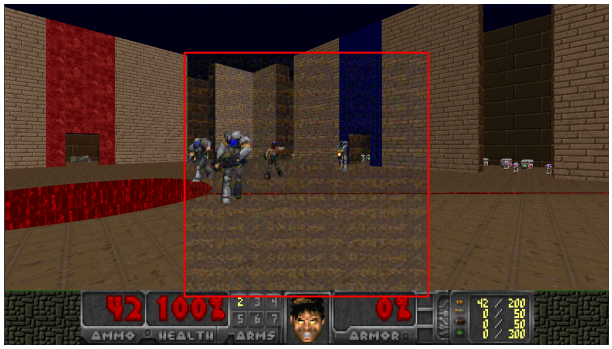


図 9　ゲームフレーム全体に対して Mist を付与する範囲（画像は強度が 16 の Mist）

下の式 (1) で表される [11].

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{e^2} \tag{1}$$

$$e^2 = \frac{1}{N} \sum_i (x[i] - y[i])^2 \tag{2}$$

ここで $x[i]$, $y[i]$ とは元の画像および加工した画像の i 番目の画素位置における輝度値であり，e^2 は元の画像と加工した画像の輝度値の平均二乗誤差である．実際には，各々の画像を `cv2.imread()` 関数で読み込んだ上で，`cv2.PSNR()` 関数に引数として渡して計算している [12].

4.4　実験方法

本節では本稿の実験方法を詳述する．本稿では，Mist の推奨出力サイズが 512 × 512 の正方形であるため，1280 × 720 のゲームフレームのうち，学習，予測で使用する範囲として図 9 の赤枠に示した 512 × 512 の正方形の領域のみを使う場合と，フレーム全体を使う場合で実験を行った．

4.4.1　学習および予測に中心部のみを用いた実験

本稿では図 10 から 14 に示す 5 種類の ViZDoom の敵をゲームのフレーム 141 枚を用いて YOLOv8 に学習させた．その内，Train と Valid に用いたフレーム数は 121 枚と 20 枚であり，学習は 100 エポックで行った．その後，予測用に 121 枚の画像を用意し，作成した YOLOv8 モデルを用いて元画像と，強度を 8, 16, 32 で Mist を付与した画像を予測し，各予測結果の mAP を，IoR=0.50, 0.70, 0.90 として計算した．学習と予測に用いた敵の数を表 2 に示す．また，元画像と強度を 8, 16, 32 で Mist を付与したときの画像の PSNR についても測定した．さらに，学習に用いた 141 枚の画像に強度を 16, 32 の Mist を付与したまま YOLOv8 を学習させて，予測に用いた 121 枚の画像と強度を 8, 16, 32 で Mist を付与した画像を予測し，各予測結果の mAP を，IoR=0.50, 0.70, 0.90 として計算した．

本稿では，5 種類の敵を学習した場合に加えて，1 種類の敵のみを学習した YOLOv8 のモデルに対しても実験を行った．ここでは図 15 に示す Enemy0 をゲームのフレーム 112 枚を用いて YOLOv8 を学習させた．その内，Train と Valid に用いたフレーム数は 100 枚と 12 枚であり，学習は 100 エポックで行った．その後，予測用に 120 枚の画像を用意し，作成した YOLOv8 モデルを用いて元画像と，強度を 8, 16, 32 で Mist

図 10　Enemy1

図 11　Enemy2

図 12　Enemy3

図 13　Enemy4

図 14　Enemy5

図 15　Enemy0

表 2　学習および予測にフレームの中心部のみを用いたときの 5 種類の敵の数

	学習 [体]	予測 [体]
Enemy1	196	169
Enemy2	85	13
Enemy3	68	48
Enemy4	33	12
Enemy5	23	21
総数	405	263

表 3　学習および予測にフレーム全体を用いたときの 5 種類の敵の数

	学習 [体]	予測 [体]
Enemy1	306	231
Enemy2	125	54
Enemy3	77	51
Enemy4	42	36
Enemy5	29	32
総数	579	404

を付与した画像を予測し，各予測結果の mAP を，IoR=0.50, 0.70, 0.90 として計算した．学習と予測に用いた敵の数は，それぞれ 105 体，120 体である．また，元画像と強度を 8, 16, 32 で Mist を付与したときの画像の PSNR についても測定した．さらに，学習に用いた 112 枚の画像に強度を 16, 32 の Mist を付与したまま YOLOv8 を学習させて，予測に用いた 120 枚の画像と強度を 8, 16, 32 で Mist を付与した画像を予測し，各予測結果の mAP を，IoR=0.50, 0.70, 0.90 として計算した．

4.4.2　学習および予測にフレーム全体を用いた実験

本稿では，Mist の推奨出力サイズではないが，より現実的な場面を想定して，4.4.1 項に示した 5 種類の敵を用いる実験を，ゲームのフレームの中心部だけではなく，フレーム全体を用いて行った．本実験で学習と予測に用いた敵の数を表 3 に示す．

4.5　実験結果

本節では Mist による物体検出の阻害効果および Mist を付与した際の画質低下についての結果をまとめる．

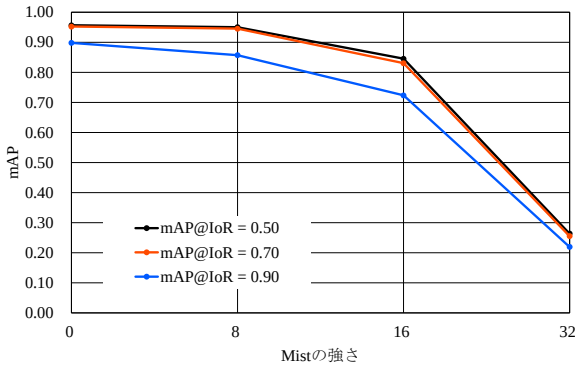


図 16　元の画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

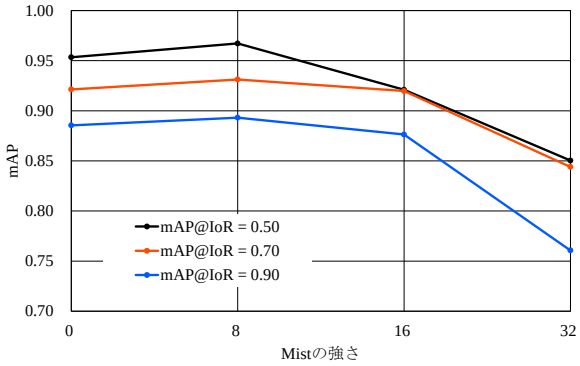


図 17　強度を 16 で Mist を付与した画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

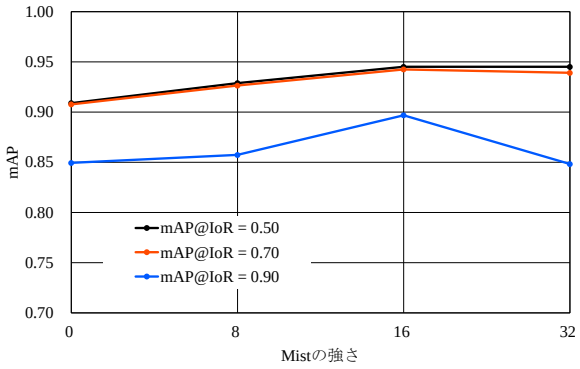


図 18　強度を 32 で Mist を付与した画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

4.5.1　Mist による物体検出の阻害効果

4.4.1 項で示した 5 種類の敵の元の画像を学習したモデルに対して，予測用の 121 枚の画像を予測させたときの mAP は，図 16 のように，フレームに付与する Mist の強さを強くするにつれて低くなった．また，図 17 のように，強度を 16 で Mist を付与した画像を学習したモデルに対して，予測用の 121 枚の画像を予測したときの mAP は，元画像（Mist が 0 の画像）を予測したときよりも，Mist の強さを 8 にしたときの方が高くなった．しかし，フレームに付与する Mist の強さが 32 の画像を予測したときの mAP は Mist の強さが 0, 8, 16 の画像を予

測したときの mAP に比べて低くなった．さらに，図 18 のように，強度を 32 で Mist を付与した画像を学習したモデルに対して，予測用の 121 枚の画像を予測したときの mAP は，フレームに付与する Mist の強さが 0, 8, 16, 32 の全ての場合において 0.80 以上と高い値を維持したままであった．

この結果から，元画像を学習したモデルを用いた CV チートは，Mist をかけることで阻害することができると考えられる．また，強度を 16 で Mist を付与した画像を学習したモデルを用いた CV チートには，強度が 32 の Mist を付与することで阻害できると考えられる．しかし，Mist を付与した画面を学習されると，図 18 に示したように，Mist を付与した画像に対してだけでなく，元画像に対する予測の mAP も高いモデルが作られる．

さらに，1 種類の敵の元画像を学習したモデルを用いて，予測用の 120 枚の画像を予測したときの mAP は，図 19 のように Mist の強度が 0, 8, 16 のときは下がらなかったが，Mist の強度が 32 のときは，0.20 以下と非常に低くなった．しかし，強度を 16 および 32 で Mist を付与した画像を学習させたモデルを用いて，予測用の 120 枚の画像を予測したときの mAP は，図 20, 21 のように Mist の強さに関わらず 0.95 以上と非常に高い値を維持したままであった．

また，4. 4. 2 項に示した 5 種類の敵の元画像をフレーム全体を用いて学習したモデルに対して，予測用の 121 枚の画像を予測させたときの mAP は図 22 のように，中心部のみを用いて学習したモデルを用いたときと比べると変化が小さいものの，Mist の強さを強くするにつれて低くなった．さらに，強度が 16 の Mist を付与した 5 種類の敵の画像をフレーム全体を用いて学習したモデルに対して，予測用の 121 枚の画像を予測させたときの mAP は図 23 のように，中心部のみを用いて学習したモデルと同様に，フレームに付与する Mist の強さが 32 の画像を予測したときの mAP は，Mist の強さが 0, 8, 16 の画像を予測したときの mAP に比べて低くなった．また，強度が 32 の Mist を付与した 5 種類の敵の画像をフレーム全体を用いて学習したモデルに対して，予測用の 121 枚の画像を予測させたときの mAP は図 24 のように，mAP@IoR=0.50 および mAP@IoR=0.70 はほとんど変化しなかったが，mAP@IoR=0.90 は Mist の強さが 32 のとき 0.51 と Mist の強さが 0, 8, 16 の画像を予測したときの mAP に比べて低くなった．加えて，フレーム全体を用いて学習したモデルを用いた全ての実験において mAP@IoR=0.90 の値が 0.50, 0.60 付近と小さくなった．これは，フレーム全体を学習した際に図 9 に示したフレーム下部のプレイヤーの顔などが学習の妨げになったからと考えられる．

これらより，元画像を学習したモデルを用いた CV チートは，Mist をかけることで阻害することができると考えられるが，Mist を付与した後のゲーム画面を学習したモデルを用いた CV チートは，Mist をかけることで阻害することができず，再学習したモデルの Mist を付与した画像に対する予測は，元画像に対する予測よりも mAP が高くなることもあるため，結果的に CV チートの精度を向上させてしまい，再びチートされると考えられる．

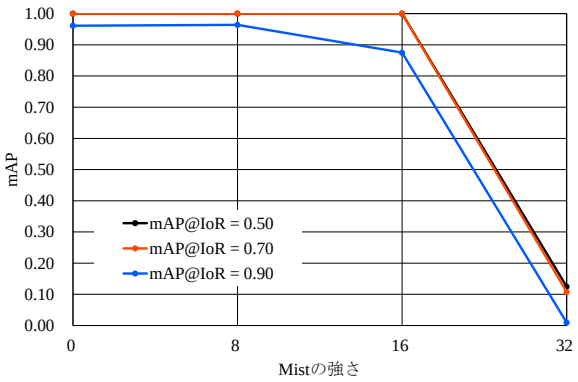


図 19　1 種類の敵の元画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

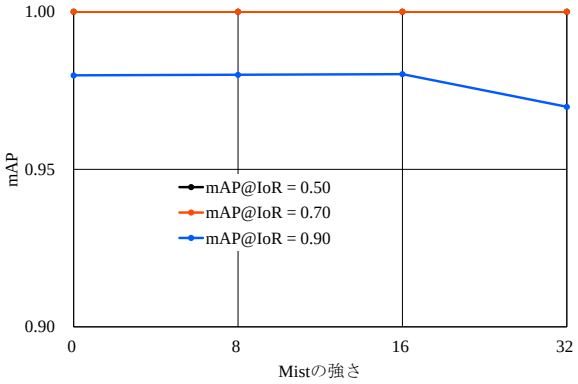


図 20　強度を 16 で Mist を付与した 1 種類の敵の画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

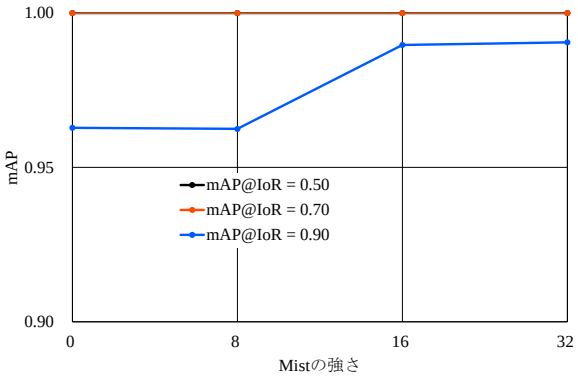


図 21　強度を 32 で Mist を付与した 1 種類の敵の画像を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

4. 5. 2　Mist による画質低下

図 25 に示したゲーム画面に Mist を付与すると，Mist の強度が 8, 16, 32 のとき，それぞれ図 26，図 27，図 28 のようになった．このときの PSNR は表 4 のように Mist の強度を強くするに従って小さくなった．PSNR の値は 35 以上であれば高品質と言われている [13]．したがって，Mist の強度が 8 のときは高品質に近いと言えるが，Mist の強度が 16 や 32 となったときは高品質とは言えない．

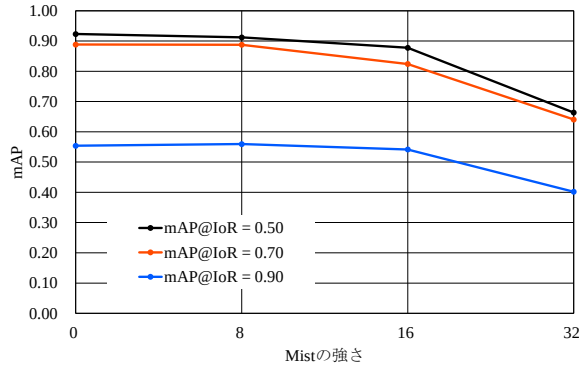


図 22　元の画像の全体を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

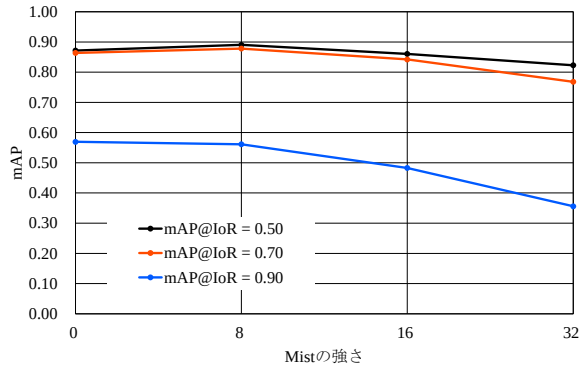


図 23　強度を 16 で Mist を付与した画像の全体を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

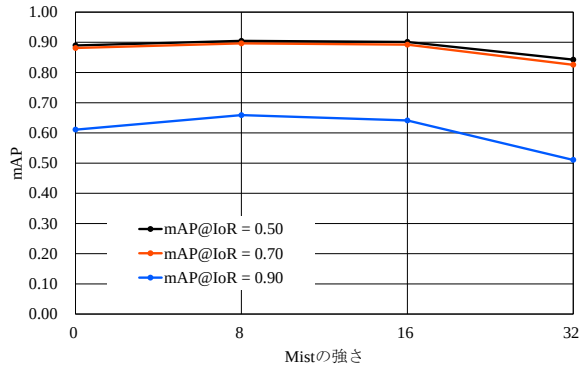


図 24　強度を 32 で Mist を付与した画像の全体を学習したモデルに対して Mist の強度を強くしたときの mAP の変化

ここで，強度が 8 の Mist を付与したゲーム画面は本稿で実験した，どのモデルを用いて予測しても，Mist が 0 のときから mAP がほとんど変化していないことから，CV 阻害効果がほとんどないと言える．また，強度 32 の Mist は 5 種類の敵の元画像を学習したモデルや強度が 16 の Mist を付与したゲーム画面を学習したモデル，1 種類の敵の元画像を学習したモデルに対して CV 阻害効果があるが PSNR が 22.13 であることから，実際の CV チート対策として Mist を付与する阻害手法は有効ではないと考えられる．

図 25　Mist を付与していないゲーム画面

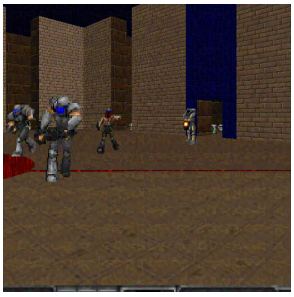


図 26　強度が 8 の Mist を付与したゲーム画面

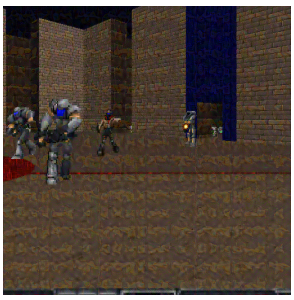


図 27　強度が 16 の Mist を付与したゲーム画面

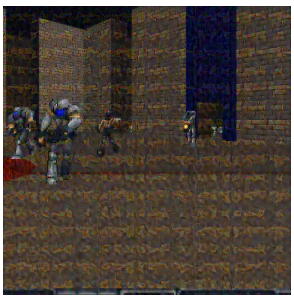


図 28　強度が 32 の Mist を付与したゲーム画面

表 4　Mist を付与した際の PSNR の平均値

Mist の強度	8	16	32
敵が 5 種類のときの PSNR（$n = 121$）	32.40	26.98	22.13
敵が 1 種類のときの PSNR（$n = 120$）	32.20	26.93	22.13

4.6　IoR をしきい値とした mAP の評価

ここで 4.5.1 項で述べた，5 種類の敵の元画像を学習したモデル，強度を 16 で Mist を付与した画像を学習したモデル，強度を 32 で Mist を付与した画像を学習したモデルの 3 種類のモデルに対して Mist の強度を強くして，IoR をしきい値とした mAP がどのように変化するかを示した実験を，従来の IoU をしきい値とした mAP に変更して行うと，結果は，図 29，図 30，図 31 のようになっており，図 16 に示した結果より元の画像を学習したモデルを用いて，元画像（Mist の強度を 0）を予測したときの mAP@IoR=0.90 の値は 0.90 であったので，CV を用いた FPS チートを考えると，予測に用いた 263 体の敵のうち 90%の敵の予測範囲は，9 割以上が正解領域に重なっているので，チートは成功すると考えられるが，図 29 に示した結果では，mAP@IoU=0.90 の値は 0.58 と小さくなっており，この値を見るとチートが失敗するかのように捉えられるので，IoU をしきい値とした mAP は，CV を用いた FPS チートの阻害を評価する指標として適切でないと言える．加えて，図 18 に示したように，強度を 32 で Mist を付与した画像を学習したモデルにおける mAP@IoR=0.90 は 0.80 を下回ることはないため，CV チートは阻害できていないと考えられるが，図 31 では，Mist の強さを 32 にすると，mAP@IoU=0.90 は 0.42 と，Mist の強さが 0, 8, 16 のときと比べて小さくなっており，CV チートの阻害効果があるように捉えられるため，IoU をしきい

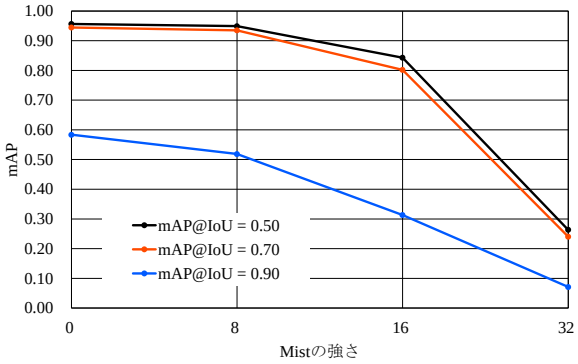


図 29　元の画像を学習したモデルに対して Mist の強度を強くしたときの IoU をしきい値とした mAP の変化

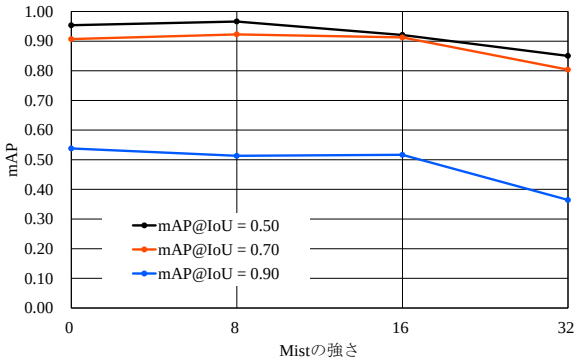


図 30　強度を 16 で Mist を付与した画像を学習したモデルに対して Mist の強度を強くしたときの IoU をしきい値とした mAP の変化

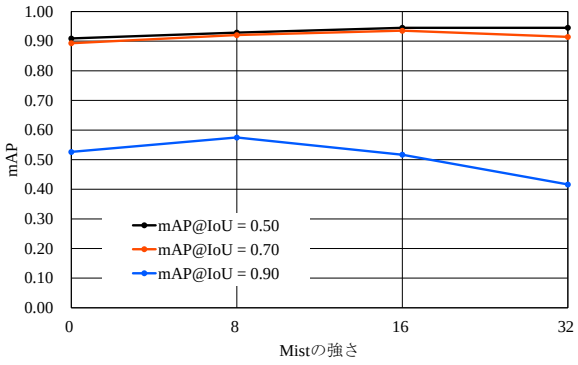


図 31　強度を 32 で Mist を付与した画像を学習したモデルに対して Mist の強度を強くしたときの IoU をしきい値とした mAP の変化

値とした mAP は，CV を用いた FPS チートの阻害を評価する指標として適していないと言える．

これは図 32 のように，予測したバウンディングボックスの領域が正解の領域の内側に入っていて，正解の領域よりも小さくなっているために起こると考えられる．実際に，図 32 において，IoU および IoR を計算すると，左の敵は IoU が 0.84, IoR が 0.95, 右の敵は IoU が 0.71, IoR が 0.97 となっていて，この図において IoR と IoU のしきい値をどちらも 0.90 としたとき，mAP@IoU ではどちらの敵も誤検出となるが，mAP@IoR

図 32　検出の位置が正解領域の内側にある例

ではどちらも正解となる．また，この検出結果の IoR が 0.95, 0.97 と高いことから，予測領域のうち大部分が敵に弾が当たる範囲であるので，CV を用いた FPS チートは阻害できていないと言える．このことから，本稿で新たに導入した IoR をしきい値とした mAP は，従来の IoU をしきい値とした mAP に比べて，CV を用いた FPS チートの阻害を評価する指標として適切であるとわかる．

5　結論および今後の展望

本稿ではゲーム画面に Mist を付与することが CV チートに対して阻害効果があるのかを検証した．結論として，Mist をゲームのフレームに付与する阻害手法は，ゲームの元画像を学習したモデルに対しては効果があるが，画質が低下してしまう点や，Mist を付与した後のゲーム画面を学習したモデルに対しては効果がないだけでなく，逆効果になることもある点から，実際のチート対策には有効ではないと言える．また，本稿では CV チートに対する阻害効果を評価する指標として，IoR をしきい値とした mAP を新たに導入した．IoR をしきい値とした mAP は，従来の IoU をしきい値とした mAP に比べて，予測領域が正解領域の中に入っている状況において，CV チートが成功したと正しく判定できるため，CV を用いた FPS チートの阻害を評価する指標として適切であると言える．

5.1　技術的貢献および社会的貢献

本稿によって，ゲーム画面を学習した YOLOv8 のモデルは Mist を付与することで検出を阻害できるが，Mist を付与した後のゲーム画面を学習したモデルに対しては，Mist を付与することでは検出を阻害できないことがわかった．CV チートを阻害する観点からは，阻害効果のあるノイズを画面に施しても，その画面を再学習されることで，ノイズのない画像に対してもノイズのある画像に対しても，阻害効果がなくなるといった望ま

しくない結果となったが，YOLOv8 の精度向上の観点からは，敵対的サンプルのような悪意のある攻撃をされた際や，ノイズによって検出が困難になる状況において，そのノイズを含めて YOLOv8 のモデルを再学習することで，検出が阻害されないようにすることができる可能性があり，結果的に，YOLOv8 の検出精度を向上することに貢献できたと考えられる．

また，2.1 節でも述べたように，CV ツールを用いたエイムボットは，CV 技術の検出精度と検出速度が向上したことによって出現した新しいチートであるため，まだ対策手法が提案されていない．本稿は，CV を用いたエイムボットの対策手法を提案する初めての論文として，ゲーム業界の健全な発展に貢献していると考えられる．また，日本においては，ゲームのチートを防止するための論文が少なく，筆者が調べた限りでは，CV を用いた FPS チートについて言及している日本語の論文は見つからなかった．本稿は，世界的に産業規模の大きいゲーム業界のチート問題について言及し，CV を用いた FPS チートなど新しい技術を用いた不正行為が存在することを示すことで，日本におけるゲーム業界の発展に貢献していると考えられる．

5.2 今後の課題

今後の本研究の課題は，引き続き CV チートに対する阻害手法を検討することである．そのために，今後の展望として，本稿のように CV チートを阻害できる加工を施した後に，その画面を学習されると再び敵を検出されてしまうことの対策として，加工済み画像で学習したモデルに対して阻害効果のある加工を，敵対的サンプルを用いることで生成することを検討していく．

文　献

[1] 株式会社ラック，“オンラインゲームを悩ませるチート行為とは？CEDEC2022 講演情報も公開，” https://www.lac.co.jp/lacwatch/people/20220725_003052.html (参照 2023-12-27).

[2] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan, “Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples,” International Conference on Machine Learning, PMLR, vol. 202, pp. 20763–20786 (2023).

[3] 株式会社 FFRI セキュリティ，“Adversarial Example，” https://www.ffri.jp/assets/files/research/research_papers/adversarialexample.pdf (参照 2024-01-04).

[4] Anssi Kanervisto, Tomi Kinnunen, and Ville Hautamaki, “GAN-Aimbots: Using Machine Learning for Cheating in First Person Shooters，” IEEE Transactions on Games, vol. 15, no. 4, pp. 566–579 (2022).

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection，” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016).

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu, “Ultralytics YOLOv8，” (Version 8.0.0), https://github.com/ultralytics/ultralytics (参照 2024-01-04).

[7] 原亨, 佐藤敬, “多数の不連続点に起因する画像の誤認識とその CAPTCHA への応用，” 情報処理学会論文誌, vol. 60, no. 12, pp. 2139–2146 (2019).

[8] Van Ranst, Wiebe, Simen Thys, and Toon Goedemé, “Fooling automated surveillance cameras: adversarial patches to attack person detection，” CVPR Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security, pp. 49–55 (2019).

[9] Padilla. R, Passos. W.L., Dias. T.L.B, Netto. S.L, and da Silva. E.A.B, “A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit，” Electronics, vol. 10, no. 3, #279 (2021).

[10] M. Wydmuch, M. Kempka, and W. Jaśkowski, “ViZDoom Competitions: Playing Doom from Pixels，” IEEE Transactions on Games, vol. 11, no. 3, pp. 248–259 (2019).

[11] 電子情報通信学会知識ベース，“9 章 符号化画質，” https://www.ieice-hbkb.org/files/02/02gun_05hen_09.pdf (参照 2023-12-27).

[12] OpenCV, “Operations on arrays: PSNR()，” https://docs.opencv.org/4.x/d2/de8/group__core__array.html#ga3119e3ea73010a6f810bb05aa36ac8d6 (参照 2024-2-13).

[13] 安藤英俊, 望月雄太, “GPU によるウェーブレット変換を用いた高精細映像の低遅延ネットワーク配信，” 情報処理学会研究報告, vol. 2009-CG-136, no. 10 (2009).

表構造情報を利用した棒グラフの自動生成の一手法

田上 歩夢† 金澤 輝一†† 上野 史††† 太田 学†††

† 岡山大学工学部情報系学科 〒 700–8530 岡山県岡山市北区津島中 3-1-1
†† 国立情報学研究所 〒 101–8430 東京都千代田区一ツ橋 2-1-2
††† 岡山大学学術研究院環境生命自然科学学域 〒 700–8530 岡山県岡山市北区津島中 3-1-1
E-mail: †pg4g0a4w@s.okayama-u.ac.jp, ††tkana@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし 学術論文では実験結果や統計情報がしばしば表にまとめられるが，表よりもグラフの方が数値の比較や変化を視覚的に読み取りやすい．そこで本研究では，青柳らの表構造解析手法を用いて生成される表構造情報を利用して，数値を含む表から棒グラフを自動生成する方法を提案する．本研究では，棒グラフ生成の前処理として，罫線を利用して誤った表構造情報を修正する処理と表のサブヘッダ行/列を判別する処理を提案する．さらに，表のキャプション抽出も提案する．実験では，棒グラフ生成手法を用いて ICDAR 2013 の Table Competition のテスト用データセットに含まれる表から棒グラフを生成し，青柳らの手法によって得られる表の構造情報とコンテンツから誤りや欠損なくグラフを生成できるかどうかを評価した．棒グラフ生成の実験の結果，構造が比較的単純な EU の表データで成功率が 0.93，構造が比較的複雑な US の表データで成功率が 0.74 となった．提案した前処理によって EU のデータの成功率が 22 ポイント，US のデータの成功率が 5 ポイント上昇した．また，同じ表データを用いて棒グラフ以外の種類のグラフも生成できることを示した．

キーワード 表構造情報，グラフ生成，グラフ自動生成，表構造解析

1 はじめに

学術論文には，実験結果や統計情報をまとめた表がよく利用される．しかし，実験結果をより視覚的に示すには，表よりもグラフが適している．そのため，論文内の表からグラフを自動生成することができれば，論文の内容をより容易に理解することができる．

表の書き方は書き手によって異なる．そのため，表からグラフを生成するには，まず表の構造を解析する必要がある．本稿では，表の構造解析後の情報を基にして棒グラフを生成する．また，表の構造解析には，青柳らが提案したグラフニューラルネットワーク（GNN）によるエンドツーエンド表構造解析手法 [1] を利用する．

青柳らは，トークンの水平結合，補助罫線推定，セル生成の 3 つの NN モジュールを用いて表の構造を解析した．実験では ICDAR 2013 の Table Competition のテスト用データセットを使用し，セルの隣接関係の再現性に基づいて評価した．その結果，セルの隣接関係の再現性の F 値が 0.984 となり，この結果は OCR で文書のデジタル化を行う商用ソフトウェアである ABBYY FineReader PDF[1] と比較して 2.4 ポイント高かった．

本論文では，高精度な青柳らの表構造解析手法を用いて生成される表構造情報を利用した棒グラフの自動生成手法を提案する．この表構造情報は，解析後に出力されるセルの 2 次元配列の情報である．さらに本研究では，罫線の情報を用いて，誤った表構造情報を修正する処理と表のサブヘッダ行/列を判別する処理を棒グラフ生成の前処理として実装した．

1：https://pdf.abbyy.com

評価実験では，ICDAR 2013 の Table Competition [2] のテストデータを評価データセットとして用い，表から棒グラフを生成し，適切な棒グラフが生成できるかを評価する．

2 関連研究

Kardas らは，機械学習に関する論文から結果を抽出するためのエンドツーエンドのモデル Automatic Extraction of Results from Machine Learning Papers (AxCell) を提案した [3]．抽出される結果は（タスク，データセット，評価指標名，評価指標に基づく値）という形式のタプルである．論文から抽出された結果は，他の論文で同様の（タスク，データセット，評価指標名）を持つ結果と比較するために 1 つの表（リーダーボード）に集約される．

Kardas らの手法は，論文の LaTeX ソースコードを入力とし，結果を抽出するタスクは以下の 3 つのサブタスクに分割される．

（1） 表のタイプ分類：論文内の表が関連する結果を持つかどうかを識別する．

（2） 表の分割：表のセルを，評価指標，データセット，モデルなどを保持しているかどうかに応じて分割し，分類する

（3） 結果のリーダーボードへの結び付け：結果のタプルを取り出し，既存の結果のリーダーボードに対応させる．

Kardas らは手法の性能を NLP-TDMS データセット [4] と PWC Leaderboards [3] データセットで評価した．NLP-TDMS データセットでは，タスク，データセット，評価指標，スコアの抽出結果の F1 値が 25.8 となり，Hou らの手法 [4] を 18.3 ポイント

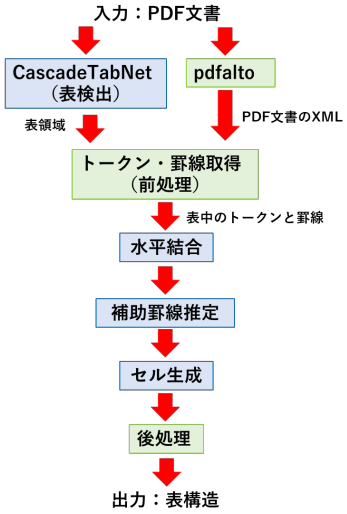


図 1　青柳らの表構造解析手法 [1] の概要

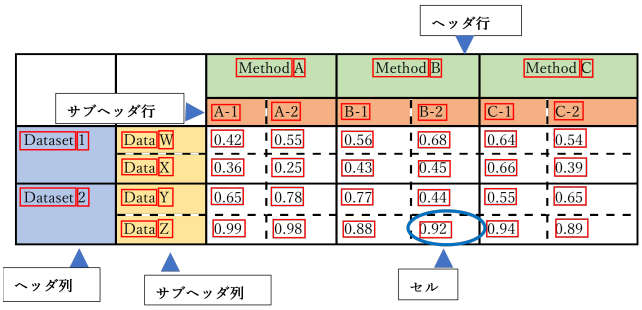


図 2　表と表の構成要素

	Method A	Method B
Data A	34.5	43.3
Data B	64.4	12.4

図 3　グラフ生成で利用する表構造情報の例

上回った．PWC Leaderboards データセットではタスク，データセット，評価指標，スコアの抽出結果の F1 値が 28.7 となった．

Herzig らは，表に対する質疑応答を行う手法 TAble PArSing(TAPAS) を提案した [5]．TAPAS は，質問に対する答えとなる表のセルの部分集合を選択し，質問からその部分集合に適用する合計や平均などの演算子を予測することで表に対する質問に応答する．表に対する質問は，表に対する単文の質問と，1 つ目の質問と 2 つ目の質問に関連がある会話形式の質問の 2 つがある．TAPAS のアーキテクチャは，BERT のエンコーダーをベースにしており，表構造を符号化するために埋め込みが追加されている．追加した埋め込みには，平坦化されたシーケンスにおけるトークンの索引を示す位置 ID やトークンが出現する列/行の索引を示す列/行 ID などがある．さらに，BERT のエンコーダーに，表のセルを選択するための 2 つの分類層と，セルに適用する集約演算子を追加する．

実験では，Wikipedia の表に関する質問で構成される WIKITQ，SQA，WIKISQL の 3 つのデータセットで評価した．WIKISQL に対しての解答の accuracy が 0.836 となり，Min らの手法 [6] に近い性能となった．WIKITQ に対しての解答の accuracy は 0.426 となり，類似する手法である Neelakantan ら [7] を 8.4 ポイント上回った．SQA に対して accuracy が 0.672 となり，比較手法の中で最も優れていた Müller ら [8] を 12.1 ポイント上回った．

青柳らの表構造解析手法の概要を図 1 に示す．青柳らの手法では，CascadeTabNet により検出された表の領域と，pdfalto[2] により PDF 文書から変換した XML ファイルを入力として表構造を解析する．表領域を利用して，XML ファイルから表中のトークンを取得する．同時に，PDFMiner[3] と OpenCV[4] を用いて文書の PDF から表中の罫線を取得する．次に，表中の水平方向に隣接する 2 トークンを，トークンの特徴を用いて再帰的に結合する水平結合を行う．結合し終えたら，トークンの位置関係に基づいて補助罫線を推定する．そして，垂直方向と水平方向の隣接 2 トークンを交互に再帰的に結合してセルを生成する．最後に，後処理で行や列の結合とセルの拡張を行い，最終的な表構造を決定する．

ICDAR 2013 で開催された Table Competition にて提供されたテスト用データセット [2] を用いて表構造解析を行った結果，セルの隣接関係の再現性を示す再現率，適合率，F 値がそれぞれ 0.984，0.985，0.984 となり OCR で文書のデジタル化を行う商用ソフトウェアである ABBYY FineReader PDF を再現率が 3.4 ポイント，適合率が 1.4 ポイント，F 値が 2.4 ポイント上回った．本稿は，この手法により得られる表構造情報を利用してグラフを生成する．

3　提案する棒グラフ生成手法

3.1　表の構成要素と表構造情報

図 2 に本稿で扱う表の例を示す．図 2 中の実線と点線はそれぞれ罫線，補助罫線を表している．補助罫線は実際には引かれていないが，セルを分割するために必要な線のことであり，青柳らの手法により推定される表構造情報の一部である．罫線もしくは補助罫線に囲まれているものをセルと呼び，セル内の赤い矩形で囲まれているものがトークンであり，表中の単語などに対応している．表の行と列のうちハイライトされているものは，緑色がヘッダ行，オレンジ色がサブヘッダ行，青色がヘッダ列，黄色がサブヘッダ列である．サブヘッダ行は 1 行とは限らず，サブヘッダ行がない，または複数行ある表も存在する．サブヘッダ列も同様である．

表構造情報は，青柳らの表構造解析によって出力されるセルの 2 次元配列が持つ情報である．グラフ生成で利用する表構造情報を図 3 の **Data A** のセルを例に説明する．なおすべての線分は罫線である．1 つ目はそのセルの開始行，開始列，終了行，終了列の索引である．図 3 の **Data A** のセルの索引はそれぞれ，開始行が 1，開始列が 0，終了行が 2，終了列が 1 と

2：https://github.com/kermitt2/pdfalto
3：https://github.com/pdfminer/pdfminer.six
4：https://opencv.org

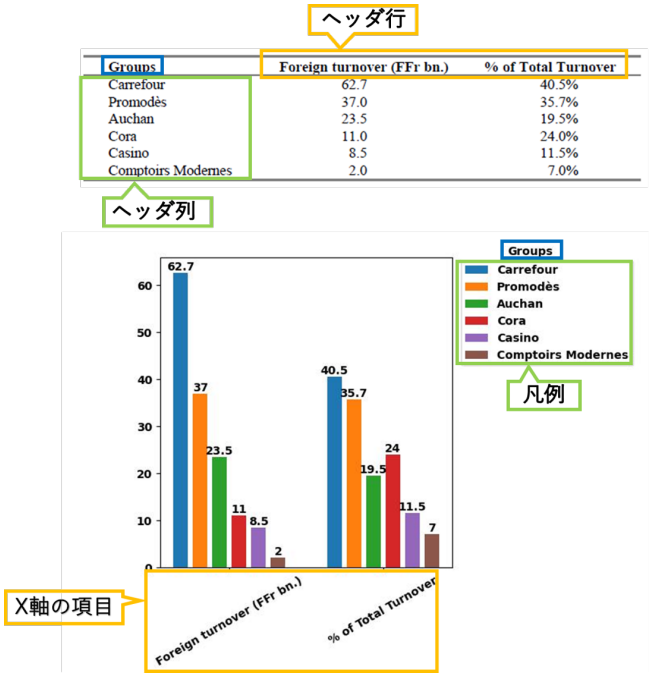


図 4　データセットの表 [2] と生成される棒グラフ

なる．2 つ目は，セルの上下左右に隣接する罫線の情報である．**Data A** のセルではそれぞれ，黄色の線が上部に隣接する罫線，赤色の線が左に隣接する罫線，青色の線が右に隣接する罫線，緑色の線が下部に隣接する罫線である．3 つ目は，PDF 文書内におけるセルの左端と右端の X 座標，上端と下端の Y 座標である．**Data A** ではそれぞれ，このセルを囲う黒線上の赤色と青色の点の X 座標，黄色と緑色の点の Y 座標となる．この座標は PDF 文書の左上を原点とする座標である．

3.2　棒グラフ生成手法

提案する棒グラフ生成手法について述べる．また，棒グラフ生成の前処理を提案する．前処理では，罫線を利用して表構造情報を修正し，表のサブヘッダ行/列を判別する．これらの前処理はそれぞれ 3.4，3.5 で述べる．

表と，その表から生成される棒グラフを図 4 に示す． 提案手法では，表のヘッダ行のセルの内容を棒グラフの X 軸の項目，ヘッダ列のセルの内容を凡例とする．

提案する棒グラフ生成手法では，Python の GUI アプリケーション作成のための標準ライブラリの 1 つである `TKinter` モジュールによって生成した PDF ビューアで PDF 文書を開き，表をドラッグし，表構造を解析する．その後，生成された表構造情報を利用して，棒グラフ画像を生成して表示する．

棒グラフ生成では，表の 2 次元配列の各セルを，ヘッダ行，ヘッダ列，数値のいずれかの配列に排他的に分類する．この処理によって表の 2 次元配列の 1 行目がヘッダ行，1 列目がヘッダ列，それ以外のセルが数値と分類する．数値は列の配列として代入する．ただし，2 次元配列の 1 行目 1 列目のセルはヘッダ行とヘッダ列のいずれにも分類しない．

数値を棒グラフにプロットするために，数値と分類されたセルに含まれる数値以外の文字列は取り除き，1 つのセルに複数

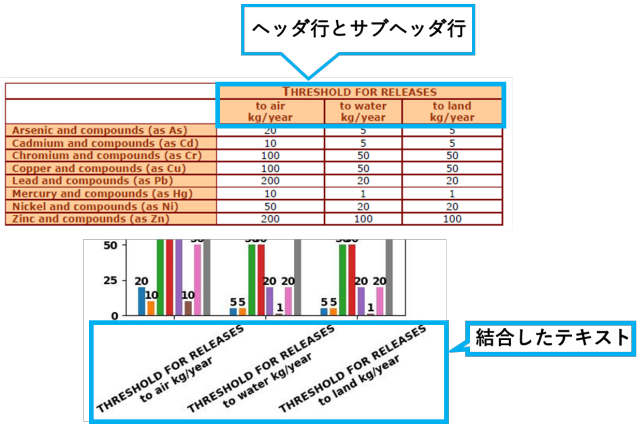


図 5　サブヘッダ行を持つ表 [2] と生成される棒グラフの X 軸の項目

の数値が入っている場合先頭の数値をそのセルの数値とする．

棒グラフの生成は，図 4 に示すように，表の列ごとに数値をプロットし，その列のヘッダ行のセルのテキストを数値の下部に X 軸の項目としてのせる．ヘッダ列のテキストは，凡例として数値がプロットされたグラフ領域の右側にのせる．1 行 1 列目のセルのテキストは多くの表でヘッダ列の説明をするものが多いため，凡例の上にのせる．

サブヘッダ行を持つ表と生成される棒グラフの X 軸の項目の例を図 5 に示す．表のサブヘッダ行を棒グラフ生成で使用する際は，ヘッダ行に結合し，1 行のヘッダ行とする．表のサブヘッダ列も同様にヘッダ列に結合し，1 列のヘッダ列とする．

3.3　表のキャプションの抽出とその利用

本研究において，表のキャプションは，生成するグラフにおいても重要な意味を持つと考えたため，表のキャプションを抽出する．表のキャプションの情報は，表構造解析の前処理で取得する表領域内のトークンと罫線の情報を利用して抽出する．取得するトークンの Y 座標と表の水平な罫線の Y 座標を利用し，表の最も上の水平な罫線よりも上にあるトークンが存在する場合は，そのトークンを表のキャプションとして抽出する．

上部から表のキャプションが抽出できなかった場合，表の最も下の水平な罫線の Y 座標とキャプションの Y 座標を比較し，罫線よりも下にあるトークンが存在する場合はそのトークンを表のキャプションとして抽出する．表の上部から抽出できたかった場合のみ，表の下部からキャプションを抽出する．

3.4　前処理 1：罫線を利用した過分割セルの結合処理

3.2 の方法では，表構造情報に誤りが含まれる場合，正確な棒グラフが生成できない．そのため，棒グラフ生成の前処理として，表の水平に引かれている罫線の数が表の行数以上ある表に対して，表中の水平方向の罫線の情報を利用して，セルの過分割を修正する．この処理は，まず，セルの上下にある水平方向の最近傍の罫線を求める．セルとその直下のセルにおいてそれらの罫線を比較し，どちらも等しい場合はそれらを一つのセルとして結合する．この処理を前処理 1 とする．

図 6 の表は，表構造情報では，第 1 列 2 行目のセルが赤線の所で上下に過分割されている．しかしセルの結合処理によって，

	All companies analysed	FTSE Eurotop 100 companies analysed
Number of member states in the analysis	21	8
Number of member states where one or more of the financial companies applied the amendment	11	3

図 6　表構造解析結果に誤りを含む表 [2] の例

Correlation coefficients of the I(1)* spread and yield differences for selected European countries

Differences with respect to Germany	Portugal	Greece	Spain	Italy	France
Q2 2006 – Q1 2010**	0.33	0.51	0.31	0.18	0.05
Q1 2009 – Q4 2009	-0.01	0.45	0.21	0.18	0.05
Q4 2009	0.17	0.70	0.26	0.09	-0.01
Q1 2010	0.64	0.72	0.56	0.43	0.25

Source: Bloomberg (CESR computations); data from the beginning to the end of the quarters; * the series are integrated of order 1, I(1), i.e. first differences are used to take account of non stationary. ** Considering the period Q1 2008 – Q1 2010 during which the sovereign CDS markets were better developed gives results which are very close to the ones presented.

図 7　水平な罫線がほとんど引かれていない表 [2] の例

サブヘッダ行

	number of franchisors		number of franchisees	
	1993	1994	1993	1994
Austria	80	170	2500	2700
Belgium/Luxembourg	90	135	3200	2495
Denmark	42	42	500	500
Finland	..	..	..	..
France	500	500	30000	30000
Germany	370	420	15500	18000
Greece	..	..	..	..
Ireland	20	..	..	..
Italy	318	361	16100	17500
Netherlands	331	340	12640	12120
Portugal	55	70	..	..
Spain	117	250	14500	20000
Sweden	200	200	900	900
UK	373	396	18100	24900
EU Total	2496	2884	113940	129115

図 8　判別処理 1 ではサブヘッダ行を適切に判別できない表 [2]

セルの過分割が修正される．

また，この処理は表の水平方向の罫線の数が表の行数以上の表に対してのみ行うため，図 7 の例のような表は，この修正処理の対象とならず，セルの誤結合は抑制される．

3.5　前処理 2：サブヘッダ行/列の判別

本研究では，棒グラフ生成の前処理として，表の各行および各列がサブヘッダ行/列かどうか判別する処理を提案する．この判別処理では，各表の 1 行目，1 列目はヘッダ行，ヘッダ列とし，サブヘッダ行やサブヘッダ列は，ヘッダ行，ヘッダ列と隣接しているものとする．また，提案する判別処理は 2 つの工程からなり，順に判別処理 1，判別処理 2 として説明する．判別処理 1 によるサブヘッダ行の判別では，ヘッダ列を除く各列で数値を 1 つも取り出せなかったセルが何行目に含まれるかを求める．その後，数値を含まないセルが最も多い行を探して，ヘッダ行のすぐ下からその行までをサブヘッダ行とする．この処理では，図 8 の第 6 行目や第 9 行目のテキスト . . のような数値を持たないことを示すテキストの場合は，そのセルでは数値を取り出せたものとして例外的に処理する．

判別処理 2 では，数値を含む行と隣接するサブヘッダ行，又はヘッダ行内に単一のセルがない場合，その行に隣接する数値を含む行もサブヘッダ行と判別する．

サブヘッダ列の判別処理は，判別処理 1，2 で行と列を入れ替えて同様に行う．

判別処理 1 では，図 8 の赤線で囲われた 2 行目のように，サブヘッダ行が西暦などの数値の場合，その行をサブヘッダ行と判別できない．多くの表では，表内の数値データと隣接するサブヘッダ行，またはヘッダ行のセルは 1 つの行，1 つの列に属するセルである．判別処理 2 はこれを利用した．

次に判別処理 1 で数値を取り出すために利用する正規表現 `[+-]?\d+[.]?\d*`の意味を示す．

（1）　`[+-]?`: + または − を 0〜1 個含む．
（2）　`\d+`: 0〜9 の数字を 1 回以上繰り返す．
（3）　`[.]?`: 小数点「.」を 0〜1 個含む．
（4）　`\d*`: 0〜9 の数を 0 回以上繰り返す．

この正規表現と，文字列内の条件に合致するすべての部分文字列を見つけてリストとして返す `re.findall` 関数を利用して各セルのテキストから数値を取り出す．

4　グラフ生成の実験

4.1　表データセット

グラフ生成の評価には，ICDAR 2013 Table Competition [2] のテスト用データセットを用いる．この表データセットは，EU および米国（US）政府の発行した文書から収集した 156 の表，それらの表を含む PDF 文書，表構造の付与された XML からなる．このデータセットの中で，棒グラフ生成の対象とする数値を含む表は，EU の 73 表，US の 71 表の合計 144 表である．

4.2　棒グラフの自動生成実験

4.1 のデータセットの表を対象として棒グラフの自動生成を行い，その精度を式 (1) で求められる成功率で評価した．

$$\text{成功率} = \frac{\text{グラフ化に成功した表の数}}{\text{グラフ化対象の表の数}} \quad (1)$$

また，3.4 と 3.5 で述べた 2 つの前処理の有無による成功率の変化を調べた．棒グラフの成功の定義は，グラフ上の数値と表の数値の対応が正しく，かつ，凡例，X 軸の項目のテキストが表のテキストと比較して適切であることとした．判定は著者が行った．

前処理 1 と前処理 2 をどちらも行った場合の棒グラフの自動生成の結果を表 1 に示す．どちらの前処理も行わなかった場合，前処理 1 のみ行った場合，前処理 2 のみ行った場合の結果をそれぞれ，表 2，表 3，表 4 に示す．2 つの前処理を行った場合の結果である表 1 では，EU で成功率 0.93，US で成功率 0.74 となった．表 1 と表 2 を比較すると，提案した前処理 1 と前処理 2 によって，成功率が EU で 22 ポイント，US で 5 ポイント上昇した．表 1 と表 3 を比較すると，前処理 2 によって，成功率が EU で 19 ポイント，US で 5 ポイント上昇したことがわかる．しかし，表 1 と 4 の比較からは，前処理 1 を使用しても，US の成功率は変わらず，EU の成功率も 3 ポイントしか上昇しなかったことがわかる．

また，表 1 から表 4 に示したどの結果でも US の成功率は EU と比べて低かった．表 1 の結果では，EU では成功率が 0.93

表 1　前処理 1 と 2 を両方行った場合の棒グラフ生成の成功率

データ	グラフ化対象の表の数	グラフ化に成功した表の数	成功率
EU	73	68	0.93
US	71	53	0.74

表 2　前処理 1 と 2 のいずれも行わなかった場合の棒グラフ生成の成功率

データ	グラフ化対象の表の数	グラフ化に成功した表の数	成功率	表 1 の成功率との差
EU	73	52	0.71	0.22
US	71	49	0.69	0.05

表 3　前処理 1 のみを行った場合の棒グラフ生成の成功率

データ	グラフ化対象の表の数	グラフ化に成功した表の数	成功率	表 1 の成功率との差
EU	73	54	0.74	0.19
US	71	49	0.69	0.05

表 4　前処理 2 のみを行った場合の棒グラフ生成の成功率

データ	グラフ化対象の表の数	グラフ化に成功した表の数	成功率	表 1 の成功率との差
EU	73	66	0.90	0.04
US	71	53	0.74	0.00

だったのに対して，US での成功率は 0.74 であり，EU と比較して 19 ポイント低かった．EU と比較して US の結果が悪くなった一因として，US は大規模な表が多く，表構造解析でも完全に解析に成功している表が少ないことが挙げられる．表構造解析に完全に成功した表の数は，EU で 73 のうち 67 表，US で 71 のうち 39 表となっており，US は EU と比べて，表構造解析に完全に成功した表の数が少ない．

表 5 に，1 表あたりの平均行数と平均列数を示す．平均行数と平均列数のどちらも US の方が大きいことから，US の表の方が EU の表と比較して大規模なものが多いと言える．

また，青柳らの方法で構造解析に失敗している表は罫線が少ないため，前処理 1 の修正処理が適用されなかった．図 9 は US で棒グラフの生成に失敗した表の例である．この表では，赤線で囲われたヘッダ列のセルが 2 列目にかかっていることにより，このヘッダ列のセルが複数列にまたがるセルであると解析された．また，図 9 の表は垂直な罫線が少なく，サブヘッダ行内の青線で囲われた 2 つの異なるセルは，セル間の距離も狭いため一つのセルであると解析された．

図 10 は，追加した棒グラフ生成の前処理 1 の修正処理によって表構造情報を適切に修正できた表の例である．修正処理の前は，表の第 1 列 2 行目のセルが赤線で上下に過分割され，2 つの異なるセルと解析されていた．そのため，本来は 2 行の表であるにもかかわらず，3 行の表と解析されていた．前処理 1 によって適切に修正が行われており，ヘッダ列のこれらのテキストが 1 つのセルに結合された．

表 5　表の平均行数と平均列数

データ	表の平均行数	表の平均列数
EU	11.9 行	4.6 列
US	16.7 行	6.4 列

2列目

Student and institutional characteristics	Percent who borrowed: Neither	Under-graduate only	Graduate only	Both	Average amount borrowed (by borrowers): Under-graduate only	Graduate only	Both
Total	31.4	23.7	17.7	27.2	$9,600	$36,900	$41,700
Type of degree-granting institution							
Public 4-year	32.6	23.0	19.5	24.9	8,100	36,500	36,900
Non-doctorate-granting	34.4	25.6	15.9	24.1	7,800	29,100	31,900
Doctorate-granting	31.7	21.7	21.3	25.2	8,200	39,300	39,300
Private not-for-profit 4-year	29.7	24.2	15.3	30.8	12,500	37,500	48,900
Non-doctorate-granting	27.5	26.1	14.5	32.0	12,000	31,500	43,200
Doctorate-granting	32.1	22.3	16.1	29.5	13,000	43,200	55,500
Other	26.9	31.8	7.1	34.2	‡	‡	39,600
Undergraduate major							
Business and management	39.6	26.4	13.0	21.1	8,900	33,000	38,800
Education	34.6	31.0	12.8	21.7	10,100	28,600	30,800
Engineering, mathematics, or science	29.4	23.3	19.1	28.2	10,000	47,200	57,400
Humanities or social sciences	28.6	18.4	21.8	31.1	9,600	38,200	40,000
Other	29.3	23.4	17.8	29.5	9,500	31,000	37,400
Amount borrowed (undergraduate)							
Did not borrow	64.0	†	36.0	†	†	36,900	†
Less than $5,000	†	54.4	†	45.7	2,700	†	32,700
$5,000–9,999	†	43.7	†	56.3	6,900	†	36,200
$10,000–14,999	†	45.3	†	54.7	11,300	†	40,200
$15,000 or more	†	41.6	†	58.4	22,400	†	57,700
Highest enrollment after bachelor's degree by 2003							
Master's degree	34.3	27.8	13.6	24.3	9,500	22,600	29,000
Doctoral degree	23.6	13.2	28.3	34.9	10,400	44,300	54,900
First-professional degree	20.4	7.6	33.6	38.5	‡	68,000	82,000
Highest degree earned by 2003							
Bachelor's degree	35.5	32.0	10.8	21.7	9,500	25,500	32,500
Master's degree	31.8	22.8	17.5	27.9	9,500	25,200	32,500
Doctoral degree	23.7	11.8	31.2	33.3	‡	46,600	52,900
First-professional degree	17.8	3.5	37.6	41.1	‡	73,900	88,900

図 9　表構造解析の失敗が原因で棒グラフ生成に失敗した US の表 [2] の例

Years	1990	1992	1993	1994	1995	1996
Budget Coefficient for Coffee, Tea, Herb Tea	0.323	0.272	0.265	0.290	0.340	0.346

図 10　表構造情報を修正できた表 [2] の例

4.3　キャプションの抽出実験

キャプション抽出実験を行った．抽出されたキャプションと本来の表のキャプションを比較し，成功率で評価する．実験の対象は，数値を含んでいてキャプションを持つ，EU の 67 表と US の 65 表の合計 132 表とした．キャプション抽出の実験結果を表 6 にまとめる．結果は EU で成功率 0.98，US で 1.00

表 6　キャプション抽出の成功率

データ	キャプション抽出対象の表の数	キャプション抽出に成功した表の数	成功率
EU	67	66	0.98
US	65	65	1.00

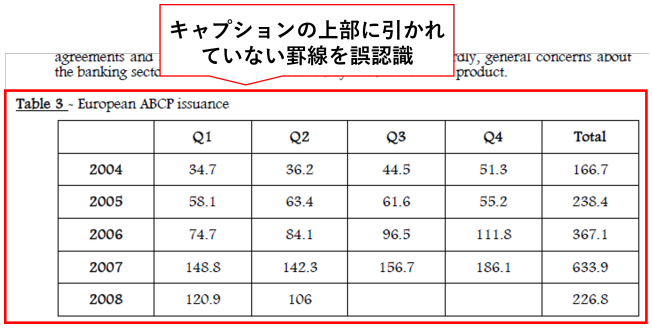


図 11　EU データでキャプションの抽出に失敗した表 [2]

となり，キャプション抽出に失敗した表は EU のデータの 1 件のみだった．

図 11 は，EU のデータでキャプション抽出に失敗した表である．この表では，キャプションの上部に罫線はないが，表構造解析により誤認識された．そのためキャプション抽出に失敗した．

4.4　表構造解析に関する考察

表 1 の結果で棒グラフ生成に失敗した 23 表は，すべて表構造解析に失敗しており，正しい表構造情報があればすべて適切な棒グラフを生成することができた．表構造解析に失敗したため，棒グラフが適切に生成できなかった 23 表について考察する．失敗した 23 表は，以下の特徴のいずれかに該当するものであった．

（1） 垂直な罫線が表の列数よりも少ない
（2） 表内に 3 行以上のテキストで構成されるセルがある
（3） セル内のテキストのトークン間の間隔が大きい
（4） 隣接するセルの間隔が小さい

(1) に該当する，図 9 の表のように垂直な罫線がほとんど引かれていない表では，水平方向に隣接する異なるセルを 1 つのセルであると解析した．EU で失敗した 5 表の内 1 表，US で失敗した 18 表の内 15 表がこれに該当した．さらに，US の該当する 15 表の内 13 表は，垂直な罫線が 1 本もなかった．

(2) の，3 行以上のテキストで構成されるセルを持つ表では，1 つのセルのテキストの行間に補助罫線が推定され複数の異なるセルに過分割していた．失敗した 23 表の内，EU の 1 表がこれに該当した．

(3) の，セル内のテキストのトークン間の間隔が大きい表も，(2) と同じように単語間に補助罫線が推定され複数の異なるセルに過分割していた．EU で失敗した 5 表の内 3 表，US で失敗した 18 表の内 1 表がこれに該当した．

(4) の，隣接するセルの間隔が小さい表では，異なるセルが 1 つのセルと解析することが多かった．失敗した 23 表の内，US の 2 表がこれに該当した．

4.5　棒グラフ以外のグラフ生成

4.2 では，提案手法で棒グラフを生成したが，表の内容によっては，他のグラフが適している場合がある．そのため，円グラフと折れ線グラフとレーダーチャート生成する処理について述べる．図 12 と図 13，図 14 に表とそれを用いて生成される折れ線グラフ，円グラフ，レーダーチャートを示す．折れ線グラフでは，棒グラフと反対にヘッダ列のテキストを X 軸の項目に，ヘッダ行のテキストを凡例とする．円グラフでは，ヘッダ行のテキストを各円グラフの上にラベルとして，ヘッダ列のテキストを各円グラフの軸ラベルとする．レーダーチャートではヘッダ列のテキストを軸ラベルとして，ヘッダ行のテキストを凡例とする．

5　おわりに

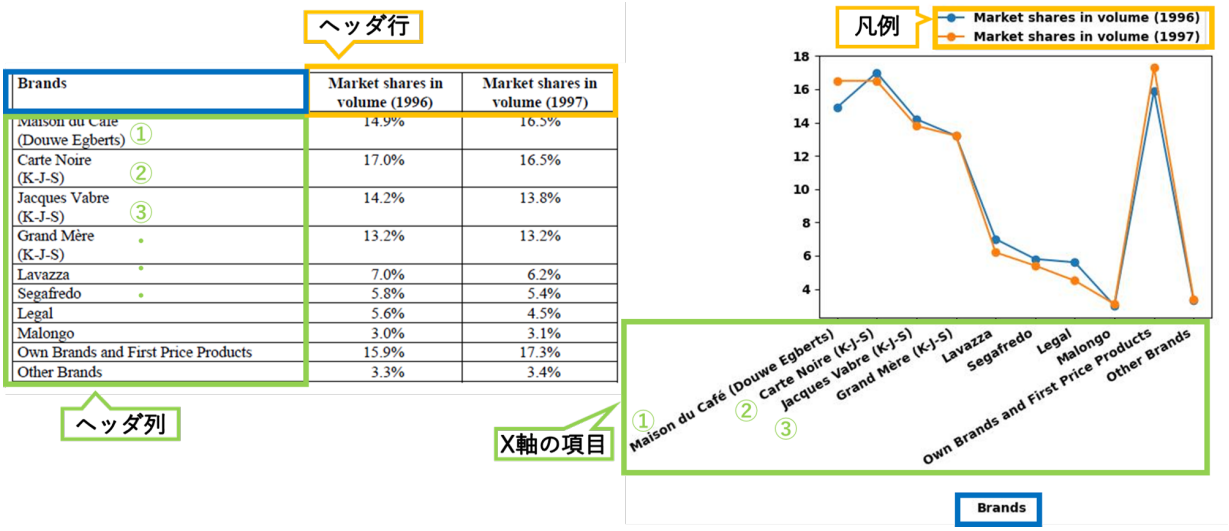
本稿では，表構造情報を利用した棒グラフ生成手法を提案した．提案手法では棒グラフの生成において，サブヘッダ行やサブヘッダ列をそれぞれそれらに隣接するヘッダ行やヘッダ列に結合して処理することで，サブヘッダ行やサブヘッダ列を持つ表にも対応した．また，罫線の情報を用いてセルの過分割を修正する前処理と表のサブヘッダ行/列を判別する前処理を提案し，グラフ生成の成功率を向上させた．ICDAR 2013 の Table Competition のテスト用データセットの表を用いて評価し，構造が比較的単純な EU のデータにおいて棒グラフ生成の成功率は 0.93，構造が比較的複雑な US のデータで 0.74 となった．キャプション抽出の成功率は EU のデータで 0.98，US のデータでは 1.0 だった．今後の課題として，グラフ生成とキャプション抽出ともに，ICDAR 2013 Table Competition のテスト用データセット以外のデータセットにおいて精度を評価することが挙げられる．また，論文内の表から実験の結果を抽出する `AxCell` [3] や表に対する質疑応答を行う `TAPAS` [5] のように表の情報をグラフ生成以外へ活用する方法の検討が挙げられる．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B)(課題番号 22H03904)．および 2023 年度国立情報学研究所公募型共同研究 (23FC02) の援助による．

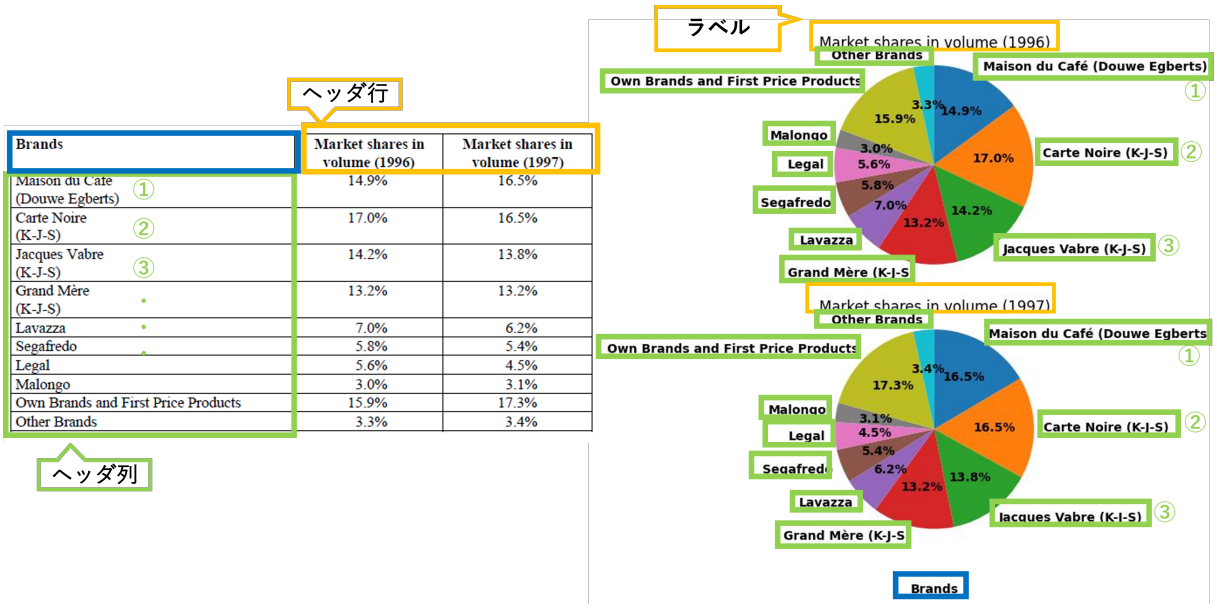
文　献

[1] 青柳拓志ほか，グラフニューラルネットワークを用いたエンドツーエンド表構造解析手法の提案，第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023)，1b-8-4，2023.
[2] Max Göbel, et al., ICDAR 2013 table competition, In Proceedings of the 2013 International Conference on Document Analysis and Recognition, pp. 1449-1453, 2013.
[3] Marcin Kardas, et al., AxCell: Automatic Extraction of



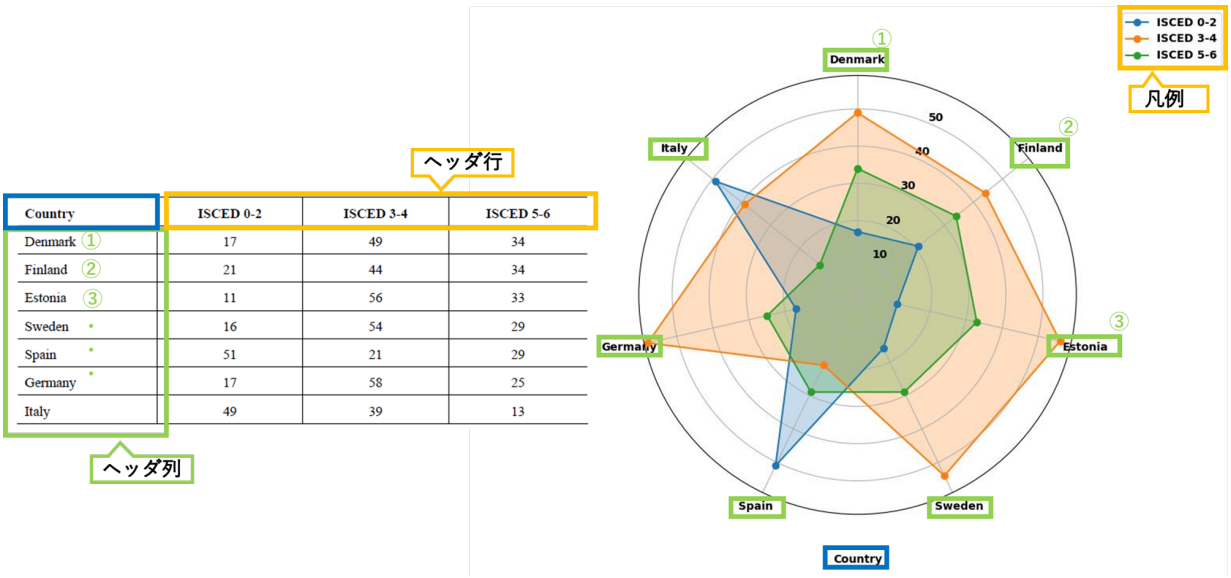
Brands	Market shares in volume (1996)	Market shares in volume (1997)
Maison du Café (Douwe Egberts) ①	14.9%	16.5%
Carte Noire (K-J-S) ②	17.0%	16.5%
Jacques Vabre (K-J-S) ③	14.2%	13.8%
Grand Mère (K-J-S)	13.2%	13.2%
Lavazza	7.0%	6.2%
Segafredo	5.8%	5.4%
Legal	5.6%	4.5%
Malongo	3.0%	3.1%
Own Brands and First Price Products	15.9%	17.3%
Other Brands	3.3%	3.4%

図 12　データセットの表 [2] と生成される折れ線グラフ



Brands	Market shares in volume (1996)	Market shares in volume (1997)
Maison du Café (Douwe Egberts) ①	14.9%	16.5%
Carte Noire (K-J-S) ②	17.0%	16.5%
Jacques Vabre (K-J-S) ③	14.2%	13.8%
Grand Mère (K-J-S)	13.2%	13.2%
Lavazza	7.0%	6.2%
Segafredo	5.8%	5.4%
Legal	5.6%	4.5%
Malongo	3.0%	3.1%
Own Brands and First Price Products	15.9%	17.3%
Other Brands	3.3%	3.4%

図 13　データセットの表 [2] と生成される円グラフ



Country	ISCED 0-2	ISCED 3-4	ISCED 5-6
Denmark ①	17	49	34
Finland ②	21	44	34
Estonia ③	11	56	33
Sweden	16	54	29
Spain	51	21	29
Germany	17	58	25
Italy	49	39	13

図 14　データセットの表 [2] と生成されるレーダーチャート

Results from Machine Learning Papers, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8580-8594, 2020.

[4] Yufang Hou, et al., Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction, In Proceedings of ACL 2019, pp. 5203-5213, 2019.

[5] Jonathan Herzig, et al., TAPAS: Weakly Supervised Table Parsing via Pre Training, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp.4320-4333, 2020.

[6] Sewon Min, et al., A discrete hard EM approach for weakly supervised question answering, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2851-2864, 2019.

[7] Arvind Neelakantan, et al., Neural programmer: Inducing latent programs with gradient descent, CoRR, abs/1511.04834, 2015.

[8] Thomas Müller, et al., Answering conversational questions on structured data without logical forms. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5902-5910, 2019.

条件付きトランスフォーマーとテキストリフレーサーに基づくクワッドローター用視覚言語ナビゲーション

Zhe CHEN†, Jiyi LI†, Fumiyo FUKUMOTO†, Peng LIU††, and Yoshimi SUZUKI†

† University of Yamanashi（山梨大学）, Japan
†† Hangzhou Dianzi University（杭州電子科技大学）, China
E-mail: †{chenzhe,perryliu}@hdu.edu.cn, ††{jyli,fukumoto,ysuzuki}@yamanashi.ac.jp

Abstract 自然言語の指示によるドローンの制御は、ビジョン・アンド・ランゲージ・ナビゲーション（VLN）における重要な課題である。しかしながら、これまでのモデルは、マルチモーダルな特徴を効果的に統合してドローンを導くことができなかった。それらのモデルの多くは、指示と環境の文脈との相関関係を活かさず、また自然言語を理解するモデルの能力を考慮していない。したがって、我々は新しい言語強化型クロスモーダルモデルを提案する。このモデルには、条件付き Transformer があり、マルチモーダルな特徴（テキスト指示と視覚的コンテキスト）を効果的に統合する。言語表現能力を向上させるために、SentenceBERT も採用している。さらに、同じナビゲーションタスクに対してユーザーがさまざまなテキスト指示を提供できるという問題に対処するため、LLM ベースの中間コンポーネント（LLMIR）を導入したプロンプトベースのアプローチを提案した。我々はクアッドコプターシミュレータを用いてこれらの手法を評価した。提案方法は、絶対的なタスク完了率を 1.39%向上させている。LLMIR を評価するために、元のテストセットから重要で最小限の指示を抽出して新しいテストセットを作成した。LLM を使用することで、タスク完了率は 1.51%向上し、新しいテストセットと元のテストセットとの性能差が 34.83%縮小した。この論文は ACM Multimedia Asia 2023 で発表された [7]。ソースコードは `https://github.com/MOZhuai/DroneVLN` で共有されている。

Key words 視覚言語ナビゲーション、クワッドローター、マルチモーダル、大規模言語モデル

1 Introduction

Asking a robot to execute natural language navigation instructions requires addressing challenges in language comprehension, visual perception, action planning, and multimodal fusion. The quadcopter is one of the most flexible robots and has promising prospects [3, 13]. A quadcopter that can understand human language and navigate intelligently would significantly benefit human society. To follow the instruction, the quadcopter must interpret visual scenes accurately, translate high-level language instructions into low-level control commands, and effectively integrate textual and visual information to adapt its navigation strategies over time [2, 3, 20]. VLN [1, 9, 10, 13] is a multimodal research field that combines computer vision and natural language processing. Its primary goal is to enable intelligent agents to autonomously navigate and execute tasks in complex environments with natural language instructions. Most of the VLN work is focused on the ground or indoor agents like Room-to-Room (R2R) [1] and its various extended versions [15, 17, 31], while there are limited studies [2, 3, 20] in the drone domain.

In this paper, we investigate the problem of using a crowdsourced natural language navigation instruction corpus in a continuous action space to navigate the quadcopter. Fig. 1 shows the demonstration of task. We refer to the structure of PVN [3], which decomposes into two stages: trajectory prediction and decision making. PVN employs a conventional LSTM [12] recurrent neural network to process natural language instructions, which is not effective in capturing long-range dependencies and dealing with complex navigation instructions. When integrating language and visual information, PVN calculates the semantic map using a convolution kernel

Figure 1 Demonstration of VLN task in real quadcopter simulator. The nature language instruction, third- and first-person views, and global trajectories in a top-down view is shown. The agent follows the instruction by the first-person observations and poses. The two paths represent trajectories executed by expert (red) and agent (blue), respectively.

generated by the instruction embedding. It reduces computational complexity, but the model can't learn and distinguish the relations between the specific words and image regions autonomously, which restricts its ability to capture complex correlations between images and text, as information can only be propagated through size-limited convolution kernels.

To address these challenges, we propose a language-enhanced cross-modal model. Our model employs a novel conditional Transformer (VLCT), which regards the natural language as a condition for visual features to effectively integrate the multimodal features and help the model understand the global interdependence between vision and language. Additionally, to make the model better represent the navigation instruction sentence, we also employ a pre-trained language model, Sentence-BERT [22] (SBERT), which can provide sentence-level embeddings to effectively improve the language representation and generalization ability of the model.

On the other hand, language models pre-trained on large-scale data and fine-tuned on specific domains perform very well on specific tasks [8]. We consider that if the system is deployed in practical scenarios, due to the different understanding of the system, users may give out-of-distribution (OOD) [19] instructions, such as a few simple words or incomplete instructions. It could potentially lead to the wrong representation of the model.

To solve this problem, we additionally introduce an LLM-based intermediary component as an instruction rephraser. When users provide OOD instructions, this component can rephrase and align these instructions with model required instruction style. Even when users provide only partial information, the model can endeavor to generate complete instructions, thereby making interactions more convenient and user-friendly.

To summarize, our contributions are mainly three-fold. (1) We propose a novel conditional Transformer (VLCT) to facilitate the fusion of vision and language features. (2) We adopt a pre-trained language model SentenceBERT (SBERT) to enhance the model's representation ability. (3) We design a prompt-based intermediary component, LLM-based Instruction Rephraser (LLMIR), to address the issue of OOD instructions from practical scenarios.

We evaluate our approach on the LANI corpus. By applying the SBERT and VLCT, our model improves the absolute task success rate (SR) by 1.39%, and a relative improvement by 4.30% on average stopping distance (AD), 1.44% on median stopping distance (MD). To evaluate the LLMIR, we keep the original target of each task and create a new test set by rephrasing the test instructions into sequences of triplets. After usin LLMIR, our model improves the SR of 1.51% on the new test set, and gets a 6.18% improvement on AD, a 5.63% improvement on MD. LLMIR narrows the performance gap between the new test set and the original test set by 34.83%. we have released the source code to facilitate the reproduction.

2 Related Work

Vision-and-Language Navigation VLN, which focuses on developing algorithms and models to navigate robots based on both visual input and textual instructions, has received significant attention [1, 3, 6, 15–17]. MTST [33] leverages external resources to mitigate data scarcity in outdoor VLN tasks. Robo-VLN [14] reconstructs a continuous 3D environment for VLN, proposing a Hierarchical Cross-Modal (HCM) agent that outperforms existing baselines by employing layered decision making, decoupling reasoning and imitation learning. PVN [3] presents a novel approach for controlling a quadcopter through natural language instructions and observations, using a two-step model that predicts position-visit distributions. Our model is based on the architecture of PVN [3].

Cross-modal Attention Cross-modal attention [29] is crucial for capturing relationships between different data modalities. It plays a pivotal role in multimodal information processing by enhancing fusion and understanding of diverse data sources. Some methods focus on aligning different modalities. CASC [30] CASC utilizes cross-modal attention for local alignment and multilabel prediction for global semantic consistency in image-text matching. In the image generation field, stable diffusion [23] employs cross-attention and domain-specific encoders to enhance conditioning capabilities. Our work refers to the conditioning mechanism from stable diffusion to align and fuse the textual and visual features.

Large Language Model (LLM) LLMs [4, 28] have shown impressive reasoning capabilities in natural language processing tasks. VELMA [26] utilizes an LLM to generate discrete navigation actions based on textual description, which combines historical navigation information. Control-GPT [32] leverages the precise code generation ability of GPT-4 [21] to generate programmatic sketches to enhance diffusion-based text-to-image pipelines, thereby improving user control over image attributes. These applications of LLMs often exhibit strong practical utility, which inspires our proposed LLM-based instruction rephraser aiming to overcome the challenge brought by OOD instructions in practical scenarios.

3 Proposed Method

3.1 Problem Setting

Given a nature language instruction u consisting of a sequence of words. At each time step t, the agent obtains a first-person view observation O_t from its current pose P_t. The agent should infer an action a_t, a tuple of forward velocity υ, and angular velocity ω that transfers the agent from state s_t to s_{t+1}. Each state contains all relevant information in the current environment, including the history visual information, poses and instruction. The state of agent at time step t cam be expressed as $s_t = (u, \mathbb{O}_t, \mathbb{P}_t)$, where $\mathbb{O}_t = \{O_i \mid i = 1, \ldots, t\}$ and $\mathbb{P}_t = \{P_i \mid i = 1, \ldots, t\}$ are the sequences of history observations and poses. Each P_i is a 6-DOF pose that encapsulates the spatial position p_i and orientation o_i. The agent

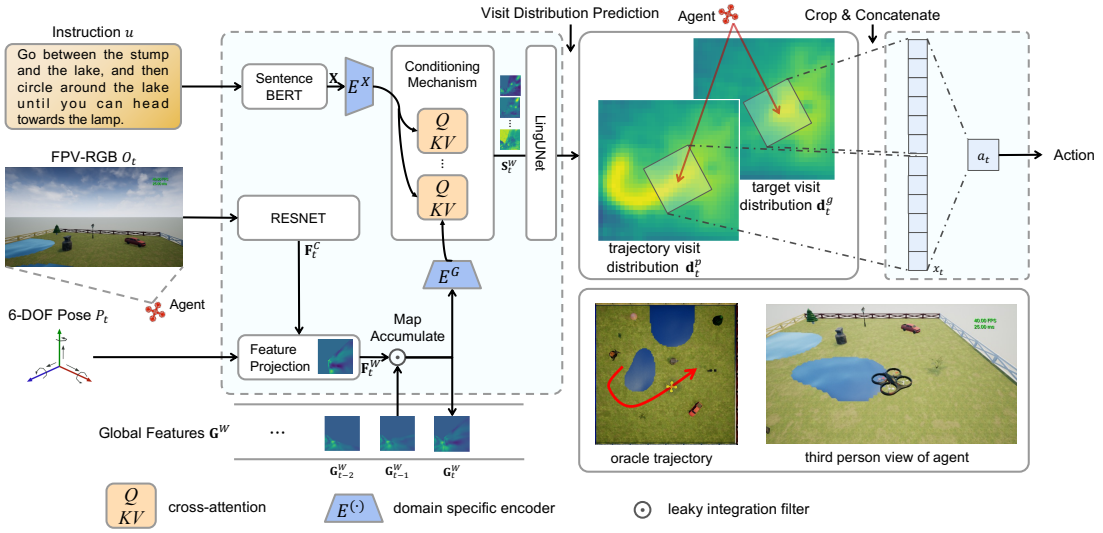


Figure 2 Schematics of Our Model. The instruction is encoded using a SentenceBERT network (Sec. 3.4). At each time step t, the FPV observation feature is projected to the world reference frame to accumulate the global feature map $\mathbf{G}_t^W$. The feature $\mathbf{G}_t^W$ is conditioned with instruction embedding to calculate the semantic map $\mathbf{S}_t^W$ (Sec. 3.5). The semantic map is used to generate two visit distributions (Sec. 3.6), which indicate the agent where to go and stop to predict the action (Sec. 3.7).

follows the instructions by generating a sequence of actions. The agent completes the task when it generates the STOP action.

3.2 Overview of Framework

The model architecture is illustrated in Fig. 2. We use a neural network to build the navigation model π. At time step t, the model's input is the current state s_t, and the outputs are an action a_t and a probability p_t^{stop}. The action a_t is a continuous variable that describes the forward velocity v_t and angular velocity ω_t, and p_t^{stop} denotes the probability of STOP action. Our model refers to the structure of the PVN [3] model. Based on this structure, we integrate the conditional Transformer to fuse instruction features and visual contexts and adopt a pre-trained language model SBERT to enhance the representation ability. The model comprises two primary elements: visit prediction and decision making.

At each time step t, visit prediction first computes a top-down semantic map $\mathbf{S}_t^W$, which combines the features of accumulated observations and natural language instructions. The map $\mathbf{S}_t^W$ is used to compute two probability distributions: trajectory visit distribution $\mathbf{d}_t^p$ and target visit distribution $\mathbf{d}_t^g$. The distribution $\mathbf{d}_t^p$ predicts the probability of visiting each position in the map, and $\mathbf{d}_t^g$ predicts the probability of choosing the STOP action at each position. The model computes the distributions every T_d time steps.

In the stage of decision making, we apply an affine transformation to align the distributions with the current pose P_t of the agent. The new distributions $\mathbf{d}_t^{g,ego}$ and $\mathbf{d}_t^{p,ego}$ will be used to predict the current action a_t and stopping probability p_t^{stop}.

3.3 Feature Projection and Map Accumulation

To predict the visit distributions, the model firstly uses a 13-layer residual neural network [11] to process the current observation O_t, obtaining a feature $\mathbf{F}_t^C = \text{RESNET}(O_t)$. Then compute the feature map in the world coordinate frame $\mathbf{F}_t^W$ by projecting the feature map $\mathbf{F}_t^C$ onto the ground plane at zero elevation.

We update the global feature map $\mathbf{G}^W$ in the world coordinate frame using the projected feature map $\mathbf{F}_t^W$. The new global feature map $\mathbf{G}_t^W$ is computed by applying a leaky integration filter [2] on the history feature map $\mathbf{G}_{t-1}^W$ and $\mathbf{F}_t^W$:

$$\mathbf{G}_t^W = (1-\lambda)\,\mathbf{G}_{t-1}^W + \lambda(\mathbf{F}_t^W \odot \mathbf{M}_t^W) + \lambda(\mathbf{G}_{t-1}^W \odot (1-\mathbf{M}_t^W)) \qquad (1)$$

where $\mathbf{M}_t^W$ is a binary value mask that indicates which pixels in $\mathbf{G}_t^W$ are within the field of view of the agent, $\odot$ is element-wise multiplication, λ is used to control the weight of the projected feature map $\mathbf{F}_t^W$ in the global feature $\mathbf{G}^W$. The leaky integration filter ensures that unobserved positions are not updated, and integrates new visual information with the existing information in the map.

3.4 Instruction Embedding

To improve the representation capacity, we introduce a pre-trained language model Sentence-BERT [22], an extension of the BERT [8] model designed for sentence-level semantic understanding. The instruction embedding is expressed as $\mathbf{X} = \text{SBERT}(u)$.

3.5 Conditioning Mechanism

We enhance the fusion ability with an attention-based conditioning mechanism [23], Condition Transformer (VLCT). Attention mechanism [29] is a widely used technique in deep learning. It allows the model to assign different weights according to different parts of the input. The attention mechanism calculates the similarity between query $\mathbf{Q}$ and key $\mathbf{K}$, and assigns these similarities as weights to the corresponding value $\mathbf{V}$. The output matrix of atten-

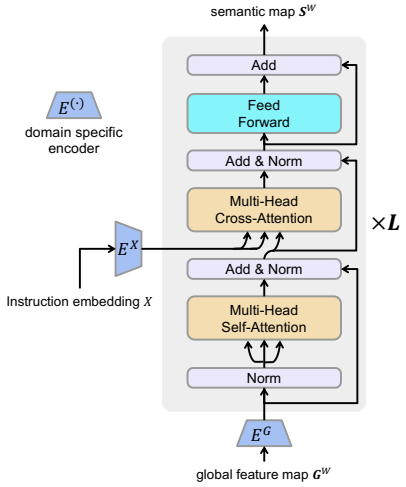


Figure 3 Condition Transformer.

tion mechanism is expressed as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where d_k is the dimension of key.

VLCT is a variant of multi-layer Transformer [29] treating the instruction features $\mathbf{X}$ as a condition to control the generation of semantic map S^W. It consists of multiple multi-head self-attention and multi-head cross-attention modules, with several domain-specific encoders that project the features to an intermediate representation. As shown in Fig. 3, it takes instruction features $\mathbf{X}$ and previous fusion features $\mathbf{H}^{l-1}$ as inputs and uses residual connections between each two attention modules:

$$\mathbf{H}^l = \text{VLCT}(\mathbf{H}^{l-1},\mathbf{X}), l \in [1,L] \tag{3}$$

where L is the layer number of VLCT.

Domain specific encoder. To preprocess the features $\mathbf{y}$ extracted from different modalities, we perform feature mapping to bring these two modalities into a common feature space, ensuring effective interaction. Specifically, we apply a domain-specific encoder $E(\cdot)$ that projects $\mathbf{y}$ to an intermediate representation $E(\mathbf{y}) \in \mathbb{R}^{N_E \times d_E}$, where $N_E \times d_E$ is the shape of the intermediate representation. Specifically, we preprocess the instruction feature and global feature map as $E^X(\mathbf{X}) \in \mathbb{R}^{N_E^X \times d_E^X}$ and $E^G(\mathbf{G}^W) \in \mathbb{R}^{N_E^G \times d_E^G}$. Here $E^G(\mathbf{G}^W)$ is the first input $\mathbf{H}^0$ of VLCT in Eq.(3).

Self-Attention on Visual Features. In each layer, VLCT uses self-attention to process the feature map from the previous layer at first. For the layer l, we calculate the query, key and value as:

$$\begin{cases} \mathbf{Q}_l^A = \mathbf{H}^{l-1}\mathbf{W}_l^{Q,A} \\ \mathbf{K}_l^A = \mathbf{H}^{l-1}\mathbf{W}_l^{K,A} \\ \mathbf{V}_l^A = \mathbf{H}^{l-1}\mathbf{W}_l^{V,A} \end{cases} \tag{4}$$

where $\mathbf{W}_l^{Q,A}, \mathbf{W}_l^{K,A}, \mathbf{W}_l^{V,A} \in \mathbb{R}^{d_E^G \times d^I}$ are learnable projection matrices, and d^I denotes the dimension after projection.

After that, the representation of the new feature $\mathbf{A}_l$ obtained by self-attention can be expressed as:

$$\mathbf{A}_l = \text{Attention}(\mathbf{Q}_l^A, \mathbf{K}_l^A, \mathbf{V}_l^A) \tag{5}$$

Cross-Attention on Visual&Language Features. We use cross-attention mechanisms to calculate the attention weights between the visual features and language features, where the key and value are calculated by the instruction embedding $\mathbf{X}$. For the layer l, its query, key and value can be calculated as:

$$\begin{cases} \mathbf{Q}_l^A = \mathbf{A}^l\mathbf{W}_l^{Q,C} \\ \mathbf{K}_l^A = E^x(\mathbf{X})\mathbf{W}_l^{K,C} \\ \mathbf{V}_l^A = E^x(\mathbf{X})\mathbf{W}_l^{V,C} \end{cases} \tag{6}$$

where $\mathbf{W}_l^{Q,C} \in \mathbb{R}^{d_E^G \times d^I}, \mathbf{W}_l^{K,C}, \mathbf{W}_l^{V,C} \in \mathbb{R}^{d_E^x \times d^I}$ are learnable projection matrices. After that, we calculate the output of the cross-attention in the layer l as:

$$\mathbf{C}^l = \text{Attention}(\mathbf{Q}_l^C, \mathbf{K}_l^C, \mathbf{V}_l^C) \tag{7}$$

Last, the model outputs the feature map of the current layer l by a feed-forward network:

$$\mathbf{H}^l = \text{FeedForward}_l(\mathbf{C}^l) \tag{8}$$

Furthermore, the residual connection followed by a layer normalization is also applied between each two sub-layers. The semantic map of time step t can be expressed as $\mathbf{S}_t^W = \mathbf{H}^L$, where $\mathbf{H}^L$ is the output of the last VLCT layer.

3.6 Visit Distribution Prediction

We predict the visit distributions $\mathbf{d}_t^p$ and $\mathbf{d}_t^g$ by the LingUNet [20], a language-conditioned image-to-image encoder-decoder model.

The inputs of LingUNet are the accumulated global feature map $\mathbf{G}^W$ and the semantic map $\mathbf{S}^W$, both having the same height and width. The model process the input feature $\mathbf{F}_0^D = \left[\mathbf{G}_t^W, \mathbf{S}_t^W\right]$ through N cascaded convolutional layers, generating a series of feature maps $\mathbf{F}_k^D = \text{CNN}_k(\mathbf{F}_{k-1}^D), k = 1 \ldots N$. After that, we filter each feature $\mathbf{F}_k^D$ with 1×1 convolution kernels with weight $\mathbf{W}_k^D$, which is calculated by the instruction feature $\mathbf{X}$ with a linear transformation. Then we obtain N intermediary features $\mathbf{G}_k^D = \mathbf{F}_k^D \circledast \mathbf{W}_k^D$. After that, the model performs N deconvolution operations to obtain N feature maps:

$$\mathbf{F}_k^G = \begin{cases} \text{DECONV}_k([\mathbf{F}_{k+1}^G, \mathbf{G}_k^D]), & 1 \leqq k < N-1 \\ \text{DECONV}_k(\mathbf{G}_k^D), & k = N \end{cases} \tag{9}$$

where $[\cdot,\cdot]$ denotes the concatenation of two vectors along the channel dimension [24]. The output of the LingUNet $\mathbf{F}_1^G$ consists of two channels representing the distributions $\mathbf{d}_t^p$ and $\mathbf{d}_t^g$ respectively.

3.7 Decision Making

At the Decision Making stage, the model predicts the action a_t and the stopping probability p_t^{stop}. We first align these two distributions with the egocentric frame of the agent through an affine transformation. Then we crop [3] the center of each distribution into a $K \times K$ region. After that, the cropped region will be flattened and concatenated to a vector x_t with size $2K^2$. We calculate a_t and p_t^{stop} through a feed-forward network $a_t, p_t^{stop} = ACT(x_t)$

The agent will take the action $a_t = (v_t, \omega_t)$ if the stopping probability p_t^{stop} is above a threshold κ_{stop}. Otherwise, the agent will execute the STOP command.

3.8 LLM-Based Instruction Rephraser for Out-Of-Distribution Challenges

Language models pre-trained on large-scale data and fine-tuned on specific domains perform very well on specific tasks [8].

Out-Of-Distribution Challenges. When we deploy the system in practical scenarios, users may provide various textual instructions even for the same navigation task. The instructions received by the model are often uncontrollable. Specifically, the input instruction to the model may potentially be Out-Of-Distribution (OOD) [19] samples, such as simple words or incomplete instructions.

LLM-Based Instruction Rephraser. To address this challenge, based on the model proposed in previous sections, we additionally introduce an LLM as an intermediate component of our model. This component is capable of rephrasing OOD instructions, thereby reducing the disparities between OOD and training instructions.

As an example, we consider a subpar scenario where the model receives the OOD instructions formatted as a sequence of triplets. This sequence is not a complete sentence, but retains the essential and minimal content of the full instruction. We express this OOD instruction sequence as $u' = \{r_1, \cdots, r_L\}$, where L is the number of triplets of the sequence, r_i is a triplet with structure:

$$r = < action, object, relative\ position > \tag{10}$$

each triplet r contains three elements, where the first element *action* denotes how the robot should move, the second element *object* indicates the agent on which objects or landmarks need to focus, and the last element *relative position* indicates the position related to the objects. Each element can be a word or a phrase. We also allow the default value of elements if the information of some elements is not necessary or does not exist.

Once the model is trained, we can integrate the LLM into the instruction encoding process. Then the embedding of rephrased instruction will be expressed as:

$$\mathbf{X}' = SentenceBERT(LLM(u')) \tag{11}$$

Fig. 4 shows the process of dealing with OOD instructions.

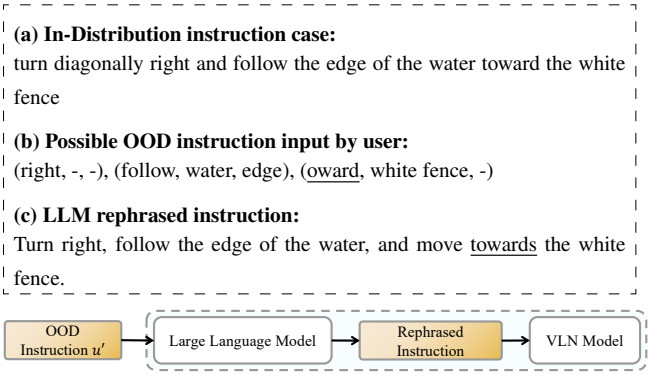


Figure 4 The process of dealing with OOD instructions. For demonstration, instruction (a) is selected from the training dataset to show an in-distribution case. The OOD instruction (b) has three triplets with some default elements and a misspelled word (oward -¿ toward). The instruction (c) is an intermediate output of our approach by rephrasing the OOD instruction (b). All of them have the same navigation target.

4 Experiments

4.1 Training

We train our network with imitation learning and reinforcement learning. We generate a sequence of agent state $S = \{s_1, \ldots, s_T\}$ by executing an oracle policy π^*, which can follow the expert trajectory by a simple control rule. Then we generate the ground truth trajectory visit distribution $\mathbf{d}_*^p$ by assigning high probability for positions around the expert trajectory, and assign high probability around the target position to get ground truth target visit distribution $\mathbf{d}_*^g$. Specifically, for each position $\widetilde{p} = (x, y)$ in the semantic map, we calculate visiting and stopping probability as:

$$\mathbf{d}_*^p(\widetilde{p}) = \frac{1}{Z_p} \sum_{\widetilde{p}_j \in \mathbf{S}^W} g(\widetilde{p} \mid \widetilde{p}_j, \sigma), \quad \mathbf{d}_*^g(\widetilde{p}) = \frac{1}{Z_g} g(\widetilde{p} \mid \widetilde{p}_T, \sigma) \tag{12}$$

where Z_p and Z_g denote the normalization terms, and $g(\cdot|\mu, \sigma)$ denotes the Gaussian probability density function with mean μ and variance σ^2. We minimize the KL divergence between the ground truth and predicted distributions: $D_{KL}(\mathbf{d}_*^p \parallel \mathbf{d}^p) + D_{KL}(\mathbf{d}_*^g \parallel \mathbf{d}^g)$.

Furthermore, we apply a set of auxiliary losses from Blukis et al. [2] to aid in more effective feature learning and accelerating convergence, including (a) object recognition loss $J_{percept}$ to classify the visible objects using their relevant positions in the semantic map; (b) the grounding loss J_{ground} to classify the visible objects mentioned in the instruction; (c) the language loss J_{lang} to classify the objects mentioned in the instruction. So the complete loss function can be expressed as:

$$J(\theta) = D_{KL}(\mathbf{d}_*^p \parallel \mathbf{d}^p) + D_{KL}(\mathbf{d}_*^g \parallel \mathbf{d}^g) + \alpha J_{percept} + \beta J_{ground} + \gamma J_{lang} \tag{13}$$

where α, β, γ represent the hyper-parameters weighting the contribution of each auxiliary loss.

For reinforcement learning, we use DAggerFM [2], a low-memory variant of DAgger [25], which improves the policy by combining oracle policy trajectories with those trajectories generated by the model during training. The model begins training with supervised learning on a dataset $\mathcal{D}^*$ collected by the oracle policy π^*. After that, We sample the initial dataset $\mathcal{D}$ from $\mathcal{D}^*$ and then perform K iterations of training. For each iteration k, we discard N_d trajectories from $\mathcal{D}$ and replace them with N_d new trajectories collected by the mixture policy of oracle policy π^* and current policy π_i.

4.2 Environmental Setup

For experiments, we use a quadcopter simulator developed by Blukis et al. [2] based on Unreal Engine 4 and AirSim plugin [27].

Dataset. We evaluate our approach on the LANI corpus [20], which

Table 1 Test Results and Ablation Studies.

Method	test set		
	SR↑	AD↓	MD↓
ORACLE	100	1.38	1.29
CHAPLOT	21.34	11.2	9.35
GSMN	24.36	9.94	8.28
PVN	41.21	8.68	6.26
OA w/o VLCT	41.49	8.37	6.27
OA w/o SBERT	35.69	9.14	7.22
OA	**42.60**	**8.30**	**6.17**

contains 27,965 crowd-sourced instructions for navigation in an open environment. Each piece of data includes an instruction, a demonstration trajectory, and an environment defining a $50m \times 50m$ grassland with randomly placed 6-13 landmarks. The dataset is split into train, test, and development sets with 19,758/4,135/4,072 samples. To evaluate the performance of LLMIR, we create an OOD test set by rephrasing the instructions of the original test set. The instruction in the OOD set is formatted as a sequence of triplets, which retain the essential and minimal content of full instructions. The specific definition of instructions can be found in Sec. 3. 8.

Evaluation Metrics. We assess the task completion based on the stopping distance between the stop and the target location: $\| p_g - p_T \|$, where p_g is the end-point of the human-annotated trajectory, and p_T is the position where the agent outputs the STOP action. If the stopping distance is less than the threshold $\kappa_d = 5m$, the task is considered successfully completed. To assess the approximate distribution of the stopping distances, we also compare the average values (AD) and median values (MD) of stopping distances.

Baselines. We compare our method against PVN [3], CHAPLOT [5] and GSMN [2]. CHAPLOT uses a Gated-Attention mechanism for multimodal fusion of verbal and visual modalities. GSMN constructs a semantic map and infers the target location using simple language-guided convolutional filters. PVN predicts interpretable position-visit distributions indicating the agent where to go and where to stop. We present the performance of ORACLE as an upper bound. For rephrasing the instructions, we introduce gpt-3.5-turbo [4] as our instruction rephraser and evaluate it on the OOD test set.

Hyper-Parameters. All models are trained on the A6000. We configure the hyper-parameters based on experience and relevant work [2, 3]. We use the AdamW [18] optimizer with the learning rate set to 10^{-3} and the weight decay set to 5×10^{-5}. And we set the threshold of the stopping probability κ_{stop} as 0.07. The model computes the visit distributions every $T_d = 6$ time steps. We conduct experiments to determine the optimal hyper-parameters for our VLCT model. We vary the number of attention heads (1-4) and layers (1-6). The model achieves the best performance when we set head to 2 and layer to 4. We also explore dropout rates P_{drop} from 0 to 0.5, with the best performance observed at a rate of 0.1.

Table 2 OOD Evaluation of LLMIR.

Insturcion Type	Rephrased by LLM	OA w/o VLCT			OA w/o SBERT			OA		
		SR↑	AD↓	MD↓	SR↑	AD↓	MD↓	SR↑	AD↓	MD↓
Original		41.49	8.37	6.27	35.69	9.14	7.22	42.60	8.30	6.17
OOD		37.04	9.22	6.93	27.63	10.2	8.53	37.69	9.25	6.81
	✓(Our)	**38.59**	**8.65**	**6.54**	**30.00**	**9.68**	**8.11**	**38.95**	**8.80**	**6.61**

Figure 5 Top-down view trajectories of executing test set instructions. Compare our model's (blue) trajectories with PVN (yellow) and Oracle (red).

4.3 Results

Table 1 shows our main results and ablation on the test set.

Main Results. Compared to PVN [3], our approach demonstrates the improvement in the absolute task completion rate by 1.39%, a relative improvement of 4.30% in average stopping distance, and 1.44% improvement in median stopping distance. This indicates our approach is more efficient in navigation.

Ablation Studies. Our ablations experiments demonstrate the performance improvement brought by the approaches we proposed for the model. After removing the VLCT or SBERT, both of them show lower performance in the three evaluation metrics of the task, which verifies their contributions.

Evaluation of LLMIR. Table 2 shows the improvement achieved by the intermediary component LLMIR (gpt-3.5-turbo) of our model. On the OOD test set, our full model improves the absolute task-completion of 1.51% compared to the version without LLMIR, and gets 6.18% improvement on average stopping distance, 5.63% improvement on median stopping distance. The approach of LLMIR narrows the performance gap between new test set and original test set by 34.83%. To demonstrate the effectiveness, we also evaluate LLMIR on other ablation models. The results show that the model integrating with LLMIR has greatly improved the ability to process OOD instructions. In addition, we find that the model removing SBERT (OA w/o SBERT) shows destructive drops on the OOD set. After using LLMIR, the performance improved a lot, which demonstrates the generalization ability of SBERT for various types of inputs and the effectiveness of LLMIR.

To assess the effectiveness of our model, we show three examples of navigation tasks (Fig. 5). When the agent is deployed in a perspective that cannot obtain any information from the environment (left), our model adapted to this scenario and successfully executed the task (blue), whereas PVN (yellow) quickly declares failure

in this situation. The other two examples highlight the challenges posed by complex instructions and long-distance navigation. When navigation instructions provide only concise destination information (mid), PVN can't handle scenario information that is not described in the instructions. In contrast, our model demonstrates the ability to more efficiently learn relationships between long-distant visual features and shows a strong capacity to represent complex instructions and generate accurate execution strategies.

5 Conclusion

In this paper, we propose a novel conditional Transformer (VLCT) regarding the instruction embeddings as a condition for visual features to facilitate the fusion of vision and language features. We adopt a pre-trained language model SentenceBERT (SBERT) to enhance the model's representation ability. We design a prompt-based intermediary component, LLM-based Instruction Rephraser (LLMIR), to address the issue of OOD instructions from practical scenarios. Sufficient experiments have demonstrated the effectiveness of the proposed model.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP23H03402.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2018.

[2] Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A. Knepper, and Yoav Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Proceedings of the Robotics Science and Systems Conference (RSS)*, 2018.

[3] Valts Blukis, Dipendra Misra, Ross A. Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 505–518. PMLR, 29–31 Oct 2018.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[6] Howard Chen, Alane Suhr, Dipendra Kumar Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[7] Zhe Chen, Jiyi Li, Fumiyo Fukumoto, Peng Liu, and Yoshimi Suzuki. Vision-language navigation for quadcopters with conditional transformer and prompt-based text rephraser. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, MMAsia '23, New York, NY, USA, 2024. Association for Computing Machinery.

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

[9] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1634–1643, October 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[13] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln-bert: A recurrent vision-and-language bert for navigation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

[14] Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, page 13238–13246. IEEE Press, 2021.

[15] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872, Florence, Italy, July 2019. Association for Computational Linguistics.

[16] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing.

[17] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, Online, November 2020. Association for Computational Linguistics.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[19] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-

distribution detection for vision-language models? *International Journal of Computer Vision*, 2023.

[20] Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. Mapping instructions to actions in 3D environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2667–2678, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[21] OpenAI. Gpt-4 technical report, 2023.

[22] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015.

[25] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[26] Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. 2023.

[27] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[30] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5412–5425, 2020.

[31] An Yan, Xin Wang, Jiangtao Feng, Lei Li, and William Yang Wang. Cross-lingual vision-language navigation. In *International Conference on Learning Representations*, 2019.

[32] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4, 2023.

[33] Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221, Online, April 2021. Association for Computational Linguistics.

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

生成系マルチメディア

座長:井手 一郎(名古屋大学)

コメンテータ:野宮 浩揮(京都工芸繊維大学)

2024年2月28日(水) 16:00 〜 18:10 T4-A (オンライン（Zoom Events))

17:45 〜 18:10

[T4-A-3-05]【技術報告】ヒトと AIの協調によるデザイン創出支援

山口喬弘 (株式会社豊田中央研究所)

©一般社団法人日本データベース学会

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

2024年2月28日(水) 16:00 ~ 18:10 | T4-B オンライン（Zoom Events)

可視化によるユーザ支援

座長:北山 大輔(工学院大学)
コメンテータ:荘司 慶行(静岡大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T4-B-3-01] レビュー動画内の商品特徴量分析システムにおける情報提示方法の検討

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

16:20 ~ 16:35
[T4-B-3-02] 日中同形異義語の意味分析と可視化：言語横断的意味研究の一環として

*韓 禾[1]、伊藤 貴之[1] (1. お茶の水女子大学　人間文化創成研究科　情報コース　伊藤研究室)

16:35 ~ 17:00
[T4-B-3-03] 散布図選択による多次元データ可視化における配置の最適化

*松尾 深景[1]、伊藤 貴之[1] (1. お茶の水女子大学)

17:00 ~ 17:25
[T4-B-3-04] 粒度の異なる商品比較を実現する評判情報可視化インタフェースの提案

*板垣 紫音[1]、上野 史[1]、太田 学[1] (1. 岡山大学大学院　環境生命自然科学研究科　太田研究室)

17:25 ~ 17:50
[T4-B-3-05] Webブラウザのタブ群のタスクごとへのグループ化手法

*久冨 瑞稀[1]、田島 敬史[1] (1. 京都大学情報学研究科)

レビュー動画内の商品特徴量分析システムにおける情報提示方法の検討

小林　愛空†　　山口　史弥††　　上田真由美†††,††††　　中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
††† 流通科学大学 経済学部　〒 651–2188 兵庫県神戸市西区学園西町 3-1
†††† 大阪大学 サイバーメディアセンター　〒 560–0043 大阪府豊中市待兼山町 1 番 32 号
E-mail: †{g2153551,nakajima}@cc.kyoto-su.ac.jp, ††i2286211@cc.kyoto-su.ac.jp,
†††,††††Mayumi_Ueda@red.umds.ac.jp

あらまし　オンラインショッピングにおける商品情報収集のためのレビュー動画視聴の効率化を目指して，我々はレビュー動画内における商品特徴量分析手法の開発を行ってきた．この研究で開発されたシステムは，レビュー動画とコメントから商品の特徴について言及している箇所とその割合を分析し，動画の視聴時に言及箇所を，視聴前に言及割合をユーザに提示している．しかし，現状では割合以外に言及時間や言及内容の詳しさなどの提示できていない情報がある．本稿では，この問題の解決を目的として，ユーザに提示する情報の検討とその提示方法の提案を行う．

キーワード　レビュー動画，情報抽出，自動スコアリング，コメント分析，YouTube

1　はじめに

近年，Amazon[1]や楽天市場[2]を中心としたオンラインショッピングの利用者が増えている．(2020 年 3 月以降 2 人以上世帯の 50%が利用 [1]) しかし，オンラインサイトは実店舗と異なり，商品を手に取って確認することができない（図 1 参照）．オンラインショッピングでは，商品の説明画像，商品を購入したユーザが投稿するレビューを参考にすることが一般的であるが，商品画像やレビュー情報だけでは商品情報が伝わり難い．視覚的な大きさの比較ができない情報等を参考にして商品を購入した場合，届いた商品が期待していたものと異なるリスクがある．そこで最近は，レビュー動画を参考にするユーザが増えている．従来のレビューよりも使用感をより的確に把握できるためである．レビュー動画からは，実際に商品が使用されている様子を観ることができ，商品画像や文字のみでは伝わらない情報が得られる．ただし，図 2 に示すように動画の内容を事前に把握することは困難で，目的の動画のみを効率的に視聴することは難しい．

この考えに基づき，我々は，レビュー動画分析システム [2] の開発に取り組んだ．オンラインショッピング時には，該当商品を購入するための判断基準となる複数の特徴の項目（以下,「特徴項目」と呼ぶ）のうち，ユーザが関心を持つ特徴項目に対する言及の有無が重要である．本システムは，各動画が商品のどの特徴項目について言及されているかを，動画を再生する前にユーザに提示することができ，効率的に動画を視聴することができる．

しかしながら，ここで開発されたレビュー動画システムについては，システムのユーザインタフェース（UI）に関する議論が十分になされていない．分析結果をどのように提示することが適切であるのか，限られたスペース上に提示する情報量とその方法について検討する．実験を行いながら本検討を行うことで，より効率的なシステムを実現し，本システムの利用を促進する．本稿では，先行研究の UI に着目しつつ，その問題点を取り上げ，これを解決するようユーザに提示する情報の検討とその提示方法の提案を行う．

本稿の構成は以下の通りである．2 節ではレビュー動画分析システムの関連研究と UI についての関連研究を紹介する．3 節ではシステムの分析手法，分析結果の提示方法の問題点を述べる．そして，問題点の解決と提示内容の検討のための，情報提示方法を述べる．4 節では 3 節で提案した提示方法の評価実験を述べ，最後に 5 章で本稿のまとめを述べる．

図 1　オンラインショッピングの問題

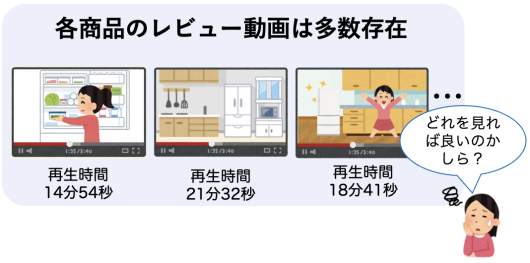


図 2　レビュー動画閲覧時における問題

1：Amazon, https://www.amazon.co.jp/
2：楽天市場, https://www.rakuten.co.jp/

2 関連研究

2.1 商品レビュー動画の分析および商品スコアリングシステムに関する関連研究

以下に，商品レビュー動画の分析および商品スコアリングシステムに関する関連研究をあげ，本研究との差異を示す．

動画の内容を事前に把握することは困難で，目的の動画のみを効率的に視聴することは難しい．そこで，動画に付与されたコメントを基に，ユーザの直感に基づく動画検索を可能にする先行研究が行われている．鴇田ら [3] は，YouTube の動画に対して，動画のタイトルや説明文などに含まれないユーザの直感に基づく動画検索のための評価項目別自動スコアリングを提案している．具体的には「じわる」「萌える」「エモい」などの動画の検索の実現のために，評価項目別に動画に付与されるコメントを分析し，動画を再生する前に真に求めている動画か否かを判断する手法を提案している．この研究では，コメント分析に基づき直感に基づく動画検索を目的としているが，本研究では分析対象に字幕を加え，さらにレビュー動画に特化して分析を行う．

また，コメント文の極性に基づき文章の分類を行う先行研究が行われている．Muhammad ら [4] は，動画に付与されたコメントから肯定的な意見と否定的な意見を分類する取り組みを行っている．Pradhan [5] らは，同様に自然言語処理を用いたコメント分類によって，当該動画の視聴者による評判を分かりやすくする取り組みを行っている．

Siersdorfer ら [6] は，機械学習手法を用いた YouTube の動画等に対するコメント分析手法を提案している．包括的な意見の極性を抽出したり，将来高い評価を受ける可能性のあるコメントの予測，コメントからコミュニティに悪影響を及ぼす可能性のあるユーザーの分析を提案している．これらの研究では肯定的な意見と否定的な意見の分類や極性分析を目的としているが，本研究では特徴量に基づく分析を行っている．

また，オンラインショッピングにおいてレビューや口コミは商品購入の重要な判断材料となっている．しかし，全てのレビューに目を通すことは非常に手間がかかる．そこで，内容や投稿者などの情報を用いて，自動で評価を算出する研究が行われている．Haque ら [7] は，Amazon データセットを用いて教示あり学習手法によるレビューの極性化に取り組んでいる．

上述のように商品レビュー分析により，商品の評判情報の可視化を可能とする取り組みもある．Basani ら [8] は，商品レビューをポジティブ・ネガティブ．ニュートラルに分類し，商品レビューの要約を行っている．Zhang ら [9] は，商品レビューにおける極性分析と意見抽出において，テキストの主観性を考慮した検討を行っている．これらの研究は，商品レビューにおける極性の分類やレビューの要約を目的としているが，本研究ではレビュー動画を対象とした特徴量分析を行っている．

加えて，オンラインショッピングでは，ユーザの目的に沿った機能を持つ製品を探すには非常に時間がかかる．これを解決すべく，Scaffidi ら [10] は商品の特徴をレビューから抽出しスコアをつけることによって，オンラインショッピングを利用するユーザの要求を満たす製品の自動検索システムを提案している．この研究では，製品の自動検索を目的としているが，本研究では商品の実際のイメージや雰囲気といった感覚的な情報を把握できるようなレビュー動画の推薦を目的としている..

以上より，先行研究において，動画の内容やレビューの要約を目的とした取り組みは盛んに行われている．しかしながら，レビュー動画そのものの商品特徴量分析を行う取り組みは十分に行われているとはいえない．このような背景から，我々はオンラインショッピング支援を目的としたレビュー動画の商品特徴量分析を行った．

2.2 UI についての関連研究

鈴木ら [12] は，データの可視化において，ユーザが見たい階層を選択することができる機能を加えた階層型ドーナツグラフを例に挙げて，インタラクティブなデータ可視化は，変化（推移) が把握しやすくなくなること等の便益をもたらすとしている．また，円グラフを例に挙げて，馴染み深いグラフには新たに読み取り方を学習する負担が少ないという性質があると述べている．

隅谷ら [13] は，円グラフと帯グラフの読み取り精度や大小関係の判定精度について 800 人を対象とした実験を行っている．その結果，読み取り精度については円グラフの方がやや高く，その処理時間については帯グラフの方が短いことが明らかになっている．また，大小関係の判定精度については，帯グラフの方が処理時間が短いことが明らかになっている．

矢部ら [11] は，購入済みコスメアイテムと未使用の購入を検討しているアイテムの相違点を可視化するシステムの開発において，これらのアイテムについてのスコアリング結果を表ではなく横棒グラフを使用し，グラフにつける色に意味を持たせることで，一目でこれらのアイテムの評価項目別の相違点をわかるような手法を提案している．

以上の先行研究より，本研究ではグラフが持つ性質を考慮した上でわかりやすい情報提示方法の検討を行う．

3 レビュー動画内の商品特徴に関する言及量分析結果の提示方法の検討

3.1 現在のレビュー動画分析システム

レビュー動画分析システムでは，検索結果に表示されたレビュー動画から字幕とコメントを取得し，これらについて，特徴項目ごとに用意したキーワードの出現に基づき，動画内の特徴項目について言及している割合 (以下,「言及割合」と呼ぶ) を求める (図 3 参照)．用意したキーワードには特徴項目との関連性に基づいた重要度を示す点数が設定されており，商品カテゴリごとに用意した各特徴項目について点数の合計値を算出する．そして，全特徴項目の点数の合計のうち，その特徴項目の点数が占める割合を計算することによって言及割合は求められる．

3.2 情報提示方法の問題点

レビュー動画分析システムでは現在，図 4 のように，動画の

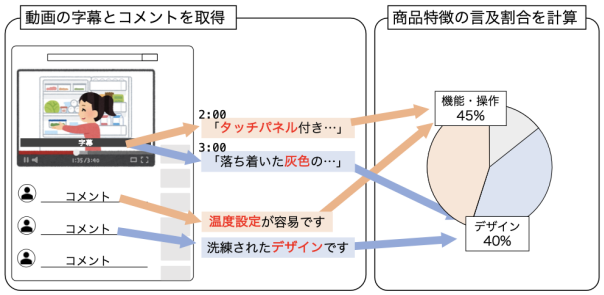


図 3　レビュー動画分析システムにおける特徴項目に関連する単語の抽出イメージ

図 4　レビュー動画分析システムの実際の画面

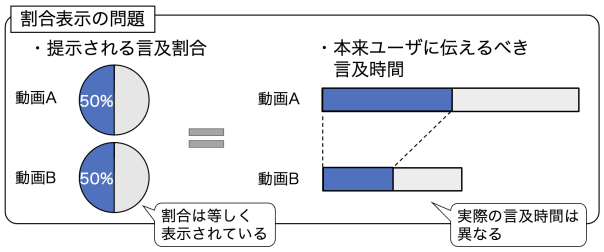


図 5　言及割合のみを提示することの問題

視聴前に言及割合が提示されているようになっているが，特徴項目について言及している時間 (以下,「言及時間」と呼ぶ) や，その内容，密度などのユーザに伝えられていない情報がある．言及割合のみを提示していると，同じ割合で表示されているにもかかわらず，実際の言及時間は異なるということが起きてしまい，ユーザは特徴項目についての言及が多い動画と少ない動画を区別することができないと考えられる (図 5 参照)．そこで本研究では，言及時間に着目して，これをユーザにわかりやすく提示する方法を検討する． まず，動画内の特徴項目について言及している時間を言及割合の代わりに提示するべきなのか，両方を提示する方法を新たに考えるべきなのか，など検討すべき項目がある．そのために，言及割合の提示と言及時間の提示における利点と欠点を挙げることにする．

言及割合は合計で 100%となるため，計算した結果がどのような値であっても，円グラフを用いて全ての動画に対して共通の形で提示することができる．そして，特徴項目の個数にもよるが，円グラフは図 4 から分かるように，面積によって値の違いを表すことができる．つまり，言及割合を提示することには，その動画が何に重点を置いているか一目でわかるように表現できるという利点がある．そして，3.2 で述べたように，言及時間の違いを表現できないという欠点がある．対して言及時間を提示することには，動画が商品のどの特徴項目について多く述べられているかを示しつつ，言及割合のみを提示する場合での欠点を補うことができるという利点がある．しかし，合計や各特徴項目についての値の最大値は一定ではないため，言及時間を円グラフに提示すること，各動画について等間隔のグラフで提示することは難しい．つまり，その動画が何に重点を置いているか一目でわかるように表現することが難しいという欠点がある．したがって，これらの提示方法には互いに補い合うような利点・欠点があり，言及割合，言及時間の両方を提示する必要があると考える．

3.3　レビュー動画分析システムにおける情報提示方式の提案

3.3.1　言及時間の分析

これまでの研究で，特徴項目に関連するキーワードを取得できているため，予め設定した特徴項目に関連がない事柄について言及している時間を算出することが可能である．

3.3.2　言及割合・言及時間の提示方法

割合を提示できる方法の中で馴染み深いものには円グラフや帯グラフ，棒グラフなどがある．どのグラフを用いるのが適切であるかを調べ，そのグラフと共に言及時間を提示した場合と比較するため，言及割合を円グラフ, 帯グラフで提示する場合とそれぞれについて言及時間を提示する場合，そして棒グラフで言及時間を提示する場合と，リストで言及時間を提示する場合の提示方法を提案する．

a)　円 グ ラ フ

言及割合を円グラフで提示する場合のイメージを図 6 に，言及割合・言及時間を円グラフに提示する場合のイメージを図 7 に示す．図 7 について，異なる情報を提示する場所は分けるべきだと考えて，言及時間を円グラフの外に載せている．

b)　帯 グ ラ フ

言及割合を帯グラフで提示する場合のイメージを図 8 に，言及割合・言及時間を円グラフに提示する場合のイメージを図 9 に示す．円グラフと同様に言及時間は帯グラフの外に載せている．

c)　棒 グ ラ フ

棒グラフを割合の提示に用いることは一般的ではないと考えられるため，言及時間のみを図 10 のように棒グラフに提示する．ただし，動画によって総時間が大きく異なる場合については，総時間が長い動画の影響で，短い動画のグラフが小さくなりすぎる問題があるため，一定の範囲を超える言及時間については特殊な提示をする方法をとることにする．

d)　リ　ス　ト

グラフに提示しない場合についても比較するため，言及割合をリストで提示する場合を図 11 のように，言及割合・言及時間をリストで提示する場合を図 12 のようにする．

4　評価実験計画

本節では，3.3.2 節で提案した情報提示方法の評価実験計画について述べる．

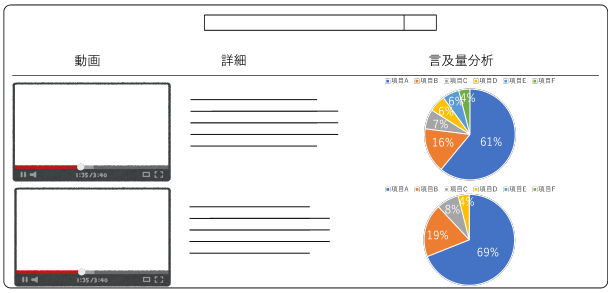


図 6　言及割合を円グラフに提示するイメージ

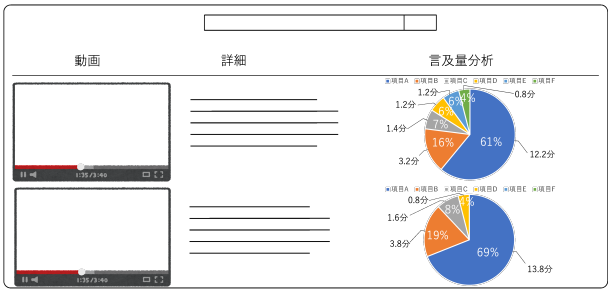


図 7　言及割合・言及時間を円グラフに提示するイメージ

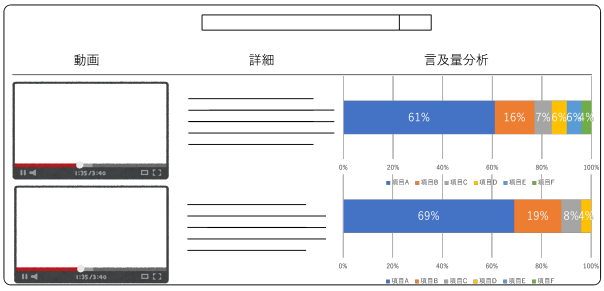


図 8　言及割合を帯グラフに提示するイメージ

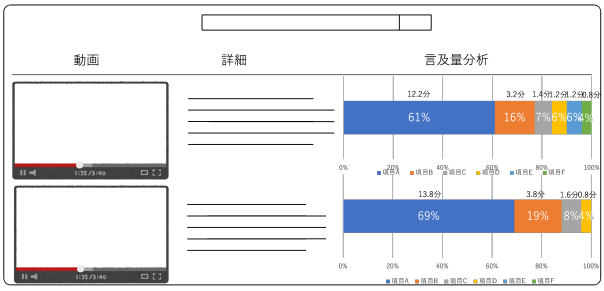


図 9　言及割合・言及時間を帯グラフに提示するイメージ

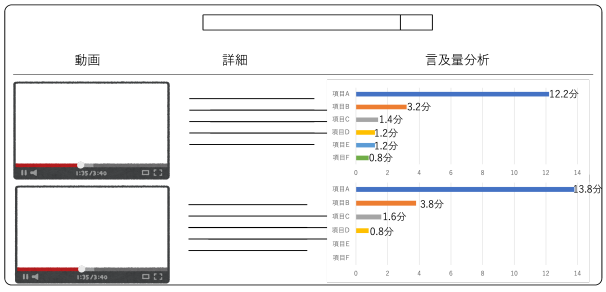


図 10　言及時間を棒グラフに提示するイメージ

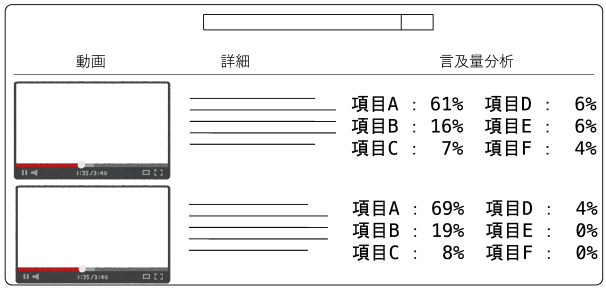


図 11　言及割合をリストに提示するイメージ

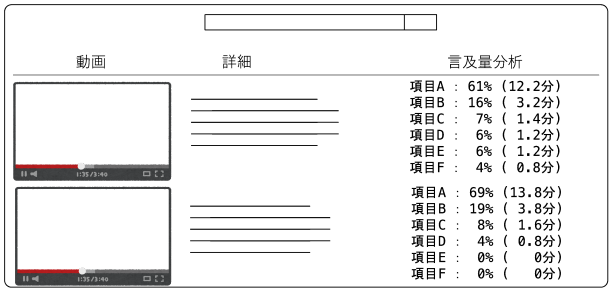


図 12　言及割合・言及時間をリストに提示するイメージ

4.1　概　　要

実施予定の評価実験は，レビュー動画分析システムにおける分析結果の提示方法の比較評価である．3.3.2 節で説明した提示方式について，ユーザビリティ評価・情報収集に係る時間・情報収集に係るストレス尺度に基づく比較を行う．それぞれの評価方法については 4.2 節で述べる．

これらを提案した提示方法全ての結果で比較し，ユーザビリティの観点で提案手法における情報提示方法の変更の有効性を評価する．

本比較評価にあたっては，ユーザが重要視している点を考慮した評価を行う予定である．ユーザごとに，重視する内容（レビュー動画内の言及割合・言及時間・まとまり具合・質など）は異なる．重視する内容が異なる場合，求められる UI も異なると考えられるため，4.2 節で説明する評価方法に加えて，ユーザが重視する内容についても同時に収集する．

4.2　レビュー動画分析システムの評価方法

4.2.1　ユーザが重視する内容の調査

ユーザが重視する内容の調査では，言及割合・言及時間の選択肢からオンラインショッピングのためにレビュー動画を視聴する場合に知りたいと思う方を答えてもらい，その後に割合・時間の両方，または片方だけを知りたいかを答えてもらった．また，図 6・8・11 に示した言及割合を円グラフ・帯グラフ・リストに提示するイメージ同士で比較，図 7・9・10・12 に示した円グラフ・帯グラフ・棒グラフ・リストに提示するイメージ同士で比較を行ってもらい，動画内容の内訳が分かりやすいと思う順番にイメージを並べてもらった．そして円グラフ・帯グラフ・リストについて言及割合のみを提示するイメージと言及割合・言及時間の両方を提示するイメージの比較を行って，使いたいと思うものを選んでもらった．

調査の結果，実験参加者 20 代 13 名の内，61.5%が言及割合を重視しており，76.9%が言及割合・言及時間の両方を知りた

いと思っていることがわかった．提示する内容を固定した異なるグラフ同士での比較について，分かりやすいと思う順番から算出した平均順位が最高だったのは，言及割合のみを提示する場合では円グラフ，言及割合・言及時間の両方を提示する場合では帯グラフとなった．提示する内容は異なった同じグラフ同士での比較について，両方を提示する場合のイメージを使いたいと答えた参加者は円グラフで 76.9%，帯グラフで 92.3%，リストで 92.3%だった．したがって，言及割合に加えて言及時間を提示することは，ユーザの要求に沿っており，その方法として帯グラフを用いることが適していると考えられる．

4.2.2 ユーザビリティ評価

システムユーザビリティ評価実験では，提案システムの有効性を System Usability Scale(SUS) [15] を用いて評価する．

SUS とは，John Brooke が開発したシステムユーザビリティの数値的な評価が可能な尺度である．質問項目は肯定的な質問と否定的な質問を交互に並べた 10 個の質問で構成され,「全く同意できない」の 1 から「非常に同意できる」の 5 段階で回答してもらう．回答の結果は，100 点満点で算出され，システムのユーザビリティを評価することが可能である．

4.2.3 情報収集に係る時間

被験者に実際のレビュー動画を視聴してもらうことで，レビュー動画内から指定された情報を発見できるまでの時間を用いて評価する．

まず，被験者に対し商品に関する特定の情報についてレビュー動画から情報を発見するように指示する．指定された情報を発見できるまでの時間を計測することで，提示方式別の発見時間を算出できる．発見時間が短くなるほど，その提示方式は有効であると言うことができる．

4.2.4 情報収集に係るストレス尺度

システムのストレス評価実験では， Stress Response Scale-18(SRS-18) を用いる．SRS-18 は鈴木らによって開発された心理的ストレス反応尺度である [16]．簡便で，かつ日常で多く経験される心理的ストレス反応の測定が可能である．レビュー動画分析システムを使用する作業を行った前後で受けるストレスの量が小さいか負であるほど，その提示方式が有効であると言うことができる．

4.3 実験結果の算出方法と考察・今後について

4.2.1 節の結果から，言及割合に加えて言及時間を提示することはユーザの要求に沿っており，その方法として帯グラフを用いることが適していると考察した．ただ，帯グラフには，グラフ上に提示する言及割合・言及時間の値が重なってしまう問題があるため，4.2 節の実験を行う際には，この問題について対処する必要があると考える．なお，4.2 節の評価については，被験者全員の平均値を算出して，各提示方式のスコアとする．

また，本評価実験ののちに，被験者に対してヒアリング調査を設けることで，本稿で提案した提示方式における更なる課題や問題点を抽出する．

今後は本実験結果に基づき，レビュー動画分析システムの UI の更なる改善，レビュー動画の視聴効率化および商品情報の効率的な収集のために必要とされる新たな分析項目を決定する．

5 おわりに

本稿では，レビュー動画の商品特徴に関する言及量分析手法における情報提示方法の検討を行った．レビュー動画の商品特徴量分析システムは，閲覧すべきレビュー動画の判定を容易にする．より良い UI はシステムの利用を促進し，レビュー動画分析システム自体の意義を高める．今後は，検討を行った UI を実装した上で，実際のシステムを利用した評価実験を行い，得られた結果からさらなるシステムの改善を行う．

謝　辞

本研究の一部は，科研費 (課題番号：19K12243，20H04293，22K12281) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　献

[1] デジタルで支える暮らしと経済，令和 3 年　情報通信白書，総務省
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/html/nd121310.html

[2] 山口 史弥, 上田 真由美, 中島 伸介, レビュー動画内における商品特徴量分析手法の開発, DEIM2023 1b-1-3, 2023.

[3] 鴇田 和士, 上田 真由美, 中島 伸介, 動画に対する評価項目別スコアを用いた直感的動画検索システム, DEIM2022 C21-5, 2022.

[4] Abbi Nizar Muhammad et al., “Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier,” International Conference on Computer Science, Information Technology, and Electrical Engineering(ICOMITEE), 1570571406, 2019.

[5] Rahul Pradhan, “Extracting Sentiments from YouTube Comments,” 2021 Sixth International Conference on Image Information Processing (ICIIP), pp. 142-145, 2021.

[6] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl, “Analyzing and mining comments and comment ratings on the social Web,” ACM Transactions on the Web, Volume 8 Issue 3 Article No. 17, pp. 1-39, 2014.

[7] Tanjim Ul Haque et al., “Sentiment Analysis on Large Scale Amazon Product Reviews,” IEEE International Conference on Innovative Research and Development, 2018.

[8] Yuniarta Basani et al., “Application of Sentiment Analysis on Product Review E-Commerce,” International Conference on Advance and Scientific Innovation (ICASI), 1175 (2019) 012103, 2019.

[9] Zhu Zhang, Balaji Varadarajan, “Utility Scoring of Product Reviews,” Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM ’06), pp. 51–57, 2006.

[10] Christopher Scaffidi et al., “Red Opal: Product-Feature Scoring from Reviews,” Proceedings of the 8th ACM conference on Electronic commerce (EC ’07), pp. 182–191, 2007.

[11] 矢部 沙也加, 上田真由美, 中島伸介, 評価項目別スコアを用いたコスメアイテム間の相違点可視化システム, DEIM2021, C14-2, 2021.

[12] 鈴木 雅彦，鈴村 嘉右. データ可視化の必要性と意義 ーデータ

ビジュアライゼーションとはー, 情報の科学と技術 65 巻 11 号, p.470-475, 2015.
[13] 隅谷 孝洋, 坂下 潤一郎. 円グラフと帯グラフの量的把握の比較検討, 情報教育シンポジウム論文集, p.183-189, 2023.
[14] YouTube Data API https://developers.google.com/youtube/v3/getting-started
[15] John Brooke, “SUS - A Quick and Dirty Usability Scale”, Taylor & Francis, Usability evaluation in industry, pp.189-194, 1996.
[16] 鈴木 伸一, 嶋田 洋徳, 三浦 正江, 片柳 弘司, 右馬埜 力也, 坂野 雄二, 新しい心理的ストレス反応尺度 (SRS-18) の開発と信頼性・妥当性の検討, 行動医学研究 4(1), 22-29, 1997.

日中同形異義語の意味分析と可視化：
言語横断的意味研究の一環として

HAN HE † 伊藤 貴之‡

†お茶の水女子大学 〒112-0012 東京都文京区大塚 2-1-1
‡お茶の水女子大学 〒112-0012 東京都文京区大塚 2-1-1
E-mail: †han.he@is.ocha.ac.jp, ‡itot@is.ocha.ac.jp

あらまし 中国語と日本語は多くの共通漢字を持ち，同じ漢字を使用した熟語も存在する．例えば，「検討」，「質問」，「連想」などがある．しかし，これらの同形異義語は，言語間で異なる意味を持つことがある．本研究では，中日同形異義語の意味の違いを可視化し，そのパターンを探求する．本研究は，同形異義語の意味の列挙，SimCSE モデルによる文埋め込みの生成，そして PCA と K-Means クラスタリングによる文埋め込みの可視化の三段階で構成される．特に，文埋め込みにもとづくクラスタリングにより，中日同形異義語の意味の違いが明確に示された．何種類かのクラスタを観察することで，言語間の意味の違いを示す興味深い結果が得られた．本報告では，これらのクラスタから得られた結果と具体的な事例を紹介する．

キーワード 日中同形異義語，文埋め込み，次元削減と可視化，多言語分析

1. 序 論

中国語と日本語は多くの漢字を共有し，同じ漢字の熟語も多い．例えば，「検討」「質問」「連想」などがある．だが，同じ漢字の熟語でも，中国語と日本語で意味が同じとは限らない．このような単語は同形異義語と呼ばれる．本研究の目的は，中国語と日本語の同形異義語の意味の違いにどのようなパターンがあるかを観察することであり，そのために可視化を適用する．

本研究は 3 つのステップで構成される．まず，中国語と日本語の同形異義語とその意味をリストアップする．次に，SimCSE モデルを使って，これらの意味の文埋め込みを作る．そして，これらの文埋め込みを次元削減し，可視化する．この可視化を通じて，中国語と日本語の同形異義語の意味の違いにどのようなパターンがあるかを理解しやすくなる．さらに，可視化した結果をもとに，クラスタ分析を行い，その結果を分析する．本報告では，可視化とクラスタ分析から得られた面白い事例を紹介する．

2. 先行研究

2.1 中国語と日本語の比較の研究

中国語と日本語の比較研究は多数存在し，その範囲は文学から文法，漢字文化の違いにまで及んでいる．たとえば，中国語と日本語の小説の比較からみると，物語の構成や描写方法，感情の表現などに違いがみられる．これは，文化的な背景や文学の伝統にもとづいている．

翻訳文法の比較[1]も，言語間の違いを理解するための重要な視点を提供する．日本語の文法と中国語の文法は異なるため，直訳するだけでは適切な意味を伝えられない．このため，翻訳者は文化的な背景や言語の特性を理解した上で，適切な表現を選択する必要がある．

また，漢字文化の違い[2]も深く研究されている．同じ漢字でも，中国語と日本語で意味や使い方が異なることがある．例えば，「大丈夫」は中国語では「一人前の男」を意味し，日本語では「問題ない」を意味している．これらの違いを理解することは，言語学習や国際交流の観点から非常に重要である．

中国語と日本語の漢字の類似点と相違点を詳細に論じた論文[3]も多数発表されており，漢字の歴史的な発展[4]や，同形異義語が生まれた理由[5]などを研究することで，言語間の深いつながりを理解することができる．

一方で，これらの課題を情報処理によって分析する手法にはまだ検討の余地がある．そこで本研究では，中国語と日本語の同形異義語を題材として，可視化によってその意味の違いを分析する．

2.2 自然言語処理における「文埋め込み」の研究

「文埋め込み」は自然言語処理の重要な手法であり，短期間で大きく進展し，多様な方法が開発されている．

Arora らが 2017 年に提出した単語ベクトル加重平均法[6]はその一つである．このアプローチは，文中の各単語のベクトルを重み付けし，それらを平均化することで文全体のベクトル表現を得るというシンプルで直感的な方法であった．この手法の利点は，その明瞭さと直感性であるが，しかし，そのシンプルさゆえに，文の順序に関する情報の多くが失われてしまうという欠点もあった．

2018 年に提案された BERT[7]は，この情報の損失問

題に対する解決を試みた．BERT は左右の文脈を全層共有することで深層的な双方向表現を事前に学習し，文の順序情報を維持することを可能にした．これにより，より長い文脈を入力とし，長距離の意味理解を改善することが可能となった．BERT の出現はこの分野における大きな前進であり，その結果，多くのプリトレーニング(PLM)型の手法が開発され，自然言語処理技術が大きく進展した．

2021 年に Gao らは SimCSE 対比学習モデル[8]を提案した．これは比較学習を用いてモデルを学習する新たなアプローチであり，教師あり学習と教師なし学習の両方で利用可能である．実験結果によると，SimCSE 対比学習法は BERT よりも性能の高い「文埋め込み」を生成できることが示された．

2.3 次元削減と可視化などの研究

次元削減と可視化は，高次元のデータを理解しやすい形に変換するための重要な技術である．それらはデータマイニングや機械学習において広く利用されている．

主成分分析(PCA)は次元削減の目的でもよく用いられる．データの分散を最大化する方向を特定し，その方向を新たな「主成分」として使用する．これにより，データの本質的な構造を保持しながら次元を減らすことができる．しかし，PCA は線形な手法であり，非線形な構造を持つデータに対しては効果的ではないことがある．

それに対して，非線形の次元削減手法として t-SNE や UMAP などが知られている．これらの手法を用いることで，高次元データの複雑な構造を保持しながら 2 次元または 3 次元にマッピングすることが可能である．これにより，可視化の際にデータのクラスタやパターンを理解しやすくなる．

次元削減を適用した後のデータは，一般に散布図やヒートマップなどを用いて可視化される．これらの可視化手法は，データのパターンや傾向を一目で把握するのに便利である．

3. 提案手法

3.1 データの収集

本研究では，「日中同形異義語 1500」[9]という重要な文献を活用して，中国語でも日本語でも使われる熟語を列挙する．この文献は，日本語と中国語で同形でありながら意味が異なる約 1500 の熟語を紹介しており，本研究ではその中から本研究で最も関連性が高いと思われる 265 の例を選び出し，詳細に分析している．そして，中国語の意味と日本語の意味をいずれも日本語で記述する．表 1 には，その分析過程で明らかになった 5 組の代表的な例を示している．

表 1　日中同形異義語の例

単語	中国語の意味	日本語の意味
大丈夫	一人前の男	問題ない
得体	言動がその場にふさわしいこと	本当の性質や姿
汽車	エンジンによって路上を走る車	蒸気動力によって動く鉄道車両
前年	昨年の前の年	ある年の前の年
入口	口の中に葉入れること	ものごとの最初の段階

3.2 「文埋め込み」の計算

本研究では，「文埋め込み」の生成に焦点を当て，単語ベクトル加重平均法，BERT，SimCSE 対比学習法という 3 つの方法で「文埋め込み」の算出を試行した．

単語ベクトル加重平均法は最も非常にシンプルで実装しやすい方法であるが，文構造を失うことで単語の意味分析の情報が失われる可能性がある．結果として，異なる意味を持つが形式的に類似した文が，類似した文埋め込みになる可能性を生む可能性があるため，本研究ではこの方法は採用しなかった．

BERT 法と SimCSE 法はどちらも簡単に利用が可能である．BERT 法は Google が提案するオープンソースの言語モデルであり，学習や実装が簡単で，さらには中国語や日本語などの多言語モデルに対応している．SimCSE 法のモデルは GitHub で公開されており，Python で呼び出して文埋め込みを取得することができる．

BERT 法と SimCSE 対比学習法の結果を比較すると，同じ文でも，SimCSE 法を使って生成された「文埋め込み」は，BERT 法で生成されたそれよりも高次元を持つため，情報量が多いとされている．この高次元性は，情報の損失を抑えつつ，より詳細な意味表現を可能になる．一方，SimCSE は対比学習という方法を用いている．これは，同じ意味を持つ異なる形式の文を互いに近づけ，異なる意味を持つ文を互いに遠ざけるという方法である．これにより，SimCSE は，BERT が持つような文脈を考慮した強力な表現力を維持しつつ，より具体的な文の意味を捉えることができる．これらの理由から，本研究では SimCSE 対比学習法を適用して「文埋め込み」を算出する．ただし SimCSE 対比学習法の公開モデルは現状では英語しかサポートしていない．そこで現時点での本研究では，すべての意味を英語に翻訳してから「文埋め込み」を導出している．中国語や日本語への対応は検討中である．

3.3 次元削減と可視化

本研究では，同型異義語の分析において重要な要素であるデータの降維方法について，t-SNE（t 分布型確率的近傍埋め込み）と PCA（主成分分析）の 2 つを比較検討した．t-SNE は，特に高次元データの可視化に適しており，データの局所的な構造を保持する能力が高い．一方で，PCA は線形の手法であり，データセットの最大の分散を保持しながら次元を減少させる．この方法は，データの内在する構造を解釈するのに適している．

計算の複雑さの観点から，t-SNE は計算コストが高く，特に大規模なデータセットに対して大きな計算時間が必要とされる．また，初期値やハイパーパラメータの選択に敏感である．これに対して，PCA は t-SNE よりも計算量が小さく，大規模なデータセットをより迅速に処理できる．

解釈可能性に関しては，t-SNE はデータの局所的構造をよく表現するが，その結果の解釈性には欠ける．一方で，PCA はより高い解釈可能性を提供し，主成分はデータの重要な特徴として理解され，データの内在する構造の理解に寄与する．

以上の理由から，本研究では PCA を次元削減手法として採用することに決定した．この選択は，計算効率の高さ，データの全体的な構造を維持する能力，結果の解釈の容易さ，そして安定性にもとづいている．

本稿では次章にて，以下のような可視化を紹介する．

- i 番目の単語における中国語の意味の m 次元ベクトルをc_iとし，日本語の意味の m 次元ベクトルをd_iとしたときに，中国語の意味のベクトルの集合$C=\{c_1,c_2,\dots,c_n\}$と日本語の意味のベクトルの集合$D=\{d_1,d_2,\dots,d_n\}$を 1 つのデータセットとして混合して次元削減を適用する．ここで n は単語数とする．
- 必要に応じて，c_iとd_iをエッジで連結する．
- 必要に応じて，中国語と日本語の同一単語の意味のベクトルを結合した 2m 次元のベクトルに対して，Kmeans 法によるクラスタリングを適用する．

4. 実行結果と分析

4.1 「文埋め込み」のコサイン類似度

表 1 に示した 5 個の熟語に対して，中国語の意味と日本語の意味の文埋め込み結果のコサイン類似度を SimCSE 対比学習により算出した．結果を表 2 に示す．コサイン類似度が高いほど，その意味が似ていることを示している．その中で最も意味の似ていない熟語が「大丈夫」，最も意味の近い熟語が「前年」である．

続いて，265 個の単語を対象にして算出した余弦類似度の分布を図 1 に示す．この結果から，中国語と日本語の同型異義語間には，一定の意味的距離が存在することが明らかになった．特に，余弦類似度が中程度の範囲に集中しており，これは異なる言語間での用法の部分的な共通性を示唆している．しかし，全体的には，同型異義語が同じ形を持ちながらも，それぞれの言語で独自の意味を持つことが多いことが確認された．

表 2　5 個の例の「文埋め込み」のコサイン類似度

単語	中国語の意味と日本語の意味の「文埋め込み」のコサイン類似度
大丈夫	0.122
得体	0.299
汽車	0.412
前年	0.781
入口	0.177

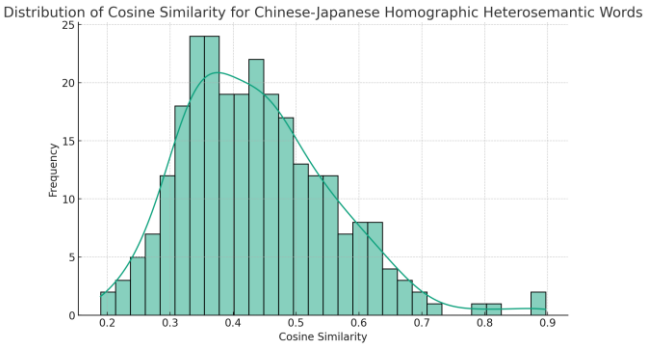


図 1 265 個の熟語に対する余弦類似度のヒストグラム

4.2 5 個の単語に対する可視化結果

図 2 は，5 個の単語を対象にして次元削減を適用して可視化した結果である．同じ色のドットは，同じ単語の中国語と日本語の熟語を示している．意味の違いが大きい単語は，2 つのドットからも離れていることが確認できる．

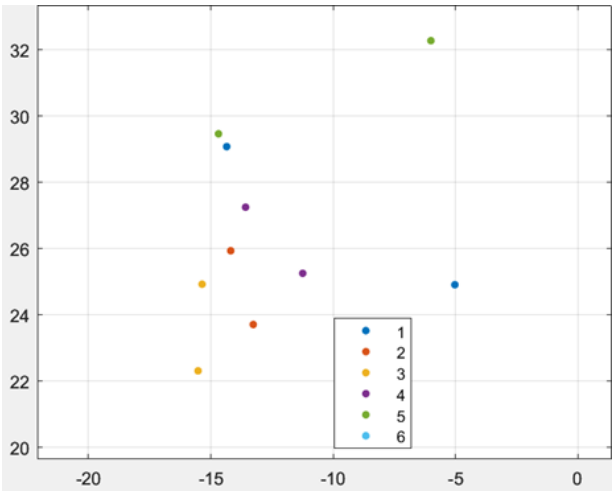


図 2 5 つの例に対するの可視化結果

4.3　265 個の単語に対する可視化結果

265 個の中国語と日本語の同型異義語について，その意味の文ベクトルを主成分分析（PCA）によって次元削減した結果を図 3 に示す．ここで，図 3(左)の赤色の点は中国語の文ベクトルを，図 3(右)の青色の点は日本語の文ベクトルを表す．

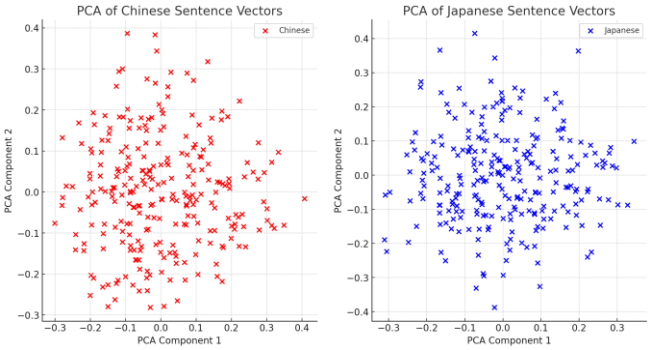


図 3　265 個の単語に対する次元削減結果

続いて，中国語と日本語の意味のベクトルを同一の画面に表示した図 4 の可視化結果から，同型異義語がそれぞれの言語においてどのように分布しているかを観察した．この可視化結果において，同一単語の中国語と日本語の意味のベクトルはエッジで連結されている．また，中国語と日本語の意味のベクトルを結合した 2m 次元のベクトルに対してクラスタリングを適用している，ここでクラスタ数は 5 であり，各点および各エッジはクラスタごとに色分けされている．5 つのクラスタを形成するこれらの文ベクトルは，言語間での意味論的な類似性を示している．同一クラスタ内でのエッジの隣接関係は，中国語・日本語ともに意味が近い同型異義語の存在を示唆している．

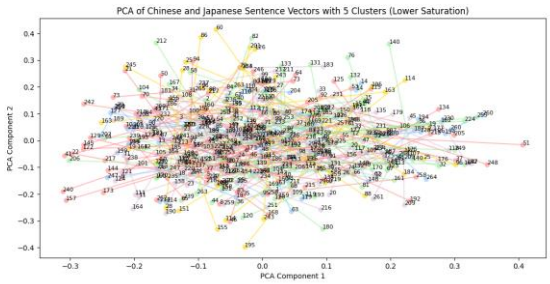


図 4 265 つの単語に対するの可視化結果

続いて，265 組の単語を対象として，K-means クラスタリングによって中国語と日本語のベクトルを 100 個のクラスタに分割した結果のうち，いくつかのクラスタについて図 5 に示す．

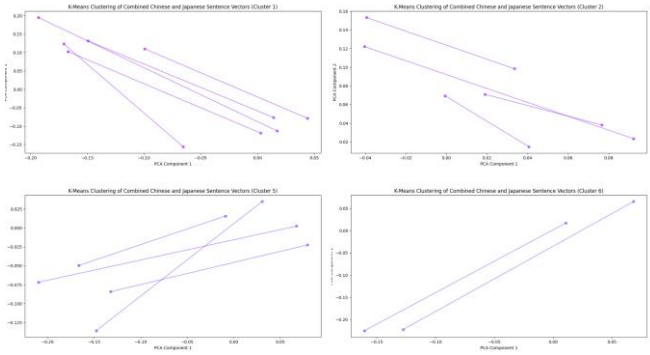


図 5　一部のクラスタを抽出した可視化結果

各クラスタの分析から，日中同型異義語の意味の進化と言語間の意味の変遷に関する興味深い観察が可能である．まず図 5(左上)に示すクラスタ 1 には，「帰還」「裁判」「純粋」「審判」「絶後」といった語が含まれており，これらはいずれも重要な社会的，法的概念に関連している．特に「裁判」と「審判」は，法の執行と判決に関連する言葉として両言語に共通する意味を持ちながらも，細かなニュアンスの違いが存在する．

図 5(右上)に示すクラスタ 2 には「官能」「群衆」「喧嘩」「談判」という語が含まれており，これらは社会集団や公共の場の行動に関連する言葉である．例えば，「官能」は日本語では感覚や感性を指すことが多いのに対し，中国語では官僚制度や能力を連想させることがある．

図 5(左下) に示すクラスタ 5 には「作業」「整理」「陳情」「通過」という語があり，これらはプロセスや手続きに関連する活動を指す言葉である．例えば，「通過」は両言語で「通過する」という意味が共通しているが，使用される文脈によってその意味合いに違いが出ている．

各クラスタの中での単語の選定は，それぞれの同型異義語が持つ独特の意味の変遷や文脈の違いにもとづいている．中日同型異義語は，漢字文化圏共通の語彙に起源を持ちながらも，時間の経過とともに，それぞれの言語と文化の中で独自の進化を遂げてきた．この歴史的な発展は，現在の意味の違いに大きな影響を与えていると考えられる．

例えば，クラスタ 1 に含まれる「審判」という単語は，中国語では主に裁判という意味で使われる一方，日本語ではより広い範囲の判断や評価という意味で使われることがある．このような違いは，文化や法制度の差異によって生じたと考えられる．

本研究におけるクラスタリング分析の結果から，日中同型異義語の意味の進化と言語間の意味の変遷に関する興味深い洞察が得られた．各クラスタに分類された語彙は，文化的背景や社会的慣習の違いにより，そ

の意味が独自の進化を遂げていることが明らかになった．特に，法的，社会的概念に関連する語彙や，プロセスに関連する活動を指す語彙は，その使用文脈の違いによって意味合いに顕著な変化が見られる．これらの発見は，言語が単にコミュニケーションの手段であるだけでなく，文化と社会の進化の証として機能していることを示している．

5. まとめと今後の課題

本研究では，中国語と日本語の同形異義語の意味の違いのパターンを可視化する試みを提案した．本研究では，中国語と日本語の同形異義語の意味を列挙し，SimCSE モデルを用いて，これらの意味を表す「文埋め込み」を形成する．そしてこれらの「文埋め込み」を次元削減して可視化する．本研究は．二つの言語間の微妙な意味の違いにどのようなパターンがあるかを明らかにしようとするものである．それは．私たちが異なる文化をより深く理解する手助けになるかもしれない．

現段階の本研究では，複数の意味を有する単語について，1 つの意味だけを適用している．しかし，複数の意味を有する単語の「文埋め込み」表現は，今後の研究における重要な課題として挙げられる．これを解決するためには，複数の意味を持つ単語の文脈による意味変化を的確に捉えるアルゴリズムやモデルが必要となる．

さらに，意味の違いのパターンをより明確に理解しやすくするために，新たな可視化手法の開発も求められる．より高度な可視化手法を用いることで，意味の違いのパターンをより詳細に，そして直感的に理解することが可能となる．

同形異義語の中には，中国語と日本語で異なる感情を表す単語も存在することがわかった.例えば,「緊張」という単語は，日本語ではネガティブな感情を示す一方で，中国語ではポジティブな感情を示す．可視化結果にこのような違いを加えることができれば，さらに新しいパターンを発見できることが期待される．

参　考　文　献

[1] 王悦, and 张积家. "不熟练中-日双语者同形词和非同形词的隐蔽翻译启动效应." 心理学报 46.6 (2014): 765.

[2] 徐娟. 日语汉字词在日译汉中的反向迁移现象. MS thesis. 南京农业大学, 2015.

[3] 李京. 中日同形词的对比研究——以检讨和 [検討] 为例[J]. 品位· 经典, 2022.

[4] 施建军, and 许雪华. "再论中日两国语言中的同形词问题." 解放军外国语学院学报 37.6 (2014): 132-139.

[5] 潘钧. "中日同形词词义差异原因浅析." 日语学习与研究 3 (1995): 19-23.

[6] Arora S., Liang Y., Ma T. A., Simple but tough-to-beat baseline for sentence embeddings, International conference on learning representations. 2017.

[7] Devlin J., Chang M. W., Lee K., et al., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] Gao T., Yao X., Chen D., SimCSE: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821, 2021.

[9] 郭 明輝 (著), 谷内 美江子 (著), 磯部 祐子 (著), 小谷 俊介 (編集). 日中同形異義語 1500. 国際語学社, 2011.

散布図選択による多次元データ可視化における表示の最適化

松尾　深景†　　伊藤　貴之†

†お茶の水女子大学　〒112-8610 東京都文京区大塚２丁目１－１

E-mail:　†{matsuo.mikage, itot} @is.ocha.ac.jp

あらまし

現代社会におけるデータの多くは多次元・多変数データである．これらの多くは散布図などを並べたダッシュボードを通じて可視化される. 多次元データからは多くの散布図が生成可能であるが, これらのすべてから適切な知見が得られるとは限らない. そのため, 我々はユーザが多次元データから迅速かつ効果的に情報を把握できるような重要な散布図を抽出するための新たな可視化手法を提案している．本研究では,「平安京ビュー」と呼ばれる階層型データを長方形の入れ子構造で類似度の高い散布図を近くに配置する手法を提案し，その効果について考察する．この手法により，多次元データの複雑な関係性が明瞭に可視化され，ユーザはより深い理解を得られることが期待される．

キーワード　ビジュアリゼーション，多次元多変数データ，散布図選択

1. はじめに

現在，一般的に扱われるデータは多次元多変数であることがほとんどである．このような多次元多変数のデータは，データの解析や理解の複雑性を増長させる要因となる．また，データ分析やデータの可視化は近年，企業や研究機関において非常に重要な役割を果たしている．そのため，多次元・多変数のデータを分析する際には，これらのデータを効果的に可視化することが重要である．

本研究ではこの手法を改良し，類似度の高い散布図をユーザが容易に比較できるような手法を実装している．具体的には，先行研究[1]で定義された散布図間の類似度を参照し，類似度の高い散布図が画面上で近くに配置されるように実装している．散布図間の類似度から算出された距離行列を用いて階層クラスタリングを適用し，その結果を利用して散布図を配置する手法を採用している.階層クラスタリングには最短距離法，重心法，ウォード法の 3 つの計算手法を実装し，それらの結果を比較した．さらに，階層クラスタリング結果から算出された散布図のクラスタ群の画面配置手法として，階層構造を視覚的に把握できる平安京ビューを採用する．

この改良には以下の 2 つの視点での効果がある．1 つ目は類似した散布図クラスタごとに階層的に整理された形で表示されるという点である．これにより，ユーザの関心に合った現象を表現する散布図を重点的に観察し，ユーザの関心に合わせてデータの理解を深めることができる．2 つ目は，類似している散布図が隣接して表示されるため，類似している散布図の中から微細な差異を見つけやすくする点である.これにより，データの意味やパターンについて新たな洞察をもたらす可能性がある．例えば，類似して見える散布図において，特定の点やクラスタに差異を見出すことで，データ分析の新しい方向性や仮説を生み出すことができる．

2. 関連研究

多次元多変数のデータの可視化の課題に対処するために，多次元データ分析のための効果的なデータ可視化手法が提案されている[1, 2, 3]．これらの手法を用いることで，インタラクティブな視覚的探索を可能にし，必要なダッシュボードを提供し，多次元データを直感的かつ明瞭に表現することができる．

多次元データ分析のための有効な手法として，次元削減技術が活用されている[4, 5]．次元削減は，複数の変数をより少ない次元に圧縮することで，データの特徴を保持しつつ解析の効率を向上させる手法であり，これらの手法を組み合わせることで，多次元データの解析と可視化をより効果的に行うことができる．それに対して，散布図行列に代表されるように，任意の 2 変数を選択して生成される多様な散布図を表示し，任意の変数間の相関を網羅的に表現する可視化手法も知られている．この手法の応用として本報告では，任意の 2 変数から生成される散布図のうち特定の条件を満たす散布図を抽出して表示する可視化手法の改良を提案する．

多次元データ可視化のための散布図選択手法は多数発表されているが，その中でも伊藤らの先行研究[1]では，多様な指標を同時に参照することで有限個の多様な散布図を選ぶ手法を提案している．我々はこの手法を改良し，類似度の高い散布図をユーザが容易に比較できるような手法を実装している．

伊藤らの先行研究では，各散布図に対して特徴ベクトルを生成し，コサイン類似度を用いて散布図間の類

似度を算出する．そしてグラフ彩色問題を適用することで，多様な散布図を選出する．本研究は伊藤らの研究を改良する形で実装されている．

3. 散布図の表示順の改良

3.1 提案手法の概要

本手法では，単一の多次元データから生成される複数の散布図を視覚的に比較する作業を効率化するために，類似して見える散布図を視覚的に比較しやすくするように，類似している散布図群を画面上で近くに配置する．そのために本手法では，散布図間の類似度にもとづいて階層クラスタリングを適用し，散布図のクラスタを生成する．類似した散布図を画面上で近くに表示するために，本論文では階層型データを視覚的に理解できる「平安京ビュー」を活用してクラスタ単位で散布図を表示する方法を紹介する．これにより，ユーザは効果的にデータのパターンや傾向を把握しやすくなることを期待できる．

3.2 散布図の指標

提案手法では，散布図の評価指標として先行研究[1]で使用された 4 つのスコアを使用している．それぞれの指標については以下で説明する．

1) 相関:

この指標は，散布図の 2 つの軸の間のスピアマンの順位相関係数を計算することで算出する．これは変数間の関連の強さと方向を測定するもので，相関係数の絶対値が大きいほど，正または負の相関が高いことを示し，変数間の関係性を明確に示し，解釈や分析に有用な散布図と評価される．

2) 細さ:

この指標は，散布図上の点の集まりが細長い領域を形成している場合に，その点の集まりがモデル化しやすいことを示す Wilkinson ら[6]で実装されたものである．まず，点の集まりを接続するドロネー三角メッシュを生成する．ドロネー三角メッシュは，与えられた点の集まりを結ぶ三角形の集まりである．生成された三角メッシュの中で，長い辺を持つ三角形から閾値以上の長さを持つ辺を持つ三角形を削除し，残った三角メッシュを用いて，三角メッシュの総面積と三角形の辺の長さの合計値から，面積に対する外周の比率を用いて，細長さを算出している．

この指標を使用することで，点の集まりの形状や分布の特徴を数値化し，モデル化の難易度を評価することができる．

3) クラスタや例外点の分離性:

この指標では，散布図の点の配置がいくつかのクラスタに分かれる，または明らかな例外点が存在する場合に，その特徴を把握する．前項と同様，点の配置を連結するドロシー三角メッシュを生成する．そして，メッシュを構成する三角形の辺を長さで最も長いものから順にソートする．

ソートされた辺の中で, i 番目と i+1 番目の辺の長さを使用して，以下の値

1- (i 番目の辺の長さ) / (i+1 番目の辺の長さ)

の最大値を求める．

この値が大きいほど，点群が明確に分離していることを示す．また，この値が小さい場合や，例外的な点が存在する場合は，点群の分布に関して特異なパターンや特徴があることがわかる．

4) クラスの分離性:

この指標では，特定のクラスを持つ点の集まりが他の点群から明確に分離されているかどうかを評価する．この手法は，情報エントロピーにもとづいてクラスの分離性を数値化し，散布図全体を複数の部分領域に分割する．各部分領域におけるエントロピーを計算し，最大エントロピーとの差を正規化した値がスコアとなる．このスコアは，点群の分離性を表し，値が高いほど点群がより明確に分離されていることを示す．

3.3 類似度にもとづく階層クラスタリング

ユーザが視覚的に比較し，意図に適した散布図を注視しやすくするようにするために，類似度の高い散布図を近くに配置する手法を導入する．この処理では，3.2 節で示した 4 種類の数値指標を 4 次元ベクトルとして，コサイン類似度にもとづいて階層クラスタリングを実行する．階層クラスタリングは，類似度の高い要素から順に連結し，次第に低い類似度を持つ要素も連結することで, 要素間の階層構造(デンドログラム)を形成する．階層クラスタリングで扱う要素間の距離として, 1 から類似度(-1 から 1 までのコサイン類似度)を引いたものを使用している．この結果を用いて類似度の高い散布図を画面上で近くに配置することで，類似度の高い散布図を視覚的に比較しやすくする．

本手法では最短距離法，重心法，およびウォード法を用いた階層クラスタリングの計算手法を採用している．最短距離法は，各クラスタの最も近い要素間の距離をクラスタの距離とするクラスタリング手法である．重心法は，クラスタの中心（重心）間の距離をクラスタの距離とし，ウォード法はクラスタの重心と散布図点の分散を考慮した距離をクラスタの距離とする．

例として，最短距離法を採用した階層クラスタリングによって形成されるデンドログラムは図 1 の通りである．各数字は散布図の番号を示している．

以上の処理によって得られたデンドログラムから，ユーザ指定のクラスタ数で散布図を分割ですることができる．例えば 9 個のクラスタに分割したい場合，図

1 に示したデンドログラムから次のようにクラスタを生成できる：
{23}, {1, 15, 5, 22}, {3}, {7, 9, 8, 14}, {16, 17}, {18}, {0, 25}, {2, 10, 1, 13, 19, 20}, {21, 4, 6, 12, 24}

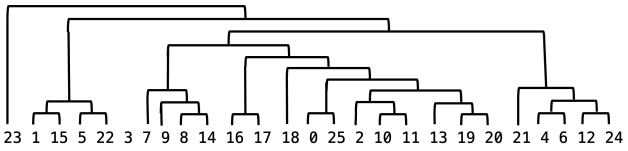


図 1. 最短距離法でのデンドログラム.

2.4 類似度にもとづく散布図群の画面配置手法の提案

2.3 節で算出された類似度にもとづいて，散布図を所定の順序に，あるいはクラスタごとに表示することで，類似度の高い散布図を近くに配置することが可能である．この配置により，ユーザは可視化の目的に沿って類似度の高い散布図を比較しやすくなる．また，データの分析時において類似する散布図間に見られる小さな差異が新たな視点として役立つ可能性がある．類似度の高い散布図群を画面上で近くに配置する方法として，先行研究[7]では，図 2 のように行列形式での表示を行っていたが，本手法では「平安京ビュー」を使用して散布図クラスタを表示する．

図 2. 行列形式でクラスタごとに表示した例.

伊藤ら[8]が考案した「平安京ビュー」は，階層型データを長方形の入れ子構造で可視化する手法である．図 3 のように，階層型データの葉ノードを長方形のアイコンで，枝ノードを長方形の枠で表現する．このようにして，階層構造を二次元の長方形群の入れ子構造で表現し，その全体を一画面に表現することを目標とした手法である．

この手法を応用して，類似度の高いクラスタごとに散布図を表示する．具体的には，葉ノードを散布図として考え，散布図の類似度の高いクラスタを枝ノードとして考える．これにより，各散布図がどの類似度のクラスタに属するかどうかを視覚的に確認できる．

以上の表示方法により，ユーザは効果的かつ直感的に類似した散布図を理解し，必要に応じてこれらを詳細に観察できるようになる．

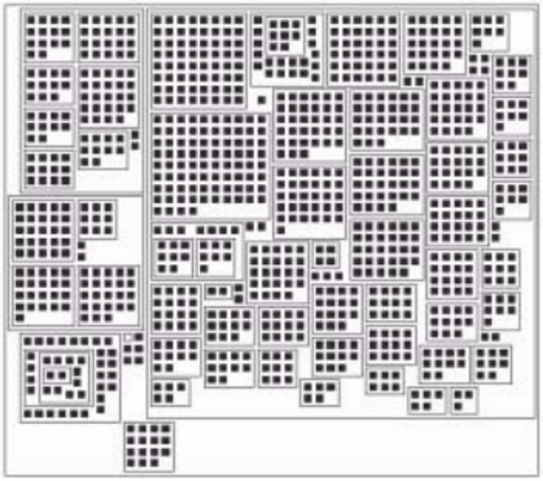


図 3. 「平安京ビュー」による大規模階層型データの可視化の例.

4. 実行例

3.3 節および 3.4 節で述べた通り，多次元データから生成された散布図の中から重要かつ多様な散布図を抽出し，その散布図群の特徴量の類似度に基づいて，最短距離法，重心法，ウォード法の 3 つの計算方法で階層クラスタリングを実行した．以下では，それぞれの計算手法と平安京ビューでの散布図群の配置手法に関する実行結果を示す．

表 1. 使用したデータの変数.

気象データ	
MinTemp	最低気温
MaxTemp	最高気温
SumRain	降水量
SumSunTime	日照時間
MaxWind	最大風速

販売データ	
Revenue	総売上額
Guest1	買上者数
Guest2	訪問者数
Ratio	訪問者数に対する買上者数の比
PerGuest	買上者あたり売上額
AveUnit	買上商品あたり価格
AveNum	買上者あたり商品数

データとして伊藤らの手法[1]と同じく，表 1 に示す気象値や販売値などの 12 個の変数で構成される多次

元データを用いる.

散布図群の距離行列をもとに，散布図の類似度の関係を多次元尺度構成法 (MDS: Multi-Dimensional Scaling) で示した結果を図 5 に示す. この図において，各点のラベルは 26 個の散布図の番号（図 4）を表している.

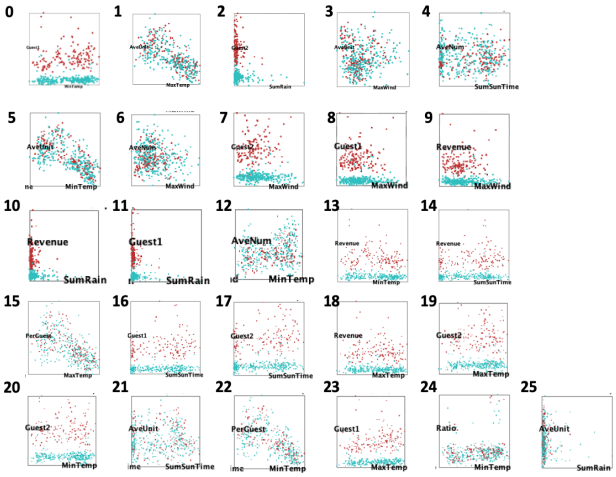


図 4 ラベルと散布図の対応

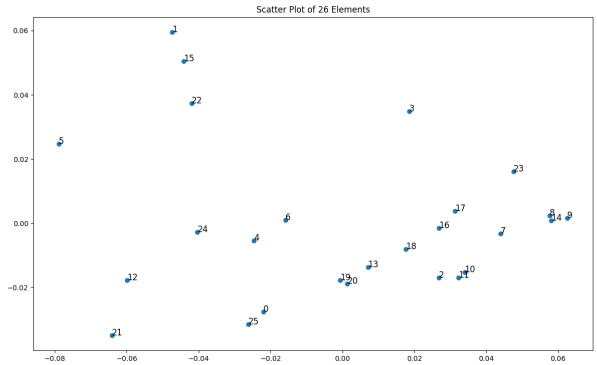


図 5 散布図間の距離行列から表した散布図の位置関係.

1. 最短距離法を用いた階層クラスタリングの結果

最短距離法を適用した結果として生成されるデンドログラムを図 1 に示す.

生成された 10 個のクラスタは以下の通りである：
[3], [5], [21], [22], [23], [1, 15], [16, 17], [7, 9, 8, 14], [4, 6, 12, 24], [18, 0, 25, 2, 10, 11, 13, 19, 20]

これらの 10 個のクラスタを多次元尺度構成法による次元削減結果に重ねて表示した例を図 6 に示す. また, 生成されたクラスタにもとづいて「平安京ビュー」で散布図を表示した例を図 7 に示す.

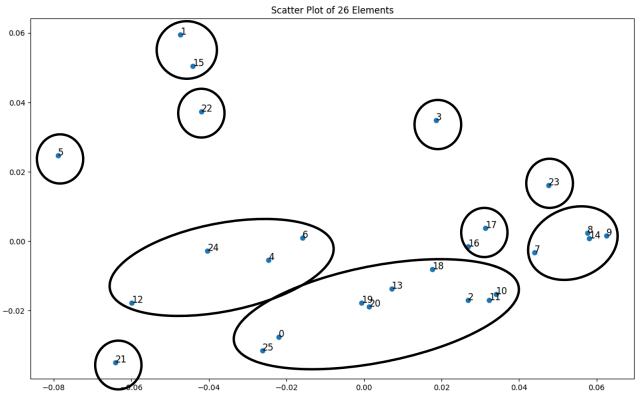


図 6 最短距離法を使った階層クラスタリング結果.

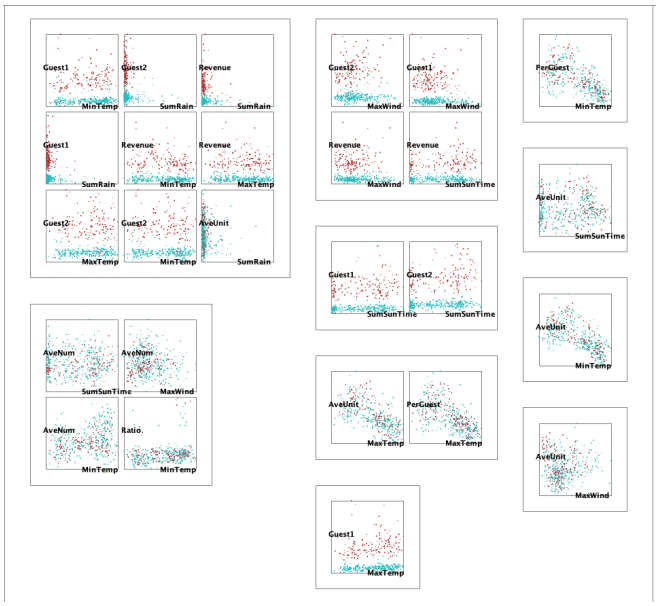


図 7 最短距離法を使った平安京ビューでの表示.

2. 重心法を用いた階層クラスタリングの結果

重心法を適用した結果として生成されるデンドログラムを図 8 に示す.

生成された 10 個のクラスタは以下の通りである：
[3], [5], [12], [21], [13, 19, 20], [22, 1, 15], [0, 25], [18, 2, 10, 11], [24, 4, 6], [23, 16, 17, 7, 9, 8, 14]

これらの 10 個のクラスタを多次元尺度構成法による次元削減結果に重ねて表示した例を図 9 に示す. また, 生成されたクラスタにもとづいて「平安京ビュー」で散布図を表示した例を図 10 に示す.

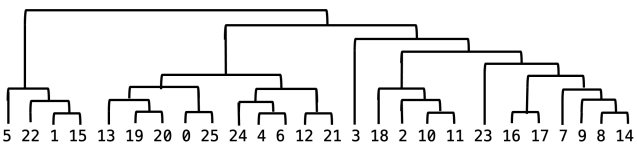


図 8. 重心法でのデンドログラム.

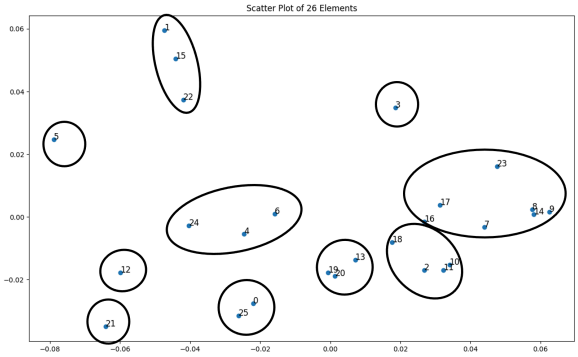


図 9 重心法を使った階層クラスタリング結果.

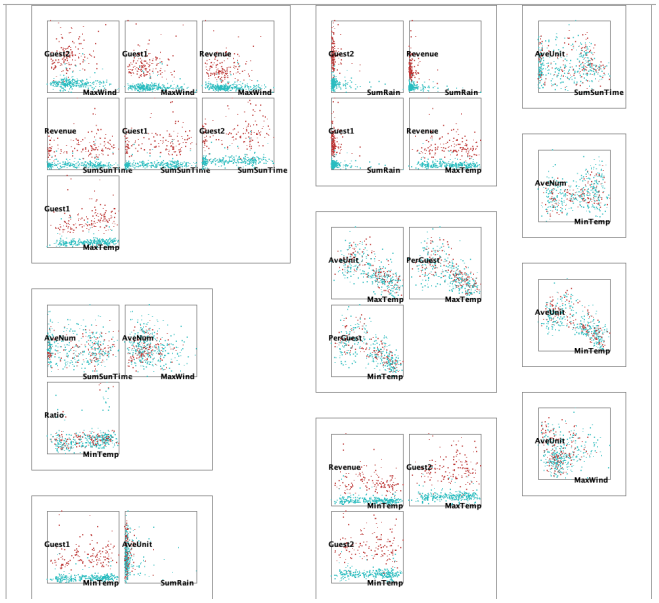


図 10 重心法を使った平安京ビューでの表示.

3. ウォード法を用いた階層クラスタリングの結果

ウォード法を適用した結果として生成されるデンドログラムを図 11 に示す.

生成された 10 個のクラスタは以下の通りである：

[[3], [5], [4, 6], [13, 19, 20], [22, 1, 15], [0, 25],
[7, 9, 8, 14], [18, 2, 10, 11], [23, 16, 17], [21, 12, 24]]

これらの 10 個のクラスタを多次元尺度構成法による次元削減結果に重ねて表示した例を図 12 に示す. また, 生成されたクラスタにもとづいて「平安京ビュー」で生成されたクラスタにもとづいて「平安京ビュー」で散布図を表示した例を図 13 に示す.

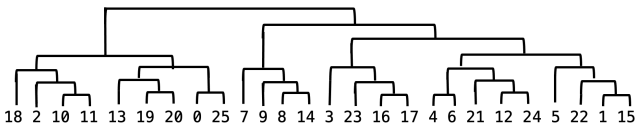


図 11. ウォード法でのデンドログラム.

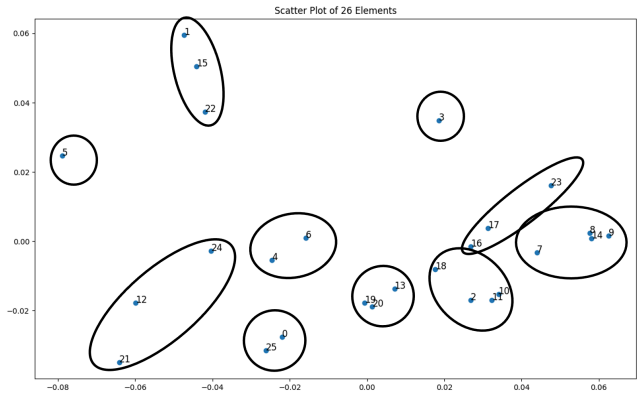


図 12 ウォード法を使った階層クラスタリング結果

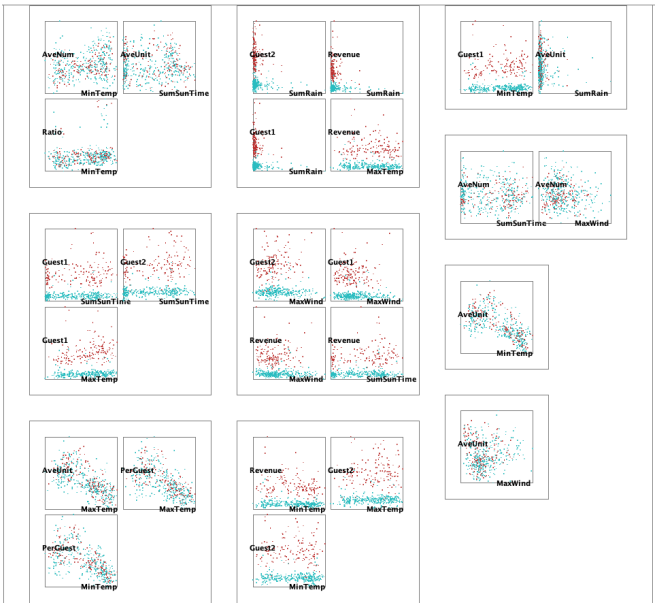


図 13 ウォード法を使った平安京ビューでの表示

以上の 3 種類の計算手法で階層クラスタリングを実施した上で, 散布図群の可視化に平安京ビューを適用したが, 図 4 に示す行列形式での表示に比べ, 類似度が高いとされた散布図によるクラスタの構造が明確である. そのため, 平安京ビューでの可視化は, 散布図群のクラスタ構成をより明瞭に表現できるといえる.

クラスタリングの性能を比較するために, 最短距離法, 重心法, ウォード法の 3 つのクラスタリング結果に対してシルエット分析を適用した. その結果を図 14 に示す. シルエット分析はクラスタリングの性能を評価する手法の一つであり, 優れたクラスタリングの条件として,「クラスタ内が密に凝集されている」かつ「異なるクラスタが遠く離れている」という 2 点が用いられる. シルエット係数はこの評価手法で用いられる指標で, その値は−1 から 1 の範囲になる. 値が 1 に近いほどクラスタリングの性能が高く, 負の値は所属クラスタの判別が誤っている可能性を示唆している.

本研究では, Python のライブラリである sklearn の

シルエット係数を採用し，3 つの計算手法それぞれで 2 から 10 個のクラスタの場合を算出した．結果として，重心法で 3 個のクラスタに分割した時にシルエット係数が最大となった．また，重心法はどのクラスタ数においても安定的に 0.5 以上の数値になっていることがわかる．ウォード法では，クラスタの分割数が増えるにつれて性能が向上することが示されている．特に，10 個のクラスタで分割した場合に，シルエット数が一番高くなった．一方で，最短距離法は分割数が多くなっても，1 つのクラスタに所属する要素の数が他の 2 手法に比べて多いため，シルエット係数は低い傾向にあった．

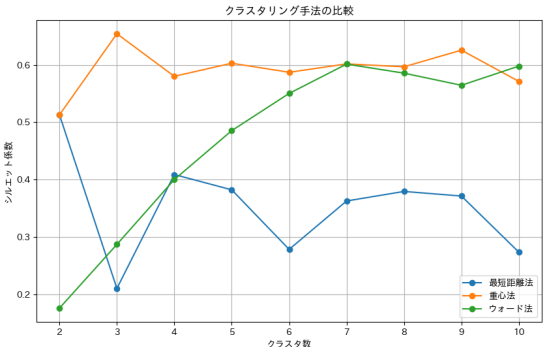


図 14 シルエット係数を用いたクラスタリング計算手法の性能比較

4. まとめ・今後の課題

本報告では，多次元データから生成される多数の散布図を選択的に表示する先行研究の改良手法として，選択された散布図群に階層クラスタリングを適用し，さらに平安京ビューを適用して画面配置する手法を提案した．可視化結果とシルエット分析から，重心法やウォード法を用いた可視化結果では，より類似度の高い散布図が近くに配置されていることが観察された．しかし，これらは定量分析結果および著者自身の主観にもとづく考察であり，第三者によるユーザ主観評価はまだ実施していない．

今後の課題として，提案手法の有効性を確認するためのユーザ評価が必要である．まず，ベースライン手法となる先行研究との比較実験を通して，提案手法の優位性や改善点を明確に示すことが求められる．さらに，実際のユーザにタスクを課する形での評価実験も，提案手法の実用性を確かめる上で不可欠である．提案手法を実際のユーザがどのように活用しうるかを理解するために，ユーザフィードバックをもとに，提案手法の改善点や調整が必要な部分を特定し，ユーザビリティの向上に焦点を当てたい．

多次元データには多様なものが存在し，データごとに異なる数値分布を有している．提案手法の柔軟性を確認するためには，本研究とは異なるデータセットやデータタイプにも適用できるかを検証する必要がある．異なる環境や条件下での実験結果を考察し，提案手法の汎用性を向上させるための調整や拡張を検討したい．

参 考 文 献

[1] T. Itoh, A. Nakabayashi, M. Hagita, Multidimensional data visualization applying a variety-oriented scatterplot selection technique, Journal of Visualization, 26(1):199-201, 2023.

[2] Q. V. Nguyen, N. Miller, D. Arness, et al. Evaluation on interactive visualization data with scatterplots, Visual Informatics, 2020.

[3] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, W. Magnorm. Quality-based visualization matrices, Vision, Modeling and Visualization (VMV), 2009.

[4] J. H. Lee, K. T. McDonell, A. Zelenyuk. D. Imre, K. Muller. A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. IEEE Transaction on Computer Graphics. 20(3):351-364. 2013.

[5] M. Ali, M. W. Jones, Dimension reduction applied to temporal data for visual analytics, The Visual Computer, 35:1013-1026, 2019.

[6] L. Wilkinson, A. Anand, R. Grossman. Graph-theoretic scagnostics. IEEE Symposium on Information Visualization, 157-164. 2005.

[7] 松尾，伊藤，散布図選択による多次元データ可視化における表示順の改良，可視化情報シンポジウム 2023, OS9-4, 2023.

[8] 伊藤，山口，小山田，長方形の入れ子構造による階層型データ可視化手法の計算時間および画面占有面積の改善，可視化情報学会論文集，26，6，51-61，2006.

DEIM Forum 2024 T4-B-3-04

粒度の異なる商品比較を実現する評判情報可視化インタフェースの提案

板垣　紫音† 　上野　史†† 　太田　学††

† 岡山大学大学院環境生命自然科学研究科　〒 700–8530 岡山市北区津島中 3–1–1
†† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山市北区津島中 3–1–1
E-mail: †p3w53x7t@s.okayama-u.ac.jp, ††{uwano, ohta}@okayama-u.ac.jp

あらまし　EC サイトでは，商品のスペックや評判情報を含む多くの商品情報が調べられる．しかし，商品情報が多くなると多数の商品を比較することは難しくなる．そこで，本稿では粒度の異なる 2 段階の商品比較を可能にする可視化インタフェースを提案する．具体的には，まず 2 種類の評価項目を軸にして多数の商品を散布図にプロットし，次にユーザが選択した商品に対して詳細な商品情報をレーダーチャートと商品の特徴を記載した表で提示する．なお，比較する情報として既定の評価項目と商品レビューから抽出した商品独自の評価項目を利用する．提案インタフェースの各機能の有用性や比較に適した商品数を評価するため，本稿では岡山大学の大学生と大学院生 10 名を対象とした被験者実験を実施した．その結果，商品ジャンル「ビール」において比較対象とした「価格.com」のインタフェースより提案インタフェースが優れていたとした被験者が多かった．

キーワード　可視化，商品レビュー分析，評判情報，インタフェース

1 はじめに

EC サイトにはユーザによる商品レビューを含む評判情報が多く投稿されている．これに加えて商品のスペック，値段，商品画像など商品の詳細情報も記載されている場合が多い．これらの情報のおかげで，商品を実際に目で見たり，手で触れたりしなくとも商品についての情報を得ることができる．ユーザはこれらの情報によって複数の商品を比較し，購入する商品を選んでいる．例えば商品比較サイトである「価格.com」では，ユーザは 1 つの商品ジャンルの商品群の中から比較したい商品を複数選択する．すると，選択した商品の商品画像やスペック，評価項目の評価点数などの比較項目を記載した表が表示される．この表によりユーザは一度に複数の商品を比較することができる．しかし，商品を比較する際に商品ジャンルや商品によっては比較したい情報が多くなることがある．例えば「価格.com」の比較表において商品ジャンル「ビール」の商品では，比較に用いる比較項目数が 24 項目ある．また，商品ジャンル「掃除機」では比較項目が 42 項目もある．そのため例えば「掃除機」の 10 商品をすべての比較項目で比較すると最大で 420 もの項目を確認する必要がある．このように比較したい情報が多い商品を複数個比較する際にユーザは多数の情報を確認せねばならず，比較したい商品数が多くなると商品比較が手間となり，場合によっては比較そのものが難しくなる．

そこで本稿では，ユーザが比較したい評価項目に絞って多数の商品を比較し，その中から自分が求めている特徴を持つ商品を選び，その詳細情報を比較できる可視化インタフェースを提案する．提案インタフェースでは，ユーザが自身で選択した評価項目を軸とした散布図によって多数の商品を比較する．さらにその散布図から詳細を比較したい商品を選択すると，商品間で共通するすべての評価項目を軸としたレーダーチャートや商品のスペックなどの情報が記載された表によって商品の詳細情報を比較できるようになる．このように，比較する商品情報の粒度を段階的に変えることにより，ユーザが気になる情報を持つ商品を多数比較することを容易にしている．また，詳細情報の評価項目は既定の評価項目と商品独自の評価項目である独自評価項目で構成しており，各商品固有の情報も確認できる．

本稿の構成は以下の通りである．2 節では関連研究について述べる．3 節では提案する評判情報の可視化インタフェースについて述べ，4 節では提案する評判情報の可視化インタフェースの評価実験について述べる．最後に 5 節で本稿をまとめる．

2 関連研究

2.1 商品比較に関する研究

Zhen ら [9] はオンラインレビューに基づいた商品比較ネットワークにより商品の競合分析を行う手法を提案した．Zhen らの手法ではまず，Latent Dirichlet Allocation (LDA) とセンチメント分析により，オンラインレビューから商品の特徴を抽出し，その特徴性能を評価した．特徴は共通特徴と商品固有の特徴を抽出する．次に，特徴性能を集計し，比較ネットワークを構築した．そして，製品の総合性能を求め，比較ネットワークを分析することで，製品比較，競合特定，製品ランキング，ブランド比較，市場構造分析を行った．Zhen らの手法では商品比較ネットワークにより商品の総合評価で複数の商品を比較できる．一方本稿では，商品レビューを用いて，複数の評価項目の評判を可視化する．これにより，商品の総合評価だけでなく，商品を複数の評価項目で比較することができる．

我々は [1] で商品レビューから抽出した単語を用いて生成した評価項目をレーダーチャートにより可視化し，2 商品を比較する手法を提案した．その手法ではまず，Google Natural

Language API の analyzeEntitySentiment[1]メソッドを使い，比較したい 2 商品のレビューから文章中の感情極性の対象となる語 Entity とその語の感情極性の値 Score を抽出する．そして Word2vec[7] を利用して抽出した Entity の単語分散表現を獲得し，獲得した単語分散表現を Ward 法によりクラスタリングする．生成されたクラスタから代表語を選出して主評価項目を生成し，主評価項目のクラスタの中から主評価項目に関連する評価項目である副評価項目とそれぞれの商品で評価されている独自評価項目も生成する．生成した評価項目は評価項目のクラスタに含まれる語の Score を用いて評判を数値化する．そして，生成した評価項目を軸，その評価項目の評判の数値を値としたレーダーチャートで可視化する．このレーダーチャートでは，ユーザが主評価項目をその主評価項目に関連する副評価項目に入れ替えることができ，自分が比較したい評価項目で 2 商品を比較することができる．しかし，この手法では評価項目以外の情報はほとんど可視化しておらず，商品のスペックや具体的なレビューなどを見ることができないため商品を比較するには不十分であった．また，重要とはいえない評価項目も多く生成された．本稿では，商品の比較にレーダーチャートだけでなく，散布図や商品のスペックなどを記載した表を組み合わせることでより詳細な商品比較が可能である．

2.2 商品の評判情報の可視化に関する研究

西川ら [4] は，5 つの特徴軸の内ユーザが指定した 3 つを組み合わせた 3 次元空間にレビューを可視化するシステムを提案した．西川らはまず，評判情報を探す上で有用な特徴軸として，「極性軸」，「信頼性軸」，「希少性軸」，「評価項目関連性軸」，「投稿日時軸」の 5 つを定義した．そして，レビューそれぞれの各軸の値を計算し，レビューの特徴ベクトルを生成した．ユーザが 5 つの特徴軸の内のレビューの探索に利用したい軸を 3 つ選ぶとその 3 つの軸で構成された 3 次元空間にレビューがプロットされる．この手法はユーザにとって有用なレビューを見つけやすくすることが目的のため，1 つの商品の評判情報を容易に比較できるが，2 つ以上の商品の評判情報の比較は考慮されていない．一方本稿では，複数商品の評判情報を比較できるように評価項目を軸とした散布図やレーダーチャートによる可視化を行う．これにより評価項目の評判を商品間で比較することもできる．

Warih ら [8] は商品のレビュー集合から抽出した評価項目とその評判により商品の総合評価を可視化するシステムを提案した．そのシステムではまず，レビューデータセットからアスペクトと肯定や否定などの意味を持つオピニオンのセットを抽出する．そして，アスペクトを含んだレビューの数と投稿したユーザの有効性から商品の総合評価を 5 段階で計算する．そして総合評価と各アスペクトを含む肯定的オピニオンと否定的オピニオンの数，割合を可視化する．本稿では，総合評価だけでなく各評価項目の評判を数値で可視化する．また，各評価項目についての評価をレビューの数やレビュアーの評価ではなく，レビューから抽出した単語の感情極性値から計算する．

3 商品比較のための評判情報可視化インタフェース

3.1 可視化インタフェースの概要

提案する評判情報可視化インタフェースの実装には，javascript のライブラリである React[2]，グラフの描画には Recharts[3] と ApexCharts[4] を用いる．

提案する評判情報可視化インタフェースを図 1 に示す．図 1 に示すように提案する評判情報可視化インタフェースは，「商品可視化部」，「総合情報比較部」，「評判情報比較部」の 3 つの機能で構成される．インタフェースの利用の流れは，まず，「商品可視化部」で散布図に多数表示されている商品の中からユーザが比較したい商品をいくつか選択する．次に，選択した商品の画像や値段などの詳細情報を表形式で表示する「総合情報比較部」や選択した商品に共通する評価項目を軸としたレーダーチャートで可視化する「評判情報比較部」で商品を比較する．そして，さらに比較したい商品があれば，「商品可視化部」で詳細を比較したい商品を再び選択する．このようにして，少ない評価項目での多数の商品比較，多数の評価項目での少数の商品比較，商品ごとのスペックなどの特徴による商品比較を行い，多数の商品比較を容易にする．

3.2 商品可視化部の実装

3.2.1 商品可視化部の概要

図 1 左上の「商品可視化部」では，著者らがあらかじめ設定した比較する複数商品に共通する評価項目 2 つを軸とし，その評判値を散布図で可視化する．散布図において点は商品を表している．また，マウスポインタを点の上に移動すると図 1 左上のように商品の画像や値段，総合評価，評価項目の評価値などの商品情報が表示される．そして点をクリックすることにより，レーダーチャートと表に選択した商品の評判情報が表示される．この散布図により，ユーザは気になる任意の 2 つの評価項目を利用して，多数の商品を比較できる．

また，散布図の軸は上部にあるプルダウンリストにより他の評価項目に変更することができる．そして，その下のスライダーバーにより選択した軸の評価項目の評判の数値で商品をフィルタリングすることができる．また，散布図に上部にあるフィルターのアイコンをクリックすると図 2 のようなタブが表示される．このタブ上で商品のブランドや商品の値段，機能の有無による商品のフィルタリングができる．この機能により，散布図に表示される商品を絞り，自分が比較したい特徴を持つ商品だけで商品比較ができる．

これらの機能によりユーザは自分が比較したい 2 つの評価項目を選んで多数の商品を比較できる．そして，任意の数の商品情報をレーダーチャートや表で可視化しながら詳細を比較した

1：https://cloud.google.com/natural-language/docs/reference/rest/v1/documents/analyzeEntitySentiment

2：https://ja.legacy.reactjs.org/

3：https://recharts.org/en-US/

4：https://apexcharts.com/

商品名	GOOSE IPA 355mlx24缶	スーパードライ 350mlx24缶	銀河高原ビール 小麦のビール 350mlx24缶	ザ・プレミアム・モルツ マスターズドリーム 350mlx24缶	ザ・プレミアム・モルツ 香るエール 350x24缶	ザ・プレミアムモルツ 350mlx24缶	ピルスナーウルケル 330mlx24本	イネディット 750ml 1本
ブランド	グースアイランド	アサヒビール	銀河高原ビール	サントリー	サントリー	サントリー	ピルスナーウルケル	Estrella Damm
画像								
値段	￥6500	￥4614	￥7033	￥5621	￥4978	￥4878	￥7473	￥1327
レビュー数	13	392	38	65	85	171	118	156
総合評価	3.52	3.6	3.25	3.86	4	3.7	3.83	3.75
風味	4.11	3.73	3.63	4	4.09	3.66	3.92	3.72
デザイン	3.98	3.43	2.87	3	3.95	3.67	4.07	3.93
のどごし	3	3.76	3.32	4.8	3.85	3.76	3.5	3.66
プレゼント	3	3.48	3.2	3.64	4.1	3.7	3.83	3.7
独自評価項目	-	-	-	-	-	-	-	-
度数	6%	5%	5%	5%	-	5.5%	4.4%	5%

図 1　提案手法のユーザインタフェース

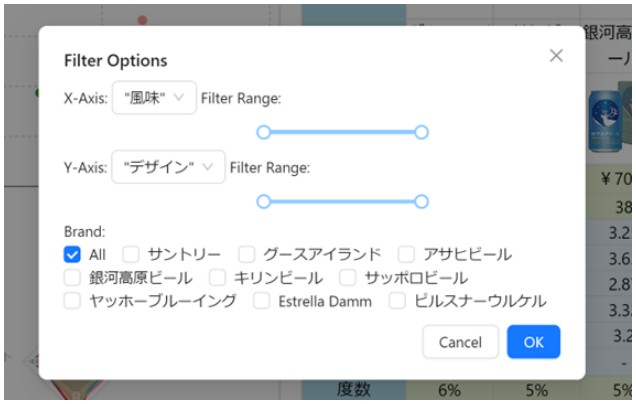


図 2　商品可視化部のフィルタリングタブ

い商品を絞ることができる．

3.2.2　既定の評価項目とその評判値

「商品可視化部」の散布図で可視化するため商品レビューから抽出した単語の既定の評価項目への分類とその評判の数値化を行う．まず，商品レビューから感情極性が示されている単語 Entity とその感情極性値 Score を抽出する．そして，抽出した Entity の単語分散表現を獲得する．その単語分散表現を利用して，Entity を既定の評価項目に分類する．既定の評価項目は EC サイトである「楽天市場」，「Amazon」，商品比較サイトである「価格.com」の 3 つのサイトを確認し，共通している評価項目や重要性を考慮して選出する．また，Entity の感情極性値 Score を利用して既定の評価項目の評判を数値化する．

商品レビューからの Entity と Score の抽出には Google Natural Language API の analyzeEntitySentiment メソッドを用いる．また，単語分散表現の獲得には Word2vec の事前学習済みモデルである chiVe[3] を用いる．chiVe は約 1 億のウェブページ文章を含む国立国語研究所の日本語ウェブコーパス (NWJC) を学習している Word2vec モデルである．

Google Natural Language API によりレビューから抽出される Entity には商品の評価項目として適切でない単語が存在する．また，Entity には代名詞など一般語のように頻出する単語も多数存在する．そのため，そのような単語をストップワードとして除去する．除去する語は以下の語である．

- 「発送」や「梱包」などの配送に関する語 13 語
- 「商品レビュー」や「星 5」などの商品レビューそのものに関する語 11 語
- 「掃除機」，「コードレス掃除機」など比較したい商品の商品ジャンルを示す語
- API において分かち書きに失敗している語
- SlothLib ライブラリ [5] が公開している不要語 310 語

API において分かち書きに失敗している語は Entity を再び分かち書きし，接続詞，助詞，助動詞を含んでいるかどうかの基準により自動で判定する．

Entity の既定の評価項目への分類には Support Vector Machine (SVM) を用いる．SVM は教師あり学習によってデータ群を分類または回帰する機械学習アルゴリズムである．本稿では，Python のライブラリである scikit-learn の SVM を利用する．SVM による分類では，まず，抽出した Entity のうち出現頻度が 5 回以上の Entity を人手で，著者らが設定した既定の評価項目とそれ以外の評価項目にタグ付けをする．タグ付けした Entity の単語分散表現を入力として SVM で教師あり学習を行う．そして，学習した SVM を分類器として，出現頻度が

5 回未満の Entity を既定の評価項目に分類する．

既定の評価項目の評判値を計算するため，まず分類されたすべての Entity の Score の平均を求める．そして，Score の平均を Min-max 正規化する．この計算によって得られた値を既定の評価項目の評判値とする．また，商品の総合評価の評判の数値化は既定の評価項目の評判値の平均とする．

3.3 総合情報比較部

3.3.1 総合情報比較部の概要

図 1 右の「総合情報比較部」には，「商品可視化部」で選択した商品のブランド名と商品名，商品の画像，値段，総合評価，各評価項目の評判値，商品独自の評価項目とその評判値，商品のサイズや重量などの基本スペックなどの商品の総合的な情報をまとめた表を表示する．選択した商品の数が増えると右に列が増え，下のスライダーバーを動かすことにより表の全体を確認する．

これによりユーザは商品のスペックなどの情報で多数の商品を比較することができる．また，商品の独自評価項目により，ユーザは商品固有の特徴も把握できる

3.3.2 商品情報の抽出

「総合情報比較部」に表示するために商品名，商品ブランド，商品画像，値段，商品のスペックを抽出する．商品の重さやサイズなどのスペックは商品説明文や商品レビューの内容から単位を含めて自動抽出する．また，商品の比較に必要な商品のスペックの項目は「価格.com」のスペックの一部の項目を利用する．例えば商品ジャンル「掃除機」では本体寸法，質量，集じん方式などを抽出する．

3.3.3 独自評価項目とその評判値

「総合情報比較部」に表示する商品レビューから商品独自の評価項目を生成し，その評判値を計算する．そのため，まず 3.2.2 項で既定の評価項目に分類できなかった Entity の単語分散表現を利用して Entity を Ward 法でクラスタリングする．そして，クラスタを代表する語を生成し，その単語を商品独自の評価項目とする．また，評判の数値化は 3.2.2 項と同様の方法で数値化する．

階層型クラスタリングを行う Ward 法では最小で 1 つのクラスタ，最大でデータの総数のクラスタを得ることができ，クラスタを結合する基準とその閾値を設定するとクラスタの結合を止め，任意のクラスタ数にすることができる．Bessel の補正を用いたクラスタの分散が最小になるようにクラスタを生成するため，比較的均一なサイズであり外れ値の影響を受けにくいクラスタを生成することができる．これにより，商品ごとに名前が異なるが似た機能，意味を持つ Entity を 1 つの評価項目にまとめることができる．また，本稿では，クラスタ結合の基準を不整合度を基に計算し，クラスタのリンクが閾値以上の不整合度を持つものは結合しない．

階層型クラスタリングによって生成されたクラスタからそのクラスタを代表する単語を生成する．単語の生成には ChatGPT[5] を用いる．ChatGPT にクラスタのリストとリストを代表する語を生成させる旨のプロンプトを入力し，生成された語を独自評価項目とする．プロンプトの例は以下である．

> 指示：今から渡す掃除機のレビューから抽出した語のクラスタを利用し，そのクラスタの意味をまとめたクラスタを代表する語を生成してください．
> #クラスタ： {entity1, entity2, entity3, ... }

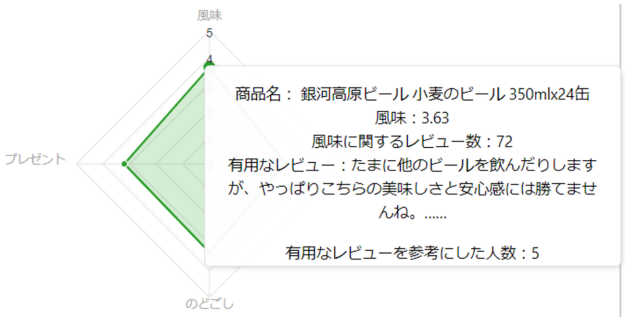


図 3 レーダーチャートの評価項目詳細表示テキストボックス

3.4 評判情報比較部

3.4.1 評判情報比較部の概要

図 1 左下の「評判情報比較部」に，軸を既定の評価項目，軸の値をその評判の値とするレーダーチャートを表示する．各軸の値の範囲は 1 から 5 である．このレーダーチャートによりユーザは「商品可視化部」で選択した複数の商品を評価項目の評判で比較できる．

また，各軸の商品の値を示す点の上にマウスポインタを移動させると図 3 のようにテキストボックスを表示する．このテキストボックスでは商品名と評価項目の評判値，関連するレビューの数，評価項目について記載されている重要なレビューを表示する．関連するレビューは各評価項目に分類される語が含まれるレビューである．重要なレビューは，その中から参考になったレビューとして投票された数が多いレビューを 1 つ選出する．このテキストボックスにより，ユーザは商品ごとにそれぞれの評価項目に関するレビューを確認し，評価項目の評判値だけでなく，その評判値になった理由などを知ることができる．著者らの研究 [1] によると，ユーザにとって商品の比較にはレビューの件数や実際のレビューが重要である．そのため，本稿では評判情報として関連するレビューの数，評価項目について記載されいる重要なレビューを表示する．

4 評価実験

まず，4.1 節で実験に用いる Word2vec モデルのファインチューニングについて説明し，4.2 節で提案システムを定性的に評価し，4.3 節で被験者による提案システムの有用性を評価する実験について述べる．

5：https://chat.openai.com

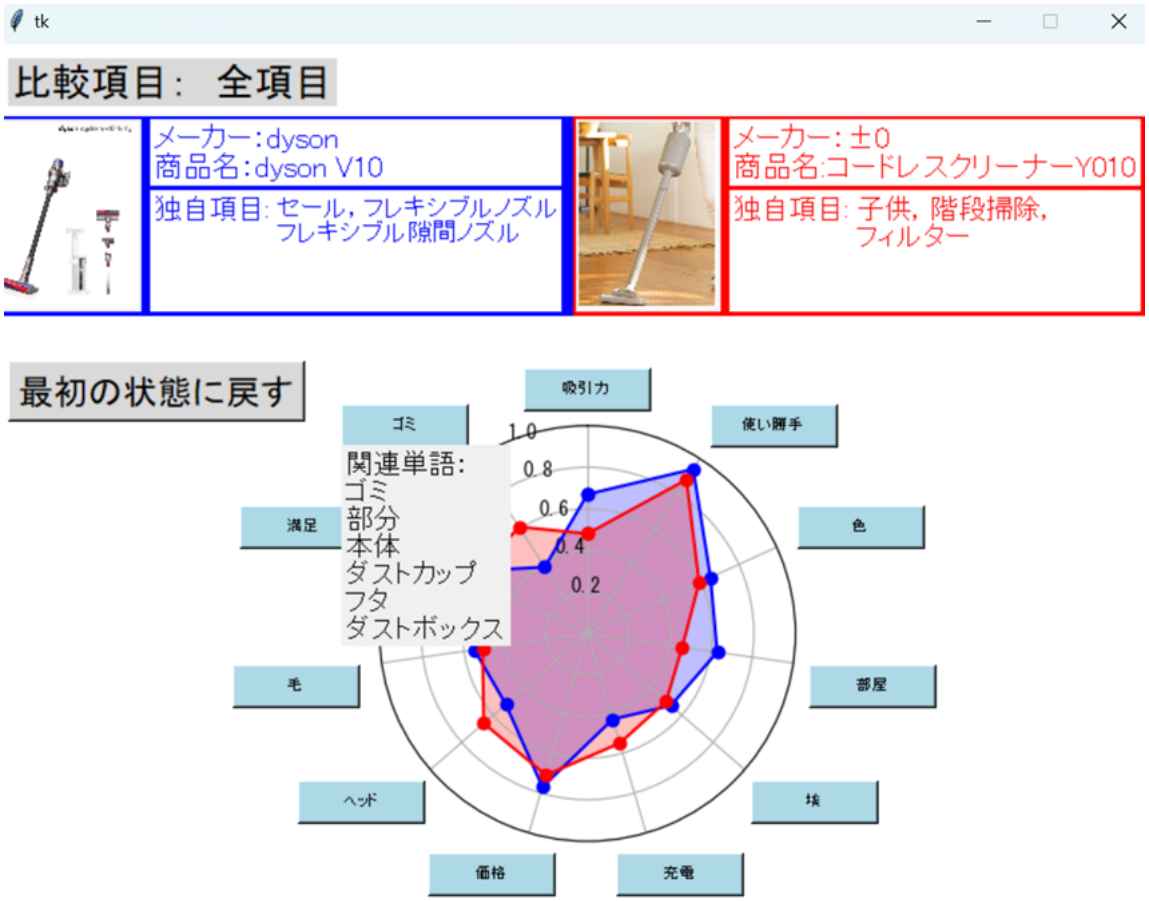


図 4　[1] で提案したインタフェース

4.1　Word2vec のファインチューニング

実験では，楽天グループ株式会社が提供している楽天市場データセットの中の「みんなのレビュー・口コミ情報」 [6] を使用して，Word2vec の事前学習モデルである chiVe をファインチューニングする．ファインチューニングには，2015 年から 2019 年の間に楽天市場に投稿された商品ジャンル「掃除機」の 130,4525 件のレビューと「ビール」の 64,808 件のレビューを形態素解析器 sudachi[2] によって分かち書きしたデータを用いる．

4.2　提案システムの定性的評価実験

4.2.1　実 験 概 要

実験では，提案インタフェースと著者らが [1] で提案した 2 商品比較を行うインタフェースで商品ジャンル「掃除機」の 2 商品の情報を可視化して定性的に分析する．商品ジャンル「掃除機」の商品として，「Dyson V10」と「± 0 Y010」を利用する．[1] で提案したインタフェースを図 4 に示す．

4.2.2　生成する評価項目の比較

提案インタフェースでは，商品ジャンル「掃除機」の共通評価項目として 6 つの既定の評価項目を利用している．それに対し，[1] のインタフェースでは比較する 2 商品のレビューから Entity を抽出し，Entity をクラスタリングし共通評価項目を生成している．実験で利用したこの 2 商品では 11 の評価項目が生成された．それらの評価項目を表 1 にまとめる．

表 1　提案インタフェースと著者ら研究 [1] のインタフェースの商品ジャンル「掃除機」の評価項目一覧

提案インタフェース	著者らの研究 [1] のインタフェース
吸引力，静音性，使いやすさ， サイズ，デザイン， 手入れのしやすさ	吸引力，使い勝手，色， 部屋，埃，充電，価格， ヘッド，毛，満足，ゴミ

表 1 より，共通する評価項目として「吸引力」があり，意味的に類似している評価項目として「使いやすさ」と「使い勝手」がある．また，提案インタフェースの「デザイン」に内含されている評価項目として [1] の「色」がある．これらの似た意味を持つ評価項目はどちらのインタフェースでも利用されており，掃除機の評価項目として必要な評価項目であるといえる．また，「静音性」については [1] において掃除機の評価項目として重要な評価項目であると評価されており，「サイズ」，「手入れのしやすさ」については「楽天市場」，「Amazon」，「価格.com」のすべてのサイトで共通して評価項目として設定されていることから掃除機の評価項目として必要であるといえる．[1] の実験から，「充電」，「価格」は [1] の重要な評価項目であることが分かっている．しかし，商品ジャンルの「掃除機」は充電式でないコード式の掃除機も含んでいることから「充電」は提案インタフェースの共通の評価項目として適切でない．「価格」はセールがあったり，時期により変動する要素であることから評価が一定でない．そのため提案インタフェースの評価項目とし

表 2　提案インタフェースと著者ら研究 [1] のインタフェースの「Dyson V10」と「± 0 Y010」の独自評価項目一覧

商品名	提案インタフェース	著者らの研究 [1] のインタフェース
Dyson V10	収納場所，アタッチメント	セール，フレキシブルノズル，フレキシブル隙間ノズル
± 0 Y010	ストレス，スイッチ，アタッチメント	子供，階段掃除，フィルター

表 3　実験に利用する商品ジャンル「掃除機」と「ビール」の商品一覧

No.	掃除機	ビール
1	CL100DW	イネディット 750mlx1 本
2	CL102DW	一番搾り 350x24 缶
3	CL105DW	一番搾り 500mlx24 缶
4	CL105DWN	エビス 350mlx24 缶
5	CL107FDSHW	エビス プレミアム ブラック 350mlx24 缶
6	CL141FDRFW	銀河高原ビール 小麦のビール 350mlx24 缶
7	CL182FDRFW	黒ラベル 350mlx24 缶
8	Dyson Digital Slim	ザ・プレミアム・モルツ 350mlx24 缶
9	IC-C102	ザ・プレミアム・モルツ 香るエール 350mlx24 缶
10	SY-089	ザ・プレミアム・モルツ マスターズドリーム 350mlx24 缶
11	XJC-B021	水曜日のネコ 350mlx24 缶
12	XJC-C030	スーパードライ 350mlx24 缶
13	XJC-Y010	ピルスナー ウルケル 330mlx24 本
14	ZC-SS24	GOOSE IPA 355mlx24 缶

てはふさわしくない．それ以外の評価項目は [1] で不適切と評価された．提案インタフェースではそれらの評価項目を除くことができている．

また，提案インタフェースと [1] で生成した「Dyson V10」と「± 0 Y010」の独自評価項目をそれぞれ表 2 にまとめる．この独自評価項目を比較すると，提案インタフェースでは「セール」と「子供」のように適切でない評価項目がなく，おおむねよい独自評価項目が得られている．

4.3　提案インタフェースを用いた商品比較の被験者実験

4.3.1　実 験 概 要

本実験では，岡山大学の情報系の学部生と院生 10 名に，「価格.com」と提案インタフェースを用いて商品ジャンル「掃除機」の 14 商品と「ビール」の 14 商品を比較させ，提案インタフェースの有用性について考察する．実験では，被験者は商品を比較し終えたら，購入する商品を 14 商品の中から 1 商品選択する．その後，提案インタフェースの有用性やインタフェースにおける商品比較に適した商品数，両インタフェースの優劣を被験者に尋ねる．

実験では被験者を 5 人ずつの 2 グループに分け表 3 に示す商品ジャンル「掃除機」と「ビール」の商品を提案インタフェース，「価格.com」で比較させる．提案インタフェースでの商品比較終了後，提案インタフェースの「商品可視化部」，「評判情報比較部」，「総合情報比較部」についての有用性の質問に回答させる．この質問の内容と回答形式は表 4 にまとめる．また，「価格.com」で商品比較終了後には，「価格.com」の評価項目の有用性に関して表 4 の 5 から 8 の質問に回答させる．また，各インタフェースを使った商品比較の終了後に，表 5 の質問に回答させ，最後に「見やすさ」，「使いやすさ」，「総合的な商品比較のしやすさ」の 3 つの観点でどちらのインタフェースが優れていたかの表 6 の質問に回答させる．実験の実施手順を表 7 にまとめる．

表 4　提案インタフェースの各表示部に関する質問と回答方式

No.	質問内容	回答方式
1	「商品可視化部」による商品比較は商品購入にどのくらい役立ったか.	5 段階評価
2	「商品可視化部」のデータのクリックによる商品の選択機能はどのくらい役に立ったか.	5 段階評価
3	「商品可視化部」のデータのマウスオーバーによる商品情報の表示はどのくらい役に立ったか.	5 段階評価
4	「評判情報比較部」による評価項目の比較はどのくらい商品購入に役立ったか.	5 段階評価
5	どの評価項目が役に立ったか.	選択式
6	それはなぜか.	記述式
7	どの評価項目が役に立たなかったか.	選択式
8	それはなぜか.	記述式
9	データをマウスオーバーして表示される有用なレビューは商品比較にどのくらい役立ったか.	5 段階評価
10	「総合情報比較部」は商品比較にどのくらい役立ったか.	5 段階評価

表 5　比較する商品数に関する質問と回答方式

No.	質問内容	回答方式
1	このインタフェースを使った商品比較にあたって 14 商品は多いと感じたか.	3 択
2	このインタフェースは商品数を増やしても問題なく使用できると思うか.	5 段階評価
3	このインタフェースによる商品比較ではどのくらいの商品数が適切だと思うか.	6 択

表 6　2 つのインタフェースの優劣に関する質問と回答方式

No.	質問内容	回答方式
1	商品比較をする上で総合的にどちらのインタフェースの方が見やすいか.	2 択
2	商品比較をする上で総合的にどちらのインタフェースの方が使いやすいか.	2 択
3	総合的にどちらのインタフェースの方が商品比較をしやすいか.	2 択

表 7　被験者の各グループが実施する実験の順番

実施順	被験者グループ A	被験者グループ B
1	提案インタフェースで「掃除機」の 14 商品比較	価格.com で「掃除機」の 14 商品比較
2	提案インタフェースの各表示部や機能についての質問	価格.com の評価項目についての質問
3	インタフェースで比較する商品数に関する質問	
4	価格.com で「ビール」の 14 商品比較	提案インタフェースで「ビール」の 14 商品比較
5	価格.com の評価項目についての質問	提案インタフェースの各表示部や機能についての質問
6	比較する商品数に関する質問	

4.3.2　実 験 結 果

表 8 に表 4 の質問の被験者の回答の平均をまとめる．表 8 を見るとどちらの商品ジャンルでも全ての評価が 5 段階評価の 3 を上回っている．特に「ビール」では，質問 3 と 9 以外の評価が 4 を上回った．

表 9 に提案インタフェースと「価格.com」の評価項目の評価を商品ジャンルごとにまとめる．なお，評価項目の評価は，評価項目が役立ったか否かの回答数である．表 9 を見ると，提案インタフェースの「掃除機」では，「吸引力」，「静音性」，「使いやすさ」が比較的高評価だった．これの評価項目を役立つとし

表 8　提案インタフェースの各表示部関する質問と回答

質問番号	商品ジャンル	
	掃除機	ビール
1(商品可視化部)	3.8	4.4
2(データ選択機能)	3.6	4.4
3(データ表示)	3.8	3.6
4(評判情報比較部)	3.2	4.2
9(有用なレビュー)	3.8	3.2
10(総合情報比較部)	4.2	4.0

表 9　提案インタフェースの評価項目の評価 (表 4 の質問 5，7 の回答数)

	商品ジャンル	評価項目	表 5 の質問番号	
			5(役に立った)	7(役に立たなかった)
提案インタフェース	掃除機	吸引力	4	0
		静音性	5	0
		使いやすさ	3	0
		サイズ	0	3
		デザイン	1	4
		手入れのしやすさ	1	1
	ビール	風味	2	1
		デザイン	0	3
		のどごし	5	0
		プレゼント	0	3
価格.com	掃除機	吸引力	4	0
		静音性	0	0
		使いやすさ	3	0
		サイズ	0	1
		デザイン	0	4
		手入れのしやすさ	1	0
		取り回し	1	1
	ビール	苦味	5	0
		キレ	0	3
		コク	0	4

表 10　商品比較にあたって 14 商品は多いか (表 5 の質問 1) への回答数

回答	提案インタフェース		価格.com	
	掃除機	ビール	掃除機	ビール
少ない	0	0	0	0
ちょうどよい	1	3	3	3
多い	4	2	2	2

表 11　商品数を増やしても問題なく使用できるか (表 5 の質問 2) への回答数

回答	提案インタフェース		価格.com	
	掃除機	ビール	掃除機	ビール
1(使いづらくなる)	1	0	0	0
2	1	2	5	0
3	2	2	0	3
4	1	0	0	1
5(問題ない)	0	1	0	1

た被験者の意見として「3 つの表項目とも掃除機にとって重要な要素である」という意見が多かった．それに対し，「サイズ」，「デザイン」は役に立たない評価項目とした被験者が多かった．その理由として「どちらも個人の感性によるところが多く評価することが難しい」という意見があった．また，提案インタフェースの「ビール」では「のどごし」が比較的高評価だったが，「デザイン」，「プレゼント」を役に立たない評価項目とした被験者が半数以上いた．この理由としては「人に渡す場合では有用であるが，個人で楽しむ際には必要ないから」などの意見があった．一方，「価格.com」では提案インタフェースと同様に「掃除機」の「吸引力」と「使いやすさ」を役に立つ評価項目とした被験者が多かった．しかし，提案インタフェースと違い「静音性」を役に立った評価項目とした被験者はいなかった．役に立たない評価項目としては「デザイン」を挙げた被験者が多かった．「ビール」では，「苦味」を多くの被験者が役に立った評価項目とした一方，「キレ」と「コク」は半数以上の被験者が役に立たない評価項目とした．この理由としては「キレやコクは飲んでいても分かりづらい」という意見が多かった．

表 12　商品比較における適切な商品数 (表 5 の質問 3) への回答数

回答	提案インタフェース		価格.com	
	掃除機	ビール	掃除機	ビール
10 以下	4	2	3	4
10〜20	1	2	2	1
20〜30	0	1	0	0
30〜40	0	0	0	0
40〜50	0	0	0	0
50 以上	0	0	0	0

表 13　グループごとの 2 つのインタフェースの優劣に関する質問の回答

質問番号	被験者グループ A		被験者グループ B	
	提案インタフェース	価格.com	提案インタフェース	価格.com
1(見やすさ)	5	0	4	1
2(使いやすさ)	2	3	2	3
3(商品比較のしやすさ)	2	3	3	2

表 5 の質問 1「このインタフェースを使った商品比較にあたって 14 商品は多いと感じたか」に対する回答を表 10 に示す．両インタフェースともに 14 商品を少ないと答えた被験者はいなかった．提案インタフェースの「掃除機」以外は，14 商品を商品比較する上でちょうどよいと答えた被験者が過半数だった．しかし，提案インタフェースの「掃除機」では 5 人中 4 人の被験者が多いと答えた．次に表 5 の質問 2「このインタフェースは商品数を増やしても問題なく使用できると思うか」に対する回答を表 11 に示す．提案インタフェースについては，商品数を増やすと使いづらくなる被験者が問題ないという被験者より少し多かった．「価格.com」の商品ジャンル「掃除機」では全員が使いづらくなると回答した．しかし，「ビール」では使いづらくなると回答した被験者はいなかった．次に表 5 の質問 3「このインタフェースによる商品比較ではどのくらいの商品数が適切だと思うか」に対する回答を表 12 に示す．おおむね 10 以下の商品数が適切であると答えた被験者が多かった．しかし，提案インタフェースの「ビール」では，10〜30 の商品数が適切であると答えた被験者がいた．

次に 2 つのインタフェースを比較する表 6 の 3 つの質問に対する回答数を表 13 にまとめる．どちらのグループでも半数以上の被験者が「価格.com」と比べ提案インタフェースが見やすい (質問 1) と答えた．しかし，どちらのグループでも半数以上

が提案インタフェースより「価格.com」の方が使いやすい (質問 2) と答えた．総合的な比較 (質問 3) については，両インタフェースの評価は拮抗していた

4.3.3 考　　察

実験結果から商品ジャンル「ビール」では「価格.com」より提案インタフェースの方が多数商品の商品比較に適しているとの評価が得られた．表 8 から提案インタフェースの各表示部の 5 段階評価の平均は「掃除機」で 3.73,「ビール」で 3.97 と比較的「ビール」の評価が高く，これは「ビール」では質問番号 1，2，4，10 の評価が高いためである．質問番号 1 の「商品可視化部」の評価が高い理由として，質問番号 2 の回からのデータの選択機能が商品比較に役立ったことが挙げられる．また，質問番号 9 は「掃除機」の評価が「ビール」より高くなっているが,「ビール」では参考になったとされるレビューが少なく有用なレビューがあまり獲得できなかった．そのため，有用なレビューの取得方法を参考になったレビュー数だけでなく，レビューを投稿したユーザーの評価など基準を利用して獲得する必要がある．

また，実験結果から「ビール」と比較して「掃除機」は両インタフェースで多数商品の比較が難しいと評価された．これは「掃除機」が「ビール」と比較して商品比較に影響を与える商品のスペックや評価項目が多いことが一因である．また，[1] から質量や運転時間，充電時間などの数値で表せられる商品のスペックが商品比較において重要であり，実験結果からは表形式である「価格.com」よりグラフを組み合わせた提案インタフェースが「見やすさ」の観点で優れているとされたことから，数値で表せられる商品のスペックのグラフ形式での可視化の検討が必要である．

5 ま と め

本稿では，ユーザが多数の商品を比較する状況を想定し，商品レビューから生成した評価項目の評判値を，散布図やレーダーチャート等で段階的に可視化するインタフェースを提案した．提案手法では，レビューから抽出した単語を分類したり，クラスタリングしたりして，商品に共通する既定の評価項目の評判と各商品独自の評価項目とその評判値を生成した．また提案する商品レビューの可視化インタフェースでは，ユーザは任意の 2 種類の評価項目を軸とする散布図を見て比較した多数の商品の中から任意の商品を選択し，選択した商品をレーダーチャートと表でさらに詳細に比較できる．

提案インタフェース全体の有用性を検証するために，大学生と大学院生あわせて 10 名による被験者実験を実施した．定性的な分析から，提案インタフェースの共通評価項目と独自評価項目として適切な評価項目を設定できた．被験者実験では被験者に提案インタフェースと既存の商品比較サイト「価格.com」で「掃除機」と「ビール」でそれぞれ 14 商品を比較させた．被験者実験の結果として商品ジャンル「ビール」では，提案インタフェースの各機能が役立ったため商品比較がしやすかったとした被験者が多かったが，有用なレビューは十分に得られなかった．また十分でなかった．提案インターフェース今後の課題として，レビュアー評価による有用なレビューの獲得や，商品のスペックのグラフ形式での可視化が挙げられる．

謝　　辞

本稿では国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供された「楽天データセット」(https://rit.rakuten.com/data_release/) を利用した．ここに記して感謝する．

文　　献

[1] 板垣紫音, 上野史, 太田学,“商品比較のための商品レビュー可視化の一手法,” DEIM2023, 4B-6, 2023.
[2] Kazuma Taoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, Yuji Matsumoto,“Sudachi: a Japanese Tokenizer for Business,” Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 2246–2249, 2018.
[3] 真鍋陽俊, 岡照晃, 海川祥毅, 高岡一馬, 内田佳孝, 浅原正幸,“複数粒度の分割結果に基づく日本語単語分散表現,” 言語処理学会第 25 回年次大会, pp. 8-5, 2019.
[4] 西川天帆路, 上野史, 太田学,“評判情報の特徴軸を考慮した可視化システム,” DEIM2021, D13-1, 2021.
[5] 大島裕明, 中村聡史, 田中克己,“SlothLib: Web サーチ研究のためのプログラミングライブラリ,”日本データベース学会 Letters, Vol. 6, No. 1, pp. 113–116, 2007.
[6] 楽天グループ株式会社,“楽天市場データ. 国立情報学研究所情報学研究データリポジトリ (データセット),” https://doi.org/10.32130/idr.2.1 (2022 年 6 月 19 日閲覧).
[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean,“ Distributed Representations of Words and Phrases and their Compositionality,” Proceedings of the NIPS 2013, Vol. 2, pp. 3111–3119, 2013.
[8] Warih Maharani, Dwi H.Widyantoro, Masayu L.Khodra,“Dynamic aspect-based rating system and visualization,” User Modeling and User-Adapted Interaction, pp. 1-24, 2022.
[9] Zhen He, Lu Zheng, Shuguang He,“A novel approach for product competitive analysis based on online reviews,” Electron Commerce Res 23, pp. 1–32, 2023.

Webブラウザのタブ群のタスクごとへのグループ化手法

久冨　瑞稀†　　田島　敬史†

† 京都大学情報学研究科　〒606–8501 京都市左京区吉田本町 36-1
E-mail: †hisatomi@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　Web ブラウザの使用において，タブなどの複数のページを並行して開く機能は複数タスクを並列で行うために不可欠である．しかし，ブラウザ内に様々なタブが乱立していると，開きたいタブを見つけるのが困難になる上，他のタブが目に入って気が散るためにタスクに集中しにくくなるという課題もある．解決策としてタブをタスクごとに分類してグループ化することが挙げられ，その為のツールも多く存在しているが，分類自体は手動で行う必要があり，そのコストは高い．機械学習を用いて手動分類を支援する方法も考案されてきたが，既存手法では自動分類するのではなくあくまで手動分類を支援するのに留まる上，学習データとしてアクセスしたページの URL といったプライバシーに関わるデータが必要になるという課題がある．本研究では，プライバシーに関わるデータを用いずにタブをタスクごとに分類する手法を提案する．具体的には，タブをノード，タブ間の遷移をエッジとする有向グラフを作成し，そのグラフにコミュニティ抽出の手法を適用して同じコミュニティに分類されたタブ群を同じタスクに属するものとすることで分類を行う．実験用データにこの提案手法と既存手法を拡張した手法を適用して，その結果を比較することで提案手法の評価を行った．

キーワード　ユーザー支援，操作ログ，タスク推定，web ブラウザ，タブ選択支援，グループ化

1　はじめに

現在，ウィンドウやタブなどの複数のページを並行して開く機能は Web ブラウジングに欠かせないものとなっている．ユーザーは，スケジュールの確認やドキュメントの作成など，様々なタスクを同時に行うために複数のウインドウやタブを開き，行き来することができる．

ここではタブについて考える．タブの利用に関する課題の一つとして，タブを多く開きすぎてしまい，ブラウザ内に様々なタスクのタブが混ざって乱雑になってしまう場合があることが挙げられる．多くのタブが乱立していると，一つのタブあたりの表示部分の幅が小さくなり得られる情報量が減少するため，その中から取り組みたいタスクに関連するタブを選択することが困難になる．また，取り組むタスクに関係ないタブも常に表示されているため，途中で他のタスクのことを思い出してしまったりしてタスクに集中することが難しくなる．

この課題を解決する方法の一つとして，タブ群をタスクごとに分類し，グループ化して表示させる方法が考えられる．多くの主要な Web ブラウザにはタブをグループ化する機能がすでに存在する．また，ユーザーがタブ群を分類する作業を支援するブラウザの拡張機能なども多数開発されている [1], [3], [12], [13]．このようなツールはユーザーがタスクごとにタブを手動分類するのを支援するものではあるが，分類作業自体のコストが高く，グループ化の機能だけでは十分に解決されたとは言えない．

このコストを下げるため，タブ群のタスクごとへ分類を支援する手法が考えられてきた．例えば，既存の拡張機能の中には，URL のドメインごとにタブを自動的に分類することによってユーザーを支援するものがある [10], [11]．しかし，一つのタスクを遂行する際には様々な URL のページを確認するのが一般的であり，ドメインごとの分類がタスクごとの分類に一致するとは考えにくい．

その他，タブ間の移動の前後のタブの情報から，その移動がタスクの切り替わりかどうかを推定し，ユーザーの分類を支援する手法が考案されている [3]．しかし，この手法はあくまで同じタスクに属すると予測されるタブをユーザーに提案するもので，最終的なタブのグルーピングはユーザーに委ねられている．さらに，この手法では事前に学習させた機械学習のモデルをユーザーに配布するという形式を取っており，かつ事前の学習にはページのタイトルなどプライバシーに関わるデータを用いている．ある程度の性能を出すためには十分な学習データが必要だが，プライバシーに関わるデータは複数のユーザーから収集することが困難である．実際，この提案手法で用いられたモデルの学習には，著者らのデータのみが用いられている．モデルを事前に学習させず，ユーザー自身のデータを用いてブラウザ内でモデルを学習させるという方法も考えられるが，この場合，学習過程でユーザーがタブを手動で分類をし，十分な量の学習データを提供する必要があるため，ユーザーの負荷が高くなってしまう．

本研究では，プライバシーに関わる情報を用いずに，タブ群を自動でタスクごとにグルーピングする手法を提案する．具体的には，タブをノード，タブ間の遷移をエッジとする重み付き有向グラフを作成し，そのグラフにコミュニティ抽出の手法を適用して，同じコミュニティと判断されたタブ群をグルーピングする方法である．[3] で提案された手法を，プライバシーに関わるデータを含まないようにした上で，タスクの切り替わりを検知した結果をもとにタブのグルーピングまで行うように拡張し，提案手法と性能を比較するための実験を行った．さらに，

分類性能を向上させるためのデータの事前処理の方法も考案し，その事前処理が性能向上に有効であるかについても検証した．

2 関連研究

2.1 Web ブラウザ内のタブの乱立

Web ブラウザの一部のユーザーがタブを多く開きすぎてしまい，ブラウザ内が乱雑になってしまう現象についての研究は過去に複数行われている．

Kulkarni らの調査 [1] では，インタビューの参加者の約半数がタブの乱立を問題視していることがわかった．

その他，Ma らの研究 [2] では，Web ブラウザを使用している間に感じる乱雑さについて，その形態や原因，対処法が調査された．その調査結果の中で，乱雑さの形態の 1 つとしてウィンドウやタブの数の多さが挙げられている．インタビューでは，参加者の多くが不快に感じ始めるタブ数の目安を持っており，それを超えるとストレスを感じることや，タブを開きすぎて視覚的にタブを区別できなくなるとブラウジング体験が悪化することが判明した．

このように，タブの乱立が Web ブラウザの使用における課題の 1 つであることが複数の研究で示されている．

2.2 タブの整理を支援する拡張機能

2.1 節で紹介した研究以外でもタブの乱立は問題視されており，タブの整理を支援するためのブラウザ拡張機能は数多く開発されている．ここでは機能が異なるもののいずれもタブの整理に関連するいくつかの拡張機能を紹介する．

OneTab [12] は，開いている複数のタブを 1 クリックで 1 つのタブグループにまとめたり，まとめたタブを復元することが出来る拡張機能である．また，複数のタブのグループを作成したり，それぞれのタブグループに名前をつけて管理することが出来る．

Acid Tabs [10] も OneTab と同様にタブのグループ化を支援する拡張機能だが，URL のドメインを利用した自動分類機能によってユーザーをより強力にサポートする．具体的には，各タブで開いている URL のドメインを取得し，ドメインが同じタブについては自動的にグループ化を行う．さらに，ユーザーは URL のドメインに基づいたカスタムルールを作成することができ，例えば複数の SNS のドメイン（facebook.com や instagram.com など）を 1 つのグループにまとめるカスタムルールを作成することが出来る．

これらの拡張機能はいずれもタブ管理をサポートするものではあるが，タブをタスクごとに分類するには手動での作業が必要になる．

2.3 タスクベースの Web ブラウザインタフェース

2.2 節で紹介した拡張機能以外にも，乱立するタブやページをタスクごとに分類することによってユーザーの支援を行う「タスクベースの Web ブラウザインタフェース」が複数の研究で開発されてきた．

Rajamanickam らは，Web ページのタスクごとへの分類や注釈の付与，タスク内の Web ページ群の共有を可能にする TabFour を開発した [4]．実際に TabFour を使用したユーザーへのインタビューでは，参加者全員から「整理整頓がしやすい」といった好意的な意見を得られている．

Chang らが開発した Tabs.do [3] では，タブのタスクごとへの分類や，複雑なタスク構造の整理，タスクの優先順位づけなどを行うことが出来る．Tabs.do を実際に使用したユーザーのログデータから，これらの機能をユーザーが積極的に使用し，ウィンドウやタブの乱雑さの減少などの変化があったことがわかった．

このようにタスクベースの Web ブラウザインタフェースもユーザーのタスク遂行の効率化に貢献しているが，どのインタフェースでもタスクの分類自体はユーザーが行う前提となっている．

2.4 タスクごとへの分類支援

Web ブラウザを利用してタスクを遂行する際には，複数のタブを使用したり複数の検索エンジンのクエリを発行することが多いことを認識し，タブや検索エンジンのクエリを自動的にタスクごとに分類する手法を提案した研究も存在する．

2.3 節でも述べた Tabs.do [3] ではタスクの分類をユーザーが手動で行う前提ではあるものの，その分類を支援するための機能も実装されている．この機能は，ユーザーが Web ブラウザを使用している間に収集した履歴データから同じタスクに属するタブを推定する．具体的には，タブの切替前後の各タブの情報を用いてその切替が新しいタスクセッションの開始かどうかを判断し，新しいタスクセッションだと判断された場合には切替前後のタブは別のタスクに属するとみなす．これをタスクの切替とし，タブの利用履歴の中で 2 つのタスク切替の間にあるタブ群は同じタスクに属すると推定してそのタブ群をユーザーに提示することで，ユーザーの手動分類を支援する．

この切り替えの判断にはニューラルネットワークが用いられており，著者ら 4 人の Web ブラウザの閲覧履歴データから学習を行った学習済みモデルがユーザーに配布される．なお，このネットワークの特徴量として以下を使用している．

- 現在のタブのページと直前のタブのページの間のタイトルの類似性
- 現在のタブのページと直前のタブのページの間の URL 間の正規化レーベンシュタイン距離
- 現在のタブのページと直前のタブのページの間のドメイン間の正規化レーベンシュタイン距離
- 現在のタブのページが Google 検索かどうか
- 直前のタブのページが Google 検索かどうか
- 過去にその URL にアクセスした回数
- アドレスバーにその URL が入力された回数
- 現在のタブから前のタブまでの秒数

この手法はタスクの分類支援に貢献するものではあるが，ページのタイトルや URL といったプライバシーに関わる情報を直接利用する必要があるという課題がある．また，この機能はあくまで同じタスクに属すると推定されるタブをユーザーに提示

することでユーザーの手動分類を支援するものであり，タスクごとへのグルーピング自体は行わない．本研究で提案する手法では，プライバシーに関わるデータを用いることなく分類を行い，タブのタスクごとへのグルーピングまで行う．

2.5 ネットワークのコミュニティ抽出

提案手法では，タブのタスクごとへの分類のためにコミュニティ抽出の手法を用いる．コミュニティ抽出とはグラフから互いに密に接続されたノード群（コミュニティ）を特定することで，ソーシャルネットワークなどの大規模なグラフデータの分析においてよく用いられる．ここでは，本研究で用いるコミュニティ抽出の手法について述べる．

コミュニティをどれだけ上手く抽出できているかを測る指標として，Newman ら [5] が提案したモジュラリティがよく用いられる．モジュラリティとは，同じコミュニティ内のノードを接続するエッジ数の割合を測定し，同じコミュニティ分割を持つがノード間の接続がランダムなネットワークにおける同じ量の期待値を差し引いたものである．よって，コミュニティ内のエッジが密な場合にモジュラリティが高い値をとる．モジュラリティQ は以下の式で表される．

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

ここで，A_{ij} はノード i とノード j の間のエッジの重み，$k_i = \sum_j A_{ij}$ はノード i に接続されたエッジの重みの合計，m は全てのエッジの重みの合計，c_i はノード i が属するコミュニティ，関数 $\delta(u, v)$ は $u = v$ の場合に 1，それ以外の場合は 0 となる関数である．

このモジュラリティの最大化を目指すことによってコミュニティ検出を行う手法はこれまでに盛んに考案されてきた．その中でも計算速度や抽出精度の点で優れている Louvain 法 [6] はよく用いられている．この手法では，まず全てのノードを別々のコミュニティに割り当て，以下のステップをモジュラリティが改善しなくなるまで繰り返す．

（1） あるノード i について，隣接するコミュニティと統合した場合のモジュラリティの増加分 ΔQ を算出し，ΔQ が最大となるコミュニティに統合する．これをグラフ内の各ノードについて行い，モジュラリティが改善しなくなるまで繰り返し行う．

（2） ステップ 1 で得られたコミュニティそれぞれについて，コミュニティ内のノードを集約して新しい重み付きグラフを作成する．

頂点 i をコミュニティC に追加した場合の ΔQ は以下の式で算出される．

$$\Delta_Q = \frac{k_i^C}{2m} - \frac{\sum_{tot}^C \cdot k_i}{2m^2} \tag{2}$$

ここで，k_i^C はノード i とコミュニティC 内のノードの間のエッジの重みの合計，$\sum_{tot}^C$ はコミュニティC 内のノードに接続されたエッジの重みの合計を表す．

Louvain 法で算出される ΔQ は [5] で提案されたモジュラリティに基づいているが，これは無向グラフに適用することを前提としたものである．Dugué らの研究 [7] では，Louvain 法を有向グラフに拡張した有向 Louvain 法が考案された．この手法では，ΔQ の計算に以下の式を用いる．

$$\Delta_Q = \frac{k_i^C}{m} - \left[\frac{k_i^{\text{out}} \cdot \sum_{\text{tot,in}}^C + k_i^{\text{in}} \cdot \sum_{\text{tot,out}}^C}{m^2} \right] \tag{3}$$

ここで，k_i^{out} と k_i^{in} はそれぞれ，ノード i を始点/終点とするエッジの重みの合計，$\sum_{\text{tot,in}}^C$ と $\sum_{\text{tot,out}}^C$ はそれぞれ，コミュニティC 内のノードを終点/始点とするエッジの重みの合計を表す．

本研究で作成するタブの使用履歴のグラフは有向グラフであるため，有向 Louvain 法を用いる．

3 提 案 手 法

本章では，プライバシーに関わるデータを除いたタブの利用履歴データからグラフを作成し，コミュニティ抽出を行うことでタスクごとへの分類を行う手法を提案する．また，提案手法の比較対象として先行研究で用いられて手法を拡張したものを用いるが，その手法の詳細についても述べる．どちらの手法においても，分類性能の向上のために利用履歴のデータに対して事前処理を行うが，その内容についても説明する．

以下，コミュニティ抽出を用いた提案手法を手法 1，既存手法を拡張した比較のための手法を手法 2 とする．

3.1 使用するデータの詳細

今回提案する手法ではユーザーのタブの利用履歴のデータを利用する．1 つの履歴には各タブに一意に割り振られた ID とタブを開いた日時，匿名化した URL，匿名化したドメインが含まれる．

3.2 データの事前処理

利用履歴データに各手法を適用する前に，事前処理として取り除くべき履歴データを削除する．具体的には，以下のいずれかに該当するものを取り除くべき履歴データとする．

- 開きたいタブを開くまでの過程で，そのタブを探すために他のタブを開いた場合や，ショートカットキーで移動したときに経由したことによって他のタブを開いた場合の履歴．これを通過の履歴と呼ぶ．
- メールの確認やスケジュールの確認など，他のタスクの遂行中に割り込んで行われる可能性が高いタスクのためのタブを開いた履歴．これを割り込みタスクタブの履歴と呼ぶ．

通過の履歴は，そのタブを利用した履歴とは見なすべきではないと考えたため除外対象とした．通過の履歴の判定基準にはそのタブでの滞在時間を用い，滞在時間が 0.5 秒未満の履歴、もしくは滞在時間が 1 秒未満かつその履歴前後の履歴のタブの滞在時間も 1 秒未満の履歴を通過の履歴と判定する．

割り込みタスクタブの履歴は，別タスクの遂行中にタスクと無関係なタブを開いた履歴として残り分類に支障が出ると考えたため除外対象とした．割り込みタスクタブの判定基準には何

種類のタブからそのタブに遷移してきたかやそのタブから何種類のタブに遷移したかの種類数を用いる．利用履歴に記録された全てのタブの種類数に対する，遷移元と遷移先のタブの種類数の合計数の割合を算出し，0.1 以上のものを割り込みタスクタブと判定する．

3.3 手法 1：グラフのコミュニティ抽出による分類

1 つ目の手法は，タブの利用履歴から，各タブをノード，タブ間の遷移をエッジで表現するグラフを作成した後グラフのコミュニティ抽出を行って，同じコミュニティに分類されたタブを同じタスクに属するとみなす手法である．以下で詳細を説明する．

3.3.1 タブの利用履歴のグラフの作成

上記で述べたように，利用履歴から，各タブをノード，タブ間の遷移をエッジとする有向グラフを作成する．具体的には，タブ A からタブ B に遷移し，タブ B からタブ C に遷移した場合には，A → B → C というグラフが作成される．タブ A からタブ B への遷移が複数回行われた場合には，その回数をノード A，B 間のエッジの重みとする．

なお，グラフの作成にあたって，過度に長い時間を要したタブ間の遷移のエッジはノイズになり得るため除外した．このような遷移は，しばらく Web ブラウザを使用していなかった時間があった場合などに起こる．このような理由で，あるタブを開いた時刻と次のタブを開いた時刻の間に大きな差がある場合は，タブ間の遷移というよりはタスクの終了と開始と見なすほうが自然であると考えられる．よって，タブ間の遷移に一定の時間より長い時間を要していた場合は，エッジで接続しない．本研究ではこの一定の時間を 5 分に設定した．

3.3.2 グラフのコミュニティ抽出

作成したグラフに 2.5 節で紹介した有向 Louvain 法を用いてコミュニティ検出を行う．なお，タブの利用履歴には分類時より前にすでに閉じられていてブラウザ内に存在しないタブの履歴も含まれているが，この過去のタブの使用履歴も含めたままでグラフを作成した上で有向 Louvain 法を適用する．その後，分類時で開いているタブとそれらに割り当てられたコミュニティのデータを取り出してコミュニティの検出結果とする．

3.4 手法 2：ランダムフォレストによる分類

2 つ目の手法は，2.4 節で述べた方法と同様に，タブの利用履歴内の連続した 2 つの履歴に対して，そのタブの切り替わりがタスクの切り替わりになっているかを機械学習を用いて判定し，それを元に分類を行う手法である．本研究では，データ提供者からのデータを用いて事前にモデルを学習させておき，実際の分類においてはそのモデルを使用することを想定している．以下でモデルの詳細と学習させたモデルを用いてタスクごとに分類する手法について述べる．なお，データ提供者からのデータを用いてモデルを学習させる方法については 4 章で説明する．

3.4.1 モデルの詳細

本研究では，タスクの切り替わりを判定するためにランダムフォレスト法を用いて学習させた分類器を用いる．特徴量として，タブの利用履歴の中の連続した 2 つの履歴についての以下の情報を用いる．

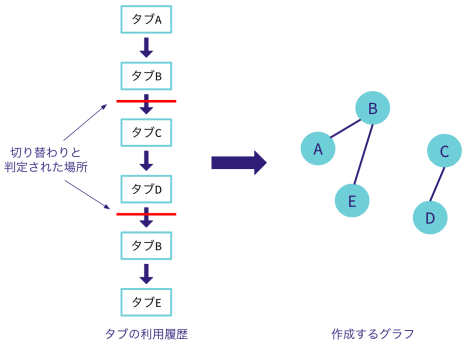


図 1　手法 2 におけるグラフの作成方法

（1） 2 つのタブを開いた時刻の差
（2） 1 の値から，2 つのタブを開いた時刻の差の平均を引いた値
（3） 2 の値を，2 つのタブを開いた時刻の差の平均で割った値
（4） 2 つのタブの URL が同じかどうか
（5） 2 つのタブの URL のドメインが同じかどうか
（6） 過去に 2 つのタブ間での遷移が起こった回数
（7） 6 の値を利用履歴内の全ての遷移の回数で割った値

3.4.2 モデルを用いたタスクごとへの分類方法

タスクごとに分類する際には，学習させたモデルと，タブをノード，タブ間の遷移をエッジとするグラフを用いる．

まず利用履歴内の全ての連続したペアについてタスクの切り替わりかどうかを学習済みモデルで判定する．その後，図 1 のように，タスクの切り替わりと切り替わりの間にある履歴で開かれたタブは同じタスクに属すると考え，タブのノード間に遷移のエッジを作成する．この処理を利用履歴全体に対して行い，完成したグラフにおいて連結成分になっているタブ群を同じタスクに属するタブとみなす．

4 実験方法

本章では，提案手法の有効性を評価するための実験方法について述べる．この実験では，事前処理の有効性を検証することおよび，手法 1 と手法 2 のどちらの方がより適切にタブの分類を行えるかを比較検証することを目的とする．

実験を行うにあたり，8 人のデータ提供者から 1 週間程度の実際のタブの利用履歴のデータを取得した．また，データの収集終了時に開いているタブを手動でタスクごとに分類してもらい，それぞれのタブにラベリングしてもらった．なお，提供者のうち 7 人からは 1 週間の履歴の収集およびラベリングを 2 回行ってもらったため，合計で 15 個のデータを収集することができた．これらのデータに事前処理あり・なしと手法 1・2 を組み合わせた 4 パターンの処理を施し，後述する指標を用いて評価する．

4.1 手法 2 の分類器の学習方法

3.4 節で述べた通り，手法 2 では利用履歴の連続したペアに対して，タスクの切り替わりかどうかを分類器を用いて判定する．分類器の学習には，利用履歴のデータに加えてタスクの切り替わりかどうかのラベルが必要だが，今回のデータ収集では利用履歴のペアがタスクの切り替わりかどうかのラベリングは行っておらず，収集終了時点で開いていたタブ群に対してのみタスクごとに手動分類したラベルが付与されている．そのため，利用履歴内の連続したペアのうち，両方とも手動分類されたタスクのラベルが付与されており，タスクの切り替わりが判定できるもののみを学習用のデータセットとして用いた．

また，3.4 節で述べたように，本研究では，データ提供者からのデータを用いて事前にモデルを学習させておき，そのモデルを利用して新たなユーザーのタブの分類を行うことを想定している．そのため，今回の評価実験では，データ提供者のそれぞれについて，本人以外のデータを用いて学習させたモデルを本人のデータに適用し，それによって得られた分類結果を評価する．

4.2 各手法の評価方法

各手法によって分類された結果は以下の 3 つの指標で評価を行う．指標 1 と 2 は，提案手法によって解決したい課題の解消に貢献しているかを測るために独自に導入するものである．指標 3 はクラスタリング結果の評価に使用される一般的な指標の 1 つである．

（1） 同じタスクのタブへの移動に必要なコスト
（2） タスクの純度
（3） 調整 rand 指数

4.2.1 指標 1: 同じタスクのタブへの移動に必要なコスト

以下，データ提供者によって手動分類されたグループをタスク，各手法によって分類されたグループを分類グループと呼ぶ．また，提案手法によって分類された結果，同じタスクに属するタブ群が複数の分類グループに分割された場合，一番多くのタブが含まれる分類グループをそのタスクのメイン分類グループとする．1 つの分割グループが複数のタスクのメイン分類グループとなる場合も許容する．

コストは，メイン分類グループ内のあるタブを開いている場合に同じタスクに属する他のタブ全てに移動するのに必要なクリック数の期待値によって計測する．各タスクについて，メイン分類グループ内かつそのタスクに属するあるタブから同じタスクに属する他の各タブに移動するコストを計算し，全てのタスクの計算値を合計することでコストを算出する．

あるタスクに着目してコストを計算するにあたり，必要なコストは，メイン分類グループ内で同タスクのタブを探すコストとメイン分類グループ外から同タスクのタブを探すコストの 2 つに分類出来る．それぞれについて計算方法を以下に示す．

タブ，分類グループの総数をそれぞれ n，m とする．また，着目しているタスクに属するタブの総数を T とする．各分類グループを C_1，C_2，$\dots C_m$ とし，C_i に含まれるタブの数を n_i，C_i に含まれるタブのうち着目しているタスクに属するタブの数を t_i とする．着目しているタスクのメイン分類グループを C_{main} とし，C_{main} に含まれるタブの数とそのうち着目しているタスクに属するタブの数をそれぞれ n_{main}，t_{main} とする．

まず，メイン分類グループ内で同タスクのタブを探すコストについて考える．C_{main} 内において，そのタスクに属するあるタブ A を開いているとき同じタスクに属する他のタブ B を探す場合を考える．C_{main} 内のタブを順に開いていったときに，1 回目でタブ B を開く確率は $1/(n_{\text{main}}-1)$ であり，その場合に必要なクリック数は 1 である．2 回目でタブ B を開く確率は $(n_{\text{main}}-2)/(n_{\text{main}}-1)\cdot 1/(n_{\text{main}}-2)$ であり，その場合に必要なクリック数は 2 である．同様に考えていくと，タブ A を開いているときにタブ B を探すのに必要なクリック数の期待値は

$$\begin{aligned}&(\frac{1}{n_{\text{main}}-1}\cdot 1)+(\frac{n_{\text{main}}-2}{n_{\text{main}}-1}\cdot\frac{1}{n_{\text{main}}-2}\cdot 2)\cdots\\ =&\frac{1}{n_{\text{main}}-1}\cdot\{1+2+\cdots+(n_{\text{main}}-1)\}\\ =&\frac{1}{n_{\text{main}}-1}\cdot\frac{n_{\text{main}}(n_{\text{main}}-1)}{2}\\ =&\frac{n_{\text{main}}}{2}\end{aligned}\tag{4}$$

と計算できる．同じタスクに属するタブのうち B 以外のタブについても同様に計算できるので，C_{main} 内で同タスクのタブを探すコストは

$$\frac{n_{\text{main}}}{2}\cdot(t_{\text{main}}-1)\tag{5}$$

と表せる．

次に，メイン分類グループ外から同タスクのタブを探すコストについて考える．メイン分類グループ外から同タスクのタブを探す場合，分類グループ群から 1 つグループを選択し，その中からタブを開くという 2 つのステップがあると考え，1 つ目のステップにも 1 クリックのコストがかかることとする．また，1 つ目のステップで選んだ分類グループに探しているタブが含まれていない場合，その分類グループに探しているタブが無いことに気付くためのコストもかかると考えられる．探しているタブがないことに気付くのには分類グループ内の全てのタブを開く必要があると考えると，必要なクリック数は選択した分類グループ内のタブの数に等しいと考えられるが，ここでは簡単にするために，必要なクリック数は分類グループに含まれるタブ数の平均，すなわち n/m とする．いまメイン分類グループ C_{main} 内のタブ A を開いており，C_{main} ではないある分類グループ C_i に含まれているタブ C を探す場合を考える．メイン分類グループ以外の $m-1$ 個の分類グループを順に調べていった時に 1 回目で C_i を開く確率は $1/(m-1)$ である．C_i 内でタブ C を開く確率や必要なクリック数はメイン分類グループ内で考えたものと同様だが，分類グループを開くのにクリック数 1 が必要なのでこれも考慮する必要がある．同様に考えると，k 回目に開いた分類グループが C_i であった場合，そのタブを見つけるのに必要なクリック数の期待値は，

$$\begin{aligned}
&\frac{m-2}{m-1}\cdots\frac{1}{m-k}\cdot\Big[\{\frac{1}{n_i}\cdot(1+k+\frac{n(k-1)}{m})\}\\
&\qquad+\{\frac{n_i-1}{n_i}\cdot\frac{1}{n_i-1}\cdot(2+k+\frac{n(k-1)}{m})\}\\
&\qquad\vdots\\
&\qquad+\{\frac{n_i-1}{n_i}\cdots\frac{1}{1}\cdot(n_i+k+\frac{n(k-1)}{m})\}\Big]\\
&\qquad=\frac{1}{m-1}\cdot\frac{1}{n_i}\cdot\Big[\frac{n_i(n_i+1)}{2}+n_i\{k+\frac{n(k-1)}{m}\}\Big]\\
&\qquad=\frac{1}{m-1}\cdot\{\frac{n_i+1}{2}+k+\frac{n(k-1)}{m}\}
\end{aligned}\tag{6}$$

と計算できる．C_i 内のタブに移動するために必要なクリック数の期待値を求めるには，$k=1,2,\cdots,m-1$ の全てについて (6) 式の値を計算し合計すればいいので，

$$\begin{aligned}
&\frac{1}{m-1}\cdot\{\frac{n_i+1}{2}+1\}\\
&+\frac{1}{m-1}\cdot\{\frac{n_i+1}{2}+2+\frac{n}{m}\}\\
&\vdots\\
&+\frac{1}{m-1}\cdot\{\frac{n_i+1}{2}+(m-1)+\frac{n(m-2)}{m}\}\\
&=\frac{1}{m-1}\cdot\Big\{\frac{(m-1)(n_i+1)}{2}+\frac{m(m-1)}{2}+\frac{n}{m}\cdot\frac{(m-1)(m-2)}{2}\Big\}\\
&=\frac{1}{2}\{m+n_i+1+\frac{n(m-2)}{m}\}
\end{aligned}\tag{7}$$

となる．C_i 内の着目しているタスクの属するタブ全てに対して同様に計算できるので，C_i 内にある同タスクのタブを探すコストは，

$$\frac{1}{2}\{m+n_i+1+\frac{n(m-2)}{m}\}t_i \tag{8}$$

と表せる．メイン分類グループ外から同タスクのタブを探すコストは，メイン分類グループ以外の全ての分類グループのこの値を合計したものなので，

$$\sum_{i=1}^{m}\frac{1}{2}\{m+n_i+1+\frac{n(m-2)}{m}\}t_i\phi(i) \tag{9}$$

と表せる．ただし関数 $\phi(i)$ は

$$\phi(i)=\begin{cases}1 & (C_i\neq C_{\mathrm{main}})\\ 0 & (C_i=C_{\mathrm{main}})\end{cases} \tag{10}$$

である．

(5) 式と (9) 式を足したものが求めるコストである．このコストがより小さい方がより良い分類であると考える．

4.2.2 指標 2: タスクの純度

タスクを行っている時に集中を削ぐ要因の 1 つとして，誤って分類された別タスクのタブを開いてしまうことが挙げられる．よって，あるタスクに着目する場合，そのタスクに属するタブを含む全ての分類グループ内のタブの合計に対するそのタスクに属するタブの割合，すなわち，そのタスクに属するタブの純度が高いほど，より良い分類であると考えられる．この純度は

$$\frac{\sum_{i=1}^{m}t_i}{\sum_{i=1}^{m}n_i\sigma(i)} \tag{11}$$

によって計測できると考えられる．ただし，

$$\sigma(i)=\begin{cases}1 & (t_i\neq 0)\\ 0 & (t_i=0)\end{cases} \tag{12}$$

である．この値を全てのタスクに対して算出し，その平均を取った値を評価指標とする．この値がより大きい方がより純度が高くより良い分類であると考える．

4.2.3 指標 3: 調整 rand 指数

調整 rand 指数は，Rand [8] によって考案された rand 指数を拡張したものである．rand 指数とは，2 つのクラスタリング結果の類似度を測る指標である．n 個の点を K 個のクラスタにクラスタリングした結果のクラスタ群を

$$Y=\{Y_1,Y_2,\cdots,Y_K\} \tag{13}$$

で表すこととする．同じ n 個の点に対する異なる 2 つのクラスタリング結果 Y と Y' を考える．n 個の点の中のあるペア o_i, o_j に着目する時，このペアの Y と Y' における状態としては以下の 3 つのパターンがある．

（1） Y と Y' の両方において，同じクラスタに分類される
（2） Y と Y' の両方において，異なるクラスタに分類される
（3） Y と Y' の一方では同じクラスタに分類され，一方では異なるクラスタに分類される

パターン 1 と 2 は，そのペアの分類において Y と Y' の間で意見が一致していることを表し，パターン 3 は Y と Y' の間で意見が不一致であることを表していると考えられ，1 と 2 に当てはまるペアが多いほど Y と Y' の類似度が高いと考えられる．Y と Y' の rand 指数 $c(Y,Y')$ は，全ての点のペアのうちパターン 1 と 2 に当てはまるものの数の割合で定義される．

調整 rand 指数は，Hubert ら [9] が rand 指数を拡張した指数で，ランダムにクラスタリングされた場合の rand 指数の値を考慮して補正したものである．調整 rand 指数は以下で定義される．

$$\frac{\text{rand 指数}-\text{rand 指数の期待値}}{\text{rand 指数の最大値}-\text{rand 指数の期待値}} \tag{14}$$

今回の実験では，手法 1 と手法 2 のそれぞれの分類結果に対して，データ提供者によって手動分類された結果との調整 rand 指数を算出して比較する．この値がより大きい方が実際のタスク分類結果との類似性が高くより優れていると考える．

5 実験結果

本章では，4 章で説明した方法で実験を行った結果について述べる．まず事前処理の有効性を検証するための実験結果について述べ，その後，手法 1 と手法 2 のどちらがより良い指標かについての実験結果について説明する．

なお，D1〜D15 はデータ提供者からのデータを表す．また，手法 1 と手法 2 はそれぞれ 3.3 節と 3.4 節で説明したものに対応する．コスト，純度はそれぞれ 4.2.1 節，4.2.2 節で述べた各指標を指す．

表 1　事前処理の有無ごとの手法 1 の各指標の値

データ	コスト 処理あり	コスト 処理なし	純度 処理あり	純度 処理なし	rand 処理あり	rand 処理なし
D1	**305**	318	0.319	**0.348**	0.0607	**0.0957**
D2	81.6	**68.7**	0.408	**0.495**	0.210	**0.341**
D3	**308**	343	**0.668**	0.548	**0.0204**	-0.0234
D4	59.8	**58.8**	**0.568**	0.361	0.280	**0.346**
D5	**5.50**	9.50	**0.633**	0.55	**0.167**	0.00
D6	**68.3**	72.8	**1.00**	0.883	0.148	**0.764**
D7	**776**	1060	0.736	**0.738**	0.00673	**0.00764**
D8	**145**	304	0.712	**0.786**	0.239	**0.334**
D9	**119**	160	**0.789**	0.652	0.388	**0.480**
D10	**555**	636	0.821	**0.866**	**0.0332**	0.0265
D11	**378**	411	0.750	**0.809**	0.0145	**0.0322**
D12	**167**	397	**0.710**	0.576	-0.0152	**-0.00438**
D13	**48.6**	61.5	**0.881**	0.756	**0.118**	0.0531
D14	**391**	467	**1.00**	0.798	**0.0908**	0.0627
D15	7.50	**6.50**	0.389	**0.417**	-0.176	**0.0367**
勝数	12	3	8	7	5	10

表 2　事前処理の有無ごとの手法 2 の各指標の値

データ	コスト 処理あり	コスト 処理なし	純度 処理あり	純度 処理なし	rand 処理あり	rand 処理なし
D1	450	**413**	**0.417**	0.125	**0.0529**	0.00
D2	**88.7**	100	**0.595**	0.553	0.210	**0.211**
D3	**171**	203	**0.307**	0.159	**0.0570**	0.0334
D4	**60.0**	87.5	0.143	**0.261**	**0.00**	-0.00990
D5	**7.50**	12.0	0.500	0.500	0.00	0.00
D6	**87.9**	138	**1.00**	0.586	0.0299	**0.0602**
D7	**560**	729	0.500	**0.501**	**0.467**	0.274
D8	**197**	254	**0.623**	0.575	0.122	**0.202**
D9	**110**	176	**0.620**	0.369	**0.190**	0.170
D10	**417**	448	**0.558**	0.250	-0.0102	**0.00**
D11	**252**	322	**0.201**	0.120	-0.0618	**0.00**
D12	**111**	177	**0.250**	0.209	**0.00**	-0.00835
D13	**43.8**	46.5	**0.467**	0.424	-0.0377	**-0.0137**
D14	**490**	514	**1.00**	0.711	**0.0195**	-0.0182
D15	**11.6**	18.0	1.00	1.00	**0.328**	0.00
勝数	14	1	11	2	8	6

5.1　事前処理の有効性

手法 1 と手法 2 のそれぞれについて，事前処理の有無ごとのコスト，純度，調整 rand 指数の値をそれぞれ表 1 と表 2 に示す．また，各指標について事前処理ありの場合となしの場合を比較したときにより良い方の値を太字にしている．4.2 節で述べた通り，コストは小さい方が良く，純度と調整 rand 指数は大きい方が良い．また，より良い値を示したデータの個数を勝数として記載した．

表 1 を見ると，コストと純度については 15 個のデータのうち半数以上が事前処理ありの場合により良い値を示している．一方，調整 rand 指数については半数以上が事前処理なしの場合により良い値を示している．また表 2 からは，3 つの指標全てについて，事前処理ありの場合により良い値を示すデータが半数以上あることが分かる．

事前処理の有無によって各指標がどのくらい変化するのかを確認するために，各データのコストと純度について，事前処理ありの場合となしの場合の値の比率を算出した．比率は，事前処理ありの場合となしの場合の値のうちより良い値を 1 とし，その場合の悪い方の値の比率を算出した．その上で，15 個のデータの比率の平均値を算出し比較した．手法 1 と手法 2 のそれぞれについてこの比率および平均値を算出した結果をそれぞれ表 3，表 4 に示す．表 3 および表 4 の平均値を見ると，いずれの表でも，コストと純度の両方で事前処理ありの場合の方が良い値を示している．また，調整 rand 指数については事前処理ありの場合となしの場合それぞれの値の平均を算出した．その結果を表 5 に示す．表 5 を見ると，手法 1 では事前処理なしの方がより良い値を示しているが，手法 2 では事前処理ありの方がより良い値を示している．

5.2　手法 1 と手法 2 の比較

事前処理なしの場合とありの場合のそれぞれについて，手法 1 と手法 2 の 3 つの指標の値をそれぞれ表 6 と表 7 に示す．5.1 節と同様に，各指標について手法 1 と手法 2 を比較したときにより良い方の値を太字にし，勝数も記載している．

表 3　手法 1 の事前処理あり・なしの場合のコスト・純度の値の比率とその平均値

データ	コスト 処理あり	コスト 処理なし	純度 処理あり	純度 処理なし
D1	1	1.04	0.917	1
D2	1.19	1	0.824	1
D3	1	1.11	1	0.820
D4	1.02	1	1	0.636
D5	1	1.73	1	0.869
D6	1	1.06	1	0.883
D7	1	1.37	0.997	1
D8	1	2.10	0.906	1
D9	1	1.34	1	0.826
D10	1	1.15	0.948	1
D11	1	1.09	0.927	1
D12	1	2.38	1	0.811
D13	1	1.27	1	0.858
D14	1	1.19	1	0.798
D15	1.15	1	0.933	1
平均	1.02	1.32	0.963	0.900

表 4　手法 2 の事前処理あり・なしの場合のコスト・純度の値の比率とその平均値

データ	コスト 処理あり	コスト 処理なし	純度 処理あり	純度 処理なし
D1	1.09	1	1	0.300
D2	1	1.13	1	0.929
D3	1	1.19	1	0.518
D4	1	1.46	0.548	1
D5	1	1.60	1	1
D6	1	1.57	1	0.586
D7	1	1.30	0.998	1
D8	1	1.29	1	0.923
D9	1	1.60	1	0.595
D10	1	1.07	1	0.448
D11	1	1.28	1	0.597
D12	1	1.59	1	0.836
D13	1	1.06	1	0.908
D14	1	1.05	1	0.711
D15	1	1.05	1	1
平均	1.01	1.28	0.970	0.757

表 5　調整 rand 指数の平均値

	事前処理あり	事前処理なし
手法 1	0.106	0.170
手法 2	0.0911	0.0600

表 6　手法 1 と手法 2 の各指標の比較（事前処理なし）

データ	コスト 手法 1	コスト 手法 2	純度 手法 1	純度 手法 2	rand 手法 1	rand 手法 2
D1	**318**	413	**0.348**	0.125	**0.0957**	0.00
D2	**68.7**	100	0.495	**0.553**	**0.341**	0.211
D3	343	**203**	**0.548**	0.159	-0.0234	**0.0334**
D4	**58.8**	87.5	**0.361**	0.261	**0.346**	-0.00990
D5	**9.5**	12.0	**0.550**	0.500	0.00	0.00
D6	**72.8**	138	**0.883**	0.586	**0.764**	0.0602
D7	1060	**729**	**0.738**	0.501	0.00764	**0.274**
D8	304	**254**	**0.786**	0.575	**0.334**	0.202
D9	**160**	176	**0.652**	0.369	**0.480**	0.170
D10	636	**448**	**0.866**	0.250	**0.0265**	0.00
D11	411	**322**	**0.809**	0.120	**0.0322**	0.00
D12	397	**177**	**0.576**	0.209	**-0.00438**	-0.00835
D13	61.5	**46.5**	**0.756**	0.424	**0.0531**	-0.0137
D14	**467**	514	**0.798**	0.711	**0.0627**	-0.0182
D15	**6.50**	18.0	0.417	**1.00**	**0.0367**	0.00
勝数	8	7	13	2	12	2

表 7　手法 1 と手法 2 の各指標の比較（事前処理あり）

データ	コスト 手法 1	コスト 手法 2	純度 手法 1	純度 手法 2	rand 手法 1	rand 手法 2
D1	**305**	450	0.319	**0.417**	**0.0607**	0.0529
D2	**81.6**	88.7	0.408	**0.595**	0.210	0.210
D3	308	**171**	**0.668**	0.307	0.0204	**0.0570**
D4	**59.8**	60.0	**0.568**	0.143	**0.280**	0.00
D5	**5.50**	7.50	**0.633**	0.500	**0.167**	0.00
D6	**68.3**	87.9	1.00	1.00	**0.148**	0.0299
D7	776	**560**	**0.736**	0.500	0.00673	**0.467**
D8	**145**	197	**0.712**	0.623	**0.239**	0.122
D9	119	**110**	**0.789**	0.620	**0.388**	0.190
D10	555	**417**	**0.821**	0.558	**0.0332**	-0.0102
D11	378	**252**	**0.750**	0.201	**0.0145**	-0.0618
D12	167	**111**	**0.710**	0.250	-0.0152	**0.00**
D13	48.6	**43.8**	**0.881**	0.467	**0.118**	-0.0377
D14	**391**	490	1.00	1.00	**0.0908**	0.0195
D15	**7.50**	11.6	0.389	**1.00**	-0.176	**0.328**
勝数	8	7	10	5	10	4

表 6 と表 7 のいずれにおいても，3 つの指標の全てで，15 個のデータのうち半数以上が手法 1 を用いた場合により良い値を示している．

5.1 節と同様に，手法 1 と手法 2 で各指標がどのくらい変化するのかを確認するために，各データのコストと純度について，手法 1 と手法 2 の値の比率を算出した．比率は，こちらも手法 1 と手法 2 の値のうちより良い値を 1 とし，その場合の悪い方の値の比率を算出した．その上で，15 個のデータの比率の平均値を算出し比較した．事前処理ありの場合と事前処理なしの場合のそれぞれについてこの比率および平均値を算出した結果をそれぞれ表 8，表 9 に示す．表 8 の平均値を見ると，コスト，純度のいずれも手法 1 の方が良い値を示している．表 9 については，コストについては僅差ではあるものの手法 2 の方が良い値を示しており，純度については手法 1 の方が良い値を示している．また，調整 rand 指数については表 5 に示した平均値を参照する．手法 1 と手法 2 で比較すると，事前処理ありの場合となしの場合のいずれにおいても手法 1 がより良い値を示している．

表 8　手法 1 と手法 2 のコスト・純度の値の比率とその平均値（事前処理なし）

データ	コスト 手法 1	コスト 手法 2	純度 手法 1	純度 手法 2
D1	1	1.30	1	0.359
D2	1	1.46	0.895	1
D3	1.69	1	1	0.290
D4	1	1.49	1	0.723
D5	1	1.26	1	0.909
D6	1	1.90	1	0.664
D7	1.45	1	1	0.679
D8	1.20	1	1	0.732
D9	1	1.10	1	0.566
D10	1.42	1	1	0.289
D11	1.28	1	1	0.148
D12	2.24	1	1	0.363
D13	1.32	1	1	0.561
D14	1	1.10	1	0.891
D15	1	2.77	0.417	1
平均	1.24	1.29	0.954	0.612

表 9　手法 1 と手法 2 のコスト・純度の値の比率とその平均値（事前処理あり）

データ	コスト 手法 1	コスト 手法 2	純度 手法 1	純度 手法 2
D1	1	1.48	0.765	1
D2	1	1.09	0.686	1
D3	1.80	1	1	0.460
D4	1	1.00	1	0.252
D5	1	1.36	1	0.790
D6	1	1.29	1	1
D7	1.39	1	1	0.679
D8	1	1.36	1	0.875
D9	1.08	1	1	0.786
D10	1.33	1	1	0.680
D11	1.50	1	1	0.268
D12	1.50	1	1	0.352
D13	1.11	1	1	0.530
D14	1	1.25	1	1
D15	1	1.55	0.389	1
平均	1.18	1.16	0.923	0.711

6　考　　察

6.1　手法 1 のエッジ削除基準の値の変化による各指標の変化

手法 1 では，コミュニティ抽出を適用するグラフを作成する際，ノード間にエッジを作成しない基準として履歴のペア間の経過時間 5 分，すなわち 300 秒という値をおいていた．しかし，この基準値に明確な根拠は特にない．そこで，この値を 10 秒から 1800 秒まで 10 秒単位で変化させた時の各指標の値をグラフにし，どのように変化するのかおよびどのような時に最も良い値を取るのかを確認した．なお，今回は事前処理ありの場合のグラフを 15 個のデータそれぞれについて作成した．図 2〜7 にそのグラフの一部を示す．横軸が基準値である経過時間，縦軸が各指標である．

まずコストのグラフでは，どのデータにおいても 50 200 秒付近で最小値または極小値を取ることを確認した．図 2〜3 でもそのことが確認できる．

次に純度のグラフでは，経過時間が短いほど大きくなるという傾向は見られたが，それ以外の特徴についてはデータによっ

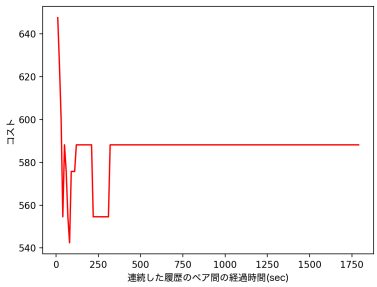


図 2　エッジ削除基準値を変化させた時のコストの値の変化 1

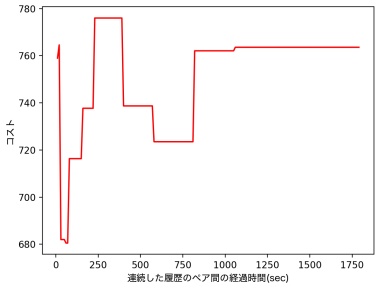


図 3　エッジ削除基準値を変化させた時のコストの値の変化 2

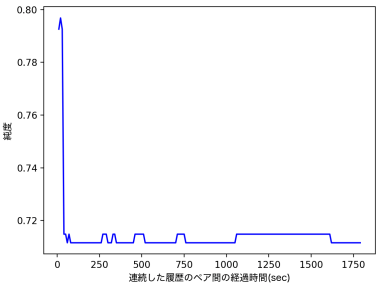


図 4　エッジ削除基準値を変化させた時の純度の値の変化 1

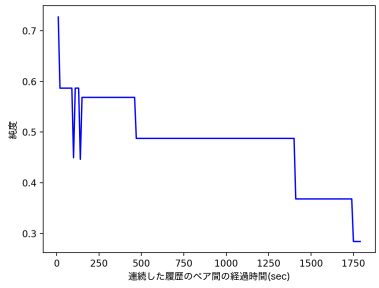


図 5　エッジ削除基準値を変化させた時の純度の値の変化 2

て様々で，全体に共通した特徴はあまり確認できなかった．例として図 4 と図 5 に純度のグラフを示す．

最後に調整 rand 指数のグラフでは，全てではないものの，複数のデータにおいて 50 200 秒付近で極大値を取ることが確認できた．50 200 秒付近での極大値が確認できたデータを図 6 〜7 に示す．

6.2　ランダムフォレストの精度について

次に手法 2 で用いたランダムフォレストの分類器の精度について考察する．事前処理ありの場合となしの場合それぞれの分類器の精度を表 10 に示す．なお，この精度は本人以外のデータ

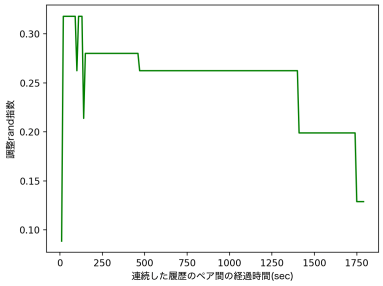


図 6　エッジ削除基準値を変化させた時の調整 rand 指数の値の変化 1

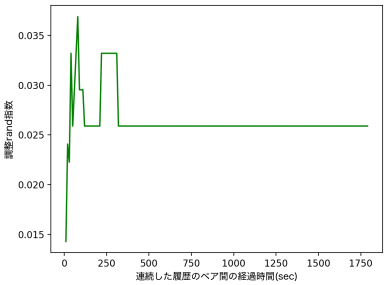


図 7　エッジ削除基準値を変化させた時の調整 rand 指数の値の変化 2

で学習させた分類器に対して各データを適用した際の精度である．事前処理ありの場合となしの場合で比較すると，半数以上のデータで事前処理なしの場合の方が精度が高い．5.1 節の実験結果では 3 つの指標全てにおいて事前処理ありの場合の方が良い値を示していたが，表 10 の内容を考慮すると，事前処理によって改善したのは分類器自体の性能ではないと考えられる．

表 10　ランダムフォレストの分類器の精度について

データ	事前処理あり	事前処理なし
D1	**0.634**	0.624
D2	0.441	**0.683**
D3	0.523	**0.551**
D4	0.486	**0.596**
D5	0.556	**0.600**
D6	**0.250**	0.192
D7	**0.850**	0.797
D8	**0.663**	0.580
D9	**0.828**	0.577
D10	0.655	**0.805**
D11	**0.739**	0.667
D12	0.353	**0.419**
D13	0.429	**0.444**
D14	0.436	**0.606**
D15	**0.667**	0.600

また，ランダムフォレストの分類器の精度について，偽陽性と偽陰性のどちらが多いかを調査するため，事前処理なし・ありの場合それぞれについて混合行列の値を算出した．それぞれ表 11，表 12 に示す．なお，ここでは「タスクの切り替わりであること」と陽性としている．つまり，偽陽性はタスクの切り替わりではないのに切り替わりだと判定された場合を指し，偽陰性はタスクの切り替わりなのに切り替わりでないと判定された場合を指す．表 11 と表 12 のいずれにおいても，偽陽性と偽陰性のどちらが多いかはデータによってばらつきがあることが分かり，精度を上げるためには個人の特性を考慮できるようなパラメータを追加するべきだと考えられる．

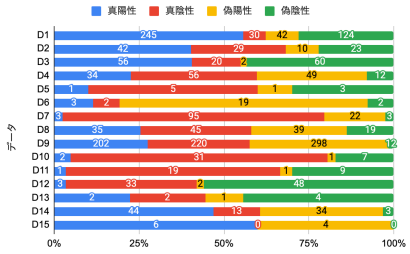


図 8　混合行列の値の帯グラフ（事前処理なし）

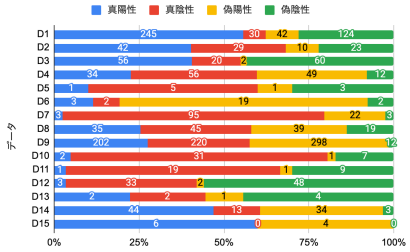


図 9　混合行列の値の帯グラフ（事前処理あり）

表 11　ランダムフォレストの各特徴量の重要度（事前処理なし）

提供者	特徴量 1	特徴量 2	特徴量 3	特徴量 4	特徴量 5	特徴量 6	特徴量 7
P1	0.124	0.195	0.129	0.00127	0.0677	0.222	0.261
P2	0.124	0.242	0.119	0.000668	0.0900	0.178	0.245
P3	0.118	0.237	0.115	0.00114	0.0943	0.192	0.243
P4	0.120	0.203	0.117	0.000798	0.122	0.204	0.232
P5	0.129	0.321	0.148	0.00141	0.0957	0.123	0.182
P6	0.113	0.216	0.114	0.000879	0.101	0.191	0.264
P7	0.113	0.225	0.116	0.000982	0.0864	0.191	0.266
P8	0.114	0.222	0.113	0.000883	0.101	0.181	0.268

表 12　ランダムフォレストの各特徴量の重要度（事前処理あり）

提供者	特徴量 1	特徴量 2	特徴量 3	特徴量 4	特徴量 5	特徴量 6	特徴量 7
P1	0.144	0.205	0.150	0.00354	0.0363	0.173	0.288
P2	0.141	0.229	0.154	0.00416	0.0399	0.169	0.263
P3	0.132	0.254	0.147	0.00338	0.0407	0.141	0.282
P4	0.140	0.210	0.154	0.00289	0.0447	0.179	0.271
P5	0.153	0.293	0.184	0.00563	0.0506	0.0669	0.247
P6	0.134	0.208	0.151	0.00379	0.0390	0.172	0.292
P7	0.135	0.211	0.153	0.00332	0.0378	0.156	0.304
P8	0.136	0.223	0.150	0.00278	0.0398	0.160	0.289

6.3　ランダムフォレストの各特徴量の重要度

最後に，ランダムフォレストの分類器において，履歴のペア間の特徴量のうちのどれが重要度が高いのかを確認するために，事前処理あり・なしの場合それぞれについて，各特徴量の値をそれぞれ表 13 と表 14 にまとめた．なお，P1〜P8 は 8 人のデータ提供者それぞれを表す．また，各特徴量の番号は節の特徴量の番号に対応する．

表 13 と表 14 を見ると，いずれにおいても，タブを開いた時刻の差に関連する特徴量 (特徴量 1〜3) や過去の起こった遷移の回数に関連する特徴量 (特徴量 6，7) の重要度が比較的高く，URL が同じかどうか (特徴量 4) やドメインが同じかどうか (特徴量 5) についての特徴量の重要度は低いことが分かる．

文　献

[1] M. Kulkarni and K. Kapoor and D.S.V Madala and S. Bansal and S. Hangal, “Compartmentalizing web browsing with Sailboat,“ Proceedings of the 10th Indian Conference on Human-Computer Interaction, pp. 1–8, 2019.

[2] R. Ma and H. Lassila and L. Nurgalieva and J. Lindqvist, “When Browsing Gets Cluttered: Exploring and Modeling Interactions of Browsing Clutter, Browsing Habits, and Coping,“ Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–29, 2023.

[3] J.C. Chang and Y. Kim and V. Miller and M.X. Liu and B.A. Myers and A. Kittur, “Tabs.do: Task-Centric Browser Tab Management,“ The 34th Annual ACM Symposium on User Interface Software and Technology, pp. 663–676, 2021.

[4] M.R. Rajamanickam and R. MacKenzie and B. Lam and T. Su, “A task-focused approach to support sharing and interruption recovery in web browsers,“ CHI ’10 Extended Abstracts on Human Factors in Computing Systems, pp. 4345–4350, 2010.

[5] M.E.J. Newman and Kristina M. Girvan, “Finding and evaluating community structure in networks,“ Phys. Rev. E Vol. 69, 2004.

[6] V.D. Blondel and J.L. Guillaume and R. Lambiotte and E. Lefebvre, “Fast unfolding of communities in large networks,“ Journal of Statistical Mechanics: Theory and Experiment, 2008.

[7] N. Dugué and A. Perez, “Direction matters in complex networks: A theoretical and applied study for greedy modularity optimization,“ Physica A: Statistical Mechanics and its Applications Vol 603, 2022.

[8] Rand, William M, “Objective Criteria for the Evaluation of Clustering Methods,“ Journal of the American Statistical Association vol. 66, pp. 846–50, 1971.

[9] Hubert, L and Arabie, P, “Comparing partitions,“ Journal of Classification 2, pp. 193–218 1985.

[10] Acid Tabs, https://chrome.google.com/webstore/detail/acid-tabs/hgceopemmcmigbmhphbcgkeffommpjfc?hl=ja.

[11] Auto-Group Tabs, https://chrome.google.com/webstore/detail/auto-group-tabs/danncghahncanipdoajmakdbeaophenb?hl=ja.

[12] OneTab, https://chrome.google.com/webstore/detail/onetab/chphlpgkkbolifaimnlloiipkdnihall?hl=ja.

[13] Tabs Outliner, https://chrome.google.com/webstore/detail/tabs-outliner/eggkanocgddhmamlbiijnphhppkpkmkl?hl=ja.

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

2024年2月28日(水) 16:00 ~ 18:10 | T5-A オンライン（Zoom Events)

教育・学習 (3)

座長:木村 昌臣(芝浦工業大学)
コメンテータ:橋本 敦史(オムロンサイニックエックス)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T5-A-3-01] 生成AIを用いた教え合い学習のための間違いを模倣するAgent

*森部 想水[1]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府、2. 九州大学大学院芸術工学研究院)

16:20 ~ 16:45
[T5-A-3-02] ピクタグラミング −ピクトグラムの作成を通じたマークアップ言語学習アプリケーションの開発とその評価−

*木下 倭[1]、高橋 伶奈[1]、伊藤 一成[1] (1. 青山学院大学 社会情報学部 伊藤一成研究室)

16:45 ~ 17:10
[T5-A-3-03] ピクトグラムを用いた教育向けネットワークシミュレータ「Pictocol Simulator」の開発

*前田 祐杜[1]、伊藤 一成[2] (1. 青山学院大学社会情報学部　伊藤一成研究室、2. 青山学院大学社会情報学部)

17:10 ~ 17:35
[T5-A-3-04] 図形への書き込みと手書き解答文字の自動認識を用いた数学学習者の解答戦略推定

*林 凱博[1]、宮﨑 公彦[2]、工藤 雅士[2]、山名 早人[3] (1. 早稲田大学 基幹理工学部、2. 早稲田大学 基幹理工学研究科、3. 早稲田大学 理工学術院)

生成 AI を用いた教え合い学習のための間違いを模倣する Agent

森部　想水†　　牛尼　剛聡††

† 九州大学大学院芸術工学府 〒 815-8540 福岡県福岡市南区塩原 4 丁目 9-1
†† 九州大学大学院芸術工学研究院 〒 815-8540 福岡県福岡市南区塩原 4 丁目 9-1
E-mail: †moribe.sosui.695@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年，アクティブラーニングをはじめとして発信力のある人材育成のために教え合い学習が注目されている．様々な研究によってその有効性が確認されているが，人間同士の教え合いは必ずしも常に有効に実施できるとは限らない．そこで本研究ではいつでもどこでも教え合いを実現する学習仲間としての Agent の開発することを目的とする．効果的な教え合い学習のためには学習者と同程度の習熟度の学習仲間を Agent として実現する必要がある．そのために本研究では，具体例として英作文の誤り再現に焦点を当て，(1) 学習目標 (ルーブリック) からの逆算をさせる手法と，(2) 文法誤り訂正で用いられる Prompt Insertion に着想を得た手法という 2 つの手法を実施し，その有効性を検証する．

キーワード　教え合い学習，Peer tutoring, 文法誤り訂正，GEC，Prompt Insertion

1　はじめに

教育の分野においては、発信力のある人材育成のために学習者の能動的な参加を促すアクティブラーニングの重要性が高まっている．特に教え合い学習はその有効性が多くの研究により確認されており，実際の教育現場においてもグループディスカッションやペア学習といった形態で積極的に導入されている．しかし，人間同士の教え合いは必ずしも常に有効に実施できるとは限らない．そこで本研究ではいつでもどこでも教え合いを実現する学習仲間としての Agent の開発をすることを目的とする．効果的な教え合い学習のためには学習者と同等の習熟度を持つ学習仲間の存在が不可欠であると考えられる．本研究では，同等の習熟度であれば同様の間違いを行うと仮定し，このような学習仲間を提供するために典型的な誤りを模倣する Agent を開発する (図 1)．本研究では具体例として英作文の誤り再現に焦点を当て，学習目標 (ルーブリック) からの逆算をさせる手法と文法誤り訂正で用いられる Prompt Insertion に着想を得た手法の 2 つの手法により，具体的かつ個別化された誤り模倣を実施し，その有効性を検証する．

2　関連研究

2.1　教え合い学習 (Peer Tutoring)

人間同士の教え合い学習の効果を示した研究は数多く存在する．特に Abdelkarim ら [1] や Gubalani ら [2] は数学と英語読解の分野において教え合い学習の有効性を示している．教え合い学習をすることによって，自分が学んだことを他人に説明することで言語化し定着を図るのが教え合い学習の利点である．本手法のアーキテクチャを以下の図 1 に示す。

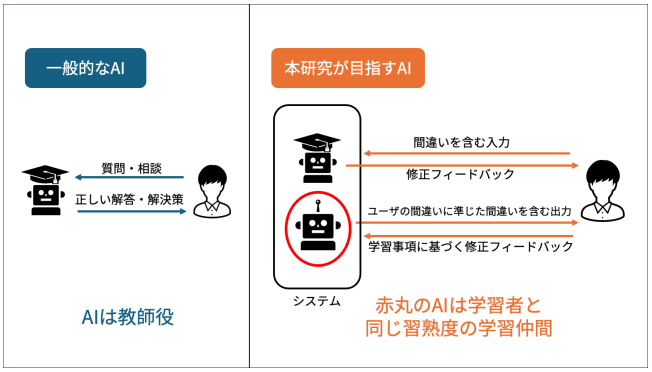


図 1　一般的な AI と本研究の目指す AI の違い

2.2　教え合い学習の問題点

前節のように人間同士の教え合いには様々な科目で学習効果が確認されているが，その恒常的な実施を妨げる問題がある．具体的には (i) 時間的制約 (スケジュールが合わないなど)(ii) 空間的制約 (同じ場所に集合できるとも限らない)(iii) 人前で間違うのが恥ずかしい (iv) 学習者同士の習熟度に乖離がある場合が挙げられる．特に (iv) に関して，Zou[5] は認知論の観点から，子供たちはメタ認知能力に欠けている場合があり，自分がどのくらいその課題に関して理解しているかを自身で把握することが難しいことがあるため，ピア・チューターとして教え合い学習を実施するのに不適であるケースに注意すべきであることを示唆している．対人に特有な問題を避ける解決策として，Agent を用いた学習も多方面で導入が見られる．

2.3　Agent を用いた学習

Agent とのコミュニケーションを通して学習ができるようなシステムとして Matsuda ら [4] の SimStudent が挙げられる．SimStudent は数学の方程式の解法を学習者が Agent に教えるシステムである．SimStudent を通した学習では，Agent が方

程式を解き，その解き方が合っているか間違っているかを学習者が判断して間違っている場合はその部分を指摘して正しい解法を提案する．学習者が方程式の解き方を間違って理解している場合は Agent を正しい解答に導くことができないため，SimStudent ではこの手順を踏むことによって，学習者自身の理解が細部まで正しく身についているかを確認している．

2.4 本研究の位置付け

本研究はいつでもどこでも教え合い学習を可能にすることを最終的な目標に据えている．本研究では具体的な課題として英作文を用いる．英作文は (i) 模範解答がなく解答が一つに定まらない (ii) 多様な種類の間違いが起こりうる (iii) 複雑な知識の融合が問われるためルールベースではシステム化が難しいと考えられる．本研究では LLM を用いることにで上記の条件下において教え合いを実現するための仕組みを提案する．提案にあたって，効果的な教え合い学習には学習者と同程度の習熟度の学習仲間が必要であることは 2.2 節で述べた通りである．同程度の習熟度とは，ある学習課題に関して学習者ペアの両方が同じ程度の知識と技能を有しており一方が教えるだけ/教わるだけの関係にならない状態のことを指す．また，本研究では同等の習熟度の学習者は似通った間違いを行うと仮定する．よって，Agent を「教わる側」として教え合い学習に導入するためには，学習者と同程度の習熟度，つまり学習者がするものと同じような間違いを出力できる機構が必要であると考えられる．

3 提 案 手 法

本研究では，英作文において学習者がしそうな間違いを再現する機構を実現するために (i) ルーブリックを用いてユーザの英作文のレベルを推定し間違いの度合いを再現する (ii) 文法誤り訂正 (GEC) で有効性が提案されている Prompt Insertion に着想を得た手法の 2 つの手法を提案する．以下にその詳細を述べる．

3.1 ルーブリックベースの手法

ルーブリックとは，ある学習科目に関して講義の指導者または資格試験の運営者が学習目標を習熟度ごとに明文化したものである．本手法では，国際的な英語の資格試験の一つである IELTS のルーブリックを参考に，入力された英作文の習熟度を分類するためのリストを作成した．IELTS と同様に留学審査で用いられる TOEFL や TOEIC を運営する ETS も人間の採点者とともに自動採点を採用している．本研究ではこのような背景から，LLM を活用したシステムによって学習者の英文をレベル分けし，得られたレベルから類似した特徴を持った英文を生成することを目標とする．本手法の手順は以下の通りである。
(A) ルーブリックベースの手法の手順
(i)LLM で学習者の入力文をルーブリックにおいてどのレベルに属するか推定する．
(ii)1 文あたりの単語数を最小〜最大まで記録する．
(iii)(i) で推定したレベルを基に，LLM にその習熟度と推定される架空の人物になりきって題に答えてもらう．

本手法のアーキテクチャを以下の図 2 に示す．

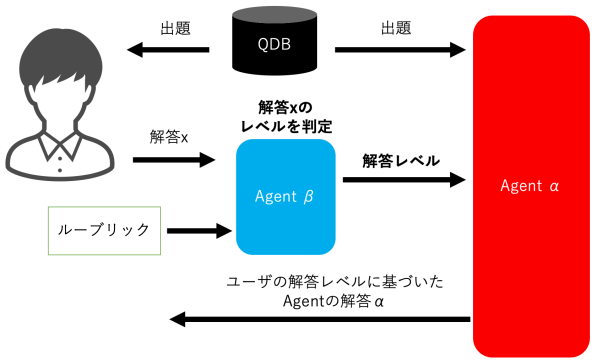


図 2 ルーブリックベース手法のアーキテクチャ

以下に本手法で用いたプロンプトを示す．

(I-1) ルーブリックに基づきユーザの英作文のレベル推定を実施するプロンプト
Evaluate the level of Introduction according to the rubric, and return the number of the level you estimated. The output must include only the number.

(I-2) ユーザの英作文の各文の長さを数えるプロンプト
Count the number of words included in the introduction essay for each sentence. Output the range of the length of the sentences. For example, if the essay has sentences that contain a minimum of 5 words to a maximum of 9 words, the output should be ”5 〜 9”. Do not contain except the range, and only include the ”minimum number 〜 maximum number”.

(I-3) 推定したレベルに基づいて Agent に英作文を書いてもらうプロンプト
Please set up a fictitious profile, and pretend to be that person. Introduce yourself at the designated level in the rubric. 6 to 8 sentences maximum. Each sentence should be written within a specified word count and should be randomized within that range. Note that the output sentence’s English level should follow the indicated level in the rubric: For example, if the specified level is 3, then the text must have the characteristics of the score of 3 as described in the rubric. It mustn’t contain any greetings or exclamatory sentences such as ”Hello” or ”Oh,” and so on. The output should contain only the essay itself.

3.2 Prompt Insertion ベースの手法

文のレベル分けはカテゴリ化の一種と考えられるが，実際の

学習者の入力文を基にしたカテゴリ分類から学習者がしそうな解答を逆算する方法には問題がある．具体例として，学習者がしうる間違いの数や種類・傾向が多様であるのに対して，レベル (習熟度) は文の性質の一側面に過ぎず，そのような抽象的な層から各人の多種多様な文に寄せるような出力が可能なのかという指摘がある．そこで本手法では間違いそのものを基にしてその学習者特有の英作文を再現する．我々は，Kanekoら [3] の Prompt Insertion に着目する．Prompt Insertion とは Grammatical Error Collection(文法誤り訂正，GEC) において LLM による誤り訂正の後，訂正点を自動的に抽出し，それらをプロンプトとして LLM の出力に挿入する手法のことである．これを基に本研究では以下のフローを提案する．

(B)Prompt Insertion ベースの手法の手順
(i) 学習者の入力文を LLM で校正する．(校正 Agent)
(ii)LLM で校正前後の差分を”校正前”→”校正後”の形で抽出し，リスト化させる．(差分 Agent)
(iii) 作成したリストの各項目について，LLM でそれがどういった誤りなのかを記述させ，その誤りの種類ごとに数えさせる．(明文化 Agent)
(iv) 数えた分だけその種類の間違いを新たに LLM に書かせた作文に入れ込むように指示する．(練り込み Agent)

本手法のアーキテクチャを以下の図 3 に示す．

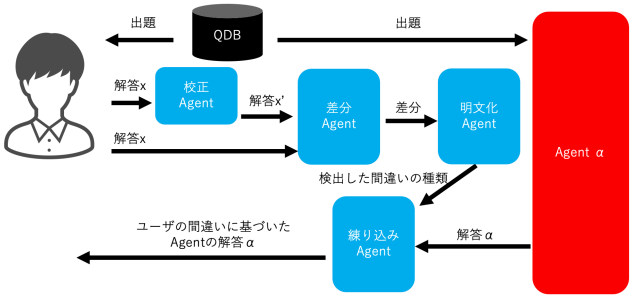


図 3　Prompt Insertion ベース手法のアーキテクチャ

以下に用いたプロンプトを示す．
(II-1) 校正 Agent のプロンプト
Correct any grammatical or structural errors in the sentences.

(II-2) 差分 Agent のプロンプト
List the changes made in the corrected essay compared to the original, in the format ’original text’ → ’corrected text’. Each item in the list should be extracted only for the changed portion and not the entire sentence.
In other words, there should be no multiple changes within a single list item.
For example, the list should look like
”Me” → ”My”, ”of” → ”for”, ”apples” → ”an apple”, etc.

(II-3) 明文化 Agent のプロンプト
The list of changes is a list of the areas that need to be corrected for the text.
Categorize each correction by grammatical element and count how many errors were made.
For example, the list should look like
singular/plural errors: 3, to-infinitive when it should be a verbal noun: 2, tense agreement: 1, incorrect articles: 3, typo:4, etc.
At the end of the list, the total number of corrections should be given, e.g., ”Total: 8”.

(II-4) 練り込み Agent のプロンプト
In the given essay, mix the specified number of errors of the specified type into the given essay to produce a grammatically incorrect essay.
Make sure that the number of grammatical errors in the final output sentence is equal to the total number of errors listed at the end of the list of specification errors to be inserted.
For example, if the list says the Total correction number is 15, your output must contain 15 grammatical errors.

4　評 価 実 験

本研究では 2 つの評価実験を行った．1 つ目の実験では 3 章で提案した 2 つの手法に関して基本的な誤りが複数含まれているサンプルの英文を用いてどちらがより誤りを再現できるかを検討した．2 つ目の実験では，1 つ目の実験でより精度が高いと思われた方の手法を用いて複数のエッセイに対し対照実験を行った．なお，モデルは gpt-4-1106-preview，Agent の管理には LangChain を用いた．間違いが再現されているかの確認手段として，Grammarly[1] を用いて修正提案の種類と個数を確認した．

4.1　手法の検討のための評価実験

4.1.1　用いる文章

本研究では，間違いを再現する対象とする文章として以下を用いる．

(ア) 間違いを模倣する対象の例文
Me name is Brian.
I am belonging to a Harvard Universities.
I major with design of architecture.
I respect is Frank Lloyd Wright, by whom is famous of its outstanding building such as “ Fallingwater ”.

1：https://www.grammarly.com/

He was also created Teikoku Hotel.
I hope I can stay their.
My dreams are creating interior of good design as him.

本例文はある学生の自己紹介文という設定で，主語の格の間違いや単数/複数の間違い，時制の間違い，of と for の使い分けなどの初歩的なよくある間違いを網羅的に入れ込んだものである．全文の合計単語数は 56 語で，各文の平均語数は 8 語である．

4.1.2 ルーブリックベースの手法

ルーブリックベースの手法で得られた文章を以下に示す．

I am Mai, an avid reader with a fascination for historical novels and the preservation of time-honored traditions.
Completing my master's thesis on the influence of Renaissance art on modern design, I spend countless hours analyzing paintings and architectural marvels.
Being an environmental advocate, I volunteer at local parks, assisting in maintaining their natural elegance.
Cycling through the countryside is my chosen method of unwinding, appreciating the simplicity of rural life.
While managing a blog dedicated to sustainable living, I connect with a global community sharing eco-friendly practices.
I am also honing my skills in traditional Japanese calligraphy, finding solace in the brush strokes that connect me to my heritage.
My future endeavors include curating a gallery exhibit that bridges the gap between historical artistry and contemporary environmentalism.

Grammarly によって提案された修正は，4 文目”unwinding, appreciating the simplicity”となっているところを，”unwinding, and appreciating the simplicity”とするべきとされた一件のみだった．

出力された文章には avid, advocate, hone skill, dedicate, heritage, endeavor,curate など比較的難易度が高く，日本においては高年次に学習するような語句が複数見受けられた．また，全文は 127 単語で構成されており，各文に含まれる単語数は平均約 18.14 語だった．模倣対象の文章の平均語数は 8 語だったため，2 倍以上の長さになっている．

4.1.3 Prompt Insertion ベースの手法

Prompt Insertion ベースの手法で得られた文章を以下に示す．

My name is Troy, me am a student fascinated by science and the stars.
At schools, my favorite subject is mathematics because it's like a puzzle.
I'm twelve years old.
I love playing chess.
When I grow up, I dream of becoming an astronaut and exploring distant galaxy. My best friend is my dog, Max, who are always by my side.
Even though science homework can being tough, I enjoy the challenge it present.
Reading fantasy novels under the tree in our backyard is my ideal way to relaxing.

Grammarly では 7 箇所修正提案があった。1 文目の me は I，2 文目の schools は school,5 文目の galaxy は galaxies,6 文目の are は is,7 文目の being は be,present は presents,8 文目の relaxing は relax に変更するべきである．主語の格の間違い，単数/複数の間違い，動詞の原形/進行形の間違いが模倣できていることが分かる．

特に高年次に学習し，原文に用いられている単語と大きく難易度が異なる語彙は見受けられなかった。また，全文は計 89 語で各文章は平均約 11.13 語だった．

4.1.4 考　　察

ルーブリックベースの手法の結果から，LLM は入力された文章をルーブリックを元にレベル分けすることが可能であることが示唆された．その一方で，各レベルに属する人間の学習者がどのような文章を書きがちであるかという抽象的な概念を有していないため推定されたレベルから英作文に含まれる間違いの模倣がうまくいかなかったと推察される．Prompt Insertion ベースの手法について，前節で述べたように主語の格の間違い，単数/複数の間違い，動詞の原形/進行形の間違いが模倣できた一方で，今回の例では冠詞の間違い・whom/who の間違い，前置詞の間違い，類似発音の単語に間違ってしまう間違いについては再現が見られなかった．各文の語長に関しては前節の通り手法 2 によってより原文に近い語数が得られた．総合すると，Prompt Insertion ベースの手法がより原文の間違いを模倣できていると言える．

4.2 Prompt Insertion ベース手法の有効性

上記の実験を踏まえ，Prompt Insertion を用いた手法の性能をさらに調べるために検証実験を行った。4 年制大学に在学中または卒業したことのある日本人 5 名を対象とし，4 つの簡単なテーマでエッセイを書いてもらい，計 20 本のエッセイデータを集めた．その後全てのエッセイに関し Grammarly を通して (a) 文法性に大きく関与する基本的な間違い (Basic)，(b) 発展的な修正提案 (Advanced)，(c) 質的な指標として Grammarly の提示するエッセイのスコア (Score) を記録した．その後全てのエッセイに対し (実験手法)Prompt Insertion により誤りを再現したエッセイ，(比較手法) 渡したエッセイを書く英語力であることを仮定して新たにエッセイを書くよう指示して作らせたエッセイを制作した．以下に本実験で用いた問題と比較手法で用いたプロンプトを示す。

(I) 用いたテーマ
a. 行ってみたい国や地域について紹介してください.
b. 好きな映画を紹介してください.
c. アウトドアアクティビティが好きか，インドアで遊ぶのが好きか教えてください.
d. これまで受けた中で一番印象に残った授業を紹介してください.

(II) 用いたプロンプト
Assuming that you are the person who has English skills and would write the given statement, answer the question in English as if you were that fictional character. (Pretend to be another different person, do not refer to the original sentence) Question: 入力する問題

4.2.1 結　　果

本節ではこの実験で得られた結果を記載する。各手法によって得られたエッセイの中に含まれる，Grammarly によって検出された誤りの個数とエッセイスコアの平均を以下の図 4 と図 5 に示す．なお，Basic + Advanced = Total である．

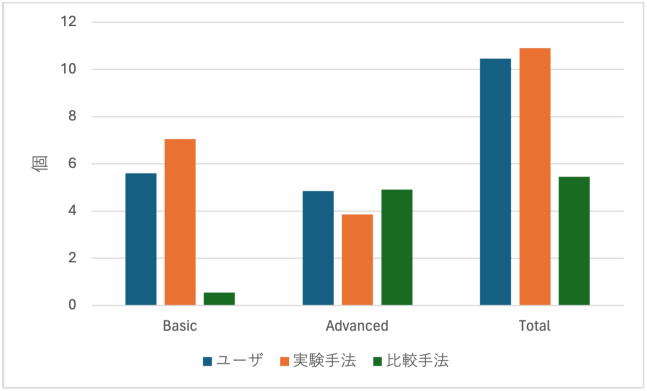


図 4　エッセイに含まれる誤り個数の平均値

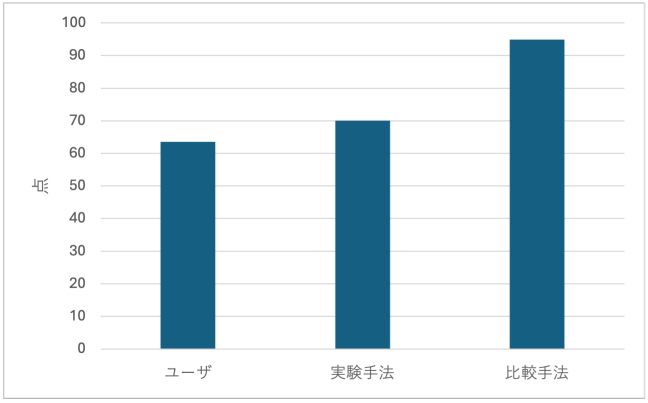


図 5　Grammarly Score の平均値

発展的な修正の個数 (図 4 の Advanced) については実験手法と比較手法の間で大きな差は見られなかった．基本的な誤り (図 4 の Basic) では比較手法が個数を再現できていないのに対してユーザの文と実験手法が近い誤り個数を有しており，それが誤り個数の合計 (図 4 の Total) に影響を与えていることが分かる．また，質的な指標 (図 5) についても，平均値の観点では実験手法の方が比較手法よりもユーザの文の質に近いことが示されている．この結果を踏まえ，エッセイごとに，実験手法とユーザの文，比較手法とユーザの文について各項目の差の絶対値を取りその平均を算出した．その結果を以下の図 6 と図 7 に示す．

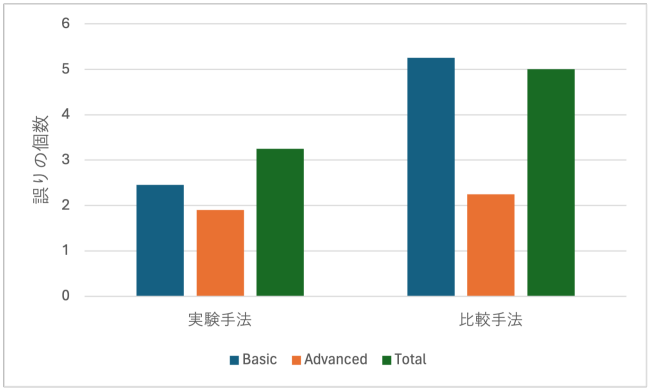


図 6　誤り個数の差の絶対値

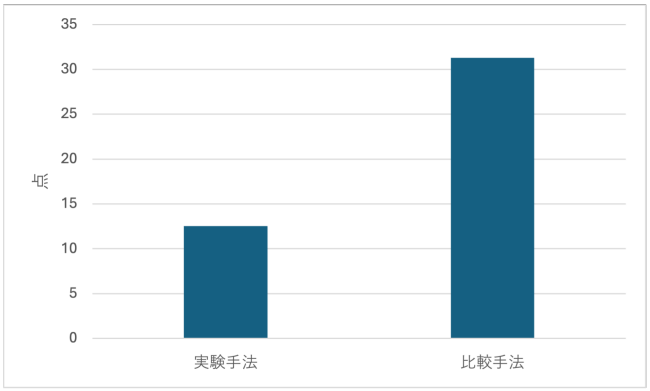


図 7　Grammarly Score の差の絶対値

差の絶対値の観点では，基本的な誤りの個数 (Basic) と質的な指標 (Grammarly Score) に関して実験手法は比較手法のおおよそ半分程度の誤差に抑えられている．続いて，実験手法と比較手法の各項目について，素の値と差の絶対値について t 検定を行った結果を以下の表 1 に示す.

4.2.2 考　　察

表 3 より，実験のスケールでは Prompt Insertion を用いた実験手法は比較手法に対して基本的な誤り (Basic) の個数，誤りの合計数，質的な指標である Grammarly の Score に関して有意差が得られた．特に Basic と Score に関しては素の値と差

表 1　t 検定の結果

検定項目	p 値 (素の値)	p 値 (差の絶対値)
Basic	0.0000000075	0.0041446770
Advanced	0.0504664270	0.2802144510
Total	0.0000003300	0.0484816100
Score	0.0000000045	0.0000197320

の絶対値の両方で強い有意差を確認した．発展的な誤り個数 (Advanced) の再現について，素の値に関しては有意水準に近い値が得られたため，サンプル数によっては更なる検討が必要である．

5 ま と め

本研究では Grammarly で検出できるような，非文法的な文になってしまう明らかな間違いを含んだ文を対象に生成 AI で模倣できるかの実験を行った．しかし，英作文において起きうる間違いの中には更に高度な間違いも存在し，それらに関しては別途対応が必要である．

(A)Sarah unwisely was praising her children too highly.
(B)Sarah was unwisely praising her children too highly.
は両方とも文法的な文である。よってこれらの違いは Grammarly で修正が提案されないが，通常考えられる「サラが子供達を褒めすぎたことが不用意 (愚か) であった」ことを表現したい文脈では，B の書き方では「サラの子供達の褒め方が愚かだった」ことを意味してしまうため不適である．この違いのような高度なものについては別の対策が必要である．また本研究では文中の誤り個数を対象に検証を行ったため，誤りの種類の内訳については更なる検証が必要である．本研究では分かりやすい間違いの模倣を試みるために自己紹介文などの一般的な題材で実験を行った．上に示したような高度な間違いと今回再現できなかった間違いに対応するために，今後は扱う題材を増やしてより多くの種類の間違いについて模倣が可能かを Prompt Insertion ベースの手法を基に改善しながら検証する予定である．

文　献

[1] Abdelkarim, R., Abuiyada, R. (2016). “The Effect of Peer Teaching on Mathematics Academic Achievement of the Undergraduate Students in Oman,” International Education Studies, 9, 124-132.
[2] Gubalani, J., Basco, J., Bulig, M. R., Bacatan, J. (2023). “The Effectiveness of Peer Tutoring in Enhancing Reading Comprehension of Ninth Grade Students,” Canadian Journal of Language and Literature Studies, 3(4), 75–89.
[3] Kaneko, M., Okazaki, N. (2023). “Controlled Generation with Prompt Insertion for Natural Language Explanations in Grammatical Error Correction, arXiv ID: 2309.11439v1.
[4] Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Cohen, W. W., Stylianides, G. J., Koedinger, K. R. (2013). “cognitive anatomy of tutor learning: Lessons learned with SimStudent,” ournal of Educational Psychology, 105(4), 1152–1163.
[5] Zou, N.R. (2023). Peer Tutoring in K-12 Education. Lecture Notes in Education Psychology and Public Media.

ピクタグラミング
－ピクトグラムの作成を通じたマークアップ言語・スタイルシート言語学習アプリケーションの開発とその評価－

木下 倭† 高橋 伶奈† 伊藤 一成†
†青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1

あらまし 高等学校必履修科目情報 I の「コミュニケーションと情報デザイン」では，Web コンテンツの制作を題材とした実習が想定される．しかし，時数が限られているため，マークアップ言語 HTML やスタイルシート言語 CSS の理解に加え，商用 Web サイトの品質に匹敵するデザインの作成は困難である．そのため，簡易的に Web コンテンツの制作が可能な学習環境が求められている．そこで筆者は，ピクトグラムの作成を題材とし，マークアップ言語 HTML やスタイルシート言語 CSS の記法の学習に特化しつつも，デザインに関しても学習することができるアプリケーション「ピクタグラミング」を開発している．今回ピクタグラミングを用いた授業実践を行い，評価を行なったため報告する．

キーワード マークアップ言語，スタイルシート言語，ピクトグラム，情報デザイン，情報 I

1. 背景

ピクトグラムとは，世界共通の記号表現として，意味するものの形状を使ってその意味概念を理解させる記号である[1]．

ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果があると言われている．有名な「非常口」ピクトグラムのデザインの策定に関わった太田は“走る人型を囲む空間が見る人を包む空間とつながって走る人は見る人の投影になる”と述べ[2]，実際に避難中の人が如何に出口へ向かって走る人型ピクトグラムを見たときに，自身に投影し，同一視するかにデザインの労力が払われた．

このピクトグラムの特徴を利用して，情報デザインやプログラミングの学習に用いるピクトグラミング及びその派生アプリケーション[3]，ソートアルゴリズムに関する学習[4]，データ活用に関する学習[5]など各種学習用アプリケーションを開発し，インターネット上で公開している．人型ピクトグラムを様々な学習単元に横断的かつ統一的に用いることで，新たな単元を学習する際に前提となる共通知識や情報を継承する方法について研究している．

令和 4 年度より，高等学校情報科では，「情報 I」が共通必履修科目となり「（2）コミュニケーションと情報デザイン」に関して，「効果的なコミュニケーションを行うための情報デザインの考え方や方法を理解し表現する技能を身に付けること．」，「効果的なコミュニケーションを行うための情報デザインの考え方や方法に基づいて表現し，評価し改善すること．」といった事項を身につけることが求められている[6]．この単元では，Web コンテンツの制作を題材にした実習も想定されている．Web コンテンツの制作に関する実習では，マークアップ言語 HTML で Web ページの構造を記述し，スタイルシート言語 CSS で Web ページを装飾する．しかし，タグ構造化文書の仕組みの理解や，HTML タグや CSS プロパティの仕様の理解にかかる時間は膨大であり，初学者が理解するのに一定の期間を要する．主に中高生を対象としたプログラミング学習サービスを利用した授業実践では，そのようなサービスの教育効果が一部認められたとする報告もあるが[7]，未だ教育効果について書かれた文献は少ない．このことから，教育効果や費用に関する懸念から，サービスを導入することをためらう教員も存在すると考えられる．そのため，会員登録やインストール等も必要なく，簡易的に Web コンテンツの制作が可能な学習環境が求められる．

そこでピクトグラムの作成を題材とし，タグ構造化文書（HTML）とスタイルシート言語（CSS）の記法の学習に特化しつつも，デザインに関しても学習できるアプリケーション「ピクタグラミング」を開発している[8]．今回「ピクタグラミング」の機能追加および授業実践を行い，評価したので報告する．以下，2 章で開発アプリケーションを説明する上で前提となるピクトグラミングについて説明する．3 章で「ピクタグラミング」の概要，実装方式，ピクトグラム作成のためのタグ要素集合 HPML（Human Pictogram Markup Language）について解説する． 4 章で授業実践，評価について述べ，5 章で，まとめと今後の展望について述べる．

2. ピクトグラミング

第 3 著者は，ピクトグラムを用いてプログラミングの諸概念を学習できる「ピクトグラミング」及びその派生アプリケーション群を開発している[3]．ピクトグラミングのスク

リーンショットを図 1 に示す.

ピクトグラミングには，視覚的なブロックを使って記述可能な Pictoch（ピクトッチ），Python 言語で記述できる Picthon（ピクソン），JavaScript 言語で記述できる JavaScpict（ジャバスクピクト），Ruby 言語で記述できる Picby(ピクビー)などの派生アプリケーションがあり，いずれもインターネット上で公開している．これらをピクトグラミングシリーズと総称している．いずれも Web アプリケーションであり，主要なブラウザ上で利用できる．詳細は，論文[3]を参照されたい.

図 1　ピクトグラミングのスクリーンショット

3.　ピクタグラミング

3.1　概要

ピクタグラミングは，タグ構造化文書やスタイルシート言語を記述することによって，ピクトグラムを様々に変化させることができる．ピクタグラミングは Web アプリケーションであり，ブラウザ上で利用できる．ピクタグラミングのスクリーンショットを図 2 に示す．ピクタグラミングでは，2 章で解説したピクトグラミングの UI を継承している.

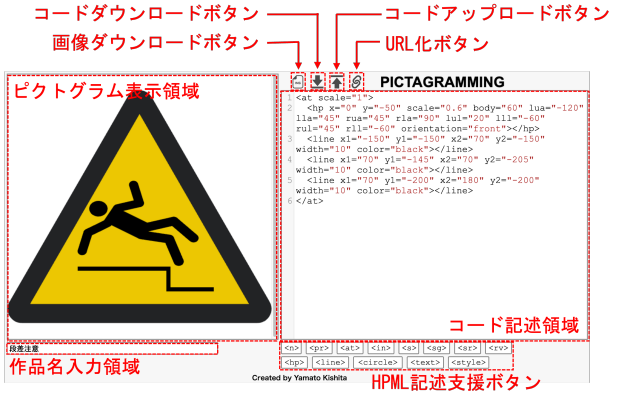


図 2　ピクタグラミングのスクリーンショット

画面左上部はコードの記述を反映する「ピクトグラム表示領域」，画面右中部はコードを記述する「コード記述領域」，画面右下部はピクトグラム作成のためのタグ要素集合 HPML(Human Pictogram Markup Language)の記述を支援する「HPML 記述支援ボタン」である．HPML については，3．2 節で解説する．また，画面右上部には，「ピクトグラム表示領域」の内容を画像としてダウンロードする「画像ダウンロードボタン」や，コードダウンロードあるいはアップロードをする「コードダウンロードボタン」，「コードアップロードボタン」，コードを含んだ情報を URL 形式の文字列で生成し，利用できる「URL 化ボタン」がある．画面左下部には，作品名を入力する「作品名入力領域」がある.

ピクタグラミングでは，「コード記述領域」に記述したコードに対応する表示は，「ピクトグラム表示領域」にリアルタイムに反映されるライブコーディングをサポートしている.

3.2　HPML(Human Pictogram Markup Language)

ピクトグラムの作成を簡便に実現するためのタグ要素集合を定義し，それを HPML(Human Pictogram Markup Language)と呼称する．ピクタグラミングでは，HTML で定義されている要素集合に加え，HPML をサポートする．HPML のタグ要素集合の一覧を表 1 に示す.

表 1　HPML タグ要素集合一覧

タグセット	処理
<n>	ノーマルのセーフティマークを付与する．このタグをセーフティマークタグとする．
<pr>	禁止のセーフティマークを付与する．このタグをセーフティマークタグとする．
<at>	注意のセーフティマークを付与する．このタグをセーフティマークタグとする．
<in>	指示のセーフティマークを付与する．このタグをセーフティマークタグとする．
<sa>	安全のセーフティマークを付与する．このタグをセーフティマークタグとする．
<sg>	安全（緑）のセーフティマークを付与する．このタグをセーフティマークタグとする．
<sr>	安全（赤）のセーフティマークを付与する．このタグをセーフティマークタグとする．
<rv>	反転のセーフティマークを付与する．このタグをセーフティマークタグとする．
<hp>	セーフティマークタグの中に記述することで，人型ピクトグラムを表示する．
<line>	セーフティマークタグの中に記述することで，線を描画する．
<circle>	セーフティマークタグの中に記述することで，円を描画する．
<text>	セーフティマークタグの中に記述することで，テキストを描画する．text タグの中の文字列が，テキストとして描画される．

これらの HPML タグには，独自の属性を定義している．属性は，先行研究のピクトグラミングの命令や変数に基づいており，ピクトグラミングと共に利用することで，学習者

の負担を軽減することを狙いとしている．次に，セーフティマークタグ，hp タグ，line タグ，circle タグ，text タグの属性の一覧を，表 2～6 にそれぞれ示す．

表 2　セーフティマークタグ属性一覧

属性	処理
scale="arg1"	セーフティマークの大きさを指定する．幅 arg1×640 px，高さ arg1×640 px の大きさとなる．初期値は"1"である．
background-color="arg2"	背景色を指定する．この属性は n, pr, at, in, sa タグのみ有効で，sg, sr, rv タグには無効である．初期値は"white"である．

表 3　hp タグ属性一覧

属性	処理
x="arg1"	水平方向の位置を指定する．セーフティマークの中心を x="0"とし，水平方向に arg1 px 移動する．なお，右方向を正の方向とする．初期値は"0"である．
y="arg2"	垂直方向の位置を指定する．セーフティマークの中心を y="0"とし，垂直方向に arg2 px 移動する．なお，下方向を正の方向とする．初期値は"0"である．
scale="arg3"	大きさを指定する．縦横それぞれ arg3 倍にする．初期値は"1"である．
angle="arg4, arg5, arg6, arg7, arg8, arg9, arg10, arg11, arg12"	身体の部位ごとの角度を指定する．arg4 から順に，body（体），lua（左上腕），lla（左前腕），rua（右上腕），rla（右前腕），lul（左大腿），lll（左下腿），rul（右大腿），rll（右下腿）の角度を変更する引数である．arg 度だけ反時計回りに回転する．引数を全て指定しない場合，body（体），lua（左上腕），lla（左前腕），rua（右上腕），rla（右前腕），lul（左大腿），lll（左下腿），rul（右大腿），rll（右下腿）の順に優先して指定される．初期値は，"0, 0, 0, 0, 0, 0, 0, 0, 0, 0"である．
body="arg13", lua="arg14", lla="arg15", rua="arg16", rla="arg17", lul="arg18", lll="arg19", rul="arg20", rll="arg21"	身体の部位ごとの角度を指定する．angle は複数の部位の角度をまとめて指定する属性であったが，これらは 1 つずつ角度を指定する属性である．angle と同様に，arg 度だけ反時計回りに回転する．angle で角度を指定している場合も，これらの属性で角度を指定した部位は，これらの属性の値で上書きされる．初期値は全て"0"である．
color="arg22"	色を指定する．初期値はセーフティマークによって変化する．
orientation="arg23"	向きを指定する．"front"の場合は正面向きに，"side"の場合は横向きにする．初期値は"front"である．

表 4　line タグ属性一覧

属性	処理
x1="arg1", y1="arg2", x2="arg3", y2="arg4"	描画する範囲を指定する．セーフティマークの中心を x="0"，y="0"とし，x="arg1"，y="arg2"の点から，x="arg3"，y="arg4"の点にかけて描画する．初期値は全て"0"である．
width="arg5"	太さを指定する．初期値は"15"である．
color="arg6"	色を指定する．初期値はセーフティマークによって変化する．

表 5　circle タグ属性一覧

属性	処理
x="arg1", y="arg2"	中心を指定する．セーフティマークの中心を x="0"，y="0"とし，x="arg1"，y="arg2"の点を円の中心とする．初期値は全て"0"である．
r="arg3"	半径を指定する．半径を arg3 px とする．初期値は"25"である．
color="arg4"	色を指定する．初期値はセーフティマークによって変化する．

表 6　text タグ属性一覧

属性	処理
x="arg1", y="arg2"	描画する位置を指定する．セーフティマークの中心を x="0"，y="0"とし，x="arg1"，y="arg2"の点をテキストを描画する左下の点とする．初期値は全て"0"である．
color="arg3"	色を指定する．初期値はセーフティマークによって変化する．
font-size="arg4"	フォントの大きさを指定する．初期値は"50"である．
font-family="arg5"	フォントを指定する．初期値は"Arial"である．

ピクタグラミングでは HPML に属性を指定することで，様々なピクトグラムを作成することができる．ここで，at タグと hp タグ，line タグを用いて作成した，段差に注意するように促すピクトグラムを図 3 に，対応するコードを図 4 に示す．

図 3　段差に注意するように促すピクトグラム

```
<at>
  <hp y="50" scale="0.6" body="60" lua="-120"
lla="45" rua="45" rla="90" lul="20" lll="-60"
rul="45" rll="-60"></hp>
  <line x1="-160" y1="150" x2="80" y2="150">
</line>
  <line x1="80" y1="140" x2="80" y2="210">
</line>
  <line x1="80" y1="200" x2="160" y2="200">
</line>
</at>
```

図 4　段差に注意するように促すピクトグラムを作成するコード

図 4 のコードでは，at タグで attention，注意のセーフティマークを描画している．そして，at タグの子要素に hp タグと 3 つの line タグを追加し，人型ピクトグラムと線を描画している．なお，hp タグと line タグには属性を指定している．hp タグには y 属性，scale 属性，body 属性，lua 属性，lla 属性，rua 属性，rla 属性，lul 属性，lll 属性，rul 属性，rll 属性を指定することで，ピクトグラムの垂直方向の位置，大きさ，身体の部位ごとの角度を変更し，図 3 の転倒する人型ピクトグラムを表現している．また，3 つの line タグに，それぞれ x1 属性，y1 属性，x2 属性，y2 属性を指定することで，描画する範囲を変更し，図 3 にあるピクトグラム下部の段差を表現している．このように，ピクタグラミングでは HPML を記述することで，多様なピクトグラムを表現することができる．

また，HPML タグは HTML タグやスタイルシートと併用することができる．それにより，図 5 に示すような，文章中心のコンテンツに挿絵的にピクトグラムを含めた Web ページも作成することができる．

図 5　文章とピクトグラムの両方を含むコンテンツ例

3.3 CSS の HPML 要素への適用

ピクタグラミングでは HPML タグに対しても CSS を適用可能である．CSS で記述するプロパティ名とプロパティ値は，HPML の属性名と属性値と同様であり，HPML タグの属性と同様の処理を行う．なお，HPML タグの属性とプロパティを両方記述した場合には，HPML タグによる記述が優先される．ここで，図 4 の段差に注意するように促すピクトグラムを作成するコードを，HPML の属性ではなく，CSS プロパティを用いて記述したものを図 6 に示す．

HPML の属性を用いた記述と比較すると，id 属性や class 属性を指定することで，再利用性の高いコードを記述することが可能である．このように，ピクタグラミングでは，CSS プロパティを HPML タグに適用することで，ピクトグラム作成のテーマであっても，タグ構造化文書とスタイルシート言語の記述方法の両方が学習できる．

```
<at>
  <hp  id="slipped"></hp>
  <line id="line-top"></line>
  <line id="line-center"></line>
  <line id="line-bottom"></line>
</at>

<style>
#slipped {
  x: 0;
  y: 50;
  scale: 0.6;
  body: 60;
  lua: -120;
  lla: 45;
  rua: 45;
  rla: 90;
  lul: 20;
  lll: -60;
  rul: 45;
  rll: -60;
}
#line-top {
  x1: -160;
  y1: 150;
  x2: 80;
  y2: 150;
}
#line-center {
  x1: 80;
  y1: 140;
  x2: 80;
  y2: 210;
}
#line-bottom {
  x1: 80;
  y1: 200;
  x2:160;
  y2: 200;
}
</style>
```

図 6　段差に注意するように促すピクトグラムを CSS プロパティを用いて記述したコードの例

4. 実践と評価

4.1 概要

ピクタグラミングの評価のために，2023 年 9 月 22 日（金曜日）に授業実践を行なった．対象者は青山学院大学社会情報学部 2 年生以上が対象の講義及び実習「ウェブテクノロジー」を履修する学生，計 72 名である．1 回の授業時間は 90 分で，およそ 40 分をピクタグラミングを用いた実習にあてた．授業は，一般的な PC 実習室が配当されている．評価の観点は，1. システムユーザビリティスケールによるユーザビリティの評価，2.ピクタグラミングを使用した授業設計及び実施が可能かどうかという実施可能性に関する評価の 2 点を行なう．実習の概要を表 7 に示す．

表 7　実習の概要

内容	時間（分）
1. ピクタグラミングの使い方の説明	20
2. オリジナル作品制作	15
3. アンケート	5

20 分でピクタグラミングのテキストを用いて使い方を説明し，15 分でオリジナル作品を制作させた．オリジナル作品のテーマは自由とし，課題として提出させた．提出の締切りは授業から 6 日後の 2023 年 9 月 28 日とした．

4.2 アンケート評価

アンケートは，アプリケーションのユーザビリティを評価するシステムユーザビリティスケールによる項目[9]と，授業に関する評価や，アプリケーションのコンセプトに関する質問項目の 2 つを行なった．システムユーザビリティスケールによる項目を表 8 に示す．

表 8　システムユーザビリティスケールによる項目

	内容
質問 1	このアプリケーションをもっと頻繁に使いたい
質問 2	このアプリケーションは必要以上に複雑だと感じる
質問 3	アプリケーションがシンプルで使いやすいと思う
質問 4	このアプリケーションを使うには、技術的なサポートが必要だ
質問 5	アプリケーションがスムーズに機能し、うまく統合されていると感じる
質問 6	イレギュラーなことも多いと思う
質問 7	このアプリケーションは、ほとんどの人がすぐに習得できると思う
質問 8	このアプリケーションは手間がかかると思う
質問 9	このアプリケーションを使っている間は自信がある
質問 10	このアプリケーションを使い始めるまでには、学ぶべきことがたくさんあると思う

アンケート結果を表 9 に示す．回答選択肢の内容は，「5.強く同意する」，「4.同意する」，「3.どちらとも言えない」，「2.否定する」，「1.強く否定する」とした．システムユーザビリティスケールのスコアは 65.0 となった．最低限ユーザの満足度を満たすには，スコアは一般的に 70 以上であることが望まれる．そのため，今後ピクタグラミングのシステムユーザビリティを改善する必要がある．質問 4，質問 9，質問 10 が特にスコアが低かったため，その理由について考察した．

表 9　システムユーザビリティスケールによるアンケート結果

	強く否定する	否定する	どちらとも言えない	同意する	強く同意する	スコア
質問 1	0	0	10	51	11	75.3
質問 2	9	43	13	7	0	68.8
質問 3	0	1	8	39	24	79.9
質問 4	4	26	17	24	1	52.8
質問 5	0	3	9	40	20	76.7
質問 6	7	32	21	10	2	61.1
質問 7	1	7	17	32	15	68.4
質問 8	8	32	22	10	0	63.2
質問 9	1	15	34	19	3	52.8
質問 10	6	19	23	20	4	51.0
合計						65.0

その理由とは，エラーに関する警告を出す機能がないためである．ピクタグラミングでは，HPML タグ，HTML タグともに半角で記述する必要があり，全角で記述すると意図した通りに表示がされなくなる．また，HPML タグや属性の綴りを間違えた場合や，属性値を囲うシングルクォーテーション，またはダブルクォーテーションの記述を忘れた場合も意図した通りに表示されなくなる．このようなエラーを防ぐために，HPML 記述支援ボタンやサンプルコードを用意したが，オリジナル作品制作の際には，作品を制作していくうちにエラーを起こしてしまったと考えられる．これらは，質問 6，質問 8，質問 9 のスコアの低下につながったと考えられる．

授業に関する評価や，アプリケーションのコンセプトに関する質問項目は表 10 に示す 5 項目である．質問 1〜4 は，いずれも 6 段階の回答選択肢を用意した．回答選択肢の内容は，「6.とてもそうである」，「5. かなりそうである」，「4.どちらかというとそうである」，「3.どちらかというとそうでない」，「2.あまりそうでない」，「1. 全くそうでない」とした．これらの選択肢に対して，それぞれ 6 点，5 点，4 点，3 点，2 点，1 点として評価した．質問 5 は自由記述形式である．質問 1〜4 の回答結果を表 11 に示す．質問 5 の自由記述の回答から一部抜粋したものを表 12 に示す．

表 10　アンケート項目

Q	内容
1	今回の授業は楽しかった．
2	コンセプトに共感した．
3	HTML の学習のモチベーションが向上した．
4	Web コンテンツ制作への興味が深まった．
5	今回の授業の感想を自由に書いてください．（自由記述）

表 11　授業に関する評価や，アプリケーションのコンセプトに関する質問のアンケート結果

質問	1 点	2 点	3 点	4 点	5 点	6 点	平均
Q1	0	1	0	21	28	22	4. 97
Q2	0	0	0	26	29	17	4. 88
Q3	0	3	1	20	31	17	4. 81
Q4	0	1	2	16	32	21	4. 97

表 12　質問 5 ”今回の授業の感想を自由に書いてください” 一部抜粋

	内容
回答 1	ピクトグラミングは直感的に操作できるため標識を楽しく作ることができました。
回答 2	HTML と同じような要領でコードを書けるので簡単で面白かったです。
回答 3	機能を理解するのが早く、使いやすかった。理解してからはサクサクと作業を進められた。
回答 4	あらかじめピクトグラムを生成するボタンがあるので HTML のルールさえ理解できれば簡単に使用することができるのがよいと感じた。

どの質問に関しても，90%以上の学生が「どちらかというとそうである」，「かなりそうである」，「とてもそうであ

る」という肯定的な回答をしたことがわかる．質問 2 に関しては，肯定的な回答が 100%を占めた．また，質問 5 の自由記述では，「楽しい」や「面白い」，「容易にピクトグラムを作成できた」といった回答が多かった．この結果から，ピクタグラミングを用いることで，短時間で楽しみながら Web コンテンツ制作の実習を行うことが可能だと考えられる．

5. まとめと今後の展望

本稿では，ピクトグラムの作成を題材にした，マークアップ言語 HTML とスタイルシート言語 CSS の学習を支援するピクタグラミングを実装し，評価した．

今後もピクタグラミングのシステムユーザビリティを改善していくと同時に，主に高校生を対象とした授業実践を実施し，繰り返し評価，分析を行っていく予定である．また，高等学校情報科「情報 I」における，情報デザインの分野での有効性についても検証していく予定である．

謝辞

本研究は JSPS 科研費 21H03560 の助成を受けたものです．

参 考 文 献

[1] 太田幸夫：国際安全標識のピクトグラムデザインの研究：http://www.tamabi.ac.jp/soumu/gai/hojo/seika/2003/kyoudou-ota1.pdf （参照 2024-01-03）

[2] 太田幸夫 : FORUM8：ユニバーサル・コミュニケーションデザインの認識と実践：https://www.forum8.co.jp/topic/universal107.htm （参照 2024-01-03）

[3] 伊藤一成 : 複数のプログラミング言語で記述可能なピクトグラムコンテンツ作成環境の提案と実装，情報処理学会論文誌 TCE，Vol.7，No.3，pp.1-11（2021）

[4] 渡辺大智，御家雄一，伊藤一成：人型ピクトグラムを用いたソートアルゴリズムを学ぶアプリケーション「人型ピクソートグラム」の実装，2022 年度情報処理学会関西支部大会（2022）

[5] 高橋伶奈，御家雄一，伊藤一成：人型ピクトグラムを構成素とするピクトグラフ生成アプリケーション「Human Pictograph」の実装，2022 年度情報処理学会関西支部大会 (2022）

[6] 文部科学省：【情報編】高等学校学習指導要領（平成 30 年告示）解説：https://www.mext.go.jp/content/000166115.pdf （参照 2024-01-03）

[7] 井手広康：高等学校におけるプログラミング教育サービスを活用したプログラミングの授業実践と教育効果の検証，日本情報科教育学会誌，Vol. 14，No. 1，pp.39-44.（2021）

[8] 木下倭，高橋伶奈，伊藤一成：ピクタグラミング -ピクトグラムの作成を通じたマークアップ言語学習アプリケーションの試作-，2023 年度情報処理学会関西支部支部大会（2022）

[9] Brooke, J.; SUS: A "quick and dirty" usability scale., Usability Evaluation in Industry. London:Taylor and Francis.（1996）

ピクトグラムに着目した教育向けネットワークシミュレータ「Pictocol Simulator」の試作

前田 祐杜† 伊藤 一成†

† 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5-10-1

あらまし 高等学校の学習指導要領の教科「情報」に必履修科目の「情報Ⅰ」が設定されたことで，多くの高校生が科目「情報Ⅰ」の学習をすることになった．コンピュータネットワークは学習項目の一つであるが，他の領域に比べ，専門用語の理解や技術的側面が強い．そこで今回，コンピュータネットワークの構築や実際の通信プロトコルを日常生活でのアクティビティに関連づけることで，コンピュータネットワークの繋がりの概念の理解を促すアプリケーション「Pictocol Simulator」を試作した．アプリケーションの特徴の一つに，ピクトグラムが有すると言われている投射の効果に着目した点が挙げられる．情報機器を人型ピクトグラムで擬人化して表現し，またネットワーク上の流れるデータや概念等の不可視の情報もピクトグラムで表現している．本論文では「Pictocol Simulator」の設計・実装方式及び，「Pictocol Simulator」を使用した学習方法の提案について解説する．

キーワード コンピュータネットワーク，情報教育，人型ピクトグラム，情報可視化

1. はじめに

2022 年度から，高等学校の学習指導要領の教科「情報」に必履修科目の「情報Ⅰ」が設定された．また，2025 年度の大学入試共通テストに教科「情報」加わったことで，科目「情報Ⅰ」の内容は，多くの大学が学力評価にも利用するようになる．しかし，現状，科目「情報Ⅰ」は導入されてまもなく，教授法も確立しているとは言い難い． 通常，他の大半の科目と同様に，教科書や副読本を用いた教授法が確立されると考えられる．しかし，科目「情報Ⅰ」で取り扱われる単元の中には，普段から使用していながらも意識し辛い分野が多く含まれている．また，現代の人々は IT 機器や IT サービスに多く触れていることから，普段の生活からこれらが解決する問題に気づくことが可能である．このことから教科「情報」は高等学校普通科目の中では実学に近い分野になると考えられる．そのため科目「情報Ⅰ」では IT 技術への関心を引くことができるコンテンツを使用することが重要である．情報分野の概念を体験的に学習できるコンテンツを用いることで，学習対象の技術がどのような仕組みであり，どのように使用できるのかを考えることができる．既に様々な学習コンテンツが誠意開発されており，プログラミング単元においては多く見受けられる一方，コンピュータネットワークの単元は比較的少ない．また，高校生が，コンピュータネットワークの概念をイメージするのは必ずしも容易ではない．そのため，コンピュータネットワークの概念を高校生の身近にある馴染み深い概念で表現することで，コンピュータネットワークの概念を容易にイメージできると考えられる．

そこで，我々は高校生が日常生活の中で馴染みがある概念を用いて，コンピュータネットワークの繋がりの概念を視覚的に表現することで，本単元を体験的に学習することができるアプリケーション「Pictocol Simulator」（ピクトコルシミュレータ）を試作したので報告する．

以下 2 章では研究の背景と動機，3 章で関連研究について説明する．4 章で試作したアプリケーションについて解説し，5 章では本アプリケーションを使用した授業内での実習方法について提案し，その新規性や有用性について考察する．6 章でまとめる．

2. 背景と動機

コンピュータネットワークの単元を教科書のみを用いて学習した場合，ネットワーク機器の名称やプロトコルなどの基本的な知識を得ることはできる．しかし，それぞれの機器が実際のコンピュータネットワーク上でどのような役割を担い，どのように繋がることでコンピュータネットワークが構成されているのかをイメージし辛い．特に，一般的な家庭に普及しているネットワーク機器は，ルータ，スイッチそして，アクセスポイントなどの複数の機能が一体化した製品であることから，それぞれの機能の違いについて考える必要がない．また，プラグアンドプレイに対応しているネットワーク機器が大半を占めており，多くの人にとってネットワークは機器のケーブルをプラグに挿入すればつながるという認識を持っているため，その裏でネットワークがどのように繋がっているのかを意識せずに使用していることも，コンピュータネットワークを理解し難い原因の一つであると考えられる．

高等学校学習指導要領では，科目「情報Ⅰ」の情報

通信ネットワークとデータの活用の内容の取扱いに関して，小規模ネットワークを設計する活動を取り入れると記載している[1]．また，教員研修資料では，ドローツールを用いてネットワーク図を作成する手法が紹介されている[2]．しかし，ネットワーク設計にとどまらず，実際にネットワーク機器の設定やトラブルシューティングを体験することは，コンピュータネットワークをより具体的に理解する有効な手法である．しかし，ネットワーク機器を用いて実習環境を整えるのはコストやセキュリティの観点から難しい．そのため，ネットワークの構築を体験できるソフトウェアが必要である．

3. 関連研究

北村らは TCP/IP の仕組みを体験的に学ぶことができる教材である「ProtoSim」を提案している[3]．この教材は，クライアントと web サーバ間での通信時，それぞれの内部の通信プロトコルの各層で流れる情報を視覚的することで，初学者が，TCP/IP の仕組みや重要性を学習できる．また，荒井らは，実際の通信のパケットをモニタリングし，視覚化することで，通信手順やデータの流れを学習できる教材を開発した[4]．これらは OSI 参照モデルで示されている各層の理論の理解に有効である一方，ネットワークの繋がりに対して関連する理解を得るのは難しい．

吉原らは，2 つの Rasberry Pi 間で通信を行い，受診時に LED を点灯させることで，直感的にコンピュータネットワークを理解することができる教材を開発した[5]．また，石川らは，吉原らの研究を発展させ，OpenFlow を用いて，Rasberry Pi にルータやファイアウォールなどの機能を実装することとで，より深くコンピュータネットワークを学習できる教材を開発した[6]．しかし，これらの教材は，Rasberry Pi やケーブルなどの機器を用意する必要があることに加え，構築できるネットワークが非常に小規模であり，学習者が独自に変更を加えることが難しい．

鈴木は，ネットワーク機器を使用せずにネットワークを理解させるロールプレイの実践を行っている[7]．学習者が，実習環境の内部に身を置ける点は，学習者が主体的に考えることに有効である．

ネットワークを構築するためのソフトウェアとしては Cisco Systems 社が提供している Packet Tracer がある[8]．このソフトウェアは，Cisco Systems が提供しているネットワーク機器を用いて PC の画面上で自由にネットワーク構築ができる．また，そのネットワーク上を流れるパケットを可視化できる．しかし，機器の設定にはベンダ独自のコマンドを習得する必要があり，学習コストが高い．立岩らも，ネットワーク構築を体験できるとともに，そのネットワーク上に流れるパケットの情報を可視化するシステムを開発した[9]．このシステムは，Packet Tracer と同様に，画面上で仮想ネットワークを構築できる．しかし，コマンドではなく，GUI でネットワーク機器の設定を行うことから，学習コストが比較的低い．これらの学習教材はネットワーク機器の機能を幅広く網羅していることに加え，様々なプロトコルを用いた通信経路を視覚化できる．このような機能の豊富さから，コンピュータサイエンスを専攻している学生が，ネットワークを学習する際に使用するのには非常に適している．しかし，充実度の高さから，学習には，ネットワークの単元に対する一定以上の理解が必要である．そのため，各分野を 1 時間程度で完結させる科目「情報 I」で採用するのは難しい．

taidalab のネットワークシミュレータ[10]は，限定された機能の中で，ネットワーク構築を行い，そのネットワーク上で通信を体験できる．このアプリケーション上で使用できるネットワーク機器はルータ，ハブ，そしてエンドポイントとなる PC のみである．また，設定できる情報はネットワーク機器のホスト名，IP アドレス，そしてサブネットマスクのみである．このように機能が制限されていることから，短時間で当単元の学習を行うことができる．しかし，このアプリケーションではエンドポイント間の通信経路の視覚化をサポートしていない．

これらの関連研究を元に，我々は短時間でネットワーク構築を体験でき，構築したネットワーク上での通信を可視化することで，ネットワークの仕組みや繋がりを意識できるアプリケーションを試作した．また，学習者が実習環境に自身を置き，身近に感じることができる施策を加えることで，学習者が主体的に学習できる環境の構築を目指す．

4. 提案アプリケーション

試作したネットワーク構築を体験することができる Web アプリケーション「Pictocol Simulator」の設計方針，搭載機能及び使用方法について解説する．

4.1. 概要・設計方針

本節では，「Pictocol Simulator」を作成するにあたり，考慮したことやどのような目的のもと設計したのかを解説する．

コンピュータネットワークにおいては，ノード間の通信がどのような経路を通って行われているのかを見ることが難しい．そのため，コンピュータネットワークの単元の初学者である高校生がコンピュータネットワーク上のパケットの流れを理解するのは難しい．ま

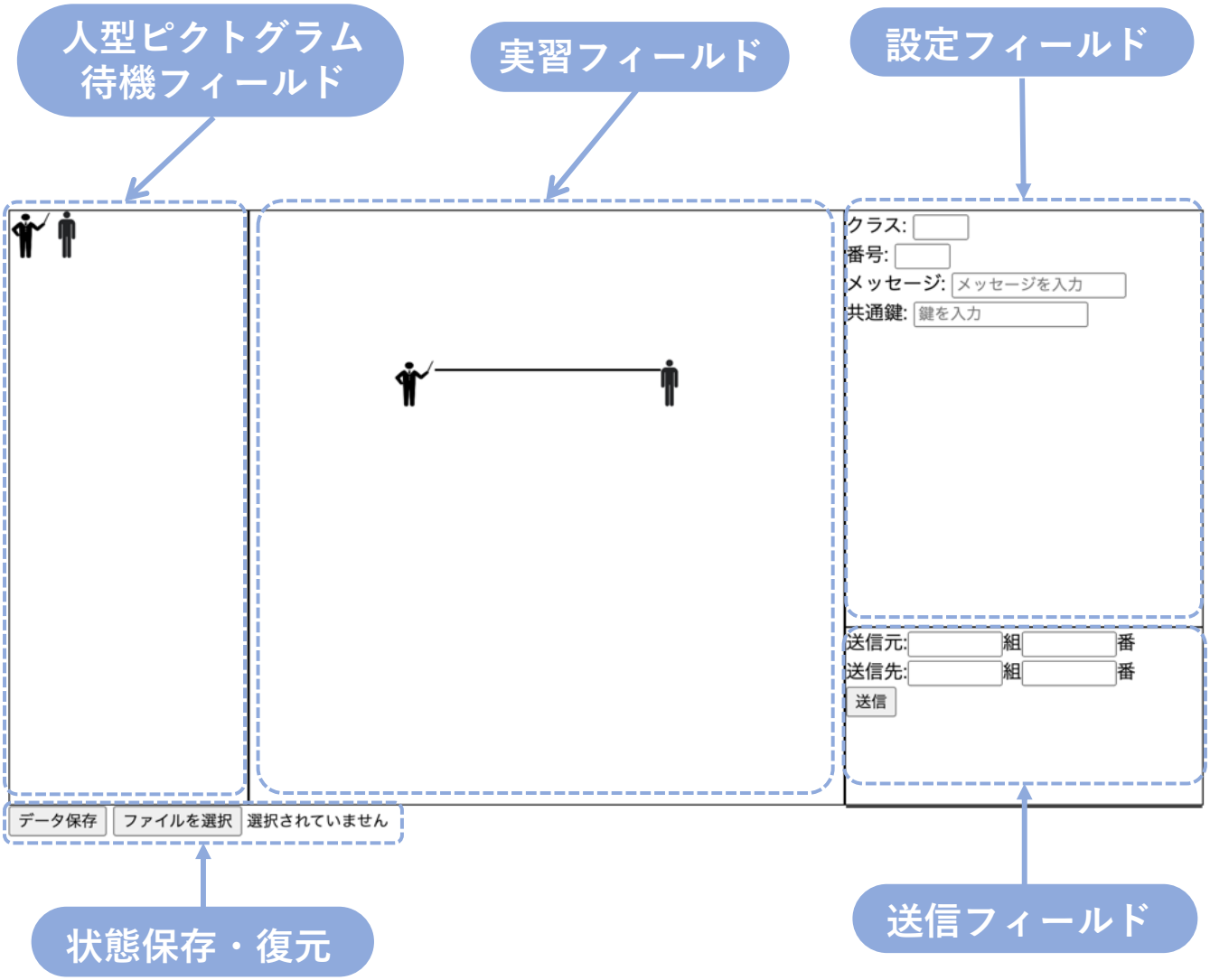


図 1　全体のスクリーンショット

た，単にコンピュータネットワークの繋がりを視覚的に表現するだけでなく，主体的に学習できるコンテンツにする必要が望ましい．そこで，コンピュータネットワークと日常生活との結びつきを高校生に理解させ，主体的に学習できるコンテンツにすることにした．また，その中でも，コンピュータネットワークの繋がりの概念と乖離が少ない概念で表現する．

本アプリケーションにおいては，コンピュータネットワークの繋がりの概念を学校におけるクラスを用いて表現することにした．コンピュータネットワークでは一つの大きなネットワークの中で，複数の小規模なネットワークが存在する事が多い．学校においては，学校という大きな組織の中にクラスが複数存在することから，構造上の観点に着目し，コンピュータネットワークを学校におけるクラスで表現するのは妥当であると考えた．また，本アプリケーションのメインの対象者は高校生であることから，概念をイメージする容易さの観点に着目したとき，学校に関する概念が適していると考えた．

本アプリケーションにおける，学校の構造とコンピュータネットワークの対応関係を解説する．コンピュータネットワークにおけるネットワーク部はクラス名（A 組，1 組などの学級を識別する概念）で表現した．また，ホスト部はクラス内の出席番号で表現した．クラスとクラス内での番号を組み合わせた値は学校全体での識別子として機能することから，IP アドレスと同じ役割を果たす．

学習者がネットワークの繋がりの概念をより身近に感じさせるために，本アプリケーション上ではネットワーク上に存在するノードを人型ピクトグラムで置き換えている．また，本アプリケーションでは人型ピクトグラムは先生と生徒の二種類が存在する．

人型ピクトグラムを用いることで，本アプリケーションの使用者に自己投影を促す．これにより，本アプ

リケーション上で構築したネットワークを学習者が所属する学校として考えることができる．そこで，登場するノードを現実世界での学習者と，学習者の周辺人物ととらえることで，より具体的にイメージを掴むことができる．

また，本アプリケーションで構築したネットワークと実際のコンピュータネットワークの概念との間に乖離が生じないように次のような制約を設けている．

(1) 先生もクラス内での出席番号を所有する．
(2) 先生は，同クラスの生徒か他のクラスの先生と繋がることができる．
(3) 生徒は先生と繋がることができる．
(4) 生徒間でのメッセージのやり取りは必ず先生を経由する．

これらの前提により，実際のネットワークがどのように繋がっているのかを体感し考えさせることで，理解を促進させることができる．

4.2. アプリケーション

4.2.1. アプリケーション概要

本アプリケーション全体のスクリーンショットを図 1 に示す．画面左側には人型ピクトグラム待機フィールドがある．画面中央部には実習フィールドがある．画面右側上部には設定フィールドがある．画面右側下部には送信フィールドがある．

本アプリケーションを用いた実習の流れは次のとおりである．

1. 人型ピクトグラム待機フィールドから実習フィールドに人型ピクトグラムを移動させる
2. 実習フィールド内の人型ピクトグラム同士をリンクでつなぐ
3. 設定フィールドで実習フィールド内の各人型ピクトグラムに属性を与える
4. 実習フィールド内の任意の人型ピクトグラムから他の人型ピクトグラムにメッセージを送信する

これらの機能の詳細及び使用方法は次節より解説する．また，補助的な機能として次のような機能も使用できる．

- メッセージ暗号化・復号機能
- トポロジ保存・復元機能

補助的な機能については，4.3 節で解説する．

4.2.2. 人型ピクトグラムの配置

本項では，実習で構築するトポロジに必要となるノード（人型ピクトグラム）を配置する方法について解説する．

図 2 に示すように，人型ピクトグラム待機フィールドには二体の人型ピクトグラムを配置している．一つは先生ピクトグラムでもう一つは生徒ピクトグラムである．実際のネットワーク機器との対応関係としては，先生ピクトグラムがルータ兼スイッチの役割を果たし，生徒ピクトグラムは PC やスマホなどのネットワークに接続するエンドデバイスの役割を果たす．

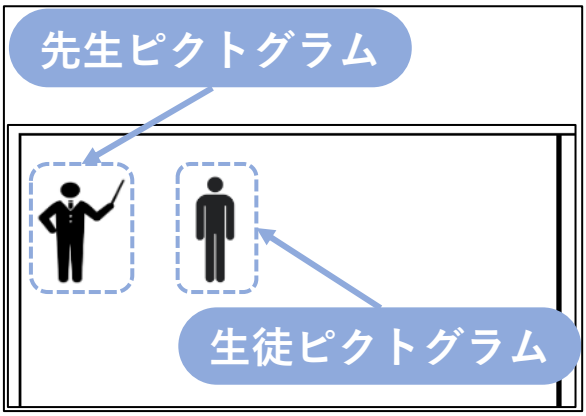


図 2　人型ピクトグラム待機フィールド

本アプリケーション上の人型ピクトグラムはドラッグアンドドロップで移動可能である．図 3 のように，構築するネットワークに必要となる人型ピクトグラムをそれぞれ人型ピクトグラム待機フィールドから実習フィールドにドラッグアンドドロップで移動する．

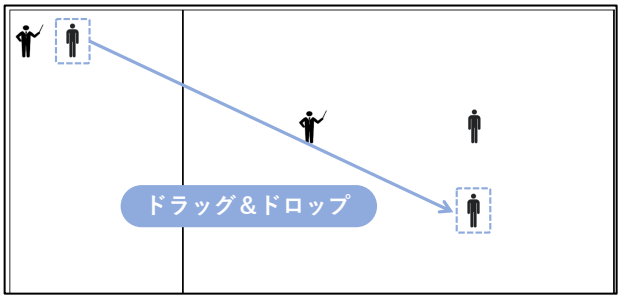


図 3　実習フィールドへの配置

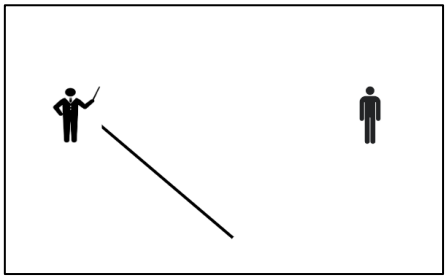
4.2.3. 人型ピクトグラム間のリンキング

本項ではトポロジの構成にあたり，ノード間の接続状態を表現する方法について解説する．

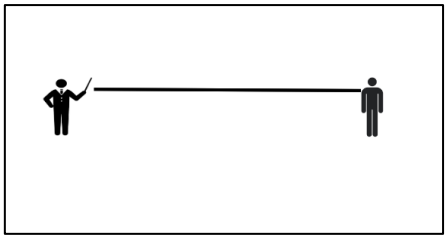
本アプリケーションでは，ノード同士の繋がりをコンピュータネットワークにおける優先 LAN のみで表現する．実際のコンピュータネットワークでは，エンドポイント同士を直接接続するのではなく一度スイッチやルータを経由させることが一般的である．そのため，本アプリケーションにおいても，エンドポイントである生徒ピクトグラム同士の接続はできない．本アプリケーションでは以下の人型ピクトグラム間の接続が可能である．

- 生徒ピクトグラムと先生ピクトグラム間
- 先生ピクトグラムと先生ピクトグラム間

任意の人型ピクトグラムにマウスカーソルを合わせて右クリックすると図 4 の(1)で示すように，リンクの片端を人型ピクトグラム上に固定する．この状態でマウスカーソルを動かすと，リンクのもう片端がマウスカーソルに追随し，一度目にクリックした人型ピクトグラムとマウスカーソルの座標を両端とするリンクを設定する．その状態でマウスカーソルを動かし別の接続可能な人型ピクトグラム上で再度右クリックすると図 4 の(2)で示すように，二体の人型ピクトグラム間をリンクで接続する．



(1) リンクの片側の固定



(2) 人型ピクトグラム同士の接続

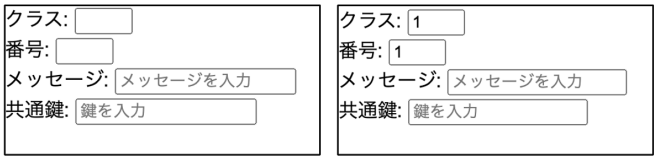
図 4　リンキングの状態

4.2.4. 人型ピクトグラムへの属性付与

本項では，ネットワーク上で通信をする際に必要となる識別情報の設定方法について解説する．

ネットワーク上のノード同士が通信するためには,ノードそれぞれに一意の識別子を与える必要がある．4.1 節で記したように本アプリケーションにおいては，人型ピクトグラムの所属クラスと出席番号が複合識別子である．人型ピクトグラム間で通信するために，有効なクラスと出席番号を人型ピクトグラムに与える必要がある．

クラス及び出席番号を人型ピクトグラムに与える方法について解説する．図 5 に設定フィールド上に表示される属性の設定入力前と設定入力後の例を示す．実習フィールド内の人型ピクトグラム上にマウスカーソルを動かし，左クリックで選択する．画面右側の設定フィールドには，最後に選択した人型ピクトグラムが所有している現在の属性とその値が表示される．初期状態は，属性に対する値を設定していないので属性名のみが表示されている（図 5 の(1)）．この設定フィールド内の属性に対するテキストエリアに値を入力することで,人型ピクトグラムの属性に値を設定できる(図 5 の(2))．学習者は，構築したいネットワークトポロジに従うように，設定フィールドから正しいクラスや出席番号の値を人型ピクトグラムに与える．このとき，4.1 節で示した制約(1)に従い，先生ピクトグラムにも出席番号を与える必要がある．現実では，出席番号は生徒のみに割り振られるが，本アプリケーションにおいては，先生ピクトグラムもネノードの一つである．ネットワーク上のノードは識別子が必要であるため，先生ピクトグラムにも番号を与える．



(1) 設定入力前　　(2) 設定入力後

図 5　属性の値の設定

4.2.5. 人型ピクトグラム間の通信

本項では，構築したトポロジや設定した内容をもとにノード間で通信を行う方法について解説する．

本アプリケーションでは，ネットワークを構築するだけでなく,ノード間での通信を体験する機能がある．コンピュータネットワーク上では通信ノード間でパケットが移動するが，本アプリケーションでは，通信対象の人型ピクトグラム間で手紙を受け渡しすることで,表現している．そして，通信を開始すると通信対象の人型ピクトグラム間で手紙が移動する状態を視覚化する．そのため，コンピュータネットワーク上では，パケットが通信対象のノード間をどのような経路を通っているのかに対するイメージを深めることができる．

使用方法について解説する．図 6 に実際に通信を行う例を示す．実際のアプリケーションの画面では，図 6 中の(a)～(d)の記号自体は表示しないが，本稿全体ではそれぞれの人型ピクトグラムの識別を容易にするために表示している．

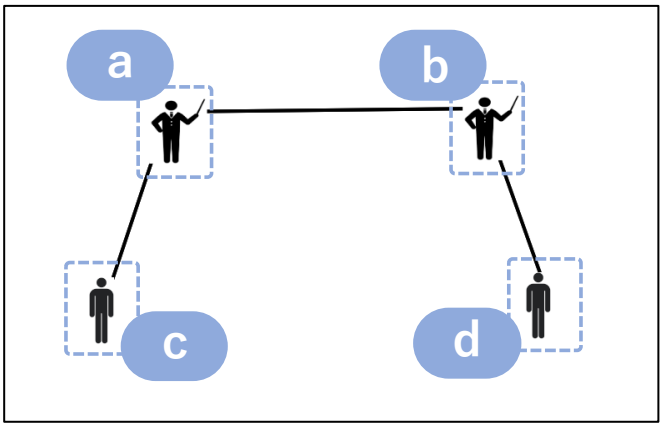


図 6　トポロジ例

図 6 中の(a)～(d)の人型ピクトグラムそれぞれに与えた設定内容を図 7 に示す．本アプリケーションにおけるクラスがコンピュータネットワークのネットワーク部に該当することから，先生ピクトグラム(a)と生徒ピクトグラム(c)，そして先生ピクトグラム(b)と生徒ピクトグラム(d)がそれぞれのネットワークに所属している．

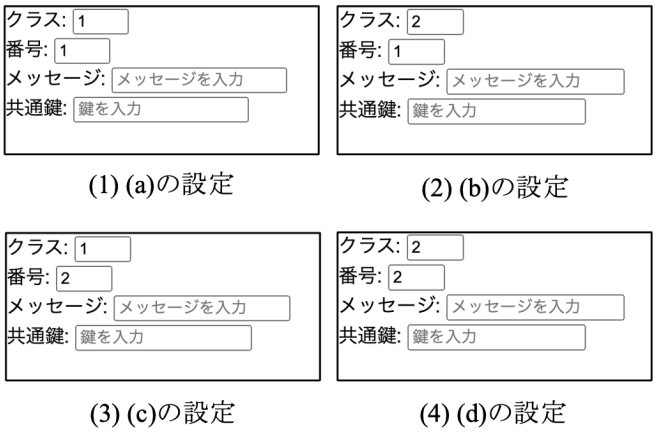


図 7　図 6 上の(a)~(d)の設定

次に手紙を送信する手順を解説する．学習者はメッセージ送信元の人型ピクトグラムと送信先の人型ピクトグラムを選定する．図 6 では，図 6 中の(c)の生徒ピクトグラムを送信元人型ピクトグラム，(d)の生徒ピクトグラムを送信先人型ピクトグラムにそれぞれ選定している．実習フィールド内の送信元人型ピクトグラムにカーソルを合わせクリックし，設定フィールドを表示する．設定フィールド内にはクラスや番号の他にメッセージという属性がある．そこに送信先人型ピクトグラムに対して送信したいメッセージを記述する．他にもメッセージ送信の過程で使用できる属性として共通鍵が存在する．これは，本アプリケーションにおける補助的な機能で，送信するメッセージを暗号化する場合のみ入力が必要となる．暗号化しない場合は入力が不要の属性である．この暗号化機能に関する詳細は，4.3.1 で解説する．

設定フィールドのメッセージ入力欄に送信メッセージを入力した後，画面右側の下部にある送信フィールドを使用しメッセージを送信する．このフィールドでは，送信元人型ピクトグラムと送信先人型ピクトグラムのそれぞれのクラスと番号を該当の入力欄に記述する．両方のクラスと番号の入力が完了した後，送信ボタンをクリックする．(c)が送信元人型ピクトグラムで，(d)が送信先人型ピクトグラムであるため，図 7 の(3), (4)に示した内容をもとに送信フィールドに適切な値を入力する．

図 8 に入力後の送信フィールドの状態を示す．送信ボタンを押すと，実習フィールド内の送信元人型ピクトグラムの場所に手紙の画像を表示し，設定したネットワークの経路を通って送信先人型ピクトグラムの場所まで手紙が移動する．このときのスクリーンショットを図 9 に示す．

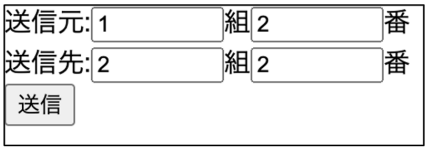


図 8　送信フィールド入力例

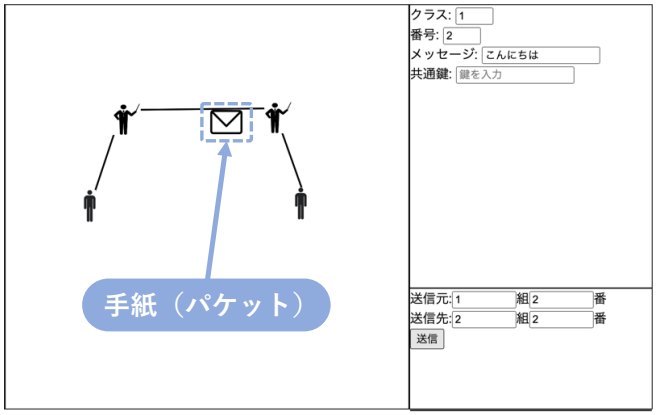


図 9　送信中の画面

手紙のアイコンが送信先人型ピクトグラムに到達した後，送信先人型ピクトグラムの設定フィールドを開くと，受信メッセージ属性の値に受信したメッセージを表示する．このときの例を図 10 に示す．

通信経路上の途中に存在する中継ピクトグラムの設定に誤りがある場合，正しく設定した経路まで手紙が進み，誤っている場所で手紙が停止する．そのため，原因箇所の特定が比較的容易である．途中で終了した場合は，手紙が停止した箇所を参考に，原因箇所を訂

正し,再度送信ボタンを押すことで検証できる.ただ,途中で手紙が停止する情報のみでは原因を特定できない学習者が一定数出てくることが想定される．そのため，今後は現状に加えて，原因をより詳細に説明するためのガイドを表示し，学習者が必要であれば参照できる機能の実装を検討している.

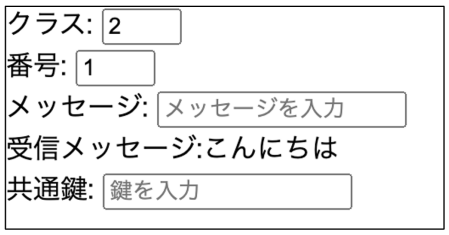


図 10　メッセージの受信

4.3. 補助的機能の解説

4.3.1. メッセージ暗号化・復号機能

本設では，メッセージ送信の際にオプションで使用できる機能であるメッセージの暗号化・復号機能を解説する.

辰己らは，情報科の入試問題への考察として，コンピュータネットワークの単元は暗号通信の単元と関連させる問題として出題する必要があると記している[11]．そこで，本アプリケーションでは，コンピュータネットワーク上で鍵暗号方式を体験できる機能を実装している.

本アプリケーション上でメッセージを送信するとき，手紙を中継する中継ピクトグラムはメッセージの内容を傍受できる仕様である．図 11 には，図 6 の (c)の生徒ピクトグラムから(d)の生徒ピクトグラムにメッセージを送信した後の(b)の先生ピクトグラムの設定フィールドの内容を示す．受信メッセージ属性に，(d)の生徒ピクトグラムの受信メッセージと同じ内容を表示している．このことから，送信元人型ピクトグラムからのメッセージを経由した中継ピクトグラムはメッセージ内容を取得できることがわかる.

そこで，暗号化・復号機能を使用することで，中継ピクトグラムが平文を取得することを防ぐ．これにより実際のコンピュータネットワーク上での暗号化の有効性を体験できる.

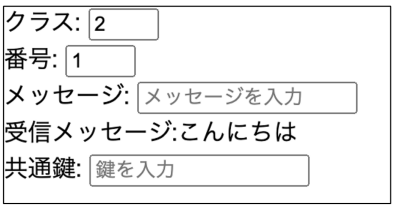


図 11　中継ピクトグラムの受信メッセージ

使用方法について解説する．4.2.5 で説明した方法で,送信元人型ピクトグラムのメッセージ属性に送信メッセージを入力するとともに，図 12 で示すように，共通鍵属性に共通鍵パスフレーズを入力する．また，受信者も共通鍵を知っていることから，送信先人型ピクトグラムの共通鍵属性にも送信者と同じ共通鍵パスフレーズを入力する.

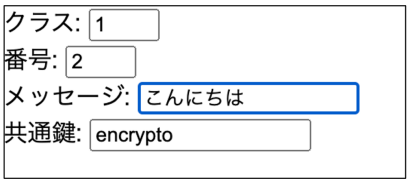
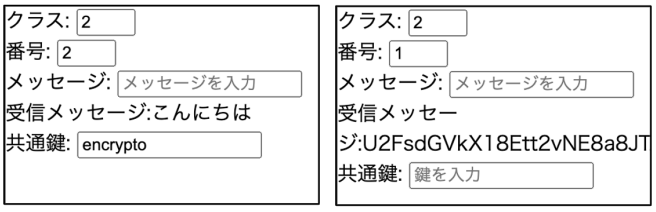


図 12　共通鍵の入力

送信元人型ピクトグラムから送信先人型ピクトグラムに対してメッセージを送信する．すると，メッセージを暗号化し，手紙の画像が送信先人型ピクトグラムまで移動する．送信先人型ピクトグラムの共通鍵属性に暗号化に使用したものと同じ共通鍵パスフレーズを入力していた場合，受信した暗号文を復号する．図 13 の(1)で示すように受信メッセージ属性には平文のメッセージを格納する.

次に,手紙が通過する経路上の中継ピクトグラムは暗号文を格納していることを確かめる．暗号化前と同じく，(b)の生徒ピクトグラムをクリックし設定フィールド内の受信メッセージ属性を開く．図 13 の(2)で示すように暗号文を表示している．このことから，ネットワーク上で通信を行う際に，メッセージを暗号化することの有効性がわかる.



(1)　復号成功　　(2) 復号失敗

図 13　メッセージの受信（暗号化通信）

このように，本アプリケーション上で暗号化通信を体験することで，暗号化の必要性や使用用途を体験的に学ぶことができる.

4.3.2. 状態保存・復元機能

本アプリケーションは学習者が Web ブラウザを閉じる他，ページの再読み込みをしたとき，実習環境の状態を保持しない．これによる使用性の低下を防ぐた

めに，本アプリケーションでは，実習環境上のトポロジや人型ピクトグラムが持つ属性の情報などを保存及び復元できる機能を実装している．

本項では，状態保存・復元機能の使用法を解説する．

1) 状態保存機能

図 14 で示すように，画面下部に保存ボタンがある．このボタンをクリックすると，現在の実習フィールドのトポロジの状態と，それぞれの人型ピクトグラムが所有している属性の情報を JSON ファイルに保存する．そして学習者の端末に JSON ファイルをダウンロードする．

2) 状態復元機能

図 14 で示すように，画面下部にファイルを選択するフィールドがある．このフィールドに状態保存機能を用いてダウンロードした JSON ファイルをアップロードする．すると，状態保存機能を使用したときの実習フィールドのトポロジと，そこに存在する人型ピクトグラムが所有する属性の情報を復元する．

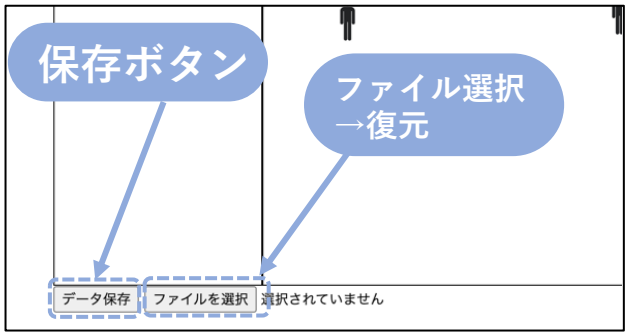


図 14　状態保存・復元機能

5. 学習方法の提案

本章では本アプリケーションを用いた学習方法の提案を行う．想定学習環境としては，各生徒に対し一台の PC が割り当てられている PC 教室である．また，実習の前半は生徒が自ら使用するのではなく，教師の誘導のもと実習を行うことを想定しているため，何らかの形で教師の，本アプリケーション内での動作を共有できる環境があるとなお良い．生徒が使用する PC とは別に教師の画面を共有する専用のモニタやプロジェクタがあることが望ましい．

本アプリケーションを用いた実習を行う前に，教科書レベルの講義が行われていることを前提とする．実習の前半は教師の実習を生徒が真似する形式で実習を進めていく．所属クラスの異なる先生ピクトグラム 2 体を経由するトポロジを作成し，それぞれの人型ピクトグラムに設定を行い，異なるクラスの生徒ピクトグラム間でメッセージの送信ができることの確認までは，教師主導のもとで行うことを想定している．その後は，生徒が各々トポロジや設定にアレンジを加えていく方法を想定している．

また，教師が事前に一通りのトポロジと設定を終えた状態で保存し，生徒に JSON ファイルを共有する方法も考えられる．教師が配布した状態を学習者変更し，実現してほしい要件を生徒に伝えることで，教師が提案した要件を満たすネットワークの作成をする方法有効である．また，教師が一部の設定を誤ったものにし，あえては通信できない状態で配布する方法も考えられる．これにより，学習者にネットワークが繋がらないときの原因を調査する機会を提供できる．

6. まとめ

本稿では，コンピュータネットワークの単元を体験的に学習することで理解を促進し，コンピュータネットワークの繋がりを意識させるアプリケーションである「Pictocol Simulator」の機能や使用方法について解説した．今後は，実際のコンピュータネットワークと人型ピクトグラムを用いたネットワークの関連に対する学習者の意識をより深めるために，モードを切り替えることで，実習フィールドの人型ピクトグラムをネットワーク機器のイラストに変更すると同時に，設定フィールド内のクラスや番号を IP アドレスに変更する機能を実装する予定である．

謝辞

本研究は JSPS 科研費 21H03560 および 20K03160 の助成を受けたものです．

参　考　文　献

[1] 高等学校学習指導要領　p191 https://www.mext.go.jp/content/20220324-mxt_kouhou02-000021499_1.pdf

[2] 高等学校情報科「情報 I」教員研修用教材　第 4 章　情報通信ネットワークとデータの活用 https://www.mext.go.jp/content/20200722-mxt_jogai02-100013300_006.pdf

[3] 北村祐稀, et al. 初学者を対象とした TCP/IP の仕組みと重要性を体験的に学ぶシミュレータ教材の開発. 情報教育シンポジウム論文集, 2023, 2023: 249-253.

[4] 荒井正之，田村尚也，渡辺博芳，小木曽千秋，武井惠雄（2003）TCP/IP プロトコル学習ツールの開発と評価．情報処理学会論文誌，44(12)：3242-3251.

[5] 吉原和明; 井口信和; 渡辺健次. 物理的可視化と物理的直接操作による IP アドレスの仕組みを学習するための教材の開発と評価. 日本産業技術教育学会誌, 2018, 60.2: 73-80.

[6] 石川有彩, et al. OpenFlow を用いたネットワーク学習教材の開発. 研究報告インターネットと運用技術 (IOT), 2021, 2021.2: 1-7.

[7] 鈴木大助, et al. 通信の仕組みを理解するためのロールプレイ演習の開発と実践. 研究報告コンピュータと教育 (CE), 2017, 2017.10: 1-7.

[8] Cisco Packet Tracer サイト
https://www.netacad.com/ja/courses/packet-tracer

[9] 立岩佑一郎, et al. 仮想環境ソフトウェアに基づく LAN 構築技能と TCP/IP 理論の関連付け学習のためのネットワーク動作可視化システムの開発. 情報処理学会論文誌, 2007, 48.4: 1684-1694.

[10] taidalab「情報 I」学習サイト
http://taidalog.html.xdomain.jp/

[11] 辰己丈夫, et al. 情報科の入試問題に関する一考察. 情報教育シンポジウム論文集, 2023, 2023: 261-268.

図形への書き込みと手書き解答文字の自動認識による数学学習者の解答戦略推定

林 凱博† 宮﨑 公彦‡ 工藤 雅士‡ 山名 早人§

†早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: †‡§{yhayashi0908, loup, kudoma34, yamana}@yama.info.waseda.ac.jp

あらまし 近年，遠隔学習の普及により，オンライン学習プラットフォームや学習アプリケーションの利用が急速に増えている．これに伴い，学習者の状況を的確に理解し，適切なフィードバックを提供すること，すなわちテーラーメイド教育の重要性が高まっている．また，中学校の学習指導要領では，数学幾何を学ぶ必要性が強調されている．そこで本研究では，実際の証明を書き込んだ解答欄のオンライン手書き解答データを使用して解答戦略の自動推定を行った．具体的には，OCR（Optical Character Recognition）を使用して証明文を書き込んだ解答欄の画像から，特徴量となる式や文字を抽出し，これを特徴量として分類器を学習させる．また，精度向上を図るために，解答欄に書き込まれた証明の手書きデータだけでなく，図中に書き込まれた記号の解析も活用し，さらなる精度向上に挑戦する．解答欄に書き込まれた証明の手書きデータのみを特徴量として学習させた分類器で推定を行った結果，3 種類の解法がある問題において，従来の手法では 79.2%の正解率であったのに対し，提案手法では正解率 91.7%を得ることができた．

キーワード 学習分析，e-Learning，物体検出，文字認識

1. はじめに

近年の技術の進歩や利便性の向上，多様な学習機会の提供，そして災害時への対応により，オンライン学習の需要がますます増加している．特に，2020 年に始まった COVID-19 の影響により，多くの教育機関がオンライン学習への移行を強いられた．また，文部科学省は 2019 年 8 月から「個別最適な学びと協働的な学び」を実現するために，児童生徒に対して「1 人 1 台端末」等の ICT 環境を整備する「GIGA スクール構想[1]」を開始した．内閣府が行っている「青少年のインターネット利用環境実態調査[2]」の最新結果（2023 年 3 月 31 日発表）では，中学生の 70.0%が GIGA 端末を利用していると発表している．このような状況を受けて，遠隔学習に対する重要性が再確認されつつある．

加えて，中学校学習指導要領[3]においては，数学幾何を学ぶ必要性が強調されている．幾何学を学ぶことは日常生活や社会の事象を図形の概念や性質，関係として理解し，同時に論理的な考察に基づいて現実の出来事をモデル化するのに有用である．したがって，オンラインを通じた幾何学の学習は重要であるが，一方で学習者一人一人に適切な教育を提供することが課題となっている．近年の技術進歩により，学習者の状態を正確に把握するのに利用可能な多種多様なデータを生成することが可能になった．こうした現状から，学習者のデータを利用して，学習者の状態を適切に把握し，学習の質や効果を向上させるための研究が盛んに行われている．

数学の手書き解答データを用いた研究では，2018 年に Goh ら[1]が幾何学証明問題の教育と学習に関する課題を理解することを目的に，幾何数学問題の手書き解答データを手動で解析し，採用した解法を特定した．2022 年には Cirillo ら[2]が同様に手動で学生の手書き解答データを解析し，学生の戦略の使用に焦点を当て，それらが問題解決にどのように貢献するかを調査した．また，図への書き込みを自動認識できることを前提として解法の分類を行った研究[3]もあるが，これらの解答戦略推定手法はすべて手動での解析に留まっている．そこで，岡田ら[4][5]は図中に書き込まれた記号の内容と種類を自動判定すると共に，記号の書き込み順を解析して解答戦略の自動推定を試みた．岡田ら[4][5]の研究によって，自動での解答戦略推定が行えるようになった一方で，課題も残されている．図形への書き込みを行わない，または書き込みが少ない学習者や，図形には書き込んでいるが，その記号を解答戦略として採用しない学習者がいるため，岡田ら[4][5]の手法ではそのような学習者の解答戦略を推定することが難しい．実際に，答案は完成されているのに，図への書き込みが無い（あるいは少ない），または図への書き込みが乱

[1] https://www.mext.go.jp/content/20210608-mxt_jogai01-000015850_003.pdf
[2] https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/html/nd122230.html
[3] https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_004.pdf

雑で記号の物体検知が行えないような被験者データがあった．

そこで，先行研究の課題を解決するために，本研究では解答欄に書き込まれた証明の手書きデータを，OCR（Optical Character Recognition：光学文字認識）を用いてテキストデータとして抽出し，これを特徴量として分類器を学習させることで，図形への書き込みを行わない学習者に対しても解答戦略の推定を可能にする．また，図形への書き込みと証明文の書き込みの 2 つの特徴量を 2 パターンの方法で統合して推定を行う．一方は，両方の特徴量を一度に学習することを目的に，もう一方はそれぞれの特徴量において，検出できなかった記号や式を相互に補完した 1 つの特徴量で学習することを目的に行った．また，図形への書き込みを特徴量として学習させた分類器と証明文の書き込みを特徴量として学習させた分類器の 2 つの分類器において，各分類器による推定ミスを相互に補うことを目的に，予測確率を重み付き和の形で統合した．

本稿の構成は次の通りである．まず，2 節で関連研究について述べ，3 節で研究目的について説明する．4 節では提案手法について説明し，5 節ではデータ収集実験について説明する．6 節では評価実験の結果および考察を示し，最後に，7 節で本稿をまとめる．

2. 関連研究

著者ら[6]は，学習中に取得できる学習者の履歴データから学習状態を推定する論文 40 件を 6 種類の推定対象に分類し，手法やデータの傾向についてまとめた．その結果，全体の 28%が学習状態推定に関する論文であった．使用データの傾向としては，オンライン学習中の学習履歴データが最も多く使用されていることが明らかになった．また，データ解析手法としては，深層学習系の手法よりも古典的な機械学習手法が多く利用されている．以下では，理解度の推定と解法分類に関する研究を紹介する．

2.1 理解度の推定

Yoshitake ら[7]は，2020 年にペンストロークデータ解析を活用した教育支援システムを提案した．具体的には，二次方程式に関する問題のペンストロークデータを使用し，k-means 法により学生の解答パターンをクラスタリングした．ペンストロークデータは解答にかかった時間，ストローク数，ストローク間の間隔，および消しゴムの使用などの情報が含まれている．この分析に基づいて，異なるスキルや知識を持つ学習者が同じグループになる異質なグループ，または同じスキルや知識を持つ学習者が同じグループになる同質なグループ，またはランダムなグループを形成した．

三浦ら[3]は，2022 年にオンライン手書きデータを使用して学習の定着度を推定する手法を提案した．具体的には，52 名の被験者から幾何数学問題の解答データを収集し，オンライン手書きデータから得られる特徴量を使用して学習の定着を予測した．特徴量としては，ストローク残存率，速度，筆圧，停滞時間率などを使用した．SVC（Support Vector Classification）を用いた分類の結果，80.5%の F 値を達成した．さらに，易問・難問を対象外として分類を行うと，82.0%の F 値を達成した．これは学習者自身が定着度を判断する場合と比較して 2.7%の F 値の改善が見られた．

2.2 解法分類に関する研究

村上ら[8]は，2019 年にオンライン教育での学習者への適切なフィードバック提供を目指し，手書き解答を解法ごとに分類し，思考過程を推定する手法を提案した．具体的には，幾何図形問題において，32 人の被験者に対して 2 題の幾何学図形問題を出題し，収集した手書き解答データから事前に用意した 3 種類の解法および「ヒントを見て解いたかどうか」についての自動分類を試みた．SVM（Support Vector Machine）を用いた分類の結果，解法の分類において平均 59.6%の F 値，ヒントの使用有無の分類で平均 81.9%の F 値を達成した．さらに，図への書き込みを自動認識できることを前提として解法の分類を行うと，F 値は平均で 79.2%となり，問題によっては完全に誤りなく分類できることが確認された．

岡田ら[4][5]は，2023 年にオンライン手書き解答データを用いた解答戦略の自動推定手法を提案した．幾何問題を題材として，図中に書き込まれた記号の内容と種類を自動判定すると共に，書き込み順を解析することで，解答戦略の分類を行った．実証実験では，38 人の被験者を対象に幾何問題を解いてもらい，事前に用意した 3 種類の解答への自動分類を試みた．XGBoost を用いた分類の結果，uni-gram，bi-gram，1-skip-gram 全ての特徴量を使用した場合，正解率 71.0%を達成した．

3. 研究目的

本稿の目的は，幾何数学問題において，学習者がどのような解答戦略を選択しているかを自動で推定する手法を提案することである．具体的には，OCR 技術を活用して抽出した証明文と図形への書き込み情報を特徴量とし，学習者の解答戦略を推定する．

4. 提案手法

4.1 データ収集方法

中学の幾何数学問題のオンライン解答データを収集するために，宮﨑ら[9]の研究で使用された Web アプリケーションを使用する．解答の入力には，ペンタブレッ

ト（Wacom Intuos ProMedium[4]）を使用した．問題解答時の解答画面を図 1 に示す．

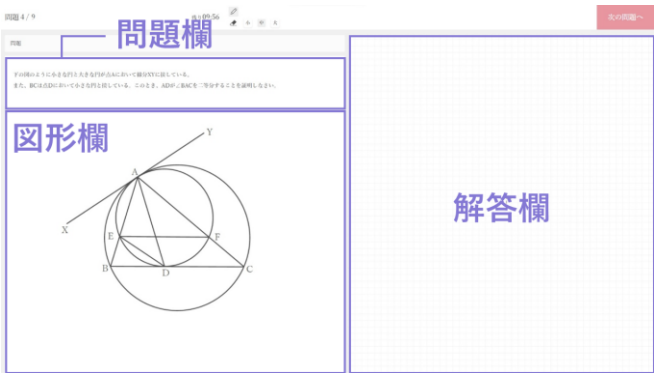


図 1　Web アプリケーションの解答画面

解答画面の左側に問題欄と図形欄，右側に証明を書き込む解答欄がある．本稿では，図形欄に書き込んだ記号と解答欄に書き込んだ証明文の両方から特徴量を生成し，解答戦略の推定を行う．

4.2 提案手法の概要

本稿では，学習者の解答戦略を推定する問題として，岡田ら[4][5]の研究で使用された三角形の相似を証明する問題を使用する．使用した問題を図 2 に示す．

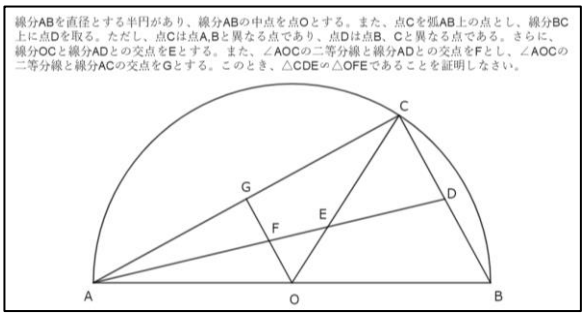


図 2　使用した問題

図 2 の問題において用意した 3 つの解法 A，B，C を表 1[4][5]に示す．

表 1　解法の一覧（[4][5]と同様）

解法 A	∠ECD = ∠OBC = ∠EOF の利用
解法 B	∠DCE = 90° - ∠GCO = ∠FOE の利用
解法 C	∠DCE = ∠FOE(錯角)の利用

表 1 に示す解法 A, B, C の自動分類を行うために，以下の 4 つの手法を使用する．

1. 図形内の記号検出手法（4.3 項）
2. 解答欄内の手書き文字の検出手法（4.4 項）
3. 図形への書き込み特徴量と解答欄内の手書き文字特徴量の統合（4.5 節）
4. 解答戦略の推定手法（4.6 項）

本手法では，岡田らの手法[4][5]（4.3 節）に対して 1) OCR を使用して解答欄の手書き文字特徴量を生成（4.4 節），2) 図形への書き込み特徴量と解答欄内の手書き文字特徴量を 1 つの特徴量として統合し学習（4.5 節），3) 図形への書き込み特徴量と解答欄内の手書き文字特徴量をそれぞれ使用して学習させた分類器をアンサンブル（4.6 節）させ，解法戦略の自動推定において性能向上を目指す．

4.3 図形内の記号検出手法

解答中に図形に書き込まれた記号を特徴量として抽出する．幾何問題を解く際，図形の性質を記号として書き込む．これらの記号は，角度，辺の長さ，相似性などの幾何学的な関係性を表すため，問題の解法に関連しており，解法の選択において重要な手がかりとなる．解答中に図形に書き込まれた記号を特徴量として抽出するために，以下の 2 つのステップを行う．

1. YOLOv8 を使用して画像から記号を自動認識する．
2. 認識した記号が図形のどの領域に書き込まれたかを判定し，特徴量を生成する．

4.3.1 YOLOv8 を使用した記号の自動認識

YOLOv8 を使用して，画像から記号を自動認識できるモデルを作成した．ここで YOLO とは，Joseph Redmon 氏ら[10]が 2016 年に提案した畳み込みニューラルネットワークを利用した物体検出手法である．認識対象とした記号は岡田ら[4][5]の研究で定義された 14 の記号（図 3 の 1~2 行目）に，比の記号を 3 種類（図 3 の 3 行目）加えた合計 17 種類である．アノテーション（記号へのタグ付け）を行った記号の種類と表記を図 3 に示す．

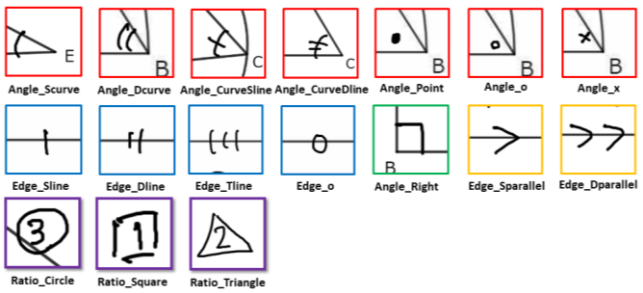


図 3　アノテーションを行った記号

モデルの学習には，推定に使用した解答データとは別に岡田ら[4][5]の研究で記号検出モデルの学習に使用された画像データと同一の 2188 枚の画像を使用した．これらの画像を，学習用とテスト用に 8 : 2（学習用：1750 枚，テスト用：438 枚）で分割して学習を行った．モデルの評価には AP（Average Precision）を使用した．各記号の AP と，全記号の AP の平均を表す mAP（Mean Average Precision）を表 2 に示す．さらに，物体検知モデルを使用して検出した記号の例を図 4 に示す．

[4] https://www.wacom.com/ja-jp/products/wacom-intuos-pro

表 2　各記号の AP と mAP

記号の分類	記号の種類	記号の総数		AP
		訓練	テスト	
等角	A_Scurve	3120	267	0.694
	A_Dcurve	157	66	0.747
	A_CurveSline	162	69	0.882
	A_CurveDline	73	41	0.709
	A_Point	264	108	0.950
	A_o	219	94	0.872
	A_x	156	68	0.846
直角	A_Right	646	258	0.909
等辺	E_Sline	266	187	0.959
	E_Dline	530	409	0.857
	E_Tline	10	35	0.893
	E_o	38	89	0.925
平行	E_Sparallel	247	251	0.972
	E_Dparallel	51	53	0.879
比	R_Circle	306	223	0.971
	R_Square	16	21	0.878
	R_Triangle	7	21	0.851
mAP				0.870

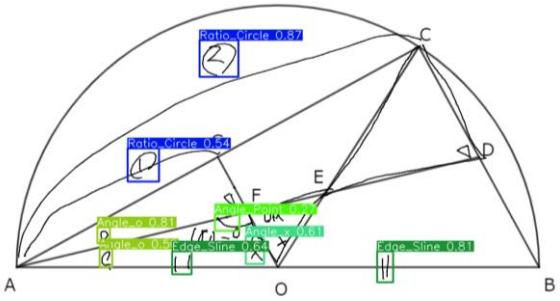


図 4　検出した記号の例

物体検知モデルに画像を渡すと，検出した記号の Bounding Box（境界ボックス）と Confidence スコアを含んだ画像，そして 6 つの項目（クラスラベル（物体の種類），BoundingBox の中心の X 座標，BoundingBox の中心の Y 座標，BoundingBox の幅，BoundingBox の高さ，信頼度（Confidence）スコア）を含んだテキストファイルが得られる．

4.3.2　書き込み領域の判定

4.3.1 項で YOLOv8 を使用して検出した記号と座標をもとに，岡田ら[4][5]の研究で使用した手法と同様の手法で記号の書き込み箇所を判定する．具体的には，YOLOv8 で検出した記号の Bounding Box を使用して書き込み領域の判定を行った．検出した Bounding Box があらかじめ設定した領域内に存在すれば，書き込みがあったものとして記録する．実際に，書き込みの判定を行った記号と特徴量として生成したときの表記を表 3 に示す．

表 3　書き込み判定を行った記号

記号の意味	判定を行った記号	表記
等辺	AG=GC	AGeqGC
	AO=OB	AOeqOB
	OC=AO	OCeqAO
	OB=OC	OBeqOC
平行	BC//OG	BCparaOG
直角	ACB=90°	ACBright
	AGO=90°	AGOright
	CGO=90°	CGOright
等角	∠OBC=∠OCB	OBCeqOCB
	∠COG=∠AOG	COGeqAOG
	∠OBC=∠AOG	OBCeqAOG
	∠OCB=∠COG	OCBeqCOG
	∠AOG=∠OCB	AOGeqOCB
	∠COG=∠OBC	COGeqOBC
	∠CDE=∠OFE	CDEeqOFE
	∠OAC=∠OCA	OACeqOCA
	∠CED=∠OEF	CEDeqOEF
	∠DCE=∠FOE	DCEeqFOE

1 ストローク毎に取得した画像データに対し，YOLOv8 を使用して記号の検出を行い，検出された記号に関して書き込み領域を判定する．このとき，2 ストローク連続で同一領域内に同じ記号の書き込みがあった場合のみ，書き込みがあったと判定する．これは図 3 に示すように「A_Scurve」が次のストロークで 1 画書き足されて「A_Dcurve」などになる可能性を考慮するためである．このようにして，書き込みが検出された記号をリストに順番に格納していき，書き込みの順序も記録しておく．そして，書き込み順をリストに格納していく段階で，表 3 に示す式が完成された時点でその式が出現したと判定する．つまり，AG=GC を例に取ると，辺 AG に E_sline が書き込まれ，その後に辺 GC に同じ記号である E_sline が書き込まれたら AG=GC が出現したと判断する．

4.3.3　N-gram を使用した特徴量の生成

4.3.1, 4.3.2 項で得られた図形に書き込まれた記号の種類とその順序を表す特徴量を生成するために uni-gram, bi-gram, 1-skip gram を用いて特徴量を生成する．

Uni-gram では，各特定の記号が図形に書き込まれたかどうかを表す二値ベクトルとして表現する．つまり，特定の記号が含まれていれば 1，含まれていなければ 0 とする．例えば，図形への書き込みの種類が AG=GC, AO=OB，AO=OC の 3 種類だけであると仮定する．このとき，AG=GC，AO=OC という図形への書き込みがあった場合，uni-gram は(1, 0, 1)となる．このベクトルは，AO=OB が含まれていないことを示し，AG=GC, AO=OC が含まれていることを示している．

Bi-gram は，書き込みの順番を 2 つの記号の組み合わせを表現する．上記の例を同様に使用すると，3 つの記号から 2 つの組み合わせを考えると次のように 3

×2=6 通り存在する．(AG=GC, AO=OB), (AG=GC, AO=OC), (AO=OB, AG=GC), (AO=OB, AO=OC), (AO=OC, AG=GC), (AO=OC, AO=OB).

このとき，AO=OB，AO=OC という順番で図形への書き込みがあったとき(AO=OB, AO=OC)の 1 つの組み合わせが存在する．よって，bi-gram では(0, 0, 0, 1, 0, 0)というベクトルで表すことができる．

1-skip gram は，bi-gram のアプローチを拡張したもので, bi-gram と同様に 2 つの記号の組み合わせを表現する．ただし，1-skip gram では，連続した 2 つの記号だけでなく，1 つの記号をスキップして組み合わせを考える．1-skip-gram は誤検出があった場合のパターンを考慮するために使用する．例えば, AG=GC, AO=OB, AO=OC という順番で書き込みがあったときに，(AG=GC, AO=OC)という組み合わせが 1 つ存在する．よって，1-skip gram では(0, 1, 0, 0, 0, 0)というベクトルで表すことができる．

4.4 解答欄内の手書き文字の検出手法

解答欄に書き込まれた証明文の中に含まれている文字を特徴量として抽出する．特に，アルファベットを用いて等角や等辺を表す式と図形の特徴を表す単語の抽出を行う．証明文を自動で認識し，特徴量として抽出するために，以下の 2 つのステップを行う．

1. OCR を使用して画像から文字を認識し, テキストデータを生成する．
2. テキストデータから，特徴量となる式や文字を抽出する．

ここで OCR とは，光学文字認識（Optical Character Recognition）の略で，画像などからテキストデータに変換する技術である．

4.4.1 OCR を使用した文字認識

画像から文字を認識し，テキストデータに変換する方法として Google Drive API の OCR 機能を使用した．実際に Google Drive API を使用して画像をテキストデータに変換した例を図 5 に示す．

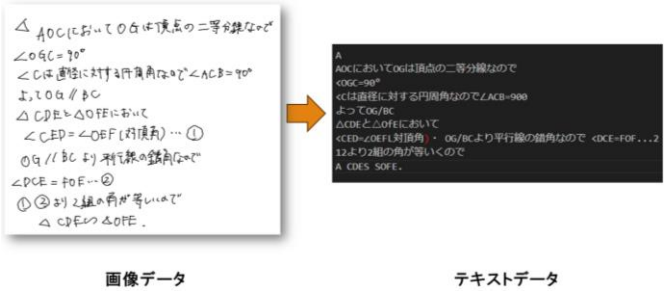


図 5　画像からテキストデータへの変換例

4.4.2 テキストデータから式や文字の抽出

OCR を用いて取得されたテキストデータ（図 5）から式の抽出を行った．まず，式の抽出を簡単にするために，前処理として以下の処理を順番に行った．

1. 全アルファベットを大文字にする．
2. 各行で日本語が続いた後にアルファベットが現れたら改行する．
3. アルファベット，数字，記号（=, /, °）以外をすべて削除する．

各処理を行った後のテキストデータの例を図 6 に示す．

図 6　前処理後のテキストデータ

次に，表 3 に示された 18 種類の式がテキストデータに含まれているかを確認した．この抽出プロセスは，以下の 2 つのステップから構成される．

1. 式の出現検索
   - 各行において，表 3 に示す 18 種類の数式の出現を検索し，出現が確認された場合，その行の番号を記録する．
   - 数式の出現を検索する際，アルファベットの順番も考慮する．例えば,「CDE=OEF」と「EDC=FEO」はアルファベットの順番が違うが，同じ意味である．
2. レーベンシュタイン距離による比較
   - ステップ 1 で出現が確認できなかった行に対して，全く関係のないアルファベットの挿入，削除，置換が生じている可能性があるため，各行と表 3 に示す 18 種類のターゲットの数式を比較し，レーベンシュタイン距離が 2 以下であったものに対してその行番号を記録する(距離は 3 以上だと間違えて抽出することが多くなったため 2 とした).

テキストデータから式の抽出を行った結果の例を図 7 に示す．

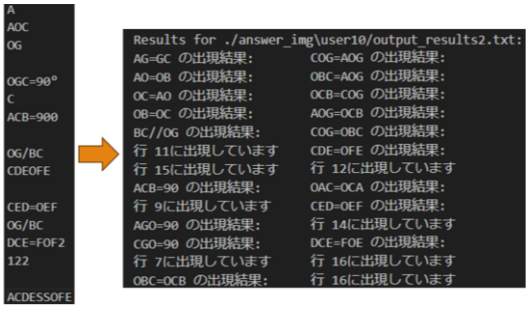


図 7　式の抽出を行った結果の例

次に，図形の特徴を表すような単語を抽出する．本

問題では，「直角」，「錯角」，「平行」，「二等辺」という 4 つの単語とその単語に対応する辺や角をセットとしてテキストデータから抽出する．抽出方法は以下の 2 ステップで行う．なお，式の抽出と同様にテキストデータに対して前処理を行ってから以下のステップで出現検索を行った．

1. 単語の出現検索：4 つの単語の出現を検索し，もし出現が確認されれば，その行の番号を記録する
2. 単語に対応する角や辺の出現検索：単語に対応する辺と角度の出現を検索し，もし出現が確認されれば，その行番号を記録する．なお，単語に対応する辺と角度の出現は上記で示した式の出現検索と同様の方法で行っている．
3. 単語と単語が示す辺や角の組み合わせ（セット）の検索：単語と単語が示す辺や角の行番号の差が 1 行以内であればセットの出現があったと記録する．

単語とその単語が示す辺と角のセットを抽出した例を図 8 に示す．

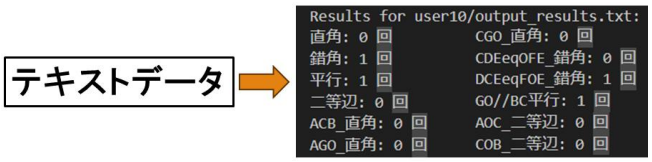


図 8　抽出単語の例

ここで，抽出した単語とその単語に対応する辺や角度のセットと同様の意味を持つ式は同じ特徴量として扱う．例えば，証明の過程において∠ACB が直角であることを明示する場合，「ACB=90°」または「ACB は直角である」という 2 パターン（式で表す，または言葉を使って説明する）がある．これらは表現が違うもののどちらも同じ意味を表しているため，同一の情報を重複して考慮することなく，効果的に特徴を捉えるために 1 つの説明変数としてまとめる．一方で，「直角」，「錯角」，「平行」，「二等辺」の単語単体での出現有無に関しては特徴量として残す．これは，単語単体の出現有無が，OCR の精度によって抽出できなかったセットが存在するデータに対して補足の情報を与えるからである．

4.4.3 OCR を使用した特徴量の生成

4.4.2 項で抽出した文字と式を特徴量として組み合わせる．具体的には，説明変数を表 3 に示されている 18 種類の式と 4.4.2 項で抽出した 4 つの単語を加えた以下の通りである：

(AGeqGC, AOeqOB, OCeqAO, OBeqOC, BCparaOG, ACBright, AGOright, CGOright, OBCeqOCB, COGeqAOG, OBCeqAOG, OCBeqCOG, AOGeqOCB, COGeqOBC, CDEeqOFE, OACeqOCA, CEDeqOEF, DCEeqFOE, 直角, 錯角, 平行, 二等辺)

特徴量の生成手法としては uni-gram を使用した．図 7, 8 を例にすると，特徴量は以下のようになる：

(0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0)

4.5　YOLO を使用して生成した特徴量と OCR を使用して生成した特徴量の統合

4.3 節，4.4 節でそれぞれ図形への書き込み記号と証明文の書き込み文字による特徴量を生成した．精度向上を目的にそれぞれの特徴量を特徴量レベルで統合して 1 つの分類器に学習させるために，以下の 2 パターンで特徴量を統合する．

パターン 1: YOLO を使用して生成した特徴量と OCR を使用して生成した特徴量の結合

1 つ目の特徴量統合手法は，YOLO を使用して生成した特徴量（図への書き込み）と OCR を使用して生成した特徴量（式と単語の情報）をそのまま結合する手法である．このように結合することで両方の特徴量を一度に学習できる．ここで，OCR の特徴量は文字と式に分かれていて，式の部分は YOLO を使用して生成した特徴量の説明変数と同じであるが，パターン 1 ではそれらを別の説明変数として統合する．例えば，問題 2 においては，YOLO を使用した特徴量の説明変数（uni-gram の場合）は表 3 に示されている 18 個あり，OCR を使用した特徴量の説明変数は 22 個（表 3 の 18 個と単語 4 個）あるため，それらをそのまま結合し，40 個の説明変数がある特徴量を生成する．

パターン 2: YOLO を使用して生成した特徴量と OCR を使用して生成した特徴量の統合による各特徴量の精緻化

2 つ目の特徴量統合手法は，YOLO による特徴量と OCR による特徴量において，同じ意味を持つ説明変数を 1 つの説明変数に統合する手法である．このように統合することで，それぞれの特徴量において，検出できなかった記号や式を相互に補完した 1 つの特徴量を作ることができる．問題 2 を例に取ると，YOLO を使用した特徴量の説明変数（uni-gram の場合）と OCR を使用した特徴量の説明変数（4 つの単語以外の部分）は意味が同じであるため，OR を取り，一つの説明変数とする．よって，統合後の説明変数は 22 個（18 個の式と単語 4 個）となる．つまり，図形への書き込み，もしくは，証明文のどちらか一方に出現すれば，当該特徴量が出現したとするような特徴量となる．なお，YOLO による特徴量において，bi-gram と 1-skip-gram でも特徴量を生成しているが，この 2 つは図の書き込み順番を表す特徴量であり，OCR を使用した特徴量に同じ意味を持つ説明変数は無いため，OCR による特徴量にそのまま結合する．

4.6 解答戦略の推定方法

4.6.1 XGBoost によるモデルの学習

4.3.3 項，4.4.3 項で生成した特徴量を基にモデルを学習する．本研究では，先行研究との比較のために先行研究で採用されていた XGBoost を使用して，図形への書き込みを特徴量とする分類器と，証明文から抽出した式と単語を特徴量とする分類器を生成する．各分類モデルは以下の手順で学習を行った．

1．UMAP を使用した次元削減：

作成した特徴量を，UMAP を使用して次元削減する．次元数の決定には，YOLO を使用して生成した 3 種類の特徴量，OCR を使用した 1 種類の特徴量，YOLO と OCR でそれぞれ生成した特徴量を統合させた 6 種類の特徴量をそれぞれ学習させた計 10 種類の分類器毎に異なる次元数（2, 3, 4, 5, 10, 15, 20, 25）でグリッドサーチを行い決定した．

2．MinMax スケーリング：

UMAP で削減されたデータに対して MinMax スケーリングを行い，データを正規化した．

3．オーバーサンプリング：

Random over sampler を使用して，各解法のサンプル数を同数に揃えた．

4.6.2 予測確率の統合

YOLO による特徴量と OCR による特徴量の 2 つの分類器において，各分類器による推定ミスを相互に補うことを目的に，2 つの分類器の結果を統合し，最終的な予測を行った．統合の手法としては，各分類器の予測確率を重み付き和の形で統合した．具体的には，YOLO を使用した特徴量の分類器には重み w を，OCR を使用した特徴量の分類器には 1-w を重み付けし，加算する．最適な重み w を見つけるためには，w を 0.1 から 0.9 までの範囲で 0.025 刻みで変化させ，その中で正解率が最も良かった w をユーザ毎に決定した．

5. データ収集

5.1 収集データの内容と形式

図 1 に示した Web アプリケーションを使用してデータ収集を行う．本アプリケーションでは，図形にストロークが書き込まれる毎に，図形と書き込まれたストロークを組み合わせた画像を PNG 形式で出力する．

5.2 データ収集実験

中学幾何数学問題の解答データを収集するために，早稲田大学の学生 39 名（理系：13 人，文系：26 人）を対象に以下に示すような実験手順でデータ収集実験を行った．

1. 実験手順の確認
2. ペンタブレットの説明
3. 5 分間のチュートリアル
4. 幾何問題 9 問の解答
5. 問題解答後のアンケート（解法のラベル付け）

5.3 データ収集実験の結果

収集した 9 問のデータの内，岡田ら[4][5]の研究で使用されていた図 2 に示す問題のみを使用する．各被験者が本問題に対してアンケートで回答した自身の解法を表 4 に示す．なお，模範解答に含まれていない解法を採用した答案，解法が途中で切り替わった答案，問題が解けなかった答案は「その他」とした．

表 4　解法ごとの答案数

解法	答案数
解法 A	13
解法 B	7
解法 C	14
その他	5

6. 評価実験

収集したデータの内，解法を「その他」と分類したデータと解答が完成していないデータを除外して提案手法の評価実験を実施した．評価実験に用いたデータセットの各解法における答案数を表 5 に，特徴量毎に作成した分類器の予測性能を表 6 にそれぞれ示す．

表 5　評価実験に使用したデータセットの答案数

解法	答案数
解法 A	7
解法 B	4
解法 C	13

表 6　各分類器の予測性能

特徴量の説明		特徴量	正解率	適合率	再現率	F1 値
図中への書き込み		Uni-gram	75.0%	73.9%	66.5%	68.9%
		Uni&bi-gram	79.2%	83.8%	72.6%	73.9%
		Uni&bi&skip-gram	66.7%	63.6%	63.6%	63.6%
証明文への書き込み		OCR	91.7%	95.6%	83.3%	86.5%
図中への書き込みと証明文への書き込みの特徴量の統合	パターン 1	(OCR)+(Uni-gram)	70.8%	78.6%	60.3%	61.7%
		(OCR)+(Uni&bi-gram)	70.8%	78.6%	60.3%	61.7%
		(OCR)+(Uni&bi&skip-gram)	87.5%	86.7%	84.3%	85.9%
	パターン 2	(OCR)or(Uni-gram)	75.0%	72.2%	68.7%	70.0%
		(OCR)or(Uni&bi-gram)	83.3%	89.2%	72.4%	73.0%
		(OCR)or(Uni&bi&skip-gram)	75.0%	68.9%	65.1%	66.5%
予測確率を重み付き和の形で統合		(OCR)×(Uni-gram)	87.5%	93.8%	75.0%	76.6%
		(OCR)×(Uni&bi-gram)	91.7%	95.6%	83.3%	86.5%
		(OCR)×(Uni&bi&skip-gram)	87.5%	93.8%	75.0%	76.6%

なお，従来の方法（図形への書き込みを特徴量とした手法）では n-gram を使用して特徴量を作成しており，uni-gram，uni-gram と bi-gram を組み合わせた uni&bi&gram，さらに 1-skip-gram を加えた uni&bi&skip-gram の 3 つがある．証明文の情報のみを特徴量として使用した分類器は OCR と記載する．図形への書き込みを使用した特徴量と証明文への書き込みを使用した特徴量を結合したもの（パターン 1）を「(OCR)+(Uni-gram)」のように「+」で表記し，統合による各特徴量の精緻化を行ったもの（パターン 2）を「(OCR)or(Uni-gram)」のように「or」で表記する．図形への書き込みを使用した特徴量で学習した分類器と証明文への書き込みを使用した特徴量で学習した分類器がそれぞれ出力する予測確率を重みつき和の形で統合して予測を行ったものに関しては「(OCR)×(Uni-gram)」のように「×」で表記する．

表 6 より，OCR を使用した分類器が最も高い正解率を示していることが分かる．これは，正解した答案において，証明文の中に，解答者が採用した解法が詳細に記述されているからであると考えられる．一方で，図形への書き込みも活用した手法（特徴量の統合，予測確率の統合）は有効ではないという結果を示した．証明文に書き込む文字は必要最低限の文字であるのに対して，図形に書き込まれた記号は問題を解くにあたって全て使用しているわけでは無いため，特徴量の統合は単に余分な情報を与えるだけであるためと考えられる．予測確率の統合は，YOLO を使用した分類器の予測確率の分布が 50-60%に多く分布しているため，OCR を使用した分類器の結果を覆すことができず，精度が向上しなかったと考えられる．

7. おわりに

本稿では，学習者に対して個別最適な学習環境を提供することを目的として，証明を書き込んだ解答欄のオンライン手書き解答データを使用して解答戦略の自動推定を行った．先行研究では，図形への書き込みが無い場合に，解答戦略を推定するのが難しいという課題があった．そこで，本研究では，岡田ら[4][5]の手法による課題を解決するために従来の手法に加えて，証明文を書き込んだ解答欄の画像から，特徴量となる式や単語を抽出し，分類器を作成した．また，精度向上を図るために，解答欄に書き込まれた証明の手書きデータだけでなく，図中に書き込まれた記号の解析も活用し，さらなる精度向上に挑戦した．

従来の手法では，3 種類の解法がある問題で正解率が 79.2%，F 値 73.9%であったのに対し，OCR を使用して解答欄に書き込まれた証明の手書きデータを特徴量として学習させる手法では，正解率 91.7%，F 値 86.5%を達成した．図中に書き込まれた記号と組み合わせた手法では精度の向上が見られなかった．

今後の課題としては，答案が完成されていないデータに対して YOLO を使用した分類器と OCR を使用した分類器の 2 つの分類器を上手く統合させることやリアルタイムでの解法推定が挙げられる．また，解法が移り変わっている答案に対して解法の推定を行うことも今後の課題として挙げられる．

謝辞

本研究の一部は，JSPS 科研費 20H04292 の助成によるものである

参 考 文 献

[1] S. E. Goh, “Guided problem solving at primary level: students’ strategies for solving challenging geometry problems”, Master Thesis of Education, Nanyang Technological University, Singapore, 2018.

[2] M. Cirillo, and J. Hummer, “Competencies and behaviors observed when students solve geometry proof problems: an interview study with smartpen technology”, ZDM Mathematics Education, 53, pp.861–875, 2021.

[3] 三浦将人，工藤雅士，岡田一洸，中山祐貴，山名早人，“オンライン手書きデータを用いた定着問題の自動推定－幾何問題を対象として－”，DEIM2022 第 14 回データ工学と情報マネジメントに関するフォーラム，pp.1-8，2022.

[4] 岡田一洸，工藤雅士，工藤雅士，山名早人，“オンライン手書きデータを用いた解答戦略の自動推定-幾何問題を題材として-”DEIM2023 第 15 回データ工学と情報マネジメントに関するフォーラム，pp.1-8，2023.

[5] K. Okada，M. Kudo, and H. Yamana, “Estimating Answer Strategies using Online Handwritten Data: A Study using Geometry Problems”, in Proc. of the 15th Int’l Conf. on Education Technology and Computers (ICETC 2023), pp.1-7 (2023.9)

[6] 林 凱博，宮崎公彦，工藤雅士，山名早人: 学習履歴データを用いた学習状態推定に関するサーベイ，信学技報, ET2023-13, pp.1-8 (2023.9)

[7] D. Yoshitake, B. Flanagan, and H. Ogata, “Supporting Group Learning Using Pen Stroke Data Analytics”, In Proc. of the 28th Int’l Conf. on Computers in Education, Asia Pacific Society for Computers in Education, pp.634-639, 2020.

[8] 村上統馬，山名早人．“オンライン手書きデータを用いた幾何学図形問題に対する答案の自動分類と解答者の理解度推定”，早稲田大学基幹理工学部情報理工学科卒業論文, 2019.

[9] 宮崎公彦，青柳宏紀，岡田一洸，山名早人，“眼球運動を用いた解答戦略有無の判定-数学問題を対象として-” DEIM2023 第 15 回データ工学と情報マネジメントに関するフォーラム，pp.1-8, 2023.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, in Proc. of 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788, 2016.

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

2024年2月28日(水) 16:00 ~ 18:10 | T5-B オンライン（Zoom Events)

地理情報分析・シミュレーション

座長:大塚 真吾(神奈川工科大学)
コメンテータ:義久 智樹(大阪大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T5-B-3-01] 平均二乗誤差を利用した明るさが変動する天体の周期解析手法

*小野 晴樹[1]、中西 宏和[1]、藤原 靖宏[2]、鬼塚 真[1] (1. 大阪大学大学院情報科学研究科マルチメディア工学専攻鬼塚研究室、2. NTTコミュニケーション科学基礎研究所)

16:30 ~ 16:55
[T5-B-3-02] 風況を考慮したEVのエネルギーベースラインマップの作成と
長期間実走行データによる精度検証

*劉 屹[1]、曽我 紗代子[1]、何 新[1]、田中 悠斗[1]、富井 尚志[2] (1. 横浜国立大学　大学院環境情報学府　情報環境学専攻、2. 横浜国立大学　大学院環境情報研究院)

16:55 ~ 17:20
[T5-B-3-03] PVとEVを統合する負荷平準指向VGIDBを用いたスマートグリッド導入効果の多角的定量化

*廣居 樹[1]、片平 昇輝[1]、鈴木 博人[1]、小松 虎太郎[2]、本藤 祐樹[3]、富井 尚志[3] (1. 横浜国立大学 大学院環境情報学府 情報環境専攻、2. 横浜国立大学 理工学部 数物・電子情報系学科、3. 横浜国立大学 大学院環境情報研究院)

17:20 ~ 17:45
[T5-B-3-04] 水害時における市民の避難行動に関する意識の分析

*石井 樹[1]、妹尾 考[2]、関 洋平[3]、川島 宏一[4]、鈴木 誉幸[5] (1. 筑波大学大学院理工情報生命学術院システム情報工学研究群社会工学学位プログラム博士後期課程、2. 筑波大学大学院人間総合科学学術院人間総合科学研究群情報学学位プログラム博士前期課程、3. 筑波大学図書館情報メディア系、4. 筑波大学システム情報系、5. 筑波大学大学院理工情報生命学術院システム情報工学研究群社会工学学位プログラム博士前期課程)

17:45 ~ 18:10
[T5-B-3-05] 人口動態を基盤としたソーシャルセンサのパフォーマンス分析

*大西 駿太朗[1]、矢田 竣太郎[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室)

平均二乗誤差を利用した明るさが変動する天体の周期解析手法

小野　晴樹† 　中西　宏和† 　藤原　靖宏†† 　鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565-0871 大阪府吹田市山田丘 1-5
†† NTT コミュニケーション科学基礎研究所　〒 243-0198 神奈川県厚木市森の里若宮 3-1
E-mail: †{ono.haruki,nakanishi.hirokazu,onizuka}@ist.osaka-u.ac.jp, ††yasuhiro.fujiwara@ntt.com

あらまし　変動天体は宇宙の謎を解き明かす重要な情報を提供する．その変動天体の周期解析の既存手法としては Stringlength 法と Lomb-scargle periodogram 法が主流である．しかし，Stringlength 法は異常値が周期解析の精度に与える影響が大きい，また計算コストが大きいという問題がある．また Lomb-scargle periodogram 法はスパースな天体データに対しては精度が低いという問題がある．また共通の問題として，離散的な候補周期で周期解析を行うため，適した候補周期を選択できないという問題がある．これらの問題を解決するため，本稿では，効率的でより適切な候補周期の設定と平均二乗誤差を用いた周期解析手法を提案する．具体的には，計算コストを小さくするために計算コストの小さい LombScargle 法を用いて候補周期の絞り込みを行い，絞り込んだ候補周期をスプライン補間を用いて連続値の候補周期に変換を行う．それらの候補周期に対して平均二乗誤差を用いた周期解析を行い，予測周期を決定する．実験では実際の天体データに対して周期解析を行い，既存手法よりも周期特定の精度が向上し，既存手法と同程度の実行時間であることを確認した．

キーワード　周期解析

1　はじめに

変動天体とは明るさが時間とともに変化する天体であり，i) 位置が変動する天体，ii) 光度が変動する天体の 2 種類存在する．変動については，長期的なものもあれば，局所的なもの，それらが混在しているものに加え，一過性であるものや周期性を持つものもなど多岐にわたる [1]．フレア星，新星，超新星，セファイド星，こと座 RR 型などはその代表例である．変動天体の多くは周期的に明るさが変化しており，これは宇宙の構造や進化の解明において役立つ重要な現象である．具体的には，変動天体を調査することによって，巨大なブラックホールが引き起こす質量吸収について調べることができる [2–5]．さらに，変動天体の周期は天体までの距離などの重要な情報を提供することがあり，これによって近くの暗い銀河の特定や銀河構造の調査につながる [6–9]．したがって，多くの研究者が変動天体について調査している．一方で大型天体望遠鏡によって変動天体を含んだ膨大なデータセットを生成され続けており，迅速かつ正確に天体の周期性を特定し変動天体を選別することは宇宙の飛躍的な解明において不可欠だといえる．

周波数特定の典型的な方法として広く使われてきたのはフーリエ変換である．フーリエ変換は時間領域を周波数領域に変換し、周波数ごとの成分に分解することができる．周期的な信号に対しては特定の周波数の成分の割合が大きくなり，周期を特定することができる．しかし，天体の観測データに対してはフーリエ変換は有効ではない．なぜなら天体の観測データには多数の欠損値が含まれており，非常にスパースなためである．図 1 にある天体の時系列データの例を示す．天体の観測データは天候の影響であったり，他の天体の観測に用いられていたりなどによって観測を行うことに制限があることが多いため，このような非一様サンプリングなデータになってしまうことは避けられない．また観測値は大気によるノイズ，観測機器によるノイズ，他の恒星によるノイズなど様々なノイズの影響を受けている．これらの理由によって，変動天体の周期特定においてフーリエ変換は有効的な手段ではない．そんななかで主流な天体の周期解析手法として，Stinglength 法 [10] と Lomb-scargle perodogram 法 [11–13] がある．Stringlength 法はスパース性の問題を解決することができる手法であるが，異常値による周期解析の精度への影響が大きいといった問題がある．Lomb-scargle perodogram 法は計算コストを抑えて周期解析を行うことができる手法あるが，スパースなデータに対しての精度が他の手法よりも低いといった問題がある．また，いずれの手法においても，全周期を細かく探索するため計算コストが大きいこと，離散的な候補周期で探索するため適した候補周期を選択できないことという問題がある．

本研究では，上記の問題を解決すべく，Folding 法と平均二乗誤差を利用した周期解析手法を提案する．Folding 法とは，Stinglength 法で用いられている技術であり，スパース性の問題を解決する役割を担っている．本提案手法は 3 部構成になっている，第一部では周期候補の絞り込む．具体的には Lomb-scargle perodogram 法を用いて候補周期全体に対して周期解析を行い，周波数領域への変換を行う．その結果においてパワースペクトルが高い上位周期を決定する．第 2 部では連続値による候補周期の決定を行う．具体的には，第一部で決定した上位周期とその周辺の候補周期に対して，自然スプライン補間を用いて離散値の候補周期を連続値に変換し，新たな候補周期を決定する．第 3 部では平均二乗誤差を用いた周期解析を行う．具体的には，第 2 部で決定した各候補周期に対して Folding 法を行い，折り

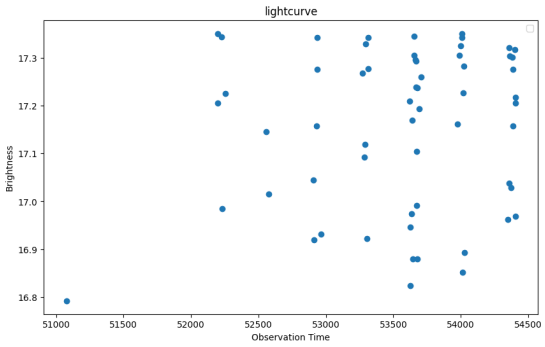


図 1: 天体の時系列観測データの例の図．横軸は観測日時 (日)，縦軸は天体の明るさ (等級) を表している．非常に長い期間で非一様間隔でのサンプリングが行われており，非常にスパースなデータとなっている．

込んだグラフを作成する．そのグラフに対して多項式近似による近似曲線を引き，グラフと近似曲線の平均二乗誤差を計算することによってその候補周期のスコアリングを行い，予測周期を決定している．

本稿の構成は以下の通りである． 2 で関連研究について説明し， 3 章で事前準備を行い， 4 章で提案手法について詳細に述べた後， 5 章で実験結果を示し，考察を行う．最後に， 6 章で結論を述べる．

2 関 連 研 究

本研究では明るさが変動する天体の周期解析を行っている．一般的にデータの周期解析を行う手法として広く使われてきたのはフーリエ変換である．しかし，天体の観測データは非一様サンプリングであり，ノイズの影響を大きく受けているため，フーリエ変換は有効ではない．このような天体データに対してより効果的な周期解析手法の開発が研究されている．主流な手法として Stinglength 法と Lomb-Scargle Perodogram 法が挙げられる．

Stinglength 法 [10] は時系列データを適切に並び変える Folding を行い，各観測点間のユークリッド距離の総和でスコアリングを行う手法である．Folding を行うことによってデータのスパース性の問題を解決するができる手法であるが，異常値の存在が周期解析の精度に影響を与えるという課題がある．

Lomb-Scargle Perodogram 法 [11–13] は非一様サンプリングにも適応できるようにフーリエ変換を拡張した手法である．計算コストを抑えつつ周期解析を行うことが可能だが、スパースなデータに対する精度が他の手法よりも低いという問題がある．

また，これらの従来の手法にはいくつかの共通の課題がある．まず，全周期を細かく探索するために必要な計算コストが大きいことが挙げられる．さらに，離散的な候補周期で探索するため，適切な候補周期を選択することが難しいという点も課題である．

さらに，観測データの品質向上やデータ収集方法の改善により，周期解析の精度向上に貢献する研究も行われている．これらの取り組みにより，天体の周期解析における課題の解決や新たな洞察の獲得が期待されている．

3 事 前 準 備

本章では提案手法の基礎となる手法について説明する．

3.1 Folding 法

天体の周期性を確認するのによく使われる Folding 法という技術について説明する.Folding 法とはある周期 (以降試行周期と呼ぶ) に対する位相情報を抽出するように観測データの時間情報を変換し並べなおす技術である．各天体のデータの観測日時を t，試行周期を P_c とすると，Folding したあとの試行周期に対する位相 φ は以下の式のように表される．

$$\varphi = \frac{t \bmod P_c}{P_c} \tag{1}$$

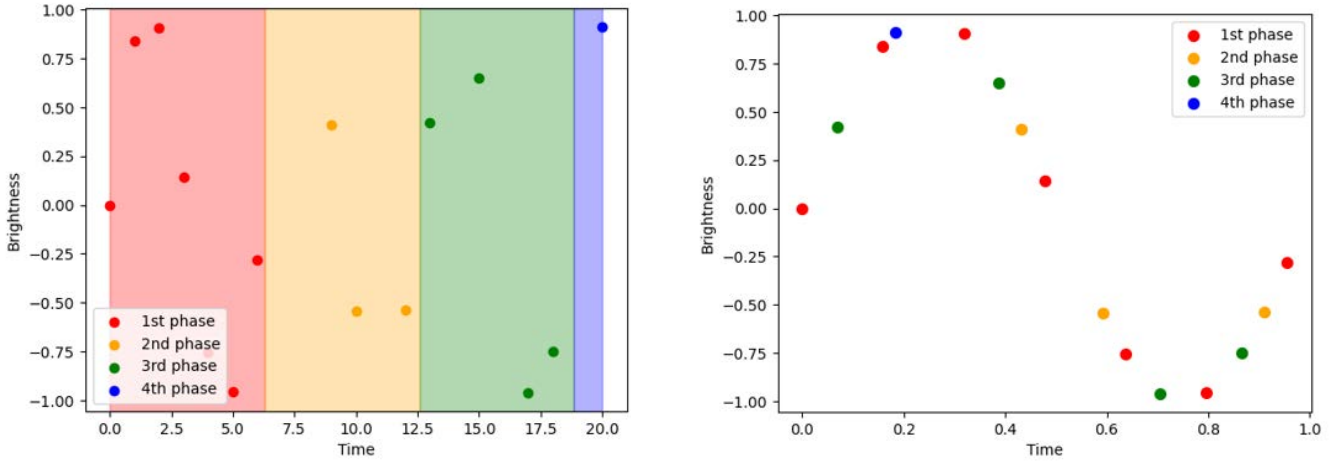
具体的に Folding 法の手順について説明する．まずはじめに，天体の時系列データを試行周期で等間隔に分割する．その次に分割した区間をすべて一つに重ね合わせる．縦軸は明るさのまま横軸を phase とし試行周期を 1 として重ねることによってグラフを作成している．その例を図 3 に示す．試行周期がその天体の正しい周期にかなり近いとき，グラフは一波長分の波に近い形となりまとまった形になる．反対に試行周期が正解周期に近い周期ではない場合は非常にばらついた形になる．

3.2 Stringlength 法

Stringlength 法 [10] は天体の周期解析の手法として広く使われてきた手法である．その手法について具体的に説明する．はじめに候補周期の一つ一つを試行周期として 2.1 で説明した Folding 法を行う．次にその Folding したグラフの各観測点ごとに隣接する観測点間のユークリッド距離の総和を計算する．同様の計算を候補周期すべてに対して行い，その和が最小となった周期を最良の周期とする手法である．この手法はデータ観測日時のスパースさの影響を受けることなく周期解析を行うことができ，図 1 のように非常にスパースなデータである天体のデータを扱うことに適しているといえる．しかし，異常値がある場合，一つの異常値に対してその両隣の観測点との距離がスコアリングに影響してしまう．そのため異常値があることでによる周期解析の精度への影響が大きいという問題がある．また全周期を細かく探索するため，計算コストが大きいこと．候補周期の周期値を離散的に探索するため，適した候補周期を選択できないという問題がある．

3.3 Lomb-scargle periodogram 法

Lomb–Scargle periodogram 法 [11–13] は，不均一にサンプリングされたデータの周期的な信号を検出して特性評価するためのよく知られたアルゴリズムであり，天文以外にも自動車業界、通信および医学など様々な分野で用いられている手法である．時間 t_k に取得される N 個 の観測値 x_k $(k = 1,.., N,)$ に対して Lomb-Scargle periodogram 法は以下の式によって定義



(a) 非一様サンプリング時系列データの例　　**(b)** Folding したグラフの例

図 2: Folding 法の例の図．グラフの横軸は時間，縦軸は天体の明るさを表している．まずはじめに観測データを試行周期で等間隔に分割する．分割した各区間を重ね合わせる．図では y=sin(x) の非一様サンプリングのグラフである．正解周期である 2 πで区切って重ね合わせることによってきれいに 1 周期分の波になる．逆に試行周期が正解周期から離れているときは元の非一様サンプリングのグラフのようにバラバラのままである．

されている．

$$P_{LS}(f) = \frac{1}{2\sigma^2}\left[\frac{\left(\sum_{k=1}^{N}(x_k-\bar{x})\cos(2\pi f(t_k-\tau))\right)^2}{\sum_{k=1}^{N}\cos^2(2\pi f(t_k-\tau))}\cdot + \cdot\frac{\left(\sum_{k=1}^{N}(x_k-\bar{x})\sin(2\pi f(t_k-\tau))\right)^2}{\sum_{k=1}^{N}\sin^2(2\pi f(t_k-\tau))}\right], \tag{2}$$

ただし $\bar{x}$ と σ^2 はそれぞれ観測値の平均と分散を表しており，τ は以下の式を満たす値である．

$$\tan(2(2\pi f)\tau) = \frac{\sum_{k=1}^{N}\sin(2(2\pi f)t_k)}{\sum_{k=1}^{N}\cos(2(2\pi f)t_k)} \tag{3}$$

これらの式を使って周波数 (周期) に対してのパワースペクトルを計算することができる．Lomb–Scargle periodogram 法は候補周期の数を膨大にする必要がなく，計算コストを抑えたまま周期解析ができる手法である一方，データのスパースさの影響によって精度が大きく下がってしまうという問題がある．

4 提 案 手 法

4.1 提案手法モデル

提案手法の概要図を図 3 に示す．

本章では，変動天体の周期解析の提案モデルについて説明する．提案手法は以下の 3 つの STEP から構成されている．

- 候補周期の絞り込み：LombScargle 法を用いて天体の時系列データを周波数領域へ変換し，その結果のパワースペクトルが高い上位周期を決定する．
- 連続値による新たな候補周期の決定：上位周期の周辺の候補周期とそのパワースペクトルに対して自然スプライン補間を用いて近似曲線を引き，連続値に変換して新たな候補周期を決定する．
- 平均二乗誤差を用いた周期解析：各候補周期に対して Folding 法を行い、時系列データをグラフに変換する．その Folding 法を行ったグラフに対して多項式近似を用いて近似曲線を引き，グラフと近似曲線の平均二乗誤差を計算する．平均二乗誤差が最も小さかった候補周期を予測周期とする．

4.2 候補周期の絞り込み

周期解析を行う際，その候補となる周期を設定しておく必要がある．変動天体ではその周期が数時間程度のものから数百日のものまで幅広く存在している．そのため変動天体における周期解析では候補周期の数が膨大になってしまう．周期解析はその膨大な候補周期の一つ一つに対して処理を行うため，計算コストが大きくなってしまうという問題がある．この問題を解決するために，計算コストが比較的小さな周期解析手法を用いて膨大な候補周期の中から正解周期の可能性が高い k 個の候補周期に絞り込むことを考えた．その具体的な手順について述べる．まず, 天体の時系列データに対して LombScargle 法と用いて周期解析を行う．これによって時系列の明るさの変動のグラフを周波数領域のパワースペクトルのグラフに変換することができる．次にそのパワースペクトルのグラフの結果からパワースペクトルが高い順に k 個の候補周期（以下上位周期）を決定する．ここで LombScargle 法を用いた理由としては，計算コストが小さいためである．

4.3 連続値による新たな候補周期の決定

候補周期を設定する際，その候補周期の周期値は離散的になってしまうことは避けられない．これによって適切な候補周期を設定できず，正解周期を見逃してしまうという問題がある．この問題を解決するために，近似曲線を用いて候補周期を連続

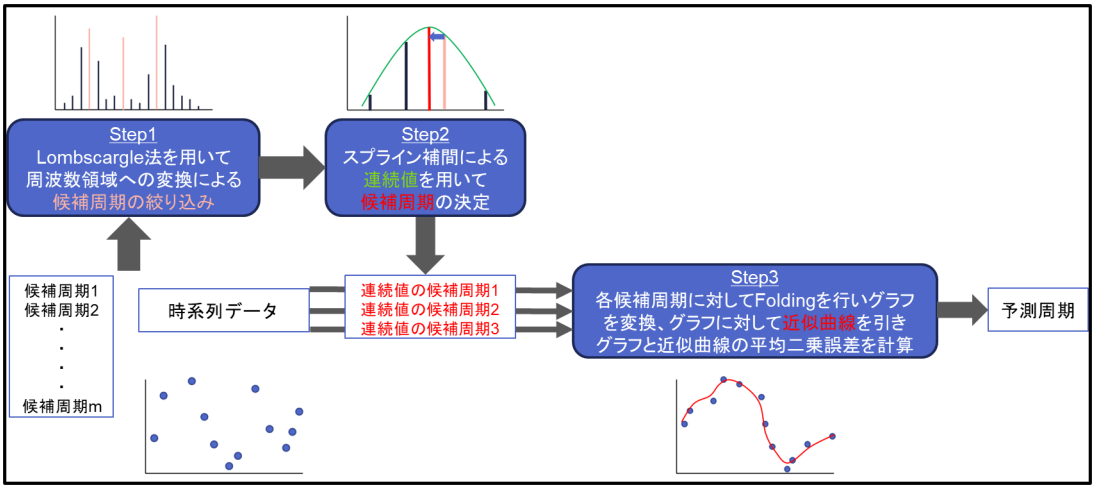


図 3: 提案手法のモデル概要図

値に近似することを考えた．その具体的な手順について述べる．3.2 の出力である上位周期とその周辺の周期とそのパワースペクトルの値に対して自然スプライン補間を用いて近似曲線を引く．これによって候補周期を連続値に変換し，その局所的な曲線においてパワースペクトルが最大となる候補周期を新たな候補周期とする．ここで，近似曲線を引く際に自然スプライン補間を用いた理由を説明する．自然スプライン補間はデータのポイント間を滑らかに結ぶ曲線を生成することができ，低次の多項式を用いるため計算コストが小さい，すべての点を通る曲線を引くためデータの局所的な表現に適しているといった特徴がある，他の候補としてはラグランシュ補間や最小二乗法に基づく回帰曲線があるが，前者は高次の多項式を用いるため計算コストが大きい，後者はすべての点を通るとは限らない曲線を引くためデータの局所的な表現には適していない．

4.4 平均二乗誤差を用いた周期解析

Stringlength 法では Folding 法を行ったあと，各観測点間のユークリッド距離の総和を計算し，その総和が最小となる候補周期を予測周期としている．しかし異常値がある場合，異常値とその異常値の両隣との距離に影響を与えてしまうため，異常値がスコアリングに与える影響が大きくなっていまい，周期解析の精度が下がってしまうという問題がある．この問題を解決するために，異常値がスコアリングに与える影響がより小さいスコアリング手法を用いることを考えた．その具体的な手順について述べる．まず各候補周期に対して Folding 法を行い，天体の時系列データを Folding 法を行ったグラフ（以下折り込みグラフ）に変換する．その折り込みグラフに対して多項式近似を用いて近似曲線を引き，グラフと近似曲線の平均二乗誤差を計算する．平均二乗誤差が最小となる候補周期を予測周期とする．ここで平均二乗誤差を用いた理由について説明する．平均二乗誤差とは曲線とグラフの y 座標の誤差の二乗平均である，異常値がある場合，異常値と曲線の y 座標の誤差のみに影響を与えるため，異常値がスコアリングに与える影響が小さくなる．

このようにして近似曲線を作成し，その近似曲線と Folodingしたグラフの平均最小二乗誤差 (mse) を計算する．平均最小二乗誤差の大きさは Folding したグラフが近似曲線からどのくらいずれているのかを表している．正解周期に近い周期で Folding したグラフが波長に似た形でまとまるため平均最小二乗誤差が小さくなり，逆に正解周期に近くない周期で Folding したグラフでは平均最小二乗誤差が大きくなる．Stringlength 法も Folding したグラフを用いて周期解析を行っているが，長い周期の方を最良の周期と判定してしまうという問題がある．平均最小二乗誤差を用いたこの手法ではその問題を解決することを期待する．

4.5 アルゴリズム

Algorithm 1 平均二乗誤差を用いた周期解析

Input: i) 候補周期リスト: *period_ls*
ii) 観測時刻リスト: *times*
iii) 観測した光度リスト: *brightness*
Output: 予測周期: *best_period*
// 各周期についての処理
mse_ls ← [] // 平均二乗誤差のリスト
for 各試行周期 in *period_ls* **do**
    // Folding 処理
    相関 = (*time* を *trial_period* で割った余り) / *trial_period*
    並べたインデックス = np.argsort(相関)
    折り返し時刻 = 相関 [並べたインデックス]
    折り返し光度 = *brightness*[並べたインデックス]
    // 多項式近似による曲線を作成
    $y_fit \leftarrow fitting_lightcurve(trial_period, times_folded, brightness_folded)$
    // 最小二乗誤差の計算
    mse_ls.append(np.mean($(brightness_folded - y_fit)^2$))
end for
// 平均最小二乗誤差が最小となる候補周期を選択
min_index ← 最小値のインデックス (*mse_ls*)
best_period ← *candidate_period_ls*[*min_index*]
return *best_period*

Algorithm 1 に STEP3 の平均二乗誤差を用いた周期解析を

表 1: データセットの詳細．天体数と平均観測期間，平均観測回数

dataset	Objects	Average Observation Period(day)	Average Number of Observations
RR Lyrae	483	2872.7	56.2
ASAS	46712	1386.1	517.5
CSDS	37745	2977.0	223.9

図 4: 周期解析における周期特定精度の比較

	String length	Lomb-Scargle	提案手法
RR Lyrae	0.369	0.383	0.569
ASAS	0.106	0.624	0.822
CSDS	0.049	0.122	0.163

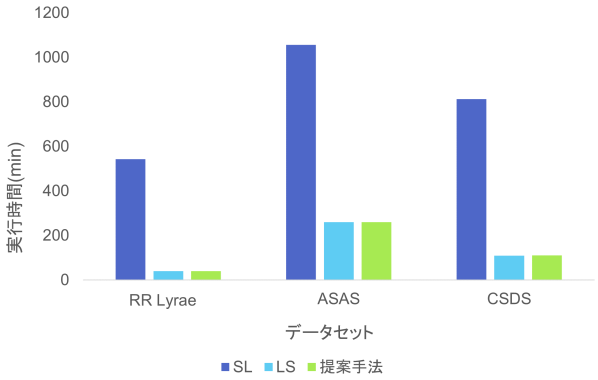


図 5: 周期解析における実行時間の比較

表現した疑似コードを掲載する．入力は STEP1 と STEP2 で決定した候補周期と天体の観測時刻と明るさの時系列データである．候補周期の各周期に対して Folding を行い，観測点を並べ替える (7〜10 行目)．次に Folding を行ったグラフに対して多項式近似を用いて近似曲線を作成し，グラフと近似曲線の平均二乗誤差を計算する (11〜15 行目)．各周期に対して計算した平均二乗誤差を比較して最小となる周期を予測周期として出力する (16〜19 行目)．

5 評価実験

本章では，以下の二つの実験を行う．一つ目の実験は実際の天体データをに対して周期解析を行い，提案手法と既存手法の周期特定精度と効率性について比較を行う．二つ目の実験は提案手法の各ステップを変えて周期特定精度の変化を見る．

5.1 データセット

本実験では RR Lyrae，ASAS，CSDS といったの異なる調査から得られた 3 つのデータセットを用いる．RR Lyrae [6] は 483 天体の光度曲線についてのデータであり，Adelman-McCarthy によって 2008 年に作成されたものである．ASAS Catalog of Variable Stars [14] は 46712 天体の光度曲線についてのデータであり，ojmanski、Pilecki、および Szczygiel らによって 2005 に作成されたものである．CSDS (The Catalina Surveys Periodic Variable Star) Catalog [15] は AJ Drake, MJ Graham らによって 2014 年に作成されたものである．3 つのデータセットにおいて，正解周期は各天体に一つずつあり，ハーバードタイムシリーズセンターで決められたものを用いており，AOV アルゴリズムを利用し，専門家によって視覚的に確認された周期を真の値と見なしている．

5.2 実験設定

実験では 3 つのデータセットから観測日時とその日時の明るさを抽出し，各天体についての明るさの時間変化を示す光度曲線を作成している．その光度曲線に対して周期解析を行い，各天体の明るさが変動する周期の特定を試みる．周期解析を行う際，予測周期の候補となる周期（以下候補周期）を設定する必要がある．今回の実験では候補周期は 0〜200(日) の範囲を 1 × 10E-4 間隔で分割した各値とした．

精度評価については，各天体について周期解析によって出力された予測周期と真の周期の差が真の周期の誤差が 0.5% 以内であればその天体の周期を正しく特定することができたとする．Recall@1 は正しく周期を特定できた天体の割合である．

5.3 比較手法

本実験で用いるそれぞれの比較手法について説明する．

Stinglength 法 (SL): Folding を用いて時系列データを並べ替えたあとに各観測点間のユークリッド距離の総和をスコアリングに用いた手法

LombScargle 法 (LS): 非一様サンプリングにも適応できるようにフーリエ変換を拡張した手法

LS+ MSE: 提案手法に対して，STEP2 のスプライン補間を用いた連続値による候補周期に決定を行わなかった場合の手法

LS + spline+ SL: 提案手法に対して，STEP3 のスコアリングを平均二乗誤差 (MSE) ではなく各観測点間のユークリッド距離の総和で行った手法

5.4 結果と分析

表 4 は実際のデータに対して，既存手法及び提案手法の周期特定の精度をまとめたものである．いずれのデータセットに対しても，提案手法は既存手法と比較して，最も高い精度を達成

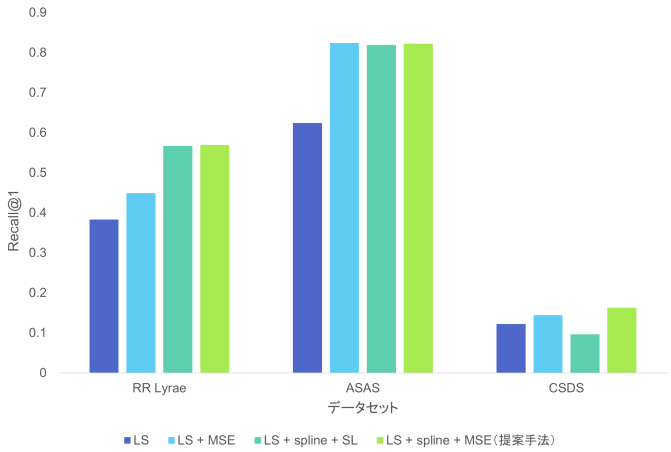


図 6: 各ステップの周期精度の比較

した．

また，図 5 は既存手法及び提案手法の実行時間をまとめたものである．ただし，ASAS と CSDS はランダムに 500 天体サンプリングし，周期解析を行った結果になっている．提案手法は Stringlength 法と比較して大幅な実行時間の短縮を達成し，LombScargle 法と同程度の実行時間となった．データセットによって実行時間が異なる主な原因は 1 天体当たりの観測データの数の違いである．また，提案手法は LombScargle 法を STEP1 として用いているため，STEP2 と STEP3 の分 LombScargle 法よりも実行時間が長くなってしまうが，STEP1 の実行時間は STEP2 と STEP 3 の実行時間より非常に長いため，LombScargle 法と提案手法は同程度の実行時間を達成している．また Stringlength 法と提案手法はともに各候補周期に対して Folding を行い，スコアリングを行っているが，提案手法では LombScargle 法を用いて候補周期の絞り込みを行っているため、大幅な実行時間の短縮を達成している．

また図 6 は提案手法に対して各 STEP を変更した手法と提案手法の周期特定の精度をまとめたものである．CSDS に対しての LS+spline+SL 手法を除いて，LombScargle 法のみの場合の精度は最も低くなっている．CSDS に対しての LS+spline+SL 手法は Stringlength 法を用いており，ユークリッド距離の総和をスコアリングとして予測周期を決定しているため，表 4 の結果と同様に精度が低くなったと考えられる．また STEP2 のスプライン補間を用いて連続値による候補周期の決定に関しては，データセット RR Lyrae と CSDS において，STEP2 を行った場合の方が精度が向上した．ASAS では STEP2 を行わない場合の方がわずかに精度が高いという結果が得られた．これに関しては，STEP 2 による候補周期の決定の効果が小さいことを示している．この要因としては真の周期が長い天体が多いため，STEP2 の周期候補の変化が相対的に小さくなったためだと考えられる．

6 終わりに

本稿では，明るさが変動する天体の周期解析手法を提案した．提案手法では，計算コストを小さくするために計算コストの小さい LombScargle 法を用いて候補周期の絞り込みを行い，絞り込んだ候補周期をスプライン補間を用いて連続値の候補周期に変換を行った．それらの候補周期に対して平均二乗誤差を用いた周期解析を行った．実験では，実際の天体データに対して周期解析を行った場合，既存手法よりも高い精度を達成し，同程度の実行時間であることを確認した．

今後の課題としては，周期解析の精度向上のためにパラメータの調整を行ったり，他の手法を導入したりすること，周期解析の結果を利用して変動天体検出を行うことなどが考えられる．

謝　　辞

本研究は JSPS 科研費 JP20H00583 の助成を受けたものです。

文　　献

[1] AC Becker, JJ Bochanski, SL Hawley, Ž Ivezić, AF Kowalski, B Sesar, and AA West. Periodic variability of low-mass stars in sloan digital sky survey stripe 82. *The Astrophysical Journal*, Vol. 731, No. 1, p. 17, 2011.

[2] Natalie M Gosnell, Michael A Gully-Santiago, Emily M Leiner, and Benjamin M Tofflemire. Observationally constraining the starspot properties of magnetically active m67 sub-subgiant s1063. *The Astrophysical Journal*, Vol. 925, No. 1, p. 5, 2022.

[3] Tang Yan-Ke, Gai Ning, Li Zhi-Kai, Yang Hai-Lian, and Dong Wen-Hui. Research on periodicity of single sector variable star of tess space satellite. *Chinese Astronomy and Astrophysics*, Vol. 46, No. 1, pp. 63–81, 2022.

[4] L Eyer and C Blake. Automated classification of variable stars for all-sky automated survey 1–2 data. *Monthly Notices of the Royal Astronomical Society*, Vol. 358, No. 1, pp. 30–38, 2005.

[5] Eran O Ofek, Maayane Soumagnac, Guy Nir, Avishay Gal-Yam, Peter Nugent, Frank Masci, and Shri R Kulkarni. A catalogue of over 10 million variable source candidates in ztf data release 1. *Monthly Notices of the Royal Astronomical Society*, Vol. 499, No. 4, pp. 5782–5790, 2020.

[6] Branimir Sesar, Željko Ivezić, Skyler H Grammer, Dylan P Morgan, Andrew C Becker, Mario Jurić, Nathan De Lee, James Annis, Timothy C Beers, Xiaohui Fan, et al. Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82. *The Astrophysical Journal*, Vol. 708, No. 1, p. 717, 2009.

[7] VF Braga, J Crestani, M Fabrizio, G Bono, C Sneden, GW Preston, J Storm, S Kamann, M Latour, H Lala, et al. On the use of field rr lyrae as galactic probes. v. optical and radial velocity curve templates. *The Astrophysical Journal*, Vol. 919, No. 2, p. 85, 2021.

[8] Gaochao Liu, Yang Huang, Sarah Ann Bird, Huawei Zhang, Fei Wang, and Haijun Tian. Probing the galactic halo with rr lyrae stars- iii. the chemical and kinematic properties of the stellar halo. *Monthly Notices of the Royal Astronomical Society*, Vol. 517, No. 2, pp. 2787–2800, 2022.

[9] A Katherina Vivas, Clara E Martínez-Vázquez, Alistair R Walker, Vasily Belokurov, Ting S Li, and Denis Erkal. Variable stars in the giant satellite galaxy antlia 2. *The Astrophysical Journal*, Vol. 926, No. 1, p. 78, 2022.

[10] D Clarke. String/rope length methods using the lafler-kinman statistic. *Astronomy & Astrophysics*, Vol. 386, No. 2, pp. 763–774, 2002.

[11] Nicholas R Lomb. Least-squares frequency analysis of

unequally spaced data. *Astrophysics and space science*, Vol. 39, pp. 447–462, 1976.

[12] Jacob T VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, Vol. 236, No. 1, p. 16, 2018.

[13] Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. Astrophysical Journal,Part 1,vol. 263,Dec. 15,1982,p. 835-853., Vol. 263, pp. 835–853, 1982.

[14] G Pojmanski, B Pilecki, and D Szczygiel. The all sky automated survey. the catalog of variable stars. v. declinations 0 deg-28 deg of the northern hemisphere. *arXiv preprint astro-ph/0508017*, 2005.

[15] AJ Drake, MJ Graham, SG Djorgovski, M Catelan, AA Mahabal, G Torrealba, D García-Álvarez, C Donalek, JL Prieto, R Williams, et al. The catalina surveys periodic variable star catalog. *The Astrophysical Journal Supplement Series*, Vol. 213, No. 1, p. 9, 2014.

風況を考慮したEVのエネルギーベースラインマップの作成と長期間実走行データによる精度検証

劉　　屹† 曽我紗代子† 何　　新† 田中　悠斗† 富井　尚志††

† 横浜国立大学大学院環境情報学府情報環境専攻　〒240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
†† 横浜国立大学大学院環境情報研究院　〒240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
E-mail: †{liu-yi-tk,soga-sayoko-cz,he-xin-fw,tanaka-yuto-kt}@ynu.jp, ††tommy@ynu.ac.jp

あらまし　我々の先行研究では, 道路勾配が説明する必須となるエネルギー消費を地図上に可視化した EV のエネルギーベースラインマップ (Energy Baseline Map: EBM) を提案した.EBM には風の影響を考慮しないという欠点があった. 本稿では, 公開されている気象データを利用して, EBM の空気抵抗算出部分を改善し, 風の影響を考慮したエネルギーベースラインマップ (Energy Baseline Map-Corection of Airspeed: EBM-CA) を提案する. 長期間に渡って蓄積した EV の実走行データを用いて, EBM-CA と EBM を比較したところ, EBM-CA の予測精度が向上したことが確認された. さらに EBM-CA を利用して場所ごとに可視化し, 季節風による EV のエネルギー消費への影響を示した.

キーワード　電気自動車, モビリティ, CAN データ, 気象データ

1 はじめに

近年, 気候変動問題が世界的な課題となり, 脱炭素化および省エネルギー化が求められている. その中で, 現在世界中で CO_2 の排出が最終需要部門の 37%を占める運輸部門 [1] について対策が必要である. その手段として, 電気自動車（Electric Vehicle：EV）が注目されている [2]. EV は走行時に CO_2 を排出せず, ライフサイクルにおける CO_2 排出量も低いという特徴がある [3]. それゆえ, 次世代の乗り物として, EV はエネルギー節約と脱炭素化を実現する候補と考えられている [4].

近年では, EV のエネルギー消費を予測することが注目されている. EV の構造はシンプルであるため, 物理方程式を用いて電力消費を予測することができる [5]. EV のエネルギー消費は, 道路の勾配, 加速と減速, そして対気速度によって大きく異なる [5,6]. この特徴を利用して, 我々の先行研究では EV のエネルギーベースラインマップ (Energy Baseline Map: EBM) を提案した [7]. EBM は仮想定速走行により生成されたもので, 道路勾配が説明する EV 走行時に必ず必要なエネルギー消費を地図上に可視化したものである.

本稿では, 公開されている気象データを利用して, EBM の空気抵抗を予測する部分を改善した, 風を考慮したエネルギーベースラインマップ (Energy Baseline Map-Corection of Airspeed: EBM-CA) を提案する. 提案手法では, 公開されている気象データの中の都市全体の風速データを利用して, 電気自動車の各時点での対気速度を計算する. この方法により, 電気自動車のエネルギー推定の精度と結果の合理性を効果的に向上させるだけでなく, 公開データを有効活用することも可能になった. また, EBM-CA を利用して季節風がエネルギー消費に及ぼす影響の定量評価を行った.

本稿の構成を以下に示す. 2 章では, 本論文と他の研究との関連性および相違点について説明する. 3 章では, 提案手法について説明する. 4 章では, 精度検証の結果について説明する. 5 章では, EBM-CA を用いた季節風がエネルギー消費に対する影響の定量評価について説明する. 6 章では, 本論文のまとめ, および今後の課題について説明する.

また, 本稿の数式とアルゴリズムで使用された変数の説明は表 1 に示す.

2 関連研究

本章では, 他の研究と比較することによって, 本論文の位置付けを示す. 近年, 電気自動車のエネルギー消費予測モデルについて多くの研究が行われている. 大まかに言えば, これらは次の二つのカテゴリーに分けられる：

① 物理方程式に基づく理論モデル

② 過去の経験データに基づくモデル

Hirota ら [5], Wu ら [6], Tran ら [8], Fiori ら [9], Yi ら [10,11], Tawfiq ら [12] は, 物理方程式を使用して EV のエネルギー消費を推定した. これらの研究は我々の先行研究 [13–16] と同様に, 運転時の GPS 時系列ログを使用した. しかし, [8,10–12] はシミュレーションの結果のみを比較し, 実際の運転を使用した検証は行われていなかった. [6,9,13–16] は運転ログを使用して推定結果の精度を示したが, 各要因が消費エネルギーに及ぼす影響についてのさらなる分析は行われていなった.

Ullah ら [17] は, 愛知県の EV の実走行データを用いて重回帰（MLR）, 人工ニューラルネットワーク（ANN）, Extreme Gradient Boosting（XGBOOST）, Light Gradient Boosting Machine（LightGBM）の四つの機械学習モデルをそれぞれ構築して精度を比較した. Pan ら [18] は, S-TSD という短距離のドライビングセグメント分割アルゴリズムを提案し, 深層ニューラルネットワーク（DNN）を基盤とした電気自動車のドライビングセグメントレベルのエネルギー消費予測モデルを構築し,

表 1: 変数説明

変数	単位	意味
ρ	kg/m^3	空気密度
μ	-	転がり抵抗係数
g	m/s^2	重力加速度
C_d	-	空気抵抗係数
A	m^2	車両前面投影面積
M	kg	車両重量
M_l	kg	積載重量
M_i	kg	慣性重量
lat	-	緯度
lon	-	経度
v	m/s	車両速度
$v_{tailwind}$	m/s	進行方向の相対風速
v_{wind}	m/s	風速
Θ_{wind}	rad	風向
Θ_{car}	rad	車両の進行方向
θ	rad	道路勾配
R	m	タイヤ周
u	-	減速機の減速比
v_c	m/s	定速速度
d_c	m	速度 v_c で 1 秒間に進む距離
$Series^{rl}$	-	道路リンク端点座標の集合
r_d	m	現在地から現在の道路リンクの終端までの距離
t	s	スタートしてからの経過時間 (t = 0, 1,..., n)
p_t	-	時刻 t における座標
p_o	-	出発地の座標
p_d	-	目的地の座標
s_t	-	仮想の走行ログ ($s_t = (t, p_t)$)
i	-	道路リンク番号 ($i = 1, 2, ..., k$)
k	-	終点の道路リンク番号
j	-	現在の道路リンク内の走行回数
l_i	m	道路リンク i の長さ
TS_{start}		走行開始時点の時刻
$TS_{current}$		走行開始後 t 時点の時刻

XGBOOST と強化学習により構築したモデルと比較した．Qi ら [19] は, 実際の交通状況下での電気自動車のエネルギー消費分析モデルを提案した. 彼らは, 減速と加速が動力エネルギーに与える影響を表す 2 つの複合指標, NKE と PKE を導入し, これらを使用して, 道路区間上での電気自動車のエネルギー消費を分解した. これらの指標を用いて, 正確な推定モデルを構築し, 回帰モデル, ニューラルネットワークモデル, 回帰ニューラルネットワークモデルを用いて比較検証を行った．これらの研究 [17–19] はいずれも EV の実走行データを利用して機械学習モデルで消費エネルギー予測モデルの構築を行っている．我々は, 物理モデルを使用してエネルギーを推定することで, 様々な原因によるエネルギー消費を効果的に推定できると主張している. よって立場が異なる.

前記の先行研究と比較して本研究は，EV のエネルギー推定モデルを用いて，EBM-CA を作成し，エネルギー消費を地図上に示す．さらに，その応用までを論じており，その点で他の研究と異なる．

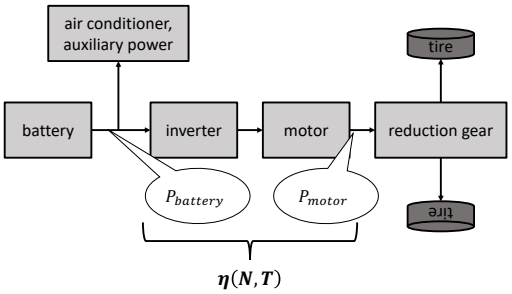


図 1: EV 内部のエネルギーの流れ

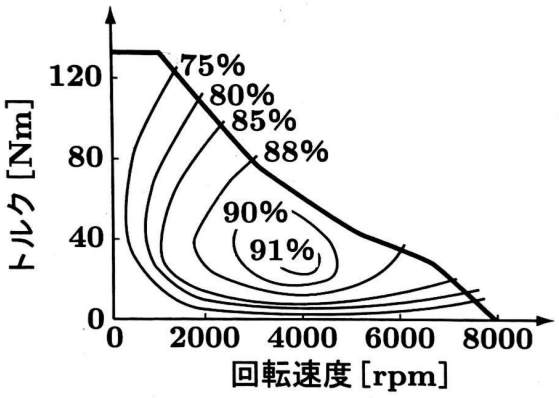


図 2: モータ・インバータ総合効率（文献 [5] より引用）

3 提 案 手 法

本章では, EV の内部構造に基づいた EV のエネルギー推定モデル及びこれを用いた EBM-CA の作成方法について説明する.

3.1 EV の内部構造

EV 内部でのエネルギーの流れを図 1 に示す．まず, バッテリの電気エネルギーがインバータで直流から交流に変換される. 次に, モータが電気エネルギーを力学的エネルギーに変換する. さらに, モータの作り出した力学的エネルギーを, 減速機によって走行に適した速度とトルクに変換してタイヤの駆動力にしている．この時, インバータとモータのエネルギーの変換過程で変換ロスが生じる．本稿では, モータ・インバータ総合効率は, この 2 つの変換機構における変換ロスをまとめたものを指す. この変換ロスを考慮した変換効率はモータ・インバータ総合効率と呼ばれ, 文献 [5] によって図 2 に示すグラフで説明されている．モータ・インバータ総合効率は車種ごとに固有の分布を持つ.

3.2 EV のエネルギー推定モデル

我々のモデルは図 1 の流れと逆の方向, すなわち駆動力から

モーター出力, バッテリー出力の方向でエネルギー消費を計算する. まず EV 走行時の駆動力を考える時に, エネルギー消費の要因別の抵抗とそれぞれに釣り合う力の和を駆動力として推定する. 運動方程式に基づく車の走行に必要な駆動力 $F_{driving}(t)$ [N] は式 (1) のように, 空気抵抗 $F_{air}(t)$, 登坂抵抗 $F_{slope}(t)$, 転がり抵抗 $F_{roll}(t)$, 加速抵抗 $F_{acc}(t)$ の 4 種類の抵抗と釣り合う力から構成される [5]. これらはそれぞれ式 (2)〜(5) と表される.

$$F_{driving}(t) = F_{\mathrm{acc}}(t) + F_{\mathrm{slope}}(t) + F_{\mathrm{roll}}(t) + F_{\mathrm{air}}(t) \quad (1)$$

$$F_{\mathrm{acc}}(t) = (M + M_l + M_i)\frac{dv(t)}{dt} \quad (2)$$

$$F_{\mathrm{slope}}(t) = Mg\sin\theta(t) \quad (3)$$

$$F_{\mathrm{roll}}(t) = \mu Mg\cos\theta(t) \quad (4)$$

$$F_{\mathrm{air}}(t) = \frac{1}{2}\rho C_d A(v(t) - v_{tailwind}(t))^2 \quad (5)$$

ただし, $v_{airspeed}(t) = v(t) - v_{tailwind}(t)$ は車両の t 時刻の対気速度を表す. $v_{tailwind}(t)$ は車両の進行方向の相対風速を表し, 車両の進行方向と風向と風速により計算できる. 風向と風速は気象のオープンデータから取得でき, 車両の進行方向は t 時刻の経緯度と $t-1$ 時刻の経緯度により計算できる.

$$\Theta_{car} = arctan(\frac{lon(t) - lon(t-1)}{lat(t) - lat(t-1)}) \quad (6)$$

$$v_{tailwind}(t) = v_{wind}(t) \times cos(\Theta_{wind} - \Theta_{car}) \quad (7)$$

次に, モータ端の出力は駆動力そのものであるため, モータ端の力学的エネルギーの仕事率 $P_{driving}(t)$ [W] は駆動力と車両速度から計算できる.

$$P_{driving}(t) = F_{driving}(t) \times v(t) \quad (8)$$

最後にバッテリー端の出力 $P_{battery}(t)$ [W] を考える. 図 1 で示したように, EV は電気的エネルギーを力学的エネルギーに変換して走行する. その際のエネルギー変換損失をモータ・インバータ総合効率 $\eta(N,T)$ として表す. 力行時の EV の瞬時消費電力 $P_{battery}(t)$ は $P_{driving}(t)$ と変換効率 $\eta(N,T)$ の商で表すことができる.

$$P_{battery}(t) = \frac{P_{driving}(t)}{\eta(N,T)} \quad (9)$$

また, EV には回生ブレーキにより, 自身の運動エネルギーを電気エネルギーとして回収する仕組みがある [5]. エネルギーの流れは図 1 の逆になり, $P_{driving}(t)$ が負のときは回生ブレーキによりバッテリーに充電する. 簡易的に、回生時にバッテリーに充電される電力 $P_{regene}(t)$ [W] は, このときの $-P_{battery}(t)$ と同等と見なす. また, 回生時もモータ・インバータ総合効率 $\eta(N,T)$ に影響される [5].

$$P_{regene}(t) = -P_{battery}(t) = -P_{driving}(t) \times \eta(N,T) \quad (10)$$

ただし, $\eta(N,T)$ はモータの回転速度 $N(t)$ [rpm] とモータ出力トルク $T(t)$ [Nm] によって決定される 2 変数関数ととらえることができる. $T(t)$ は駆動力と減速比とタイヤ周により計算できる. $N(t)$ は車両速度 $v(t)$ と減速比 u とタイヤ周 R により計算できる.

$$T(t) = \frac{F_{driving}(t) \times R}{u} \quad (11)$$

$$N(t) = \frac{v(t) \times 60 \times u}{2\pi \times R} \quad (12)$$

3.3 EBM-CA の作成

この節では, EBM-CA の作成方法を説明する. 先行研究 [7] では, 式 (5) の風による影響を無視した. 本稿では気象データから風の情報を取得して風の影響を考慮したものを EBM-CA とする. EBM-CA の生成プロセスは次の通りに示す:

Step 1. 走行のルート, 定速速度, 走行の開始時点を指定して仮想定速走行 (Virtual Constant Velocity Running: VCVR) による時刻情報付きの時系列の GPS ログ $TSVCVRLog$ を生成する.

Step 2. $TSVCVRLog$ を走行ログとみなして, 風況データ, 標高データと車両の諸元と一緒にシミュレーションの仕組みに入力して各時刻のエネルギー消費を計算する. シミュレーションは式 (1)〜式 (12) をベースにしたプログラムのことを指す.

Step 3. 計算されたエネルギー消費の時系列データを, 地系列のデータに変換して地図上に示す.

Step1 では, 我々の先行研究 [20] の手法をベースにして, 開始時刻を取得した時刻情報付きの $TSVCVRLog$ を生成した. 手法のイメージを図 3 に示す. この先行研究 [20] では, 「交差点を端点（ノード）として現実の道路をデータとして表現するためのフォーマット」を道路リンク（RoadLink）と定義した. 開始時刻, 出発地と目的地を指定し, 走行ルート上を一定速度で走行する仮想の走行ログを生成する. 走行ルートは連続した道路リンクの集合 $Series^{rl}$ と考えることができる. 入力である出発地と目的地から, 既存のナビシステム（例えば, Google map など）で走行ルートが決定できるので, 連続した道路リンクが取得できる. 走行ルート上を速度 v_c で走る EV を想定した等間隔 (d_c) の座標時系列として仮想の走行ログを生成する. また, この座標時系列の位置情報と時刻情報を参照することで気象データからその時の風況情報を取得できる. この風況情報付きの時系列データを $TSVCVRLog$ と呼ぶ.$TSVCVRLog$ 生成時の各パラメタを表 1 に, アルゴリズムを Algorithm 1 に示す. Algorithm 1 の各行の詳細を以下に列挙する. 番号は, Algorithm 1 の各行番号に該当する.

- 1: 時刻 $t = 0$ における初期値を設定する.
- 2–4: 出発時点と出発場所を $TSVCVRLog$ の最初のレコードとして記録する.
- 5: 現在地が目的地であるかを判別する.
- 6–8: 現在の道路リンクの座標を取得し, ヒュベニーの公式 $Hubeny()$ を用いて道路リンク長を計算する.

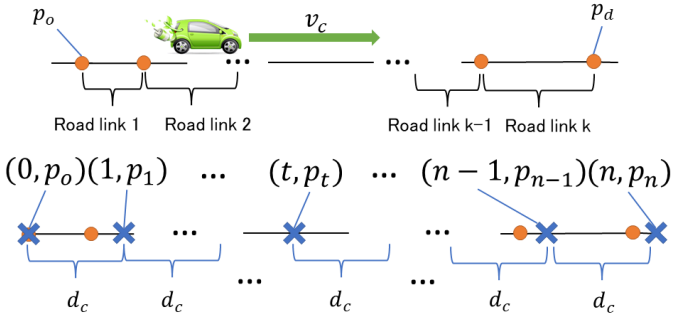


図 3: 先行研究 [20] の手法のイメージ

Algorithm 1 Generate $TSVCVRLog$

Require: $Series^{rl}, v_c, TS_{start}$
Ensure: $TSVCVRLog$
1: $t \Leftarrow 0, i \Leftarrow 1, d_c \Leftarrow v_c \times 1[s], r_d \Leftarrow d_c, TS_{current} \Leftarrow TS_{start}$
2: $p_o \Leftarrow Series_0^{rl}$
3: $s_t \Leftarrow (TS_{current}, p_t)$
4: $TSVCVRLog.add(s_t)$
5: **while** $i < k$ **do**
6: $(lat_{now}^{rl}, lon_{now}^{rl}) \Leftarrow Series_i^{rl}$
7: $(lat_{last}^{rl}, lon_{last}^{rl}) \Leftarrow Series_{i-1}^{rl}$
8: $l_i \Leftarrow Hubeny(lat_{last}^{rl}, lon_{last}^{rl}, lat_{now}^{rl}, lon_{now}^{rl})$
9: **if** $r_d > l_i$ **then**
10: $r_d \Leftarrow r_d - l_i$
11: **else**
12: $j \Leftarrow 0$
13: **while** $r_d + d_c \times j < l_i$ **do**
14: $lat^{vcvr} \Leftarrow lat_{last}^{rl} + \frac{(lat_{now}^{rl} - lat_{last}^{rl}) \times (r_d + d_c \times j)}{l_i}$
15: $lon^{vcvr} \Leftarrow lon_{last}^{rl} + \frac{(lon_{now}^{rl} - lon_{last}^{rl}) \times (r_d + d_c \times j)}{l_i}$
16: $p_t \Leftarrow (lat^{vcvr}, lon^{vcvr})$
17: $TS_{current} \Leftarrow TS_{start} + t$
18: $s_t \Leftarrow (TS_{current}, p_t)$
19: $TSVCVRLog.add(s_t)$
20: $j \Leftarrow j + 1$
21: $t \Leftarrow t + 1$
22: **end while**
23: $r_d \Leftarrow r_d - l_i + d_c \times j$
24: **end if**
25: $i \Leftarrow i + 1$
26: **end while**
27: **return** $TSVCVRLog$

- 9: 現在地から d_c 進むと道路リンクをまたぐかどうか判別する.
- 10: 9: で道路リンクをまたぎ，5: で k には至っていない場合，r_d を更新し，道路リンク番号をインクリメント.
- 11–13: 9: で道路リンクをまたがない場合，またぐまで d_c の整数 j 回を進み続ける.
- 14–19: またぐまで 1 秒ごとの移動軌跡と時刻を算出する. そして，その時の場所情報と時刻情報を記録する.
- 20–22: 経過時間 t を 1 秒，移動回数 j を 1 回追加する. これを現在の道路リンクをまたぐまで繰り返す.
- 23–26: 10: で道路リンクをまたぎ，6: で k には至っていない場合，r_d を更新し，道路リンク番号をインクリメント.
- 27: 目的地に到達した場合，$TSVCVRLog$ を返す.

Step2 では，この $TSVCVRLog$ を走行ログとして扱い，式 (1)〜(12) までに示されたモデルに入力する. このとき，道路勾配データ，車両の仕様データ，定速速度，風況を含むデータ (例えば，気象庁が公開した気象データ) も同時に入力する. 道路勾配データは，外部の地図データ（例えば，ゼンリン，Google Earth，国土地理院の ADAS データ）を使用する. これにより，EV でそのルートを走行した際の時系列のエネルギー消費データ $VCVRCAEnergySeries$ が得られる.

図 4: Segment 割り振りのイメージ

Step3 では，$VCVRCAEnergySeries$ を地系列データに変換して地図上に可視化する. 通常，EV のエネルギー消費データは瞬時の電力消費 [W] の時系列で表される. 一方で，エネルギー消費の時間積算値 [J] を地図上に示し，地点ごとのエネルギー消費を可視化したいという需要もある. そのため，単位距離当たりのエネルギー消費である [J/m] の次元で電費を可視化する. 本節では，集約の単位として利用する地図情報として，Semantick Link と segment を定義する. まず Semantick Link (SL) は，「運転者やデータ解析者にとって有意な道路区間」である [13]. 例えば，高速道路の特定区間や渋滞が起こりやすい区間，複数のランドマークを結ぶ道などが挙げられる. また、本稿では、SL は精度検証時の道路のことを指す. さらに,「セマンティックリンクをある一定間隔 [m] で分割した区間単位」をセグメントと定義する. $VCVRCAEnergySeries$ は $TSVCVRLog$ から受け継いた GPS 座標情報を使って，各データ点がどのセグメントに属するかを決定する. このプロセスは最近傍法によって行われ，セグメントの中心と $VCVRCAEnergySeries$ のデータ点のユークリッド距離を基に割り振りする. ラベリングプロセスのイメージは図 4 に示す. セグメントの長さは，データ分析の目的に応じて決定する. 例えば，短距離の SL には 10 メートルや 50 メートル，長距離には 500 メートルや 1 キロメートルを使う. Segment 割り振り後，$VCVRCAEnergySeries$ を segment ID を用いて集約することによって時空間変換を行う. 最後に変換後の地系列エネルギー消費データを地図上に可視化したら EBM-CA になる. そのイメージを図 5 に示す.

4 EBM-CA と EBM との精度比較

本章では，EBM の精度検証結果を説明し，EBM-CA の精度向上を示す.

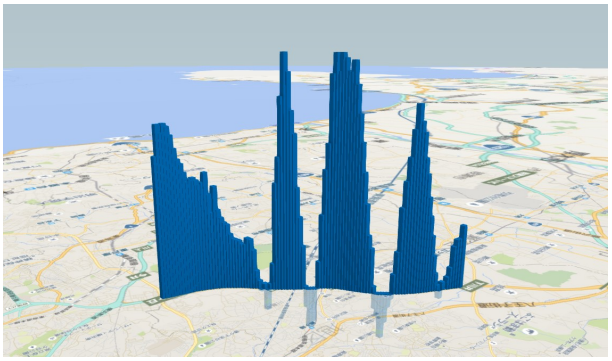


図 5: EBM-CA のイメージ

4.1 EBM の精度検証結果

我々の先行研究 [7, 21] では, EBM の精度検証を様々な特徴の異なる道路で行った. EBM は先行研究 [7] の手法で作成し, 正解データは日常走行の中に存在する ACC（Adaptive Cruise Control）を用いた定速走行時の CAN データを使用した. 実験用の EV として日産 LEAF（ZAA-ZE1 型 2020 年式）を用いた. CAN データの取得には Android アプリ LeafSpy Pro[1]を使用した. Android 端末には SONY 製 XPERIA 1 を用いた. また, OBD2（On Board Diagnosis second generation）接続用の Bluetooth 通信端末として Kiwi3 OBD2 Wireless Bluetooth Diagnostic Scannar[2]を使用した. 表 2 に示した道路について, 精度検証の結果を表 3 に示す. ただし, 正解データには電装品とエアコンのエネルギー消費を除いたものを用いた. いずれの道路についても精度が高いことが示された.

4.2 EBM と EBM-CA の精度比較

EBM と EBM-CA の精度を比較する. 精度評価に使用した正解データは, 節 4.1 と同じ手法で取得され, 2020 年 6 月から 2023 年 11 月までの被験者の日常的な通勤データである. ただし, 正解データは道路ごとに定速走行のデータのみを抽出したものである. 精度比較に使用した対象道路は表 4 に示す. また, EBM-CA 作成では, 出発時刻は被験者の日常的な通勤データの通行時間に合わせて時刻を設定し, 風況を取得する際はその時刻を利用し気象庁の公開データ [22] を参照した. EBM と EBM-CA の精度を比較するために, 比較対象道路における合計消費エネルギー量 [kWh] を, EBM と正解データ, EBM-CA と正解データの間の二乗平均平方根誤差（Root Mean Square Error: $RMSE$）および相関係数 (R) を計算して評価する. それぞれの式を次に示す.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{13}$$

$$R = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{14}$$

ただし, y_i が正解データで,$\hat{y}_i$ が EBM と EBM-CA の予測値を表し,$\bar{y}$ と $\bar{\hat{y}}$ がそれぞれの平均を表す.

SL1 における精度比較結果では, EBM の場合の RMSE と相関係数がそれぞれ 0.0912[kWh] と 0 に対して, EBM-CA では 0.0702[kWh] と 0.6 に向上した. 精度の向上が確認できた.

SL2 における精度比較結果では, EBM の場合の RMSE と相関係数がそれぞれ 0.0285[kWh] と 0 に対して, EBM-CA では 0.0289[kWh] と 0.51 であり, わずかながらも精度の向上が確認できた.

また, 図 6a と図 6b に SL1 の精度検証結果の散布図を示す. 横軸に正解データ, 縦軸に EBM と EBM-CA の合計消費エネルギー量 [kWh] が表示されている. 同様に, 図 6c と図 6d に SL2 の精度検証結果の散布図を示す.

EBM-CA は風況の情報を利用した結果, EBM と比較して合理性と精度の向上が確認された.

5 EV エネルギー消費における季節風の定量評価

本章では, EBM-CA を使用して, 季節風が EV の走行中のエネルギー消費に与える影響を定量的に評価する.

SL1 は横浜市内で北西方向に向かって走行する道路である. 横浜市の風況は, 夏季に南風が, 冬季には北風が吹く傾向を示す. 風速は夏季に比較的弱く, 冬季や初春にやや強くなっている. 気象庁が公開している 2020 年 2023 年の風況データ [22] をもとにして, SL1 を走る際の進行方向の相対風速の月ごとの分布を図 7 に示す. SL1 では夏季には南風がやや弱まり, 図 7 に示されているように, 6 月から 8 月はやや追い風になる. 一方, 冬季には北風がやや強まり, 10 月から翌年の 1 月にかけては顕著な向かい風となる. 図 8 では, 2020 年から 2023 年までの 3 年間, 平日の夜に SL1 を走行するシナリオにおいて, EBM-CA を日毎に生成し, 月ごとの平均エネルギー消費を示している. 季節風の影響が明らかに現れており, 夏季には追い風の影響でエネルギー消費が少なくなる傾向が見られる. 特に, 向かい風が最も強い 2 月は, 追い風の最も強い 6 月と比較して 14 %増加している. EBM-CA による定量分析および結果は, EV の走行を計画する際の充電に対する意思決定に役立つと考える.

6 ま と め

本研究では, 公開されている気象データを活用し, EBM の空気抵抗計算部分を改善し, 風の影響を考慮した EBM-CA を提案した. 複数の対象道路で EBM と EBM-CA の精度を比較し, 比較対象道路における合計消費エネルギー量 [kWh] の RMSE および相関係数 R を計算し評価した.

SL1 における精度比較結果では, EBM の場合の RMSE と相関係数がそれぞれ 0.0912[kWh] と 0 であるのに対し, EBM-CA では 0.0702[kWh] と 0.6 に向上し, 精度の向上が確認された. また, SL2 における精度比較結果でも, EBM の場合の RMSE

1：”LeafSpy Pro”, Google Play, https://play.google.com/store/apps/details?id=com.Turbo3.Leaf_Spy_Pro&hl=ja&gl=US（参照 2023-12-17）.
2：Inc. PLX Devices, ”PLX Devices model Kiwi3” https://www.plxdevices.com/Kiwi-3-OBD-Car-to-Smartphone-Connection-p/897346002832.html（参照 2023-12-17）.

表 2: 対象道路

名称	道路種類	距離 [km]	特徴	*Segment* 間隔 [m]	定速条件 [km/h]
SL1	高速道路 1	5.4	緩やかな勾配変化あり	50	80
SL2	高速道路 2	1.9	緩やかな勾配変化あり	50	70
SL3	高速道路 3	4	山道	50	50
SL4	高速道路 4	13.8	環状線	50	50
SL5	一般道路 1	1.8	山道で下り坂	50	30
SL6	高速道路 5	31.7	緩やかな勾配変化あり	500	70
SL7	一般道路 2	170	勾配変化が少ない	1000	60

表 3: 精度検証結果

名称	合計消費エネルギー (EBM)[kWh]	合計消費エネルギー（正解データ）[kWh]	$RMSE$[kWh]
SL1	0.758	0.773	0.001
SL2	0.358	0.333	0.002
SL3	1.707	1.61	0.002
SL4	1.717	1.763	0.003
SL5	-0.098	-0.134	0.003
SL6	4.54	4.723	0.013
SL7	21.615	20.737	0.049

表 4: EBM と EBM-CA の精度比較対象道路

道路名	定速条件 [km/h]	定速走行トリップ数
SL1	80	183
SL2	70	306

と相関係数がそれぞれ 0.0285[kWh] と 0 であるのに対して, EBM-CA では 0.0289[kWh] と 0.51 であり, わずかながらも精度の向上が確認された.

さらに, 季節風が EV の走行中のエネルギー消費に与える影響を定量的に評価した. 横浜の北西方向に向かって走行する道路 SL1 において, 季節風による影響を分析した結果, 向かい風が最も強い 2 月は, 追い風の最も強い 6 月と比較して 14 %増加していることが明らかになった.

今後の課題として, より広範な実験を行うことや, EV のバッテリーの温度特性の定量評価を行うことが挙げられる.

謝　　辞

本研究の一部は JSPS 科研費 (課題番号 22H03810) の支援による．道路勾配データの作成については，株式会社ゼンリンデータコムのサポートを得た．アルゴリズムの実装と分析結果の作成は同じ研究室に所属する中島誠敬氏と吉川遥氏の協力を得た.

文　　献

[1] International Energy Agency. Analysis - Transport. `https://www.iea.org/topics/transport`, 2022. Last accessed: 12 December 2023.

[2] 経済産業省資源エネルギー庁. 令和 2 年度エネルギーに関する年次報告 (エネルギー白書 2022), Last accessed: 22 July 2022.

[3] T. R. Hawkins, B. Singh, G. Majeau-Bettez, and A. H. Strømman. Comparative environmental life cycle assessment of conventional and electric vehicles. *Journal of Industrial Ecology*, Vol. 17, No. 1, pp. 158–160, 1 2013.

[4] International Energy Agency. Energy Efficiency. `https://www.iea.org/reports/energy-efficiency-2022/executive-summary`, 2022. Last accessed: 12 December 2023.

[5] Yukitsugu Hirota and Satoshi Ogasawara. *Electric Automotive Engineering*. MORIKITA PUBLISHING, 7 2017. In Japanese.

[6] Xinkai Wu, David Freese, Alfredo Cabrera, and William A. Kitch. Electric vehicles' energy consumption measurement and estimation. *Transportation Research Part D: Transport and Environment*, Vol. 34, pp. 52–67, 2015.

[7] 劉屹, 曽我紗代子, 七里祐輔, 富井尚志. 仮想定速走行シミュレーションデータに基づく ev のエネルギーベースラインマップの作成と精度評価. WebDB 夏のワークショップ 2022, 情報処理学会, 研究報告データベースシステム（DBS）, No. 2022-DBS-175(28), pp. 1–6, 9 2022.

[8] Trung Bach Tran, Ilya Kolmanovsky, Erik Biberstein, Omar Makke, Marina Tharayil, and Oleg Gusikhin. Wind sensitivity of electric vehicle energy consumption and influence on range prediction and optimal vehicle routes. In *2023 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*, pp. 112–123, 2023.

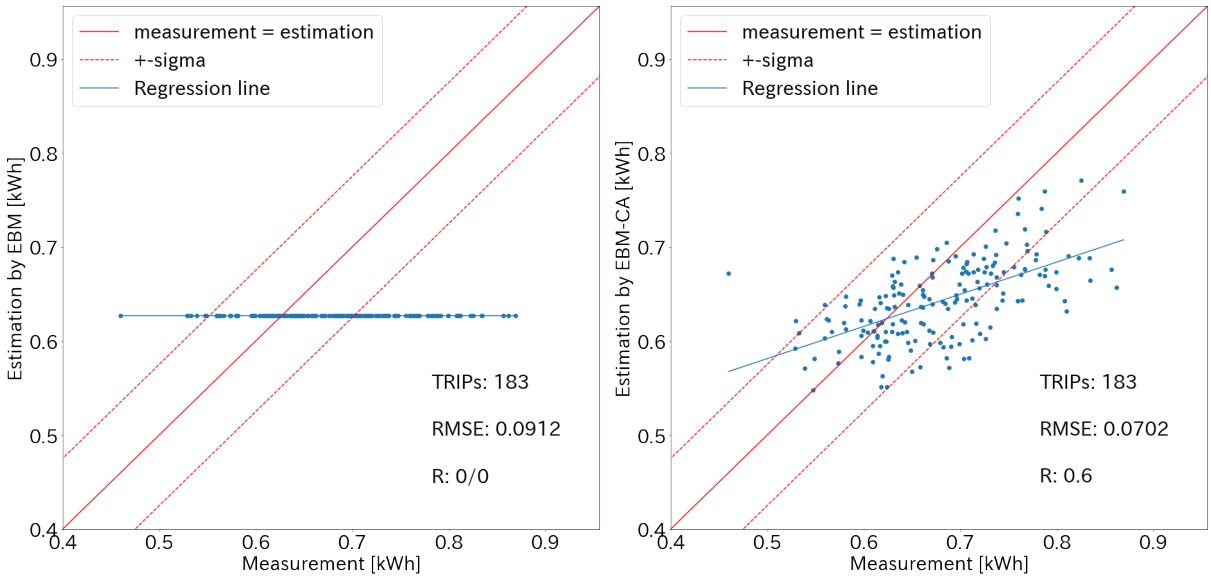
[9] Chiara Fiori, Kyoungho Ahn, and Hesham A. Rakha. Power-based electric vehicle energy consumption model: Model development and validation. *Applied Energy*, Vol. 168, pp. 257–268, 2016.

[10] Zonggen Yi and Peter H. Bauer. Sensitivity analysis of environmental factors for electric vehicles energy consumption. In *2015 IEEE Vehicle Power and Propulsion Conference (VPPC)*, pp. 1–6, 2015.

[11] Z. Yi and H. P. Bauer. Energy consumption model and charging station placement for electric vehicles. In *Proceedings of the 3rd International Conference on Smart Grids and Green IT Systems (SMARTGREENS-2014)*, pp. 150–156. SCITEPRESS, 2014.

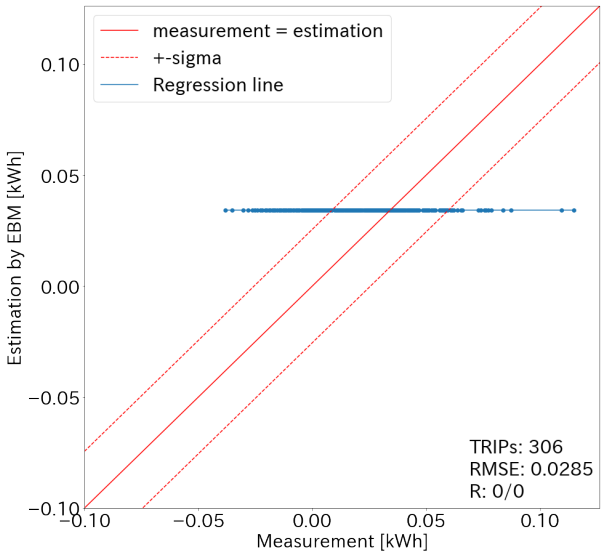
[12] Tawfiq M. Aljohani, Ahmed Ebrahim, and Osama Mohammed. Real-time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and markov chain model. *Electric Power Systems Research*, Vol. 192, p. 106962, 2021.

[13] Daiki Kawanuma, Yuta Kashiwabara, Toshiaki Uemura, and Takashi Tomii. Data analysis framework for visualizing

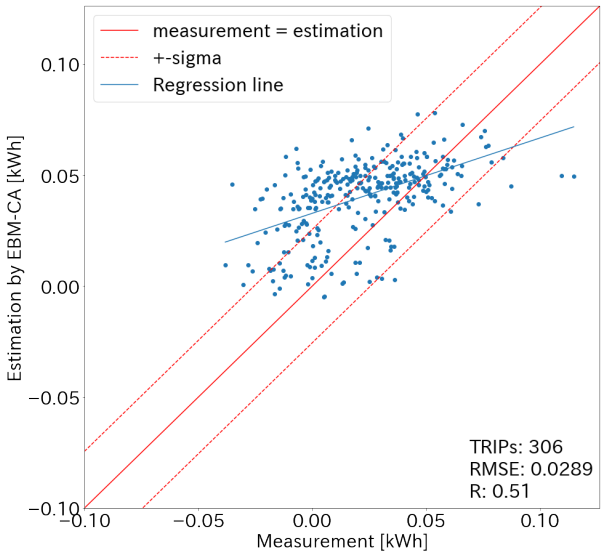


(a) SL1 における EBM の精度検証

(b) SL1 における EBM-CA の精度検証



(c) SL2 における EBM の精度検証



(d) SL2 における EBM-CA の精度検証

図 6: 精度検証結果

correlation of energy consumption and transit time in road sections using the ecolog database. In *Adjunct Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services*, pp. 207–212, 2016.

[14] Takashi TOMII, Shintaro HAGIMOTO, Naoki FUEDA, Toru DEGUCHI, Makoto IDENAWA, and Takuya HAYASHI. Long-term experiment of the ecolog database capability of estimating v2x effect replacing with evs. In *Proceedings of 20th ITS World Congress Tokyo 2013*, No. paper#3162, p. 10, Tokyo, Japan, 10 2013.

[15] Toshiaki Uemura. Pre-estimation of electric vehicle energy consumption on unfamiliar roads and actual driving experiments. In Takashi Tomii, editor, *Proceedings of the VLDB 2019 PhD Workshop, co-located with the 45th International Conference on Very Large Databases (VLDB 2019)*, Vol. 2399 of *CEUR Workshop Proceedings*, pp. 1–4, Los Angeles, 8 2019.

[16] Toshiaki Uemura, Daigo Nojo, Yudai Kichise, and Takashi Tomii. Electric vehicles' estimated energy consumption data analysis system enabling aggregations by road sections in analysts' curiosity. *IPSJTOD*, Vol. 14, No. 4, pp. 70–85, 10 2021. In Japanese.

[17] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal. A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability. *Energy & Environment*, Vol. 33, No. 8, pp. 1583–1612, 2022.

[18] Yingjiu Pan, Wenpeng Fang, and Wenshan Zhang. Development of an energy consumption prediction model for battery electric vehicles in real-world driving: A combined approach of short-trip segment division and deep learning. *Journal of Cleaner Production*, Vol. 400, p. 136742, 2023.

[19] Xuewei Qi, Guoyuan Wu, Kanok Boriboonsomsin, and Matthew J. Barth. Data-driven decomposition analysis and estimation of link-level electric vehicle energy consumption under real-world traffic conditions. *Transportation Research Part D: Transport and Environment*, Vol. 64, pp. 36–52, 2018. The contribution of electric vehicles to environmental challenges in transport. WCTRS conference in summer.

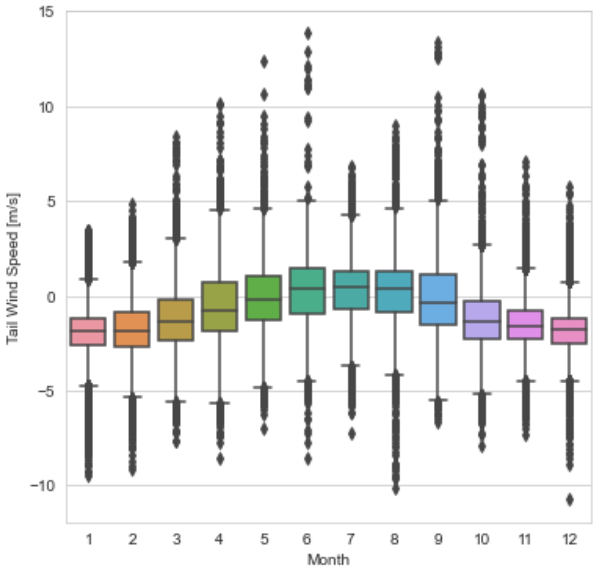


図 7: SL1 における風況

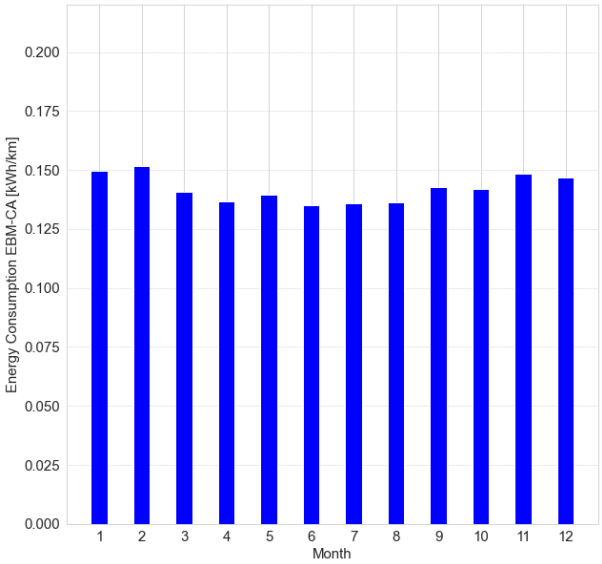


図 8: SL1 における EBM-CA による電費

[20] 磯部康太, 有長拓海, 植村智明, 齊藤祐亮, 吉瀬雄大, 富井尚志. 未知道路に対する ev の消費電力量推定システムの実走行データを用いた評価. 第 11 回データ工学と情報マネジメントに関するフォーラム (DEIM2019), No. H6-3, pp. 1–8, 3 2019.

[21] 曽我紗代子, 劉屹, 七里祐輔, 田中悠斗, 何新, 富井尚志. 実走行データを用いた ev のエネルギーベースラインマップの精度評価. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), No. 5C-8-3, pp. 1–8, 3 2023.

[22] 国土交通省気象庁. 過去の気象データ. `https://www.data.jma.go.jp/stats/etrn/index.php`, 2023. Last accessed: 12 December 2023.

PV と EV を統合する負荷平準指向 VGIDB を用いたスマートグリッド導入効果の多角的定量化

廣居　樹† 片平　昇輝† 鈴木　博登† 小松虎太郎†† 本藤　祐樹††† 富井　尚志†††

† 横浜国立大学大学院環境情報学府情報環境専攻　〒 240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
†† 横浜国立大学理工学部数物・電子情報系学科　〒 240–8501 神奈川県横浜市保土ケ谷区常盤台 79–5
††† 横浜国立大学大学院環境情報研究院　〒 240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
E-mail: †{hiroi-tatsuki-cf,katahira-shoki-gm,suzuki-hiroto-rs}@ynu.jp, ††{komatsu-kotaro-mk}@ynu.jp, †††{hondo,tommy}@ynu.ac.jp

あらまし　本論文では, 太陽光発電（PV）と電気自動車（EV）を統合する新しいスマートグリッドのあり方である VGI（Vehicle Grid Integration）を定量評価するためのデータベースを構築する. データのパラメータとして, PV の設置面積, EV の台数, および, 定置型バッテリー容量を設定可能とした. また, そのパラメータ設定の下で生成されたグリッドの消費電力, PV 発電電力, EV のバッテリー状態, および, EV のグリッドへの接続状態, 等の時系列データを用いた. これらのデータに基づいて VGI シミュレーションを行った. 負荷平準指向 VGI の実現性を評価するために, 提案データベースに定型質問処理を行い, 効果を多角的に定量化した. 結果として次に示されることが確認された. (1)PV だけを導入したグリッドと比較し, 負荷平準指向 VGI は発電余剰電力を削減可能であること. (2)EV の年間総充電量のうち PV 発電電力からの充電量は PV 及び EV が増加するほど増加し，その割合が最大で 69%となること. (3) 実質電力需要の変動量及び変動のある日数が EV が増加するほど減少すること. (4) 定置型バッテリー容量に関して適切なバッテリー容量が存在すること.

キーワード　ライフログ，太陽光発電，負荷平準化，スマートグリッド，VGI: Vehicle-Grid Integration，再生可能エネルギー，電気自動車，オープンデータ

1 はじめに

センサ技術やストレージ技術の発展により，現実世界の膨大かつ多種多様なデータの取得・蓄積が可能になった [1]. 中でも, スマートデバイスや IoT（Internet of Things）機器の普及により, 個人の生活の記録である「ライフログ」が容易に利用できるようになった. また, 国や地方公共団体がオープンデータに関する取り組みを進めたことにより, 気象データなどの様々な場所固有のデータが利用できるようになった. これらのライフログやオープンデータを充分に活用するためのデータ管理方法と適切な活用方法を確立することは重要である.

一方で, 地球温暖化対策として, 温室効果ガスの一種である CO_2 の排出を抑えつつ, 効率的にエネルギーを利用していくことが求められている [2] [3]. 東京電力グループの「長期的な安定供給とカーボンニュートラルの両立に向けた事業構造変革について」[4] によれば，「カーボンニュートラル社会を実現していくためには，相対的に安価な需要側の取り組み (太陽光・電化・蓄電池の設置等) が有効である」とされている.

需要側の取り組みのうち太陽光に関して，日本では再生可能エネルギーの中でも特に太陽光発電（PV：Photovoltaic power generation）が多く導入されており「国土面積当たり・平地面積当たりの太陽光発電の導入量は、主要国の中でも最大」と報告されている [5]. しかし, PV の導入に伴う課題が存在する. まず, PV の発電量は季節や気象に依存し不安定である. そのため, 発電電力を需要電力に一致させることが難しく, 消費しきれずに余ってしまう「発電余剰」が生じることがある. さらに, PV を導入しても太陽が夕方に沈んでしまうため, 夕方以降の需要のピークに対応できず, 実質電力需要（太陽光発電電力と需要の差）のピークは PV 導入前とかわらない. また, PV の導入量が増えると, 発電する時間帯としない時間帯で電力供給側に対する実質電力需要に急激な変動が生じ（ダックカーブ現象）, 電力供給側の発電効率に悪影響を及ぼすことが懸念される [**?**] [6].

次に, 需要側の取り組みのうち電化に関して, 世界的に電気自動車（EV）へのシフトが始まっている. それに伴い, バッテリー技術の研究が進んでいる. バッテリーの大容量化や全個体電池の研究によって, 将来はさらなる容量増加が予想される. しかし, EV シフトにも課題が存在する. EV によるカーボンニュートラルは電源の脱炭素化が前提となっているため, EV の走行電力が再生可能エネルギー由来である必要がある. また, EV が新たな電力需要を生じさせるという問題もある.

これらの PV・EV の問題に対し, PV と EV を統合する新しいスマートグリッドのありかたである VGI（Vehicle-Grid-Integration）によって両者の問題を同時に解決する試みがある [3]. VGI では EV のバッテリーを活用して効果的に充給電を

行う．さらに, 電力の需要と供給のマッチングを図ることで, 太陽光発電電力を有効活用できることが期待される．加えて, 実質電力需要の変動を抑えること（負荷平準化）[3] も実現可能である．

このとき, VGI によってもたらされる影響（スマートグリッド導入効果）は CO2 排出削減量を含む費用対効果, メリットに対するデメリットなど, 多くの競合条件が存在する．また, それらスマートグリッド導入効果は再生可能エネルギー発電設備や EV の導入量, その場所の天候など様々な要因により変化する．そのため, 設備を投入する前に, 現地で収集される多様な日常データからその効果を算出し, 複数の観点から多角的に定量化を行い, 意思決定者を支援する枠組みが必要である．

本論文では, 人の生活に伴う場所固有のデータを蓄積する．蓄積したデータを用い, 多種類のパラメータの任意の組み合わせに基づく VGI シミュレーションを行い, そのデータを独自に設計したデータベースに蓄積する．これらに対して定型質問処理を行い, パラメータを変更した場合のスマートグリッド導入効果を多角的に定量化する．結果として次に示されることが確認された．

(1). PV だけを導入したグリッドと比較し, 負荷平準指向 VGI は発電余剰電力を削減可能であること．

(2). EV の年間総充電量のうち PV 発電電力からの充電量は PV 及び EV が増加するほど増加し，その割合が最大で 69%となること．

(3). 実質電力需要の変動量及び変動のある日数が EV が増加するほど減少すること．

(4). 定置型バッテリー容量に関して適切なバッテリー容量が存在すること．

以上のことから, 提案データベースがスマートグリッド導入を計画する意思決定者に役立つ知見を与えることを示す．

2 関 連 研 究

本研究における関連研究として主に以下の 3 つが挙げられる．

（1） 電力供給側のスマートグリッドに関する研究

（2） 電力需要側のスマートグリッドに関する研究

（3） VGI および V2G（Vehicle to Grid）に関する研究

スマートグリッドに関する研究は，大別して電力供給側の立場から設計する研究と電力需要側の立場から設計する研究が見られる．

2.1 電力供給側のスマートグリッドに関する研究

電力供給側の立場から, 大規模な需要変動に対して周波数変動を抑制することを目的としてスマートグリッドを設計する研究が行われている．Zhang ら [7] は，カリフォルニア州を対象として，供給側の立場から EV の大量導入に伴う電力系統負荷の増大という課題に対して，EV を管理充電することの利点を様々な評価軸から示した．Zhang ら [8] もまた, 供給側の立場から, アメリカ中西部地域を対象に 2020 年から 2040 年にかけての EV の増加に伴う電力系統負荷の増大をシミュレーションから示した．加えて, 非制御充電と比較して, 管理された充給電による V2G が効果的であることを示した．小嶋ら [9] は, 供給側の観点から，戸建て需要をまとめることでより効率的に電力融通ができることと, PV・EV・蓄電池を導入することでエネルギーコストが低減可能であることを示した．

これらの研究は電力供給側の立場から大規模なスマートグリッドのあり方を議論したものである．一方で, 我々の研究は電力需要側の立場からスマートグリッドを考えるという点でこれらの研究と立場が異なる．

2.2 電力需要側のスマートグリッドに関する研究

前節と異なり電力需要側の立場からも, スマートグリッドを設計する研究が行われている．Narottam ら [10] は, 住居用負荷を最適に管理する V2G シミュレーションを行い, 容量の大きい EV をバッテリーとして活用することでピーク負荷の削減が可能であることを示した．Arab ら [11] は EMS（エネルギー管理システム）と PV 及び EV を各家庭に設置し効果的に運用することで, 外部供給電力及び電気料金の削減が可能であることを示した．

また, 再生可能エネルギーとバッテリーを組み合わせたスマートグリッドに関する研究は, 多種多様な分野から学際的に行われている．横浜市・豊田市・けいはんな・北九州市の 4 都市 [12] では, EMS や蓄電池, デマンドレスポンスを組み合わせたスマートシティプロジェクトの実例を挙げ, 供給側への影響を評価した．Simm ら [13] は再生可能エネルギーの発電電力と需要電力を同期させることを目的に, ユーザの電力消費意識を改善させるシステムを提案し, 電力系統が独立した離島で実証実験を行った．

これらの研究では需要体として各家庭を想定している．本論文では通勤先を需要体として想定したスマートグリッドでのシミュレーションを行う点が異なる．

2.3 VGI/V2G に関する研究

EV は非走行時に電力を電力網/家に供給すること（V2G および V2H:Vehicle to Home）ができ, 活用が期待される．これらの実験や効果検証に関する研究が近年多く見られる．V2G の導入効果の検討として Kempton ら [14] は, V2G をアメリカの電力市場に投入した際にどの程度の効果と利益が得られるのかの検証を行った．Jansen ら [15] は複数の EV 群を 1 つの仮想的な発電所としてみなす EV-VPP（EV Virtual Power Plant）を提案し, V2G が果たす役割について論じた．大澤ら [16] は, EV 単独で使用するよりも EV と V2H を組み合わせることで CO_2 排出量を多く削減できることと, EV と V2H を組み合わせれば PV が増えれば増えるほど CO_2 排出量を削減できることを示した．Ghosh ら [17] は，EV バッテリーを用いた V2G サービスの適切な価格設定方法を提案した．Tran ら [18] は, PV を組み合わせた V2H を行うための新しい技術を提案した．Turker ら [19] は, 電力網/PV/EV を組み合わたスマートグリッドのための, PV の自己消費量を最大化し EV への充電を最適

化するエネルギー管理方法として, 線形計画法を用いた手法を提案した. Donadee ら [20] は, V2G 技術を発電/配電容量に制約のある場所に導入すると, EV バッテリーの放電エネルギーが 47%増加してしまうが, 電力網の価値を 32%向上できることをシミュレーションから示した.

EV をスマートグリッド間を移動するバッテリーとして用いるアイディアは上記の研究から妥当な方法であると言える. 特に我々の研究は PV と EV に関するライフログデータを組み合わせて VGI のシミュレーション評価を行う点において優位性がある.

2.4 本研究の立場

前節までに示した研究に対し, 本研究では, 需要側の立場から, 不安定な再生可能エネルギーを地産地消しつつ外部供給電力を平準化すること, すなわち実質電力需要の平準化を通じて供給側の負担を減らす方法をとる. また, データ工学の立場から EV や家庭, 職場などで生じる日常的な生活の記録（ライフログ）を用いる. ライフログに限らず気象情報などのオープンデータを組み合わせて適切にデータ管理することによって, 効率的なエネルギーの活用が可能であることを示す.

また, VGI は, PV と EV を組み合わせるものである. PV は場所や天気によって発電量が大きく異なる. また, EV も場所や時間によってその数は大きく異なる. さらに, 電力需要も場所や季節, 時間によって大きく異なる. よって, VGI が有効であるかは場所や時間に大きく依存する. 今あるその場所固有のデータから VGI が有効であるかどうかを定量化することは有用である.

そこで, VGI に関する一連のデータを総合的に蓄積するためのデータベースを提案する. このデータベースは, スキーマに適合するデータ形式であれば仮想データ・実データの区別なく VGI における様々な影響を定量化できる. 本研究では, 現時点で入手可能な実データおよび妥当に生成された仮想データを用いることによって, 再生可能エネルギーと EV バッテリーを活用した VGI シミュレーションを行う. また, VGI では需要側や電力供給側, 各 EV ユーザなど様々な立場の人が関わる. したがって, VGI によって得られるメリットは複雑であり, 様々な捉え方がある. 本研究ではそれらを提案データベースを用いて多角的に定量化する.

3 需要側の負荷平準化を実現するスマートグリッド

3.1 想 定 環 境

本研究では日常的に移動する電気自動車（EV）を移動するバッテリーとみなしスマートグリッドの一部として利用する. 特に電力需要側の立場に立ち，オフィスビル群などの中小規模な地域電力網（マイクログリッド）を形成することを想定する. 図 1 に示すように，マイクログリッドは電力需要設備，PV 設備，バッテリー，および充給電設備（ES：Energy Station，以下 ES）により構成される. バッテリーには通勤者の EV とマイクログリッド内に設置される定置型バッテリーの 2 種類がある.

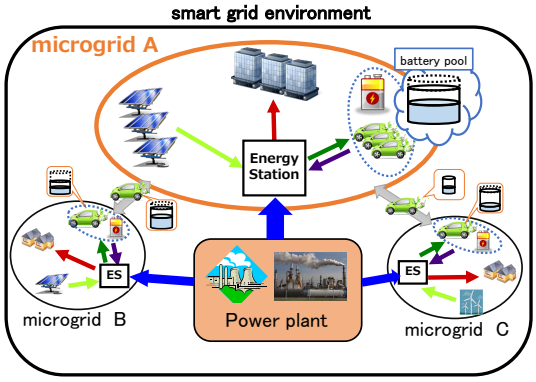


図 1 本研究で想定するスマートグリッド環境

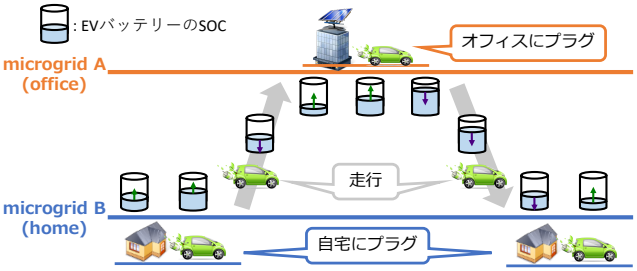


図 2 通勤に使用される EV の一日の EV の動き

ここで，通勤のためにオフィスと自宅を日常的に移動する EV を考える. EV の一日の動きの例を図 2 に表す. 自宅では充給電，移動の際には走行による放電，マイクログリッド内では充給電を行う. マイクログリッドにおいて，この 2 種類のバッテリーを論理的に多数束ねた「バッテリープール」が構成されるとみなす. バッテリープールの満充電容量（以下，バッテリープール容量）と充電率（SOC）はバッテリーの充給電や EV のマイクログリッドへの出入りによって常に変化する. また，ES は PV のパワーコンディショナー，PV 設備/バッテリー/電力系統間の AC/DC 変換，後述するエネルギーフローの制御（どこにどれだけの電力を流すか），充給電制御といった役割を持つものを想定する [1]. そのような ES が，VGI を行うマイクログリッド 1 つに対して 1 つ存在する. なお，マイクログリッド内部の発電だけでマイクログリッドの需要を賄うのは不可能である. したがって，マイクログリッド外には東京電力のような「外部発電所」が存在し，マイクログリッドに対して電力を供給している（外部供給電力）.

3.2 太陽光発電エネルギーと EV バッテリーを活用する負荷平準化

本研究では，外部供給電力に対してピークカット・ピークシフトを究極的に行った結果として外部発電所からの供給電力が一定になることを理想とする. このような状態は外部電源にとっての負荷が変動しないことから「負荷平準化」と呼ばれる.

図 3 に想定する電力の流れを，図 4 に太陽光発電エネルギー

1：この想定と同等の機能を持つシステムが，パナソニック株式会社などから HEMS として社会実装されている.

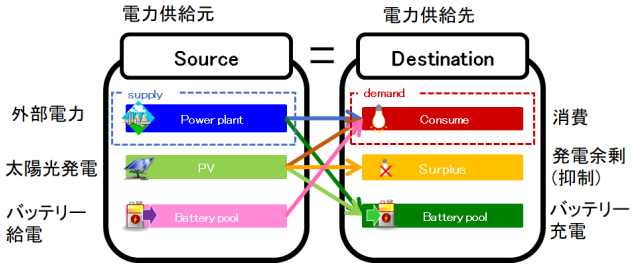


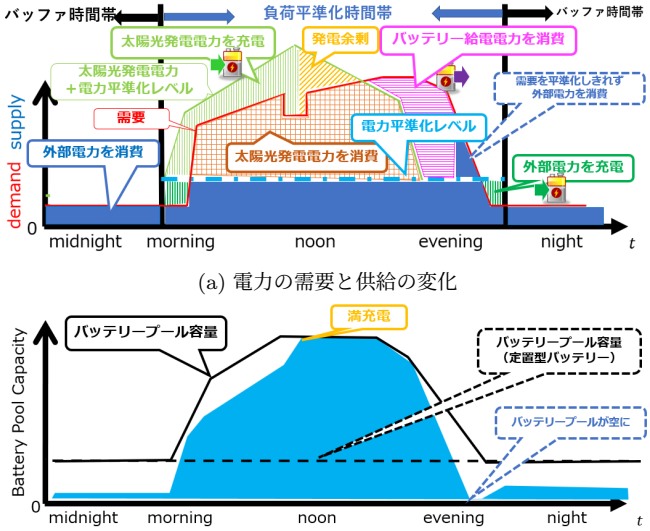
図 3　エネルギーフロー

と EV バッテリーを活用する負荷平準化の概念図を示す．ここで図 4（a）はあるマイクログリッドの一日の電力の流れの状態を表し，図 4（b）が対応するバッテリープールの状態を表す．図 4（a）の塗りつぶされた部分と吹き出しの色が図 3 の矢印の色に対応する．EV はマイクログリッド内外へ移動するため，時間ごとにバッテリープールの容量は変化する（図 4（b））．

負荷平準化を行うためには大容量のバッテリープールが必要となるため，1 日を通して負荷平準化を実現することは現実的に難しい．そこで，1 日のうち特定の時間帯のみを負荷平準化の対象とする．この時間帯を負荷平準化時間帯と定義する．それ以外の時間帯をバッファ時間帯と定義する．改めて，図 4（a）の負荷平準化時間帯に注目する．横軸に 1 日の時間，縦軸に電力をとった．赤線の概形（需要）は建物の電力需要の推移を表す．黄緑線の概形（太陽光発電電力＋電力平準化レベル）は電力平準化レベルを基準として太陽光発電電力をその上に積み上げたもので時間帯ごとの総供給量を表す．ここで外部からの供給電力を一定にするための基準となる電力平準化レベルをあらかじめ定める．このときの電力の流れは以下のようになる．

- 需要（赤）が電力平準化レベル（水色一点鎖線）以下の時間帯
  - 外部供給電力を消費（青）
  - 外部供給電力をバッテリープールへ充電（緑）
  - 太陽光発電電力をバッテリープールへ充電（黄緑）
- 需要（赤）が電力平準化レベル（水色一点鎖線）以上の時間帯
  - 需要（赤）以下の太陽光発電電力（黄緑）
    - 太陽光発電電力を消費（茶色）
  - 需要（赤）以上の太陽光発電電力（黄緑）
    - バッテリープールに充電可能の場合
      - バッテリープールへ充電（黄緑）
    - バッテリープールが満充電の場合
      - 発電余剰（オレンジ）
  - 太陽光発電電力（黄緑）以上の需要（赤）
    - バッテリープールが給電可能の場合
      - バッテリー給電電力を消費（ピンク）
    - バッテリープールが残量不足の場合
      - 外部供給電力を消費（青）

また，図 4 の青点線の吹き出しが指し示すようにバッテリープールの充電量が不足した場合には一時的に負荷平準化不可能となる．



(a) 電力の需要と供給の変化

(b) バッテリープールの容量と蓄積エネルギーの変化

図 4　太陽光発電エネルギーと EV バッテリーを活用する負荷平準化

4　太陽光発電と EV を統合する負荷平準指向スマートグリッド DB

4.1　VGI スキーマ

3 章に示したスマートグリッドに関する一連のデータを統合して蓄積するデータベースについて説明する．概念データモデルを図 5，6 に示す． この提案データベースは，VGI を行った結果を多角的に定量化できるものである．また，スキーマに適合するデータ形式であれば，仮想データ・実データの区別を問わず定量評価することが可能である．このとき，仮想データを用いればシミュレーションに，すべて実データを用いれば振り返り評価となる．

図 5，6 に示したスキーマをもとに，RDB スキーマに変換し，実装システムを構築した．提案するデータベースの RDB スキーマを図 7 に示す．図 5 中の赤字番号と図 7 中の赤字番号はそれぞれ対応している．概念スキーマの各エンティティ及び RDB スキーマの詳細は文献 [21], [22] に示したため，紙面の都合上ここでは省略する．

4.2　VGI スキーマを用いた定量評価

本研究では，スマートグリッドにおける負荷平準化の有用性を示すために，以下の 3 つのシナリオを考える．

シナリオ 0（太陽光パネルなし・バッテリーなしシナリオ）
現在の多くの電力需要側の環境に最も近いシナリオである．

シナリオ 1（太陽光パネルあり・バッテリーなしシナリオ）
シナリオ 0 に加え，PV を建物の屋上に設置したシナリオである．

シナリオ 2（太陽光パネルあり・バッテリーありシナリオ）
PV を建物の屋上に設置することに加え，スマートグリッドに出入りする車を EV とし，負荷平準化を行う．

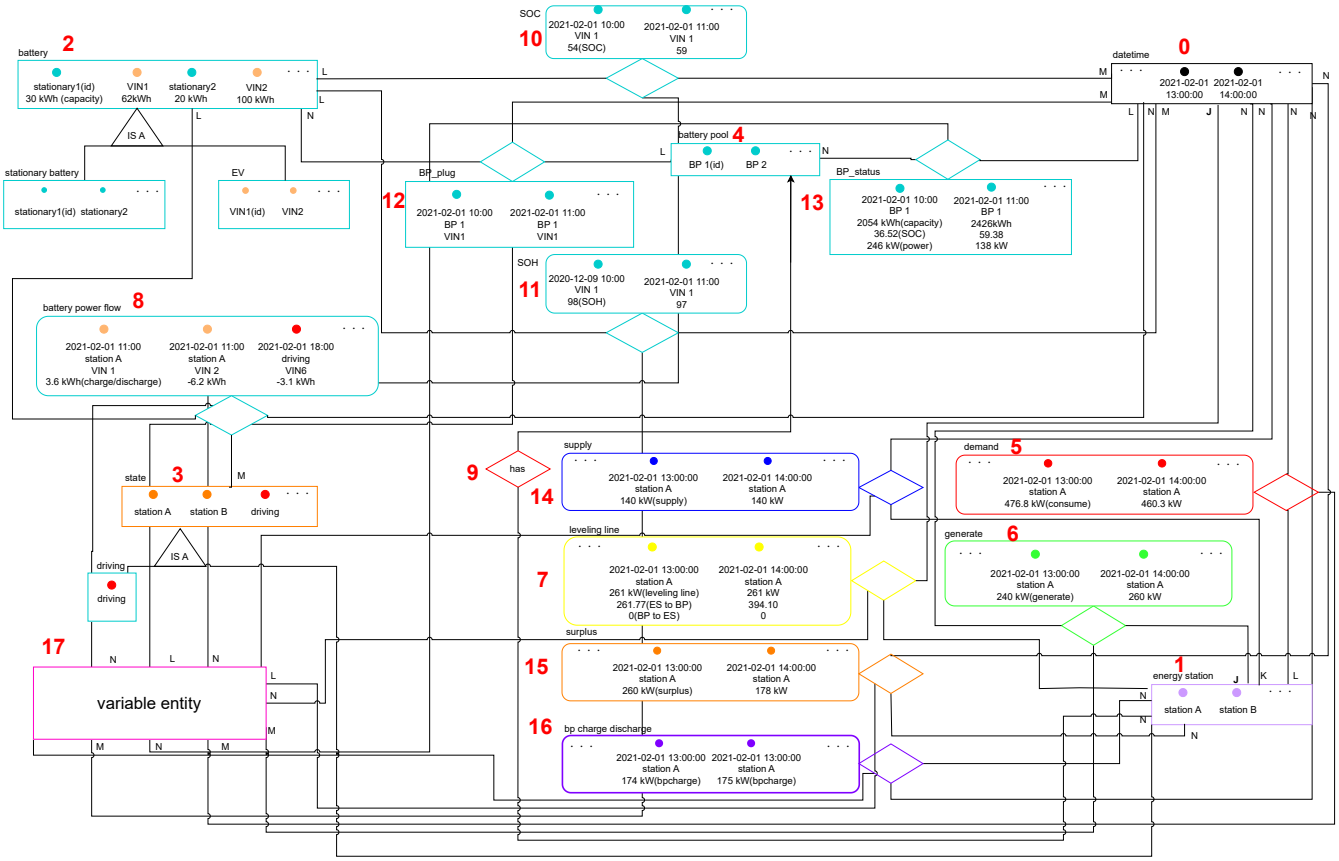


図 5　概念スキーマ

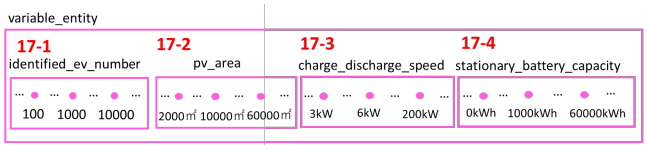


図 6　variable entity

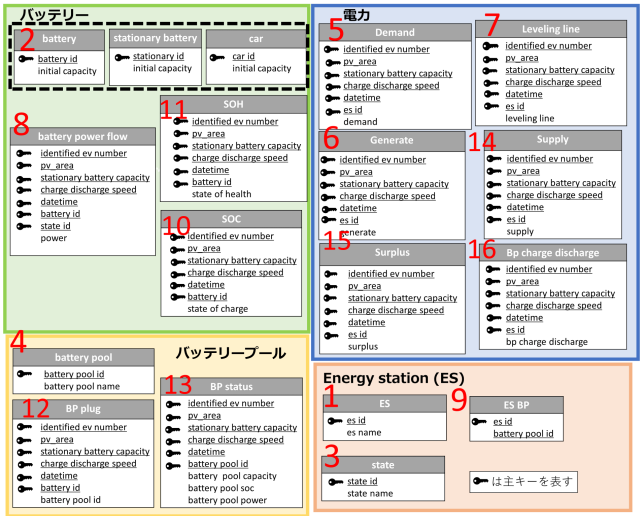


図 7　RDB スキーマ

5　VGI シミュレーションとスマートグリッド導入効果の多角的定量化

本章では, 3 章で述べた想定環境における VGI シミュレーションを行う．さらに, 4 章で提案したデータベーススキーマに後述する表 2 に記載するパラメータの全組み合わせのシミュレーションデータを格納する．データベースに対して定型質問処理を行い, パラメータを変更した場合のシミュレーション結果の定量化を行う．またその影響を多角的に検討する．

5.1　使用データ

VGI シミュレーションにあたり, 図 7 で示した提案データベーススキーマに実際にデータを挿入した．想定環境の例題として横浜国立大学常盤台キャンパス（横浜市保土ヶ谷区, 学生・教職員総数約 10000 人）の電力需要データ, および, 横浜市の気象データを基にシミュレーション用データ（一部実データ）を生成した．すなわち, この VGI シミュレーションでは横浜国立大学常盤台キャンパスを 1 つのマイクログリッドとした場合にどのようなことが起きるのかを多角的に定量化して検討する例題となる．

以下, 使用データの詳細について図 7 中の赤字番号順に説明していく．このとき説明に使用した文字式について表 1 に示す．また VGI シミュレーションに使用したパラメータを表 2 に示す．

0. datetime

τ_0（初期値）を 2021-01-01 0:00, T_{step}（サンプリング間隔）を 1 時間とした．

$$\mathbb{T} = \{2021\ 01\text{-}01\ 0{:}00 + (n-1) | n \in \mathbb{N} \cap n < 24 * 365 hour\} \tag{1}$$

なお, 実際には取得データの欠損により 355 日間のデータを使用した.

1. ES

ES を有するマイクログリッドとして, 通勤先のような電力需要及び駐車場に EV が存在するマイクログリッドと EV ユーザの各自宅で構成されるマイクログリッドがあることとした. なお, EV ユーザの自宅に関しては, シナリオ 0, 1, 2 に関わらず PV がないものとし, 負荷平準化・VGI を行わない. 4.2 節で述べたシナリオは, 通勤先のみに適用する.

2. battery(car,stationary battery)

シナリオ 2 でのみ存在する. 通勤先に存在するとした架空の定置型バッテリーと, 通勤者などによって出入りする EV のデータを保持する. 各 EV のバッテリー容量（capacity）は全て 100kWh として仮想データを生成した. EV の入退構データについては横浜国立大学に実在する研究棟一棟の実際の駐車場の自動車の出入り時刻を観測し, そこから EV50 台分の仮想データを生成した. さらに, 50 台を複製拡張し, 500 台, 1000 台, 1500 台の EV が存在する仮想のデータセットを生成した.

EV 台数の設定については, ある一日の横浜国立大学常盤台キャンパスの自動車の入構時刻を観測し, その総台数が約 600 台であったことを参考にした. 定置型バッテリー容量の設定については, 0kWh, 2500kWh,5000kWh とした.

3. state

EV は通勤のために EV ユーザの自宅と通勤先間を走行する. 各 EV については, 全国の平均通勤距離が約 10km である [23] ことから, EV ユーザの自宅と通勤先の距離は約 10km として走行消費エネルギーに関する仮想データを生成した.

4. Battery Pool : BP

シナリオ 2 でのみ存在する. BP は, 通勤先に 1 つ, 各自宅に一つずつ存在するとした.

5. Demand

シナリオ 0, 1, 2 で共通のデータを用いた. 10000 人規模の人が生活する中小規模のスマートグリッドにおける VGI を想定する. データ取得の観点から, 横浜国立大学の施設部が公開している, 構内の各建物の電力データを [1] 通勤先グリッドの需要電力データ (実データ) として用いた.

6. Generate

太陽光発電データはシナリオ 1, 2 で共通である. シナリオ 0 では存在しない. 日射量のオープンデータから以下の式（JIS C 8907:2005）を用いて算出 [2] した仮想データを生成した. H_{Am} は自治体が [2] 公開している日射量から算出した.

$$E_p = K * P_{AS} * H_{Am}/G_s \tag{2}$$

太陽光パネル設置面積について 30000m^2, 40000m^2, 50000m^2, 60000m^2 をパラメータとした. パラメータの設定について, 横浜国立大学の全建物の屋上面積の合計が約 60000m^2 であったため最大パネル面積を 60000m^2 に設定した.

7. Leveling line

あるマイクログリッドの電力平準化レベルの高さ（leveling line[kW] を日ごとに定める. [3] このとき T_{start} を 0 時, T_{end} を 23 時とした. また, 負荷平準化開始時刻を太陽光発電開始時刻とし, 負荷平準化終了時刻を 23 時とした.

$$L(\tau) = \frac{1}{T_{end} - T_{start}} \sum_{t=T_{start}}^{T_{end}} (Destination_c(t) - Source_{pv}(t)) \tag{3}$$

8. battery power flow

バッテリーの充給電は以下の条件のもとで発生しうる仮想データを生成した.

- 通勤先における負荷平準化時間帯の負荷平準化を達成するための充給電
- EV ユーザの各自宅における次の日の通勤先到着時に SOC が 55% [4] になるための EV への充電
- EV ユーザの各自宅と通勤先間における出退勤のための EV の走行

通勤先での充給電速度を 6kW, 自宅での充給電速度を最大 3kW とした. また, 各自宅, 通勤先ともに充給電効率を考慮し, 充電効率を 90%給電効率を 85%とした.

9. ES BP

通勤先の ES が唯一バッテリープールを 1 つ, 各自宅の ES が唯一バッテリープールを一つずつもつ.

10. SOC (State of Charge : バッテリー充電率)

EV の SOC は 30〜80% [4] の間に収まるように充給電を行うこととした仮想データを生成した.

11. SOH (State of Health : バッテリー健全度)

SOH は各バッテリーの出荷時の健全度を 100%としたときの各バッテリーの現在の状態を仮想データで生成した.

12. BP plug

定置型バッテリーは常に通勤先のバッテリープールに属しているものとした. EV は入退構ログから通勤先に停車中は常に通勤先のバッテリープールに属しているものとし, 自宅に到着すると自宅のバッテリープールに属しているものとし仮想データを生成した.

13. BP status

その時間にバッテリープールに属している個々のバッテリーに状態の総和で求められる仮想データを生成した.

14. Supply

Supply は Demand, Generate, Leveling line, BP status から決まる仮想データを生成した. 負荷平準化時間帯は Supply

1：横浜国立大学施設部 (学内限定),
http://shisetsu.ynu.ac.jp/gakugai/shisetsu/

2：横浜市環境創造局, http://www.city.yokohama.lg.jp/kankyo/

3：式（3）は 1 日の電力需要と供給量が定まらないと計算できない. 本来であれば, 電力平準化レベルは前もって決定しておかなければならない値である. 式（3）からは「最も理想的な」電力平準化レベルが算出される. すなわち, 本稿における VGI シミュレーションは, 理想的な電力平準化レベルを定めたという条件の下で行われたものである. 電力平準化レベルを事前に決定すること, および, 電力平準化レベルが理想値から外れた場合の影響については今後の課題とする.

4：EV のバッテリー損耗が最も小さくなるとされる SOC80%から 30%の範囲で充放電を行うことを想定する. すなわち, 1 日の開始 SOC を $\left(\frac{80\%-30\%}{2}\right) + 30\% = 55\%$ と定めた.

表 1　説明に使用した文字式

項目	単位	説明
t	なし	連続時間
$\mathbb{T}$	なし	離散時間の全体集合
τ	なし	離散時間
T_{step}	s	サンプリング間隔
T_{start}	hour	1 日の開始時刻
T_{end}	hour	1 日の終了時刻
$L(t)$	kWh/h	負荷平準化レベルの高さ
E_p	kWh	PV 発電電力量
K	なし	システム出力係数
P_{AS}	kW	標準太陽電池アレイ出力
H_{Am}	kWh/m^2	傾斜面日射量
G_s	kW/m^2	日射強度

は Leveling line に一致するように制御される．BP への充給電に過不足が生じると Supply は Leveling line からそれだけずれる．負荷平準化時間帯以外では，定置型バッテリーへの充給電と Demand の値から定まる．

15. Surplus

BP へ充電しきれないと Surplus が発生する仮想データを生成した．

16. Bp charge discharge

ES から計測した BP との充給電を記録した仮想データを生成した．

17. variable entity 使用したパラメータを表 2 に示す．なお, 充給電速度のパラメータに関しては, 普通充電を想定した定格 6kW のみとした．

5.2　スマートグリッド導入効果の多角的定量化

提案データベースに定型質問処理を行うことにより得られた結果を以下に示す．

5.2.1　シナリオごとの年間評価

図 8 にシナリオごとの年間の電力収支積み上げグラフを示す．また, このときのパラメータ設定はシナリオ 1 では, PV 設置面積 60000m^2 とした．また, シナリオ 2 では, EV 台数 1500 台, PV 設置面積 60000m^2, 定置型バッテリー容量 5000kWh とした．横軸が想定シナリオ，縦軸が電力量 [MWh] を表しており，シナリオごとに電力の供給量 (source: 左側の積み上げ棒グラフ) と供給先 (destination: 右側の積み上げ棒グラフ) を並べた．この時，グリッド内のエネルギーフローの収支は一致するので，同一シナリオの供給元と供給先の棒グラフの高さは一致する．供給元の積み上げ棒グラフにおいて，各色は以下を表す．

- 青色: 外部供給電力量の年間積算値
- 黄緑色: 太陽光発電電力量の年間積算値
- 紫色: バッテリープールからの給電量の年間積算値

供給先の積み上げ棒グラフにおいて，各色は以下を表す．

- 赤色: 消費電力量の年間積算値（基準値とする．そのため各シナリオで同じ値となる．）
- 橙色: 発電余剰電力量の年間積算値
- 緑色: EV バッテリーへの総充電量の年間積算値
- オリーブ色：定置型バッテリーへの総充電量の年間積算値

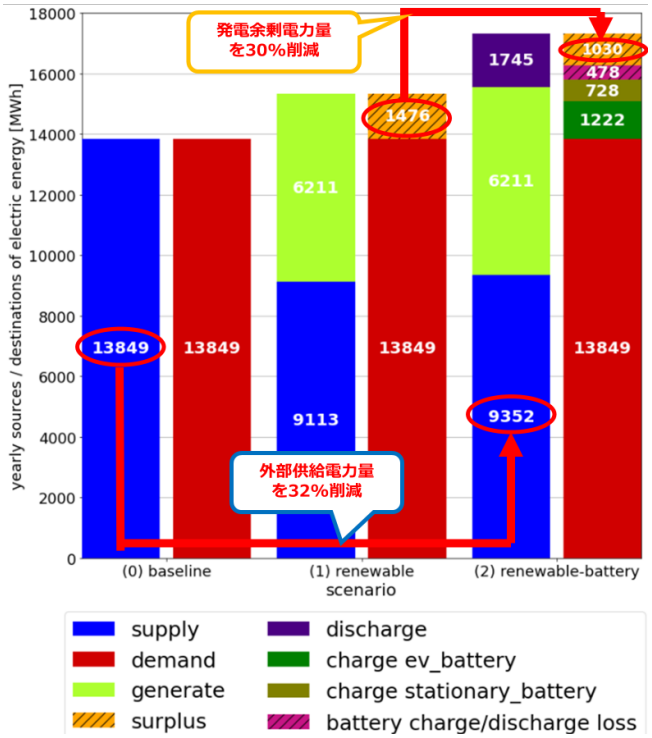


図 8　シナリオごとのエネルギーフローの年間収支

- 赤紫色: 充給電ロスの年間積算値

図 8 のシナリオ 1 と 2 の発電余剰電力量の年間積算値 (図 8 橙色) を比較すると，シナリオ 2 では 1 に比べてバッテリーの充給電管理により発電余剰電力を約 30%削減できたことがわかる．このことはシナリオ 1 では PV の発電余剰の問題が顕在化していて，一方シナリオ 2 では，VGI により問題が低減できることを示す．さらに，シナリオ 0 と 2 の外部供給電力量の年間積算値を比較すると，シナリオ 2 では 0 に比べて 1 年間で 32%の外部供給電力量を削減できたことがわかる．

以上より以下のことが示された．

(1)-1. PV だけを導入したグリッドと比較し, 負荷平準指向 VGI は発電余剰電力を 30%削減可能であること．

(1)-2. PV も EV も導入しないグリッドと比較し, 負荷平準指向 VGI は外部供給電力量を 32%削減可能であること．

5.2.2　EV への太陽光発電電力の充電量

シミュレーションごとの EV 一台当たりの一年間の総充電量に対する太陽光発電電力の充電割合を表にまとめたものを図 9 に示す．単位は%である．PV 及び EV が増加するほど太陽光発電由来充電の割合が高くなることが分かった．PV 及び EV のパラメータが最大のシミュレーションで, 69.3%の太陽光発電由来充電が可能であることが示された．この結果はカーボンニュートラルに向けた EV の充電電源の脱炭素化への貢献を示す．さらに, EV が増加すればするほど絶対量的に太陽光エネルギーで消費される電力量が増えることからも EV が増加するメリットが示されたと考えられる．

以上より以下のことが示された．

(2)-1. 負荷平準指向 VGI では PV および EV が同時に増加することで EV への太陽光発電電力の充電割合が増加すること．

表 2　パラメータ

エンティティ名	説明	パラメータ
identified ev number	EV 台数 (台)	500, 1000, 1500
pv area	太陽光パネル面積 (m^2)	30000, 40000, 50000, 60000
charge discharge speed	充給電速度（kW）	6
stationary batterry capacity	定置型バッテリー容量 (kWh)	0, 2500, 5000

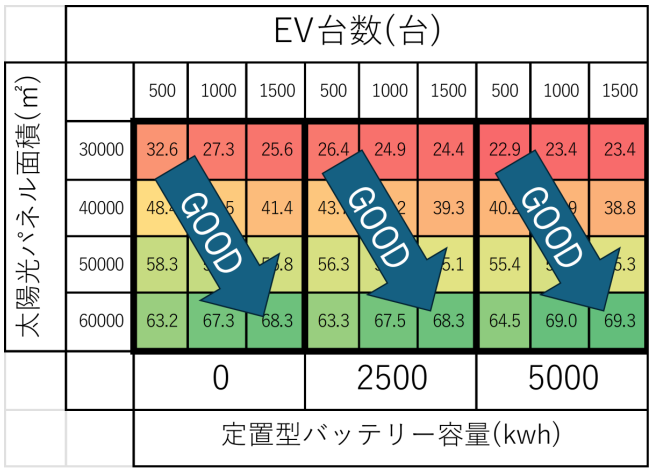


図 9　シミュレーションごとの EV 一台当たりの一年間の総充電量に対する太陽光発電電力の充電割合

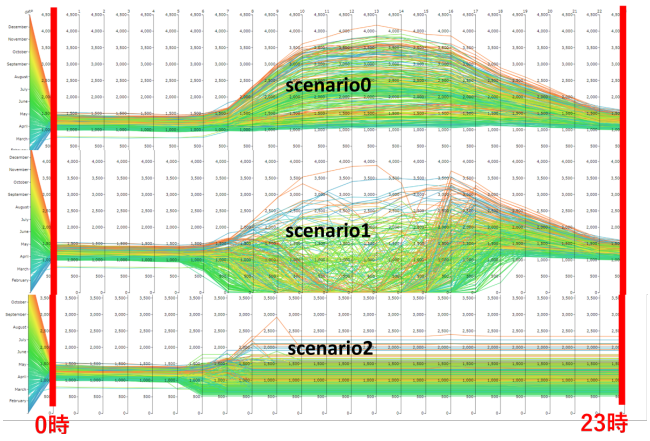


図 10　シナリオごとの実質電力需要の概形

(2)-2. 負荷平準指向 VGI により, EV への総充電量のうち 69.3%が太陽光発電電力による充電とすることが可能であること.

5.2.3　実質電力需要の変動

各シナリオの一日の実質電力需要の概形一年間分をシナリオ 0, 1, 2 の順で並べたグラフを図 10 に示す. なお可視化にあたっては我々とは別の研究グループの提案している可視化ツール (PC)^{2}DV [24] を活用した. シナリオ 2 のパラメータ設定は EV 台数 1500 台, PV 設置面積 60000m^2 定置型バッテリー容量 5000kWh である. シナリオ 1 では昼の太陽光発電電力を建物の消費に使うため昼の外部供給電力の消費が少なくダックカーブ現象を生じさせてしまっている. 一方, シナリオ 2 ではほとんどの日で完全な直線となった. 昼の太陽光エネルギーを駐車場に止まっている EV に充電し夕方のピークに対し効果的に給電が行えていると言える.

次に EV 台数を説明変数とし, 太陽光パネル面積が 60000m^2, 定置型バッテリー容量 0kWh の時の一年間の日ごとの実質電力需要の概形を図 11 に示す. この図から EV の台数が多いほど実質電力需要の推移が完全な直線である日が多いことが確認できる. つまり EV の台数が増えれば増えるほど負荷平準化が達成されやすいことがわかる.

最後に定置型バッテリー容量を説明変数とし, EV 台数が 1500 台, 太陽光パネル面積が 60000m^2, のときの一年間の日ごとの実質電力需要の概形を図 12 に示す. この図から定置型バッテリー容量 0kWh シミュレーションにくらべて定置型バッテリー 2500kWh, 5000kWh のシミュレーションではほとんどの日で負荷平準化できていることが確認できる. これは定置型バッテリーがないと EV がいない時間帯の太陽光エネルギーを有効活用できないことが原因と考えられる.

また, 結果が最も良いシミュレーションでほとんどの日 (323 日/355 日 = 約 91%) において切片の高さが一定に保たれていることが確認された. 以上より以下のことが示された.

(3)-1. 実質電力需要の変動において, PV だけを導入したグリッドと比較し, 負荷平準指向 VGI は 91%の日で切片の高さを一定に保つことが可能であること.

(3)-2. 実質電力需要の変動において, 負荷平準指向 VGI は EV が増加するほど負荷平準化が達成されやすい.

5.2.4　定置型バッテリー容量ごとの評価

定置型バッテリー容量に関して, 太陽光発電電力の充電割合の観点から 5.2.2 節図 9 では定置型バッテリー容量による差はほとんどないことが確認された. 実質電力需要の変動の観点から, 5.2.3 節図 12 では 0kWh シミュレーションと比較して, 2500kWh, 5000kWh での両シミュレーションにおいてほとんどの日で負荷平準化できていることが確認できる. しかし, 2500kWh, 5000kWh シミュレーションの間に大きな差は確認できなかった. 定置型バッテリー容量を増やすことはコストがさらに追加されることを意味する. 容量増加に伴う大きな改善効果がみられないことからコストを考慮した場合には, 負荷平準指向 VGI には適切な定置型バッテリー容量の値が存在することが考えられる. よって以下のことが確認された.

(4)-1. 負荷平準指向 VGI には適切な定置型バッテリー容量の値が存在すること.

6　ま　と　め

本論文では提案 DB を用いて, VGI シミュレーションによる

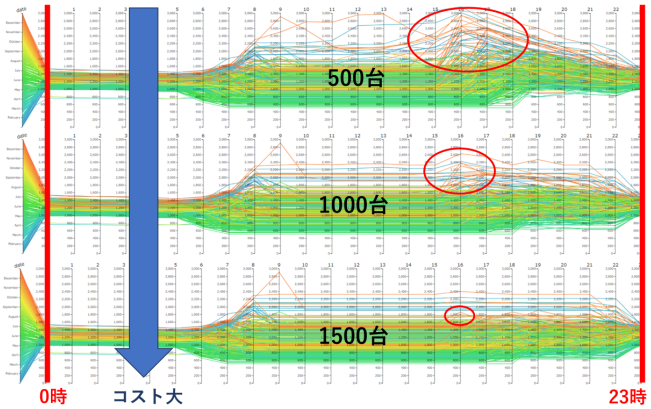


図 11　EV 台数ごとの実質電力需要の概形

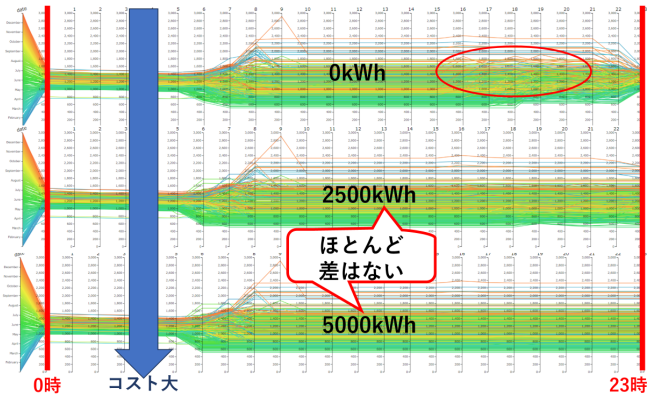


図 12　定置型バッテリー容量ごとの実質電力需要の概形

スマートグリッド導入効果の多角的定量化を行った．その結果，以下のことが明らかになった．

(1)-1. PV だけを導入したグリッドと比較し，負荷平準指向 VGI は発電余剰電力を 30%削減可能であること．

(1)-2. PV も EV も導入しないグリッドと比較し，負荷平準指向 VGI は外部供給電力量を 32%削減可能であること．

(2)-1. 負荷平準指向 VGI では PV および EV が同時に増加することで EV への太陽光発電電力の充電割合が増加すること．

(2)-2. 負荷平準指向 VGI により，EV への総充電量のうち 69.3%が太陽光発電電力による充電とすることが可能であること．

(3)-1. 実質電力需要の変動において，PV だけを導入したグリッドと比較し，負荷平準指向 VGI は 91%の日で切片の高さを一定に保つことが可能であること．

(3)-2. 実質電力需要の変動において，負荷平準指向 VGI は EV が増加するほど負荷平準化が達成されやすい．

(4)-1. 負荷平準指向 VGI には適切な定置型バッテリー容量の値が存在すること．

以上より提案 DB によってスマートグリッド導入効果の多角的定量化が可能であることを示した．また，シミュレーション結果より，このスマートグリッドの想定はある程度有用な見込みがあることがわかった．さらに，中小規模のスマートグリッドにおいても，先行研究である建物一棟のような小規模なマイクログリッドでの VGI シミュレーションと同様な結果を得られることを示した．ただし，シミュレーションに用いたデータはどちらも横浜国立大学固有のものであった．別の場所のデータを用いれば，異なる結果が示される．適用するデータの多様性を増やしていくことは今後の課題である．

謝　　辞

本研究の一部は JSPS 科研費 (課題番号 22H03810) の支援による．

文　　献

[1] 総務省. 令和元年版 情報通信白書. https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2019/2019-index.html. Last accessed: 24 December 2023.

[2] 経済産業省資源エネルギー庁. 令和元年度エネルギーに関する年次報告 (エネルギー白書 2020). https://www.enecho.meti.go.jp/about/whitepaper/2020pdf/. Accessed on 12/24/2023.

[3] 合田忠弘, 諸住哲. スマートグリッド教科書. インプレスジャパン, 初版, 2011.

[4] 東京電力ホールディングス株式会社. 長期的な安定供給とカーボンニュートラルの両立に向けた事業構造変革について. https://www.tepco.co.jp/press/release/2022/pdf2/220428j0101.pdf, 4 2022. (Accessed on 12/24/2023).

[5] 経済産業省資源エネルギー庁. もっと知りたい！エネルギー基本計画 再生可能エネルギー (1) コスト低減、地域の理解を得てさらなる導入拡大へ. https://www.enecho.meti.go.jp/about/special/johoteikyo/energykihonkeikaku2021_kaisetu01.html. Last accessed: 24 december 2023.

[6] 三菱総合研究所, 環境省. 平成２７年度低炭素社会の実現に向けた中長期的再生可能エネルギー導入拡大方策検討調査委託業務 報告書. https://www.env.go.jp/earth/report/h29-02/h27_all.pdf. Accessed on : 12/24/2023.

[7] Jiucai Zhang, Jennie Jorgenson, Tony Markel, and Kevin Walkowicz. Value to the grid from managed charging based on california's high renewables study. *IEEE Transactions on Power Systems*, Vol. 34, No. 2, pp. 831–840, 2019.

[8] Cong Zhang, Jeffery B. Greenblatt, Pamela MacDougall, Samveg Saxena, and Aditya Jayam Prabhakar. Quantifying the benefits of electric vehicles on the future electricity grid in the midwestern united states. *Applied Energy*, Vol. 270, p. 115174, 2020.

[9] 小嶋祐輔, 荒井綾希子, 阿部力也, 岩船由美子. 電気自動車・蓄電池を用いた戸建て住宅地域の電力融通システムの設計方法の確立およびその評価. エネルギー・資源学会論文誌, Vol. 42, No. 2, pp. 50–57, 2021.

[10] Narottam Das, Akramul Haque, Hasneen Zaman, Sayidul Morsalin, and Syed Islam. Domestic load management with coordinated photovoltaics, battery storage and electric vehicle operation. *IEEE Access*, Vol. 11, pp. 12075–12087, 2023.

[11] Marwa Ben Arab, Mouna Rekik, and Lotfi Krichen. Suitable various-goal energy management system for smart home based on photovoltaic generator and electric vehicles. *Journal of Building Engineering*, Vol. 52, p. 104430, 2022.

[12] 池田伸太郎, 大岡龍三. 日本国内におけるスマートシティ・スマートコミュニティ実証事業の最新動向. 生産研究, Vol. 66, No. 1, pp. 69–77, 2014.

[13] Will Simm, Maria Angela Ferrario, Adrian Friday, et al. Tiree energy pulse: exploring renewable energy forecasts on the edge of the grid. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1965–1974, 2015.

[14] Willett Kempton and Jasna Tomić. Vehicle-to-grid power

fundamentals: Calculating capacity and net revenue. *Journal of power sources*, Vol. 144, No. 1, pp. 268–279, 2005.

[15] Bernhard Jansen, Carl Binding, Olle Sundstrom, et al. Architecture and communication of an electric vehicle virtual power plant. In *2010 First IEEE International Conference on Smart Grid Communications*, pp. 149–154. IEEE, 2010.

[16] 理弘大澤, 保夫桑沢, 孝之秋元, 大介田嶌, 洋園江, 聡史上森, 正之小松. 自動車，蓄電システムを含めた住宅エネルギー性能指標とＶ２Ｈによる co2 排出量削減の効果. エネルギー・資源学会論文誌, Vol. 40, No. 3, pp. 78–84, 2019.

[17] Arnob Ghosh and Vaneet Aggarwal. Menu-based pricing for charging of electric vehicles with vehicle-to-grid service. *IEEE Transactions On Vehicular Thechnology*, Vol. 67, pp. 10268–10280, 2018.

[18] Viet Thang Tran, Md. Rabiul Islam, Kashem M. Muttaqi, and Danny Sutanto. An efficient energy management approach for a solar-powered ev battery charging facility to support distribution grids. *IEEE Transactions on Industry Applications*, Vol. 55, pp. 6517–6526, 2019.

[19] Harun Turker and Ilhami Colak. Multiobjective optimization of grid- photovoltaic- electric vehicle hybrid system in smart building with vehicle-to-grid (v2g) concept. In *2018 7th International Conference on Renewable Energy Research and Applications (ICRERA)*, pp. 1477–1482, 2018.

[20] Jonathan Donadee, Robbie Shaw, Oliver Garnett, Eric Cutter, and Liang Min. Potential benefits of vehicle-to-grid technology in california: High value for capabilities beyond one-way managed charging. *IEEE Electrification Magazine*, Vol. 7, No. 2, pp. 40–45, 2019.

[21] 石毛大貴, 能條太悟, 廣居樹, 本藤祐樹, 富井尚志. ライフログデータベーススキーマに基づくマイクログリッドにおける電力負荷平準化の価値の定量化. 研究報告データベースシステム（DBS）, 第 2022-DBS-175 巻, pp. 1–6, 2022.

[22] 廣居樹, 片平昇輝, 鈴木博登, 小松虎太郎, 本藤祐樹, 富井尚志. 太陽光発電と EV を統合する負荷平準指向スマートグリッド DB を用いた試行錯誤を伴う VGI シミュレーション. 研究報告データベースシステム（DBS）, 第 2023-DBS-177 巻, pp. 1–6, 2023.

[23] 国土交通省総合政策局. 交通基本法案検討小委員会 資料. `https://www.mlit.go.jp/common/000128770.pdf`, 11 2010. (Accessed on 12/28/2022).

[24] 能條太悟, 杉本航洋, 富井尚志. SQL ライクな操作言語を用いた可視化システムの応用とライフログ分析のためのデータ操作高速化. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), 4b-7-3, pp. 1–10, 2023.

水害時における市民の避難行動に関する意識の分析

石井 樹† 妹尾 考†† 関 洋平††† 川島 宏一†††† 鈴木 誉幸†, ‡

†筑波大学大学院システム情報工学研究群 〒305-0006 茨城県つくば市天王台１丁目１番地
††筑波大学大学院人間総合科学研究群 〒305-8550 茨城県つくば市春日１丁目２番地
†††筑波大学図書館情報メディア系 〒305-8550 茨城県つくば市春日１丁目２番地
††††筑波大学システム情報系 〒305-0006 茨城県つくば市天王台１丁目１番地
E-mail: †s2330111@u.tsukuba.ac.jp, ††s2221652@s.tsukuba.ac.jp, †††yohei@slis.tsukuba.ac.jp,
††††hkawashima@sk.tsukuba.ac.jp, ‡s2220429@u.tsukuba.ac.jp

あらまし 水害は発生の初期段階で対策を講じることが効果的であり，対策を講じる上では被災現場における市民の意識を把握することが重要である.一方で,市民の避難行動に関するアンケート調査は事後的に分析されるため，即時性が低く市民の意識を把握するには課題がある．本研究では，市民ポストの分析を通じて，被災時の市民の意識をより正確に理解する新たな手法を提案する．まず，避難行動研究に基づいて，市民意識の分析のための属性とラベルを定義し，水害時の市民ポストを都市別に分析する．分析には，人手で付与されたラベルを用いて，大規模言語モデルを微調整する方法を検討する．さらに，ラベル付きポストを利用して時系列分析を行い，都市間での比較を通じて，被災時の市民避難の現象の説明可能性について検討を行う.

キーワード 市民ポスト，災害・防災，避難行動，大規模言語モデル

1. はじめに

水害は，事前に災害の被害や規模等が想定できることから進行型災害と呼ばれており，災害発生前の初期段階で対策を講じることが効果的とされている[1]．対策を講じる上では，被災時における市民の意識を把握することが重要である．被災時の市民の意識に関しては，災害時に情報が活発に報道される状況下で危機意識を持ち始め，自身の置かれた状況を把握し避難行動の必要性を判断するために，より積極的に災害時の情報を取得する行動を取ることが明らかとなっている[2]．しかし，市民の意識を把握する上で，アンケート調査は事後的に分析されるため即時性は低い．一方，SNS は即時性が高く，中でも気軽に投稿することが可能なポスト（旧ツイート）は，市民の率直な反応を観測する上で適している．従って，ポストを適切な定義に基づいて抽出することで，市民の意識をより正確に理解することが可能になると考えられる．

そこで本研究では，水害時に市民が投稿したポストの分析を通じて，被災時の市民の意識をより正確に理解する新たな手法を提案する．まず，従来の避難行動研究に基づき市民の意識を分析するための属性とラベルを定義し，ポストを都市別に分析する．分析には，人手で付与されたラベルを用いて，大規模言語モデルを微調整する手法を検討する．次に，ラベル付きポストを利用した時系列分析を行い，都市間での比較を通じて，被災時における市民の避難行動の現象の説明可能性について検討を行う．最後に，大規模言語モデルを用いた自動分類の実用性についても検討を行う．

2. 関連研究

2.1. 水害時における市民の避難行動の調査方法

市民の避難行動は，アンケート調査によって実態が把握されることが多い．田中ら[3]は，豪雨，水害，避難などをキーワードに論文検索を行い，大雨災害時の避難行動に関する学術論文を網羅的に収集した．その結果，128 編が収集され，それらの調査方法の内訳（重複あり）は，アンケート 98 編，ヒアリング 27 編，その他 16 編であり,住民や関係機関へのアンケートによる調査が多く行われていることを明らかにした.一方，柳森ら[4]は，住民に対するアンケート調査は記憶に依存するため，回答精度が低下する点を指摘している．

2.2. 災害時におけるポストの利活用と処理技術

災害時におけるポストの利活用に関する研究では，市民による利用実態把握及び処理技術の開発が行われている．佐藤[5][6]は，災害対応に SNS は有効であるという立場と有効でないという立場に分けて既存事例をレビューし，それらの議論が，処理技術，質と両の問題（質が高いポストが少ない，不要不急のポストの量が多い），に帰着することを示している．その上で，SNS 利用者に対する啓発を行っている．柳森ら[4]は，台風接近期間のポストを大規模に収集し，避難情報に対する住民の反応を主観的に手作業で分類することで，避難情報発令前後での住民の反応度合いを観察している．貝塚ら[7]は，災害時におけるポストの中から，人に行動を促すポストを抽出するための 3 つの手法を提

案し，精度の評価を行い最も優れた手法を検証している．いずれの研究を踏まえても，災害時におけるポストの有効活用においては，リテラシー，分類基準，抽出手法といった一定の規則が示されている．

2.3. 水害時における市民の避難行動の分析方法

市民による避難行動の促進を図るため，従来の研究では，市民の避難行動を把握するために次の分析方法が用いられてきた．片田ら[2]は，災害情報の取得，危機意識，情報取得態度・行動，対応行動の観点から分析を行い，水害時における市民の情報取得行動が一連の循環構造にあることを明らかにした．竹内[8]は，危険認識，行動決定，行動実施の観点から分析を行い，台風接近に伴う市民の災害対応の行動順序が 6 パターンに分類されることを明らかにした．柿本ら[9]は，防災情報の取得，脅威評価，避難意志，避難行動実施の観点から分析を行い，豪雨時に市民が避難意思を持てなかった場合でも，地域の呼びかけが避難行動を促すきっかけとなり得ることを明らかにした．いずれの研究においても，市民の避難行動に関する分析方法は，災害情報・防災情報といった情報の取得，情報の取得を踏まえた危険意識の形成，危険意識に基づく避難行動の検討・実施に大別される．

2.4. 大規模言語モデルによる災害時のポストの自動分類

大規模言語モデルを用いたテキスト分類の技術は，災害に関する市民のポストの分類に応用されている．Zahera ら[10] は，大規模言語モデルの一種である BERT [11] を用いて，台風や地震発生時のポストに対して「救助要請」や「ボランティア志望」等の 24 値のマルチラベル分類を行った．また，Madichett ら[12] や Adwaith ら[13] は，BERT の後継モデルである RoBERTa [14] を用いて，災害時のポストに被害報告等の有益な情報が含まれるか否かの 2 値分類を行った．このように，災害時の市民ポストの自動分類に関する研究では，主に救助要請や被害報告等の災害対応において重要なポストを抽出することに焦点を当てている．一方で，本研究においては，被災した市民の避難行動を分析するための属性を定義して，大規模言語モデルによる自動分類を行う点で，独自の試みとなっている．

また，石田ら[15] や Senoo ら[16] は，行政課題に対する市民の意見や立場の分類に大規模言語モデルの T5 [17] が有効であることを示した．加えて，近年は ChatGPT [18] に代表される大型の大規模言語モデルの利活用が注目されている．本研究では RoBERTa に加えて T5，および，日本語対応のオープンソース大型大規模言語モデルの Stockmark-13B（以下，Stockmark）のいずれが市民の避難行動の分類により適しているか検証するため，それぞれの事前学習モデルを人手で作成したデータセットを用いて微調整して分類モデルを構築し，分類精度を比較する．

3. 提案手法

2.3 節の整理を踏まえ，本研究では，水害に対する市民の状況認識，避難行動，他者への呼びかけの意識を把握するため,『状況認識』,『避難行動』,『呼びかけ』,『その他』を属性として定義する．更に，各属性の中に複数のラベルを設定することで，市民の意識及び行動をより正確に把握することを試みる．以下の表 1 は，本研究で定義する属性及びラベルの対応関係を示したものである．

表 1: 属性・ラベルの対応

属性	ラベル
状況認識	防災情報，周辺環境，危険認識
避難行動	検討，決定，避難準備， 安全確保，振り返り
呼びかけ	市民，行政
その他	

3.1. 属性・ラベルの定義

- **状況認識**

台風の進行及び災害の危険性に対する市民の状況認識を把握するために定義する．ラベルには『防災情報』,『周辺環境』,『危険認識』を設定する．『防災情報』には，避難指示，大雨特別警報など，公的機関が発令した情報を取得しているものが該当する．『周辺環境』には，雨量の強さ，家族や友人の避難状況及び被害状況など，台風進行に伴う周辺環境の変化を把握したものが該当する．『危険認識』には，台風の進行状況を把握した上で，災害の危険性を危険又は安全と判断しているものが該当する．

- **避難行動**

市民の避難行動に伴う一連の意識を把握するために定義する．ラベルには『検討』,『決定』,『避難準備』,『安全確保』,『振り返り』を設定する．『検討』には，避難行動に対する躊躇，避難の候補先及び避難に伴う必要物品の確認を行っているものが該当する．『決定』には，避難行動を決意したものが該当する．『避難準備』には，食料品の準備及び車両・ペットを移動させているものが該当する．『安全確保』には，避難所又は自宅及び親戚宅に移動し，身の安全性を確保しているものが該当する．『振り返り』には，自身の避難行動及び台風による被害・影響を顧みているものが該当する．

- **呼びかけ**

市民による他者及び公的機関に対する情報の発信状況を把握するために定義する．ラベルには『市民』,『行政』を設定する．『市民』には，自身の家族・友人に対

し，避難行動を促しているもの，気遣いの言葉を掛けているものが該当する.『行政』には，自治体の発令した避難情報及び避難所の開設に関して，疑問・意見・要望を主張しているものが該当する.

- **その他**

上記以外のものが該当する．具体的には，台風及び避難行動と関係の無い話題について言及しているもの，避難所リストのみの情報や，新聞記事の転載など，投稿者の意識が判別困難であったものが該当する.

4. 実験：市民ポストに対する自動分類の検証

4.1. データセットの構築

本研究で使用するデータセットは，つくば市，仙台市，横浜市，名古屋市の計 4 都市における市民ポストを基に構築される．各都市で 500 件の市民ポストを収集・サンプリングし，全体で 2,000 件のデータセットとなる．本データセットを用いて，人手によるアノテーションを実施し，大規模言語モデルを用いた自動分類，ラベル付きポストを利用した時系列分析を行う.

4.1.1. 対象事例

本研究では，2019 年に東日本を縦断し人的被害を生んだ災害として令和元年台風第 19 号（以降，令和元年台風と呼ぶ）を対象事例とする．令和元年台風は，日本国内で 2019 年 10 月に甚大な被害をもたらし，特に静岡県や新潟県，関東甲信地方，東北地方の多くの地点で記録的な大雨になったと記録されている．この大雨に対して，10 月 12 日 15 時 30 分から静岡県，神奈川県，東京都，茨城県，宮城県等の 1 都 12 県に大雨特別警報が発表された[19].

4.1.2. 対象都市・期間

本研究の対象都市は，令和元年台風による影響・被害を受けたとされており，かつ市民ポストが十分に収集可能な都市として，人口 20 万人以上の横浜市，仙台市，つくば市，名古屋市とする．対象ポストは，台風が接近した 2019 年 10 月 11 日 0 時 00 分から同年 10 月 13 日 23 時 59 分までの 3 日間で,各都市に居住していると推定された市民が投稿したものとする.

4.1.3. 市民ポストの収集

X（旧 Twitter）からプロフィール情報に基づき収集された横浜市民，仙台市民，つくば市民，名古屋市民の市民ポストを Twitter の Streaming API を用いて収集した. なお, リツイートは排除した. クエリは,「避難」という単語を含んだ市民ポストを収集した. その結果, 横浜市 4,157 件，仙台市 1,137 件，つくば市 805 件，名古屋市 912 件, 計 7,011 件の市民ポストを収集した.

表 2: 令和元年台風期間内の市民ポスト件数

横浜市	仙台市	つくば市	名古屋市	総計
4,157	1,137	805	912	7,011

4.1.4. 市民ポストの抽出

本稿では，市民ポストを用いた都市間比較を行うことを踏まえ, 各都市 500 件かつ 4 都市分となる計 2,000 件の市民ポストを抽出する．抽出に当たっては，標本に偏りが生じることを防ぐため，日数（3 日間），時間帯（24 時間）のセグメント（72 区分）を作成する．その後，各セグメントに含まれる市民ポストの件数に対し，抽出比率（500 件を 3 日間の市民ポスト総数で除した値）を乗じることで，抽出する件数を算出する．例えば，横浜市における 10 月 11 日 0 時の抽出件数を算出する場合，当該時刻に実際に投稿された市民ポストが 8 件に対し，抽出比率となる 0.12（500 件を横浜市の 3 日間の市民ポストの総数 4,157 件で除すため）を乗じると，抽出する件数は 1 件となる．以上の抽出作業をセグメント毎及び都市毎に行い，市民ポストを抽出する.

4.1.5. 人手による各属性のアノテーション

抽出した市民ポストに対し，第 3 章で定義したラベルを人手で付与することで,データセットを構築する. なお，本研究では 1 つの市民ポストに対し複数のラベルを付与する．本稿では，著者 1 名が行ったアノテーション 2,000 件を用いて実験を行う．また，著者のアノテーションの一貫性を確認するため，作業協力者 2 名と共通の市民ポスト 300 件をアノテーションし，Cohen の κ 係数[20]を算出した．また，同時に Fleiss の κ 係数をシングルラベル分類[21]及びマルチラベル分類[22]で算出した．その結果，全ての属性において Cohen の κ 係数（協力者 1 及び協力者 2 とも）及び Fleiss の κ 係数（シングルラベル分類）の値は 0.6（Substantial Agreement [23]）を上回った．また，Fleiss の κ 係数（マルチラベル分類）の値は 0.678 となった.

表 3: 各属性の κ 係数

属性	κ 係数 (対協力者 1)	κ 係数 (対協力者 2)	κ 係数 (Fleiss)
状況認識	0.791	0.718	0.784
『防災情報』	0.820	0.785	0.812
『周辺環境』	0.720	0.644	0.074
『危険認識』	0.725	0.610	0.709
避難行動	0.809	0.799	0.813
『検討』	0.708	0.951	0.750
『決定』	0.728	0.946	0.819
『避難準備』	0.788	0.848	0.747
『安全確保	0.775	0.695	0.651
『振り返り』	0.755	0.681	0.830
呼びかけ	0.831	0.729	0.830
『市民』	0.877	0.785	0.842
『行政』	0.764	0.683	0.727
その他	0.752	0.634	0.796

4.2. 実験：大規模言語モデルを用いた自動分類

4.2.1. 目的

本節では，大規模言語モデルを微調整して分類モデルを構築し，市民の避難行動を分析するための属性ラベルに対する分類精度を属性ごとに確認し，大規模言語モデルを用いた自動分類の有効性について検証する．

4.2.2. 方法

4.1 節で作成したデータセットを用いて，5 分割交差検証により分類モデルの訓練と評価を行う．分類の有効性は，5 分割交差検証の各回におけるテストデータのマクロ平均 F 値の平均に基づき評価する．

分類モデルは，大規模言語モデルに分類用の線形変換層を接続したものを使用し，各モデルを微調整した結果について分類精度を比較する．大規模言語モデルには日本語で事前学習済みの RoBERTa[1]と T5[2]，Stockmark[3]を用いる．各モデルのパラメータ数は RoBERTa と T5 が約 1 億，Stockmark が約 130 億である．大規模言語モデルが出力する埋め込み表現を線形変換層への入力ベクトルに変換する方法については，①*[CLS]*トークンの埋め込みベクトルを抽出，②最大値プーリング，③平均値プーリングの 3 種類のプーリング方法の中から，後述する予備実験で最良の結果となる方法を適用する．また，本研究では，Zahera [10] のように，1 つのポストに同時に割り当てられる複数の属性を推定するため，マルチラベル分類を実現する．そこで，線形変換層の出力ベクトルの各要素をシグモイド関数で各ラベルへの予測確率に変換し，確率が 0.5 以上のとき当該ラベルに該当すると判断する．

訓練のための最適化関数には AdamW [24] を使用する．学習率はスケジューラを用いて線形に減衰[4]させる．損失関数は各ラベルに対するバイナリクロスエントロピーの平均である．さらに，各ラベルの損失は以下の式(1)で計算した係数により重みづけする．

$$w_i = \log\left(\frac{\sum_j f_j}{f_i}\right) \quad (1)$$

ここで，w_iはラベルiの重み，f_iはラベルiが割り当てられた訓練データ数である．

その他のハイパーパラメータは予備実験により決定した．具体的には，バッチサイズや学習率等を変更しながら訓練を繰り返し，検証データ[5]に対する F 値のマクロ平均が最大となる組み合わせを採用した．採用したハイパーパラメータを表 4 に示す．

表 4: 大規模言語モデルの微調整用ハイパーパラメータ

項目	RoBERTa	T5	Stockmark
学習率	0.00004	0.0005	0.00004
スケジューラの Warmup ステップ数	50	100	0
プーリング方法	*[CLS]*	最大値	平均値
最大トークン長	128		
バッチサイズ	16		
最大エポック数	50（検証データの損失の最小値が 2 エポック連続で更新されなかった場合，早期終了）		
シード値	100		

4.2.3. 結果

RoBERTa と T5，Stockmark の各ラベルの F 値とマクロ平均 F 値を表 5 に示す．

表 5: 実験結果（F 値）

ラベル	RoBERTa	T5	Stockmark
状況認識（防災情報）	**0.840**	0.807	0.812
状況認識（周辺環境）	**0.496**	0.453	0.483
状況認識（危険認識）	**0.643**	0.562	0.516
避難行動（検討）	**0.661**	0.629	0.572
避難行動（決定）	**0.531**	0.498	0.392
避難行動（避難準備）	**0.558**	0.487	0.401
避難行動（安全確保）	**0.689***†	0.513	0.461
避難行動（振り返り）	**0.546**	0.512	0.446
呼びかけ（市民）	**0.811***†	0.757	0.714
呼びかけ（行政）	**0.648**	0.544	0.637
その他	**0.598**	0.578	0.569
マクロ平均 F 値	**0.638***†	0.576	0.546

* T5 に対して両側t検定による有意差あり（5%水準）

† Stockmark に対して両側t検定による有意差あり（5%水準）

また，各モデルのマクロ平均適合率とマクロ平均再現率を表 6 に示す．

表 6: 実験結果（マクロ平均適合率・再現率）

評価指標	RoBERTa	T5	Stockmark
適合率	0.693	0.659	**0.745**
再現率	**0.610***†	0.536	0.457

* T5 に対して両側t検定による有意差あり（5%水準）

† Stockmark に対して両側t検定による有意差あり（5%水準）

4.2.4. 考察

表 5 から，マクロ平均 F 値は RoBERTa が 0.638，T5 が 0.576，Stockmark が 0.546 となり，RoBERTa が最も精度の高い結果となった．2 番目に精度が高い T5 と RoBERTa の差は両側t検定により統計的に有意差がある（$t(4) = 4.276,\ p = 0.013,\ r = 0.906$）．ここで，$r$ は効果量（Pearson's r）を指し，0.5 以上であることから F 値の向上の程度が大きい [25] ことを示している．

また，ラベル別に見ると，全てのラベルで RoBERTa が最も高い値を示しているが，とりわけ『避難行動（安全確保）』と『呼びかけ（市民）』は，T5 と Stockmark

[1] https://huggingface.co/rinna/japanese-roberta-base
[2] https://huggingface.co/sonoisa/t5-base-japanese（Encoder 部分のみを用いる）
[3] https://huggingface.co/stockmark/stockmark-13b
[4] Hugging Face Transformers ライブラリの *get_linear_schedule_with_warmup* 関数を使用
[5] テストデータを除いた 20%のデータを検証データとして使用

と比べて両モデルとも両側t検定により有意差が確認された．F 値の向上の幅が最も大きい『避難行動（安全確保)』について，RoBERTa でのみ正しくラベルを推定できた例[6]を以下に示す．

- 「昨日は自分の地域で避難指示が出たので、子供が入学予定の学校に避難。さっき帰宅したけど、家にまで水来てなくてホッとした。家の前は冠水してたけど、今朝はもう引いてた。ひとまず、良かった。」

上記の例は「学校に避難」と述べているため『避難行動（安全確保)』に該当する．一方で，避難指示が出ていることを報告していることから『状況認識（防災情報)』，家の前の冠水を報告していることから『状況認識（周辺環境)』，台風通過後に「良かった」と所見を述べていることから『避難行動（振り返り)』にも該当する．こうした複数のラベルが割り当てられるデータは文章の表現が複雑になるため，正確な推定にはより高度な言語能力が必要となる．実際に，RoBERTa でのみ正しく『避難行動（安全確保)』と推定できた 24 件のポストのうちの 5 件については，3 つ以上のラベルが割り当てられたポストであったのに対し，RoBERTa，T5，Stockmark が共通して『避難行動（安全確保)』と推定できた 27 件については，3 つ以上のラベルが割り当てられたポストは存在しなかった．以上から，RoBERTa が，市民の避難行動を正確にマルチラベルで推定するための言語能力を有したモデルであるといえる．

一方，Stockmark について表 6 を見ると，3 つのモデルの中で再現率が最も低いが，適合率が最も高いことから，他のモデルと比べて傾向が大きく異なる．これは Stockmark が各ラベルへ該当すると推定した件数が少ない一方，判断が正確で，かつ，非該当のラベルを正しく非該当と推定できたためである．以下に『避難行動（振り返り)』について，Stockmark でのみ正しく非該当と推定できた例を示す．

- 「お陰で旦那は無事に帰宅しました。お騒がせしました…ご心配下さった皆さま、ありがとうございます🙇‍♀️でも…仙台市全域避難指示になってしまった…💦」

上記の例には『避難行動（振り返り)』に該当するような情報は見受けられない．しかし，「しました」や「しまった」などの過去形の表現を多く含むことから，RoBERTa と T5 は避難行動を振り返っていると誤って推定した可能性がある．上記の例以外にも RoBERTa と T5 が推定を誤ったポストには「無事だった」「避難した」等の振り返りとは異なる意図の過去形の表現が多く見られた．Stockmark はこうした微妙な文意の違いを正確に読み取り慎重に各ラベルの判断を行った結果，該当とした件数が少ない一方で正確な推定ができたと予想される．よって，Stockmark を市民の避難行動の分類モデルとして運用するには，慎重なラベル推定の能力を維持しつつ，各ラベルへ該当と推定する件数を増やし再現率を向上させることが課題といえる．

5. 属性とラベルを用いた避難に関するポストの時系列推移の分析と都市間比較

5.1. 分析：人手によるラベルの集計・比較

5.1.1. 目的・方法

都市間で避難行動に関する意識に差異が見られたかを分析するため，4.1 節で構築したデータセットを用いて，各属性及び各ラベルの内訳を都市別に算出し，結果を比較する．なお，本研究では 1 つのポストに対し複数のラベルを付与しており，各属性及び各ラベルの総和が各都市のポストの件数を上回ることがある．

5.1.2. 結果・考察

図 1 に示すように，横浜市，仙台市，つくば市において，ラベルの比率の高い順は『状況認識』，『避難行動』，『呼びかけ』となっている．一方，名古屋市においては，ラベルの比率の高い順は『呼びかけ』，『避難行動』，『状況認識』であり，『避難行動』及び『状況認識』に関する投稿の比率は低い．このことから，名古屋市では，他の 3 都市と比較して台風による市民の避難行動に関する意識への影響度合いが相対的に高くなく，『状況認識』及び『避難行動』に関する投稿が行われない傾向にあったと考えられる．下記は，名古屋市で『呼びかけ』に分類されたポストの一例である．

- 「名古屋市内、本当に大型台風だったの？というくらい、酷くならなかった。ありがたいことだ。今も関東方面は避難勧告出たり、被害が出てるとこもあるようなので、お気をつけてください。」

上記の例からは，名古屋市内における台風による被害は大きくなく，他の都市の市民の台風による被害を心配する様子が伺われることから，自身の避難行動に関する意識よりも，他者及び他地域に対する『呼びかけ』を行うことが，名古屋市民の意識に上っていた可能性があると考えられる．

[6] ユーザーの特定回避のため，文面の表現を一部変更した．以下の例も同様．

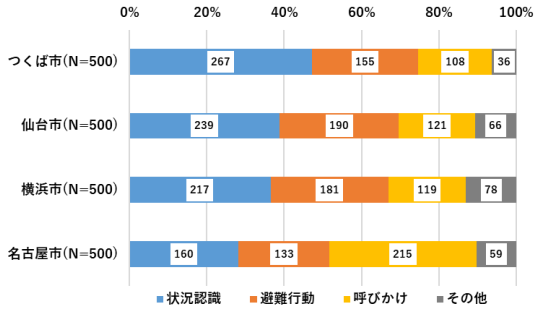


図 1: 4 つの属性の内訳

5.1.3. 状況認識ラベルの比較

図 2 に示すように，横浜市，仙台市，つくば市では，『防災情報』,『周辺環境』,『危険認識』の順に比率が高い．一方，名古屋市では,『周辺環境』,『防災情報』,『危険認識』の順に件数が低い．以上より，台風による影響・被害が相対的に高くなかったと考えられる名古屋市を除き，他の 3 都市に関しては，市民が『状況認識』に関する投稿を行った内訳として,『周辺環境』よりも『防災情報』に関する投稿が行われる傾向にあったことが示された．

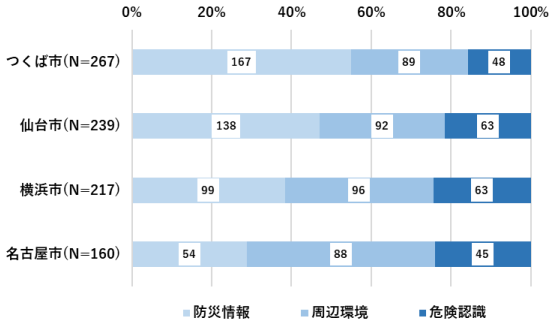


図 2: 『状況認識』ラベルの内訳

5.1.4. 避難行動ラベルの比較

図 3 に示すように，横浜市，仙台市，つくば市では，『検討』のラベルの比率が最も高く，名古屋市では，『振り返り』に関するラベルの比率が最も高い．従って，比較的台風による影響・被害が高かったと推察される都市では,『避難行動』に関する投稿の内訳として,『検討』に関する投稿が行われる傾向にあることが示された．

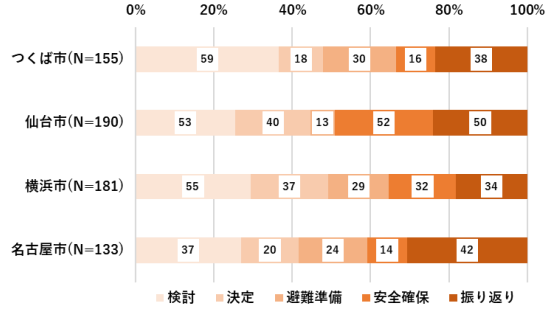


図 3: 『避難行動』ラベルの内訳

5.1.5. 呼びかけラベルの比較

図 4 に示すように,『呼びかけ』のラベルの内訳については，全ての都市において,『市民』,『行政』の順にラベルの比率が高く,『行政』よりも『市民』に対する『呼びかけ』の方が多く投稿される傾向にあることが示された．

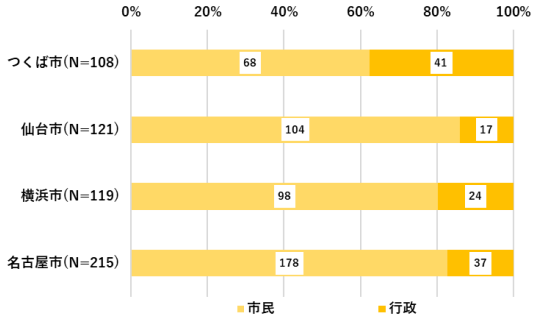


図 4: 『呼びかけ』ラベルの内訳

5.2. 分析 : 市民ポストの件数の時系列推移

5.2.1. 目的・方法

時間の経過に伴う市民ポストの件数の推移を把握し，都市間及びラベル間で比較を行うため，市民ポストに付与されたラベルの件数を時系列に集計・可視化し，都市間及びラベル間の観点から考察を行う．

5.2.2. 結果・考察 1 : 都市別における件数の推移

図 5 は都市別における件数の推移を示した結果であり，左上が横浜市，右上が仙台市，左下がつくば市，右下が名古屋市となっている．

まず,各ラベルの累積の件数に着眼すると,横浜市，仙台市，つくば市では,『状況認識』,『避難行動』,『呼びかけ』,『その他』の順に件数が多く，名古屋市では,『呼びかけ』,『状況認識』,『避難行動』,『その他』の順に件数が多くなった．従って，横浜市，仙台市，つくば市では，特に『状況認識』及び『避難行動』という自分自身が避難行動を取る際の事柄を含むポストが多く投稿される傾向にあったと考えられる．

次に，ポストの件数が増加した時点に着眼すると，各都市では 10 時 11 日 0 時から件数が横ばいで推移後，名古屋市及び横浜市では 10 月 12 日 9 時時点を境として,つくば市及び仙台市では 10 月 12 日 12 時時点を境として件数が増加していることから，名古屋市及び横浜市の方が，つくば市及び仙台市よりも早期にポストの件数が増加する傾向にあったことが示された．並びに，名古屋市及び横浜市，つくば市及び仙台市において，両グループ内の都市で件数の増加がほぼ同一時刻であった点に関しては，全国版の報道を介して台風に関する情報が発信され，それを受信した市民がほぼ同一時刻に投稿を行った可能性があると考えられる．

更に，横浜市，仙台市，つくば市における各ラベル

の件数の順位に着眼すると，各都市においても，『避難行動』又は『呼びかけ』の件数が先に増加した後に，『状況認識』がそれらの件数を上回る傾向が見られており，他の市民が投稿した『避難行動』又は『呼びかけ』に反応する形で，『状況認識』に関する投稿が行われた可能性があると考えられる．

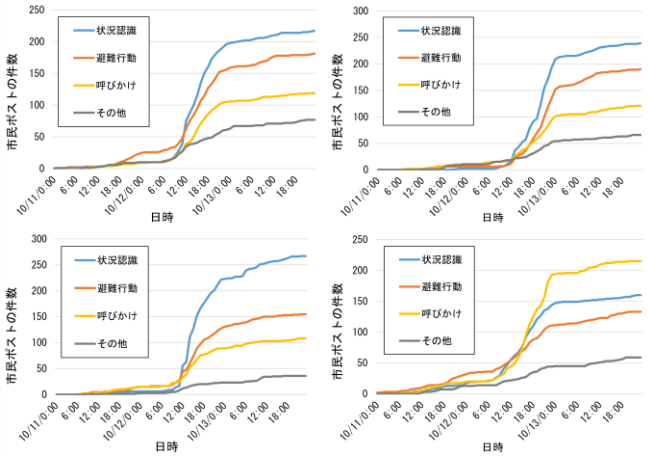


図 5: 横浜市・仙台市・つくば市・名古屋市
(左上・右上・左下・右下)

5.2.3. 結果・考察 2：ラベル別における件数の推移

図 6 はラベル別に件数の推移を示した結果であり，左上が『状況認識』，右上が『避難行動』，左下が『呼びかけ』，右下が『その他』となっている．

まず，『状況認識』に着眼すると，10 月 12 日 12 時時点では名古屋市，横浜市，つくば市の順にポストの件数が多く，その後，横浜市，つくば市，仙台市の順に名古屋市の件数を上回っている．また，『避難行動』に関しても，10 月 12 日 12 時時点では名古屋市の件数が最も多く，時間が経過するにつれ，横浜市，つくば市，仙台市の順に，名古屋市の件数を上回っている．一方，『呼びかけ』に関しては，常に名古屋市の件数が他の 3 都市を上回っている．

以上より，『状況認識』及び『避難行動』という自分自身が避難行動を取る際の事柄を含むラベルに関しては，ポストの波形が名古屋市から，横浜市，つくば市，仙台市の向きに平行移動したと解釈することも可能であり，時系列モデルを用いた分析を行い，ポストという波形の変動を予測することについても検討の余地があると考えられる．

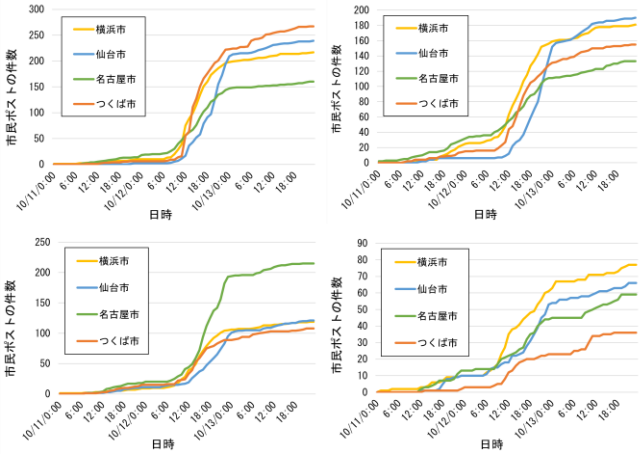


図 6: 状況認識・避難行動・呼びかけ・その他
(左上・右上・左下・右下)

6. おわりに

6.1. 総括

本研究の成果は次の通りである．1 点目は，従来の避難行動研究に基づき，市民の避難行動に関する意識を把握するための新たな属性とラベルを定義した点である．2 点目は，大規模言語モデル RoBERTa を用いてポストのマルチラベル分類を行う枠組みを実現し，各属性についての分類の効果と課題を示した点である．3 点目は，人手で付与されたラベル付きポストを用いた分析を行い，水害時における市民の避難行動に対する意識の把握を都市別に試み，違いを明らかにした点である．

6.2. 自動分類の実用性

本研究の成果を踏まえ，自動分類の実用性について検討を行う．本研究では 2,000 件のデータセットを用いた分析を行ったが，今後更にデータの規模を増やした分析を行う上では，アノテーションの作業協力者を募る代わりに，大規模言語モデルを用いた自動分類が実用的と考えられる．実際，作業協力者とのアノテーション方針の認識統一を図るには時間的コストが必要であり，かつ避難行動分野は一定の専門知識を有することが望まれるため，作業協力者には当該分野を学習するコストも発生し得る．従って，避難行動分野においてアノテーション作業を行ってゆく上では，専門家が非専門家に対しアノテーションの訓練を行う代わりに，専門家が機械に対し訓練を行う方が，必要となるコストが相対的に小さくなる可能性を念頭に置くことが重要と考える．

6.3. 今後の課題

本研究の課題は次の通りである．1 点目は，ラベル付きポストの件数を追加するため，大規模言語モデルを用いた自動分類も念頭に置いた上で，アノテーション作業に伴うコストの削減を図る点である．2 点目は，

日本語に特化した大規模言語モデルを用いた検証を行い，マルチラベルで正確に推定する上で適したモデルを検討する必要がある点である．3 点目は，今回定義したラベルが，他の都市及び他の自然災害の種別においても汎用性があるかを検証する点である．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 B（課題番号 23H03686），挑戦的研究（萌芽）（課題番号 22K19822）の助成を受けて遂行されました．また，筑波大学公共イノベーション研究室の皆様には，人手によるアノテーション作業をお手伝いいただきました．心より御礼申し上げます．

参　考　文　献

[1] 国土交通省，水災害に関する防災・減災対策本部，防災行動計画ワーキンググループ，“タイムライン（防災行動計画）策定・活用指針(初版)”，2016，p.4．https://www.mlit.go.jp/river/bousai/timeline/pdf/timeline_shishin.pdf

[2] 片田敏孝，児玉真，及川康，“水害進展過程における住民の災害情報の取得構造に関する実証的研究”，土木学会論文集，2005，No.786，IV-67，p.77-88.

[3] 田中皓介，梅本通孝，糸井川栄一，“既往研究成果の系統的レビューに基づく大雨災害時の住民避難の阻害要因の体系的整理”，地域安全学会論文集，2016，Vol.29，p.185-195.

[4] 柳森和真，井料隆雅，“Twitter データを用いた水害時の避難情報に対する住民の反応の分析”，情報処理学会第 77 回全国大会講演論文集，2015，Vol.1，p.149-150.

[5] 佐藤翔輔，“災害対応における SNS の有効性と限界: 東日本大震災の発生から 7 年をふりかえる”，災害情報，2018，No.16-2，p.143-148.

[6] 佐藤翔輔，邑本俊亮，立木茂雄，“2016 年～2020 年の災害事例にみる被災地内における災害時の SNS の利用実態”，自然災害科学，2022，No.41-2，p.107-120.

[7] 貝塚圭一，米田吉希，鈴木優，灘本明代，“災害時の行動促進ツイート抽出手法”，第 11 回データ工学と情報マネジメントに関するフォーラム（DEIM2019），2019，I7-4.

[8] 竹之内健介，“令和元年東日本台風における住民の災害対応の特徴と時間変化に基づく行動分類”，自然災害科学，2022，Vol.41，No.1，p.5-22.

[9] 柿本竜治，吉田護，“状況認識を考慮した令和 2 年 7 月豪雨時の避難行動意思決定モデルの推定”，土木学会論文集 D3（土木計画学），2022，Vol.78，No.2，p.45-57.

[10] Hamada M. Zahera, Ibrahim Elgendy, Rricha Jalota, Mohamed Ahmed Sherif. “Fine-tuned BERT Model for Multi-Label Tweets Classification”. The 28th Text REtrieval Conference (TREC), 2019, pp. 1-7.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Proc. of the 2019 Conf. of the North American Chapter of the ACL, Volume 1 (Long and Short Papers). ACL, 2019, pp. 4171-4186.

[12] Sreenivasulu Madichetty, Sridevi M. A Neural-Based Approach for Detecting the Situational Information from Twitter During Disaster. IEEE Transactions on Computational Social Systems. 2021, Vol. 8, No. 4, pp. 870-880.

[13] Divakaran Adwaith, Ashok Kumar Abishake, Siva Venkatesh Raghul, Elango Sivasankar. Enhancing Multimodal Disaster Tweet Classification Using State-of-the-art Deep Learning Networks. Multimedia Tools and Applications, 2022, Vol. 81, pp. 18483-18501.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. arXiv:1907.11692 [cs.CL], 2019.

[15] 石田 哲也，関 洋平，欅 惇志，柏野 和佳子，神門 典子．都市を横断した市民意見抽出の評価．自然言語処理，2023. Vol. 30, No. 2, pp. 586-631.

[16] Ko Senoo, Yohei Seki, Wakako Kashino, Noriko Kando. “Visualization of the Gap Between the Stances of Citizens and City Councilors on Political Issues”. From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries (ICADL 2022). Lecture Notes in Computer Science. Springer, 2022, Vol. 13636, pp. 73-89.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. The Journal of Machine Learning Research, 2020, Vol. 21, No. 1, pp. 5485-5551.

[18] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, et al. “GPT-4 Technical Report”. arXiv: 2303.08774 [cs.CL], 2023.

[19] 気象庁，“災害をもたらした気象事例台風第 19 号による大雨，暴風等”，2019．https://www.data.jma.go.jp/obd/stats/data/bosai/report/2019/20191012/20191012.html

[20] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960, Vol. 20, No. 1, p. 37-46.

[21] Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin. American Psychological Association, 1971, Vol.76，No.5，p.378‐382.

[22] 東中竜一郎，荒木雅弘，塚原裕史，水上雅博，“雑談対話システムにおける対話破綻を生じさせる発話の類型化”，自然言語処理，2022，Vol.29，No.2，p.445-466.

[23] J. Richard Landis, Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. Biometrics, 1977, Vol. 33, No. 1, p. 159-174.

[24] Ilya Loshchilov, Frank Hutter, “Decoupled Weight Decay Regularization”. arXiv:1711.05101 [cs.LG], 2019.

[25] Jacob Cohen. A Power Primer. Psychological Bulletin, 1992, Vol. 112, No. 1, pp. 155-159.

人口動態を基盤としたソーシャルセンサのパフォーマンス分析

大西駿太朗† 矢田竣太郎† 若宮 翔子† 荒牧 英治†

† 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室
〒 630-0192 奈良県生駒市高山町 8916 番地-5
E-mail: †{onishi.shuntaro.or3,s-yada,wakamiya,aramaki}@is.naist.jp

あらまし ソーシャルセンシングとは，ソーシャルメディア上のユーザ活動を計測することで社会的変化を探知することであり，主に，X（旧，Twitter）をプラットフォームとして用いている．ソーシャルセンシングの研究は多くあるが，基本となる前提は，投稿を観測することにより，観測対象の規模（参加者数や社会的影響）を推定するというものが多い．しかし，実際に投稿数が観測対象の規模に対応しているという仮説自体の検証は，厳密にはされていない．そこで，本研究では，ある地域の投稿数から，その地域に滞在する人数（人口動態）を推定するタスクをベースに，相関検定による仮説の検証を行う．結果，仮説は常に成り立つわけではないこと．また，仮説の成立条件としては，人口動態に関しては,「観光魅力度」と「居住者数」が大きな要因となることを明らかにした．

キーワード 地理データ活用， ソーシャルメディア， ユーザモデル, 行動分析

1 はじめに

現代人の多くがスマートフォンを持ち，何らかのソーシャルメディアを利用している．この結果，人々の情報発信や情報収集が活発になった．同時に，これらの情報をもとにした，ソーシャルセンシングが注目されている．ソーシャルセンシングとは，ソーシャルメディア上のユーザ活動を計測することで社会的変化を探知することである [1]．ソーシャルセンシング自体は，特定のソーシャルメディアに依存するものではないが，通常は Twitter（現，X）を用いるものが多い．Twiiter を用いたソーシャルセンシングには，主に 3 つの利点がある．1 つ目は，ユーザ数が膨大であり，大規模な観察が可能であること．2 つ目は，Twitter の即時性を利用した迅速な情報収集が可能であること．3 つ目は，位置情報を利用することで地理的位置を考慮した分析が可能であることが挙げられる．

観測対象は多岐にわたるもののソーシャルセンシングでは一般的に，Twitter での投稿数をもとに，観測対象の規模，例えば，人数や社会的影響の大きさを推定することは共通している．しかし，この大前提である投稿数が観測対象の規模に対応しているという仮説そのものを直接検証した研究は，我々の知る限り存在しない．代表的な先行研究としては，地震や台風の検知 [2], 人々の気分の推定 [3], [4], インフルエンザの患者数を推定 [5], 交通渋滞把握 [6] が挙げられる．さらに，最近の研究では，投稿数, コンテンツ, ユーザプロファイル情報に基づいて, コミュニティの性別や人種構成を推定している [7]．どの研究も投稿数をもとに，台風, インフルエンザ, 交通渋滞など観測対象の規模を推定している．

本研究では，ソーシャルセンサ研究が前提としていた投稿数が観測対象の規模に対応するという仮説を検証する．実験 1 ではピアソンの相関検定の結果をもとに仮説の検証を行う．実験 2 では同時に得られる相関係数を「ソーシャルセンサパフォーマンス」としてその値の変動要因を分析する，これまで提案されたソーシャルセンサの種類は多くあるが，本研究では，最も素朴なソーシャルセンサとして，ある地域の投稿数から，その地域に滞在する人数（人口動態）を推定するタスクをベースに検証を行う．

2 実験とその目的

2.1 実験 1： 仮説の検証

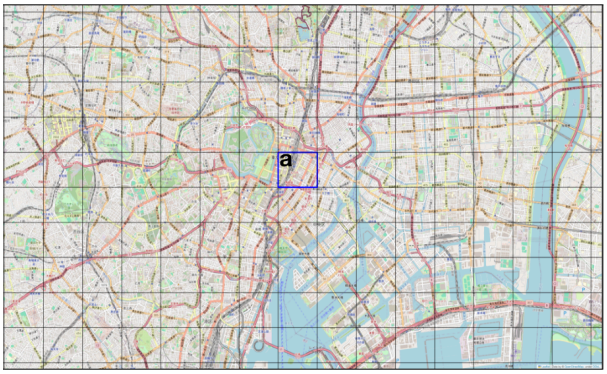
実験 1 の目的は．投稿数が観測対象の規模に対応しているという仮説の検証である．本章では，観測対象としてその地域に滞在する人数 (人口動態) を採用した．コンサートへの参加人数や被災者数などとは異なり，属性条件を持たず，同時にそれらの基礎となる変数であるため採用した．仮説は，人数と投稿数のピアソン相関検定によって検証する．帰無仮説は「人数と投稿数は無相関である」とする．帰無仮説が棄却されたすべての地域, 時間帯で相関係数が大きい正の値であるとき仮説は成立する．逆に，帰無仮説が棄却されたすべての地域で大きい正の相関がなかったり，値が条件によって大きく変動するとき仮説は成立しない．本稿では以後この相関係数を「ソーシャルセンサパフォーマンス」と呼称する．

2.2 実験 2：ソーシャルセンサパフォーマンスの値の変動要因

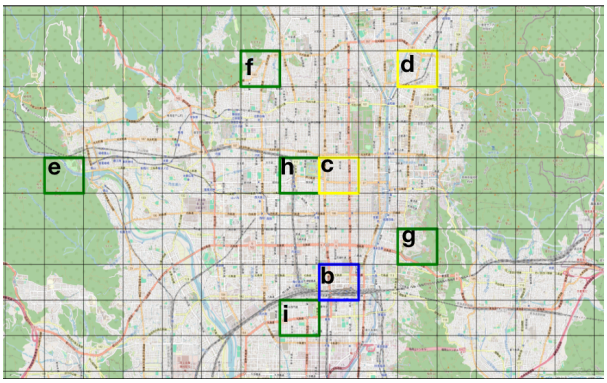
実験 2 の目的は，ソーシャルセンサパフォーマンスの値の変動要因を明らかにすることである．ソーシャルセンサパフォーマンスは理想的には 1 に近い値であることが望ましい．しかし，現実的にソーシャルセンサパフォーマンスが 1 である場面は殆どなく，値に変動があると考える．一般的には，人数が多くかつ投稿の多い都市部のほうがソーシャルセンシングが盛んに行われている．こうした状況を踏まえ，変動要因として地域ごとの人数と投稿数の多寡や，その地域が都市部であるか観光地であるかなどを候補として分析する．

表 1: 調査対象地域

名称	属性	地域メッシュID	アンケート内での名称
(a) 東京駅	駅	5339-46-11	都心部
(b) 京都駅	駅	5235-36-80	京都駅周辺
(c) 中京区	居住地域	5235-46-10	京都御所周辺
(d) 京都大学学生街	居住地域	5235-46-42	銀閣寺・哲学の道・百万遍周辺
(e) 嵐山	観光地	5235-45-13	嵯峨嵐山
(f) 金閣寺	観光地	5235-45-48	きぬかけの路
(g) 清水寺	観光地	5235-36-92	清水・祇園周辺
(h) 二条城	観光地	5235-45-19	二条城・壬生周辺
(i) 東寺	観光地	5235-35-79	京都駅周辺



(a) 東京



(b) 京都

図 1: 調査対象地域: 青枠は駅，黄枠は居住地域，緑枠は観光地を表す

3 実験 1： 仮説の検証

3.1 データ収集

3.1.1 調査対象地域

表 1 および図 1 に示すように，9 つの地域を調査した．具体的には都市の主要駅である東京駅と京都駅．さらに，居住地域として京都府中京区と京都大学学生街．くわえて，有名な観光地として嵐山, 金閣寺, 清水寺, 二条城, 東寺を選んだ．この選定は以下の予想に基づくものである．まず，駅は人数が多いものの，人々の往来が激しく投稿が少ないと考えた．したがって，ソーシャルセンサパフォーマンスの値については予想がつかない．次に，観光地は人数も投稿数も多いと考えられ，ソーシャルセンサパフォーマンスが大きい正の値になると予想した．居住地域は人数も投稿も少ないと考えられ，ソーシャルセンサパフォーマンスが弱い正の値になると予想した．

地域の分割方法として，各種統計でよく採用され，データの取得が容易である地域メッシュを採用する．地域メッシュとは，総務省統計局位が策定した緯度経度に基づく地域の分割方法である [1]．まず，地域は第 1 次地域区画と呼ばれる一辺がおよそ 80km の正方形で分割される．その後，第 2 次地域区画, 基準地域メッシュ, 2 分の 1 地域メッシュ, 4 分の 1 地域メッシュに分割される．それぞれ一辺の長さがおよそ 10km, 1km, 500m, 250m の正方形である．本章では基準地域メッシュを元に地域を選択した．分割された地域メッシュにはそれぞれ固有の ID が割り当てられ，分割が細かくなるに連れてその桁が増えていく．選択した地域と地域メッシュID の対応を表 1 に示す．

3.1.2 人数と投稿数の取得

東京駅のデータは 2021 年，それ以外の 8 つの京都府の地域は 2022 年のデータである．

人数のデータは NTT ドコモが提供するモバイル空間統計 [2] から入手した．この人数は，各地域メッシュにおける携帯端末数に基づいて 1 時間毎に推計され，1 台の端末が 1 時間該当の地域メッシュに留まった場合 1 人とカウントされ，端末が 30 分間滞在した場合，1/2 人としてカウントされる．

投稿数は Twitter のアプリケーション・プログラミング・インターフェース（API）[3]を利用し，位置情報が付与された投稿を取得した．この API では指定されたクエリに一致する過去の投稿を取得可能である．クエリは，言語情報と位置情報を含むように構築した．具体的には `lang:ja bounding_box:[w s e n]` とクエリを構築した．ここで `lang:ja` は言語を日本語に設定する部分である．また，`bounding_box:[w s e n]` は `w s e n` にそれぞれの地域メッシュの西, 南, 東, 北の境界の緯度と経度の値をそれぞれ入力することで位置情報を設定する部分である．取得したデータには，特定の時間帯に過剰に投稿するユーザの影響を軽減するため，1 時間以内に行われた連続した投稿をまとめて 1 投稿として集計した．

3.2 ソーシャルセンサパフォーマンスの定義

人数 x と投稿数 y はそれぞれ時間 ($i = 24$ 時間) 日付 ($j = 365$ 日) 場所 ($k = 9$ 地点) で抽出可能な 3 次元データであり，これを x_{ijk}, y_{ijk} と表す．

ソーシャルセンサパフォーマンスは，地域ごとに SSP_{area} と定義する．同時に，時間を考慮した SSP_{hour} を定義する．くわえて，平日と休日では人々の行動が変容する可能性が高いと考え，分析のために全日, 平日, 休日で人数と投稿数を分割した．ここで，休日は，土曜日, 日曜日, 祝日, 年末年始（12 月 29 日 ~1 月 3 日) を指す．

SSP_{area} は各地域ごとに人数と投稿数を日毎で集計し，その後，相関係数を求めたものである．任意の場所 k における地域

1：https://www.stat.go.jp/data/mesh/m_tuite.html
2：https://mobaku.jp/
3：https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

ごとのソーシャルセンサパフォーマンス SSP_{area} は以下のように定義される．

$$SSP_{area} = R_k = \frac{\mathrm{cov}_j(S_{jk}, T_{jk})}{\mathrm{std}_j(S_{jk})\ \mathrm{std}_j(T_{jk})} \quad (1)$$

(1) において，S_{jk}, T_{jk} は x_{ijk}, y_{ijk} をそれぞれ日単位で集計したものであり，一日の合計人数と合計投稿数を表す．それぞれ以下のように定義される:

$$S_{jk} = \sum_i x_{ijk}, T_{jk} = \sum_i y_{ijk}$$

(1) において，cov_j は添字 j に関する共分散であり，std_j は添字 j に関する標準偏差である．それぞれ以下のように定義される:

$$\mathrm{cov}_a(u_{ab\ldots}, v_{ab\ldots}) = \frac{1}{N_a}\sum_a (u_{ab\ldots} - \mathrm{mean}_a(u_{ab\ldots}))(v_{ab\ldots} - \mathrm{mean}_a(v_{ab\ldots})) \quad (2)$$

$$\mathrm{std}_a(u_{ab\ldots}) = \sqrt{\frac{1}{N_a}\sum_a (u_{ab\ldots} - \mathrm{mean}_a(u_{ab..}))^2} \quad (3)$$

$$\mathrm{mean}_a(u_{ab\ldots}) = \frac{1}{N_a}\sum_a u_{ab\ldots} \quad (4)$$

(2), (3), (4) において，$u_{ab\ldots}$ の次元数に関わらず，添字 a に関して計算する．このとき，N_a は全日の場合は 365 (日) である，平日と休日で分割したとき，2021 年はそれぞれ 240 (日) と 125 (日)，2022 年はそれぞれ 239 (日) と休日 126 (日) である．あわせて人数 x_{ijk} と投稿数 y_{ijk} も平日のデータのみ，休日のデータのみを用いる．

時間ごとのソーシャルセンサパフォーマンス SSP_{hour} は人数と投稿数の日毎の集計をおこなわず，相関係数を求めたものである．任意の場所 k における任意の時刻 i のソーシャルセンサパフォーマンス SSP_{hour} は以下のように定義される．

$$SSP_{hour} = R_{ik} = \frac{\mathrm{cov}_j(x_{ijk}, y_{ijk})}{\mathrm{std}_j(x_{ijk})\ \mathrm{std}_j(y_{ijk})}$$

このとき，平日と休日に関する分割は地域ごとのソーシャルセンサパフォーマンスと同じように行う．

3.3 結　　果

3.3.1 地域ごとのソーシャルセンサパフォーマンス

表 2 に示すように，全日における地域ごとのソーシャルセンサパフォーマンスは，最小-0.38 から最大 0.74 と広い範囲をとる．大きい正の値を示した地域は京都駅, 嵐山, 清水寺, 二条城と約半数にとどまり，逆に東京駅や中京区のように，負の値を示した地域もあった. 平日と休日を比べると東京駅，京都駅，清水寺，二条城ではおよそ 0.1 程度の差がある．

最も興味深い地域は東京駅である．表 2 に示すように，地域ごとのソーシャルセンサパフォーマンスは全日で-0.38 と負の値である．しかし，平日 0.62 と休日 0.79 はどちらも大きい正の値を示している．図 2 (a) に示すように，母集団での相関と，母集団を分割した集団での相関が，異なっている．図 2 (b) を見ると，東京駅と異なり，京都駅では全ての条件で正の値となっている．表 2 が示すように，地域ごとのソーシャルセンサパフォーマンスは地域によってさまざまである．そこで， 4.2 節では地域ごとのソーシャルセンサパフォーマンスの値の変動要因について分析する．

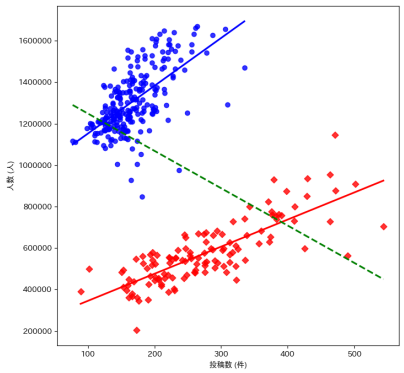
表 2: 地域ごとのソーシャルセンサパフォーマンス（太字は p 値 < 0.05）

	全日 (p 値)	平日 (p 値)	休日 (p 値)
(a) 東京駅	**-0.38 (0.00)**	**0.62 (0.00)**	**0.79 (0.00)**
(b) 京都駅	**0.69 (0.00)**	**0.81 (0.00)**	**0.72 (0.00)**
(c) 中京区	**-0.19 (0.00)**	0.06 (0.39)	**0.23 (0.01)**
(d) 京都大学	**0.14 (0.01)**	0.01 (0.84)	0.04 (0.63)
(e) 嵐山	**0.65 (0.00)**	**0.60 (0.00)**	**0.62 (0.00)**
(f) 金閣寺	**0.02 (0.00)**	-0.02 (0.75)	-0.09 (0.35)
(g) 清水寺	**0.74 (0.00)**	**0.63 (0.00)**	**0.73 (0.00)**
(h) 二条城	**0.48 (0.00)**	**0.26 (0.00)**	**0.36 (0.00)**
(i) 東寺	**0.18 (0.00)**	**0.22 (0.00)**	**0.30 (0.00)**

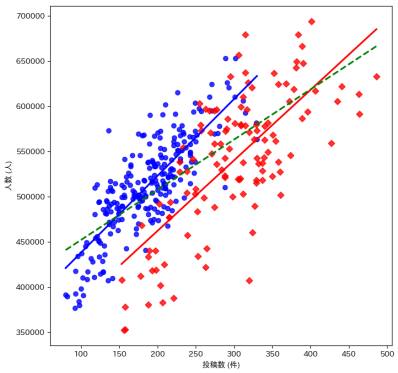
3.3.2 時間ごとのソーシャルセンサパフォーマンス

図 3 は，p 値 < 0.05 であった時間ごとのソーシャルセンサパフォーマンスを折れ線グラフで可視化したものである．0 時から 5 時の間の p 値は多くの地域で 0.05 を超えている．また，中京区と金閣寺の p 値は多くの時間帯で 0.05 を超えている．ソーシャルセンサパフォーマンスは，日中の明るい時間帯は正の値である. しかし，その増減は地域によって異なる．

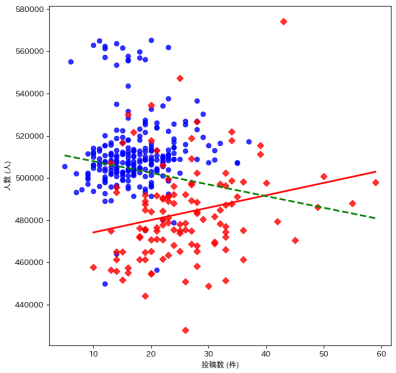
これらの結果から，本稿の結論を示す. 表 2 は場所や平日, 休日などの条件によるソーシャルセンサパフォーマンスが負の値から正の値まで大きく変動することを示している. さらに図 3 は，時間の違いによるソーシャルセンサパフォーマンスの変動を示している．すなわち，人数と投稿数の相互関係の強弱は条件によってさまざまである．したがって，投稿数が観測対象の規模に対応するという仮説は常に成り立つわけではない．



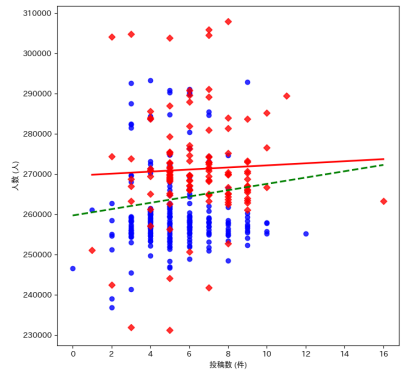
(a) 東京駅



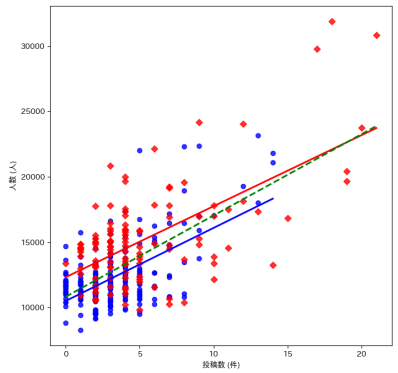
(b) 京都駅



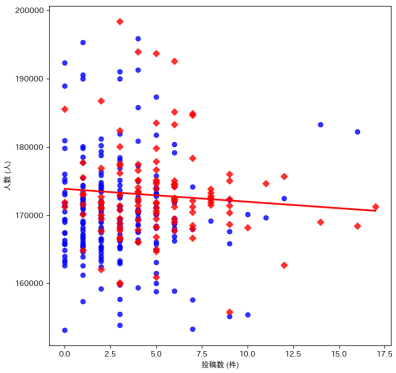
(c) 中京区



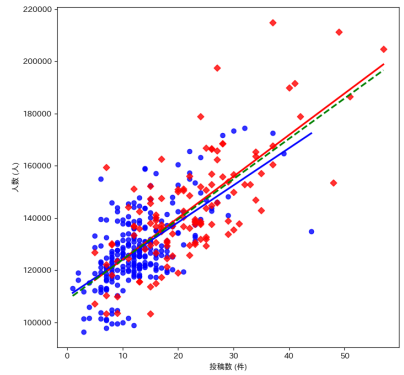
(d) 京都大学



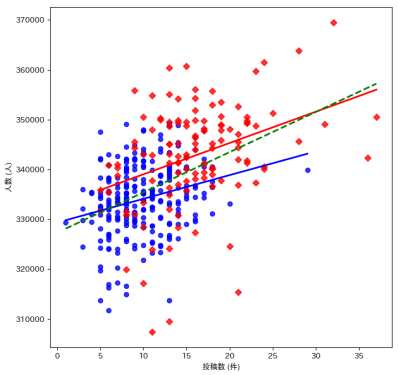
(e) 嵐山



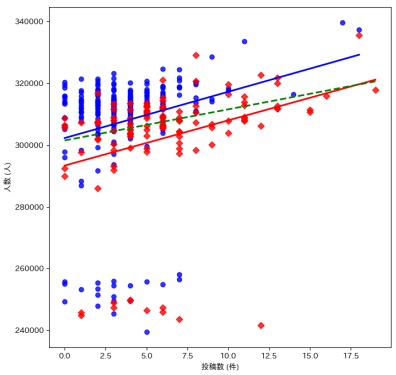
(f) 金閣寺



(g) 清水寺

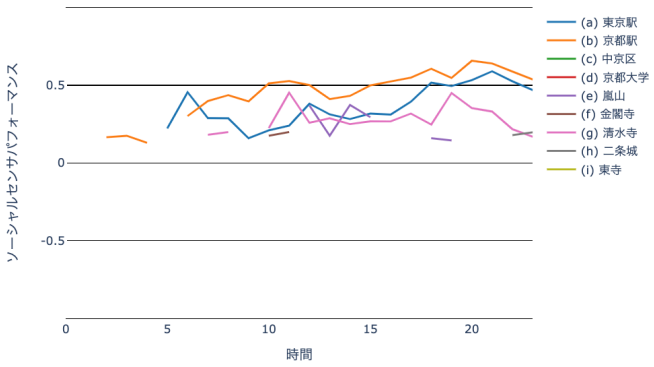


(h) 二条城

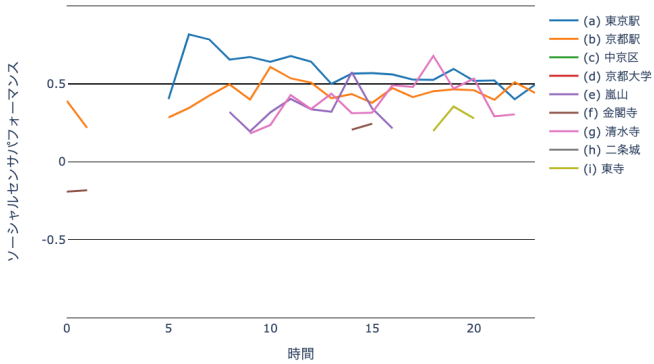


(i) 東寺

図 2: 地域ごとのソーシャルセンサパフォーマンス： 青丸は平日，赤ひし形は休日のデータ点を示す．青線は平日，橙線は休日，緑破線は全日のソーシャルセンサパフォーマンスを示す



(a) 平日



(b) 休日

図 3: 時間ごとのソーシャルセンサパフォーマンス

4 実験 2： ソーシャルセンサパフォーマンスの値の変動要因

4.1 地域の特性を示すデータの取得

実験 1 に追加して，人数と投稿数の他にソーシャルセンサパフォーマンスの値の変動要因を分析するため，地域の特性を反映していると考えられるデータを取得した．具体的には，施設の数, 最寄り駅からの距離, 居住者数, 観光魅力度などを取得した．

まず，施設の数について，地域ごとにコンビニエンスストア，飲食店，学校，ホテルの件数を Overpass turbo [4]を用いて取得した．Overpass turbo は，OpenStreetMap データのウェブベースのデータマイニングツールであり無料で利用できる．クエリは，`shop=convenience`, `amenity=restaurant`, `amenity=school`, `tourism=hotel`, `highway=bus_stop` である．データの取得は 2023 年 10 月 1 日に行われた．

最寄り駅からの距離は，geopy (version 2.3.0) を用いて算出した．地域メッシュの代表点は地域メッシュの中心座標し，地域メッシュに駅が含まれる場合は 0km とする．

居住者数は，日本政府が実施した 2020 年人口動態調査の結果から取得した [5]．この調査は 5 年ごとに実施されており，2024 年現在，2020 年版が最新である．

観光客数の適当なデータが発見できなかったため，観光魅力度を採用した．観光魅力度は東京都と京都府で実施されたアンケート調査から推計する．それぞれ「東京都データカタログ: 都民の観光に対する意識調査 [6]」と「京都府市産業観光局: 観光客の動向等に係る調査 [7]」を採用した．東京駅の観光魅力度は，東京都の魅力のあるエリアとして「東京都心部（千代田区，中央区，港区，新宿区，文京区)」をあげた人の割合とした．京都府では，回答者のうちその観光スポットに実際に訪れた人の割合に基づいている．ただし，回答エリアはそれぞれの地域メッシュと完全に一致しないことに注意が必要である．表 1 に，地域メッシュと対応するアンケート内での名称を示す．

4.2 手法と結果

ソーシャルセンサパフォーマンスの変動要因を明らかにするために，人数と投稿数の代表値 (平均, 中央値, 合計, 最大値, 最小値) と地域特性を示すデータ (施設の数, 最寄り駅からの距離, 居住者数, 観光魅力度) について地域ごとのソーシャルセンサパフォーマンスとの相関を求めた．人数と投稿数に関しては，データを日毎に集計した．その後，それぞれのデータを平日と休日で分割し，代表値を算出した．

それぞれの代表値も地域特性を示すデータも全てデータ長は地域数 (=9) である．したがって区別せずすべて a_k と表すと，相関 R' は以下のように定義される:

$$R' = \frac{\mathrm{cov}_k(SSP_{area}, a_k)}{\mathrm{std}_k(SSP_{area})\ \mathrm{std}_k(a_k)} = \frac{\mathrm{cov}_k(R_k, a_k)}{\mathrm{std}_k(R_k)\ \mathrm{std}_k(a_k)}$$

表 3 に示すように，観光魅力度と正の相関が居住者数とは負の相関が確認できた．

5 考　　察

一般的に，居住者は観光客よりも投稿が少ないと考えられる．特に，人気のある観光地では，観光客が自分の体験を共有する意欲が高いことが多い. 言い換えれば，SNS 上で居住者が非活動的な傾向があるのに対し，観光客は活動的な傾向にある. つまり，SNS 上で活動的な人が多くかつ非活動的な人が少ない条件下で，ソーシャルセンサパフォーマンスが大きくなると言える．同じ理由で，コンサートやフェスティバルのような特別なイベント時にソーシャルセンサパフォーマンスが高くなると考えられる．意外なことに，人数と投稿数の代表値との相関は見られなかった. 人口が密集し，ソーシャルメディアの利用が盛んな都市部では，ソーシャルセンサパフォーマンスが大きくな

4：Overpass turbo https://overpass-turbo.eu/

5：https://www.e-stat.go.jp/gis/statmap-search?page=1&type=1&toukeiCode\
\=00200521&toukeiYear=2020&aggregateUnit=S&serveyId=S002005112020\
\&statsId=T001100

6：https://data.tourism.metro.tokyo.lg.jp/data/ishiki/

7：https://www.kyokanko.or.jp/wp/wp-content/uploads/kyoto_tourism_stat_2022.pdf

表 3: ソーシャルセンサパフォーマンスとの相関： p 値 < 0.05 を太字で示す

	平日 (p 値)	休日 (p 値)
人数平均	0.24 (0.54)	0.25 (0.52)
人数中央値	0.23 (0.55)	0.23 (0.54)
人数合計	0.24 (0.54)	0.25 (0.53)
人数最大値	0.29 (0.45)	0.43 (0.25)
人数最小値	0.12 (0.75)	-0.16 (0.69)
投稿数平均	0.65 (0.06)	0.63 (0.07)
投稿数中央値	0.66 (0.05)	0.62 (0.07)
投稿数合計	0.65 (0.06)	0.62 (0.07)
投稿数最大値	0.66 (0.05)	0.64 (0.06)
投稿数最小値	0.64 (0.05)	0.58 (0.10)
コンビニ軒数	0.32 (0.40)	0.41 (0.28)
レストラン軒数	0.52 (0.15)	0.59 (0.09)
学校数	0.45 (0.23)	0.58 (0.10)
バス停数	0.40 (0.29)	0.35 (0.36)
ホテル軒数	0.58 (0.10)	0.46 (0.21)
観光魅力度	**0.69 (0.04)**	**0.74 (0.02)**
居住者数	**-0.83 (0.01)**	**-0.73 (0.02)**
最寄り駅からの距離	0.06 (0.87)	-0.11 (0.78)

ることが期待される．それに反して今回の結果は期待に沿うものではなかった．

今後の課題は 3 つある．まず，投稿数が観測対象の規模に対応する事例を報告した先行研究との詳細な比較が必要である．例えば，投稿数と歩行者数の相関が報告されている [8]．加えて，投稿数と観光客の流れの相関も報告されている [9]．さらに，いくつかの先行研究では，投稿テキストに単語抽出, case-folding, 正規化, 単語の除去などの前処理が施されている [5], [10], [11]．一方，実験 1 では投稿に対してほとんど前処理をせずそのまま利用している．観測対象の違いや，前処理の違いによる影響を十分に議論していく必要がある．

次に，実験 2 で用いた手法には改善の余地がある．実験 2 では,「観光魅力度」と「居住者数」がソーシャルセンサパフォーマンスの変動要因である可能性を示した．ただし，相関関係は因果関係を示すものではないという点に注意が必要である．さらに，手法自体もソーシャルセンサパフォーマンスとしてピアソン相関係数を採用し，そのソーシャルセンサパフォーマンスと相関関係にあるデータを探索した．つまり，相関係数の変動要因を相関検定により推定した．同じ方法を再帰的に利用している．このようなアプローチは，これまでの研究ではほとんど行われておらず，その有効性と解釈可能性については議論と改良の余地がある．今後の研究では，ソーシャルセンサパフォーマンスを従属変数とする線形回帰のような別の手法を用いることで，より有効な分析ができる可能性がある．

最後に，我々はソーシャルセンサのパフォーマンスという概念を導入した．ソーシャルセンサパフォーマンスは 1 であることが理想だが，効果的なソーシャルセンシングのための閾値を明らかにする必要がある．

6 おわりに

本研究では投稿数が現実世界の規模に対応しているという仮説を検証した．人数と投稿数のピアソンの相関係数に基づく，地域ごと, 時間ごとのソーシャルセンサパフォーマンスの分析によると，相関がある条件と無相関である条件が混在し，相関係数の大小も様々であった．したがって，投稿数が現実世界の規模に対応するという仮説は常に成り立つわけではないという結論を得た．同時に第 4.2 節では,「観光魅力度」と「居住者数」がソーシャルセンサパフォーマンスの変動要因である可能性を示した．

謝　辞

本研究は，JSPS 科研費 JP22H03648, JP22K12041, JST SICORP JPMJSC2107 の支援を受けたものである．

文　献

[1] Dong Wang, Boleslaw K. Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. The Age of Social Sensing. *Computer*, Vol. 52, No. 1, pp. 36–45, 2019.

[2] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 851–860, New York, NY, USA, 2010. Association for Computing Machinery.

[3] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter, 2010. https://www.ccs.neu.edu/home/amislove/twittermood/.

[4] Mark E. Larsen, Tjeerd W. Boonstra, Philip J. Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. We Feel: Mapping Emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, Vol. 19, No. 4, pp. 1246–1252, 2015.

[5] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1576, 2011. https://aclanthology.org/D11-1145.

[6] Senzhang Wang, Lifang He, Leon Stenneth, Philip S. Yu, and Zhoujun Li. Citywide Traffic Congestion Estimation with Social Media. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Association for Computing Machinery, 2015. https://doi.org/10.1145/2820783.2820829.

[7] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and Rosenquist. Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5, No. 1, pp. 554–557, Aug. 2021. https://ojs.aaai.org/index.php/ICWSM/article/view/14168.

[8] Debjit Bhowmick, Stephan Winter, and Mark Stevenson. Using Georeferenced Twitter Data to Estimate Pedestrian Traffic in an Urban Road Network. In Krzysztof Janowicz and Judith A. Verstegen, editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part I*, Vol. 177 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 1:1–1:15, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://drops.dagstuhl.de/entities/document/10.4230/\\LIPIcs.GIScience.2021.I.1.

[9] Johannes Scholz and Janja Jeznik. Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria. *ISPRS International Journal of Geo-Information*, Vol. 9, No. 11, 2020. https://www.mdpi.com/2220-9964/9/11/681.

[10] Rudy Arthur, Chris A. Boulton, Humphrey Shotton, and Hywel T. P. Williams. Social sensing of floods in the uk. *PLoS ONE*, Vol. 13, , 2017. https://api.semanticscholar.org/CorpusID:5460691.

[11] Syifa Khairunnisa, Adiwijaya Adiwijaya, and Said Al Faraby. Pengaruh text preprocessing terhadap analisis sentimen komentar masyarakat pada media sosial twitter (studi kasus pandemi covid-19). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 2021. \\https://api.semanticscholar.org/CorpusID:235251521.

appendix

表 4: 平日の時間ごとのソーシャルセンサパフォーマンス： 括弧内の値は p 値であり，p 値 < 0.05 を太字で示す

時間 (時)	0	1	2	3	4	5
(a) 東京駅	0.03 (0.59)	0.07 (0.29)	-0.07 (0.30)	-0.01 (0.85)	-0.05 (0.44)	**0.22 (0.00)**
(b) 京都駅	**0.39 (0.00)**	0.08 (0.20)	**0.17 (0.01)**	**0.18 (0.01)**	**0.13 (0.04)**	0.08 (0.19)
(c) 中京区	-0.05 (0.46)	-0.02 (0.70)	-0.00 (0.98)	-0.05 (0.43)	-0.04 (0.58)	0.02 (0.72)
(b) 京都大学	-0.01 (0.93)	No post	-0.05 (0.40)	0.03 (0.65)	-0.11 (0.09)	**0.15 (0.02)**
(e) 嵐山	0.10 (0.11)	0.04 (0.58)	-0.05 (0.43)	0.11 (0.10)	0.00 (0.94)	0.08 (0.23)
(f) 金閣寺	0.01 (0.90)	-0.00 (0.98)	-0.05 (0.47)	0.09 (0.14)	0.04 (0.56)	-0.01 (0.92)
(g) 清水寺	0.08 (0.19)	0.07 (0.31)	0.02 (0.77)	0.02 (0.81)	**-0.13 (0.04)**	0.02 (0.78)
(h) 二条城	**0.17 (0.01)**	-0.05 (0.46)	-0.02 (0.71)	0.04 (0.56)	0.05 (0.48)	-0.03 (0.68)
(i) 東寺	0.09 (0.17)	0.05 (0.41)	0.04 (0.57)	0.04 (0.50)	-0.07 (0.26)	-0.07 (0.31)
時間 (時)	6	7	8	9	10	11
(a) 東京駅	**0.46 (0.00)**	**0.29 (0.00)**	**0.29 (0.00)**	**0.16 (0.01)**	**0.21 (0.00)**	**0.24 (0.00)**
(b) 京都駅	**0.30 (0.00)**	**0.40 (0.00)**	**0.44 (0.00)**	**0.40 (0.00)**	**0.51 (0.00)**	**0.53 (0.00)**
(c) 中京区	-0.04 (0.49)	0.03 (0.64)	0.09 (0.17)	0.03 (0.63)	-0.04 (0.55)	0.05 (0.48)
(b) 京都大学	-0.04 (0.49)	0.04 (0.57)	0.06 (0.37)	-0.00 (0.96)	0.01 (0.89)	-0.01 (0.87)
(e) 嵐山	-0.02 (0.77)	**0.16 (0.01)**	0.10 (0.12)	**0.27 (0.00)**	0.01 (0.92)	0.12 (0.06)
(f) 金閣寺	0.08 (0.22)	0.09 (0.18)	0.00 (0.99)	0.09 (0.18)	**0.18 (0.01)**	**0.20 (0.00)**
(g) 清水寺	0.09 (0.16)	**0.18 (0.00)**	**0.20 (0.00)**	-0.06 (0.35)	**0.23 (0.00)**	**0.45 (0.00)**
(h) 二条城	0.04 (0.56)	0.05 (0.44)	0.04 (0.49)	**0.14 (0.03)**	0.02 (0.72)	**0.14 (0.03)**
(i) 東寺	0.08 (0.19)	0.10 (0.14)	**0.15 (0.02)**	0.06 (0.34)	0.09 (0.16)	0.04 (0.56)
時間 (時)	12	13	14	15	16	17
(a) 東京駅	**0.38 (0.00)**	**0.31 (0.00)**	**0.28 (0.00)**	**0.32 (0.00)**	**0.31 (0.00)**	**0.39 (0.00)**
(b) 京都駅	**0.50 (0.00)**	**0.41 (0.00)**	**0.43 (0.00)**	**0.50 (0.00)**	**0.53 (0.00)**	**0.55 (0.00)**
(c) 中京区	0.08 (0.24)	-0.04 (0.56)	0.00 (0.98)	0.03 (0.68)	-0.04 (0.57)	0.05 (0.41)
(b) 京都大学	-0.03 (0.69)	0.05 (0.48)	0.12 (0.06)	**-0.19 (0.00)**	0.03 (0.67)	0.01 (0.91)
(e) 嵐山	**0.37 (0.00)**	**0.18 (0.01)**	**0.37 (0.00)**	**0.30 (0.00)**	0.11 (0.08)	0.07 (0.29)
(f) 金閣寺	0.02 (0.74)	0.07 (0.30)	0.02 (0.78)	0.11 (0.09)	0.11 (0.09)	0.05 (0.47)
(g) 清水寺	**0.26 (0.00)**	**0.29 (0.00)**	**0.25 (0.00)**	**0.27 (0.00)**	**0.27 (0.00)**	**0.32 (0.00)**
(h) 二条城	0.02 (0.80)	0.07 (0.29)	0.09 (0.15)	0.04 (0.52)	0.09 (0.14)	0.11 (0.08)
(i) 東寺	0.08 (0.22)	0.21 (0.00)	-0.04 (0.56)	0.04 (0.51)	0.07 (0.31)	0.08 (0.21)
時間 (時)	18	19	20	21	22	23
(a) 東京駅	**0.52 (0.00)**	**0.50 (0.00)**	**0.53 (0.00)**	**0.59 (0.00)**	**0.53 (0.00)**	**0.47 (0.00)**
(b) 京都駅	**0.61 (0.00)**	**0.55 (0.00)**	**0.66 (0.00)**	**0.64 (0.00)**	**0.59 (0.00)**	**0.54 (0.00)**
(c) 中京区	0.02 (0.72)	0.04 (0.53)	0.07 (0.28)	**0.13 (0.04)**	0.07 (0.27)	0.06 (0.38)
(b) 京都大学	0.06 (0.34)	-0.02 (0.70)	0.04 (0.58)	-0.02 (0.77)	0.00 (0.97)	-0.00 (0.96)
(e) 嵐山	**0.16 (0.01)**	**0.15 (0.02)**	0.04 (0.54)	0.06 (0.34)	0.06 (0.32)	**0.14 (0.03)**
(f) 金閣寺	0.06 (0.36)	-0.12 (0.06)	0.10 (0.11)	**-0.14 (0.03)**	-0.09 (0.17)	**-0.14 (0.02)**
(g) 清水寺	**0.25 (0.00)**	**0.45 (0.00)**	**0.35 (0.00)**	**0.33 (0.00)**	**0.22 (0.00)**	**0.17 (0.01)**
(h) 二条城	**0.18 (0.01)**	0.08 (0.23)	0.18 (0.00)	0.02 (0.72)	**0.18 (0.01)**	**0.20 (0.00)**
(i) 東寺	0.12 (0.06)	**0.21 (0.00)**	0.10 (0.11)	0.05 (0.44)	-0.00 (0.97)	0.10 (0.11)

表 5: 休日の時間ごとのソーシャルセンサパフォーマンス： 括弧内の値は p 値であり，p 値 < 0.05 を太字で示す.

時間 (時)	0	1	2	3	4	5
(a) 東京駅	**0.21(0.02)**	-0.03(0.72)	-0.11(0.21)	**-0.19(0.04)**	-0.06(0.52)	**0.40(0.00)**
(b) 京都駅	**0.39(0.00)**	**0.22(0.02)**	0.10(0.25)	**0.26(0.00)**	0.13(0.17)	**0.29(0.00)**
(c) 中京区	0.10(0.27)	0.01(0.91)	-0.01(0.95)	**0.21(0.02)**	0.01(0.95)	0.03(0.72)
(b) 京都大学	0.03(0.71)	-0.04(0.64)	**-0.20(0.03)**	-0.04(0.67)	No post	-0.08(0.37)
(e) 嵐山	-0.05(0.59)	0.06(0.48)	0.17(0.07)	No post	-0.00(0.96)	0.09(0.34)
(f) 金閣寺	**-0.19(0.03)**	**-0.18(0.04)**	0.02(0.80)	-0.05(0.59)	-0.01(0.92)	No post
(g) 清水寺	-0.04(0.64)	0.03(0.77)	0.13(0.15)	0.01(0.93)	0.04(0.67)	-0.03(0.78)
(h) 二条城	-0.04(0.65)	0.10(0.29)	0.05(0.55)	0.01(0.90)	-0.00(0.96)	-0.00(1.00)
(i) 東寺	-0.06(0.54)	0.08(0.40)	-0.03(0.75)	0.04(0.66)	0.08(0.41)	0.11(0.21)
時間 (時)	6	7	8	9	10	11
(a) 東京駅	**0.82(0.00)**	**0.79(0.00)**	**0.66(0.00)**	**0.67(0.00)**	**0.64(0.00)**	**0.68(0.00)**
(b) 京都駅	**0.35(0.00)**	**0.43(0.00)**	**0.50(0.00)**	**0.40(0.00)**	**0.61(0.00)**	**0.54(0.00)**
(c) 中京区	-0.04(0.64)	0.12(0.18)	0.11(0.21)	0.04(0.64)	**0.32(0.00)**	0.15(0.09)
(b) 京都大学	0.01(0.88)	-0.06(0.50)	0.00(1.00)	0.09(0.30)	0.03(0.74)	-0.02(0.82)
(e) 嵐山	0.03(0.76)	0.01(0.89)	**0.32(0.00)**	**0.20(0.03)**	**0.32(0.00)**	**0.40(0.00)**
(f) 金閣寺	-0.02(0.81)	-0.11(0.21)	-0.16(0.09)	-0.02(0.85)	0.03(0.72)	**0.21(0.02)**
(g) 清水寺	0.11(0.22)	**0.28(0.00)**	0.05(0.56)	**0.18(0.04)**	**0.24(0.01)**	**0.43(0.00)**
(h) 二条城	-0.07(0.45)	-0.08(0.40)	0.04(0.70)	**0.19(0.03)**	0.09(0.34)	**0.22(0.02)**
(i) 東寺	0.18(0.05)	0.08(0.38)	0.08(0.38)	**0.23(0.01)**	0.16(0.08)	0.13(0.16)
時間 (時)	12	13	14	15	16	17
(a) 東京駅	**0.64(0.00)**	**0.50(0.00)**	**0.57(0.00)**	**0.57(0.00)**	**0.56(0.00)**	**0.53(0.00)**
(b) 京都駅	**0.51(0.00)**	**0.41(0.00)**	**0.44(0.00)**	**0.38(0.00)**	**0.47(0.00)**	**0.42(0.00)**
(c) 中京区	**0.29(0.00)**	0.18(0.05)	0.17(0.05)	0.08(0.38)	-0.02(0.85)	0.02(0.79)
(b) 京都大学	0.05(0.59)	-0.01(0.87)	0.04(0.65)	0.03(0.78)	0.03(0.71)	-0.04(0.65)
(e) 嵐山	**0.34(0.00)**	**0.32(0.00)**	**0.58(0.00)**	**0.34(0.00)**	**0.22(0.02)**	0.01(0.87)
(f) 金閣寺	0.11(0.21)	-0.01(0.94)	**0.21(0.02)**	**0.25(0.01)**	0.12(0.17)	-0.12(0.20)
(g) 清水寺	**0.34(0.00)**	**0.44(0.00)**	**0.31(0.00)**	**0.32(0.00)**	**0.49(0.00)**	**0.48(0.00)**
(h) 二条城	0.07(0.46)	0.07(0.47)	0.09(0.31)	0.10(0.27)	**0.23(0.01)**	-0.03(0.71)
(i) 当時	**0.19(0.04)**	0.06(0.50)	0.01(0.91)	-0.09(0.33)	0.07(0.47)	0.13(0.16)
時間 (時)	18	19	20	21	22	23
(a) 東京駅	**0.53(0.00)**	**0.60(0.00)**	**0.52(0.00)**	**0.52(0.00)**	**0.40(0.00)**	**0.50(0.00)**
(b) 京都駅	**0.45(0.00)**	**0.47(0.00)**	**0.46(0.00)**	**0.40(0.00)**	**0.51(0.00)**	**0.44(0.00)**
(c) 中京区	0.15(0.10)	0.13(0.14)	0.02(0.82)	**0.28(0.00)**	0.05(0.58)	0.17(0.06)
(b) 京都大学	-0.10(0.27)	0.01(0.87)	-0.03(0.72)	0.08(0.38)	-0.14(0.12)	0.01(0.92)
(e) 嵐山	**0.32(0.00)**	0.10(0.28)	0.09(0.30)	0.13(0.17)	-0.11(0.23)	0.13(0.16)
(f) 金閣寺	-0.17(0.06)	-0.01(0.90)	-0.09(0.34)	-0.05(0.57)	-0.05(0.59)	-0.07(0.46)
(g) 清水寺	**0.68(0.00)**	**0.47(0.00)**	**0.53(0.00)**	**0.29(0.00)**	**0.30(0.00)**	0.01(0.94)
(h) 二条城	0.17(0.07)	**0.21(0.02)**	0.17(0.06)	0.08(0.39)	0.14(0.11)	0.07(0.42)
(i) 東寺	**0.20(0.03)**	**0.36(0.00)**	**0.28(0.00)**	0.04(0.65)	-0.09(0.33)	**0.26(0.00)**