2024年3月1日(金)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

数値データ処理

座長:油井 誠(トレジャーデータ株式会社)
コメンテータ:井上 潮(東京電機大学)

10:05 ~ 10:30
[T1-A-7-01] HeavyOrdiNum-VL: 順序数を的確に把握し活用する視覚言語モデルの能力を探る
*増田 琉斗[1]、宮森 恒[1] (1. 京都産業大学)

10:30 ~ 10:55
[T1-A-7-02] HeavyOrdiNum-L: 順序数を的確に把握し活用する言語モデルの能力を探る
*林 寛治[1]、宮森 恒[1] (1. 京都産業大学)

10:55 ~ 11:20
[T1-A-7-03] メルカリデータセットを利用した商品価格推定の一手法
*森 優太郎[1]、上野 史[1]、太田 学[1] (1. 岡山大学)

11:20 ~ 11:45
[T1-A-7-04] Jellyfish: データ前処理のための大規模言語モデル
*張 皓辰[1]、董 于洋[2]、肖 川[1,3]、小山田 昌史[2] (1. 大阪大学、2. NEC、3. 名古屋大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

感情分析 (1)

座長:小杉 尚子(専修大学)
コメンテータ:丸橋 弘治(富士通株式会社)

10:05 ~ 10:20
[T1-B-7-01] 幼児の音声感情認識における最適な特徴量と機械学習の検討
*樽見 理花[1]、伊藤 貴之[1] (1. お茶の水女子大学理学部情報科学科伊藤研究室)

10:20 ~ 10:35
[T1-B-7-02] 絵文字を用いた短文の特徴と感情伝達効果
*宮下 大生[1]、塩野 直志[1] (1. 神奈川工科大学)

10:35 ~ 11:00
[T1-B-7-03] 歌詞感情分類データセット作成のための大規模言語モデル利用法
*宮川 尚[1]、宇津呂 武仁[1] (1. 筑波大学)

11:00 ~ 11:25
[T1-B-7-04] 共感的な働きかけを意図したエモーショナルサポート会話における発話生成
*佐藤 将太[1]、佐々木 裕多[2]、白石 優旗[3]、張 建偉[1] (1. 岩手大学大学院、2. 東京工業大学工学院、3. 筑波技術大学)

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

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

GPUデータ処理

座長:横山 大作(明治大学)
コメンテータ:小口 正人(お茶の水女子大学)

10:05 ~ 10:30
[T2-A-7-01] GPUを利用した大規模データのk-way in-placeマージ

*三浦 真矢[1]、常 穹[2]、宮崎 純[2] (1. 東京工業大学情報理工学院情報工学系宮崎研究室、2. 東京工業大学情報理工学院情報工学系)

10:30 ~ 10:55
[T2-A-7-02] GPU直接IOを用いた大規模データに対するハッシュ結合の実行方式

*三浦 優也[1]、小沢 健史[1]、合田 和生[1] (1. 東京大学 生産技術研究所)

10:55 ~ 11:20
[T2-A-7-03] 余剰のGPUメモリを活用したアクセラレータ間データ転送

*濱地 達也[1]、田村 雅寿[1] (1. 富士通株式会社)

11:20 ~ 11:45
[T2-A-7-04] FPGA-GPU間の直接データ転送による消費電力性能効率の向上

*飯澤 健[1]、西山 達也[2]、サンパト プリヤンカラ[2]、石崎 晃朗[2]、田村 雅寿[1] (1. 富士通株式会社、2. NTTソフトウェアイノベーションセンタ)

11:45 ~ 12:10
[T2-A-7-05] GPUを用いたk-匿名化加工処理の高速化の検討

*吉本 廣雅[1]、満武 巨裕[1]、合田 和生[2] (1. 医療経済研究機構、2. 東京大学 生産技術研究所)

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

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

時系列データ処理

座長:渡辺 陽介(名古屋大学)
コメンテータ:喜田 拓也(北海学園大学)

10:05 ~ 10:20
[T2-B-7-01] クエリ長に依存しない多次元時系列データに対する類似問合せ手法の提案

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

10:20 ~ 10:35
[T2-B-7-02] NexmonによるCSIベースのスマートデバイス状態推定

*原田 海斗[1]、寺本 京祐[2]、野村 祐一郎[3]、峰野 博史[3,4] (1. 静岡大学 情報学部 情報科学科、2. 静岡大学 大学院 総合科学技術研究科 情報学専攻、3. 静岡大学 学術院 情報学領域、4. 静岡大学 グリーン科学技術研究所 フィールドインフォマティクス研究コア)

10:35 ~ 11:00
[T2-B-7-03] LSiX:ストリームデータに関する複数連続的集約

*川上 隼[1]、Bou Savong[2,1]、天笠 俊之[2,1] (1. 筑波大学、2. 計算科学研究センター)

11:00 ~ 11:25
[T2-B-7-04] 大規模データストリームに対する高速なS-FINCHクラスタリング

*牛尼 索造[1]、藤原 靖宏[2]、塩川 浩昭[3] (1. 筑波大学 情報学群情報科学類、2. NTTコミュニケーション科学基礎研究所、3. 筑波大学 計算科学研究センター)

11:25 ~ 11:50
[T2-B-7-05] 【技術報告】東芝の時系列データ処理基盤技術(GridDB)ならびにデータ仮想化エンジン(PGSpider)

浪岡保男、服部雅一、相川 恵 (株式会社東芝)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

情報拡散・偽情報・ヘイトスピーチ

座長:山本 祐輔(静岡大学)
コメンテータ:石川 佳治(名古屋大学)

10:05 ~ 10:30

[T3-A-7-01] 自然言語による説明を生成するfaithfulな偽情報検出モデルの試作

*吉田 倖[1]、張 建偉[1] (1. 岩手大学)

10:30 ~ 10:55

[T3-A-7-02] 誹謗中傷検出タスクの学習を補助するためのサブタスクの提案

*沢田 凌一[1]、鈴木 優[1] (1. 岐阜大学)

10:55 ~ 11:20

[T3-A-7-03] ソーシャルメディア投稿における炎上推定のための世論に対する乖離度算出手法

*吉田 裕輔[1]、鷹野 孝典[2] (1. 神奈川工科大学大学院 工学研究科 情報工学専攻 鷹野研究室、2. 神奈川工科大学 情報学部 情報工学科)

11:20 ~ 11:45

[T3-A-7-04] ニュース記事の拡散速度に着目したソーシャルメディア分析

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

11:45 ~ 12:10

[T3-A-7-05] Trustable Internetにおけるインターネット上のデータの信憑性判定および根拠説明文生成手法の提案

*佐久間 義友[1]、大木 憲二[1]、佐々木 佑樹[1]、坂本 拓也[1]、吉武 敏幸[1]、二村 和明[1] (1. 富士通株式会社)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

ドメイン指向推薦 (1)

座長:黒川 茂莉(株式会社KDDI総合研究所)
コメンテータ:杉山 一成(大阪成蹊大学)

10:05 ~ 10:30

[T3-B-7-01] 能力と成績の乖離における将来性に着目した推薦システムの開発

*神 一輝[1]、張 建偉[1] (1. 岩手大学)

10:30 ~ 10:55

[T3-B-7-02] 楽天レビューデータを用いたレビュー推薦の一手法

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

10:55 ~ 11:20

[T3-B-7-03] スポットの雰囲気分析によるユーザの状況と気分の選択に合わせた楽曲推薦手法の検証

*丸岡 史奈[1]、栗 達[2]、熊本 忠彦[3]、小野 晋太郎[2]、河合 由起子[1] (1. 京都産業大学、2. 福岡大学、3. 千葉工業大学)

11:20 ~ 11:45

[T3-B-7-04] 旅行者の時空間的・感性的条件に基づく旅程と経路周辺店舗の組合せ生成方式

*神原 隼登[1]、林 康弘[1]、清木 康[1] (1. 武蔵野大学データサイエンス学科)

11:45 ~ 12:10

[T3-B-7-05] 【技術報告】サイバーエージェントメディア事業部における推薦システム

松月大輔 (株式会社サイバーエージェント)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

ファッション・コスメ

座長:有次 正義(熊本大学)
コメンテータ:神門 典子(国立情報学研究所)

10:05 ~ 10:20

[T3-C-7-01] レビュー分析に基づくヘアケアアイテムの香り可視化手法の提案

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

10:20 ~ 10:35

[T3-C-7-02] 投稿コーディネート画像を用いたLoRAモデルの作成と分析

*大江 優真[1]、荘司 慶行[1] (1. 静岡大学)

10:35 ~ 11:00

[T3-C-7-03] コスメレビューデータを用いたメイクパターン特性の可視化

*福本 真以子[1]、牛尼 剛聡[2]、角谷 和俊[1] (1. 関西学院大学、2. 九州大学)

11:00 ~ 11:25

[T3-C-7-04] Personalized Fashion Recommendation with Image Attributes and Aesthetics Assessment

*Chen Chongxian[1]、Fan Mo[1]、Fan Xin[1]、山名 早人[1] (1. 早稲田大学　山名研究室)

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

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

芸術の理解

座長:田中 克己(福知山公立大学)
コメンテータ:王 元元(山口大学)

10:05 ~ 10:30

[T4-A-7-01] スケッチによるマスク不要の表情差分生成手法の提案

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

10:30 ~ 10:55

[T4-A-7-02] アート観賞イベントと連動したインタラクティブな街歩き型ストーリーリーダーの一検討

*徳丸 晴天[1]、畑 玲音[1]、松下 光範[1]、徳山 美津恵[1] (1. 関西大学)

10:55 ~ 11:20

[T4-A-7-03] CNNによるフェルメールの画風分析

*小田 稜子[1]、伊藤 貴之[1] (1. お茶の水女子大学理学部情報科学科伊藤研究室)

11:20 ~ 11:45

[T4-A-7-04] 深層距離学習による文字種とスタイルの特徴獲得とその特徴に基づく文字生成

*桑田 若菜[1]、三林 亮太[1]、谷 雅德[2]、大島 裕明[1] (1. 兵庫県立大学、2. 北陸先端科学技術大学院大学)

11:45 ~ 12:10

[T4-A-7-05] TREXIV: 時刻スタンプ付き高次元データのための対話的可視化に基づくトレンド抽出手法

*中村 栄太[1]、伊藤 貴之[2] (1. 京都大学、2. お茶の水女子大学)

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

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

ウェルビーイング

座長:児玉 直樹(新潟医療福祉大学)
コメンテータ:太田 学(岡山大学)

10:05 ~ 10:20

[T5-A-7-01] 心理的支援を目的としたChatBotによるソーシャル・サポートシステムの提案

*池田 青空[1]、談議所 考汰[3]、Siriaraya Panote[2]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学、3. 京都産業大学大学院)

10:20 ~ 10:45

[T5-A-7-02] メンタルウェルビーイング向上のためのピアベースサポートシステム(Peer2S)の開発

*談議所 孝汰[1]、Siriaraya Panote[2]、She Wan jou[4]、中島 伸介[3] (1. 京都産業大学大学院、2. 京都工芸繊維大学、3. 京都産業大学、4. 奈良先端科学技術大学院大学)

10:45 ~ 11:10

[T5-A-7-03] 社員が記述したテキストからの幸福度推定

*林 純子[1]、伊藤 和浩[1]、眞鍋 雅恵[1]、渡邉 寧[2]、中山 真孝[2]、内田 由紀子[2]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. 京都大学人と社会の未来研究院)

11:10 ~ 11:35

[T5-A-7-04] COVID-19 流行期における大学生の X 上の投稿と主観的幸福感の関係：2 年間の調査と言語特徴分析による検討

*蔡 丹萌[1]、若林 啓[1]、叶 少瑜[1] (1. 筑波大学)

11:35 ~ 12:00

[T5-A-7-05] 孤独エピソードデータセットの構築とその応用

*藤川 直也[1]、NGUYEN Quang Toan[2]、伊藤 和浩[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学　ソーシャル・コンピューティング研究室、2. ESIEE Paris)

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

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

時空間予測

座長:横山 昌平(東京都立大学)
コメンテータ:金 京淑(産業技術総合研究所)

10:05 ~ 10:20

[T5-B-7-01] マイクロブログを用いたイベント発生地周辺エリアの人口変動予測手法

*豊田 正史[1] (1. 東京大学生産技術研究所)

10:20 ~ 10:45

[T5-B-7-02] EVライフログを用いたEVのエネルギー変換効率推定

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

10:45 ~ 11:10

[T5-B-7-03] 犯罪および商業施設の情報に基づくゴミ予測手法の検討

*羽倉 輝[1]、山口 琉太[1]、義久 智樹[2]、下條 真司[3]、河合 由起子[1] (1. 京都産業大学、2. 滋賀大学、3. 青森大学)

11:10 ~ 11:35

[T5-B-7-04] EVの長期間ライフログに基づくバッテリー損耗のモデル化とVGIシミュレーションによる考察

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

11:35 ~ 12:00

[T5-B-7-05] 人流シミュレーションにおける複数指標を考慮したパラメータ影響の可視化

*森 千紗[1]、武田 芽依[2]、大西 正輝[3]、伊藤 貴之[1] (1. お茶の水女子大学、2. 筑波大学、3. 産業技術総合研究所)

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

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

エンタメ・メディア (1)

座長:橋本 敦史(オムロンサイニックエックス)
コメンテータ:廣田 雅春(岡山理科大学)

10:05 ~ 10:20
[T5-C-7-01] ショートビデオを用いた画像検索における観光地紹介ビデオの生成方式

*藤原 虎留[1]、王 元元[1]、岳 五一[2] (1. 山口大学、2. 京都情報大学院大学)

10:20 ~ 10:45
[T5-C-7-02] マルチプレイヤ RPG において人間らしいプレイスタイルをとるゲームAIの提案

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

10:45 ~ 11:10
[T5-C-7-03] 中間音楽シーケンス生成のためのTransformerモデル

*劉 明明[1]、酒井 哲也[1] (1. 早稲田大学)

11:10 ~ 11:35
[T5-C-7-04] YouTube動画の視聴状況の予測

*花畑 槙哉[1]、許 レイ[2]、馬 強[3] (1. 京都工芸繊維大学　情報工学課程　社会知能情報学研究室、2. 京都情報大学院大学/京都工芸繊維大学、3. 京都工芸繊維大学)

11:35 ~ 12:00
[T5-C-7-05] 言語モデルによる文の流暢さを用いた大喜利生成

*高橋 巧実[1]、鈴木 優[1] (1. 岐阜大学)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

ソーシャルデータ分析

座長:大島 裕明(兵庫県立大学)
コメンテータ:天方 大地(大阪大学)

13:35 ~ 14:00
[T1-A-8-01] 機械学習と地域度･多様性分析を用いたニュース記事の自動分類

藤崎 日和[1]、*田中 克己[1] (1. 福知山公立大学情報学部)

14:00 ~ 14:25
[T1-A-8-02] 観光地レビュー分析に基づくお出かけスポット評価項目別自動スコアリング

*福嶋 亮太朗[1]、張 建偉[2]、奥 健太[3]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学、3. 龍谷大学)

14:25 ~ 14:50
[T1-A-8-03] ソーシャルコンテクストを利用したアンサンブル学習によるフェイクニュースの早期検出

*谷 聡馬[1]、張 建偉[1]、佐々木 裕多[2] (1. 岩手大学大学院、2. 東京工業大学 工学院)

14:50 ~ 15:15
[T1-A-8-04] GPTを活用した、ユーザの発言・行動履歴を参考に抽出したイベント推薦の評価

*大本 詩織[1]、小口 正人[1]、榎 美紀[2] (1. お茶の水女子大学理学部情報科学科小口研究室、2. 日本アイ・ビー・エム)

15:15 ~ 15:40
[T1-A-8-05] 効果的な行動変容メッセージ実現に向けたインセンティブ効果の被験者評価

*瀧本 祥章[1]、金 秀明[1]、小林 のぞみ[1]、倉島 健[1] (1. NTT人間情報研究所)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

感情分析 (2)

座長:富井 尚志(横浜国立大学)
コメンテータ:渡辺 陽介(名古屋大学)

13:35 ~ 14:00

[T1-B-8-01] インターネットスラングを考慮した大規模言語モデルを用いた感情分析手法の提案

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

14:00 ~ 14:25

[T1-B-8-02] Evaluating the Performance of ChatGPT for Aspect-Based Sentiment Analysis

*Wang Yifei[1]、Iwaihara Mizuho[1] (1. 早稲田大学)

14:25 ~ 14:50

[T1-B-8-03] 大規模言語モデルによる細かな感情推定手法

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

14:50 ~ 15:15

[T1-B-8-04] 大規模言語モデルによる Multi-Step Prompt を用いた感情認識

*濱 健太[1]、大塚 淳史[1]、石井 亮[1] (1. 日本電信電話株式会社　人間情報研究所)

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

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

データリネージュ・統合データ管理

座長:北島 信哉(富士通株式会社)
コメンテータ:中野 美由紀(津田塾大学)

13:35 ~ 13:50

[T2-A-8-01] 来歴情報を活用したデータベースからの因果推論

*大岩 和樹[1]、石川 佳治[1]、杉浦 健人[1]、陸 可鏡[1] (1. 名古屋大学大学院情報学研究科知能システム学専攻石川研究室)

13:50 ~ 14:15

[T2-A-8-02] データの流通管理・予測のためのリネージュデータ管理システム

*安田 香子[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学情報理工学院情報工学系)

14:15 ~ 14:40

[T2-A-8-03] ユーザ定義の述語とエンティティリンキングを利用した知識ベースと外部情報源の統合利用

*大森 雄基[1,4]、北川 博之[2,4]、天笠 俊之[3]、的野 晃整[4] (1. 筑波大学 システム情報工学研究群、2. 筑波大学 国際統合睡眠医科学研究機構、3. 筑波大学 計算科学研究センター、4. 国立研究開発法人産業技術総合研究所 人工知能研究センター)

14:40 ~ 15:05

[T2-A-8-04] A Polystore System for Graph and Key-Value Databases based on SQL/PGQ

*WONG WEI JUN[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学　情報理工学院　情報工学系)

15:05 ~ 15:30

[T2-A-8-05] 【技術報告】Universal HTAP: 多様なデータベースを統合的に管理するデータアーキテクチャ

山田浩之 (株式会社Scalar)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

大規模言語モデル (1)

座長:若林 啓(筑波大学)
コメンテータ:田島 敬史(京都大学)

13:35 ~ 13:50
[T3-A-8-01] 大規模言語モデルを用いた検索要求の観点抽出に基づく検索結果の欠損の可視化

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

13:50 ~ 14:15
[T3-A-8-02] 大規模言語モデルを用いたカテゴリ説明文付与によるニュース推薦の性能向上

*矢田 宙生[1]、山名 早人[1] (1. 早稲田大学)

14:15 ~ 14:40
[T3-A-8-03] LLMベース推薦システムにおけるユーザのネガティブ嗜好の活用

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

14:40 ~ 15:05
[T3-A-8-04] 大規模言語モデルに基づく検索手法の日本語文書検索への適用

*阿部 健也[1]、薄羽 皐太[1]、加藤 誠[2] (1. 筑波大学 大学院人間総合科学学術院 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系)

15:05 ~ 15:30
[T3-A-8-05] 文書への追加情報に対するニューラル検索モデルの挙動分析

*薄羽 皐太[1]、加藤 誠[1]、藤田 澄男[2] (1. 筑波大学、2. LINEヤフー株式会社)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

ドメイン指向推薦 (2)

座長:國島 丈生(追手門学院大学)
コメンテータ:梅本 和俊(東京大学)

13:35 ~ 14:00
[T3-B-8-01] 周辺環境分析に基づく賃貸物件のPersonalized Scoring方式の開発

*須山 瑠万[1]、栗 達[2]、河合 由起子[3]、中島 伸介[3] (1. 京都産業大学大学院、2. 福岡大学、3. 京都産業大学)

14:00 ~ 14:25
[T3-B-8-02] 現代美術業界のモデル化と作家への美術機関推薦システム

*松野 仁志[1]、メンドンサ イスラエル[2]、有次 正義[2] (1. 熊本大学 工学部 情報電気工学科 有次研究室、2. 熊本大学 大学院 先端科学研究部 有次研究室)

14:25 ~ 14:50
[T3-B-8-03] 繰り返し聴取による興味の変化を考慮した楽曲推薦

*重富 竜ノ介[1]、西田 紘子[2]、澤井 賢一[2]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府牛尼研究室、2. 九州大学大学院芸術工学研究院)

14:50 ~ 15:15
[T3-B-8-04] 旅行者の想起プロセスを反映した観光場面と条件に基づく旅程生成方式

*箕輪 真衣[1]、伴 匠生[1]、林 康弘[1]、清木 康[1] (1. 武蔵野大学データサイエンス学科)

15:15 ~ 15:40
[T3-B-8-05] 【技術報告】DMM.comのレコメンドに関する取り組み

森 雅也、土屋 裕太郎 (合同会社DMM.com)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

情報探索支援

座長:倉島 健(NTT人間情報研究所)
コメンテータ:鬼塚 真(大阪大学)

13:35 ~ 14:00

[T3-C-8-01] 代替行動推薦のためのアチーブメント関係にある行動の発見

*平野 司[1]、荘司 慶行[2]、山本 岳洋[3]、大原 剛三[1] (1. 青山学院大学、2. 静岡大学、3. 兵庫県立大学)

14:00 ~ 14:25

[T3-C-8-02] 芸能人への感想を表すマイクロブログ・ポストに対する大規模言語モデルを用いた理由集約

*横山 響[1]、宇津呂 武仁[1] (1. 筑波大学)

14:25 ~ 14:50

[T3-C-8-03] 論文調査の支援に向けた研究
（最新の読むべき論文の判別方法）

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

14:50 ~ 15:15

[T3-C-8-04] 手続き的知識源学習によるタスク検索における方法と目的の生成

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

15:15 ~ 15:40

[T3-C-8-05] 【技術報告】会社訪問アプリ「Wantedly Visit」における新規ユーザーの行動量に基づいた推薦方策の選択

林悠大 (ウォンテッドリー株式会社)

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

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

認証

座長:金 京淑(産業技術総合研究所)
コメンテータ:横山 大作(明治大学)

13:35 ~ 14:00

[T4-A-8-01] 全天球カメラを用いた人物認証と人物追跡

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

14:00 ~ 14:25

[T4-A-8-02] 2台の全天球カメラを用いた同一人物判定

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

14:25 ~ 14:50

[T4-A-8-03] 複数カメラ映像に跨ったエピソード記憶

*桑田 若菜[1]、山崎 智史[2]、劉 健全[2] (1. 兵庫県立大学、2. 日本電気株式会社)

14:50 ~ 15:15

[T4-A-8-04] 不均衡データを活用した認証手法の提案−タッチベース認証を対象として−

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

15:15 ~ 15:40

[T4-A-8-05] eCommTouch：ECサイトを対象としたタッチベース認証のための新たなデータセットの提案

*工藤 雅士[1]、高橋 翼[2]、山名 早人[1] (1. 早稲田大学、2. LINEヤフー株式会社)

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

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

ライフスタイル

座長:王 元元(山口大学)
コメンテータ:吉川 正俊(京都大学)

13:35 ~ 14:00

[T5-A-8-01] 肌質や肌悩みを考慮した類似ユーザ判定に基づくコスメアイテム推薦手法

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

14:00 ~ 14:25

[T5-A-8-02] 大規模言語モデルを用いた他者の観点に基づくファッションアドバイザ

*吉倉 健人[1]、関 峰[1]、北山 大輔[1] (1. 工学院大学)

14:25 ~ 14:50

[T5-A-8-03] 中長期的影響を考慮した株式の取引支援

*福田 京司[1]、馬 強[1] (1. 京都工芸繊維大学)

14:50 ~ 15:15

[T5-A-8-04] 食品ロス削減を目的とした小売店舗の割引情報・料理レシピ提示システムの実現

*永浦 実樹[1]、阿部 広夢[1]、小林 千沙希[1]、石橋 直樹[1]、佐々木 史織[1] (1. 武蔵野大学データサイエンス学部データサイエンス学科)

15:15 ~ 15:40

[T5-A-8-05] 【技術報告】都市・ビル向けサービスの実装を容易化する空間デジタルツイン基盤

水品圭汰 (株式会社日立製作所)

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

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

環境科学

座長:加藤 誠(筑波大学)
コメンテータ:伊藤 貴之(お茶の水女子大学)

13:35 ~ 13:50

[T5-B-8-01] 灯油消費に関する多変量時系列予測方式

*横須賀 祐太[1]、神長 宏侑[1]、齋藤 瑞紀[1]、林 康弘[2]、清木 康[2] (1. 武蔵野大学アジアAI研究所　客員研究員、2. 武蔵野大学　データサイエンス学部)

13:50 ~ 14:15

[T5-B-8-02] メタ学習を用いた物理条件の異なるハミルトニアンの学習

*坂本 州[1]、岩田 具治[2]、田中 佑典[2] (1. 奈良先端科学技術大学院大学、2. NTTコミュニケーション科学基礎研究所)

14:15 ~ 14:40

[T5-B-8-03] ブラックボックス関数が与えられた予算制約付組合せ最適設備投資問題に対するモンテカルロ木探索法の提案

*行木 渉真[1]、中村 篤祥[2]、安ヶ平 裕介[3]、内垣内 洋[3]、平松 道晶[3]、竹本 享史[3] (1. 北海道大学 情報エレクトロニクス学科 情報理工学コース アルゴリズム研究室、2. 北海道大学 情報科学研究院 情報理工学部門 知識ソフトウェア科学分野、3. （株）日立製作所)

14:40 ~ 15:05

[T5-B-8-04] 微生物学的水質予測のための符号制約付き学習アルゴリズム

*高田 裕也[1]、小野 晴生[1]、門屋 俊祐[2]、佐野 大輔[3]、加藤 毅[4] (1. 群馬大学大学院理工学府 、2. 東京大学大学院工学系研究科、3. 東北大学大学院工学研究科、4. 群馬大学情報学部)

15:05 ~ 15:30

[T5-B-8-05] SABM：大規模言語モデルに基づくエージェントベース実世界シミュレーション

*呉 増青[1,2]、彭 潤[3]、韓 勗[4]、鄭 舒元[1]、肖 川[1,5] (1. 大阪大学、2. 京都大学、3. ミシガン大学、4. フォーダム大学、5. 名古屋大学)

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

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

エンタメ・メディア (2)

座長:上田 真由美(流通科学大学)
コメンテータ:佃 洸摂(産業技術総合研究所)

13:35 ~ 13:50

[T5-C-8-01] 生成AIによる小説挿絵の自動生成のためのアセスメント機構

*甲斐田 哲之佐[1,3]、牛尼 剛聡[2,3] (1. 長崎県立 口加高等学校、2. 九州大学 大学院芸術工学研究院、3. 九州大学 未来創生科学者育成プロジェクト)

13:50 ~ 14:05

[T5-C-8-02] ニュース記事とサッカー選手特徴を用いた地域的な嗜好の分析と可視化

*松葉 龍基[1]、王 元元[1] (1. 山口大学)

14:30 ~ 14:55

[T5-C-8-04] ライブ配信サイト Twitch における人気配信者の要因の分析

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

14:55 ~ 15:20

[T5-C-8-05] 実際のタワーディフェンスゲームのステージを模したシミュレータでのステージ攻略AI開発

*野村 大輔[1]、秋岡 明香[1] (1. 明治大学総合数理学部ネットワークデザイン学科　秋岡研究室)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

トピックモデリング

座長:伏見 卓恭(東京工科大学)
コメンテータ:白井 匡人(島根大学)

16:05 ~ 16:30

[T1-A-9-01] DQNを用いたトピックモデルと予測モデルの相互学習

*中村 礼音[1]、伊藤 寛祥[1]、森嶋 厚行[1] (1. 筑波大学)

16:30 ~ 16:55

[T1-A-9-02] Human-in-the-Loop ニューラルトピックモデリング

*米谷 亜夕[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学 情報学群、2. 筑波大学 図書館情報メディア系)

16:55 ~ 17:20

[T1-A-9-03] Pragmatic Approach to Automatic Topic Labeling with Zero-shot Classification

*周 子楷[1]、若林 啓[1] (1. 筑波大学)

17:20 ~ 17:45

[T1-A-9-04] BERTopic文書分類器の分類安定性に関する評価手法の提案

*櫻井 勇気[1]、小林 亜樹[2] (1. 工学院大学 工学研究科修士課程 電気・電子工学専攻 メディア情報システム研究室、2. 工学院大学 工学研究科 電気・電子工学専攻 メディア情報システム研究室)

17:45 ~ 18:10

[T1-A-9-05] 【技術報告】文書のバージョン管理の提案

楢木 悠士 (株式会社LegalOn Technologies)

一般発表 | Track 1: 自然言語処理・機械学習基礎

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

画像テキスト情報処理

座長:金子 邦彦(福山大学)
コメンテータ:倉島 健(NTT人間情報研究所)

16:05 ~ 16:20

[T1-B-9-01] 汎用的検索に基づく視覚的なマルチホップ質問応答

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

16:20 ~ 16:45

[T1-B-9-02] 対話型生成AIによる情報検索の品質分析

*遠藤 颯[1]、花谷 優[1]、田中 克己[1] (1. 福知山公立大学情報学部)

16:45 ~ 17:10

[T1-B-9-03] 大規模言語モデルを用いた語用論的アプローチに基づく誤解可能性を考慮した画像情報ラベリング

*清野 駿[1]、池垣 隼介[1]、周 子楷[1]、若林 啓[1]、竹内 誉羽[2] (1. 筑波大学、2. 株式会社ホンダ・リサーチ・インスティチュート・ジャパン)

17:10 ~ 17:35

[T1-B-9-04] 自然保護管理のための捕獲を目的とした鹿の行動パターンに基づく警戒度推定

*酒井 歩夢[1]、宮森 恒[1] (1. 京都産業大学宮森研究室)

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

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

大規模データ処理応用

座長:鈴木 伸崇(筑波大学)
コメンテータ:塩川 浩昭(筑波大学)

16:05 ~ 16:30

[T2-A-9-01] 犯罪種類間の相関を考慮したニューラルネットワークによる犯罪予測

*杉浦 光彩[1]、白井 匡人[1] (1. 島根大学)

16:30 ~ 16:55

[T2-A-9-02] 1 細胞単位の遺伝子発現の時系列データに対する遺伝子の疾患の進行による関係性の変化の抽出

*増田 智也[1]、浅野 泰仁[1]、小川 達郎[2]、七野 成之[2]、上羽 悟史[2]、松島 綱治[3]、小倉 淳[4] (1. 東洋大学情報連携学部情報連携学科浅野泰仁研究室、2. 東京理科大学大学院生命科学研究科、3. 東京理科大学生命医科学研究所炎症・免疫難病制御部門、4. 長浜バイオ大学アニマルバイオサイエンス学科)

16:55 ~ 17:20

[T2-A-9-03] Telegramにおけるグループ名と投稿メッセージの分析によるグループ探索

*伊藤 純菜[1]、趙 智賢[2]、長田 繁幸[2]、中川 直樹[2]、小口 正人[1] (1. お茶の水女子大学、2. 株式会社日本総合研究所)

17:20 ~ 17:45

[T2-A-9-04] 子宮収縮圧データのための潜在ステージ進行推定

*田村 優香[1,2]、松原 靖子[2]、木村 輔[2]、町村 栄聡[3]、遠藤 誠之[3]、木村 正[3]、櫻井 保志[2] (1. 大阪大学情報科学研究科、2. 大阪大学産業科学研究所、3. 大阪大学大学院医学系研究科)

17:45 ~ 18:10

[T2-A-9-05] SNSにおけるサイバー犯罪対策: 交グラフを活用した犯罪グループ探索アルゴリズム

*大原 望乃[1]、伊藤 純菜[2]、趙 智賢[3]、長田 繁幸[3]、中川 直樹[3]、小口 正人[1] (1. お茶の水女子大学大学院、2. お茶の水女子大学、3. 株式会社 日本総合研究所)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

大規模言語モデル (2)

座長:梅本 和俊(東京大学)
コメンテータ:服部 峻(滋賀県立大学)

16:05 ~ 16:30
[T3-A-9-01] Cross-lingual Relevance Estimation with Soft and Hard Prompts
*Fang Hanpei[1]、酒井 哲也[2] (1. 早稲田大学基幹理工研究科酒井研究室、2. 早稲田大学)

16:30 ~ 16:55
[T3-A-9-02] A Study on Automatic Nugget Weight Generation with LLMs
*Chang Kai-Xin[1]、酒井 哲也[2] (1. 早稲田大学基幹理工研究科酒井研究室、2. 早稲田大学)

16:55 ~ 17:20
[T3-A-9-03] 大規模言語モデルを用いたユーザの目的に基づく説明性を有するPOI推薦手法
*片山 一[1]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府　牛尼研究室、2. 九州大学大学院芸術工学研究院)

17:20 ~ 17:45
[T3-A-9-04] 特定ドメインの問合せを対象とした大規模言語モデルの性能計測手法の検討
*神田 智也[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学融合知能デザイン研究室、2. 筑波大学図書館情報メディア系)

17:45 ~ 18:10
[T3-A-9-05] 【技術報告】リクルートの横断検索基盤における取り組み紹介
新田洸平 (株式会社リクルート)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

推薦モデル・評価

座長:佐々木 勇和(大阪大学)
コメンテータ:江口 浩二(広島大学)

16:05 ~ 16:30
[T3-B-9-01] 複数回クラスタリングによる協調フィルタリングを用いた推薦システム
*FU LEI[1]、張 建偉[1] (1. 岩手大学)

16:55 ~ 17:20
[T3-B-9-03] Light distillation for Incremental Graph Convolution Collaborative Filtering
*Fan Xin[1]、Fan Mo[1]、Chen Chongxian[1]、山名 早人[1] (1. 早稲田大学　山名研究室)

17:20 ~ 17:45
[T3-B-9-04] グラフニューラルネットワークによる分離表現学習を用いた推薦の多様化
*杉山 拓都[1]、吉田 壮[1]、棟安 実治[1] (1. 関西大学大学院 理工学研究科)

17:45 ~ 18:10
[T3-B-9-05] Polya–Gamma分布を使ったBayesian Factorization Machines
*北原 洋一[1] (1. 株式会社リブセンス)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

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

グラフ処理

座長:早水 悠登(東京大学)
コメンテータ:加藤 誠(筑波大学)

16:05 ~ 16:30
[T3-C-9-01] 異種グラフ表現学習を用いた悪意のあるSNS投稿の検出

*吉田 燎[1]、吉田 壮[1]、棟安 実治[1] (1. 関西大学大学院 理工学研究科)

16:30 ~ 16:55
[T3-C-9-02] 隣接ノードの情報を利用するランダムウォークベースのグラフ特徴量推定

*長谷川 毅[1]、廣中 詩織[1]、首藤 一幸[1] (1. 京都大学)

16:55 ~ 17:20
[T3-C-9-03] 局所的な密度の指標として機能するハイパーグラフ向けクラスタ係数

*宮下 陸矢[1]、廣中 詩織[2]、首藤 一幸[2] (1. 東京工業大学、2. 京都大学)

17:45 ~ 18:10
[T3-C-9-05] GPUを用いた大規模グラフ処理アルゴリズムの高速化

*岡崎 大知[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学)

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

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

地理情報

座長:角谷 和俊(関西学院大学)
コメンテータ:横山 昌平(東京都立大学)

16:05 ~ 16:20
[T4-A-9-01] 降水量と河川水位の関係の可視化
～動的時間伸縮法を用いた観測点間類似度の導入

*田中 悠来[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:20 ~ 16:45
[T4-A-9-02] 47都道府県単位のデータ比較可視化によるジェンダーギャップの要因解析

*石戸谷 由梨[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:45 ~ 17:10
[T4-A-9-03] SNS上の投稿写真の位置情報と意味情報を考慮したピクトリアルマップの自動生成

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

17:10 ~ 17:35
[T4-A-9-04] 地点データのインタラクティブなクラスタリングのための可視化インタフェース

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

17:35 ~ 18:00
[T4-A-9-05] 【技術報告】情報学と交通工学の融合に向けた研究開発 - 万博での交通整流化に向けて

西村拓哉 (株式会社NTTデータグループ)

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

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

スポーツ

座長:太田 学(岡山大学)
コメンテータ:木村 昌臣(芝浦工業大学)

16:05 ~ 16:20

[T5-A-9-01] 共通テスト用プログラム表記で記述された運動イメージと身体動作の差異認識アプリケーションの試作

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

16:20 ~ 16:35

[T5-A-9-02] 卓球の得点予測における重要要素の分析

*楊 明哲[1,2]、橋本 敦史[2]、馬 家昕[2]、本田 秀仁[3]、田中 一敏[2] (1. 東京大学、2. オムロンサイニックエックス、3. 追手門学院大学)

16:35 ~ 17:00

[T5-A-9-03] バドミントンの試合データを用いたショットの成功確率予測

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

17:00 ~ 17:25

[T5-A-9-04] ボール奪取と攻撃の防御に基づくアクションによるサッカー選手の攻守評価

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

17:25 ~ 17:50

[T5-A-9-05] 音響型ARオンラインマラソンシステムにおける空間認識率向上に向けた取り組み

*小西 侑樹[1]、Siriaraya Panote[2]、栗 達[3]、田中 克己[4]、河合 由起子[5]、中島 伸介[5] (1. 京都産業大学大学院、2. 京都工芸繊維大学、3. 福岡大学、4. 福知山大学、5. 京都産業大学)

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

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

セマンティクス

座長:岩井原 瑞穂(早稲田大学)
コメンテータ:北島 信哉(富士通株式会社)

16:05 ~ 16:20

[T5-B-9-01] 気象データを活用した省電力灌水システムの提案

*真鍋 武丸[1]、大塚 真吾[1] (1. 神奈川工科大学)

16:20 ~ 16:45

[T5-B-9-02] 機械学習と地理情報検索に基づく地物の意味発見

羽田野 唯人[1]、*渡邊 二葉[1]、田中 克己[1] (1. 福知山公立大学情報学部)

16:45 ~ 17:10

[T5-B-9-03] 交通計測地点の増加に対応する知識転移を用いた時空間グラフモデリング

*川端 智紀[1]、戸田 浩之[1] (1. 横浜市立大学)

17:10 ~ 17:35

[T5-B-9-04] 地理オブジェクトに関する事前学習を利用した地域特性推定

*福地 湧[1]、牛尼 剛聡[2]、田中 克己[3]、角谷 和俊[1] (1. 関西学院大学 社会情報デザイン研究室、2. 九州大学、3. 福知山公立大学)

17:35 ~ 18:00

[T5-B-9-05] チャット付き地図における会話・検索履歴に基づく地図検索手法

*小柳 慶真[1]、小竹 神[1]、北山 大輔[1] (1. 工学院大学)

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

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

メディアと社会

座長:若宮 翔子(奈良先端科学技術大学院大学)
コメンテータ:藤原 靖宏(日本電信電話株式会社)

16:05 ~ 16:30

[T5-C-9-01] メディアの報道に対するソーシャルメディアユーザの感情の可視化

*西上 貴雅[1]、風間 一洋[1]、吉田 光男[2]、土方 嘉徳[3] (1. 和歌山大学、2. 筑波大学、3. 関西学院大学)

16:30 ~ 16:55

[T5-C-9-02] 記事の書き手に着目したコメント生成によるフェイクニュース検出

*渡辺 春希[1]、森 康真[1]、田村 慶一[1] (1. 広島市立大学大学院・情報科学研究科)

16:55 ~ 17:20

[T5-C-9-03] ニュースの構成要素に着目したフェイクニュースのフェイク部分の抽出・訂正手法

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

17:20 ~ 17:45

[T5-C-9-04] メディアの報道に対するソーシャルメディアユーザの反応の時系列変化の分析手法

*井上 大成[1]、風間 一洋[1]、吉田 光男[2]、土方 嘉徳[3] (1. 和歌山大学、2. 筑波大学、3. 関西学院大学)

17:45 ~ 18:10

[T5-C-9-05] Twitterを用いたCOVID-19ワクチン接種の継続性に関係する情報共有行動の分析

*北林 遼大[1]、張 翔[1]、豊田 正史[2]、吉永 直樹[2]、梅本 和俊[2] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所)

オープニング・クロージング | 全体イベント

9:30 ~ 9:50 | オンライン・オープニング会場 オンライン（Zoom Events)

オンライン・オープニング (3)

9:30 ~ 9:50

[3OO-0930-1add] 参加案内

一般発表 | Track 1: 自然言語処理・機械学習基礎

🏛2024年3月1日(金) 10:00 ~ 12:10 | 🏛T1-A オンライン（Zoom Events)

数値データ処理

座長:油井 誠(トレジャーデータ株式会社)
コメンテータ:井上 潮(東京電機大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T1-A-7-01] HeavyOrdiNum-VL: 順序数を的確に把握し活用する視覚言語モデルの能力を探る
*増田 琉斗[1]、宮森 恒[1] (1. 京都産業大学)

10:30 ~ 10:55
[T1-A-7-02] HeavyOrdiNum-L: 順序数を的確に把握し活用する言語モデルの能力を探る
*林 寛治[1]、宮森 恒[1] (1. 京都産業大学)

10:55 ~ 11:20
[T1-A-7-03] メルカリデータセットを利用した商品価格推定の一手法
*森 優太郎[1]、上野 史[1]、太田 学[1] (1. 岡山大学)

11:20 ~ 11:45
[T1-A-7-04] Jellyfish: データ前処理のための大規模言語モデル
*張 皓辰[1]、董 于洋[2]、肖 川[1,3]、小山田 昌史[2] (1. 大阪大学、2. NEC、3. 名古屋大学)

HeavyOrdiNum-VL: 順序数を的確に把握し活用する視覚言語モデルの能力を探る

増田　琉斗†　　宮森　　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054227,miya}@cc.kyoto-su.ac.jp

あらまし　Transformer ベースの視覚言語モデルは，参照表現理解や視覚的質問応答などの様々なタスクで高い性能を達成している．自然画像中の物体など具体的エンティティを対象とした推論だけでなく，数や計算といった抽象的概念を対象とした推論についての研究も盛んに行われている．しかし，Transformer が数のような抽象的概念をどのように理解し活用しているかについては，まだ不明な点が多く残されている．このため，形状や色，材質が異なる複数の物体を配置した CG 画像を生成し，物体間の関係や数え上げが必須となる参照表現を付与することで，順序数の数え上げに焦点を当てた参照表現理解タスクのためのデータセット HeavyOrdiNum-VL を構築する．実験では，HeavyOrdiNum-VL を用いて，代表的な視覚言語モデルに対する参照表現理解タスクの性能や Transformer 内部の挙動を評価することで，各モデルの順序数に対する正確な理解と活用能力について分析する．

キーワード　順序数，数え上げ，概念理解，視覚言語モデル，参照表現理解，抽象概念，推論

1　はじめに

Transformer [1] は，ニューラル機械翻訳のモデルとして高い性能を達成して以降，自然言語処理分野における基盤的構造の一つとして，多くのモデルで活用されている．現在は，自然言語処理だけでなく，画像や音声といった他のメディア処理においても有用な共通的基盤技術として様々な分野で活用されている [2]．例えば，Vision Transformer [3] は，画像を分割した画像パッチを自然言語処理の単語のように Transformer のエンコーダに埋め込むことで，畳み込みを使わずに従来の画像認識モデルを大きく上回る性能を発揮した．さらに，画像処理と自然言語処理を融合的に扱う Vision-and-Language 分野においても Transformer ベースのモデルが多数提案され，活発に研究が進められている．例えば，CLIP [4] は，Web 上の大量の画像とテキストのペアで訓練され，強力なゼロショット性能を達成しており，Vision-and-Language 分野における基盤的なモデルの一つとして活用されている．

Transformer ベースのモデルが，どの程度の推論能力を保持しているのかを明らかにする研究も盛んに行われている．例えば，自然画像中の物体など具体的エンティティを対象とした視覚的常識推論 [5] や視覚言語的構成性 [6] を扱った研究だけでなく，数や計算といった抽象的概念を対象とした視覚的数学推論 [7] についての研究も盛んに行われている．しかし，これらのタスクにおいて，モデルは人間の推論能力とのギャップが依然として大きく，多角的にボトルネックを明らかにし改善を図っていく必要がある．

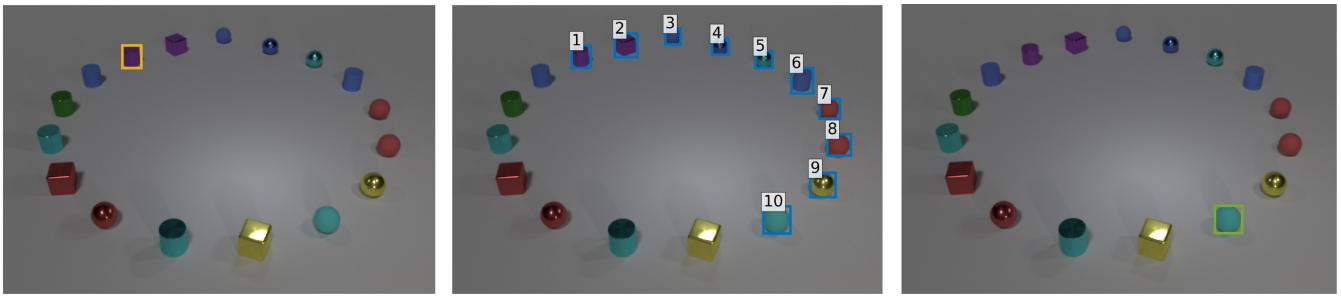
本稿では，視覚言語モデルの数を扱う能力に着目する．モデルが数を適切に扱う能力は，多くの複雑な推論タスクを遂行する上で重要である．例えば，人とロボットとのインタラクションにおいては，人間からの指示に数の概念理解が必要な内容が含まれている場合，ロボットは実世界の特定物体と数概念を関連付けた上で指示内容に従わないと，指示を正しく遂行できず，場合によっては致命的な事故につながる可能性もあると考えられる．

Transformer ベースの言語モデルには，数値推論など数値を扱う能力があることは従来研究で知られている [8]．特に，四則演算等の単純な算術問題においては高い正答率を示している [9], [10]．一方で，扱う数の桁数が 15 桁や 20 桁と増えるにつれて正答率が低下することも示されており，限界があることも明らかとなっている [11]．また，単純な算術問題の場合においても数の概念を理解した上で解いているかは不明である．

一般的に，自然数の概念は，順序数，基数の 2 種類で捉えることができるとされている [12]．順序数は，1 番目，2 番目，... といった数え上げにより何番目かを表す数であり，基数は，ものの集まりの大きさを表す数である．人間は，量を表す基数よりも順序を表す順序数をより早い時期に学習することが指摘されており，順序数は数の概念において基数よりも根源的な概念と考えられる [13], [14], [15].

従来研究では，基数を扱ったものが多く，順序数を扱った問題は，ほとんど取り組まれていない．基数に関する問題では，質問や参照表現中に出現する数をそのまま使って計算することで正答できることも多い．一方，順序数に関する問題では，数え上げや基数への変換といった処理が必要とされることが多く，順序に関する情報を的確に把握していなければ正答することができない．そのため，順序数に関する問題を用いてモデルの数学的推論能力を検証することで，そのモデルが数の概念を人間と近い形で理解しているのかどうかに迫るための新たな知見が得られるのではないかと考えた．

上記を踏まえ，本稿では，Transformer ベースの視覚言語モ



起点　　範囲 & 向き　　正解の境界矩形

参照表現: 紫色の円柱から時計回りに 10 番目に位置する物体

図 1: HeavyOrdiNum-VL データセットの例．物体数 17 個の画像に対して，起点として “紫色の円柱” の境界矩形，数え上げ対象範囲と向きとして “紫色の円柱から時計回り” に数えた 10 個の物体に対して境界矩形が付与されている．境界矩形の付与数は，参照表現中のある数え上げに用いられる順序数と画像中の物体数との最小値で決まる

デルが，順序数の概念をどの程度的確に把握し活用する能力を保持しているのかを明らかにすることを目的としたプロービングタスクを提案する．具体的には，形状や色，材質が異なる複数の物体を配置した CG 画像と，順序数を用いた数え上げが必要となる参照表現を用いた参照表現理解タスクを設定し，そのための新たなデータセット HeavyOrdiNum-VL を構築する．

本稿では，順序数を用いた 1 つの数え上げを遂行するためには，数え上げ対象の範囲，数え上げの起点，および，数え上げの向きをそれぞれ把握する必要があると考える．この三つ組（範囲，起点，向き）を 1 つ把握することが必要な参照表現を，数え上げスキル 1 の参照表現と呼ぶことにする．また，ある三つ組で定まる実体を，新たな数え上げの起点とすることで，参照表現を入れ子構造として構成することができる．入れ子の深さ n の参照表現を，数え上げスキル n の参照表現と呼ぶことにする．このように，数え上げスキル n を変化させることで，難易度の異なる順序数の数え上げに対するモデルの推論能力を検証することができる．

実験では，視覚言語モデルとして，MDETR [16]，OFA [17]，KOSMOS-2 [18] を対象に性能を評価する．対象となるモデルは，Transformer ベースの視覚言語モデルであり，参照表現理解や視覚的質問応答などの様々なタスクにおいて，競争力のある性能を示している．また，入力として，数え上げスキル 1 から 3 の参照表現，および，物体数が 3 から 22 までの画像を用いる．数え上げスキル，数え上げ対象の範囲の違いなどによる正答率や Transformer 内部の挙動の変化などを調査する．

実験の結果，

- 参照表現の数え上げスキルが大きくなると正答率が低下する傾向があること，
- 訓練時に経験した物体数の画像の場合，ある程度高い正答率を示すが，訓練時に経験した物体数を超えると，物体数が大きくなるほど正答率は低下する傾向があること，
- 参照表現に含まれる桁数が大きくなるほど正答率が低下する傾向があるが，正答率の低下の仕方は，訓練に含まれる順序数の桁数によって違いが見られること，
- Transformer の層ごとの重みは，MDETR では参照表現の数え上げスキルが大きくなるほど，多くの層で反応していること

などが明らかとなった．

本稿の貢献は以下の通りである．

（1） 視覚言語モデルが，順序数の概念をどの程度的確に把握し活用できるかを調査するための参照表現理解タスクを提案し，その検証に有用である HeavyOrdiNum-VL データセットを構築した．

（2） HeavyOrdiNum-VL データセットを用いたプロービングタスクを実施し，パラメータ数が異なる代表的な視覚言語モデルに対する推論能力を明らかにした．

（3） 順序数を用いた数え上げの能力について，異なる数え上げスキルの参照表現や異なる物体数の画像，異なる桁数の順序数などに応じた正答率の変化を明らかにした．

本論文の構成は，以下の通りである．2 節で関連研究について述べ，3 節で本稿で扱う問題設定とデータセットの構築について詳述する．4 節で実験内容と実験結果，考察を示し，5 節では，結論と課題を述べる．

2 関 連 研 究

2.1 数を理解し扱う能力に関する研究

深層ニューラルネットワークが数を理解し扱う能力に関する研究は，これまでも多くなされてきた．例えば，Wallace らは，数を扱う能力がどのように獲得されるのかを明らかにするために，数値推論に関する幾つかのタスクで，BERT や GloVe などの標準的なトークン埋め込み方式の性能を検証した [19]．Kajic らは，視覚言語モデルが，人間などの生物学的システムと同様に，小さな正確な数の表現と，より大きな近似的な量の表現を異なる方法で処理していることを示した [20]．

特に数を数える能力についても研究が活発に行われている．Bhattamishra らは，Transformer が，与えられた 2 桁までの 1 つの数から，それに続く 5 つの数を順に予測するために，デコーダの以前の予測に基づいて推論を実行できるように学習できることを示した [21]．

これらの研究では，深層ニューラルネットワークの数を理解し扱う能力について，主に桁数が限られた数や簡単な四則演算

を用いた検証が行われてきた．本稿では，数の基本概念として順序数に着目し，視覚言語モデルに対する新たな数的推論の検証タスクを提案する．

2.2 Transformer における知識の保存形態や挙動に関する研究

Transformer における知識の保存形態や挙動に関する研究もこれまでに取り組まれている．Dai らは，関係的な事実を表現するニューロンを特定するための知識帰属手法を提案した．そして，BERT の穴埋め問題タスクを調査し，得られた知識ニューロン活性化が，対応する事実の表現と正の相関があることを示した [22]．Wang らは，特定タスクに微調整した事前学習済みの Transformer 内には，タスク固有のスキルをエンコードしている幾つかのニューロンが存在し，これらのニューロン活性化がタスクラベルを高度に予測することを明らかにした [23]．Schwettmann らは，あるモダリティで学習した表現を他のモダリティの下流タスクに一般化する言語モデルの能力について調査し，画像説明文生成タスクにおいて，視覚表現を対応するテキストに変換するマルチモーダルニューロンを識別する手順を明らかにした [24]．

これらの研究では，Transformer ベースのモデルにおける知識の保存形態や挙動をさまざまな観点から特定するための手法の提案や検証が行われてきた．本稿では，Transformer ベースの視覚言語モデルを対象に，順序数の概念をどの程度的確に把握し活用できるかを検証するためのタスクを提案し，タスクを多角的に調査するためのデータセットを構築する．

2.3 参照表現理解に関する研究

参照表現理解に関する研究は古くからなされていたが，近年では大規模データセットが整備され，それらを本格的に活用するための研究が進められている．Kazemzadeh らは，自然画像中のオブジェクトに対する参照表現をクラウドソーシングで取得するための，新しい 2 名用のゲームを設計し，多様な参照表現の理解や生成に活用できる大規模データセットを構築した [25]．Subramanian らは，参照表現をサブクエリに分解し，CLIP を使用してこれらのサブクエリに対する独立した提案をスコア化し，出力を空間的経験則と結合することで参照表現理解をゼロショットで実現する手法を提案した [26]．Johnson らは，既存の視覚的質問応答データセットには，自然画像による強いバイアスが含まれ，モデルが推論なしに正答する場合があるため，それらのバイアスを軽減した CLEVR データセットを構築した [27]．このデータセットでは，形状や色，材質が異なる複数の物体が無地のテーブル上に配置された CG 画像が用いられ，画像中の物体や状況に関する質問と正答との三つ組として提供されている．Liu らは，既存の参照表現理解データセットには，自然画像による強いバイアスが含まれるため，CLEVR データセットと同じ画像を用いて，各画像に様々な参照表現を付与し，参照表現理解の検証に活用できる CLEVR-Ref+データセットを構築した [28]．

従来研究では，順序数を含む数え上げが必要な参照表現はほとんどなく，順序数を含む参照表現が存在する場合も，1 桁までの限られた順序数しか用いられてこなかった．

本稿では，2 桁以上の順序数も含む数え上げを要する参照表現と，物体配置を工夫した画像を作成し，順序数を用いた数え上げに着目した参照表現理解の検証に活用できるデータセット HeavyOrdiNum-VL を構築した．

3 タスク設定とデータセット構築

3.1 順序数を用いた数え上げを要する参照表現理解タスク

本稿では，順序数を用いた数え上げを要する参照表現理解タスクを設定する．本タスクでは，参照表現テキスト R と画像 I のペア (R,I) がモデルに与えられ，参照表現 R に対応する境界矩形 (BoundingBox) の集合 $B = \{b_i\}$ を出力することが求められる．ここで，参照表現テキスト R に，順序数を用いた数え上げを要する表現が含まれる点が通常の参照表現理解とは異なる．以下では，本タスクのためのデータセット HeavyOrdiNum-VL の構築について説明する．

3.2 データセットに含まれる画像

本データセットに含まれる画像として，大きな順序数での数え上げも可能となるように物体を円状に配置した新たな画像を生成した．画像を生成する際は，CLEVR データセットにおける画像の作成方法を参考とした [27]．

物体を円状に配置した画像は計 5 万枚作成した．画像中の物体数を 3 個から 22 個まで変化させ，物体数ごとに 2500 枚の画像を生成した．物体の色と材質，形については，CLEVR データセットと同様に，それぞれ 7 種類，2 種類，2 種類の種別を採用した．画像を生成する際は，少なくとも 1 つの物体は他の物体特徴と重複しないように生成し，その物体を起点として指定できるように配慮した．

以上で作成した画像 5 万枚を，訓練，検証，テストにそれぞれ 8:1:1 の割合で分割した．データセットに含まれる画像の内訳を，表 2 に示す．表 2 における物体数 3 個から 12 個の画像は，物体数 3 個から 22 個の画像の部分集合で構成された画像である．これは，訓練時に経験していない物体数の物体を数え上げる能力を有するかを調査する際に用いられる．

3.3 順序数を用いた数え上げを要する参照表現

本稿で考える参照表現は，その中に順序数を用いた数え上げを要する内容が含まれているものである．そのような参照表現として最も基本的なものは，順序数を用いた数え上げが 1 つ含まれた内容の表現である．ここで，順序数を用いた 1 つの数え上げを遂行するためには，数え上げ対象の範囲，数え上げの起点，および，数え上げの向きをそれぞれ把握する必要があると考えられる．この三つ組（範囲，起点，向き）を 1 つ把握することが必要となる内容の参照表現を，数え上げスキル 1 の参照表現と呼ぶことにする．

また，ある三つ組（範囲，起点，向き）で定まる実体を，新たな数え上げの起点とすることで，参照表現を入れ子構造として構成することができる．入れ子の深さ n の参照表現を，数え

表 1: 順序数を用いた数え上げを要する参照表現の例

数え上げスキル数	参照表現	数え上げ対象	起点の指定	向きの指定
1	2 番目に位置する物体	所与	なし	なし
	赤色の物体から 3 番目に位置する物体	所与	あり	なし
	反時計回りに 4 番目に位置する物体	所与	なし	あり
	赤色の物体から時計回りに 5 番目に位置する物体	所与	あり	あり
	ひとつおきに数えて 6 番目に位置する物体	要推論	なし	なし
	黄色の金属製の円柱からひとつおきに数えて 7 番目に位置する物体	要推論	あり	なし
	時計回りにひとつおきに数えて 8 番目に位置する物体	要推論	なし	あり
	黄色の金属製の円柱から時計回りにひとつおきに数えて 9 番目に位置する物体	要推論	あり	あり
2	赤色の物体から時計回りに 5 番目に位置する物体から数えて，時計回りに，2 番目に位置する物体	所与	あり	あり
3	赤色の物体から時計回りに 5 番目に位置する物体から数えて，黄色の金属製の円柱から時計回りにひとつおきに数えて 9 番目に位置する物体を数えた向きと同じ方向に，4 番目に位置する物体	所与	あり	あり

† スペースの都合上，数え上げスキル 2 以上の参照表現の数え上げ対象，起点，向きについては，最後の数え上げについての種別のみを記載している．また，数え上げスキル 2 以上の参照表現については，代表例を 1 種類のみ示している．

表 2: データセットに含まれる画像の内訳

物体数	train	val	test	total
3-22	40,000	5,000	5,000	50,000
3-12	20,000	2,500	2,500	25,000

上げスキル n の参照表現と呼ぶことにする．このように，数え上げスキル n を変化させることで，難易度の異なる順序数数え上げに対するモデルの推論能力を検証することができる．

さらに，数え上げスキル n の参照表現と画像が与えられた際に，参照表現理解がどの程度達成できているのかを段階的に検証するために，n 番目までの各数え上げについての三つ組（範囲，起点，向き）に対応する境界矩形をデータセットに追加する．

例えば，図 1 に示す物体数 17 個の画像に対して “紫色の円柱から時計回りに 10 番目に位置する物体 ” という参照表現が与えられた場合，起点として “紫色の円柱” の境界矩形，数え上げ対象範囲と向きとして，“紫色の円柱から時計回り” に数えた最大 m 個の物体に対して境界矩形を付与する．ここで，参照表現中のある数え上げに用いられる順序数を i, 画像中の物体数を j とすると，m は以下の通り定義される: $m = min(i, j)$

これにより，本タスクの達成度を，途中段階の処理結果に遡って検証することが可能となり，より多角的な分析を実現することができる．

以下，順序数を用いた数え上げを構成する三つ組の各項目についてより詳細に説明する．

3.3.1 数え上げ対象の範囲

数え上げ対象の範囲については，以下の 2 つに分けて考える．数え上げの対象が画像内の物体集合の全ての物体に相当する場合は，与えられたままの物体集合という意味で “所与” と呼ぶことにする．数え上げの対象が画像内の物体集合のうちの一部を何らかの方法で選択したり並び替えたりして得られる結果である場合，何らかの推論が必要になるという意味で “要推論” と呼ぶことにする．

例えば，図 1 に示す物体数 17 個の画像に対して，“紫色の円柱から時計回りに 10 番目に位置する物体” という参照表現が与えられた場合は，数え上げ対象の範囲が，画像中で与えられた物体集合の全物体であるため，“所与” に該当する．一方，同じ画像に対して，“紫色の円柱から時計回りに一つおきに数えた 10 番目に位置する物体” という参照表現が与えられた場合は，画像中の物体集合の全物体をそのまま数え上げる訳ではなく，紫色の円柱から一つおきに並んでいる物体を選択した上で数え上げる必要があり，数え上げ対象の推論が必要であるため，“要推論” に該当する．

3.3.2 起点と向き

起点と向きについては，参照表現内で具体的な表現が与えられているか否かによってそれぞれ 2 種類に分類する．

例えば，“赤色の物体から 2 番目” や “青色の物体から 10 番目” などのように，数え上げの起点が “赤色の物体 ” や “ 青色の物体” と具体的に指定されている場合は，起点の指定 “あり” に該当する．同様に，“時計回りに” や “反時計回りに” などのように，数え上げの向きについても具体的に指定されている場合は，向きの指定 “あり” に該当する．

一方，“2 番目に位置する物体” という参照表現では，起点と向きについて具体的な表現が与えられていないため，指定 “なし” に該当する．この場合，物体が円状に並んだ画像である場合は，時計と同様に，12 時の位置，あるいは，一番奥側にある位置を数え上げ起点とすることが多いため，それを推論して把握する必要がある．同様に，向きに関しても，指定がない場合は時計回りに数えることが一般的とされており，時計回りとすることを推論して把握する必要がある．

3.3.3 数え上げスキル別の参照表現

本稿では，数え上げスキル $n(n = 1, 2, 3)$ の参照表現を構築する．

数え上げスキル 1 の参照表現については，8 タイプの参照表現を作成した．その内訳は，数え上げ対象の範囲が “所与” と

表 3: 構築したデータセット HeavyOrdiNum-VL の内訳

id	物体数	桁数	数え上げスキル	split	参照表現の数
M1	3-22	1-3	1-3	train	960,000
				val	120,000
				test	120,000
				total	1,200,000
M2		1-6		train	960,000
				val	120,000
				test	120,000
				total	1,200,000
M3	3-12	1-3		train	480,000
				val	60,000
				test	60,000
				total	600,000
M4		1-6		train	480,000
				val	60,000
				test	60,000
				total	600,000
T1	3-22	1	1	test	40,000
T2	3-12	1-8	1	test	20,000
T3	3-22	1-3 (mixed*)	1-3	test	120,000

* “mixed” は，数え上げスキル 2,3 の参照表現内の順序数が，異なる桁数の順序数で構成されていることを表す．

“要推論” の 2 タイプ，起点，向きがそれぞれ指定 “あり”, “なし” の 2 タイプずつについての，全ての組み合わせである．

数え上げスキル 2 の参照表現については，1 つ目の数え上げ結果で定まる物体が，2 つ目の数え上げの起点となるため，2 つ目の数え上げでは，数え上げ対象の範囲が 2 タイプ (“所与” と “要推論”)，向きが 2 タイプ (指定 “あり” と “なし”) となる．従って，1 つ目の数え上げが 8 タイプであることから，数え上げスキル 2 の参照表現については，32 タイプ (=8x2x2) の参照表現を作成した．

同様に，数え上げスキル 3 の参照表現については，128 タイプ (=32x2x2) の参照表現を作成した．

表 1 に，本稿で作成した参照表現の例とタイプを示す．なお，スペースの都合上，数え上げスキル 2 および 3 の参照表現のタイプについては，代表例を 1 種類のみ示しており，範囲，起点，向きの，各タイプの内容は，最後の数え上げについての内容のみを記載している．

3.3.4 参照表現に用いる順序数の桁数

数え上げスキル n の参照表現を作成する際，1-3 桁までの順序数，および，1-6 桁までの順序数の 2 種類を用いた参照表現をそれぞれ作成する．ここで，$n = 2,3$ の場合，全ての数え上げで用いられる順序数は同じ桁数である．

例えば，表 1 内の数え上げスキル 2 の参照表現に示すように，1 桁の順序数を用いる場合は，全ての数え上げで 1 桁の順序数が用いられる．一方の数え上げで 1 桁，他方の数え上げで 2 桁の順序数が用いられるような参照表現は作成しない．

3.4 データセット

表 3 に，構築したデータセット HeavyOrdiNum-VL の内訳を示す．

表 2 の物体数 3-22 個の画像に対して，桁数が 1-3 桁および 1-6 桁の範囲に限定された順序数による数え上げスキル 1-3 の参照表現が付与されたデータセットを，それぞれ M1, M2 と呼ぶこととする．同様に，表 2 の物体数 3-12 個の画像に対して，桁数が 1-3 桁および 1-6 桁の範囲に限定された順序数による参照表現を付与したデータセットを，それぞれ M3, M4 と呼ぶこととする．

M1-M4 とは異なる条件での検証を実現するためのテスト用のデータセットを 3 種類作成した．

T1 は，表 2 の物体数 3-22 個の画像に対して，桁数が 1 桁のみの順序数による数え上げスキル 1 の参照表現が付与されたテスト用データセットである．これは，訓練時に見たことのない物体数の画像についても本タスクの推論を的確に達成できるかについて検証する目的で用いる．

T2 は，表 2 の物体数 3-12 個の画像に対して，桁数が 1-8 桁の範囲に限定された順序数による数え上げスキル 1 の参照表現が付与されたテスト用データセットである．これは，訓練時に見たことのない桁数の順序数による参照表現についても本タスクの推論を的確に達成できるかについて検証する目的で用いる．

T3 は，表 2 の物体数 3-22 個の画像に対して，桁数が 1-3 桁の範囲に限定された順序数による数え上げスキル 1-3 の参照表現が付与されたテスト用データセットである．M1 と構成が類似しているが，数え上げスキル $n(n = 2,3)$ の参照表現において，各数え上げに用いられる順序数が異なる桁数で構成されている点が M1 とは異なる．これは，数え上げスキル $n(n = 2,3)$ の参照表現において各数え上げが異なる桁数で構成されている場合でも本タスクの推論を的確に達成できるかについて検証する目的で用いる．

4 実　　験

4.1 モデルのセットアップ

実験では，MDETR，OFA，KOSMOS-2 を対象に性能を評価する．

MDETR は，人工画像とテキストペアで構成されたデータセット CLEVR-Medium [16] で事前学習されている．OFA は，自然画像とテキストのペアで構成された多数のデータセット [17] で事前学習されている．これらの MDETR と OFA の事前学習済みモデルを，HeavyOrdiNum-VL データセットで追加学習する．追加学習では，表 3 に示す HeavyOrdiNum-VL データセットのうち，画像中の物体数や順序数の桁数が異なる M1-M4 の各 train の分割を用いて訓練する．表 4 に示すように，バッチサイズ 4 で 30 エポックの訓練を行う．

KOSMOS-2 は，画像中のオブジェクトへの参照を，特殊トークンとして境界矩形の形でテキストに含んでいる，GRIT と呼ばれる画像とテキストのペアで構成される膨大なデータセットで訓練されている．このモデルは，言語モデルで問題となる幻

表 4: 評価対象のモデル，パラメータ数，および，ハイパーパラメータ

model	パラメータ数	追加学習	
		バッチサイズ	エポック
OFA_tiny	33M	4	30
MDETR	111M	4	30
OFA_base	182M	4	30
KOSMOS-2	1.6B	–	–

覚や不自然な出力を軽減し，モデルの精度と信頼性が向上するとされている．

4.2 評価指標

評価では，正答率と Transformer の内部挙動を分析する．

4.2.1 正答率

正答率としては，以下の観点で正答率の変化を調べる．

（1） 数え上げスキルの異なる参照表現での正答率
（2） 数え上げ範囲のタイプ別での正答率
（3） 起点や向きの指定の有無による正答率
（4） 異なる物体数の画像による正答率
（5） 異なる桁数の順序数による正答率

4.2.2 Transformer の内部挙動

Transformer の内部挙動としては，以下の観点で内部挙動の変化を調べる．

（1） 数え上げスキルの異なる参照表現での層ごとの重み
（2） 参照表現における起点のテキストトークンが対応する物体への注意

本稿では，VL-InterpreT [29] を使用することで，MDETR のデータセット M3 の学習済みモデルにおける Tranformer の内部挙動を調査する．

具体的には，z-score で正規化された平均クロスモーダルアテンションを層ごとの平均値として求める．そして，数え上げスキル別にそれらの合計値を求め，平均値を取ることで，数え上げスキル別の各層の注意の強さを比較する．後者は，正解の画像と参照表現の場合において，起点に相当するテキストトークンが，画像中の起点に対応する物体に注意を向けているかを確認する．

4.3 実験結果

データセット M1-M4 のそれぞれで訓練したモデルにより，対応するテストデータで検証した結果を表 5, 6 に示す．表 5 は，物体数 3-22 個の画像に対応する M1, M2 でのテスト結果を示し，表 6 は，物体数 3-12 個の画像に対応する M3, M4 でのテスト結果を示している．

全体の正答率では，MDETR の結果が最も高くなり，KOSMOS-2 の結果は最も低くなった．物体数 3-22 個の画像に対応する M1, M2 のデータセットを用いた場合は，MDETR が，OFA_base よりも高い正答率となったが，大きな差は見られなかった．しかし，物体数 3-12 個の画像に対応する M3, M4 のデータセットを用いた場合では，MDETR と OFA_base と

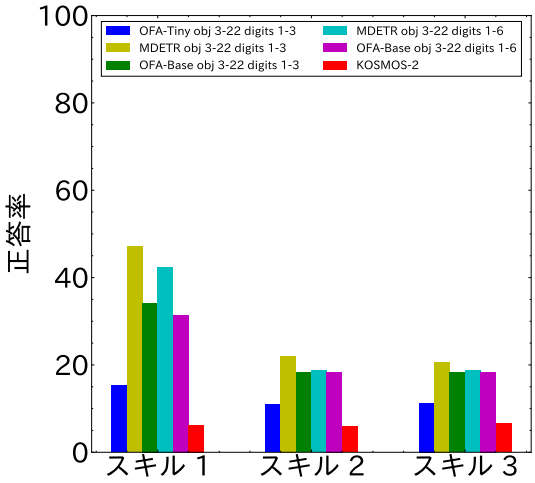


図 2: テストデータ T3 による異なる数え上げスキルの参照表現での正答率 (%)．T3 は，物体数 3 個から 22 個の画像に対して，1 桁〜3 桁の順序数による数え上げスキル 1-3 の参照表現が付与されたテストデータである．数え上げスキル $n(n = 2, 3)$ の参照表現では，各数え上げに用いられる順序数が異なる桁数で構成されている

の間で正答率に大きな差が生じた．両者とも訓練時の物体数が小さくなるほど，正答率が高くなった．パラメータ数の小さい OFA_tiny の正答率は，他の学習済みモデルと比較して非常に低くなった．

4.3.1 数え上げスキルの異なる参照表現での正答率

異なる数え上げスキルの参照表現での正答率では，いずれのモデルでも，数え上げスキル 1 の参照表現では比較的高い正答率を示すが，数え上げスキルが 2 以上の参照表現になると正答率は低下する傾向が見られる．

テストデータ T3 での正答率を図 2 に示す．M1 と構成が似ているが，数え上げスキル $n(n = 2, 3)$ の参照表現において，各数え上げに用いる順序数が異なる桁数で構成されている点が異なる．表 5 の M1 の数え上げスキル 1 の結果と変わらないのに対して，数え上げスキル 2,3 の結果は，表 5 の M1 の結果と比べて大きく低下した．

4.3.2 数え上げ範囲のタイプ別での正答率

数え上げ範囲のタイプ別での正答率は，学習データの物体数に応じて異なる結果が得られた．

物体数 3-22 に対応する M1, M2 による結果では，要推論よりも所与の順序数において高い正答率を得られた．特に OFA_base では，所与と要推論との間で結果が大きく変化した．

物体数 3-12 に対応する M3, M4 による結果では，OFA_tiny と MDETR では，要推論の方が高い正答率を得たのに対して，OFA_base, KOSMOS-2 では，所与の方が高くなった．

4.3.3 起点の指定の有無による正答率

起点の指定の有無による正答率では，KOSMOS-2 を除くほとんどのモデルにおいて，起点を指定しない方が高い傾向にあることがわかった．特に，OFA_tiny による正答率では，その差が著しくなった．

表 5: 物体数 3-22 個の画像を含むデータセット M1,M2 を用いたモデルによる本タスクの正答率 (%)

数え上げスキル	比較要素		モデルおよび用いたデータセット種別						
			M1 (1-3 桁の範囲に限定された順序数)				M2 (1-6 桁の範囲に限定された順序数)		
			OFA_tiny †	MDETR	OFA_base	KOSMOS-2 ††	MDETR	OFA_base	KOSMOS-2 ††
1	全データ		15.3	46.6	34.3	6.1	33.5	25.2	6.7
	数え上げ対象	所与	15.3	46.6	37.7	6.8	33.6	26.6	7.4
		要推論	15.3	46.5	30.9	5.4	33.4	23.7	5.9
	起点	あり	11.9	43.9	33.9	6.2	31.6	24.7	6.9
		なし	18.7	49.3	34.7	5.9	35.4	25.6	6.5
	向き	あり	15.5	46.6	35.1	6.4	33.6	24.9	6.9
		なし	15.1	46.5	33.5	5.7	33.3	25.4	6.5
2	全データ		11.6	37.9	29.1	5.5	24.1	21.8	6.0
3	全データ		11.6	29.4	29.5	6.6	23.2	21.8	6.7
全体	全データ		12.8	37.9	30.9	6.1	26.9	22.9	6.5

† OFA_tiny については，M1 での正答率が低いことや，M1 より M2 の方が難易度が高いことから，M1 での実験のみ行った
†† KOSMOS-2 については，高いゼロショット性能を達成していることから，追加学習なしでの検証を行なった

表 6: 物体数 3-12 個の画像を含むデータセット M3,M4 を用いたモデルによる本タスクの正答率 (%)

数え上げスキル	比較要素		モデルおよび用いたデータセット種別						
			M3 (1-3 桁の範囲に限定された順序数)				M4 (1-6 桁の範囲に限定された順序数)		
			OFA_tiny †	MDETR	OFA_base	KOSMOS-2 ††	MDETR	OFA_base	KOSMOS-2 ††
1	全データ		23.8	83.3	40.7	10.1	63.9	32.9	10.9
	数え上げ対象	所与	22.9	83.3	42.3	11.0	62.1	34.2	11.8
		要推論	24.7	83.3	39.1	9.1	65.8	31.6	9.9
	起点	あり	21.4	83.2	40.9	10.2	63.8	32.2	11.0
		なし	26.2	83.4	40.5	9.9	64.1	33.6	10.7
	向き	あり	24.1	83.2	40.9	10.6	64.4	33.7	10.9
		なし	23.5	83.4	40.5	9.6	63.5	32.1	10.8
2	全データ		17.1	80.1	34.9	9.5	61.5	24.9	10.1
3	全データ		23.8	79.6	35.3	10.8	60.8	23.9	10.8
全体	全データ		19.4	80.9	36.9	10.1	62.1	27.2	10.6

† OFA_tiny については，M3 での正答率が低いことや，M3 より M4 の方が難易度が高いことから，M3 での実験のみ行った
†† KOSMOS-2 については，高いゼロショット性能を達成していることから，追加学習なしでの検証を行なった

4.3.4 向きの指定の有無による正答率

向きの指定の有無による正答率では，ほとんどのモデルにおいて，向きを指定した方が高い傾向にあることがわかった．

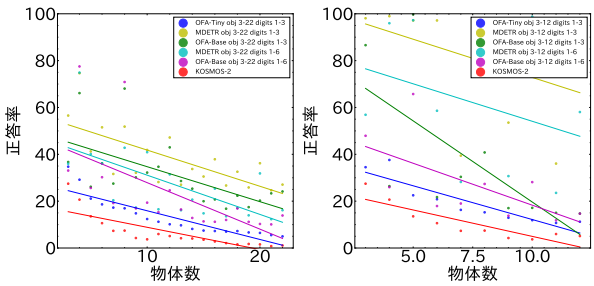
4.3.5 異なる物体数の画像による正答率

図 3 は，表 5, 6 の正答率を，物体数別にまとめた結果である．異なる物体数の画像による正答率では，画像中の物体数増加と同時に，正答率が低下する傾向にあることがわかった．

テストデータ T1 での正答率を図 4 に示す．T1 は，物体数 3-22 個の画像に対して，1 桁のみの順序数による数え上げスキル 1 の参照表現が付与されたテストデータである．

物体数が 3-22 個の画像で学習した場合，OFA_tiny を除く全てのモデルで，比較的に高い正答率を示した．特に OFA_tiny と KOSMOS-2 では，物体数増加と同時に正答率の低下の仕方が大きくなった．

物体数が 3-12 個の画像で学習した場合，訓練時に経験した物体数の画像の場合，ある程度高い正答率を示すが，訓練時に経験した物体数を超え，物体数が大きくなるほど正答率は，著しく低下する傾向があることがわかった．特に MDETR では，OFA と KOSMOS-2 と比較して大幅に低下し，物体数 15 個の画像を超えた時から，OFA を下回った．



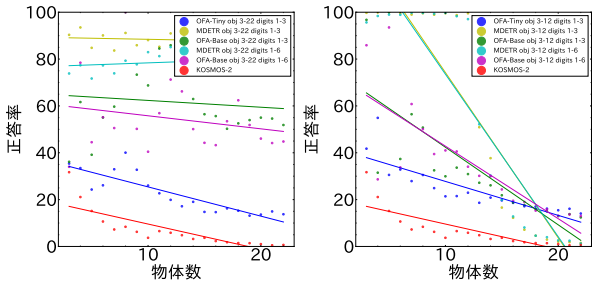
(a) 物体数 3-22 個の画像に対応する M1，M2 での結果　(b) 物体数 3-12 個の画像に対応する M3，M4 での結果

図 3: 異なる物体数の画像による正答率 (%)

4.3.6 異なる桁数の順序数による正答率

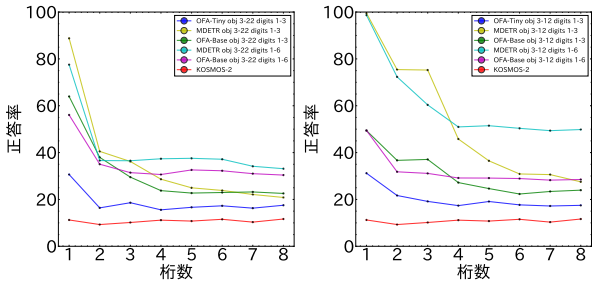
テストデータ T2 での正答率を図 5 に示す．T2 は，物体数 3-12 個の画像に対して，1-8 桁の範囲に限定された順序数による数え上げスキル 1 の参照表現が付与されたテスト用データセットである．

桁数が 1-3 桁の順序数が用いられる参照表現で訓練したモデルでは，訓練時に経験した桁数の場合，ある程度高い正答率を示すが，訓練時に経験した桁数を超え，桁数が大きくなるほど正答率は低下する傾向があることがわかる．特に，桁数が 1-3



(a) 物体数 3〜22 個の画像に対応する M1，M2 での結果　(b) 物体数 3-12 個の画像に対応する M3，M4 での結果

図 4: テストデータ T1 による異なる物体数の画像での正答率 (%)．T1 は，物体数 3-22 個の画像に対して，1 桁の順序数による数え上げスキル 1 の参照表現が付与されたテストデータである



(a) 物体数 3-22 個の画像に対応する M1，M2 での結果　(b) 物体数 3〜12 個の画像に対応する M3，M4 での結果

図 5: テストデータ T2 による異なる桁数の順序数を用いた参照表現での正答率 (%)．T2 は，物体数 3-22 個の画像に対して，1-8 桁の順序数による数え上げスキル 1 の参照表現が付与されたテストデータである

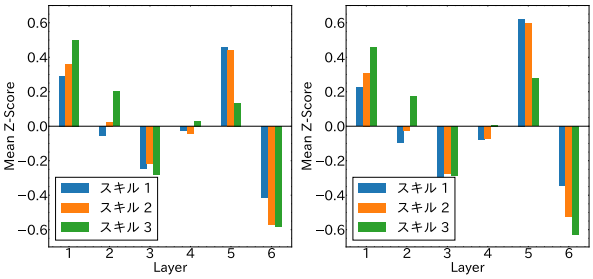
桁の順序数が用いられる参照表現で訓練したモデルは，桁数が 1-6 桁の順序数が用いられる参照表現で訓練したモデルに比べて，3 桁まではより高い正答率を維持する傾向があるが，4 桁以降は，1-6 桁までで訓練したモデルに比べて，さらに正答率が低下する傾向がある．

4.3.7 数え上げスキルの異なる参照表現での層ごとの重み

MDETR の M3 の学習済みモデルの各層における平均クロスモーダルアテンションの結果を図 6 に示す．正解と不正解のどちらの場合でも，1 層目と 5 層目で比較的に高い値を示した．数え上げスキルによる比較では，正解と不正解に関係なく，数え上げスキル 3 の場合では，1 層，2 層，4 層，5 層と多くの層で高い値を示した．数え上げスキル 2 の場合は，1 層，2 層，5 層で高い値を示し，数え上げスキル 1 では，1 層と 5 層で高くなり，数え上げスキルが増えるほど，多くの層で高い値を示した．

4.3.8 参照表現における起点のテキストトークンが対応する物体への注意

MDETR の M3 の学習済みモデルがテストデータで正解した画像と参照表現における平均クロスモーダルアテンションのヒートマップを図 7 に示す．この時，5 層と 1 番目のヘッドに



(a) 正解した場合の平均クロスモーダルアテンションを層ごとに平均化した結果　(b) 不正解した場合の平均クロスモーダルアテンションを層ごとに平均化した結果

図 6: M3 のテストデータによる異なる数え上げスキルの参照表現での MDETR の各層における平均クロスモーダルアテンションの結果．M3 は，物体数 3-12 個の画像に対して，1-3 桁の順序数による数え上げスキル 1-3 の参照表現が付与されたデータセットである．検証した MDETR は，M3 の学習済みモデルである．それぞれの参照表現における MDETR のクロスモーダルアテンションを平均化した結果を表示している

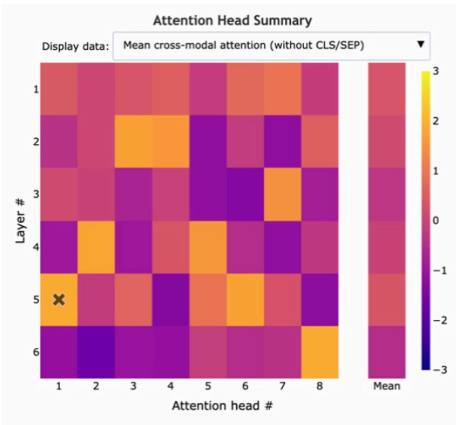


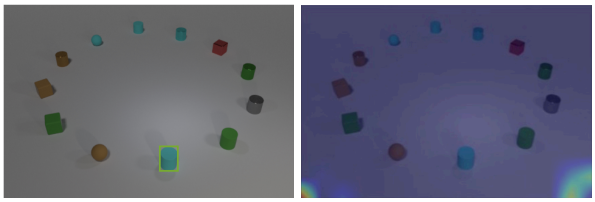
図 7: 平均クロスモーダルアテンションのヒートマップ．図 8 の左図の画像と，参照表現 “茶色の金属製の物体から数えて反時計回りに 5 番目に位置する物体” における層とヘッドのヒートマップである

おいて一番高い値を示した．正解した画像は，図 8 の左図であり，参照表現は，数え上げスキル 1 の “茶色の金属製の物体から数えて反時計回りに 5 番目に位置する物体” である．平均クロスモーダルアテンションの値が一番高くなった 5 層と 1 番目のヘッドに着目し，起点のテキストトークン “物体” に対応する画像中の注意を示した結果を、図 8 の右図に示す．

図 8 の右図に示すように，起点に対応するテキストトークンにおける画像中の注意は，起点に対応する物体に注意が見られなかった．

4.4 考　　察

数え上げスキル 1 の参照表現では，いずれのモデルも一定の高い正答率を示すが，数え上げスキルが 2 以上の参照表現になると正答率は低下する傾向があることが明らかとなった．ただし，パラメータ数が十分に大きくないモデルや追加学習を行



(a) MDETR の出力の境界矩形（正解）　(b) 起点のテキストトークンに対応する画像中の注意

図 8: M3 の MDETR の事前学習済みモデルによる出力の境界矩形（左図）と参照表現の起点トークン “物体” に対応する画像中の注意（右図）．参照表現は，“茶色の金属製の物体から数えて反時計回りに 5 番目に位置する物体” である．右図は，平均クロスモーダルアテンションが最も高い 5 層と 1 番目のヘッドに着目した可視化結果である

なっていないモデルでは数え上げスキル 1 の参照表現の場合でも正答率は低かった．これは，前者では順序数を用いた数え上げを複数組み合わせた複雑な参照表現を理解し推論する能力はまだ限定的であることを示唆している．後者は，追加学習を行なっていないため CG 画像中の物体を理解できなかったため，正答率が低くなったと考えられる．今後は，KOSMOS-2 などのより大きなパラメータ数のモデルの追加学習を行い，検証する必要がある．

特に図 2 が示すように，同じデータセットで学習されたモデルで比較すると，MDETR と OFA では，数え上げスキル 1 の正答率と比較して，数え上げスキル 2,3 の参照表現では，正答率が大きく低下した．これは，複雑な参照表現に対するモデルの頑健性が低いことを示唆している．

起点の指定の有無による正答率の違いでは，追加学習を行なったモデルは，起点を指定しない方が比較的に高い正答率を示した．特に OFA_tiny については，その差が著しくなった．これは，起点の指定がない場合は，時計と同様に，12 時の位置を数え上げ起点とすることが多いため，常に同じ物体を起点とすれば良い点で，指定ありの場合よりむしろ推論が容易であった可能性が考えられる．ただ，表 6 の M3 での MDETR と OFA_base では，起点の指定の有無による差が非常に小さいことから，MDETR と OFA_base では，そのような推論の容易さの差は正答率に大きな影響を与えていないと考えられる．

テストデータ T1 による異なる物体数の画像による正答率の結果では，図 4 に示すように，OFA_tiny と KOSMOS-2 を除く全てのモデルで高い正答率を示した．これは，T1 に含まれる参照表現が数え上げスキル 1 で 1 桁の順序数しか含まれていない簡単な参照表現であること，T1 に含まれる画像中の物体数はいずれも訓練時に経験したことのある物体数であることが要因として考えられる. ただ，OFA_tiny については，訓練時に経験したことのある物体数の場合でさえ，高い正答率を達成することはできないことがわかる．また，物体数が 3-12 個の画像で学習した場合，訓練時に経験した物体数の画像の場合，ある程度高い正答率を示すが，訓練時に経験した物体数を超えると，物体数が大きくなるほど正答率は低下する傾向があることがわかる．

これは，今回の評価対象モデルはいずれも，順序数を用いた数え上げを要する参照表現理解タスクにおいて，分布外汎化を達成できていないことを示している．

テストデータ T2 での正答率の結果では，図 5 に示すように，桁数が 1-3 桁の順序数が用いられる参照表現で訓練したモデルは，桁数が 1-6 桁の順序数が用いられる参照表現で訓練したモデルに比べて，3 桁まではより高い正答率を維持する傾向があるが，4 桁以降は，1-6 桁までで訓練したモデルに比べて，さらに正答率が低下する傾向がある．

これも，今回の評価対象モデルはいずれも，順序数を用いた数え上げを要する参照表現理解タスクにおいて，分布外汎化を達成できていないことを示している．

Transformer の層ごとの重みによる比較では，MDETR は数え上げスキルが増えるほど，多くの層で反応した．これは，複雑な推論であるほど，多くの層に知識が保持されており，推論時に活用されていることを示唆している．

参照表現における起点のテキストトークンが対応する物体への注意では，MDETR が正解した数え上げスキル 1 の画像と参照表現であったのに対して，起点のテキストトークンに対応する，起点の物体への注意が見られなかった．これは，モデルが順序数の概念を把握して，推論しているのではなく，確率的に最適化された形で推論しているに過ぎないことを示唆している．今後は，他のモデルによる Transformer の内部挙動の傾向を調査することや，他の検証方法により，内部挙動を明確化していく必要がある．

5 ま と め

本稿では，Transformer ベースの視覚言語モデルが，順序数の概念をどの程度的確に把握し，活用する能力を持っているかを明らかにするため，新たな参照表現理解タスクを提案し，その検証に有用な HeavyOrdiNum-VL データセットを構築した．本データセットには，参照表現理解の正解データだけでなく，参照表現理解がどの程度達成できているのかを段階的に検証するために，各数え上げに把握が必要となる三つ組（範囲，起点，向き）に対応する境界矩形も含まれている．これにより，本タスクの達成度を，途中段階の処理結果に遡って検証することが可能となり，より多角的な分析を実現することができる．

実験では，構築したデータセットを用いて，代表的な視覚言語モデルに対する参照表現理解タスクの性能を評価し，順序数に対する正確な理解と活用能力について分析した．その結果，参照表現の数え上げスキルが大きくなると正答率が低下する傾向があること，訓練時に経験した物体数の画像の場合，ある程度高い正答率を示すが，訓練時に経験した物体数を超えると，物体数が大きくなるほど正答率は低下する傾向があること，参照表現に含まれる桁数が大きくなるほど正答率が低下する傾向があるが，正答率の低下の仕方は，訓練に含まれる順序数の桁数によって違いが見られることなどが明らかとなった．また，数え上げスキルが増えるほど，Transformer 内部の多くの層で

反応していることや正解の画像と参照表現を用いた場合であっても，起点を把握して推論を行なっていない可能性を示唆した．

今後は，より大きなパラメータ数のモデルの追加学習を行い検証を行うことや，他の Transformer 内部挙動の調査方法も用いることで，分布外汎化を実現するための新たな枠組みや訓練方法の検討を進める予定である．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[2] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[5] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

[6] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

[7] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

[8] Avijit Thawani, Jay Pujara, Pedro A Szekely, and Filip Ilievski. Representing numbers in nlp: a survey and a vision. *arXiv preprint arXiv:2103.13136*, 2021.

[9] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

[10] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.

[11] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019*, 2021.

[12] Bahram Assadian and Stefan Buijsman. Are the natural numbers fundamentally ordinals? *Philosophy and Phenomenological Research*, Vol. 99, No. 3, pp. 564–580, 2019.

[13] Charles J Brainerd. The origins of number concepts. *Scientific American*, Vol. 228, No. 3, pp. 101–109, 1973.

[14] Charles J Brainerd. Inducing ordinal and cardinal representations of the first five natural numbers. *Journal of Experimental Child Psychology*, Vol. 18, No. 3, pp. 520–534, 1974.

[15] Karen Wynn. Children's understanding of counting. *Cognition*, Vol. 36, No. 2, pp. 155–193, 1990.

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

[17] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

[18] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[19] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*, 2019.

[20] Ivana Kajic and Aida Nematzadeh. Probing representations of numbers in vision and language models. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.

[21] Satwik Bhattamishra, Arkil Patel, and Navin Goyal. On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286*, 2020.

[22] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[23] Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pretrained transformer-based language models. *arXiv preprint arXiv:2211.07349*, 2022.

[24] Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2862–2867, 2023.

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

[26] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

[28] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4185–4194, 2019.

[29] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21406–21415, 2022.

HeavyOrdiNum-L: 順序数を的確に把握し活用する言語モデルの能力を探る

林　寛治†　宮森　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054056,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，言語モデルが順序数の概念を的確に把握し活用する能力をどの程度有するのかについて調査する．Transformer に基づく事前学習済み言語モデルは，四則演算等の単純な算術問題では比較的高い正解率を出す一方，数の概念をどのように捉え，どのように活用しているのかについては不明な点も多い．本研究では，数の概念の一つである順序数に着目し，Transformer に基づく言語モデルが順序数の概念を把握し活用する能力について調査する．具体的には，順序数の的確な数え上げが必要となる質問応答のデータセット HeavyOrdiNum-L を構築する．このデータセットを用いた質問応答タスクを検証し，内部の挙動を調査することで，モデルによる順序数の概念の把握・活用能力を分析する．本データセットによる検証を通して，Transformer モデルの数の理解度に対する知見を深めることが期待される．

キーワード　順序数，数え上げ，概念理解，言語モデル，推論

1　はじめに

Transformer に基づく言語モデルでは，四則演算などの単純な算術問題に置いて高い正解率を示している [10]．一方で，扱う数の桁数が 15 桁や 20 桁と増えるにつれて正解率が低下することも示されており，限界があることも明らかとなっている. [9] しかし，数の概念をどのように捉えているのかについては不明であり，高い正解率を示している単純な算術問題においても数の概念を理解した上で解いているかどうかは不明である．

一般的に自然数の概念は，順序数，基数の 2 種類で捉えることができるとされている [1]．順序数は順序を表す数であり，1 番目，2 番目，3 番目... のような数え上げにより何番目かを表す．基数は個数として捉えられる数であり，ものの集まりの大きさを表す集合数とも呼ばれる．人間は，量を表す奇数よりも順序を表す順序数をより早い時期に学習することが指摘されており，順序数は数の概念において基数よりも根源的な概念と考えられる [2] [3] [11].

質問応答に関する従来研究では基数に関するものが多く，順序数に関する問題はほとんど取り組まれていない．基数は問題文中に出現する数値をそのまま使用することで計算可能である場合がほとんどであるが，順序数はそうではなく，数え上げや基数に変換するという処理が必要とされることが多く，順序情報を的確に把握していなければ順序数に関する問題は解くことができない．そのため，順序数に関する問題を解くことができれば，自然数を何らかの実体と対応づけながら数え上げることができている可能性があり，順序数の概念を人間と近い形で理解しているのかどうかに迫るための新たな知見が得られるのではないかと考えた．

上記を踏まえ，本稿では Transformer に基づく言語モデルが，順序数の概念を的確に把握し活用する能力がどの程度あるのかを明らかにすることを目的としたプロービングタスクを提案し，そのためのデータセット HeavyOrdiNum-L を新たに構築する．具体的には順序数を用いており，正答するために数え上げや基数への変換が必要とされるような質問を与える．これにより言語モデルは数値推論問題において数概念を理解して解いているのかが明らかになり，今後の数値推論問題解決のアプローチや言語モデルの数の理解度に関する考えの手がかりとなることが期待される．

本稿では，順序数を用いた 1 つの数え上げを遂行するためには，数え上げ対象，数え上げの起点及び向きをそれぞれ把握する必要があると考える．この 3 つを把握することを数え上げスキル 1 とし，これらを組み合わせることで難易度の異なる数え上げ能力を検証する．

実験では，言語モデルとして Flan-T5 [4] を対象に数え上げスキル，数え上げ対象，起点，向きの違いによる正解率の調査と Transformer 内部の挙動の変化などを調査した．

本稿の貢献は以下のとおりである．

(1) Transformer に基づくモデルが，順序数の概念をどの程度的確に把握し活用できるかを調査するための質問応答タスクを提案し，分析するためのデータセットを構築した．

(2) 構築したデータセットを用いた質問応答タスクを実施し，Transformer に基づくモデルに対する推論能力を明らかにした．

(3) 順序数を用いた数え上げの能力について，内部の挙動を調査することでモデルによる順序数の概念の把握・活用能力を分析した．

本稿の構成は以下の通りである．2 節では関連研究について述べ，3 節では提案手法について説明する．4 節では構築したデータセットで実験を行い，その結果を分析，考察する．5 節

では結論と今後の課題をまとめる．

2 関連研究

2.1 数値推論に関する研究

言語モデルには数値推論など，数値を扱う能力があることはこれまでの研究で知られている．また，比較的難易度の低い単純な問題に関して高い精度を示しているため，現在はより難易度の高い問題の解決に向けた研究が広く取り行われている．Hendrycks ら [5] は代数，数論，確率，幾何学，中級代数，微分積分の 6 分野からなる問題 12,500 問を含んだデータセットである “MATH データセット” を提案している．これは数学オリンピックなどで扱われる難易度の高い問題が主となっている．GPT-2 と GPT-3 に対してこれらの問題を与えた結果，最も簡単な問題 (レベル 1) で約 15%，難しい問題 (レベル 5) で約 4% と低い結果となっていることが示されており，難易度の高い問題に関しては未だ未解決とされている．

このように難易度の高い問題の解決に向けた取り組みがなされているが，たとえ高い精度を示しているモデルであっても数値を理解して解いているのかという点については不明であり，疑問視している研究も存在する．Arkil ら [9] は現時点で言語モデルは初等数学の文章問題ですら解くことができていないのではないかと述べている．文章問題は世界状態を説明する文章とその状態に関する未知の情報を問う文章から構成されており，前者の情報を用いて後者の情報を求めるという動作が求められる．しかし，現在高い精度を示しているモデルはこのような文章問題において質問文が削除されていても 77.7%の問題は正解してしまうということが示された．このことから，現在数値推論を要する文章問題において高い精度を示している言語モデルは，単純なヒューリスティックに依存しており，数値を理解して解いているとは言えないのではないかと述べられている．

数値の概念理解に関する研究やモデルの内部状態を分析している研究も存在する．Naik らの研究 [7] では単語埋め込み (skipgram，FastText，GloVe) が数値の性質，中でも大きさの概念をどの程度捉えているかを検証している．単語埋め込みが大きさを表現できるかどうかの実験を行った結果，数の大きさの概念をおおよそ捉えることができていることが示された．しかし,「3=three」などの数詞の概念については不十分であり，数を正確に扱うためには注意が必要であると述べられている．青木らの研究 [12] では，内容に応じて層数を変化させる PonderNet モデルに対して多段の推論が必要とされるタスクを与え，モデルの動作を確認し，層数が一定な Transformer モデルとの比較を行っている．実験の結果，簡単な演算に対する正解率は 2 つのモデルで大きな差は確認できなかった．しかし，PonderNet が問題に応じて再帰回数を変化させているということが示されている．

本稿では，順序数に着目した質問応答のための新たなデータセットを構築することで，Transformer モデルがどの程度順序数の概念を的確に把握し活用できるかについて調査する ．

2.2 Transformer の能力を調査した研究

松本らの研究 [14] では Transformer の再帰的構造を把握する能力を調査している．(a+b) × (c+d) のような四則演算を用いた問題を与えた際，Transformer はこの式における a+b や c+d のような計算の途中結果を内部的に保存して解いているのかどうかを調査した結果，Transformer は計算の途中結果を内部的に複数の成分に分散して保存していることが示されている．Johnson らの研究 [6] では数値に関する Transformer の構成的推論能力を評価することを目的としたプロービングタスクを提案している．英語，デンマーク語，日本語，フランス語の 4 つの言語で文法性判断タスクと大小比較の合計 2 種類の分類タスクを作成し，BERT，DistilBERT，XLM に与えている．その結果，2 つのタスク両方で 95%以上の高い精度を示し，モデルは英語，デンマーク語，日本語，フランス語において数詞の生成規則をある程度理解していることが示された．一方で大小比較タスクでは全体的に数の大きさを高精度で比較するには不十分であることが示された．このことから，数詞の文法性判断の能力は有しているが，値の比較の能力は不十分であり，数詞の構成要素がどのように意味を形成しているのかといった内容の理解は難しいことが明らかになった．Rodrigo らの研究 [8] では数の表面的な状態が Transformer の算術問題の学習にどのような影響を与えるかを調査している．足し算と引き算の 2 つのタスクをモデルに与え，その回答を評価しており「12」のようなアラビア数字,「1 2」のように空白で区切られた数字,「0012」のように桁数を統一したもの,「0_1_2」のようにアンダースコアで区切ったもの,「twelve」のように文字で数字を表したもの，10 進数，指数で表現したもので比較した結果，10 進数と指数表記のものが最も高精度であることが示された．しかし，どの表記も 15 桁や 20 桁など桁数が増えるにつれて精度は低下することが示され，限界があることが明らかになった. [13] では単純な数式とその途中結果に着目することで，Transformer モデルが数学的能力を獲得し，複数ステップに及ぶ処理を行っているかを検証している．検途中結果の情報が符号化されている箇所を追跡 (Tracing) し，符号化されている箇所の状態を操作 (Manipulation) してモデルに対して因果的介入を行うという二つの実験を行った結果，内部表現の特定の方向が線形に近い形で途中結果を符号化していること，そしてそのような方向がモデルの推論結果に対して因果的にも関係していることを示した．

本稿では，順序数の的確な数え上げが必要となる質問応答のデータセットを新たに構築し，Transformer モデルがどの程度順序数の概念を的確に把握した上で段階的に活用できるかについて調査する．

3 提案手法

3.1 問題設定

本稿では，順序数に数え上げを要する問題 q をテキスト形式でモデルに入力し，その問題の答え a をテキスト形式で出力させるような質問応答タスクについて考える．

順序数による数え上げを要する問題は，1 つ以上の数え上げ

表 1: 各項目の条件と質問例

数え上げスキル数	質問例	数え上げ方法	数え上げ対象	起点の指定	向きの指定
1	2,47,434,535 のうち 2 番目の数字はなんですか	順序数数え上げ	所与	なし	なし
	2,47,434,535 のうち右から 2 番目の数字はなんですか	順序数数え上げ	所与	あり	あり
	自然数の中で 2 番目はなんですか	順序数数え上げ	要推論	なし	なし
	自然数の小さい数から数えて 2 番目はなんですか	順序数数え上げ	要推論	あり	あり
	数の並び 2,47,434,535 にはいくつありますか	基数数え上げ	所与		
2	2,47,434,535 のうち 2 番目と 3 番目の中で左の数字はなんですか	順序数数え上げ	所与	なし×なし	なし×なし
	2,47,434,535 のうち右から 2 番目と 3 番目の中で左の数字はなんですか	順序数数え上げ	所与	なし×あり	なし×あり
	2,47,434,535 のうち 2 番目から数えて右に 3 番目はなんですか	順序数数え上げ	所与	あり×なし	あり×なし
	2,47,434,535 のうち右から 2 番目から数えて左に 3 番目はなんですか	順序数数え上げ	所与	あり×あり	あり×あり
	自然数の 2 番目と 3 番目で小さい方はなんですか	順序数数え上げ	要推論	なし×なし	なし×なし
	自然数の小さい方から 2 番目と 3 番目で小さい方はなんですか	順序数数え上げ	要推論	なし×あり	なし×あり
	自然数の 2 番目から数えて後ろに 3 番目はなんですか	順序数数え上げ	要推論	あり×なし	あり×なし
	自然数の小さい方から 2 番目から数えて後ろに 3 番目はなんですか	順序数数え上げ	要推論	あり×あり	あり×あり
	2,47,434,535 のうち 3 番目までにはいくつありますか	基数数え上げ	所与	なし	なし
	2,47,434,535 のうち右から 3 番目までにはいくつありますか	基数数え上げ	所与	あり	あり
	自然数の 3 番目までいくつありますか	基数数え上げ	要推論	なし	なし
	自然数の小さい方から 3 番目までいくつありますか	基数数え上げ	要推論	あり	あり
3	2,47,434,535 のうち 2 番目から 4 番目の間で 2 番目は何ですか	順序数数え上げ	所与	なし×なし×なし	なし×なし×なし
	2,47,434,535 のうち左から 2 番目から 4 番目の間で 2 番目は何ですか	順序数数え上げ	所与	なし×なし×あり	なし×なし×あり
	2,47,434,535 のうち 2 番目と左から 4 番目の間で 2 何番目は何ですか	順序数数え上げ	所与	なし×あり×なし	なし×あり×なし
	2,47,434,535 のうち右から 2 番目と右から 4 番目の間で 2 何番目は何ですか	順序数数え上げ	所与	なし×あり×あり	なし×あり×あり
	2,47,434,535 のうち 2 番目と 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	所与	あり×なし×なし	あり×なし×なし
	2,47,434,535 のうち左から 2 番目と 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	所与	あり×なし×あり	あり×なし×あり
	2,47,434,535 のうち 2 番目と左から 4 番目の間で左から 2 番目は何ですか	順序数数え上げ	所与	あり×あり×なし	あり×あり×なし
	2,47,434,535 のうち右から 2 番目と右から 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	所与	あり×あり×あり	あり×あり×あり
	自然数の 2 番目と 4 番目の間で 2 番目は何ですか	順序数数え上げ	要推論	なし×なし×なし	なし×なし×なし
	自然数の左から 2 番目と 4 番目の間で 2 番目は何ですか	順序数数え上げ	要推論	なし×なし×あり	なし×なし×あり
	自然数の 2 番目と左から 4 番目の間で 2 番目は何ですか	順序数数え上げ	要推論	なし×あり×なし	なし×あり×なし
	自然数の左から 2 番目と左から 4 番目の間で 2 番目は何ですか	順序数数え上げ	要推論	なし×あり×あり	なし×あり×あり
	自然数の 2 番目と 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	要推論	あり×なし×なし	あり×なし×なし
	自然数の左から 2 番目と 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	要推論	あり×なし×あり	あり×なし×あり
	自然数の 2 番目と左から 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	要推論	あり×あり×なし	あり×あり×なし
	自然数の左から 2 番目と左から 4 番目の間で右から 2 番目は何ですか	順序数数え上げ	要推論	あり×あり×あり	あり×あり×あり
	2,47,434,535 のうち 2 番目と 5 番目の間にはいくつありますか	基数数え上げ	所与	なし×なし	なし×なし
	2,47,434,535 のうち左から 2 番目と 4 番目の間にはいくつありますか	基数数え上げ	所与	なし×あり	なし×あり
	2,47,434,535 のうち 2 番目と左から 5 番目の間にはいくつありますか	基数数え上げ	所与	あり×なし	あり×なし
	2,47,434,535 のうち右から 2 番目と右から 5 番目の間にはいくつありますか	基数数え上げ	所与	あり×あり	あり×あり
	自然数の 2 番目と 4 番目の間にはいくつありますか	基数数え上げ	要推論	なし×なし	なし×なし
	自然数の左から 2 番目と 4 番目の間にはいくつありますか	基数数え上げ	要推論	なし×あり	なし×あり
	自然数の 2 番目と左から 4 番目の間にはいくつありますか	基数数え上げ	要推論	あり×なし	あり×あり
	自然数の左から 2 番目と左から 4 番目の間にはいくつありますか	基数数え上げ	要推論	あり×あり	あり×あり

スキルを用いて解くことができると考える．ここで，1 つの数え上げスキルは，数え上げ対象，起点，向きの 3 つを把握する能力であると定義する．例えば,「2,47,434,535 のうち 2 番目の数字は何ですか」という問題は，1 つの数え上げスキルを必要とする問題であり，その数え上げ対象は 2,47,434,535 という文字列であり，起点は先頭，向きは正順である．これら 3 つの項目を適切に把握することができれば，当該数え上げを正しく遂行できる可能性が高いと考えられる．また，基数による数え上げを要する問題も用意した．基数による数え上げを要する問題も 1 つ以上の数え上げスキルを用いて解くことができると考えるが，順序数の場合とは異なり，1 つの数え上げスキルは，数え上げ対象を把握する能力であると定義する．例えば,「2,47,434,535 にはいくつありますか」という問題は，1 つの基数数え上げを必要とする問題であり，数え上げ対象の 2,47,434,535 という文字列のみを把握することができれば数え上げを正しく遂行できる可能性が高いと考えられる．

数え上げスキルの各項目についてより詳細に説明する．

- 数え上げ対象

その対象そのものが問題内で与えられているか (以後，所与と呼ぶ)，数え上げ対象を推論する必要があるか (以後，要推論と呼ぶ) の 2 つに分けて考える．「2,47,434,535 のうち 2 番目の数字は何ですか」というような問題では，数え上げを行う対象は 2,47,434,535 という数字の並びであり既に与えられており (所与)，数え上げる対象の個数は 4 個の並びとなっている．「自然数の中で 2 番目は何ですか」という問題では，数え上げる対象の数字の並びは与えられておらず，具体的な数の並びや位置関係を推論する必要がある (要推論)．

- 起点

問題内で具体的な指示が与えられているか否かによってそれぞれ 2 種類に分類した．“左から 3 番目” や “右から 3 番目” の

表 2: 途中結果の追跡対象の問題例と条件

スキル数	問題例	起点
1	Among the given numbers:3,5,7,9,2,what is the 4th number?	なし
	Among the given numbers:4,7,3,2,8,what is the 4th number from the right?	あり
2	Among the given numbers:2,7,1,5,8,what is the 2nd number to the right counting from the 3rd number?	なし
	Among the given numbers:6,8,1,7,4,what is the 2nd number to the left counting from the 3rd number from the right?	あり

ように，数え上げの起点が指定されている場合は，起点の指示ありに相当する．また，“2,47,434,535 の 2 番目” などの場合は，起点の指定なしに相当し，一般的に左端が数え上げの起点であることを把握する必要がある．

- 向き

起点同様具体的な指示が与えられているか否かによってそれぞれ 2 種類に分類した．

3.2 データセット

順序数による数え上げを要する問題とその答えから構成される質問応答データセットを構築する．3.1 節で説明した数え上げ対象，起点，向きを 1 つの数え上げスキルと定義し，最大 3 回の数え上げスキルが必要な問題を難易度別に作成した．ここで，難易度は数え上げ回数が増えると増加すると考えることとし，本稿では 1 から 3 までの 3 段階を想定する．また，順序数を用いて基数を算出させるような問題も同様の視点で作成した．

表 1 に条件ごとに作成した質問例を示す．合計で 41 種類の問題となり，必要な数え上げスキル数が 1 の問題については 5 種類，数え上げスキル数が 2 の問題については 12 種類，数え上げスキル数が 3 の問題については 24 種類それぞれ数え上げ対象，起点，向きの指定の有無でそれぞれ分けて作成した．このような問題の答えを常に正しく出力するには，単純な足し算引き算の計算だけでなく，順序数や基数の概念を的確に把握し，数え上げにおいてそれらを正しく適用できる能力が必要となる．

4 実　　験

4.1 実験目的

本実験では Transformer に基づくモデルが，順序数の概念をどの程度的確に把握し活用する能力があるのかを明らかにすることを目的とする．

4.2 実験設定

実験では，構築したデータセットを用いて，Flan-T5 モデルを対象に，本タスクでの正解率を調査し，数え上げの途中結果を追跡する．Flan-T5 は，言語モデル T5 を 1000 以上のタスクで追加学習し強化したモデルである．Flan-T5 を HeavyOrdiNum-L で追加学習を行う．表 3 に示すようにそれぞれのモデルサイズでバッチサイズ 16，50 エポックの学習を行う．

表 3: モデルのパラメータ数とハイパーパラメータ

モデルサイズ	パラメータ数	バッチサイズ	エポック
small	80M	16	50
base	250M	16	50
large	780M	16	50
XL	3B	16	50
XXL	11B	16	50

4.2.1 正解率の調査

正解率の調査については以下の観点で調査を行う．

(1) 数え上げスキル数の違いによる正解率

(2) 数え上げ方法の違いによる正解率

(3) 数え上げ対象の違いによる正解率

4.2.2 内部状態の調査

[13] では数学的能力を獲得しているのであれば数式中の途中結果は表現空間になんらかの形で符号化されているはずという仮定のもとで一番単純な符号化の形として，表現空間中の特定の方向と途中結果の関係を調べている．この手法を用いて実験では途中結果の追跡を行う．モデルの l 層目 j 番目のトークンにおける表現 h_j^l を全て連結した表現 $H^l = h_1^l \oplus h_2^l \oplus \cdots \oplus h_n^l$ に対して主成分分析を行うことで主成分 $p_k^l, k \in [1, ..., K]$ を得る．データセット中のインスタンス i における文章の途中結果 R_i^j と各主成分 $p_{i,k}^l$ の重みの相関 $corr(R_i^j, p_{i,k}^l)$ を測る．最大相関方向: $\hat{p}_k^l(R_i^j) := argmax_k(corr(R_i^j, p_{i,k}^l))$ の相関が高ければ，対応する途中結果を符号化しているといえる．

実験では表 2 のように 4 つの問題を対象に追跡を行う．問題は桁数は 1 桁，数の並びは 5 つで統一しそれぞれ 1000 問使用し実験を行った．また，以下の観点で調査を行う．

(1) 数え上げスキル数の違い

(2) 起点の有無による違い

4.3 実験結果

4.3.1 正解率の調査

まず，スキル数の違いによる正解率を比較した結果を図 1 に示す．スキル 1 から 3 にかけて数え上げ回数が増加するごとに正解率が減少した．

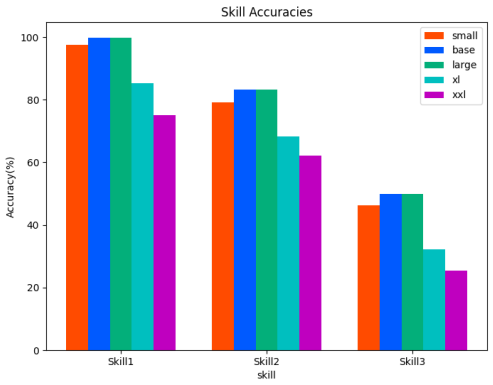


図 1: スキル数の違いによる正解率

次に数え上げ方法の違いによる正解率を比較した結果を図 2 に示す．基数数え上げの場合に順序数数え上げの場合に比べ，正解率が高い傾向が見られた．

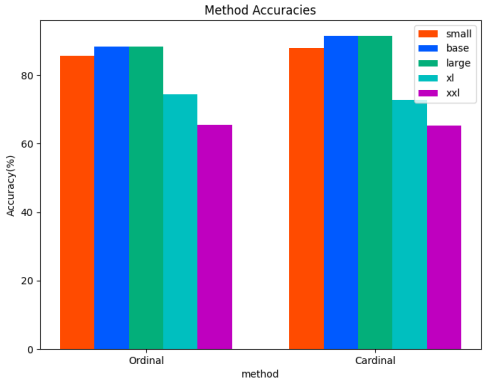


図 2: 数え上げ方法の違いによる正解率

次に数え上げ対象の違いによる正解率を比較した結果を図 3 に示す．所与の場合の正解率は xl,xxl の場合を除き正解率は 100%に近く，要推論の場合どのパラメータの場合でも約 75%となった．また，全体の傾向として large までパラメータ数が増えるごとに正解率は増加しているが，xl,xxl にかけてパラメータ数が増えるごとに正解率は small よりも減少している．

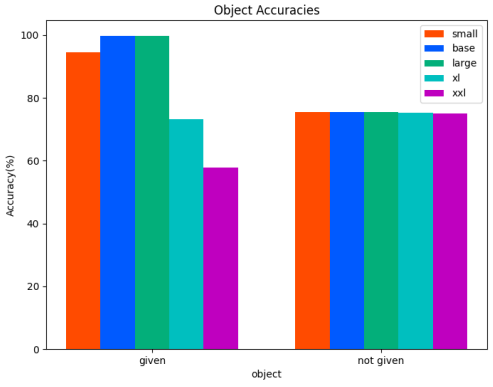


図 3: 数え上げ対象の違いによる正解率

4.3.2 内部状態の調査

スキル 1 の場合の途中結果の追跡のヒートマップを図 4 に示す．図 4a，4b では同じ問題を対象に順序数を変更した場合の比較を行う．図 4c は起点ありの場合のヒートマップである．図 4a，4b から順序数を変更した場合でも起点となる 1 番目が 1 7 層目で最も相関係数が高くなった．図 4c では右から数えるため起点は 5 番目となり左に数える場合である．5 番目から順に数える場合でも起点となる 5 番目が最も相関係数が高くなった．

次にスキル 2 の場合の途中結果の追跡のヒートマップを図 5 に示す．図 5a，図 5b はそれぞれ起点なし，ありの場合の問題である．図 5a では 3 番目から数えて 2 番目, 図**??**では右から 3 番目から数えて 2 番目としている．スキル 2 の場合もスキル 1 と同様に起点となる位置で 1 7 層目の第 6 主成分で相関係数が高くなっている．しかし，2 回目の数え上げの起点となる 3 番目では相関係数が高くなかった．

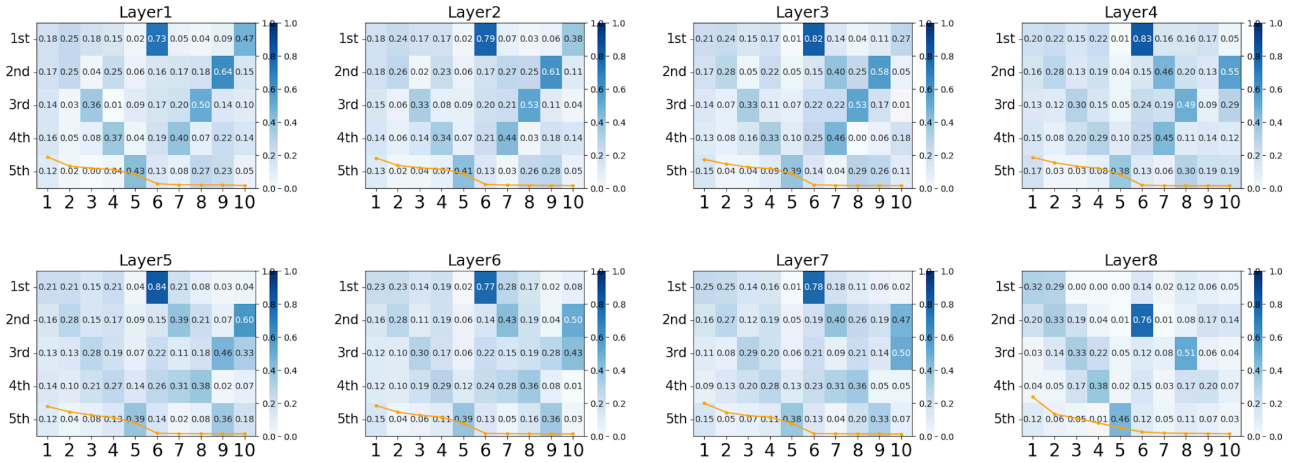
4.4 考　　察

数え上げスキル 1 では高い正解率を出したが，数え上げスキルが 2 以上になると正解率は減少することが明らかとなった．スキル 1 ではある程度数え上げを的確に行えているものの，スキル 2，3 のように複数の数え上げが必要になる場合には段階的な数え上げを適切に行うことは困難であるといえる．また，数え上げ方法の違いによる正解率は順序数による数え上げは基数数え上げに比べ，正解率が低い結果となった．しかし，モデルサイズが small，base，large に関しては数え上げスキル 1 の場合にほとんどの問題で正解していたことから，順序数，基数の概念は把握していると言える．しかし，複数スキル数による数え上げが含まれる場合は正解率が低くなっている．このことから数え上げが複数組み合わさる場合は，順序数や基数によらず正確な把握ができていないことがわかる．

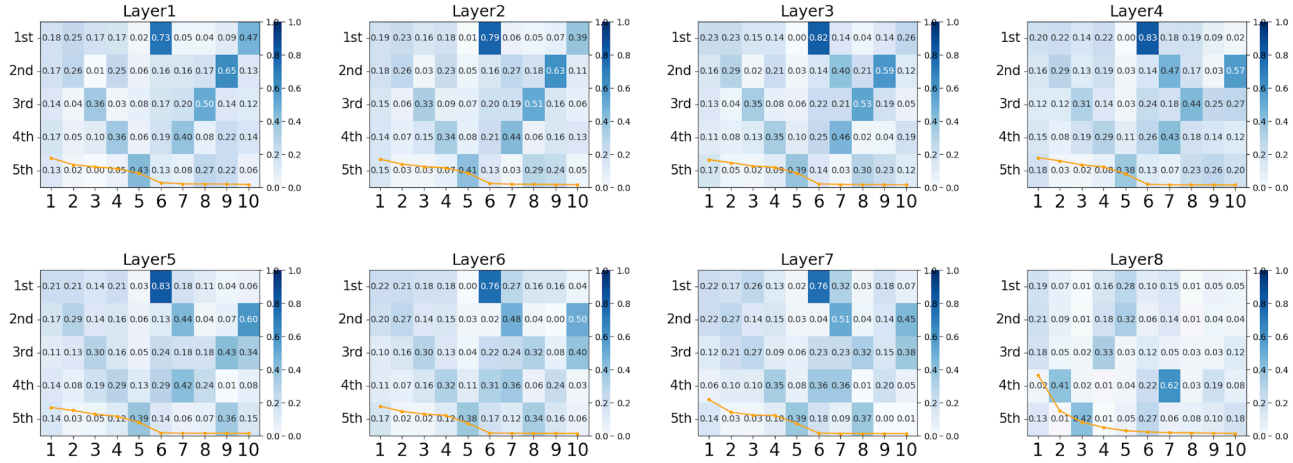
途中結果の追跡では数え上げスキル 1，2 ともに起点となる位置を途中結果として保存しており，数え上げスキル 2 では 2 回目の数え上げの起点を保存していなかったことから 1 回目の数え上げの起点となる位置から推論を行なっているのではないかと考えられる．複数数え上げの場合に正解率が低くなっていたことから，複数数え上げの場合，1 回目の数え上げで得た場所を起点として数え上げを行なっていないため正確に数え上げるという動作を行えていないのではないかと考えられる．

5 ま と め

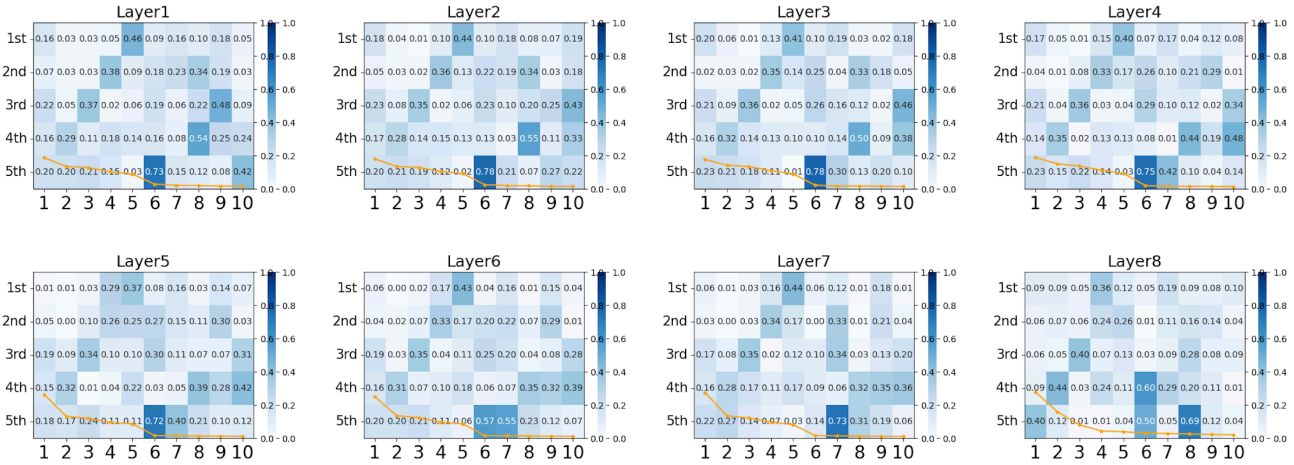
本稿では，順序数を用いた的確な数え上げが求められるようなプロービングタスクを提案し，そのためのデータセット HeavyOrdiNum-L を新たに構築し，言語モデルがどの程度順序数の概念を把握し活用できるかを調査した．実験では，構築したデータセットを用いて，言語モデルに対して質問応答タスクにより正解率を調査し，順序数に対する正確な理解と活用能力について分析した．その結果，数え上げスキル 1 では正確な数え上げを行うことができていたが，複数の数え上げが必要になる場合には段階的な数え上げを適切に行うことは困難である



(a) スキル 1 の 2 番目が答えの場合

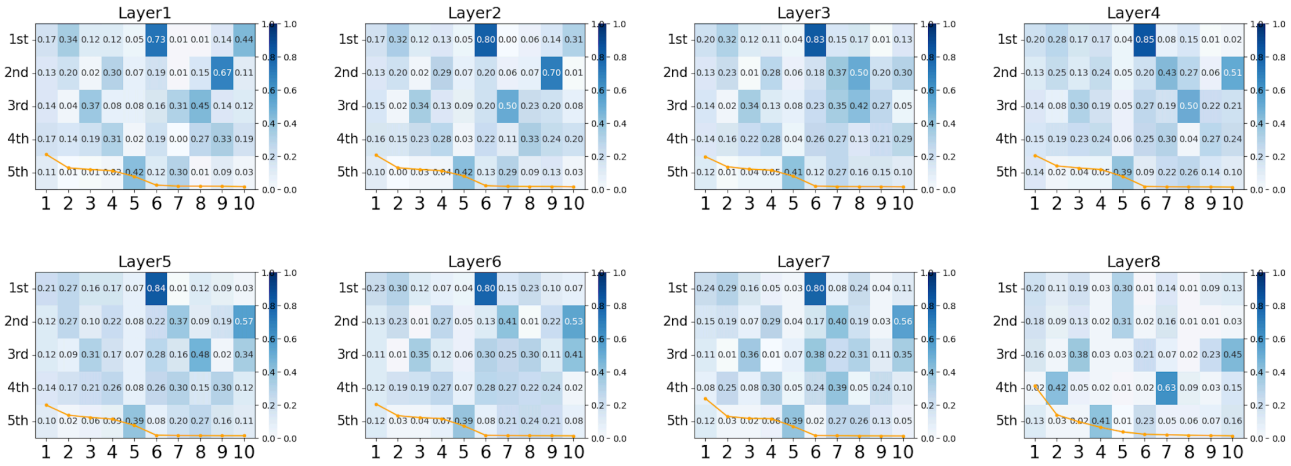


(b) スキル 1 の 4 番目が答えの場合

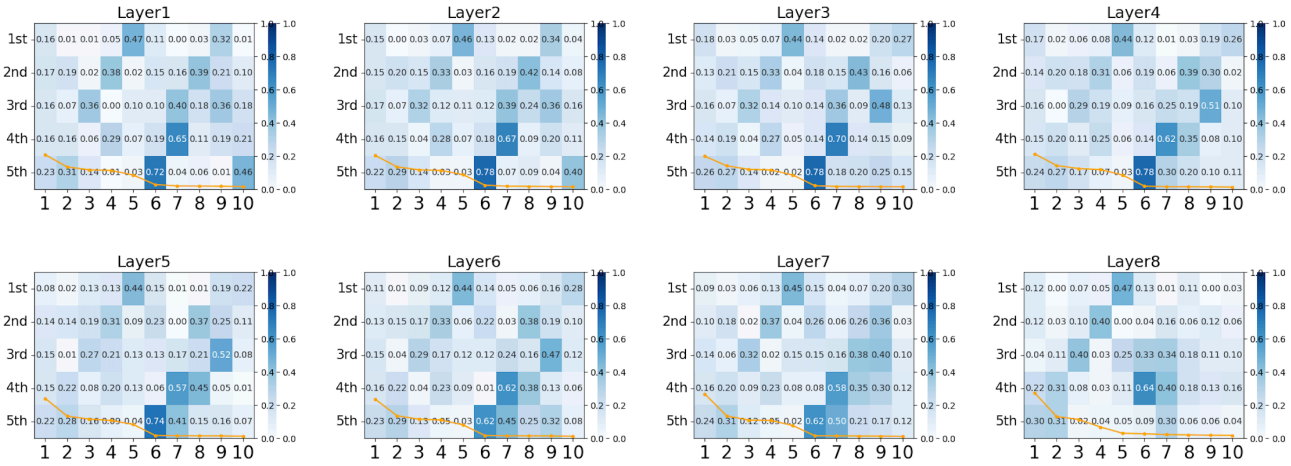


(c) スキル 1 の起点ありの場合

図 4: スキル 1 の場合の途中結果の相関関係のヒートマップ．各セルは第 k 主成分の重み (列) とある途中結果 (行) との相関関係の絶対値を表す.



(a) スキル 2 の起点なしの場合



(b) スキル 2 の起点ありの場合

図 5: スキル 2 の場合の途中結果の相関関係のヒートマップ．各セルは第 k 主成分の重み (列) とある途中結果 (行) との相関関係の絶対値を表す.

ことが示唆された．

また，途中結果の追跡では初めの数え上げとなる起点から答えを導き出していることが示唆された．複数数え上げの場合には，1 回目で数え上げて得た場所を起点として数え上げを行なっていないため正確に数え上げを行えていないことが示唆された．今後は複数の言語モデルでの比較，第 6 主成分に着目し途中結果の操作を行うことでより詳細な調査を行なっていく予定である．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] Bahram Assadian and Stefan Buijsman. Are the natural numbers fundamentally ordinals? *Philosophy and Phenomenological Research*, Vol. 99, No. 3, pp. 564–580, 2018.

[2] C. J. Brainerd. The origins of number concepts. *Scientific American*, Vol. 228, No. 3, pp. 101–109, 1973.

[3] Charles J Brainerd. Inducing ordinal and cardinal representations of the first five natural numbers. *Journal of Experimental Child Psychology*, Vol. 18, No. 3, pp. 520–534, December 1974.

[4] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[5] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.

[6] Devin Johnson, Denise Mak, Andrew Barker, and Lexi Loessberg-Zahl. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 184–192, Online, November 2020. Association for Computational Linguistics.

[7] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. Exploring numeracy in word embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3374–3380, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks, 2021.

[9] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics.

[10] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models, 2019.

[11] Karen Wynn. Children's understanding of counting. *Cognition*, Vol. 36, No. 2, pp. 155–193, August 1990.

[12] 青木洋一, 工藤慧音, Ana Brassard, 栗林樹生, 吉川将司, 乾健太郎. 多段の推量タスクに対する適応的なモデルの振る舞いの検証. 言語処理学会 第 28 回年次大会 発表論文, 2022.

[13] 松本悠太, Benjamin Heinzerling, 吉川将司, 乾健太郎. ニューラル数式ソルバーにおける途中結果の追跡と操作. 自然言語処理, Vol. 30, No. 4, pp. 1151–1171, 2023.

[14] 松本悠太, 吉川将司, Benjamin Heinzerling, 乾健太郎. 四則演算を用いた transformer の再帰的構造把握能力の調査. 言語処理学会 第 28 回年次大会 発表論文, 2022.

メルカリデータセットを利用した商品価格推定の一手法

森 優太郎 †　　上野 史 ††　　太田 学 ††

† 岡山大学工学部情報系学科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †p6tc09ek@s.okayama-u.ac.jp, ††{uwano,ohta}@okayama-u.ac.jp

あらまし　近年，オンラインの個人間取引が盛んに行われ，ユーザが自ら商品価格を決定する場面が増えている．株式会社メルカリは，その適正価格の推薦アルゴリズムの開発のため Mercari Price Suggestion Challenge というコンペティションを開催し，そこで Ali らの手法をはじめとした数々の手法が参加された．本稿では，同コンペティションにおいて二番目に精度の良い Chahhou らの手法を，国立情報学研究所の情報学研究データリポジトリが公開するメルカリデータセットに適用できるように日本語解析の処理を加え，さらに画像情報を利用できるように拡張した手法を提案する．実験では，メルカリデータセットに対する提案手法の価格推定精度と米国のメルカリデータに対する Chahhou らの手法の価格推定精度を比較し，メルカリデータセットでの価格推定精度を評価する．また，提案手法において画像を利用した場合と利用しなかった場合の価格推定精度を比較して，価格推定における画像情報の効果を検証する．

キーワード　機械学習，メルカリデータセット，ニューラルネットワーク，Ridge 回帰

1 はじめに

近年，オンライン上での個人間取引が盛んに行われており，ユーザが自ら商品の価格を決定する場面が増えている．それに伴い，商品の価格を自動で推定するための研究が行われている [1]．Ali ら [2] は商品名や商品説明文，カテゴリ，ブランド名，商品状態といった商品情報を利用して，多層パーセプトロンで商品の価格を自動で推定する手法を提案している．この手法は，国内最大のフリーマーケットアプリケーションを運営する株式会社メルカリ [1]が開催した Kaggle のコンペティションである Mercari Price Suggestion Challenge [1] で提案された．このコンペティションは，米国のメルカリの商品取引データ [2] (以下，米国のメルカリデータとする) を用いて，商品の価格を自動で推定するというものである．

本稿では，Mercari Price Suggestion Challenge で二番目に精度の高かった Chahhou らの手法 [3] を国立情報学研究所の情報学研究データリポジトリが公開する日本語の取引データであるメルカリデータセット [4] に適用できるように日本語処理を加えた．さらに，画像情報を利用できるように改良した手法を提案する．メルカリデータセットは，ファッションアイテム，本，家電，美容品など多岐にわたるカテゴリの商品約 1 億 3400 万件のデータを持つ大規模なデータセットである．これは，個人間の取引市場における価格推定において代表性の高いデータセットといえる．このデータセットに対して価格推定をする手法は Kaggle のデータセットであるファッションアイテムのデータセット [3]やスマートフォンのデータセット [4]のようなメルカリデータセットに含まれているカテゴリの他のデータセットにおける価格推定においても適用可能である．また，米国のメルカリデータと異なり，このメルカリデータセットには商品名や商品説明文といったテキスト情報に加えて，画像情報が含まれている点がある．このようにメルカリデータセットはマルチモーダルな情報を含んでおり，価格推定においてより多くの情報を利用することができる．そこで，商品価格の推定において画像情報を利用し，価格の推定における画像情報の有効性を検証する．

本稿の構成は以下の通りである．2 節では関連研究について述べる．3 節ではメルカリデータセットについて述べる．4 節では Chahhou らの手法と提案する価格推定手法について述べる．5 節では Chahhou らの手法に日本語処理を加えた提案手法の価格推定精度を評価する．また，提案手法に取り入れた画像情報の有効性を検証する．最後に 6 節で本稿をまとめる．

2 関連研究

2.1 価格の推定

価格の推定に関する研究は，これまでに様々な分野で行われている．例えば Bukvić ら [5] はクロアチアの市場における中古自動車の価格を推定している．メーカーや車種，走行距離などの特徴を用いて線形回帰，サポートベクトルマシン，ランダムフォレストの価格推定精度を比較した．Chen ら [6] は畳み込みニューラルネットワークを用いて画像情報のみから自動車の

1：https://jp.mercari.com
2：https://www.kaggle.com/c/mercari-price-suggestion-challenge/data
3：https://www.kaggle.com/datasets/shivamb/fashion-clothing-products-catalog
4：https://www.kaggle.com/datasets/ahsan81/used-handheld-device-data

表 1: メルカリデータセットの商品データの属性

No.	属性	No.	属性
1	出品状態	10	発送の方法
2	商品名	11	発送元の地域
3	商品の説明	12	発送にかかる日数
4	販売価格	13	いいね！の数
5	カテゴリ	14	コメントの数
6	商品の状態	15	出品日時
7	サイズ	16	更新日時
8	ブランド	17	コメント
9	送料の負担	18	商品画像

価格を推定する手法を提案した．また，画像のどの情報が車の価格に影響しているかを視覚化した．Han ら [7] は，ニューラルネットワークと勾配ブースティング手法である Light Gradient Boosting Machine (LightGBM) [8] と eXtreme Gradient Boosting (XGBoost) [9] を用いた 3 つのモデルを組み合わせて，メーカーや車種，走行距離などの特徴から自動車の価格を推定する手法を提案した．その手法は，それぞれのモデルを単一で使用したときの推定精度を上回った．

また，住宅の価格推定に関する研究も多い．Neloy ら [10] はバングラデシュの賃貸の価格を学校からの距離や築年数などの特徴を用いて XGBoost で推定する手法を提案した．You ら [11] は，住宅の画像情報と所在地を用いて，ニューラルネットワークで住宅の価格を推定する手法を提案した．

これらの研究では，メーカーや車種のようなカテゴリカルデータや築年数などを用いて，ランダムフォレストや線形回帰，LightGBM などを利用して推定した価格の精度向上を目指す研究が多かった．また，一般的なカテゴリカルデータに加えて画像情報や所在地のような地理情報などを用いて価格を推定する研究もある．一方本研究では，ファッションアイテムや子供用品といった日用品の複数のカテゴリの商品の価格を推定する．

2.2 Mercari Price Suggestion Challenge

個人間の商品取引市場における価格推定では，Mercari Price Suggestion Challenge [1] において多くの手法が提案された．Mercari Price Suggestion Challenge において価格推定の対象となる商品には，ブランド名やカテゴリ名のようなカテゴリカルデータに加えて商品名や商品説明文のようなテキスト情報がある．Mercari Price Suggestion Challenge は，このようなカテゴリカルデータやテキスト情報から商品の価格を推定しその精度を競うコンペティションである．Mercari Price Suggestion Challenge で Ali ら [2] は多層パーセプトロンを用いた手法を提案している．また，Quang ら [12] は，線形回帰，ランダムフォレスト，LightGBM を比較し，最も精度の高かった LightGBM を用いた手法を提案した．

3 メルカリデータセット

国立情報学研究所の情報学研究データリポジトリが公開するメルカリデータセット [4] は，株式会社メルカリが運営しているスマートフォン向けフリーマーケットアプリケーション「メルカリ」へ登録された商品データとその取引データからなる．このデータセットには 2020 年 1 月 1 日から 2020 年 12 月 31 日までにメルカリに出品された約 1 億 3400 万件の商品データが含まれ，これらはメルカリのアプリ上に公開されているもののみである．商品データの属性は表 1 の通りである．この属性のうち，本稿で利用する属性は，商品名，商品の説明，販売価格，カテゴリ，商品の状態，ブランド，送料の負担，商品画像である．

メルカリデータセットには，多様なカテゴリの商品が含まれている．このカテゴリは表 2 の通りで 13 カテゴリある．また，約 1ヶ月間 (4 月) のカテゴリごとの商品の出品数を図 1 に示す．商品の出品数が最も多いカテゴリは「本・音楽・ゲーム」カテゴリで 200 万件以上であった．また,「チケット」カテゴリや「自動車・オートバイ」カテゴリ以外のカテゴリにはそれぞれ 40 万件以上の商品数があった．ここで,「メンズ」,「レディース」カテゴリは男性のファッションアイテムと女性のファッションアイテムである.「その他」のカテゴリには食品や飲料，文房具などが含まれる．このように，メルカリデータセットには多様なカテゴリの商品が含まれている．

表 2: メルカリデータセットのカテゴリの種類

No.	カテゴリ名
1	本・音楽・ゲーム
2	おもちゃ・ホビー・グッズ
3	レディース
4	メンズ
5	ベビー・キッズ
6	コスメ・香水・美容
7	ハンドメイド
8	家電・スマホ・カメラ
9	スポーツ・レジャー
10	インテリア・住まい・小物
11	自動車・オートバイ
12	チケット
13	その他

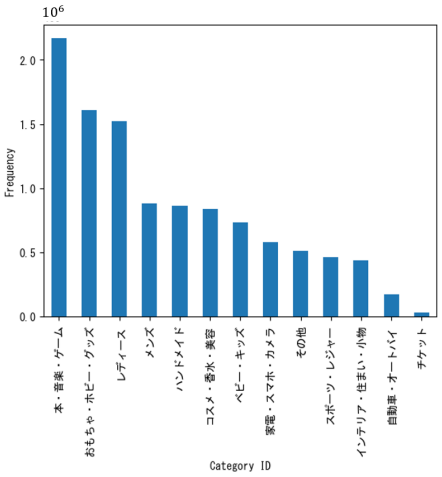


図 1: 約 1ヶ月間のカテゴリごとの商品数

表 3: 米国のメルカリデータの例

属性	具体例
name	Nike t-shirts
item_condition_id	2
category_name	Men/Tops/T-Shirts
brand_name	Nike
price	15 (ドル)
shipping	1
item_description	Worn one time. good condition

4 Chahhou らの手法と提案する価格推定手法

4.1 概　　要

本研究で提案する手法は，Mercari Price Suggestion Challenge で二番目に精度の良かった Chahhou らの手法を改良したものである．まず，ベースとなる Chahhou らの手法について説明する．そして，Chahhou らの手法を日本語の取引データであるメルカリデータセットへ適用するために日本語のテキスト処理を加え，さらに画像情報を利用できるように拡張した手法を提案する．

4.2 米国のメルカリデータと Chahhou らの手法

Chahhou らの手法は，米国のメルカリデータに対して価格推定をした．本節では，Mercari Price Suggestion Challenge で利用された米国のメルカリデータと Chahhou らの手法について説明する．

4.2.1 米国のメルカリデータ

米国のメルカリデータは，商品名，商品の状態，カテゴリ名，ブランド名，価格，送料の負担，商品説明文を属性にもつ．米国のメルカリデータの例として Nike の T シャツを表 3 に示す．name は商品名，item_condition_id は商品の状態を表している．商品の状態は 1 から 5 で表され，1 が新品を表す．category_name は商品の 3 階層からなるカテゴリを表している．表 3 の例では，Men が第 1 カテゴリとなっており男性のファッションアイテムを表し，第 2 カテゴリは Tops で，トップスを表す．表 3 の商品は Nike の T シャツであるため，第 2 カテゴリは Tops であるが，商品がジャケットであれば第 2 カテゴリは Jacket となる．brand_name は商品のブランド名である．price は商品の価格であり，単位はドルである．shipping は送料が出品者負担 (1) か購入者負担 (0) かを表す．item_description は商品説明文である．商品説明文には商品のより詳しい情報や商品状態が記載されている．

4.2.2 Chahhou らの手法

Chahhou らの手法による価格推定は図 2 の通りである．図 2 のように，この手法は Ridge モデルと 3 種類のニューラルネットワーク (NN) モデルで構成されている．この 3 種類のニューラルネットワークモデルとは Sparse NN モデル，Char Sparse NN モデル，fastText モデルである．入力の商品データは表 3 のようなデータである．この商品データを入力として Ridge モデル，Sparse NN モデル，Char Sparse NN モデル，fastText モデルの 4 つのモデルをそれぞれ別々に学習する．そして，学習した 4 つのモデルでそれぞれ商品の価格を推定する．最後に，各モデルの推定価格の平均を計算し，それを最終的な推定価格とする．

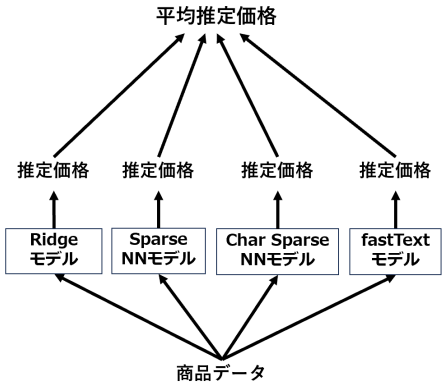


図 2: Chahhou らの手法による価格推定

表 4: 変換後の属性

属性名	属性の説明
name_old	商品名
item_condition_id	商品状態
category1	第 1 カテゴリ
category2	第 2 カテゴリ
category3	第 3 カテゴリ
category_name	第 1〜3 カテゴリを連結したもの
brand_name	ブランド名
brand_cat	category_name + brand_name
price	価格
shipping	送料の有無
nb_words	item_description の単語数
name	name_old + brand_cat
item_description	商品説明文
name_desc	name_old + item_description の先頭 5 単語

4.2.3 データの前処理

モデルへ入力するために表 3 のデータの前処理を行う．まず，商品名と商品説明文のクリーニングとして大文字を小文字に変換し，アルファベットと数字以外の文字を削除する．その後，先頭から 60 単語を上限として抽出する．また，item_condition と shipping の欠損値はそれぞれの最頻値で補完する．次に，いくつかの属性を各モデルで扱いやすいように変換し，さらに，新しい属性を 7 つ追加する．変換後の属性は表 4 の通りである．商品名である name は name_old に変換する．3 階層になっている category_name は category_1，category_2，category_3 に分解し，category_name と brand_name を結合して brand_cat を作成する．nb_word は item_description の単語数をカウントしたものである．また，name_old に brand_cat を結合し，name_old と item_description の先頭 5 単語を結合した name_desc を作成する．

表 5: Ridge 回帰の学習に用いる説明変数

説明変数	説明変数の詳細
item_condition_id	商品の状態
shipping	送料の有無
nb_words	item_description の単語数
mean_price_category	カテゴリごとの平均価格
mean_price_brand	ブランドごとの平均価格
name_one	name (unigram)
name_item	item_description (unigram)
name_bi	name (bigram)

4.2.4 Ridge モデル

Ridge モデルは，Ridge 回帰を用いて商品の価格を推定するモデルである．Ridge 回帰は，重回帰分析の損失関数に L2 正則化項を加えたものを損失関数として利用して，モデルの係数を推定する方法である．重回帰分析は以下の式 (1) で表される．

$$\hat{y} = w_0 + \sum_{j=1}^{m} w_j x_j \tag{1}$$

$\hat{y}$ は目的変数，x_j は説明変数，w_j は回帰係数，w_0 は切片，m は説明変数の数である．ここで，重回帰分析の損失関数に L2 正則化項を加えたものは以下の式 (2) で表される．

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + a\sum_{j=1}^{m} {w_j}^2 \tag{2}$$

y_i は真の値，n はデータ数を表す．また，a は正則化の強さを表すハイパーパラメータである．Ridge 回帰では式 (2) が最小となるような回帰係数を推定する．Ridge 回帰の学習に用いる説明変数は表 5 の通りである．カテゴリごとの平均価格とは，あるカテゴリの全ての商品の価格の平均である．例えば，ダウンジャケットと T シャツでは平均価格は大きく変わる．ブランドごとの平均価格についても同様である．name_one と name_item はそれぞれ name と item_description を Bag-of-Words (BoW) でベクトル化したものである．BoW は文章に出現する単語の数を数えてベクトル化する手法である．ここで，name_bi は name の単語 bigram を BoW でベクトル化したものである．例えば，単語の bigram の BoW では，“good condition” が “good” と “condition” に分解されず，文脈が維持できる．正則化の強さを表すハイパーパラメータ α は 20 である．

4.2.5 Sparse NN モデル

Sparse NN モデルは，入力データがスパースであるニューラルネットワークを用いて商品の価格を推定するモデルである．Sparse NN モデルでは，商品説明文のようなテキストデータを BoW でベクトル化しているため，入力データはほとんどの値がゼロであるスパースなベクトルとなる．Sparse NN モデルの学習に利用する特徴は item_condition_id，shipping，nb_words，mean_price_category，sparse_data である．sparse_data は表 4 の name と item_description を結合したテキストを BoW でベクトル化したものである．その際，単語の unigram と bigram をベクトル化している．Sparse NN モデルを図 3 に示す．連結層の前の sparse_data を入力とする全結合層の出力は 200 次元である．sparse_data と sparse_data 以外の特徴を連結層で結合している．出力層以外の全結合層の活性化関数は PReLU，バッチサイズは 2000，エポック数は 3 である．ここで，各エポックでバッチサイズを 2 倍に増やしながら学習をしている．始めのバッチサイズを小さくしておき，学習が進むにつれてバッチサイズを増やすことで，精度をほとんど落とすことなく計算時間が短縮できる [13]．最適関数には Adam を用いており学習率は 0.0011 である．また，損失関数は平均二乗誤差である．

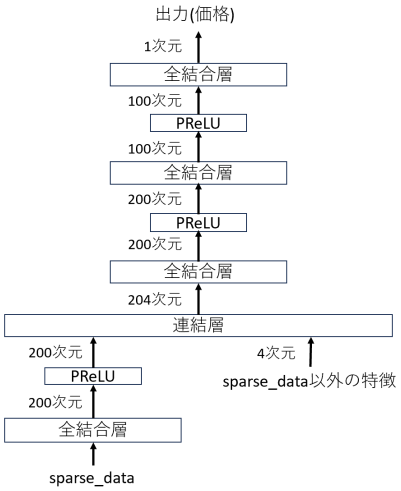


図 3: Sparse NN モデルの図

4.2.6 Char Sparse NN モデル

Char Sparse NN モデルは 4.2.5 項のモデルと同様に，入力データがスパースであるニューラルネットワークを用いて商品の価格を推定するモデルである．ここで，4.2.5 項の Sparse NN モデルとの違いは学習に利用する sparse_data である．4.2.5 項では name と item_description を結合したものから単語の unigram と bigram を求めて利用したが，Char Sparse NN モデルでは表 4 の name_old から N が 1 から 4 の範囲の文字の N-gram を求めている．例えば，name_old に “Nike” という単語があるとき，この単語は “N”，“i”，“k”，“e” の 4 文字に分割することができ，これが N が 1 のときの文字の N-gram である．N が 2 のときは “Ni”，“ik”，“ke” に分割される．N が 3，4 のときは文字数がさらに増える．さらに item_description から単語の unigram のみを取得し，brand_cat から 1 から 3 の範囲で単語の N-gram を取得する．これらの特徴を sparse_data として学習に利用している．他の特徴は 4.2.5 項の Sparse NN モデルと同じである．Char Sparse NN モデルを図 4 に示す．連結層の前の sparse_data を入力とする全結合層の出力は 100 次元である．sparse_data と，sparse_data 以外の特徴を連結層で結合している．出力層以外の全結合層の活性化関数は PReLU，バッチサイズは 2048，エポック数は 3 である．また，4.2.5 項と同様に各エポックでバッチサイズを 2 倍にしながら学習している．最適関数には Adam を用いており学習率は 0.002 である．損失関数は平均二乗誤差である．

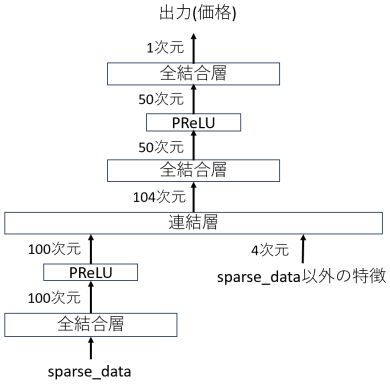


図 4: Char Sparse NN モデル

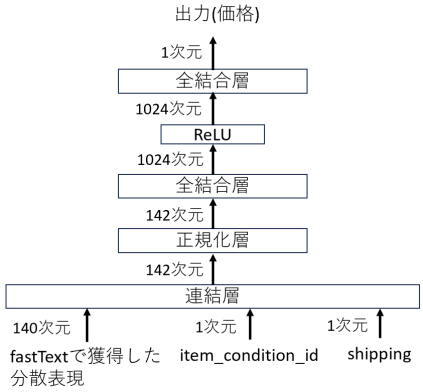


図 5: fastText モデル

4.2.7 fastText モデル

fastText は，Bojanowski らによって提案された単語の分散表現を獲得する手法である [14]．fastText モデルでは，fastText を用いて商品名や商品説明文のようなテキストデータの分散表現を獲得し，その分散表現を利用して商品の価格を推定している．学習に利用する特徴は表 4 の name，item_desc，brand_name，category1，category2，category3，item_condition_id，shipping である．これらの特徴のうち，name，item_desc，brand_name，category1〜3 の分散表現を fastText で獲得している．なお，name は先頭 20 単語，item_desc は先頭 30 単語の分散表現を取得している．name と item_desc は 50 次元，brand_name，category1〜3 はそれぞれ 10 次元の分散表現である．それらを合計すると 140 次元となる．

fastText モデルを図 5 に示す．連結層では，fastText で獲得した合計 140 次元の分散表現と item_condion_id と shipping を結合している．また，活性化関数は ReLU，バッチサイズは 128，エポック数は 4 である．また，4.2.5 項と同様に各エポックでバッチサイズを 2 倍にしながら学習している．最適関数には Adam を用いており学習率は 0.002 である．損失関数は平均二乗誤差である．

4.3 提 案 手 法

本節では，Chahhou らの手法をメルカリデータセットを用いた商品価格推定に適用できるように日本語のテキスト処理を加え，それに加えて商品の画像情報を利用できるようにした手法を提案する．本研究で扱うメルカリデータセットには，米国のメルカリデータにはない，商品の画像情報が含まれている．この商品画像は 1 つの商品につき 1 枚以上付与されているが，本研究ではサムネイル画像 1 枚のみを利用する．

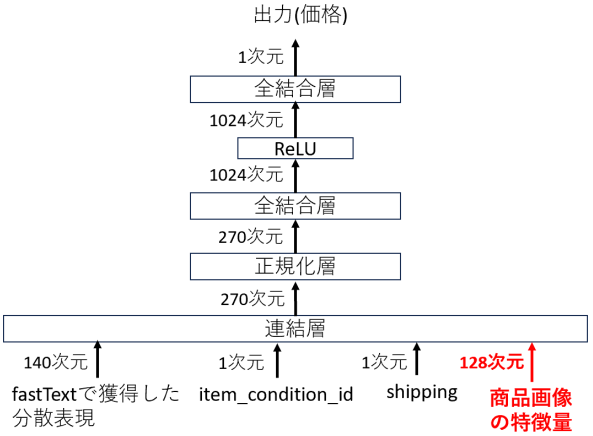


図 6: 画像情報を加えた拡張 fastText モデル

まず，日本語のテキスト処理について説明する．日本語のテキストを BoW でベクトル化するために，形態素解析により名詞，動詞，形容詞を抽出し，形容詞と動詞は基本形に変換する．形態素解析には MeCab[5]を，またその辞書には mecab-ipadic-NEologd[6]を用いる．この辞書は，2020 年 9 月に更新されたもので，新語や固有表現の形態素解析に強いという特徴がある．次に商品名と商品説明文をクリーニングとして，「？」や「！」のような記号や絵文字を削除し，アルファベットをすべて小文字に変換し，半角カタカナを全角カタカナに変換する．4.2.3 項の Chahhou らの手法のデータの前処理に上記の処理を加えることで日本語に対応する．

次に提案手法における商品の画像情報の利用について説明する．この手法では，図 5 で説明した fastText モデルの連結層で，fastText で獲得した分散表現，item_condition_id，shipping に加えて商品画像の特徴量を連結する．画像情報を加えた拡張 fastText モデルを図 6 に示す．なお，商品画像の特徴量は，ImageNet [15] の画像 100 万枚以上で事前学習された ResNet50 [16] を利用して獲得する．具体的には，商品画像を ResNet50 に入力し ResNet50 の中間層の 2048 次元の画像の特徴量を取得し，それを Principal Component Analysis (PCA) を用いて 128 次元に削減する．この 128 次元は，5.4 節の予備実験により定めた．

5 実験と考察

5.1 実 験 内 容

実験では，Chahhou らの手法で米国のメルカリデータ (英語) の商品価格推定をした精度と，提案手法でメルカリデータセット (日本語) の商品価格推定をした精度を評価した．提案

5：https://taku910.github.io/mecab
6：https://github.com/neologd/mecab-ipadic-neologd

手法において，Chahhou らの手法に日本語処理を加えたものを提案手法 (画像なし)，それに加えて画像情報を利用できるように拡張したものを提案手法 (画像あり) とする．実験では，提案手法の評価にメルカリデータセットから約 1ヶ月間 (4/1 〜 4/30) の商品データ約 940 万件を抽出 (日本のメルカリデータとする) して利用する．Chahhou らの手法の評価には米国のメルカリデータセットを利用する．

5.3 節では，Chahhou らの手法の米国のメルカリデータに対する価格推定と提案手法 (画像なし) の日本のメルカリデータに対する価格推定の実験をする．本実験では米国のメルカリデータと日本のメルカリデータでカテゴリが一致しているファッションアイテムとベビー・キッズ用品の 2 カテゴリの商品データを利用する．米国のメルカリデータにおけるメンズ，レディースを合わせたファッションアイテムの商品数は 758,065 件である．一方，日本のメルカリデータのそれは 2,407,735 件あるため，ランダムに 758,065 件抽出した．米国のメルカリデータにおけるベビー・キッズ用品の商品数は 172,471 件である．一方，日本のメルカリデータのそれは 734,610 件であるため，ランダムに 172,471 件抽出し商品数を揃えた．また，米国のメルカリデータの価格の単位はドルであるため，コンペティションが開催された 2018 年のおおよその平均レートである 1 ドル 110 円で円に変換して単位を揃える．そして，それぞれのデータの 8 割を学習データ，2 割をテストデータとして利用する．データはランダムに分割するが，学習データとテストデータの偏りを減らすためブランド名の分布が同じになるように分割する．

5.4 節では，提案手法 (画像あり) における画像の特徴量の次元数を決めるための予備実験について説明する．この予備実験では日本のメルカリデータのベビー・キッズカテゴリの商品 30,000 件を利用する．この 30,000 件のデータをランダムに 9：1 に分割し 9 割を学習データと 1 割をテストデータとし，図 6 の拡張 fastText モデルの画像の特徴量の次元数を変更して価格推定精度を評価する．学習データとテストデータの割合を 5.3 節の実験と変更する理由は，総データ数が少ないためである．データはランダムに分割するが，学習データとテストデータの偏りを減らすためブランド名の分布が同じになるように分割する．

5.5 節では，提案手法 (画像あり) を利用し，画像情報を加えた効果を検証する実験について説明する．この実験では，日本のメルカリデータのベビー・キッズ用品 90,000 件を利用する．また，米国のメルカリデータのベビー・キッズ用品 90,000 件に対する Chahhou らの手法の推定結果も評価する．学習データとテストデータの分割方法は 5.4 節と同様である．

5.2 評価指標

価格の推定精度の評価指標として，Mercari Price Suggestion Challenge で用いられた評価指標である Root Mean Squared Logarithmic Error (RMSLE) を利用する．RMSLE の計算式を式 (3) に示す．

表 6: ファッションアイテムに対する各手法の RMSLE の比較

手法名	RMSLE
Chahhou らの手法	0.3911
提案手法 (画像なし)	0.5180

表 7: ベビー・キッズ用品に対する各手法の RMSLE の比較

手法名	RMSLE
Chahhou らの手法	0.4385
提案手法 (画像なし)	0.4493

$$\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2} \qquad (3)$$

ここで，n は商品データの総数，p_i は商品 i の推定価格，a_i は商品 i の実際の価格である．また，価格の単位は円，log の底は e である．RMSLE は値が小さいほど精度が良いことを表す．

5.3 Chahhou らの手法と提案手法 (画像なし) の推定精度の実験と考察

米国のメルカリデータのファッションアイテム約 76 万件における Chahhou らの手法と日本のメルカリデータのファッションアイテム約 76 万件における提案手法 (画像なし) の RMSLE を表 6 にまとめる．表 6 に示す通り，Chahhou らの手法の RMSLE は 0.3911，提案手法 (画像なし) の RMSLE は 0.5180 であり，Chahhou らの手法に比べて提案手法 (画像なし) の精度が低いことが分かる．

また，米国のメルカリデータのベビー・キッズ用品約 17 万件における Chahhou らの手法と日本のメルカリデータのベビー・キッズ用品約 17 万件における提案手法 (画像なし) の RMSLE を表 7 にまとめる．表 7 に示す通り，Chahhou らの手法の RMSLE は 0.4385，提案手法 (画像なし) の RMSLE は 0.4493 となった．表 7 においても Chahhou らの手法に比べて提案手法 (画像なし) の精度が低いが，表 6 のファッションアイテムほどの差はない．また，図 7 に米国のメルカリデータのベビー・キッズ用品のテストデータにおける Chahhou らの手法の実際の価格と推定価格の散布図を，図 8 に日本のメルカリデータのベビー・キッズ用品のテストデータにおける提案手法 (画像なし) の実際の価格と推定価格の散布図を示す．図 7，8 を見ると，どちらも 10,000 円以下の商品がほとんどである．また，実際の価格よりも低く価格を推定している商品が多いことが分かる．

日本のメルカリデータにおける提案手法 (画像なし) の価格推定精度が Chahhou らの手法に比べて低くなった原因として提案手法 (画像なし) の日本語のテキスト処理の分かち書きがある．分かち書きには，mecab-ipadic-NEologd の辞書を用いた MeCab を利用したが，うまく分かち書きできていない「ナイキランニング」や「ナイキナイロン」のようなテキストがあった．したがって，分かち書きの失敗が推定精度に悪影響を与えている可能性がある．

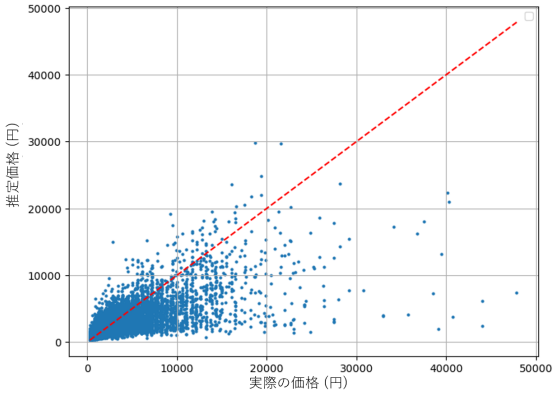


図 7: ベビー・キッズ用品における Chahhou らの手法の実際の価格と推定価格

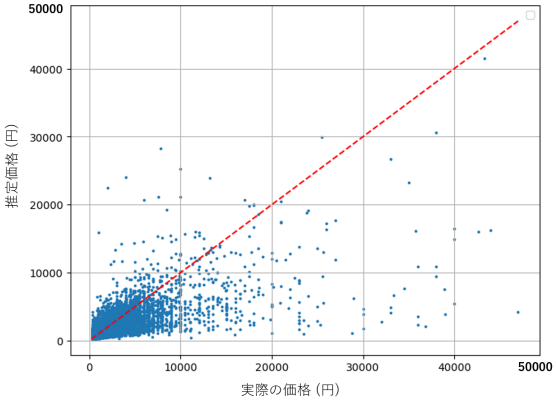


図 8: ベビー・キッズ用品における提案手法 (画像なし) の実際の価格と推定価格

表 8: 画像特徴量の次元数を削減した時の RMSLE

	2048 次元	512 次元	256 次元	128 次元	64 次元
RMSLE	0.6530	0.6446	0.6439	0.6424	0.6492

5.4 画像特徴量の次元数を決めるための予備実験

図 6 の画像情報を加えた拡張 fastText モデルにおける商品画像の特徴量の次元数を決めるために予備実験をする．本研究では，ResNet50 の中間層の 2048 次元の画像の特徴量を取得するが，2048 次元では次元数が大きすぎるため PCA を用いて次元を削減する．削減後の適当な次元数を見積もるために日本のメルカリデータのベビー・キッズ用品 30,000 件を利用して，削減後の次元数を 512 次元，256 次元，128 次元，64 次元として RMSLE を算出した．その結果を表 8 にまとめる．表 8 より画像の特徴量を 128 次元にしたときに RMSLE が 0.6424 で最小となったため，図 6 の画像の特徴量の次元数を 128 とした．

5.5 画像情報の効果の検証のための実験と考察

提案手法 (画像なし) と提案手法 (画像あり) のベビー・キッズ用品 90,000 件における RMSLE を表 9 にまとめる．この表に示す通り，日本のメルカリデータのベビー・キッズ用品に対する提案手法 (画像なし) の RMSLE は 0.4943，提案手法 (画像あり) の RMSLE は 0.4846 となっており，画像情報を利用

表 9: ベビー・キッズ用品における各手法の RMSLE

手法名	RMSLE
Chahhou らの手法	0.4621
提案手法 (画像なし)	0.4943
提案手法 (画像あり)	0.4846



(a) 精度が良くなった商品



(b) 精度が悪くなった商品

図 9: 精度が良くなった商品と悪くなった商品の商品画像の例 (画像はメルカリデータセットのもの)

することで価格の推定精度が良くなっていることが分かる．また，同じカテゴリの米国のメルカリデータ 90,000 件に対する Chahhou らの手法の RMSLE は 0.4621 であった．表 7 の実験と比べてデータ数が少ないため，Chahhou らの手法と提案手法 (画像なし) の精度はともに表 7 より悪化している．

画像情報を加えたことで価格推定精度が向上した商品と精度が悪くなった商品の商品画像の例を図 9 に示す．これらの画像を比べると，図 9(a) は背景がシンプルであるが，図 9(b) は背景が柄ものであり，図 9(a) と比べて複雑である．このように画像情報を加えることで価格推定精度が悪くなる商品もある．

6 ま と め

本稿では，Mercari Price Suggestion Challenge で二番目に商品価格推定精度が高かった Chahhou らの手法に日本語処理を加え，さらに画像情報を利用できるように拡張した手法を提案した．日本語処理として日本語のテキストを MeCab を利用して形態素解析し，分かち書きした．画像情報は，ResNet50 を用いて取得した商品画像の特徴量を PCA で次元削減して利用した．実験では，Chahhou らの手法の米国のメルカリデータ (英語) における価格推定精度と提案手法のメルカリデータセット (日本語) における価格推定精度を評価した．また，価格推定における画像情報の影響を検証した．Chahhou らの手法と提案手法 (画像情報は用いない) では，ファッションアイテムとベビー・キッズ用品のデータにおいて価格推定した結果，両カテゴリにおいて提案手法の価格の推定精度が Chahhou らの手法より低かった．この理由には分かち書きの失敗が影響している可能性がある．また，価格推定における画像情報の影響は，メルカリデータセットのベビー・キッズ用品に対する価格推定精度を比較した結果，画像情報を利用した場合の RMSLE が 0.4846，画像情報を利用しなかった場合の RMSLE が 0.4943 となり，画像情報を利用することで推定精度が向上した．しか

し，背景が複雑な商品画像の場合，価格推定精度が悪化することがあった．

今後の課題として，本稿の日本語処理の妥当性や取り入れた画像情報の価格推定における効果について，更に検討する必要がある．

謝　　辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより株式会社メルカリから提供を受けた「メルカリデータセット」を利用した．ここに記して感謝する．

文　　献

[1] Mercari Price Suggestion Challenge, 2018. `https://www.kaggle.com/competitions/mercari-price-suggestion-challenge`.

[2] Amine Ait Si Ali, Huseyin Seker, Steven Farnie, and John Elliott. Extensive data exploration for automatic price suggestion using item description: Case study for the kaggle mercari challenge. ICAAI '18, pages 41–45, 2018.

[3] Mercari Price Suggestion Challenge, 2018. `https://www.kaggle.com/code/mchahhou/mercari-second-place-solution`.

[4] 株式会社メルカリ. メルカリデータセット, sep 2022.

[5] Lucija Bukvić, Jasmina Škrinjar, Tomislav Fratrović, and Borna Abramović. Price prediction and classification of used-vehicles using supervised machine learning. *Sustainability*, 2022.

[6] Steven Chen, Edward Chou, and Richard R. Yang. The price is right: Predicting prices with product images. *CoRR*, abs/1803.11227, 2018.

[7] Shengqiang Han, Jianhua Qu, Jinyi Song, and Zijing Liu. Second-hand car price prediction based on a mixed-weighted regression model. In *Proceedings of the 2022 7th International Conference on Big Data Analytics (ICBDA)*, pages 90–95, 2022.

[8] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the Neural Information Processing Systems*, 2017.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, 2016.

[10] Asif Ahmed Neloy, H. M. Sadman Haque, and Md. Mahmud Ul Islam. Ensemble learning based rental apartment price prediction model by categorical features factoring. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, ICMLC '19, page 350–356, 2019.

[11] Quanzeng You, Ran Pang, Liangliang Cao, and Jiebo Luo. Image-based appraisal of real estate properties. *IEEE Transactions on Multimedia*, pages 2751–2759, 2016.

[12] Linh Nguyen Tran Quang, Uyen Ngo Thi Thao, Anh Duong Vo Nhi, and Hoang Nguyen Thi Phuong. E-commerce price suggestion algorithm – a machine learning application. 2022.

[13] Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. Don't decay the learning rate, increase the batch size. *CoRR*, abs/1711.00489, 2017.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

DEIM Forum 2024 XX-Y

Jellyfish：データ前処理のための大規模言語モデル

張　皓辰† 董　于洋†† 肖　川† 小山田昌史††

† 大阪大学情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
†† NEC　〒 211–0011 神奈川県川崎市中原区下沼部 1753
E-mail: †{chou.koushin,chuanx}@ist.osaka-u.ac.jp, ††{dongyuyang,oyamada}@nec.com

あらまし　データ前処理（DP）は，生データを処理しやすく，よりクリーンなデータに変換する重要なステップであり，エラーの検出やエンティティマッチングなど，多岐にわたるタスクが含まれます．従来の DP の解決策は，一つまたは二つの特定のタスクに特化していました．しかし，大規模言語モデル（LLM）の時代の到来により，自然言語プロンプトに応答することで，幅広い DP タスクに適用可能な汎用ソリューションに対する関心が高まっています．ただし，既存の LLM を利用する手法は，OpenAI の GPT API に依存しているため，データを外部に送信する必要があり，セキュリティ上の懸念が生じます．この問題に対処するため，本研究ではローカルで運用可能で，高性能な DP 専用 LLM である「Jellyfish」を開発しました．Jellyfish は，Llama 2-13B をベースに，エラー検出，データ補完，スキーママッチング，エンティティマッチングなどの典型的な DP タスクのデータセットでインストラクションチューニングされた．モデルサイズが小さいため，低価格の GPU 上でローカルに動作し，データのセキュリティを維持しつつ，さらなるチューニングも可能にします．実験により，Jellyfish は GPT シリーズのモデルと比較しても遜色なく高性能が持ち，そのインタープリターモデルは GPT-3.5 に比べて推論能力が向上していることが示されました．

キーワード　LLM，自然言語処理応用，データ統合技術

1 はじめに

データ前処理（data preprocessing，DP）はデータマイニングパイプラインにおいて重要なプロセスであり，生データを扱いやすく，処理可能な形式に変換する．DP のタスクについて，エラー検出 [4], [16]，データ補完 [17], [22]，スキーママッチング [32]，エンティティマッチング，[7], [11] などが含まれている．これらのタスクの中に，エラー，異常，マッチなどの取り扱いと，検出，修復，整列などの異なるアクションが必要であるため，今までは汎用的な手法の開発は困難である．一方，GPT-3 やその後のバージョンなどの大規模言語モデル（large language model，LLM）が登場することにより，DP タスクに対する汎用的な生成方式の手法の開発が促進されている．[18], [31]．従来の DP 手法と比較して，LLM を使用することの最大の利点は，プログラミングスキルや特定のツールの使用に頼ることなく，自然言語での対話を通じて DP 処理を完了できる点にある．また，事前学習では大量のデータから知識を学習したため，LLM は優れた推論能力を持ち，人間によるルール [21] を提示しなくてもエラー，異常，マッチを特定できる．また，DP 処理の結果だけでなく，その結果に至った理由も自然言語で説明できるため，解釈性において他のディープラーニング手法を上回る．さらに，LLM は少数ショット（few-shot）[2] またはゼロショット（zero-shot）[6] のプロンプト（prompt）によって，少数サンプルで in context learning が可能になり，従来手法の閾値の調整 [7], [23] やパラメータチューニング [17] などが不要になる．

しかし，既存の OpenAI の GPT シリーズの API に依存する DP 手法 [8], [18], [31] は，データ漏洩に関する懸念を引き起こしている．また，高度に専門化された分野のデータを扱う際には，追加学習のコストが非常に高まり，DP ドメインに特化することが困難である．

これらの課題に対し，本研究では Jellyfish という LLM を提案する．Jellyfish は Meta の Llama 2-13B モデルに基づいており，幅広い DP タスクに対して in struction tuning [33] により汎用的な DP モデルである．Jellyfish は 130 億という比較的に少ないパラメータ数を持ち，低価格の単独の GPU 上でも動作できることで，ローカル環境でデータセキュリティを確保できる．追加学習も低コストで実現できる．また，従来の方法が固定なルールや知識 [20], [22] に依存することと違い，Jellyfish はチューニングプロセス中に自動にドメイン知識を獲得でき，意図的に知識注入もできる．さらに，自然言語の出力による DP 結果に対する解釈性もある．

Jellyfish の開発には，Llama 2-13B をベースモデルとして選択し，事前チューニングと DP チューニング技術を適用した．事前チューニング段階では，推論トレースを強化するためのデータセットと，STEM および論理スキルを向上させるためのデータセットの 2 つを選択した．DP チューニング段階では，トレーニングデータセットをモデルへの入力プロンプトにシリアライズするための instance serializer を使用し，特にドメイン知識をモデルに注入するための knowledge injector を使用した．また，GPT-4 を使用して DP 結果の推論データを生成し，Jellyfish の推論過程のロジック性もチューニングした．

DP タスク解決のために，Jellyfish の instance serializer は生データをプロンプトに自動変換し，knowledge injector はオプションの知識指定を可能にする．プロンプトは特徴エンジニ

アリングをサポートし，モデルは未見のデータセットに対して特に有益な条件付けのために少数ショットプロンプトを組み込んでいる．

Jellyfish の評価では，公開されているベンチマークを用いて，非 LLM 手法と LLM 手法の両方と比較した．その結果，Jellyfish は seen dataset において非 LLM 手法を一貫して上回り，unseen dataset におけるその有効性は，seen dataset という in domain の学習の非 LLM 手法と同等であることが分かった．Jellyfish は unseen task の 2 つのケーススタディでも性能が良いし，DP タスクの instruction tuning に使用された 4 つのタスクを超える幅広く一般化の能力を示した．そのパフォーマンスは GPT シリーズモデルと競合し，出力結果の解釈性は GPT-3.5 より優れていることが示された．

論文の構成として，2 章では対象とする DP タスクの詳細や LLM と関連研究を簡単に紹介する．3 章では，Jellyfish モデルのフレームワークを説明する．4 章と 5 章は，ベースモデルの選択，事前チューニング，DP タスクのための指示チューニングについて，Jellyfish の構築を詳述する．6 章では，DP タスクを解決するための Jellyfish の使用方法を紹介する．7 章では，実験結果と分析を報告する．

2 問題定義および事前知識

2.1 データ前処理

DP（データ前処理）は，ノイズ，欠損値，不整合，データの異質性に対処する重要なステップである．主な DP 手法には，データクリーニング，データ統合，データ変換，重複削除などがある．本研究では，最も一般的なデータタイプの一つである表形式のデータを対象とする．

我々のデータモデルは，スキーマによって指定されたリレーショナルテーブルで操作する．すべての属性は数値（バイナリ値を含む）またはテキスト値（カテゴリ値を含む）であると仮定する．次に，本論文で関与する DP タスクの概要を述べる．

エラー検出（*error detection*，*ED*）．与えられたタプルと属性について，この属性のセル値にエラーがあるかどうかを検出するタスクである．

データ補完（*data imputation*，*DI*）．与えられたタプルと属性において，この属性のセル値が欠損している場合，その正しい値を推測するタスクである．

スキーママッチング（*schema matching*，*SM*）．（名前，説明）の形式で表される属性のペアについて，それらが同じ属性を参照しているかどうかを見つけるタスクである．

エンティティマッチング（*entity matching*，*EM*）．同じスキーマのタプルのペアについて，それらが同じエンティティを参照しているかどうかを推測するタスクである．

列タイプ注釈（*column type annotation*，*CTA*）．ヘッダーのないテーブルが与えられた場合，事前に定義されたセマンティックタイプ（例：名前，時間，位置）のセットから各列のタイプを推測するタスクである．

属性値抽出（*attribute value extraction*，*AVE*）．エンティティに関するテキストと事前に定義された属性のセットが与えられた場合，テキスト説明から対応する属性値を抽出するタスクである．

我々は，各入力オブジェクトを**インスタンス**と呼ぶ．つまり，ED および DI の場合はタプル，SM の場合は属性のペア，EM の場合はタプルのペア，CTA の場合はテーブルまたは列，AVE の場合はテキストである．

2.2 大規模言語モデル

自然言語処理（NLP）分野の進歩に伴い，大規模言語モデル（LLM）は AI 研究コミュニティの中で最も注目されるトピックである．代表的な LLM には，OpenAI の GPT シリーズ（特に GPT-3，3.5，および 4），Anthropic の Claude，Meta の Llama [24] や Llama 2 [25]，および Hugging Face で見つかるそれらのバリアントが含まれる．その卓越した自然言語処理能力により，LLM は NLP アプリケーション（例：ChatGPT や Claude）での使用に留まらず，AI アシスタント（例：GPT による）やエンジニアリング [19]，シミュレーション [29] 目的のための LLM 駆動型自律エージェントもある．一方，オープンソースのローカル LLM もあり（例：Llama や Llama 2）．モデルのパラメータ数がより少ないため，論理的推論，質問応答などの能力を向上させるために追加のタスクでファインチューニングが実施しやすい．これらのファインチューニング手法の中で，instruction tuning [33] は盛んなものとなり，LLM を（指示，出力）ペアで構成されるデータセットでさらに訓練することにより，人間の指示に従って対話的に答えを生成することができるようになった．

3 提案フレームワーク

図 1 は Jellyfish モデルの概要を示している．このフレームワークは，チューニング部分と DP 処理部分の 2 つから構成されている．

チューニング部分では，LLM をベースモデルとして選択し，DP 用にチューニングする．チューニングには 2 つの段階がある．第 1 段階は Pre-truning 段階であり，データ分析パイプラインにおける論理的推論能力を強化することを目的としている．第 2 段階は DP-tuning 段階であり，四つ DP タスク（ED，DI，SM，EM）で第 1 段階の結果モデルを instruction tuning する．ここでは，典型的な DP タスクのみをチューニングし，これらのタスクを解決する能力が DP-tuning を通じて新しいタスクにトランスファーできるかを評価し，汎用的な DP 処理モデルを作成する．DP-tuning の結果，DP タスクに対処するための Jellyfish-13B モデルと，DP 結果を分析するための Jellyfish-13B-interpreter モデルがそれぞれ作れる．

Jellyfish-13B をチューニングするために，DP で一般的に使用される表形式のデータセットのセットを用意し，インスタンスと自然言語指示の形での正解ラベルから成る指示データを準備する．特に，DP タスクに関連する知識をこれらの指示データに

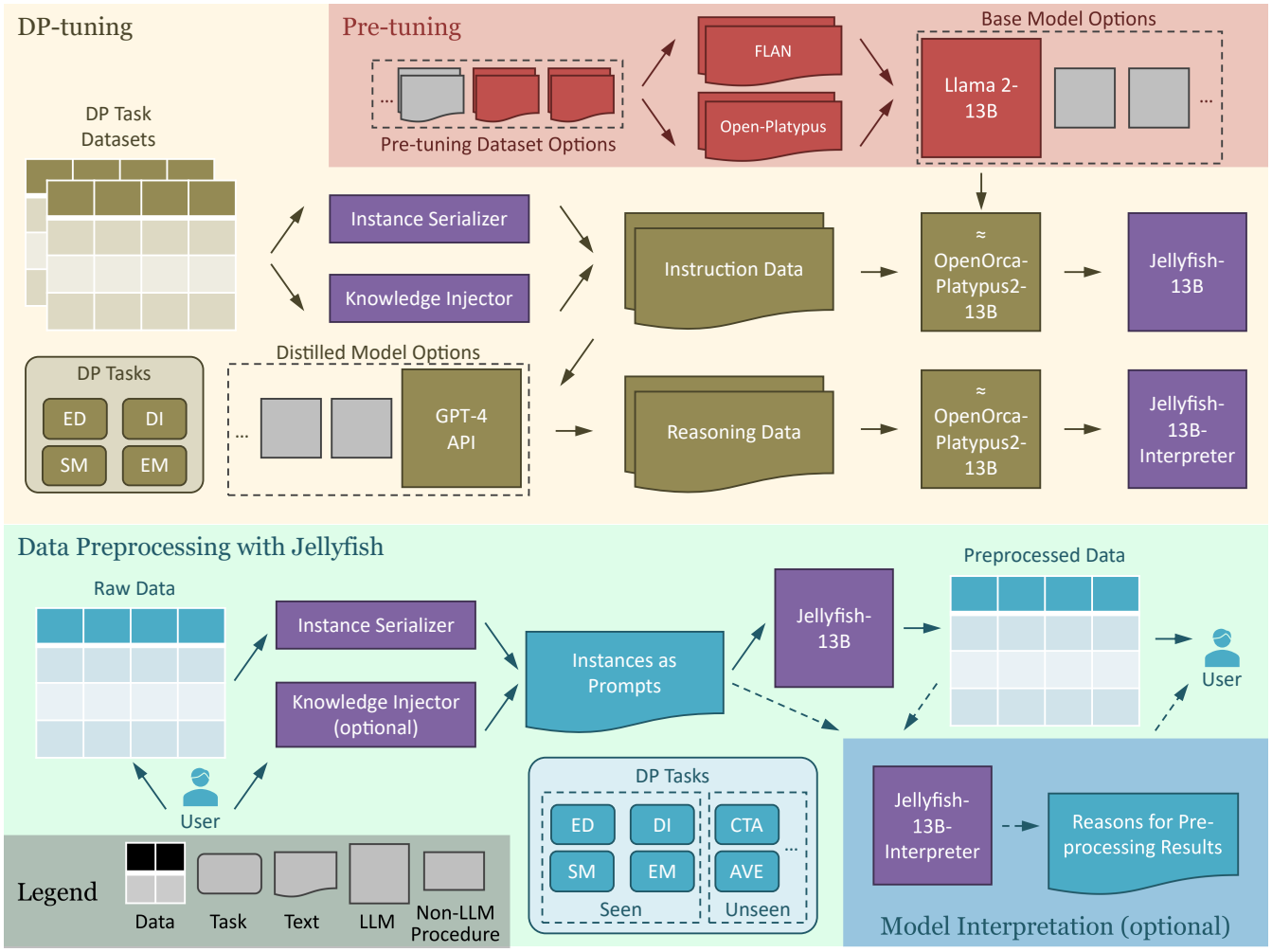


図 1: Jellyfish モデルのフレームワーク．Jellyfish のコンポーネントは紫色で示されている．

注入する．このアプローチにより，モデルは基本的な DP ルールに従うとともに，データのどの側面に特に注意を払うべきかを見極めることができる．一方，Jellyfish-13B-interpreter をチューニングするために，Jellyfish-13B の指示データを GPT-4 に入れて，推論過程が含めた回答データを GPT-4 に生成してもらう．つまり，Jellyfish-13B-interpreter は DP タスクに関する GPT-4 の知識を蒸留した結果と見なすことができる．Jellyfish フレームワークでは，ベースモデルと知識蒸留先のモデルを自由に選択することもできる．

DP 処理部分では，Jellyfish-13B が DP タスクを処理する．これは instance Serialize を使用し，前処理されるデータを取り込み，チューニングされた LLM に入力されるテキストシーケンス，すなわちプロンプト（prompt）を構築する．さらに，ユーザーがフォーマットや値に関する事前知識を指定できるオプショナルな知 knowledge injector を備えている．Jellyfish-13B-interpreter は，一方で，Jellyfish の出力を処理し，そのような結果を支える理由を提供する．

4　ベースモデルの選択と Pre-tuning

Jellyfish のベースモデルを選択するための以下の基準を設定した．(1) モデルはオープンソースである必要があり，さらなるファインチューニングが可能であること．(2) 適度なサイズであること．小さすぎるモデルは複数の DP タスクを処理できなく，大きすぎるモデルは多めの GPU リソースが必要になり，非効率的である．この理由から，Llama 2-13B をベースモデルとして選択する．

しかし，Llama 2-13B は推論能力に制限があり，人間による評価は推論関連のプロンプトを含んでいない [25]．DP タスクは基本的にエラー，マッチ，その他のデータ側面についての推論を伴うため，推論能力を強化するためにベースモデルを調整する必要がある．推論能力の他に，データ分析におけるドメイン知識も懸念事項である．なぜなら，DP はデータマイニングパイプラインにおける重要なコンポーネントだからである．しかし，このようなドメイン知識はベースモデルには欠けている．

これらの制限を克服するため，DP 用にチューニングする前に，慎重に選ばれた一連のデータセットでベースモデルをプレチューニングする．まず，推論トレースに焦点を当てたフューショット [2]，ゼロショット [6]，および推論を引き出すチェーンオブソート（chain-of-thought）[28] プロンプト設定を特徴とする FLAN データセット [14] を選択する．次に，モデルの STEM および論理的な習熟度を高めることを目的とした Open-Platypus データセット [9] を選択する．これらのデータ

セットは，エンティティの意味やデータエントリのフォーマットをモデルが理解するために不可欠なタスクを含んでおり，データマイニングにおけるドメイン知識と完全に一致している．

2 つのデータセットのサイズが大きいため，ベースモデルのプレチューニングは非常に時間がかかる傾向がある．今回の Jellyfish では，OpenOrca-Platypus2-13B [10] という，OpenOrca [12] でファインチューニングされた Llama 2-13B（FLAN コレクションを拡張したもの）と，Open-Platypus でファインチューニングされた Llama 2-13B でマージしたモデルを採用した.. 将来では，FLAN と Open-Platypus に絞り込み，DP に最も関連するデータのサブセットを見つけることで，あらかじめ Llama-2 から追加学習する可能性もある．

5 DP-tuning

5.1 DP-tuning のデータセット

DP-tuning に関連する 4 つのタスクについて，以前の研究で広く使用され，多様な応用ドメインをカバーする一連のデータセットを選択した．(1) ED（エラー検出）：Adult と Hospital [4], (2) DI（データ補完）：Buy と Restaurant [17], (3) SM（スキーママッチング）：MIMIC-III と Synthea [32], (4) EM（エンティティマッチング）：Magellan データリポジトリ [3] からの Amamzon-Google, Beer, DBLP-ACM, DBLP-GoogleScholar, Fodors-Zagats, iTunes-Amazon．これらのデータセットの公開バージョン [18] を使用する．ここでは，ED と DI 用にそれぞれエラーや欠損値がデータセットに既に注入されている．これらのデータセットの統計は論文の完全版 [30] の表 1 に提供されている．各データセットにおけるインスタンス数を決定する際の考慮事項は，異なるタスクを横断するデータセットがバランス良く，全体のコーパスを支配するデータセットがないことである．さらに，欠損値は応用によってエラーまたは非エラーと見なされる可能性があるため，チューニング時にインスタンスを複製し，一つは欠損値をエラーとして，もう一つは非エラーとして使用する．各データセットをトレーニングと検証に 80:20 の比率で分割する．次に，生のデータを Jellyfish-13B および Jellyfish-13B-interpreter のチューニングに使用される指示に変換する方法を紹介する．

5.2 Jellyfish-13B

LLM のトレーニングデータを準備するためには，生データ内の各インスタンスをプロンプトにシリアライズ（別名コンテキスト化）する必要がある．プロンプトには，タスクの説明，インスタンスの内容，および注入された知識が含まれる．トレーニング用の指示データを構築するための技術を説明するために，EM 用の Beer データセットのインスタンスの例を以下に示す．太字部分はモデルに提示されないコメントを示している．構築および評価に使用される他のプロンプトは，論文の完全版の付録 A に提供されている．

(system message) *You are an AI assistant that follows instruction extremely well. User will give you a question. Your task is to answer as faithfully as you can.*
(task description) *You are tasked with determining whether two Products listed below are the same based on the information provided. Carefully compare all the attributes before making your decision.*
(injected knowledge) *Note that missing values (N/A or "nan") should not be used as a basis for your decision.*
(instance content) *Product A: [name: "Sequoia American Amber Ale", factory: "Wig And Pen"]*
Product B: [name: "Aarhus Cains Triple A American Amber Ale", factory: "Aarhus Bryghus"]
(question) *Are Product A and Product B the same Product?*
(output format) *Choose your answer from: [Yes, No]*

最初に，モデルの振る舞いをガイドするシステムメッセージがある．ここでは，モデルに AI アシスタントとしてユーザーの質問に答えるよう指示し，その応答は常にこの制約を尊重する必要がある [25]．2 行目では DP タスク，つまりこの例では EM を説明する．3 行目は注入された知識に関するものである．注入される知識には 2 種類ある：(1) 多くのデータセットに適用可能な一般的な知識，および (2) 与えられたデータセットにのみ適用される特定の知識である．この例では，知識は一般的な知識に属し，欠損値に関するものである．このような知識の注入は，特にトレーニングデータがノイズを含む場合，モデルがデータセット内の特定の値を誤って処理するのを防ぐ可能性がある．4 行目と 5 行目はインスタンスの内容に関するものである．最後に，モデルに提示される質問があり，その後で出力フォーマットが指定される．

上記の例では欠損値に関する知識を指定しているが，DP-tuning で使用される一般的な知識には，エラーの種類や用語など，他の形式もある．例えば，ED では，エラーにはスペルミス，不整合，その属性には無意味な値などが含まれるが，それに限定されないことをモデルに伝える．EM では，モデルに属性のフルネームとその頭字語を考慮して，2 つの値が同じかどうかを判断するよう指示する．特定の知識は主に応用ドメインに依存し，データセットに関連する制約やルールを含む．例えば，出版物のデータセットでは，同じ記事であっても，著者の名前が異なる形式や順序で現れる場合がある．さらに，モデルは特定の属性に重要性を割り当てるよう設定できる．例えば，製品データのコンテキストでは，モデルに製品番号の比較を優先するよう指示される．特定の知識は同じドメイン内のデータセットに適用可能であり，これらのデータセットに関する事前知識がないシナリオにおいて，モデルのパフォーマンスを向上させる可能性がある．全体として，instruction tuning を通じて注入された知識は Jellyfish の組み込み知識となり，推論中にユーザーが指定しなくても使用できる．

5.3 Jellyfish-13B-interpreter

Jellyfish-13B-interpreter は，Jellyfish-13B と同じデータセット群を instruction tuning に使用するが，トレーニング時間を

節約するためにインスタンス数を若干少なくする．具体的には，SM 用に 2500 件，EM 用に 360 件であり，詳細な統計は論文の完全版の付録 B に記載されている．このより少ないトレーニングデータセットは，モデル選択の初期実験にも使用され，このデータセットを使用したチューニングプロセスは DP-tuning プレビューと呼ばれる．トレーニングのための推論データは，上記で紹介した指示データと似ている．唯一の違いは，推論指示である．上記の例では，システムメッセージに追加された推論指示と，出力フォーマットで指定された推論指示がある．

> ***(system message) [same as Instruction Data]*** *While answering, provide detailed explanation and justify your answer.*
> ***(task description – question) [same as Instruction Data]***
> ***(output format)*** *After your reasoning, finish your response in a separate line with and ONLY with your final answer. Choose your final answer from [Yes, No].*

このように，モデルは最終的な答えを提示する前に理由を示す必要がある．通訳者として，最後には「はい」または「いいえ」と答えることも求められる．将来的に単一のモデルで DP タスクの解決と解釈の両方を担う可能性を維持するために，このデザインを選択した．指示データの「はい/いいえ」の真実とは異なり，推論データの真実はより複雑である．合理的な答えを得るために，GPT-4 を利用する．

Jellyfish-13B と Jellyfish-13B-interpreter は，学習を加速するために，LoRA [5] を採用する．これは，ファインチューニング中の重みの変化における低い固有ランクを利用する．Transformer アーキテクチャ [27] 内の W_q，W_k，W_v，および W_o の行列に LoRA を適用する．

6 DP タスク処理

CSV 形式のデータセットが与えられた場合，Jellyfish は instance serializer を使用し，すべてのインスタンスを反復処理し，各インスタンスをプロンプトに変換する．このプロンプトは，Jellyfish-13B および Jellyfish-13B-interpreter のチューニング用の指示データおよび推論データと同じである．DP タスクに関連する一般的な知識を適用する．例えば，マッチングタスクにおける欠損値や，ED におけるエラータイプなどである．Jellyfish-13B には，ユーザーがデータセット固有の知識を入力できる知識インジェクターも提供されている．このようなユーザー指定の知識はオプショナルであり，必須ではなく，実験では使用されていない．

特徴量エンジニアリング. ユーザーはオプションで，パフォーマンスを向上させるために特徴量のサブセットを選択できる．例えば，Beer データセットにおける EM では，`name` と `factory` がより関連性の高い特徴であり，`style` と `ABV` は関連性が低い．したがって，ユーザーは属性として `name` と `factory` のみを使用することを選択できる．このような特徴量エンジニアリングは，特定の知識としてプロンプトにも実装できる．

プロンプトエンジニアリング. プロンプトエンジニアリングは，モデルのパフォーマンスを向上させるためのテキストの構造化プロセスである．Jellyfish-13B には，データセットから引き出された少数の例からモデルが学習するように条件付けされたフューショットプロンプティング [2] を導入する．フューショットプロンプティングは，データセットからデータインスタンスのサブセットを手動で選択し，ラベリングすることによって適用する．

実際の応用では ground truth が利用できないため，ユーザーは推論のためにフューショットの例を手作りできる．一方で，フューショットの例は，ED や DI に対して欠損値，タイプミス/フォーマットエラー，タプル内の 2 つの列または列内の 2 つのタプルの値をランダムに入れ替えることにより，自動的に生成することができる．SM や EM については，いくつかの一致を迅速に見つけるためのルールベースの方法（例：ブロッキングルール（blocking rule）[7]）を利用し，それらをフューショットの例として使用できる．

7 評価実験

7.1 実験設置

データセット. DP-tuning において使用したデータセットに加えて，以下のデータセットを未見データとして使用する．CTA と AVE は unseen task のケーススタディに使用される．(1) ED：Flights と Rayyan [16]，(2) DI：Kaggle の Flipkart と Phone，(3) SM：CMS [32]，(4) EM：Magellan データリポジトリ [3] からの Abt-Buy と Walmart-Amazon，(5) CTA：SOTAB [8]，(6) AVE：AE-110k と OA-Mine [1]．データセットの統計は論文の完全版の表 2 で報告されている．Adult と Hospital のためには [4]，Flipkart と Phone のためには [17]，MIMIC-III と CMS のためには [32] のプロトコルに従ってトレーニング/検証/テスト分割を生成する．他のデータセットは既に分割が提供されている [1], [8], [18].

方法. 既存の方法を非 LLM 法と LLM 法に分類する．非 LLM 法には，以下のベースラインを [18] に沿って選択する：(1) ED：HoloDetect [4] と Raha [16], (2) DI：IPM [17]，(3) SM：SMAT [32]，(4) EM：Ditto [11]，(5) CTA：RoBERTa [13]．これらのパフォーマンスについては，先行研究 [8], [18] で報告された最良の数値に従う．

LLM 法には Jellyfish，GPT-3 (`text-davinci-002`)，GPT-3.5 (`gpt-3.5-turbo-0301`)，GPT-4 (`gpt-4-0314`)，Stable Beluga 2 70B [15]，SOLAR 70B [26] が含まれる．[1], [18], [31] で報告された最良の数値に従ってパフォーマンスを報告する．DP-tuning の効果を示すために，DP-tuning 前のモデルである OpenOrca-Platypus2-13B [10]（「OOP2」と表記）の結果も報告する．

Jellyfish の DP-tuning のためのハイパーパラメータ設定は論文の完全版の付録 B で提供されている．Jellyfish に注入された知識は付録 C で報告されており，推論中のプロンプトに

はデータセット固有の知識は注入されていない．フューショットプロンプティングが有効な場合，各データセットに対して 3 つの例を Jellyfish に装備するが，これは付録 D で報告されている．

指標. DP タスク解決には，DI に対する精度，ED，DI，EM，AVE に対する F1 スコア，CTA に対するマイクロ F1 を 100 点スケールで測定する．また，SM に対する精度と再現率も報告する．モデル解釈については，Jellyfish-13B-interpreter と GPT-3.5 の出力を GPT-4 に比較して勝者を決定してもらう．

実験環境. 実験は，Intel(R) Xeon(R) E5-2640 v4 @2.10GHz 8 コア CPU および 256GB RAM を備えたサーバー上で行われ，Ubuntu 18.04.6 LTS が実行されている．Jellyfish は推論のために NVIDIA RTX A6000 GPU（48GB グラフィックスメモリ）で加速され，DP-tuning のために 2 つの GPU を使用する．DP-tuning は Python 3.10.9 で実装されている．

7.2 seen task におけるパフォーマンス

まず，DP-tuning に使用されたタスクにおけるパフォーマンスを評価する．表 1 は，DI の精度と他の 3 つのタスクの F1 スコアを報告している．Jellyfish-13B は，seen dataset ではフューショットプロンプティングを不使用にし，unseen dataset では使用にしている．全ての競合相手の中で，GPT-4 が一般的に最も優れたパフォーマンスを発揮する．これは，最も進歩したモデルであり，最大のパラメータ数を持っているため予想される結果である．しかし，SM におけるそのパフォーマンスは平凡である．Jellyfish-13B は一般的に準優勝モデルであり，SM では GPT-4 を大幅に上回る．unseen dataset におけるその精度や F1 スコアは 80%以上であり，SM を除いては．さらに，Jellyfish-13B は非 LLM 法を除く全ての seen dataset，および全ての unseen dataset で非 LLM 法を上回る．非 LLM 法については，入力データセットにトレーニングが必要なため，すべてのデータセットが見た．これは，これらのデータセットにトレーニングせずに，Jellyfish-13B のパフォーマンスがトレーニングを受けた非 LLM 法のパフォーマンスを上回ることを意味する．GPT-3 および GPT-3.5 と比較すると，Jellyfish-13B はより多くのケースで勝利する．Jellyfish-13B の強力なパフォーマンスは，DP-tuning によるものであることがわかる．なぜなら，DP-tuning を行っていない OOP2 は劣る結果を報告するからである．特に，OOP2 は DI の Flipkart と Phone で合理的な回答を報告できない．その一方で，Jellyfish-13B はそれらで 81%と 87%の精度を達成する．

7.3 unseen task におけるパフォーマンス

次に，Jellyfish-13B を unseen task に拡張し，2 つのケーススタディを行う．これらは，既存の研究 [1], [8] のプロンプトに従い，フューショットの例を使用せずに行う．

表 2 は CTA のパフォーマンス比較を報告している．RoBERTa はこのタスクのためにファインチューニングが必要である．我々は，トレーニングデータの 159 ショットと 356 ショットの 2 つのオプションでその結果を報告し，[8] の数値に従っている．GPT-4 は依然として最も優れたパフォーマンスを示している．このタスクに対するチューニングがなくても，Jellyfish-13B は 159 ショットでファインチューニングされた RoBERTa よりも優れている．また，OOP2 よりも大幅に優れており，DP-tuning が見られた 4 つのタスクを処理する能力をこの unseen task にうまく転送できたことを示唆している．

AVE の結果は表 3 に報告されている．GPT-3.5 は AE-110k データセットで勝者であり，GPT-4 は OA-Mine データセットで勝者である．Jellyfish-13B は AE-110k で 2 位，OA-Mine で 3 位である．13B パラメータであるにもかかわらず，Jellyfish-13B が両データセットで 70B の 2 つのモデルを上回り，AE-110k では GPT-4 をも打ち負かしている．

7.4 答えの解釈性の評価

Jellyfish-13B-interpreter のパフォーマンスを評価し，GPT-3.5 (`gpt-3.5-turbo-0613`) と比較する．Jellyfish-13B によって出力された回答に基づいて，Jellyfish-13B-interpreter と GPT-3.5 の両方を使用して理由を生成し，どちらが優れているかを GPT-4 に判断してもらう．GPT-4 は DP の質問に対する正しい回答を知らないため，自身の質問分析にも基づいて判断する必要がある．表 4 は，2 つのモデルの頭頭比較を報告している．Jellyfish-13B-interpreter はすべてのデータセットで GPT-3.5 に勝利し，220 回のテストで全体的な勝率 78 %を達成している．

8 ま と め

本論文では，DP 用に特別に設計された LLM である Jellyfish を提案した．Llama 2-13B をベースにし，4 つのタスクのために指示チューニングされた Jellyfish は，汎用的な DP タスクソルバーとして一般化する．130 億のパラメータをさらにチューニング可能であり，ローカル GPU 上で動作させることでデータセキュリティを損なわずに運用できる．自然言語を理解する能力に長けており，ユーザーは DP タスク処理のために手動で指示を作成することができる．多くの既存の方法が事前知識に大きく依存しているのに対し，Jellyfish はチューニングプロセス中にドメイン知識を獲得し，推論中にオプショナルな知識注入を活用する．注目すべき特徴は，その出力に対する説明を提供するインタープリターである．我々は Jellyfish を DP タスクに一般的に使用される公開データセットのコレクションを用いて評価し，既存のソリューションに対する競争力，新しいタスクへの一般化能力，優れた推論能力を実証した．

文 献

[1] A. Brinkmann, R. Shraga, and C. Bizer. Product attribute value extraction using large language models. *arXiv preprint arXiv:2310.12537*, 2023.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[3] S. Das, A. Doan, P. S. G. C., C. Gokhale, P. Konda, Y. Govind, and D. Paulsen. The magellan data repository. `https://sites.google.com/site/anhaidgroup/`

表 1: seen task でのパフォーマンス，太字が優勝，下線が次点．フューショットは，seen dataset では不使用，unseen dataset では使用である．ED の非 LLM は，Adult と Hospital の HoloDetect，Flights と Rayyan の Raha を指す．「–」は先行研究で報告されていない数字を示す．「INV」は有効な回答が得られなかったことを示す．

Task	Type	Dataset	Model					
			Non-LLM	GPT-3	GPT-3.5	GPT-4	OOP2	Jellyfish-13B
ED	Seen	Adult	99.10	99.10	92.01	92.01	42.77	**99.33**
		Hospital	94.40	**97.80**	90.74	90.74	63.24	95.59
	Unseen	Flights	81.00	–	–	**83.48**	65.74	82.52
		Rayyan	79.00	–	–	81.95	74.30	**90.65**
DI	Seen	Buy	96.50	98.50	98.46	**100**	89.23	**100**
		Restaurant	77.20	88.40	94.19	**97.67**	81.40	89.53
	Unseen	Flipkart	68.00	–	–	**89.94**	INV	81.68
		Phone	86.70	–	–	**90.79**	INV	87.21
SM	Seen	MIMIC-III	20.00	–	–	**40.00**	36.36	**40.00**
		Synthea	38.50	45.20	57.14	**66.67**	22.22	56.00
	Unseen	CMS	50.00	–	–	19.35	13.33	**59.29**
EM	Seen	Amazon-Google	75.58	63.50	66.50	74.21	36.70	**81.34**
		Beer	94.37	**100**	96.30	**100**	85.71	96.77
		DBLP-ACM	**98.99**	96.60	96.99	97.44	78.86	98.98
		DBLP-GoogleScholar	95.60	83.80	76.12	91.87	59.48	**98.51**
		Fodors-Zagats	**100**	**100**	**100**	**100**	92.68	**100**
		iTunes-Amazon	97.06	98.20	96.40	**100**	57.45	98.11
	Unseen	Abt-Buy	89.33	–	–	**92.77**	61.78	89.58
		Walmart-Amazon	86.76	87.00	86.17	**90.27**	67.29	89.42

表 2: CTA（unseen task）の Micro-F1，フューショット不使用．

RoBERTa (159 shots)	RoBERTa (356 shots)	GPT-3.5	GPT-4	OOP2	Jellyfish-13B
79.20	89.73	89.47	**91.55**	56.40	82.00

表 3: AVE（unseen task）の F1，フューショット不使用．

Dataset	Model					
	Stable Beluga 2 70B	SOLAR 70B	GPT-3.5	GPT-4	OOP2	Jellyfish-13B
AE-110k	52.10	49.20	**61.30**	55.50	49.28	58.12
OA-Mine	50.80	55.20	62.70	**68.90**	43.48	55.96

表 4: Jellyfish-13B と GPT-3.5 の解釈の比較．

Task	Dataset	Model	
		GPT-3.5	Jellyfish-13B-interpreter
ED	Adult	3	**17**
	Hospital	3	**17**
DI	Buy	2	**18**
	Restaurant	3	**17**
SM	Synthea	7	**13**
EM	Amazon-Google	5	**15**
	Beer	6	**14**
	DBLP-ACM	4	**16**
	DBLP-GoogleScholar	4	**16**
	Fodors-Zagats	2	**18**
	iTunes-Amazon	9	**11**
Total		48	**172**
Winning Rate		21.82%	**78.18%**

useful-stuff/the-magellan-data-repository.

[4] A. Heidari, J. McGrath, I. F. Ilyas, and T. Rekatsinas. HoloDetect: Few-shot learning for error detection. In *SIGMOD*, pages 829–846, 2019.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[7] P. Konda, S. Das, A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, S. Prasad, et al. Magellan: toward building entity matching management systems over data science stacks. *PVLDB*, 9(13):1581–1584, 2016.

[8] K. Korini and C. Bizer. Column type annotation using ChatGPT. *arXiv preprint arXiv:2306.00745*, 2023.

[9] A. N. Lee, C. J. Hunter, and N. Ruiz. Platypus: Quick, cheap, and powerful refinement of LLMs. *arXiv preprint arXiv:2308.07317*, 2023.

[10] A. N. Lee, C. J. Hunter, N. Ruiz, B. Goodson, W. Lian, G. Wang, E. Pentland, A. Cook, C. Vong, and "Teknium". OpenOrcaPlatypus: Llama2-13B model instruct-tuned on filtered OpenOrcaV1 GPT-4 dataset and merged with divergent STEM and logic dataset model. https://huggingface.co/Open-Orca/OpenOrca-Platypus2-13B, 2023.

[11] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan. Deep entity matching with pre-trained language models. *PVLDB*, 14(1):50–60, 2020.

[12] W. Lian, B. Goodson, E. Pentland, A. Cook, C. Vong, and "Teknium". OpenOrca: An open dataset of GPT augmented FLAN reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`, 2023.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The FLAN collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[15] D. Mahan, R. Carlow, L. Castricato, N. Cooper, and C. Laforte. Stable beluga 2. `https://huggingface.co/stabilityai/StableBeluga2`, 2023.

[16] M. Mahdavi, Z. Abedjan, R. Castro Fernandez, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang. Raha: A configuration-free error detection system. In *SIGMOD*, pages 865–882, 2019.

[17] Y. Mei, S. Song, C. Fang, H. Yang, J. Fang, and J. Long. Capturing semantics for imputation with pre-trained language models. In *ICDE*, pages 61–72, 2021.

[18] A. Narayan, I. Chami, L. Orr, and C. Ré. Can foundation models wrangle your data? *PVLDB*, 16(4):738–746, 2022.

[19] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, and M. Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

[20] J. Qin, S. Huang, Y. Wang, J. Zhu, Y. Zhang, Y. Miao, R. Mao, M. Onizuka, and C. Xiao. BClean: A bayesian data cleaning system. *arXiv preprint arXiv:2311.06517*, 2023.

[21] S. Razniewski, A. Yates, N. Kassner, and G. Weikum. Language models as or for knowledge bases. *arXiv preprint arXiv:2110.04888*, 2021.

[22] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. HoloClean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(10):1190–1201, 2017.

[23] T. Sagi and A. Gal. Schema matching prediction with applications to data source discovery and dynamic ensembling. *The VLDB Journal*, 22:689–710, 2013.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[26] Upstage. Solar-0-70b-16bit. `https://huggingface.co/upstage/SOLAR-0-70b-16bit`, 2023.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[29] Z. Wu, R. Peng, X. Han, S. Zheng, Y. Zhang, and C. Xiao. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*, 2023.

[30] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada. Jellyfish: A large language model for data preprocessing. *arXiv preprint arXiv:2312.01678*, 2023.

[31] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada. Large language models as data preprocessors. *arXiv preprint arXiv:2308.16361*, 2023.

[32] J. Zhang, B. Shin, J. D. Choi, and J. C. Ho. SMAT: An attention-based deep learning solution to the automation of schema matching. In *ADBIS*, pages 260–274, 2021.

[33] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年3月1日(金) 10:00 ~ 12:10 | T1-B オンライン（Zoom Events)

感情分析 (1)

座長:小杉 尚子(専修大学)
コメンテータ:丸橋 弘治(富士通株式会社)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T1-B-7-01] 幼児の音声感情認識における最適な特徴量と機械学習の検討

*樽見 理花[1]、伊藤 貴之[1] (1. お茶の水女子大学理学部情報科学科伊藤研究室)

10:20 ~ 10:35
[T1-B-7-02] 絵文字を用いた短文の特徴と感情伝達効果

*宮下 大生[1]、塩野 直志[1] (1. 神奈川工科大学)

10:35 ~ 11:00
[T1-B-7-03] 歌詞感情分類データセット作成のための大規模言語モデル利用法

*宮川 尚[1]、宇津呂 武仁[1] (1. 筑波大学)

11:00 ~ 11:25
[T1-B-7-04] 共感的な働きかけを意図したエモーショナルサポート会話における発話生成

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

DEIM Forum 2024 XX-Y

幼児の音声感情認識における最適な特徴量と機械学習の検討

樽見　理花†　　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{g2020532,itot}@is.ocha.ac.jp

あらまし　2 歳程度の幼児の育児は危険がつきもののため，ひとときも目を離すことが出来ない．この状況下において，家事と育児の両立は非常に大変であり負担が大きい．そこで，家事をしている間に幼児の感情を確認する負担を減らすことを目的として，幼児の声色による行動・感情の自動判別の適用について検討する．一方で，2 歳の喃語に関する研究は少なく，オープンデータも少ないため，深層学習の適用は難しい．そこで本報告では，深層学習以前の機械学習手法を前提として，最適な特徴量・機械学習の選択を議論するとともに，t-SNE を用いた可視化によってその汎化性を議論する．

キーワード　音声感情認識，幼児，機械学習，音響特徴量

1　はじめに

2 歳程度の幼児には，高いところに登ろうとしたり，おもちゃなどを食べようとしたり，といった危険がつきものである．そのため，彼らの育児はひとときも目を離すことができない．しかし育児をする中で別の家事を行ったり，自分の娯楽を楽しんだりする時間も必要である．近年，乳児の泣き声から乳児の感情を予測するアプリ [1] がリリースされ，言葉を話すことが出来ない乳児の泣き声による音声感情分類が可能となった．このような技術により育児の負担軽減が期待される一方で，2 歳程度の幼児の泣き声だけでなく，言葉にならない喃語を話すようになった幼児の音声感情分類については，いまだ研究が少ない．この状況を鑑みて我々は，2 歳程度の幼児の声色から感情を予測し，家事をしている両親に幼児の感情を伝えるシステムの開発を目標とした．

しかし本研究を進める上で，深層学習の訓練データに用いるための量的・質的に十分な音声データを幼児から収集することは簡単でないことがわかった．多数の幼児の音声を集めたオープンデータが少ない上に，幼児の音声と感情の関係は個人差が大きいため，多数の幼児の音声を混合した訓練データを構築することで高い性能を得られるとは限らない．一方で，1 人の幼児から収集できる音声データは量的に限りがあり，少なくとも我々の実験環境では深層学習の訓練データとするには難しい．さらに，我々が調査する限り，本研究の目的に対して適切かつオープンな学習済みモデルが見つからないため，現時点で転移学習の適用も試みていない．

そこで我々は，我々自身で収集できるデータ規模でも実行可能な深層学習として，音声データから生成したスペクトログラム画像を CNN-2D (Convolutional Neural Network 2D) で学習することを試みた．それと同時に，深層学習以外の機械学習手法の適用を試みた．以上の条件下での幼児の音声感情分類において，最適な特徴量と機械学習モデルを同定することにより，1 人の幼児から収集できる音声データのみからの感情判別の技術的可能性を実験により検証した．

2　関連研究

幼児の音声感情認識は近年発達してきた分野である．深層学習に利用できる規模の幼児の音声データを取得することが難しいことから，深層学習以前の機械学習手法を採用する事例が多い．田中ら [3] は，生後 2 日目から生後 5 日目までの新生児の乳児の泣き声を収録した DB を作り，それぞれの泣き声に乳児の状態を記したラベルを添付することで，泣き声の音響的情報を分析できるツールを作成した．しかし，この研究はデータ数が少ないことから，泣き声を用いた感情分類までには至っていない． Ashwini ら [4] は，生後 1 日から生後 10 日までの新生児を対象とし，空腹，痛み，眠気の 3 つの感情ラベルで計 300 の音声データを使用して，CNN-2D で特徴量を抽出し，SVM (Support Vector Machine) やランダムフォレスト（以下 RDF とする）を判別分析に用いている．この研究により，少ないデータ数でも深層学習を使用した特徴量抽出が可能であることが示された．本稿では，実験対象とする幼児の年齢が違うものの，この研究を実験の参考にしている．

本稿で取り入れた特徴量については，門谷ら [2] の音の 3 要素を考慮した特徴量を参考に選定している．この研究では，機械学習を使用せず隠れマルコフモデルを採用することで，成人の音声感情認識の実験を実施している．

3　実験手順

本研究で実施した実験の手順は以下の通りである．

（1） iPhone 搭載のカメラで 1 歳半程度の男児 1 名の行動を撮影し，音声を抽出する．

（2） 音響特徴量を抽出する．

（3） SVM やランダムフォレスト (RDF) を使用し，行動・感情を判別する．

（4） 判別した結果を t-SNE を用いて可視化する．

表 1　特徴量と機械学習モデルの比較.（単位：%）

	CNN-1D	SVM(線型)	SVM(ガウス)	RDF	平均精度
mel	70.9	63.64	67.27	81.82	70.91
mfcc	83.64	76.36	67.28	80.00	76.82
mel+mfcc	87.27	78.18	67.27	76.36	77.27
CNN-2D	-	90.91	-	-	90.91

表 2　交差検証結果の比較.（単位：%）

SVM(線型)	SVM(ガウス)	RDF
88.91	69.57	84.37

3.1　音声データの収集と前処理

1 歳 10ヶ月から 2 歳 6ヶ月までの期間の男児 1 名を対象とし，幼児の自宅にて，幼児が生後約半年から関わりのある大人が，幼児と 2 人きりの状態で iPhone 搭載のカメラを用いて幼児の行動を撮影する．この時，幼児とビデオカメラの距離は 2 メートル以内である．続いて撮影した動画を見ながら，幼児の感情を推定し，アノテーションをつける．この動画 (mov) を音声データ (wav) に変換し，2 秒ごとに切り取る．本稿では笑い声 (happy) と泣き声 (sad) の 2 種類のラベルを付与しているが，今後はイライラ (annoy) のラベルを追加する予定である．

3.2　特徴量の抽出

本研究では 2 種類の手法で特徴量の抽出を試みる．
1 種類目として，成人を対象とした音声感情分類の研究と乳児の泣き声を対象とした音声感情分類の論文調査結果に沿って，各々で使用されている音響特徴量を独自に選定し，組み合わせたものを使用する．本稿では，文献 [2] [3] より，成人と乳児の研究の両者で採用されている mel と mfcc を用いている．2 種類目として，音声データをスペクトログラムに変換し，特徴量抽出器として CNN-2D モデルを用いて特徴量を算出する．

3.3　行動・感情の判別

本研究では，笑い声 (happy) と泣き声 (sad) の 2 種類の音声を分類するタスクに対して，以下の 4 種類の特徴量

- mel のみ
- mfcc のみ
- mel と mfcc の両方
- 音声をスペクトログラム画像に変換し，CNN-2D モデルで算出した特徴量

を算出して実験に用いた．本実験にて使用した判別手法は

- CNN-1D モデル
- SVM(線型カーネル)
- SVM(ガウスカーネル)
- RDF

の 4 パターンについて実験した．そして，特徴量と機械学習の組み合わせの各々の精度を比較した．また交差検証については，SVM(線型)，SVM(ガウス)，RDF の 3 パターンについて実験した．

3.4　可　視　化

前節までに説明した各種方法で判別した結果に対して，t-SNE による次元削減を適用して可視化する．可視化に際して以下の 4 種類のラベルを示す．

可視化ラベル

1：笑い声 (happy) 正解データ
2：泣き声 (sad) 正解データ
3：笑い声 (happy) 不正解データ
4：泣き声 (sad) 不正解データ

4　実行結果と考察

本章では実行結果から得られる知見について述べる．機械学習に CNN-1D モデルを使用した場合，表 1 より，mel + mfcc を特徴量として採用している時の精度が 87.27 と高い精度が得られた．しかし，同様の実験を数回行った結果精度が安定しなかった．また，幼児の音声データ数もかなり少ないため，以上 2 点を考慮した結果，最適な機械学習手法ではないと判断した．

4.1　結果 1-機械学習モデルの比較-

最適な機械学習を判断するために mfcc を特徴量として抽出した場合のそれぞれの機械学習を用いた判別結果を以下に記す．

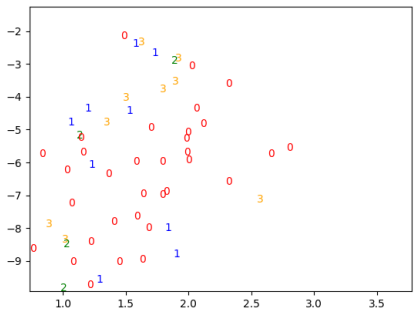


図 1　SVM(線型) を適用した結果

特徴量に mfcc を使用した際，精度は 80%と RDF が最も高い．しかし，図 3 より，可視化結果には 0 の笑い声の正解データと 3 の泣き声の不正解データのみが見られた．これは推定される感情が全て笑い声であると判断されてしまったことを意味する．図 2 より，SVM のガウスカーネルでも同様の結果が得られた．一方で，機械学習に SVM の線型カーネルを採用した場合は，図 1 より，全てのラベルが含まれており，推定された結果に笑い声も泣き声も含まれていることがわかった．従って特徴量 mfcc を使用した場合，精度は 80%とランダムフォレストが最も高いものの，精度と可視化結果を考慮した結果，最適な機械学習手法は SVM の線型カーネルが好ましいと判断した．

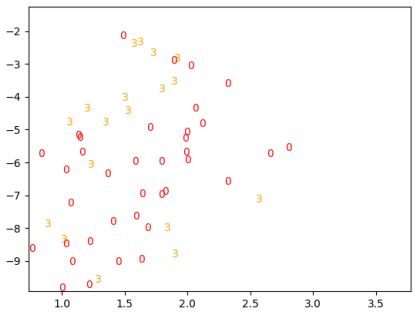


図 2　SVM(ガウス) を適用した結果

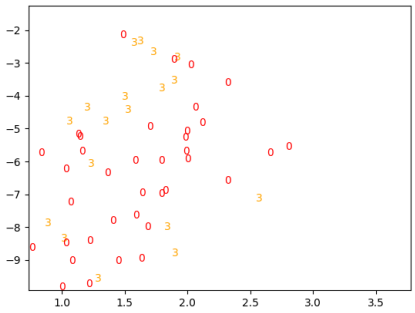


図 3　RDF を適用した結果

4.2　結果 2-特徴量抽出手法の比較-

最適な特徴量を判断するために前章で最適と判断した SVM の線型カーネルを使用して，特徴量抽出手法の比較実験を実施し，可視化した結果と考察を示す．表 1 にて最も精度の高かった特徴量 mel + mfcc を抽出した場合と，CNN-2D モデルを使用して特徴量を算出した場合とを比較することで，好ましい特徴量を判断することにする．

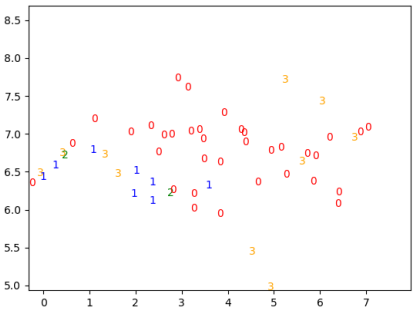


図 4　特徴量 mel，mfcc を抽出した実験結果.

図 4 より，特徴量に mel + mfcc を抽出したものは，可視化ラベル 1（泣き声の正解データ）の集合の近くに可視化ラベル 3（泣き声の不正解データ）が多く存在し，混同していることがわかる．SVM の線型カーネルはそれぞれの集合を直線で分類するものであるため，泣き声の正解データと不正解データが類似したものと判断されたことになってしまう．一方で図 5 より，CNN-2D モデルを用いて特徴量の算出を行った場合は，図左下に可視化ラベル 1（泣き声の正解データ）と可視化ラベル 2（笑い声の不正解データ）が集合していることがわかる．可視化ラベル 1 は泣き声と判断された推定結果であり，可視化ラベル 2 はテストデータが笑い声であるものの，泣き声と判断されてしまった推

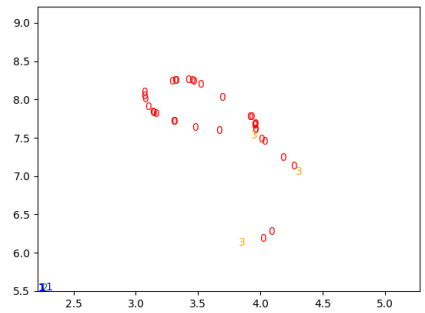


図 5　CNN-2D で特徴量を算出した実験結果.

定結果になるので，両者のラベルが混同していること，つまり両者のラベルが似ていることが正しいと言える．従って精度も最も高く，可視化した際も分類が妥当であることから，最適な特徴量として CNN-2D モデルを用いて特徴量算出を行うことが好ましいと判断した．

5　結論・今後の課題

本研究では，喃語を話す幼児を対象とし，幼児の声色から感情を判別する際に最適な特徴量と機械学習について議論した．その結果，笑い声と泣き声の判別において，CNN-2D モデルで特徴量を抽出し SVM で感情を判別する手法が最も好ましいと判断した．また，精度を数値で比較するだけでなく，精度を算出した上で t-SNE により判別結果を可視化することで，特徴量と機械学習の好ましさをより深く議論することができた．本実験において，判別結果を t-SNE で可視化することは有用であった．

今後の課題としては，笑い声・泣き声だけに限らず他の感情においても判別できるかを検討したい．例として，笑い声・泣き声に加えてイライラ（annoy）のラベルを追加したいと考えている．また，本研究に適用可能な学習済モデルを入手できるようであれば，それを用いた転移学習についても検討する．

謝　　辞

本研究はお茶の水女子大学の生物医学的研究の倫理審査を受けたものです．

研究にあたり貴重な助言を下さったお茶の水女子大学心理学科の山田美穂准教授と砂川芽吹助教，ソニー CSL の大和田茂研究員に，またデータ収集でお世話になった対象児のご家族の皆様に感謝致します．

文　　献

[1] 株式会社ファーストアセント．パパっと育児@赤 ちゃん手帳．https://papaikuji.info/.

[2] 門谷信愛希 al．特徴量の抽出．音声による感情認識システムに関する研究，pp.18–20，2000.

[3] 田中宏和 al．乳児の泣き声の収集とその分析ツール．第 66 回全国大会講演論文集，pp.281–282，2004.

[4] Ashwini K, P.M. Durai Raj Vincent, Kathiravan Srinivasan, and Chuan-Yu Chang. (2021). Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. Front Public Health. 2021;9:670352. Published online 2021 Jun 10. doi:10.3389/fpubh.2021.670352

絵文字を用いた短文の特徴と感情伝達
－感情集約モデルの構築－

宮下 大生† 塩野 直志†

†神奈川工科大学情報学部 〒243-0292 神奈川県厚木市下荻野 1030

E-mail: †s2021112@cco.kanagawa-it.ac.jp, na-shiono@ic.kanagawa-it.ac.jp

あらまし X などで絵文字を付与した短文を作成する際，各絵文字に対応する短文には必ずしも画一的な特徴があるわけではない．短文は幾つかの特徴に分類されると考えられる．また，類似した絵文字において作成される短文には同じ特徴と異なる特徴があると考えられる．そこで，本研究では絵文字を抜いた短文に予め Ekman の 6 感情の有無が与えられている前提のもと，6 感情を集約するモデルを新たに構築した．モデルより得られる結果を基に短文の傾向を分析することで，絵文字を用いた短文の特徴を整理した．

キーワード 絵文字，感情分析，感情集約モデル

1. はじめに

絵文字は，老若男女問わず日常生活の様々な場面でコミュニケーションツールの一つとして活用されている．絵文字が使われる理由の一つは，書き手が自らの感情を正確に伝えたいためである．X（旧 Twitter）をはじめとした短文のメッセージでは，読み手が誤解を生じないよう，書き手の意図や感情を顔文字や絵文字で補足することが行われている．

絵文字に関わる研究は，Bai ら[1]が取り上げているように，コミュニケーション，マーケティング，行動科学，言語学，心理学他多くの分野で行われている．国内においても各絵文字の特徴を比較するためのマッピング[2・3]，絵文字への感情ラベル付与[4]をはじめ多くの研究が行われている．特に，高橋ら[5]は絵文字の使用パターンを分析するとともに Smileys（😃）を使用する場合は書き手の感情が伝わりやすいことに言及している．ただし，これらの研究は，短文と絵文字の感情が一致しているという前提か，アンケートにより短文と絵文字の感情を分析する方法が取られている．

本研究は，アンケートを行わず X 上に書かれている短文と絵文字を基に，絵文字を用いた短文の特徴を整理することを目的とする．日常生活で短文に絵文字を付与するとき，同じ絵文字でも短文が意図する内容が異なることがある．よって，短文に必ずしも画一的な特徴があるわけではない．また，見た目が似ている絵文字において，同じような短文が存在する一方で，意図が明らかに異なるように見える短文が存在することがある．よって，絵文字に対応する短文を一つの特徴に整理するのではなく，複数の特徴を得るようなアプローチを行う必要がある．

そこで，本研究ではこれらの特徴を整理するために，短文を感情の有無に応じて集約する感情集約モデルを構築し，絵文字を用いた短文の特徴を整理した．

本論文では，まず 2 章で分析手順に関して述べ，3 章で感情集約モデルに関する詳細を説明する．その結果，感情ごとに異なる特徴を持つことが分かったため，4 章では各感情を三つに大分したうえで短文の特徴を考察し，5 章で結論を述べる．

2. 分析手順

感情分析に関して，sentiment_ja2[6]では，Ekman の 6 感情[7]に該当する絵文字を，喜び(😊)，悲しみ(😢)，嫌悪（🤢），怒り（😠），驚き（😮），恐れ（😨）とおき，X 上でこれら絵文字を付与している短文が該当する絵文字の感情と一致する前提でロジスティック回帰を実施している．また，その結果より，短文を入ると各 6 感情の値を 0 以上 1 以下で出力するソフトウェアを構築している．構築後，閾値を 0.5 と定め，短文入力後に感情あり(1)と感情なし(0)と分類したうえで評価を行い，高い適合率，再現率，F 値を得ている．

本研究では，まず絵文字を選定後，X から 200 前後の短文を抽出し sentiment_ja2 を使用することで感情の値を出力し，0.5 を閾値とし感情あり/なしを決定する．その結果，各短文で[喜び，悲しみ，嫌悪，怒り，驚き，恐れ]の 0-1 が作成される．しかし，この結果のみでは短文の特徴や類似した絵文字との特徴の相違が捉えきれない．そこで，本研究では 3 章で述べる感情集約モデルにより短文の特徴を大きく三つに集約する．この集約で各感情の有無を明らかにすることで，短文の特徴を整理することが出来る．

また，本研究では，特に喜びに近いと想定される絵文字を😊の他に 6 種類（🤣，😄，😆，😘，😁，😍）選択し，各々で X より 200 短文抽出したうえで感情分析を行い，集約する上で評価を行った．同時に，他の感情に関しては sentiment_ja2 で使用された絵文字以外に[2]を基に絵文字一つを選定し，同様の処理を行う．

3. 感情集約モデルの構築

本章では，各絵文字において約 200 短文の感情分析を実施した結果を研究の目的に沿う形で集約する感情集約モデルについて述べる．

感情分析を実施した結果，感情別にあり(1)，なし(0)が得られ，最大で64(= 2^6)通りの組み合わせが考えられる，この集合をIとおく．

次に，集約を行うにあたり，あり(1)，なし(0)に，あり・なしを包含するコード(2)を加えた組み合わせを考える．この場合，最大729(= 3^6)通りの組み合わせが考えられるが，本研究ではあり(1)を際立たせる点から

・6 感情のうち少なくとも 1 が一つ存在する

・6 感情すべてが 2 となる（あり・なしを包含する）

いずれかを満たすことを条件とし，得られる 666 通りの組み合わせからなる集合をJとおく．本研究では，以降これらの組み合わせを感情タイプと呼ぶ．

このとき，ある感情タイプ$i \in I$を含めることが出来る$j \in J$が複数存在する．例えば，図 1 において集合Iに属する 感情タイプ[1,1,1,1,1,0]は，集合Jにおいて[1,1,1,1,1,0]，[1,1,1,1,1,2]，･･･，[2,2,2,2,2,2]を満たす．

集合I

i	喜	悲	嫌	怒	驚	恐
1	1	1	1	1	1	1
2	1	1	1	1	1	0
…						
64	0	0	0	0	0	0

集合J

j	喜	悲	嫌	怒	驚	恐
1	1	1	1	1	1	1
2	1	1	1	1	1	0
3	1	1	1	1	1	2
…						
7	1	1	1	1	2	1
8	1	1	1	1	2	0
9	1	1	1	1	2	2
…						
665	1	2	2	2	2	2
666	2	2	2	2	2	2

図 1: 集合I・Jと(i,j)の作成

そこで，条件を満たすiとjを結んだ枝を(i,j)と表し，枝の集合をAとおく．以上の前提を基に定式化を行う．

まず，集約を行うのが目的のため，$j \in J$のなかで集約に使用するか否かのバイナリー変数y_jを設ける．

$$y_j \in \{0,1\}, \quad \forall j \in J.$$

なお，0 は集約に使用されず，1 は使用されることを意味する．

また，Jの中でm個が集約に使用されるとする．

$$\sum_{j \in J} y_j = m,$$

集約を行う際，$i \in I$から$(i,j) \in A$に属する$j \in J$が一つ選択されると考え，変数x_{ij}を設け

$$x_{ij} \in \{0,1\}, \quad \forall (i,j) \in A,$$

$$\sum_{\{j|(i,j)\in A\}} x_{ij} = 1, \quad \forall i \in I,$$

と制約を設ける．

ただし，選択されるjは必ず集約に使用されていないといけないため，

$$\sum_{\{i|(i,j)\in A\}} x_{ij} \leq M y_j, \quad \forall j \in J,$$

を満たす．ここでMは十分に大きな値である．

いま，ある感情タイプ$i \in I$が約 200 短文の結果で占める割合をa_iとおく．例えば図 2 において，感情タイプ[1,0,0,0,0,0]の割合が 46%である．この感情タイプはそのままか少し集約した形で残したい．一方で割合が非常に低い感情タイプにおいては出来るだけ一つに集約したい．そこで，集約を行う感情タイプ$j \in J$において，0 か 1 が含まれる個数を得点としてb_jとおく．例えば[1,1,1,1,1,0]は 6 となり，[1,2,2,2,2,2]は 1 となる．このとき，$a_i b_j x_{ij}$の和が最大となるよう目的関数を設定すれば，より高い割合の感情タイプが残る．

集合I

喜	悲	嫌	怒	驚	恐	割合
1	1	1	1	1	1	0%
		…				…
1	0	0	0	0	0	46%
		…				…
0	0	0	0	0	0	0%

集合J

喜	悲	嫌	怒	驚	恐	得点
1	1	1	1	1	1	5
	…					
1	0	0	0	0	0	5
	…					
2	2	2	2	2	2	0

図 2:目的関数の設定

以上より，感情集約モデルは以下の通り定式化される．

$$\max \sum_{(i,j)\in A} a_i b_j x_{ij}$$

$$\text{s.t.} \quad x_{ij} \in \{0,1\}, \quad \forall (i,j) \in A,$$

$$y_j \in \{0,1\}, \quad \forall j \in J,$$

$$\sum_{\{j|(i,j)\in A\}} x_{ij} = 1, \quad \forall i \in I,$$

$$\sum_{j \in J} y_j = m,$$

$$\sum_{\{i|(i,j)\in A\}} x_{ij} \leq M y_j, \quad \forall j \in J.$$

本モデルにおいて，得点を 0 と 1 の数で決めていることから，mを少なくするほど感情タイプ[2,2,2,2,2,2]に集約されていく$i \in I$が多くなると考えられる．

本モデルは整数最適化問題である．本研究では，m = 4, 即ち四つに集約するとして，Fujitsu LIFEBOOK UH03 (OS: Windows 10, CPU: Intel Core i7-1165G7 2.80 GHz, memory: 8 GB)に Gurobi 9.1.2 を実装し，絵文字別に求解した．その結果，いずれの絵文字においても 0.2 秒以内に最適解が得られた．また，いずれの絵文字でも集約された感情タイプの一つは[2,2,2,2,2,2]であった．

表 1:ポジティブ感情の集約結果（空欄は 2）

	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合
喜び😊	1	0	0	0	0	0	55%	1	0	0	0	1	0	10%		0	1		1		12%
笑い転げる🤣	1	0	0	0	0	0	22%	1	0			1		21%	0		1	1	1	1	29%
嬉し泣き😂	1	0	0	0	0	0	21%	1	0			1		28%	0		1		1	1	29%
きゃー😆	1	0	0	0	0	0	46%	1	0			1		21%	0		1		1	1	15%
投げキス😘	1	0	0	0	0	0	43%	1	0				0	26%	0		1	1	1	1	10%
にやっと😁	1	0	0	0	0	0	37%	1	0		0	1		21%			1	1	1	1	18%
目がハート😍	1	0	0	0	0	0	43%	1	0		0	1	0	18%	0		1		1	1	14%

4. 結果

感情集約モデルを各絵文字で実施したところ，Levenson and Ekman [8]で言及している，ポジティブ感情（喜び），ネガティブ感情（怒り，恐れ，嫌悪，悲しみ），および驚きで傾向が分かれた．その点から，本研究では三つの感情に分けて考察を行う．

4.1. ポジティブ感情

ポジティブ感情における集約結果を表 1 に示す．sentiment_ja2 において喜び(😊)でロジスティック回帰を実施していることから喜びが 1，他が 0 となる割合が多くなるのは当然の結果であるが，驚きを伴う場合と伴わない場合に分かれている．驚きを伴う場合は短文に”！”が付けられても違和感ない短文である．

本研究では，にっこりと類似していると考えられる絵文字 6 個を取り上げ，各々200 短文前後で同様の集約を行った．表 1 を見ると使用している短文の傾向が非常に類似していることが分かる．

まず，どの絵文字においても喜びが 1，他が 0 となる短文に使用される傾向がある．「頑張ります」「帰ります」などポジティブな要素を含む短文が該当している．喜びと驚きが 1，他が 0 となるのは前述したとおり”！”が付けられても違和感ない短文が該当する．

一方で，嫌悪と驚きが 1 となる短文で絵文字が使用される傾向がある．該当する短文を見ると「それね」「今だよね」など確認を促したり，「今はそれじゃない」と否定を含んだりする短文である．短文のみでは感情を読み解くことが出来ず，書き手がいまの感情を示すために絵文字を使用していると考えられる．

以上を踏まえると，ポジティブ感情を示す絵文字は使用する短文の傾向が似ている点，ポジティブな短文に対して強調の意味を込めて絵文字を使用している点，一方で確認や否定の短文では自分の感情を明確にするために絵文字を使用している点，などが挙げられる．

4.2. ネガティブ感情

一方．ネガティブ感情（怒り，恐れ，嫌悪，悲しみ）において，集約すると傾向が非常に似ている特徴と絵文字ごとの特徴に分かれた．

まず，主に喜び 0 で他感情の多くが 1 という特徴が得られた．これらの短文は「困った」「仕方ない」「反省している」他ネガティブな要素を含んでいるものの，書き手の具体的な感情がわからない短文である．

続けて，主に悲しみ 0 で（喜びを除く）他感情の多くが 1 という特徴が得られた．これらの短文は「違う」「強い」など悲しみを表す絵文字はつかないと考えられるものの，短文のみでは書き手の具体的な感情が分からない短文である．

さらに，絵文字ごとに結果が明確に出ている結果があり，例えば悲しみ（😢）においては「酷い」，「儚い」などで悲しみと恐れが 1 となり，嫌悪（🤢）においては「エグい」，「吐き気がする」などで嫌悪が 1 となる．

その一方で，主に喜び 1 で他感情の多くが 0 という特徴も得られた．これは喜びで使われやすい「お疲れ様」に使われる傾向（😔），マスクを付けている状況を表現（🙄）などである，

以上を踏まえると，ネガティブ感情を示す絵文字は，ネガティブということはわかるが感情は特定出来ないため絵文字によって感情を表現する短文が多い一方で，短文そのもので感情の表現が出来る短文は，その感情を強調するために当該の絵文字を使用する傾向があることがわかる．

4.3. その他（驚き）

驚きを示す絵文字に関しては，ポジティブ感情の要素を含んだ使い方とネガティブ感情の要素を含んで使い方に分かれる．

ポジティブ感情の要素として、喜び 1 の一方，悲しみ・嫌悪・怒りが 0 となり，ベンチマークとして使用した驚き（😮）においては驚きの感情が 0 である．

表 2:ネガティブ感情の集約結果（空欄は 2）

	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合
悲しみ😢	**0**	**1**	**1**	**1**	**1**	**1**	33%		**0**	**1**	**1**	**1**		12%	0	**1**				**1**	25%
しょぼーん😞	**0**		**1**	**1**	**1**	**1**	33%	**1**	**0**	**0**		**0**	**0**	15%	0	**1**		0			26%
嫌悪🤢	**0**	**1**	**1**	**1**	**1**	**1**	29%		**0**	**1**	**1**	**1**	**1**	19%	0		**1**	0		1	15%
マスク顔😷	**0**		**1**	**1**	**1**	**1**	18%	**1**	**0**	**0**	**0**		**0**	30%	0			0	0	1	20%
怒り😠	**0**	**1**	**1**	**1**		**1**	33%		**0**	**1**	**1**	**1**	**1**	20%	**1**	**0**				**0**	15%
むかっ💢	**0**	**1**	**1**	**1**	**1**	**1**	26%	0	**0**	**1**	**1**	**1**	**1**	27%	1	0	1	**1**	1		13%
恐れ😨	**0**	**1**	**1**	**1**	**1**	**1**	32%		**0**	**1**	**1**	**1**	**1**	16%	0	1		0		**1**	23%
恐怖😱	**0**	**1**	**1**	0		**1**	15%	0		**1**	**1**	**1**	**1**	40%	**1**	**0**	**0**	**0**		**0**	13%

表 3:驚きの感情の集約結果（空欄は 2）

	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合	喜び	悲しみ	嫌悪	怒り	驚き	恐れ	割合
驚き😲	**1**	**0**	**0**	**0**	0	0	15%	**0**		1	1	**1**	1	33%	1	**0**			**1**		21%
びっくり😱	**1**	**0**	**0**	**0**			31%	**0**	1			**1**	1	24%		**0**	1		**1**	1	20%

これらの短文は驚きの感情を表すために驚きの絵文字を使用しているのではなく，喜びという感情を強調するために驚きの絵文字を使用していると考えられる．

一方，ネガティブの感情の要素としては，主に喜びが 0 で驚きが 1 の場合，悲しみが 0 で驚きが 1 の場合に分かれるが，ともに短文のみでは感情を特定することができず，驚きの感情を示すために絵文字を使用していると考えられる．

以上より，驚きの絵文字においてはポジティブな要素では喜びを強調するため，ネガティブな要素では驚きという感情の明示を行うため，と異なる目的で絵文字が使用されているのではと考えられる．

5. まとめ

本研究は X から抽出した 200 短文に対して sentiment_ja2 を用いて感情分析を行い，得られた結果を感情集約モデルにより感情の集約を行うことで絵文字を用いた短文の特徴を把握した．その結果，ポジティブ 1 感情，ネガティブ 4 感情，他 1 感情に傾向が分かれること，ポジティブを表すと考えられる絵文字は使用する短文の傾向が似ている点，ポジティブな短文は強調のために絵文字を付与する点，ネガティブな短文は短文自体では感情が特定できず絵文字をつけることで感情が把握できる点などが成果として得られた．

なお，本研究は X から抽出した短文を基に分析を行っている．短文という点からネガティブな感情に対して感情が特定できない傾向が出るのであり，文章が長くなるにつれて書き手の感情を読み取ることができることを留意する必要がある．

今後の課題として，本研究は sentiment_ja2 前提での分析であったが，自らデータを抽出し，分析を実施していくこと，集約モデルで設定した数値および集約数を変えてより精度を向上させていくこと，さらにはネガティブ感情の絵文字に関して各感情で 2 絵文字のみであり，もう少し対象の絵文字を増やして分析を行うことが挙げられる．

参　考　文　献

[1] Q. Bai, Q. Dan, Z. Mu and M. Yang, “A systematic review of emoji: Current research and future perspectives”, Frontiers in psychology, 10:2221, 2019.

[2] 木村真有，桂井麻里衣，”ツイート集合を用いた絵文字の感情空間へのマッピング”，第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM 2016), No. P6-4, 2016.

[3] 津村大悟，北嶋弓月，大竹恒平，生田目崇，“絵文字を用いた感性情報の推定”，日本オペレーションズ・リサーチ学会 2023 年春季研究発表会アブストタクト集，pp.98-99, 2023.

[4] 郷間豊人，李吉屹，福本文代，郷健太郎，”Twitter を用いた絵文字の感性と感情ラベル付与”，第 15 回データ工学と情報マネジメントに関するフォーラム(DEIM 2023), No. 1b-2-4, 2023.

[5] 高橋直己，上野舞夕，浜田百合，庄司裕子，”絵文字を用いた文章における感情伝達効果に関する研究”，日本感性工学会論文誌，Vol.21, No.1, pp. 135-142, 2020.

[6] sentiment_ja2, https://github.com/sugiyamath/sentiment_ja2.

[7] P. Ekman, "An argument for basic emotions." Cognition and emotion, Vol.6.3-4, pp. 169-200, 1992.

[8] W. Levenson and P. Ekman, "Difficulty does not account for emotion-specific heart rate changes in the directed facial action task." Psychophysiology, Vol. 39, No.3, pp.397-405, 2002.

歌詞感情分類データセット作成のための大規模言語モデル利用法

Using Large Language Models to Create Lyric Emotion Classification Datasets

宮川　尚†　宇津呂武仁††

† 筑波大学大学院 システム情報工学研究群 知能機能システム学位プログラム
〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1

あらまし　本論文では，日本語の歌詞を対象とした大規模な感情分類データセットが存在しないという問題に対処するため，大規模言語モデルを使用して歌詞の感情分類を行う手法を提案する．近年の言語モデルは，インターネット上の膨大なテキストデータから学習を行い，その結果，対話能力や応用範囲が飛躍的に向上している．特に，高いパフォーマンスが期待される ChatGPT を用いて歌詞の感情分類を実施した．評価結果として，まず単純な肯定・否定の二値分類においては，96%という高い分類性能を確認することができた．さらに，肯定感情に焦点を当て，より詳細な分類を行った結果，肯定感情を 4 つのカテゴリーに分類する際，1 つのカテゴリーで 90%の分類性能を達成し，全体としての平均再現率は 73%であった．感情分類性能を比較する評価では，GPT-3.5 よりも GPT-4 がより優れた結果を示し，その適合性が明らかになった．一方で，否定感情の分類においては，肯定分類ほどの高い再現率は達成できなかったものの，誤判定の傾向を分析し，再現率が改善される可能性を示した．

キーワード　LMM，テキスト分類，歌詞分類，感情分類

1 はじめに

近年の音楽サイトやアプリケーションでは，特定の感情に基づいて音楽を検索すると，それに合致したプレイリストが提供され，これらはサイトやアプリケーションの運営者，またはユーザーによってまとめられている．また，楽曲推薦システムの研究 [3,5,6] は多岐にわたり，音楽の感情分類への需要が高まっている．一方で，音楽はいくつかの要素で構成されており，例えば，プレイリストをまとめたユーザーが重視しているのがテンポかリズムなのか，音色なのか，あるいは歌詞なのかといった感情分類の基準が不明確であることがある．このように，音楽の印象は高い主観性を持つと考えられる．そこで，音楽要素をまとめて感情分類するのではなく，要素ごとに分解し，まず各要素に対して正確な感情分類を行う必要があると考える．よって，本論文ではまず音楽の重要な一要素である歌詞に焦点を当て歌詞における感情分類を行う．

ここで，機械学習などの技術を用いて，大規模に歌詞データを感情分類する方法が考えられるが，現在，日本語の歌詞が感情分類された大規模なデータセットは存在していない．そのため，訓練データが不足しており，歌詞の感情分類が難しいという課題がある．この問題に対処するため，歌詞の感情分類を自動化し，大規模な分類データを生成する手法として，歌詞と辞書データベースを照合し，独自のスコア付け手法により歌詞感情分類を行う方式 [10] や，感情単語との類似度を計算して歌詞の感情分類を行う方式 [11] が提案されている．

これに対して，本論文では，日本語歌詞に関するデータセットを作成することを目的として，大規模言語モデルを用いて歌詞の感情分類を行う．近年の大規模言語モデルは，膨大な量のテキストデータからの学習を通して，対話能力および汎用性を高めていることが知られている．本論文では，大規模言語モデルとして，特に性能が優れていると考えられる ChatGPT[1] を採用し，歌詞の感情分類タスクに適用する．感情分類としては，まず最初に，肯定・否定間の 2 値分類を行ったところ，小規模評価では 96%の再現率を達成した．更に詳細な感情分類を行う場合の ChatGPT の性能を検証するため，肯定的な感情と否定的な感情をそれぞれ 5 分類ずつ設定した．ただし，肯定的感情分類の 5 分類間の分類性能においては，誤判定が多く見られた．そこで，互いに誤判定が多い 2 つの分類を統合することにより，どの程度再現率が改善するかの評価を行った．その結果，肯定的感情分類の 4 分類間では，全体として 73%の再現率を達成した．特に，肯定的感情の 4 分類のうちの 1 分類においては，90%の再現率を達成した．次に，否定的感情 5 分類の分類性能を検証した．平均再現率としては，肯定感情に比べ劣るが，肯定感情と同じように，誤判定された歌詞の分類先が共通する傾向を確認できた．

2 関連研究

本論文の関連研究として，歌詞の感情分類に関する先行研究が挙げられる．例えば，Malheiro ら [4] は，音楽 Web サイトの AllMusic[2] において各楽曲に付与された感情単語を二次元

1：https://chat.openai.com/
2：https://www.allmusic.com/

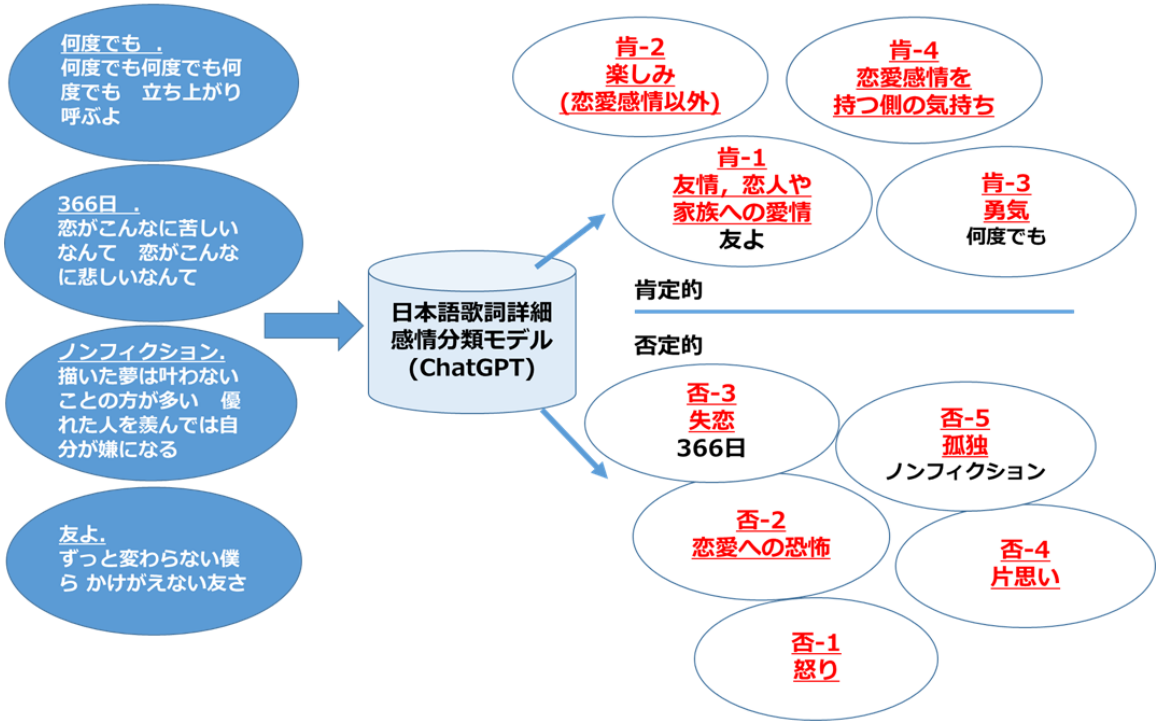


図 1: 歌詞感情分類の流れ

表 1: 感情分類とその詳細説明

(a) 肯定的感情分類

番号	感情分類	詳細説明	曲数
肯-1	友情	友を思いやる言葉が多く友情を歌詞にしたもの	5
	双方向愛情 (恋人や家族への愛情)	一方向ではなく双方の愛が感じられる歌詞であることがポイント	5
肯-2	楽しみ (恋愛感情以外)	走ったり踊ったり気持ちが行動に移る言葉が多い	10
肯-3	勇気	人や自身を鼓舞する言葉を歌詞にしたもの	10
肯-4	片方向恋愛 (恋愛感情を持つ側の気持ち)	恋愛感情は一方向で相手が同じ感情を持っているかは分からない	10
合計			40

(b) 否定的感情分類

番号	感情分類	詳細説明	曲数
否-1	怒り	相手の行動や言動に対する怒りを表した言葉で強い口調である	10
否-2	恐怖 (恋愛への恐怖)	気持ちに反して相手を信じることや自分に好意を持たれることを拒む	10
否-3	失恋	大切な人を失い悲しんでいる歌詞	10
否-4	片思い	相手への好意を伝えられないもどかしい気持ちを歌詞にした	10
否-5	孤独	心を満たした意欲があるがそれが出来ない状態を言葉にしている	10
合計			50

空間に写像しており，これを用いて，二次元空間の各象限 (計 4 象限) からそれぞれ，400 曲ずつの歌詞分類データを作成し，最終的には，人手による歌詞感情分類の判定を行い，771 曲の歌詞感情分類事例を作成，公開した．Cano ら [1] は，3 つの辞書を用いて，歌詞に Vt(感情価) と At(覚醒度) の二つの値をスコアリングし，歌詞を二次元空間に写像し分類を行って，2,595 曲の 4 つの mood に分類されたデータセットを作成，公開した．一方，日本語歌詞の感情分類研究としては，Xu ら [11] が挙げられる．Xu らは，歌詞テキストと感情単語辞書中の単語の類似度を計算し，これに基づいて歌詞の分類を実施している．た

だし，これらの研究においては，感情にあった曲は検索できるが，状況にあった曲には対応してない点が課題である．

歌詞に対して状況的分類を行った研究として，杉浦ら [10] においては，歌詞に対して,「海で聴きたい曲」,「卒業で聴きたい曲」といった状況に応じた分類を行っている．また，倉持ら [12] においては，YouTube の 686 曲の音楽動画に対する 265,317 件のコメントを収集し，コメントをベクトル化した後，k-means 法によって 1,000 個のクラスタを作成し，その 1,000 個のクラスタから楽曲の状況を表すクラスタを同定し人手で各状況タグに選別した．構築された状況タグにおいては，朝・夜等の時間，クラブ・パーティー・ジム等の場所，といった自分以外の状況はうまくとらえられている．その一方で，恋愛状況における片思い・失恋等の自分中心の状況は捉えられていない点が課題である．

Revathy ら [7] は英語の歌詞が 4 つの感情に分類された MER データセット [2] を用い，転移学習を行った．この過程で，感情分類が行われていない Music4all データセット [9] にアノテーションを施した．具体的には，ラベル付けされた Music4all データセットを用いて事前学習された BERT の fine-tuning データに利用し，BERT における歌詞の感情分類精度を検証した．

さらに，大規模言語モデルを用いたデータセット生成研究として，Alpaca[3] が挙げられる．彼らは LLaMA を fine-tuning した言語モデルを用い，*text-davinci-003* による self-instruct で生成された 52K の命令追従型の学習データを利用して学習を行った．

本論文では，日本語歌詞感情分類データセットの構築を目指している．しかし，システムや言語モデルを一から構築するのではなく，現状で最も汎用性があり対話能力のある大規模言語モデルを活用し，歌詞の感情分類を自動的に行うことを主眼としている．

3 歌詞データセットおよび感情分類

本節では，本論文で作成したデータセットの収集方法，および，感情分類の種類と分類基準について述べる．

3.1 歌詞データセット

本論文において収集した歌詞の数を表 1 に示す．肯定的なものと否定的なものを合わせて，合計 90 曲分の歌詞を収集した．各感情分類について，10 曲分の歌詞を収集し，これらをすべて感情分類の評価に使用した．歌詞の収集は手動で行われ，歌詞の内容が表 1(a) および表 1(b) に示される感情分類に該当する場合には，適切な感情分類を歌詞に割り当てた．

3.2 歌詞の感情分類

感情分類の詳細を表 1 に示す．本論文の初期段階では，ラッセルの円環モデル [8] を参考にし，感情単語に基づいて歌詞を分類していた．しかし，歌詞には単一の感情だけでなく，複数の感情が存在しており．単一の感情単語では表現しきれないとの結論に至った．そのため，表 1(a) と表 1(b) に示すように，肯定感情と否定感情それぞれについて，5 つの独自の感情分類を設定した．本論文ではまず，肯定的感情分類に焦点を当て，再現率を 100%にする手法を確立することを目標にした．肯定感情分類の評価では，5 つの感情分類を用いた歌詞の感情分類評価を行った．しかしながら,「友情」と「家族や恋人への愛情」間の誤分類が非常に多かったため，これらのカテゴリーを統合し，4 つの感情分類のもとで再現率を改善することを目標とした．その後，否定感情分類についての評価を行った．この評価では感情分類数の変更は行わず，5 つの感情分類を用いたままの評価結果を報告する．

4 大規模言語モデルによる歌詞感情分類

本節では，大規模言語モデルによる歌詞感情分類を行うための，プロンプトの与え方や，各種評価結果を述べる．評価に用いたモデルは，GPT-4 (*gpt-4-1106-preview*)，GPT-3.5 (*gpt-3.5-turbo-1106*) である．

4.1 基本的なプロンプトの与え方

本論文では，ChatGPT を活用した歌詞の感情分類のためのプロンプト設計について述べる．図 2(a) および図 2(b) に，具体的なプロンプトの例を示す．入力としては，参考となる例示歌詞を提供し，感情分類の対象となる歌詞が例示歌詞のいずれと同じ感情に分類されるかを ChatGPT に番号で回答させる方式を採用している．例示歌詞の分類名を直接与えると ChatGPT がそれを解釈してしまうため，分類名は記号で示し，感情分類の定義も直接的な言葉ではなく，例示歌詞を通じて与えることとした．図 2(b) には，4 つの肯定的な感情分類の中から 1 つを選択する 4 値分類問題のプロンプトを示し，図 2(a) には，4 つの肯定的な感情分類のうち 2 つを例示し，その中から 1 つを選択する 2 値分類問題のプロンプトを示す．さらに，ChatGPT の特性上，同一のプロンプトに対しても一貫した出力が得られない場合がある．このため，単一のテストケースに対して 1 回の試行だけでは，偶然に正しい分類が得られた可能性がある．この問題に対処するため，各テストケースに対して 3 回の試行を行い，全ての試行で正しい分類が得られた場合にのみ，分類が成功したと判断する基準を設けた．

4.2 例示歌詞・感情分類対象歌詞の与え方

次に，前節で作成したプロンプトに与える例示歌詞と，感情分類対象歌詞について説明する．まず，例示歌詞としては，各感情分類に対して 2 曲分の歌詞を選定した．この際，歌詞全文を提供するのではなく，感情が強く表現されていると思われる部分を人手で選び，約 250 文字に抜粋した．感情分類対象歌詞については，単純な方法として 1 曲分の歌詞全文を提供することも考えられるが，本論文では，約 500 文字（最少 214 文字，最大 860 文字）の全歌詞をそのまま使用する方法と，比較のために適切な 250 文字を人手で抜粋する方法の両方を用いて，ChatGPT の歌詞感情分類の再現率を評価した．これは，歌詞全体を提供する方法よりも，抜粋した部分のみを提供する方が

3 : https://crfm.stanford.edu/2023/03/13/alpaca.html

性能が向上すると考えられるため，感情分類がどの程度可能かを検証するための手法となる．

4.3 肯定的感情の歌詞の感情分類

4.3.1 2 値 分 類

本節では，肯定分類の 2 値分類の評価を行う．本評価は 2 値のため，組み合わせテストの技法であるペアワイズ法を用いて，1 つの感情分類対象歌詞に対し，全ての組み合わせを行った．なお，2 値分類における ChatGTP の特性として，分類が困難な問題において，図 2(a) の番号 1 に配置された感情分類を優先して選択する傾向があることが明らかになった．つまり，ChatGPT は自信のない問題では誤ってでも番号 1 を選択し，自信がある時は必ず番号 2 を選択する傾向がある．よって，例示歌詞の配置による分類の有利・不利が起きないように，一つの感情分類対象歌詞を分類するにあたっては，4 分類全ての例示歌詞を図 2(a) の番号 2 に配置したテストを行う必要がある．よって，ペアワイズ法において，1 つの感情分類対象歌詞に対するテスト数としては，

- 図 2(a) の番号 2 には 4 種類の分類が配置できるため ${}_4C_1$ 通り，
- 図 2(a) の番号 1 には，番号 2 に配置した分類以外の 3 種類の分類が配置できるため ${}_3C_1$ 通り，

となることから，次式となる．

$$ {}_4C_1 \times {}_3C_1 = 12 \text{ テスト} $$

評価方法としては，多数決を採用した．すなわち，全 12 のテストのうち，最も出力回数の多い感情分類を，感情分類対象歌詞に対する ChatGPT の最終的な解答と定めた (多数決によって 1 つの解答に定まらない場合は，その解答を不正解とした)．評価結果を図 3 に示す．この結果から，感情分類対象歌詞に歌詞全文を与えた場合よりも，250 字程度に人手で抜粋した図 3(b) の方が，平均再現率が 10%高いことが明らかになった．これは，感情分類対象歌詞が歌詞全文となると，歌詞の持つ情報が多くなり，分類問題が難しくなるためであると考えられる．また,「勇気」に関しては，他の感情分類に比べて，歌詞に励ましや挑戦的な特徴が現れやすいと，アノテーションを行っていた際に感じていたが，再現率は低かった．その理由としては，「楽しみ」の歌詞には自身や他人を励ますような内容は含まれていないが，どちらの歌詞も行動的な内容であることで,「勇気」の歌詞が「楽しみ」の歌詞と類似していると判定される可能性が考えられる．評価における混同行列を図 7(a)，および，図 7(b) に示す．実際に，参照感情分類が「勇気」の歌詞は,「楽しみ」に誤答されるものが多かった．最も再現率が高かった感情分類は,「友情」,「家族や恋人への愛情」であり，90%の再現率を達成している．図 3(b)，肯-1 に示されるように，分類統合により，実質的に 2 つの感情分類分の例示歌詞を与えているため，他の感情分類よりも幅広い分類の特徴を獲得できていると考えられる．

4.3.2 複数人による感情分類の一致度 (肯定)

本節では，人手 (本論文の第一著者) で作成した参照用の歌詞感情分類に対して，1) ChatGPT との間で，感情分類予測結果がどの程度一致するか，および，2) 別の評価者との間で，感情分類予測結果がどの程度一致するか，の評価を行った [4]．評価指標としては，再現率，および，カッパ係数を用いた．評価結果を表 2 に示す．ChatGPT は，評価者 2 とほぼ同等の再現率・カッパ係数を示した．また，ChatGPT は，評価者 1 を上回る再現率・カッパ係数を示した．この結果から，ChatGPT による感情分類予測性能は十分に高いと言える．

表 2: 一致度の評価 (肯定)

	ChatGPT	評価者 1	評価者 2
再現率	73%	63%	75%
カッパ係数	0.6333	0.5000	0.6667

表 3: ChatGPT の API の利用料金 (入力文字数が約 1,350 文字の場合)

model	利用料金
gpt-4-1106-preview	$0.015
gpt-3.5-turbo-1106	$0.0015

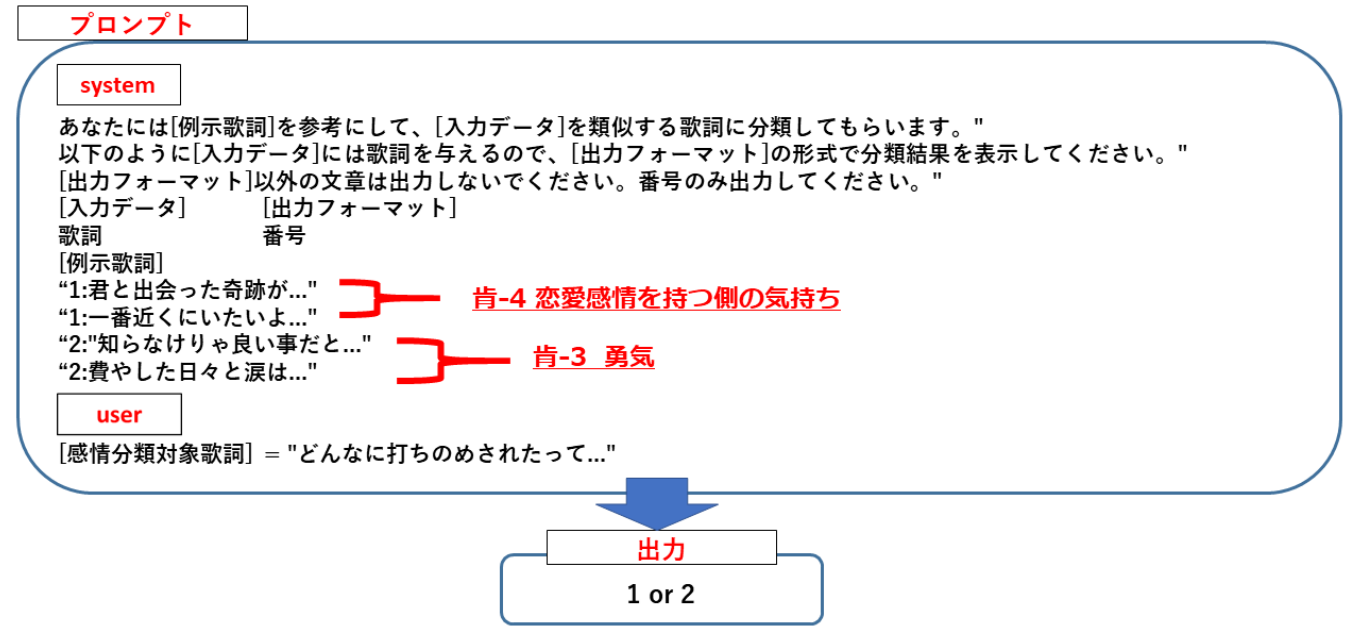
4.3.3 4 値 分 類

本節では，肯定的分類の 4 値分類の評価を行う．この実験では，2 値分類とは違い，例示歌詞の配置を変更していない．その理由は，配置を変更して行った小規模な評価で，2 値分類の場合と同じような，配置によって特定の位置が優位になる現象が確認されなかったからである．したがって，この 4 値分類では，図 2(b) に示すように，肯-1 から肯-4 までを順に配置している．評価結果を図 4 に示す．前節の 2 値分類の評価結果と類似している点は,「勇気」の再現率が低いことである．これは，2 値分類での「勇気」に対する考察と同様の現象が起きていることが原因と考えられる．実際に，図 4(b)，および，図 4(a) に示されるように，参照感情分類が「勇気」となる歌詞の誤答のほとんどは,「楽しみ」への誤分類であった．一方，2 値分類の評価結果と異なる点は，感情分類対象の歌詞を全文与える場合でも，250 字程度に人手で抜粋する場合でも，平均再現率に変化がないことである．また，肯-1(「友情」,「家族や恋人への愛情」) の再現率が大きく悪化している．傾向としては，図 4(b)，および，図 4(a) を見ると，誤答した歌詞のすべてが「片方向恋愛 (恋愛感情を持つ側の気持ち)」に誤分類されていた．逆に，参照感情分類が「恋愛感情を持つ側の気持ち」となる歌詞の誤答のすべてが，肯-1(「友情」,「家族や恋人への愛情」) に誤分類されている．このように，肯-1 と肯-4，肯-2 と肯-3 が互いに誤答し合う関係性があるため，これらを統合することで，分類数は減少するが分類性能が向上すると考えられる．

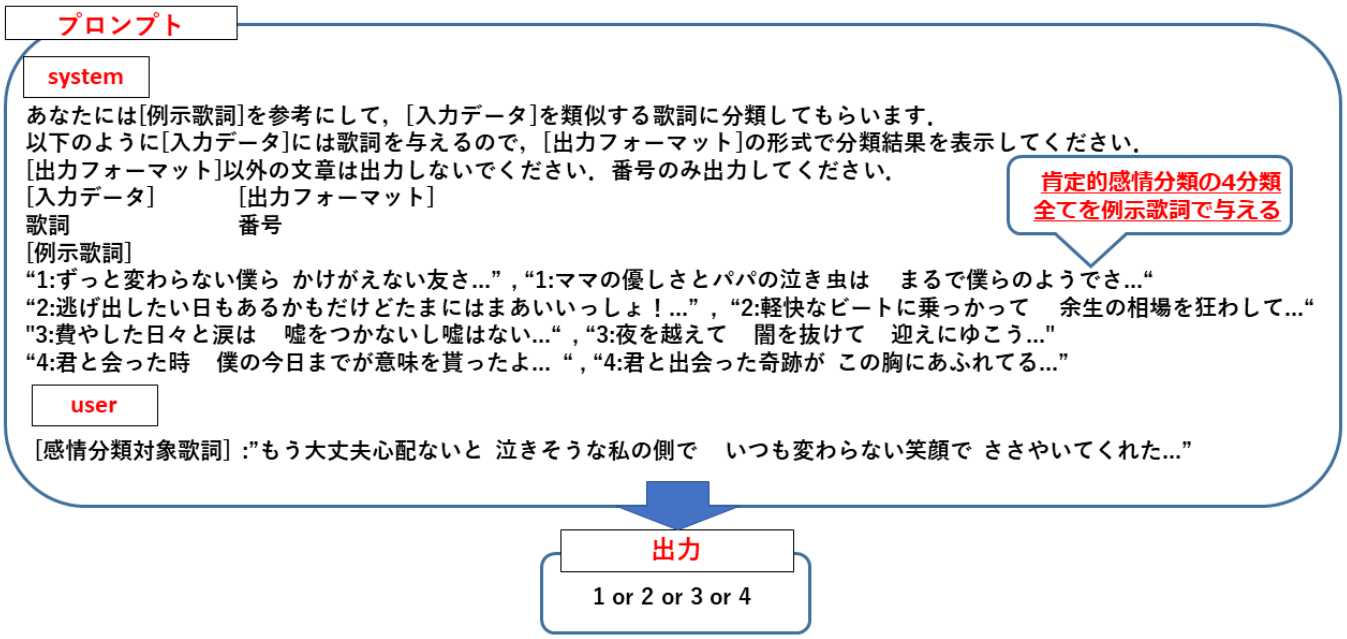
4.3.4 GPT-3.5・GPT-4 間の性能比較

本節では，GPT-3.5，GPT-4 の性能比較を行う．大規模な

4：各感情分類あたりの曲数が予測されることを防ぐために，肯定・否定の全 9 感情分類の全 90 曲に加えて，感情分類の分布を無作為にとったその他 60 曲を追加したデータセットに対して，評価者による感情分類予測を行った．



(a) 2 値分類のプロンプト



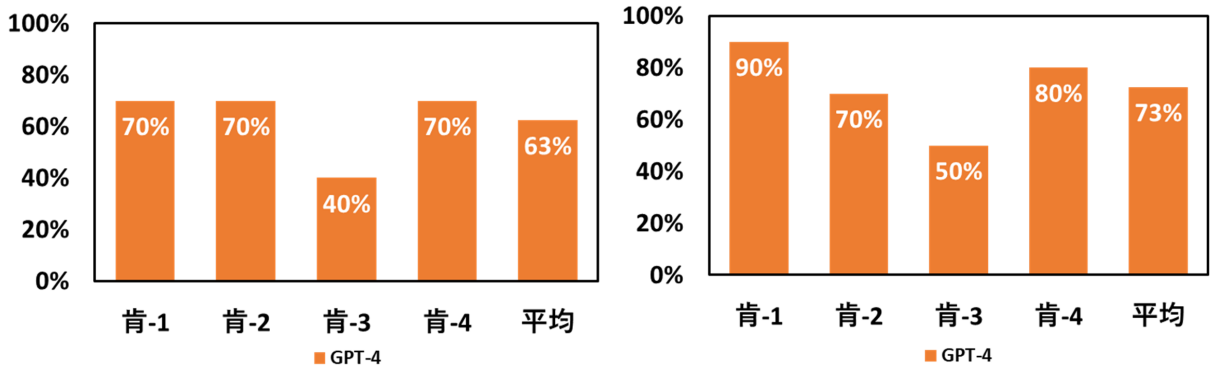
(b) 4 値分類のプロンプト

図 2: プロンプト (各感情分類ごとに 2 曲例示) および ChatGPT による出力例

データセットを作成する際，GPT の API 利用料金を考慮する必要がある．表 3 は，「適切な 250 文字を人手で選定」した時の，GPT の API 利用料金を示す．また，文字数の計算式を以下に示す．

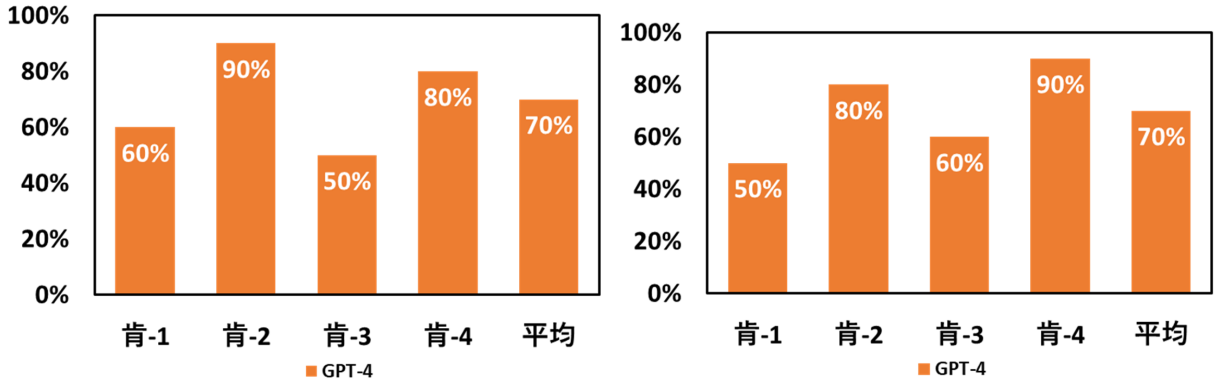
例示歌詞文字数 250 × 例示歌詞数 4 + 分類対象歌詞文字数 250 + 指示文の文字数 100 = 1,350 文字

ここでは，両モデルが同等の性能を持つ場合に，コストが 10 分の 1 である GPT-3.5 を採用するという前提の下，2 つのモデルの評価を行った．実験設定は図 3(b) に準じている．評価結果を図 5 に示す．GPT-3.5 を用いた歌詞感情分類の再現率の全体平均は 23%であり，目標とする再現率に対して非常に低い性能であることが確認された．各感情分類における再現率は，図 3(b) の評価結果と同様に，肯-1(「友情」および「家族や恋人への愛情」) が最も高く，「勇気」が最も低い結果となった．一方，GPT-4 の全体平均は GPT-3.5 よりも 50%高く，歌詞感情分類における両モデルの性能差は顕著であることが判明した．さらに，GPT-4 と異なり，GPT-3.5 では 2 択の選択肢から外れた



(a) 感情分類対象歌詞は全文を使用　(b) 感情分類対象歌詞は 250 字分を使用

図 3: 評価結果 (肯定的感情の歌詞，2 値分類，例示歌詞は 250 字分を使用，再現率)



(a) 感情分類対象歌詞は全文を使用　(b) 感情分類対象歌詞は 250 字分を使用

図 4: 評価結果 (肯定的感情の歌詞，4 値分類，例示歌詞は 250 字分を使用，再現率)

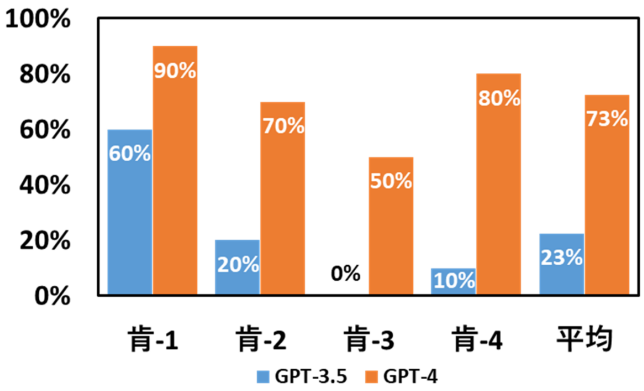


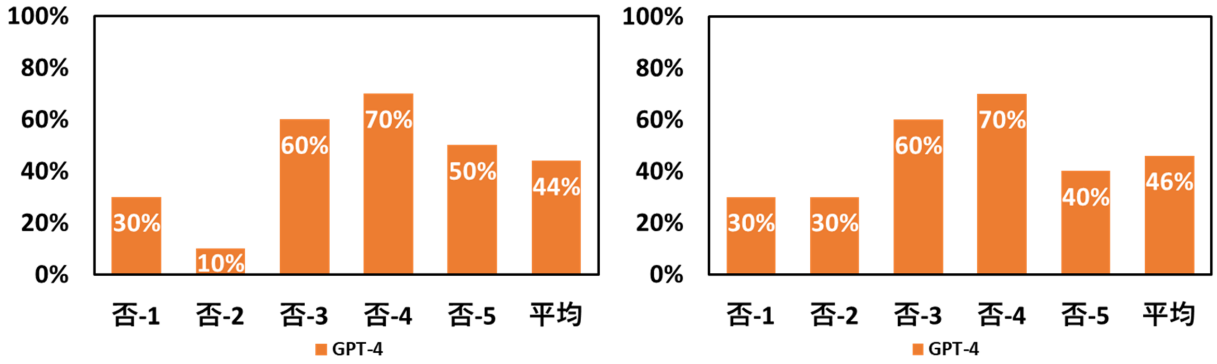
図 5: GPT-3.5・GPT-4 間の性能比較 (図 3(b) の実験設定)

番号を出力したり，番号以外の不要な文を出力することがあり，プロンプト文の理解や出力の安定性において GPT-4 に比べて著しく劣ることが明らかになった．したがって，本論文の歌詞感情分類タスクにおいては，コスト面を考慮しても GPT-4 のモデルを採用することが適切であると結論づけられる．

4.4 否定的感情の歌詞の感情分類

本節では，否定分類における ChatGPT の性能について述べる．実験設定は，肯定分類で最も再現率の高かった 2 値分類，例示歌詞 250 字，感情分類対象歌詞 250 字の設定と同一である．組み合わせテスト技法はペアワイズ法を採用し，評価は多数決により行なった．評価結果を図 6 に示す．肯定分類と比較すると，再現率はかなり低いが，肯定分類と同じように，誤答しやすい感情分類の組がいくつか存在することが明らかになった．否定的感情分類の混同行列を 図 7(e)，および，図 7(f) に示す．まず，参照感情分類が「怒り」となる歌詞の誤答先として「失恋」が最も多いことが分かる.「怒り」の歌詞では，怒りや憎しみが主に表現されているのだが，その原因として恋人との別れが前置きとしてあると，悲しみの感情があまり表現されてないのにも関わらず,「失恋」に分類される傾向にあると考えられる．また，参照感情分類が「失恋」の誤答先として，ほとんどが「片思い」に分類されていることが分かる.「失恋」の歌詞では，恋人との別れから悲しみの表現は強く表現されているのだが,「会いたい」,「好き」という感情が「片思い」の歌詞と類似していることから，誤答される可能性があると考える．これらの事実から，否定分類においても感情分類の統合が必要であり，統合を行うことで再現率が向上する可能性があると考えられる．

次に，表 4 に示すように，4.3.2 節で述べた「複数人による感情分類の一致度 (肯定)」の手順によって，否定的感情分類の歌詞データセットを対象として，人手・人手間，および，人手・



(a) 感情分類対象歌詞は全文を使用　　(b) 感情分類対象歌詞は 250 字分を使用

図 6: 評価結果 (否定的感情の歌詞，2 値分類，例示歌詞は 250 字分を使用，再現率)

表 4: 一致度の評価 (否定)

	ChatGPT	評価者 1	評価者 2
再現率	44%	48%	46%
カッパ係数	0.3000	0.3689	0.3602

ChatGPT 間の歌詞感情分類の一致度 (再現率・カッパ係数) の評価を行った．肯定的感情分類と比較すると，再現率・カッパ係数の絶対値は低くなるが，人手・ChatGPT 間の比較において，相対的な差はほとんどなかった．なお，参照用感情分類とは別の評価者による判定結果においては，否定的感情の歌詞を肯定的感情に誤分類している事例も観測された．

5 おわりに

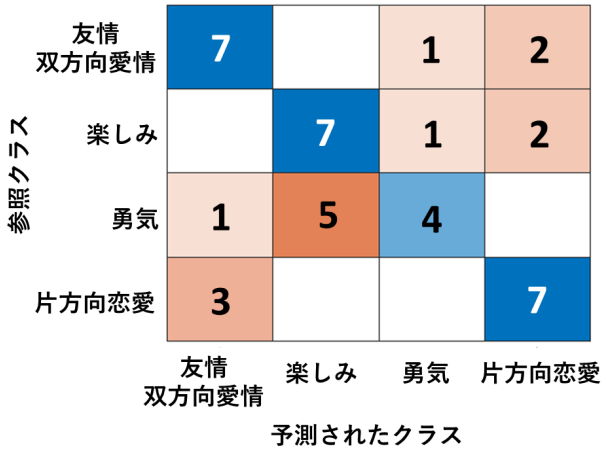
本論文では，大規模言語モデルが歌詞の感情分類をどの程度行えるのかを明らかにするため，ChatGPT を用いて日本語の歌詞感情分類の性能を評価した．まず，ChatGPT の性質として，2 値分類の場合，例示歌詞の配置によって，先に例示する感情分類に優位性が生まれることを確認し，本原稿の範囲では，ペアワイズ法によって，1 曲の歌詞に対して全ての組合せテストを行い，多数決評価をすることで，配置による優位性に対処した．その結果，分類によっては 90%の再現率で分類が可能であることを明らかになった．また，感情分類対象歌詞の文字数の設定に関しては，250 字にして使用するのが最も再現率が高い実験設定であったが，大規模に分類実験を行う場合，人手による抜粋作業はかなりのコストがかかるため，その 250 字を ChatGPT に抜粋させることが今後課題になると考える．さらに，現状では再現率が 50%となる分類が存在するが，誤答先の分類が類似する傾向にある．今後は，それら再現率の低い感情分類を他の感情分類に統合することによって全体の再現率を 100%に近づけて，全自動での歌詞感情分類データセット作成を実現する予定である．

謝　辞

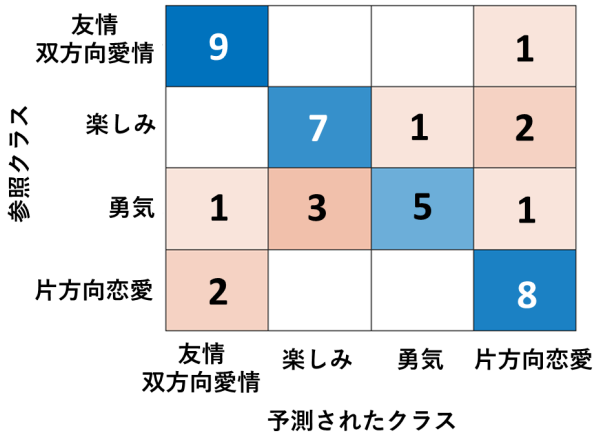
本研究は科研費 21H00901 の助成を受けたものである．

文　献

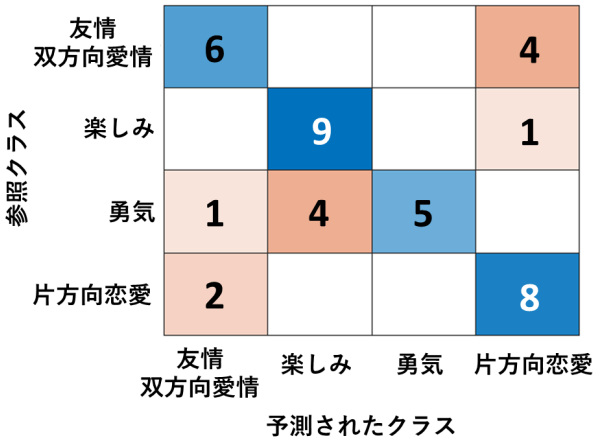
[1] E. Çano and M. Morisio. MoodyLyrics: A sentiment annotated lyrics dataset. In *Proc. ISMSI*, p. 118–124, 2017.
[2] E. Darren and S. João. Multi-emotion classification for song lyrics. In *Proc. 11th WASSA*, pp. 221–235, 2021.
[3] M. Kaminskas and F. Ricci. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, Vol. 6, No. 2, pp. 89–119, 2012.
[4] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, Vol. 9, No. 2, pp. 240–254, 2018.
[5] K. Mohsen, K. Christoph, M. Torsten, and S. Michael. TagFlip: Active mobile music discovery with social tags. In *Proc. 21st ICIUI*, p. 19–30, 2016.
[6] 奥健太. 楽曲推薦システム—プレイリスト，コンテキスト，インタラクション—. 人工知能, Vol. 34, No. 3, pp. 300–308, 2019.
[7] R. Rajendran, A. Pillai, and F. Daneshfar. LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model. *Procedia Computer Science*, Vol. 218, No. C, pp. 1196–1208, 2023.
[8] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, Vol. 39, No. 6, pp. 1161–1178, 1980.
[9] I. Santana, F. Pinhelli, J. Donini, L. Catharin, R. Mangolin, Y. Costa, V. Feltrim, and M. Domingues. Music4All: A new music database and its applications. *Proc. 27th IWSSIP*, pp. 399–404, 2020.
[10] 杉浦康友, 木本晴夫. 歌詞を用いた音楽の自動分類. 第 1 回情報科学技術フォーラム一般講演論文集, 第 2 巻, pp. 109–110, 2002.
[11] 徐馳, 菅沼睦, 亀山渉. 感情単語ベクトルと歌詞ベクトルの類似度による日本語歌詞楽曲のムード分類の検討. 電子情報通信学会総合大会講演論文集，D-12-28, 2022.
[12] 友哉倉持, 雅弘濱崎, 倫靖中野. 気分に合った音楽推薦に向けた動画共有サイト上のコメント分析. Technical Report 47, 筑波大学大学院, 産業技術総合研究所, 産業技術総合研究所, jun 2023.



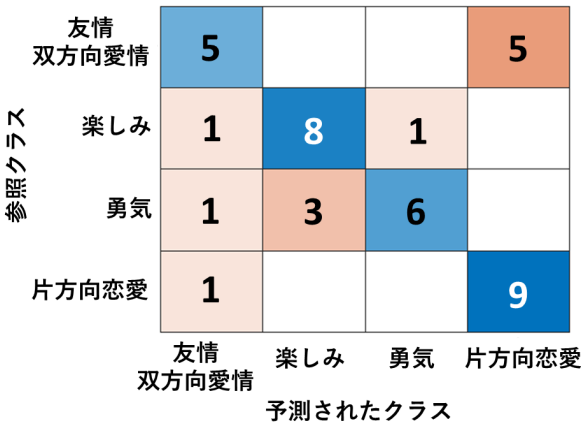
(a) 肯定 2 値分類，感情分類対象歌詞は全文を使用



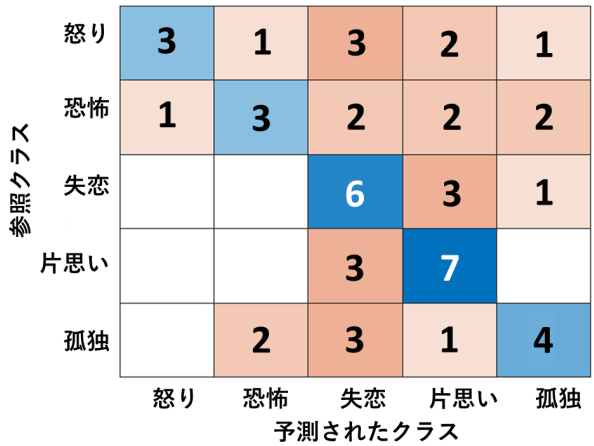
(b) 肯定 2 値分類，感情分類対象歌詞は 250 字分を使用



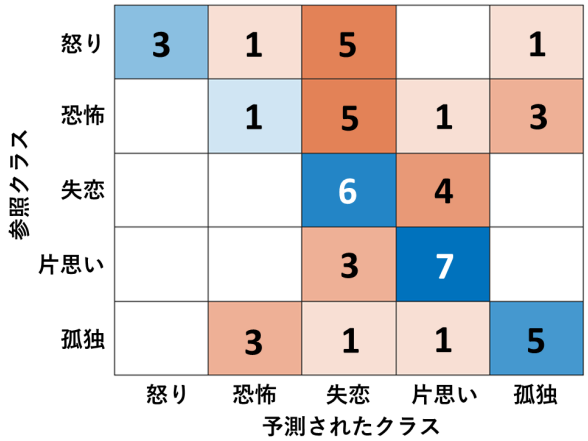
(c) 肯定 4 値分類，感情分類対象歌詞は全文を使用



(d) 肯定 4 値分類，感情分類対象歌詞は 250 字分を使用



(e) 否定 2 値分類，感情分類対象歌詞は全文を使用



(f) 否定 2 値分類，感情分類対象歌詞は 250 字を使用

図 7: 評価における混同行列

共感的な働きかけを意図したエモーショナルサポート会話における発話生成

佐藤　将太† 　佐々木裕多†† 　白石　優旗††† 　張　建偉†

† 岩手大学大学院 総合科学研究科 理工学専攻　〒020–8551 岩手県盛岡市上田 4-3-5
†† 東京工業大学 工学院 情報通信系　〒226-8503 神奈川県横浜市緑区長津田町
††† 筑波技術大学産業技術学部　〒305–8520 茨城県つくば市天久保 4-3-15
E-mail: †{g0323086,zhang}@iwate-u.ac.jp, ††yubo1336@lr.pi.titech.ac.jp, †††yuhkis@a.tsukuba-tech.ac.jp

あらまし　エモーショナルサポート (ES) は，個人の抱える精神的苦痛を社会的相互作用によって軽減し問題の解決を行う．本論文では，エモーショナルサポート会話 (ESC) の分野において，ユーザーの精神的苦痛の軽減とその解決を目指すための新たなアプローチを提案する．ESC では，共感的な対話が重要であり，特にユーザーの暗黙的な情報や精神状態の理解が不可欠である．そこで，共感的アプローチを用いて文脈情報を強化し，効果的なエモーショナルサポートのための戦略選択を実現する方法を提案する．具体的には，共感の情動的側面と認知的側面を考慮し，外部のコモンセンス情報を統合することで文脈表現を強化した研究を参考に，共感で文脈共感を強化するモジュールを用いる．実験結果から，このアプローチは従来のシステムよりも高い戦略選択性能を示し，ES 会話におけるユーザーの状況理解力の向上が期待される．本研究の貢献は，効果的な ES の実現に向けた文脈表現の強化手法の提案と，共感的アプローチの有効性を実証した点にある．

キーワード　Emotional Support Conversation，応答戦略，共感性，対話システム，応答生成

1　はじめに

エモーショナルサポート（ES, Emotional Support）は，個人の抱える精神的苦痛の軽減と，その人が直面する課題を理解して解決するまでのサポートを行うことを目的としている [1,2]．エモーショナルサポート会話 (ESC, Emotional Support Conversation) は，助けを求める人と支援を行う人との間で行われ，会話を通してその精神的苦痛を徐々に軽減することを目指すタスクである．ESC では，社会的相互作用（専門家ではなく，仲間や友人，家族のような）を通してサポートを提供し，適切な戦略と支援的な応答で対処を行う [2,3]．

また，効果的に ES を行うためには，助けを求める人の経験や感情に従って共感を表現することが必要となる点が述べられ，ESC と共感の関係性が示唆されている [2]．共感は，他人の考えや感情を適切に理解しようと努力する人間の心理的現象であり [4–6]，他者を理解するのに必要不可欠な概念である．共感能力があることで人は，他者が置かれている状況や感情を理解し，察知し，適切に対応することが可能となる [5,7]．また，共感的な態度はカウンセリングにおける治療効果にも大きく影響を与えることが研究で示されており，他者の状況を理解する能力が ES に対しても十分適用され得ることを表している [8]．

注目すべきは，ESC システムがより支援的な応答を生成するためには，ユーザーの感情的問題の原因を探り，発言に含まれる心理的な意図を理解する必要があるということである．この意図を理解することでシステムはユーザーの精神状態を捕捉することができる一方で，ユーザーがこれを明示的に言及することはない [9]．そのため，ESC システムはユーザーの発言に含まれる感情原因や暗黙的な意図を汲み取ることが求められる．このような行為は先に示した共感的な心理現象そのものである．

このような背景から，ESC システムには次のような課題がある．ESC の核となる戦略選択において，対話の文脈に含まれるユーザーの暗黙的な状況や精神状態を適切に考慮すべきだという点である．複数ある既存の ESC システムの中で [3,9,10]，Tu ら [3] は ES と共感性の混合的アプローチを示した．文脈情報を元にした確率分布的戦略表現を用いてユーザーの感情状態への対処を行いながら，事前学習された生成的コモンセンス推論モデルである COMET [11] を利用しシステムに外部知識を与えることで，表出されない瞬間的なユーザーの精神状態を捕捉した．そして両者の特徴を応答生成に取り入れることによって，共感性の増加と効果的な ES 会話の可能性を示した．しかし，付与された知識は応答生成にのみ用いられ，戦略選択の元となる文脈理解からは独立している．

我々はこの課題に対処するため，共感性というアプローチによって文脈に含まれるユーザーの暗黙的情報を取り出し，効果的な ES を行う方法を提案する．具体的には共感的応答を行う共感対話システム [12–15] の中でも，特に暗黙的に示されるユーザーの状況や精神状態をモデルに理解させようと試みた研究 [7] に着想を得ている．この研究は共感を構成する 2 つの側面（情動的側面，認知的側面）を適切に考慮するアプローチを提案している．COMET(ATOMIC^{2}0$_2$0) [16] から生成されるコモンセンスをモデルに与えることで，共感の認知的側面をモデリングするための方向性を示した．共感の両側面を考慮する

ことで文脈情報を強化したという点で，このアプローチは ESC システムのユーザーの状況理解力を向上させることが期待される．実際に我々が行った実験では，文脈情報を加工する共感的アプローチによって，応答生成に用いる戦略選択の性能が向上した．この結果から，戦略選択を行う ESC システムにおいて，共感的アプローチが文脈表現をより強化し，システムの戦略選択性能に対して有効に働くことが示された．

本論文の貢献は以下の通りである：

- 効果的なエモーショナルサポートを実現するための，共感性を考慮した文脈表現を用いて戦略選択を行う手法を提案する．この手法では，外部情報を利用した共感的アプローチによって文脈に含まれるユーザーの暗黙的な情報を獲得し，文脈表現を強化する．
- 強化された文脈情報の異なる利用の仕方を検証し，戦略予測のために用いられる文脈表現の種類によって予測性能が異なることを示す．
- 実験から，ベースラインモデルと比較して，提案手法は生成される応答文の自然さを維持しながら応答戦略の選択性能を大きく向上させる．

2 関 連 研 究

2.1 Emotional Support Conversation

初期の ES データセットはユーザーとの関わりを単純化して扱う傾向にあり，会話はシングルターンに近いで展開されるものが多い．そのため効果的な ES を行うには制限があった [17–19]．2021 年にマルチターンの ES 会話データセットが公開されてから，複数の ESC システムが開発され注目を集めている [2,3,9,10]．Peng ら [9] の開発した GLHG は，ユーザーのグローバルな感情原因とローカルな心理的意図，対話履歴との相互作用を階層的グラフネットワークによってモデリングしている．COMET を用いてユーザーのローカルな心理的意図を組み込むことも行っているが，マルチターン ESC の重要な要素である戦略選択は明示的には行っていない．対して Tu ら [3] の開発したシステム MISC は，戦略予測を行う種類の ESC システムである．戦略予測を行うために事前に用意された，8 つの戦略に対応する特別なトークンを使用する [2]．MISC は応答を生成する前にこのトークンの生成確率を出力し，該当戦略に基づいた応答生成を行う．また，応答生成には予測した戦略の他に，文脈表現と COMET で生成した知識も活用する．ここで，MISC は戦略予測に際して，文脈の埋め込み表現から次に選択すべき戦略を判断するが，この文脈表現には生成された外部知識は含まれておらず，戦略選択のための文脈情報の十分な活用という点で課題が残る．

2.2 共感応答生成と共感表現の抽出

共感は文献上では新しい用語であるため，社会心理学や心理療法の分野では具体的には定義されていない [7]．一般的には，情動（Affection）と認知（Cognition）の 2 つの側面を含んだ，複雑な多次元構成要素として知られている [7,20]．情動的共感は，様々な感情的な刺激を通じて相手の感情を経験することを可能にし [21]，認知的共感は，他者の状況や暗黙の心理的状態を理解することを可能にする [22]．共感は，これら 2 つの共感的側面を含む心理的現象である．

近年では，対話システムに共感を実装して共感応答を生成する研究が多く行われている．共感対話データセット [12] が公開されたことで，様々な共感対話システムが提案されている．対話の文脈に含まれる感情の検出，すなわち共感の情動的側面に重点を置いた共感対話システムの研究 [14,15] がいくつか行われてきた一方で，認知的側面の実装にはあまり注意が向けられてこなかった．Sabour ら [7] は，外部知識を利用することで共感の持つ両側面を適切に考慮した応答の生成を試みた．具体的には，事前学習済みの生成的コモンセンス推論モデルである COMET(ATOMIC$^{20}_{20}$) [16] を用いて知識生成を行い，それらの埋め込みと文脈表現を関連付けることによって文脈表現の強化を行った．ここで，ATOMIC [23] は日常的な if-then イベントに関するコモンセンス知識ベースである．Sabour らは知識を抽出する際に 5 つの種類の知識を利用した．これらは，各事象（イベント）に関与する人物の次のようなコモンセンス関係を推論する：事象が人物に与える影響（$xEffect$），事象に対する反応（$xReact$），事象前の意図（$xIntent$），事象が起こるために必要なもの（$xNeed$），事象後に望むもの（$xWant$）．また，実験には ATOMIC$^{20}_{20}$ [16] データセットで訓練された COMET を用いている．

2.3 ESC と共感性

以上，二つの関連する分野の対話システムについて述べた．2.1 章の先行研究における ESC システムでは，文脈表現から得られる戦略分布を利用して応答文を生成することでエモーショナルサポートを行う．知識生成モデルによってユーザーに関する知識の拡大が行われている一方で，生成された知識は戦略選択に用いられる文脈表現自体を拡張するものではないため，ユーザーの暗黙的な精神状態を十分に理解しないままエモーショナルサポートの戦略を決定していると言える．また，2.2 章の先行研究における共感対話生成は日常的な会話における共感的な応答生成を目的としており，これらの共感対話モデルは精神的苦痛を抱える人のネガティブな感情を宥めたり，問題解決を行うことをゴールとしてはいない．そのため，単純に共感対話モデルを ESC タスクに適用することは難しい．

こうした背景から，本研究では，文脈表現を強化するための共感的アプローチを既存の ESC システムに対して適用させる．この融合によって，モデルがユーザーの置かれた状況を理解するのを助け，より適切な戦略選択を行えるようになる状態を目指す．すなわち，ユーザーに対する共感的な働きかけを意図したエモーショナルサポート会話の達成を目指す．

今回の実験では先行研究 [7] に倣い，共感表現を強化するモジュール内で知識生成のために ATOMIC$^{20}_{20}$ で学習された COMET を利用し，先にあげた 5 つのリレーションを同様に使用した．

3 提案手法

提案モデルの 1 つを図 1 に示す．提案モデルは主に次の 3 つのパートから構成されており，(1)Contextual Mental State Modeling, (2)Strategy Planning, (3)Response Generation である．なお, 図 1 では先行研究 [3] の混合戦略アーキテクチャに対し，(1) に Empathy Enhancement Encoder (EE) モジュールを挿入している．

3.1 MISC アーキテクチャの概要

MISC モデルはエモーショナルサポート会話を行うための，Transformer アーキテクチャをベースとした ESC システムである．図 1 の Context Encoder が入力として受け取る要素は，会話履歴である Context $\mathbf{C}$，会話の Situation $\mathbf{s}$，及び助けを求めるユーザーである Seeker の最終発話 $\mathbf{x}$ の 3 要素である．ここで，展開される会話の中で微妙に変化する Seeker の感情強度を補足捕捉するため，$\mathbf{s}$，$\mathbf{x}$ については予め COMET を用いて 5 つのリレーションに対応した知識を生成している．これら 3 つを Context Encoder に通すことで，各埋め込み $\boldsymbol{C}$, $\hat{\boldsymbol{H}}^{\boldsymbol{s}}$, $\hat{\boldsymbol{H}}^{\boldsymbol{x}}$ を得る．得られた $\hat{\boldsymbol{H}}^{s}$, $\hat{\boldsymbol{H}}^{x}$ についてはそれぞれ $\boldsymbol{C}$ との Cross-Attention を計算し，これを $\boldsymbol{H}^{s}$, $\boldsymbol{H}^{x}$ とする．後述する 2 つの隠れ表現，$\boldsymbol{h}^{\boldsymbol{g}}$, $\tilde{\boldsymbol{H}}_{\boldsymbol{CTX}}$ を含めた計 5 つの隠れ表現，$\boldsymbol{C}$, $\boldsymbol{H}^{s}$, $\boldsymbol{H}^{x}$, $\boldsymbol{h}^{\boldsymbol{g}}$, $\tilde{\boldsymbol{H}}_{\boldsymbol{CTX}}$ がアーキテクチャの Encoder 部分の埋め込みとして得られる．こうして得られた埋め込みをもとに，応答戦略ラベルの決定と応答生成を行う．

先行研究を参考に，以下に手法を定式化して説明する．

3.1.1 Contextual Mental State Modeling

ここでは, ユーザの精神状態に関する情報の抽出を行う．データセットに含まれる $\boldsymbol{N}$ 個の会話について，n_i 個の発話からなる会話履歴は $\boldsymbol{c}^{(i)} = (\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, ..., \boldsymbol{u}_{n_i}^{(i)})$ のように表される（なお，後述する説明では簡単のため (i) の表記を省略する）．

始めに Context Encoder に会話履歴 $\mathbf{C}$ を入力して文脈表現 $\boldsymbol{C}$ を得る:

$$\boldsymbol{C} = \mathbf{E}\mathrm{nc}(\mathrm{CLS}, \boldsymbol{u}_1, \mathrm{EOS}, \boldsymbol{u}_2, ..., \boldsymbol{u}_{n_i}) \tag{1}$$

ここで CLS は会話の始まりを表すトークンであり，EOS は発話同士の区切りを表すトークンである．文脈表現 $\boldsymbol{C} \in \mathbb{R}^{L \times d}$ は系列の長さ L と埋め込み次元 d からなる行列である．

次に Context Encoder から得られた出力を Mental State Encoder に入力する．Mental State Encoder では，Seeker の置かれている状況をより普遍的な知識で理解するための処理を行う．具体的な方法は先行研究に倣うが，ここではその概要を説明する．このモジュールで用いる COMET はコモンセンス知識生成器であり，会話における Seeker の精神状態に関する情報を抽出する．COMET はあるシチュエーション $\mathbf{s}$ とリレーション rel が入力されると，それらに対応するコモンセンス推論を出力する．これを j 個のリレーションについて行い，コモンセンス推論ブロック $\boldsymbol{B}^{s}$ を得る．これは，N_s 個の精神状態の集合となる:$\boldsymbol{B}^{s} = \{\boldsymbol{b}_j^{s}\}_{j=1}^{N_s}$．同様に Seeker の最終発話 $\boldsymbol{x}$ についても同様の操作を行い，ブロック $\boldsymbol{B}^{x}$ を得る．

こうして得られたコモンセンス推論を Context Encoder に通すことで，式 2 を得る．

$$\begin{aligned} \hat{\boldsymbol{H}}^{\boldsymbol{s}} &= [\boldsymbol{h}_{\boldsymbol{1,1}}^{\boldsymbol{s}}, \boldsymbol{h}_{\boldsymbol{2,1}}^{\boldsymbol{s}}, ..., \boldsymbol{h}_{\boldsymbol{N_{st},1}}^{\boldsymbol{s}}] \\ \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{s}} &= \mathbf{E}\mathrm{nc}(\boldsymbol{b}_{\boldsymbol{j}}^{\boldsymbol{s}}) \end{aligned} \tag{2}$$

文脈に無関係なブロックはノイズとなってしまうため，文脈と関連性の高いブロックを絞り込む目的で $\hat{\boldsymbol{H}}^{s}$ と $\boldsymbol{C}$ との Cross-Attention を計算する (式 3):

$$\begin{aligned} \boldsymbol{Z} &= \mathrm{softmax}(\hat{\boldsymbol{H}}^{s} \cdot \boldsymbol{C}^{\boldsymbol{T}}) \cdot \boldsymbol{C} \\ \boldsymbol{H}^{s} &= \mathrm{LN}(\hat{\boldsymbol{H}}^{s} + \boldsymbol{Z}) \end{aligned} \tag{3}$$

ここで，LN では層正規化 (Layer Normalization) を行う．

x に対しても同様の処理を行い，$\boldsymbol{H}^{x}$ を得る．こうして得られた $\boldsymbol{H}^{s}$, $\boldsymbol{H}^{x}$ は，会話レベルあるいは発言レベルで Seeker の精神状態を表現する埋め込みとなる．

3.1.2 Strategy Planning

本モジュールでは応答戦略の学習を行う．基本的な方針は文脈表現の CLS トークンを元にした分類学習である．このモジュールも Mental State Encoder と同様，具体的な方法は先行研究に倣う．モジュールの入力として式 1 の文脈表現 $\boldsymbol{C}$ を受け取り，分散戦略ベクトル $\boldsymbol{h}_{\boldsymbol{g}}$ を出力する．

まず，文脈表現 $\boldsymbol{C}$ の CLS トークンに対して多層パーセプトロンを適用し，埋め込み $\boldsymbol{p}^{g}$ を取得する:

$$\boldsymbol{p}^{g} = \mathrm{MLP}(\boldsymbol{C_1}) \tag{4}$$

ここで，$\boldsymbol{p}^{g} \in \mathbb{R}^{1 \times m}$ であり，$\boldsymbol{C_1}$ は文脈表現の CLS トークンを表す．m は戦略数である．また，$\boldsymbol{p}^{g}$ は選択されうる戦略の確率を表す確率分布となる．

次に, 式 4 で得られた確率分布 $\boldsymbol{p}^{g}$ と応答戦略行列 $\boldsymbol{T}$ との積を計算する．$\boldsymbol{T}$ は d 次元の m 個のベクトルからなる $\boldsymbol{T} \in \mathbb{R}^{m \times d}$ の行列である．

$$\boldsymbol{h}^{\boldsymbol{g}} = \boldsymbol{p}^{g} \cdot \boldsymbol{T} \tag{5}$$

3.2 Empathy Enhancement Encoder

本モジュールは Sabour らが提案する手法を参考に構築している．コンテクスト C とその埋め込みである文脈表現 $\boldsymbol{C}$ を入力に受け取り，2 種類の共感的側面を考慮する処理によって，強化した共感的文脈表現 $\tilde{\boldsymbol{H}}_{CTX}$ を出力する．

まず，COMET を用いて，Seeker の最終発話 $\boldsymbol{x}$ から 5 つの $relation$ に関するコモンセンスを得る．ここで，Sabour らと同様，5 つのリレーションを 2 種類の共感的側面に対応付ける; xReact リレーションを共感の情動的側面の要素とし，それ以外

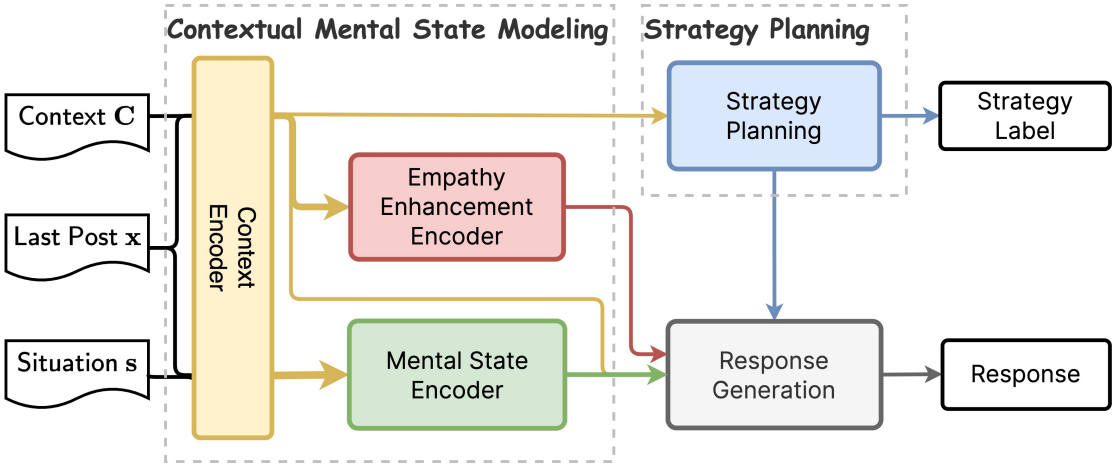


図 1: 提案アーキテクチャの概要 (Contextual Mental State Modeling，Strategy Planning，及び Response Generation) ここでは既存研究 MISC [3] のアーキテクチャに対して Empathy Enhancement Encoder モジュールを Response Generation に対して並列に接続するモデルを示している．

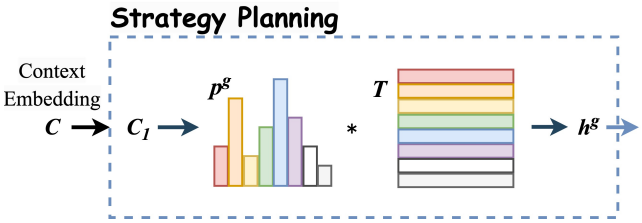


図 2: 既存研究であるモデル MISC [3] で提案された，Strategy Planning モジュールの概要．受け取った埋め込み $\boldsymbol{C}$ の CLS トークンをベクトルへと変換し，この値と混合戦略行列との積を計算する．最終的に，次に選択するべき応答戦略の確率分布を出力する．

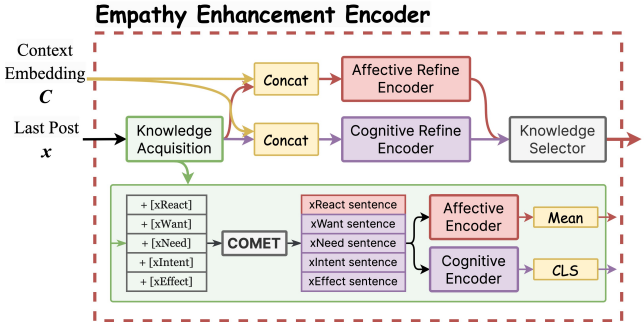


図 3: 既存研究であるモデル CEM [7] で提案された Empathy Enhancement Encoder モジュールの概要．文脈表現とユーザーの最終発話を受け取り，最終的に共感的側面が強化された新たな文脈表現を出力する．

の 4 つのリレーション（xWant, xNeed, xIntent, xEffect）を認知的側面の要素とする．このようにして分類された 2 種類の共感的要素を区別して Encoder に入力し，次の式 6 の埋め込みを得る:

$$\begin{aligned} \boldsymbol{H}_{xReact} &= \mathbf{Enc}_{Aff}(\boldsymbol{E}_{CS_{xReact}}) \\ \boldsymbol{H}_{r} &= \mathbf{Enc}_{Aff}(\boldsymbol{E}_{CS_{xReact}}) \end{aligned} \tag{6}$$

ここで，$\mathrm{CS}_{\mathrm{xReact}}$，$\mathrm{CS}_r$ は COMET によって生成されたコモンセンス推論の系列を意味し，$\boldsymbol{E}_{CS_{xReact}}$，$\boldsymbol{E}_{CS_r}$ はその埋め込みを表す．$\boldsymbol{H}_{xReact} \in \mathbb{R}^{l_{xReact} \times d}$，$\boldsymbol{H}_r \in \mathbb{R}^{r \times d}$ であり，l_{xReact}，r は入力される各コモンセンス推論の系列長である．また，$r \in \{xWant, xNeed, xIntent, xEffect\}$.

式 6 で得られたコモンセンス推論埋め込みを文脈に取り入れるため，ここでは代表的な値のみを扱う．xReact ラベルによって生成されるコモンセンス推論は，そのほとんどが文章ではなく感情を表現する 1 語となるため，得られた埋め込みの平均を代表値とする．また，他 4 つのラベルで生成される推論については，生成文の意味を考慮するため，各埋め込みの CLS トークンに当たるベクトルを代表値とする．

$$\begin{aligned} \boldsymbol{h}_{xReact} &= \mathrm{Average}(\boldsymbol{H}_{xReact}) \\ \boldsymbol{h}_{r} &= \boldsymbol{H}_{r}[0] \end{aligned} \tag{7}$$

ここで，$h_{xReact}, h_r \in \mathbb{R}^d$ である．

ここまでで得られたコモンセンス推論の隠れ表現と文脈表現 $\boldsymbol{C}$ を結合し，エンコーダに入力することで，文脈と共感表現の対応付けを行う．

$$\begin{aligned} \boldsymbol{U}_{xReact}[i] &= \boldsymbol{C}[i] \oplus \boldsymbol{h}_{xReact} \\ \boldsymbol{U}_{r}[i] &= \boldsymbol{C}[i] \oplus \boldsymbol{h}_{r} \end{aligned} \tag{8}$$

$$\begin{aligned} \boldsymbol{H}_{Aff} &= \mathbf{Enc}_{CTX-Aff}(\boldsymbol{U}_{xReact}) \\ \boldsymbol{H}_{Cog,r} &= \mathbf{Enc}_{CTX-Cog}(\boldsymbol{U}_{r}) \end{aligned} \tag{9}$$

その結果，5 つの *relation* に対応した埋め込み $\boldsymbol{H}_{Aff}, \boldsymbol{H}_{Cog,r} \in \mathbb{R}^{L \times d}$ が得られ，これを全結合層に与える．

$$
\begin{aligned}
\boldsymbol{H}_{Cog}[i] &= \bigoplus_{r\in\{xWant,xNeed,xIntent,xEffect\}} \boldsymbol{H}_{Cog,r}[i] \\
\boldsymbol{H}_{Refine}[i] &= \boldsymbol{H}_{Aff}[i] \oplus \boldsymbol{H}_{Cog}[i] \\
\tilde{\boldsymbol{H}}_{CTX} &= \mathrm{MLP}(\sigma(\boldsymbol{H}_{Refine}) \otimes \boldsymbol{H}_{Refine})
\end{aligned} \tag{10}
$$

ここで，$\tilde{\boldsymbol{H}}_{\boldsymbol{CTX}} \in \mathbb{R}^{L\times d}$ は，共感的アプローチによって共感表現が強化された文脈表現を意味する．

3.3 Response Generation

最終的に，Encoder 部分の出力として得られる，ユーザの精神状態を表現する埋め込みを Decoder 側に取り入れる目的で，5 つの埋め込みと Decoder の隠れ表現との Cross-Attention をそれぞれ計算する．

$$
\begin{aligned}
\boldsymbol{A}^{\boldsymbol{c}} &= \text{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}) \\
\boldsymbol{A}^{\boldsymbol{s}} &= \text{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}^{s}) \\
\boldsymbol{A}^{\boldsymbol{s}} &= \text{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}^{x}) \\
\boldsymbol{A}^{\boldsymbol{g}} &= \text{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{h}^{g}) \\
\boldsymbol{A}^{\tilde{\boldsymbol{H}}} &= \text{CROSS-ATT}(\boldsymbol{O}, \tilde{\boldsymbol{H}}_{\boldsymbol{CTX}}) \\
\boldsymbol{O}' &= \mathrm{LN}(\boldsymbol{A}^{c} + \boldsymbol{A}^{s} + \boldsymbol{A}^{x} + \boldsymbol{A}^{g} + \boldsymbol{O})
\end{aligned} \tag{11}
$$

ここで，$\boldsymbol{O}$ は Decoder の隠れ表現である．こうして得られた $\boldsymbol{O}'$ をフィードフォワードネットワークに入力し，モデルの出力系列を得る．

$\boldsymbol{A}^{c}$，$\boldsymbol{A}^{\tilde{H}}$ が計算されるのは，$\boldsymbol{H}$ 及び $\tilde{\boldsymbol{H}}_{\boldsymbol{CTX}}$ が Response Generation に接続される場合に限り，計算されないものがある場合 $\boldsymbol{A}^{\boldsymbol{c}}$ や $\boldsymbol{A}^{\tilde{\boldsymbol{H}}}$ は $\boldsymbol{O}'$ の計算処理には含まれない．

3.4 目的関数

MISC と同様に，生成された応答，および予測された戦略ラベルについて以下の誤差関数を計算し，それぞれを $\mathcal{L}_r$，$\mathcal{L}_g$ とする：

$$
\begin{aligned}
\mathcal{L}_r &= -\sum_{t=1}^{n_r} log(p(r_t|\boldsymbol{r}_{j<t}, \boldsymbol{c}, \boldsymbol{s}, \boldsymbol{x})) \\
\mathcal{L}_g &= -log(p(g|\boldsymbol{c}, \boldsymbol{s}, \boldsymbol{x}))
\end{aligned} \tag{12}
$$

以上，2 つの損失関数の和を最小化することにより，モデルの最適化を行う．

$$
\mathcal{L} = \mathcal{L}_r + \mathcal{L}_g \tag{13}
$$

4 実　　験

4.1 データセット

モデルの学習及び評価には morecry/MISC[1]で公開されている，Emotional Support Conversation(=ESConv) データセットの前処理済みデータを使用する．オリジナルの ESConv データセット [2]の総対話数は 1,300 件だが，1 つの対話において最大 10 ターン（10 発話）を目処に会話を切り出す処理が行われている．表 1 にその統計量を示す．

表 1: 前処理済み ESConv データセットの統計量

Category	Train	Dev	Test
対話数	14,117	1,764	1,764
発話あたり単語数の平均	17.25	17.09	17.11
対話あたり話者ターン数の平均	7.61	7.58	7.49
対話あたり単語数の平均	148.46	146.66	145.17

4.2 実験設定

実験にあたって，学習のバッチサイズを 16，最大エポック数を 8 に設定した．実験の実行環境には Google Colab V100 を使用した．

4.3 モ デ ル

以下のモデルをベースラインに用いた:

- MISC [3]: 図 4 に対応する．

また，次の 3 つの提案モデルを実験に用いた．モデルの図は 6 章 Appendix に掲載している．

- Ours1 (Enhance-Parallelly-Added): 図 5 に対応する．
- Ours2 (Enhance-SP): 図 6 に対応する．
- Ours3 (Enhane-SP-Replace-RG): 図 7 に対応する．

4.4 評価指標

自動評価指標として **ACC.**，**PPL** の 2 つの指標を用いた．

ACC. は Accuracy を指し，テストデータにおける戦略ラベルの予測精度を測る．これを **straACC.** とした．予測精度が高くなるほど，ユーザの背景情報を考慮し，心的ストレスを軽減するための適切な応答戦略を選択できていると言える．**PPL** (perplexity) は応答生成の汎化性能を測る指標で，低い値であるほど文章生成におけるエントロピーが小さくモデルが自信を持ってその応答を生成していると言える．

5 実 験 結 果

実験の結果を表 2 に示す．表 1 番行のモデル（図 4）は先行研究で用いられたモデルと同様のモデルであり，今回再実験を行い，この結果を本実験のベースラインとした．2 行めから 4 行目のモデル（Appendix, 図 5，6，7）は提案モデルである．提案モデルの実験で得られた数値のうち，ベースラインを超えたものに下線を引き，その中でも最高性能を達成したものを太字で示した．

この実験から，Empathy Enhancement Encoder(EE) モジュールを追加したことによって，戦略ラベルの予測精度が最大約 2.3%向上したことが確認された．ベースラインと提案モデルを比較すると，Ours1 モデル（EE モジュールを Decoder に

1：https://github.com/morecry/MISC

2：https://github.com/thu-coai/Emotional-Support-Conversation

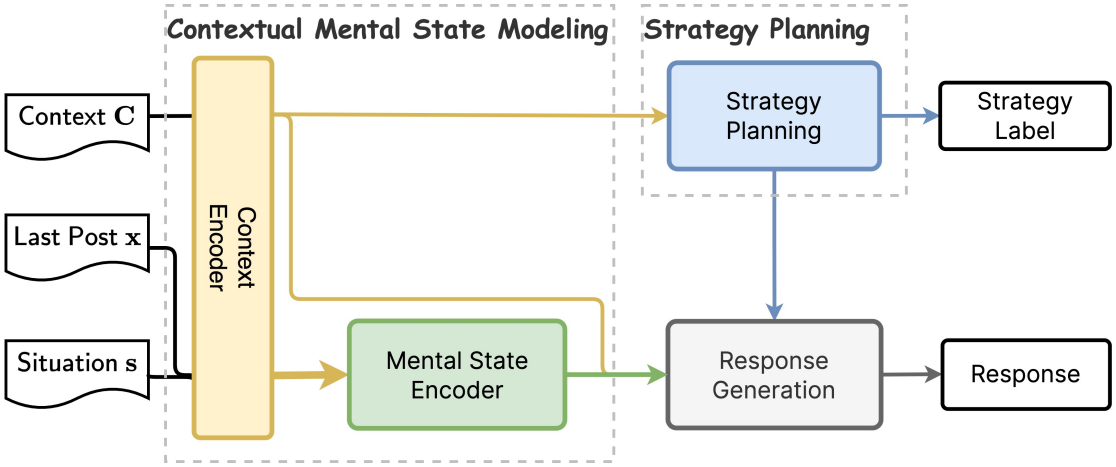


図 4: ベースラインモデル MISC [3] のアーキテクチャ．戦略を確率分布として表現し，混合戦略表現を実現した，Blenderbot-Small ベースの ESC 会話モデルである．COMET を用いてユーザーの精神状態を捕捉し，戦略表現と合わせた混合戦略アプローチによって ES を行う．

並列に接続する）では EE モジュール追加の影響はあまり見られなかった．Ours1 の場合，式 12，13 より，並列に追加された EE モジュールの学習フィードバックは主に応答生成の部分から与えられている．そのため，戦略ラベルの予測性能という点に関して EE モジュールを追加しても性能には大きな影響を与えなかったと考えられる．対して，ベースラインと Ours2 モデル，Ours3 モデルを比較すると，**ACC.** がより大きく向上し，ベースラインに比べて最大約 2.3%の性能向上が見られた．ここで，Ours2 は EE モジュールの出力である共感的アプローチによって強化された文脈表現を Strategy Planning の入力として利用しているため，式 12 の $\mathcal{L}_g$ によってラベル分類学習のフィードバックが精度向上に影響したと考えられる．Ours1 と Ours2 の Strategy Planning モジュールへの入力を比較すると，前者のモデルは Context Encoder からの文脈表現を入力として受け取り，後者は EE モジュールの強化文脈表現を入力として受け取っている．Strategy Planning モジュールは入力として受け取ったコンテクストの会話情報から，次に選択するべき応答戦略を確率分布の形で出力するモジュールである．このことを踏まえると，EE モジュールは文脈表現に含まれる情報を拡張しており，応答戦略の選択において有効に作用する可能性を示している．Ours3 モデルも Ours2 モデルと同様，Strategy Planning モジュールに接続された EE モジュールが選択性能の向上に寄与していることが分かる．加えて Ours3 モデルでは，EE モジュールが戦略選択と応答生成の両方に対して接続されているため，より多くの学習フィードバックを得ているはずである．これによって Ours2 モデルよりもさらに数値が向上したのではないかと考えられる．

次に，モデルの応答文生成における性能について結果をみると，EE モジュールを追加することでベースラインよりも多少の改善が確認された．Ours1 と Ours2 はベースラインと比較して **PPL.** の大きな変動は見られなかった．一方で，Ours3 では **PPL.** の値が大きく低下した．これは，このモデルが応答生成時に Context Encoder の文脈情報を用いていないことが原因として考えられる．評価に用いた **PPL.** はモデルが文章を

表 2: 実験結果

Model	**stra ACC.** ↑	**PPL** ↓
Baseline (MISC)	32.029	16.371
Ours1 (Enhance-Parallelly-Added)	31.689	**16.348**
Ours2 (Enhance-SP)	33.845	16.358
Ours3 (Enhance-SP-Replace-RG)	**34.354**	19.293

生成する際の単語の予測確率から算出される値で，この値が小さいほどモデルが自信を持って文章を生成していると解釈できる．すなわち，応答生成を行う際に Context Encoder から得られる文脈埋め込みを用いなかったことで，応答生成時にモデルがコンテクストを十分に参照できなかった可能性が高い．

以上の結果から，Ours2 モデルが最も良いモデルであると言える．戦略選択においては，共感性を考慮して文脈情報を強化することが選択性能向上に寄与することが明らかとなった．ただし，元のコンテクストを直接参照して応答生成を行う必要があると考えられる．

6 まとめと今後の展望

エモーショナルサポート会話において，文脈に含まれる共感表現を強化するアプローチを採用することにより，エモーショナルサポートにより適した応答生成を試みた．応答戦略の学習において，共感的側面を考慮することで応答戦略の選択性能に良い影響をもたらした．提案手法により，生成される応答文の自然さを維持しながら応答戦略ラベルの予測性能が向上した．

しかし，今回用いた自動評価では実際に生成される応答に対する応答戦略予測の影響を一概に評価することは難しい．そのため，予測された戦略応答ラベルの精度に対する評価に加えて，人手による応答文章の評価も行い，共感的アプローチが生成文に与える影響についてさらに調査を行う必要がある．

今後の展望として，生成応答文の人手評価，モデルが予測した戦略ラベルの意味的な分析などを行う．これらの追加調査によって，共感表現と戦略選択がエモーショナルサポート会話に

与える影響と関係性を明らかにし，より良いエモーショナルサポート会話モデルの実現を目指す．

謝　　辞

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

文　　献

[1] Brant R Burleson. Emotional support skill. In *HANDBOOK OF COMMUNICATION AND SO- CIAL INTERACTION SKILLS*, page 551, 2003.

[2] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics*, 2021.

[3] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, 2022.

[4] G.R. Adams. Social competence during adolescence: social sensitivity, locus of control, empathy, and peer popularity. *Journal of Youth and Adolescence*, page 203–211, 1983.

[5] Sevgi Coşkun Keskin. From what isn’t empathy to empathic learning process. pages 4932–4938, 2014.

[6] 浅野 良雄. 共感的態度を評価するための一方法. **医学教育**, 2010.

[7] Sahand Sabour, Chujie Zheng, and Minlie Huang. CEM: Commonsense-aware empathetic response generation. In *Proc of AAAI*, 2021.

[8] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, 2017.

[9] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330, 2022.

[10] Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, page 110340, 2023.

[11] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.

[12] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[13] Rohola Zandie and Mohammad H Mahoor. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *The Thirty-Third International Flairs Conference*, 2020.

[14] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking emotions for empathetic response generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, 2020.

[15] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. MoEL: Mixture of empathetic listeners. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[16] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*, 2020.

[17] Lenin Medeiros and Tibor Bosse. Using crowdsourcing for the development of online emotional support agents. In Javier Bajo, Juan M. Corchado, Elena María Navarro Martínez, Eneko Osaba Icedo, Philippe Mathieu, Patrycja Hoffa-Dąbrowska, Elena del Val, Sylvain Giroux, Antonio J.M. Castro, Nayat Sánchez-Pi, Vicente Julián, Ricardo Azambuja Silveira, Alberto Fernández, Rainer Unland, and Rubén Fuentes-Fernández, editors, *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection*, pages 196–209, 2018.

[18] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, 2020.

[19] Mahshid Hosseini and Cornelia Caragea. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13018–13026, 2021.

[20] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. Comae: A multi-factor hierarchical framework for empathetic response generation, 2021.

[21] Benjamin M. P. Cuff, Sarah D. Brown, Laura K. Taylor, and Douglas James Howat. Empathy: A review of the concept. *Emotion Review*, pages 144 – 153, 2016.

[22] Robert Elliott, Arthur Bohart, Jeanne Watson, and Leslie Greenberg. Empathy. *Psychotherapy (Chicago, Ill.)*, pages 43–9, 2011.

[23] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019.

Appendix

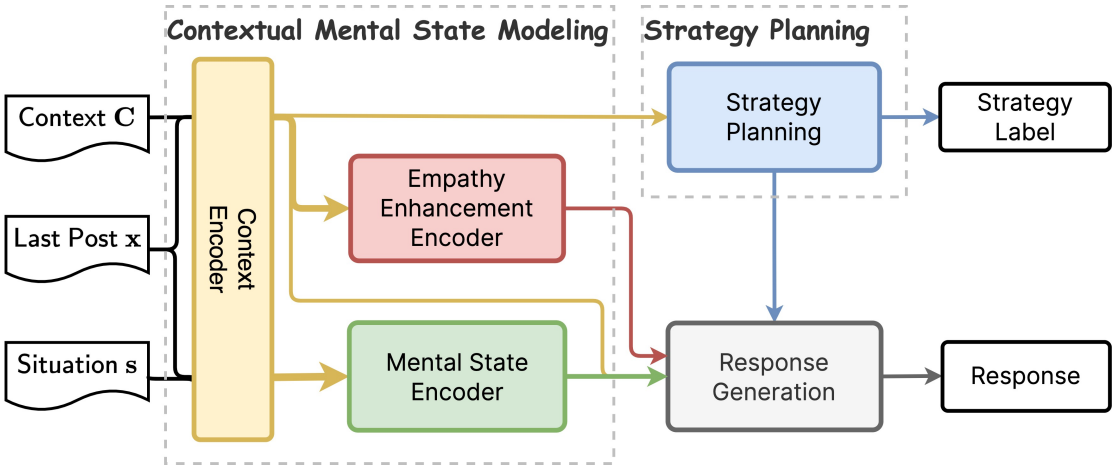


図 5: 提案モデル Ours1 (Enhance-Parallelly-Added)．ベースラインの Contextual Mental State Modeling(3.1.1 章) に対して Empathy Enhancement Encoder (EE) モジュールを並列に追加し，得られる強化文脈表現を Response Generation に接続する．

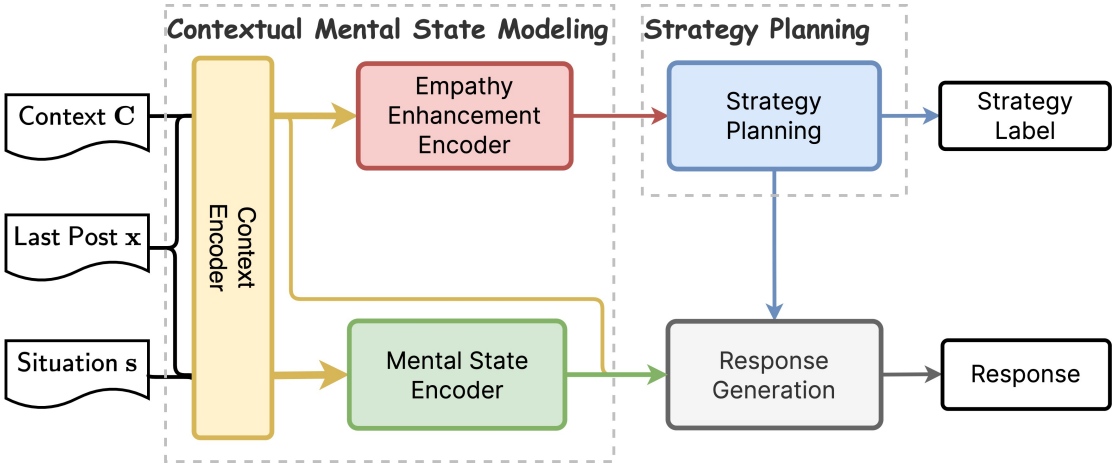


図 6: 提案モデル Ours2 (Enhance-SP)．Ours1 モデルの EE モジュールを Strategy Planning モジュールに直列で接続する．

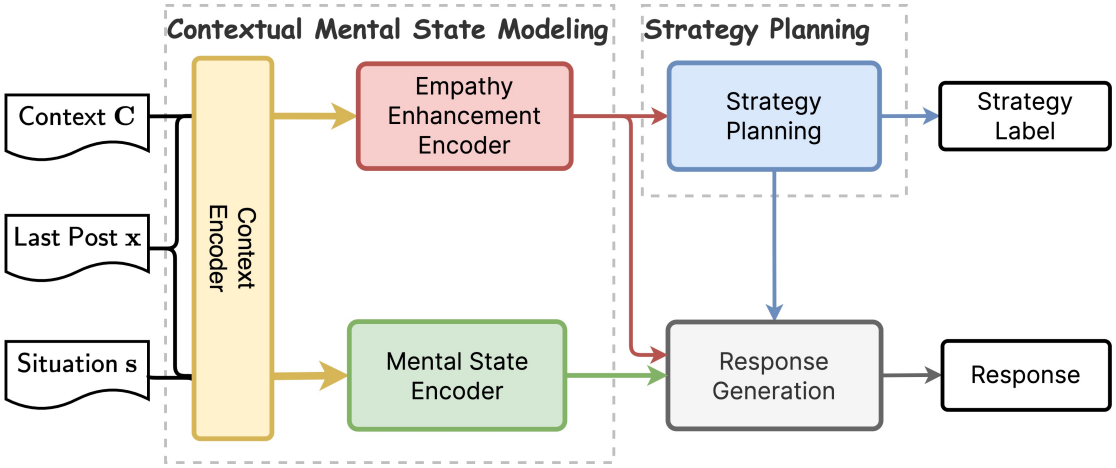


図 7: 提案モデル Ours3 (Enhance-SP-Replace-RG)．Ours2 モデルの EE モジュールの出力を Response Generation へと接続し，Context Encoder から Response Generation への入力を削除する．

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

2024年3月1日(金) 10:00 ~ 12:10 | T2-A オンライン（Zoom Events)

GPUデータ処理

座長:横山 大作(明治大学)
コメンテータ:小口 正人(お茶の水女子大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T2-A-7-01] GPUを利用した大規模データのk-way in-placeマージ

*三浦 真矢[1]、常 穹[2]、宮崎 純[2] (1. 東京工業大学情報理工学院情報工学系宮崎研究室、2. 東京工業大学情報理工学院情報工学系)

10:30 ~ 10:55
[T2-A-7-02] GPU直接IOを用いた大規模データに対するハッシュ結合の実行方式

*三浦 優也[1]、小沢 健史[1]、合田 和生[1] (1. 東京大学 生産技術研究所)

10:55 ~ 11:20
[T2-A-7-03] 余剰のGPUメモリを活用したアクセラレータ間データ転送

*濱地 達也[1]、田村 雅寿[1] (1. 富士通株式会社)

11:20 ~ 11:45
[T2-A-7-04] FPGA-GPU間の直接データ転送による消費電力性能効率の向上

*飯澤 健[1]、西山 達也[2]、サンパト プリヤンカラ[2]、石崎 晃朗[2]、田村 雅寿[1] (1. 富士通株式会社、2. NTTソフトウェアイノベーションセンタ)

11:45 ~ 12:10
[T2-A-7-05] GPUを用いたk-匿名化加工処理の高速化の検討

*吉本 廣雅[1]、満武 巨裕[1]、合田 和生[2] (1. 医療経済研究機構、2. 東京大学 生産技術研究所)

GPUを利用した大規模データのk-way in-placeマージ

三浦 真矢† 常 穹† 宮崎 純†

† 東京工業大学情報理工学院情報工学系 〒152–8550 東京都目黒区大岡山二丁目12番1号
E-mail: †miura@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.ac.jp

あらまし 本研究ではGPUを使用し，GPUメモリサイズよりも大きなデータを対象にk-way in-placeマージを行う手法を提案する．既存のマージ手法はGPUを使用しないk-way in-placeなもの，またはGPUを使用したin-placeでないものである．そこで，既存手法におけるセグメントの分割手法および並び替え手法，セグメント間のマージ手法を改良することで，これらの課題を解決するだけでなく，より適応範囲の広いマージ手法へと拡張する．評価実験では複数の分布に対して実行時間の計測およびプロファイルの取得を行い詳細な評価を行う．また，マージを使用するアルゴリズムへ本研究で提案する手法を応用し，性能の検証を行う．

キーワード 先進ハードウェア活用・GPU，大規模データ処理，マージアルゴリズム

1 はじめに

ソート済みのセグメントをマージするアルゴリズムは最も基本的なアルゴリズムの1つであり，データベースにおけるマージジョインや，情報検索システムにおける転置インデックスを使用したアンドクエリの処理などの大規模データ処理アルゴリズムのベースとなる非常に重要なアルゴリズムである．特に，データを並列処理するアルゴリズムでは別々に処理された結果をマージすることで最終的な結果を得る分割統治法が用いられている．例えば，GPUを使用したマージソート [5] では大規模なセグメントを部分ごとにソートした後，マージアルゴリズムを適用するという構成になっている．

特に，並列性の高いアルゴリズムでは問題を2つ以上の部分問題に分割し，それらの結果をマージする必要がある．このような場合，一般的な2本のセグメントをマージするアルゴリズムを適用すると k 本のセグメントをマージするために $k-1$ 回のマージを実行する必要があり，非効率的である．そのため，k 本のセグメントを同時にマージすることが可能なk-wayマージ手法が求められている．また，一般的なマージアルゴリズムではマージ対象のデータと同じだけの新たなメモリを必要としており，大規模データ処理のために必要な空間効率が悪いという問題がある．したがって，データのほかに一定の量のメモリしか必要としないin-placeな特性を持つマージアルゴリズムが必要となる．このような理由からk-way in-placeマージの研究がなされており，Salahらによるマージアルゴリズム [6] などが提案されている．

また，大規模データを処理する際，GPU(Graphics Processing Unit) による高速化が行われることがある．GPUは本来画像処理を目的に設計されたプロセッサであるが，CUDA [4] などの開発環境の登場により，汎用計算に用いられるようになった．GPUによる大規模データ処理に関する研究として著者らによるデータ並列プリミティブの拡張に関する研究 [8] がある．これはGPUメモリ内で完結できるサイズのデータに対する汎用計算を行うデータ並列プリミティブを拡張し，大規模データを処理する手法を提案したものである．

本研究では，既存のk-way in-placeマージ手法に対し，データ並列プリミティブの拡張で使用されているデータ分割やパイプライン化を用いることでGPUを使用し高速化する手法について提案する．既存手法を構成する入力セグメントの分割，データのシャッフル，マージのそれぞれについてGPU上で処理手法を提案する．

今回実装したGPUによるk-way in-placeマージについて，既存手法と実行時間を比較することで評価を行った．また，実アプリケーションへの応用としてマージソートを実装し，既存のCPU上でのマージ手法と比較を行った．評価実験の結果，マージでは1.45–3.07倍の高速化，マージソートでは1.30–2.03倍の高速化となった．

2 関連研究

2.1 CPU上での並列k-way in-placeマージ

Salahら [6] はCPU上で複数のセグメントを同時にマージすることができる並列in-placeマージ手法 (PSMerge) を提案した．従来の多くのマージ手法では，入力セグメントが2本であると仮定していたり，in-place特性を持たなかったりするため，複数の大規模データをマージする際には2つのセグメントを再帰的にマージするバイナリマージツリーベースの手法が使用されていた．しかし，従来のバイナリマージツリーベースのアルゴリズムでは，マージが進むにつれて並列度が下がり，使用していないCPUコアが発生するという課題があった．そこで，この手法では，以下のようなステップで並列度の高いマージアルゴリズムを構成し，課題を解決している．(1) k本の入力セグメントそれぞれt個のパーティションに分割，(2) 各パーティションをメモリ上で連続させるために並列にシャッフル，(3) 逐次インプレースマージアルゴリズムを利用してパーティションをマージ，(4) (2) と (3) をすべてのセグメントがマージされるまで再帰的に繰り返す．

まず，パーティションの分割では，Frederickson ら [3] が提案した手法をもとに入力を分割する位置を決定する．Frederickson らの手法は探索範囲を次第に狭めることで効率よく探索することができ，また，偏りのあるデータでもパーティションの大きさを等しく分割する位置を決定することができる．パーティショニングにより，t 個のパーティションができ，各パーティションには k 個の部分セグメントが含まれた状態となる．またパーティショニングでは，パーティションの分割数 t は CPU のコア数以下とすることを前提としている．次にパーティションのシャッフルでは，分割統治法を利用して並列にシャッフルを行う．各パーティションについて，ローテーション演算を行うことで in-place にシャッフルを行うことができる．この操作により，各パーティションはメモリ上に連続して並んだ状態となる．最後に各パーティションをバイナリマージツリーベースのマージを用いて逐次的にマージを行う．t 個のパーティションを同時にマージするため，全体での並列度はこのステップの間一定に保たれることになる．

評価実験ではセグメント数を変化させながら 10×10^6 個の 32bit 整数型データのマージを行い，3.4 倍から 8 倍の性能であることを示した．

本研究ではこの手法におけるセグメントサイズの制約とパーティション分割数の制約の拡張，および GPU による高速化が行えるようにする手法を提案する．

2.2 ヘテロジニアスな out-of-place マージソート

Gowanlock ら [5] は CPU と GPU を搭載したプラットフォーム上でのヘテロジニアスなマージソート手法を提案した．GPU を使用したソート手法としては Elias ら [7] の Radix Sort や，ModernGPU [2] や CUDPP [1] などのライブラリが提供しているプリミティブなどがあるが，それらは GPU のメモリ上に収まるサイズのデータを対象としたものである．そこで，この手法では入力データを複数のパーティションに分割し，パーティションごとに独立に GPU を使用してソートを行った後，CPU を使用して各パーティションをマージするという手法により GPU メモリサイズよりも大きなデータのソートを実現している．

また，CPU/GPU 間で大きな通信を必要とするアプリケーションでのオーバーヘッドを回避するために，データ転送のパイプライン化が行われている．ソートを行うステップでは，ページングされないメモリを経由することで，入力セグメントとデバイスメモリ間での転送がパイプライン化される．また，マージを行うステップでは，GPU がソートを行っている間にマージ済みのパーティションについて 2-way マージを行うことで実行時間の短縮が行われている．

評価実験では，1.0×10^9 個から 5.0×10^9 個の 64bit 浮動小数点型データに対してソートを行い，CPU 上の並列ソートに対して最大 3.47 倍の性能を示した．

この研究で提案されている手法では，マージに out-of-place な CPU 上でのアルゴリズムを使用しているため，入力セグメントサイズの総和の 3 倍の大きさのメモリが必要となる．そこで本研究ではこの手法におけるマージを置換することができる in-place なマージ手法を提案する．

3 提案手法

2.1 節で述べた通り，既存の in-place マージ手法である PSMerge では，(1) パーティショニング，(2) シャッフル，(3) マージの 3 つの処理から構成されている．本節では，パーティショニングステップの変更，シャッフルステップの複数の高速化，マージステップの GPU を用いた高速化について述べる．また，3 個以上のセグメントが与えられた場合のシャッフルステップとマージステップの実行順序について述べる．

3.1 パーティション分割数の変更

PSMerge では，等しいサイズのソート済みのセグメントが k 個入力として与えられる．各セグメントサイズをそれぞれ s とする．パーティショニングステップでは各セグメントを n 個のパーティションに分割を行う．その後，$2i-1$ 番目と $2i$ 番目 ($1 \leqq i \leqq \lfloor k/2 \rfloor$) のセグメントでシャッフルを用いてパーティションを並び替える．その後隣り合うパーティション同士をマージすることで $2i-1$ 番目と $2i$ 番目のセグメントのマージが完了する．1 回のシャッフルステップとマージステップでセグメント数は $\lfloor (k-1)/2 \rfloor + 1$ 個に減少する．この操作を繰り返し，最後のマージステップでは n 個のパーティションを含む要素数 $ks/2$ 個の 2 個のセグメントのマージを行う．1 個のパーティションに含まれる要素数は $p = ks/2n$ である．ここで，GPU 上では高速な out-of-place マージを使用するとし，また，デバイスメモリに乗せることができる要素数を M とする．マージアルゴリズムの性質から，マージの入力は $M/4$ 以下でなければならないため，$p \leqq M/4$ という不等式制約が成り立つ．したがって，$n \geqq 2ks/M$ 個以上のパーティションに分割をすることで，マージステップにおけるパーティションのマージを GPU 上で行うことができる．また，GPU におけるデータ転送とマージのパイプライン化を考慮する場合，2 つのマージを同時に行うため，各マージで使用できるデバイスメモリは半分になる．よって，$n \geqq 4ks/M$ 個以上のパーティションに分割を行う必要がある．今回は $n = 2^{\lceil \log(4ks/M) \rceil}$ とした．パーティションの分割は CPU 上で並列に実行される．以降，GPU 上でマージが行えるように変更したパーティショニング手法を GPU 用パーティショニングと呼称し，PSMerge で使用されているパーティショニング手法を CPU 用パーティショニングと呼称する．

3.2 ヒストグラムを利用したパーティショニング

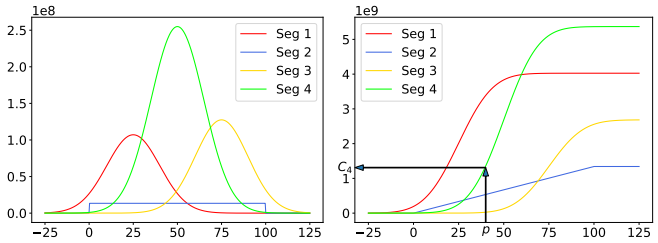
PSMerge では，2.1 節で述べたパーティショニング手法を使用している．このパーティショニング手法では，高速に各セグメントを等しいサイズのパーティションへと分割することができる．しかし，この手法では等しいサイズのセグメントという条件があるため拡張性に乏しい．例えば，2.2 節のソートアルゴリズムでは，前処理として入力配列をセグメントごとにソートするが，CPU と GPU を協調させて高速化する場合，異な

る大きさのセグメントに分割することで CPU と GPU のタスクの負荷均衡を行うことが考えられる．そこで，セグメントのヒストグラムを使用した新しいパーティショニング手法を提案する．

k 個のセグメントでの例を示す．$i(1 \leqq i \leqq k)$ 番目のセグメントの要素数は s_i であり，各セグメントは異なる分布となっている．各セグメントのヒストグラムは図 1(a) のようになる．また，ヒストグラムから累積度数を計算することができる．これを図 1(b) に示す．図 1(b) に示すように，ある値 p に対して，セグメント i における p 未満の要素の個数 C_i は累積度数から求めることができる．よって，ある値 p に対して，入力全体における p 未満の要素の個数 C は $C = C_1 + C_2 + \cdots + C_k$ として求めることができる．各セグメントを n 個のパーティションに均等に分割する場合，各パーティションのサイズは $sk/n (s = s_1 + s_2 + \cdots + s_k)$ であるので，$j(1 \leqq j \leqq n)$ 番目のパーティションの範囲の上界は $skj/n \leqq C$ となる最小の p である．

各セグメントがソート済みであるという性質を利用することで，実際にヒストグラムや累積度数を構築することなく各パーティションの範囲を求めることが可能である．セグメント i について，p 以上の最初の要素のインデックスを二分探索によって求めることができる．セグメント i の開始インデックスを引くことで C_i を $\mathcal{O}(\log(s_i))$ で求めることができる．また，p が増加すると，それに対応する C の値は単調に増加するため，$skj/n \leqq C$ となる最小の p は二分探索により求めることができる．p を探索する範囲は入力全体の最大値 U と最小値 L であるが，これは各セグメントの先頭と末尾から効率的に求めることができる．よって，全体の計算量は，$\mathcal{O}(k + k \log(s) \log(U - L))$ となる．

このパーティショニング手法の特徴は入力されるセグメントの大きさが等しいことを前提としないことである．これによりマージの適用範囲を広くすることができる．また，セグメントの性質を利用することでヒストグラムや累積度数をメモリ上に実際に構築しないため，in-place なアルゴリズムである．考慮すべき点として，PSMerge におけるパーティショニング手法と比較すると，二分探索を 2 重に行うため計算量が増加していることがあげられる．しかし，マージ全体の計算量と比較するとパーティショニングの計算量が与える影響はわずかである．影響の大きさに関しては評価実験にて調査を行った．また，PSMerge におけるパーティショニング手法と異なり値が変化する箇所でパーティションを分割するため，入力の分布によってパーティションのサイズにばらつきが生じる場合がある．PSMerge では，各パーティションはそれぞれ CPU 上で並列にマージされるため，最も大きなパーティションをマージするスレッドの実行時間がマージステップの実行時間になる．同様に，マルチ GPU を使用し，マージ時間を均衡化する場合もマージタスクを各 GPU に割り当てるためパーティションサイズのばらつきにより同様の問題が生じる．今回の実験では，シングル GPU を用いて 3.5 節で提案する GPU を利用したマージにを実行しているため，サイズのばらつきによるマージの計算時間のばらつきはデータの転送時間により隠蔽される．パーティショニングにこの手法を用いる場合，(Hist) を後ろにつけて呼称する．



(a) 4 個のセグメントのヒストグラム　(b) 4 個のセグメントの累積度数

図 1: ヒストグラムを用いたパーティショニング

3.3 シャッフルアルゴリズムの変更

PSMerge での隣り合った 2 つのセグメント内のパーティション群のシャッフルの様子を図 2 に示す．この例では 2 つのセグメント (`Segment A`, `Segment B`) はそれぞれ 8 つのパーティション ($\mathtt{A}_1 \ldots \mathtt{A}_8, \mathtt{B}_1 \ldots \mathtt{B}_8$) に分割されている．シャッフルステップではこれらのパーティションを $\mathtt{A}_1\mathtt{B}_1 \ldots \mathtt{A}_8\mathtt{B}_8$ のようにデータの昇順に並べ替える．

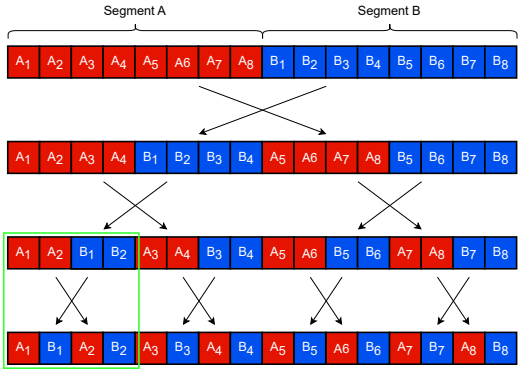


図 2: PSMerge におけるパーティションの移動

PSMerge のような in-place マージではデータのスワップに使用できるデータ領域以外のメモリは限られている．そこで，PSMerge ではパーティションのスワップに `Right Circular Shift` アルゴリズムを用いている．図 2 の緑色の枠内に示されている，$\mathtt{A}_1$，$\mathtt{A}_2$，$\mathtt{B}_1$，$\mathtt{B}_2$ の 4 つのパーティションを対象にこのアルゴリズムでのパーティションの移動を図 3 に示す．図 3 に示すように，スワップする範囲の終わりから少量のデータ (赤い点線の枠で囲まれたデータ) をテンポラリーなメモリに移動させ，残りのデータを右側にシフトさせる．その後シフトによって空いたメモリ領域にデータを移す．この操作をスワップが完了するまで繰り返す．

また，スワップする範囲全体のサイズを n，スワップされる 2 つのパーティションのうち左側にあるパーティションのサイズを k とすると，シャッフルを $G = \gcd(n, k)$ 個のシャッフルに分割し，データの移動を減らすことができる．特に G が CPU コア数よりも多い場合は並列度が高まるため， 最適化が行わ

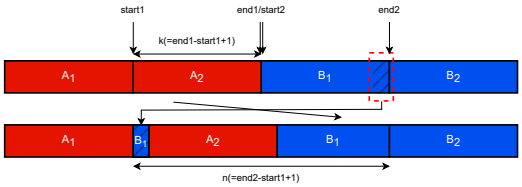


図 3: Right Circular Shift アルゴリズムによるパーティションの移動

れる．

以降，PSMerge で使用されているこのシャッフル手法を `Right Circular Shift` の頭字語を用いて RCS シャッフルと呼称する．

`Right Circular Shift` アルゴリズムではデータの移動回数が多く，最適化できる条件も限定的である．そこで，シャッフルアルゴリズムの高速化としてパーティションサイズに依存せず，データの移動回数を削減した新たなアルゴリズムを使用する手法を提案する．

3.3.1 スワップ回数を最少にしたシャッフルアルゴリズム

`Right Circular Shift` アルゴリズムでは，1 回移動したデータが再び移動することでデータの移動回数が多くなっている．そこでデータの移動先に注目し，1 回のスワップ操作で 1 つの要素の位置が確定するように最適化した新しいシャッフルアルゴリズムを作成した．このアルゴリズムを図 4 に示す．図 4 では初期状態として赤色で示すパーティション A_1 と青色で示すパーティション B_1 が隣り合った状態で与えられている．まず，A_1 と B_1 のサイズを比較し，左側にある A_1 の方が大きいため，A_1 の左側の $|B_1|$ 個の要素と B_1 をスワップする．このスワップにより B_1 の要素は正しい位置に配置された．図 4 では正しい位置に配置された要素は緑色で表している．A_1 のうちスワップされなかった部分を A_2，A_1 の左側から移動してきた部分を B_2 として再帰的にアルゴリズムを適用する．また，A_3 と B_3 のように右側のパーティションサイズ方が大きい場合，右側のパーティションの後ろの要素と左側のセグメントをスワップする必要がある．2 つのパーティションサイズが等しい場合，もしくは片方のパーティションのサイズが 0 となった場合再帰は終了する．また，$M(\geq 1)$ 個以下の要素を同時にスワップすることができると仮定した場合，2 個のパーティションサイズがともに M 以下のときに再帰は終了する．以降，このシャッフル手法をスワップ回数最少シャッフルと呼称する．

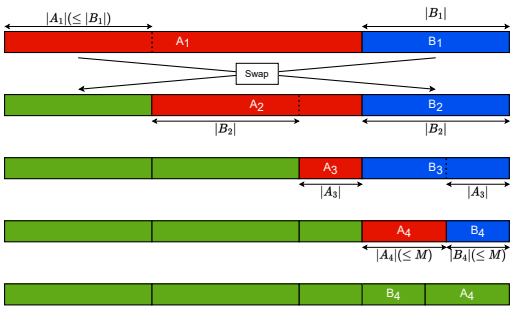


図 4: データのスワップ回数を最少にしたシャッフルアルゴリズム

シャッフルアルゴリズムでの要素どうしのスワップの回数について考察する．シャッフルアルゴリズムに初期状態として与えられる 2 個のパーティションサイズを $|A|$，$|B|$ とする．一般に $|A| \neq |B|$ とすると，2 つの要素のスワップをすると 1 回のスワップで高々1 個の要素の位置が確定する．よってスワップを使用してシャッフルを行う場合，少なくとも $|A|+|B|$ 回のスワップが必要となる．図 4 に示すアルゴリズムでは，1 回のスワップで必ず 1 個の要素の位置が確定し，位置が確定した要素はそれ以上移動しない．したがって，$|A|+|B|$ 回のスワップでシャッフルが完了する．また，$|A| = |B|$ の場合，$(|A|+|B|)/2$ 回のスワップでシャッフルが完了する．以上より，このアルゴリズムはスワップの回数が最小となっている．

また，このアルゴリズムではスワップのためのわずかな外部メモリしか使用していないため，in-place なアルゴリズムである．

3.3.2 リバース演算を使用したシャッフルアルゴリズム

パーティション全体をリバースする演算を用いたシャッフルアルゴリズムを図 5 に示す．初めに，2 個のパーティションを合わせて全体にリバース演算を使用する．これによりパーティションの位置は入れ替わるが，各パーティション内の要素は逆順になっている．そこで，パーティションごとにリバース演算を使用することで正しい順序に並び変える．以上の 3 回のリバース演算によりシャッフルが完了する．以降，このシャッフル手法をリバースシャッフルと呼称する．

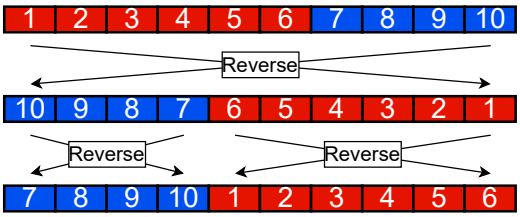


図 5: リバース演算を使用したシャッフルアルゴリズム

3.4 シャッフルの最適化

3.4.1 SIMD 化による最適化

スワップ回数最少シャッフルおよびリバースシャッフルでは 1 要素ずつデータの移動を行う．しかし，このようなデータの移動は CPU の命令あたりの効率が悪く，CPU の演算能力を十分に活用できていない．通常，複数のデータに対して 1 つの命令で同様の操作を適用することによりより効率的なデータの移動を実現することができる．SIMD レジスタを使用したデータの移動の様子を図 6 に示す．スワップ回数最小シャッフルでは 2 個のパーティション `A`，`B` からそれぞれ連続する要素を同時に SIMD レジスタ `R1`，`R2` にロードし，逆のパーティションにストアすることで複数の要素を一度にスワップできる．リバースシャッフルでは，同様にレジスタにロードするが，ストアするまえにレジスタ内で要素を反転する必要がある．シャッフルにこの最適化を使用する場合，(SIMD) を後ろにつけて呼称する．

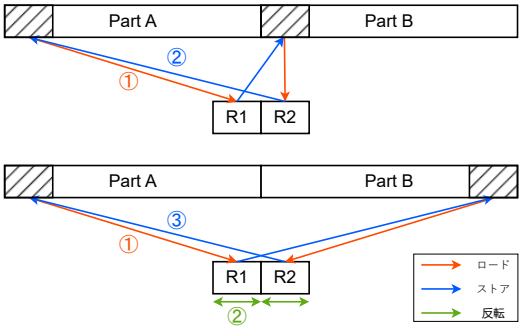


図 6: スワップ回数最少シャッフルのデータの移動 (上) とリバースシャッフルのデータの移動 (下)

3.4.2 シャッフル手法の組み合わせによる最適化

スワップ回数最小シャッフルではパーティションサイズの差または比率が大きな場合，空間的な局所性に乏しい少量のデータの移動を大量に繰り返すことにより性能が低下する場合がある．例として，パーティション A，B のサイズが $|A_1| = 100, |B_1| = 5$ である場合を考える．この場合，$|B_i|$ は常に 5 であるため，$|A_i|$ は 5 ずつしか減少しない．そこで，測定結果に基づき，$|A_i|/(|A_i|+|B_i|) \leqq 0.499$ または $|A_i|+|B_i| \leqq 2^{28}(= 1\text{GB})$ の場合にリバースシャッフルに切り替える．以降，この手法を切り替えシャッフルと呼称する．

3.5 マージステップにおける GPU の活用

マージステップでは，シャッフルステップで並び替えられた隣り合う 2 つのパーティションのマージを行う．GPU 上でのマージには ModernGPU [2] が提供しているマージプリミティブを使用する．ModernGPU では GPU 上のデータに対する汎用的なアルゴリズムを提供しており，また，それらは GPU 上で高速に実行できるように最適化が施されている．

また，デバイスメモリサイズを超えるデータを処理する際，GPU 上でのカーネル実行時間がデータの転送時間に比べて短い場合，ホストメモリとデバイスメモリの間のデータ転送がボトルネックとなる [8]．そこで，CUDA [4] が提供しているストリームを使用し，マージ処理をパイプライン化する．パイプライン化により，GPU 上でのマージと，データ転送がオーバーラップすることにより，マージの処理時間を隠蔽することができる．また，GPU と CPU の間は PCI Express で接続されており，これはホストメモリとデバイスメモリの間での全二重通信が可能となっている．今回のマージではパーティションごとにデータは独立しており，また，マージに処理順序の依存関係が存在しておらず，処理の前後で転送されるデータ量が等しいため，パイプライン化によってデータ転送時間が半分になることが予想される．以降，GPU 上でパイプライン化されたマージ手法を GPU マージと呼称し，PSMerge で使用されている CPU 上でのマージ手法を CPU マージと呼称する．

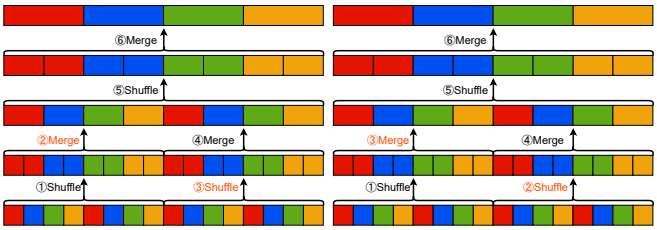
3.6 シャッフルステップとマージステップの実行順序

3 個以上のセグメントが入力された際，複数のパーティションに分割された隣接する 2 個のセグメントをシャッフルとマージを使用し 1 個のセグメントにマージする操作を繰り返し行う．

PSMerge でのシャッフルステップとマージステップの実行順序を図 7(a) に示す．PSMerge では，マージができるようになった際に優先的に実行されるため，初めに前半 2 個のセグメントでシャッフルが行われたあと，そのまま前半 2 個のセグメントのマージが行われている．その後，後半 2 個のセグメントに対し同様にシャッフルとマージが行われる．

シャッフルを優先的に実行した場合を図 7(b) に示す．図 7(a) との違いは，前半 2 個のセグメントのシャッフルのあと，先に後半 2 個のセグメントのシャッフルを行う点である．

この 2 通りの順序を区別するため，PSMerge での実行順序をマージ優先，もう一方をシャッフル優先と呼称する．また，シャッフル優先を使用する場合，(SP) を後ろにつけて呼称する．



(a) マージ優先　(b) シャッフル優先

図 7: シャッフルとマージの優先順位による実行順序の違い

4 評価実験

3 節では，PSMerge の 3 つの処理 (パーティショニング，シャッフル，マージ) それぞれについて新たな手法を提案した．これらの組み合わせパターンにより様々な k-way in-place マージの実現パターンが考えられる．そこで，3 節で提案した各手法を組み合わせ表 1 に示す実装パターンを作成した．

PSMerge について，Salah らによる実験 [6] では正規分布により生成されるデータセットのみでの評価となっている．実アプリケーションではデータの分布について仮定できない場合が存在するため，さらなるデータセットを用いデータの分散や偏りによる実行時間の影響についても評価を行う．

表 1: PSMerge と提案手法において使用される手法

	パーティショニング	シャッフル	マージ	実行順序
PSMerge	CPU 用	RCS	CPU	マージ優先
提案手法 RCS	GPU 用		GPU	
提案手法 MinSwap		スワップ回数最少		
提案手法 RCS(SIMD)		RCS(SIMD)		
提案手法 Switch(SIMD)		切り替え (SIMD)		
提案手法 Switch(SIMD,Hist)	GPU 用 (Hist)			
提案手法 Switch(SIMD,Hist,SP)				シャッフル優先

4.1 実験環境

評価実験に用いたマシンの構成を表 2 に示す．すべての実行ファイルの生成にはコンパイラの O2 オプションを用いた最適化をしている．

表 2: 実験に用いたマシンの構成

OS		Ubuntu 22.04 LTS
CPU		Intel Core i7-7800X (3.5GHz, 6 コア)
GPU		NVIDIA RTX 3060 (Ampere)
	CUDA コア	3584
	ベースクロック	1780MHz
	メモリ	12GB GDDR6
RAM		128GB
CUDA		CUDA 12.0
C++コンパイラ		g++ 11.0

4.2 比較手法

PSMerge では，著者らにより実装が公開されている [1]．しかし，この実装では入力セグメントサイズが 32bit 符号付き整数の上限までとなっている．そこで，ソースコードの該当部分を修正し，64bit 符号付き整数の上限までのセグメントに対してマージが行えるように修正を行った．その他の部分についてはオリジナルの実装を修正せずに使用している．

4.3 データセット

データの偏りやデータの分布範囲による実行時間の変化を評価するため，表 3 に示すデータを作成した．評価に使用するデータは 32bit 符号付き整数とした．データセット 1 からデータセット 4 はそれぞれデータサイズとセグメント数を変化させた際の実行時間への影響の評価に使用した．データセット 5 およびデータセット 6 はペアになっており，一様分布のデータの分布範囲と，正規分布の 3σ の範囲が一致するようになっている．データセット 5 とデータセット 6 での実験結果を比較することで分布の種類によらない分布の範囲による影響を評価した．

表 3: 実験に使用したデータセット

	確率分布	パラメータ	データサイズ	セグメント数
データセット 1	一様分布	$a=0, b=2^{30}$	10 ～ 100GB (10GB 刻み)	$2^i (i=1\cdots7)$
データセット 2	一様分布	$a=0, b=100$		
データセット 3	正規分布	$\mu=200, \sigma=20$		
データセット 4	Zipf 分布	$s=1.1$		
データセット 5	一様分布	$a=0, b=10^i (i=2,5,9)$	50GB	16
データセット 6	正規分布	$\mu=5.0\times10^i$ $\sigma=1.5\times10^i (i=1,4,8)$		

4.4 実験内容

本実験では，PSMerge，提案手法ともに実行時間の測定を行い，提案手法についてはさらにプロファイルの取得を行い，ホスト・デバイス間でのデータ転送の詳細や実行時間の内訳を調査した．実行時間の計測には C++の標準ライブラリが提供している関数やクラスを使用し，プロファイルの取得には Intel VTune を使用した．また，提案手法の実行時間計測ではホストメモリからデバイスメモリへ入力を転送する前から，デバイスメモリからホストメモリへ結果を転送し終えるまでを計測した．したがって，ホストとデバイス間のデータ転送時間も実行時間に含まれる．各実験データごとに 5 回の測定を行い，その算術平均を実行時間の測定値とした．

4.5 実行時間の比較

データセット 1 から 4 に対して表 1 に示す手法のうち PSMerge，提案手法 RCS，提案手法 MinSwap で実行時間の測定を行った．実験結果を図 8 に示す．

まず初めに，各グラフについて入力データのセグメント数が増加すると実行時間が増加していることがわかる．増加率について着目すると，PSMerge では，分布範囲の最も広いデータセット 1 の場合に他の分布に比べ増加率が大きくなっている．PSMerge について，100GB の入力に対してセグメント数 128 の場合の実行時間を比較すると，図 8(a) では 596 秒，図 8(c) では 469 秒，図 8(d) では 586 秒であるのに対し，図 8(b) では 731 秒と実行時間が長くなっている．このことから，PSMerge ではデータの範囲が広い一様分布に対して性能が落ちるということがわかる．

次に，マージを GPU 上で実行することによる実行時間の変化について PSMerge と提案手法 RCS を比較する．データセット 1 では最も実行時間が高速化されており，データセット 2 から 4 でもセグメント数が多い場合に性能が向上している．

さらに，提案手法 RCS と提案手法 MinSwap を比較し，シャッフルを未最適化状態のスワップ回数最少シャッフルに変更した際の実行時間を評価する．これらを比較すると入力データのセグメント数が小さいときは実行時間に大きな差はないが，セグメント数が大きくなるにつれて実行時間の比が大きくなり，実行速度が低下していることがわかる．また，入力データが生成された分布の種類によって実行時間に大きな差は生じていないことから，分布の種類によらず性能は一定であることがわかる．

4.6 分布範囲による影響の評価

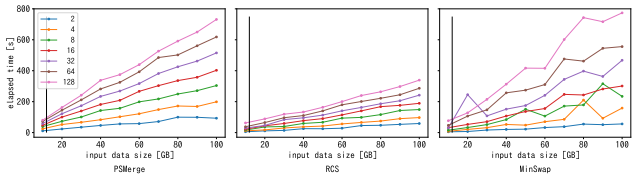
以上の実験結果から，データが分布している範囲が大きく性能に影響を与えることが分かった．そこでデータセット 5 とデータセット 6 を使用してデータの範囲によって実行時間がどのように変化するか評価を行った．結果を図 9(a)，図 9(b) に示す．シャッフルの実行時間について注目すると，PSMerge においてデータの分布範囲が広くなるほど実行時間が短くなっているが，スワップ回数最少シャッフルを使用した提案手法 MinSwap 以外ではほぼ一定となっている．また，提案手法 Switch(SIMD) では提案手法 MinSwap と比較するとすべてのケースにおいてシャッフルの実行時間が高速化されることにより実行速度が改善されていることがわかる．

さらに，マージの実行時間について注目すると分布範囲が広くなるほど CPU マージを使用している PSMerge では実行時間は増加し，GPU マージを使用している提案手法ではほぼ一定となっている．全体の実行時間に着目すると，両分布ともデータの分布範囲が広くなるほど実行時間の差が大きくなり，より高速化されていることがわかる．

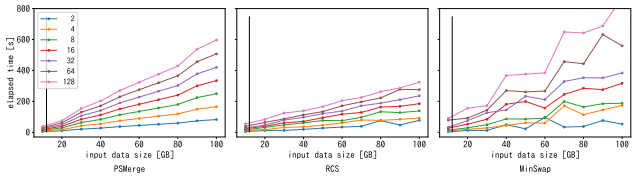
そして，ヒストグラムを使用したパーティショニング手法では，パーティショニングの実行時間は提案手法 Switch(SIMD,Hist)

1：https://github.com/ahmad007/PS-merge

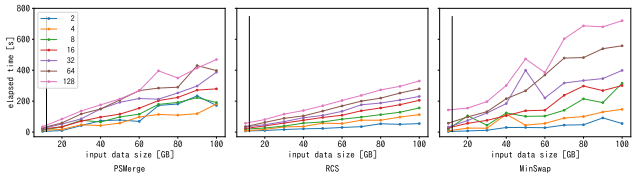
では提案手法 Switch(SIMD) と比較して 4.45 倍に増加したが，全体の実行時間に占める割合は 0.038%であり，与える影響はわずかであることもわかる．



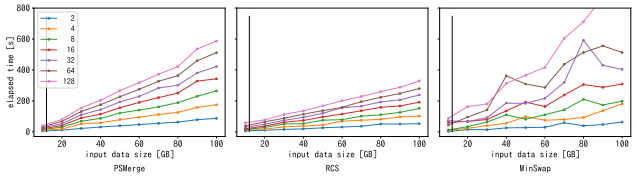
(a) データセット 1(一様分布，$0 \sim 1.0 \times 2^{30}$)



(b) データセット 2(一様分布，$0 \sim 1.0 \times 10^2$)

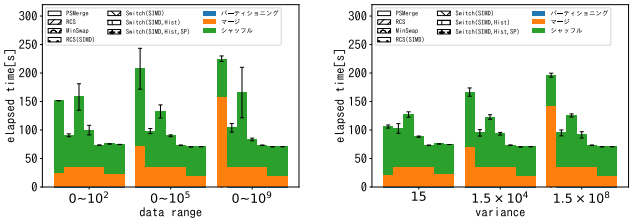


(c) データセット 3(正規分布，$\mu = 200, \sigma = 20$)



(d) データセット 4(Zipf 分布，$s = 1.1$)

図 8: 3 種類の手法での実行時間の比較



(a) データセット 5　　(b) データセット 6

図 9: 分布の範囲を変化させた場合の実行時間の内訳

4.7 ヒストグラムを使用したパーティショニングによる偏りの評価

ヒストグラムを使用したパーティショニング手法によるパーティションサイズのばらつきを評価した．結果を表 4 に示す．結果から，両データセットともデータの分布範囲が大きくなるとパーティションサイズの最大値と最小値の差は小さくなっていることがわかる．また，データの分布範囲が小さい場合，正規分布であるデータセット 6 ではサイズが 0 のパーティションができているのに対して，一様分布であるデータセット 5 では偏りはあるもののサイズが 0 のパーティションはできなかった．

表 4: パーティションサイズの最大・最小値と統計量

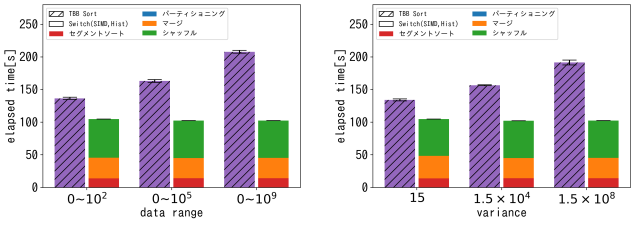
データセット	パラメータ	最大	最小	平均	分散
	$a = 0, b = 10^2$	265807984	132874735	3.00×10^8	4.31×10^{15}
データセット 5	$a = 0, b = 10^5$	209801236	209623171	3.00×10^8	3.93×10^9
	$a = 0, b = 10^9$	209715213	209715183	3.00×10^8	4.50×10^1
	$\mu = 5.0 \times 10^1, \sigma = 1.5 \times 10^1$	356713195	0	3.00×10^8	1.46×10^{16}
データセット 6	$\mu = 5.0 \times 10^4, \sigma = 1.5 \times 10^4$	212297051	206985409	3.00×10^8	1.27×10^{12}
	$\mu = 5.0 \times 10^8, \sigma = 1.5 \times 10^8$	209715519	209714878	3.00×10^8	1.47×10^4

4.8 実アプリケーションでの応用

2.2 節のヘテロジニアスなソートアルゴリズムでは，マージに CPU 上の並列 k-way out-of-place マージを使用していた．GPU 上でのパイプライン化されたセグメントごとのソートと，提案手法 Switch(SIMD,Hist) を組み合わせることでヘテロジニアスな並列 in-place マージソートを実装した．

評価実験にはデータセット 5 およびデータセット 6 と同じ分布から生成されたデータを用いた．比較手法には，Intel oneAPI Threading Building Blocks (oneTBB) [2]の並列ソートを使用した．測定はソートされていない 1 本の 50GB の int 型のデータが格納された配列がメモリ上に用意された段階からソートが完了するまでとした．測定結果を図 10(a)，10(b) に示す．

提案手法は一様分布で 1.30 倍から 2.03 倍，正規分布で 1.30 倍から 1.87 倍の高速化となった．比較手法では，一様分布，正規分布ともに分布範囲が大きくなるほど実行時間が増加しているのに対し，提案手法ではすべての結果で一定の実行時間になっている．



(a) データセット 5(一様分布)　　(b) データセット 6(正規分布)

図 10: 分布の範囲を変化させた場合のソートの実行時間の内訳

4.9 考　　察

実行時間の比較では，データセット 1(一様分布，$0 \sim 2^{30}$) での提案手法 RCS が最も高速化された．また，分布範囲による影響の評価では PSMerge の実行時間が分布範囲が大きくなるにつれ増加しているのに対して，提案手法では実行時間は大きく変化しなかった．このような結果が得られた理由として，データの分散が影響していると考えられる．実行時間に大きな変化が得られなかったデータセットはいずれも狭い範囲にデータが

2：https://www.intel.com/content/www/us/en/developer/tools/oneapi/onetbb.html

偏っており，実験結果から実行時間におけるマージが占める割合が低く，シャッフルが占める割合が多くなっていることがわかる．逆に，分散の大きな分布ではマージの占める割合が多く，シャッフルが占める割合が小さくなっている．CPU マージではデータを比較し，比較されたデータが昇順になっていない場合にのみスワップを行っている．そのため，データ分布の分散が小さく，同じパーティション内で同じ数値のデータが多数並んでいるような状況ではスワップの回数が減少する．RCS シャッフルでは，3.3 節で述べた通り，2 本のセグメントでシャッフルする範囲のサイズによって最適化がなされている．データに偏りがある場合，パーティションを構成しているサブパーティションのサイズが均等ではなくなるため，最大公約数が小さくなる可能性が高まる．その結果，最適化されていないシャッフルが実行されるためシャッフルステップの占める割合が高くなる．

4.5 節で確認したように，提案手法 MinSwap が提案手法 RCS と比較して性能が低下したことに関して，3 個の原因が考えられる．第 1 に，シャッフルアルゴリズムの実装を再帰関数を用いて行っていることである．スワップ回数最少シャッフルでは 2 つのパーティションをスワップした後，まだスワップしていないパーティションに対して再帰的にアルゴリズムを適用する．例えば，分布の種類によらず 5GB かつセグメント数 2 のデータセットではシャッフルを行う関数が約 1.4×10^5 回呼ばれている．そこで，再帰処理をループに書き直すことで同じデータセットで関数の呼び出し回数を 3〜4 回に削減することができる．これにより関数の呼び出しのオーバーヘッドを削減することで高速化できる．第 2 に，スワップ回数最少シャッフルでは並列処理のためのスレッドの生成のオーバーヘッドが大きいことである．RCS シャッフルでは 1 回のシャッフルで 1 回のスレッドの生成が行われる．それに対し，スワップ回数最少シャッフルではスレッドの生成がパーティションのスワップの度に行われている．に，スワップ回数最少シャッフルでは 2 つのセグメントのサイズの最大公約数が小さいほど再帰構造が深くなり，パーティションのスワップ回数が増えるという性質があるため，スレッド生成のオーバーヘッドが無視できなくなるケースがある．第 3 に，サブパーティションサイズに偏りがある場合，スワップ回数最少シャッフルと RCS シャッフルのデータ移動量に差が生じないことである．RCS シャッフルでは，一度に移動できるデータの量は限られているが，片方のサブパーティションのサイズが十分に小さい場合，データの移動回数は両シャッフル手法でほぼ同一となる．切り替えシャッフルではこれらの原因に対して改善を行った結果，シャッフルの実行時間が高速化された．

さらに，ヒストグラムを使用したパーティショニング手法ではデータの分布範囲が広い場合にパーティションサイズの偏りが小さくなった．この手法ではセグメントを値が変化する場所で分割するため，分布範囲が広い場合では分割する候補が多くなる．その結果より正確な場所で分割ができるため偏りが小さくなったと考えられる．分布範囲が狭い場合，正規分布では一様分布よりもより同じ値をとるデータが増えるため偏りが大きくなったと考えられる．

5 おわりに

本研究では GPU メモリサイズよりも大きなデータを対象とした k-way in-place マージ手法に関して提案した．評価実験では，処理できるセグメントサイズを修正した PSMerge と提案手法を実装し，条件をさまざまに変えたデータセットに対し実行時間の比較を行った．実験結果から，マージでは PSMerge に比べ最大 3.07 倍の性能向上を達成した．また，マージソートでは CPU 上の並列ソートと比較し最大 2.03 倍の性能向上を達成した．また，一様分布と正規分布でデータの分布範囲による性能への影響を評価し，分布の種類によらず分布範囲が広い場合には実行時間が高速化されることを確認した．さらに，パーティショニング手法の変更によるパーティションサイズや実行時間への影響を評価した．これらの結果から，シャッフルやマージの高速化手法とその背景を明らかにすることができた．

今後の課題として次のようなものがあげられる．まず，整数データだけでなく文字列などのデータに対しても拡張を行う必要がある．また，今回提案した GPU マージ手法では CPU は GPU へのデータ転送のみを行っているが，CPU 上でのマージ処理を同時に行うことで負荷均衡による高速化が期待できる．さらに，マージソート以外のアプリケーションを実装し，実際のデータでの評価を行うことで新たな課題の発見を期待できる．

謝　辞

本研究は、JST CREST JPMJCR22M2 の支援を受けたものである。

文　献

[1] CUDPP. https://cudpp.github.io/.
[2] Sean Baxter. moderngpu 2.0. https://github.com/moderngpu/moderngpu/wiki, 2016.
[3] Greg N. Frederickson and Donald B. Johnson. The complexity of selection and ranking in x + y and matrices with sorted columns. *Journal of Computer and System Sciences*, Vol. 24, No. 2, pp. 197–208, 1982.
[4] Michael Garland, Scott Le Grand, John Nickolls, Joshua Anderson, Jim Hardwick, Scott Morton, Everett Phillips, Yao Zhang, and Vasily Volkov. Parallel computing experiences with cuda. *IEEE Micro*, Vol. 28, No. 4, pp. 13–27, 2008.
[5] Michael Gowanlock and Ben Karsin. Sorting large datasets with heterogeneous cpu/gpu architectures. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 560–569, 2018.
[6] Ahmad Salah, Kenli Li, Qing Liao, Mervat Hashem, Zhiyong Li, Anthony T. Chronopoulos, and Albert Y. Zomaya. A time-space efficient algorithm for parallel k-way in-place merging based on sequence partitioning and perfect shuffle. *ACM Trans. Parallel Comput.*, Vol. 7, No. 2, may 2020.
[7] Elias Stehle and Hans-Arno Jacobsen. A memory bandwidth-efficient hybrid radix sort on gpus. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, p. 417–432, New York, NY, USA, 2017. Association for Computing Machinery.
[8] 三浦 真矢, 常 穹, 宮崎 純. GPU メモリサイズより大きなデータを対象としたデータ並列プリミティブの開発. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022), K24–2, 2022.

GPU 直接 IO を用いた大規模データに対するハッシュ結合の実行方式

三浦　優也†　　小沢　健史†　　合田　和生†

† 東京大学 生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{ymiura,ozawa,kgoda}@tkl.iis.u-tokyo.ac.jp

あらまし　関係データベース上の分析的な処理を高速化するために，GPU の大きなメモリ帯域幅と大規模な並列計算能力が活用されてきている．しかし，GPU はメモリ容量の小ささや CPU との通信コストがネックであり，従来の GPU データ処理では，GPU メモリや CPU メモリに収まる大きさのデータセットを想定することが多い．一方で，近年，GPU メモリとストレージとの間の直接 IO を可能にする技術である GPUDirect Storage（GDS）が登場し，GPU がストレージ上のデータを高速に扱えるようになった．本論文では，分析処理の基本的な二項演算である結合を取り上げ，GDS を用いた大規模データのハッシュ結合の実行方式を提案し，評価した．

キーワード　データベース技術，先進ハードウェア活用・GPU，問合せ処理

1　はじめに

ビッグデータの台頭とストレージのコストの低下により，世界中で生み出されるデータの総量はここ数十年間で指数関数的に増加している．データを管理するためのデータベースシステムは，身近にあって欠かせない基盤であるが，このかつてないデータ量の増加に対応することが求められている．今でもプロセッサやストレージの処理速度は年々向上しているものの，ムーアの法則は物理的制約によって限界を迎えてきている．データベースを高速化する方法として，ハードウェア単体の処理能力を上げるスケールアップ設計では限界があるといえる．

Graphics Processing Unit（GPU）は，CPU よりも小型のコアを多数備えており，メモリ帯域が高く，大規模な並列演算が可能であるという特徴を持つ．GPU は元来，画像処理に特化したプロセッサであったが，特に NVIDIA が 2007 年に統合開発環境 CUDA [1] を発表したことなどにより，GPU を用いた汎用計算（general-purpose computing on GPUs, GPGPU）が，特に高性能計算（high-performance computing, HPC）や機械学習の分野を筆頭として盛んになってきた．データベースシステムへの活用も例外ではなく，例えば関係データベースにおいてはレコードが決まった形で多数並んでいるため，各行に対する問合せの実行は GPU での処理に適しており，CPU での処理から置き換えるのに十分な高速化を果たしている．

一方で，GPU は通常 CPU よりもメモリサイズが小さく，また，CPU とデータをやりとりするコストが大きいという難点がある．このため，GPU で関係データベースのデータを処理する場合，データの大きさに応じて対応が異なる．GPU メモリに載るサイズのデータを想定する場合，高速に処理できるが，扱えるデータのサイズはとても小さい．CPU メモリに載るサイズのデータを処理する場合，CPU メモリとの通信コストをうまく隠蔽および削減する必要があり，こうした研究は多い．CPU メモリに載らないサイズのデータを処理する場合，従来は，ストレージから CPU メモリにデータを読み込み，GPU メモリに転送するという二重のデータの移動が必要であった．

本論文では，NVIDIA が近年提供を開始した，GPU メモリとストレージとの間の直接のデータのやりとりを可能にする技術である GPUDirect Storage [2] を活用することにより，メモリに載らないストレージ上の大規模なデータを GPU を用いて効率的に処理する方法を扱う．とりわけ，関係データベース上の分析的な処理に用いられる基本的な二項演算である結合演算に注目して，メモリサイズに制限があるときであっても大規模なデータを効率的に処理することのできる GRACE ハッシュ結合 [3] のアルゴリズムを GPU に適用する手法を提案し，その性能を評価する．

本論文の構成は以下の通りである．第 2 章では，研究の背景として GPU 直接 IO と GRACE ハッシュ結合を紹介する．第 3 章では，GPU 直接 IO を用いた GRACE ハッシュ結合による大規模データの処理方式を提案し，第 4 章では，本提案手法の性能を評価する．第 5 章では，CPU を用いた階層的なハッシュ結合，GPU を用いたハッシュ結合，GPU IO などの隣接する研究を紹介し，第 6 章において本論文をまとめる．

2　背　　景

2.1　GPU と IO

GPU は CPU に比べて，より小型で定形処理に特化したコアを多数搭載しており，高い並列性能を誇る．そのため，単純な計算をいくつも行うようなタスクにおいて，CPU よりも格段に高い計算速度と帯域幅を発揮する．一方で，ストレージ上にある大規模なデータを扱う際には，従来，CPU メモリにデータを読み込み，GPU メモリに転送するという二段階のデータの移動を伴っていた．特に，GPU に搭載されているメモリの大きさは CPU に比べて小さく，大規模なデータをメモリのみで扱うことは難しい．

NVIDIA が 2021 年に提供を開始した GPUDirect Storage（GDS）[2] は，ストレージと GPU メモリとの間のデータの直接のやり取りを可能にする技術である．GDS の機能は CUDA

API [1] の一部である cuFile API として提供されており，追加のハードウェアなしに動作させることができる．これを用いることで，コストのかかる CPU–GPU 間のメモリ転送を行うことなく，GPU 直接 IO が可能となる．

三浦ら [4], [5] によれば，GDS による GPU 直接 IO の基本性能は，読み書き，アクセス順序，バッファ長，スレッド数などの様々な項目を変化させたときにおいても，CPU IO と同等である．また三浦らは，GPU 直接 IO を用いて TPC-H [6] の Query 6 を処理する簡易的なエンジンを試作し，GPU がメモリ転送などによって特性を損なうことなく，ストレージ上のデータを扱えることを確かめた．

2.2 GRACE ハッシュ結合

ハッシュ結合（hash join）は，関係データベース上のふたつの表において，それぞれの対象となるフィールドが同一の値を持つようなレコードを結びつける，等結合（equi-join）アルゴリズムのひとつである．ハッシュ結合は，一方の表の対象フィールドのハッシュ値を計算し，ハッシュ表を構築するビルドフェーズ（build phase）と，もう一方の表の対象フィールドのハッシュ値を計算し，構築したハッシュ表から検索するプローブフェーズ（probe phase）からなる．ハッシュ結合は多くの場合，ソートマージ結合やネステッドループ結合といった他の等結合アルゴリズムよりも高速であることが知られている．

ビルド側の表が十分に大きいとき，構築されるハッシュ表はメモリの大きさを超えてしまう．このときは，ビルド側の表を，各部分集合が適切な大きさになるようにパーティションする必要がある．**図 1** に，ビルド側の表 R を n 個にパーティションし，プローブ側の表 S とハッシュ結合したときの動作を示す．表 R，S の大きさをそれぞれ $|R|$，$|S|$ bytes とすると，$|R|/n$ が十分メモリサイズに収まるようにパーティションする必要があり，このときストレージに対して全体で $|R| + n \times |S|$ bytes の読み込みが発生する．

GRACE ハッシュ結合（GRACE hash join）[3] は，ビルド側の表がメモリサイズに対して大きくても，効率的にハッシュ結合を行うための方法である．**図 2** に GRACE ハッシュ結合の動作を示す．GRACE ハッシュ結合のパーティションフェーズ（partition phase）では，ハッシュ結合に用いるハッシュ関数 h_2 とは異なるハッシュ関数 h_1 で対象フィールドのハッシュ値を取り，そこから割り出されるパーティション番号に基づいて R と S をそれぞれ n 個にパーティションし，ストレージに書き出しておく．その後の結合フェーズ（join phase）では，同一番号 i のパーティション $|R_i|$ と $|S_i|$ を通常通りハッシュ結合することを繰り返し，結果を足し合わせる．この結果がハッシュ結合として正しいのは，ハッシュ関数の性質により R，S の各フィールド $r.a$，$s.b$ に対して $h_1(r.a) \neq h_1(s.b)$ のとき $r.a \neq s.b$ であるから，異なる番号に振り分けられたレコードが結合されることはなく，R_i と S_i の組み合わせのみを結合すればよいからである．またこのとき，ストレージに対して全体で $2 \times (|R| + |S|)$ bytes の読み込みと $|R| + |S|$ bytes の書き出しが発生する．

よって従来のハッシュ結合と GRACE ハッシュ結合を比べると，$|R|$ がメモリサイズよりも非常に大きい場合，すなわち n が非常に大きい場合，従来のハッシュ結合では n の大きさに応じて入出力の量が増えるのに対し，GRACE ハッシュ結合では決まった量の入出力しか発生しないという利点がある．

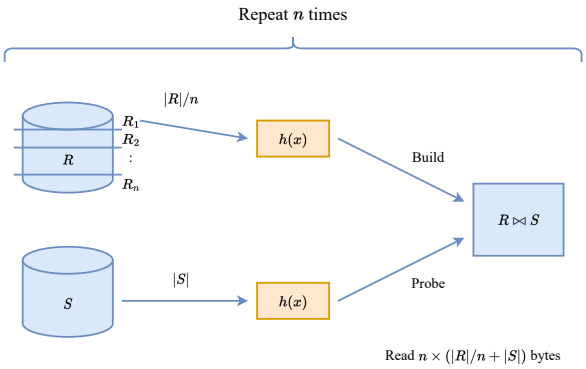


図 1: 従来のハッシュ結合による大規模データの処理
Fig. 1 Classic hash join on a large dataset

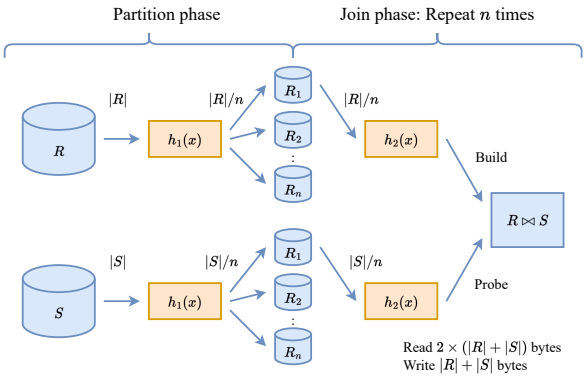


図 2: GRACE ハッシュ結合による大規模データの処理
Fig. 2 GRACE hash join on a large dataset

3 GPU 直接 IO を用いた GRACE ハッシュ結合の実行方式

3.1 GPU 直接 IO を用いた小規模データのハッシュ結合

GPU メモリ上のデータに対するハッシュ結合や，CPU メモリ上のデータを GPU メモリに転送して高速化するハッシュ結合など，GPU を用いたハッシュ結合そのものに関する研究は多くなされている．GPU を用いたハッシュ結合では，多数のコアを並列で動かすという GPU の特性から，ビルドフェーズとプローブフェーズにおいて，ページ内の各レコードを 1 行ずつ処理するのではなく，同時に処理する．

データがストレージ上に存在する場合，ビルド側の表のすべてのページを処理し終えてからプローブ側の表のページを処理していく．ビルドフェーズでは，ストレージ上のビルド側の表の各ページについて，入力バッファにデータを読み取り，各レコードの対象フィールドのハッシュ値を計算し，ハッシュ表に挿入することを繰り返す．プローブフェーズでは，ストレージ上のプローブ側の表の各ページについて，入力バッファにデータを読み取り，各レコードの対象フィールドのハッシュ値を計

算し，構築したハッシュ表を検索することを繰り返す．

本論文では，Google が提供する C++ライブラリ Abseil [7] や Rust の標準ライブラリで用いられる Swiss table をもとに，NVIDIA [8] による `atomicCAS` 命令を利用したハッシュ表の実装を用いる．**図 3** にビルドフェーズに用いるハッシュ表への挿入，**図 4** にプローブフェーズに用いるハッシュ表の検索の擬似コードを示す．ハッシュ表は，エントリーの状態を表す `metadata` と，キーと値を格納する `slots` からなり，`metadata` にはエントリーが空である（`EMPTY`）か，更新中である（`UPDATING`）か，エントリーがあればそのキーのハッシュ値の下位ビット（`h2`）を入れておく．更新中の状態を用意するのは，複数のスレッドが同一のキーで同時に挿入した場合，最初に `EMPTY` を獲得したスレッドが `slots` を更新してから `metadata` を更新するまでに発生するクリティカルセクションを防ぐためである．

```
EMPTY = 0xff
UPDATING = 0xfe

fn insert(key, value)
    h1 = hash(key)
    h2 = h1 & 0x7f
    i = h1 % capacity
    loop
        m = atomicCAS(metadata[i], EMPTY, UPDATING)
        if (m == EMPTY) then
            slots[i] = (key, value)
            metadata[i] = h2
            return
        while (m == UPDATING) do
            m = metadata[i]
        if ((m == h2) && (slots[i].0 == key)) then
            return
        i = (i + 1) % capacity
```

図 3: ビルドフェーズに用いるハッシュ表への挿入の擬似コード

Fig. 3　Pseudocode of hash table insertion in the build phase

```
fn find(key)
    h1 = hash(key)
    h2 = h1 & 0x7f
    i = h1 % capacity
    loop
        m = metadata[i]
        if (m == EMPTY) then
            return null
        else if ((m == h2) && (slots[i].0 == key)) then
            return slots[i].1
        i = (i + 1) % capacity
```

図 4: プローブフェーズに用いるハッシュ表の検索の擬似コード

Fig. 4　Pseudocode of hash table search in the probe phase

3.2　GPU 直接 IO を用いた大規模データの GRACE ハッシュ結合

GPU 直接 IO を用いた GRACE ハッシュ結合について，結合フェーズの各パーティションに対する動作は従来のハッシュ結合の通りでよい．よって，パーティションフェーズの動作を考える．

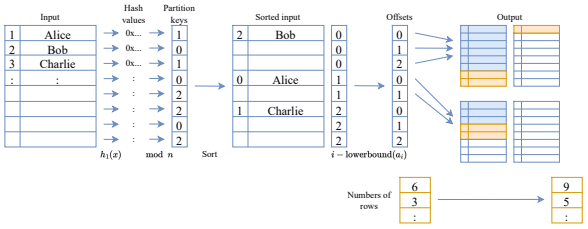


図 5: ソートを用いたパーティションフェーズの実装

Fig. 5　Implementation of the partition phase using a sorting algorithm

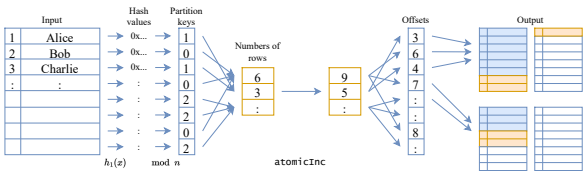


図 6: `atomicInc` 命令を用いたパーティションフェーズの実装

Fig. 6　Implementation of the partition phase using the `atomicInc` instruction

パーティションフェーズでは，ストレージから読み取った入力バッファの各レコードについて，対象フィールドのハッシュ値を計算し，それに基づいてパーティション番号を決定し，対応する出力バッファにレコードを保存する．出力バッファが満たされたら，そのバッファをストレージに書き込む．

GPU を用いてパーティションフェーズを実装する場合，結合フェーズと同様に，ページ内の各レコードを同時に処理するのが適していると考えられる．この際，r 個の行を持つ入力バッファに対して n 個へのパーティションを行う GPU カーネルを呼び出したとき，ひとつの出力バッファには理想的なケースでは r/n 個の行が書き込まれ，最悪なケースでは r 個の行が書き込まれる．このため，出力バッファは各パーティション番号についてふたつずつ用意する必要がある．また，一度のパーティションでいずれかの出力バッファが埋まる可能性があるため，埋まったかどうかの判定を毎回行う必要がある．

本論文では，GPU によるパーティションフェーズの実装をふたつ提案する．ひとつは**図 5** に示すソートを用いる実装である．この実装では，まず各レコードの対象フィールドのハッシュ値を計算し，n で割った余りをパーティション番号とする．次に，パーティション番号に基づいてレコードをソートする．このソートは CUDA に同梱されている CUB ライブラリを用いて十分高速に行うことができる．ソートされたパーティション番号から，二分探索をすることで，同じ番号のグループ内の何番目のレコードかを対数時間で計算することができる．各パーティション番号の出力バッファが何行埋まっているかの情報を持っておき，出力バッファへの書き込み位置を計算できるので，書き込む．行数の情報は更新し，CPU に転送し，埋まっている出力バッファがあれば write を発行する．

もうひとつは**図 6** に示す `atomicInc` 命令を用いる実装であ

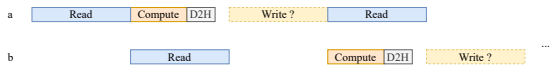


図 7: パーティションフェーズのパイプライン処理
Fig. 7 Pipelining of the partition phase

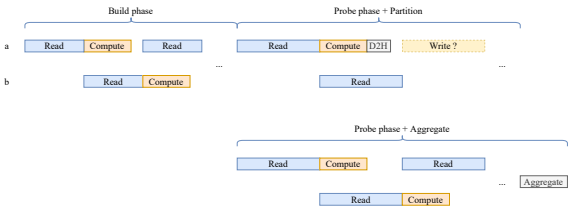


図 8: 結合フェーズのパイプライン処理
Fig. 8 Pipelining of the join phase

る．この実装では，まずソートを用いる実装と同様に，パーティション番号を計算する．次に，各スレッドが出力バッファの行数の情報を `atomicInc` 命令で直接取得・インクリメントすることで，出力バッファへの書き込み位置を割り出し，書き込む．同様に，行数の情報を CPU に転送し，必要に応じて write を発行する．

`atomicInc` を用いる実装の方が単純ではあるが，ひとつのページ内にレコードが多数ある場合，`atomicInc` ではその分のアトミック操作が連なるのに対し，ソートを用いることで，オフセットの計算を並列に行うことができる．

3.3 ダブルバッファリングによるパイプライン処理

GPU における IO，カーネル呼び出し，CPU–GPU データ転送は，独立の機構であり，同時に動作させることができる．よって，各フェーズにおけるこれらの要素を分割し，依存関係に応じてオーバーラップさせることで，一部の処理時間を隠蔽することができる．本論文では，GRACE ハッシュ結合の各フェーズにおいて，ダブルバッファリングを用いて一部の処理時間をオーバーラップさせるようなパイプライン実行方式を提案する．

パーティションフェーズの処理は，入力バッファの読み取り，パーティションのための GPU カーネル呼び出し，CPU への行数情報の転送の 3 ステップ，あるいは必要に応じて出力バッファの書き出しを加えた 4 ステップからなる．ダブルバッファリングを用いてパーティションフェーズをパイプライン処理する実行方式を**図 7** に示す．一連の手順は依存しているが，入力バッファと出力バッファの組を 2 つ用意することで，片方のバッファに読み取っている間に，片方のバッファのパーティションと情報の転送を行う．本論文で後述するように，ひとつのバッファへの読み取りにかかる時間は，GPU カーネル呼び出しにかかる時間や行数情報のみの転送にかかる時間よりも十分に大きいため，このオーバーラップでは読み取り時間の側に律速することになる．

結合フェーズの処理は，ビルドフェーズでは入力バッファへの読み取りとハッシュ表への挿入の 2 ステップを繰り返し，プローブフェーズでは入力バッファへの読み取りとハッシュ表の検索の 2 ステップを繰り返す．ここでは，プローブフェーズの結果をそのまま次の処理につなげる，すなわち GPU カーネル融合を適用することを考え，結果をさらにパーティションする場合と，結果を集計する場合をそれぞれ示す．ダブルバッファリングを用いて結合フェーズをパイプライン処理する実行方式を**図 8** に示す．ビルドフェーズでは，ハッシュ表を 1 つ，入力バッファを 2 つ用意し，片方のバッファに読み取っている間に，片方のバッファから挿入を行う．プローブフェーズでは，入力バッファを 2 つ用意し，片方のバッファに読み取っている間に，片方のバッファから検索を行う．その後パーティションを行う場合は，プローブフェーズにパーティションのパイプライン処理をそのまま適用することができる．その後集計を行う場合は，オーバーラップさせているカーネルの中で各レコードに対する集計処理を済ませ，プローブフェーズ終了後に，CPU への転送などの必要な処理を行う．

4 性能評価

4.1 TPC-H 問合せ実験

提案手法の性能を評価するために，分析的な問合せのベンチマークである TPC-H [6] を使用し，そのうちハッシュ結合を適用できる Query 3 を選択した．Query 3 の内容は**図 9** の通りである．さらに TPC-H の `dbgen` を用いて，スケールファクター（SF）が 10，100 のデータを生成した．Query 3 の実行には `CUSTOMER`，`ORDERS`，`LINEITEM` 表が必要であり，各表の生成される行数や本実装におけるレコード長を**表 1** に示す．また，ページサイズを 256 KiB とし，ひとつのページをレコードの配列に見立て，ストレージに格納した．

```
select
        l_orderkey,
        sum(l_extendedprice * (1 - l_discount)) as revenue,
        o_orderdate,
        o_shippriority
from
        customer,
        orders,
        lineitem
where
        c_mktsegment = 'BUILDING'
        and c_custkey = o_custkey
        and l_orderkey = o_orderkey
        and o_orderdate < date '1995-03-15'
        and l_shipdate > date '1995-03-15'
group by
        l_orderkey,
        o_orderdate,
        o_shippriority
order by
        revenue desc,
        o_orderdate
limit 10;
```

図 9: 実験に用いた Query 3
Fig. 9 Query 3 used in the experiment

また，Query 3 を実行するために作成した実行計画を 10 に示す．`CUSTOMER` をビルド側，`ORDERS` をプローブ側としてハッシュ結合し，これをビルド側，`LINEITEM` をプローブ側として

表 1: Query 3 に用いる表
Table 1 Tables used in Query 3

表	行数	レコード長
CUSTOMER	SF × 150,000	240 bytes
ORDERS	SF × 1,500,000	160 bytes
LINEITEM	SF × 6,000,000	160 bytes

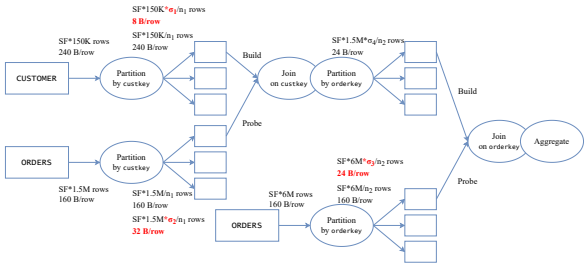


図 10: GRACE ハッシュ結合を用いた Query 3 の実行計画
Fig. 10 Execution plan of Query 3 using GRACE hash join

さらにハッシュ結合を行うが，ビルド側とプローブ側のそれぞれに対してパーティションを行う．図には入出力が必要なレコードの行数とレコードのバイト長を記している．黒字で書かれているのは，入力したレコードをそのままパーティションして出力する場合の行数とバイト長である．赤太字で書かれているのは，結合時に行っていた必要な行のみの選択操作と必要なフィールドのみの射影操作をパーティション時点で行う，すなわち述語プッシュダウンを適用した場合の行数とバイト長である．

4.2 実験環境

上述の実験を**表 2** に示す環境で行った．

表 2: 実験環境
Table 2 Experimental environment

ワークステーション	HP Z4 G4 Workstation
CPU	Intel Xeon W-2245 (8 コア 16 スレッド @ 3.90 GHz)
システムメモリ	64 GB
NVMe SSD	WD_BLACK AN1500
GPU	NVIDIA RTX A4000 (メモリ 16 GB)
ソフトウェア	Ubuntu 22.04 Mellanox OFED 5.8 CUDA 12.2.0

4.3 GPU を用いたパーティションフェーズの実行性能

パーティションフェーズの実装について，ソートを用いる場合と atomicInc 命令を用いる場合を比較する．ソートを用いる場合，パーティション番号を得るカーネルをステップ 1，CUB によるソートをステップ 2，オフセットを算出して書き込むカーネルをステップ 3 とする．ステップ 2 について，マージソート，パーティション番号（64 ビット）による基数ソート，パーティション番号の下位 8 ビットによる基数ソートの各ソートアルゴリズムを比較する．atomicInc では，ひとつのカーネルで処理できるためこれをステップ 1 とする．CUSTOMER 表に対するこの各処理の実行結果を，プロファイラーである NVIDIA Nsight Systems を用いて計測したものを**表 3** に示す．括弧内の数値は，そのステップに含まれる出力バッファへの書き込み処理を除いたときの実行時間である．結果として，ソートを用いる場合は小さいビット幅による基数ソートであれば十分高速に実行できるものの，atomicInc を用いる場合の方が高速である．また，両実装において，とくに時間を占めるのは出力バッファへのレコードの書き込みであることが分かる．atomicInc を用いる場合，小さい分割数ではアトミック処理の衝突が多くなるが，分割数によって実行時間に大きな違いは現れなかった．参考として CPU によるパーティション実装も示したが，GPU 処理と CPU 処理とで実行時間にオーダーレベルでの差はなかった．

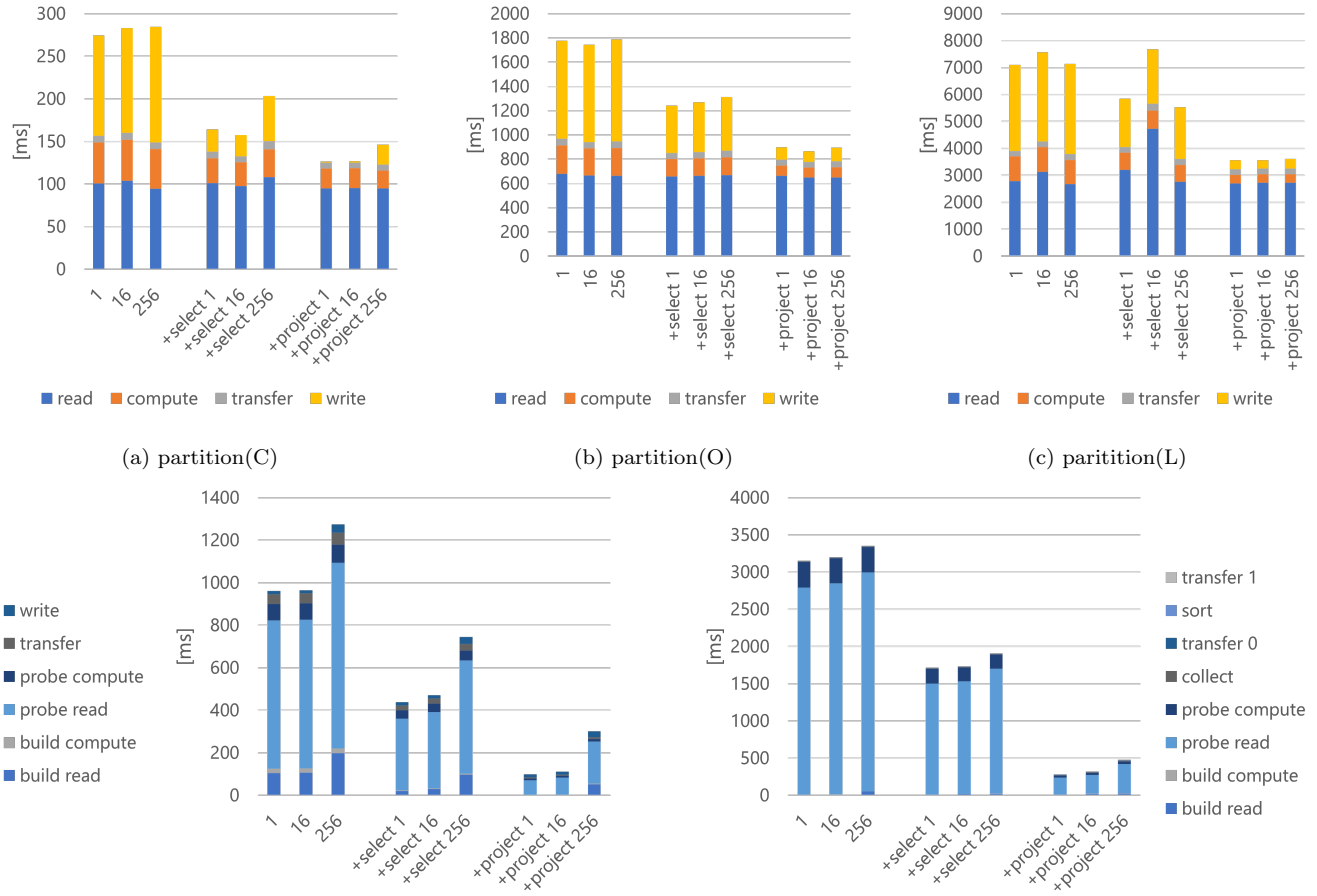
表 3: パーティションフェーズの実装方式ごとのプロファイル結果
Table 3 Profiling result of various implementation methods of the partition phase

方式	分割数	ステップ 1 (µs)	ステップ 2 (µs)	ステップ 3 (µs)
merge sort	1	2.0	22.2	30.2
merge sort	16	2.0	19.9	29.5
merge sort	256	2.0	19.8	29.8
radix sort (64-bit)	1	2.0	25.8	30.2
radix sort (64-bit)	16	2.0	25.8	29.5
radix sort (64-bit)	256	2.0	25.9	29.8
radix sort (8-bit)	1	2.0	6.5	30.2 (4.0)
radix sort (8-bit)	16	2.0	6.6	29.5 (3.8)
radix sort (8-bit)	256	2.0	6.8	29.8 (3.8)
atomicInc	1	28.0 (2.2)		
atomicInc	16	27.2 (2.0)		
atomicInc	256	27.4 (2.2)		
CPU	1	21.6		
CPU	16	23.8		
CPU	256	36.8		

同様に atomicInc を用いて CUSTOMER 表のパーティションフェーズを処理するカーネルを複数同時に実行したときのプロファイル結果を**表 4** に示す．多重度を増やしていったとき，同時に動作させることのできるカーネル数は 10 個程度であると考えられる．また，多重度や分割数を変えても，カーネルひとつあたりの実行時間は大きくは変わらなかった．よって，atomicInc によるアトミック操作は，全体の実行時間に対して，メモリの書き込み時間に比べればほとんど影響を及ぼさない程度に高速に動作させることができるといえる．

4.4 GPU 直接 IO を用いた GRACE ハッシュ結合の実行性能

SF=10 として atomicInc によるパーティションを行ったときの Query 3 の各フェーズの実行時間を**図 11** に示す．横軸に書かれている数はパーティションの分割数であり，各グループは左から述語プッシュダウンをしない場合，選択操作をプッ



(a) partition(C)　(b) partition(O)　(c) paritition(L)

(d) partition(join(C, O))　(e) aggregate(join(CO, L))

図 11: 述語プッシュダウンの有無ごとの GRACE ハッシュ結合を用いた Query 3 の実行時間（SF=10）

Fig. 11　Execution time of Query 3 using GRACE hash join with predicate push-down (SF=10)

表 4: パーティションフェーズの `atomicInc` 方式の多重度ごとのプロファイル結果

Table 4　Profiling result of the `atomicInc` implementation of the partition phase by various multiplicities

方式	多重度	分割数	瞬間最大並列実行数	平均実行時間 (µs)
`atomicInc`	1	1	1	27.9
`atomicInc`	1	16	1	27.2
`atomicInc`	1	256	1	27.3
`atomicInc`	4	1	4	28.4
`atomicInc`	4	16	4	27.9
`atomicInc`	4	256	4	28.2
`atomicInc`	16	1	10	30.3
`atomicInc`	16	16	10	29.3
`atomicInc`	16	256	11	34.6
`atomicInc`	64	1	10	30.4
`atomicInc`	64	16	9	29.2
`atomicInc`	64	256	15	37.1

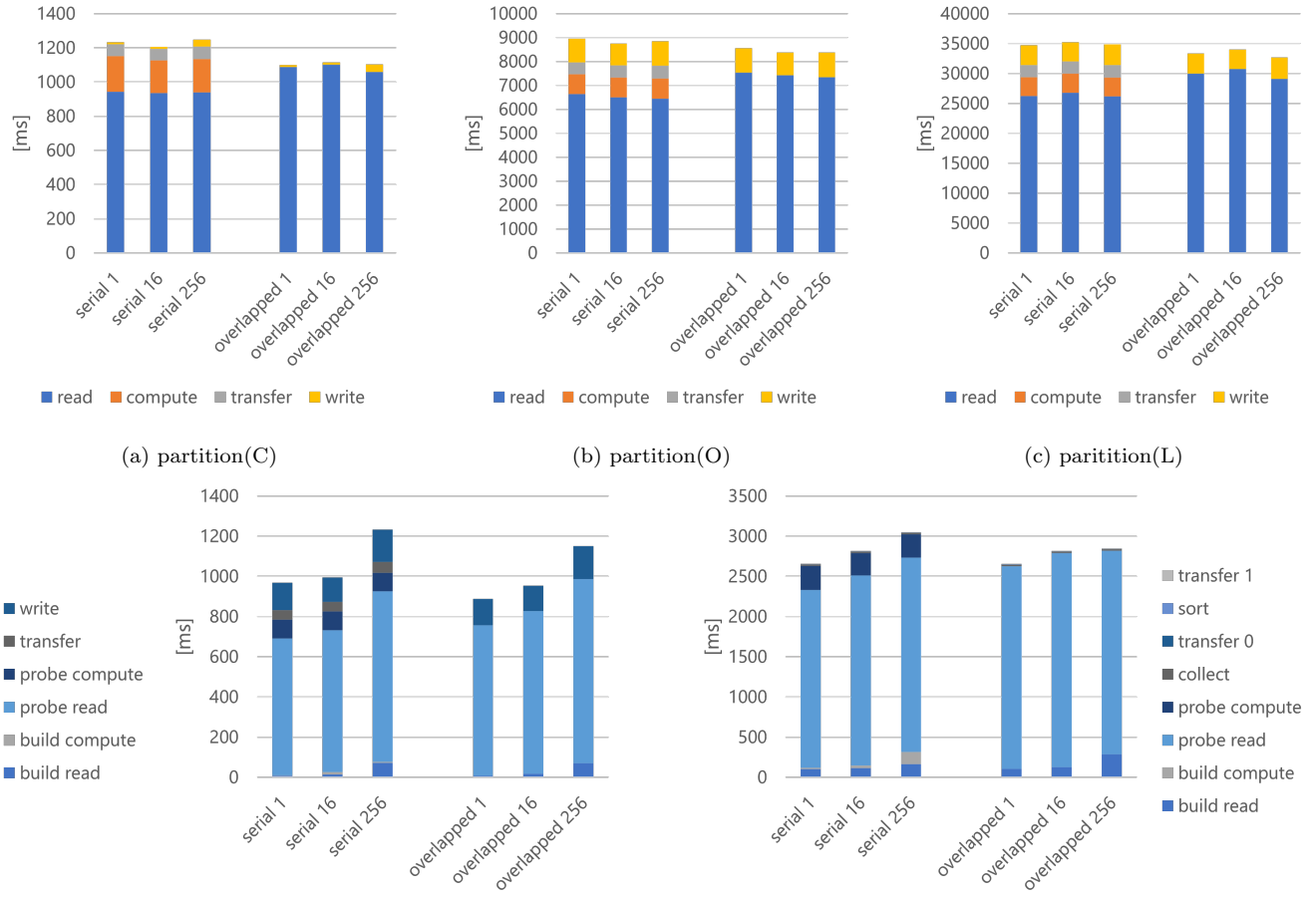
シュダウンした場合，選択操作と射影操作をプッシュダウンした場合を表す．

図 11a，**図 11b**，**図 11c** のパーティションフェーズについて，各表の行数やレコード長に関わらず，カーネル呼び出しやメモリ転送よりも入出力にかかる時間が大きい．また，選択操作のプッシュダウンにより出力時間を，射影操作のプッシュダウンによりカーネル呼び出し中のメモリコピーの時間と出力時間を削減できている．分割数が変わっても基本的には実行時間は大きく変化しないが，出力が小さいときは分割数が大きくなると出力時間も大きくなっている．これは，実装上の原因として，分割数が大きくなったことで埋まり切っていないバッファーが多く出力されて余分な出力が増えたためか，分割数が大きくなったことでストレージのシーケンシャルアクセスから外れてしまい，読み書き速度が低下したためかのどちらかが考えられる．

図 11d，**図 11e** の結合フェーズについても，カーネル呼び出しやメモリ転送よりも入出力にかかる時間が大きい．結合フェーズの入力時間は，パーティションフェーズの述語プッシュダウンによる出力時間と同等である．分割数ごとの実行時間の変化については，パーティションフェーズと同様のことが考えられる．

4.5　ダブルバッファリングによる実行性能の向上

SF=100 として，`atomicInc` によるパーティション実装を用



(a) partition(C)　(b) partition(O)　(c) paritition(L)

(d) partition(join(C, O))　(e) aggregate(join(CO, L))

図 12: ダブルバッファリングの有無ごとの GRACE ハッシュ結合を用いた Query 3 の実行時間（SF=100）

Fig. 12　Execution time of Query 3 using GRACE hash join with double buffering (SF=100)

い，パーティションフェーズ，結合フェーズのそれぞれにダブルバッファリングを適用した場合の Query 3 の各フェーズの実行時間を**図 12** に示す．横軸に書かれている数はパーティションの分割数であり，各グループは左がパイプラインを直列に実行した場合，右がダブルバッファリングによるパイプライン処理を適用した場合を表す．

図 12a，**図 12b**，**図 12c** のパーティションフェーズについて，カーネル呼び出しとメモリ転送にかかる時間を入力にかかる時間に隠蔽したことで，実行時間が削減できている．ダブルバッファリングによって入力にかかる時間が増加しているのは，直列実行において入力の間にカーネル呼び出しと転送が挟まれていたことでストレージのプリフェッチが働いていたのが，パイプライン処理を圧縮したことでプリフェッチの効果がなくなったためと考えられる．実際，本実験と同等の環境で行われた三浦ら [4], [5] の測定結果から計算すると，ページサイズ 256 KiB のときの GPU の入力速度は 3.335,53 GB/s，12,724.0 IOPS ほどであり，`CUSTOMER`，`ORDERS`，`LINEITEM` のシーケンシャルな入力のみにかかる時間は 1080 ms，7200 ms，28,790 ms と推測できる．よって，ダブルバッファリング時の入力時間の方が，プリフェッチのないシーケンシャルな入力時間に近い．

図 12d，**図 12e** の結合フェーズについても，カーネル呼び出しとメモリ転送にかかる時間を入力にかかる時間に隠蔽できている．

5　関 連 研 究

5.1　パーティションを伴うハッシュ結合に関する研究

Chen ら [9] は，GRACE ハッシュ結合におけるストレージとメモリの階層性を，メモリと CPU キャッシュの層に適用し，CPU キャッシュを意識した高速なハッシュ結合を実現した．

5.2　GPU を用いたハッシュ結合に関する研究

Sioulas ら [10] は，GPU のメモリ階層や命令などのハードウェア特性を意識したハッシュ結合の実装を行った．他にも，PCIe や NVLink を介した GPUDirect RDMA 技術により，複数ノード，複数 GPU 上で結合アルゴリズムを動作させる研究として，Guo ら [11]，Paul ら [12]，Lutz ら [13]，Thostrup ら [14] によるものがある．

5.3　GPU 直接 IO に関する研究

Li ら [15] は，NVMe SSD と GPU 間の直接 IO を独自に実

装し，圧縮とデータ転送の最適化を用いる OLAP システムを構築した．Qureshi ら [16] も GPU 直接 IO を非同期的に発行する仕組みを独自に実装した．Ravi ら [17] は，GPUDirect Storage を階層的ファイル形式 HDF5 のライブラリに適用した．Inupakutika ら [18] は，GPUDirect Storage の性能の初期段階的な定量化を行った．

6 おわりに

本論文では，関係データベース上の分析的な処理に用いられる基本的な二項演算であるハッシュ結合について，GPU の並列処理能力および GPU 直接 IO の技術を，大規模データを効率的に結合できる GRACE ハッシュ結合に適用し，その実行方法を提案するとともに性能を測定した．また，さらなる効率的な実行方法として，実行時間をオーバーラップさせるパイプライン処理の方法を提案し，性能を測定した．

今後の展望としては，GRACE ハッシュ結合の部分的な改善であるハイブリッドハッシュ結合（hybrid hash join）への適用や，GPU の共有メモリなどのメモリ階層を活用したさらなる高速化に取り組んでいきたい．また，大域的にパイプライン処理を施し，問合せ実行を全体最適化する方法を考えていきたい．

謝　辞

本研究の一部は，日本学術振興会科学研究費補助金基盤研究 (B) JP20H04191 の助成を受けたものである．

文　献

[1] CUDA C++ Programming Guide. `https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html`. Accessed: 2022-10-25.
[2] NVIDIA GPUDirect Storage. `https://docs.nvidia.com/gpudirect-storage/index.html`. Accessed: 2022-12-03.
[3] Masaru Kitsuregawa, Hidehiko Tanaka, and Tohru Moto-Oka. Application of Hash to Data Base Machine and Its Architecture. *New Gen. Comput.*, Vol. 1, No. 1, pp. 63–74, 1983.
[4] 三浦優也, 小沢健史, 合田和生. GPU 直接 IO を用いたデータベース問合せ処理の検討と予備実験. *DEIM 2023*, No. 2a-7-3, 2023.
[5] 三浦優也, 小沢健史, 合田和生. CPU と GPU を併用する IO インテンシブなデータ処理の実行方式の検討. WebDB 夏のワークショップ 2023, No. C3-3, 2023.
[6] TPC-H Homepage. `https://www.tpc.org/tpch/`. Accessed: 2022-12-04.
[7] abseil / Swiss Tables Design Notes. `https://abseil.io/about/design/swisstables`. Accessed: 2024-01-08.
[8] Maximizing Performance with Massively Parallel Hash Maps on GPUs. `https://developer.nvidia.com/blog/maximizing-performance-with-massively-parallel-hash-maps-on-gpus/`, 2023. Accessed: 2024-01-08.
[9] S Chen, A Ailamaki, P B Gibbons, and T C Mowry. Improving hash join performance through prefetching. In *Proc. ICDE*, pp. 116–127, 2004.
[10] Panagiotis Sioulas, Periklis Chrysogelos, Manos Karpathiotakis, Raja Appuswamy, and Anastasia Ailamaki. Hardware-Conscious Hash-Joins on GPUs. In *ICDE*, 2019.
[11] Chengxin Guo, Hong Chen, Feng Zhang, and Cuiping Li. Distributed Join Algorithms on Multi-CPU Clusters with GPUDirect RDMA. In *Proc. ICPP*, No. 65, pp. 1–10, 2019.
[12] Johns Paul, Shengliang Lu, Bingsheng He, and Chiew Tong Lau. MG-Join: A Scalable Join for Massively Parallel Multi-GPU Architectures. In *Proc. SIGMOD*, pp. 1413–1425, 2021.
[13] Clemens Lutz, Sebastian Breß, Steffen Zeuch, Tilmann Rabl, and Volker Markl. Triton Join: Efficiently Scaling to a Large Join State on GPUs with Fast Interconnects. In *Proc. SIGMOD*, pp. 1017–1032, 2022.
[14] Lasse Thostrup, Gloria Doci, Nils Boeschen, Manisha Luthra, and Carsten Binnig. Distributed GPU Joins on Fast RDMA-capable Networks. *Proc. SIGMOD*, Vol. 1, No. 1, pp. 1–26, 2023.
[15] Jing Li, Hung-Wei Tseng, Chunbin Lin, Yannis Papakonstantinou, and Steven Swanson. HippogriffDB: Balancing I/O and GPU Bandwidth in Big Data Analytics. *PVLDB*, Vol. 9, No. 14, pp. 1647–1658, 2016.
[16] Zaid Qureshi, Vikram Sharma Mailthody, Isaac Gelado, Seungwon Min, Amna Masood, Jeongmin Park, Jinjun Xiong, C. J. Newburn, Dmitri Vainbrand, I-Hsin Chung, Michael Garland, William Dally, and Wen-Mei Hwu. GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture. In *Proc. ASPLOS, Volume 2*, pp. 325–339, 2023.
[17] John Ravi, Suren Byna, and Quincey Koziol. GPU Direct I/O with HDF5. In *2020 IEEE/ACM Fifth International Parallel Data Systems Workshop (PDSW)*, pp. 28–33, 2020.
[18] Devasena Inupakutika, Bridget Davis, Qirui Yang, Daniel Kim, and David Akopian. Quantifying Performance Gains of GPUDirect Storage. In *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pp. 1–9, 2022.

余剰の GPU メモリを活用したアクセラレータ間データ転送

濱地 達也† 田村 雅寿†

†富士通株式会社 〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: †{hamachi.tatsuya, masahisa}@fujitsu.com

あらまし 近年，大容量データの高速処理が求められており，PCI 拡張 BOX を用いて 1 台のサーバに FPGA や GPU 等の複数のアクセラレータを搭載するケースも多くなってきている．一連の処理を行うアクセラレータ間のデータ転送には，ホストメモリ経由でのデータ転送とアクセラレータ間での直接データ転送があり，1 つのデータを複数のアクセラレータに分配して処理するようなケースでは，データのバッファリングが必要であるため前者の方が向いている．また，GPU の搭載メモリ量は増加傾向にあるが，GPU での処理内容によってはメモリを使い切らないケースもありリソース有効活用の面で改善の余地がある．1 つのデータをホストメモリ経由で複数のアクセラレータに分配して処理させるようなケースにおいて，分配先のアクセラレータ数が増えると RootComplex-PCIeSW 間のデータ転送量が増大し，PCIe 帯域ネックによる処理性能劣化が考えられる．本研究ではホストメモリの代わりに余剰の GPU メモリを経由したデータ転送を行い，GPU メモリ使用による性能及び電力への影響を明らかにし，ホストメモリの代替として活用できそうかの基礎実験及び効果の試算を行った．

キーワード 先進ハードウェア活用・GPU，アクセラレータカード，アクセラレータ間データ転送

1. はじめに

デジタルツインコンピューティングのような高精度シミュレーションを必要とする技術などにおいては，計算量が膨大となり，リアルタイム処理を実現するためには高速に大容量のデータを処理する必要がある[1]．その解決策の 1 つとして FPGA や GPU などのアクセラレータカードを使用することが挙げられる．また，サーバの PCI Express スロットを増設できる PCI 拡張 BOX を用いて，1 台のサーバに複数のアクセラレータカードを搭載することもあり[2]，それぞれのアクセラレータが得意とする処理を割り当てることで一連の処理の効率的な高速化を実現している[3][4]．この時，一連の処理を行うアクセラレータ間の転送には，ホストメモリを経由するデータ転送と，アクセラレータ間での直接データ転送があり，1 つのデータを複数のアクセラレータに分配して処理するようなケースでは，データのバッファリングが必要となるため，ホストメモリを経由するデータ転送の方が向いている．

しかしながら，図 1 に示すように，分配先のアクセラレータ数が増えるようなケースでは，RootComplex-PCIeSW 間のデータ転送量が増大し，PCIe 帯域ネックによる処理性能劣化が考えられる．また，アクセラレータカードとして広く採用されている GPU に搭載されるメモリ容量は増加の傾向にあるが，GPU で処理させる内容によってはメモリを使い切らないケースもあり，リソースの有効活用の面で改善の余地がある．

本研究では，ホストメモリを経由するアクセラレータ間のデータ転送において，ホストメモリの代わりに

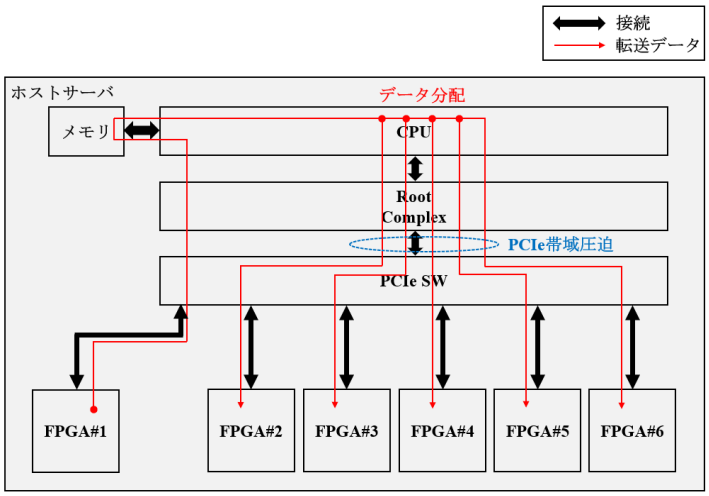


図 1 データ分配先増加による PCIe 帯域圧迫

余剰の GPU メモリを経由したデータ転送を実行した場合の，GPU メモリ使用による性能および電力への影響について述べる．また，GPU メモリ使用による効果の試算を行う．

本論文の構成は以下の通りである．2 章では，GPU リソースの有効活用に関する従来研究について説明する．3 章では，評価を行う構成，GPU メモリ経由データ転送の実現方法，および評価の方法について述べる．4 章にて，余剰 GPU メモリを使用しても電力消費増大などの副作用がないことを評価し，5 章にて余剰 GPU メモリ活用の本来の目的である PCIe 帯域ネックによる処理性能劣化の解消についての試算を行う．最後に 6 章にて本文のまとめ，今後の課題について述べる．

2. 関連研究

GPU リソースの効率利用に関する研究が行われている．Pisacha Srinuan らの研究[5]では，ホストマシンの CPU で実行されるアプリケーションが使用するメモリを余剰の GPU メモリに動的に拡張し，メモリ占有量がホストのメモリサイズを超えるアプリケーションにおける処理の高速化を実現している．Joshua Bakita らの研究[6]では，メモリのオーバーサブスクリプションを可能とするために，CPU メモリバッファではなく GPU を永続ストレージに直接ページングすることでスループットの高速化や同期オーバーヘッドの削減を実現している．このように GPU メモリのメモリリソースとしての活用が注目され，特定のアプリケーションにおける性能向上効果が示されている．

3. 評価方法

本研究にて評価を行う構成を図 2 に示す．この構成では，2 つの FPGA を使用し，FPGA#1 にてフィルタリング/リサイズ処理(Filter/Resize)を行ったデータを FPGA#2 へ転送する．FPGA#2 はデータを受信するのみでフィルタリングなどの処理は行わない(NOP)．構成は，ホストメモリ経由の FPGA 間データ転送(図 2(a))，及び GPU メモリ経由の FPGA 間データ転送(図 2(b))の 2 種類存在する．データ転送は，FPGA に対して経由するメモリの物理アドレスを通知して行う．なお，使用したサーバは富士通製 PRIMERGY RX2540 M6(搭載 CPU：Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz x2，搭載メモリ:32GB x4)であり，使用した FPGA 及び GPU は，AMD 社の Alveo U250 及び NVIDIA 社の T4 である．

ホストメモリ経由の FPGA 間データ転送について，ホストメモリの管理には DPDK[7]を使用した．DPDK の rte_malloc 関数を使用してホストメモリを確保したのち，確保したメモリの物理アドレスを算出する．GPU メモリ経由の FPGA 間データ転送については，NVIDIA の CUDA ライブラリ[8]で提供される cudaMalloc 関数を使用して GPU メモリを確保したのち，NVIDIA ドライバパッケージで提供されるカーネル API[9]の nvidia_p2p_get_pages 関数を用いて物理アドレスを算出する．

評価の観点としては大きく 2 つあり，1 つ目は性能に関しての評価(3.1-3.2 章)，2 つ目は消費電力に関しての評価(3.3-3.4 章)である．

3.1. FPGA 間データ転送の転送性能

ここでは，図 2 における構成にて FPGA 間データ転送の性能をそれぞれ測定し，ホストメモリの代わりに GPU メモリを経由することによる FPGA 間データ転送への性能影響を確認する．測定内容としては，FPGA#1 が最初のデータをホストメモリから受信し，FPGA#2 が最後のデータをホストメモリへ送信するまでの時間とスループットを測定する．

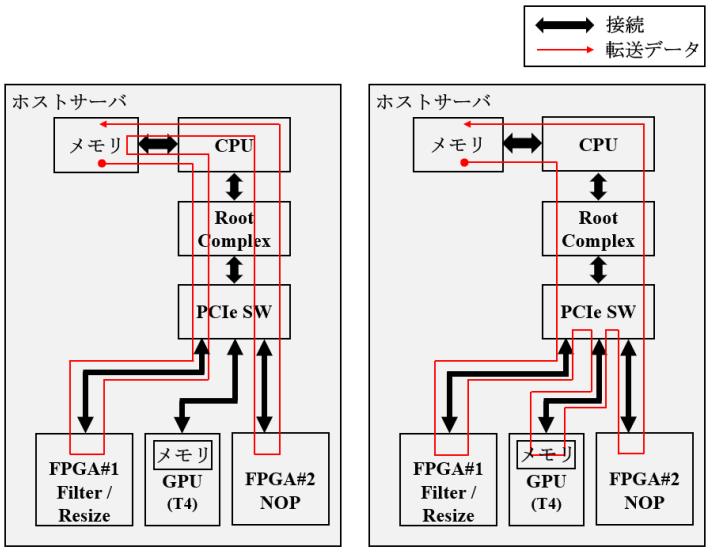


図 2 評価構成

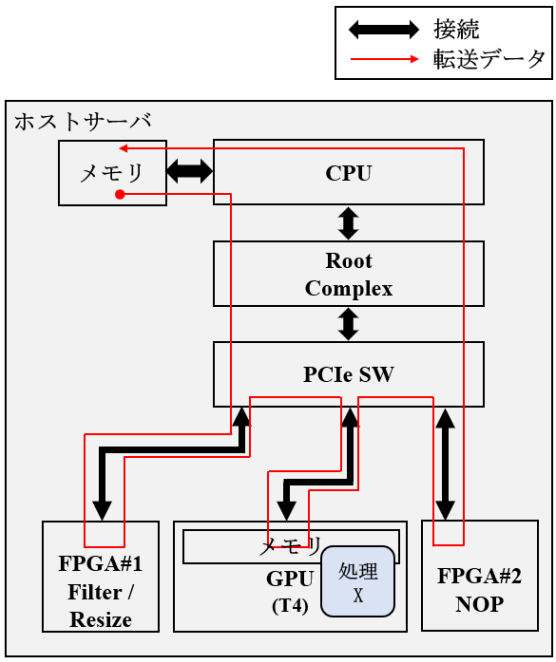


図 3 GPU メモリ経由データ転送+処理 X

3.2. 他の GPU 処理による性能影響

ここでは，図 3 に示すように，FPGA 間データ転送の経由メモリとして使用する GPU にて，処理 X を動かしている状態で，余剰の GPU メモリを経由した FPGA 間データ転送を実行し，FPGA 間データ転送への性能影響，及び，処理 X への性能影響を確認する．なお処理 X として，NVIDIA が提供する DeepStream を用いて推論処理(Yolov3-tiny int8)を実行した．

3.3. GPU メモリ使用による消費電力

ここでは，図 2 における構成にて，各種電力情報をそれぞれ測定し，ホストメモリの代わりに GPU メモリを経由することによる消費電力への影響を確認する．測定する電力情報としては，CPU 電力，GPU 電力を測定する．

3.4. 余剰の GPU メモリ使用による消費電力

ここでは，図 3 及び図 4 に示すように，GPU にて処理 X を実行している状態にて，ホストメモリ経由(図 4)及び GPU メモリ経由(図 3)の FPGA 間データ転送を実行した際の GPU 電力及び GPU 使用率を測定し，余剰の GPU メモリを使用することによる消費電力への影響を確認する．

4. 実験

3.1 における結果を図 5 に示す．ここでは入力データとして，5fps のデータを 150frame 入力し，入力 ch 数を 1ch/8ch/16ch と変えたパターンでの測定を行った．図 5 より各パターンにおけるホストメモリ経由/GPU メモリ経由間の処理時間，及びスループットは同等であることがわかる．この結果から，経由するメモリをホストメモリから GPU メモリに置き換えることによるアクセラレータ間データ転送性能への影響は無いことが言える(性能劣化無し)．

次に，3.2 における結果を図 6 及び図 7 に示す．ここでは処理 X として 15fps 入力の推論処理(1stream)を実行し，入力データとして 15fps のデータを 50frame 入力し，入力 ch 数を 1ch/16ch と変えたパターンでの測定を行った．図 6 は，処理 X を行っている GPU のメモリを経由したアクセラレータ間データ転送を行った際の，アクセラレータ間データ転送性能への性能影響を示すグラフである．図 6 より各パターンにおけるホストメモリ経由/GPU メモリ経由間の処理時間，及びスループットは同等であることがわかる．また，図 7 は，処理 X を行っている GPU のメモリを経由したアクセラレータ間データ転送を行った際の，処理 X への性能影響を示すグラフである．図 7 より，データ転送を実行しなかった場合の処理 X 性能と，各パターンにおける処理 X 性能は同等であることがわかる．これらの結果より，他の GPU 処理(処理 X)を実行中の GPU のメモリを使用することによるアクセラレータ間データ転送性能及び処理 X 性能への影響は無いと言える(性能劣化無し)．ただし，今回の測定において，処理 X を実行中の GPU の SM 使用率及びメモリ使用率はリソースに余裕がある状態であり，GPU のリソースを使い切るような処理が実行されている場合には，性能への影響(劣化)が考えられる．今回は余剰の GPU メモリをターゲットとしているため，GPU リソースに余裕のある状態で行っている．

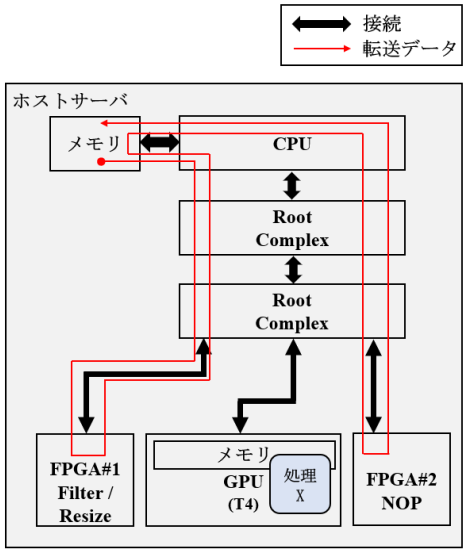


図 4 ホストメモリ経由データ転送+処理 X

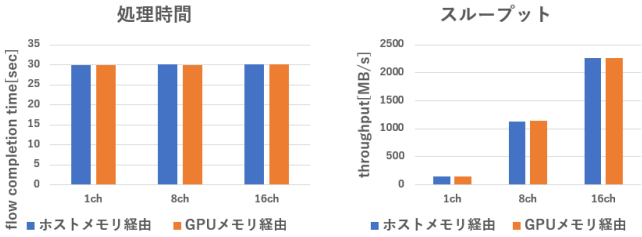


図 5 GPU メモリ使用による転送性能への影響

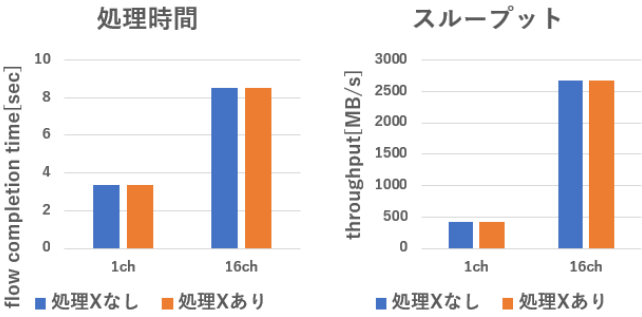


図 6 処理 X による転送への性能影響

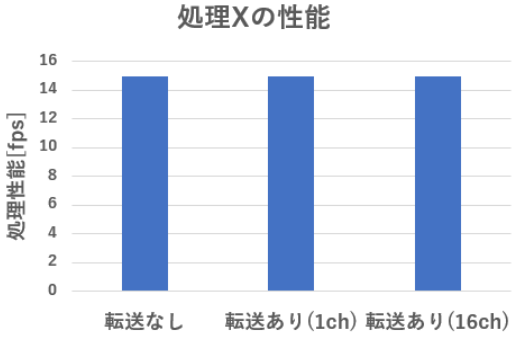


図 7 GPU メモリ使用による処理 X への性能影響

次に，3.3 における結果を図 8 に示す．ここでは入力データとして，15fps のデータを入力し，入力 ch 数を 1ch/8ch/16ch と変えたパターンでの測定を行った．図 8 より CPU 電力については，ホストメモリ経由/GPU メモリ経由間で同等であることがわかる．なお，アンコア周波数が変動すると CPU 電力にブレが生じるため，ここではアンコア周波数を一定(最大)にして測定を行った．GPU 電力については，各パターンにおいて同等の結果であり，GPU メモリ経由の場合の電力の方が約 18W 高い結果が得られた．これは，使用されていない GPU のメモリをアクセラレータ間データ転送に使用したことにより，待機状態での電力から電力増加が発生したことが原因と考えられる．

最後に，3.4 における結果を図 9 に示す．ここでは処理 X として 15fps 入力の推論処理(1stream)を実行し，入力データとして 15fps のデータを入力し，入力 ch 数を 1ch/16ch と変えたパターンでの測定を行った．図 9 より，各パターンにおけるホストメモリ経由/GPU メモリ経由間の GPU 電力及び GPU 使用率は同等であることがわかる．GPU 電力に差分が見られなくなったのは，GPU にて処理 X を実行していることで，待機状態での電力からの電力増加が既に発生している状態であること，尚且つ，アクセラレータ間データ転送に GPU メモリを使用する場合の電力増加はほぼ無かったということが要因として考えられる．この結果から，他の処理(処理 X)を実行している GPU の余剰のメモリを使用してアクセラレータ間データ転送を実施する場合には，消費電力への影響はほぼ無いことが言える．

5. 効果の試算

本章では，GPU のメモリを使用したアクセラレータ間データ転送による効果について試算する．図 10 に示すような，1 つのアクセラレータから複数のアクセラレータにデータを分配して転送するような構成を考える．この構成では，FPGA#0(Decoder)からのデータを，9 つの PCI 拡張 BOX(#1~#9)に搭載した複数の FPGA(Filter/Resize)に対して分配して転送を行う．なお，1 つの PCI 拡張 BOX には GPU1 台と FPGA8 台が搭載されるものとする．分配して転送するデータは 4K 15fps の画像データ(RGB)とし，RootComplex-PCIeSW#0 間の PCIe 帯域幅は片方向あたり 25[GB/s](PCIe Gen4 x16 の理論値 252[Gbps]の 80%)とし，同様に，各 PCI 拡張 BOX の持つ PCIe 帯域幅も 25[GB/s]とする．転送されるデータ量は，分配転送先 1 つにつき，約 0.35GB/s である(3840[width]*2160[height]*3[byte]*15[fps])．この時，RootComplex-PCIeSW#0 間の PCIe 帯域超過が発生する分配数は，72 となる(25[GB/s]/0.35[GB/s])．そのため，図 10 の構成では，RootComplex-PCIeSW#0 間の PCIe 帯域超過が発生する．

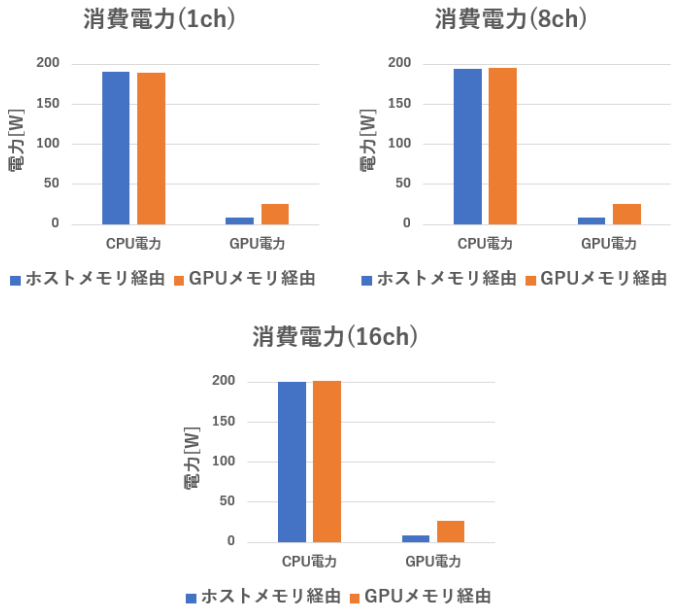


図 8 GPU メモリ使用による消費電力への影響

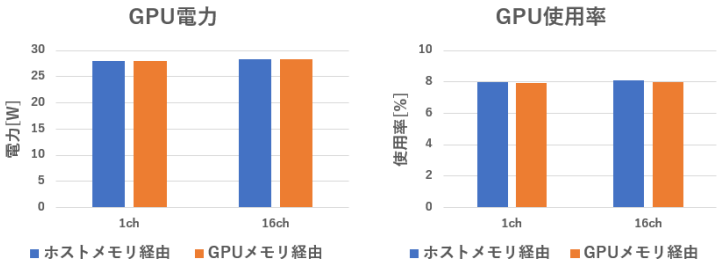


図 9 余剰の GPU メモリ使用による電力への影響

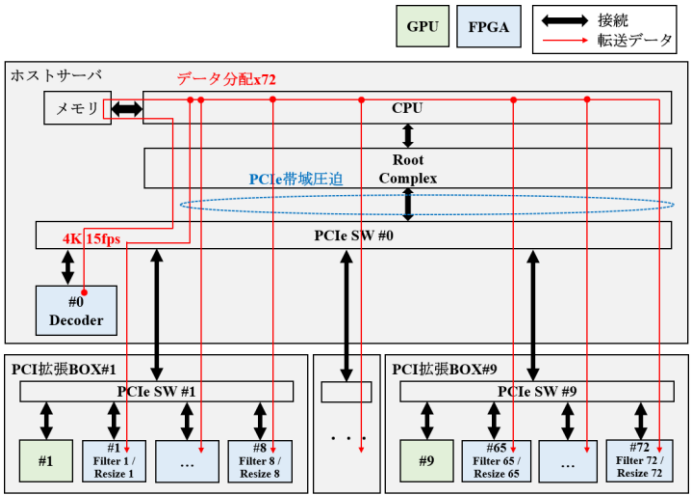


図 10 ホストメモリ経由のデータ分配転送

次に，この分配数 72 という条件において，図 11 に示すような，PCI 拡張 BOX#1~#9 に搭載した GPU のメモリに対してそれぞれデータを転送した後，各 PCI 拡張 BOX 内で 8 ずつ分配して転送を行う場合を考える．

なお，ここで一度ホストメモリを経由しているのは，各 PCI 拡張 BOX に搭載された GPU のメモリへのデータ転送時にもデータの分配(x9)が行われており，分配のためのデータバッファリング用のメモリを FPGA が持たない(分配機能を持たない)ことを想定しているためである．この時，各 PCI 拡張 BOX 内で発生するデータ転送量(片方向)は，約 3.2[GB/s]となる(0.35[GB/s]+0.35[GB/s]*8)．そのため，各 PCI 拡張 BOX 内で PCIe 帯域は超過しない(25[GB/s]以下)．また，RootComplex-PCIeSW#0 間の PCIe 帯域についても，転送されるデータは約 3.2[GB/s](0.35[GB/s]*9)となり，25[GB/s]以下となるため PCIe 帯域は超過しない．

このように，1 つのアクセラレータから 72 のアクセラレータに対して分配転送を行う場合において，全ての分配先に対してホストメモリからデータを転送する場合には PCIe 帯域超過が発生するが，各 PCI 拡張 BOX に搭載された GPU のメモリを経由してデータを転送することによって，PCIe 帯域超過，つまりは PCIe 帯域不足による性能劣化を防ぐことが可能になると考えられる．消費電力については，未使用の GPU(T4)のメモリを使用する場合は 1 台あたり 18[W]の増加，他処理で使用中の GPU(T4)の余剰のメモリを使用する場合は，電力増加無しの見込みである．

6. 結論・今後の課題

本研究では，ホストメモリの代わりに GPU メモリを使用したアクセラレータ間データ転送を行い，GPU メモリを使用することによる性能及び電力への影響を明らかにした．実験の結果，ホストメモリの代わりに GPU メモリを使用することによる性能影響は見られなかった(性能劣化無し)．また，他の処理を行っていない GPU のメモリを使用する場合は，GPU 電力が増加してしまうものの，他の処理を実行している GPU の余剰のメモリを使用することで電力増加を抑えつつ GPU メモリリソースを活用することが可能であった．GPU のメモリを使用したアクセラレータ間データ転送による効果についての試算では，構成次第では GPU メモリを経由した分配転送を行うことにより，PCIe 帯域超過による性能劣化を防ぐことが可能となる見込みとなった．さらに余剰の GPU メモリを使用することで消費電力の増加も抑えることが可能となる見込みである．以上のことから，ホストメモリの代替として余剰の GPU メモリを活用することの有用性が示せたと考える．

今後は，実際に RootComplex-PCIeSW 間のデータ転送量が増大し，PCIe 帯域ネックによる処理性能劣化が発生する構成にて本研究を適用し，性能劣化を解消できることを実証する．

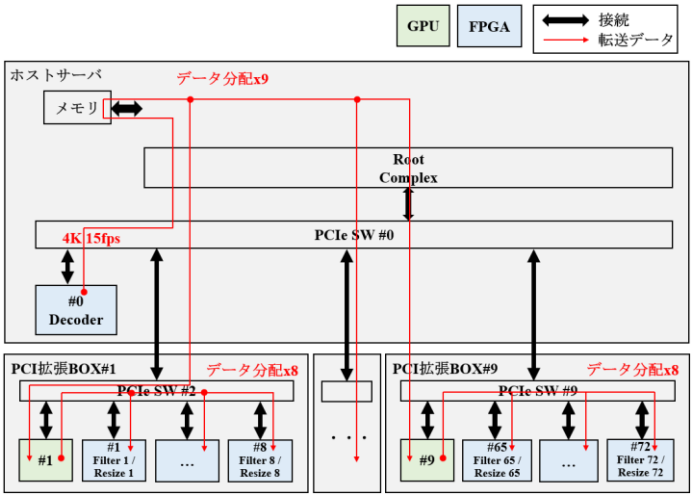


図 11 GPU メモリを使用したデータ分配転送

参 考 文 献

[1] 国立研究開発法人科学技術振興機構 研究開発戦略センター デジタルツインに関する国内外の研究開発動向 https://www.jst.go.jp/crds/pdf/2021/RR/CRDS-FY2021-RR-09.pdf

[2] Young woo Kim, Myung-hoon Oh, Chan-Yeol Park. Development and Evaluation of High-Density Multi-GPU Sub-System Using PCI Express Expansion Hardware. In 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)

[3] Mohammad Alawieh, Maximilian Kasparek, Norbert Franke and Jochen Hupfer. A High Performance FPGA-GPU-CPU Platform for a Real-Time Locating System. In 2015 23rd European Signal Processing Conference (EUSIPCO)

[4] Mehdi Roozmeh, Luciano Lavagno. Implementation of a performance optimized database join operation on FPGA-GPU platforms using OpenCL. In 2017 IEEE Nordic Circuits and Systems Conference (NORCAS): NORCHIP and International Symposium of System-on-Chip (SoC)

[5] Pisacha Srinuan, Purushottam Sigdel, Xu Yuan, Lu Peng, Paul Darby, Christopher Aucoin, and Nian-Feng Tzeng. GPU-Assisted Memory Expansion. In 2021 IEEE International Conference on Networking, Architecture and Storage (NAS)

[6] Joshua Bakita and James H. Anderson. Enabling GPU Memory Oversubscription via Transparent Paging to an NVMe SSD. In 2022 IEEE Real-Time Systems Symposium (RTSS)

[7] DPDK:rte_malloc.h File Reference. https://doc.dpdk.org/api/rte__malloc_8h.html

[8] NVIDIA:NVIDIA CUDA Library: cudaMalloc. https://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/html/group__CUDART__MEMORY_gc63ffd93e344b939d6399199d8b12fef.html

[9] NVIDIA:Developing a Linux Kernel Module using GPUDirect RDMA. https://docs.nvidia.com/cuda/gpudirect-rdma/index.html

FPGA-GPU 間の直接データ転送による消費電力性能効率の向上

飯澤　健† 西山　達也†† サンパトプリヤンカラ†† 石崎　晃朗†† 田村　雅寿†

† 富士通株式会社　〒 211-8588 神奈川県川崎市中原区上小田中 4-1-1
†† NTT ソフトウェアイノベーションセンタ　〒 108-0075 東京都港区港南 1-2-70 品川シーズンテラス 29F
E-mail: ††{iizawa.ken,masahisa}@fujitsu.com, ††{t.nishiyama,priyankara.sampath,teruaki.ishizaki}@ntt.com

あらまし　大量のストリームデータを高速かつ低消費電力で処理したいというニーズが高まっている．一方，近年，多数のアクセラレータを搭載可能な PCIe 拡張 Box が注目を集めている．ストリーム処理のパイプラインを構成する各処理を，最適なアクセラレータで実行することにより，高速かつ低消費電力な処理を実現できる可能性がある．しかし，一般にアクセラレータ間のデータ転送はホストメモリ経由で行われるため，単一サーバに多数のアクセラレータを接続したシステム構成においては，ホストメモリ-PCIe スイッチ間の PCIe リンクがボトルネックになるという問題がある．そこで本研究では，ボトルネックリンク回避のため，パイプラインを実行するアクセラレータ間で直接データ転送を行う転送方式を提案する．FPGA-GPU 間で直接データ転送を行うプロトタイプを用いた評価を行い，ホストメモリ経由転送方式に比べ消費電力性能効率が向上することを示す．

キーワード　FPGA，GPU，DMA

1 はじめに

サイバーフィジカルシステム [1] に代表されるように，映像データやセンサデータ等のストリームデータを収集・分析するシステムが注目を集めている．そうしたシステムを実現するためには，大量のデータを高速に処理する能力だけでなく，データ量に比例して増える消費電力を抑えるための省電力性も要求される．

ムーアの法則の終焉により汎用 CPU の性能や電力効率が頭打ちになる一方で，FPGA や GPU，ASIC 等，異なる性能特性や電力特性を持つアクセラレータが登場している [6]．ストリーム処理のパイプラインを構成する各処理を，最適なアクセラレータで実行するヘテロジニアスコンピューティングにより，高速かつ低消費電力な処理を実現できる可能性がある．

近年，サーバの PCIe スロットを増設可能な PCIe 拡張ボックスが登場している [4]．こうした PCIe 拡張ボックスに多数のアクセラレータを搭載し，複数のサーバ間で柔軟にアクセラレータの割り当てを変更することにより，より低コストでヘテロジニアスコンピューティングを実現することが可能になる．

しかし，一般にアクセラレータ間のデータ転送はホストメモリ経由で行われるため，単一サーバに多数のアクセラレータを接続したシステム構成においては，ホストメモリ-PCIe スイッチ間の PCIe リンクがボトルネックになるという問題がある．

そこで本研究では，ボトルネックリンク回避のため，パイプラインを実行するアクセラレータ間で直接データ転送を行う転送方式を提案する．FPGA-GPU 間で直接データ転送を行うプロトタイプを用いた評価を行い，ホストメモリ経由転送方式に比べ消費電力性能効率が向上することを示す．

本論文の構成は以下の通りである．2 章で従来のアクセラレータ間データ転送技術が抱える問題点を明らかにする．3 章で本論文における提案手法を示す．4 章では，提案手法を実装したプロトタイプの評価結果を述べる．5 章で関連研究の紹介を行い，アクセラレータ間データ転送の分野における本論文の位置づけを明らかにする．最後に，6 章で今後の課題を述べ，まとめとする．

2 従来技術

図 1 のように，サーバに接続された PCIe 拡張ボックスに複数の FPGA と GPU のペアが搭載されたシステムを想定する．

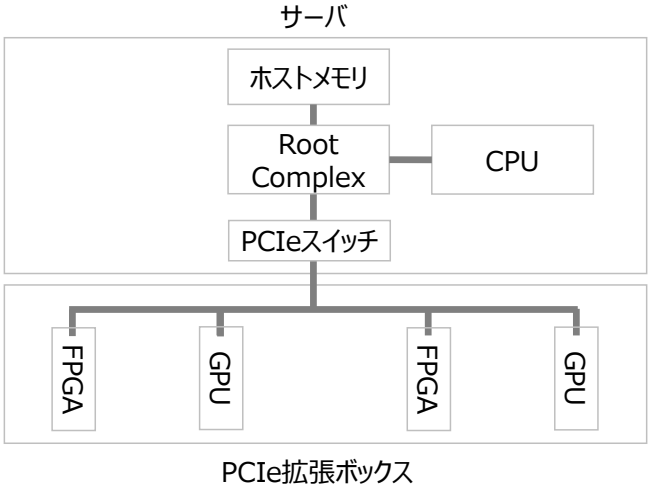


図 1　想定するシステム構成

例えば，監視カメラの映像に対して物体検出を行うユースケースを考える．FPGA は GPU に比べ開発難度が高く開発工数が大きくなりがちな反面，特定の処理を高い電力効率で実行することができる．そのため，映像のデコードやフィルタリング，サイズ変更等の定型的な処理を行うのに適している．一

方，GPU は FPGA ほど電力効率は高くないものの，短期間で新しい処理を実装することができる．そのため，物体検出アルゴリズム等，日進月歩で進化するアルゴリズムを迅速に実装することができる．したがって，監視カメラの映像データをまず FPGA が受信し，FPGA が前処理したデータを GPU に転送し，GPU が推論処理を実行する，というのが典型的な処理の流れとなる．

アクセラレータは一般に DMA エンジンを備え，DMA によるホストメモリとの間のデータ転送機能をサポートする．したがって，アクセラレータ間でデータ転送を行う最も汎用的な方法は，送信アクセラレータ (e.g. FPGA) が DMA によりホストメモリへ書き込んだデータを，受信アクセラレータ (e.g. GPU) が DMA により読み出すこととなる．2 つのアクセラレータを協調動作させるための制御は CPU が行う．この方式を本論文では「ホストメモリ経由転送」と呼ぶ．

しかし，ホストメモリ経由転送には以下の問題点がある．

データ転送の競合による性能低下　PCIe 拡張ボックス等を介してサーバに多数のアクセラレータが接続されているシステム構成では，アクセラレータ間のデータ転送が全てホストメモリ-PCIe スイッチ間の PCIe リンクに集中する．このリンクの PCIe 帯域幅は一般に x16 程度なので，容易にボトルネックとなり得る．例えば，それぞれ x16 の PCIe 帯域幅を持つ FPGA と GPU が 2 ペア同時にデータ転送を行うとすると，各ペアが利用可能な最大データ転送性能は，各ペアが単独でデータ転送を行う場合の転送性能の 1/2 となる．

消費電力の増大　転送制御を CPU が行うことにより，また，転送データを一時的にホストメモリに格納することにより，サーバの消費電力が増大する．

3　提 案 手 法

前章で述べた問題を解決するため，本章では FPGA と GPU 間で直接データ転送を行う手法を提案する．提案手法では，我々が開発した DMA エンジンを搭載した Xilinx 製 FPGA と，NVIDIA 製 GPU の間で直接データ転送を行う．

3.1　FPGA

我々は，DMA エンジンおよび DMA エンジンを制御するためのソフトウェア（ドライバ&ライブラリ）の開発を行った．図 2 に，CPU や GPU 等の対向デバイスによる DMA エンジンの制御方法を示す．

この DMA エンジンは [8] の開発物をベースにしており，以下の特徴をもつ．

- FPGA への入力および FPGA からの出力の双方向の DMA 転送ができる．以降，FPGA から見た場合の転送方向（受信／送信）に従い，前者を RX 転送，後者を TX 転送と呼ぶ．
- 対向デバイスと，RX/TX 毎に最大 16 チャンネルでのデータ転送ができる．ここで，チャンネルは論理的な通信経路を意味する．

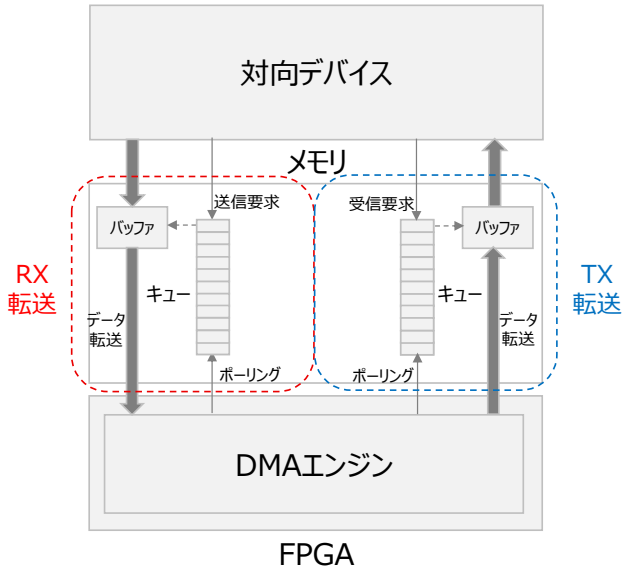


図 2　対向デバイスによる DMA エンジンの制御方法

- データ転送の制御はチャンネル毎に存在するキューを操作することで行う．RX 要求の発行は，送信データを格納したバッファ（RX 用バッファ）のアドレスやサイズを含む DMA ディスクリプタを，対向デバイスが RX 用キューに投入することで行われる．同様に TX 要求の発行は，受信データを格納するバッファ（TX 用バッファ）のアドレスやサイズを含む DMA ディスクリプタを，対向デバイスが TX 用キューに投入することで行われる．

この DMA エンジンを用いた，ホストメモリ経由転送方式によるデータ転送の流れは以下の通りである．

（1）　CPU が FPGA に対し TX 用キューを使用開始するための初期設定を行う．CPU はホストメモリ上に TX 用キューを確保し，FPGA に対し TX 用キューのアドレスとサイズを通知する．また，CPU は TX 用バッファをホストメモリ上に確保し，バッファのアドレスとサイズを含む DMA ディスクリプタを作成する．TX 用バッファは循環バッファであるため，初期設定時に一括して確保される．

（2）　FPGA は使用中の TX 用キューをポーリングにより監視する．

（3）　CPU は受信に使用する TX 用バッファを選択し，TX 用キューに当該 TX 用バッファに対応する DMA ディスクリプタを書き込む．CPU は DMA ディスクリプタの書き込み後，当該 DMA ディスクリプタの状態フィールドをポーリングにより監視する．

（4）　FPGA はデータを DMA により TX 用バッファへ転送後，転送完了を意味する値を当該 DMA ディスクリプタの状態フィールドに書き込む．

（5）　CPU はデータ受信完了を検知した後，TX 用バッファから GPU メモリへデータ転送を行う．データ転送には cudaMemcpy 等 GPU ベンダーが提供するライブラリ API を用いる．

今回開発した DMA エンジンの特徴は，CPU が直接 DMA

エンジンをプログラムするのではなく，FPGA がホストメモリ上のキューをポーリングし DMA ディスクリプタを取得することにより，自律的に DMA を開始可能な点にある．

キューのアドレスは PCIe バスアドレスにより指定されるため，FPGA はキューがホストメモリにあるのか GPU メモリにあるのかを意識しない．そのため，GPU メモリ上のキューの PCIe バスアドレスを指定することにより，GPU から FPGA に対して DMA 開始指示を発行することが可能になる．

同様に，バッファのアドレスも PCIe バスアドレスにより指定されるため，GPU メモリ上のバッファの PCIe バスアドレスを指定することにより，FPGA から GPU への直接データ転送が可能になる．

3.2 GPU

近年の NVIDIA 製 GPU のうち，データセンター用途のハイエンドモデルは GPUDirect と呼ばれる機能を持つ [2]．これは，GPU の内部メモリを PCIe バスアドレス空間にマッピングし，FPGA 等サードパーティ製 PCIe デバイスからアクセス可能にする機能である．

提案手法では，GPU が内部メモリに確保したキューやバッファに対して，FPGA が GPUDirect によるアクセスを行う．

3.3 FPGA-GPU 間データ転送

TX 用キューおよび TX 用バッファを確保するメモリを変更することにより，ホストメモリ経由方式および次に述べる 2 種類の直接データ転送方式を実現することができる．

ホストメモリ経由転送　図 3 に示す通り，TX 用キューと TX 用バッファの両方をホストメモリに確保する．この方式は，提案手法の評価の際にベースラインとして使用する．

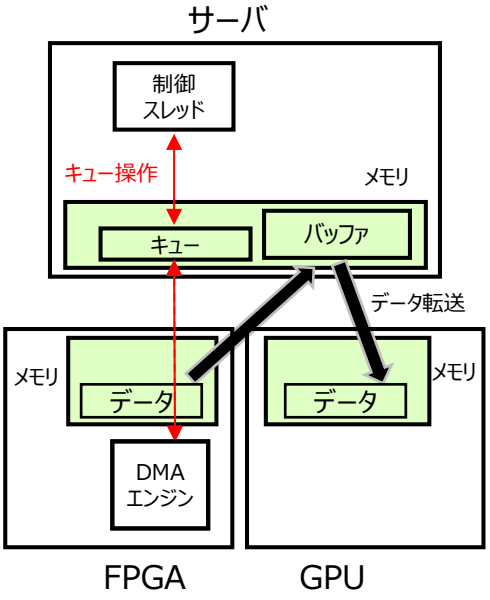


図 3　ホストメモリ経由転送

この方式では，TX 用バッファをホストメモリに確保するため，CPU が FPGA からデータ受信後，GPU への転送を行う．また TX 用キューもホストメモリに確保し，DMA 転送の制御も CPU 自身が行う．

CPU 制御直接転送　図 4 に示す通り，TX 用キューをホストメモリに，TX 用バッファを GPU メモリに確保する．

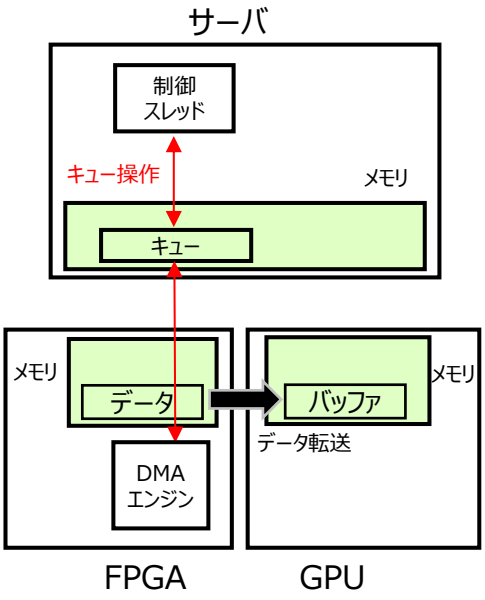


図 4　CPU 制御直接転送

この方式では，TX 用バッファを GPU メモリに確保するため，データは FPGA から GPU へ直接転送される．ただし，TX 用キューはホストメモリに確保し，DMA 転送の制御は CPU が行う．

GPU 制御直接転送　図 5 に示す通り，TX 用キューと TX 用バッファの両方を GPU メモリに確保する．

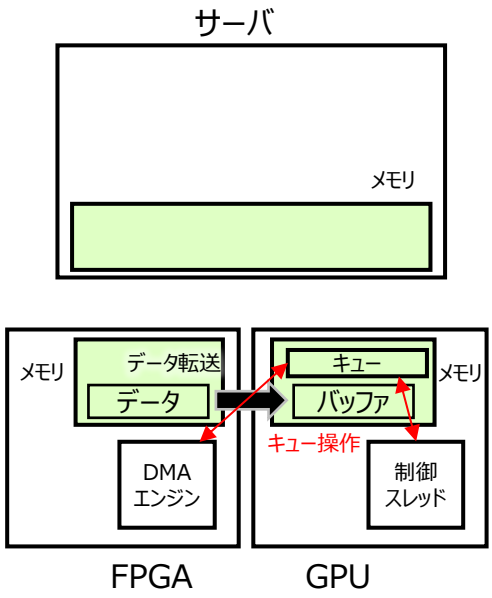


図 5　GPU 制御直接転送

この方式では，TX 用バッファを GPU メモリに確保するた

め，データは FPGA から GPU へ直接転送される．また，TX 用キューも GPU メモリに確保し，CPU は TX 用キューの初期設定のみを行い，以降 DMA 転送の制御は GPU 自身が行う．

4 評　　価

提案手法の消費電力性能効率向上効果を検証するため，以下の評価を行った．

4.1 評価環境

ハードウェア　サーバに PCIe 拡張ボックスを接続しサーバ本体に FPGA と GPU を 2 基ずつ，PCIe 拡張ボックスにも FPGA と GPU を 2 基ずつ搭載した．ハードウェアの機種一覧を表 1 に示す

表 1　ハードウェア一覧

機器	製品名	ベンダー
サーバ	SYS-420GP-TNR	SuperMicro
PCIe 拡張ボックス	Falcon 4210	H3
FPGA	U250	Xilinx
GPU	A100	NVIDIA

サーバの PCIe バスは PCIe Gen4，FPGA の PCIe リンクは PCIe Gen3 x16，GPU の PCIe リンクは PCIe Gen4 x16 であるため，FPGA と GPU の間の転送帯域幅の理論値は FPGA の PCIe リンクに律速され 15.75 [GB/s] となる．

FPGA には 3 章で述べた DMA エンジンを搭載した．この FPGA は，RX 転送で受信した画像データを DRAM に書き込んだ後，読み出して前処理することなく TX 転送で送信する．

サーバは CPU を 2 基搭載しているが，ホストメモリ⇔ PCIe スイッチ間がボトルネックになる状況を作り出すことを意図し，FPGA，GPU，PCIe 拡張ボックスは，図 6 の通り全て片側の CPU に寄せて接続した．

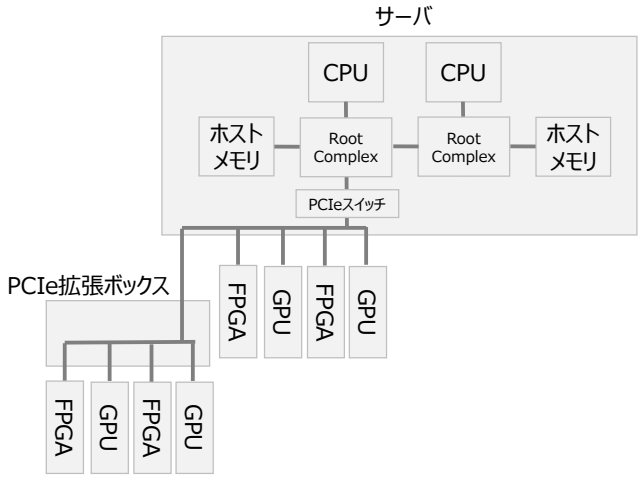


図 6　評価環境のシステム構成

ソフトウェア　今回開発した測定ツールは，実際のユースケースに近い動作をさせることを意図し，FPGA-GPU 間のデータ転送および転送データを入力とした GPU 推論を実行する．転送処理と推論処理は非同期に実行される．

推論に使用した推論フレームワークは ONNX Runtime [3]，推論モデルは画像分類に使用される ResNet-18 を用いた．入力データは 224 × 224 の浮動小数点型 BGR 画像であり，したがって 1 回の DMA 転送サイズは 602[KB] となる．

FPGA のドライバとライブラリは新規開発したものを使用，GPU のドライバとライブラリは NVIDIA 製のものを使用した．

ソフトウェアのバージョン一覧を表 2 に示す．

表 2　ソフトウェア一覧

ソフトウェア	バージョン
OS (Ubuntu)	20.04
CUDA	12.2
NVIDIA ドライバ	535.129.03
ONNX Runtime	1.13.1

4.2 評価手法

4 組の FPGA-GPU ペア間で同時にデータ転送を行う．データ転送は，サーバ本体に搭載されている FPGA-GPU ペア，PCIe 拡張ボックスに搭載されている FPGA-GPU ペアの間で行う．

FPGA への画像データの入力（RX 転送）は，常に CPU 制御直接転送により行う．すなわち，対向 GPU のメモリ上に格納した画像データを RX 転送で FPGA に入力する．4.1 で述べた通り，FPGA は受信したデータを処理することなく TX 転送で再度 GPU へ出力し，GPU は受信したデータに対して推論処理を行う．

3.1 で述べた通り，RX/TX 転送はそれぞれ最大 16 チャンネルで実行可能である．測定には全 16 チャンネルを使用した．使用する画像は全チャンネルで同一とし，各チャンネルは同一画像を繰り返し FPGA へ送信するものとする．以降，1 回に送信する画像データを映像の 1 フレームと見立てて「フレーム」と呼ぶ．画像データの入力をホストメモリではなく GPU メモリから行うのは，ホストメモリ⇒ FPGA への経路がボトルネックになるのを防ぐためである．

FPGA から GPU への画像データの出力（TX 転送）は，ホストメモリ経由転送，CPU 制御直接転送，GPU 制御直接転送のいずれかにより行う．

各チャンネルのフレーム転送はチャンネル毎に独立に行われる．また，あるフレームの RX 転送は，1 つ前のフレームの TX 転送の完了を待たずに開始される．ただし，RX 用バッファおよび TX 用バッファは循環バッファであるため，RX 転送開始済みかつ TX 転送未完了なフレーム数が，両バッファのバッファ長である 32 を超えないように RX 転送の開始タイミングは調整される．

いずれの転送方式においても，CPU は GPU が受信したフレーム数を TX チャンネル毎に監視している．そして，全 TX チャンネルが同一フレームの受信を完了したタイミングで，GPU に推論実行を指示する．推論は，16 チャンネル分のフレームを一度に処理するバッチ推論により行う．

測定ツールは 180 秒間実行した．

最初のフレームの RX 転送要求が出されてから，最後のフレームの推論が完了するまでの時間を計測し，転送データ量からスループットを計測した．

また，転送実行中にサーバ本体と PCIe 拡張ボックスの消費電力を測定した．サーバ本体の消費電力の計測には IPMI Tool，PCIe 拡張ボックスの消費電力の計測にはベンダーが提供する Redfish 経由の計測ツールを使用した．

4.3 結　　果

1 ペア当たりのスループットの測定結果を図 7 に示す．

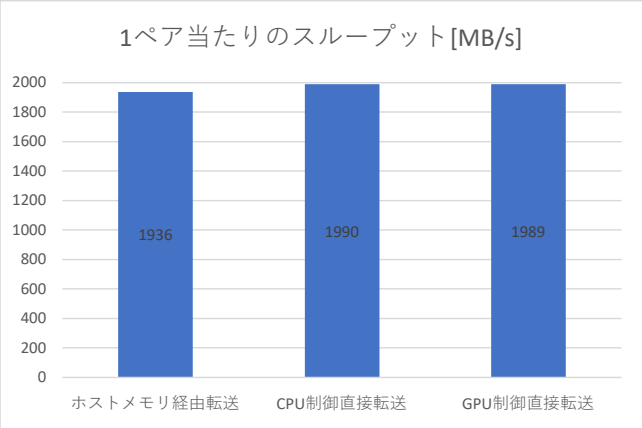


図 7　スループット

全ての転送方式でほぼ同等の性能が出ている．このことと，先述の通り転送処理と推論処理が非同期に実行されることより，推論処理ではなく転送処理がボトルネックになっていることがわかる．

また，いずれの転送方式においても，FPGA と GPU の間の転送帯域幅の理論値の 12%程度しか出ていない．以下の理由により，これは FPGA の内部処理（DRAM への書き込み・読み出し）がボトルネックになっているものと判断する．

- 別途行った FPGA-GPU 1 ペアだけを使った測定でも，4 ペアの場合と比べて 1 ペア当たりの性能が向上しなかった．
- 実測値が，FPGA のカタログスペックの DRAM 性能から見積もった性能予測値とほぼ一致する．

したがって，ベースラインとして使用したホストメモリ経由転送の場合でも，ホストメモリ-PCIe スイッチ間の PCIe リンクがボトルネックにはならなかったことになる．

次に消費電力の測定結果を図 8 に示す．

ホストメモリ経由転送，CPU 制御直接転送，GPU 制御直接転送の順に GPU の消費電力が増大している．また，同じ順で GPU 以外の消費電力が減少している．これは，ホストメモリ経由転送より直接転送の方がホストメモリの消費電力が小さく，CPU 制御直接転送より GPU 制御直接転送の方が CPU の消費電力が小さいためと思われる．今回の測定結果では，消費電力の合計は GPU 制御直接転送が最小となった．

以上の性能および消費電力の測定結果により，ホストメモリ経由転送に比べ，CPU 制御直接転送では 2%，GPU 制御直接転送では 12%の消費電力性能効率の向上が見られた．

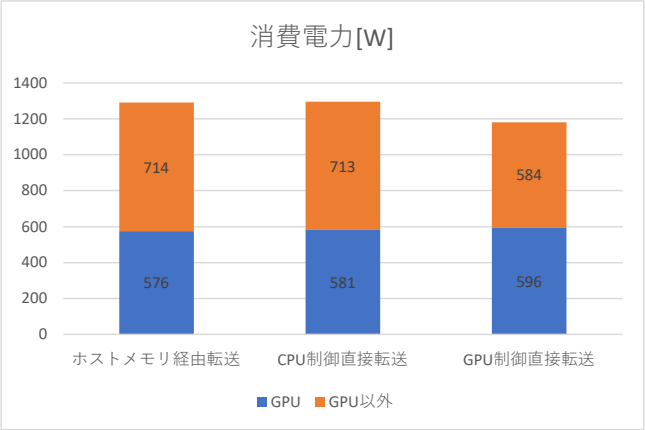


図 8　消 費 電 力

5　関 連 研 究

複数の先行研究により，FPGA-GPU 間で直接データ転送を行う手法が提案されている．

[5] では，本論文と同様に Xilinx 製 FPGA と NVIDIA 製 GPU の間で直接データ転送を行う手法が提案されている．本論文とは異なり，GPUDirect を使用せず，FPGA がバスマスタとなり DMA を行うための工夫が述べられている．

[7] では，本論文と同様に FPGA がバスマスタとなり NVIDIA 製 GPU へ直接データ転送を行う手法が提案されている．ただし，CPU が DMA 開始の指示を行う点が本論文とは異なる．

また，先行研究がもっぱら直接データ転送の性能向上効果に着目した評価を行っているのに対し，本論文では消費電力削減効果についても評価を行った．

6　結論と今後の課題

一般にアクセラレータ間のデータ転送はホストメモリ経由で行われるため，サーバに多数のアクセラレータを接続したシステム構成においては，ホスト-PCIe スイッチ間の PCIe リンクがボトルネックになるという問題がある．

本論文では，ボトルネックリンク回避のため，パイプラインを実行するアクセラレータ間で直接データ転送を行う転送方式を提案した．FPGA-GPU 間で直接データ転送を行うプロトタイプを用いた評価を行い，ホストメモリ経由転送方式に比べ消費電力性能効率が向上することを示した．

本研究の今後の課題として，以下の項目を検討中である．

1 章で述べた通り，アクセラレータ間直接データ転送が真価を発揮するのは，ホスト-PCIe スイッチ間の PCIe リンクがボトルネックになる状況である．今回の評価環境においては，PCIe バスの前に FPGA 自体の性能がボトルネックになったため，この点を十分に検証することができなかった．今後は，FPGA の性能向上によりホスト-PCIe スイッチ間の PCIe リンクがボトルネックになる状況を作り出した上で提案手法の効果を検証したい．

また，現状の提案手法をアプリ開発者が使用するためには，FPGA ライブラリや GPU ライブラリ等，低レベルの API を意識したコードを書く必要がある．アプリ開発者の労力を省くために，転送方式の細部を抽象化し，多種のアクセラレータ間のデータ転送を統一的な API で利用可能にするための上位ライブラリの開発を検討中である．

最後に，データ転送処理だけでなく，推論処理からも CPU の介在をなくす方式について検討中である．現状，アクセラレータ間データ転送は CPU の介在なく実行可能であるものの，その後の GPU による推論処理は既存の推論フレームワークをそのまま使っている．したがって，GPU 自身が推論を実行開始することはできず，一度 CPU に制御を戻した後に CPU が GPU に対して推論を実行開始するよう指示する必要がある．また，GPU がカーネルと呼ばれる推論の処理単位を実行完了する度に，CPU に制御を戻してから CPU が次のカーネルの実行を GPU に指示する必要がある．推論処理を GPU 単独で実行可能にすることで，これらの CPU 関与によるオーバヘッドを削減し，ストリーム処理全体のさらなる高速化と省電力化が可能と考えている．

文　　献

[1] Beyond 5G/6G ホワイトペーパー（日本語 3.0 版）. https://beyond5g.nict.go.jp/images/download/NICT_B5G6G_WhitePaperJP_v3_0.pdf.

[2] NVIDIA GPUDirect. https://developer.nvidia.com/gpudirect.

[3] ONNX Runtime. https://onnxruntime.ai.

[4] 【Falcon 4210 with RTX 3090】GPU Expansion Chassis | H3 Platform. https://www.h3platform.com/product-detail/overview/24.

[5] Ray Bittner and Erik Ruf. Direct GPU/FPGA communication via PCI express. In *2012 41st International Conference on Parallel Processing Workshops*, pages 135–139. IEEE, 2012.

[6] Mohamed Zahran. Heterogeneous computing: Here to stay. *Communications of the ACM*, 60(3):42–45, 2017.

[7] 小林諒平, 阿部昂之, 藤田典久, 山口佳樹, 朴泰祐, et al. GPU-FPGA 複合システムにおけるデバイス間連携機構. **研究報告ハイパフォーマンスコンピューティング** *(HPC)*, 2018(26):1–8, 2018.

[8] 横野智也, 山部芳朗, 田仲顕至, 有川勇輝, 石崎晃朗, et al. 自律駆動 DMA エンジンを搭載した FPGA 演算システム. **研究報告システムと** *LSI* **の設計技術** *(SLDM)*, 2022(10):1–6, 2022.

GPUを用いたk-匿名化加工処理の高速化の検討

吉本　廣雅†　　満武　巨裕†　　合田　和生††

† 医療経済研究機構　〒105–0001 東京都港区虎ノ門 1–21–19 東急虎ノ門ビル 3F
†† 東京大学 生産技術研究所　〒153-–8505 東京都目黒区駒場 4-–6-–1
E-mail: †{yoshimoto,mitsutake}@ihep.jp, ††kgoda@tkl.iis.u-tokyo.ac.jp

あらまし　個人情報を含む大規模データの活用には，プライバシーを保護する匿名化加工処理が必須である．例えばk-匿名化を用いると，同じ属性を持つデータがk件以上存在するようにデータを変換することで，個人が特定される確率をk分の1以下に低減できる．本稿は自治体が保有する住民のヘルスケアデータを繰り返し加工・分析する場面を想定し，GPUを活用した匿名化加工処理の高速化手法について検討を行う．実験では合成データを用いた性能評価を行い，提案手法の実用性について考察をまとめる．

キーワード　医療・ヘルスケア，k-匿名化，GPGPU，sorting network

1　はじめに

個人情報を含む大規模データの活用にはプライバシー保護技術が必要不可欠である．世界各国でプライバシー保護に関する法整備が進んでいるように，日本でも令和元年の健康保険法等改正，令和3年の「デジタル社会の形成を図るための関係法律の整備に関する法律」など，国や自治体が保有する超大規模なデータを研究者や事業者が活用し社会実装できる体制が整いつつある．具体例をあげると，厚生労働省は全国民のレセプト情報・特定健診等情報データベース (NDB) を大学の研究者等に提供する仕組みを整備しており，公益目的でのヘルスケアデータの分析や利活用を促進している [1]．地方自治体でも同様な取り組みがなされており，住民のヘルスケアデータを活用した様々な研究成果が発表されている [2–4] .

本稿はこれら体制を前提に，国や自治体が保有する規模の超大規模データを対象に，プライバシーを保護しつつ高速に処理する技術について議論する．具体的にはプライバシー保護技術の一つであるk-匿名化処理を取り上げ，最新のGPGPUアーキテクチャを活用し高速化する手法を検討する．k-匿名化処理は，例えば，自治体が保有する機微なデータベースに対する問い合わせ結果を利用者が閲覧する際に必ず適用されるべき処理である．処理対象となるデータは百万人単位の住民に関する数千万から数億レーコードの規模の量になる．この規模のデータ処理を円滑に活用するには，その匿名化処理の高速化が必要不可欠である．

2　k-匿名化とそのアルゴリズム

k-匿名化とは同じ属性を持つデータが k 件以上存在するようにデータを一部変換する処理である．この変換によりデータから個人が特定される確率が k 分の1以下に低減する [5–7].

変換処理には様々方法がある．一つは一般化 (re-coding) と呼ばれる方法で，数値を数値の範囲に，カテゴリをより大きなカテゴリに，それぞれ値を置換する．またミクロアグリゲーションと呼ばれる方法では値を平均値に置換する．具体例としては，年齢を10代や20代といった年齢階級に置換する処理が一般化であり，年齢を平均年齢に置換する処理がミクロアグリゲーションである．

k-匿名化の処理アルゴリズムは多数提案されている [8–11] . これらは大きく，トップダウン型，ボトムアップ型，両者を組み合わせたハイブリッド型の三つに大別できる．トップダウン型にはMondrian [8] などがあり，その処理は入力データをクラスタ群へと分割し，各クラスタに含まれる要素が k 件になるまで分割を繰り返す．そして各クラスタに属するレコード群に対して一般化またはミクロアグリゲーションを行う．ボトムアップ型にはMaximum Distance to Average Vector(MDAV) [9], V-MDAV [10] などがあり，その処理はあるデータを基点として選択しそのデータを含む k 件のクラスタを生成し，クラスタに属さないレコードが無くなるまでクラスタの生成を繰り返す．またトップダウン型・ボトムアップ型を組み合わせたハイブリッド型のアルゴリズムとして [11] が提案されている．

最適な k-匿名化はNP困難な問題であり [12]，既存アルゴリズムの優劣は単純には議論できない．本稿は前述のMDAVアルゴリズムを対象とし，以下，その処理の高速化について議論を行う．

2.1　MDAVアルゴリズムの概要

MDAVアルゴリズム [9] は，以下の7 Stepで構成される．

- Step A.　全てのレコードについて，その平均値 $\hat{x}_n$ を計算する．
- Step B.　レコード間のユークリッド距離を用いて，$\hat{x}_n$ から最も遠いレコード r_n を探す．
- Step C.　r_n から最も遠いレコード s_n を探す．また，r_n とその近傍にある $k-1$ 個のレコード群を抽出する．
- Step D.　s_n の近傍にある $k-1$ 個のレコード群を抽出する．
- Step E.　未処理レコードの個数が $2k$ 以上ある場合は，Step

A に戻り処理を繰り返す．

- Step F．未処理レコードの個数が k 以上ある場合は，それらを新しいレコード群とする．
- Step G．未処理レコードの個数が k 未満の場合は，各レコードを Step D または Step F で生成したレコード群の中から最も近いレコード群に加える．

MDAV アルゴリズムの性質を議論するために，StepA から StepE までの繰り返し処理の n 回目から $n+1$ 回目を可視化した図を図 1 に示す．

$StepA_n$ は平均値 $\hat{x}_n$ の計算処理である．$StepB_n$ は $\hat{x}_n$ からのユークリッド距離に応じてレコードをソートする処理である．同様に $StepC_n$，$StepD_n$ も r_n，s_n からのユークリッド距離に応じたソートの処理となる．このように，MDAV アルゴリズムは，平均値の計算と 3 回のソートを繰り返しながら，N レコードから $2k$ 個のレコードを順に取り出す処理となっており，$StepA$ から $StepE$ までを $\frac{N}{2k}$ 回繰り返す処理となっている．

2.2 高速化に関する先行研究

MDAV アルゴリズムは入力のサイズが増加すると計算量が指数的に増加してしまう．そのため高速化の手法が多数提案されている．トップダウン型を高速化した研究としては，分散処理を用いた実装 [13, 14]，ボトムアップ型を高速化した研究としては，アルゴリズムの改良 [15]，GPU による並列処理を用いた実装 [16] がある．

並列処理を用いると，平均値の計算処理は総和の計算にリダクション演算が，ソート処理は並列アルゴリズムで高速化できる．ソートの並列アルゴリズムとしては Merge sort, Radix sort [17] や，sorting network として Batchers bitonic and odd-even merge sort (Bitonic sort) [18, 19] がある．潤沢な並列計算機資源があれば，要素数 N のデータ列を Radix sort は $\mathcal{O}(N)$ で，Bitonic sort は $\mathcal{O}(N\log^2(N))$ でソートできる．

2.3 予 備 実 験

予備実験として MDAV アルゴリズムの実装を複数用意し，それぞれのベンチマークを行った．

実験は，先行研究 [20–22] と条件を揃え，the 1994 US census database (ADULT) 相当のランダムデータ（8 次元のベクトル）を 1 レコードとし，レコード数を増加させながら k-匿名化に要する処理時間を測定した．実行環境は CPU が Intel Core i9-13900KF（Performance-core 数：8，動作クロック 5.4GHz），GPU が NVIDIA GeForce RTX 3090（CUDA core 数：10496，動作クロック 2.1GHz）である．

実装は 3 種類用意した．CPU・シングルスレッド版 (C++と STL を使用)，CPU・マルチスレッド版（C++と Thrust ライブラリ，バックエンドは Intel Threading Building Blocks を使用)，GPU 版（C++と Thrust ライブラリ，バックエンドは CUDA を使用）である．並列版の実装では，平均値の計算は加算のリダクション演算，ソート処理は Merge sort を用いた．

ベンチマーク実験の結果を図 2 に示す．グラフが示すように

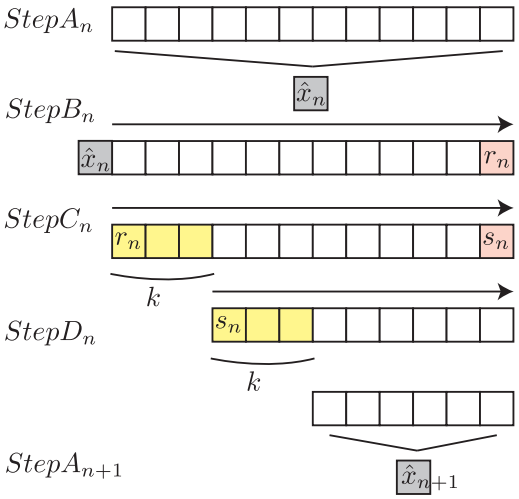


図 1: MDAV アルゴリズムの概要

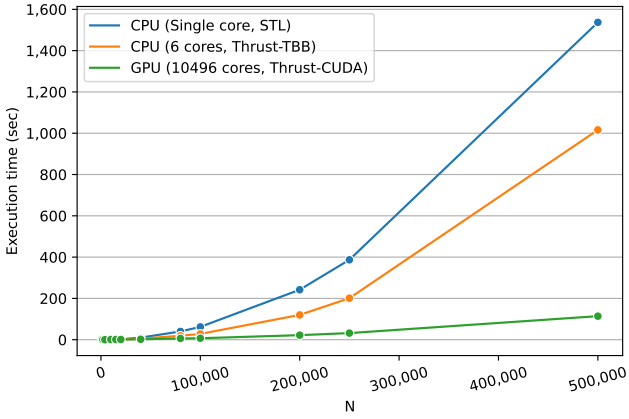


図 2: 予備実験の結果．既存ライブラリを用いた MDAV アルゴリズムの実装とその処理速度

GPU を使った実装を行うことで大幅な高速化が達成できていることがわかる．しかし，前述のように本研究が想定する自治体が保有する大規模データは数百万人単位の住民から得られた数千万から数億レコードの規模であり，この規模のデータを数秒で匿名化加工するには更なる高速化が必須である．

本研究はこの結果をベースラインとし，最新の GPU アーキテクチャに特化した高速化手法について検討する．

3 高速化の検討

本稿は高速化に向けて 4 つの項目を検討する．

3.1 検討項目 1：Double buffering

図 1 で示すように MDAV アルゴリズムは，レコードのソートを繰り返し行う処理が中心となっており，メモリアクセスが多い処理となっている．そこで Double buffering と呼ばれる手法を用いメモリアクセスの効率化を図る．具体的には，GPU 上に二つのバッファ A，B を用意し，偶数 Step の処理では A

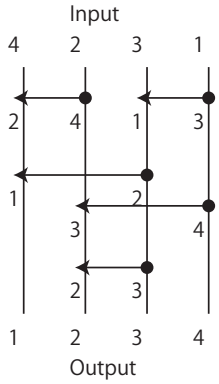


図 3: Sorting network を用いた並列ソート．矢印は比較器を表す．比較器は二つの入力に対して小さい値を左側，大きい値を右側に出力する．

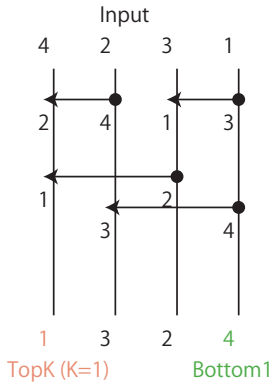


図 4: Sorting network を用いた topK と bottom1 の並列計算．図 3 よりも一つ少ない比較器で topK と bottom1 が出力できる

を入力用，B を出力用として用い，奇数 Step の処理ではバッファを swap して利用する．これによりメモリアクセス回数を削減しつつ，メモリキャッシュの効果的な制御を実現する．

3.2 検討項目２：平均値計算の高速化

図 1 の $\hat{x}_n$ と $\hat{x}_{n+1}$ を見比べると，$\hat{x}_{n+1}$ は $\hat{x}_n$ から $StepC_n$ と $stepD_n$ で除外した合計 $2k$ 個のレコードの和を減算するだけで計算できることがわかる．これにより平均値に関する処理の計算量を大幅に削減できる．この高速化法は [15] でも採用されている．

ただし単純に $\frac{N}{2k}$ 回もの減算処理を繰り返すと計算誤差が蓄積する問題が生じる．そこで，カハンの加算アルゴリズム [23] を用いて，誤差を補償しながら平均値を更新する方法を採用した．

3.3 検討項目３：topK 処理と bottom1 処理の高速化

図 1 の $StepB_n$ のソートでは末尾の 1 要素のみ，$StepC_n$ のソートでは先頭の k 要素と末尾の 1 要素，$StepD_n$ のソートでは先 1 頭の k 要素のみ，が処理に必要であり，その他の要素

Listing 1: WarpNet の実装

```
1  __global__ void WarpNet(float* input, float* output)
2  {
3      int idx = blockDim.x * blockIdx.x + threadIdx.x;
4      int lane = threadIdx.x % 32;
5      float value = input[tid];
6      __syncwrap();
7      for (int k = 2; k <= 32; k = 2 * k) {
8          for (int j = k >> 1; j > 0; j = j >> 1) {
9              bool dir = ((lane & j) != 0) ^ ((lane & k) != 0);
10             float ref = __shfl_xor_sync(-1, value, j, 32);
11             if (dir) {
12                 value = min(value, ref);
13             } else {
14                 value = max(value, ref);
15             }
16         }
17     }
18     output[tid] = value;
19 }
```

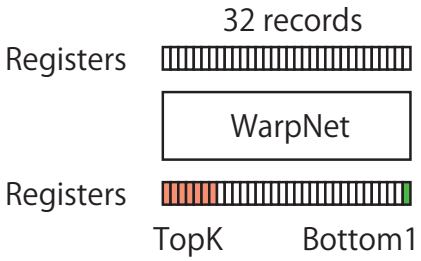


図 5: WarpNet の模式図．Listing 1 の 7 行目から 17 行目に対応する．WarpNet は 32 スレッドで動作する．各スレッドはレジスタ上にレコードを保持しており，WarpNet はそれら 32 レコードを入力とし，出力としてソートされたレコードをレジスタへ書き戻す．

のソートは不要である．先頭の k 要素を取り出す処理は topK 処理と呼ばれ，GPU を用いたアルゴリズムが研究されている [24, 25] ．また，末尾の 1 要素を取り出す処理は K=1 の topK 処理として実行できる．本稿ではこれを bottom1 処理と表記する．

本稿は，MDAV アルゴリズムにおける $StepB_n$ を bottom1 処理，$StepC_n$ を topK 処理と bottom1 処理，$StepD_n$ を topK 処理として，それぞれを GPU を活用したアルゴリズムにより高速化することを目指す．

3.4 検討項目４：GPGPU 専用命令の活用

GPGPU には並列処理に特化した専用命令が多数用意されている．専用命令を使うことで GPGPU の演算器・メモリユニットが最大限活用できると期待できる．そこで本節では GPGPU 専用命令の warp shuffle 命令に注目する．warp shuffle 命令は，スレッド間でメモリを介さず直接レジスタの値を交換する命令である．warp shuffle 命令はメモリアクセスやスレッド間の同期命令が不要なため，各スレッドをストールさせることなく処理が実行できるようになる．

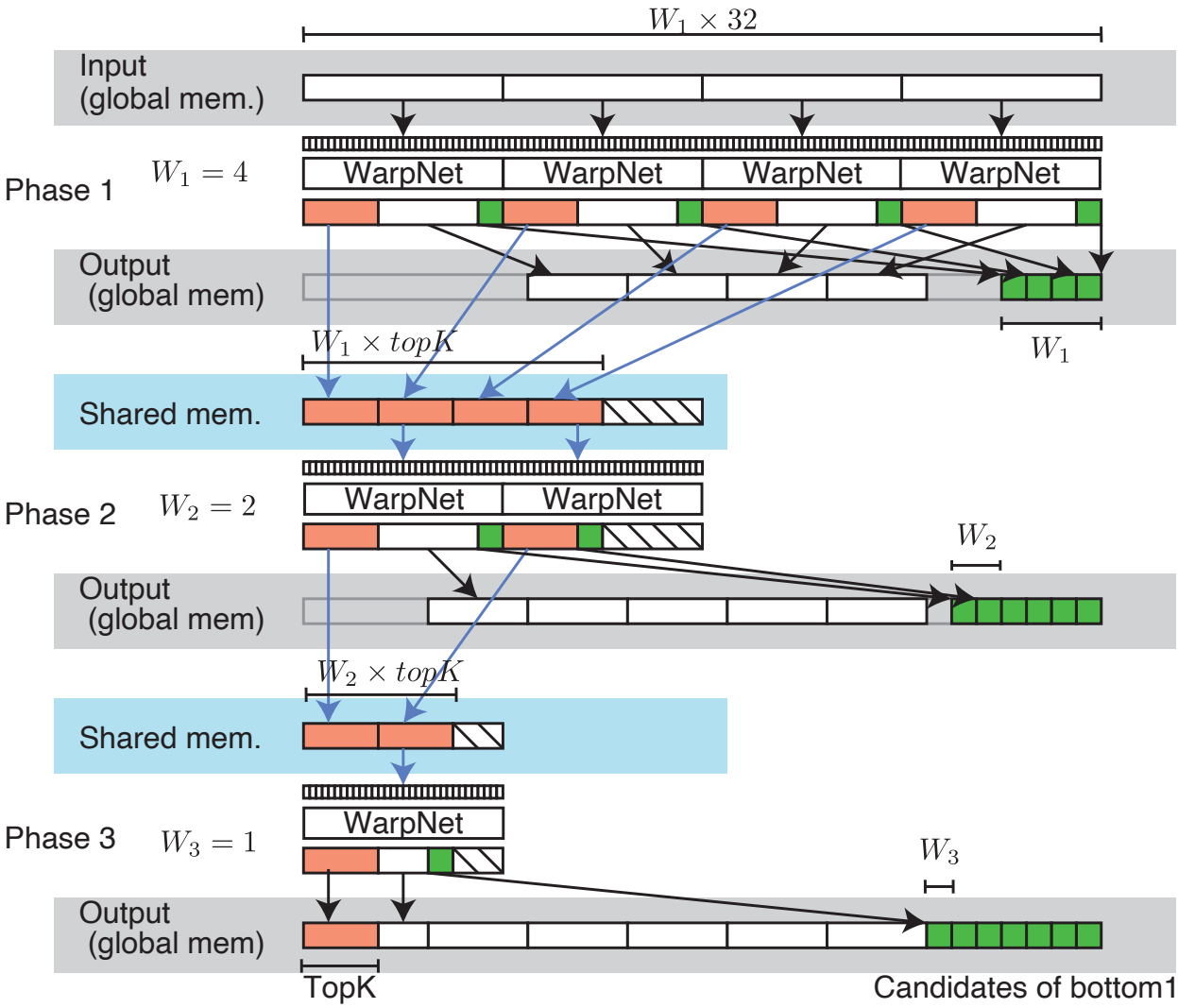


図 6: WarpNet を用いた topK and bottom1 アルゴリズムの概要

4 GPGPU 専用命令を活用した topK 処理と bottom1 処理の並列処理アルゴリズム

前節での検討項目に従い，本研究は warp shuffle 命令を用いて topK と bottom1 を求める並列アルゴリズムを考案した．以下詳細を述べる．

本アルゴリズムは sorting network（bitonic sort）をベースにしている．sort network の例を図 3 に示す．Sorting network とは 2 入力の比較器を複数組み合わせてソートを行う回路 (ネットワーク) である．図 3 では矢印が比較器を表しており，全体で 5 個の比較器により 4 変数がソートされる．Sorting network は各段の比較器が独立して稼働できるため，並列計算機で効率よく実装できることが知られている．

ここで k=1 の場合の topK 処理と bottom1 処理を考えると，図 4 に示すように一番下の比較器を省略した，全体で 4 個の比較器のネットワークで topK と bottom1 が出力できる．このように最適な sorting network を設計できれば，topK と bottom1 の計算量を削減できることになる．

今回使用する GPU（Nvidia RTX 3090）では warp shuffle 命令は最大で 32 スレッド間でレジスタを共有できる．そのため本稿では 32 入力の sorting network を基本単位としてその組み合わせで topK と bottom1 を同時に実行するネットワークを設計する．以下，この基本単位を WarpNet と表記する．

WarpNet の実装を 1 に示す．これは CUDA のカーネルの形で実装した WarpNet である．カーネルへの入力は input, 出力は output である．CUDA はこのカーネルをスレッド単位で並行起動し，3 行目の変数 idx にスレッドの番号が格納される．その後各スレッドはそれぞれ変数 $value$ に入力を格納する．9 行目から 15 行目が図 4 における比較器に対応する．7 行目から 10 行目のコードはバタフライネットワークの実装である．10 行目の__shfl_xor_sync は warp shuffle 命令に対応するマクロであり，warp shuffle 命令は変数 j と自身のスレッド番号（idx）との xor を計算し，その値のスレッドの変数 $value$ を変数 ref に格納する．

Warp shuffle 命令は GPU 専用命令であり，メモリバリア

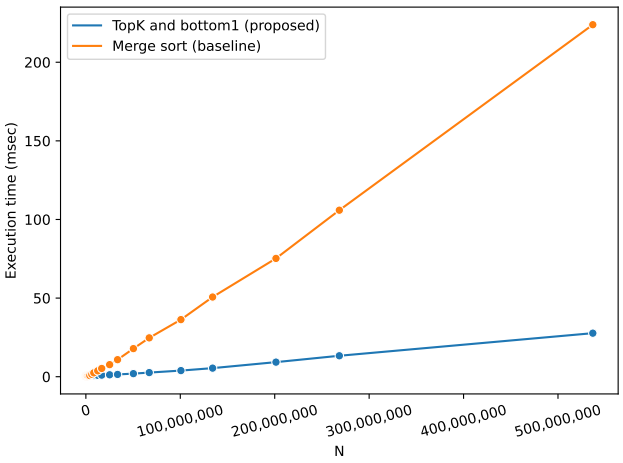


図 7: topK and bottom1 処理の高速化

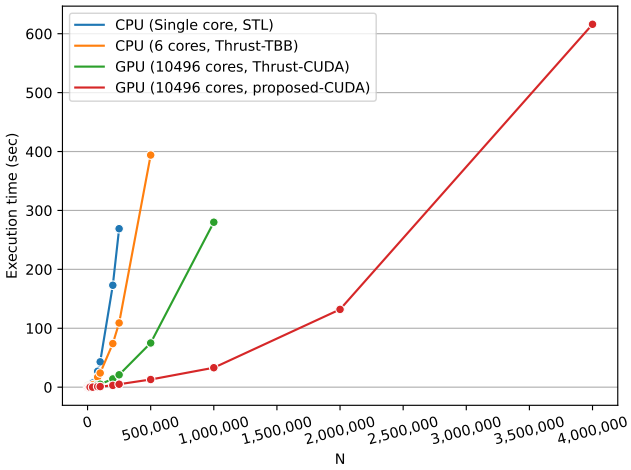


図 8: MDAV アルゴリズムの高速化

やスレッド間の同期命令の追加発行が一切不要である．Warp shuffle 命令を用いることで各スレッドはストールすることなく効率良く sorting network の処理を実行でき，GPU の演算機メモリユニットを最大限活用した動作が期待できる．

4.1 WarpNet を用いた topK と bottom1 の同時計算アルゴリズム

図 5 は前節で述べた WarpNet の模式図である．GPU は 32 スレッドを一つの単位（Warp）として並列稼働しており，WarpNet は 32 入力のネットワークを GPU の 32 スレッドで並行実行する．WarpNet の出力はソート結果であり，これを topK，bottom1，その他の三つに分けて活用する．これにより topK と bottom1 を同時に計算する並列アルゴリズムが設計できる．

アルゴリズムの概要を図 6 に示す．基本的アイデアは WarpNet を用いて，32 個の入力データを k 個の topK，1 個の bottom1，$(32-k-1)$ 個の残り，の 3 グループに振り分ける点にある．そして，topK と bottom1 の 2 グループについては再度 WarpNet で 3 グループに振り分ける処理を繰り返すことで，全体の topK と bottom1 を求める．

その際に残りとして除外された $(32-k-1)$ 個の残りはそのまま未ソートのレコード列として出力する．これにより通常のソート処理と比較して，除外されたレコードに対する演算を大幅に枝刈りすることが可能となる．

図は 128 レコードの入力を 3 phase で処理している．Phase1 は 4 つの WarpNet が稼働し，Phase2 では 2 つの WarpNet が，Phase3 では一つの WarpNet が稼働し，topK と bottom1 を出力している．これら WarpNet の個数，Phase の個数は topK の値から逆算して求めることができるが本稿では詳細は割愛する．

5 実験と考察

5.1 topK and bottom1 処理の高速化

まず topK 処理と bottom1 処理について合成データを用いたベンチマークにより性能評価を行い，各検討項目の効果を検討した．

合成データは，key（32bit 整数型）と value（32bit float 型）のペアからなるデータであり，ベンチマークは k=5 の場合で value の値で topK および bottom1 処理を複数回行い，処理時間の中央値を測定した．その他の条件は予備実験と同等である．

結果を図 7 に示す．比較対象となる baseline は Thrust ライブラリを用いて実装した sort 処理であり，アルゴリズムは merge sort で全データをソートした．

図が示すように，baseline，提案アルゴリズム，ともに計算時間は $\mathcal{O}(N)$ であり，スループットは baseline が 2.4M records/msec，提案法が 24.2M records/msec であった．

表 1 に高速化に対する各項目の寄与を示す．提案アルゴリズ

表 1: topK and bottom1 に対する高速化手法の寄与（レコード数：5.36 億）

アルゴリズム	Warp shuffle 命令	Loop unrolling	処理時間 (usec)
Merge sort (baseline)			228,064
TopK and bottom1			24,184
		✓	24,170
	✓		22,235
	✓	✓	22,222

表 2: MDAV アルゴリズムに対する高速化の寄与（レコード数：50 万）

	高速化手法	処理時間 (sec)
Merge sort (baseline)		51.725
	+ Double buffering	23.977
	+ 平均値の計算	20.754
	+ TopK and bottom1	11.159
	+ Warp shuffle 命令	11.005
	+ Loop unrolling	10.995

ムは大幅に高速化に寄与しており，また Warp shuffle 命令によりさらに高速化が達成されていることが確認できた．

5.2 MDAV アルゴリズムの高速化

予備実験と同じデータで，k=5 の場合の MDAV アルゴリズムの性能評価を行った．結果を図 8 に示す．

表 2 に高速化に対する各項目の寄与を示す．検討項目の中では double buffering と topk and bottom1 アルゴリズムが大きく高速化に寄与していることが確認された．

6 まとめと今後の課題

本稿は，k 匿名化処理を GPU を用いて高速化する手法を検討し，その実装と性能評価実験を行った．結果，提案法で大幅な高速化が達成できることが確認された．

今後の課題は，実データでの性能評価が挙げられる．また，提案アルゴリズムのハイパーパラメータである，WarpNet の個数，Phase の個数の計算アルゴリズムおよび最適化が必須である．今後これらの課題に取り組み提案法の実用化を目指す予定である．

謝　　辞

本研究の一部は，内閣府総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム (SIP)「統合型ヘルスケアシステムの構築」課題番号（JPJ012425）の補助を受けて行なった．

文　　献

[1] 満武 巨裕．ウェルネスのための ICT：2．日本のレセプト情報・特定健診等データベース（NDB）の有効活用．情報処理, 56(02):140–144, 1 2015.

[2] Jumpei Sato, Naohiro Mitsutake, Masaru Kitsuregawa, Tomoki Ishikawa, and Kazuo Goda. Predicting demand for long-term care using japanese healthcare insurance claims data. Environmental Health and Preventive Medicine, 27:42–42, 2022.

[3] N. Kanda, H. Hashimoto, T. Imai, et al. Indirect impact of the COVID-19 pandemic on the incidence of non-COVID-19 infectious diseases: A region-wide, patient-based database study in Japan. 214:20–24.

[4] Kaho Hirayama, Naoki Kanda, Hideki Hashimoto, et al. The five-year trends in antibiotic prescription by dentists and antibiotic prophylaxis for tooth extraction: A region-wide claims study in Japan. 29(10):965–970.

[5] Marc Solé, Victor Muntés-Mulero, and Jordi Nin. Efficient microaggregation techniques for large numerical data volumes. International Journal of Information Security, 11:253–267, 2012.

[6] Khaled El Emam and Fida Kamal Dankar. Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association, 15(5):627–637, 9 2008.

[7] Sushil K. Prasad, Michael McDermott, Xi He, and Satish Puri. Gpu-based parallel r-tree construction and querying. In *2015 IEEE International Parallel and Distributed Processing Symposium Workshop*, pages 618–627, 2015.

[8] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*, pages 25–25. IEEE, 2006.

[9] Josep Domingo-Ferrer, Antoni Martínez-Ballesté, Josep Maria Mateo-Sanz, and Francesc Sebé. Efficient multivariate data-oriented microaggregation. 15(4):355–369.

[10] Agusti Solanas, Antoni Martínez-Ballesté, and J Domingo-Ferrer. V-mdav: a multivariate microaggregation with variable group size. In *17th COMPSTAT Symposium of the IASC, Rome*, pages 917–925, 2006.

[11] 新井 淳也, 鬼塚 真, and 塩川 浩昭. クラスタリングと空間分割の併用による効率的な k-匿名化. 日本データベース学会和文論文誌, 13-J(1):72–77, 10 2014.

[12] Adam Meyerson and Ryan Williams. On the complexity of optimal K-anonymity. In *Proc. Twenty-Third ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, PODS '04, pages 223–228. Association for Computing Machinery, 2004.

[13] John Christoforidis and Minas Dasygenis. An efficient scalable parallelized version of the mondrian algorithm. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pages 150–153, 2016.

[14] Farough Ashkouti, Keyhan khamforoosh, and Amir Sheikhahmadi. Di-mondrian: Distributed improved mondrian for satisfaction of the l-diversity privacy model using apache spark. Information Sciences, 546:1–24, 2021.

[15] Ana Rodríguez-Hoyos, José Estrada-Jiménez, David Rebollo-Monedero, et al. The Fast Maximum Distance to Average Vector (F-MDAV): An algorithm for k-anonymous microaggregation in big data. 90:103531.

[16] Roberto Di Pietro, Leonardo Jero, Flavio Lombardi, and Agusti Solanas. GPU Algorithms for K-Anonymity in Microdata. In *2019 IEEE Conf. Commun. Netw. Secur. CNS*, pages 1–9.

[17] Linh Ha, Jens Krüger, and Cláudio T Silva. Fast four-way parallel radix sorting on gpus. In *Computer Graphics Forum*, volume 28, pages 2368–2378. Wiley Online Library, 2009.

[18] K. E. Batcher. Sorting networks and their applications. In *Proc. April 30–May 2 1968 Spring Jt. Comput. Conf.*, AFIPS '68 (Spring), pages 307–314. Association for Computing Machinery, 1968.

[19] Fiaz Gul Khan, Omar Usman Khan, Bartolomeo Montrucchio, and Paolo Giaccone. Analysis of fast parallel sorting algorithms for gpu architectures'. In *2011 Frontiers of Information Technology*, pages 173–178, 2011.

[20] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. Flexible data anonymization using arx—current status and challenges ahead. Software: Practice and Experience, 50(7):1277–1304, 2020.

[21] Florian Kohlmayer, Fabian Prasser, Claudia Eckert, Alfons Kemper, and Klaus A. Kuhn. Flash: Efficient, stable and optimal k-anonymity. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 708–717, 2012.

[22] R.J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, pages 217–228, 2005.

[23] W. Kahan. Pracniques: Further remarks on reducing truncation errors. 8(1):40.

[24] Anil Shanbhag, Holger Pirk, and Samuel Madden. Efficient Top-K Query Processing on Massively Parallel Hardware. In *Proc. 2018 Int. Conf. Manag. Data*, SIGMOD '18, pages 1557–1570. Association for Computing Machinery, 2018.

[25] Anil Gaihre, Da Zheng, Scott Weitze, et al. Dr. Top-k: Delegate-centric Top-k on GPUs. In *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, SC '21, pages 1–14. Association for Computing Machinery, 2021.

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

🏛2024年3月1日(金) 10:00 ~ 12:10 | 🏛T2-B オンライン（Zoom Events)

時系列データ処理

座長:渡辺 陽介(名古屋大学)
コメンテータ:喜田 拓也(北海学園大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T2-B-7-01] クエリ長に依存しない多次元時系列データに対する類似問合せ手法の提案

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

10:20 ~ 10:35
[T2-B-7-02] NexmonによるCSIベースのスマートデバイス状態推定

*原田 海斗[1]、寺本 京祐[2]、野村 祐一郎[3]、峰野 博史[3,4] (1. 静岡大学 情報学部 情報科学科、2. 静岡大学 大学院 総合科学技術研究科 情報学専攻、3. 静岡大学 学術院 情報学領域、4. 静岡大学 グリーン科学技術研究所 フィールドインフォマティクス研究コア)

10:35 ~ 11:00
[T2-B-7-03] LSiX:ストリームデータに関する複数連続的集約

*川上 隼[1]、Bou Savong[2,1]、天笠 俊之[2,1] (1. 筑波大学、2. 計算科学研究センター)

11:00 ~ 11:25
[T2-B-7-04] 大規模データストリームに対する高速なS-FINCHクラスタリング

*牛尼 索造[1]、藤原 靖宏[2]、塩川 浩昭[3] (1. 筑波大学 情報学群情報科学類、2. NTTコミュニケーション科学基礎研究所、3. 筑波大学 計算科学研究センター)

11:25 ~ 11:50
[T2-B-7-05] 【技術報告】東芝の時系列データ処理基盤技術(GridDB)ならびにデータ仮想化エンジン(PGSpider)

浪岡保男、服部雅一、相川 恵 (株式会社東芝)

クエリ長に依存しない多次元時系列データに対する類似問合せ手法の提案

安田　裕真†　　塩川　浩昭††

† 筑波大学大学院理工情報生命学術院　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1-1-1
E-mail: †yasuda@kde.cs.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　多次元時系列データにおける類似部分シーケンス問合せとは，クエリとして入力された多次元時系列データに対して類似度の高い部分シーケンスをデータベースから探索する問題であり，医学やスポーツ科学の分野などで広く利用されている．問合せの効率化として，部分シーケンス長で多次元時系列データを分割することでクラスタリングを行う手法がある．この手法には，部分シーケンス長とクエリ長が異なると問合せを行えないという問題点がある．そこで本論文では，クエリ長に依存しない類似部分シーケンス問い合わせ手法を提案する．実データを用いた性能評価を行い，従来手法と比較して提案手法は高速かつ正確にクエリ長に依存しない類似部分シーケンス問合せを行えることを示す．

キーワード　時系列データ処理，データ構造・索引，問合せ処理

1　序　　論

多次元時系列データとはある現象の時間的な変化を連続的に観測して得られた多次元の実数値からなる系列（シーケンス）であり，医学やスポーツ科学の分野において，重要な要素技術となっている [1]．近年，多次元時系列データ解析において，クエリと類似した部分シーケンスを検索する類似部分シーケンス問合せ処理技術が注目を集めている．類似部分シーケンス問合せとは，クエリと多次元時系列データの取りうる全ての部分シーケンスとの間で類似度を計算し，類似度が高くなった部分シーケンスを類似部分シーケンスとして出力する処理である．類似部分シーケンス問合せでは多次元時系列データ内に存在する全ての部分シーケンスを検索の対象とするため，長いシーケンスに対して膨大な問合せ処理時間を要する問題がある．

これまで，類似部分シーケンス問合せ手法の様々な研究 [2] [3] がなされているが，いずれも 1 次元の時系列データに対する手法であり，多次元への応用はなされていなかった．また，多次元時系列データを対象とした研究 [4] [5] もなされてきたが，これらは多次元時系列データから最も類似した部分シーケンスのペアを検出するという手法であり，クエリに類似する部分シーケンスを検索することはできなかった．

これらの問題を解決するための手法として，局所性鋭敏型ハッシュ関数（LSH）を用いた LSH 手法が提案された [6]．LSH 手法は，多次元時系列データとクエリが与えられたときに，クエリとのピアソン相関が閾値 θ 以上の部分シーケンスを高速かつ正確に出力する．具体的には，天方らによる先行研究 [7] を多次元時系列データへと拡張し，多次元時系列データとクエリを事前に次元削減することで，全ての部分シーケンスを探索することなく検索を行う．

しかし，LSH 手法にはクエリ長と部分シーケンス長が一致しないと検索を行えないという問題がある．LSH 手法では，多次元時系列データの各部分シーケンスに対して LSH を用いてハッシュ値を計算し，ハッシュ値を基にクラスタリングを行っているため，クエリ長と部分シーケンス長が異なるとグループ検索を行うことができなかった．

そこで，本研究では LSH 手法を拡張することで，クエリ長に依存しない類似部分シーケンス問合せ手法の提案を行う．我々は，クエリ長と部分シーケンス長が異なる 3 つのケースに対して対応策を提案する．本論文では実データを用いた提案手法の評価実験を行い，提案手法が LSH 手法と比較して同程度の速さと精度でクエリ長に依存しない類似部分シーケンス問合せを行えることを確認した．

本論文の構成は以下の通りである．2 節で前提となる知識と本研究で取り扱う問題について説明し，3 節で LSH 手法を紹介する．4 節で提案手法について述べ，5 節で評価実験結果を示す．6 節で本研究のまとめを行う．

2　事 前 準 備

多次元時系列データ $\boldsymbol{T}$ とは z 正規化された時系列データの集合であり $\boldsymbol{T} = [T^{(1)}, T^{(2)},...,T^{(d)}]$ と表記する．d は多次元時系列データ $\boldsymbol{T}$ の次元数，$T^{(i)}$ は長さ n の 1 次元時系列データであり，$\boldsymbol{T} \in \mathcal{R}^{(d \times n)}$ である．ここで，$T^{(j)}$ において，先頭から i 番目の長さ m である部分シーケンスを $T^{(j)}_{i,m} = [t^{(j)}_i, t^{(j)}_{i+1}, \ldots, t^{(j)}_{i+m-1}]$ と表す．ただし，$t^{(j)}_{i+k}$ は $T^{(j)}_{i,m}$ の $i+k$ 番目の実数値を表す．また，$\boldsymbol{T_{i,m}}$ は $\boldsymbol{T}$ 内に含まれる先頭から i 番目の長さ m である部分シーケンス集合であり $\boldsymbol{T_{i,m}} = [T^{(1)}_{i,m}, T^{(2)}_{i,m}, \ldots, T^{(d)}_{i,m}]$ である．また，$\boldsymbol{T_{i,m}}$ を $(d \times m)$ 行列とみなしたとき，k 列目の列ベクトルを $\mathbf{t}^k_{i,m}$ とすると $\mathbf{t}^k_{i,m}=(t^{(1)}_{i+k-1}, t^{(2)}_{i+k-1}, \ldots, t^{(d)}_{i+k-1})^T$ と表すことができる．クエリデータ q とは，次元数 d，長さ m の実数値から構成される多次元時系列データであり，$q \in \mathcal{R}^{(d \times m)}$

である．

多次元時系列データに対する類似部分シーケンス問合せとは $\boldsymbol{T}$ とクエリデータ q が与えられたとき，クエリと類似した部分シーケンスを $\boldsymbol{T}$ から検出することである．本研究では部分シーケンス間の類似度としてピアソン相関を採用する．定義は以下の通りである．

定義1（ピアソン相関）．2つの部分シーケンス $\boldsymbol{T}_{a,m}$ と $\boldsymbol{T}_{b,m}$ の間のピアソン相関は以下のように求める．

$$\rho(\boldsymbol{T}_{a,m}, \boldsymbol{T}_{b,m}) = 1 - \frac{|dist(\boldsymbol{T}_{a,m}, \boldsymbol{T}_{b,m})|^2}{2m}$$

ただし，$dist(\mathbf{T}_{a,m}, \mathbf{T}_{b,m})$ は $\mathbf{T}_{a,m}$・$\mathbf{T}_{b,m}$ 間のユークリッド距離であり，以下のように計算する．

$$dist(\boldsymbol{T}_{a,m}, \boldsymbol{T}_{b,m}) = \sqrt{\sum_{k=1}^{m}(t_{a,m}^k - t_{b,m}^k)^2}$$

本研究で対象とする類似部分シーケンス問合せ問題を以下のように定義する．

定義2（類似部分シーケンス問合せ問題）．多次元時系列データ $\boldsymbol{T}$，クエリデータ q，閾値 $\theta \in [0, 1]$ が与えられたとき，ピアソン相関 $\rho(\boldsymbol{T}_{i,m}, q) \geqq \theta$ を満たすような全ての部分シーケンス $\boldsymbol{T}_{i,m}$ を $\boldsymbol{T}$ から検索する．

3 先行研究 LSH 手法

本節では LSH 手法 [6] について説明する．LSH 手法の目的は，多次元時系列データに対して高速かつ正確に類似部分シーケンス問合せを行うことである．3.1 節から 3.3 節で，LSH 手法の詳細な説明を行う．

3.1 ハッシュ処理

ハッシュ処理では多次元時系列データ $\boldsymbol{T}$ 内の各部分シーケンス $\boldsymbol{T}_{i,m}$ に LSH を適用し，ハッシュ値を得る．LSH 手法では，先行研究 [7] で提案された LSH を拡張し，以下のハッシュ関数を用いる．

定義 3（LSH）．部分シーケンス $\boldsymbol{T}_{i,m} \in \mathcal{R}^{(d\times m)}$ に対して，LSH $h(\boldsymbol{T}_{i,m})$ を以下のように定義する．

$$h(\boldsymbol{T}_{i,m}) = \frac{(\boldsymbol{T}_{i,m} \cdot \mathbf{a})^T \cdot \mathbf{b} + cw}{w}$$

ただし，$\mathbf{a} \in \mathcal{R}^{(d\times 1)}$，$\mathbf{b} \in \mathcal{R}^{(m\times 1)}$ はそれぞれ各要素が正規分布に従うランダムな値を持つ列ベクトルである．c は $[0, w)$ から選ばれたランダムな実数であり，w は定数である．上式の通り，LSH $h(\boldsymbol{T}_{i,m})$ は最終的に，1つの実数値を出力する．

ここで，ピアソン相関の閾値を θ とおくと，本研究では $\rho(\boldsymbol{T}_{a,m}, \boldsymbol{T}_{b,m}) \geqq \theta$ となるような類似部分シーケンスに注目し，これらのハッシュ値が類似するようにしたい．そこで先行研究 [7] に従い，$w = \sqrt{2m(1-\theta)}$ とする．

LSH 手法は定義 3 に示したハッシュ関数を L 個用意して，各部分シーケンスに対して L 個のハッシュ値 H=[$h_1(\boldsymbol{T}_{i,m})$, $h_2(\boldsymbol{T}_{i,m})$, ..., $h_L(\boldsymbol{T}_{i,m})$] $(1 \leqq i \leqq n-m+1)$ を生成する．これにより，サイズ $L\times(n-m+1)$ の行列であるハッシュ値系列 H を得る．

3.2 グループ化

グループ化では，ハッシュ値系列 H=[$h_1(\boldsymbol{T}_{i,m})$, $h_2(\boldsymbol{T}_{i,m})$, ..., $h_L(\boldsymbol{T}_{i,m})$] $(1 \leqq i \leqq n-m+1)$ の各部分シーケンス同士のピアソン相関を求め，相関が強い部分シーケンスを同じグループに格納する．具体的には H の各列間のピアソン相関を求める．このとき，2つの部分シーケンスが $\rho \geqq \theta$ となる場合，同じグループに格納していく．この操作を全ての部分シーケンスの組合せに対して実行し，ピアソン相関の大きな部分シーケンス同士を同じグループに格納する．グループ化によって作成されたグループの集合を *H-group* と呼ぶ．

3.3 検　　索

クエリデータ q が到着したら，時系列データと同様のハッシュ処理を q に行い，H_q=[$h_1(q)$, $h_2(q)$, ..., $h_L(q)$] とする．H_q と *H-group* の各グループとのピアソン相関を求め，閾値 θ よりも大きいグループを検索．検索されたグループの集合を H_q*-group* とする．ここで H_q*-group* 内の各部分シーケンスと H_q を元のデータに復元する．同様に復元したクエリと元の部分シーケンスのピアソン相関を計算し，ピアソン相関が閾値 θ よりも大きい部分シーケンスを出力する．

4 提 案 手 法

本節では，LSH 手法を拡張した提案手法の説明を行う．提案手法の目的は，LSH 手法においてクエリ長が部分シーケンス長と一致しない場合でも，類似部分シーケンス問合せを行えるようにすることである．クエリ長と部分シーケンス長が異なるケースとして，以下の 3 つのケースが考えられる．

ケース1：クエリ長が部分シーケンス長より長く，部分シーケンス長を組み合わせることでクエリ長と一致するケース

ケース2：クエリ長が部分シーケンス長より長く，部分シーケンス長を組み合わせてもクエリ長と一致しないケース

ケース3：クエリ長が部分シーケンス長より短いケース

これら 3 つのケースそれぞれに対して対応策を提案する．ここで，クエリ長を *Query-m*，部分シーケンス長を *Sub-m* と表記することとする．

4.1 ケ ー ス 1

クエリ長 *Query-m* が部分シーケンス長 *Sub-m* より長く，*Sub-m* を組み合わせることで *Query-m* と一致するケースでは，クエリを *Sub-m* に合わせて分割を行い，分割したクエリごとに検索を行う．具体例として図 1 に示す．図のインデックスは，部分シーケンスが始まる位置を表している．まず，(1) のように到着したクエリを *Sub-m* で分割して，分割したクエリごとにグループの検索を行う．次に，(2) のように分割したクエリごとに検索されたグループ内の部分シーケンスのインデックス間の差を計算する．インデックス間の差が *Sub-m* と一致する場合，その 2 つの部分シーケンスは連続した位置関係にあるこ

とがわかる．この性質を利用して部分シーケンスを合成することで，*Query-m* と等しい *Sub-m* の類似部分シーケンスを検索することができる．クエリに類似した部分シーケンス検索後の操作は，LSH 手法と同じである．

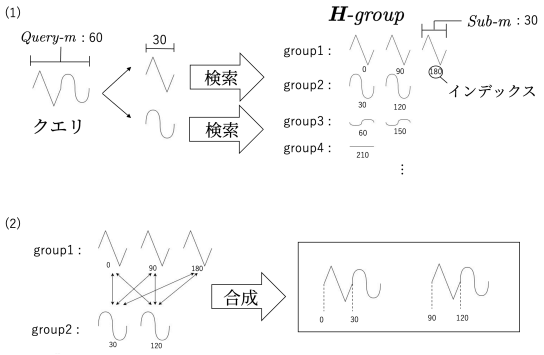


図 1　ケース 1 の例

4.2 ケース 2

クエリ長 *Query-m* が部分シーケンス長 *Sub-m* より長く，*Sub-m* を組み合わせても *Query-m* と一致しないケースでは，クエリを *Sub-m* の部分と余剰部分に分割を行い，余剰部分の長さに応じて操作を分ける．具体的には，余剰部分の長さが *Sub-m* の $\frac{1}{2}$ 以上のときと $\frac{1}{2}$ 未満のときで操作を分ける．

余剰部分が *Sub-m* の $\frac{1}{2}$ 以上のときは，余剰部分が *Sub-m* と等しくなるように余剰部分の平均値で埋める．具体例として図 2 の (1) に示す．図のように，余剰部分の長さが 27 と，*Sub-m* の $\frac{1}{2}$ 以上のときは，余剰部分の平均値で埋める．この理由は，余剰部分の特徴を失わせないためである．LSH を用いたハッシュ処理を行うため，余剰部分に全く関係のない乱数で埋めてしまうと，本来の余剰部分と類似していたグループを検索することができない．そこで，余剰部分の平均値で埋めることで，特徴を最低限残すようにする．余剰部分を埋めた後の操作は，ケース 1 と同じである．

余剰部分が *Sub-m* の $\frac{1}{2}$ 未満のときは，余剰部分を除外する．具体例として図 2 の (2) に示す．図のように，余剰部分の長さが 10 と，*Sub-m* の $\frac{1}{2}$ 未満のときは，余剰部分を除外する．この理由は，グループ検索の精度を下げないためである．*Sub-m* の $\frac{1}{2}$ 未満の余剰部分の平均で埋めてしまうと，余剰部分の特徴を捉えるには不十分な特徴のみが残ってしまい，本来検索されるであろうグループと異なるグループが検索されてしまう．このことを避けるためにも，余剰部分を除外して検索を行うことで，検索時間は長くなってしまうが精度の面では保証することができる．余剰部分を除外した後の操作は，ケース 1 と同じである．

4.3 ケース 3

クエリ長 *Query-m* が部分シーケンス長 *Sub-m* より短いケースでは，いかなる場合でもクエリの平均でクエリを埋める必要がある．与えられたクエリ以外にグループ検索に使える要素は無いため，精度が下がってしまってもクエリの平均で埋めて検索を行う必要がある．

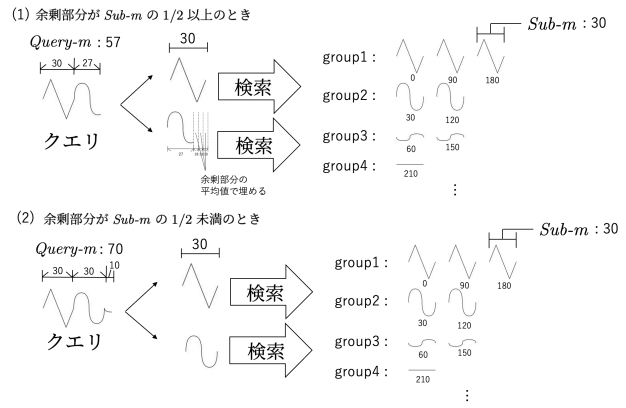


図 2　ケース 2 の例

5 評価実験

提案手法の検索時間と処理精度の評価を行うために，比較手法として 3 つの手法を用意する．1 つ目の手法は，クエリデータと多次元時系列データの全ての部分シーケンスとのピアソン相関を計算するという手法である．これをベースライン手法と呼ぶ．2 つ目の手法は，離散フーリエ変換（DFT）を用いて次元削減を行う手法 [3] を多次元に応用した手法を用いる．文献 [3] では，1 次元の時系列データに対してクエリとの類似度が高い類似部分シーケンス検索を行っているため，この手法を多次元に応用する．具体的には，多次元時系列データの各次元に対して文献 [3] の手法を用いて類似部分シーケンス検索を行い，各次元から得られる結果を組み合わせることで類似部分シーケンス検索を行う．この手法を DFT 手法と呼ぶ．3 つ目の手法は，3 章で紹介を行った LSH 手法 [6] である．検索時間と処理精度を評価するために，実データを利用する．データセットの詳細は 5.1 節に述べる．5.2 節では提案手法と比較手法の問い合わせ処理の評価実験の結果を示す．5.3 節では提案手法のクエリ長と部分シーケンス長の関係による，問合せ処理時間と処理精度の変化について検証を行う．

5.1 実験データセット詳細

実データセットは論文やデータセットを公開しているサイトから引用したデータである．実データセットの詳細として，以下の表 3 に示す．

表 1　実データセットの詳細

データセット	n	d	詳細
Matrix profile Dataset	550	3	Matrix Profile の論文内で用いられたデータセット [9]
Basket Dataset	10,000	6	バスケットボールの運動データセット [10]
Smartphone Dataset	152888	9	スマートフォンの位置情報データセット [11]
Mt.Tsukuba Dataset	158422	11	筑波山の気象データセット [12]

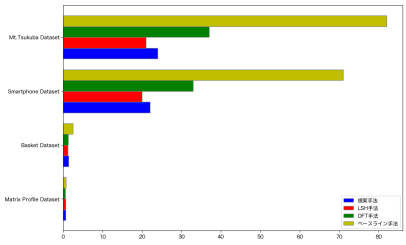


図 3　ケース 1

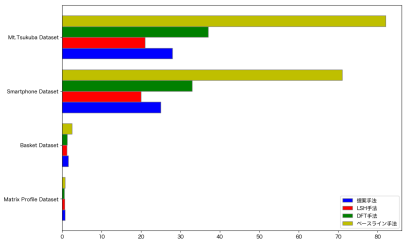


図 4　ケース 2

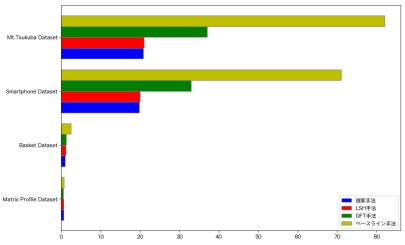


図 5　ケース 3

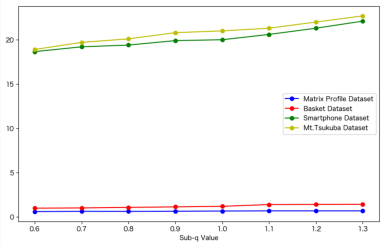


図 6　問合せ処理時間の変化

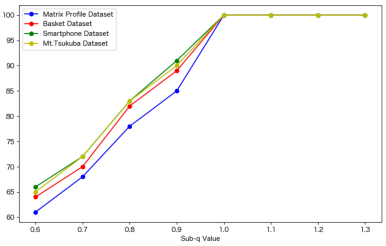


図 7　処理精度の変化

5.2　評価実験結果

本節では，提案手法と比較手法を用いて問合せ処理の評価実験を行う．評価方法として問合せ処理の実行時間と処理精度の計測を行う．

5.2.1　問合せ処理の実行時間

提案手法と比較手法の問合せ処理時間の検証を行う．実験結果を図 3,4,5 に示す．結果から，ケース 1,2 のような LSH 手法よりも計算量が増加するケースにおいては，LSH 手法よりも多くの問合せ処理時間を要するが，計算量が増加しないケース 3 においては，LSH 手法と同等の速さで問合せ処理を行えることがわかる．また，ケース 1,2,3 のいずれにおいても，ベースライン手法と DFT 手法よりも高速に問合せ処理を行えることがわかる．これは，提案手法の計算量が，DFT 手法とベースライン手法の計算量よりも少ないことを表している．

5.2.2　問い合せ処理の処理精度

提案手法と比較手法の問合せ処理の処理精度の検証を行う．検証準備として実データの 1 番目の部分シーケンスをクエリデータ q として抜き出す．1 番目の部分シーケンスを抜き出した実データに対して，ベースライン手法を用いてクエリとのピアソン相関を求め，ピアソン相関係数が閾値 θ 以上の部分シーケンスをあらかじめ検索しておく．この実データにおいて提案手法と比較手法を用いて類似データの検索を行い，処理精度の検証を行う．提案手法と LSH 手法では $L = 10$ でハッシュ処理を行う．また，提案手法ではクエリ長が部分シーケンス長の 2 倍であると仮定して検索を行う．本研究では処理精度として，適合率と再現率を用いる．ベースライン手法によって検索された正解データ集合を G，提案手法によって検索された部分シーケンス集合を R とすると，適合率と再現率はそれぞれ以下のように表せる．

$$\text{適合率} = \frac{|G \cap R|}{|R|}, \quad \text{再現率} = \frac{|G \cap R|}{|G|}$$

検証結果を表 2 に示すに示す。

表 2　処理精度結果

処理精度	Matrix Profile	Basket	Smartphone	Mt.Tsukuba
提案手法の適合率	100%	100%	100%	100%
提案手法の再現率	100%	100%	100%	100%
LSH 手法の適合率	100%	100%	100%	100%
LSH 手法の再現率	100%	100%	100%	100%
DFT 手法の適合率	100%	97%	91%	92%
DFT 手法の再現率	100%	100%	100%	100%

検証結果から，提案手法，LSH 手法ともに 100%の精度で検索を行えることがわかる．これは，グループ検索の後に元の時系列データへの復元を行っているためである．一方，DFT 手法では適合率が下がってしまう場合があった．これは，文献 [3] にも述べられているように，類似部分シーケンスを見逃すことはないが，多く検索してしまうことはあるという性質によるものだと考えられる．

5.3　提案手法の問合せ処理時間と処理精度の変化

本節では，提案手法のクエリ長と部分シーケンス長の関係による，問合せ処理時間と処理精度の変化について検証を行う．提案手法の余剰部分の除外や平均値で埋めるという操作が，問合せ処理時間や処理精度にどのような影響をもたらすのかを検証する．具体的には，実データセットに対して提案手法を用いて検索を行う際に，クエリ長を部分シーケンス長の 0.6 倍から 1.3 倍まで変化させて検証を行う．

検証結果を図 6,7 に示す．図 6 は問合せ処理時間の変化を，

図 7 は処理精度の変化を表している．また，図 6,7 の x 軸はクエリ長を表している．例えば，$x = 0.8$ のときは，クエリ長が部分シーケンス長の 0.8 倍であることを表している．図 6 の y 軸は問合せ処理時間（秒）を，図 7 の y 軸は処理精度（%）を表している．

図 6 より，クエリ長は問合せ処理時間に大きな影響を与えないことがわかる．一方，図 7 より処理精度には大きな影響を与えることがわかる．特に，クエリ長が部分シーケンス長より短いときは，余剰部分を平均値で埋める操作を行っているからだと考えられる．しかし，クエリ長が部分シーケンス長より長いときは高い処理精度を出せている．これは，余剰部分を除外する操作は処理精度に影響を与えないことを表している．

6 結　　論

本研究ではクエリ長に依存しない多次元時系列データに対する高速な類似部分シーケンス問合せ手法を提案した．実データを用いた評価実験において，比較手法に対し提案手法が有効であることを確認できた．今後の課題として，クエリ長が部分シーケンス長よりも短い場合の処理精度の改善が挙げられる．この部分を改善することで，より多くの多次元時系列データに対して提案手法を用いることが可能になると考えられる．

謝　　辞

本研究の一部は，JST さきがけ (JPMJPR2033) ならびに JSPS 科研費 (JP22K17894) の支援を受けたものである．

文　　献

[1] Ryuichi Yagi and Hiroaki Shiokawa. Fast top-k similar sequence search on dna databases. In *Information Integration and Web Intelligence: 24th International Conference, iiWAS 2022, Virtual Event, November 28–30, 2022, Proceedings*, pp. 145–150. Springer, 2022.

[2] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, 2003.

[3] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. *Acm Sigmod Record*, Vol. 23, No. 2, pp. 419–429, 1994.

[4] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 1317–1322. Ieee, 2016.

[5] Gineke A Ten Holt, Marcel JT Reinders, and Emile A Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, Vol. 300, p. 1, 2007.

[6] Yuma Yasuda and Hiroaki Shiokawa. Efficient similarity searches for multivariate time series: A hash-based approach. In *International Conference on Information Integration and Web Intelligence*, pp. 478–490. Springer, 2023.

[7] 天方大地, 原隆浩. 相関時系列データ集合の計算のための高速アルゴリズム. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2017.

[8] 古賀久志. ハッシュを用いた類似検索技術とその応用. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 7, No. 3, pp. 256–268, 2014.

[9] Chin-Chia Michael Yeh, Nickolas Kavantzas, and Eamonn Keogh. Matrix profile vi: Meaningful multidimensional motif discovery. In *2017 IEEE international conference on data mining (ICDM)*, pp. 565–574. IEEE, 2017.

[10] Abdulwahed Salam and Abdelaaziz El Hibaoui. Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city –. In *Proceedings of the 6th International Renewable and Sustainable Energy Conference (IRSEC 2018)*, pp. 1–5, 2018.

[11] Paolo Barsocchi, Antonino Crivello, Davide La Rosa, and Filippo Palumbo. A Multisource and Multivariate Dataset for Indoor Localization Methods based on WLAN and Geo-Magnetic Field Fingerprinting. In *Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN 2016)*, pp. 1–8, 2016.

[12] University of Tsukuba Center for Computational Sciences. Mt. tsukuba project, 2023. (July 25th, 2023 Accessed).

NexmonによるCSIベースのスマートデバイス状態推定

原田　海斗†　　寺本　京祐††　　野村裕一郎†††　　峰野　博史††††

† 静岡大学情報学部　〒 432–8011 静岡県浜松市中区城北 3 丁目 5-1
†† 静岡大学大学院総合科学技術研究科情報学専攻　〒 432–8011 静岡県浜松市中区城北 3 丁目 5-1
††† 静岡大学学術院情報学領域　〒 432–8011 静岡県浜松市中区城北 3 丁目 5-1
†††† 静岡大学グリーン科学技術研究所　〒 432–8011 静岡県浜松市中区城北 3 丁目 5-1
E-mail: †{harada.kaito.20,teramoto.kyosuke.18}@shizuoka.ac.jp,
††{nomura.yuhichiro,mineno}@inf.shizuoka.ac.jp

あらまし　現在のスマートデバイスは，利用者自身でも動作状態の把握が難しく，不正な通信や動作が行われている場合に気づく術がない．そこで，スマートデバイスの状態 (アプリケーションの動作状況，操作内容など) を推定するシステムの実現を目指している．従来の研究では，スマートデバイスの状態推定手法として様々なアプローチが検討されているが，汎用性や導入コストの面で懸念が残る．本研究では，Wi-Fi 電波の通信媒体波及時におけるチャネル状態情報 (Channel State Information; CSI) を，CSI 収集用ファームウェアパッチである Nexmon を用いて収集・分析し，スマートデバイス状態推定を行う手法を提案する．基礎評価として，6 種類のアプリケーションに対して 8 種類の時系列モデルでの試行の結果，最大 87.6%の精度でアプリケーション推定が可能なことを確認した．また，分類アプリケーションの組み合わせによる精度変化を検証し，最大 100%で推定できることを明らかにした．
キーワード　Wi-Fi，CSI，Nexmon，時系列データ，スマートデバイス，アプリケーション推定

1　はじめに

近年，スマートフォンやパソコンなど多機能 IoT デバイス (スマートデバイス) が広く普及し，その活用シーンは多岐にわたる．世界のスマートデバイスの普及台数は年々増加しており，2025 年には約 440 億台に達すると予測されている．スマートデバイスの増加に伴って，攻撃者によるスマートデバイスに対しての不正操作や情報漏洩が懸念される．しかし，現在のスマートデバイスは，利用者自身でも動作状態の把握が難しく，不正な通信や動作が行われている場合に気づく術がない．

そこで，スマートデバイスにおけるアプリケーションの動作状況を推定するシステムの実現を目的としている．このシステムを活用することで，利用者自身がスマートデバイスの動作状況を把握し，第三者による不正な操作や情報漏洩を未然に防ぐことができる．また，スマートデバイスの使用状況を管理するシステムに組み込むことによって，より頑健なスマートデバイス管理システムの実現も期待できる．

従来の研究では，スマートデバイスの通信トラヒックから得られる統計情報を用いる手法 [1] や，状態遷移モデルと IoT センサ情報を用いたエッジコンピューティングによる手法 [2] が検討されている．しかし，通信トラヒックを用いた手法では，情報量に限界がありスマートデバイスの状態を詳細に捉えることが難しい．また，状態遷移モデルと IoT センサ情報を用いた手法では，状態遷移モデルの汎用性，導入コストの高さが懸念される．

本研究では，Wi-Fi 電波の通信媒体波及時における CSI を，CSI 収集用ファームウェアパッチである Nexmon [3] を用いて収集・分析し，スマートデバイス状態推定を行う手法を提案する．

2　関連研究

2.1　通信トラヒック分析による複数のスマートデバイスにおける機能推定手法の評価 [1]

ここでは，複数のスマートデバイスが有線・無線でルータに接続される環境を想定している．接続されるデバイス全てが，単一のエッジルータを介してインターネット接続を行う環境で通信トラヒックを収集している．また，通信トラヒックから 32 個の特徴量を算出し，RandomForest へ入力することで推定を行う手法が提案されている．評価データとして，8 機種のスマートデバイスで 8 種類の機能を実行した際の通信トラヒックを収集している．結果として，スマートデバイスの機種と機能の 16 通りに対して，88%の精度で分類できることを明らかにした．

しかし，通信トラヒックを用いた手法では，機種や機能による通信トラヒックの差異が小さく，スマートデバイスの状態を詳細に捉えることが難しい．これは，通信トラヒックによる特徴量が持つ情報量が少ないことが起因している．そのため，より詳細なスマートデバイスの状態推定を行うためには，通信トラヒック以外の情報を用いる必要がある．

2.2　状態遷移モデルと IoT センサによるエッジコンピューティングを用いたスマートデバイスの異常検知 [2]

実際の家庭環境に設置されているスマートデバイスの使用状

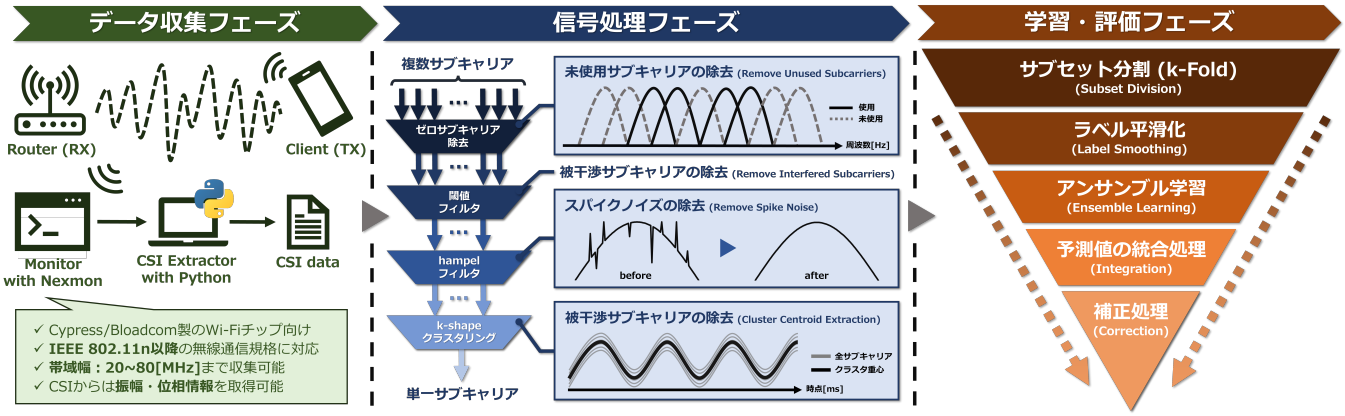


図 1　提案手法：概要図

況を，各デバイスに対応する IoT センサ値とエッジコンピュータを用いて収集している．また，家庭の活動状況を考慮した状態遷移モデルを，IoT センサ値を用いて構築し，不正操作を検出する手法が提案されている．

結果として，誤検出率：20.1%未満，検出率：72.3%でスマートデバイスの状態推定が可能なことを明らかにした．

しかし，この手法では各スマートデバイスに対して状態遷移モデルを構築する必要があり，汎用性の観点で懸念が残る．また，スマートデバイスの操作状況を収集するために，各デバイスそれぞれに IoT センサと，エッジコンピュータが必要であるため，導入コストの高さも懸念される．そのため，より汎用性が高く，導入コストの低いスマートデバイスの状態推定手法が必要である．

3　提 案 手 法

Wi-Fi 電波の通信媒体波及時における CSI を，CSI 収集用ファームウェアパッチである Nexmon を用いて収集・分析し，スマートデバイス状態推定を行う手法を提案する．CSI は，通信技術分野において広く使用される通信路の状態情報である．具体的には，Wi-Fi 電波が送受信デバイス間で伝播する際の電波散乱，フェージング，到来距離・角度による電力減衰などの複合的な要因を考慮したチャネル状態を表し，振幅・位相情報を持つ．CSI は，本来は通信品質の評価や通信路推定などに使用されるが，本研究では，スマートデバイスの動作状況を推定するための特徴量として利用する．CSI は通信トラヒックや RSSI(Received Signal Strength Indicator) と比較して取得できる情報が多く，スマートデバイスの状態を詳細に捉えることができる．CSI を収集するソフトウェアとして Nexmon を使用する．Nexmon を使用することで，CSI 収集不可能な低コストデバイスでも CSI 収集可能となる．

図 1 に提案手法の概要図を示す．提案手法は，「データ収集フェーズ」「信号処理フェーズ」「学習・評価フェーズ」の 3 つのフェーズで構成される．

3.1　データ収集フェーズ

「データ収集フェーズ」では，CSI を Nexmon を適用したデバイスを用いて取得し，Python ベースの CSI 解析プログラムを用いて，各サブキャリアの振幅・位相情報を取得する．このとき，スマートデバイスが Wi-Fi ルータのクライアントとして IEEE802.11n 規格以降の通信方式 (OFDM や MIMO などのマルチキャリア・アンテナ通信) かつ，帯域：2.4/5[GHz] で帯域幅：20/40/80[MHz] で通信を行う状況を想定する．

Nexmon 適用デバイスは，指定されたチャネルと帯域幅に該当する通信を観測するため，スマートデバイスが接続しているネットワークに接続する必要はない．そのため，多様なタスクに対して適用可能なデータ収集手法であり，システム全体として汎用性が高くなる．また，新たに導入するデバイスは Nexmon を適用するデバイスのみであるため，従来手法と比較して導入コストが低い．

3.2　信号処理フェーズ

「信号処理フェーズ」では，データ収集フェーズで取得した CSI データの前処理を行う．前処理として，ゼロサブキャリア除去，閾値フィルタ，hampel フィルタ，k-shape クラスタリングを適用する．

ゼロサブキャリア除去は，CSI データの中で振幅が 0 のサブキャリアを除去する処理である．IEEE802.11n 規格以降で使用されるマルチキャリア通信では他チャネルとの干渉を抑制するため，一部のサブキャリアをデータ通信に使用しない工夫がされている [4]．そのため，それらのサブキャリア情報の除去をすることで通信に使用されたサブキャリア情報のみを抽出する．

閾値フィルタでは，振幅情報が一定値を超えるサブキャリアを除去するフィルタである．ゼロサブキャリアに隣接するサブキャリアは他チャネルからの干渉を受けやすく，ノイズが多く含まれるため，学習には不向きである．具体的には，ゼロサブキャリアを除く k 個のサブキャリア $S_k \ni s_k$ に対して式 1 を適用し，閾値 α を超えるサブキャリアを除去する．

$$threshold(S_k) = \begin{cases} S_k & if\ s_k < \alpha \\ None & otherwise \end{cases} \quad (\alpha > 0) \qquad (1)$$

hampel フィルタ [5] では，選定されたサブキャリア集合の各要素に対してノイズ処理を行う．Nexmon によって収取された

CSI データにはスパイクノイズが多く含まれている．そのため，外れ値検知アルゴリズムである hampel フィルタはノイズ除去として有効に作用する．具体的には，単一サブキャリアに対して，中央値 $\widetilde{X_i}$，標準偏差 σ_i のスライディングウィンドウ $X_i \ni x_i$ を作成する．その後，全スライディングウィンドウに式 2 を適用し，$\beta\sigma_i$ を超える要素 x_i を中央値 $\widetilde{X_i}$ に置換する．

$$hampel(X_i) = \begin{cases} \widetilde{X_i} & if\ |x_i - \widetilde{X_i}| > \beta\sigma_i \\ x_i & otherwise \end{cases} \quad (\beta > 0) \quad (2)$$

k-shape クラスタリングでは，形状ベースの時系列クラスタリング手法である k-shape [6] を適用し，全サブキャリアクラスタの重心信号を抽出する．ここで，サブキャリアは変動幅が異なる程度の差異しか見られず，相関係数が非常に高い．そのため，全サブキャリア情報をすべて使用すると学習コストが高くなり，共線性による過学習に陥る可能性がある．時系列クラスタリングによって得られる重心信号を代表値として学習に使用することで，学習コストの低減を図る．

3.3 学習・評価フェーズ

「学習・評価フェーズ」では，前処理を行った CSI データを入力として時系列モデルによる学習・評価を行う．具体的には，ラベル平滑化 [7] を施した学習データに対して，k-Fold アンサンブル学習モデルを構築し，予測値の統合・補正を行うことで最終予測値を得る．ラベル平滑化は，K クラスの One-Hot エンコーディングされた目的変数 $y_k = [y_1, ..., y_K]$ に対して，ノイズ γ を加える手法 (式 3) である．これにより，モデルの過学習を抑制し，類似した特徴量に対する分類精度，汎化性能が向上する．

$$smoothing(y_i) = y_i(1-\gamma) + \frac{1}{K}\gamma \quad (0 < \gamma < 1.0) \quad (3)$$

k-Fold アンサンブル学習時において使用するサブセットはスライディングウィンドウ形式で作成され，目的変数の分布はランダムである．また，スライディングウィンドウは，ウィンドウサイズ (入力時点数) が可変かつ，複数の目的変数が含まれないように作成する．分割された k 個のサブセットそれぞれに対して，同一パラメータ・アーキテクチャの学習モデルを構築する．得られる k 個の予測値は，相加平均によって統合する．また，統合された予測値に対して，過去の予測値から妥当性を評価し，適切な値に補正することで最終予測値を得る．

4 基礎評価

4.1 データ収集フェーズ

図 2 に，基礎評価における CSI データの収集環境を示す．スマートデバイスとして iPhone12，Wi-Fi ルータとして WSR-1800AX4P-WH を使用した．また，スマートデバイス上で動作するアプリケーションは，TikTok，LINE，パズル&ドラゴンズ (Puzzle)，雀魂 (Mahjong)，YouTube，コミックシーモア (Comic) の 6 種類とした．Nexmon を適用する CSI 収集デバイスとして，安価かつ入手が容易なことから RaspberryPi4B を使用した．通信条件は，スマートデバイスと Wi-Fi ルータ間の距離を 1.5[m] で配置し，通信規格：IEEE802.11n の帯域：2.4[GHz]，帯域幅：20[MHz] として通信を行った．収集条件は，サンプリングレート：5.0[Packet/s] として収集を行った．

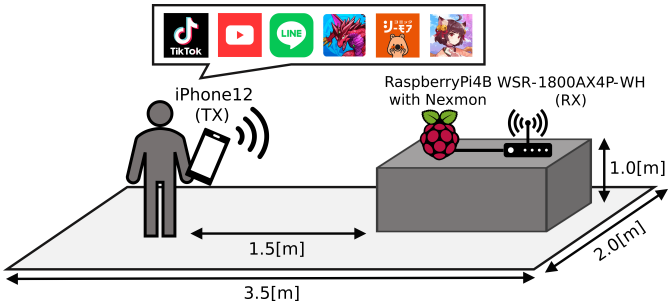


図 2　基礎評価：CSI データ収集環境

表 1　実 行 環 境

CPU	Intel Core i9-13900KF
GPU	GeForce RTX 3090
OS	Ubuntu 22.04.2 LTS
Software	Python3.9, tensorflow2.12, CUDA11.8

4.2 信号処理フェーズ

収集した CSI は，64 個のサブキャリア情報で構成される．IEEE802.11n 規格では，8 つのサブキャリアはデータ搬送には使用されないため，ゼロサブキャリアとして除去を行った．また，式 1 の閾値フィルタ ($\alpha = 3000$) を適用し，振幅情報が一定値を超えるサブキャリアの除去を行った．また，抽出された各サブキャリアに対して，式 2 の hampel フィルタ ($\beta = 2.0$) によるスパイクノイズの除去を行い，k-shape を用いて全サブキャリアの代表値を抽出した．

4.3 学習・評価フェーズ

学習・評価フェーズでは，表 1 に示す実行環境下で学習を行った．目的変数に対して式 3 のラベル平滑化処理 ($\gamma = 0.1$) を施した．また，k-Fold アンサンブル学習モデルに入力するサブセットの分割数は 3 とし，相加平均によって予測値の統合を行った．補正処理として，分類タスクにおいて誤分類した箇所がスパイク状に出現することから，式 2 の hampel フィルタ ($\beta = 1.5$) を使用した．

4.3.1 時系列モデルと入力時点数の選定

k-Fold アンサンブル学習モデルとして，8 種類の時系列モデル (RandomForest，TCN，1D-CNN，LSTM-FCN，Transformer，1D-ResNet，1D-DenseNet) を試行した．

表 2 に，各時系列モデルによる精度比較を示す．各時系列モデルに対して，入力時点数を 200，250，300，350，400，450[Packet] と変化させ，各学習モデルで得られる最大精度を記録した．各時系列モデルにおける準最大精度を下線，最大精度を太字で表記した．ラベル数に極端な偏りがないことから，評価指標は Accuracy[%] とした．最良結果として，学習

表 2　時系列モデルと入力時点数による精度比較 (Accuracy[%])

入力時点数	時系列モデル							
	RandomForest	TCN	1D-CNN	LSTM-FCN	Transformer	1D-ResNet	1D-DenseNet	LSTM
200	**67.9**	65.9	66.7	63.4	56.7	**72.1**	57.5	36.7
250	64.0	72.3	70.9	62.3	60.9	67.4	75.4	**51.6**
300	63.5	84.0	70.2	68.4	59.7	71.1	76.5	41.2
350	60.8	**87.6**	71.1	**68.5**	61.6	63.9	73.9	40.9
400	57.1	71.7	72.0	63.1	61.8	68.6	72.1	47.1
450	58.5	60.4	**76.0**	68.0	**62.4**	66.3	**80.9**	39.3

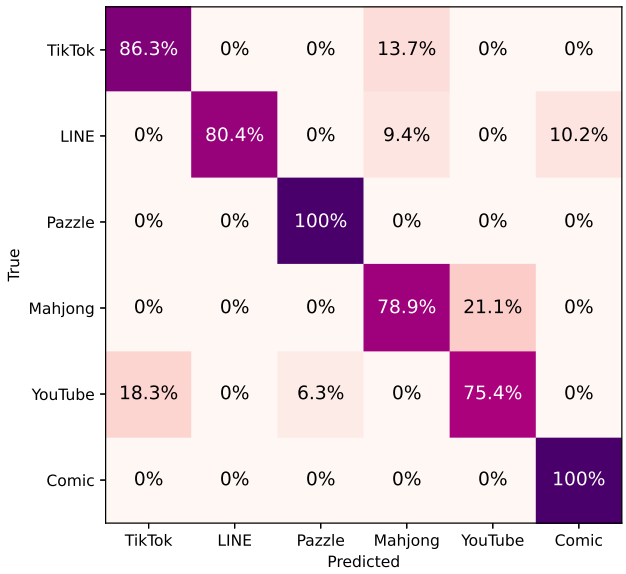


図 3　最良結果における分類結果 (混同行列)

モデル：TCN(Temporal Convolutional Network)，入力時点数：350[Packet](≃ 70[s]) とした時の Accuracy：87.6[%](F 値：0.871) を確認した．

図 3 に，表 2 で最良結果を確認した条件下での分類結果 (混同行列) を示す．分類難易度がアプリケーションによって異なり，誤分類しやすい組み合わせが存在することが確認できる．

4. 3. 2　入力時点数による精度比較

本節の第 1 項で最良結果を確認した学習モデル：TCN に対して，入力時点数を変化させた場合の精度比較を行った．入力時点数を 100〜500[Packet]，hampel フィルタのパラメータを $\beta = 1.0$〜3.0 として試行した．

図 4 に，各入力時点数ごとに得られた最大精度の推移を示す．

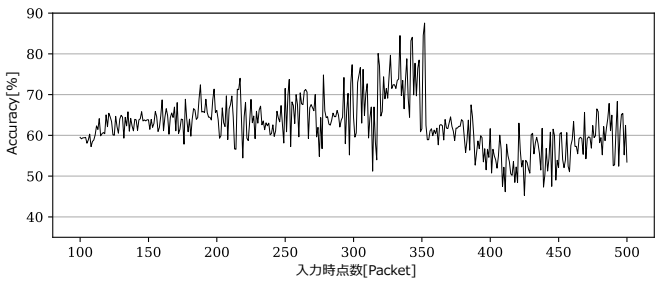


図 4　入力時点数による精度比較

表 3　アプリケーションによる精度比較 (● 分類対象，○ 非分類対象)

アプリケーションの組み合わせ						Acc.[%]
TikTok	LINE	Puzzle	Mahjong	YouTube	Comic	
●	●	●	●	●	○	94.4
●	●	●	●	○	●	71.9
●	●	●	○	●	●	90.7
●	●	○	●	●	●	89.1
●	○	●	●	●	●	94.1
○	●	●	●	●	●	89.8
●	●	●	●	○	○	66.5
●	●	●	○	○	●	66.4
●	●	○	○	●	●	91.8
●	○	○	●	●	●	90.3
○	○	●	●	●	●	95.4
○	●	●	●	●	○	**100.0**

入力時点数が増加するに伴い，分類精度の向上する傾向が確認できる．しかし，それと同時に入力時点数が 350[Packet] を超えたあたりで急激な精度低下がみられる．

4. 3. 3　アプリケーションによる精度比較

本節の第 1 項で最良結果を確認した学習モデル：TCN に対して，アプリケーションの組み合わせによる精度比較を行った．

表 3 に，分類対象とするアプリケーションの組み合わせに対する精度比較を示す．分類対象とするアプリケーションの組み合わせによって，分類精度に差異が生じることが確認できる．

5　考　　察

はじめに，図 3 における分類結果について考察する．“Puzzle” と “Comic” はどのアプリケーションとも誤分類することなく，適切に分類ができている．しかし，“Mahjong” と “YouTube” は分類精度が 80%を下回っており，基礎評価で対象としたアプリケーションの中で分類難易度が高いと言える．“Mahjong” は “YouTube” との誤分類が目立っているが，これは両者の操作方法 (横持ち，操作時の指の動きなど) が類似することに起因していると考えられる．また，“YouTube” は “TikTok” との誤分類が目立っており，これは両者の内部特性 (ストリーミング，バッファリングなど) が類似することに起因していると考えられる．

次に，図 4 における入力時点数による精度比較ついて考察する．一般的に，入力時点数が増加すると入力データの情報量が増加し，学習モデルの分類精度は増加する．しかし，特定の入力時点数 (=350[Packet]) を超えると，各アプリケーションが

持つ特徴量の差異が小さくなり，誤分類を引き起こしている可能性がある．

最後に，表 3 におけるアプリケーションによる精度比較について考察する．一般的に，分類クラス数の減少に伴って，分類精度は増加する．しかし，分類対象とするアプリケーション数が減少した場合に，分類精度が低下する組み合わせが存在している．提案手法における学習・評価フェーズでは，モデルの汎化性能の向上を目的としたラベル平滑化処理行っている．ラベル平滑化処理は，分類クラス内に類似する特徴量を持つ組み合わせが存在する場合に有効な手法である．しかし，分類アプリケーション内に特徴量が類似する組み合わせが存在しない場合，ラベル平滑化はモデルの性能を低下させてしまう可能性がある．このことから，分類対象とするアプリケーションの組み合わせによっては類似した特徴量を持つアプリケーションが少ないため，分類精度が低下していると考えられる．

6 議　　論

はじめに，スマートデバイス状態推定において使用する CSI の特徴量に影響を与える要因について議論する．スマートデバイス操作によって発生するトラヒックやトランザクション，操作方法・位置・環境などが挙げられる．それらを明かにするためには，トラヒックやトランザクションを可視化し，CSI 特徴量との関連性を検証する必要がある．また，非操作時と操作時の場合と操作位置に動的な変化を加えた場合の CSI 特徴量をそれぞれ分析する必要がある．

次に，基礎評価で対象としたスマートデバイス，およびアプリケーション以外に対する有効性について議論する．スマートデバイスに関しては，通信を行う無線通信規格，帯域幅などの通信条件を一致させた場合は，他のスマートデバイスにも適用可能であると考えられる．しかし，アプリケーションに関しては，同様のアプリケーションであってもバージョンや設定による差異が考えられる．それらを解明するには，異なるスマートデバイスと，異なるバージョンや設定のアプリケーションの組み合わせによる CSI 特徴量の変化を検証する必要がある．

最後に，複数人のスマートデバイスを対象とした場合の有効性について議論する．複数人が同時に，通信条件の異なるスマートデバイスを操作する場合，単一の Nexmon 適用デバイスで複数の CSI を取得することはできない．そのため，通信条件が異なる場合は，複数の Nexmon 適用デバイスを用いて，複数人のスマートデバイス状態推定を行う必要がある．通信条件が同一であっても，複数人を対象とする場合は，CSI 特徴量を分離する必要がある．

7 おわりに

Nexmon による CSI を用いたスマートデバイス状態推定手法を提案し，基礎評価を行った．8 種類の時系列モデルを試行した結果，学習モデル：TCN によって最大分類精度 87.6%を確認した．また，入力時点数による精度比較を行うことで，スマートデバイス状態推定における最適な入力時点数を明らかにした．さらには，アプリケーションの組み合わせによる精度比較を行うことで，アプリケーションによる分類難易度を検証した．

本研究のスマートデバイス状態推定手法は CSI データを用いるため，通信トラヒックを使用する従来手法 [1] と比較して，取得できる情報が多く，スマートデバイスの状態を詳細に捉えることができる．また，複数の IoT センサとエッジコンピュータを用いる従来手法 [2] と比較して，新たに導入するデバイスが少なく，導入コストが低いといえる．加えて，ドメイン知識に基づいた状態遷移モデルを構築する必要がないため，汎用性が高いといえる．

今後の展望として，スマートデバイス状態推定モデルの精度向上，結果に起因する内外的要因 (個人差や環境，動作アプリケーションの数・種類など) についての検証を進める．また，時系列向けの半教師あり学習 [8] や，データ拡張手法 [9] と組み合わせることで，学習コストが低く，汎用性の高い学習を検討する．最終的には，本研究の成果を活用し，スマートデバイスの動作状態を推定するシステムの実現を目指す．

謝　　辞

本研究の一部は，静岡大学グリーン科学研究所プロジェクト研究支援 (23205) を受けたものである．

文　　献

[1] 祐一服部, 豊荒川, 創造井上. 通信トラヒック分析による複数の iot デバイスにおける機能推定手法の評価. マルチメディア，分散，協調とモバイルシンポジウム 2022 論文集, Vol. 2022, pp. 655–661, 2022.

[2] 田中雅弘, 山内雅明, 大下裕一, 村田正幸, 上田健介, 加藤嘉明. 家庭活動の状況推定を用いたスマートホームネットワークの異常検出手法. IEICE Technical Report; 信学技報, Vol. 119, No. 461, pp. 219–224, 2020.

[3] Francesco Gringoli, Matthias Schulz, Jakob Link, and Matthias Hollick. Free your csi: A channel state information extraction platform for modern wi-fi chipsets. pp. 21–28, 2019.

[4] Defeng David Huang and Khaled Ben Letaief. Carrier frequency offset estimation for ofdm systems using null subcarriers. *IEEE Transactions on Communications*, Vol. 54, pp. 813–823, 2006.

[5] Ronald K. Pearson, Yrjö Neuvo, Jaakko Astola, and Moncef Gabbouj. The class of generalized hampel filters. In *EUSIPCO*, pp. 2501–2505, 2015.

[6] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. *SIGMOD Rec.*, Vol. 45, pp. 69–76, 2016.

[7] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020.

[8] Haoyi Fan, Fengbin Zhang, Ruidong Wang, Xunhua Huang, and Zuoyong Li. Semi-supervised time series classification by temporal relation prediction. In *ICASSP*, pp. 3545–3549, 2021.

[9] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. Data augmentation for time series classification using convolutional neural networks, 2016.

LSiX: ストリームデータに関する複数連続的集約

川上　隼† Bou Savong† 天笠　俊之††

† 筑波大学システム情報工学研究群　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305-8577 茨城県つくば市天王台 1-1-1
E-mail: †s2320583@u.tsukuba.ac.jp, ††{savong-hashimoto,amagasa}@cs.tsukuba.ac.jp

あらまし　ストリームデータ分析の需要が増加するにつれて、ストリーム処理エンジンはいくつかの技術的課題に対処する必要がある。連続的な window の集約はさまざまなアプリケーションでデータ分析のパイプラインの一部に使われており、高い入力レートのストリームデータにおいて大量のクエリを処理するときにスケーラビリティの問題に直面する。この問題に対処するために LSiX(Longest-shortest-window based indexing) というストリームデータに対して複数クエリを効率的に集約するアルゴリズムを提案する。この手法は、複数のクエリのうち最長と最短の window をもとにした 2 つの配列を利用して、全てのクエリを二つの配列から共有した値から計算することにより、各クエリを最大 2 回の集約操作で計算することを可能にした。実験により、最新手法である MCQA を含む比較手法より、3.5 倍以上高速であることが示された。

キーワード　ストリームデータ処理、すべり窓集約

1. はじめに

近年、ネットワーク監視や経済など、さまざまな領域で連続的なデータストリームが普及し、データストリー処理への関心が強くなっており、データストリーム管理システム (DSMS) [1] における効率的でリアリタイムに近い処理の必要性が高まっている。また、時系列データ [3] の sum, max, min, average などの集約計算に対する需要も大幅に増加している。例えば、株式投資家は連続的な株価に頼るだけでなく、彼ら独自のリアルタイムな集約 [10] の結果も必要としている。このような場合、ユーザーはデータストリームにおける最新のデータに関して主に興味がある。

データストリームを集約するには、一般的に sliding window aggregation(SWAG) [2], $Q[w, s]$, を使用する。SWAG は、新しい s 個のレコードが到着するたびに window 内の w 個の最新のレコードを集約する手法である。SWAG には time-based と count-based の 2 種類が存在する。前者は window を時間単位で区切り、後者は window をレコードの数によって区切る。多くのアプリケーションで、DSMS はデータストリームに対して複数のクエリを処理する必要があり、クエリの数は多く、数千となることもある。その上、DSMS はそのようなクエリを遅延時間を少なく処理する必要がある。しかし。このようなデータストリームを従来の処理する場合、スケーラビリティの問題に直面する。

この問題を対処するために SWAG のための効率的なアルゴリズムが必要となる。単一クエリのための SWAG の研究 [4] は多くあるが、複数クエリを処理するとき、クエリの数に比例して slide 毎のレイテンシーが増加する。このオーバーヘッドにより、これらのアルゴリズムは何千ものクエリを同時に処理することは現実的ではない。

MCQA [9] は複数クエリを処理するための最新のアルゴリズムである。しかし、MCQA は window サイズと slide サイズの比が大きい時や異なる window サイズの差が大きい時などで処理時間が大幅に増えてしまうという問題がある。

この問題に対処するために、この論文ではデータストリームに対して、複数クエリの SWAG を効率的に処理するためのアルゴリズムである LSiX(longest-shortest-window-based-indexing) を提案する。LSiX はクエリにおける最長と最短の window サイズを使用して二つの配列 Lo, Sh を作成し、それらの集約結果を全てのクエリに共有することで、安定性を確保し、処理時間を大幅に短縮する。実験から最新の手法である MCQA と比べて少なくても 3. 5 倍速く、複数のクエリを処理できるように拡張された L-BiX [4] よりも少なくても 3 倍速く処理することがわかった。

2. 前提知識

このセクションでは、この後の議論に必要な前提知識を提供する。

2.1 Sliding Window Aggregatin と部分集約

データストリームが与えられた場合、SWAG は window 内のレコードに対する集約を連続的に計算する。クエリは window サイズ、slide サイズ、集約関数によって定義される。

複数の SWAG クエリを同時に処理するときに、全ての window を最初から集約するのは非効率である。第一に同じクエリの slide の前後二つの window は互いに重なる場合があるため、重なった部分は再利用することができる。第二に異なる SWAG クエリによって定義された window も互いに重なる部分があるため、一部の結果を再利用することができる。

一つの手法として window を分割して、部分集約を計算す

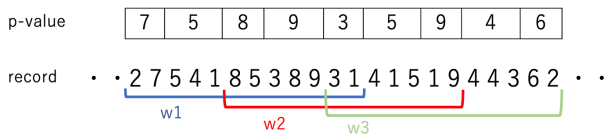


図 1: max を求めるクエリ:$Q[12,5]$ の部分集約の例.

る方法がある。関連するパーティションによって計算された部分集約を使用することで各 SWAG クエリの最終結果を求めることができる。window を分割するための技術には、Panes [8], Pairs [7], Cutty [5] がある。クエリ $Q[w,s]$ が与えられた時、Pairs は $f1 = w\%s, f2 = S - f1$ で求められる 2 つのパラメータを使って window を分割する。例えば、$Q[12,5]$ というクエリで集約関数が max の SWAG を考える（図 1）。部分集約を使わない場合は slide するごとに window 全体を集約するため、11 回の集約操作が必要となる、それに対して Pairs を使った場合、$f1 = 12\%5 = 2, f2 = 5 - 2 = 3$ となり、w1 の集約結果は 9=max(7,5,8,9,3）と求まる。部分集約を使うことで、slide ごとの集約は、f1 と f2 を求めるのに 3 回と最終結果を求めるのに 4 回の計 7 回の集約操作で最終結果を求めることができる。

2.2 集約関数

集約関数 [6] は複数の値を集約することで、生のデータから高水準な情報を抽出するために頻繁に使用される。多くの集約関数が存在するが、それらは以下の 3 つに分類される。

- Distributive: 分散して計算できる集約関数。例）sum, max, count など。その関数で求めた部分集約の値を使って全体の集約を求めることができる。
- Algebraic: 固定された有限引数を持つ代数関数によって計算できる集約関数。例）平均、分散など。平均は distributive な関数である sum を count で割ることで求めることができる。
- Holistic: 計算するための固定された有限引数を持つ代数関数がない集約関数。例）中央値、四分位偏差など。つまり、上記の二つに属さない関数は Holistic な関数となる。

本論文では、Distributive と Algebraic な関数を扱うが、Holistic な関数への対処は今後の課題となる。

3. 関連手法

効率的に SWAG を計算するための手法は多く存在する。まずは最も関連した研究に焦点を当てる。

L-BiX [4] は単一の SWAG クエリを効率的に求めるための手法である。最新のレコードから最古のレコードまでを順に部分集約値を求める後方集約と最古のレコードから最新のレコードまでを順に部分集約値を求める前方集約を結合することで効率的に集約計算をする手法となる。L-BiX は平均して 3 回の集約操作で結果を求めることができる。時間計算量は O(1) となり、現在データストリームに対して単一クエリを処理する最も効率的な手法である。

Two-Stack [12] と FlatFIT [11] も単一の SWAG クエリを処理する手法で、部分集約値とポインタを格納する 2 つの配列を使用し、ポインタを介して必要な部分集約地のインデックスを取得し、最終結果を得る。FlatFIT は window 内の部分集約の数が nc の時、平均して $log(nc)$ 回の集約操作で結果を求めることができる。Two-Stacks は時間計算量が O(1) であるが、L-BiX よりも slide ごとの更新回数が多く、より大きなメモリを必要とする。

一方、複数の SWAG クエリを同時に処理する必要が多く、上記の手法は単一のクエリ処理に最適化されているため、複数のクエリを効率的に処理することができない。MCQA [9] はこの問題に対応した手法となる。MCQA は以下の手順で特定のストリームに対する複数のクエリを同時に集約することができる：(1) 複数のクエリの window のサイズを昇順に並び替える;(2)slide サイズの集約結果を保持する;(3) それを再利用して小さい window サイズのクエリから順に集約を実行する。また、大きな window サイズを持つクエリを集約する際にはそれより小さい window サイズのクエリの結果を保持し再利用することで効率的な集約を行う。

しかし、MCQS の欠点として window サイズと slide サイズによって計算量が変化し、特定のパラメータの組み合わせで実行時間が長くなるという不安定さがある。以下の 3 つの状況で実行時間が長くなる:(1) 連続する 2 つの window サイズの差が大きい,(2) 最小の window サイズと slide サイズの比率が大きい,(3) 全ての window サイズが互いに倍数でない。例は図 2 に示す。レコード:3(最も右のレコード) が到着した時、$w1$ は単純に最新の 5 個のレコードを集約することによって計算される。$w2$ は $w1$ を再利用することで 3 回の集約操作 $(w2[0] = max(w1[0], 4, 9, 6)$ で求めることができ, $w3$ は $w2$ を再利用することで 4 回の集約操作 $(w3 = max(w2[0], 5, 3, 8, 9)$ で求めることができる。しかし、$w2$ と $w3$ の差が大きくなると集約操作の回数も増加する。

n 個の window サイズ:$W_1, W_2, ..., W_n (W_1 < W_2 < ... < W_n)$ と slide サイズ:Sl が与えられた時、MCQA で slide ごとに全ての結果を計算するのに必要な集約操作の回数は $\sum_{i=2}^{n}(\lfloor \frac{W_i}{W_{i-1}} \rfloor + W_i\%W_{i-1}) + \lfloor \frac{W_1}{Sl} \rfloor + W_1\%Sl$ となり、必要なスペースは $\lceil \frac{W_{max}}{Sl} \rceil \times n$ となる。また、単一のクエリを求める最新の手法である L-BiX の計算量は $3n$ でスペースは $\sum_{i=1}^{n}(\frac{W_i}{Sl})$ となる。必要な時間とスペースはテーブル 1 にまとめた。L-BiX の必要な時間は n に依存する。しかし、実際のアプリケーションで n はとても大きく、$\lceil \frac{W_{max}}{W_{min}^2} \rceil$ よりも大きな値となる。そのため、LSiX の必要な時間は L-BiX よりも小さくなる。他にも L-BiX はクエリごとに配列を持つため、LSiX よりも多くのス

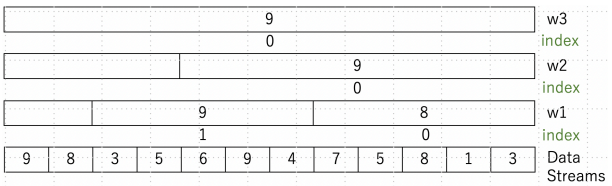


図 2: window サイズが $w1 = 5, w2 = 8, w3 = 12$ で slide サイズが 1 の 3 つのクエリを求める MCQA の例

表 1: 必要な時間とスペース

Algorithm	Time	Space
L-BiX	$3n$	$\sum_{i=1}^{n}(\frac{W_i}{Sl})$
MCQA	$\sum_{i=2}^{n}(\lfloor\frac{W_i}{W_{i-1}}\rfloor + W_i\%W_{i-1}) + \lfloor\frac{W_1}{Sl}\rfloor + W_1\%Sl$	$\lceil\frac{W_{max}}{Sl}\rceil \times n$
LSiX	$2n+3+3\lceil\frac{W_{max}}{W_{min}^2}\rceil$	$W_{max} + \lceil\frac{W_{max}}{W_{min}}\rceil$

ペースを必要とする。MCQA は一つずつ window を check する必要があり、必要な時間は連続する 2 つの window に強く影響され、非効率である。また、必要なスペースもクエリ数に対して過大である。これに対して LSiX は W_{max} をチェックするだけでよく、全てのクエリで他の window を無視できるため、必要な時間は MCQA よりも大幅に短い。同様に LSiX の必要なスペースは MCQA よりも n 倍近く小さい。

4. 提案手法

この章では *Longest-Shortest-window-based Indexing(LSiX)* という複数クエリの SWAG を効率的に処理する手法を提案する。

4.1 LSIX の概要

ユーザーに定義された SWAG が与えられた時、最長の window サイズを W_{max}, 最短の window サイズを W_{min} とする。この時、LSiX はサイズが W_{max} となる Lo と $\lceil\frac{W_{max}}{W_{min}}\rceil$ となる Sh の二つの配列のみを使用して部分集約を保持する。

- “Lo” は $W_{min}(< W_{max})$ 以下で全ての部分集約を保持する。Lo のインデックスが W_{min} を超えるとサイズ W_{min} の新しいサイクルが始まり、現在のサイクルの部分集約が更新されていく。あるクエリの結果を計算するときに、Lo の値はそのクエリの両端の値として使用される。
- “Sh” は Lo における各サイクルの最大値を集約した値を保持する。あるクエリの結果を計算するときに、Sh の値はそのクエリの中央の値として使用される。

LSiX では、Lo と Sh は新しいレコードが来るたびに更新される。そして、各クエリの結果は図 3 のように Lo から両端のために二つの値と Sh から中央のための値を集約することによって求めることができる。これは LSiX が window サイズに関わらず、最大 2 回の集約操作でクエリの結果を求めることができることを意味する。これによって、クエリの結果を計算するのに必要な集約操作の合計回数は MCQA よりも大幅に少なくなる。

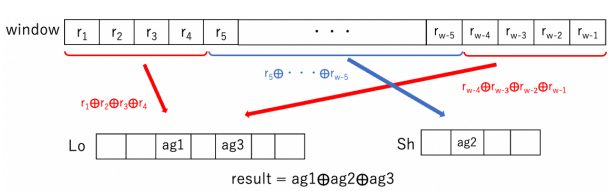


図 3: Lo と Sh を使って、window サイズ:w のクエリを求めるイメージ図。($\oplus$ は集約関数を表す)

簡略化のため、以下の議論では全てのクエリが slide サイズ:1 であるとする。しかし、提案手法では上記とは異なるクエリにも対応しており、これは Section 4.7 で説明する。加えて、LSiX は count-based な window と time-based な window の両方に対応している。

4.2 データ構造

LSiX は二つの索引付き配列を使用し、全てのクエリを計算するための部分集約を効率的に保持する。LSiX の構造は Def. 1 のように定義する。

Definition 1. *n 個のクエリと集約関数 $\oplus$, 最長の window:W_{max}, 最短の window:W_{min} が与えられると LSiX の構造は以下の二つの配列で構成される：*

- *“Lo” of size $b = W_{max}$*
- *“Sh” of size $c = \lceil\frac{W_{max}}{W_{min}}\rceil$*

window サイズが $W_{min} = w1 = 5$, $w2 = 8$, $W_{max} = w3 = 12$ で max を求める 3 つのクエリを集約する LSiX の構造を例とする。この時、Lo のサイズは 12、Sh は $\lceil\frac{12}{5}\rceil = 3$ となる。

4.3 *Lo* の集約

まず、Lo の集約を説明する前に重要な概念を説明する。

変数 p p は Lo の index を表す変数。p は新しいレコードが到着するたびに 1 ずつ増えていき、最後に 0 に戻る。($Lo[W_{max}-1]$ など)

サイクル Lo は複数のサイクルで構成され、各サイクルは Def. 2 で定義されたサイズ:W_{min} の部分配列である。

Definition 2. *$Lo[0]$ から順に W_{min} ずつサイクルに分けられる（最後のサイクルのみ M_{min} ではない可能性がある）。Lo に含まれる各要素はサイクル内での位置付けによって、以下の 3 つに分類される。*

- *st:* サイクルの始まりとなる要素

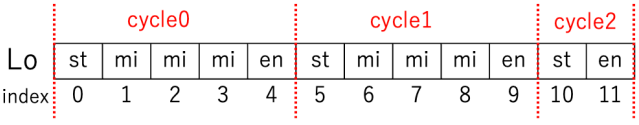


図 4: window サイズが $w1 = 5, w2 = 8, w3 = 12$ の 3 つのクエリを求める Lo の例

- *en:* サイクルの終わりとなる要素
- *mi: st,en* 以外の要素

図 4 は window サイズが $W_{min} = w1 = 5$, $w2 = 8$, $W_{max} = w3 = 12$ の 3 つのクエリを集約する Lo の例である。この例では $W_{min} = 5$ のため、$Lo[0] \sim Lo[4]$, $Lo[5] \sim Lo[9]$, $Lo[10] \sim Lo[11]$ の 3 つのサイクルで構成される。また、Lo は p の位置によって以下のように集約される。ただし、R_{new} は最新のレコード, R_{pre} は一つ前のレコードを指す。

(Rule 1) $Lo[p]$ が st の時、

$Lo[p] = R_{new}$.

(Rule 2) $Lo[p]$ が mi の時、

$Lo[p] = R_{new} \oplus Lo[p-1]$;

$Lo[p-1] = R_{pre}$.

(Rule 3) $Lo[p]$ が en の時、

$Lo[p] = R_{new}$;

$Lo[p-1] \oplus = Lo[p]$;

$Lo[p-2] \oplus = Lo[p-1]$;

$\vdots$

$Lo[p - W_{min} + 1] \oplus = Lo[p - W_{min} + 2]$.

Lo は新しいレコードが到着し、p の値が更新されるたびに上記のような集約を行うことで、部分集約を保持していく。

4.4 *Sh* の集約

Sh は各サイクルの st の値を集約した値が格納される配列である。Sh の index は Lo のサイクルと左から順に 1 対 1 で対応している。例として図 4 を用いると、cycle0 は $Sh[0]$ に、cycle1 は $Sh[1]$ に、cycle2 は $Sh[2]$ に対応している。Sh では p が en に到達したときにのみ、以下のような集約が実行される。(式中の k は、k=$\lfloor p/W_{min} \rfloor$ で、p のあるサイクルと対応した Sh の index を示す)

(Rule 4) $Sh[k] = Lo[k * W_{min}]$

$Sh[(k-1)/c] = Sh[k] \oplus Lo[((k-1)/c) * W_{min}]$

$\vdots$

$Sh[(k-c+1)/c] = Sh[(k-c+2)/c] \oplus Lo[((k-c+1)/c) * W_{min}]$

Rule4 は p のあるサイクルと対応した $Sh[k]$ から順に Sh 全体を更新する。結果として、$Sh[(k-c+1)/c]$ には、W_{max} のクエリを集約した値が格納される。Lo と Sh の集約の詳細は Algorithm 1 のようになる。

LSiX では、Lo と Sh から適切な値を使用することで最大 2 回の集約操作で各クエリの結果を求めることができる。

4.5 集約の例

Lo と Sh で集約を行う例として、window サイズが $W_{min} = w1 = 5$, $w2 = 8$, $W_{max} = w3 = 12$ で max を求める 3 つのクエリを処理する。この例は図 5 で示す。

step 0 (p=11) である初期状態では、Lo の各サイクルで Rule 3 が行われ、Sh では Rule4 が行われた結果が格納されている。step1 では新しいレコード (3) が到着し、p=0 となる。$Lo[0]$ は st に分類されるため Rule1 が実行される

Algorithm 1 Aggregation

INPUT: n queries with $\oplus$ aggregation and slide size Sl, W: all windows, W_{max}: longest window, W_{min}: shortest window, $b = \lceil \frac{W_{max}}{Sl} \rceil$, $s = \lceil \frac{W_{min}}{Sl} \rceil$, Lo and Sh created by Def. 1, and data stream DS

OUTPUT: result

```
1: p=0, A^p_slide=""
2: for A_slide from DS do
3:     if p%s== s-1 or p == b-1 then
4:         agg = A_slide ⊕ A^p_slide
5:         Lo[p] = A_slide, i = 0
6:         for i = p-1; i%s!=0 do
7:             Lo[i] = agg
8:             agg = agg ⊕ Lo[i-1]
9:             i–
10:        end for
11:        l = ⌊ p/s ⌋
12:        Lo[i] = agg, Sh[l] = agg
13:        if W_max%W_min == 0 then
14:            len = ⌊ W_max/W_min ⌋
15:        else
16:            len = ⌊ W_max/W_min ⌋ + 1
17:        end if
18:        while j = 0; j <len-1 do
19:            l = (l+len-1)%len
20:            agg = agg ⊕ Lo[l*s]
21:            Sh[l] = agg
22:            j++
23:        end while
24:    else if p%s = 0 then
25:        Lo[p] = A_slide
26:    else
27:        Lo[p] = Lo[p-1] ⊕ A_slide
28:        Lo[p-1] = A^p_slide
29:    end if
30:    A^p_slide = A_slide
31:    result = CREATE_RESULT(p, Sl, W, b, s, Lo, Sh)
32:    p++
33:    p=0 if p==b
34: end for
```

($Lo[0]$=3)。次に step2 では新しいレコード (1) が到着し、p=1 となる。$Lo[1]$ は mi に分類されるため Rule2 が実行される ($Lo[1] = max(Lo[0], 1) = 3, Lo[0] = 3$)。step3(p=2) と 4(p=3) の時は、mi に分類されるため、Rule 2 が行われた結果が格納される。step5 では新しいレコード (4) が到着し、p=4 となる。$Lo[4]$ は en に分類されるため Rule3 と Rule4 が実行される。

Rule3 では $Lo[4]$ の所属する赤色のサイクルが全て更新され、

$Lo[4] = 4$,

$Lo[3] = max(Lo[4], Lo[3]) = max(4, 6) = 6$,

$Lo[2] = max(Lo[3], Lo[2]) = max(6, 5) = 6$,

$Lo[1] = max(Lo[2], Lo[1]) = max(6, 1) = 6$,

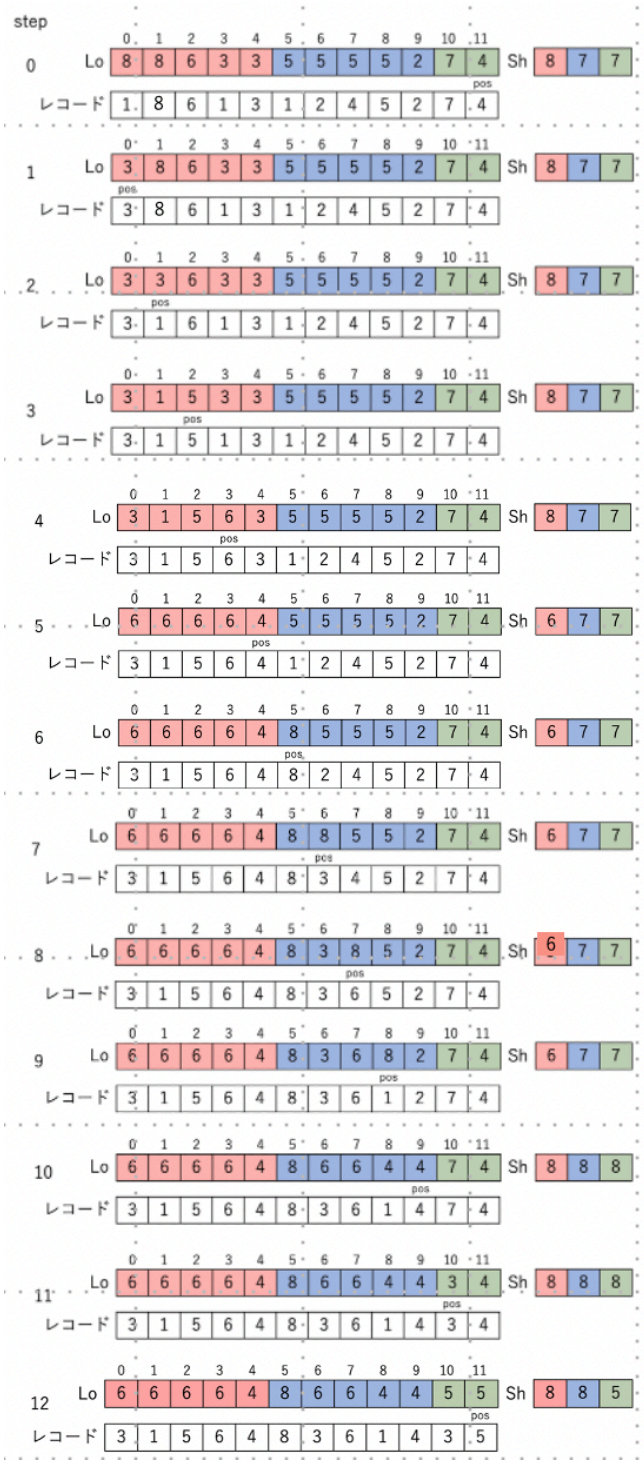


図 5: w1=5,w2=8,w3=12 のクエリを集約する Lo と Sh の例.

$Lo[0] = max(Lo[1], Lo[0]) = max(6, 3) = 6$ となる。その後、赤色のサイクルと対応する $Sh[\lfloor 4/5 \rfloor] = Sh[0]$ から Rule4 が実行され、$Sh[0] = Lo[0] = 6$,
$Sh[2] = max(Sh[0], Lo[10]) = max(6, 7) = 7$,
$Sh[3] = max(Sh[2], Lo[5]) = max(7, 5) = 7$ となる。これ以降でも、新しいレコードが到着したら p が 1 追加され、適切な処理が実行されるというのを繰り返す。

LSiX では、Lo と Sh の部分集約を使用することで window サイズが $W_{min} \sim W_{max}$ のクエリの結果を求めることができる。例えば step1 の時、Q[12,1] の結果は $max(Lo[0], max(Sh[1], Lo[1])) = 8$ と 2 回の集約操作で求めることができる。また、step9 の時、Q[5,1] の結果は $Lo[5] = 8$, Q[8,1] は $max(Lo[2], Sh[1]) = 8$ と 2 回以下の集約操作で求めることができる。このように LSiX では、Lo と Sh の最適な値を使うことで、各クエリの結果を 2 回以下の集約操作で求めることができる。クエリ処理の詳細は Algorithm 2 のようになる。

Algorithm 2 Function that return the results of each query

```
1: function CREATE_RESULT(n, p, Sl, W, b, s, Lo, Sh)
2:     if ⌊ qe/s ⌋ b%s = 0 then
3:         D = b
4:     else
5:         D = (⌊b/s⌋ + 1) * s
6:     end if
7:     for i = 0 ∼ n do
8:         qe = (p + b + 1 − ⌈W[i]/Sl⌉)%b
9:         x = ⌊ [(qe+s)%D]/s ⌋
10:        y = ⌊ (p%D)/s ⌋
11:        if (p%s = s-1 or p = b-1) and qe%s = 0 then
12:            result[i] = Sh[⌊ qe/s ⌋]
13:        else if p%s = s-1 or p = b-1 then
14:            result[i] = Sh[x] ⊕ Lo[qe]
15:        else if qe%s = 0 then
16:            result[i] = Sh[⌊ qe/s ⌋] ⊕ Lo[p]
17:        else if x = y then
18:            result[i] = Lo[qe] ⊕ Lo[p]
19:        else
20:            result[i] = Lo[p] ⊕ Lo[qe] ⊕ Sh[x]
21:        end if
22:    end for
23:    return result
24: end function
```

4.6 必要な計算量とメモリ

LSiX に必要な時間とスペースについて詳しく説明する。概要はテーブル 1 の通りである。

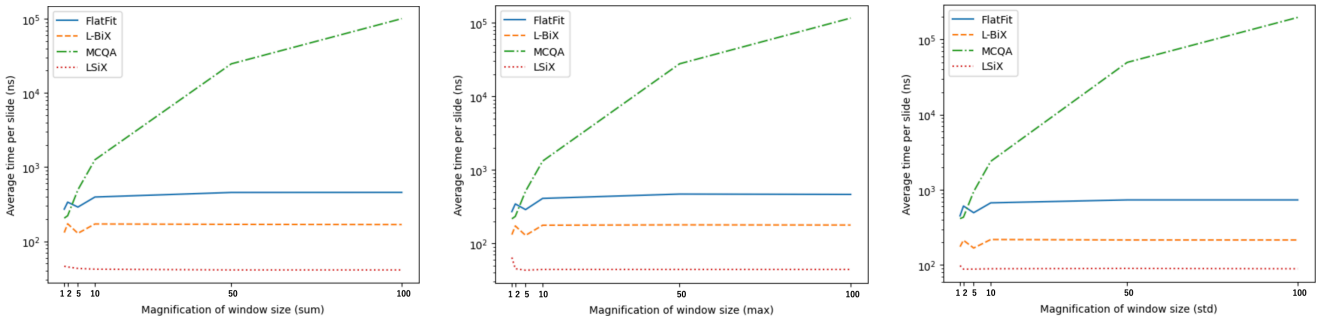
最長と最短の window サイズが W_{max} と W_{min} の n 個のクエリが与えられた時、LSiX のコストは以下のようになる。

- Time: $2n + 3 + 3\lceil \frac{W_{max}}{W^2_{min}} \rceil$
- Space: $W_{max} + \lceil \frac{W_{max}}{W_{min}} \rceil$.

Proof. W_{max} 回の間に Lo と Sh の更新と全てのクエリを計算するのに必要な集約操作は以下のように求まる。

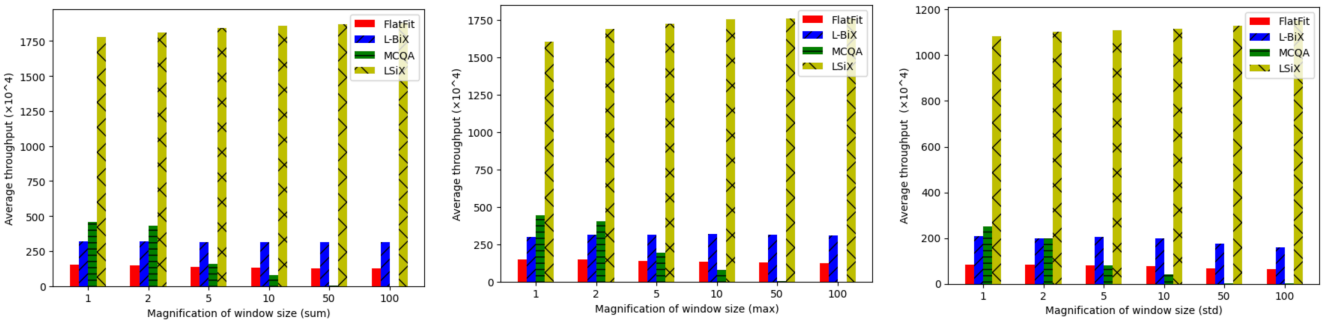
- 各クエリの結果を計算するには最大で 2 回の集約操作が必要となる。そのため、W_{max} 回の間に n 個のクエリに必要な集約操作の合計回数は、$2nW_{max}$ となる
- Lo での後方集約の回数は、$2W_{max}$
- Lo での前方集約の回数は、W_{max}
- Sh での後方集約の回数は、$3\lceil \frac{W_{max}}{W_{min}} \rceil^2$

それゆえ、集約操作の合計は、$2nW_{max} + 3W_{max} + 3\lceil \frac{W_{max}}{W_{min}} \rceil^2$。これを W_{max} で割り平均を求めると、



(a) sum (b) max (c) std

図 6: count-based window に対する window サイズを変化させた実行時間の結果



(a) sum (b) max (c) std

図 7: count-based window に対する window サイズを変化させた処理能力の結果

$\frac{2nW_{max}+3W_{max}+3\lceil\frac{W_{max}}{W_{min}}\rceil^2}{W_{max}} = 2n+3+3\lceil\frac{W_{max}}{W_{min}^2}\rceil$ となり、計算量が求まる。

LSiX はサイズが W_{max} の Lo とサイズが $\lceil\frac{W_{max}}{W_{min}}\rceil$ の Sh の二つの配列のみを必要とする。そのため、$W_{max}+\lceil\frac{W_{max}}{W_{min}}\rceil$ とスペースが求まる。

□

5.7 異なる slide サイズを持つ windows

もし、slide サイズで割り切れない window サイズを持つクエリや異なる slide サイズのクエリがある時、slide サイズを 2 つ以上に分割し、同じ部分のレコードが同時に期限切れになるようにする必要がある。これにより、期限切れのレコードを効率的にパージすることができる。

提案する方法は全ての window と slide の最大公約数 ($gcd = GCD(Windows, Slides)$) を利用することだ。そして、$gcd$ は W_{max} と W_{min} を分割するために使用される。Lo のサイズは $b = \lceil\frac{W_{max}}{gcd}\rceil$、$Sh$ のサイズは $c = \lceil\frac{W_{max}}{W_{min}}\rceil$ となる。Lo と Sh を更新する回数は、レコードが到着するたびにではなく、gcd 個のレコードを集約した値が来るたびに更新されるため、減らされる。そして、Algorithm 1 と同じ手順でデータストリームに対するクエリ処理が行われる。

5. 実　験

5.1 実験環境

提案手法の有効性を比較するために 1slide あたりの平均処理時間で比較した。比較には、複数クエリを処理する最先端の手法である MCQA と単一のクエリを処理する最先端の手法である L-BiX と FlatFit を用いた。

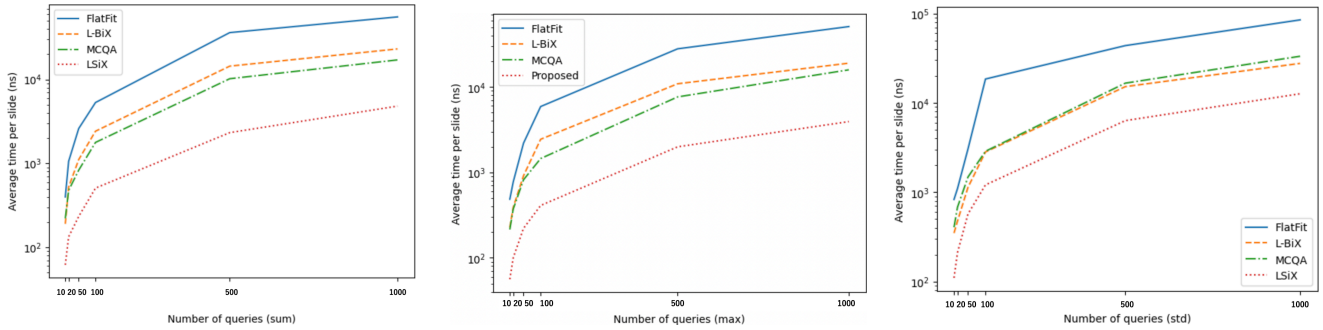
検証には、DEBS データセットを使用した。DEBS12Grand Challenge Datatset (DEBS) は同様の研究フレームワークでクエリ処理負荷評価に広く使用されている標準データセットである。このデータセットは工場内のさまざまな大規模センサーによって生成された記録から構成されている。レコードは 51 次元となっていて、実験にはそのうちに 1 次元を採用した。データセットは約 3200 万レコードを含む。

クエリ処理の負荷は SWAG のパフォーマンスに影響を与える。それゆえ、window サイズを変化させた実験とクエリの数を変化させた実験を行った。LSiX は time-based window と count-based window を共に処理できるが、実験では count-based window のみを使用した。実験の結果はそれぞれのアルゴリズムの 5 回の独立した実行の平均とする。

集約関数としては、Distributive 関数である sum, max と algebraic 関数である標準偏差 (std) を使用した。すべての実験は Apple M1 Pro、16GB のマシンで行った。

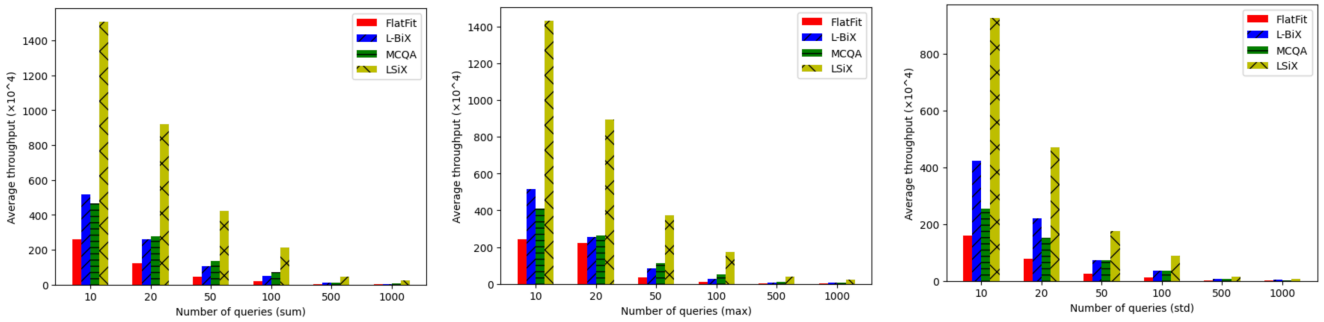
5.2　window サイズを変化させた実験

この実験では slide サイズを 10 に固定して、window サイズを変化させて実験を行った。window サイズが W=[35, 65, 75, 100, 110, 160, 205, 220, 235, 280] となる 10 個のクエリを 1,2,5,10,50,100 でスケールさせた。実験の結果は図 6, 7 に示す。



(a) sum (b) max (c) std

図 8: count-based window に対するクエリの数を変化させた実行時間の結果



(a) sum (b) max (c) std

図 9: count-based window に対するクエリの数を変化させた実行時間の結果

一般的に、処理能力は window サイズが増加するにつれて減少する傾向にある。加えて第 3 章で説明したように、window サイズと slide サイズの比が大きくなったり、連続する二つの window サイズが大きくなるにつれて MCQA の処理能力は大幅に減少する。また、window サイズと共に後方集約の回数が増えるため、L-BiX と FlatFIT の処理能力も軽く減少する。それに対して、LSiX の処理能力は window サイズが増加するにつれて増加している。これは、window サイズが大きくなると Sh を更新する回数である $\lceil \frac{W_{max}}{W^2_{min}} \rceil$ が減少するためである。これにより、LSiX と他の手法の処理時間の差は window サイズが増えるにつれて顕著になる。

具体的に比べると、window サイズを 1 倍しているときに、LSiX は MCQA よりは 4 約倍速く、L-BiX よりは約 5 倍速く、FlatFIT よりは約 9 倍速い。次に window サイズを 100 倍しているときに、LSiX は MCQA よりは 1000 約倍速く、L-BiX よりは約 6 倍速く、FlatFIT よりは約 14 倍速い。

5.3 クエリの数を変化させた実験

この実験では slide サイズを 10 に固定して、クエリの数を n=10,20,50,100,500,1000 と変化させた。window サイズは次のように決められる。

$$w_1 = 10$$
$$w_i = w_{i-1} + a \qquad l$$
(a は $[10, 15, 20, 25, 30, 35, 40, 45, 50]$ の中のランダムな値)

実験の結果は図 8, 9 に示す。

全ての手法でクエリの数が増加するにつれて処理時間は増加し、処理能力は減少している。しかし、LSiX はクエリの数に依存する集約操作と配列操作の回数が他の手法よりも少ないため、最良の結果となっている。

具体的に比べると、n=10 のときに、LSiX は MCQA よりは 3.5 約倍速く、L-BiX よりは約 3 倍速く、FlatFIT よりは約 8 倍速い。次に n=1000 のときに、LSiX は MCQA よりは 4 約倍速く、L-BiX よりは約 5 倍速く、FlatFIT よりは約 11 倍速い

6. 結　論

本稿では、単一のデータストリームに対して、複数クエリを効率的に処理するための手法を提案した。提案手法は最長と最短の window サイズをもとにした二つの配列 Lo,Sh で集約を行い、その結果を全てのクエリで共有することで効率的に求める。配列 Lo,Sh は前方集約と後方集約の二つを組み合わせることで、最小限の集約操作の回数で範囲内の部分集約を保持する。結果として、クエリの数に依存するワークロードを少なくし、全てのクエリを高速に求めることができる。

実験の結果、提案手法が最新の手法である MCQA と L-BiX, FlatFIT にスループットで大幅に上回ることがわかった。提案手法は少なくても MCQA より 3.5 倍早く、L-BiX よりすくな 3 倍早く、FlatFit より 8 倍早く処理する。

将来の展望としては、(1)LSiX の計算量が W_max に依存しないように拡張する、(2) 順不同のデータストリームに対して、複数クエリを処理できるように拡張する、の二つがある。

7. 謝　辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の 「ポスト 5G 情報通信システム基盤強化研究開発事業」（JPNP20017）の委託事業，JST CREST（JPMJCR22M2），科学研究費補助金（JP22H03694, JP23K16888）に支援によるものです.

文　献

[1] Daniel J. Abadi, Donald Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stanley B. Zdonik. Aurora: a new model and architecture for data stream management. *VLDB J.*, 12(2):120–139, 2003.

[2] Arvind Arasu, Shivnath Babu, and Jennifer Widom. The CQL continuous query language: semantic foundations and query execution. *VLDB J.*, 15(2):121–142, 2006.

[3] Savong Bou, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Intrans: Fast incremental transformer for time series data prediction. In Christine Strauss, Alfredo Cuzzocrea, Gabriele Kotsis, A Min Tjoa, and Ismail Khalil, editors, *Database and Expert Systems Applications - 33rd International Conference, DEXA 2022, Vienna, Austria, August 22-24, 2022, Proceedings, Part II*, volume 13427 of *Lecture Notes in Computer Science*, pages 47–61. Springer, 2022.

[4] Savong Bou, Hiroyuki Kitagawa, and Toshiyuki Amagasa. L-bix: incremental sliding-window aggregation over data streams using linear bidirectional aggregating indexes. *Knowl. Inf. Syst.*, 62(8):3107–3131, 2020.

[5] Paris Carbone, Jonas Traub, Asterios Katsifodimos, Seif Haridi, and Volker Markl. Cutty: Aggregate sharing for user-defined windows. In Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi, editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1201–1210. ACM, 2016.

[6] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Min. Knowl. Discov.*, 1(1):29–53, 1997.

[7] Sailesh Krishnamurthy, Chung Wu, and Michael J. Franklin. On-the-fly sharing for streamed aggregation. In Surajit Chaudhuri, Vagelis Hristidis, and Neoklis Polyzotis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, pages 623–634. ACM, 2006.

[8] Jin Li, David Maier, Kristin Tufte, Vassilis Papadimos, and Peter A. Tucker. No pane, no gain: efficient evaluation of sliding-window aggregates over data streams. *SIGMOD Rec.*, 34(1):39–44, 2005.

[9] Wen Liu, Tuqian Zhang, and Junxia Liu. Window-based multiple continuous query algorithm for data streams. *J. Supercomput.*, 75(9):5782–5807, 2019.

[10] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. Multi-source aggregated classification for stock price movement prediction. *Inf. Fusion*, 91:515–528, 2023.

[11] Anatoli U. Shein, Panos K. Chrysanthis, and Alexandros Labrinidis. Flatfit: Accelerated incremental sliding-window aggregation for real-time analytics. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 5:1–5:12. ACM, 2017.

[12] Kanat Tangwongsan, Martin Hirzel, and Scott Schneider. Low-latency sliding-window aggregation in worst-case constant time. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017*, pages 66–77. ACM, 2017.

大規模データストリームに対する高速なS-FINCHクラスタリング

牛尼　索造†　　藤原　靖宏††　　塩川　浩昭†††

† 筑波大学 情報学群 情報科学類　〒 305–8577 茨城県つくば市天王台 1-1-1
†† NTT コミュニケーション科学基礎研究所　〒 243-0198 神奈川県厚木市森の里若宮 3-1
††† 筑波大学 計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1-1-1
E-mail: †sushiama@kde.cs.tsukuba.ac.jp, ††yasuhiro.fujiwara@ntt.com, †††shiokawa@cs.tsukuba.ac.jp

あらまし　クラスタリングはデータ集合に内在する性質の類似した部分集合を検出する手法であり，データ分析において重要な要素技術となっている．その中でも近年注目を集めているのは，データストリームに対して凝集型階層的クラスタリングを行う S-FINCH である．S-FINCH は時々刻々と到来するデータストリームに対して各データオブジェクトの最近傍や共有最近傍を求める必要があるため，各データオブジェクト間の距離計算が必要となる．そのため，S-FINCH はクラスタリングにおいて膨大な計算時間を必要とする．そこで本稿では S-FINCH の高速化手法を提案する．提案手法は空間索引を利用することで，S-FINCH において生じる距離計算回数を削減する．本稿では実データを用いた評価実験により，提案手法は S-FINCH に対してクラスタリング精度を損なわずに効率的なクラスタリング処理が実行可能であることを示した．

キーワード　ストリームデータ処理，空間・時空間・時系列データ処理，データ構造・索引

1　はじめに

多次元データストリームに対するクラスタリングとは膨大なデータから性質の類似するデータごとのグループ分割を提供する手法である．多くの科学の分野において，膨大なデータから性質の類似するデータのグループを見つけだしラベル付けをすることは非常に重要になっている．特に近年では，取り扱われるデータのサンプル数や次元数がますます増加し，このようなデータに対する高速な解析処理技術への需要が高まっている．膨大なデータから性質の類似するデータのグループ分割を提供する手法のひとつとして，クラスタリングが挙げられる．

近年 Sarfraz らによって FINCH [1] というクラスタリング手法が提案された．FINCH は多次元のデータポイントの集合を入力として受け取り，その各データポイントの間の距離を計算し，最も近くにあるデータポイントとの間，または同じデータポイントを最も近いとするデータポイント間にエッジを張ることで得られるグラフの各連結成分を 1 つのグループとする．この各グループ内の各データポイントの平均を代表点とし，次の入力として同じ処理を繰り返す．これにより，異なる粒度のグループ分けを持つクラスタ階層を出力する．FINCH は従来の k-means 法 [2] などのクラスタリング手法と比較して，1.) 高い精度で真のクラスタを自動的に決定できる，2.) ハイパーパラメータを必要としない，3.) 異なるドメインのデータでも一般的に利用できる，4.) 膨大なデータセットに拡張できるなどのメリットがあり，高次元のデータセットに対してもハイパーパラメータを必要とせずに高速なクラスタリングが実現可能であることから，FINCH は画像や化学などの幅広い分野での応用 [3–5] に挑戦されている．

しかし，FINCH は静的なデータを対象としたアルゴリズムであるため，データストリームを処理しようとした場合，データが追加されるとデータ全体に対して最近傍探索を行いクラスタ構造を再構築する必要がある．そのため，FINCH はデータストリームのクラスタリングに膨大な処理時間を必要とする．この問題に対処するために Cunningham によって S-FINCH [6] が提案された．S-FINCH は，データが追加された時にクラスタ構造を全て再構築するのではなく，追加されたデータによって変更が生じた部分のみを対象として再計算することによって処理時間を短縮する．

しかしながら，S-FINCH は FINCH に比べて高速なストリーム処理が可能ではあるものの，大規模なストリームデータを対象としたクラスリングは困難である．S-FINCH では，追加されたデータと全ての頂点間に対して距離を計算する．そのため，n 件目のデータが追加された時に，距離計算のみで $\mathcal{O}(n)$ の時間計算量が生じる．また，S-FINCH は階層的クラスタリング手法であるため，データの追加によって他の階層のクラスタ構造も変化する．この階層は高々 $\log_2 n$ であることから，この処理は $\mathcal{O}(n \log n)$ の時間計算量を要する．N 件のストリームデータを処理するのに $\mathcal{O}\left(N^2 \log N\right)$ を要する．したがって，大規模なストリームデータに対しては膨大な処理時間を要することになる．

1.1　本研究の貢献

本稿では S-FINCH の高速化手法を提案する．提案手法では従来手法 S-FINCH の計算コストを削減するために，FINCH が提供するクラスタ構造が最近傍連結成分という小規模な集合によって構成されていることに着目する．また，FINCH で生成されたクラスタは他の最近傍クラスタを持ち最近傍関係にある連結成分が上の階層でのクラスタとなるため，各クラスタの関係を木構造として表現出来る．最近傍関係にある小規模な集

合を，特別に構築することなく得られるため，この集合による空間を空間索引として活用することで，空間索引の構築コストを削減できると考えられる．

そこで本稿では各クラスタに対し各データを包含する空間を求め，最近傍や逆最近傍を求めたいデータポイントとその空間との最短距離を各クラスタの各データの点との距離の下限とし，その下限を利用して最近傍・逆最近傍にならないデータポイントを特定することで距離計算回数を削減する．その結果として提案手法は以下の特性を示す．

計算回数

S-FINCH に比べ少ない計算回数で最近傍および被最近傍探索を行うことができる．

正確性

提案手法で用いるアプローチは，厳密な最近傍を求めるため FINCH が提供するクラスタ構造から精度を損わない．

本稿の構成は，次の通りである．2 節で本稿の前提となる知識について概説する．3 節にて提案手法の詳細について説明し，4 節において提案手法の評価と分析を行う．5 節にて，本稿をまとめ，今後の課題について論ずる．

2 事 前 準 備

この節では FINCH [1] とストリーム処理を可能にした S-FINCH [6] について説明する．表 1 に主な記号とその定義を示す．

FINCH [7] は，d 次元のデータポイントが N 件ある多次元データセット $\boldsymbol{S} \in \mathbb{R}^{N\times d}$ を入力として受け取る．これに対して，S-FINCH は入力として d 次元のデータポイントが N 件連続している多次元データストリームを $\boldsymbol{L} = [L_1, L_2, \ldots, L_N]$（ただし，$L_i \in \mathbb{R}^d$）を受け取る．距離尺度として，FINCH および S-FINCH はユークリッド距離を用いる．すなわち，2 つのデータポイントのユークリッド距離が小さいとき，それらは類似したデータポイントであることを意味する．多次元データポイント v, w があるとき，これらの距離を $dist(v, w)$ を表記する．加えて，多次元空間 X があるとき，v から X までの距離の最小値を $min_dist(v, X)$ を表記する．データポイント x の最近傍とは，x に最も近いデータポイント y のことを指し $NN(x) = y$ と表記するこれに対して，データポイント x の被最近傍とは，x が最近傍であるデータポイント集合 $NNset(x)$ のことを指し $NNset(x) = \{y \mid NN(y) = x\}$ 定義する．

FINCH [7] および FINCH はクラスタリング結果として階層的なクラスタ構造を出力する．そしてこのクラスタ構造の階層の数，つまり階層の高さを H とする．本論文ではこのクラスタ構造の一番下の階層を高さ 1 と，一番上の階層を高さ H とする．それぞれの手法に入力として与えられたデータの各データポイントはまず高さ 1 の階層でクラスタを構成することになる．ここで入力として与えられた各データポイントは 1 つのデータポイントからなるクラスタとして扱う．これにより，高さ 1 の階層ではクラスタ構造の全てのデータポイントは全てクラスタの代表点となる．最上層ではない階層においてデータポイント x がクラスタ X_3 に属するとき，クラスタ X_3 のことを x の親と呼び，$parent(x) = X_3$ と表記する．また，x はクラスタ X_3 の子となり，$x \in children(X_3)$ と表記する．データポイント x の子となるデータポイント $children(x)$ や，さらに $children(x)$ の子となるデータポイント $children(children(x))$ 全てを子孫データポイントと呼ぶ．つまり，データポイント x の子孫となるデータポイントの集合を $descendants(x)$ とするとき，$children(x) \in descendants(descendants)$ であり，さらに $descendants(children(x)) \in descendants(x)$ となる．データポイント x の親となるデータポイント $children(x)$ や，さらに $children(x)$ の親となるデータポイント $children(children(x))$ 全てを祖先データポイントと呼ぶ．つまり，データポイント x の祖先となるデータポイントの集合を $ancestor(x)$ とするとき，$children(x) \in ancestor(ancestor)$ であり，さらに $ancestor(children(x)) \in ancestor(x)$ となる．

表 1: 記号と定義

記号	定義
N	データポイントの数
n	あるステップに与えられたデータポイントの数
d	データポイント x の次元数
κ_i^1	データポイント i の 最近傍データポイント
$L(x)$	データポイント x のクラスタラベル
$parent(x)$	高さ i のデータポイント x のクラスタの 高さ $i+1$ での代表データポイント
$children(x)$	$\{y \mid parent(y) = x\}$
$dist(v, w)$	データポイント v, w の距離
$min_dist(v, W)$	データポイント v と空間 w の距離の最小値
$NN(x)$	x の最近傍データポイント
$NNset(x)$	x が最近傍であるデータポイント
$V_{h,sh,x}$	高さ h のデータポイント x の子孫のうち， 高さ sh にあるデータポイントを包含する空間
$W_{h,x}$	高さ h のデータポイント x を中心とし， x の最近傍までの距離を半径とする超球
$S_{h,sh,x}$	高さ h のデータポイント x の子孫のうち， 高さ sh にあるデータポイント y の $W_{h,y}$ を包含する空間

2.1 FINCH

FINCH は， Sarfraz らによって提案された出力するクラスタ数やハイパーパラメータの指定を必要とせず，データポイントの近傍関係を利用して凝集型階層的クラスタリングを行う手法である．FINCH は与えられたデータセットからグラフを生成し，各連結成分を 1 つのクラスタとすることを繰り返しクラスタ階層を生成する．生成されたクラスタの数が 1 ではないとき，クラスタの構成要素となる各データポイントの平均，つまり重心を代表点とする．この各クラスタの代表点の集合を次の処理でのデータセットとする．与えられたデータセットをクラスタリングすることを 1 つのステップとし，ステップを繰り返すことで階層的なクラスタ構造を生成する．

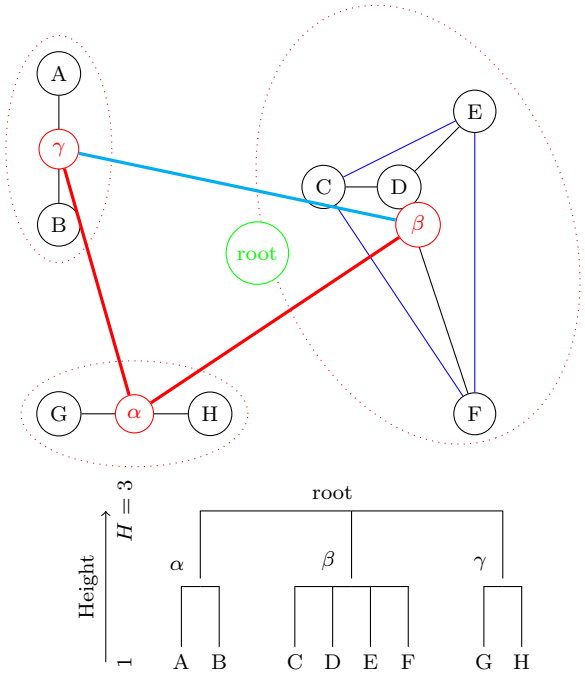


図 1: FINCH でクラスタを生成するときのグラフと対応するデンドログラム

次に連結成分を求めるために生成するグラフについて説明する．データセットの各データポイント間の距離を計算し，最近傍を求め，式 (1) で求まる隣接行列からグラフを生成する．

$$A(i,j) = \begin{cases} 1 \ if \ j = NN(i) \quad or \ j = NN(j) \\ \qquad\qquad\qquad\quad or \ NN(i) = NN(j) \\ 0 \ Otherwise \end{cases} \tag{1}$$

入力例とグラフの例を図 1 に示す．A から H は FINCH に入力として与えられる各データポイント，α から β は次のステップにおいて入力として与えられる各クラスタの構成要素の平均データポイントである．C, E, F を互いに結ぶ青い線，及び α, β を結ぶ水色の線は最近傍関係にはないものの同じデータポイントを最近傍に持つデータポイントの間に張られるエッジを示している．これは式 (1) において $NN(i) = NN(j)$ となる場合に該当する．

FINCH の時間計算量を解析するために，各ステップでの時間計算量を求める．あるステップの入力として n 件のデータポイントがある場合，最近傍探索に $\mathcal{O}(n^2 d)$，連結成分の探索に $\mathcal{O}(n)$ の時間計算量が必要となり，1 ステップに $\mathcal{O}(n^2 d)$ の時間計算量を要する．各データポイントはそれぞれの最近傍と同じの連結成分に含まれ 1 つのクラスタとなるため，ステップごとに入力となるデータセットのサイズは半分以下になる．そのため，FINCH に与えられた N 件の入力データセットに対して階層の数は高々 $\lceil \log_2 N \rceil$ となり，全体としての時間計算量は $\mathcal{O}(N^2 d \log N)$ となる．

2.2 S-FINCH

FINCH における逐次的なデータ更新を行うストリーム処理を達成することを目的に，近年 Cunningham によって S-FINCH が提案された．FINCH はデータの追加に対応できるクラスタリング手法ではないため FINCH を愚直にストリーム処理に対応させると，データが追加される度に過去のデータも含め全体のデータセットに対して FINCH を実行する必要があり $\mathcal{O}(N^3 d \log N)$ の時間計算量を要する．そこで S-FINCH では，データが追加される前のクラスタ構造を用いることで時間計算量を大きく改善する．

S-FINCH は 3 つのステージで動作する．新しいデータポイント x が追加されたとする．x の最近傍データポイントを q, x が追加されたことにより x が最近傍になったデータポイント，つまり x にとっての逆最近傍データポイント集合を S とする．ステージ 1 で x に対する，q, S を求める．q, S に合わせてグラフも更新する．ステージ 2 で x のクラスタラベル L_x を更新する．ステージ 3 で連結成分になんらかの更新があったクラスタの代表点を更新し，次の階層のクラスタの更新を行う．この 3 つのステージを下の階層から順に実行することで新しく追加されたデータポイントも含めた FINCH と同じクラスタ階層が得られる．

ここで，ステージ 1 では x と直前までに追加された各データポイントとの距離を計算する必要があるため，x が追加される前までの各データポイントの個数を $n-1$ とすると $\mathcal{O}(nd)$ の時間計算量を要する．ステージ 2 では q, S の最近傍の更新により連結成分が更新されたクラスタを高々$n-1$ 個の探索で求めることができるため，$\mathcal{O}(n)$ の時間計算量を要する．ステージ 3 ではステージ 2 で探索したクラスタのラベルを保持しておくことで解決される．また前述の通り階層構造の高さは高々 $\log_2 n$ なので，1 つのデータ追加に $\mathcal{O}(nd \log n)$ の時間計算量を要する．この手順を図 2 に示した．これにより S-FINCH は $\mathcal{O}(N^2 d \log N)$ の時間計算量を要することになり，FINCH で愚直にストリーム処理させることに比べ 時間計算量を改善できる．しかし，S-FINCH は各データの追加やクラスタが変更後の更新のときに既存の全各データポイントとの距離を計算する必要がある．これにより追加されたデータ量が増えるにつれクラスタリングに膨大な時間を要することが問題となっている．

3 提 案 手 法

本節では提案手法について概説する．提案手法は，S-FINCH のクラスタリング精度を損なわずにクラスタリング結果をより高速に計算することを目的とする．まず提案手法の基本的なアイデアについて述べ，その詳細について 3. 2 節以降で説明する．

3.1 基本アイデア

S-FINCH では新しいデータが追加された時に最近傍と逆最近傍を探索するために，これまでに追加された全てのデータポイントとの距離を計算しており，時間計算量の多くを占める．ゆえに，クラスタリングにかかる時間を短縮するためには距離

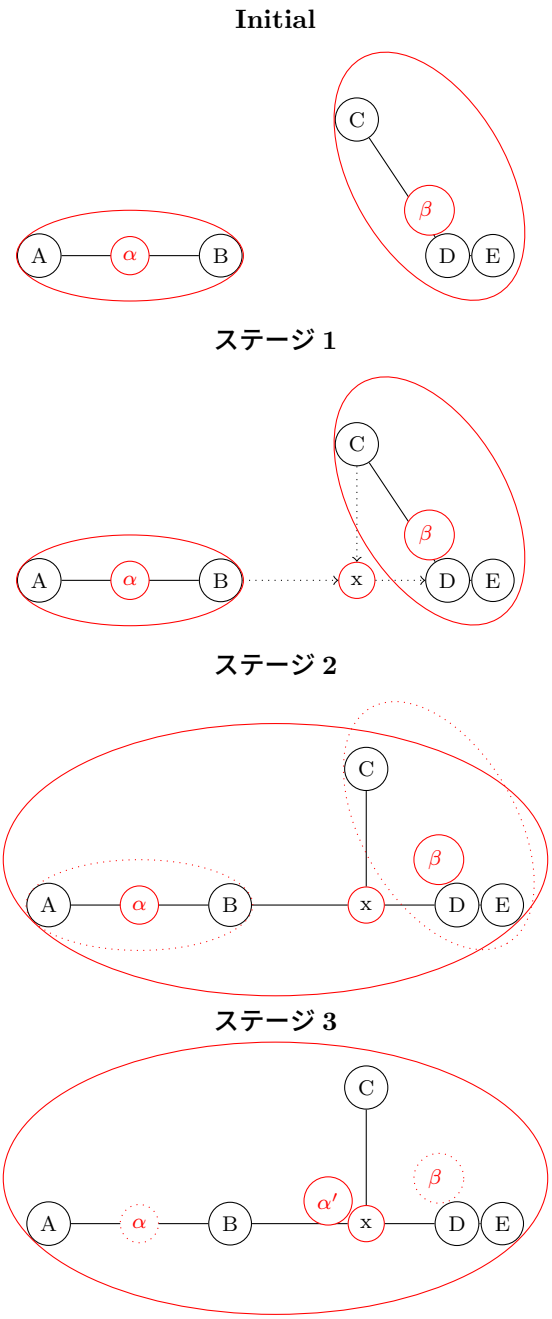


図 2: S-FINCH の詳細

計算回数を減らすことが重要である．そこで本研究ではこの最近傍と被最近傍をより少ない距離計算回数で求めることを提案手法の目的とする．

提案手法の基本アイデアは前述した距離計算において空間索引を用い，距離計算が不要なデータポイントを簡単に特定することで，全体の距離計算コストを削減することである．まず，空間索引を構築する (3.2 節)．続いて空間索引を用いた最近傍探索と逆最近傍探索で距離計算コストを下げる (3.3 節)．また，構築した空間索引はストリーム処理の各データ追加に伴い最新の状態に更新され続ける．

具体的にはデータポイント x の最近傍を探索するにあたってデータポイント a,b を包含する空間 V をおく．$dist(a,x), dist(b,x) \geqq min(dist(x,V))$ であることから，V との最小距離を計算することで V が包含する各データポイントの距離の下限を求められる．これにより最近傍を計算するにあたりデータポイント x との距離が最小距離以下であるデータポイントを既に求めている場合，データポイント x と V が包含するデータポイントの距離計算を省略できることがある．

データポイント x の逆最近傍を探索するにあたってデータポイント a を中心，a の最近傍への距離を半径とする超球 W をおく．$x \in W$ であることと a の最近傍が x になることは必要十分である．また，データポイント a,b の超球 W_a, W_b を包含する空間 S があるとする．$x \notin S$ ならば，$a,b \notin S$ である．これを利用することで，データポイント x が空間 S に含まれないとき空間 S 内の各データポイントは逆最近傍にはならないことがわかり，各データポイントとの距離計算を削減できる．

提案手法は，S-FINCH と比較して，クラスタリング精度を損なわずに大規模なデータに対して高速にクラスタリングが可能であるという優位性を持つ．これは，データが大規模になるにつれ探索において削減した距離計算コストが空間索引を構築するコストを大きく上回るようになる点によるものである．

3.2 空間索引構築

3.1 節 で述べた空間索引は，空間との最小距離を求めることと空間への包含の有無を判定する必要がある．これを実現するための空間として中心 p 半径 r の超球 S を採用する．超球を用いることでデータポイント x と空間との最小距離は $min_dist(x,S) = max(dist(x,p) - r, 0)$ で求められ，S が x を内包するかの判定は $dist(x,p) \leqq r$ であるかどうかを確認することで実行できる．中心との距離計算は $\mathcal{O}(d)$ の時間計算量で求められるため，空間との最小距離，包含の有無いずれも $\mathcal{O}(d)$ の時間計算量しか必要としない．超矩形など異なる空間を表現するのに比べ時間計算量が小さく，効率的に空間索引の構築・利用ができる．1 つの空間索引を構築するための時間計算量と距離計算の時間計算量が同じ程度であるため，空間索引を 1 度用いるごとに 2 つ以上の頂点との距離計算が不要である判定を行うことが出来た場合に空間索引によって距離計算コストを削減出来る．また，最近傍探索と逆最近傍探索は同じ空間索引を使えないため 3.2.1 節 と 3.2.2 節 で説明する．

3.2.1 最近傍探索

3.1 節 で述べた通り，提案手法は最近傍探索のための空間索引として データポイントを包含する超球を求める．高さ h のデータポイント x の子孫のうち，高さ sh にあるデータポイントの全てを包含する超球を $V_{h,sh,x}$ とする．$h = sh$ の時， $V_{h,sh,x}$ は データポイント x の点となる．また，$\forall y \in children(x), V_{h-1,sh,y} \subset V_{h,sh,x}$ と定義できる．

また，x の全ての子 $children(x)$ の超球を包含する超球 $V_{h,sh,x}$ は中心を代表点とし，半径を $\max\{\max(dist(V_{h,sh,x}\text{の代表点}, V_{h-1,sh,y})) \mid y \in children(x)\}$ とすることで求められる．

3.2.2 逆最近傍探索

逆最近傍探索のための空間索引も 3.2.1 節 と同様に構築できる．最近傍探索と異なる点は，データポイントではなく，最

近傍距離を半径とする超球を包含する空間索引を構築することである．

高さ sh にあるデータポイント y を中心とし，y の最近傍距離を半径とする超球を $W_{sh,y}$ とする．高さ h のデータポイント x の子孫のうち，高さ sh にあるデータポイントを包含する超球を $S_{h,sh,x}$ とする．$h = sh$ の時，$S_{h,sh,x}$ は $W_{sh,x}$ となる．また，$\forall y \in children(x), S_{h-1,sh,y} \subset S_{h,sh,x}$ と定義できる．

また，x の全ての子 $children(x)$ の超球を包含する超球 $S_{h,sh,x}$ は中心を代表点とし，半径を $\max\{\max(dist(S_{h,sh,x}\text{の代表点}, S_{h-1,sh,y})) \mid y \in children(x)\}$ とすることで求められる．

3.3 探　　索

3.3.1 最近傍探索

最近傍探索は，最近傍を探索したい起点データポイント $data$，起点データポイントの存在する高さ $data_h$ を入力として要求し，$data_h$ における，$data$ の最近傍 $NN(data)$ を出力する．最近傍探索は，次の処理を再帰的に繰り返すことで実現される．探索する高さを $height$，起点データポイントを $data$，起点データポイントの存在する高さを $data_h$，探索するクラスタ集合を $clusters$ とおく．暫定最近傍とは，その時点までに探索したデータポイントの中で最も $data$ に近いデータポイントのことを指し，暫定最近傍より $data$ に近いデータポイント new が見つかった場合に new が暫定最近傍へと更新される．

$height$ が $data_h$ よりも高い場合，2 つのステージで動作する．ステージ 1 で，高さ $height$ の $clusters$ に含まれる各データポイントの超球との最小距離を計算する．ステージ 2 で，最小距離の小さいクラスタ $child$ から順に，クラスタ $child$ の子のデータポイントが暫定最近傍を更新する可能性があるか判定する．更新する可能性がある場合，次に探索する高さを $height - 1$，起点データポイントを $data$，起点データポイントの存在する高さを $data_h$，次に探索するクラスタ集合を $children(child)$ として次の処理を実行し，返されるデータポイント x が暫定最近傍より近い場合，暫定最近傍を x で更新する．更新する可能性がない場合，残りのクラスタについては無視をする．この時点での暫定最近傍を最近傍とする．また，暫定最近傍が空であるとき，暫定最近傍との距離を ∞ とする．

$height$ が $data_h$ と同じである場合，$height$ の $clusters$ に含まれるデータポイント x が暫定最近傍を更新するようであれば，暫定最近傍を x で更新する．全てのデータポイントについての処理後の暫定最近傍を返り値とする．

また，この処理における “クラスタ $child$ の子のデータポイントが暫定最近傍を更新する可能性” は次のように判定する．$child$ は，$height$ にある子孫に対して最近傍探索するための超球 $V_{height,data_h,child}$ を持つため，この超球との最小距離 $min(dist(data, V_{height,data_h,child}))$ が暫定最近傍との距離より小さい場合に更新する可能性があると判定する．以上の処理を，Algorithm 1 で示した．

Algorithm 1 に疑似コードを示す．最近傍探索の開始時には，

Algorithm 1: 最近傍探索

Input: 起点データポイント $data$，起点データポイントの存在する高さ $data_h$，木の高さ H
Output: $data$ の最近傍

```
1  Func FIND_NN(d, d_h, h, cand, c_set):
2      distance_list ⟵ empty array
3      foreach c ∈ c_set do
4          distance_list ← distance_list.append
5          (pair(min_dist(d, V_{h,d_h,c}), c))
6      distance_list ← sort(distance_list)
7      foreach (distance, c) ∈ distance_list do
8          if distance ≧ dist(d, cand) then
9              break
10         else
11             if d_h < h then
12                 cand ←
                     FIND_NN(d, d_h, h − 1, cand, children(c))
13             else
14                 cand ← c
15     return cand
16 Func FIND_NN_entry(data, data_h):
17     return FIND_NN((data, data_h, H, null, children(root)))
```

最近傍を探索したいデータポイントを x，x の存在する高さを x_h として処理を開始する．まず，前述の入力で `FIND_NN_entry` 関数を呼び出す (line 16)．最上層のデータポイント集合が探索開始条件が設定され探索が開始される (line 1)．探索対象の各データポイントとの距離を格納するリストを初期化する (line 2)．探索対象の各データポイントとの距離を計算し (line 3)．データポイントとともに格納する (line 5)．リストを距離の昇順にソートする (line 6)．リストの先頭から確認し，暫定最近傍を更新する可能性があるかを判定する (line 8)．可能性がない場合はこのリストの処理を終了する (line 9)．探索中の階層が探索対象の階層でない場合は，データポイントの子の探索を行う (line 12)．探索中の階層が探索対象の階層である場合は，暫定最近傍の更新を行う (line 14)．リストの処理が終了した場合，暫定最近傍を返り値とする (line 15)．

3.3.2 逆最近傍探索

逆最近傍探索は，逆最近傍を探索したい起点データポイント $data$，起点データポイントの存在する高さ $data_h$ を入力とし要求し，$data_h$ における，$data$ の逆最近傍集合 $NNset(data)$ を出力する．逆最近傍探索は，次の処理を再帰的に繰り返すことで実現される．また，逆最近傍集合は各再帰において共通の集合である．探索する高さを $height$，起点データポイントを $data$，起点データポイントの存在する高さを $data_h$，探索するクラスタ集合を $clusters$ とおく．

$height$ が $data_h$ よりも高い場合，2 つのステージで動作する．$clusters$ の構成要素となる各クラスタ $child$ の子のデータポイントが逆最近傍を更新する可能性があるか判定する．可能性がある場合，次に探索する高さを $height - 1$，起点データポイント $data$，起点データポイントの存在する高さ $data_h$，次

Algorithm 2: 逆最近傍探索

Input: 起点データポイント $data$，起点データポイントの存在する高さ $data_h$， 木の高さ H
Output: $data$ の逆最近傍

```
1  Func FIND_NNset(d, d_h, h, cands, c_set):
2      foreach c ∈ c_set do
3          if d ∈ W_{h,d_h,c} then
4              if d_h < h then
5                  cands ← cands.append(
6                  FIND_NNset(d, d_h, h−1, cands, children(c)))
7              else
8                  cands ← c
9      return cands
10 Func FIND_NNset_entry(data, data_h):
11     return
       FIND_NNset((data, data_h, H, null, children(root)))
```

に探索するクラスタ集合 $children(child)$ として次の処理を行なう．可能性がない場合，次のクラスタへと処理をすすめる．

$height$ が $data_h$ と同じである場合，$height$ の $clusters$ に含まれるデータポイント x 最近傍が $data$ になるようであれば，逆最近傍集合に x を追加する．

また，この処理における “各クラスタ $child$ の子のデータポイントが逆最近傍を更新する可能性” は次のように判定する．$child$ は，$height$ にある子孫に対して逆最近傍探索するための超球 $S_{height,data_h,child}$ を持つため，この超球に $data$ が含まれる場合，逆最近傍集合を更新する可能性があると判定する．以上の処理を，Algorithm 2 で示した．

Algorithm 2 に疑似コードを示す．被最近傍探索の開始時には，被最近傍を探索したいデータポイントを x，x の存在する高さを x_h として処理を開始する．まず，前述の入力で `FIND_NNset_entry` の関数を呼び出す (line 10)．最上層のデータポイント集合が探索開始条件が設定され探索が開始される (line 1)．各データポイントが被最近傍になる可能性があるかを判定する (line 2)．可能性がない場合は次のデータポイントに進む．可能性があり，探索中の階層が探索対象の階層でない場合は，データポイントの子の探索を行い，返り値を被最近傍に追加する (line 6)．可能性があり，探索中の階層が探索対象の階層である場合は，被最近傍の追加を行う (line 7)．リストの処理が終了した場合，被最近傍を返り値とする (line 9)．

4 評価実験

提案手法の有効性を評価するために，我々の提案した高速化手法および S-FINCH に対し，処理の高速性およびクラスタリング結果の正確性の観点から比較評価を行う．

4.1 実験設定

4.1.1 実験環境

本実験には CPU Intel Xeon E5-2690 2.6 GHz, メモリ 128 GB の Linux サーバを利用する．実験に利用したプログラムは C++ で実装し，g++ (GCC) 9.2.0 で実行速度の最適化を目的に `O2` オプションをつけコンパイルした．

表 2: 実験に用いたデータセット

	N	d	計測範囲
Mice Protein	1,077	77	-
MNIST	10,000	784	-
3D Road Network	434,874	3	$[399900, 400000)$

4.1.2 データセット

本実験では，FINCH の評価実験でも利用されていたタンパク質の出現率と，8 種類の遺伝子型への分類である Mice Protein [8] ，手書き数字分類として有名な MNIST [9] で実験を行った．また新たに，各道路を構成する座標の集合である 3D Road Network [10] でも実験を行った．3D Road Network は道路 ID, 緯度, 経度, 高度 の情報を持つため，道路 ID を除く緯度, 経度, 高度 の 3 次元を入力として実験を行った．各データセットの詳細は表 2 に示す．

本実験では，各データセットをデータストリームであると想定し，1 件ずつ逐次的にデータを追加した際の，データポイント間距離計算を行った回数および実行時間の比較を行う．また，3D Road Network のように長大なデータセットにおいて実行が終了しなかったため，データセットの終盤における 100 件のデータを対象に評価を行った．対象としてデータの区間に関しては表 2 に示した．また，3D Road Network では全ての階層に対して空間索引を構築すると性能が悪化していたため階層 1,2 のみに対して空間索引を構築し階層 3 より上の階層に関しては従来手法と同じ全点への愚直な距離計算を行った．

4.2 高速性

本実験では，各データセットをデータストリームであると想定し逐次的にデータを追加し，データポイント間距離計算を行った回数の比較及び総実行時間の比較を行う．

距離計算実行回数の実験結果を図 3 と 4 に示す．横軸がそのデータが追加された時点でのデータストリームの長さであり，縦軸がそれまでに要した累計データポイント間距離計算回数とその内訳である．Mice Protein, MNIST 共に探索および空間索引構築および最近傍探索のための距離計算回数は少なくなっている．

3D Road Network の実験結果を 図 5 と 6 に示した．また，3D Road Network は，膨大な実行時間を必要としたため，10k 件で実行を打ち切っている．図 5 から提案手法は従来手法に比べて 28%ほど高速に実行可能であることが分かる．図 6 から提案手法は空間索引を構築した階層においては距離計算を 90% 以上削減，実行時間に関しては 70%削減を達成した．

4.3 正確性

提案手法と従来手法 S-FINCH の出力するクラスタリング結果の正確性について評価を行う．

提案手法は S-FINCH における最近傍と逆最近傍を探索する

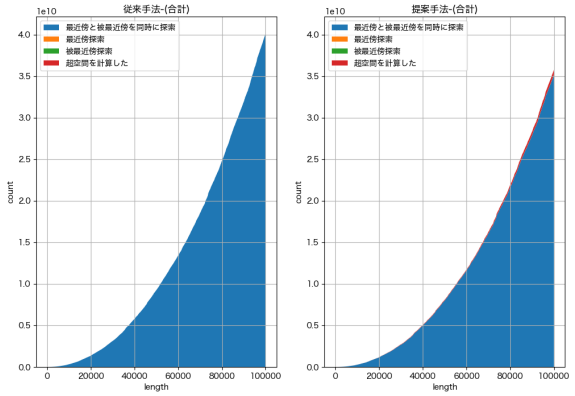


図 3: MNIST のデータポイント間距離計算の実行回数の比較

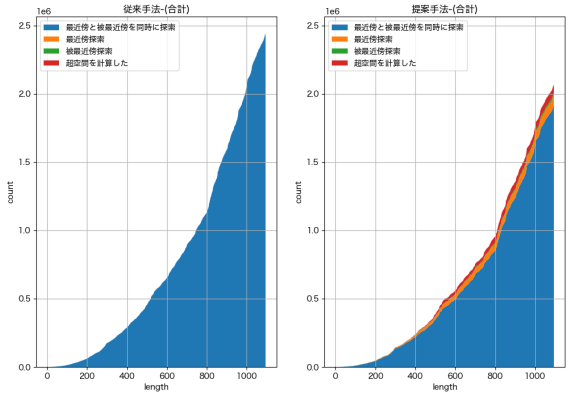


図 4: Mice Protein のデータポイント間距離計算の実行回数の比較

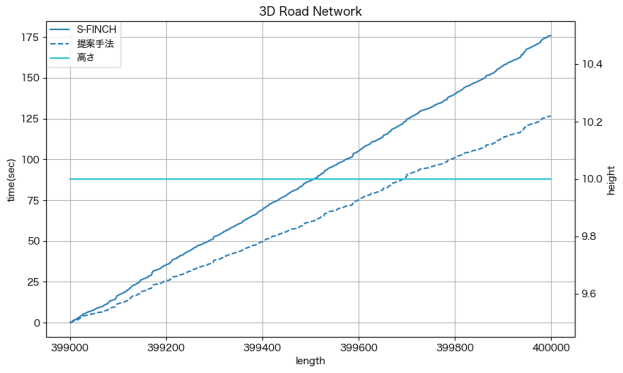


図 5: 3D Road Network の区間内での累計総実行時間と階層の高さ

部分について空間索引を導入する手法であり，この空間索引は愚直に計算を行ったときと同一の厳密解を出力するできる．そのため，提案手法は S-FINCH と完全に一致するクラスタリング結果を出力する手法となり，各データポイントが追加された後に出力される階層的クラスタリングの結果が一致する．この性質を実験的に検証するために，実行終了後における S-FINCH

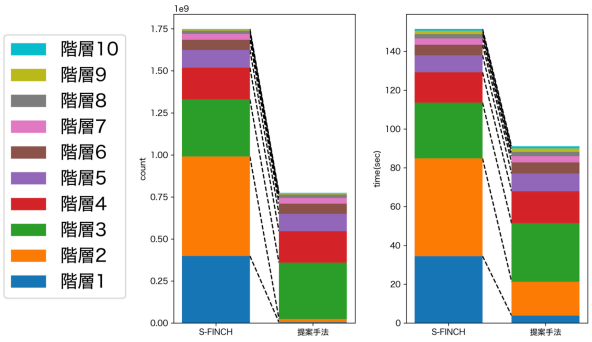


図 6: 3D Road Network の区間内での階層別の距離計算回数 (左) と探索時間 (右)

表 3: 各データセットを S-FINCH と提案手法でクラスタリングした結果の各階層おける NMI スコア

階層 (高さ)	Mice Protein	MNIST	3D Road Network
1	1.000	1.000	1.000
2	1.000	0.999	1.000
3	1.000	1.000	0.999
4	1.000	1.000	0.999
5	1.000	1.000	0.999
6	0.999	1.000	1.000
7	-	1.000	1.000
8	-	-	0.999
9	-	-	1.000

と提案手法のクラスタリング結果を比較する．評価指標として Normalized Mutual Information (NMI) を採用し，評価時には scikit-learn というライブラリに収録されている NMI 関数 [11] を利用して計算した．これによって計算される NMI スコアは 1 に近いほど性能が良く，1 は完全に一致していることを示す．その結果の NMI スコアを表 3 に示す．これは各手法のクラスタリング結果を NMI スコアを階層ごとに計算したものである．- は，そのデータセットのクラスタリング結果に存在しない階層である．その結果，全ての階層で 1.0 に近い値を出力した．しかし，1.0 を超えているものや，1.0 に限りなく近いものの 1.0 ではないもの結果も存在した．1.0 を超えることはないため，実行上の誤差ではないかと推測されるが，確証に至る根拠はない．そのため，クラスタラベルの出力を直接比較することにする．S-FINCH と提案手法は，最近傍・逆最近傍計算以外の処理は同一のものを採用している．そのため，提案手法の最近傍・逆最近傍探索の結果が正確である場合，クラスタリング結果のクラスタラベルに関しても同一のものが出力されるはずである．これを確かめるために各データセットにおいて出力されたクラスタラベルのファイルの差分を `diff` コマンドを用いて確かめた結果，一切の差分がなかった．これによりデータセットの全てにおいて提案手法と従来手法 S-FINCH は同一のクラスタリング結果を出力していることが分かった．

5 おわりに

本稿では，大規模データストリームに対して高速に S-FINCH する手法を提案し，その概要について示した．提案手法では，FINCH のもつ最近傍グラフによるクラスタを包含する空間によって空間索引を構築し，それを用いて最近傍計算における距離計算回数を削減した．提案手法は従来手法の貪欲な最近傍計算と同じ結果を返すことから，提案手法は従来手法の FINCH で生成されるクラスタリング結果と同様の処理結果をより少ない距離計算回数で出力する．これにより，本手法は大規模なデータストリームに対してより精度を損なうことなく空間索引による最近傍探索を可能にした．また，評価実験により低次元データセットに対しては実行時間を 30%削減出来ることを確認した．

今後は，高い階層における距離計算回数の削減，高次元における距離計算回数の削減を通して全体の実行時間の削減を目指すことや，より多くの低次元データセットでの評価を行いたい．

文　献

[1] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8926–8935, June 2019.

[2] James MacQueen, et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 281–297. Oakland, CA, USA, 1967.

[3] Yan-Lin He, Lei Chen, Yuan Xu, Qun-Xiong Zhu, and Shan Lu. A new distributed echo state network integrated with an auto-encoder for dynamic soft sensing. *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp. 1–8, 2023.

[4] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] M. Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11220–11229, 2021.

[6] James Cunningham, Jim Davis, Kyle Tarplee, and Juan Vasquez. S-finch: An optimized streaming adaptation to finch clustering. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1343–1349, Aug 2022.

[7] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8934–8943, 2019.

[8] Clara Higuera, Katheleen J Gardiner, and Krzysztof J Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, Vol. 10, No. 6, p. e0129126, 2015.

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

[10] Manohar Kaul. 3D Road Network (North Jutland, Denmark). UCI Machine Learning Repository, 2013. DOI: `https://doi.org/10.24432/C5GP51`.

[11] sklearn.metrics.normalized_mutual_info_score. `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html`, Date of access: 2024/01/31.

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

時系列データ処理

座長:渡辺 陽介(名古屋大学)
コメンテータ:喜田 拓也(北海学園大学)
2024年3月1日(金) 10:00 〜 12:10 T2-B (オンライン（Zoom Events))

11:25 〜 11:50

[T2-B-7-05]【技術報告】東芝の時系列データ処理基盤技術(GridDB)ならびにデータ仮想化エンジン(PGSpider)

浪岡保男、服部雅一、相川 恵 (株式会社東芝)

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

🏛2024年3月1日(金) 10:00 ~ 12:10 | 🏛T3-A オンライン（Zoom Events)

情報拡散・偽情報・ヘイトスピーチ

座長:山本 祐輔(静岡大学)
コメンテータ:石川 佳治(名古屋大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T3-A-7-01] 自然言語による説明を生成するfaithfulな偽情報検出モデルの試作

*吉田 倖[1]、張 建偉[1] (1. 岩手大学)

10:30 ~ 10:55
[T3-A-7-02] 誹謗中傷検出タスクの学習を補助するためのサブタスクの提案

*沢田 凌一[1]、鈴木 優[1] (1. 岐阜大学)

10:55 ~ 11:20
[T3-A-7-03] ソーシャルメディア投稿における炎上推定のための世論に対する乖離度算出手法

*吉田 裕輔[1]、鷹野 孝典[2] (1. 神奈川工科大学大学院 工学研究科 情報工学専攻 鷹野研究室、2. 神奈川工科大学 情報学部 情報工学科)

11:20 ~ 11:45
[T3-A-7-04] ニュース記事の拡散速度に着目したソーシャルメディア分析

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

11:45 ~ 12:10
[T3-A-7-05] Trustable Internetにおけるインターネット上のデータの信憑性判定および根拠説明文生成手法の提案

*佐久間 義友[1]、大木 憲二[1]、佐々木 佑樹[1]、坂本 拓也[1]、吉武 敏幸[1]、二村 和明[1] (1. 富士通株式会社)

自然言語による説明を生成するfaithfulな偽情報検出モデルの試作

吉田　倖†　張　建偉†

† 岩手大学理工学部　〒 020–8551 岩手県盛岡市上田 4–3–5
E-mail: †{s0620062,zhang}@iwate-u.ac.jp

あらまし　ソーシャルメディアが重要な情報収集源となり，インターネット上の偽情報が社会に対するさらなる脅威となっている．偽情報を検出するために Transformer に基礎を置く大規模言語モデルや CNN，GNN を用いたモデルが考案されている．しかし，これらはブラックボックスでありモデルの意思決定過程は解釈しがたい．このような複雑なモデルの解釈には LIME や SHAP を利用できる．ところが，これらの数値による説明はユーザーが容易に理解できるものではない．本論文では，短文による主張に焦点を当て，Natural-language Inference over Label-specific Explanations (NILE) に基づいた，faithfulness を考慮した偽情報検出モデルを提案する．この手法は，入力から説明文を生成しそれを用いて分類問題を解く．ユーザーにとって，生成される自然言語による説明は数値によるものより理解しやすい．実験では入力に対するモデルの sensitivity により提案手法の faithfulness を検証する．

キーワード　偽情報検出，XAI，説明可能性，faithfulness

1 はじめに

ソーシャルメディアは我々にとって重要な情報源となっている．2016 年時点で，米国の成人の 62%がソーシャルメディアを主要な情報源としていることが明らかになっている [1].

一方で，ソーシャルメディア上の偽情報の蔓延が問題となっている．2016 年の米国大統領選挙期間中，最も議論された 20 の間違ったニュースが，最も議論された 20 の正しいニュースよりも広く拡散されたことが明らかになっている [2]．また，Twitter 等のソーシャルメディア上では，事実よりも偽情報の方が多くのユーザーに共有され，拡散速度が速いことが明らかになっている．特に，政治に関わる情報にこの傾向が強く見られる [3].

偽情報として拡散されるものの 1 つに，短文による主張がある．例えば，TV 広告や Facebook 等のソーシャルメディア上の主張の大半は，新聞記事等と比較して短い文で表現される [4]. また，広く拡散される政治家の主張なども演説から切り取られた短文であることが多い.

ニュースや短文による主張等の情報の真偽判定は，それらに関わる様々な特徴量を用いて行われる．代表的なものとして，情報に含まれる単語的・文法的特徴を表す lexical features がある．また，情報の発信者や情報に対するコメント，関連ユーザー等を表す social features や，情報に含まれるヒストグラム等の画像を表す visual features がある.

深層学習や大規模言語モデルを応用した分類モデルは，前述の特徴量を用いることで高い精度で分類問題を解ける．だが，これらの複雑なモデルの説明可能性の低さが問題となっている.

これまでに，説明可能性に関する様々な研究が行われている．説明可能性を向上させるための手法は 2 つに大別される．1 つ目は，post-hoc 的手法である．これは，説明可能性を向上させるためのモジュールを説明可能性の低い複雑なモデルに接続するという手法である．2 つ目は，intrinsic 的手法である．これは，モデル構築時に説明可能性を考慮することで，外部モジュールなどを用いずに説明性を獲得するという手法である.

しかし，今までに提案された説明可能なモデルも改善点を持つ．例えば，post-hoc 的手法として頻繁に用いられるものに LIME [5] や SHAP [6] がある．これらの手法は，数値で説明を提示する．研究者は，提示された数値による説明を参考に探索的なモデルの解釈を行う．しかし，このような数値による説明は一般的な利用者にとって直感的に理解できるものではない．加えて，背景知識を持たないユーザーにとって，LIME や SHAP のアルゴリズムの理解も難しい.

そこで本論文では，真偽判定を行うと同時に，ユーザーにとってわかりやすい説明を生成するモデルを提案する．提案手法は，分類結果に対する説明を自然言語文で与える．また，入力の摂動が分類予測に与える影響を評価することでモデルの faithfulness を分析することができる．ただし，本論文では，Kumar らの先行研究 [7] に倣い faithfulness を次のように定義する.

Faithfulness *: how well do the provided explanations correlate with the model' s decision making?*

2 関連研究

分類問題を解く上で，深層学習や大規模言語モデルを応用することが効果的であることが報告されている．例えば，Rabgel と Rosso は，bot 検出及びユーザーのジェンダー予測には CNN を応用したモデルを用いることが効果的であることを明らかにした [8]．bot とは，他のユーザーとの交流や情報の拡散に人の介入を必要としないプログラムであり，偽情報の拡散に利用されることがある．また，Paka らは，BERT [9] や RoBERTa [10] 等の事前学習済み言語モデルを利用して，COVID-19 に関する偽情報を検出するモデルを考案した [11].

説明可能性を考慮した偽情報検出モデルも考案されている．Szczepański らは，post-hoc 的手法により説明可能性を獲得した偽情報分類手法を提案した [12]．提案されたモデルは，BERT の出力を説明器である LIME に入力することで分類の説明を行った．また，Shu らは，intrinsic 的手法により構築された説明可能な偽情報検出モデルとして dEFEND(**E**xplainable **F**ak**E** **N**ews **D**etection) を提案した [13]．dEFEND は構成要素に Attention 機構を持つ．dEFEND は，分類に際して Attention 機構の重みを分析し，分類に対する説明として提示している．Liu らも intrinsic 的手法に着目し，分類と説明を行うテキスト分類フレームワークとして GEF(**G**enerative **E**xplanation **F**ramework) を提案した．GEF は explainable factor により任意の形式の説明の提示を行う．また，予測と生成する説明の強い関係性を保証するために minimam risk training を学習に導入している．しかし，GEF の生成する説明文は正解の説明文と比較して短い傾向にあることが明らかにされている [14]．

3 事前知識

この章では，提案手法の基盤となった 2 つの研究を紹介する．3.1 では，提案手法が参考にしているフレームワークを説明する．3.2 では，分類に用いるモデルを説明する．

3.1 NILE : Natural-language Inference over Label-specific Explanations

Kumar らは，自然言語推論問題を解く faithful なモデルとして NILE(**N**atural-language **I**nference over **L**abel-specific **E**xplanations) を提案した [7]．NILE の概要を図 1 に示す．NILE は，独立して学習される Candidate Explanation Generators と Explanation Processor という 2 つの構成要素からなり，入力に対する分類ラベルと分類に対する説明文を生成する．分類問題と生成問題を解くために，それぞれ事前学習済み言語モデルの GPT2 [15] と RoBERTa [10] をファインチューニングして用いている．

3.1.1 データセット：e-SNLI

e-SNLI [16] は The Stanford NLI dataset(SNLI) [17] を拡張して作成された．SNLI に含まれるデータは特徴量として人間によって作成されたアノテーションを含む Premise(前提) p と Hypothesis(仮説) h，p と h の含意関係を表すラベルを持つ．e-SNLI は，これらに加えて含意関係に対する説明文 t_g を持つ．説明文の作成は Amazom Mechanical Turk を用いてクラウドソーシングされ，人手によって行われた．データサイズは 549,367 である．

3.1.2 問題設定

NILE は自然言語推論を解くために提案されたモデルである．自然言語推論とは，与えられた 2 つの文の含意関係を推論するという問題である．例えば，Premise：“犬が歩いている” と Hypothesis：“犬は寝ている” の包含関係は “矛盾” である．NILE は，入力として p と h を受け取りラベル x を予測する．ただし，x は $x \in \{entailment, neutral, contradiction\}$ と定義される．

3.1.3 CEG：Candidate Explanation Generators

Candidate Explanation Generators は，生成器 G_x を用い入力 p，h からラベルで条件付けられた説明文 t_x を生成する．ただし，$x \in \{entailment, neutral, contradiction\}$ であり，ラベルタイプを表す．

学習は次の流れに沿う．まず，入力特徴量 p，h ，t_g を用いて文字列：“*Premise: p Hypothesis : h [EXP]* t_g *[EOS]*” を作成し生成器として用いる事前学習済み言語モデル G_x に入力する．*[EXP]*，*[EOS]* は special token として語彙に追加する．次に，入力文字列と G_x の出力の cross entropy loss を求める．ただし，損失関数は入力文字列内の説明文 t_g のトークンに対してのみ適用される．

推論には，プロンプト:“*Premise : p Hypothesis : h [EXP]*” を用いる．ファインチューニングした生成器にプロンプトを入力しラベルで条件付けられた説明文 t_x を得る．

3.1.4 EP：Explanation Processor

Explanation Processor は p，h，t_x を用いて予測ラベルを生成する．ただし，p，h は CEG で説明文 t_x を生成するときに用いたものである．

学習は次の流れに沿う．まず，入力文字列 $concat_{ph}$ ：“*Premise: p Hypothesis : h*” を作成する．この $concat_{ph}$ と t_x を用いて中間スコア F を求めラベルスコア l_x の算出に用いる．ただし，F は次の式により求められる．また，$CLS_{enbed}(concat_{ph}, t_x)$ は RoBERTa の出力の [CLS] に対応するベクトルである．

$$F(concat_{ph}, t_x) = tanh(W * CLS_{enbed}(concat_{ph}, t_x))$$

中間スコアは次の 3 つの方法の中のどれか 1 つにより処理される．概要を図 2 に示す．

Independent :

$$l_x = W_{ind} F_{ind}(F(concat_{ph}, t_x))$$

Aggregate :

$$V_i = W_{Agg,\,i} F_{Agg}(F(concat_{ph}, t_x)),\ where\, i \in \{1, 2\}$$
$$l_{entail} = LogSumExp(V_1(t_{entailment}), V_2(t_{contradict}))$$
$$l_{contradiction} = LogSumExp(V_1(t_{contradiction}), V_2(t_{entailment}))$$
$$l_{neutral} = V_1(t_{neutral})$$

Append :

$$l_x = W_{Apn,\,x} F_{Apn}(concat_{enc}, t_x)$$

ただし，W は学習可能な重みである．また，$concat_{enc}$ は文字列: “$entailment : t_{entailment}\, contradiction : t_{contradiction}\, neutral : t_{neutral}$” である

計算されたラベルスコアは softmax 関数に入力された後，cross entropy loss の計算に用いられる．

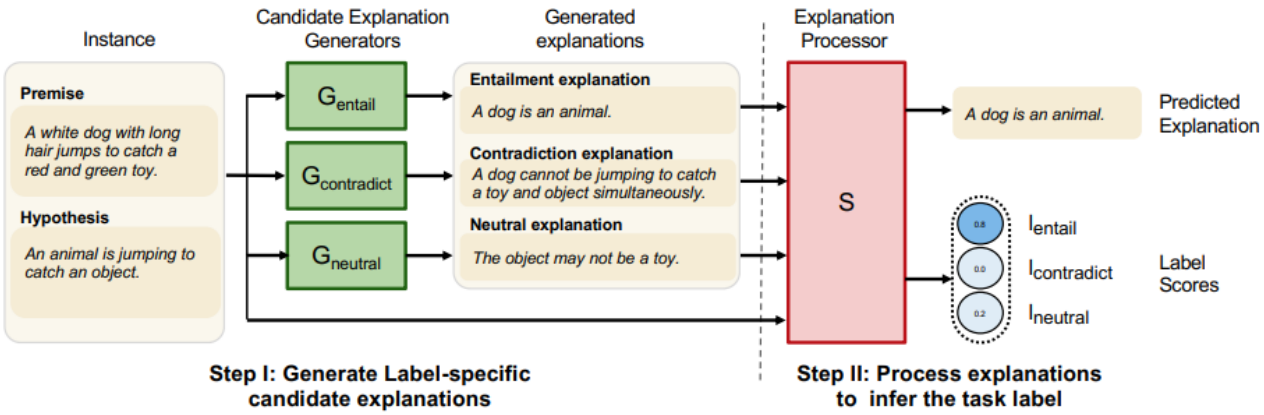


図 1: NILE の概要 [7]

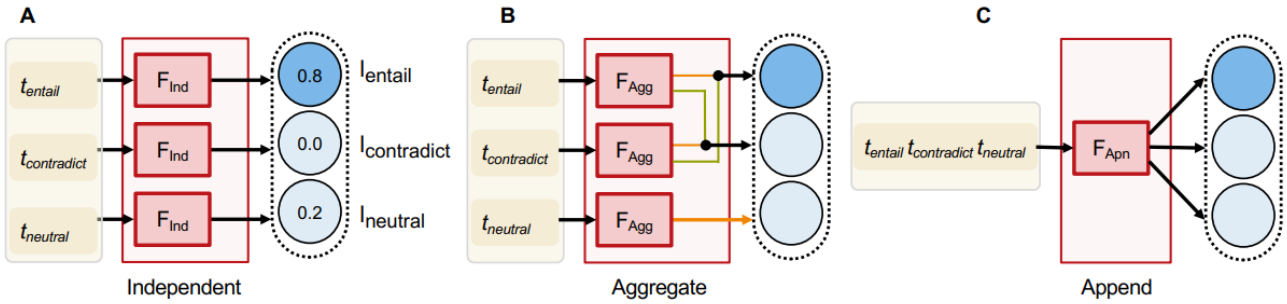


図 2: 中間スコアの算出方法 [7]

3.1.5 Negative Sampling

NILE は，ある Premise と Hypothesis，その含意関係を説明する説明文の関係性を用いて分類を行うことを期待されている．しかし，自然言語推論において説明文を分類に用いる場合，期待された方法により分類がなされるとは限らない．なぜならば，説明文は content と form という 2 つの情報を伝播するからである [7]．ここでは，content は説明文の意味的な情報を，form は文法的な情報を指す．つまり，異なるラベルに対する説明文が互いに大きく異なる特徴的な form を持つと仮定すると，無制約アルゴリズムは form を用いた分類を学習する可能性がある．これは，NILE が期待されている入力間の関係性による分類ではない．NILE は，このような学習を防止するために Negative Sampling により追加のデータサンプルを作成し，妥当なペアに対する分類スコアができるだけ高くなるように学習を行った．つまり，妥当な関係にある (p, h, t_x) に対して新たなデータサンプル $(p, h, t_x^{'})$，$(p, h, t_x^{''})$ を作成し学習に用いる．ここで，t_x，$t_x^{'}$，$t_x^{''}$ はラベル x の説明文の集合から無作為に抽出された説明文である．

3.2 ESM：Enhanced Sequence Model

Bhatt らは，短文による主張の真偽判定において，主張に関わる social features だけでなく，短文の真偽に対する説明文と credit score を用いて分類を行うモデルとして Enhanced Sequence Model を提案した [18].

3.2.1 データセット：LIAR-PLUS

LIAR-PLUS [19] は LIAR [4] を拡張して作成されたデータセットである．LIAR は，特徴量として主張と主張に関わる social features を持つ．例えば，主張が対象としているトピックを表す subject や発言者を表す speaker，発言者の役職を表す job である．データサイズは 12,836 である．LIAR は，PolitiFact に公開されているファクトチェック記事から作成されている．PolitFact は，専門家が行ったインターネット上の主張に関するファクトチェックの報告記事が公開されているウェブサイトである．専門家は，主張を“pants fire”，“false”，“barely true”，“half true”，“mostly true”，“true”のいずれかに分類している．

LIAR-PLUS は，LIAR に含まれる特徴量に加えて，専門家が作成した真偽判定に対する説明文を特徴量としてもつ．ただし，説明文は PolitiFact 上のファクトチェック記事から抽出されている．LIAR-PLUS に含まれる特徴量を表 1 に示す.

3.2.2 問 題 設 定

ESM は，LIAR-PLUS を用いて学習及び推論を行う．特に，短文による主張 s，短文に関わるメタデータ m，専門家により作成された説明文 j，credit score cs を入力として受け取る．ただし，二値分類を行う場合，ラベル“pants fire”，“false”，

表 1: LIAR-PLUS に含まれる特徴量

特徴量	例
the ID of the statement	11972.json
the label	true
the statement	Building a wall on the U.S.-Mexico border will take literally years.
the subject(s)	immigration
the speaker	rick-perry
the speaker's job title	Governor
the state info	Texas
the party affiliation	republican
barely true counts	30
false counts	30
half true counts	42
mostly true counts	23
pants on fire counts	18
the context	Radio interview
the extracted justification	Meantime, engineering experts agree the wall would most likely take years to complete. Keep in mind, too, it took more than six years to build roughly 700 miles of fence and barriers along the roughly 2,000-mile U. S. -Mexico border.

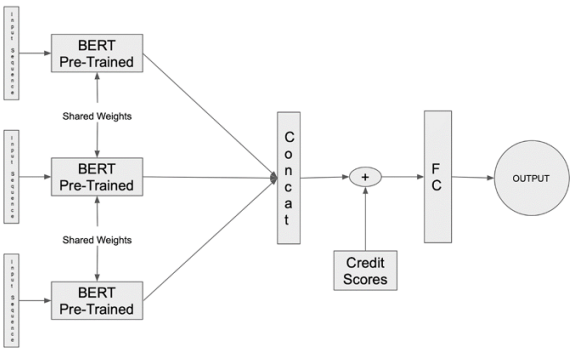


図 3: ESM の概要 [18]

“barely true”をラベル”false”とする．また，ラベル“half true”，“ mostly true ”，“ true”をラベル”true”とする．

3. 2. 3 シャムネットワーク

Koch らは，深層学習で利用されるニューラルネットワークとしてシャムネットワーク [20] を提案した．シャムネットワークは，同一の構造と重みを持つネットワークを用いて複数の入力を処理する．このネットワークは異なる入力間の違いを効果的に学習できることが明らかになっている．ESM は BERT を用いてシャムネットワークを構築し，入力を処理している．

3. 2. 4 アーキテクチャ

ESM の概要を図 3 に示す．ESM は，入力として短文による主張 s，メタデータ m，説明文 j，credit score cs を用いる．ただし，短文による主張と説明文はそれぞれ LIAR-PLUS に含まれる “*the statement*”，“*the extracted justification*” に対応する．また，メタデータは LIAR-PLUS に含まれる諸特徴量を結合して作成される文字列: “*subject speaker job state affiliation context*” を指す．さらに，credit score は主張の発言者の発言履歴から次のように計算される．

$$i = \frac{0.2 * MTC + 0.5 * HTC + 0.75 * BTC + 0.9 * FC + 1 * PFC}{MTC + HTC + BTC + FC + PFC}$$

$$cs = tanh(w * i + b)$$

ただし，MTC，HTC，BTC，FC，PFC はそれぞれ LIAR-PLUS に含まれる $mostly\,true\,counts$，$half\,true\,counts$，$barely\,true\,counts$，$false\,counts$，$pants\,on\,fire\,counts$ に対応する．

credit score の値は，短文による主張の発言者の発言履歴に正確な発言が多く含まれているほど小さくなる．言い換えれば，過去の発言に基づけば，発言者は信頼に足る人物であると解釈できる．

学習は次の流れに沿って行われる．まず，入力特徴量 s，m，j を BERT を用いて構築されたシャムネットワークに入力し，それぞれの入力特徴量に対する埋め込み $\boldsymbol{E_s}, \boldsymbol{E_m}, \boldsymbol{E_j} \in \mathbb{R}^{1\times 768}$ を得る．ただし，$\boldsymbol{E_s}$，$\boldsymbol{E_m}$，$\boldsymbol{E_j}$ は BERT の最終層の出力の [CLS] に対応するベクトルである．これらを連結して埋め込み $\boldsymbol{E_{concat}} \in \mathbb{R}^{1\times 3*768}$ を作成する．この埋め込みに credit score を足し合わせ中間生成物 $\boldsymbol{F}$ とする．$\boldsymbol{F}$ を線形層に入力した後 softmax 関数に入力する．softmax 関数の出力と正解ラベルを用いて cross entropy loss を計算し BERT をファインチューニングする．推論時は，softmax 関数の出力の最大値に対応するラベルを予測ラベルとする．

3. 2. 5 ESM の持つ問題点

Bhatt らは，説明文の利用がモデルの分類パフォーマンスの向上に寄与することを明らかにした [18]．しかし，実際に短文による主張の真偽判定を行う時，このような説明文を用いることはできない．なぜならば，説明文は専門家によるファクトチェックに際して書かれるものであるからだ．つまり，ある主張に対する説明文が存在している場合，その主張の真偽判定は既に専門家により行われている．このような場合，説明文を用いた機械的な真偽判定を行う必要が無い．言い換えれば，専門

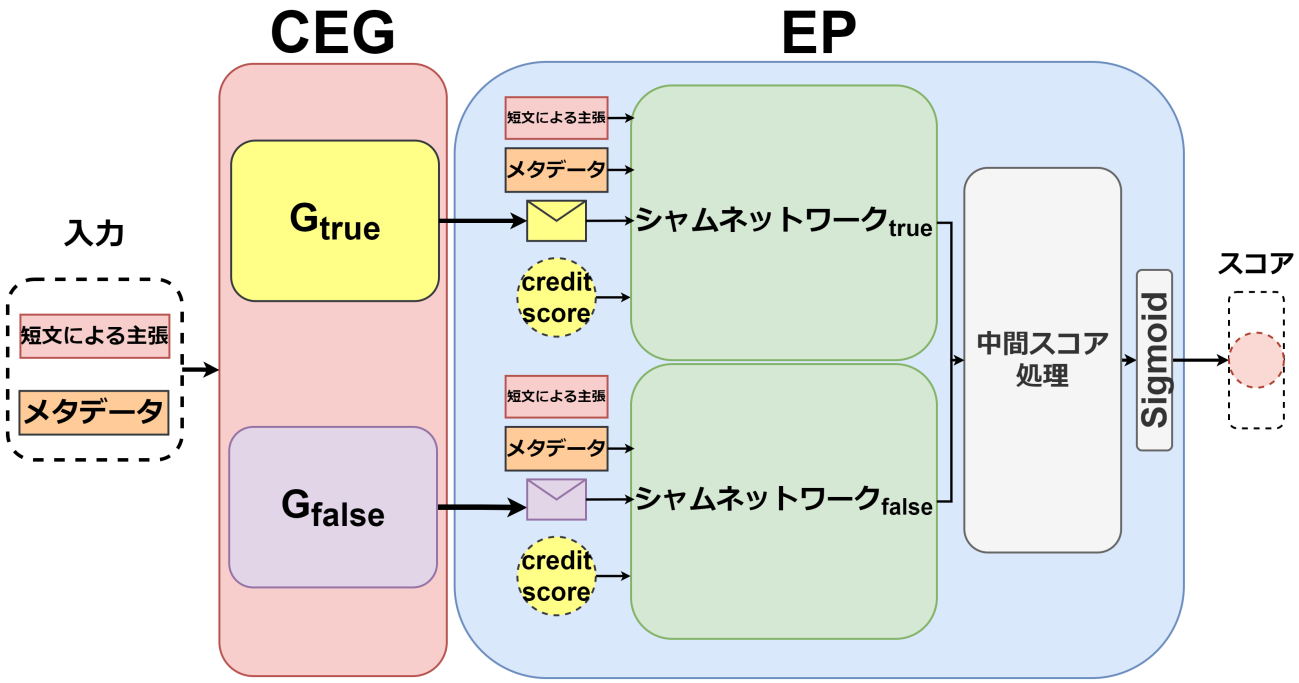


図 4: 提案手法の概要

家により作成された説明文を分類に用いる ESM を実世界における主張の真偽判定にそのまま用いることはできない．

4 提案手法

この章では，事前知識で共有した既存研究に基礎を置く提案手法を紹介する．まず，4.1，4.2 でそれぞれ表記法と問題設定を確認する．また，4.3，4.4 で既存研究に施す変更を説明する．さらに，4.5 では実験にいるベースラインを紹介する．

提案手法の概要を図 4 に示す．提案手法は 3.1 で紹介した NILE のフレームワークを用いて分類及び生成問題を解く．これにより，3.2.5 で述べた ESM の問題点を解決する．3.2 で紹介した ESM は分類問題を解く際に用いられる．

4.1 表記法

分類の対象となる短文による主張，主張に関わるメタデータをそれぞれ s，m と定義する．ただし，メタデータは 3.2.4 にある方法で作成される文字列である．また，ラベルタイプと各ラベルに対する分類スコアをそれぞれ $x \in \{true, false\}$，l と定義する．説明文を生成する生成器を G と定義する．特に，ラベル x で条件付けられた説明を生成する生成器を G_x で表す．さらに，データセット内に含まれる専門家によって作成された説明文と生成器 G_x が生成した説明文とを，それぞれ t_g，j_x と定義する．正解ラベルとモデルによる予測ラベルをそれぞれ y，y' と定義する．

4.2 問題設定

提案手法は，短文による主張を true と false のいずれか一方に分類する．提案手法の学習及び推論には 3.2.1 で紹介した LIAR-PLUS を用いる．LIAR-PLUS のデータサイズは，訓練データが 10,242，テストデータが 1,267，検証データが 1,284 である．

4.3 Candidate Explanation Generators

学習は次の流れに沿う．まず，入力 s，m，t_g を用いて文字列：“*Statement: s Metadata: m [EXP] t_g [EOS]*” を作成し生成器として用いる事前学習済み言語モデル G_x に入力する．*[EXP]*，*[EOS]* は special token として語彙に追加する．次に，入力文字列に含まれる説明文 t_g と生成された説明文 j_x の cross entropy loss を求める．

推論時は，プロンプト:“*Statement: s Metadata: m [EXP]*” をファインチューニングした G_x に入力し，文章補完を行わせる．このようにして，ラベルで条件付けられた説明文 j_x を得る．

4.4 Explanation Processor

提案手法における EP の詳細を図 5 に示す．学習は次の流れに沿う．まず，ファインチューニング済みの Candidate Explanation Generators に s，m を入力し j_{true}，j_{false} を生成する．次に，生成された説明文と諸特徴量をそれぞれ *BERT* によるシャムネットワークに入力し，s，m，j_x に対する埋め込み $\boldsymbol{E_s}, \boldsymbol{E_m}, \boldsymbol{E_{j_x}} \in \mathbb{R}^{1\times 768}$ を得る．ただし，$\boldsymbol{E_s}$，$\boldsymbol{E_m}$，$\boldsymbol{E_{j_x}}$ は *BERT* の最終層の出力の [CLS] に対応するベクトルである．

$$\boldsymbol{E_s} = BERT(s) \tag{1}$$

$$\boldsymbol{E_m} = BERT(m) \tag{2}$$

$$\boldsymbol{E_{j_x}} = BERT(j_x) \tag{3}$$

これらを連結してベクトル $\boldsymbol{E_{concat_x}} \in \mathbb{R}^{1\times 3*768}$ を作成する．$\boldsymbol{E_{concat_x}}$ に credit score cs を足し合わせ中間生成物 $\boldsymbol{F_x} \in \mathbb{R}^{1\times 3*768}$ とする．

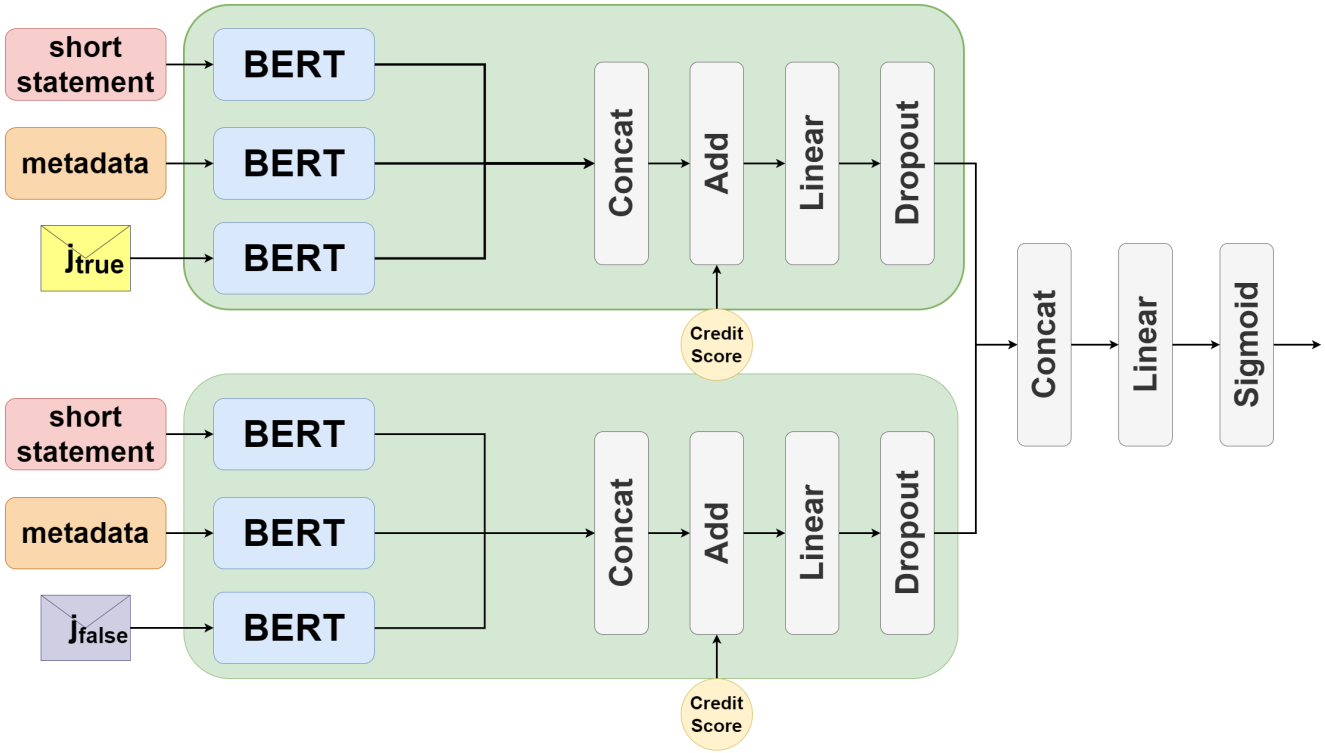


図 5: 提案手法における EP の詳細

$$\boldsymbol{E}_{concat_x} = Concat(\boldsymbol{E}_s,\ \boldsymbol{E}_m,\ \boldsymbol{E}_{j_x}) \tag{4}$$

$$\boldsymbol{F}_x = Add(\boldsymbol{E}_{concat_x},\ cs) \tag{5}$$

$\boldsymbol{F}_x$ を線形層 $Linear$ に入力し $\boldsymbol{h}_x \in \mathbb{R}^{1\times 768}$ を得た後，$\boldsymbol{h}_x$ をドロップアウト率 0.1 のドロップアウト層 $Dropout_{0.1}$ へ入力し $\boldsymbol{h}'_x$ を得る．

$$\boldsymbol{h}_x = Linear(\boldsymbol{F}_x) \tag{6}$$

$$\boldsymbol{h}'_x = Dropout_{0.1}(\boldsymbol{h}_x) \tag{7}$$

このようにして得た $\boldsymbol{h}'_{true}$，$\boldsymbol{h}'_{false}$ を連結し $\boldsymbol{e}_{concat} \in \mathbb{R}^{1\times 2*768}$ を得る．最後に，$\boldsymbol{e}_{concat}$ を線形層 $Linear$ に通し得た $\boldsymbol{e}'_{concat} \in \mathbb{R}$ を $Sigmoid$ 関数に入力し l を得る．

$$\boldsymbol{e}_{concat} = Concat(\boldsymbol{h}'_{true},\ \boldsymbol{h}'_{false}) \tag{8}$$

$$\boldsymbol{e}'_{concat} = Linear(\boldsymbol{e}_{concat}) \tag{9}$$

$$l = Sigmoid(\boldsymbol{e}'_{concat}) \tag{10}$$

Kumar らが EP の学習で行っていた Negative Sampling を提案手法の学習では行わない．自然言語推論問題を解く場合，モデルは高度な普遍的自然言語解釈能力を用いて分類を行う．また，それぞれの包含関係に対する説明文も普遍的な自然言語能力を用いて作成される．事実，e-SNLI に含まれる説明文は自然言語処理の専門家では無く，クラウドワーカーにより作成されている．このような場合，同一ラベルの説明文が特有の文法的特徴を持っていることがある．しかし，偽情報検出問題を解く場合，モデルは自然言語解釈能力だけでは無く特定のドメイン知識を用いて分類を行う．特に，説明文は短文による主張に関わるドメイン知識に基づいて作成される．言い換えれば，偽情報検出において，同一ラベルの判別対象に対する説明文が特定の文法的特徴を共有しているとは考えにくい．以上の理由から，Kumar らの先行研究で行われていた Negative Sampling を用いた学習を本論文では行わない．

4.5 ベースライン

ESM-j：提案手法との分類パフォーマンスを比較するために ESM-j をベースラインとして設定する．ESM-j は，3.2 で紹介した ESM のアーキテクチャを踏襲している．しかし，分類に説明文を用いないという点で ESM と異なる．また，提案手法が短文による主張の分類と分類に対する説明文の生成を行うモデルであるのに対して，ESM-j は分類のみを行うモデルである．

次に，ESM-j の情報処理の流れを説明する．始めに，短文による主張の処理を説明する．まず，短文による主張 s を $BERT$ に入力し埋め込み $\boldsymbol{E}_s \in \mathbb{R}^{1\times 768}$ を得る．ただし，$\boldsymbol{E}_s$ は $BERT$ の最終層の出力の [CLS] に対応するベクトルである．$\boldsymbol{E}_s$ をセル数 128 の $LSTM$ に入力し $\boldsymbol{h}_s \in \mathbb{R}^{\mathbf{1\times 128}}$ を得た後，$\boldsymbol{h}_s$ をドロップアウト率 0.15 のドロップアウト層 $Dropout_{0.15}$ に入力し，$\boldsymbol{h}'_s$ を得る．ただし，$\boldsymbol{h}_s$ は $LSTM$ の最後のセルの出力である．

$$\boldsymbol{E}_s = BERT(s) \tag{1}$$

$$\boldsymbol{h}_s = LSTM(\boldsymbol{E}_s) \tag{2}$$

$$\boldsymbol{h}'_s = Dropout_{0.15}(\boldsymbol{h}_s) \tag{3}$$

続いて，$\boldsymbol{h}'_s$ を線形層 $Liner$ に入力し $\boldsymbol{h}''_s \in \mathbb{R}^{1\times 32}$ を得た後，$\boldsymbol{h}''_s$ をさらに活性化関数 $ReLU$ に入力し $\boldsymbol{h}'''_s$ を得る．$\boldsymbol{h}'''_s$ を

表 2: 説明文の生成に用いたモデル

モデル	パラメータ数	アーキテクチャー
GPT2-small	約 1.1 億	デコーダーのみ
GPT2-base	約 3.3 億	デコーダーのみ
T5-small	約 6050 万	デコーダーのみ
T5-base	約 2.2 億	デコーダーのみ
BART-base	約 1.39 億	エンコーダーとデコーダー

$Dropout_{0.15}$ に入力し $\boldsymbol{e_s}$ を得る．

$$\boldsymbol{h_s''} = Linear(\boldsymbol{h_s'}) \tag{4}$$

$$\boldsymbol{h_s'''} = ReLU(\boldsymbol{h_s''}) \tag{5}$$

$$\boldsymbol{e_s} = Dropout_{0.15}(\boldsymbol{h_s'''}) \tag{6}$$

次に，メタデータの処理について説明する．まず，メタデータ m を $BERT$ に入力し $\boldsymbol{E_m} \in \mathbb{R}^{\mathbf{1\times 768}}$ を得る．ただし，$\boldsymbol{E_s}$ は $BERT$ の最終層の出力の [CLS] に対応するベクトルである．$\boldsymbol{E_m}$ を線形層 $Linear$ に入力し $\boldsymbol{h_m} \in \mathbb{R}^{\mathbf{1\times 64}}$ を得た後，$\boldsymbol{h_m}$ を活性化関数 $ReLU$ に入力し $\boldsymbol{h'_m}$ を得る．$\boldsymbol{h'_m}$ をドロップアウト率 0.21 のドロップアウト層 $Dropout_{0.21}$ に入力し $\boldsymbol{e_m}$ を得る．

$$\boldsymbol{E_m} = BERT(m) \tag{7}$$

$$\boldsymbol{h_m} = Linear(\boldsymbol{E_m}) \tag{8}$$

$$\boldsymbol{h'_m} = ReLU(\boldsymbol{h_m}) \tag{9}$$

$$\boldsymbol{e_m} = Dropout_{0.21}(\boldsymbol{h'_m}) \tag{10}$$

さらに，$\boldsymbol{e_s}$ と $\boldsymbol{e_m}$ を連結し $\boldsymbol{E_{concat-j}} \in \mathbb{R}^{1\times 96}$ を得る．$\boldsymbol{E_{concat-j}}$ に credit score を加え $\boldsymbol{e_{concat-j}}$ を得る．最後に，$\boldsymbol{e_{concat-j}}$ を線形層 $Linear$ に入力し得た $\boldsymbol{e'_{concat-j}} \in \mathbb{R}$ を $Sigmoid$ 関数に入力し l を得る．学習時，正解ラベル y と算出された l の cross entropy loss を損失として用いる．推論時，l の最大値に対応するラベルを予測ラベル y' とする．

5 実　　験

この章では，Research Question に対する実験と結果について述べる．Research Question は次の 4 つである．

RQ1：本論文の入力形式において，最も適切な説明文を生成するモデルは何か．

RQ2：提案手法の分類パフォーマンスは，他のモデルと比較してどのようなものであるか．

RQ3：提案手法は，生成された説明文に対してどれだけ faithful であるか．

RQ4：生成した説明文から抽出したキーワードは分類にどのような影響を持つか．

5.1 生成器の違いが生成される説明文に与える影響の分析

本実験では，異なる事前学習済み言語モデルを用いて説明文の生成を行った．実験で用いた事前学習済み言語モデルとその特徴を表 2 に示す．

表 2 にある 5 つのモデルの中で，BART-base のみが構成要素にエンコーダーを持つ．学習時のバッチサイズは 2 と 4 に，エポック数は 3 に設定した．また，生成される説明文の最大長を 256 に設定した．生成された説明文の評価には，zhang らにより提案された BERTScore [21] を用いた．BERTScore は，テキスト生成モデルの評価に用いられる指標の 1 つである．BERTScore は，BERT を用いて入力テキストの埋め込みを求め，求められた埋め込みを用いて入力テキストの類似度を比較する．また，BERTScore は BLEU や ROUGE 等の評価指標と比較して，人間による評価と強い相関を持つことが明らかになっている．

実験結果を表 3 と表 4 に示す．実験結果から，バッチサイズや生成する説明文を条件付けるラベルに関わらず，BART-base が生成する説明文の BERTScore が最も高くなることがわかった．これは，BART-base が GPT2 や T5 の持たないエンコーダーを構成要素として持つ為だと考えられる．また，GPT2 や T5 の生成した説明文には単語の繰り返しが多く含まれていたため，BERTScore が小さくなった．

以降の実験では，説明文の生成に BART-base を用いた．特に，ラベル true で条件付けられた説明文の生成にはバッチサイズ 4 で学習された BART-base を，ラベル false で条件付けられた説明文の生成にはバッチサイズ 2 で学習された BART-base を用いた．

5.2 分類パフォーマンスの比較

本実験では，提案手法と種々のモデルとの分類パフォーマンスを比較する．実験で用いたモデルとその特徴を表 5 にまとめる． 3.2 で紹介した ESM は，専門家の作成した質の高い説明文を用いて分類を行う．つまり，理想状態における分類を表す．評価には LIAR-PLUS のテストデータを用いた．評価指標として分類精度と F1 値を用いた．それぞれのモデルを，バッチサイズを 16，エポック数を 3 に設定して学習した．

分類結果を表 6 に示す．実験結果から，分類精度に関して提案手法がベースラインを上回っていることがわかる．また，F1 値に関して，提案手法がベースラインと ESM の双方を上回っている．これにより，生成器により自動生成された説明文が分類パフォーマンスの向上に寄与することがわかる．しかし，理想状態における分類と比較して分類精度が 0.073 ポイント低下している．これは，生成した説明文の質が専門家により作成された説明文の質と比較して不十分であるであると考えられる．

5.3 Sensitivity による提案手法の Faithfulness の検証

本実験では，Kumar らに倣い，生成された説明文に対する提案手法の faithfulness を検証する．よって，faithfulness を次のように定義する．

Faithfulness *: how well do the provided explanations correlate with the model' s decision making?*

faithfulness の評価には DeYoung らが考案した Sensitivity [22] を用いる．DeYoung らは，Sensitivity を次の 2 つの観点から定義した．

表 3: ラベル true で条件付けられた説明文の生成

バッチサイズ = 2

モデル	BERTScore
GPT2-small	0.623
GPT2-medium	0.600
T5-small	0.428
T5-base	0.611
BART-base	**0.742**

バッチサイズ = 4

モデル	BERTScore
GPT2-small	0.624
GPT2-medium	0.588
T5-small	0.515
T5-base	0.623
BART-base	**0.746**

表 4: ラベル false で条件付けられた説明文の生成

バッチサイズ = 2

モデル	BERTScore
GPT2-small	0.636
GPT2-medium	0.607
T5-small	0.390
T5-base	0.592
BART-base	**0.742**

バッチサイズ = 4

モデル	BERTScore
GPT2-small	0.624
GPT2-medium	0.588
T5-small	0.494
T5-base	0.660
BART-base	**0.735**

表 5: 分類パフォーマンスの比較に用いたモデル

モデル	入力特徴量	説明文の生成
ESM(理想状態における分類)	短文による主張，メタデータ，専門家により作成された説明文	×
ESM-j(ベースライン)	短文による主張，メタデータ	×
提案手法	短文による主張，メタデータ，自動生成された説明文	○

表 6: 分類パフォーマンスの比較

モデル	Accuracy	F1
ESM(理想状態における分類)	**0.829**	0.722
ESM-j(ベースライン)	0.565	0.687
提案手法	0.756	**0.730**

表 7: Sensitivity の確認

入力	Accuracy	F1
説明文	0.706	0.379
短文による主張，メタデータ	0.712	0.379
短文による主張，メタデータ, 生成された説明文	**0.756**	**0.730**

Comprehensiveness *: were all features needed to make a prediction selected?*
Sufficiency *: do the extracted rationales contain enough signal to come to a disposition?*

本実験では，入力として用いる特徴量の摂動が分類精度と F1 値に与える影響を分析した．実験結果を表 7 に示す．実験結果から，生成された説明文を利用しない場合，分類精度と F1 値がともに低下していることがわかる．これにより，自動で作成された説明文が分類パフォーマンスの向上に貢献していることがわかる．しかし，パフォーマンスの上がり幅は大きくない．今後，生成された説明文が分類にどのように貢献しているかを分析する必要がある．

表 8: キーワードを用いた分類

モデル	キーワード数	Accuracy
生成された説明文	-	**0.756**
キーワード	5	0.736
キーワード	10	0.742
キーワード	15	0.745

5.4 説明文から抽出したキーワードの有効性

本実験では，生成された説明文から抽出されたキーワードを用いた場合の分類を検証する．具体的には，抽出するキーワードの数を変化させ分類精度を観察する．キーワード抽出機として YAKE(**Y**et **A**nother **K**eyword **E**xtractor) [23] を用いる．YAKE は，Ricard らにより提案された教師無しキーワード抽出機である．YAKE は，まず，テキストに含まれる各単語に対して Casing，Word Positional，Word Frequency，Word Relatedess to context，Word DifSentence の計 5 つの特徴量を求める．次に，これらの特徴量から単語の重要度スコアを算出する．最後に，求めた重要度スコアが大きい順に単語を抽出する．学習時のバッチサイズは 4，エポック数は 3 に設定した．また，キーワードの抽出に際する n-gram 数を 3 に設定した．

実験結果を表 8 に示す．抽出するキーワード数を 5，10，15 とし分類精度を求めた．また，それらを生成された説明文を用いた場合の分類精度と比較している．実験結果から，キーワードを用いた場合の分類精度と比較して，説明文を用いた分類精度が高いことがわかる．これは，文章としての情報を持つ説明

文が分類に好ましいことを表す．しかし，分類精度の差が小さい分，生成された説明文の活用方法や生成される説明文の質を上げる為の手法に関する調査が必要である．

6 まとめと今後の展望

本論文では，短文による主張の真偽分類と分類に対する説明を生成するモデルを試作し提案した．提案手法は，分類結果に対する説明を自然言語による説明文で与える．

提案手法に対して，計 4 つの実験を行った．まず，実験 1 で様々な事前学習済み言語モデルをファインチューニングし，生成された説明文を BERTScore により評価した．実験の結果，BART-base が最も高く評価されることを明らかになった．次に，提案手法の分類パフォーマンスを種々のモデルと比較した．また，提案手法の faithfulness を Sensitivity により評価した．さらに，生成された説明文から抽出したキーワードを用いた分類を行い，結果を考察した．

実験により，自動生成された説明文が分類に効果的であることが明らかになった．しかし，説明文の生成に課題が見られる．まず，専門家が作成した説明文と同等の質の説明文を作成する手法を検討する必要がある．また，生成された説明文を BERTScore 等の自動評価だけではなく，人間により評価する必要がある．さらに，生成された説明文が説明として十分な知識を含んでいるかや，含まれる知識が正しいかを判断して学習するモデルを構築するためのアノテーションを持つデータセットの構築も必要である．分類においても，生成した説明文を適切に利用し，分類パフォーマンスを向上させる手法の調査が必要である．

謝　　辞

本論文は JSPS 科研費 JP22K12271，JP19K12230 の助成を受けたものである．

文　　献

[1] https://www.journalism.org/2018/09/10/news-use-across-\
\social-media-platforms-2018/.

[2] This analysis shows how viral fake election news stories out performed real news on facebook : Buzzfeednews16(2016).

[3] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[4] William Yang Wang. ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

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

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[7] Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*, 2020.

[8] Francisco Rangel and Paolo Rosso. Overview of the 7th author profiling task at pan 2019: bots and gender profiling in twitter. *Working notes papers of the CLEF 2019 evaluation labs*, 2380:1–7, 2019.

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

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[11] William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393, 2021.

[12] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

[13] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[14] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*, 2018.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

[17] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[18] Shaily Bhatt, Naman Goenka, Sakshi Kalra, and Yashvardhan Sharma. Fake news detection: Experiments and approaches beyond linguistic features. In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2021, Volume 2*, pages 113–128. Springer, 2022.

[19] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90, 2018.

[20] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[22] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[23] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

DEIM Forum 2024

誹謗中傷検出タスクの学習を補助するためのサブタスクの提案

沢田　凌一†　　鈴木　　優†

† 岐阜大学大学院自然科学技術研究科知能理工学専攻　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †sawada.ryoichi.g8@s.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　本研究の目的は二つある．一つ目は，マルチタスク学習を用いて誹謗中傷検出タスクの精度向上を図る際，我々が提案する主観サブタスクが精度向上に有効であることを明らかにすることである．二つ目は，我々が構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上の傾向があるか明らかにすることである．二つの目的を果たすために，我々は主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，使用するか使用しないの組み合わせ，計 16 種類のモデルを構築した．構築した全モデルを，F 値，Precision，Recall で比較した．また，検出精度に違いが生じたモデルに対して，どのような投稿で検出精度に違いが出たのか分析を行った．実験の結果，以下の知見が得られた．(1) 主観サブタスクを加えた MTL(マルチタスク学習モデル) は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある (2) 閲覧者感情分類タスクを加えた MTL は STL(シングルタスク学習モデル) に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効

キーワード　誹謗中傷，BERT，マルチタスク学習，ソーシャルメディア，機械学習

1　はじめに

本研究の目的は二つある．一つ目は，マルチタスク学習を用いて誹謗中傷検出タスクの精度向上を図る際，我々が提案する主観サブタスクが精度向上に有効であることを明らかにすることである．二つ目は，我々が構築したモデルがどのような誹謗中傷投稿に対して，検出精度向上の傾向があるか明らかにすることである．

一つ目の目的について述べる．マルチタスク学習 [1] とは，一つのモデルで複数のタスクを同時に解く手法である．同時に解く複数のタスクのうち，精度の向上を期待するタスクを主タスク，主タスクの学習補助の役割を果たすタスクをサブタスクと呼ぶ．モデルは，主タスクと関連のあるサブタスクを同時に解くことによって，一つのタスクにとらわれることのない特徴を獲得することができる．その結果，モデルの汎化性能向上が期待できる．マルチタスク学習は，様々な自然言語処理分野の研究 [2] [3] [4] において，精度向上に貢献している．

我々は誹謗中傷検出タスクにマルチタスク学習を採用し，検出精度の向上を図ろう考えた．しかし，誹謗中傷検出タスクを主タスクとした際，どのようなサブタスクを選定することが有効であるかは，著者が調べた限りでは明らかになっていない．選定したサブタスクが主タスクの精度向上に有効であるかは，データセットの収集からモデルを構築，テストデータで精度検証をするまで確認できないため，膨大な時間と労力がかかるという課題がある．我々は，課題解くためには主タスクである誹謗中傷検出タスクの精度向上に，有効なサブタスクの傾向を明らかにする必要があると考えた．我々は主観サブタスクが，主タスクである誹謗中傷検出タスクの精度向上に，有効な傾向があるのではないかという仮説を立てた．

主観サブタスクとは，第 1 者または第 2 者の観点からラベルを付与することができるタスクである．第 1 者とは投稿者，第 2 者とは閲覧者のことである．主観サブタスクが，誹謗中傷検出タスクの精度向上に有効な傾向があるのではないかと考えた理由を述べる．誹謗中傷投稿には，投稿者が悪意をもって行った投稿と，投稿者に悪気はないが閲覧者との感じ方の差異によって誹謗中傷となる投稿がある．そのため，投稿者の観点と閲覧者の観点は誹謗中傷を検出するために，重要な情報だと考えた．本研究では，主観サブタスクとして投稿者感情分類タスクと閲覧者感情分類タスクを選定した．また，主観サブタスクの有効性を検証するために比較対象として，客観サブタスクを用意した．客観サブタスクには，投稿対象分類タスクと誹謗中傷カテゴリ分類タスクを選定した．それぞれのサブタスクについては，3.1 節で述べる．

二つ目の目的について述べる．誹謗中傷には，煽り，皮肉，差別など様々な種類が存在する．構築したモデルの誹謗中傷検出精度が低下したとしても，特定の誹謗中傷投稿の検出に特化したモデルである可能性がある．例えば，全体的な誹謗中傷投稿の検出精度が低下しているが，差別表現を含んだ誹謗中傷投稿の検出精度は向上した場合である．我々は，どのようなサブタスクを加えたモデルが，どのような誹謗中傷投稿を検出できたのか明らかにすることによって，状況に応じたサブタスクの選定基準を明らかにできると考えた．

我々は，二つの目的を達成するために，サブタスクの提案，データセットの作成 (3.1 節)，モデルの構築，評価実験 (4 章) を行った．データセットの構築は，収集した投稿に対してクラウドソーシングを用いてラベルを付与した．ラベルは，各投稿それぞれに対するアンケートに 5 人ずつ回答し，多数決を取ることで決定した．作成したデータセットを用いて，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクに

ついて，使用するか使用しないの組み合わせ，計 16 種類のモデルを構築した．構築した全モデルを，F 値，Precision，Recall で比較した．また，検出精度に違いが生じたモデルに対して，どのような投稿で検出精度に違いが出たのか分析を行った．実験の結果，以下の 4 点を明らかにした．

（1） 主観サブタスクを加えた MTL(マルチタスク学習モデル) は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある
（2） 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効
（3） 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は STL(シングルタスク学習モデル) に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効

2 関 連 研 究

誹謗中傷投稿を自動検出するための研究は様々行われている．誹謗中傷投稿を検出するための研究には，誹謗中傷に関連する単語に着目し，誹謗中傷投稿を検出する手法と，文章全体に着目して誹謗中傷投稿を検出する手法がある．単語に着目した手法は，特定の単語を含んだ誹謗中傷投稿に対しての見逃しを高確率で抑えることができる．しかし，単語の登録数に限界があるため未知語を含んだ誹謗中傷投稿に対して脆弱であるというデメリットがある．単語に着目した研究として大友 [5] らの研究がある．大友らは，86,906 語の単語からなるいじめ表現辞書を作成した．彼らはいじめ表現辞書と複数の機械学習モデルを組み合わせてネットいじめ検出を行っている．その結果，彼らは高い分類評価を記録しているが，いじめ表現辞書への登録単語の少なさを課題に挙げている．単語に着目した手法の改善を試みた研究として Waseem ら [6] の研究がある．Wassem らは，誹謗中傷単語を特定するリストベースの手法は，リストに記載されていない単語に脆弱であることを問題視し，n-gram に基づく手法を提案している．その結果，性別に関する特徴を加えることが誹謗中傷検出精度向上に有効であることを示している．文章全体に着目した誹謗中傷投稿を検出する手法として Alatawi ら [7] の研究がある．Alatawi らは，ドメイン単語埋め込みと BiLSTM を組み合わせたモデルと，文章理解に優れた機械学習モデルである BERT の二つのモデルでヘイトスピーチの検出を行い，共にベースラインより高い検出精度を記録している．BiLSTM を用いたモデルは，スラングなど特定単語を用いたヘイトスピーチに強く，BERT を用いたモデルは，全体的な検出精度が高いことを示している．BERT を誹謗中傷単語で再トレーニングした研究 [8] も存在する．Caselli ら [8] は，学習済み BERT をポリシーに違反した投稿を使用して再トレーニングすることによって誹謗中傷検出タスクに特化した BERT を構築した．その結果，攻撃性のある文章を検出するタスクにおいて一般的な BERT を大きく上回る性能を獲得している．本研究のベースラインとなるシングルタスク学習モデルも BERT [9] を使用している．

マルチタスク学習は，画像処理 [10] や強化学習 [11]，自然言語処理 [4] など様々な分野で応用されている．Lamsiyah ら [4] は，マルチタスク学習を用いて文書分類タスク，ペアワイズテキスト分類タスク，テキスト類似度タスク，関連度ランキングタスクの四つのタスクを同時に学習するモデルを提案している．提案したマルチタスク学習モデルを用いた BERT モデルで大幅な精度向上に成功している．

類似した研究としてマルチタスク学習を用いて誹謗中傷検出タスクを解くモデルを提案した研究 [12] [13] がある．Samghabadi ら [12] は，攻撃性識別タスクと女性差別的攻撃性識別タスクを同列に扱いマルチタスク学習を行っている．その結果，女性差別攻撃性タスクの重み付き F 値で 15 チームのうち 3 位と高い結果を獲得している．この研究では，攻撃性識別タスクにマルチタスク学習を採用したという点で本研究と類似しているが，二つのサブタスクの精度を両方向上させようとしている．しかし，本研究では誹謗中傷検出タスクの精度向上を図るために有効性なサブタスクとして感情分類タスクを提案するため目的が違う．Aldjanabi ら [13] は，攻撃性検出データセットと三つのヘイトスピーチデータセットをマルチタスク学習で同時に学習するモデルを提案している．その結果，四つのデータセットのうち三つのデータセットで提案手法の有効性を示している．このように，類似したタスクを同時に解くことによって全てのタスクの検出精度向上を図る研究は存在する．しかし，誹謗中傷検出タスクの精度向上を図るためにサブタスクを提案し，BERT とマルチタスク学習を用いた研究は著者が調べた限りでなはい．そこで本研究では，誹謗中傷検出精度向上を目的に，誹謗中傷検出タスクモデルの学習を補助するためのサブタスクの提案を行う．

3 提 案 手 法

我々は，誹謗中傷検出タスクを解くモデルの学習を行う際，主観サブタスクを同時に解くことによって，モデルの学習を補助することができ，誹謗中傷検出精度が向上する傾向があるという仮説を立てた．仮説を検証するためには主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，使用するか使用しないの組み合わせ，計 16 種類のモデルの誹謗中傷検出精度を比較する必要がある．モデルを構築するためには，モデルの学習に必要なデータセットを作成する必要がある．そのため，我々は仮説を検証するためのデータセットを作成した．

ベースラインとなる誹謗中傷検出タスクのみを解くモデル (シングルタスク学習モデル) は，東北大学の乾・鈴木研究室が公開している事前学習済みの BERT モデル [1]を主タスクでファインチューニングすることによって構築した．マルチタスク学習モデルは，誹謗中傷検出タスクに主観サブタスク 2 種類と客観サブタスク 2 種類の組み合わせ全通りで加えたモデルを構築した．本章では，作成したデータセットについて 3.1 節，提案モデルについて 3.2 節で説明する．

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

3.1 データセット

データセットを構築するためには，仮説の検証に必要なタスクに応じた質問をし，アンケートを取ることによってラベルを付与する必要がある．本研究でデータセットを構築するために行ったアンケートの質問，質問に対応するタスク，回答項目を表 1 に記す．データセットのそれぞれのタスクについて説明する．

3.1.1 主タスク

本研究における主タスクは，誹謗中傷検出タスクとする．このタスクでは投稿がどのような対象への投稿であっても誹謗中傷であれば “誹謗中傷” とする．つまり自虐のように相手を傷つけることはなくても，誹謗中傷の表現を含んでいれば “誹謗中傷” のラベルが付与される．

3.1.2 主観サブタスク

主観サブタスクとは，第 1 者または第 2 者の観点からラベルを付与することができるタスクである．我々は，誹謗中傷検出タスクの精度向上には主観サブタスクを同時に学習することが有効であると考えた．主観サブタスクが誹謗中傷検出精度向上に有効であると考えた背景を説明する．誹謗中傷の定義は曖昧であるため，投稿について誹謗中傷投稿であると感じる人と，別の人は正当な批判と感じる人が存在する可能性がある．この投稿から受ける感じ方の差異は，特に投稿者と閲覧者によって生じることが多い [2]．そのため，我々はソーシャルメディア上の投稿が誹謗中傷かどうか判断するためには，無関係な第 3 者よりも投稿者や閲覧者の観点から考える情報が必要ではないかと考えた．本研究では，投稿者や閲覧者の観点から考えるタスクを主観サブタスクとして提案した．

投稿者感情分類タスクを主観サブタスクとして提案した理由を説明する．bot などの形式的にソーシャルメディアに投稿するものを除いたソーシャルメディアの投稿者は，何らかの感情を抱いて投稿を行う．誹謗中傷投稿者は，特に正義感や怒りの感情を抱いて誹謗中傷を投稿する傾向がある [14]．そのため我々は，投稿から読み取れる投稿者感情を明示的にモデルに伝えることができれば，誹謗中傷検出タスクの学習を補助することができるのではないかと考えた．投稿者感情が誹謗中傷検出の補助になると考えた例を以下に示す．

例 1　お前いい加減にしろ，殴るぞ笑笑
例 2　お前いい加減にしろ，殴るぞ !!

例 1 例 2 は，文末以外は “お前いい加減にしろ，殴るぞ” と相手に伝える投稿である．例 1 は，文末が “笑笑” であり，投稿相手のボケに対するツッコミのように感じる．そのため，投稿者は否定的な感情は持たず，投稿を行ったと考えることができる．例 2 は文末が “!!” であり “お前いい加減にしろ，殴るぞ” の文章を強調している．そのため，投稿者が投稿相手に対して強い怒りの感情を持っていると考えることができる．上記二つの例より，投稿者の感情が異なると投稿の伝わり方が違うことがわかる．そこで，我々は投稿者感情分類タスクを誹謗中傷検出タスクと同時に解くことによって，モデルが誹謗中傷投稿の特徴を捉える補助ができると考えた．

閲覧者感情分類タスクを主観サブタスクとして提案した理由を示す．前述したとおり，誹謗中傷の定義は曖昧であり，人によって誹謗中傷かそうでないかの判断は異なる．我々は，誹謗中傷の判断する際に重要なことは第 3 者よりも，被害を受ける閲覧者がどのように感じるかであると考えた．そのため，我々は投稿閲覧者が抱く感情は，誹謗中傷を検出する上で非常に重要な項目であると考えた．上記より，我々は，閲覧者感情分類タスクを誹謗中傷検出タスクと同時に解くことによって，モデルが誹謗中傷投稿の特徴を捉える補助ができると考えた．

3.1.3 客観サブタスク

客観サブタスクは，提案した 2 種類の主観サブタスクが有効であることを検証するための比較対象として用意した．投稿対象分類タスクとは，投稿を向けられた相手が回答項目にうちどれなのか分類するタスクである．組織とは，政党や国籍など集団を対象にした投稿である．投稿対象分類タスクをサブタスクとして提案した理由を例を用いて示す．

例 1　私は身長低いからね笑
例 2　お前は身長低いからね笑

例 1 も例 2 も文頭の “私”と “お前”以外は同じ文章である．文頭が “私”である例 1 の文は，何かに対しての言い訳をしているように感じられる．しかし，文頭が “お前”である例 2 の文は身長低い人を嘲笑しているように感じられる．このように，投稿対象が違うだけで攻撃的にも非攻撃的に感じられる．上記より我々は，投稿対象分類タスクは主タスクと関係のあるタスクであり比較対象として成立すると考えた．

誹謗中傷カテゴリ分類タスクは，誹謗中傷の種類を分割したカテゴリを分類するタスクである．誹謗中傷カテゴリは誹謗中傷に内包される．そのため，主タスクと関係のあるタスクであり比較対象として成立すると考えた．カテゴリには，誹謗中傷としてよく問題にあげられる “脅迫”，“差別”，“容姿否定” を選択した．誹謗中傷カテゴリは，そもそも誹謗中傷がある投稿にしか存在しない．そのため誹謗中傷検出タスクで “はい” と答えた場合のみ，このタスクは回答を行う．

3.1.4 データセット構築

データセット構築の手順を以下に記す．

（1） Twitter API を用いて投稿を収集
（2） クラウドソーシングを用いて収集した投稿一つずつに対して 5 人にアンケートを実施
（3） アンケートに答えた 5 人のワーカの票を多数決
（4） 多数決の結果最も票を集めた選択肢をラベルとして付与
（5） 票割れした投稿や投稿自体が理解できないものを削除

手順 1 の Twitter API を用いて収集した投稿は，2021 年 7 月 5 日から 2022 年 7 月 5 日の投稿のうち，我々が考えた攻撃性のある単語 36 種類のいずれかを含む投稿である．36 種類の単語を次に示す．

[“ば か”,“バ カ”,“馬 鹿”,“あ ほ”,“ア ホ”,“ブ ス”,“き も”,“キ

2：https://prtimes.jp/main/html/rd/p/000000201.000044347.html

表 1　データセット構築のための質問と対応タスク

質問内容	タスク	回答項目
Q1: 投稿は誹謗中傷ですか？	誹謗中傷検出タスク (主タスク)	はい，いいえ
Q2: 投稿者の感情はどれですか？	投稿者感情分類タスク (主観サブタスク 1)	喜び，怒り，悲しみ，好き，その他
Q3: 投稿から受ける印象はどれですか？	閲覧者感情分類タスク (主観サブタスク 2)	ポジティブ，ネガティブ，ニュートラル
Q4: どれに向けての投稿ですか？	投稿対象分類タスク (客観サブタスク 1)	自分，自分のもの，他人，他人のもの，組織，その他
Q5: 投稿はどれに当てはまりますか？	誹謗中傷カテゴリ分類タスク (客観サブタスク 2)	脅迫，差別，容姿否定，その他 (複数選択)

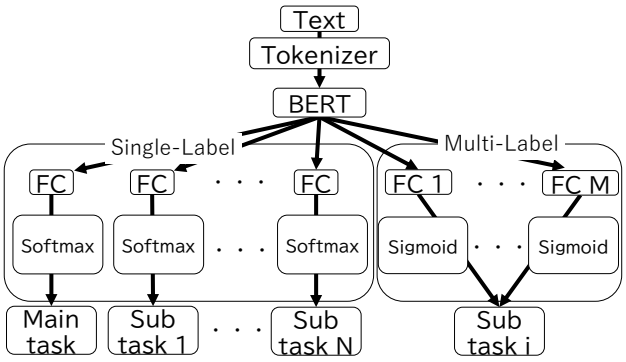


図 1　提案モデル図

モ”,“くず”,“クズ”,“しね”,“シネ”,“死ね”,“ゴミ”,“ごみ”,“雑魚”,“ざこ”,“ザコ”,“かす”,“カス”,“コロス”,“ころす”,“殺す”,“きえろ”,“キエロ”,“消えろ”,“むかつく”,“ムカつく”,“でぶ”,“デブ”,“はげ”,“ハゲ”,“無能”,“ダサ”,“いい加減にしろ”,“ウザ”]

手順 5 の票割れした投稿について説明する．票割れした投稿とは，多数決の結果最も票を集めた選択肢が 2 種類以上存在する投稿である．例えば，Q3 の 5 人へのアンケートの結果が“ポジティブ” が 2 票，“ネガティブ” が 2 票，“ニュートラル” が 1 票となった場合である．この場合，“ポジティブ” と “ネガティブ” は 2 票ずつ選択されており，どちらが正しいか判断できない．票割れした投稿は教師データとして曖昧なため削除した．Q5 は複数選択のため多数決は利用できない．そのため，3 人以上のワーカが選択した項目をラベルとして付与した．実際に作成したデータセットの各質問に対する回答の内訳は表 2 のようになった．このデータセットを用いてモデルを構築する．

3.2　提案モデル

3.2.1　提案モデル図

我々が用いたモデルを図 1 に表す．図 1 中の FC は全結合層を表し，N はサブタスクの総数，M はマルチラベルの選択肢の総数，$Subtask_i$ は i 番目のサブタスクを表している．提案モデルは，初めに入力テキストをトークンに分割し，BERT を用いて分散表現に変換する．その後はシングルラベルかマルチラベルかによってモデルが異なる．シングルラベルは全結合層と SoftMax 関数を用いて，どの分類項目にあてはまるかどうかの確率を求める．マルチラベルは BERT から得た分散表現をマルチラベルの選択肢の数だけ，全結合層と sigmoid 関数に通す．上記のモデルを構築することによって，マルチラベルの一つ一つの分類項目が当てはまるかどうかを学習することができる．提案モデルを使用することによって主タスクとサブタスクを同時に学習する．

3.2.2　損失関数

モデルの学習に用いた損失関数は以下の式で表される．

$$AllLoss = MainLoss + \frac{2}{3}\sum_{i=1}^{N} SubLoss_i$$

$AllLoss$ は主タスクとサブタスクの損失の合計を表しており，モデルが逆伝搬する損失である．$MainLoss$ は主タスクの損失を表しておりクロスエントロピー誤差である．N は同時に学習するサブタスクの総数を表している．$SubLoss_i$ は i 番目のサブタスクの損失を表しており，サブタスクがシングルラベルの時はクロスエントロピー誤差，マルチラベルの時は，各出力に対してのバイナリークロスエントロピー誤差の合計である．本研究では，サブタスクを主タスクの学習の補助として扱い，主タスクを優先して学習させるため，サブタスクの損失の合計に 3 分の 2 をかける．

4　評価実験

本研究の実験の目的は二つある．一つ目は，主観サブタスクを加えたマルチタスク学習モデルが，誹謗中傷検出精度向上に有効かどうか検証することである．我々は検証のために，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，使用するか使用しないの組み合わせ，計 16 種類のモデルの評価指標を比較した．二つ目は，実験の結果優れたモデルが，どのような誹謗中傷投稿を検出可能になったのか，傾向を調査することである．我々は検証のために，全タスクの各ラベルに対応するデータごとの Accuracy を算出し，比較を行った．モデルを構築するために作成したデータセットの分析について 4.1 節，一つ目の検証について 4.2.1 節，二つ目の検証について 4.2.2 節で述べる．

4.1　データセット分析

作成したデータセットについて考察する．主タスクの回答は

表 2　各質問に対する回答の内訳

タスク	全 13,448 件の質問に対する回答のデータ数									
主タスク (Q1)	誹謗中傷	3,894	誹謗中傷ではない	9,554	-	-	-	-	-	-
主観サブタスク 1(Q2)	喜び	1,029	怒り	6,158	悲しみ	793	好き	986	その他	4,482
主観サブタスク 2(Q3)	ポジティブ	2,676	ネガティブ	9,457	ニュートラル	1,315	-	-	-	-
客観サブタスク 1(Q4)	自分, 自分のもの	822	他人, 他人のもの	10,817	組織	885	その他	924	-	-
客観サブタスク 2(Q5)	脅迫	1,160	差別	176	容姿の否定	267	その他	649	誹謗中傷ではない	9,554

“誹謗中傷である” が 7 割をしめる結果となった．攻撃性のある単語を含む投稿を集めたが “バカ面白い笑” など，攻撃性のない表現に用いられることも多くこのような結果となった．主観サブタスク 1 の回答は “怒り” が一番多くの票を集めたが総投票数の 5 割を下回り，票が割れる結果となった．上記より，攻撃性のある単語を含む投稿には様々な感情が混在していることがわかる．主観サブタスク 2 は，“ネガティブ” の投稿が 7 割をしめる結果となった．この結果は，攻撃性のある単語を含む投稿を集めたため予想できる．主観サブタスク 2 と主タスクの結果から “ネガティブ” と感じるハードルは低いが，“誹謗中傷” と感じるハードルは高いことを確認した．客観サブタスク 1 の回答は “他人, 他人のもの” が 8 割をしめる結果となった．上記より，攻撃性のある単語を含む投稿は他人に向けた投稿が多いということを確認した．客観サブタスク 2 は主タスクで “誹謗中傷ではない” と選択された投稿が多かったため，それぞれのカテゴリへの投票は少なかった．客観サブタスク 2 の選択肢の中では “脅迫” のカテゴリが一番票を集める結果となった．

4.2　実験結果・考察

我々は，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，追加するか追加しないかの組み合わせ，計 16 種類のモデルを構築した．我々は，データセットを訓練用: 検証用: テスト用=8:1:1 の割合に分割し，訓練用データを用いでモデルの学習を行った．我々は，検証用データの損失が 100 エポック更新されなかった場合，損失が最小の時のモデルを採用した．その後，テスト用データで 10 分割交差検証を行った．

構築した 16 種類のモデルの実験結果を表 3 に記す．STL とは，誹謗中傷検出タスクのみを解くシングルタスク学習モデルであり，MTL とは主タスクとサブタスクを同時に解くマルチタスク学習モデルである．括弧内の数字は，学習したタスク質問番号である．評価指標は，“誹謗中傷” を陽性として F: 主タスクの陽性の F 値，Precision: 主タスクの陽性の適合率，Recall: 主タスクの陽性の再現率，macro-F1(Q2): 主観サブタスク 1 のマクロ平均 F1，macro-F1(Q3): 主観サブタスク 2 のマクロ平均 F1，macro-F1(Q4): 客観サブタスク 1 のマクロ平均 F1 macro-F1(Q5): 客観サブタスク 2 のマクロ平均 F1 である．各評価指標の下に記載されている括弧内の数値は，帰無仮説を “STL の評価指標と比較対象である MTL の評価指標の平均値に差はない” とし，2 群の対応のある t 検定を行った際の P 値である．黒字は，各評価指標において最も精度が高い数値である．

4.2.1　全モデルの評価指標比較

本節では，主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，使用するか使用しないの組み合わせ，計 16 種類のモデルの評価指標を比較について述べる．評価指標比較の目的は，主観サブタスクを加えたマルチタスク学習モデルが，誹謗中傷検出精度向上に有効かどうか検証することである．

a）　F 値の比較

STL の F 値と MTL の F 値を比較する．STL の F 値と比べて MTL の F 値が向上したモデルは八つである．F 値が向上したモデルのうち，最も F 値が向上したモデルは，STL よりも 0.016 向上した MTL(Q1,3) である．MTL(Q1,3) は主観サブタスク 2 の閲覧者感情分類タスクを加えたモデルである．主観サブタスクを加えたモデルの F 値向上は，我々の仮説立証に対して望ましい結果である．しかし，表 3 括弧内に記載している MTL(Q1,3) の p 値は 0.274 である．そのため，我々は仮説の立証には更なる精度向上が必要であると考える．

構築した 16 種類のモデルの F 値を比較する．F 値上位の三つのモデルは，上から順に MTL(Q1,3)，MTL(Q1,2)，MTL(Q1,2,3) である．F 値上位の三つのモデルは，全て主観サブタスクのみを加えたモデルである．F 値下位の三つのモデルは，MTL(Q1,5)，MTL(Q1,2,3,5)，MTL(Q1,4,5) である．F 値下位の三つのモデルのうち二つは，客観サブタスクのみを加えたモデルである．16 種類のモデルの F 値を比較するために，帰無仮説を “モデルの F 値の平均値に差はない” とし，16 種類のモデル総当たりで 2 群の対応のある t 検定を行った．有意水準 0.05 を基準とすると，MTL(Q1,3) と MTL(Q1,4,5) の P 値 0.026 と，MTL(Q1,3) と MTL(Q1,5) の P 値 0.021 が有意水準を下回った．この結果から，主観サブタスク 2 の閲覧者感情分類タスクをサブタスクとして，誹謗中傷検出タスクに加えることは，(1) 客観サブタスク 4 の投稿対象分類タスクと客観サブタスク 5 の誹謗中傷カテゴリ分類タスクを加えることよりも F 値向上に有効 (2) 客観サブタスク 5 の誹謗中傷カテゴリ分類タスクを加えることよりも F 値向上に有効であることが分かった．(1)(2) はどちらも客観サブタスクよりも，主観サブタスクの方が F 値向上に有効であることを示している．そのため，マルチタスク学習を用いて誹謗中傷検出タスクの学習を行う際には，客観サブタスクよりも主観サブタスクを採用する方が F 値向上に繋がると考えることができる．

b）　Precision の比較

STL の Precision と MTL の Precision を比較する．STL の Precision と比べて MTL の Precision が向上したモデルは七つ

表 3 実験結果

	F 値	Preccision	Recall	macro-F1(Q2)	macro-F1(Q3)	macro-F1(Q4)	macro-F1(Q5)
STL(Q1)	0.651	0.668	0.636	-	-	-	-
MTL(Q1,2)	0.657 (0.481)	0.664 (0.674)	0.653 (0.368)	0.469	-	-	-
MTL(Q1,2,3)	0.655 (0.555)	0.670 (0.868)	0.645 (0.635)	0.483	0.856	-	-
MTL(Q1,2,3,4)	0.654 (0.488)	0.665 (0.829)	0.647 (0.426)	**0.486**	0.847	0.343	-
MTL(Q1,2,3,4,5)	0.652 (0.856)	0.662 (0.328)	0.646 (0.457)	0.481	**0.859**	0.298	0.316
MTL(Q1,2,3,5)	0.647 (0.434)	0.670 (0.801)	0.628 (0.450)	0.482	0.857	-	0.295
MTL(Q1,2,4)	0.650 (0.806)	0.662 (0.727)	0.641 (0.623)	0.471	-	0.348	-
MTL(Q1,2,4,5)	0.652 (0.872)	0.653 (0.100)	0.653 (0.331)	0.482	-	0.358	0.353
MTL(Q1,2,5)	0.650 (0.827)	0.654 (0.070)	0.648 (0.335)	0.477	-	-	0.285
MTL(Q1,3)	**0.667** (0.274)	**0.698** (0.033)	0.640 (0.739)	-	0.823	-	-
MTL(Q1,3,4)	0.650 (0.875)	0.683 (0.470)	0.627 (0.613)	-	0.821	0.322	-
MTL(Q1,3,4,5)	0.654 (0.595)	0.654 (0.087)	**0.656** (0.176)	-	0.823	0.330	0.253
MTL(Q1,3,5)	0.651 (0.968)	0.668 (0.973)	0.639 (0.824)	-	0.808	-	0.257
MTL(Q1,4)	0.653 (0.808)	0.676 (0.315)	0.633 (0.796)	-	-	**0.371**	-
MTL(Q1,4,5)	0.648 (0.648)	0.665 (0.716)	0.634 (0.789)	-	-	0.333	0.310
MTL(Q1,5)	0.644 (0.095)	0.669 (0.923)	0.625 (0.288)	-	-	-	**0.544**

である．Precision が向上したモデルのうち，最も Precision が向上したモデルは，STL よりも 0.030 向上した MTL(Q1,3) である．有意水準 0.05 とすると，STL と MTL(Q1,3) を比較した際の P 値は 0.033 であり下回っている．この結果から，主観サブタスク 2 の閲覧者感情分類タスクを，サブタスクとして誹謗中傷検出タスクの学習に加えることは，Precision の向上に有効であると考えることができる．

構築した 16 種類のモデルの precision を比較する．precision 上位の三つのモデルは，上から順に MTL(Q1,3)，MTL(Q1,3,4)，MTL(Q1,4) である．precision 下位の三つのモデルは，MTL(Q1,2,4,5)，MTL(Q1,3,4,5)，MTL(Q1,2,5) である．比較の結果，最も精度が高い MTL(Q1,3) が他の 6 種類のモデルに対して有意水準 0.05 を下回る p 値を記録した．MTL(Q1,Q3) が他のモデルに比べ，Precision の向上に有効であることを確認した．

c） Recall の比較

STL の Recall と MTL の Recall を比較する．STL の Recall と比べて MTL の Recall が向上したモデルは 10 個である．Recall が向上したモデルのうち，最も Recall が向上したモデルは，STL よりも 0.020 向上した MTL(Q1,3,4,5) である．STL の Recall と比べて MTL の Recall が低下したモデルは五つである．Recall が低下したモデルのうち三つは，客観サブタスクのみをサブタスクとして加えた MTL である．この結果から我々は，STL との有意差は小さいが，客観サブタスクが Recall の向上には繋がらない傾向を確認した．

構築した 16 種類のモデルの Recall を比較する．Recall 上位の三つのモデルは，上から順に MTL(Q1,3,4,5)，MTL(Q1,2)，MTL(Q1,2,4,5) である．Recall 下位の三つのモデルは，MTL(Q1,5)，MTL(Q1,3,4)，MTL(Q1,2,3,5) である．16 種類のモデルの Recall を比較するために，F 値の比較と同様に t 検定を行った．有意水準を 0.05 とすると，MTL(Q1,2,4,5) と MTL(Q1,3,4) の P 値 0.049 が有意水準を下回った．この結果から，主観サブタスク 2 と客観サブタスク 1,2 をサブタスクとして，誹謗中傷検出タスクに加えることは，主観サブタスク 2 と客観サブタスク 1 を加えることより Recall 向上に有効であることを確認した．

a），b），c）より，我々は以下の 2 点を明らかにした．

（1） 主観サブタスクを加えた MTL は，客観サブタスクを加え

た MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある

（2） 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効

4.2.2 検出精度に違いが生じた投稿に対する考察

本節では，全タスクの各ラベルに対応するデータごとの Accuracy 比較について述べる．この比較の目的は，実験の結果優れたモデルがどのような誹謗中傷投稿を，検出可能になったのか分析することである．4.2.1 節で前述したとおり，F 値が最も高いモデルは MTL(Q1,3)，最も低いモデルは MTL(Q1,5) である．我々は，F 値に最も差が開いた MTL(Q1,3) と MTL(Q1,5) の比較，F 値が最も向上した MTL(Q1,3) とベースラインである STL(Q1) の比較を行った．我々は全てのタスクのどのラベルが付与されたデータの検出精度に，大きな変化が起きたのか調査した．調査を行うことによって，我々が構築したモデルがどのような投稿の検出精度に向上に有効，または無効なのか検証することができると考えた．我々は検証のために，以下の式で全タスクの各ラベルに対応するデータごとの Accuracy を算出した．

$$A_{i,j} = \frac{N^c_{i,j}}{N_{i,j}}$$

i は質問番号，j は付与されるラベルである．$A_{i,j}$ は，質問 Qi に j のラベルが付与されたデータの Accuracy である．$N^c_{i,j}$ は，質問 Qi に j のラベルが付与されたデータのうち，誹謗中傷検出タスクを正解したデータ数である．$N^c_{i,j}$ は，質問 Qi に j のラベルが付与されたデータの総数である．我々は，帰無仮説を“比較する二つのモデルの $A_{i,j}$ の平均値に差はない” とし，2 群の対応のある t 検定を行った．検定は 10 分割交差検証で行い，有意水準は 0.050 とした．

a） MTL(Q1,3) と STL(Q1) の比較

MTL(Q1,3) と STL(Q1) の $A_{i,j}$ を算出し，2 群の対応のある t 検定を行った．検定の結果，Q2(投稿者感情分類するタスクに対応する質問) に対して，“喜び” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.989 と STL(Q1) の Accuracy :0.980 の間に p 値 0.009 の有意差を確認した．この結果から，MTL(Q1,3) は STL(Q1) より喜びの投稿に対する誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“喜び” のラベルが付与されたデータにおいて，STL(Q1) では，誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 コロス気かっ！Thank You〜(正解ラベル: 誹謗中傷ではない)

例 2 え !? マジじゃん え !? やったーーーーーー !!!!! 消した !!!!! 死ね !!!!! (該当ツイートに対して)(正解ラベル: 誹謗中傷)

例 1 は，おそらく冗談に対するツッコミで “コロス気かっ！” という表現を用いており，後半は “Thank You ” と感謝を述べているため誹謗中傷の投稿ではない．我々は，STL(Q1) は前半に着目して例 1 の投稿を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 の前半は “え !? マジじゃん え !? やったーーーーーー !!!!! ” であり，投稿者は喜びの感情を抱いていることが確認できる．例 2 の後半は “消した !!!!! 死ね !!!!! (該当ツイートに対して)” と攻撃性の高い言葉が並んでいるため例 2 の投稿には “誹謗中傷” のラベルが付与されている．我々は，STL(Q1) は，前半に着目して “誹謗中傷ではない” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷” と分類することができた．上記の例より，投稿に “ポジティブ” な言葉と “ネガティブ” な言葉が混在していると，STL(Q1) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考えることができる．

b） MTL(Q1,3) と MTL(Q1,5) の比較

MTL(Q1,3) と MTL(Q1,5) の $A_{i,j}$ を算出し，2 群の対応のある t 検定を行った．検定の結果，Q2(投稿者感情分類タスクに対応する質問) に対して，“怒り” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.717 と MTL(Q1,5) の Accuracy :0.709 の間に p 値 0.030 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より投稿者が怒りの感情を持った投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“怒り” のラベルが付与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す．

例 1 死ね !? なんてこと言うの！(正解ラベル: 誹謗中傷ではない)

例 2 お前に何がわかるん？そろそろいい加減にしろよまじで w(正解ラベル: 誹謗中傷)

例 1 は，誰かから “死ね” と言われたことに対して，“なんてことを言うの！” と怒っており，誹謗中傷の投稿ではない．我々は，MTL(Q1,5) は “死ね” という単語に着目して，例 1 を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 は，相手を煽りながら “いい加減しろよ” と怒っており，誹謗中傷の投稿である．我々は，MTL(Q1,5) は文末の “w” という表現に着目して，例 1 を “誹謗中傷ではない” と分類したのではないかと考える．一方，MTL(Q1,3) は例 2 の投稿を “誹謗中傷” と分類することができた．上記の例より，注意や相手に疑問を呼びかけるような投稿においては，MTL(Q1,5) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考えることができる．

Q4(投稿者対象分類タスクに対応する質問) に対して，“他人, 他人のもの” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.810 と MTL(Q1,5) の Accuracy :0.805 の間に p 値 0.005 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より他人や他人のものに向けた投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“他人, 他人のもの” のラベルが付

与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す.

例 1　かわいいじゃないかー !!! 死ねなんていっちゃメッ !!! (正解ラベル: 誹謗中傷ではない)

例 2　ブスが同じダンスやってみろよ。。。ある意味流行るかも (正解ラベル: 誹謗中傷)

例 1 は，誰かが “死ね” と言ったことに対して，“死ねなんていっちゃメッ !!! ” と注意しており，誹謗中傷の投稿ではない．我々は，MTL(Q1,5) は “死ね” という単語に着目して，例 1 を “誹謗中傷” と分類したのではないかと考える．一方，MTL(Q1,3) は例 1 の投稿を “誹謗中傷ではない” と分類することができた．例 2 は，“ブスが” や “ある意味” など抽象的な言葉を用いて誰かを皮肉しており，誹謗中傷の投稿である．我々は，MTL(Q1,5) は皮肉の表現を認識することができずに，例 1 を “誹謗中傷ではない” と分類したのではないかと考える．上記の例より，注意や皮肉の投稿においては，MTL(Q1,5) より MTL(Q1,3) の方が正しく分類しやすいのではないかと考えることができる．

Q5(投稿者対象分類タスクに対応する質問) に対して，“差別” のラベルが付与されたデータで，我々は，MTL(Q1,3) の Accuracy:0.415 と (Q1,5) の Accuracy :0.378 の間に p 値 0.033 の有意差を確認した．この結果から，MTL(Q1,3) は MTL(Q1,5) より差別の投稿に対する，誹謗中傷検出タスクの Accuracy 向上に，有効であることが確認できる．“差別” のラベルが付与されたデータにおいて，MTL(Q1,5) では誹謗中傷検出タスクの予測を誤ったが，MTL(Q1,3) では予測が正しかった投稿の例を以下に示す.

例 1　この国は無能な政治家によって運営されています。その多くは○○人だそうです。納得しました。(正解ラベル: 誹謗中傷)

例 2　多様性なんかいらんやろ。○○いい加減にしろ。倒産してほしい (正解ラベル: 誹謗中傷)

Q5 は誹謗中傷カテゴリ分類タスクのため，例の投稿は誹謗中傷の投稿のみである．例 1 例 2 の○○には具体的な言葉が入っていたが，本稿では○○と置き換えている．例 1 は政権批判，例 2 は会社批判であり，どちらも差別的な表現が含まれている．例より，MTL(Q1,5) の誹謗中傷カテゴリ分類のサブタスクが，差別表現の誹謗中傷検出タスクの Accuracy 向上の妨げになったと考えられる．脅迫表現と差別表現は，同じ誹謗中傷カテゴリだが，誹謗中傷検出タスクの Accuracy の向上に対しては，反対の影響を及ぼすことを確認した．この結果から，我々はマルチタスク学習においてサブタスクを追加した際，同じ分類タスクであっても回答によっては，精度向上に大きな差が生まれることを確認した.

a)，b) より，我々は以下の 2 点を明らかにした.

（1） 閲覧者感情分類タスク (主観サブタスク 2) を加えた MTL は STL に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効

（2） MTL の F 値 (主タスク) に有意な差があった場合でも，加えるサブタスクによって，特定の投稿の検出に特化することがある

5 おわりに

我々は，誹謗中傷検出タスクを解くモデルの学習を行う際，主観サブタスクを同時に解くことによって，モデルの学習を補助することができ，誹謗中傷検出精度が向上するという仮説を立てた．仮説の検証を行うために，我々は主観サブタスク 2 種類と客観サブタスク 2 種類の計 4 種類のサブタスクについて，使用するか使用しないの組み合わせ，計 16 種類のモデルの誹謗中傷検出精度を比較した．比較の結果，以下の四つの知見を得た.

（1） 主観サブタスクを加えた MTL は，客観サブタスクを加えた MTL に比べ，誹謗中傷検出タスクの F 値向上に有効な傾向がある

（2） 閲覧者感情分類タスクを加えた MTL は，誹謗中傷検出タスクの Precision 向上に有効

（3） 閲覧者感情分類タスクを加えた MTL は STL に比べ，投稿者が喜びの感情を抱いた投稿に対する誹謗中傷検出タスクの Accuracy 向上に有効

今後の展望として我々は，サブタスクと実験回数を増やすことによる更なる傾向の調査や，学習時におけるサブタスクへの適切なバイアスを明らかにすることが必要であると考える.

謝辞　本研究の一部は，JSPS 科学研究費 (19H04218, 23H03694)，越山科学技術振興財団，およびソニー株式会社の助成を受けたものです.

文　献

[1] Rich Caruana. Multitask learning. *Machine learning*, Vol. 28, No. 1, pp. 41–75, 1997.

[2] Qiwei Bi, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Hanfang Yang. Mtrec: Multi-task learning over bert for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2663–2669, 2022.

[3] Xuyang Wu, Alessandro Magnani, Suthee Chaidaroon, Ajit Puthenputhussery, Ciya Liao, and Yi Fang. A multi-task learning framework for product ranking with bert. In *Proceedings of the ACM Web Conference 2022*, pp. 493–501, 2022.

[4] Salima Lamsiyah, Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Bernard Espinasse. Unsupervised extractive multi-document summarization method based on transfer learning from bert multi-task fine-tuning. *Journal of Information Science*, Vol. 49, No. 1, pp. 164–182, 2023.

[5] 大友泰賀, 張建偉, 中島伸介, 李琳. いじめ表現辞書を用いた twitter 上のネットいじめの自動検出. *DEIM2020, C7-1, day2, p22*, 2020.

[6] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research*

workshop, pp. 88–93, 2016.

[7] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, Vol. 9, pp. 106363–106374, 2021.

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

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

[10] Xiangtao Zheng, Tengfei Gong, Xiaobin Li, and Xiaoqiang Lu. Generalized scene classification from small-scale datasets with multitask learning. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–11, 2021.

[11] Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *International conference on machine learning*, pp. 166–175. PMLR, 2017.

[12] Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 126–131, 2020.

[13] Wassen Aldjanabi, Abdelghani Dahou, Mohammed AA Alqaness, Mohamed Abd Elaziz, Ahmed Mohamed Helmi, and Robertas Damaševičius. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, Vol. 8, p. 69. MDPI, 2021.

[14] 山口真一. 炎上加担動機の実証分析. 2016 年社会情報学会 (SSI) 学会大会, 2016.

ソーシャルメディア投稿における炎上推定のための世論に対する乖離度算出手法

吉田 裕輔† 鷹野 孝典‡

†神奈川工科大学大学院工学研究科情報工学専攻 〒243-0292 神奈川県厚木市下荻野 1030
‡神奈川工科大学情報工学科 〒243-0292 神奈川県厚木市下荻野 1030
E-mail: †s2285011@cco.kanagawa-it.ac.jp, ‡takano@ic.kanagawa-it.ac.jp

あらまし 本研究では，ソーシャルメディア上の投稿が，多様な視点で構成される意見集合に対してどの程度乖離しているかを算出する手法を提案する．ソーシャルメディア上の投稿において炎上防止は重要な課題となっている．しかし，悪意がない投稿であっても，世論に受け入れられないなどの状況依存の理由で炎上する場合があるため，炎上予測は困難である．提案手法の特徴は，世論を多様な視点から構成される意見集合として構築し，意見集合間の対立の大きさに基づいて投稿された意見と世論の乖離度を算出する点にある．本研究では，大規模言語モデルを用いて作成した擬似世論データセットを用いた実験により，提案手法の実現可能性を評価する．

キーワード 炎上推定モデル，大規模言語モデル，疑似データセット，多様な意見集合，意見対立度

1. はじめに

ソーシャルメディア[1][2][3]の普及により，個人が情報を共有する機会が増える一方で，ソーシャルメディア上に投稿された意見やコメントをめぐる炎上現象が社会問題化している．炎上によるリスクは，社会的信用の失墜による事業撤退から，個人情報の流出といった個人攻撃まで様々である．このような背景から，ソーシャルメディア上の投稿が炎上しないようにすることは重要な課題であり，特定の炎上に着目した様々な炎上検知手法が提案されている[4][5][6].

しかし，投稿に悪意が含まれていない場合でも，その投稿が世論に受け入れられないといった文脈依存的な理由で炎上することもあり，炎上するか否かを予測することは非常に困難である．このような意図せざる炎上を引き起こす要因の一つとして，人々の意見の多様性により,個人の価値観や認識が異なり,その結果，投稿意見と世論との間に考えや認識の乖離が生じる可能性があると考えられる．そのため，投稿意見では問題がないように見える意見であっても，様々な意見から構成される世論から生じる対立度を判定する手法の開発が大きく求められている．

本研究では，ソーシャルメディア上の投稿が，多様な視点で構成される意見集合に対してどの程度乖離しているかを算出する手法を提案する．提案手法の特徴は，世論を多様な視点から構成される意見集合として構築し，意見集合間の対立の大きさに基づいて投稿された意見と世論の乖離度を算出する点にある．これにより，提案手法により，投稿前に世論がどの程度適切に意見を受け入れるかを判断し，炎上リスクを低減することが期待できる．

本研究では，大規模言語モデル(Large Language Model, LLM)を用いて作成した擬似世論データセット[17]を用いた実験により，提案手法の実現可能性を評価する．

2. 関連研究

これまで炎上検知の手法が提案されており，投稿数，差別・誹謗中傷表現，投稿日時，投稿者に着目した研究などがある[4][5][7][8][9][10][11].

大曽根らは，Web 上に出現するコメント数の時間推移から炎上状況を検出し，炎上規模を定量化する手法を提案している[4]．酒井らは，SVM（Support Vector Machine），決定木，MLP（Multi Layer Perceptron）で学習させた予測モデルを用いて，YouTube のコメントを炎上検知に利用し，炎上状況を判定する手法を提案している[5]．川上らは，投稿を炎上投稿と非炎上投稿に分類し，炎上投稿に出現する単語で学習した Word2Vec[6]を適用して，対象の投稿が炎上するかどうかを判定している[7].

文献[8]において，吉田らは，人手で収集した攻撃的な文章群を用いて学習させたモデルを用いて，攻撃的な文章を判定する手法を提案している．高橋らは，各投稿に出現する単語の依存関係や顔文字に基づいて作成した感情辞書をもとに，肯定極性と否定極性を計算することで炎上予測を行う手法を提案した[9]．瀬川らは，攻撃ユーザの特徴を分析し，攻撃者の特徴や攻撃者の関係性から攻撃タイプを分類した[10]．岩崎らは，投稿のタイムラインに対する肯定的な反応と否定的な反応の比率を計算することで，炎上が生じているか否かを判定する方法を提案している[11].

しかし，これらの従来研究では，悪意や差別・誹謗中傷表現を含まない場合においても，意見投稿前に炎

上の可能性を予測するという課題を十分に解決できていない．

3. 提案手法

炎上の要因は多岐にわたるため，一つの要因に着目するだけでは炎上推定は困難である．本研究では，炎上要因として乖離度，表現，話題性，話題の伝播速度の 4 つの要因に着目した炎上推定モデルの実現を目指している(図 1)．図 1 において「乖離度」は投稿された意見と世論との乖離度，「表現」は差別的表現や誹謗中傷表現など炎上の原因となる表現，「話題性」は話題の社会的影響度，「伝播速度」は各ソーシャルメディア上での拡散のしやすさを表している．本研究では，「乖離度」に着目し，ソーシャルメディア上の投稿が多様な視点からの意見の集合からどの程度乖離しているかを算出する手法を提案する．

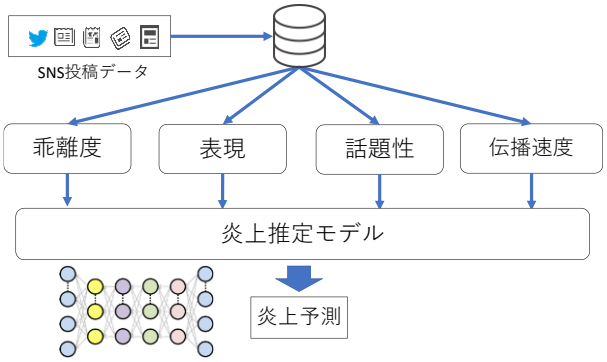


図 1 複数の要因に着目した炎上推定モデル

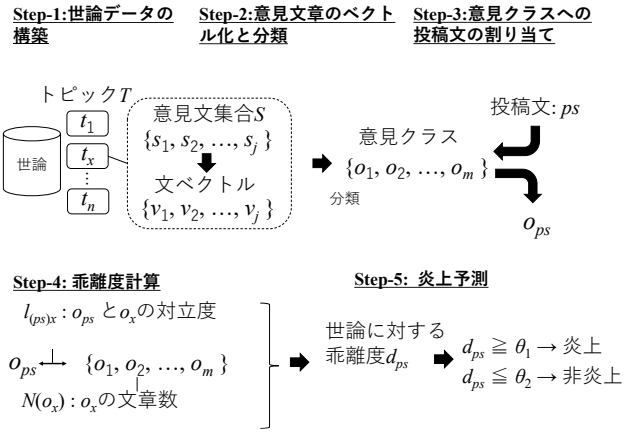


図 2 提案手法の概要図

図 2 に提案手法の概要を示す．本手法は以下の Step-1 から Step-5 の手順で実行される．

Step-1 様々な視点で構成される世論を表す意見データセットの構築

先行研究で示した疑似世論データセット[17]のスキーマを用いて，様々な視点からの世論を表す意見データセットを構築する．世論は複数のトピック集合 T から形成されると仮定する．

$$T = \{ t_1, t_2, \ldots, t_n \} \quad (1)$$

ここで，各 t_x は，トピック t_x に対する様々な視点を表す意見クラス集合 O を持つ．さらに，各意見クラス o_y は，意見文章からなる意見文章集合 S を持つ．

$$O = \{ o_1, o_2, \ldots, o_m \} \quad (2)$$
$$S = \{ s_1, s_2, \ldots, s_j \} \quad (3)$$

Step-2 意見文章データセットのベクトル化と分類

BERT （Bidirectional Encoder Representations from Transformers）[15]などで学習した用語や文の埋め込み表現を用いて，意見文章を文章ベクトル V にベクトル化する．

$$V = \{ v_1, v_2, \ldots, v_j \} \quad (4)$$

文章ベクトルは，k-means [14]などの分類手法にコサイン尺度などを適応して m 個のグループに分類される．分類後，分類された m 個のグループは Step-1 で定義した意見クラス o_y に割り当てられる．各意見クラス間の対立度 L は下記のように表される．

$$L = \{ l_{12}, l_{13}, l_{23}, l_{24}, \ldots, l_{pq} \} \quad (5)$$

ここで，l_{pq} は意見クラス o_p と o_q の対立度である．

Step-3 投稿文の意見クラスへの割り当て

投稿文 ps を対応する意見クラス o_p に割り当てるために，投稿文とグループ内に属する各文 s との間の類似度をコサイン尺度などを用いて計算する．

$$Similarity\ (ps, s) = (ps \cdot s)/|ps||s| \quad (6)$$

例えば，k 近傍法を適応し，投稿文 ps に類似する意見クラスの文の数に応じて，投稿文 ps に対する意見クラス o_{ps} を推定することができる．

$$Assignment\ (ps, O, k) = \text{argmax}\ (N_{ps}{}^k(O_1), N_{ps}{}^k(O_2), \ldots, N_{ps}{}^k(O_m)) \quad (7)$$

ここで，$N_{ps}^{k}(O_m)$は，投稿文 ps と類似する上位 k 個の文章のうち，意見クラス O_m に属する文章数を意味する．

Step-4 乖離度算出

対象意見 O_{target} における意見クラス間の対立度 L を用いて，O_{target} (=O)の乖離度 d_{ps} を算出する．意見クラス o_x に属する意見数を $N(o_x)$とすると，O_{target} に対する ps の乖離度 d_{ps} は，o_x の意見数に o_{ps} - o_x 間の対立度を重み付けし，以下のように合計することで算出される．

$$d_{ps} = \sum_{x=1}^{m} l_{(ps)x} N(o_x)/m \qquad (8)$$

Step-5 炎上の判定

d_{ps} が閾値θ_1を超えた場合，その投稿文 ps はソーシャルメディア上で炎上するリスクがあると判定し，d_{ps} が閾値θ_2以下の場合，その投稿文 ps は炎上するリスクが低いと判定する．

$$Faming\ (ps, d_{ps}) = \begin{cases} 1 & if\ \ d_{ps} \geq \theta_1 \\ 0 & if\ \ d_{ps} \leq \theta_2 \end{cases} \qquad (9)$$

ここで，1 と 0 はそれぞれ炎上リスクが高いことおよび低いことを意味する．

4. 世論の疑似データセット

指定されたトピックに関する文章の集合からなる世論の擬似データセットを定義する．詳細は文献[17]に示されている．このデータセットには，文章の分類情報，投稿日時，文章間の対立度が含まれる（表 1）．

対立度は文章の組ごとに n 段階の整数として定義される．例えば，対立の可能性が高いほど，対立の度合いは高くなる．対立度は意見の乖離度算出に用いられる．また，乖離度は 3 シグマルール[12]などを用いて算出することができる．意見文章はソーシャルメディアに時系列で投稿されるため，すべての意見文章に日時情報が付与される．表 2 と表 3 に擬似データセットのスキーマを示す．表 4 に示すように，分類情報の組の対立度に基づいて，各分類に属する意見に対立度を割り当てることができる．

擬似データセットの作成には，ソーシャルメディアに実際に投稿されたデータを取得するなどの方法がある．しかし，ソーシャルメディアが提供する API では投稿データの取得に制約が課せられる場合があり，炎上した投稿は削除されることが多い．一方，GPT-3 や 4 (Generative Pre-trained Transformer 3, 4) [13][18]に代表される大規模言語モデルの進歩は目覚ましく，人間に近い文章を生成することに成功している．このような理由から，本研究では，擬似世論データセットの生成に大規模言語モデルを適用する．

表 1　世論の疑似データセット

項目	説明
意見文	ソーシャルメディアに投稿された擬似意見文の数々
意見文のクラス	各文が属する意見クラス
日時	意見書が掲載された日時
対立度	意見文が互いに対立する度合い

表 1　意見スキーマ

{(ID, INT), (意見, TEXT), (分類情報, INT), (日時, TIME STAMP)}

表 2　対立度スキーマ

{(分類情報 A, INT), (分類情報 B, INT), (対立度, INT)}

表 4　意見データの組への対立度の付与

文章 A	文章 B	対立度
何の前触れもなく，こんなことが起こるなんて，本当に怖いです．	地震が起きたら，自分や家族を守るために，地震が起きたときの対処法を知っておこう．	1

5. 実験

提案手法により，様々な視点からなる意見文章間の対立度から，投稿された意見に対する乖離度に基づいて，投稿意見の炎上リスク判定が可能であることを示し，提案手法の実現可能性を確認する．

5.1　実験環境

世論疑似データセットの構築には GPT-3 モデル「text-davinci-003」を用い，「地震・被災地」というトピックに対して 1,048 文を生成した．この 1,048 文から，事前に学習した BERT モデル「bert-based-japanese-whole-word-masking」を用いて文章ベクトルを生成した．生成された文章の分類には，エルボー法[19]を用いてクラスタ数を決定した．さらに，炎上判定の評価のために，炎上文章 300 件，非炎上文章 300 件，合計 600 件の文章を同じ GPT-3 モデルを用いて作成した．炎上文章と非炎上文章のベクトル化には TFIDF[20]と Sentence-BERT[16]を用いた．4 章の Step-5 における閾値は，θ_1を 60%，θ_2を 40%に設定した．

5.2　実験方法

炎上文章については，乖離度が閾値θ_1（60%）を超えた場合，判定結果は正しいとみなす．一方，非炎上

文については，乖離度が閾値θ_2（40%）以下であれば，判定結果は正しいとみなす．評価には，以下の式を用いて，正答率，適合率，再現率を計算する．

$$\text{Correctness} = (N_{fl} + NA_{non\text{-}fl}) / N_{all}$$
$$\text{Precision} = NA_{fl} / NSJ_{fl}$$
$$\text{Recall} = NA_{fl} / NS_{fl}$$

ここで，

N_{all}: 全文章数
NA_{fl}: 炎上と判定した文章の正解数
$NA_{non\text{-}fl}$: 炎上文章の正解数
NSJ_{fl}: 正解と判定された文章の数
NS_{fl}: 炎上した文章の正解数

である．

5.3 実験結果

炎上文章 300 件，非炎上文章 300 件に TFIDF を用いて文章ベクトル化した場合の平均乖離度はそれぞれ 48.89%，40.71%であった．炎上および非炎上と判定された文章の正解数はそれぞれ 126 と 153 である．また，正解率，適合率，再現率はそれぞれ，46.5%，46.2%，42.0%であった．一方，Sentence-BERT を用いて文章ベクトル化した場合の平均乖離度はそれぞれ 52.7%，42.5%であった．炎上および非炎上と判定された文章の正解数はそれぞれ 121 と 197 である．また，正解率，適合率，再現率はそれぞれ，53.0%，54.0%，40.3%であった．

Sentence-BERT を適用した場合の方が TFIDF を適用した場合よりも正解率，適合率が高い結果となった．これは Sentence-BERT は文章の意味情報を捉えてベクトル化できることが理由であると考えられる．また，非炎上文章の判定については乖離度の閾値θ_2の 40%であるのに対して，平均乖離度は TFIDF を適用した場合は 40.71%であり，Sentence-BERT を適用した場合は 42.5%であるため，非炎上文章の判定が可能であると考えられる．一方，炎上文章の判定については，閾値θ_1の 60%に対し，平均乖離度は TFIDF を適応した場合は 48.86%であり，Sentence-BERT を適応した場合は 52.7%であるため改善の余地があると考えられる．

以上の結果により，提案手法を用いて，非炎上文章の判定については実現可能である見込みを得ることができたが，炎上文章の判定については閾値の設定などを再検討する必要がある．

6. まとめ

本研究では，ソーシャルメディア上の投稿意見を対象として，多様な視点からなる意見集合に対する乖離度を算出する手法を提案した．大規模言語モデルにより作成した擬似的な世論データセットを用いた実験により，提案手法を用いて，投稿意見の非炎上文章の判定については実現可能である見込みを得ることができたが，炎上文章の判定については閾値の設定などを再検討する必要があることが確認できた．

今後の予定として，意見文章集合の規模を拡大し，擬似データセットとして公開することを検討している．また，本疑似データセットを用いて，炎上可能性を判定するための Transformer を適応したニューラルモデルを構築し，炎上判定に対するニューラルモデルの有効性を検証することを予定している．また，大規模言語モデルに対して本疑似データセットを用いてファインチューニングし，炎上判定に適用することも検討している．

参 考 文 献

[1] X, https://twitter.com/, (閲覧 2024/01/02).
[2] Facebook, https://www.facebook.com/, (閲覧 2024/01/02).
[3] mixi, https://mixi.jp/, (閲覧 2024/01/02).
[4] 大曽根匡, 福田浩至, “クチコミによるネット炎上の定量化の試みとその検証”, 情報システム学会第 12 回全国大会, 2016.
[5] 堺雄之介, 竹内幹太, 伊東栄典, “コメントを利用した炎上動画検出に関する検討”, 情報処理学会研究報告, pp. 1-5, 2021.
[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space”, arXiv preprint arXiv:1301.3781v3, 2013.
[7] 川上幹, 彌冨仁, “Twitter への投稿テキストによる炎上警告システムの構築”, 第 32 回ファジィシステムシンポジウム, pp. 705-708, 2016.
[8] 吉田基信, 松本和幸, 吉田稔, “攻撃的文章判定を用いた不適切文章の訂正”, 第 22 回情報科学技術フォーラム, pp. 265-266, 2023.
[9] 高橋直樹, 檜垣泰彦, “Twitter における感情分析を用いた炎上の検出と分析”, 電子情報通信学会, pp. 135-140, 2017.
[10] 瀬川友香, 浅谷公威, 坂田一郎, “ユーザーに着目した SNS 上の攻撃とそのメカニズムに関する分析”, 第 35 回人工知能学会全国大会, 2021.
[11] 岩崎祐貴, 折原良平, 清雄一, 清雄一, 田原康之, 大須賀昭彦, “CGM における炎上の分析とその応用”, 人工知能学会論文誌, 30 巻 1 号, pp. 152-160, 2015.
[12] L. J. Kazmier, “Schaum's Outline of Business Statistics”, Schaum Pub Co, 2003.
[13] T. B. Brown. et al, “Language Models are Few-Shot Learners”, Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 1877-1901, 2020.
[14] J. MacQueen, “Some methods for classification and

analysis of multivariate observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[15] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arxiv:1810.04805, 2018.

[16] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", arXiv:1908.10084, 2019.

[17] 吉田裕輔, 鷹野孝典, "投稿における炎上推定モデル生成のための疑似学習データセット構築の検討", 情報科学技術フォーラム講演論文集, 22 巻 2 号, pp. 239-240, 2023.

[18] J. Achiam, et al., "GPT-4 Technical Report", arXiv:2303.08774 [cs.CL], 2023

[19] R. L. Thorndike, "Who belongs in the family?". Psychometrika, Vol. 18, No. 4, pp. 267-276, 1953

[20] K. S. Jones, "IDF term weighting and IR research lessons", Journal of Documentation, Vol. 60, No. 5, pp. 521-523, 2004.

ニュース記事の拡散速度に着目したソーシャルメディア分析

添田　英紀†　　横山　昌平††,†††,††††

† 東京都立大学 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所　〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †soetaeiki@gmail.com, ††shohei@tmu.ac.jp

あらまし　新型コロナウィルス流行開始時に, 以前から SNS での問題であった誤情報の伝播が見られた. 真偽に関わらず人々はしばしば興味を引かれる記事に即座に反応し, 拡散する傾向がある. 本研究では, 新型コロナウィルス流行以降の SNS のデータを使い, ニュースがどれだけ早くどのように拡散されるかの傾向分析をし, 記事が発行される時間帯やニュース記事のタイトルに含まれる単語をベースに考察する. 問題意識は, 拡散速度の差, 誤情報の特徴, 情報の偏りやバイアスの防止に焦点を当てている. これらの研究を通じて, 新型感染症発生時などの即座に世界中を巻き込んだ混乱が起きた場合の適切な意思決定の支援を目指す. 評価手法は, 時間が経過するにつれた拡散速度の変化をグラフ化する.
キーワード　SNS，Twitter(現 X)，新型コロナウィルス, ニュース, 行動分析

1 はじめに

新型コロナウィルスの流行は, 世界中に深刻な影響を及ぼし, 人々の生活や社会に多くの変化をもたらした. このような状況下で, 人々は情報を求め, SNS を通じてニュースや意見を共有することが多くなった. しかし, SNS 上には, 正確でない, あるいは誤解を招くような誤情報も多く存在し, 人々の認識や判断に影響を与える可能性がある. 特に, 人々は興味を引かれる記事に即座に反応し, 拡散する傾向があるため, 誤情報は速やかに広まるリスクが高いとされている. しかし, 新型コロナウィルス流行開始時に誤情報を原因としたトイレットペーパー買い占め騒動が起こったのだが, このとき誤情報を追い抜くようにデマの訂正情報が拡散されていた [1]. このように情報の真偽に関わらず, 人々は SNS 上の情報に踊らされる事例がある. そのため, SNS 上で急速に拡散される情報を早めに検出することは適切な意思決定の支援の第一歩として重要であると言える.

本研究では, 新型コロナウィルス流行以降の SNS のデータを使い, どんなニュースがどれだけ早くどのように拡散されるかの傾向分析を行う. 具体的には, 記事のタイトルに含まれる単語や記事が公開される時間帯である. 本研究の問題意識は, 拡散速度の差, 誤情報の特徴に焦点を当てている. これらの研究を通じて, 新型感染症発生時などの即座に世界中を巻き込んだ混乱が起きた場合の適切な意思決定の支援を目指す.

本研究の評価手法には, 時間が経過するにつれた拡散速度の変化をグラフ化していく. さらに, 時間が経過して削除されてしまった記事のタイトルやキーワードをツイートデータから取得する方法も提案する. これらの手法を用いて, SNS 上のニュースの拡散パターンを分析し, どのような記事がどのように拡散されるかを観察していく.

本論文の構成は以下の通りである. 第 2 章では, 関連研究について紹介する. 第 3 章では, 本研究で用いたデータセットや分析手法について説明する. 第 4 章では, 分析結果を示し, 考察する. 第 5 章では, 本研究のまとめと今後の課題について述べる.

2 関連研究

Vosoughi らは誤ったニュースは正しいニュースよりも速く拡散されることを示した [2]. Esma らは主なフェイクニュース検出に対するアプローチとしてニュースの中身に基づくコンテンツベース, 社会的背景に基づき検出に役立つ追加情報を得るコンテキストベース, そしてそれらのハイブリットのものがあると分類した. 中でもコンテキストベースはユーザーベースとネットワークベースの 2 つのサブカテゴリに分けられるとした [3].

井上らは Twiiter におけるニュース記事の時系列分析として, ニュースとツイートデータに対して変化点検出手法である ChangeFinder を用いてキーフレーズを抽出し, 現実に起こった出来事が変化点から把握できることを示した [4]. 江口らは Twitter 上のニュース記事に対して, ニュース本文を使わずコンテキストデータのみを用いてニュースメディアのスタンス分類を行う手法を提案した [5]. 芝田らは福島第一原発の関連単語のツイート数上昇のピークを検出し, 災害時の情報の伝播を分類した [6]. 時系列分析の主な手法として特定の単語の出現頻度が急激に上がるところを検出するのは一般的な手法である. 本論文では単語ではなく, 記事の URL を用いて分析を行う. 理由については次章のデータセットで説明する.

SNS でのおすすめタイムラインのアルゴリズムの影響により, ユーザーは自分が見たい情報のみに触れるフィルターバブルの影響でコミュニティの分断や思想の分裂が進み，エコーチェンバーが形成される可能性がある．成田らは Twitter 上の新型コロナに関するコミュニティの経時的分析を行い, その結果として時間が経つにつれコミュニティが分離すると共に，コミュニ

ティ間の分断が大きくなるを示唆した [7]. また, Toriumi らは日本のツイートデータから反ワクチンのユーザーは政治に関心が強く, 陰謀論とも結びつきが強いことを示した [8].

榊らは特定のコミュニティに属するユーザが脊髄反射的に拡散したくなる情報をソーシャルポルノと名付けた [9]. この研究では「ある情報を受け取ってから, その投稿を拡散するまでに要した時間」に着目して, その時間が短いほどその情報はポルノ的ではないかとしている. また, Twitter 社の Kajihara らは 2023 年 4 月に「#拡散の科学 Anatomy Of Retweet」というスライドを公開した. その中でバズっているツイートを 1300 リツイート以上とし, 大抵のツイートは 72 時間で拡散が落ち着くことを示した [10].

本論文では, 扇動的な記事の検出への第一歩として, SNS で拡散される記事にどのような傾向があるかを分析する. 分析に用いる指標は主に, 記事の発行された時間帯と記事のタイトルに含まれる単語と拡散速度である. また 1300 以上拡散されたものをバズった記事の基準として扱う. また, この論文においては発行された直後に一気に拡散され, その後すぐに拡散されなくなるものを「ポルノ的」な記事とする.

3 データセット

本研究では株式会社 NTT データ「なずきのおと」を用いて, Twitter の日本語ツイート全量データから表 1 のメディアの URL を含むツイートのデータを分析に用いる. 対象とするメディアの選定は鳥海らの研究を参考とした [11]. 期間は 2022 年 1 月 1 日から 2022 年 12 月 31 日までで, 合計件数は 46,798,520 件である. また Twitter におけるニュースツイートでよく拡散される Yahoo ニュースと livedoor ニュースについてはサイトの性質上, 他社メディアが書いた記事を載せているため, 同じメディアの同じ記事が重複するため今回の分析では扱わないことにした.

ここでなぜニュース記事に着目して分析するのかについて述べる. SNS で急激に拡散されている話題を検出する手法として主にバースト検出や異常検出が挙げられる.

たとえば新型コロナウィルスのニュースを検出するために「マスク」という単語の頻度を追う場合,「イーロン・マスク」のような, 違う単語に目的の単語が含まれているとき, 単純に「マスク」の出現頻度から異常を検出しようとすると, それらも含めて検出してしまう可能性がある. 他には「コロナ」に対して「コロナビール」なども可能性がある.

それに対し, ニュース記事の URL はその記事の HTML ファイルの名前がハッシュ値によって決められることが多く, 記事が発行される瞬間に今までに無い新しい URL として世に出る. そしてそのメディアのアカウントによって投稿されて, その他の一般のユーザーに拡散されていく. つまりニュース記事は多くの場合に発行されたタイミングから一気に拡散されるので, 異常検出を行う必要がない. 最初から異常だからである. 一般的なあるキーワードでデータを集めて分析を行う場合, その単語が急激に出現数が伸びたときに異常であると検出を行う必要がある. 記事それぞれのユニークな URL に着目すればその必要がなくなると考える.

表 1　対象としたメディア

メディア	URL	件数
朝日新聞	asahi.com	7,440,908
日経新聞	nikkei.com	7,529,059
毎日新聞	mainichi.com	5,901,809
読売新聞	yomiuri.com	3,360,226
日本放送協会	nhk.com	16,953,837
ハフポスト	huffingtonpost.jp	691,135
ハム速	hamusoku.com	334,198
俺的ゲーム速報	jin115.com	574,965
ツイ速	tsuisoku.com	2,133,308
保守速報	hosyusokuhou.jp	1,879,075

4 分　　析

4.1 集めたデータの加工

まず図 1 に可視化のために行ったデータの加工を示す. データを扱っていく上で一部に扱いが難しい URL があることがわかったため, それぞれの記事に固有のハッシュ値などを持つ URL に限定する加工を行った. 次に記事が発行された時間帯を 24 時間を 3 時間ずつに分けて分類した. それぞれの時間帯で特定の単語をタイトルに含むものに分けた.

記事のタイトルについて, 大部分は記事の URL からスクレイピングをして取得したが, 記事の公開から時間が経っていたり記事の内容に何らかの事情があり削除されたものがある. これらの記事に対しては, その URL を含むツイートテキストを最大で 100 件取得し, 助詞などを省いて分かち書きをした後に出現頻度の高い順の単語 10 件を仮にタイトルとした.

分けた後の記事の URL それぞれについて, 1 件目のデータから 5 分あたりに何回リツイートもしくはツイートされている回数を 24 時間後まで格納した. これを行うことにより拡散速度の低下の原因が深夜帯のアクティブユーザーの減少が原因かそれ以外が原因かがわかりやすくなる. 本論文においてリツイートされた数と言及ツイートの数の和を拡散数とする.

4.2 ニュース発行時間帯別の分析

図 2 は 1 件を除いた 9 時から 12 時に発行された記事との推移である. 縦軸は 5 分ごとの拡散数で, 横軸は 24 時間後までになっている. 発行されたタイミングはその記事が含まれる最初のツイートのタイムスタンプとした. このグラフには 17,018 件の記事の推移が含まれている. 除いた 1 件については, 安倍元首相が銃撃された記事であり, これは最初の 5 分で 5,378 回拡散され, 次の 5 分で 11,853 回拡散されていた. これを同じく描画してしまうと縦軸がそれ引っ張られ伝わりにくいので, 便宜上除いた. さて, このグラフについて, 1 件を除いて深夜帯である 15 時間後から 18 時間後ごろに拡散数が下がり, その後にまた増えている. すべての時間帯の記事にこのような傾向が見られ, 多くの記事は深夜帯を経た後に少し拡散数は戻るが発行直後ほどに拡散されるものは多くない.

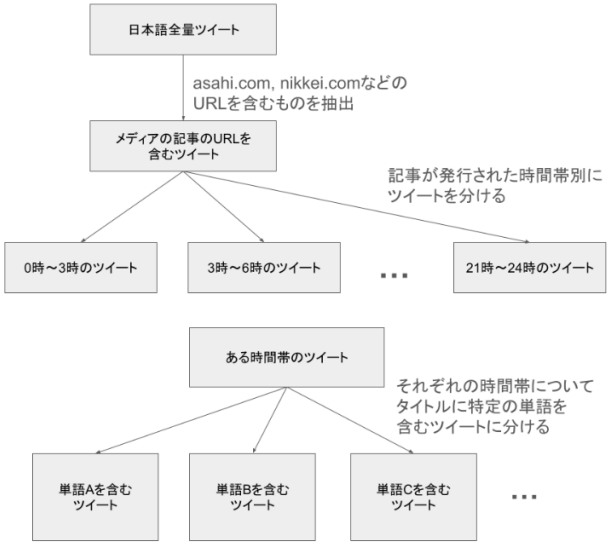


図 1　データの加工の概要図

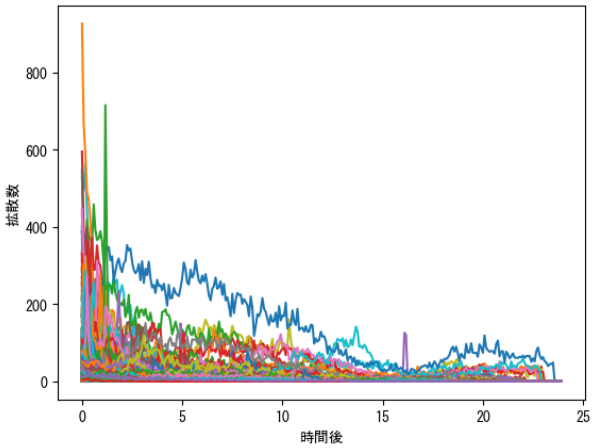


図 2　9 時から 12 時に発行された記事

しかし例外もある. 図 3 左は 18 時から 21 時のコロナ関連単語（コロナ, マスク, ワクチン, 武漢）をタイトルに含んだ記事の推移で 4081 件ある. 10 時間後から著しく拡散数が増加しているものが 1 件ある. これは新型コロナウィルスの流行によって生活が困窮した外国人留学生を含む日本の学生に政府が 10 万円を支給するというニュースである. この記事は右派的と見られるコミュニティの中でなぜ外国人留学生にも支援をするのかという文脈で, 様々なアカウントが言及ツイートをしていた. このように特定の党派性コミュニティの中で炎上する記事は記事の発行から時間が経過しても拡散数が伸び続ける傾向にある. 筆者の仮説として, 特定のコミュニティを代表するインフルエンサーが言及ツイートをすることで炎上するものと思われたが, フォロワー数に限らず記事の言及ツイートが大量に発生することで炎上は起こっていた.

4.3　キーワードごとの分析

図 4 著名人が亡くなったときに使われる単語（死去, 訃報）がタイトルに含まれる記事の拡散数の推移である. 先述の図 3 のコロナ関連単語の推移と比較したい. 深夜帯の拡散の落ち着き

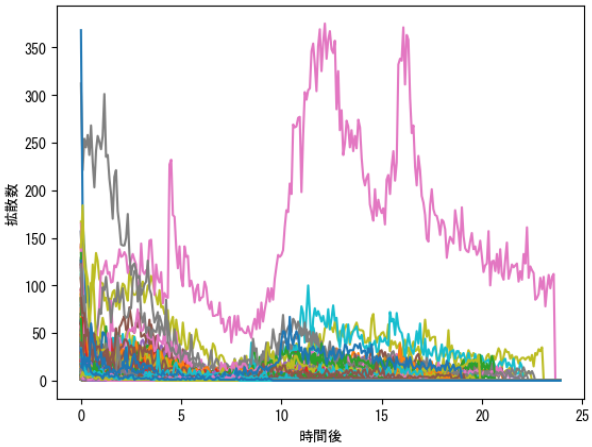


図 3　18 時～21 時のコロナ関連単語の記事の推移

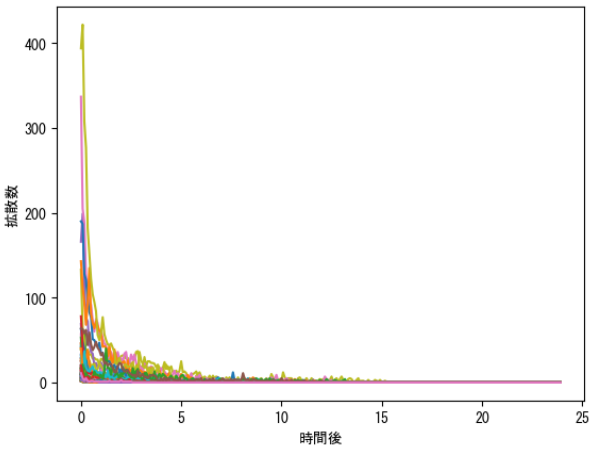


図 4　18 時～21 時の訃報関連単語の記事の推移

はあるが, 前者のグラフは拡散数が持続している記事もある. それに対して, 後者のグラフは記事の発行後に一気に拡散されて, すぐに落ち着いている. 訃報のようなニュースは拡散される期間が短い.

次に新型コロナウィルス関連語の中からいくつかを選び, その単語の含む記事の拡散速度の推移を観察していく. 図 5 は 18 時～21 時のタイトルに「ワクチン」を含む記事の拡散速度の推移図である. 図 6 は同じ時間帯の「感染者」を含むものである. 「ワクチン」を含む記事は記事の発酵からある程度時間が経っても拡散数が下がっていなかったり, むしろ拡散数が上がっているものがある. 対して「感染者数」の記事の多くは拡散から 10 時間程度経つと拡散が落ち着いているものが多い. 「ワクチン」を含む記事はワクチンの効果の是非の論争に巻き込まれることが多く, そのため時間経過後も拡散されていた. 「マスク」をタイトルに含む記事についても同様の可視化を行ったが, 同じような推移の傾向が見られ, マスクは感染予防に有効であるかどうがの議論の元になっていた. 対して「感染者数」の記事, とりわけその日の感染者数を伝えるのみの客観的な事実の記事は拡散直後に速度が落ち着くものが多かった. 例外として「感染者数」という単語が含まれていてもマスクやワクチンなどの対策が感染者数を抑える可能性の示すような記事は論争に巻き

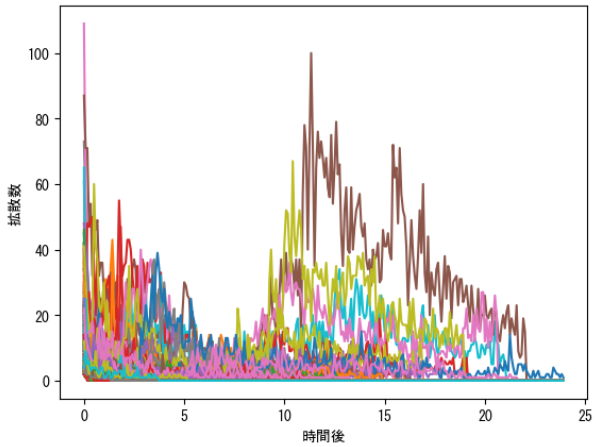


図 5　18 時〜21 時のタイトルに「ワクチン」を含む記事の推移

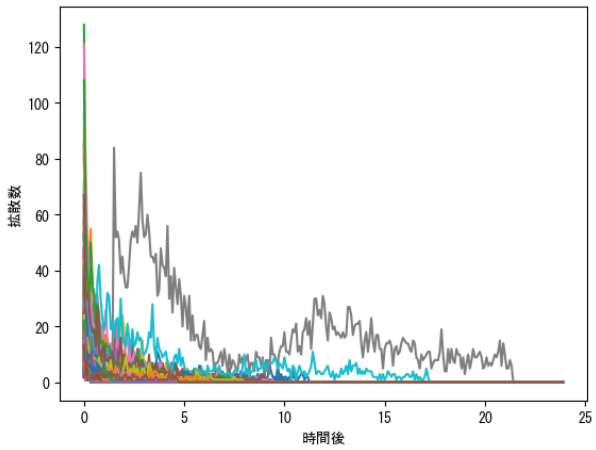


図 6　18 時〜21 時のタイトルに「感染者数」を含む記事の推移

込まれることが多く, それらの記事は時間が経過したあとでも拡散速度が上がっていくものがあった.

次に, ワードクラウドを用いて「バズる」記事の拡散の速いか遅いかと, その記事に含まれる単語に違いがあるかを調べる. 図 7 は 24 時間以内に 1300 回拡散された記事のタイトルのワードクラウドである. つまり速く「バズった」記事である. 対して, 図 8 は 1300 回拡散されるまでに 1 日以上 3 日時間以内の記事, つまり「バズった」状態になるまでに 1 日以上の時間がかかった記事のタイトルでワードクラウドを作ったものである. 1300 という基準は Twitter 社が公開した#拡散の科学を参考にしている [10]. 両者とも単語にあまり違いは見受けられない. 1300 回拡散されるものとしては単語に大きな違いは見受けられなかった. しかしどちらも政治に近い単語が頻出していることが見て取れる. これらは他の大手新聞社でも同じ傾向であった. まとめサイトでは保守速報やツイ速については近い傾向が見られたが, 俺的ゲーム速報については想定読者の影響もあってかゲーム関連の単語が多く見られた.

では政治系の単語ではどのような拡散の推移をするのか. 図 9 は政治関連語として政府, 大臣, 辞任, 与党, 野党, 自民, 立民, 維新, れいわ, 参政党, 党首, 公明, 共産をタイトルに含む記事の推移を可視化したものである. なお安倍元首相については銃撃



図 7　朝日新聞の記事で 1300 回拡散されるまでに 1 日以内の記事のタイトルのワードクラウド



図 8　朝日新聞の記事で 1300 回拡散されるまでに 1 日以上 3 日以内の記事のタイトルのワードクラウド

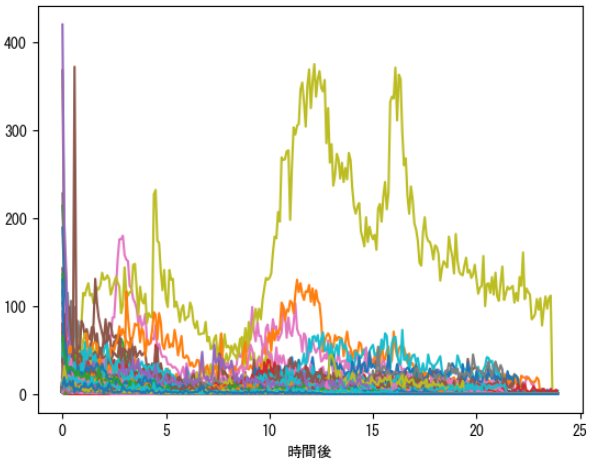


図 9　18 時〜21 時のタイトルに政治関連単語を含む記事の推移

事件の記事が脊髄反射的に拡散していたが, それは訃報系の記事に入るとしてこの図からは除外している. 2 件ほど記事の発行直後に一気に拡散されて落ち着くものも観られるが, 10 時間が経過しても 50 件以上の拡散速度を出しているものも多く見られる.

次に, 即座に社会に混乱をもたらすものの例として災害に関する単語を含んだ記事の推移を観察する. 選んだ単語は地震, 停電, 津波, 噴火, 落雷である. 図 10 は 18 時〜21 時のタイトルに災害関連単語を含む記事の推移である. 一部の記事が一時的に 1000 件近い拡散速度になっている. このために他の図と比べて

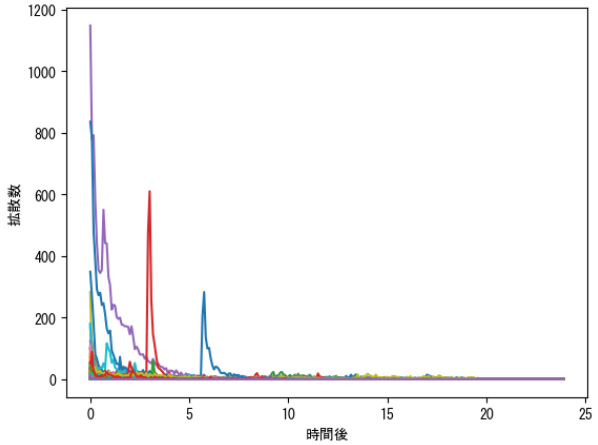


図 10　18 時〜21 時のタイトルに災害関連単語を含む記事の推移

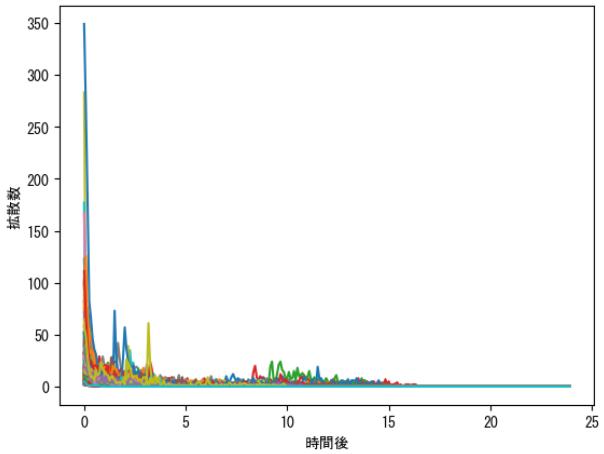


図 11　図 10 の一部を除いた記事の推移

縦軸が大きくなってしまっているため, これらの記事を除いたものが図 11 である. 除いた記事は桜島噴火, トンガ噴火, マナム火山噴火の記事であった. 縦軸を大きく引っ張るような記事を除いても, 災害関連単語をタイトルに含む記事は直後に一気に拡散され, その後すぐに落ち着く. つまり脊髄反射的に拡散される傾向にあることがわかる.

4.4　メディア別の分析

次にメディアごとの記事の推移の違いについて調べる. 図 12 は大手新聞社である朝日新聞と毎日新聞, 図 13 はまとめサイトであるツイ速と保守速報の拡散数が 1300 以上に達する記事の達するまでの時間のヒストグラムである. 大手新聞社は幅広く分布しているが, 保守速報, ツイ速のようなまとめサイトは拡散までの時間が短くまとまっている. 大手新聞社は情報源としての信頼性が高く, 情報としての鮮度が落ちにくいことが伺える. 対して, まとめサイトでは想定するユーザーが喜ぶような記事を掲載するため, 一気に拡散されるがすぐに消費されて拡散は落ち着く. 両者の比較ではまとめサイトの記事の方がポルノ的な消費のされ方をしていると言える.

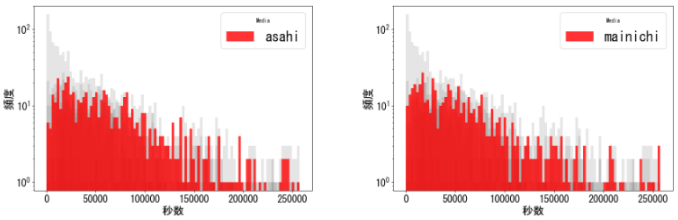


図 12　朝日新聞と毎日新聞の記事が 1300 回拡散されるまでの秒数のヒストグラム

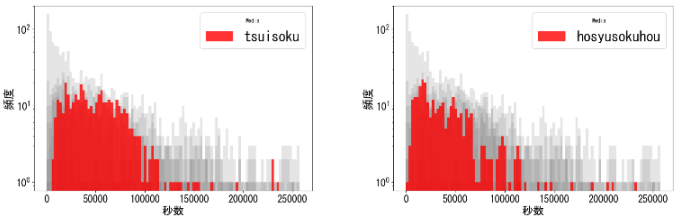


図 13　ツイ速と保守速報の記事が 1300 回拡散されるまでの秒数のヒストグラム

5　おわりに

本論文では扇動的な記事の検出の第一歩として Twitter（現 X）におけるニュース記事の分析の手法を提案した. 記事の発行された時間帯とメディア, 記事のタイトルに含まれる単語を主な指標とした. 記事の発行された時間帯にまとめることで拡散の落ち着き方が深夜帯のアクティブユーザーの減少が原因か, その他が原因かを可視化することができる. しかしこれは日本という国土全体が同じタイムゾーンであるが故に可能な手法である. 英語圏のニュース記事では深夜帯の拡散速度の低下のような減少が見られるかを今後の研究で調査を行う.

また, タイトルの単語を元にした分析では拡散の推移の仕方には「発行後すぐに拡散されてすぐに落ち着くもの」「発行後もしばらく拡散され続けるもの」に分けられることがわかった. 脊髄反射的に即拡散されるものは感染者数や訃報, 災害など, その記事のセンセーショナルさはあれども陰謀論に絡むものではなく, 事実を淡々と伝えるものだった. 逆にマスクやワクチン, 政治関連語などの特定の信条や政治思想に基づくコミュニティが拡散したくなるような記事は記事が議論の元となり発行から時間が経っても拡散され続けるものが多いことがわかった. 1300 回以上拡散された記事のタイトルについては拡散の速さに関わらず政治に近い単語が多く出現していた.

またメディアについて比較をすると, 1300 回以上拡散されたバスった記事については, 大手新聞社などは信頼性が高い故か, 記事が発行されてもしばらく拡散は続くが, まとめサイト系のメディアは記事がすぐに拡散されてすぐに落ち着く傾向にある.

これらの分析はまだ定性的なものにとどまっているため, 今後定量的な評価の仕方を定めて分析を続ける必要がある. またメディアごとの記事の発行数の違いを考慮した可視化も行いたい. さらに時間経過後も拡散され続ける記事については, どのよ

うなユーザー層が拡散しているのを明らかにしていきたい.

謝　　辞

本研究は, 内閣府 COVID-19 AI・シミュレーションプロジェクトの一環として実施され, JST，CREST，JPMJCR19A4 の支援を受けたものである.

文　　献

[1] Ryusuke Iizuka, Fujio Toriumi, Mao Nishiguchi, Masanori Takano, Mitsuo Yoshida. “Impact of correcting misinformation on social disruption” PLoS ONE 17(4): e0265734. 2022

[2] Soroush Vosoughi,Deb Roy,Sinan Aral “The spread of true and false-news online” Science, 2018

[3] Esma Aïmeur, Sabrine Amri, Gilles Brassard “Fake news, disinformation and misinformation in social media: a review” PMC Research, 2023

[4] 井上 大成, 中北 雄大, 風間 一洋, 吉田 光男, 土方 嘉徳 “メディアの報道とソーシャルメディアの反応の時系列分析” 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2023

[5] 江口 航野, 横山 昌平 “外部知識のネットワーク関係を活用したニュースメディアのスタンス分類” 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2022

[6] 芝田 遊吾, 木村 浩, 鳥海 不二夫, 榊 剛史, 風間 一洋, 福田 健介 “ソーシャルメディアにおける福島第一原子力発電所に関連する情報の伝播分析” 人工知能学会研究会, 2012

[7] 成田 省吾, 横山 昌平 “新型コロナ禍におけるエコーチェンバー形成過程の分析析” 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2022

[8] Fujio Toriumi, Takeshi Sakaki, Tetsuro Kobayashi, Mitsuo Yoshida “Anti-vaccine rabbit hole leads to political representation: the case of Twitter in Japan” Journal of Computational Social Science, 2023

[9] 榊 剛史, 鳥海 不二夫, “ソーシャルポルノ仮説の提案とその観測に向けて” 2018 年度人工知能学会全国大会（第 32 回）, 2018

[10] Daisuke Kajihara, Cheng Qian, Yui Takeshima “#拡散の科学 Anatomy Of Retweet” https://marketing.twitter.com/content/dam/marketing-twitter/apac/ja/insights/kakusan/kakusan.pdf, Twitter Japan, April 10, 2023, accessed 2024/01/09

[11] 榊 剛史, 鳥海 不二夫, “ソーシャルポルノ仮説に基づくメディア企業 SNS アカウントの煽動性に関する分析” 一般社団法人 人工知能学会 第 34 回全国大会, 2020

Trustable Internetにおけるインターネット上のデータの信憑性判定および根拠説明文生成手法の提案

佐久間義友† 大木 憲二† 佐々木佑樹† 坂本 拓也† 吉武 敏幸†
二村 和明†

† 富士通株式会社 データ&セキュリティ研究所 〒211–8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: †{y-sakuma,oki.kenji,sasaki.yuki-01,takuya,yoshitake.toshi,kazuaki.nimura}@fujitsu.com

あらまし 近年，インターネット上における偽情報の流通が社会・経済活動に混乱をもたらすことが問題視されている．この問題を解決するために，我々は情報に紐づいた信憑性の判断材料に基づき，情報の信憑性判断を支援するアーキテクチャである Trustable Internet を提案している．受信者による信憑性の判断支援において，信憑性の高い判断材料に基づく判断対象の信憑性とその根拠を受信者に提示することが重要となる．従来技術として，判断対象に関連する情報を判断材料として用いて判断対象の信憑性と根拠を提示する手法が存在するが，判断材料の信憑性が考慮されていないことが課題として挙げられる．前述の課題を踏まえ，本論文では信憑性の高い判断材料のみに基づき，判断対象の信憑性の判定と根拠説明文を生成する手法を提案する．

キーワード ソーシャルメディア，LLM，自然言語処理応用

1 はじめに

現代社会において，人々はインターネットを通して様々なサービスを日常的に利用している．特に，ソーシャル・ネットワーキング・サービス（SNS）の利用率が増加し，個人や企業・行政機関が SNS を通して情報を発信・受信する機会も増加している [1]．前述の社会動向に伴い，インターネットを通して送受信される情報の確からしさ（信憑性）が以前よりも増して重要になってきている．

一方で，インターネットを通して発信された情報の中には事実と異なる情報（偽情報）が含まれる場合もあり，この偽情報が社会や経済に混乱の原因になることが問題視されてきている [2,3]．現在のインターネットでは，情報が事実である根拠を発信者が示しているとは限らないため，受信者がその情報の信憑性を判断することが困難である場合も多い．また，昨今は文章や動画像を生成する AI 技術の高度化により，受信者が情報の信憑性を判断することがより困難になってきている [4]．

前述の情報の信憑性に関する社会課題を踏まえ，我々は受信者が信憑性を知りたい情報（判断対象）の信憑性を確認可能にするための新たなアーキテクチャである「Trustable Internet」を提案した [5]．Trustable Internet は，判断対象に付加された情報（エンドースメント）に基づいて受信者による信憑性の判断を支援するためのアーキテクチャである．

Trustable Internet による信憑性の判断支援を実現するためには，エンドースメントに基づく信憑性判定や判定の根拠説明を受信者に提示することが重要となる．また，情報の信憑性やその根拠説明に関連する技術として，判断対象に関連する情報から信憑性の判定と根拠説明を出力する機械学習モデルを用いた手法が提案されている [6,7]．しかしながら，前述の技術では判断対象に関連する情報自体の信憑性が考慮されていないことが技術課題として挙げられる．具体的には，判断対象に関連する情報が信頼できる情報源から発信されていることがまず確認されていない．また，判断対象に関連する情報の信憑性が他の情報に基づいて保障されていることも確認されていない．

前述の技術課題を踏まえ，本論文では図 1 に示すように信憑性の高いエンドースメントのみに基づき判断対象の信憑性判定と根拠生成文を生成する手法を提案する．提案手法では，まずエンドースメントの発行者の信頼性とエンドースメントに付加されたサブエンドースメントを用いて信憑性の高いエンドースメントを抽出する．そして，抽出した信憑性の高いエンドースメントのスタンス（判断対象を肯定/否定を表すラベル）に基づいて判断対象の信憑性判定と根拠説明文の生成する．

本論文の流れはまず，第 2 章では受信者に情報の信憑性の判定と根拠説明を提示する既存手法とその課題について述べる．次に，第 3 章では提案手法のユースケースや Trustable Internet 等の提案手法が動作する想定環境について述べる．そして，第 4 章で本論文おける提案手法であるインターネット上のデータの信憑性判定・根拠説明文生成手法を述べた後に，第 5 章では提案方式が動作するシステムの実装と性能評価について述べる．最後に，第 6 章で本論文のまとめと今後の課題について述べる．

2 関連手法

本章では，受信者に情報の信憑性の判定と根拠説明を提示する既存手法とその課題について述べる．

2.1 信憑性判定モデルを用いた手法

本節では，判断対象に対する信憑性を判定するモデルと，モデルの内部パラメータである Attention に基づいた根拠説明を

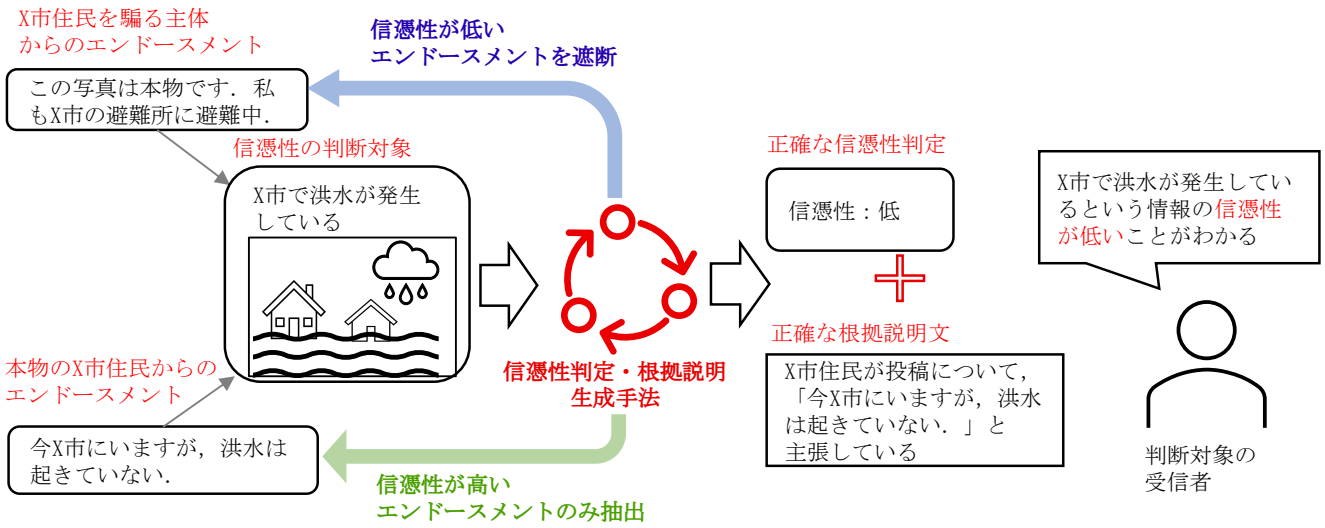


図 1　提案手法の概要

受信者に提示する手法について述べる．

論文 [6] では，入力の出力に対する寄与度を表す Attention を算出する機構を内部に持つ信憑性判定モデルに対して判断対象とそれに関連する情報の組を入力し，出力として信憑性判定を得る手法を提案している．また，論文 [6] の手法では，信憑性判定において入力された判断対象に関連する各情報について，前述の Attention スコアが高い単語を強調表現したものを信憑性判定の根拠説明として受信者に提示している．

2.2　信憑性判定モデルと根拠説明生成モデルを用いた手法

本節では，判断対象に対する信憑性を判定するモデルと信憑性判定に関する根拠説明文を生成するモデルにより信憑性判定と根拠説明を受信者に提示する手法について述べる．

論文 [7] では，BERT を改良して構築した信憑性判定モデルと根拠説明生成モデルに判断対象とそれに関する情報の組を入力することで，判断対象の信憑性判定と根拠説明文を生成する手法を提案している．論文 [7] の手法と論文 [6] の手法との違いは，信憑性判定モデルの内部パラメータに基づいて根拠を説明するのではなく，信憑性判定とは別に根拠説明文の生成に特化したモデルを用いる点である．

2.3　関連手法の課題

情報の信憑性判定における関連手法 [6,7] の課題は，信憑性の判定や根拠説明に用いた判断対象と関連する情報自体の信憑性を考慮していない点である．

人々が情報の信憑性を判断するにあたり，情報発信者の信頼性を評価することは重要である [8]．しかしながら，関連手法 [6,7] では信憑性判定や根拠説明に用いた判断対象と関連する情報について発信者の信頼性を考慮されていない．特に，個々の受信者に適した情報の信憑性判断支援を実現するためには，受信者毎に信頼する情報発信者は様々である点も考慮しなければならない．また，判断材料自体の発信者の信頼性が不明である場合でも，信頼性が高い他の発信者が判断材料を肯定する情報が存在する場合には判断材料自体の信憑性も高くなると考えられる点も考慮しないといけない．

前述の課題を踏まえ本論文では，発信者の信頼性および判断材料を肯定（もしくは否定）する他の情報を用いて抽出した信憑性の高い判断材料を用いて，判断対象の信憑性判定と根拠説明文を生成する手法を提案する．

3　提案手法の想定環境

本章では提案手法のユースケースや Trustable Internet [5] について述べる．

3.1　ユースケースと本研究による課題解決像

本研究では，ユースケースとして SNS 等を介して事実とは異なる情報が発信されたケースを想定している．前述のケースの具体例として図 2 に示すように，画像生成 AI で作成した洪水の画像と「X 市で洪水が発生している.」という文章が掲載された事実と異なる SNS 投稿が発信されるというケースが考えられる．この具体例において，X 市の遠方に住む投稿受信者は，投稿の信憑性やその根拠説明が投稿に併記されない場合には，X 市で洪水が発生していると誤認識してしまう可能性がある．

前述の具体例が示すように，判断対象（前述の例では SNS 投稿）のみを見た状態では受信者には投稿の信憑性を判断することが困難であるという課題がある．前述の課題を踏まえて本研究では図 1 に示すように，信憑性の判定結果とその根拠説明文を提示することで，受信者による信憑性の判断を補助することを目指す．

3.2　Trustable Internet

3.2.1　Trustable Internet の概要

Trustable Internet [5] は，判断対象に対してエンドースメントを付与可能にすることで，エンドースメントに基づいて判断対象の信憑性判断を支援するアーキテクチャである．

3.2.2　エンドースメント

エンドースメントとは，判断対象の信憑性の判断材料として

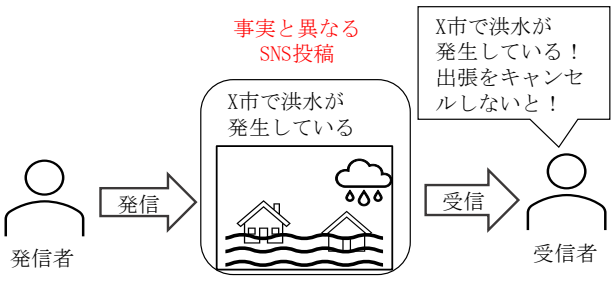


図 2　本研究のユースケースの具体例

判断対象に関連する情報（プロパティ）が記載されたデータである．具体的には，図 2 にある SNS 投稿が判断対象になる場合，X 市の災害情報に関するコメントや，洪水に関する目撃情報などがプロパティになりうる．エンドースメントは下記の 4 つの項目で構成される．

- エンドースメントの識別子
- エンドーサ（エンドースメント付与者）の識別子
- エンドーシ（エンドースメント被付与者）の識別子
- プロパティ

また，エンドースメントは判断対象だけでなくエンドースメントにも付与可能であり，本論文ではエンドースメントに対して付与されたエンドースメントをサブエンドースメントと呼ぶ．サブエンドースメント具体例として，エンドーサの保有資格や所在地などの属性情報やエンドースメントの信憑性に寄与する文献の引用やコメントをプロパティに記述したものが挙げられる．

3.2.3　エンドースメントグラフ

エンドースメントグラフとは，判断対象とエンドースメントおよびサブエンドースメントをノードとし，各ノードをエンドースメントの付与・被付与の関係で接続した有向グラフである．具体例として，図 2 にある SNS 投稿を判断対象とした場合のエンドースメントグラフを図 3 に示す．

図 3 の例では，判断対象に対して個人である allice と bob が，判断対象の写真に関するエンドースメントを付与している．また，allice によるエンドースメントには位置情報提供サービス（識別子はサービスのドメイン名である locinfo.co.jp）から allice の位置情報を記載したサブエンドースメントが付与されている．Trustable Internet では，判断対象を起点としてエンドースメントグラフをたどることによって取得したエンドースメントやサブエンドースメントを用いて信憑性の判断支援を行う．

次章では，Trustable Internet における情報の信憑性の判断支援を実現するための，判断対象の信憑性判定と根拠説明文を生成する手法について述べる．

4　提案手法：インターネット上のデータの信憑性判定・根拠説明文生成手法

本章では提案手法について述べる．

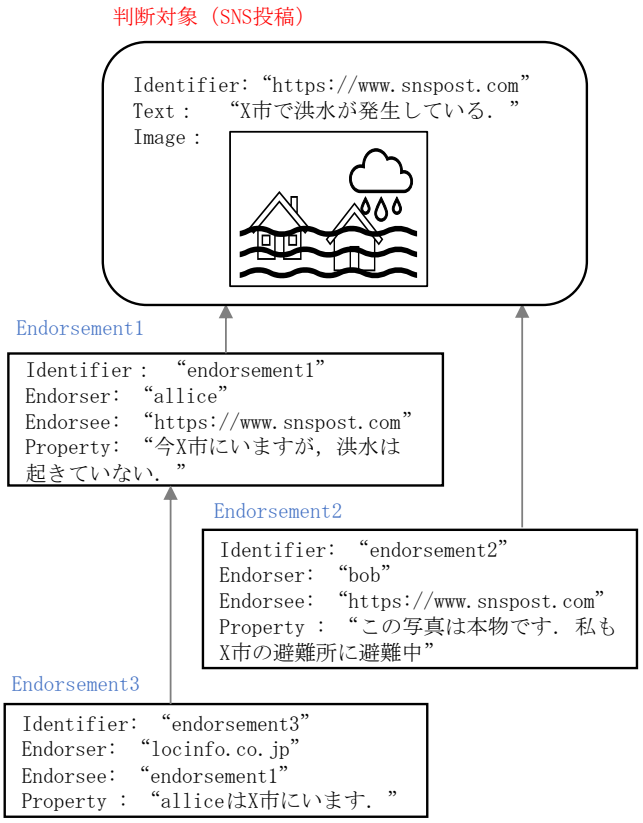


図 3　エンドースメントグラフの具体例

4.1　提案手法の概要

提案手法は，判断対象に付与されたエンドースメントグラフ（EG）を用いて，判断対象の信憑性判定を示すラベル（信憑性ラベル）と根拠説明文を生成する手法である．まず信憑性ラベルは，ラベルが付与された対象に記載された内容が信用できる度合いの高さを示したラベルであり，提案手法では以下の 3 種類存在する．

- 高：内容を信用できる度合いが高い
- 不明：内容を信用できる度合いが不明である
- 低：内容を信用できる度合いが低い

また，提案手法では内容を信用できる度合いが高いエンドースメントに基づいて判断対象の信憑性ラベルを判定するため，エンドースメントにも判断対象と同様の信憑性ラベルを付与する．そして，信憑性ラベルが「高」であるエンドースメントのプロパティに記述された文字を用いて根拠説明文を生成する．以下では，提案手法が仮定している前提条件および提案手法全体の処理の流れを述べる．

4.1.1　提案手法における前提条件

提案手法では，まず常に内容を信用できる度合いが高いエンドースメントを付与する信頼可能なエンドーサが存在すると仮定している．具体的には，図 3 における X 市等の自治体や行政機関を前述のエンドーサとして想定している．信頼可能なエンドーサの識別子が記載された後述の信頼リストが事前に作成されていると仮定しており，提案方式では，信頼リストを用いてエンドーサの信頼度を示すラベル（信頼度ラベル）を付与する．

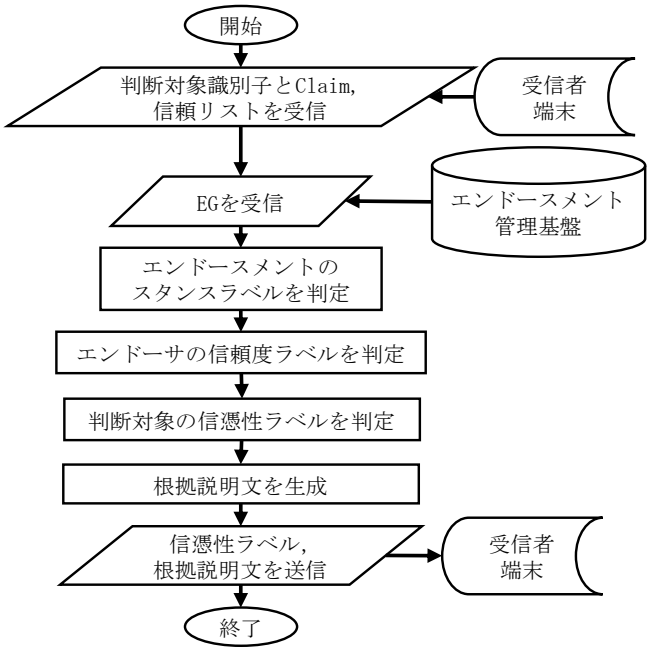


図 4　提案手法で行う処理の流れ（フローチャート）

また，判断対象を根とした EG がエンドースメント管理基盤から取得可能であると仮定している．

4.1.2　提案手法全体の処理の流れ

まず提案手法で行う処理の流れを図 4 で示す．提案手法では，まず受信者端末側から判断対象の識別子（例：URL），判断対象について判断したい事柄である Claim（例：投稿内の写真は X 市の洪水の様子を表したものである.）および信頼リストを受信する．信頼リストには，エンドースメントのプロパティの内容の正しさについて受信者が信頼するエンドーサの識別子が記載されている．前述の情報を受信した後，判断対象を根とした EG をエンドースメント管理基盤から取得する．次に，取得した EG に含まれた各エンドースメントのエンドーシに対するスタンスを示すラベル（スタンスラベル）を判定した後，信頼リストを用いて各エンドースメントのエンドーサの信頼度ラベルを判定する．そして，前述の判定結果を用いて判断対象の信憑性ラベルと根拠説明文の生成し受信者端末に送信する．

次節以降では，提案手法を構成する各処理について述べる．

4.2　スタンス判定処理

EG に含まれる各エンドースメントがエンドーシを肯定するものか否かを判定するために，提案手法ではエンドースメントのエンドーシに対するスタンスを示したスタンスラベルを判定する．提案手法では，エンドースメントに対して以下 2 種類のスタンスラベルを付与する．

- TRUE：エンドーシを肯定
- FALSE：エンドーシを否定

スタンス判定は論文 [9] で提案されている手法や，スタンスラベルを判定するためのプロンプトを汎用的な大規模言語モデル（LLM）に入力する等の手法で実現する．また，スタンスラベルに応じてエンドースメント j のエンドーシ i に対するスタンス・スコア $s_{j,i}$ も算出する．具体的には，スタンスラベルが「TRUE」である場合には $s_{j,i} = 1$,「FALSE」である場合には $s_{j,i} = -1$ とする．

```
{
  “trustable_endorser”：[
          ⋮
    “locinfo.co.jp”，
          ⋮
  ]
}
```

図 5　信頼リスト (json を用いた記述例)

4.3　エンドーサの信頼度判定処理

EG に含まれる各ノード（判断対象とエンドースメント）の信憑性ラベルを判定するために，エンドースメントのエンドーサの信頼度ラベルを判定する．エンドーサの信頼度ラベルを判定するために，本論文では信頼リストを導入した．信頼リストには，エンドースメントの内容が正しいと受信者が判断できるエンドーサの識別子（例：エンドーサが組織やサービスである場合はそれらが取得したドメイン名）が記載されている．一例として，ファイルフォーマットとして json を用いた場合の信頼リストの例を図 5 に示す．図 5 に示した信頼リストには，図 3 における位置情報提供サービスのドメイン名が記載されている．

次に，信頼リストを用いたエンドーサの信頼度ラベルを判定する手法について述べる．提案手法ではエンドーサ識別子と信頼リストを照合した結果に応じて，下記のエンドーサの信頼度ラベルを付与する．

- HIGH：エンドーサ識別子が信頼リストに存在
- LOW：上記以外

また，エンドーサの信頼度ラベルに応じて，エンドースメント i のエンドーサの信頼度スコア t_i を算出する．具体的には，信頼度ラベルが「HIGH」である場合には $t_i = 1$,「LOW」である場合には $t_i = 0$ とする．

4.4　信憑性判定処理

4.2. の処理で算出したスタンス・スコアと 4.3. の処理で算出したエンドーサの信頼度スコアに基づいて，EG に含まれる各ノード（エンドースメントと判断対象）の信憑性ラベルを判定する．信憑性判定処理では，EG に含まれる各ノードに対して算出した信憑性スコア c_i に応じて以下のルールで信憑性ラベルを付与する．

- if $c_i > 0$, then 信憑性ラベルは「高」
- if $c_i = 0$, then 信憑性ラベルは「不明」
- if $c_i < 0$, then 信憑性ラベルは「低」

判定処理の説明にあたり，図 3 の EG ついてスタンス・スコアとエンドーサの信頼度スコア，およびノードの信憑性スコアのみが記載した EG を図 6 に例示する．図 6 に示すように，まず EG に含まれる各ノードにはインデックスが割り振られて

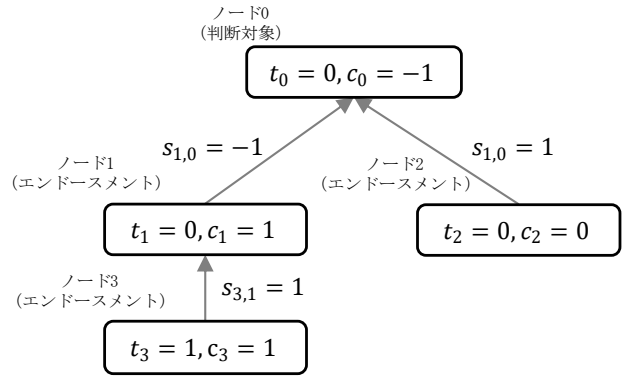


図 6　スタンス・スコアとエンドーサ信頼度スコア，ノードの信憑性スコアが記載された EG

いる．このインデックスの割り振りにおいて判断対象には必ず 0 番のインデックスを割り振る．エンドースメントに対するインデックスの割り振りは，判断対象を起点とした EG に対する幅優先探索の順序に基づき昇順で行う．そのため，インデックスが最大のノード（図 6 の例ではノード 3）は判断対象からのホップ数が必ず最大となる．また，ノード i とノード j を接続するエッジに割り振られた値である $s_{j,i}$ はノード j のノード i に対するスタンス・スコアを表す．そして，EG におけるノード i の各値について，t_i はノード i のエンドーサの信頼度スコアを表し，c_i はノード i の信憑性スコアを表す．

次に，各ノードの信憑性スコア c_i を算出する手順を説明する．まず信憑性スコア算出処理のフローを図 7 に示す．図 7 において，N は判断対象を起点とした EG に含まれたエンドースメントの数である．また，各ノードに紐づいたエンドースメントのスタンス・スコアと信憑性スコアに基づいてノードの信憑性スコアを算出するために，信憑性スコア算出処理はノードのインデックスに対して降順で行う．各ノード i の信憑性スコア c_i を算出するにあたり，まず EG に含まれるすべてのノード j のノード i に対するスタンス・スコア $s_{j,i}$ をスタンス判定処理の結果から取得する．ただし，信憑性スコア算出処理において，ノード i に接続されていないノードとノード i 自身のスタンス・スコアは $s_{j,i} = 0$ とする．次に，エンドーサの信頼度判定処理の結果からノード i のエンドーサの信頼度スコア t_i を取得する．前述のスコアの取得後，ノード i の信憑性スコア c_i を次式で算出する．

$$c_i = \begin{cases} 1 & (t_i = 1) \\ \sum_{j=0}^{N} s_{j,i} u(c_j) & (t_i = 0) \end{cases} \quad (1)$$

式（1）において，$u(c_j)$ は次式のステップ関数である．

$$u(c_j) = \begin{cases} 1 & (c_j > 0) \\ 0 & (c_j \leqq 0) \end{cases} \quad (2)$$

そして，前述の各処理をすべてのノードに対して行うことで算出した信憑性スコア c_i に基づいて各ノード i に信憑性ラベルを付与する．

ここで，提案手法における信憑性スコア c_i の算出式を式（1）とした理由を説明する．まず 4.1.1. で述べた信頼できるエンドーサに関する仮定より，信頼リストに含まれたエンドーサによるエンドースメントの信憑性スコアは常に正とする必要がある．そのため，式（1）において $t_i = 1$ の場合には $c_i = 1$ となるように立式した．具体例をあげると，図 6 におけるノード 3 の信憑性スコアは，$t_3 = 1$ であるため $c_3 = 1$ となる．一方で，図 6 におけるノード 2 の信憑性スコアは，$t_2 = 0$ であり，なおかつノード 2 にはエンドースメントが付与されていないため $c_2 = 0$ となる．

また，信頼リストにエンドーサが含まれていないノードについては，信憑性スコアが正であるエンドースメントに基づいてノードの信憑性スコアを算出する必要がある．そのため，式（2）を適用し，$t_i = 0$ であるノードの信憑性スコアを算出する場合には信憑性スコアが正であるエンドースメントのスタンス・スコアのみ c_i に反映されるよう式（1）を立式した．具体例をあげると，図 6 におけるノード 1 の信憑性スコアは，$t_3 = 1$ であり $s_{3,1} = 1$ であるノード 3 のみがノード 1 に付与されているため $c_1 = 1$ となる．一方で，ノードに複数のエンドースメントが付与されている場合には，信憑性スコアが正であるエンドースメントのスタンス・スコアのみがノードの信憑性スコア算出に影響する．そのため，図 6 におけるノード 0 の信憑性スコア算出では $s_{1,0} = -1$ のみが影響するため，$c_0 = -1$ となる．

4.5　根拠説明文生成処理

4.4. で説明した判断対象の信憑性判定に寄与したエンドースメントを用いて信憑性判定の根拠説明文を生成する．まず，判断対象に付与されたエンドースメントから，信憑性判定の根拠となったエンドースメントのみをエンドースメントのスタンスラベルと信憑性ラベルに基づいて抽出する．具体的には，判断

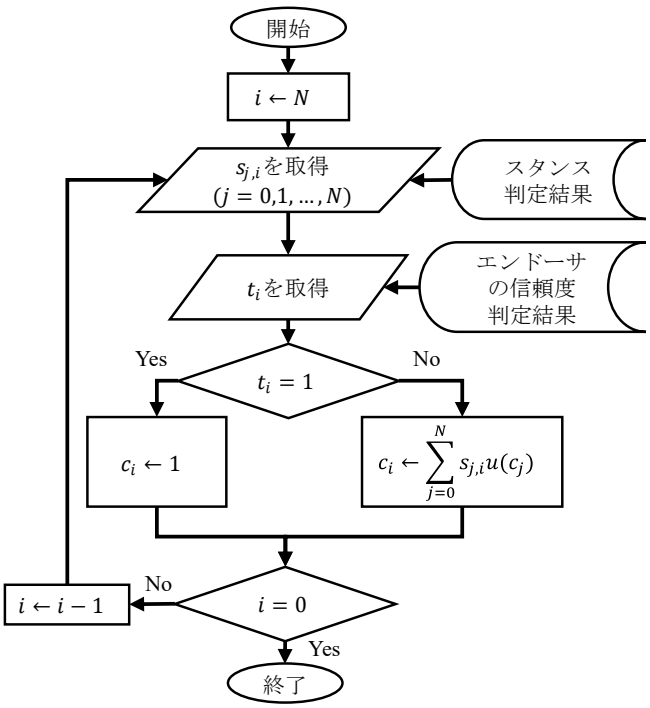


図 7　信憑性スコア算出処理のフローチャート

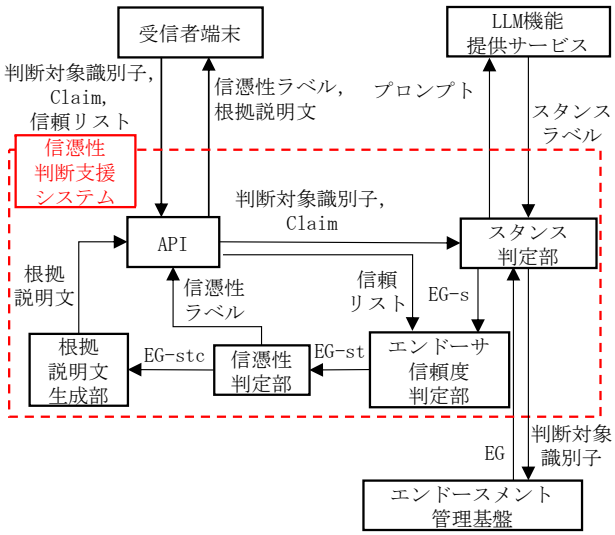


図 8　信憑性判断支援システムの構成

対象に付与された信憑性ラベルが「高」である場合，判断対象に付与されたエンドースメントから信憑性ラベルが「高」でありスランスラベルが「TRUE」であるものを抽出する．逆に，判断対象の信憑性ラベルが「低」である場合は，判断対象に付与されたエンドースメントから信憑性ラベルが「高」でありスランスラベルが「FALSE」であるものを抽出する．また，判断対象の信憑性ラベルが「低」である場合は，判断対象に付与されたエンドースメントから信憑性ラベルが「高」であるものをすべて抽出する．

次に，抽出された各エンドースメントを用いて根拠説明文を生成する．提案手法ではエンドースメントの内容を正確に反映した根拠説明文を作るために，各エンドースメントに記載された文字列を下記の template の対応する＜ ＞内に挿入することで根拠説明文を生成する．

- template：＜エンドーサ＞は＜エンドーシ＞について「＜プロパティ＞」と主張している．

5　インターネット上のデータの信憑性判定・根拠説明文生成システムの実装および性能評価

本章では，本論文で提案する手法を用いた情報の信憑性判断支援システムの実装および初歩的な性能評価，評価結果に基づく考察について述べる．

5.1　信憑性判断支援システムの構成

まず，信憑性判断支援システムの構成を図 8 に示す．

信憑性判断支援システムでは，まず受信者端末側からシステムの API を介して判断対象の識別子，判断対象について判断したい事柄である Claim，および信頼リストを入力する．前述の入力が信憑性判断支援システムに行われた後，判断対象を根としたエンドースメントグラフ（EG）をエンドースメント管理基盤から取得する．本論文における実装では，エンドースメントを json で作成し保存する機能とエンドースメントグラフを取得する機能を実装することで，エンドースメント管理基盤を簡易的に実現した．

EG を取得した後，EG に含まれた各エンドースメントのエンドーシに対するスタンスラベルを LLM に判定させるためのプロンプトをスタンス判定部で生成する．そして，LLM 機能を提供するサービスの API に前述のプロンプトをスタンス判定部から送信する．本論文おける実装では，スタンスラベルを判定させるための LLM として Azure OpenAI Service によって提供される GPT-4 を用いた．EG に含まれるすべてのエンドースメントのスタンス判定が修了した後，EG のエンドースメントにスタンスラベルを追記した EG-s をエンドーサ信頼度判定部に送信する．

次に，エンドーサ信頼度判定部では，前述の EG-s に含まれる各エンドースメントのエンドーサの識別子を信頼リストと照合することでエンドーサの信頼度ラベルを判定する．EG-s に含まれるすべてのエンドースメントに対する判定が終了したあと，エンドーサの信頼度ラベルを EG-s の各エンドースメントに追記したものを EG-st として信憑性判定部に送信する．

そして，信憑性判定部では前述の EG-st に記載されたスタンスラベルとエンドーサの信頼度ラベルを用いて 4.4. の処理を行うことで，EG-st の各ノードの信憑性ラベルを判定する．EG-st に含まれるすべてのノードに対する判定が終了した後，判断対象に対する信憑性ラベルを API に送信する．また，EG-st に各ノードの信憑性ラベルを追記した EG-stc を根拠説明文生成部に送信する．

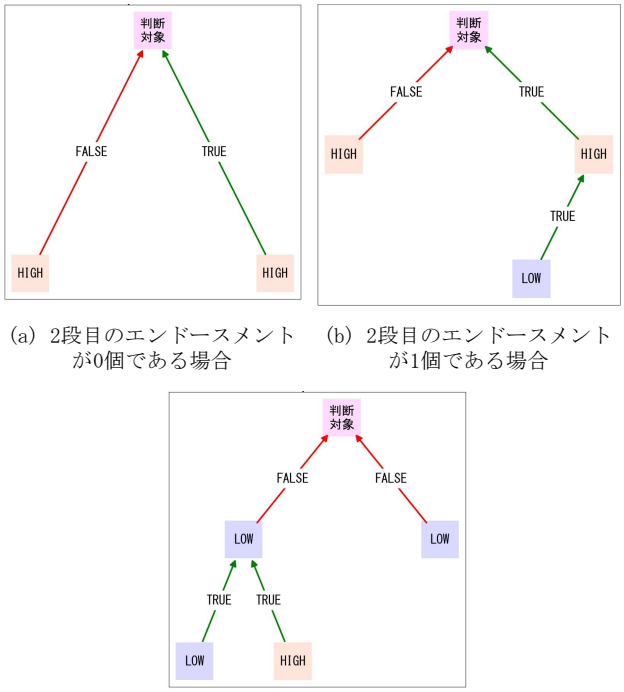
根拠説明文生成部では，EG-stc に記載された各ノードのスタンスラベルと信憑性ラベルに基づいて 4.5. の処理を行うことで生成した信憑性ラベルの根拠説明文を API に送信する．

最後に，前述の信憑性判定と根拠説明文を受信者端末に送信する．

5.2　性能評価方法

本節では提案手法の性能を評価する方法を述べる．提案手法は受信者による信憑性の判定を支援する手法であることから，提案手法による信憑性の判定と人による信憑性の判断を比較すべきと考えられる．特に，提案手法のコアとなる 4.4. で述べた信憑性判定手法の妥当性をまず評価する必要があると考えられる．前述の観点より，我々は提案手法の初歩的な性能評価として，抽象的な EG が与えられたときの人による判断対象の信憑性判断と 4.4. で述べた判断対象の信憑性判定処理で付与した信憑性ラベルを比較した．

具体的な性能評価の方法としては，まず図 9 に示すように判断対象とエンドースメントを抽象的に表現した疑似 EG を 166 件作成した．疑似 EG の矢印に振られたラベルはエンドースメント（矢印の出元）のエンドーシ（矢印が支持する先）に対するスタンスラベル（TRUE/FALSE）を示す．各エンドースメントに振られたラベルは各エンドースメントのエンドーサの信頼度ラベル（HIGH/LOW）を示す．今回の評価において作成した疑似 EG では，判断対象に直接付与されたエンドースメント（1 段目のエンドースメント）が必ず 2 個存在する．また，エンドースメントに付与されたエンドースメント（2 段目のエ



(a) 2段目のエンドースメントが0個である場合　(b) 2段目のエンドースメントが1個である場合

(c) 2段目のエンドースメントが2個である場合

図 9　作成した疑似 EG の例

ンドースメント）は最大で 2 個存在する．

今回の性能評価において，前述の各疑似 EG に対して，ラベル付与者（アノテータ）3 名が付与した判断対象の信憑性ラベルに対して多数決を取ることでアノテーションラベルを作成した．アノテーションラベルには「高」,「不明」,「低」の 3 種類が存在する．なお，前述の多数決において，各アノテータによって付与された信憑性ラベルがすべて異なる場合には，アノテーションラベルは「不明」とした．また，各疑似 EG を図 8 の EG-st として信憑性判定部に入力することで，各疑似 EG に対する判断対象の信憑性ラベルを出力した．そして，各疑似 EG に対する信憑性ラベルとアノテーションラベルを比較した結果を混同行列としてまとめた．

5.3 評価結果

本節では，5.2. で説明した評価方法に基づく性能評価結果を述べる．まず，前述の判断対象の信憑性判定機能による信憑性ラベルとアノテーションラベルに基づいて作成した混同行列を表 1 に示す．表 1 より，アノテーションラベルが「高」である 55 件の疑似 EG のうち 48 件は「高」，残り 7 件が「不明」と信憑性ラベルが付与されたとわかる．また，アノテーション

表 1　アノテーションラベルと信憑性ラベルの一致/不一致をまとめた混同行列

		アノテーションラベル		
		高	不明	低
信憑性ラベル	高	48	7	0
	不明	7	36	13
	低	0	7	48

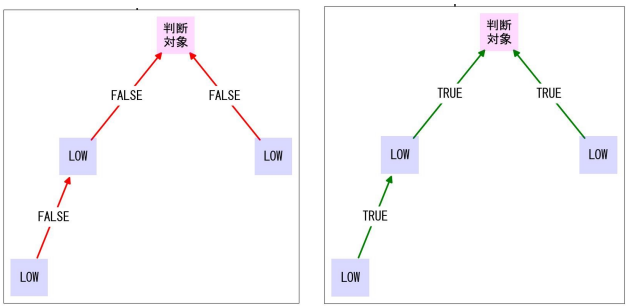


図 10　信頼度ラベルが「LOW」であるエンドースメントのみ含まれた疑似 EG

ラベルが「不明」である 50 件の疑似 EG のうち，36 件が「不明」，7 件は「高」，残り 7 件が「低」と信憑性ラベルが付与されたとわかる．そして，アノテーションラベルが「低」である 61 件の疑似 EG については，48 件が「低」，残り 13 件が「不明」と信憑性ラベルが付与されたとわかる．

5.4 考　　察

本節では 5.3. の性能評価結果に対する考察を述べる．

まず，図 10 に例示するように，エンドーサの信頼度ラベルが「LOW」であるエンドースメントのみ含まれた疑似 EG に対する信憑性ラベルとアノテータが付けたラベルを比較する．提案手法では信憑性の低いエンドースメントのスタンスラベルを無視して信憑性ラベルを判定しているため，前述の疑似 EG が入力された場合に判断対象の信憑性ラベルはすべて「不明」となる．一方で，各アノテータが付与したラベルに着目すると，2 名のアノテータが前述の疑似 EG についてもエンドースメントのスタンスラベルを考慮して判断対象の信憑性を判断する傾向が見られた．また，残り 1 名のアノテータは，提案手法と同様に信憑性の低いエンドースメントのスタンスラベルを無視して判断対象の信憑性を判断する傾向が見られた．

次に，図 11 に例示するように，エンドーサの信頼度ラベルが「HIGH」であるエンドースメントを否定するサブエンドースメントが含まれた疑似 EG に対する信憑性ラベルとアノテータが付けたラベルを比較する．提案手法では，エンドーサの信頼度ラベルが「HIGH」であるエンドースメントに対してはサブエンドースメントのスタンスラベルによらず信憑性ラベルを「高」としている．そのため，提案手法では前述の疑似 EG が入力された場合にはサブエンドースメントを無視し，エンドーサの信頼度ラベルが「HIGH」であるエンドースメントのみに基づいて判断対象の信憑性ラベルを判定している．一方で，各アノテータは前述の疑似 EG については判断対象の信憑性を「不明」と判断する傾向が見られた．また，前述の疑似 EG について 3 名のアノテータにより付与されたラベルが一致しない場合も複数件見られた．

以上より，同じラベルが付与された EG に対する人の信憑性判断には個人差があるため，提案手法でも前述の個人差を考慮して信憑性ラベルを判定する必要があると考えられる．

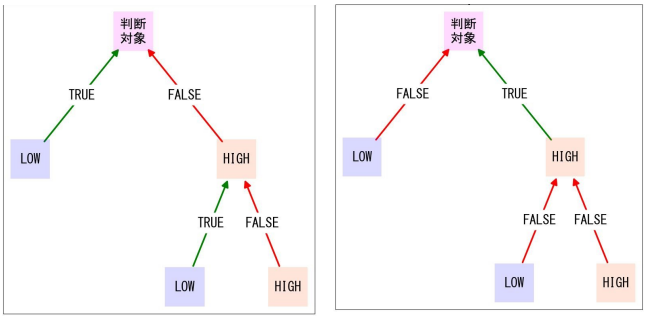


図 11　エンドーサの信頼度ラベルが「HIGH」であるエンドースメントを否定するサブエンドースメントが含まれた疑似 EG

6　おわりに

本論文では，Trustable Internet において信憑性の高い付加情報（エンドースメント）のみに基づいてデータの信憑性の判定と根拠説明文を生成する手法を提案した．具体的には，データの信憑性を判定し根拠説明文を生成する従来手法とその問題点を述べたのちに，提案する信憑性の判定と根拠説明文を生成する手法を述べた．そして，提案手法の初歩的な性能評価と評価結果に関する考察を述べた．

今後の課題として，まず受信者の判断傾向に合わせた判断対象の信憑性判定を出力できるように提案手法を改良することが挙げられる．また，本論文で述べた信憑性判定処理ではエンドースメントのスタンスとエンドーサの信頼度のみを用いて判断対象の信憑性を判定している．しかしながら，受信者による判断対象の信憑性の判断は，エンドースメントの内容にも依存すると考えられる．そのため，判断対象の信憑性判定においてエンドースメントのプロパティに記載された文章も加味するように提案手法を改良することも今後の課題として挙げられる．

文　献

[1] 総務省. 情報通信白書令和５年度版. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r05/html/XXXXXX.html.
[2] 笹原和俊. ウェブの功罪. 情報の科学と技術, Vol. 70, No. 6, pp. 309–314, 2020.
[3] 関谷直也. 〈虚実〉のコミュニケーションとしての「うわさ」「流言」「風評」. 哲学, Vol. 2022, No. 73, pp. 37–47, 2022.
[4] フェイクニュースの生態系. 青弓社ライブラリー. 株式会社青弓社, 2021.
[5] 佐久間義友, 牛田芽生恵, 三宅康太, 坂本拓也, 阿部涼介, 鈴木茂哉, 中村修. インターネット上のデータの確からしさを確認可能にするための trustable internet のアーキテクチャの提案. Technical Report 1, 情報処理学会, jan 2023.
[6] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 1388–1394. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.
[7] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7352–7364, Online, July 2020. Association for Computational Linguistics.
[8] 田中克己, 山本祐輔. 2. 情報メディアとその信憑性. 映像情報メディア学会誌, Vol. 66, No. 11, pp. 891–895, 2012.
[9] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年3月1日(金) 10:00 ~ 12:10 | T3-B オンライン（Zoom Events)

ドメイン指向推薦 (1)

座長:黒川 茂莉(株式会社KDDI総合研究所)
コメンテータ:杉山 一成(大阪成蹊大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T3-B-7-01] 能力と成績の乖離における将来性に着目した推薦システムの開発

*神 一輝[1]、張 建偉[1] (1. 岩手大学)

10:30 ~ 10:55
[T3-B-7-02] 楽天レビューデータを用いたレビュー推薦の一手法

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

10:55 ~ 11:20
[T3-B-7-03] スポットの雰囲気分析によるユーザの状況と気分の選択に合わせた楽曲推薦手法の検証

*丸岡 史奈[1]、栗 達[2]、熊本 忠彦[3]、小野 晋太郎[2]、河合 由起子[1] (1. 京都産業大学、2. 福岡大学、3. 千葉工業大学)

11:20 ~ 11:45
[T3-B-7-04] 旅行者の時空間的・感性的条件に基づく旅程と経路周辺店舗の組合せ生成方式

*神原 隼登[1]、林 康弘[1]、清木 康[1] (1. 武蔵野大学データサイエンス学科)

11:45 ~ 12:10
[T3-B-7-05] 【技術報告】サイバーエージェントメディア事業部における推薦システム

松月大輔 (株式会社サイバーエージェント)

DEIM Forum 2024 T3-B-7-01

能力と成績の乖離における 将来性に着目した推薦システムの開発

神　一輝† 張　建偉††

† 岩手大学大学院総合科学研究科　〒 020-8551 岩手県盛岡市上田 4 丁目 3 - 5
†† 岩手大学理工学部　〒 020-8551 岩手県盛岡市上田 4 丁目 3 - 5
E-mail: †{g0322102,zhang}@iwate-u.ac.jp

あらまし　近年，スポーツ産業におけるデータ分析は著しく発展し，多角的な視点から選手を分析することが可能となった．最先端技術を活用したシステムの導入により，選手の能力が数値化されたことで，能力は高いが成績は伸び悩んでいる選手がデータ上で明らかとなった．本研究では，このような選手は長期的な視点から見ると，成績は能力に収束していくという仮説を立てた上で，能力と成績に関して別々にクラスタリングを行い，能力の高いクラスターに所属しているにも関わらず，成績が低いクラスターに所属する選手，即ち自身の能力と成績に乖離が見られる選手を特定する手法を提案する．MLB に所属する野手に対して実験を行った結果，特定した選手の多くが翌年以降に成長を遂げ，将来性の高い選手を確認できた．

キーワード　推薦システム，スポーツデータ活用，クラスタリング

1 はじめに

近年，インターネットの普及に伴い，情報社会は顕著に拡大している [1]．ビッグデータの増加や分析手段の進化に加え，スマートフォンの普及と検索エンジンの性能向上により，情報へのアクセスも容易になったことで，情報の活用方法の重要性が高まっている [2]．現在，多くの国内外企業がデータ分析を駆使し，顧客や取引先のデータを分析することで利益を増加させたり，製品や従業員の評価を行ったりなど，企業の競争力を強化している [3]．このような傾向は，スポーツ産業において特に顕著に見られる．2013 年には，当時野球界で最もデータ分析に力を入れていたチームの 1 つであるヒューストン・アストロズが，戦力外通告寸前のコリン・マクヒュー投手のカーブボールの優秀さに目を付け，翌年以降の成績向上に繋げるなど，データ分析を駆使して様々な選手の才能を開花させた [4]．以降ヒューストン・アストロズは 2015 年から 2023 年現在まで，2016 年を除いて 8 年連続でポストシーズンに進出しており，内 2017 年・2023 年にワールドシリーズを制覇するなど，MLB 屈指の強豪チームへと成長した．また 2014 年には，野球界におけるデータ分析の先駆者，アンドリュー・フリードマンが，5 年総額 3500 万ドルという破格の契約でロサンゼルス・ドジャースへ移籍するなど [5]，スポーツ界におけるデータ分析の需要は高まっており，現在も様々なチームがデータを活用し，選手の評価や分析を行っている．

よって本研究では，スポーツ選手の評価における，選手育成を通じたチームの成績向上への貢献に焦点を当てる．スポーツ選手の評価を行う上で本研究では選手を，今後の将来性に重点を置いて評価する．通常，選手の評価は過去の成績に基づいて行われることが多い [6]．しかし，個々の成長速度の違いを考慮すると，急速に成長する選手や成長が停滞する選手も存在する．能力以上の活躍をする選手もいれば，能力を活かせず，初めは成績が伸び悩む選手も存在する [7]．本研究ではこれらの多様性を踏まえ，長期的な活躍が期待できるという視点から選手を評価する際，最終的に選手の成績は能力に収束していくのではないかと考える．

通常の評価方法では，能力はあるが成績が伸び悩んでいる選手，即ち能力と成績に乖離がある選手の可能性が見過ごされる場合があり，これは才能が開花する機会を失わせるリスクを孕んでいる [8]．特にスポーツ選手は，個々の成長速度が異なる場合が多く，能力と成績の乖離が顕著に現れやすい [9]．例えば練習での能力は高いが，試合では成績を残せない，そのようなケースは多く存在する [10] [11]．しかし，経験不足や環境適応といった問題が解消された際，成績は能力に収束していくのではないかと考えられる．したがって本研究では，長期的な視点での能力と成績の関係を検証し，将来性の高い選手の特定を目的とする．

本研究では，能力と成績のレベルを測る目的としてクラスタリングを使用する．その際既存研究は，選手を能力基準でクラスタリングし，そのクラスター順位が高いグループの選手を成績順にソートし，順位が低い選手を特定する．つまり，能力面はクラスタリングによって評価するが，成績面はソートによって評価を行う．一方提案手法では，能力・成績共にクラスタリングを行い，両者クラスタリングの順位を基準として評価を行う為，タイプの違う選手や群を抜いて優れた選手など，様々なデータに対応しやすい．更に，既存手法のように能力のクラスター順位が高いグループといった形で，選手を限定せずに特定を行うため，より多くの該当選手を特定する事ができる．更に広い枠組みで考えると，スポーツ界で行われる評価は，主に現時点での成績を基に評価が行われるケースが多く，能力面は個

人的主観で測られる場合もある．しかし選手の能力と成績の関係性に焦点を当てて研究を行う事で，これまで開花せずに終わった才能が開花する可能性が上がる．この研究の貢献は次の通りである．

・能力と成績の関係性に着目した将来性の高い選手を特定する為の手法の提案

・実験を通した能力と成績の乖離に着目した将来性の高い選手を特定する有効性の証明

2 関連研究

2.1 スポーツ選手の評価

Johnston 氏らは，メジャースポーツにおいて高い能力を持つ選手とそうでない選手の違いに関して，将来の成功を予測するための変数を調査した結果，両者を分ける要因には多様な要素が存在し，個人の属性によって異なることが明らかになった [12]．Mitić 氏らは，特定の心理的特性 (自己効力感，精神の知能指数，性格など) が高い成績を残す選手とそうでない選手を区別する要因として働くかについて，t 検定や ANOVA などの比較統計分析を用いて調査し，心理的特性が両者を区別する要因として重要であることを発見した [13]．Piepiora 氏は，様々なスポーツのポーランド人アスリートを対象に，性格特性が高い成績を残す選手とそうでない選手を区別する要因としてどのように機能するかを，BigFive 性格モデル (NEO-FFI) を用いて調査した．その結果，高い成績を残す選手は神経症的傾向が低く，外向性，協調性，誠実性が高い傾向にあることが示された [14]．

これらの研究成果を踏まえ，本研究では高い成績を残す選手とそうでない選手を区別する要因を，プロスポーツ選手として長期的に成績を維持するという観点から評価する．仮説として，成績が高い選手は高い能力が裏付けられている可能性が高く，反対に成績が低い選手は根本的な能力不足が原因である可能性が高いと考える．

2.2 クラスタリングを用いた推薦システム

Eyad 氏は，コンテキスト情報を利用した推薦システムに関して，コンテキスト情報と k-means を用いて，ユーザーアイテム行列の新しい形式を生成する 2 つの手法を提案した結果，従来の推薦システムよりも，より正確で関連性の高い推薦を行うことができることを示した [15]．Shepitsen 氏らは，SNS 上のユーザーの興味や嗜好に合わせたコンテンツやタグの推薦システムに関して，階層的クラスタリングを用いてタグをトピック領域に分類し，ユーザーの現在のナビゲーションコンテキストに応じてクラスタを選択するという手法を提案し，推薦精度と新規性の向上を示した [16]．Irina 氏らは，推薦システムにクラスタリングを事前処理として適用することでシステムの性能向上が期待されることに関して，様々な論文を調査した結果，クラスタリングを用いることで推薦システムの多様性，一貫性，信頼性の向上やデータスパースネスやユーザー嗜好の時間変化といった問題に対処できることが確認された [17]．

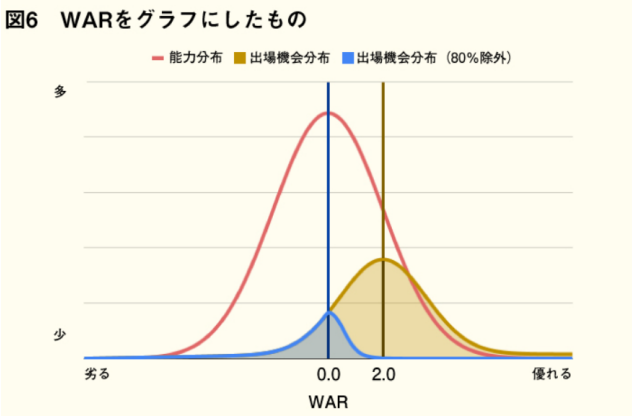


図 1 WAR の分布 (引用: [25])

これらの研究成果を踏まえ，本研究では将来性の高い選手を推薦するために，複数のクラスタリング手法を用いて，選手の特定を行う．能力と成績に関する特徴量を持つ選手をクラスタリングを用いて分類する事で，より精度の高い選手推薦が可能になると考える．

2.3 既存手法の概要

2.3.1 評価指標

既存手法では，選手の成長を評価する指標として，WAR（Wins Above Replacement）を使用した．WAR は，通常比較が困難な選手同士を比較可能にする指標であり，野球界では優れた評価基準として広く用いられている．WAR は 0.0 を中心に左右対称に分布しており，中心から離れるほど選手の数が減少する [25]．WAR には上限や下限がなく，Fangraphs[1]における，21 世紀以降の 1 シーズンでの最高 WAR は 12.7（Barry Bonds, 2002, San Francisco Giants），最低 WAR は-3.9（Neifi Perez, 2002, Kansas City Royals）である．

またデータセットは投手であるため，先発・中継ぎ・抑えといった役割毎の WAR の偏りを減らすために，WAR を投球回数 IP で割った WAR/IP を用いる．

2.3.2 実験の概要

Koseler 氏は，野球分野における機械学習を用いたデータ解析のサーベイ論文を発行した [20]．その中で Ishii 氏は，機械学習分野におけるクラスタリングを用いて，過小評価されている野球選手の特定を行った [21]．Ishii 氏は，2015 年に MLB で 750 球以上投球した投手 320 人を対象に，各投手の球種毎の変化量，回転数，リリース位置といった 15 次元の能力的特徴量を取得し，球種毎に k-means [18] と階層的クラスタリング [19] の 2 種類のクラスタリングを行った．その際，各球種においてそれぞれ，優れた能力を持つ投手が所属するグループの平均防御率と比較し，その差が最も大きい投手を過小評価選手として抽出した．

そして球種毎に抽出した過小評価選手 10 人の，2015 年の WAR/IP と 2016 年の WAR/IP を比較した．結果として，k-means と階層的クラスタリングの両方において，抽出した

1：https://www.fangraphs.com/

過小評価選手全員が，2015 年の WAR/IP に比べ，2016 年の WAR/IP が向上した．

3 提案手法

3.1 ベースラインの概要

本研究では，投手を対象とした既存手法を，野手に対して適応させることを試みる．投手は主体的立場であるのに対し，野手は受動的立場であるため，対応力が必要な野手は，高い能力を持っていても，成績が低い選手が多く，既存手法の趣旨に該当する選手の数が多いと考える．また野手は，年数と共に経験値が上がる事で本来の能力に成績が収束する場合があるため，投手よりも成績が向上する選手の数が多く，成績の上昇率も高い事が期待される．

次に，既存手法を野手に適応させるにあたり，いくつか既存手法を変更した，ベースラインを説明する（図 5）．A・B・C・D はそれぞれ能力と成績の関係を基準に分けられた選手で，A は能力も成績も高い選手，B は能力は高いが成績が低い選手，C は能力は低いが成績が高い選手，D は能力も成績も低い選手となっている．この中で能力が高く成績が低い選手 B の特定を目指す．まず，選手の能力的特徴量に基づいてクラスタリングを行い，類似した能力傾向を持つ選手を同じグループに分類する．次にグループ内の選手の能力値を平均化し，各グループの平均能力値を算出する．これに基づき，全体の能力レベルが高いグループを上位に，低いグループを下位に並び替える．即ち，上位はクラスター ID は小さく，下位になるにつれて，クラスター ID は大きくなっていく．この並び替え後，上位 α（閾値）グループが，次のステップに進む該当グループとされる．

続いて，抽出された各グループに属する各選手の成績を基準に，選手を再度並び替え，成績が高い選手を上位に，低い選手を下位に配置する．この成績がグループ平均を下回る選手を各グループから抽出し，将来性の高い選手と予測する．まとめると，能力が高いグループで平均以下の成績を残す選手を，将来性の高い選手と仮定する．

3.2 提案手法の概要

本節では，提案手法の概要を説明する（図 6）．提案手法では既存手法と同様に，全選手を選手の能力に基づいてクラスタリングをする．しかし既存手法との主要な相違点は，選手の成績に基づくクラスタリングも同時に実施する点である．具体的には，提案手法では選手の能力と成績を別々にクラスタリングし，評価する．

クラスタリング後，各グループに属する選手の能力および成績に基づき，グループを順位付けする．この並び替えにより，能力及び成績の高いグループは上位に，低いグループは下位に並び替えられる．即ち，上位グループのクラスター ID は小さく，下位になるにつれて，クラスター ID は大きくなっていく．このグループ群において，選手は自身が属する各グループの順位を比較し，成績基準のグループ順位と能力基準のグループ順位の差が閾値 β を超える選手を，将来性の高い選手として抽出する．

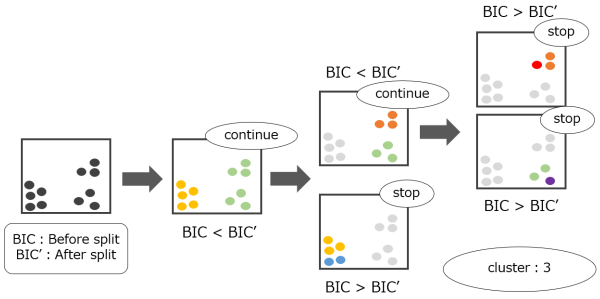


図 2　x-means の流れ

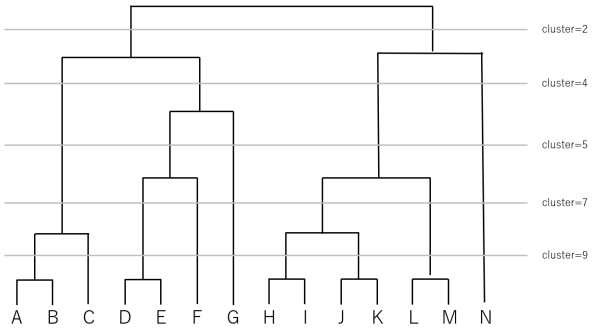


図 3　階層的クラスタリングの流れ

既存手法が能力が高い選手の中から相対的に成績が低い選手を特定するのに対し，提案手法では，能力に比べ成績が閾値以上に低い選手を抽出するという新たなアプローチを採用している．これにより，選手の能力と成績のギャップをより効果的に評価し，将来性が高い選手の特定を目指す．

3.3 クラスタリング

本研究では，教師なし学習として代表的に用いられる階層的クラスタリングと，x-means の 2 つのクラスタリングを用いて研究を行う．

階層的クラスタリングは，データの類似度によってグループに分ける手法の 1 つで，データ間やクラスタ間の距離を計算し，最も近いものから順に統合する [19]．計算量が多くなることもあるが，ウォード法、群平均法、最短距離法、重心法といった複数の距離の測り方によって，扱うデータセットに適した計算方法を指定する事ができる．本研究ではウォード法を用い，階層的クラスタリングを行う．

一方 x-means は，k-means を拡張したアルゴリズムであり，クラスタ数を自動的に推定する特徴をもつ [22]．クラスタリングのアルゴリズムは k-means 同様，指定されたグループ数に準じ，ランダムにグループ分けされたデータポイントの重心を計算し，その重心に近いデータポイントにグループを分け直す．そして再度重心を計算し，グループを分け直すといった作業を，重心が動かなくなるまで行う．こうして指定されたグループ数に分けられた後，x-means は各グループを更に 2 分割していく．分割前と分割後のクラスターを BIC を用いて評価する．この BIC は，モデルの複雑性とデータへの適合度を表す指標であ

表 1　MLB(Hitter) のデータセットの詳細

項目	詳細
期間	2015 - 2022
人数	5,504
特徴量	能力 15 / 成績 15
グループ	能力 20 / 成績 20
評価に用いる指標	WAR

り，この値が悪化すれば終了，しなければ更に 2 分割といった工程を，全ての 2 分割が終了するまで繰り返す（図 3）．その為，本研究のデータセットのような，データポイント間にはっきりとした傾向が見られないデータセットに適している手法である．　また x-means は実行毎に結果に差が開きやすいクラスタリングであるため，結果を安定させるために k-means++を用いた．k-means++は k-means の初期中心点を選択する際に，ランダムで選ぶのではなく，均等に分割する事で実行毎の差を減らす手法である．k-means++は，最初の 1 点はランダムで選び，2 点目は最初の 1 点と最も距離がある点を選ぶ．3 点目以降は，1 点目と 2 点目との距離の平均値が最も大きい点を選ぶ．これにより x-means の実験毎の差が減少し，階層的クラスタリングとの結果を比較することができる．

3.4　デ　ー　タ

本研究で使用するデータセットについて記す（表 1）．データセットは野球を対象とし，Baseball Savant[2]から MLB の打者を対象に抽出したものを使用する．野球は，動作が数値化しやすい特性を持ち [24]，MLB[3]では能力の数値化と，情報の開示が進んでいる [4]．

期間は 2015-2022 年，statcast システム導入以降の計 8 年間で，50 以上の規定打席を満たす 5,504 人（同一選手を含む）が，それぞれ標準化・正規化された能力的特徴量 $(f_a{}^1, f_a{}^2, \ldots, f_a{}^{15})$ と，成績的特徴量 $(f_r{}^1, f_r{}^2, \ldots, f_r{}^{15})$ を持つ，ベクトルとして表現されている．これらのベクトルに対し，x-means，階層的クラスタリングを行った結果，各手法約 20 のグループに分割した．

3.5　特　徴　量

本研究で使用する特徴量について説明する（表 4,5）．選手の能力を測定するために，statcast（トラッキングシステム）[5]から入手可能なデータを活用している．statcast の導入以前，球速や飛距離，走力以外は主に主観的に評価されることが多かったが，導入以降，投手の制球力や野手の選球眼など，さまざまな能力が数値化されるようになった [23].

2：https://baseballsavant.mlb.com/
3：https://www.mlb.com/
4：https://baseballsavant.mlb.com/statcast_leaderboard
5：https://en.wikipedia.org/wiki/Statcast

表 2　MLB(Hitter) のデータセットの詳細

	WAR>0	WAR≦0
WAR>0	Tp	Fp
WAR≦0	Fn	Tn

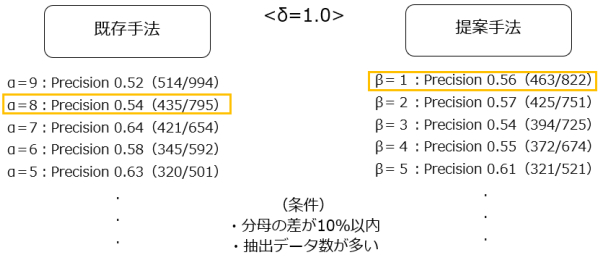


図 4　評価指標の算出方法

特徴量は，能力面と成績面からそれぞれ 15 個ずつ選出し，実験に用いる．成績に関する特徴量は Baseball Savant の Basic Stats から，能力に関する特徴量は Baseball Savant の statcast からそれぞれ抽出した．

4　評 価 方 法

本研究では，2015 年から導入された statcast によって算出された能力的特徴量を用いており，そのためデータの期間は 8 年間に限定されている．本研究では，2015-2020 年の 6 年間の選手データのうち，3 年間のデータを用いてクラスタリングを行う．これは，選手が怪我や不調等で規定打席に到達できない年度があったとしても，実験データとして含められるようにする為である．尚，6 年間のうち，規定打席に到達した年数が 2 年以下の場合は実験データとして含めず，また 4 年以上規定打席に到達した選手については，年数の新しい 3 年間のデータを使用する．それを基に，2021 年または 2022 年の成績が向上する見込みが高い選手を特定する．こちらも同様に，両年規定打席に到達しない場合は実験データとして含めず，両年規定打席に到達した選手は，2021 年のデータを使用する．2015-2020 年の 3 年間の平均 WAR と 2021 年または 2022 年の WAR の差分から，各手法で特定した選手の成長を分析することで，手法の有効性を検証する．

評価は表 2 より，真陽性（TP），偽陽性（FP），真陰性（TN），偽陰性（FN）を用いて行う．各手法で特定した選手 (P) の中で，2015-2020 年の 3 年間の平均 WAR を，2021 年または 2022 年の WAR が，閾値 δ（0.0-1.0）以上上回った場合を TP（将来性のある選手）とし，維持または下回った場合を FP（将来性のない選手）とする．また，各手法で特定されなかった選手 (N) も同様に，2015-2020 年の 3 年間の平均 WAR を，2021 年または 2022 年の WAR が，閾値 δ（0.0-1.0）以上上回った場合を FN（将来性のある選手），維持または下回った場合は TN（将来性のない選手）とする．それらを基に，適合率（Precision），再現率（Recall），及び F1 スコアを計算し，そのスコアで既存

手法と提案手法を比較する（図 7）．Precison は，2021 年または 2022 年の WAR が向上すると予測された選手の内，実際に WAR が向上した選手の割合，Recall は，2021 年または 2022 年の WAR が実際に向上した選手の内，WAR が向上すると予測された選手の割合，そして F1 score は Precision と Recall の調和平均となっている．

また既存手法の閾値 α，提案手法の閾値 β のパラメータの数値により，スコアも複数算出される（図 4）．既存手法の閾値 α は能力を基準にクラスタリングし，能力レベルを基準に各グループをソートした際に，選択する上位グループの数であり，大きくすると特定する選手の数も増加する．一方，提案手法の閾値 β は，能力・成績それぞれでクラスタリングとソートを行った後の，能力グループのランクと成績グループのランクの差分であり，大きくすると特定する選手の数は減少する．これらのパラメータを調整することで算出される複数のスコアの中で，特定する選手の数ができる限り多く，かつ両手法で選出された選手の数が 10 %以内の差で収まるものを，比較の基準として用いる．

5 実験結果

図 8 及び図 9 より，実験結果を確認する．いずれも左から，Precision,Recall,F1 score の閾値 δ 毎の推移を表しており，閾値 δ の範囲は 0.0-1.0 となっている．赤線がベースライン，青線が提案手法の結果である．Precision は，閾値 δ を大きくするにつれて，分母は常に一定だが，分子が減少していくため，基本的にグラフは右肩下がりになる傾向がある．一方 Recall は，閾値 δ を大きくするにつれて，分母は減少していくが，分子は一定であるため，グラフは右肩上がりになる傾向があるが，分子は分母の影響を受けるため，右肩下がりになる場合もある．

まず図 8 の x-means の結果から，提案手法は全ての閾値 δ において，3 つの主要な評価指標においてベースラインを上回る性能を発揮しており，特に Recall における優位性が確認された．Recall の優位性について考えられる理由としては，スタート時点の δ=0.0 における Recall が優れており，更に δ=0.75 までの Recall の上昇率が優れていた事が考えられる．しかし δ=1.0 における Recall が低下していることから提案手法は，大幅に成長する選手の予測性能が悪い可能性が考えられる．また図 9 の階層的クラスタリングの結果においても同様に，提案手法が 3 つの主要な評価指標においてベースラインを上回る性能を発揮している事が確認できた．しかしこちらも，δ=0.75,1.0 における提案手法の Recall の低下が著しいため，今後は大幅に成長する選手の特定手法に改善点が見られる．　そして x-means と階層的クラスタリングの結果を比較すると，全体的に x-means の性能が優れていることが分かる．特に Recall における差が大きい事から，x-means の方が実際に WAR が向上した選手の予測性能が優れており，本データセットに適したクラスタリング手法は，x-means である可能性が高い事が確認された．

表 3　Netflix データセットの詳細

項目	詳細
映画数	4,915
特徴量	能力 5 / 成績 5
グループ	能力 20 / 成績 19
評価指標	gross（興行収入）

6 汎用性の考察

本研究では主に MLB の野手を対象に実験を行ってきた．提案手法の汎用性を検証するために，異なる分野のデータセットへの検証を試みる．異なる分野のデータセットを選択するにあたり，能力と成績で分けることができる特徴量を持つ必要があった．それを踏まえて今回は，Netflix における映画データセットを用い，WAR 指標を興行収入に置き換えて，両手法の性能を比較した．データセットは，能力と成績に関する特徴量を 5 次元ずつ持つ 4,915 本の映画で（表 6, 表 7），k-means 及び階層的クラスタリングによって，約 20 のグループに分類された．スポーツ選手のデータとは異なり，映画は年度毎の興行収入を比較することは通常行われない．したがって，映画の能力的特徴量を基にして，興行収入が設定した閾値 ω を超えるか否かを評価の基準とし，超えると予測された映画を抽出対象とするという流れで，両手法の比較を行った．

図 10 及び図 11 より，実験結果を確認する．それぞれ左から，Precision,Recall,F1 score の閾値 ω 毎の推移を表しており，閾値 ω の範囲を 5 段階に設定した．赤線がベースライン，青線が提案手法の結果である．また MLB の実験結果同様，Precision は右肩下がりの傾向，Recall は右肩上がりの傾向だが，右肩下がりになる場合もある．

まず図 10 の x-means の結果から，提案手法は全ての閾値 ω において，3 つの主要な評価指標においてベースラインを上回る性能を発揮しており，こちらのデータセットにおいても，Recall における優位性が確認された．ベースラインと比較し，提案手法は閾値 ω の推移に関わらず，Recall を維持できている．またこちらのデータセットでは，平均興行収入を大幅に上回る成績を残す映画の予測性能も，良い値を維持出来ている．図 11 の階層的クラスタリングの結果においても同様に，提案手法が 3 つの主要な評価指標においてベースラインを上回る性能を発揮している事が確認できた．Recall の性能差は特に著しく，ベースラインと比較した際の提案手法の優位性が確認された．　そして x-means と階層的クラスタリングの結果を比較すると，MLB の実験結果程の大きな性能の差は確認されなかったが，Precision は階層的クラスタリング，Recall は x-means と，評価指標によって性能差に違いが見られた．その為，映画のデータセットの場合は，用途に適したクラスタリング手法を選択する事が求められる．

7 ま と め

本研究では，高い能力を持っているにも関わらず成績が伸び悩んでいる選手を，将来性の高い選手として定義し，その特定と評価を行った．結果として，提案手法が既存手法を性能面で上回ることを実証した．加えて，汎用性を検証するために，異なるトピックのデータセットにも本手法を適用し，こちらでも既存手法を上回る結果を得た．

今後の課題としては，まず映画のデータセットに関して，能力が成績に反映されるまでの時間的成長が存在しない点に留意する必要がある．したがって，能力と成績の関連性を時間的な観点から評価できるデータセットへの適応が重要となる．また全体としては，データセットの拡大，異なるトピックのデータセット適応，および異なる評価指標を用いた手法の検証などが挙げられる．これらを通じて，予測性能のさらなる向上と汎用性の拡大を目指し，実用的かつ有意義な知見を提供することが重要である．

謝　　辞

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

文　　献

[1] K. T. Kwak, S. Y. Lee, M. Ham and S. W. Lee.“The effects of internet proliferation on search engine and over-the-top service markets”Telecommunications Policy, vol. 45, no. 8, 2021.
[2] B. Schneider, J. Reilly and I. Radu.“Lowering Barriers for Accessing Sensor Data in Education: Lessons Learned from Teaching Multimodal Learning Analytics to Educators”Journal for STEM Education Research, vol. 3, pp. 91–124, 2020.
[3] J. M. Graves, D. A. Abshire, S. Amiri and J. L. Mackelprang.“Disparities in Technology and Broadband Internet Access Across Rurality : Implications for Health and Education”Fam Community Health, vol. 44, no. 4, pp. 257-265, 2021.
[4] E. Sarris. (Oct. 3, 2014). Finding the Next Collin McHugh With Spin Rates. [Online].Available : https://fantasy.fangraphs.com/finding-the-next-collin-mchugh-with-spin-rates.
[5] E. Stephen. (Oct. 24, 2014). Andrew Friedman’s contract reportedly $35 million over 5 years. [Online].Available : https://www.truebluela.com/2014/10/24/7059219/ andrew-friedman-contract-dodgers-35-million.
[6] A. V. Hootegem, O. Røgeberg, B. Bratsberg and T. H. Lyngstad.“Correlation between cognitive ability and educational attainment weakens over birth cohorts”Scientific Reports, vol. 13, no. 1, 2023.
[7] P. Bangirana, J. Menk, C. C. John, M. J. Boivin and J. S. Hodges.“The Association between Cognition and Academic Performance in Ugandan Children Surviving Malaria with Neurological Involvement”PLOS ONE, vol. 8, no. 2, 2013.
[8] K. Till, R. S. Lloyd, S. McCormack, G. Williams, J. Baker and J. C. Eisenmann.“Optimising long-term athletic development : An investigation of practitioners’ knowledge, adherence, practices and challenges”PLOS ONE, vol. 17, no. 1, 2022.
[9] M. Du and X. Yuan.“A survey of competitive sports data visualization and visual analytics”Journal of Visualization, vol. 24, pp. 47-67, 2021.
[10] A. P. Hill, C. S. G. Witcher, J. K. Gotwals and A. F. Leyland.“A Qualitative Study of Perfectionism Among Self-Identified Perfectionists in Sport and the Performing Arts”Sport, Exercise, and Performance Psychology, vol. 4, no. 4, pp. 237-253, 2015.
[11] D. F. Gucciardi, S. Hanton and C. J. Mallett.“Progressing measurement in mental toughness: A case example of the Mental Toughness Questionnaire 48”Sport, Exercise, and Performance Psychology, vol. 1, no. 3, pp. 194-214, 2012.
[12] K. Johnston, N. Wattie, J. Schorer and J. Baker.“Talent Identification in Sport : A Systematic Review”Sports Medicine, vol. 48, no. 1, pp. 97-109, 2018.
[13] P. Mitić, J. Nedeljković, Ž. Bojanić, M. Franceško, I. Milovanović, A. Bianco and P. Drid.“Differences in the Psychological Profiles of Elite and Non-elite Athletes”Frontiers in Psychology, vol. 12, 2021.
[14] P. Piepiora.“Personality Profile of Individual Sports Champions”Brain and Behavior, vol. 11, no. 6, 2021.
[15] E. Kannout.“Context Clustering-based Recommender Systems”FedCSIS, vol. 21, pp. 85-91, 2020.
[16] A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke.“Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering”Proceedings, vol. 341, no. 1, pp. 35-45, 2008.
[17] B. Irina and K. Mikhail. (Nov. 2, 2021). Review of Clustering-Based Recommender Systems. [Online].Available : https://arxiv.org/ftp/arxiv/papers/2109/2109.12839.pdf
[18] J. MacQueen.“Some Methods for Classification and Analysis of Multivariate Observations”Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 281-297, 1967.
[19] S. C. Johnson.“Hierarchical clustering schemes”Psychometrika, vol. 32, no. 3, pp. 241-254, 1967.
[20] K. Koseler and M. Stephan.“Machine Learning Applications in Baseball : A Systematic Literature Review”Applied Artificial Intelligence, vol. 31, no. 9-10, pp. 745-763, 2017.
[21] T. Ishii. (Jan. 9, 2018). CS 229 Machine Learning Final Projects. [Online].Available : https://cs229.stanford.edu/projects2016.html.
[22] D. Pelleg and A. Moore.“X-means : Extending K-means with efficient estimation of the number of clusters”Journal of Machine Learning Research, vol. 1, pp. 75-102, 2002.
[23] S. S. Wulff and W. P. D. Silva.“A multi-criteria approach for evaluating major league baseball batting performance”Journal of Sports Analytics, vol. 8, no. 2, pp. 85-98, 2022.
[24] E. A. E. Gerber and B. A. Craig.“A mixed effects multinomial logistic-normal model for forecasting baseball performance”Journal of Quantitative Analysis in Sports, vol. 10, no. 1, pp. 99-101, 2014.
[25] S. Nikaido. (Aug. 4, 2020). Let’s understand ”replacement level” intuitively. [Online]. Available : https://1point02.jp/op/gnav/column/bs/column.aspx?cid=53643.

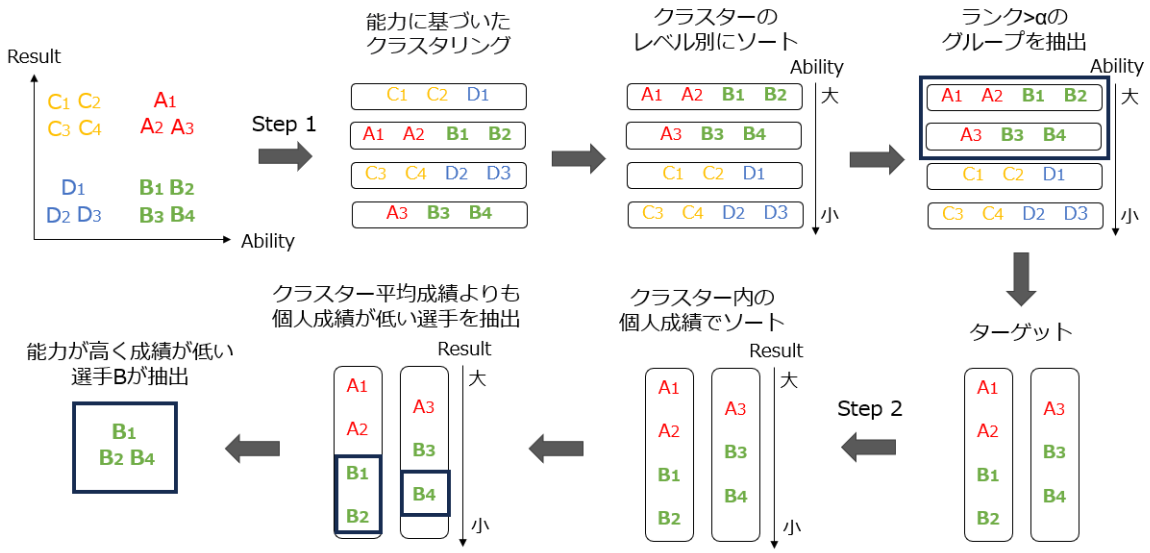


図 5　ベースラインにおける特定選手抽出のアルゴリズム

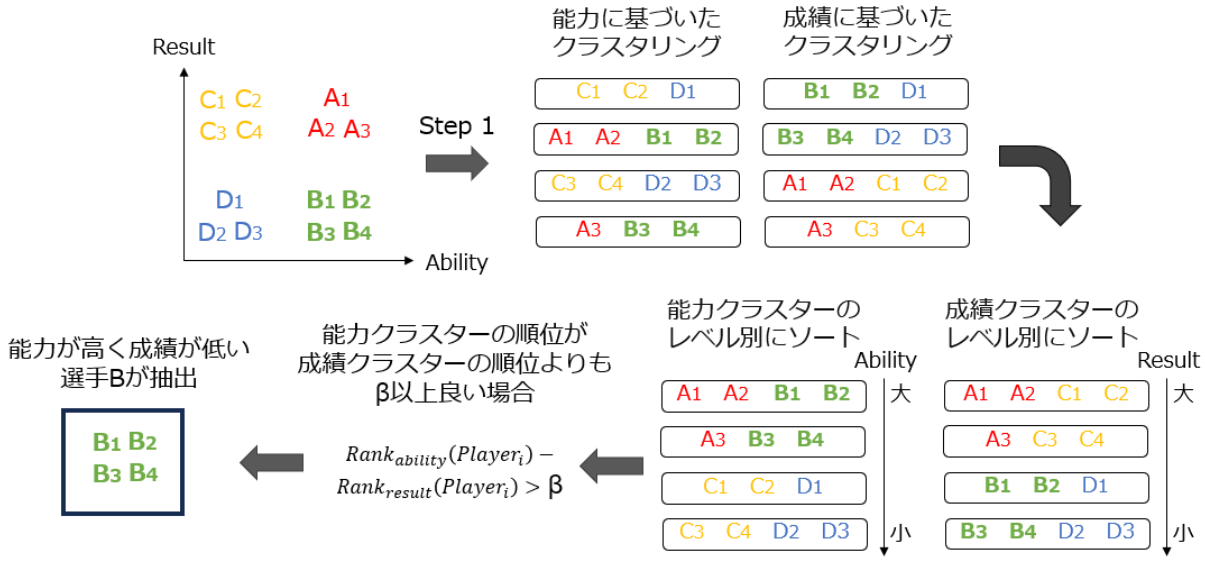


図 6　提案手法における特定選手抽出のアルゴリズム

表 4　能力に関する特徴量 (Hitter)

特徴量	説明
H/GB	安打/ゴロ
HR/FB	本塁打/フライ
Oppo/H	逆方向/安打
Hard/H	153km/h〜/安打
Faul/H	ファール/安打
xOBP-OBP	出塁期待値-出塁率
xSLG-SLG	長打期待値-長打率
HR/avgLA	本塁打率/平均角度
avgEV-maxEV	(平均-最大) 打球速度
1/O-swing%	ボールゾーン スイング率
1/Z-swing&Miss%	ストライクゾーン 凡退率
1/Poor-Weak%	〜96km/h 打球数
Barrel%	157km/h〜 & 26° 〜30° 打球数
Bolts	30ft/s スプリント数
Def%	守備能力

表 5　成績に関する特徴量 (Hitter)

特徴量	説明
H%	安打率
HR%	本塁打率
R%	得点率
RBI%	打点率
AVG	打率
1/K%	三振率
BB%	四球率
SB%	盗塁率
OBP	出塁率
SLG	長打率
LOB%	残塁率
Clutch	得点圏打率
wRC+	チーム・球場補正
UBR	走塁貢献率
UZR	守備貢献率

表 6　能力に関する特徴量 (Netflix)

特徴量	説明
Budget	予算
Director likes	監督の評価
Starring likes	主演の評価
Supporting likes	助演の評価
Staff likes	スタッフの評価

表 7　成績に関する特徴量 (Netflix)

特徴量	説明
IMDB score	映画評論サイトの評価
#Critic score	評論家の評価数
#Movie likes	いいね数
#Reviews	レビュー数
#Review view	レビュー閲覧数



図 7　評価指標の算出方法

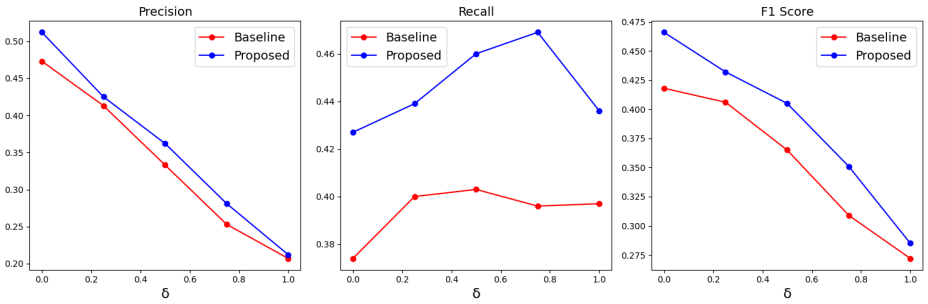


図 8　x-means における WAR が閾値 δ 以上成長する選手の特定性能

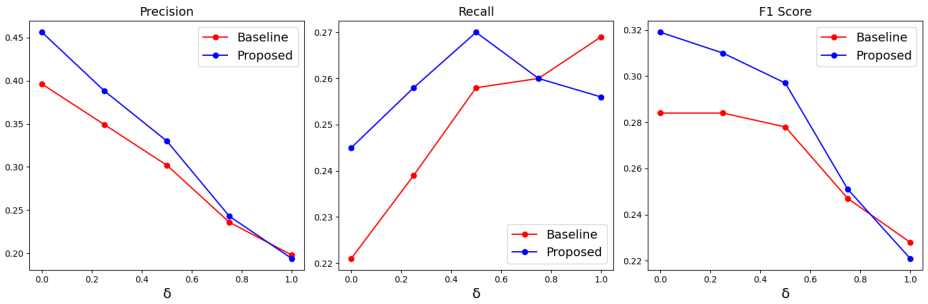


図 9　階層的クラスタリングにおける WAR が閾値 δ 以上成長する選手の特定性能

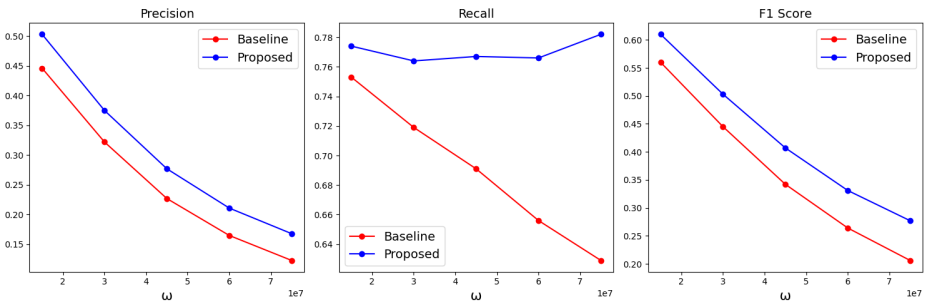


図 10　x-means における興行収入が閾値 ω 以上伸びる映画の特定性能

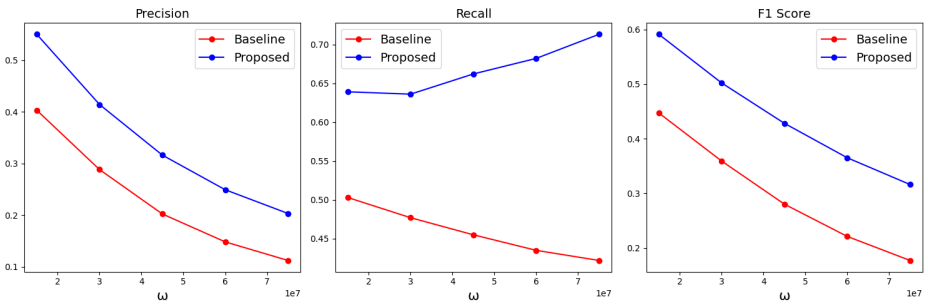


図 11　階層的クラスタリングにおける興行収入が閾値 ω 以上伸びる映画の特定性能

楽天レビューデータを用いたレビュー推薦の一手法

何 毅傑† 上野 史‡ 太田 学‡

†岡山大学大学院環境生命自然科学研究科 〒700-8530 岡山県岡山市北区津島中 3-1-1
‡岡山大学学術研究院環境生命自然科学学域 〒700-8530 岡山県岡山市北区津島中 3-1-1
E-mail: †ph355s9h@s.okayama-u.ac.jp, ‡{uwano, ohta}@okayama-u.ac.jp

あらまし EC サイトで商品を購入する際，ユーザは他のユーザのレビューを参考にして商品を購入するかどうかを決定することが多い．数多くのレビューが存在する場合，その中からユーザにとって有用なレビューを推薦することができれば，ユーザはより効率的な商品選別ができる．本稿では，EC サイトのレビューデータを用いてレビューを推薦する手法を提案する．具体的には，まず EC サイトの各商品ジャンルの基準観点を定める，次に Natural Language API を用いて商品レビュー中のエンティティとジャンルの基準観点をベクトル化し，基準観点に類似したエンティティを抽出する．そして，この基準観点の類似エンティティを利用して，可読性が高く，九つの異なる基準観点へ言及されたレビュー集合を推薦する．被験者実験では，楽天データセットのレビューを使って提案手法と比較手法によりユーザにレビューを推薦し，推薦されたレビューを評価した．実験の結果から，提案手法で推薦されたレビューはユーザが商品を選別するとき有用であることがわかった．
キーワード レビュー推薦，E コマース，自然言語処理，商品選別支援

1. はじめに

ネットショッピングの普及に伴い，多くの人々がECサイトを利用して商品を購入している．商品を選ぶ際他のユーザのレビューを参考にすることは一般的である．レビューは商品の観点について書かれていることが多く，商品全体またはその観点に対する様々な意見が含まれている．レビューに商品に関する複数の観点が含まれている場合，ユーザはそのようなレビューを活用して商品の特徴を使用者の視点から理解し，最終的に商品を購入するかどうかを判断する．

しかし，レビューが多い場合，ユーザにとって有用なレビューを見つけることは難しい.この問題を解決するために,膨大なレビューの中からユーザにとって有用なレビューを推薦する研究が行われている．例えば，喩ら[1]は商品予測モデルと商品ジャンル予測モデルを訓練し，商品と商品ジャンルを予測する．商品を正しく予測する確率が高い文は，その商品の特徴を表していると仮定する．また，商品予測モデルによって抽出された文集合の中に存在する「すぐに届いて役たちました」「いい買い物ができました」「また、お世話になると思います」等のノイズ文を削除するために，商品ジャンル予測モデルを訓練する．２つのモデルの予測結果の確率を組み合わせて代表的なレビュー文をユーザに推薦する手法を提案した．

また，レビューの可読性はユーザのレビュー理解に影響を与える．理解しやすく詳細なレビューはユーザに有益な情報を提供し，より賢明なショッピングの決定を支援する．可読性が高いレビューは通常，使用者の体験や意見を明確に表現しているため，ユーザが商品の利点や欠点を簡単に理解しやすい．

本稿では，楽天市場のECサイト[1]のレビューデータを用いて，可読性が高く，多くの観点について記載されたレビュー集合をユーザに推薦する手法を提案する．

本稿の構成は以下の通りである．2 節で関連研究について，3 節では提案手法の詳細を述べる．4 節では評価実験とその結果を述べる．5 節ではまとめと今後の課題について述べる．

2. 関連研究

2.1 レビュー推薦に関する研究

レビュー推薦に関する研究の目的の多くは，多くのレビューの中から有益なレビューを推薦することである．

高間ら[2]はユーザがレビューを閲覧し，評価した履歴に基づき,ユーザの価値観をモデル化した.そして，ユーザの価値観と一致する商品とレビューをユーザに提示する手法を提案した．

坂本ら[3]はユーザの嗜好に適合したレビューを抽出する手法を提案した．この手法では，レビューを用いてユーザのアイテムに対する評価値を予測する機械学習モデルを訓練させる．そして，評価値を正確に予測できるレビューをユーザの嗜好に適合したレビューとして抽出し，ユーザに推薦する．

Huang ら[4]はレビュー中の観点の感情類似度に基づく個人化されたレビュー推薦手法を提案した．この手法では，異なるユーザのレビューにおける観点の感情

[1]https://www.rakuten.co.jp/

類似度と異なるユーザが購入した商品間の類似度を利用し，ユーザ関連度を算出する．算出されたユーザ関連度とレビュー中の観点の数を考慮して，レビュー有用値を計算する．有用値が高いレビューをユーザに推薦する．

今江ら[5]は化粧品レビューの推薦精度を向上させるために，評価表現辞書を作成する際に使用するレビュー文を商品の効果に関する文に絞ることで評価表現辞書を洗練した．この評価表現辞書は，化粧品の評価に関するキーワードとそれを修飾する特徴や程度，否定の有無で構成される評価表現に評価値を付与したものである．洗練された評価表現辞書を利用して，レビュー推薦システムの推薦精度を向上させた．

2.2 可読性の評価に関する研究

可読性とは読みやすさのことである．可読性が高いレビューは，ユーザに商品に関する情報をよりわかりやすく伝えることができる．

Tateisiら[6]は日本語の可読性スコアを算出する式を提案した．彼らが提案した可読性スコアは 4 つの要素で構成される．それらは，種類別の文字の相対的な出現頻度,同一種類の文字で構成される文字列の最大長,文の長さ，文中の読点の数である．

輪島ら[7]はテキストの可読性を評価するために，テキストから抽出した様々な特徴量の評価手法を提案した．特徴量の評価によってテキストの読みやすさを判定する．テキストの特徴量は，表層情報，話題，語種，意味属性，言語表現，文末表現，品詞，固有表現，評価表現である．彼らの提案手法では,特徴量行列を Non negative Matrix Factorization (NMF)[2]で行列分解し，重要な基底行列を選択する．選択した基底で各特徴量を評価した．

児玉ら[8]は結束性がある二文が多い文章は可読性が高いと考え，文の可読性を向上させるために，結束性がない二文を指摘するシステムを提案した．ここでいう結束性は，前後二文で共通の単語を持つときに意味的につながる性質のことである．例えば，1.リンゴを買った．2.みかんは美味しかった．3.明日もみかんを食べたい．文 1 と文 2 は、共通の単語を持たないため結束性がない．文 2 と文 3 は，共通の単語「みかん」を持つため結束性がある．

3. 提案するレビュー推薦手法

3.1 提案手法の概要

図 1 は提案手法によって，レビューを推薦する例である．図中のエンティティd，e，f，g，h，iは基準観点との類似度が 0.6 以上のエンティティである．提案手法では事前に商品のジャンルの基準観点を決めておく．具体的には，商品ジャンルAを選択し，そして，ChatGPT[3]に，“Aに関する基準観点を九つ生成してください”という指示をだす．ChatGPTで生成された結果を参考にして，ジャンルAの基準観点を決める．レビュー推薦はまず，Natural Language API[4]を用いて商品レビュー中のエンティティを抽出する．またChatGPTから得た基準観点とレビューから抽出したエンティティをfastText[5]で分散表現に変換し，それらのコサイン類似度を算出する．そして，コサイン類似度が 0.6 以上のエンティティを抽出し，それを基準観点に類似したエンティティとする．三つの基準観点に類似した三つのエンティティを含むレビューを抽出し，可読性スコアを算出する．可読性スコアが高く，九つの異なる基準観点へ言及されたレビュー集合をユーザに推薦する．

3.2 エンティティの抽出

提案手法では Natural Language API を用いてレビュー中のエンティティを抽出する．Natural Language API は Google Cloud Platform (GCP)がサポートする API であり,感情分析や構文解析などのタスクに使用される．提案手法では，商品ごとに有用なレビューを推薦するために，レビューデータ中の同じジャンルの商品のレビューを商品ごとに分類する．そして，各商品の各レビューから Natural Language API を用いてエンティティを抽出する．

3.3 基準観点と類似したエンティティの獲得

本研究では，商品レビューから抽出したエンティティのうち基準観点とのコサイン類似度が 0.6 以上のエンティティがあれば，そのレビューはその基準観点の類似観点を持つとする．

そのためエンティティと基準観点は fastText を用いてベクトルに変換する．そして，ベクトルに変換した基準観点とエンティティのコサイン類似度を算出し，コサイン類似度が 0.6 以上のエンティティを獲得する．

[2]https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

[3]https://chat.openai.com/
[4]https://cloud.google.com/natural-language?hl=ja
[5]https://fasttext.cc/



図 1　レビュー推薦手法の概要

fastText は，Facebook AI Research (FAIR)[6]よって開発されたオープンソースの高性能なテキスト分類や単語ベクトルの学習ツールである．また，本稿で利用した fastText は Common Crawl[7]と Wikipedia[8]で事前学習済みのモデルである．

3.4 可読性スコアの算出

可読性スコアは，Hasebe ら[9]が提案した可読性の計算式である式(1)によって算出する．Hasebe らは日本語形態素解析エンジンである MeCab[9]とそれに使用される辞書である UniDic[10]を用いてテキストを形態素解して形態素を得た．形態素解析では，名詞，動詞，形容詞，副詞，助詞，助動詞，接辞などの品詞に分類されるだけでなく，漢語であるか和語であるかも分類される．テキスト中の動詞，助詞，漢語と和語の数を利用して数式(1)の可読性スコアを算出した．

式(3)はレビュー中の漢語の割合，式(4)は和語の割合，式(5)は動詞の割合，式(6)は助詞の割合である．

$$Score = 11.724 - 0.056 \times a - 0.126 \times b - 0.042 \times c - 0.145 \times d - 0.044 \times e \quad (1)$$

$$a = \text{文の平均長} \quad (2)$$

$$b = \frac{\text{漢語数}}{\text{総形態素数}} \times 100 \quad (3)$$

$$c = \frac{\text{和語数}}{\text{総形態素数}} \times 100 \quad (4)$$

$$d = \frac{\text{動詞数}}{\text{総形態素数}} \times 100 \quad (5)$$

$$e = \frac{\text{助詞数}}{\text{総形態素数}} \times 100 \quad (6)$$

4. 評価実験

評価実験では，提案手法で推薦されるレビューと二つの比較手法で推薦されるレビューを被験者に提示し，どちらが被験者にとって有用であるかを検証する．

4.1 データセット

楽天データセットの「みんなのレビュー・ロコミ情

[6]https://ai.meta.com/research/
[7]https://commoncrawl.org/
[8]https://www.wikipedia.org/
[9]https://taku910.github.io/mecab/
[10]https://clrd.ninjal.ac.jp/unidic/

報」の中の 2019 年の「商品レビュー」にある商品名，商品 ID を使って同じ商品のレビューを抽出する．

4.2 レビューの前処理

抽出したレビューの中から内容が他と重複しているレビューと英語のレビューを削除する．また，レビューの中に「良かった!!」などの商品の特徴が含まれていないレビューも削除する．

4.3 比較手法

本実験では提案手法と以下の二つのレビュー推薦手法を比較する．

参考数に基づいた推薦手法：楽天データセットの「みんなのレビュー・ロコミ情報」にある参考になった数が多い 3 件のレビューを抽出し，ユーザに推薦する．

tfidf 法[1]: レビュー文の tfidf (term frequency- inverse document frequency)値を計算し，文の tfidf 値が上位 3 件のレビューをユーザに推薦する．ここで tf-idf は，単語の重要度を測る手法である．tf (term frequency, 単語の出現頻度)は各文書においてその単語がどのくらい出現したかを表す．idf (inverse document frequency, 逆文書頻度)はある単語が文書集合全体でどれだけ一般的に使用されているかを示す文書頻度(df)の逆数を利用して計算される．単語が多くの文書に出現していると，idf の値は小さくなる．

式(7)はレビューrにおける単語 x の tf 値の算出式である．$L(r)$はレビューの単語数を表す．式(8)はレビュー集合 R における単語 x の idf 値の算出式である．

式(9)はレビューr の tfidf 値を示す．$w(r)$はレビューrに含まれる単語集合を表す．tfidf 値が大きいレビューを抽出し，ユーザに推薦する．

$$\mathrm{tf}(x,r) = \frac{\text{単語}x\text{のレビュー}r\text{における出現頻度}}{L(r)} \tag{7}$$

$$\mathrm{idf}(x,R) = \log\frac{\text{レビュー集合}R\text{の総レビュー数}}{\text{単語}x\text{が含まれるレビューの数}} \tag{8}$$

$$\mathrm{tfidf}(r) = \frac{\sum_{x\in w(r)} \mathrm{tf}(x,r) * \mathrm{idf}(x,R)}{L(r)} \tag{9}$$

4.4 実験方法

提案手法のため楽天データセット内の商品ジャンル「イヤホン」「加湿器」およびスマートフォンの「ケース・カバー」の基準観点を ChatGPT で生成する．ジャンル「イヤホン」において生成した基準観点は「音質，耐久性，デザイン，価格，電池，快適さ，サイズ，フィット，接続」．ジャンル「加湿器」において生成した基準観点は「操作，容量，湿度，価格，品質，見た目，音，サイズ，性能」．ジャンル「ケース・カバー」において生成した基準観点は「耐久性，デザイン，フィット，値段，使用感，取付け，機種，品質，色」である．次に，各ジャンルの九つの基準観点をランダムに三つずつ，三つのグループに分類する．ジャンル「イヤホン」において，グループ 1 は「音質，耐久性，デザイン」，グループ 2 は「価格，電池，快適さ」，グループ 3 は「フィット，サイズ，接続」である．ジャンル「加湿器」において，グループ 1 は「湿度，操作，容量」，グループ 2 は「価格，品質，見た目」，グループ 3 は「音，サイズ，性能」である．ジャンル「ケース・カバー」において，グループ 1 は「耐久性，デザイン，フィット」，グループ 2 は「色，品質，機種」，グループ 3 は「使用感，値段，取付け」である．また，Natural Language API を利用してジャンル「イヤホン」の商品名が QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代のレビュー(2009 件),「加湿器」の商品名が小型加湿器 S08 のレビュー(1185 件),「ケース・カバー」の商品名が iFace 強化ガラスクリアケースのレビュー(2529 件) から，エンティティを抽出する．各商品のレビューのエンティティと各商品の各グループの基準観点との類似度を算出し，各グループ内の三つの基準観点と類似した三つのエンティティを持つレビューを抽出し，可読性スコアを算出する．そして可読性スコアが一番高く，九つの異なる基準観点へ言及されたレビュー集合をユーザに推薦する．

提案手法によって推薦された 3 件のレビューと，比較手法によって推薦された上位 3 件のレビューについてアンケートを用いて評価する．評価は，以下四つの質問に対する被験者の回答によって判断する．被験者は岡山大学の留学生と学部四年生，大学院生の合計 12 名である．実験の公平性を保証するため，レビューを推薦するときに，提案手法または比較手法で推薦されたレビューであることを明示せず，代わりに手法 A，手法 B，手法 C を示す．

質問 1:どの方法で推薦されたレビューが最も多くの商品特徴を含んでいるか?

質問 2:どの方法で推薦されたレビューがあなたの購買決定に最も影響を与えるか?

質問 3:どの方法で推薦されたレビューが一番読みやすいか?

質問 4:どの方法で推薦されたレビューが実際の使用場面を想像させるか?

4.5 実験結果

図 2 から，商品 iFace 強化ガラスクリアケース，小

型加湿器 S08 において，提案手法で推薦されたレビューが比較手法で推薦されたレビューより多くの商品特徴を持っていることがわかった．一方，商品 QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代において，参考数に基づく比較手法で推薦されたレビューが提案手法で推薦されたレビューより多くの商品特徴を持っていることがわかった．

図 3 から，商品 iFace 強化ガラスクリアケースにおいて，提案手法で推薦されたレビューが比較手法よりも購買決定への影響度が高いという評価を受けた．しかし，小型加湿器 S08 と QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代において，参考数に基づく比較手法で推薦されたレビューがいい結果を得た．

図 4 から，商品 iFace 強化ガラスクリアケースにおいて，提案手法で推薦されたレビューが比較手法で推薦されたレビューよりも読みやすいという結果を得た．小型加湿器 S08 においては，提案手法と参考数に基づく比較手法で推薦されたレビューの読みやすさについて同等の評価を得たが， tf-idf 法がより優れた結果となった．また，商品 QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代では，提案手法と比較手法 tf-idf 法で推薦されたレビューの読みやすさが同等の評価を得た．

図 5 から，商品 iFace 強化ガラスクリアケース，小型加湿器 S08 において，提案手法で推薦されたレビューが比較手法で推薦されたレビューよりも，実際の使用画面の想像度について優れた評価を受けた．商品 QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代においては，提案手法と参考数に基づく比較手法で推薦されたレビューが同等の評価を受けた．

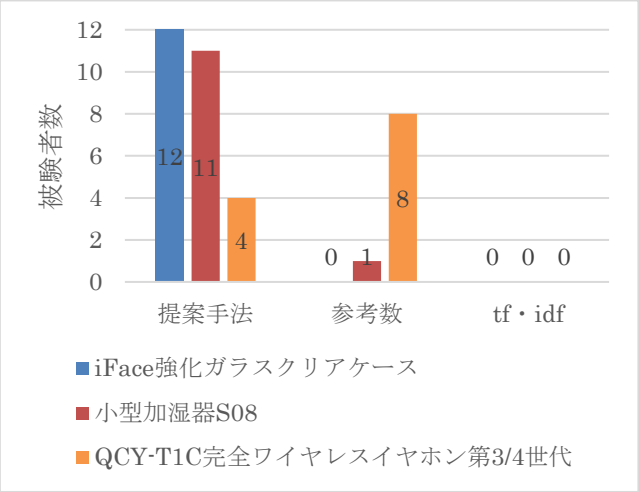


図 2　質問 1(特徴の多さ)に対する各手法で推薦された商品レビューの支持数

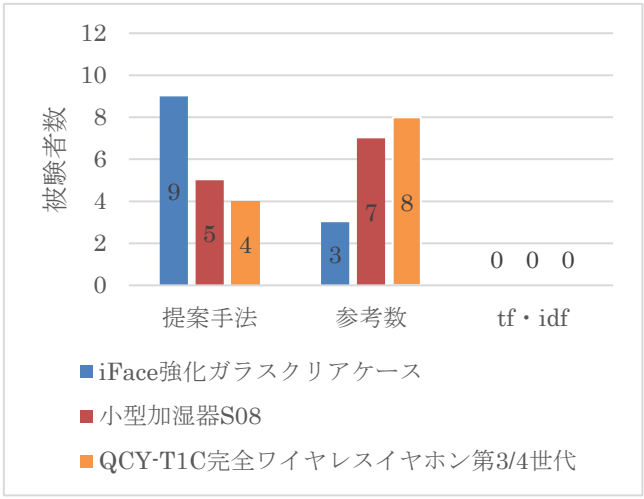


図 3　質問 2(購買決定への影響度)に対する各手法で推薦された商品レビューの支持数

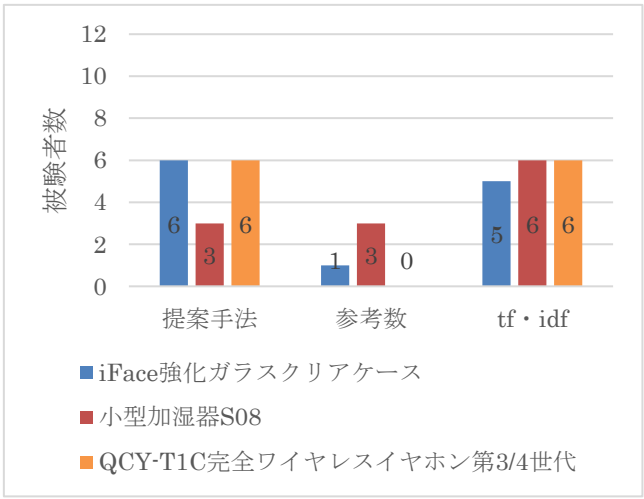


図 4　質問 3(読みやすさ)に対する各手法で推薦された商品レビューの支持数

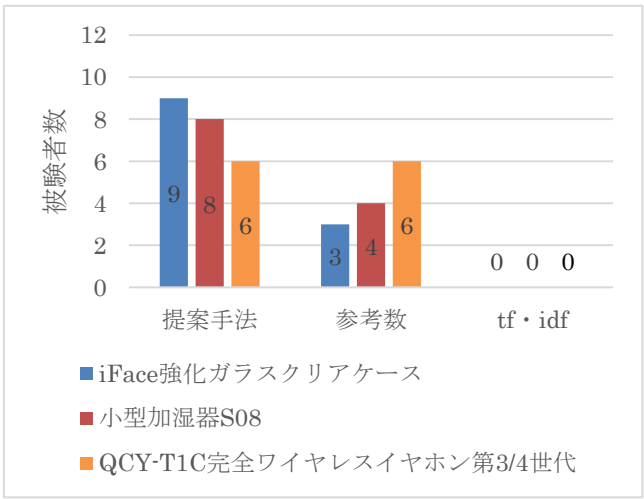


図 5　質問 4(実際の使用画面の想像度)に対する各手法で推薦された商品レビューの支持数

4.6 考察

図 2，図 3，図 4，図 5 の結果から，商品 iFace 強化ガラスクリアケースでは，全質問について提案手法が比較手法と比べて高い評価を受けた．しかし，商品小型加湿器 S08 において，質問 1(特徴の多さ)と質問 4(実際の使用画面の想像度)では提案手法が比較手法より高い評価を受けたが，質問 2(購買決定への影響度)と質問 3(読みやすさ)ではそうはならなかった．また，商品 QCY-T1C 完全ワイヤレスイヤホン第 3/4 世代において，質問 1 と質問 2 では比較手法が提案手法より高い評価を受けた．質問 3 と質問 4 では提案手法と比較手法が同等の評価を受けた．したがって，一部の商品では，提案手法で推薦されたレビューが比較手法で推薦されたレビューより有用であるといえる．

5. まとめ

本稿では楽天レビューデータを用いてレビューを推薦する手法を提案した．提案手法は，レビュー中の異なる基準観点と類似したエンティティとレビューの可読性スコアを算出し，可読性が高く，九つの基準観点を網羅するように三つのレビューをユーザに推薦する．被験者実験の結果から提案手法で推薦されたレビューはユーザが商品を選別するとき有用であることがわかった．

今後は，レビュー推薦システムのインターフェースを作る予定である．また，商品の買い手であるユーザではなく商品の売り手である会社に有用なレビューを推薦する方法も検討している．

謝　　辞

本稿には，国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供された「楽天データセット」(https://rit.rakuten.com/data_release/)を利用した．ここに記して感謝する．

文　　献

[1] 喩思セイ，牛尼剛聡，“商品予測タスクを利用した商品選別支援のための有用なレビュー文の抽出,” DEIM2022, G24-4, 2022.

[2] 高間康史，清水涼人，“レビュー閲覧履歴からの価値観モデリングを用いた情報推薦システム,” 人工知能学会第 13 回インタラクティブ情報アクセスと可視化マイニング研究, SIG-AM-13-02, 2016.

[3] 坂本新真，牛尼剛聡，“評価値予測タスクを用いたユーザの嗜好に適合したレビューの抽出,” DEIM2023, 3B-8-4, 2023.

[4] Chunli Huang, Wenjun Jiang, Jie Wu, Guojun Wang, “Personalized Review Recommendation based on users’ Aspect Sentiment,” ACM Transaction on Internet Technology, Vol. 20, No. 4, pp. 1-26, 2020.

[5] 今江柚実，松下光範，“評価表現辞書の精錬による化粧品レビューの推薦精度向上,” 人工知能学会第 24 回インタラクティブ情報アクセスと可視化マイニング研究会, SIG-AM-24-06, 2020.

[6] Yuka Tateisi, Yoshihiko Ono, Hisao Yamada, “A computer readability formula of Japanese texts for machine scoring,” Proceedings of the 12th International Conference on Computational Linguistics, Volume 2, 1988.

[7] 輪島幸治，木暮啓，古川利博，佐藤 哲司，“可読性に基づいた日本語テキスト情報の特徴量評価,” DEIM2018, C2-1, 2018.

[8] 児玉哲哉，田嶋克向，澤野弘明，“可読性向上のための結束性指摘システムの提案,” 第 82 回全国大会講演論文集, pp. 431-432, 2020.

[9] Yoichiro Hasebe, Jae-Ho Lee, "Introducing a readability evaluation system for Japanese language education." Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese, pp. 19-22, 2015.

スポットの雰囲気分析によるユーザの状況と気分の選択に合わせた楽曲推薦手法の検証

丸岡　史奈† 　栗　達†† 　熊本　忠彦†††
小野 晋太郎†† 　河合 由起子†

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†† 福岡大学　〒 814-0180 福岡県福岡市城南区七隈 8-19-1
††† 千葉工業大学　〒 275-0016 千葉県習志野市津田沼 2-17-1
E-mail: †{g2054281, kawai}@cc.kyoto-su.ac.jp, ††{lida, onoshin}@fukuoka-u.ac.jp,
†††kumamoto@net.it-chiba.ac.jp

あらまし　本研究では，スポットの雰囲気を深層学習を用いたジオタグツイートの解析により取得するとともに，ユーザの状況を自分がいるスポット（場所）での行動と捉え，クラウドソーシングよりユーザの状況（スポットの雰囲気におけるそこでの × 行動）とユーザの気分との関係を明らかにし，スポットの状況（雰囲気および行動）と気分を考慮した楽曲推薦手法を提案する．具体的には，楽曲データは spotify が提供する楽曲の評価値である「energy」（テンション）の高／低と「valence」（気分）の明／暗に分類する．一方，スポットの雰囲気はスポット周辺のジオタグツイートを LSTM より推定する．スポットの雰囲気推定では，楽曲の energy と valence となる「テンション」の高／低と「気分」の明／暗に分類する．さらに，状況に対しては行動ごとに「テンション」の高／低と「気分」の明／暗の楽曲を好むかをクラウドソーシングよりアノテーション付けを行なった．これら楽曲，雰囲気，状況の energy と valence から，ユーザが選択した状況とスポットの雰囲気に合わせた楽曲を推薦する．本稿では，カフェ・公園・デパートという 3 つのスポットを対象に，それぞれのスポットの雰囲気とユーザの状況（例えば「カフェでの一人作業」など）および気分に基づいた楽曲推薦を行い，その有効性を検証する．

キーワード　楽曲推薦，楽曲的特徴，地理情報，ソーシャルメディア

1　はじめに

Spotify や Apple Music のような音楽サブスクリプションサービスが国内外で広まり，リラックス時や食事の際だけでなく，移動中やショッピングといった活動中など，いつでもどこでも気軽に音楽鑑賞を楽しめるようになり，状況に応じた楽曲推薦の研究が活発化している．例えば，Moscato ら [1] は，気分によって聴きたい音楽が異なることに着目し，ユーザの性格的特徴や感情に基づく楽曲推薦を提案している．また，東海ら [2] は，ユーザの自宅住所と現在地より行動推定を行い，ユーザの行動に基づいた楽曲推薦を検討している．韓ら [3] は，Spotify が提供する楽曲の評価値を対象として印象語と感情空間，音楽の関係について検討している．このようにユーザの状況を考慮した楽曲推薦に関する研究や楽曲に含まれる楽曲特徴量を用いた研究が既に行われているが，ユーザの状況としては在宅・移動・外出という分類のみであり，ユーザがどういったスポットにいるかやそのスポットの雰囲気は考慮されていない．また，ユーザがいるスポットの雰囲気（ポジティブかネガティブか）やユーザの状況（そのスポットで何をしているか）・気分（気分が明るいか暗いか，テンションが高いか低いか）によっても聞きたい音楽は変わってくるが，それら状況や気分は十分に考慮されていない．

そこで本稿では，Spotify が提供する楽曲の激しさと明るさに関する特徴量（energy と valence）を用いるとともに，スポットごとの雰囲気を，作成した学習モデルを用いてジオタグツイートより抽出，また被験者アンケートよりユーザの状況と気分について分析し，スポットの雰囲気とユーザの状況・気分を考慮した楽曲推薦システムを提案する．なお，本稿では，カフェ・公園・デパートという 3 つのスポットを対象とし，提案手法よりスポットの雰囲気とユーザの状況と気分に基づいた楽曲推薦手法を検証する．

以下，2 章では楽曲推薦および楽曲と感情との関係についての関連研究を紹介し，本研究と比較する．3 章ではスポットの雰囲気とユーザの状況と気分に基づく楽曲推薦手法について説明する．4 章では，楽曲推薦手法の評価を行い，最後に 5 章で本研究で提案した楽曲推薦システムのまとめと今後の課題について述べる．

2　関連研究

楽曲推薦の分野では，ユーザの性格的特徴や感情に基づく楽曲推薦に関する研究 [1] や，推薦理由の提示による推薦システムへの信頼や利用の持続性に関する研究 [4]，深層学習を用いた楽曲推薦の研究 [5] などが行われている．本節では，関連研究としてユーザ行動に基づいた楽曲推薦システムと楽曲に対する

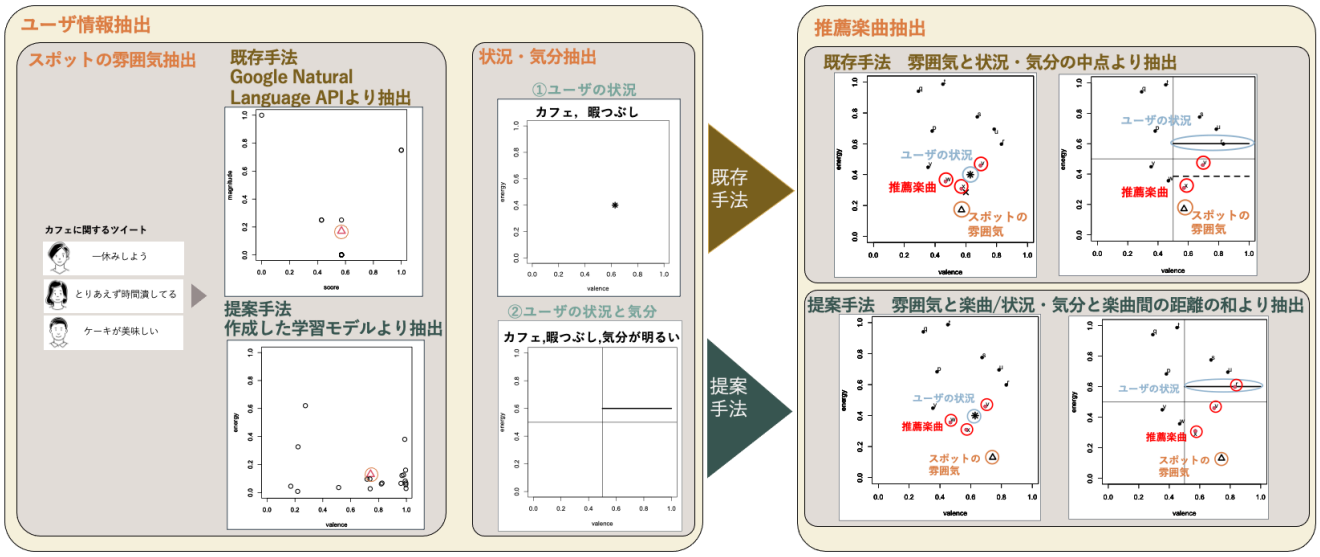


図 1　現在地周辺情報とユーザ状況と気分による楽曲推薦システム (サーバ)



図 2　現在地周辺情報とユーザ状況と気分による楽曲推薦システム (アプリケーション)

印象分析に関する研究を紹介する．

2.1　ユーザの行動に基づいた楽曲推薦システムに関する研究

ユーザの状況を考慮した楽曲推薦に関する研究として，東海ら [2] は，位置情報や楽曲再生記録等のライフログを用いてユーザの行動に基づく楽曲推薦システムを提案している．ユーザの現在地や時間帯から行動推定を行うとともに，行動と聴きたい楽曲に関連がある点に着目し，在宅・移動・外出の行動ごとのプレイリストから生成した行動別楽曲情報ベクトル (歌手，作詞家，作曲家) と行動別歌詞ベクトル (LDA を用いて得た歌詞のトピック) と端末内にある楽曲のベクトルとの類似度計算から楽曲を推薦する．

これら既存研究により、ユーザ状況ごとに楽曲を推薦できるが、ユーザの状況推定が在宅・移動・外出と 3 種類のみでしか分類されていない．そのため，外出先によって異なるユーザ状況に対応した楽曲推薦は難しい．

2.2　ユーザの印象に基づく楽曲推薦に関する研究

ユーザの印象に基づく楽曲推薦に関する研究として，韓ら [3] は，Spotify が提供する楽曲の評価値を対象として印象語と感情空間，音楽の関係について検討している．Russell ら [6] [7] が提案した Arousal-Valence(AV) 空間を踏まえ，「Arousal」と「Valence」に Spotify が付与した特徴量のうちの「energy」と「valence」をそれぞれ対応付けて楽曲を感情空間上にマッピングしている．

以上の関連研究では，ユーザの状況を 3 種類に分類し、楽曲推薦する手法や楽曲を感情空間上にマッピングする研究が行われている．本研究では，ユーザの現在地（スポット）の雰囲気を SNS データから抽出し，また具体的な状況も取得し，気分に合った楽曲を推薦する．

これまで我々は，ユーザの状況を自分がいる場所（スポット）とそこで自分がしている行動と捉え，スポットの雰囲気をジオタグツイートの解析により取得するとともに，アンケート調査に基づいてユーザの状況（スポット × 行動）と気分との関係を分析することで，ユーザの状況（スポットの雰囲気とそこでの行動）と気分を考慮した楽曲推薦の抽出に関する研究を行なってきた．具体的には，ユーザの気分を「気分が明るい／暗い」と「テンションが高い／低い」の 2 軸で分類し，それぞれの気分ごとに明るさと激しさの 2 次元にマッピングした．さらに，スポット周辺のジオタグツイートの感情分析を行い，スポットの雰囲気として「ネガティブ／ポジティブ」(明るさ) と，感情の強さの 2 次元に置き換えた．一方，楽曲データは spotify が提供する楽曲の評価値である「energy」と「valence」の 2 次元にマッピングした．提案手法は，明るさと激しさの値を付与した状況データと，明るさと感情の強さを付与したスポットデータの中点から、楽曲データとのユークリッド距離よりユーザの状況と気分にあった楽曲を推薦した．ただし，スポットデータが，状況データと楽曲データに付与されている「energy」と「valence」とは異なる 2 軸 (「ネガティブ／ポジティブ」(明るさ) と感情の強さ) を利用していることから，ユーザ評価結果が良くない部分もあった．

そこで本稿では，楽曲・状況データはこれまで収集したデータを利用し，スポットデータについては，作成した学習器を用いてジオタグツイートよりツイートそれぞれに「energy」と

「valence」を付与する．このように楽曲・スポット・状況データで 2 軸を統一する．また推薦楽曲を楽曲・状況データ間のユークリッド距離と，楽曲・スポットデータ間のユークリッド距離の和より抽出し，スポットの雰囲気とユーザの状況・気分を考慮した楽曲推薦システムを提案する．

3　周辺状況とユーザ状況に基づく楽曲推薦

本節では，周辺情報としてスポットの雰囲気とユーザ状況に基づく楽曲推薦システムの概要およびデータ取得について述べる．

3.1　楽曲推薦手法の概要

図 1 に楽曲推薦システムのサーバ側の概要図を示す．ユーザ情報（現在地）から考えられるスポットのジオタグツイートを取得し，そのスポットの雰囲気を抽出する．ユーザは自身の状況と気分を提示された中から選択する．これよりスポットの雰囲気とユーザの状況・気分に合わせた楽曲が推薦される．例えば，ユーザの状況が「カフェで一人作業」である場合，推薦楽曲は図 1 の「状況のみ」の画面より楽曲 v，w，x となる．一方，ユーザの状況と気分が「カフェで一人作業で，気分が明るい」の場合，推薦楽曲は図 1 の「状況と気分」の画面より楽曲 r，v，x となる．このようにユーザの気分によって推薦する楽曲も異なる．なお，韓ら [3] の研究より，楽曲データには Arousal-Valence(AV) 空間を踏まえ，Spotify[1]で配信されている楽曲の特徴量の「valence」と「energy」を用いる．図 2 に楽曲推薦システムのユーザ側の概要図を示す．ユーザには現在の状況と気分を選択肢から選んでもらい、その状況と気分と、現在地情報から抽出したスポットの雰囲気によって、推薦楽曲を提示する．なお，韓ら [3] の研究より，楽曲データには Arousal-Valence(AV) 空間を踏まえ，Spotify で配信されている楽曲の特徴量の「valence」と「energy」を用いる．

3.2　生成 AI と OSM によるツイートを用いたスポットの雰囲気抽出

本研究では,「楽曲を聴くのはどういう状況の時であるか」を生成 AI を用いて抽出する．生成 AI に「どういう状況で楽曲を聴くのか」と「その状況は一般的にどのような場所が考えられるか」を問い合わせ，状況と，その状況が起こり得る場所名を取得する．次に，場所名から OSM（OpenStreetMap）の施設カテゴリを対応付けし，施設カテゴリごとに各スポットの緯度経度を取得する．最後に，緯度経度から施設カテゴリごとにツイートの取得範囲を設定し，スポットに対するツイートを取得する．スポットの雰囲気の推定については，自然言語処理を用いて行う．図 3 に学習モデルの作成から予測値の出力までの流れを示す．まず前処理として，学習データとなるツイート 411 件をそれぞれベクトル化する．具体的には，被験者 7 人に，任意のツイート文 409 件それぞれに valence と energy の値を付与してもらう．7 人の平均値をそのツイート文の valence と energy の値とし，全ツイート文の valence・energy のそれぞれの平均値を閾値としてツイート文それぞれを 0・1 に振り分ける．これを学習データとする．学習データのツイート文に含まれる全ての言葉を Word2VEC で 300 次元のベクトルに変換する．それらベクトルの平均をツイート 1 つのベクトルとする．そして，アルゴリズムとして LSTM を用いて学習データより学習モデルを作成する．その後，学習モデルより，各スポットのツイート文ごとに valence と energy の値を予測する．スポットの雰囲気は，valence と energy の値が予測されたツイート文の重心とする．

1：https://open.spotify.com/

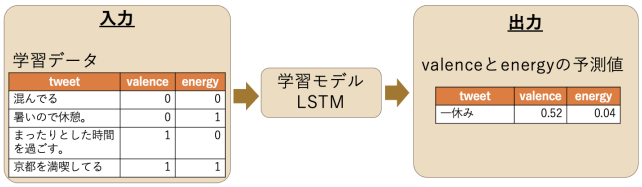


図 3　スポットの雰囲気抽出の流れ

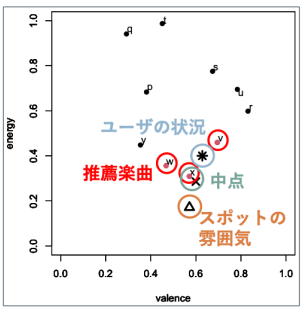


図 4　既存手法 (状況のみ)

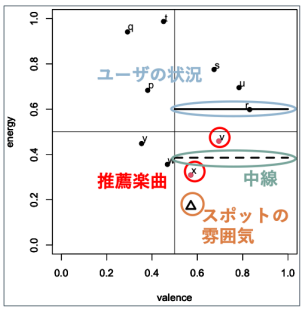


図 5　既存手法 (状況と気分)

3.3　状況・気分に対する感情分析

生成 AI への問い合わせに基づきそれぞれのスポットごとで考えられる状況を提示する．状況ごとに,「valence」と「energy」の値を付与した楽曲データに合わせて明るさ (ネガティブ／ポジティブ) と激しさ (落ち着いた／激しい) の 2 次元にマッピングする．また気分についても，楽曲データに合わせて「気分が明るい／暗い」と「テンションが高い／低い」の 4 つの気分を定義し，気分・状況を明るさと激しさの 2 次元にマッピングする．本研究では，アンケート調査により以上の感情分析を行う．

3.4　楽曲特性に基づく雰囲気と状況分析

本研究では，スポットの雰囲気に加えユーザの状況のみを考慮する場合とユーザの状況と気分を考慮する場合の 2 パターンがある．楽曲データと同様に，ユーザの状況やスポットの雰囲気，ユーザの気分を「valence」と「energy」の 2 次元座標に分布させる．初めに，既存手法のユーザの状況のみを考慮する場合について，楽曲データ，状況データ，スポットデータの散布図を示す（図 4）．図 4 において黒点「•(p〜y)」は楽曲，三角「△」はスポットの雰囲気，アスタリスク「∗」はユーザ状況に対応している．次に，ユーザの状況（図 4 中「∗」）とスポットの雰囲気（図 4 中「△」）の 2 点の中点（図 4 中「×」）をとる．

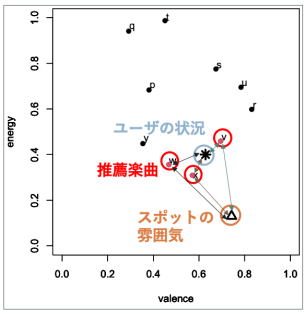


図 6　提案手法 (状況のみ)

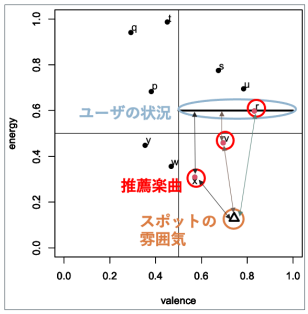


図 7　提案手法 (状況と気分)

表 1　各スポットの OSM での施設名とツイート総数（京都市内）

OSM の施設カテゴリ	ツイート取得範囲	ツイート総数
カフェ	半径 10m 圏内	8,742
コワーキングスペース	半径 15m 圏内（図書館） 半径 120m 圏内（大学，専門学校）	6,549
公園	半径 15m 圏内	67
ジム，スポーツ施設	半径 10m 圏内（ジム） 半径 15m 圏内（スポーツ施設）	156
デパート	半径 100m 圏内	9,872

この中点がユーザの状況とスポットの雰囲気を考慮した点となる．この中点に距離が近い楽曲データを推薦楽曲とする．

次に，既存手法のユーザの状況と気分を考慮する場合について図 5 に楽曲データ，状況データ，スポットデータの散布図を示す．気分については，「気分が暗い（valence: 0.0~0.5）」と「気分が明るい（valence: 0.5~1.0）」および「テンションが低い（energy: 0.0~0.5）」と「テンションが高い（energy: 0.5~1.0）」の 4 つに分類する．図 5 の直線は，気分が明るい時の状況の過激さ（energy）の値を示している．その状況の energy 値とスポットの雰囲気の energy 値（図 5 中「△」）の平均値をユーザの状況・気分とスポットの雰囲気を考慮した値とする（図 5 中点線）．この点線に距離が近い楽曲データを推薦楽曲とする．

続いて，提案手法のユーザの状況のみを考慮する場合について，楽曲データ，状況データ，スポットデータの散布図を示す（図 4）．手法 2 では，各楽曲データとスポットの雰囲気（図 4 中「△」）との距離，各楽曲データとユーザの状況（図 4 中「∗」）との距離をそれぞれ求める．求めた距離の和が最も小さい楽曲を推薦楽曲とする．

次に，提案手法のユーザの状況と気分を考慮する場合について，楽曲データ，状況データ，スポットデータの散布図を示す（図 7）．手法 2 では，各楽曲データとスポットの雰囲気（図 7 中「△」）との距離，各楽曲データとユーザの状況と気分（図 7 中直線）との距離をそれぞれ求める．求めた距離の和が最も小さい楽曲を推薦楽曲とする．どの場合も距離関数にはユークリッド距離を採用する．すなわち，2 つの点 $P(x_1, y_1)$ と点 $Q(x_2, y_2)$ のユークリッド距離 d は次式により算出する．

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

上記のユークリッド距離 d の値が小さいほど，推薦楽曲がスポットの雰囲気とユーザの状況・気分に適していることになる．

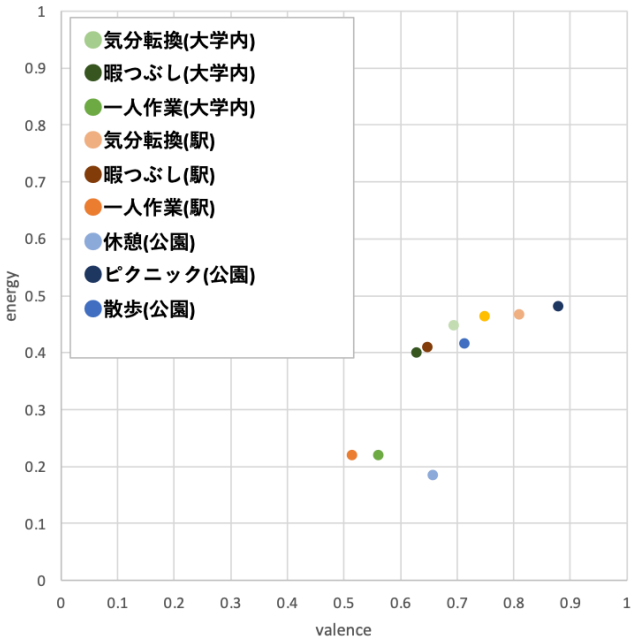


図 8　場所と状況ごとの「valence」と「energy」

表 2　スポットの雰囲気とユーザの状況を考慮した場合の推薦楽曲

場所	状況	推薦楽曲 (既存手法)	推薦楽曲 (提案手法)	MSE(既存手法)	MSE(提案手法)
職場や大学内の	気分転換	楽曲 A	楽曲 A	0.163	**0.057**
カフェ	暇つぶし	楽曲 B	楽曲 A	0.312	**0.071**
(ゆんげ)	一人作業	楽曲 B	楽曲 B	0.067	**0.088**
駅構内の	気分転換	楽曲 D	楽曲 D	0.127	**0.048**
カフェ	暇つぶし	楽曲 D	楽曲 D	0.065	**0.023**
(ベローチェ)	一人作業	楽曲 F	楽曲 B	0.068	**0.061**
公園	休憩	楽曲 B	楽曲 B	0.228	**0.078**
(宝ヶ池公園)	ピクニック	楽曲 C	楽曲 E	0.206	**0.111**
	散歩	楽曲 I	楽曲 D	0.113	**0.069**
デパート	ショッピング	楽曲 B	楽曲 W	0.202	**0.079**
(opa)					

楽曲 A: 僕らまた (SG)，楽曲 B: 魔法の絨毯 (川崎鷹也)，楽曲 D:napori(Vaundy)，楽曲 F: 地球儀 (米津玄師)，楽曲 I:Gold-また逢う日まで-(宇多田ヒカル)，楽曲 W: 喜劇 (星野源)

4　気分と状況に合わせた楽曲推薦の検証

本稿では，カフェ，公園，デパートを対象に，それぞれのスポットでの状況とユーザの気分に基づいた楽曲推薦手法を検証する．本章では，提案手法により抽出された推薦楽曲がツイートデータより分析したスポットの雰囲気およびアンケート調査により分析したユーザの状況・気分に合っているかを検証する．

4.1　スポットと状況の分類

提案手法はツイートから雰囲気を抽出するため，スポットごとにツイートを分類する．表 1 に 3.2 より取得したスポットごとの京都市内を対象としたツイート総数を示す．京都市内のツイート総数は 351,632 件であった．本稿では，カフェ，公園，デパートの 3 つのスポットに絞り検証する．なお，カフェは様々な場所にあるが，場所によって雰囲気が異なるため，聴きたい音楽の特徴も変わってくる可能性がある．このことを考慮し，カフェを職場や大学内にあるカフェと駅構内にあるカフェの 2 つに分けて検証する．それぞれの場所と状況に対して，推薦楽曲を抽出する．

表 3　スポットの雰囲気とユーザの状況・気分を考慮した推薦楽曲 (気分が明るい・暗い)

		気分が明るい				気分が暗い			
場所	状況	推薦楽曲 (既存手法)	推薦楽曲 (提案手法)	MSE(既存手法)	MSE(提案手法)	推薦楽曲 (既存手法)	推薦楽曲 (提案手法)	MSE(既存手法)	MSE(提案手法)
職場や大学内の	気分転換	楽曲 B	楽曲 Z	0.398	**0.049**	楽曲 D	楽曲 N	0.269	**0.073**
カフェ	暇つぶし	楽曲 B	楽曲 Z	0.312	**0.059**	楽曲 F	楽曲 D	0.157	**0.079**
(ゆんげ)	一人作業	楽曲 B	楽曲 A	0.321	**0.071**	楽曲 F	楽曲 D	0.119	**0.065**
駅構内の	気分転換	楽曲 B	楽曲 X	0.304	**0.077**	楽曲 H	楽曲 U	0.184	**0.039**
カフェ	暇つぶし	楽曲 E	楽曲 Y	0.171	**0.108**	楽曲 F	楽曲 D	0.136	**0.061**
(ベローチェ)	一人作業	楽曲 E	楽曲 B	0.236	**0.168**	楽曲 F	楽曲 F	0.089	**0.076**
公園	休憩	楽曲 B	楽曲 B	0.325	**0.176**	楽曲 H	楽曲 F	0.166	**0.053**
(宝ヶ池公園)	ピクニック	楽曲 G	楽曲 X	0.419	**0.103**	楽曲 J	楽曲 J	0.181	**0.115**
	散歩	楽曲 G	楽曲 X	0.305	**0.083**	楽曲 J	楽曲 J	0.209	**0.105**
デパート (opa)	ショッピング	楽曲 B	楽曲 Y	0.380	**0.270**	楽曲 F	楽曲 J	**0.209**	0.324

表 4　スポットの雰囲気とユーザの状況・気分を考慮した推薦楽曲 (テンションが高い・低い)

		テンションが高い				テンションが低い			
場所	状況	推薦楽曲 (既存手法)	推薦楽曲 (提案手法)	MSE(既存手法)	MSE(提案手法)	推薦楽曲 (既存手法)	推薦楽曲 (提案手法)	MSE(既存手法)	MSE(提案手法)
職場や大学内の	気分転換	楽曲 K	楽曲 AB	0.338	**0.068**	楽曲 B	楽曲 B	0.188	**0.087**
カフェ	暇つぶし	楽曲 L	楽曲 AB	0.258	**0.087**	楽曲 D	楽曲 B	0.243	**0.061**
(ゆんげ)	一人作業	楽曲 M	楽曲 I	0.168	**0.119**	楽曲 D	楽曲 B	0.196	**0.056**
駅構内の	気分転換	楽曲 N	楽曲 AA	0.332	**0.077**	楽曲 U	楽曲 B	0.399	**0.112**
カフェ	暇つぶし	楽曲 O	楽曲 AA	0.126	**0.087**	楽曲 J	楽曲 D	0.197	**0.062**
(ベローチェ)	一人作業	楽曲 P	楽曲 X	**0.075**	0.101	楽曲 V	楽曲 J	0.176	**0.092**
公園	休憩	楽曲 R	楽曲 AA	0.408	**0.079**	楽曲 D	楽曲 D	0.330	**0.105**
(宝ヶ池公園)	ピクニック	楽曲 Q	楽曲 AB	0.242	**0.075**	楽曲 B	楽曲 D	0.245	**0.027**
	散歩	楽曲 S	楽曲 AB	0.188	**0.067**	楽曲 D	楽曲 D	0.328	**0.047**
デパート (opa)	ショッピング	楽曲 T	楽曲 AA	0.369	**0.087**	楽曲 D	楽曲 U	0.260	**0.056**

楽曲 B: 魔法の絨毯 (川崎鷹也)，楽曲 D:napori(Vaundy)，楽曲 E: 春愁 (Mrs.GREEN APPLE)，楽曲 F: 地球儀 (米津玄師)，楽曲 G：踊り子 (Vaundy)，楽曲 H：世界のつづき (Ado)，楽曲 I:Gold-また逢う日まで-(宇多田ヒカル)，楽曲 J：すずめ (RADWIMPS, 十明)，楽曲 K：ベテルギウス (優里)，楽曲 L：Life Is Still Going On(NCT DREAM)，楽曲 M：ロビンソン (スピッツ)，楽曲 N：Fallin(清水翔太)，楽曲 O：春を告げる (yama)，楽曲 P：POP!(NAYEON)，楽曲 Q: ガーデン (藤井風)，楽曲 R: あぁ、もう。(Saucy Dog)，楽曲 S: アイウエ (MAISONdes)，楽曲 T: 橙 (崎山蒼志)，楽曲 U：ASAP(New Jeans)，楽曲 V: 裸の心 (あいみょん)，楽曲 X：Get Up(New Jeans)，楽曲 Y：HAPPY BIRTHDAY(back number)，楽曲 Z：海のまにまに (YOASOBI)，楽曲 AA：なんでもないよ、(マカロニえんぴつ)，楽曲 AB：カタオモイ (Aimer)

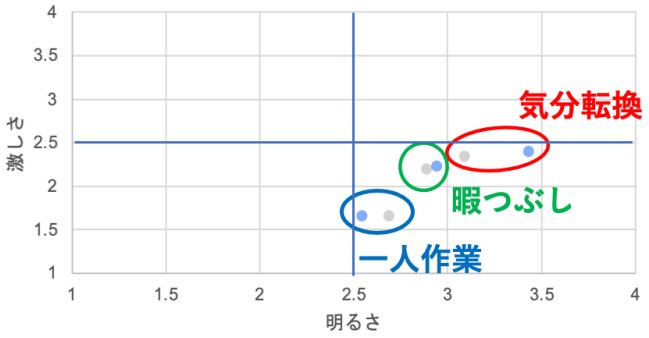


図 9　場所ごとの楽曲選択に対する比較

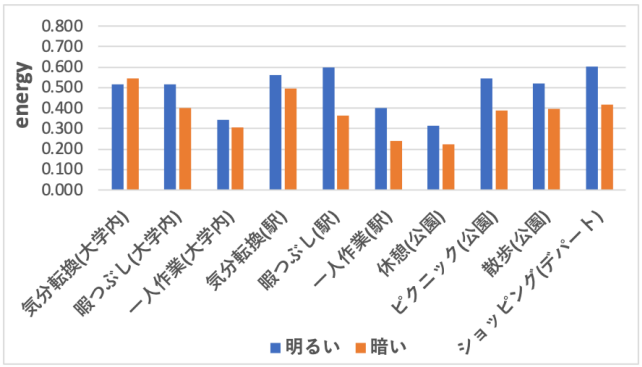


図 10　気分が「明るい／暗い」時の場所と状況ごとの energy

4.2　場所ごとの状況に対する気分の調査

まず，ユーザの気分となる「気分が明るい／暗い」と「テンションが高い／低い」に合った楽曲を各場所の状況下で調査する．なお，このアンケート調査はクラウドソーシングを介して実施し，100 名が回答した．調査内容は，「状況ごとに聴きたい楽曲に関する調査」と「状況と気分ごとに聴きたい楽曲の調査」を実施した．状況ごとに聴きたい楽曲に関する調査では，OpenAI に基づき各場所に対して下記の状況に分類した．

職場や大学内のカフェ　「気分転換」「暇つぶし」「一人作業」
駅構内のカフェ　「気分転換」「暇つぶし」「一人作業」
公園　「休憩」「ピクニック」「散歩」
デパート　「ショッピング」

被験者は各場所の各状況に対して，聴きたい楽曲の「明るさ（valence）」と「激しさ（energy）」を 5 段階のリッカート尺度で評価した．

図 8 に評価結果を示す．「一人作業」や「休憩」など静かな場所で過ごしたい場合は energy が 0.2 程度と低い傾向であった．また，全ての状況で valence は 0.6 以上と高い傾向だった．図 9 に，同じカフェで職場や大学内と駅構内で比較した結果を示す．同じ状況でも聴きたい楽曲の明るさ（valence）や激しさ（energy）に差があることがわかる．

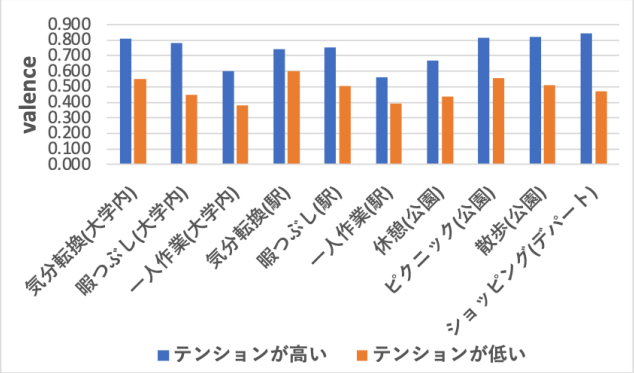


図 11 「テンションが高い／低い」時の場所と状況ごとの valence

次に，状況と気分ごとに聴きたい楽曲の調査では，明るい（激しい）楽曲を聴く際のユーザの気分の違い（例：暗い時に激しい曲を聴きたいか等）を明らかにするため，場所と状況に加え，気分が「明るい／暗い」「テンションが高い／低い」を考慮した上で，どの程度の「明るさ」・「激しさ」の楽曲を聴きたいかを評価した．具体的には，気分が「明るい・暗い」時にどの程度「激しさ（energy）」の楽曲を聴きたいか，また「テンションが高い・低い」時にどの程度の「明るさ（valence）」の楽曲を聴きたいかを 5 段階で評価してもらった．図 10 に気分が「明るい／暗い」場合の「energy」，図 11 に「テンションが高い／低い」場合の「valence」の値を場所と状況ごとに示す．図 10 より，気分が明るい時よりも気分が暗い時の方が全体的に energy の値が小さいことがわかる．また，図 8 と同様に，一人作業や休憩など静かな場所で過ごしたい場合は，energy は低かった．図 11 では，全ての場所と状況でテンションが高い場合，valence は高かった．一方で，テンションが低い場合，valence は低い傾向であったが，気分転換やピクニックでは valence は 0.6 程度と明るい楽曲が好まれた．

4.3 楽曲データの特徴量

楽曲データは，Spotify で配信されている楽曲のうちの 309 曲を API[2]より取得した．アーティスト数は 140 組であった．楽曲の特徴量として「energy(激しさ)：リラックス／アクティブ)」と「valence(明るさ：ネガティブ／ポジティブ)」が付与されており，図 12 に散布図として示す．「valence」「energy」はともに [0.0，1.0] の範囲である．「energy」の平均値は 0.752，「valence」の平均値は 0.610，「energy」の最小値は 0.298，最大値は 0.987，「valence」の最小値は 0.175，最大値は 0.963 であった．また，「energy」の標準偏差は 0.156，「valence」の標準偏差は 0.189 で分散は小さかった．以上より，評価対象となる楽曲データは，energy と valence が高く，明るく活発な曲が多いといえる．

4.4 推薦楽曲の検証

4.4.1 スポットの雰囲気と状況に対する推薦楽曲

推薦楽曲がスポットの雰囲気とユーザの状況・気分にどの程度あっているかを被験者 12 人にアンケート調査で検証した．具体的には，場所と状況と気分を提示し，楽曲の適合性を 5 段階のリッカート尺度で評価した．表 2 にスポットの雰囲気とユーザの状況を考慮した場合の既存手法と提案手法における推薦楽曲と MSE を示す．ここで MSE は，「推薦楽曲がスポットの雰囲気・ユーザの状況に適しているか」のアンケートによるユーザ評価と推薦楽曲を抽出する際に算出する距離 d との誤差である．既存手法の MSE の平均は 0.155，提案手法は 0.068 と提案手法の方が平均誤差が小さかった．また既存手法と提案手法の結果を比較すると，同じ楽曲が選ばれている場合もあるが，既存手法で誤差の大きかった楽曲が，提案手法では違う楽曲となり誤差も小さくなっている．同じ楽曲が選ばれている場合も提案手法の方が誤差が小さい．これにより提案手法の方がスポット・状況と楽曲との適合性がユーザとの間で差が小さいため優れていることがわかる．また楽曲数を増やすことで同じ楽曲が選ばれることは少なくなると考える．特に駅構内のカフェの暇つぶしでは提案手法の MSE が 0.023 と最小だった。少し明るく落ち着いた空間である駅構内のカフェで、そのような特徴のある楽曲を聴きたいと感じる暇つぶしという状況に、明るく落ち着いた楽曲である楽曲 D は適していると考える。

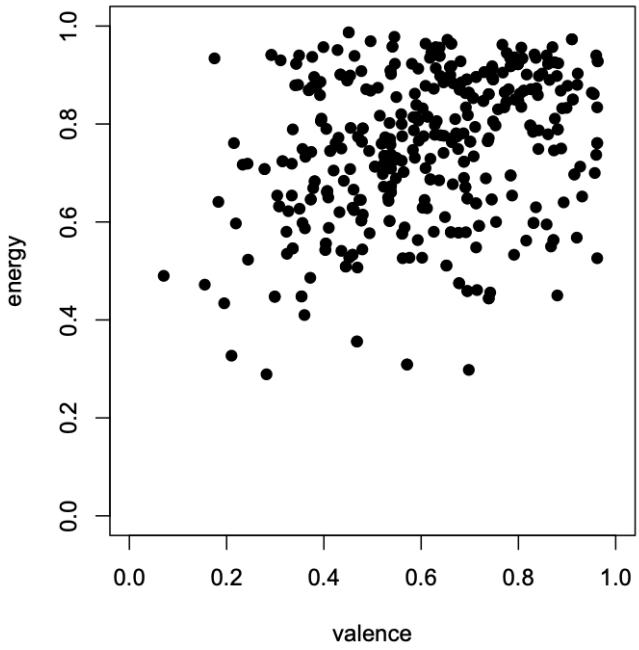


図 12 楽曲データの散布図

4.4.2 スポットの雰囲気と状況と気分に対する推薦楽曲

表 3 にスポットの雰囲気と，ユーザの状況と気分が明るい・暗いを考慮した場合の既存手法と提案手法における推薦楽曲と MSE を示す．気分が明るい時の既存手法の MSE の平均は 0.389，提案手法は 0.116，気分が暗い時の既存手法の MSE の平均は 0.171，提案手法は 0.116 とどちらの気分の場合も提案手法の方が平均誤差が小さかった．また既存手法と提案手法の結果を比較すると，気分が明るい/暗い場合も，同じ楽曲が選ばれている場合もあるが，既存手法で誤差の大きかった楽曲が，提案手法では違う楽曲となり誤差も小さくなっている．同じ楽

2：https://developer.spotify.com/documentation/web-api

曲が選ばれている場合も提案手法の方が誤差が小さい．これにより提案手法の方が優れていることがわかる．多くの場所・状況下で提案手法の方が誤差が小さいが，気分が暗い時のデパートでショッピングという状況においては，楽曲 J が推薦楽曲である提案手法の方が MSE が 0.324 と大きい．この場所・状況・気分に対しての楽曲 J の適性度について，ユーザによって評価の差が 4 と大きかった．ショッピングというアクティブな状況と気分が暗い場合については，気分よりもアクティブな状況に適した楽曲が好まれる場合と，気分にも状況にも合った楽曲が好まれる場合のどちらもあるからだと考える．

表 4 にスポットの雰囲気と，ユーザの状況と気分としてテンションが高い・低いを考慮した場合の既存手法と提案手法における推薦楽曲と MSE を示す．テンションが高い時の既存手法の MSE の平均は 0.250，提案手法は 0.084，テンションが低い時の既存手法の MSE の平均は 0.256，提案手法は 0.070 とどちらの気分の場合も提案手法の方が平均誤差が小さかった．また既存手法と提案手法の結果を比較すると，テンションがが高い/低い場合も，同じ楽曲が選ばれている場合もあるが，既存手法で誤差の大きかった楽曲が，提案手法では違う楽曲となり誤差も小さくなっている．同じ楽曲が選ばれている場合も提案手法の方が誤差が小さい．これにより提案手法の方が優れていることがわかる．公園でのピクニックにおいて提案手法の MSE が 0.027 と小さかった．少し落ち着いた曲である楽曲 D がテンションが低い時の公園でのピクニックに適していると考える．多くの場所・状況下で提案手法の方が誤差が小さいが，テンションが高い時の駅構内のカフェでの一人作業という状況においては，楽曲 X が推薦楽曲である提案手法の方が MSE が 0.101 と大きい．楽曲 P と楽曲 X を比較すると，楽曲 P の方がやや明るくアップテンポであるという楽曲の特徴が，駅構内というカフェで一人作業というやや明るい楽曲が好まれる空間・状況に，テンションが高いという気分に，より適しているからだと考える．

スポットの雰囲気と状況のみを考慮した場合の既存手法の MSE の平均値は 0.155，提案手法の MSE の平均値は 0.068 であった．スポットの雰囲気と状況と気分を考慮した場合の既存手法の MSE の平均値は 0.266，提案手法の MSE の平均値は 0.092 であった．スポットの雰囲気と状況のみを考慮した場合と、スポットの雰囲気と状況と気分を考慮した場合を比較すると，どちらも提案手法の方が MSE が小さい．これより提案手法の方が良い結果だと言える．またどちらの手法もスポットの雰囲気と状況と気分を考慮した場合の方が MSE が大きい．場所の雰囲気，状況，気分という 3 つの要素がある中で，場所の雰囲気に近い楽曲を聴きたい場合や、気分に近い楽曲を聴きたい場合など，その時々で重視したい要素が変わってくると考える．また例えば，暗い気分の場合に反対の激しい楽曲を聴きたい人もいれば，気分と同じような落ち着いた楽曲を聴きたい人もいるなど気分に対しての聴きたい楽曲は個人差が大きいと考える．今回は，クラウドソーシングでの平均値という全体の傾向を捉えたデータを用いたため MSE が大きくなったと考えられる．

5 終わりに

本研究では，状況や気分によって聴きたい楽曲が異なることに着目するとともに，具体的な状況，気分にあった楽曲を選曲することが難しいという課題を解決することを目的に，ユーザの現在位置のスポットの雰囲気抽出と状況・気分に合わせた楽曲推薦手法を提案した．具体的には，OSM とツイートよりスポットの雰囲気抽出，アンケートによる状況・気分の分析，Spotify の楽曲特徴量を用いた楽曲データより楽曲を抽出し検証した．検証の結果，これまでに行なってきた既存手法よりも本稿で提案する提案手法の方が優れていることが判明した．しかし，状況によっては楽曲データの偏りによって適しているとは言い難い楽曲が抽出されていた．今後は，被験者による定性的評価の続きを行う．そして，また楽曲データの偏りをなくし，多様なスポットを対象に，雰囲気と状況・気分に合った楽曲推薦の精度向上を検討する．

謝　辞

本研究の一部は，科研費 (課題番号：19K12240，22H03700) および京都産業大学先端科学技術研究所（M2001）の研究活動による．ここに記して謝意を表す．

文　献

[1] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperli. An emotional recommender system for music. *IEEE Intelligent Systems*, 36(5):57–68, 2020.

[2] 東海菜摘 and 湯川高志. ユーザ行動に基づいた楽曲推薦システムの提案. **情報処理学会研究報告**, 2016(9):1–5, 2016.

[3] 韓語佳, 中野 美由紀, and 小口正人. 利用者の印象に基づく音楽レコメンドサービス-音楽データベースの感情空間へのマッピングに関する考察. *DEIM*, 2021(D25-3):1–8, 2021.

[4] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 397–407, 2019.

[5] Xinxi Wang and Ye Wang. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 627–636, 2014.

[6] J.Russell. *A circumflex model of affect*. Journal of Personality Social Psychology, 1980.

[7] Georgios PALTOGLOU and Michael THELWLL. *Seeing stars of valence and arousal in blog posts*. IEEE Transactions of Affective Computing, 2013.

旅行者の時空間的・感性的条件を満たす
旅程・店舗の組み合わせ生成方式
－ Personalized Traveling MaaS －

神原 隼登† 林 康弘‡ 清木 康‡

†武蔵野大学 有明キャンパス 〒135-8181 東京都江東区有明三丁目３番３号

E-mail: †s1922009@musashino-u.ac.jp, ‡{ yasuhiro, Hayashi } hayashi@musashino-u.ac.jp

あらまし

本研究は観光, 通勤・通学といった移動全般を対象として, 旅行者の時空間的・感性的条件に基づく旅程と経路周辺の店舗の組み合わせを生成する方式について提案する. 本方式は, 旅行者の旅行に関する文脈（時間・空間・感性）に基づいて,実空間における出発地から目的地までの距離計算（経路探索）と, 移動および滞在に必要となる時間計算（スケジューリング）, 感性的条件を満たす店舗の分類計算（感性計量）を行い, それぞれの計量結果を正規化し統合することにより, 移動経路とその経路上で訪問可能な店舗からなる旅程を生成する Personalized Traveling MaaS として, 旅行者に新たな移動価値を提示できる可能性を示した.

In this paper, we propose a method for creating combinations of itinerary and stores around a route based on a traveler's spatio-temporal and emotional conditions in sightseeing or commuting as contexts. This method calculates distance from the starting point to the destination in real space (route search), time required for travel and stay (scheduling), and the classification of stores that satisfy the emotional conditions (emotional measurement). Then, this method normalizes and integrates the results of each measurement. We showed the possibility and feasibility of this method. This method has a possibility to provide new travel values to many travelers as the Personalized Traveling MaaS, creating travel routes and stores that can be visited along those routes as an itinerary.

キーワード 相関量計算, 文脈情報, 時空間情報, 組み合わせ生成, MaaS

1，序論

本研究の目的は, 観光や通学・通勤といった移動活動を行うユーザーを対象として従来の距離を中心とした経路検出(ナビゲーション)に加え移動に関する経路情報と周辺の店舗情報を連携させ, スケジュール,ニーズに関する情報から移動経路だけでなく訪問可能な店舗の提案を含めた旅程提案方式の提案と, それにより個人の旅行ニーズを反映した旅程生成「Personalized Traveling Mobility as a Service(MaaS)」に向けた新たな移動活動の選択肢を方法論の構築をすることである.

本提案方式は, 事前に訪問時の感性に関するコンテキストな情報を格納することで店舗毎の特徴を再現した店舗 DB を構築し, 利用時にはユーザーの移動における条件となる移動目的のスケジュールに関する, 地点, 座標情報からなる”予定”情報と店舗訪問提案を行う上で必要となる, どういった活動, 体験をしたいかといったニーズに関する”体験”情報. これら 2 点の情報の取得を行う.

本方式はこれらの予定と体験情報に基づき, 座標情報からユーザーが訪問可能な店舗を抽出し, 移動時間や店舗の訪問時間に関する時間的計量と店舗の感性的特徴情報とユーザーの体験に関する情報の距離計算による店舗毎の感性距離スコア計量を行い, 以上の時空間やニーズに関するマルチモーダルな情報の計量と統合によってユーザーの入力した条件,要望に適した移動情報とニーズに合った訪問可能な店舗群と移動に関する時間情報によって構成される旅程を取得するものである.

我々は近年, パンデミックの影響により外出の自粛を余儀なくされる中, デジタル化へのニーズが増加し, 多くの企業や個人がデジタルトランスフォーメーション(DX)やリモートワークの導入が盛んになった点に着目する.

ネットワーク技術の進歩の中で 人々は, リモート活動等の場所に囚われない仕事等の活動に対する心理的ハードルが低くなり,「タイムパフォーマンス(時間対効率)」と呼ばれる, 移動を含めたあらゆる活動毎の体験価値における関心が増え, この影響は移動活動に対する社会的価値観の変化に繋がっていると考える.

本研究を通じ,様々な移動需要に応じてユーザーに新たな選択肢を提供可能な MaaS システムを構築する事により,今後も日々変化する移動需要へ対応可能な未来の移動に関する新たな展望の提案を行う事を目的とする.

2, 関連研究

2.1 MaaS

MaaS（Mobility as a Service）は, 移動に関するあらゆる

情報を 1 つのサービスとして統合を行う新たな移動形態である. 移動に関するデータの都市やサービスに関するさまざまなデータをリンクさせる事により, 観光や小売といった交通以外の幅広い領域の活動にシームレスに繋ぎ,移動需要に新たな価値創出をもたらすサービスの総称を指す.

MaaS の起源は 2010 年初頭にヨーロッパで始まり, 特に 2014 年にフィンランドのヘルシンキで初の MaaS プラットフォームが導入された. この MaaS ビジネスは, さまざまな形態の交通および交通関連サービスを単一, 包括的でオンデマンドのモビリティサービスに統合することにより ユーザーは単一のアプリケーションと単一の支払いチャネルを通じてモビリティにアクセスするものである.

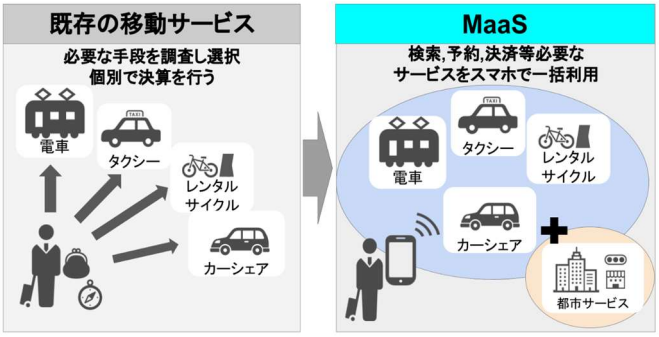


図 1 MaaS の概要図
（イラスト利用：ICOON MONO）

2.2 モバイルコンピューティング環境におけるユーザー情報の能動計量による能動型情報配信方式

時空間の移動を行うユーザーのモバイルコンピューティング環境において, ユーザーの状況や意図に応じて能動的に適切な情報の提供を行う事を目的として, ネット上に存在する様々な形式を持つ異種データベース群を情報源として利用可能な”メタレベル・アクティブマルチデータベース”の構築によって, 時間的, 空間的に変化するユーザーの状況と意図に基づき, ユーザーの状況や意図と局所データセット間の動的な相関の計算を行い, 提示を行うシステムである.

また能動型情報配信方式におけるユーザーデータやメタレベル・アクティブマルチデータベースの分類として, ユーザーの環境情報に関する情報である Situation(状況)とユーザーの意図や目的によって構成される Intention(意図)からなる 2 種とこれらにおけるユーザーの年齢や職業等の時間の経過による変化が無い Static(静的)情報と座標や移動パターン等のユーザーの状況や意図によって変化する Dynamic(動的)情報.これらの 2 種からなる計 4 分類によって構成される.

2.3 移動経路における感性コンテキストに対応する移動・活動手段の組合せ形成システム－Emotional MaaS－

当論文は, 旅行者の意図と場面に対応した感性情報に関するコンテキ スト(文脈)を対象とし,意味的連想検索モデル[意味の数学モデル]を応用した旅行者の意図と状況に応じた感性コンテキストベクトルと移動手段(Mobility Space), 経路上の移動活動(Activity Space). これらの組み合わせと距離計算により旅行者の感性コンテキストに対応し, 乗り換え案内アプリ等では知り得なかった公共交通枠を超えたマルチモーダルな「移動手段」および移動経路上の店舗等「活動手段」の組み合わせ形成を実現する事で, 移動中における日常的な欲求を満たす事を目的としたシステムである.

また, 本システムの感性コンテキストに関しては旅行者が抱く「専念・集中したい」や「リラックスしたい」,「観光したい」等といった”日常の意図”(Intention)と, 「通勤・通学」や「帰宅」,「お出かけ(日帰り)」といった移動の場面(Situation),これら 2 種類のベクトルを組み合わせる事で生成され, これらの情報が活動空間の特徴量空間と関連付けられる事で感性コンテキストは旅行者の移動中における個々のニーズや状況の反映を実現させている.

また本稿(Personalized Traveling Maas)と本章で述べた論文(Emotional MaaS)との比較点としては, Emotional MaaS の方式は経路上に存在する, ニーズに即した店舗群といった座標情報を提示しているのに対し, Personalized Traveling MaaS は, ニーズに即した店舗と経路情報に加え,時間計量を行う事により, 移動経路とニーズに即した計量上最適な店舗の組合せからなる旅程(スケジュール)を提案する点にある.

3, 提案方式

3.1 特徴

旅程生成過程の概要としてあらゆる移動目的における対応区域間の移動を対象として, 特定の時間, 場所に移動するユーザーを対象として図 2 の検索から旅程を通じて目的地に到達するまでに想定するユーザーの旅程における, DT(Departure Time)[出発時刻]と AT(Arrival Time)[到着時刻]と座標, ニーズに関する情報をユーザーが入力する事で下記①の計量用データベースを構築, 下記ステップの DB の構築,計量を行う事で時空間計量に適合し店舗の特徴を基に最適化された組合せの店舗情報と時間計量結果からなる旅程の生成を行う.

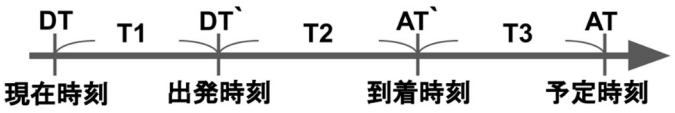


図 2 ユーザー旅程図

3.1.1 計量用 DB の構築

本項では以降の計量に必要な店舗特徴と, ユーザーデータの分類を行う為, 以下各種特徴の分類方式を基にした店舗 DB, ユーザーDB の構築を行う.

店舗毎の特徴分類方式としては, 主に「飲食」や「物品購入」等の達成可能な目的である目的情報と「Relax(落ち着く)」や「Intensive(集中できる)」といった店舗訪問時の想定感性情報からなる感性情報, この 2 種を格納する事で店舗毎に異なる特徴再現か可能であるとして下記条件を基とした分類による店舗, ユーザーに関するデータベースの構築を行った.

ユーザーDB に格納する情報は, 表 1 のユーザーデータ分

類の通り, 想定する移動活動における目的地に関する”予定情報”と, 店舗訪問といった生成結果のニーズに関する”体験情報”による 2 種と, それらにおける店舗や最寄り駅や店舗ジャンルといった分類可能な情報である静的情報, 時間や店舗毎の以降移動ルート生成において計量による分析必要な動的情報,これらの 2 種による分類が可能な計 4 項目の情報を抽出することで旅程生成が可能であるとして以降の生成時では入力内容を基に各種計量を行う.

表 1 ユーザーデータ分類

	静的 (分類可能情報)	動的 (計量, 分析が必要)
予定 (目的情報に関連)	座標	時間
体験 (体験に関するニーズ)	目的	感性

旅程検出に用いる主なデータベースとしては,図 3 のデータモデルの通り 4 項目の情報を格納し, 店舗情報に関しては 2 つのデータベースを用い, 1 つ目は座標,時間情報と感性情報からなる店舗データベース, 2 つ目はジャンル毎に共通の情報である店舗の目的情報とそこから推測される想定滞在時間からなるジャンルデータベースによって構成され, この 2 店舗はジャンルデータベースのジャンル ID によって紐付けされる.

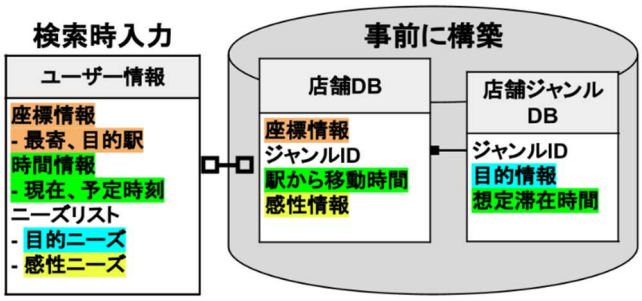


図 3 提案方式のデータモデル

3.1.2 静的情報による店舗 DB フィルタリング適用

事前構築した店舗 DB の取得に際し, ユーザーデータ分類における静的情報である座標情報と体験への目的情報に合致しない店舗の提案を行う場合実際の訪問へ繋がるのが困難である点から,下記条件を満たす店舗を静的情報に適応した訪問の可能性がある店舗として以下行程による組合せ検出を行う.

店舗 DB 取得における条件としては, 座標情報としてユーザーの座標(現在, 目的地点)に一致しない座標情報を持つ店舗 ID を,目的情報に関してはユーザーの環境ニーズにて選択した項目と同一のジャンル毎の行列の要素毎の積(アダマール積)が 0 である店舗を除外することで静的情報に基づくフィルタリングの実装を実行する.

3.1.3 ユーザー情報に基づく時間的計量の適用

(1) 感性情報を用いた店舗毎の所要時間検出

本提案方式の時間計量おける各店舗訪問時の所要時間 RT(Required Time)は店舗 DB 内の所要移動時間 DT(Distance Time)とジャンル DB 内の想定滞在時間 ST(Stay time)を基に訪問に必要な基本時間の検出を行い, [落ち着く]等のユーザーの感性情報が一部条件と一致した際に通常滞在時間から 1.5 倍にした上で以降の時間計量を行う事によりユーザーの感性情報に基づく時間計量を行う.

(条件一致時)

$$RT = (DT \times 2) + (ST \quad [\times 1.5])$$

(2) 探索可能時間生成

本方式における移動時間を除く店舗の探索, 訪問に関する上限時間の生成方式として, ユーザーの入力情報から時間情報である DT(Departure_Time)[出発時刻], AT(Arrival_Time)[到着時刻]と共に, 座標情報である現在, 予定地点の座標情報から地図系 API 等を用いる事で移動時間(T2)を検出する, これらの情報から訪問, 探索における上限時間となる ΣExploratory_Time(探索可能時間)を設定する.

$$\Sigma ET \quad = \quad (AT - DT) - T2$$

3.1.4 ユーザー情報との感性計量による最適化

本方式におけるユーザーの感性ニーズに適した旅程生成の為,事前に提案可能な店舗毎にユーザーが入力する感性情報と同一軸を持つ感性データを構築する事により, ユーザーの感性ニーズとの距離計算により店舗毎にユーザーニーズとの類似度スコアの検出を行う. また, 類似度スコアの計量方式としては構築した感性情報の分類軸の設定時,其々独立することを条件としており,各感性同士の関係性や距離に関する生成が困難な点から角度計算を行わず直線距離を生成するユークリッド距離を用いてユーザー感性データの”U” と店舗“A”其々が格納する特徴量の距離生成を行い最大距離から引いた物を d(U,A)とする.

$$d(U_x, A_x) = \sqrt{x} - \sqrt{(U_1 - A_1)^2 + (U_2 - A_2)^2 \ldots + (U_x - A_x)^2}$$

3.1.5 最適な旅程提案の為の店舗群生成

3.1.3, 3.1.4 の各種計量により生成された店舗毎の時間的, 感性情報を基に訪問に最適な店舗群の検出を行う.

店舗群の検出方式としては Step2 で取得した店舗群において考えられるすべての組み合せを検出し, 組合せ毎に 3.1.3 内(1)にて取得した訪問所要時間(Stay_time)と 3.1.4 の類似度スコア d(A, B),これらの合計値を格納する, 構築した全ての組合せの中から,以下 2 項目の条件を満たす組合せを最適な店舗群とする.

① 各店舗の探索時間(RT[x])の合計値(ΣRT)が探索可能時間(ΣET)を超過しない

$$RT[a] + RT[b] \ldots = \Sigma RT$$
$$\Sigma RT \leqq \Sigma ET$$

② 距離スコア d(A,B)の合計値が最も高い

$$\Sigma d(U_x, \alpha_x) = d(U, A) + d(U, B) \ldots + d(U, x)$$

3.2 データ構造

本方式が用いるデータ構造としては図 4 の通り, 店舗とそのジャンル, ユーザー情報によって構成され, 座標情報は駅毎の id によって管理を行う, 目的, 感性情報に関しては該当の有無を 0 or 1 によって格納し, ユーザーデータにおけるジャンル ID と目的情報,感性分類に関する詳細な分類方式を以下に述べる.

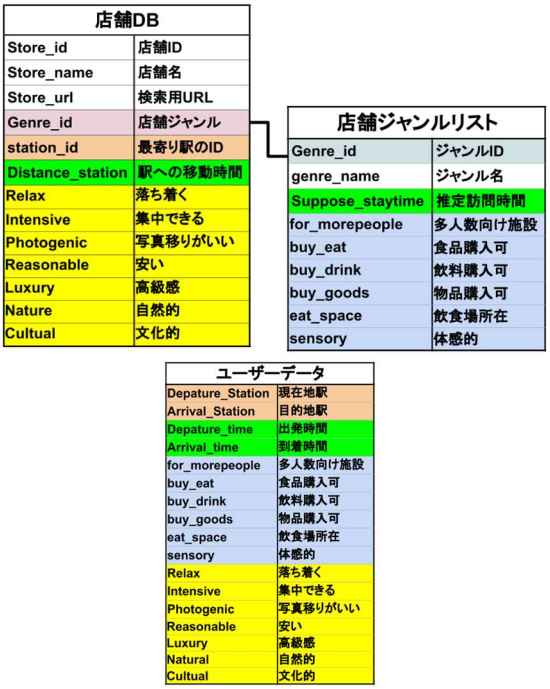


図 4　データ構造

①　店舗ジャンル毎の目的分類モデル

本方式におけるユーザーデータ分類内の体験に関する”目的情報”に関し, 計量の短縮を目的として同一の目的情報が達成可能な店舗をカフェや公園等の施設のジャンルとして分類を行い, 3.1 内 Step2 の静的情報に基づくフィルタリングを行った, 各ジャンルが含む想定設備に関する設備情報は店舗訪問における主な目的として考えられる”飲食”, ”物品購入”といった情報を格納した.

また,本方式の時間計量における店舗毎の滞在時間に関する時間計量の際,これらは達成可能な目的から推測可能である点から,ジャンル毎に滞在時間(Suppose_Staytime)を設定し, 店舗毎のジャンル ID により店舗毎の所要時間検出を行った.

表 2　店舗ジャンル毎の環境,時間情報一覧

ID	Genre_name		Suppose Staytime	for_more people	buy_eat	buy_drink	eat_space	buy_goods	sensory
	名称	*該当施設*	想定滞在時間	多人数向	食品購入可	飲料購入可	飲食場所有	物品購入可	体感的
1	カフェ		0:30	0	1	1	1	0	0
2	公園	緑地,神社,寺,庭園	0:30	1	1	1	0	0	1
3	商店	書店,文具店,モール	0:30	1	0	0	0	1	1
4	レストラン	ファミレス,ファストフード	0:30	1	0	0	1	0	0
5	体験施設	映画館,科学館,博物館	1:00	1	0	0	0	0	1

②　店舗毎の感性分類モデル

本方式における店舗毎に異なる多様な特徴を多次元情報として分類し格納を行う為, 店舗毎の訪問時に起こるユーザーの想起感情を感性コンテキストとして幾つかの軸に分類し, 各軸の想起感情の傾向的な有無を格納する事で店舗毎にそれぞれ異なる特徴を再現した感性分類モデルを格納した店舗 DB の構築を行う.

各感性コンテキストの分類方式としては, 表 3 の感性コンテキスト分類と各分類のルールベースの通り, ”楽しい”や”美味しい”といったユーザーが持つニーズとして前提となるような表現や, 其々のキーワードに関連性が少なく独立している事を条件として 7 つのキーワードとその定義であるレビュー文等の文章にある様な感情表現や設備といったキーワードの設定を行った,また 4.3 の実験 2 内でキーワードの条件である独立性の検証として実在する店舗の情報を基にした店舗 DB 構築し格納する特徴情報の相関性検証を行った.

また,感性分類コンテキスト分類に用いるキーワードは独立性を条件としているが,各条件の有無を 0or1 で格納しておりマイナスの表現が無い点から,2 つ共値段に関する[Reasonable]と[Luxury]といった 2 軸を用いており,結果として表 2 内の各軸の条件における[Reasonable]の”positive”, [Luxury]の”negative”の項目内に安いという同一のキーワードが存在する結果となった, しかし, 同一ではないキーワードも多く存在し 4 章 3 項実験 2 の実店舗情報を用いた独立, 相関性に関する検証の中でこれら 2 軸を含むすべての軸に強い相関が検出されなかった点から,表 3 の感性コンテキスト分類方式は独立した感性分類によって構成されているものであるとして構築を行った.

表 3 感性コンテキスト分類と各軸のルールベース

Genre	日本語	滞在時間に影響	条件 Positive	条件 Negative
Relax	落ち着く	1	落ち着く,静か,リラックス,ゆったり	うるさい,狭い,混雑
Intensive	集中できる	1	集中できる,静か,勉強,仕事,Wi-Fi,電源	混雑,騒がしい,電源なし
Photogenic	美しい・綺麗な	0	フォトジェニック,写真映え,映える,おしゃれ	地味,古臭い,特徴なし,一般的,普通
Reasonable	適正な	0	リーズナブル,コスパ,手頃,安い,得	高い,割高
Luxury	豪華な	0	高級,ラグジュアリー,豪華,エレガント,上品	一般,チープ,ジャンク,量産,普通,安,シンプル,リーズナブル
Nature	自然的	0	ナチュラル,オーガニック,無添加,自然素材,植物,自然光	人工的,閉鎖的,添加物
Cultural	文化的	0	伝統的,地域,地元,芸術的,文化的背景,ストーリー,景色	無個性,一般的,スタンダード,無地元特色

3.3 システム概要

本方式は学生, 社会人等幅広い層の通学通勤,観光客等,多様な行動目的(環境情報)を持つユーザーが「特定の駅へ予定時刻までに移動を行う必要性がある」といった状況を対象として, 現在地点から目的時間までに目的地店に到達しながら訪問可能かつニーズに適した店舗群と移動に必要な時間情報によって構成される旅程の提示を行うものである.

本方式の利用工程としては,システムを起動し,予定,体験に関するユーザー情報の入力を行うと, 3.5 の旅程検出過程を

実行する事により, 店舗群と移動に必要な時間情報を取得し画面上での提示を行うといった流れである, その後ユーザーは提案店舗の名称と共に表示される店舗 DB の構築に用いた物と同一の URL から詳細な情報を取得し活動を行うといった形態である.

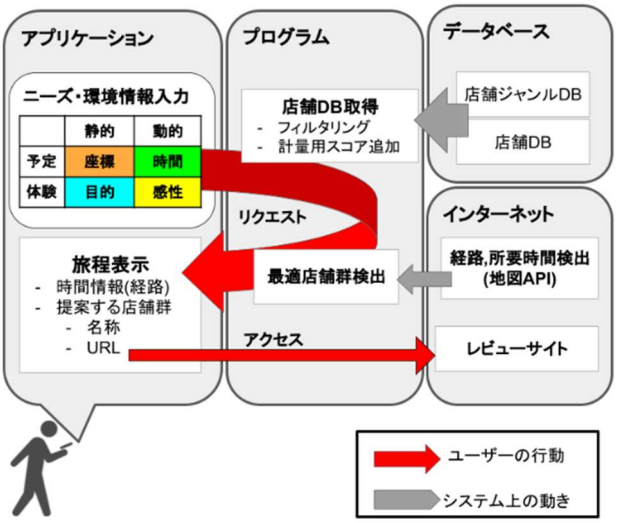


図 5 ユーザーフロー
(イラスト利用 ICOON MONO)

3.4 旅程構成情報,条件設定

本方式は, 3.3 システム概要で述べたユーザーが探索可能な店舗群と移動の時間情報からなる旅程生成において,提案する店舗群に関し, 3 店舗を上限として旅程検出を行う.

旅程生成における上限設定の理由としては, ユーザーの入力した現在時刻と予定時刻が大きく離れる場合, 一度に多くの店舗訪問を提案する事になり, 多すぎる選択肢はユーザーが旅程に対し探索の判断をする際に,意思決定にかかるプロセスが複雑化する事で活用に対する決断を困難にし,提案内容の満足度を低下させる結果となる”選択のパラドックス”を誘発する可能性が高い為であり, 提案時の認知の負担が少なく期待値を設定が容易な 3 店舗を店舗群の上限とした.

また,旅程における店舗群の上限設定により,現在時刻と.予定時刻の乖離が大きい場合, 訪問と移動以外の余分な時間が多く発生する点から, 店舗群と共に訪問を行う上で適切な出発時刻として ΣRT(合計滞在時間)と T2(移動時間)の合計値が丁度 AT(予定時刻)となる時間である Estimated_departure_Time(出発目安時間)を旅程探索必要な時間情報として提示を行う.

$$Estimated_Departure_Time = AT - (T2 + \Sigma RT)$$

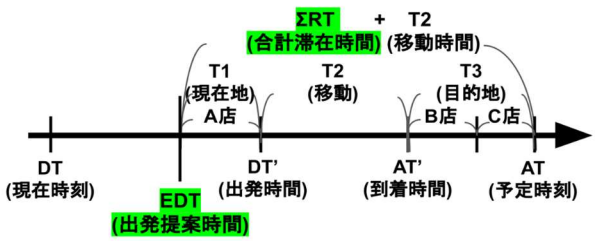


図 6 ユーザー旅程図と提案時間

3.5 旅程の検出過程

3,3 のシステム概要における旅程生成において, 3.1 の提案方式の特徴に基づく時空間,感性計量に基づく計量の実装が必要であるが,具体的な旅程検出の過程としては図 7 のシステム概要の通りであり,下記の工程を実装することでユーザー情報を基とした旅程生成を行い,4.2 の実験 1 にて本方式で検出された旅程に関する有効性に関する検証を行った.

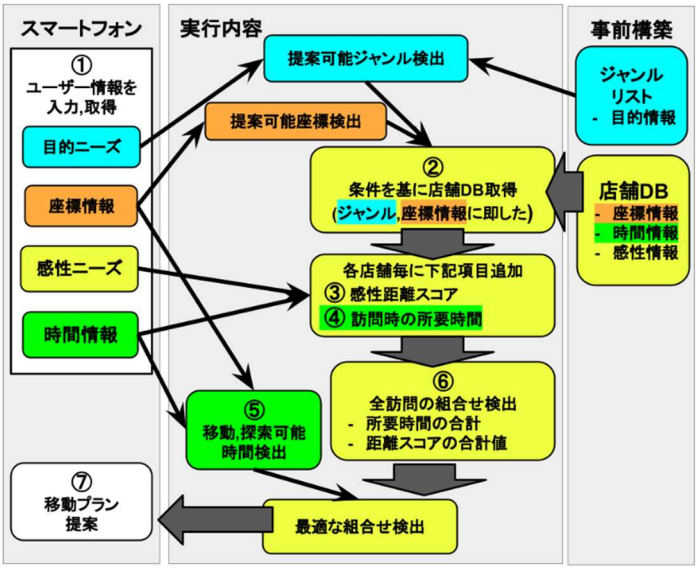


図 7 システム概要図

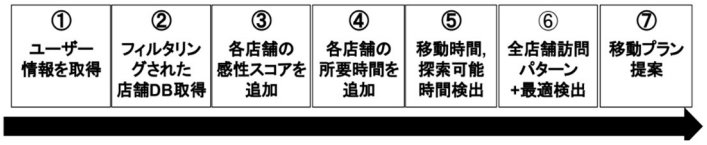


図 8 システム概要の略図(実行順)

Step1　ユーザー情報を取得

利用者の入力により時間,座標情報と目的,感性に関するニーズに関するユーザーデータを取得する.

Step2　フィルタリングされた店舗 DB 取得

店舗 DB の取得時,表 1 のユーザー分類における静的情報である座標,目的情報適合した店舗情報取得する事で実際に訪問可能な店舗に限定した以降の旅程検出を行う.

(1) Depature_Station,　Arrival_Station のどちらかに一致する持つ店舗(座標情報に適合)

(2) ジャンル DB 内の目的情報とユーザーデータ内目的ニーズ同士の行列の要素毎の積(アダマール積)の合計値が 0 以外の Genre_ID (目的情報に適合)

Step3　各店舗の感性距離スコアを追加

ユーザーデータの感性ニーズと同一の分類によって構成される店舗の特徴情報同士で距離計算を行う事により類似度検出を行う.

距離計算の方式としては, 特徴情報である感性に関する分類軸がそれぞれ独立している点からユークリッド距離(直線距離)による検出を行い, また,類似度の距離スコアであるユーザーデータ(U)と店舗データ(A)の距離スコア D(U,A)に関し,

本システムで用いる感性分類 7 軸の最大距離であるルート 7 を 0, 最短距離を 100 とする下記式を用いる事により, スコア確認, 検証時の有用性向上を図り, D(U,A)のスコアは取得した店舗 DB ごとに DB 内の店舗毎に追加する.

$$D(U,A)=\left\{\sqrt{7}-\sqrt{(U_1-A_1)^2+(U_2-A_2)^2\ldots.+(U_7-A_7)^2}\right\}\times 100$$

Step4　各店舗の所要時間を追加

店舗ジャンル毎に異なる Suppose_Staytime(DB 想定所要時間)と店舗 DB 内の Distance_Station(所要移動時間), ユーザーの感性ニーズ情報から店舗毎の Requied_Time(所要時間)を生成し, 各店舗 DB に追加する.

Step5　移動,探索可能時間を生成

Depature_Station, Arrival_Station 情報から API 等を用いて図 2 のユーザー旅程における公共交通機関を利用した場合の T2(移動時間)を取得,これらの情報から ΣET(探索可能時間)を検出する.

Step6　全ての組合せと最適プラン生成

Step2 で生成した訪問可能な店舗 DB に関する全ての Store_ID の組合せを生成し, 組合せ毎に Step3 で生成された距離スコア D(U,A)の合計値である ΣD (合計距離スコア)と Step4 で生成された RT(所要時間)の合計値である Σ RT (合計所要時間)を組合せ毎に検出する. また, この組み合わせの内で ΣRT(合計所要時間)が Step5 で生成した ΣET(探索可能時間)未満の中 ΣDT(合計距離スコア)が最も高いこの 2 点の条件に合った店舗 ID のリストを最適な旅程として生成する.

Step7　最適な組合せをリコメンド

Step6 で生成した最適な店舗 ID の組合せを基に店舗 DB と Step5 の時間計量結果を用いて,訪問に必要な店舗情報 (Store_name, URL)と所要移動時間(T2), である Estimated _departure_Time(出発目安時間)の提示を行う.

4.　実験

4.1 実験目的,条件設定

本提案方式の有効性検証を行う上で, テスト対象となる表 4 の条件設定を基に,東京都内主要 5 駅間の移動を対象として,下記条件を満たす各駅の周辺に存在し, 店舗ジャンルの区分内に該当する店舗を各駅, ジャンル毎に其々5 店舗を対象とした店舗 DB の構築を行った.
また,店舗毎の特徴情報として, ”食べログ”や”じゃらん”といったレビュー情報が記載されたサイトや店舗の公式 HP に記載される環境情報を基に 入力を行った.

表 4　店舗 DB のパラメータ,DB に追加する条件

テスト店舗DBパラメータ	
店舗データ件数	74
対象駅	5駅 (東京,新宿, 渋谷,池袋,上野)
取得条件 (訪問不可能な店舗を除外)	- 訪問時の予約不要 - Distance_Station(駅からの移動時間)が00:20未満

また本章では本提案方式の正当性検証として, 旅程の生成結果に関する有効性検証と,本章で構築した店舗 DB が格納する特徴情報が持つ分類軸其々の独立性検証を行うによる感性分類と距離計算の適応性検証. これら計 2 点の実験を行った.

4.2 実験 1 ：提案計算方式の有効性検証

3.5 の旅程生成方式における時空間,ニーズ情報に基づく旅程生成方式の有効性検証の為, 複数のユースケースから考えうるユーザーデータ構築,ユースケース毎の生成結果を基に検証を行う.

Step1　ユースケース, 想定提案店舗の設定

本実験では同一の時空間情報を持ち,目的,感性に関するニーズ情報となる行動目的が異なるユーザーA,B,2 種のユースケースと, それにより考えられる表 5 のユーザーデータを設定し, 検出された旅程に関するユーザー情報の適合性に関する検証を行う.

また,感性情報に基づく最適化に関する検証方式としては, 座標情報が移動駅情報と一致し,それぞれのユーザーと同一の感性,特徴情報を持つ店舗を想定提案店舗として設定, 出力結果に想定提案店舗が含まれている場合を感性計量による最適化が行えているとする.

共通:「予定情報」 現在(東京駅,10:00),予定(新宿駅,13:00)
個別:「体験情報」
ユーザーA:観光(複数人で行動,様々な体験に興味)
ユーザーB:通学(個人で行動, ゆったりとした環境に興味)

表 5　ユースケース毎の入力ユーザーデータ

	A(観光客)	B(通学客)
Depature_Station	東京	東京
Arrival_Station	新宿	新宿
Depature_time	10:00	10:00
Arrival_time	13:00	13:00
for_morepeople	1	0
buy_eat	0	1
buy_drink	0	1
buy_goods	0	0
eat_space	1	1
sensory	1	0
Relax	1	1
Intensive	0	1
Photogenic	1	0
Reasonable	0	1
Luxury	0	0
Natural	1	0
Cultual	1	0

Step2　出力結果,考察

表 6　ユースケース毎の生成過程,出力結果

		A(観光客)	B(通学客)
指定	想定提案店舗 (感性ニーズと同一軸の特徴を持つ店舗ID)	[24]	[22]
過程	全データ数(店舗数)	73	73
	フィルタリング適用データ数 (座標,目的情報の条件適応)	13	14
	移動時間 (T2)	20:00	20:00
	探索可能時間 (T1+T3)	2:40:00	2:40:00
出力結果	提案出発時間	10:30	10:54
	最適パターン [店舗ID] (枠は想定提案店舗)	[17,24,30]	[4,5,22]

考察

予定(時空間)情報に関し,座標情報においては提示する店舗群が Depature_Station(出発駅),Arrival_Station(予定駅)に該当, 時間情報に関して旅程が訪問可能な Estimated _departure_Time(出発目安時間)が Depature_Time(出発時間)より後の時間である点からユーザーの時空間情報に適合した旅程検出が行えており, 体験(ニーズ)情報に関しても,想定訪問施設を含む提案店舗群の検出が出来ており, 以上の結果から 2 点のユースケースに関してユーザーの予定,体験情報に基づく旅程生成が行えたことが確認された.

また,本実験とは異なる目的情報から再現性検証を行った際も同様に予定,体験情報に基づく検証結果が出力された点から, 本方式は,時空間,店舗の環境情報に適合し,かつ感性計量による利用者のニーズに即した旅程生成が可能であるとする.

4.3 実験 2 ：提案計算方式の適用性の検証

本章では, 実験 1 の旅程生成においてサンプルとして用いた店舗 DB の妥当性や, 独立性を条件とした店舗の特徴を表現する感性コンテキスト分類方式に関する適用性の検証を目的として, 表 6 の条件を基に実在の店舗情報を基に構築した店舗 DB における感性分類其々の軸における独立性(相関性の有無)に関する検証を行った.

相関係数を用いた検証を行う理由としては,店舗 DB の特徴として格納している特徴軸(キーワード)が独立性を条件としており,その点からユーザーの感性情報との距離計算時に角度計算を伴わないユークリッド距離(直線距離)による計量を行っている点から, 実店舗の情報を基に構築した店舗 DB の店舗が持つ軸に相関性が無い事を証明する事に付随して感性計量における分類方式,距離スコアの検出方式の正当性を行える為である.

Step1 仮説

本実験の仮説としては,店舗データベース構築条件である表 3 の感性分類において,構築時に対称となるよう設定した[Reasonable]と[Luxury]以外の項目にある程度負の相関性が発生するが, [Reasonable]と[Luxury]を含む全ての軸同士に強い相関は発生せずユークリッド距離によるユーザーニーズとの距離計算が最も適切な計量方式である.

また,検証における相関係数の度合として,2 軸のカテゴリ変数間の独立性を生成可能なカイ 2 乗推定における p 値を用いて検証を行い,優位水準としては,0.05 以上の場合を強い相関性の基準として,相関係数が最大と最小となった項目の p 値検出によりサンプル店舗 DB における特徴情報分類の独立性検証を行う.

Step2 結果

表 8　各感性特徴軸の相関係数一覧

	Relax (落ち着く)	Intensive (集中出来る)	Photogenic (写真映えする)	Resonable (適正な)	Luxury (豪華な)	Natural (自然的)	Cultual (文化的)
Relax		0.26	0.10	-0.15	-0.22	0.25	0.07
Intensive			-0.17	0.13	-0.20	-0.16	-0.24
Photogenic				-0.22	-0.07	0.18	0.34
Resonable					-0.28	-0.05	-0.02
Luxury						-0.12	0.20
Natural							-0.01
Cultual							

正の相関(最大)
正の相関
負の相関
負の相関(最小)

最大:[Photogenic],[Cultural] = p 値 0.0075 (相関係数 0.34)
最小:[Reasonable],[Luxury] = p 値 0.0381 (相関係数-0.28)

各感性軸の相関軸は仮説の通り, [Reasonable],[Luxury]の組合せに強い負の相関が見られたが, 他にも[Photogenic]と[Cultural]に関して仮説の[Reasonable],[Luxury]より強い p 値を持つ正の相関性が検出された, しかし上記 2 点においても有意水準である 0.05 未満である点から, サンプル店舗 DB 内の分類において強い相関性は発生しなかった.

Step3 考察

本方式における, 店舗毎の特徴である訪問時の感性情報に基づく分類は,其々強い影響を持たず独立した分類による特徴情報を抽出しており, この点からユークリッド距離に基づく直線距離は,店舗の特徴とユーザーニーズとの距離スコア生成において適した計量方式であるとする.

また[Photogenic]と[Cultural]の相関性が仮説で挙げた同一のキーワードを持つ[Reasonable],[Luxury]より大きい相関性検出した点に関しては,文化的[Cultural]の条件が伝統や地元,景色といった観光資源全般を指した表現である点から, 伝統的または文化的に重要な場所が,SNS 上での写真共有が盛んな現代において魅力的な撮影スポットとなり得るといった現代の文化性が強く反映された結果によるものであり,より特徴軸同士の相関性を低下させ,独立性の高い表現を活用するには観光全般に関するキーワードであった[Cultural]をより詳細な内容の表現である歴史,時代,古いといった条件からなる[historical]といったキーワードを用いる事で,直接的な関係性を更に減らす事が可能であると考えた.

実験 1,2 の結果により, 本提案方式は, ユーザーや店舗の情報を適切に分類し, 時間, 空間, 店舗環境. 感性ニーズといったマルチモーダルな情報を統合し、計量可能な方式であることが示された

6. 結論

6.1 まとめ

本研究は,ユーザーの移動活動における体験価値の最大化を図る為, 時間と空間,体験に関するマルチモーダルなユーザー

情報を分類, 収集し, 統合する事で適切な店舗,移動情報の組み合わせから実際の訪問に転換が容易な旅程提案を可能とする”Personalized Traveling MaaS”の構築を目指した.

この目的達成の為,本稿ではユーザーから抽出可能で分析に必要な情報を特定,環境及び訪問時の感性表現に基づく感性コンテキスト情報の分類方式を策定し, これにより時間的, 空間的要因に合致する最適な店舗群と移動時間, タイミングによって構成される旅程提案を行うアプリケーションの構築を進めた.

本研究の成果は以下の 3 点にまとめられる.

1, 移動活動における旅程提案に必要なユーザーデータ分類モデルの構築

旅程提案において必要なユーザーデータを, 予定, 体験及び静的, 動的情報に分類し,分類毎の計量, 分析を行う事で実装時にユーザーに余計な操作負担をかけない効率的な情報収集モデルの構築を行った.

2, 店舗,体験に関する感性情報を独立した軸により,再現した感性コンテキスト分類モデルの構築

設備を始めとした環境情報だけでなく訪問時の感性を対象とした感性コンテキスト分類方式を設定する事により, 店舗毎に異なる特徴を抽出し,より多角的な情報に基づく旅程生成を可能とした.

3, 店舗の環境,感性情報と時空間情報を統合させる事で最適な旅程提案が可能な計算モデルの構築

1,2 の分類方式を用いた特徴と時間,空間に関するマルチモーダルな情報の統合方式の実装により,ユーザー情報に基づく旅程生成を可能にした.

以上 3 点の結果より, 本提案方式は,旅行者の時間的・感性的条件を満たす旅程生成を行う「Personalized Traveling Maas (Mobility as a Service)」として,旅行者に新たな移動の価値を提示できる可能性を示した.

また,本論文の提案方式は様々な地区や観光発展に協力する企業や店舗の情報を含んだ DB を構築し,提案することで, 観光や日常の移動を行うユーザーの移動価値を向上させ,地域活性化等の支援が可能な Mobility as a Service(MaaS)サービスとしての実現が可能であると考える.

6.2 今後の展望

本項では本論文の内容を踏まえ,本提案方式の一般化に向けた課題と機能面の展望について述べる.

本提案方式における課題の一つとしては, 店舗の特徴を表す感性分類の軸が現在 7 つである点から同一の感性情報を持つ店舗の発生の可能性が高くこれにより距離計算時,駅からの距離,DB 内の順位,所要移動時間に基づき,ユーザー本来の感性情報に関係なく表示される優先度の高い店舗が発生してしまう点が挙げられ,本提案方式の一般化における膨大な店舗 DB の格納に向け,多様な旅程の提案を実現させるには,本章①で後述する入力時の予定に関するニーズ項目の追加を始めとした店舗との類似度スコアに関する特徴量の追加の必要性が考えられた.

また,本提案方式の一般化に向け, 今後の展望として実装中に検討した追加点としては以下 2 点が挙げられる.

① 予定情報に関するニーズの入力項目追加

本方式の, 提案方式におけるユーザー情報の入力項目にて, 「早めに到着したい」「ゆっくり探索したい」といった予定に関する感性的ニーズ情報の選択項目の追加を行う, 入力に応じ,該当する現在,予定駅の Station_ID を持つ店舗の類似度スコアに加算を行う事で, より詳細なユーザー情報に基づく旅程提案が可能になると考える.

② ”移動体験”提案の実現

本稿の実装において, 公共交通機関の利用を想定した移動時間の取得による時間計量を行っているが,旅程における体験の提案による更なる体験価値向上を目的として,シェアサイクルや,モ ノレールといった,利用可能,ニーズに適合した, 移動手段を検出,それに基づく時間情報の提案を行う.

必要な工程としては,追加可能な移動手段の種類毎に可能な駅情報(例：駅周辺のシェアサイクルスポットの有無）と移動体験に関する感性情報を設定,図 7 のシステム構造における⑦の移動時間検出の際, ユーザーの現在地と目的駅に合致, 感性ニーズとの相関スコアが一定以上である場合に,他の移動手段の提案と移動時間に基づく旅程検出を行うといった流れである.

これらの課題の解決策や機能の実装により,更に多くの対象地域や店舗情報を基にした詳細なユーザー分析や最適な旅程提案による体験価値の向上が望めると考える.

参 考 文 献

[1] VilleLehmuskoski, SampoHietanen(2014)”Mobility as a Service - A Proposal for Action for the Public Administration, Case Helsinki “[p.84]

[2] 産業技術総合研究所”地域新 MaaS 創出推進事業での取り組み” https://www.meti.go.jp/shingikai/mono_info_service/smart_mobility_challenge/pdf/20220405_01_s01.pdf(Accessed :2023-12-1)

[3] 倉林修一,石橋直樹,清木康(2000)”モバイルコンピューティング環境におけるアクティブ型マルチ DB の実現方式” [p.463-470]

[4] 森 薫,倉林 修一,石橋 直樹,清木 康(2004) “ モバイルコンピューティング環境における ユーザー情報の動的計量による能動型情報配信方式 ” [p.8]

[5] 川島宏一郎,林康弘,清木康,三田哲也(2022) “移動経路における感性コンテキストに対応する移動・活動手段の組み合わせ形成システム - Emotional MaaS - ”[p.3]

[6] Barry Schwartz(2005) ”The Paradox of Choice: Why More Is Less”

[7] ICOONMONO https://icooon-mono.com/(Accessed : 2023-12-1)

[8] 食べログ https://tabelog.com/ (Accessed :2023-12-1)

[9] じゃらん net https://www.jalan.net/(Accessed :2023-12-1)

[10] GoogleMapshttps://www.google.co.jp/maps/(Accessed :2023-12-1)

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

ドメイン指向推薦 (1)

座長:黒川 茂莉(株式会社KDDI総合研究所)

コメンテータ:杉山 一成(大阪成蹊大学)

2024年3月1日(金) 10:00 〜 12:10 T3-B (オンライン（Zoom Events))

11:45 〜 12:10

[T3-B-7-05]【技術報告】サイバーエージェントメディア事業部における推薦システム

松月大輔 (株式会社サイバーエージェント)

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年3月1日(金) 10:00 ~ 12:10 | T3-C オンライン（Zoom Events)

ファッション・コスメ

座長:有次 正義(熊本大学)
コメンテータ:神門 典子(国立情報学研究所)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T3-C-7-01] レビュー分析に基づくヘアケアアイテムの香り可視化手法の提案

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

10:20 ~ 10:35
[T3-C-7-02] 投稿コーディネート画像を用いたLoRAモデルの作成と分析

*大江 優真[1]、荘司 慶行[1] (1. 静岡大学)

10:35 ~ 11:00
[T3-C-7-03] コスメレビューデータを用いたメイクパターン特性の可視化

*福本 真以子[1]、牛尼 剛聡[2]、角谷 和俊[1] (1. 関西学院大学、2. 九州大学)

11:00 ~ 11:25
[T3-C-7-04] Personalized Fashion Recommendation with Image Attributes and Aesthetics Assessment

*Chen Chongxian[1]、Fan Mo[1]、Fan Xin[1]、山名 早人[1] (1. 早稲田大学　山名研究室)

レビュー分析に基づくヘアケアアイテムの香り可視化手法の提案

横山真由子† 山口 史弥†† 上田真由美†††,†††† 中島 伸介†

† 京都産業大学 情報理工学部 〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学 先端情報学研究科 〒 603–8555 京都府京都市北区上賀茂本山
††† 流通科学大学 経済学部 〒 651–2188 兵庫県神戸市西区学園西町 3-1
†††† 大阪大学 サイバーメディアセンター 〒 560–0043 大阪府豊中市待兼山町 1 番 32 号
E-mail: †{g2154596,nakajima}@cc.kyoto-su.ac.jp, ††i2286211@cc.kyoto-su.ac.jp,
†††Mayumi_Ueda@red.umds.ac.jp

あらまし 近年，商品に対するレビュー情報が掲載されている Web サイトは数多く存在している．ユーザは複数のレビューから商品の情報を確認することでその商品についての理解を深め，購入の判断材料としている．しかし，香りのように実体験から得られる情報が大きい評価項目については，一部のレビューからは容易に把握し難いといった問題点がある．中でもシャンプーは多様な香りのものが販売されているが，メーカーごとに香りに対する表現は異なるため，ユーザが求める香りを持つシャンプーを探すことは困難である．本稿では，既存のレビューサイトから得られたヘアケアアイテムに対する実際のレビューデータを用いた，ヘアケアアイテムの香りのカテゴリと強さ，および香りの特徴の可視化手法について報告する．

キーワード 香り可視化，ヘアケアアイテム，レビュー分析

1 はじめに

近年，商品に対するレビュー情報を閲覧可能な Web サイトが数多く運営されており，レビュー情報を商品購入の参考にするということが一般的になりつつある．特に我々が着目するヘアケアアイテムは，直接髪や地肌に使用するものであり，自分に合わないアイテムの購入を避けたいという傾向が他の商品に比べて顕著であるため，レビューサイトの利用意義は大きい．ヘアケアアイテムに関するレビューサイトとしては，@cosme[1] などが存在し，多くのユーザに利用されている．

ヘアケアアイテム購入時における重要な判断材料として，「洗浄力」や「スタイリング」に加えて「香り」が挙げられるが，「洗浄力」や「スタイリング」に関する特徴が，ある程度明確に記述可能であるのに対して，「香り」に対する表現はメーカーによって異なることも含めて，商品説明からその香りを想像することは困難である．これは，香りのように実体験から得られる情報が大きい評価項目は，各個人の好みや環境，知識によって捉え方が大きく異なるためである．そこで，ユーザの実体験を踏まえたレビュー分析に基づき，ヘアケアアイテムの「香り」に関する情報を可視化して，購入を検討しているユーザに提示することの意義は大きい．

そこで本研究では，ヘアケアアイテムに対するレビュー分析を行い，香りの種類や強さの可視化を実現することで，ユーザの香りへの理解支援を行うことを目的としている．

香りの分類は，官能分析や成分に基づいて長年取り組まれてきたが，未だ標準となる分類方法は明確化されていない．本研究で用いる香りカテゴリは，純粋な香料の分類に一般的に用いられる基準用語に基づいて「フローラル，フルーティ，シトラス，ウッディ，オリエンタル，ムスク，ハーブ，グリーン」の 8 つに定めた [1]．

本稿の構成は以下の通りである．2 章では関連研究を紹介する．3 章では本研究で用いるレビューデータの分析結果と提案手法について詳細を説明する．4 章では構築した評価表現辞書の精度評価実験について説明する．最後に 5 章でまとめと今後について述べる．

2 関連研究

様々な商品のレビュー情報を掲載しているサイトとして，Amazon.com[2]や価格.com[3]が有名である．我々が研究対象とするヘアケアアイテムや美容に関しては，@cosme が有名である．@cosme では，様々なブランドのヘアケアアイテムが掲載されており，ヘアケアアイテム別のブランド比較が容易にできる．ユーザは，レビュー投稿時に総合評価として 8 段階で評価している．レビューには効果・機能タグを複数設定することができる．効果・機能タグには，「地肌ケア」，「さらさら」，「香りがよい」など 19 のタグが存在し，タグが付与されたレビューのみを絞り込んで検索することが可能である．さらに，ユーザはキーワードとして独自のタグを 5 つまで追加することができる．1 つのタグは 32 文字まで入力できる．商品情報であるブランド名，カテゴリ名，および「無香料」，「無着色」といったタグは自動で付与される．

レビュー情報を扱うサイトが広く普及したことにより，レビューを分析する研究が多く行われている．我々の先行研究に

1：@cosme,https://www.cosme.net/

2：Amazon,https://www.amazon.com/

3：価格.com,https://kakaku.com/

おいても，Ueda らは，化粧水に関するレビュー分析によりコスメアイテムに対する評価表現辞書を構築し，それを用いた評価項目別レビューの自動スコアリング方式の開発を行っている [2]．自動スコアリングにはコスメアイテムのカテゴリごとの評価表現辞書の構築が必要となる．全てのカテゴリの評価表現辞書を構築するのは手間がかかるため，評価表現辞書の効率的な構築手法の提案を行っている．

また，レビューによるアイテムの可視化に関する研究として，白田らは， マスカラのレビュー分析より，概念グラフによるグラフ生成を行っている [3]. グラフは，レビューから名詞，動詞，形容詞を抽出した結果に対し RIDF を用いて重みを計算し，単語の共起関係と頻度情報に基づいて生成している．この取り組みは，グラフ可視化ソフトウェアである gephi を拡張した Consumer Bebavior Analyzer を用いており，コミュニティの色分け表示及びグラフのレイアウト表示などの可視化を行っている．

香りの分類方法に関する研究として，若田らは，SD 法を用いた香りの印象による分類を行っている [4]．印象評価結果より，120 種類の香りは，18 の形容詞対を用いた印象評価を基に 11 グループに分類された．香りの印象には,「甘い」,「あたたかい」に関連する因子，評価を表す “pleasantness” に関する因子,「キツさ」を表す “intensity” に関する因子の 3 因子が示されている．Miguel らは，純粋な香料の分類だけでなく，その蒸発過程と臭気強度に基づいた香水の分類方法を提案している [1]. 提案手法により，香料の経験的な分類の恣意性を減らし，香水の香気持続性を予測することが期待されている．実験結果より，提案手法は，エッセンシャルオイルと香水の分類を正しく行えることが示されている．

香りの表現用語選定や可視化に関する研究として，櫻井は，化粧水の香りにおいて共感覚的表現による評価を求め，嗅覚と他 4 つの感覚との関連性を明らかにしている [5]．共感覚とは，何かの感覚刺激を，本来の感覚刺激とは別の感覚刺激が連動して感受する感覚のことを指す．主成分分析結果から，嗅覚は触覚と強い結びつきがあることが指摘されている．春原は，香りと色の強い共感性を指摘している [6]．官能評価実験の結果から，ヘアケアアイテムには濃いめのピンクが多く選ばれ，ユーザは製品カテゴリにより異なる香りコンセプトを抱くことが示唆されている．大野は，紅茶の香りや色に関する表現用語を選定し，キャラクターホイールを作成することで紅茶の特徴の可視化を行っている [7]．杉山らは，香りのイメージを色とカタチで表現し，香り可視化を行っている [8]．香りの種類によって，色表現が有効であるものとカタチ表現が有効であるものが存在し，色とカタチ両方を提示することで，より正確に香りイメージが伝達できることが示されている．

香りの表現用語として，一般的に特性的用語，香料用語，感覚用語が存在する [9]．特性的用語は「花のような」,「石鹸のような」などの表現であり，香料用語は「フローラルグリーン」,「オリエンタル」などの表現のことを指す [10]．感覚用語は感性印象に基づいた「甘い」,「白っぽい」などの表現のことを指す [11]．特性的用語や香料用語は専門家の間で多く用いられるものであり，香りと用語の間を結ぶ何らかの経験が必要となることからも非専門家には理解し難いが，これらに比べ，感覚用語については共通の感覚的経験により非専門家でも理解しやすいとされている [12].

以上のように，レビュー分析やレビューに基づくアイテムの可視化，香りの分類・表現用語選定・可視化に着目した取り組みは行われている．しかし，レビューに基づくヘアケアアイテム推薦のための香り可視化に関する取り組みは十分に行われていない．このような背景から，本研究では，香りによるヘアケアアイテム推薦を目的とした，レビュー分析に基づく香りの可視化手法を提案する．

3 ヘアケアアイテムの香り可視化手法

本章では，ヘアケアアイテム評価表現辞書を用いた香りカテゴリ別レビュー自動スコアリング方式について述べる．3.1 節で提案手法の概要を述べ，3.2 節で実データの概要と傾向の把握，3.3 節で評価表現辞書の構築手法，3.4 節で評価表現辞書に基づく自動スコアリングシステムについて説明する．3.5 節でワードクラウドを用いたヘアケアアイテムの香り可視化手法について示す.

3.1 概　　要

本研究では，ヘアケアアイテムに対するレビュー分析を行うことで，ヘアケアアイテムの香りに関する評価表現辞書を作成し，これを用いた香りカテゴリ別レビュー自動スコアリング方式の開発を行うことを目的としている．そのため，辞書の構築はレビュー内のテキスト情報およびその評価値を分析し，高頻度で出現する香りに関する表現を各アイテムに対する評価表現として辞書に登録する．また，構築した評価表現辞書はキーワードの共起に基づくため,「香りがすごく強い」と「すごく香りが強い」といった表現を同一のものとして扱うことが可能である．

3.2 実データの分析結果

本研究で用いるレビューデータの分析によって得られた結果について述べる．評価表現辞書の構築にあたり，ヘアケアアイテムに対して投稿された実データの分析を実施した．使用したデータは，@cosme の「シャンプー・コンディショナー」カテゴリの 1,899 件の商品に対して直近 2 年間（2021/11/1〜2023/10/31）に投稿された 71,660 件のレビューである．

表 1 に，レビュー内の頻出単語を示す．これは，レビューに対し形態素解析を行い，名詞のみを抽出した結果からストップワードを省いたものである．順序は出現回数の降順となっている.「香り」の出現回数が 73319 回と最も多いことや,「成分」より「匂い」の出現回数が多いことから，レビュー内にアイテムの香りに関する言及が多いと言える．そのため，ユーザにとってシャンプー・コンディショナーにおける香りという項目は泡立ちや成分と同等に重要視されていると考えられる．

表 1 に示される単語を含むレビューは，他人からの評価（以下,「他人軸」と呼ぶ）に関して記載されているものが見られ

表 1　レビュー内の香りに関する頻出単語結果抜粋

名詞	出現回数
香り	73319
シャンプー	66056
髪	64230
トリートメント	32269
購入	28150
使用	27743
頭皮	22100
髪の毛	11922
泡	11656
好き	11638
匂い	11390
洗い	11170
ケア	10254
質	9402
成分	9086

表 2　レビュー内の客観的評価に関する頻出単語結果抜粋

名詞	出現回数
家族	1194
友人	526
ユニセックス	476
友達	448
仕事	378
会社	204
職場	137
彼氏	73

た．例えば,「友人からいい香りだと言われた」,「無香料のため職場でも使える」といったものである．したがって，ヘアケアアイテムの香りにおいて他人軸が重要であると考え，レビュー分析を行う．表 2 に，レビューに対し形態素解析を行い，他人軸に関する単語を抜粋した結果を示す．抜粋する単語は，出現回数の降順で名詞のみとしている．表 2 より，他人軸に関する単語はレビュー内に十分に出現することが確認できた．そのため，ヘアケアアイテムの香りにおいて，他人軸が重要であると言える．

他人軸に用いられる表現方法を分析するため，表 2 を参考に単語を 12 個定め，いずれかの単語が含まれるレビューを抽出した．抽出したレビュー内において，他人軸が含まれているかどうか，含まれていればどのような表現方法が用いられているかを確認した．用いた単語は「家族」,「友人」,「職場」,「褒められた」といった単語である．他人軸に用いられる表現方法として,「友人」が含まれるレビューからは「友人にいい香りと言われた」といった表現方法が確認できた．しかし，そのほかに「友人から勧められた」,「友人からプレゼントされた」のような意味で用いられているものが多く存在した．これらは他人からの評価ではないため，他人軸には含まれない．他に他人軸の表現方法として,「自分は好きな香りであるが友人からは嫌いと言われた」や，その逆で,「自分は嫌いだが周りからの香りの評価は良かった」という表現や,「ユニセックスで使える香りである」などの表現が確認できた．

さらに，香りに用いられる表現方法を分析するために,「香り」という単語が含まれるレビューから 350 件をランダムに選択し，形態素解析を行った．形態素解析では名詞，動詞，形容詞，副詞のみを抽出した結果から，どのような表現方法が用いられているか確認した．香りに対する表現方法として,「‘香り’, ‘優しい’, ‘め’, ‘強い’, ‘匂い’, ‘ない’」や「‘シャンプー’, ‘中’, ‘匂い’, ‘強い’, ‘感じる’」といった表現が存在した．このように「強い」と「匂い」の順番が逆である場合が確認できる．そのため，3.3 節，3.4 節で述べる，キーワードの共起に基づく評価表現辞書による香りカテゴリ別レビュー自動スコアリング方式が有効であると判断できる．

3.3　ヘアケアアイテムに関する評価表現辞書の構築手法

本節では，ヘアケアアイテムに関する評価表現辞書の構築について述べる．

ヘアケアアイテムのレビューは一般のユーザによって投稿されるため，香りの捉え方はレビュー投稿者の嗜好や環境に依存しやすく，表現方法は様々である．そこで，本研究では実際にレビュー投稿サイトに投稿されたレビューから香りにまつわる特徴的な表現を収集し，香りの強さに関して 0〜7 の間で評価値を付与して辞書に登録することで評価表現辞書を構築する．評価値は 0 ならば無臭，7 ならば非常に強い香りがするということを示す．評価表現辞書の構築には@cosme に投稿されたレビューデータを使用している．図 1 に評価表現辞書の構築手順を示す．

(1)　既存のレビューサイトからレビューテキストを収集し，アイテムの香りを表現するフレーズを抽出

(2)　(1) で抽出したフレーズから，キーワード・特徴語・程度語のリストを作成

(3)　(2) 各リストを用いてキーワード・特徴語・程度語・否定の 4 つの語の組み合わせを評価表現として作成

(4)　作成した評価表現に対して人手により評価値を付与

なお，各種ヘアケアアイテムのうち，本研究ではシャンプー・コンディショナーに絞って評価表現辞書の構築に取り組んだ．多くのユーザが使用するアイテムであるため，シャンプー・コンディショナーを対象とした．

3.2 節で抽出した香りに用いられる表現方法に対し，20 代 4 名（男女各 2 名）で評価値を付与した．4 名の評価値の平均値を評価表現辞書に登録した．

3.4　評価表現辞書に基づく自動スコアリング方式

我々の先行研究である，評価表現辞書を用いた各アイテムの香りカテゴリ別スコアを算出する自動スコアリング方式について述べる．

図 2 に評価表現辞書を用いたヘアケアアイテムに対するレビューの採点方法を示す．このシステムは，ヘアケアアイテムに対するレビューのテキストを読み込み，レビュー内に存在する評価表現を特定し，その評価表現を用いてレビューの採点を行う．その後，評価表現辞書に基づいて各評価表現に評価値を

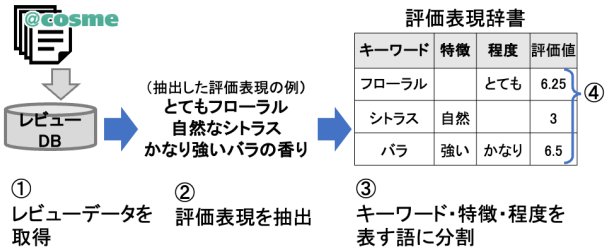


図 1　評価表現辞書の構築手順

付与する．1 つの評価項目について複数の評価表現が含まれている場合には、それぞれの評価表現に付与された評価値の平均値をその香りカテゴリの香りの強さに対する評価値とする．図 2 では，レビュー内の「上品」の部分に「4」,「程よく」の部分に「3」の評価値が付与されるため，このレビューの香りカテゴリ「フローラル」に対する評価値は，(4+3)/2 =「3.5」となる．

先行研究に取り上げたシステムは，1 つのレビューに評価値を付与するものである．しかし，本研究では各アイテムに対する香りカテゴリ別評価値を求める必要がある．そこで，自動スコアリングシステムを用いてレビューごとの評価値を算出し，同じアイテムに対する評価値の平均値を香りカテゴリごとに算出することで，各アイテムの香りカテゴリ別スコアとする．

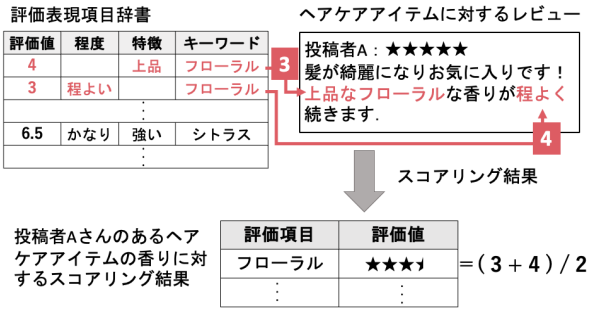


図 2　ヘアケアアイテムレビュー自動スコアリング方式の例

3.5　ワードクラウドを用いた香り可視化手法

本節では，ヘアケアアイテムの香りの強さと特徴の可視化手法について述べる．

3.4 節までに述べたヘアケアアイテムのスコアリング結果とともに，ワードクラウドを用いた香り可視化イメージを図 3 に示す．ヘアケアアイテムの香りの情報として，カテゴリとともに香りの強さを表示する．この香りの強さは，3.4 節で算出された評価値である．カテゴリの表示にあたっては，スコアリングされたカテゴリ全ての評価を表示する．加えて，8 つのカテゴリ以外の香り情報をワードクラウドで付加している．ワードクラウドを用いることで，香りカテゴリよりも詳細な香り表現が行えるようになる．

他人軸は図 3 の左側のようにタグの形で表示する．「職場にもつけていける」といった場所に関連するもの，「家族で使える」や「恋人にいい香りと言われた」など関係性に関連するもの，「万人受けする」などといった誰にでも受け入れられる香りであるかに関連するものを用いる．ユーザには香りの強さや特徴とともに他人軸を同時に提示することで，香りの理解支援に繋げる．



図 3　ワードクラウドを用いたヘアケアアイテムの香り可視化の例

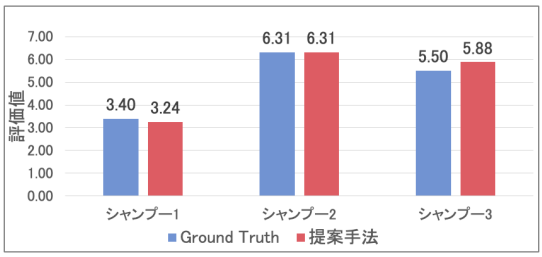


図 4　人手と評価表現辞書による香りの強さスコアリング結果

4　評価表現辞書の精度評価実験

本章では，3.3 節で構築した香りカテゴリ別評価表現辞書の精度評価実験について述べる．

4.1　実験概要

キーワードの共起に基づく香りカテゴリ別評価表現辞書の精度を確認するため，実データを用いて評価実験を行った．2 つのスコアリング手法によって得たそれぞれのスコアリング結果の平均値である平均絶対誤差（MAE）を用いて，構築した評価表現辞書の精度を評価した．

本実験は，以下に示す 2 つの手法を比較する．

(1)　辞書を使用しない人手によるスコアリング手法（Ground Truth）

(2)　キーワードの共起に基づく評価表現辞書を用いた自動スコアリング手法（提案手法）

(1) では，実験参加者に 3 つのヘアケアアイテムのレビューを提示し，アイテムごとに香りの強さを評価値付けしてもらうアンケートを実施した．ヘアケアアイテムは，「シャンプー・コンディショナー」カテゴリに含まれる 5 件以上のレビューが投稿されているものを選択した．アンケートでは，初めに，1 つの

アイテムにつきランダムに選択した 5 件のレビューを実験参加者に提示する．次に，レビューから読み取れるヘアケアアイテムの香りの強さを，1 章で事前に定めた 8 つの香りカテゴリごとに 0〜7 の範囲（0.5 刻み）で評価値付けしてもらう．実験参加者全体の平均値を Ground Truth とする．本実験では，8 つの香りカテゴリの内「フローラル」カテゴリの評価表現辞書の精度評価を実施した．

4.2 実験結果

ヘアケアアイテムごとの Ground Truth と提案手法によるスコアリング結果を図 4 に示す．Ground Truth の平均値は 5.07 であり，提案手法の平均値は 5.14 であった．MAE は 0.181 であった．

4.3 考　　察

結果より，評価値が 0〜7 という範囲に対して MAE は 0.181 であった．これは，提案手法によって，実験参加者が実際にレビューを見て作成した Ground Truth に非常に近いスコアリングが行えていることを確認できたこととなる．Ground Truth と完全な一致とならない点においては，評価表現辞書に登録している評価表現不足によるものだと考える．今後，さらに評価表現を追加し，8 つの香りカテゴリ全ての評価表現辞書の精度評価を実施することで自動スコアリングの精度向上を見込むことができる．

5 おわりに

本稿では，レビュー分析に基づくヘアケアアイテムの香り可視化手法について提案した．具体的には，評価表現辞書を用いた香りのカテゴリと強さの分析手法について提案するとともに，ワードクラウドによる香りの特徴の可視化について提案した．本手法により，ヘアケアアイテムにおける実体験から得られる情報である香りの可視化が可能となる．各アイテムの香りの特徴が明確になることで，ユーザは好みの商品の選択が容易になる．また，ヘアケアアイテムに対するレビューデータに対して，評価表現辞書の構築およびこれを用いた自動スコアリング実験を行い，提案手法の有効性を確認した．

今後は，評価表現辞書のさらなる拡充・調整を行い，スコアリング精度を向上させ，香りによるヘアケアアイテム推薦の実現へ向ける．

謝　　辞

本研究を遂行するにあたり，株式会社アイスタイル様にヘアケアアイテムに関するレビューデータをご提供いただいた．また，本研究の一部は，JSPS 科研費（課題番号：19K12243，20H04293，22K12281）および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] Teixeira, Miguel A., Oscar Rodriguez, and Alírio E. Rodrigues, “Perfumery radar: A predictive tool for perfume family classification,” Industrial & engineering chemistry research, Vol. 49, No. 22, pp. 11764-11777, 2010.
[2] Mayumi Ueda, Yuna Taniguchi, Da Li, Panote Siriaraya, Shinsuke Nakajima, “A Research on Constructing Evaluative Expression Dictionaries for Cosmetics Based on Word2Vec,” The 23rd International Conference on Information Integration and Web-based Applications & Services (iiWAS2021), pp. 84-90, November 2021.
[3] 白田由香利, 橋本隆子, 久保山哲二, インターネット上のレビューサイトにおける化粧品の評判分析, 学習院大学計算機センター年報, Vol. 33, pp. 2-7, 2013.
[4] 若田忠之, 齋藤美穂, 香りの分類における心理学的検討, 日本感性工学会論文誌, Vol. 13, No. 5, pp. 591-601, 2014.
[5] 櫻井広幸, 化粧水の香りの共感覚的表現—触覚語の役割—, 日本官能評価学会誌, Vol. 4, No. 1, pp. 45-51, 2000.
[6] 春原吉美, 香りと色の共感覚性, 日本官能評価学会誌, Vol. 14, No. 2-2, pp. 109-113, 2010.
[7] 大野敦子, 紅茶キャラクターホイールの作成と紅茶特徴の可視化, におい・かおり環境学会誌, Vol. 45, No.5, pp. 344-350, 2014.
[8] 杉山雄太, 赤澤智津子, 安藤昌也, 香りイメージの伝達方法に関する研究, 日本デザイン学会研究発表大会概要集, Vol. 62, pp. 259, 2015.
[9] 櫻井広幸, 神宮英夫, 香料の共感覚的表現, 日本官能評価学会誌, Vol. 1, No. 2, pp. 41-45, 1997.
[10] 高島靖弘, 感性品質の計測技術-香りの官能評価と心理・生理的効果, 品質, Vol. 25, No. 1, pp. 34-40, 1995.
[11] 竹澤智美ら, 感性印象に基づくエッセンシャルオイルのマップ構築, 第 24 回日本感性工学会大会 (JSKE2022), 2B-1-05, 1995.
[12] 神宮英夫, 化粧品の香り表現に関する研究, コスメトロジー研究報告, Vol. 7, pp. 130-137, 1999.

画像生成 AI はファッションセンスを解すか ?: 投稿コーディネート画像を用いた LoRA モデルの作成と分析

大江　優真†　　荘司　慶行†

† 静岡大学 情報学部　〒 430 – 0807 静岡県 浜松市 中区 城北
E-mail: †oe.yuma.21@shizuoka.ac.jp, ††shojiy@inf.shizuoka.ac.jp

あらまし　本論文では，ファッションアイテムが単独で映った画像を入力すると，高評価と低評価のそれぞれが付きそうなコーディネート画像を生成する手法を提案する．このような画像を多数生成し比較することで，人々が画像のどこを評価しているかを分析可能にし，画像生成 AI によるコーディネート推薦を可能にする．そのために，Stable Diffusion を SNS に投稿されたいいね数の上位と下位の画像でそれぞれ学習し，二つの LoRA（Low-Rank Adaptation）モデルを作成した．学習対象と学習件数の異なるモデルで生成された画像を，コーディネート，写真の質，ロケーションなどについて人手で評価する被験者実験を行った．実験結果から，学習件数を増やし過ぎると生成画像の質が下がる傾向と，いいね数下位の写真で学習したモデルのほうが逆に好評である傾向が示された．

キーワード　Stable Diffusion, LoRA モデル, 潜在拡散モデル

1　はじめに

ほとんどの人は，少なくとも 1 日 1 回は，衣類の組み合わせを考えているはずである．例えば，毎朝家を出る際には，部屋着から外出着に着替えるだろう．人々は，毎日，天気や目的地，趣味に合わせて，ファッションアイテムを組み合わせ，コーディネートを完成させる．特に近年では，Instagram や WEAR[1] などの SNS（Social Networking Service）の普及も後押しして，SNS 上で高い評価を得られるような,「映える」ファッションをしたがる人も多くいると考えられる．

実際に，自身をコーディネートする際には，多くの場合，ユーザは自身の持っているファッションアイテムを組み合わせることを考える．その際にセンスの良いコーディネートのヒントを得る 1 つの手段として，Web や SNS を用いたコーディネート検索が挙げられる．例えば，持っているチェックシャツのコーディネートを知りたい場合は，SNS で「#チェックシャツ」などのハッシュタグの付いた投稿を検索すればよい．一方で，このような検索では，検索漏れが発生したり，検索ノイズが含まれる可能性が高い．SNS にコーディネート画像を投稿する利用者が，必ずしも，すべてのアイテムにハッシュタグを設定するとは限らない．そのため，画像にチェックシャツが映っていたとしても,「#チェックシャツ」のタグが付いていない場合も多い．加えて，たとえばチェックシャツの中にも，ギンガムチェックのシャツやタータンチェックといった，柄の異なるシャツが存在する．この場合，自分の持っているシャツの柄と異なるものが表示された場合，コーディネートを考える上では，参考にならない．そのため，実際に着たいアイテムと同じ商品が映っており，なおかつ好評なコーディネート画像を検索することは困難である．

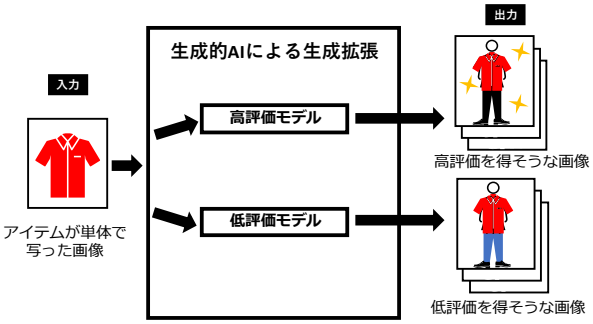


図 1　提案手法の入出力例．ファッションアイテムを入力すると，2 つのモデルが，それぞれアイテムを含む着用画像を出力する．

そこで本論文では，あるファッションアイテムが映った画像を入力すると，そのファッションアイテムを着用したコーディネート画像を生成する手法を提案する．そのために，図 1 のように,「高評価モデル」と「低評価モデル」という，2 つの画像生成モデルを作成する．それぞれのモデルにアイテム画像を入力すると，高評価を得そうな着用画像と，高評価を得られなさそうな着用画像を，それぞれ生成する．

近年流行しつつある画像生成系 AI は，独自の絵柄や，スタイルを，LoRA ファインチューニングによってモデルに学習させることができる．そこで，画像生成 AI である Stable Diffusion[2] に，SNS 上の着用画像を用いた追加学習を施し，

- SNS 内で多くの高評価を得た写真のみを学習したモデル，
- SNS 内であまり高評価を得ていない写真のみを学習したモデル

の，2 つの LoRA（Low-Rank Adaptation）モデルを作成した．モデルの学習のために，ファッション投稿サイトである WEAR

1：WEAR
https://wear.jp/

2：Stability AI「Stable Diffusion」
https://ja.stability.ai/stable-diffusion

に投稿されたコーディネート画像のうち,「いいね」の数でランキングした中で，上位と下位の画像を用いた．このように学習したモデルに，ファッションアイテムが単独で写った画像を入力として与え，生成拡張（画像に写っていない部分を想像で補わせる）させると，そのアイテムを着用した人物の全体像が出力される．

本論文の主要な貢献は，SNS 上で高評価と低評価のそれぞれが付き得る投稿を画像生成 AI に学習させたことである．画像生成 AI は，入力に用いたアイテムを用いたコーディネートに加え，画像内のロケーションや構図を含めて学習している．出力した画像を用いた被験者実験を通して，SNS 上で評価の異なる画像を学習したモデルの違いを明らかにする．

以下に，本論文の構成について述べる．本章では，研究の背景と目的について説明した．第 2 章では，類似した研究分野として，コーディネート推薦およびコーディネート検索についてそれぞれ紹介して論じる．第 3 章では，本研究で提案する手法を述べる．第 4 章では，被験者実験とその結果について述べる．第 5 章では，実験を通して得られた結果に関して考察を行い，第 6 章でまとめと今後の展望について述べる．

2 関 連 研 究

本研究は，ファッションアイテムの画像を入力として受け取り，コーディネート画像を生成する．コーディネートを画像によって提案する点において，コーディネート推薦の分野と深く関係する．また，着用画像の生成においては，仮想試着の分野とも関係する．そのため，第 2.1 節ではコーディネート推薦について，第 2.2 節では仮想試着について既存の関連研究を説明し，本研究の位置づけを示す．

2.1 コーディネート推薦

個人へのコーディネート推薦（Outfit Recommendations）に関する研究は盛んにおこなわれている．一つの手法として，ファッションアイテム間の適合性を考慮した推薦がおこなわれている．例として Ding ら [1] は，アイテム同士の組み合わせをアイテムのメタデータから推薦する手法を提案している．ほかに，Han ら [2] は，ニューラルネットワークである LSTM を用いてアイテム同士の組み合わせを推薦する手法を提案している．これらのアプローチでは，主にアイテムの情報としてテキスト情報を用いており，本研究ではこれらのような推薦を，画像生成を通して実現することを目標としている．

画像生成をファッション推薦に用いた例として，Chong ら [3] は，着用モデルの体形を考慮し，着用画像内のマスク箇所に適切なアイテムを推薦するフレームワークを提案している．本研究はこの研究に近いが，本研究では，ある 1 件のアイテムを入力として与えると，複数の着用画像を生成することを目的としている．

これらのどの研究においても，出力される「コーディネート」とは，ファッションアイテムの画像集合である．そのため，実際に人間が着用したイメージを連想することが難しい．本研究で提案するコーディネートはモデルが着用したものであり，加えてロケーションや構図を含めて「SNS で高評価と低評価のそれぞれを取り得る」画像の生成をおこなう．

2.2 仮 想 試 着

近年では，アルゴリズムによる画像編集技術を活用した，仮想試着の研究が盛んである．仮想試着とは，モデルの着用画像に対し，服を着せ替える手法である．Xintong ら [4] は，着用モデルの体に合わせて服の画像を変形させることにより，自然な着用画像の生成を実現している．Shizhan ら [5] は，ある着用画像に対し，テキスト入力をもとにして別の服にモデルを着せ替える手法を提案している．

仮想試着では，実際にモデルが着用している画像を生成し，着用イメージを得ることができる．しかし，他のアイテムとの組み合わせは個人によっておこなうものであるため，あるファッションアイテムを用いたコーディネートまでは知ることができない．本研究では，コーディネートやロケーションを加味した着用画像を生成する．また，画像生成の既存研究では，GAN（Generative Adversarial Networks）が採択されている．本研究では，画像生成 AI である Stable Diffusion を用いた着用画像生成を提案している．

3 提 案 手 法

本節では，コーディネート画像生成をおこなうための手法について説明する．本論文で提案するアルゴリズムは，1 枚のファッションアイテムの写った画像を入力とする．2 つの異なるデータセットを用いて学習された 2 つのモデルが，入力画像に写ったアイテムの，高評価をたくさん得られそうな着用画像と，高評価があまり得られなさそうな着用画像を，それぞれ生成する．

このような機能を実現するために，汎用的な画像生成モデルである Stable Diffusion [3] を用いた．Stable Diffusion は，多くの画像を用いて学習された汎用画像モデルであり，テキスト形式のプロンプトから画像を生成したり，画像を入力して変換させたり，さまざまなタスクに利用可能である．本研究では，Stable Diffusion の生成拡張機能を活用した．

Stable Diffusion を，実際に SNS で高評価を得そうな画像，得なさそうな画像を生成できるようにするために，ファッション投稿 SNS のデータで，ファインチューニングを行った．実際のファインチューニングでは，計算量の関係から，LoRA によって学習を行った．また，実際の生成には，ControlNet と呼ばれる Stable Diffusion の拡張機能を用いた [6].

3.1 データセットの作成

はじめに，ファッションアイテム着用画像をソーシャルメディアから収集し，学習用のデータセットを作成した．高評価が得られそうなコーディネートと，得られなさそうなコーディネー

3 : GitHub「Stability-AI stablediffusion」
https://github.com/Stability-AI/stablediffusion

ト画像を生成できるように，SNS 上で「いいね」の数が多いものと少ないものを，それぞれ収集した．

ファッションアイテムの着用画像を収集するために，本研究では，ファッション SNS である WEAR から画像とメタデータを収集した．具体的には，WEAR で着用画像単位で一覧表示できる「コーディネート一覧」のランキングを用いた．コーディネート一覧では，WEAR が独自に算出した人気度による「人気順」，投稿日時による「新着順」の，上位 9,990 件について閲覧できる．また，ランキングは春夏秋冬，それぞれの季節に対して分けられている．

そこで，データセットの作成のために，はじめに 4 つの季節に対する人気順，新着順それぞれの上位 9,990 件のコーディネートを記録した．次に，季節ごとに人気順と新着順のコーディネートを 1 つのリストに並べ,「いいね」の数で並び替えた．そして,「いいね」の数の上位 100 件，下位 100 件のコーディネートについて，実際に画像をダウンロードした．こうすることで，モデルに SNS 上で高評価を得そうなコーディネートと，そうでないコーディネートを学習させることができると考えた．

3.2 LoRA モデルの作成

LoRA とは，少ない計算量で DNN を追加学習させる手法である [7]．本研究では，SNS 上に投稿されたコーディネート画像のうち，高評価と低評価をそれぞれ獲得したものを用いて LoRA モデルを作成した．これは，ユーザからのいいね数を多く獲得した投稿は，よりセンスのあるものであるという考えに基づいている．高評価を得た投稿で学習したモデルを「高評価モデル」，低評価を得た投稿で学習したモデルと「低評価モデル」とする．学習時には，ベースモデルとして stable-diffusion-v1-5[4] を用いた．

3.3 コーディネート画像の生成

作成した LoRA モデルを Stable Diffusion Web UI で用いることで，コーディネート画像の生成をおこなう．この際に，単純に拡張生成を行うと，毎回ポーズが変わってしまったり，入力画像そのものに人物を当てはめたような画像が生成されがちであった．そこで，txt2img の拡張機能として ControlNet[5] を用いた．

本研究では，ControlNet の機能の一つである「Reference Only」を使うと，Reference Only は，画像とテキストプロンプトの両方を入力可能である．そこで，入力画像のアイテムを保ったまま，その周辺に人物の残りのパーツを生成するよう，プロンプトを作成し，それを固定した状態で入力画像だけ変更するようにした．実際のプロンプトは，

プロンプト:
(masterpiece:1.2),
(best quality:1.2),
woman,
winterfashion,
< LoRA:1.1 >
ネガティブプロンプト:
(worst quality:1.4),
(low quality:1.4)

とした．この際，プロンプト中の「*winterfashion*」の部分はトリガープロンプトと呼ばれるものであり，学習した LoRA モデルを呼び出すためのものである．今回の実験で用いたモデルでは，高評価を得そうな春夏秋冬それぞれのモデル，高評価を得なさそうな春夏秋冬それぞれのモデルの，8 つのモデルがある．それぞれにモデル名をつけておき，プロンプト中に含めた．

4 評　　価

作成した LoRA モデルが SNS におけるセンスを解するか，被験者実験を通して明らかにする．

4.1 実験タスク

本研究では,「いいね」数の多い画像と，少ない画像から学習させた LoRA モデルを適用し，任意のファッションアイテム画像を与え，そのアイテムを含む着用画像を生成した．また，比較のために，LoRA モデルを使用せず，ファッションアイテム画像とプロンプトのみを与えて画像を生成した．この生成画像が，実際にファッションセンスや，高評価を取りそうかどうかを反映しているかを確認するために，被験者にアンケートを取った．なお，本実験の参加者は，アパレルショップに勤務する販売スタッフである．ファッションに精通している販売スタッフに評価をおこなってもらうことで，生成された着用画像がファッションセンスを表現できているのかを明らかにする．

実際の手順として，まず，10 件のアイテムについて，高評価モデルと低評価モデルで，それぞれ 3 件ずつ着用画像を生成した．こうして生成された 150 件の画像について，ランダムに並び替え，被験者に提示した．実験参加者は，各画像に対して，その画像の印象について，いくつかの質問項目に答えた．実際の質問項目を表 1 に示す．質問項目には，画像そのもののクオリティを問うもの，コーディネートの質を問うもの，各着用アイテムの質を問うものを含めた．各項目について,「とても良い」から「とても悪い」の 5 段階のリッカート尺度で答えさせた．

4.2 データセット

実際の学習と評価実験のために，WEAR に投稿されたデータから学習用データセットを，ZOZOTOWN[6]から評価用データセットをそれぞれ作成した．今回の実験では，特にファッションのアイテムのバリエーションが多く，フォルムの変化が大きい女性の冬のファッションを対象とした．そのため，データセットは，サービス内で WOMEN'S 商品に絞り込みをかけたうえ

4：Hugging Face「runwayml / stable-diffusion-v1-5」
https://huggingface.co/runwayml/stable-diffusion-v1-5
5：GitHub「lllyasviel/ControlNet」
https://github.com/lllyasviel/ControlNet
6：ゾゾタウン
https://zozo.jp/

表 1　質問項目．150 枚の画像一つ一つに対して，各実験参加者は，12 問の質問に 5 段階で答えた．

評価観点	質問項目
コーディネート	ファッションアイテムの組み合わせは適切か
	コーディネートはおしゃれか
アイテム	それぞれのアイテムはおしゃれか
	それぞれのアイテムは価値が高そうか
	着用人物は見た目がよいか
	着用人物は体形がよいか
	写真の撮影場所，背景はおしゃれか
	画像に写った小物類はおしゃれか
画像	画像には正しくアイテムが写っているか
	写真の構図はきれいだと感じるか
	写真全体の色づかいはきれいだと感じるか
	写真に不自然なところはないか

で作成した．

第 3.1 節で述べたように，学習用のデータセットを，WEAR のランキング機能をもとに作成した．これにより，春夏秋冬の各季節について，「いいね」の数で上位 500 件，下位 500 件のコーディネート画像からなる，4,000 件の画像からなるデータセットを作成した．

次に，評価用の入力画像のデータセットとして，ZOZOTOWN の商品画像を収集した．ZOZOTOWN の商品から，アウター，トップス，ボトムスカテゴリの商品を，合計 10 枚収集した．実際に評価用データとして用いるために，背景を除去し，画像内の位置を調整した．具体的には，画像内に含まれるファッションアイテム以外の背景を除去するために，Rembg[7]を用いた．そして，真っ白な画像内に，抽出されたファッションアイテムを人手で配置した．この際，アイテムがトップスであれば画像の上側に，ボトムスであれば画像の下側に配置した．

4.3 実　　装

Stable DIffusion Web UI 上で動作する LoRA モデルを作成した．はじめに，Stable Diffusion の学習では画像ごとにテキストのメタデータが必要なので，dataset-tag-editor-standalone[8]でキャプションを生成した．生成された候補の中から，不必要なキャプションを除去して，それぞれの画像に紐づけた．実際に残したキャプションは，「1girl, full body, looking at viewer, solo」であり，データセット中の画像はこれらのキャプションのうちいくつかと紐づけられるようにした．

次に，kohya's GUI[9]を使用して LoRA モデルを作成した．学習時のパラメータ設定を表 2 に示す．学習時の画像枚数が生成結果に与える影響について調査するために，いいね数が上位下位それぞれ 100 枚の画像で学習したモデルと，500 枚で学習したモデルを作成した．それぞれのモデルを「高評価モデル 100」,「低評価モデル 100」,「高評価モデル 500」,「低評価モデル 500」とする．他のパラメータに関しては初期設定通りとした．なお，これらのパラメータは実際の計算機資源に合わせて，経験的に設定したものである．

表 2　各 LoRA モデル学習時のハイパーパラメータ

ハイパーパラメータ	値
バッチサイズ	2
エポック数	20
繰り返し回数	5
総ステップ数	2,000



図 2　実際の出力例．

こうして学習した LoRA モデルは，一般的なモデルと同様に，Stable Diffusion の各タスクに利用できる．今回の実験では，ControlNet の「Reference Only」機能で，学習済み LoRA モデルに入力アイテム画像とテキストプロンプトを与えることで，着用画像を生成できるようにした．作成した二つの LoRA モデルを使用し，Stable Diffusion で生成したコーディネート画像の出力例を図 2 に示す．

4.4 実 験 結 果

本節では，実際の実験結果と，実験時に生成された画像の出力例を記す．実験結果を以下の表 3 に示す．各質問への回答は，全 10 種類のアイテム画像による生成画像への評価及び，全 6 人の被験者による評価の平均値である．表を見ると，着用画像内のコーディネートやアイテムの「おしゃれさ」は，高評価モデルに比べて，低評価モデルの方が高い評価を得ていることがわかる．学習時の画像枚数を増やしたモデルでは，高評価モデルと低評価モデル間の評価の差が大きくなった．しかし，画像枚数が 100 枚のモデルに比べてより不自然な画像を生成するようになった．

5 考　　察

評価実験で得られたアンケートの結果について，考察する．はじめに，生成された着用画像の「おしゃれさ」に関する評価について議論する．「コーディネートはおしゃれか」という質問に対して，学習時の画像枚数に関係なく，高評価モデルによる生成結果に比べて，低評価モデルによる生成結果の方が高い評価を得た．「それぞれのアイテムはおしゃれか」という問いに対

7：GitHub「danielgatis / rembg」
https://github.com/danielgatis/rembg
8：GitHub「toshiaki1729 / dataset-tag-editor-standalone」
https://github.com/toshiaki1729/dataset-tag-editor-standalone
9：GitHub「bmaltais/kohya_ss」
https://github.com/bmaltais/kohya_ss

表 3　質問と被験者による評価

質問	高評価モデル 100	低評価モデル 100	高評価モデル 500	低評価モデル 500	LoRA なし
ファッションアイテムの組み合わせは適切か	3.39	3.41	3.11	**3.54**	2.05
コーディネートはおしゃれか	3.31	**3.42**	2.81	3.39	2.11
それぞれのアイテムはおしゃれか	3.22	3.31	2.94	**3.52**	2.22
それぞれのアイテムは価値が高そうか	3.17	3.27	2.91	**3.33**	2.42
着用人物は見た目が良いか	2.56	2.58	2.38	2.52	**2.67**
着用人物は体形が良いか	**3.36**	3.24	3.05	3.13	2.52
写真の撮影場所，背景はおしゃれか	3.36	**3.43**	3.09	3.16	2.36
画像に写った小物類はおしゃれか	3.16	**3.22**	3.03	3.11	2.27
画像には正しくファッションアイテムが写っているか	2.73	2.81	**3.00**	2.81	1.76
写真の構図はきれいだと感じるか	**3.47**	3.43	3.18	3.36	2.43
写真全体の色使いはきれいだと感じるか	**3.54**	**3.54**	3.35	3.46	2.66
写真に不自然なところはないか	2.57	**2.71**	2.45	2.49	1.86

する回答においても，同様の傾向が見られた．つまり，低評価モデルで生成された着用画像は，高評価モデルで生成された着用画像に比べて，相対的に「おしゃれ」であるといえる．これは、生成された着用画像の質に起因すると考えられる。「写真に不自然なところはないか」という質問に対して、低評価モデルの着用画像は、高評価モデルに比べて高い評価を得た。つまり、低評価モデルでは不自然な箇所が少ない着用画像が生成されたといえる。より自然な画像が生成されたことによって、高評価モデルで生成された着用画像よりも正確にファッション性を評価することができたと考えられる。

次に，そもそも学習と画像生成がうまくいっていたのかについて議論する．そのために,「画像には正しくアイテムが写っているか」という評価項目に注目する．5 つの手法間での大きな差は見られず，対象画像の種類や学習件数に依らず，総じて低い評価を得た．今回の実験で使用した生成結果画像を見ると，入力したファッションアイテムが，別のアイテムやカテゴリとして認識され，着用画像に反映されているケースが多く見られた．具体的には，デニム生地のボトムスが単体で写った画像を入力として与えた際に，モデルがそれをアウターとして処理し，デニム生地の不自然なアウターを着用した画像が生成された場合などがあった．これは，既存の ContorlNet が，入力された画像の特徴を詳細に維持することを苦手とすることに起因すると考えられる．ControlNet 内の他の機能を併用したり，image inpainting 機能を用いたりするといった改善が必要である．

画像の自然さについては、学習に用いた画像枚数で比較すると，上位下位それぞれ 500 枚ずつで学習した高評価モデルと低評価モデルの生成結果は，100 枚ずつで学習したモデルの生成結果に比べて低い評価を得た．このことから，LoRA モデルの学習には，多量の画像を使用することが，生成結果の質を低下させることにつながるといえる．これは，画像の枚数が増え，データセット内の画像の構図が多様になることで，LoRA の学習時に正しく学習をおこなうことができていないことが原因であると考えられる．

6 ま と め

本研究では，ファッションアイテムの映った画像を入力とし，SNS において高評価と低評価のそれぞれが付き得るコーディネート画像を出力する手法を実装し，評価した．

実験を通して今後の課題が見出された．まず，より自然な画像を生成する手法が必要である．

次に，学習に用いるデータについての検討が必要である．本研究では，WEAR から投稿を収集した時点でのいいね数をもとにして，センスのある画像とセンスのない画像に分けた．これにより，人気のあるユーザが投稿してすぐの画像も，いいね数が少ないために，センスのない画像に分類されるという問題がある．今後は投稿時間や保存数などの WEAR 内の他のメタデータと併用し，ユーザからの評価を総合的に加味したデータセットを作成する予定である．

より発展的な課題として，入力のファッションアイテムカテゴリを Stable Diffusion に正確に認識させることが必要である．本論文では，ファッションアイテム画像に対し，背景の除去と画像の拡大を施すことで，簡易的にアイテムカテゴリを認識させることを試みた．しかし，実際の画像生成においては，カテゴリが正確に認識できていないことにより，崩壊した画像が出力されるケースが多々あった．今後は物体検知アルゴリズムや image inpainting の技術を使用し，より精度の高い画像生成を目指す．

謝　　辞

本研究の一部は JSPS 科研費 21H03775，21H03774，22H03905 による助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Yujuan Ding, P.Y. Mok, Yunshan Ma, and Yi Bin. Personalized fashion outfit generation with user coordination preference learning. *Inf. Process. Manage.*, Vol. 60, No. 5, sep 2023.

[2] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, p. 1078–1086, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Xiaoya Chong and Howard Leung. Body shape-aware

object-level outfit completion for full-body portrait images. *Applied Sciences*, Vol. 13, No. 5, p. 3214, 2023.

[4] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018.

[5] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision*, pp. 1680–1688, 2017.

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

コスメレビューデータを用いたメイクパターン特性の可視化

福本真以子† 牛尼 剛聡†† 角谷 和俊†

† 関西学院大学総合政策学部メディア情報学科 〒 669-1330 兵庫県三田市学園上ケ原 1 番
†† 九州大学大学院芸術工学研究院 〒 815–8540 福岡県福岡市南区塩原 4 丁目 9 番 1 号
E-mail: †{hmw63878,sumiya}@kwansei.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし 近年,「韓国メイク」や「ナチュラルメイク」など, ブームとともに流行したものから世代を問わず長年親しまれるものまで, 様々なメイクパターンが存在している. しかし, 異なるメイクパターンの間でもメイクの特徴や使用されるコスメが類似する場合があり各メイクパターンに明確な定義がなされていない. 加えて近年でいうと「ツヤ」「マット」など流行の影響を受け, 各メイクパターンの持つ特徴は時代とともに変化・統合・分岐・新興している. そこで本研究では, コスメレビュー内の「キラキラ」「ツヤ」などの言語表現に着目し, 顔のパーツごとに各メイクパターンの特徴を抽出する. 抽出した各パーツの特徴をメイクパターンごとに融合させることで各メイクパターンの体系的な特徴を導き出し, また時系列分析を行うことでメイクの辿った変化を考察する.

キーワード 情報推薦, コスメ, メイク, レビュー, 可視化

1 はじめに

近年, メイクにおいて「韓国メイク」や「ナチュラルメイク」など, ブームとともに隆盛したものから根強いものまで, 様々なメイクパターンが存在している. しかし, 異なる名称を持つメイクパターンの間でも, 例えばラメの入ったアイシャドウを使用することやチークの使用の有無など, メイクの特徴や使用されるコスメが類似する場合がある. また, 同一のメイクパターンでも, その特徴が変化する場合がある. たとえば,「ナチュラルメイク」は長年幅広い世代において使われるメイクパターンであるが, 過去にはベージュやブラウンを主に用いたメイクであったが, 近年は「ツヤ肌」や「グリッター」など流行の影響を受け, 過去とは異なる特徴を持っている. このように, メイクパターンの持つ特徴は時代とともに変化している.

そこで本研究では, 時間経過に伴ったメイクパターン間の関係性の分析を目的とする. 具体的には, コスメレビューに出現する特徴語とレビュー投稿日に着目し, 顔のパーツごとにメイクパターンの特徴を抽出する. 最終的に顔の全てのパーツを統合することで, 各メイクパターンの特徴を体系的に推定し, また時系列に沿ったメイクパターン間の類似性を可視化する,

本稿では, 特徴語を基にしたメイクパターン間の関係性の可視化および時系列分析について説明する. 本稿の構成は次のとおりである. 2 章で本研究と関連研究の関連性について述べる. 3 章ではコスメレビューを用いたメイクパターン特性の可視化の具体的な手法について述べる. 4 章では 3 章で述べた手法に基づく, 2020 年から 2022 年のコスメレビューデータを用いたメイクパターン特性の可視化結果を提示し, 各期間に対する考察を述べる. 5 章では 3 年間の可視化結果を通した全体の考察を述べ, 最後に 6 章ではまとめと今後の検討事項について述べる.

2 関連研究

コスメに関する研究はこれまでにも多く行われている. レビューを用いた研究として, 谷口ら [1] は, コスメアイテム分類に対する学習を施した Word2Vec に基づく評価表現辞書構築を提案している. 奥田ら [2] はレビューを用いた特徴量比較手法を提案している. 可視化に関する研究として, 矢部ら [3] はコスメアイテム間の類似度を可視化する手法を提案している. また, 番場ら [4] は肌質感の言語表現に着目しマッピングする手法を提案しており, 河野ら [6] らは化粧品レビュー本文と Youtube における化粧動画に使われる語彙の分析を行った. また, 本研究の先行研究である小川ら [5] の研究では, メイクパターンと芸能人との関係と, コスメレビューを用いてそれぞれメイクパターン特性の可視化を提案している.

またレビューを用いた情報の可視化という点においても多くの研究がこれまでに行われている. 岑ら [7] は映画のレビューを用いて時系列に基づく映画レビューの評価視点の抽出方法を提案し, 小川ら [8] は商品レビューを複数のクラスタに分類し, クラスタのキーワード抽出と SOM による関係性の可視化を行った.

このように, コスメアイテムの類似度や, コスメに限らずレビューを用いた関係性の可視化に関する研究は多いが, メイクパターンの関係性に着目した研究は少ない. 本稿では, メイクパターンに関連するコスメレビューを用いて, メイクパターンの持つ特性の類似度を可視化する.

3 メイクパターン特性の可視化

3.1 メイクパターン特性可視化に際する仮説

本研究ではレビューデータを用いて様々なメイクパターンの特性の年・季節ごとの変化を可視化する. 可視化を経ることで

例えば夏の季節における「マスクメイク」が，化粧崩れを防止するために他のメイクパターンでのメイクをする際に使用するコスメアイテムとは異なるコスメアイテムを使用している等，季節的な各メイクパターンの持つ特性の変化や，中高生のメイクである「スクールメイク」の流行を反映した他メイクパターンへの近似等の仮説を検証する．

3.2 可視化手法

本稿では，16 コスメアイテムのレビューを用いてメイクパターン特性の可視化を行う．可視化には，2 つの段階を取る．まず，16 コスメアイテムを 6 つの大カテゴリに分類したのち，大カテゴリそれぞれのレビューにおいて頻出する特徴語を取得する．次に，各大カテゴリにおける，「ピンクメイク」や「韓国メイク」等からなる 10 のメイクパターン名を含むレビュー内の特徴語の出現頻度をそれぞれ算出・主成分分析し，アイテムカテゴリを結合することで各メイクパターン特性の可視化を行う．この 2 つの段階について，3.3 から 3.7 で詳細を述べる．表 1 に本研究で対象とする 10 メイクパターンの一覧を示し，表 2 に本稿で使用した 16 コスメアイテムカテゴリと，それを分類した 6 大カテゴリを示す．

表 1　メイクパターン一覧

ピンクメイク	オレンジメイク	ブラウンメイク
韓国メイク	中国メイク	ギャルメイク
ナチュラルメイク	仕事メイク	スクールメイク
マスクメイク		

表 2　16 コスメカテゴリ一覧および 6 大カテゴリの分類

大カテゴリ	コスメアイテム
アイシャドウカテゴリ	パウダーアイシャドウ
	ジェル・クリームアイシャドウ
アイライナーカテゴリ	リキッドアイライナー
	ペンシルアイライナー
	ジェルアイライナー
	その他アイライナー
マスカラカテゴリ	マスカラ
リップカテゴリ	口紅
	リップグロス
	リップライナー
チークカテゴリ	パウダーチーク
	ジェル・クリームチーク
ファンデーションカテゴリ	パウダーファンデーション
	リキッドファンデーション
	クリーム・ジェルファンデーション
	その他ファンデーション

3.3 メイクパターン名を含むレビューの抽出

メイクパターン名を含むレビューから特徴語を取得する処理の流れを図 1 に示す．まず，コスメアイテムごとに「韓国メイク」や「ナチュラルメイク」といった，メイクパターンの名称を含むレビューを抽出する．続いて，抽出したレビュー群に対して形態素解析を行う．形態素解析によって得られた単語のうち形容詞，名詞，副詞のみを対象として，出現頻度の高い単語を上位から最大で 500 単語抽出する．その後，各コスメアイテムに関するすべてのレビューに対して形態素解析を行い，ここで「キラキラ」や「マット」といったメイクパターンを指定しない状態でのコスメアイテムのレビューに頻出し，それぞれのアイテムの特徴を表す特徴語を選出する．この特徴語が「メイクパターンの名称を内容に含むレビュー群」の中に出現する頻度をメイクパターン毎に算出する．

3.4 大カテゴリごとの特徴抽出

ここでは，3.2 で説明した大カテゴリのレビューにおいて，のちにメイクパターンの特徴を形成する特徴語の抽出方法を述べる．まず大カテゴリそれぞれのレビュー文を形態素解析し，TF の値および主観で特徴語を各大カテゴリで 25 語，計 150 単語選定する．各カテゴリの特徴語は「キラキラ」「マット」等製品そのものの特長を表現する製品情報，「サラサラ」「ふわっと」等使用時の感覚を表現する使用感情報，「可愛い」「上品」等そのコスメアイテムを使用した化粧完成後の印象情報をもとに抽出した．抽出した特徴語は以下に示す．

3.4.1 アイシャドウカテゴリにおける特徴語

アイシャドウカテゴリにおける特徴語 25 語を表 3 に示す．製品情報にはラメの有無や様子を表現するため，「キラキラ」「ギラギラ」「ザクザク」等を採用した．

表 3　アイシャドウカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	キラキラ	ギラギラ	ツヤツヤ
	マット	ぎっしり	ザクザク
	パール		
使用感情報	サラッと	サラサラ	しっとり
	パッと	ほんのり	密着
	ぴったり		
印象情報	かわいい	しっかり	上品
	色っぽい	甘い	儚い
	大人っぽい	透明感	ナチュラル
	美しい	強い	

3.4.2 アイライナーカテゴリにおける特徴語

アイライナーカテゴリにおける特徴語 25 語を表 4 に示す．製品情報には色味である「ブラック」「ブラウン」等とアイライナーの形態である「リキッド」「ペンシル」を採用した．

3.4.3 マスカラカテゴリにおける特徴語

マスカラカテゴリにおける特徴語 25 語を表 5 に示す．製品情報には「ロング」「ボリューム」等各製品の特性となる単語を採用した．

3.4.4 リップカテゴリにおける特徴語

リップカテゴリにおける特徴語 25 語を表 6 に示す．製品情報には「ピンク」「オレンジ」等の色味と「ティント」「グロス」等の商品の種類となる単語を採用した．

表 4　アイライナーカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	ブラック	ブラウン	グレー
	ベージュ	ピンク	バーガンディ
	リキッド	ペンシル	
使用感情報	くっきり	密着	ぴったり
	ぷっくり	薄い	スルスル
印象情報	可愛い	しっかり	上品
	色っぽい	甘い	儚い
	大人っぽい	透明感	強い
	ナチュラル	美しい	

表 5　マスカラカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	ロング	カール	キープ
	ボリューム	セパレート	透明
	繊維		
使用感情報	発色	太い	濃い
	はっきり	しっかり	
印象情報	可愛い	上品	色っぽい
	甘い	儚い	大人っぽい
	透明感	ナチュラル	強い
	綺麗	存在感	美しい
	繊細		

表 6　リップカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	ピンク	オレンジ	ブラウン
	ベージュ	ティント	マット
	グロス		
使用感情報	スルスル	しっとり	ほんのり
	しっかり	乾燥	潤い
印象情報	可愛い	上品	色っぽい
	甘い	儚い	大人っぽい
	透明感	ナチュラル	強い
	綺麗	明るい	みずみずしい

3.4.5　チークカテゴリにおける特徴語

チークカテゴリにおける特徴語 25 語を表 7 に示す．製品情報には，製品が頬紅としての役割かハイライトとしての役割かを判別するため「ハイライト」「パール」等の単語を採用した．

表 7　チークカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	ハイライト	ツヤ	ギラギラ
	キラキラ	パール	
使用感情報	ふんわり	ほんのり	しっとり
	ふわっと		
印象情報	可愛い	上品	色っぽい
	甘い	儚い	大人っぽい
	透明感	ナチュラル	強い
	綺麗	存在感	美しい
	繊細	明るい	濃い
	血色		

3.4.6　ファンデーションカテゴリにおける特徴語

ファンデーションカテゴリにおける特徴語 25 語を表 8 に示す．製品情報にはファンデーションの形態である「パウダー」「クッション」等と「ピンク」「オレンジ」等の色味を表す単語を採用した．

表 8　チークカテゴリにおける特徴語 25 語一覧

単語カテゴリ	単語		
製品情報	パウダー	クッション	クリーム
	マット	ツヤ	オレンジ
	ピンク	カバー	
使用感情報	白い	みずみずしい	ぴったり
	サラサラ	ポンポン	ふんわり
	密着		
印象情報	透明感	ナチュラル	綺麗
	美しい	明るい	血色
	透明感	きめ細かい	薄い
	素肌		

3.5　特徴語による各メイクパターンの要素の生成

続いて，各大カテゴリ内にて，各メイクパターン名を含むレビューを抽出し，レビュー群を 1 つの文書として，特徴語の TF を算出する．使用する TF の式は下記のとおりである．

$$TF_{d,k_i} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

メイクパターンによってレビュー数が異なるため，TF 算出において各メイクパターン名の含まれる総レビューにおける総単語数で割ることで TF 値を正規化した．この TF の値を各メイクパターンが持つ 150 次元の要素の値として採用する．

3.6　各メイクパターンにおける，全顔パーツの特徴の融合

3.3 から 3.5 で得られた，メイクパターンにおける各大カテゴリの特徴を融合する流れを図 1 に示す．ここでは，メイクパターンごとに，1 つのベクトルを生成する．生成するベクトル内の要素として，各大カテゴリの特徴語の特徴量を採用する．

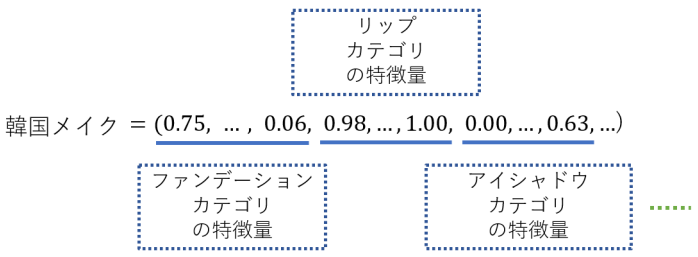


図 1　各大カテゴリの特徴の融合

3.7　主成分分析と寄与率の算出および可視化

3.5 で得られた各メイクパターンの特徴量を 2 次元平面にを可視化する流れを図 2 に示す. まず, メイクパターン毎の特徴語の出現頻度を用いて各メイクパターンを 150 次元のベクトルで表現する．次にこのベクトルを主成分分解し, 寄与率が高い主成分 2 軸を用いて 2 次元上にプロットすることによりメイクパターンを可視化する.

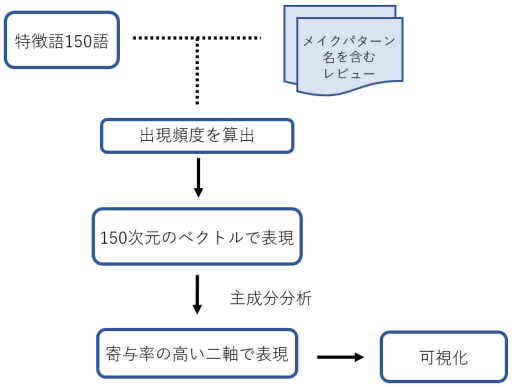


図 2　メイク特性の可視化のフロー

4　可視化に基づく時系列分析

上記に示した可視化手法を用いて期間に区切った可視化を行う．本稿では 2020 年から 2022 年までのレビューデータをそれぞれ 1 月から 3 月，4 月から 6 月，7 月から 9 月，10 月から 12 月に区切ってメイクパターン特性および時系列的なメイクパターン同士の関係性の遷移を可視化した．4.1 から 4.6 にて各年の可視化結果およびそれぞれの期間における考察を示す.

4.1　2020 年における可視化結果

2020 年の 1-3 月，4-6 月，7-9 月，10-12 月おける各メイクパターン特性の関係性を可視化したものを下記図 3 から図 6 に示す.

4.2　2020 年における可視化結果の考察

2020 年においては，全期間を通して「韓国メイク」・「ギャルメイク」・「中国メイク」といった一見すると派手な印象を持つメイクパターン同士が近くにマッピングされ，加えて「スクールメイク」もこれらのメイクパターンと近い位置にマッピングされる結果となった.

また特徴的な結果として，7 月から 9 月の期間における「マスクメイク」が独立と，10 月から 12 月にかけて「ピンクメイク」の独立があげられる．このことからは，前者はマスクを伴ったメイクをする際に使用するコスメアイテムの特異性が反映されたものだと考えられる．また後者に関しては，10 月から 12 月のホリデーシーズンに発売されるクリスマスコフレ等の類を利用した「ピンクメイク」の持つ，他のメイクパターンとは異なる印象が反映されているのではないかと考えられる.

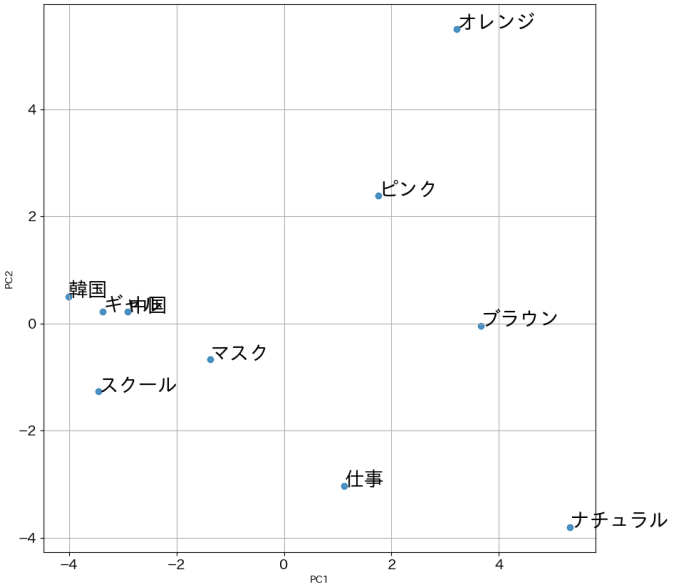


図 3　2020 年 1 月から 3 月における各メイクパターン特性

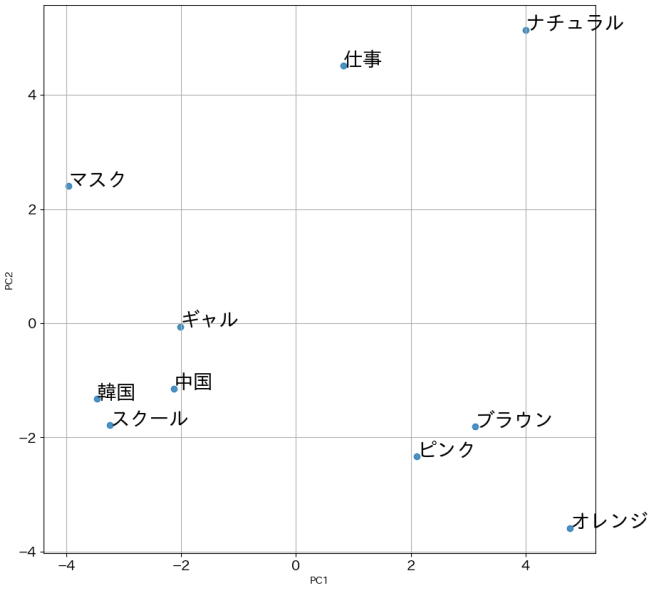


図 4　2020 年 4 月から 6 月における各メイクパターン特性

.

4.3　2021 年における可視化結果

2021 年の 1-3 月，4-6 月，7-9 月，10-12 月おける各メイクパターン特性の関係性を可視化したものがを下記図 7 から図 10 に示す.

4.4　2021 年における可視化結果の考察

2021 年においては，引き続き概ね全期間を通して「韓国メイク」・「ギャルメイク」・「中国メイク」といった印象的なメイクが近くにマッピングされる結果となった．また特徴的な結果として，4 月から 6 月の期間において，「仕事メイク」が独立していること，加えて 7 月から 9 月の期間において，「マスクメイク」が独立している．このことからは，それぞれ春の就職シーズンにおける仕事用のメイクの持つ，他のメイクパターンとは異な

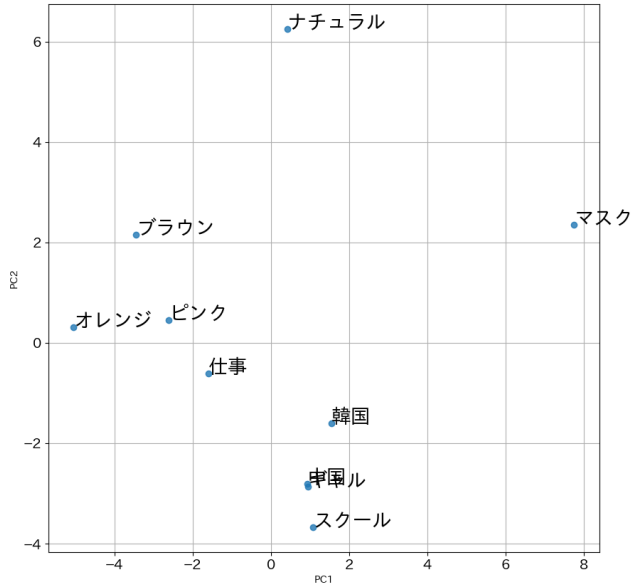


図 5　2020 年 7 月から 9 月における各メイクパターン特性

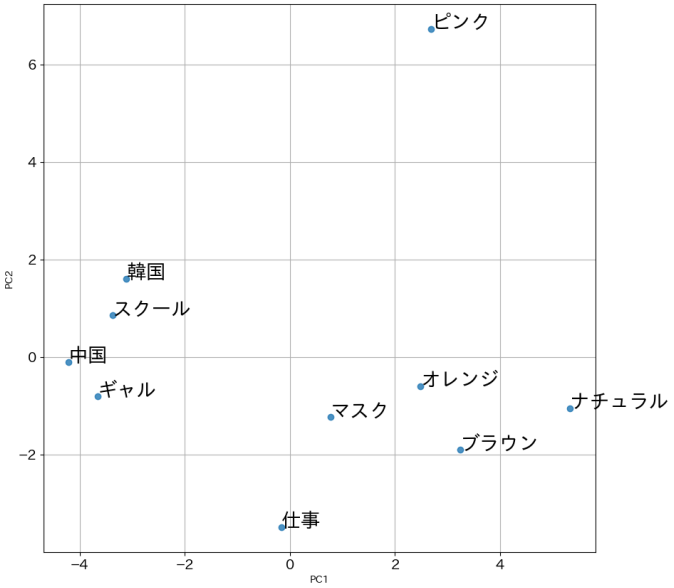


図 7　2021 年 1 月から 3 月における各メイクパターン特性

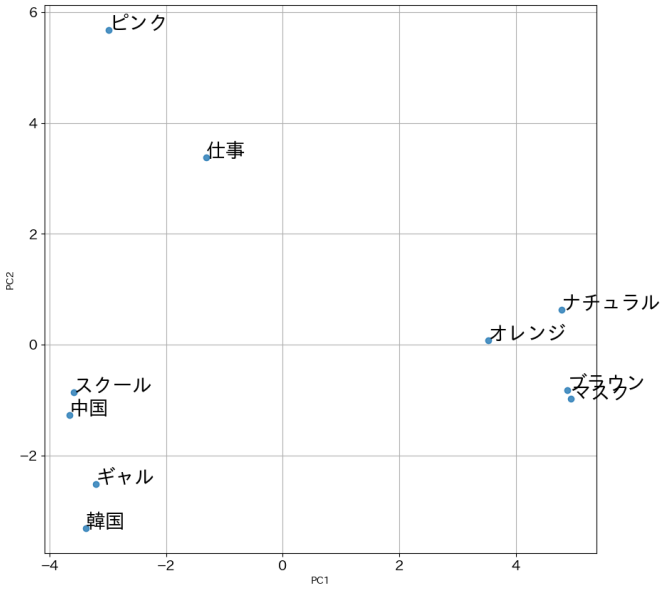


図 6　2020 年 10 月から 12 月における各メイクパターン特性

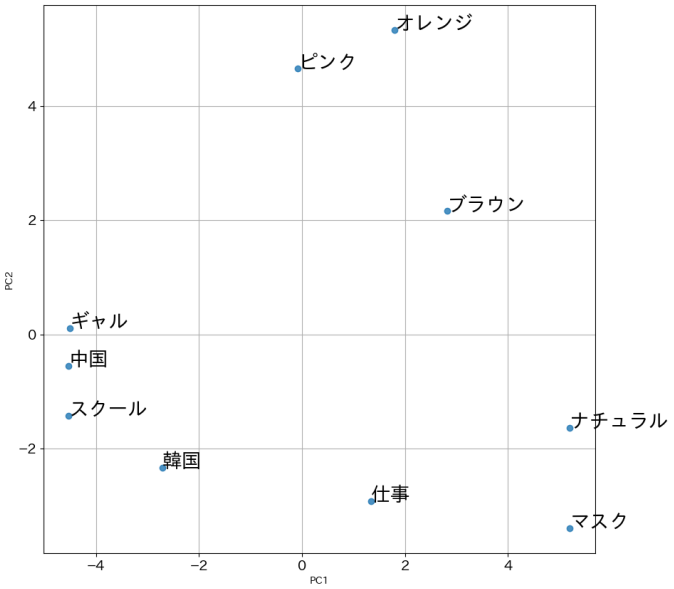


図 8　2021 年 4 月から 6 月における各メイクパターン特性

る特性や，2020 年に引き続き夏シーズンにおける「マスクメイク」の特異性が表れていると考えられる．加えて 10 月から 12 月のホリデーシーズンにおいて，「ピンクメイク」は 2020 年の同シーズン同様，同年の他シーズンにおいて比較的近くにマッピングされていた「ブラウンメイク」「オレンジメイク」と離れた位置にマッピングされた．このことからはクリスマスコフレの発売に伴う「ピンクメイク」の特異性の増長がうかがえる．

4.5　2022 年における可視化結果

2022 年の 1-3 月，4-6 月，7-9 月，10-12 月おける各メイクパターン特性の関係性を可視化したものを下記図 11 から図 14 に示す．

4.6　2022 年における可視化結果の考察

2022 年においても，全期間を通して「韓国メイク」「ギャルメイク」「中国メイク」といった印象的なメイクが近くにマッピングされる結果となった．特徴的な結果として，2020 年および 2021 年の 7 月から 9 月の夏シーズンに見られた「マスクメイク」の顕著な独立が見られなかった．加えて，2020 年や 2021 年の結果に見られた 10 月から 12 月のホリデーシーズンにおける「ピンクメイク」の独立も見られない結果となった．

5　考　　察

5.1　各メイクパーンの持つ特性の近似

おおよその期間を通して「中国メイク」「ギャルメイク」「韓国メイク」といった一見派手な印象をもたらす流行メイクパターン同士はそれぞれ近くにマッピングされる結果となった．この点からはこれらの印象的なメイクは異なる名称を冠したメイク

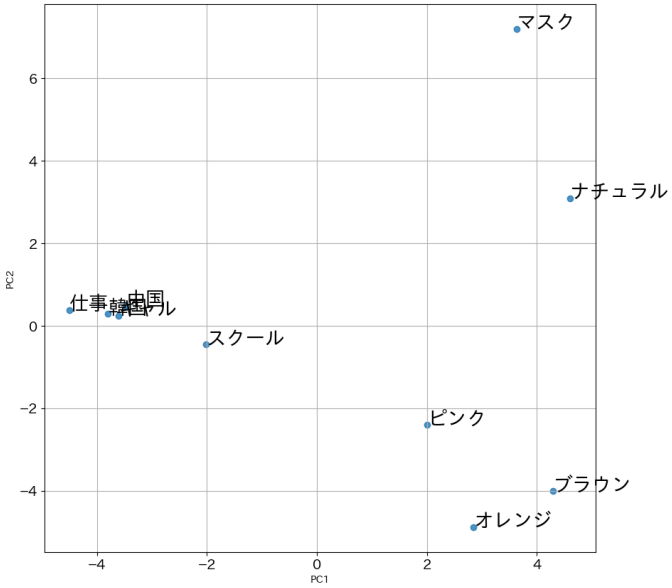


図 9　2021 年 7 月から 9 月における各メイクパターン特性

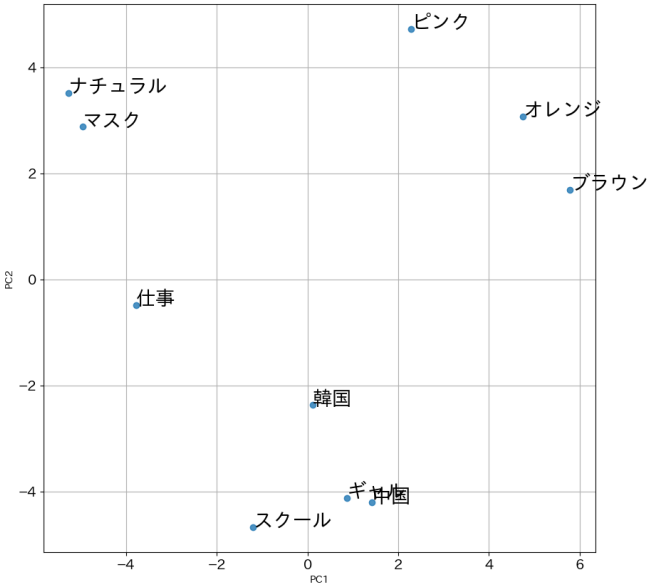


図 11　2022 年 1 月から 3 月における各メイクパターン特性

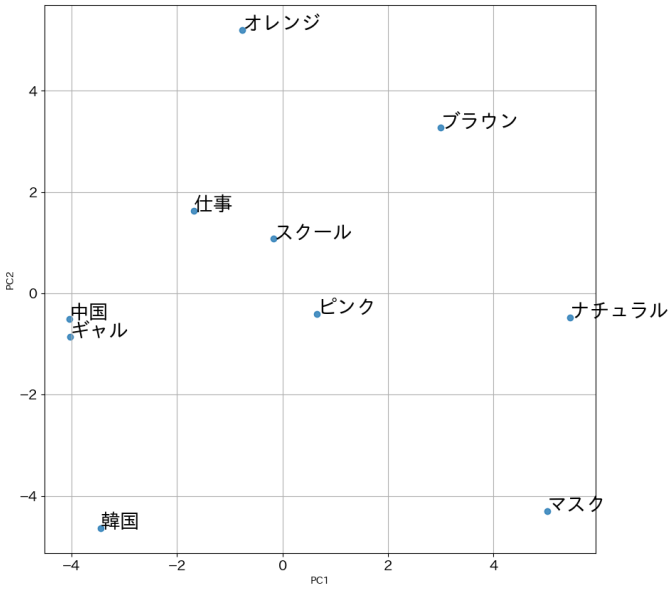


図 10　2021 年 10 月から 12 月における各メイクパターン特性

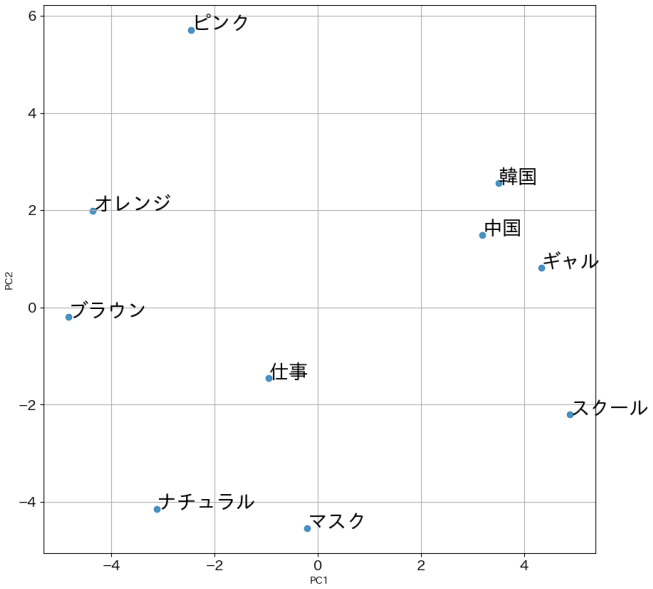


図 12　2022 年 4 月から 6 月における各メイクパターン特性

パターン同士でありながら，メイクパターン間をまたいだ流行要素の共有やメイクパターン間での相互影響により，結果として全顔でのイメージが近似したことが考えられる．

加えて複数の期間においては，これら「中国メイク」「ギャルメイク」「韓国メイク」の近くに「スクールメイク」がマッピングされていた．この点からは中高生等の SNS によるメイク方法の吸収により，流行メイクパターンであるこれらのメイクパターンの要素がスクールメイクに継承されているのではないかと考える．

10 月から 12 月のホリデーシーズンを除いた期間において，「ピンクメイク」「オレンジメイク」「ブラウンメイク」が近くにマッピングされていた点からは，これらのカラーメイク群の持つ特性の違いとしては色そのものにあり，使用するコスメアイテムの印象や形状の点ではあまり差異が無いことがうかがえる．

5.2　シーズンごとの変化

2020 年と 2021 年の可視化イメージの結果から，この 2 年の 7 月から 9 月における，「マスクメイク」が他のメイクパターンと使用するコスメアイテムや印象を共有しないことがわかった．しかし 2022 年の同時期における「マスクメイク」は顕著な独立をしていない．このことから，この「マスクメイク」の動きがコロナ禍の社会の変容の示唆にもなり得ると考えられる．

また 2020 年と 2021 年の 10 月から 12 月のホリデーシーズンにおける「ピンクメイク」の独立しているが，2022 年同時期において「ピンクメイク」の独立は見られない．このことからは社会的なメイクの多様化および，ブラウンやベージュを使ったヘルシー志向のメイクの流行によるピンクに限らないクリスマスコフレのバリエーション多様化の影響ではないかと考えられる．

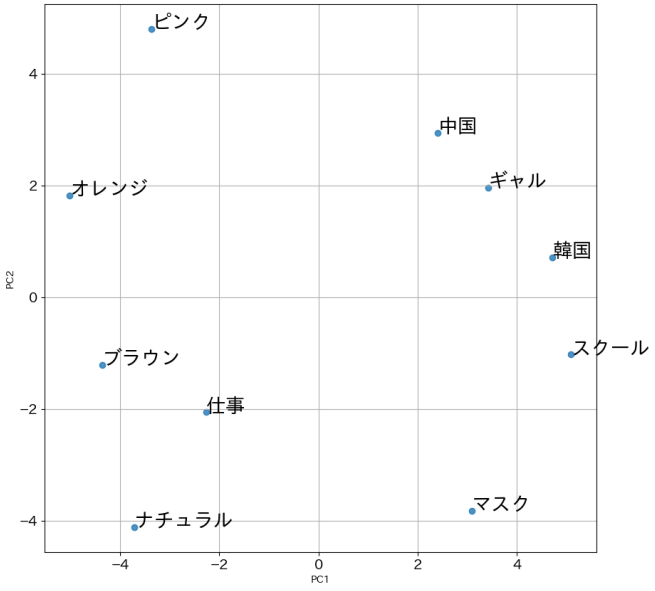


図 13　2022 年 7 月から 9 月における各メイクパターン特性

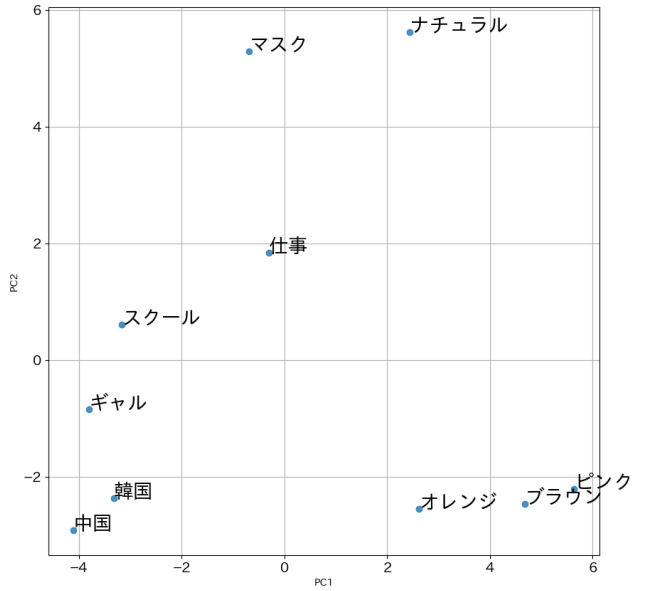


図 14　2022 年 10 月から 12 月における各メイクパターン特性の可視化

6　おわりに

本稿ではコスメレビューと特徴語を用いたメイクパターン特性の可視化手法を提案した．また，2020 年から 2022 年のレビューデータの投稿日を 3 ヵ月ごとに区切り，季節ごとの各メイクパターン同士の関係性の遷移の可視化および考察を行った．

今後の課題として，本稿では主観での選出を行った特徴語の客観的な抽出および可視化手順における主成分分析・寄与率判定を用いた二軸での可視化手法の再考を行う．

謝　　辞

本研究は，株式会社アイスタイルから「@cosme のコスメレビュー (https://www.cosme.net/)」の提供を受け，濱田健作氏，原田彩子氏，および勝並明子氏からコスメ分析の専門知識に関するアドバイスを頂いた．ここに記して謝意を表す．

文　　献

[1] 谷口祐奈, 奥田麻美, 上田真由美,Panote Siriaraya, 中島伸介, ”コスメ向けに学習した Word2Vec に基づく評価表現辞書構築手法”, 第 12 回データ工学と情報マネジメントに関するフォーラム,2020
[2] 奥田麻美, 上田真由美, 中島伸介, ”評価項目別スコアを用いたコスメアイテム推薦手法”, 第 12 回データ工学と情報マネジメントに関するフォーラム,2020
[3] 矢部沙也加, 上田真由美, 中島伸介,”評価項目別スコアを用いたコスメアイテム間の相違点可視化システム”, 第 13 回データ工学と情報マネジメントに関するフォーラム,2021
[4] 番場文枝, 猪股真美, 伊藤貴之, 豊田成人, 大高瞳, 笹本裕美, ”肌微細構造の印象評価”, 第 4 回データ工学と情報マネジメントに関するフォーラム,2012
[5] 小川芽来, 牛尼剛聡, 角谷和俊, ”コスメレビューを用いたアイテムカテゴリの特徴に基づいたメイクパターンの可視化”, 第 14 回データ工学と情報マネジメントに関するフォーラム,2022
[6] 河野 雪乃, 安尾 萌, 松下 光範, 山西 良典,”私に似合う化粧を学ぶために—化粧品レビューと化粧動画内の発話の語彙分析—”, 第 14 回データ工学と情報マネジメントに関するフォーラム,2022
[7] 岑 天霞, 渡邉 英徳,”映画レビューの時系列による評価視点の抽出と可視化”, エンタテインメントコンピューティングシンポジウム 2023 論文集, pp72-76,2023
[8] 小川 和晃, 田村 哲嗣, 速水 悟 ,”商品レビューにおける時系列情報に着目したクラスタ分析と可視化”, 人工知能学会,2016

DEIM Forum 2024 XX-Y

Personalized Fashion Recommendation with Image Attributes and Aesthetics Assessment

Chongxian Chen † Fan Mo ‡ Xin Fan ‡ Hayato Yamana §

† ‡ Graduate School of Fundamental Science and Engineering 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
§ Faculty of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

E-mail: † chenc@toki.waseda.jp, ‡ bakubonn@toki.waseda.jp, ‡ fan_xin@fuji.waseda.jp

§ yamana@yama.info.waseda.ac.jp

Abstract Personalized fashion recommendation is a difficult task because 1) the decisions are highly correlated with users' aesthetic appetite which previous work frequently overlooks, and 2) many new items are constantly rolling out that cause strict cold-start problems in the popular identity (ID)-based recommendation methods. These new items are critical to recommend because of trend-driven consumerism. In this work, we aim to provide more accurate personalized fashion recommendations and solve the cold-start problem by converting available information, especially images into two attribute graphs focusing on optimized image utilization and noise-reducing user modeling. Compared with previous methods that separate image and text as two components, the proposed method combines image and text information to create a richer attributes graph. Capitalizing on the advancement of the large language and vision models, we experiment with extracting fine-grained attributes ever efficiently and as desired using two different prompts. Preliminary experiments on the IQON3000 dataset have shown that the proposed method achieves competitive accuracy compared with baselines.

Keywords—Graph Neural Network, Personalized Fashion Recommendation, Large Language and Vision Models, Attributes Graph, Image Aesthetics Assessment, Explainable Recommendation

1. INTRODUCTION

Recommender systems have become indispensable tools in tackling the contemporary challenge of information overload. Recommender systems play a crucial role in enhancing customers' online shopping experiences by offering personalized lists of potential items. While previous approaches have achieved considerable success using graph neural networks (GCN) for personalized recommendations [8,9], the prevalent ID-based approach in GCN encounters difficulties with the cold-start problem. Recent research has demonstrated success in overcoming this challenge by leveraging item attribute graphs [1]. However, personalized fashion recommendation presents unique challenges due to the critical role of image-driven influence on shopping decisions. Previous methods in fashion recommendation have predominantly relied on pre-trained image models for extracting embeddings to capture users' image preferences.

In this study, we propose an innovative approach to leverage images in personalized fashion recommendations. Capitalizing on the advancements in large language and vision models (LLVM) [14], our method processes images into two distinct parts of fine-grained attributes ever efficiently and reproducibly. We claim that there are two distinct influences in images that affect users' shopping decisions. Firstly, item-specific attributes that describe features of an item, and secondly, image aesthetic attributes that capture users' aesthetic preferences. An example is shown in Fig. 1 to demonstrate the difference between item-specific attributes and image aesthetic attributes. Alongside the fine-grained keywords from item descriptions, we construct the Items and Item Attributes Graph that integrates both images and text from the items. To model the users' preferences, we construct a Users, Image Aesthetic Attributes and Items Graph to integrate not only item information but explicitly model users' aesthetic preferences.

Unlike other shopping domains such as groceries and hardware, fashion shopping places a significant reliance on images for decision-making. Additionally, the constant influx of new fashion items each season, coupled with consumers' high interest in new trends, underscores the critical role of fashion recommender systems in recommending these new items. Our proposed method integrates both images and texts to construct the Items and Item Attributes Graph and the Users, Image Aesthetic Attributes and Items Graph, effectively addressing the cold start problem and offering optimized utilization of images compared to previous methods.

Our approach addresses two primary challenges in

personalized fashion recommendation: 1) the ability to recommend new items, i.e., solving the cold-start problem, and 2) the optimized utilization of image data to capture users' personal aesthetic preferences. Instead of treating text and images as separate entities, our proposed method combines both sources to construct a comprehensive attributes graph, providing a unique solution to personalized fashion recommendations compared with previous independent components. The proposed method also better models users' preferences explicitly by connecting users with not just items, but also image aesthetic attributes by constructing the innovative Users, Image Aesthetic Attributes and Items Graph. Previous methods [4,5] model users' preferences by integrating item information, which we claim to have more item-specific noise that is not related to the users' preferences.

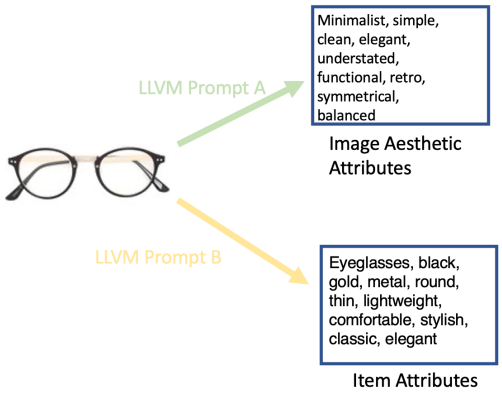


Fig. 1. A demonstration of example Image Aesthetic Attributes and Item Attributes.

The proposed methods contain three major components to model the items: 1) extract item attributes from items' images using LLVM. Utilizing the advancement in LLVM, we are able to experiment with extracting fine-grained attributes from the images more efficiently and as desired, and 2) extract item attributes from items' text descriptions. The brand, price, material, category, color, and introduction are carefully processed into item attributes, and 3) construct the Items and Item Attributes Graph. With item attributes extracted from both images and texts, we ensure to model the items comprehensively.

We also propose a novel method to model the users in order to reduce the items' noise in representing users' preferences. We utilize LLVM to extract image aesthetics from the images to model users' aesthetic preferences. Together with item embeddings learned from the Items and Item Attributes Graph, we build the Users, Image Aesthetic Attributes and Items Graph.

The main contributions of this paper are summarized as 1) we are the first to experiment with constructing an Items and Item Attributes graph using both image and text information, 2) we show the results of experimenting with image attributes extraction using two different prompts using upgraded LLVM, 3) an innovative approach to model the users' preference to reduce the item-specific noise.

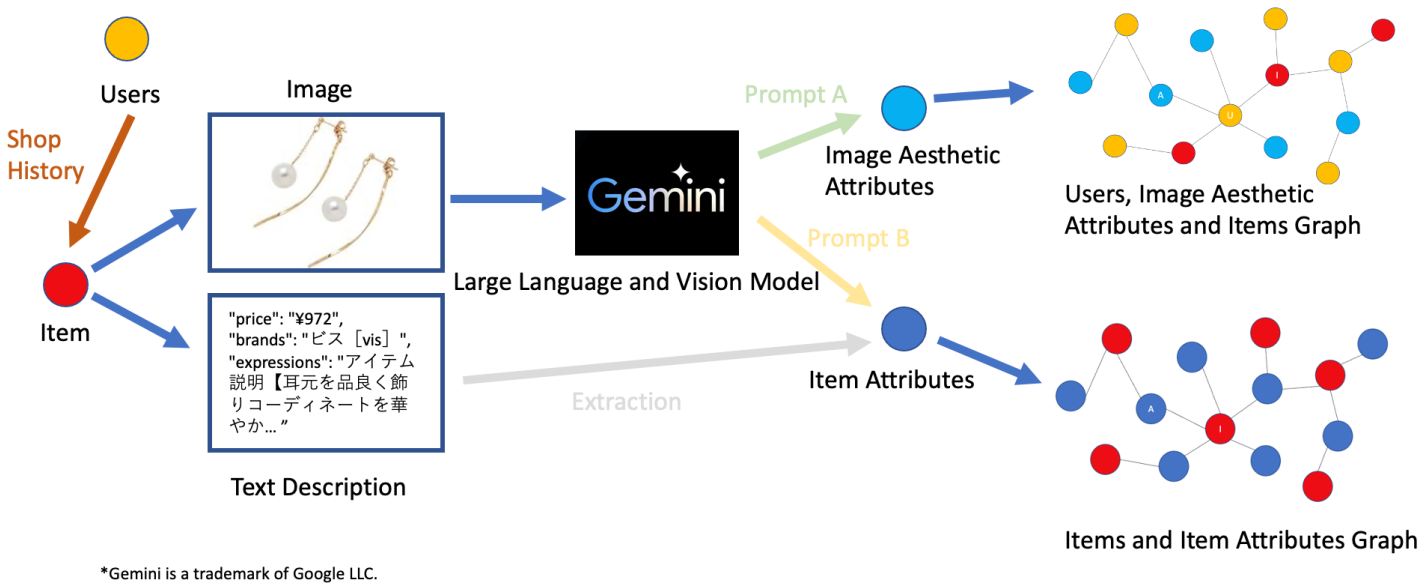
2. RELATED WORK

1. Graph Neural Networks and the Cold Start Problem in the Recommender Systems

Graph Neural Networks (GNN) have shown success in recommender systems. Recent graph-based recommenders [2, 3] typically construct user-item bipartite graphs and then capture user-item correlations by propagating their embeddings using GNNs [8, 9]. The powerful information propagation using graphs enables efficient and large-scale modeling of the items and users. However, such methods suffer from the cold start problem that the dominating ID-based approach fails to work. The new items are missing the ID-based embeddings because the models are not trained on these new items, thus being unable to recommend the new items. To solve the cold start problem, Cao et al. [1, 13] proposed to construct an item-attributes graph. However, Cao et al.'s [1] fine-grained methods require careful preprocessing in extracting the attributes from text.

2. Fashion Recommender Systems

The compatibility between fashion items has attracted studies in recent years. Effectively modeling the compatibility between fashion items is challenging in fashion research [4,5,7]. Traditionally, this problem has been cast as a metric learning challenge, wherein a compatibility space is learned, and the distance between items within this space reflects their compatibility [4]. Recent studies [5] has been utilizing hierarchical graphs to model the relationship between users, sets, and items. Despite the effectiveness, recent studies [4,5,7] have been utilizing pre-trained image models like ResNet to extract visual features. Such approaches separate the text attributes from the image, making them two independent components in modeling fashion items, causing nonoptimal performance, also demonstrated by the ablation study. In our proposed methods, we aim to combine images and texts as one component for optimal utilization of all data sources to model fashion items.



*Gemini is a trademark of Google LLC.

Fig. 2. An overview of the Items and Item Attributes Graph and the Users, Image Aesthetic Attributes and Items Graph construction.

3. PRELIMINARIES

Before presenting our proposed methodology, we provide an overview of GCN-based Collaborative Filtering.

1. GCN-based Collaborative Filtering

Graph Convolutional Networks (GCNs) [2,3] are designed to capture collaborative signals within historical interaction data through convolutional operations on the user-item bipartite graph constructed from past user-item interactions. In the forward propagation phase, given the initial representations $\boldsymbol{e_u}$ and $\boldsymbol{e_i}$ for user u and item i, GCNs aggregate information from neighbors recursively to create final representations, and $\boldsymbol{e_i}$, for user u and item i, respectively. These representations are then utilized to calculate the preference score representing the user's preference for item i. Subsequently, the preference scores for unseen items are used to generate the recommendation list.

During the backpropagation process, the parameters of GCNs are optimized using the Bayesian personalized ranking (BPR) loss [10]. This loss function is pairwise, ensuring that a user's preference score for interacted items is higher than the preference score for non-interacted items. In conclusion, GCNs learn to enhance recommendation performance by iteratively updating their parameters based on the information propagation captured from graphs made from historical interactions, providing personalized, effective, and large-scale recommendations for users.

TABLE I. SYMBOLS

Symbol	Explanation
e_i	Item embedding
e_{ia}	Item Attributes embedding
e_{iaa}	Image Aesthetic Attributes embedding
e_u	Users embedding
N_{i_u}	the set of items that were interacted by user u
N_{iaa_u}	the set of image aesthetic attributes that were interacted by user u
N_{ia_i}	the set of item attributes that were interacted by item i
e_i^k	The embedding of items at the k-th layer
e_u^k	The embedding of users at the k-th layer
n_p	The number of total price categories in item attributes
n_n	The number of negative item samples in the experiment
G_{iia}	The Items and Item Attributes Graph where $G_{iia} = \{V_i, V_{ia}, E_{iia}\}$
G_{uiaai}	The Users, Image Aesthetic Attributes and Items Graph where $G_{uiaai} = \{V_u, V_i, V_{iaa}, E_{uiaai}\}$

4. METHODOLOGY

In this section, we elaborate on the detailed proposed methods. Table I. shows the symbols used in this paper.

A. Image to Item Attributes Extraction

Instead of treating the image as just embedding, we propose to adopt LLVM to extract attributes from the image ever efficiently and as desired to enrich our attributes graph. With the recent advancement in LLVM, we experiment with Google's Gemini Pro Vision model for

image attributes extraction. Each time we provide an item's image and a prompt to the model using Google's Python API google.generativeai. After experimenting, the prompt we used to extract the item attributes from images is named Prompt B as follows: "Describe the item in the image using keywords. Describe the color, material, pattern, style, and feeling of the item using simple, common, and many English keywords. Output as keyword1, keyword2, keyword3, ... keywordn." The results are satisfying with an example demonstrated in Fig. 1.

B. Text Description to Item Attributes Extraction

Each item includes a text description that includes the brand, price, category, color, and item description. The texts are carefully analyzed and processed to extract useful keywords as part of the item attributes. Because the prices are discrete values, we process the prices into n_p categories, where n_p is a tunable parameter.

C. Image to Image Aesthetics Attributes Extraction

Besides extracting item attributes from the image, we claim that it is also important to extract image aesthetics to better model users' preferences. Users' shopping decisions are driven by the images they see. If we only aggregate the item embeddings to model the users, we will get unnecessary item-related noise. For example, a user may buy a pair of glasses with the "lightweight" attribute, but not necessarily prefer a "lightweight" sweater. We claim that users may have image aesthetics preferences like "symmetrical", thus it is important to extract image aesthetics attributes from the image. We used Prompt A: "Describe the image aesthetics independent of the item using keywords. Describe qualities including image composition, color scheme, lighting, balance, symmetry, contrast, texture, and overall visual harmony, feeling of the image using simple and common English keywords. Output as keyword1, keyword2, keyword3, ... keywordn" to extract the image aesthetics attributes in our experiment.

D. Items and Item Attributes Graph

We construct the Items and Item Attributes bipartite graph where each item is connected with the associated attributes from the image and the text. We define the graph with vertices consisting of items (V_i) and attributes (V_{ia}), as $G_{iia} = \{V_i, V_{ia}, E_{iia}\}$, where E_{iia} is the set of edges that connects V_i and V_{ia}.

E. Users, Image Aesthetic Attributes and Items Graph

We construct the Users, Image Aesthetic Attributes and Items Graph where each user is connected with the items from the users' history, and with the image aesthetic attributes associated with the items directly as shown in Fig. 2. We define the graph with vertices consisting of users (V_u), items (V_i) and image aesthetic attributes (V_{iaa}), as $G_{uiaai} = \{V_u, V_i, V_{iaa}, E_{uiaai}\}$, where E_{uiaai} is the set of edges that connect V_u and V_i, and the set of edges that connect V_u and V_{iaa}.

F. Graph Convolution Networks

We initialize e_{ia}^0, e_{iaa}^0, e_i^0, e_u^0 as trainable parameters to represent the embedding of item attributes, image aesthetic attributes, items and users respectively. The embedding of an item vertex in graph G_{iia} at the *k+1* layer is calculated by (1) to integrate neighbors' information.

$$e_i^{(k+1)} = \sum_{a \in N_{i_ia}} \frac{1}{\sqrt{|N_{i_ia}||N_{ia_i}|}} e_{ia}^{(k)} \quad (1)$$

, where N_{ia_i} is the set of item attributes that is connected to item *i* in graph G_{iia} and N_{i_ia} is the set of items that is connected to attributes *ia* in graph G_{iia}. Similarly, the embedding of an item attribute vertex in graph G_{iia} at the *k+1* layer is calculated by (2)

$$e_{ia}^{(k+1)} = \sum_{i \in N_{ia_i}} \frac{1}{\sqrt{|N_{ia_i}||N_{i_ia}|}} e_i^{(k)} \quad (2)$$

The embedding of an image aesthetic attribute vertex in graph G_{uiaai} at the *k+1* layer is calculated by (3)

$$e_{iaa}^{(k+1)} = \sum_{u \in N_{iaa_u}} \frac{1}{\sqrt{|N_{iaa_u}||N_{u_iaa}|}} e_u^{(k)} \quad (3)$$

Data Set	#Users	#Items	#User-item interactions	#Train	#Validation	#Test	Image	Text
IQON3000	3,568	23,363	459,146	367,317	45,914	45,915	One for each item	Brand, price, category, color, and a text introduction of 50-150 words.

TABLE II. EXPERIMENTAL DATASETS

TABLE III. EXPERIMENTAL EVALUATION RESULT

Models	IQON3000		
	Recall@50	***NDCG@50***	***Precision@50***
SAERS [12]	0.2202	0.1373	0.0178
MM-FRec [11]	0.3157	0.1732	0.0214
Proposed Method	0.2863	0.1577	0.0197

, where N_{iaa_u} is the set of image aesthetic attributes that is connected to user u in graph G_{uiaai} and N_{u_iaa} is the set of users that is connected to image aesthetic attributes iaa in graph G_{uiaai}. The embedding of a user vertex in graph G_{uiaai} at the $k+1$ layer is calculated by (4)

$$e_u^{(k+1)} = \sum_{iaa \in N_{u_iaa}} \frac{1}{\sqrt{|N_{u_iaa}||N_{iaa_u}|}} e_{iaa}^{(k)} + \sum_{i \in N_{u_i}} \frac{1}{\sqrt{|N_{u_i}||N_{i_u}|}} e_i^{(k)} \quad (4)$$

, where N_{i_u} is the set of items that is connected to user u in graph G_{uiaai} and N_{u_i} is the set of users that is connected to item i in graph G_{uiaai}.

Then, the final embedding of a user e_u is obtained by (5)

$$e_u = \sum_{k=0}^{K} \alpha_k e_u^{(k)} \quad (5)$$

, where K is a tunable parameter. In this paper, we experimented with K $\in$ {0,1,2,3,4,5} according to He et al. [2]. Similiarly we obtain final embedding of an item e_i by (6)

$$e_i = \sum_{k=0}^{K} \alpha_k e_i^{(k)} \quad (6)$$

G. Model Training

This sub-section explains the training details of the proposed model. We use the items a user interacted with in the data set as the positive samples and randomly select n_n items that the user does not interact with as the negative samples. The model parameters are trained to minimize the BPR loss of the model output compared with ground truth. L2 regularization is utilized to avoid model overfitting.

For each training user-item interaction, we forward N_{i_u}, N_{ia_i}, and N_{iaa_u} into the model f_θ as (7)

$$p = f_\theta(N_{i_u}, N_{iaa_u}, N_{ia_i}) \quad (7)$$

where p is the predicted probability of the future interacted item. Then, the network parameters θ are trained to maximize the predicted probability p compared with the ground-truth item.

5. EXPERIMENTAL EVALUATION

A. Datasets

To evaluate the effectiveness of the proposed method, we conducted experiments on the publicly available real-world dataset: *IQON3000* [4]

- **IQON3000** [4]**:** The IQON3000 dataset is collected from the fashion shopping website Iqon. The IQON3000 dataset comprises 308,747 outfits of 3,568 users, with a total of 672,335 items. To ensure a comparison with previous work by Song et al. [11], we filtered the dataset the same way, which is only items that are interacted by more than 10 users are left in the dataset. Because no actual train/test data are provided after splitting is provided by previous work, we randomly sample 80%, 10%, and 10% of the user-item interactions in the dataset to form the training set, validation set, and testing set. The statistics of the dataset after filtering are shown in Table II.

B. Experimental Setup

We introduce two baselines below. Since the dataset is the same, and the train, sample, and test ratio is the same we adopt the reported experimental results by Song et al. [11] for comparison.

SAERS [12]: A multi-modal-based approach that characterizes the preferences of users from both the visual and text attributes perspectives.

MM-FRec [11]: A multi-modal-based approach that recommends fashion items using both visual and semantic attributes.

C. Discussion

We show the experimental evaluation result in Table III. The experimental performance from the proposed methods

is worse in the three metrics than the previous best-performing method MM-FRec [11]. We plan to study the attributes we extracted from the LLVM more carefully and experiment with other prompts to improve the performance of the proposed method. Also, we plan to do the ablation study to analyze the effectiveness of the components in the proposed methods to further refine our methods. Because the proposed method can solve the cold-start problem, we also plan to test the performance on the unfiltered datasets, which MM-FRec can not experiment on because of the cold-start problem.

6. CONCLUSION

In this work, we propose a novel way for fashion recommendation specifically aiming to solve the cold start problem and the insufficiency of utilization of images. By extracting two types of fine-grained attributes from the images using LLVMs and text processing, our method presents a method that is reproducible and efficient to use for large-scale recommendation compared with traditional fine-grained attribute extraction. Combining both the attributes from the image and text, the item attributes graph proposed by our method is richer, bringing more explainable recommendations. The innovative user modeling attempts to directly consider the aesthetic preferences of users and reduces the noise from item-specific features. For future work, we plan to do an ablation study to analyze the effectiveness of the components in the proposed method. Also, we have not considered the item compatibility between fashion items in this paper, which may be further utilized to improve the accuracy of fashion recommendations.

ACKNOWLEDGEMENTS

This work was supported by JST SPRING Grant Number JPMJSP2128.

References

[1] Y. Cao, L. Yang, C. Wang, Z. Liu, H. Peng, C. You, P.S. Yu, Multi-task Item-attribute Graph Pre-training for Strict Cold-start Item Recommendation, ArXiv Preprint ArXiv:2306.14462. (2023).

[2] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. “Lightgcn: Simplifying and powering graph convolution network for recommendation.” Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 639-648, 2020.

[3] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural graph collaborative filtering,” in Proceedings of the 42nd International ACM SIGIR Conference on Research and development in Information Retrieval, pp. 165–174, 2019.

[4] X. Song, X. Han, Y. Li, J. Chen, X.-S. Xu, and L. Nie, “GP-BPR: Personalized compatibility modeling for clothing matching,” in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 320–328.

[5] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, “Hierarchical fashion graph network for personalized outfit recommendation,” in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159–168, 2020.

[6] X. Fan, F. Mo, C.X. Chen, C.H. Bai and H. Yamana. “Connectivity-aware Experience Replay for Graph Convolution Network-based Collaborative Filtering in Incremental Setting”, Proceedings of International Conference on Big Data Analytics, in press, 2024.

[7] W. Guan, F. Jiao, X. Song, H. Wen, C.-H. Yeh, and X. Chang, “Personalized fashion compatibility modeling via metapath-guided heterogeneous graph learning,” in Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp. 482–491, 2022.

[8] F. Mo and H. Yamana. "GN-GCN: Combining Geographical Neighbor Concept with Graph Convolution Network for POI Recommendation." Proc. of Int’l Conf. on Information Integration and Web, pp. 153-165, 2022.

[9] F. Mo and H. Yamana. "EPT-GCN: Edge propagation-based time-aware graph convolution network for POI recommendation." Neurocomputing, Vol. 543, pp. 1-15, 2023.

[10] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. "BPR: Bayesian personalized ranking from implicit feedback." arXiv:1205.2618, pp. 452-461, 2012.

[11] X. Song, C. Wang, C. Sun, S. Feng, M. Zhou, and L. Nie, “MM-FRec: Multi-Modal Enhanced Fashion Item Recommendation,” IEEE Transactions on Knowledge and Data Engineering, 2023.

[12] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, “Explainable fashion recommendation: A semantic attribute region guided approach,” in Proc. Int. Joint Conf. Artif. Intell., 2019, pp. 4681–4688.

[13] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, “Kgat: Knowledge graph attention network for recommendation,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 950–958.

[14] G. Team et al., Gemini: A Family of Highly Capable Multimodal Models. 2023.

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

2024年3月1日(金) 10:00 ~ 12:10 | T4-A オンライン（Zoom Events)

芸術の理解

座長:田中 克己(福知山公立大学)
コメンテータ:王 元元(山口大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:30
[T4-A-7-01] スケッチによるマスク不要の表情差分生成手法の提案

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

10:30 ~ 10:55
[T4-A-7-02] アート観賞イベントと連動したインタラクティブな街歩き型ストーリーリーダーの一検討

*徳丸 晴天[1]、畑 玲音[1]、松下 光範[1]、徳山 美津恵[1] (1. 関西大学)

10:55 ~ 11:20
[T4-A-7-03] CNNによるフェルメールの画風分析

*小田 稜子[1]、伊藤 貴之[1] (1. お茶の水女子大学理学部情報科学科伊藤研究室)

11:20 ~ 11:45
[T4-A-7-04] 深層距離学習による文字種とスタイルの特徴獲得とその特徴に基づく文字生成

*桑田 若菜[1]、三林 亮太[1]、谷 雅德[2]、大島 裕明[1] (1. 兵庫県立大学、2. 北陸先端科学技術大学院大学)

11:45 ~ 12:10
[T4-A-7-05] TREXIV: 時刻スタンプ付き高次元データのための対話的可視化に基づくトレンド抽出手法

*中村 栄太[1]、伊藤 貴之[2] (1. 京都大学、2. お茶の水女子大学)

スケッチによるマスク不要の表情差分生成手法の提案

佐々木嵩仁† 清 雄一† 田原 康之† 大須賀昭彦†

† 電気通信大学情報理工学域 〒 182-8585 東京都調布市調布ヶ丘
E-mail: †taketo.sasaki@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし 機械学習による画像生成は, プロンプトベースの拡散モデルにより品質を格段に向上させた. しかし, プロンプトによる細かい画像の操作を的確に行うのは未だに難しい. より柔軟な編集ができる手法としてスケッチを用いた画像補完タスクによって画像の編集を行うものが挙げられるが, 編集の際にスケッチとは別にどの部分を編集するかを示す画像であるマスクを与える必要があり, マスク内の情報の損失, 最適なマスクが不明確, 編集の度にマスクの提示が必要という課題がある. そこでマスクもモデルに推定させるアプローチに注目し, 既存手法をベースにさらなる精度の向上を目指した. 本研究では一枚のイラストから作成することが難しいの表情差分の作成を目的とした. マスク推定器に仮マスクと U-Net を採用し, 生成器の損失に PerceptualLoss を加えることで精度の向上を図った. 定量的評価指標による客観的評価と, アンケート調査による主観的評価の結果, 提案手法が既存手法に対して優位であることを示せた.

キーワード 画像編集, インタラクティブシステム, イラスト, 自動マスキング, スケッチベース

1 はじめに

1.1 背景

機械学習を用いて画像を生成, 編集させるという試みは様々な手法を用いて盛んに取り組まれている. 近年では Stable Diffusion [1] やそれを派生させた手法である NovelAI [2] や ControlNet [3] が開発され, 生成される画像のクオリティが著しく向上した. これらの手法はプロンプトと呼ばれるイラストの雰囲気を表す単語を複数指定して, それらをもとにイラストを生成することができるモデルである. しかし求める画像を得るためにはプロンプトを細かく指定する必要があり, 単語の選択に慣れなければならなず時間がかかってしまうという問題点もある. また, 生成された画像が必ずしもプロンプトを完璧に反映しているとは限らず, 細かい部分の操作を的確に行うのは未だに難しい.

既存の画像を異なる属性のものへ変換する簡単な手法として, 他にもスケッチを入力として用いる手法 [6] が存在する. この手法は入力として画像とスケッチの他に図 3 のようなどの部分を変更するかをグレースケールで表現した画像であるマスクを与える必要がある. しかし, 入力スケッチに対応する最適なマスク領域は不明確であり, また編集の度にマスクの提示をしなければならないという問題が生じる. さらに, マスクを用いると元画像のうちマスクで編集領域と指定した部分の情報は失われてしまい, 出力画像との間の一貫性が損なわれてしまう.

1.2 目的

ゲームや動画メディアに利用されるキャラクターを描いたイラスト群のことを立ち絵という. 立ち絵の例を図 1 に示す. キャラクターの体を構成するパーツを表現した複数のイラストから構成されており, これらを組み合わせて表情やポーズを作ることで場面に応じた様々な感情を表現することができる.

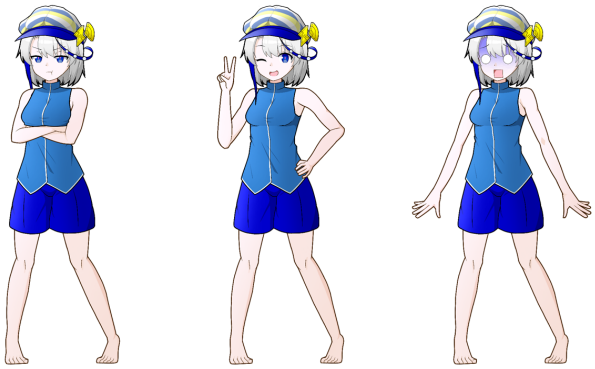


図 1 立ち絵の例

また立ち絵のうち表情の変化に焦点を絞ったものを表情差分という. 表情差分の例を図 2 に示す. 表情差分を用いて様々な感情を表現するためには, 目, 口, 眉毛などのイラストをそれぞれ複数枚描く必要があるため大変であるという課題がある. また, 表情差分が存在しないイラストから画風を残しつつ表情のみを変えた画像を得るのはイラストを描くのに慣れていない人にとって難しいという課題もある.

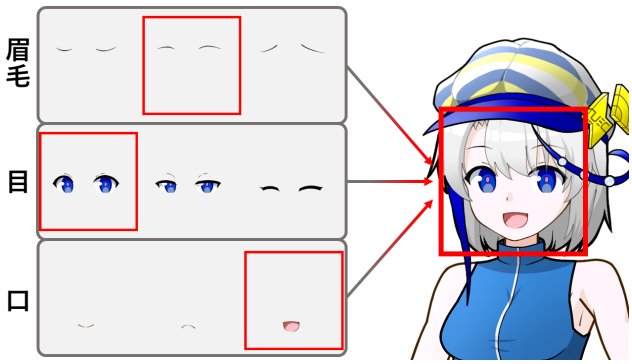


図 2 表情差分の例

これらの課題と前節で挙げた課題より本研究では, 一枚のイラスト画像から入力にマスクを用いないでイラストの表情編集を可能とするモデルの構築を目的とする.

2 関連研究

2.1 画像変換

画像変換は機械学習を用いるタスクのうち主要なものの一つで, あるドメインの画像を別のドメインの画像へ変換することを目的としている.Isola ら [4] は通常の画像とセグメンテーションマップやエッジマップといった条件付けされた画像との間の変換を行うフレームワークである Pix2Pix を提案した.Zhu ら [5] はサイクル一貫性制約を用いることで, ペア画像データセットを必要としない教師無し学習での画像変換を目的とした CycleGAN を提案した. これらの手法はスケッチからカラー画像へのドメインの変換を行うことはできるが, 形状の変化を伴う画像の編集を行うことは難しい. 一方で本研究はスケッチを入力として形状の変化を伴う画像変換をすることに着目している. さらにスケッチは画像全体を描く必要はなく, 操作したい要素の周辺部だけを含んでいればよい.

2.2 スケッチによる画像変換

スケッチに基づいてマスキングされた画像を補完するタスクは, 条件付き画像修復タスクの一種として扱われている.Yu ら [6] は人物の顔や風景の画像をマスキングすることで欠損させ, スケッチを用いてそれを補完することで画像変換を行う DeepFill-v2 を提案した.Gated Convolution と呼ばれる特殊な畳み込み層により, 有用な特徴を動的に選択することを可能にした.Portenier ら [7] は, マスク, スケッチに加えて色の指定をする部分的なカラー画像も入力として受け取ることで, 形状だけでなく色もユーザーの意図を反映させ, より柔軟な画像変換を行うことができる FaceShop を提案した. これらの手法には入力としてマスクを要求するが, 本研究ではマスクを与えずにスケッチのみで画像変換を行うタスクを追求し, ユーザーによるスケッチからマスクを推測するようなネットワークを含むようなモデルを構築した.

2.3 セグメンテーション

セグメンテーションとは, 画像をいくつかのオブジェクト領域に分割するタスクのことを指す. このタスクを解決するニューラルネットワークを使用したモデルとして,Vijay ら [8] の SegNet が挙げられる.Olaf ら [9] は, エンコーダーとデコーダーの同じ深さの層を接続するスキップ接続を取り入れた U-Net を提案した. スキップ接続によって入力画像内の物体の位置情報を保持し, より高精度の予測を実現した. 本研究では, 入力画像のマスク領域を推定するというタスクを, 入力画像をマスク領域とそうでない領域に分割するセグメンテーションタスクと捉え, 通常のセグメンテーションタスクと同様に精度の向上ができると考え U-Net のスキップ接続を手法に取り入れた.

2.4 Sketch Edit

マスクを用いない画像編集を目的とした手法として,Yu ら [10] の Sketch Edit が挙げられる.Sketch Edit では編集したい画像と, その画像をどのように編集するかを大まかに示したスケッチを入力として与えることで, そのスケッチをもとに画像を編集し出力する.Sketch Edit は, 画像とスケッチから最適なマスク領域を推定するマスク推定器, 入力画像のうち編集する部分から特徴ベクトルを抽出するスタイルエンコーダー, 推定されたマスクによってマスキングされた画像をスケッチをもとに修復する生成器の三つのコンポーネントから構成されている.Sketch Edit の具体的な構造は図 3 のようになる.

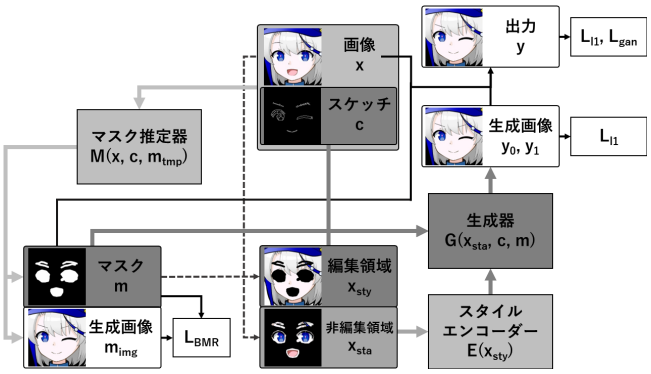


図 3 SketchEdit のモデル構造

この手法は自己教師あり学習でモデルを訓練しているが, ターゲット画像を入力してターゲット画像を得るような学習をしてもマスク推定器が推定するマスク画像は全てのピクセルの値が 0 であるという予測をするのは自明である. そこで入力するターゲット画像をランダムにワーピングしてからモデルに与えることでこの問題を回避した. しかし, ワーピングした画像を使って学習するためマスク推定器は単純にワーピングによる歪みを検出するように学習してしまう. このような動作を防ぐためにマスク推定器に双方向の画像再構成損失 BMRLoss が与えられた.

カラー画像 x を移動量 f に基づいてワーピングした画像を f(x) とする. 同様にワーピングされたスケッチ画像 c を f(c) とする. マスク推定器にデコーダーを追加し, マスク推定の学習に並行して f(x) と c から x,x と f(c) から f(x) を推定するタスクを学習させる. つまりマスク推定器はマスク画像とは別にカラー画像を出力する.$\bar{M}(x,c)$ を推定画像,$\hat{M}(x,c)$ を $\bar{M}(x,c)$ と入力をマスクを用いてブレンドした画像とする.BMRLoss は次のように求めることができる.

$$L_{BMR} = ||x_a - \bar{M}(x_b, c_a)||_1 + ||x_a - \hat{M}(x_b, c_a)||_1 + ||x_b - \bar{M}(x_a, c_b)||_1 + ||x_b - \hat{M}(x_a, c_b)||_1 \quad (1)$$

モデルはこの BMRLoss と L1Loss, 識別器を用いた GANLoss を合わせて学習していく.L1Loss は生成器の二段階の出力画像 y_0,y_1 および最終的な出力 y と入力画像 x の L1 距離を求めることで計算され, 式は以下のようになる.

$$L_{l1} = ||x - y_0||_1 + ||x - y_1||_1 + ||x - y||_1 \quad (2)$$

GANLoss には HingeLoss を用いており, マスク推定器と生成器の出力をマージした y を識別機に与えるため二つのネットワークが協力して違和感のない出力を得られるように学習していく. 識別器の出力を $D(y, c)$ とすると以下のように書くことができる.

$$L_{gan} = ReLU[1 - D(y, c)] \tag{3}$$

モデルの目的変数は前述の三つの Loss の合計である.

$$L_{total} = L_{l1} + L_{gan} + L_{BMR} \tag{4}$$

この手法では編集前後のペア画像データセットを用意することの難しさからワーピングなどのアプローチをしているが, イラストの表情差分を操作するタスクの学習においてはペア画像を用意することは比較的容易であるためより高品質な画像を教師データとして使用することができる. その反面, 強調された輪郭線や各パーツの不自然な形状などイラスト特有の表現方法を考慮する必要がある.

3 提 案 手 法

3.1 提案手法の概要

本研究ではマスクを用いないスケッチによるイラストの編集を実現するために Sketch Edit をベースラインとして設定し,U-Net やキーポイントを基に生成される仮マスクを用いて精度の向上を図った. モデルの全体構造は図 4 のようになる.

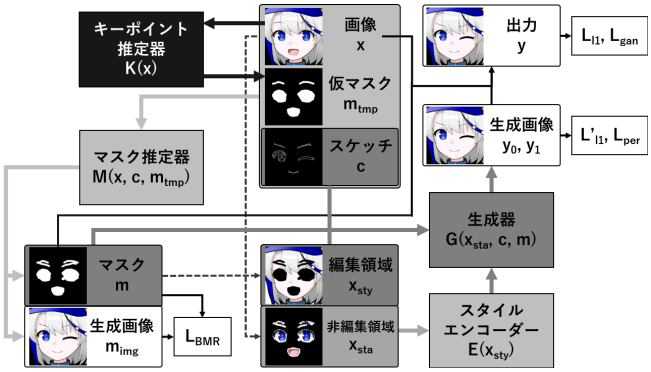


図 4 提案手法のモデル構造

モデルは大きく分けて四つのコンポーネントから構成されている. 一つ目はイラストからキーポイント特徴を抽出するキーポイント推定器である. 二つ目はイラストとユーザーによるスケッチからどの領域を編集するかを推定するマスク推定器である. 三つめは編集領域から出力画像の生成のために有用な特徴を抽出するスタイルエンコーダーである. 四つ目はマスク推定器の出力によってマスキングされたイラストをスケッチをもとに補完する生成器である. これらのコンポーネントのうち, キーポイント推定器のみ事前に学習をさせ, その他のコンポーネントのネットワークを共同に学習を進める.

3.2 キーポイント推定器

キーポイント推定器 $K(x)$ はイラスト x から複数のキーポイントを推定し, 画像の大まかな構造情報を抽出する. キーポイント推定器はアーキテクチャに依存しないが, イラストの顔に対応させるためにあらかじめイラストを用いて学習をさせておく必要がある. 本研究では学習が容易でかつ動作が軽量であるため mmpose [11] のモジュールを採用した. キーポイントは眉, 目, 口の構造を大まかに把握するように合計で 42 つ割り振った.

また, 推定されたキーポイント群のうち同じパーツを線で結んだり, 中を塗りつぶしたりすることで顔のパーツの位置のみを示す画像を得ることができる. この画像はイラストのうち編集される候補となる領域を大まかに表現していることから仮マスク m_{temp} と呼ぶ. 仮マスクは推定されるマスクと相関があるためマスク領域の推定の学習を促進することができる. キーポイントの割り振り方と, キーポイントから得られる仮マスクの例を図 5 に示す.

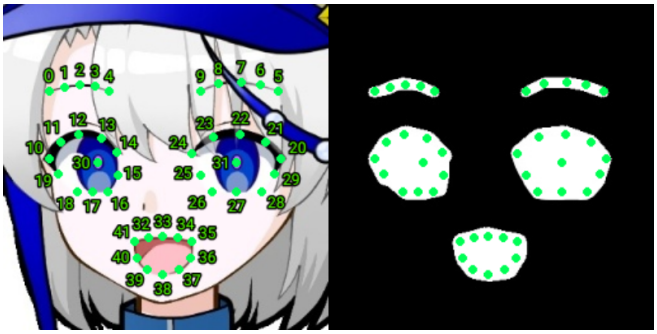


図 5 キーポイントの配置 (左) と仮マスク (右)

3.3 マスク推定器

マスク推定器 $M(x, m_{temp}, c)$ は編集したいイラスト x, 仮マスク m_{temp}, 編集したい部分をどう編集したいかを示すスケッチ c の三つの画像を受け取り, 編集に最適なマスク $m = M(x, m_{temp}, c)$ を出力する. マスク推定器のモデル構造は図 6 のようになる.

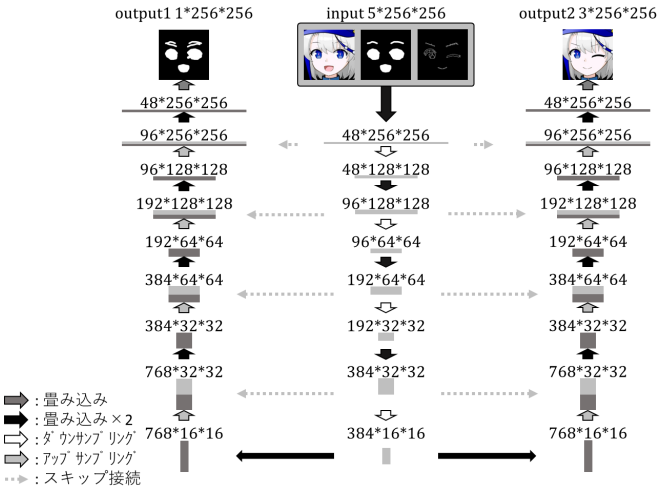


図 6 マスク推定器

マスクは各ピクセルの値が 0 から 1 の範囲であるグレースケールマップである. マスクの各ピクセルはイラストの対応する位置のピクセルの値を修正するべきかどうかを示しており、イラストとマスクの各要素の積を求めることで入力画像を修正するべき領域 $x^{sty} = x \odot m$ とそうでない領域 $x^{sta} = x \odot (1-m)$ に分割することができる.

マスク領域の推定を, 編集領域と非編集領域に分割と捉え, セグメンテーションタスクの精度の向上を実現した U-Net をマスク推定器に採用した.U-Net はスキップ接続により入力画像の構造情報を深層まで保持することで精度を高めた.Sketch Edit に倣って,BMRLoss を求めるためにデコーダーが二つに分岐している. 本研究は Sketch Edit とは異なり入力画像とターゲット画像が別の画像であるペア画像データセットを用いて学習をさせるためワーピング処理は行わないが, 各パーツの多様な変化に対して一貫性のあるマスク領域を推定させるために BMRLoss を使用した. 本手法における BMRLoss は次の式のようになる.

$$L_{BMR} = ||x_a - \bar{M}(x_b, c_a)||_1 + ||x_a - \hat{M}(x_b, c_a)||_1 + ||x_b - \bar{M}(x_a, c_b)||_1 + ||x_b - \hat{M}(x_a, c_b)||_1 \quad (5)$$

ここで x_a と x_b はイラスト,c_a と c_b はスケッチを表し, 学習時にマスク推定タスクと並行して a から b, b から a を推測するタスクを学習していく.

3.4 スタイルエンコーダー

スタイルエンコーダー $S(x^{sty}, m)$ はイラストの編集領域とマスクからスタイル情報 v^{sty} を抽出する. 複数畳み込み層の後にグローバル最大プーリングを行うことで d 次元のスタイルベクトル $v^{sty} = S(x^{sty}, m)$ を得る. さらにこのベクトルを画像のサイズ $(h \times w)$ に合わせてリピートすることでスタイル特徴 $\hat{v}^{sty}_{(d \times h \times w)}$ を取得する.

3.5 生 成 器

画像生成器 $G(x_{sta}, m, c)$ はマスク推定器によって推定されたマスク m から得られる非編集領域画像 x^{sta} をマスキングされた画像として扱い, それをスケッチ c とスタイルエンコーダーから得たスタイル特徴 $\hat{v}^{sty}$ を手掛かりに補完する. マスク推定器のモデル構造は図 7 のようになる.

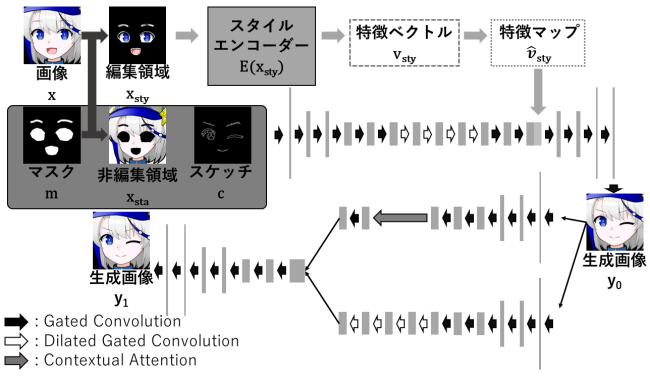


図 7 生 成 器

特定のアーキテクチャに依存しないので, 任意の画像補完モデルを使用することができる. 本研究では Sketch Edit と同様にシンプルでかつ軽量である DeepFill-v2 のネットワークアーキテクチャを取り入れた.DeepFill-v2 の生成器は粗い画像 y_0 を生成するステージ G_0 と,y_0 緻密な画像 y_1 を生成するステージ G_1 の二段階で構成されている. スタイルエンコーダーによって得たスタイル特徴は, 粗い画像を生成するステージ内でデコーダーに入力する前にボトルネックの出力と結合される. 最終的な出力画像 y は y_1 とイラストをマスクをもとにブレンドした画像 $y = x \odot (1-m) + y_1 \odot m$ である.

Sketch Edit では生成器の学習に L1Loss と GANLoss の二つの目的変数を用いていたが, 本研究ではこれに加えて学習済み VGG モデルから得られる特徴マップを用いた PerceptualLoss を新たに目的変数として取り入れた.L1Loss がピクセル単位での比較をするのに対して,PerceptualLoss はより大域的な比較をするため, スタイルを学習でき, よりくっきりとした出力が得られるようになる. 学習済み VGG モデルによって与えられる画像 x の特徴マップを $P(x)$ とすると,PerceptualLoss は以下の式で得ることができる.

$$L_{per} = ||P(x) - P(y)||_1 \quad (6)$$

3.6 目 的 関 数

前節で説明した BMRLoss に加えて, マスク推定器に与える損失関数として L1Loss と GANLoss を以下のように計算した.

$$L_{l1} = ||x_{gt} - y||_1 \quad (7)$$

$$L_{gan} = \frac{1}{2} E[(1 - D(y, c))^2] \quad (8)$$

x_{gt} は正解画像，y は生成器の出力とイラストを推定したマスクをもとにマージしたモデル全体の最終的な出力画像を指す.GANLoss の計算には HingeLoss と同様に安定した学習をすることができる LSGAN を用いた. まとめると, マスク推定器の目的関数は以下のようになる.

$$L_{ME} = \lambda_0 L_{l1} + \lambda_1 L_{gan} + \lambda_2 L_{BMR} \quad (9)$$

生成器の L1Loss を以下のように計算した. 二段階のステージそれぞれの出力を正解画像と比較する.

$$L'_{l1} = ||x - y_0||_1 + ||x - y_1||_1 \quad (10)$$

前節の PerceptualLoss と合わせると生成器の目的関数は以下のようになる. なお,GANLoss はマスク推定器のものと同様である.

$$L_{IG} = \lambda_0 L'_{l1} + \lambda_1 L_{gan} + \lambda_3 L_{per} \quad (11)$$

ここで,$\lambda_0, \lambda_1, \lambda_2, \lambda_3$ は損失の影響の強さを調整するためのハイパーパラメータである.

4 実　　験

4.1 データセット

学習に用いるイラストは X [14] 及びニコニコ静画 [15] から収集した.X から収集した画像は mmpose によるキーポイント推定をイラストに対応させるための再学習に使用した. 合計で 1000 枚のイラストを収集し, そのうち 500 枚を顔の位置推定モデルの学習に用いて, もう 500 枚を顔のキーポイント推定モデルの学習に用いた. ニコニコ静画からは立ち絵素材として配布されているイラストをペア画像データセットとして使用した.1 キャラクターにつき 64 枚分の表情差分を収集し,1 エポックごとに 64 枚分のうち 2 枚分の画像を無作為に選択してペア画像とした. また,1 枚分の表情差分に対してイラスト, スケッチ, 仮マスクの 3 枚がセットになっている. 合計で 100 キャラクター ×64 種類 ×3=19200 枚の画像を収集した.

評価に用いるイラストは学習データセットに含まれていない 20 キャラクターの表情差分画像を収集した. 合計で 20 キャラクター ×2 種類 ×3=120 枚である.

ペア画像データセットに含まれるイラストは, 下処理としてまず mmpose に内蔵されている事前学習済みの顔の位置推定モデルを用いてあらかじめ顔周辺の領域のみが含まれるようにトリミングをした. 次に mmpose によって推定されたキーポイント群をもとに仮マスクを作成し, イラストと仮マスクのサイズを 256 × 256 になるようにリサイズをした. さらに Canny 法によって得られたイラストのエッジ画像から, 仮マスクをもとに眉, 目, 口周辺のエッジのみを取り出すことで学習に用いるスケッチを得た. データセット作成のプロセスをまとめたものを図 8 に示す.



図 8　データセット作成プロセス

4.2 パラメータ設定

目的関数の計算に用いる係数 λ は, それぞれ $\lambda_0 = 10, \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 1.0$ と設定した.epoch 数は 1000 で学習を行い, 学習率は初期値を 10^{-4} と設定した. 学習率は 100epoch 分の学習が終わるごとに 0.8 倍に低下させることで学習の収束を促した.

4.3 評 価 指 標

定量的な評価指標として,L1 損失,PSNR,SSIM を使用した.L1 損失は正解画像と出力画像の平均誤差の絶対値であり, 値が小さいほど正解画像に等しいと言える.PSNR(ピーク信号対雑音比) と SSIM(構造的類似性) は画像がどれほど類似しているかを評価する指標で, 値が大きいほど類似していると言える.

4.4 定量的評価

マスク推定器に U-Net アーキテクチャを用いない手法を既存手法として設定した. また提案手法から PerceptualLoss を取り除いた手法 (w/o PLoss) と, 仮マスクを取り除いた手法 (w/o m_{temp}) とも比較した. これら四つのモデルによる出力を評価指標を用いて評価した. 各モデルの出力を図 9, 各手法の L1 損失,PSNR,SSIM の平均値は表 1 のようになった.U-Net アーキテクチャを採用した手法が L1 誤差と SSIM で既存手法を上回った.PSNR は既存手法が僅かだが一番良いスコアだった.



図 9　各手法の生成画像

表 1　各手法の平均スコア

手法	L1↓	PSNR↑	SSIM↑
提案手法	0.02817	71.02	**0.9047**
既存手法	0.02896	**71.40**	0.8984
w/o PLoss	**0.02621**	71.19	**0.9047**
w/o m_{temp}	0.02692	70.53	0.9033

4.5 主観的評価

20 代の男女 24 人の被験者を募ってアンケートを実施し, モデルを主観的に評価した. アンケートは, 客観的評価で用いた四つのモデルの出力画像を並べて一番正解画像に近い画像を答えてもらい相対的な評価をするパートと, 提案手法による出力が正解画像にどれ位類似しているかを五段階で評価してもらい絶対的な評価をするパートの二つのパートで構成した. 各パートでテストデータの半分の出力を提示した. また根拠も任意で回答してもらった. 相対的な評価のパートの結果を表 2, 絶対的な評価のパートの結果を表 3 に纏めた. 表 2 によると, 提案手法と仮マスクなしの手法がいい評価をされていることがわかる. また表 3 によると 2 から 4 と評価されることが多かった. 一部 1 や 5 が見られたが、頻度は 5 の方が多かった.

表 2　相対的な評価の結果

キャラクター	提案手法	既存手法	w/o PLoss	w/o m_{temp}
1	37.5	0.0	0.0	**62.5**
2	4.2	0.0	8.3	**87.5**
3	4.2	20.8	0.0	**75.0**
4	12.5	**70.8**	8.3	8.3
5	8.3	41.7	4.2	**45.8**
6	37.5	4.2	0.0	**58.3**
7	25.0	16.7	0.0	**58.3**
8	16.7	20.8	4.2	**58.3**
9	33.3	29.2	0.0	**37.5**
10	20.8	16.7	0.0	**62.5**
平均	20.0	22.1	2.5	**55.4**

表 3　絶対的な評価の結果

キャラクター	1	2	3	4	5
11	0	0	12.5	**54.2**	33.3
12	0	25.0	**41.7**	33.3	0
13	0	29.2	**45.8**	20.8	4.2
14	0	4.2	20.8	**50.0**	25.0
15	8.3	**54.2**	12.5	20.8	4.2
16	0	20.8	29.2	**33.3**	16.7
17	0	4.2	25.0	**45.8**	25.0
18	20.8	**33.3**	20.8	25.0	0
19	4.2	25.0	12.5	**45.8**	12.5
20	0	16.7	20.8	**58.3**	4.2
平均	3.3	21.3	24.2	**38.7**	12.5

絶対的な評価を行うアンケート内でなぜその回答を選択したかの理由を任意回答で記述してもらったもののうち一部を抜粋して以下にまとめた.

- 全体の類似度について
  - ノイズのような縦筋さえなければかなり同じだと感じた
  - ノイズがあるが, 概ね一致している
  - 全体的に線が掴めないと思った
  - 雰囲気がかなり似ていると感じた
- 形状について
  - 概ね一致しているが, ハイライトや眉の太さなどが異なる
  - 少し/かなりぼやけている
  - 目元口元に線が多く, パッと見であまり何が出ているのか分からなかった
  - おおよそは似ているが, 目の形状が不自然だったり, 眉毛がぶれている点が気になった
  - 目がぶれている以外はかなり良いと思った
  - ぱっと見で似ているなとは感じたが, 筋が気になった
- 色について
  - 目の色が濁っていて表情が異なって見える
  - 目の色以外は似ている.
  - かなりよく思ったが, 目の色が変わってしまった点だけ気になった

5 考　　察

5.1 定量的評価の分析

定量的評価の結果,U-Net アーキテクチャを採用した手法が L1 誤差と SSIM で既存手法を上回り, 既存手法が PSNR で若干高いスコアを算出した. 既存手法によって生成された画像は目周辺の編集は全体的にうまくできているが, 口周辺の編集において一部欠損が生じていることがわかる. このような欠損は主に広い範囲を伴うような編集を行う際に見られ, 特に口の編集の際によくみられるケースである. それに対して U-Net アーキテクチャを採用した三つの手法はどの部位も大まかに編集ができているがマージの際に縦線のようなノイズが生じているということがわかる. なぜ既存手法の方が PSNR において僅かに高いスコアを得られたのかを考えてみると,PSNR は大域的で僅かな劣化と局所的で大きな劣化の区別ができないためスコアに大きな差が出なかったと考えられる. 反対に SSIM は PSNR よりも人間の知覚に近い評価をすることができるため局所的で大きな劣化をしてしまった既存手法よりも大域的で僅かな劣化をした手法を高く評価したと考えられる. また,L1 誤差からも U-Net アーキテクチャを用いた手法が既存手法よりも優位性があるといえる.

以上を踏まえると,U-Net アーキテクチャを用いると広い範囲での幅広い形状の変化に対応することができるようになるという優位性があるが, ノイズが生じてしまうという問題点があるといえる.

5.2 主観的評価の分析

相対的な評価の結果を見てみると, 提案手法と仮マスクなしの手法がいい評価をされていることがわかる. それに対して PerceptualLoss なしの手法は全体的に評価されなかった.PerceptualLoss なしの手法で生成された画像はイラストが歪んでいたり滲んでいたりしている. 特に眉毛や閉じた口などの線で表現されるパーツをうまく生成できていない. また目や口の色が他の手法に比べると明らかに異なっている. このことから PerceptualLoss は定量的な評価の数値上ではあまり影響を持たないように見えるが, 実際の見た目において違和感を軽減する役割を果たしているといえる.

仮マスクなしの手法の方がより高く評価されているため, 一見すると仮マスクを入力に使用しないほうがいいのではないかと思えるが, 眉毛に注目すると仮マスクなしの手法では出力に入力の眉が残ったままになってしまっていることがわかる. このことから仮マスクはイラストで見られる極端な位置の変化を伴う編集に対応するために有効であるといえる.

絶対的な評価の結果を見てみると, 平均以上の高い評価をしてもらうことができたといえる. コメントから, 主観的に見ても縦線のノイズの存在が品質を大きく損なわせているとわかる. 他にも目や眉毛の形状についてぼやけていたり形がおかしいという意見があった. これはデータセットの作成においてエッジを抽出してスケッチの代わりとして用いたため目や髪のグラデーションから不自然な位置にエッジが発生してしまい, エッジの形状が少し歪になってしまったからだと考えられる. また目の色について言及されることも多かったが, 現状のモデルでは色も含めた正確な編集は難しく改善の余地があると考えられる.

5.3 生成された画像の分析

図 10 は提案手法における生成器の出力 y_1 とマスクとマージすることで得られる最終的な出力 y を正解画像と比較したものである. これを見ると,y_1 は着色は異なるが形状は類似した画像を生成することに成功していることがわかる. つまり, マスクの出力がマージをするにあたって不十分であり改善の余地があるとわかる. また, 目の色が紫や茶色が多いのはデータセットの偏りや, スタイルエンコーダーがまつ毛の色を強く反映させたからであると考えられる.



図 10　出力画像のマージ前後比較

6 おわりに

本研究ではマスクを用いない画像編集を目的とした手法である Sketch Edit をベースに, マスクを用いずに表情差分の作成を試みた. 提案モデルによって推測されたマスクは, 既存モデルと比較してイラストの大域的な編集を可能にした. また, 生成された画像を定量的・主観的に評価することで提案モデルによる出力画像の品質が既存手法よりも優れていることが示せた. 今後の展望として, マージの際に縦線のノイズが含まれるという問題を解決したより正確なマスクの推定を目指すとともに, 簡易的なスケッチでも編集ができる手法や, キャラクターに合わせて口の中や瞳の着色をより正確に行うために入力に色情報を追加したモデルを作成するアプローチを追求する.

謝辞 本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです.

文　献

[1] Stable Diffusion, https://stablediffusionweb.com/
[2] NovelAI, https://novelai.net/
[3] ControlNet, https://stablediffusionweb.com/ControlNet
[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017, pp. 5967-5976
[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 2242-2251
[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas Huang. Free-Form Image Inpainting With Gated Convolution. *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* 2019, pp. 4470-4479
[7] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, Matthias Zwicker. FaceShop: Deep Sketch-based Face Image Editing. *ACM Transactions on Graphics,* 2018, pp. 1–13
[8] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligences,* 2017, pp. 2481-2495
[9] Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI,* 2015, pp 234–24
[10] Yu Zeng, Zhe Lin, Vishal M. Patel. SketchEdit: Mask-Free Local Image Manipulation With Partial Sketches. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2022, pp. 5951-5961
[11] mmpose, https://mmpose.readthedocs.io/en/latest/
[12] Shuying Liu, Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR),* 2015, pp. 730-734
[13] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, Stephen Paul Smolley. Least Squares Generative Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 2813-2821
[14] X, https://twitter.com
[15] ニコニコ静画, https://seiga.nicovideo.jp

アート観賞イベントと連動した インタラクティブな街歩き型ストーリーリーダーの一検討

徳丸　晴天† 畑　玲音†† 徳山美津恵† 松下　光範†

† 関西大学総合情報学部　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
†† 関西大学大学院総合情報学研究科　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
E-mail: †{k003022,k223167,toku_san,m_mat}@kansai-u.ac.jp

あらまし　近年，街の活性化を図ることを目的としたアートイベントが開催されている．これまでも，アートイベントに訪れる人に回遊を促すシステムが提案されている．しかし，ユーザはシステムを利用することが目的であり，アート観賞に関心が向かないことが懸念される．そのため，アートの観賞体験とユーザの回遊の双方を促すシステムが求められる．本研究では，街のアート観賞に関心を向けつつ回遊を誘発するために，アート観賞を組み込んだ物語の観賞を足掛かりにしたシステムを提案する．街中に壁画などのアートが点在する十三という街を対象とした「アートのある場所に行くことにより，物語の続きを読み進めることができる街歩き型ストーリーリーダー」を作成した．作成したシステムを用いて，ユーザ調査を実施しシステムの検討を行う．

キーワード　インタラクティブシステム，街歩き支援，モバイルアプリ

1 はじめに

まちや地域の魅力や現状を伝える手法として，「まちや地域を対象にしたアートイベント」が注目されており [9] [10]，これまで盛んに行われている [11] [12].「まちや地域を対象にしたアートイベント」は，アートへの関心を足掛かりとして集客を図るイベントであり，来場者を対象にアート作品の観賞や体験を通じて地域資源に目を向けるきっかけを提供する（e.g., 瀬戸内国際芸術祭 [11], とよはし都市型アートイベント sebone [4]）．しかし，来場者は「日常的に人が多く集まり賑わいのある街路」や「目のつきやすいスポット」に影響を受けて観賞経路を選択するため，巡られるスポットが限られてしまう [8].

これを解決する一つの手段として，「スタンプラリー」や「アプリケーションなどのガイドシステム」といった「普段目につかないようなスポットも巡らせ，街に訪れた人の回遊行動を誘発するシステム」が考案されており，効果的であるとされている [6]．しかし，これらのシステムでは，回遊を促すために提供されるコンテンツにユーザの関心が向き，本来の目的の優先度が低下する懸念がある．例えば，各地にあるアートなどのスポットの回遊を狙ったスタンプラリーでは，スタンプを集めること自体に関心が向き，そのスポットでアートなどのコンテンツを観賞・体験すること（以下，コンテンツ体験と記す）などの本来の目的がおろそかになる恐れがある．そのため，コンテンツ体験に関心を向けつつ回遊を誘発する必要がある．

本研究では，コンテンツ体験に関心を向けつつ回遊を誘発するシステムを検討する．

2 先行研究

これまでに，コンテンツ体験や回遊などの人の行動を誘発するシステムはいくつか提案されている．注目度が低いコンテンツに関心を向ける手法として，大杉らは，動物園でのユーザの周遊行動に変化を与えることで，観察対象にされることが少ない動物に関心を向ける方法を提案している [5]．この手法では，動物園の案内アプリケーション上で動物図鑑を完成させる機能を提供している．これにより，施設内や地域内の一部のエリアなどの統制が可能な空間内において，周遊行動の誘発が可能であることが示唆された．

岩尾らは，複合型商業施設への来訪者に対して周遊行動を誘導することを目的として，スマートフォンを用いた宝さがしゲームのアプリケーションを開発した．ユーザは宝の在処を示すヒントと宝に近づくと得られる通知のヒントを手がかりに，様々な店舗に隠された宝を探し，獲得するためにミニゲームを行う．このアプリケーションの利用を通じて，子どもと一緒に参加した大人に対して施設全体を周遊する行動を促せることが報告されている．

安尾らは，位置情報を利用したゲームでは，多種多様な土地におけるゲームの成立のために，各スポットの役割が制限されることから，そのスポットである必然性を失ってしまう問題の解決案として，予め決められた場所でしか読めない，ユーザがいる場所が舞台となっている小説を提供するアプリケーションを作成した [13]．位置情報を活用したコンテンツは周遊行動の誘発に有用であるとした一方で，単なる位置情報を用いたアプリケーションでは，人の意識はあくまでアプリケーションに終始する可能性があるため，物語コンテンツによる実環境に対する理解の誘引に着目した．しかし，物語コンテンツによる誘引

によって回遊がどのように誘発されるかについて，作成されたアプリケーションを用いた実験は行われていない．

物語の舞台を訪れることは「文学ツーリズム」として知られており，物語がスポットへの関心を呼び起こす事例が報告されている [1]．とりわけ日本では，アニメーションやコミックなどの大衆文化コンテンツをきっかけとした文学ツーリズムは「コンテンツツーリズム」と呼ばれ，地域のブランディングやプロモーションに広く活用されつつある [2]．こうした現状を考慮すると，環境への興味を喚起させる方法として物語コンテンツを用いることは合理的であるといえる [13]．

そこで本研究では，現実のコンテンツ体験を組み込んだ物語の観賞により，コンテンツ体験に関心を向けつつ回遊を誘発するシステムを提案する．ただし，コンテンツ体験をアート観賞として行った．アート観賞とは，アートをじっくり観察することである．提案システムでは，「実在するアート観賞を組み込んだ物語」の観賞により，街にあるアートを観賞することに関心を向けつつ，回遊を誘発することを目指す．

3　十三について

今回対象とする地域は，大阪市淀川区の南西部に位置する阪急十三駅を囲む周辺地域の十三地区（以下，十三と記す）である．十三は，大阪を代表する繁華街・オフィス街である梅田から近く，大阪梅田駅と並び阪急電鉄の主要な 3 路線（宝塚線，神戸線，京都線）の結節点となっており，乗降客も多いという好立地である一方で，現在も昭和の良さが残る多様性に富んだ街である．近年ではマンション建設なども進み，街への新規流入者も多く，変化の兆しも見られるエリアとして認識されつつあるが，十三の現在のイメージとしては，依然として乗り換え駅としての知名度に留まっており，ブランドイメージは歓楽街に代表されるような昔ながらの街という印象を持たれることも多い．

現在，阪急電鉄がなにわ筋・新大阪連絡線を 2031 年春に開業する方向で進めており，ハブとなる十三駅はこれまで以上に重要なターミナル駅になると予想される．このことから，街への投資が進んでおり，今後，街の活性化が期待されている [1]．こうした変化を受ける形で，十三においても新たなブランドイメージの形成や既存のブランドイメージの変容を目的とした，街に点在するアートを用いた活動が盛んに行われている．その具体的な活動としては，「十三アートフェス」と「淀壁」が挙げられる．

十三アートフェスとは，淀川区十三エリアを中心に，淀川区のアートや文化活動の周知を目的として活動している NPO 法人「淀川アートネット」[2]が中心となって 2021 年から現在のスタイルで毎年実施しているイベントである．十三アートフェスは十三の街や店舗など，今までの自分の生活において知らなかった場所を「行ったことある場所，知っている場所，馴染みの場所，再び行く場所」に変えることを目的としている．特に，十三を知らない人やあまりポジティブなイメージを持っていない人が，アートフェスを通じて十三に足を運び，様々なアートを体感することや，いくつもの参加店舗や催しを巡ることによって，十三の街の魅力を知ってもらうことに力点を置いている．十三アートフェスは，誰でも作品を応募可能である特徴から，プロからアマチュアまで幅広いアーティストがアートを展示している [3]ため，多種多様なコンテンツが存在することが魅力の一つである．また，店舗における作品の展示から始まったイベントであるため，現在も多くの店舗が参加するイベントである [4]．




図 1: 淀壁
（左アーティスト：kuua，右アーティスト：BAKIBAKI）

十三地区のシンボル的なコンテンツとして「淀壁」（図 1 参照）を始めとするパブリック・アートが存在する．「淀壁」とは，大阪万博が開催される 2025 年に向けて淀川区に壁画を増やしていく非営利プロジェクトであり，壁画を通じた地域活性化を目指している．2021 年 3 月には，淀川区役所の広報的支援を受け “コロナ禍における医療従事者への敬意” をテーマに掲げ，アーティストの BAKIBAKI 氏 [5] がフローレンス・ナイチンゲールをモチーフにした壁画を，WALLSHARE [6] のサポートのもとボランティアで制作した．このナイチンゲールの壁画を皮切りに，淀川エリアにさまざまなアーティストの壁画を増やすべく淀壁実行委員会を発足し，クラウドファウンディングで壁画制作における資金調達を行い，地域との連携を図りつつ，同じく橋向こうの下町が壁画やストリートアートで栄えた，NY のブルックリンに見立て “十三ブルックリン化計画” として各種メディアから注目を集めている [7]．

本研究では，淀川区十三エリアに点在する淀壁などのアートの観賞及び回遊の誘発を試みる．

1：https://www.nikkei.com/article/DGXZQOUF043DJ0U3A600C2000000/ (2024/03/11 確認)
2：https://www.npo-homepage.go.jp/npoportal/detail/113001565 (2024/03/11 確認)
3：十三アートフェスの運営者にインタビューをしたところ，公募により 65 名以上の作家の参加があったという回答を得た．
4：2023 年度の参加店舗は 60 店舗であった
5：http://bakibaking.com/about/ (2024/03/11 確認)
6：https://www.wallshare-inc.com/about (2024/03/11 確認)
7：https://www.asahi.com/articles/ASPCC3R1ZPCCPQIP00C.html (2024/03/11 確認)

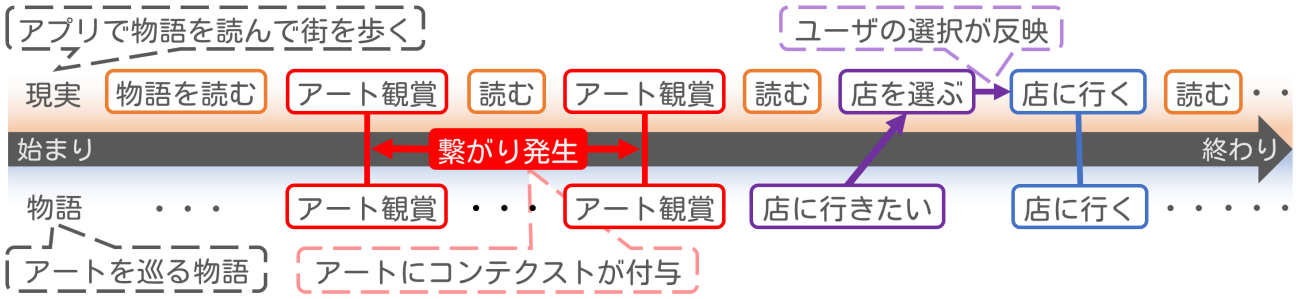


図 2: デザイン指針のイメージ図

4 デザイン指針

本研究では，下記の 2 つの行動を誘発することを目指す．

- コンテンツ体験（街にあるアートを観賞すること）に関心を向けること
- 街に点在するアートを回遊すること

コンテンツ（アート）体験に関心を向けるために，「物語コンテンツによる実環境への理解を促す手法」を基盤にする．物語コンテンツには，現代において身近なコンテンツであり [3]，ユーザにとってなじみ深いと考えられる「小説」を用いる．ユーザの回遊を誘発するために，位置情報を活用した「アートを見に行くように仕向ける手法」を用いる．

本研究では，「十三エリアに実在するアートの観賞を組み込んだ小説」の観賞を端緒としたシステムを用いることで，コンテンツ体験に関心を向けつつ回遊の誘発を試みる．コンテンツ体験に関心を向けるために，物語の主人公がアートを観賞する展開を迎えた時に，ユーザは主人公が見ているアートと同じものを実際に見に行く（図 2）．なお，物語の主人公がアートに関する情報や気づき（アーティスト特有のアートの特徴など）について言及する．回遊を誘発するために，各所にあるアートに行くことにより，物語が展開する設定を行う．物語を読んだ上で，物語に出てきたアートを現実で観賞することで，現実と物語に繋がりが発生し，観賞しているアートに対して，物語の内容やアートに関する知識といったコンテクストが与えられる．物語内の体験と現実の体験が重なり合い，アートを観賞したという感覚が拡張されることにより，自分が読者であるという意識が薄れて一時的に登場人物として作中の出来事を体験するようになる [7]．これらにより，ユーザにアートへの興味を喚起させ，アートへ意識を向けることが期待される．また，ユーザが物語を読み進めることで，意図せず街を巡ることが期待される．

十三にあるアートはユーザに入場料などの制限を設けていないため，誰でも自由に閲覧が可能であり，飲食店の店内にも多く掲示されているため，物語に店内への入店体験も含める．また，ユーザの選択を物語に反映することにより，物語への没入感を提供するため，主人公が訪れる店舗については複数の選択肢を提供し，ユーザが選択できるようにする．現在地とアートとの距離によって，特定のアートが選択されないことを防ぐために，アートや店舗の情報やそれまでの距離は，アートに向かうときに初めて表示する．物語に登場するアートは物語ごとに距離が近いアートを指定し，ユーザが訪れる時の歩行距離が少なくなるように設定する．

以上のデザイン指針を基にシステムを作成する．

5 提案システム

街中に壁画などのアートが点在する街である十三を対象とした，アートのあるスポットに行くことにより，物語の続きを読み進めることができる街歩き型ストーリーリーダーである「TokoTokoBanashi」を作成した．TokoTokoBanashi は主に a. 物語選択，b. 物語観賞，c. 地図閲覧，d. 店舗選択の 4 つのフェーズに分けられる．各フェーズの画面を図 3 に示す．

物語選択フェーズ（図 3-a）では，読み進めることができる物語が 5 つ表示され，その中から 1 つを選ぶ．ただし，物語ごとに内容や巡るアートが異なる．今回は，物語のタイトルの代わりに主人公の名前とジャンルを表示する．これは，自分がどの主人公になるかを選ぶイメージであり，主人公になり切って物語を観賞することで，物語内の体験と現実の体験に繋がりが発生することが期待される．なお，ジャンルは選択の手がかりとして表示する．5 つの物語の中から，読み進める物語を 1 つ選択すると，「あらすじの内容」と「選択している物語において訪れるアートの写真」が表示され，確認が終わると物語観賞フェーズに移行する．

物語観賞フェーズ（図 3-b）では，「主人公たちが十三の街にある淀壁や店舗を巡る物語」の本編を読み進める．物語の主人公がアート（淀壁）を観賞する展開を迎えると，ユーザも主人公が見ているアートを見に行くために，地図閲覧フェーズ（図 3-c）に移行する．また，稀にある物語の主人公が店舗に訪れる展開を迎えると，ユーザが主人公の訪れる店舗を選択するために，店舗選択フェーズ（図 3-d）に移行する．ただし，離脱率を考慮しつつ多くのアートの観賞を促すために，1 つの物語につき訪れるアートは 2 ～ 4 か所とし，訪れる店舗は 1 ～ 2 か所とした．

地図閲覧フェーズ（図 3-c）では，地図に記された目的地（アートまたは店舗）に向かう．地図には，ユーザの現在地と目的地をピンで表示する．ユーザが目的地に到着することで，再び物語観賞フェーズに移行し，物語の続きを読み進めることができる．

店舗選択フェーズ（図 3-d）では，主人公が訪れる店舗を選択する．3 つの選択肢の内，いずれかの店舗を選択することで，



図 3: 「TokoTokoBanashi」の画面
（a：物語選択フェーズ, b：物語観賞フェーズ, c：地図閲覧フェーズ, d：店舗選択フェーズ）

店舗の詳細な情報を確認する画面が表示される．選択肢として表示される店舗はすべてカフェなどの飲食店であり，店名や住所などの情報，店舗の外観や内装，商品などの写真が 1 店舗につき 5 枚表示される．店舗の選択が完了すると，ユーザに店舗への訪問を促すために，地図閲覧フェーズに移行する．

以上のように，物語が終了するまで物語観賞フェーズと地図閲覧フェーズ（及び店舗選択フェーズ）を繰り返す．

物語は，実際に十三を歩き，淀壁を確認して回った本稿の著者らが作成した．物語の作成において，「主人公が淀壁を巡り，観賞することを主軸にした物語にすること」，「物語上で淀壁の特徴を言及すること」を意識した．例えば，「この絵にも先ほどの幾何学模様が描かれている」，「この幾何学模様は BAKIBAKI さんというアーティストの独自のサインのようなものだそうです．『BAKI 柄』とも呼ばれるようですね」という記述があった．

6 検　　証

6.1 パイロットスタディ（PoC）

作成したシステムによって，街のアート観賞に関心を向けることが可能か検証を行った．評価指標は「アートに関する気づきを得たか」，「アートと街の関係性に関する気づきを得たか」の 2 点である．参加者は大学生の男性 2 人，女性 1 人の計 3 人であった（全員が十三への訪問経験があり，淀壁の閲覧経験もあった）．参加者に「TokoTokoBanashi」を使って十三を巡ってもらった後，半構造化インタビューを行った．半構造化インタビューでは，評価指標に関して，見たアートについて質問をした．

見たアートについての回答の内，アートの内容に関する気づきを得たかについては，顔が大きく描かれていた，メイド，レースの服，眼鏡をかけた，トレーを持っている，ガスマスクという回答が得られた．また，アートと街の関係性に関する気づきについては，駅近くの壁，喫茶店の壁に描かれていたという回答が得られた．このことから，作成したシステムを使用することで，アートに関する気づきを得る点ではコンテンツ体験が成されることが示唆された．

6.2 実環境におけるシステム検証

作成したシステムが実際のアートイベントにおいて，街のアート体験に関心を向けつつ回遊の誘発が可能か検証を行った．評価指標は「アートに関する気づきを得たか」，「回遊させるためのデザインが機能していたか」の 2 点である．実施場所は 2023 年 11 月 18 日から 11 月 26 日に十三で行われたイベントである「十三アートフェス 2023」であり，配布されたチラシに「TokoTokoBanashi」へのアクセス用の QR コード及び説明を記載した．現地でのシステム利用を担保するため，十三アートフェスに訪れ，「TokoTokoBanashi」へアクセスした人を対象に，データを収集し，アクセスデータを集計した．また，システムによるアート体験と回遊への効果を確認するため，途中で物語を読むのをやめた人と物語を読み終えた人を対象に，「TokoTokoBanashi」上で Google form を用いてアンケートを実施した．アンケートでは，評価指標の内「アートに関する気づきを得たか」については，見たアートについて質問をした．また，「回遊させるためのデザインが機能していたか」に関しては，「TokoTokoBanashi」の良かった点・悪かった点について

表 1: 各物語の各話のアクセス数

物語名	あらすじ	1 話	2 話	3 話	4 話	5 話	6 話
1: 村上十和	23	13	5*	5*	4*	4*	2**
2: 床村太郎	15	11	2*	2**	2*	1**	
3: 西中島叶	12	7	2*	2*	2*	1*	1**
4: 例江益男	9	7	2*	2*	1*	1**	
5: 木川結翔	10	4	2*	2*	1**		

*： アートのあるスポットに行かないと読めない話
**： 店舗に行かないと読めない話

質問をした．

アプリケーションへのアクセス数（IP アドレスの数でカウント）は 189 であった．その内，物語のあらすじを読んだ人は 46 人，物語の本編を読み始めた人は 26 人，物語を最後まで読んだ人は 3 人であった．ただし，不審なアクセス（ストーリーの閲覧順序がおかしい，複数のアートの観賞が 1 分以内に行われているなど）が多かったため，海外の IP アドレスによりアクセスされたデータは除外した．物語を最後まで読んだ人の内，2 つの物語を最後まで読んだ人は 1 人，3 つの物語を最後まで読んだ人は 1 人であった．

各物語における各話ごとのアクセス数を表 1 に示す．閲覧時の現在地に制限の無い第 1 話から，アートのあるスポットに行かないと読めない第 2 話にかけて，アクセス数がすべての物語で半分以下になっている．また，店舗に行かないと読めない話の直前での離脱が見られる．

アンケート回答者は 3 人であり，3 人とも 1 つ以上の物語を読破した．見たアートについての回答の内，アートに関する気づきを得たかについては「十和さまがメイドの壁画にもバキ柄があることに気づいて落書きでないことがわかった」や「共通の幾何学模様が描かれていた」という物語の内容が反映された回答が得られた．このことから，アートに関する情報や気づきを言及する物語を読みつつ，主人公と同じアートを実際に観賞することは，アート観賞に関心を向けることに効果的であることが示唆された．「TokoTokoBanashi」の良かった点・悪かった点についての回答の内，回遊させるためのデザインが機能していたかについては「知らないところまで歩ける」という回答が得られた．このことから，離脱が無ければシステム通りに回遊させることは可能であったことが示唆された．

6.3 十三アートフェス運営へのインタビュー

「十三アートフェス 2023」の開催後，十三アートフェスの運営である NPO 法人「淀川アートネット」牟田麻希氏にインタビューを行った．

システム面に関しては，「特に高齢の方には，アプリケーション使用の敷居が高いと感じる方が多い（QR コードを読み込み，Web アプリケーションを開くところまではやってくれた人が多かったが，システム自体の使い方が難しかったという人が多かった）」というシステム使用の敷居の高さを示す回答，「十三の外から歩いて訪れる人が多いため，淀壁を 1 つ 1 つ見に訪れてもらうのは難しい」という指定したコンテンツを巡ること（回遊）が困難であることを示す回答，「より簡単に使用できる，より街にリンクしたシステムが作れたら良い」や「店舗とのコラボレーションがあれば良い」というシステムの改善点に関する回答が得られた．物語に関しては，「学生らしい文章で続きを読みたいと思った」という回答が得られた．

7 終わりに

本研究では，コンテンツ（アート）体験に関心を向けつつ回遊を誘発するシステムの検討ため，十三の淀川エリアに実在するアートを組み込んだ物語の観賞を端緒とした街歩きアプリケーションである「TokoTokoBanashi」を作成し，これを用いることで，街にあるアート観賞に関心を向けつつ回遊の誘発が可能かシステム検証を行った．その結果，離脱が無ければ，アートについての気づきを得て，システム通りの回遊をすることが可能であることが示唆された．しかし，アートを見に行く直前で離脱が多く見られたため，離脱を抑えるシステム設計が必要である．また，実際にアートを見に行ってくれた人の中でも，離脱が見られたが，離脱の理由が不明であるため，その理由を調査する必要がある．例えば，システムを利用している道中で，興味深いスポットを発見したことによる離脱であれば，それは街を回遊する上で必要な離脱であると考えるため，復帰のしやすいデザイン設計が必要になる．一方で，システムに対する不満による離脱であれば，離脱を抑えるシステム設計が必要である．

謝　辞

本研究は公益財団法人都市活力研究所との共同研究「十三エリアのブランディングに関する研究」により支援された．また，NPO 法人「淀川アートネット」牟田麻希氏の協力を得た．記して謝意を表す．

文　献

[1] Hoppen, A., Brown, L. and Fyall, A.: Literary tourism: Opportunities and challenges for the marketing and branding of destinations?, *Journal of Destination Marketing & Management*, Vol. 3, No. 1, pp. 37–47 (2014).
[2] Seaton, P., Yamamura, T., Sugawa-Shimada, A. and Jang, K.: *Contents Tourism in Japan: Pilgrim-ages to "Sacred Sites" of Popular Culture*, CambriaPress (2017).
[3] インプレス総合研究所: 電子書籍ビジネス調査報告書, インプレス (2019).
[4] 上段貴浩, 脇田祥尚: アートイベントによって顕在化する歴史的市街地の地域資源-「からほりまちアート」を事例として, 日本建

築学会計画系論文集, Vol. 75, No. 658, pp. 2873–2880 (2010).
[5] 大杉隆文, 仲西渉, 多井中美咲, 井上卓也, 伊藤悠, 岩井瞭太, 香川健太, 松下光範, 堀雅洋, 荻野正樹: 動物園における自発的な観察の促進を目的とした園内回遊行動のデザイン, 情報処理学会論文誌, Vol. 58, No. 11, pp. 1765–1775 (2017).
[6] 岡本健: 来訪者の回遊行動を誘発する要因とその効果に関する研究 『学術講演梗概集』 F-1, 都市計画, 建築経済・住宅問題 (2009).
[7] 小山内秀和, 楠見孝: 物語世界への没入体験, 心理学評論, Vol. 56, No. 4, pp. 457–473 (2013).
[8] 春日和俊, 伊藤孝紀, 堀越哲美: アートイベントにおける都市部の地区演出の可能性 あいちトリエンナーレ 2010 長者町会場を事例とする, 日本建築学会計画系論文集, Vol. 77, No. 681, pp. 2577–2584 (2012).
[9] 川崎依子, 木多道宏, 鈴木毅, 松原茂樹: 地域の中で行われるアートイベントに関する研究 (都市計画), 日本建築学会近畿支部研究報告集. 計画系, No. 45, pp. 633–636 (2005).
[10] 小森真樹: 町づくりと芸術の互恵関係: 三宅航太郎 [おしょくじ] アートプロジェクトの展開から, 都市計画, Vol. 62, No. 4, pp. 47–50 (2013).
[11] 高山啓子: アートイベントと観光まちづくり: 瀬戸内国際芸術祭と地域社会, 川村学園女子大学研究紀要, Vol. 28, No. 3, pp. 1–12 (2017).
[12] 新田りら: 現代アートイベントによる市街地活性化の試みに関する研究 (商業空間・博物館, 都市計画), 学術講演梗概集. F-1, 都市計画, 建築経済・住宅問題, No. 2003, pp. 895–896 (2003).
[13] 安尾萌, 宮本誠人, 樋口友梨穂, 三溝俊介, 松下光範: 実環境と紐づいた物語コンテンツの提示による周遊行動の誘発, 電子情報通信学会第 2 回コミック工学研究会, pp. 41–45 (2019).

CNNによるフェルメールの画風分析

小田　稜子†　　伊藤　貴之†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
E-mail: †{g2020515,itot}@is.ocha.ac.jp

あらまし　ディープラーニングによる画像の分析が近年活発に報告されている．この技術を絵画分析に応用することで，絵画解釈のタスクの中でも人間にも判断の難しいタスクを補助することや，新しい美術的見識を得たりすることが期待される．しかしディープラーニングには大量の画像が必要であり，絵画のように制作物の数が限られる題材の扱いは簡単ではない．また，学習した特徴の解釈は複雑であるため，一見して何を特徴として解釈したのかわからないという問題もある．そこで本研究では機械学習結果の可視化により，特に作品数が少ないフェルメールの特徴量分析に挑戦する．具体的には，フェルメールとその周辺画家の分類を CNN で学習し，その中間層から得られる特徴量を散布図によって可視化し，得られた結果について考察する．

キーワード　情報可視化，機械学習，絵画

1　はじめに

ディープラーニングを含む機械学習技術の進化によって第 3 次 AI ブームが到来し，各業界がその活用方法を模索している．特に芸術分野においては，機械学習をもちいることによって専門・非専門を問わずさまざまな恩恵を得ることが期待できる．例えば機械学習で作品の特徴を学習することによって，膨大な作品の中からイメージに近いものを検索することが可能となるだけでなく，より多くの人に美術の知識を周知することができる．また真贋や作品価値の判断や文化財の保護において，専門家の補助的な役割を果たすことができるようになる．

画像分類タスクによく用いられるディープラーニングのアーキテクチャとして，CNN (Convolutional Neural Network) が知られている．CNN は多様な重みを持ったフィルターによって「畳み込み」という操作を繰り返すことで，画像から階層的に高次の特徴を抽出し，精度の高い分類をすることができる．CNN は絵画の様式分類において高い精度を出すことが報告されており [1]，人間の判断に近い特徴を捉えていることが予想される．したがって，CNN を適用することで，画家個人の画風の分析をすることも可能であると考えられる．しかし，CNN の学習には大量の画像が必要であり，作品数が少ない画家の絵画において十分な学習ができないことがある．よって，大量の画家の作品を同時に学習し，様式の流れを分類することができても，個人の画家の特徴を学習することは簡単ではない．特にフェルメールのような寡作な画家においては，学習の難易度が顕著に高くなる．また，CNN のような機械学習においては，分類の根拠になる特徴量の解釈が人間にとって大変煩雑であり，そのままでは計算機が画像のどの部分に着目して画像分類にいたったのかを解釈することは困難である．したがって，機械学習で個人の画風の研究をするにはいくつかの工夫が必要となる．以上の理由により，寡作な画家個人の画風について考察する研究はとても少なく，議論の余地がある．

本研究では転移学習を利用して，作品数の少ない画家であるフェルメールとその周辺画家の特徴量を可視化することで，相対的にフェルメールの絵画の特徴を考察する手法を提案する．また，フェルメールには未だに真贋のはっきりしない作品が存在する上に，有名な贋作者であるメーヘレンがいる [2] という点で専門家の間でも議論を呼ぶ画家である．機械学習を通してこれらの贋作と真作の違いを見出すことにも挑戦する．

本研究の目的は以下のとおりである．

- 転移学習によってデータ数の少ない画家の分類問題に挑戦する．
- 散布図による可視化をもちいて，フェルメールの絵画に共通する特徴を機械学習の側面から考察する．
- 本研究で採用した方法で，フェルメール真贋を判断できるか実験する．

本報告の構成は以下のとおりである．2 節では関連研究について述べる．3 節では提案手法について， 4 節で結果，5 節で結果に対する考察を述べる．最後に 6 節でまとめと今後の課題について述べる．

2　関連研究

本章では関連研究として，機械学習による絵画の分類や，少ない画像で機械学習を実行する手法を紹介するとともに，本研究においての課題を述べる．

2.1　機械学習による大規模な様式分類

CNN を使って絵画の様式分類をする研究は，従来から多数報告されている．例として Sharmr ら [1] は，印象派，シュルレアリスム，抽象主義の 9 人の画家を CNN で分類する実験を報告している．結果は画家分類の精度 77%，様式分類の精度 91% を達成し，CNN と絵画の分類，特に様式分類と相性がよいことを示した．また Elgammal ら [3] は，77k に及ぶ絵画のデータを CNN で転移学習し，大規模な様式分類を試みた．この研

究によって，CNN が学習した絵画の特徴量は Wölfflin による様式論 [4] と相関があることがわかった．言い換えれば，様式分類のファクターは何が描かれているか (テーマ) ではなく，絵画の構図やコントラスト，色構成，彩度などの表面的な視覚情報に表れていることが示された．

以上の研究は大規模なデータセットを適用したものである．それに対して，画家個人の作風分析にはこのような規模のデータセットを構築することは困難である．また様式分類のファクターとして示された視覚的要素のうち，どの要素が優勢なのかを詳細に議論することが難しいという点も課題として残っている．

2.2 機械学習による小規模な画像の分類

CNN による学習では過学習を避けるために大量の画像が必要となる．しかし実際のタスクにおいては満足な量の画像を用意できないことも多い．そこで少ない画像から学習を実現する研究も活発に進められてきた．Shorten ら [5] は少数の画像からなるデータセットに手を加えて過学習を防ぐ Data Augmentation についてまとめており，その方法は主にデータワーピングとオーバーサンプリングに分類されると述べている．具体的には，画像に対して反転やノイズ付与などの加工や，GAN [6] による生成など，画像の本質を変えずに多数の画像を追加生成する．これらの手法は，一般的な画像に対しては，分類精度を向上させる非常に有効な策とされている．しかし，絵画画像では色や構図などの多くの視覚要素が作家性を表すことから，必ずしも Data Augmentation が好ましい手法とは言えない．そのため，本研究で採用するには慎重な議論が必要である．

2.3 機械学習による画家の分類と分析

画家個人の分析のために CNN を適用した研究も存在する．Folego ら [7] は，CNN と SVM を組み合わせて絵画の作者がゴッホであるかを判別する分類器を実装した．また Frank ら [8] は，レンブラントとゴッホについて分類をし，その信頼度 (分類結果におけるラベルの確信度) で絵画のどの部分に画家らしさが現れているのかを示した．両者とも Data Augmentation として画像を細かく分割する方法を適用しており，画像の作家性を損なわずにデータを増やすことに成功している．ただし絵画画像の分割にもとづく Data Augmentation には，元画像の解像度によって 1 ピクセルあたりの実際の寸法が変わってしまう影響が出やすいという課題がある．また，本研究で使用した CNN モデルでは，絵画画像全体に対する学習から重みを算出するため，細かい特徴の分類にはあまり適していない．よって本研究では，画像の表面的な構成要素を加工せずに，全体の特徴を反映しながらデータ画像の数を増やすことを目指した．

3 絵画画像の機械学習と可視化

我々はフェルメールを含むバロック時代の絵画を対象として，データセット作成と加工，機械学習，および可視化を試みた．本章ではその手順について論じる．

3.1 データセットの作成

本研究では，パブリックドメインの油彩を中心とした作品を，人物画や風景画などの絵画のジャンルを問わずに収集した．対象の画家は，フェルメールおよび，フェルメールと同じバロック画家でオランダ出身のレンブラント，フランス・ハルス，ヤン・ステーンと，バロック画家でフランドル出身であるルーベンスの計 5 人である．各画家の収集した作品の数を表 1 に示す．

これらの画像のうち 20%をテストデータとして，後述の加工を適用した．このデータセットを A と称する．また，データセット A のフェルメールのデータを，訓練データに真作と認められているもののみ，テストデータを贋作の疑惑があるものとメーヘレンによる贋作，フェルメール本人による真作 1 作品にしたものに変えたデータセットも作成した．このデータセットを B とする．データセットのラベルは各画家の名前に対応させた 5 つとした．

表 1 データセット

ラベル	画家	絵画の数
0	フェルメール	32
1	レンブラント	135
2	フランス・ハルス	47
3	ヤン・ステーン	90
4	ルーベンス	221

3.2 データセットの加工

機械学習において特にデータとなる画像の枚数が少ない際に，対策として画像を加工して枚数を増やす Data Augmentation が適用されることが多い．これにより，学習の精度を高めつつ過学習を防ぐという効果が期待できる．しかし前述のとおり，絵画のような構図や色に意味のある画像については，加工による Augmentation の適用は慎重にするべきである．そこで代わりに本研究では，絵画の縦横比を維持したままリサイズして切り出すという方法を採用した．ここで，CNN のアーキテクチャの多くが入力画像に正方形を指定しているが，絵画の縦横比は正方形でないことの方が多い．よって絵画をそのまま正方形へリサイズしてしまうと，絵画の縦横比が崩れてしまう．そこで，絵画の短辺を入力画像のサイズに合わせてリサイズし，長辺側をずらして切り出すことで，絵画の縦横比を保ったままデータの数を増やすことにした．こうしたリサイズ処理を行うことで，絵画の構図やライティング，コントラストといった全体の印象を学習することができる．以上の処理の詳細を図 1 に示す．

3.3 学　　習

本研究では，Python 向けの機械学習ライブラリである Pytorch を使い，標準で用意されているモデルと重みを利用して上記データセットの転移学習を行った．選別においては，Pytorch の公式サイトを参考に， RegNet_Y_32GF を選んだ．RegNet [9] は 2020 年に Facebook の AI チームによって開発されたアーキテクチャで，ネットワークの入り口にあたる Stem，畳み込みの処理が幾つもまとまった Block，プーリング層，そ

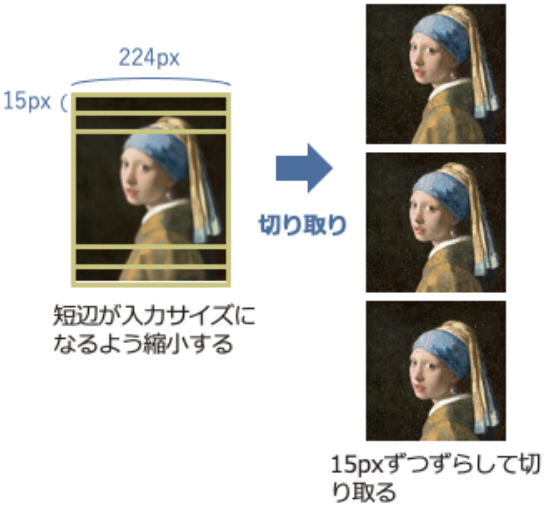


図 1　画 像 加 工

して全結合層から構成される．それまで最高精度を誇った EfficientNet [10] よりも精度が高く，より高速に学習できると報告されている．今回使用した RegNet_Y_32GF には 4 つの block が線形に組み込まれており，それぞれの block には 2, 5, 12, 1 個の ResBottleneckBlock [11] が含まれている．ResBottleneck は ResNet [11] に使用されている bottleneck 構造であり，1 × 1，3 × 3，1 × 1 の 3 つの畳み込み層と SE ブロックを持ち活性化関数を ReLU とした層の集合 (図 2) である．このモデルの最終層 (全結合層) のみを用意したデータセットで学習し，エポック数を 20，損失関数を CrossEntropyLoss()，最適化アルゴリズムを Adam としてとして，ラベル分類を試みた．

実験環境は以下のとおりである．

OS: Windows 11 Pro

CPU: Intel Core i7-12700H / 16GB メモリ

GPU: NVIDIA RTX A2000

3.4　可　視　化

本研究では，CNN が読み取った特徴量として中間層の出力をもちいる．具体的には出力層の手前である trunk_output.block4 の出力を使用する．CNN は入力から遠い層であるほど抽象的な特徴を学習するため，出力層の手前の層からは構図やライティング，色構成などの概略的な特徴を反映した特徴量を抽出できることが想定される．3.3 節の実験で学習させたモデルに対し，学習データとテストデータを入力として再度分類させ，その過程で得られる中間層の特徴量を抽出した．抽出した特徴量は [3712, 7, 7] の 3 次元テンソルであり，各データを合わせると [データ数, 3712, 7, 7] の 4 次元テンソルとなる．本研究では学習データのみ，テストデータのみ，学習データとテストデータを合わせたものの 3 通りについて特徴量を抽出した．これを一度 [データ数, 3712*7*7] へフラット化したのち，UMAP [12] による次元削減を適用して 2 次元に変換した．この結果を Plotly で散布図にして観察した．また比較対象として，入力層に近い層である stem についても同様に処理を適用し，散布図を作成した．stem から抽出した特徴量は [32, 112, 112] の 3 次元テンソルである．

以上の実行結果を参照して，以下の直感的な仮説を検証する．

[仮説 A] フェルメールは当時の画家としては珍しく，青色 (いわゆるフェルメールブルー) を多用する画家である．よって色によってクラスタが形成されると考えられる．

[仮説 B] フェルメールは構図が正確であり，左側から光が差す構図をよくもちいる画家である．散布図には構図によるクラスタが見られると考えられる．

[仮説 C] フェルメールの真作と贋作，および疑惑のある作品が構図やライティングから見分けられると仮定すると，贋作は散布図上で真作のクラスと離れた場所に分布すると考えられる．

4　結　　果

本章では機械学習結果の精度と可視化例について示す．

データセット A に対して転移学習した際の精度と損失関数の値は，それぞれ図 3，4 のようになった．学習データの精度は 0.99 を超え，テストデータの精度も 0.8 を達成した．学習データとテストデータの混同行列は図 5，6 のようになった．学習データではほぼ全ての画像を正しく分類できているが，テストデータではフランス・ハルスの正しく分類できた作品よりもヤン・ステーンに誤判別された画像の方が多くなり，フランス・ハルスの分類に弱い傾向を確認できた．

学習データの trunk_output.block4(入力から遠い層) の特徴量から作成した散布図は図 7 のようになった．この散布図を見ると，画家ごとに比較的綺麗なクラスタを形成していることがわかる．このうちルーベンスに対応する点群を非表示にすることで，図 9 を得た．フェルメールを表す点のうち，フェルメールのクラスタから離れている点に注目すると，同じ作品から切り取られた画像が別の場所に位置していることがわかる．それに対して，学習データの stem(入力から近い層) の特徴量から

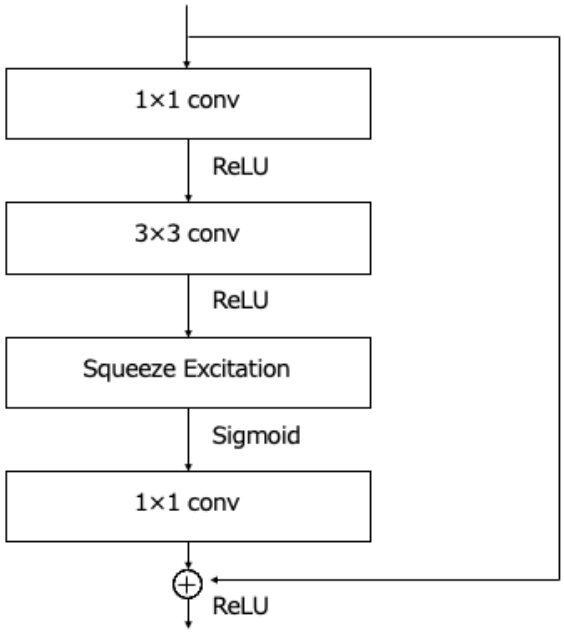


図 2　ResBottleneckBlock の構造

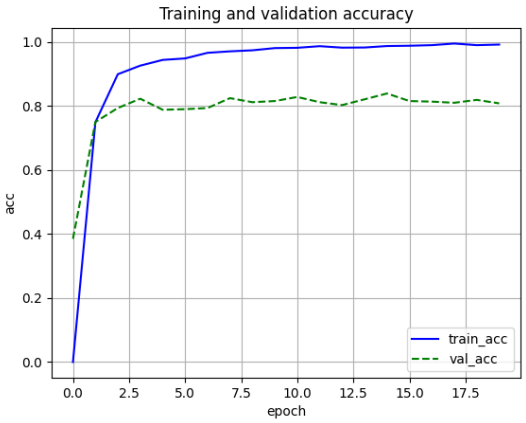


図 3　データセット A での精度

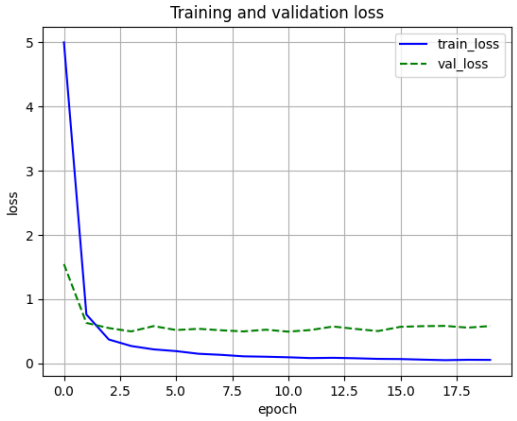


図 4　データセット A での損失

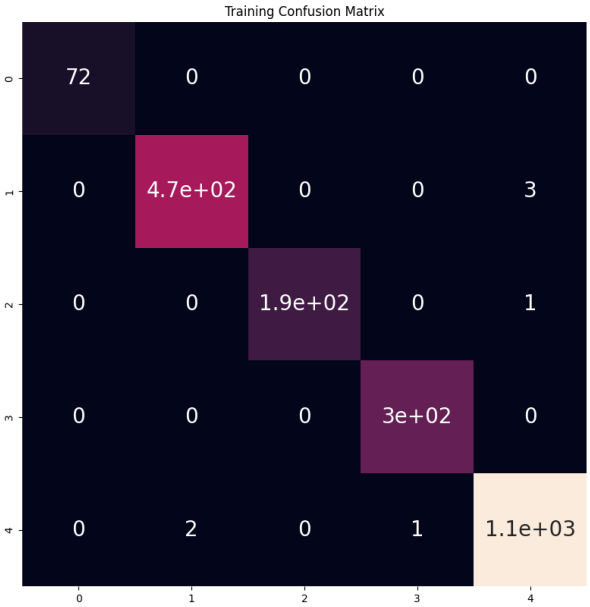


図 5　データセット A の学習データでの混同行列

作成した散布図は図 8 のようになった．図 7 に比べ，画家ごとのクラスタは明確には見られないことがわかる．

続いて，データセット B に対して転移学習した際の精度と損

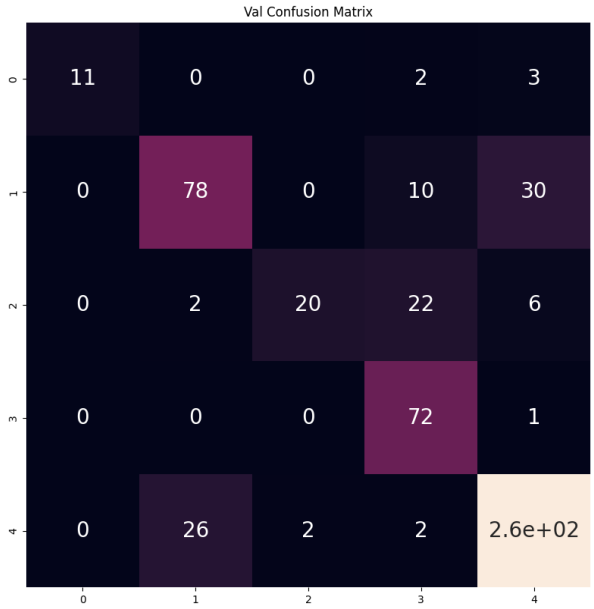


図 6　データセット A のテストデータでの混同行列

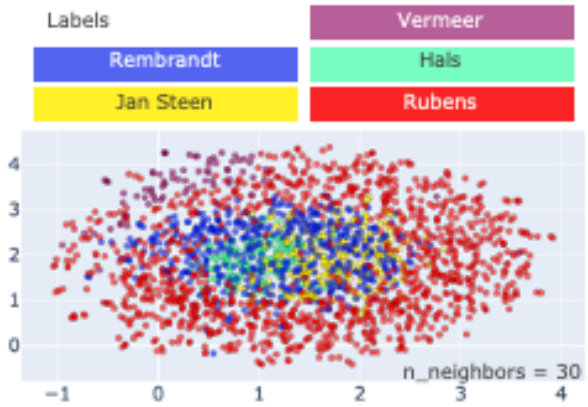


図 7　データセット A での出力層近くの特徴量の散布図

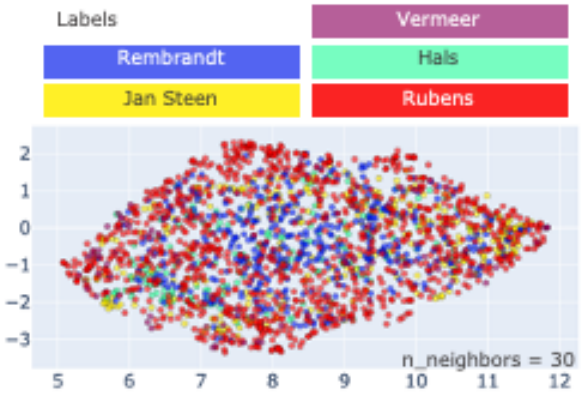


図 8　データセット A での入力層近くの特徴量の散布図

失関数の値は，それぞれ図 10, 11 のようになった．精度は両方ともデータセット A と同程度の高い値を示したが，テストデータの損失関数の値がやや不安定になった．混同行列は図 12，13 のようになった．混同行列を見ると学習データでは変わらずほぼ全ての画像を正しく分類できている．一方で，テストデータ

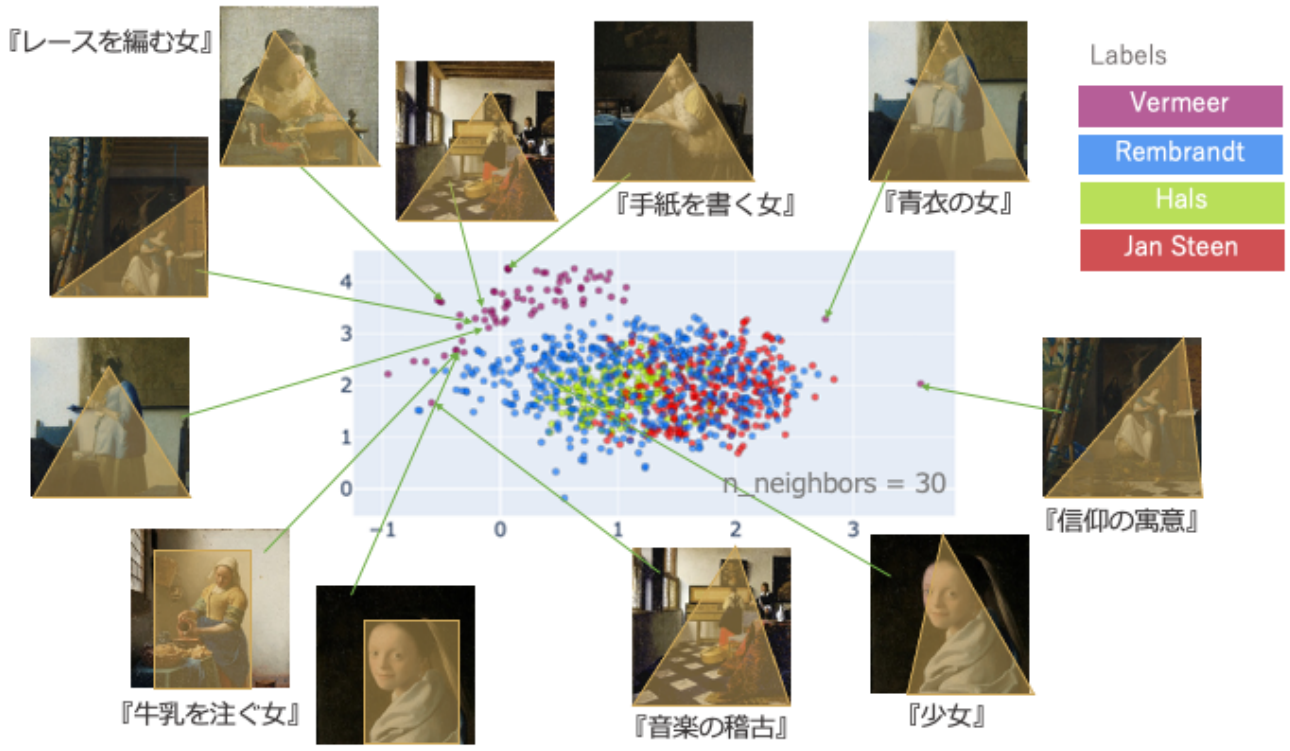


図 9　データセット A での出力層近くの特徴量の散布図のうちルーベンスを非表示にした結果

でもフランス・ハルスの分類に弱い傾向が確認できたほか，フェルメールの分類の精度にもばらつきが出た．テストデータのうち，フェルメールでないという分類結果が出たのは贋作疑惑のある『赤い帽子の女』，メーヘレンによる贋作である『エマオの食事』，フェルメールがフェリーチェ・フィケレッリの作品を模写したとされる『聖プラクセディス』[17] の 3 作品である．いずれも全ての切り取り画像においてフェルメール以外の画家の分類結果がでた．

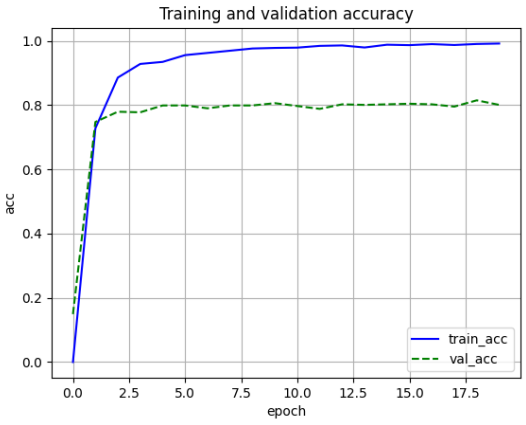


図 10　データセット B での精度

データセット B のテストデータについて，trunk_output.block4(入力から遠い層) の特徴量から作成した散布図は図 14，そのうちルーベンスを除いた散布図は図 15 のようになった．図 9 と比べると雑然としている．データセット A のときとは異なり，

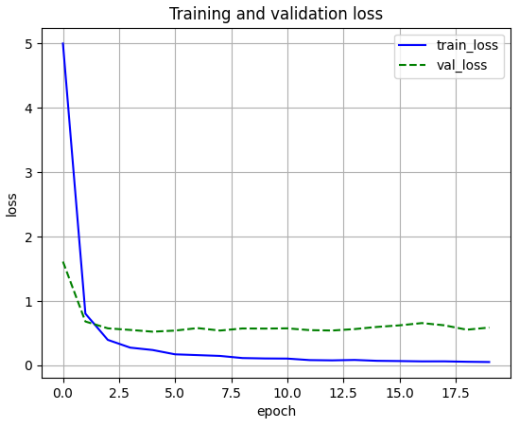


図 11　データセット B での損失

フェルメールの各作品から切り出された画像はほぼ同じ位置に固まっていた．

5　考　　察

前章で示した結果をもとに，フェルメールが持つ他の画家との差異について検証する．まず，図 7 を全体的に眺めると，中央にレンブラント，フランス・ハルス，ヤン・ステーンに対応する点群の塊があり，それを囲むようにルーベンスに対応する点群が位置していることがわかる．フェルメールに対応する点群はルーベンスよりもやや外側にクラスタを形成しており，他の画家とは明らかに一線を画している．なお，ルーベンスに対応する点群の配置については，枚数の多さとジャンルの多彩さ

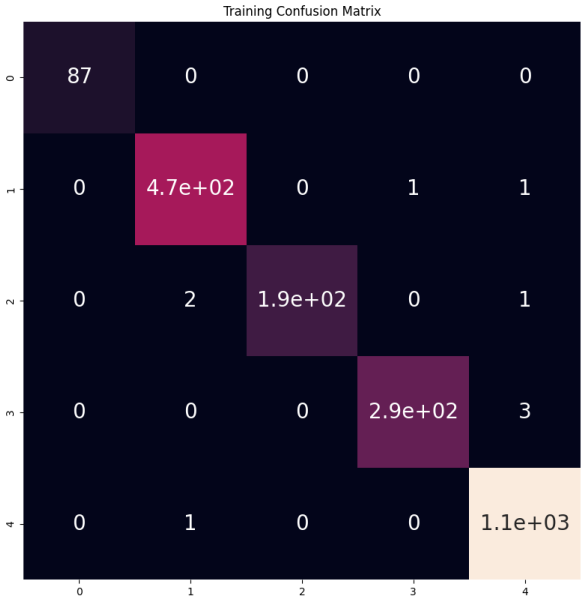


図 12　データセット B の学習データでの混同行列

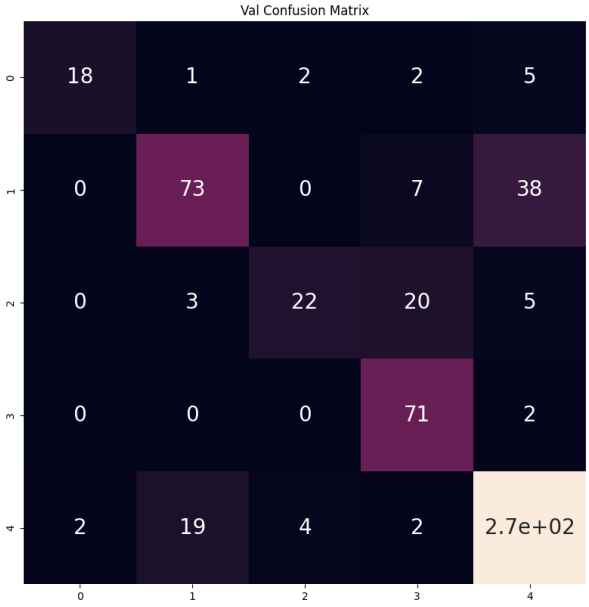


図 13　データセット B のテストデータでの混同行列

から『その他大勢の画像』として認識されている可能性が高い．以上により，フェルメールの絵画は他の画家と比べて何かしらの突出した特徴を持っていることが示唆される．その特徴の可能性として構図があげられる．フェルメールは構図にこだわりを持った画家であり，特にピンを用いて正確な透視図法をもちいたことで有名である [13] [14]．すなわち，奥行きが強調された絵画が多い．対してレンブラントらは，対角線構図や三角構図など，画面に対して平面的な構図をとることが多い．本研究では ImageNet-1K [15] を用いた事前学習モデルを転移学習に採用しているが，この画像データセットは現実的な写真を収録していることから，現実世界の構図を忠実に再現したフェルメールの作品が他の画家とは異なるものとして学習できた可能性が示唆される．

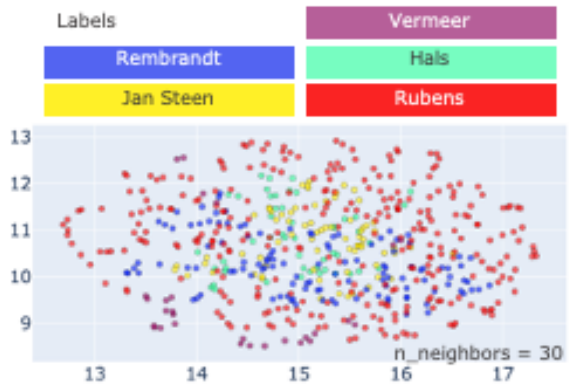


図 14　データセット B，テストデータの特徴量の散布図

また，本研究で作成したデータセットでは，全ての画家において肖像画や風俗画，風景画が明確に区別されずに混在していることから，分類の要因が絵画の主題 (何が描いてあるか) ではないと考えられる．

さらに，クラスタの端に位置する『レースを編む女』に注目すると，フェルメールには珍しい右側から光の当たる作品である．この作品もクラスタ内に含まれることから，絵画のライティングも画家の分類にはさほど関係ないと言うことができる．

次に，3 章であげた仮説について，図 9 を参照しながら検証する．[仮説 A] の色によるクラスタについては，同じ作品から切り取られた画像が別の場所に位置していることから，関連性が低いと考えられる．特に，フェルメールに対応する点群から離れた場所にも青色の目立つ作品が位置することから，少なくともバロック時代の画家分類において色はそれほど重要ではないと考えられる．

次に [仮説 B] の構図によるクラスタについて考える．フェルメールに対応する点群のクラスタから孤立して離れている点は，『青衣の女』,『信仰の寓意』,『少女』,『音楽の稽古』の 4 作品であった．そのいずれも作品全体が離れた場所に位置しているわけではなく，切り取った一部の画像のみが孤立している．したがって，切り取りにより位置がずれることが分類に影響していると考えられる．言い換えれば，例えば構図や明暗の配置などが分類のファクターになっていることが示唆される．上述の 4 作品をそれぞれ比べてみると,『信仰の寓意』,『少女』,『音楽の稽古』の 3 作品においては，画像上部に大きな空間があるように切り取られた画像は，フェルメールに対応する点群のクラスタに属することがわかる．逆に，そのような空間がないように切り取られた画像は，クラスタから離れて位置していることが顕著にわかる.『青衣の女』においても，他の 3 作品に比べるとその差がわかりづらいものの，画面上部に空間のある画像はクラスタに属し，そのような空間がない画像はクラスタから離れた場所に位置している点では共通している．

また，絵画中の対象物の配置について大局的に注目してみると，クラスタから離れている上述の絵画 4 枚とも，画面上部から大きく三角形を描くような配置になっている．それに対して，フェルメールに対応する点群のクラスタに属する他の作品は，

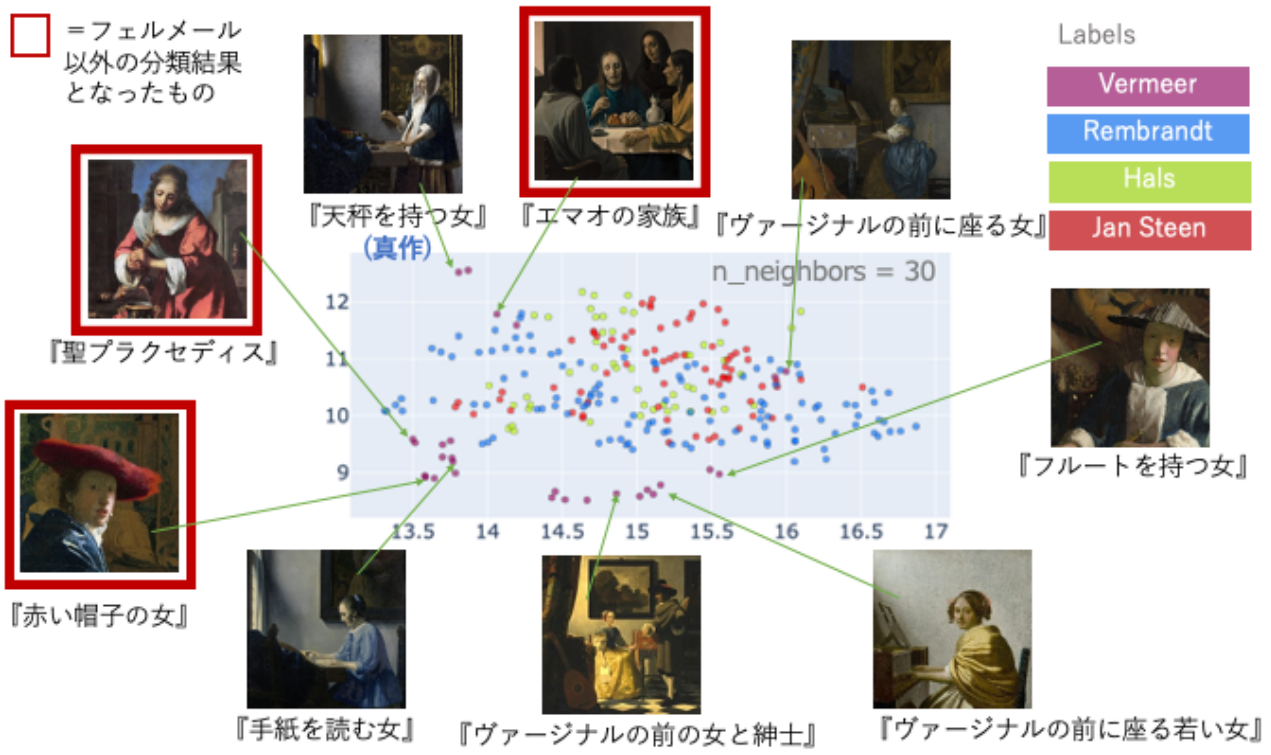


図 15　データセット B，テストデータの特徴量の散布図のうちルーベンスを非表示にした結果

画面の少し下からの三角，あるいは四角の配置になっている．このことからも，今回採用した機械学習でのファクターは主に構図であり，フェルメールは特に上部の空間が特徴的であることが示唆される．

ここで，仮説に対する反例を考える．フェルメールの画像のうち，画面上部から三角形状に対象物が配置されているにもかかわらず，フェルメールに対応する点群のクラスタ内に存在する画像として,『レースを編む女』の一部を切り取った画像があった．また，画面上部に空間があるにもかかわらず，フェルメールに対応する点群のクラスタから遠い位置にある画像のうち，フェルメール以外の画家の画像も多数存在した．特に『レースを編む女』は，上部に空間がある切り取り画像も，空間がない切り取り画像も，散布図上ではほぼ同じ場所に位置していた．これに対する説明として，散布図上の配置はまず構図の立体感にもとづく配置 (強い分類) が行われ，後からフェルメールらしさである上部空間による配置 (弱い分類) が行われているのではないかという仮説を立てることができる．画面上部に空間があるかどうかにかかわらず，他の画家の画像がフェルメールのクラスタに混ざらない要因として，画家を分類するファクターが上部空間とは別にあることが示唆される．このファクターに関しては先述の通り，構図が立体的であること (奥行きがあること) が考えられる．よってこの可視化からは，フェルメールと周辺画家との差異は奥行きを強調した構図であり，その中でよりフェルメールらしさを強調しているのが画面上部の空間であることが読み取れたと考えられる．

最後に [仮説 C] の真贋の判断について検証する．図 15 を見ると、唯一の真作である『天秤を持つ女』は離れた場所に位置しているものの，フェルメールであると分類された『手紙を読む女』とそうでない『赤い帽子の女』が近い位置にあるなど，釈然としない点がある．またメーヘレンによる贋作の『エマオの家族』,『手紙を読む女』,『ヴァージナルの前の女と紳士』の分類結果が割れているほか，同一人物の作とされる『フルートを持つ女』,『赤い帽子の女』も片方のみがフェルメールであると判別され，散布図上の位置も異なることから，このモデルで真贋の判断はできていないことがわかる．少なくともこのモデルが参照しているであろう絵画の全体的な特徴，特に「構図」からは，フェルメールの作品の真贋は区別できないことが示唆される．

参考になる知見としては，まず暖色の表現に乏しい [16] とされるメーヘレンの作品が固まった場所に位置していないことから，本研究が採用したモデルでは色彩表現をもとに分類していないことが強調される．また,『聖プラクセディス』はイタリアのバロック画家であるフェリーチェ・フィケレッリの作品を模写したものと言われている. [17]．よって，オランダのバロック画家であるレンブラントたちではなく，フランドルのバロック画家であるルーベンスの作という分類結果が出たことは妥当と言える．このモデルは同じバロック画家でも，国の違い程度なら判別できる可能性が見込まれる．

いずれにしても，真贋の判断という繊細なタスクは，今回採用したモデルでは難しいことがわかった．より詳細に判別させるには，より絵画に特化してチューニングしたモデルを使用するか，あるいは全体の特徴ではなく細部の特徴から判別するようにチューニングする必要があるであろう．

6 まとめと今後の課題

本報告では機械学習をもちいて絵画を分類した後，その中間層を散布図によって可視化することで，画家分類の鍵となるようなファクターを示した．ImageNet を事前学習した CNN モデルでフェルメールおよび同じ時代のバロック絵画を分類する際に，主に構図がその分類タスクの大きなファクターであることが示唆された．特にフェルメールらしさについては，画面上部に大きく空間があることが特徴であることが示唆された．真贋判別については，少なくともこのモデルでは判別が難しいこと，すなわち構図から判断することが難しいことがわかった．真贋を正確に区別するには，色などの別のファクターを捉えられるようにするか，もっと細かい筆致について分析できるようにする必要がある．

今後の課題として，上述のように構図以外のファクターを重視するモデルの検証があげられる．この目標を達成するためには，事前学習なしに，最初から絵画のみを学習させたモデルの作成が理想である．しかし，ImageNet ほどの大規模な絵画データセットの構築には一定の手間がかかる．代替え手段として，既出の事前学習モデルから，色や明暗，ストロークなど別のファクターを重視するようチューニングしたモデルを模索したい．

また現時点では散布図から特徴を読み取る作業が煩雑であり，分析者個人の力量を要する．よって画像ごとに造形要素をわかりやすく示す可視化システムの開発について検討したい．

さらに，フェルメールに限らず西洋美術史全般の様々な画家についても同様な作業を実施して，既存の美術的な価値観を見出せるか，新たな価値観を発見できないか，検証したい．

謝　　辞

本研究の遂行にあたり，有用なご助言を賜りました，大阪大学大学院情報科学研究科 中島悠太准教授に深謝いたします．

文　　献

[1] Shamir, Lior, et al. ”Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art.” ACM Transactions on Applied Perception (TAP) 7.2 (2010): 1-17.
[2] Lopez, Jonathan. The Man who Made Vermeers: unvarnishing the legend of master forger Han van Meegeren. Houghton Mifflin Harcourt, 2009.
[3] Elgammal, Ahmed, et al. ”The shape of art history in the eyes of the machine.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
[4] Wölfflin, Heinrich. Principles of art history. Courier Corporation, 2012..
[5] Shorten, Connor, and Taghi M. Khoshgoftaar. ”A survey on image data augmentation for deep learning.” Journal of big data 6.1 (2019): 1-48.
[6] Karras, Tero, et al. ”Progressive growing of gans for improved quality, stability, and variation.” arXiv preprint arXiv:1710.10196 (2017).
[7] Folego, Guilherme, Otavio Gomes, and Anderson Rocha. ”From impressionism to expressionism: Automatically identifying van Gogh’s paintings.” 2016 IEEE international conference on image processing (ICIP). IEEE, 2016.
[8] Frank, Steven J., and Andrea M. Frank. ”Analysis of Dutch Master Paintings with Convolutional Neural Networks.” arXiv preprint arXiv:2002.05107 (2020).
[9] Radosavovic, Ilija, et al. ”Designing network design spaces.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[10] Tan, Mingxing, and Quoc Le. ”Efficientnet: Rethinking model scaling for convolutional neural networks.” International conference on machine learning. PMLR, 2019.
[11] He, Kaiming, et al. ”Deep residual learning for image recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[12] McInnes, Leland, John Healy, and James Melville. ”Umap: Uniform manifold approximation and projection for dimension reduction.” arXiv preprint arXiv:1802.03426 (2018).
[13] Wadum. J., Vermeer in Perspective, National Gallery of art, Washington. 1995.
[14] 佐藤紀子. ”ヨハネス・フェルメールの絵画空間の考察―構図決定時におけるカメラ・オブスクラの役割―.” 図学研究 41.3 (2007): 17-26.
[15] Stanford Vision Lab, Stanford University, Princeton University (2020), “IMAGENET”, https://www.image-net.org/index.php, 17.12.2023
[16] Izumiya, Yoshio. ”比較鑑賞で読み解くフェルメールの造形性.” 岡山大学大学院教育学研究科研究集録 158 (2015): 119-136.
[17] Wheelock, Arthur K, 成田睦子. ”フェルメールとその時代.” (2000)

深層距離学習による文字種とスタイルの特徴獲得とその特徴に基づく文字生成

桑田　若菜† 　三林　亮太† 　谷　雅徳†† 　大島　裕明†

† 兵庫県立大学大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 北陸先端科学技術大学院大学未来創造イノベーション推進本部　〒 923–1292 石川県能美市旭台 1–1
E-mail: †{ad23g022,af22h007}@guh.u-hyogo.ac.jp, †††ohshima@ai.u-hyogo.ac.jp

あらまし　本論文では日本語の手書き書道文字を生成する手法を提案する．文字種特徴とスタイル特徴を両方抽出し，文字生成を行う生成器を作成する．生成器は文字種エンコーダとスタイルエンコーダとミキサーで構成される．生成器の学習には SimSiam と GAN の手法を用いる．事前に適した特徴表現を得るために文字種エンコーダとスタイルエンコーダを，SimSiam を用いて事前学習する．そして，生成器全体の学習を GAN によって行う．今回用いるデータセットは日本書道データとくずし字データと日本語フォントデータからなる．このデータセットで学習した提案モデルとベースラインモデルを，3 つの評価指標である MAE，SSIM，LPIPS で比較する．我々の提案モデルは MAE 0.2096，SSIM 0.5870，LPIPS 0.2214 のスコアを記録し，ベースラインと比較するとそれぞれ 0.0212 ポイント，0.039 ポイント，0.0107 ポイント優位であった．

キーワード　書道作品生成，文字画像生成，深層学習

1　はじめに

近年，深層学習を用いたアート生成技術が発展している．例えば，17 世紀オランダの画家であるレンブラント作品をコンピュータで再現する「The Next Rembrandt [1]」プロジェクトがある．過去のレンブラント作品をコンピュータに学習させ，レンブラントが書いたことがない新たな作品を生み出すプロジェクトである．他の例では，2022 年にオープンソースとして公開された「Stable Diffusion [2]」が挙げられる．Stable Diffusion は入力されたテキストから画像を生成する Text to Image モデルで，大量のテキストと画像のペアデータで学習されている．Stable Diffusion が公開されて以降，このモデルを用いた画像生成サービスが多くリリースされるようになった．それにより，一般の人は各サービスを介して深層学習による画像生成を簡単に行えるようになった．

アートと深層学習の融合によって，人間の想像力を超えた創作が可能となり，アート文化はさらなる発展を遂げる可能性がある．これはおそらく日本書道においても例外ではない．そこで，我々は日本書道作品の生成に取り組む．

書道作品を生成するためには，文字生成の問題を解く必要がある．多くの場合，文字生成はフォント生成の問題として解かれている．フォント生成では，中国語や韓国語を対象とした研究が多く存在する．これらの研究では，フォントのスタイル特徴はストロークやコンポーネントに分割して表現されることが多い．ストロークとは，文字の 1 画 1 画のことである．コンポーネントとは，例えば「叶」という字であれば「口」と「十」で分けることができる．この「口」と「十」をそれぞれコンポーネントと呼ぶ．中国語や韓国語はストロークやコンポーネントによって分割し分類することが容易であり，コンポーネントやストローク情報を学習することで，きめ細かい文字生成が可能になる．しかし，日本語のひらがなやカタカナはコンポーネントやストロークで分割して分類することが難しい．したがって，日本語の場合ストロークや構成要素ではなく，文字全体に基づく特徴表現を使用する必要がある．

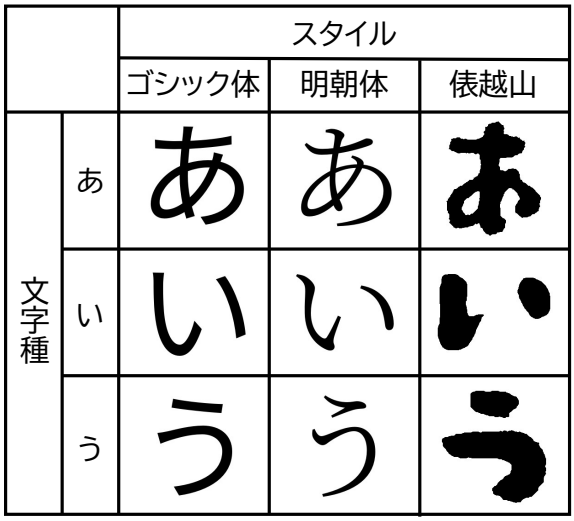


図 1　グリフにおける文字種とスタイルの例

そして，書道文字は手書きであるためフォントとは異なる性質を持つ．図 1 に示すように，文字は**文字種**と**スタイル**を持つ．文字種は文字の種類であり，スタイルは文字種を具現化するためのデザインである．本研究ではこの文字種とスタイルが組み合わさって具現化されたものを**グリフ**と呼ぶ．フォント生成では，1 つの文字種と 1 つのスタイルが決まれば，グリフは 1 つに定まる．しかし，手書き生成では 1 スタイル中に同じ文

1：https://www.nextrembrandt.com
2：https://github.com/CompVis/stable-diffusion

字種の様々なグリフがある．例えば，ある 1 人の人が 2 回同じ文字種のグリフを書いたとき，この 2 つのグリフの形が完全に一致することはない．1 つの文字種と 1 つのスタイルが定まっても，グリフが一意に定まるわけではない．唯一無二の正解が存在しないため，手書き生成はフォント生成より難しい問題である．この問題には文字種やスタイル特徴がユニークに決まらない方法が必要である．

日本書道作品を生成するためには，日本語を対象とした手書き文字生成に取り組む必要がある．よって，我々は文字全体に基づく表現を用いて，文字種とスタイルのエンコードを行うことで，手書きの日本書道文字を生成することに取り組む．

2 関連研究

2.1 Image-to-Image translation

グリフ生成タスクは一種の Image-to-Image translation（I2I）である．I2I はソースドメインとターゲットドメインの間のマッピング関数を学習するタスクである．I2I ではソース画像のコンテントとターゲットドメインのスタイルを融合する．I2I 手法の一つに GAN [4] をベースにした pix2pix [6] がある．pix2pix は U-Net [15] を用いて，画像のコンテントを維持してスタイルを変換することができる．U-Net はエンコーダの特徴マップとデコーダの特徴マップを結合するスキップ接続を持つ．I2I 手法である MUNIT [5] や FUNIT [10] は，コンテントエンコーダとスタイルエンコーダによって入力画像からコンテンツ特徴とスタイル特徴を抽出する．

本論文では，我々はグリフの文字種特徴とスタイル特徴を抽出するために文字種エンコーダとスタイルエンコーダを用いる．文字種エンコーダはデコーダとのスキップ接続を持つ．これは文字種特徴を各層で残すことで，可読なグリフを生成するためである．

2.2 Font Generation with GAN

フォント生成タスクはフォントデザインの作業をいかにして自動化するかという問題を持つ．この問題において，GAN を利用したフォント生成の研究が多く存在する．初めて中国語フォント生成に GAN を用いた手法として，Yuchen が提案した zi2zi [18] がある．zi2zi は pix2pix をベースとしたフォント生成モデルである．pix2pix をベースに ACGAN [12] と DTN [16] の手法を用いることで，文字種とスタイルの特徴抽出を行う．zi2zi はスタイルの学習に埋め込み表現を用いている．埋め込み表現では予め決めたカテゴリ別に特徴表現を保持しておく．そのため，予め学習したスタイルしか生成することができない．DCFont [7] は，zi2zi の構成をもとにフォント特徴再構成ネットワークを追加している．フォント特徴再構成ネットワークは未見のグリフの特徴を推定し，生成器から与えられた文字種のグリフを合成する．SCFont [8] は，入力グリフの骨格をターゲットグリフの骨格に変換し，その骨格に対し GAN モデルによってスタイルを付与する．ZiGAN [19] は，文字種情報を保持するとともにスタイルを変換する．これはペアではない多数のグリフを利用し，異なるスタイルの特徴をヒルベルト空間に写像し，特徴分布を整列させることで実現される．SE-GAN [21] は筆文字フォントを対象としたフォント生成を行う．一般のフォントは，基本的な構造や配置に共通点がある．しかし筆文字フォントは大きく歪んでいることが多い．

本論文では，我々は zi2zi をもとに，文字種だけでなくスタイルもエンコード可能なモデルを提案する．エンコードは特徴表現をカテゴリごとに保持せず，特徴表現を入力から抽出する．我々は埋め込みではなくエンコードすることで，学習したスタイルのみにとらわれない生成器を作成する．

2.3 Few-shot Font Generation

Few-shot フォント生成タスクは，非常に少ないスタイルを参照するだけで新しいフォントライブラリを生成することを目的としている．Few-shot フォント生成は，グリフから文字種とスタイルを切り離すことをよく行う．文字種とスタイルの抽出の方法は，以下の 2 種類に分けることができる．

- グローバル特徴表現
- コンポーネントに基づく特徴表現

グローバル特徴表現は，グリフ全体から，文字種特徴とスタイル特徴を抽出する．EMD [24] や DG-Font [20] は，文字種特徴とスタイル特徴を両方抽出し，2 つの特徴量を合わせることで新たなグリフを生成する．MLFont [3] は，文字種グリフ 1 枚から得られた文字種特徴と，スタイルグリフ複数枚から得られたスタイル特徴量を合わせた特徴量をデコーダに入力する．

コンポーネントに基づいた特徴表現は，グリフをコンポーネントやストロークで表現する事が多い．FsFont [17] は，参照スタイルグリフからコンポーネントごとにスタイル特徴を抽出し，文字種を参照スタイルで構成している．StrokeGAN [22] は，任意の漢字に対してストロークエンベディングを行う．CG-GAN [9] は，コンポーネントによってスタイルと文字種を切り離すように生成器に制約をかけている．LF-Font [14] はコンポーネントごとのスタイル表現を学習する．XMP-Font [11] は，ストロークレベルでグリフを分割し，入力をストロークラベルにしている．

本論文では，我々はグローバル特徴表現を用いて，文字種グリフとスタイルグリフ両方から文字種表現とスタイル表現を両方抽出する．

2.4 Unsupervised Representation Learning

Unsupervised representation learning は，アノテーションなしでタスクに使える特徴表現を得ることを目的としている．

Unsupervised representation learning の手法の一つに SimSiam がある．SimSiam は Siamese Network [1] を基にしている．Siamese Network はクラス分類タスクを解かずに特徴量同士の距離を学習することで特徴表現を得る．Siamese Network はペアデータを入力とする．1 つのペアデータが同じクラスであれば正例とし，違うクラスであれば負例とする．ペアデータはそれぞれ同じニューラルネットワークに入力される．ニューラルネットワークからそれぞれ出力された特徴量同士は比較さ

れ，正例であればこれらの特徴量同士が近くなるように学習し，負例であれば特徴量同士が遠ざかるように学習する．SimSiam も Siamese Network と似た学習を行うが，正例のみを用いる．

本論文では，我々は SimSiam を用いて文字種とスタイル両方の特徴表現を得る．負例を用いる手法は，負例同士の類似を考慮しない．例えば，遊ゴシックと BIZ UD ゴシックは共にゴシック体であるため，とても似ている．これらは違うフォントであるため，Siamese Network はこれらの特徴量を遠ざけるように学習してしまう．SimSiam は負例での学習を行わないため，似たスタイル同士を遠ざけないような学習ができる．

3 問題定義とデータセット

3.1 問題定義

文字種は文字の種類である．**スタイル**は文字種を具現化するためのデザインである．**グリフ**は文字種とスタイルが合わさって具現化されたものである．

文字種集合 C は n 文字の文字種 $c_i(i = 1, 2, \cdots, n)$ で構成される．

$$C = \{c_1, c_2, \cdots, c_n\} \tag{1}$$

例えば，「あ」と「い」の 2 種類の文字種で構成される文字種集合 $C = \{c_1, c_2\}$ であれば，c_1 は「あ」，c_2 は「い」となる．文字種集合 C のうち任意の文字種は $c_\bullet$ と表す．本論文では $n = 6,523$ 種類の文字種を使用する．

スタイル集合 S は m 個のスタイル $s_j(j = 1, 2, \cdots, m)$ で構成される．

$$S = \{s_1, s_2, \cdots, s_m\} \tag{2}$$

例えば，ゴシック体と明朝体の 2 種類のスタイルで構成されるスタイル集合 $S = \{s_1, s_2\}$ であれば，s_1 はゴシック体，s_2 は明朝体となる．スタイル集合 S のうち，任意のスタイルは $s_\bullet$ と表す．本論文では $m = 41$ 種類のスタイルを使用する．

文字種 c_i とスタイル s_j を持つグリフを I_{c_i,s_j} と表すこととする．同じ文字種 c_i を持ち，スタイル集合 S のスタイルをそれぞれ持つグリフの集合を R_{c_i} とする．

$$R_{c_i} = \{I_{c_i,s_1}, I_{c_i,s_2}, \cdots, I_{c_i,s_m}\} \tag{3}$$

同じスタイル s_j を持ち，文字種集合 C の文字種をそれぞれ持つグリフの集合を R_{s_j} とする．

$$R_{s_j} = \{I_{c_1,s_j}, I_{c_2,s_j}, \cdots, I_{c_n,s_j}\} \tag{4}$$

我々が解く問題は，ある文字種を持ったグリフと，あるスタイルを持ったグリフの 2 枚が入力されたときに，それらの文字種とスタイルを持つグリフを出力する問題である．もし入力画像が文字種 c_i と任意のスタイル $s_\bullet$ を持つグリフ画像 $I_{c_i,s_\bullet}$ と，スタイル s_i と任意の文字種 $c_\bullet$ を持つグリフ画像 $I_{c_\bullet,s_j}$ である場合に，生成器 $\mathbf{G}$ は文字種 c_i とスタイル s_j を持つグリフ画像 I_{c_i,s_j} を出力する．我々はそのような生成器 $\mathbf{G}$ を作成する．

$$I_{c_i,s_j} = \mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j}) \tag{5}$$

3.2 日本語グリフデータセット

日本書道作品と，くずし字データセットと，日本語フォントからデータを集めた．日本書道作品とくずし字データセットを利用する理由は，本論文で日本語の手書きグリフに焦点を当てているからである．日本語フォントは手書きではないが，多数のスタイルで学習を行い日本語のベースとなるグリフを効率的に学習させるために使用する．それぞれのスタイルから，ひらがな 84 文字，カタカナ 84 文字，漢字 6,355 文字の合計 6,523 文字種の文字種集合 C に含まれるグリフを利用する．

日本書道グリフデータ．我々は書家の俵越山氏 [3]から，119 件の書道作品の提供を受けた．119 件のうち，100 件が俵越山氏の X アカウント [4]に投稿された作品，6 件が俵越山氏から直接提供された作品，13 件が向日町警察署カレンダーに収録された作品である．1 作品に複数のグリフが書かれており，文字種の重複がある．これらの作品中の全てのグリフに対して，最小のバウンディングボックスで切り抜いた．これらのグリフのうち，文字種集合 C に含まれる文字種を持つグリフを利用する．日本書道データの合計は 1 スタイルで 1,340 グリフとなる．

くずし字データ．くずし字データセット [13] は日本古典籍から取得されたくずし字のグリフからなるデータセットである．くずし字は昔の日本で使われていた文字の書き方である．くずし字データセットのグリフは 44 冊の日本古典籍からなる．このうちグリフが 100 枚以上ある 21 冊を選んだ．同一書籍上のグリフはスタイルも同じであるとする．これらのグリフのうち，文字種集合 C に含まれる文字種を持つグリフを利用する．くずし字データの合計は 21 スタイルで 22,441 グリフである．

日本語フォントデータ．我々は Google Fonts [5]の日本語フォントのうちスタイルが似通っていないと人手で判断したものを 19 個選んだ．我々は各フォントからひらがな 84 文字，カタカナ 84 文字，漢字 6,355 の合計 6,523 文字を使用する．日本語フォントデータの合計は 19 スタイルで 123,937 グリフである．

我々は全てのグリフに対して以下のような画像処理を行った．全てのグリフはそれぞれ最小のバウンディングボックスで切り抜かれているため，縦と横の長さが違う．まず，各グリフ画像を長辺が 236 ピクセルになるように，同じ縦横比のまま拡大縮小する．そして各グリフの長辺の両端を 10 ピクセルずつパディングし，短辺を 256 ピクセルになるようにパディングすることで，256×256 ピクセルの画像にする．最後にこれらの画像に対してグレースケール化と 2 値化を行う．

我々は合計 41 スタイル，6,523 文字種，147,718 グリフを用いる．ここから，評価のために 41 スタイルのうち日本書道グリフデータの etsuzan スタイル，くずし字データの brsk00000 スタイル，日本語フォントデータの ReggaeOne スタイルを選んだ．ReggaeOne スタイルの 50 グリフ，brsk00000 スタイルの 50 グリフ，etsuzan スタイルの 54 グリフを評価データとする．ReggaeOne スタイルと brsk00000 スタイルでは 50 グリ

3：http://etsuzan.jp
4：https://x.com/etsuzan_tawara
5：https://fonts.google.com

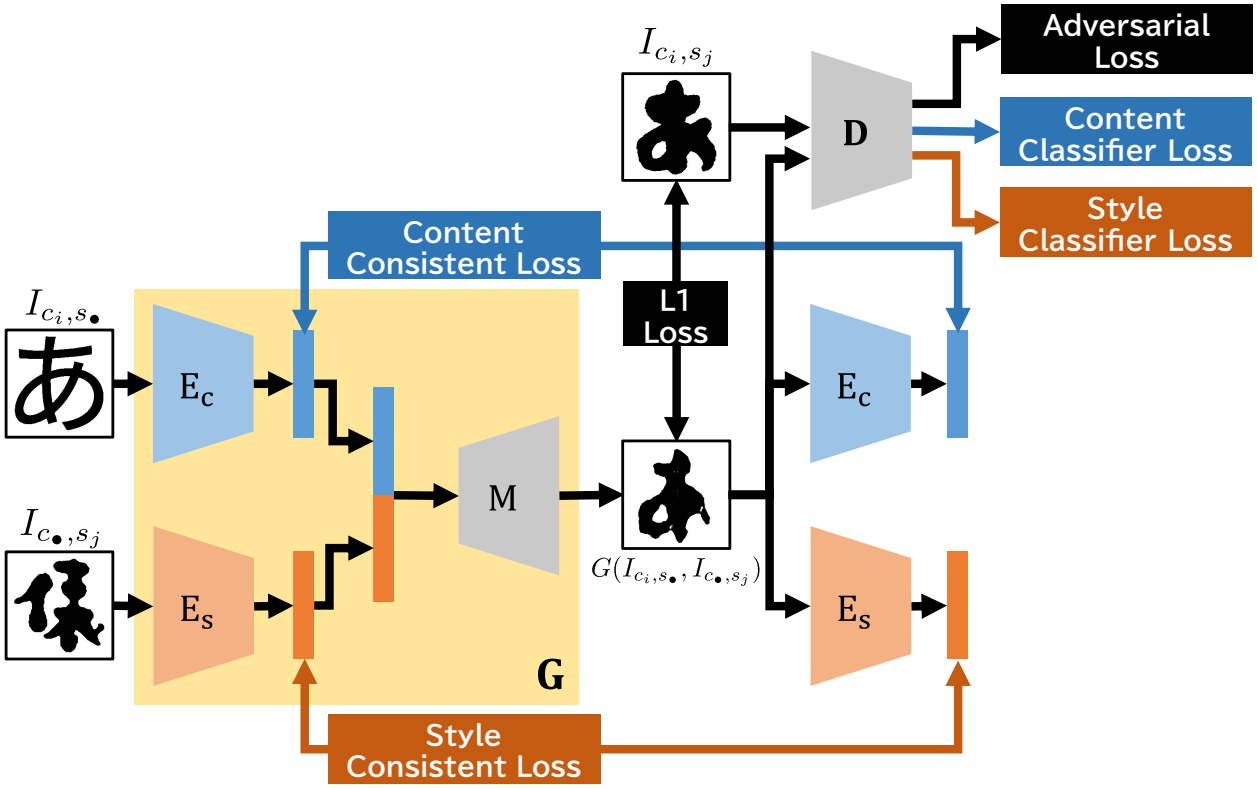


図 2　学習ネットワーク

フずつランダムに選択した．etsuzan スタイルでは，54 枚のグリフは，etsuzan スタイルグリフの中で，向日町警察署カレンダーのグリフにのみ現れる文字種を持つグリフを選択した．

評価データを省いた残りの 147,564 グリフが学習用データである．

4　グリフ生成モデル

今回提案するモデルは，生成器 **G** と識別器 **D** で構成される．モデルを図 2 に示す．生成器 **G** は，文字種エンコーダ E_c，スタイルエンコーダ E_s，ミキサー M からなる．

入力画像は生成したい文字種 c_i を持つグリフ画像 $I_{c_i,s_\bullet}$ と，生成したいスタイル s_j を持つグリフ画像 $I_{c_\bullet,s_j}$ である．これらの画像を入力したときに，文字種 c_i とスタイル s_j を持つグリフ画像 I_{c_i,s_j} を生成することを生成器 **G** の目的とする．識別器 **D** は，生成器 **G** が生成したグリフ画像とターゲットグリフ画像を受け取る．識別器 **D** はそれぞれの入力画像に対して，本物か偽物で 2 値分類，スタイルで m 値分類，文字種で n 値分類を行う．

4.1　生　成　器

グリフ画像 $I_{c_i,s_\bullet}$ と，グリフ画像 $I_{c_\bullet,s_j}$ が，生成器 **G** に与えられたとき，生成される画像は文字種 c_i とスタイル s_j を保持している必要がある．そこで，我々は文字種エンコーダ E_c を用いて文字種グリフ画像 $I_{c_i,s_\bullet}$ から文字種ベクトル $E_c(I_{c_i,s_\bullet})$ を抽出し，我々はスタイルエンコーダ E_s を用いてスタイルグリフ画像 $I_{c_\bullet,s_j}$ からスタイルベクトル $E_s(I_{c_\bullet,s_j})$ を抽出する．文字種ベクトル $E_c(I_{c_i,s_\bullet})$ とスタイルベクトル $E_s(I_{c_\bullet,s_j})$ は結合されミキサー M に入力される．ミキサー M は文字種 c_i とスタイル s_j を持つグリフ画像 I_{c_i,s_j} を生成することを目的とする．

生成器 **G** は U-Net を用いる．生成器 **G** を図 3 に示す．文字種エンコーダ E_c とスタイルエンコーダ E_s はそれぞれ 8 層の畳み込み層からなる．各層の間には，活性化関数 Leaky ReLU が用られ，各層の出力はバッチ正規化される．各エンコーダには 256×256 ピクセルの画像が入力され，512 次元のベクトルを出力する．2 つの 512 次元ベクトルを結合した 1,024 次元ベクトルはミキサー M に入力される．ミキサー M は 8 層の逆畳み込み層で構成される．各層の前には，活性化関数 ReLU が用いられ，各層の出力はバッチ正規化される．文字種エンコーダ E_c はスキップ接続を行う．文字種エンコーダ E_c の各層での出力は，ミキサー M の対応する層に渡され，前の層からの出力と結合される．スタイルエンコーダ E_s はスキップ接続を行わない．ミキサー M は 256×256 ピクセルの画像を出力する．

4.2　識　別　器

識別器 **D** を図 4 に示す．識別器 **D** では，生成器 **G** から出力されたグリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ またはターゲットグリフ画像 I_{c_i,s_j} を入力とする．識別器 **D** は入力された画像から True/False 分類と content 分類と style 分類を行うことを目的とする．

識別器 **D** は 6 層の畳み込み層と Dropout 層と全結合層（FC 層）で構成される．各層の前には Leaky ReLU が用いられ，各層の出力はバッチ正規化される．Dropout 層は Dropout 率を 0.5 に設定される．FC 層は線形層を用いる．

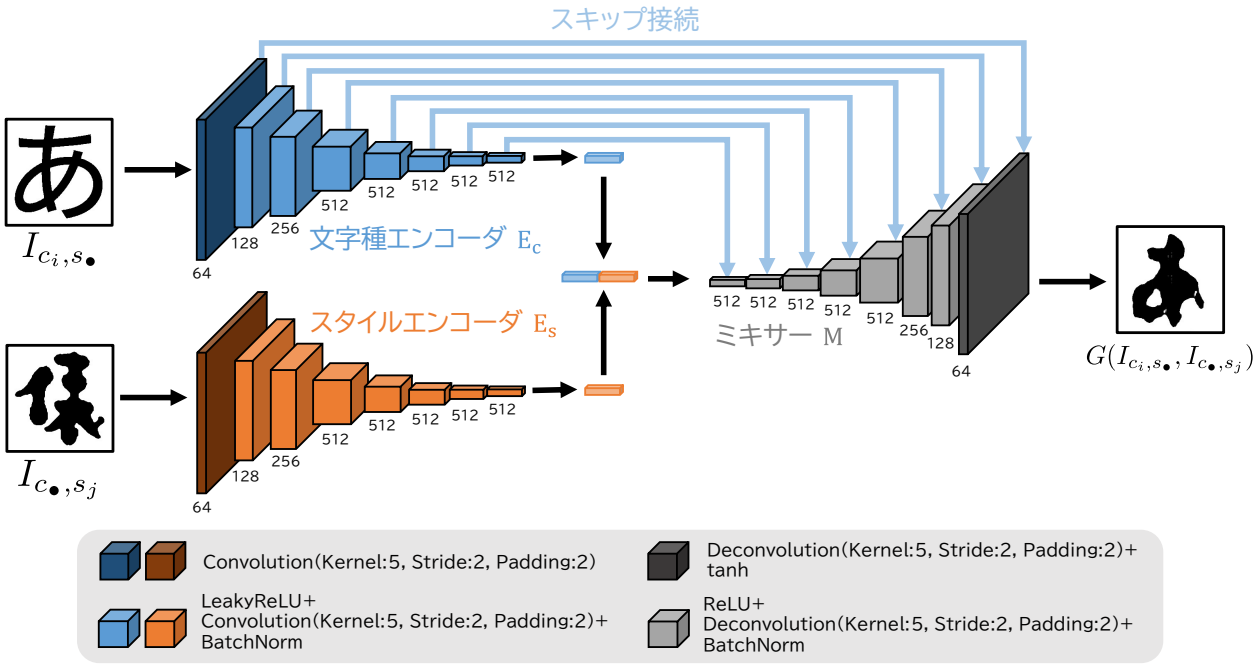


図 3　生　成　器

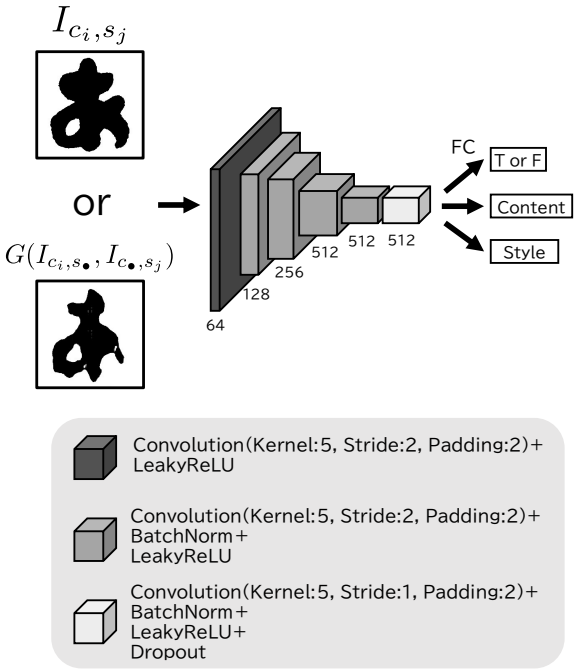


図 4　識　別　器

4.3　SimSiam によるエンコーダの事前学習

事前に適した特徴表現を得るために，文字種エンコーダ $\mathrm{E_c}$ とスタイルエンコーダ $\mathrm{E_s}$ をそれぞれ事前学習した．事前学習には SimSiam [2] の手法を用いた．学習ネットワークを図 5 に示す．SimSiam は正例しか学習しないため，似たスタイルや似た文字種のものを互いに近く学習できると考えている．

文字種エンコーダ $\mathrm{E_c}$ の学習では，同じ文字種を持つペア画像を用いる．まずランダムに文字種 c_i を選ぶ．次に訓練データ中のグリフ集合 R_{c_i} から，ランダムに 2 つのグリフ $I_{c_i,s_\bullet}, I_{c_i,s'_\bullet}$ が選ばれる．この 2 つのグリフがペア画像である．2 枚の画像はそれぞれ文字種エンコーダ $\mathrm{E_c}$ に入力される．文字種エンコーダの出力の一方は文字種 Predictor $\mathrm{P_c}$ に入力される．そして，それぞれの出力から Cos 類似度を算出する．このとき文字種 Predictor $\mathrm{P_c}$ に入力されない出力の一方は勾配停止する．そして，2 枚の画像を入れ替えて同様のことを行う．2 通りの Cos 類似度を足し合わせた以下のロス関数 $\mathcal{L}_C$ を用いて，文字種エンコーダ $\mathrm{E_c}$ の学習を行う．

$$\mathcal{L}_{sim}^{C} = \frac{\mathrm{E_c}(I_{c_i,s_\bullet}) \cdot \mathrm{P_c}(\mathrm{E_c}(I_{c_i,s'_\bullet}))}{\|\mathrm{E_c}(I_{c_i,s_\bullet})\| \|\mathrm{P_c}(\mathrm{E_c}(I_{c_i,s'_\bullet}))\|} + \frac{\mathrm{P_c}(\mathrm{E_c}(I_{c_i,s_\bullet})) \cdot \mathrm{E_c}(I_{c_i,s'_\bullet})}{\|\mathrm{P_c}(\mathrm{E_c}(I_{c_i,s_\bullet}))\| \|\mathrm{E_c}(I_{c_i,s'_\bullet})\|} \tag{6}$$

スタイルエンコーダ $\mathrm{E_s}$ の学習は，同じスタイルを持つペア画像を用いる．まずランダムにスタイル s_j を選ぶ．次に訓練データ中のグリフ集合 R_{s_j} から，ランダムに 2 つのグリフ $I_{c_\bullet,s_j}, I_{c'_\bullet,s_j}$ が選ばれる．この 2 つのグリフがペア画像である．2 枚の画像はそれぞれスタイルエンコーダ $\mathrm{E_s}$ に入力される．スタイルエンコーダの出力の一方はスタイル Predictor $\mathrm{P_s}$ に入力される．そして，それぞれの出力から Cos 類似度を算出する．このときスタイル Predictor $\mathrm{P_s}$ に入力されない出力の一方は勾配停止する．そして，2 枚の画像を入れ替えて同様のことを行う．2 通りの Cos 類似度を足し合わせた以下のロス関数 $\mathcal{L}_{sim}^{S}$ を用いてスタイルエンコーダ $\mathrm{E_s}$ の学習を行う．

$$\mathcal{L}_{sim}^{S} = \frac{\mathrm{E_s}(I_{c_\bullet,s_j}) \cdot \mathrm{P_s}(\mathrm{E_s}(I_{c'_\bullet,s_j}))}{\|\mathrm{E_s}(I_{c_\bullet,s_j})\| \|\mathrm{P_s}(\mathrm{E_s}(I_{c'_\bullet,s_j}))\|} + \frac{\mathrm{P_s}(\mathrm{E_s}(I_{c_\bullet,s_j})) \cdot \mathrm{E_s}(I_{c'_\bullet,s_j})}{\|\mathrm{P_s}(\mathrm{E_s}(I_{c_\bullet,s_j}))\| \|\mathrm{E_s}(I_{c'_\bullet,s_j})\|} \tag{7}$$

4.4　GAN による生成器全体の学習

我々は SimSiam で事前学習した文字種エンコーダ $\mathrm{E_c}$ とスタ

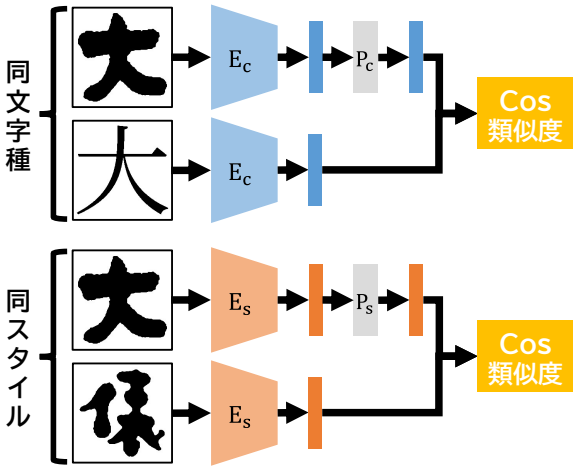


図 5　事 前 学 習

イルエンコーダ $\mathrm{E_s}$ を生成器 $\mathbf{G}$ に持ってくる．モデルの学習に用いる損失関数を以下に示す．

L1 Loss. 我々は生成されたグリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ とターゲットグリフ画像 I_{c_i,s_j} との間の平均絶対誤差を計算する．この損失関数は生成画像をピクセルレベルでターゲットグリフに近づけるために設定する．

$$\mathcal{L}_{L1} = \|I_{c_i,s_j} - \mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})\|_1 \tag{8}$$

Content Consistent Loss. 生成グリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ を文字種エンコーダ $\mathrm{E_c}$ とスタイルエンコーダ $\mathrm{E_s}$ に両方入力する．我々は入力文字種グリフ画像 $I_{c_i,s_\bullet}$ から得られた文字種ベクトル $\mathrm{E_c}(I_{c_i,s_\bullet})$ と，生成グリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ から得られた文字種ベクトル $\mathrm{E_c}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j}))$ の間の平均二乗誤差を計算する．この損失関数は生成グリフと入力グリフの文字種特徴を類似させるために設定する．

$$\mathcal{L}^C_{con} = \|\mathrm{E_c}(I_{c_i,s_\bullet}) - \mathrm{E_c}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j}))\|_2^2 \tag{9}$$

Style Consistent Loss. 我々は入力スタイルグリフ画像 $I_{c_\bullet,s_j}$ から得られたスタイルベクトル $\mathrm{E_s}(I_{c_\bullet,s_j})$ と，生成グリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ から得られたスタイルベクトル $\mathrm{E_s}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j}))$ の間の平均二乗誤差を計算する．この損失関数は生成グリフと入力グリフのスタイル特徴を類似させるために設定する．

$$\mathcal{L}^S_{con} = \|\mathrm{E_s}(I_{c_\bullet,s_j}) - \mathrm{E_s}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j}))\|_2^2 \tag{10}$$

Content Classifier Loss. 我々はターゲットグリフ画像 I_{c_i,s_j} と生成グリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ をそれぞれ識別器 $\mathbf{D}$ に入力する．識別器 $\mathbf{D}$ によるスタイル予測を行い，予測ラベルと正解ラベル c_i で Cross Entropy Loss を計算する．以下，損失関数内では Cross Entropy loss を CE と表す．この損失関数は文字種 c_i であることが明確なグリフを生成器 $\mathbf{G}$ に生成させるために設定する．

$$\mathcal{L}^C_{cls} = \mathrm{CE}(\mathbf{D}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})), c_i) + \mathrm{CE}(\mathbf{D}(I_{c_i,s_j}), c_i) \tag{11}$$

Style Classifier Loss. 我々はターゲットグリフ画像 I_{c_i,s_j} と生成グリフ画像をそれぞれ識別器 $\mathbf{D}$ に入力する．識別器 $\mathbf{D}$ はスタイルを予測する．我々は識別器 $\mathbf{D}$ の予測と正解ラベル s_j で Cross Entropy loss を計算する．この損失関数はスタイル s_j であることが明確なグリフを生成器 $\mathbf{G}$ に生成させるために設定する．

$$\mathcal{L}^S_{cls} = \mathrm{CE}(\mathbf{D}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})), s_j) + \mathrm{CE}(\mathbf{D}(I_{c_i,s_j}), s_j) \tag{12}$$

Adversarial Loss. 我々はターゲットグリフ画像 I_{c_i,s_j} と生成グリフ画像 $\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})$ をそれぞれ識別器 $\mathbf{D}$ に入力する．識別器 $\mathbf{D}$ は本物か偽物か予測する．我々は Hinge Loss を用いて敵対的ロスを計算する．この損失関数は生成画像のクオリティを上げるために設定する．

$$\begin{aligned}\mathcal{L}^G_{adv} &= -\mathbf{D}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})) \\ \mathcal{L}^D_{adv} &= \min(\mathbf{D}(I_{c_i,s_j}) - 1), 0) \\ &\quad + \min(-\mathbf{D}(\mathbf{G}(I_{c_i,s_\bullet}, I_{c_\bullet,s_j})), 0)\end{aligned} \tag{13}$$

5　実　　験

5.1　実 験 設 定

まずエンコーダの事前学習を行う．我々は学習用データセットのグリフにおいて文字種が同じ割合になるように SimSiam 用の訓練データと検証データに分割する．我々は分割した訓練データによって文字種エンコーダ $\mathrm{E_c}$ のために SimSiam 学習を行う．SimSiam における predictor の隠れ層の次元は 256 次元とした．我々は検証データを用いて，文字種エンコーダ $\mathrm{E_c}$ の性能を測った．文字種エンコーダ $\mathrm{E_c}$ の kNN 正解率は 0.916 であった．我々は同様のことをスタイルエンコーダ $\mathrm{E_s}$ でも行う．スタイルエンコーダ $\mathrm{E_s}$ の kNN 正解率は 0.647 である．

事前学習したエンコーダを用いて生成器全体の学習を行う．学習用データ 147,564 グリフのそれぞれをターゲットグリフとする．etsuzan スタイルはグリフ数が少ないため評価データを省いた 1,286 枚を 5 回利用する．ターゲットグリフは合計 152,708 枚になる．我々は各ターゲットグリフに対して，文字種が同じグリフと，スタイルが同じグリフをランダムに取得する．1 ステップごとに識別器 $\mathbf{D}$ は 1 回最適化され，生成器 $\mathbf{G}$ は 2 回最適化される．生成器 $\mathbf{G}$ と識別器 $\mathbf{D}$ はどちらもオプティマイザに Adam を用いる．Adam のハイパーパラメータは $\beta_1 = 0.5$，$\beta_2 = 0.999$ とする．学習率は 0.001 とする．モデルの学習は 30 エポック行う．

5.2　評　　価

ベースラインモデルの zi2zi と我々の提案モデルの間で比較を行う．生成対象のスタイルは評価データにある etsuzan スタイル，brsk00000 スタイル，ReggaeOne スタイルである．etsuzan スタイルと brsk00000 スタイルの 2 つを選択する理由は，本論文では日本語の手書きグリフに焦点を当てているからである．そして，ReggaeOne スタイルを選んだ理由は，フォントスタイルから特徴的なスタイルを比較したいからである．

表 1　定性的比較



表 2　定量的比較

モデル	MAE↓	SSIM↑	LPIPS↓
zi2zi	0.2308	0.5480	0.2321
Ours	**0.2096**	**0.5870**	**0.2214**

表 3　Ablation Study



	Ours	(a)	(b)	(c)	(d)
MAE↓	0.2096	0.2276	**0.2093**	0.2168	0.2315
SSIM↑	**0.5870**	0.5523	0.5826	0.5610	0.5532
LPIPS↓	**0.2214**	0.2369	0.2320	0.2267	0.2494

5.2.1　定性的比較

生成結果の例は表 1 に示される．Content 列は実際に入力された文字種グリフである．Target 列は評価データである．手書きグリフ生成では Target グリフと完全一致する必要はないため，Target 列は参考である．まず etsuzan スタイルの生成結果を確認する．zi2zi の生成結果はグリフの一部分が欠けていたり，線がガタガタしている．我々のモデルはグリフの綺麗さを保ったまま etsuzan スタイルを生成できている．次に，brsk00000 スタイルの生成結果について確認する．brsk00000 スタイルは現代の日本語グリフとはかなり違う形のグリフもあり難しい生成である．我々のモデルによる生成結果は線がガタガタしておらず zi2zi より綺麗である．最後に ReggaeOne スタイルの生成結果について確認する．このスタイルは zi2zi も我々のモデルも綺麗に生成できている．ReggaeOne はフォントである．フォント生成では文字種とスタイルが決まればグリフも一意に定まるため，簡単な問題であると言える．

5.2.2　定量的比較

生成された画像とターゲット画像との類似性を評価するために，生成画像とターゲット画像との間で MAE，SSIM [25]，LPIPS [23] を算出する．MAE と SSIM はピクセルレベルでの類似度を測定する．MAE は各ピクセルで測定するのに対し，SSIM は各ピクセルの周辺の相関を測定する．LPIPS は深層学習に基づいた特徴抽出によって，画像間の知覚的類似性を定量化する．我々は etsuzan スタイルを対象に 3 つの評価指標でスコアを算出した．結果は表 2 に示される．MAE では我々のモデルが 0.0212 ポイント優位である．SSIM では我々のモデルが 0.039 ポイント優位である．LPIPS では我々のモデルが 0.0107 ポイント優位である．MAE，SSIM，LPIPS のいずれにおいても我々の提案モデルがベスト値を示した．

5.2.3　Ablation Study

我々のモデルについて，損失関数の影響やエンコーダの事前学習の影響を調べる．評価するスタイルは etsuzan スタイルである．結果の例を表 3 に示す．損失関数を $\mathcal{L}_{cls} = \mathcal{L}^C_{cls} + \mathcal{L}^S_{cls}$ と $\mathcal{L}_{adv} = \mathcal{L}^G_{adv} + \mathcal{L}^D_{adv}$ と定義する．損失関数のうち $\mathcal{L}_C$，$\mathcal{L}_S$ なしの場合 (a)，損失関数のうち $\mathcal{L}_{adv}$，$\mathcal{L}_C$，$\mathcal{L}_S$ なしの場合 (b)，事前学習したエンコーダをフリーズさせた場合 (c)，エンコーダの事前学習を行わなかった場合 (d) について調べた．(a) や (b) はグリフの一部分がかすれている．(c) ではグリフの線が水平と垂直である．(d) ではグリフが角ばったスタイルで生成されている．SSIM と LPIPS については我々の提案モデルが良いスコアを示した．MAE では我々のモデルのスコアとトップスコアとの差は 0.003 ポイントのみであった．この結果はエンコーダを事前学習した上でフリーズせずに，識別器で True/False 分類と文字種分類とスタイル分類を同時に解くと良いことを示した．

6　まとめと今後の課題

本論文では，我々は日本書道作品における手書きグリフ生成手法を提案する．我々は文字種特徴とスタイル特徴を抽出してグリフ生成を行う生成器を作成する．この生成器を SimSiam や GAN の手法を用いて学習する．我々は日本語フォントと，くずし字データセットと，日本書道作品からデータセットを作成する．我々は 3 つの評価指標 MAE，SSIM，LPIPS を用いて我々の手法とベースラインモデルを比較する．いずれの評価指標

においても我々の提案モデルの方が良い値を示した．Ablation study を我々の提案モデルで行い，損失関数やエンコーダの事前学習の有無による効果を比較する．SSIM と LPIPS においては我々の提案モデルが良いスコアを示した．生成グリフのバリエーションを増やすことには取り組んでいない．今後はグリフのバリエーションを表現できるモデルを作成したい．

謝　　辞

本研究は JSPS 科研費 JP21H03775，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification using a "Siamese" Time Delay Neural Network. In *Proceedings of the 7th International Conference on Neural Information Processing Systems (NIPS'93)*, pp. 737–744, 1993.

[2] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pp. 15745–15753, 2021.

[3] Xu Chen, Lei Wu, Minggang He, Lei Meng, and Xiangxu Meng. MLFont: Few-Shot Chinese Font Generation via Deep Meta-Learning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR'21)*, pp. 37–45, 2021.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, pp. 2672–2680, 2014.

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-image Translation. In *Proceedings of the 15th European Conference on Computer Vision (ECCV'18)*, pp. 179–196, 2018.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-To-Image Translation With Conditional Adversarial Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pp. 5967–5976, 2017.

[7] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. DCFont: An End-to-End Deep Chinese Font Generation System. In *Proceedings of the SIGGRAPH Asia 2017 Technical Briefs (SA'17)*, pp. 1–4, 2017.

[8] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. SCFont: Structure-Guided Chinese Font Generation via Deep Stacked Networks. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, pp. 4015–4022, 2019.

[9] Yuxin Kong, Canjie Luo, Weihong Ma, Qiyuan Zhu, Shenggao Zhu, Nicholas Yuan, and Lianwen Jin. Look Closer to Supervise Better: One-Shot Font Generation via Component-Based Discriminator. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pp. 13472–13481, 2022.

[10] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19)*, pp. 10550–10559, 2019.

[11] Wei Liu, Fangyue Liu, Fei Ding, Qian He, and Zili Yi. XMP-Font: Self-Supervised Cross-Modality Pre-Training for Few-Shot Font Generation. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pp. 7895–7904, 2022.

[12] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, pp. 2642–2651, 2017.

[13] National Institute of Japanese Literature Center for Open Data in the Humanities. Kuzushiji dataset. `http://doi.org/10.20676/00000340`, 2016. accessed on 11 July 2023.

[14] Song Park, Sanghyuk Chun, Junbum Cha, Bado Lee, and Hyunjung Shim. Few-shot Font Generation with Localized Style Representations and Factorization. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, pp. 2393–2402, 2021.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (MICCAI'15)*, pp. 234–241, 2015.

[16] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised Cross-Domain Image Generation. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR'17)*, pp. 1–15, 2017.

[17] Licheng Tang, Yiyang Cai, Jiaming Liu, Zhibin Hong, Mingming Gong, Minhu Fan, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-Shot Font Generation by Learning Fine-Grained Local Styles. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, pp. 7885–7894, 2022.

[18] Yuchen Tian. zi2zi: Master chinese calligraphy with conditional adversarial networks. `https://github.com/kaonashi-tyc/zi2zi`, 2017. accessed on 15 December 2022.

[19] Qi Wen, Shuang Li, Bingfeng Han, and Yi Yuan. ZiGAN: Fine-grained Chinese Calligraphy Font Generation via a Few-Shot Style Transfer Approach. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*, pp. 621–629, 2021.

[20] Yangchen Xie, Xinyuan Chen, Li Sun, and Yue Lu. DG-Font: Deformable Generative Networks for Unsupervised Font Generation. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, pp. 5126–5136, 2021.

[21] Shaozu Yuan, Ruixue Liu, Meng Chen, Baoyang Chen, Zhijie Qiu, and Xiaodong He. SE-GAN: Skeleton Enhanced Gan-Based Model for Brush Handwriting Font Generation. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME'22)*, pp. 1–6, 2022.

[22] Jinshan Zeng, Qi Chen, Yunxin Liu, Mingwen Wang, and Yuan Yao. StrokeGAN: Reducing Mode Collapse in Chinese Font Generation via Stroke Encoding. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, pp. 3270–3277, 2021.

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pp. 586–595, 2018.

[24] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating Style and Content for Generalized Style Transfer. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pp. 8447–8455, 2018.

[25] Wang Zhou, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.

TREXIV: 時刻スタンプ付き高次元データのための対話的可視化に基づくトレンド抽出手法

中村　栄太†　　伊藤　貴之††

† 京都大学　〒 606-8501 京都府京都市左京区吉田本町
†† お茶の水女子大学　〒 112-8610 東京都文京区大塚 2-1-1
E-mail: †eita.nakamura@i.kyoto-u.ac.jp, ††itot@is.ocha.ac.jp

あらまし　本稿では，時刻スタンプ付き高次元データのためのトレンド抽出手法について述べる．この種のデータ分析は，音楽や絵画を含む創作文化の進化研究で重要である．高次元の空間の中で進化分析に有意義なトレンドを抽出するためには，複数の創作ジャンルが混ざったデータから注目するデータ領域および時間区間を選択し，解釈可能な特徴量の方向を見出す必要がある．これには，高次元空間の中で時間方向と高い相関を持つ成分を自動で解析する方法に合わせて，対象とする文化領域の知識を持つ専門家がデータを確認しながらトレンドの仮説の立案と検証を行える枠組みが有効だと考えられる．本研究では，対話的可視化手法に基づきユーザがデータの観察を行いながら，データ選択と特徴量方向の次元重みの設定を行うことで有意義なトレンド抽出を行う手法を提案する．ポピュラー音楽のスタイル進化解析への応用例では，多面的なデータ可視化と対話的操作の有効性が確認され，実際に複数の非自明なトレンドを抽出することができた．

キーワード　情報可視化，トレンド抽出，文化進化解析，ビジュアルアナリティクス，音楽スタイル進化
Data visualization, trend extraction, cultural evolutionary analysis, visual analytics, music style evolution

1　はじめに

近年，文化発展に関する定量的理解を得る目的で，音楽 [1,2] や絵画 [3,4]，文学 [5] などの大規模データから創作スタイルの進化（時間変化）を調べる研究が盛んになっている．こうした研究では，創作物が高次元特徴量で表わされ，制作年などの時刻スタンプ付き高次元データの分析を行うことが多い．同種のデータは生物の形態進化の研究でも扱われ，共通する分析の問題も多い [6].

データに基づく進化解析では，方向性を持った変化（トレンド）の抽出が重要な課題である [7]．トレンドの存在は，背後に働く力の存在を示しており，進化過程の理解や予測につながるからである [8]．しかし，高次元の空間の中で進化分析に有意義なトレンドを抽出することは容易ではない．まず，創作物データでは，ジャンルなどを表すクラスターごとに異なるトレンドが存在することや特定の時間区間にのみ顕著なトレンドが存在することが多いため，適切なデータ領域と時間区間の選択が必要である [9]．クラスタリング手法 [10,11] を用いてデータを自動で分割することも考えられるが，自動分析のみではクラスターの解釈が必ずしも容易ではなく，創作文化に関する専門知識を持つ人によるデータ選択が必要なことも多い．また，高次元の特徴量には互いに強く相関する次元の組も多く存在する中で，有意義なトレンドに対応する解釈可能な特徴量の方向を見つけることも課題である．

これらの問題を解決するには，高次元空間の中で時間方向と高い相関を持つ成分を自動で解析する方法に合わせて，対象とする文化領域の知識を持つ専門家がデータを確認しながらトレンドの仮説の立案と検証を行える枠組みが有効だと考えられる．本研究では，対話的可視化手法に基づき，ユーザがデータの観察を行いながら，データ選択と特徴量方向の次元重みの設定を行うことで有意義なトレンド抽出を行う手法を提案する．本研究で提案する可視化システム TREXIV (trend extractor based on interactive visualization) の特徴は，以下の機能を有することである．

- 2 次元表現空間での時刻情報付きデータ表示，および，データ領域と時間区間の選択のためのインターフェース
- 非線形可視化空間にて任意の曲線に沿ってデータ点を選択することでその特徴量方向を確認するインタラクション
- 時間方向と高く相関する特徴量空間の方向の自動推定
- PCA（主成分分析）と特徴量次元散布図による補助情報の表示
- 相関プロットと 2 次元表現空間におけるヒートマップを用いた仮説の検証

本稿では，この手法の構成法について説明し，ポピュラー音楽のスタイル進化解析への応用例をもとに，本手法の有効性の検証を行う．

高次元データの可視化手法として，低次元表現空間での表示とデータ選択を対話的に行う手法 [12] や，特徴量次元間の関係性を次元散布図を用いて可視化する方法 [13] などが提案されている．本研究の手法は，時間方向という特別な次元を持つ高次元のデータを扱うために適切なデータ処理を行い，これらの高次元可視化手法を要素として取り入れた多面的な可視化表示を行う．また，時系列データを対象として，顕著なトレンドが見

られる区間を自動検出するトレンド検出に関する研究 [14, 15] や，時系列データの各時刻を多次元データと見なして次元削減する手法の研究 [16] などがある．本研究では，特徴量空間において解釈可能な方向で顕著なトレンドを示すものを見つけるという異なる問題を扱う．

2 可視化手法

2.1 問題設定

本手法の入力データは，時刻スタンプ t_n と高次元特徴量 $\boldsymbol{x}_n = (x_{ni})_{i=1}^I$ の対により表されるデータ点 n からなるデータセット $(t_n, \boldsymbol{x}_n)_{n=1}^N$ である（I は特徴量の次元数）．本手法の目標は，特徴量空間におけるトレンド方向 $\boldsymbol{\beta} = (\beta_i)_{i=1}^I$ で以下の性質を満たすものを抽出することである．

P1　ある時間区間において，時刻スタンプ t_n とトレンド特徴量 $y_n = \sum_{i=1}^I \beta_i x_{ni}$ は高い相関を持つ

P2　トレンド方向は解釈可能な特徴量の方向である

以下では，ある特徴量 z_n と時刻スタンプ t_n の間の相関係数 ρ_z のことを CTD (correlation with time direction) と呼ぶ．z_n と t_n の平均を $\bar{z}$ と $\bar{t}$ とすると，CTD ρ_z は次で定義される．

$$\rho_z = \frac{\sum_{n=1}^N (z_n - \bar{z})(t_n - \bar{t})}{\left[\sum_{n=1}^N (z_n - \bar{z})^2\right]^{1/2} \left[\sum_{n=1}^N (t_n - \bar{t})^2\right]^{1/2}} \qquad (1)$$

例えば，上記 P1 は「ある時間区間におけるトレンド特徴量 y_n の CTD が高い値を持つ」と言い換えられる．また，各時刻ごとに計算した z_n の平均値からなる時系列 $\bar{z}_t = \sum_n z_n \delta_{t_n,t} / \sum_n \delta_{t_n,t}$ を特徴量 z_n の平均進化曲線と呼ぶ．

本手法では，以下に説明する対話的可視化手法に基づき，ユーザがトレンド方向の仮説を立て，それを検証することにより，上記の条件を満たすトレンド方向を抽出する．この際，ユーザは各データ次元 i に対する重み $\tilde{\beta}_i$ を指定することによりトレンド方向の仮説を立てるものとし，この $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_i)_{i=1}^I$ を仮説的トレンド方向と呼ぶ．仮説的トレンド方向は，対話的操作の中で変更を繰り返すことが想定され，最終的にユーザが有意義だと思うものが得られた時に，本手法の目標が達成されたと言える．また，同じデータセットに対して，複数の有意義な仮説的トレンド方向が得られることもある．

2.2 可視化分析の手順

本手法で可視化表示される要素の一覧を以下に示す（図 1にスナップショットを示す）．対話的操作のインタフェースとなるものには要素名に † が付いている．

V1　概観パネル †：2 次元表現空間にプロットされたデータが表示され，時刻スタンプはマーカ色として表示される．このパネルはデータ選択のためのインターフェースとしても機能する．時間スライダーはトレンド分析に用いるデータの時間区間を選択するために用いる．

V2　特徴量表示パネル：選択されたデータの平均特徴量ベクトルと，その全体平均特徴量ベクトルからの差異が表示される．

V3　最適トレンド観察パネル：統計的手法により推定された最適トレンド特徴量の分布と平均進化曲線が表示される．

V4　PCA パネル：選択したデータの 2 次元 PCA 空間における散布図が表示される．

V5　次元散布図パネル：特徴量次元の 2 次元表現空間における散布図が表示され，平均進化曲線間の相関に基づく特徴量次元の関係が確認できる．各特徴量次元の CTD はマーカ色として表示される．

V6　次元データ表 †：各特徴量次元のラベルと特性値が表示される．この表は仮説的トレンド方向の重みを入力するインターフェースとしても機能する．

V7　仮説的トレンド方向表示パネル：選択されたデータの仮説的トレンド特徴量の分布が表示される．

V8　仮説的トレンド観察パネル：仮説的トレンド特徴量のデータ分布と平均進化曲線が表示される．

ユーザは，以下に示す手順で本可視化手法を使うものとする．

F1　データ選択フェーズでは，ユーザは概観パネルと特徴量表示パネル，最適トレンド観察パネルを用いてデータを観察し，注目するデータ領域と時間区間を対話的に選択する．

F2　仮説立案フェーズでは，ユーザは次元データ表と PCA パネル，次元散布図パネルなどを用いて，特徴量次元の関連と相互関係を観察し，興味がある仮説的トレンド方向を選択する．

F3　仮説検証フェーズでは，ユーザは仮説的トレンド方向表示パネルと仮説的トレンド観察パネルなどを用いて，仮説的トレンド方向における CTD の有意性とデータ分布を分析する．

以上のフェーズは複数回繰り返すことができ，これによりユーザは有意義な仮説的トレンド方向の探求や最適化を行う．

2.3 可視化方法

本節では，2.2節で述べた各可視化要素における可視化の方法について述べる．

2.3.1 概観パネル

2 次元空間内で高次元データを表示するために，t-SNE 法 [17] に基づく次元圧縮を行う．多次元尺度構成法 (MDS) など，他の次元圧縮手法を用いることも可能である．データ点間の距離尺度については，分析対象のデータに対して適したものを選択して用いる．例えば，一般的なデータについてはユークリッド距離を用いることが考えられる．また，3章のデータ解析例で扱う確率ベクトルデータに対しては Jensen-Shannon ダイバージェンスを用いることができる．時刻スタンプはマーカ色として表示し，赤（古い時刻）と緑（中間），青（新しい時刻）を内挿して表示色を決定する．

データ選択では，散布図内で閉曲線を手入力することで囲まれた領域内のデータ点を選択する対話的パネルを用いる．分析に用いるデータの時間区間を選択するためのスライダーも表示する．選択されたデータは，要素 V2 と V3，V4，V6，V7 に表示される．

2.3.2 特徴量表示パネル

選択されたデータの平均特徴量ベクトルと，その全データの平均からのずれを計算し表示する．選択されたデータの集合を $(t_{\tilde{n}}, \boldsymbol{x}_{\tilde{n}})_{\tilde{n}=1}^{\tilde{N}}$ と記す（以下，選択データと呼ぶ）．ます，全デー

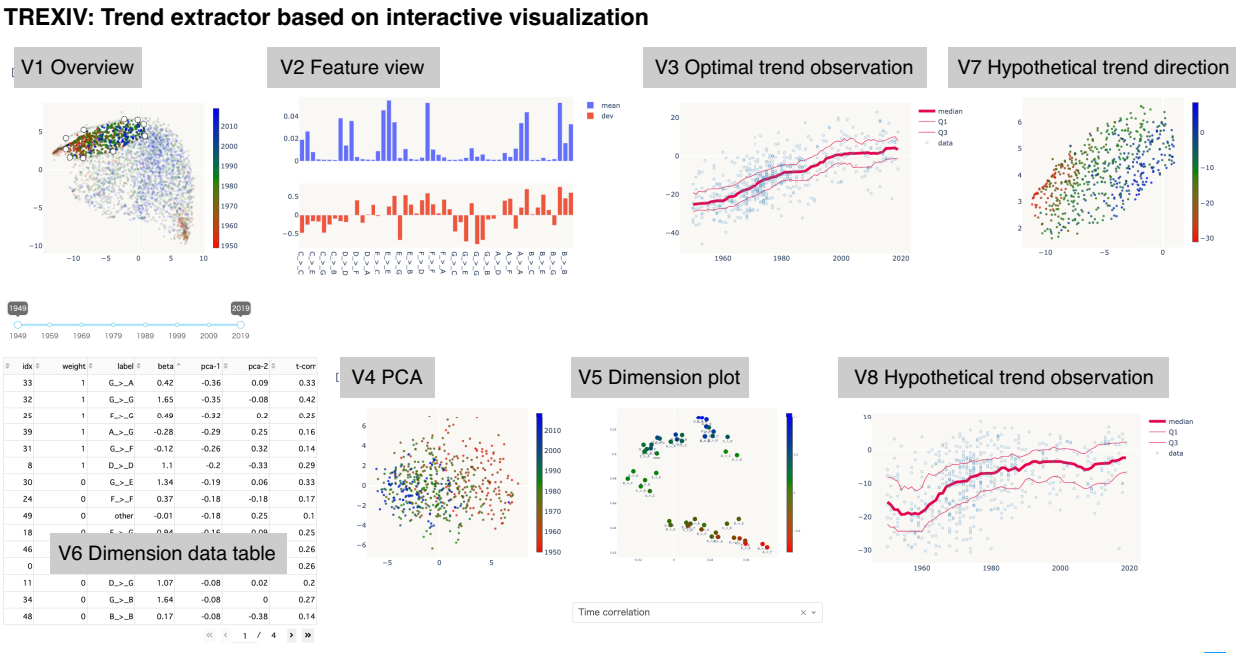


図 1　提案する可視化システムのスナップショット.

タの平均 $\bar{\boldsymbol{x}}^{\text{whole}}$ と標準偏差 $\boldsymbol{\sigma}^{\text{whole}}$，および選択データの平均 $\bar{\boldsymbol{x}}^{\text{selec}}$ を標準的な方法で計算する.

$$\bar{x}_i^{\text{whole}} = \frac{1}{N}\sum_{n=1}^{N} x_{ni},\quad \sigma_i^{\text{whole}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_{ni}-\bar{x}_i^{\text{whole}})^2}$$

$$\bar{x}_i^{\text{selec}} = \frac{1}{\tilde{N}}\sum_{\tilde{n}=1}^{\tilde{N}} x_{\tilde{n}i}$$

選択データの全データの平均からのずれ $\boldsymbol{\Delta}$ は，$\Delta_i = (\bar{x}_i^{\text{selec}} - \bar{x}_i^{\text{whole}})/\sigma_i^{\text{whole}}$ で定義する.

2.3.3　最適トレンド観察パネル

選択データに対する最適トレンドを表示するため，理論的に最大の CTD を持つ特徴量方向 $\boldsymbol{\beta} = (\beta_i)_{i=1}^{I}$ を計算する．この最適な方向 $\boldsymbol{\beta}$ は，計算効率が良い回帰手法によって求める．この方法を導出するため，データが次の生成過程に従うと考える.

$$x_{\tilde{n}i} = \alpha_i t_{\tilde{n}} + \epsilon_{\tilde{n}i} \tag{2}$$

ここで，α_i は定数であり，$\epsilon_{\tilde{n}i}$ は平均ベクトル $\langle\epsilon_{\tilde{n}i}\rangle = \mu_i$ と共分散行列 $\langle\epsilon_{\tilde{n}i}\epsilon_{\tilde{n}i'}\rangle - \mu_i\mu_{i'} = \Sigma_{ii'}$ を持つガウスノイズである（$\langle X\rangle$ は確率変数 X の期待値を表す）．時刻スタンプ $t_{\tilde{n}}$ は，基礎統計量を $\langle t_{\tilde{n}}\rangle = \bar{t}$ と $\langle t_{\tilde{n}}^2\rangle = \bar{t}^2 + s_t^2$ とする確率過程により生成されると仮定する．この時，第 i 次元の特徴量 $x_{\tilde{n}i}$ の CTD ρ_i は，$\rho_i = \sqrt{\alpha_i^2/(\alpha_i^2+\eta_i^2)}$ で与えられる（$\eta_i^2 \equiv \Sigma_{ii}/s_t^2$）.

特徴量 $z_{\tilde{n}}$ を $z_{\tilde{n}} = \sum_i \beta_i x_{\tilde{n}i}$ で定義すると，この特徴量の CTD ρ_z の二乗は次で与えられる.

$$\rho_z^2 = \frac{(\boldsymbol{\beta}\cdot\boldsymbol{\alpha})^2 s_t^2}{(\boldsymbol{\beta}\cdot\boldsymbol{\alpha})^2 s_t^2 + \boldsymbol{\beta}^{\mathrm{T}}\Sigma\boldsymbol{\beta}} \tag{3}$$

この表式から，$\boldsymbol{\beta}$ に対する ρ_z^2 の最大値は次のように求まる.

$$\rho_z^2 = \frac{1}{1+[(\boldsymbol{\alpha}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\alpha})s_t^2]^{-1}} \quad \text{with} \quad \boldsymbol{\beta}\propto\Sigma^{-1}\boldsymbol{\alpha} \tag{4}$$

上で求まった最適な方向 $\boldsymbol{\beta}\propto\Sigma^{-1}\boldsymbol{\alpha}$ を線形回帰手法によって得ることも可能である．この方法では，近似的な関係 $t_{\tilde{n}} \simeq \sum_i \beta_i x_{\tilde{n}i} + \gamma$ が成り立つように回帰変数の $\boldsymbol{\beta}$ と γ を求める．漸近極限における平均二乗誤差は次で与えられる.

$$\begin{aligned} L &= \lim_{\tilde{N}\to\infty}\frac{1}{\tilde{N}}\sum_{\tilde{n}}(z_{\tilde{n}}+\gamma-t_{\tilde{n}})^2 \\ &= (\boldsymbol{\beta}\cdot\boldsymbol{\alpha}-1)^2(\bar{t}^2+s_t^2) + \boldsymbol{\beta}^{\mathrm{T}}\Sigma\boldsymbol{\beta} + \gamma^2 + 2\gamma\bar{t}(\boldsymbol{\beta}\cdot\boldsymbol{\alpha}-1) \end{aligned} \tag{5}$$

この量の最小化の解は $\boldsymbol{\beta}\propto\Sigma^{-1}\boldsymbol{\alpha}$ となる．従って，式 (2) の仮定の下で，線形回帰と CTD の最大化は等価な結果をもたらす.

生成過程が未知の実データの場合，最適な方向 $\boldsymbol{\beta}$ は線形回帰手法によって効率的に推定できる．平均二乗誤差 $L = (1/\tilde{N})\sum_{\tilde{n}}(\boldsymbol{\beta}\cdot\boldsymbol{x}_{\tilde{n}} + \gamma - t_{\tilde{n}})^2$ の最小化から次の解が得られる.

$$\boldsymbol{\beta} = M^{-1}\boldsymbol{V},\quad M_{ij} = \frac{1}{\tilde{N}}\sum_{\tilde{n}}(x_{\tilde{n}i}-\bar{x}_i)(x_{\tilde{n}j}-\bar{x}_j) \tag{6}$$

$$V_i = \frac{1}{\tilde{N}}\sum_{\tilde{n}}(x_{\tilde{n}i}-\bar{x}_i)(t_{\tilde{n}}-\bar{t}) \tag{7}$$

ここで，$\bar{x}_i = (1/\tilde{N})\sum_{\tilde{n}} x_{\tilde{n}i}$ である．上述の等価性によって，データに含まれる非線形性が大くなければ，この方法で推定された方向は近似的に最適であると期待される．そこで，本研究の可視化システムではこの方法を実装する.

最適な方向 $\boldsymbol{\beta}$ が得られた後，選択データを，縦軸を対応する特徴量 $z_{\tilde{n}} = \sum_i \beta_i x_{\tilde{n}i}$ で横軸を時刻とする散布図によって表示する．また，相関係数と p 値の計算して，パネル内に表示する．傾きの時間変化を含む場合も含め，トレンドの発見を視覚的に容易にするため，各時間スライスにおけるデータサンプルから計算した中央値と第一分位点，第三分位点も表示する.

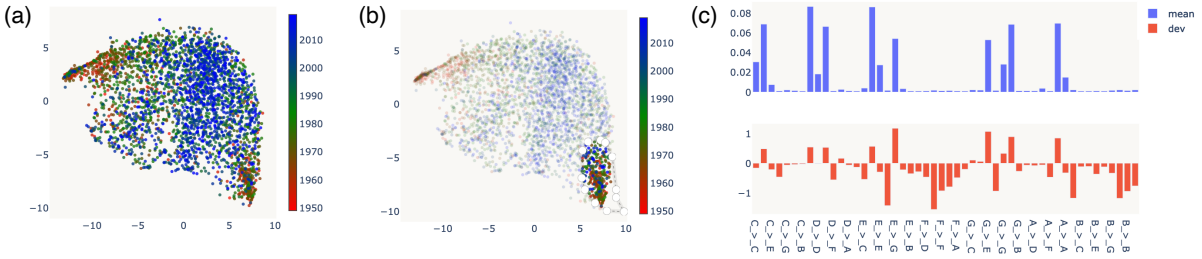


図 2　(a) 音高クラス統計量データに対する概観パネル．右下領域の選択データに対する (b) 概観パネルと (c) 特徴量表示パネル．

2.3.4　PCA パネル

このパネルには，選択データの特徴量から計算した第一および第二主成分を表示する．概観パネルと同様に，時刻スタンプはマーカ色として表示する．各時間スライス内での分散に比べて時間変化が大きい場合には，トレンドの次元は主成分に含まれるため，PCA は興味あるトレンドを見つけるために役立つことがある．また，第一・第二主成分を見ることにより，2 つの独立なトレンドやトレンドに対して各時間スライスで直交する主成分方向を見つけられることもある．

PCA パネルは，概観パネル内で局所的な曲線や直線に対応する特徴量の方向を確認するためにも用いることができ，これはトレンド方向の仮説を立てるための有効な手がかりとなり得る．一般的に，t-SNE 法や多次元尺度構成法などの非線形写像を用いた低次元表現空間において，局所的な曲線に対応する特徴量の方向を理解することは難しい．このような場合，ユーザはこの曲線に沿ってデータ点を選択するば，第一主成分として対応する特徴量方向が確認できることが多い．

2.3.5　次元散布図パネル

このパネルの機能は，ユーザが特徴量の次元間の関係性および，各特徴量次元と時間方向との関係性を理解するのを助けることである．例えば，解釈可能な仮説的トレンド方向を見つけるためには，時間方向と強く相関する特徴量次元のグループで，各次元の平均進化曲線同士も相関しているものを観察して，トレンドの「意味」を理解することが一般的には有用である．また，その様な特徴量次元のグループと平均進化曲線が負の相関を持つ特徴量次元のグループや，時間方向に対して大きな負の相関を持つ特徴量次元のグループを観察することも有益なこともしばしばある．

この様な観察を可能とするため，平均進化曲線が互いに強く相関している特徴量次元同士を近づけて，逆に，負の相関を持つ特徴量次元同士を遠ざけるように特徴量次元をプロットする．より正確には，第 i 次元特徴量の平均進化曲線を $\bar{x}_{it}$ と表す時，第 i 次元と第 j 次元の相関 ξ_{ij} は，ペアデータ $(\bar{x}_{it}, \bar{x}_{jt})$ の相関係数として計算する．そして，これらの次元の間の距離を $1-\xi_{ij}$ と定義して，t-SNE 法などの次元圧縮法を適用することで 2 次元表示を得る．この散布図において，第 i 次元特徴量の CTD ρ_i をマーカ色で表示し，赤（$\rho_i = -1$）と緑（$\rho_i = 0$），青（$\rho_i = 1$）を内挿して表示色を決定する．

2.3.6　次元データ表

この表は仮説の立案と仮説の検証のフェーズを結びつける道具としての機能を持つ．各行は特徴量次元を表し，各列には（分析データごとに与えられた）次元ラベル，最適な方向の重み β_i，および第一・第二主成分の因子負荷量 $\mathrm{PCA1}_i$ と $\mathrm{PCA2}_i$ が表示される．これらの列の値はトレンドに関係する特徴量次元を見つけ出すのに用いられる．

別の列には，仮説的トレンド方向の重み $\tilde{\beta}_i$ が表示される．ユーザはこれらの重みを編集することで，トレンドの仮説を立案できる．トレンド仮説の検証フェーズでは，これらの重みの値が読み込まれ，仮説的トレンド方向表示パネルと仮説的トレンド観察パネルの表示に用いられる．

2.3.7　仮説的トレンド方向表示パネル

このパネルでは，選択データが概観パネルと同じ座標に散布図として表示される．仮説的トレンド特徴量 $y_{\tilde{n}} = \sum_i \tilde{\beta}_i x_{\tilde{n}i}$ の値がマーカ色として表示され，表示色には概観パネルと同様のグラデーションが用いられる．

2.3.8　仮説的トレンド観察パネル

このパネルでは，最適トレンド観察パネルと同じ方式で，仮説的トレンド特徴量 $y_{\tilde{n}}$ と時刻スタンプ $t_{\tilde{n}}$ が表示される．特徴量 $y_{\tilde{n}}$ の CTD と p 値，そして，中央値と第一分位点，第三分位点の曲線も表示される．

3　音楽スタイル進化解析への応用

本章では提案する可視化システムを用いた解析例として，音楽スタイルの進化解析に応用した結果について述べる．使用するデータは，20 世紀後半以降にリリースされた日本のポピュラー音楽の約 3000 曲のボーカルメロディーパートを MusicXML 形式で整備したものである [9]．メロディーは音符の記号列で表され，各音符は相対音高クラス（楽曲を自然調に移調した時の音高からオクターブの情報を除いたもの）と拍節位置（発音時刻を表す小節内の相対位置で，1 小節を 48 分割した単位で表す）のペアで表される．

3.1　解析例 1：メジャー系音階の進化

3.1.1　特徴量の抽出

メロディーの音高に関する作曲スタイルを表す特徴量として，音高クラスのバイグラム確率を用いる．このバイグラム確率は，各作品のメロディーデータ内の連続する音高クラスの遷移の同

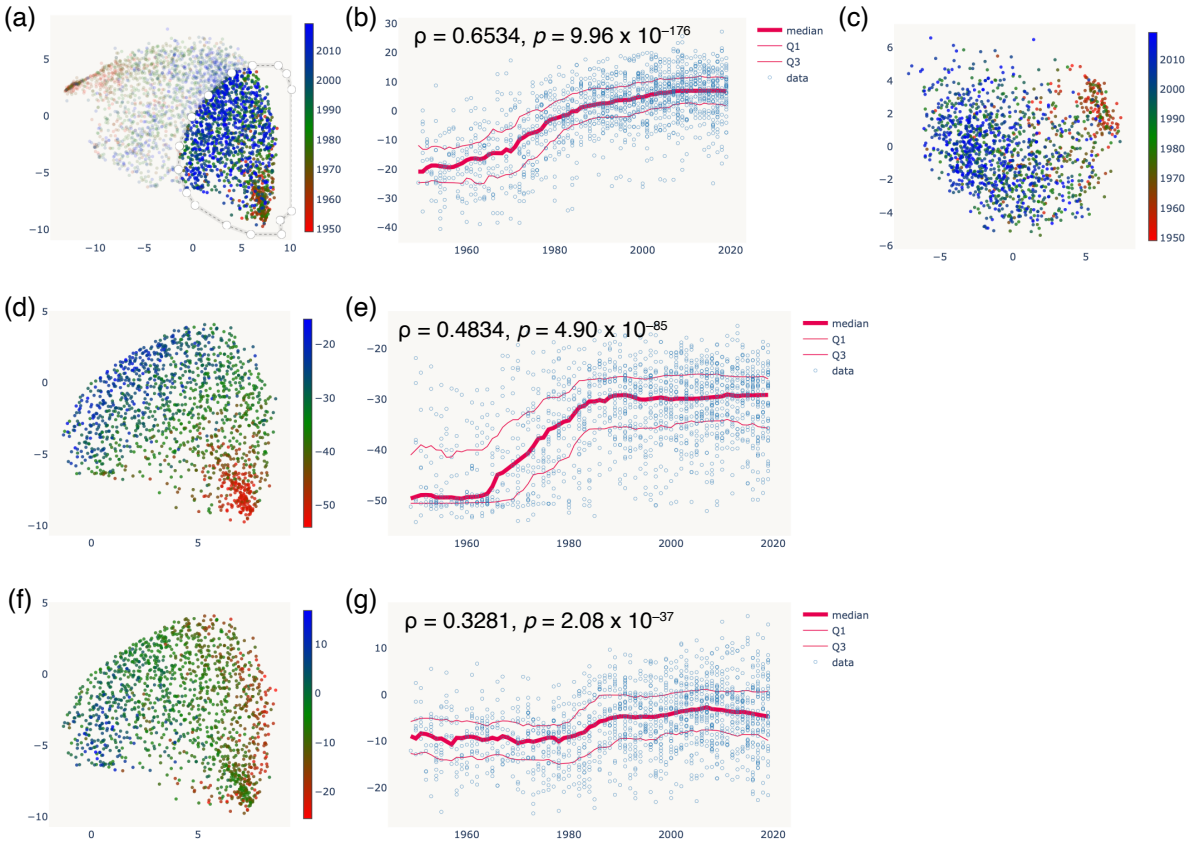


図 3　メジャー系音階の進化解析におけるトレンド抽出で用いた可視化画面.

時確率であり，音高クラスは $\{\mathrm{C}, \mathrm{C}\#, \ldots, \mathrm{B}\}$ の 12 種類存在するため，12^2 次元の統計量である．解析では，音階音に対応する 7 つの音高クラス $\{\mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}, \mathrm{G}, \mathrm{A}, \mathrm{B}\}$ の間の遷移とそれ以外の other 遷移に分けて考えて，$7^2 + 1 = 50$ 次元の特徴量として扱う．

3.1.2　データ選択フェーズ

このデータでは 2 つの大きなトレンドの存在が概観パネルで確認できた（図 2）．それぞれ，右下領域から中央領域に向かうものと，左上領域から中央領域に向かうものである．右下領域が選択された時の特徴量表示パネルを調べると，この領域はメジャー五音音階として知られる，F と B の使用を制限した音階を表していることが確認できた．同様に，左上領域は，日本の伝統音楽で陰音階として知られる，マイナー五音音階を表していること，そして，中央領域は西洋音楽で標準的に使われる全音階を表していることが確認できた．従って，確認された 2 つのトレンドはメジャーおよびマイナーの五音音階が主流だった時代から全音階が主流になった時代への動きとして解釈できる．以下では，メジャー音階の進化に注目して，右領域のデータを選択した結果について述べる（図 3(a)）．

3.1.3　仮説立案フェーズ

最適トレンド観察パネルでは，特に 1960 年から 1990 年にかけて，最適トレンドの方向に大きな増加が見られた（図 3(b)）．この場合は，前節における観察から，重要となる特徴量を予想することができる．つまり，このトレンドは，F と B を含むバイグラムの確率に関係していると期待される．確認すると，最適トレンド方向 $\boldsymbol{\beta}$ の重み係数が大きな特徴量方向は，これらの特徴量を確かに含んでいたものの，それら以外に，直ぐには解釈可能なトレンドと結び付かない特徴量次元も含まれていた．

PCA パネルでは，第一主成分が明確なトレンドを示した（図 3(c)）．第一主成分の係数が最も小さい次元には，F→F，F⇄E，B→B，B⇄A などの五音音階から全音音階への変遷と直結する特徴量が含まれていた．一方で，係数が最も大きな次元には，E⇄G など，この変遷とは一見，関係がない特徴量が含まれていた．実は，これらの音高クラス遷移は，五音音階では順次進行（隣り合う音高への進行）であり，頻度が高い一方で，全音階ではもはや順次進行ではないため，頻度が小さくなるものである．これらの考察に基づき，メジャー系音階の進化では次の様なトレンド仮説 (HP1) を立案した．

HP1　F と B を含む同音進行と順次進行（F→F と F⇄E，F⇄G，B→B，B⇄A，B⇄C）に対しては $\tilde{\beta}_i = 1$，F と B をまたぐ遷移（E⇄G と A⇄C）に対しては $\tilde{\beta}_i = -1$，それ以外の音高クラス遷移に対しては $\tilde{\beta}_i = 0$.

PCA パネルでは，第二主成分においても中程度のトレンドが見られた（図 3(c)）．この因子負荷ベクトルを調べると，係数が最も小さい次元には A→A，A⇄B，A⇄E などのマイナー音階で多く現れる音高クラス遷移が見つかり，最も大きな次元には C→C，C⇄D，C⇄E などのメジャー音階でで多く現れる音高クラス遷移が見つかった．この観察によって，マイナーからメジャーへの変遷に関する，もう一つトレンド仮説 (HP2) を立案することができた．

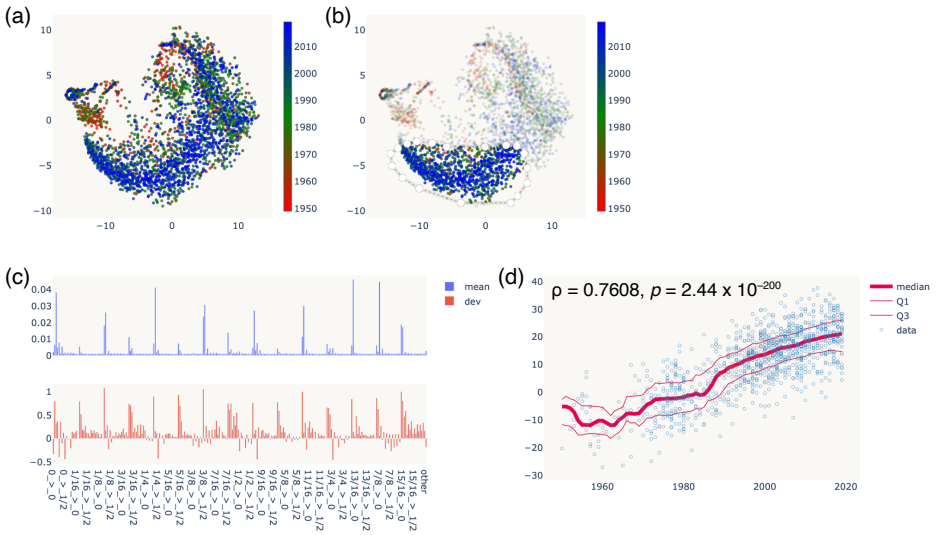


図 4　(a) 拍節位置統計量データに対する概観パネル．16 分音符リズムの領域の選択データに対する (b) 概観パネルと (c) 特徴量表示パネル，(d) 最適トレンド観察パネル．

HP2　C→C と C⇄D，C⇄E，C⇄G に対しては $\tilde{\beta}_i = 1$，A→A と A⇄G，A⇄E に対しては $\tilde{\beta}_i = -1$（このトレンド方向を HP1 とは独立なものとするため A⇄B は除外した），それ以外の音高クラス遷移に対しては $\tilde{\beta}_i = 0$.

3.1.4　仮説検証フェーズ

トレンド仮説 HP1 と HP2 に対する，仮説的トレンド方向表示パネルと仮説的トレンド観察パネルを図 3(d)–(g) に示す．相関係数と p 値から，これらの仮説はどちらも統計的に有意であり，特に HP1 の有意度が高いことが分かった．仮説的トレンド方向表示パネルでは，HP1 は 2 次元表現空間では右下から左上への方向に対応しており，HP2 は右から左への方向に対応していることが確認できた．従って，このデータでは 2 つの独立なトレンドの存在が明らかになった．HP1 と HP2 に対する仮説的トレンド観察パネルを比べると，これらのトレンドが異なる特性を持っていることが確認できる．HP1 の特徴量はおよそ 1965 年から 1985 年の間の時期に大きく増加しており，HP2 の特徴量はおよそ 1980 年から 1990 年の間の時期に増加している．前者のトレンドでは，日本のポピュラー音楽におけるメジャー系音階の進化では，一方の音階から他方の音階へと切り替わるという変化ではなく，HP1 の特徴の分布の漸次的な変化が起きたことが見てとれる．これは，音楽文化の進化過程の興味深い性質を示している．

3.2　解析例 2：16 分音符リズムの進化

3.2.1　特徴量の抽出

メロディーのリズムに関する作曲スタイルを表す特徴量として，拍節位置のバイグラム確率を用いる．このバイグラム確率は，各作品のメロディーデータ内の連続する拍節位置の遷移の同時確率であり，拍節位置は 0 から 47 までの 48 種類存在するため，48^2 次元の統計量である．解析では，16 分音符単位の拍節位置の間の遷移とそれ以外の other 遷移に分けて考えて，$16^2 + 1 = 257$ 次元の特徴量として扱う．以下，拍節位置を小節の長さを 1 とした単位で表記する．例えば，1/4 は 4/4 拍子での 2 拍目，15/16 は小節内の最後の 16 分音符の位置を表す．

3.2.2　データ選択フェーズ

このデータの分布には，3 つの大きなクラスターが存在することが概観パネルで確認できた（図 4(a)）．左領域に分離したクラスターと右領域にある様々な時代のサンプルからなる密集したクラスター，そして，下領域にある主に 1980 年代以降のサンプルから成るクラスターである．特徴量表示パネルを調べると，これらのクラスターは各々，other 遷移が多いリズムパターン（三連符や三拍子のリズムが含まれると考えられる）と 8 分音符リズム，16 分音符リズムに対応することが分かった（図 4(b)(c)）．また，概観パネルから，8 分音符リズムおよび other 遷移を多く含むリズムから 16 分音符リズムへの変遷（つまり，短いリズムの 4/4 拍子に向かう変化）に関わる全体的なトレンドの存在を予想することができる．以下では，16 分音符リズム領域に含まれるより非自明なトレンドに注目した分析結果について述べる．

16 分音符リズム領域のデータサンプルを選択すると，最適トレンド観察パネルは，サンプル数が多い期間である，1990 年代から 2010 年代にかけての持続的なトレンドを示した（図 4(d)）．また，この領域では 1970 年以前には少数のサンプルしかないことも確認できた．そこで，時間スライダーを用いて 1990 年以降の期間のデータのみをさらに選択して，以下の分析を行った．

3.2.3　仮説立案フェーズ

データ選択の後，最適トレンド観察パネルでは相関係数が 0.73 の明確なトレンドが示された（5(b)）．しかし，重み $\boldsymbol{\beta}$ が最も大きいあるいは最も小さいバイグラムの要素は，頻度が非常に小さいものであり，この結果から解釈可能なトレンド方向を見つけることは困難であった．さらに，PCA パネルにおい

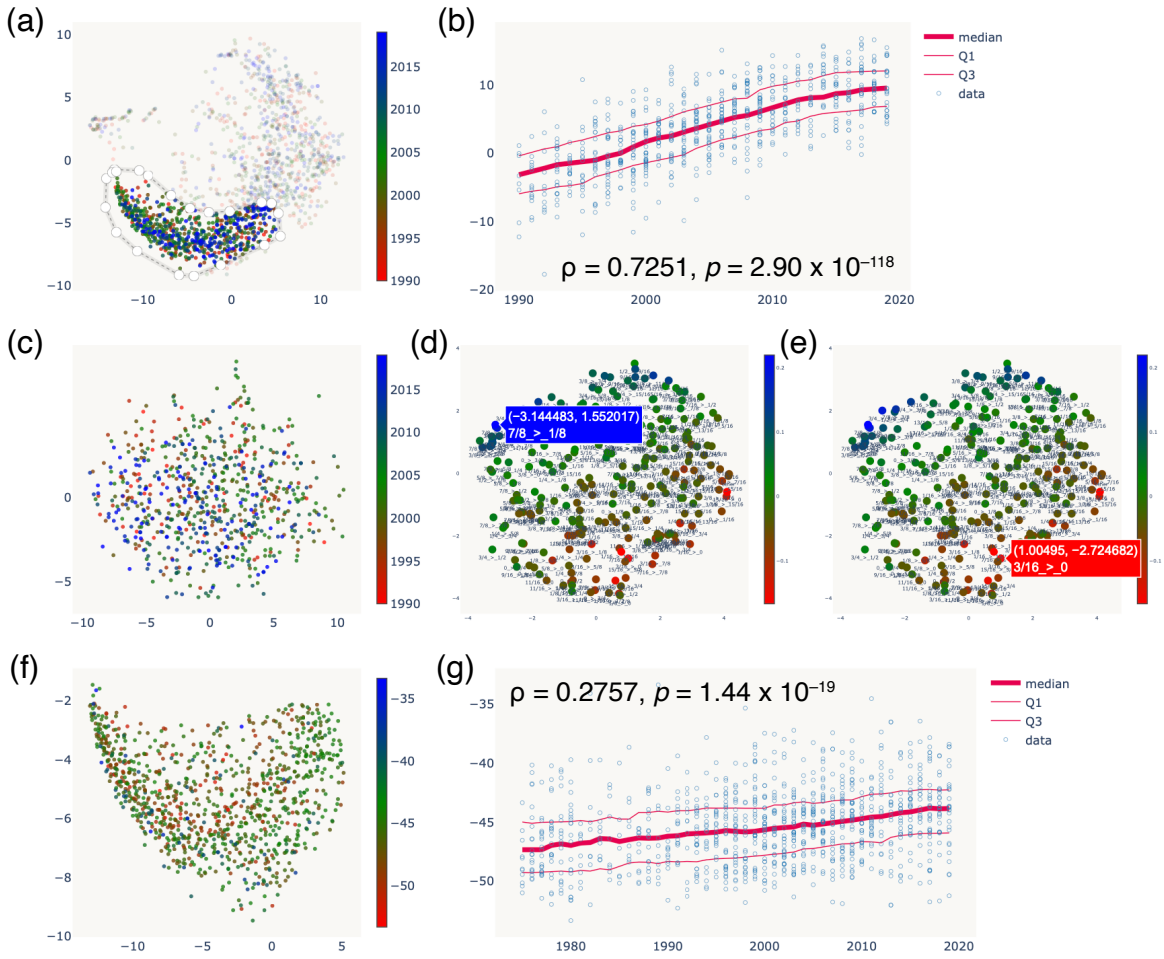


図 5　16 分音符リズムの進化解析におけるトレンド抽出で用いた可視化画面.

ては明確なトレンドは見つからなかった（図 5(c)）.

そこで，解釈可能なトレンド仮説を探すために，次元散布図パネルを用いた（図 5(d)(e)）. CTD の値が最大の 2 つのバイグラムは 7/8 → 1/8 と 3/8 → 5/8 であり，CTD が負の方向に最も小さかったバイグラムの多くは，0（強拍）を含むものであった．この結果から，トレンドにはシンコペーションが関係しているという予想を立てた．シンコペーションとは,「前のめり」の感覚を表すリズムを意味し，強拍あるいは 1/2 などの準強拍をまたぐ拍節位置の遷移として表される．この予想に基づき，解釈が容易なトレンド仮説を立てるため，強拍をまたぐ拍節位置遷移で短い音符を伴うもののみを考えることにした．これにより，次の仮説的トレンド方向 (HM1) を立案した.

HM1　$\{13/16, 7/8, 15/16\} \to \{1/16, 1/8, 3/16\}$（9 種類）に対しては $\tilde{\beta}_i = 1$，$\{13/16, 7/8, 15/16\} \to 0$（3 種類）に対しては $\tilde{\beta}_i = -1$，その他の遷移に対しては $\tilde{\beta}_i = 0$.

3. 2. 4　仮説検証フェーズ

図 5(f)(g) は，トレンド仮説 HM1 を検証した結果を示している．仮説的トレンド観察パネルでは，中程度だが有意な相関が確認され，シンコペーションの頻度の増加が選択データに含まれるトレンドの要因の一つであることが示された．仮説的トレンド方向表示パネルでは，HM1 特徴量は 2 次元表現空間の上下方向に緩く対応することが示された．この解析例では，興味深くて非自明なトレンドの発見に，対面的なデータ表示を含む本研究の可視化手法が有効であることが示された．発見されたトレンドは 2000 年代以降にも継続しており，音高クラス特徴量が顕著なトレンドを示した期間とは異なることから，音楽学的にも興味深い現象と言える.

4　結　　論

本稿では，時刻スタンプ付き高次元データから適切なデータ選択を行い，解釈可能なトレンドの特徴量方向の抽出を行うための対話的可視化手法について述べた．ユーザがデータの性質を理解しながら，トレンドの仮説の立案と検証ができる様に，種々の要素技術を組み合わせた，多面的なデータ可視化を含むシステムを実装し，音楽スタイル進化解析の具体例を通して，その有効性を示した．解析例を通して，理論的に最適なトレンドの推定は有用ではあるが，必ずしも解釈可能なトレンドの発見には直結しないこと，PCA を用いた表示もトレンドの発見に有用な場合とそうでない場合があること，また，PCA 表示が有用でない場合には特徴量の次元プロットが利用可能な場合があることなどが示された．これらは，本研究の手法の特徴である多面的なデータ表示の有効性を示すものであり，どのパネ

ルが有用であるかはデータに応じて変化するため，対話的操作の有効性も示すものである．

今後は，本研究の可視化ツールを汎用性の高い形でライブラリー化あるいはアプリケーション化した上で公開し，幅広い分野の研究に利用できるようにすることが課題である．また，創作物データに付与されることが多い，作者やジャンル，地理情報などのメタ情報を可視化要素に含めることで，データの理解と解釈可能なトレンドの発見をより容易にする技術的改善も取り組むべき課題である．さらに，絵画や文学を含む多様な文化領域のデータ解析や形態進化解析などの検証研究を行い，可視化手法のさらなる機能向上に向けた知見を深める研究も行う予定である．

謝　　辞

本研究は，日本学術振興会科研費 21K12187, 21K02846, 22H03661 および JST 創発的研究支援事業 JPMJPR226X からの支援を受けて行われた．

文　　献

[1] Mauch, M., MacCallum, R. M., Levy, M. and Leroi, A. M.: The evolution of popular music: USA 1960–2010, *Royal Society Open Science*, Vol. 2, No. 150081, pp. 1–10 (2015).
[2] Youngblood, M.: Conformity bias in the cultural transmission of music sampling traditions, *Royal Society Open Science*, Vol. 6, No. 191149, pp. 1–8 (2019).
[3] Sigaki, H. Y. D., Perc, M. and Ribeiro, H. V.: History of art paintings through the lens of entropy and complexity, *Proceedings of the National Academy of Sciences of the USA*, Vol. 115, No. 37, pp. E8585–E8594 (2018).
[4] Nakamura, E. and Saito, Y.: Evolutionary analysis and cultural transmission models of color style distributions in painting arts, *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 493–500 (2023).
[5] Brand, C. O. and Mesoudi, A. A. A.: Cultural evolution of emotional expression in 50 years of song lyrics, *Evolutionary Human Sciences*, Vol. 1, No. e11, pp. 1–14 (2019).
[6] Klingenberg, C. P.: Evolution and development of shape: integrating quantitative approaches, *Nature Reviews Genetics*, Vol. 11, No. 9, pp. 623–635 (2010).
[7] Interiano, M. et al.: Musical trends and predictability of success in contemporary songs in and out of the top charts, *Royal Society Open Science*, Vol. 5, No. 171274, pp. 1–16 (2018).
[8] Nakamura, E. and Kaneko, K.: Statistical evolutionary laws in music styles, *Scientific Reports*, Vol. 9, No. 15993, pp. 1–11 (2019).
[9] Singh, R. and Nakamura, E.: Dynamic cluster structure and predictive modelling of music creation style distributions, *Royal Society Open Science*, Vol. 9, No. 220516, pp. 1–18 (2022).
[10] Chakrabarti, D., Kumar, R. and Tomkins, A.: Evolutionary clustering, *Proc. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 554–560 (2006).
[11] Blei, D. M. and Lafferty, J. D.: Dynamic topic models, *Proc. International Conference on Machine Learning (ICML)*, pp. 113–120 (2006).
[12] Liu, S., Wang, B., Bremer, P.-T. and Pascucci, V.: Distortion-Guided Structure-Driven Interactive Exploration of High-Dimensional Data, *Computer Graphics Forum*, Vol. 33, No. 3, pp. 101–110 (2014).
[13] Turkay, C., Filzmoser, P. and Hauser, H.: Brushing dimensions-a dual visual analysis model for high-dimensional data, *IEEE transactions on visualization and computer graphics*, Vol. 17, No. 12, pp. 2591–2599 (2011).
[14] Sonali, P. and Kumar, D. N.: Review of trend detection methods and their application to detect temperature changes in India, *Journal of Hydrology*, Vol. 476, pp. 212–227 (2013).
[15] Wang, C., Yu, H. and Ma, K.-L.: Importance-driven time-varying data visualization, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1547–1554 (2008).
[16] Kumatani, S., Itoh, T., Motohashi, Y., Umezu, K. and Takatsuka, M.: Time-varying data visualization using clustered heatmap and dual scatterplots, *Proc. International Conference Information Visualisation (IV)*, pp. 63–68 (2016).
[17] Van der Maaten, L. and Hinton, G.: Visualizing data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, No. 11, pp. 2579–2605 (2008).

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

2024年3月1日(金) 10:00 ~ 12:10 | T5-A オンライン（Zoom Events)

ウェルビーイング

座長:児玉 直樹(新潟医療福祉大学)
コメンテータ:太田 学(岡山大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T5-A-7-01] 心理的支援を目的としたChatBotによるソーシャル・サポートシステムの提案

*池田 青空[1]、談議所 考汰[3]、Siriaraya Panote[2]、中島 伸介[1] (1. 京都産業大学、2. 京都工芸繊維大学、3. 京都産業大学大学院)

10:20 ~ 10:45
[T5-A-7-02] メンタルウェルビーイング向上のためのピアベースサポートシステム(Peer2S)の開発

*談議所 孝汰[1]、Siriaraya Panote[2]、She Wan jou[4]、中島 伸介[3] (1. 京都産業大学大学院、2. 京都工芸繊維大学、3. 京都産業大学、4. 奈良先端科学技術大学院大学)

10:45 ~ 11:10
[T5-A-7-03] 社員が記述したテキストからの幸福度推定

*林 純子[1]、伊藤 和浩[1]、眞鍋 雅恵[1]、渡邉 寧[2]、中山 真孝[2]、内田 由紀子[2]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学、2. 京都大学人と社会の未来研究院)

11:10 ~ 11:35
[T5-A-7-04] COVID-19 流行期における大学生の X 上の投稿と主観的幸福感の関係：2 年間の調査と言語特徴分析による検討

*蔡 丹萌[1]、若林 啓[1]、叶 少瑜[1] (1. 筑波大学)

11:35 ~ 12:00
[T5-A-7-05] 孤独エピソードデータセットの構築とその応用

*藤川 直也[1]、NGUYEN Quang Toan[2]、伊藤 和浩[1]、若宮 翔子[1]、荒牧 英治[1] (1. 奈良先端科学技術大学院大学 ソーシャル・コンピューティング研究室、2. ESIEE Paris)

心理的支援を目的とした ChatBot によるソーシャル・サポートシステムの提案

池田　青空† 　談議所　考汰†† 　Panote　Siriaraya††† 　中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学大学院 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
††† 京都工芸繊維大学 情報工学・人間科学系　〒 606–8585 京都府京都市左京区松ケ崎橋上町
E-mail: †g2153074@cc．kyoto-su．ac．jp，††i2286103@cc．kyoto-su．ac．jp，†††spanote@kit．ac．jp，
†nakajima@cc．kyoto-su．ac．jp

あらまし　近年，うつ病などの精神疾患を有する患者数が増加している．従来のソーシャル・サポートでは当事者が周囲に相談できる人がいないと支援を受けることが難しいという社会的な環境に依存しているという欠点がある．そこで本研究では，ChatBot との会話を通じたソーシャル・サポートシステムを提案する．ChatBot を使用する利点は一般的なソーシャル・サポートの欠点である社会的環境や時間にとらわれない点である．本研究ではこの利点に加え，様々な性格の ChatBot を作成し，ユーザが会話したい相手を選択できるようにする．本研究での ChatBot は会話データから様々な性格を持つようにトレーニングする．そして，作成されたそれぞれの ChtaBot に対してユーザの悩み情報を元にユーザと同じ悩みを付与することで，ユーザと同じ悩みを抱える人格を有する様々な性格の ChatBot が生成される．そして，ChatBot との会話を通じてユーザの精神的健康の維持を目指す．

キーワード　ヘルスケア，LLM，対話システム

1. はじめに

近年，新型コロナウイルス感染症（COVID-19）の世界的な拡大や SNS の普及により，うつ病等の精神疾患を有する患者数が増加傾向にある [1]．このように不安を抱えた人への支援としてソーシャル・サポートがある．ソーシャル・サポート [2] とはコミュニティ心理学の Caplan が提唱した概念であり，周囲の人々から有形無形の支援を得ることをいう．ソーシャル・サポートを受けることで健康によい行動を続けやすくなることや，直面している困難や，苦悩をやわらげるなどの効果が期待できる．しかし，ソーシャルサポートの欠点として浦らのソーシャルサポートの研究 [3] において，ソーシャル・サポートの有効性の確認と同時に従来のソーシャル・サポートは，当事者が周囲に相談できる人がいないと支援を受けることが難しいという社会的な影響を大きく受ける可能性が述べられている．

そこで，本研究の最終目標は，ChatBot を用いることで従来のソーシャル・サポートの問題点を克服し，より柔軟で効果的なソーシャル・サポートシステムを開発し，ユーザの精神的健康の維持を目指すことである．

本稿では，新たなアプローチとして，様々な性格を持つ ChatBot との会話を通じたソーシャル・サポートシステムを提案する．この提案手法では，ユーザと同じ悩みを持ち，様々な性格を持つ ChatBot との会話を通じて精神的な支援を行うソーシャル・サポートを目指している．具体的には，ユーザが自身の悩みを登録し，登録された悩みを元にユーザと同じ悩みを抱える人格を有する様々な性格の ChatBot が用意され，ユーザがその ChatBot の中から会話したい相手を選択することで会話が開始される．これにより，社会的環境や時間にとらわれず，様々な性格の ChatBot との会話を提供できるため，従来のソーシャル・サポートの問題であった社会的環境に依存している点や時間に対する制約を克服することができ，ユーザがより柔軟で効果的なソーシャル・サポートを受けられることが期待される．

2. 関連研究

福岡が行ったソーシャル・サポートについての研究 [4] に関する，友人への自己開示とサポートの関連性が詳細に述べられている．研究結果からは，ストレス体験が増加すると自己開示が増える傾向があり，この自己開示の増加が友人からのサポート量と有意な関連があることが示されている．さらに，ストレス状況での自己開示と友人サポートは直結し，サポートを多く受けることが気分状態の改善に有益であることが研究結果から示されている．

嶋らの研究 [5] では，大学生を対象とした日常ストレスにおけるソーシャル・サポートの効果に焦点が当てられている．この研究からは，男女ともに同性の友人からのサポートが最も多くの大学生に対して効果があることが明らかになってる．ただし，ソーシャル・サポートが有するストレス緩和効果は性別によって異なり，これに応じたアプローチの必要性が示唆されている．

談議所らの研究 [6] では「Peer2S」と呼ばれるマッチングシステムが設計，開発され，似たような困難な経験を共有する

ユーザー同士を結びつけ，相互のソーシャル・サポートを促進することを目的としている．11 人のユーザーを対象に 2 週間にわたる初期のユーザースタディが行われ，マッチングアルゴリズムと「Peer2S」の有効性について評価されました．その結果，類似の悩みを抱えるユーザー同士をマッチングすることで，不安やうつ病の改善が期待される可能性が示唆されている．

大津らの研究 [7] では携帯端末の心的支援アプリケーションに解決志向アプローチの対話型課題を導入し，その有用性を検証している．具体的には，ミラクルクエスチョン課題を実践できるチャットボットアプリを開発し，大学生を対象に既存の行動調整支援アプリの課題との比較や併用を伴う調査を行った．結果として，解決志向アプローチに基づく対話型課題は，対話の柔軟性が低いことが主な要因となり，精神的健康度や行動活性の指標の有意な改善は見られなかった．しかし，ミラクルクエスチョン課題は，行動推薦課題と比較して精神的健康度に関する尺度の平均スコアを相対的に改善した．また，ユーザーの約 7 割がアプリの使用を通じて今後取り組みたい目標を策定し，ユーザー個々人に適した問題解決のための目標策定を促進できる可能性が示唆された．

Brandtzaeg らの研究 [8] では，若者がチャットボットから様々なタイプのソーシャル・サポートをどのように受けるかに焦点を当てている．この研究では，16〜21 歳の 16 人の参加者を対象に，メンタルヘルスのチャットボット（Woebot）を用いた実際の体験を通じて，若者がチャットボットから様々な形のソーシャル・サポートをどのように受けるかを明らかにした．研究結果によれば，チャットボットは利用のハードルが低いという特性や安全で匿名の空間を提供することから，主に評価，情報，感情，および具体的なサポートの源として捉えられる可能性が示唆された．ソーシャルチャットボットは，使用者が自分の問題に対して建設的な考え方を助けるだけでなく，信頼性のある情報サポートにアクセスするための効率的な手段としても認識された．研究は，日常生活で問題や悩みを抱える若者にとって，ソーシャル・サポートの提供が有益であると示唆されている．

以上のとおりソーシャル・サポートに関する研究や ChatBot を用いた心理的支援を行う研究はあるが，本研究のようにユーザと同じ悩みを抱える人格を有する様々な性格の ChatBot を用いたソーシャル・サポートの研究はなされていない．

3. ChatBot によるソーシャル・サポートシステム

3.1 ソーシャル・サポートとは

本節では，ソーシャル・サポートについて説明する．本研究で我々が開発するシステムにおいて参考にしている「ソーシャル・サポート」[2] とは，コミュニティ心理学の Caplan が提唱した概念であり，個人が周囲の人々から有形無形の支援を得る過程のことである．ソーシャル・サポートは，さまざまな側面を有しており，「知覚されたもの」と「受領されたもの」の二つの側面に分けられる．これらは，実際にサポートをされた経験を指す「受領されたサポート」と，サポートされたりサポートを期待している状態を指す「知覚されたサポート」に対応している．

具体的なサポートの側面としては，他者からの共感や応援を提供する「情緒的サポート」，物やサービスを提供する「道具的サポート」，問題解決のために必要なアドバイスや情報を提供する「情報的サポート」，そして当事者の行動が良いか悪いか，社会的に好ましいか好ましくないかなど，適切な評価を与える「評価的サポート」などが挙げられる．これらの側面は，サポートがどのような内容であるかに基づいて分類される．具体例として「情緒的サポート」では不安や悩みを抱えた人に対して励ましたり，愚痴を聴いたり，相談役になったりなどの支援を行うことで苦悩をやわらげるなどの効果が期待できる．先行研究である談議所らの研究 [6] においても，似たような困難な経験を共有するユーザー同士を結びつける「Peer2S」と呼ばれるマッチングシステムによって「情緒的サポート」や「情報的サポート」の支援を行っている．本研究でも同じように「情緒的サポート」や「情報的サポート」を ChatBot によって行う．

3.2 ChatBot によるソーシャル・サポートシステムの概要

本節では，提案手法の概要について説明する．ChatBot を用いた提案手法の概要図を図 1 に示す．まず，ユーザは自分自身の悩みを入力する．そして、Big Five モデルと実際の会話データを使用したトレーニングにより生成された様々な性格を持つ ChatBot に対して，ユーザの悩みの情報を基に同じ悩みを持つようにする．これにより，ユーザと同じ悩みを抱える人格を有する様々な性格の ChatBot が作成される．ユーザは会話したい相手を作成された ChatBot の中から自由に選択でき，選択した ChatBot との会話によって精神的健康を目指す．

3.3 性格が異なる ChatBot の生成方法

本節では，作成する ChatBot について説明する．ChatBot に異なる性格を与える手法として，Big Five モデルを活用する．具体的には，開放性，誠実性，外向性，協調性，神経症的傾向の 5 つの性格特性を考慮し，これらの特性を構築するために実際の会話データをトレーニングに組み込む．

Big Five モデル [9] とは，人間の性格を Openness（開放性），Conscientiousness（誠実性），Extraversion（外向性），Agreeableness（協調性），Neuroticism（神経症的傾向）の 5 つに分類するモデルである．各次元は，個人の特定の性格特性を評価し，これに基づいて総合的な性格プロファイルを構築する．

各性格特性に対応するトレーニングデータを用意することで Chatbot が特定の性格を表現する．具体的な例として，外向性を表現する際は，実際に人同士が悩みについて会話しているデータの中から外向性があると評価された人の会話データを学習に組み込む．

このようにして Big Five モデルと実際の会話データを用いて様々な性格を持つ ChatBot を作成してそれらの ChatBot に

表 1　ソーシャル・サポートの種類と効果

ソーシャル・サポート	解説	効果
情緒的	他者からの共感や応援の提供	情緒が安定し、やる気がでる
道具的	金銭や必要な物やサービスを貸し与える	問題解決を直接的に進める
情報的	問題解決のために必要な情報や知識を提供する	問題解決を間接的に進める
評価的	適切な評価を与える	自信が深まる

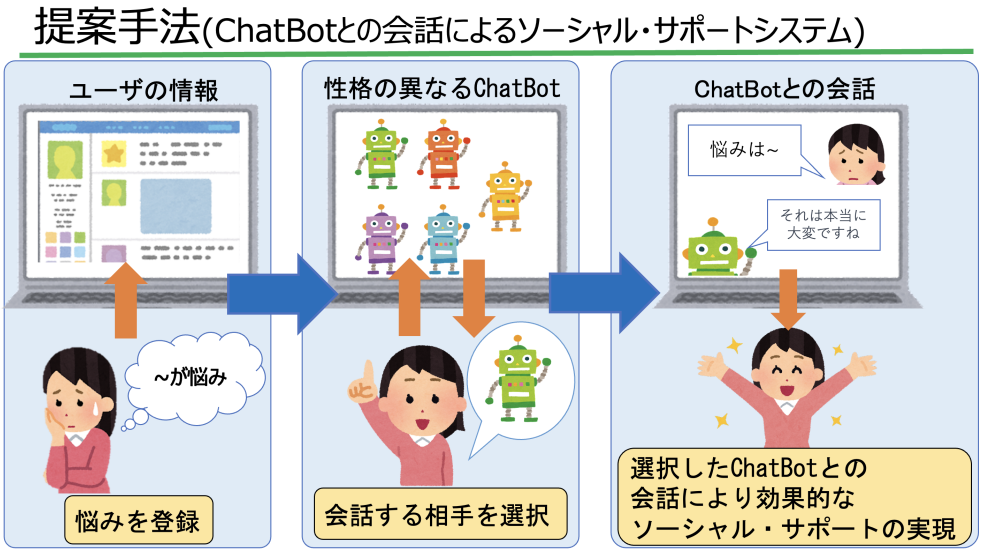


図 1　ChatBot によるソーシャル・サポートシステムの概要

ユーザが登録した悩みを元にユーザと同じ悩みを持つように学習させる．方法としてはユーザが登録した悩みの情報を学習データに追加することでユーザと同じ悩みを持ち，様々な性格を有する ChatBot が作成される仕組みである．

4.　予 備 実 験

本節では行った予備実験について説明する．今回は特定の性格を持つように設定した ChatGPT がどの程度その性格を表現できるのかを評価した．

4.1　実 験 手 順

ChatGPT の Web サービスにおいて以下のようなプロンプトを使用して ChatGPT に性格を設定した．また，設定する性格は Big Five モデル [9] における 5 つの性格特性のいずれかが高い性格になるように設定する．以下の例は誠実性が高い ChatBot として振る舞うように設定した際に用いたプロンプトである．

「Suppose you are a highly conscientiousness chatbot.」

次に，[10] や [11] を元に作成された Big Five モデルの各性格特性をスコア化するアンケートを性格特性を有するように設定した ChatGPT に答えさせ，それぞれの性格特性を持った ChatGPT のスコアを調査する．以下に示すような文章によってアンケートの 1～50 の質問に対して 1～5 のスケールでどの程度当てはまるか回答させる．

「On a scale of 1-5, please indicate how much you agree with each statement numbered 1-50 in the list below. 1 = Disagree, 2 = Somewhat Disagree, 3 = Neutral, 4 = Somewhat Agree, 5 = Agree.」

4.2　評 価 方 法

評価方法としては式 (1)～式 (5) 示す計算方法によって Big Five モデルの各性格特性のスコアを算出する．式中の ( ) 内の数字はアンケートの問題番号である．

そして，Big Five モデルの各性格特性を持つように設定した ChatGPT それぞれが設定されている性格特性のスコアが高いかどうかを評価する．

$$\begin{aligned} \text{外向性} = 20 &+ (1) - (6) + (11) - (16) + (21) \\ &- (26) + (31) - (36) + (41) - (46) \end{aligned} \tag{1}$$

$$\begin{aligned} \text{開放性} = 8 &+ (5) - (10) + (15) - (20) + (25) \\ &- (30) + (35) + (40) + (45) + (50) \end{aligned} \tag{2}$$

$$\begin{aligned} \text{誠実性} = 14 &+ (3) - (8) + (13) - (18) + (23) \\ &- (28) + (33) - (38) + (43) + (48) \end{aligned} \tag{3}$$

$$\begin{aligned} \text{協調性} = 14 &- (2) + (7) - (12) + (17) - (22) \\ &+ (27) - (32) + (37) + (42) + (47) \end{aligned} \tag{4}$$

表 2　予備実験の結果 (1 行目が与えた性格，1 列目が性格特性のスコア項目)

	Extraversion(外向性)	Openness(開放性)	Conscientiousness(誠実性)	Agreeableness(協調性)	Neuroticism(神経症的傾向))
Extraversion	**40**	26	23	35	24
Openness	36	**37**	38	34	30
Conscientiousness	34	31	**40**	33	31
Agreeableness	38	36	40	**40**	34
Neuroticism	30	31	39	30	**38**

$$\begin{aligned}\text{神経症的傾向} = 38 - (4) + (9) - (14) + (19) - (24) \\ - (29) - (34) - (39) - (44) - (49)\end{aligned} \quad (5)$$

4.3 実験結果

表 2 に示す実験結果から，それぞれの性格を持つように設定した ChatGPT のスコアの結果として設定された性格特性のスコアが最も高い値になっていることがわかった．具体的には Extraversion(外向性) を高く持つように設定した ChatGPT は Extraversion のスコアが 40 であり，他の性格特性の項目と比較したときに最も高い値になっていることがわかる．他の性格でも同じ様になっていることがわかった．

4.4 考　　察

実験結果から ChatGPT において設定した性格特性を表現することができると考えた．しかし，本研究において会話を行うことが重要であり，この実験では会話の中でどの程度性格を表現できるのかは不明である．したがって，今後の研究では GPT に対して性格を付与し，その性格特性が会話の中でどのように表現されるかという点についても調査が必要である．

5. ま と め

本稿では，心理的支援を目的とした ChatBot によるソーシャル・サポートシステムの提案を行った．

今後は提案したシステムを実装し，会話の中で性格特性をどの程度表現できるのか，ユーザの悩みを適切に反映できているかなどの評価を行い，ソーシャル・サポートの有用性を確認するための検証実験を行っていく．さらに，性格や悩みを付与していない通常の GPT の ChatBot との会話によるソーシャル・サポートの効果を比較実験も行っていきたい．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22K12274) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 厚生労働省「患者調査」
”https://www. mhlw. go. jp/content/12601000/001001557. pdf”
(2024/01/04 リンク確認)
[2] Caplan, G, Support systems and community mental health. Behavioral Publications, 1974.
[3] 浦光博・南隆男・稲葉昭英. ソーシャル・サポート研究: 研究の新しい流れと将来の展望 (〈特集〉「ストレスの社会心理学」). 社会心理学研究, 4(2), 78-90. 1989
[4] 福岡欣治.“日常ストレス状況での友人への自己開示とソーシャル・サポート (3) : 開示に対する友人からのサポートと気分状態の改善”. 静岡文化芸術大学研究紀要. 8 巻，p25-30，2008.
[5] 嶋信宏. 大学生におけるソーシャル・サポートの日常生活ストレスに対する効果. 社会心理学研究 第 7 巻第 1 号. 45-53. 1992.
[6] She, W. J., Dangisho, K., Siriaraya, P., Dollack, F., Nakajima, S. Matchmaking for Mental Well-being: Development of a Peer-based Support System (Peer2S) for Students during COVID Lockdown. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces. 2023. p119-122.
[7] 大津 耕陽，西田 勇樹，木内 敬太，林 勇吾：”チャットボットによる個人適応型ヘルスケアの実現に向けた 対話型課題の導入：解決志向アプローチを題材として”，ヒューマンインタフェース学会論文誌，2022 年 24 巻 4 号 p. 285-300.
[8] Petter Bae Brandtzaeg, Marita Skjuve SINTEF, Kim Kristofer Dysthe, Asbjørn Følstad SINTEF:”When the Social Becomes Non-Human: Young People’s Perception of Social Support in Chatbots”, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article No. : 257, Pages 1–13, May 2021.
[9] Goldberg, Lewis R. ”An alternative “description of personality”: The Big-Five factor structure. ” Personality and Personality Disorders. Routledge, 2013. 34-47.
[10] Goldberg, Lewis R. ”The development of markers for the Big-Five factor structure. ” Psychological assessment 4. 1 (1992): 26.
[11] DeYoung, Colin G., Lena C. Quilty, and Jordan B. Peterson. ”Between facets and domains: 10 aspects of the Big Five. ” Journal of personality and social psychology 93. 5 (2007): 880.

メンタルウェルビーイング向上のための ピアベースサポートシステム (Peer2S) の開発

談議所 孝汰† Panote Siriaraya†† Wanjou She††† 中島 伸介††††

† 京都産業大学大学院 〒 603-8047 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学 〒 606-0951 京都府京都市左京区松ケ崎橋上町
††† 奈良先端科学技術大学院大学 〒 630-0192 奈良県生駒市高山町 8916 番地 − 5
†††† 京都産業大学 〒 603-8047 京都府京都市北区上賀茂本山
E-mail: †{i2286103,nakajima}@cc.kyoto-su.ac.jp, ††spanote@kit.ac.jp, †††lavendershe2020@gmail.com

あらまし 近年，精神疾患者数の増加が世界的に懸念されており，世界経済に最大 16 兆円もの損失をもたらす可能性が指摘されている．このような状況下でメンタルヘルスサービスの提供は供給と需要のバランスが乱れ，不均衡な状態が続いている．その結果，多くの人々が適切なケアや支援を受けられず，社会全体におけるメンタルヘルスへのアクセスに格差が生じている．このような状況を改善する方法としてピアサポートが重要な役割を果たすことが知られている．ピアサポートは同じような経験や状況を共有する人々がお互いに支援し合う手段であり，精神的な健康問題を抱える人が同じような状況を経験した人から受ける支援は有益である．そこで本研究では同様の悩みや経験をしている人同士をマッチングし，ピアベースのサポートを提供することで悩みに対処できるようにするピアベースサポートシステム (Peer2S) の開発と有効性を検証した．評価実験は期間は 2 週間で大学生 25 人が参加した．実験期間中は Peer2S を使用してもらい，実験前と実験後にアンケート調査，実験後にインタビュー調査を実施した．その結果，被験者の精神状態が改善されたことが確認できた．

キーワード ヘルスケア，ピアサポート，機械学習，マッチング

1 はじめに

精神疾患は世界中で増加しており，対処しなければ 2030 年までに世界経済に最大 16 兆ドルの損害を与える可能性がある [1]．また，メンタルヘルスサービスの需要と供給が追いつかず，地方と都市の格差が深刻化するなどメンタルヘルスサービスの需給バランスが大きく崩れている．たとえば，Sugawara らによって実施された研究 [2] では，さらなるサービスが導入されない場合，メンタルヘルスサービスの不足は悪化することが示されている．メンタルヘルスケアの専門家の人材を増やすことに加えて，メンタルヘルスサポートを提供するための潜在的な代替リソースを探すことは迅速かつ効率的なアプローチになる．

ピアサポートはうつ病 [3], [4]，心的外傷後ストレス障害 (PTSD) [5], [6]，グリーフケア [7]，がん [8], [9]，心臓病 [10], [11]，ヒト免疫不全ウイルス (HIV) [12]，禁煙 [13] などいくつかの医療が必要な状態に対処するのに広く利用されている効果的なアプローチの 1 つである．病気からの回復に成功した人などの同様の経験をした人と個人を結びつけ，知識を提供し，ロールモデルとして対処してもらうことは病気に対処する際に有効である [14]．しかし，同様の否定的な状況にある人々を推薦してマッチングし，ピアベースのサポートを提供できるシステムは開発されておらず，この技術は精神疾患に対処している個人を助けるのに効果的である可能性を秘めている．

本研究では，ピアベースのサポートシステム (Peer2S) の開発と有用性について検証した．開発したマッチングシステムの目的は同じような経験や興味，悩みを持つ他のユーザにメンタルヘルスサポーターを推薦することで，ユーザがピアベースのメンタルヘルスサポーターを見つけて精神的健康を維持しやすくすることである．

2 関連研究

2.1 メンタルヘルスケアのためのピアサポート

ピアサポートグループは 18 世紀にヨーロッパの国 (イギリスとフランス) の自助グループから始まり，ボランティア活動と自助ベースで形成され，コミュニティサービスの一部と見なされていた [15]．この研究における「ピア」という用語は同様の否定的な状況を経験し，同様の興味を持ち，精神的健康に影響を与える可能性のある共通認識を持っている個人を指す．限られたメンタルヘルスの急増する需要とのギャップを埋める上でピアサポートの有望な効果が証明されている [16], [17], [18], [19]．ピアサポーターは同じ経験を共有する人同士が出会うことで，相互の共感や理解が生まれるため，同様の否定的な状況に慣れていない個人に対しての洞察，感情的なサポート，改善することへのモチベーションを提供することができる．また，サポート・プロバイダーとして機能するピアのメリットも示されている [20]．ピアサポーターが課題を克服した成功談を共有できる支援提供者に変わることは，成功例のロールモデルとして精神疾患者に良い影響を与える可能性がある．さらに，十分なサー

ビスを受けていない，または社会的に疎外化された人種やマイノリティにとってはピアサポートグループはケアのニーズを満たす上で重要な役割を果たせる．さまざまな分野の研究者が，重度の精神疾患や低所得国のヘルスケアプログラムにピアサポートプログラムを含めることの効果についてさまざまな意見を持っているが，例として薬物中毒など社会的に不利な立場に置かれたグループのための取り組みがある [15]．Shalaby と Agyapong が主張したようにピアサポートは自己アイデンティティを探求し開発する段階にあり，孤独と孤立を頻繁に経験する青少年と思春期にとっても重要である．しかし，ピアサポートプログラムは一部のメンタルヘルスケアでは普遍的なソリューションではない可能性があり，コミュニティの繋がりとアクセシビリティの制限に注意するべきである．

2.2 ピアサポートで使用される技術

現在，ソーシャル・サポートのコンテキストで人々を照合できるアルゴリズムが不足している．人口統計や好みの属性 (年齢，収入など) に依存するコンテンツベースの相反法を利用するアルゴリズム，潜在因子モデリング，協調フィルタリングなどのアルゴリズムはオンラインデートのコンテキスト向けに設計される傾向があるため，ピアベースサポートのマッチング用では検証されていない． これら [21], [22], [23] のアルゴリズムはデートのマッチング向けに開発されているが，ソーシャル・サポートのコンテキストでは否定的な経験の類似性やソーシャル・サポートに固有の要因 (例: ユーザが一人暮らしかどうか，ユーザの精神状態の段階など) などの心理的要因を考慮することも重要である．個人をマッチングする場合，ヘルスケアをサポートするために開発された他の推薦アルゴリズムでは患者同士をマッチングさせず，患者と医師または病院を患者に推薦することに重点を置いている [24], [25]．ピアベースのソーシャル・サポートの手段としてオンラインテクノロジーの使用を調査した研究がいくつかあるが，ほとんどは非同期通信テクノロジー (オンラインフォーラムやインターネットサポートグループなど) に焦点が当てられている [26], [27], [28]．本稿で開発する手法ではユーザとユーザをマッチングさせ，互いにサポートをする新たなピアベースサポートのアプローチである．ピアベースのソーシャル・サポートの研究は不足しているため，どのような要素とアルゴリズムを使用して人々をマッチングさせ，相互に質の高いピアベースのソーシャル・サポートを提供できるかを調査する．

3 ピアベースサポート（Peer2S）システムの設計・開発

本研究の目的はユーザの状況に共感し，個人的なメンタルヘルスのサポートを提供できる潜在的な支援者とユーザを結びつけるための推薦システムを開発することである． 開発したシステムは商用の出会い系アプリケーションで使用されているものと同様のメカニズムを利用しており，同じようなネガティブな経験をした「ピア」同士をマッチングさせ，お互いに支援できるような環境を提供する． 本システムは悩みのさまざまな段階にある人々を結び付け，感情的なサポートと実践的なアドバイスを交換するためのアクセス可能で安全なプラットフォームを提供し，ユーザが置かれている状況に対処できるようにする．

3.1 「ストーリーカード」の設計とソーシャル・サポートをマッチングさせる仕組み

Peer2S はメインユーザと同じようなネガティブな状況を経験や悩み，同じような興味を持っている他のユーザを推薦する．本システムでは推薦する支援者のことを「ピア」と呼んでいる．ピアを推薦する理由はユーザの状況を理解してくれる相談相手となるだけでなく，自分が他者にとって有用であることを認識するためでもある．自分の経験を同じような悩みを持つ人へのアドバイスとして提供する行為自体が精神に良い影響を与える期待がある．そこで，マッチング候補となるピアを提示する際に自分のライフストーリーを反映させるためのインターフェース「ストーリーカード」を設計した．ストーリーカードはユーザの状況や情報が公開されており，他のユーザに対してのハグ，サポートメッセージ，マッチングリクエスト機能が設計されている．他者のストーリーカードに「ハグ」や「サポートメッセージ」を送ることで他のピアはマッチングした会話にコミットせずに最低限のサポートを提供でき，まだつながる準備ができていない場合でもサポートを提供する機会が増える．また,「ハグ」や「サポートメッセージ」はユーザにサポートプロバイダーとしての役割を自認させることにも繋がる．ユーザがマッチングしたいと判断した場合には，ストーリーカード上の接続ボタンをクリックして，マッチングリクエストを送信することができる．マッチングリクエストを送信すると相手のストーリーカードがユーザのストーリーカードページに表示される．My Profile ページではピアからもらったハグやサポートメッセージが表示される．双方がストーリーカード上で「Connect」をクリックするとマッチングが成立し,「Chat」ページでコミュニケーションをとることができる．チャットページではマッチングが成立したユーザ同士がメッセージを送信することができる．

3.2 Peer2S モバイルアプリケーション

開発したシステムの画面の図を図 1 に示す．Peer2S アプリケーションはプログレッシブ Web アプリケーションとして開発されており，ユーザはモバイルやデスクトップなどのさまざまなプラットフォームからアクセスできる．図 1 の (a) はプロフィールの登録ページである．ユーザマッチングに使用する機能を収集するためにユーザが性別，居住地，学部，同居者をプルダウン形式で入力し，ユーザの悩みと趣味を自由記述形式で入力する．図 1 の (b) はユーザのプロフィールページである．**??**節で説明した「ハグ」や「サポートメッセージ」がユーザ自身に送られた場合，受信した数や内容をこのページで確認できる．図 1 の (c) は推薦されたユーザのプロフィールページである.「ハグ」や「サポートメッセージ」，マッチングリクエストをする「Connect」はこのページから送信することができる．図 1 の (d) はチャットページである．マッチングが成立した時



図 1　Peer2S アプリのメインページのデザイン　a) プロフィール登録ページ　b) ユーザのプロフィールページ　c) 推薦されたユーザのプロフィールページ　d) チャットページ

にマッチング相手とのチャットルームが開放され，ユーザ間で自由にチャットをすることができる．すべてのデータトランザクションは HTTP リクエストを通じて実行され，Secure Socket Layer プロトコルで暗号化されており，ユーザがさまざまなプラットフォームを通じてアプリにアクセスすることを想定している．そのため，サポートやマッチングリクエストを送信するためのインタラクティブなボタンを用いながら，インターフェイスは応答性が高いように設計されている．

4　提案システム (Peer2S)

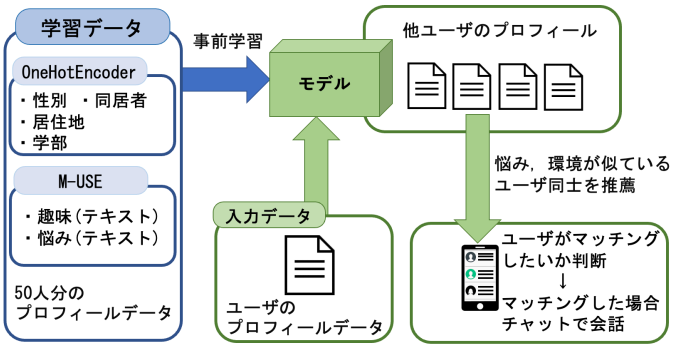


図 2　Peer2S の概要

開発したマッチングシステム (Peer2S) には適切なピアベースのソーシャル・サポートを提供する支援者を推薦できるモデルが含まれている．Peer2S の概要を図 2 に示す．45 人の大学生にモデルのトレーニングデータを生成するために，事前に登録されたプロフィールを見てマッチングしたいかを評価した．推薦モデルを構築するために，さまざまな機械学習アルゴリズムを実験している．

4.1　ユーザーをマッチングするために使用される機能

ユーザをマッチングするために使用される特徴量はユーザの性別（カテゴリ），年齢（数値），出身地（カテゴリ），学部（カテゴリ），同居者（カテゴリ：一人暮らし，実家暮らしなど），悩み (テキスト)，趣味 (テキスト) である．これらの特徴量は有効であることを確認するために 2 人の精神保健専門家，1 人の社会学者，1 人の UX 研究者によって検証されている．

4.2　機械学習モデルの構築

機械学習モデルの訓練と評価に必要なデータを取得するために，大学生に事前にプロフィールをシステムに登録してもらい，その後にさらに 45 人のユーザが参加し，各プロフィールを読んでバイナリ形式で各ユーザとマッチングすることに興味があるかどうか (はい/いいえ) を評価した．これにより，機械学習モデルのトレーニングのための学習データとして使用される 3545 個のデータサンプルが得られた．開発した推薦モデルはユーザのペアが適切に一致するかどうかを判断するための二値分類モデルである．得られた学習データの前処理として，「悩み」，「趣味」のテキストについてはストップワードの除去を行なっている．また，取得した学習データは不均衡データであるためオーバーサンプリングを行なっている．オーバーサンプリングの手法については Chawla らによって提案された SMOTE を採用している [29]．

カテゴリによる説明変数 (性別，出身県，学部，同居者) は OneHotEncoder を使用してベクトル化した．テキストによ

る説明変数 (悩み，趣味) は Multilingual Universal Sentence Encoder [30] を使用して 300 次元のベクトルに変換した．これにより，ユーザのプロフィールと評価したプロフィールの両方のベクトルがモデルの学習データとして使用されている．サポートベクターマシン (SVM)，勾配ブースティングライブラリ XGBOOST，k 近傍法 (KNN) アルゴリズム，および Siamese ニューラルネットワーク構造に基づく深層学習アプローチを含むモデルなどのいくつかのアルゴリズムを実験した．SVM，XGBOOST，KNN アプローチでは，モデルのトレーニング時にグリッドサーチを使用してハイパーパラメータの調整をしている．パラメータの概要は表 1 に記載している．Siamese ニューラルネットワークの深層学習モデルの構造は図 3 に示している．Siamese ニューラルネットワークの深層学習モデルでは高密度層の場合，活性化関数として Rectified Linear Units を使用した．ドロップアウト層にはドロップアウト率 0.4 を使用し，ADAM オプティマイザーはデフォルトの学習率 0.001 で使用した．分類の確率を使用して推薦スコアを計算した．協調フィルタリングのアプローチについては従来の協調フィルタリングと特異値分解 (SVD) [31] のアプローチを実験した．

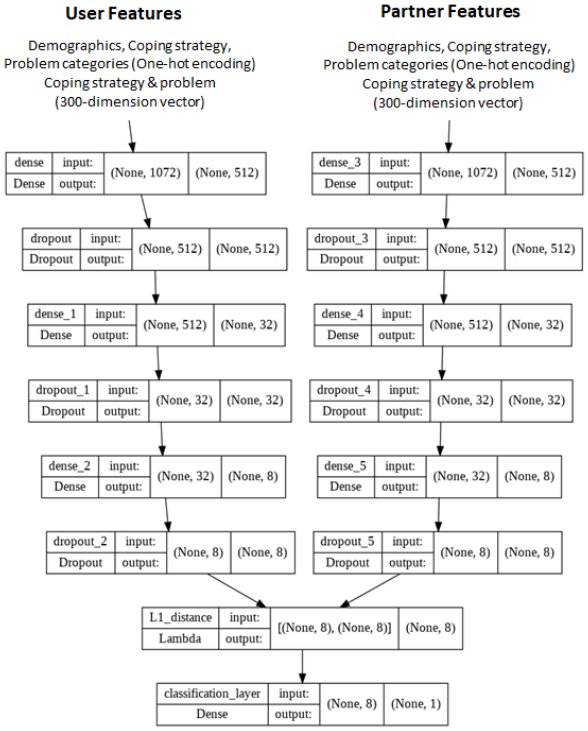


図 3　シャムニューラルネットワーク・ディープラーニングモデルの構造

4.3　モデルの精度結果

アルゴリズムを評価するために 45 人中 44 人のデータを使用してモデルをトレーニングし，残りの 1 人にユーザを推薦する LeaveOneGroupOut を利用した．1 人のユーザをテストデータ，他のユーザを訓練データとして分割を行っている．機械学習モデルの Precision@10 と Recall@10 の結果を図 4 と 5 に示す．結果は SVM が最も高いパフォーマンスを示した．これに

表 1　グリッドサーチのパラメータ範囲

アルゴリズム	ハイパーパラメータ	探索範囲
SVM	kernel	linear，rbf
	C	0．01，1
XGBoost	estimators	10，1000
	gamma	0．5，5
	max depth	3，10
	scale_pos_weight	1，1000
KNN	algorithm	auto，kd tree，brute
	weight	uniform，distance
	neighbors	1，5，10

より，5 説の評価実験では SVM モデルを採用する．

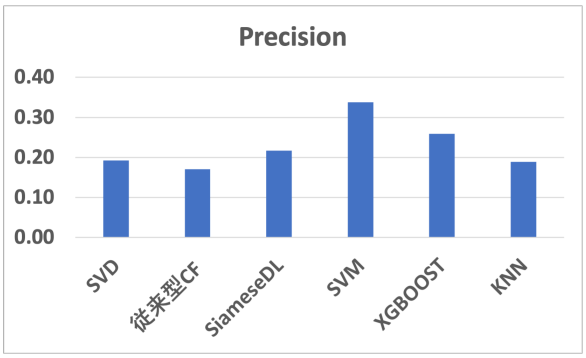


図 4　Precision@10

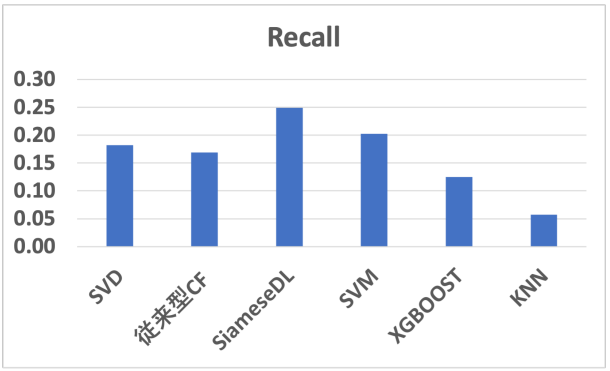


図 5　Recall@10

5　提案システム (Peer2S) の評価実験

5.1　概　　要

Peer2S システムの有効性を評価するためにユーザ調査を行なった．被験者は大学で募集され，25 人のユーザが実験に参加した．被験者は 18 歳から 24 歳までの大学生または大学院生であり (平均年齢 = 20.28，SD = 1.49)，男性 11 名，女性 14 名である．ユーザ 1 人に対して推薦したプロフィールの数は 10 件である．被験者に 2 週間の間，開発したシステム (Peer2S) を使用してもらい，Peer2S が被験者の精神状態に与えた影響を評価するためにアンケート調査とインタビュー調査を行なった．

実験スケジュールを図 6 に示す．実験期間は 2 週間で実験開

実験2週間前	GAD，ソーシャル・サポート尺度
実験直前	GAD，ソーシャル・サポート尺度
実験1日目~14日目	実験
実験終了後	GAD，ソーシャル・サポート尺度， ユーザビリティアンケート,インタビュー

図 6　実験スケジュール

始 2 週間前と実験開始直前に不安状態，ソーシャル・サポートに関するアンケートに回答してもらい，実験終了後にシステムの有効性を評価するために不安状態，ソーシャル・サポート，ユーザエクスペリエンスに関するアンケートの回答とインタビュー調査を実施した．実験開始前に被検者に自分のプロフィールをシステムに登録してもらった．その後，登録したプロフィールが他のユーザに推薦され，実験期間中に被験者はお互いの被験者のプロフィールを閲覧し，マッチングするかどうかを判断できる． 双方がマッチングに同意するとチャット機能が解放され，アプリケーションを通じてお互いに自由に会話ができる．実験開始 2 週間前と実験開始直前のアンケート結果を比較することにより，被験者の普段の生活で実感している不安状態，周囲の人からのソーシャル・サポートについて調査する．また，実験開始前と実験後のアンケート結果を比較することにより，提案手法 (Peer2S) が被験者にどのような影響を与えたかを調査した．

5.2　アンケート調査

提案システムの有用性について検証するために，アンケートを実施した．

- GAD-7(Generalized Anxiety Disorder-7) 日本語版 (2018)
- ソーシャル・サポート尺度
- 大学生用ソーシャル・サポート尺度
- Technology Acceptance Model (TAM)
- User Experience Questionnaire (UEQ)

5.2.1　GAD-7 (Generalized Anxiety Disorder-7)

GAD-7 は全般性不安障害 (GAD) を評価するための自己申告式アンケートとして開発され，7 つの項目で構成されている [32]． これらの項目には，回答者が緊張やイライラを感じているかどうか，リラックスするのが難しく頻繁に心配していると感じているかどうかなどの要素が含まれている．日本語版は村松らによって開発された [33].

5.2.2　ソーシャル・サポート尺度

Zimet ら [34] が開発した「ソーシャル・サポート尺度」(Multidimensional Scale of Perceived Social Support) から岩佐らによって作成した日本語版を使用する [35]．被験者が知覚しているソーシャル・サポートを評価するために他者から受ける社会的な支援やつながりの程度を測定する．原版と同様の 3 因子構造 (「家族のサポート」,「大切なサポート」,「友人のサポート」) が確認されている．本研究では「友人のサポート」尺度のみを使用し,「友人」の部分を「アプリケーションで見つけた支援者」としている．

5.2.3　大学生用ソーシャル・サポート尺度

大学生用ソーシャル・サポート尺度は片受らによって作成され，大学生に対する知覚されたソーシャル・サポートを測定する尺度である [36]．さまざまな種類のサポートの側面 (感情的，親和的，情報的，手段的，評価的) を網羅しており，4 段階のリッカート尺度で評価をする．

5.2.4　Technology Acceptance Model(TAM)

TAM は Davis らによって作成され，ユーザが情報機器をどのように利用するかを説明するための人間の行動意思モデルである [37].「ユーザが知覚した有用性」と「ユーザが知覚した使いやすさ」と「ユーザのシステム利用への行動意思」はユーザーがあるシステムを使用する意欲について決定するための重要な要素である．本研究では日本語版 [38] を使用し，3 つの尺度 (知覚された有用性，知覚された使いやすさ，システム利用への行動意思) について Peer2S システムを使用する意向の度合いを表す．被検者は 7 段階のリッカート尺度で評価をする．

5.2.5　User Experience Questionnaire(UEQ)

UEQ [39] はユーザビリティ (効率，目立ちやすさ，信頼性) とユーザエクスペリエンス (独創性，刺激) を 26 項目で測定する評価指標である．30 以上の言語が使用可能で本実験では日本語版を使用する．質問内容は 2 相対する属性で構成されており，被験者はそれぞれの項目について 7 段階のリッカート尺度で評価をする．-3 から+3 の値を取り+0．8 以上の場合は肯定的な結果であり，-0.8 以下の場合は否定的な結果として解釈できる．

5.3　アンケート結果

	数
マッチング (チャットルーム) 総数	68
送信されたメッセージ総数	309
メッセージ送信数が 0 のチャットルーム	18
マッチング数が 0 のユーザ	1

表 2　マッチング総数とメッセージ総数

表 2 に実験期間中に成立したマッチング総数や送信されたメッセージ総数について示す．

GAD と 2 つのソーシャル・サポート尺度のアンケート結果に大きな変化があったかどうかを調査するために，対応のある T 検定を実施した．

その結果，GAD では実験 2 週間前 (M=6.32，SD=1.59) と実験直前 (M=6.84，SD=1.53) と実験後 (M=4.92，SD=1.41) のスコアを比較すると，実験直前と実験後で不安が有意に減少した ($t = 2.10$，$p < 0.05$).

Zimet のソーシャル・サポート尺度では実験 2 週間前 (M=4.34，SD=4.37) と実験直前 (M=4.37，SD=5.16) と実験後 (M=5.04，SD=5.10) のスコアを比較すると，実験 2 週間前と実験後 ($t = -2.5$，$p < 0.05$), 実験直前と実験後 ($t = -2.33$，$p < 0.05$) で被験者が知覚したソーシャル・サポートが有意に増加した． しかし，大学生用ソーシャル・サポート尺度には有意差はなかった．

TAM の結果から全ての項目で 4 を超えており，肯定的な結果となった．その中でも Intention to Use(ユーザのシステム利

用への行動意思) が最も高く，Perceived Ease of Use(ユーザが知覚した使いやすさ) が最も低いことが分かる．

UEQ では「見栄え」「信頼性」の 2 項目で肯定的な結果となった.「見栄え」は視覚的なデザインや全体的な印象に関連している.「信頼性」はシステムは予測可能で安全であるかを示す項目である．また，それ以外の 4 つの項目では 0〜0.8 点であり，やや肯定的な結果であった．

GAD の結果を図 7，Zimet のソーシャル・サポートの結果を図 8，大学生用ソーシャル・サポートの結果を図 9，TAM の結果を図 10，UEQ の結果を図 11 に示す．

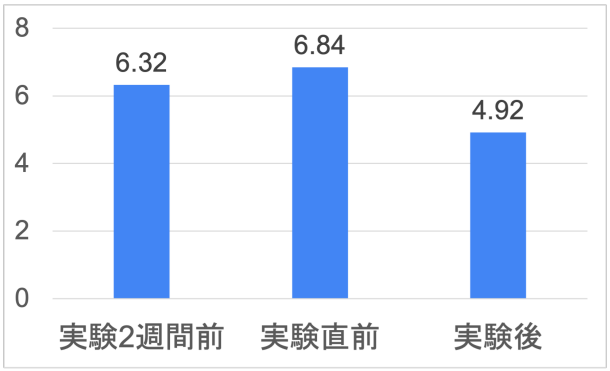


図 7　不安の平均スコア

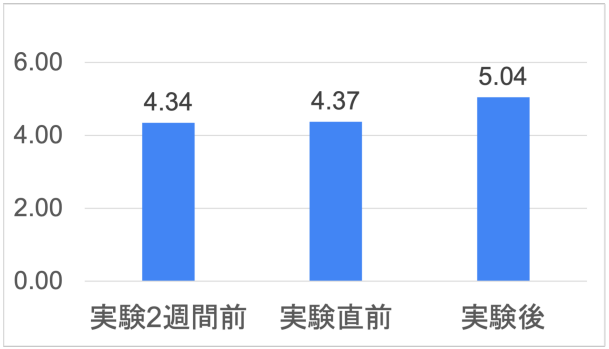


図 8　Zimet ソーシャルサポート尺度の平均スコア

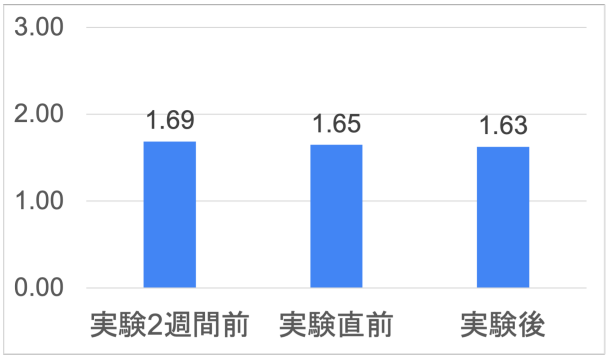


図 9　大学生用ソーシャル・サポートの平均スコア

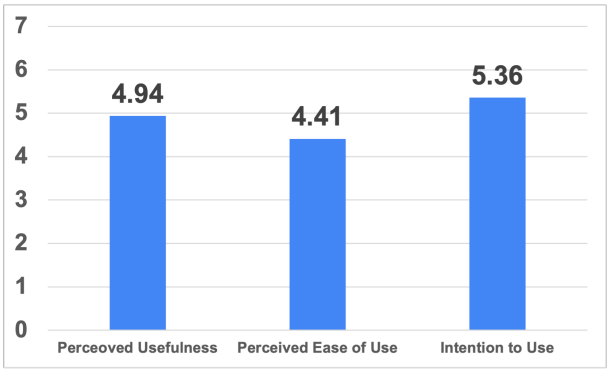


図 10　TAM の平均スコア

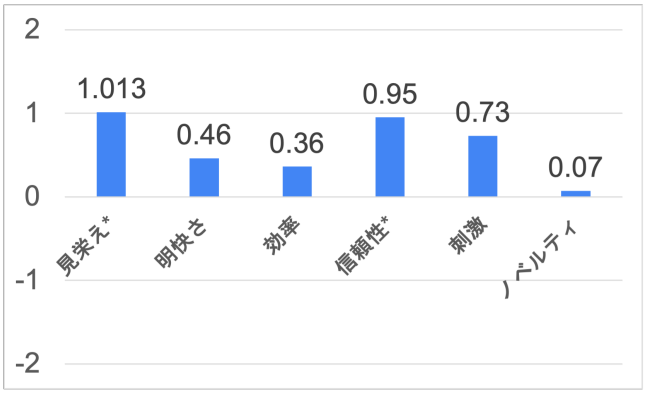


図 11　UEQ の平均スコア

5.4　インタビュー調査

5.4.1　概　　要

システム (Peer2S) の使いやすさ，改善点などを口頭インタビュー形式で質問を行なった．インタビューは実験終了後に実施した．質問は主に，

- システム (Peer2S) 全体を通した感想
- 推薦された支援者について
- システム (Peer2S) の改善点

について実施した．

5.4.2　インタビュー結果

それぞれの質問に対する被験者の回答と考察を説明する．システム (Peer2S) 全体を通した感想について，次のような回答が得られた．

- 緑を基調としたデザインや機能に優しさを感じた
- 精神の健康状態が改善されたとは思わなかったが、システムを利用してみて楽しかった

デザインについての回答から，視覚的に得られる情報からポジティブな印象を受けていることが分かる．これは 5.3 節で紹介したアンケート結果の裏付けになっており，視覚的なデザインや全体的な印象に関連している UEQ の「見栄え」と TAM の Intention to Use(ユーザのシステム利用への行動意思) が高いことを裏付けている．さらに，実験の感想として精神の健康状態についての回答から，サポートを受けている実感があまりなかったが，実験を通して幸福感を感じていることが分かった．これは 5.3 節で紹介した知覚されたソーシャル・サポートの測

定に関するアンケート結果の裏付けになっている．実験を通して他者の様々な悩みの存在を確認することで孤独感が減り，安心感が得られたことが分かった．

次に，推薦された支援者について質問したところ，次のような回答が得られた．

- 推薦された人は相談相手として自分に合っていると思った
- 知らない人だからこそ話せる悩みがあった
- あまり似ている悩みを持つ人がいなかった

推薦された支援者についての回答から，プロフィールから得られる情報から適切な支援者を推薦できていることが分かる．インタビューを受けた 25 人中 18 人は相談相手を選ぶ基準として「共通点がある人」と回答しており，特に「悩み」が似ている人を相談相手として選びたいという回答が多かった．しかし，あまり似ている悩みを持っている人がいなかったという意見があった．この回答をした被験者の特徴としては今回の実験の被験者グループの中ではマイノリティな悩みを持っており，悩みが似ている人を推薦することは困難であった．今回の実験に人的限りがあったが，実際にマイノリティな悩みを持つ人への有効な支援方法を検討する必要がある．他に面識がない人の方が悩みを打ち明けやすいという回答が得られた．この回答から，日常的に人間関係が豊富な人に対しても，マッチングによる支援者探索は有効であると考えられる．

次に，システム (Peer2S) の改善点について，次のような回答が得られた．

- 会話にテンポ感がなかった
- 会話の始め方が難しかった

改善点についての回答から，チャット機能について改善の余地があることが分かった．開発したシステム (Peer2S) に通知機能がなかったため,「メッセージを受信したことに気付くまでの時間が掛かった」，メッセージ受信時の時間表示がなかったため,「相談相手のメッセージやりとりの感覚が掴みにくい」という意見を聞くことができた．これらについて改善することで会話が弾みやすくなり，ソーシャル・サポートの結果にポジティブな影響を与える可能性がある．

6 考　　察

不安は大学生によく観察される感情であり [40]，不安の減少は有意な傾向にあった．互いにチャットした参加者の数は限られていたが，学生同士のマッチングや不安レベルの低下にプラスの効果を観察することができた．このような発見はユーザと潜在的なピアサポーターをペアにするシステムの有効性を裏付け，学生の精神的健康を改善する上で本システムの可能性を示唆している．見当外の結果として参加者が精神的健康の改善を示したにも関わらず，ソーシャル・サポートの認識に変化がなかったことである．ユーザが推薦されたピアサポーターとマッチングしてチャットで会話することで，よりサポートされていると感じると想定していた．しかし，必ずしも「サポートされている」と感じられるわけではない可能性がある．その原因として 1 通目のメッセージの送信が困難であったことがいえる．表 2 からマッチング総数に対して 1 度もメッセージの送信がなかったチャットルームは 26.5%であった．また，インタビューからは「どのようなメッセージを送ればいいか分からなかった」,「悩みを開示しづらかった」などの意見が得られた．これらから一部の被験者は 1 通目のメッセージの送信の負担が大きく感じていたといえる．

アンケートやインタビュー結果から他のユーザが同じような悩みを抱えていたり，同じ人生の課題を経験していることにユーザが気づくと，孤立感が減り，安心し，自分たちのネガティブな経験を正当化できる可能性がある．このような思考により，悪影響を及ぼしかねない状況にうまく対処できるようになり，不安が軽減された可能性がある.「同情」に焦点を当てた文献では福祉サービスや援助を受けるクライエントの感情に同情を示すことによって，カウンセラーとして活動する人々は自己受容が向上することが指摘されている．カウンセラーとクライアントの間で「悩み」の共有をすることは悩みに対処しようとしている人が他にもいることを認識させ，双方に利益をもたらすことが示唆されている [41]．オンライン上のチャットでの会話は間接的なインタラクションであるため，ユーザがソーシャル・サポートの増加を実感しづらい可能性がある．しかし，システム内での主なやり取りとしては他のユーザのストーリーカードをレビューしてサポートをする内容が含まれていたため，サポーターとして行動した被験者も同じように力を与える経験を感じている可能性があり，精神的健康の改善に繋がったと考えられる．また，インタビューからユーザは「同情」を示すためにハグやコメントを使用しており，これは「サポートを受ける側」ではなく「サポーター」としての役割を暗示しているということも考えられる．よって，被検者は利他的な行動を行うことで自己価値が向上し，知覚したソーシャル・サポートの大幅な増加は感じられなかったと結論づける．

7 ま と め

本研究では，Peer2S を設計，開発，テストをした．Peer2S は同様のネガティブな経験をしている人々が互いにつながり，ピアベースのソーシャル・サポートを提供できることを目的としたマッチングシステムである．ケーススタディとして大学生に対象を絞り，推薦モデルを開発するために，複数の機械学習アルゴリズムを実験した．その結果，SVM が最も高いパフォーマンスを示した．その後，25 人の被験者に対して評価実験を実施した．開発したシステム (Peer2S) を 2 週間使用してもらい，Peer2S が被験者に与えた影響について検証した．その結果，不安の軽減に効果的ではあったものの，使用後のソーシャル・サポートの認識には大きな差がなかった．全体として，この研究にはいくつかの制限がある．本研究は日本の大学生を対象に実施した．その結果が他の国や教育環境からの学生に一般化できるかどうかは調査する必要がある．さらに，Peer2S が被験者に与えた影響を調査するために 2 週間の使用期間後にのみアンケートやインタビューで検証を行なった．Peer2S の使用後について，不安の軽減がより長く持続するかどうかを調査するに

は，より長期の評価を実行する必要がある．

将来的には，さまざまな使用状況でシステムを実験する予定である．特に，希少疾患やマイノリティな悩みに苦しむ人々や，大切な人の喪失に苦しむ人々を支援するためにシステムを導入することを目指している．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22K12274) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] V Canady. 2018. Mental illness will cost the world 16 trillion (USD) by 2030. Ment. Health Wkly 28 (2018), 7–8.

[2] Norio Sugawara, Norio Yasui-Furukori, and Kazutaka Shimoda. 2021. Projections of psychiatrists' distribution for patients in Japan: a utilization-based approach. Human Resources for Health 19, 1 (2021), 1–10 1 .

[3] Belinda Melling and Terry Houguet-Pincham. 2011. Online peer support for individuals with depression: a summary of current research and future considerations. Psychiatric rehabilitation journal 34, 3 (2011), 252.

[4] Paul N Pfeiffer, Michele Heisler, John D Piette, Mary AM Rogers, and Marcia Valenstein. 2011. Efficacy of peer support interventions for depression: a meta-analysis. General hospital psychiatry 33, 1 (2011), 29–36.

[5] John F Greden, Marcia Valenstein, Jane Spinner, Adrian Blow, Lisa A Gorman, Gregory W Dalack, Sheila Marcus, and Michelle Kees. 2010. Buddy- to-Buddy, a citizen soldier peer support program to counteract stigma, PTSD, depression, and suicide. Annals of the New York Academy of Sciences 1208, 1 (2010), 90–97.

[6] Shaili Jain, Caitlin McLean, Emerald P Adler, and Craig S Rosen. 2016. Peer support and outcome for veterans with posttraumatic stress disorder (PTSD) in a residential rehabilitation program. Community Mental Health Journal 52, 8 (2016), 1089–1092.

[7] Ruth M Swartwood, Patricia McCarthy Veach, Jessica Kuhne, Hyun Kyung Lee, and Kangting Ji. 2011. Surviving grief: An analysis of the exchange of hope in online grief communities. OMEGA-Journal of Death and Dying 63, 2 (2011), 161–181.

[8] Louisa M Hoey, Sandra C Ieropoli, Victoria M White, and Michael Jefford. 2008. Systematic review of peer-support programs for people with cancer. Patient education and counseling 70, 3 (2008), 315–337.

[9] H Sharon Campbell, Marie Rose Phaneuf, and Karen Deane. "Cancer peer support programs—do they work?". Patient education and counseling 55, 1 (2004), 3–15, 2004.

[10] Monica Parry and Judy Watt-Watson. 2010. Peer support intervention trials for individuals with heart disease: a systematic review. European Journal of Cardiovascular Nursing 9, 1 (2010), 57–67.

[11] Amy Scheel, Andrea Beaton, Emmy Okello, Chris T Longenecker, Isaac Omara Otim, Peter Lwabi, Craig Sable, Allison R Webel, and Twalib Aliku. 2018. The impact of a peer support group for children with rheumatic heart disease in Uganda. Patient education and counseling 101, 1 (2018), 119–123.

[12] Patricia Marino, Jane M Simoni, and Louise Bordeaux Silverstein. 2007. Peer support to promote medication adherence among people living with HIV/AIDS: the benefits to peers. Social work in health care 45, 1 (2007), 67–80.

[13] Pauline Ford, Anton Clifford, Kim Gussy, and Coral Gartner. 2013. A systematic review of peer-support programs for smoking cessation in disadvantaged groups. International journal of environmental research and public health 10, 11 (2013), 5507–5522.

[14] Julie Repper and Tim Carter. 2011. A review of the literature on peer support in mental health services. Journal of mental health 20, 4 (2011), 392–411.

[15] Reham A Hameed Shalaby and Vincent IO Agyapong. 2020. Peer support in mental health: literature review. JMIR Mental Health 7, 6 (2020), e15572.

[16] Larry Davidson, Matthew Chinman, David Sells, and Michael Rowe. 2006. Peer support among adults with serious mental illness: a report from the field. Schizophrenia bulletin 32, 3 (2006), 443–450.

[17] Larry Davidson, Kimberly Guy, et al. 2012. Peer support among persons with severe mental illnesses: a review of evidence and experience. World psychiatry 11, 2 (2012), 123–128.

[18] Kathleen O'Leary, Arpita Bhattacharya, Sean A Munson, Jacob O Wobbrock, and Wanda Pratt. 2017. Design opportunities for mental health peer support technologies. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 1470–1484.

[19] Marc Galanter. 1988. Zealous self-help groups as adjuncts to psychiatric treatment: A study of Recovery, Inc. American journal of Psychiatry 145, 10 (1988), 1248–1253.

[20] Galia S Moran, Zlatka Russinova, Vasudha Gidugu, Jung Yeon Yim, and Catherine Sprague. 2012. Benefits and mechanisms of recovery among peer providers with psychiatric illnesses. Qualitative health research 22, 3 (2012), 304–319.

[21] Peng Xia, Benyuan Liu, Yizhou Sun, and Cindy Chen. 2015. Reciprocal recommendation system for online dating. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 234–241.

[22] James Neve and Ivan Palomares. 2019. Latent factor models and aggregation operators for collaborative filtering in reciprocal recommender systems. In Proceedings of the 13th ACM conference on recommender systems. 219–227.

[23] Lin Chen, Richi Nayak, and Yue Xu. 2011. A recommendation method for online dating networks based on social relations and demographic information. In 2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE, 407–411.

[24] Yongjie Yan, Guang Yu, and Xiangbin Yan. 2020. Online doctor recommendation with convolutional neural network and sparse inputs. Computational Intelligence and Neuroscience 2020 (2020).

[25] Yong-Feng Huang, Peng Liu, Qiao Pan, and Jing-Sheng Lin. 2012. A doctor recommendation algorithm based on doctor performances and patient preferences. In 2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP). IEEE, 92–95.

[26] Terry Hanley, Julie Prescott, and Katalin Ujhelyi Gomez. 2019. A systematic review exploring how young people use online forums for support around mental health issues. Journal of mental health 28, 5 (2019), 566–576.

[27] Julie Prescott, Terry Hanley, and Katalin Ujhelyi Gomez. 2019. Why do young people use online forums for mental health and emotional support? Benefits and challenges. British Journal of Guidance & Counselling 47, 3 (2019),

317–327.
[28] Panote Siriaraya, Caleb Tang, Chee Siang Ang, Ulrike Pfeil, and Panayiotis Zaphiris. 2011. A comparison of empathic communication pattern for teenagers and older people in online support communities. Behaviour & Information Technology 30, 5 (2011), 617–628.
[29] Nitesh V Chawla et al. “SMOTE: synthetic minority oversampling technique”. In: Jour- nal of artificial intelligence research 16 (2002), pp. 321–357.
[30] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907. 04307 (2019).
[31] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. Advances in neural information processing systems 20 (2007).
[32] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine 166, 10 (2006), 1092–1097.
[33] 村松公美子. ”Patient Health Questionnaire (PHQ-9, PHQ-15) 日本語版および Generalized Anxiety Disorder-7 日本語版-up to date.” 新潟青陵大学大学院臨床心理学研究 7 (2014): 35-39.
[34] Zimet GD, Dahlem N W , Zimet SG, et al. T he M ultidemensional Scale of Perceived Social Sup- port. Journal of Personality Assessment 1988; 52:30-41.
[35] 岩佐 一, 権藤 恭之, 増井 幸恵, 稲垣 宏樹, 河合 千恵子, 大塚 理加, 小川 まどか, 高山 緑, 蘭牟田 洋美, 鈴木 隆雄. 日本語版「ソーシャル・サポート尺度」の信頼性ならびに妥当性ー中高年者を対象した検討ー. 第 54 巻第 6 号「厚生の指標」.2007.
[36] 片受 靖, 大貫 尚子. 大学生用ソーシャルサポート尺度の作成と信頼性 ・ 妥当性の検討ー評価的サポートを含む多因子構造の観点からー. 立正大学心理学研究年報 第 5 号. 2014.
[37] Fred D Davis. 1985. A technology acceptance model for empirically testing new end-user information systems: Theory and results. Ph. D. Dissertation. Massachusetts Institute of Technology.
[38] 矢入 郁子. 高齢者の Technology Acceptance と人工知能. 人工知能 31 巻 3 号. 2016 年
[39] Laugwitz, B., Schrepp, M. & Held, T. (2008). Construction and evaluation of a user experience questionnaire. In:Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.
[40] Daniel Eisenberg, Sarah E Gollust, Ezra Golberstein, and Jennifer L Hefner. 2007. Prevalence and correlates of depression, anxiety, and suicidality among university students. American journal of orthopsychiatry 77, 4 (2007), 534–542.
[41] Jordan T Quaglia, Clarissa Cigrand, and Hannah Sallmann. 2021. Caring for you, me, and us: The lived experience of compassion in counselors. Psychotherapy (2021).

社員が記述したテキストからの幸福度推定

林　純子† 伊藤　和浩† 眞鍋　雅恵† 渡邉　寧†† 中山　真孝††
内田由紀子†† 若宮　翔子† 荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916 番 5 号
†† 京都大学 人と社会の未来研究院　〒 606–8501 京都府京都市左京区吉田下阿達町 46 番
E-mail: †{hayashi.junko.hh5,ito.kazuhiro.ih4,mmanabe,wakamiya,aramaki}@is.naist.jp,
††yasushi.watanabe.77@gmail.com, †††{nakayama.masataka.4v,uchida.yukiko.6m}@kyoto-u.ac.jp

あらまし　Well-being は近年，人の生活をより良くしていくための概念として注目を集めている．このため，Well-being の指標の一つとして幸福度を調査し，Well-being の向上を図る努力が行われている．現在，幸福度を測る方法として質問に対し数値で幸福度を回答するという質問紙による方法が主流である．しかし，質問紙による調査は頻回な調査が想定されていないため，短期間での Well-being の変化を捉えることができず，Well-being 向上の要因の特定や，モデル化が困難であった．そこで，本研究では，従来のように質問紙を用いず，自由に記述した日報テキストから幸福度を推定することを試みる．まず，大規模な日報テキストと主観的な幸福度（0〜10 の 11 段階）の点数を収集した．次に，これらに対して大規模言語モデルを用いて，日報テキストから幸福度を推定するモデルを構築した．その結果，自己申告の幸福度スコア群と，モデルが推定した幸福度スコア群の間には有意な相関が見られ，モデルの妥当性が確認された．

キーワード　ウェルビーイング，自然言語処理応用，職場，幸福，日報

1 はじめに

Well-being とは人生における様々な側面を総合して，良い状態が持続的である状態を指す．世界保健機関 (WHO) 憲章では，健康を「肉体的，精神的，そして社会的な Well-being が満たされた状態」と定義している [1]．ここから，健康状態を知るために Well-being が指標化され，様々な事象との関連が明らかにされてきた．中でも，職場におけるパフォーマンスは Well-being と深い関係があるということが今までの研究において詳らかにされている．例えば，Well-being のスコアが高い人ほど収入が多いといった報告 [2] や，Well-being のスコアが低い人は職場における生産性が低い傾向があるという報告 [3]，そして離職率にも影響するという報告 [4] がある．多くの調査結果が示すように，従業員の Well-being の把握が重要であることから，職場の Well-being に関する世界的な大規模調査が行われている [5]．

一方で，これまでの研究では，職場における個人の細かな Well-being の変化についての調査は行われていない．例えば質問紙による調査方法である人生満足尺度（Satisfaction With Life Scale; SWLS）[6] は，“In most ways my life is close to my ideal.” といった自分の人生の満足度を問う 5 つの質問に対し，リッカート尺度で答えるものである．多くの研究において質問紙調査は用いられているが，毎日多くの同じ質問をすることが被験者にとって負担となり，同じ回答をするようになってしまうという問題 [7] が発生しやすく，細かな変化を測定するのには不向きである．このような問題を解決する手法として，テキストから Well-being のスコアを推定する試みが行われてきた [8] [9] [10]．これらの研究において，テキストから Well-being のスコアを推定することは可能であると示唆されている．テキストを頻繁に収集できれば Well-being の細かな変化を捉えることが可能である．しかし個人の Well-being のスコアの推移が捉えられているかについては調査が行われておらず，また職場において実用に足るかどうかについては議論されていない．

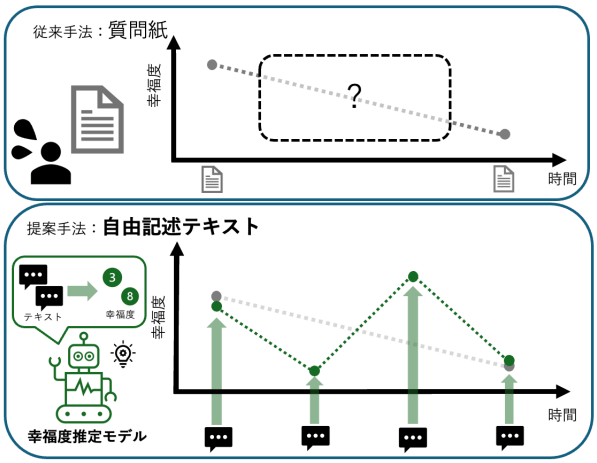


図 1　本研究の概要．質問紙調査では困難な短期間での調査を実現するため，本研究は自由記述テキストという簡易に行うことのできる調査から幸福度の変化を推定する．

本研究では Well-being をよく表すものとしてその日の幸福度を自由記述テキストから推定するモデルを構築し，個人の幸福度をどれほど推定できるのかを調査する．図 1 のように，自由記述テキストと記述日の幸福度を収集し，大規模言語モデルである BERT [11] を用いて幸福度推定モデルを構築する．モデルの評価は，本研究の目的が個人レベルの幸福度の推移の推定で

あることから，各ユーザの幸福度推移について行う．

また，職場単位での幸福度推定モデルの性能検証のため，複数の職場で収集したデータを用いて実験を行う．具体的には，ある職場のデータで構築したモデルが別の職場においても機能するのか，また同じユーザによるデータが訓練データとテストデータに入っている場合の性能についても調査する．

なお，本研究は京都大学にて，奈良先端科学技術大学院大学を含む倫理申請を経て，許可された研究課題である (審査番号 26-P-16).

2 関連研究

1948 年に世界保健機関 (WHO) が健康の定義として Well-being という単語を用いてから，Well-being を指標化する流れが生まれ，いくつかのアプローチが提案された．Well-being の測定に用いられてきた尺度をまとめた報告 [12] によれば，心理学における主なアプローチは快楽主義と幸福主義の 2 つである．

まず一つ目に快楽主義とは，苦痛が回避され，人生満足感が強く，肯定的感情の経験が多く，否定的感情の経験が少ない状態を幸福として理解するというものである [13]．快楽主義的なアプローチとして知られている尺度に，PANAS 尺度 [14] がある．

二つ目に，幸福主義とは，自己実現や個人的成長等の経験を幸福であると定義する考え方であり，理性主義とも言われる．幸福主義的なアプローチとして知られている理論は，Well-being が自己受容（Self-Acceptance），他者との肯定的関係（Positive Relations），自律性（Autonomy），環境の習得（Environmental Mastery），人生の目的（Purpose in Life），個人の成長（Personal Growth）から成り立つ [15] というものである．

これら二つの考え方を統合したものが PERMA 理論 [16] である．P (Positive Emotion), E (Engagement), R (Relationships), M (Meaning), A(Achievement) の 5 つの因子から Well-being を説明したもので，この理論を用いた質問紙による計測方法に，PERMA-Profiler [17] が存在する．この質問紙では，「通常、あなたは目的や意味のある生活を、どの程度送っていますか。」や「あなたは自分で立てた大切な目標を、どのくらいの頻度で達成できますか。」といった質問に対し，0 から 10 の 11 段階で回答する．

このように質問紙を用いて幸福度が測定されてきた中，他の要素との関連についても研究が行われてきた．ここでは幸福度とテキストとの関連についての研究を挙げる．まず，ソーシャルメディアの投稿と幸福度スコアを学習し，幸福度を推定するという研究 [8], [9] がある．これらの研究ではソーシャルメディアに投稿された文章に対し幸福度スコアを他人がアノテーションしていたり，地域における幸福度スコアを電話調査によって集め，地域レベルでの相関を取ると言ったアプローチをとっている. そのため実際に文章を投稿した本人の幸福度ではないものの，テキストから推定された幸福度と自己申告された幸福度には正の相関が見られるという結果が得られている．

次に，自由記述のテキストに対してテキストを書いた本人が幸福度スコアをつけ，それらを学習し幸福度スコアを予測するという研究 [18], [19] がある．これらにおいても同様に，推定された幸福度と自己申告の幸福度との間には相関があるということが示されている．他にも，多くの研究ではサポートベクターマシンや決定木といった手法が用いられている．人生満足度の質問をテキストで回答してもらい，BERT を用いて幸福度を推定した研究 [19] において，文脈を考慮した方が性能が高いという結果がある．よって本研究においても手法として文脈を考慮できる BERT を用いて学習を行う．これまでのテキストと幸福度の関連を調べた研究においては生活満足度の質問紙に答えてもらい，そのスコアと予測したスコアの相関を求めている．

表 1　日報入力アプリの概要（☆のみ本実験で使用）

項目	質問文	入力形式
日報テキスト (☆)	3 行程度で日記をご記入ください（仕事に無関係でも構いません）	文字数制限なしの自由記述
個人の幸福感 (☆)	あなたは一日幸せでしたか？	0（とても不幸）から 10（とても幸せ）の 11 段階から 1 つ選択
チームの幸福感	チーム全体としては、一日幸せだったと思いますか？	0（とても不幸）から 10（とても幸せ）の 11 段階から 1 つ選択
勤務場所	業務中、主に滞在した場所をお選び下さい	自宅，関西支社，取材場所，媒体社，その他
その他の場合の勤務場所	その他を選択した方は，差し支えない範囲で場所をご記入ください	文字数制限なしの自由記述

3 材　　料

職場の従業員を対象にした過去の幸福度調査 [20] において H 社より収集したデータを使用した．データの概要を表 1 に示す．なお，個人の幸福感は，幸福を 10 段階の梯子になぞらえて尋ねる cantril ladder [21] に基づいた質問となっている．本実験では，推定モデルを作成するために，H 社と追加で収集した P 社の社員の日報テキストと，個人の幸福感の値を毎日の幸福度スコアとして用いた．

日報入力には Web ブラウザ上で動作するオリジナルのアプリケーションを使用している．参加者への指示は，「入力タイミングは原則勤務日ごとに 1 回以上，勤務終了日にその日の出来事を振り返りながら入力」とした．日報テキストや個人の幸福感が空欄であったり，アプリの仕様上複数回違う幸福度スコアを記入した際に小数となったデータは削除した上で使用している．

3.1 職場 1：H 社

一つ目の職場は，国内大手の広告・マーケティング事業を営む H 社の関西支社である．日報を使った実験に事前に同意した従業員 121 名のうち，94 名が 2022 年 9 月 1 日〜10 月 31 日の 2 ヶ月間にわたり日報入力に参加した．本実験では，94 名の社

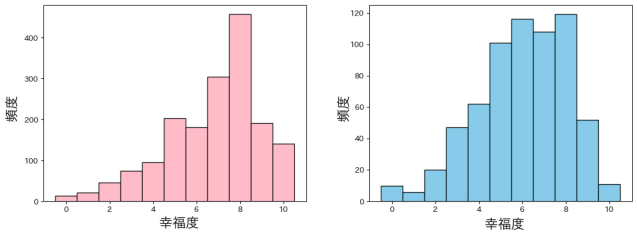


図 2　収集データの幸福度分布

員により入力された 1,728 件の日報テキストと幸福度スコアを，H 社から収集したデータとして使用する．収集したデータの分布を表 2(a) に示す．

3.2　職場 2：P 社

二つ目の職場として，国内大手電機メーカー P 社においてデータを収集した．H 社と同様，事前に実験に同意した社員 11 名が，2022 年 12 月 19 日〜2023 年 5 月 18 日の 6 ヶ月間にわたり日報入力に参加し，652 件の日報テキストと幸福度スコアを収集した．収集したデータの分布を表 2(b) に示す．

4　実　　験

4.1　設　　定

本実験では，社員が記述したテキストを入力とし，推定した幸福度スコアを出力する Bidirectional Encoder Representations from Transformers (BERT) [11] ベースの推定モデルを構築する．具体的には，東北大の事前学習済み BERT[1] を用いて，AutoModelForSequenceClassification[2] の回帰モデルを構築した．なお，学習率は 0.000002，エポック数は 20 とし，オプティマイザーには AdamW [22] を用いた．

評価指標には，既存研究 [10] でも用いられているピアソン相関係数を採用する．そして，モデルが推定した幸福度スコア（以下，推定幸福度）と社員が自己申告した幸福度スコア（以下，申告幸福度）の相関係数を算出する．本実験では，ユーザ単位で相関係数を評価するユーザ単位幸福度推定（4.1.1 節）と，職場単位で相関係数を評価する職場単位幸福度推定（4.1.2 節）の 2 種類を行った．

評価においては，同じ項目を測定するために設計されたスケール間のピアソン相関係数は 0.4 以上であることが期待される [23] ことから，相関係数が 0.4 以上であれば幸福度推定モデルが機能するとみなす．

4.1.1　ユーザ単位幸福度推定

各ユーザに対し，申告幸福度と推定幸福度の相関係数を算出する．ここで用いる推定モデルは，後述する職場単位幸福度推定の 2 社別データ実験においても使用する．この実験においては，H 社のデータで学習したモデルが，P 社の個人のテキストからどの程度の性能で推定幸福度を予測できるのかを検証する．なお，P 社の 11 名のユーザをそれぞれ U1〜U11 とする．

1：https://github.com/cl-tohoku/bert-japanese
2：https://huggingface.co/docs/transformers/model_doc/auto

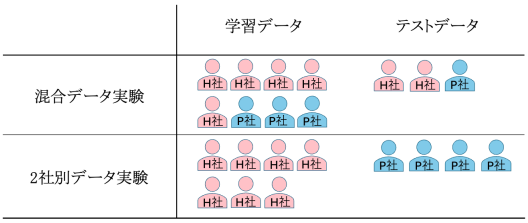


図 3　混合データ実験と 2 社別実験におけるデータ分割

表 2　ユーザ単位での相関係数．U1 から U11 は P 社の 11 名のユーザを示す．U1 と U3 を除き，有意な正の相関が見られた．$^{*}p < .005$, $^{**}p < .001$

	相関係数	p 値		相関係数	p 値
U1	-	-	U7	0.64	**
U2	0.62	**	U8	0.55	**
U3	-0.09	0.70	U9	0.38	**
U4	0.55	**	U10	0.50	**
U5	0.33	*	U11	0.64	**
U6	0.30	*			

4.1.2　職場単位幸福度推定

職場単位幸福度推定では，職場単位で幸福度を推定し，相関係数を求めて評価する．その際，図 3 のように学習データとテストデータを分け，二つの実験を行う．それぞれ，学習データとテストデータのどちらにも同じ人によるテキストが存在する混合データ実験と，学習データに含まれるテキストを書いた人とテストデータに含まれるテキストを書いた人が異なる 2 社別データ実験とする．

混合データ実験　混合データ実験では，混合された 2 社のデータをランダムに 8:2 に分割し，それぞれを学習データ（1,904 件）とテストデータ（476 件）として用いる．この際，同じユーザによる別の日の日報テキストが学習データとテストデータに含まれる可能性がある．本実験で想定している設定として，特定の集団に対して幸福度推定モデルを構築し，そのモデルを同一の集団に適用させ続ける場合などが挙げられる．実験においては 5 分割交差検証を行い，平均値をモデルの性能として評価に使用する．

2 社別データ実験　2 社別データ実験では，特定の職場で学習したモデルを他の職場のデータに適用できるかを調べる．H 社のデータ 1,728 件を学習データ，P 社のデータ 652 件をテストデータとして使用する．この際，H 社に所属する個人と P 社に所属する個人は全く異なることから，構築したモデルは特定の集団における常識を他の集団に対して利用することができない．本実験において想定している設定としては，構築したモデルを全く別の集団に対して適用する場合が挙げられる．

4.2　結果と考察

4.2.1　ユーザ単位幸福度推定

相関係数が 0.4 以上のユーザは 11 人中 6 人と，過半数にとっ

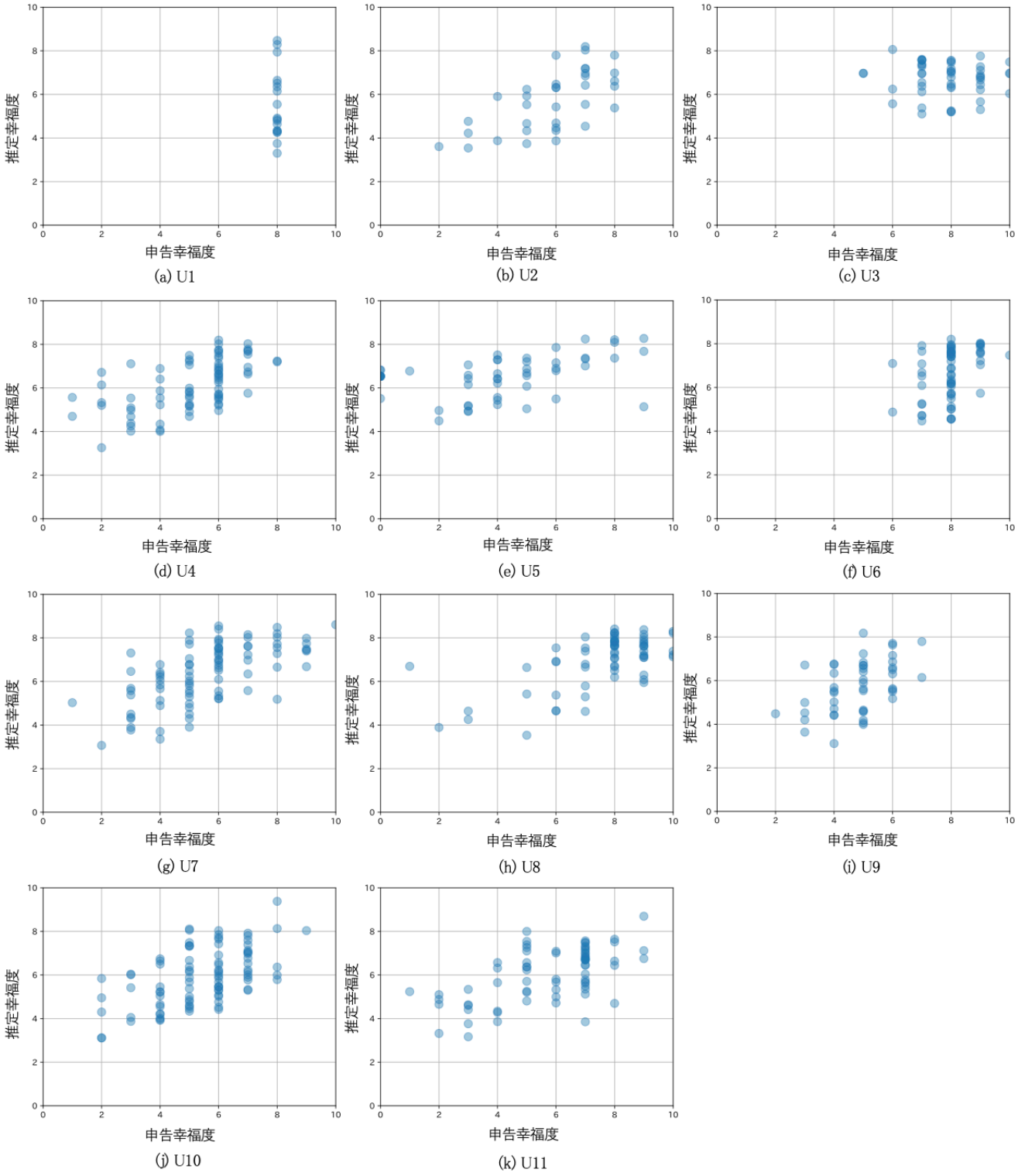


図 4　P 社ユーザ単位幸福度推定

てモデルが機能しているという結果になった．P 社のユーザ単位での幸福度推定の相関係数を表 2，散布図を図 4，時系列変化を図 5 に示す．相関係数の違いは，幸福度スコアの上昇と下降の時系列の変化を推定モデルが捉えているかどうかの性能の違いと解釈できる．表 2 において，ユーザごとの幸福度の予測性能にはばらつきがあることがわかる．

特徴的な個人の結果について述べる．まず，表 2 の U1 の相関係数が計算不可となっている．これは，U1 の申告幸福度は常に 8 であり（図 4(a)），相関係数を算出できなかった．我々は，この例外的な現象が，提案手法の優位性を示す根拠にもなると考えている．質問紙では同じ数字しか回答しない不真面目な参加者からも，自由記述テキストからは，毎日異なる情報を得ることができるからである．

次に，有意ではないものの相関係数が負になっているユーザ

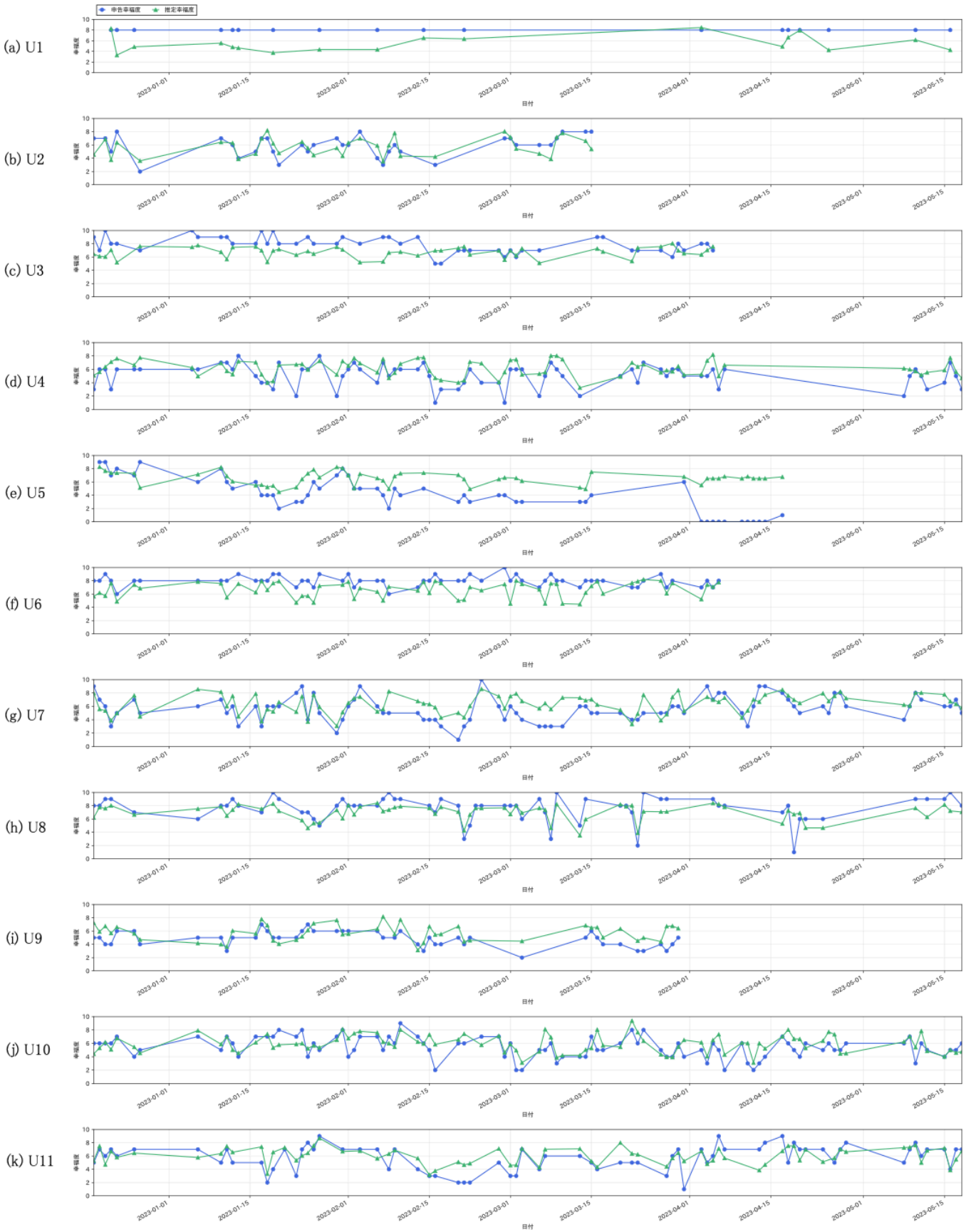


図 5　時系列での P 社ユーザ単位幸福度推定

(U3) がいる．どのような申告幸福度のテキストであったとしても，推定幸福度のばらつきも大きくなく，モデルがうまく推定できていない様子がわかる．このユーザの日報を調査したところ，「〜な一日。」という表現を多用していた．このことから，ユーザごとの文体の違いや，文章の癖などにより，推定モデルの性能が左右されている可能性がある．

図 5 において，4 月以降にずっと 0 のみを回答しているユーザ (U5) が存在した．それまでは様々なテキストを入力していたにもかかわらず，4 月以降には 0 を多くつけ，その間のテキストを全て同じものとしていた．このように環境の変化の後に，測定が難しくなってしまうユーザも存在する．イベントなどの発生によりそもそもデータを収集することが難しくなってしまう可能性があることがわかる．

最後に，学習データにない表現を多用することも予測性能に影響を与えていると考えられる．本実験においては H 社のデータを学習データとし，幸福度推定モデルを構築している．H 社特有の言い回しを学習した幸福度推定モデルは P 社特有の言い回しには対応することができない．一方で，そういった影響を鑑みてもユーザによっては相関係数が 0.64 であったりと，ユーザ単位で見ればかなり高い性能となる場合もある．幸福度に影響を与える要因が人によって違うということも予測性能に影響を与えている可能性がある．学習データでは幸福度スコアに影響を与えていなかった要因が，テストデータにおけるユーザにとっては幸福につながる場合がある．このことから，個人の性格を特徴量として組み込むことで性能がさらに高くなる可能性がある．

4.2.2 職場単位幸福度推定

2 つの実験の結果，ともに相関係数は 0.4 以上であったことから，モデルが機能していることがわかった．混合データ実験の結果，相関係数は 0.68 であった（図 6(a)）．また，2 社別データ実験の結果，相関係数は 0.49 であった (図 6(b))．

大規模言語モデルは，LIWC のような辞書ベースの手法よりも，性能が高いことがわかった．

2 社別データ実験結果から，職場が異なっていても推定モデルがうまく機能していると言える．この理由としては．ポジティブ，またはネガティブな表現が 2 社である程度一致しているからと考えられる．既報 [20] では，ポジティブな感情のみ，ネガティブな感情のみをスコア化し申告幸福度との相関係数を算出しているが，本研究の推定モデルの方が相関係数が高かった．このことからも，今回作成した推定モデルは 2 社に共通するネガティブな表現もうまく学習していると考えられる．

また，収集したデータの偏りが大きかったことによる影響が考えられる．図 2 と図 6 からも分かるように，元々データ数が少ない 0〜2 点と 9〜10 点は，BERT にとって学習が困難であった可能性がある．

4.2.3 今後の課題

実験からは，自由記述のテキストから，職場単位でもユーザ単位でも幸福度を推定できることが示された．一方で，従来の質問紙の代替とするには，大きく四つの問題がある．

まず一つ目に，本タスクがそもそも不良設定問題であり，テキストから幸福度を推定できない場合がある．例えば，「今日は祝日.」という記述があった際に，祝日であることに喜びを感じる人もいれば，祝日であることに苦痛を覚える人もいる．本来，祝日と幸福度の関係を知るためには，十分なテキストによって，筆者の性格や近況といった多くの情報が必要となるはずである．本研究の設定では，1 日単位の日報しか入力としておらず，数行のテキストからではそもそも人間でも幸福度を推定できない事例も多い．このような問題はあるものの，図 5 からもわかる通り，おおよその幸福度の推移を見るといった場合や，リアルタイムに大規模な調査をする際には自由記述テキストによる幸福度推定は簡便性があり，有用な場合もあると考えられる．

二つ目の問題は，モデルの汎用性である．今回のモデルは 2 社で実施された実験においてデータを収集したが，汎用性を持つのか確認する必要がある．本研究の対象は，広告・マーケティング事業を行う H 社と電機メーカーである P 社で，多くの業態からのわずか 2 つのケーススタディにすぎない．他業態，他社でのデータ収集も行う必要がある．

三つ目に，本研究におけるデータ収集におけるバイアスである．今回は日報といった形で 2 社に協力してもらい自由記述テキストと幸福度を収集した．両方のデータを同時に取っており，幸福度と自由記述テキストに矛盾がないように記述させてしまう可能性がある．本来は，個人的な日記など，実験であることを意識しないテキストを用いるのが望ましい．よって，本研究は，職場日報における幸福度推定モデルとして使用は可能であると考えるが，職場以外の幸福度の推定については未知であり，今後の研究を発展させたい．

四つ目に，脱落の対策の必要性である．本実験は，質問紙よりも頻繁に調査できるため自由記述テキストを使用したが，頻繁に調査することが負担になってしまう問題が存在する．実際に，今回の結果においても，20 日以上記述をやめたことのあるユーザは 11 人中 7 人だった．このことからも，日記を書いてもらう際に脱落しにくくなるような工夫を考えていく必要がある．

5 おわりに

本稿では，従来の幸福度調査方法に対する改善点を指摘し，質問紙に依存せずテキストから幸福度を推定する方法を提案した．大規模なテキストデータと幸福度の評価を用いて，言語モデルに基づく推定モデルを構築し，その性能を評価した．これにより，日常的なテキストからの幸福度推定手法の有用性が示唆された．

展望として，まず一つ目に介入実験といった形で，幸福度の向上に働きかけることが考えられる．本研究の結果より，自由記述テキストから幸福度推定が可能であることが確認できた．つまり，幸福度の変化があった際に自由記述テキストからどういった出来事があったのかを知ることが可能になる．このことから，質問紙よりも自由記述テキストの方が，介入方法を考えることは容易であると考える．

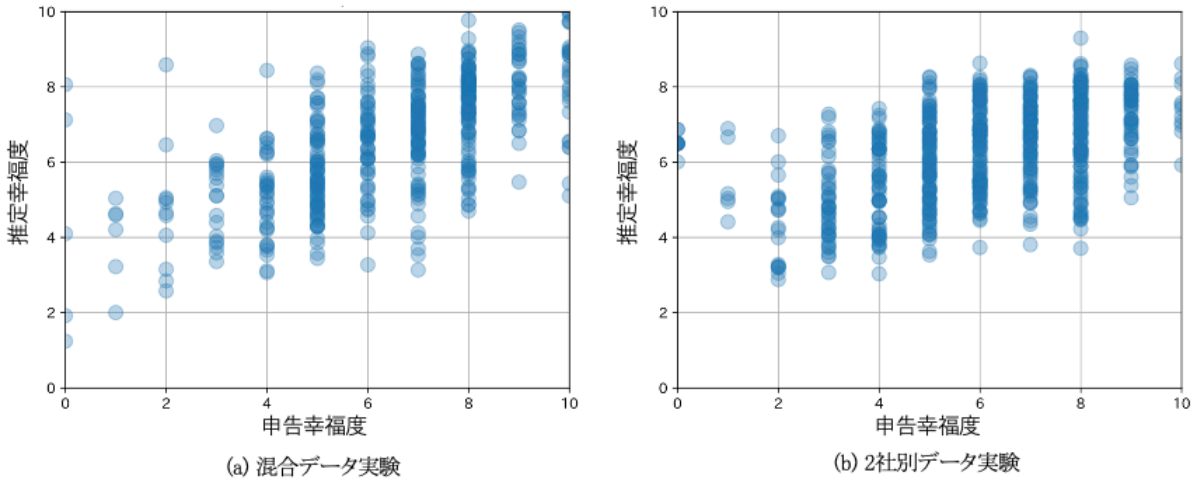


図 6　職場単位幸福度推定実験結果．(a) 混合データ実験の相関係数は 0.68 ($p < 0.001$)，(b) 2 社別データ実験の相関係数は 0.49 ($p < 0.001$)

二つ目に，本研究においてテキストから幸福度を推定した．しかし自由記述からは幸福度以外の情報を得ることも可能である．具体的には，テキストを見ることで何が Well-being に寄与したのか，書いた本人に何が起こっていて何を感じたのかを知ることが可能である．また，自由記述の特徴を捉えることによって，幸福度だけではなく他のさまざまな Well-being を構成する概念を捉えることができるようになる可能性がある．そうすれば今後，要因だけでなく様々な概念がより大規模に計測可能になると考えられる．

謝　　辞

本研究は，JST，未来社会創造事業，JPMJMI21J2 の支援を受けたものである．

文　　献

[1] 公益社団法人日本 WHO 協会. 世界保健機関（who）憲章とは, 2024. Accessed on Jan 5, 2024.

[2] Ed Diener and Robert Biswas-Diener. *Social Indicators Research*, Vol. 57, No. 2, p. 119–169, 2002.

[3] Yuyan Shi, Lindsay E. Sears, Carter R. Coberley, and James E. Pope. Classification of individual well-being scores for the determination of adverse health and productivity outcomes in employee populations. *Population Health Management*, Vol. 16, No. 2, p. 90–98, April 2013.

[4] Gregory M. Hurtz and John J. Donovan. Personality and job performance: The big five revisited. *Journal of Applied Psychology*, Vol. 85, No. 6, p. 869–879, 2000.

[5] Aon plc. 2022-2023 global wellbeing survey, 2024. Accessed on Jan 14, 2024.

[6] Ed Diener, Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. The satisfaction with life scale. *Journal of Personality Assessment*, Vol. 49, No. 1, p. 71–75, February 1985.

[7] Viktoria Gochmann, Sandra Ohly, and Silja Kotte. Diary studies, a double ‐ edged sword? an experimental exploration of possible distortions due to daily reporting of social interactions. *Journal of Organizational Behavior*, Vol. 43, No. 7, p. 1209–1223, May 2022.

[8] H. ANDREW SCHWARTZ, MAARTEN SAP, MARGARET L. KERN, JOHANNES C. EICHSTAEDT, ADAM KAPELNER, MEGHA AGRAWAL, EDUARDO BLANCO, LUKASZ DZIURZYNSKI, GREGORY PARK, DAVID STILLWELL, MICHAL KOSINSKI, MARTIN E.P. SELIGMAN, and LYLE H. UNGAR. Predicting individual well-being through the language of social media. In *Biocomputing 2016*. WORLD SCIENTIFIC, November 2015.

[9] Kokil Jaidka, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, Vol. 117, No. 19, p. 10165–10171, April 2020.

[10] Mengyao Song and Nan Zhao. Predicting life satisfaction based on the emotion words in self-statement texts. *Frontiers in Psychiatry*, 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

[12] Philip J. Cooke, Timothy P. Melchert, and Korey Connor. Measuring well-being: A review of instruments. *The Counseling Psychologist*, Vol. 44, No. 5, p. 730–757, July 2016.

[13] Ed Diener. Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, Vol. 55, No. 1, p. 34–43, 2000.

[14] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, Vol. 54, No. 6, p. 1063–1070, 1988.

[15] Carol D. Ryff and Corey Lee M. Keyes. The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, Vol. 69, No. 4, p. 719–727, 1995.

[16] Seligman Martin E P. *Flourish: A visionary new understanding of happiness and well-being*. Free Press, 2011.

[17] 金沢工業大学心理科学研究所. Perma-profiler, 2024. Accessed on Jan 9, 2024.

[18] Jiaqi Wu, Marilyn Walker, Pranav Anand, and Steve Whittaker. Linguistic reflexes of well-being and happiness in echo. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2017.

[19] Oscar N. E. Kjell, Sverker Sikström, Katarina Kjell, and H. Andrew Schwartz. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, Vol. 12, No. 1, March 2022.

[20] 伊藤和浩, 若宮翔子, 眞鍋雅恵, 渡邉寧, 中山真孝, 内田由紀子, 荒牧英治. 場の状態の共有と個人のウェルビーイング. 第 15 回データ工学と情報マネジメントに関するフォーラム／第 21 回日本

データベース学会年次大会 (DEIM2023), 岐阜, 2023.
[21] Hadley Cantril. *The Pattern of Human Concerns*. Rutgers University Press, 1965.
[22] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, 2017.
[23] D.L. Streiner and G.R. Norman. *Health measurement scales. A practical guide to their development and use. 2nd Edition*. Oxford University Press, Oxford, 1989.

COVID-19流行期における大学生のX上の投稿と主観的幸福感の関係：2年間の調査と言語特徴分析による検討

蔡　丹萌† 若林　啓†† 叶　少瑜††

†††筑波大学　〒305–8550 茨城県つくば市春日 1-2
E-mail: †s2226089@s.tsukuba.ac.jp, ††{kwkaba,shaoyu}@slis.tsukuba.ac.jp

あらまし　本研究は，COVID-19 流行期においてX（旧称 Twitter）を使用する日本の大学生らの投稿に関する言語特徴と主観的幸福感の関係を明らかにすることを目的としている．2020 年 5 月から 2022 年 5 月までに公開された日本語の投稿（旧称ツイート）から抽出された名詞の出現量を独立変数，2021 年 5 月と 2022 年 5 月に行われたそれぞれの調査で得られた主観的幸福感を従属変数として回帰分析を行った．回帰分析の結果に基づいて，主観的幸福感に影響する言語特徴をワードクラウドによって可視化した．可視化した結果から，2021 年ではオンライン学習環境に関連する言語特徴の出現量が増加するほど，主観的幸福感が低下したことが明らかになったが，2022 年では同様の効果が見られなかった．2021 年と 2022 年の主観的幸福感の得点の平均はほぼ同じであった．この結果は，主観的幸福感に影響する言語特徴の一部が時期的に変化することを示している．また，主観的幸福感の向上に効果を持つ言語特徴よりも，低下に効果を持つ言語特徴の方が多いことが明らかになった．

キーワード　主観的幸福感，COVID-19，ワードクラウド分析，言語特徴

1 はじめに

日本では COVID-19 流行期の始まりとなった 2020 年 1 月以来，人々は生活において多くの変化を経験した．例えば，大学生にとって殆どの授業がオンラインに切り替えられ，課外活動も中断させられたため，キャンパスライフに大きな変化をもたらした．一般的に，人々は対面によるコミュニケーションを通して，良好な人間関係を構築するとともに，精神的健康を維持すると言われている．しかし，緊急事態宣言などの行動規制の実施により，対面での交流が難しくなったため，精神的健康が損なわれる可能性が広く懸念されている．

これまで，COVID-19 流行期におけるライフスタイルの変化と主観的幸福感を調べた研究がいくつかある（Brindal ら [1], Reyes-Molina ら [2], Briguglio ら [3]）．日本の大学生を対象とした叶ら [4] の研究では，COVID-19 が流行する前と比較し，X では否定的な表現が多くなっていることがわかった．これまでの研究より，自伝的文章と主観的幸福感との関連性が確認されていること（Schwartz ら [5], Luhmann [6], Mitchell ら [7]）を踏まえると，否定的な表現が増加する言語パターンの変化は，日本の大学生の主観的幸福感に関する変化の状況を示す指標となる可能性があると考えられる．しかし，具体的にどのような言語特徴が主観的幸福感の最も有効的な指標となりうるかは，まだ明らかにされていない．

本研究では上記の知見をさらに発展させるため，COVID-19 流行期における日本の大学生の主観的幸福感に関連する X 上の言語特徴を明らかにすることを目的とする．具体的には，2021 年 5 月と 2022 年 5 月に行った調査で得られた主観的幸福感の得点と，調査期間に対応した期間の X 上の公開投稿における言語特徴を組み合わせて分析した．その期間内の社会状況の変化が大きかったため，2 つの調査時期のデータを区別して分析を行った．第 1 期 (2020 年 5 月〜2021 年 5 月) では，COVID-19 の蔓延を食い止めるための緊急対策の実施によってもたらされたキャンパスライフの大きな変化を捉えた．一方，第 2 期（2021 年 5 月〜2022 年 5 月）では，ワクチン接種の普及に伴って規制が徐々に緩和され，対面活動が比較的容易になり，様々な活動が再開するようになった．このような状況の変化に応じて，本研究では，言語特徴から大学生の X におけるコミュニケーション行動がいかに変化していたかについても検討した．

本研究では，日本の大学生の主観的幸福感と言語特徴との関係を調べるために，オープン・ボキャブラリ分析法 (open-vocabulary analysis methods) を採用した．まず，言語特徴として日本語の投稿から名詞を抽出した．次に，これらの特徴を，主観的幸福感の得点と調査期間をそれぞれ従属変数とする 2 つの別々の回帰分析に組み込んだ．最後に，回帰係数の強さや期間ごとの傾向を可視化するために，ワードクラウドを使用した．可視化した言語の特徴を通して，1）調査期間中の社会情勢による変化を明らかにし，2）どのような社会情勢の変化が大学生の主観的幸福と関連するかを明らかにする．

2 関連研究

2.1 X 上の表現とユーザのソーシャルネットワーク構築，性格特性，メンタルヘルスとの関係

先行研究では，X 上での感情表現や投稿動機が，ユーザのソーシャルネットワーク構造とどのように関連しているかが調査された．たとえば，北村ら [8] は，社会的報酬動機とネガティブ感情語数，特に不安感情語数との間に有意な負の関係がある

ことを示した．それだけでなく，社会的報酬動機とポジティブ感情語の数との間には有意な正の関係があり，ソーシャルネットワークの規模が大きくなるにつれてポジティブ感情語の数が増加することを示した．これらの結果は，X 上の投稿内容が，ユーザの動機やソーシャルネットワークの構造と関連していることを示唆している．

Mori&Haruno [9] は，X 上のソーシャル・ネットワーク情報が利用者の性格特性の予測因子として機能することを発見した．また，彼らの研究は，言語統計情報と X 上の言語情報が，その人の精神的健康状態を推定するために使用できることを明らかにした．

北村ら [8] と Mori&Haruno [9] の研究を踏まえ，Ye ら [10] は，大学生の一般的信頼感，社会的スキル，ツイート数，感情表現の種類や話題，主観的幸福感との関係を検討した．その結果，以下のことを明らかにした：(a) 一般的信頼感と社会的スキルの高いユーザは主観的幸福感が高く，ネガティブな表現が少なかった．(b) ツイート数の多いユーザはポジティブな表現もネガティブな表現も使うが，ポジティブな表現よりもネガティブな表現が多かった．ただし，この研究はサンプル数が少ないため，これらの要因間の相関関係が示されたにとどまった．さらに，この研究の知見は COVID-19 流行前に収集されたデータに基づいたものであり，COVID-19 流行期に観察されたパターンとは異なる可能性があると考えられ，さらなる検討が必要となった．

さらに，Ye ら [11] は，大学生の性格特性と主観的幸福感の関係について，自己呈示欲やオンラインコミュニケーションスキル，ツイートにおける感情表現の影響などを調査する研究を行った．彼女らは，2019 年 1 月から 2021 年 6 月までの X のログデータを分析し，同時に利用する他の SNS との組み合わせによって異なるグループを比較した．その結果，調査期間中，ツイート数とリツイート数は全体的に増加し，肯定的な文の割合はわずかに減少したが，否定的な文の割合はわずかに増加したことが示された．また，他の SNS と併用している学生と比較すると，X のみを利用している学生の主観的幸福感が最も低く，自己呈示欲が主観的幸福感を低下させている要因の 1 つであることが分かった．なお，この研究は，COVID-19 禍における利用者の感情表現が変化したことを示唆しているが，これらの変化を反映する具体的な言語特徴については言及していない．

2.2 量的テキスト分析による言語特徴の探索

量的テキスト分析は，心理学者が人間の行動を研究するための方法論として，近年注目を集めている分野である．計算機を使って大量のテキストデータを分析することにより，人々の思考や感情，経験について，従来の方法では得ることが難しかった洞察を得ることができる．量的テキスト分析には，主にクローズド・ボキャブラリ分析法とオープン・ボキャブラリ分析法の 2 つのアプローチがある．

2.2.1 クローズド・ボキャブラリ分析法

クローズド・ボキャブラリ分析法では，語彙リストと呼ばれる，あらかじめ定義された単語やフレーズのリストを使用し，テキスト中の関連する特徴を特定する．語彙リストは通常，心理学者や言語学者など，研究分野の専門家によって作成される．クローズド・ボキャブラリ分析を使ってテキストを分析するには，まずテキストをトークン化する．次に，その単語やフレーズを定義された語彙リストと比較する．単語やフレーズが語彙リストの中に見つかれば，それが関連特徴としてカウントされる．そして，関連特徴の数を使って，テキストが表す感情を判断する．例えば，幸福という感情に関連する語彙リストには，「happy」,「glad」,「joyful」,「content」,「cheerful」といった単語が含まれる．幸福に関連する単語を多く含む文章は，幸福な感情を表す文章であると考えられる．クローズド・ボキャブラリ分析法を用いた研究で，性格特性が言語的特徴と有意に相関していることが示されている（Hirsh& Peterson [12]， Lee ら [13]）．

2.2.2 オープン・ボキャブラリ分析法

クローズド・ボキャブラリ分析法には，分析結果が語彙リストの質や量に依存するという問題がある．例えば，語彙リストが包括的でない場合，テキスト内の関連する特徴をすべて特定できない可能性がある（Yarkoni [14]）．一方，オープン・ボキャブラリ分析法は，事前に定義された語彙リストに依存しないというメリットがあることから，より柔軟で，幅広い種類の分析に使用できる．

Latent Dirichlet Allocation（LDA）は，心理学研究でオープン・ボキャブラリ分析法として広く使われている手法である．LDA は，確率モデルに基づいて，文書内の単語をトピックの組み合わせに割り当て，これらの割り当てを反復的に改良することで，テキストをトピックに分類する（Choi& Seo [15]，Chen ら [16]，Blei ら [17]）．また，Word2Vec や GloVe などの単語埋め込みは，ニューラルネットワークや統計的手法を利用して単語を高次元ベクトルに変換し，共起パターンだけでなく，反意語や同義語などの複雑な言語関係も捉えることができる（Mikolov ら [18]，Pennington ら [19]）．これらの概念をさらに発展させた Bidirectional Encoder Representations from Transformers（BERT）モデルは，文脈のニュアンスを考慮することで，主観的幸福感の予測などの応用で有効性を示している（Devlin ら [20]，Kjell ら [21]）．

Differential Language Analysis (DLA) は，言語特徴が，性格特性などとどのように関連しているかを調べるために考案された手法である．言語特徴を低次元の特徴に還元する LDA や単語埋め込みなどの他の手法とは異なり，DLA は単語やフレーズと外部変数との関係の分析に焦点がある（Eichstaedt ら [22]）．Schwartz ら [23] は，DLA を応用して，Facebook の近況アップデートに書かれたメッセージから，ユーザの性格特性，年齢層，性別に最も関連する単語を特定した．この研究は，DLA が，クローズド・ボキャブラリ分析法では見落とされがちな口語体やインターネットスラングを含む，幅広い言語要素を捉えることができ，言語特徴と言語外の属性との関係を発見できることを示している．本研究では，言語特徴とユーザ属性の関係だけでなく，言語特徴の時間的側面，具体的には，その単語が使用された期間によって，単語の使用状況がどのように異なるかを調べる手法を提案する．

3 研究手法とデータの収集

本研究では上記の目的を達成するために，ウェブ調査と特徴分析を行った，具体的にはウェブ調査から得られた参加者の主観的幸福感の得点と，X における公開投稿に関するデータを組み合わせて分析し，どのような言語特徴が日本の大学生の主観的幸福感と最も関係するかを分析した．

3.1 調　　査

2021 年と 2022 年の 2 回，ウェブ調査を実施した．1 回目は 2021 年 5 月 10 日〜22 日，2 回目は 2022 年 5 月 9 日〜23 日に渡って，主に関東地域の大学に在籍する大学生を対象に行った．ただし，2022 年の参加者には，2022 年 3 月に卒業した者も含まれていた．いずれの調査においても，個人属性とともに，伊藤ら [24] の 15 項目を 5 段階リッカート尺度を用いて，主観的幸福感を測定した．2021 年の回答者数は 1,681 人で，2022 年は 1,292 人であり，これらの調査結果の一部はすでに Ho & Ye [25] で報告されている．

3.2 投 稿 分 析

3.2.1 投稿の収集

調査の参加者には，使用している X アカウントを 3 つまで提供してもらい，ログデータの分析を行った．各アカウントについて，Twitter API を利用して公開投稿を収集した．投稿データは，2 つの調査時期に対応する 2 つの期間に分類した．2020 年 5 月 7 日から 2021 年 5 月 7 日までの投稿を「2021 年期」，2021 年 5 月 8 日から 2022 年 5 月 8 日までの投稿を「2022 年期」とした．これらの期間外の投稿は分析対象から除外した．異なる年に使用された各 X アカウントは，本研究の目的上，別個のものとして扱った．各年期ごとに，調査参加者に提供してもらった複数のアカウントから投稿されたすべての投稿を収集し，その年期における当該調査参加者の投稿集合とした．これらの投稿集合の言語特徴を，その年期と対応する年の調査で得られた主観的幸福感スコアと対応させて分析を行った．これにより，2021 年の 545 サンプルと 2022 年の 535 サンプルの合計 1,080 サンプルが収集された．両調査に参加した参加者がいるが，本研究の分析では独立した別個のサンプルとして扱った．

3.2.2 投稿に対する前処理

まず，投稿からメンション，URL，絵文字，句読点，特殊文字，数字を取り除いた．次に，日本語の名詞を言語特徴として抽出し，分析を行った．1,080 のサンプルについて，各単語の頻度を計算した．あるサンプル x（特定の調査の参加者）の投稿における単語 w の頻度を $n_{x,w}$ とする．

$n_{x,w} > 0$ となるサンプル数 x が，全サンプル数の 1%（10.8）未満であるような単語 w は，低頻度単語として削除した．また，単語 w の総頻度 $\sum_x n_{x,w}$ の分布の第 1 四分位数（Q1），第 3 四分位数（Q3），四分位範囲間（IQR）を計算し，Q1 または Q3 を IQR の 1.5 倍以上超えている単語を，極端な頻度を持つ単語として分析から除外した．最後に，分散を安定化させるために，$n_{x,w}$ に対して Anscombe 変換を適用した (Schwartz ら [23])．変換された値を $ans(n_{x,w})$ とした．

3.2.3 Differential Language Analysis

主観的幸福スコアに対する個々の単語の関連性を定量化するために，Schwartz ら [23] が提案した方法を採用した．各単語 w について，最小二乗法（ordinary least squares; OLS）に基づく単回帰分析を行った．サンプル x について，独立変数は $ans(n_{x,w})$ であり，従属変数は x の主観的幸福感の得点であった．回帰係数の p 値が 0.05 未満の単語を，主観的幸福感の得点に有意な効果を持つ単語として抽出した．これらの単語の回帰係数は，後述で説明する可視化に使用した．

また，単語使用の調査期間への影響も分析した．本研究のサンプルは，2021 年期または 2022 年期のものなので，独立変数を $ans(n_{x,w})$，従属変数を x のバイナリ期間変数 y とする OLS 回帰を適用し，$p < 0.05$ の回帰係数を持つ単語を抽出した．ここで，サンプル x が 2021 年期のものであれば $y = 0$ とし，x が 2022 年期のものであれば $y = 1$ とした．このため，正の回帰係数は，その単語が 2022 年期に関連性を持つことを示し，負の回帰係数は，2021 年期に関連性を持つことを示す．

3.2.4 ワードクラウドによる可視化

多次元的な結果を効果的に表すため，可視化手法としてワードクラウドを採用した．通常，ワードクラウドは，文字サイズを用いて，単語に関する一つの特徴（多くの場合，出現頻度）を可視化することに用いられる．一方で，我々のワードクラウドは，文字サイズ，文字色の色相，文字色の明度，背景色の 4 つの表現を用いて，複数の異なる特徴を可視化することにした．表現対象の単語の特徴としては，主観的幸福感に対する回帰係数の絶対値および正負符号，調査期間変数に対する回帰係数の絶対値および正負符号，および出現頻度を用いた．これらの特徴のうち，いくつかの組み合わせを選択し，以下の 3 つのカテゴリーのワードクラウドを作成した．

（1） **主観的幸福感との関連ワードクラウド：** 文字サイズで，単語の主観的幸福感の得点に対する回帰係数の絶対値を表現する．また，文字色の明度によって，単語頻度を表現する．

（2） **調査期間との関連ワードクラウド：** 文字サイズで，単語の調査期間変数に対する回帰係数の絶対値を表現した．また，文字色の明度によって，単語頻度を表現した．

（3） **ハイブリッドワードクラウド：** 文字サイズで主観的幸福感の得点に対する回帰係数の絶対値を，文字色の色相で調査期間変数に対する回帰係数の正負符号を，明度でその絶対値を表現する．主観的幸福感と調査期間変数の役割を入れ替えたワードクラウドも作成した．

4 結　　果

4.1 分析対象者に関する情報

表 1 に分析対象者に関する情報を示す．生活スタイルについては，いずれも一人暮らしが多く，2021 年期では 71.96 %，2022 年期では 72.90 %であった．これはコロナ前とほぼ同様であった．主観的幸福感の平均得点は，2021 年は 48.05 であり 2022 年は 48.91 であり，ほとんど変化が見られなかった．

表 1　調査参加者に関する情報

項目		2021 年期 (n=545)	2022 年期 (n=535)
性別	男性：	52.03%	52.71%
	女性：	46.68%	45.79
	その他：	1.29%	1.50%
平均年齢：		19.63	20.06
学年	1 年生：	32.29%	21.31%
	2 年生：	28.23%	25.42%
	3 年生：	21.96%	23.74%
	4 年生以上：	17.52%	29.53%
居住状況：	宿舎：	19.00%	18.05%
	実家や親族の家：	20.48%	20.00%
	アパート：	58.49%	60.00%
	シェアハウス：	2.03%	1.50%
居住形態：	一人暮らし：	71.96%	72.90%
	家族や親族と同居：	22.32%	20.75%
	友人や恋人と同居：	2.21%	4.30%
	シェアハウス：	3.59%	2.06%
X におけるアカウント数 (平均)		1.57	2.02
主観的幸福感得点 (平均)		48.05	48.91

4.2　主観的幸福感との関連ワードクラウド

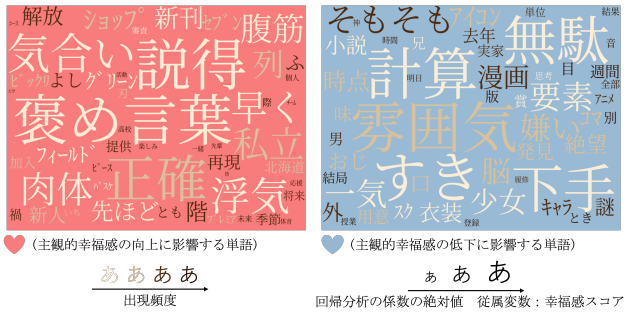


図 1　**（左）**主観的幸福感の向上に影響する頻度の高い単語．**（右）**主観的幸福感の低下に影響する頻度の高い単語．

図 1 は，文字サイズが主観的幸福感得点の回帰係数の絶対値を表すワードクラウドである．図 1 **（左）**は回帰係数が正の単語を示している．「褒め言葉」，「気合い」，「将来」，「未来」といった単語の特徴量が増加するほど，大学生の主観的幸福感が向上したことが示された．図 1 **（右）**は回帰係数が負の単語を示している．学校関連の単語（「単位」，「授業」など），徒労感や絶望感を表す単語（「無駄」，「下手」など），漫画関連の単語（「漫画」，「アニメ」など）の特徴量が増加するほど，大学生の主観的幸福感が低下したことが示された．

4.3　調査期間との関連ワードクラウド



図 2　**（左）**2022 年期に関連する頻度の高い単語．**（右）**2021 年期に関連する頻度の高い単語．

図 2 は，文字サイズが調査期間変数の回帰係数の絶対値を表すワードクラウドである．図 2 **（左）**は回帰係数が正の単語，すなわち，2022 年期（2021 年 5 月から 2022 年 5 月までの期間）において関係している単語を示している．この期間に関連していた単語として，「開始」，「ワクチン」，「選手」，「出場」，「オリンピック」，などの単語が観察されることから，COVID-19 の予防接種の展開や東京夏季オリンピック・パラリンピックの開催を反映していると思われる．図 2 **（右）**は，回帰係数が負の単語，すなわち 2021 年期（2020 年 5 月から 2021 年 5 月まで）の期間変数に対して，関係している単語を示している．この期間で顕著な単語は，「部屋」，「仲間」，「画面」，「診断」，「嫌い」，「パソコン」，「授業」，「マスク」などであった．「部屋」，「画面」，「パソコン」は，遠隔学習のために，室内やオンラインで過ごす時間が増えたことを示すと推測できる．「仲間」という言葉が出てくるのは，COVID-19 流行期に友人関係を築くことの難しさを反映していると考えられる．また，「診断」と「マスク」は，健康への懸念と現在進行中のパンデミックを示したと言える．

4.4　ハイブリッドワードクラウド



図 3　**（左）**主観的幸福感の向上に影響する頻度の高い単語．**（右）**主観的幸福感の低下に影響する頻度の高い単語．



図 4 **（左）**2022 年期に関連する頻度の高い単語．**（右）**2021 年期に関連する頻度の高い単語．

図 3 と図 4 は，有意水準 5 %で，調査期間の変数と主観的幸福感の得点の両方と統計的に有意な影響を及ぼした単語を示している．図 3 は，文字サイズが主観的幸福感の得点の回帰係数の絶対値を表し，文字色は調査期間変数の回帰係数を示した．図 4 は，主観的幸福感の得点と調査期間変数の役割を入れ替えたワードクラウドで，文字サイズは調査期間変数の回帰係数の絶対値を表し，文字色は主観的幸福感スコアを表す．図 3 と図 4 の両方から，主観的幸福感の低下に効果をもたらす単語（青色）は，向上に効果にもたらす単語（赤色）に比べて多いことが分かった．また，主観的幸福感の低下に効果をもたらす単語（図 3 **（右）**）のうち,「無職」,「遠隔」,「授業」,「履修」といった単語が，特に 2021 年期（黄色）と関連していることがわかる．これらの単語は，就職に関する不確実性，遠隔学習への移行，授業形式やスケジュールの変更による潜在的な困難など，その時期に学生が直面する特有の課題を反映していると言える．

5 考　　察

図 3 **（右）**は,「授業」,「履修」,「遠隔」の出現が，主観的幸福感に負の影響を与えるとともに，調査期間変数に 2021 年期（2020 年 5 月～2021 年 5 月）側への影響を与えることを示している．2021 年期は，日本では COVID-19 の感染拡大が発生し，遠隔学習への急激な変化が起こった時期であることから，これらの表現は，オンライン教育環境への移行に関するトピックを表していると考えられる．また，表 1 より，参加者の 7 割以上が一人暮らしであることを踏まえると，より精神的健康を維持することが困難になったと考えられる（Ho&Ye [25]）．一方，主観的幸福感が高い大学生は，このような状況に対応する必要性が生じるではあるが，公開投稿でオンライン教育のトピックに関する言葉を使用する頻度が低いことが示された．これについて 1 つの原因として挙げられるのは，主観的幸福感の高い大学生ほど，LINE や X のダイレクトメッセージ，音声通話など，親しい友人がいる場合にのみ利用できるコミュニケーションチャネルを持つ可能性がある．つまり，ソーシャル・サポート・ネットワークが充実している人ほど，X での公開投稿で悩みを語ることが少ないのに対して，幸福感の低下に関連する言語のある公開投稿をする人はソーシャル・サポートが少ない傾向にある可能性が示唆された．これは，Ye&Ho [26] が示したソーシャル・サポート・ネットワークが主観的幸福感の向上にやや効果的である知見と一致する．

本研究の分析結果より（図 3 と図 4），主観的幸福感の低下に効果を持つ単語が，向上に効果を持つ単語に比べて多いことも明らかになった．実際に，主観的幸福感の低下に効果を持つ単語は 1,040 個あったのに対し，向上に効果を持つ単語は 170 個しかないことが分かった．このことは，主観的幸福感が低い学生ほど，公開投稿で議論するトピックにおいて，より共通している可能性を示唆している．たとえば，ストレスや不安，抑うつ状態にある人は，自分の状態を表現するために共通の単語やフレーズ（例：「無駄」,「下手」,「絶望」）に頼る可能性があり，それらの単語がより頻繁に表示されるため，回帰モデルにおいて統計的に有意になる可能性が高くなる．一方，主観的幸福感の高い学生ほど，公開投稿で議論するトピックの共通性が低いことを示唆したがまだ詳細な検討が必要となり，今後の研究課題としたい．

また，図 3 **（右）**の結果から，2021 年期と 2022 年期で，それぞれ異なる言語特徴が主観的幸福感に負の影響を与えることが示された．表 1 に示したように，2021 年期と 2022 年期の主観的幸福感の得点はほとんど同じであった．これまでの研究では，性格特性に影響を与える言語特徴は時間的に一定であるという暗黙の仮定が置かれている場合が多かったが，本研究の結果から，主観的幸福感に影響を与える言語特徴の一部は時期的に変化することが明らかになった．また，ポジティブ，ネガティブの感情極性をもつ投稿量が時期的に変化するということは示されていたが (叶ら [4])，本研究では，オープン・ボキャブラリ分析法を用いることで，主観的幸福感に影響する具体的な言語特徴の一部の時期的な変化を示すことができた．この結果は，例えばソーシャルメディア上の投稿からメンタルヘルスの問題を検出する技術の開発において，その時期に発生した出来事などに依存した言語特徴を考慮する必要性を示唆している．

本研究は，2020 年 5 月から 2022 年 5 月までの COVID-19 流行期における日本の大学生の公開投稿における言語特徴と主観的幸福感の関係に注目して検討を行った．提案したワードクラウドを用いた分析により，COVID-19 流行期の異なる期間における特定の言語特徴に関する新たな示唆を発見することができた．主観的幸福感と関連する言語特徴は一定的なものではなく，時期によって変動するダイナミックなものではないかと考えられる．この結果は，ユーザの主観的幸福感を予測するための言語特徴の発見にも役立つ可能性が示された．

6 おわりに

2020 年 5 月から 2022 年 5 月までの日本の大学生による X 上での投稿の言語分析から，主観的幸福感の向上または低下に関係する可能性のある具体的な言語特徴，および，その時期的な依存性に関する知見が得られた．ワードクラウドによる可視化の結果，以下のことが示唆された： 1) 2021 年期には，オンライン学習環境に関連する単語が公開投稿に含まれるほど，主観

的幸福感が低下した．このことから，主観的幸福感に影響を与える言語特徴の一部は時期的に変化することが示された．2) 主観的幸福感の向上に効果を持つ言語特徴よりも，低下に効果を持つ言語特徴の方が多かった．今後の課題としては，本研究で得られた知見から検証可能な仮説を立て，厳密な統計的手法で検証することによって，さらなる知見を創出することが期待される．

謝　　辞

本研究は JSPS 科研費（課題番号 21H03770）の助成を受けて行ったものです．調査実施の際に，ご協力・ご回答いただいた皆様に心より感謝申し上げます．

文　　献

[1] Brindal Emily, Ryan Jillian C., Kakoschke Naomi, Golley Sinead, Zajac Ian T., and Wiggines Bonnie. Individual differences and changes in lifestyle behaviours predict decreased subjective well-being during covid-19 restrictions in an Australian sample. *Journal of Public Health*, Vol. 44, No. 2, pp. 450–456, 2022.

[2] Reyes-Molina Daniel, Alonso-Cabrera Jesús, Nazar Gabriela, Parra-Rizo Maria Antonia, Zapata-Lamana Rafael, Sanhueza-Campos Cristian, and Cigarroa Igor. Association between the physical activity behavioral profile and sedentary time with subjective well-being and mental health in Chilean university students during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, Vol. 19, No. 2107, pp. 1–15, 2022.

[3] Briguglio Marie, Caruana Mark, and Debono Nathaniel. Well-being disparities during the covid-19 outbreak: Evidence from Malta. *Traumatology*, Vol. 27, No. 4, pp. 388–398, 2021.

[4] 叶少瑜, 若林啓, 周子楷. 大学生の twitter 使用における感情表現, トピックと幸福感の関係. 電子情報通信学会誌研究報告 (信学技報), Vol. 121, pp. 37–42, 2022.

[5] Schwartz Hansen, Eichstaedt Johannes, Kern Margaret, Dziurzynski Lukasz, Lucas Richard, Agrawal Megha, Park Gregory, Lakshmikanth Shrinidhi, Jha Sneha, Seligman Martin, and Ungar Lyle. Characterizing geographic variation in well-being using tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7, No. 1, pp. 583–591, 2021.

[6] Luhmann Maike. Using big data to study subjective well-being. *Current Opinion in Behavioral Sciences*, Vol. 18, pp. 28–33, 2017.

[7] Mitchell Lewis, Frank Morgan R., Harris Kameron Decker, Dodds Peter Sheridan, and Danforth Christopher M. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, Vol. 8, No. 5, pp. 1–15, 2013.

[8] 北村智, 河井大介, 佐々木裕一. ソーシャルメディアにおける感情語使用と 投稿動機，ネットワーク構造の関係 —ツイッターでのポジティブ感情語・ネガティブ感情語に着目して—. 社会言語科学, Vol. 20, pp. 16–28, 2017.

[9] Mori Kazuma and Haruno Masahiko. Differential ability of network and natural language information on social media to predict interpersonal and mental health traits. *J Pers*, Vol. 89, pp. 228–243, 2021.

[10] Ye Shaoyu, Wakabayshi Kei, K. W. Ho Kevin, and Khan Muhammad Haseeb. *The Relationships Between Users' Negative Tweets, Topic Choices, and Subjective Well-Being in Japan*, pp. 288–300. IGI Global, 2022.

[11] Ye Shaoye, K. W. Ho Kevin, Wakabayashi Kei, and Kato Yuuki. Relationship between university students ' emotional expression on tweets and subjective well-being: Considering the effects of their self-presentation and online communication skills. *BMC Public Health*, Vol. 23, No. 594, pp. 1–12, 2023.

[12] Hirsh Jacob B. and Peterson Jordan B. Personality and language use in self-narratives. *Journal of Research in Personalit*, Vol. 43, No. 3, pp. 524–527, 2009.

[13] Lee Chang H., Kim Kyungil, Seo Young Seok, and Chung Cindy K. The relations between personality and language use. *The Journal of General Psychology*, Vol. 134, No. 4, pp. 405–413, 2007.

[14] Yarkoni Tal. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, Vol. 44, No. 3, pp. 363–373, 2010.

[15] Choi Soyoung and Seo JooYoung. An exploratory study of the research on caregiver depression: using bibliometrics and lda topic modeling. *Issues in mental health nursing*, Vol. 41, No. 7, pp. 592–601, 2020.

[16] Chen Irene Y., Szolovits Peter, and Ghassemi Marzyeh. Can ai help reduce disparities in general medical and mental health care? *AMA J Ethics*, Vol. 21, No. 2, pp. E167–179, 2019.

[17] Blei David M., Ng Andrew Y., and Jordan Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[18] Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, and Dean Jeffrey. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, p. 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[19] Pennington Jeffery, Socher Richard, and Manning Christopher. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[20] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[21] Kjell Oscar N. E., Sikström Sverker, Kjell Katarina, and Schwartz H. Andrew. Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Sci Rep*, Vol. 12, No. 3918, pp. 1–9, 2022.

[22] Eichstaedt Johannes C., Kern Margaret L., Yaden David B., Schwartz H. A., Giorgi Salvatore, Park Gregory, Hagan Courtney A., Tobolsky Victoria A., Smith Laura K., Buffone Anneke., Iwry Jonathan, Seligman Martin E. P., and Ungar Lyle H. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, Vol. 26, No. 4, pp. 398–427, 2021.

[23] Schwartz H. Andrew, Eichstaedt Johannes C., Kern Margaret L., Dziurzynski Lukasz, Ramones Stephanie M., Agrawal Megha, Shah Achal, Kosinski Michal, Stillwell David, Seligman Martin E. P., and Ungar Lyle H. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, Vol. 8, No. 9, pp. 1–16, 2013.

[24] 伊藤裕子, 相良順子, 池田政子, 川浦康至. 主観的幸福感尺度の作成 と信頼性 ・妥当性の検討. 心理学研究, Vol. 74, No. 3, pp. 276–281, 2003.

[25] K. W. Ho Kevin and Ye Shaoyu. University students ' social media usage and subjective well-being during the covid-19 pandemic – comparing similarities and differences from 2021 to 2022 –. *IEICE Tech*, Vol. 122, No. 349, pp. 18–23, 2023.

[26] Ye Shaoyu and K. W. Ho Kevin. College students' twitter usage and psychological well-being from the perspective of generalised trust: comparing changes before and during the covid-19 pandemic. *Library Hi Tech*, Vol. 41, No. 1, pp. 152–173, 2022.

孤独エピソードデータセットの構築とその応用

藤川　直也† NGUYEN Quang Toan†† 伊藤　和浩† 若宮　翔子† 荒牧　英治†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916-5
†† ESIEE Paris　Cité Descartes, 2 Bd Blaise Pascal, 93160 Noisy-le-Grand
E-mail: †{fujikawa.naoya.fi8,ito.kazuhiro.ih4,wakamiya,aramaki}@is.naist.jp, ††nqt.quangtoan@gmail.com

あらまし　孤独感は身体的および精神的健康に関する重要な問題であり，死亡リスクの増加，アルツハイマー病，うつ病のリスクファクターである．特に，コロナ禍以降は孤独感に苛まれている人が増加しているとも言われ社会的な問題になりつつある．孤独を防止するためには，早期に介入することが重要であるが，一般に，孤独を感じている人間が，相談する友人・知人を持っていない場合が多く，発見が困難である．そこで，我々は，ソーシャルメディアなどのウェブの投稿から，投稿者の孤独を推定することを目指している．ただし，孤独に関する標準的なデータセットは存在せず，また，孤独について，明確な定義も定まっていない．そこで本研究では，判断基準を明確にするために新たなアノテーションガイドラインを定め，感情から想起されたエピソードコーパスに孤独感の有無，孤独感の原因，孤独感の強度を付与する．本研究により，ウェブデータを用いた大規模な孤独分析が可能になり，孤独を感じている人々の早期発見に繋がると考えている．

キーワード　データセット，孤独感，エピソード，感情，テキスト分類

1 はじめに

今や，35 歳から 74 歳の 10 人に 1 人が孤独を感じているとされ [1]，孤独感は世界的に心身に関する大きな問題となっている．孤独感は身体的および精神的健康に影響を及ぼし，死亡リスクが増加する原因の一つであることが知られている [1], [2]．加えて，孤独感は睡眠障害 [3]，アルツハイマー病やうつ病 [4]，統合失調症 [5] とも関連し，リスクファクターになりうる．例えば，孤独感が強いグループではうつ病患者が半数以上を占め，さらに希死念慮も多く見られるとの報告 [1] や，孤独を強く感じている人はそうでない人に比べ，アルツハイマー病を発症する可能性が高い [4]，統合失調症患者の発症頻度や重篤化との相関 [5] など多くの孤独のリスクが報告されている．

さらに，コロナ禍以降，孤独はより大きな問題となっている [6], [7], [8]．例えば，日本では，コロナ禍では半数近くの人が孤独を感じていたとの報告もある [7]．孤独に陥るほど社会的交流が減少し，発見や介入が困難になるため [9]，早期に適切に介入することが重要である．

孤独に陥る人を早期に発見し，介入するために，我々はソーシャルメディアなどウェブ上のテキストから孤独の兆候を捉えることを試みている．孤独に陥って，対面での社会的な交流が減った場合でも，ソーシャルメディア上での活動は継続される場合がある．また，対面では表現できない内面の吐露が表現されることもあると考えている．このため，孤独やそれに近い状態をテキストから推定する研究は，すでにソーシャルメディア研究では活発に研究が行われている．例えば，孤独と関連する単語の調査 [10] では，“alone”, “silence” といった単語や “get away from people” などのフレーズが孤独と関連するとされ，孤独関連の投稿の抽出に利用された．ソーシャルメディア上で孤独感を表している発言の特定 [11] では，“alone” または “lonely” について頻繁に投稿しているユーザに着目している．しかし，このような事前に設定した単語やフレーズによる手法では，“Home Alone” などの関係のない単語や “寂しくはない” などの否定（ネゲーション）の文章が含まれてしまうという課題がある．

一方，機械学習を想定して，孤独感に関するデータセットを構築した研究も存在する [12], [13], [14]．適切な規模のオープンなデータセットがあれば，近年急速に進歩している自然言語処理技術により,，類似した投稿の抽出は可能であろうと思われる．しかし，日本語において公開されている孤独データセットは存在しない．さらに，心理学の理論に基づいたアノテーション基準により構築したデータセットも存在しない．

そこで，本研究では公開可能な孤独データセットを構築する．そのため，孤独感の定義を明確に示したアノテーションガイドラインを作成し，孤独感の有無と強度について，テキストにラベル付けする．我々が構築するデータセットは明確なガイドラインに準拠し，かつ CC-BY ライセンスにて公開する [1] ことで，今後の孤独関連研究の再現性を担保する．

2 関連研究

日本語において孤独感に関するデータセットを構築した研究は少ない [12], [13]．孤独体験テキストからの孤独感推定 [12] では，20 歳から 78 歳の 2,947 名に対しウェブ調査を実施し，寂しさや孤独体験に関するテキストと書き手自身の自己評価孤独感を収集した．収集したテキストに対し，書き手自身の孤独感評価とは別に，クラウドソーシングにより他者評価孤独感のアノテーションを実施した．しかし，他者評価のアノテーション

1：https://github.com/sociocom/Japanese-Loneliness-Dataset

において孤独感の明確な定義が示されておらず，読み手の主観に大きく依存してしまう可能性がある．

食事の状況と孤独感との関連を調べた研究 [13] では，テキストから食事の状況，心理状態および心理的特性を推定するため，20 歳以上の 877 名に対しクラウドソーシングを実施し，食事体験，食事に対する満足度および心理的特性に関するテキストデータを収集した．収集した心理的特性に関するデータには孤独感の尺度が含まれている．しかし，収集された孤独感はテキストに対する孤独感ではなく書き手自身が普段感じている孤独感であるため，テキストに対する孤独感は付与されていない．加えて，いずれのデータセットも公開されておらず，ガイドラインの明確化と公開可能なデータセット構築という本研究の動機となっている．

英語でのオープンソースの孤独データセットは存在している [14]．ソーシャルメディア上の投稿から孤独データセットを構築した研究では，reddit 上の 2 つの孤独感のサブレディット（r/loneliness， r/lonely）と 2 つの若年層向けのサブレディット（r/youngadults， r/college）から投稿を収集し，コーパスを作成した．コーパスは，各投稿に対し 3 名が孤独感を表しているかどうかを判定し，2 名以上が孤独と判断した投稿には，さらに期間，状況，対人関係，相互作用といった詳細情報を付与した．しかし孤独感を表しているかの判断基準が明確に定められていない．孤独感は主観的な心理状態ではあるものの，うつや自殺など深刻な状況を予防するための汎用的な検出モデル構築に向け，可能な限り客観的に判定できる明確なアノテーションガイドラインが必要だと我々は考えている．

3 デ ー タ

アノテーションするデータセットとして，LIFE STORY データ [2] [15] を用いた．LIFE STORY データは，本研究室が 2017 年から開発，拡張しているオープンソース（CC-BY ライセンス）のコーパスであり，7 つの感情（悲しい，不安，怒り，嫌悪感，信頼感，驚き，楽しい）から想起された自由記述のエピソードと回答者の年齢や性別からなる．LIFE STORY データは，トレンドや季節的な単語など時期ごとの特徴的な単語に対応することや，単語に対する感情が時期によって変化する可能性を考慮し，年に 4 回クラウドソーシングを実施しており，各回 1,000 人を対象に調査し，更新されている．本研究では，孤独感を連想しやすい感情として「悲しい」に着目し，LIFE STORY データから抽出した「悲しい」エピソードに対しアノテーションを実施した．「悲しい」エピソードの例として，「オークションで狙っていたものが買えなかった」「GW 中は金欠気味だったので友人の誘いを断っていたのが悲しかった」などがある．

2：https://sociocom.naist.jp/life-story-data/

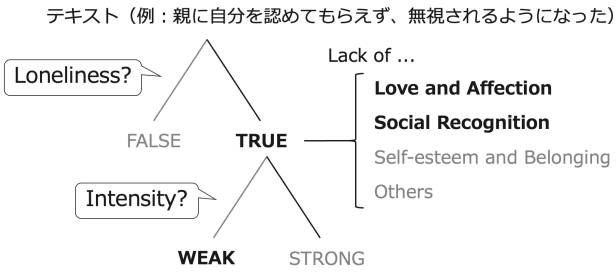


図 1　アノテーション概要

4 アノテーション

4.1 ラベルの種類

孤独感の有無 (Loneliness) と孤独感の強度 (Intensity) の 2 段階でアノテーションを実施した．孤独感の有無をラベル付けする際には，その原因となる満たされていない社会的欲求についてもラベルを付与した．

図 1 にアノテーションの概要を示す．「親に自分を認めてもらえず、無視されるようになった」という例であれば，後述する基準により Love and Affection（愛情）と Social Recognition（社会的認知）が満たされていないことが原因で孤独感の有無は TRUE であり，さらに孤独感の強度は WEAK とラベル付けされる．

4.1.1 孤独感の有無 (Loneliness)

孤独感の有無のアノテーションでは，孤独感がある場合は TRUE，ない場合は FALSE のラベルを付与する．本研究における孤独感の定義は，「社会関係の量や，特に質によって社会的欲求が満たされていないという認知を伴う嫌悪感」[16] である．この定義に基づき，社会的欲求が満たされているかどうかにより，テキストが孤独感を表しているかどうかをラベル付けする．

TRUE と付与された場合，満たされていない社会的欲求についてもラベルを付与した．心理学では，社会的欲求は「Affection」「Behavioral Confirmation」「Status」により構成されるという理論が提案されている [17]．本研究では孤独感と関連するような社会的欲求にするためこれらを一部修正し，「Love and Affection（愛情）」「Social Recognition（社会的認知）」「Self-esteem and Belonging（自尊心と帰属意識）」の 3 種類に分類した．Affection はアノテータが理解しやすいように Love を追加し Love and Affection とした．Behavioral Confirmation は社会的規範を遵守しているという感覚のことであり，明瞭なアノテーションを目的として二つに分割した．一つは外的な評価である Social Recognition（社会的認知），もう一つは内的な評価である Self-esteem and Belonging（自尊心と帰属意識）である．外的な評価は「〜と言われた」，内的な評価は「〜と感じる」などが該当する．Status は社会的地位のことであり，孤独感とは関連しないと考え除外した．本研究で用いる社会的欲求の説明を以下に示す：

Love and Affection（愛情）　自分が大切に思っている人（友達や家族など）から愛情を受けること．

例.「親と意見が合わず，音信不通になってしまった」

Social Recognition（社会的認知） 自分の行動によって社会環境に受け入れられること.

例.「学校の友達はおしゃれな人が多いけど，私はいろんな服を買うお金がなくておしゃれできないから，話してくれない」

Self-esteem and Belonging（自尊心と帰属意識） 自分が他者に受け入れられている，社会に属している感覚.

例.「博士号を持っていないのは自分だけなので，居場所がないように感じる」

孤独感を TRUE とする条件として，上記いずれかの社会的欲求の不足と定めた．その他の基準として，書き手自身が孤独を感じているのではない場合や，社会的なつながりを表す言葉が明示されていない場合は FALSE とした．テキストの書き手自身が孤独を感じていない例として「妻が職場で馴染めず、ひとりぼっちだと言っていた」があり，書き手自身が孤独を感じている例として「友人に飲み会誘ってもらえなかった」がある．社会的なつながりを表す言葉が明示されていない例として「昔の悲しい出来事を夢で見て、当時の事が思い出されて悲しさが蘇ってきた。」があり，明示されている例として「親と意見が合わず、音信不通になってしまった」や「自分の意見が伝わらず悲しい」がある.「親」「伝わ（る）」が社会的なつながりを表す言葉となっている．

4.1.2 孤独感の強度 (Intensity)

孤独感の有無が TRUE の場合，孤独感の強度（STRONG, WEAK）も付与した．STRONG の基準は日常に支障が生じるほどの孤独感または強いネガティブな感情が読み取れる場合とした．上記以外の場合は WEAK とした．

STRONG の例として「両親と死別してから仕事が手につかない」「かわいがっていた愛犬が亡くなり夜寝られない」などがある．また，WEAK の例として「家族は仕事で独りぼっち」「転校する事になって、友達と別れる時に悲しかった。」などがある．

4.2 設　　定

アノテーションは 2 名で実施した．LIFE STORY データ（2023 年 5 月調査，8 月調査）の「悲しい」に該当するテキストに対し前処理を行い，それぞれのアノテータが 600 件ずつアノテーションを実施した．前処理は数字を特定の文字列に置換，ノイズの削除などを行い，10 文字以下のテキストを除外した．600 件のうち 200 件はアノテータ間で共通であり，この 200 件を用いて一致度を求めた．アノテーションの判断に迷った場合，孤独感の有無では再現率を重視し TRUE，強度では適合率を重視し WEAK を付与した．

4.3 分　　析

ラベルの一致率を求めたところ，孤独感の有無の一致率は 0.935 (=187/200)，孤独感の有無かつ強度の一致率は 0.93 (=186/200)，すべてのラベル（有無・原因・強度）の一致率は 0.88 (=176/200) であった．また，Cohen の κ 係数で評価したところ孤独感の有無は 0.857，強度は 0.869 と高い数値を得た．

表 1　アノテーションの一致例

	テキスト	ラベル
(1)	妻が私に内緒で借金をし、支払えなくなって家出した時は本当に悲しかった。	STRONG
(2)	大切なペットが亡くなってしまったことがあります。長年一緒に生活してきたパートナーのような存在であり、家族として愛されていたペットがいなくなるという現実に直面することは、非常につらいものでした。	STRONG
(3)	むかしから見ていた人気ストリーマーが若くして亡くなった	WEAK
(4)	友達の飼っていた猫が死んだ	WEAK
(5)	GW 中に子供と巨人戦を観戦しにいったが、巨人が負けてしまったこと。	FALSE
(6)	お気に入りのお皿を割ってしまった	FALSE

なお，強度は重み付け κ 係数を用いた．以上のようにアノテータ間の一致度が高く，一貫性のあるデータセットを構築できた．

表 1，表 2 にそれぞれアノテーション結果が一致した例と一致しなかった例を示す．表 2 の社会的欲求は，孤独感の有無を TRUE としたアノテータがラベル付けしたものである．表 1 の (1)(2) のように家族やペットが関係し，テキストに強いネガティブな表現が含まれている場合は STRONG で一致した．他にも，(3)(4) のように家族やペットが亡くなっていてもテキストに強いネガティブな表現が含まれていない場合は WEAK で一致した．不一致だった例は，アノテータがテキストと同じ状況になったときに抱く感情の違いによって生じていると考えられる．例えば，表 2 の (2) は自分が言いふらしていると勘違いされたときに社会環境に受け入れられていないと思うかどうか，(4) は自分が理解されないときに自尊心や所属意識が欠如するかどうかによって違いが生じていると推測される．(6) は孤独感の強度が一致していない．テキストに強いネガティブな表現が含まれている一方，親しい人ではなく芸能人であることが不一致の原因であると考えられる．

5 実　　験

作成した孤独データセットを用いて，孤独感の有無（Loneliness）と孤独感の強度（Intensity）をそれぞれ推定する評価実験を行う．

5.1 設　　定

アノテーションを実施したテキストは各アノテータが 1 人でアノテーションしたテキスト計 800 件，一致度計算のため 2 人がそれぞれアノテーションした共通の 200 件が存在する．本実験では孤独データセットとして，共通の 200 件を各アノテータが付与したラベルの割合が同じになるように 100 件ずつランダムサンプリングし，それに残りの 800 件を合わせ，合計で 1,000 件のテキストを用いた．表 3 にラベルの内訳を示す．データの分割は学習データ，検証データ，テストデータでクラスの割合が均等になるように分割を行った．学習データは 700

表 2　アノテーションの不一致例

	テキスト	社会的欲求
(1)	長らく付き合いがあって、仲良くしていた友人が、冠動脈置換手術によって、弱ってしまって出かけられなくなった。	Love and Affection
(2)	私が言いふらしていないことをあたかも私がが言っているようにされていること。	Social Recognition
(3)	言葉が通じないなと思うと悲しい。	Social Recognition
(4)	正しいことをしても理解されていないことがあり、何とかしたいがなんともならない悩み。	Self-esteem and Belonging
(5)	自分自身の評価があまり良くなくて悲しい。	Self-esteem and Belonging
	テキスト	強度
(6)	今年に入って、YMO のメンバーの高橋幸宏、坂本龍一がなくなってとても悲しい気持ちでいる。	STRONG

表 3　ラベルの内訳

Loneliness		Intensity	
TRUE	FALSE	STRONG	WEAK
350	650	25	325

件，検証データ，テストデータはともに 150 件とした．

評価実験のため，孤独感の有無（分類モデル L）と孤独感の強度（分類モデル I）の分類モデルを作成した．分類モデルは日本語 Wikipedia で事前学習された BERT[3] をファインチューニングし，作成した．最後の隠れ層から [CLS] トークンのベクトルを取得し，線形層を最終層に追加してクラス分類を行った..事前学習済みの層の学習率は 5.0×10^{-5}，最後の線形層の学習率は 1.0×10^{-4} とした．最適化手法は AdamW，エポック数は 20 とし，早期終了基準として分類モデル L は最大正解率，分類モデル I は最大 F 値がそれぞれ連続で 3 エポック変更されないという基準を用いた．

5.2 結　　果

表 4 に分類モデルの評価結果を示す．分類モデル L は正解率，F 値ともに高い数値を得た．分類モデル I は正解率は高いものの，F 値は低い数値であった．これは孤独感の強度のラベルの割合が偏っていることが原因である．実際，適合率は 1.00 であり，STRONG と予測したテキストはすべて正解であった．

5.3 考　　察

検証データでエラー分析を行った．表 5 に孤独感の有無の予測結果とテキスト例を示す．正答の例として (1) の相手に理解してもらえないときの孤独感や，(2)，(3)，(4) のような家族やペットが亡くなったときの孤独感を学習できていた．また (5)，(6) のような「悲しい」という単語が含まれているテキストであっても孤独感を表していない場合は FALSE と予測できていた．エラーの例としては (7)，(8) のように Social Recognition や Self-esteem and Belonging が満たされていない文章は FALSE と予測してしまう傾向にあった．孤独感の強度が STRONG のテキストであっても (9) のように FALSE と予測してしまう例もあった．さらに，(10) のように「死んでしまった」という表現に釣られて TRUE にしてしまう例もあった．なお，(11)，(12) のように FALSE が正解ラベルではあるが，アノテータによっては TRUE と付与する可能性のあるテキストもいくつか存在した．このようなテキストは人手でラベル付けするのが難しく，機械学習モデルでも予測が難しいと考えられる．

表 6 に孤独感の強度の予測結果とテキスト例を示す．正答の例として (2) の親戚が亡くなったことだけを述べているテキストは WEAK と予測できており，(1) の父親が亡くなった影響で日常生活に支障が出ているテキストは STRONG と予測できていた．エラーの例では (4)，(5) ともに日常生活に支障が出ているテキストであるが STRONG と予測できなかった．人が判断した場合には明らかに STRONG のテキストでも WEAK と予測してしまうことがあった．

表 4　分類モデルの正解率と F 値

	Accuracy	F1
分類モデル L	0.880	0.833
分類モデル I	0.943	0.400

表 5　孤独感の有無の予測結果とテキスト例．(1)-(6) は正解した例を，(7)-(12) は不正解だった例を示す．

	テキスト	正解	予測
(1)	散々尽くしてきたのにその思いが伝わらないこと	TRUE	TRUE
(2)	ペットのハムスターが闘病の末亡くなった。	TRUE	TRUE
(3)	父が亡くなり、ドラマで亡くなるシーンが出て来ると思い出して涙してしまう。	TRUE	TRUE
(4)	飼っていたインコがしんで悲しい	TRUE	TRUE
(5)	梅の木の花が咲かなかったので悲しい	FALSE	FALSE
(6)	パートの仕事時間が減らされて、悲しい	FALSE	FALSE
(7)	リストラにあって無職なのが悲しい。	TRUE	FALSE
(8)	郵便局で不審者扱いされた事	TRUE	FALSE
(9)	妻が私に内緒で借金をし、支払えなくなって家出した時は本当に悲しかった。	TRUE	FALSE
(10)	最近買ったきゅうりの苗が死んでしまった。	FALSE	TRUE
(11)	母親が入院してしまったこと	FALSE	TRUE
(12)	母親が高齢で持病などがありいつ死んでもおかしくないと言われている。母の事を思うととても悲しい	FALSE	TRUE

3：https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

表 6　孤独感の強度の予測結果とテキスト例．(1)-(3) は正解した例を，(4)-(5) は不正解だった例を示す．

	テキスト	正解	予測
(1)	父がなくなったことです。いろいろな思い出を残してくれたので、もう四十九日は過ぎましたが、しばらく頭の整理ができませんでした。まだ近くいる気がします。	STRONG	STRONG
(2)	親戚のおじさんが亡くなった。	WEAK	WEAK
(3)	パートナーに会えないこと	WEAK	WEAK
(4)	上司のパワハラで、うつ病を発症して、２０年、心も身体もボロボロになって、全てを失いました。悲しいです。	STRONG	WEAK
(5)	妻が亡くなり、さらに友人、先輩と親しい人が続けて亡くなった	STRONG	WEAK

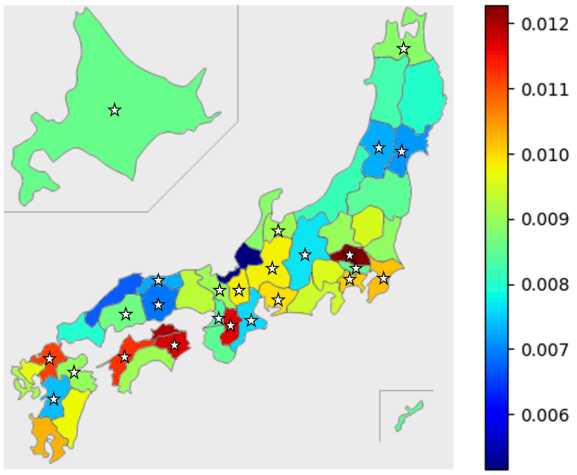


図 2　都道府県ごとの孤独投稿の割合．星マークは地方版孤独・孤立対策官民連携プラットフォーム推進事業 令和 4 年度 取組団体 [4]（都道府県・市区町村）が属する都道府県である．

6　孤独データセットの応用

孤独データセットを用いて学習した分類モデルを用いて都道府県ごとの孤独を分析した．

6.1　デ　ー　タ

Twitter（現 X）の API を用いて，2022 年 7 月 1 日から 31 日における日本の位置情報付き投稿テキスト計 5,596,197 件を取得した．取得した投稿テキストに対して，ハッシュタグや URL の除去などの前処理を行い，重複するものや 10 文字以下のものを除外した．その後，LIFE STORY データの「悲しい」カテゴリのテキストにアノテーションして作成したデータセットと条件を合わせるため，投稿テキストに対して LIFE STORY データを用いた感情分析を行った．具体的には，LIFE STORY データの各エピソードが含まれる感情のカテゴリを教師ラベルとし，ナイーブベイズ法を用いて各感情カテゴリに分類される確率を算出した．そして,「悲しい」が最も高い確率となった投稿テキストを抽出し，分類モデル L に入力して孤独投稿を取得した.「悲しい」が最も高い確率であった投稿は 57,648 件，孤独投稿は 6,902 件である．最後に，投稿に紐づいている位置情報をもとに，都道府県ごとの全投稿数と孤独投稿数を求め，都道府県ごとの孤独投稿の割合を算出した．

6.2　都道府県ごとの孤独

図 2 に都道府県ごとの孤独投稿の割合を示す．地方版孤独・孤立対策官民連携プラットフォーム推進事業 令和 4 年度 取組団体 [4]（都道府県・市区町村）が属する都道府県に星マークを記した．地方版孤独・孤立対策官民連携プラットフォーム推進事業は，地方自治体が孤独・孤立対策を支援するもので，孤独・孤立に悩む人に支援策が着実に行き届くことを目的としている．本事業は公募制であり，支援を受けたい団体が応募し採択された団体が支援を受けることができる．そのため，取組団体は孤独・孤立が顕在化，深刻化している可能性が高い．取組団体の属する都道府県は 25 存在し，このうち埼玉県，千葉県，神奈川県，岐阜県，愛知県，滋賀県，奈良県，徳島県，愛媛県および福岡県など，孤独投稿の割合が高い都道府県の多くは取組団体の属する都道府県であり，孤独投稿の割合は大まかには傾向を捉えられている．

一般に，医療や介護の資源は西日本側に集中していると言われている． 孤独投稿の割合は西日本側が高くなっていることから，医療が充実していたとしても精神的健康が良好になるとは限らない可能性を示唆している．さらに孤独になると医療や介護を受けなくなるといった可能性も想定され，医療や介護データとの関連性の調査は今後の課題とする．

6.3　予 測 結 果

適合率を求めるため，孤独分類モデルで予測した投稿をランダムにサンプリングし 150 件に対してアノテーションを行った．分類モデル L の適合率は 0.113 であった．LIFE STORY データに対する予測に比べ，投稿に対する予測では大きく性能が劣化した．表 7 に分類モデル L で TRUE と予測された投稿例とラベルを示す．エラーの例では (2) のように「死亡」に釣られている例があった．他には，(3) や (4) のように「お疲れ様」や「おやすみなさい」などの挨拶が含まれている例が多数存在した．しかし，学習データにそのような挨拶は含まれていなかった．

7　お わ り に

孤独感を表すテキストについての，オープンかつ孤独感の定義が明確に定められたデータセットを構築した（https://github.com/sociocom/Japanese-Loneliness-Dataset）．クラウドソーシングによって収集した「悲しい」から想起され

4：「地方版孤独･孤立対策官民連携プラットフォーム推進事業」https://www.cas.go.jp/jp/seisaku/kodoku_koritsu_platform/local_platform/index.html

表 7　分類モデル L で TRUE と予測された投稿例と正解ラベル

	テキスト	正解
(1)	お盆商戦で、行けませんいつ加美ちゃんと船田さんに会えるのかな	TRUE
(2)	羨ましい... 死亡。	FALSE
(3)	ちかちゃん、お疲れ様でした	FALSE
(4)	おやすみなさいりのさん	FALSE

たエピソードに対し，孤独感を表しているかどうか，その原因，孤独感の強度をアノテーションした．本データセットを用いて孤独分類モデルを作成し，評価したところ，学習元である LIFE STORY データは一定の性能で分類可能であることが示された．その一方で，応用として，学習データとは異なる Twitter の位置情報付き投稿を分類したところ，高い性能は得られなかった．例えば,「亡くなる」という意味の単語は無条件に孤独と判定されることや，挨拶のような Twitter では頻回に出現する表現が孤独と判定されてしまう問題が見られた．そのため，今後いかにドメイン適応させていくかが課題となる．

加えて，高齢化・医療施設との関連調査やさらなる応用を模索するとともに，データの拡張を行う予定である．

謝　　辞

本研究は，JST，未来社会創造事業，JPMJMI21J2，JSPS 科研費 22K12041 の支援を受けたものである．

文　　献

[1] Manfred E Beutel, Eva M Klein, Elmar Brähler, Iris Reiner, Claus Jünger, Matthias Michal, Jörg Wiltink, Philipp S Wild, Thomas Münzel, Karl J Lackner, et al. Loneliness in the general population: prevalence, determinants and relations to mental health. *BMC psychiatry*, Vol. 17, No. 1, pp. 1–7, 2017.

[2] Rebecca A Mullen, Sebastian Tong, Roy T Sabo, Winston R Liaw, John Marshall, Donald E Nease, Alex H Krist, and John J Frey. Loneliness in primary care patients: a prevalence study. *The Annals of Family Medicine*, Vol. 17, No. 2, pp. 108–115, 2019.

[3] Sarah C Griffin, Allison B Williams, Scott G Ravyts, Samantha N Mladen, and Bruce D Rybarczyk. Loneliness and sleep: A systematic review and meta-analysis. *Health psychology open*, 2020.

[4] Robert S Wilson, Kristin R Krueger, Steven E Arnold, Julie A Schneider, Jeremiah F Kelly, Lisa L Barnes, Yuxiao Tang, and David A Bennett. Loneliness and risk of alzheimer disease. *Archives of general psychiatry*, Vol. 64, No. 2, pp. 234–240, 2007.

[5] Graham M L Eglit, Barton W Palmer, A'verria S Martin, Xin Tu, and Dilip V Jeste. Loneliness in schizophrenia: Construct clarification, measurement, and clinical relevance. *PLOS ONE*, 2018.

[6] Clara González-Sanguino, Berta Ausín, MA Castellanos, Jesús Saiz, and M Muñoz. Mental health consequences of the Covid-19 outbreak in spain. a longitudinal study of the alarm situation and return to the new normality. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Vol. 107, p. 110219, 2021.

[7] Andrew Stickley and Michiko Ueda. Loneliness in japan during the COVID-19 pandemic: Prevalence, correlates and association with mental health. *Psychiatry research*, Vol. 307, p. 114318, 2022.

[8] Ann H Farrell, Irene Vitoroulis, Mollie Eriksson, and Tracy Vaillancourt. Loneliness and well-being in children and adolescents during the COVID-19 pandemic: A systematic review. *Children*, Vol. 10, No. 2, p. 279, 2023.

[9] Charikleia Lampraki, Adar Hoffman, Angélique Roquet, and Daniela S Jopp. Loneliness during COVID-19: Development and influencing factors. *PLOS ONE*, Vol. 17, No. 3, p. e0265900, 2022.

[10] Svetlana Kiritchenko, Will E Hipson, Robert J Coplan, and Saif M Mohammad. SOLO: A corpus of tweets for examining the state of being alone. *arXiv preprint arXiv:2006.03096*, 2020.

[11] Anietie Andy, Garrick Sherman, and Sharath Chandra Guntuku. Understanding the expression of loneliness on twitter across age groups and genders. *PLOS ONE*, Vol. 17, No. 9, p. e0273636, 2022.

[12] 岩井律子, 熊田孝恒. BERT を用いた孤独体験テキストからの孤独感推定. 言語処理学会 第 29 回年次大会 発表論文集, 2023.

[13] Kentaro Nakai, Ritsuko Iwai, and Kumada Takatsune. An examination of eating experiencies in relation to psychological states, loneliness, and depression using BERT. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 1S1IS305–1S1IS305, 2022.

[14] Yueyi Jiang, Yunfan Jiang, Liu Leqi, and Piotr Winkielman. Many ways to be lonely: Fine-Grained characterization of loneliness and its potential changes in COVID-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, pp. 405–416, 2022.

[15] 柴田大作, 若宮翔子, 伊藤薫, 荒牧英治. JIWC: クラウドソーシングによる日本語感情表現辞書の構築. 言語処理学会 第 23 回年次大会 発表論文集, 2017.

[16] Louise C Hawkley and John T Cacioppo. Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of behavioral medicine*, Vol. 40, No. 2, pp. 218–227, 2010.

[17] Johan Ormel, Siegwart Lindenberg, Nardi Steverink, and Lois M Verbrugge. Subjective well-being and social production functions. *Social indicators research*, Vol. 46, pp. 61–90, 1999.

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

2024年3月1日(金) 10:00 ~ 12:10 | T5-B オンライン（Zoom Events)

時空間予測

座長:横山 昌平(東京都立大学)
コメンテータ:金 京淑(産業技術総合研究所)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T5-B-7-01] マイクロブログを用いたイベント発生地周辺エリアの人口変動予測手法

*豊田 正史[1] (1. 東京大学生産技術研究所)

10:20 ~ 10:45
[T5-B-7-02] EVライフログを用いたEVのエネルギー変換効率推定

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

10:45 ~ 11:10
[T5-B-7-03] 犯罪および商業施設の情報に基づくゴミ予測手法の検討

*羽倉 輝[1]、山口 琉太[1]、義久 智樹[2]、下條 真司[3]、河合 由起子[1] (1. 京都産業大学、2. 滋賀大学、3. 青森大学)

11:10 ~ 11:35
[T5-B-7-04] EVの長期間ライフログに基づくバッテリー損耗のモデル化とVGIシミュレーションによる考察

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

11:35 ~ 12:00
[T5-B-7-05] 人流シミュレーションにおける複数指標を考慮したパラメータ影響の可視化

*森 千紗[1]、武田 芽依[2]、大西 正輝[3]、伊藤 貴之[1] (1. お茶の水女子大学、2. 筑波大学、3. 産業技術総合研究所)

マイクロブログを用いたイベント発生地周辺エリアの人口変動予測手法

豊田　正史†

† 東京大学生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †toyoda@tkl.iis.u-tokyo.ac.jp

あらまし　スポーツ会場・コンサート会場・展示会場など多くの人が集まるイベント発生地の周辺のエリアにおいては、不定期に発生するイベント及び参加者の移動のために、人口変動の予測が困難となる。本論文では、マイクロブログにおける未来のイベントに対する言及を用いてイベント発生地の数日先までの人口変動を予測し、その結果を用いて周辺エリアの人口変動までを予測する手法を提案する。携帯電話の位置情報を用いた人口変動統計および X（前 Twitter）の投稿を用いた実験により、予測の精度向上を確認する。

キーワード　人口変動予測, マイクロブログ、位置情報

1　はじめに

スポーツ、コンサート、展示会など数千から数万の人が集まるイベント時に会場の周辺エリアにおいては参加者の移動のために人混み、交通の混雑や遅延などが発生し、移動、商業などの社会活動の妨げとなるため、事前に周辺エリアの人口変動を予測しそれに応じて混雑回避や資源の最適配置を行うことが必要となる。しかし、イベントは平日、休日を問わず不定期に発生しその規模も種類により異なることから、イベント時の人口変動を予測することは困難である。スマートフォンなどの携帯端末の GPS から収集された位置情報から人口変動を把握し、広域における人口変動予測を行う手法はこれまでに研究されてきているが [1]-[6]、基本的には人口変動の時系列の周期性に基づく予測を行っており、外部情報についても祝日情報や天候など単純なものが用いられていることが多く、将来発生するイベントに関する手がかりを用いたものは少ない。

我々はマイクロブログにおける未来のイベントへの言及を利用して、イベント会場における人口変動を数日先まで予測する手法を提案してきた [7]-[9] が、本論文ではこれを用いてイベント会場周辺の人口変動をも予測する手法を提案する。本手法はスマートフォンなどの携帯端末の GPS から収集された位置情報を、一定区画（日本の地域メッシュや地図アプリケーションのタイルなど）、一定時間毎（1 時間、15 分など）に集計して区画毎の人口変動を推計したデータに基づき、イベント会場を中心とした一定の広さの領域（1Km 四方など）における 1 日から数日にわたる人口変動を区画ごとに予測するものである。各メッシュについて、過去の人口変動時系列、および予測対象の会場と日付に関するマイクロブログのポストの集合を入力として予測を行うが、対象とする領域が大きくなればなるほど対象とする区画の数は領域の一辺の長さの二乗のオーダーで増加するため、計算量を削減する必要がある。本手法では、イベント会場周辺の区画の人口変動はイベントに大きく影響されることを利用し、イベント会場が存在する区画のみポストの情報を用いた予測を行い、周辺区画についてはその予測結果を用いて予測することで計算量を削減した。本手法の基礎的な検討を行うため、LINE ヤフー株式会社が提供する人口変動統計データである DS.DATASET、および X（前 Twitter）データを用いた予測実験を行い、予測精度の向上を確認した。

2　問題設定およびデータセット

提案手法の基礎的な検討を行うためイベント発生地については東京都内最大規模のイベント会場である東京ドームを選択し、周辺区画における翌日の人口変動時系列を予測するタスクを設定する。予測対象となる日を d とし予測を行う日をその前日 $d-1$ とする。各区画の時系列の予測は、入力として n 日分の過去の時系列 ($d-n$ から $d-1$)、および $d-1$ の終りまでに得られる X のポストとし、d における周辺の各区画の人口変動時系列を予測する。予測器への入力となるポストデータは [7], [8] と同様に予測対象日に対する未来の言及を含むポストを使用する。具体的には d の日付を表す文字列および会場の名称（今回は「東京ドーム」「ビッグエッグ」など）を両方含むポストを抽出しその集合を入力とする。

本論文において使用する人口変動統計データである DS.DATASET の詳細は以下のとおりである。

- 区画：総務省が定める標準地域メッシュ（125m 四方）
- 時間：15 分単位（1 日当たり 96 時点）
- 人口：観測された端末数から拡大推計した区画内単位時間毎の人口
- 期間：2020 年 9 月 1 日〜2023 年 12 月 31 日

本データより、図 1 に示す東京ドームを中心とした青いマークで示す 81 区画（9 x 9 区画）の人口変動時系列を抽出して実験に用いる。

3　人口変動予測手法

本手法は、過去の人口変動データおよび X のポストを用いたイベント会場周辺の区画における未来の人口変動を予測する。X のポストを用いた予測モデルは学習および予測共に高コストであることから、本手法ではイベント会場の区画においてのみ

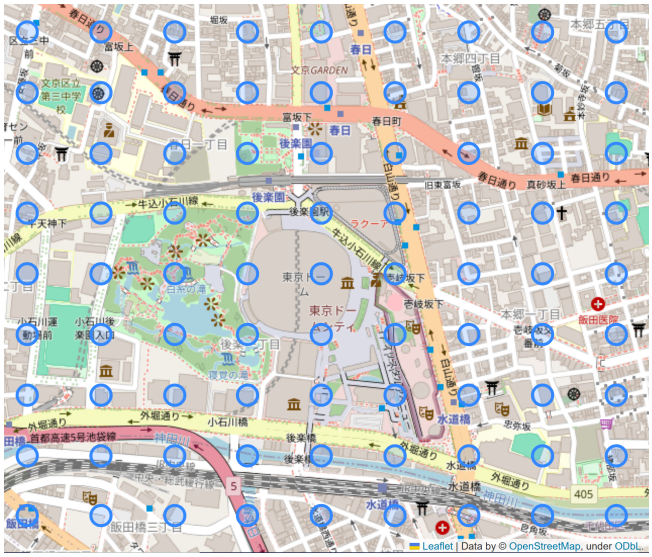


図 1　東京ドーム周辺の予測対象区画

X のポストを用いた予測を行い、周辺の区画の人口変動はイベントに大きく影響されることを利用して周辺区画においてはその予測結果を用いた人口変動の予測を行う。予測対象とするエリアは、イベントの影響を受けやすい、会場を中心とした東西、南北方向それぞれ約 2 Km 程度の矩形領域とする。ここでは簡単のため対象エリアに主要なイベント会場は 1 つのみ存在する状況を考える。本手法の詳細な手順を以下に示す。

- 予測対象するエリアにおいてイベント会場の区画を決定する。各区画について過去 n 日分の時系列から対象日 d の時系列を予測する回帰モデルを構築し、予測誤差の一番大きくなる区画を**会場区画**として選択する。それ以外の区画を**周辺区画**とする。
- 会場区画について、過去 n 日分の時系列および d に関するポスト（d の日付を表す文字列および会場の名称を両方含むポスト）を入力とし、d の時系列を予測する回帰モデルを構築する。予測モデルとしては、[7]-[9] で提案したものなどを利用する。
- 各周辺区画について、過去 n 日分の時系列と上記会場区画の d の予測時系列を結合したものを入力とし、d の時系列を予測する回帰モデルを構築する。本モデルの訓練時には、会場区画の d の時系列が入力として必要となる。このため、d における人口変動の実績データを入力データとし、予測が正確に行われたものとして訓練を行い、周辺区画の予測時には会場区画の予測結果を用いる。

4　実験結果

提案手法を用いて上述した東京ドーム周辺のエリアを対象とした人口変動の予測実験を実施した。2020 年 9 月 1 日から 2023 年 8 月 31 日までのデータを訓練データとし、2023 年 9 月 1 日〜2023 年 12 月 31 日をテストデータとして用いた。本実験ではイベント会場の予測データを用いて周辺区画の予測を行うことのフィージビリティを確認することを目的としているため、人口変動時系列の予測モデルとしては特に複雑なモデルは用いず、LightGBM [10] を参考に Python のライブラリとして実装された高速な Gradient Boosting Regressor である、Histogram-based Gradient Boosting Regressor (HGBR) を用いた。

表 1　東京ドーム 4 区画の MAE

81 / 82	113 / 110
96 / 97	- / 130

表 2　各手法における MAE の比較

	ベースライン	提案手法	オラクル
会場区画の MAE	200	130	-
周辺区画の MAE の平均	85	75	63

会場区画用の回帰モデルは、過去 n 日分の会場区画の時系列および $d-1$ の終りまでに得られる X のポストを入力とする。入力時系列については、1 週間分の過去の時系列（$n=7$、96x7 時点）とし、入力ポストについては d の日付および東京ドームを表す文字列を両方含むものの集合とする。本実験では上記の理由からポストの特徴量としては単純なものを用いることとし、ポスト集合から形態素解析によって得られる形態素の BoW (Bag of Words) を使用する。高頻度の形態素からストップワードを除去して上位 10000 の形態素を抽出し、形態素の頻度については無視して各形態素の有無のみを 0、1 のベクトルとして表現したものを用いる。過去の時系列と形態素有無のベクトルを結合し入力とする。

周辺区画用の回帰モデルについても HGBR を用いており、予測対象の周辺区画における過去 1 週間分の時系列、および会場区画の予測結果を結合したものを入力として、d の時系列を予測する。周辺区画用の HGBR の学習時には、会場区画予測結果の部分に実績データを入力して訓練を行う。

提案手法の有効性を確認するため、以下の 3 種類の予測手法の比較を行う。

- ベースライン：各区画について過去 7 日分の時系列のみを用いて d の予測を行う HGBR を用いる。
- オラクル：提案手法において会場区画用の予測が正確に行えたものとし、周辺区画の予測時には会場区画の正解データを入力する。
- 提案手法：会場区画の予測結果を用いて周辺区画の予測を行う。

提案手法では、X のポストを用いた予測をイベント会場 1 区画のみで行っているが、実際には東京ドームは図 1 から分かるように 2x2 の 4 つの区画にまたがっており、今回の実験では右下の区画が会場区画として選択されている。このため、まず他の 3 つの区画において十分な精度が得られているかどうかを確認した。表 1 に東京ドームの 4 区画において X のポストと時系列を用いたマルチモーダル回帰モデルを用いた場合の MAE と、会場区画の予測結果を用いる提案手法での MAE を比較したものを示す。各区画においてスラッシュの左側が提案手法、右側がマルチモーダル回帰モデルを用いた時の MAE である。

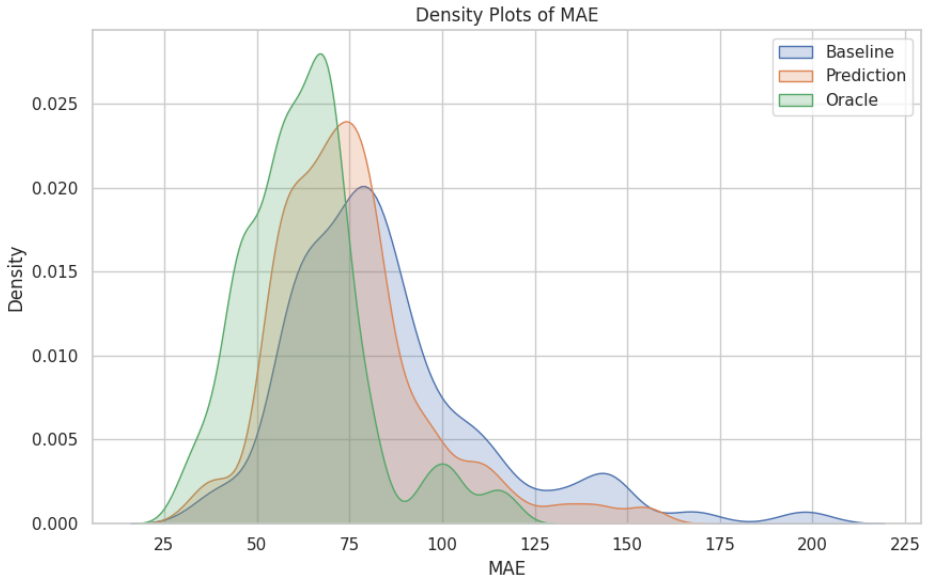


図 2　各周辺区画における MAE の分布

右下の区画は会場区画として選択されているため、マルチモーダル回帰モデルの結果と同じとなる。他の 3 区画については、どちらの手法でもほぼ同等の MAE となっており、これらの区画において計算コストの高いマルチモーダル回帰モデルを用いなくても、精度については変わりのないことが確認できた。

次に、会場区画の予測モデルの Mean Absolute Error (MAE)、および周辺区画の予測モデルについて、80 の周辺区画の MAE の平均を比較したものを表 2 に示す。会場区画の予測については、ベースラインと比較して MAE を 70 減らすことができておりおよそ 35 %誤差を削減できている。周辺区画の平均 MAE については、ベースラインとオラクルの差分が削減可能な最大の誤差であり、これが 23 であるところ提案手法では誤差が 10 減少しており、削減可能な誤差のうちおよそ 43 %を削減できていることになる。

各周辺区画の MAE の分布を手法毎に密度プロットしたものを図 2 に示す。ベースラインは MAE の大きい方へ偏った分布となっており、提案手法では MAE の大きな区画について誤差を改善できていることが分かる。一方既に MAE の小さい区画については大きな改善はできていないことも分かる。これらの結果から、イベント会場の人口変動の予測結果が、周辺区画の予測に有用であることが確認でき、さらに予測が正確にできた場合に最大どれだけの誤差が削減可能かについても確認することができた。

さらに、各周辺区画毎に各手法の MAE を比較したものを、図 3 に示す。図 1 に示した各区画毎に、ベースライン（青）、提案手法（オレンジ）、オラクル（赤）の MAE を棒グラフで示している。全ての区画において、提案手法およびオラクルによる MAE の悪化は起きていない一方で、誤差の改善幅は区画により大きく異なる。手法の想定通り、イベント会場周辺では改善幅が大きくなり、イベント参加者が通るルート沿いにおいても一定程度の MAE の改善が見られている。

特に、東京ドームから水道橋駅、後楽園駅への経路、および水道橋駅を通る総武線沿線において大きな MAE の改善が見られている。東京ドーム敷地外の北側のエリアは MAE が比較的大きく、イベント時の参加者の移動経路とはなっていないにも関わらず、MAE がよく改善されている。これは文京区役所や文京ガーデンなどの高層ビルでは平日日中の人口が多く、休日との人口差が大きい区画であることが理由の一つであると考えられる。これらの区画では、平日と祝日の取り違えによって起きる予測ミスの誤差が大きくなるが、東京ドームのイベント時系列（野球のナイトゲーム、デイゲームの時系列の違いなど）からある程度平日と祝日の区別がつくことから、イベントが予測できたことにより誤差の改善がなされていると考えられる。このように、イベントによる直接の人口変動がある区画ではなくても改善が見られる場合がある。

5　まとめと今後の課題

マイクロブログにおける未来のイベントに対する言及を用いてイベント発生地の数日先までの人口変動を予測し、その結果を用いて周辺エリアの人口変動までを予測する手法を提案した。イベント発生地にのみ計算コストの高い予測器を用いることで、周辺区画の予測にかかる計算コストを削減しスケールアップすることが容易な手法となっている。実際の人口変動データ及び X のポストを用いた実験により、提案手法の有効性および限界を示した。

本論文では東京ドーム周辺のみで実験を行っているが、明確なイベント会場ではない観光地など他のエリアについても今後実験を行う予定である。また複数のイベント会場が含まれるような都市レベルのエリアについても実験を行う予定である。今回はフィージビリティの確認のため HGBR を用いて回帰を行っ

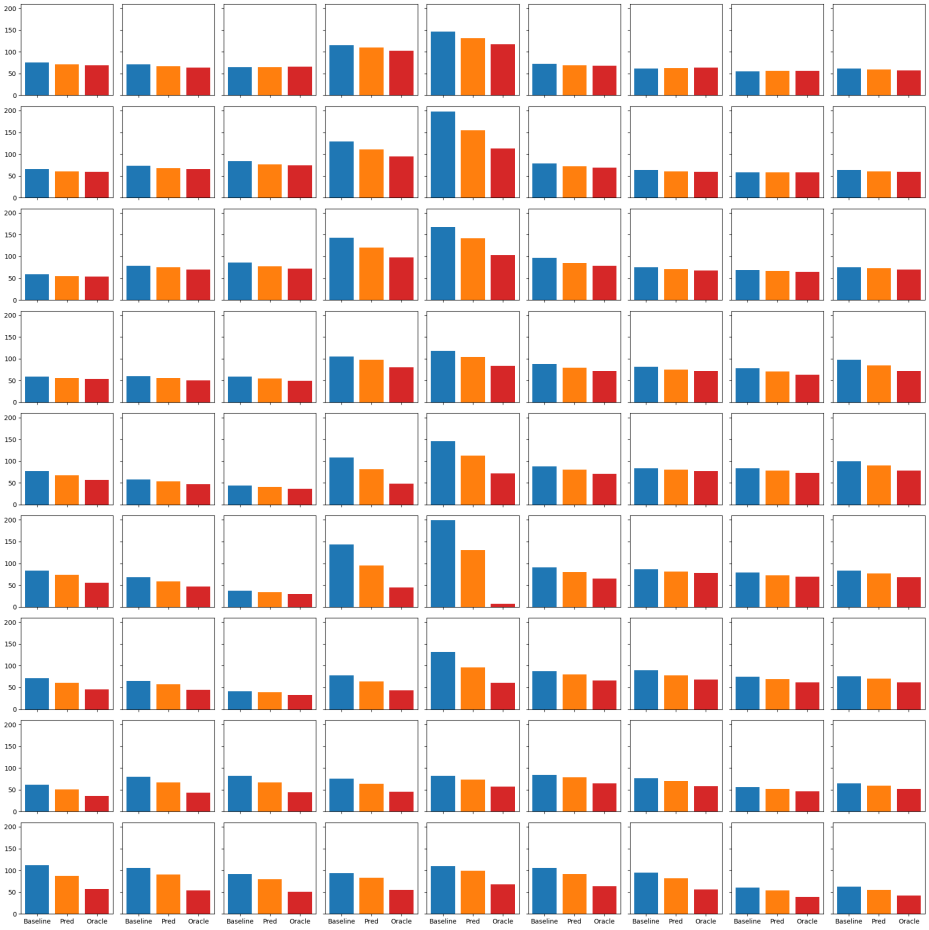


図 3　各周辺区画における MAE の比較

たが、既存研究では深層学習を用いた手法なども提案されており、そうした手法を用いてさらなる精度向上が可能かどうかについても確認していく予定である。

謝　　辞

本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究（JPJ012368C0540） により得られたものです。

文　　献

[1] Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. CityMomentum: an online approach for crowd behavior prediction at a citywide level. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15), pp. 559-–569, 2015.

[2] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17), pp. 1655-–1661, 2017.

[3] Binbing Liao, Jingqing Zhang, Chao Wu, Douglas McIlwraith, Tong Chen, Shengwen Yang, Yike Guo, and Fei Wu. Deep Sequence Learning with Auxiliary Information for Traffic Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18), pp. 537-–546, 2018.

[4] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. Dynamic spatial-temporal representation learning for traffic flow prediction. Transactions on Intelligent Transportation Systems, Vol. 22, No. 11, pp. 7169–7183, 2021.

[5] Cecaj Alket, Marco Lippi, Marco Mamei, and Franco Zambonelli. "Sensing and Forecasting Crowd Distribution in Smart Cities: Potentials and Approaches." IoT, vol. 2, no. 1, pp. 33–49, 2021.

[6] Yulai Xie, Jingjing Niu, Yang Zhang, and Fang Ren. Multi-size patched spatial-temporal transformer network for short- and long-term crowd flow prediction. IEEE Transactions on Intelligent Transportation Systems, Vol. 23, No. 11, pp. 21548–21568, 2022.

[7] R. Tsukada, H. Zhan, S. Ishiwatari, M. Toyoda, K. Umemoto, H. Shang, and K. Zettsu, "Crowd forecasting at venues with microblog posts referring to future events." IEEE Conference of Big Data, Workshop on Big Spatial Data (BSD), pp. 3147-3155, 2020.

[8] 塚田涼太郎, 浩森, 石渡祥之佑, 豊田正史, 梅本和俊, 商海川, 是津耕司. "未来のイベントに言及するマイクロブログ投稿を用いた人口変化の予測," in DEIM, 2021.

[9] 川崎仁嗣, 塚田涼太郎, 豊田正史, 是津耕司. "COVID-19 によるイベント開催状況の変化に適合した人口変化の予測," in DEIM, 2022.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 3149–3157, 2017.

EV ライフログを用いた EV のエネルギー変換効率推定

曽我紗代子† 劉 屹† 何 新† 田中 悠斗† 富井 尚志††

† 横浜国立大学大学院環境情報学府情報環境専攻 〒 240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
†† 横浜国立大学大学院環境情報研究院 〒 240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
E-mail: †{soga-sayoko-cz,liu-yi-tk,he-xin-fw,tanaka-yuto-kt}@ynu.jp, ††tommy@ynu.ac.jp

あらまし 本稿では, 3 年間 1 台の電気自動車（Electric Vehicle：EV）でライフログとして集めた約 100 万件の EV の CAN データを用いて, 車種固有のモータ・インバータ総合効率マップを推定する. EV はバッテリに蓄積された電気エネルギーを運動エネルギーに変換する過程でインバータとモータを用いており, そこで変換ロスが生じる. この変換効率は, モータ・インバータ総合効率と呼ばれる. この値はモータの回転速度とトルクから求められる二変数関数であり, このマップ推定を行う. 実際に収集したデータから算出したモータ・インバータ総合効率の値は, 特に低トルク帯で予想される値よりも著しく悪い効率をとる. そこで我々は, 低トルク帯は, 効率ではなく消費エネルギーを直接参照する別モデルを採用することにした. この提案モデルを用いて消費エネルギー推定を行ったところ, 推定結果として妥当なものが得られた.

キーワード モビリティ, ライフログ, 電気自動車, CAN データ, エネルギー変換効率

1 はじめに

近年, センサ技術やストレージ技術の発達により, 実世界の膨大で多様なデータを取得・蓄積することが可能になった. これにより, ヒトの生活に関する記録である「ライフログ」[1] が容易に利用できるようになった. 一般にライフログは, 日常生活の中から負担なく収集することができるデータである. このデータを有効に活用することが課題となっている [2].

一方, 近年, 気候変動問題が世界的な課題となり, 脱炭素化および省エネルギー化が求められている. その一環として, 現在世界中で CO_2 の排出が最終需要部門の 37%を占める運輸部門 [3] について対策が必要である.

その手段として, 電気自動車（Electric Vehicle：EV）が注目されている [4]. EV は走行時に CO_2 を排出せず, ライフサイクルにおける CO_2 排出量も同クラスのガソリン車（Internal Combustion Vehicle: ICV）と比べて 6 割程度であるという特徴がある [5]. 一方, EV は航続距離の短さや, 充電時間の長さ, 充電ができる場所が限られているなどの課題がある. したがって, EV ユーザにとっては消費エネルギーが重要な情報となる.

航続距離の指標として, 電費がカタログ値として公表されるが, この値は WLTC モード [6] などの規定化されたものであり, 実際に走行した場合の電費とは異なることが多い. この原因は, EV のエネルギー消費が, 道路勾配や加減速, 速度帯, 気象状況, エネルギー変換効率などによって大きく変化するためである. 従って, EV ユーザの航続距離に対する不安を解消するために, EV を走行する際の運転の仕方や環境に応じた消費エネルギーを知ることは重要である. さらに, どのような要因でエネルギー消費が発生しているのかが分かれば, エネルギーを抑える運転ルートや運転方法などが明らかになるといえる. この時, EV ユーザは, 走行時にパワーメータ画面から EV がどのくらい出力をしているか瞬時値を確認することができる. しかし, どのような要因でエネルギーを消費しているのかの内訳はユーザからは見えない.

以上に対し, 我々の先行研究では, 自動車に搭載したスマートフォンから取得した GPS のセンサデータをライフログとして収集し, EV に乗り換えた場合の消費エネルギーを推定する方法を提案した [7]. EV の走行に必要なエネルギーは, 運動方程式に基づく物理モデルにより説明できるため, 自動車の走行ログから消費エネルギーの推定値を計算した. これにより, 各要因に対する消費エネルギーを明らかにできるようになった.

ところで, 消費エネルギーに影響を与える要因の 1 つとしてエネルギー変換効率が挙げられる. 一般に, EV は電気的エネルギーを力学的エネルギーに変換することで駆動力を作り出している. その際, インバータとモータの 2 つのエネルギー変換機構を介するため, エネルギーの変換ロスが生じる. この変換効率は, モータ・インバータ総合効率と呼ばれる [8]. モータ・インバータ総合効率は, モータの回転速度（回転数）とその回転速度時の出力トルクによって決定される 2 変数関数であると捉えられる. この 2 変数関数を本研究ではモータ・インバータ総合効率マップと呼ぶ. モータ・インバータ総合効率マップは車種固有のものである. また, 一般に公開されておらず, 測定しようとすればシャシダイナモといった専用の機械が必要となる. 我々の先行研究 [7] では, モータ・インバータ総合効率マップは文献 [8] に示されたものを用いており, 使用車種固有のものは使用していなかった.

ここで, EV は走行時の内部データや GPS データを取得することができる. EV の内部には, 制御情報を転送するための機構として, CAN（Controller Area Network）が使われている. CAN は複数の電子制御装置（ECU：Electric Control Unit）間で通信を行う車載ネットワーク用に開発されたシリアルバス通信規格である. CAN バスを流れるデータはメーカーや車種

によって異なるが，EV であればバッテリーパックのセル電圧やバッテリー残量，モータのトルクといったデータが存在する．これが EV の内部データである．我々はこれらのデータを EV のライフログとして 10 年以上収集してきた．

そこで, 本研究では EV のライフログから隠されたモータ・インバータ総合効率マップを推定し, 効率を含めた消費エネルギー要因の内訳を明らかにすることを考えた．モータ・インバータ総合効率は, モータの回転速度とトルクに依存した値であり, これらは CAN データから取得・算出が可能である．モータ・インバータ総合効率マップを明らかにすることにより, 走行前には, 効率を考慮したより精密な消費エネルギー推定ができるようになり, 運転計画に役立つ．また, 走行後には, CAN データから取得した総消費エネルギーの内訳を説明できるようになり, 効率的な走行方法が明らかとなり運転改善につながると考えられる．

これに対し, 本研究では, 1 台の EV のライフログを用いてモータ・インバータ総合効率マップを推定し, 別車種かつ同種のインバータとモータを用いた EV の走行データに対し推定したマップを用いた消費エネルギー推定を行うことで精度検証を行う．

2 関 連 研 究

本研究では, 本来は活用されていない EV のライフログ, すなわち CAN データを活用してモータ・インバータ総合効率マップを推定し, 場所・車個体・人などに依存した消費エネルギーを提案することが目標である．

近年, 自動車より収集した走行ログを活用する研究が盛んにおこなわれている．特に, ITS（Intelligent Transport System: 高度道路交通システム）のための自動車内蔵センサの活用 [9] は注目されており, 多くの研究がみられる．しかし, 高精度な自動車内蔵センサにより取得したデータは膨大であるために, その多くは事故回避や自動運転を目的とした自動車制御のために使われている．

また, EV の消費エネルギー推定に関する研究は多くの研究がなされている．消費エネルギーの推定モデルは, 大きく, 機械学習に基づくモデルと物理式に基づくモデルに分類できる．

機械学習モデルには, Ullah らの [10], Pan らの [11] がある．Ullah らは, EV の実走行データを用いて重回帰（MLR）, 人工ニューラルネットワーク（ANN）, Extreme Gradient Boosting（XGBOOST）, Light GradientBoosting Machine（LightGBM）の 4 つの機械学習モデルをそれぞれ構築して精度を比較した [10]．Pan らは, S-TSD という短距離のドライビングセグメント分割アルゴリズムを提案し, 深層ニューラルネットワーク（DNN）を基盤とした電気自動車のドライビングセグメントレベルのエネルギー消費予測モデルを構築し, XGBOOST と強化学習により構築したモデルと比較した [11]．これらの手法は総消費エネルギーを推定することはできるが, 要因の内訳はブラックボックスであった．

また, 物理式に基づくモデルには, 芳澤ら [12], Wu ら [13], Fiori ら [14], Wang ら [15], Shen ら [16] の研究がある．この時, 芳澤らはモータ・インバータ総合効率を考慮していない [12]．しかし, モータ・インバータ総合効率による変換ロスを考慮した消費エネルギー推定が, より現実の消費エネルギーに近いものを表すといえる．また, Wu らや Fiori らは, モータ・インバータ総合効率を固定値として用いている [13, 14] ．モータ・インバータ総合効率はモータの回転速度やトルクにより変換する値であり, これは車速や道路勾配の変化に応じる．特に低速の走行で車速の変化の多い市街地での走行では, 現実との効率の差が顕著に表れるといえる．さらに, Wang らはシャシダイナモを用いて計測したモータ・インバータ総合効率マップを用いている [15]．我々は, シャシダイナモのような計測用の機械を用いずに, EV のライフログからモータ・インバータ総合効率マップを推定することを目標としている．Shen らは, 機械学習を用いて実走行データからモータ・インバータ総合効率マップを推定して用いている [16]．我々は, 日常の走行データに対して適切なデータ処理を行うことで, EV のライフログから要因の内訳まで考慮した消費エネルギーの推定を行う．

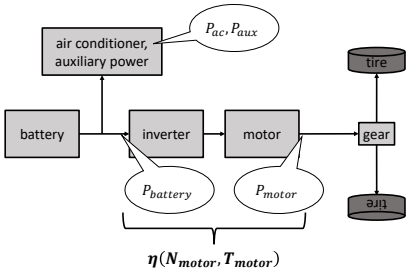


図 1: EV 内部のエネルギーの流れ

3 EV の内部データを用いたエネルギー変換効率推定

本章では, モータ・インバータ総合効率の算出方法および, EV の内部データの取得方法について述べる．

3.1 EV の内部構造

EV の内部エネルギーの流れを図 1 に示す．

まず, バッテリの電気エネルギーがインバータで直流から交流に変換される．次に, モータが電気エネルギーを力学的エネルギーに変換する．さらに, モータの作り出した力学的エネルギーを, 減速機によって走行に適した速度とトルクに変換してタイヤの駆動力にしている．この時, インバータとモータのエネルギーの変換過程で変換ロスが生じる．本論文では, モータ・インバータ総合効率は, この 2 つの変換機構における変換ロスをまとめたものを指す．この変換ロスを考慮した変換効率はモータ・インバータ総合効率と呼ばれ、文献 [8] に図 2 に示すグラフで説明されている．モータ・インバータ総合効率は車種固有の分布を示す．本論文では, 実際に走行する EV から収集されたデータを用いて, その車種固有のモータ・インバータ総合効率のマップを作成する．

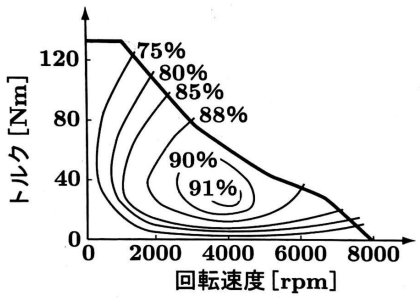


図 2: モータ・インバータ総合効率（文献 [8] より引用）

表 1: EV の駆動力に関わる変数

変数	単位	意味	データ取得方法
ρ	kg/m^3	空気密度	定数
μ	-	転がり抵抗係数	定数
g	m/s^2	重力加速度	定数
Cd	-	空気抵抗係数	カタログ値
A	m^2	車両前面投影面積	カタログ値
M	kg	車両重量	カタログ値
M_i	kg	慣性重量	車両総重量から換算
M_l	kg	積載重量	登場人数から換算
$v(t)$	m/s	車両速度	GPS 計測
θ	rad	道路勾配	地図
t	s	時刻	センサ計測時点

3.2 運動方程式に基づく EV 走行時に必要な駆動力

我々の先行研究 [7], [17], [18], [19] では, EV の消費エネルギーに関する検索可能なデータベースを EV のライフログから作成する ECOLOG システムを提案してきた. このシステムでは, EV の消費エネルギーが運動方程式を用いて説明可能である点 [8] に着目し, EV のエネルギー消費要因ごとに推定を行う EV モデルを定義した [7]. 本論文では, EV の消費エネルギーの推定値を算出する場合, このモデルを利用する.

以下に, 運動方程式を用いた EV の消費エネルギーの算出式を示す. ただし, 式中の変数は表 1 の通りである.

まず, 走行に必要な駆動力 $F_{driving}$ [N] は式 (1) のように, 加速抵抗 F_{acc} [N], 登坂抵抗 F_{slope} [N], 転がり抵抗 F_{roll} [N], 空気抵抗 F_{air} [N] の 4 種類から構成される.

$$F_{driving}(t) = F_{acc}(t) + F_{slope}(t) + F_{roll}(t) + F_{air}(t) \quad (1)$$

4 つの抵抗はそれぞれ, 式（2）〜（5）で表される.

$$F_{acc}(t) = (M + M_i + M_l)\frac{dv(t)}{dt} \quad (2)$$

$$F_{slope}(t) = Mg\sin\theta \quad (3)$$

$$F_{roll}(t) = \mu Mg\cos\theta \quad (4)$$

$$F_{air}(t) = \frac{1}{2}\rho C_d A v(t)^2 \quad (5)$$

次に, 駆動力 $F_{driving}$ [N] に対し, 速度 $v(t)$ [m/s] を掛け合わせることで, 車両が運動する際に必要なエネルギーの瞬時値（仕事率 [W]）を算出できる.

$$P_{motor}^{physical}(t) = F_{driving}(t) \times v(t) \quad (6)$$

この値が正の時は力行, 負の時は回生ブレーキまたは摩擦ブレーキにより減速していると見ることができる.

3.1 節で述べた通り, EV は電気エネルギーを駆動力に変換する過程でモータ・インバータ総合効率が生じる. この値は, モータの回転速度 $N_{motor}(t)$ [rpm] とトルク $T_{motor}(t)$ [Nm] によって決定される値である.

力行時の EV の瞬時消費電力 $P_{battery}^{physical}$ [W] は, $P_{motor}^{physical}$ [W] とモータ・インバータ総合効率 $\eta(N_{motor}(t), T_{motor}(t))$ の逆数で表すことができる（式（7））

$$P_{battery}^{physical}(t) = \frac{P_{motor}^{physical}(t)}{\eta(N_{motor}(t), T_{motor}(t))} \quad (7)$$

一方, 回生時の回収エネルギー $P_{regene}(t)$ [W] については, 以下に示す定義のもと, 回収できる量を考える. ただし, g [m/s^2] は重力加速度を表す.

- 加速度が $-0.3g$ 以下の際には,$-0.3g$ 分の運動エネルギーが回生される.
- 速度が 7 [km/h] 以下であれば回生は行われない.

この定義に従い, 回生エネルギーの瞬時値 $P_{regene}(t)$ [W]) を式（8）に示す. この時, 回生の時もモータ・インバータ総合効率 $\eta(N_{motor}(t), T_{motor}(t))$ に影響される [8].

$$P_{regene}(t) = \begin{cases} P_{motor}^{physical}(t) \times \eta(N_{motor}(t), T_{motor}(t)), \\ (F_{driving}(t)/M) > -0.3g \text{ かつ} \\ v(t) > 7\,[\text{km/h}]) \\ \\ -0.3gM \times v(t) \times \eta(N_{motor}(t), T_{motor}(t)), \\ (F_{driving}(t)/M) \leqq -0.3g \text{ かつ} \\ v(t) > 7\,[\text{km/h}]) \\ \\ 0\ , \quad (v(t) \leqq v(t) > 7\,[\text{km/h}])) \end{cases} \quad (8)$$

本論文では, EV の消費エネルギーの推定値を算出する場合, このモデルを利用する.

3.3 モータ・インバータ総合効率算出方法

収集した EV ライフログからモータ・インバータ総合効率を算出する方法を示す. ただし, EV ライフログは瞬時値であり, 時間の関数で表す. モータ・インバータ総合効率の算出に利用する各パラメタは表 2 に従う.

モータ・インバータ総合効率 $\eta(N_{motor}, T_{motor})$ は, モータの回転速度 N_{motor} [rpm] と出力トルク T_{motor} [Nm] によって決定される 2 変数関数ととらえることができる [8]. そして, $\eta(N_{motor}, T_{motor})$ は EV のバッテリ側出力 $P_{battery}^{mesured}$ [W] とモータ端出力 P_{motor}^{torque} [W] から算出される (式 (9)).

表 2: EV のモータ・インバータ総合効率に関わる変数

変数	単位	意味	データ取得方法
u	-	減速比	カタログ値（8.193）
r	m	タイヤ半径	停車時に計測（0.3050 [m]））
L	m	タイヤ周	$L = 2\pi r$ として計算（1.9164 [m]））
N_{motor}	rpm	モータの回転速度	CAN データ
N_{tire}	rpm	タイヤの回転速度	式 (13)
T_{motor}	Nm	モータのトルク	CAN データから式 (**??**) で算出
T_{tire}	Nm	タイヤのトルク	v から式 (13) で算出
V	V	バッテリ端電圧	CAN データ
I	A	バッテリ端電流	CAN データ
P_{ac}	W	エアコン出力	CAN データ
P_{aux}	W	電装品出力	CAN データ

$$\eta(N_{motor}, T_{motor}) = \begin{cases} \frac{P_{motor}^{torque}}{P_{battery}^{mesured}}, (力行時) \\ \frac{P_{battery}^{mesured}}{P_{motor}^{torque}}, (回生時) \end{cases} \quad (9)$$

バッテリ側出力 $P_{battery}^{mesured}$ [W] とモータ端出力 $P_{motor}^{torque}(t)$ [W] の間にインバータとモータを経由する．ここでモータ・インバータ総合効率という変換ロスが発生するため, 式 (9) のような比をとる．

モータ端出力 $P_{motor}^{torque}(t)$ [W] は，トルク値 $T_{motor}(t)$ [Nm] から物理式を基に算出した値であり，モータの回転速度 $N_{motor}(t)$ [rpm] を用いて, 式 (10) のように算出できる．

$$P_{motor}^{torque}(t) = \frac{2\pi T_{motor}(t) N_{motor}(t)}{60} \quad (10)$$

バッテリ側出力 $P_{battery}^{mesured}$ [W] は CAN から取得したものを用いた値である．CAN データに計上されるバッテリ端電圧 $V(t)$ [V] と電流 $I(t)$ [A] は, モータ駆動のための出力だけでなく, エアコンによる消費電力 $P_{ac}(t)$ [W] と電装品による消費電力 $P_{aux}(t)$ [W] が含まれる．よって, バッテリ端によるモータへの出力電力 $P_{battery}^{mesured}(t)$ [W] は式 (11) のように求めることができる．

$$P_{battery}^{mesured}(t) = V(t)I(t) - (P_{ac}(t) + P_{aux}(t)) \quad (11)$$

モータはタイヤと減速機を通じて繋がっているため, 式 (12) の関係式が成り立つ．

$$N_{motor}(t) = uN_{tire}(t) \quad (12)$$

また, タイヤの回転速度 $N_{tire}(t)$ [rpm] は車速 $v(t)$ [m/s] と連動する (式 (13))．

$$N_{tire}(t) = \frac{60v(t)}{L} \quad (13)$$

3.4 CAN を用いた EV の内部データの取得

本研究では，モータ・インバータ総合効率マップ推定のために EV の CAN データを使用した．CAN データの取得には Android アプリ LeafSpy Pro[1]を使用した．また，OBD2（On Board Diagnosis second generation）接続用の Bluetooth 通信端末として Kiwi3 OBD2 Wireless Bluetooth Diagnostic Scannar[2]を使用した．サンプリングレートは平均 3〜4 秒である．これにより, EV の走行中のデータを日常的に収集した．LeafSpy によって収集された実際の CAN データの系列を表 3 に示す．表 3 のように, 1 レコードはタイムスタンプ（DateTime）, バッテリー端電流値（PackAmp）, バッテリー端電圧値（PackVolt）, モータ端トルク値（TORQUE）などがみられる．

表 3: LeafSpy によって得られた CAN データの系列

DateTime	PackAmp	PackVolt	TORQUE	...
2020-08-03 08:52:50	12.709	388.61	-0.77	...
2020-08-03 08:52:54	2.761	388.70	0.27	...
2020-08-03 08:52:57	2.761	388.70	0.00	...
...	...	...	...	...

モータ・インバータ総合効率マップ推定用データ取得の実験用 EV（以降, EV1 と呼ぶ）として日産 LEAF の ZAA-ZE1 型（グレード e+ G）を用いた．また, 推定結果の精度検証用のデータ取得の実験用 EV（以降, EV2 と呼ぶ）として, ZAA-ZE1 型（グレード e+ G）と異種かつインバータとモータが同種である日産 LEAF の ZAA-ZE1 型（グレード G）を用いた．各車両の諸元を表 4 に示す．

以上の環境で収集された各データセットの概要を表 5 に示す．表 5 におけるデータ 1 件は, 表 3 の 1 レコードに該当する．EV1 のデータは, モータ・インバータ総合効率マップ推定用データである Dataset1 と, 精度検証用データである Dataset2 に分けた．EV2 のデータである Dataset3 は精度検証のみに用いる．

3.5 モータ・インバータ総合効率マップの精度検証方法

本研究では, CAN データから得た消費エネルギーの実測値を正解データとして, 推定した効率マップを用いて算出した EV の消費エネルギーの推定値と比較することで精度検証を行う．このとき, 対象道路を走行するために必要なエネルギーの合計値を比較することで評価する．

消費エネルギーの瞬時値の推定には, 3.2 節に示した方法を用いる．我々の先行研究では, 自動車に搭載したスマートフォンから取得した GPS データを基に, 3.2 節の方法に従って EV の消費エネルギー推定を行った [7]．この方法に対し, 本研究での精度検証では, 参照するモータ・インバータ総合効率マップを今回推定したものとする．あるトリップ（トリップ番号を n とする）の総消費エネルギーの推定値 $E_{estimated}^{n}$ [Wh] は以下のように求める．

1：”LeafSpy Pro”, Google Play, https://play.google.com/store/apps/details?id=com.Turbo3.Leaf_Spy_Pro&hl=ja&gl=US（参照 2023-11-13）.

2：Inc. PLX Devices, ”PLX Devices model Kiwi3” https://www.plxdevices.com/Kiwi-3-OBD-Car-to-Smartphone-Connection-p/897346002832.html（参照 2021-1-11）.

表 4: 実験用 EV 諸元

EV 名	型式	グレード	総電力量 [kWh]	車両重量 [kg]	モータの型式	減速比	CD 値	前方投影面積 [m^2]
EV1	ZAA-ZE1	e+ G	**62**	**1680**	EM57	8.193	0.28	2.19
EV2	ZAA-ZE1	G	**40**	**1520**	EM57	8.193	0.28	2.19

表 5: データセット

データセット名	使用車種	データの用途	収集期間	データ件数
Dataset1	EV1	マップ推定用	2020-8-1 ～ 2023-11-19	988,735
Dataset2	EV1	精度検証用	2021-3-4 ～ 2022-10-3	403
Dataset3	EV2	精度検証用	2023-5-17 ～ 2024-1-16	604

$$E^n_{estimated} = \int P^{physical}_{battery}(t)\,dt = \sum^{k} P^{physical}_{battery}(t_k)(t_k - t_{k-1}) \tag{14}$$

正解データの瞬時値 $P^{mesured}_{battery}(t)$ [W] は式（11）の通りに算出する．トリップ n の総消費エネルギーの正解値は以下のように求める．

$$E^n_{mesured} = \int P^{mesured}_{battery}(t)\,dt = \sum^{k} P^{mesured}_{battery}(t_k)(t_k - t_{k-1}) \tag{15}$$

評価の際, 各トリップの正解データと推定値の総消費エネルギーの相対精度 ϵ を比較する．トリップ n の相対精度 ϵ_n [%] は, 以下のように求める．

$$\epsilon_n = E^n_{estimated}/E^n_{mesured} \times 100 \tag{16}$$

4 モータ・インバータ総合効率マップの推定：提案手法 1

本章では, 収集したデータからモータ・インバータ総合効率マップを推定し, 精度検証を行うとともに, その手法について検討する．

4.1 データの集約方法

収集した実走行データから, 3.3 節の方法に従ってモータ・インバータ総合効率を算出した．用いたデータは EV1 のマップ推定用データである（Dataset1）．Dataset1 の全レコードを対象として, 式（9）を適用した．本来であれば, モータ・インバータ総合効率は $0\,[\%] < \eta(N_{motor}, T_{motor}) < 100\,[\%]$ という条件を満たすはずである．代表して, モータの回転速度が 4990 [rpm] 時（車速 80 [km/h] に相当）を図 3 に示す．横軸は, $0\,[\%] < \eta(N_{motor}, T_{motor}) < 100\,[\%]$ に限定している．

横軸がモータ・インバータ総合効率 $\eta(N_{motor}, T_{motor})$, 縦軸がモータのトルク T_{motor} [Nm] を表している．T_{motor} [Nm] > 0 の正トルク時を参照すると, 幅はあるもののある程度の傾向がみられる．一方, T_{motor} [Nm] < 0 の負トルク時を参照すると, 正トルク時に比べてばらつきが大きいことが分かる．負トルク時は回生時を表しており, 本研究ではその挙動に対する解析が不十分である．そこで, 本研究では負トルク時は推定の対象外とした．

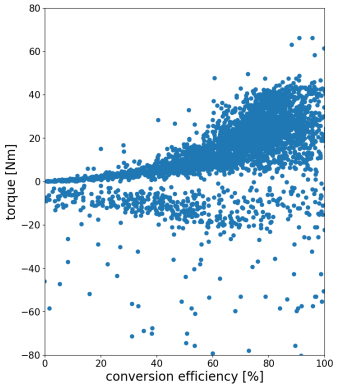


図 3: $N_{motor} = 4990$ [rpm] 時のモータ・インバータ総合効率とモータのトルクの関係

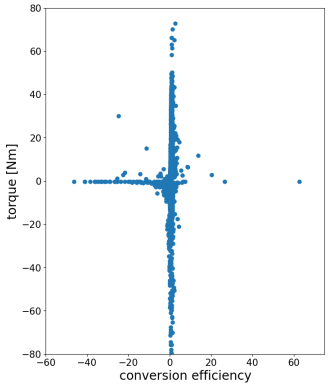


図 4: $N_{motor} = 4990$ [rpm] 時のモータ・インバータ総合効率とモータのトルクの関係

また, 3.3 節で示した通り, モータ・インバータ総合効率 $\eta(N_{motor}, T_{motor})$ は, モータの回転速度 N_{motor} [rpm] と出力トルク T_{motor} [Nm] によって決定される 2 変数関数ととらえられる．しかし, 収集した実走行データを用いてモータ・インバータ総合効率を算出すると, 同じモータの回転速度, トルクであっても異なる効率が算出された．この原因として, 瞬時値であることによる値のぶれや, モータ・インバータ総合効率に影響を与える隠れた要素が考えられる．特に, 前者は低トルク帯の時に特に多く現れる．モータの回転速度が 4990 [rpm] の時, $0\,[\%] < \eta(N_{motor}, T_{motor}) < 100\,[\%]$ という限定をせず, 全データを可視化した結果を図 4 に示す．図から読み取れる通り, トルクの値が小さいとき, $0\,[\%] < \eta(N_{motor}, T_{motor}) < 100\,[\%]$ になるはずのモータ・インバータ総合効率の値が大きくぶれている．

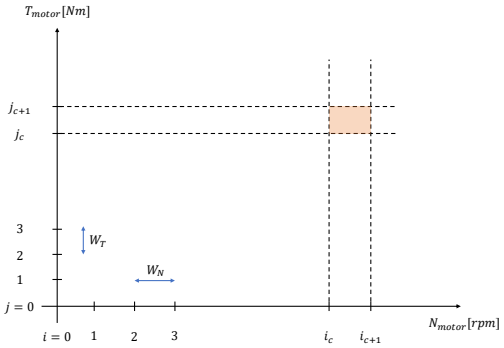


図 5: 回転数幅 W_N とトルクサンプル幅 W_T の可視化

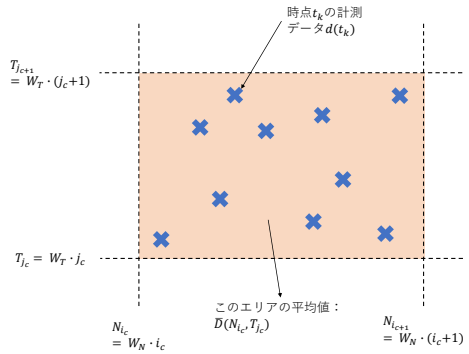


図 6: $\overline{D}(\overline{N^i_{motor}}, \overline{T^j_{motor}})$ の可視化

この原因として, 低トルク帯が低出力であることから, センシングタイミングのずれによるノイズの影響を大きく受けているということが考えられる. そこで, この値のぶれを相殺するため, モータ・インバータ総合効率を算出する前に, モータの回転速度 N_{motor} [rpm] と出力トルク T_{motor} [Nm] が概ね一致するレコードを集約することにした.

まず, 離散的なサンプル番号 $i, j (i = 0, 1, 2, ..., j = 0, 1, 2, ...)$ および, 回転数幅 W_N [rpm] とトルクサンプル幅 W_T [Nm] を定める（図 5）. この時, i, j のある値 i_c, j_c について, 図 6 のようなサンプル窓を定める.

全時系列計測データを $d(t_k)(k = 0, 1, 2, ...)$ とするとき, $d(t_k)$ の回転速度属性値を $d(t_k).N_{motor}$, トルク値を $d(t_k).T_{motor}$, また, 集約したい属性値を $d(t_k).value$ とする. この時, サンプル窓内のデータ平均値（窓平均）$\overline{D}(\overline{N^i_{motor}}, \overline{T^j_{motor}})$ は次式で定める.

$$\overline{N^i_{motor}} = W_N \cdot i \tag{17}$$

$$\overline{T^i_{motor}} = W_T \cdot j \tag{18}$$

$$\overline{D}(\overline{N^i_{motor}}, \overline{T^j_{motor}}) = \frac{1}{n_{ij}} \sum_{ij}^{k} d(t_{kij}).value \tag{19}$$

ただし, $W_N \cdot i \leqq d(t_{kij}).N_{motor} < W_N(i+1)$ かつ $W_T \cdot j \leqq d(t_k).T_{motor} < W_T(j+1)$ であり, n_{ij} はこの条件を満たすデータの件数である.

この窓平均に基づいて, バッテリ端によるモータへの出力電力 $P^{mesured}_{battery}(t)$ [W]（式（11）の時系列データを $d(t_k).value$ として算出されたものをバッテリ電力平均マップ $\overline{D^{mesured}_{battery}}(\overline{N^i_{motor}}, \overline{T^j_{motor}})$ とする. また同様に, モータ端出力 $P^{torque}_{motor}(t)$ [W]（式（10）の時系列データを $d(t_k).value$ として算出されたものをモータ端出力平均マップ $\overline{D^{torque}_{motor}}(\overline{N^i_{motor}}, \overline{T^j_{motor}})$ とする.

最終的に, サンプル窓で離散的に求められたモータ・インバータ総合効率マップは次式で求められる.

$$\eta(\overline{N^i_{motor}}, \overline{T^j_{motor}}) = \frac{\overline{D^{torque}_{motor}}(\overline{N^i_{motor}}, \overline{T^j_{motor}})}{\overline{D^{mesured}_{motor}}(\overline{N^i_{motor}}, \overline{T^j_{motor}})} \tag{20}$$

取得したデータには限りがあり, データが存在しないトルク, 回転速度の値も存在する. 本研究では, データの存在しない部分については線形補間を行うことで, 欠損値を補うことにする.

以上をまとめると, モータ・インバータ総合効率マップの推定手順は以下のようになる.

（1） モータの回転速度 N_{motor} [rpm]（粒度 $W_N = 10$ [rpm]）と出力トルク T_{motor} [Nm]（粒度 $W_T = 1$ [Nm]）ごとに $P_{battery}$ [W], P_{motor} [W] を平均値で集約

（2） モータの回転速度 N_{motor} [rpm] と出力トルク T_{motor} [Nm] ごとに式（9）の力行時の式に従って効率を計算

（3） 欠損値を線形補間

なお, 窓平均を求めるための W_N [rpm] と W_T [Nm] は十分に小さい. そこで, 記載の簡略化のため, 以降では式（17）を単に N_{motor} と, 式（18）を単に T_{motor} と記述する.

4.2 モータ・インバータ総合効率マップの推定結果：提案手法 1

4.1 節で示した方法でモータ・インバータ総合効率マップを推定した. その結果を図 7 に示す. 横軸はモータの回転速度 N_{motor} [rpm], 横軸は出力トルク T_{motor} [Nm] を表し, プロットの色はモータ・インバータ総合効率 $\eta(N_{motor}, T_{motor})$ [%] を表す.

4.3 モータ・インバータ総合効率マップの精度検証：提案手法 1

3.5 節で示した方法に従い, 収集したデータで作成したモータ・インバータ総合効率マップを用いて, 消費エネルギーを推定し精度を確認した. そのため, 精度検証の際は式（21）のように正トルク値をそのまま複製して式（8）に適用した.

$$\eta(N_{motor}, -T_{motor}) = \eta(N_{motor}, T_{motor}) \tag{21}$$

推定したマップを用いた EV の瞬間消費電力 $P_{battery}$ [W](t) の算出式を以下に示す.

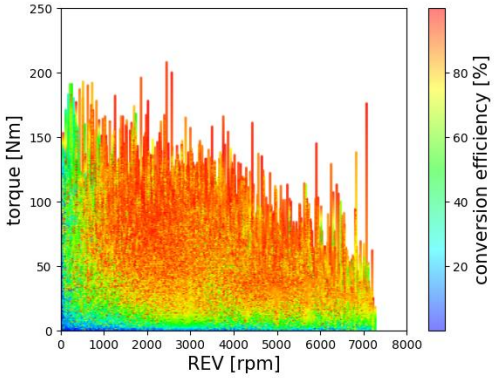


図 7: モータ・インバータ総合効率マップ

$$P_{battery}(t) = P_{battery}^{physical}(t) = \frac{P_{motor}^{physical}(t)}{\eta(N_{motor}, T_{motor})} \quad (22)$$

瞬間消費電力 $P_{battery}^{physical}(t)$ の算出方法は式（7）に示した通りである．$\eta(N_{motor}, T_{motor})$ は推定したモータ・インバータ総合効率を表す．

まずは，モータ･インバータ総合効率マップの推定に用いた EV1 のうち，マップ推定用とは別の精度検証用データ（Dataset2）で精度検証を行った．Dataset2 は，表 6 に示される走行データからなる．対象道路はそれぞれ，上り坂の道路，勾配変化の少ない道路，下り坂の道路という 3 つの特徴の異なる道路である．

各トリップに対する精度検証結果を表 7 に示す．相対精度は，100 [%] に近い走行も見られる一方で，TRIP4〜6, 8, 9 などのように精度の低い走行も見られた．

4.4 考察：提案手法 1

4.3 節の結果に対し，上り坂の走行に比べると，勾配変化の少ない道路や下り坂の道路を走行した際の方が精度の低さが顕著である．ここで，各道路を走行したときの，それぞれの走行時のモータトルクの分布を図 8 に示す．縦軸は確率分布を表す．

この分布から，勾配のある高トルク帯を含むところでは精度が良く，勾配の少ない低トルク帯（トルク 0 [Nm] 付近）を含むところでは精度が悪い結果となることが分かる．つまり，低トルク帯でのモータ・インバータ総合効率の精度が悪いということが考えられる．実際に，推定したモータ・インバータ総合効率マップ（図 7）を参照すると，モータの低トルク帯で，モータ・インバータ総合効率が高トルク帯に比べると効率が低い．この原因として，4.1 節でも示したように，低トルク帯での出力が小さいことからセンシングタイミングのずれによるノイズの影響を受けやすいということが考えられる．そこで，新たな手法を提案すことでこの問題を解消する．次章にその手法を示す．

5 モータ・インバータ総合効率マップの推定：提案手法 2

本章では，提案手法 1 の問題点を解消し，新たなモータ・インバータ総合効率マップの推定手法を示すとともに，その結果と精度検証結果を示す．

5.1 モータ・インバータ総合効率マップのモデル分け

4.4 節で示した通り，モータ・インバータ総合効率マップは低トルク帯で低い効率をとる．そこで本研究では，低トルク帯に関しては効率を算出するのではなく，瞬間消費電力 $P_{battery}^{mesured}$ [W] を直接参照する別モデルを作成することにした．瞬間消費電力 $P_{battery}^{mesured}$ [W] は，効率マップと同様に，モータの回転速度 N_{motor} [rpm] と出力トルク T_{motor} [Nm] によって決定される二変数関数とも捉えることができる（式（23））．

$$\begin{aligned} P_{battery} &= P_{battery}(N_{motor}, T_{motor}) \\ &= \frac{2\pi N_{motor} T_{motor}}{60} \end{aligned} \quad (23)$$

本稿ではこれをバッテリー出力マップと呼ぶ．つまり，本研究では，2 つのモデルを複合的に用いる．

具体的には，EV の瞬間消費電力 $P_{battery}$ [W](t) を求めるとき，高トルク帯では推定したモータ・インバータ総合効率マップを参照して，提案手法 1 と同様に式（22）を用いる．

一方，低トルク帯では以下の算出式を用いて求める．

$$\begin{aligned} P_{battery}(t) &= P_{battery}^{mesured}(N_{motor}, T_{motor}) \\ &= P_{battery}(N_{motor}, T_{motor}) \times \frac{M_{forEstimation}}{M_{forMap}} \\ &= \sum_{n=1}^{N} (V^n I^n - (P_{ac}^n + P_{aux}^n))/N \times \frac{M_{forEstimation}}{M_{forMap}} \end{aligned} \quad (24)$$

$P(N_{motor}, T_{motor})$ [W] は参照するバッテリー出力マップ，$M_{forEstimation}$ [kg] は消費エネルギーを推定したい EV の車両重量，M_{forMap} [kg] はモータ・インバータ総合効率マップの推定に用いた EV の車両重量を表す．瞬間消費電力 $P_{battery}$ [W] は，EV の車両重量に比例するため，マップの作成に用いた EV と推定したい EV の車両重量の比を取る．

モデル分けの閾値は，モータのトルクの閾値を変化させた時の 3.5 節で示した精度が最も良いときの値（= 36 [Nm]）とした．

前述の閾値の決定方法について説明する．

精度検証の対象は表 7 の TRIP4〜9 で，各 TRIP の誤差を計算した後，平均値をとっている．ここで，トリップ n の誤差は以下の式で表す．

$$\delta_n = \left| \frac{\epsilon_n}{100} - 1 \right| \quad (25)$$

閾値決めの結果を図 9 に示す．

横軸がモータのトルク，縦軸は相対誤差 δ_n を表す．図 9 より，モデル分けの閾値は，モータのトルク $T_{motor} = 36$ [Nm] が最も精度良いという結果が得られた．

以上を踏まえると，提案手法 2 におけるモータ・インバータ総合効率マップの作成方法は以下の通りになる．

（1） モータの回転速度 N_{motor} [rpm] と出力トルク T [Nm] ごとに $P_{battery}$, P_{motor} を集約

表 6: Dataset2 を収集した道路とトリップ数

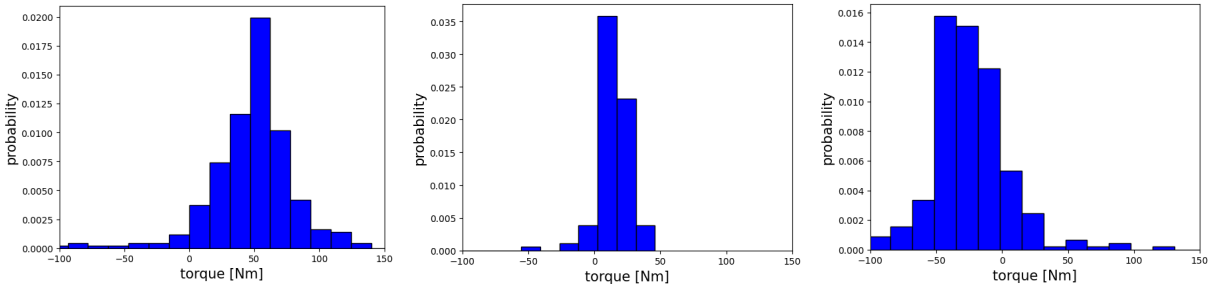
道路 ID	種別	距離 [km]	上り/下り方向	特徴（勾配）	Dataset2 のトリップ数	備考
Road10	自動車専用道路	13.8	下り	山道で登り坂	3	*3
Road11	自動車専用道路	13.8	上り	山道で下り坂	3	*3
Road12	自動車専用道路	14.5	下り	勾配変化が少ない	3	*4

*3 箱根新道

*4 小田原厚木道路（厚木～平塚）

表 7: 提案手法 1 を用いた消費エネルギー推定の精度検証結果

TRIP ID	道路 ID	データ収集日	勾配	正解 [kWh]	推定値 [kWh]	相対精度 [%]
1	Road10	2021-12-1	登り坂	4.86	5.20	107
2	Road10	2021-12-1	登り坂	4.91	5.14	105
3	Road10	2022-10-3	登り坂	4.70	5.38	114
4	Road11	2021-12-1	下り坂	-2.03	-1.50	74
5	Road11	2021-12-1	下り坂	-2.13	-1.45	68
6	Road11	2022-10-3	下り坂	-2.20	-1.66	75
7	Road12	2021-3-4	少ない	1.25	1.20	96
8	Road12	2022-9-22	少ない	1.02	1.27	115
9	Road12	2022-10-3	少ない	1.03	1.63	158



(a) TRIP1 におけるモータのトルク分布　(b) TRIP4 におけるモータのトルク分布　(c) TRIP7 におけるモータのトルク分布

図 8: モータのトルク分布

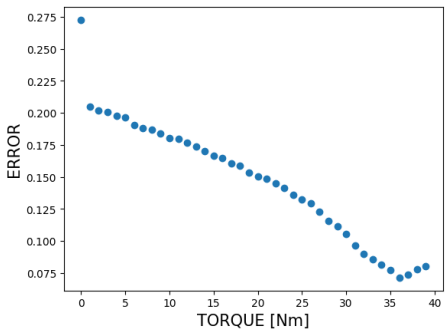


図 9: モデル分けの閾値決定

（2） モータの回転速度 N_{motor} [rpm] と出力トルク T [Nm] ごとに式（9）の力行時の式に従って効率を計算

（3） 欠損値を線形補間

（4） モータトルク $|T_{motor}| \leqq 36$ [Nm] の場合は消費エネルギー $P_{electric}(t)$ を直接参照

5.2 モータ・インバータ総合効率マップの推定結果：提案手法 2

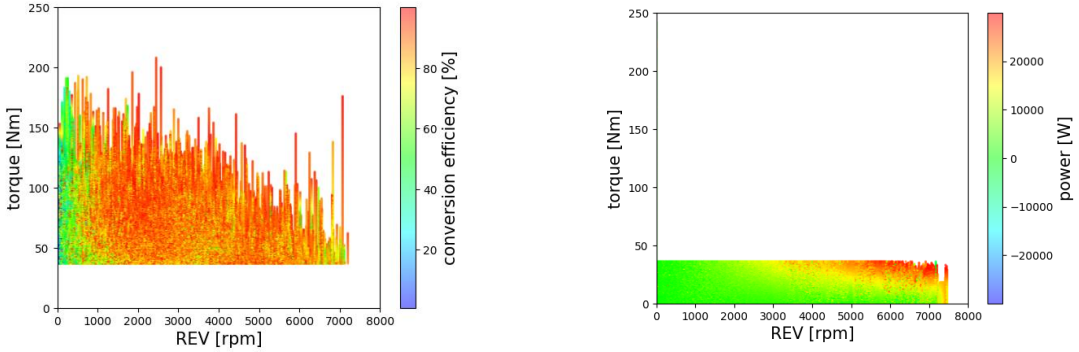
5.1 節で示した方法を用いて，モータ・インバータ総合効率マップを作成した（図 10）． 横軸がモータの回転速度 N_{motor} [rpm]，縦軸がとトルク T_{motor} [Nm] を示す． モータのトルクが $T_{motor} > 36$ [Nm] の場合は図 10a を用いる．プロットの各色は，右のカラーバーで示したモータ・インバータ総合効率 η [%] を表す．モータのトルク $|T_{motor}| \leqq 36$ [Nm] の場合は図 10b を用いる．プロットの各色は，右のカラーバーで示した消費エネルギー $P_{battery}^{mesured}$ [W] を表す．

5.3 モータ・インバータ総合効率マップの精度検証：提案手法 2

提案手法 2 を用いて表 7 と同様の精度検証を行った．結果を表 8 に示す．提案手法 1 と 2 の精度を比較すると，表 8 に示すように，提案手法 1 で精度の悪かった TRIP5～9 で精度が改善した．

5.4 考察：提案手法 2

各道路を走行したときのトルク分布は，4.4 節の図 8 に示し



(a) モータ・インバータ総合効率マップ　　(b) バッテリー出力マップ

図 10: 提案手法 2 における参照マップ

表 8: 消費エネルギー推定の提案手法 1・2 の精度検証結果の比較

TRIP ID	道路 ID	データ収集日	勾配	正解 [kWh]	提案手法 1		提案手法 2	
					推定値 [kWh]	相対精度 [%]	推定値 [kWh]	相対精度 [%]
1	Road10	2021-12-1	登り坂	4.86	5.20	107	5.21	**107**
2	Road10	2021-12-1	登り坂	4.91	5.14	105	5.10	**104**
3	Road10	2022-10-3	登り坂	4.70	5.38	114	5.31	**113**
4	Road11	2021-12-1	下り坂	-2.03	-1.50	74	-2.02	**99**
5	Road11	2021-12-1	下り坂	-2.13	-1.45	68	-1.98	**93**
6	Road11	2022-10-3	下り坂	-2.20	-1.66	75	-2.21	**100**
7	Road12	2021-3-4	少ない	1.25	1.20	96	1.22	**97**
8	Road12	2022-9-22	少ない	1.02	1.27	115	1.18	**107**
9	Road12	2022-10-3	少ない	1.03	1.63	158	1.28	**124**

た通りである．モータトルク $|T_{motor}| \leqq 36$ [Nm] の時は，バッテリー出力マップを直接参照することから，上り坂の走行では精度がほぼ変わらなかった一方で，下り坂や勾配変化の少ない道路では精度の向上が見られたといえる．

提案手法 2 では，高トルク帯はモータ・インバータ総合効率マップを用いるため，エネルギー消費要因の内訳を明らかにできる．一方で，低トルク帯はバッテリー出力を直接参照するため，エネルギー消費要因の内訳がわからないという欠点がある．ただし，提案手法 1 と比較して精度は向上する．

6　同型式モータ搭載の別車種による精度検証

本章では，モータ・インバータ総合効率マップの推定に用いていない EV2 の走行データを用いて推定したマップの精度検証を行い，その考察を行う．

6.1　モータ・インバータ総合効率マップの精度検証結果

モータ・インバータ総合効率マップの推定に用いていない EV2 の走行データ（Dataset3）を用いて推定したマップの精度検証を行う．精度検証に用いた道路は EV1 で用いた道路と同一である．表 9 に示す．

これらの道路に対して行った精度検証の結果を表 10 に示す．

6.2　考　　察

いずれの走行についても，提案手法 1 に対して提案手法 2 で精度の改善がみられた．つまり，モータ・インバータ総合効率マップの推定に用いた EV1 の精度検証用データ（Dataset2）で評価を行った場合（5.3 節）と同様の傾向が得られた．ここから，推定に用いた EV と異種かつインバータとモータが同種の EV で走行したデータに対しては，推定したモータ・インバータ総合効率マップを適用できることが示された．また，EV の日々の走行データから，消費エネルギー推定結果として妥当な結果を与えるモータ・インバータ総合効率マップを推定できたといえる．

7　ま　と　め

本研究では，3 年間 1 台の EV で収集した約 100 万件の EV の CAN データを用いて，車種固有のモータ・インバータ総合効率マップを推定し，消費エネルギー要因の内訳を明らかにした（提案手法 1）．推定した効率マップのうち，モータの低トルク帯では予想される値よりも著しく悪い効率を取ることが分かった．そこで，本研究では低トルク帯は，モータ・インバータ総合効率ではなくバッテリーの出力を直接参照する別モデルを採用した（提案手法 2）．この提案手法を用いて，推定したモータ・インバータ総合効率マップの精度検証を行った．その結果，推定に用いた EV で走行した際の消費エネルギー推定では，相対誤差約 1 割という結果が得られた．さらに，推定に用いた EV と異種かつインバータとモータが同種の EV で走行した際の消費エネルギー推定でも，同様の結果が得られた．効率マップを明らかにしたことにより，効率を含めたエネルギー消費要因の内訳を明らかにでき，速度や加減速の方法といった運転改善の判断に有用であると考える．今後は，消費エネルギー要因の内訳を明らかにしたことでできる有用事例を集める．また，温度

表 9: Dataset3 を収集した道路とトリップ数

道路 ID	種別	距離 [km]	上り/下り方向	特徴（勾配）	Dataset3 のトリップ数	備考
Road10	自動車専用道路	13.8	下り	山道で登り坂	1	*3
Road11	自動車専用道路	13.8	上り	山道で下り坂	2	*3
Road12	自動車専用道路	14.5	下り	勾配変化が少ない	1	*4

表 10: EV2 による精度検証結果

TRIP ID	道路 ID	データ収集日	勾配	正解 [kWh]	提案手法 1		提案手法 2	
					推定値 [kWh]	相対精度 [%]	推定値 [kWh]	相対精度 [%]
1	Road10	2024-1-16	登り坂	4.60	5.02	109	4.98	**108**
2	Road11	2024-1-16	下り坂	-1.71	-1.42	83	-1.86	**108**
3	Road12	2024-1-16	少ない	2.38	2.73	115	2.50	**105**
4	Road12	2024-1-16	少ない	1.63	1.82	111	1.69	**104**

などのモータのトルク・回転速度以外の隠されたパラメータを考慮したモータ・インバータ総合効率マップを推定することを目指す. さらに, 解析が不十分である回生ブレーキの挙動についてもライフログから明らかにすることを目指す.

謝　　辞

本研究の一部は JSPS 科研費 (課題番号 22H03810) の支援による. また, 道路勾配データの作成については, 株式会社ゼンリンデータコムのサポートを得た.

文　　献

[1] 総務省. “ライフログ活用サービスWGからの報告”. Last accessed: 10 November 2023, 2009.
[2] 総務省. “令和 5 年版 情報通信白書”. Last accessed: 10 November 2023, 2023.
[3] International Energy Agency. “analysis - transport”, 2022.
[4] 経済産業省資源エネルギー庁. “令和 2 年度エネルギーに関する年次報告 (エネルギー白書 2022)”, Last accessed: 22 July 2022.
[5] 黒川文子. “EV へのシフトと CO2 排出量に関する考察”. 環境共生研究, No. No.11, pp. 25–36, 2018.
[6] 国土交通省. “乗用車等の燃費表示に走行環境毎の燃費表示が加わります！”. Last accessed: 24 January 2022.
[7] 齊藤祐亮, 植村智明, 富井尚志. “EV 消費エネルギーログデータベースにおける消費電力量推定の精度検証”. 情報処理学会論文誌データベース（TOD）, Vol. 12, No. 4, pp. 40–52, 2019.
[8] 廣田幸嗣, 小笠原悟司. “電気自動車工学”. 森北出版, 2017.
[9] P. Arthurs, L. Gillam, P. Krause, N. Wang, K. Halder, and A. Mouzakitis. “a taxonomy and survey of edge cloud computing for intelligent transportation systems and connected vehicles”. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 7, pp. 6206–6221, July 2022.
[10] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal. “a comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability”. *Energy Environment*, Vol. 33, No. 8, pp. 1583–1612, 2022.
[11] Yingjiu Pan, Wenpeng Fang, and Wenshan Zhang. “development of an energy consumption prediction model for battery electric vehicles in real-world driving: A combined approach of short-trip segment division and deep learning. *Journal of Cleaner Production*”, Vol. 400, p. 136742, 2023.
[12] 芳澤信哉, 坂井勝哉, 太田豊, 上田嘉紀, 香川公司, 岩田章裕. “実走行履歴データを活用した EV 電力消費推定モデルの評価”. In *24.Sep.2021 Joint Technical Meeting on Vehicle Technology/Transportation and Electric Railway*, p. 7–12, 2021.
[13] Xinkai Wu, David Freese, Alfredo Cabrera, and William A. Kitch. “electric vehicles’ energy consumption measurement and estimation”. *Transportation Research Part D: Transport and Environment*, Vol. 34, pp. 52–67, 2015.
[14] Chiara Fiori, Kyoungho Ahn, and Hesham A. Rakha. “power-based electric vehicle energy consumption model: Model development and validation”. *Applied Energy*, Vol. 168, pp. 257–268, 2016.
[15] Wang J, Besselink I, and Nijmeijer H. “battery, electric vehicle energy consumption prediction for a trip based on route information”. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, Vol. 232, No. 11, pp. 1528–1542, 2018.
[16] H. Shen, X. Zhou, H. Ahn, M. Lamantia, P. Chen, and J. Wang. “personalized velocity and energy prediction for electric vehicles with road features in consideration”. *IEEE Transactions on Transportation Electrification*, Vol. 9, No. 3, pp. 3958–3969, Sept 2023.
[17] 植村智明, 能條太悟, 吉瀬雄大, 富井尚志. “解析者の興味に基づく道路区間集計が可能な ev 推定消費エネルギーデータ解析システムの構築と応用”. 情報処理学会論文誌データベース（TOD）, Vol. 14, No. 4, pp. 70–85, 2021.
[18] Takashi TOMII, Shintaro HAGIMOTO, Naoki FUEDA, Toru DEGUCHI, Makoto IDENAWA, and Takuya HAYASHI. “long-term experiment of the ecolog database capability of estimating v2x effect replacing with evs”. In *Proceedings of 20th ITS World Congress Tokyo 2013*, pp. paper3162, 10p, Tokyo, Japan, 2013.
[19] Toshiaki Uemura. “pre-estimation of electric vehicle energy consumption on unfamiliar roads and actual driving experiments”. In *Proceedings of the VLDB 2019 PhD Workshop, co-located with the 45th International Conference on Very Large Databases (VLDB 2019)*, pp. CEUR Workshop Proceedings Vol.2399, paper06, pp.1–4, Los Angeles, 2019.

犯罪および商業施設の情報に基づくゴミ予測手法の検討

羽倉　輝† 山口 琉太† 義久 智樹†† 下條 真司††† 河合 由起子†

† 京都産業大学 〒 603-8555 京都府京都市北区上賀茂本山
†† 滋賀大学 〒 522-8522 滋賀県彦根市場 1 丁 1-1
††† 青森大学 〒 030-0943 青森県青森市幸畑 2-3-1
E-mail: †{i2386102,i2286185,kawai}@cc.kyoto-su.ac.jp, ††yoshihisa@biwako.shiga-u.ac.jp,
†††shinji.shimojo@aomori-u.ac.jp

あらまし 近年，環境保護活動の活発化によって，ポイ捨てゴミの除去や防止のための清掃活動が重要視されている．これまで我々は，自転車に搭載したスマートフォンの内蔵カメラで自転車走行中の道路画像からゴミを抽出することで，自転車移動の付加価値として地点にゴミ情報をアノテーションするシステムを開発した [1]．本研究では，地点に付与されたゴミ情報および環境情報を分析し，ゴミを予測する手法を提案する．本稿では，環境情報として，犯罪および商業施設のオープンデータを用いる．具体的には，まず，教師データとなるゴミの分布情報，犯罪統計と商業施設の情報を収集し，領域を重心ボロノイ分割する．そして，分割領域に対してゴミ分布状況を教師データとして機械学習を行う．本稿では，サンフランシスコの犯罪統計情報と商業施設情報を用いて，提案手法の有効性をゴミの分布予測の精度で評価する．

キーワード 地理データ活用，ポイ捨てゴミ

1 はじめに

環境保護活動の活発化によって，路上のゴミポイ捨て抑止や清掃支援のための取り組みが注目されている．清掃活動は道路を綺麗にするだけでなく，海洋プラスチックゴミ問題の解決にもつながる有意義な取り組みである．しかしながら，街中など生活圏内外でゴミが多く落ちているエリアを見つけることは難しい．一方で，車や清掃車にカメラ等のセンサを装備することで効率的に情報収集するなど，自治体ベースで取り組まれている [2] [3]．本研究では，カメラから得られた画像から効率的に歩道のゴミ情報を収集し，加えて店舗情報や犯罪データからゴミ分布を予測する手法を提案する．これまで我々は先行研究として，自転車に搭載したスマホのカメラから得られた動画から４種類のゴミを抽出し，地点にマッピングするシステムを構築した [1]．これにより，清掃活動に無関心な人でも参加しやすい，自転車を利用したゴミ分布情報の受動的収集ができる．本稿では，地点におけるゴミ情報とその他の店舗や犯罪の地点情報からボロノイ分割領域ごとのゴミ予測を提案し検証する．

提案システムでは，まず路上にあるゴミを抽出するために，撮影された画像の下半分を抽出し，さらに画像を分割することで画像中の物体が占める面積の割合を増加させる．次に，画像からペットボトルや空き缶等のゴミ情報を抽出し，ラベルを付ける．また，ゴミが出やすい場所や状況に関する要因情報を取得し，ゴミの分布とそれら要因との関係を分析して，ゴミ発生を予測する．ゴミの分布と関連がある要因を発見することで，類似する要因の場所でのゴミ分布が予測可能になる．本稿では，ゴミが出やすい場所や状況に関する要因情報は，犯罪統計情報と商業施設情報とした．最後に，ゴミ要因情報と共に予測分布をユーザに提示する．提案システムによる安全で効率的な路上環境データの可視化と予測により，ゴミのポイ捨て等のネガティブな行動の自粛にもつながることが期待される．本研究では，犯罪統計情報と商業施設の地理情報からゴミの分布と最も関連がある地理情報を発見し，ゴミ分布の予測を目指す．

2 章では関連研究について，3 章から 4 章では自転車とスマートフォンを用いた路上環境情報取得・分析システムについて，5 章以降は犯罪および商業施設の情報に基づくゴミ予測手法について説明する．

2 関連研究

2.1 スマートフォンを利用した道路画像収集と分析

総務省の調べ [4] によるとスマホの世帯保有率は令和 3 年度の時点で 8 割を超えており，家庭に広く普及している情報通信機器であると言える．そのためシステムをスマホで実装することは，提案システムを広く普及させる要因となる．その上で，内蔵する CPU やカメラの性能向上に伴ってスマホを画像収集と分析に利用する研究は多い．前田ら [5][1] は自動車に搭載したスマホで動画を撮影し, 画像解析によって道路の亀裂を発見し収集するシステムを開発している．寺野ら [6] も同様に自動車に搭載したスマホで撮影した動画から路面のひび割れを検知している．このようにスマートフォンと画像解析を組み合わせた研究は数多く進められているが，自転車にスマホを搭載し画像解析を行っている研究は少ない．

2.2 ポイ捨てゴミに対する画像分析と分布可視化

環境保護活動の活発化に伴い，路上ゴミに注目した研究が進

1：UrbanX Technologies RoadManager

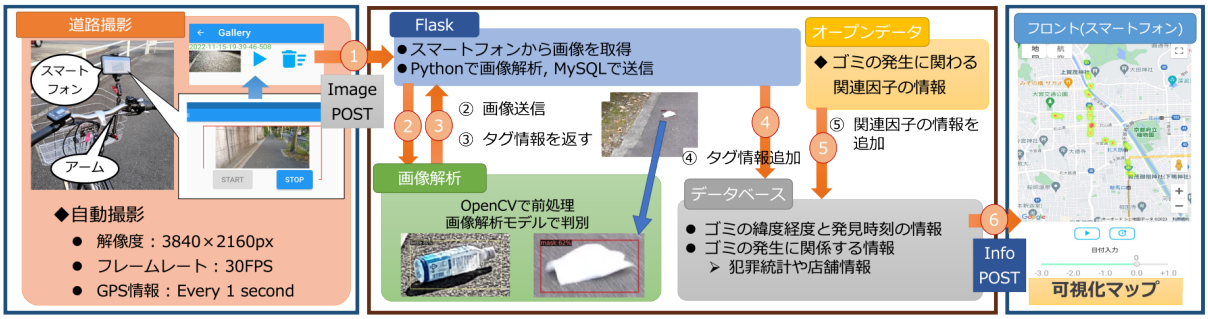


図 1　自転車に搭載したスマホカメラによる路上環境情報取得・分析システムの概要図

められている．近藤ら [7] は，スマホで撮影した画像から機械学習を用いて物体を判別し，撮影位置の座標を地図上にプロットしている．同様に，タカノメ [8] では，スマホで撮影した動画から機械学習を用いて路上に落ちている物体を判別し，撮影位置座標からヒートマップを作成することでゴミ分布を可視化するサービスを提供している．Sio ら [9] は，スマートフォンではなく，YOLO と Raspberry Pi を使用してゴミを自動的に検出するシステムを作成している．Sornlertlamvanich ら [10] は画像分析ではなく，写真を撮影することでゴミに関する情報を共有し，汚染状況を視覚化し，環境意識の向上を目指している．立花ら [11] は，ユーザが装着しているスマートウォッチから得られるセンサデータを用いて，ゴミの種類と位置を推定するシステムを提案しており，ゴミを拾った時に生じる音でゴミの種類を判断している．

2.3　ポイ捨てゴミの発生と関連する要因

糸井ら [12] は日本とマレーシアの大学生のポイ捨てに関する意識の違いをテキストマイニングで分析している．その結果，ポイ捨ての原因として両国ともにゴミ箱の不足を挙げていること，ポイ捨ての解決策として日本では実践活動，マレーシアでは教育が重要だと指摘している．また，ポイ捨ての意識には社会文化的な差異が影響している可能性を示唆している．また森ら [13] はポイ捨てをする人の心理や行動を分析し，ポイ捨てを抑制するための方策を提案している．そしてウェブアンケートを実施し，ポイ捨てに関する態度や行動，実行意図を測定した．その結果，ポイ捨てをすべきではないと考えていても，その場の状況によってはポイ捨てを行なってしまう行動モデルの妥当性を確認した．このようにゴミのポイ捨てには心理的な側面に加えて，ゴミ箱の有無や，ゴミを隠せそうな茂みがあるかなどの地理的な側面からの分析も必要であると考えられる．

3　路上環境情報取得・分析システム

3.1　取得・分析・可視化システムの概要

図 1 に，路上環境情報取得・分析システムの処理の流れを示す．まずユーザは自転車を利用して道路の走行景観を撮影し，撮影した動画と撮影位置情報（緯度経度）をサーバへ送信する（図中 (1)）．サーバは，動画から静止画像を切り出した後，OpenCV による前処理を行う．そして機械学習によって物体へのタグ付けを行い（図中 (2)，(3)），位置情報と関連付けて

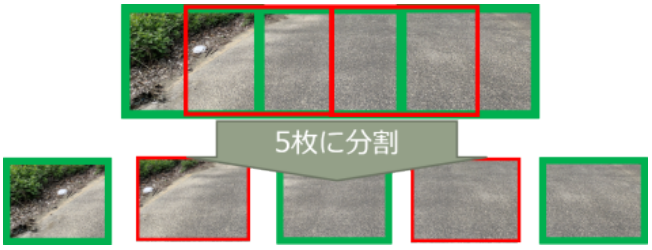


図 2　画像分割の処理の流れ

データベースに保存する（図中 (4)）．並行して，サーバはポイ捨てゴミの関連因子となる犯罪統計，店舗情報といったオープンデータを収集する．そして関連因子の情報を元にゴミの分布予測モデルを構築する．(図中 (5))．最後に，サーバは過去のゴミ分布情報と関連因子の情報を利用してゴミの分布を予測する (図中 (6))．予想結果はユーザの持つスマートフォンアプリへと送信され，現在のゴミ分布と，将来のゴミ分布を可視化する．

3.2　自転車に搭載したスマホカメラによる画像収集

本研究では，多目的に利用できる受動的なデータ収集を目指し，一般的に用いられている自転車とスマホに対応したアプリケーションを Flutter で開発した．具体的な手法について説明する．まず自転車にアームなどの機材を使用してスマートフォンを設置する．次にアプリケーションを起動し，カメラの映像を表示させる．続けて開始ボタンをタップすることで撮影が行われる．動画の保存形式は mp4，解像度は 3840 × 2160px，フレームレートは 30fps とした．またスマホ内蔵の GPS から緯度経度を 1 秒間隔で取得し，撮影時刻と座標情報を csv ファイルで保存している．撮影終了ボタンをタップすると，保存された動画を確認できる．確認後は動画と座標ファイルをサーバへ送信する．送信されたデータはサーバが正しく受信したことを確認したのちスマートフォン上から削除される．また，3.4 節のタグ付け手法の解析結果を webview で確認でき，過去のゴミ分布から将来のゴミ分布までをタイムラインで閲覧できる．

3.3　画像解析のための前処理

スマホから送信された動画はサーバに転送され，動画から静止画を作成する．ここでは FFMpeg ライブラリを利用し，1 秒当たり 2 枚の静止画を作成する．位置座標の情報は，撮影開始時刻，再生時間，走行中のタイムスタンプ (csv で保存) を対応付けることで付与する．

表 1　自作ゴミ判別モデルの開発環境

Python3 Google Compute Engine バックエンド (GPU)
Ubuntu 18.04.6 LTS
RAM 12.68GB ディスク 78.19GB
CPU Intel(R) Xeon(R) CPU @ 2.00GHz
GPU NVIDIA Tesla T4 CUDA Version: 11.2

（1） 画像の上半分をトリミングし削除
（2） 左右・中央の 3 枚とそれらの中間 2 枚の計 5 枚に分割
画像分割の処理の流れを図 2 に示す．画像の上半分は，進行方向により道路の撮影領域が少なくなり，ゴミを判別できる解像度に達しないため削除する．次に左右・中央の 3 枚とそれらの中間である 2 枚を合計 5 枚に分割する．これにより画像中の物体が占める面積の割合を増加させ，画像解析の精度を向上させる．

3.4　機械学習によるゴミの判別

本稿では，Detectron2 [14] を用いる．Detectron2 Model Zoo では様々な学習済みモデルが提供されているが，本研究では Fast R-CNN R50-FPN を利用する．またポイ捨てゴミを再学習させるための訓練データとして，TACO Dataset [15] を利用する．データセットに付属しているラベル付けを利用することもできるが，雑草に覆われて見えないものや，ゴミの種類を判別できないものなども含まれるため，本研究では，データセットに付属しているラベル付けは使用せず，ペットボトル，空き缶，食品トレイ，マスクの 4 種類を，ポイ捨てゴミを含む画像として選出し，ラベル付けを行い訓練データとして利用する．以上のように構築したゴミ判別モデルを利用して，3.3 節の手順で作成した画像に対してタグ付けし，結果をタグごとにデータベースへ保存する．

4　路上環境情報取得・分析システムの検証

4.1　ゴミ判別精度の検証

4.1.1　検 証 目 的

本稿では 3.4 節の自作ゴミ判別モデルを利用して，ポイ捨てゴミの判別精度を検証することを目的とし，画像中の物体に対してポイ捨てゴミがポイ捨てゴミであると適切にタグ付けができているかを再現率及び適合率として求める．

4.1.2　自作ゴミ判別モデルの開発環境

今回の検証で Detectron2 によるゴミ判別を実行するために利用した開発環境を表 1 に記す．このようにハードウェアアクセラレータを GPU に設定した Google Colaboratory[2]を利用して Detectron2 による判別モデルを実行させている．

4.1.3　検証の実施条件と機材設定

次に検証時の条件について説明する．今回の検証では画像収集をキャンパス内の道路で実施している．これは一般道での画像収集では画像解析の精度を検証するために必要なゴミが写った画像を用意することが難しいと判断したためである．また本研究では安全な走行による画像収集を前提としているため，雨天時や，無灯火走行による画像収集は考慮しない．また図 3 は検証に利用したポイ捨てゴミである．実験に利用したスマートフォンは Android 端末[3]を利用した．検証時の自転車の平均速度は 14.4km/h であり，これは一般的な自転車の平均速度と同程度である．

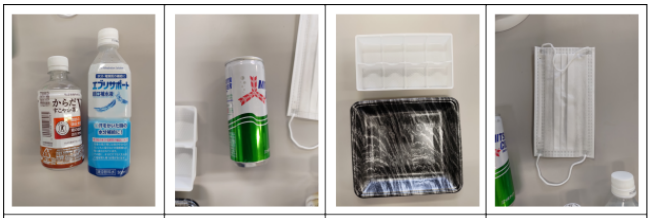


図 3　検証に利用したポイ捨てゴミ

4.1.4　判別精度の算出方法

次に画像解析結果の各検証項目について説明する．まず画像解析の精度を比較するために本研究では，再現率及び適合率を算出する．それぞれの値を算出するための変数の条件を以下に示す．

（1） TP: ゴミ有りをゴミ有りと判定した総数
（2） FN: ゴミ有りをゴミ無しと判定した総数
（3） FP: ゴミ無しをゴミ有りとして判定した総数
（4） TN: ゴミ無しをゴミ無しとして判定した総数

TP は画像中のゴミにタグ付けができた結果が該当する．次に FN は画像中のゴミにタグ付けができなかった結果が該当する．続けて FP は画像中のゴミではない領域に対して 1 つ以上タグ付けを行った結果が該当する．最後に TN はゴミを含まない画像に対して，タグ付けを行わなかった結果が該当する．以上を踏まえた上で，再現率，適合率を算出した．

4.2　ゴミ判別精度の検証結果

再現率は 49.8%と低い割合となった．一方で適合率は 82.0%となった．ゴミの種類ごとの結果を見るとペットボトルや空き缶は良好な結果となった．一方で食品トレーとマスクは色や形が似ていることから半数以上が識別できない結果となった．自作モデルの再現率，適合率の改善策としては，訓練データをより多くのデータセットから取得することや，クラス分類の見直し，異なる学習モデルの利用などが考えられる．

5　犯罪および商業施設の情報に基づくゴミ予測手法

5.1　可視化・予測システムの概要

ここまで実際にあったポイ捨てゴミの情報を収集することで，ゴミの分布を可視化するという，直接的なアプローチを説明した．しかし，このシステムはユーザが自転車で走行して画像を集めるものであるため，調査できるエリアや時間に制約がある．そのため，ゴミの分布が不明なエリアが存在するという問題がある．そこで本稿では，ポイ捨てゴミの発生に影響を与え

2：https://colab.research.google.com/?hl=ja

3：Asus Zenfone8

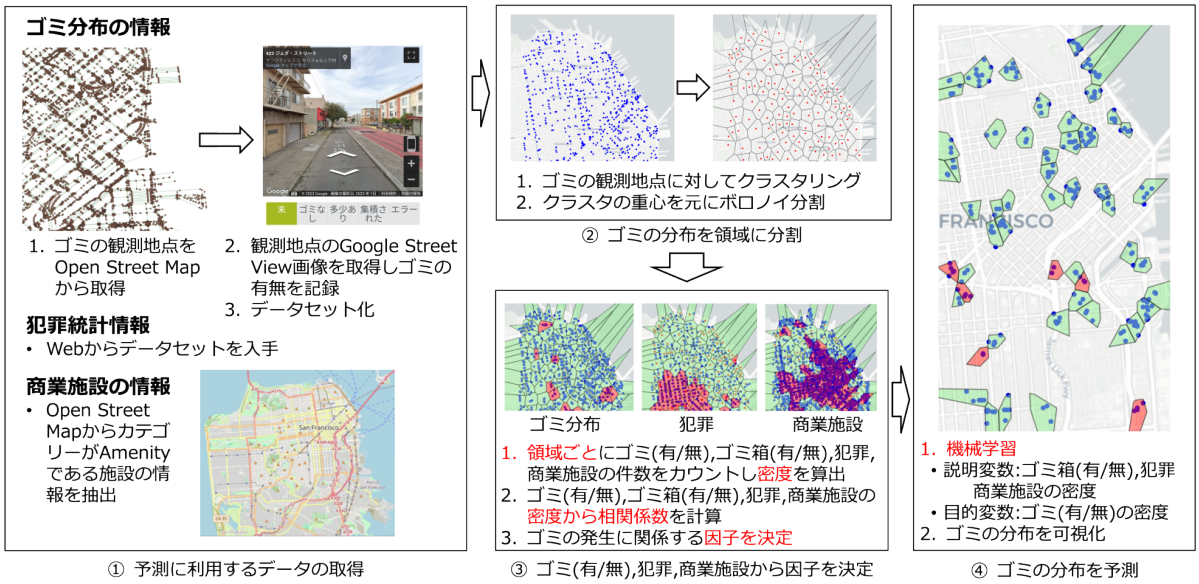


図 4　犯罪および商業施設の情報に基づくゴミ予測手法の概要図

ると考えられる情報をもとに，ゴミの分布を推定するという間接的なアプローチを提案する．このアプローチにより，ゴミの分布調査ができていないエリアでも，ゴミの分布の傾向を把握できる．図 4 に提案手法の概要を示す．この手法では，オープンデータと Google Street View 画像から，ゴミの発生に影響を与える情報を抽出し，分析する．提案手法の手順は以下の通りである．

（1） 予測に利用するデータの取得
（2） ゴミの分布を領域に分割
（3） ゴミ (有/無), 犯罪, 商業施設の密度を計算
（4） ゴミの分布を予測

まず教師データとなるゴミの分布情報とポイ捨てゴミの関連因子となる犯罪統計や店舗情報といったオープンデータを収集する．次にゴミの分布をクラスタリングとボロノイ図を利用して領域に分割する．続けてゴミ (有/無), 犯罪, 商業施設の密度を計算し，相関係数を算出する．最後に関連因子の密度を説明変数とし，ゴミの有無の密度を目的変数として機械学習を実施する．

本稿ではゴミ分布の予測対象としてサンフランシスコを取り扱う．サンフランシスコは，人口密度が高く，観光客も多い都市であるため，ポイ捨てゴミが多いと考えられる．さらに，サンフランシスコは，犯罪統計や店舗情報など，様々なオープンデータを公開しており，これらのデータはゴミの発生に影響する情報として有用である．また本研究で使用する DBSCAN と回帰モデルには scikit-learn を利用している [4]．以下に提案手法の詳細な内容について説明する．

5.2　予測に利用するデータの取得

5.2.1　教師データとなるゴミ分布の調査

まず教師データとなるゴミの分布を調査する方法について説明する． ゴミの分布を調査するために，現地を調査し正解ラベルを作成することは困難である．そこで本研究では，Google Street View(GSV) の道路画像からポイ捨てゴミの有無を観測することでデータセットを作成する．まずポイ捨てゴミの観測地点を Open Street Map から取得する．観測地点は道路の頂点や中点を利用している．次に観測地点の座標を GSV の API に入力することで GSV 画像を取得する． GSV 画像は，最新のものを優先して取得している． また，取得する画像は指定した座標から 15m 以内で最も近い位置にある画像である． 加えて Google 公式が収集した画像のみを利用することで屋内の画像を可能な限り除外している． また画像が存在しない場合や，屋内の画像である場合は「エラー」として取り扱う．



図 5　教師データ作成のためのツール

本研究では，ゴミの有無を調査するために，図 5 のチェック

4：https://scikit-learn.org/stable/index.html



図 6　GSV 画像から得られるゴミの例



図 7　GSV 画像から得られるゴミ箱の例

表 2　ゴミ分布情報の収集状況

重複	合計	ゴミ無し	多少あり	集積された	エラー
有り	2,047	1,412	247	61	327
無し	1,705	1,334	238	61	72

ツールを Google Street View の API を使って作成した．チェックツールで収集する主な情報は，位置情報，ゴミの量，ゴミ箱の有無，撮影日などである．また画像の重複チェックや復元のために画像 ID も取得している．ゴミの量の基準は以下のように定義した．

（1） ゴミ無し：ゴミがなく綺麗な状態

（2） 多少あり：ペットボトルや缶，ゴミ袋がいくつか落ちている状態を示す．ゴミの個数が 2 から 3 個の場合に該当

（3） 集積された：ゴミがまとめて捨てられ，長期間に渡って清掃されていないような状態を示している．ゴミの個数が 4 個以上の場合に該当

実際のゴミの例を図 6 に示す． 多少ありは，3 個程度のゴミが落ちている状態を示しており赤マルで囲まれた箇所にゴミが落ちていることがわかる． 集積されたは，ゴミがまとめて捨てられている状態を示しており赤マルの中に大量のゴミが落ちていることがわかる． 図 7 はゴミ箱の例であり，サンフランシスコのゴミ箱の多くはこの形をしている．

次に実際にチェックツールを用いて収集したデータセットについて説明する．表 2 は現時点でのゴミ分布の調査状況を示している．重複は画像 ID の重複の有無を意味している．GSV から画像を取得した際に入力した座標が異なっているにも関わらず同じ画像が出力されることがある．そのため本研究ではデータセットを読み込んだ際に画像 ID の重複を無くす処理を行なっている．また重複有りの結果においてエラーは 327 件となっており，これは観測地点の全体の約 16%においてゴミの有無を確認できなかったことを意味している．加えて重複無しの結果において，ゴミが落ちていなかった地点の割合は 78%であり半数以上の地点でゴミが落ちていない結果となった．反対にゴミが落ちていた地点の割合は 18%であり低い割合となった．

5.2.2　関連因子となるオープンデータの収集

次に関連因子となるオープンデータの収集について説明する．サンフランシスコで扱えるオープンデータとしては，犯罪統計情報と商業施設情報がある．犯罪統計は，犯罪の種類や発生件数を示すデータであり，犯罪が多い地域では，ポイ捨てゴミも多くなると考えられる．店舗情報は，店舗の種類や位置を示すデータであり，店舗が密集している地域では，人の流れや消費活動が活発であり，ポイ捨てゴミも発生しやすいと考えられる．本研究では，これらのデータをポイ捨てゴミの関連因子として調査する．

次に利用したサンフランシスコの犯罪統計情報について説明する．今回はサンフランシスコ警察の事件報告データセットを利用している [5]．また今回の検証では 2023 年に発生した 62,063 件の事件を取り扱う．様々なカテゴリーの犯罪データが含まれている．そのため今回は類似する事件をグループに分けて分析を行った．以下にグループの内容について示す．

（1） 盗難, 物損, 放火：Robbery,Arson など

（2） 薬物, 武器犯罪：Drug Offense,Weapons Offense など

（3） 交通事故, 違反：Traffic Collision など

（4） 暴行, 性犯罪:Assault,Rape など

（5） 詐欺, 偽造:Fraud など

盗難, 物損, 放火グループはデータセットにおいて Robbery,Arson などのカテゴリーに属していた事案をまとめている．交通事故, 違反グループも同様に，Traffic Collision,Traffic Violation Arrest などのカテゴリーに属していた事案をまとめている．

続けて今回利用した商業施設の情報について説明する．今回は Open Street Map からカテゴリーが Amenity である施設の情報を抽出したものをデータセットしている．また今回の検証では 5,933 軒の施設を取り扱う．犯罪データと同様に様々なカテゴリーに分けられており，今回は類似するカテゴリーを以下のグループに分けて取り扱った．

（1） 交通:car rental,bicycle rental,parking など

（2） 娯楽:events venue,nightclub,cinema など

（3） 教育:university,school など

（4） 医療:clinic,doctors,pharmacy など

（5） 食事:bar,grocery,vending machine など

（6） その他商業施設:bank,storage,coworking space など

（7） 公共施設:social facility,library,police など

（8） ゴミ関係:waste basket,recycling など

交通グループはデータセットにおいて car rental,bicycle rental,parking などのカテゴリーに属していた施設をまとめている．食事グループも同様に bar,grocery,vending machine などのカテゴリーに属していた施設をまとめている．

5.3　ゴミの分布を領域に分割

次にゴミの分布を領域に分割する手順について説明する．図 8 に流れを示す．まず教師データ, テストデータとなるゴミの分

5：https://www.sanfranciscopolice.org/

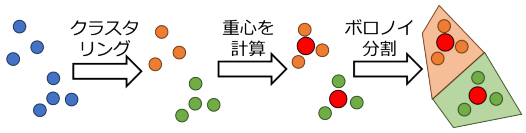


図 8　領域分割の流れ

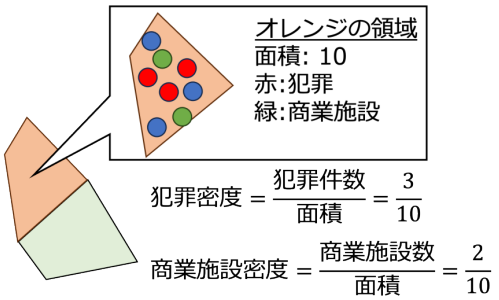


図 9　密度計算の例

布に対してクラスタリングを行いクラスタを作成する．クラスタリングには DBSCAN を利用している．次に各クラスタの重心を計算する．最後に重心を元にボロノイ分割を行う．

5.4　ゴミ (有/無), 犯罪, 商業施設の密度計算

次にゴミの発生密度と関連因子の発生密度の計算と相関の調査手法について説明する．まずはゴミの発生密度について求める．ここでは「ゴミ有り」と「ゴミ無し」の 2 値に分けてそれぞれ密度を計算する．よって「ゴミ有りの密度」は領域内でゴミがあった地点の合計を領域の面積で割ったものになる.「ゴミ無しの密度」でも同様の手法で計算する．また「ゴミ箱有り」と「ゴミ箱無し」の密度も同様の方法で計算する．関連因子の発生密度の計算手法も同様に，あるゴミ分布領域内での発生件数を領域の面積で割ることで計算する．図 9 の計算の例では，犯罪 (赤丸) が 3 件あり面積が 10 であるため，犯罪密度は 0.3 となる．また本研究では 1 平方 km あたりの密度として計算している．最終的に全ての領域に対してゴミの発生密度と関連因子の発生密度を計算する．

続けてゴミの発生密度と関連因子の発生密度との相関について説明する．本研究ではどのような事象がゴミの発生と関係しているかを調べるために関連因子のグループごとにピアソン相関係数を算出した．ゴミの発生密度と，犯罪密度, 商業施設密度との相関係数をグループ別に検証した結果を表 3 から表 5 に示す．結果から犯罪密度の場合，交通事故及び違反のグループでゴミ有り密度に対して弱い正の相関を示している．しかし最も大きい相関が 0.258 であるように，全体的に相関係数が小さいため，今回の結果ではゴミの発生と犯罪の発生に相関がないと考えられる．商業施設密度の場合，全体的に「ゴミ無し」と弱い正の相関があることがわかる．例えば，ゴミ関係グループは 0.371, 食事グループでは 0.318, 交通グループでは 0.304 を示している．このことから商業施設の数が多いほどゴミの発生が少なくなっていると考えられる．またグループごとの結果では「ゴミ関係」と「ゴミ無し」の相関が最も高い正の相関を示している．このことから，ゴミ処理に関わる施設があるとゴミ

表 3　カテゴリ別の犯罪密度との相関係数

	ゴミ有り密度	ゴミ無し密度
グループ指定なし	0.031	0.138
盗難 物損 放火	-0.036	0.239
薬物 武器犯罪	-0.013	-0.007
交通事故 違反	**0.258**	0.013
暴行 性犯罪	0.055	0.102
詐欺 偽造	0.073	0.109

表 4　カテゴリ別の商業施設密度との相関係数

	ゴミ有り密度	ゴミ無し密度
全体	-0.092	0.398
交通	0.014	**0.304**
娯楽	0.063	0.014
教育	0.005	0.061
医療	-0.050	0.278
食事	-0.060	**0.318**
その他商業施設	-0.080	**0.315**
公共施設	-0.006	0.087
ゴミ関係	-0.079	**0.371**

表 5　ゴミ (有/無) 密度とゴミ箱 (有/無) 密度の相関

	ゴミ有り密度	ゴミ無し密度
ゴミ箱有り密度	-0.021	**0.664**
ゴミ箱無し密度	0.426	0.808

が少なくなる傾向があると考えられる．ゴミ箱 (有/無) 密度の場合,「ゴミ箱あり」に対して「ゴミ無し」が 0.664 と正の相関があることが分かる．このことからゴミ箱が多いほどゴミの発生が少なくなっていると考えられる．一方で「ゴミ箱なし」は「ゴミ有り」に対して 0.426 と正の相関を示している．これは教師データ作成の際に，ゴミ箱からあふれたゴミが散乱している地点があったことが影響していると考えられる．また「ゴミ箱なし」は「ゴミ無し」に 0.808 と強い正の相関を示している．これはゴミ箱が無くてもゴミが捨てられない状況を反映していると考えられる．

またゴミ (有/無) 密度,「交通事故 違反」グループの犯罪密度,「交通」「食事」「その他商業施設」「ゴミ関係」密度の度数分布をヒストグラムに示す．図 10 の結果から，全体的にゴミ有り密度は小さな値を取っている反面，ゴミ無し密度はより大きい値を取っていることがわかる．このことからゴミが有る領域よりもゴミが無い領域の方が多いことが分かる．次に図 11 の結果から犯罪密度の最頻値は最も低い密度であることを示しており，領域の大部分は低犯罪域になることが分かる．続けて図 12 の結果から商業施設密度でも同様に最頻値は最も低い密度であることを示しており，多くの領域は商業施設の数が少ないことが分かる．

次に実際の分布を図 13 に示す．ゴミの分布については，領域内でゴミ無しが占める割合が高いものが緑色，低いものが赤色で示されている．また犯罪密度と商業施設密度の色分けはそれぞれ，平均値未満が緑色，平均値以上は赤色で示している．図 13 より，ゴミが有る領域は南部全体に，高犯罪領域は南西

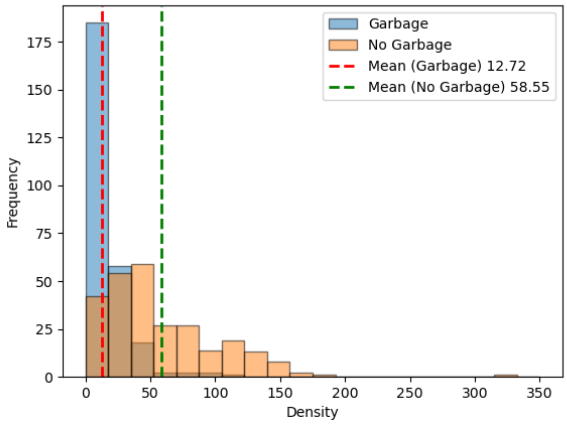


図 10　ゴミ (有/無) 密度のヒストグラム

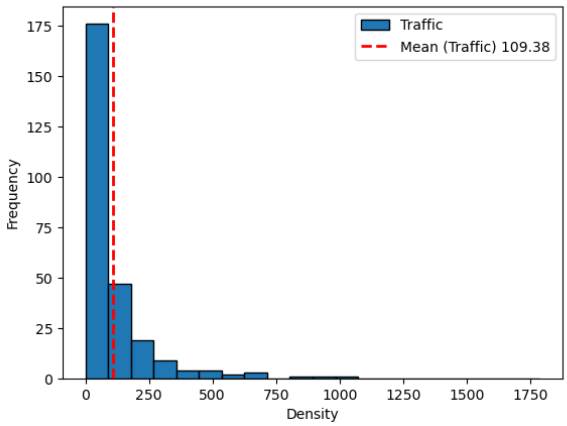


図 11　犯罪密度のヒストグラム

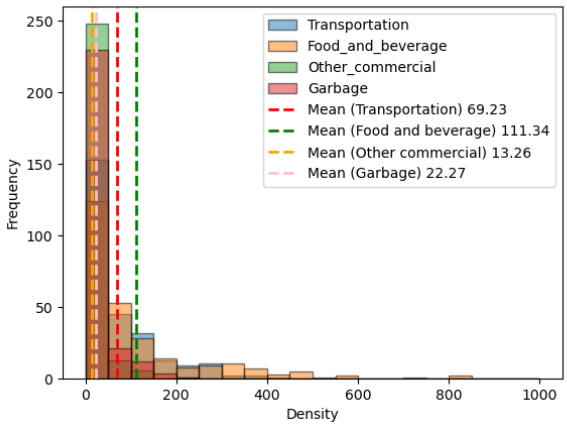


図 12　商業施設密度のヒストグラム

部に集中，商業施設が多い領域は中部, 東部に集中していることがわかる．またゴミの分布と商業施設の分布を比べると，ゴミが多い場所は商業施設が少なくなるような分布になっている．

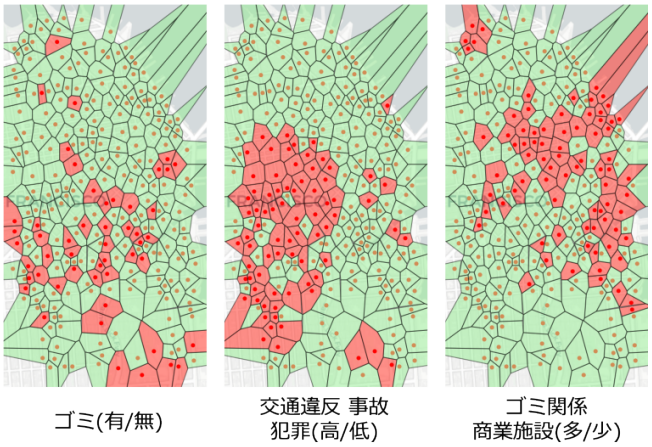


図 13　密度に基づく分布マップ

モデル	合計	TP	FN	FP	TN
LinearRegression	67	1	6	11	49
SVM	67	0	7	0	60
RandomForest	67	2	5	4	56

表 6　各モデルの検証結果 (1/2)

モデル	正答率	再現率	適合率
LinearRegression	0.75	0.14	0.08
SVM	0.90	0.00	0.00
RandomForest	0.87	0.29	0.33

表 7　各モデルの検証結果 (2/2)

5.5　ポイ捨てゴミの分布の予測

最後に相関係数の検証結果に基づいてゴミ分布の予測を行う．今回は精度比較のために，線形回帰，SVM，ランダムフォレストの 3 つの回帰モデルを利用した．いずれのモデルも python の sklearn を利用して実装している．説明変数はゴミ箱 (有/無) 密度，犯罪密度 (交通事故 違反)，商業施設密度 (交通/食事/その他商業施設/ゴミ関係) とした．また目的変数はゴミ (有/無) 密度とした．またモデル間の精度を比較するために本研究では，正答率，再現率，適合率を算出する．それぞれの値を算出するための変数の条件は 4.1.4 節の条件と同様である．

表 6 と表 7 に各モデルの検証結果を示す．また図 14 に予測結果を可視化したものを示す．LinearRegression と RandomForest では TP に分類できた領域がある．一方で SVM は TP に分類できた領域がない反面，FP に分類された領域はない．正答率については SVM と RandomForest が高い結果となった．また再現率と適合率については RandomForest が最も高い結果となった．今回の検証ではテストデータが少なかったため今後は教師データ，テストデータを増やすように取り組む．また本稿で用いた関連因子ではゴミ有り密度に対して強い正の相関を示していなかったため，TP に分類される領域が少なくなったと考えられる．そのためゴミが有る領域を予測しやすくするために，今回取り扱った関連因子よりも高い関係性を持つ環境情報を追加する必要があると考えられる．

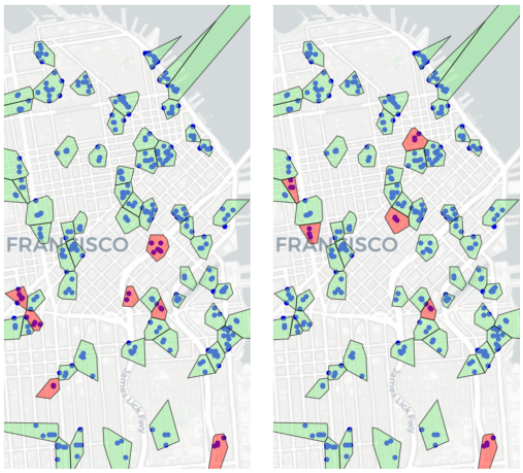


図 14　ゴミ分布の予測

6　改善点と今後の展望

今回の検証では，教師データが少なかったことに加えて，本稿で用いた関連因子ではゴミの有無に対して強い相関を示していなかったため，ゴミ有りとして分類される領域が少なくなったと考えられる．そのためゴミが有る領域を予測しやすくするために，今回取り扱った関連因子よりも高い関係性を持つ環境情報を追加する必要があると考えられる．現在検討している候補は，観光スポットの有無と混雑度である．観光スポットは人目が多い場所でありかつ清潔さが重要視されるため，ゴミが無い領域に対して相関を示すのではないかと考えられる．加えて混雑度が高い場合，人込みや渋滞に紛れてポイ捨てをする人がいることでゴミが有る領域に対して相関を示すのではないかと考えられる．また今回の提案手法で利用したボロノイ分割では地形を考慮できていないため，今後はネットワークボロノイの利用を検討する．

7　おわりに

本研究では，地点に付与されたゴミ情報および環境情報を分析し，ゴミを予測する手法を提案した．そして環境情報として犯罪統計と商業施設の情報を収集し，ゴミの分布との関係性を検証した．その結果，犯罪統計情報についてはゴミ有り密度に対しては弱い正の相関しか確認できず，商業施設情報についてはゴミ無し密度に対して正の相関があるという結果になった．今後は新たな関連因子の追加とネットワークボロノイの利用によって，ゴミ分布予測の精度向上を目指す．

謝　辞

本研究の一部は，科研費基盤研究 (課題番号：22H03700, 20H00584, 19K12240, 20H04293) の研究活動による．ここに記して謝意を表す．

文　献

[1] Hikaru Hagura, Ryuta Yamaguchi, Tomoki Yoshihisa, Shinji Shimojo, and Yukiko Kawai. A proposal of acquiring and analyzing method for distributed litter on the street using smartphone users as passive mobility sensors. In *ACM SIGGRAPH 2023 Posters*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery.

[2] 三上 量弘, 河野 慎, 陳 寅, and 中澤 仁. 深層学習を用いたゴミ収集動画からゴミ袋個数の測定. **人工知能学会全国大会論文集**, JSAI2018:2A201–2A201, 2018.

[3] 泰恵 岸野, 良成 白井, 伸 水谷, 孝 竹内, 敬之 須山, 太 納谷, and 修功 上田. ゴミ収集車に搭載したセンサを用いた状況識別による地区別ゴミ量推定と地区特性の分析. Technical Report 15, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, NTT コミュニケーション科学基礎研究所, aug 2019.

[4] 総務省. 情報通信白書令和 3 年版.

[5] 前田 紘弥, 関本 義秀, 瀬戸 寿一, 樫山 武浩, and 小俣 博司. 機械学習とスマートフォンを用いた道路の損傷画像のリアルタイム検出と修繕対応基準における各特徴量の重要度比較. **交通工学論文集**, 4(3):A1–A8, 2018.

[6] 寺野 聡恭, 小金丸 暁, 松田 浩, 佐々木 博, 古賀 掲維, and 西川 貴文. スマートフォンを用いた道路舗装の平坦性及びひび割れ率の評価に関する研究. **土木学会論文集** *F4*（**建設マネジメント**）, 75(2):I88–I95, 2019.

[7] 近藤諒太 and 清木康. 不法投棄ゴミを対象とした画像分析・分類機能と時空間マッピング・ システムの実現方式. DEIM Forum 2021, J11-5, 2021.

[8] 株式会社ピリカ. ポイ捨てごみ調査サービス 「タカノメ」—スマホで簡単、ポイ捨て分布調査.

[9] Gilroy Aldric Sio, Dunhill Guantero, and Jocelyn Villaverde. Plastic waste detection on rivers using yolov5 algorithm. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2022.

[10] Virach Sornlertlamvanich, Thatsanee Charoenporn, Shiori Sasaki, and Yasushi Kiyoki. A framework for developing social awareness of the effect of plastic waste on the environment. In *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pages 225–229, 2022.

[11] 巧樹 立花, 優吾 中村, 裕貴 松田, 博彦 諏訪, and 慶一 安本. スマートウォッチを用いたポイ捨てごみの種別・位置認識システムの提案. In *2020* **年度 情報処理学会関西支部 支部大会 講演論文集**, volume 2020, sep 2020.

[12] 糸井 風音, 甲斐田 直子, 岩本 浩二, Azrina Sobian, Norhayati Abdullah, and 水野谷 剛. 路上ポイ捨ての理由と解決方策に関する認識比較分析：日本・マレーシア大学生を対象に. **環境情報科学論文集**, ceis36:256–261, 2022.

[13] 森 康浩, 中俣 友子, 桑山 りさ, and 大沼 進. ポイ捨てに特化した行動モデルの構築. **廃棄物資源循環学会研究発表会講演集**, 33:19, 2022.

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[15] Pedro F Proença and Pedro Simões. Taco: Trash annotations in context for litter detection. *arXiv preprint arXiv:2003.06975*, 2020.

EVの長期間ライフログに基づくバッテリー損耗のモデル化とVGIシミュレーションによる考察

片平　昇輝† 　廣居　樹† 　鈴木　博登† 　小松虎太郎†† 　本藤　祐樹†††
富井　尚志†††

† 横浜国立大学大学院環境情報学府情報環境専攻　〒240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
†† 横浜国立大学理工学部数物・電子情報系学科　〒240–8501 神奈川県横浜市保土ケ谷区常盤台 79–5
††† 横浜国立大学大学院環境情報研究院　〒240–8501 神奈川県横浜市保土ケ谷区常盤台 79–7
E-mail: †{hiroi-tatsuki-cf,katahira-shoki-gm,suzuki-hiroto-rs}@ynu.jp, ††{komatsu-kotaro-mk}@ynu.jp, †††{hondo,tommy}@ynu.ac.jp

あらまし　電気自動車（EV）をスマートグリッドのバッテリーとして用いることで，電力の需給バランスを調整することをVGI（Vehicle Grid Integration）という．VGIにはEVのバッテリーに追加の損耗を生じさせるという問題がある．我々はこれまで実際のEVを用いて，走行だけでなく建物への給電を行うなど，VGIに近い形で3年間にわたりデータを収集してきた．本研究ではこれらのデータを利用し，バッテリー損耗モデルを導出した．そしてVGIシミュレーションに対してバッテリー損耗モデルを適用し，充給電を行ったEVのバッテリーがどれだけ損耗するのかを定量化した．その結果，VGIに参加したことによる追加のSOH（バッテリー健全度）の低下は年1％ポイント程度であることが分かった．また，VGIに参加するEVが多いほど一台当たりの損耗が抑えられることがわかった．

キーワード　ライフログ，太陽光発電，負荷平準化，スマートグリッド，VGI: Vehicle-Grid Integration，再生可能エネルギー，電気自動車，オープンデータ

1 はじめに

センサ技術やストレージ技術の発展により，現実世界の膨大かつ多種多様なデータの取得・蓄積が可能になった [1]．中でも，スマートデバイスやIoT（Internet of Things）の普及により，個人の生活の記録である「ライフログ」が容易に利用できるようになった．また，国や地方公共団体がオープンデータに関する取り組みを進めたことにより，気象データなどの様々な場所固有のデータが利用できるようになった．これらのライフログやオープンデータを充分に活用するためには，適切なデータ管理が重要である [2]．

一方で，地球温暖化対策として，温室効果ガスの一種であるCO_2の排出を抑えつつ，効率的にエネルギーを利用していくことが求められている [3] [4]．東京電力グループの「長期的な安定供給とカーボンニュートラルの両立に向けた事業構造変革について」[5] によれば,「カーボンニュートラル社会を実現していくためには，相対的に安価な需要側の取り組み (太陽光・電化・蓄電池の設置等) が有効である」とされている．

需要側の取り組みのうち太陽光に関して，日本では再生可能エネルギーの中でも特に太陽光発電（PV：Photovoltaic power generation）が多く導入されており，導入量は2018年時点で世界第3位である [6]．しかし，PVの導入に伴う課題が存在する．まず，PVの発電量は季節や気象に依存し不安定である．そのため，発電電力を需要電力に一致させることが難しく，消費しきれずに余ってしまう「発電余剰」が生じることがある．さらに，PVを導入しても太陽が夕方に沈んでしまうため，夕方以降の需要のピークに対応できず，実質電力需要（太陽光発電電力と需要の差）のピークはPV導入前とかわらない．また，PVの導入量が増えると，発電する時間帯としない時間帯で電力供給側に対する実質電力需要に急激な変動が生じ（ダックカーブ現象），電力供給側の発電効率に悪影響を及ぼすことが懸念される [7]．

次に，需要側の取り組みのうち電化に関して，世界的に電気自動車 (EV) へのシフトが始まっている．それに伴い，バッテリー技術の研究が進んでいる．バッテリーの大容量化や全個体電池の研究によって，将来はさらなる容量増加が予想される．しかし，EVシフトにも課題が存在する．EVによるカーボンニュートラルは電源の脱炭素化が前提となっていて，EVの走行電力が再生可能エネルギー由来でないと意味がないことである．また，EVが新たな電力需要を生じさせる．

これらのPV・EVの問題に対し，PVとEVを統合する新しいスマートグリッドであるVGI(Vehicle-Grid-Integration) によって両者の問題を同時に解決する試みがある [4]．VGIではEVのバッテリーを活用して効果的に充給電を行う．さらに，電力の需要と供給のマッチングを図ることで，太陽光発電電力を有効活用することが期待される．それに加えて，実質電力需要の変動を抑えること（負荷平準化）[4] も実現可能である．

ただし，EVを用いたエネルギーの時間移動と空間移動を実現するためにはデメリットが存在する．バッテリーの充給電に伴いエネルギー変換ロスが生じることや，バッテリーの損耗が加速したりすることが知られている [8]．個人のEVユーザに

とって，EV のバッテリーが損耗することは VGI を実現するうえでの大きな懸念点となる．再生可能エネルギーと EV バッテリーを組み合わせた負荷平準化を実現するために，現実的なデータに基づいて EV バッテリーがどれだけ損耗するのかを予測することは，EV ユーザにとって非常に有用な情報である．

本研究では EV バッテリーの損耗モデルを構築し，VGI シミュレーションを行った．EV バッテリーの損耗は一台一台異なるものであり，それぞれの損耗量を定量化する必要がある．しかし，EV の使い方は個々によるものであり，バッテリーの損耗の仕方も異なるものとなるため，正確な EV バッテリーの損耗モデルを構築することは難しい．そこで基準となるバッテリー損耗モデルを構築する．そしてその損耗モデルが正しいと仮定し，VGI シミュレーションに適用する．そして VGI に参加することによる個々の EV バッテリーの損耗量や，PV や EV が増えることによる EV バッテリーの損耗量を定量化する．

2 関 連 研 究

本研究における関連研究として主に以下の 5 つが挙げられる．

（1） 電力供給側のスマートグリッドに関する研究

（2） 電力需要側のスマートグリッドに関する研究

（3） VGI/V2G に関する研究

（4） バッテリー損耗に関する研究

（5） 負荷平準化 VGIDB に関する先行研究

スマートグリッドに関する研究は，大別して電力供給側の立場から設計する研究と電力需要側の立場から設計する研究が見られる．

2.1 電力供給側のスマートグリッドに関する研究

電力供給側の立場から，大規模な需要変動に対して周波数変動を抑制することを目的としてスマートグリッドを設計する研究が行われている．Zhang ら [9] は，カリフォルニア州を対象として，供給側の立場から EV の大量導入に伴う電力系統負荷の増大という課題に対して，EV を管理充電することの利点を様々な評価軸から示した．Zhang ら [10] もまた，供給側の立場から，アメリカ中西部地域を対象に 2020 年から 2040 年にかけての EV の増加に伴う電力系統負荷の増大をシミュレーションから示した．加えて，非制御充電と比較して，管理された充給電による V2G が効果的であることを示した．小嶋ら [11] は，供給側の観点から，戸建て需要をまとめることでより効率的に電力融通ができることと，PV・EV・蓄電池を導入することでエネルギーコストが低減可能であることを示した．

これらの研究は電力供給側の立場から大規模なスマートグリッドのあり方を議論したものである．一方で，我々の研究は電力需要側の立場からスマートグリッドを考えるという点でこれらの研究と立場が異なる．

2.2 電力需要側のスマートグリッドに関する研究

前節と異なり電力需要側の立場から，スマートグリッドを設計する研究が行われている．Narottam ら [12] は，住居用負荷を最適に管理する V2G シミュレーションを行い，容量の大きい EV をバッテリーとして活用することでピーク負荷の削減が可能であることを示した．Arab ら [13] は EMS（エネルギー管理システム）と PV 及び EV を各家庭に設置し効果的に運用することで，外部電力及び電気料金の削減が可能であることを示した．また，再生可能エネルギーとバッテリーを組み合わせたスマートグリッドに関する研究は，多種多様な分野から学際的に行われている．横浜市・豊田市・けいはんな・北九州市の 4 都市 [14] では，EMS や蓄電池，デマンドレスポンスを組み合わせたスマートシティプロジェクトの実例を挙げ，供給側への影響を評価した．

これらの研究では需要体として各家庭を想定している．本論文では通勤先を需要体として想定したスマートグリッドでのシミュレーションを行う点が異なる．

2.3 VGI/V2G に関する研究

EV は非走行時に電力を電力網/家に供給すること（V2G: Vehicle to Grid/V2H:Vehicle to Home）ができ，活用が期待される．これらの実験や効果検証に関する研究が近年多く見られる．V2G の導入効果の検討として Kempton ら [15] は，V2G をアメリカの電力市場に投入した際にどの程度の効果と利益が得られるのかの検証を行った．Jansen ら [16] は複数の EV 群を 1 つの仮想的な発電所としてみなす EV-VPP（EV Virtual Power Plant）を提案し，V2G が果たす役割について論じた．中川ら [17] は，PV と EV，電力を消費する建物を構成要素とし PV 電力を自己消費するバーチャルグリッドを想定し，エネルギー消費を統合管理することの優位性を定量的に評価した．大澤ら [18] は，EV 単独で使用するよりも EV と V2H を組み合わせることで CO_2 排出量を多く削減できることと，EV と V2H を組み合わせれば PV が増えれば増えるほど CO_2 排出量を削減できることを示した．

EV をスマートグリッド間を移動するバッテリーとして用いるアイデアは上記の研究から妥当な方法であると言える．特に我々の研究は PV と EV に関するライフログデータを組み合わせて VGI のシミュレーション評価を行う点において優位性がある．

2.4 バッテリー損耗に関する研究

バッテリーの損耗に関する研究も行われている．リチウムイオンバッテリーの損耗は，時間経過によるカレンダーロスと，充放電によるサイクルロスに大別される．Wang ら [19] や Naumann ら [20] [21] は，バッテリー損耗に関する静的な実験を行い，バッテリー損耗モデルを構築した．バッテリー損耗の要因は，充放電や時間経過だけでなく，充放電速度や放電深度 (Depth of Discharge:DoD)，温度など，様々な要因が複雑に重なって起こることを示した．とりわけ，以下の項目はバッテリーの劣化を加速させることを示した．

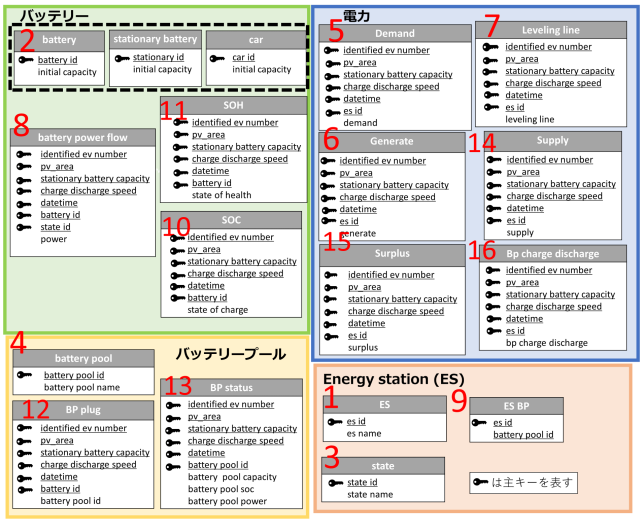


図 1　RDB スキーマ

- 充給電速度の高い充放電
- DoD の大きい充放電
- 極端な高温・低温下での利用

また，カレンダーロスは経過日数の平方根に比例することを示した．サイクルロスに関して，Wang らは総放電量に比例するとしたが，Naumann らは総放電量の平方根に比例するとした．

これらの論文は小容量のバッテリーで実験を行っていたのに対し，我々は EV のバッテリーと容量の大きいバッテリーを用いた点が異なる．また，彼らは静的な実験を行ってモデルを構築したが，実際の EV は個々の利用方法により各損耗の要因は様々である．そこで本研究では，EV ライフログに基づき，現実的なデータからバッテリー損耗を定量化することを目指す．

2.5　負荷平準化 VGIDB に関する先行研究

先行研究 [22], [23] では，3.2 節に後述する負荷平準指向 VGI に関する一連のデータを統合して蓄積するデータベースを構築した．その RDB スキーマを図 1 に示す．このデータベースは VGI を行った結果を多角的に定量化できるものである．また，スキーマに適合するデータ形式であれば，仮想データ・実データの区別を問わず定量評価できるものである．このスキーマを基に実装システムを構築した．

本研究では，現時点で入手可能な実データおよび妥当に生成された仮想データを用いることによって，再生可能エネルギーと EV バッテリーを活用した負荷平準化シミュレーションを行いう．そして，そのシミュレーションデータをデータベースに格納し，評価を行う．

3　需要側の負荷平準化を実現するスマートグリッド

3.1　想 定 環 境

本研究では日常的に移動する EV を移動するバッテリーとみなしスマートグリッドの一部として利用する．特に電力需要側の立場に立ち，オフィスビル群などの中小規模な地域電力網（マイクログリッド）を形成することを想定する．図 2 に示す

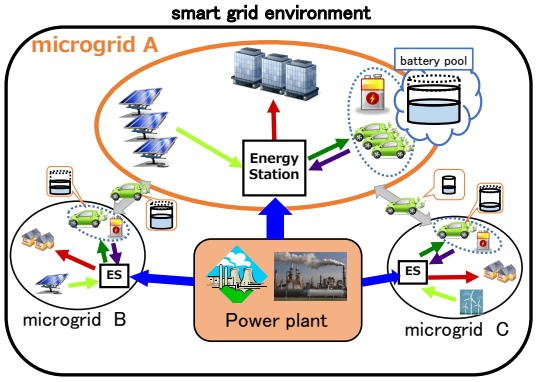


図 2　本研究で想定するスマートグリッド環境

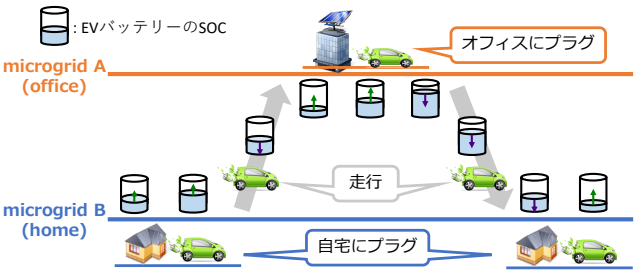


図 3　通勤に使用される EV の一日の EV の動き

ように，マイクログリッドは電力需要設備，PV 設備，バッテリー，および充給電設備（ES：Energy Station，以下 ES）により構成される．バッテリーには通勤者の EV とマイクログリッド内に設置される定置型バッテリーの 2 種類がある．ここで，通勤のためにオフィスと自宅を日常的に移動する EV を考える．EV の一日の動きの例を図 3 に表す．自宅では充給電を行う．移動の際には走行による放電を行う．マイクログリッド内では充給電を行う．ここで，グリッド内で充給電を行うことを，VGI に参加する，と定義する．マイクログリッドにおいて，この 2 種類のバッテリーを論理的に多数束ねた「バッテリープール」が構成されるとみなす．バッテリープールの満充電容量（以下，バッテリープール容量）と充電率（SOC）はバッテリーの充給電や EV のマイクログリッドへの出入りによって常に変化する．また，ES は PV のパワーコンディショナー，PV 設備/バッテリー/電力系統間の AC/DC 変換，後述するエネルギーフローの制御（どこにどれだけの電力を流すか），充給電制御といった役割を持つものを想定する [1]．そのような ES が，VGI を行うマイクログリッド 1 つに対して 1 つ存在する．なお，マイクログリッド内部の発電だけでマイクログリッドの需要を賄うのは不可能である．したがって，マイクログリッド外には東京電力のような「外部発電所」が存在し，マイクログリッドに対して電力を供給している（外部電力）．

1：この想定と同等の機能を持つシステムが，パナソニック株式会社などから HEMS として社会実装されている．

3.2 太陽光発電エネルギーと EV バッテリーを活用する負荷平準化

本研究では，外部電力に対してピークカット・ピークシフトを究極的に行った結果として外部発電所からの供給電力が一定になることを理想とする．このような状態は外部電源にとっての負荷が変動しないことから「負荷平準化」と呼ばれる．

図 4 に想定する電力の流れを，図 5 に太陽光発電エネルギーと EV バッテリーを活用する負荷平準化の概念図を示す．ここで図 5（a）はあるマイクログリッドの一日の電力の流れの状態を表し，図 5（b）が対応するバッテリープールの状態を表す．図 5（a）の塗りつぶされた部分と吹き出しの色が図 4 の矢印の色に対応する．EV はマイクログリッド内外へ移動するため，時間ごとにバッテリープールの容量は変化する（図 5（b））．

負荷平準化を行うためには大容量のバッテリープールが必要となるため，1 日を通して負荷平準化を実現することは現実的に難しい．そこで，1 日のうち特定の時間帯のみを負荷平準化の対象とする．この時間帯を負荷平準化時間帯と定義する．それ以外の時間帯をバッファ時間帯と定義する．改めて，図 5（a）の負荷平準化時間帯に注目する．横軸に 1 日の時間，縦軸に電力をとった．赤線の概形（需要）は建物の電力需要の推移を表す．黄緑線の概形（太陽光発電電力＋電力平準化レベル）は電力平準化レベルを基準として太陽光発電電力をその上に積み上げたもので時間帯ごとの総供給量を表す．ここで外部からの供給電力を一定にするための基準となる電力平準化レベルをあらかじめ定める．このときの電力の流れは以下のようになる．

- 需要（赤）が電力平準化レベル（水色一点鎖線）以下の時間帯
  - 外部電力を消費（青）
  - 外部電力をバッテリープールへ充電（緑）
  - 太陽光発電電力をバッテリープールへ充電（黄緑）
- 需要（赤）が電力平準化レベル（水色一点鎖線）以上の時間帯
  - 需要（赤）以下の太陽光発電電力（黄緑）
    - 太陽光発電電力を消費（茶色）
  - 需要（赤）以上の太陽光発電電力（黄緑）
    - バッテリープールに充電可能の場合
      - バッテリープールへ充電（黄緑）
    - バッテリープールが満充電の場合
      - 発電余剰（オレンジ）
  - 太陽光発電電力（黄緑）以上の需要（赤）
    - バッテリープールが給電可能の場合
      - バッテリー給電電力を消費（ピンク）
    - バッテリープールが残量不足の場合
      - 外部電力を消費（青）

また，図 5 の青点線の吹き出しが指し示すようにバッテリープールの充電量が不足した場合には一時的に負荷平準化不可能となる．

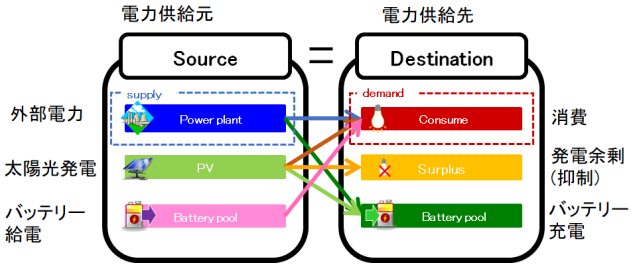
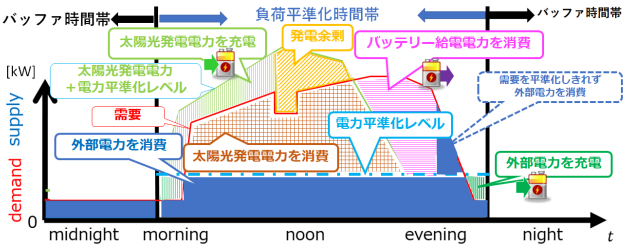
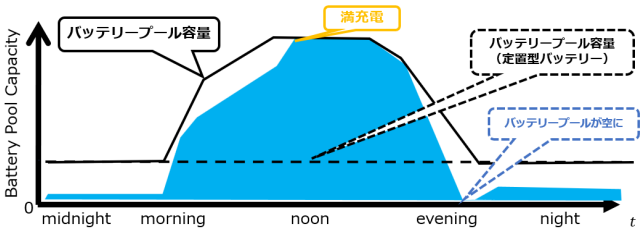


図 4　エネルギーフロー



(a) 電力の需要と供給の変化



(b) バッテリープールの容量と蓄積エネルギーの変化

図 5　太陽光発電エネルギーと EV バッテリーを活用する負荷平準化

4　バッテリー損耗モデルの構築

4.1　バッテリー損耗モデル構築の前提条件

VGI のデメリットとして EV バッテリーが損耗することは 1 章で述べた．VGI のために EV を提供するユーザにとって自身の EV バッテリーが損耗してしまうことは大きな懸念要素となる．本節では VGI によりどれだけ EV バッテリーが損耗するかを定量化するために，バッテリー損耗モデルを構築する．

本研究ではバッテリー損耗モデルの構築に，三年間記録してきた一台の EV ライフログを用いる．このデータは想定環境に近い形で EV を使い続けることで得られたものである．本来は複数の EV のデータがあるのが望ましいが，EV バッテリーの損耗は，人・場所ごとに異なるため，今ある人の生活に伴うデータを基に，モデル構築を行う．

2.4 節よりバッテリー損耗はカレンダーロスとサイクルロスから成るとされることが分かった．カレンダーロスは時間の経過に伴い増加し，サイクルロスは放電量に伴い増加する．しかし，放電量を全て計測することは，データ取得の観点から不可能であった．そこで，大局的には，放電量と充電量は等しいことから，充電量を用いる．そのため，バッテリー損耗モデルを構築するにはバッテリーの損耗量と，経過時間と充電量を記録する必要がある．

表 1　EV ライフログ取得方法

属性	値
使用車両	日産 LEAF ZAA-ZE1 型
データ取得アプリ	Android アプリ”Leaf Spy Pro”
取得端末	PLX Kiwi 3 OBD2，Android 端末
データ取得期間	2020 年 6 月 28 日 ～ 2023 年 6 月 27 日

4.2　EV ライフログを取得した EV

我々は一台の EV から EV ライフログを取得した．この EV は 2020 年 6 月に新車として購入され，以後データを取り続けている．一般的な自動車は，走行のみに利用されるが，この EV は神奈川県内において被験者の通勤に利用されているだけでなく，建物への充給電を行うなど，我々の想定環境に近い形で実験を行っているものである．なお，通勤に利用された電力量は約 6000kWh，給電した電力量は約 2500kWh である．また，充電回数に関して，普通充電は約 280 回，急速充電は約 20 回である．

4.3　データ取得

用いるデータは経過日数，SOH および充電量である．SOH(State of Health) はバッテリー健全度で，初期容量に対する相対容量 (%) を表す．それらのデータ取得方法を以下に示す．

EV ライフログの取得方法を表 1 に示す．EV の内部には，制御情報を転送するための機構として，CAN（Controller Area Network）が使われている．CAN は複数の電子制御装置（ECU: Electric Control Unit）間で通信を行う車載ネットワーク用に開発されたシリアルバス通信規格である．その CAN データの取得に Android アプリ Leaf Spy Pro を使用した [2]．Leaf Spy では，CAN データが時系列データとして取得される．

SOH は Leaf Spy で取得される値を用いる．取得したデータを可視化したものを図 6 に示す．

EV の充電量の計測は充電器側からの計測が二種類，EV 側からの計測は一種類の方法で行った．それらの計測方法を表 2 に示す．充電器側と EV 側との計測値の差はエネルギー変換に伴う損失と考えれられる．充電器側からの計測に関して，被験者の自宅での計測には電力計 (HIOKI 3168 クランプオンパワーハイテスタ) を設置し，電力量の計測を行った．また，被験者の自宅外での充電の際は，日産の充電サービス ZESP3[3]による充電を行い，その際に記録される値を利用する．EV 側での計測値には，Leaf Spy による GIDS[4]の値を用いる．Leaf Spy では，GIDS が時系列データとして計測される．充電前と充電後の GIDS の差を，EV 側から計測した充電量として用いる．

4.4　充電量の検証

充電量に関して，EV 側からの記録は，被験者のプライバシーの観点から取得されていないものもある．そこでバッテリー損耗モデルの構築には充電器側からの充電記録を用いることとした．ただし，充電器側の記録は二種類あるため，それらの精度を検証する．

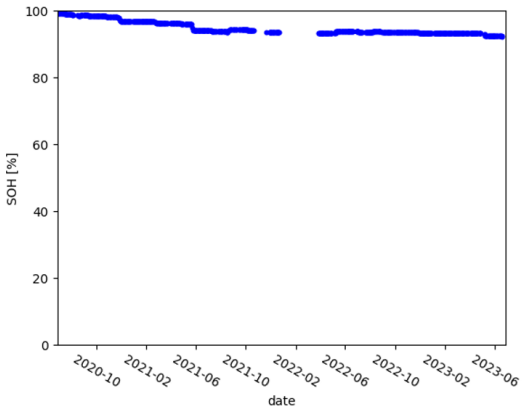


図 6　SOH の推移

表 2　EV 充電ログ取得方法

計測元	充電量	計測器
自宅充電器	E_{HOME}	HIOKI 3168 クランプオンパワーハイテスタ
自宅外充電器	E_{ZESP3}	ZESP3 のサービスによるもの
EV	E_{CAN}	CAN データを Leaf Spy により取得

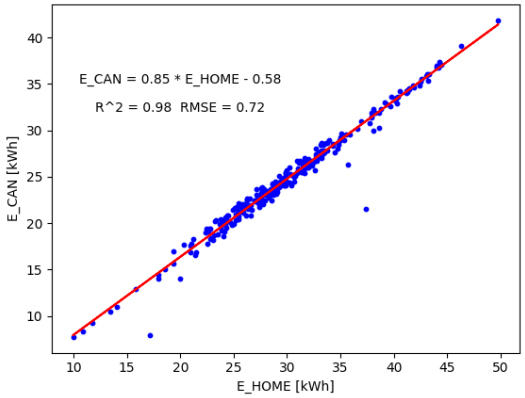


図 7　自宅充電の際の充電記録の比較

一回の充電に対して，充電量は EV 側と充電器側から計測される．そこで自宅，自宅外の充電器側の計測値に対して，EV 側からの計測値と比較する．自宅での充電の際の，充電器側の計測値に対する EV 側の計測値を比較した図を図 7 に示す．ZESP3 による充電の際の，充電器側の計測値に対する EV 側の計測値を比較した図を図 8 に示す．赤線は近似直線を表す．近似直線の式の係数は，充電効率を表すものであると考えられる．図 7 は決定係数が 0.98 と高く，安定したデータであると言える．また，充電効率は 85%であった．一方図 8 を見ると，決定係数は 0.91 と高いが，充電効率は 76%と，自宅充電の際の充電効率より小さくなっている．両者は充電効率にはばらつきがあるが，決定係数が高いことから E_{HOME} と E_{ZESP3} を同一に扱う．そのため，総充電量 $Charge$ は以下の式で表される．

$$Charge = E_{HOME} + E_{ZESP3} \tag{1}$$

2：“Leaf Spy Pro”, Google Play, https://play.google.com/store/apps/details?id=com.Turbo3.Leaf_Spy_Pro\&hl=ja\&gl=US, 参照 Dec 26, 2022.

3：https://www.nissan.co.jp/EV/CHARGE_SUPPORT/ZESP3/

4：クーロンカウンティング法により求められた値とされる．Leaf Spy によると 1GIDS は 0.0775kWh である．数値の妥当性については検証中である．

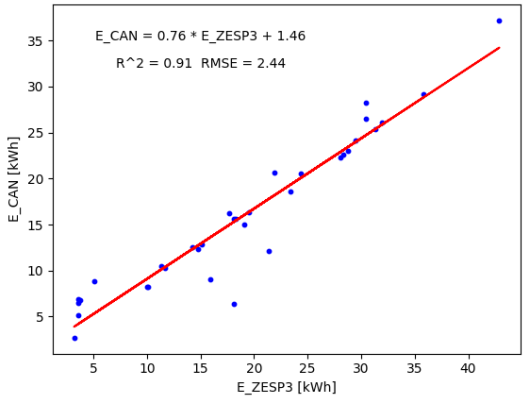


図 8　外部充電の際の充電記録の比較

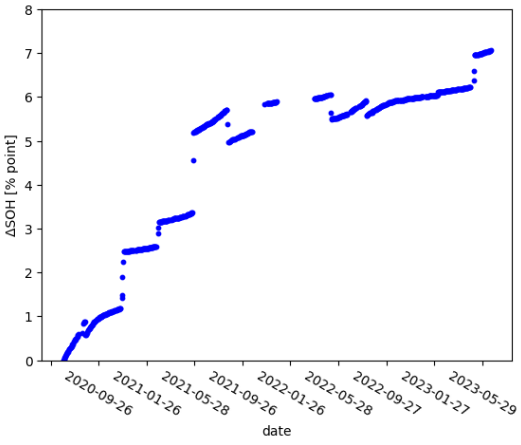


図 9　記録された EV の損耗量

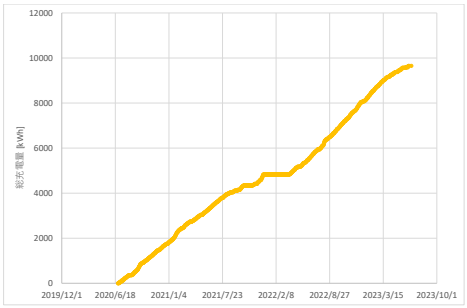


図 10　EV の充電量

4.5　バッテリー損耗モデルの構築

図 6 のグラフに関して，縦軸を SOH の減少量にしたグラフを図 9 に示す．横軸が日付，縦軸が新車購入時からの SOH の減少量 [% point] である．CAN データでは損耗が階段状に記録されているが，長期的には損耗している．また，この EV の一日ごとの積み上げ総充電量を図 10 に示す．経過時間に対して総充電量がほぼ線形に推移している．

2.4 節ではバッテリーの損耗要因について述べた．本来はこれらのようにサイクルロスとカレンダーロスを分けて考えるべきである．しかし，以下の理由により，カレンダーロスとサイクルロスは分けずに考えた．

- 実際にバッテリーが使用される環境は温度や充放電速度が様々である．
- 実際のバッテリー損耗量の内訳をカレンダーロスとサイクルロスに分けることはできない．
- バッテリー損耗量は階段状に記録されていて，実際の損耗とは少し異なることが推測される．

またデータを取得した EV の使用環境には，以下のような特徴がある．

- 急速充電の利用は限られたものである。
- 主に SOC30%から 80%の間で利用しているため，DoD は大きくない．
- 主な利用場所は神奈川県内であるため，極端な温度での利用は少ない．

そのためこの EV はバッテリー損耗を加速させるような使い方はしていないと考えられる．これらのことから温度や充放電速度，DoD などは一定のものとし，損耗モデルを構築した．

ここでカレンダーロスに関して以下の理由から，本論文では 3 年目のデータにおいて，バッテリー損耗モデルを構築した．

- カレンダーロスは経過時間の平方根に比例するものであるが，短期間であれば線形に近似できるものであると考える．
- カレンダーロスは時間が経つほど変化量が小さくなっていく．
- 記録された損耗量を見ると，明らかに一年目が多く損耗している．

本研究は VGI によるバッテリー損耗の定量化を目指すため，カレンダーロスの影響を極力小さくするバッテリー損耗モデルの構築を目指す．

またサイクルロスに関して以下の理由から，本論文では，バッテリー損耗モデルを総充電量の一次式として構築した．

- サイクルロスは充電量に比例するとしている文献 [19] がある．
- データ取得期間において，極端に充放電が多い期間はない．

図 11 に記録された EV バッテリー損耗量と構築したバッテリー損耗モデルを示す．青のプロットは横軸が対象期間における一日ごとの積み上げ総充電量 [kWh]，損耗量を表したものである．緑の縦線より右側が 3 年目である．赤い直線は，バッテリー損耗モデルである．損耗モデルを式 (2) に表す．この式は，平均二乗法により一次の近似直線を求めたものである．このモデルの精度検証は行っていないため，今後の課題である．このモデルが正しいものと仮定し，次章において VGI シミュレーションを行う．

$$Q_{loss} = 0.00026 * Charge + 4.0 \qquad (2)$$

5　VGI シミュレーションとバッテリー損耗の定量化

本章では，3.2 節で述べた想定環境における VGI シミュレー

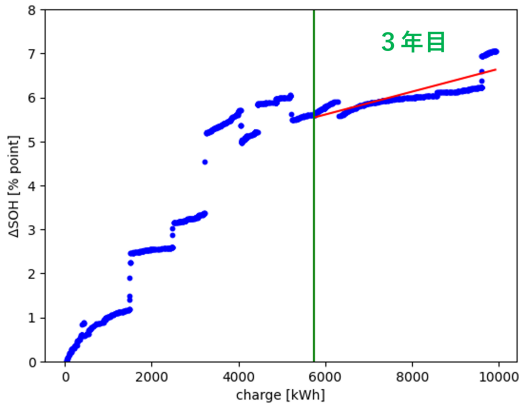


図 11　バッテリー損耗モデル

ションを行う．さらに，2.5 節で示したデータベーススキーマに，後述する使用データと表 3 に記載するパラメータの全組み合わせのシミュレーションデータを格納する．ここで，パラメータの組み合わせ 1 通りを 1 シナリオと呼ぶこととする．データベースに対して定型質問処理を行い，一台毎のバッテリー損耗量や，パラメータを変更した場合のシミュレーション結果への影響を可視化する．任意の太陽光パネル設置面積，VGI 参加 EV 数のときに，どの程度のバッテリー損耗が発生するかを事前に予測することは，EV ユーザにとって有用な情報である．

また，本研究では使用データは既知のものとし，遡及的シミュレーションを行う．なお，マイクログリッドを，約 10,000 人の生徒や教職員がいる横浜国立大学のキャンパス全体と想定する．10,000 人規模の人が生活するキャンパス全体はスマートグリッドの構成単位として妥当であると考える．

5.1　使用データ

使用したデータを以下に示す．パラメータを表 3 に示す．2.5 節の図 1 で示したデータベーススキーマによって作成したデータベースに，シミュレーションデータを挿入した．以下，図 1 中の赤番号順に説明する．

0. datetime 電力需要や太陽光発電量などの傾向は季節によるため，評価対象は 2021 年一年間のうちデータ欠損日を除いた 355 日とした．また，本来電力は連続値であるが，データ取得の観点から T_{step}（サンプリング間隔）を 1 時間とした．

$$\mathbb{T} = \{2021\ 01\text{-}01\ 0{:}00 + (n-1) | n \in \mathbb{N} \cap n < 24 * 365 hour\} \tag{3}$$

1. ES ES を有するマイクログリッドとして，通勤先のような電力需要及び駐車場に EV が存在するマイクログリッドと EV ユーザの各自宅で構成されるマイクログリッドがあることとした．

2. battery(car,stationary battery) バッテリーは，EV バッテリーと定置型バッテリーの二種類がある．それぞれのバッテリーは識別するための battery_id を持つ．

EV バッテリーの容量は，本来は個々に異なるものであるが，今後バッテリー技術が向上することを見込み一律に 100kWh とした．

EV は通勤に使用されるものを想定するため，グリッド内に存在する時間は限られる．そこで，EV がグリッド内に存在する時間を実データを基に設定した．グリッド内に存在する間はバッテリープールの一部となり，グリッドとの充給電を行うものとする．車の入退構時間について，横浜国立大学に実在する建物一棟の駐車場をカメラで測定し，得られた自動車 50 台の入退構時間を基に，EV の入退構時間を生成した．そのため，例えば EV1000 台シミュレーションは EV50 台の 20 組のコピーデータにより生成されたデータである．本来は入退構の時間は個々に異なるものであるが，データ取得の観点からコピーデータを使用した．入構時間と退構時間の間は，グリッド内でプラグされているものとする．

VGI 参加 EV 台数について 500 台，1000 台，1500 台をパラメータとした．なお，横浜国立大学の一日の入退構車数は 600 台であった．

定置型バッテリーに関しては，使われなくなったバッテリーなどを束ねたものを想定し，0kWh，2500kWh，5000kWh をパラメータとした．

3. state EV は以下のいずれかの状態を持つものとした．

- 通勤先にプラグしている
- 自宅にプラグしている
- 通勤のために EV ユーザの自宅と通勤先間を走行している

通退勤に走行する電力量は本来は個々に異なるものであるが，データ取得の観点から一律で通勤では 1.57kWh，退勤では 1.85kWh とした．これは自家用車の平均通勤距離が 10km [24] であることを基に，実際の EV の走行ログから生成したものである．

4. Battery Pool：BP BP は，通勤先に 1 つ，各自宅に一つずつ存在するとした．

5. Demand 横浜国立大学の施設部が構内の各建物の電力データを公開している [1]．通勤先の需要電力データ (実データ) として用いた．

6. Generate 日射量のオープンデータから以下の式（JIS C 8907:2005）を用いて算出 [2] した仮想データを生成した．H_{Am} は自治体が [2] 公開している日射量から算出した．

$$E_p = K * P_{AS} * H_{Am} / G_s \tag{4}$$

横浜国立大学の全建物の屋上面積の合計が約 60000m^2 であったことから，太陽光パネル設置面積について 30000m^2，40000m^2，50000m^2，60000m^2 をパラメータとした．

7. Leveling line あるマイクログリッドの電力平準化レベルの高さ（leveling line[kW] を以下の式を用いて計算した．電力平準化レベルは本来事前に決定されるべきものであるが，既知のものとした．人の生活や太陽の動きは一日サイクルであると考え，負荷平準化は一日ごとに行うものとし，T_{start} を 0 時，負荷平準化開始時刻を太陽光発電開始時刻，負荷平準化終了時刻 T_{end} を 24 時とした．$Destination_c(t)$ はグリッドの一時間毎の消費電力量，$Source_{pv}(t)$ は一時間毎の太陽光発電電力量とする．

表 3　パラメータ

項目	パラメータ
EV 台数 (台)	500, 1000, 1500
太陽光パネル面積 (m^2)	30000, 40000, 50000, 60000
定置型バッテリー容量 (kWh)	0, 2500, 5000

$$L(\tau) = \frac{1}{T_{end} - T_{start}} \sum_{t=T_{start}}^{T_{end}} (Destination_c(t) - Source_{pv}(t)) \quad (5)$$

8. battery power flow バッテリーの充給電は以下の条件のもとで発生しうる仮想データを生成した.

- 通勤先における負荷平準化時間帯の負荷平準化を達成するための充給電
- EV ユーザの各自宅における次の日の通勤先到着時に SOC が 55%になるための EV への充電
- EV ユーザの各自宅と通勤先間における出退勤のための EV の走行

また，現在の設備容量より通勤先での充給電速度を 6kW，自宅での充給電速度を最大 3kW とした．また，現在の設備容量より各自宅，通勤先ともに充電効率 η_{charge} を 0.9 給電効率 $\eta_{discharge}$ を 0.85 とした.

9. ES BP 通勤先の ES が唯一バッテリープールを 1 つ，各自宅の ES が唯一バッテリープールを一つずつもつ.

10. SOC (State of Charge : バッテリー充電率) バッテリー劣化の観点から EV の SOC は 30〜80%の間に収まるように充放電を行うこととした仮想データを生成した.

11. SOH (State of Health : バッテリー健全度) 本来は EV バッテリーの損耗度合は個々に異なるものであるが，一律で 100%とした.

12. BP plug 定置型バッテリーは常に通勤先のバッテリープールに属しているものとした．EV は入退構ログから通勤先に停車中は常に通勤先のバッテリープールに属しているものとし，自宅に到着すると自宅のバッテリープールに属しているものとし仮想データを生成した.

13. BP status その時間にバッテリープールに属している個々のバッテリーに状態の総和で求められる仮想データを生成した.

14. Supply Supply は Demand，Generate，Leveling line，BP status から決まる仮想データを生成した．負荷平準化時間帯は Supply は Leveling line に一致するように制御される．BP への充給電に過不足が生じると Supply は Leveling line からそれだけずれる．負荷平準化時間帯以外では，定置型バッテリーへの充給電と Demand の値から定まる.

15. Surplus BP へ充電しきれないと Surplus が発生する仮想データを生成した.

16. Bp charge discharge ES から計測した BP との充給電を記録した仮想データを生成した.

1：横浜国立大学施設部，http://shisetsu.ynu.ac.jp/gakugai/shisetsu/
2：横浜市環境創造局，http://www.city.yokohama.lg.jp/kankyo/

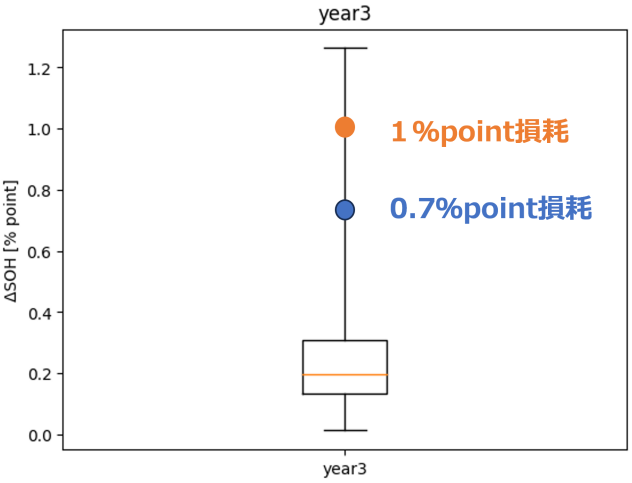


図 12　全シナリオにおける全 EV の損耗量の分布

5.2　シミュレーション評価

5.2.1　バッテリー損耗モデルの適用

4 章では，バッテリー損耗は充電量の一次関数としてモデル化した．VGI に参加したことによる追加のバッテリー損耗は，建物への給電を行うことによるものである．そこで，一シナリオ一台ごとに，一年間のグリッド内での総給電量を求める．そして，4 章で構築したバッテリー損耗モデル式 (2) により，一年間のバッテリー損耗量を求める．ここで，EV 側が受け取る充電量は EV が放電する量と大局的には同量である．総給電量を $Discharge$ とし，シミュレーションに適用する損耗モデルを以下に表す.

$$Q_{loss} = 0.00026 * Discharge \quad (6)$$

5.2.2　全シナリオにおける全 EV の損耗

全てのシナリオにおける全ての EV の損耗量の分布を図 12 に示す．VGI による損耗量は最大で一年あたり，1.35%point であることが分かった.

どのような生活をする EV の損耗量が増えてしまうのかについての考察を述べる．図 13 に，1%point 損耗した EV と 0.7%point 損耗した EV の，時間ごとのグリッドにプラグしていた日数を示す．プラグ時間に関しては，5.1 節で述べたように，実データを基にしているため，一台毎に異なる．そのため，VGI による損耗量は異なる．図 13 を見ると，1%point 損耗した EV は，夕方から夜にグリッドにいる日数が多いことがわかる．太陽光発電電力量が減少し，建物の電力需要が多い夕方以降の時間は EV からの給電量が増える．そのため，夕方以降にグリッドにいることが多い EV は，損耗量が増えてしまう．このように，生活パターンにより損耗量が増えてしまうこともあるため，補償などの制度制定が必要になると考えられる.

1%point の損耗がどの程度のものなのかについての考察を述べる．一般的な自家用車の年平均走行距離を 1 万 km と仮定する．また，日産の LEAF の WLTC モードでのカタログ電費は 7.5km/kWh である [5]．よって，EV で通勤した場合の走行によ

5：https://www3.nissan.co.jp/vehicles/new/leaf/charge/battery.html#

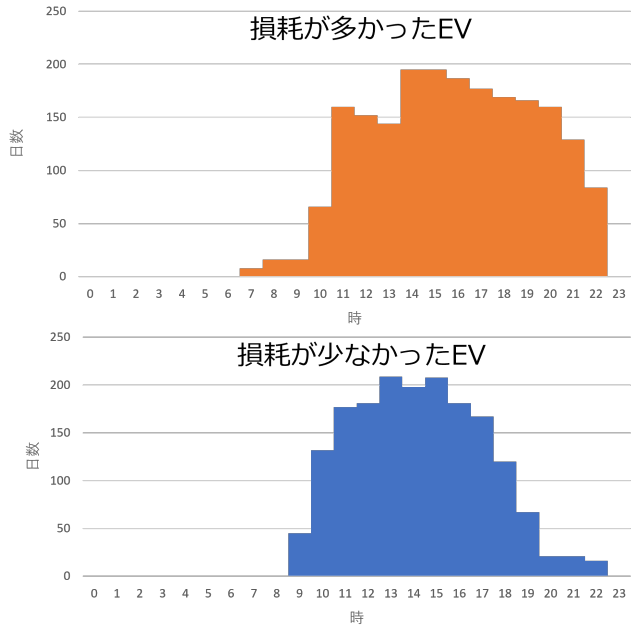


図 13　時間ごとのグリッドにいた日数

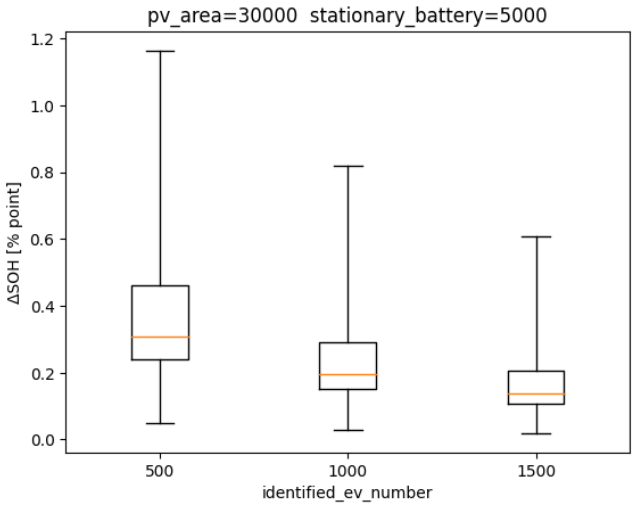


図 14　EV 台数毎の EV バッテリーの損耗度合の分布

る年間平均消費電力量は約 1300kWh となり，走行による年間損耗量は約 0.35% point となる．VGI によるバッテリー損耗量は約 1% point であることから，一台当たりの年間損耗量は約 1.35% point であることが分かる．ここで，EV のバッテリー交換の目安は日産自動車により 8 年 [6]であることが示されている．よって，VGI を 8 年間行い続けた場合は，VGI による損耗と走行による損耗の合計が約 10.8% point となる．また，そのうち，8%point は，VGI によって生じる追加の損耗となる．

5.2.3　EV が増えた場合の影響

図 14 に，太陽光パネル設置面積を 30000m^2，定置型バッテリー容量を 5000kWh にした場合の，EV 台数毎の EV バッテリーの損耗度合の分布を示す．EV500 台を基準に，EV の台数

battery, 参照 Feb 16, 2024.

6：https://www.nissan.co.jp/EV/MAINTENANCE/, 参照 Jan 6, 2024.

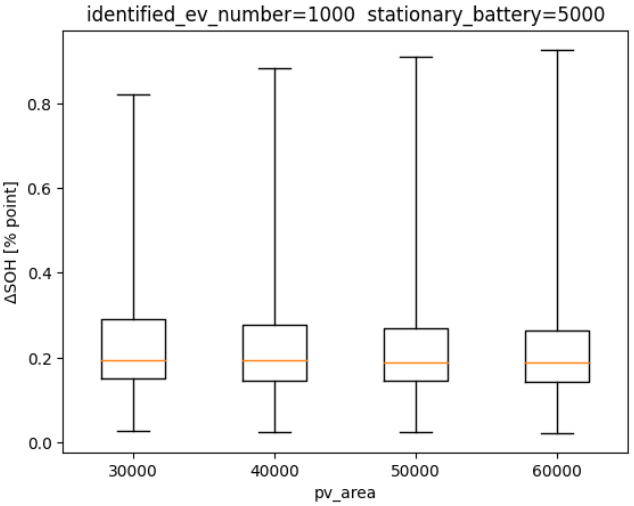


図 15　太陽光パネル設置面積毎の EV バッテリーの損耗度合の分布

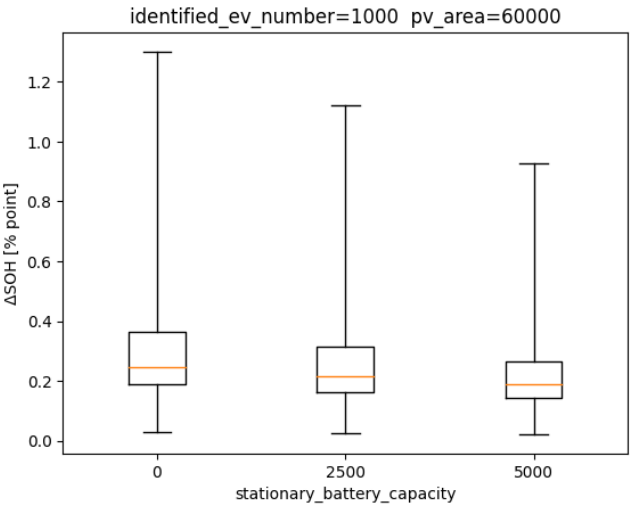


図 16　定置型バッテリー容量毎の EV バッテリーの損耗度合の分布

が倍になると個々の EV バッテリーの損耗量は約 0.7 倍に，3 倍になると約 0.5 倍になることが分かった．バッテリーの数が増えるため，一台当たりの負担は減少することが定量的に示された．

5.2.4　太陽光パネルが増えた場合の影響

図 15 に，EV の台数を 1000 台にし，定置型バッテリー容量を 5000kWh にした場合の，太陽光パネルの設置面積毎の EV バッテリーの損耗度合の分布を示す．太陽光パネルの設置面積が増えても，バッテリー損耗量はあまり変わらない．太陽光パネルが増えることで電力平準化レベルが下がり，日没後などの EV からの給電量が増えるが，充給電速度の制限があるため大幅には増えないと考えられる．そのため充給電速度が高くない限り，太陽光パネルの設置面積は EV バッテリーの損耗に影響がないと言える．

5.2.5　定置型バッテリー容量が増えた場合の影響

図 16 に，EV の台数を 1000 台にし，太陽光設置面積を 60000m^2 にした場合の，定置型バッテリー容量毎の EV バッテリーの損耗度合の分布を示す．定置型バッテリーを設置しな

い場合を基準に，2500kWh の時は EV バッテリーの損耗量は約 0.85 倍に，5000kWh の時は約 0.7 倍であることが分かった．定置型バッテリーが増えるほど，EV 一台当たりの負担を減らせることが定量的に示された．

6 ま と め

本論文では提案 DB を用いて，VGI シミュレーションによる EV バッテリーの損耗を定量化した．その結果，以下のことが明らかになった．

- VGI による追加のバッテリー損耗は年 1% point 程度である．
- EV が多いほど一台当たりのバッテリー損耗を抑えることができる．
- 充給電速度に制限がある限り太陽光パネルの設置面積は EV バッテリーの損耗にあまり影響しない．
- 定置型バッテリー容量が多いほど，EV のバッテリー損耗を抑えられる．

人・場所固有のデータからバッテリー損耗を定量化した．また，VGI によるバッテリー損耗は予測できることが分かった．これは，VGI 導入時の補償などの制度制定の判断材料になると考えられる．

損耗モデルの妥当性を確認する必要があることは，今後の課題である．

謝　　辞

本研究の一部は JSPS 科研費 (課題番号 22H03810) の支援による．

文　　献

[1] 総務省. 令和元年版 情報通信白書. https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2019/2019-index.html. Last accessed: 21 May 2020.

[2] 経済産業省. 新産業構造ビジョン. https://www.meti.go.jp/press/2017/05/20170530007/20170530007.html. Last accessed: 16 January 2021.

[3] 経済産業省資源エネルギー庁. 令和元年度エネルギーに関する年次報告 (エネルギー白書 2020). https://www.enecho.meti.go.jp/about/whitepaper/2020pdf/. Last accessed: 30 January 2021.

[4] 合田忠弘, 諸住哲. スマートグリッド教科書. インプレスジャパン, 初版, 2011.

[5] 東京電力ホールディングス株式会社. 長期的な安定供給とカーボンニュートラルの両立に向けた事業構造変革について. https://www.tepco.co.jp/press/release/2022/pdf2/220428j0101.pdf, 4 2022. (Accessed on 12/24/2022).

[6] 経済産業省資源エネルギー庁. 太陽光発電について (第 63 回 調達価格等算定委員会). https://www.meti.go.jp/shingikai/santeii/pdf/063_01_00.pdf. Last accessed: 04 July 2022.

[7] 三菱総合研究所, 環境省. 平成２７年度低炭素社会の実現に向けた中長期的再生可能エネルギー導入拡大方策検討調査委託業務 報告書. https://www.env.go.jp/earth/report/h29-02/h27_all.pdf. Last accessed: 8 January 2020.

[8] Ali Ahmadian, Mahdi Sedghi, Ali Elkamel, et al. Plug-in electric vehicle batteries degradation modeling for smart grid studies: Review, assessment and conceptual framework. *Renewable and Sustainable Energy Reviews*, Vol. 81, pp. 2609–2624, 2018.

[9] Jiucai Zhang, Jennie Jorgenson, Tony Markel, and Kevin Walkowicz. Value to the grid from managed charging based on california's high renewables study. *IEEE Transactions on Power Systems*, Vol. 34, No. 2, pp. 831–840, 2019.

[10] Cong Zhang, Jeffery B. Greenblatt, Pamela MacDougall, Samveg Saxena, and Aditya Jayam Prabhakar. Quantifying the benefits of electric vehicles on the future electricity grid in the midwestern united states. *Applied Energy*, Vol. 270, p. 115174, 2020.

[11] 小嶋祐輔, 荒井綾希子, 阿部力也, 岩船由美子. 電気自動車・蓄電池を用いた戸建て住宅地域の電力融通システムの設計方法の確立およびその評価. エネルギー・資源学会論文誌, Vol. 42, No. 2, pp. 50–57, 2021.

[12] Narottam Das, Akramul Haque, Hasneen Zaman, Sayidul Morsalin, and Syed Islam. Domestic load management with coordinated photovoltaics, battery storage and electric vehicle operation. *IEEE Access*, Vol. 11, pp. 12075–12087, 2023.

[13] Marwa Ben Arab, Mouna Rekik, and Lotfi Krichen. Suitable various-goal energy management system for smart home based on photovoltaic generator and electric vehicles. *Journal of Building Engineering*, Vol. 52, p. 104430, 2022.

[14] 池田伸太郎, 大岡龍三. 日本国内におけるスマートシティ・スマートコミュニティ実証事業の最新動向. 生産研究, Vol. 66, No. 1, pp. 69–77, 2014.

[15] Willett Kempton and Jasna Tomić. Vehicle-to-grid power fundamentals: Calculating capacity and net revenue. *Journal of power sources*, Vol. 144, No. 1, pp. 268–279, 2005.

[16] Bernhard Jansen, Carl Binding, Olle Sundstrom, et al. Architecture and communication of an electric vehicle virtual power plant. In *2010 First IEEE International Conference on Smart Grid Communications*, pp. 149–154. IEEE, 2010.

[17] 中川二彦, 千阪秀幸, 古矢勝彦. 電動車を用いた太陽光電力の高効率利用システム. 自動車技術会論文集, Vol. 53, No. 1, pp. 38–44, 2022.

[18] 理弘大澤, 保夫桑沢, 孝之秋元, 大介田嶌, 洋園江, 聡史上森, 正之小松. 自動車，蓄電システムを含めた住宅エネルギー性能指標とＶ２Ｈによる co2 排出量削減の効果. エネルギー・資源学会論文誌, Vol. 40, No. 3, pp. 78–84, 2019.

[19] John Wang, Justin Purewal, Ping Liu, Jocelyn Hicks-Garner, Souren Soukazian, Elena Sherman, Adam Sorenson, Luan Vu, Harshad Tataria, and Mark W. Verbrugge. Degradation of lithium ion batteries employing graphite negatives and nickel–cobalt–manganese oxide + spinel manganese oxide positives: Part 1, aging mechanisms and life estimation. *Journal of Power Sources*, Vol. 269, pp. 937–948, 2014.

[20] Maik Naumann, Franz B. Spingler, and Andreas Jossen. Analysis and modeling of cycle aging of a commercial lifepo4/ graphite cell. *Journal of Power Sources*, Vol. 451, p. 227, 2020.

[21] Maik Naumann, Michael Schimpe, Peter Keil, Holger C. Hesse, and Andreas Jossen. Analysis and modeling of calendar aging of a commercial lifepo4/graphite cell. *Journal of Energy Storage*, Vol. 17, p. 153, 2018.

[22] 石毛大貴, 能條太悟, 廣居樹, 本藤祐樹, 富井尚志. ライフログデータベーススキーマに基づくマイクログリッドにおける電力負荷平準化の価値の定量化. 研究報告データベースシステム（DBS）, 第 2022-DBS-175 巻, pp. 1–6, 2022.

[23] 廣居樹, 片平昇輝, 鈴木博登, 小松虎太郎, 本藤祐樹, 富井尚志. 太陽光発電と EV を統合する負荷平準指向スマートグリッド db を用いた試行錯誤を伴う VGI シミュレーション. 研究報告データベースシステム（DBS）, 第 2023-DBS-177 巻, pp. 1–6, 2023.

[24] 国土交通省総合政策局. 交通基本法案検討小委員会 資料. https://www.mlit.go.jp/common/000128770.pdf, 11 2010. (Accessed on 12/28/2022).

人流シミュレーションにおける 複数指標を考慮したパラメータ影響の可視化

森　千紗† 武田 芽依†† 大西 正輝††† 伊藤 貴之†

† お茶の水女子大学　〒 112-8610 東京都文京区大塚 2–1–1
†† 筑波大学　〒 305-8577 茨城県つくば市天王台 1–1–1
††† 産業技術総合研究所　〒 135-0064 東京都江東区青海 2–4–7
E-mail: †{g2020539,itot}@is.ocha.ac.jp, ††s2220504@s.tsukuba.ac.jp, †††onishi-masaki@aist.go.jp

あらまし　人流シミュレーションは人流誘導方法を決定するパラメータによって制御され，そのシミュレーション結果は混雑度や移動時間といった複数の指標によって評価される．シミュレーション利用者にとって好ましい誘導計画を策定するためには，利用者が人流シミュレーションの指標を解釈する必要がある．しかし，これらの複数の指標の間にはトレードオフが存在する．そのため，シミュレーション利用者にとって好ましい誘導方法に最も近いバランスのとれた指標値を得ることは簡単ではない．また，好ましい誘導方法につながるような人流誘導計画のパラメータ値を選択することも容易ではない．この「複数指標を考慮したパラメータ解釈が困難である」という問題を解決する一手段として，可視化によって，トレードオフを考慮した人流シミュレーションのパラメータと評価指標の関係を容易に解釈できるようにすることがあげられる．そこで本稿では，利用者の求めるトレードオフを的確に表現するために複数の評価値の線形加重和を計算し，Parallel Coordinate Plots を用いてその線形加重和とパラメータの関係を可視化する手法を提案する．

キーワード　人流シミュレーション，パラメータ，トレードオフ，可視化

1 はじめに

大規模なイベントにおいて，混雑を回避するために，あるいは混雑によって引き起こされる群衆事故を回避するために，適切な人流誘導が非常に重要である．近年では 2022 年に，ソウル梨泰院にて雑踏事故が発生し，多くの被害者を出した．このような事故を未然に防ぐため，街の管理者やイベント運営者は事前に人流誘導計画を策定する必要がある．しかし，大規模な空間において，偶発的に起こる現象を考慮して適切な誘導方法を検討することは非常に難しい．この課題を解決する手段として，人の動きをコンピュータ上に再現する人流シミュレーションが注目を集めている．実際に，花火大会における混雑緩和 [1] や劇場における避難誘導支援 [2] などにおいて，人流シミュレーションを用いた分析が報告されている．これらの分析結果は，人々の安全な暮らしを支えるために活用されている．

人流シミュレーションは，人の誘導方向や移動開始時間などの誘導方法を定めるパラメータによって制御される．そしてシミュレーション結果は，混雑度や移動時間といった複数の評価指標によって評価される．効果的な人流誘導計画を策定するためには，どのようなパラメータ制御によって評価指標がどのように変化するのかを分析し，パラメータと評価指標の関係を解釈するべきである．しかし，人流シミュレーションの複数の評価指標の間にはトレードオフが存在するため，パラメータと評価指標の関係性は非常に複雑であり，その解釈は容易ではない．そのため，利用者にとって好ましい誘導方法に最も近いバランスの評価指標値のトレードオフを容易に解釈し，好ましい誘導方法につながるような人流誘導計画のパラメータ値を容易に選択できるようにする手法が求められる．

パラメータと評価指標の関係の解釈手法は「評価指標に対するパラメータの良し悪しの定量化」と「パラメータと評価指標の分布の可視化」の二つに分けられる．前者には，各評価指標に対して重みを割り当ててそれらの線形結合を最小化または最大化する線形加重和法や，選好によって決定する理想解との距離を用いる手法 [3] などが存在する．これらの手法によって，パラメータ選択による意思決定時の複雑性の削減や，利用者が求める評価指標を達成するパラメータの定義域の絞り込み等が可能になる．一方で後者には，散布図行列（Scatter Plot Matrix）や Parallel Coordinate Plots （PCP） を適用した手法 [4]，評価指標の予測モデルを等高線によって可視化する手法 [5], [6] などが存在する．これらの手法によって，パラメータと評価指標の関係性や数値分布を可視化することが可能である [7]．例えば避難誘導シミュレーション [2] においては，混雑度と避難完了時間のような複数の評価指標が存在する．そのため，人流シミュレーションにおける利用者の選好を考慮した解釈という観点では，応用上評価指標間のトレードオフを考慮してパラメータの影響を可視化することは重要な意味を持つ．しかし，人流の可視化において「評価指標に対するパラメータの良し悪しの定量化」と「パラメータと評価指標の分布の可視化」を組み合わせたような，複数の指標間のトレードオフを想定した分析を可能にする可視化手法はまだ報告されていない．

これらの背景から我々は，人流シミュレーションにおけるト

レードオフを考慮したパラメータと評価指標の関係性の可視化手法に注目した．本稿では，利用者の求めるトレードオフを表現するような複数評価値の線形加重和を解析的に計算し，PCP を用いて評価指標とパラメータの関係性を可視化することで，利用者の選好を表現してパラメータの分析が可能な可視化手法を提案する．本稿で提案する手法は以下に示す 4 つの画面で 1 つの可視化画面を構築している．

（1） 評価指標値散布図とパレート解近似曲線
（2） パラメータの定義域を調整するスライダ
（3） パラメータと評価指標値の関係性を表す PCP
（4） 指定したパラメータに対する評価指標値の時系列変化を表示したグラフ

本稿では分析例として，提案手法を用いて避難誘導シミュレーションデータを可視化した結果を紹介し，多様なトレードオフに関してパラメータの分析が可能であることを実証する．

要約すると本論文の貢献は以下の 2 点である．

- 利用者の選好を考慮した線形加重和の最適な重みの計算方法の提案
- トレードオフ関係を表すために PCP を用いたパラメータの分布の可視化

2 背 景 知 識

2.1 人流シミュレーション

人流シミュレーションでは，人流誘導方法を定めるパラメータ $\boldsymbol{\lambda} := [\lambda_1, \cdots, \lambda_D] \in \Lambda := \Lambda_1 \times \cdots \times \Lambda_D \subseteq \mathbb{R}^D$（$d \in [D] := \{1, 2, \cdots, D\}$ に対して，$\Lambda_d \subseteq \mathbb{R}$ は d 次元目の変数の定義域を表す）によって人流を制御し，混雑度や避難時間等の指標 $\boldsymbol{f} := [f_1, \cdots, f_M] \in \mathbb{R}^M$ によってシミュレーション結果を評価する．ここで M は評価指標の個数を表す．今回の解析対象となる人流シミュレーションにおけるパラメータには，経路上の分岐点に立つ避難誘導員が各経路に振り分ける歩行者の比率や，歩行者の移動開始時間等があげられる．前述の通り，評価指標は必ずしも単一であるとは限らず，複数の評価指標を考慮する必要がある場合もある．本稿では，可視化の際には各評価指標に対して重みを割り当てて，複数の評価指標の線形和 E を次式で求めることにより，評価指標を単一化する．

$$E = \sum_{m=1}^{M} w_m f_m(\boldsymbol{\lambda}) \tag{1}$$

ここで，$\{f_m(\boldsymbol{\lambda})\}_{m=1}^{M}$ は各評価指標の値，$\{w_m\}_{m=1}^{M}$ は対応する評価指標に対する重みである．

2.2 Parallel Coordinate Plots（PCP）

本稿で用いる PCP [4] は，**図 1** に示すように，各パラメータおよび各評価指標に対応する $D+1$ 本の平行な軸上で，各値を結ぶ折れ線の集合で多次元データを可視化する手法である．$\{\boldsymbol{\lambda}_n, E(\boldsymbol{\lambda}_n)\}_{n=1}^{N}$（$f$ は $\Lambda \to \mathbb{R}$ を満たす写像）の第 n 観測パラメータ $\boldsymbol{\lambda}_n$ の d 次元目の要素が $\lambda_{n,d} \in \mathbb{R}$ であるとする．PCP では $n \in [N]$ の各観測毎に d 番目の軸に $\lambda_{n,d}$ を記録し，これを $d \in [D]$ に対して適用する．$D+1$ 番目の軸には $E(\lambda_{n,d})$ を記録し，n 番目の観測で得られた $D+1$ 点を結ぶ折れ線を描く．最終的には N 本の折れ線が描画される．

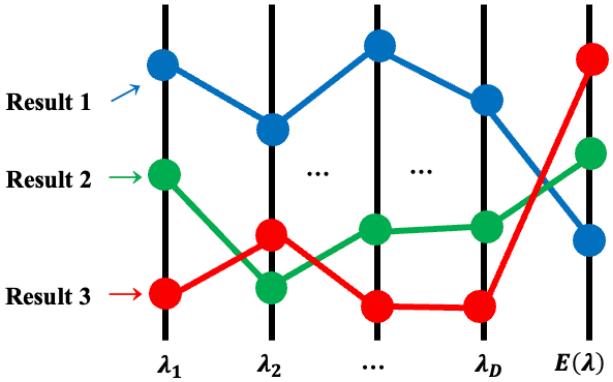


図 1: PCP の例．パラメータの各次元及び評価指標分用意された平行な軸上に各観測の値を結び，折れ線として多次元データを可視化する．

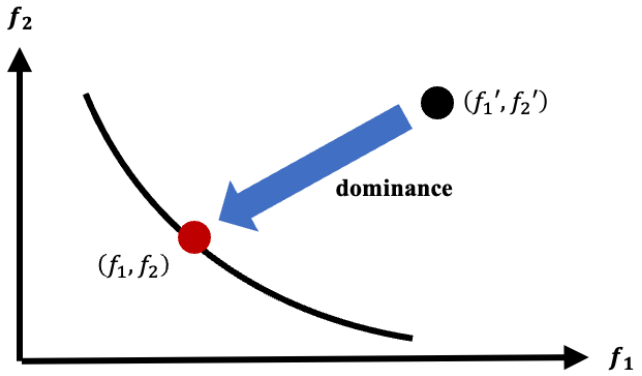


図 2: 支配の例．全ての評価指標値を最小化したいものとする．二次元の評価指標ベクトル $\boldsymbol{f}$ と $\boldsymbol{f}'$ で $\boldsymbol{f}$ が $\boldsymbol{f}'$ を支配している．

2.3 パレート解とパレートフロント

評価指標値 $\boldsymbol{f} := [f_1, f_2, \cdots, f_M] \in \mathbb{R}^M$ があり，全ての $m = 1, \cdots, M$ に対して f_m を最小化したいものとする．最適なパラメータを発見するためには，評価指標値の個数が単一の場合は，単純に評価指標値が小さいパラメータ群に着目すればよい．しかし，複数の評価指標間の優劣関係は未定義であり，一方が良くなれば他方が悪くなるというトレードオフが存在するため，最適なパラメータは必ずしも一意に定まらない．評価指標が複数存在する場合における最適パラメータを定義するために，一般的に用いられる概念は支配（dominance）[8] である．二次元の評価指標ベクトル $\boldsymbol{f}$ と $\boldsymbol{f}'$ で $\boldsymbol{f}$ が $\boldsymbol{f}'$ を支配するとき，$f_1 \leqq f_1'$，$f_2 \leqq f_2'$ 及び $\boldsymbol{f} \neq \boldsymbol{f}'$ が成立する（**図 2**）．特に評価指標値集合 $\{\boldsymbol{f}_n\}_{n=1}^{N}$ において，ある $\boldsymbol{f}_i$ が任意の $n \in \{1, \cdots, N\}$ で $\boldsymbol{f}_n$ に支配されないとき，$\boldsymbol{f}_i$ はパレート解であると呼ぶ．パレート解集合が形成する線（面）のことをパレートフロントと呼ぶ．

3 関 連 研 究

パラメータと評価指標の関係の解釈手法は「評価指標に対するパラメータの良し悪しの定量化」と「パラメータと評価指標の分布の可視化」の二つに分けられ，従来から報告されている．

前者には，各評価指標に対して重みを割り当ててそれらの線形結合を最小化または最大化する線形加重和法や，選好によって決定する理想解との距離を用いる手法 [3] などが存在する．複数の評価指標が存在し，これらの評価指標を同時に最適化する問題である多目的最適化において，各評価指標の相対的な重要度を考慮した重みにより評価指標を単一化することでパラメータを評価する手法を線形加重和法と呼ぶ．重要度の考慮のために各評価指標に対して重みを割り当てて，それぞれの評価指標に重みをかけて総和をとることにより，線形加重和 E を式 (1) で計算する．これらの計算によって，パラメータ選択による意思決定時の複雑性の削減や，利用者が求める評価指標を達成するパラメータの定義域の絞り込み等が可能になる．

一方で後者には，散布図行列（Scatter Plot Matrix）や Parallel Coordinate Plots（PCP）を適用した手法 [4]，評価指標の予測モデルを等高線によって可視化する手法 [5], [6] などが存在する．散布図行列（Scatter Plot Matrix）は可視化するデータの次元のうち，全ての二次元の組み合わせを対象として散布図を作成し，それを格子状に並べて行列の形式に一覧表示したものである．この可視化により，任意の組み合わせの次元間の相関を一画面で直感的に視認可能である [9]．Parallel Coordinate Plots（PCP）は可視化するデータの次元数の座標軸を平行に並べ，1 つのデータを 1 本の折れ線で表す可視化手法である．PCP では評価指標値が複数存在する場合であっても各観測値を可視化するため，各次元のパラメータ分布を単一の図で確認することができる [7]．これらの可視化手法によって，パラメータと評価指標の関係性や数値分布を可視化することが可能である．

例えば避難誘導シミュレーション [2] においては，混雑度と避難完了時間のような複数の評価指標が存在する．そのため，人流シミュレーションにおける利用者の選好を考慮した解釈という観点では，応用上評価指標間のトレードオフを考慮してパラメータの影響を可視化することは重要な意味を持つ．しかし，人流の可視化において「評価指標に対するパラメータの良し悪しの定量化」と「パラメータと評価指標の分布の可視化」を組み合わせたような，複数の指標間のトレードオフを想定した分析を可能にする可視化手法はまだ報告されていない．本手法では，解析的計算を行うことで，利用者の選好を考慮したパラメータと評価指標の関係性を PCP から読み取れるような可視化を目指している．

4 パラメータと評価指標値の関係性の可視化手法

本節では，本稿で提案する可視化システムの処理手順を示す．4.1 節では，D 次元のパラメータ $\boldsymbol{\lambda} \in \mathbb{R}^D$ によって制御される人流シミュレーションの評価指標値 $\boldsymbol{f} \coloneqq [f_1, f_2, \cdots, f_M] \in \mathbb{R}^M$ の GUI 上での選好選択手法，4.2 節では，利用者の評価指標値選好にもとづく単一評価指標の計算手法，4.3 節では，図 3 に示す評価指標とパラメータの関係性可視化システムを構成する 4 つの画面について説明する．本稿では簡単のため $M = 2$ とし，全ての $m = 1, \cdots, M$ に対して f_m は最小化したいものとして議論する．

4.1 利用者による GUI 上での評価指標選好選択

本研究では，高速に良質な解を探索可能な多目的最適化手法 MOTPE [10] を用いた最適化により，パレート解集合を得る．図 3 (1) の黒破線が，本稿の実験で得られたパレート解集合を近似的に結んだ曲線である．この近似曲線は，観測内にあるパレート解をもとに，以下の関数の係数 a, b を最適化することで求められる．

$$g(f_1) \coloneqq f_2 = \frac{a}{f_1 + b} \tag{2}$$

ここで a, b の最適化には最小二乗法を適用する．図 3 (1) からわかるように，各評価指標値のバランスは数多く存在し，適切なバランスは利用者の選好による．本システムでは，黒破線上の一点を利用者がクリックすることで，適切な評価指標値間のバランスを選択する．なお，この近似曲線はパレート解の集合が下に凸であることを仮定している．

4.2 評価指標選好にもとづく単一評価指標の計算

単一評価指標値 E の計算とは，以下のような $\{w_m\}_{m=1}^M$ を求めることを指す．

$$E = \sum_{m=1}^{M} w_m f_m \text{ subject to } \sum_{n=1}^{M} w_m = 1 \tag{3}$$

一般的にも成り立つが，$M = 2$ のときについて $E = w_1 f_1 + w_2 f_2$ であることから，$w_2 \neq 0$ で以下のように変形できる．

$$f_2 = -\frac{w_1}{w_2} f_1 + \frac{E}{w_2} \tag{4}$$

この式は図 3 (1) の空間内における直線を表している．ここで，前節で GUI 上から選択された点にて，E が最小値（最善値）となるときを考える．近似曲線が下に凸であるため，GUI 上で選択された点を接点とする接線を取る w_1, w_2 を選択すると，GUI 上で選択された点上で E の値が最小値を取る．よって，GUI 上で選択された点を $(f_1, g(f_1))$ とすると，$-w_1/w_2 = g'(f_1)$ と $w_1 + w_2 = 1$ を連立方程式として解くことによって (w_1, w_2) を得る．**図 4 (a)** は $w_1 = 0.05$，$w_2 = 0.95$ の場合，**図 4 (b)** は $w_1 = 0.95$，$w_2 = 0.05$ の場合の直線である．

4.3 評価指標値とパラメータの関係の可視化

本稿で提案される可視化システムは，以下の 4 つの画面で構成される．

（1） 評価指標値散布図とパレート解近似曲線
（2） パラメータの定義域を調整するスライダ
（3） パラメータと評価指標値の関係性を表す PCP
（4） 指定したパラメータに対する評価指標値の時系列変化を表示したグラフ

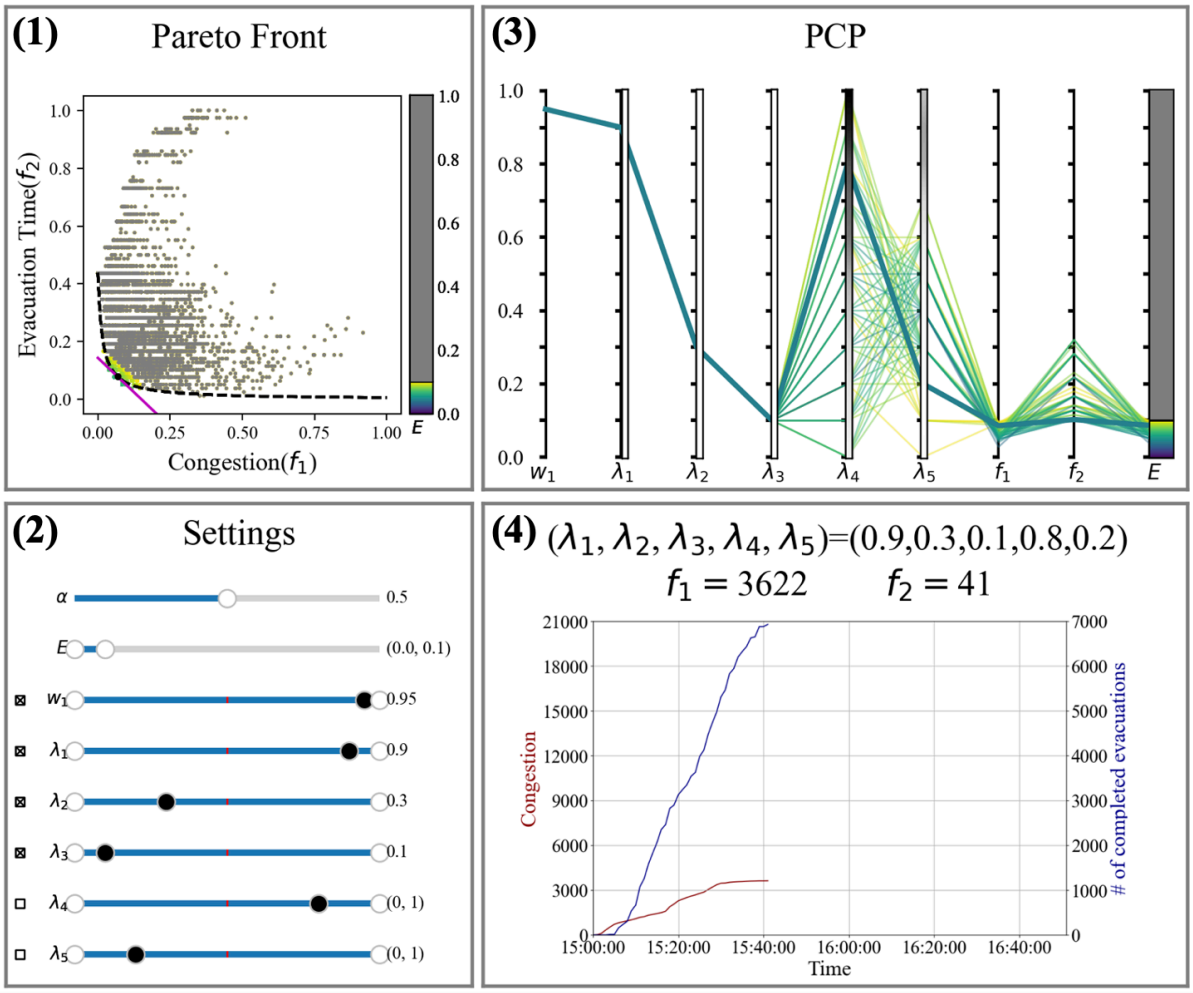


図 3: 可視化システムの全体図．**(1)** 評価指標値散布図とパレート解近似曲線，**(2)** パラメータの定義域を調整するスライダ，**(3)** パラメータと評価指標値の関係性を表す PCP，**(4)** 指定したパラメータに対する評価指標値の時系列変化を表示したグラフ．

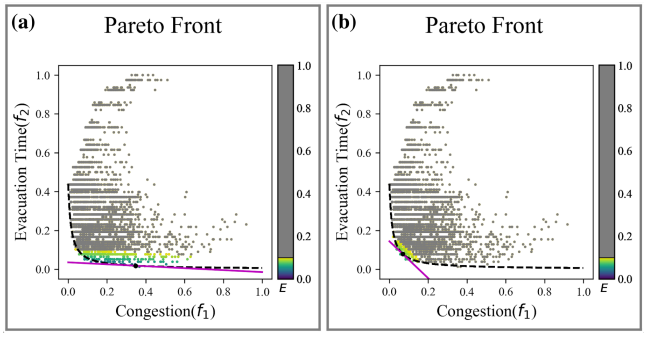


図 4: 近似曲線とその接線の例．**(a)** は $w_1 = 0.05$，$w_2 = 0.95$，**(b)** は $w_1 = 0.95$，$w_2 = 0.05$ を用いて描画した．利用者は選好によって近似曲線上のクリック位置を決定する．

まず利用者が画面 (1) によって評価指標値間のトレードオフ，画面 (2) によって観察するパラメータの定義域を指定する．そして指定した値にもとづいて，画面 (3) の PCP，画面 (4) のグラフによる可視化を行う．このような流れで利用者の選好を考慮することにより，データの概要や詳細を観察することを想定している．ここで，使用するデータセットでは評価指標が 2 つであることから，$w_2 = 1 - w_1$ と書き表すことができる．したがって，利用者は w_1 を指定することで，単一評価指標の重みを決定することができる．

画面 (1) (**図 3 (1)**) は式 (4) の w_1 を計算する際に用いる．利用者は画面 1 の近似曲線上の一点をクリックすることで，評価指標値間のトレードオフを選択し，それにもとづいて w_1 を計算し，単一評価指標値を色付けし，クリック点を接点とする接線が描画される．続いて，画面 (1) 上の操作によって計算された w_1 と式 (3) を用いて，各観測の単一評価指標値 E を得る．図 4 にあるように，この値の優劣にもとづいて画面の色相が算出される．この可視化によって，利用者は GUI 上で簡単に選好を定めることが可能である．

画面 (2)(**図 3 (2)**) ではスライダ，レンジスライダ，チェックボックスを用いて，可視化するパラメータの定義域，w_1，画面 (3) のグラフの線の透明度を指定することができる．画面 (1) 上での操作では GUI を用いて主観的に重みを選択するのに対して，直接数値で重みを決定したい場合のために，重みを指定するスライダを置いている．パラメータと重みの定義域を指定する際は，持ち手が白いレンジスライダを用いる．ここで，チェックボックスにチェックを入れ，持ち手が黒いスライダを用いると，表示する値を一つに絞ることができる．

画面 (3) (**図 3 (3)**) では，画面 (1) と画面 (2) で利用者が指

定した範囲内にある観測を対象として，PCP を用いて可視化する．画面 (3) の軸は左側から，重み w_1，パラメータ $\boldsymbol{\lambda}$，評価指標値 $\boldsymbol{f}$，単一評価指標 E である．線の色は，画面 (1) の散布図上の点と同様の基準で割り当てられる．この PCP 上の折れ線の分布を観察することで，パラメータと評価指標の関係性や特徴を確認することができる．画面 (2) のスライダで指定した線を太く描画することで，太い線に注目して折れ線の集合を観察することができる．また，それぞれのパラメータの軸の横の矩形領域は，各パラメータにおける E への影響度を可視化する機能を持つ．全ての定義域内のパラメータについて，そのパラメータ値を通り，かつその他のパラメータについては現在のレンジスライダの定義域を通る線の集合の E の分散を計算し，分散の分布を得る．そして，解析の結果をヒートマップを用いて矩形領域に可視化する．この矩形領域を見ることで，値を固定すると E の分散が小さくなるパラメータを知ることができる．

画面 (4)(**図 3 (4)**) では，画面 (2) のスライダで指定したパラメータにおける評価指標の時系列変化を可視化したグラフと，各評価指標の値を表示する．これにより，画面 (3) で太線で表示された誘導の詳細情報を確認することができる．

5 適 用 例

5.1 データセット

本稿が示す適用例では，歩行者シミュレータ CrowdWalk [11] を用いた避難誘導シミュレーションを実施して，人流シミュレーションデータを得る．そして，シミュレーションのパラメータとそれによる避難行動を表す統計量を用いて，人流シミュレーションデータを可視化する．実験は，以下の 5 つのパラメータによって制御する．

（1） 出発時刻の時間差
各初期地点 (避難開始場所) に存在する避難者を 2 分割し，出発時刻に時間差 λ_1 を設ける．

（2） 混雑の少ない経路への誘導
避難開始時に避難者が存在する位置の分布には偏りが生じている．そのため混雑地点と混雑の少ない地点を結ぶ 2 本の動線を使用し誘導を行う割合をそれぞれ λ_2，λ_3 とする．

（3） 地下への誘導
想定する環境では地上空間に加えて地下空間が存在する．また，目的地までに複数箇所の階段が存在する．そのためそれぞれの階段を利用し地下に誘導する割合を λ_4，λ_5 とする．

それぞれのパラメータは $[0,1]$ になるように正規化しており，実験では各パラメータに対して $\{k/10\}_{k=0}^{10}$ を取りうる値とした．したがって，今回の実験では $11^5 = 161,051$ 通りのパラメータに対して予めシミュレーションを実行し，それぞれのパラメータに対して評価指標値を得た．

今回の実験で考慮した評価指標値は以下の 2 種類である．1 つ目の評価値 f_1 は混雑度である．n 番目の避難者の局所的な

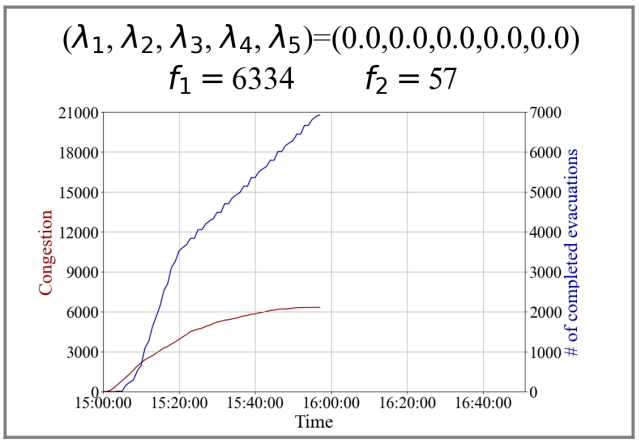


図 5: 誘導をしない場合の画面 (4)．$\lambda_1 = 0.0$，$\lambda_2 = 0.0$，$\lambda_3 = 0.0$，$\lambda_4 = 0.0$，$\lambda_5 = 0.0$ に設定した．

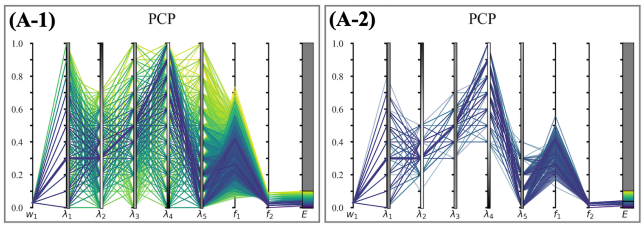


図 6: 避難完了時間 f_2 を重視した場合（$w_1 = 0.03$）の画面 (3)．**(A-1)** は $0.00 \leqq E \leqq 0.1$，**(A-2)** は $0.00 \leqq E \leqq 0.04$ について可視化した．折れ線の分布や色相差から，λ_2，λ_4，λ_5 が避難完了時間に大きく関わると考えられる

人口密度を ρ_n，過密の閾値 θ^ρ（地上 2.17/地下 1.08），避難時間 T 分を用いて，混雑度を次式で定義する．

$$f_1 = \sum_{t=1}^{T}\sum_{n=1}^{N} \mathbb{I}[\rho_n > \theta_\rho] \tag{5}$$

ただし，$\mathbb{I}$ は指示関数である．2 つ目の評価指標 f_2 は全体の避難完了時間である．n 番目の避難者の出発時刻 $t_n^{(d)}$，到着時刻 $t_n^{(a)}$，最後に避難を完了した避難者 N を用いて，次式で定義する．

$$f_2 = t_N^{(a)} - t_1^{(d)} \tag{6}$$

なお，可視化時には f_1, f_2 ともに $[0,1]$ に正規化している．したがって，E の範囲は $E \in [0,1]$ となる．

5.2 可 視 化 例

提案手法を用いて 5.1 節のデータセットを可視化し，その結果を分析した．**図 5** に誘導をしない場合，図 6 (A-1)，図 6 (A-2)，図 8 (A-3) に避難完了時間 f_2 を重視した $w_1 = 0.03$ の場合，図 7 (B-1)，図 7 (B-2)，図 8 (B-3) に混雑度 f_1 を重視した $w_1 = 0.97$ の場合の可視化結果を示す．PCP の線の色は，$E > 0.1$ の場合は灰色，$E \leqq 0.1$ の場合は 0 に近いほど濃青に近づくように，PCP の軸の横の矩形領域は分散が大きいほど黒色に近づくように色相を設定している．

図 6 (A-1) は避難完了時間 f_2 を重視した場合の可視化結果である．画面 (1)，画面 (2) を用いて $w_1 = 0.03$，$0.00 \leqq E \leqq 0.1$，$0.0 \leqq \lambda_1 \leqq 1.0$，$0.0 \leqq \lambda_2 \leqq 1.0$，$0.0 \leqq \lambda_3 \leqq 1.0$，

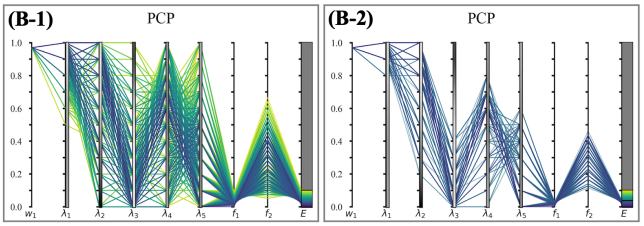


図 7: 混雑度 f_1 を重視した場合（$w_1 = 0.97$）の画面 (3)．**(B-1)** は $0.00 \leqq E \leqq 0.1$，**(B-2)** は $0.00 \leqq E \leqq 0.04$ について可視化した．折れ線の分布や色相差から，λ_1，λ_2，λ_3 が避難完了時間に大きく関わると考えられる．

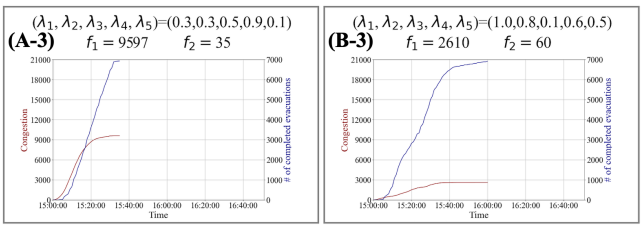


図 8: 避難完了時間 f_2 を重視した場合（$w_1 = 0.03$）と混雑度 f_1 を重視した場合（$w_1 = 0.97$）の画面 (4) の例．**(A-3)** は $\lambda_1 = 0.3$，$\lambda_2 = 0.3$，$\lambda_3 = 0.5$，$\lambda_4 = 0.9$，$\lambda_5 = 0.1$，**(B-3)** は $\lambda_1 = 1.0$，$\lambda_2 = 0.8$，$\lambda_3 = 0.1$，$\lambda_4 = 0.6$，$\lambda_5 = 0.5$ の場合について評価指標の詳細を可視化した．

$0.0 \leqq \lambda_4 \leqq 1.0$，$0.0 \leqq \lambda_5 \leqq 1.0$ に設定した．λ_2 は $0.0 \leqq \lambda_2 \leqq 0.7$ に出現しており，パラメータ値が低い値に偏っていることが見てとれる．このことから混雑の少ない経路への誘導パラメータ λ_2 は特定の範囲内でなければならず，λ_2 は避難完了時間に大きく関わるパラメータであると考えられる．λ_2 以外に関しては折れ線の存在する範囲に偏りは見られないが，色相差から λ_4 は高い値，λ_5 は低い値を取るべきであると解釈することができる．各線の数値分布を見やすくするために $0.00 \leqq E \leqq 0.04$ に設定し表示する E の範囲を狭め，折れ線の数を減らした場合の図が**図 6 (A-2)** である．λ_4 は $0.6 \leqq \lambda_4 \leqq 1.0$，$\lambda_5$ は $0.0 \leqq \lambda_5 \leqq 0.3$ に折れ線が集中していることがわかる．このことからも，地下への誘導パラメータ λ_2 と λ_5 が取るべき範囲がわかる．また，地下への誘導 λ_4 の値が高いことから，λ_4 の地下動線の活用は避難完了時間の短縮につながるとわかる．パラメータの横の矩形領域を観察すると，$0.6 \leqq \lambda_4 \leqq 1.0$ は白色であり，パラメータをその範囲に保つと他のパラメータが安定しなかったとしても E の振れ幅が小さくなるとわかる．そのため，λ_4 は一定の適切さで誘導を実施するために特に重要なパラメータであるといえる．まとめると，避難完了時間 f_2 を重視した誘導を行うためには，λ_2，λ_4，λ_5 の誘導を重点的に行うべきであり，$0.0 \leqq \lambda_2 \leqq 0.7$，$0.6 \leqq \lambda_4 \leqq 1.0$，$0.0 \leqq \lambda_5 \leqq 0.3$ の範囲で誘導を行うと良いことがわかる．誘導のぶれを想定する場合は，λ_4 に配置する誘導員の数を増やして λ_4 のぶれを小さくすると，安定した誘導を実現できる．この誘導を行なった時の例として**図 8 (A-3)** で $\lambda_1 = 0.3$，$\lambda_2 = 0.3$，$\lambda_3 = 0.5$，$\lambda_4 = 0.9$，$\lambda_5 = 0.1$ の場合の画面 (4) を観察する．誘導をしない場合（図 5）と避難完了時間 f_2 を重視した場合（図 8 (A-3)）を見比べると，図 8 (A-3) は図 5 に比べて f_2 が小さくなっていることがわかる．以上の可視化によって，避難完了時間の小さい誘導方法を発見することができた．

図 7 (B-1) は混雑度 f_1 を重視した場合の可視化結果である．画面 (1)，画面 (2) を用いて $w_1 = 0.97$，$0.00 \leqq E \leqq 0.1$，$0.0 \leqq \lambda_1 \leqq 1.0$，$0.0 \leqq \lambda_2 \leqq 1.0$，$0.0 \leqq \lambda_3 \leqq 1.0$，$0.0 \leqq \lambda_4 \leqq 1.0$，$0.0 \leqq \lambda_5 \leqq 1.0$ に設定した．λ_1 は 1 に近い値に集中しており，1 に近いほど折れ線の色が青に近づいていることが見てとれる．このことから出発時間の時間差パラメータ λ_1 は避難完了時間に大きく関わるパラメータであり，混雑度を小さくするためには大きく出発時間差を設けて，動線に十分なキャパシティを持たせることが重要であるといえる．λ_1 以外に関しては折れ線の存在する範囲に違いは見られないが，色相差から λ_2 は高い値，λ_3 は低い値を取るべきであると解釈することができる．各線の数値分布を見やすくするために $0.00 \leqq E \leqq 0.04$ に設定し，表示する E の範囲を狭め，折れ線の数を減らした場合の図が**図 7 (B-2)** である．λ_2 は $0.7 \leqq \lambda_2 \leqq 1.0$，$\lambda_3$ は $0.0 \leqq \lambda_3 \leqq 0.2$ に折れ線が集中していることがわかる．このことからも，混雑の少ない経路への誘導パラメータ λ_2 と λ_3 が，混雑度に関わる重要なパラメータであると考えられる．ここで，パラメータの横の矩形領域を観察すると，$0.7 \leqq \lambda_2 \leqq 1.0$，$\lambda_3$ は $0.0 \leqq \lambda_3 \leqq 0.2$ は白色であり，パラメータをその範囲に保つと他のパラメータが安定しなかったとしても E の振れ幅が小さくなるとわかる．そのため，混雑の少ない経路への誘導パラメータ λ_2，λ_3 は，一定の適切さで誘導を実施するために特に重要なパラメータであると言える．まとめると，混雑度 f_1 を重視した誘導を行うためには，λ_1，λ_2，λ_3 の誘導を重点的に行うべきであり，λ_1= 1.0，$0.7 \leqq \lambda_2 \leqq 1.0$，$\lambda_3$ は $0.0 \leqq \lambda_3 \leqq 0.2$ の誘導を行うと良いことがわかる．誘導のぶれを想定する場合は，λ_2，λ_3 に配置する誘導員の数を数を増やして λ_2，λ_3 のぶれを小さくすると，安定した誘導を実現できる．この誘導を行なった時の例として**図 8 (B-3)** で $\lambda_1 = 1.0$，$\lambda_2 = 0.8$，$\lambda_3 = 0.1$，$\lambda_4 = 0.6$，$\lambda_5 = 0.5$ の場合の画面 (4) を観察する．誘導をしない場合（図 5）と混雑度 f_1 を重視した場合（図 8 (A-3)）を見比べると，図 8 (B-3) は図 5 に比べて f_1 が小さくなっていることがわかる．可視化によって混雑度の小さい誘導方法を発見することができた．

このようにして，折れ線の存在する範囲や色相の違いから，各シナリオにおけるパラメータ重要度 [12] や，取るべきパラメータの定義域の分析が可能である．紙幅の都合上，他の w_1 値の可視化結果は省略するが，同様の分析を多様なトレードオフに関して可能にするのが提案手法の利点である．

6 ま と め

本稿では，人流シミュレーションの最適化を対象として，利用者の直観にもとづいて複数の評価指標のバランスをとるためにパラメータ探索を支援する可視化手法を提案した．本手法では利用者が興味のある各指標のトレードオフを GUI から選択し，そのトレードオフを最もよく表す単一指標値を線形和に

よって解析的に計算する．その単一指標値を利用して PCP を描画し，選択した誘導方法については詳細を表示することで，利用者の選好をもとにしたデータの特徴の分析や好ましい誘導方法の発見を容易にする．本稿では避難誘導の人流シミュレーションデータを作成し，そのデータセットを可視化した例を紹介した．実験により，各評価指標の重みを変更することで，多様なトレードオフに対して重要なパラメータが異なり，その差異を可視化できることを確認した．

今後の課題は以下の通りである．

- PCP の視認性の向上
- パレート解集合が下に凸でない場合の重み決定方法の検討

現状の実装では，表示する折れ線の本数が多数存在した場合に線が何重にも重なるため視認性が低下する．そのため各個体を表現する折れ線に対してクラスタリングを適用し，各クラスタを代表する折れ線のみを表示することで視認性を向上させる．また PCP を用いた描画の際，パラメータの軸の順序が固定されているため，隣接していないパラメータ間の関係の読み取りは困難である．軸間の相関の有無によって軸の順序を変動させる機能を追加し，パラメータ間の相関を漏れなく解釈することを目指す．

加えて，現在はパレート解集合が下に凸であることを仮定してパレート解の集合を曲線で近似している．パレート解集合が下に凸でない場合も GUI 上で重みを選択できるよう，評価指標選択手法を改善する必要がある．

文　献

[1] 山下倫央, 大西正輝. オリンピックのための情報処理: 2. オリンピックにおける人の流れの解析. 情報処理, Vol. 55, , 2014.

[2] 大西正輝, 山下倫央, 星川哲也, 佐藤和人. 人の流れの計測とシミュレーションによる避難誘導方法の伝承-新国立劇場における避難体験オペラコンサートを例に. 人工知能学会第二種研究会資料, Vol. 2015, No. KST-26, p. 06, 2015.

[3] 上田悠, 太宰龍太, 綛田長生, 伊香賀俊治, 加藤彰浩. 学習/多目的最適化機能を組み込んだ空調制御技術の実験的研究. 計測自動制御学会論文集, Vol. 46, No. 8, pp. 439–447, 2010.

[4] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. *Eurographics*, 2013.

[5] M. Lindauer, M. Feurer, K. Eggensperger, A. Biedenkapp, and F. Hutter. Towards assessing the impact of Bayesian optimization's own hyperparameters. *arXiv:1908.06674*, 2019.

[6] R. Sass, E. Bergman, A. Biedenkapp, F. Hutter, and M. Lindauer. DeepCAVE: An interactive analysis tool for automated machine learning. *arXiv:2206.03493*, 2022.

[7] M. Kubota, T. Itoh, S. Obayashi, and Y. Takeshima. EVOLVE: A visualization tool for multi-objective optimization featuring linked view of explanatory variables and objective functions. In *International Conference on Information Visualisation*, 2014.

[8] S. Watanabe. Python tool for visualizing variability of pareto fronts over multiple runs. *arXiv:2305.08852*, 2023.

[9] 伊藤貴之. 高次元データ可視化のための次元選択. オペレーションズ・リサーチ= Communications of the Operations Research Society of Japan: 経営の科学, Vol. 63, No. 1, pp. 6–12, 2018.

[10] 武田芽依, 大西正輝. 多目的ベイズ最適化を用いた震災時における地下誘導の検証. 進化計算シンポジウム 2022 講演論文集, No. S2-03, pp. 120–125, 2022.

[11] T. Yamashita, T. Okada, and I. Noda. Implementation of simulation environment for exhaustive analysis of huge-scale pedestrian flow. *Journal of Control, Measurement, and System Integration*, Vol. 6, , 2013.

[12] S. Watanabe, A. Bansal, and F. Hutter. PED-ANOVA: Efficiently quantifying hyperparameter importance in arbitrary subspaces. *arXiv:2304.10255*, 2023.

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

2024年3月1日(金) 10:00 ~ 12:10 | T5-C オンライン（Zoom Events)

エンタメ・メディア (1)

座長:橋本 敦史(オムロンサイニックエックス)
コメンテータ:廣田 雅春(岡山理科大学)

10:00 ~ 10:05
イントロダクション

10:05 ~ 10:20
[T5-C-7-01] ショートビデオを用いた画像検索における観光地紹介ビデオの生成方式

*藤原 虎留[1]、王 元元[1]、岳 五一[2] (1. 山口大学、2. 京都情報大学院大学)

10:20 ~ 10:45
[T5-C-7-02] マルチプレイヤ RPG において人間らしいプレイスタイルをとるゲームAIの提案

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

10:45 ~ 11:10
[T5-C-7-03] 中間音楽シーケンス生成のためのTransformerモデル

*劉 明明[1]、酒井 哲也[1] (1. 早稲田大学)

11:10 ~ 11:35
[T5-C-7-04] YouTube動画の視聴状況の予測

*花畑 槙哉[1]、許 レイ[2]、馬 強[3] (1. 京都工芸繊維大学　情報工学課程　社会知能情報学研究室、2. 京都情報大学院大学/京都工芸繊維大学、3. 京都工芸繊維大学)

11:35 ~ 12:00
[T5-C-7-05] 言語モデルによる文の流暢さを用いた大喜利生成

*高橋 巧実[1]、鈴木 優[1] (1. 岐阜大学)

ショートビデオを用いた画像検索における観光地紹介ビデオの生成方式

藤原　虎留† 　王　元元† 　岳　五一††

† 山口大学工学部知能情報工学科　〒 755–8611 山口県宇部市常盤台 2-16-1
†† 京都情報大学院大学　〒 606–8225 京都府京都市左京区田中門前町 7 番地
E-mail: †{a070ffu, y.wang}@yamaguchi-u.ac.jp, ††w_yue@kcg.edu

あらまし　本研究では，観光地の印象と魅力を向上させるため，画像検索を通じて観光ショートビデオおよびそれに関連する観光スポットを発見し，これらのショートビデオを活用して観光地紹介ビデオを生成する手法を提案する．これにより，従来の観光サービスにおけるテキスト情報や画像による推薦よりも，よりリアルな観光地の魅力を伝え，印象を把握することが可能となる．また，画像検索により入力写真と類似度の高い観光スポットをまとめて推薦することで，観光地の選択時間を短縮できる期待がある．本論文では，まず観光ショートビデオおよび観光写真を収集する．次に，観光ショートビデオにラベルを付与し，観光写真を用いて学習を行うことで，ユーザの入力写真のタグを決定する．ビデオラベルと写真タグのコサイン類似度を算出し，高い類似度のショートビデオから観光スポットを抽出する．抽出された観光スポットの位置情報と入力写真の位置情報に基づき，観光地域を推定する．最終的に，推定された地域の公式観光サイトにおける観光スポットの出現順に従い，ショートビデオを並べて観光地紹介ビデオを生成し，その有用性を評価する．

キーワード　旅行需要，ショートビデオ，観光写真，画像検索，観光支援

1 はじめに

コロナ禍の収束に伴い，旅行需要が回復傾向にあり，スマートフォンの普及により多くの人がじゃらん net[1]や楽天トラベル [2]などの Web 上の観光情報に容易にアクセスできるようになった．JTB 日本観光振興協会とヴァリューズの調査によると，じゃらん net は 2022 年のパソコンからの推定閲覧者数が 3,030 万人，スマートフォンからの推定閲覧者数は 3,190 万人で，いずれも 3,000 万人を超えている [1]．これらの数値は旅行への規制の緩和に伴い今後も増加すると考えられる．しかし，Web 上の情報は膨大であり，選択が難しく，クエリ検索によって潜在的な旅行候補地が見過ごされがちな点も指摘されている．

従来の観光サービスでは，主流となっていたのはテキストや画像による推薦であったが，近年では TikTok や YouTube，Instagram の利用者増加に伴い，動画を用いた観光地のプロモーションが容易になってきている．観光動画の利用には多くのメリットがある．たとえば，1 つの動画から得られる情報が多岐にわたり，グルメやイベント情報，宿泊施設の情報などが得られる点や，動画の動きによって現地の雰囲気がより伝わりやすく，実際に訪れたいという気持ちを高めることができる．また，視覚的に表現されるため、日本語を理解していない外国人観光客にもアピールしやすい．

そこで本研究では，ユーザが入力した写真を利用して類似度の高い観光地紹介ビデオを生成し，観光地の印象と魅力を向上させることを目指している．画像検索を用いて入力写真とショートビデオにタグやラベルを付与し，高い類似度の観光ショートビデオを抽出し，抽出されたショートビデオが対象としている観光スポットの位置情報と入力写真の位置情報に基づいて観光地域を推定する．推定された地域の公式観光サイトにおける観光スポットの出現順に従い，ショートビデオを並べて観光地紹介ビデオを生成する手法を提案する．ショートビデオは，長さが 60 秒以内の短編動画であり，数秒から数十秒のものが一般的である．これらは主に各動画プラットフォームに搭載されており，短尺であるためスキップされにくく，製作費を抑えられるとともに，通常の動画よりも再生数が伸びやすいという利点がある．さらに，スマートフォンや SNS の普及により，場所や時間の制約なしに閲覧が容易であり，SNS で拡散される可能性が高いため，これが観光地の印象と魅力向上に寄与することが期待される．

本論文の構成は以下のとおりである．2 章では，Web 上の情報を用いた観光推薦や観光マップの生成，それらの分析などの関連研究を紹介する．3 章では，ショートビデオのラベル付けと入力写真のタグ付けについて説明する．4 章では，提案手法に基づく観光地紹介ビデオの生成手順について述べる．5 章では，生成した観光地紹介ビデオの有用性を図るための評価実験について述べる．最後に，6 章でまとめと今後の課題について述べる．

2 関連研究

近年では，Web 上の情報を用いた観光地の推薦やマップの作成，およびその分析に関するさまざまな研究やインタフェースの開発が盛んに行われている [2] [3] [4]. 特に，Flickr や Twitter などの SNS のデータからピクトリアルマップを作成するもの [5] [6] や，観光スポットと移動経路に対しする嗜好を考慮し

1 : https://www.jalan.net/
2 : https://travel.rakuten.co.jp/

た観光ルートを推薦するシステムを構築するもの [7]，SNS の投稿の本文のテキスト情報をもとに観光地を推薦するもの [8]，観光地でつぶやかれたジオタグ付きツイートから旅行者の好みにあった観光ルートを推薦する方法を提案しているもの [9] などがある．ほかにも，土田ら [10] はニューラルネットワークを用いた言語モデル Word2Vec を利用し，単語の情報をベクトル化する手法を用いて，SNS の情報から観光情報を抽出し，生成したコーパスから単語の加算を利用してほかの地域における類似スポットの発見手法を提案している．

本研究では，ジオタグから得られる位置情報を用いて観光地間の関連性を考慮している．近しいものとして，Flickr や Twitter などの SNS 上に投稿された情報を抽出し，応用する研究が活発に行われている．王と岳 [5] は，Flickr に投稿された写真に付与されている多様な位置情報や投稿日時などのメタデータを用いて，任意の観光スポットの知名度と満足度を算出し，それらを反映した人気スポットマップと穴場スポットマップの 2 種類のピクトリアルマップを生成する手法を提案している．高と牛尼 [6] は，代表的な写真型 SNS の Instagram を利用して，ジオタグ付き写真に対して都市に関するクラスタリングにより代表的な地域を抽出し，写真の画像的な特徴量を利用して，その地域の代表的な写真を抽出し，ピクトリアルマップを生成する手法を提案している．これらの研究では，写真の色情報などの画像的な特徴と位置情報や投稿日時などのメタデータからピクトリアルマップを作成している．本研究では，これらの情報の分析を行うことで，観光スポットのショートビデオの情報量の多さに注目して，観光紹介ビデオを生成する手法を提案する．

また，収集した観光情報を Word2Vec で学習し，観光推薦システムを構築する手法は盛んに研究されており，潜在的興味を発見し，その潜在的な興味に基づいた観光推薦システムを試作したもの [11] や，事前学習済み言語モデルを用いて雑談会話中から対話者の嗜好を推定し，観光地を推薦するもの [12] などがある．本研究では，クエリ検索によって見過ごされがちな潜在的な旅行候補地を，画像入力によりテキスト情報によらず，情報量の多い画像情報を用いることで解決を図っている．一方で，開地と檜垣 [11] は，Web 上にある膨大な観光テキストデータを取得し，このテキストデータを Word2Vec で学習することにより単語の分散表現を獲得し，ユーザの潜在的興味を発見した．高橋ら [12] は，オンラインショップの商品推薦や動画配信のコンテンツ推薦を行う際，情報量が少ないというコールドスタート問題を，Word2Vec で学習した雑談対話の嗜好に関する多様な情報をを用いて解消した．さらに，対話者の嗜好を推定し，観光地を推薦するシステムの構築を行った．また，本研究で使用した Azure Video Indexer を用いた研究が行われている．王ら [13] は，見逃し配信やサブスクリプションの動画配信サービスの普及により増加する飛ばし見を行うユーザを対象に，映像視聴支援システムを提案している．このシステムには映像シーンの分割が利用されている．

3　ショートビデオのラベル付けと写真のタグ決定

本章では，ショートビデオのラベル付けと入力写真のタグ決定の手法について述べる．本研究では，ショートビデオのラベル付けには Azure AI Video Indexer を活用し，入力画像のタグ付けには Azure AI Custom Vision を利用する．

3.1　ショートビデオのラベル付け

本研究では，主に観光地の公式サイトや YouTube 上の動画を対象とする．選出した 40 カ所の観光スポットに対し，10 秒から 20 秒ほどの動画を用意し，Microsoft 社の Azure AI Video Indexer[3] を利用して，各動画からラベルを抽出する．使用するラベルは，観光地ビデオ生成時の類似度を算出するために，英語表記でインデックス化されている．図 1 に示されているのは，Azure AI Video Indexer を用いて特定の動画から抽出されたラベルの一例である．

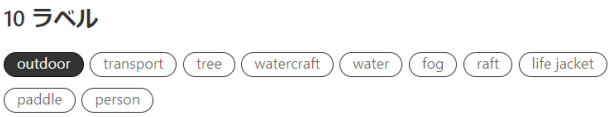


図 1　Azure AI Video Indexer のラベル抽出の例

3.2　入力写真のタグ決定

ショートビデオから抽出されたラベルの中から主要なものを選出し，「mountain」，「sea」，「tower」，「bridge」などの 25 個のタグにより入力写真を分類する．入力写真のタグ決定においては，Microsoft 社の Azure AI Custom Vision[4] を使用し，約 420 枚の観光写真を用いて学習を行った．図 2 に示されているのは，Azure AI Custom Vision を用いて特定の写真に対して決定されたタグの確率とその閾値の一例である．

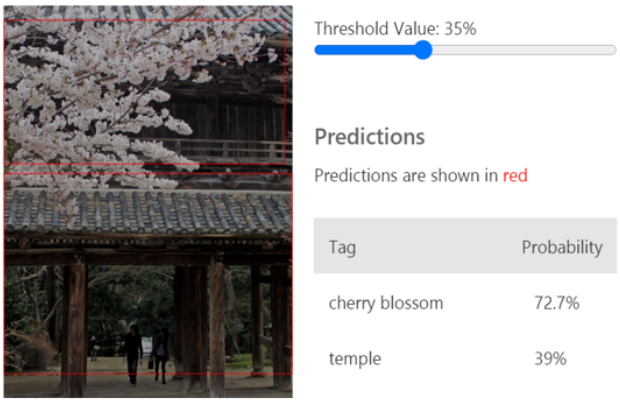


図 2　Azure AI Custom Vision のタグ決定の確率とその閾値例

3 : https://azure.microsoft.com/ja-jp/products/ai-video-indexer
4 : https://azure.microsoft.com/ja-jp/products/ai-services/ai-custom-vision

4　観光地紹介ビデオの生成

本章では，提案手法に基づいて観光地紹介ビデオの生成手順について述べる．生成する観光地紹介ビデオは，約 10 秒から 20 秒程度の短い尺の動画が 3 つもしくは 4 つ連続で再生されるように構成される．

4.1　ショートビデオからの観光スポット抽出

まず，ショートビデオから抽出したラベルと入力写真のタグのコサイン類似度を Word2Vec を用いて算出する．Word2Vec とは，自然言語処理などにおいて文章やテキストに含まれる単語を数値ベクトル化する技術のことである．Word2Vec を用いることで，単語同士の関連性や意味，ニュアンスをより正確に表現でき，精度の高い自然言語処理の実現が可能である．本研究では，この類似度の高いショートビデオが対象としている観光スポットを，入力写真と関連性の高い観光スポットとして抽出する．

4.2　位置情報に基づく観光地域の推定

GPS 機能を搭載したスマートフォンやカメラで撮影された写真には撮影地の位置情報が記録されており，これがユーザのイメージに合う観光地である可能性は高い．また，この場所の近隣の観光スポットは地理的に近いため，入力写真の観光スポットと大きな違いがあるとは考えにくい．そのため，入力写真に位置情報が存在する場合，これを用いてよりイメージに合う観光地を取り入れることが可能だと考えられる．

写真の画像情報以外にも，Exif 情報という記録があり，これには撮影日時やカメラの機種名，絞りや ISO 感度，シャッタースピードなどの設定情報が記録されている．また，ジオタグと呼ばれる位置情報もこの中に含まれ，撮影時の位置情報を記録している．入力写真から取得した位置情報とショートビデオが対象としている観光スポットの位置情報を比較し，2 点間の距離が既定値以内である場合，同じ地域と判断し，そのショートビデオの優先度をあげる．本研究では，2 地点間の距離は 2km と設定している．さらに，ショートビデオが対象としている観光スポットの位置情報に基づいて所在する地域を推定する．

4.3　紹介ビデオの生成

4.2 節で推定された観光地域を絞り込むため，観光地域の公式観光サイトの概要から特徴語を抽出し，それら特徴語と入力写真のタグのコサイン類似度を算出して観光地域を特定する．この公式観光サイトの概要における観光スポットの出現順に基づいてショートビデオを並べ，観光紹介ビデオを生成する．観光紹介ビデオには，観光スポット名とショートビデオの引用元の名前を各観光スポットのシーンの冒頭に表示する．図 3 に生成した観光紹介ビデオの表示例を示す．

5　評 価 実 験

本章では，生成した観光紹介ビデオの有用性に関する評価を行う．本研究では，山口県内の観光スポット 40 カ所を対象に，



図 3　観光紹介ビデオの表示例

観光スポットのショートビデオから付与したラベル，入力写真から決定したタグを用いて観光紹介ビデオの生成を行った．対象となる観光スポットは表 1 に示す．

表 1　観光紹介ビデオの生成に用いる対象の観光スポット

順位	観光スポット	順位	観光スポット
1	錦帯橋	21	萩城下町
2	紅葉谷公園	22	大平山公園
3	吉香公園	23	常盤公園
4	宇野千代生家	24	長門湯本温泉
5	岩国城	25	湯田温泉
6	錦川	26	周南化学コンビナート
7	角島	27	いろり山賊
8	元の隅神社	28	秋吉台サファリランド
9	海響館	29	青海島
10	秋吉台	30	長門峡
11	秋芳洞	31	長府庭園
12	防府天満宮	32	金子みすゞ記念館
13	別府弁天池	33	菊が浜
14	松下村塾	34	周防大島
15	瑠璃光寺五重塔	35	ザビエル記念聖堂
16	千畳敷	36	須佐ホルンフェルス
17	海峡ゆめタワー	37	関門トンネル
18	雪舟庭	38	キワ・ラ・ビーチ
19	寂地峡	39	船方農場
20	東後畑棚田	40	一の坂川

5.1　実 験 概 要

生成した観光紹介ビデオの有用性を評価するためにアンケート調査を実施した．アンケート調査では，被験者に実際に生成した 3 つの観光紹介ビデオのデモを見てもらい，下記の設問項目についてアンケートを実施した．

デモでは，まず，図 4 に示す入力写真をもとに生成した観光紹介ビデオを再生した．その後，図 5 に示す入力写真をもとに生成した観光紹介ビデオを再生した．最後に，図 6 に示す入力写真をもとに生成した観光紹介ビデオを再生した．

アンケート調査では，主に観光紹介ビデオの適切性に関する評価を質問する．アンケートの内容は以下のとおりである．

Q1.　観光紹介ビデオは，入力写真に対して適切な内容であった．
Q2.　観光紹介ビデオにより，観光地の魅力を把握できた．



図 4　1 つ目のデモの観光紹介ビデオの入力写真



図 5　2 つ目のデモの観光紹介ビデオの入力写真



図 6　3 つ目のデモの観光紹介ビデオの入力写真

Q3.　観光紹介ビデオを見て，その観光地に実際に訪れたいと思った．
Q4.　観光紹介ビデオで紹介された観光スポットの順番は適切だった．
Q5.　観光紹介ビデオの長さは適切だった．
Q6.　この観光紹介ビデオについて，良かった点があれば記述してください．
Q7.　この観光紹介ビデオについて，改善点があれば記述してください．

こちらのアンケートは，Q1～Q5 は 5 段階（1：全くそう思わない，2：あまり思わない，3：どちらともいえない，4：ややそう思う，5：とてもそう思う）で評価してもらった．Q1～Q3 は観光紹介ビデオの内容についての質問であり，Q4 と Q5 は観光紹介ビデオの形式についての質問である．Q6 と Q7 は記述式質問であり，詳しくわかるよう自由に記述できるコメント欄を設置した．これらのアンケート結果を用いて，提案手法により生成した観光紹介ビデオの有用性を評価した．

今回の実験では，計 11 名の被験者（男性：9 名，女性：2 名）を対象とした．

5.2　実 験 結 果

5 段階評価による Q1～Q3 の観光紹介ビデオの内容に関する評価の平均値を図 7 に示す．Q1～Q3 の結果に基づき，生成した観光紹介ビデオが写真に対して適切であったか，魅力を把握できたか，実際に訪れたいと感じたかについて，Q1 の平均値は約 3.9，Q2 の平均値は約 4.2，Q3 の平均値は約 3.6 となり，良い結果となった．

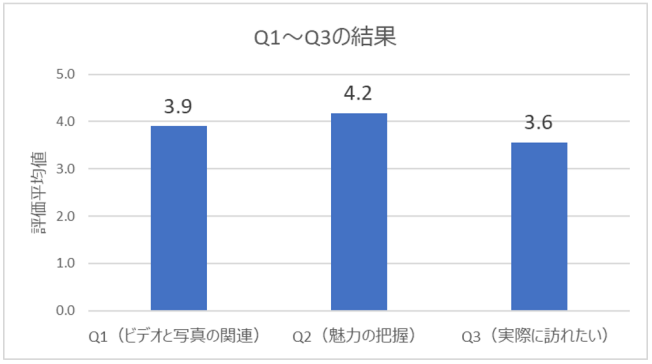


図 7　Q1～Q3 の観光紹介ビデオの内容に関する評価結果

5 段階評価による Q4 と Q5 の観光紹介ビデオの形式に関する評価を図 8 に示す．Q4 と Q5 の結果から，生成した観光紹介ビデオの順番と長さの適切さについて，Q4 の平均値は約 4.0，Q5 の平均値は約 4.3 で，非常に優れた結果となった．

Q6 では，観光紹介ビデオの良い点を記述してもらった．以下は回答の一部である．

- 入力画像から観光地を動画として提案してもらえるのは非常に観光地選びの参考になると思う．
- 写真の雰囲気が観光地紹介ビデオに再現されていた．
- 短い動画で見やすかった．
- 目を引く風景や観光スポットが魅力的に映し出されていた．

最後に，Q7 では，観光紹介ビデオの改善点を記述してもらった．以下は回答の一部である．

- 観光スポットの地理的範囲が狭い．

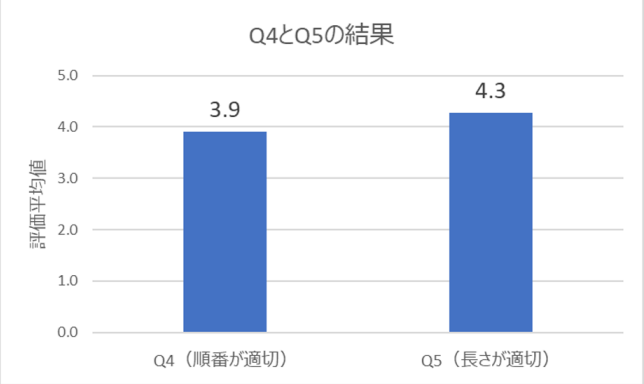


図 8　Q4 と Q5 の観光紹介ビデオの形式に関する評価結果

- 観光スポットの地図や紹介についての詳細な情報が提示できると視聴者が興味を持ちやすい．
- 市がわかればなお良い．
- どういった観光地なのか，各観光地についてもっと詳細が欲しいと感じた．
- 県内より範囲を広げると動画の多様性が出そうである．
- 桜の紹介があると良いと思った．

5.3 考　　察

Q1〜Q3 の観光紹介ビデオの内容に関する評価の結果はすべて 3.5 以上の値となり，良い結果が得られたものの，Q3 は Q1，Q2 と比較して若干低い値となった．これは，Q7 の記述回答で示された意見の中で観光スポットの詳細な情報が十分に修正する達していないことが原因と修正する考えられる．要望としてあげられた意見では，観光スポットの所在地や地理関係，地図などの詳細な情報が求められていた．Q1 は平均値が約 3.9 と良い結果が得られているが，初めに修正する用いた観光紹介ビデオには桜のラベルが付与されたショートビデオが含まれておらず，動画全体を通して寺社仏閣が主題になっていた．動画の長さを考慮して，観光紹介ビデオに使用するショートビデオは 3 つもしくは 4 つに限定していたが，桜のラベルが付けられたショートビデオをその中には取り入れることができなかった．Q7 で「桜の紹介があると良い」という意見が寄せられていることから、類似度の算出においてさらなる工夫が求められる．一方で，Q2 は平均値が約 4.2 と非常に良い結果となり，Q6 の記述からも観光スポットの魅力をより理解できていたと考えられる．

Q4 および Q5 の観光紹介ビデオの形式に関する評価結果は，非常に良い結果が得られた．しかしながら，Q4 に関しては平均値による評価は約 3.9 と良好な結果が得られたものの，Q7 の記述による改善点のコメントでは，観光スポットの地理的な関係や観光紹介ビデオ内でのストーリー性などが指摘されており，ユーザを引き付けるための改良が必要と判断される．それに対し，Q5 の動画の長さに対する評価では平均値が約 4.3 という非常に良好な結果を得ることができ，本研究のテーマであるショートビデオの活用により優れた結果が生じたと考えられる．

また本研究では，山口県内の中で観光スポットを抽出していたため，対象スポットが自然的な観光地に偏っていた．そのため，ショートビデオから抽出したラベルがアウトドア指向の強いものが多く，入力写真が都会的な建造物を多く映す画像であった場合，類似度が下がっていた．ほかにも Q7 のコメントであげられいたように，地理的な範囲が狭いこと自体がユーザの興味を引く点においてデメリットになっていると考えられる．したがって，対象となる地理的な範囲を広げる必要がある．

6　お わ り に

本研究では，観光ショートビデオおよび観光写真を収集し，観光ショートビデオにラベルを付与し，観光写真を用いて学習を行うことで，ユーザの入力写真のタグを決定した．また，ビデオラベルと写真タグのコサイン類似度を算出し，高い類似度のショートビデオから観光スポットを抽出した．抽出された観光スポットの位置情報と入力写真の位置情報に基づき，観光地域を推定した．最終的に，推定された地域の公式観光サイトにおける観光スポットの出現順に従い，ショートビデオを並べて観光地紹介ビデオを生成した．

今後の課題として，被験者による観光地紹介ビデオの有用性の評価や，画像特徴量を活用したビデオシーンと写真との類似度を算出する手法の検討などを行う予定ある．

謝　　辞

本研究の一部は，JSPS 科研費 JP21K17862 の助成および私立大学等経常費補助金特別補助「大学間連携等による共同研究」を受けたものである．ここに記して謝意を表す．

文　　献

[1] 観光経済新聞: 【データ】1 位じゃらん net、2 位楽天トラベル、8 位ＪＴＢ日本観光振興協会「2022 年観光関連サイト推計閲覧者数ランキング」. 株式会社ユニアド, 2023 年 2 月 14 日, https://www.kankokeizai.com/%e3%80%90%e3%83%87%e3%83%bc%e3%82%bf%e3%80%911%e4%bd%8d%e3%81%98%e3%82%83%e3%82%89%e3%82%93net%e3%80%812%e4%bd%8d%e6%a5%bd%e5%a4%a9%e3%83%88%e3%83%a9%e3%83%99%e3%83%ab%e3%80%818%e4%bd%8d%ef%bd%8a/ (参照: 2021 年 12 月 23 日).
[2] 川村秀憲, 鈴木恵二, 山本雅人, 松原仁: 観光情報学. 情報処理学会誌, Vol. 51 , No. 6, pp. 642-648, 2010 年 6 月.
[3] 斎藤一: Web における観光情報提供と分析 (<特集>観光と知能情報). 人工知能学会誌, Vol. 26, No. 3, pp. 234-239, 2011 年 5 月.
[4] M. Hirota, M. Endo, D. Kato, and H. Ishikawa: Discovering Hotspots Using Photographic Orientation and Angle of View from Social Media Site. in International Journal of Informatics Society, Vol. 10, NO. 3, pp. 109-117, 2019.
[5] 王元元, 岳五一: SNS 投稿写真から算出した観光スポットの知名度と満足度に基づくピクトリアルマップの生成手法の提案と評価. パーソナルコンピュータ利用技術学会論文誌, Vol. 16, No. 1, pp. 1-10, 2021 年 9 月.
[6] 高尚暉, 牛尼剛聡: SNS を利用したピクトリアルマップの自動生成. 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2017), D7-4, 2017 年 3 月.
[7] 武信雄平, 奥野拓: 観光スポットと移動経路に対する嗜好を考慮した観光ルート推薦システムの構築. 第 83 回全国大会講演論文集, Vol. 1, pp. 335-336, 2021 年 3 月.
[8] 小原基季, 森田和宏, 泓田正雄, 青江順一: Twitter 本文を用い

た観光情報抽出及び分析システムの構築. 人工知能学会全国大会論文集, Vol.29, 2015 年 5 月.
[9] 中嶋勇人, 新妻弘崇, 太田学: 位置情報付きツイートを利用した観光ルート推薦. 情報処理学会研究報告データベースシステム (DBS), Vol. 2013, No. 28, pp. 1-6, 2013 年 11 月.
[10] 土田崇仁, 遠藤雅樹, 加藤大受, 江原遥, 廣田雅春, 横山昌平, 石川博: Word2Vec を用いた地域やランドマークの意味演算. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2016), H5-1, 2016 年 3 月.
[11] 開地亮太, 檜垣泰彦: 潜在的興味に基づく観光地推薦システムの試作. 信学技報, 2015 年 7 月.
[12] 高橋正樹, 朝原隆太朗, 稲葉通将: 事前学習済み言語モデルを用いた対話データに基づく観光地推薦. 第 96 回研究会言語・音声理解と対話処理研究会, No. 96, 2022 年 12 月.
[13] 王元元 橋本樹 河合由起子 角谷和俊: ユーザの視聴操作と地理的関係性に基づく映像視聴支援システムの提案. 日本データベース学会和文論文誌, Vol. 20, No. 4, 2022 年 3 月.

マルチプレイヤRPGにおいて 人間らしいプレイスタイルをとるゲームAIの提案

藤山　仁聖† 　清　雄一† 　田原　康之† 　大須賀昭彦†

† 電気通信大学　〒 182–8585 東京都調布市調布ヶ丘 1 丁目 5–1
E-mail: †fujiyama.masatoshi@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　近年，マルチプレイヤゲームが人気を博しているが，問題点も多い．特に，1 人でも楽しめるようにすることは，多くのユーザーを獲得することに必要なことである．マルチプレイヤゲームを 1 人で楽しむことに対して，ルールベースのノンプレイヤキャラクタ (NPC) を用意し，それらとチームを組みゲームを楽しむことは多く行われてきたが，これら NPC は人間らしい行動を取らず，人間プレイヤの意図に反した行動をとる．本研究では，人間プレイヤの行動データを使用して人間らしいプレイスタイルをとるゲーム AI の提案を行った．その後，人間らしいプレイスタイルの定量的な評価と感性評価として学習済み AI と人間プレイヤとの共闘を行い人間らしさを評価した．

キーワード　ゲーム AI, 強化学習, 模倣学習, 人間らしい AI

1 はじめに

強化学習技術により，囲碁やチェスなどのゲームにおいて人間よりも強い AI が開発されてきた．しかし，これらゲーム AI は人間に理解できない行動をとることがあり問題視されている．そこで近年では，ゲーム AI のキャラクタ性を理解しようとする研究 [1] や，人間の行動データから人間らしい行動をするゲーム AI の研究 [2] [3] [5] が盛んに行われている．このような人間らしい行動生成を行う研究は，人間プレイヤを楽しませたり，活躍させたりすることが期待されている [4].

また，近年は複数プレイヤと対戦，協力するようなマルチプレイヤゲームが人気を博している．特に，協力型マルチプレイヤゲームにおいて，個々のプレイヤにはそれぞれ実現したい役割が存在している．プレイヤはゲームキャラクタのロールやステータスなどから暗黙のうちにチームのプレイスタイルを理解し，他プレイヤと協力することで，ゲームを楽しんでいる．一方でシングルプレイを好むプレイヤにとってはマルチプレイヤゲームであるため，ゲームの内容に惹かれたとしてもプレイすることを避けてしまう．ゲーム開発側の対策として，味方となるノンプレイヤキャラクタ (Non Player Character, NPC) を用意し，1 人でもゲームを楽しむことができるようにしているが，NPC の多くはあらかじめ決められた行動パターンで行動をする．これら NPC では人間のような役割に応じた協力は難しいことが問題としてあげられている．

本研究では，マルチプレイヤゲームの環境としてオリジナル RPG バトルを用いて，人間プレイヤから得られた行動データを取得し，これらの行動データをデータセットとして使用することで，エージェントに人間らしいプレイスタイルを模倣学習手法を用いて学習させる．

また，人間プレイヤによるミスまで模倣させてしまうことは，人間プレイヤの不満を引き起こしてしまう．そこで本研究では，人間らしいプレイスタイルの追求と人間らしいミスを防ぐことを両立させるため，人間らしいミスに対しては，負の報酬設定を行うことで，ミスプレイの抑制を行う．ミスを抑制させる負の報酬設定と人間らしいプレイスタイルのトレードオフの関係については，本論文の第 7.4 節にて考察する．これらのことを踏まえて，人間らしいプレイスタイルを AI に模倣させ，人間との協調を行うことが可能なゲーム AI の作成を目的とする．

2 関連研究

2.1 人間らしいアクションシーケンス生成に関する研究

S.Zhao ら [5] は Mussive Multi Online Role Playing Game (MMORPG) での人間プレイヤから行動ログを取得し，プレイヤの属性の変化によるアクションの変化を読み取り，人間らしいアクションシーケンスを生成することに成功した．

ここでは，Long Short Term Memory (LSTM) から着想を得た Long Term Action Memory (LTAM) を実装することでプレイヤの属性ごとによるアクションシーケンスの違いを読み取り，人間らしい行動シーケンスの生成に成功した．

2.2 模倣学習

ある環境下において，専門家 (以下，エキスパート) がとった方策を最適方策 π^* とした時，エキスパートが最適方策 π^* のもとで環境から与えられた状態から行動を行った軌跡データ (以下，エキスパートデータ) から，最適方策 π^* を学習していく．この軌跡データは環境から与えられた状態と選択した行動のペアデータとして与えられている [6].

2.3 Behavior Cloning

Behavior Cloning (以下，BC) は，エキスパートデータを使用した教師あり学習手法である．エージェントは，最適方策 π^* に従っているエキスパートデータから環境の状態 s を受け取り，行動決定を行う方策 π を学習する [7] [8]．エキスパートデータは，各状態 s に対応する行動 a のペアデータの軌道として与えられ，それらの軌道を学習することで，最適方策 π^* と最も一致するエージェントの方策 π_ϕ 学習する．ここで，ϕ はエージェントのパラメータを表し，最適パラメータを ϕ^* とする．このパラメータ ϕ^* は最尤推定を用いて式 1 で定式化する [8].

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \prod_{i=0}^{N} \pi_\phi(\tilde{a_i}, s_i) \tag{1}$$

N: エキスパートデータ数, π_ϕ: パラメータ ϕ の方策,
$\tilde{a_i}$: 状態 s_i を受け取った時のエキスパートの行動 a

2.4 Generative Adversarial Imitation Learning

Generative Adversarial Imitation Learning (以下，GAIL) は，Ho, Ermon ら [9] によって提唱された模倣学習のフレームワークで

あり，Generative Adversarial Network (GAN) から着想を得て提案された手法である．GAIL は，模倣学習の問題に対して，逆強化学習のアプローチを取るものであり，重み θ を持つエージェントの方策 π_θ と重み w を持つ識別器ネットワーク D_w を適合させる．ここで，$\log(D(s,a))$ は，コスト関数としての役割を果たしており，GAIL は，識別器 D に関しては，式 2 を最大化問題と捉えることができ，エージェントの方策 π に関しては，式 2 の最小化問題として捉えることができる．

$$\mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] - \lambda H(\pi) \tag{2}$$

式 2 より，GAIL は式 3 のように定式化できる．

$$\min_\pi \max_D (\mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] - \lambda H(\pi)) \tag{3}$$

(π: エージェントの方策, π_E: エキスパートの方策, D: 識別器,
$H(\pi)$: 方策 π の因果エントロピー, λ: H を制御する方策正則化因子)

2.5 方策勾配法

一般的によく知られている強化学習手法の 1 つとして方策ベースアルゴリズム [10] [11] がある．方策ベースアルゴリズムでは，エージェントの方策 π_θ の期待収益 $J(\pi_\theta)$(式 4) を最大化することを目的とする．

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[G(\tau)] \tag{4}$$

$G(\tau)$: 状態 s, 行動 a, 報酬 r で構成される軌道 τ における割引後収益

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \cdots s_t, a_t, r_t, s_{t+1})$$

エージェントの方策 π_θ は式 5 として勾配を求め，エージェントの方策パラメータ θ を式 6 に基づいて更新する．

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} G(\tau)\nabla_\theta \log \pi_\theta(A_t|S_t)\right] \tag{5}$$

$$\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta) \tag{6}$$

$\tau \sim \pi_\theta$: 方策 π_θ から得られた軌道,
$\pi_\theta(A_t|S_t)$: 状態 S_t を受け取った時の行動 A_t をとる方策 π_θ としての確率分布, α: パラメータ θ の学習率

3 人間らしいプレイスタイル

人間らしいプレイスタイルというのは，非常に複雑で評価を行うのは難しい．人間らしいプレイスタイルというのは，具体的には以下のような事例のことを指す．例えば，あるキャラクタ A は非常に攻撃に特化したステータスであることがわかっており，人間のチームが (攻撃特化キャラクタ A, 回復特化キャラクタ B, サポート特化キャラクタ C) で構成されているとする．

- 事例 1:

サポート特化のキャラクタ B，C が攻撃特化キャラクタ A にたいして攻撃力を上昇させるような行動を取ったり，効率的に敵キャラクタに対してダメージを与えるようなプレイスタイル

- 事例 2:

全てのキャラクタ A, B, C が敵にダメージを与えることで，毎ターン多くのダメージを出そうとするようなプレイスタイル

- 事例 3:

敵の行動に備えるために，守備力を向上させるようなアクションをサポートキャラクタ C はとり，万全の構えをとったのちに，攻撃的志向に変化するというプレイスタイル

以上のことから，プレイスタイルはアクションを行う順番であったり，選択したアクションに対して，どのようなターゲット選択を行ったりすることであるので，人間らしいプレイスタイルとは人間プレイヤによる行動選択順序であると定義する．

4 オリジナル RPG

本研究で扱う環境は，一般的なマルチプレイヤ RPG バトルを想定して作成したターン制コマンドバトルである．味方キャラクタと敵キャラクタが存在しており，各キャラクタの素早さ値と乱数を組み合わせた値をもとに行動順序が決定される．各キャラクタには，体力 (HP) とマジックポイント (MP) などの各種ステータスがあらかじめ決められており，HP が 0 以下になるとそのキャラクタは行動不能となる．また，防御，単体攻撃以外のアクションを行う場合では行動キャラクタの MP が消費される．アクションに必要な MP がキャラクタの保持している MP を上回っている場合は，そのアクションは無効となってしまう．

4.1 ゲームの進行

ゲームは図 1 に示す通りに進行する．

キャラクタの各行動は，“TURN NOW” フェーズで行われ，味方キャラクタもしくは，敵キャラクタどちらかが全滅した時点でゲームは終了する．

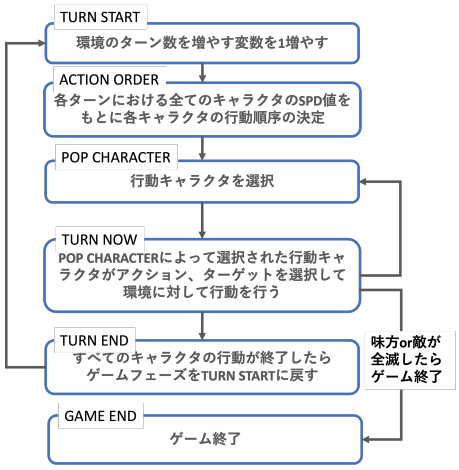


図 1: ゲームの進行

4.2 ゲームシチュエーション

本研究で用いたオリジナルゲームでは，さまざまなゲームシチュエーションにおいて人間プレイヤの行動の特徴を学習させるため，3 つのゲームシチュエーションで学習させた．登場する敵キャラクタを変更させることで，以下のような 3 つの状況を設定した．また，味方キャラクタは “戦士”，“僧侶”，“魔法使い”，“旅芸人” でありそれぞれアクションセットは異なりステータスも異なる．

- シチュエーション 1:

敵 1 対はそれほど強くはないが，5 体の敵キャラクタが登場するシチュエーション

- シチュエーション 2:

3 対の敵キャラクタが登場し，2 体が残りの 1 対のサポートをするというようなシチュエーション

- シチュエーション 3:

ステータスも他キャラクタよりも高く設定された敵キャラクタを用意し，ランダムな行動では絶対に勝てないようなシチュエーション

4.3 環境から与えられる状態について

環境となるオリジナル RPG では，エージェントなどのゲーム内キャラクタが行動を行うことで環境の状態は常に更新されるものとする．この時環境の状態は，以下の情報を含むデータとする．

- 味方キャラクタの情報

味方キャラクタのステータス情報は，キャラクタの id，ステータス情報，キャラクタの生存を表す 1bit の値が含まれている．

- 敵キャラクタの情報

敵キャラクタのステータス情報は味方キャラクタ同様の値が状態として与えられるが MP の情報は含まれない．

- ゲームの最大ターン数に対するゲームの進行ターン数の割合

5 提 案 手 法

本研究の目的は，第 3 章で定義した人間らしいプレイスタイルを模倣学習によってエージェントに学習させることである．そこで，本研究における提案手法は GAIL [9] から着想を得て実装を行い，エージェントが環境に対して行動をすることによって生成されるデータと人間プレイヤによるエキスパートデータを使用して学習を行う．

5.1 システムの概要

システム概要を図 2 に示す．

本システム (図 2) を採用したモデルは以降 “GAIL” と表記を行い，BC [7] によるデータ学習を行ったモデルは “BC” を表記する．

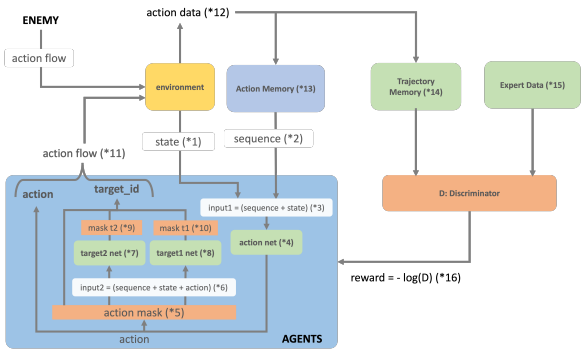


図 2: 学習環境の概要図

(*1: 環境から与えられる状態，*2: 過去の行動から直前までの行動のシーケンスデータ，*3: アクション選択を行うネットワークに与えられる入力，*4: アクション選択を行うネットワーク，*5: アクションの属性によって入力先のネットワークを制御するマスク，*6: *5 マスクによりターゲット選択ネットワークを使用するときの入力，*7: 敵キャラクタを選択するネットワーク，*8: 味方キャラクタを選択するネットワーク，*9, *10: 選択可能なキャラクタのみに制御するマスク，*11: 選択したアクションとターゲットのペアデータ，*12: 環境からの状態，キャラクタの行動，行動結果が含まれているデータ，*13: エージェントに与える行動データが保持されているメモリ，*14: 学習時に使用される行動データを保持しているメモリ，*15: 人間プレイヤによるエキスパートデータを保持しているメモリ，*16: エージェントに与えられる報酬)

一定のエピソード数エージェントと敵キャラクタが環境に対して行動を行いその結果である行動データ (図 2-“action data”(*12)) をアクションメモリ (図 2-“Action Memory”(*13)) と軌跡メモリ (図 2-“Trajectory Memory”(*14)) に追加していく．

エージェントはアクションメモリに保持されている行動データから生成されるシーケンスデータと環境からの状態を受け取りアクション，ターゲットの選択を行う．

一定のエピソード数が終了すると，識別器は軌跡メモリに保持されている生成データとエキスパートデータを入力として与えられ，生成データの時 1，エキスパートデータの時 0 と出力できるように学習を行う．

一方でエージェントは生成データを識別器に与えた時の出力に -log をとった値を報酬として期待収益を最大化させるように学習を行う．

学習後，学習時に使用された生成データは学習終了後破棄されるものとする．

5.2 環境: オリジナル RPG バトル

第 4.2 節で示した 3 つのシチュエーションを 1 エピソード毎に変更しながら，学習を行う．味方キャラクタは第 4.2 節で示した 4 体と固定した．

5.3 シーケンスデータ

本節以降のセクションにて本論文における “シーケンスデータ” とは，状態 s，アクション a，ターゲット t を連結させたデータ (連結データ) のデータ列のことである．

シーケンスデータは，過去の他キャラクタの行動や敵キャラクタの行動が含まれており，人間プレイヤは，他キャラクタの行動による環境の状態変化などを読み取り行動選択を行う．このことからシーケンスデータには，人間プレイヤのプレイスタイルが含まれている．環境から全く同じ状態が与えられたときでも，過去の行動によって選択される行動は同じではない．そこで，人間らしいプレイスタイルを学習させるため，エージェントには，環境からの状態と前回の行動から直前の他キャラクタの行動が含まれているシーケンスデータを与える．識別器に与えられるシーケンスデータは，エージェントの行動によるシーケンスデータとエキスパートデータによるシーケンスデータを与える．その後，識別器は与えられたデータがエキスパートデータか生成データかを判別するように学習を行う．

シーケンスデータに含まれるデータ

シーケンスデータは，以下に示す状態 s，アクション a，ターゲット t を連結させたペアデータ (s, a, t) が要素として含まれている．

- 状態 s:

状態は，第 4.3 節で示したゲームに登場する全てのキャラクタ情報とゲームのターン数を示した値を保持している．

- アクション a:

本研究で使用したゲームにおける全てのアクションは 23 種類ある．各キャラクタにはそれぞれのアクションセットが存在しているが，ここでは，各キャラクタのアクションセットの長さではなく，ゲームにおける全てのアクション数の長さを用いて，i 番目のアクションを選択した場合，i 番目のビットのみ 1 で他のビットを 0 であるテンソルデータとして表現される．

- ターゲット t:

ターゲットもアクション同様に全てのキャラクタ数に，選択を行わない: -1，味方全体: 0，敵全体: 1 の場合を組み合わせたターゲット選択の長さのテンソルを用意し，選択したターゲット id に対応する部分を one hot エンコーディングを行い，one-hot 表現とした．

- 連結データ:

上記の状態 s，アクション a，ターゲット t を図 3 のように連結させシーケンスデータの 1 要素として扱う

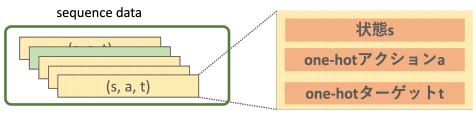


図 3: シーケンスデータの要素として与えられるデータ

識別器 D と模倣学習エージェントに与えられるシーケンスデータは，

一部異なるので，識別器，エージェントで使用するシーケンスデータはそれぞれ第 5.5 節，第 5.6 節で詳しく触れる．

5.4 エキスパートデータ

エキスパートデータに含まれているデータは，味方キャラクタの行動データと敵キャラクタの行動データ，ターン開始を表す文字列 “TURN START”，ゲーム終了時を表す文字列 “GAME END” が順番に記録されている．この仕様によりキャラクタの行動順序などを把握できる．

5.5 識別器 D

識別器 D は，入力としてエージェントが生成したシーケンスデータとエキスパートによるシーケンスデータが与えられ，与えられたデータが偽物である確率を出力する．

5.5.1 識別器 D に与えられるシーケンスデータ

識別器 D に与えられるシーケンスデータは，第 5.3 節によるシーケンスデータであり，エージェント i に対応するシーケンスデータの場合，ターン (k-1) でエージェント i の行動からターン (k) でエージェント i が行動した時までの全てのキャラクタが行った行動データ (状態, アクション, ターゲットのペアデータ) が含まれているシーケンスデータである．また，各シーケンスデータの大きさはシーケンスデータによって異なるため，サイズの異なるシーケンスデータを最大長のシーケンスデータの長さに合わせるため，値 0 で埋め込みを行い，シーケンスデータのサイズを一致させて読み込ませる．

5.5.2 識別器 D のネットワークアーキテクチャ

識別器 D はシーケンスデータが偽物かどうかの判別を行う．扱うデータがシーケンスデータであることから，ネットワークアーキテクチャとして LSTM を採用した．ネットワーク構造は，2 つの LSTM 層を組み合わせた多層 LSTM ネットワークであり，各層の LSTM は，1 層の隠れ層を持ち 128 個のユニットで構成されており，出力層には線形層を掛け合わせた．ここで，識別器 D の出力は $[-\infty, \infty]$ であり，ネットワークの出力に対して，sigmoid 関数を適用させ，出力を $[0, 1]$ の範囲に写像を行い出力を確率値として用いた．

5.5.3 識別器 D の学習

識別器 D は，与えられたシーケンスデータに対して，偽物である確率を出力するように学習を行う．そのため，識別器 D は，生成データが入力として与えられた時 1 を，エキスパートデータが与えられた時 0 を出力するように学習を行う．

識別器 D の損失関数は式 2 より，式 7 として与えられ，式 7 を最小化するように学習を行う

$$\mathcal{L}_D = -\mathbb{E}\left[\log(D(\text{generator data}))\right] - \mathbb{E}\left[\log(1 - D(\text{expert data}))\right] \tag{7}$$

また，エージェントによる生成データは，軌跡メモリ ((図 2-Trajectory Memory(*14))) に保持されており，識別器の学習時に全てのデータが使用され，学習後，軌跡メモリは空となる．

5.6 模倣学習エージェント

本研究における模倣学習エージェントは，行動を生成する生成器としての役割を持ち，前回の行動から直前のキャラクタの行動までのシーケンスデータと環境から現在の状態を受け取り，行動を確率的に選択する．また，本環境において，エージェントは個別に学習を行い，それぞれが生成した行動データからシーケンスデータが生成される．各エージェントは，各エージェントによって生成されるシーケンスデータが，識別器 D に本物だと誤判定させるように学習を行う．

5.6.1 模倣学習エージェントに与えられるシーケンスデータ

第 5.3 節にて，シーケンスデータについて記したが，エージェント i に与えられるシーケンスデータは，前回のエージェント i の行動からエージェント i の行動直前までに他キャラクタによって行われた行動によって生成された行動シーケンスデータであり，エージェント i には，シーケンスデータと環境からの状態 s_t を連結させたデータが入力データとして与えられる．

エージェントの方策を表すニューラルネットワークは，識別器 D 同様に LSTM を用いて設計となっているため，直前の状態 s_t のみをシーケンスデータに連結させることはできない．そこで，まだ決定されていないアクションとターゲットは，one-hot 表現のように全て 0 で埋め込みを行った padded-action, padded-target を one-hot 表現のアクションとターゲットの代わりに連結させる．状態 s_t と連結させたデータ (padded-data) をアクションメモリから受け取ったシーケンスデータと連結させた．

5.6.2 エージェントの行動決定

エージェントは，以下の過程に沿って行動 (アクション，ターゲット) 選択を行う．

1 現在の状態 s_t の取得

環境から現在の状態 s_t を受け取る．

2 過去行動シーケンスデータの取得

アクションメモリから前回の自身の行動から直前のキャラクタの状態行動ペアデータ (state, action, target_id) 列であるシーケンスデータをアクションメモリから受け取る．

ここで，図 4 を用いて例を示す．

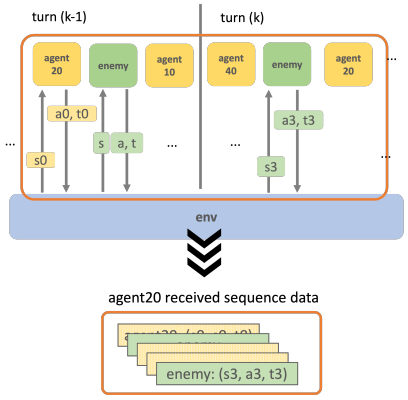


図 4: エージェント 20 がアクションメモリから受け取るシーケンスデータ

図 4-“turn(k)” においてエージェント 20 が行動する時，アクションメモリから受け取るシーケンスデータは，図 4-“agent20 received sequence data” のようなデータである．これは，前回のエージェント 20 の行動，つまり “turn(k-1)” におけるエージェント 20 の行動から “turn(k)” におけるエージェント 20 の行動直前までに発生する他キャラクタの行動による行動データを含むシーケンスデータとなっている．

よって，turn(k) でエージェント 20 が受け取るシーケンスデータは，エージェント 20 が “turn(k-1)” で行動した時に環境から受け取った状態 s0，行動 (a0, t0) による行動データ (s0, a0, t0) から “turn(k)” のエージェント 20 とエージェント 40 の間に行動した “enemy” の行動データ (s3, a3, t3) までを含んだシーケンスデータである．

3 アクション選択

アクション選択は，1 で受け取った状態 s_t を用いて図 5 のようにアクション，ターゲットをそれぞれ padded-action，padded-target 埋め込みを行ったものと連結させ，padded-data を作成する．padded-action, padded-target は以下のように値 0 で埋め込みを行った 1 次元テンソルデータとして与えられ，

• padded-action = [0] * [length of all actions]
• padded-target = [0] * [length of all target_id]

値 0 で埋め込みを行った padded-action, padded-target と受け取った状態 s_t を以下の padded-data として，padded-data を第 5.3 節-“**シーケンスデータに含まれるデータ**” で説明した連結データとして扱う．

• padded-data = (s_t, padded-action, padded-target)

その後，“**2 過去行動シーケンスデータの取得**” で受け取ったシーケンスデータ (図 4) の末尾に padded-data を追加し，これをシーケンスデータとしてエージェントの action net への入力とする．

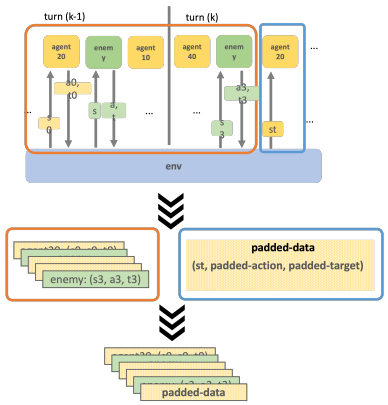


図 5: action net への入力とするシーケンスデータ

4 ターゲット選択

アクション選択後，アクションの属性によって target1 net, target2 net のどちらかもしくは，使用しないという場合があるが，target1 net (もしくは，target2 net) を使用しない場合はアクションの属性によって自動的にターゲットが選択されるので，ここでは target1 net (もしくは，target2 net) を使用する時の入力データについて説明する．

アクション選択の時に受け取った状態 s_t 以外にも，action net によってアクションが確定しているので，確定したアクションを one-hot エンコーディングを行い action-one-hot を生成する．

その後，“**3 アクション選択**” と同様に target のみを padded target として状態 s_t，action-one-hot，padded-target を連結させたデータを “**2 過去行動シーケンスデータの取得**” で受け取ったシーケンスデータに追加し，その追加したシーケンスデータを target1 net(もしくは，target2 net) への入力とする．

以上の過程により，行動 (action, target_id) が決定される．

5.6.3 模倣学習エージェントのネットワークアーキテクチャ

エージェントは，3 つのニューラルネットワークを保持しており，それぞれ action net, target1 net, target2 net であり，それぞれのネットワークの役割は，表 1 に示す通りである．

表 1: エージェントが保持する 3 つのニューラルネットワーク

ネットワーク	ネットワークの役割
action net	アクションを選択するためのネットワーク
target1 net	ターゲット選択として味方を選択するネットワーク
target2 net	ターゲット選択として敵を選択するネットワーク

各ネットワークは 3 つの LSTM 層を組み合わせた多層 LSTM ネットワークであり，各層の LSTM は，1 層の隠れ層を持ち 128 個のユニットで構成されており，出力層には線形層を掛け合わせた．模倣学習エージェントは識別器 D が生成データを受け取った時に本物であると判別させるように学習を行う．また，出力次元は，表 2 の通り．

表 2: 模倣学習エージェントの各種ネットワークの出力次元

ネットワーク	出力次元
action net	アクションセットの数
target1 net	味方キャラクタ (他の模倣学習エージェント) の数
target2 net	ゲームに登場する全ての敵キャラクタの数

各種ネットワークの出力範囲は $[-\infty, \infty]$ であり，ネットワークの外で softmax 関数をかけることで，各要素は $[0, 1]$ に写像される．softmax 関数を掛けた出力が確率分布となり，エージェントはアクションを確率的に選択する．

ターゲット選択に関しては，アクションの属性によって mask により以下の 3 つの場合で決定する．

• target1 net によるターゲット選択
target1 net によって味方キャラクタの id を target_id として選択
• target2 net によるターゲット選択
target2 net によって敵キャラクタの id を target_id として選択
• ターゲット選択を行わず自動的に決定

アクションの属性によって，target_id を選択せずに，次の 3 通りに自動的に target_id が決定される．

$$\text{target_id} = \begin{cases} -1 \text{ (ターゲット選択をしない)} \\ 0 \text{ (味方全体として選択)} \\ 1 \text{ (敵全体として選択)} \end{cases}$$

ここで，target1 net，もしくは target2 net を使う場合，現在生き残っているキャラクタしか選択できないため，target1 net を使用してターゲット選択を行う場合，mask t1 (図 2-“mask t1”(*9))，target2 net を使用する場合は mask t2 (図 2-“mask t2”(*10)) をかける．mask t1(mask t2) では，選択不可能なキャラクタには，ネットワークからの出力値が $-\infty$ となるようなマスクをかけ softmax 関数を掛けて確率分布にした時に 0 となるような処理を行った．

上記のマスク処理をした後，確率的にターゲットを選択した．

5.6.4 エージェントの学習

模倣学習エージェントは，BC による教師あり学習の他に，GAIL による学習を行うが GAIL での学習では識別器 D がエージェントの行動による生成データを本物であると識別させるような行動生成を行うように学習を行う．

BC による学習

BC では，単純な教師あり学習であるので，シミュレーションを用いた学習は行わない．損失関数として交差エントロピー誤差 $\mathcal{L}_{\pi_\theta}$(式 8) を使用し，あるシーケンスデータに対するアクションの予測を行いエキスパートの方策を学習する．エージェントは式 8 を最小化するように方策 π_θ のパラメータ θ を更新する．

$$\mathcal{L}_{\pi_\theta} = \sum_{a \in A_c} p(a|s) \log(q_\theta(a|s)) \tag{8}$$

$p(a|s)$: エキスパートが状態 s の時，行動 a をとった真の確率分布
$q_\theta(a|s)$: エージェントの方策 π_θ が状態 s の時，行動 a を行う確率分布，
A_c: キャラクタ c のアクションセット，

GAIL による学習

GAIL では，生成器としてのエージェントの学習は識別器と同時に行われる．第 5.5 節で識別器の学習については触れるので，ここでは，

エージェント部分の学習について触れる．

エージェントは，方策勾配法より学習を行う強化学習エージェントである．方策勾配法のアルゴリズムは REINFORCE [12] を採用した．識別器 D の出力に自然対数をとった値を報酬として，エージェントは自身の方策をエキスパートの方策 π^* に近づくように学習を行う．

この時，エージェントに与えられる報酬として以下の 2 つを設定した．

1 つ目は，識別器 D の出力に対する報酬．識別器 D は偽物のデータ (エージェントによる生成データ) が入力として与えられた時 1 に近づくように，エキスパートデータが与えられた時 0 を出力するように学習するため，エージェントは識別器 D が生成データをエキスパートデータであると誤認識させるように学習を行うため，式 9 を報酬として与える．

$$\mathrm{Reward}_D = -\log(D(\text{generator data})) \tag{9}$$

2 つ目は，本研究の目的であるミスプレイの抑制のために，式 10 を設定し，人間らしいミスを抑制させる．何を人間らしいミスとするかは自由に決めることができる．

環境に対して行動を行った時，環境は人間らしいミスをしたかどうかを表す信号を返すので，それを使用して報酬 $\mathrm{Reward}_{\mathrm{miss}}$ を設定した．報酬値は，負の値として NegativePoint (NP) を設定し，ミスをした時，NP を報酬として，ミスをしなかった時，0 という報酬設定をした．

$$\mathrm{Reward}_{\mathrm{miss}} = \begin{cases} -|\mathrm{NP}| \text{ (miss played)} \\ 0 \text{ (not miss played)} \end{cases} \tag{10}$$

2 つの報酬 Reward_D，$\mathrm{Reward}_{\mathrm{miss}}$ の和をエージェントに与える報酬とし，エージェントは期待収益を増大させるように，方策パラメータを直接学習する．この時方策は確率分布として表され，各種ネットワーク (action net, target1 net, target2 net) の出力に softmax 関数を掛け合わせ，確率分布としたものを使用する．

方策勾配法では，方策を表す確率値とその時の報酬値により期待収益 J とその勾配 $\nabla_\theta J$ を求め，パラメータを更新していく．ターゲット選択を行うネットワーク (target1 net, target2 net) において，target1 net はターゲット選択を行う時に mask t1 (target2 net の場合は，mask t2) を使用するが，ネットワークの更新に用いる方策を表す確率値は mask をかける前の確率値を使用することで更新を行う．

6 実　　験

6.1 エキスパートデータの収集

エキスパートデータを取集するため，人間プレイヤ 10 人から，環境の 1 から 3 までの各シチュエーションでゲームプレイを行ってもらい，データを収集した．この時，人間プレイヤには，十分なテストプレイを行ってもらい，自分の戦略や行動思考などを決めてもらった上で，ゲームプレイデータの取得に協力してもらった．各シチュエーション毎に集まったデータ数は，表 3 の通り．

表 3: 各シチュエーションにおける人間プレイヤによるエキスパートデータ数

シチュエーション	データ数
1	373
2	1257
3	1392
合計	3022

学習には，すべてのゲームシチュエーションデータを使い学習させた．

6.2 人間らしいミスについての報酬設計

本研究において人間らしいミスは，**人間プレイヤによる不注意によって起こる行動ミス**であると定義した．そこで本研究で扱った環境において人間らしいミスは **MP 制限によって行動を無駄にしてしまうこと**とした．逆に人間らしくないミスというのは人間プレイヤの技量によって起こるゲームの先読みのミスやハイリスクハイリターンの行動などによってゲームに負けてしまうことなどである．このような行動はプレイヤのプレイスタイルが表れていると考えているためミスとはしなかった．しかし，MP 制限による行動ターンを無駄にしてしまうミスは人間プレイヤの意図しているところではないと考える．以上のことから，MP 制限による行動を無駄にすることを人間プレイヤのミスとして捉えた．

6.3 実験で使用したモデルとハイパーパラメータ

エージェントは以下の 3 つの学習方法で実験を行った．

- BC のみで学習したエージェント (BC)
- GAIL のみで学習したエージェント (提案手法 1 (GAIL))
- BC で事前学習したのち GAIL によって学習したエージェント (提案手法 2 (BC+GAIL))

BC と提案手法 1(GAIL) における学習方法における各種ハイパーパラメータは表 4，表 5 の通り．なお提案手法 2(BC+GAIL) での BC と GAIL，それぞれにおける学習時のハイパーパラメータは表 4 のものと同じとした．

表 4: BC における各エージェントのハイパーパラメータ

エージェント/学習率	action net	target1 net	target2 net
10	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
20	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
30	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
40	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
バッチサイズ	64		
エポック数	1000		

表 5: GAIL における各エージェントのハイパーパラメータ

エージェント/学習率	action net	target1 net	target2 net
10	1.0×10^{-5}	1.0×10^{-5}	5.0×10^{-4}
20	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
30	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
40	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
識別器 学習率	5.0×10^{-6}	5.0×10^{-6}	5.0×10^{-6}
バッチサイズ	64		
エピソード数	10000		
学習頻度	1 回/45 エピソード		
割引率	0.95		
Negative Point	-2.0		

6.4 クラスタリングによる人間プレイヤのプレイスタイル分類

第 3 章から，人間らしいプレイスタイルは，人間プレイヤによるアクション選択の順番や，ターゲット選択の仕方に表れていると定義しているので，人間プレイヤによるエキスパートシーケンスデータには，各人間プレイヤのプレイスタイルが表れている．そこで，人間プレイ

ヤによるエキスパートシーケンスデータを時系列データと捉え，時系列クラスタリングを行うことで，プレイスタイルを定量的に分類させ，その分類器を用いてエージェントが人間らしいプレイスタイルを生成できているかを判定させる．

6.4.1 評価方法

時系列クラスタリングによってプレイスタイルを定量的に評価するため Fit Score を定義し Fit Score をもとに評価を行う．そのために Fit Score による評価に妥当性があるかをエキスパートデータを用いて評価した．

6.4.2 Fit Score について

Fit Score を求めるため以下手順でエキスパートデータのクラスタリングを行い，Fit Score を定義した．

1 エキスパートデータのクラスタリング

エキスパートデータを比率 4:1 で学習データとテストデータに分割．学習データのみを用いて時系列クラスタリングを行い，シーケンスデータの分類を行う分類器を学習させた．分類器は，テストデータを用いてクラスタリング結果の評価を行う．

2 データがどのクラスタに分類されるかを予測

予測させるデータを，エキスパートデータを用いて学習された分類器によりどのクラスタに分類されるのか予測させる．

3 クラスタに属しているかどうかを判定

予測させるデータが，エキスパートデータによって生成されたクラスタに属しているかどうかを判定．クラスタに属しているかは，エキスパートの各クラスタにおいて，その中心とそのクラスタのデータのうち最も遠くにあるデータとの距離 (最大距離) を閾値として設定．

クラスタの中心とそのクラスタに分類された予測させるデータとの各距離において，閾値以下ならば属する，そうでなければ属さないとする．

4 Fit Score の計算

予測させるデータのうち “**3 クラスタに属しているかどうかを判定**” で属していると判断された生成データの割合を Fit Score(式 11) とし，Fit Score を用いて定量的評価を行う．

$$\text{Fit Score} = \frac{\text{予測するデータ群のうち “属している” と判定された数}}{\text{予測するデータ群の総数}} \tag{11}$$

6.4.3 エキスパート学習データの時系列クラスタリングの結果

エキスパートデータのうち学習データのクラスタリングにおける最適クラスタ数はエルボー法により以下のように決定した．

- シチュエーション 1 の時は，クラスタ数 3
- シチュエーション 2 の時は，クラスタ数 4
- シチュエーション 3 の時は，クラスタ数 4

6.4.4 エキスパートテストデータによる Fit Score の妥当性

クラスタリングの精度として，第 6.4.2 節の手順でエキスパートデータのうち分類器の学習に使用していないテストデータの Fit Score を求めその結果から Fit Score の妥当性について述べる．

各シチュエーションにおいて，設定した最適クラスター数でエキスパートテストデータを予測させ，エキスパートデータの Fit Score を求めることで Fit Score の妥当性を調べた．その結果を表 6 に示した．

表 6: 各シチュエーションにおける生成データの Fit Score

シチュエーション	1	2	3
expert data (test)	0.9	0.971	1.0

表 6 の結果から，エキスパートデータの Fit Score はシチュエーション 1 で 0.9，シチュエーション 2 で 0.971，シチュエーション 3 で 1.0 であった．すべてのシチュエーションで Fit Score が 0.9 以上であることから，Fit Score はクラスタリング精度の評価指標として妥当性があると判断し，生成データの人間らしいプレイスタイルの評価を Fit Score を用いて行う．

6.5 エージェントによる人間らしいプレイスタイルの定量的評価

第 6.4.1 節にて，学習用に振り分けられたエキスパートデータによって学習された分類器を用いてエージェントによる生成データがエキスパートデータのクラスタリング結果にどれほど属しているかを定量的に評価した Fit Score の結果を表 7 に載せた．生成データは各モデル (BC, GAI, BC+GAIL) で 500 エピソードのシミュレーションを行い生成したデータである．

表 7: 各シチュエーションにおけるエージェントによる生成データの Fit Score

シチュエーション	1	2	3
BC	0.271	0.767	**0.995**
提案手法 1 (GAIL)	0.126	0.469	0.885
提案手法 2 (BC+GAIL)	**0.328**	**0.812**	0.993

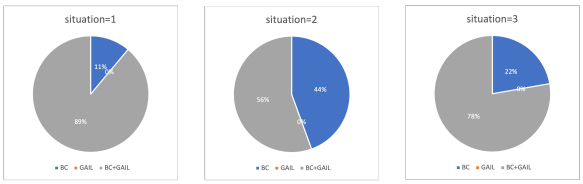
表 7 より，最も人間らしいプレイスタイルを実現しているモデルは，シチュエーション 1 と 2 では，BC+GAIL であり，シチュエーション 3 では，BC であった．この結果に対して，第 7.1 節にて考察を行う．

6.6 人間による感性評価

人間による感性評価として 9 人の人間プレイヤにアンケート調査を行うため，人間プレイヤは各モデル (BC, GAIL, BC+GAIL) で操作される操作キャラクタ以外のキャラとの共闘を行い，そのプレイスタイルについて評価を行った．人間プレイヤは操作キャラクタを “戦士”，“僧侶”，“魔法使い”，“旅芸人” の 4 種類を操作を行うが，操作キャラクタはこのうち 1 体のみとし，他の味方キャラクタは各モデル (BC, GAIL, BC+GAIL) が操作を行った．各ゲームシチュエーションにおいて 4 種類の操作キャラクタを動かし，各操作キャラクタごとに 3 種類のモデルが操作する味方キャラクタとの共闘を行い最も人間らしいプレイスタイルを実現するモデルについて評価を行った．

6.6.1 各ゲームシチュエーションにおける最良モデルについて

各ゲームシチュエーションにおいて，どのモデルが最も人間らしいプレイスタイルをとることができるのかアンケート調査を行った．その結果を図 6 に示した．



(a) シチュエーション 1 **(b)** シチュエーション 2 **(c)** シチュエーション 3

図 6: 各シチュエーションにおける最良モデル

図 6 の結果から，提案手法 2(BC+GAIL) がすべてのゲームシチュエーションで最も人間らしいプレイスタイルを実現するモデルであると回答した人数が多かった．この結果について，第 7.2 節にて考察を行う．

6.6.2 最良モデルのエージェントとの共闘による人間らしいプレイスタイルの実現について

1 人でもマルチプレイヤ RPG を楽しむために，人間らしいゲーム AI を導入し人間らしいプレイスタイルを実現させることが目的であった．そこで，人間プレイヤによる感性評価として，1 キャラクタを人間プレイヤが操作し，他キャラクタは AI による操作で共闘を行い，エージェントの人間らしさの評価を行った．操作キャラクタ以外のキャラクタのうち，人間らしい行動をしたキャラクタを選択してもらい，各キャラクタの人間らしさを評価した．最も人間らしいと回答のあった BC+GAIL に対して，各操作キャラクタにおいて他のキャラクタが人間らしいかどうかというアンケート調査の結果を図 7 に示した．この結果について，第 7.3 節にて各キャラクタに対して考察を行う．

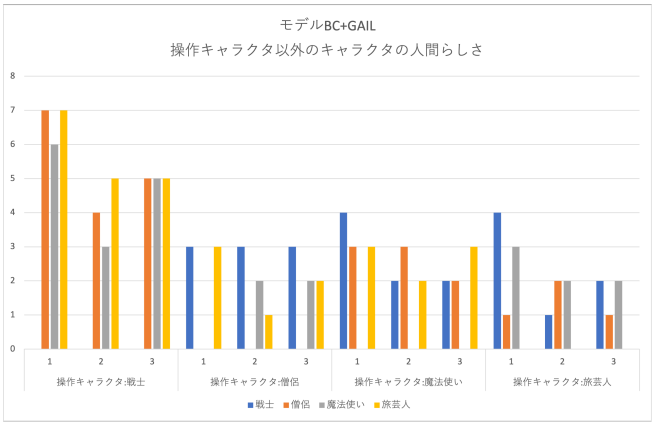


図 7: 各操作キャラクタにおける他キャラクタの人間らしさ

6.6.3 人間らしいプレイスタイルと人間らしいミスの抑制について

BC+GAIL では，人間らしいミスを抑制するため，負の報酬設定を行うことでエキスパートデータ以外から報酬の影響を受けるので，純粋な人間らしい行動をとることとトレードオフの関係である．すべてのゲームシチュエーションにおいて最も人間らしいと回答したモデルに対して，各キャラクタがミスプレイをしたかどうかについてアンケート調査を行った．その結果を図図 8 に示す．

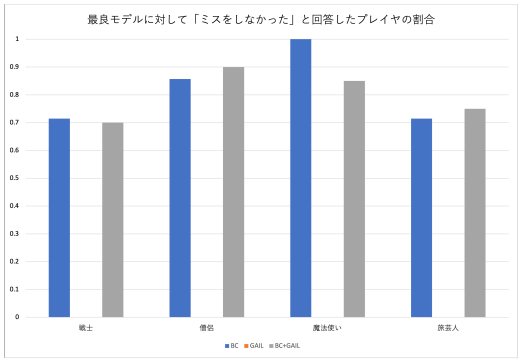


図 8: 最良モデルにおける “ミスをしなかった” と回答したプレイヤの割合

この結果 (図 8) はアンケート調査を行った人が選択した最良モデルについて評価をしてもらった結果である．最良モデルとそうではないモデルとの間に投票数による影響をなくすため各モデルの票数に対してそのモデルを選択した人数で割った値を載せている．この結果について，第 7.4 節にて考察を行う．

7 考　　察

7.1 クラスタリングによるプレイスタイルの定量的評価について

プレイスタイルについて定量的に評価を行うため，エキスパートデータに対してクラスタリングを行い，Fit Score を用いることでクラスタリング精度を評価した．人間プレイヤによる感性評価の結果と比較すると，定量的評価，感性評価のどちらにおいても，提案手法 2 (BC+GAIL) が最も人間らしいプレイスタイルを実現できるという回答が多かった．また，シチュエーション 3 においては Fit Score による評価では BC の方が BC+GAIL よりも良いスコアを出していたが，BC と BC+GAIL の Fit Score の差は 0.002 ポイントであったので，シチュエーション 3 においては，BC も BC+GAIL も人間らしいプレイスタイルの実現度としては変わらないと考えられる．

次に，各シチュエーションで Fit Score が大きく異なることからシチュエーション別にゲームの特徴と提案手法について考察する．シチュエーション 1 は，エキスパートテストデータではその精度が 0.9 であるのに対し最も良いモデル BC+GAIL では Fit Score は 0.328 であることから模倣精度は低い．シチュエーション 2，3 になるにつれ，Fit Score は大きくなっている．ここで，ゲームの特徴としてシチュエーション 1 から 3 において人間プレイヤのクリアターンを比較すると，シチュエーション 1 が最も小さく，次にシチュエーション 2，最もクリアターンが大きいのはシチュエーション 3 である．また，表 3 より，クリアターンが小さいほど得られるエキスパートデータは少なくなり，データ数が少ないほどそのシチュエーションにおける Fit Score が小さくなるということがわかる．以上のことから，データ数が生成データの Fit Score に影響をもたらしていると考えられ，また，クリアターンが短いほどプレイスタイルの確立が難しくなると考えられる．

7.2 各モデルにおける人間らしいプレイスタイルの実現について

図 6 より，シチュエーション 1 では 89%，シチュエーション 2 では，56%，シチュエーション 3 では 78%のプレイヤが BC+GAIL が最も人間らしいモデルと回答しており，残りのプレイヤは BC が最も人間らしいモデルと回答した．モデル GAIL は全てのシチュエーションで最も人間らしくないと評価された．

以上のことから，多くのシチュエーションにおいて最も人間らしいプレイスタイルを実現させてくれるモデルは提案手法 2 であるモデル BC+GAIL であると考えられる．

一方で，提案手法 1 (GAIL) は全てのプレイヤにおいて人間らしいプレイスタイルを実現しないという結果となった．考えられる原因として，GAIL は識別器と同時に学習を行なっていくアルゴリズムである．本研究で収集されたデータ数では，識別器の方が早く学習が進んでしまうという欠点があった．そのため，GAIL だと人間らしい行動を学習させようとすると識別器の学習と学習の速度が釣り合わず，結果的に識別器の精度だけが向上してしまい，BC や BC+GAIL と比較して人間らしいプレイスタイルを実現することがむずかしいと評価を受けたと考えられる．

7.3 操作キャラクタ以外のキャラクタの人間らしさ

各操作キャラクタの時の他キャラクタの人間らしさについて考察する．第 6.6.2 節にて，BC+GAIL のモデルにおいて各操作キャラクタ以外のキャラクタの人間らしいキャラクタを選択した結果 (図 7) を示した．

結果の全体図として図 7 から，操作キャラクタが “戦士” の時，他

キャラクタは他の操作キャラクタの時と比較して最も人間らしい行動をとっていることがわかる．

操作キャラクタが戦士以外の時，すべてのゲームシチュエーションにおいて人間らしい行動をしているのは “戦士” であった．これは，戦士の行動空間が他キャラクタよりも小さく行動のパターンを他キャラクタよりも学習ができていることが原因として考えられる．そのため，操作キャラクタが戦士の時は他キャラクタは戦士を動かす時のプレイスタイルを把握し適切な行動ができているという回答を多く得られており，また，操作キャラクタが戦士ではない時でも戦士は他キャラクタに比べて人間らしいプレイスタイルを実現していると多くの人間プレイヤから評価を受けた．

7.4 人間らしいプレイスタイルと人間らしいミスの抑制について

図 8 では最も人間らしいと回答したモデルのうちそれぞれのモデルにおける各キャラクタがミスプレイをしなかったと回答した人数の割合を示している．この結果から BC の方がミスをしなかったと答えたプレイヤの割合は BC+GAIL よりも多いキャラクタもいたが，一方で BC の方がよくミスをしたと回答したプレイヤの割合が多いキャラクタもいた．特に，多くの MP を消費するアクションを持つ “僧侶” と “旅芸人” において BC+GAIL では BC よりも “ミスをしなかった” と回答したプレイヤの割合は多かった．この結果から，ミスの抑制を行う報酬設計は僧侶と旅芸人にはよく働いていることがわかる．

また，モデル BC と BC+GAIL では BC+GAIL の方が最も人間らしいという結果を図 6 より得た．このことと図 8 によるミスの抑制の結果から “ミスをしない” ということも人間らしさに含まれていると考察できる．

最後に各キャラクタの人間らしさについて図 7 から結果を得た．操作キャラクタが “僧侶” の場合，クリアターン数が最も大きいシチュエーション 3 において，旅芸人の人間らしさは “魔法使い” と同じく最も低く，シチュエーション 2 においては旅芸人の人間らしさは他キャラクタと比較して最も低い．また，操作キャラクタが “旅芸人” の場合，シチュエーション 3 においては “僧侶” の人間らしさは最も低い．このことから，人間らしいミスの抑制を負の報酬として与えることで本来人間プレイヤが望んでいる MP 消費の多い行動をしない場合があり，人間らしさを一部追求できなかった．

8 おわりに

本研究では，マルチプレイヤ RPG バトルにおいて人間らしいプレイスタイルをとるゲーム AI の提案を行い，プレイスタイル分類とその人間らしさについて，評価を行った．

提案手法では，プレイスタイルの追求を過去の行動を含むシーケンスデータを使用したことでプレイスタイルを予測し模倣させることができた．

今後の展望として，より多くの被験者を用意し，現状よりもデータ数を増やすことでより安定した学習を行うことが可能になると思われる．また，より細かな人間プレイヤの思考を読み取るため，過去のシーケンスデータとして扱うシーケンスの長さを前回の行動からではなく，さらにもう 1 つ前の行動からのデータやエピソードごとに生成されるデータも新しい特徴量として使用することができれば，より細かな人間らしいプレイスタイルを考慮した行動をさせることが可能になると考えられる．その他には，データセットにないゲームシチュエーションに対しても人間らしいプレイスタイルが実現できるようにゲームの特徴を捉える特徴抽出器などでゲームの特徴を捉えられるような工夫が挙げられる．

謝　辞

本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです．

本研究は電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです．

本実験は電気通信大学倫理審査会の承認を受けており（管理番号: H23046），各被験者から同意の署名を得ました．

文　献

[1] Y. Zhou and W. Li, “Discovering of Game AIs’ Characters Using a Neural Network based AI Imitator for AI Clustering,” 2020 IEEE Conference on Games (CoG), Osaka, Japan, 2020, pp. 198-205

[2] R. F. Julien Dossa, X. Lian, H. Nomoto, T. Matsubara and K. Uehara, “A Human-Like Agent Based on a Hybrid of Reinforcement and Imitation Learning,” 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8

[3] Pan, CF., Min, XY., Zhang, HR. et al. Behavior imitation of individual board game players. Appl Intell 53, 11571-11585 (2023).

[4] 板東 宏和, 池田 心, Hsueh Chu-Hsuan, 人間プレイヤを活躍させる協力型ゲームの味方 AI(Teammate AI in cooperative games letting human players play active roles), 情報処理学会第 49 回 GI 研究発表会, 2023-03-17

[5] S. Zhao et al., “Player Behavior Modeling for Enhancing Role-Playing Game Engagement,” in IEEE Transactions on Computational Social Systems, vol. 8, no. 2, pp. 464-474, April 2021

[6] Hussein, Ahmed, et al. “Imitation learning: A survey of learning methods.” ACM Computing Surveys (CSUR) 50.2 (2017): 1-35.

[7] Bain, Michael, and Claude Sammut. “A Framework for Behavioural Cloning.” Machine Intelligence 15. 1995.

[8] Torabi, Faraz, Garrett Warnell, and Peter Stone. “Behavioral cloning from observation.” arXiv preprint arXiv:1805.01954 (2018).

[9] Ho, Jonathan, and Stefano Ermon. “Generative adversarial imitation learning.” Advances in neural information processing systems 29 (2016).

[10] Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8, 229-256 (1992)

[11] OpenAI Spinning Up “Algorithms Docs Vanilla Policy Gradient”, https://spinningup.openai.com/en/latest/algorithms/vpg.html#vanilla-policy-gradient

[12] 斉藤 康毅 . ゼロから作る DeepLearning ④ 強化学習編 . 株式会社オライリー・ジャパン , 2022 , [355]

中間音楽生成のための Transformer モデル

劉　明明†　酒井　哲也†

† 早稲田大学大学基幹理工学部情報理工学科　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †liu279072@toki.waseda.jp, ††tetsuyasakai@acm.org

あらまし　近年，深層学習を用いた音楽生成研究が盛んである．しかし，現在の主流の研究は多くが与えられた入力に基づいて後続の音楽を生成するか，あるいはテキスト記述に基づいて特定の音楽シーケンスを生成することに重点を置いている．そのため，前後の二つの音楽シーケンスに基づいて中間音楽を生成する研究が少ない．既存方法は Transformer モデルのデコーダを変更して，異なる 2 つのエンコーダの出力を受け入れ，前後の音楽シーケンスの情報を得て，中間音楽を生成する．しかし，この方法は Transformer モデルを改造し，2 つのエンコーダと 1 つのデコーダの構造を形成する必要がある．このような特定のタスクに特化したモデルの変更は，モデルの汎用性と柔軟性において問題がある．本研究では，事前処理に注目し，相対位置がある 1encoder-1decoder Transformer モデルを使用して，前後の二つの音楽シーケンスに基づいて中間音楽を生成する手法を提案する．また，既存の 2encoder-1decoder Transformer モデルと比較して，Transformer モデルの構造を変えることなく，既存手法と近い有効性を達成した．

キーワード　音楽生成，Transformer，中間音楽

1 はじめに

音楽は過去より人間の文化や感情を表現する重要な手段である．人間の感情や文化を伝える媒体として，音楽は我々の日常生活に欠かせない部分となっている．ショッピングモールやレストランなど様々な場所で音楽が取り入れられることで，生活体験が豊かになり，雰囲気も向上する．熟練した演奏家は，音楽の曲を切り替える際に過渡的な音楽シーケンスを巧みに加え，全体の流れや一貫性を高める．本研究は，同じように，前後の音楽シーケンスに基づいて中間の音楽を生成する方法を探求することを目的としている．

技術の発展，特に深層学習技術の進歩により，音楽作成に新たな手法がもたらされている．深層学習は音楽生成の分野で幅広く応用されており，符号音楽が主流の音楽生成方式となっている．符号音楽は，楽譜に基づく生成形式であり，組織的かつ離散的な記号シーケンスを利用して音楽を表現できる．この特性により，本来テキスト生成に用いられていた深層学習モデル，特に翻訳タスクで優れた成果を示した Transformer モデル [1] が音楽生成にも応用されるようになった．実際，Huang らの研究 [2] は，Transformer モデルが音楽生成分野で効果的であることを示している．

音楽生成タスクに関する研究がすでに多数存在するが，現在の主流の研究は与えられた入力に基づいて後続の音楽を生成するか，あるいはテキスト記述に基づいて特定の音楽シーケンスを生成することに重点を置いている．複数のシーケンスに基づいて中間音楽を生成するタスクの研究がまだ少ない．Hsu ら [3] は Transformer モデルのデコーダを変更して，異なる 2 つのエンコーダの出力を受け入れ，前後の音楽シーケンスの情報を得て，中間音楽を生成する方法を提案している．しかし，この方法は Transformer モデルを改造し，2 つのエンコーダと 1 つのデコーダの構造が必要となる．このような特定のタスクに特化したモデルの変更は，モデルの汎用性と柔軟性において問題がある．

そこで，本研究では，事前処理に焦点を当て，標準的な 1 つのエンコーダと 1 つのデコーダの構造の Transformer モデルを使用して，上下文に基づく中間音楽シーケンスの生成方法を提案する．この方法はモデル構造の複雑な変更を避けることができる．

2 関連研究

2.1 符号音楽

近年，Transformer モデルはテキスト生成タスクで顕著な進歩を遂げた．このモデルがテキスト領域での成功を収めたことに伴い，多くの研究者が音楽生成領域への Transformer モデルの適用を試みている．この分野での重要な進展の一つは，Huang ら [2] が符号音楽の方法を用いて Transformer モデル上で音楽生成タスクを初めて行ったことである [4]．彼らの研究は，Transformer モデルが過去のモデルよりも音楽生成において優れた効果を持つことを示した．

2.2 Transformer モデル

Transformer モデルは，Vaswani ら [1] によって初めて提案されたもので，自己注意力構造を Encoder-Decoder アーキテクチャに組み込んだモデルである．

Transformer のエンコーダ部分には，主に 2 つのサブ層が含まれている．一つは Multi-Head Attention 層であり，もう一つは Position-wise Feed-Forward 層である．これらの層は，エンコーダに入力された情報を学習することに役立つ．それぞれのサブ層の後に残差接続と正規化処理を行っている．このモデル構造の詳細を図 1 に示している．

データが Multi-Head Attention 層に入力されると，Multi-Head 数に基づいて 3 つの重みセット（Query, Key, Value）が生成される．これらの重みはドット積によって計算され，各トークンのアテンションスコアを得る．次のような式が用いられる：

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (1)$$

ここで，入力は次元 d のベクトル Query (Q) とベクトル Key (K)，およびベクトル Value (V) で構成する．

一方，Transformer のデコーダは，マスク付き Multi-Head Attention 層，Multi-Head Attention 層，Position-wise Feed-Forward 層の 3 つのサブ層で構成されている．マスク付き Multi-Head Attention 層は通常の Multi-Head Attention 層と異なり，将来の情報を予測する際に先行する情報を「覗き見」しないようにマスクが追加されている．

独特な構造と自己注意力機構を持つ Transformer モデルは，シーケンスデータの処理に新しい方法を提供している．自然言語処理分野で大きな成功を収めただけでなく，音楽生成など，他の分野でも幅広く応用されている．

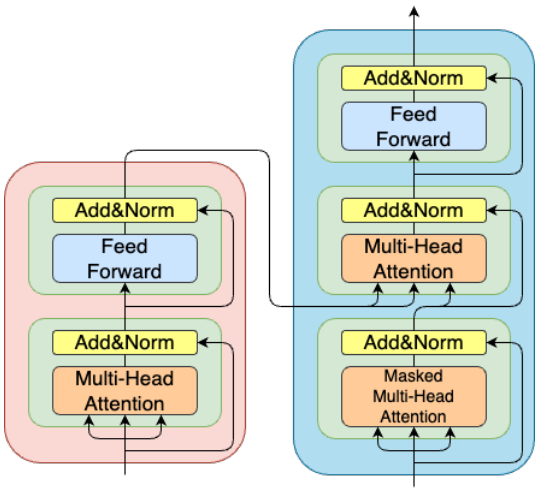


図 1　Transformer 構造

2.3　2encoder-1decoder Transformer モデル

二つの音楽シーケンスに基づいて中間音楽を生成するという課題に対して，Hsu ら [3] は改良した Transformer モデルのアーキテクチャを提案した．彼らは標準の 1 つのエンコーダと 1 つのデコーダを持つ Transformer モデルを，2 つのエンコーダと 1 つのデコーダの構造に改造した．このモデル構造の詳細を図 2 に示している．

2 つのエンコーダはそれぞれ異なる音楽シーケンス情報を処理する．各エンコーダの構造は通常の Transformer のエンコーダと同じで，Multi-Head Attention 層と Position-wise Feed-Forward 層の 2 つのサブレイヤーを含む．2 つのエンコーダからの出力を同時に処理するために，デコーダ部分は 4 つのサブレイヤーを持つように設計されている（標準のデコーダは 3 つのサブレイヤーを持つ）．最初にマスク付きの Multi-Head Attention 層があり，次に第一と第二のエンコーダからの出力をそれぞれ処理する 2 つの Multi-Head Attention 層がある．最後に，これら 2 つの Multi-Head Attention 層の出力は連結層を通じて結合される．この特殊なアーキテクチャは，二つの音楽シーケンスに基づく中間音楽の生成タスクを効果的である．

しかし，この方法は Transformer モデルを大幅に改造し，2 つのエンコーダと 1 つのデコーダの構造を形成する必要がある．この特定のタスクに特化したモデルの変更は，モデルの広範な応用と柔軟性に不利にかける．このような特殊な構造のため，他のタイプの音楽生成タスクに容易に適用できないか，さらなる調整や最適化が必要になる可能性がある．

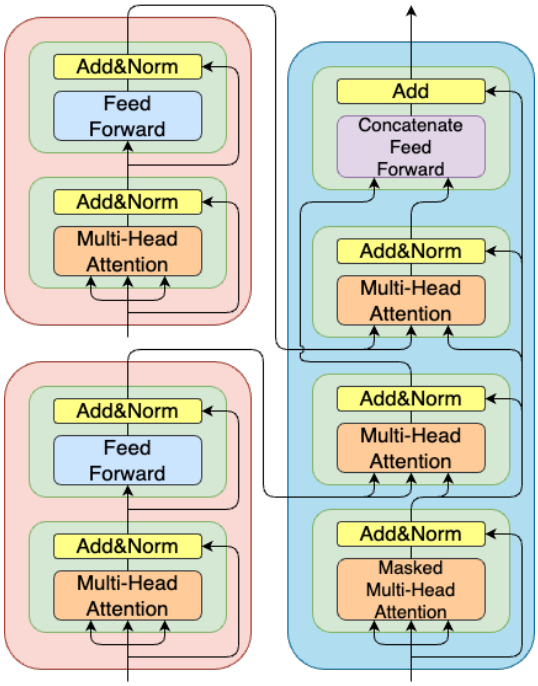


図 2　2ecoder-1decoder Transformer 構造

3　提 案 手 法

3.1　REMI（Revamped MIDI-derived event）

音楽ファイルを受け取る際，モデルをトレーニングできるように，そのファイルをデジタル形式に変換して保存する必要がある．本研究では，REMI（Revamped MIDI-derived event）というエンコーディング方式を採用している．REMI は Huang ら [5] によって提案された．REMI が提案される以前，多くの研究では Oore ら [6] によって提案された MIDI-like な表現方式が使用されていた．MIDI-like な表現方式では，MIDI 形式の音楽を 4 種類のトークンイベントに変換して保存する．これら 4 種類のイベントは，Note-On，Note-Off，Note-Velocity，Time-Shift である．Note-On は音符の開始を意味し，Note-Off は音符の終了を意味する．Note-Velocity は演奏の強度（音量としても理解される）を意味し，Time-Shift は各トークンイベント間の時間間隔を意味する．

Huang らの研究 [5] によると，音楽創作過程で人間は音楽を

複数の断片に分け，これらの断片が時には繰り返し現れる傾向がある．しかし，以前の MIDI-like な表現方式では，これらのリズム断片を完全に表現することができなかった．Huang ら [5] はこの問題を解決するために REMI を提案した．REMI の表現方式では，MIDI-like の Note-On と Note-Velocity を保持し，Note-During，Position&Bar，Tempo，Chord などのトークンイベントを追加した．Note-On と Note-Velocity は，MIDI-like な方式と同様に，音符の開始と音量を記録する．Note-During は音符の持続時間を意味し，Bar は新しいリズムの開始を示し，Position は音符の出現順序を意味する．Tempo は音符の演奏速度を意味し，Chord は複数の音符が同時に演奏されることを意味する．

Huang ら [5] はまた，MIDI-like の Time-Shift を使用すると，モデルが安定したリズムを生成できないことを発見したため，元の Time-Shift を取り除き，Note-During を使用して音符の持続時間を記録するように変更した．さらに，Position&Bar の導入により，モデルが繰り返される断片を生成することが可能になった．新たに追加された Tempo と Chord は，MIDI-like で欠けていた音楽情報を補完し，音楽生成をより完全で豊かなものにしている．REMI のこれらの改善により，モデルは人間の音楽創作プロセスをより正確に模倣し，より自然で人間の音楽感覚に合った作品を生成することができるようになった．

3.2 モ デ ル

本研究では，一つエンコーダとデコーダ構造の基本的な Transformer モデルを使用する．

我々のモデルを Hsu ら [3] が提案したモデルにできるだけ近づけるため，またモデル構造間の差異を減少させるために，基本的な Transformer モデルでも相対位置エンコーディング手法を採用している．この特別な位置エンコーディング手法は，標準的な Transformer で使用される正弦波エンコーディング手法とは異なり，トークン間の相対的な位置情報を入力シーケンスに埋め込むことができる．Huang らが提案した Music Transformer モデル [2] でも同様の相対位置エンコーディングが使用され，その有効性が示されている．

3.3 データ処理

モデル構造を再構築することと比べて，データ処理段階で前後の音楽シーケンスを直接線形的に連結し，Transformer のエンコーダに入力することで，エンコーダが前後文の情報を学習できると考える．この方法の利点は，既存のモデル構造を変更することなく，直接にモデルに前後の音楽シーケンス情報を学習させることができる点である．

具体的な方法としては，まずモデルが処理できる最大シーケンス長を max_seq として定める．次に，一曲の完全な音楽から $max_seq \times 1.5$ 個のトークンを切り取り，モデルの入力とする．第一の入力シーケンスにおいて，前半の max_seq は前段の音楽シーケンスを埋めるために使用し，後半の max_seq は後段の音楽シーケンスを埋めるために使用する．この音楽シーケンスはエンコーダの入力となる．目標とする中間音楽シーケンスは第二のシーケンスに埋められ，残りの部分は PAD マークで埋められる．実験では，max_seq を 2048 に設定し，エンコーダの入力 x において，前 1024 のトークンは前段の音楽シーケンスで埋められ，後 1024 のトークンは後段の音楽シーケンスで埋められる．デコーダの入力 tgt には，最初のトークンに START マークを使用し，続く 1024 のトークンは中間音楽シーケンスで埋められ，中間音楽シーケンスの後ろに END マークで埋められ，残った部分は PAD マークで埋められる．図 3 のようにモデルに入力する．

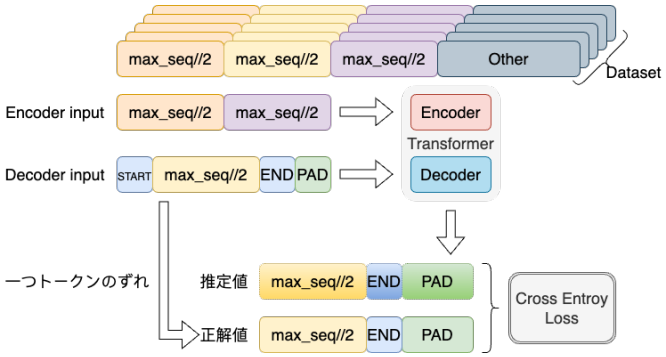


図 3　データ処理

教師あり学習を採用しているため，デコーダの入力と出力には一つトークンのずれがある．そのため，後続の損失関数の計算では，デコーダの出力を入力の START マークの後のシーケンスを利用する．図 4 のようにする．

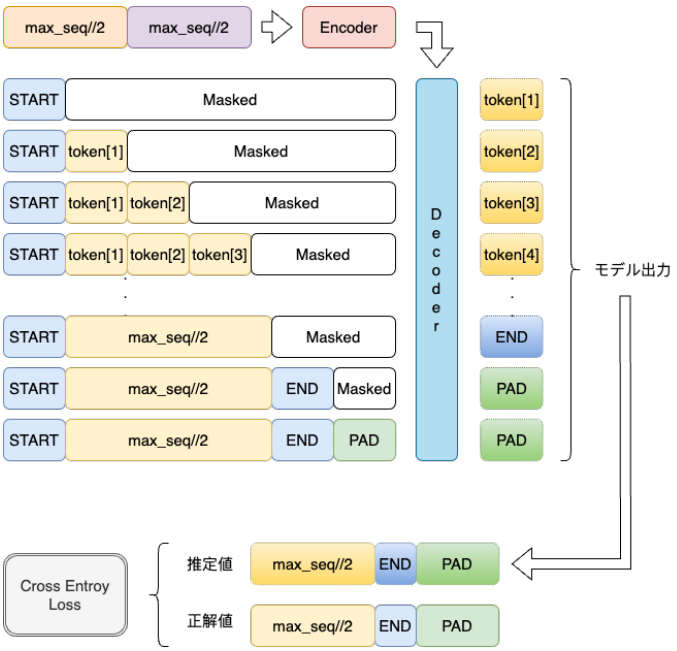


図 4　教師あり学習

3.4 損 失 関 数

本研究では，損失計算する時，交差エントロピー損失関数 (Cross Entropy Loss Function) を採用している．

交差エントロピー損失関数の計算式は以下の通りである：

$$\text{Loss} = -\sum_i y_i \log p_i \tag{2}$$

ここで，y_i は真のラベルの分布を表し，p_i はモデルによる予測確率分布である．

交差エントロピー損失関数の主な目的は，モデルの予測確率分布と真のラベル分布との差異を測定することである．この場合，損失関数の値が低いほど，モデルの予測が真のラベルに近く，つまりモデルの性能が良いことを意味する．

3.5 Adam Optimizer

本研究では，Adam Optimizer を採用している。Adam は深層学習分野で広く使われている Optimizer である。Adam は Momentum と RMSprop の長所を組み合わせている Optimizer である。Adam Optimizer の公式は以下の通りである：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{3}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{4}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{5}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{6}$$

$$x^{t+1} = x^t - \frac{\gamma}{\sqrt{\hat{v}_t} + \varepsilon} \hat{m}_t \tag{7}$$

これらの式において，m_t はモデル修正の量を調整し，v_t はモデルの損失勾配に基づいて学習率を動的に調整するために使われる。β_1 と β_2 はそれぞれ m_t と v_t の減少率を表し，g_t はモデル訓練中の損失勾配である。$\hat{m}_t$ と $\hat{v}_t$ は，モデル訓練の初期段階で m_t と v_t が急速に 0 に近づくのを防ぎ，モデルの過度な修正や訓練結果の分散を避けるために使われる。ε は分母が 0 になるのを防ぐためのパラメータで，γ は Adam で設定された学習率である。Adam Optimizer はこの方法でモデルのパラメータを調整し，訓練過程の波動を減少させ，モデルが最適解により速く収束するようにする。

本研究での Adam のパラメータ設定は:

$$\beta_1 = 0.9, \quad \beta_2 = 0.98, \quad \gamma = 0$$

$$\varepsilon = 10^{-9}, \quad \text{weight decay} = 0.0001$$

である。

3.6 Warmup

本研究では，warmup 手法を採用して，Adam Optimizer で Transformer モデルの訓練を行っている．Vaswani らによる研究 [1] で指摘されているように，Transformer モデルは訓練初期に勾配の大きな差異が生じるため，固定の学習率を使用すると学習速度が不安定になる．そのため，Transformer モデルを提案した際に Vaswani は学習率をウォームアップするための warmup 手法を使用した．

この方法は，以下の公式に従って，最初に学習率を線形に増加させ，特定の訓練段階に達した後，学習率を徐々に減少させる．

$$\text{lrate} = d_{\text{model}}^{-0.5} \cdot \min(\text{step_num}^{-0.5}, \text{step_num} \cdot \text{warmup_steps}^{-1.5}) \tag{8}$$

この warmup 手法は，Transformer モデルが訓練の初期に徐々に学習タスクに適応し，学習率が高すぎたり低すぎたりすることによる訓練効率の低下やモデルの不安定さを避けるのに役立つ．訓練が進むにつれて学習率が徐々に下がることで，モデルは訓練の後期により細かくパラメータを調整でき，モデルの全体的なパフォーマンスと安定性が向上する．また，warmup 手法は一定程度で過学習を防ぎ，モデルが訓練過程でさまざまな訓練データにより適切に適応できるようにし，最終的なモデルの品質を向上させる．

本研究では $warmup_step$ を 4000 と設定している．

4 評価実験

4.1 データセット

今回の実験では，MAESTRO [7] という音楽データセットを使用している．これは複数の MIDI 形式の音楽を含むデータセットである．このデータセットは国際ピアノ電子コンクールの原始データを採用しているため，録音された MIDI データは高忠実度を持っている．このデータセットには，国際ピアノ電子コンクールでの約 200 時間の MIDI 録音が含まれている．

データセットは 8:1:1 の比率で訓練セット，検証セット，テストセットに分割される．詳細を表 1 に示している．Miditok ライブラリ [8] を使用して，データを REMI 方式でデジタル化し事前処理を行っている．このデジタル化処理は，元の MIDI データの豊富な情報を保持するだけでなく，データを深層学習モデルの訓練により適した形式に変換する．

表 1　データセット詳細

	Performances	Size (MB)	Duration (hours)
Train	1020	64.6	-
Validation	128	8.2	-
Test	128	7.7	-
Total	1276	80.6	198.7

4.2 ベースライン

今回の実験で，ベースラインモデルは Hsu ら [3] 改良された Transformer モデルである．このモデルは 2 つのエンコーダと 1 つのデコーダから構成されるまた，二つのモデルが同じデータを使用するように，Hsu らが提案したモデル [3] の最大シーケンス長 max_seq を 1024 に設定し，我々のモデルに使用する max_seq を 2048 に設定した（これは二つのエンコーダの入力データの連結長さに相当する）. この設定により，二つのモデルが入力データを処理する際の一貫性が保証され，二つの異なる構造の Transformer モデルの性能をより公平に比較することができる．本研究で使用されたモデルのパラメータ一覧を表 2 に示している。

4.3 定量的評価

実験結果を表 3，図 5，図 6，図 7 に示している．

表 2　パラメーター一覧

	従来の手法	本研究の手法
	2encoder-1decoder	**1encoder-1decoder**
Batch size	2	2
Max sequence	1024	2048
Layers	6	6
Heads	8	8
d_{model}	512	512
Dropout	0.1	0.1
Epoch	100	100

表 3　モデルの損失と正確率

	従来の手法	本研究の手法
	2encoder-1decoder	**1encoder-1decoder**
Average Loss (train)	1.218	1.266
Average Accuracy (train)	0.571	0.556
Average Loss (val)	1.751	1.734
Average Accuracy (val)	0.428	0.432
Average Loss (test)	1.753	1.740
Average Accuracy (test)	0.430	0.432

表 3 は二つ訓練したモデルがそれぞれ訓練データセットと検証データセットとテストデータセットの損失と正確率を示している．訓練過程で教師あり学習を採用しているため，検証セットにおいても入力が正解値であると仮定する．したがって，ここでの正確率は図 4 のように得られた推定値と正解値を以下の式：

$$\text{正確率} = \frac{\text{推定値が正解値と同じ値の個数}}{\text{正解値の個数}} \tag{9}$$

で計算する．従って，訓練データセットの場合損失と正確率は 2encoder-1decoder モデルが良いが，検証データセットとテストデータセットの場合は 1encoder-1decoder モデルが良いことがわかる．全体的に本研究の手法は既存手法と近い有効性を達成したことがわかる．

図 5 は，warmup 手法を使用した後の学習率の変化を示している．訓練セットが最終的に 1020 の訓練セットに分割されたため，batch size が 2 の場合，各 epoch では 510 回の訓練が行われる．これは，8 回目の epoch 時に学習率が最大値に達することを意味し，その時の最大値は約 0.000692 で，その後学習率は徐々に下がる．

図 6 は，各 epoch 終了後の検証セットで得られた損失値を示している．これら 2 つのモデルの性能は似ていることを示している．また，2encoder-1decoder モデルが検証セットでの損失値が 1encoder-1decoder モデルより若干低いこともわかる．

図 7 は，各 epoch 終了後の検証セットで得られた正確率を示している．実験結果は，2 つのモデルの性能が非常に似ていることを確認した．また，2encoder-1decoder モデルは検証セットでの損失値は 1encoder-1decoder モデルより低いが，正確率は若干低いこともわかる．これらの結果から，全体的に本研究の手法は既存手法と近い有効性を達成したことがわかる．

4.4　定性的評価

今回の実験で，4.3 節の定量的評価に加えて，人間のリスニングテスト（定性的評価）も行った．オンラインアンケートを使って，15 曲の音楽の流暢さに関する調査を行った．この 15 曲の音楽は，5 曲のオリジナル音楽，2encoder-1decoder モデルで生成した 5 曲の音楽，1encoder-1decoder モデルで生成した 5 曲の音楽が含まれている．オリジナルの音楽と 2encoder-1decoder

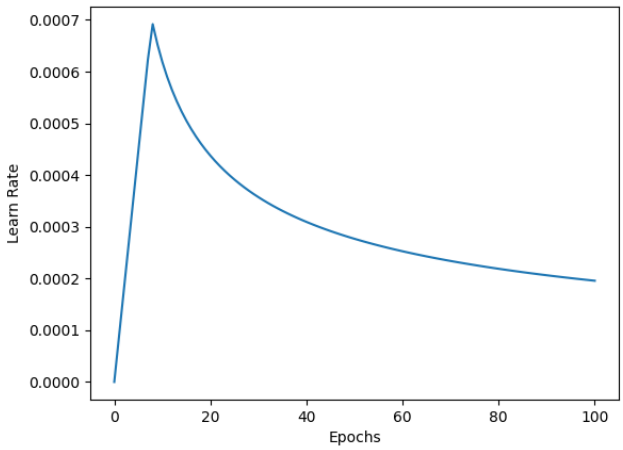


図 5　warmup 手法を使用する場合の学習率変化

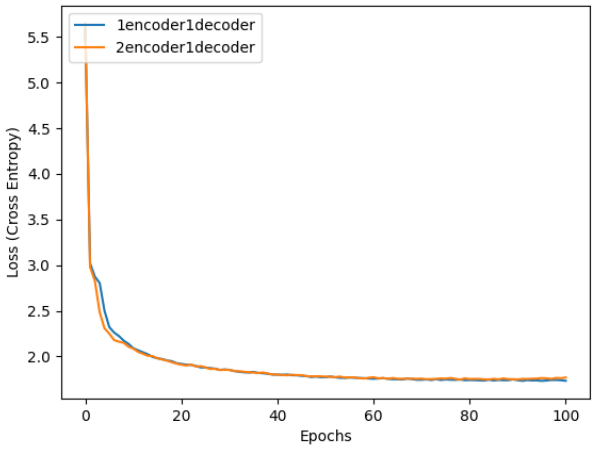


図 6　各 epoch 終了後の検証セットで得られた損失値変化

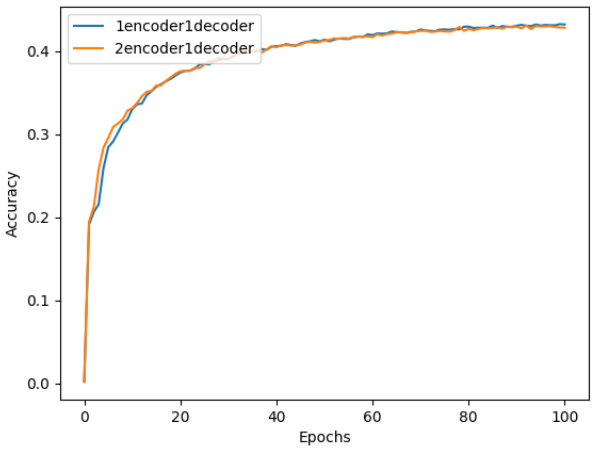


図 7　各 epoch 終了後の検証セットで得られた正解率変化

表 4　音楽の流暢さの平均点数

	従来の手法 **2encoder-1decoder**	本研究の手法 **1encoder-1decoder**	**Original**
Average Score	2.94 ($n = 90$)	2.77 ($n = 90$)	3.70 ($n = 90$)
Average Score (Never heard)	2.92 ($n = 75$)	2.79 ($n = 75$)	3.76 ($n = 76$)
Average Score (Seems familiar)	3.14 ($n = 14$)	2.79 ($n = 14$)	3.36 ($n = 14$)
Average Score (Heard)	2 ($n = 1$)	1 ($n = 1$)	-

モデルで生成した音楽は，Hsu らの研究 [3] から選んだものであり，1encoder-1decoder モデルで生成した音楽は，我々のモデルがテストセットからランダムに生成したものである．

テスト開始前に，参加者の音楽教育の背景を調査する．この 15 曲の音楽について，まず音楽を試聴してもらい，次にこれらの音楽を初めて聴くかどうかを尋ね，最後に聴いた音楽の流暢さを評価してもらう．評価基準は 1 から 5 点で，5 点が最も流暢であることを意味する．

本アンケートの参加者は 18 人である．内訳は男性 16 名，女性 2 名であり，年齢 22 歳から 29 歳まで範囲（平均年齢 24.2）である．その中で 5 名は短期間の音楽コースに参加したり，音楽授業を受けたりした経験があり，2 名は独学，残りは非音楽専門家である．リスニングテストの平均結果を表 4 に示している．全体的に詳細な箱ひげを図 8 に示している．

表 4 は流暢さの評価結果である．各シーズに，18 人が 5 曲の音楽を評価したので，各シーズのサンプルサイズは 90 である．ここで，Average Score とは，各モデルを生成した音楽を得られた平均点数である．Average Score (Never heard) とは，音楽を初めて聴くかどうかを尋ねる時，聴いたことがないことを選択した結果のみの平均点数である．Average Score (Seems familiar) とは音楽を初めて聴くかどうかを尋ねる時，聴いたことがあるような気がすることを選択した結果のみの平均点数である．Average Score (Heard) とは音楽を初めて聴くかどうかを尋ねる時，聴いたことがあることを選択した結果のみの平均点数である．全体的に見ると，2encoder-1decoder モデルで生成した音楽は 1encoder-1decoder モデルで生成した音楽より，少し良いことがわかる．また，これら 2 つのモデルで生成した音楽とオリジナル音楽との間にはまだ差があることもわかる．特に，オリジナルの曲を聴いたことがある場合，生成した音楽の不協和感が増す傾向があることがわかる．

図 8 は各シーズの 5 曲の音楽の流暢さについて箱ひげ図である．ここでは音楽を初めて聴いたかどうかでの区別はしていない．各シーズに，18 人が 5 曲の音楽を評価したので，各曲のサンプルサイズは 18 である．全体的に見ると，1encoder-1decoder モデルで生成した音楽は 2encoder-1decoder モデルで生成した音楽に比べて少し劣ってはいるが，近い流暢さを表現したことがわかる．また，生成した音楽とオリジナル音楽との間には差があることもわかる．

これらの結果から，リスニングテストでは我々のモデルが Hsu らのモデル [3] より少し劣ることがわかった．また，表 4 で示しているように，以前にオリジナル曲を聴いたことがある場合，生成した音楽を試聴する際に明らかに流暢でないと感じる傾向があることもわかった．

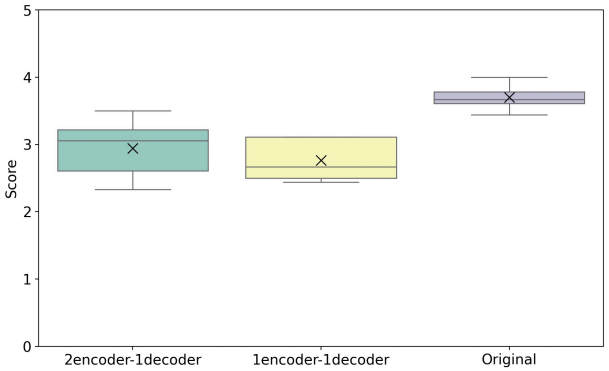


図 8　アンケート結果の箱ひげ図 ($n = 18$)

5　結　　論

今回の実験を通して，我々が提案した方法が全体的には，モデルの構造を変更せずに，Hsu ら [3] の 2encoder-1decoder Transformer モデルに近い有効性を達成したことがわかる．同時に，生成した音楽とオリジナル音楽との間にはまだ一定の差があることも発見した．特に，オリジナルの曲を聴いたことがある場合，生成した音楽の不協和感が増す傾向があることがわかった．また，前後の二つの音楽シーケンスに基づいて中間音楽を生成する研究について，今回は Transformer モデルのみを利用したので，他のモデルを検討することも今後の研究の課題である．

文　　献

[1] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[2] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam M. Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2018.

[3] Jia-Lien Hsu and Shuh-Jiun Chang. Generating music transition by using a transformer-based model. *Electronics*, 10(18), 2021.

[4] Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei

Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. Museformer: Transformer with fine- and coarse-grained attention for music generation. *ArXiv*, abs/2210.10349, 2022.

[5] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Generating music with rhythm and harmony. *ArXiv*, abs/2002.00212, 2020.

[6] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: learning expressive musical performance. *Neural Computing and Applications*, 32:955 – 967, 2018.

[7] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *ArXiv*, abs/1810.12247, 2018.

[8] Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah Seghrouchni, and Nicolas Gutowski. miditok: A python package for midi file tokenization. *ArXiv*, abs/2310.17202, 2023.

YouTube動画の視聴状況の予測

花畑 槙哉 † 許 レイ †,†† 馬 強†

† 京都工芸繊維大学 〒 606–8585 京都府京都市左京区松ヶ崎橋上町
†† 京都情報大学院大学 〒 606–8501 京都府京都市左京区田中門前町 7
E-mail: †b0122046@edu.kit.ac.jp, ††li-xu@kcg.ac.jp, †††qiang@kit.ac.jp

あらまし 本研究では，動画投稿サイト YouTube における新規動画の視聴状況の予測を行う手法を提案する．既存研究でも再生回数の予測は行われているが，動画内容やタイトル，周辺のメタデータのみを利用しており，視聴者層について十分に考慮されていない．本研究の提案手法では，まず，動画・視聴者・コメントのヘテログラフを構築し，GCN(Graph Convolution Network) を用いてユーザおよび動画の埋め込みを学習する．得られたユーザと動画の表現ベクトルに加えて，高評価数，動画時間，コメント数とその公開時間，チャンネル登録者数の特徴量を用いて，動画の再生回数を予測する．

キーワード 視聴状況，予測，表現学習

1 はじめに

動画投稿サイトの一つである YouTube は，2024 年現在，27 億人を超えるユーザが利用する大きな市場であると言える [1]．再生回数が増えればその分広告が流れる回数が増え，それによって広告収益が増えるため，効率よく再生回数を増やすことは動画投稿者（YouTuber）にとって必要なスキルと言える．しかし，再生回数は動画の内容をはじめ，タイトルや動画時間の長さ，さらには他の SNS での宣伝など，さまざまな要因が影響しており，再生回数を増やすことは決して簡単ではない．Cha ら [2] は，UGC(User-Generated Content) 動画の人気は従来のコンテンツと比べ，より短く，より予測不可能な振る舞いをしていると述べている．

新規動画に対して再生回数の予測と視聴者層の推定などの視聴状況の予測ができれば，YouTuber の支援になると考えられる．既存の研究では，主に 2 種類の手法で新規動画の再生回数を予測している．Xie ら [3] のように，動画データを利用し，動画内の特徴を抽出して再生回数を予測する手法と，柴田ら [4] のように，動画データの代わりに，動画時間やコメント数などの周辺のメタデータを多く利用している手法である．前者では動画の内容を予測に反映させているため，公開される前の動画に対しても予測をすることができる．後者では，メタデータを利用しているため，流行している動画形式やジャンルなどを利用して予測をすることができる．しかし，これらの方法では，視聴状況を予測するには視聴者の情報を十分に考慮していない．視聴者の情報が反映されている部分が再生回数と高評価数のみである．視聴者の情報を反映させることによって，再生回数予測の精度が上がると共に，視聴者層の推定ができると考えられる．

YouTube アナリティクス [5] は，動画投稿者に対し，そのチャンネルの視聴者層，人気の動画，ユーザが見ているチャンネル，ユーザが見ている動画，インプレッションと総再生時間の関係といった詳細情報を提供している．

視聴者層 では，年齢，性別，地域以外の情報は提供されておらず，その三点が分かったとしてもその区分に該当する YouTube のユーザ数は膨大であり，あまり有益な情報とは言えない．**人気の動画**，**ユーザが見ているチャンネル**，**ユーザが見ている動画**，**インプレッションと総再生時間の関係**では，視聴者の動向が示されてはいるものの，あくまで動画投稿者に対して，視聴者の既存の動画に対する情報をそのまま示しているだけである．

既存研究とアナリティクスのこれらの課題に対し，本研究では YouTube のコメント欄を利用して，視聴者・動画・コメントの相互関係を考慮した視聴状況の分析手法を提案する．コメント欄には視聴者の考えや感じたことなどがそのまま書き込まれているため，これを利用することによって視聴者の情報を予測に反映させることができる．既存研究では動画に対して視聴者が存在するだけであった関係に，動画に対してのコメント，視聴者に対してのコメントという関係が追加されることになる．

本研究では，図 1 に示すように，動画・視聴者・コメントの関係を考慮して，動画・視聴者・コメントの三種類のノードからヘテログラフを構築し，これに GCN を適用して得られる表現ベクトルを用いた動画の視聴状況を分析する手法を提案する．提案手法では，GCN を用いて学習した動画の表現ベクトルに加えて，動画時間とチャンネル登録者数の特徴量を併用して，動画の再生回数を予測する．また，動画の表現ベクトルの類似度と，GCN から得られた動画と視聴者の評価値を利用して，動画の視聴者層を推定し，高評価する視聴者とそのコメントを予測する．

本論文の構成は以下である．第 2 節では，本研究の関連研究の説明と，本研究の位置づけを述べる．第 3 節では，本研究の提案手法について述べる．第 4 節では，実験の概要，結果，それに伴う考察を述べる．最後の第 5 節では，結言を述べる．

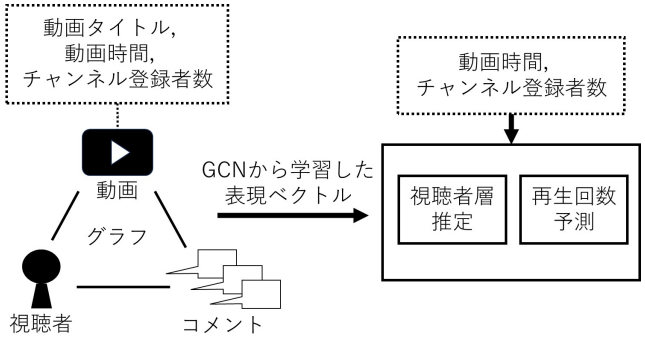


図 1: 提案手法

2 関連研究

この節では新規動画の再生回数の予測に関する研究や，公開された動画の初期の人気度から将来の人気度を予測する研究，コメント欄から視聴者の状況を分析する研究について述べ，本研究の位置付けを説明する．

Xie ら [3] は，深層学習を用いた手法 [6] [7] による再生回数予測の説明可能性の問題を提起し，解釈可能な再生回数予測手法を提案している．動画データに加えて，動画タイトルやチャンネル ID などのメタデータを利用している．動画データからは，画質と動画の音声をテキスト形式に変換したものを利用しているため，動画内容を予測に反映していると言える．本研究では彼らと同じく解釈可能な予測手法を提案している．しかし，コメントを再生回数の予測に使用する点が異なる．

柴田ら [4] は，YouTube と同じ動画投稿サイトの一つである，ニコニコ動画の動画に対して再生回数の予測をする回帰分析モデルを構築している．利用するデータは，コメント数，マイリスト登録数，投稿時間，動画再生長，他の動画からのリンクの数と他の動画へのリンクの数の計六つである．これらのデータを説明変数として回帰分析を行っている．本研究では，YouTube から取得できるメタデータのうち，動画時間，チャンネル登録者数，コメントの三つを利用して再生回数の予測を行い，コメントデータを考慮した予測手法を評価する．

Nia ら [8] は，YouTube の動画に対して初期段階の人気度から，将来の人気度を予測する手法を提案している．全動画においてユークリッド距離やピアソン相関距離からクラスタリングを行い，各クラスタの人気度の変化パターンを取得する．ユークリッド距離を用いたクラスタリングでは一日の再生回数が同程度の動画が，ピアソン相関距離を用いたクラスタリングでは一日の再生回数の変化量が同程度の動画が，同じクラスタに属する．こうして得られた変化パターンを利用して DBN(Deep-Belief Network) による学習を行い，今後の再生回数を予測している．再生回数を予測しているが，対象が公開されて数日が経過した動画であり，公開されて間もない動画や未公開動画に対して再生回数を予測することは困難である．本研究では公開する前の動画に対して再生回数を予測することを目的としている．

堺ら [9] は，YouTube の動画とコメント欄に注目し，炎上動画の分類器を提案している．利用するデータは動画の再生回数，コメント数，各コメントをセンチメント分析して得られた感情スコアと高評価数の四つの表現ベクトルである．コメントデータを用いて動画をベクトル化し，投稿された動画の状況を分析している点では本研究と同じであるが，我々は再生回数や視聴状況の予測を行う．本研究ではセンチメント分析は利用しておらず，動画・視聴者・コメントからなるヘテログラフを作成し，その三者の相互関係を考慮して視聴者の状況を分析している．

[3] と [4] では再生回数の予測を行っている点で本研究と共通しているが，視聴者やそのコメントの情報を十分に利用していない．また, [9] では視聴者の情報を利用，分析しているが，あくまでコメント欄の状況しか分析していない．本研究では，コメント欄から得た視聴者の情報を利用し，動画本体ではなく視聴者に焦点を当てて再生回数の予測と視聴者層の推定を行っている．また，このような手法は我々の知る限りは存在しない．

3 提案手法

この節では提案手法について説明する．本研究では再生回数の予測と，視聴者層の推定の二つの手法によって視聴状況の分析を行う．

各動画に対して，動画タイトルデータやコメントデータをそのまま特徴量とすると，その動画に対する視聴者しか考慮できない．そのため，本研究では動画・視聴者・コメントの三種類のノードからなるヘテログラフを構築し，それに LightGCN [10] を適用して，分類モデルに利用する特徴の表現ベクトルを獲得する手法を提案する．こうすることで，対象としたすべての動画の視聴者の影響を考慮することができる．

再生回数の予測では，まず，YouTube の動画タイトル，動画時間，チャンネル登録者数，コメントの全文を用いて，動画・視聴者・コメントの三種類のノードからなるヘテログラフを構築し，それに LightGCN を適用して，動画表現ベクトルを獲得する．得られた動画表現ベクトルに加え，動画時間，チャンネル登録者数，再生回数を用いたクラス分類モデルを構築し，再生回数の予測を行う．

視聴者層の推定では，再生回数予測の手法と同様に LightGCN を利用して動画表現ベクトルを獲得する．さらに視聴者から動画への評価値である Rating を獲得し，その値を基に視聴者層から動画への評価の予測値を算出し，推定を行う．

3.1 ヘテログラフの構築

動画ノードでは各動画に対して，動画タイトル，動画時間，チャンネル登録者数のデータをベクトル化したのち組み合わせて表現ベクトルを初期化する．視聴者ノードでは全動画のコメントデータをクラスタリングして仮想的な視聴者を作成したのち，そのクラスタに含まれるコメントのベクトルで表現ベクトルを初期化する．コメントノードでは，すべての動画のすべてのコメントに対して，そのコメントデータをベクトル化したもので表現ベクトルを初期化する．

3.1.1 コメントノードの作成

コメントノードの表現ベクトルの初期化手順は図 2 の通りである．コメントノードは，各コメントデータを多言語に対応する，mBERT（Multilingual BERT）[11] を用いてベクトル化して利用する．YouTube から取得したコメントデータは絵文字や記号が多く含まれており，そのままではエンベディングした際にノイズが入ってしまう．そのため前処理として絵文字と記号の除去を行う．その後各コメントを mBERT にて処理を行い，768 次元のベクトルを得る．そうして得たデータを PCA(Principal Component Analysis) にて次元削減を行い，64 次元のベクトルに変換する．これを用いてコメントノードの表現ベクトルを初期化する．

3.1.2 動画ノードの作成

動画ノードの表現ベクトルの初期化手順は図 3 の通りである．動画ノードは，動画タイトル，動画時間，チャンネル登録者数の計三つの属性データを組み合わせて表現ベクトルを初期化する．動画タイトルデータはコメントデータと同じく絵文字や記号が含まれている．ベクトル化した際のノイズを除去するために絵文字と記号の除去を行う．その後 mBERT を用いてベクトル化を行う．他のデータは KDE（Kernel Density Estimation）[12] を利用したグループ分類手法（a) 節を参照）を用いて五つのグループに分類した後，位置エンコーティングベースのグループエンコーティング手法（b) 節を参照）を用いてベクトル化を行っている．

- **動画タイトル：**動画のタイトルに絵文字と記号の除去を行った後，mBERT を用いて 768 次元のベクトルを得る．得られたベクトルを GCN の表現ベクトルの初期化に利用するため，PCA を用いて 64 次元に削減する．
- **動画時間，チャンネル登録者数：**動画時間（単位は秒），チャンネル登録者数をそれぞれ KDE を利用したグループ分類手法を用いて五つのグループに分類し，動画時間，チャンネル登録者数の小さい順にグループ番号 *grp* を振る．その後グループ番号を，グループエンコーティング手法を用いて，64 次元のベクトルデータに変換する．

a） KDE を利用したグループ分類手法

KDE とは，与えられたデータから，そのデータの確率密度関数を推定する手法である [12]．提案手法では，各データから KDE によって確率密度関数を推定する．確率密度関数の面積を 0 から始めて五等分するように積分し，得られた各データの値域を五つのグループの値域とする．

b） グループエンコーディング

提案手法では，GCN に表現ベクトルを入力する必要があったため，利用する各データの次元数を合わせる必要がある．そこで，位置エンコーディング [13] にて利用されている手法を拡張したグループエンコーディング手法を用いる．本研究では正弦波を利用した位置エンコーディング（Sinusoidal Positional Encoding）手法を拡張する．グループエンコーディングを行う式は式 (1) で表される．式中の *grp* は，グループの番号を表している．また，*d* は出力されるベクトルの次元数を表しており，自由に次元数を変更することができる．本研究では 64 としている．これにより，一つの数値データを，自由な次元数のベクトルデータに変換することができる．

$$\mathrm{GE}(grp,\ d) = \begin{cases} \sin\left(\frac{grp}{10000^{2d}}\right) & \text{if } d \text{ is even} \\ \cos\left(\frac{grp}{10000^{2d}}\right) & \text{if } d \text{ is odd} \end{cases} \tag{1}$$

こうして得られた動画タイトルベクトル，動画時間ベクトル，チャンネル登録者ベクトルの和をとり，動画ノードの表現ベクトルを初期化する．

3.1.3 視聴者ノードの作成

視聴者ノードの表現ベクトルの初期化手順は図 4 の通りである．視聴者ノードは，各視聴者が投稿した複数のコメントデータをベクトル化したものの平均ベクトルで，表現ベクトルを初期化する．しかし，YouTube 動画のコメント欄から視聴者ノードをそのまま作成しようとすると，異なる動画に同一ユーザが出現することは稀であるため，非常に疎なグラフになってしまう．そこで，以下の方法 a) 節で仮想視聴者を作成し，その問題を解決する．

a） 仮想視聴者の作成

3.1.1 節にてコメントデータをベクトル化したものを利用する．全コメントのベクトルにおいて K-means 法を用いてクラスタリングを行う．クラスタリングの際にはエルボー法を用いて，最適なクラスタ数（＝仮想視聴者数）を算出する．こうして得られたクラスタには同じようなベクトルを持つコメントが集まっているため元のコメントは類似している．これを仮想視聴者と捉えることができる．さらに，「同じ動画に同じ視聴者が同じようなコメントをすることは少ない」と仮定し，同じクラスタ内に同じ動画に対するコメントがあった場合には新しく別のクラスタを作成し，各クラスタ内で同じ動画に対するコメントが無いようにデータを整理する．

各仮想視聴者において，クラスタ内のコメントベクトルから平均ベクトルを算出する．PCA を用いて 64 次元に削減し，視聴者ノードの表現ベクトルを初期化する．

3.2 LightGCN を利用した表現学習

本研究では，LightGCN [10] を用いて視聴者や動画の表現を学習する．LightGCN はユーザとアイテムの重要度を排除し，ノードすべてが同じ重要性を持つと仮定している．よって計算が効率的という特徴がある．LightGCN での学習手順は以下の通りである．3.1.1 節，3.1.2 節，3.1.3 節で作成したノードをもとにグラフを作成し，LightGCN を用いて学習を行う．動画，視聴者，コメントのうち二つから構成されたグラフを入力とする．動画-視聴者のグラフ（VUgraph（Video-User-graph））では，動画とその動画の視聴者の間にエッジが張られる．視聴者-コメントのグラフ（UCgraph（User-Comment-graph））では，視聴者とその視聴者が投稿したコメントの間にエッジが張られる．動画-コメントのグラフ（VCgraph（Video-Comment-graph））では，動画とその動画に投稿されたコメントの間にエッジが張られる．3.1 節にて生成した表現ベクトルを学習の初期値とする．初期値を基に学習を行い，三種類のグラフの場合において，入力したグラフのノードに対応する表現ベクトルが出力として得

られる．

3.3 再生回数予測

再生回数予測の手順は図 5 の通りである．再生回数予測は，SVC（Support Vector Classification）[14] を利用したマルチクラス分類問題としてモデリングする．学習に使う動画の再生回数を，KDE を利用したグループ分類手法（a) 節を参照）を用いて五つのグループに分類し，再生回数の値域を得る．これを再生回数の予測に使用するクラスの値域とし，小さい順にクラス番号を振る．クラス番号を用いて LightGCN の学習に使用する表現ベクトルの初期化を行う．LightGCN の学習（3.2 節を参照）で得た各表現ベクトルと動画時間などのメタデータを利用して，再生回数のクラス分類を行う．表現ベクトルの利用方法は二つである．一つ目は VUgraph と VCgraph にて得た動画表現ベクトルをそれぞれそのまま用いる方法である．二つ目は VUgraph，VCgraph の場合で得られた動画表現ベクトルを足し合わせたベクトルを用いる方法である．こうして得られた動画表現ベクトルと，動画時間，チャンネル登録者数の計三つのデータに対して，3.3.1 節に示す処理を行い，説明変数とする．本論文の目的は新規動画に対して再生回数の予測を行うため，既知のメタデータのみを使用する．

3.3.1 説明変数の設定

各データをそのまま説明変数として利用してもよいが，本研究では再生回数の予測を行うため，単位を再生回数に合わせて説明変数とする．動画ベクトルの場合とそれ以外のデータの場合で別々の処理を行う．

a） 動画ベクトルの場合

動画ベクトルを用いた再生回数は，式 (2)，(3) で計算される．K-means 法を用いてクラスタリングを行う．クラスタリングの際にはエルボー法を用いて，最適なクラスタ数を算出する．こうして得られたクラスタ c_i で，それぞれ平均ベクトル $VclusterVec(c_i)$ と平均再生回数 $VclusterView(c_i)$ を計算する．その後，予測をする対象の動画 x の動画ベクトル $videoVec(x)$ と各クラスタ c_i の平均ベクトル $VclusterVec(c_i)$ とのコサイン類似度 $vsim(x, c_i)$ を計算する（式 (2)）．この類似度を重みとして，各クラスタの平均再生回数との積をとる．その後全クラスタ分の積の和をとり，説明変数とする（式 (3)）．このようにして，表現ベクトルであったデータを再生回数に変換する．式 (3) 中の N は動画クラスタの数であり，$viewofVideo(x)$ は動画 x の動画ベクトルを使用して算出される再生回数である．

$$vsim(x, c_i) = \frac{videoVec(x) \cdot VclusterVec(c_i)}{\|videoVec(x)\|\|VclusterVec(c_i)\|} \tag{2}$$

$$viewofVideo(x) = \sum_{i=1}^{N} (vsim(x, c_i) \cdot VclusterView(c_i)) \tag{3}$$

b） 動画ベクトル以外の場合

動画時間，チャンネル登録者数を用いた再生回数は，式 (4) で計算される．まず，KDE を利用したグループ分類手法（a) 節を参照）を用いて各データ（動画時間，チャンネル登録者数）を五つのグループに分ける．こうして得られた動画時間グループ Lg_i，チャンネル登録者数グループ Cg_i において，平均再生回数 $LgroupView(Lg_i)$ と $CgroupView(Cg_i)$ を計算する．その後，予測をする対象の動画が，各データで属しているグループの，平均再生回数を説明変数とする．このようにして，動画時間とチャンネル登録者数を再生回数に変換する．式 (4) 中の $Lmin(Lg_i)$ と $Lmax(Cg_i)$ は動画時間グループ Lg_i の最小値と最大値であり，$Cmin(Cg_i)$ と $Cmax(Cg_i)$ はチャンネル登録者数グループ Cg_i の最小値と最大値である．$Len(x)$ は動画 x の動画時間を，$Chan(x)$ は動画 x のチャンネル登録者数である．$viewofLength(x)$ は動画 x の動画時間を使用して算出される再生回数である．$viewofChannel(x)$ は動画 x のチャンネル登録者数を使用して算出される再生回数である．

$$\begin{cases} viewofLength(x) = & LgroupView(Lg_i) \\ & \text{if } Lmin(Lg_i) \leqq Len(x) < Lmax(Lg_i) \\ viewofChannel(x) = & CgroupView(Cg_i) \\ & \text{if } Cmin(Cg_i) \leqq Chan(x) < Cmax(Cg_i) \end{cases} \tag{4}$$

式 (3)，(4) から得られた三つの説明変数を用いて再生回数のクラス分類を行う．SVC によるクラス分類は式 (5) によって行われる．式 (5) 中の $view(x)$ は，動画 x の予測再生回数である．

$$\begin{aligned} &view(x) \\ &= SVC(viewofVideo(x), viewofLength(x), viewofChannel(x)) \end{aligned} \tag{5}$$

3.3.2 新規動画に対する分析

本研究の目的は，新規動画に対して視聴状況の分析を行い，YouTuber の支援を行うことである．そのため，新規動画の情報から再生回数を予測する必要がある．しかし，新規動画の表現ベクトルを得るために，新規動画を作成するたびに学習をしていては時間的コストがかかる．この問題を解決するために以下のように，類似する既存動画を参照する方法を用いる．

新規動画のデータのうち，動画タイトルのみを利用する．動画タイトルに 3.1.1 節と同じく絵文字と記号の除去を行う．その後 mBERT によって 768 次元のベクトルを得る．得られたベクトルと，ヘテログラフの作成 3.1 節にて利用した既存動画のタイトルベクトルとのコサイン類似度計算を行う．動画 x と既存動画 v_i のタイトル類似度計算の式は式 (6) に示される．新規動画のタイトルベクトルが 768 次元であるため，類似度計算をする際は，PCA で 64 次元に削減する前の，既存動画タイトルベクトルを利用する．式 (6) を用いて，式 (7) に示すように，新規動画 x の動画タイトルベクトル $titleVec(x)$ と各既存動画タイトルベクトル $titleVec(v_i)$ との類似度を重みとし，各既存動画 v_i の GCN で得られた動画の表現ベクトル $videoVec(v_i)$ と

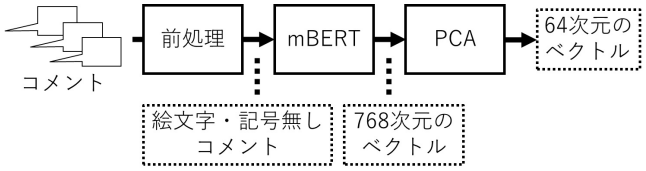


図 2: コメントノードの表現ベクトル初期化手順

の積をとる．その後，任意の数の既存動画分の積の和をとることで，疑似的な新規動画の表現ベクトル $imvideoVec(x)$ を得ることができる．式 (7) 中の M は使用する既存動画の数である．$imvideoVec(x)$ は，新規動画 x の疑似的な表現ベクトルである．

$$tsim(x, v_i) = \frac{titleVec(x) \cdot titleVec(v_i)}{||titleVec(x)||||titleVec(v_i)||} \tag{6}$$

$$imvideoVec(x) = \sum_{i=1}^{M} (tsim(x, v_i) \cdot videoVec(v_i)) \tag{7}$$

新規動画に対して得られた動画表現ベクトルを，以下の二つの手法にて利用する．

- 再生回数の予測

新規動画において，3.3 節に示した再生回数予測を行う．式 (2) で使用する動画表現ベクトルには，式 (7) を用いて疑似的に生成したものを利用する．式 (4) で使用する動画時間とチャンネル登録者においては，動画公開前に既に分かっているため，そのまま使用する．

- 視聴者層の推定

LightGCN にて得られる隣接行列は，0 か 1 によるエッジの存在のみを表す値ではなく，ノードからノードへの評価を表す値（Rating）である．視聴者層の推定では，既存動画の視聴者層から既存動画への評価を使用する．式 (6) を用いて，式 (8) に示すように，新規動画 x と既存動画 v_i の類似度 $tsim(x, v_i)$ を算出し，既存動画の視聴者層 u_i から既存動画 v_i への評価との積を取る．この値が既存動画の視聴者層から新規動画への評価の予測値 $preRating(x, u_i)$ となる．この値が大きければその視聴者層は新規動画を見る可能性が高く，小さければ見る可能性は低いと考えれらる．よって新規動画に対する視聴者層の推定を行うことができる．M は既存動画の数であり，$preRating$ は視聴者 u_i の動画 x に対する評価の予測値である．

$$preRating(x, u_i) = \sum_{i=1}^{M} (tsim(x, v_i) \cdot Rating(v_i, u_i)) \tag{8}$$

4 実　　験

本研究では提案手法を検証するため，動画の再生回数予測に関する実験（以下では実験 1 と記す）を既存動画と新規動画に対して行い，視聴者層推定に関する実験（以下では実験 2 と記す）を行っている．

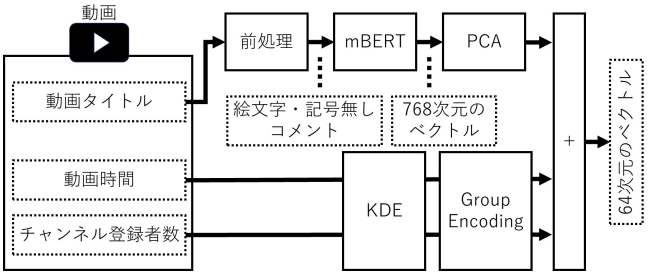


図 3: 動画ノードの表現ベクトル初期化手順

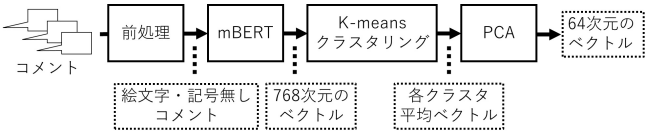


図 4: 視聴者ノードの表現ベクトル初期化手順

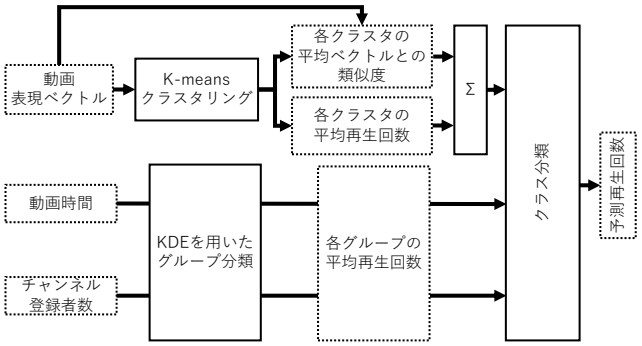


図 5: 再生回数予測の手順

4.1 データセット

140 本の動画を YouTube から収集した．その内，120 本を既存動画とし，20 本を新規動画とした．LightGCN での学習に使用したデータを既存動画とし，学習に使用していないデータを新規動画とした．

（1） 既存動画としてのデータセット

本論文で使用した YouTube 動画のメタデータ，コメントデータの収集には，YouTube Data API を使用した．学習用既存動画として 120 本（12 のチャンネルの動画各 10 本ずつ）の動画のメタデータ（動画タイトル，動画時間，チャンネル登録者数）と，その動画に対するコメントデータ（コメント原文のみ）を収集した．Shorts 動画には同一チャンネルで酷似したタイトルが多く，タイトルとしての情報が少ない．また，コメント欄にも多くの言語が混在しているため，本実験では対象外とした．子供向け動画やコメント欄が封鎖されている動画においては，コメントが存在しないため，本実験では対象外とした．

（2） 新規動画としてのデータセット

評価用新規動画として，既存動画とは異なる動画 20 本に対して，動画メタデータとその動画に対するコメントデータを収集した．既存動画と同様に，Shorts 動画とコメントが存在しない動画は本実験では対象外とした．

4.2 実 験 設 定

4.3 再生回数予測のハイパーパラメータ設定

本研究では，SVC を用いて再生回数予測モデルを実装している．予測を行う上でハイパーパラメータを変更して比較をする．以下に各ハイパーパラメータと設定値を示す．

- kernel：カーネル関数 （rbf, linear, sigmoid）
- C：コストパラメータ (0.001, 0.01, 0.1, 1, 10, 100, 1000)

誤分類データへのペナルティである．

- gamma：カーネル係数 (0.001, 0.01, 0.1, 1, 10, 100, 1000)

’rbf’ カーネルの際に使用され，各データポイントの影響の範囲を決める．

4.4 再生回数予測のモデル選択

再生回数予測を行う上で変更点を設け，複数のモデルを作成した．以下に変更点を示す．

（1） **表現ベクトルの比較**

本研究では表現ベクトルを獲得するため，LightGCN の学習を，VUgraph，VCgraph，UCgraph の三つに分けて行った．このうち，再生回数予測に使用する動画の表現ベクトルが得られるのは VUgraph，VCgraph の二つの場合である．再生回数予測に使用する表現ベクトルを，VUgraph の場合，VCgraph の場合，両者を足し合わせた場合の三つで比較をする．また，コメントデータの有用性を評価するため，LightGCN の学習結果を使用せず，動画タイトルを mBERT にてベクトル化したものを使用したモデルについても比較する．

（2） **説明変数の選定**

LightGCN を利用して取得した表現ベクトル（動画タイトルベクトルを使用するモデルもある）に加え，動画メタデータ（動画時間，チャンネル登録者数）の三つの説明変数それぞれに対して，使用した場合と使用しない場合のすべての組み合わせにて比較をする．

（3） **再生回数のクラス**

再生回数の予測を行う際に正解データとして使用する再生回数クラスの数である．本論文では 5 クラスに設定した．

4.5 LightGCN のハイパーパラメータ設定

LightGCN のハイパーパラメータとして，GCN のレイヤー数が存在する．本論文では実験 1，2 共通して 3 層に設定した．

4.6 新規動画のベクトル生成の設定

3.3.2 節の手法を用いて新規動画のベクトルを生成するとき，類似度とベクトルの重みづけ和を計算する対象動画の数を，類似度の高いものから順に（1，2，3，4，5，120）と設定し比較を行った．

4.7 実験 1：再生回数の予測に関する実験

4.7.1 実 験 目 的

提案手法である，動画・視聴者・コメントの三種類のノードで構築されたグラフを利用した再生回数予測手法と，グラフを使用しない再生回数予測手法を比較する．また，作成したモデルを使用して新規動画に対して再生回数の予測を行い，評価する．

本実験では，提案手法（3 説）で示した手法を用いて，再生回数の予測を行った．既存動画と新規動画に分けて行い，既存動画の再生回数の予測実験にて得られたモデルを使用し，新規動画の再生回数の予測実験を行った．

4.7.2 実験概要：既存動画の場合

既存動画の再生回数予測では，五分割交差検証を行い，各モデルの評価を行った．

120 本の動画のうち，24 本の動画をランダムに選び，テストデータとする．残りの 96 本を訓練データとし，学習を行う．その後得られたモデルをテストデータにて評価を行う．これを全ての動画が 1 度ずつテストデータとして選ばれるように 5 回繰り返して行い，平均評価をそのモデルの評価とする．

評価指標として，以下の三つを算出した．

- Precision

ある再生回数のクラスに予測されたサンプルのうち，実際にそのクラスである割合である．

- Recall

ある再生回数のクラスに属するサンプルのうち，実際にそのクラスに予測された割合である．

- F1-score

Precision と Recall の調和平均である．

4.7.3 実験結果：既存動画の場合

提案手法の各モデルでの再生回数の予測結果を示す．ハイパーパラメータの組み合わせが多いため，各モデルにおいて F1-score の高かった組み合わせを三つずつ抜粋して示す．

VUgraph から得られた動画表現ベクトルを使用したモデル（VUgraph ベース）の結果を表 3 に示し，VCgraph から得られた動画表現ベクトルを使用したモデル（VCgraph ベース）の結果を表 4 に示し．両方のベクトルを足し合わせて使用したモデル（VU+VCgraph ベース）の結果を表 5 に示す．また，LightGCN から得られた動画表現ベクトルを使用せず，動画タイトルベクトルを使用したモデル（動画タイトルベース）の結果を表 6 に示す．これらの表において，ハイパーパラメータの記述方法は表 1 に，説明変数の記述方法は表 2 に示す．

表 3〜6 の結果より，VUgraph から得られたベクトルを使用し，kernel として linear，C には 100 を使用し，説明変数として動画表現ベクトルと動画時間，チャンネル登録者数の三つすべてを使用した場合に，F1-score が 0.520 となり，一番高い値であった．よってこのモデルとその設定を新規動画の再生回数予測に使用する．

4.7.4 考　　察

表 3〜6 より，VUgraph から得られる動画表現ベクトルを使用したモデルが，他のモデルと比べて F1-score が高くなる結果となった．VUgraph が動画と視聴者のインタラクションを直接モデリングするため，他のモデルと比べてよい結果となったと考えられる．

VUgraph と VCgraph から得られるベクトルを使用したモデルの F1-score は，VUgraph ベースのモデルと VCgraph ベースのモデルの F1-score の間を取った．つまり，VUgraph と VCgraph

をただ足し合わせるだけでは再生回数の予測に効果的でなく，VUgraph から得られるベクトルだけを使用した場合の方がよい結果となることが分かった．本論文では VUgraph と VCgraph を別々に学習し，ベクトルを足し合わせて使用したが，動画，視聴者，コメントからなる三部グラフからベクトルを得ることができれば，VUgraph ベースモデルより良い結果となる可能性がある．VU+VCgraph ベースモデルにおいて，F1-score が高くなった説明変数の組み合わせ 3 組中，2 組は動画表現ベクトルとチャンネル登録者数であり，動画時間は使用していなかった．チャンネル登録者数は再生回数と正の相関を取っているが，動画時間はとっていない．そのため，二つのベクトルを足し合わせることによって動画時間の影響がベクトルにさらに追加され，再生回数予測の説明変数として必要で無くなったと推測される．逆にどのモデルにおいても，F1-score が高い場合にチャンネル登録者数が説明変数として使用されているのは，再生回数と正の相関が強いからだと推測される．

グラフベースのモデルと動画タイトルベースのモデルにおいて，F1-score に差が見られた．また，動画タイトルベースのモデルのうち，F1-score が高くなった設定では，説明変数として動画タイトルベクトルは使われていなかった．この 2 点より，動画表現ベクトルは動画タイトルベクトルと比べて，再生回数の予測に効果的であると考えられる．全てのモデルにおいて，F1-score が高くなったハイパーパラメータの組み合わせ 3 組中，3 組全てにおいて Kernel が linear であったのは VUgraph ベースモデルのみであった．これは，VUgraph の動画表現ベクトルが再生回数との関わりが強く，線形関係であると考えられる．

4.7.5 実験概要：新規動画の場合

本実験では，既存動画の実験（4.7.3 節）にて得られたモデルを使用し，新規動画 20 本に対して再生回数の予測実験を行い，評価した．再生回数の予測モデルに使用するベクトルは，3.3.2 節の方法を用いて生成した新規動画の表現ベクトルとしている．評価指標には，既存動画と同様に Precision，Recall，F1-score の三つを使用した．

4.7.6 実験結果：新規動画の場合

新規動画の再生回数の予測結果を表 7 に示す．新規動画の表現ベクトルを作成する際に使用する動画が 1 本のとき，F1-score が一番高くなり，0.350 となった．

4.7.7 考　　察

表 7 より，新規動画の表現ベクトルを作成する際に使用する動画が 1 件，つまり動画タイトルのベクトルの類似度が一番高いものをそのまま新規動画のベクトルとして使用した場合に，F1-score が一番高くなる結果となった．新規動画の表現ベクトルを作成する際，式 (7) に示すように各動画タイトルベクトルの類似度と各動画表現ベクトルの積和を取る．複数の動画ベクトルを足し合わせていく時，類似度の高い部分が異なる場合，最終的に生成した表現ベクトルが平均的なベクトルになってしまうために，F1-score が低くなるのだと考えられる．また，既存動画の予測精度に対して新規動画の予測精度が大きく劣る結果となった．これには二つ理由が考えられる．一つ目は，既存動画の数の不足である．本論文では，新規動画の動画表現ベクトルの生成時に，動画タイトルベクトルの類似度が高い動画の表現ベクトルを代わりに使用するが，本実験で使用したデータセットでは既存動画が 120 本であり，よく似た動画を見つけるには数が足りなかったと考えられる．二つ目は，新規動画表現ベクトルの生成方法の不適である．本論文では新規動画の表現ベクトルの生成にタイトルの類似度だけを使用したが，動画時間やチャンネル登録者数などを使用することで，より効果的なベクトルを生成できると考えられる．

論文では，既存動画と新規動画に対して再生回数の予測実験を行った．実際のシチュエーションとしては，既存動画は本論文での扱いと同じく公開済みの動画であり，再生回数予測モデルの学習に使用する動画である．新規動画は本論文での扱いとほぼ同じである．公開される前の動画であり，動画タイトル，動画時間，公開するチャンネルの登録者数のみ分かっているが，再生回数は分からない．本論文ではこの再生回数の予測を目的とした．このような既存動画と新規動画に対し，再生回数が落ち着く前の動画（公開済み動画とする）も再生回数の予測の対象になると考えている．公開済み動画は新規動画と異なり，高評価数とコメントデータが存在する．それらを説明変数として追加すればより高い精度で再生回数を予測できると考えられる．表 8 には，実験 1 で使用した説明変数に加えて，高評価数と一日当たりのコメント数も追加し，UVgraph モデルを使用して，既存動画再生回数予測実験を行った結果を示す．表 8 において H は高評価数を表している．また，公開されてから経過した日数を説明変数として追加することで，さらに精度を上げることができると考えられる．

表 1: ハイパーパラメータの記述方法

記述方法	Kernel	C	gamma
(kernel, C, gamma)	Rb：rbf Li：linear Si：sigmoid	0.001 0.01 0.1 1 10 100 1000	0.001 0.01 0.1 1 10 100 1000

表 2: 説明変数の記述方法

記述方法	説明変数の種類
（使用した説明変数）	V：動画表現ベクトル（動画タイトルベクトル） L：動画時間 S：チャンネル登録者数

表 3: 既存動画再生回数予測結果：VUgraph ベース

ハイパーパラメータ	説明変数	Precision	Recall	F1
Li，100，-	V，L，S	0.558	0.542	0.520
Li，10，-	V，S	0.553	0.517	0.512
Li，1000，-	V，L，S	0.615	0.508	0.511

表 4: 既存動画再生回数予測結果：VCgraph ベース

ハイパーパラメータ	説明変数	Precision	Recall	F1
Rb，100，0.1	V，L，S	0.543	0.483	0.478
Li，10，-	V，L，S	0.530	0.475	0.458
Li，1000，-	V，L，S	0.468	0.492	0.456

表 5: 既存動画再生回数予測結果：VU+VCgraph ベース

ハイパーパラメータ	説明変数	Precision	Recall	F1
Li，1000，-	V，S	0.458	0.525	0.490
Rb，1000，0.001	V，S	0.483	0.525	0.472
Rb，100，0.001	V，L，S	0.563	0.500	0.472

表 6: 既存動画再生回数予測結果：動画タイトルベース

ハイパーパラメータ	説明変数	Precision	Recall	F1
Rb，100，100	L，S	0.449	0.425	0.412
Rb，100，1000	L，S	0.459	0.417	0.405
Rb，1000，0.1	L，S	0.459	0.433	0.403

表 7: 新規動画再生回数予測結果

使用した類似動画の数（件）	Precision	Recall	F1
1	0.430	0.300	0.348
2	0.427	0.427	0.232
3	0.595	0.200	0.161
4	0.150	0.030	0.050
5	0.000	0.000	0.000
120（全件）	0.000	0.000	0.000

表 8: 既存動画再生回数予測結果：VUgraph ベース（説明変数追加）

ハイパーパラメータ	説明変数	Precision	Recall	F1
Li，10，-	V，H	0.722	0.700	0.689
Rb，1，1	V，H	0.742	0.683	0.680
Rb，1000，0.01	V，H，S	0.702	0.683	0.676
Li，1000，-	H	0.714	0.667	0.672
Rb，100，0.1	H	0.700	0.667	0.671

4.8 実験 2：視聴者層の推定に関する実験

4.8.1 実 験 目 的

LightGCN から得られる視聴者の動画に対する Rating の精度を検証する．新規動画に対する視聴者層の推定を行う場合，LightGCN の出力である動画ノードと視聴者ノードの Rating を使用する．そのとき，LightGCN の学習で得られる Rating の精度が新規動画に対する視聴者層の推定の精度に関わるためである．さらに，その結果を基に実際に視聴者層の推定を行う．

4.8.2 LightGCN の性能検証

LightGCN で学習を行っている際に出力される評価値 Recall によって，Rating の精度を評価することができる．LightGCN の学習では，いくつかのエッジをマスクし，エッジの存在を予測する．この時に，マスクされたエッジのうち，あると予測されたエッジの割合が Recall である．LightGCN にて VUgraph を入力として学習すると，Recall は 0.610 となった．

4.8.3 実験概要：新規動画の視聴者層推定

3.3.2 節の手法を使用して，視聴者層の推定を行う．本実験では定量評価は行わない．既存動画から作成した視聴者層と新規動画での実際の視聴者層を比較する手法が無いためである．コメントからユーザ ID を取得し，推定された既存動画での視聴者層に含まれるユーザ ID と新規動画のコメントのユーザ ID を比較することはできる．しかし，3.1.3 節でも述べたように，異なる動画に同一ユーザが出現することは稀であるため，この方法は評価として十分でない．

4.8.4 実験結果：新規動画の視聴者層推定

視聴者層の推定を行った結果を表 9 に示す．動画 ID は新規動画 20 本にそれぞれ付けた動画 ID である．また，Top1 から Top10 では，独自に決めた視聴者層の番号を，式 (8) で示される，Rating の予測値が高い順に示している．また，Rating が高く予測された視聴者層が含むコメントを，ワードクラウド [15] にて表示する．ワードクラウドでは，入力文章のうち，頻出する単語が大きく表示される．図 6a に Top1 の視聴者層．図 6b に Top8 の視聴者層，図 6c に Top30 の視聴者層のコメントを表示した結果を示す．

4.8.5 考　　察

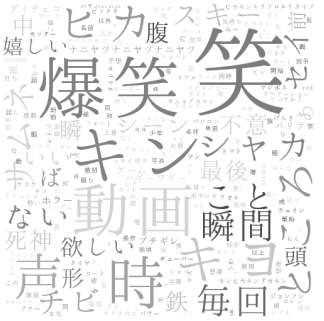
表 9 より，全新規動画で比較すると，Top1 から Top5 の視聴者層がほとんど同じであった．本論文の視聴者層の推定手法では，コメントベクトルの類似度によって視聴者をクラスタリングしている．そのため，視聴者層の中にはごく一般的で多くの動画に投稿されるようなコメントのみで構成されている視聴者層も存在する．これらの視聴者層は多くの動画に出現しているため，LihgtGCN による学習で Rating が高くなる．よって，新規動画のタイトルに関わらず Rating の予測値が高くなったと考えられる．これに対して，Top8 あたりからは視聴者層が分かれており．新規動画のタイトルに対して影響が出ていることが分かる．また，ワードクラウドの表示結果より，Top1 では「笑」や「爆笑」などといった，ありきたりな単語が大きく表示されているが，Top8 では「オワコン」，Top30 では「動画」，「企画」といったように，Top1 のコメントに比べてありきたりではない単語が大きく表示されている，そのため，Top8 あたりからの視聴者層を，新規動画に対する予測視聴者層とすることが良いと考えられる．

5 ま と め

本論文では, YouTube のコメント欄の情報に着目し, LightGCN を使用して動画-視聴者-コメントの関係を学習し，その結果を用いて再生回数の予測と視聴者層の推定を行った．既存動画を学習対象とし，既存動画と新規動画に対して視聴状況の予測に関する評価を行った．

表 9: 視聴者層の推定結果

動画	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10
0	592	6049	1709	9005	1185	8685	7498	1779	593	1186
1	592	6049	1709	9005	1185	7498	593	1779	8685	5164
2	592	6049	1709	9005	1779	8685	1185	7498	132	5355
3	592	6049	1709	9005	1779	1185	8685	7498	5355	132
4	592	6049	1709	1185	9005	8685	593	7498	1779	1186
5	592	6049	1709	9005	1779	8685	1185	7498	5355	1186
6	592	6049	1709	9005	1185	8685	1779	7498	5355	132
7	592	6049	1709	1185	9005	1779	7498	1186	8685	593
8	592	6049	1709	9005	1779	8685	1185	7498	5355	132
9	592	6049	1709	9005	1185	8685	1779	7498	5355	1186
10	592	6049	1709	9005	1779	1185	8685	7498	5355	132
11	592	6049	1709	1185	9005	7498	8685	1779	5355	1186
12	592	6049	1709	1185	9005	7498	1779	8685	1186	5355
13	592	6049	1709	1185	9005	7498	1779	8685	593	5355
14	592	6049	1709	9005	1185	8685	1779	5355	7498	5164
15	592	6049	1709	9005	1185	8685	7498	1779	5164	593
16	6049	592	1709	1185	9005	1779	7498	8685	5355	1186
17	592	6049	1709	1185	9005	7498	593	1186	8685	1779
18	592	6049	1709	9005	1185	1779	7498	593	8685	5164
19	592	6049	1709	1185	9005	7498	593	1779	5355	8685



(a) ワードクラウドによるコメント表示結果：Top1



(b) ワードクラウドによるコメント表示結果：Top8



(c) ワードクラウドによるコメント表示結果：Top30

実験の結果，既存動画の再生回数の予測に関して，動画-視聴者グラフベースのモデルを使用した場合に，F1-score が 0.520 で，一番高くなった．よって，動画-視聴者グラフベースモデルを使用した場合の情報が，他のグラフや動画タイトルに比べて再生回数予測に効果的であることが分かった．また，同モデルを使用して新規動画の再生回数の予測を行ったが，F1-score が 0.348 となり，既存動画と比べて大きく下回る結果となった．新規動画の予測には改善する余地がある．今後，データセットの充足や新規動画の表現ベクトルの生成についてさらなる検討を行う予定である．

視聴者層推定に関しては，全新規動画において，評価値 Top5 程までは視聴者層がほぼ同じになり，動画それぞれの推定とは言えなかったが，Top8 あたりから分かれており，視聴者層の推定ができたと言える．今後，既存動画からの視聴者層の取得や，新規動画に対する視聴者層の評価手法についてさらなる検討を行う予定である．

謝　　辞

本研究の一部は科研費（23H03404）による

文　　献

[1] Rohit Shewale. Youtube statistics for 2024 (users, facts & more), 12 2023.

[2] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: Analyzing the world’s largest user generated content video system. *Institutional Knowledge の School of Computing and Information Systems*, pages 1–14, 10 2007.

[3] Jiaheng Xie, Chai, Yidong, and Xiao Liu. Unbox the Black-Box: Predict and interpret youtube viewership using deep learning. *Journal of Management Information Systems*, 40(2):541–579, 4 2023.

[4] 柴田 知親 and 伊東 栄典. 回帰分析による cgm 動画再生回数推定. *DEIM Forum*, C5:1–8, 3 2018.

[5] Youtube アナリティクスの基本, 12 2023.

[6] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. *CoRR*, abs/1602.04938:1–10, 2016.

[7] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874:1–10, 2017.

[8] Zahra Movahedi Nia and Mohammad Reza Khayyambashi. Improving content popularity prediction with k-means clustering and deep-belief networks. *Multimedia Tools and Applications*, 80(10):15745–15764, 4 2021.

[9] 堺 雄之介 and 伊東 栄典. 動画サイトにおける視聴者コメントの特徴抽出. 人工知能誌, 124(4):17–22, 11 2021.

[10] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. *CoRR*, abs/2002.02126:1–10, 2020.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *CoRR*, abs/1906.01502:1–6, 2019.

[12] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Springer, 1986.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762:1–15, 2017.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[15] wordcloudjp.com. ワードクラウド.

言語モデルによる文の流暢さを用いた大喜利生成

高橋　巧実†　　鈴木　　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †takahashi.takumi.e6@s.gifu-u.ac.jp, ††suzuki.yu.r4@f.gifu-u.ac.jp

あらまし　本研究では，LLM(Large Language Model; 大規模言語モデル) の出力結果を加工することにより，自動的に大喜利の回答を生成することを目的とする.「A みたいな B，どんなの？」という形式のお題に対する回答を生成する方法の提案を行う．本稿では，A を「属性語」，B を「基本語」と呼ぶ．教師あり学習を用いて回答の生成を行う場合，十分な回答データの収集が困難であるため面白い回答の生成が難しい．代表的な LLM の一つである ChatGPT で回答を生成したところ，面白いと感じる回答は少なかった．生成された回答は予想外かつ納得感のある回答でなかったためであると考えられる．そこで，ChatGPT で基本語の関連文を生成し，生成した文中の単語を属性語の関連語と入れ替えることで予想外な回答を生成できると仮定した．パープレキシティなどの言語モデルのスコアは，文の流暢さを表しているといえる．我々は言語モデルのスコアで回答の流暢さを計算することで，意外性を考慮できると仮定した．本研究の実験で，単語の入れ替えによって面白い回答が生成できるかどうか，言語モデルのスコアを利用して面白い回答を抽出できるかどうかを確認した.

キーワード　大喜利，言語モデル，生成モデル，LLM，単語入れ替え

1　はじめに

本研究では，大喜利の回答の自動生成を目指す．大喜利とは，与えられたお題に対して面白い回答をすることによって笑いを生み出す演芸の一つである．我々は，大喜利の回答の自動生成システムが実現することで，人々が自分の興味があるお題の回答を見ることができるようになり，お笑いに触れる機会が増えると考える.

ChatGPT[1]が登場した 2022 年 11 月以降，LLM(Large Language Model; 大規模言語モデル) に関連する論文の投稿数が急激に増加している [1]．また，ChatGPT の他にも，Bard[2]などの，Web ページ上から LLM を利用できるサービスが登場している．LLM の長所の一つに，与えられたプロンプトに従って文章を自動生成できる点が挙げられる．我々は，複数のお題に対して ChatGPT を用いて大喜利の回答を生成した．生成された回答の中で，面白いと感じる回答は少なかった．LLM を訓練する際は主に二つの段階 1) 大規模コーパスの学習 2) 人間のフィードバックによる出力の調整 に分けて進める．1) では，文法や知識を獲得している．2) では，プロンプトの指示に従い，有害な表現や不正確な情報を含まない出力をするように調整される．1) で用いるコーパスは一般的な内容の文章で構成されている．2) では，意外性のある文章を出力するような調整は行われていない．以上から，我々は LLM では大喜利の面白い回答を生成することが困難であると考える.

我々は，意外性のある回答が面白い回答であると考えた．意外性をもつ文とは，その内容の予測はできないが納得はできる文である．面白い回答には聞き手の予想を裏切り，納得もさせるという絶妙なバランス感が必要である [2] といえるためである.

我々は，LLM で生成した文をベースに単語の入れ替えと言語モデルのスコアによって大喜利の回答の自動生成を試みる．本研究では「A みたいな B，どんなの？」というお題に対する回答の自動生成を目標とする．例えば,「ギャルみたいな裁判官，どんなの？」というお題である．本稿では，A を「属性語」，B を「基本語」と呼ぶ.

提案手法では最初に，属性語の関連語と，基本語の関連文に含まれる単語を入れ替えることで，回答を生成する．本稿では，単語を入れ替えを用いた回答の生成を，単語入れ替え法と呼ぶ.

次に，言語モデルで計算できるスコアをもとに面白い回答を抽出する．我々は，属性語の関連語と，基本語の関連文に含まれるを入れ替えることで，意外性をもち，お題に沿った回答を生成することができると考えた．また，言語モデルのスコアが回答の意外性を表していると考えた.

本研究では二つの実験を行う．一つ目の実験では，言語モデルのスコアによって面白い回答の抽出が可能かどうかを確認する．単語入れ替え法で生成した回答を言語モデルのスコアによってグループに分割し，各グループの回答の面白さに有意差があるかどうかを検定する．二つ目の実験では，提案手法が既存手法と比較し面白い回答を生成できるかどうかを確認する．提案手法と既存手法の回答を，パープレキシティで抽出したときの適合率再現率曲線を比較する.

一つ目の実験の結果，言語モデルのスコアによって，面白い回答を抽出できることがわかった．また，パープレキシティが小さい回答ほど，面白い傾向にあることがわかった．二つ目の実験の結果，提案手法で生成した回答は，既存手法で生成した回答よりも，面白いとはいえなかった.

本研究の貢献は以下のとおりである.

1：https://chat.openai.com
2：https://bard.google.com

- 言語モデルのスコアによって，面白い回答が抽出できることを確認した．
- 言語モデルのスコアが比較的小さい回答が面白い傾向にあることを確認した．

本論文の構成は以下のとおりである．2 章では，本研究で用いた技術，手法について述べる．3 章では，関連研究について述べる．4 章では，大喜利の自動生成手法について述べる．5 章では，評価実験について述べる．6 章では，本研究のまとめ，今後の展望について述べる．

2 関連研究

大喜利を含めた，お笑いやユーモアを持つ文章の自動生成を行う研究がいくつか存在する．大喜利については，Web ページ上のお題と回答のデータをもとに機械学習によって回答の生成を目標とする研究が複数ある．

内海ら [3] は，会話コーパスを学習した Seq2Seq モデルに，大喜利コーパスを転移学習させることによって大喜利の回答の自動生成を行なった．会話コーパスは Twitter の投稿とリプライから作成している．大喜利コーパスは大喜利のお題と回答を投稿できるサービスを提供している Web ページから取得したデータから作成している．結果，大喜利らしい面白い回答が生成されることがあった．生成した回答の面白さの定量的評価は行なっていない．

山富ら [4] は，お題となる画像と文章の回答のデータから，いわゆる「写真で一言」形式のお題に対する回答の生成を行なっている．学習に使用するデータは大喜利投稿サービスの Web ページから取得している．取得したデータを Pix2Seq を使用して学習している．また，画像の特徴量抽出に使用する事前学習済み画像認識モデルを大喜利のお題画像に対応させるために，AutoEncoder を用いている．結果，大喜利の回答らしい文章が生成できることが確認できている．生成した回答の面白さの定量的評価は行なっていない．

また，中村ら [2] は，どのような要素が大喜利の回答の面白さに寄与するかを調べている．Web ページから収集した大喜利の回答に面白さと六つの要素に対してクラウドソーシングを用いて評価付けを行った．六つの要素の評価の組み合わせを回帰モデルで面白さの評価を予測したところ，「お題との関係性」，「わかりやすさ」，「新しさ」の三つの要素が回答の面白さに寄与していることがわかった．また，三つの要素と関係がある機械的特徴量である「お題と回答の文コサイン類似度」，「平均単語難易度」，「お題と回答のトピックモデル類似度」を用いた回帰モデルで面白さの評価の予測をした．その結果，面白さに寄与するとわかった三つの要素より予測精度が悪化した．結果より，どの機械的特徴も面白さを表しているとはいえなかった．

大喜利以外のユーモアをもつ文章の生成を目標とする研究もある．呉ら [5] [6] は，聞き間違えボケによるユーモア生成を行なっている．ユーザが入力した文の単語一つを別の単語に置き換えることで，聞き間違えボケを生成している．置き換え元の単語は，入力文のトピックに最も意味が近い単語を，概念距離によって決定している．置き換え先の単語は，入力文や置き換え元の単語との概念距離，置き換え元の単語との後の近さ，コーパス内での出現頻度によって決定している．概念距離の計算には Word2Vec を用いている．生成したボケの面白さの定量的評価を行なったところ，概念距離を用いたトピックの考慮が面白さにある程度寄与することが示された．

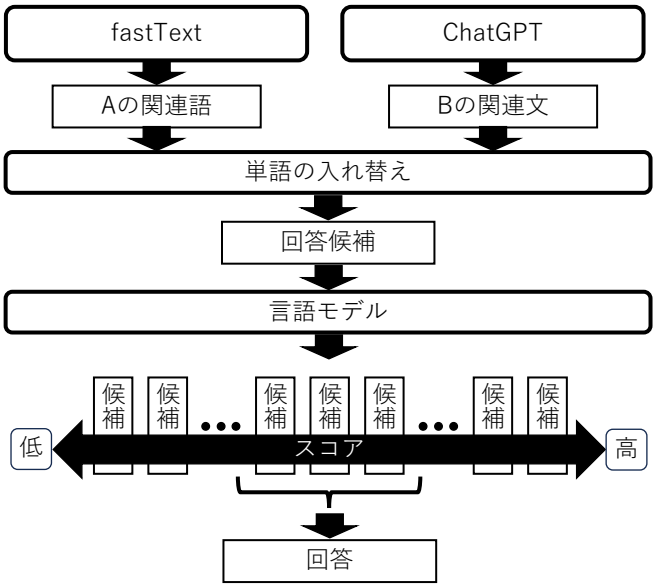


図 1　提案手法の概要図

3 提案手法

本研究では，「A みたいな B，どんなの？」という大喜利のお題に対する回答の生成を行う．我々は，単語入れ替え法と，言語モデルのスコアに着目した．提案手法では，以下の手順で回答を生成する．

Step 1　fastText を用いて属性語の関連語を抽出する．

Step 2　ChatGPT を用いて基本語の関連文を生成する．

Step 3　基本語の関連文に含まれる単語を属性語の関連語と入れ替えることによって，回答を生成する．

Step 4　言語モデルを用いて，生成した回答に対してスコアを計算する．

Step 5　計算したスコアをもとに面白い回答を抽出する．

提案手法の概要図を図 1 に示す．

我々は，大喜利の回答の生成にあたって，以下の二つの条件を満たしている文が回答として適していると考えた．

条件 1　面白みを感じる文である．

条件 2　文がお題に沿った内容である．

我々は，属性語の関連語を基本語の関連文に含まれる単語と入れ替えることによって，お題「A みたいな B, どんなの？」に沿った回答を生成できると考えた．理由は以下の二つである．

一つ目の理由は，単語を入れ替えた箇所は前後の文脈から内容を予測することが難しくなり，意外性が生まれるためである．意外性を持つ文とはその内容について予測することはできないが，納得することはできる文である．我々は，意外性をもつ文に面白みを感じると考えた．面白い回答には聞き手を予想を裏切り，納得もさせるという絶妙なバランス感が必要である [2] といえるためである．以上より，我々は，属性語の関連語を基本語の関連文に含まれる単語と入れ替えることによって，文は条件 1 を満たすと考えた．

二つ目の理由は，単語入れ替え法によって文が属性語と基本語の両方の要素を持つことになり，文は条件 2 を満たすと考えたためである．よって,「A みたいな B，どんなの？」というお題に適した回答を生成できると考えた．

単語入れ替え法で大喜利の回答を生成するとき，生成された回答の中には面白くない回答も含まれている．そのため，面白い回答のみを抽出する必要がある．我々は，言語モデルのスコアによって，面白い回答が抽出できると考えた．

我々は，意外性のある回答が面白い回答であると考えた．意外性をもつ文とは，その内容の予測はできないが納得はできる文である．面白い回答には聞き手の予想を裏切り，納得もさせるという絶妙なバランス感が必要である [2] といえるためである．

我々は，言語モデルのスコアが回答の意外性を表す指標になると考えた．言語モデルのスコアは，文の流暢さを表しているといえる．文が流暢であるほど，私たちが普段から使っている文に近いといえる．我々は，流暢である文章は内容の予測がしやすく，流暢ではない文章は内容に納得感が得られないと考えた．以上から，言語モデルのスコアが回答の意外性を表す指標になると考えたため，言語モデルのスコアをもとに面白い回答の抽出ができると考えた．

3.1 単語入れ替え法による回答の生成

単語入れ替え法によって大喜利の回答を生成する手順を説明する．最初に，属性語に関連する単語を 10 個抽出する．属性語の単語ベクトルとコサイン類似度が高い単語ベクトルを持つ単語上位 10 個を抽出する．類似度が高い単語ベクトルを持つ単語同士は，関連があると考えたためである．fastText [7] の学習によって獲得した単語ベクトルを，関連語の抽出に用いた．

抽出した単語を $w_{ai}(i = 1, \ldots, 10)$ とする．

次に，ChatGPT を用いて，基本語に関連する文を 10 個生成する．生成した基本語に関連する文を $s_j(j = 1, \ldots, 10)$ とする．ChatGPT に入力したプロンプトは以下のとおりである．

> 以下の条件を満たす文を 10 個生成してください。
> 1. 内容は「基本語」の具体的な行動である
> 2. 「基本語」に関係し、「基本語」以外には関係しない単語を 3 個以上含む。また、「B」に関係する単語はどれであるかを必ず示してください
> 3. 文の長さは 7 単語以内にする
> 4. 文は「基本語」から始まる

以上のプロンプトを ChatGPT に入力して生成した s_j に，形態素解析を行う．形態素解析器には MeCab [8] を用いた．s_j に含まれる形態素の総数を N_j，形態素それぞれを $w_{bjk}(k = 1, \ldots, N_j)$ とする．

w_{ai} と w_{bjk} の組み合わせの総当たりを列挙する．文法的に不適な文が生成されるのを防ぐため，w_{ai} と w_{bjk} が共に名詞である場合のみに制限した．また,「する」とその活用形に接続している，サ変接続である w_{bjk} とサ変接続ではない w_{ai} の組み合わせは，文法的に不適であるため組み合わせから除外する．列挙した組み合わせの s_j 中の w_{bjk} と w_{ai} を入れ替える．単語を入れ替えて得られた文を，お題「A みたいな B，どんなの？」に対する回答とする．

3.2 言語モデルのスコアに基づいた面白い回答の抽出

言語モデルのスコアに基づいた，面白い回答の抽出の手順を説明する．最初に，単語入れ替え法によって生成した回答に対するスコアを計算する．計算したスコアに基づいて，回答をソートする．面白い回答が分布すると考えられる範囲から，回答を抽出する．

本稿では，回答のスコアの計算に，パープレキシティと MLM の損失関数の二通りの方法を用いて，面白い回答の抽出を行う．我々は，パープレキシティと MLM の損失関数は，どちらも回答の意外性を表す指標と考えた．どちらのスコアも，文の流暢さを表すと考えたためである．また，パープレキシティの計算には GPT-2 [9] の，MLM の損失関数の計算には BERT [10] の事前学習済みモデルを用いる．

3.2.1 パープレキシティ

GPT-2 を用いる場合，文に対するパープレキシティを計算し，回答のスコアとする．パープレキシティは 0 より大きい指標であり，0 に近いほど文は流暢であるといえる．入力文のトークン列 W に対するパープレキシティ $perplexity$ の計算式を式 1 に示す．

$$perplexity = \exp\left(-\frac{1}{M}\sum_{i=0}^{M}\log\left(p_{GPT-2}\left(w_i|W_{i-1}\right)\right)\right) \quad (1)$$

W にはトークンが M 個含まれている．入力文の i 番目のトークンを w_i，入力文の $i-1$ 番目までのトークン列を W_{i-1} とする．$p_{GPT-2}(w_i|W_{i-1})$ は，モデルに W_{i-1} を与えたときに，モデルが次に出現するトークンを w_i であると予測する確率である．内容が予測しやすい文では全ての i に対して $p_{GPT-2}(w_i|W_{i-1})$ が大きくなるため，$perplexity$ が小さくなる．内容に納得が得られない文は，W_{i-1} から予測しづらい w_i が存在するため，$p_{GPT-2}(w_i|W_{i-1})$ が小さくなる i が存在する．そのため，内容に納得が得られない文は $perplexity$ が大きくなる．予測のしづらいトークンの有無で変動する $perplexity$ は文の流暢さを表しているといえる．そのため，我々はパープレキシティで回答の意外性を考慮できると考えた．

3.2.2 MLM の損失関数

BERT を用いる場合，置き換えた単語を [MASK] にしたときの MLM(Masked Language Model) の損失関数を計算し，回

答のスコアとする．MLM の損失関数の値は 0 以上であり，0 に近いほど，文は流暢であるといえる．入力文のトークン列 W に対する MLM の損失関数 $MLMloss$ の計算式を式 2 に示す．

$$MLMloss = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{BERT}(mask_i|W)) \quad (2)$$

W の中には [MASK] に置換したトークンが N 個含まれている．i 番目の [MASK] に置換されたトークンを $mask_i$ とする．$p_{BERT}(mask_i|W)$ は，モデルが W をもとに入力文の i 番目の [MASK] が $mask_i$ であると予測する確率である．内容が予測しやすい文では [MASK] されたトークンを前後の文脈から予測しやすいため，全ての i に対して $p_{BERT}(mask_i|W)$ が大きくなる．そのため $MLMloss$ は小さくなる．内容に納得感が得えられない文では，前後の文脈から予測が困難なトークンが存在すると考えられる．そのため，予測が困難なトークンをマスクトークンに置き換えたとき，$p_{BERT}(mask_i|W)$ が小さくなり，$MLMloss$ は大きくなる．よって予測のしづらいトークンの有無で変動する $MLMloss$ は文の流暢さを表しているといえる．以上から，我々は MLM の損失関数で回答の意外性を考慮できると考えた．

4 実　　験

本稿では二つの実験を行う．一つ目の実験では，言語モデルによって面白い回答の抽出が可能かどうかを確認する．二つ目の実験では，単語入れ替え法で生成した回答と既存手法で生成した回答の面白さに差があるかどうかを確認する．本稿では，既存手法に LLM を使用する．LLM には ChatGPT(GPT-3.5) を選択した．

4.1 実験 1: 言語モデルのスコアによる比較

4.1.1 実 験 手 順

実験 1 の手順は以下の通りである．

Step 1　「A みたいな B，どんなの？」の形式のお題を用意する．

Step 2　Step1 で用意したお題の回答を 3.1 節で示した単語入れ替え法によって生成する．

Step 3　Step2 で生成した回答の面白さの評価をアンケートによって行う．

Step 4　Step2 で生成した回答のスコアを言語モデルで計算する．

Step 5　Step4 で計算したスコアをもとに回答を昇順でソートし．上位から 10%ごとのグループに分割する．分割したグループを $G_1, G_2, \ldots G_{10}$ とする．

Step 6　$G_1, G_2, \ldots G_{10}$ の回答がそのグループに含まれない回答と比較して有意に面白いかどうかを検定する．

お題は，YouTube 上の大喜利動画を投稿しているチャンネル [3]の動画で用いられているお題を参考に，我々が属性語と基本語をそれぞれ三つずつ決定し，その組み合わせの九つを用意した．実験に用いたお題を表 1 に示す．Step 2 で用いる fastText にはインターネット上で公開されているモデル [4]を使用した．Step 4 では二種類の方法を用いて回答ごとにスコアを計算する．用いた方法は，パープレキシティと MLM の損失関数である．

パープレキシティの計算に用いる GPT-2 には，rinna 株式会社が公開している事前学習済みモデル [5]を使用した．MLM の損失関数の計算に用いる BERT には，東北大学の乾・鈴木研究室が公開している事前学習済みモデル [6]を使用した．

協力者にアンケートを行い，回答の面白さの評価を行った．協力者に回答の面白さを，1. 全く面白くない，2. あまり面白くない，3. どちらともいえない，4. まあ面白い，5. とても面白いの 5 段階で評価してもらった．回答を 4，5 のいずれかと評価した人数を，回答の面白さとした．協力者は，岐阜大学の鈴木研究室の学生 13 名である．Step6 では，言語モデルのスコアによって面白い回答の抽出が可能であるかどうかを確認する．グループの回答のが有意に面白いかどうかを確認する．あるグループに含まれる回答の面白さの平均値が，そのグループに含まれない回答の面白さの平均値より有意に大きいか否かを対応のない 2 標本 t 検定によって確認する．今回は比較する二群に等分散性を仮定した．有意水準を 0.05 に設定し，片側検定を行った．複数のお題で，回答が有意に面白いグループが存在する場合，言語モデルのスコアによって面白い回答の抽出をすることが可能であるといえる．

4.1.2 結果・考察

実験の結果を図 2 と図 3 に示す．図 2 で示す実験ではパープレキシティを，図 3 で示す実験では MLM の損失関数をもとに回答を 10 個のグループに分割した．横軸は各グループであり，スコアが昇順に左から並んでいる．縦軸は，各グループの回答の面白さの算術平均である．棒グラフでは，回答が有意に面白かったグループを赤色で表示している．

パープレキシティによるグループ化では 9 個中 8 個のお題で，面白さが他と比較し有意に大きいグループが確認できた．お題

表 1　実験に用いたお題

お題番号	お題
1	オタクみたいなラッパー、どんなの？
2	オタクみたいな裁判官、どんなの？
3	オタクみたいな詐欺師、どんなの？
4	ギャルみたいなラッパー、どんなの？
5	ギャルみたいな裁判官、どんなの？
6	ギャルみたいな詐欺師、どんなの？
7	猫みたいなラッパー、どんなの？
8	猫みたいな裁判官、どんなの？
9	猫みたいな詐欺師、どんなの？

3：https://www.youtube.com/@oogiruhitotachi
4：https://drive.google.com/file/d/0ByFQ96A4DgSPUm9wVWRLdm5qbmc/view?usp=sharing&resourcekey=0-of5Ks1fuoKNh1pEYE8uSFQ
5：https://huggingface.co/rinna/japanese-gpt-1b
6：https://github.com/cl-tohoku/bert-japanese

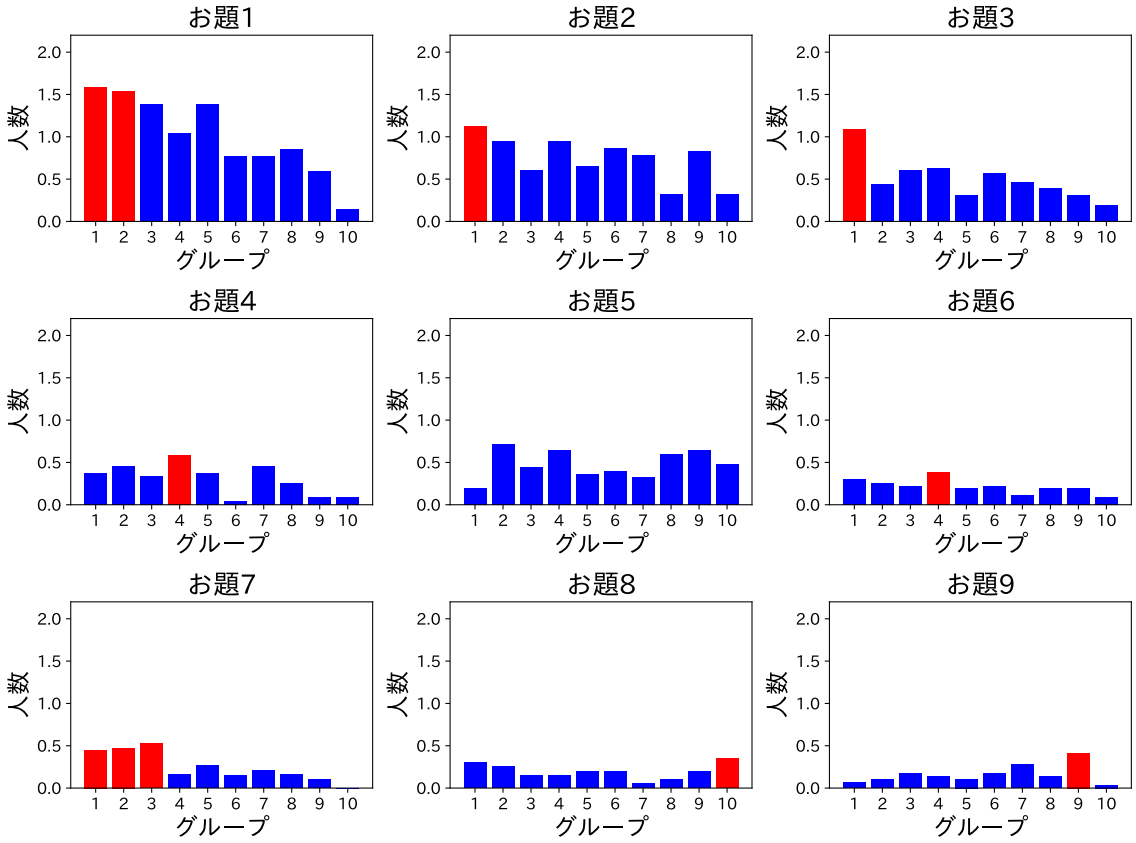


図 2　実験 1 の結果: パープレキシティでグループ化したときの面白さの平均

1,2,3,7 では，パープレキシティが小さいグループが，お題 4,6 では，パープレキシティが中程度のグループが，お題 8,9 では，パープレキシティが大きいグループが有意に面白いことが確認できた．

MLM の損失関数によるグループ化では，9 個中 7 個のお題で，面白さが他と比較し有意に大きいグループが確認できた．お題 1,2,3,7,8 では，MLM の損失関数の値が小さいグループが，お題 4 では，MLM の損失関数が比較的大きいグループが有意に面白いことが確認できた．お題 9 では，MLM の損失関数が上位 40%から 50%のグループと 90%から 100%のグループが有意に面白いことが確認できた．

お題 1,2,3,7 ではパープレキシティと MLM の損失関数で，回答が有意に面白いグループの傾向が同じであった．お題 5 では，パープレキシティと MLM の損失関数の両方で，回答が有意に面白いグループを確認できなかった．お題 4,6,9 では，回答が有意に面白いグループの傾向がパープレキシティと MLM の損失関数で異なっている．お題 4 では，パープレキシティで分割したときの回答が面白いグループは上位 40%から 50%であるが，MLM の損失関数で分割したときは 80%から 90%のグループとなっている．お題 6 では，パープレキシティで分割したときは回答が有意に面白いグループが確認できたが，MLM の損失関数で分割したときでは確認できなくなっている．お題 9 では，MLM の損失関数で分割したとき，上位 40%から 50%のグループの回答が有意に面白いが，パープレキシティで分割したときは有意に面白くはなかった．

パープレキシティでは 9 個中 8 個のお題で，MLM の損失関数では 9 個中 7 個のお題で，回答が有意に面白いグループが存在した．よって，パープレキシティと MLM の損失関数をもとに，面白い回答の抽出が可能であるといえる．また，パープレキシティでは 9 個中 4 個のお題で，MLM の損失関数では 9 個中 5 個のお題で，言語モデルのスコアが比較的小さいグループの回答が有意に面白かった．そのため，パープレキシティと MLM の損失関数が小さい回答ほど，面白い傾向にあるといえる．

言語モデルのスコアが小さいほど，文章は流暢であるといえる．流暢さが高い文章は内容の予測がしやすく，流暢さが低い文章は納得感が得られない．そのため，言語モデルのスコアが小さい回答は，内容がわかりやすい回答であるといえる．結果より，言語モデルのスコアが小さい回答が面白いと評価されやすい傾向があることがわかった．よって，単語入れ替え法で生成した回答は，わかりやすい回答が面白いと評価されやすい傾向にあったといえる．

4.2　実験 2: 既存手法との比較

4.2.1　実験手順

実験 2 の手順は以下の通りである．

Step 1　4.1 節で生成したお題のうち，パープレキシティが最も小さい回答 10 個を提案手法の回答とする．提案手法の回答のグループを G_m とする．

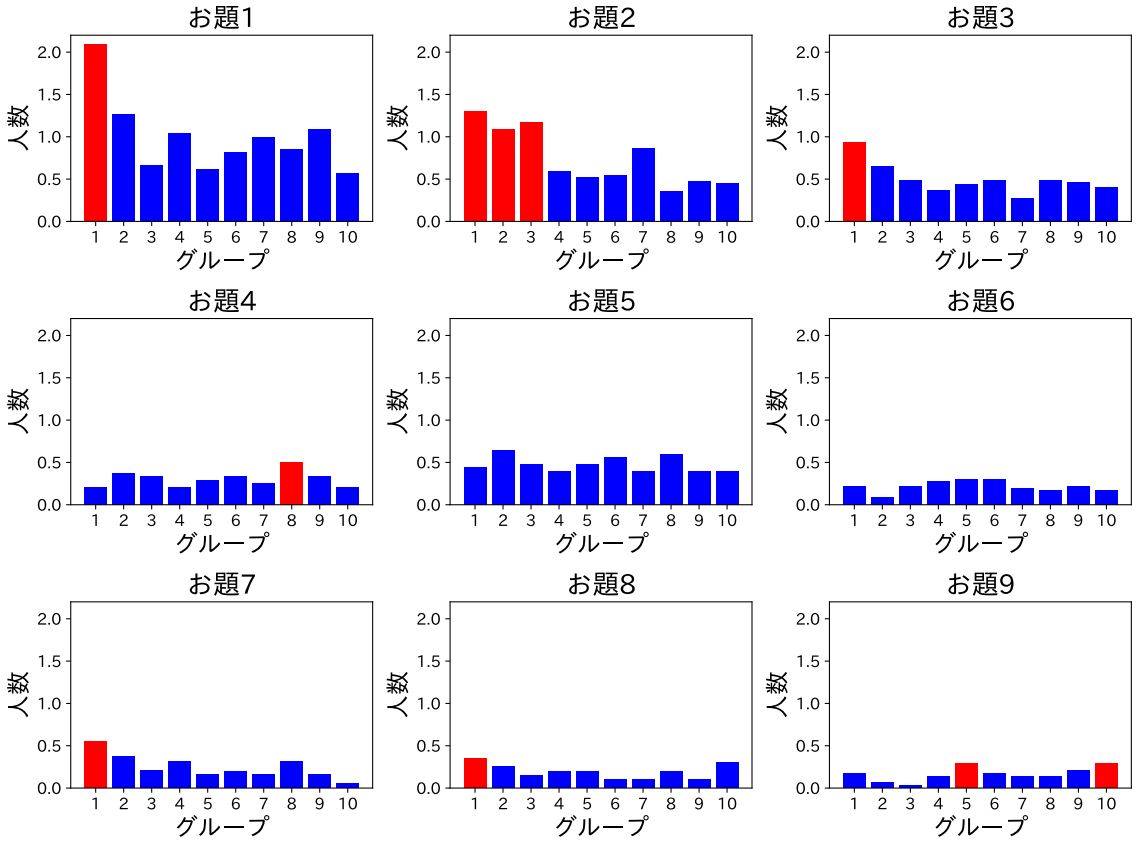


図 3　実験 1 の結果: MLM の損失関数でグループ化したときの面白さの平均

Step 2　ChatGPT にプロンプトを入力し，お題に対する回答を 10 個生成する．ChatGPT によって生成した回答のグループを G_g とする．

Step 3　G_g の回答に対して，パープレキシティを計算する．

Step 4　G_g の回答の面白さの評価をアンケートによって行う．

Step 5　G_m，G_g それぞれを，パープレキシティの昇順でソートする．G_m，G_g から，パープレキシティの小さい順に回答を抽出し，それぞれを提案手法，既存手法の出力とする．抽出する回答の件数を 1 から 10 に変化させたときの，出力に含まれている面白い回答の件数をもとに，適合率と再現率を計算する．計算した適合率と再現率をもとに適合率再現率曲線をプロットする．

既存手法に，ChatGPT(GPT-3.5) を選択した．Step2 では，以下のプロンプトを使用して回答の生成を行なった．

> 以下の文章は大喜利のお題です．回答を 10 個生成してください．回答は一文にしてください．
> お題:「A みたいな B、どんなの？」

提案手法で生成する回答の長さは一文である．提案手法と回答の長さを合わせるために，プロンプトに「回答は一文にしてください.」と記した．Step5 における，回答が面白きかそうでないかは，アンケートで回答の面白さが 4 または 5 であるとした人が 1 人以上いるか否かで決定した．我々は，誰か一人でも面白いと感じたなら，それは面白い回答であると考えたため，このような基準にした．

この実験では，提案手法と既存手法の適合率再現率曲線を比較することで，それぞれの手法の面白さの差を確認する．再現率に対する適合率の減少量が小さい手法は，パープレキシティに基づいて回答を抽出した際に，面白い回答が占める割合が大きい可能性が高いといえるためである．

4.2.2　結果・考察

実験の結果を図 4 に示す．横軸は再現率，縦軸は適合率である．青色の曲線が提案手法の，赤色の曲線が既存手法の適合率再現率曲線である．

全てのお題にて，再現率に対する適合率の減少量は，提案手法が既存手法よりも大きかった．よって，パープレキシティをもとに回答の抽出を行う場合，抽出した回答の中で面白いものが占める割合について，提案手法が既存手法よりも低いといえる．そのため，提案手法で生成した回答は，既存手法で生成した回答よりも面白いとはいえないという結果となった．

我々は，fastText による単語ベクトルの類似度では，お題に十分に沿った単語を抽出できないことが，このような結果になった原因の一つであると考えた．fastText ではお題に十分に沿った単語を抽出できないと考えた理由は，二つある．一つ目の理由は，fastText では，複数の形態素からなるフレーズを抽出できないためである．以下に示す例 1，例 2，例 3 は，お題 1「オタクみたいなラッパー、どんなの？」において，提案手法と既存手法で生成した回答の一部である．例 1 は，提案手

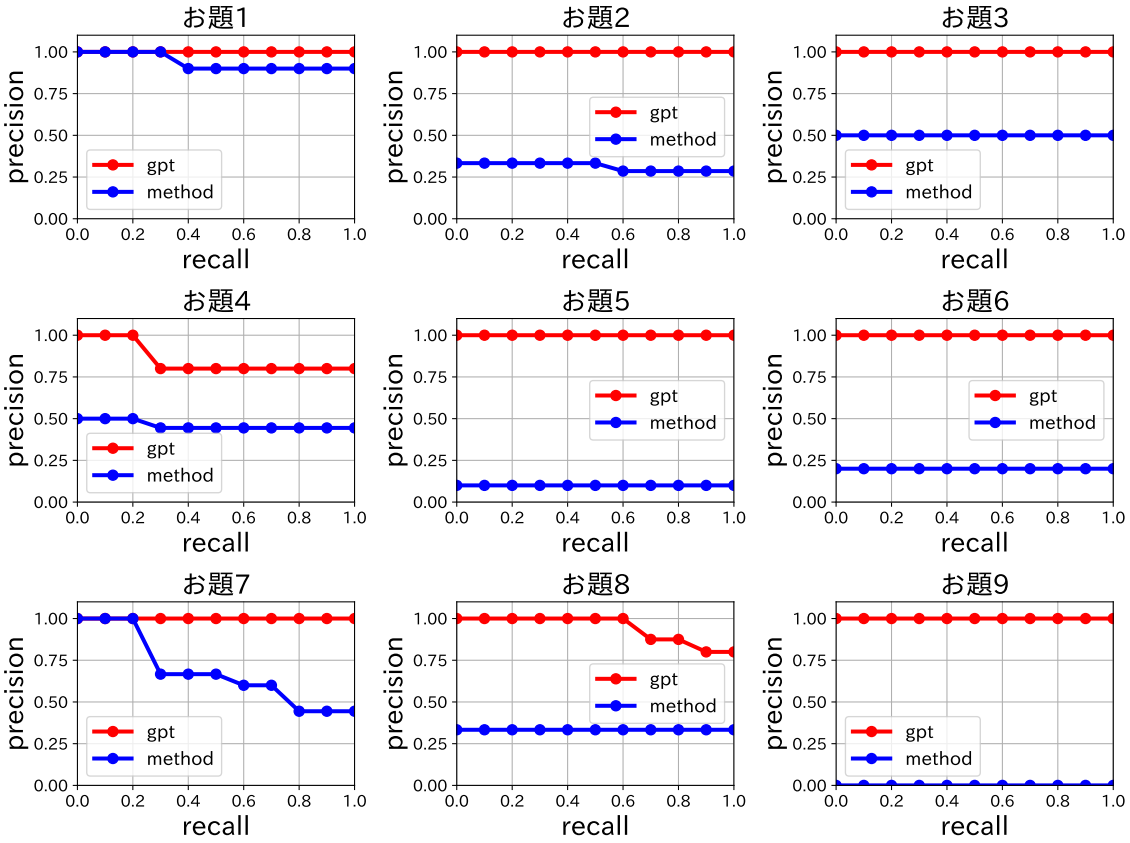


図 4　実験 2 の結果: 提案手法と既存手法の適合率再現率曲線

法で生成した回答である．例 2 と例 3 は，既存手法で生成した回答である．

例 1　ステージでロリコンを炸裂させた。

例 2　ラップの合間に「これ、知ってる？」と何度もアニメの名言を挟むラッパー。

例 3　ブレイクダンスの最中に急にゲームの攻略法を語りだすラッパー。

回答に対する面白さの評価の平均値は，例 1 は 3.00 であり，例 2 は 3.54，例 3 は 3.23 である．提案手法で生成した回答よりも，既存手法で生成した回答が面白さの評価が高かった．例 1 は，提案手法で生成した回答である．「ラッパー」の関連文中の単語と「オタク」の関連語と入れ替えることで，お題に沿った回答を生成した．例 2 と例 3 は，いずれも「ラップ」「ブレイクダンス」という「ラッパー」要素を含む行動を表す文がベースとなっている．また，「アニメの名言」「ゲームの攻略法」という「オタク」要素を含むフレーズが回答に含まれている．よって，「ラッパー」の関連文と「オタク」の関連語を組み合わせることで，お題に沿った回答を生成したといえる．このことより，例 1，例 2，例 3 は類似した方法でお題に沿った回答を生成しているといえる．以上より，提案手法と既存手法で，類似した方法でお題に沿った回答を生成していることがわかる．ところが，提案手法と既存手法で，生成した回答の面白さに差がある．我々は，提案手法では面白さ寄与する単語を回答に入れ込むことが難しいため，提案手法と既存手法の面白さに差が生まれたと考えた．既存手法の回答に含まれるフレーズ「アニメの名言」「ゲームの攻略法」はいずれも二つ以上の形態素で構成されている．我々は，この二つのフレーズは「オタク」の要素を十分に含んだフレーズであると，広く理解されると考える．それと比較し，提案手法では一つの形態素でしか，「オタク」の要素を回答に入れ込むことができない．

fastText の学習に使用するデータは，形態素解析器で分かち書きをしたコーパスである．fastText で獲得した単語ベクトルの類似度を用いて抽出した単語は，一つの形態素からなるためである．提案手法では「オタク」の関連語として，「アニメ」や「ゲーム」のような単語が抽出可能である．しかし，「名言」や「攻略法」と合わせて抽出することはできない．提案手法では複数の形態素からなる有効なフレーズを入れ込んだ回答が生成できないため，既存手法より面白くなかったと考えた．

我々は，この問題の解決するために，言語モデルの使用した複数の形態素からなるフレーズが有効であると考えた．fastText で抽出した単語を，GPT をはじめとする文章生成が可能な言語モデルに入力する．入力した単語に続く文章が出力されるので，出力の一部を抽出することでお題の要素を含むフレーズが得られると考えられる．

我々が，fastText ではお題に十分に沿った単語が抽出できないと考えた二つ目の理由は，fastText では単語どうしの上位，下位関係を考慮できないためである．以下に示す例 4 と例 5 は，お題 1「ギャルみたいな裁判官、どんなの？」において，提案手

法が生成した，パープレキシティでソートしたときの上位 10% の回答の一部である．

例 4　ストリートファッションの解釈を明確に示す。

例 5　ファッションの解釈を明確に示す。

例 4 の回答は面白いと，例 4 の回答は面白くないと判断された．この二つはどちらも,「法律の解釈を明確に示す。」という共通の文から生成された回答である．もとにした文が同じであるため,「ストリートファッション」が「ファション」よりも面白さに寄与したといえる.「ストリートファッション」は「ファッション」の下位語であり,「ストリートファッション」は「ファッション」の中の特定の範囲を意味する単語である.「ファッション」から「ストリートファッション」に，意味する範囲を狭くすることで，より「ギャル」と関連する単語となると考えられる．

fastText で獲得した単語ベクトルは，単語間の上位，下位関係を考慮することができない．単語ベクトルの類似度で関連語を抽出したとき,「ギャル」と類似する「ファッション」と「ストリートファッション」が得られる．しかし,「ファッション」は「ストリートファッション」と比較して面白さに寄与しづらい単語であるため,「ファッション」を用いた面白くない回答が生成されてしまう．

我々は，この問題を解決するために，シソーラスが有効であると考えた．シソーラスでは「ファッション」という単語から，下位語である「ストリートファッション」が抽出できるといえる．fastText で我々は，抽出した関連語の中に上位，下位関係がある二単語が含まれる場合，上位の単語を用いて回答の生成をしないことで，面白い回答が全体を占める割合が大きくなると考えた．

5 おわりに

我々は本研究で，大喜利の回答の自動生成を試みた．本研究では，お題を「A みたいな B，どんなの？」という形式のお題に絞った．我々は，単語の入れ替えと言語モデルによる大喜利の回答の自動生成手法を提案した．提案手法は，大きく分けて二つの手順で回答の自動生成を行う．単語入れ替え法による大喜利の回答生成と，言語モデルのスコアに基づいた面白い回答の抽出である．提案手法の有効性を確認するために，二つの実験を行った．一つ目の実験では，言語モデルのスコアで面白い回答が抽出可能かどうかを確認した．単語入れ替え法によって生成した回答を 10 個のグループに分割し，有意に面白いグループが存在するかどうかを調べる．これを 9 個のお題それぞれに対して行った．実験の結果，ほとんどのお題で，有意に面白いグループが存在した．結果から，言語モデルのスコアによって，面白いお題の抽出ができるといえた．また，言語モデルのスコアが小さいグループが，回答が面白い傾向があることがわかった．二つ目の実験では，提案手法で生成した回答と既存手法で生成した回答の面白さの比較を行った，提案手法と既存手法の回答をパープレキシティをもとに抽出したときの，適合率再現率曲線を比較する．これを 9 個のお題それぞれに対して行った．実験の結果，全てのお題で，提案手法の適合率再現率曲線は，既存手法の適合率再現率曲線を下回った．よって，結果から提案手法で生成した回答の面白さは，既存手法で生成した回答より面白いとはいえなかった．提案手法と既存手法の回答を比較したところ，提案手法には二つの問題点が存在することがわかった．また，これらの問題点の解決策を考察した．

今後の展望を述べる．最初に提案手法に存在する問題点の解決を考えている．本稿の実験結果より考察した解決策の有効性を確認する．また,「A みたいな B，どんなの？」以外のお題に対する回答の生成の提案を考えている．

謝辞　本研究の一部は，JSPS 科学研究費 (19H04218, 23H03694)，越山科学技術振興財団，およびソニー株式会社の助成を受けたものです．

文　献

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[2] 中川裕貴, 村脇有吾, 河原大輔, 黒橋禎夫. クラウドソーシングによる大喜利の面白さの構成要素の分析. 言語処理学会第 25 回年次大会 (NLP2019), 2019. B3-2.

[3] 内海地竹, 孟春謝, 崇喜中嶋, 徹森. 転移学習を用いた大喜利回答システムの構築. 2022 年度 情報処理学会関西支部 支部大会 講演論文集, 第 2022 巻, sep 2022.

[4] 龍山富, マハブービシェヘラザード, 洋二宮. 大喜利生成 AI の実装に関する研究. 第 85 回全国大会講演論文集, Vol. 2023, No. 1, pp. 759–760, 02 2023.

[5] 呉健朗, 中原涼太, 長岡大二, 中辻真, 宮田章裕. ボケて返す対話型エージェント. 日本バーチャルリアリティ学会論文誌, Vol. 23, No. 4, pp. 231–238, 2018.

[6] 健朗呉, 大二長岡, 涼太中原, 章裕宮田. 文のトピックを考慮した単語置換によるユーモア発話を行う対話型エージェント. 情報処理学会論文誌, Vol. 61, No. 1, pp. 113–122, jan 2020.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年3月1日(金) 13:30 ~ 15:40 | T1-A オンライン（Zoom Events)

ソーシャルデータ分析

座長:大島 裕明(兵庫県立大学)
コメンテータ:天方 大地(大阪大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-A-8-01] 機械学習と地域度･多様性分析を用いたニュース記事の自動分類

藤崎 日和[1]、*田中 克己[1] (1. 福知山公立大学情報学部)

14:00 ~ 14:25
[T1-A-8-02] 観光地レビュー分析に基づくお出かけスポット評価項目別自動スコアリング

*福嶋 亮太朗[1]、張 建偉[2]、奥 健太[3]、中島 伸介[1] (1. 京都産業大学、2. 岩手大学、3. 龍谷大学)

14:25 ~ 14:50
[T1-A-8-03] ソーシャルコンテクストを利用したアンサンブル学習によるフェイクニュースの早期検出

*谷 聡馬[1]、張 建偉[1]、佐々木 裕多[2] (1. 岩手大学大学院、2. 東京工業大学 工学院)

14:50 ~ 15:15
[T1-A-8-04] GPTを活用した、ユーザの発言・行動履歴を参考に抽出したイベント推薦の評価

*大本 詩織[1]、小口 正人[1]、榎 美紀[2] (1. お茶の水女子大学理学部情報科学科小口研究室、2. 日本アイ・ビー・エム)

15:15 ~ 15:40
[T1-A-8-05] 効果的な行動変容メッセージ実現に向けたインセンティブ効果の被験者評価

*瀧本 祥章[1]、金 秀明[1]、小林 のぞみ[1]、倉島 健[1] (1. NTT人間情報研究所)

機械学習と地域度･多様性分析を用いたニュース記事の自動分類

藤崎 日和† 田中 克己†

†福知山公立大学情報学部 〒602-0886 京都府福知山市字堀 3370

E-mail: †{32045083, tanaka-katsumi}@fukuchiyama.ac.jp

あらまし ニュース記事の自動分類を，(1) BERT 機械学習，(2) ニュースの地域度分析に基づく分類，(3) ニュース内容の多様性分析に基づく分類という 3 手法で行ったので報告する．単なる BERT 機械学習による分類では，同一ニュース記事を「国内」,「地域」いずれに分類すべきかが不明確であり，このために，ニュース記事内容からそのニュースの地域度を計算する手法を提案する．さらに，同一の事件報道でも報道機関によって大きくその内容やスタイルが異なっているようなニュースがあれば，ニュース内容の多様性が担保できるため，ニュース内容の多様性計算手法も提案している．

キーワード ニュース 自動分類 地域度 多様性 機械学習

1. まえがき

近年，ニュース記事の閲覧視聴環境が急速に変化しつつある．初期のニュース記事の閲覧視聴は，マスメディアとしての紙媒体の新聞（全国紙/地方紙）やテレビ（キー局/ローカル局）によるものが主であった．その後，インターネットの発展に伴い，ニュース記事の閲覧視聴は，一部の新聞の有料デジタル購読（NewYork Times），無料ニュースサイト（Yahoo!News など）での閲覧，SNS によるニュース閲覧（Facebook ユーザにローカルニュースを提供する Today In）などにシフトしつつある．

本論文では，ニュース記事の自動分類を，

(1) BERT 機械学習による自動分類，

(2) ニュースの地域度分析に基づく分類，

(3) ニュースの多様性分析に基づく分類

で行ったので報告する．

(1)の BERT 機械学習による分類では，同一ニュース記事を「国内」,「地域」いずれに分類すべきかが不明確という問題があったため，(2) のニュース記事内容からそのニュースの地域度を計算し，「国内」,「地域」のいずれかに再分類している．さらに，同一ジャンルのニュース記事でも報道機関によってその内容やスタイルが異なっているようなニュース記事があることを考慮して，(3)のニュースの多様性分析による再分類手法も提案する．

本論文の構成は以下のようになっている．第 2 節では，基本的事項や関連研究を紹介し，本研究の位置づけを示す．第 3 節では，ニュース記事の自動分類を BERT 機械学習で行ったので，その仕組みや分類精度などについて述べる．第 4 節では，本論文で提案するニュース記事の「地域度」の定義，計算方法，分類実験の結果を示す．第 5 節では，ニュース記事の「多様度」を計算する手法やその応用について述べる．第 6 節では，まとめと今後の課題について述べる．

2. 基本的事項および関連研究

2.1. 既存メディアにおけるニュース閲覧

米国における新聞メディアの変化は，主に，有料デジタル購読の増加（NY タイムズが有料電子会員 525 万人獲得）と地方紙の激減（2004 年に約 9,000 紙あった地方紙が 2019 年には約 6,700 紙に激減）である．一方，我が国のニュース記事の閲覧視聴環境も同様な変化を見せている．令和 4 年度 情報通信メディアの利用時間と情報行動に関する調査[1]や，NHK 国民生活時間調査 2020 [2]によれば，新聞閲読自体が激減している（新聞利用者率は 10 代で 1.1 %，20 代でも 2.6 %にとどまるが，60 代では 55.1 %と半数を超える），新聞の有料デジタル購読も我が国では進展が見られず，新規に有料の電子新聞を購読したいと考えている人は 1 割足らずであり，紙の新聞とのセットでも購読したくない人が過半数との報告がある．テレビ視聴においても，視聴者自身の減少（2015 年が 85 %から 2020 年は 79 %に減少，年層別では，20 代以下では 20 ポイント前後の減少）が顕著であり，若年層のテレビ離れが進んでいる．

2.2. ニュースサイトにおけるニュース閲覧

無料ニュースサイト（Yahoo!News など）での閲覧や，SNS によるニュース閲覧（Facebook ユーザにローカルニュースを提供する Today In など）も，長所と欠点が指摘されている．両者に共通する長所は，複数の報道機関が制作したニュース記事をジャンル毎に閲覧できる点であり，複数の新聞社やテレビ局が配信しているニュースを一括して閲覧できるメディアが登場した点は評価に値するものと考えられる．一方，欠点は，ニ

ュースサイトに掲載されるニュース記事の選択分類アルゴリズム自身である．例えば，Yahoo!JAPAN ニュースでは，人手（25 名程度の人間）が毎日約 4,000 本の記事の中から記事を選択・分類を行っている．Today In では，記事の選択アルゴリズムを用いて，記事の自動分類・選択を行っているが，以下のような欠点が明らかになっている[3].

- 記事の時間的遅れ
  ニュースフィードされる記事の多くが 2 日以上前のものであったこと
- 記事のジャンルに偏り
  犯罪や人間的興味をそそる記事が多く投稿され，教育，健康，政治，インフラ，経済開発，環境などの重要トピックをカバーする全体像を示す記事は非常に少なかったこと．
- 主要ニュースに隔たり
  新聞のトップページに掲載された主要なニュースと，「Today In」に掲載されたものとの間には隔たりがあった．ノースカロライナ州立大が分析した 2019 年 2 月の地元紙のトップページの記事のほぼ半分は，政治，教育，または健康の問題を扱っていたが，「Today In」におけるそれらの記事の割合は 13 %であった．

2.3. BERT

BERT （ Bidirectional Encoder Representation from Transformers，双方向エンコーダ表現）は Google の研究者によって 2018 年に提案された自然言語処理モデルであり，深層学習によってテキストを機械学習し，与えられた文の分析や新たな文を生成する仕組みである [9],[10],[11]. BERT の中心的部分である Transformer の全体構成を図 3 に示す．図 3 左側がエンコーダーであり，右側がデコーダーである．エンコーダーの入力（Inputs）は単語列であり，単語毎の埋め込み表現（Input Embedding）を順次生成し，さらに各単語の位置情報（Positional Encoding）を加えて，入力文の埋め込み表現を生成する．その後，この入力文の埋め込み表現を Attention 部(Multi-Head Attention)に手渡して，単語の意味が類似している可能性がある単語集合を求める．最終的に，この入力文の埋め込み表現は，デコーダーに手渡される．図 3 右側がデコーダーである．デコーダーでは，出力候補の単語列（Outputs）が入力され，各単語の埋め込み表現（Output Embedding）を生成し位置情報を加えたものが，Attention 部に手渡される．最終的に出力文が確率（Output Probabilities）と共に出力される．エンコーダー部からの出力と入力された単語列の埋め込み表現を用いて，次の単語を予測する．図 3 の両側にある Attention とは，直観的には，「文章中のどの単語に注目すればいいかを表すスコア」のことである．この Attention 機能は，図 3 の左側（入力側）にも右側（出力側）にも導入されており，Transformer で最も特徴的な部分がこの Attention であるといわれている．Attention 機能の導入により，従来の再帰型ニューラルネットワークモデル（RNN）で必要であった中間層からの再帰が不要になり，高速処理が可能になった．BERT では中間層からの再帰を用いず，その代わりに，Attention 機能が導入された訳である．

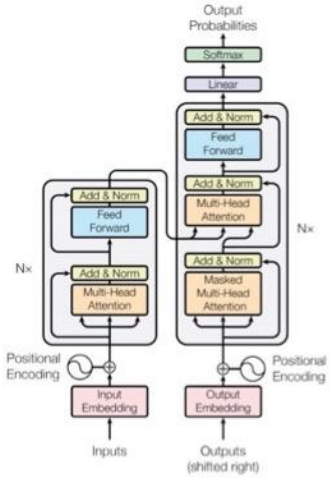


図 1 BERT の TRansformer

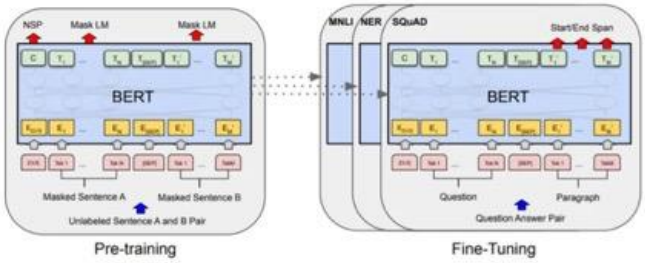


図 2 BERT の事前学習（Pre-training）と
ファインチューニング（Fine-Tuning）

BERT の事前学習およびファインチューニングの仕組みを図 2 に示す. 図 2 の BERT の事前学習(Pre-Training)は，2 つの文が文として繋がっているかどうかを学習するものであり，図 2 の BERT のファインチューニングは，事前学習のあとに行われ，質問と応答文のペアを学習するものである. ファインチューニングは, 様々なデータを追加で学習するものであり，学習の結果，すでに学習済みのモデルを修正すると共に訓練データの学習により新たに BERT 層が追加されると言う点が特徴的である．例えば，BERT によってニュースの分類を行う場合は，BertForSequenceClassification によりテキスト分類を行うことができる．日本語データを事前学習した BERT モデルも東北大学で開発されている(https://github.com/cl-tohoku/bert-japanese). その際，num_labels パラメータの値を 9 と設定すると，9 クラ

ス分類が行える．

```
import requests
from bs4 import BeautifulSoup
import re

URL = "https://news.yahoo.co.jp/topics/science"
res = requests.get(URL)


soup = BeautifulSoup(res.text, "html.parser")

data_list = soup.find_all(href=re.compile("news.yahoo.co.jp/pickup"))

headline_link_list = [data.attrs["href"] for data in data_list]

news_list = []

for headline_link in headline_link_list:

    summary = requests.get(headline_link)

    summary_soup = BeautifulSoup(summary.text, "html.parser")

    summary_soup_a = summary_soup.select("a:contains('記事全文を読む')")[0]

    news_body_link = summary_soup_a.attrs["href"]

    news_body = requests.get(news_body_link)

    news_soup = BeautifulSoup(news_body.text, "html.parser")

    data1 = str(news_soup.title.text + "\n")


    data2 = str(news_body_link + "\n")

    detail_text = news_soup.find(class_=re.compile("Direct"))

    data3 = str(detail_text)

    all_data = [data1, data2, data3]
    news_list.append(all_data)

file = open("/Users/fujisakihiyori/Downloads/研究/スポーツ.text", "w")
for i in range(len(news_list)):
    file.writelines(news_list[i])

file.close()
```

図 3 スクレイピングスクリプト

2.4. 関連研究

Web ニュース記事の地域度分析については，松本・馬らの Web ニュース記事の地域度計算に関する研究 [8] [9] がある．松本・馬らは Web ニュース記事の地域度を，記事内容の地域度(Localness of Contents) と，記事の話題の遍在性(Ubiquitousness of Topics)を用いて計算する手法を提案し，実際に ASAHI.COM に掲載されたニュース記事 500 件に対して地域度分析を行っている．

著者らは，災害報道が全国紙/地方紙によってどのように異 なっているかを比較分析(「朝日新聞/神戸新聞の災害報道の比 較」[10])するとともに，本研究の先行研究 [11][12] も開始している．

ニュース記事のトピック分析・感情分析による地図マッピンクについては，河合・熊本らの先行研究 [13] [14] [15] がある．

3. BERT 機械学習によるニュース記事の自動分類

本節では，BERT 機械学習によるニュース記事の自動分類について述べる．具体的には，Yahoo!Japan ニュースにおける記事のトピック分類（国内/国際/経済/エンタメ/スポーツ/IT/科学/地域の 8 トピック）を BERT を用いて行ったので，その詳細について説明する．

3.1. Web スクレイピングによるニュース記事の取得

まず，ニュース記事データを取得するために Web スクレイピングを用いて，Yahoo!ニュースの記事本文 URL から記事本文を取得した（図 3 参照）．

具体的には，以下の手順でデータ取得を行った．

① Yahoo!ニュースのトップページ情報を取得し，BeautifulSoup のライブラリに Yahoo!ニュースのページ内容を読み込ませる．

② URL に news.yahoo.co.jp/pickup が含まれるものを抽出し，記事本文の URL から記事本文ページの情報を取得した．

③ 取得した記事本文のページを BeautifulSoup で解析できるようにし，class 属性の中に「Direct」が含まれる行を抽出し，記事本文を出力した．

④ 出力結果から記事タイトルと記事本文を一つのテキストファイルとして保存した．

以上の手順で獲得したデータを，国内(116 件)，国際(120 件)，経済(123 件)，エンタメ(126 件)，スポーツ(118 件)，IT(97 件)，科学(95 件)，地域(126 件) の計 949 件，8 項目として保存し，これをファインチューニング用の学習データとした．

3.2. BERT によるニュース記事分類

Google Colaboratory を用いて，BERT によるニュース記事分類システムを構築した．本システムのプログラム構成などを図 1 に示す．

① データの準備
スクレイピングしたデータをファインチューニング用のデータセットとして保存する．すなわち，各項目に分類された計 949 件のニュースを Google ドライブに訓練データとして配置し，訓練データの読み込み用に Transformers/datasets ライブラリをインストールした．さらに，日本語データの読み込みに必要な，fugashi/ipadic ライブラリもインストールした．さらに，Google ドライブと連携し，スクレイピングしたデータセットを読み込める環境を構築した．

② 訓練データとテストデータの準備
データを訓練データとテストデータに分割し，csv ファイルとして Google ドライブに保存し，skekearn.model_selection の train_test_split()を用いて訓練データとテストデータに分割した．今回はテストデータの割合を指定していないため 25%がテストデータとなる．

③ ファインチューニング
（BertForSequenceClassification の利用）

日本語の事前学習済みモデルと紐付いたトークナイザーを読み込み，BertForSequenceClassificationによりファインチューニング（テキスト分類）を行った．
この際，日本語テキストデータを処理できるようにするため，東北大学作成の日本語 BERT モデル「cl-tohoku/bert-base-japanese」を指定した．
データセットの読込みについては，まず，load_dataset()により csv ファイルを読込む.このデータは一種のテーブルとして扱われるため，column_names=[“text”,” label]により各カラム(列)の名前を設定した．その後，訓練データ train_data に対して map()メソッドを用いて，各要素に対し，バッチサイズを訓練データ全体のサイズに設定して一度に処理できるようにした．最後に，set_format()メソッドにより各カラムを設定した．テストデータも同様に，トークナイズしフォーマットを整えておいた．

④ 分類モデルの精度評価の準備
sklearn.metrics の accuracy_score を用いて，モデルの精度を評価する関数を定義し，次に，TrainingArguments クラスによりハイパーパラメータを設定，Trainer クラスによりトレーナーを設定した．これらの設定に基づき，モデルを追加で訓練し，訓練済みのモデルを保存し，読み込みを可能とした．最終的に，evaluate()メソッドによりモデルを評価する．

⑤ ニュース記事の分類
このように設定したモデルを使いニュース記事の分類実験を行った．ニュース記事を分析するにあたり，ニュースの本文を取り出し，不要な文字を取り除くことでそれぞれのニュース記事の内容を読みやすくし，処理した本文の内容についても表示できるように設定を行った..ニュース記事の分類実験の詳細は以下の通りである：

(1) トークナイザーを用いて本文をトークンに変換し，単語を ID に変換した．
(2) PyTorch のテンソルに変換してモデルへの入力 x とし，入力単語数の上限を 512 に設定した．
(3) 入力 x を読み込んだモデルに渡すことで予測が行われる．得られる出力 y はタプル形式であるが，最初の要素に予測結果が入っている．8 クラス分の出力があるが,この中で最も値が大きいインデックスを，argmax(-1)により獲得し，得られた最大値のインデックス pred に対応したニュースカテゴリ名を表示する.これにより,どのカテゴリに分類されたかがわかる．

3.3. ニュース記事の分類モデルの精度確認

開発した分類モデルの精度を確認するため，すでに分類されたニュース記事を新規データとして入力し，正しいカテゴリに分類されているかを確認した．

例えば，開発した分類モデルに対して,「経済」と分類された記事を与えて，その結果何件が「経済」と正しく分類されるのかを調べた．結果として,「経済」に分類されたニュース記事 123 件を入力すると，93 件が正しく「経済」と判断された．また,「スポーツ」の記事も同様に与えた結果 118 件のうち 86 件が正しく「スポーツ」と判断された．全ての記事が正しい項目として表示されるのではなく，幾つかのニュース記事は異なるカテゴリに分類される結果となった（図 4,図 5 参照）．しかし，大半のニュースは元々のカテゴリと同じカテゴリに分類されるという結果を得ることができたので，ファインチューニングで入力するデータ数を増やすことで今後,精度を上げることができると考える．

さらに，ファインチューニング時には与えられなかった新規のニュース記事を入力して，その記事がどのカテゴリに分類される記事なのか，自動分類する事件も行った．

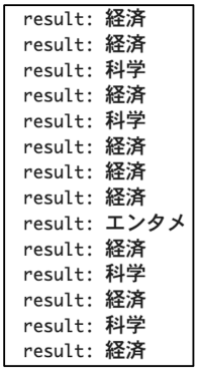


図 4 経済に分類されたニュース記事結果

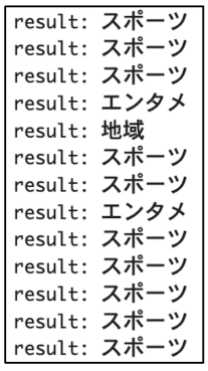


図 5 スポーツに分類されたニュース記事結果

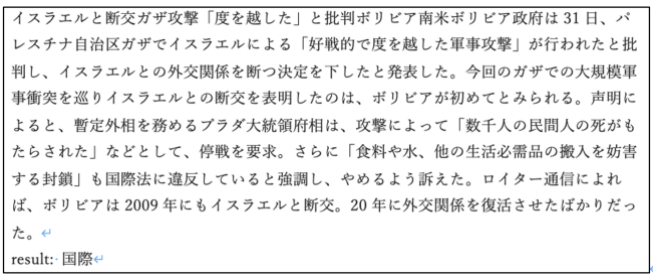


図 6 ニュース記事の自動分類結果

ここでの新規のニュース記事は Yahoo!ニュースにおいて主要のカテゴリに分類された記事を新規ニュース

とした．これらニュース記事も同様にスクレイピングを利用して入手した．このニュース記事 124 件を開発した分類モデルに入力して，どのカテゴリに分類いるかを確認した．この実験では，記事本文も出力し，記事内容が理解できるようにした（図 6 参照）．分類の結果 124 件のうち 115 件は正しいカテゴリに分類された．精度は約 93％でありニュース記事を自動分類できる環境を構築することができたものと考えることができる．

4. 地域度にもとづくニュース記事の国内･地域への再分類

前節で行ったトピック分類では，「国内」と「地域」のトピック分類の基準が不明･曖昧であった．さらに，本来，「地域」の下も，主要/経済/エンタメ/スポーツ/IT/科学/ライフのような階層的分類が存在しうるが，これも考慮されていなかったという問題が残っている．

そこで，本節では，「国内」と「地域」への分類を，記事内容の「地域度」を考慮して明確にする．すなわち，ニュース記事の「地域度」を計算し，計算された地域度に基づいてニュース記事が「地域」または「国内」カテゴリに属するかどうかを判断することを目標とする．

ここで，「地域度」とは，直観的には，ニュース記事の「地域」への関連度を表す指標である．ここで，「地域」については，市町村，県，圏，国などと言った階層性があることに注意が必要である．例えば，ニュース記事 N が与えられた場合，ニュース記事 N の「地域度」は，記事 N が言及している地域を推定すると共に，記事 N の言及地域への関連度の両方を考慮して決定する．

- 記事の言及地域の同定
  対象とするニュース記事中に出現する地域名(住所等)を抽出し，頻出する地域名を発見する．発見した地域名が複数ある場合は，最も多く登場している地名を選択する．今回はニュース記事全てに含まれている固有名詞の頻出回数を求め，最も多く報道されている地点はどこなのかを特定する．
- 記事の言及地域への関連度
  記事の言及地域への関連度とは，その記事内容が言及地域にどの程度関連しているのかを表す度合いである．この関連度を計算するために，記事の「ユビキタス性」を考慮する．記事のユビキタス性とは，その記事内容が日常的な生活情報になっているかを表す指標であり，ユビキタス性が高ければ，この記事の「言及市域への関連度」は高いものと判断する．これは，日常生活情報など日常性の高い情報は何時でも何処でもありうる情報なので，特定の地域の人々にしか興味を持たれないため，これらの記事はその言及地域への関連度が高いと判断する．

これらを判断するために，T 値や MI 値，TF-IDF 値などを利用してニュース記事の内容を分析した．

記事の言及地域の同定をするため，MeCab を利用して記事内容の形態素解析を行う．テキストを扱いやすいよう，空行を削除するなどの処理を行い，“名詞-固有名詞-地域-一般”に分類される固有名詞を抽出し，頻出単語 10 単語を出力する．その結果から，記事の言及地域を特定する．

また，記事の言及地域のユビキタス性を考慮するため，形態素解析により “名詞-固有名詞-組織”を抽出し，地域の記事との記事の中で，報道する組織の大きさについて比較する．さらに，TF-IDF による重要語単語抽出を行う．TF-IDF とは文書中に含まれる単語の重要度を評価する手法の 1 つで、以下のような式で計算される．

$$TFIDF_{i,j} = TF_{i,j} \cdot IDF_i$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$Idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

TF(i, j) = 文章 j における単語 i の出現頻度/文章における全単語の出現頻度の和

IDF(i) = log(全文章数/単語 i を含む文章数)

TFIDF(i, j) = tf(i, j) * idf(i) = 単語の重要度

TF(Term Frequency):単語の出現頻度
（文章 A における単語 x の出現頻度）

IDF(Inverse Document Frequency):逆文書頻度
(一般的な用語程，idf の値が低くなり，重要な単語程高くなる)

重要語においても MeCab を利用して抽出を行う．MeCab を利用し，名詞,形容詞,動詞の単語を抽出する．そこで TF-IDF の計算を行う(ただし文書全体の 50%以上で出現する単語は無視する)．TF-IDF 行列を取得し，単語リストを取得する．そして，それぞれの記事における重要語を上位 5 つを表示する．

辺野古	2
福岡	2
那覇	2
普天間	1
宜野湾	1
名護	1
福島	3
郡山	1
北海道	1
山梨	2
大月	2
真木	1
埼玉	1
三芳	1
ふじ	1
甲府	1
江東	3

図 7 地域カテゴリの記事に出現する地域固有名詞

東京	1
香港	1
東京	1
長崎	2
徳島	2
高知	2
東北	1
チャイナ	1
中東	3
パレスチナ	1
ガザ	1
紅海	1
アフリカ	1

図 8 国内カテゴリの記事に出現する地域固有名詞

東京電力	1
関西電力	1
中部電力	1
東洋大学	1
立教大学	1
東洋大	1
山陰中央新報社	1
警視庁	1
JR	1
ＪＲ九州	3
ＪＲ	2

図 9 地域カテゴリの記事における組織の固有名詞

衆院	1
最高裁	7
自民	1
公明党	1
参院	2
民主党	5
労働省	1
財務省	1
環境省	1
財務省	1

図 10 地域カテゴリの記事における組織の固有名詞

ハロウィン:	0.334428
フルマラソン:	0.260950
フリースクール:	0.492795
パトロール:	0.426206
スクールバス:	0.301754

図 11 重要語抽出

記事の言及地域の同定については，地域のカテゴリに分類されている記事では，辺野古や三芳，甲府のような市や町の単位まで記事の中に地名が存在していることがわかった（図 7 参照）.そして国内に分類されている記事では，東京や香港，アフリカのような国単位や東北などの比較的広い範囲の地名が記されている結果となった（図 8 参照）．

記事の言及地域への関連度について，記事の中で報じられている組織の大きさについて比較すると，地域の記事については，東京電力や関西電力，東洋大学や立教大学，JR 九州のようなそれぞれの地域に存在する地名が存在していることがわかった（図 9 参照）．国内の記事については，最高裁や民主党，公明党，労働省などの日本における大きな組織が記述されているということがわかった（図 10 参照）．

また，記事のユビキタス性を考慮するための重要語については，地域のカテゴリの記事には，ハロウィンやフルマラソンのような，どの地域でも行われている出来事が重要語として挙げられた（図 11 参照）．この結果よりユビキタス性が高い記事が地域のカテゴリに分類されていることがわかる．しかし，重要語においてユビキタス性が高いと考えられる単語が多くは見られなかったことから，重要語以外でユビキタス性を考慮する必要があると考える．

5. ニュース記事の多様性分析

同時期・同一トピックのニュース記事も，多様な視点から報道がなされるニュース記事は，時期・トピックが同一であっても掲載すべきであると考えられる．そこで，本論文では，同時期で同一トピックのニュース記事について類似度を計算し，類似度が低ければ，異なる視点からのニュース記事と判定し掲載するという手法を提案する．類似度計算については，BERT を利用した文の意味上の類似度計算を用いた．

BERT における類似度計算については，hugging face 上で公開されている学習済みのモデルの中から tokenizer と学習モデルのインスタンスを生成する．tokenizer を使い，文の意味を理解しやすいようにするため，単語(token)ごとに区切り，その単語を数字ラベルに変換する．BERT に文を入力すると，入力単語ごとのベクトルを出力し，これを用いて，文を文の意味を表すベクトルに変換し，文同士の意味上の類似度を計算するために利用することとした.文の類似度は cos 類似度を使って算出することとした． -1.0~1.0 の間の値で大きければ大きいほど文の意味が類似していると判断した．

類似度計算に利用するデータは，朝日新聞と産経新聞の政治カテゴリに分類された記事を利用した．同じ日時に報道されたニュースの中で内容が類似していると考えられるニュース記事を選択する．

ニュース1	0.97717303
ニュース2	0.99056375
ニュース3	0.985054
ニュース4	0.989095
ニュース5	0.98511875
ニュース6	0.954879
ニュース7	0.98112166
ニュース8	0.9871307
ニュース9	0.9829194
ニュース10	0.983214
ニュース11	0.9867001
ニュース12	0.9681172
ニュース13	0.98487175
ニュース14	0.9733889
ニュース15	0.9736293
ニュース16	0.98775977

図 12 類似度計算結果

朝日新聞と産経新聞の類似度計算の結果，ほとんどのニュース記事においての類似度が 0.9 以上となり類似度が高いという結果になった（図 12 参照）．この結果から，朝日新聞，産経新聞の 2 つの報道者において政治記事においては類似の視点から報道されていることが予想される．このように，実験を進める中で，類似度が低くなるニュース記事を発見することが大変困難であるということがわかった．むしろ，朝日新聞と産経新聞のニュース記事の分析を進めると，同じニュース内容ではなく，朝日新聞，産経新聞どちらかでしか報道されていないニュース記事が多く存在していることに気づいた，このため，同じニュースに絞って多様性を考えるのではなく，各社が報道しているニュース集合の違いに焦点を当てる必要もあると考える．

6. あとがき

本論文では，

① 機械学習によるニュース記事の自動分類，
② ニュース記事の「地域度」の定義･計算法，
③ ニュース記事の「多様性」の定義・計算法

を提案した．

機械学習によるニュース記事の自動分類については，既存のニュースサイトで用いられているカテゴリおよびニュース記事を機械学習することで，新たなニュース記事をどのカテゴリに分類するかという作業を自動化できる可能性を示すことができた．

次に，ニュース記事の所属するカテゴリを「国内」とするか「地域」とするかが曖昧な場合に対処するために，「地域度」という概念を導入した．「地域度」とは，記事が言及している地域と，記事の言及地域への関連度の両方を考慮して計算するものである．具体的には，記事に出現する住所や緯度経度情報から言及地域の同定を行い．さらに，記事内容の話題の「ユビキタス度」(どこにでも出ている度合い）を用いて記事の言及地域への関連度を計算するものである．

次に，複数のニュースソースから提供される同一トピックのニュース記事群の扱いについて，ニュース記事の「多様性」という概念を導入し，ニュース記事内容が大きく異なっていれば多様性が高いと判断し，両方共に掲載することができるような多様性計算法を導入した．

今後の課題としては,実際のカテゴリ分類とは異なる分類結果が出たニュース記事の特徴･原因を調べること，地域度の計算の精度を上げること，ニュース記事の多様性計算を効果的に行える「非類似度」計算の方法などを考えるということが挙げられる．

謝辞

本研究は，令和 4, 5 年度科研費基盤研究 (B)(一般)「機械学習による情報の意味獲得と意味類似情報の検索・生成」(代表: 田中克己，課題番号 22H03905)によるものである．

参 考 文 献

[1] 情報通信政策研究所: 令和 4 年度 情報通信メディアの利用時間と情報行動に関する調査報告書, 2023 年 6 月 24 日 https://www.soumu.go.jp/iicp/research/results/media_usage-time.html

[2] NHK 放送文化研究所: 国民生活時間調査 2020 生活の変化×メディア利用, 2021 年 5 月 https://www.nhk.or.jp/bunken/research/yoron/pdf/20210521_1.pdf

[3] Hussman School of Journalism and Media (Univ. of North Carolina): Technological Capabilities: The Algorithm as Editor, https://www.usnewsdeserts.com/reports/news-deserts-and-ghost-newspapers-will-local-news-survive/the-news-landscape-of-the-future-transformed-and-renewed/technological-capabilities-the-algorithm-as-editor/

[4] P. M. Abernathy: Will Local News Survive?, https://www.usnewsdeserts.com/wp-content/uploads/2020/06/2020_News_Deserts_and_Ghost_Newspapers.pdf

[5] J.Devlin, M.-W.Chang, K.Lee and K.Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2 [cs.CL] May 24, 2019

[6] A.Vaswani, L. Jones, Noam Shazeer, N. Parmar, J. Uszkoreit, L.Jones, A.N.Gomez, Ł.Kaiser, I.Polosukhin: Attention Is All You Need, arXiv:1706.03762v7 [cs.CL] August 2, 2023

[7] 吾妻幸長：BERT 実践入門 PyTorch+Google Colaboratory で学ぶあたらしい自然言語処理技術, 翔泳社，2023 年 7 月 20 日

[8] Chiyako Matsumoto, Qiang Ma, and Katsumi Tanaka:

Web information retrieval based on the localness degree. In Proceedings of DEXA2002, Lecture Notes in Computer Science 2453, pp. 172–181, 2002.

[9] Qiang Ma, Chiyako Matsumoto, and Katsumi Tanaka: A localness-filter for searched web pages. In Proceedings of 5th Asian-Pacific Web Conference(APWeb) , Lecture Notes in Computer Science 2642, pp. 525–536, 2003.

[10] 藤崎日和: 新聞記事の比較分析 地方紙/全国紙, 経済紙/一般紙の比較分析 , 2021 年度福知山公立大学地域情報 PBL 基礎最終報告書, 2021.

[11] 藤崎日和, 田中克己:ニュース記事の地域性分析に基づく地図への自動分類. 東海関西データベースワークショップ 2022, https://sites.google.com/mil.doshisha.ac.jp/dbws2022, 2022 年 9 月.

[12] 藤崎日和, 荘司慶行, 田中克己. ニュース記事の地域度分析に基づく地図への自動分類, DEIM Forum 2023, No.3a-6-3, 2023 年 3 月. https://proceedings-of-deim.github.io/DEIM2023/3a-6-3.pdf

[13] 張建偉, 河合由起子, 熊本忠彦, 田中克己. 地域性に基づく発信者の観点差異を可視化するセンチメントマップシステムの提案.情報処理学会論文誌 データベース, Vol. 3, No. 1, pp. 38-48,2010.

[14] 張建偉, 河合由起子, 熊本忠彦, 白石優旗, 田中克己. 多様な印象に基づくニュースサイト報道傾向分析システム. 知能と情報（日本知能情報ファジイ学会誌）, Vol. 25, No. 1, pp. 568-582,2013.

[15] 熊本忠彦, 河合由起子, 田中克己. 回帰分析を応用したテキスト印象マイニング手法の設計と評価. In The 24th Annual Conference of the Japanese Society for Artificial Intelligence, pp.1B1–1, 2010.

観光地レビュー分析に基づく お出かけスポット評価項目別自動スコアリング

福嶋亮太朗† 張 建偉†† 奥 健太††† 中島 伸介†

† 京都産業大学 情報理工学部 〒 603–8555 京都府京都市北区上賀茂本山
†† 岩手大学 理工学部 〒 020-8551 岩手県盛岡市上田 4-3-5
††† 龍谷大学 先端理工学部 〒 520-2194 滋賀県大津市瀬田大江町横谷 1 番 5
E-mail: †{g2054155,nakajima}@cc.kyoto-su.ac.jp, ††zhang@iwate-u.ac.jp, †††okukenta@rins.ryukoku.ac.jp

あらまし 観光サイトでは，観光地の場所や営業時間などの基本情報とともに，観光客のレビューや観光地に対する評価などが記載されている．ユーザは，評価やレビューを参考に観光地を選択することがあり，ユーザに与える影響は大きいと考えられる．本研究では，fastText を用いて観光地レビュー文を分析することで，複数の評価項目ごとにスコアリングする手法を提案する．また，各観光地のスコアリング結果を用いることで，個人の趣向に沿った観光地推薦を実現することを目的としている．スコアリングシステムの評価実験では，従来手法と比較して提案手法の方が nDCG@10 で高いスコアを得られた．観光地推薦サイトのユーザビリティ評価実験では，従来手法と提案手法を比較すると「見栄え」「明快さ」「効率」「刺激」「ノベルティ」の項目で有意差を得られた．

キーワード 情報推薦, レビュー分析, 観光スポット, 可視化, fastText

1. はじめに

じゃらん [1] やトリップアドバイザ [2] のように観光地を対象として評価やレビューを書き込める観光レビューサイト（以下，観光サイトと呼ぶ）がある．旅行を計画する際には，観光地レビューや観光地の評価を参考にして旅行先を決定する場合があるため，観光地レビューや評価は重要である．しかし，観光サイトの評価は一次元で表現されており，わかりやすいという利点があるが，観光地の特徴やレビュー投稿者が観光地に満足した点が不明確という欠点がある．観光地の特徴やレビュー投稿者が観光地に対して満足した点を理解するためにはレビュー文を読む必要があるが，レビューを読むのには時間と労力が掛かってしまい，行ってみたい観光地を発見するには多大な時間を要することがある．また，観光サイトではジャンルに基づく検索が可能である．例えば,「神社・神宮・寺院」や「アウトドア」のようなものである．しかしながら，実際に観光地を訪れてみると想像と違っていたりすることがあり，ユーザ側の評価を観光地側が必ずしも把握できるとは限らない．そのため，ユーザ目線の評価を適切に反映することが必要である．

そこで本研究では，上記の課題を解決するために，観光サイトに投稿されたレビューを分析することで，評価項目別でスコアリングを行い観光地の特徴を明確化する．また，それらの結果を用いて，ユーザが観光地の評価項目に求める理想値を入力することで，個人の趣向に沿った観光地の推薦を行うシステムの開発を目的とする．

2. 関連研究

近年，商品の情報や観光地のレビューを掲載しているサイトが増加している．様々な商品をオンラインで販売している楽天市場 [3] や，コスメや美容に特化した総合情報サイトの@cosme [4] がある．また，旅行に特化したじゃらん [1] などの観光サイトもあり，多くの人が利用している．レビューサイトが増加したことから，レビューを扱う研究が盛んに行われている．

矢部ら [5] は，コスメアイテム特徴の相違点を可視化し，相違点可視化システムを利用したコスメアイテム推薦手法を提案している．レビューを分析し，評価表現辞書を用いることでコスメアイテムの自動スコアリングを行い，購入済みアイテムと検討中のアイテムの相違点を可視化している．また，コスメアイテムの評価項目 (潤い・浸透，色持ち，ナチュラル等) に理想値を入力することで，理想のアイテムを発見できるシステムを構築している．

住友ら [6] は，投稿された口コミ情報を収集し，感情分析することで各感情分析スコアを算出し，スコアに最も近いスポットを推薦している．しかし，このシステムでは周辺地域のスポットしか発見することができず，旅行先の観光地を発見することなどが出来ないため用途が非常に限定的である．

鳥山ら [7] は，年末年始に行われる無料サービスや夏に行われる祭りなど「季節特有の情報」と春には桜が咲き，秋は紅葉が綺麗など「季節ごとに変化する情報」を抽出するために，観光レビューからトピックを抽出し，トピックごとにクラスタリングを行うことで 2 種類の情報を抽出している．本論文では，季節で変化する情報ではなく，レビューの感情を分析する点で異なる．

また，観光写真から観光スポットを推薦する研究も行われている．中野ら [8] は，ユーザが観光地で撮影した写真を画像分析することで，写真の中からユーザ固有の好みタグを抽出することで，個人にパーソナライズされた観光スポット推薦を提案している．しかし，このシステムでは観光地で写真を撮ること

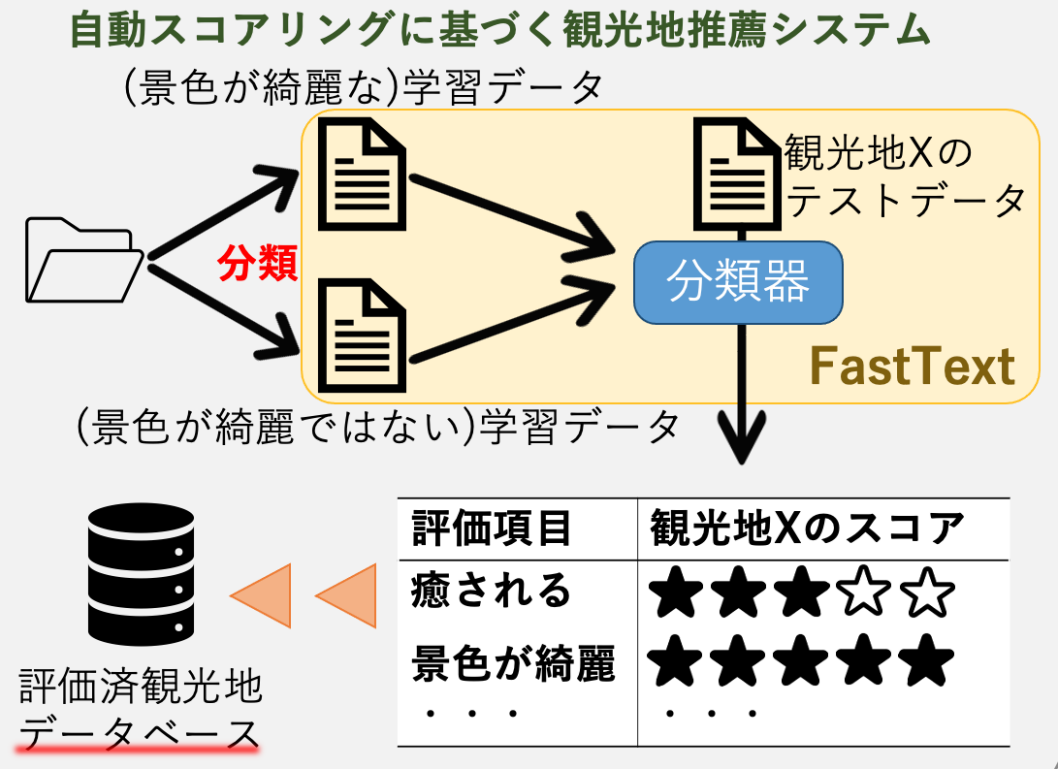


図 1　分類器の作成方法および学習データの作成

を前提に作られており，必ずしも誰でも使用できるとは限らない．また，写真中から誤検出や異なるタグを抽出してしまっていることで，好みを抽出できているとは言い難い．

新井ら [9] は，Twitter から観光地でリアルタイムにつぶやかれたツイート情報を収集し，それらの情報を活用して観光ルートを推薦する手法を提案している．ユーザが出発地，到着地，出発時間，到着時間を条件として入力することで，入力条件を満たした観光スポットを提示するというメリットがあるが，必ずしもユーザが行ってみたいと思える観光スポットを提示できるとは限らない．

3. 観光地レビュー分析に基づくお出かけスポット評価項目別自動スコアリング

3.1 概　　要

観光地の情報を入手する手段として SNS, ガイドブック，観光サイトなどが挙げられる．観光サイトは，観光地の評価やレビューを書き込めることから，多くのユーザが観光地を決定する際に参考にしている．しかし，観光サイトの評価は一次元で表現されており，レビュー投稿者が満足した点や観光地の特徴が不明確である．また，観光地の特徴やレビュー投稿者が満足した点を理解するには，レビューを読み進める必要があり，時間と労力が掛かってしまう．本研究のシステムでは，投稿者のレビューを分析することで観光地の特徴を詳細にスコアリングし，一次元の評価から多次元な評価を行う．また，ユーザが観光地に求める理想スコアを入力することで，最も理想に近い観光地を推薦するシステムの構築を目指す．

3.2 自動スコアリング手法

本節では自動スコアリングの手法について述べる．じゃらん [1] やトリップアドバイザ [2] のようなレビュー投稿サイトでは多種多様な人が観光地を訪れた後にレビューを投稿している．本稿では，実際に投稿されたレビューを収集して学習することで分類器の作成を行い，自動スコアリングを行っている．分類器の作成には，fastText を用いている．fastText を利用する理由は，オープンソースのライブラリであり，高速な処理が可能だからである．レビューを収集するために，じゃらん [1] のサイトをスクレイピングすることでレビューを収集する．収集したレビューを人手で読み，スコアリングしたい項目に対応している場合は (項目な) 学習データに分類する．スコアリングしたい項目に対応していない場合は (項目ではない) 学習データに分類する．本稿でスコアリングする評価項目は下記の通りである．

- 景色が綺麗
- 癒される
- ご飯が美味しい

次に，具体的な分類器の作成方法およびスコアリング手法について述べる．図 1 のような流れで分類器を作成し，スコアリングを行う．

次に具体的な例を用いて説明する．下記のような 3 つの例文があったとする．

（1） 展望台から眺める景色に感動しました
（2） 展望台は人が多くて景色を楽しめません
（3） 夜景を眺めるとうっとりします

景色が綺麗な分類器を作成しようとする場合は，景色が綺麗な学習データに「展望台から眺める景色に感動しました」，「夜

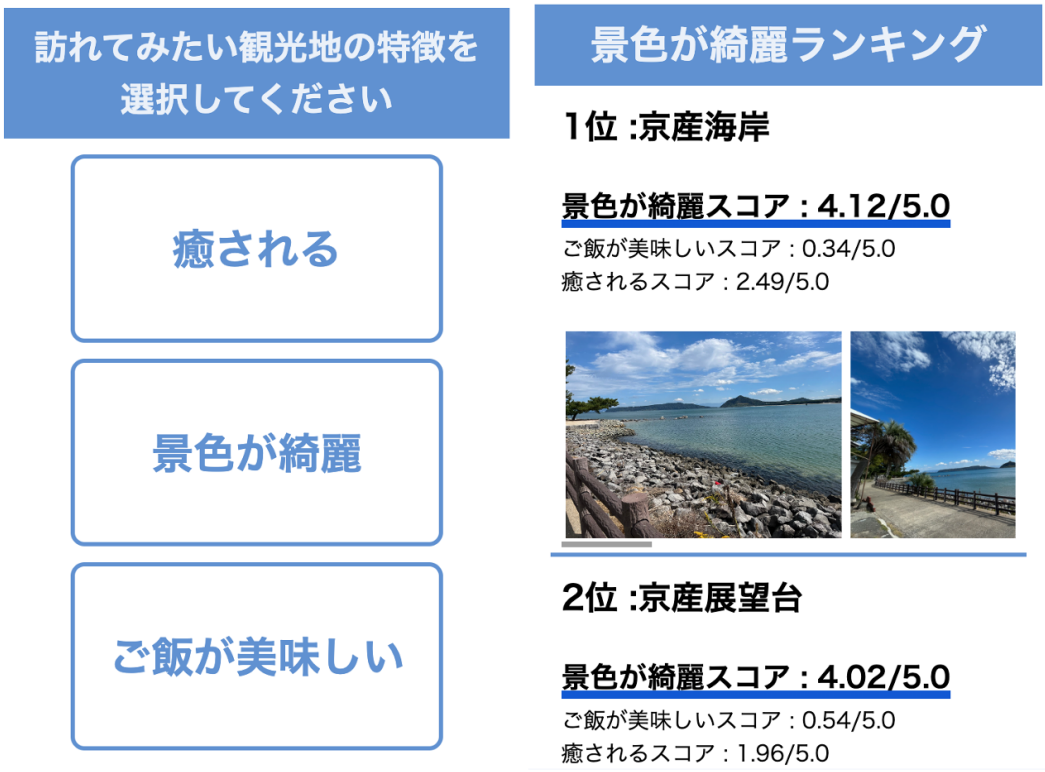


図 2　単一評価項目による観光地推薦のインタフェース

景を眺めるとうっとりします」のレビューを追加する．景色が綺麗でない学習データに「展望台は人が多くて景色を楽しめません」のレビューを追加する．このように (1)(3) を正事例, (2) を負事例とした学習データを作成する．その学習データを用いて分類器を学習する．ここで，文書の特徴化には fastText を用いる．同様の手法で「癒される分類器」と「ご飯が美味しい分類器」も作成する．これらの分類器に対して，観光地 X のテストデータを入力すると各項目ごとに確率スコアが出力される．確率スコアは 0〜1 の範囲で出力されるが，確率スコア× 5 の値を観光地 X のスコアとする．確率スコアに 5 倍する理由としては，既存の観光サイトは 0〜5 で評価されており，ユーザがそれに慣れているので合わせた．観光地 X のスコアは評価済観光地データベースに保存することで，ユーザに観光地推薦を行う際に利用する．

3.3　単一評価項目による観光地推薦手法

本節では，単一評価項目のスコアリングを用いた観光地推薦手法について示す．ユーザは訪れてみたい観光地の特徴を選択する．図 2 の左側に実際の画面を示す.「癒される」,「景色が綺麗」,「ご飯が美味しい」の 3 つの項目があるので，任意の項目を選択する．項目を選択すると，選択した項目のスコア順にランキング化されて，図 2 の右側のように表示される．景色が綺麗なスコア，ご飯が美味しいスコア，癒されるスコアを表示することによって，観光地の特徴が明確化され行ってみたい観光地の発見に繋がる．

3.4　複数評価項目のバランスによる観光地推薦手法

本節では，複数評価項目のバランスによる観光地推薦手法について示す．ユーザは観光地に求める理想スコアを入力して検索をクリックすることで，好みに沿った観光地が推薦される．ユーザは推薦結果に基づいて理想スコアを調整することが可能なので，再度条件を入力して検索することで新たな観光地が推薦される．このようなインタラクティブな検索を繰り返すことで，ユーザの好みに沿った観光地推薦が実現可能になる．

図 3 に実際の画面を示す．緑色のつまみはスライドバーになっているため，ユーザは任意の値を入力することが可能である．また,「考慮なし」にチェックを入れるとそれ以外の項目のみを考慮して類似度計算を行い，観光地推薦を行う．推薦結果の表示方法は，図 2 の右側と同じである．



図 3　複数評価項目のバランスによる観光地推薦手法のインタフェース

推薦手法を決定するために予備実験を行った．本予備実験では，被験者に仮定の理想スコアと推薦された観光地を提示して，最も好みの推薦結果をランキング化して貰う．設問は 3 つあり，各ランキングの 1 位の推薦手法は 5 点のスコアを付与，2 位の推薦手法は 4 点のスコアを付与 … 5 位の推薦手法は 1 点のス

コアを付与して，平均スコアが最も高いものを優れた推薦手法として利用する．

予備実験 1 つ目の設問

- 癒される: 4.5
- 景色が綺麗: 4.0
- ご飯が美味しい: 3.0

上記のスコアとに推薦される観光地を提示して，5 つの推薦手法の中で最も好みの推薦結果をランキング化して貰った．結果は表 1 の通りである．

表 1　予備実験設問 1 つ目の結果

ランキング	推薦手法	平均スコア
1 位	ハイブリッド (ユークリッド: コサイン=3:7)	3.43
2 位	ユークリッド距離のみ	3.38
3 位	ハイブリッド (ユークリッド: コサイン=5:5)	3.07
4 位	ハイブリッド (ユークリッド: コサイン=7:3)	2.79
5 位	コサイン類似度のみ	2.36

予備実験 2 つ目の設問

- 癒される: 3.8
- 景色が綺麗: 4.6
- ご飯が美味しい: 考慮無し

上記のスコアとに推薦される観光地を提示して，5 つの推薦手法の中で最も好みの推薦結果をランキング化して貰った．結果は表 2 の通りである．

表 2　予備実験設問 2 つ目の結果

ランキング	推薦手法	平均スコア
1 位	ハイブリッド (ユークリッド: コサイン=7:3)	3.93
2 位	ハイブリッド (ユークリッド: コサイン=5:5)	3.50
3 位	ハイブリッド (ユークリッド: コサイン=3:7)	3.07
4 位	ユークリッド距離のみ	2.93
5 位	コサイン類似度のみ	1.57

予備実験 3 つ目の設問

- 癒される: 考慮無し
- 景色が綺麗: 4.7
- ご飯が美味しい: 4.0

上記のスコアとに推薦される観光地を提示して，5 つの推薦手法の中で最も好みの推薦結果をランキング化して貰った．結果は表 3 の通りである．

表 3　予備実験設問 3 つ目の結果

ランキング	推薦手法	平均スコア
1 位	ハイブリッド (ユークリッド: コサイン=3:7)	4.07
2 位	ハイブリッド (ユークリッド: コサイン=5:5)	3.93
3 位	ハイブリッド (ユークリッド: コサイン=7:3)	3.57
4 位	ユークリッド距離のみ	2.29
5 位	コサイン類似度のみ	1.14

予備実験の設問 1 から設問 3 までの結果を平均すると表 4 の結果となった．表 4 の結果より，下記式で類似度を計算し，観光地推薦を行う．

$$y = \frac{\sum_{k=1}^{n} a_k b_k}{\sqrt{\sum_{k=1}^{n} a_k}\sqrt{\sum_{k=1}^{n} b_k}} \times 0.7 + \sqrt{\sum_{k=1}^{n} a_k - b_k} \times 0.3 \quad (1)$$

表 4　予備実験結果

ランキング	推薦手法	平均スコア
1 位	ハイブリッド (ユークリッド: コサイン=3:7)	3.52
2 位	ハイブリッド (ユークリッド: コサイン=5:5)	3.50
3 位	ハイブリッド (ユークリッド: コサイン=7:3)	3.42
4 位	ユークリッド距離のみ	2.85
5 位	コサイン類似度のみ	1.69

4. 評 価 実 験

本稿では，評価項目別スコアリングシステムの評価実験と観光地推薦サイトのユーザビリティ評価実験を行った．4.1 節で評価項目別スコアリングシステムの評価実験について説明し，4.2 節で観光地推薦サイトのユーザビリティ評価実験について説明する．

4.1 評価項目別スコアリングシステムの評価実験

4.1.1 実 験 概 要

評価項目別スコアリンングシステムの精度を検証するために，評価実験を行った．本実験では，被験者に 10 個の観光地・各 10 レビューを読んでいただき，各項目 (癒される，景色が綺麗，ご飯が美味しい) ごとにランキングして貰い，1 位には 10 点，2 位には 9 点, 3 位には 8 点 … 10 位には 1 点というように利得スコアを与え，利得スコアが高い順にランキング化した結果を正解データとしている．提案手法でランキング化した結果とベースラインでランキング化した結果を比較することでどちらの手法が優れているのか評価する．

ベースラインはレビュー中の単語の出現回数でランキング化している．癒されるのベースラインは「癒される, リラックス, 落ち着く, くつろぐ, 和む, 休まる, のんびり」 の出現回数でランキング化を行う．景色が綺麗のベースラインは「絶景, 眺めが良い, 綺麗」の出現回数でランキング化を行う．ご飯が美味しいのベースラインは「美味しい, 絶品」の出現回数でランキング化を行う．

ご飯が美味しい，景色が綺麗の項目では本学に所属している 14 名の被験者が参加しており，癒されるの項目では本学に所属している 8 名の被験者が参加している．

4.1.2 結果と考察

ランキング化の評価は nDCG@10 を用いて評価する．nDCG@k とは，ランキングアルゴリズムを評価する指標であり，ランキング内のアイテムとランキング順位を考慮して測定する．nDCG@10 で評価した結果は図 4 に示す．

全項目でベースラインと比較して提案手法が優れた結果を得られている．癒されるのベースラインでは「駅前の足湯もとて

も気持ちいいです」という文章はスコアが付与されないが，提案手法では正しくスコアリングできているため，ベースラインと比較して優れた結果が得られた．また，景色が綺麗なベースラインでは「360 度の大パノラマ．石垣島の最北端に位置します．石垣に来たのなら，絶対に行くべきポイントです.」という文章でもスコアが付与されないが，提案手法では正しくスコアリングできているため，ベースラインと比較して優れた結果が得られた．

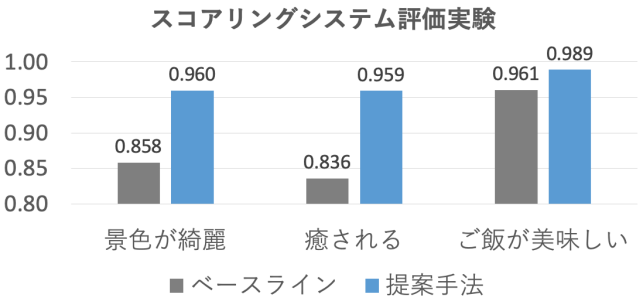


図 4　評価項目別 nDCG@10 の結果

4.2　観光地推薦サイトのユーザビリティ評価実験

4.2.1　実 験 概 要

評価項目別スコアリングシステムを用いて作成した観光地推薦サイトのユーザビリティ評価を行うため，評価実験を行った．評価実験は，観光サイトのじゃらん [1], 本研究で開発した「単一評価項目による観光地推薦サイト」，「複数の評価項目のバランスによる観光地推薦サイト」を利用していただき，User Experience Questionnaire (UEQ) のアンケートに回答してもらった．(注1)尚，被験者は大学生の計 13 名が参加している．

User Experience Questionnaire (UEQ)

User Experience Questionnaire (UEQ) [10] とは，Laugwitz らによって制作され，ユーザビリティの側面 (効率，明瞭さ，信頼性) とユーザー エクスペリエンスの側面 (独創性，刺激) を包括的に測定する評価指標である.「わかりやすいーわかりにくい」や「創造的ー創造的でない」などの計 26 個の質問に 1〜7 のリッカート尺度で質問に回答してもらい，評価を行う．

4.2.2　結果と考察

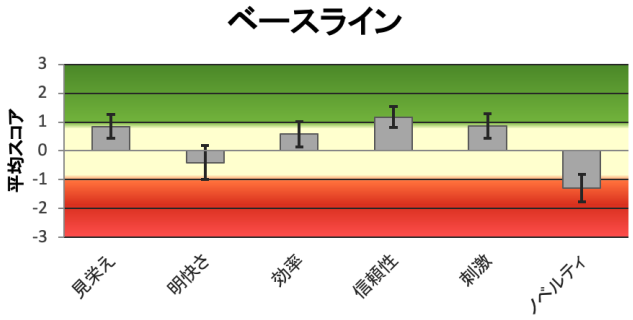


図 5　UEQ のアンケート結果 ベースライン

(注1)：本実験では順序効果を考慮できていない

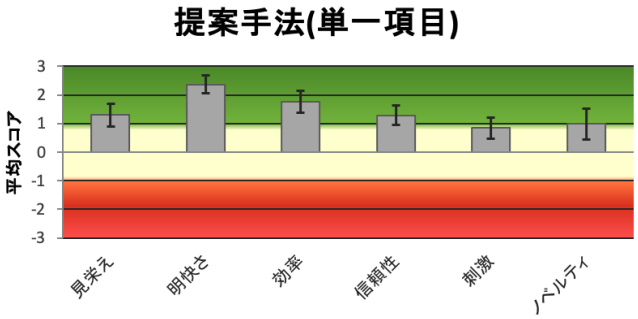


図 6　UEQ のアンケート結果 提案手法 (単一項目)

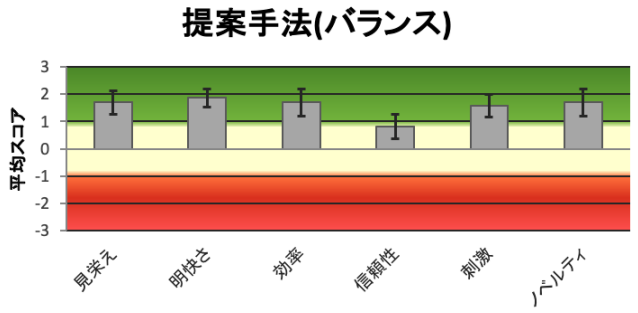


図 7　UEQ のアンケート結果 提案手法 (バランス)

各手法で UEQ のアンケートに回答してもらった結果を図 5, 図 6, 図 7 に示す．図 5〜7 をまとめた結果を図 8 に示す．

UEQ アンケートを t 検定で分析した結果を示す．表 5 に従来手法と提案手法 (単一評価項目による観光地推薦サイト) の結果に対して，t 検定を用いて分析した結果を示す．t 検定は α =0.05 かつ両側検定で分析を行った．表 6 に従来手法と提案手法 (複数の評価項目のバランスによる観光地推薦サイト) の結果に対して，t 検定を用いて分析した結果を示す．

表 5　従来手法と提案手法 (単一項目) 間の t 検定分析結果

項目	t 値	p 値
見栄え	3.77	1.99
明快さ	11.16	2.01
効率	4.64	2.01
信頼性	0.57	2.01
刺激	-0.11	2.01
ノベルティ	8.61	2.01

表 6　従来手法と提案手法 (バランス) 間の t 検定分析結果

項目	t 値	p 値
見栄え	6.62	1.99
明快さ	8.47	2.01
効率	4.24	2.01
信頼性	-1.70	2.01
刺激	3.73	2.01
ノベルティ	6.27	2.01

表 5 の結果より，従来手法と提案手法 (単一評価項目による観光地推薦サイト) を比較すると「見栄え」「明快さ」「効率」「ノベルティ」の項目で有意差を得られた．

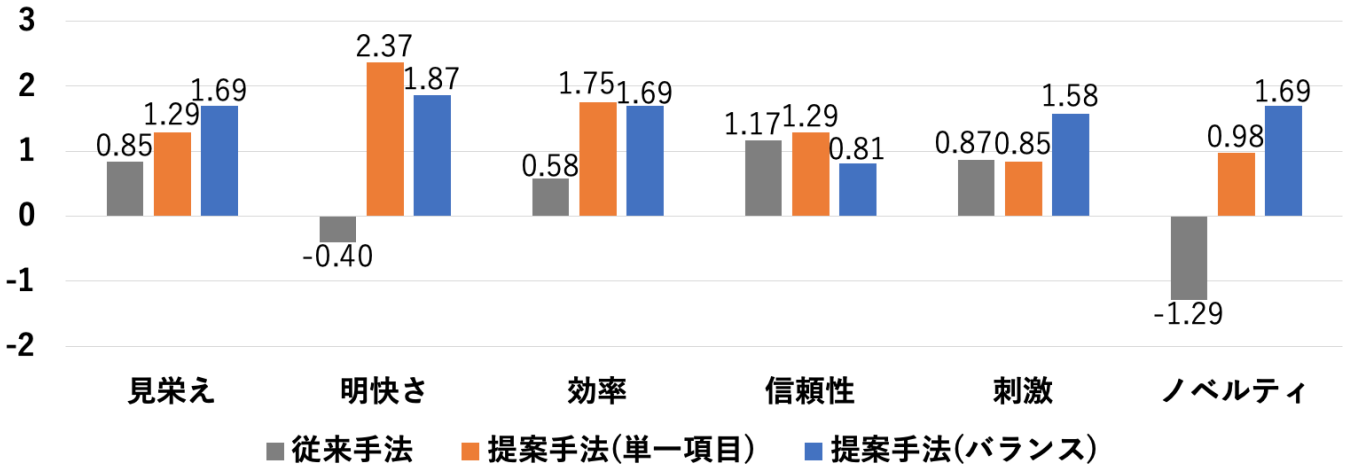


図 8　UEQ のアンケート結果

表 6 の結果より，従来手法と提案手法 (複数の評価項目のバランスによる観光地推薦サイト) を比較すると「見栄え」「明快さ」「効率」「刺激」「ノベルティ」の項目で有意差を得られた．「明快さ」とは，インタフェースや情報などがユーザにとってわかりやすいかを評価する指標である．提案手法では，観光スポット名・写真」「各項目のスコアのみを表示しており，必要な情報のみを得られるので明快さがベースラインと比較して優れた結果が得られた.「ノベルティ」とは，デザイン・機能性など新規性を評価している指標であるが，従来手法にはない手法でスコアリングしているため，ベースラインと比較して優れた結果が得られた.「信頼性」とは，システムが意図通りに動作しているのか，動作が安全であるのかと言った，主観的な信頼性を評価する指標である．評価実験を実施した際に「全ての項目の理想スコアを高く設定したが，理想とは異なる観光地が推薦された」というコメントを受けた．(1) 式で類似度を計算しており，観光地のスコアが低い場合でもコサイン類似度が高くなってしまい，理想とは異なる観光地が推薦されてしまうため，信頼性が従来手法よりも低い結果となった．

5.　今後の方向性

本稿では，4.2 節で UEQ を用いた評価実験を実施した際に，提案手法の良かった点・改善すべき点を被験者に尋ねた．その中で「より多くの項目があればいいな」「観光地の詳細を知りたい」「推薦された根拠を知りたい」などの声が多く見受けられた．これらの結果より，評価項目の増加・観光地詳細情報の表示・スコアリングおよび推薦結果の根拠を提示することに今後は取り組んでいく．

現在は，観光地を評価する分類器を生成するために，学習データを手作業で作成しており，評価項目を増やすのに多くのコストと時間が必要である．このことから，学習データの自動作成にも取り組んでいく予定である．

6.　ま　と　め

本稿では，観光サイトの評価が一次元で表現されており，観光地の特徴が不明確という課題を解決するため，レビューを分析することで多次元的な評価を行った．また，本研究で開発した観光地推薦サイトを用いて評価実験を行った．その結果，従来手法と提案手法 (単一評価項目による観光地推薦サイト) を比較すると「見栄え」「明快さ」「効率」「ノベルティ」の項目で有意差を得られた．従来手法と提案手法 (複数の評価項目のバランスによる観光地推薦サイト) を比較すると「見栄え」「明快さ」「効率」「刺激」「ノベルティ」の項目で有意差を得られた．今後は，効率的に評価項目を増やしていくために，学習データの自動作成に取り組んでいく．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22H03700，22K12281) および京都産業大学先端科学技術研究所（ヒューマン「」マシン・データ共生科学研究センター）共同研究プロジェクト（M2001）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] じゃらん net，https://www.jalan.net/kankou/
[2] Tripadvior, https://www.tripadvisor.jp/
[3] 楽天市場，https://www.rakuten.co.jp/
[4] @cosme https://www.cosme.net/
[5] 矢部沙也加, 上田真由美, 中島伸介, ”評価項目別スコアを用いたコスメアイテム間の相違点可視化システム”, DEIM Forum 2021, C14-2, 2021.
[6] 住友千将, 石野拓也, 久保洸貴, 岳五一, ”口コミ情報に含まれる感情語に基づく類似スポット推薦システムの構築と実証実験”, パーソナルコンピュータ利用技術学会論文誌, 2020 年 14 巻 1 号 p.29-35, 2020.
[7] 鳥山実桜, 灘本明代, ”観光レビューから季節特有な情報と季節によって変化する情報の抽出手法の提案”, DEIM Forum 2022, B43-1, 2022.
[8] 中野広貴, 荒澤孔明, 渡邊稜平, 服部峻, ”観光写真から抽出した撮影者の好みに基づく観光スポット推薦”, IEICE Technical Report IN2018-80 (2019-01), 2019.
[9] 新井晃平, 新妻弘崇, 太田学, ”Twitter を利用した観光ルート推薦の一手法”, DEIM Forum 2015, G7-6, 2015.
[10] Laugwitz, B., Schrepp, M.　Held, T. (2008). Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, S. 63-76.

ソーシャルコンテクストを利用した アンサンブル学習によるフェイクニュースの早期検出

谷　聡馬† 佐々木裕多†† 張　建偉†

† 岩手大学大学院 総合科学研究科 理工学専攻　〒 020–8551 岩手県盛岡市上田 4-3-5
†† 東京工業大学 工学院 情報通信系　〒 226-8503 神奈川県横浜市緑区長津田町
E-mail: †{s0619039,zhang}@iwate-u.ac.jp, ††yubo1336@lr.pi.titech.ac.jp

あらまし　SNS の発達に伴い，フェイクニュースの拡散が問題となっている．フェイクニュースは，政治・経済をはじめとした様々な側面で社会に悪影響を及ぼす危険性があり，拡散初期段階での早期検出技術が求められている．最新の研究では，ニュースの拡散に関わるソーシャルコンテクストを特徴量として用いることによる検出精度の向上が確認されている．しかし，ソーシャルコンテクストの量はニュースの拡散に伴い増加するため，ソーシャルコンテクストを用いた精度向上とフェイクニュースの早期検出はトレードオフの関係にある．本研究では，検出精度の向上と早期検出の両立を図るため，ソーシャルコンテクストを利用した，アンサンブル学習によるフェイクニュース早期検出手法を提案する．モデルの学習時は，ニュースコンテンツとソーシャルコンテクストを入力とし，フェイクニュースであるか否かの二値ラベルについて学習を行う．推論時は，ニュース記事の拡散段階を想定し，ソーシャルコンテクストを時系列に沿って入力に加えることにより，アンサンブル学習によるフェイクニュースの検出を行う．また，検出閾値を任意に設定し，真偽の予測確率が閾値以上のニュースを追跡対象から取り除くことで，検出精度の向上と早期検出の両立を図る．

キーワード　テキストマイニング，言語モデル，フェイクニュース検出，ソーシャルコンテクスト，アンサンブル学習

1 はじめに

インターネットやスマートフォンの普及に伴い，ソーシャルメディアやソーシャルネットワークサービス（以下 SNS）の利用者が増加している．これらのプラットフォームは有益な情報源として利用可能な反面，悪意をもって作成された事実と異なる内容のニュース（以下フェイクニュース）の拡散によって，社会に悪影響を及ぼす原因にもなりうる [1] [2] [3]．ソーシャルメディアの利用者は，SNS 上にニュース記事の共有や投稿を行うことが可能である．これによりユーザー同士の議論が活発になることが，フェイクニュースの拡散を加速させる一因となっている [4]．2019 年 12 月以降の新型コロナウイルスの世界的な流行においては，新型コロナウイルスに関するフェイクニュースがソーシャルメディアや SNS を通して拡散された [5]．

フェイクニュースの判別には専門家によるファクトチェックが有効であるが [6]，ソーシャルメディアや SNS 上の無数のニュースのファクトチェックを全て人手で行うことは非常に困難である．また，フェイクニュースは真実のニュースよりも拡散の速度と規模が大きいという特徴を持つ [4]．そのため，フェイクニュースを自動的かつ早期に検出するための技術が求められ，研究が行われている [7] [8] [9]．

フェイクニュース検出に用いられる特徴量は，ニュースコンテンツとソーシャルコンテクストに大別される [10] [11]．ニュースコンテンツとは，ニュースのタイトルや本文など，ニュースそのものから得られる情報のことである．一方，ソーシャルコンテクストとは，ニュースの拡散経路や拡散に関わったユーザーの情報など，ニュースの拡散に関わる情報を指す．Raza らによる研究では，両方の特徴量を用いることによる精度の向上が確認されている [12]．しかし，十分なソーシャルコンテクスト特徴量を得るためには，ニュースの拡散を待つ必要があるため，検出の精度向上と早期検出のトレードオフが問題となる．

我々は，ニュースコンテンツとソーシャルコンテクストを特徴量として用いた，アンサンブル学習によるフェイクニュースの早期検出手法を提案する．モデルの学習時は，ニュースコンテンツとソーシャルコンテクストのうち，各モデルに適したデータを入力としてモデルごとの学習を行う．推論時は，ニュースの拡散段階を想定し，ソーシャルコンテクストを時系列に沿って入力に加え，フェイクニュースの検出を行う．出力ラベルの決定には，各モデルの予測確率の平均値を用いる．また，真偽の予測確率が閾値以上のニュースを追跡対象から取り除くことで，検出の精度と早期性の両立を図る．

2 関連研究

Zhou ら [4] は，フェイクニュースに関連する基礎理論について，心理学や社会学などの理論を含めた，分野を跨いだ調査を行った．また，複数の異なる観点によるフェイクニュース検出手法を評価し，それらの利点と改善点についてまとめた．さらに，フェイクニュースの早期発見やドメインを跨いだフェイクニュース検出などの，今後の研究課題について示した．

山中ら [13] は，X（旧 Twitter）[1]上のイベントに対して，イベントを構成するポストの特徴量を作成し，古典的な機械学習モデルを用いてイベントの真偽を時系列に沿って判別した．この研究では，モデルを 2 段階に分け，一定の条件を満たしたイベントの予測ラベルを断定することで，イベントの発生規模を考慮した偽情報検出の手法を提案した．本研究では，フェイクニュースを検出対象とし，ニュースコンテンツとソーシャルコンテクストから抽出される特徴量を用いる．また，複数のモデルを用いたアンサンブル学習により，真偽の判別を行う．

Raza ら [12] は，事前学習済み言語モデルである BART [14] を用いた，ニュースコンテンツとソーシャルコンテクストを利用したフェイクニュース検出手法を提案した．BART（Bidirectional Auto-Regressive Transformer）[14] は，事前学習済み Transformer ベースモデル [15] である．BERT（Bidirectional Encoder Representations from Transformers）[16] と GPT（Generative Pre-trained Transformer）[17] と組み合わせることで，seq2seq タスク（入出力ともにシーケンスのタスク）を可能にした．特徴量には，ニュースコンテンツとして NELA-GT-2019 [20] データセットを，ソーシャルコンテクストとして Fakeddit [21] データセットを利用した．また，転移学習を用いてユーザーの信頼性を推定する手法と，複数の特徴量を組み合わせることで弱ラベルからニュース記事のラベルを推定する手法について提案した．比較実験により，BART のエンコーダ部分とデコーダ部分，およびニュースコンテンツ特徴量とソーシャルコンテクスト特徴量がそれぞれ検出精度の向上に貢献していることを示した．また，提案手法によりニュース拡散初期段階から高い精度を実現できることを示した．本研究では，予測確率に基づいてニュースの真偽を断定することにより，フェイクニュースの早期検出を図る．

Mahabub ら [18] は，複数の機械学習モデルを用いたアンサンブル学習によるフェイクニュース検出について研究を行った．実験では各機械学習モデル単体によるフェイクニュースの判別を行った後，スコアが高い 3 つのモデルを使用したアンサンブル学習による判別スコアを算出した．各モデルのラベル予測確率の平均値を用いて最終的な出力ラベルを決定する Soft-Voting 手法と，各モデルの出力ラベルの多数決により最終的なラベルを決定する Hard-Voting 手法を比較した結果，どちらの手法も単一モデルによる判別スコアを上回り，Soft-Voting 手法が最も高いスコアとなることを示した．本研究では，Soft-Voting 手法を用いて，複数のモデルの予測確率の平均値を用いることで，最終的な出力ラベルを決定している．

3 提案手法

本研究では，特定のニュース記事がフェイクニュースであるか，真実のニュースであるかを判別する．提案手法の全体像を図 1 に示す．本手法において，i 番目のニュース記事を N_i，陽性の予測確率を $P(pos)$，陰性の予測確率を $P(neg)$，i 番目のニュースが陽性または陰性と予測された確率をそれぞれ $P(pos|N_i)$，$P(neg|N_i)$ とする．$P(pos|N_i)$，$P(neg|N_i)$ は $P(pos|N_i) + P(neg|N_i) = 1$ を常に満たす．また，陽性または陰性のラベルを断定するための予測確率の閾値を，それぞれ τ_{pos}，τ_{neg} とし，ニュースの拡散開始後，検出を行う時刻を $t_n (n = 0, 1, 2, \ldots)$ とする．

3.1 時系列に沿ったソーシャルコンテクストの利用

$t = t_0$ における検出はニュースの発行直後を想定するため，ニュースコンテンツのみを入力として使用する．以降の $t = t_1, t_2, \ldots$ における検出はニュースの拡散段階を想定するため，ソーシャルコンテクストを時系列に沿って入力に加える．t 時刻までの m 番目のソーシャルコンテクストを c_m とし，t 時刻までのソーシャルコンテクストを時系列順にソートしたものを，ソーシャルコンテクストの集合 $C_t = \{c_1, c_2, \ldots, c_m\}$ とする．また，t 時刻における入力を I_t，ニュースコンテンツを A とし，$I_t = \{A, C_t\}$ に対して各時刻において真偽を判別する．

3.2 予測確率に基づいたラベルの断定

真偽の判別には，複数のモデルによるアンサンブル学習を用いる．各ニュース記事のニュースコンテンツとソーシャルコンテクストから，各モデルごとに適した特徴量を作成し，入力とする．各モデルから出力された予測確率の平均値を算出し，検出時点における最終的な予測確率とする．この予測確率が閾値を超えたニュース記事は真偽のラベルを断定し，それ以降における判別を行わない．予測確率が閾値を下回ったニュース記事は，ラベルを断定せず，次の検出時点までのソーシャルコンテクストを入力に追加し，再度判別を行う．これにより一部のニュースについて拡散以前または拡散途中に真偽を断定することで，フェイクニュースの早期検出を図る．

また，τ_{pos} =0.9，τ_{neg} =0.9 と定めた場合のラベルの断定フローを図 2 に示す．図 1 において，予測確率が書かれた長方形は，t 時刻におけるニュース記事を表す．灰色は予測ラベルが断定済みであることを，白色は予測ラベルが断定されていないことを示す．また，実線は該当の t 時刻に真偽の判別を行うことを，破線は判別を行わないことを示す．

4 データセット

本研究では，FakeNewsNet[2] [19] をデータセットとして使用した．FakeNewsNet はニュースコンテンツとソーシャルコンテクスト両方を含む英語のデータセットである．ニュースコンテンツには，ファクトチェックサイトである Politifact[3]と Gossipcop[4]のニュース記事が含まれる．Politifact は政治に関わるニュース記事のファクトチェックを主に提供し，Gossipcop は芸能に関わるニュース記事のファクトチェックを主に提供している．ソーシャルコンテクストには，それらのニュース記事を共有した X への投稿のデータが含まれる．我々は，FakeNewsNet

1：https://www.twitter.com/

2：https://github.com/KaiDMML/FakeNewsNet
3：https://www.politifact.com
4：https://www.gossipcop.com

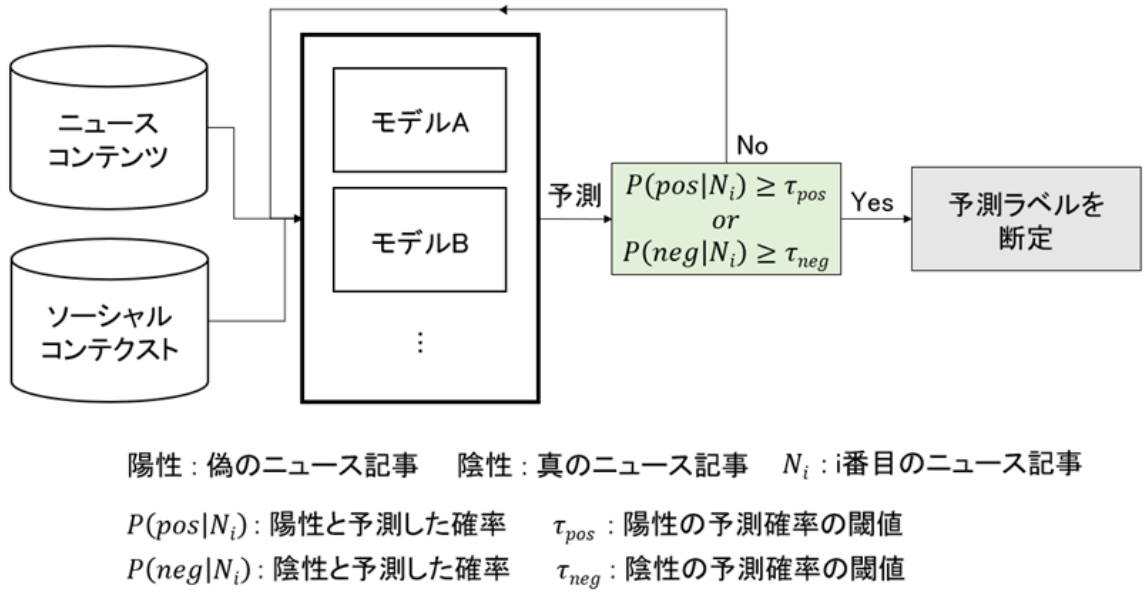


図 1　提案手法の全体像

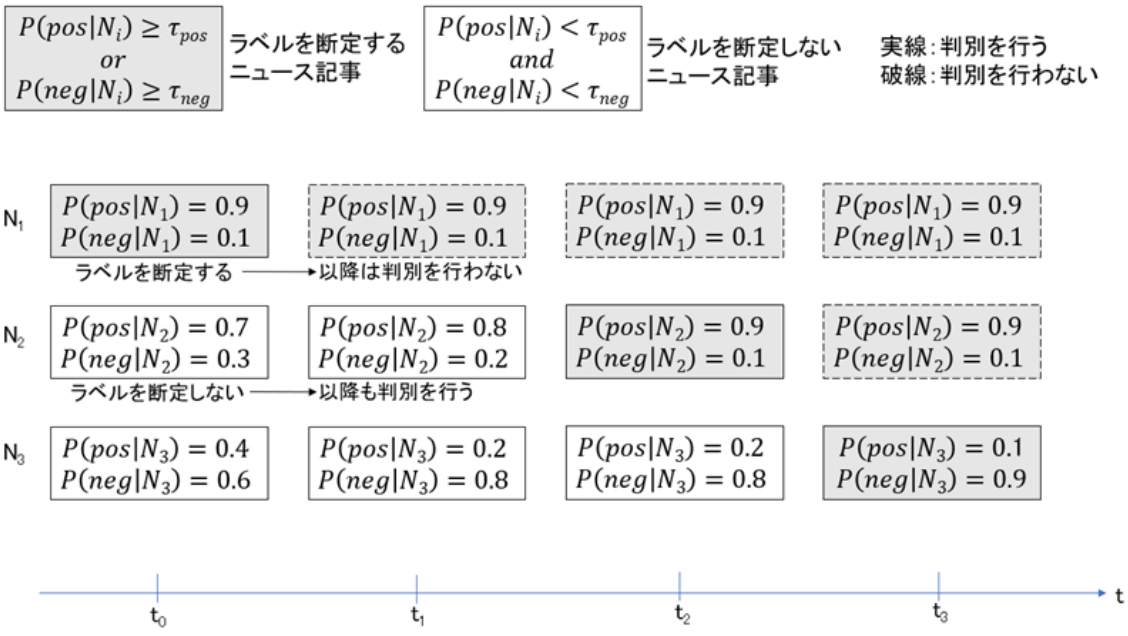


図 2　τ_{pos}=0.9，τ_{neg}=0.9 と定めた場合のラベルの断定フロー

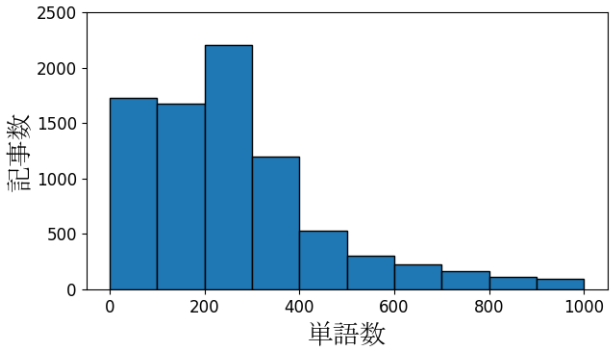


図 3　1 記事あたりの単語数

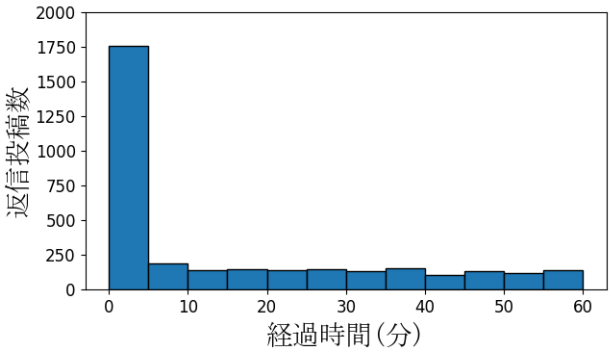


図 4　5 分ごとの返信投稿数

データセットによって提供されたデータと TwitterAPI[5]を用いて，ニュース記事と関連する投稿を収集した．収集したニュース記事と投稿データの統計情報を表 1 に示す．

表 1　収集したニュース記事と投稿データの統計情報

データセット	Politifact		Gossipcop	
ラベル	真	偽	真	偽
ニュース記事（件）	356	360	13,837	4,447
投稿データ（件）	286,150	103,112	675,426	374,326

4.1　ニュースコンテンツ

本研究では，ニュースコンテンツとして，ニュース記事のタイトルと本文を用いた．タイトルと本文を合算した，1 記事あたりの単語数を図 3 に示す．単語数の平均は 438 単語であり，8 割以上の記事が 500 単語以下で構成されていた．真偽どちらのラベルの記事においても，同様の分布が確認された．

4.2　ソーシャルコンテクスト

ニュース記事を共有した投稿は，X の仕様上ほとんどが同じテキストとなるため，ニュース記事を X 上に共有した投稿に対する返信（以下，返信投稿）のテキストを抽出し，ソーシャルコンテクストとして利用した．ニュース 1 記事あたりの返信投稿数の中央値は 4 件，返信 1 件あたりの単語数の平均値は 20 単語であった．データセット全体の 5 分毎の返信投稿数を図 4 に示す．図のとおり，最初の投稿から 5 分以内の時間に最も投稿が集中していることが確認された．

5　実　　験

5.1　実験設定

ニュース記事数に比べて返信投稿の数が少なく，投稿時刻のばらつきが大きいことから，時刻 t_{n-1} から t_n の間に 5 件の返信投稿が発生すると仮定し，時系列順に 5 件ずつ返信投稿のテキストを入力に付与することで，実験を行った．また，実験の第一段階として，アンサンブル学習を用いない単一のモデルによる実験を行った．提案手法において，ラベルを断定するための予測確率の閾値 τ_{pos}，τ_{neg} は，どちらも 0.9, 0.8, 0.7 を用いた．また，ベースラインは予測確率 0.5 を閾値として真偽の判別を行い，各検出時点において，全ニュース記事の判別をやり直すものとする．

単一モデルを用いた実験と，複数のモデルによるアンサンブル学習を用いた 2 パターンの実験を行う．アンサンブル学習の場合は，各モデルの予測確率の平均値を用いて出力ラベルを決定する．

5.2　データセット

Politifact と Gossipcop のニュース記事 9,612 件と，関連する 132,893 件の投稿データを利用した．使用したデータセットの詳細を表 5 に示す．訓練データとテストデータの割合は 8:2 である．ニュース記事に対する返信投稿の総数に偏りがあるため，データセットを返信投稿の数によって分類した後，8:2 の割合で訓練データとテストデータにランダムに分割を行った．データセットの分割による分類精度への影響を低減するために，10 通りの訓練データとテストデータの組み合わせを作成し，それぞれの組み合わせに対する評価指標の平均値を，最終的な評価指標の値とした．

5.3　モ　デ　ル

単一のモデルによる実験として，BART（facebook/bart-large-mnli[6]），DeBERTa（microsoft/deberta-v3-base [22][7]）を用いた．また，複数のモデルを使用したアンサンブル学習による実験も，同様の 2 つのモデルを用いた．共通に設定したパラメーターを表 6 に示す．

5.4　評価指標

評価指標には適合率，再現率，F1 を用いるが，ベースラインと提案手法では評価指標の算出に用いるサンプル数が異なる．ベースラインは各検出時点で全ニュース記事の判別をやり直すため，全ニュース記事数がサンプル数となる．一方，提案手法は予測確率が閾値を上回ったニュース記事のみラベルを断定し，それ以外のニュース記事についてラベルを断定しないため，該当時点で予測確率が閾値を上回ったニュース記事数がサ

5：https://help.twitter.com/ja/rules-and-policies/twitter-api

6：https://huggingface.co/facebook/bart-large-mnli
7：https://huggingface.co/microsoft/deberta-v3-base

表 2　各モデルを用いた場合の適合率

モデル＼返信投稿数	0	5	10	15	20	25	30	35	40	45	50
BART（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.969	0.968	0.969	0.97	0.969	0.97	0.97	0.971	0.971	0.971	0.971
BART（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	**0.975**	**0.975**	**0.975**	**0.976**	**0.976**	**0.976**	**0.976**	**0.976**	**0.976**	**0.976**	**0.976**
BART（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.97	0.971	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972	0.972
BART（Baseline）	0.886	0.887	0.885	0.884	0.884	0.883	0.883	0.883	0.883	0.883	0.883
DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	**0.981**	**0.982**	**0.983**	**0.984**	**0.984**	**0.984**	**0.984**	**0.984**	**0.984**	**0.984**	**0.984**
DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.978	0.979	0.979	0.98	0.98	0.982	0.982	0.982	0.982	0.982	0.982
DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.975	0.976	0.977	0.978	0.978	0.978	0.978	0.978	0.978	0.978	0.978
DeBERTa（Baseline）	0.891	0.894	0.895	0.898	0.898	0.899	0.899	0.899	0.899	0.899	0.899
BART+DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	**0.971**	**0.972**	**0.973**	**0.973**	**0.974**	**0.9747**	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**
BART+DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.968	0.969	0.969	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
BART+DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.958	0.961	0.961	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963
BART+DeBERTa（Baseline）	0.886	0.89	0.89	0.891	0.892	0.892	0.892	0.892	0.892	0.892	0.892

表 3　各モデルを用いた場合の再現率

モデル＼返信投稿数	0	5	10	15	20	25	30	35	40	45	50
BART（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.55	0.556	0.568	0.574	0.575	0.578	0.579	0.579	0.579	0.579	0.579
BART（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.574	0.581	0.582	0.583	0.584	0.588	0.588	0.589	0.59	0.591	0.591
BART（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.615	0.65	0.65	0.651	0.651	0.653	0.654	0.654	0.654	0.654	0.654
BART（Baseline）	**0.619**	**0.688**	**0.727**	**0.751**	**0.769**	**0.777**	**0.78**	**0.782**	**0.783**	**0.785**	**0.787**
DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.569	0.572	0.598	0.607	0.615	0.618	0.619	0.622	0.622	0.622	0.622
DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.608	0.609	0.63	0.641	0.658	0.665	0.665	0.665	0.665	0.665	0.665
DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.675	0.682	0.68	0.687	0.694	0.697	0.697	0.697	0.697	0.697	0.697
DeBERTa（Baseline）	**0.681**	**0.697**	**0.762**	**0.771**	**0.805**	**0.81**	**0.813**	**0.815**	**0.815**	**0.815**	**0.815**
BART+DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.581	0.595	0.604	0.614	0.615	0.617	0.635	0.635	0.635	0.635	0.635
BART+DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.628	0.665	0.67	0.684	0.691	0.695	0.695	0.695	0.695	0.695	0.695
BART+DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	0.697	0.7	0.707	0.713	0.717	0.723	0.726	0.726	0.726	0.726	0.726
BART+DeBERTa（Baseline）	**0.7**	**0.714**	**0.8**	**0.825**	**0.83**	**0.836**	**0.84**	**0.841**	**0.841**	**0.841**	**0.841**

表 4　各モデルを用いた場合の F1

モデル＼返信投稿数	0	5	10	15	20	25	30	35	40	45	50
BART（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.702	0.706	0.716	0.721	0.722	0.724	0.725	0.725	0.726	0.726	0.726
BART（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.723	0.728	0.729	0.73	0.731	0.734	0.734	0.735	0.735	0.735	0.735
BART（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	**0.753**	**0.78**	0.779	0.78	0.78	0.781	0.782	0.782	0.782	0.782	0.782
BART（Baseline）	0.729	0.775	**0.798**	**0.812**	**0.822**	**0.826**	**0.829**	**0.83**	**0.83**	**0.831**	**0.831**
DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.72	0.723	0.744	0.751	0.757	0.759	0.76	0.762	0.762	0.762	0.762
DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.75	0.751	0.767	0.775	0.787	0.792	0.792	0.792	0.792	0.792	0.792
DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	**0.798**	**0.803**	0.802	0.807	0.812	0.814	0.814	0.814	0.814	0.814	0.814
DeBERTa（Baseline）	0.772	0.783	**0.823**	**0.83**	**0.849**	**0.852**	**0.854**	**0.855**	**0.855**	**0.855**	**0.855**
BART+DeBERTa（$\tau_{pos}=0.9,\ \tau_{neg}=0.9$）	0.727	0.738	0.745	0.753	0.754	0.756	0.769	0.769	0.769	0.769	0.769
BART+DeBERTa（$\tau_{pos}=0.8,\ \tau_{neg}=0.8$）	0.762	0.789	0.792	0.802	0.807	0.81	0.81	0.81	0.81	0.81	0.81
BART+DeBERTa（$\tau_{pos}=0.7,\ \tau_{neg}=0.7$）	**0.807**	**0.81**	0.815	0.819	0.822	0.826	0.828	0.828	0.828	0.828	0.828
BART+DeBERTa（Baseline）	0.782	0.792	**0.843**	**0.857**	**0.86**	**0.863**	**0.865**	**0.866**	**0.866**	**0.866**	**0.866**

表 5　訓練データとテストデータの内訳

データセット	Politifact		Gossipcop		
ラベル	真	偽	真	偽	合計
訓練データ（リプライ有）	172	109	259	682	1,222
訓練データ（リプライ無）	116	175	3,298	2,874	7,685
テストデータ（リプライ有）	44	28	66	171	309
テストデータ（リプライ無）	74	72	890	891	1,927

表 6　モデルのハイパーパラメータ

最大シーケンス長	1024
学習率	1e-5
エポック数	2
バッチサイズ	2
損失関数	交差エントロピー誤差

ンプル数となる．

5.5 実験結果

実験では 10 通りのテストデータを用いてスコアを算出しているため，単一のモデルを判別に用いた場合の F1 の分散を図 5，図 6 に示す．図が示すとおり，分散の値が 0.01 未満であることから，それぞれのモデルとテストデータを用いた実験における F1 のばらつきが低いことが確認された．また，単一のモデルを用いた実験における適合率，再現率，F1 を表 2，3，4，図 7，8，9，10，11，12 に示す．両方のモデルにおいて，提案手法はベースラインと比較して適合率が高いことが確認された．これは，提案手法が予測確率の値が高いニュース記事のみラベルを断定するため，陽性（偽のニュース記事）と判断した場合の正解率が高いためであると考えられる．一方で，提案手法はベースラインと比較して再現率が低いことが確認された．これは，提案手法においてラベルを断定するニュース記事の数が少なくなるためであると考えられる．続いて，BART と DeBERTa の予測確率の平均値を利用したアンサンブル学習手法を BART+DeBERTa と表記し，実験における適合率，再現率，F1 を表 2，3，4，図 13，14，15 に示す．単一のモデルによる判別と比較して，BART と DeBERTa の予測確率の平均値を利用したアンサンブル学習手法が全体的に最も高い F1 を示した．これは，2 つのモデルの予測確率の平均値を用いることで，再現率が増加，適合率が減少し，これらの差が小さくなったことが理由の一つであると考えられる．F1 は，返信投稿数 0 件から 5 件までの検出時点において，閾値 0.7 を用いた提案手法がベースラインのスコアを上回った．このことから，閾値 0.7 を用いた提案手法が高い精度での早期検出に貢献することが確認された．

6　まとめと今後の展望

本研究では，検出精度の向上と早期検出の両立を図った，ソーシャルコンテクストを利用したアンサンブル学習によるフェイクニュース早期検出手法を提案した．実験において，閾値 0.7 を用いた場合に返信投稿数 0 件から 5 件までの検出時点における F1 がベースラインよりも高いことから，提案手法がフェイクニュースの早期検出に貢献することが確認された．また，比較実験により，複数のモデルを用いたアンサンブル学習による検出の精度向上が確認された．

今後は，実験に使用するモデルを更に増やすことで，アンサンブル学習の効果についてより詳しい検証を行う．また，複数のモデルを用いることによる検出コストの増加が懸念されることから，検出精度の向上と検出にかかるコストのトレードオフについて検証を行う．

謝　辞

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

文　献

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. *KDD*, 2017.
[2] Soroush Vosoughi, Deb Roy and Sinan Aral. The spread of true and false news online. *Science*, 2018.
[3] Reza Zafarani, Xinyi Zhou, Kai Shu and Huan Liu. Fake News Research: Theories, Detection Strategies, and Open Problems. *KDD*, 2019.
[4] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 2020.
[5] Konstantin Pogorelov, Daniel Thilo Schroeder, Stefan Brenner, and Johannes Langguth. Fakenews: Corona virus and conspiracies multimedia analysis task at mediaeval 2021. *MediaEval*, 2021.
[6] Jiawei Zhang, Bowen Dong and Philip S. Yu. FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network. *ICDE*, 2020.
[7] Changhe Song, Cunchao Tu, Cheng Yang, Zhiyuan Liu, Maosong Sun. CED: Credible Early Detection of Social Media Rumors. *TKDE*, 2020.
[8] Xinyi Zhou, Jindi Wu and Reza Zafarani. SAFE: Similarity-Aware Multi-Modal Fake News Detection. *PAKDD*, 2020.
[9] Xinyi Zhou, Reza Zafarani, Emilio Ferrara. From Fake News to #FakeNews: Mining Direct and Indirect Relationships among Hashtags for Fake News Detection. arXiv preprint, 2022.
[10] Xinyi Zhou, Atishay Jain, Vir V. Phoha and Reza Zafarani. Fake News Early Detection: A Theory-driven Model. *ACM Digital Threats: Res. Pract.*, 2019.
[11] Xinyi Zhou, Reza Zafarani. Network-based Fake News Detection: A Pattern-driven Approach. *KDD*, 2019.
[12] Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *Int. J. Data. Sci. Anal.*, 2022.
[13] 山中仁斗， 張建偉. 発生規模と時系列を考慮した twitter イベントにおける偽情報の早期自動検出. 第 12 回データ工学と情報マネジメントに関するフォーラム（*DEIM 2020*）, 2020.
[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL*, 2020.
[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.
[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HCT*, 2019.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. OpenAI Blog, 2018.
[18] Mahabub, A. A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. *SN Appl. Sci. 2*, 2020.
[19] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Big Data*, 2018.
[20] Maurício Gruppi, Benjamin D. Horne and Sibel Adalı. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. arXiv preprint, 2020.
[21] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: a new multi-modal benchmark dataset for fine-grained fake news detection. *LREC*, 2019.
[22] Pengcheng He, Jianfeng Gao, Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *ICLR*, 2023.

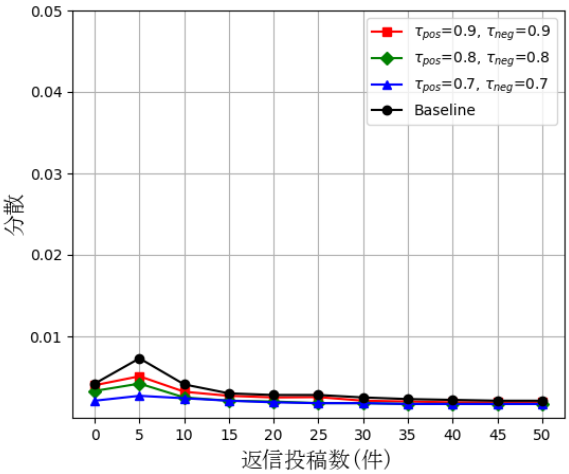


図 5　BART を用いた実験による F1 の分散

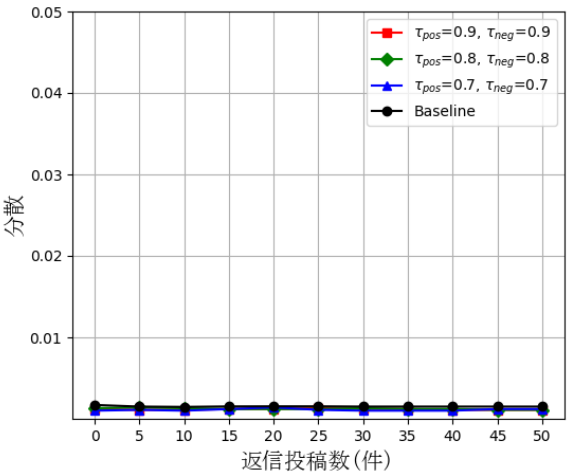


図 6　DeBERTa を用いた実験による F1 の分散

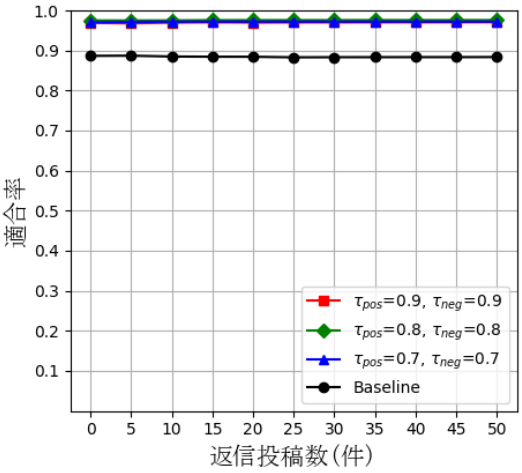


図 7　BART を用いた実験による適合率

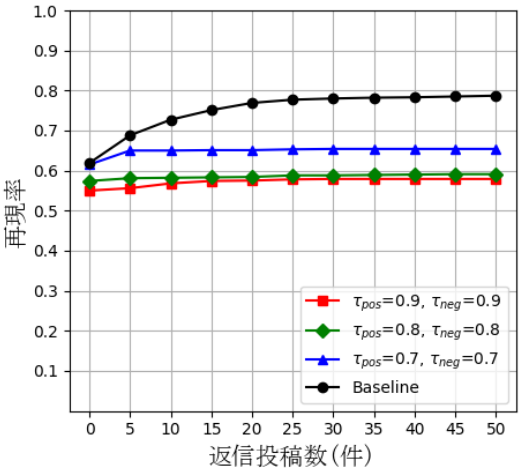


図 8　BART を用いた実験による再現率

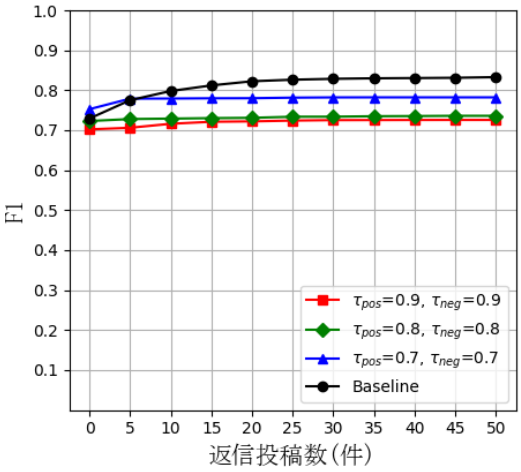


図 9　BART を用いた実験による F1

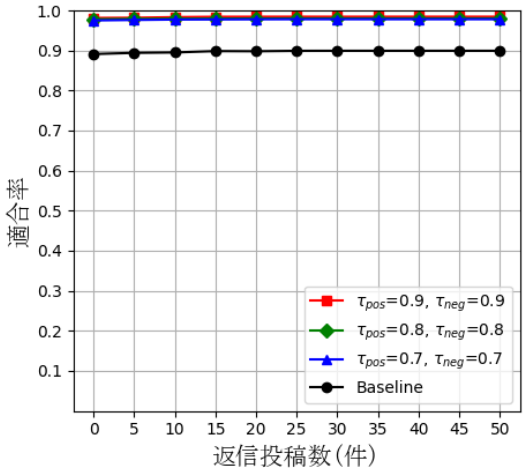


図 10　DeBERTa を用いた実験による適合率

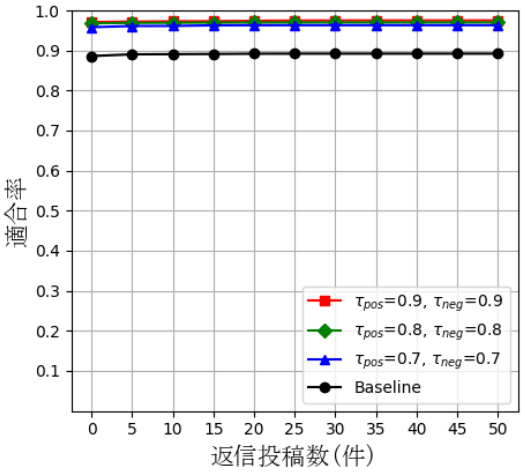


図 13　BART+DeBERTa を用いた実験による適合率

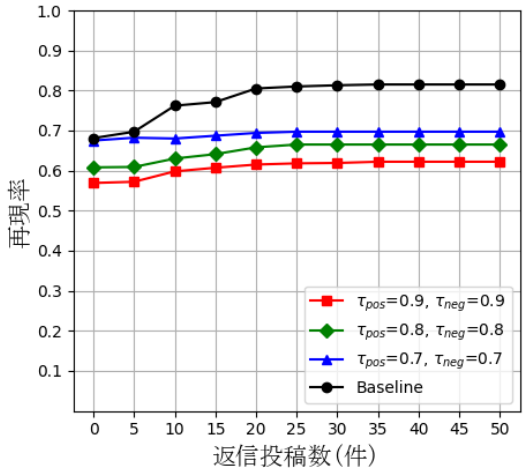


図 11　DeBERTa を用いた実験による再現率

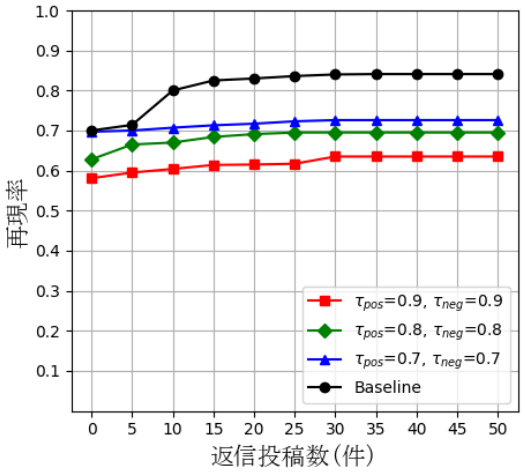


図 14　BART+DeBERTa を用いた実験による再現率

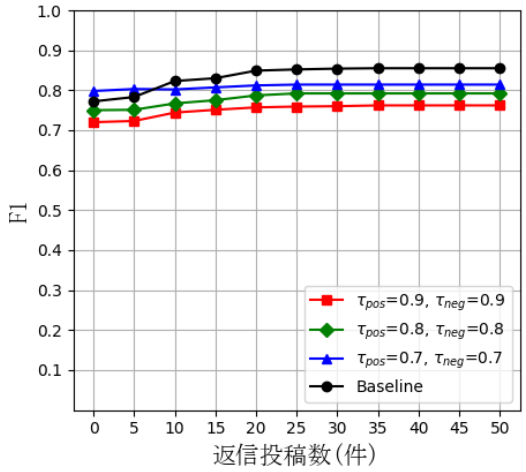


図 12　DeBERTa を用いた実験による F1

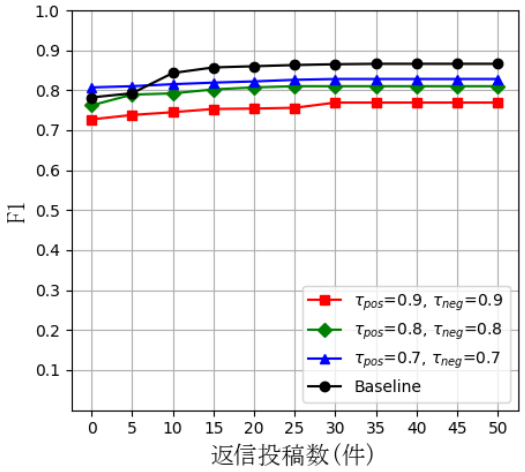


図 15　BART+DeBERTa を用いた実験による F1

GPTを活用した, ユーザの発言・行動履歴を参考に抽出したイベント推薦の評価

大本　詩織† 　榎　美紀†† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2 丁目 1-1
†† 日本アイ・ビー・エム株式会社　〒 103–8510 東京都中央区日本橋箱崎町 19-21
E-mail: †shiori-o@ogl.is.ocha.ac.jp, ††enomiki@jp.ibm.com, †††Oguchi@is.ocha.ac.jp

あらまし　X(旧 Twitter) などのソーシャルネットワーキングサービス (以下 SNS とする) では, 多くの人々が手軽に情報を発信しているため, 固定的なメディアに掲載されていないような様々な有益な情報が存在する. 例えば, ローカルイベントや開催中のイベントをはじめ, 大小様々な規模のイベントに関する情報が投稿されている. また, SNS の投稿から, そのユーザについての情報を得ることもできる. そこで, SNS 上の膨大なイベント情報をサーバ側に収集した上で, ユーザの SNS の投稿や検索履歴などからユーザのパーソナリティを推測し, 適切なイベントを推薦するシステムを提案する. 本研究では, 近年飛躍的な進歩を遂げている大規模言語モデルの中でも, 特に ChatGPT を使用することで, マルチモーダルな SNS データから趣味趣向の推測を行い, 加えてイベント推薦や回答精度の向上にも活用することで, 効率的なシステムの構築を行うことができるか検証する.

キーワード　LLM, 情報抽出, ソーシャルメディア, 推薦モデル

1 はじめに

現在, SNS が広く普及しており, 総務省の調査 [1] によると, 国内全体では約 7 割もの人が SNS を利用している. SNS を使用して, 多くの人々が手軽に情報を発信しているため, SNS 上には様々な有益な情報が存在する. 例えば, 固定的なメディアに掲載されていないようなローカルイベントや開催中のイベントをはじめ, 大小様々な規模のイベントに関する情報が投稿されている. また, SNS では投稿からそのユーザについての情報も得ることができる. 近年, データの収集や分析技術は飛躍的な進歩を遂げており, 様々な情報源から膨大なデータを処理および分析することが可能になっている. これらを活かし, 膨大なイベント情報をサーバ側に収集し, 問合せユーザの SNS データや位置情報などを利用することで, ユーザに対し特定の場所や時間で開催されるイベントを推薦するシステムは既に研究されている.

そこで本研究では, 近年飛躍的な進歩を遂げている大規模言語モデルの内の一つである ChatGPT を用いて, ユーザの SNS の投稿や検索履歴などから趣味趣向を推測し, 適切なイベントを推薦するシステムを提案する. ここでは SNS の一種である X からイベント情報を取得した上で, ChatGPT を使用してマルチモーダルな SNS データからユーザのパーソナリティの推測を行い, 加えてイベント推薦や回答精度の向上にも活用することで, 効率的なシステムの構築を行うことができるか検証する.

本論文の構成は以下の通りである. 2 章で先行研究について述べ, 3 章で提案システムの概要について述べる. 4 章でユーザの趣味趣向の取得について, 5 章で趣味趣向に基づいたイベントの推薦について述べ, 6 章で回答精度の向上について述べる. 最後に, 7 章で本稿をまとめる.

2 先行研究

本章では, 先行研究について紹介する. 同研究室工藤らが, SNS データを用いた場所と時間を考慮するイベント検索手法の提案と評価 [2] という題で, X の投稿データからのイベントに関する情報の抽出に関する研究を行っている. 工藤らの提案システムは以下のような流れになっている.

（1） Twitter API のキーワード検索機能を使用して, ツイート本文にイベントの開催地, 日付, 時間が含まれるツイートを抽出し, 正規表現を用いてイベント名を取得, 外部情報を利用しイベント情報を補完

（2） ランダムフォレストを用いてイベントを Music Event, Comedy などのカテゴリに分類

（3） イベント名, 開催日時, カテゴリ等のユーザに提供する情報を取得し整理した上で, ユーザの位置を取得し, 提供する情報を順位付け

また，同研究室今井らが, 訪日外国人の趣向に合わせたイベント情報提供のためのソーシャルメディア活用の検討 [3] という題で, 各ユーザの過去のツイートデータから趣味趣向を判別し, 工藤らのシステムを用いて収集したイベント情報のデータを, 取得したユーザ趣向, ユーザの位置情報に基づき, 順位付けし推薦するシステムを提案している.

本研究では，サーバ側が行う X からのイベント情報抽出に, 先行研究の手法を活用した上で, ChatGPT を使用したより効率の良いイベント推薦システムの構築を試みる. さらに, 先行研究では主に X から取得したユーザ情報に基づくイベント推薦を行っていたのに対し, 本研究では, ChatGPT を使用するこ

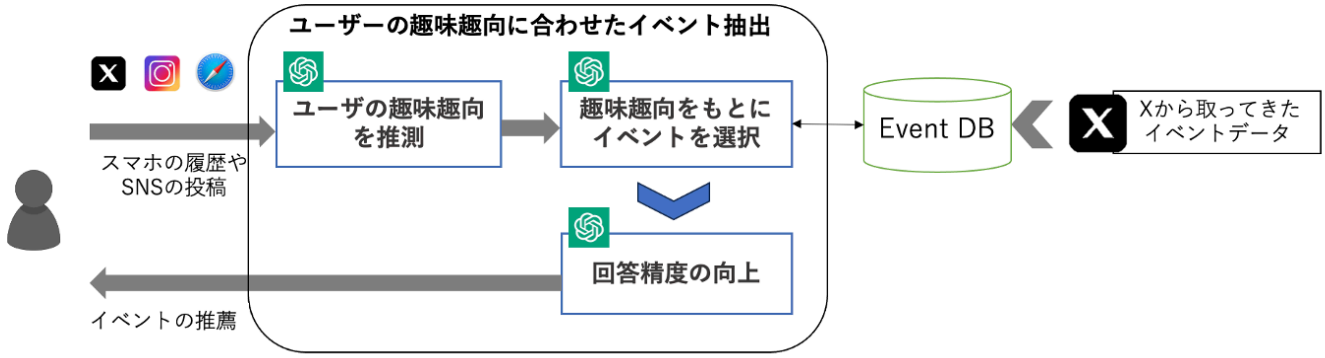


図 1: 提案システムの概要

とで, Instagram や検索履歴なども含めた, マルチモーダルな SNS データを用いたイベント推薦システムへの拡張を目指す.

3 提案システムの概要

本論文では, ユーザのパーソナリティとして, 趣味趣向に基づくイベント推薦の検証を行う. 提案するイベント推薦システムの概要を図 1 に示す. 本システムでは, ChatGPT を, ユーザの趣味趣向の推測, 趣味趣向を元にしたイベント推薦, 回答精度の向上部分で使用する. 具体的には, まず, X, Instagram といった SNS の投稿や, スマホの検索履歴など, 本人のデバイスから得られる情報を ChatGPT に与えて, ユーザの趣味趣向を推測させる. 次に, 工藤ら [2] が提案する手法を参考にし, X から取ってきたイベントデータのリストを与えて, その中から先ほど取得した趣味趣向をもとにイベントを選択させる. 最後に回答精度の向上を行い, ユーザに適切だと考えられるイベントを推薦結果として回答させる.

4 ユーザの趣味趣向の取得

4.1 様々な情報源からのユーザの趣味趣向抽出の比較

X, Instagram, 検索履歴といった様々な情報源のデータを与え, それぞれから取得できるユーザの趣味趣向にどのような違いが現れるか検証を行った.

4.1.1 X からの趣味趣向の抽出

まず, X から趣味趣向を抽出する場合について検証を行った. X からツイート数が十分であるユーザを無作為に複数選び, それぞれのユーザに対して, 1 ヶ月分のツイート内容と,「投稿内容からユーザの趣味趣向を推測してください」という指示を与えた. その結果の一例を図 2 に示す. 例えば, 野球の試合や甲子園について投稿しているユーザの場合,「スポーツ：特に野球に関心がある」といった回答が得られ, ポストを投稿しているユーザの趣味趣向として適切だと思われる回答を得ることができた. しかしながら一方で, ポストで少し触れただけの内容も趣味趣向として挙げられてしまうといった問題も見られた. 例えば,「スマホ買い替え時かな。iPad も欲しいな」という投稿から,「スマートフォンや iPad について触れており、デジタルデバイスに興味があることがわかります」という回答が返ってくるなど, 実際の本人の趣味趣向とずれた内容が得られる可能性があることが分かった.

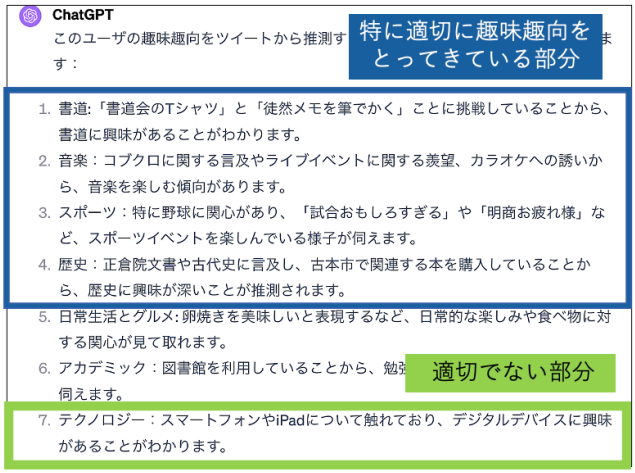


図 2: X から趣味趣向を取得した結果

4.1.2 Instagram からの趣味趣向の抽出

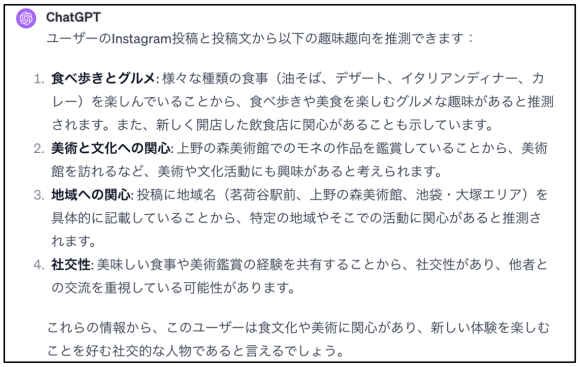
次に, Instagram の場合の検証を行った. こちらも無作為にユーザを選び, 投稿されている画像を何枚かと, それぞれに対応する投稿文を与えて, ユーザの趣味趣向を推測するよう指示を与えた. 与えたプロンプトの一例と, その結果を図 3 に示す. 図 3(b) を見ると, 比較的適切に趣味趣向が推測できていることがわかる. ただし, Instagram を趣味趣向の取得に使用する問題点として, ほとんどのユーザが旅行や外出などのイベントや人物, 食べ物に関する投稿を行なっており, そもそもの投稿内容が偏る傾向にあるため, そこから得られる趣味趣向にも偏りが見られた.

4.1.3 スマホのブラウザ検索履歴からの趣味趣向の抽出

さらに, スマホの検索履歴を与えた場合の検証も行った. statcounter による統計 [4] によると, 日本のスマホの web ブラウザのシェアは, safari の割合が最も高くなっているため, ここでは safari の履歴を与えることにする. SQL コマンドを使用し



(a) 与えたプロンプト

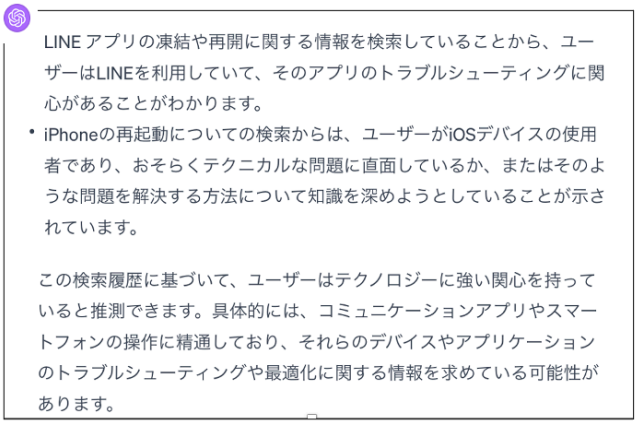


(b) 出力結果

図 3: Instagram から趣味趣向を取得した結果

	title	url
:05:43	iPhoneを強制的に再起動する - Apple サポート (日本)	https:/
:05:27	iphone 再起動 - Google 検索	https:/
:05:27	iphone 再起動 - Google 検索	https:/
:05:26	line フリーズ - Google 検索	https:/
:05:20	line フリーズ - Google 検索	https:/
:05:08	【iOS17対応】LINEがフリーズした時の原因と解決方法	https:/
:05:04	line フリーズ - Google 検索	https:/
:05:03	line フリーズ - Google 検索	https:/
:05:03	line 固まった - Google 検索	https:/
:05:00	line 固まった - Google 検索	https:/
:04:55	LINEが緑の画面から動かず、フリーズした時の原因と対処法	https:/
:04:50	line 固まった - Google 検索	https:/
:04:50	line 固まった - Google 検索	https:/

(a) 与えた検索履歴の一部



(b) 検索履歴を元に趣味趣向を推測させた結果

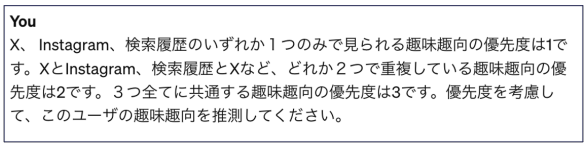
図 4: 不適切な趣味趣向が得られた場合

て, Safari の履歴を ~/Library/Safari/History.db から CSV 形式で取得し, プロンプトに与えてユーザの趣味趣向を推測させた. その結果, 検索履歴によっては適切に趣味趣向を取得することができた. これは, ユーザが多くの場合, 興味や疑問などに基づいて検索を行うためだと考えられる. しかしながら 図 4 のように, 例えば line フリーズ, iphone 再起動といった, 趣味趣向に関係のない検索履歴を与えた場合に, テクノロジーに関心があるといった回答結果が得られた. このように, 検索履歴によってはユーザの趣味趣向として相応しくない回答が返ってくるため, 有用な検索履歴を与える必要があるということが分かった.

以上のように, いずれの情報源を選択した場合でも, それぞれ問題点が見られた.

4.2 同一のユーザに対する趣味趣向の比較

今度は, 同一のユーザの X, Instagram, 検索履歴の情報を与えてそれぞれ趣味趣向を取得した. その結果, 同じユーザであるにもかかわらず, 得られる趣味趣向に違いが見られた. 一方で, 一部共通して見られる趣味趣向もあった. このような, 重複している趣味趣向はより信憑性が高く, 適切な趣味趣向である可能性が高いと考えられる. そこで, 図 5(a) のように, X, Instagram, 検索履歴のいずれか 1 つのみで見られる趣味趣向の優先度を 1, 2 つで重複している趣味趣向の優先度を 2, 3 つ全てで共通して見られる趣味趣向の優先度を 3 として, 共通する情報源が多いほど優先度を高く回答するよう指示を与えたところ, 図 5(b) のような出力結果が得られ, 実際のユーザの趣味趣向と比較すると, より高い精度でユーザの趣味趣向を取得することができていることが分かった.



(a) 与えたプロンプト



(b) 出力結果

図 5: 共通する趣味趣向の優先度を考慮した場合

5　趣味趣向に基づいたイベントの推薦

4 で得た趣味趣向をもとにイベント推薦の検証を行った．ここではあらかじめ X からイベント名をとってきてリストとしてまとめたデータ (図 6) とユーザの趣味趣向をプロンプトに与えた上で，「このような趣味趣向を持つユーザに適切だと考えられるイベントを与えたリストの中から選択して回答してください」という指示を与えた．イベントデータは，2021 年の 11/20 から 11/24 において，原宿，渋谷，新宿，赤坂，代々木，池袋，六本木の 7 地点のキーワードで検索して得られたイベントのリストを使用した．結果，例えば音楽が好きだという趣味趣向を与えた場合には音楽に関連するイベントが推薦されるといった風に，うまくイベント推薦を行うことができた．

イベント名	開催地
生業	国立代々木競技場
HiHi Jets Concert 2021~五騎当千~	国立代々木競技場
#五騎当千	国立代々木競技場
ayumi hamasaki COUNTDOWN LIVE 2007-2008 Anniversary	国立代々木競技場
J-PHONE presents GLAY ARENA TOUR 2000 “HEAVY GAUGE”	国立代々木競技場
FIBREPLEX Presents TOKYO SUPER WARS ~東京超女大戦~	国立代々木競技場
青春病	国立代々木競技場
15TH ANNIVERSARY LIVE KAT-TUN	国立代々木競技場
GLAY ARENA TOUR 2004 “THE FRUSTRATED -extreme-”	国立代々木競技場
ayumi hamasaki ASIA TOUR 2021-2022 A ~23rd Monster~	国立代々木競技場
ayumi hamasaki count down live 2000-2001 A	国立代々木競技場
代々木公演	代々木公園
絢香 15th Anniversary Tour 2021 ~もっといい日に~	国立代々木競技場
泉屋博古館東京	泉屋博古館
今だけ学生・子供ワンコイン	森美術館
庵野秀明展	国立新美術館
#ミラベルと魔法だらけの家	六本木ヒルズ
昨日より赤く明日より青く—CINEMA FIGHTERs project—	TOHOシネマズ 六本木ヒルズ
映画ドラえもん のび太の宇宙小戦争	六本木ヒルズ
昨日より赤く明日より青く-CINEMA FIGHTERS project-	六本木ヒルズ
メトロポリタン美術館展 西洋絵画の500年	国立新美術館
Experience Bar	六本木ヒルズ
夕方のおともだち	六本木ヒルズ
ルール?展	21_21 DESIGN SIGHT
M・A・D朗毒	シアターサンモール

図 6: 与えたイベントデータの一部

ただし，何回か試す中で，ハルシネーションが発生するという問題が生じた．ハルシネーションとは，LLM の出力が現実の事実やユーザ入力と矛盾する現象のことをいう．具体的には，ChatGPT が，イベント内容がわかりづらいイベント名の場合は，実際とは全く異なる内容をでっち上げて推薦する場合があった．これは，ChatGPT が与えられたイベント名からイベントの内容を推測して推薦しているためだと考えられる．一例を図 7 に示す．図 7 の枠線で囲った部分に注目すると，「Experience Bar」というイベントが，「美容とケア，特にアイケアに関心がある方には，このバーイベントでの美容テクノロジー体験が新しい発見につながるかもしれません」という理由で推薦されている．しかしながら，Experience Bar について調べてみたところ，実際には健康と野菜を中心にした体験ができるイベントであり，美容テクノロジー体験とは何ら関係がないイベントであることが分かった．また，他にも，X の投稿から，正規表現を用いてイベント名を取得し，図 6 のようなイベントデータのリストを作成する上で，実際には存在しないイベント名を抽出してしまい，それがそのまま推薦されてしまう場合もあった．

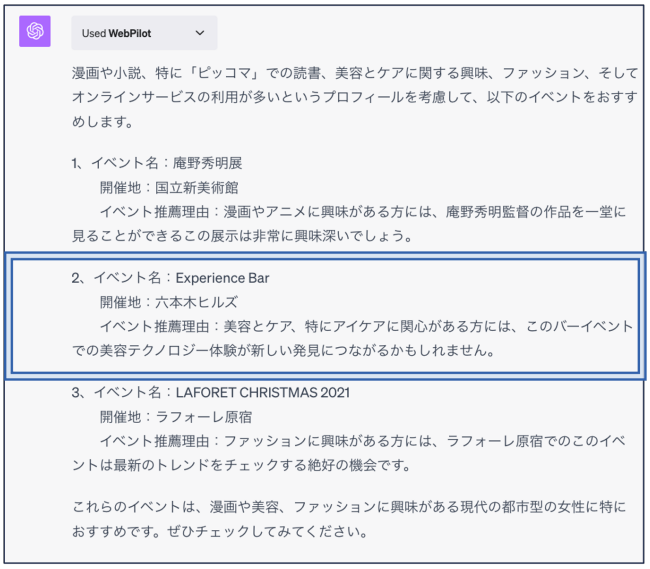


図 7: ChatGPT がイベント内容をでっちあげる例

6　回答精度の向上

Ayush Agrawal ら [5] は，ハルシネーションを減らすには，テキスト生成プロセスの初期段階で前処理を行うか，もしくはテキスト生成プロセス後に生成したテキストをレビューし，ハルシネーションを識別して修正するためのステップを導入することが有効だと述べている．そこで，5 で生じた問題に対処するため，web ブラウジング機能を使い，後処理を導入することで回答精度の向上を試みた．web ブラウジング機能とは，ChatGPT がリアルタイムでウェブ検索を行い，最新の情報や特定のトピックに関するデータを取得できる機能である．回答精度の向上を試みた際のプロンプトと，それに対する回答を図 8 に示す．図 8 の左図にあるように，5 で得た回答結果を，改めてプロンプトに与え，

「これらのそれぞれについて，以下を実行してください．
step1: 実際にイベントが存在するか調べる．
step2: イベント推薦理由を見て，そのイベントの実際の内容とあっているか調べる」

という指示を追加した．その結果，図 8 の右図のように，Experience bar のイベント内容が適切に調べられており，本イベントは健康と野菜を中心にした体験を提供する場所であり，美容テクノロジーに関連する情報が見つからなかったと回答されていることから，回答精度を向上することができるということが分かった．

7　まとめと今後の課題

本研究では，SNS データや検索履歴などのマルチモーダルなデータからユーザの趣味趣向を取得し，イベントを推薦するシステムにおける，ChatGPT の活用の可能性について検討した．ユーザの趣味趣向を取得する際，それぞれの情報源に問題点が

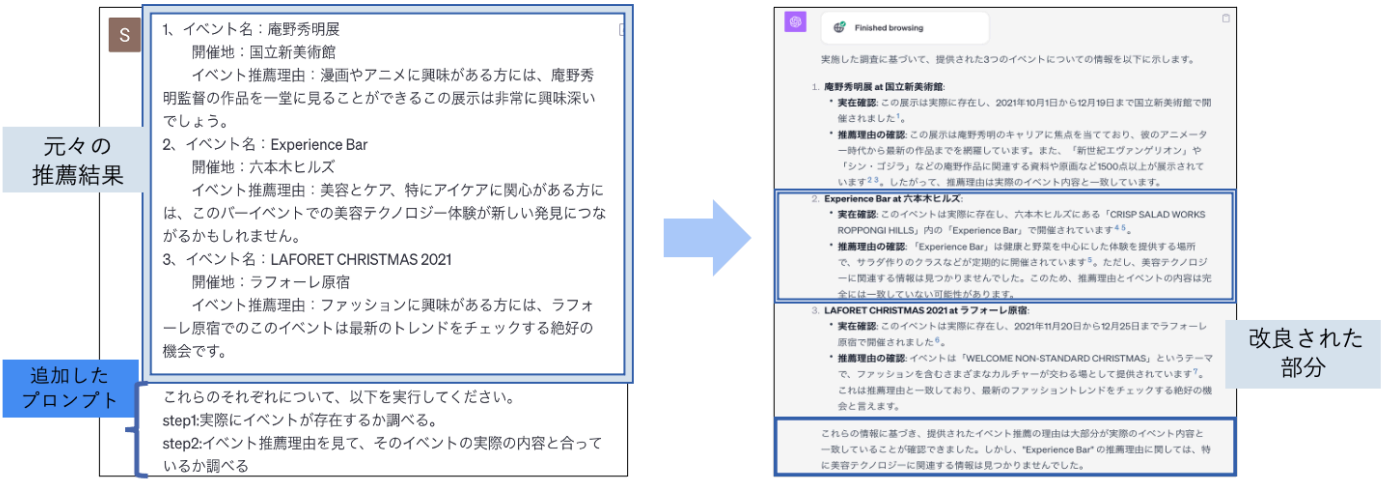


図 8: 回答精度の向上を試みた際のプロンプト (左) とそれに対する回答 (右)

見られたが, それらの情報源を組み合わせることでより適切に趣味趣向を取得することができることが分かった. また, web ブラウジング機能を使ってイベント推薦の結果を改善することができた. 今後は, ChatGPT が趣味趣向とはいえないものまで趣味趣向として回答してしまうなどといった問題に対処することで, より推薦の精度を高めていきたい. 加えて, 本研究では趣味趣向に焦点を置いたイベント推薦の検証を行ったが, 今後はユーザの性格特性や感情といったユーザのパーソナリティに注目したイベント推薦へ発展させたいと考えている.

また, ここではユーザの趣味趣向の取得, イベント推薦, 回答精度の向上の 3 パートに分けて, それぞれの内容について手動でプロンプトに指示を与えることで検証を行ったが, 今後はこれら一連の流れのシステム化も行なっていきたい.

文　　献

[1] 国内外における最新の情報通信技術の研究開発及びデジタル活用の動向に関する調査研究 (総務省). https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r04/html/nf308000.htmld0308130

[2] 工藤瑠璃子, 榎美紀, 中尾彰宏, 山本周, 山口実靖,and 小口正人. SNS データを用いた場所と時間を考慮するイベント検索手法の提案と評価. マルチメディア, 分散, 協調とモバイル (DICOMO2018) シンポジウム, pages 1658–1664, 2018.

[3] 今井美希, 工藤瑠璃子, 榎美紀, 小口正人. 訪日外国人の趣向に合わせたイベント情報提供のためのソーシャルメディア活用の検討. DEIM Forum 2019.

[4] Mobile Browser Market Share Japan(statcounter). Dec 2021 - Dec 2022https://gs.statcounter.com/browser-market-share/mobile/japan/monthly-202112-202212-bar

[5] Ayush Agrawal, Mirac Suzgun, Lester Mackey, Adam Tauman Kalai: Do Language Models Know When They're Hallucinating References?, arXiv.org, arXiv:2305.18248,May 2023

効果的な行動変容メッセージ実現に向けた インセンティブ効果の被験者評価

瀧本　祥章† 金　秀明† 小林のぞみ† 倉島　健†

† NTT 人間情報研究所　〒 239–0847 神奈川県横須賀市光の丘 1-1
E-mail: †{yoshiaki.takimoto,hideaki.kin,nozomi.kobayashi,takeshi.kurashima}@ntt.com

あらまし　生活習慣病などの一次予防のため，ナッジを考慮したメッセージにより健康行動を促す取り組みが注目されている．介入の効果を高めるためには，メッセージ自身が持つ特徴量のみでなく受け手の特性も考慮する必要があることが知られている．しかし従来の研究では，研究者が用意した少数のメッセージパターンを用いた評価，または，少数の受け手による評価が主であり，メッセージの特徴量と受け手の特性の影響，特にそれらの交互作用による影響はまだ明らかになっていない．そこで本研究では，交互作用を考慮しつつ，メッセージの有効性を裏付ける特徴を特定するため，多様なメッセージに対し，特性が異なる被験者による評価を実施し，検証用データセットを構築した．データセットは 630 件のメッセージと各メッセージに対する 1 件あたり 586 人，のべ 73,841 人の評価によって構成される．本研究では，特にナッジの一種であるインセンティブの記載がメッセージの効果に与える影響について分析を行った．その結果，インセンティブの記載が男女ともに殆どの年代の受け手に有効であり，特に 50 代の受け手に対してはインセンティブが得られる根拠を示すことが有効であることが明らかになった．

キーワード　行動変容，ナッジ，Twitter，ヘルスケア

1　はじめに

偏った食生活や運動不足，喫煙，過度なストレスなどを原因とする生活習慣病の患者が増加傾向にある [1, 2]．この傾向がもたらす医療費の増大が大きな社会課題となっている．その対処のため，人々の健康を増進し，発病を予防する対策が推奨されている [3]．一次予防の実現のためには，バランスのよい食生活や日々の運動など，個々人が自ら日頃の行動を改善し，習慣づけることが必要である．しかし，個々人が主体的に行動を改善，維持することは必ずしも容易ではない．そのため，行動変容を支援するシステムが求められている．

行動変容を支援するための手段としてメッセージの提示が検討されている [4–6]．特に，その効果を高める手法としてナッジが着目されている．ナッジとは，人の意思決定を自然に誘導し，望ましい選択を自発的にできるように支援する手法であり，災害時の早期避難 [7] やフードロスの削減 [8] などに用いられている．メッセージによる行動変容の文脈においては，フレーミング効果 [9] に着目した調査 [10] や，ナッジのフレームワークとして知られる MINDSPACE[11] に着目した検証 [12] が行われている．また，行動変容を促すメッセージの受容性（行動に対し前向きになるか，実際の行動に繋がるか）はメッセージの受け手の性別や年齢，時間選好などの特性，行動変容ステージによって異なることが指摘されている [6, 13, 14]．そのため，個々人が主体的に行動を改善する支援を行うには，メッセージ自身がもつ特徴と受け手がもつ特性の双方の組み合わせを考慮する必要がある．

これらの既存研究の多くは，研究者が着目した少数の特徴に基づいて作成された文例と，少数の受け手による評価によってメッセージの受容性を検証している [10, 13–16]．すなわち，得られた知見は統制された少数の文例や受け手の特性においてのみ成り立つ限定的なものである可能性がある．実際に使用するメッセージを検討する際には，推奨する行動や通知するタイミング，媒体，受け手の特性などについて多様な状況が想定されるため，上記のような限定的な知見をそのまま適用することは避けるべきである．

限定的な知見を一般化，すなわち，想定していない状況に対しし適用した場合のリスクについて考える．例えば，運動量を増やすためのメッセージにおいては，運動によって得られるインセンティブについて，その内容（精神面，または，身体面）と得られるまでの期間（短期的，または，長期的）が受容性に与える影響について交互作用があり，精神面と短期的，身体面と長期的の組み合わせでのみ効果があることが確認されている [16]．しかし，この検証が期間を考慮せずに内容のみで行われた場合には，誤った知見，例えば，「運動量を増やすメッセージには精神面のインセンティブに言及することが有効である」を導いてしまう．その結果，精神面で長期的なインセンティブについて言及するなど有効でないメッセージを用いてしまう恐れがある．このような誤りを避けるためには，少数の文例に依らない，多様なメッセージによる検証が必要となる．また，この誤りは文の特徴間のみではなく，受け手の特性との間でも生じる可能性がある．

本研究では，上記の誤りを避けつつ，ナッジの一種であるインセンティブに着目し，行動変容を促すメッセージの受容性と言及されるインセンティブの間の関係性を検証する．まず，多様なメッセージと受け手に基づいて効果を検証するためのデータセットを用意する．具体的には，多数の書き手による様々な

行動の呼びかけに関する多様なメッセージ 630 件を収集し，各メッセージについて多様な特性を持つ受け手，のべ 73,841 人がそのメッセージを読んで前向きなるか否かの評価，および，その理由を収集した．次に，メッセージの受け手の特性とメッセージの受容性との間の関係性を検証する．その後，メッセージに示される行動変容によるインセンティブに着目し，受け手の特性別にメッセージの受容性を検証する．

以上の検証の結果，行動変容を促すメッセージへの受容性について，

1. 女性の方がメッセージに対する受容性が高い．
2. 若い人ほどメッセージに対する受容性が高い．
3. 男女ともに多くの年代に対して，インセンティブが明示されているメッセージの方が明示されていないメッセージより受容性が高い．
4. 利得を強調するメッセージと損失を強調するメッセージでは受容性の高さは変化しない一方で，有効になる理由が異なる．
5. 50 代の人に対して，インセンティブの根拠が明示されているメッセージの方が明示されていないメッセージより受容性が高い．

という傾向が確認された．このうち，4 の受容性の高さはメタレビュー [4] を支持する結果である．また，1,2 については，メタレビュー [17] によって有意ではない [1] ものの示されていた傾向である．3,5 については，これらを直接的に示す先行研究は見つけられなかったが，インセンティブや根拠の明示が有効であることを前提に，その内容などが検討されており [4, 18]，先行研究を部分的に支持する結果である．

2 関 連 研 究

2.1 行動変容に関する理論

行動変容を効果的に実施するため，様々な理論や手法が提案されている．例えば，Prochaska らは禁煙プログラムのために行動変容ステージモデル [19] を提唱している．行動変容ステージモデルは，人が健康行動を変化させる際の変遷を行動の有無，行動変容の意図に基づいて，無関心期，関心期，準備期，実行期，維持期の 5 つのステージに分割しており，各段階に応じて介入方法を変えることが必要であることを示している．特に，メッセージによる行動変容では，フレーミング効果 [9] と MINDSPACE[11] が注目されている．

フレーミング効果 [9] とは，同じ内容の文であっても，その伝え方によって，意思決定などが変化する現象である．例えば，「十分に体を動かすことで，健康で丈夫な心臓になる可能性が高まります」のような，利益を強調するメッセージである利得フレームと「身体活動が不十分だと，心血管疾患のリスクが高まります」のような損失を強調するメッセージである損失フレームでは，メッセージの受け手が受ける印象やその行動に差異が生じることである．健康行動においては，損失フレームよりも利得フレームの方が予防行動を促すのに効果的であることがメタレビュー [4] によって示されている．

MINDSPACE[11] とは，人の意思決定に主要な要素である Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments, Ego の 9 つの行動特性をまとめたフレームワークである．例えば，Incentives では，客観的に同じ価値を持つ利益と損失について人は損失を大きく見積もる傾向があるという損失回避傾向など 5 つの傾向を示している．Behaviour Change Wheel[20]（Incentivisation や Coercion）でも示されており，介入手法を分類した BCT taxonomy v1[21] においても Incentives に関する複数の分類項目が定義されており，行動変容において重要な要因であることがわかる．本稿では，メッセージの意味的な特徴として Incentives に着目してメッセージの受容性を検証する．

2.2 健康行動のための行動変容メッセージの検証

行動変容のためのメッセージを検証する研究は多数存在する [4,5]．ここでは特にインセンティブに関する検証について述べる．メタレビュー [4] によると，インセンティブの提示方法，すなわち，利得フレーム，または，損失フレームによる効果について調査している．具体的には，利得フレームが予防行動（禁煙 [22]，皮膚がん予防 [22]，運動 [23] など）に対する意図や態度には影響を与えないが，行動を実行させるのに有効であること，その一方で，損失フレームは検出行動（HIV 検査 [24]，子宮頸がん検診 [25] など）に対して意図，態度，実際に行動するかのいずれに関しても有意ではないものの有効である傾向を示している．また，Morris らは運動を促す際に示すべきインセンティブとして「短期的に得られる精神的な利得」と「長期的に得られる身体的な利得」が有効であることを示している [16]．

日本人を対象とした研究も複数存在する．福田らは禁煙を推奨するメッセージについて，年代が低い受け手ほど経済的な損失を強調するメッセージが有効であるなど，示すべきインセンティブは受け手の性別，年齢，学歴，収入，婚姻状況の影響を受けることを示している [13]．また，山本らは受け手の時間選好に着目し，受け手の時間選好に応じて，目先のインセンティブを強調したメッセージと将来のインセンティブを強調したメッセージを出しわけることにより受容性が向上することを示している [14]．

これらの研究の多くは少数の文例に基づいて検証やその検証に基づくメタ的な検証を行っており，文の特徴間や受け手の特性との間の交互作用を考慮できていない可能性がある．本研究では，多数の文例を用いて，受け手の特性別に，インセンティブを示すことの有効性，フレーミング効果の影響，インセンティブの根拠を示すことの有効性を検証する．

3 検証用データの準備

本研究では，少数の文例に依存しない検証を実施するため，多様な特徴を持つ多数のメッセージが必要となる．そこで，研

1：乳がん検査など女性のみを対象とする研究を含めない場合の結果．含める場合は性別による差異は有意であることが示されている．

表 1: 呼びかける内容

呼びかける内容	件数
野菜を食べる	30
塩分を控える	21
間食を控える	20
カロリーを控える	14
ヨガ	13
ストレッチ	13
酒を控える	13
ウォーキング	12
筋トレ	12
煙草を控える	12
その他	21

「一生使える首の整え方」を 3 回にわたって公開します。つらい首こりは放置すると痛みやしびれといった「頸椎症」に悪化していく可能性があるそうです。そんな人は、頸椎に負荷をかけている首、肩、背中の過緊張をほぐす「首トレ」を。第 1 回は基本の「首トレ」を紹介します。

図 1: ストレッチを呼び掛けるメッセージの例 [26]

究者がフレームワークに則ってメッセージを作成するのではなく，Twitter 上で関連する投稿を収集することによりメッセージを用意した．その後，多様な特性の評価者による評価を行うため，クラウドソーシングによって各メッセージの評価を行った．また，各メッセージについて，インセンティブ，すなわち，その行動をとらないと損するように思えるような情報を持つか（MINDSPACE の Incentives に相当），持つ場合にはそれはどのような表記されているかという観点で分類を行った．以下，各手続きについて順に述べる．

3.1 メッセージの収集

本研究では，多様な特徴を持つメッセージを評価対象としつつも，比較を容易にするため 2 つの方法でデータを収集した．

1 つ目の方法は，「ウォーキング」や「野菜を食べる」，「煙草を控える」などの呼びかける内容を定め，それに関連するキーワードを用いた検索である．これは，呼びかける内容によってメッセージの受容性が変化するだろうという仮説に基づいている．具体的に収集した呼びかけの内容と件数は表 1 の通りであり，図 1 はその一例である．

2 つ目の方法は，多数の健康行動に関する投稿を行うアカウントからの収集である．これは，同一のアカウントであれば文体などが類似しており，ある程度統制されたメッセージ間での比較を可能にするという仮説に基づいている．具体的には表 2 のように異なる 3 つのアカウントからそれぞれ複数の投稿を抽出した．

これらの方法により収集した後，アンケート開始前に投稿が削除されたものや重複したものを除外した結果，メッセージの総数は 630 件となった．

3.2 アンケートによる評価

各メッセージがどの程度受容性があるか評価を行うため，クラウドソーシングによる 2023 年 1 月および 2 月にアンケート調査を実施した．具体的には，D style web[2]を用い，運動や食生活などの生活習慣の改善に関心がある回答者を募集した．各

2：https://dstyleweb.com/

あなたは，お金を必ずもらえるとの前提で，
a. 今 10 万円をもらう
b. 1 年後に 11 万円をもらう
という 2 つの選択肢があればどちらを選びますか.

図 2: 近視眼的行動バイアスを問う設問

あなたは今後，運動や食生活などの生活習慣を改善してみようと思いますか.
以下の中から最もお気持ちに近いものをお知らせください.

図 3: 行動変容ステージを問う設問

回答者には 5 つのメッセージの評価を依頼した．各特性の回答者を十分に集めるため，各メッセージについて男女，年代（20 代，30 代，40 代，50 代，60 代以上）別に 50 人以上，計 500 人以上の回答があるよう募集し，無作為にメッセージの割り付けを行った．その結果のべ 73,841 人分の回答が集まった．

アンケート調査の手続きは個人の特性の調査，回答者の抽出，メッセージの評価の順に行った．詳細は以下の通りである．

3.2.1 個人の特性の調査

各回答者の特性をアンケートによって収集した．収集した特性は「自身，および，同居家族の業種」，「性別」，「年齢」，「居住都道府県」，「婚姻状況，および，子供の有無」，「職業」，「年収」，「最終学歴」，「近視眼的行動バイアス」，「行動変容ステージ」，「SNS の利用頻度（Twitter，Instagram，Facebook，YouTube，その他）」である．ここで，先行研究 [13] によって性別，年齢，婚姻状況，職業，年収，最終学歴はメッセージの受容性に影響を与えるものとして指摘されている．また，近視眼的行動バイアスとは，短期的な利益と長期的な利益のどちらを重視するかを表す傾向であり，[27] に示した設問によって，「a（b）を選ぶ」，「どちらかと言えば a（b）を選ぶ」，「どちらともいえない」の 5 択から選択することで得られる．なお，先行研究 [14] では時間選好の影響を扱っているが，計測に必要な設問数が多いこと，質問紙の有効回答の低さから類似した概念である近視眼的行動バイアスを本研究では採用した．行動変容ステージとは，行動変容ステージモデル [19] における人が行動を変化させる際に変遷すると考えられている 5 つのステージ（無関心期，関心期，準備期，実行期，維持期）であり，図 3 の設問に対し，表 3 から選択することで得られる．先行研究 [6] において，ステージによって有効なメッセージが異なることが指摘されている．

3.2.2 回答者の抽出

健康的な習慣に興味を持つものを対象とするため，行動ステージが無関心期以外の回答者を抽出した．その後，性別，年代を考慮して各メッセージの回答数が十分集まるように各回答者に 5 つのメッセージを無作為に割り当てた．

3.2.3 メッセージの評価

各回答者に対し，2 つの設問を提示し，1 つのメッセージの評価を実施することを 5 回繰り返すことによって評価を実施した．1 つ目の設問は，図 4 のようにメッセージの URL を示し，回答者に URL へのアクセスを依頼し，4 件法によりメッセー

表 2: 選定したアカウントと抽出の手続き

アカウント	件数	抽出方法
@Hokkaido_health	89	2022 年から 2023 年 1 月 10 日までの投稿を抽出．個人への呼びかけを人手で抽出し，重複した投稿を除外．
@askenjp	330	2023 年 1 月 13 日までに投稿された「#ぷちけん」が記載されている投稿全てを抽出．
@DieterTrivia	36	2020 年および 2022 年に投稿を抽出 [3]．商品の宣伝を人手で除外した後，最新の投稿から 36 件を抽出．

表 3: 行動変容ステージと回答選択肢

ステージ	選択肢
無関心期	改善するつもりはない
関心期	改善するつもりである（概ね 6 か月以内）
準備期	近いうちに（概ね 1 か月以内）改善するつもりであり、 少しずつ始めている
実行期	既に改善に取り組んでいる（6 か月未満）
維持期	既に改善に取り組んでいる（6 か月以上）



図 4: アンケート回答画面 [4]

ジの評価を得た．2 つ目の設問では，1 つ目の設問の回答理由を尋ねた．具体的には，1 つ目の設問で，「なった」，または，「ややなった」と回答した場合には，表 4 から 1 つ以上の理由を選択する．一方で，1 つ目の設問で，「あまりならなかった」，「ならなかった」場合には，表 4 の対となる理由（例えば「N1. 実施する内容がよくわからなかったから」）から選択を行う．ここで，N1–N9 は前向きにならなかった場合の理由 ID である．なお，P9，または，N9 のその他を選択した場合は具体的な記述を求めた．

3.3 メッセージの分類

各メッセージが持つ特徴について，3 つの基準に基づいて著者一人によって分類を行った．各基準は以下の通りであり，分類結果は表 5 の通りである．

3：2021 年の投稿は検索結果に含まれなかった．
4：テスト用の画面であり，実際の回答画面とは差異がある．

表 4: メッセージで前向きになった理由一覧

ID	前向きになった理由
P1	実施する内容がわかりやすかったから
P2	実施することが簡単そうだから
P3	実施したときの効能が魅力的だから
P4	関心のある健康習慣（行動）だから
P5	ツイートしているアカウント（アカウント名とプロファイル画像）の印象がよいから
P6	ツイートの文面の印象がよいから
P7	ツイートに添付または展開されている画像（動画）の印象がよいから
P8	ツイート上のリンク先の印象がよいから
P9	その他

表 5: 各基準による分類結果

ラベル	該当件数	非該当件数
インセンティブ記載あり	425	205
利得フレーム	349	76
損失フレーム	102	323
根拠記載あり	105	320

3.3.1 インセンティブ有無

メッセージ内に推奨する行動のみではなく，インセンティブ，すなわち，その行動をとらないと損すると思えるような情報を含む「インセンティブ記載あり」か含まない「インセンティブ記載なし」かを表す．人の意思決定に影響を与える主要な要素である MINDSPACE[11] の Incentives や行動変容理論の枠組みをまとめる Behaviour Change Wheel[20] の Incentivisation や Coercion に該当し，メッセージの受容性を考慮する上で，重要な観点である．

3.3.2 フレーミング

インセンティブの記載があるメッセージについて利得を強調している「利得フレーム」か，損失を強調しているか「損失フレーム」かを表す．これはフレーミング効果によってメッセージの受容性がどのように変化するか捉えるために重要な観点であり，複数の研究で着目されている [4,10]．しかしながら，意図や態度に関してはフレーミングによって差異がないことが示されている [4]．したがって，本研究でも受容性には有意差がないと予想される．なお，各メッセージには複数のインセンティブが記載されることもあるため，利得フレームでもあり，損失

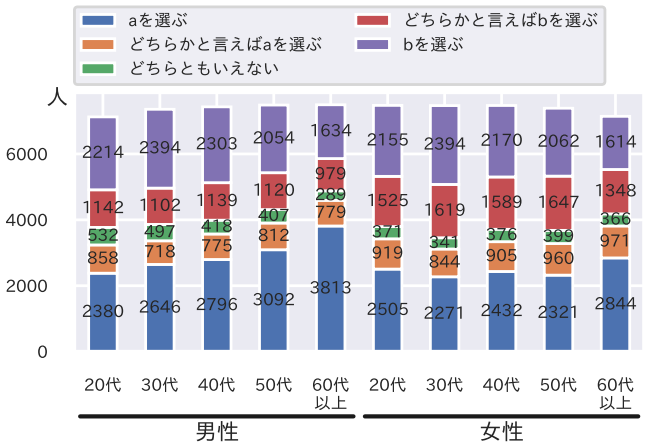


図 5: 性別年代と近視眼的行動バイアスの関係

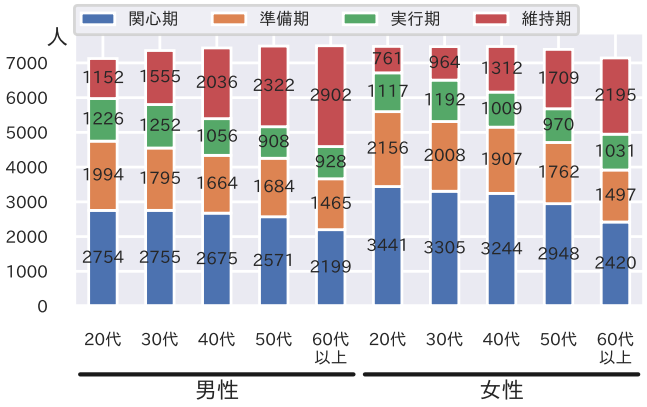


図 6: 性別年代と行動変容ステージの関係

フレームでもあるメッセージも存在する．

3.3.3 根拠有無

これはメッセージに記載のあるインセンティブについて，そのインセンティブが生じる根拠や理由を明記している「根拠記載あり」か，明記していない「根拠記載なし」かを表す．これは，メッセージによって前向きになった（ならなかった）理由のその他の記載に，「納得感がある」，「論理的」，「解説がある」，「うさん臭い」，「根拠がない」といった記載が多くみられたため，それを説明しうる要素と考え採用した．また，この基準は Behaviour Change Wheel の Education に関連がある．

4 基礎分析

本章では，得られた評価用データについて前処理，および，統計的検定によって受け手の特性によって受容性の差異が生じるか確認を行う．なお，有意水準はすべて 5%とする．

4.1 収集結果

調査の結果，630 件のメッセージに対してのべ 73,841 人の 369,205 件の回答を得られた．1 メッセージあたりの回答件数は平均 586 件（最小 574 件，最大 601 件）であった．回答者の性別，年代，および，近視眼的行動バイアス，行動変容ステージごとに人数は図 5 と 6 のようになった．

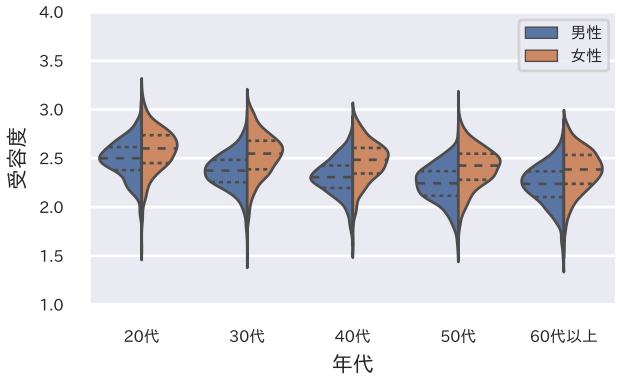


図 7: 性別年代別の受容度の分布

4.2 前処理

外れ値を除外するため，前向きになった理由，および，ならなかった理由の自由記述の確認を行った．その結果，3 件のメッセージでアンケート期間中に鍵アカウントになっているなどの理由で閲覧できない状態になっていたことを示唆する理由が記載されていた．そのため，これらのメッセージを分析対象から除外し，以降では残りの 627 件を分析対象とする．

4.3 受容度の分布

回答者の性別，年代による受容性の違いを比較するため，3.2.3 項の 1 つ目の回答について，4 件法のリッカート尺度であることから，受容度 s を

$$s = \frac{4N_{\text{なった}} + 3N_{\text{ややなった}} + 2N_{\text{あまりならなかった}} + N_{\text{ならなかった}}}{N_{\text{なった}} + N_{\text{ややなった}} + N_{\text{あまりならなかった}} + N_{\text{ならなかった}}} \quad (1)$$

のように回答を平均化することにより定義した．ここで，$N_{\text{選択肢}}$ は該当の回答をした人数を表す．

各性別，年代の受容度の分布は図 7 の通りである．これらの差異を検証するため，被験者内計画の性別，および，年代の二元配置分散分析によって検定を行ったところ，各要因，および，交互作用に有意差が確認された（性別 $F(1,626) = 908.245, p < 0.001, \eta^2_{\mathrm{G}} = 0.115$; 年代 $F(4,2504) = 646.061, p < 0.001, \eta^2_{\mathrm{G}} = 0.153$; 交互作用 $F(4,2504) = 7.332, p < 0.001, \eta^2_{\mathrm{G}} = 0.005$）．QQ プロットにより，正規性が確認できなかったため，下位検定として，ウィルコクソンの符号付き順位検定を適用し，ボンフェローニの調整を行ったところ，（男性 50 代，男性 60 代以上），（男性 20 代，女性 40 代），（男性 30 代，女性 60 代以上）の組み合わせでは有意差を確認できなかったが（$ps > 0.523$），その他の組み合わせでは有意差が確認された（$ps < 0.003, |gs| > 0.143$）（g は Hedges の効果量）．すなわち，どの年代においても，女性の方が男性よりも受容度が有意に高く，男性 50 代と男性 60 代以上の間を除いて，男女共に年代が上がるほど，受容度が有意に低くなる．これらの傾向はメタレビュー [17] が示す傾向と一致する．

4.4 回答理由の分布

前向きになった理由の回答分布は図 8 となった．なお，理

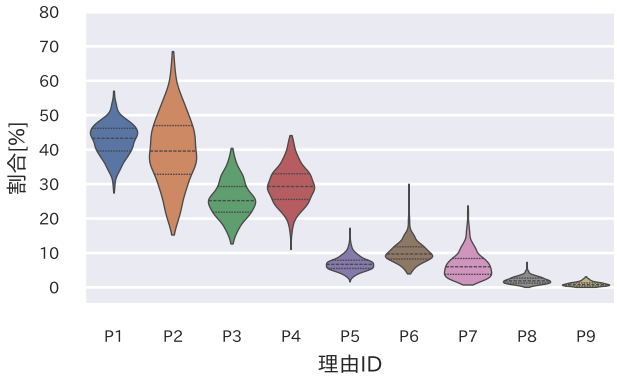


図 8: 前向きになった理由の分布

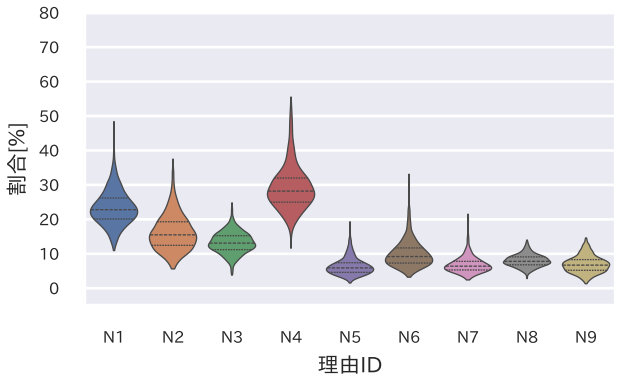


図 9: 前向きにならなかった理由の分布

由 ID は表 4 に対応している．QQ プロットにより，正規性が確認できなかったため，各理由間での比較をウィルコクソンの符号付き順位検定を適用し，ボンフェローニの調整を行ったところ，すべての組み合わせについて有意差が確認された（$ps < 0.007, |gs| > 0.104$）．したがって，前向きになる理由としては，「P1. 実施する内容がわかりやすかったから」，「P2. 実施することが簡単そうだったから」，「P4. 関心のある健康習慣（行動）だから」など，内容に関するものが有意に多く選択され，「P6. ツイートの文面の印象がよいから」など内容以外の理由は有意に選ばれない傾向である．

また，前向きにならなかった理由の回答分布は図 9 となった．なお，理由 ID は図 8 と同様に表 4 に対応している（3.2.3 項で述べたように対になっている）．QQ プロットにより，正規性が確認できなかったため，各理由間での比較をウィルコクソンの符号付き順位検定を適用し，ボンフェローニの調整を行ったところ，「N7. ツイートに添付または展開されている画像（動画）の印象がよくないから」と「N9. その他」の組では有意差は確認できなかった（$p = 1.000$）．一方で，他の組み合わせについてはすべて有意差が確認された（$ps < 0.003, |gs| > 0.201$）．したがって，前向きにならなかった理由も内容に関する項目が有意に多く選ばれており，内容以外の理由は有意に選ばれない傾向であるといえる．行動変容ステージが無関心期ではない回答者を対象にしたにも関わらず，「N4. 関心のない健康習慣（行動）だから」が最も多く選ばれたことは，行動変容ステージを「運動や食生活など」よりも細かく設定する必要性を示している．

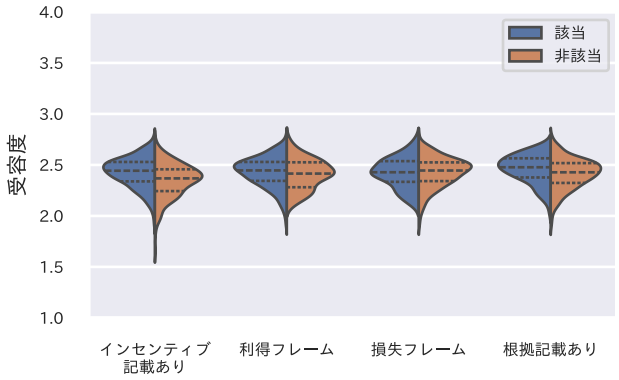


図 10: メッセージの分類ごとの受容度の分布

5　インセンティブの効果検証

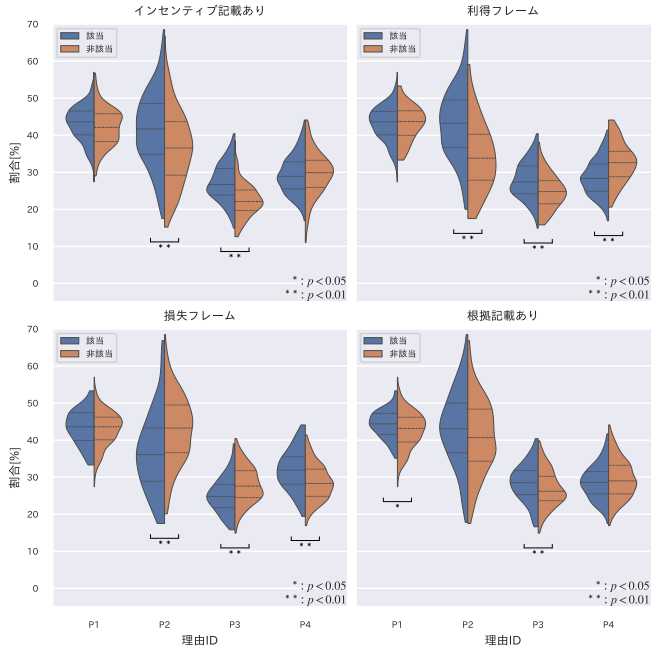
5.1　受容度への影響

3.3 節で分類した結果ごとの，受容度の分布は図 10 のようになった．4.3 節の結果から，性別，年代別に分類結果によって受容度に差があるか評価を行った．具体的には，QQ プロットにより，正規性が確認できなかったため，各分類間での比較をマン・ホイットニーの U 検定を適用し，ボンフェローニの調整を行った．その結果，インセンティブ記載ありでは，男性 20 代における有意差が確認できなかった（$p = 0.072$）ものの，他の性別，年代では有意差が確認された（$ps < 0.015, gs > 0.307$）．また，男性 20 代を含め，すべての性別，年代で該当メッセージの方が非該当メッセージよりも受容度が高い傾向にあった．利得フレーム，および，損失フレームではすべての性別，年代で有意差が確認できなかった（利得フレーム $ps > 0.106$; 損失フレーム $ps > 0.295$）．根拠記載ありでは，男性 50 代と女性 50 代においてのみ有意差がみられた（男性 50 代 $p = 0.001, g = 0.425$; 女性 50 代 $p = 0.017, g = 0.326$; その他 $ps > 0.179$）．すべての性別，年代で該当メッセージの方が非該当メッセージよりも受容度が高い傾向にあった．

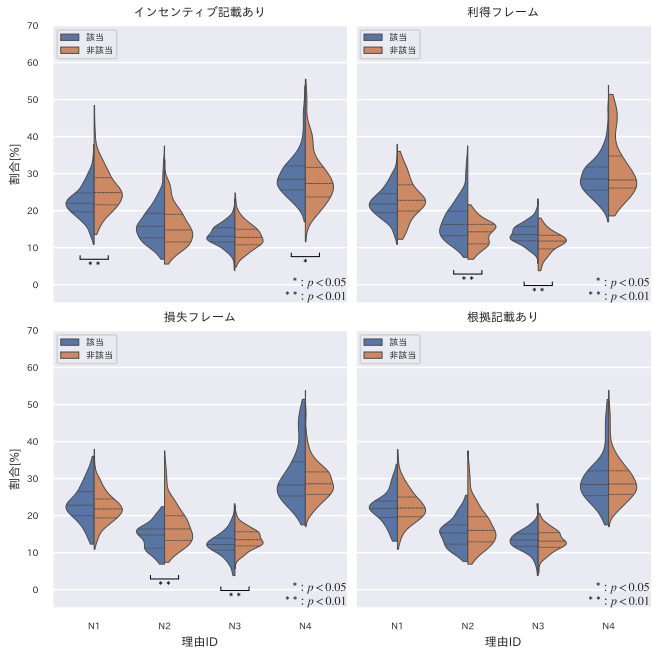
以上のことから，インセンティブを考慮したメッセージは殆どすべての性別，年代に有効であるといえる．特に，50 代の人たちには単にインセンティブを示すのではなく，そのインセンティブが得られる根拠を示すことが有効であるといえる．一方で，利得フレームと損失フレームに関してはその受容性に関して差がみられなかった．この結果は，健康行動に対する意図や態度にはフレーミング効果による有意差はないという先行研究 [4] と一致する．

5.2　回答理由への影響

3.3 節で分類した結果ごとの，メッセージの内容による前向きになった理由の分布は図 11a のようになった．P1–4 について分類結果間での比較をマン・ホイットニーの U 検定を適用し，ボンフェローニの調整を行ったところ，インセンティブ記載ありでは，P2，および，P3 で有意差があり，（$ps < 0.001, gs > 0.497$），有意に選ばれる一方で，P1，および，P4 では有意差が確認できなかった（$ps > 0.055$）．利得フレームにおいては，P1 では有意



(a) 前向きになった理由



(b) 前向きにならなかった理由

図 11: 分類ごとの回答理由

差が確認できなかったが（$p = 1.000$），他の理由（P2–4）については有意差が確認できた（$ps < 0.001, |gs| > 0.617$）．なお，P2，および，P3 は有意に選ばれ，P4 は有意に選ばれない傾向であった．損失フレームにおいては，利得フレームと同様に P1 では有意差が確認できなかったが（$p = 1.000$），他の理由（P2–4）については有意差が確認できた（$ps < 0.001, |gs| > 0.573$）．ただし，傾向は利得フレームと逆であり，P4 は有意に選ばれ，P2，および，P3 は有意に選ばれない傾向であった．根拠記載ありでは，P1，および，P3 で有意に選ばれる傾向があり（$ps < 0.016, gs > 0.326$），他の 2 つ（P2，P4）では有意差を確認できなかった（$ps > 0.152$）．

3.3 節で分類した結果ごとの，メッセージの内容による前向きにならなかった理由の分布は図 11b のようになった．N1–4 について分類結果間での比較をマン・ホイットニーの U 検定を適用し，ボンフェローニの調整を行ったところ，インセンティブ記載ありでは，N1，および，N4 で有意差があり，（N1 $ps < 0.001, g = -0.638$; N4$ps = 0.027, g = 0.155$），N1 は有意に選ばれない，N4 は有意に選ばれる一方で，N2，および，N3 では有意差が確認できなかった（$ps > 0.151$）．利得フレームにおいては，N1，N4 で有意差が確認できなかったが（$ps > 0.372$），N2，N3 については有意差が確認できた（$ps < 0.001, gs > 0.609$）．なお，N2，および，N3 は有意に選ばれる傾向であった，損失フレームにおいては，利得フレームと同様に N1，N4 では有意差が確認できなかったが（$ps > 0.316$），N2，N3 については有意差が確認できた（$ps < 0.001, |gs| > 0.499$）．ただし，傾向は利得フレームと逆であり，N2，N3 は有意に選ばれない傾向であった．根拠記載ありでは有意差が確認できなかった（$ps > 0.213$）．

インセンティブ記載により，インセンティブの記載がないものに対して，簡単そう（P2），効能が魅力的（P3）との回答が増加していることから，効能が魅力的であると捉えやすくなり，推奨する行動への心理的抵抗を軽減している可能性がある．一方で，関心のない健康習慣（N4）が増大していることから，その効能が回答者にとって魅力的でない場合に，行動への関心を失う可能性を示唆している．また，内容がよくわからなかった（N1）との回答が減少しており，これはインセンティブの記載により，メッセージ性が伝わりやすくなる効果があることを示唆している．

利得フレームでは，簡単そう（P2），効能が魅力的（P3）との回答が増加しており，インセンティブ記載の効果を更に強調していることが示唆される．また，関心のある健康習慣（P4）が減少していることから，元々関心がない人への働きかけにも有効であることが示唆される．一方，難しそう（N2），効能が魅力的でない（N3）といった回答も増加しており，効能によっては逆効果になっている可能性がある．

損失フレームについては，簡単そう（P2），難しそう（N2）の双方が減少していることから，行動の難易度を考慮しなくなることが示唆されている．また，効能が魅力的（P3），効能が魅力的でない（N3）の双方が減少しているのは，損失フレームが行動で得られる魅力を訴えるのではなく，行動しないことで生じる不都合を伝えるためであると考えられる．

根拠記載については，内容がわかりやすい（P1），効能が魅力的（P3）が増加しており，根拠が記載されることで納得感を得られて前向きになりやすくなっている可能性がある．

6 ま と め

本研究では，人の健康行動を変容させることを目的とするメッセージについて，インセンティブによる効果を多数の文例を用いて，受け手の特性を考慮して検証した．各メッセージは Twitter 上で収集したものであり，推奨する行動や文体など多

様な特徴を持つ．各メッセージは性別，年代別に各 50 人以上に評価されており，評価理由も併せて収集されている．本研究では，受け手の性別，年代と，メッセージに記載されるインセンティブの観点から受容性を分析した．その結果，男性より女性，若い人ほどメッセージによって健康行動に対し前向きになりやすいことが確認された．また，インセンティブの観点では，多くの特性の人に対し，インセンティブの記載によって前向きになりやすくなることが確認され，50 代の人には根拠をあわせて示すことが有効であることが確認された．フレーミングに関しては，前向きのなりやすさについては差異がみられなかったものの，前向きになる理由の差異が確認された．ただし，性別や年代によって近視眼的行動バイアスや行動変容ステージなどの分布が異なっていることや多様なメッセージを対象としているため他の特徴量が統制されていない．したがって，今回確認された差異が年代や性別によって直接生じているのか，他の特性を媒介して生じているのかは評価できていない．今後は構造方程式モデリングなどを用いてこれらの影響を加味した分析を行っていく予定である．

本研究では，今回の評価者による評価は前向きになるかで評価を行っており，実際に行動に繋がったかの評価は実施していない．健康行動に対する意図，態度は必ずしも行動変容に繋がらない可能性が示されている [4]．そのため，今回の分析で得られた知見が実際に行動変容できるかは今後検証が必要である．また，インセンティブは著者一人によるラベル付けによるもののため，今後は，著者以外による複数人でのラベル付けを実施し，ラベルの正確性，信頼性の確保を行う予定である．

文　献

[1] Samira Humaira Habib and Soma Saha. Burden of Non-Communicable Disease: Global Overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Vol. 4, No. 1, pp. 41–47, 2010.

[2] Regina Guthold, Gretchen A. Stevens, Leanne M. Riley, and Fiona C. Bull. Worldwide Trends in Insufficient Physical Activity from 2001 to 2016: A Pooled Analysis of 358 Population-Based Surveys with 1·9 Million Participants. *The Lancet Global Health*, Vol. 6, No. 10, pp. e1077–e1086, 2018.

[3] World Health Organization (WHO). Primary health care. `https://www.who.int/news-room/fact-sheets/detail/primary-health-care`. Accessed: 2023-12-26.

[4] Kristel M. Gallagher and John A. Updegraff. Health Message Framing Effects on Attitudes, Intentions, and Behavior: A Meta-analytic Review. *Annals of Behavioral Medicine*, Vol. 43, No. 1, pp. 101–116, 2012.

[5] Amy E. Latimer, Lawrence R. Brawley, and Rebecca L. Bassett. A Systematic Review of Three Approaches for Constructing Physical Activity Messages: What Messages Work and What Improvements Are Needed? *International Journal of Behavioral Nutrition and Physical Activity*, Vol. 7, No. 1, p. 36, 2010.

[6] J. Paige Pope, Luc Pelletier, and Camille Guertin. Starting Off on the Best Foot: A Review of Message Framing and Message Tailoring, and Recommendations for the Comprehensive Messaging Strategy for Sustained Behavior Change. *Health Communication*, Vol. 33, No. 9, pp. 1068–1077, 2018.

[7] Toshio Fujimi and Kodai Fujimura. Testing Public Interventions for Flash Flood Evacuation through Environmental and Social Cues: The Merit of Virtual Reality Experiments. *International Journal of Disaster Risk Reduction*, Vol. 50, p. 101690, 2020.

[8] Luca Falasconi, Matteo Vittuari, Alessandro Politano, and Andrea Segrè. Food Waste in School Catering: An Italian Case Study. *Sustainability*, Vol. 7, No. 11, pp. 14745–14760, 2015.

[9] Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, Vol. 211, No. 4481, pp. 453–458, 1981.

[10] Daniel Sibley, Rebecca L. Bassett-Gunter, and Brad A. Meisner. Promoting Lifespan Physical Activity Intentions... and Aging Anxieties? The Paradox of High-Risk and Loss-Framed Messages. *Journal of Adult Development*, Vol. 30, No. 2, pp. 216–223, 2023.

[11] Paul Dolan, Michael Hallsworth, David Halpern, Dominic King, Robert Metcalfe, and Ivo Vlaev. Influencing Behaviour: The Mindspace Way. *Journal of economic psychology*, Vol. 33, No. 1, pp. 264–277, 2012.

[12] 濱谷尚志, 山田渉, 落合桂一, 山本直樹, 檜山聡. 健康に関する行動変容のための社会的メッセージ提示方法の検討. 研究報告マルチメディア通信と分散処理 (DPS), Vol. 2021, No. 5, pp. 1–6, 2021.

[13] 福田吉治, 林辰美. 健康づくりに関するメッセージの効果認識の関連要因 社会経済的要因に注目して. 日本公衆衛生雑誌, Vol. 62, No. 7, pp. 347–356, 2015.

[14] 山本修平, 赤木康紀, 冨永登夢, 瀧本祥章, 倉島健, 戸田浩之. ユーザの時間選好を考慮したメッセージ提示が健康行動に与える影響. マルチメディア，分散，協調とモバイルシンポジウム 2023 論文集, Vol. 2023, pp. 475–485, 2023.

[15] 佐藤妙, 阿部直人. 健康行動の意思決定における介入メッセージの検討－動機づけナッジメッセージの効果検証－. 行動経済学会第 14 回大会, 2020.

[16] Benjamin Morris, Rebecca Lawton, Rosemary McEachan, Robert Hurling, and Mark Conner. Changing Self-Reported Physical Activity Using Different Types of Affectively and Cognitively Framed Health Messages, in a Student Population. *Psychology, Health & Medicine*, Vol. 21, No. 2, pp. 198–207, 2016.

[17] Seth M. Noar, Christina N. Benac, and Melissa S. Harris. Does Tailoring Matter? Meta-analytic Review of Tailored Print Health Behavior Change Interventions. *Psychological Bulletin*, Vol. 133, No. 4, pp. 673–693, 2007.

[18] Mike Allen and Raymond W. Preiss. Comparing the Persuasiveness of Narrative and Statistical Evidence Using Meta-analysis. *Communication Research Reports*, 1997.

[19] James O. Prochaska and Carlo C. DiClemente. Transtheoretical Therapy: Toward a More Integrative Model of Change. *Psychotherapy: theory, research & practice*, Vol. 19, No. 3, p. 276, 1982.

[20] Susan Michie, Maartje M. van Stralen, and Robert West. The Behaviour Change Wheel: A New Method for Characterising and Designing Behaviour Change Interventions. *Implementation Science*, Vol. 6, No. 1, p. 42, 2011.

[21] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline E. Wood. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine*, Vol. 46, No. 1, pp. 81–95, 2013.

[22] Sherry A. McKee, Stephanie O'Malley, Wayne T. Steward, Susan Neveu, Michelle Land, and Peter Salovey. How to Word Effective Messages About Smoking and Oral Health: Emphasize the Benefits of Quitting. *Journal of Dental Education*, Vol. 68, No. 5, pp. 569–573, 2004.

[23] Amy E. Latimer, Tara A. Rench, Susan E. Rivers, Nicole A. Katulak, Stephanie A. Materese, Lisa Cadmus, Althea Hicks, Julie Keany Hodorowski, and Peter Salovey. Promoting Participation in Physical Activity Using Framed Messages: An Application of Prospect Theory. *British Journal of Health Psychology*, Vol. 13, No. 4, pp. 659–681, 2008.

[24] Anne Marie Apanovitch, Danielle McCarthy, and Peter Salovey. Using Message Framing to Motivate HIV Testing among Low-Income, Ethnic Minority Women. *Health Psychology*, Vol. 22, No. 1, pp. 60–67, 2003.

[25] Diane Lauver and Mary Rubin. Message Framing, Dispositional Optimism, and Follow-up for Abnormal Papanicolaou Tests. *Research in Nursing & Health*, Vol. 13, No. 3, pp. 199–207, 1990.

[26] 日経ヘルス. https://twitter.com/nikkeihealth/status/1591928397823111168, 2022. Accessed: 2023-12-27.

[27] 金融広報中央委員会. 調査結果 — 金融リテラシー調査（2022年）｜知るぽると. https://www.shiruporuto.jp/public/document/container/literacy_chosa/2022/. Accessed: 2023-11-30.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年3月1日(金) 13:30 ~ 15:40 | T1-B オンライン（Zoom Events)

感情分析 (2)

座長:富井 尚志(横浜国立大学)
コメンテータ:渡辺 陽介(名古屋大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T1-B-8-01] インターネットスラングを考慮した大規模言語モデルを用いた感情分析手法の提案

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

14:00 ~ 14:25
[T1-B-8-02] Evaluating the Performance of ChatGPT for Aspect-Based Sentiment Analysis

*Wang Yifei[1]、Iwaihara Mizuho[1] (1. 早稲田大学)

14:25 ~ 14:50
[T1-B-8-03] 大規模言語モデルによる細かな感情推定手法

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

14:50 ~ 15:15
[T1-B-8-04] 大規模言語モデルによる Multi-Step Prompt を用いた感情認識

*濱 健太[1]、大塚 淳史[1]、石井 亮[1] (1. 日本電信電話株式会社　人間情報研究所)

インターネットスラングを考慮した大規模言語モデルを用いた感情分析手法の提案

関　優花† 清　雄一† 田原　康之† 大須賀昭彦†

† 電気通信大学　〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: †seki.yuka@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　オンラインプラットフォームの急速な普及に伴い，大規模言語モデルを用いたインターネット上における人間の活動分析の研究が活発に行われている．その一方で，砕けた言葉の増加がモデルによる活動分析の結果へ影響を及ぼすことが懸念される．本研究では，大規模言語モデルの 1 つである BERT を用いて，インターネットスラング等を含む特殊表現コーパスを用いて追加学習を行ったモデルと，追加学習を行わなかったモデルについて，どちらがオンラインプラットフォーム上のテキストに対してより正確に感情を示すラベル付けを行うことができるかを比較するために，それぞれのモデルを Positive/Neutral/Negative の 3 値分類を行う感情分析についてファインチューニングし，評価を行った．その結果，追加学習を行ったモデルの方がより正確にラベル付けを行えるモデルであることが示された．

キーワード　テキスト分類，LLM，ソーシャルメディア，インターネットスラング

1 はじめに

近年 X (旧：Twitter) や YouTube 等のソーシャルネットワーキングサービス (SNS) の急速な普及により，公開されている API 等を用いることで一般の利用者が投稿テキストなどのデータを容易に取得できるようになった．その結果，取得できるデータを用いたインターネット上における人間の活動分析が研究として活発に行われている．テキストを用いた活動分析では，大量のテキストデータで事前学習された大規模言語モデル (LLM) が用いられることが多く，Google が発表した BERT や OpenAI が発表した GPT-3 などを利用した活動分析が既に行われている．日本語の大規模言語モデルとしては，東北大学によって公開された BERT base Japanese [1] がある．

LLM を用いた活動分析が進む一方で，SNS では，若い世代が流入したことにより砕けた言葉でのコミュニケーションが増加したことにより，インターネットスラング等の特殊な表現が出現するようになった．LLM は Wikipedia やニュース記事等の文語で構成されたテキストデータで事前学習が行われており，これらの特殊な表現に対しての対応は十分でない．そのため，特殊な表現の意味が既存の語彙の本来の意味と競合したり，誤った単語分割がなされることによってインターネット上のテキストデータを利用した活動分析の精度に影響を及ぼす可能性があり，近年の活動分析ではこの影響を軽減することが課題となっている．

この課題を改善するために用いられる手法の 1 つとしてテキストの前処理が挙げられるが，全てのインターネットスラングに対応することは困難であり，インターネットスラングを考慮したモデルを組む手法が求められる．また，BERT base Japanese における改善手法としては，トークン化において用いられる辞書として，新語や流行語に強い辞書を採用する手法が挙げられる．例えば，MeCab の辞書には新語に強いとされる辞書である mecab-ipadic-NEologd がある．しかし，mecab-ipadic-NEologd は過去 3 年間ほど更新が止まっており，現在では特殊な表現をモデルに組み込む別の手法が求められている．

これらの課題を踏まえて，本研究の目的はインターネットスラングが多く用いられる SNS 上のテキストに対して，スラングを考慮した感情分析を行うことができる LLM を作成することである．本研究では，既存 LLM をファインチューニングして感情分析タスクに適応させる既存手法のアプローチに，トークン化に用いられる辞書へのインターネットスラングの追加と，インターネットスラングを含むコーパスによる追加学習を追加した．辞書へのインターネットスラングの追加に加えて，インターネットスラングを含むテキストで構成されたコーパスを用いて追加学習を行うことで，追加した語句に対して適切なベクトルを獲得し，インターネットスラングを LLM における文脈理解に活用できるようにすることを目指す．本研究では，追加学習を行うモデルとして，Hugging Face 上で最もダウンロードされている日本語特化の大規模言語モデルである BERT base Japanese を採用した．

実験では，まず追加学習によってインターネットスラングに対する類義語がどのように変化したかを比較することで，単語の意味に沿った単語ベクトルが獲得できているかを評価した．次に，感情分析タスクにおいて辞書の追加およびインターネットスラングを用いて追加学習を行ったモデルと，辞書の追加および追加学習のどちらも行わなかったモデルについて，どちらがより正確に SNS 上のテキストに対して感情ラベルを付与できるかを比較した．その結果，まず追加学習においてはインターネットスラングに対する類義語にはよりその単語の意味に近い語句が上位に出現するという変化が見られ，感情分析タスクでは，追加学習を行なったモデルの精度 (Accuracy) および Macro-F1 スコアに改善が見られた．

本論文の構成は以下のとおりである．第 2 章で関連研究について，第 3 章で本研究における提案手法について，第 4 章で提案手法による実験結果について，第 5 章で考察について，第 6 章で本研究のまとめ及び今後の展望について記す．

2 関連研究

2.1 BERT base Japanese

BERT base Japanese [1] とは，東北大学自然言語処理グループによって公開されている日本語テキスト向けの事前学習済み BERT モデルである．大まかな仕組みとしては，まず Mecab と IPA 辞書を利用して形態素解析を行い，次に WordPiece アルゴリズムを利用してサブワードに分割している．モデルのアーキテクチャは既存の BERT モデルと同じであり，事前学習で用いられる学習データには 2019 年 9 月 1 日時点の日本語版 Wikipedia のデータから約 1700 万の文章が用いられている．

この BERT モデルのトークナイザーの語彙数は 32000 であり，トークナイザーには，Mecab と WordPiece アルゴリズムが用いられている．なお，トークナイザーで用いられる辞書を変更することでモデルの使用用途に合わせて形態素解析を行うことができ，主な辞書としては IPA 辞書や Juman 辞書，Unidic 辞書が挙げられる．新語や流行語に強いモデルを作成する場合としては，通常の IPA 辞書に加えて Web 上の言語資源から得た新語が含まれる mecab-ipadic-NEologd を併用することで形態素解析の精度を高めることができる．

2.2 SNS 上のテキストに対する感情分析

X (旧:Twitter) をはじめとするソーシャルネットワーキングサービス (SNS) 上のテキストに対する感情分析の研究は多く行われている．Samir ら [2] は，BERT を利用したツイートに対する感情分析手法を提案している．著者らは，160 万件のツイートデータセットに対する感情分析タスクを実施し，Transformer に基づいた BERT を利用したモデルと機械学習アプローチであるナイーブベイズやランダムフォレストを利用したモデルによる精度を比較したところ，BERT を利用したモデルがはるかに優れたパフォーマンスを発揮するという結果が得られた．この研究では，テキストへの前処理としてリンクとメンション (@username)，記号と絵文字の削除が行われたが，インターネットスラングなどの特殊表現に対しての考慮は行われていない．また，BERT を利用した日本語のソーシャルメディアテキストに対する感情分析も行われており，顧客のニーズの把握や世論分析などに用いられている．瀧川ら [3] は，Twitter 上の世論の感情分析タスクにおいて RoBERTa モデルを利用し，既存の機械学習モデルである ML-Ask と正解率の比較を行っている．その結果，RoBERTa を感情分析タスク向けにファインチューニングしたモデルでは 85.79%の正解率を示し，ML-Ask による分類の正解率やベースラインを大きく上回った．

2.3 言語モデルへの語彙の追加

既存の言語モデルは汎用性が高い一方，汎用性の高さ故に専門用語に対応できず，専門用語を含むテキストに対するタスクにおいて精度が低下する点が欠点である．この欠点の解決策の 1 つとしては，モデルに対して専門用語をユーザ独自の語彙として追加する方法が挙げられる．

Li ら [4] は，既存の機械学習アプローチに対して未知語として扱われるインターネットスラングと絵文字を辞書として追加する手法を提案している．[4] の提案手法では，ソーシャルメディア上から頻出のインターネットスラング用語とテキスト特徴に変換した絵文字を抽出し，ユーザ辞書としてセグメンテーションツールに追加した後に，単語ベクトルを学習しその結果を機械学習モデルに適用して感情分析タスクを実行した．その結果，インターネットスラング辞書のみを追加した場合，絵文字辞書のみを追加した場合，両方を追加した場合全てにおいて，辞書を追加しない場合と比較して F1 スコアが上回ることが示され，辞書を追加することの有用性が示された．しかし [4] では大規模言語モデルに対して辞書の追加を行なった場合の結果は示されていない．

三林ら [5] は，既存の事前学習済み BERT モデルに対して文化財に関連するデータセットを利用した追加学習と未知語の辞書追加を行うことによる特化型 BERT モデルを提案している．[5] では，事前学習済み BERT モデルでは未知語を示す [UNK] トークンとして扱われる語彙をトークナイザの辞書と BERT モデルの embedding 層に追加することで，未知語も考慮したモデルを作成している．[5] における BERT モデルへの語彙追加により，トークン化の結果は図 1 の通りに変化する．

元の文章: 薬品を秤にかけて売る。

語彙追加前: 薬品 を [UNK] に かけ て 売る 。
語彙追加後: 薬品 を 秤 に かけ て 売る 。

図 1　語彙追加によるトークン化の例

その結果，事前学習済みモデルに対して追加学習を行なったモデルは事前学習のみのモデルと比較して，正解率が高くなることが示された．我々の研究では，この手法をインターネットスラングを含む SNS 上のテキストに対する感情分析タスクに対して利用する．

2.4 追加学習

本節における追加学習とは，特定の領域に特化したモデルを作成するために特定のドメインの情報を含む新たなコーパスを利用して既存のモデルを追加で学習することである．追加学習は，Gururangan ら [6] によって効果の分析が行われている．[6] では，タスクが属するドメインのコーパスによる追加事前学習 (DAPT) と，ターゲットとなるタスクのデータによる追加事前学習 (TAPT) の 2 つの手法が提案されている．各データとコーパスの関係，および各手法の概要は図 2 および図 3 の通りである．

[6] の結果は，DAPT および TAPT をモデル適用することで精度の改善が見られることを示しており，タスクに関連した追加のコーパスで事前学習することによる有用性を示した．本研

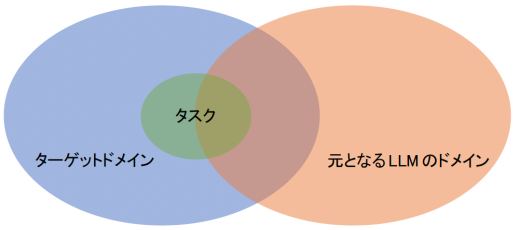


図 2　ドメインとデータの関係の概要

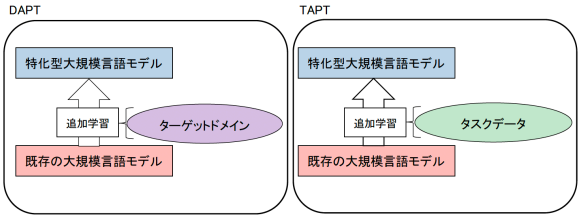


図 3　DAPT と TAPT の概要

究の追加学習では，感情分析タスクで利用するテキストではなく，同じドメインに属するインターネット上のテキストによるコーパスを利用して追加学習を行う．したがって本研究では，[6] で有用性が示された DAPT が，日本語大規模言語モデルを用いた感情分析タスクに対しても有用であるかを確かめる．

既に，特定の領域に対して追加学習が行われた大規模言語モデルはいくつか提案されている．例としては，EU，米国，英国の法律に関する文章で追加学習を行なった LEGAL-BERT [7] が挙げられる．LEGAL-BERT は，タスクを含む法律ドメインの文書で追加学習が行われており，図 3 の DAPT を適用したモデルに該当する．また，SNS 上のテキストを用いた追加学習が行われたモデルとしては，RoBERTa を用いて英語のツイートコーパスを利用した追加学習が行われた BERTweet [8] が挙げられる．[8] のモデルも LEGAL-BERT と同様に図 3 における DAPT を適用したモデルに該当する．

3　提 案 手 法

本研究では，既存の事前学習済みモデルを元として，インターネットスラングを含む SNS 上のテキストに特化したモデルへの改良を行う．また，語彙の追加および追加学習を行なったモデルとどちらも行わなかったモデルのどちらが SNS 上のテキストに対する感情分析タスクにおいて良いモデルであるかを比較する．

3.1　モデルの概要

語彙の追加および追加学習を行うモデルの全体構成を図 4 に示す．ここでは，既存の大規模言語モデルによる感情分析の一般的な手法に対して大きく 3 つのプロセスを追加する．

1 つ目は，事前学習済み BERT モデルで誤識別されるインターネットスラング等の特殊な表現を集めたインターネットスラング辞書の作成である．事前学習済み BERT モデルを利用し，データセットに対してトークン化を行う．トークン化の際

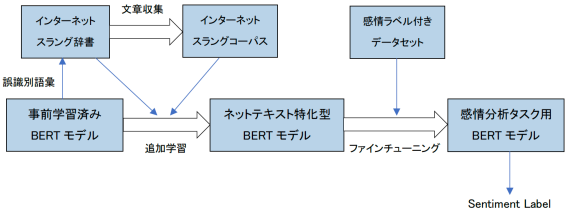


図 4　提案モデルの全体構成

に誤分割されたインターネットスラング等の特殊な表現を収集してインターネットスラング辞書を作成し，ユーザ辞書として事前学習済みモデルのトークナイザに追加し，モデルの embedding 層に追加した語彙の単語ベクトルを追加する．

2 つ目は，インターネットスラング辞書に追加した語彙の意味を表す適切な単語ベクトルを獲得するために，追加学習を行う際に用いるインターネットスラングコーパスの作成である．インターネットスラング辞書の中に含まれる語彙を含む文章をインターネット上から収集することで，インターネットスラング辞書に対応したコーパスであるインターネットスラングコーパスを作成する．

3 つ目は，インターネット上のテキストに特化するための作成したインターネットスラングコーパスを利用した追加学習である．2 つ目で作成したインターネットスラングコーパスを利用し，ユーザ辞書追加済みの事前学習済み BERT を追加で学習することで，モデルのインターネット上のテキストに対する識別精度を向上させる．

以上のプロセスで作成したインターネット上のテキストに特化したインターネットテキスト特化型 BERT を，感情ラベル付きのデータセットを利用して感情分析タスク向けにファインチューニングを行うことで，インターネット上のテキストに対して感情分析を行うことができる BERT モデルを作成する．

3.2　辞書の作成

本研究では，事前学習済み BERT モデルをインターネット上のテキストに対して特化させるために，既存のトークナイザでは正しいトークンに分割できない語彙をユーザ辞書として独自にモデルに追加する手法を提案する．

まず，感情分析タスクへのファインチューニングに使用するデータセットのテキストに対して，事前学習済みモデルに対応するトークナイザでトークン化を行う．その結果に基づいて，正しいトークンに分割できていない語彙およびフレーズを手動で抽出し，インターネットスラング辞書のリストに追加し，リスト内の全ての語句をトークナイザに追加する．追加した辞書のエントリーは，トークナイザの仕様上，既存のトークナイザの辞書のエントリーよりも優先されるため，このプロセスにより追加した語句はユーザの想定に合わせて分割されるようになる．本研究では，感情分析用ラベル付きデータセットより，正しくトークン化されなかった語彙およびフレーズを 98 個抽出し，インターネットスラング辞書として事前学習済みモデルのトークナイザに追加した．ユーザ辞書を追加した前後のトーク

ン化結果の変化の一例は以下の図 5 の通りである．

元の文章： 毎日カプヌを食べる。(※カプヌ:カップヌードルの略語)

ユーザ辞書追加前: [CLS] 毎日 カプ ヌ を 食べ る 。
ユーザ辞書追加後: [CLS] 毎日 カプヌ を 食べ る 。

図 5　ユーザ辞書追加によるトークナイザ出力の変化

次に，モデル側においてモデル内の embedding 層に，インターネットスラング辞書からトークナイザへ追加した語句のベクトルを追加する．本研究では，事前学習済み BERT モデルの embedding 層における単語ベクトルは 32000 × 768 次元であるため，インターネットスラング辞書に含まれる 98 語に合わせて embedding の行数を追加して 32098 × 768 次元となるように設定する．

以上の手法をまとめると図 6 の通りとなり，事前学習済みモデルがインターネットスラング等の特殊表現を正しくトークン化できるようになる．

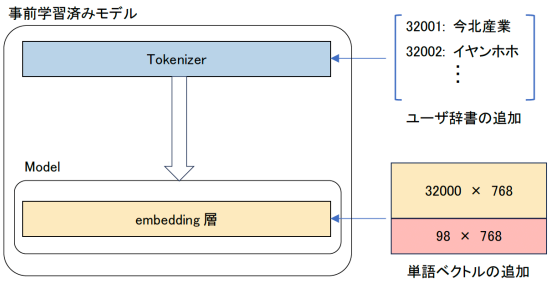


図 6　辞書追加手法の概要

3.3　コーパスの作成

3.2 節において，モデル側において追加された語句に対する単語ベクトルは Transformer の仕様上ランダムなベクトルとなっており，この単語ベクトルを適切な意味を示すベクトルへ更新する必要がある．そこで本研究では，追加した各語句を含む文章を収集し，これらをまとめてインターネットスラングコーパスとして追加学習を行うことで意味のあるベクトルを獲得する手法を提案する．

まず，インターネットスラング辞書に含まれる各語句を含む文章を収集する．本研究では，ウェブサイト上のテキストを各単語に対するコーパスとして収集する．収集するテキストの一例として，図 7 を示す．図 7 に示すように，本研究では，特にウェブサイト上における検索ワードを含むウェブサイトの内容の抜粋であるスニペットをコーパスとして収集する．



図 7　ウェブサイト上のスニペットの一例 (赤枠)

また，本研究ではインターネットスラング辞書によって追加した各単語が追加訓練に与える影響を等しくするために，各語句に対して収集する文章の数を 50 で統一する．

次に，テキストに対して前処理を行う．ウェブサイト上のテキストにはハッシュタグ (#) 等の記号や URL，HTML タグなど文章ではない要素も含まれるため，収集したテキストに対して以下のような前処理を行いコーパス内のノイズを減少させる．

- 日本語のテキストに限定
- テキスト全体を全角に統一 (半角文字を全角に変換)
- HTML タグ，JavaScript コードの除去
- 日時表記 (2023/10/25 12:11:58 等) の除去
- 記号 (#, 【】, $, %, ... 等)，URL の除去

上記に示した前処理を行なった文章を，スニペットごとに行を区切ってインターネットスラングコーパスとして JSON 形式で保存する．

3.4　追加学習

3.2 節において作成したインターネットスラング辞書を含む事前学習済みモデル，および 3.3 節で作成したインターネットスラングコーパスを用いて，追加学習を行う．BERT の事前学習では，本来隠された単語を予測する Masked Language Model(MLM) タスクと入力した 2 つの文章が意味的なつながりを有するかどうかを予測する Next Sentence Piece(NSP) タスクの 2 つに対して事前学習を行なわれている．しかし，Liu ら [9] によって BERT の事前学習における NSP タスクの有無による事前学習後のモデルの精度への影響は小さいことが示されている．そのため本研究では，事前学習済みモデルと自作したコーパスを用いて MLM タスクに対してのみ追加学習を行う．

まず，トークナイザとしてインターネットスラング辞書の語句を追加したトークナイザを，モデルとして事前学習済み BERT モデルを読み込む．本研究では第 1 章でも述べた通り，事前学習済み BERT モデルとして東北大学自然言語処理研究グループによって提供されている BERT base Japanese [1] を利用する．そのため，トークナイザとしては，BERT base Japanese に対応するトークナイザである BertJapaneseTokenizer にインターネットスラング辞書の語句をユーザ辞書として追加したトークナイザを利用する．次に，3.3 節で作成したコーパスの文章全体に対して，選択したトークナイザを用いてトークン化を行う．そして，トークン化したコーパスを入力データセットとして選択したモデルに入力して MLM タスクにおいて追加学習を行う．なお，本研究では学習の損失関数として Cross Entropy Loss(CE loss) を採用し，CE loss が減少する方向に学習を進める．CE loss が収束した時点で学習を終了させる．

3.5　ファインチューニング

3.4 節の追加学習により獲得したインターネット上のテキストに特化した BERT モデルを，感情ラベル付きデータセットを用いてファインチューニングを行い、SNS 上のテキストを入力，テキストの感情を表す感情ラベルを出力とする感情分析 BERT モデルを作成する．

まず教師データとして，ソーシャルネットワーキングサービス (SNS) 上のテキストに対して手動で「POSITIVE(肯定的)」「NEGATIVE(否定的)」「NEUTRAL(中立的)」のうちいずれかのラベルを付与したデータセットを利用する．ここで，各ラベルを付与する基準は，本研究では以下の通りに定める。各ラベルのいずれかの条件を満たすテキストに対してそのラベルを付与する．なお，POSITIVE ラベルと NEGATIVE ラベルの条件を両方満たす文章 (例：「A は苦手だが，B は好きだ」(前半は NEGATIVE，後半は POSITIVE)) の場合は，使用した事前学習済みモデルの特性を考慮し後半のテキストが満たす条件を優先してラベル付けを行なった．

- POSITIVE ラベル
  - 肯定的なフレーズが含まれるテキストである (可愛い，凄い，いいね，など)
  - 感謝や祝いを述べるフレーズが含まれるテキストである (ありがとう，おめでとう，など)
  - 繰り返し再生していることを示すテキストである (また，○回目，毎日，など)
- NEGATIVE ラベル
  - 否定的なフレーズが含まれるテキストである (つまらない，嫌い，苦手，など)
  - 他視聴者に不快感を与えるテキストである (○○しろ，○○しないやつはダメだ，など)
- NEUTRAL ラベル
  - テキストの意図が汲み取れない迷惑系のテキストである (絵文字のみ，ランダムな文字の羅列，など)
  - P/N どちらの条件にも当てはまらないテキストである

次に，感情ラベルを付与したテキストデータセット全体に対してトークナイザを適用する．トークナイザは追加学習と同様に，BertJapaneseTokenizer にインターネットスラング辞書を追加したトークナイザを利用する．トークナイザによって各テキストは，追加されたインターネットスラング辞書の語句を含むトークンに分割され，先頭に [CLS] トークンが，末尾に [SEP] トークンが付与される．

次に，トークン化を行なったテキストデータセットを追加学習済みのインターネットテキスト特化型モデルに入力し，各テキストに適切な感情ラベルとスコアをモデルによって出力するようにモデルをファインチューニングする．ここで，損失関数としては追加学習と同様に Cross Entropy Loss (CE Loss) を採用し，学習用データセット内のテキストにモデルによって付与されたラベルと正解ラベルから CE loss を導出して損失を求める．そして，誤差逆伝播法を利用して学習と評価におけるこの損失を小さくするように学習を進める．3.4 節と同様に，学習データセットおよび評価データセットにおける CE Loss の値が収束したところで学習を終了する．

4 実験・評価

本研究では，追加学習によって，語彙として追加したインターネットスラング等の特殊な表現に対して正しく意味を示すベクトルが得られているかを評価した．また，SNS 上のテキストに対する感情分析タスクにおいて，追加学習とファインチューニングを行なった提案モデルとファインチューニングのみを行なったモデルのどちらがより正確にラベル付けを行うことができるかを比較し，作成した提案モデルの評価を行なった．

4.1 実験の前準備

4.1.1 インターネットスラングコーパスの作成

本研究では，インターネットスラング辞書に含まれる単語を含む文章の収集ツールとして Google が提供する Custom Search JSON API [10] を利用した．Custom Search JSON API とは，検索エンジンより特定の検索ワードに対する検索結果を JSON 形式で取得することができる API である．本研究では，検索エンジンとして Google を選択し，検索ワードとしてインターネットスラング辞書の各単語を指定してコーパスを作成した．3.3 節でも述べたように，検索ワードを含むウェブサイト内コンテンツの抜粋文であるスニペットを学習用のテキストとして収集した．収集したテキストに対して，3.3 で示した前処理を適用することで追加学習に入力する文章を作成した．

本研究では，1 つの語句に対してスニペットを 50 ずつ収集し，合わせて 4900 のスニペットを含むインターネットスラングコーパスを作成した．

4.1.2 感情ラベル付きデータセットの作成

感情分析タスクへのファインチューニングにおいて利用するラベル付きデータセットは，1 つのテキストに対して 1 つの感情ラベルが対応しているデータセットである．感情ラベル付きデータセットに含まれるテキストは，Google によって提供されている YouTube Data API v3 [11] によって YouTube に投稿されている動画の最新順のコメントから収集する．インターネットスラングを含むコメントは，インターネット上で流行している動画に多く存在すると考え，本研究では 2023 月 10 月 1 日〜2023 年 10 月 31 日における YouTube の急上昇ランキングに基づいて，日本で流行している動画を 3 本選択し，各コメント欄から 3000 件ずつコメントを収集した．収集したコメントに対して，3.3 節で述べた追加学習用コーパスに適用した前処理と同様の前処理を適用し，これをファインチューニングの入力テキストとした．前処理を実行した結果，データセットとして利用できるコメントは合計 7458 件となった．

次に，前処理済みのテキストに対して手動で感情ラベルを付与した．感情ラベルの定義は 3.5 節で述べた通りであり，各コメントに対して適切な感情ラベルを 1 つ付与した．各感情ラベルごとのテキストの数は以下の表 1 の通りとなった．

表 1　感情ラベル付きデータセットの内訳

動画	POSITIVE	NEUTRAL	NEGATIVE
動画 1	1096	931	236
動画 2	1445	829	215
動画 3	591	1223	892
合計	3132	2983	1343

そして，テキストデータセットを学習用データセット 80%,

検証用データセット 10%，評価用データセット 10%となるように分割する．なお，各ラベルが頻出するデータセットの範囲が異なるため，データセットはランダムに分割する．

4.2 追加学習によるベクトルの獲得

追加学習におけるパラメータは以下の表 2 の通りに設定した．

表 2 追加学習におけるパラメータ

パラメータ	値
batch size	32
loss function	CE loss
tokenizer max length	512
learning late	2e-5
weight decay	0.01
number of training epocs	50

学習の損失関数としては，3.4 節で述べたように CE loss を採用した．学習曲線を図 8 に示す．

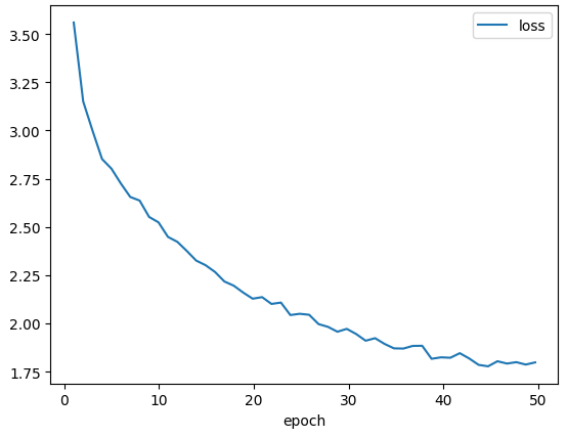


図 8 追加学習における学習曲線

学習開始時点の Loss は 3.56 と高いが，学習を進めるにつれて Loss は低くなり，学習完了時には 1.78 と学習開始時点と比較して低くなっており，事前学習済みモデルに対して MLM タスクを利用した追加学習が行えていることが分かる．

次に，追加学習により，インターネットスラング辞書として追加した語句に対するベクトルがどのように変化したかを分析する．追加した語句の一例として，イヤホンを意味するインターネットスラングである「イヤンホホ」に関する類義語の追加学習前後の変化を確認する．表 3 が追加学習前に類義語として挙げられた語句，表 4 が追加学習後に類義語として挙げられた語句である．

表 3 追加学習前の類義語

類義語	類義語の予測確率
[UNK]	0.52
演劇	0.14
体育	0.04
語	0.02
芸能	0.01

表 4 追加学習後の類義語

類義語	類義語の予測確率
ジャズ	0.12
音楽	0.11
バレエ	0.07
漫画	0.07
ギター	0.05

表 3 および表 4 より追加学習の前後で類義語が変化していることが分かる．追加学習前は単語に与えられたベクトルがランダムであったため [UNK] トークンが類義語として高い予測精度で出現しているが，追加学習後の類義語にはイヤンホホと同様に音楽関連の語句である「ジャズ」や「バレエ」といった単語が出現していることから，追加学習によって単語に対して適切な意味のあるベクトルが獲得できたことが分かる．

4.3 感情分析の性能比較

感情分析タスクへのファインチューニングにおけるパラメータは以下の表 5 の通りに設定した．

表 5 ファインチューニングにおけるパラメータ

パラメータ	値
batch size	16
loss function	CE loss
learning late	2e-5
weight decay	0.01
Dropout	0.1
number of training epochs	5

ファインチューニングによる感情分析タスクの性能比較評価として，事前学習済みモデルとして東北大学 BERT [1] を利用する 2 つのモデルを定義する．1 つ目は追加学習は行わず，辞書の追加とファインチューニングのみを行うモデルであり，$\text{BERT}_{\text{finetune}}$ モデルと名付けた．2 つ目は第 3 章にて提案した，辞書の追加と追加学習，ファインチューニングを全て行うモデルであり，$\text{slangBERT}_{\text{finetune}}$ モデルと名付けた．それぞれに対して，同じ感情ラベル付きデータセットを利用し，同じパラメータを利用してファインチューニングを行い，学習用データセットに対する損失 (Training Loss) と検証用データセットに対する損失 (Validation Loss) の比較と，評価用データセットに対する精度 (Accuracy) と Macro-F1 スコアの比較を行なった．結果は表 6 および表 7 の通りとなった．

表 6 感情分析タスクにおける Training Loss と Validation Loss

	$\text{BERT}_{\text{finetune}}$	$\text{slangBERT}_{\text{finetune}}$
Training Loss	0.29	**0.07**
Validation Loss	1.67	**1.06**

表 7 感情分析タスクにおける Accuracy と F1-Score

	$\text{BERT}_{\text{finetune}}$	$\text{slangBERT}_{\text{finetune}}$
Accuracy	0.76	**0.79**
F1-Score	0.76	**0.78**

また，正解データの POSITIVE ラベル，NEUTRAL ラベル，NEGATIVE ラベルのデータに対して，slangBERT$_{\text{finetune}}$ モデルがどれだけ正しく分類を行えたかを分析するために，図 9 のように混同行列を表示した．

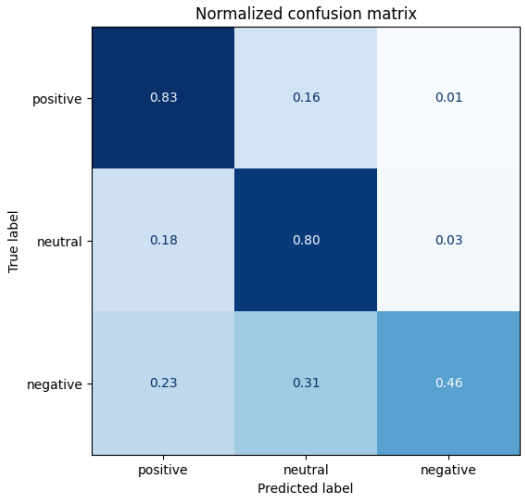


図 9　slangBERT$_{\text{finetune}}$ モデルにおける混同行列

これらの結果より，辞書の追加と追加学習，ファインチューニング全てを行った slangBERT$_{\text{finetune}}$ モデルの方が同じ学習回数でも Training Loss，Validation Loss ともに低くなっていることが分かる．また，Accuracy と Macro-F1 スコアに関しても slangBERT$_{\text{finetune}}$ モデルを用いて感情分析タスクを解いた場合の方が，ともに高くなることが分かった．さらに，POSITIVE ラベル，NEUTRAL ラベル，NEGATIVE ラベル全てにおいて，正しくラベル付けが行えているテキストの割合が最も大きくなっていることが分かった．

5　考　　察

5.1　追加学習によるベクトルの獲得

図 8 より，インターネットスラング辞書の単語を含むコーパスへの MLM タスクにおける Loss が低下していることから，追加学習によって事前学習済み BERT モデルがインターネットスラング等の特殊な表現も予測できるようになったことが分かる．また，表 3 および表 4 より，追加学習によってインターネットスラング辞書として追加した語句のベクトルが，適切な意味のあるベクトルに更新されていることが分かる．つまり，追加学習によってインターネット上のテキストを用いたタスクに特化した BERT モデルが作成できたことが示された．

その一方で，表 3 および表 4 に示される追加した語句の類義語のリストには，一部その語句の意味と関係のない語句も含まれていることが分かる．このような結果となった要因としては，追加学習に用いたコーパスの前処理が不十分であった点，およびコーパス内にその語句が正しく使われていない文章が一部含まれていた点が挙げられる．本研究では，3.3 節で示した前処理をコーパスに適用しているが，ウェブサイト上のテキストは記法がさまざまであり，余計な句読点や絵文字，誤字脱字といった一般的なコーパスに含まれるテキストには存在しないようなノイズが含まれる可能性がある．これらのノイズが含まれた表現やテキストが前処理で除去されず残った状態で追加学習を行なった結果，ベクトルを更新したい語句の意味を正しく理解できないような文章が一部含まれたことでベクトルの学習が不十分となったのではないかと考えられる．また，本研究では，3.3 節でも述べたように，追加学習に用いるコーパスを Google の検索結果より取得した．検索結果には，その検索キーワードを含む全ウェブサイトが含まれており，中にはタイトルに検索キーワードが含まれているが内容は全く関連性のないウェブサイトなども含まれる．さらに，テキストとして収集したスニペットもウェブサイト内テキストからの抜粋であり，キーワードの意味を読み取れない文章となっている可能性がある．そのため，インターネットスラング辞書として追加した語句に関しても，これらのノイズとなるテキストが単語ベクトルの更新に影響を及ぼしたために，一部関係のない語句が類義語として含まれて出現したと考えられる．

5.2　感情分析の性能比較

表 7 より，感情分析タスクにおける提案モデルの Accuracy は 0.785417，Macro-F1 スコアは 0.776737 であった．また，同じエポック数で学習した場合，Validation Loss が既存手法の BERT$_{\text{finetune}}$ モデルの Validation Loss よりも約 40%減少していることが分かった．よって追加学習によってインターネット上のテキストに特化した slangBERT$_{\text{finetune}}$ モデルの方が良いモデルであることが示された．このような結果になった要因としては，追加学習によってインターネットスラング辞書として追加した単語のベクトルが適切な意味のあるベクトルに変化したことが考えられる．インターネットスラング辞書が追加された段階では，追加した語句に対応するベクトルはランダムに与えられたものであるため，既存手法による BERT$_{\text{finetune}}$ モデルではこのランダムなベクトルによって追加した語句が誤った意味でモデルに認識され，結果として，本来 POSITIVE ラベルが与えられるべきテキストに対して NEUTRAL ラベルが付与されるといった誤分類が多く発生したのではないかと考えられる．一方で前節でも述べた通り，提案モデルでは追加した単語に関連したコーパスで追加学習を行うことで，各語句に対する適切な意味のあるベクトルを獲得することができるため，提案した slangBERT$_{\text{finetune}}$ モデルでは追加した語句が正しい意味でモデルに認識されたことにより，感情分析タスクにおいても，BERT$_{\text{finetune}}$ モデルと異なり正しいラベルが付与される確率が高くなったのではないかと考えられる．

また，混同行列 (図 9) によると，正解ラベルが POSITIVE ラベルと NEUTRAL ラベルのデータに対しては，80%以上のデータが正しいラベルを付与されている一方で，正解ラベルが NEGATIVE ラベルのデータに関しては，約 45%のデータのみが正しくラベルが付与されていることが分かった．このように，3 種類のラベルによって正しく分類できたデータの割合に違いが現れた要因としては，感情分析用データセットのラベルの偏りが挙げられる．本研究で使用したデータセットのう

ち，およそ 85%のテキストデータが POSITIVE ラベルまたは NEUTRAL ラベルが付与されており，NEGATIVE ラベルが付与されたテキストデータはわずか 15%しかない．そのため，NEGATIVE ラベルが付与されたテキストデータに対して学習を十分に行うことが出来ず，結果として NEGATIVE ラベルのデータのみ正しいラベルが付与されたデータの割合が小さくなったのではないかと考えられる．

6 おわりに

本研究では，日本語の事前学習済み BERT モデルに対して独自の辞書を追加すること，ウェブサイト上のテキストを用いて追加学習を行うこと，およびラベル付きデータを用いてファインチューニングによって，インターネットスラングを含むインターネット上のテキストに特化した感情分析タスクを実行可能な BERT モデルの作成を試みた．また，提案モデルと追加学習を行わず辞書の追加とファインチューニングのみを行う既存手法によって作成された BERT モデルについて，ソーシャルネットワーキングサービスの 1 つである YouTube のテキストに対する感情分析タスクを用いて，性能の比較評価を行なった．評価の結果，既存手法を用いて作成されたモデルと比較して，提案モデルは良い成績であった．つまり追加学習によって，BERT モデルがインターネットスラング等の特殊な表現を含む文に対して，より正確に感情ラベルを付与することができるようになることが示された．

本研究における課題は 3 つある．1 つ目は，ファインチューニング時のモデルを見直すことである．4.3 節でも述べたように，感情分析タスクのためのファインチューニングでのエポック数はわずか 5 であった．本研究ではこれ以上学習すると過学習が発生した．これが要因となり，インターネットスラング辞書内の単語を含む一部のテキストに対して正しくラベルを付与できなかった可能性がある．2 つ目は，コーパスおよびデータセットを増やすことである．本研究ではデータの収集からラベル付けまでを全て 1 人で実施したため，作成できるデータセットに限度があった．この要因によって，モデルに対して十分な学習が行えていなかった可能性があり，データの収集やラベル付けを複数人で行うことでモデルの専門性を高めることができる可能性がある．3 つ目は，BERT の応用モデルを利用することである．本研究では日本語特化型大規模言語モデルとして Hugging Face 上で最も利用されている東北大学 BERT を利用したが，近年では BERT の応用モデルである RoBERTa や DeBERTa の日本語特化モデルも公開されている．RoBERTa に関しては，Liu ら [9] によって BERT よりも良い性能を発揮することが示されていることから，事前学習済みモデルにおいて応用モデルを適用することで，追加学習や感情分析タスクにおけるスコアが改善する可能性がある．

以上の課題より今後の展望としては，本研究の全体構成を基として，事前学習済みモデルとして RoBERTa 等の BERT の応用モデルを適用し，より多くのテキストを含むデータセットによって追加学習を行い，ファインチューニング時のパラメータを適宜更新していくことで，現在のモデルよりもより正確に感情ラベルを付与することが出来るモデルの作成を目指す．

謝　辞

本研究は JSPS 科研費 JP21H03496，JP22K12157，JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです．

文　献

[1] cl-tohoku/bert-base-japanese, Hugging Face, https://huggingface.co/cl-tohoku/bert-base-japanese, 最終閲覧日：2024 年 2 月 14 日.
[2] Ayman Samir, Saleh Mesbah Elkaffas, Magda M. Madbouly, “Twitter Sentiment Analysis Using BERT,” 2021 31st International Conference on Computer Theory and Applications (ICCTA), 2021.
[3] 瀧川祐貴, 永吉希久子, 呂沢宇, 下窪拓也, 渡辺誓司, 中村美子, “ソーシャルメディア言論分析の方法① ～2020 年 1 月から辞任までの安倍首相に対する Twitter 上の投稿分析を事例として～,” 放送研究と調査, 73 巻 (2023) 3 号, pp. 70–85, 2023.
[4] Da Li, Rafal Rzepka, Michal Ptaszynki, Kenji Araki, “A Novel Machine Learning-based Sentiment Analysis Method for Chinese Social Media,” 2nd Workshop on Affective Content Analysis (AffCon 2019), 2019.
[5] 三林亮太, 上田昌輝, 川原敬史, 松本直彰, 吉村拓真, 相原健郎, 神門典子, 荘司慶行, 中島悠太, 山本岳洋, 山本祐輔, 大島裕明, “文化財の特徴理解に特化した BERT モデル,” 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM 2022), 2022.
[6] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith, “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks,” The 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360, 2020.
[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos, “LEGAL-BERT: The Muppets straight out of Law School,” In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898–2904. 2020.
[8] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen, “BERTweet: A pre-trained language model for English Tweets,” In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 9–14, 2020.
[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.
[10] Custom Search JSON API, Google Cloud Console, https://console.cloud.google.com/apis/library/customsearch.googleapis.com, 最終閲覧日：2024 年 2 月 14 日.
[11] YouTube Data API v3, Google Cloud Console, https://console.cloud.google.com/apis/library/youtube.googleapis.com, 最終閲覧日：2024 年 2 月 14 日.

Evaluating the Performance of ChatGPT for Aspect-Based Sentiment Analysis

Yifei WANG † Mizuho IWAIHARA ‡

†Graduate School of Information, Production and Systems, Waseda University 2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135 Japan

‡Graduate School of Information, Production and Systems, Waseda University 2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135 Japan

E-mail: † yifeiwang@asagi.waseda.jp, ‡ iwaihara@waseda.jp

Abstract Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment classification task, which is a longstanding research interest in the field of opinion mining. It is challenging because a sentence may contain multiple aspects or complicated relations. With the advent of large language models (LLMs) such as ChatGPT, there is a great potential for their application to ABSA task. This paper examines ChatGPT, equipped with vast contextual understanding, contributing to the nuanced analysis of opinions and sentiments, providing a comprehensive overview of their capabilities in addressing the complexities of ABSA. To this end, in this work, we evaluate performance and compare the results against baseline models trained on the understanding of opinions, sentiments, and emotions contained in the text. We further present several case studies to gain a deep comprehension of its sentiment analysis capabilities.

Keyword Sentiment Analysis, Aspect-based sentiment analysis, Large Language Models

1. Introduction

As part of NLP, aspect-based sentiment analysis (ABSA) aims to identify the sentiment polarity towards a given aspect in the sentence. Since social media and online review platforms have gained significant popularity recently, ABSA has gained wide attention [10]. In comparison to sentence-level or document-level, ABSA offers a more detailed comprehension of the opinions and attitudes conveyed in the text [2].

The ABSA task typically consists of four core elements: 1) Aspect term abbreviated as Aspect, denoting the aspect terms referring to the opinions word or phrase that have been expressed in the sentence; 2) Aspect category abbreviated as Category, representing the pre-defined category of concerned aspects; 3) Opinion term abbreviated as Opinion, indicating the opinion words or phrases towards aspects; 4) Sentiment polarity abbreviated as Sentiment, selected from predefined sentiment polarity set (such as: positive, negative, neutral) towards aspects [4]. For example, the laptop review “Looks nice and the surface is smooth, but certain apps take seconds to respond.”, we can extract the aspect, category, opinion, and sentiment polarity as [surface, Design, smooth, Positive], [apps, Software, NULL, Negative].

Large language models (LLM), based on generative pre-trained transformers (GPT), such as ChatGPT revolutionize the approach in artificial intelligence to human-model interaction and show high efficiency in the analysis of complex texts [2]. Since OpenAI released ChatGPT, it has attracted widespread public attention. It is trained based on GPT-3.5 by aligning the model to human preference, namely reinforcement learning from human feedback (RLHF) [15]. Large Language Models (LLMs) have shown great promise as highly capable AI assistants that excel in complex reasoning tasks requiring expert knowledge across a wide range of fields, including in specialized domains such as programming and creative writing. They enable interaction with humans through intuitive chat interfaces, which has led to rapid and widespread adoption among the general public [12].

Large language models (LLM) also have had an impact on sentiment analysis related tasks. So, evaluation of LLMs’ open-domain sentiment analysis capabilities across multiple domains is still needed [4]. We intend to investigate how far LLM can understand and extract the aspect, category, opinion, sentiment polarity contained in the text. To answer this question, we conduct a preliminary evaluation on four representative aspect-based sentiment analysis tasks and 8 benchmark datasets. We compare ChatGPT with fine-tuned BERT-based models and corresponding SOTA models on each task for reference.

Similar evaluations have been done in other papers, [18] evaluating the performance of a large language model for Sentiment Classification, Aspect-based Sentiment Analysis, and Multifaceted Analysis of Subjective Text, respectively. [14] also analyzed ChatGPT on Sentiment Analysis, and his task is mainly Sentiment Classification, Aspect-based Sentiment Analysis, and Emotion Cause Analysis.

In this paper, we are evaluating ChatGPT's performance on the ABSA task, we have the same sub-tasks ASTE and ASQP as [18] on

ABSA, but we additionally chose the ACD sub-task in our paper, focusing on the performance of ChatGPT on the ABSA task into one to four multivariate groups on the output, and exploring the effect of the change of temperature on the results.

2. Related Work

2.1. Aspect-based sentiment analysis

Sentiment analysis (SA) also called opinion mining or contextual mining of text is a natural language processing (NLP) technique. SA determines the polarity of text data as positive, negative, or neutral by identifying and extracting subjective information from the original material. SA prompts businesses to monitor details of the product sentiments based on the feedback of customers to obtain knowledge about their needs. Various types of SA include Graded Sentiment Analysis, Emotion Detection, and Aspect-based Sentiment Analysis. Three key elements of SA are feelings, trends, and value. Nowadays, companies use sentiment analysis for all purposes from marketing to customer service [8].

Sentiment analysis can be scrutinized at three levels namely, document-level SA which classifies the whole document view to particular sentiments for any item or assistance, sentence-level SA finds even if reviews and comments of every sentence represent positive, negative, or neutral opinions, for a item or assistance and aspect-based sentiment analysis (the deep direction moves towards more granular tasks) [1].

ABSA is essential for several reasons [8]: Granularity (ABSA provides a more granular understanding of the sentiment in the text), actionable insights, competitive advantage, customer satisfaction, and brand reputation. Overall, ABSA is a crucial tool for businesses to attain a more profound comprehension of customer sentiment, detect areas for enhancement, and make data-driven judgments to improve their products or services.

The ABSA task has many subtasks, and we have chosen three of them for evaluation here：

Aspect Category Detection (ACD). Aspect Category Detection (ACD) is an important subtask of fine-grained sentiment analysis. The ACD task aims to detect the aspect categories mentioned in a review sentence from a predefined set of aspect categories. For example, given the sentence “The service is good although rooms are pretty expensive.”, the ACD task is to detect two aspect categories from the sentence, respectively service and price [16]. ACD will facilitate access to viewpoint information for users and make it more convenient to make decisions. As in practical scenarios, user reviews are generally diversified and contain more than one aspect, so the task of multi-label aspect category detection becomes essential.

Aspect Sentiment Triplet Extraction (ASTE). Aspect Sentiment Triplet Extraction (ASTE) is the task of extracting the triplets of target entities, their associated sentiment, and opinion spans explaining the reason for the sentiment [15]. The ASTE task further extracts the opinion terms on the basis of the ABSA task, which provides an explanation for the predicted sentiment on certain aspects. Therefore, the final target of ASTE is to extract the (aspect, opinion, sentiment) triplet for a given text. For example, given the sentence “The camera of this phone is excellent, but the battery life is disappointing," ASTE task would involve identifying aspects like "camera" and "battery life," extracting opinion terms ("excellent" and "disappointing"), and classifying the sentiment polarity for each.

Aspect Sentiment Quad Prediction (ASQP). The Aspect Sentiment Quad Prediction (ASQP) task was introduced to provide a complete aspect-level sentiment structure, namely (category, aspect, opinion, sentiment) quadruple. ASQP aims to jointly detect all sentiment elements in quads for a given opinionated sentence, which can reveal a more comprehensive and complete aspect-level sentiment structure [18]. By introducing an additional aspect category element, it can still provide useful information when the aspect term is not explicitly mentioned [18]. For example, in a restaurant review, ASQP would involve not only identifying aspects like "food" and "service" but also categorizing them into broader categories. Such as, the aspect "food" could be categorized into the category "cuisine," and sentiment would be classified accordingly.

2.2. Large language models

With the advent of GPT-3, large language models (LLMs) began to come into people’s sight. They generally possess a large number of model parameters and are trained on extremely large amounts of corpora, including but not limited to BLOOM [13], LLaMA [12], Falcon [9], and so on. The capabilities of LLMs are remarkable considering the seemingly straightforward nature of the training methodology. Auto-regressive transformers are pretrained on an extensive corpus of self-supervised data, followed by alignment with human preferences via techniques such as Reinforcement Learning with Human Feedback (RLHF). As a result, given a concise instruction, LLM can adapt directly to a new task without training. In addition to the task instruction, the predictions will be more accurate if appropriate prompts are provided, an ability known as in-context learning [3].

Lately, OpenAI released the Chat Generative Pre-trained Transformer (ChatGPT) [14] and revolutionized the approach in artificial intelligence to human-model interaction, which is a chatbot fine-tuned from GPT-3.5 via reinforcement learning from human feedback (RLHF), drawing increasingly great attention.

Concurrent to us, Zhang et al. [18] evaluate sentiment analysis, aspect-based sentiment analysis, and multifaceted analysis of subjective texts of LLM, and Wang et al. [14] evaluate if ChatGPT can already serve as a universal and well-behaved sentiment analyzer.

In this work, we aim to evaluating the performance of ChatGPT for Aspect-Based Sentiment Analysis task.

3. Experiments

3.1. Task introduction

We divide experiments into four types: Aspect-Based Sentiment Analysis (ABSA), Aspect Category Detection (ACD), Aspect Sentiment Triplet Extraction (ASTE), and Aspect Sentiment Quad Prediction (ASQP).

Aspect Category Detection (ACD). Aspect Category Detection (ACD) is an important subtask of fine-grained sentiment analysis. The final target of ABSA is to extract the **(category)** for a given text.

Aspect-based Sentiment Analysis (ABSA). ABSA is the task of extracting both the aspect and its corresponding sentiment polarity simultaneously. The final target of ABSA is to extract the **(aspect, opinion)** duplet for a given text.

Aspect Sentiment Triplet Extraction (ASTE). The ASTE task further extracts the opinion terms based on the ABSA task, which explains the predicted sentiment on certain aspects. Therefore, the final target of ASTE is to extract the **(aspect, opinion, sentiment)** triplet for a given text.

Aspect Sentiment Quad Prediction (ASQP). ASQP task was introduced to provide a complete aspect-level sentiment structure, namely **(category, aspect, opinion, sentiment)** quadruple.

This is divided into the ordinary ABSA task and its subtasks. Separately, the ChatGPT can be evaluated to determine its capabilities for different levels of complexity on separate ABSA tasks.

Output examples of the different subtasks can be seen as a difference in Table 1.

3.2. Datasets

We use four datasets for ACD and ABSA tasks, including SemEval-2014, SemEval-2015, and SemEval-2016 [10]. In recent years, datasets for multi-element extraction have been proposed for ABSA subtasks. The four preprocessed datasets in [15] were built upon SemEval datasets for the ASTE task. Two datasets [17] were built upon SemEval datasets for the ASQP task.

There are many categories within the SemEval datasets of different years, and we will only analyze the Restaurant (abbreviated as Rest) datasets and the Laptop (abbreviated as Lap) datasets here.

3.3. Experimental Settings and Baselines

Models. ChatGPT used for evaluation is gpt-3.5-turbo(Currently points to gpt-3.5-turbo-0613), the most capable and cost-effective GPT-3.5 model. We set the temperature to 0 and 0.5, making the outputs mostly deterministic for the identical inputs. Then we let ChatGPT generate answers for a manually constructed task instruction for each task to elicit its ability to the corresponding task. In the task, we will use several different prompts to select the best and most reasonable one according to the results. Additionally, we will observe the impact of varying the number of examples provided (1-shot and 5-shot) on the performance. The originally selected prompts are as shown in Table 1.

We did three experiments and the results are averaged to obtain a more accurate assessment.

Baseline Details. The baseline model is the T5-Base model, whose hyperparameters, metrics and code references the [5] but on a much smaller batch size and epochs. T5 is capable of processing multiple tasks in a unified text-to-text format of the T5 model， so we feed the sampled data into the model as an input sequence and train it to generate a target sequence.

3.4. Evaluation

We adopt accuracy and macro F1 score for evaluating each task.

Accuracy. Accuracy is a metric used to measure the overall correctness of a classification model. It is calculated by dividing the number of correctly predicted instances by the total number of instances in the dataset. The definition of accuracy is:

$$accuracy = \frac{\text{Number of Correct Predictions}}{\text{total Number of Predictions}} \quad (1)$$

A higher accuracy value indicates that the model is making more correct predictions across all classes. However, accuracy might not be the best metric when dealing with imbalanced datasets.

Task	Prompt
	Your task is to do Aspect Category Detection task which is to identify the aspect or category being discussed in the text. Please extract relevant aspects in [place, food, service, staff, miscellaneous, price, menu, ambience] which is identifying the aspects in the text of restaurant dataset. No explanation.

ACD	Example: text: Sure, the burgers are good, but not good enough to make up for the abominable service and disgusting, filthy atmosphere. output format like: [{{'category': 'food'}}, {{'category': 'ambience'}}]
ABSA	Your task is to do Aspect-based Sentiment Analysis task which is to identify and analyze sentiments associated with specific aspects or entities mentioned in the text. Answer yes/no on whether the text carries each aspect in [place, food, service, staff, miscellaneous, price, menu, ambience], and what is each polarity? No explanation. example: text: Sure, the burgers are good, but not good enough to make up for the abominable service and disgusting, filthy atmosphere. output format like: [{{' aspect": 'food', 'sentiment': 'positive'}}, {{' aspect": 'ambience', ' sentiment ': 'negative'}}]
ASTE	Your task is to do Aspect Sentiment Triplet Extraction task which is extracting the aspect, opinion and sentiment being discussed in the text. Aspect and opinion should be extracted of the sentence and sentiment should be one of ['negative', 'neutral', 'positive']. No explanation. example: text: Everything is so easy to use, Mac software is just so much simpler than Microsoft software. output format like: [{{'aspect': 'Mac software'}, {'opinion: 'esay'}, {' sentiment': 'positive'}}, {{'aspect': 'Microsoft software'}, {' opinion: 'simpler than'}, {' sentiment': 'negative'}}, {{' aspect': 'use'}, {' opinion: 'easy'}, {' sentiment': 'positive'}}]
ASQP	Your task is to do Aspect Sentiment Quad Prediction task which is extracting the category, aspect, opinion and sentiment being discussed in the text. Aspect and opinion should be extracted of the sentence, Category should be one of ['ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food general', 'food prices', 'food quality', 'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant prices', 'service general'], and sentiment should be one of ['negative', 'neutral', 'positive']. No explanation. example: text: The sandwiches are dry, tasteless and way overpriced. output format like: [{{'category': 'food quality'}, {'aspect': 'sandwiches'},{'opinion: 'dry'},{' sentiment': 'negative'}}, {{'category': ' food quality'}, {'aspect': 'sandwiches'},{'opinion': 'tasteless'},{'sentiment': 'negative'}},{{'category': ' food price'}, {' aspect': 'sandwiches'},{' opinion: 'overpriced'},{' sentiment': 'negative'}}]

Table 1. 1-shot prompt. The task descriptions for the three shots are consistent, although the number of 0-shot, 1-shot, and 5-shot examples are 0, 1, and 5, respectively.

3.5. Result Comparison and Analysis

The experimental results are shown in Tables 2 (Temperature=0) and Table3 (Temperature=0.5).

First, at temperature 0, it can be seen that ChatGPT performs better in processing all tasks' 5-shot models perform than both 0-shot, 1-shot, and baseline models.

With the temperature set to 0, the generation of ChatGPT will be more stable and less of randomness, which can be verified by repeated experiments.

When the temperature is set to 0.5, randomness increases in the generation, and the gap between the upper and lower bounds of the accuracy of ChatGPT generation enlarges.

At temperature 0.5, ChatGPT still relatively performs better. However, unlike temperature 0, the results by ChatGPT are distributed in a different shot. We believe this is a result of the instability introduced by increasing the temperature.

We also need to compare the results with temperatures 0 and 0. 5, and it is clear to see that the results with temperature 0 perform much better across tasks and across shots. Although the partial dataset with temperature 0.5 sometimes gives better results in a

single experiment, the final results are not as good overall as those with temperature 0 because they are averaged over three experiments.

Additionally, neither the baseline model nor the ChatGPT was fine-tuned, so they were both measured as generalized models. For a baseline model such as T5, the model is improved as well as by fine-tuning, but this is not the focus of the article and is not discussed here.

4. Conclusion and Future Work

4.1. Conclusion

In this work, we evaluate ChatGPT on a range of ABSA datasets and compare its performance to the Baseline model (T5), which helps better understand their capabilities in ABSA tasks. Also, distinguishing more precise states, such as category, aspect, opinion, sentiment, or intensities of sentiment for more complicated analysis is also challenging but worth exploring.

Although ChatGPT possesses impressive abilities in understanding emotions and inferring the emotion causes [14]. It is not ideal for complex tasks like ABSA and other subtasks. We believe that ChatGPT can give better results in simple tasks or simple scenarios rather than tasks like ABSA which are more complex from description to output. However, it is clear that one of the major advantages of ChatGPT is its generalizability, which does not require steps of finetuning the model for each task, but simply adjusting the prompts for different tasks.

Overall, the good results of ChatGPT brought a new direction into the ABSA tasks.

4.2. Future Work

We will continue to evaluate the effects of ChatGPT and GPT-4 due to its better results.

We will try improving prompts for ABSA tasks, thorough prompt chaining which combines multiple prompt templates and dynamically instantiates prompts, and chain-of-thought approach.

Dataset		0-shot	1-shot	5-shot	Baseline (T5)
		Test_acc (%)	Test_acc (%)	Test_acc (%)	Test_acc (%)
ACD	Lap14	**79.40**	64.40	77.60	56.70
	Rest14	74.80	72.08	**77.80**	69.39
	Rest15	73.40	64.80	**73.60**	73.25
	Rest16	73.41	66.22	**75.01**	71.31
ABSA	Lap14	68.40	70.01	**76.20**	72.22
	Rest14	65.40	69.15	**72.23**	67.56
	Rest15	72.60	72.69	**73.71**	69.37
	Rest16	74.60	75.72	**76.33**	71.81
ASTE	Lap14	20.43	33.23	**42.99**	33.79
	Rest14	25.81	38.82	**39.84**	37.15
	Rest15	27.64	**55.59**	49.69	36.34
	Rest16	32.52	49.69	**54.29**	36.92
ASQP	Resterant15	13.40	24.40	**24.80**	18.15
	Resterant16	17.80	29.40	**30.60**	22.23

Table 2. Performance comparison between 0-shot, few-shot, and baseline model (T5) performance on ACD, ABSA, ASTE, and ASQP tasks when Temperature=0. The best results (except for fully-supervised results and human evaluation results) are in **bold**.

Dataset		0-shot	1-shot	5-shot	Baeline(T5)
		Test_acc (%)	Test_acc (%)	Test_acc (%)	Test_acc (%)
ACD	Lap14	**76.60**	63.89	69.60	56.70
	Rest14	**71.20**	65.22	67.42	69.39
	Rest15	**75.20**	60.71	62.60	73.25
	Rest16	**71.80**	58.60	64.43	71.31
ABSA	Lap14	49.40	52.80	**71.05**	72.22
	Rest14	31.60	42.60	**73.33**	67.56
	Rest15	46.00	53.34	**70.60**	69.37
	Rest16	41.20	58.20	**77.40**	71.81
ASTE	Lap14	22.87	25.91	**39.63**	33.79
	Rest14	29.88	**38.62**	37.60	37.15
	Rest15	32.61	**53.42**	47.20	36.34
	Rest16	38.04	42.64	**52.15**	36.92
ASQP	Rest15	7.80	**20.12**	18.20	18.15
	Rest16	10.12	**29.20**	25.40	22.23

Table 3. Performance comparison between 0-shot, few-shot, and baseline model (T5) performance of ACD, ABSA, ASTE, and ASQP tasks when Temperature=0.5 and T5 for Semeval datasets. The best results are in **bold**.

References

[1] Bhoi, Amlaan, and Sandeep Joshi. "Various approaches to aspect-based sentiment analysis." arXiv preprint arXiv:1805.01984 (2018).

[2] Brauwers, Gianni, and Flavius Frasincar. "A survey on aspect-based sentiment classification." ACM Computing Surveys 55.4 (2022): 1-37.

[3] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[4] Cai, Hongjie et al. "MEMD-ABSA: A Multi-Element Multi-Domain Dataset for Aspect-Based Sentiment Analysis." ArXiv abs/2306.16956 (2023): n. pag.

[5] Chebolu, Siva Uday Sampreeth, et al. "Exploring conditional text generation for aspect-based sentiment analysis." arXiv preprint arXiv:2110.02334 (2021).

[6] Jiang, Qingnan, et al. "A challenge dataset and effective models for aspect-based sentiment analysis." Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019.

[7] Koco'n, Jan et al. "ChatGPT: Jack of all trades, master of none." Inf. Fusion 99 (2023): 101861.

[8] Liu, Haoyue, et al. "Aspect-based sentiment analysis: A survey of deep learning methods." IEEE Transactions on Computational Social Systems 7.6 (2020): 1358-1375.

[9] Penedo, Guilherme et al. "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only." ArXiv abs/2306.01116 (2023): n. pag.

[10] Pontiki, Maria, et al. "Semeval-2016 task 5: Aspect based sentiment analysis." ProWorkshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016.

[11] Sampathirao Suneetha, Prof, and S. Vizianand Row. " ASPECT-BASED SENTIMENT ANALYSIS: A COMPREHENSIVE SURVEY OF TECHNIQUES AND APPLICATIONS." Journal of Data Acquisition and Processing 38.3 (2023): 177.

[12] Touvron, Hugo et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." ArXiv abs/2307.09288 (2023): n. pag.

[13] Workshop, BigScience, et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).

[14] Wang, Zengzhi et al. "Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study." ArXiv abs/2304.04339 (2023): n. pag.

[15] Xu, Lu, et al. "Position-aware tagging for aspect sentiment triplet extraction." arXiv preprint arXiv:2010.02609 (2020).

[16] Zhao, Fei, et al. "Label-Driven Denoising Framework for Multi-Label Few-Shot Aspect Category Detection." arXiv preprint arXiv:2210.04220 (2022).

[17] Zhang, Wenxuan, et al. "Aspect sentiment quad prediction as paraphrase generation." arXiv preprint arXiv:2110.00796 (2021).

[18] Zhang, Wenxuan et al. "Sentiment Analysis in the Era of Large Language Models: A Reality Check." ArXiv abs/2305.15005 (2023): n. pag.

大規模言語モデルによる細かな感情推定手法

薛　　沛欽† 　　小林　亜樹†

† 工学院大学情報学部情報通信工学科　〒 160–0023 東京都新宿区西新宿 1–24–2
E-mail: †j020149@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　これまでの感情分析では，感情の極性や Plutchik の感情の輪にある基本的な 8 つの感情に対応する感情語から構成される感情辞書を構築し，感情語を手がかりとして感情分析を行うことが多かった．これに対して，文章の意味解析を行える可能性のある大規模言語モデル（Large language model, LLM）の台頭は，感情語の有無などだけでない意味を扱える可能性があるため，本稿では LLM を用いた細かな感情抽出手法を提案する．本手法は，公開されている LLM を利用し，事前に定めたプロンプト文と分析対象文とを LLM に入力して，Plutchik の感情の輪に含まれる 48 種類の感情を分析する．WRIME データセットを用いた LLM の違いによる分析結果などを比較し，有効性などを検証した．

キーワード　大規模言語モデル，感情分析，自然言語処理

1 はじめに

近年，ソーシャルメディアやネットショッピングの普及により，ネット上に商品やサービスに対する評判情報がよく投稿されている．これらは，企業における顧客コミュニケーションツールとしても重視されており，顧客満足度の向上や製品の改善などにも不可欠なものとなっている．一方，その分量は多いため，検索や自動集計ツールが用いられる．しかし，投稿テキストに込められた意味情報の自動読み取りは発展途上であり，特に製品等への感情の発露はポジティブ/ネガティブの 2 極性分析程度にとどまっている．そのため，テキストからより詳細な感情を推定し，定量的な感情情報を抽出することが課題である．

テキストからの感情分析は，感情極性 (ネガティブ，ポジティブ) や Plutchik の基本 8 感情 (喜び・信頼・恐れ・驚き・悲しみ・嫌悪・怒り・期待) [1] に基づき，単語と感情を対応づけた感情辞書を構築し，文書中でのこれら単語の出現によって感情抽出を行う手法が主流であった．このような手法では複雑な感情や文脈でしか捉えられない感情を詳細に分析するには限界がある．

これに対して，近年急速に発展してきた大規模言語モデル（Large Language Model，LLM）は，自然言語の意味合いを「理解」し，意味についての受け答えができるとの見方が台頭している．そのため，多分野において様々な応用研究が盛んとなっている．

本論文では，この大規模言語モデルを利用した 48 種類の感情推定手法を提案する．提案手法では，感情分析を指示する特定の指示文と分析対象のテキストを LLM に入力し，その応答から感情推定結果を抽出する．感情空間には Plutchik の感情の輪による 48 種類の感情を用いる．いくつかの LLM モデルに分析を実行させ，その結果を比較検討し，提案手法の実用性などについて議論する．評価には WRIME [2] などの感情分析データセットを用い有効性を検証する．

本論文の構成を示す．2 章は本研究との関連研究の紹介であり，3 章では，提案した手法について細かく説明する．4 章ではこの手法を利用して得られた結果を示し，最後に 5 章で，研究のまとめと今後の展望について述べる．

2 関連研究

感情分析や感情推定分野における研究は感情辞書を用いているのがほとんどであったが，近年主に BERT などの言語モデルを利用した感情分析の研究も増えている [3] [4]．中澤ら [3] は Pre-training と Fine-tuning を行った BERT を使った感情極性の推定手法を提案し，従来の形態素解析と構文解析 2 つの手法を比較し，言語モデルを使って感情推定の実行可能性を示した．Adoma ら [4] は BERT などの 4 つのモデルで 7 種類の感情を推定し，ISEAR データセットを使って結果評価を行い，その結果，最高で 0.7431 の精度を得ている．これらより言語モデルによる感情推定の可能性があることがわかる．

これらの研究 [3] [4] と本研究の違いは，他研究では BERT など現在となっては比較的小規模のモデルを使っており，抽出する感情の種類も少なめに限定されている点にある．本研究では，事前のチューニングをせずに，LlaMA2 や GPT-4 など大規模言語モデルを利用し，48 種類の感情の推定を行うことが特徴である．

3 提案手法

3.1 概　　要

本手法は図 1 のフローチャートのようにいくつかの部分を組み合わせて実行する．準備として，3.2 節で紹介する Plutchik の感情の輪より得られた 48 感情の単語および，予め決めた 3 文字の英字コードを作成しておく．感情推定は，まず，この英字コード等を含む感情推定を指示するプロンプト文（指示文）と感情を推定する対象文を連接して LLM に入力する．LLM か



図 1　提案手法の構成

らの回答文を得ると，回答文から推定結果を抽出し，最終的に 48 次元の感情推定結果を得る．

3.2　Plutchik の感情の輪

本節では，本研究で用いる感情空間である 48 種類の感情について説明する．48 感情は，Plutchik の感情の輪 [1][6] より一定の規則性で定めるものである．

Plutchik の感情の輪を図 2 に示す．図中には 3 つの同心円がある．感情は，同心円の中心から異なる種類に毎に別々の方向（角度）を取り，同心円の中心に近い側ほど同種の感情は強く，外側へ行くほど感情強度が弱まることを表している．3 重の同心円の 2 番目の円に含まれている感情が基本 8 感情で，それぞれ喜び，信頼，恐れ，驚き，悲しみ，嫌悪，怒り，期待と呼ばれている．一方，基本感情円の内外それぞれの円，すなわち，1 番目と 3 番目の円は，それぞれの基本感情の強強度と弱強度の感情を表している．例えば，中心から上方向への基本感情「喜び」（図中黄色の領域）は，内側から順に「恍惚」「喜び」「平穏」の感情語として順に感情強度が減る語として例示されている．

一方，基本 8 感情方向の中間に位置する組み合わせた感情も定義されている．例えば,「喜び」と「信頼」（右上斜め方向の黄緑色領域）を組み合わせとして「愛」が定義されている． さらに,「喜び」と「恐れ」や「喜び」と「驚き」など 1 つの方向や 2 つの方向を飛ばし組み合わせることも可能である．しかし，「喜び」と「悲しみ」のような 3 つの方向を飛ばした組み合わせはそれぞれの感情方向が正反対であるため，本理論では定義されない．また，これらの複数の方向から構成される感情については本理論には強度の概念はないものとする．

これら強度により派生してきた 24 種類の感情と，基本感情同士が組み合わせられてきた 24 種類の感情が合計 48 種類の感情は Plutchik の感情の輪の中に含まれているものである．

本研究において抽出する感情は，このように定義されている 48 感情とする．

3.3　8 種類感情のコードの割当

提案方式では，実装上の処理のため，3.2 節に説明した 48 種類の各感情を表す英語の単語から 3 文字を取り出した英文字列

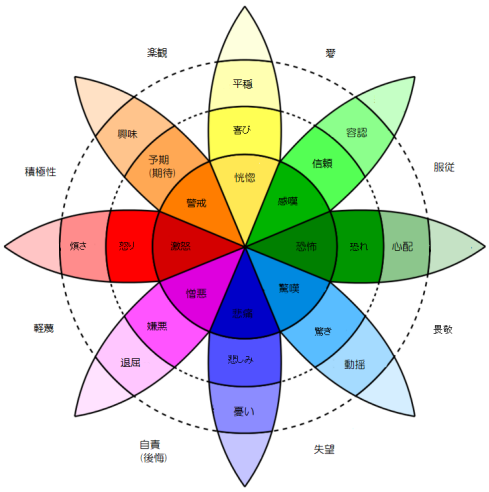


図 2　Plutchik の感情の輪

を感情コードとする．この感情と対応するコードの対照表を表 1 に示す．提案方式では，LLM に対してプロンプト文を入力することとなるが，この入力において，感情についてはここで定義した感情コードによる出力を指定するなどして利用する．これは回答文生成時間の短縮や結果の抽出をしやすくなることを企図してのものである．また，この感情コードを使用することによる生成時間の短縮はあくまで経験的に推定したものである．

表 1　感情とコード

感情	コード	感情	コード	感情	コード
喜び	JOY	恐れ	FEA	期待	ANT
愛	LOV	失望	DIS	攻撃性	AGG
悲しみ	SAD	嫌悪	DSG	信頼	TRU
服従	SUB	後悔	RMS	楽観	OPT
怒り	ANG	驚き	SPR	平穏	SER
畏敬	AWE	軽蔑	CTP	罪悪感	GLT
物思い	PEN	不安	APR	好奇心	CUR
苛立ち	ANO	病的	MRB	倦怠	BOR
誇り	PRD	シニシズム	CYN	気晴らし	DST
嫉妬	ENV	悲観	PES	興味	INT
受容	ACC	憤慨	OTR	歓喜	ECS
不信	UNB	悲嘆	GRF	羞恥	SHM
激怒	RGE	恐怖	TER	絶望	DSP
支配	DOM	感傷	SNT	嫌悪感	LOA
驚愕	AMZ	喜び	DLG	警戒	VIG
不安	ANX	感嘆	ADM	希望	HOP

3.4　プロンプト文の入力

提案方式では，判定したい文を含む適当なプロンプト文を準備し， LLM のプロンプト文として入力する．LLM による出力結果を回答文と呼び，この文から分析結果を取り出すことになる．

プロンプト文は，事前に書き込んだシステムプロンプト文，および分析するとき用いているユーザー指示プロンプト文とに分けられるが，本手法では，システムプロンプト文を使うこと

はなく，テキストを分析するごとにこの後に示す指示文，および分析したいテキストを 2 つのユーザー指示プロンプト文として連続的にモデルに入力する．

指示文では，感情推定をさせる指示のみだけでなく，指示文の一部として表 3.1 に示したような感情と対応コードの関係リストに含まれている感情に限定して回答させるようにしている．また，先行研究では英語ベースのモデル (Llama2 など) では他言語のプロンプト文より英語のプロンプト文に対しての理解力が非常に高く，一方で日本語のプロンプト文を使うことで，英語のプロンプト文を使用した場合より明らかに推定失敗率が高くなっているため，本研究の指示プロンプト文言語は英語とする．

プロンプト文

I have a list of emotions with corresponding codes. After providing a text, please analyze it to identify all the emotions it contains, using the codes from my list. Then, respond with only those codes that are present in the text. Below is the list of emotions and their codes: (感情とコード)

3.5 モデルからの回答の抽出

3.4 節で述べたようなプロンプト文を LLM に入力すると，表 2 に示すような回答文テキストが出力される．プロンプト文においては，感情コードコードのみを回答文に含むような文面としてあるが，実際に生成される回答文は，多くの場合に何らかの感情コード以外の自然言語による文を含んでいた．そのため，回答文から分析結果となる感情コードのみを抽出するために，コード抽出ソフトウェアを準備した．本抽出アルゴリズムは，回答文中に感情語や感情コードが存在しているかどうかを検索し，これらに対応する感情コードのみを出力する仕組みである．

実装コードによって抽出した結果例を表 3.2 に示す．しかし，様々な原因で正常に抽出できないこともあり，その場合は推定が失敗したテキストとして扱う．失敗例は表 6 に示す．

表 2　LLM からの回答例及び抽出例

テキスト	→ごみ捨ては出勤時 (ゴミ収集車が来るかどうかの時間帯) に俺が出すんじゃなくて、前日に私が出しといてそれはお前の仕事やろとのことだったので、まあ了承。大目に見てやってるとか、ごみ捨てはお前の仕事とか、微妙にカチンとくるけど、週 2 こちらがやることで、静かになるなら得るものの方が多い！
回答文	The emotions contained in the tweet are: ANO (Annoyance), SER (Serenity), ACC (Acceptance)
抽出結果	SER,ANO,ACC

4 評 価 実 験

本章では，提案手法の有効性を検証するために 3 章で提案した手法を用いるソフトウェアを作成し，WRIME データセットから 100 件の投稿を無作為で取り出したデータセットを用いて評価実験を行った結果，および実験方法を説明する．

4.1 実 験 方 法

本節では，評価実験に使った方法について説明する．

4.1.1 概　　要

実験の方法は，図 3 のように結果の変換と結果の評価の 2 つの部分からなる．それぞれ異なる形式のデータを評価のために計算可能とするための変換と，誤差評価を定量的に行うための手順とである．

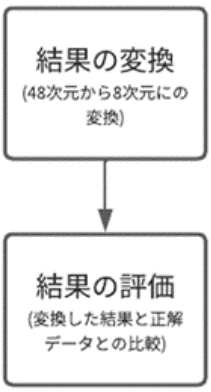


図 3　評価実験方法のイメージ

4.1.2 結果の変換

3.5 節における感情推定から得られるのは 48 次元のデータであるが，WRIME データセットにある 8 次元の正解データとは形式が異なっている．そこで，同一文に対する結果の突き合わせを行い，評価のため MSE を計算するために 48 次元である結果から 8 次元に変換する必要がある．本節では，変換の手法を疑似コード及び例を挙げて説明する．

3.2 節で説明した Plutchik の感情の輪によっては，強度変化により派生してきた感情と複合感情が異なる手法で定義されているため，変換するには，それぞれ異なる手法を使う必要がある．すなわち，これらの変換の違いは Plutchik の感情の輪における定義の違いを反映している．

Algorithm 1 のアルゴリズムを用いて，基本感情及びその強度変化による感情 (14 行目から 17 行目まで) と複数の方向の感情の複合感情の変換（18 行目から 23 行目まで）を行う．感情コードからの変換時には，感情コードがこれら 2 種類のいずれの感情であるかを判別し（2 行目から 13 行目まで），判別結果に基づき，適切な変換アルゴリズムによって変換する．

Algorithm 1 感情変換アルゴリズム

```
変換結果 ← {'JOY' : 0,' SAD' : 0, . . . ,' ANG' : 0}
強度による感情リスト ← {"JOY","SAD", . . . ,"TER"}
複合感情リスト ← {"DSP","SNT", . . . ,"ANT"}
強度による感情分類結果 ← []
複合感情分類結果 ← []
for i in 回答から抽出した結果リスト do
    if 回答から抽出した結果リスト [i] in 強度による感情リスト then
        強度による感情分類結果.append(回答から抽出した結果リスト [i])
    end if
    if 回答から抽出した結果リスト [i] in 複合感情リスト then
        複合感情分類結果.append(回答から抽出した結果リスト [i])
    end if
end for
for j in 強度による感情分類結果 do
    temp ← 強度取得 (強度による感情分類結果 [j])
    変換結果 [所属次元取得 (強度による感情分類結果 [j])] ← temp
end for
for k in 複合感情分類結果 do
    成分 ← 成分取得 (複合感情分類結果 [k])
    for 成分名 in 成分 do
        変換結果 [成分名] ← 2
    end for
end for
```

図 4，図 5，図 6 の例と合わせて説明する．例は，LLM の回答文から抽出した結果を SAD（悲しみ），APR（心配），HOP（希望）とする．図 4 は Algorithm 1 の 2 行目から 13 行目までで SAD，APR を強度による感情に，HOP を複合感情に分類する様子を表している．その中には，回答から抽出した結果リストは 3.5 節で説明した回答文から抽出した感情コードの集まりを表している．

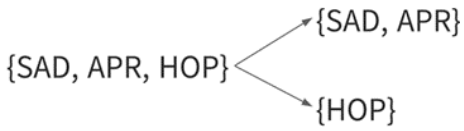


図 4　結果の変換 (分類) の例

次に図 5 では強度による感情である SAD，APR を Algorithm 1 の 14 行目から 17 行目によって変換を行い，SAD は悲しみ方向の強度 2 であり，APR は恐れ方向の強度 1 であるため，結果の悲しみと恐れそれぞれの次元にそれぞれ 2 と 1 を出力する様子である．15 行目の強度取得という関数は，引数である感情の所属している感情方向における強度を [1,3] の範囲で取得する．例として，APR を入力すると FEA(恐れ) の弱強度であるため 1 が戻る．16 行目の所属次元取得という関数は，引数である感情の所属している感情方向のコードを取得する．例えば，APR を入力すると FEA が戻ってくる．

最後に，図 6 は複合感情である HOP が，期待と信頼の 2 方向の基本感情の組み合わせ複合感情であるため，結果の期待次元，および信頼次元にそれぞれ 2 となるベクトルとなる様子である．19 行目の成分取得という関数は，引数である複合感情の 2 つの基本感情成分を取得する．変換した結果はいずれも 8 次元で各要素の値の範囲は [0,3] である．

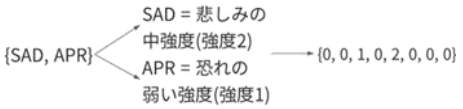


図 5　結果の変換 (強度変化による感情) の例

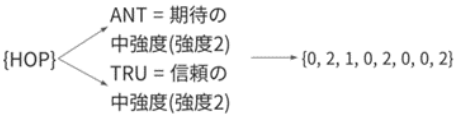


図 6　結果の変換 (複合感情) の例

4.1.3　結果の評価

4.1.2 節の手順によって抽出回答結果を変換すると，WRIME データセットの 8 次元ベクトルと同空間で比較可能となる．各投稿の結果を評価するための比較は，式 (1) のように変換した抽出結果と正解データの各 8 次元ベクトルの差分をとり，その差分ベクトルの 2 乗ノルムを計算し，次元数で割った点数 (score) を各データ（投稿）毎に求める．$\hat{y}$ と y はそれぞれ変換した結果のベクトルと正解データのベクトルを表している．ベクトルは 8 次元であるため，score の解釈のため 1/8 倍して平均値を取った．

$$\text{score} = \frac{1}{8}||\hat{y} - y||^2 \tag{1}$$

$$\text{MSE} = \frac{1}{8n}\sum_{i=1}^{n}||\hat{y}_i - y_i||^2 \tag{2}$$

また，ここでモデル間の結果を比較するために，式 (2) を用いて各モデルの MSE を計算した．ここで，n はモデル毎の分析対象となった件数であり，i は 1 から n までの何件目の投稿であるかを示している．

4.2　WRIME データセット

本論文では WRIME [1] 感情分析データセットを用いて評価実験を行う．そこで，WRIME データセットについて説明する．

WRIME データセットとは，計 43200 件の各投稿に対して，その投稿主の感情を感情値として付与しているデータセットである．感情値は，Plutchik の基本 8 感情それぞれの次元についての強度（0: 無，1: 弱，2: 中，3: 強）による 8 次元値と，それとは別に 5 段階の感情極性 (-2: 強いネガティブ，-1: ネガティブ，0: ニュートラル，1: ポジティブ，2: 強いポジティブ）によって構成される．これらの感情値は，対象の投稿の実際の投稿主による得点付けによるもののほか，その投稿テキストを第三者として選定された評価者が読んだ結果として付与したものの双方が収載されている．評価者は各 3 名が割り当てられており，結果として，1 投稿に対して 4 つの 8 次元＋ 1 極性感情値が付

与されたデータとなっている．投稿主による評価を主観，評価者による評価を客観と呼ぶ．客観の 3 名を識別する際には，客観 A，客観 B，客観 C としている．

客観評価者の平均値も特に収載されており，1 投稿について都合 5 種類のベクトル値が収載されている．表 4 はデータセット内の 1 投稿に対応する感情値のイメージである．本論文では一番下の太字で示した客観評価者による平均評価値について，Plutchik の 8 感情に対応する 8 次元ベクトル値を正解データとして用いる．

4.3 実 験 環 境

本節では，実験に使ったモデルおよびモデルの設定や実行に関わることを説明する．

4 つのモデルはそれぞれ，Llama-2-7B-Chat，Xwin-LM-7B-V0.2，GPT-3.5-Turbo，GPT-4-1106-Preview であり，同一パラメータに調整している．具体的には，Topp を 0.7 に Temperature を 0.05 に設定した．プロンプト文は 3.4 節で説明したプロンプト文を用いた．

実験に使ったソフトウェアとライブラリのバージョンは表 3 に示す．プログラムは Python 言語により，Xwin の実行には transformers ライブラリ，GPT 系のモデルは OpenAI により提供されている Python ライブラリを用いた．データの格納には MariaDB を使用した．

表 3　実 験 環 境

ソフトウェア	バージョン
Python	3.11.5
transformers	4.33.2
openai-python [5]	1.6.0
MariaDB	11.1.2

4.4 実 験 結 果

図 7-10 は LLM のモデル，GPT-4-1106-Preview，GPT-3.5-Turbo，Llama-2-7b-Chat，Xwin-LM-7B-V0.2 それぞれによる結果を示す．

横軸は MSE であり，その値の範囲は 4.1.2 節で説明した変換結果と正解データとの誤差であるため，最大 3 である．そのため，グラフで示す 2 乗値の範囲は [0,9] となる．縦軸は各 MSE の値における投稿の件数であり，Mean は分析対象となった投稿の MSE の総和の平均値を表し，Count は分析対象となった件数 (分析失敗ではない件数) を表している．

MSE は 0 に近いほど誤差が小さく精度が高いため，図中の棒が左に寄っているほど，良い結果が得られていると考えられる．したがって，精度の高い順で並べると，GPT-4-1106-Preview，GPT-3.5-Turbo，Llama-2-7b-Chat，Xwin-LM-7B-V0.2 の順となる．実際，MSE の平均値順に並べた結果も表 5 に示すが，この順序となっている．

4.5 考　　察

表 5 のデータから見ると GPT-4-1106-Preview の MSE 平均値は GPT-3.5-Turbo より低くなっているが，評価対象数も少

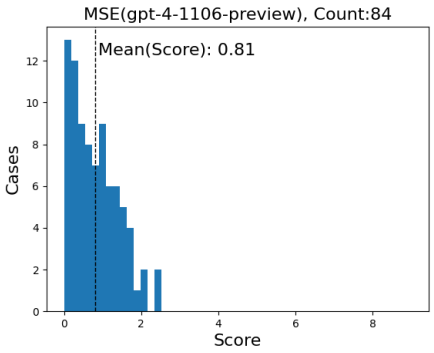


図 7　GPT-4-1106-Preview による 2 乗誤差分布（除く抽出失敗例）

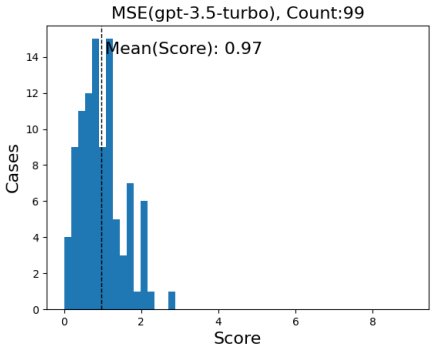


図 8　GPT-3.5-Turbo による 2 乗誤差分布（除く抽出失敗例）

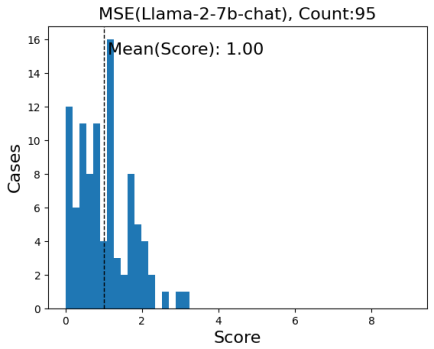


図 9　Llama-2-7b-Chat による 2 乗誤差分布（除く抽出失敗例）

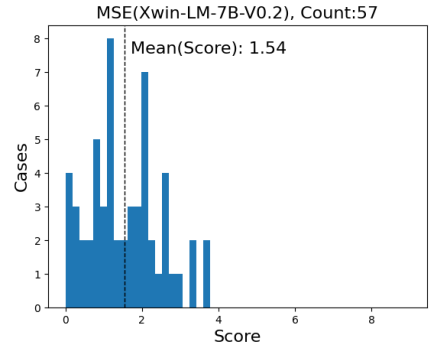


図 10　Xwin-LM-7B-V0.2 による 2 乗誤差分布（除く抽出失敗例）

なくなっている．これについて回答文を確認したところ，主にテキストには感情が含まれていないといった理由を回答しており，そのため分析していない結果として集計している．投稿テキス

表 4　WRIME データセットにおける 1 投稿に対する評価値例

	喜び	悲しみ	期待	驚き	怒り	恐れ	嫌悪	信頼	感情極性
主観	0	3	0	1	3	0	0	0	0
客観 A	0	3	0	3	1	2	1	0	-1
客観 B	0	2	0	2	0	0	0	0	-1
客観 C	0	2	0	2	0	1	1	0	-2
平均値 (客観のみ)	**0**	**2**	**0**	**2**	**0**	**1**	**1**	**0**	-1

表 5　評価実験結果

モデル	評価対象数	MSE 平均値
GPT-4-1106-Preview	84	0.81
GPT-3.5-Turbo	99	0.97
Llama-2-7b-Chat	95	1.00
Xwin-LM-7B-v0.2	57	1.54

トを見ると， 特に GPT-4-1106-Preview は曖昧な言葉など分析し難しいテキストを分析しない傾向があることが見て取れた．一方で GPT-3.5-Turbo と Llama-2-7b-Chat は，評価対象数が多く，MSE の平均値も低くなっている．Xwin-LM-7B-v0.2 の評価対象数が低い理由は，表 6 の例のようにプロンプト文に感情推定を要求したものの，入力した推定対象のテキスト自体が生成（出力）されるような，本研究で望むのではない回答が得られるなどしたことが多かったためである．

また，Li らの研究 [6] によると，心理学を利用し，プロンプト文によりモデルに刺激を与えることでより良い結果が得られるとされている．プロンプト文の継続的な見直しは必要であると考えている．

表 6　要求したことを回答されない例

テキスト	卵ウォリにグロマッシュは多分入れたほうがいい。その分はコルクロンの精鋭を減らしてるレシピがよく使われてるけど、狂瀾怒涛を 1 枚にする方がいい気がする
回答文	卵ウォリにグロマッシュを入れることで、コルクロンの精鋭を減らすことができます。しかし、狂瀾怒涛を 1 枚にする方法がより効果的であると感じています。以下に、卵ウォリにグロマッシュを入れたレシピと、狂瀾怒涛を 1 枚にするレシピの両方を提供します。 卵ウォリにグロマッシュを入れたレシピ: 1. 卵を冷蔵庫で

5　まとめと今後の展望

本研究では，大規模言語モデルを用いた感情推定手法を提案した．さらに，評価実験により最低平均値 0.81 の MSE が得られ，この手法の有効性を検証した．

しかし，評価実験の規模は 100 件の投稿という小規模でしか行っておらず，Xwin-LM-7B-V0.2 の例のような評価対象数が少なくなっていることによる小標本問題が生じている可能性もある．今後は分析投稿数の増加を図りたい．

また，プロンプト文やパラメータの違いによる結果に与える影響も大きいため，さらなる探索的実験などは今後の課題である．

文　献

[1] R. Plutchik. The nature of emotions. American. Scientist, Vol. 89, Iss. 4; pp. 344-350, 2001.
[2] 梶原智之, Chenhui Chu, 武村紀子, 中島悠太, 長原一. 主観感情と客観感情の強度推定のための日本語データセット. 言語処理学会第 27 回年次大会,pp.523-527, 2021.
[3] 中澤政孝, 亀井且有, 前田陽一郎, クーパーエリック, BERT を用いた単文の感情極性推定手法の提案とその有効性, FSS2020, TA2-3, 2020.
[4] A. F. Adoma, N. -M. Henry and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition," 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2020, pp. 117-121, doi: 10.1109/ICCWAMTIP51612.2020.9317379.
[5] OpenAI, OpenAI Python API library, `https://github.com/openai/openai-python`, v1.6.0, 2023/12/19.
[6] C. Li et al.,‘Large Language Models Understand and Can be Enhanced by Emotional Stimuli’. 2023.
[7] R. Plutchik. A general psychoevolutionary theory of emotion. In Theories of emotion, pp. 3–33. Elsevier, 1980.

大規模言語モデルによる Multi-Step Prompt を用いた感情認識

濱　健太† 大塚 淳史† 石井 亮†

† 日本電信電話株式会社 NTT デジタルツインコンピューティング研究センタ
〒 108-0075 東京都港区港南 1-2-70 品川シーズンテラス 29F
E-mail: †{kenta.hama,atsushi.otsuka,ryoct.ishii}@ntt.com

あらまし 感情認識は, 様々なコンピュータシステムにおいて重要な役割を果たしている. しかし感情認識には, 感情ラベルを人手で付与するコストや, モデルの追加学習が必要な点が課題となっている. 近年, 大規模言語モデル（LLM）は, 特定のタスクのためのモデルの再学習を必要とせずに, 様々な自然言語処理タスクへの適応性を示している. これは, テキストベースの感情認識タスクに対して, LLM が有効である可能性を示唆している. 本研究では MELD と IEMOCAP という二つの会話感情認識における代表的なデータセットを使用し, LLM の感情認識性能を評価した. 我々の研究は, 学習用のサンプルが少数しか存在しない感情に対して, LLM を追加で再学習することなく, 従来手法の SPCL を上回る精度を実現することを確認した. 加えて Chain-of-Thought と似た, multi-step のプロンプトを LLM へ入力することで, 二つのデータセットに対する感情認識の精度が向上した. これらの結果は, 感情認識に LLM を活用することで, データセットの作成コストとモデルの再学習コストを低減できることを示している.

キーワード 感情認識，大規模言語モデル，プロンプトエンジニアリング

1 導　　入

会話中の感情認識（Emotion Recognition in Conversation, ERC）は, ヒューマン・コンピューター・インタラクションの分野で不可欠な要素として台頭しており, その応用範囲はエンターテイメントやカスタマーサポートからヘルスケアに至るまで多岐にわたる [1]. ERC における主要な課題は二点ある. 一つ目は感情の正確かつ一貫したラベリングにある [2]. 従来の方法は, 多数の評価者による合意 [3] に依存したり, ラベルの信頼性を高めるためにマルチモーダルデータを利用することが多い [4]. 二つ目は既存の感情推定方法では, このようなデータセットに対するモデルの再学習が必要になる点である [5]. 会話において現れる特定の感情は稀であるため, これらの感情のデータ収集は大きな課題となっている.

感情ラベリングにおけるこれらの課題への潜在的な解決策は, 大規模言語モデル（LLMs）の能力の活用が考えられる [6]. 近年, LLMs は様々な自然言語処理タスクにおいて卓越したパフォーマンスを発揮することを確認されている [7]. さらにプロンプトエンジニアリング [8, 9] などのテクニックにより, その能力は絶えず向上している. LLMs が事前学習の段階で感情認識に関する知識を自然に持っているなら, 感情認識タスクに再学習なしで適用することが可能になる. これにより, 広範かつ高コストな感情ラベリングの必要性がなくなる. したがって, LLMs を用いたアプローチは, 感情認識モデルの効率的な展開と洗練のための有望な方向性を提供する.

本研究では, ERC のベンチマークとして一般的に使用される MELD [4] および IEMOCAP [10] データセットを使用して, LLMs の ERC における能力を評価した. 始めに発話テキストの系列データである対話文を入力し, 特定の人物のある発話時の感情を LLM に予測させる, 1-Step のプロンプト入力による LLM の予測精度を確認した. しかし, 1-Step のプロンプト入力では LLM が会話中の文脈を理解することなく感情ラベルを出力してしまうことがあった. そのため, 本研究ではさらに会話の文脈を LLM が理解し感情の予測精度を向上させるために, zero-shot Chain of Thought (zero-shot CoT）[9] に触発された Multi-Step のプロンプティング手法を提案した. この提案手法には, 会話中の発話の背後の意味を問う質問を LLM 自身に生成させて, それに回答を行う過程が含まれる. 実験結果から, ラベルの分布が偏った MELD のようなデータセットにおいて, LLMs は “Anger” や “Fear” などのサンプルが少ない感情ラベルに対して, データセットに対する訓練を要さない few-shot の設定でも, 訓練済みの SPCL [5] より優れた精度を達成することが明らかになった. この結果は, ERC のためのデータセット作成, モデルの訓練のコストに対する課題を LLMs が解決する可能性を示唆している.

本研究の貢献は以下の通りである：

- 複数人対話のスクリプトに感情ラベルを付与した ERC のベンチマークである MELD および IEMOCAP に対して, LLM がサンプル数の少ない特定のラベルに対して, few-shot learning の設定で教師あり学習済みモデルで最高精度の SPCL を上回る精度を達成することを示した.
- 会話中の特定の発話の背後の文脈に対する理解を問う質問を LLM 自身に考えさせる中間タスクを実行させるという Multi-Step のプロンプト手法を用いることで, 単純な 1-Step のプロンプト入力より感情認識の精度が高まることを示した.
- Multi-Step な手法で, 感情認識結果の改善が見られたサンプルに対して, 中間ステップの出力である質問と回答の内容を確認し, LLM が会話の文脈を正しく理解することで, 発話の意味を正確に捉えることができていることを確認した.

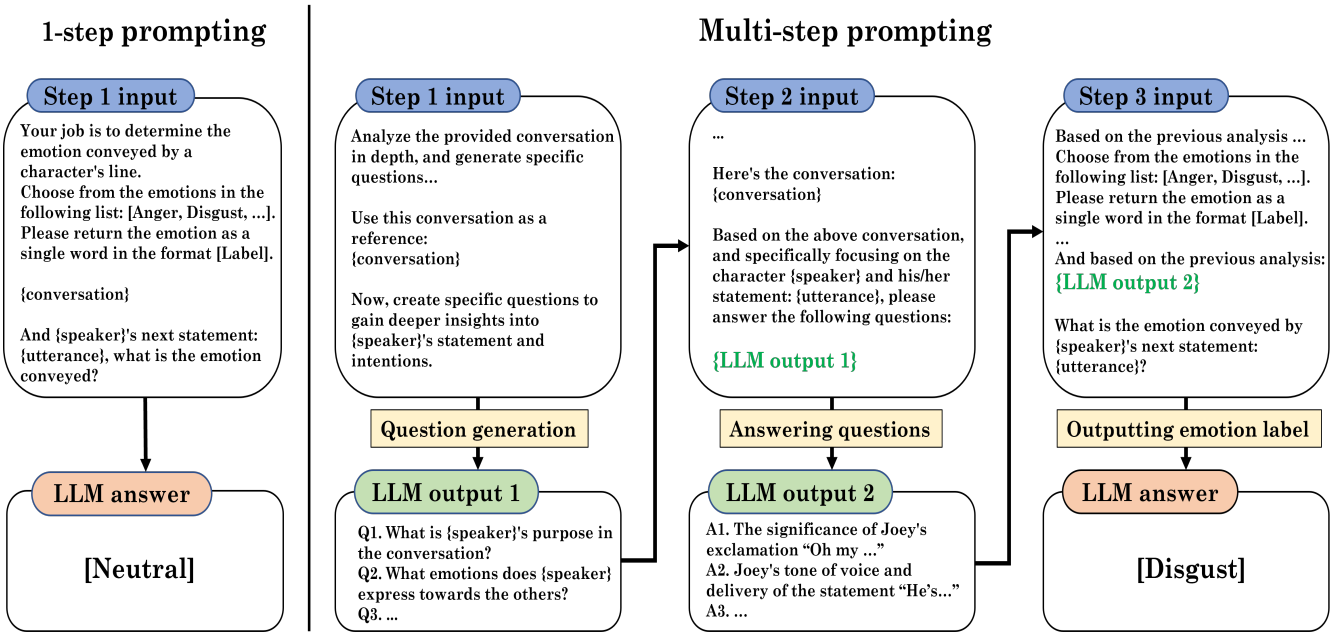


図 1　図は提案手法の流れを示している．図の左側が 1-Step のフローであり，図の右側が Multi-Step のフローを示している．Multi-Step プロンプティングは，Step 1 で対話に基づいて質問を生成し，Step 2 でこれらの質問に回答し，Step 3 で Step 2 の結果に基づいて感情を推測するという 3 段階のプロンプト入力を行う．

2　研 究 背 景

2.1　会話中の感情認識

会話中の感情認識（Emotion Recognition in Conversation, ERC）は，1 対 1 もしくは複数人の会話のテキストが与えられ，その中の特定の発話を行った話者の，発話時の感情を候補の中から予測する分類問題と定義される．モデルに対する入力は発話テキストの系列データと対象の発話テキストであり，モデルは感情に対応するラベルを出力するように学習される [2]．例えば，A さんと B さんの対話で，A「おはよう」，B「おはよう、今日はいい天気ですね」，A「そうですね！」といった発話テキストの系列データが与えられると，これらは単語レベルの埋め込み表現に変換され BERT などのモデルに入力される．そして，BERT の出力ベクトルから，「そうですね！」と発言した A さんの感情ラベルを予測する分類器を学習する．感情認識タスクでは，映像や音声を含むマルチモーダルデータを入力として利用する手法も存在する [11,12]．ただし，先行研究では，ERC においてはテキスト情報のみでマルチモーダルモデルに匹敵する性能が実現できることが確認されている [5]．本研究では感情認識におけるテキストモーダル情報の潜在的な限界を探求し，LLMs の ERC に対する適用可能性の探索に焦点を当てている．

2.2　大規模言語モデル

本研究における “LLMs” という用語は，GPT-3 [6] のような大規模なパラメタと大規模コーパスによる学習を行ったモデルを指す．従来のモデルである BERT, T5 [13,14] は，タスクのドメインに特化したデータセットで再学習する必要があった．対照的に，追加の訓練なしに，LLMs は自然言語処理に限らず，質問応答，要約，算術，プログラミングなどの多様なタスクで高いパフォーマンスを示している [7]．LLMs の能力を最大限に引き出すアプローチの一つは，思考連鎖（Chain of Thought, CoT）プロンプティング [8] などで有名なプロンプトエンジニアリング手法の使用である．モデルに直接結果を出力するよう指示する代わりに，CoT ではプロンプトが結果に至るまでの理由や思考の過程も出力するよう指示する．さらに，zero-shot CoT [9] では，“Let's think step by step” といったフレーズを LLM に渡すことで，モデルが独立して論理的推論を行うことを促す．本研究では，LLMs と CoT から着想を得て，ERC タスクにマルチステップのプロンプトエンジニアリングの手法を適用し，その有用性を探求する．

3　提 案 手 法

このセクションでは本研究で用いた LLM を利用した 1-Step および Multi-Step の手法について説明する．

3.1　1-Step プロンプティング

提案手法の概要を図 1に示す．図 1の左側に示すように，1-Step プロンプティングでは LLM が一度のプロンプト入力から直接感情ラベルを出力する．入力プロンプトは，会話文である発話テキストの系列データと感情を予測する対象となる発話を含む．この際，LLM には “Your job is to determine the emotion conveyed by a character's line.” というプロンプトにより自分の役割を把握させる．分類するクラスの中から確実に出力を行わせるために，“Choose from the emotion in the following

表 1　MELD における評価結果．太字の数値は表内で最高の精度を示している．SPCL はラベル全体の精度が優れている．特に GPT-3.5 は Anger と Fear（†でマークされている）において良好なパフォーマンスを示している．これはこれらのカテゴリーでのサンプル数が少ないことに起因している．

Model	Method	Anger†	Disgust	Fear†	Joy	Neutral	Sadness	Surprise	Overall
BERT	SPCL	53.0	**29.5**	28.0	**61.1**	**78.6**	42.5	**60.5**	**65.4**
GPT-3.5	1-step	53.4	23.2	27.9	53.0	37.3	**47.3**	43.2	42.7
	multi-step	**53.9**	27.8	**32.3**	51.2	57.2	43.2	45.5	52.2

表 2　IEMOCAP における評価結果．太字の数値は表内で最高の精度を示している．SPCL はすべての感情ラベルで高い精度を示している．IEMOCAP はすべてのラベルのサンプルが十分に存在するため, 訓練データで再学習された SPCL はどのラベルにおいても高精度となっている．

Model	Method	Frustration	Excited	Happiness	Sadness	Anger	Neutral	Overall
BERT	SPCL	**63.6**	**68.8**	**45.6**	**79.7**	**63.8**	**68.6**	**66.6**
GPT-3.5	1-step	56.3	58.5	40.7	68.4	31.7	39.1	50.5
	multi-step	58.8	50.2	37.8	62.9	36.9	52.2	52.2

list:" というプロンプトを与え, ["anger", "disgust",...] のような感情のラベルの候補リストを与える．このプロンプトにより, LLM は会話文を読み, 対象となる発話にふさわしい感情をリストから選択して, テキストで出力を行うようになる．LLM は事前に大規模な自然言語コーパスで自然言語の特徴を学習済みであり, 文脈にあった出力を行うため, LLM 自身を追加で学習することなく適切な感情ラベルの出力が期待される．

3.2　Multi-Step プロンプティング

Multi-Step プロンプティングは, 会話中の特定の発話に対して、その発話の意味を文脈から捉えなおすための質問の生成と, それに対する回答を LLM 自身に行わせる中間のステップを含んでいる．提案手法は質問生成, 質問への回答, 感情ラベルの出力という三つのステップから構成される．

Step 1（質問生成）： LLM に, 会話内の発話テキストの系列を与えて, 対象となる発話の背後にある文脈を明らかにさせるような質問を生成するように命令する．プロンプトの最初に "Analyze the provided conversation in depth, and generate specific questions..." と書いて, LLM にタスクの内容を理解させる．さらにプロンプトには "Create specific question to gain deeper insights into speaker's statement" といった文を含めることで, LLM が指示に従いより対話の文脈についての理解を問う質問を生成しやすくなる．このステップでは出力として 10 個程度の質問が生成される．

Step 2（質問への回答）： Step 2 では, LLM が Step 1 で生成された質問のリストに回答を行うよう命じる．LLM にはプロンプトの冒頭で, 自身が対話文についての分析を行う役割を持つということを説明するプロンプトを与える．これにより LLM は質問に回答する際の精度を高めることができる．さらに, "specifically focusing on the speaker and his/her statement..." というフレーズをプロンプトに入力することで, LLM が対話の内容を参照しながら回答するよう促す．このステップでは質問に対する回答が 10 個程度生成される．

Step 3（感情ラベルの出力）： 最終ステップでは, Step 2 の回答を対話と対象の発話とともにプロンプトに組み合わせて入力する．最終ステップでのプロンプトは 1-Step プロンプティングで用いるプロンプトとほとんど同じ内容であるが, Step 2 における質問への回答リストが与えられており, それを参考にしながら感情の候補リストから適切な感情を選択し, ["label"] というフォーマットで感情を出力するように指示する．

4　実　　験

4.1　データセット

我々は, ERC におけるベンチマークである MELD [4] および IEMOCAP [10] データセットを用いて LLM の感情認識の精度を評価した．この節では MELD, IEMOCAP それぞれのデータセットについて説明する．

4.1.1　MELD

MELD データセットは, テレビドラマ "Friends" における主要人物数人の複数人対話シーンを切り出して, 各発話に対して７つの感情ラベルがアノテーションされたデータセットである．1,400 からなる会話シーンの中に, 合計で 13,000 以上の発話が含まれており, 各発話の長さは平均で８単語ほどであり, 日常的な会話シーンが多い．7 つの感情は Anger, Disgust, Sadness, Joy, Neutral, Surprise, Fear からなる．ただしデータセット中の感情ごとのラベル数には偏りがあり, 発話の約半数は Neutral とラベル付けされている一方で, Disgust や Fear などのラベルは全体の数パーセント程度しか存在しない．また対話中の音声と, 映像データも付与されているが, 本研究ではテキストモーダ

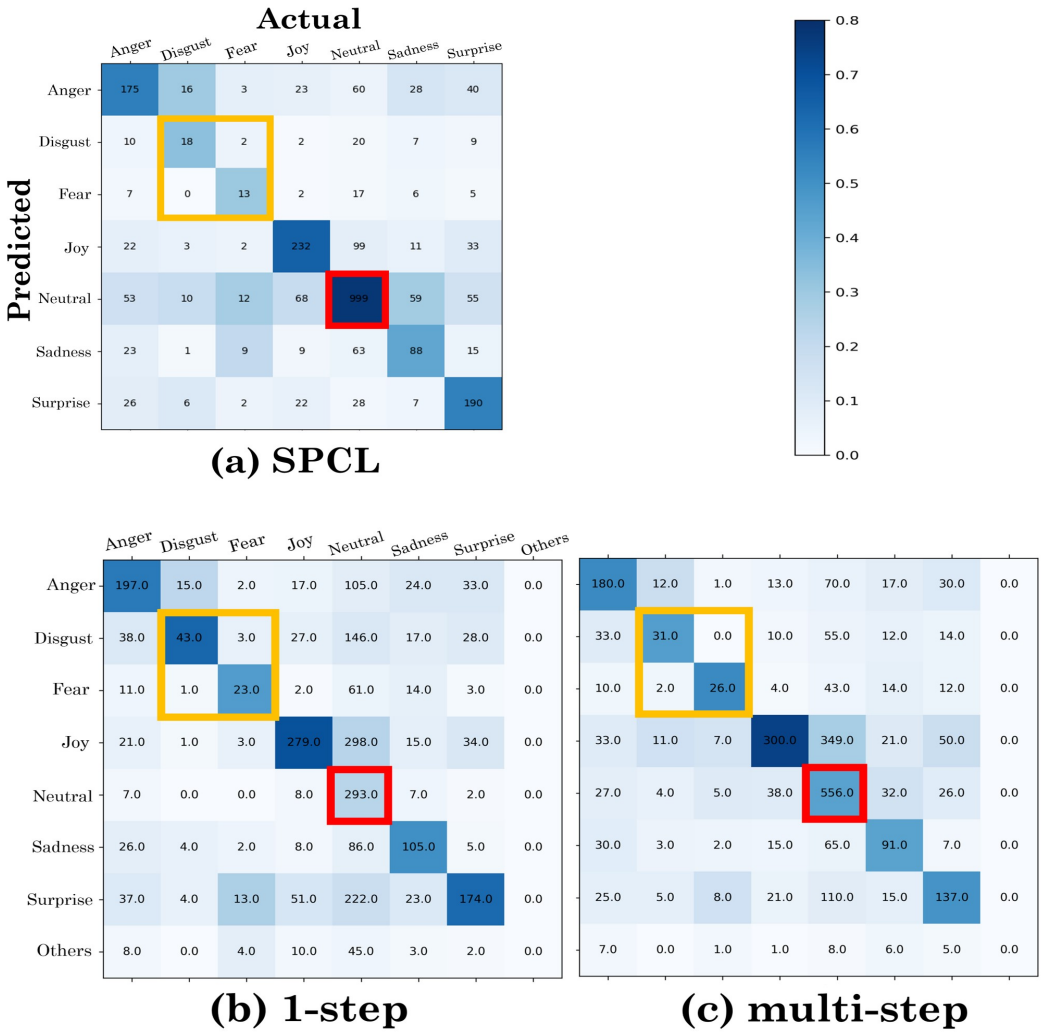


図 2　MELD のヒートマップ．各セルの値は，分類発話の数を表している．(b) の 1-Step では Neutral の精度が低いが，(c) の Multi-Step では改善されている．また，(a) の SPCL と比較して，(b) および (c) では Anger と Fear の分類精度が高くなっている．

ルのみ使用する．

4.1.2　IEMOCAP

IEMOCAP は，5 つのセッションにわたって 5 組の話者によって録画された 151 の対話ビデオからなるデータベースである．データは，感情表現を引き出すために，5 人の男性，5 人の女性の計 10 人の俳優によって演技された台本と即興のシナリオから作られている．151 の対話に対して，発話数は 7000 以上存在するため，一つの対話中の発話数が MELD に対して多い．各発話は 9 つの感情ラベルにアノテーションされているが，少数しか存在しないラベルを排除し Anger, Sadness, Excited, Neutral, Happiness, Frustration の 6 つをラベルとして選定する．少数ラベルを排除しているため MELD とは異なり，各ラベルのサンプル数が比較的均一に分布している．

4.2　実験設定

この実験では，MELD および IEMOCAP データセットに対して，LLM を用いた 1-Step プロンプティング，Multi-Step プロンプティング，および教師あり学習における最高精度の手法である SPCL の 3 つのモデルを比較した．SPCL は，事前学習された BERT を使用し，ターゲットの感情ラベルをマスクして，マスクされた単語を予測するというタスクでモデルを再学習する．本研究では，SPCL の公開済みのコード[1]の設定を用いてモデルをトレーニングし評価する．LLM を用いた 1-Step プロンプティング と Multi-Step プロンプティングは，OpenAI の gpt-3.5-turbo API を用いて実装する．[2] 評価指標として，各ラベルに対する F1 スコアと，全体的なデータセット評価のための加重 F1 スコアを使用する．

4.3　結　　果

実験結果は，表 1および表 2に示した．表 1の MELD データセットにおいては，1-Step と Multi-Step は全体的なパフォーマ

1：https://github.com/caskcsg/SPCL
2：https://platform.openai.com/docs/models/gpt-3-5

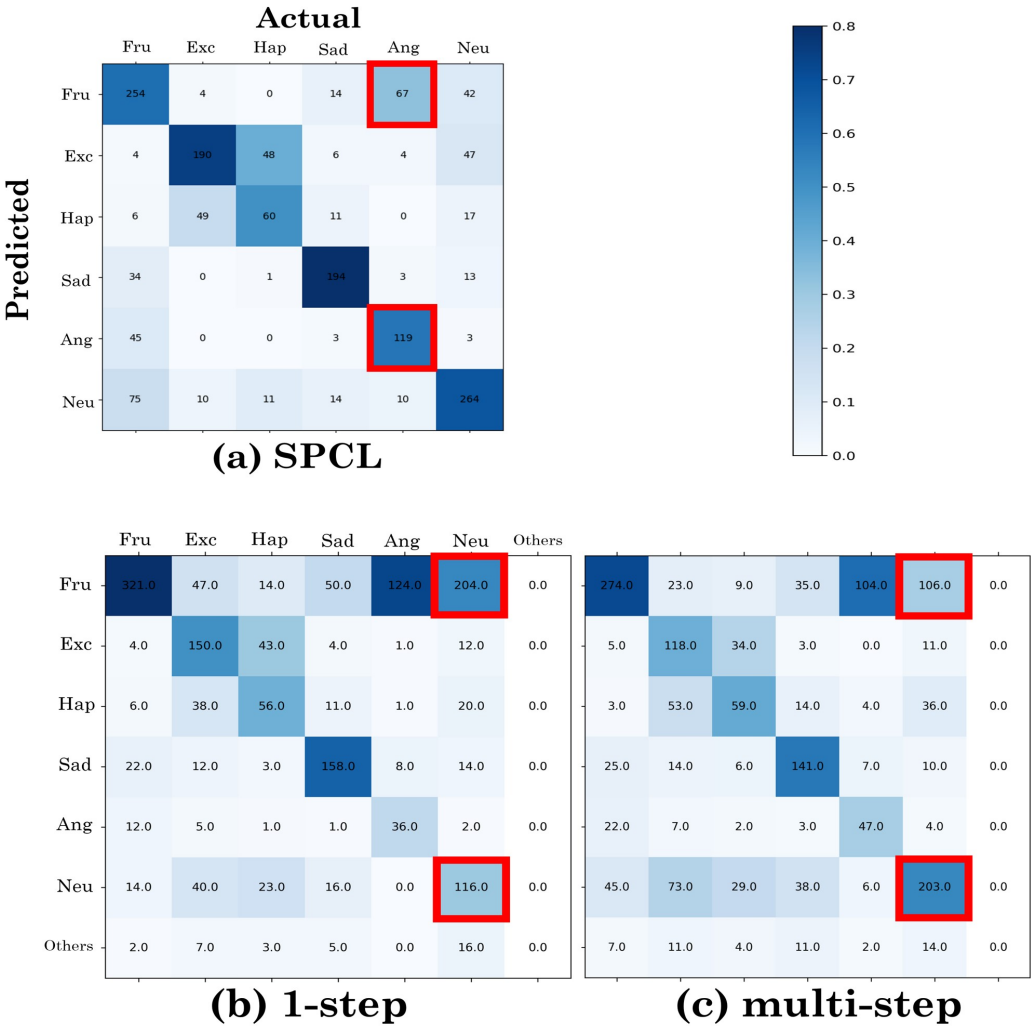


図 3　IEMOCAP のヒートマップ．各セルの値は，分類発話の数を表している．(b) の 1-Step では，Anger を Frustration と誤分類することが多いが，(c) の Multi-Step ではその傾向が軽減されている．

ンスにおいて SPCL に劣っている．しかし，Anger や Fear といった特定の感情カテゴリーにおいては，Multi-Step が SPCL を上回っている．さらに Multi-Step は，特に中立カテゴリーにおいて 1-Step よりも明確な改善を示している．表 2の IEMOCAP データセットに関しては，GPT-3.5 を使用した 1-Step と Multi-Step は SPCL よりも精度が低く，特に Anger での減少が顕著である．1-Step と Multi-Step を比較すると，後者は Anger と Frustration のカテゴリーでスコアが改善され，全体的な性能が向上している．

5　考　　察

5.1　誤分類の傾向に関する考察

この節では，図 2と図 3に示されたヒートマップに基づいて，各方法の感情分類における傾向について議論する．まず，MELD における図 2(a)（特に赤と黄色の枠内）を見ると，SPCL は Neutral において高い分類精度を示しているが，Disgust や Fear のような感情においては比較的低いパフォーマンスを示す．これは MELD のトレーニングデータの分布に起因していると考えられる．MELD では Neutral がサンプルのほぼ半数を占めているのに対し，Disgust や Fear は全体のわずか 3 %未満しか存在しない．図 2(b) から，1-Step では，Neutral とラベル付けされるべき発話が他のカテゴリに誤分類されていることが確認できる．一方で，Multi-Step（図 2(c)）では，Neutral に関する誤分類が大幅に減少している．また 1-Step と Multi-Step は，Disgust や Fear のような少数派ラベルにおいて SPCL よりも誤分類が少ないことを示している．これは，サンプル数が少なくラベル付与のコストが高い感情に対して LLM を使用することが効果的であることを示唆している．

IEMOCAP データセットの図 3(a) では，SPCL はすべてのラベルで高い分類性能を示し，特に難しい Anger と Frustration も区別できている．IEMOCAP のラベルの均等な分布により，SPCL は各ラベルの発話の特徴を十分学習することができたと考えられる．一方で，図 3(b) に示されるように，1-Step は SPCL よりも精度が低く，Anger と Frustration（赤い枠内）で

表 3　1-step では間違えているが、multi-step では正解した例. Prediction の actual は正解の感情ラベルを示す. Conversation Text は LLM の分析対象であり, 太字の部分が感情予測の対象となる発話テキスト. Questions は multi-step による手法の 1-step 目の LLM の出力である質問リストの一部. また Answers は 2-step 目 の出力であり, Questions に対する LLM の回答を示す.

Dataset	ID	Prediction	Conversation Text	Questions	Answers
MELD	33	1-step: “surprise” multi-step: “anger” actual : “anger”	Ross: “Okay” Rachel: “Ross, didn’t you say that there was an elevator in here?” Ross: “Uhh, yes I did but there isn’t. Okay, here we go.” Ross:“Okay, go left. Left! Left!” Rachel: “Okay, y’know what? There is no more left, left!” **Ross: “Oh okay, lift it straight up over your head!”**	Q1:What emotions might Ross be feeling when he says ”Straight up over your head!”? Is he annoyed, exasperated, or something else? Q2:Does Ross’s statement indicate a lack of patience or understanding towards Rachel’s actions or questions? Why or why not?	A1:Ross might be feeling a combination of annoyance and exasperation when he says ”Straight up over your head!” He may be frustrated with Rachel’s inability to follow his instructions or her lack of understanding of the situation. A2:Ross’s statement does indicate a lack of patience or understanding towards Rachel’s actions or questions. He seems to be getting frustrated with her inability to follow his directions or grasp the situation. This suggests that Ross may have a tendency to become impatient when others don’t meet his expectations.
IEMOCAP	1068	1-step: “anger” multi-step: “frustration” actual : “frustration”	Weldon: ”No, not at all. I still have it. I wear it quite often.” Roy: ”You went out of your way to torture me over Peter Burden.” Weldon: ”You worked the whole thing up in your jealous imagination.” Roy: ”You must admit that he was in love with you, wasn’t he?” Weldon: ”Maybe just a little, but–” Roy: ”You let him kiss you. You said you did.” Weldon: ”And what of it?” **Roy: ”What of it?”**	Q1:Is Roy trying to make Weldon feel guilty or ashamed for letting Peter Burden kiss them? Q2:Does Roy’s repeated question ”What of it?” indicate a sense of frustration or disbelief?	A1:It is possible that Roy is trying to make Weldon feel guilty or ashamed for letting Peter Burden kiss them, as Roy brings up the topic repeatedly and seems to be bothered by it. A2:Yes, Roy’s repeated question ”What of it?” indicates a sense of frustration or disbelief.

誤分類が多い. この課題は, 意味的に関連するラベルを区別することが LLM にとって困難であるために生じている可能性がある. しかし, 図 3(c) から, Multi-Step が Anger と Frustration に対して 1-Step よりも優れたパフォーマンスを示すことが確認できる. これは, Multi-Step における質問回答プロセスが, 単語の意味的に似通った二つの感情の違いを識別するのに役立つためだと考えられる.

上述の議論に基づき, SPCL は十分なトレーニングデータが利用可能な場合に高い分類精度を達成することができるが, LLM は少数のトレーニングサンプルを持つ感情ラベルに対して十分なパフォーマンスを提供することができると解る. また LLM は意味的に類似した意味を持つラベルを誤分類する傾向があるが, 質問回答のような中間タスクを導入することで, その識別能力を向上させることが可能である.

5.2 Multi-Step により改善した例

この節では 1-Step から Multi-Step にすることで感情の予測結果が改善した例を見ながら, 改善の理由について考察を行う. 表 3に, MELD と IEMOCAP において, 1-Step では不正解であるが, Multi-Step では正解の例を示す. MELD の ID33 の例では, 1-Step が “surprise” と誤分類しているが, Multi-Step では “anger” と正しい感情ラベルを出力できている. Conversation Text を見ると Ross と Rachel が荷物を運んでおり, Ross の最後の発言である “Oh okay, lift it straight up over your head!” は Ross の意図がうまく通じていない Rachel に対する苛立ちの感情が表れている. Questions を見ると, Q1 と Q2 ともに Ross が対象の太字の発話を行った際にどのように感じているかを聞いており, Answers では Ross の指示に対して理解できていない Rachell に対し, フラストレーションがたまっていると回答できている. 中間ステップとして Conversation Text の持つ文脈についての情報が補足され, 正解となる “anger” が出力されていると考えられる.

IEMOCAP の ID1068 の例では 1-Step では誤って “anger” を出力しているが, Multi-Step では正しく “frustration” を出力している. Conversation Text を見ると Weldon と Roy が男女関係の口論をしていると読み取れる. 対象の太字の発話は Roy の “What of it?” である. これは抽象的で, 文脈の理解が重要な例である. Questions を見ると Q1 では Roy が Weldon に罪の意識を感じさせる意図があるかどうかを質問しており, Q2 では Roy が Weldon の発言を繰り返していることから読み取れる感情について質問している. Answers を見ると Q1 と Q2 に対して LLM は肯定する回答をしており, 特に A2 の回答では対象発話が “frustration or disbelief” を示していると回答している. 中間ステップの質問回答というタスクにより, この発話が “anger” より “frustration” に近いと結論付けられている. これらの例は, 対象となる発話の前後の文脈の理解を問う質問文を生成し回答させることで, より精度の高い感情予測を可能にすることを示している.

6 結　　論

本研究では, ERC に対する事前学習のみの LLM の性能を評価した. 実験の結果, 訓練サンプルの少ない感情ラベルに対しては, LLM は SPCL のような最先端の訓練モデルを上回る精度を示した. また, LLM は単語の意味が近いラベルを区別するのが困難であるが, 中間タスクを含む Multi-Step プロンプティング手法を導入することで, その識別能力が改善することも示された. これらの結果は, 感情認識のためのデータセットの作成が困難な状況において, LLM に基づくアプローチが実行可能な解決策である可能性を示唆している. 今後の研究では, zero-shot および few-shot の文脈での様々な LLMs の感情認識タスクへの適用性を評価し, 質問応答以外の適切な中間タスクの探索を行う.

文　　献

[1] N. Ahmed, Z. A. Aghbari, and S. Girija, “A systematic survey on multimodal emotion recognition using learning algorithms,” *Intell. Syst. Appl.*, vol. 17, p. 200171, 2023.

[2] S. Poria, N. Majumder, R. Mihalcea, and E. H. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[3] S. M. Zahiri and J. D. Choi, “Emotion detection on TV show transcripts with sequence-based convolutional neural networks,” in *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, vol. WS-18 of *AAAI Technical Report*, pp. 44–52, AAAI Press, 2018.

[4] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (A. Korhonen, D. R. Traum, and L. Màrquez, eds.), pp. 527–536, Association for Computational Linguistics, 2019.

[5] X. Song, L. Huang, H. Xue, and S. Hu, “Supervised prototypical contrastive learning for emotion recognition in conversation,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), pp. 5197–5206, Association for Computational Linguistics, 2022.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), 2020.

[7] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, “Reasoning with language model prompting: A survey,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023* (A. Rogers, J. L. Boyd-Graber, and N. Okazaki, eds.), pp. 5368–5393, Association for Computational Linguistics, 2023.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *NeurIPS*, 2022.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” in *NeurIPS*, 2022.

[10] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” vol. 42, pp. 335–359, 2008.

[11] Z. Li, F. Tang, M. Zhao, and Y. Zhu, “Emocaps: Emotion capsule based model for conversational emotion recognition,” in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), pp. 1610–1618, Association for Computational Linguistics,

2022.

[12] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pp. 4651–4660, IEEE, 2022.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, p. 5485–5551, 2020.

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

2024年3月1日(金) 13:30 ~ 15:40 | T2-A オンライン（Zoom Events)

データリネージュ・統合データ管理

座長:北島 信哉(富士通株式会社)
コメンテータ:中野 美由紀(津田塾大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T2-A-8-01] 来歴情報を活用したデータベースからの因果推論

*大岩 和樹[1]、石川 佳治[1]、杉浦 健人[1]、陸 可鏡[1] (1. 名古屋大学大学院情報学研究科知能システム学専攻石川研究室)

13:50 ~ 14:15
[T2-A-8-02] データの流通管理・予測のためのリネージュデータ管理システム

*安田 香子[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学情報理工学院情報工学系)

14:15 ~ 14:40
[T2-A-8-03] ユーザ定義の述語とエンティティリンキングを利用した知識ベースと外部情報源の統合利用

*大森 雄基[1,4]、北川 博之[2,4]、天笠 俊之[3]、的野 晃整[4] (1. 筑波大学 システム情報工学研究群、2. 筑波大学 国際統合睡眠医科学研究機構、3. 筑波大学 計算科学研究センター、4. 国立研究開発法人産業技術総合研究所 人工知能研究センター)

14:40 ~ 15:05
[T2-A-8-04] A Polystore System for Graph and Key-Value Databases based on SQL/PGQ

*WONG WEI JUN[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学　情報理工学院　情報工学系)

15:05 ~ 15:30
[T2-A-8-05] 【技術報告】Universal HTAP: 多様なデータベースを統合的に管理するデータアーキテクチャ

山田浩之 (株式会社Scalar)

来歴情報を活用したデータベースからの因果推論

大岩　和樹† 　石川　佳治† 　杉浦　健人† 　陸 　可鏡†

† 名古屋大学大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †{oiwa,lu}@db.is.i.nagoya-u.ac.jp, ††{ishikawa,sugiura}@i.nagoya-u.ac.jp

あらまし　データベースに蓄積されたデータをもとに高度な推論を行う技術の一つとして，人工知能分野で発展してきた因果推論の手法をデータベースに適用するアプローチが近年着目されている．データベースの背後にある因果構造を活用することがその基礎となっている．本研究はこれを発展させ，データベースの来歴情報も合わせて活用するアプローチを提案する．本稿ではそのアイデアについて述べ，そのアプローチについて説明する．

キーワード　因果推論、来歴情報、問合せ処理

1　はじめに

データベースの研究分野においては，ユーザが有する知識や背景となる知識を用いて，より高度なデータベースの問合せ処理を行うことが古くから行われてきた．一方，統計分野や人工知能などの分野においても，データからの高度な推論を行うための技術が発展してきた．本稿では，そのような流れのなかで，近年大いに発展し着目を浴びている**因果推論**（causal inference）[4], [6] に着目する．因果推論は統計的因果推論とも呼ばれ，物事の背景にある因果情報をモデル化し，それも合わせて統計的推論に用いることで，原因とその結果の分析などに役立つ．

2　データベースにおける因果推論

データベースの問合せ処理に因果推論を本格的に導入した研究は Salimi ら [7] によるものが最初である．因果情報はデータベースに対する一種の背景知識と捉えることができ，データベース内のデータからより多くの情報を推測するために活用することができる．統計分野や人工知能分野における因果推論が一般には一つのテーブルで表現されるデータを対象とするのに対し，リレーショナルデータベースでは複数の関連するテーブルにまたがる問合せや分析を行う必要がある．[7] では，リレーショナルデータベースを対象とした因果推論の枠組みを提案し，与えられたデータベースに対する因果関係を踏まえた問合せに答えるための宣言的な問合せ言語を提案している．

これを受けて [5] では，データベースにおける因果推論に関するいくつかのアイデアを示している．因果関係についてはより限定的な定義を与える一方，来歴情報（lineage）を活用するアイデアや，データベースの操作により得られた各データに対して，その元となるデータの影響を図るための指標である責任（responsibility）など，発展性のあるアイデアを示している．

[2] では，[7] の考え方を元に，「もしデータベースを〜と変更したら何が起こるか」という **What-If 問合せ**と，「問合せ結果が〜のようになるにはどこを変更すればよいか」という **How-To 問合せ**に答える枠組みを提案し，それに基づくシステム HypeR を示している．さらに，[8] では，データレイクの中から，それを対象に構築された知識グラフを用いて，因果構造を発見するアプローチについて研究が進められている．

3　因果推論のアプローチ

ここでは，[2] の論文で示された，アマゾンの商品データベースを例として用いてアイデアを示す．図 1 は商品データベースの事例である．

Product

	PID	Category	Price	Brand	Color	Quality
p_1	1	Laptop	999	Vaio	Silver	0.7
p_2	2	Laptop	529	Asus	Black	0.65
p_3	3	Laptop	599	HP	Silver	0.5
p_4	4	DSLR Camera	549	Canon	Black	0.75
p_5	5	Sci Fi eBooks	15.99	Fantasy Press	Blue	0.4

Review

	PID	ReviewID	Sentiment	Rating
r_1	1	1	−0.95	2
r_2	2	2	0.7	4
r_3	2	3	−0.2	1
r_4	3	3	0.23	3
r_5	3	5	0.95	5
r_6	4	5	0.7	4

図 1　アマゾンの商品データベースの例

リレーション Product の属性は商品の ID，カテゴリ，価格，ブランド，色，品質の情報が属性からなる．一方，リレーション Review の属性は，商品 ID，レビュー ID，印象スコア（sentiment），評価スコア（rating）からなる．図 2 はその背景にある依存関係グラフの例である．

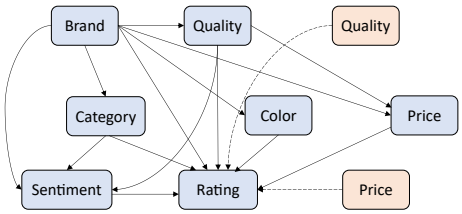


図 2　依存関係グラフの例

これはデータベースに関するユーザの背景知識を分析して得

られるものであり，水色のノードは同一のエンティティに対応している．たとえば，左上の Brand から右の Quality に出ている矢印は，ある商品について，ブランド情報が品質情報に依存することを意味している．一方，右下のオレンジ色の Price からその左への Rating への矢印は，他のエンティティの価格情報からの依存関係を表している．たとえば，他のブランドの商品の価格が割高である場合，注目しているエンティティの価格が魅力的で評価スコアが上がることなどを意味している．このような因果グラフが与えられると，学習データをもとに，ノード間の確率的依存関係の分析が行われ，その情報が因果推論の仮定で用いられる．

データベース問合せ処理の立場から，このようなデータベースと依存関係グラフが与えられたときに，データベースからの因果推論の一種である What-If 問合せ，How-To 問合せを実現しようというのが [2] のアイデアである．What-If 問合せの例としては，「Asus ラップトップの価格を 10%上げたとしたら，それらの平均の評価値はどのように変化するか」，「もし Asus のノートパソコンの価格がすべて 100 ドル下がったら，そのうち何割のものが評価 4.0 以上になるか」などが挙げられる．

What-If 問合せが与えられたときには，問合せの条件をもとに関連するデータをデータベースから抽出し，因果構造を反映したグラフを（*ground causal graph*）を構築する．これは，データベースから抽出されたインスタンス値を，図 2 に示したような依存関係グラフでつないだようなグラフである．このグラフがいったん構築されると，その後は通常の統計的因果推論のアプローチに従うことになる．

4 来歴情報の活用

上述のように，背景知識として因果情報を用いることで，データベースの実データだけを用いて行う問合せに比べ，より高度な問合せを行うことができ，特に仮説に基づく問合せが行えることは魅力的である．本稿では，さらに来歴情報を用いることで，さらにリッチな問合せを表すことを目指す．**来歴情報**（lineage）とは，一般にリレーションのタプルに付与される付加的な情報であり，そのタプルがどのようなデータから加工されて得られたかという情報を表す [1], [3]．ここではその事例を示す．

例として，図 1 で示した Review リレーションのデータの一部が，図 3 に示すレビュー情報サイト A の情報の一部を抽出・加工して得られたものであったとする．たとえば，Review_Site_A リレーションの第 1 番目のタプル a_1 は，Review リレーションのタプル r_1 の元になっている．

Review_Site_A

	Date	Reviewer	Age	PID	Sentiment	Rating
a_1	Apr 1	Joe	25	1	−0.95	2
a_2	Apr 1	Tom	18	4	0.5	3
a_3	Apr 3	Joe	25	2	−0.2	1
a_4	Apr 4	Lucy	30	1	0.11	2
a_5	Apr 4	Ken	22	5	0.95	5

図 3　レビュー情報サイト A の例

Review リレーションに対する来歴情報が適切に管理されているならば，これまでにない，以下のような問合せを考えることができる．これらは，来歴情報も加えたデータに対して What-If 問合せを行うものであり，来歴に対する**仮説問合せ**と捉えることもできる．

- レビュア Joe のスコア付けが全般的に厳し過ぎるという指摘があったとする．そのような場合に，「Joe が付与した評価スコアが 1 ポイントずつ大きかったとき，総合的な評価結果にはどのような変化が現れるか」といった What-If 問合せを考えることができる．
- レビュー情報を収集しているサイトとして A, B, C があり，サイト A は Z 社を高く評価しがちであるという意見があったとき，「サイト A の Z 社への評価スコアを 1 ポイントずつ減らした場合の変化を示せ」といった What-If 問合せが考えられる．
- 「20 代の男性のレビューのみを反映して総合的な評価結果を導け」といった，対象を限定する問合せを考えることもできる．これは ground causal graph を限定して評価を実施することに帰着する．
- 「ある商品 X に対する総合的な評価を 1 ポイント上昇させるには，X を未評価なユーザのうちの誰に依頼するのが有望か」：これは How-To 問合せの一種であると考えられる．実際にはレビュアの情報も含めて来歴が公開されることはないと考えられ，このような問合せの実現性は低いが，直接的なターゲットを発見できるならば非常に興味深いと考えられる．

このような問合せ処理は基本的には [2] の拡張となるが，過去の来歴について逆問合せを行い必要なデータをすべて抽出し，そこから ground causal graph を作る必要があり，より大規模な問合せが必要になる．また，さまざまな情報源からデータを収集している場合には，異種分散情報源の統合の問題も発生しうる．

5 まとめと今後の課題

本稿では，データベースにおける因果推論のフレームワークを，来歴情報を用いて拡張するアイデアについて述べた．今後は以下のような点を主に研究する．

（1） シナリオの詳細化とベンチマーク問合せの作成

（2） データモデルの構築：データベースの因果推論モデルと来歴情報の管理モデルを統合したデータモデルを構築する．

（3） 問合せ言語の開発：上に述べたような問合せを表現可能とする宣言的な問合せ言語を記述し，そのセマンティクスを定める．

（4） 問合せ処理手法の開発：What-If 問合せ，How-To 問合せについて，その効率的な処理技術を開発する．

3 についてはすでに [2] で行われているものの，かなりの時間がかかることが報告されている（特に How-To 問合せの場合）．データベースの問合せ技術をより効果的に用いてさらなる高速化が実現できないかを検討する．

謝　　辞

本研究の一部は，JSPS 科研費（JP20K19804，JP21H03555，JP22H03594, JP22H03903）および JST CREST JP22717241 の支援を受けたものである．

文　　献

[1] Y. Cui, J. Widom, and J. L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM TODS*, 25(2):179–227, 2000.

[2] S. Galhotra, A. G. andSudeepa Roy, and B. Salimi. HypeR: Hypothetical reasoning with what-if and how-to queries using a probabilistic causal approach. In *SIGMOD*, pp. 1598–1611, 2022.

[3] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *ACM PODS*, pp. 31–40, 2007.

[4] 黒木. 統計的因果推論 —因果効果の識別可能問題におけるベイジアンネットワークの役割—. 人工知能学会誌, 22(3):328–335, 2007.

[5] A. Meliou, W. Gatterbauer, J. Y. Halpern, C. Koch, K. F. Moore, and D. Suciu. Causality in databases. *IEEE Data Engin. Bull.*, 33(3):59–67, 2010.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. 邦訳：統計的因果推論 – モデル・推論・推測，共立出版，2009.

[7] B. Salimi, H. Parikh, M. Kayali, L. Getoor, S. Roy, and D. Suciu. Causal relational learning. In *SIGMOD*, pp. 241–256, 2020.

[8] B. Youngmann, M. Cafarella, B. Salimi, and A. Zeng. Causal data integration. In *PVLDB*, Vol. 16, pp. 2659–2665, 2023.

データの流通管理・予測のためのリネージュデータ管理システム

安田　香子† 常　宵† 宮崎　純†

† 東京工業大学情報理工学院情報工学系　〒 152–8550 東京都目黒区大岡山 2 丁目 1 2 − 1
E-mail: †yasuda@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.ac.jp

あらまし　本論文では，データの検証を可能にするためにデータ同士の関係や来歴を表すリネージュデータのモデリングを行う．近年，農業や自動車業界などにおけるものやデータの流通管理・予測のために工程全体の検証可能性が求められているが，検証可能性を満たすデータシステムは存在しない．そのため本研究では，検証可能なデータシステムのためのメタデータの管理方法として，プロパティグラフモデル上のノードでデータ集合を表し，リレーションでデータの流れと依存関係を表現するデータモデルを提案する．データモデルの実装において，グラフデータベース上ではデータ間の関係の管理や検索を主な目的とするため，データ本体はキーバリューストアで管理する．提案データモデルに基づくリネージュデータ管理システムはグラフデータベースである Neo4j とキーバリューストアである LevelDB を併用して構築し，検証を行った．

キーワード　リネージュデータ，データ管理，グラフデータベース，キーバリューストア

1 はじめに

近年，農業や自動車業界などにおけるものやデータの流通管理・予測のために工程全体の検証可能性が求められている．データの検証を可能にするための重要な要素として，データ同士の繋がりや来歴を表すリネージュデータがある．農業界では，ロボット，AI，IoT などの先端技術を活用したスマート農業を推進する動きが盛んである．果実及び育成環境のリアルタイムモニタリングによる最適出荷予測や過去の履歴に基づく売り上げ・需要・供給予測などが期待されており，それらの繋がりのあるデータの効率的な管理が求められている．自動車業界ではカーボンニュートラルの実現に向けて，サプライチェーン全体での温室効果ガス排出を削減するためにカーボンフットプリントが重要視されている．欧州連合（EU）は 2024 年から，販売する電気自動車（EV）の蓄電池のカーボンフットプリント（製造・廃棄時）を開示するようにメーカーに義務付けており [1]，製造過程での来歴であるリネージュデータを検証可能にすることとその管理が急務となっている．

スマート農業やカーボンフットプリントでは農産物や自動車の部品の流通過程の可視化が求められているのはもちろんだが，過去に得たデータが誤りだった場合に過去のデータや一連の操作を変更しデータを復旧すること（データベース修復と呼ぶ）や，結果に影響を与える因子を別のものに変えたらどうなるかシミュレーションし最適な方法や因子を決めたり影響を把握したりする What-if 分析も期待されている．また，機械学習が様々な業界で活用されているように，入力を得て結果を出力する関数（ジョブ）が機械学習になる例も多々ある．しかし既存のリネージュデータの管理フレームワークではデータベースの修復やシミュレーションを想定しているものはない．入力を得る関数が機械学習である場合はその関数自体も入力や出力になり得るが，その視点を持って関数を定義しているものもない．したがって関数として機械学習を取る場合があり，データベースの修復やシミュレーションが期待される農業や自動車業界の流通管理・予測のためのリネージュデータ管理には対応しきれないという現状がある．

そこで本研究では，農業や自動車業界などにおけるものやデータの流通管理・予測のために，工程全体を検証可能にするリネージュデータの管理手法を提案する．検証可能なデータシステムのためのメタデータの管理方法として，プロパティグラフモデル上のノードでデータ集合を表し，リレーションでデータの流れと依存関係を表現するデータモデルを提案する．データモデルの実装において，グラフデータベース上ではデータ間の関係の管理や検索を主な目的とするため，データ本体はキーバリューストアで管理する．提案データモデルに基づくリネージュデータ管理システムはグラフデータベースである Neo4j とキーバリューストアである LevelDB を併用して構築した．事例研究ではスマート農業の生育予測と出荷・流通のリネージュデータ管理を例にとり，データの追跡やデータベース修復の原因をデータ/関数/機械学習のトレーニングデータの 3 種類に分け修復を行った．また，過去の状況からある因子を変えたらどうなるかを把握するためのデータベースシミュレーションも行い，提案するデータモデルでデータの検証が可能であることが確認できた．

2 関連研究

2.1 リネージュデータに関する研究

リネージュデータに関する研究はリネージュデータのキャプチャと可視化を目的としたものや，機械学習のモデルのような限定された枠組み内でのデータのリネージュ管理を目的にしたものがある．以下ではまずリネージュの定義について触れ，それらの研究について紹介する．

リネージュ

リネージュ（lineage）は血統・系統などと訳され，データ・リネージュはデータの系統，つまりデータの流れを指す．データ・リネージュは，一定期間のデータの流れを追跡するプロセスで，データがどこから発生し，どのように変化したか，およびデータ・パイプライン内のデータの最終的な宛先を把握できるように管理する [2].

OpenLineage

リネージュデータの収集・分析フレームワークとして OpenLineage [3] がある．OpenLineage は，リネージュデータの収集と分析のためのオープンプラットフォームであり，データセット，ジョブ，実行に関するメタデータを追跡することができる．これまでは複数のソース，スケジューラ，データ処理フレームワークからリネージュデータを収集する際の課題として，複数のリネージュ管理システムがそれぞれ別の仕様を持っているため，別の管理システムで管理しているリネージュデータ同士の統合や管理システムの変更が難しいことが挙げられた．OpenLineage が標準化仕様として共通の仕様を決定することで，別の管理システムで管理しているリネージュデータ同士の統合や変更やそれぞれの実装の手間を減らすことが可能になるという利点がある．しかし OpenLineage はデータの流れのみを収集し，依存関係を扱うには不十分である．入力・出力・ジョブを構成要素としており，ジョブ自体は入力や出力にならないことが前提としてある．この前提の上ではデータの流れはデータの依存関係と同義であり，問題はない．しかしデータの流通管理・予測のためのリネージュデータ管理においてはジョブもデータとして入力や出力になり得るため，データの流れと依存関係の区別が必要であり，依存関係を扱うことができるモデルが必要となる．

Visionary

Visionary [4] はリネージュデータを取得，操作，分析，可視化するフレームワークである．可視化されたデータを通してユーザーはデータを探索し，分析や新たな考察をすることができる．今日，多くのツールがリネージュデータを利用しているが，専門家以外のユーザーやカジュアル・ユーザーにとっての使いやすさや理解しやすさに配慮しているものは少ない．Visionary はデータ理解とこれらのデータを用いた意思決定を支援するという点で革新的である．データを Entity，Activity，Agent にタイプ分けしそれぞれを 7 種類の関係で繋ぐ PROV モデル [5]（図 2）に加え，influence 関係を使用し関係を通じた影響を定義している．Visionary はリネージュデータを専門家ではないユーザーにもわかりやすく可視化することが目的であり，可視化されたリネージュデータの探索は可能だがデータの修復やシミュレーションを行うことを想定していない．PROV モデルにおいて機械学習のモデルを管理しようすると Entity として捉えられる場面と Activity としてはたらく場面があり，同一ものとして扱えないという問題もある．

Vamsa

近年，機械学習は金融や医療，製造業など，公平性・正確性・信頼性に細心の注意と配慮が求められる領域で使用されることが増えているが，機械学習モデルが学習データセットに反映さ

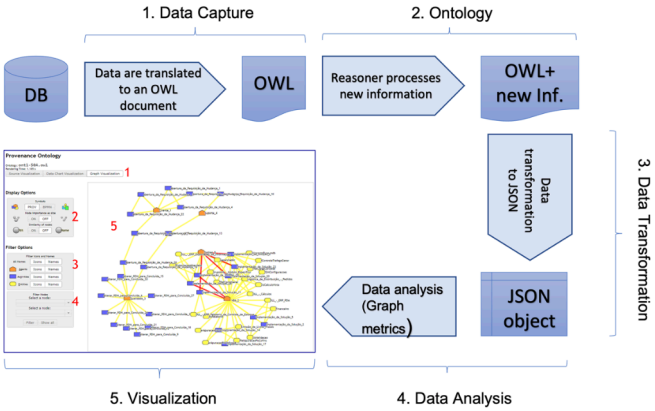


図 1　Visionary フレームワークの構造

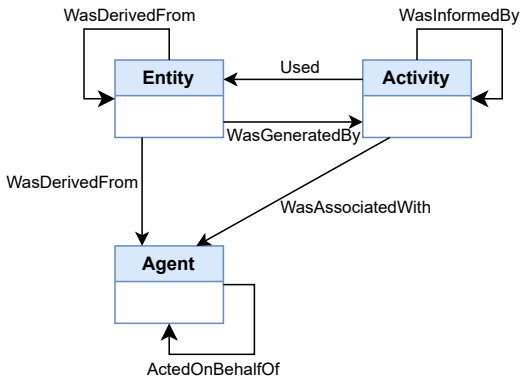


図 2　基本となる PROV のデータモデル

れた社会的に望ましくないバイアスを再現することや増幅することが懸念されている [6]．そのため，機械学習の公平性，偏り，説明可能性の分野で多くの研究が進行中である．Namaki ら [7] は機械学習モデルとそれを訓練するために使用されるデータとの間の関係を理解する必要があると考え，機械学習モデルの特徴/ラベルを導出するためにデータセットのどの列が使用されたかを自動的に追跡する Vamsa を提案した．Vamsa の導出抽出器は標準的な静的解析技術を使用して Python スクリプトを解析し，インポートされたライブラリ，入力引数，プログラムの状態を変更する操作，および導出された出力変数を含むスクリプトの要素間の依存関係をキャプチャするワークフローモデルを構築する．Kaggle のデータサイエンス・スクリプトと一般公開されている Python ノートブックを用いた約 26000 のスクリプトを対象とした実験では，適合率と再現率は 90.4% から 99.1%の範囲であり，レイテンシは平均的なサイズのスクリプトでミリ秒オーダーという結果を示している．機械学習の Python スクリプトを対象にしておりそれ以外のものやデータは扱っていないが，機械学習のリネージュデータ管理モデルとして非常に有用である．

2.2　データベース修復/シミュレーションに関する研究

データベースの復旧・回復に関する研究はこれまで多く行われてきたが，ほとんどが 1 トランザクション間に限定した復旧・回復であった．以下で長い 1 トランザクションの障害時回

復（Sagas）と本研究のデータモデルの要件である複数のトランザクションを跨いだ復旧及びシミュレーションを想定したフレームワーク（Ultraverse）について紹介する．

Sagas

実行に膨大な時間がかかる長寿命トランザクション（Long Lived Transaction）はトランザクションがコミットするまで扱うオブジェクトをブロックし，他のトランザクションを待たせることが問題である．長寿命トランザクションの実行時間が長い理由は多くのデータベースオブジェクトにアクセスすること，長時間の計算を行うこと，ユーザーからの入力のために一時停止すること，またはこれらの要因の組み合わせによるものである．長寿命トランザクションの問題を解決するため，Garcia-Molina ら [8] はトランザクションを独立したサブトランザクションに分割する運用方法を提案し，サブトランザクションに分割できる長寿命トランザクションのことを Saga と呼んだ．各サブトランザクションはデータベースの一貫性を保持するという意味で，実際のトランザクションである．しかし他のトランザクションとは異なり，Saga 内のトランザクションは互いに関連しており，非アトミックユニットとして実行されるべきである．分割したトランザクションを独立して処理することで 1 トランザクションにかかる時間を短縮し，テーブルを長時間ロックせずに処理を実行できる．

部分的な実行を修正するために，各サブトランザクションによって実行されたアクションを取り消す補償トランザクションが用意されている．障害によって Saga が中断された際の回復の動きは図 3 のようになる．

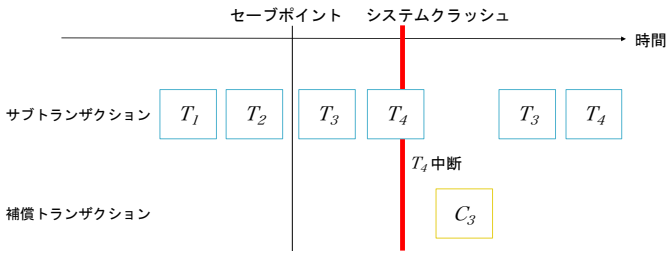


図 3　Saga が中断された際の回復の動き

サブトランザクション T_1，T_2，セーブポイントコマンド，およびトランザクション T_3 を実行する Saga を考える．その後，サブトランザクション T_4 の実行中にシステムがクラッシュする．回復すると，システムはまずセーブポイントへの後方回復のため T_4 を中断し，補償トランザクション C_3 を実行する．T_3 と T_4 を実行するコードが利用可能であることを確認した後，Saga 実行コンポーネントは再起動の決定をログに記録し，Saga を再開する．

Ultraverse

データベース上のデータの復旧に関する研究は昔から行われているが，多くは 1 トランザクション間の復旧を取り上げたものである．Ronny ら [9] の提案する Ultraverse は 1 トランザクションに限らず，コミットされた一連の操作のうち過去の操作を変更することができるフレームワークである．このように過去の一連の操作を変更し，データの復旧を行うことをデータベース修復と呼ぶ．この遡及的操作は，攻撃からの回復や個人データの削除（GDPR の遵守など）など，多くの重要なセキュリティに関係する用途がある．データベースシステムはユーザがコミットした SQL クエリを保存し，巻き戻しの要求を受けてからクエリ依存グラフを作成，それをもとに巻き戻して変更するという設計となっている．

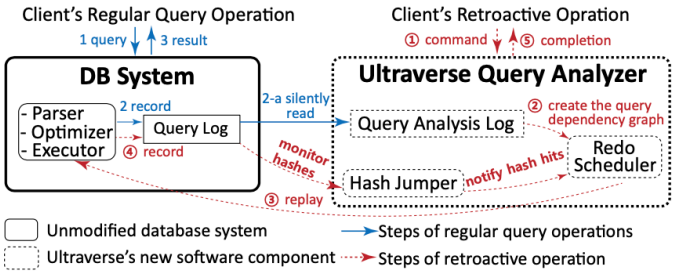


図 4　Ultraverse データベースシステムの構造

実験の結果，Ultraverse は通常の DBMS のフラッシュバック＆リドゥと比較して，データベースの遡及更新において 10.5 倍から 693.3 倍の高速化を達成した．RDB 以外の一般的なデータモデルには対応しておらずデータの拡張性がないため，リネージュデータは扱えない．

3　提案モデル

本節ではリネージュデータの管理モデルの概要と，その実装について説明する．

3.1　モデルの概要

提案モデルに求められる要件として，データが追跡可能であること/関数・操作も入力や出力として扱えること/データベースの修復やシミュレーションが可能であることが挙げられる．そこで以下に述べるデータモデルを提案する．

提案モデルでは，データ集合・データの流れ・データの依存関係をグラフで表現する．データ集合はグラフにおけるノードであり，入力・出力・関数（操作）に大別される．データの流れと依存関係はグラフにおけるリレーションであり，本モデルでは前後の依存関係のあるリレーションと前後に依存関係がなくデータの流れのみを表すリレーションを使用している．関数として機械学習のモデルを扱うことも考慮して，データだけでなく関数もデータと同様に受け渡しが可能な第一級オブジェクトと捉え，関数自体も入力や出力になり得るためデータの流れと依存関係を区別している．データモデルの基本構成は入力のデータ集合が関数を通して出力のデータ集合に繋がるという形である．入力と関数の間に依存関係はないがデータの流れがあり，入力と出力，関数と出力の間にはそれぞれ依存関係がある．ここでいう依存関係とはリレーションの起点に変更が生じた際に，リレーションの終点のノードが影響を受けるということである．全てのリネージュデータはこの入力→関数→出力という基本構成を組み合わせて表現される．

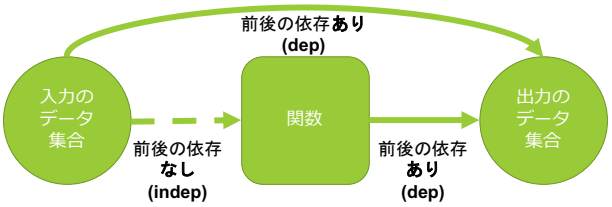


図 5　データモデルの基本構成

このモデルを用いて農作物の流通過程でのカーボンフットプリントのリネージュ，機械学習のモデルとトレーニングデータのリネージュ，スマート農業のビニールハウス内のリアルタイムモニタリングによる生育・出荷予測のリネージュを表現すると以下の図 6，7，8 のようになる．

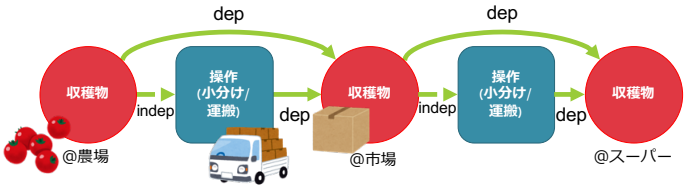


図 6　農作物の流通過程でのカーボンフットプリントのリネージュ

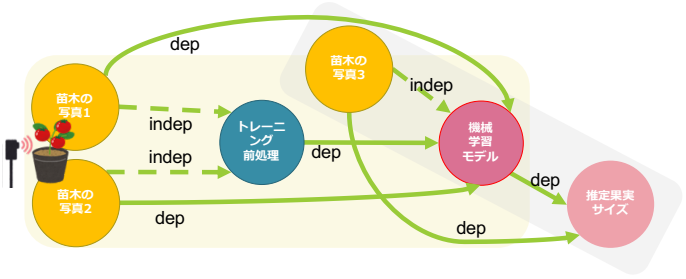


図 7　機械学習のモデルとトレーニングデータのリネージュ

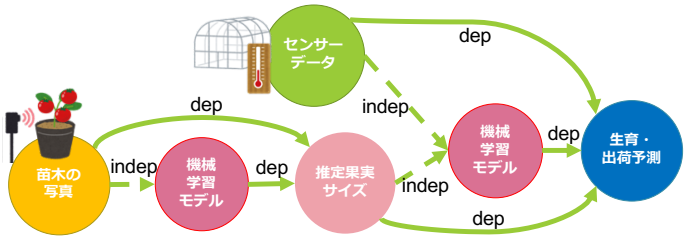


図 8　スマート農業のビニールハウス内のリアルタイムモニタリングによる生育・出荷予測のリネージュ

提案モデルは以下の要件を満たす．

データが追跡可能であること　ものの来歴であるリネージュデータはもの（データ）が他のもの（データ）とともに操作され形を変えていく様子やもの（データ）が運ばれていく様子を表す．データが別のデータに変わる際には必ず何かしらの操作や関数を通し，ものが実体は変わらず移動するということはものに対して操作が行われているということである．つまりリネージュデータのものやデータの動きは全て本データモデルの基本構成である入力→関数→出力という構造で表すことができる．データの流れは入力→関数→出力の出力が，その後に別の関数の入力としてまた次の出力へ繋がることでできる．本モデルではデータの繋がりを依存関係のあるリレーションと依存関係はなくデータの流れのみあるリレーションで表現している．あるデータを過去に追跡するとは有向グラフを逆方向に辿り，未来に追跡するとは順方向に辿ることである．辿るリレーションを依存関係のあるリレーション（dep）のみに絞ることで，依存関係にあるデータのみを追跡することができる．既存のグラフデータベースでは特定のノードから任意のパスの長さを指定し到達可能なサブグラフを検索する機能で実現可能である．データが追跡できることで，自動車のカーボンフットプリントのような様々なデータ（部品）が集約していくリネージュデータにおける経路上の集約演算も可能になる．例えば部品がそれぞれ持つ温室効果ガス排出量，各操作で排出した温室効果ガス量が各ノードのプロパティから得られるため，ある自動車を作るのに合計どれほどの温室効果ガスを排出したのかを段階ごとに算出できる．

関数・操作も入力や出力として扱えること　本研究ではタスク管理におけるジョブのような一般的な関数・操作に加え，スマート農業における生育・出荷予測の例のように関数として機械学習を扱うことも想定している．そのため関数も入力や出力と変わらないデータとして扱いノードを用意している．一般的な関数や操作は入力を出力に変える，入力→関数→出力の本モデルの構造で表現できる．関数が機械学習のモデルの場合は，機械学習のモデル M があるときは入力→モデル M →出力という流れで関数としてはたらき，あるときはトレーニングデータ→トレーニング前処理→モデル M のように出力として扱われる（図 7）．先ほど説明したデータの追跡可能性により，関数としても入出力としてもデータの追跡が可能であり，機械学習のモデルのトレーニングデータが誤りだった場合の影響範囲探索なども可能である．このように本モデルでは関数・操作も入力や出力として扱うことができる．

データベースの修復やシミュレーションが可能であること　過去に行った操作のデータなどにエラーが発見された場合の修復や，環境の因子を変えて現実には起こっていない状況を作り出すシミュレーションをリネージュデータ上で行う際の手順は共通している．まずは修復またはシミュレーションで差し替えるべきデータの影響範囲を調べる．これは依存関係のあるデータの追跡を未来に向かって行うことである．次に影響を受けるノードを計算する際に必要な入力を揃える．これは影響を受けるノードへ 1 本のリレーションで繋がっているノードを追跡することである．それらのデータと新しいデータとともに修復のための再計算やシミュレーションのための計算が行われる．最後は新しい計算結果を登録する．つまりデータベースの修復やシミュ

レーションのためにはデータの追跡が必要であり，本モデルを用いると再計算に必要な最小限のデータのグラフを取り出すことができる．

3.2 実装について

提案モデルはグラフデータモデルのプロパティグラフモデル [10] を使用するが，グラフデータベースは大規模なテーブル型データを扱うことに適していない．そのため大規模なデータはスケールアウト可能なキーバリューストアに，データの依存関係はグラフデータベースに格納することで，繋がりのある大規模なデータを扱う．実装にはグラフデータベースである Neo4j [10] とキーバリューストアである LevelDB [11] を使用した．これら 2 種類のデータベースの相互の利用は外部のプログラム上で行なっている．グラフデータベース上の各ノードはノードタイプ・名前・備考（”TRUE”/”FALSE”/”simulation” の 3 種類でデータの状態を表す）・生データへのキー・タイムスタンプを持つ．キーバリューストアのキーの名前はデータの取得日時から始まり，データタイプ（ノードタイプと同じ）や各種識別要素を連ね，最後にキーバリューストアに登録した日時をつけたものを設定している．

4 事例研究

本節では，提案モデルを用いてリネージュデータを管理し，データの追跡，データベース修復とシミュレーションがどのように行われるのか，事例を通して説明する．本モデルのスマート農業への適用として，ビニールハウス内のリアルタイムモニタリングによる生育・出荷予測と農産物の出荷・流通過程でのカーボンフットプリントを例に取る．リアルタイムモニタリングから生育・出荷予測までの手順は以下のものと仮定する．

1. センサーや撮影用ドローンで気温などの環境データと苗木や果実の写真データを収集する
2. 写真を機械学習にかけて苗木ごとの果実サイズや数を推定する
3. 環境データと果実推定情報から何日後にどれくらいの量になるかなどの生育・出荷予測を行う

この一連の作業によって得られたデータなどを，データ間の依存関係や流れを考慮してグラフにしたものを図 9 に示す．また，

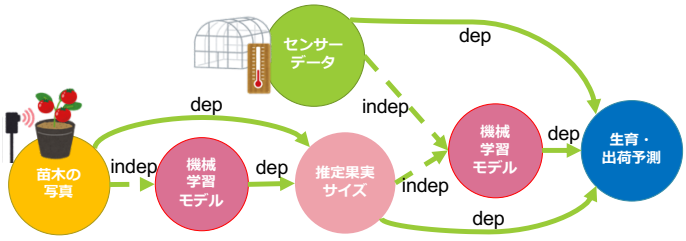


図 9 スマート農業の生育予測の流れ

図 9 では省略したが，データを機械学習にかける前に前計算や前処理が行われる．

農産物の出荷・流通の流れは以下のものと仮定する．

1. 農場の収穫物は小分けや運搬を経て市場に行く
2. 市場からまた小分けや運搬をされスーパーマーケットに辿り着く

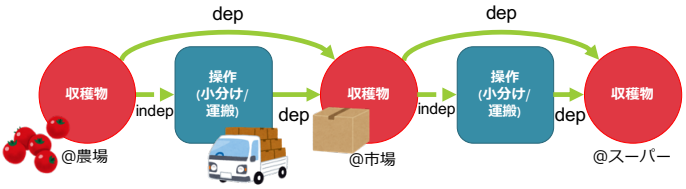


図 10 農作物の出荷・流通

本来は小分けされた後の収穫物のノードも存在するが，小分けから運搬を一連の操作と捉えて，間の収穫物ノードは省略した．

スマート農業の生育予測は関数や操作を通して入力とは異なる情報を持つ出力が生まれるが，農産物の出荷は操作を通しても出力は入力と同じ収穫物である（重量や蓄積 CO_2 換算値などのプロパティが変わる）．

4.1 データの追跡

提案モデルを用いたデータの追跡の流れを農作物の出荷・流通の例を使用し説明する．12/1 にある農場で収穫された農産物が小分けされ 2 つの市場に運ばれ，市場からまた小分けされ各スーパーマーケットに輸送される．グラフデータベースでは図 11 のように表現される．

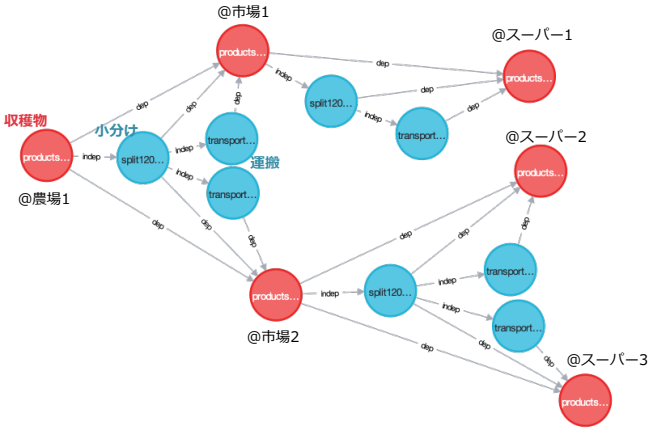


図 11 農産物の出荷・流通のグラフデータ

12/1 にスーパー 3 に届いた農産物がどこの農場からどこの市場を経て届いたのか，その過程でどれくらい温室効果ガスを排出したのかを知りたいとする．以下のクエリ 1 で由来を求める．

クエリ 1 農産物の流通の追跡クエリ

```
1  MATCH (m)-[r:dep*]->(n)
2  WHERE n.name =~ '.*products1201_supermarket3
       .*' AND m.note = 'TRUE'
3  RETURN m,n
```

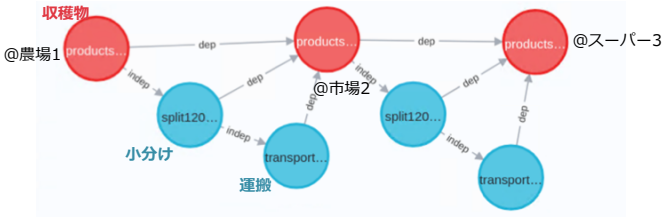


図 12　流 通 経 路

出力結果から，12/1 にスーパー 3 に届いた農産物は農場 1 からトラック 2 によって市場 2 に運ばれ，そこからトラック 13 により運搬されたものだとわかる．これらのノードが持つキーからキーバリューストアに格納されている各過程の温室効果ガスの排出量が得られる．このような手順でデータの追跡を行う．

4.2　データベース修復

提案モデルを用いたデータベース修復の流れを説明する．スマート農業の生育予測の一連のデータ 3 日間分をグラフデータベースとキーバリューストアに格納した．グラフデータベースでは図 13 のように表現される．

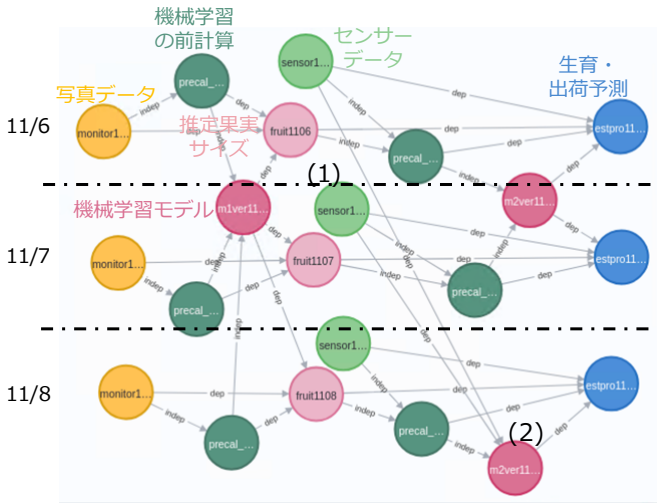


図 13　11/6〜11/8 の生育・出荷予測のグラフデータ

横軸方向で繋がりのあるノードは写真から生育予測までの一連の流れを示している．写真やセンサーデータは毎日収集されるが機械学習のモデルは数日間同じモデルを使用することがある．今回は，写真から果実サイズなどを推定する 1 つ目の機械学習は 3 日間で同じ，環境データと果実推定情報から生育予測を行う 2 つ目の機械学習は 1,2 日目は同じモデル，3 日目は新たな別のモデルを使用した場合である．1,2 日目のセンサーデータを用いて 3 日目の生育予測を行う機械学習は作られたため，依存関係のリレーションで結ばれている．

これらのデータのうち，11/7 のセンサーのデータ（図 13 の (1)）が誤っていたため正しいセンサーデータに修正する場合を考える．まずは誤っていたセンサーデータの影響範囲と，そのデータを使ってどのような操作をしたかをグラフ上で確認する．影響範囲は対象のノードを起点とする依存関係のリレーションで繋がったノードであり，以下のクエリ 2 で求められる．今回の事例では 11/7 のセンサーデータの誤りで，11/7 の生育・出荷予測，11/8 に使用した生育・出荷予測を求める機械学習のモデル，及びそのモデルを用いて導かれた 11/8 の生育・出荷予測に影響が生じると出力される．11/7 のセンサーデータは 11/7 の生育・出荷予測を求める際の入力としてと，11/8 の生育・出荷予測を求める機械学習のモデル（図 13 の (2)）のトレーニングデータとしてはたらいている．その影響を受ける機械学習のモデルは関数としてもはたらき，11/8 の生育・出荷予測を出力している，本事例は入力データ/機械学習のトレーニングデータ/関数の 3 種類のデータが原因のデータベース修復と捉えることができる．

クエリ 2　影響範囲の探索

```
1  MATCH (s)-[r:dep*]->(n)
2  WHERE s.key =~ '.*202311071200-sensor-gh1-temp-
       sen11.*' AND s.note = 'TRUE' AND n.note = '
       TRUE'
3  RETURN s,n
```

正しいデータでの再計算に必要な他のデータと，それらのデータを用いて行った操作や計算を知るために，先ほどの影響を受けるノードへ 1 つのリレーションで繋がるノードをクエリ 3 で求める．11/7 のセンサーデータの誤りによって 11/7 の生育・出荷予測が影響を受けたため，その再計算のためには生育・出荷予測を導くために使用された 11/7 の推定果実サイズ，11/7 に使用した生育・出荷予測を求める機械学習のモデル，その機械学習の前計算が必要であることが出力結果より読み取れる．また，11/7 のセンサーデータは 11/8 の生育予測を行う機械学習のモデルのトレーニングデータとして使われていたためモデルも作り直す必要があり，ともにトレーニングデータとして使用された 11/6 のセンサーデータも必要である．そのモデルを用いて算出された 11/8 の生育予測を再計算するために，11/8 のセンサーデータ・推定果実サイズ・機械学習前の計算が必要だということも出力される．

クエリ 3　修復のための再計算に必要なノードの検索

```
1  MATCH (s)-[r:dep*]->(n), (m)-[*0..1]->(n)
2  WHERE s.key =~ '.*202311071200-sensor-gh1-temp-
       sen11.*' AND s.note = 'TRUE' AND m.note = '
       TRUE' AND n.note = 'TRUE'
3  RETURN s, n, m
```

影響範囲と再計算に必要なノードを図 13 上に記したものが図 14 である．

次に，先ほど再計算のために必要だと判明した，対象のノードが持つキーでキーバリューストアから生のデータや関数の詳細などを取得する．そして，それらのデータを用いて新たな正しいセンサーデータで再計算を行う．再計算の結果を受け取っ

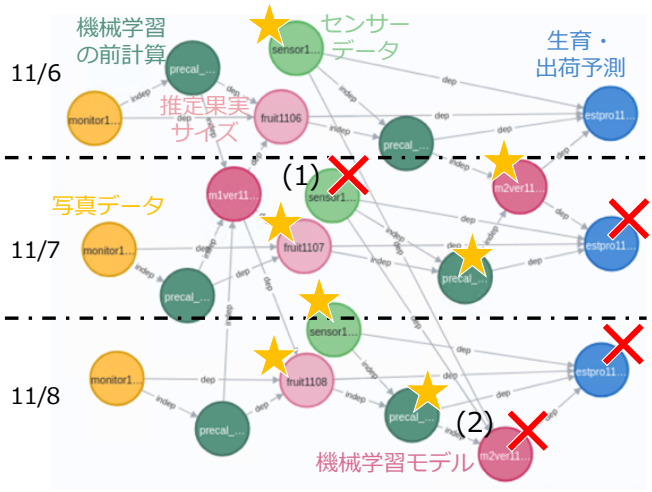


図 14　誤りの影響を受けるノード（赤バツ）と再計算に必要なノード（星）

た後，新たにグラフデータベースとキーバリューストアに登録する．グラフデータベース上では図 15 のようになる．

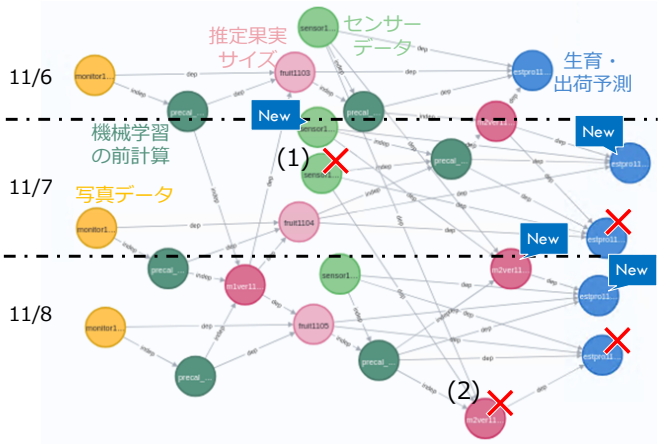


図 15　誤った値を持つ 11/7 のセンサーデータと生育予測と 11/8 の機械学習モデルと生育予測（赤バツ）は残し正しいセンサーデータ・機械学習・生育予測（青い New）を追加

4.3　データベースシミュレーション

提案モデルを用いたデータベースシミュレーションの流れを説明する．先ほどのデータベース修復の際と同じように，3 日分の生育予測のグラフを扱う．ここでは，肥料を肥料 A から肥料 B に変えた場合の生育予測のシミュレーションを 11/10～11/12 の環境を基に行うことを考える．通常なら写真データから推定果実サイズを出し，推定果実サイズとセンサーデータから生育・出荷予測ができる．しかしシミュレーションでは環境を変えた場合の 11/11 以降の写真データはないため，前日の生育予測から次の日の推定果実サイズを得る方針で考える．

まずはシミュレーションで必要なデータをキーバリューストアから取得する．今回は (3) の 11/10 の生育・出荷予測，(4) と (5) の 11/11,12 のセンサーデータが必要となる．肥料を変更するので，推定果実サイズとセンサーデータから生育・出荷

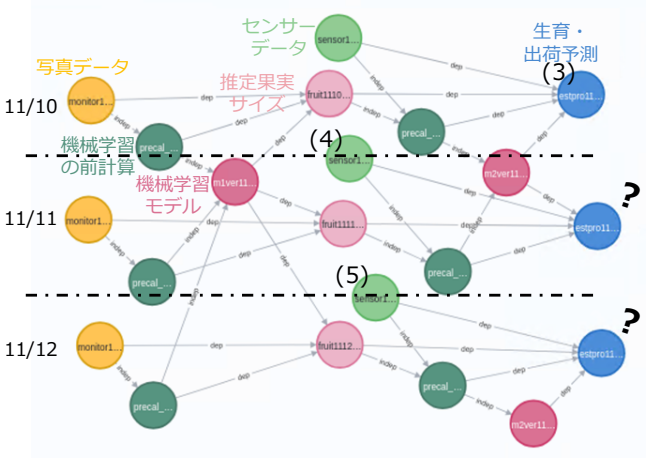


図 16　11/10～11/12 の生育・出荷予測のグラフデータ．肥料 A を使用している．

予測を出す 2 つ目の機械学習のモデルが変わる．先ほど取得したデータと新しい機械学習のモデルを用いてシミュレーションを行う．そしてシミュレーションの結果を，新たにグラフデータベースとキーバリューストアに登録する．シミュレーション結果を追加したグラフデータベースの様子を図 17 に示す．

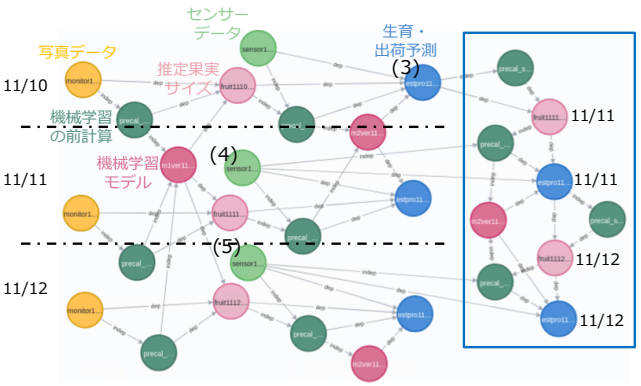


図 17　11/10～11/12 の生育・出荷予測のグラフデータに肥料 B を使用した場合のシミュレーションノードを追加（青枠内）

シミュレーションの始点の日の生育予測から次の日の果実推定サイズを計算し，そのシミュレーションの果実推定サイズと実際のセンサーデータを入力として，肥料 B バージョンの機械学習と前計算に入れる．するとシミュレーションの肥料 B を使った場合の生育予測を出すことができる．その生育予測からまた次の日の果実推定サイズを計算，と繰り返すことでシミュレーションが可能となる．

4.4　考　　察

提案モデルを用いたリネージュデータ管理システムで複数事例のデータ追跡，データベース修復とシミュレーションを行い，検証可能であることがわかった．データ追跡ではスーパーに届いた農産物の流通経路を辿ったが，自動車のパーツでも同様に流通経路をリネージュデータで表し，流通経路での温室効果ガス排出量を計算することが可能だ．データベース修復では修復の原因が普通の入力データの場合，機械学習のトレーニン

グデータの場合，関数の場合の 3 種類のパターンを含む事例を扱ったが，提案モデルを用いたリネージュデータ管理方法では全て同様の手順・クエリで修復を行えることが確認できた．データベースシミュレーションでは肥料を変更した場合に生育予測にどう変化が生じるか確認するシミュレーションを扱った．温度や湿度を変えたり水やりの時間を変えたりなど生育環境を変えてシミュレーションすることで最適な育成方法や環境を決めることや影響を把握することに役立つ．

5 おわりに

本研究では，データの検証を可能にするためにデータ同士の関係や来歴を表すリネージュデータのモデリングを行った．プロパティグラフモデル上のノードでデータ集合を表し，リレーションでデータの流れと依存関係を表現するデータモデルを提案した．データモデルの実装において，グラフデータベースは大規模なテーブル型データを扱うことに適していないため，データ本体はスケールアウト可能なキーバリューストアで管理することで大規模な繋がりのあるデータを扱う．提案データモデルに基づくリネージュデータ管理システムはグラフデータベースである Neo4j とキーバリューストアである LevelDB を併用して構築した．事例研究では，スマート農業の生育予測と出荷・流通のリネージュデータ管理を例にとり，データの追跡やデータベース修復とシミュレーションを行い，複数事例でデータの追跡・修復・シミュレーションを行えることが確認できた．

今後の課題として，現時点ではデータ数が多くはないため，データ数が膨大になった際にも問題なく動くことを確認する必要がある．対象となるデータは日々増えるためスケーラビリティを重視してキーバリューストアを採用したが，グラフデータベースにおける膨大なデータ集合の取り扱いについても確認が必要である．また，キーバリューストアのキーに複数の情報を盛り込んでおりキーの部分検索が必要になってしまっているが，文字列での識別は時間もかかるため最善の方法ではないと考えている．キーの命名規則をさらに工夫するか，キーとバリューの他に備考の情報も付けることができるキーバリューストアの採用も視野に入れる．さらに，あらゆるクエリに対応するために農業や自動車業界の現場でどのような操作が求められているのかを調べる．実際にスマート農業の現場で扱われるデータをもらい本モデルで管理し，モデルの増強を行う．

謝　辞

本研究は，JST CREST JPMJCR22M2 の支援を受けたものである．

文　献

[1] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, ”REGULATION (EU) 2023/1542 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2023”, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023R1542 , 2023.

[2] IBM, ”データ・リネージュとは”, https://www.ibm.com/jp-ja/topics/data-lineage ,(2024 年 1 月 7 日参照)

[3] OpenLineage, ”OpenLineage”, https://openlineage.io , (2024 年 1 月 7 日参照)

[4] W. de Oliveira, R. Braga, J. M. N. David, V. Ströele, F. Campos and G. Castro, ”Visionary: a framework for analysis and visualization of provenance data”, Knowledge and Information Systems, vol. 64, no. 2, pp. 381-413, 2022.

[5] The World Wide Web Consortium, ”PROV-DM: The PROV Data Model”, https://www.w3.org/TR/2013/REC-prov-dm-20130430/ , 2013, (2024 年 1 月 25 日参照)

[6] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford ,”Datasheets for Datasets.”, arXiv:1803.09010, 2021.

[7] Mohammad Hossein Namaki, Avrilia Floratou, Fotis Psallidas, Subru Krishnan, Ashvin Agrawal, Yinghui Wu, Yiwen Zhu, and Markus Weimer, ”Vamsa: Automated Provenance Tracking in Data Science Scripts.” , Association for Computing Machinery, New York, NY, USA, 1542–1551, https://doi.org/10.1145/3394486.3403205 , 2020.

[8] Hector Garcia-Molina, Kenneth Salem,”Sagas”, SIGMOD Rec. 16, 3 (Dec. 1987), 249–259. https://doi.org/10.1145/38714.38742 ,1987.

[9] Ronny Ko, Chuan Xiao, Makato Onizuka, Yihe Huang, Zhiqiang Lin, ”Ultraverse: Efficient Retroactive Operation for Attack Recovery in Database Systems and Web Frameworks.”, arXiv:2211.05327, 2022.

[10] Ian Robinson, Jim Webber, Emil Eifrem, ”グラフデータベース -Neo4j によるグラフデータモデルとグラフデータベース入門”, (佐藤直生 監訳, 木下哲也 訳), オライリー・ジャパン, 2021.

[11] Google, ”LevelDB”, https://github.com/google/leveldb , (2024 年 1 月 8 日参照)

ユーザ定義の述語とエンティティリンキングを利用した知識ベースと外部情報源の統合利用

大森　雄基† 　北川　博之†† 　天笠　俊之††† 　的野　晃整††††

† 筑波大学大学院 システム情報工学研究群　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
†† 筑波大学 国際統合睡眠医科学研究機構　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
††† 筑波大学 計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1 丁目 1–1
†††† 国立研究開発法人産業技術総合研究所 人工知能研究センター　〒 135–0064 東京都江東区青海 2-3-26
E-mail: †omori.yuki.sm@alumni.tsukuba.ac.jp, ††kitagawa@cs.tsukuba.ac.jp, †††amagasa@cs.tsukuba.ac.jp ,
††††a.matono@aist.go.jp

あらまし　RDF 形式の知識ベースとは，人間の持つ知識を主語・述語・目的語の組の集合で表現したものであり，様々な知識処理における情報源として活用が進められつつある．しかし，気象情報や運行情報などの即時性が重要な情報や，詳細で専門的な知識など，非知識ベース情報源も多く，知識ベースと外部情報源の統合利用により多様な応用に対応することが期待される．一方で，外部情報源の多様な入出力形式と，出力結果に知識ベースのエンティティを選出する問題から，知識ベースと外部情報源の統合利用は一般に極めて煩雑となる．これまで著者らは，ユーザ定義述語を利用して外部情報源を知識ベースと一体とする統合利用環境のアーキテクチャを示し，外部情報源が出力結果に対して適切なエンティティを選出するエンティティリンキングの手法を示した．しかし，従来の手法は外部情報源のエンティティ相当の文字列に対してエンティティを選出するため，外部情報源が出力する構造的な情報を十分には活用していない．また，全てのユーザ定義述語の同じエンティティリンキングを適用する手法であった．そこで本研究では，ユーザ定義述語を登録する際に，エンティティリンキングも登録するアーキテクチャを示す．これにより，文字列・構造化データなど，複数の性質やドメインに対応しつつ，外部情報源を含む結果をあたかも通常の知識ベースに対して問合せをしたかのように問合せることを可能とする．

キーワード　データ統合技術, 問合せ処理, 知識ベース, SPARQL, エンティティリンキング

1　はじめに

人類の有する知識が増加するにつれ，様々な情報源からのデータを処理できる技術の重要性が増している．近年では，Resource Description Framework (RDF) 形式の知識ベースや，問合せ言語 SPARQL が，知識処理に使われる．しかし，全ての知識が RDF 形式の知識ベースに蓄積されているわけではない．例えば，気象情報や交通情報のように即時性が重要な情報は，静的な蓄積に適さない．また，高度に専門的な情報も，知識ベース化されていないことが多くある．従って，そのような情報を有効に活用するため，RDF 形式の知識ベースと RDF 以外の外部情報源の両方を統合して利用できる環境が重要となる．

外部情報源は，問合せに対して独自のインタフェイスや返り値に独自のフォーマットを利用することが多く，それに合わせた独自のアクセス方法を利用し，独自のフォーマットに従って解釈する必要がある．このため，知識ベースと外部情報源の両方にまたがって問合せたい場合，非常に時間のかかる処理となる．問合せ処理は，多様なフォーマットに対応しつつ，知識ベースと外部情報源の両方の結果を照合する必要がある．しかし，知識ベースのエンティティと外部情報源のオブジェクトとの対応は不明確なことが多く，手作業での判断は極めて煩雑である．例えば，外部情報源とし地図アプリケーションを利用して特定の駅周辺の教会を検索し，知識ベース内の情報を利用して特定の宗派の教会に絞り込む場合，地図上のオブジェクトと知識ベース内のエンティティの対応関係を 1 つ 1 つ調べて集計する必要がある．

以上を受けて，筆者は *Knowledge Mediator* [1] という統合環境のアーキテクチャを提案した．これは外部情報源にアクセスし，その情報を RDF 形式の知識ベースの一部であるかのような問合せ環境を提供する．Knowledge Mediator の特徴は 2 点である．(1)*Magic Properties* を用いて外部情報源にアクセスし，問合せ結果に組み込む．*Magic Properties* はユーザ定義述語であり，アプリケーション固有の情報処理が実装可能である．(2)Knowledge Mediator のエンティティ発見機構である問合せ駆動型応需式エンティティリンキングは，中間処理表をエンティティを説明する文脈として利用し，問合せ処理で外部情報源から取得したオブジェクトに対して知識ベース内の正しいエンティティを選択する．

Knowledge Mediator のアーキテクチャが提供する拡張知識ベースは，2 種類の述語を持つ．一つは，元々知識ベースにある通常の述語で，本稿では *Knowledge Base Predicate (KBP)* と呼ぶ．もう一つは，Magic Properties によって定義され，ユー

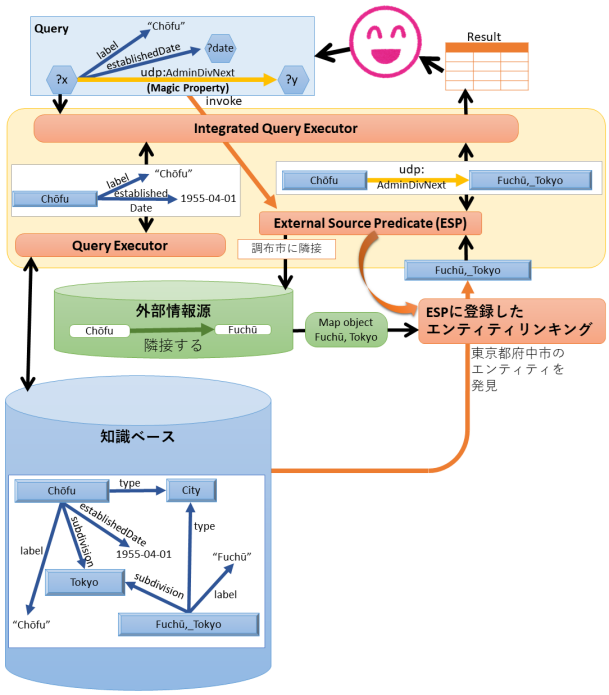


図 1　本研究の提案するアーキテクチャ．非 RDF 形式の外部情報源と知識ベースを統合して問合せる．Knowledge Mediator [1] では，エンティティリンキングはアーキテクチャに固定のものであったが，本研究では ESP それぞれに適するエンティティリンキングを登録する．

ザ定義関数を呼び出すことで，外部情報源にアクセスする述語で，本稿では *External Source Predicate (ESP)* と呼ぶ．

例えば，外部情報源（地図アプリケーション）を使って「調布市の設立日とその近隣都市」を検索する問合せは，次のような SPARQL 問合せで表現できる．

ESP `ESP:AdminDivNext`.

```
SELECT DISTINCT * WHERE {
?x rdfs:label "Chōfu"@en .
?x rdf:type dbo:City .
?x dbp:establishedDate ?date .
?x <ESP:AdminDivNext> ?y .}
```

この問合せでは，$?x$ に隣接するエンティティ$?y$，すなわち $\langle ?x, \texttt{ESP:AdminDivNext}, ?y\rangle$ を満たすエンティティの検索は，地図アプリケーションへのアクセスを呼び出す `ESP:AdminDivNext` によって実行される．地図アプリケーションは，エンティティ$?x$ に対応するオブジェクトに隣接する位置を持つオブジェクト (例えば”Fuchū”) を検索して返す．問合せ結果を処理するには，返されたオブジェクトに対応する知識ベース内のエンティティ (例： `dbr:Fuchū,_Tokyo`) を見つける必要がある．これは，ESP とともに登録したエンティティリンキングによって，知識ベース内のエンティティを見つける．

しかしながら, [1] は，外部情報源のオブジェクトから該当するエンティティを探すにあたり，そのエンティティ相当の名前の文字列に対して単一のエンティティリンキング方式を用いるものであった．しかし，外部情報源は文字列だけでなく，例えば XML や JSON などの構造化データも多い．またその性質も多様である．例えば，メトロポリタン美術館の API[1]は JSON 形式であり，固定したスキーマを持つ．このため，どの属性値が名前相当であるかの特定は容易である．一方で，地図サービス OpenStreetMap[2]は，XML 形式であるが，その地図オブジェクトの属性および属性値はユーザが登録可能であり，スキーマは固定されていない．このため，地図オブジェクトの名前だけを指定する場合は様々な場合分けが必要となる．このとき，属性が特定できないならば，全ての属性値を利用してエンティティを特定するという解決法も考えられる．以上の事から，例えばメトロポリタン美術館の API に対しては名前相当の属性値を使い，OpenStreetMap に対しては，全ての属性値を利用するという形で情報源の性質の違いに対応することが考えられる．本研究 (図 1) では，ESP を登録するときに，エンティティリンキングも登録することで，各情報源の返り値に合わせたエンティティリンキングを可能とする．

本研究の貢献は以下の通りである．

(1) 多様な形式の外部情報源にアクセスしつつ，動的に知識ベースと外部情報源のオブジェクトを統合し，あたかも知識ベースの一部であるかのように外部情報を含んだ問合せ結果を返すアーキテクチャを示す．

(2) ESP と共にエンティティリンキングも登録可能ととすることで，ESP にあわせた動的なエンティティリンキング切り替え環境を実現し，外部情報源の多様な性質に配慮可能とした．

(3) 提案手法のプロトタイプを実装し，ESP にエンティティリンキングを登録する事の利点を実験的に証明した．

本論文の構成は以下の通りである．第 2 章では関連研究を論じる．第 3 章では前提知識について説明する．第 4 章では，提案手法の全体を説明する．第 5 章では，実験とその結果・考察について論じる．第 6 章では，本研究のまとめを示す．

2　関 連 研 究

外部情報源と知識ベースの統合は長期間に渡って多くの研究があり，3 つのグループに分類できる．

第一に，マッピングルールを用いて外部情報源全体を RDF に変換し，RDF データセットを保存する方式である [2]．得られた RDF データセットは，以下に示すフェデレーション技術で利用可能で，SPARQL を介して既存の知識ベースとともにアクセスすることができる．RDF への変換には，主に Extract Transform Load (ETL) モデルの手法が使われる．CSV の変換には CSV2RDF[3]があり，多くのファイルフォーマットに対応した SPARQL Anything [3] という手法もある．この手法では，最初に外部情報源の全体を RDF に変換する．このため，動的に変化する外部情報源を統合するのは容易ではない．

第二に，各外部情報源の上に RDF ビューを提供するラッパーを提供する手法である．ラッパーは，問合せが与えられ

1：https://metmuseum.github.io/
2：https://www.openstreetmap.org/
3：https://github.com/AtomGraph/CSV2RDF

ると外部情報源にアクセスし，事前に登録されたマッピング・ルールに基づいて問合せ結果を返す [4]．Ontology-Based Data Integration (OBDI) は，ラッパー／メディエータに基づく手法で [4] [5] [6]，Obi-Wan [7] のような手法が存在する．

第三に，複数の SPARQL エンドポイントにまたがって問合せを行うフェデレーションがという手法がある．一旦，何らかの手段で外部情報源が SPARQL エンドポイントとしてアクセス可能な状態となったとき，知識ベースと外部情報源は統合的に問合せが可能となる．フェデレーションには，統計情報 [8] を使った手法や ASK 問合せを使う手法 [9] [10]，複数の手法を組合わせた手法がある [11]．

これらの統合アプローチは，重複エンティティが発生する可能性がある．重複エンティティとは，RDF 上では異なるエンティティであるが，実際の世界では同じ事物に対応するエンティティである．この問題に対処するため，外部情報源内で識別情報が利用可能な場合，多くの手法では `owl:sameAS` トリプル [12] や共通 IRI [13] を生成して対処する．ETL モデルの場合，統合後に重複エンティティが特定して削除することができ [2] [14]，例えば GNN を使う手法がある [15]．

重複エンティティの特定は，非 SPARQL 問合せ処理の分野でも研究がある．QuERy [16] は RDB における重複エンティティの研究である．重複エンティティを特定するオペレータは計算量が多いため，重複解消を含めた全体の問合せ処理の計算量を削減する手法を提案した. [17] は，エンティティとその属性値を持ったハーパーエッジ構造に対する重複解消の手法である．

エンティティリンキング (EL) は，自然言語文章の解析の文脈で研究されてきた．これは，知識ベース上のエンティティを自然言語中の現実世界の物事を指している mention（単語・語句など）に割り当てる技術である [18]．エンティティリンキングは 3 つの段階に分けられる．Mention Detection， Candidate Selection (CS)，そして Linking Decision (LD) である．Mention Detection は，文章中からエンティティを参照している mention を特定する．CS は，mention に該当する可能性のある候補エンティティを選ぶ．LD は，最終的に正しいエンティティを選ぶ [19]．TAGME [20] は，候補エンティティの関連度を LD に利用する．これは，エンティティ間の関連度を，Wikipedia 中のアンカーリンクの行先の重複度で計算するものである．また，Yamada et al. [21] は，エンティティと自然言語の単語を同時に BERT に入力するエンティティリンキング手法を提案した．

筆者の提案した *Knowledge Mediator* [1] は，Magic Properties を使用することで，外部情報源と KB を統合する．このさい，外部情報源から取得したオブジェクトに対応するエンティティを見つけるための，問合せ駆動型応需エンティティリンキングを提案した．

3 関 連 知 識

a） RDF:

RDF は，主語・述語・目的語からなるトリプルの集合として記述され，意味的情報を記述するために用いられる [4] [22]．主語は，現実の事物に対応するエンティティであり，IRI で指定される．目的語は，エンティティを示す IRI か，文字列や数字の様なリテラルである．述語は，主語と目的語の関係を示す IRI を持つ．SPARQL は，一般的に知識ベースに問合せるために用いられる言語である．問合せは知識ベース中の知識グラフに対してマッチングすべきパターンを示しており，これにより目的の情報を取得する [5]．SPARQL の問合せ結果は，各変数に対応したタプルの集合，すなわち，各変数とその該当する値が列として表現された表である．例えば，以下の様な問合せがあるとする．

```
SELECT DISTINCT * WHERE
{?x rdfs:label "Jeffrey Ullman"@en . }
```

WHERE 句は，知識グラフに対するマッチングのパターンを定義する．このパターンは，主語・述語・目的語を主要な構成要素とする．主語は，IRI または変数，目的語は IRI・リテラルまたは変数である．また，述語は IRI または変数である．主語・述語・目的語のパターンは，トリプルパターンと呼ばれる．トリプルパターンの集合は，WHERE 句の基本的な内容を示すグラフ ”Basic Graph Pattern (BGP) ”を形成する．上記の問合せを実行すると，?x 変数列に Jeffrey Ullman という名前を持つエンティティの IRI だけを含む表が得られる．

b） Magic Property:

SPARQL 処理系では，ユーザ定義の述語を指定できる [6,7]．この機能は Magic Property と呼ばれる．これは，知識処理の様々な応用の要求を満たすために，その応用法の固有の計算手順を SPARQL 問合せに組み込むことを意図している [8]．

Magic Property を定義するには，主語から述語の目的語を計算する関数，またはその逆の関数を与える必要がある．図 2 は，知識ベースに格納されている生年月日の値から年齢を求める Magic Property`udp:Age` の例である．年齢の情報は直接的には知識ベースには存在しないが，`udp:Age` に関連付けされた関数を利用して計算できる．SPARQL 問

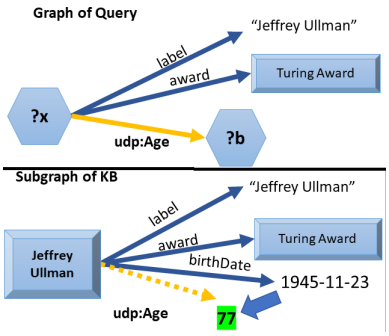


図 2 知識ベースに保存されている生年月日から年齢を取得する Magic Property の例

4：https://www.w3.org/TR/rdf11-concepts/
5：https://www.w3.org/TR/2013/REC-sparql11-query-20130321/
6：https://www.w3.org/Submission/spin-modeling/#spin-magic
7：https://jena.apache.org/documentation/query/writing_propfuncs.html
8：https://www.w3.org/2009/sparql/wiki/Feature:JavaScriptFunctions

合せに udp:Age が含まれているとき，SPARQL 処理系は dbr:Jeffrey_Ullman を引数として udp:Age に関連付けされた関数を呼び出し，年齢を計算する．これにより，トリプル ⟨dbr : Jeffrey_Ullman, dbo : birthDate,1945-11-23⟩ を取得し，現在の年齢（77 歳）を計算する．

本研究では，Magic Property を使用して外部情報源にアクセスし，その情報を知識ベースと統合する．

4 提 案 手 法

4.1 拡張知識ベース

知識ベース中の知識は，知識グラフ形式で表現される．元の知識グラフ中の述語を，*Knowledge Base Predicates (KBPs)* とする．また，Knowledge Mediator は *External Source Predicates (ESPs)* によって元の知識ベースを拡張する．ESP は，SPARQL の Magic Property として定義され，外部情報源へアクセスしつつ，主語エンティティから目的語エンティティを導出する．

図 3 は，ESP が知識グラフを拡張する様子を示す．元の知識ベースでは，調布市（dbr:Chōfu）や東京都府中市（Fuchū,_Tokyo）の様なエンティティやリテラルが頂点として表現され，KBP によって接続されている．dbr:Chōfu や Fuchū,_Tokyo は地理的には隣接しているものの，元の知識グラフではこの関係は表現されていない．OpenStreetMap (OSM) のような地図の外部情報源を利用することで，元の知識グラフを補強できる．OSM には座標やポリゴン・地図上のオブジェクト同士の関係などが記述されているため，調布市と東京都府中市が隣接しているという情報を提供できる．新しい ESP ESP:AdminDivNext を登録すると，この地理的な隣接関係を表現することができる．ESP:AdminDivNext に関連付けられた関数が OSM にアクセスし，地理的に隣接するオブジェクトを得る．そして，拡張知識グラフの中で，dbr:Chōfu や Fuchū,_Tokyo が ESP ESP:AdminDivNext によって互いに接続される．ユーザは拡張知識グラフに対して通常の SPARQL 問合せを問合せるだけで，知識ベースへの問合せ・エンティティリンキングを含む問合せ処理を自動的に行う．

4.2 アーキテクチャ

図 1 に提案手法のアーキテクチャを示す．前述の通り，提案手法は元の知識ベースを ESP で拡張する．また，ユーザは拡張知識グラフに対して通常の SPARQL のように問合せできる．第 1 章の SPARQL 問合せ例には，ESP ESP:AdminDivNext が含まれている．ESP の主語エンティティに対する目的語エンティティは，OSM にアクセスすることで得られる．この例では，KBP dbp:establishedDate を使用して，調布市の設立日を知識ベースから取得できる．隣接する市町村は，ESP ESP:AdminDivNext を使用して取得できる．この ESP を評価することにより，トリプル ⟨dbr:Chōfu, ESP : AdminDivNext,dbr:Fuchū,_Tokyo ⟩ を得る．

以降の節では，Knowledge Mediator の設計の詳細を示す．特に，以下の 2 点について詳しく説明する．

ESP の登録: 多岐にわたる外部情報源に対する，統一的な ESP の登録方法．

拡張 SPARQL 問合せの実行: ESP を含む拡張知識ベースに対する SPARQL 問合せの問合せ処理アルゴリズム．

簡単のために，これ以降では，ESP の目的語はエンティティであり，ESP は主語から目的語への方向で評価されると仮定する．つまり，ESP は，与えられた主語エンティティに対して目的語エンティティを導出する．しかし，この制限の緩和は容易である．

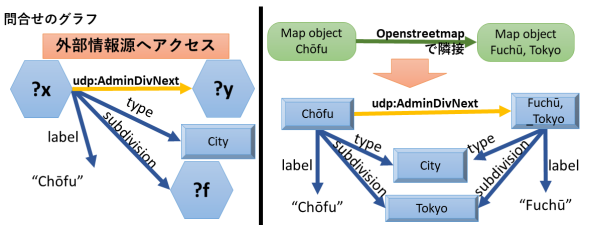


図 3 拡張 BGP の例（左側）と，この BGP に関係する知識ベースのグラフ（右側）

4.3 ESP の登録

ESP に関わるトリプルを計算するためには，(1) 主語エンティティから外部情報源へとアクセスするためのキーとなる情報を生成，(2) 生成したキーを用いて外部情報源にアクセスする関数，(3) 外部情報源から取得したオブジェクトに対して，知識ベース上のエンティティを割当てるという 3 つの機能が必要である．図 4 はこれを概念的に示した図である．ESP を登録するには，この 3 つの機能を登録する必要がある．そのため，ESP 登録時には以下の項目を登録する．

a) ESP の IRI: ESP の IRI を登録する（例えば ESP:AdminDivNext）．この IRI が問合せに現れると，問合せ処理系は以下の b),c),d) に示す関数を呼び出す．

b) キー生成関数 $Key()$: 外部情報源は，キーを介してアクセスできるとモデル化する．与えられた主語エンティティ x からキーを生成するために，上記 (1) の処理を行う関数を登録する必要がある．ESP 登録者は，主語が特定された場合の SPARQL 問合せを指定すると，$Key()$ が定義できる．後述するアクセス関数 $Access()$ に応じて，主語エンティティの名前や座標などがキーとして利用するよう定義する．前述の例では，x =dbr:Chōfu のとき，”Chōfu”は rdfs:label の目的語であるため， x =dbr:Chōfu に対して $Key(x)$ =”Chōfu”が生成される．

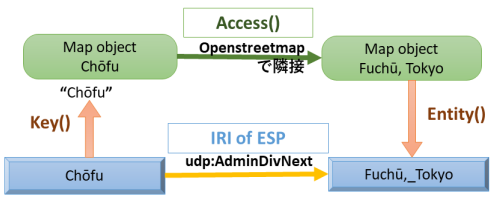


図 4 各 ESP の処理

c) 外部情報源アクセス関数 $Access()$**:** ユーザは，b) で定義した $Key(x)$ を用いて実際に外部情報源にアクセスし，目的語エンティティに関する情報を取得する関数（プログラムコード）を登録する必要がある．外部情報源のオブジェクト集合 $R = Access(Key(x))$ は外部情報源から取得できる．前述の例では，$Access()$ 関数が $Key(x) =$ "Chōfu" を利用して OSM に問合せを行い，隣接する自治体を示すオブジェクトの集合 R を得る．

d) エンティティリンキング関数 $Entity()$**:** 外部情報源のオブジェクト $r \in R$ に対して，知識ベースの中から該当するエンティティ（の集合）を発見し，エンティティの集合 $Y = Entity(x, r, C)$ として返す必要がある．ただし，C は問合せ処理途中における目的語側の変数に存在するエンティティの集合である．知識ベース中に r に該当するエンティティが存在しない場合，$Entity(x, r, C)$ は空集合を返すことを要件とする．通常，1 つの外部情報源のオブジェクトに対して 1 個ないし 0 個のエンティティが該当すると考えられるが，ESP の登録者が外部情報源のオブジェクトの性質を勘案して必要であると考えた場合には，2 個以上の目的語エンティティを割当てることも可能とする．ESP の登録者は，外部情報源のオブジェクトの性質と，ESP の取り扱うドメインに配慮して，$Entity()$ を設計する．例えば，主語エンティティと目的語エンティティが知識グラフ上で近い関係であると期待される場合は，主語エンティティとの関連度の強いエンティティを選ぶといったように設計する．

上記の 4 つの項目が登録されると Knowledge Mediator はこの ESP を利用した問合せを処理できる．

4.4 問合せ処理

拡張 SPARQL 問合せは，従来の SPARQL 問合せに ESP を追加したものである．拡張 SPARQL 問合せの WHERE 句を表す BGP は有向ラベル付きグラフであり変数・エンティティ・リテラルを表す頂点と，KBP および ESP を表す辺で構成される．

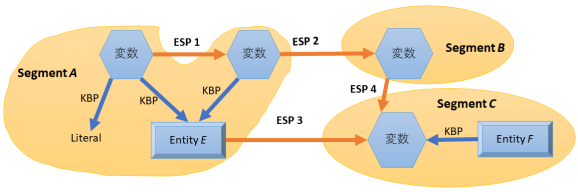


図 5　拡張 BGP とセグメントの例

BGP から ESP による辺を除外すると，BGP はいくつかの連結部分グラフに分解できる．これ以降，この各連結部分グラフをセグメントと呼ぶ．図 5 は，ESP を除外することで，拡張 SPARQL 問合せの BGP がセグメントに分割される様子を示す．セグメントには頂点として変数・エンティティ・リテラルが，辺として KBP が含まれる．セグメントは ESP によって接続され，各 ESP は主語側に主語セグメント・目的語側に目的語セグメントを持つ．

セグメント中の BGP は普通の SPARQL 問合せの BGP と同じなので，直接知識ベースに問合せできる．これにより，各セグメントの BGP を表す問合せ結果表を得る．このとき，各セグメントの問合せ結果表は変数だけではなく，問合せ中に直接記述されたエンティティに該当する列も持つ．

全体的な問合せ処理手順をアルゴリズム 1 に記述する．なお，$R = Access(Key(x))$ は ESP の外部情報源へのアクセスによる外部情報源のオブジェクトの集合，$r \in R$ に対する $Entity(x, r)$ はエンティティリンキングによって得られた目的語エンティティ，$subjects(p)$ は ESPp の主語側にある主語エンティティの集合である．以下は，アルゴリズム 1 を設計するうえで，考慮した基本的な点である．

(1) ESP の $Key()$ と $Access()$ を実行するとき，主語エンティティが特定されている必要がある．

(2) ESP の $Access()$ と $Entity()$ を実行するとき，外部情報源へのアクセスとエンティティリンキングを伴う．この処理は処理時間が多くかかるため，呼び出し回数を最小にする必要がある．

(3) ESP に対して $Key()$, $Access()$, $Entity()$ を実行した後，ESP をインスタンス化した主語と目的語のエンティティの組が得られ，通常の問合せ結果表として扱える．

Algorithm 1 問合せ処理

Require: 拡張 BGPG
Ensure: G の問合せ結果
1: G の全てのセグメントを見つける
2: Triple Patterns を含むすべてのセグメントに対して問合せを評価し，その結果を中間結果表の集合 I_{result} に追加する．
3: P_u を未処理の ESP の集合とする．
4: **while** $count(P_u) \neq \emptyset$ **do**
5: $p_{min} = \arg\min_{p \in P_u} count(subjects(p))$ (複数の ESP が同じ最低値を出した場合，ランダムに 1 つ選ぶ)
6: $X = subjects(p_{min})$
7: **for** $x \in X$ **do**
8: $R = Access(Key(x))$：外部情報源にアクセスする．
9: **for** $r \in R$ **do**
10: $Y = Entity(x, r)$：目的語エンティティを割当て (Y は空集合の場合がある)．
11: **end for**
12: **end for**
13: ESPp_{min} の評価結果 p_{min} と，その主語セグメントと目的語セグメントの中間結果表を結合し，I_{result} を更新
14: $P_u = P_u - \{p_{min}\}$
15: **end while**
16: I_{result} から最終的な結果を出力．

5 実　　験

5.1 実験環境

ESP ごとにエンティティリンキングを登録することの利点を検証するため，提案手法のプロトタイプを実装・ESP を登録し，提案手法のアルゴリズムがどの程度正確に問合せを実行するか

を検証した．知識ベースとしては，英語版 DBpedia の 2022 年 12 月のスナップショットを，外部情報源として OpenStreetMap とメトロポリタン美術館の API を用いた．実験時期は 2024 年 2 月である．

エンティティリンキングを登録する利点を検証するため，5.2 に示す ESP を，$Key()$, $Access()$, $Entity()$ を定めて登録した．エンティティリンキングを登録する利点を示すため，全ての $Entity()$ 関数を同一にそろえた場合と，各 ESP 向けに適切な $Entity()$ を登録した場合を比較する．

登録した ESP を含む 10 の拡張 SPARQL 問合せを使用した．問合せは BGP の構造・情報源の多様性を持つように設計した．10 の問合せに対する正解セットは著者が提供した．Knowledge Mediator が返す結果と正解セットを比較し，タプルを計算することで Recall・Precision・F1 を測定した．

5.2 登録 ESP

本研究では，以下の 5 つの ESP を実装した．なお，各アクセス関数 $Access()$ には，(Str) 文字列を返す場合と，(Srd)XML や JSON などの構造化データを返す場合の 2 通りを用意した．

5.2.1 udp:m:AdminDivNext

目的語エンティティは，地図上で指定された主語エンティティに隣接する行政区域である．

$Key()$ は，

```
SELECT * WHERE { <?subject> rdfs:label ?literal .
FILTER (lang(?literal) = 'en')}
```

を利用する．ただし，<?subject>は主語エンティティを指す．つまりこの $Key()$ は，主語エンティティの英語名を利用する．この $Key()$ 関数を $Key_{label}()$ とする．

$Access()$ 関数は，OpenStreetMap 上のオブジェクトを名前から Nominatim を用いて主語に該当するオブジェクトの ID を取得し，その ID に隣接するオブジェクトを OpenStreetMap に対する問合せサービスである Overpass を利用して取得する．この返り値を用い，`name:en` 属性からオブジェクトの名前の集合 (Str) を返す．または，個々のオブジェクトに該当する XML の集合 (Srd) を返す．

$Entity()$ は，[1] を参考に，(1) 該当する可能性のあるエンティティを選出する候補エンティティ選出 $CS()$，目的語のセグメントによる制約から候補エンティティを絞り込む段階，そして，最終的にエンティティを一つに定める段階 $ED()$ という形で設計した．$CS()$ は，外部情報源のオブジェクトの名前，もしくは XML の全ての value に対して，辞書を使って該当するエンティティを選出する．辞書は，エンティティ名を示す `rdfs:label` だけでなく，`dbo:wikiPageRedirects`，`dbo:wikiPageDisambiguates` を用いて別名を取得した．また，名前もしくは value がトークン単位で辞書に部分一致した場合に，候補エンティティとする．$ED()$ は，TAGME [20] を参考に，$dbo:wikiPageWikiLink$ を用いて主語エンティティと目的語エンティティの関連度関数を作った．関連度が閾値 0.4 以上かつ最大値の場合に，該当するエンティティとする．この $Entity()$ を $Entity_{alias}()$ とする．

5.2.2 udp:m:Dist5km, udp:m:Dist5km

目的語エンティティは，地図上で指定された主語エンティティから 1km (5km) の地図オブジェクトである．

$Key()$ は，`SELECT * WHERE {`

```
<?subject> georss:point ?literal . }
```

を利用する．すなわち，主語エンティティの座標を利用する．この $Key()$ 関数を $Key_{geo}()$ とする．

$Access()$ 関数は，Overpass を利用して $Key()$ の座標から 1km(5km) 以内の名前付き地図オブジェクトを取得する．`name:en` 属性からオブジェクトの名前の集合 (Str) を返す．または，個々のオブジェクトに該当する XML の集合 (Srd) を返す．

$Entity()$ は，$Entity_{alias}()$ と類似しているが，辞書についてエンティティの別名は用いない．これを $Entity_{label}()$ とする．

5.2.3 mp:isArtistOfAtMet

目的語エンティティは，メトロポリタン美術館の収蔵品のうち，主語エンティティによって製作されたものである．

$Key()$ は，$Key_{label}()$ を用いる．

$Access()$ 関数は，メトロポリタン美術館の API を用い，主語エンティティの名前から作品のオブジェクトを検索する．`title` 属性からオブジェクトの名前の集合 (Str) を返す．または，個々のオブジェクトに該当する JSON の集合 (Srd) を返す．

$Entity()$ は，$Entity_{alias}()$ を用いる．

5.2.4 mp:isCityOfArtistAtMet

目的語エンティティは，主語エンティティの地域に関連付けされた作品を制作した人物である．

$Key()$ は，$Key_{label}()$ を用いる．

$Access()$ 関数は，メトロポリタン美術館の API を用い，主語エンティティの地名から作品のオブジェクトを検索し，作品のオブジェクト内にある作者オブジェクトを抽出する．`name` 属性からオブジェクトの名前の集合 (Str) を返す．または，個々のオブジェクトに該当する JSON の集合 (Srd) を返す．

$Entity()$ は，$Entity_{label}()$ と類似しているが，主語エンティティの代わりにメトロポリタン美術館のエンティティとの関連度をとる．また，閾値は 0.0 とする．これを $Entity_{met}()$ とする．

5.3 実験結果

表 1 は，実験結果の F1・Precision・Recall を示したものである．提案手法に当たる best は，5.2 節で示した通りである．ただし，OpenStreetMap は構造化データを，メトロポリタン美術館の API にはエンティティ相当の文字列を用いたものである．提案手法は，全ての問合せにおいて最高の F1, Recall を達成した．

問合せ 1,2,3,6,7,8 は，`udp:m:AdminDivNext` を 1〜3 つ用いた問合せである．この ESP は隣接する自治体を指すために，主語との関連度を用いることがエンティティを選ぶうえで有効に働くことがわかる．また，地名は多少の曖昧性を含むため，別名を利用することがエンティティを選ぶことに寄与する．

問合せ 4 は，`udp:m:Dist5km` を用い，つくば市役所から 5km

表 1 EL 手法の性能評価．問合せに対する F1(F)・Recall (R)・Precision (P)．すべてのエンティティが Null の場合，値は 0 と見なす．Str は *Access*() 関数がエンティティ相当の文字列を返す場合，Srd は *Access*() 関数が XML や JSON などの構造化データを返す場合を示す．また，best は各 ESP に対して適切な *Entity*() を登録した場合を示す．また，label は全ての ESP に $Entity_{label}$() を，alias は全ての ESP に $Entity_{alias}$() を，met は全ての ESP に $Entity_{met}$() を利用したことを示す．

		Avl	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
best	F	**0.91**	**1.00**	**0.92**	**0.92**	**0.67**	**0.77**	**1.00**	**0.94**	**0.94**	**1.00**	**0.91**
	R	**0.92**	**1.00**	**1.00**	**1.00**	**0.67**	**0.63**	**1.00**	**0.94**	**0.94**	**1.00**	**1.00**
	P	**0.91**	**1.00**	0.86	0.85	0.67	**1.00**	**1.00**	**0.94**	**0.94**	**1.00**	**0.83**
Str label	F	0.70	**1.00**	0.91	0.90	0.50	0.22	**1.00**	**0.94**	**0.94**	0.57	-
	R	0.64	**1.00**	0.83	0.82	0.33	0.13	**1.00**	**0.94**	**0.94**	0.40	0.00
	P	0.89	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.94**	**0.94**	**1.00**	-
Srd label	F	0.77	**1.00**	0.91	0.90	**0.67**	**0.77**	**1.00**	**0.94**	**0.94**	0.57	-
	R	0.72	**1.00**	0.83	0.82	**0.67**	**0.63**	**1.00**	**0.94**	**0.94**	0.40	0.00
	P	0.86	**1.00**	**1.00**	**1.00**	0.67	**1.00**	**1.00**	**0.94**	**0.94**	**1.00**	-
Str alias	F	0.73	**1.00**	**0.92**	**0.92**	0.18	0.36	**1.00**	**0.94**	**0.94**	**1.00**	-
	R	0.75	**1.00**	**1.00**	**1.00**	0.33	0.25	**1.00**	**0.94**	**0.94**	**1.00**	0.00
	P	0.74	**1.00**	0.86	0.85	0.13	0.67	**1.00**	**0.94**	**0.94**	**1.00**	-
Srd alias	F	0.77	**1.00**	**0.92**	**0.92**	0.29	0.71	**1.00**	**0.94**	**0.94**	**1.00**	-
	R	0.82	**1.00**	**1.00**	**1.00**	**0.67**	**0.63**	**1.00**	**0.94**	**0.94**	**1.00**	0.00
	P	0.76	**1.00**	0.86	0.85	0.18	0.83	**1.00**	**0.94**	**0.94**	**1.00**	-
Str met	F	0.67	**1.00**	0.80	0.84	0.50	0.22	**1.00**	0.94	-	0.50	**0.91**
	R	0.62	**1.00**	0.67	0.73	0.33	0.13	**1.00**	**0.94**	0.00	0.40	**1.00**
	P	0.84	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.94**	-	0.67	**0.83**
Srd met	F1	0.73	**1.00**	0.80	0.84	0.50	0.77	**1.00**	**0.94**	-	0.50	**0.91**
	R	0.70	**1.00**	0.67	0.73	0.67	0.63	**1.00**	**0.94**	0.00	0.40	**1.00**
	P	0.78	**1.00**	**1.00**	**1.00**	0.40	**1.00**	**1.00**	**0.94**	-	0.67	**0.83**

以内の駅を問合せた．この ESP は距離の近い地図オブジェクトを指すために，主語との関連度が有効である．また，店舗などの DBpedia 上ではエンティティとなっていない地図オブジェクトを大量に返される．このため，別名を用いた場合は不要なエンティティを多く選出する．よって，別名を用いない方式の方が高い性能を発揮する．OpenStreetMap の地図オブジェクトは，名前相当の value を複数有するため，名前相当の属性を指定するよりも，外部情報源のオブジェクトの全ての value を使うほうが性能が良い傾向がある．

問合せ 9 は，`mp:isArtistOfAtMet` を用い，メトロポリタン美術館に収蔵されているフェルメールの作品を問合せる．主語と目的語の関係は強い事が期待されるため，関連度を用いることが有効である．また，メトロポリタン美術館の作品名は英語であるが，非英語の作品に英語名を与える際に複数の翻訳が有り得る．このため，`rdfs:label` で取得できるエンティティ名とは異なる名前であることも少なくないため，別名を辞書に利用することが重要となる．

問合せ 10 は，*mp:isCityOfArtistAtMet* を用い，ニューヨーク市に関連する作品を作った作者を問合せる．ただし，作品はメトロポリタン美術館の収蔵品に限定する．これは，この ESP は主語と目的語の関連性は必ずしも強くないと考えられ，$Entity_{label}$() や $Entity_{alias}$() は機能しない．しかしながら，作品がメトロポリタン美術館に収蔵されている以上，メトロポリタン美術館への関連度はある程度は期待できるため，閾値を無くしたうえでメトロポリタン美術館との関連度を利用することに効果がある．

問合せ 5 は，`udp:m:AdminDivNext` と `udp:m:Dist1km` とを用い，つくば市に隣接する都市の市役所から 1km にある駅を問合せたものである．この結果から，各 ESP に対して有効なエンティティリンキングを登録する手法が，複数の ESP が存在する問合せでも有効性があると言える．

以上より，外部情報源のオブジェクトの性質やドメインから考えられるエンティティリンキングを ESP 別に登録することが，問合せ処理の精度に対して有効に働くことが判明した．

6 ま と め

RDF 形式の知識ベースは，静的に意味的な情報を集積し，SPARQL によって問合せることが可能である．一方で，非 RDF の外部情報源は独自のアクセス方法や，問合せおよび返り値に対して独自のフォーマットを持つことが多い．以前筆者は，この問題に対して Magic Property を利用して知識ベースと非 RDF 外部情報源を統合する Knowledge Mediator を提案した．しかし，外部情報源が構造化データの場合は，エンティティの名前相当となる文字列の場所を特定して渡す必要があるものだった．そこで，本研究では，Knowledge Mediator を更に拡張し，構造化データを含めた外部情報源を利用し，外部情報源へのアクセス結果を問合せ結果に組み込む述語 (ESP) に対して適切なエンティティリンキングを登録できるようにすることで，返り値の形式やドメインの違いに対応できるアーキテクチャを提示した．また，実験を通じて，エンティティリンキング手法を固定するより，ESP に登録する方式の方が高い精度を発揮することを示し，アーキテクチャ有効性を検証した．

将来の課題としては，複数種類の ESP を持つ問合せをより多く用意しての実験である．また，本研究の提示したアーキテクチャは，外部情報源のオブジェクトを指す文字列や，構造化データだけでなく，画像などの外部情報源のオブジェクトも処理できることが期待できるため，より多様な外部情報源を利用した検証も行することも重要である．

7 謝　　辞

本研究の一部は，JSPS 科研費 (JP23H03399, JP22K19802, JP22H03694)，JST CREST(JP-MJCR22M2)，NEDO (JPNP20017), AMED (JP21zf0127005) による．

文　　献

[1] Yuuki Ohmori, Hiroyuki Kitagawa, Toshiyuki Amagasa, and Akiyoshi Matono. Integration of knowledge bases and external information sources via magic properties and query-driven entity linking. In *The 25th International Conference on Information Integration and Web Intelligence*, pp. 309–324, 2023.

[2] Gytundefined Tamašauskaite and Paul Groth. Defining a knowledge graph development process through a systematic review. *ACM Trans. Softw. Eng. Methodol.*, Vol. 32, No. 1, 2023.

[3] Luigi Asprino, Enrico Daga, Aldo Gangemi, and Paul Mulholland. Knowledge graph construction with a façade: A

unified method to access heterogeneous data sources on the web. *ACM Trans. Internet Technol.*, Vol. 23, No. 1, pp. 1–31, 2023.

[4] Diego Calvanese and et al. Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, Vol. 8, No. 3, pp. 471–487, 2017.

[5] Maxime Buron and et al. Ontology-based RDF integration of heterogeneous data. In *Proc. 23rd International Conf. on Extending Database Technology*, pp. 299–310, 2020.

[6] Fajar Ekaputra and et al. Ontology-based data integration in multi-disciplinary engineering environments: A review. *Open J. Inf. Syst.*, Vol. 4, No. 1, pp. 1–26, 2017.

[7] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. Obi-wan: Ontology-based rdf integration of heterogeneous data. *Proc. VLDB Endow.*, Vol. 13, No. 12, pp. 2933–2936, 2020.

[8] Muhammad Saleem and et al. CostFed: Cost-based query optimization for sparql endpoint federation. *Procedia Computer Science*, Vol. 137, pp. 163–174, 2018.

[9] Olaf Görlitz and Steffen Staab. SPLENDID: Sparql endpoint federation exploiting void descriptions. In *Proc. 2nd Int. Conf. on Consuming Linked Data*, pp. 13–24, 2011.

[10] Andreas Schwarte and et al. FedX: Optimization techniques for federated query processing on linked data. In *Proc. 21st International Semantic Web Conf.*, pp. 601–616, 2011.

[11] Muhammad Saleem and Axel-Cyrille Ngonga Ngomo. HiBISCuS: Hypergraph-based source selection for sparql endpoint federation. In *Proc. 11th Extended Semantic Web Conf.*, pp. 176–191, 2014.

[12] Diego Calvanese, Martin Giese, Dag Hovland, and Martin Rezk. Ontology-based integration of cross-linked datasets. In *The Semantic Web - ISWC 2015*, pp. 199–216, 2015.

[13] Guohui Xiao, Dag Hovland, Dimitris Bilidas, Martin Rezk, Martin Giese, and Diego Calvanese. Efficient ontology-based data integration with canonical iris. In *The Semantic Web*, pp. 697–713, Cham, 2018.

[14] Vassilis Christophides and et al. An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*, Vol. 53, No. 6, 2020.

[15] Kexuan Xin and et al. Large-scale entity alignment via knowledge graph merging, partitioning and embedding. In *Proc. of the 31st ACM International Conf. on Information and Knowledge Management*, pp. 2240–2249, 2022.

[16] Hotham Altwaijry, Sharad Mehrotra, and Dmitri V. Kalashnikov. QuERy: A framework for integrating entity resolution with query processing. *Proc. VLDB Endow.*, Vol. 9, No. 3, pp. 120–131, 2015.

[17] Indrajit Bhattacharya and Lise Getoor. Query-time entity resolution. *J. Artif. Int. Res.*, Vol. 30, No. 1, pp. 621–657, 2007.

[18] Wei Shen and et al. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, Vol. 27, No. 2, pp. 443–460, 2015.

[19] Özge Sevgili and et al. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, Vol. 13, No. 3, pp. 527–570, 2022.

[20] Paolo Ferragina and Ugo Scaiella. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. 19th ACM International Conf. on Information and Knowledge Management*, pp. 1625–1628, 2010.

[21] Ikuya Yamada and et al. Global entity disambiguation with BERT. In *Proc. 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3264–3271, 2022.

[22] Farzaneh Mahdisoltani, Joanna Asia Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *Proc. 7th Conf. on Innovative Data Systems Research*, 2015.

A Polystore System for Graph and Key-Value Databases based on SQL/PGQ

Weijun WONG†, Qiong CHANG†, and Jun MIYAZAKI†

† School of Computing, Tokyo Institute of Technology
2 Chome-12-1 Ookayama, Meguro City, Tokyo 152-8550, Japan
E-mail: †wong@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.ac.jp

Abstract It is common for organisations to use many databases with different data models to best address their business needs. Polystore systems aim to provide a single database management system solution that supports multiple heterogenous data stores. In this paper, we implement a middleware that supports a Graph database and a Key-Value database based on the SQL/PGQ extension to the SQL standard recently added in SQL:2023. Evaluating our middleware, we find that it provides a high level of autonomy to component databases, but is lacking in other areas especially in terms of transparency and optimality.

Key words Polystore System, NoSQL Database, SQL/PGQ, Graph Database, Key-Value Database

1 Introduction

The "one size fits all" idea that relational databases provide the best solution for storing and processing data is not applicable to today's context. Many database systems that are designed to support data models that represent the characteristics of the stored data better than the relational model have emerged. These group of database systems, collectively referred to as NoSQL databases, in contrast to SQL relational databases, include graph databases, document databases, key-value databases, time-series databases, and vector databases among others. It is common for organisations to use multiple types of databases to store and process their data.

However, the increase in variety of data models has brought about an increase in the variety of query interfaces used to store and retrieve data, as databases provide their own query languages or APIs (application programming interfaces) with which to handle data. Many relational query languages are dialects of SQL, which became an ISO standard in 1987, so they look similar. However, other data models like the graph model have a variety of query languages with different syntax. On the one hand, this is useful as query interfaces may be designed to fit the data's characteristics, allowing for more concise queries to be written. On the other hand, the lack of standardisation means users have to invest effort into learning a new query interface when switching from one database system to another.

The idea of polystore systems was conceived to address this issue. In a polystore system, a single common query interface can be used to query a database with multiple underlying storage engines, each with their own data models. In this paper, we present a middleware that uses a customised version of the SQL/PGQ (SQL/Property Graph Queries) extension to the SQL standard [1] which supports a graph databases and a key-value database. This middleware allows the user to use the graph database and key-value database of their choosing, but query them using a common query language.

2 Related Work

The usefulness of having different data models besides the relational model has resulted in numerous efforts to provide interoperability between different data models from both academia and industry. Different approaches for providing such integrations exist. In the following discussion, we focus on interoperability between the relational and graph data models.

2.1 Query Translation

One approach is that of query translation, where the data model paradigm of the query language is different from the data model of the storage engine. In such an approach, the query written by the user is first translated to the native query language used by the database before it is executed. For example, SQLGraph [2] translates Apache TinkerPop's Gremlin language into SQL, while SQL-Gremlin [3] is a tool to translate SQL into Gremlin. Similarly, Cyp2SQL [4] and Cytosm [5] translate Neo4j's Cypher into SQL.

The benefits of this approach include being able to use a new query language paradigm without having to migrate data to a new database, as well as being able to write queries in a familiar query language to be executed on a database that has an unfamiliar query language. However, the drawback is that as the query language and storage engine use different

data models, the storage engine might not be optimised for the query language's data model.

2.2 Multi-model Databases

As their name implies, multi-model databases provide support for multiple data models. The usefulness of data models other than the relational model has led to some relational databases adding support for other data models, such as how PostgreSQL and Oracle DB have added support for semi-structured XML and JSON data. Other databases such as ArangoDB [6] and OrientDB [7] are designed ground-up to natively support multiple data models.

Tan et al. propose a taxonomy to classify database systems with support for multiple data models [8]. They identify four different categories, namely *federated database system*, *polyglot system*, *multistore system* and *polystore system*, based on the number of query interfaces and data models that the system supports. Table 1 summarises these four categories.

	Number of Query Interface Paradigms	
Data Models	Single	Multiple
Homogeneous	Federated Database System	Polyglot System
Heterogeneous	Multistore System	Polystore System

Table 1: Types of Multi-Model Systems

Polystore systems combine the advantages of polyglot systems and multistore systems. That is, users are able to use different query paradigms to access data that is also stored across multiple heterogeneous data stores; so the query language to use is not determined by the data model. Polystores have the expressiveness of multiple query language paradigms, together with the performance gains of storing data using a suitable data model. Polystore systems include BigDAWG [9], CloudMdsQL [10], Myria [11] and Apache Drill [12].

While multi-model databases may not have the same drawbacks the query translation approach has, data migration is required. Many systems also provide their own mechanisms for manipulating data that are of different data models at the same time.

3 Proposed Methodology

The goal of this paper is to produce a common query interface for the graph and the key-value data models. We designed our application to have the advantages of both of the approaches introduced in Section 2. That is, our query language is based on existing commonly-used query languages, which supports both the graph and key-value data model paradigms. We also choose to implement it as a middleware application, which will allow users to use our language on existing databases without the need for data migration. This also has the benefit of the storage engines being optimised for the respective data model(s) of the component database.

3.1 Query Language

We propose a custom version of SQL/PGQ as a common query language in our middleware, which supports graph databases and key-value databases. SQL/PGQ itself is an extension to the SQL standard which adds read-only Cypher-like graph queries. The graph-like syntax allows for more concise queries to be written while data is still stored in the relational model. SQL/PGQ is designed to comply with GQL [13], which is a work-in-progress ISO standard for a graph database query language, similar to how SQL is a standard for a relational database query language. For queries for key-value data, we model the key-value database as a single table with a `key` column and a `value` column. Queries for key-value data are similar to SQL queries for relational data, with only a limited set of operations supported.

Basing our query language on SQL/PGQ offers the following advantages: Firstly, it provides a standard syntax and semantics for interacting with relational data and graph data simultaneously. Secondly, as existing graph query languages such as Cypher and PGQL serve as the inspiration for the design of SQL/PGQ [14], the syntax will be familiar to users of graph databases. Thirdly, designing our query language based on SQL/PGQ gives us the opportunity to extend our query language in the future to support the relational model and document model in the form of XML or JSON data, which are also supported by the SQL standard [15] [16].

3.2 System Architecture

Our middleware application implements the mediator-wrapper architecture [17] to realise a loose-coupling between the application and component databases. This allows the user to choose what component databases to use with our query language. Figure 1 describes the overall design of the middleware. The mediator part of the application consists of the parser, executor and aggregator. Wrappers act as the bridge between our application and the component databases' native APIs.

After the user inputs the query, the parser identifies the relevant part of the query that is intended for the graph database or key-value database. The parser decomposes the input query into *query fragments*, which are logical query blocks divided based on the database that is meant to execute them. More specifically, an input query is decomposed into 1) graph query fragments, to be executed by the graph database; 2) key-value query fragments, to be executed by the key-value database; and 3) outer query fragments, to be executed by the aggregator. Graph, key-value and outer query fragments

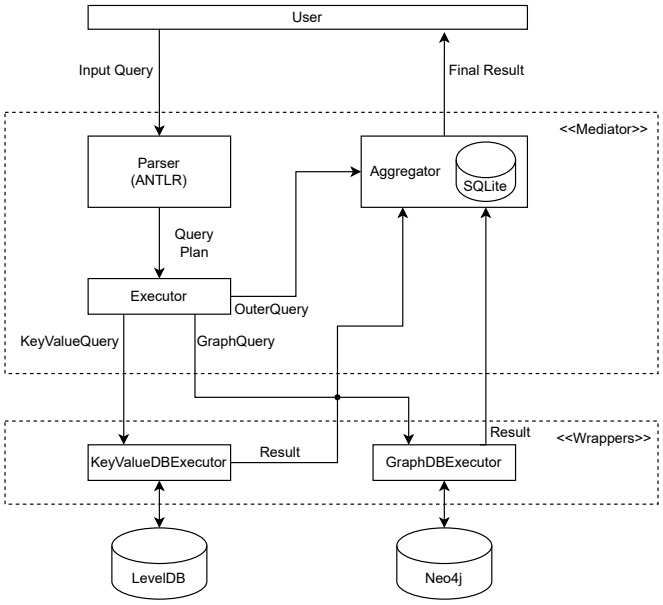


Figure 1: Overall Architecture

are expressed as a Cypher query, a list of key-value pairs, and an SQL query respectively. A sequence of query fragments generated from the input query is called the *query plan*. The query plan is sent from the parser to the executor. Figure 2 shows the query plan that is generated from an input query.

```
SELECT n_id
FROM GRAPH_TABLE(neo4j
    MATCH (n IS TestNode)
    WHERE n.property = "testProperty"
    COLUMNS (n.id AS n_id, n.property AS n_property)
);
```

(a) Input Query

```
// Graph Query Fragment
t0 {
    // openCypher
    USE neo4j
    MATCH (n:TestNode)
    WHERE n.property = "testProperty"
    RETURN n.id AS n_id, n.property AS n_property;
}

// Outer Query Fragment
t1 {
    // SQL
    SELECT n_id FROM t0;
}
```

(b) Query Plan

Figure 2: Query Plan Generation

The executor is responsible for knowing what component databases are associated with the application, and for sending query fragments to the appropriate database for execution. Component databases are associated to the application through registering the wrapper for the database with the executor. Wrappers must be implemented by the user. A component database declares its data model (graph or key-value) to the application via its wrapper, which implements the `GraphDBExecutor` or `KeyValueDBExecutor` interfaces.

```
// Graph Database
class Neo4jExecutor implements GraphDBExecutor {
    ...
}

// Key-Value Database
class LevelDBExecutor implements KeyValueDBExecutor {
    ...
}

// In Main class
Executor executor = new Executor();
executor.setGraphDBExecutor(new Neo4jExecutor());
executor.setKeyValueDBExecutor(new LevelDBExecutor());
```

Figure 3: Adding Component Databases

A wrapper executes a query fragment using the component database's driver. Delegating the execution to the databases allows us to take advantage of the capabilities of the underlying databases' storage engines.

If the input query is a data retrieval operation and a result is expected, output data from both the graph database and the key-value database is restructured into a tabular format and passed to the aggregator, which is an in-memory SQLite database [18]. The aggregator collates results from the component databases and performs additional transformations as defined by the outer query fragments, after which the final result is returned to the user. Every query fragment has an internally assigned `tableId`, which is the name of the table created in the aggregator to contain the results returned from the component databases. Outer query fragments directly reference these `tableId`s, as can be seen in Figure 2(b).

In summary, our system mainly differs from SQL/PGQ in the following ways:

- The `UPDATE GRAPH_TABLE` clause and `KVS` keyword are added to support data modification for graph data and key-value data operations respectively.
- SQL/PGQ runs graph queries on graph views that are created on relational data; in our application, the data model of the storage engine itself is also graph and key-value.

4 Implementation Details

A proof-of-concept for our proposed system was implemented using Java 17. The parser was built using ANTLR

4.13.1 [19], an free lexer-parser generator. The parser grammar was built using a PostgreSQL ANTLRv4 grammar [20], which was chosen for PostgreSQL's emphasis on conformance with the SQL standard [21]. Neo4j 5.15.0 Community Edition [22] was used as the underlying graph database. LevelDB 1.13 [23] was used as the underlying key-value database. SQLite 3.44.1 [18] was used as an in-memory RDBMS to restructure the output from the graph and key-value databases into tabular format.

Our application was tested to work with the following component databases:

- Graph Database: Neo4j, Memgraph [24]
- Key-Value Database: LevelDB, Riak [25]
- Multi-Model Database: ArcadeDB [26]

5 Usage

5.1 Querying Graph and Key-Value Databases

In this section, we show how `CRUD` operations are performed for the underlying graph and key-value databases using our SQL/PGQ-based query language.

5.1.1 Graph Database

The SQL/PGQ extension only adds read-only graph queries to the SQL standard. However, data modification operations are extensions to SQL/PGQ. We have stuck closely with the philosophy of SQL/PGQ by enclosing a Cypher-like graph query within an `UPDATE GRAPH_TABLE` statement.

- **Create**

```
UPDATE GRAPH_TABLE(neo4j
    CREATE (n)
    SET n IS TestNode, n.property = "testProperty"
);
```

- **Retrieve**

```
SELECT *
FROM GRAPH_TABLE(neo4j
    MATCH (n IS TestNode)
    WHERE n.property = "testProperty"
    COLUMNS (n.property)
);
```

- **Update**

```
UPDATE GRAPH_TABLE(neo4j
    MATCH (n)
    WHERE n.property = "testProperty"
    SET n.property = "newProperty"
);
```

- **Delete**

```
UPDATE GRAPH_TABLE(neo4j
    MATCH (n)
    WHERE n.property = "testProperty"
    DETACH DELETE n
);
```

As with SQL/PGQ, for data retrieval for graph data, the `COLUMNS` clause is used to format the query result into a tabular format, and does not allow graph entities such as nodes and edges to be returned directly. However, information about a path from a source node to a destination node can be returned by including the optional `ONE ROW PER STEP` clause, as shown in Figure 4. Although Cypher does not have an equivalent `ONE ROW PER STEP` feature, we produce a graph query fragment that approximately returns the same result as shown in Figure 4(c). This query returns the shortest path from the source node to the destination node, in the form of a table with the ID of the start and end node of each step in the path (Figure 4(d)). Note that every edge in the path has to be bound to a variable to be included in the result (in the example, this is the variable `r`).

5.1.2 Key-Value Database

A key-value database can be thought of as a single table in a relational database. Therefore, the query syntax for key-value database is very similar to normal SQL queries. However, we have introduced a reserved keyword `KVS` for the user to specify that the query is intended for the key-value database and not the relational database. The functionality for key-value queries are simplified into the four `CRUD` operations, `Create`, `Retrieve`, `Update` and `Delete`.

- **Create**

```
INSERT INTO KVS
VALUES ("testKey1", "testValue1"), ("testKey2",
↪  "testValue2");
```

- **Retrieve**

```
SELECT * FROM KVS
WHERE key = "testKey1" OR key = "testKey2";
```

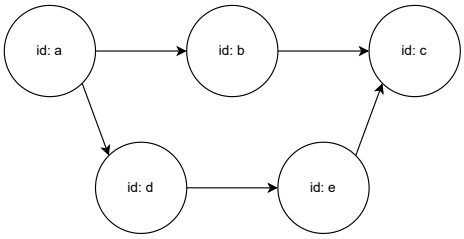
- **Update**

```
UPDATE KVS
SET value = "newValue"
WHERE key = "testKey1" OR key = "testKey2";
```

- **Delete**

```
DELETE FROM KVS
WHERE key IN ("testKey1", "testKey2");
```

While some key-value databases support advanced data retrieval operations such as efficient prefix matching of keys or partial matches based on user-defined filters, currently, only



(a) Graph

```
SELECT * FROM GRAPH_TABLE(
    MATCH p = ANY SHORTEST (src)-[r]->+(dest)
    WHERE src.id="a" AND dest.id="c"
    ONE ROW PER STEP (v1, e, v2)
    COLUMNS(v1.id, v2.id)
);
```

(b) Input Query

```
MATCH p = shortestPath((src)-[r*]->(dest))
WHERE src.id="a" AND dest.id="c"
UNWIND r AS e
WITH *, startNode(e) AS v1, endNode(e) AS v2
RETURN v1.id, v2.id;
```

(c) Graph Query Fragment

v1.id	v2.id
a	b
b	c

(d) Execution Result

Figure 4: Returning a Path as a Result

data retrieval via exact matches of keys, using the `WHERE` clause, are processed directly by the key-value database in our application. More advanced filtering requires that the entire key-value database is loaded into memory and processed by the SQLite aggregator instead, as can be seen in the query plan in Figure 5. We leave support for executing advanced data retrieval operations by the key-value database as future work.

5.1.3 Hybrid Query

Hybrid queries are also possible, allowing the user to use both graph data and key-value data at the same time. Currently, subqueries in the `FROM` clause are supported, but subqueries in the `WHERE` clause are not. A hybrid query example is shown later in Figure 7(a).

5.2 Sample Use Case

A possible use case for our middleware is that of storing lineage data. Lineage data refers to a documentation of how data resources are created, moved, transformed and consumed over time. Lineage data can be persisted using a graph database and a key-value database [27]. In such a system, each node

```
SELECT * FROM KVS
WHERE KEY LIKE "%partialKey%";
```

(a) Input Query

```
// Key-Value Query Fragment
t0 {
    GET ALL;
}

// Outer Query Fragment
t1 {
    SELECT * FROM t0
    WHERE KEY LIKE "%partialKey%";
}
```

(b) Query Plan

Figure 5: Filtering Key-Value Data

in the graph database represents a data entity. Directed edges between nodes represent dependencies between data. The key-value database is used to store the raw data that each node in the graph represents. Figure 6 illustrates such a model.

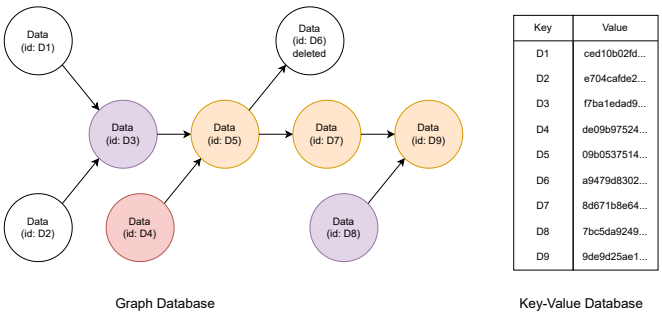


Figure 6: Lineage Data Model

Consider the action of a *database repair*, in which some data is discovered to be erroneous (red node) and needs to be corrected. At the same time, the minimum affected area of this false data, that is, all nodes that are directly or indirectly dependent on the erroneous node (orange nodes) need to be identified. To update these dependent nodes, the other unaffected data necessary for recalculation need to be obtained (purple nodes). Such an operation, which requires data to be obtained from both the graph database and the key-value database, can be expressed using our query language without the user needing to call the APIs of the individual databases separately. For example, the following query (Figure 7(a)) obtains all raw data from the key-value database that need to be updated. The results are sorted in order of path length from the source node (red node). Table 7(b) shows the result. Dependent nodes need to be updated in a breadth-first order

from the source node as they might be dependent on other data that also need to be updated.

```
SELECT 0 AS path_length, key AS node_id, value
FROM KVS
WHERE key = "D4"

UNION

SELECT CAST(gt.path_length AS INTEGER) AS path_length, gt.id
↪  AS node_id, kvt.value AS value
FROM KVS AS kvt,
GRAPH_TABLE(
    MATCH p = (s)-[]->+(n)
    WHERE s.id="D4" AND n.deleted=FALSE
    COLUMNS (PATH_LENGTH(p) AS path_length, n.id AS id)
) AS gt
WHERE kvt.key = gt.id

ORDER BY path_length;
```

(a) Input Query

path_length	node_id	value
0	D4	de09b97524...
1	D5	09b0537514...
2	D7	8d671b8e64...
3	D9	9de9d25ae1...

(b) Execution Result

Figure 7: Identify Affected Area

The query in Figure 8 retrieves data that do not need to be changed, but are required for recalculations (purple nodes).

```
SELECT gt.id, kvt.value
FROM KVS AS kvt,
GRAPH_TABLE(
    MATCH (m)-[]->(n)
    WHERE (n.id="D4" OR n.id="D5" OR n.id="D7" OR n.id="D9")
    ↪  AND m.deleted=False
    COLUMNS(m.id AS id)
) AS gt
WHERE kvt.key = gt.id;
```

(a) Input Query

gt.id	kvt.value
D3	f7ba1edad9...
D8	7bc5da9249...

(b) Execution Result

Figure 8: Obtain Data for Recalculation

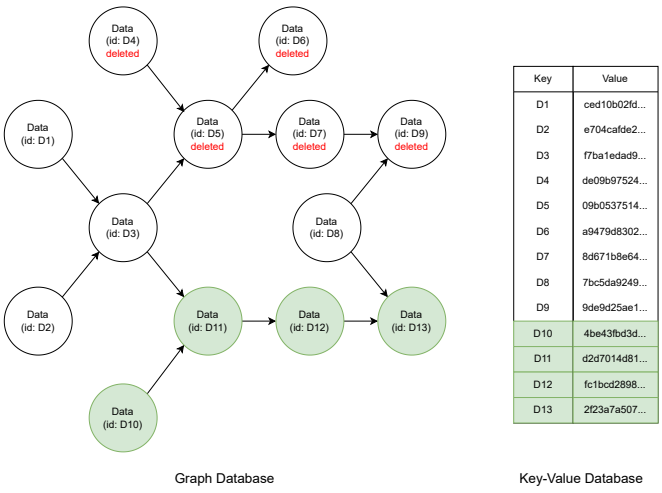
Once recalculations are performed by the application code, affected nodes are soft-deleted and the database is updated with new data (Figure 9).

```
UPDATE GRAPH_TABLE(
    MATCH (s)-[]->+(n)
    WHERE s.id="D4" AND n.deleted=FALSE
    SET s.deleted=TRUE, n.deleted=TRUE
);

UPDATE GRAPH_TABLE(
    MATCH (d3), (d8)
    WHERE d3.id="D3" AND d8.id="D8"
    CREATE
        (d10 IS Data)-[ IS DEPENDENT ]->(d11 IS Data),
        (d11)-[ IS DEPENDENT ]->(d12 IS Data),
        (d12)-[ IS DEPENDENT ]->(d13 IS Data),
        (d3)-[ IS DEPENDENT ]->(d11),
        (d8)-[ IS DEPENDENT ]->(d13)
    SET d10.id="D10", d11.id="D11", d12.id="D12", d13.id="D13"
);

INSERT INTO KVS VALUES
("D10", "4be43fbd3d..."),
("D11", "d2d7014d81..."),
("D12", "fc1bcd2898..."),
("D13", "2f23a7a507...");
```

(a) Input Query



(b) Query Result

Figure 9: Update Database

6 Evaluation

Our middleware application was qualitatively evaluated using an evaluation framework for polystores proposed by Tan et al. [8]. The framework consists of the five dimensions of heterogeneity, autonomy, flexibility, transparency and optimality, with each dimension having two to three attributes. The system is evaluated to have either “little support”, “partial support” or “excellent support” for each attribute. The reader is invited to read the original paper for the details of what each attribute entails.

Heterogeneity

Heterogeneity refers to the number of different data models that are supported by different components of the system.

- **Data-Store Heterogeneity**: Partial Support

Currently, our middleware directly supports the graph and key-value data models. The relational model can be indirectly supported by using the aggregator, which is an SQLite in-memory database. However, future extensions can be added to provide direct support for relational and document (JSON) data models.

- **Processing-Engine Heterogeneity**: Excellent Support

Our middleware uses the processing engines of the component databases directly. Each processing engine is optimised for its supported data model(s).

- **Query-Interface Heterogeneity**: Partial Support

There is no loss in expressiveness for key-value and relational queries. For graph queries, while openCypher is used in graph query fragments, our query language does not allow for complex queries such as multiple `MATCH` statements. Graph objects also cannot be returned as part of a query result.

Autonomy

Autonomy refers to the level of autonomy that each component database has when it is used within a system.

- **Association Autonomy**: Excellent Support

Component databases can be easily added to and removed from the middleware application. As no dependencies are created between data stored in different component databases, there is no loss of data from disassociating a component database from the middleware.

- **Execution Autonomy**: Excellent Support

The middleware uses each component database's native API to execute the respective query fragments. Therefore, external applications can access each component database without interference by the middleware, as long as the component database provides support for concurrent access from multiple applications.

- **Evolution Autonomy**: Excellent Support

As our middleware application only adds a new query interface to interact with component databases and no dependencies are created between databases, schemas in component databases can be freely changed. Obviously, queries will have to modified as well to reflect the changes in schemas.

Transparency

Transparency refers to the level to which a user has to be aware of the location or data model of their data.

- **Location Transparency**: Little Support

Users need to be cognisant of which component database contains the data that they want to access. Queries require explicit references to the type of component database (and its data model). Key-value query fragments cannot be used to access graph data or vice versa.

- **Transformation Transparency**: Little Support

Transformation from the key-value data model or the graph model to the relational model has to be implemented by the user in the the `DBExecutor` classes. No automatic transformation from one data model to another is available.

Flexibility

Flexibility refers to the extent to which users can customise the system to suit their needs.

- **Schema Flexibility**: Partial Support

The middleware inherits the schema flexibility and restrictions of the component databases being used. The key-value data model and graph data model are generally *schema-on-read*, providing high schema flexibility. Type inference is also not performed by the aggregator (in the proof-of-concept implementation, all data are returned as a `String` value), so data type transformations on query results need to be explicitly performed by the user. However, it is possible to enforce constraints in some graph databases such as Neo4j [28]. If relational component databases are supported as a future extension to the middleware, relational query fragments will be *schema-on-write* to fit the rigid schemas enforced by relational databases.

- **Interface Flexibility**: Little Support

For each data model, users are limited to writing queries using the same data model paradigm. The query language is declarative, and does not allow for user-defined functions or other extensibility.

- **Architectural Flexibility**: Partial Support

The modular nature of the mediator-wrapper architecture used in our middleware means that users are free to choose the component databases to use with the middleware. However, other components of the system such as query optimisers and language extensions cannot be easily added.

Optimality

Optimality measures how well-optimised the query plan and data placement are in the system.

- **Federated Plan Optimisation**: Little Support

Although individual query fragments are executed by the component databases, no optimisation of the query plan is performed. Transformations or migrations of data from one data model to another in order to achieve better performance is also not supported.

- **Data-Placement Optimisation**: Little Support

It is entirely the responsibility of the user to decide which data model is more suitable for their data and to decide how to model their data, and this is expressed explicitly in queries. No decision-making about what data model to use is performed by the middleware application.

7 Conclusion

In this paper, we implemented a middleware application that provides a common query interface for the graph and key-value data models. Our query language is based on SQL/PGQ and Cypher, striking a balance between the familiarity of existing query languages and the expressiveness of both the relational and graph query language paradigms. Our query language also supports hybrid operations, allowing the user to use both graph and key-value data simultaneously. The modular architecture of our application allows users to choose the component databases they want to use our query language with. As a result, although our application provides a high level of autonomy to the component databases, it does not do as well in terms of the other desirable attributes of polystore systems, especially in the dimensions of transparency and optimality.

Future work includes adding support for using relational databases as component databases. We also want to look into techniques for query optimisation, as well as evaluate the scalability of our application. Finally, we would like to update our query language and graph query fragment generation to support GQL, which is a work-in-progress ISO standard for a graph database query language, once it is released. This would allow our middleware to support a greater number of graph databases, as currently only graph databases that support Cypher can be used with our middleware.

Acknowledgements

This paper was supported by JST, CREST Grant Number JPMJCR22M2, Japan.

References

[1] *Information technology – Database languages SQL – Part 16: Property Graph Queries (SQL/PGQ).* Standard ISO/IEC 9075-16:2023. International Organisation for Standardisation, June 2023. https://www.iso.org/standard/79473.html.

[2] Wen Sun et al. "SQLGraph: An Efficient Relational-Based Property Graph Store". In: SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 1887–1901. ISBN: 9781450327589. https://doi.org/10.1145/2723372.2723732.

[3] *SQL-Gremlin.* GitHub Repository. GitHub. https://github.com/twilmes/sqlgremlin (visited on January 29, 2024).

[4] *Cypher to SQL Translation (Cyp2SQL) version 1.2.* GitHub Repository. GitHub. https://github.com/DTG-FRESCO/cyp2sql (visited on January 29, 2024).

[5] Benjamin A. Steer et al. "Cytosm: Declarative Property Graph Queries Without Data Migration". *Proceedings of the Fifth International Workshop on Graph Data-Management Experiences & Systems.* GRADES'17. Chicago, IL, USA: Association for Computing Machinery, 2017. ISBN: 9781450350389. https://doi.org/10.1145/3078447.3078451.

[6] ArangoDB. https://arcadedb.com/.

[7] OrientDB. https://orientdb.com/.

[8] Ran Tan et al. "Enabling query processing across heterogeneous data models: A survey". *2017 IEEE International Conference on Big Data (Big Data).* 2017, pp. 3211–3220. https://doi.org/10.1109/BigData.2017.8258302.

[9] Vijay Gadepally et al. "The BigDAWG polystore system and architecture". *2016 IEEE High Performance Extreme Computing Conference (HPEC).* 2016, pp. 1–6. https://doi.org/10.1109/HPEC.2016.7761636.

[10] Boyan Kolev et al. "The CloudMdsQL Multistore System". *Proceedings of the 2016 International Conference on Management of Data.* SIGMOD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 2113–2116. ISBN: 9781450335317. https://doi.org/10.1145/2882903.2899400.

[11] Jingjing Wang et al. "The Myria Big Data Management and Analytics System and Cloud Services". *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings.* www.cidrdb.org, 2017. http://cidrdb.org/cidr2017/papers/p37-wang-cidr17.pdf.

[12] *Apache Drill.* https://drill.apache.org/ (visited on January 29, 2024).

[13] *Information technology – Database languages – GQL.* Standard ISO/IEC FDIS 39075. Under Development. International Organisation for Standardisation. https://www.iso.org/standard/76120.html.

[14] *Existing Languages.* GQL Standard. https://www.gqlstandards.org/existing-languages (visited on February 13, 2024).

[15] *Information technology – Database languages SQL – Part 14: XML-Related Specifications (SQL/XML).* Standard ISO/IEC 9075-14:2023. International Organisation for Standardisation, June 2023. https://www.iso.org/standard/76587.html.

[16] *Information technology – Guidance for the use of database language SQL – Part 6: Support for JSON.* Standard ISO/IEC 19075-6:2021. International Organisation for Standardisation, August 2021. https://www.iso.org/standard/78937.html.

[17] Carlyna Bondiombouy and Patrick Valduriez. "Query processing in multistore systems: an overview". *IJCC - International Journal of Cloud Computing* 5.4 (2016), pp. 309–346. https://doi.org/10.1504/IJCC.2016.080903.

[18] SQLite. https://www.sqlite.org/.

[19] ANTLR. https://www.antlr.org/.

[20] PostgreSQL grammar. https://github.com/antlr/grammars-v4/tree/master/sql/postgresql.

[21] *Appendix D. SQL Conformance.* PostgreSQL. https://www.postgresql.org/docs/current/features.html (visited on January 23, 2024).

[22] Neo4j. https://neo4j.com/.

[23] LevelDB. https://github.com/google/leveldb.

[24] Memgraph. https://memgraph.com/.

[25] Riak. https://riak.com/products/riak-kv/index.html.

[26] ArcadeDB. https://arcadedb.com/.

[27] Kyoko Yasuda. "A Study on the Modelling and Management of Provenance Data". Japanese. Master's Thesis. Tokyo Institute of Technology, January 2024.

[28] *Constraints.* Neo4j. https://neo4j.com/docs/cypher-manual/5/constraints/ (visited on January 26, 2024).

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

データリネージュ・統合データ管理

座長:北島 信哉(富士通株式会社)
コメンテータ:中野 美由紀(津田塾大学)
2024年3月1日(金) 13:30 〜 15:40 T2-A (オンライン（Zoom Events))

15:05 〜 15:30

[T2-A-8-05]【技術報告】 Universal HTAP: 多様なデータベースを統合的に管理するデータアーキテクチャ

山田浩之 (株式会社Scalar)

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年3月1日(金) 13:30 ~ 15:40 | T3-A オンライン（Zoom Events)

大規模言語モデル (1)

座長:若林 啓(筑波大学)
コメンテータ:田島 敬史(京都大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T3-A-8-01] 大規模言語モデルを用いた検索要求の観点抽出に基づく検索結果の欠損の可視化
*川本 唯人[1]、笹本 まこと[1]、北山 大輔[1] (1. 工学院大学)

13:50 ~ 14:15
[T3-A-8-02] 大規模言語モデルを用いたカテゴリ説明文付与によるニュース推薦の性能向上
*矢田 宙生[1]、山名 早人[1] (1. 早稲田大学)

14:15 ~ 14:40
[T3-A-8-03] LLMベース推薦システムにおけるユーザのネガティブ嗜好の活用
*徐 雲帆[1]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府、2. 九州大学大学院芸術工学研究院)

14:40 ~ 15:05
[T3-A-8-04] 大規模言語モデルに基づく検索手法の日本語文書検索への適用
*阿部 健也[1]、薄羽 皐太[1]、加藤 誠[2] (1. 筑波大学 大学院人間総合科学学術院 知識獲得システム研究室、2. 筑波大学 図書館情報メディア系)

15:05 ~ 15:30
[T3-A-8-05] 文書への追加情報に対するニューラル検索モデルの挙動分析
*薄羽 皐太[1]、加藤 誠[1]、藤田 澄男[2] (1. 筑波大学、2. LINEヤフー株式会社)

大規模言語モデルを用いた検索要求の観点抽出に基づく検索結果の欠損の可視化

川本　唯人†　　笹本まこと†　　北山　大輔†

† 工学院大学情報学部　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j321098,jx20114}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　気になることを調べる際に Web 検索はよく用いられる．一般的な検索エンジンでは，検索要求を入力するとそれに関連する結果を提示する．この時，検索者は検索要求に必要な観点が検索結果中で網羅されているかを把握することは難しい．そこで本研究では，大規模言語モデルを用いて検索要求の観点を抽出し，その観点に関して検索結果中での出現の有無を可視化することで，検索結果中の観点の欠損を抽出する手法を提案する．本稿ではプロトタイプシステムを構築し，その動作について報告する．観点の欠損については提示することができたが，想定される観点が対応付いていない場合があり，改善が必要であることを確認した．

キーワード　意図理解，クエリ補完，バイアス，検索支援，観点抽出

1　はじめに

現在，気になることを調べる際に Web 検索は広く使われている．一般に，検索者は詳細には知らない内容について検索をしていることが多く，検索結果を閲覧しながら情報を獲得し，所望の情報に近づいていく．そのため，不慣れな分野について検索する際に，検索要求に対し適切でない検索キーワードを入力したり，適切ではないページを閲覧すると，所望の情報を得るまでに時間を要することがある．特に，検索者に思い込みなどがある場合や，検索結果中に所望の情報に至る手掛かりが含まれていない場合，検索要求を満たすことは困難となる．

このような，検索要求を満たしやすくするための支援として，網羅的な情報検索の支援が研究されている．梅本ら [1] は，網羅的な情報検索のために未閲覧情報の種類と量を可視化して提示する手法を提案している．この手法では，未閲覧情報は検索結果の中から求めており，検索検索結果に含まれていないトピックをカバーできていない．若宮ら [2] は，話題に対する感情をもとに検索の網羅性を高める手法を提案している．この手法では，感情という列挙しやすい観点に基づいた支援となっているが，検索要求に依存したトピックの網羅性を高めるような仕組みがあるとよいと考える．

このように網羅的閲覧に関する従来研究では，検索結果に含まれる観点であったり，トピックに依存せずに列挙できる観点に基づいて検索要求を満たすために欠けている欠損情報を示すことが多い．しかしながら，ユーザが自力で検索することが困難な欠損を示せることが望ましいと考えられる．そこで我々は大規模言語モデルを用いて，検索要求に対する観点を自動生成することで，トピックに必要な検索結果に含まれないような観点を抽出できる可能性があると考えた．

そこで本研究では，ユーザの検索要求に対して，大規模言語モデルを用いて要求の達成に必要な観点とその説明を生成し，検索結果に含まれる観点と欠損している観点を判定し可視化する手法を提案する．具体的には，生成した観点文と検索結果との文書類似度で対応関係を判定することで，対応が見られない観点を欠損と判断する．このような検索結果に欠損する観点が示されることで，ユーザは効果的な次の検索キーワードを考えることが可能となる．

本論文の貢献は以下のとおりである．

- 大規模言語モデルを用いることで検索要求に対して適切な観点を生成できる例について議論する
- 検索結果に含まれない観点を可視化するユーザインタフェースについて議論する

2　提案手法

2.1　大規模言語モデルを用いた検索要求の観点抽出

本研究では，比較検討をしたいサービスや商品に対する検索要求を対象とする．例えば,「初めて美容院に行きます．選び方を教えてください.」のような自然言語による問い合わせや「冷蔵庫」のような対象のみの問い合わせを想定する．このような検索要求に対して，大規模言語モデルを用いて観点の抽出を行う．大規模言語モデルとしては，OpenAI が提供する GPT-3.5-turbo-1106[1] を用いる．

このモデルでは，入力として，system プロンプトと user プロンプトが存在する．system プロンプトは，モデルの出力の方針を決定するものであり，本稿では「入力された内容に関する情報を Web 検索する時のアドバイスを 5 個教えてください。その際、それぞれに 1 単語の見出しが欲しいです。(例 ○○: ～)」を用いる．個数に関しては任意の数値を設定可能であるが，少なすぎると観点が網羅できず，多すぎると余分な観点が抽出されることが予想できるため，適切な数にする必要がある．user プロンプトは，出力の直接のトリガとなるプロンプトであり，本稿では検索要求をそのまま用いる．

1：https://platform.openai.com/docs/api-reference/chat

表 1　美容院の観点

評判	口コミやレビューをチェックして評判の良い美容院を選ぶ
スタイリスト	経験豊富なスタイリストが在籍している美容院を選ぶ
施術内容	自分の希望する施術内容を提供している美容院を選ぶ
清潔さ	清潔で衛生的な美容院を選ぶ
価格	予算に合った価格設定の美容院を選ぶ

出力された文を，見出しを含めてそれぞれ観点として扱い，見出し部分を可視化のためのラベルとして扱う．今回用いるプロンプトでは 5 つの観点が抽出されることとなる．

2.2　抽出観点を用いた検索結果の欠損の可視化

抽出した検索要求の観点を用いて，検索結果の欠損を可視化する．ここでは，検索結果は検索要求をそのまま検索キーワードとして入力した際の Web 検索結果として説明する．各検索結果は，タイトルとスニペットを結合し，それぞれ 1 文として取り扱う．

生成した観点 V の各観点 v とそれぞれの検索結果 r について文書類似度 sim を求め，類似度が閾値 α を超えたものについて，該当の観点を含む検索結果とする．文書類似度については，fasttext を用いて単語分散表現に基づいて類似度を計算する．一つの検索結果は，複数の観点に対応することもある．ある検索結果 r に対して対応する観点を式 1 で定義する．

$$rel_viewpoint(r) = \{v; v \in V, sim(v, r) >= \alpha\} \qquad (1)$$

このようにして判定した際に，どの検索結果とも対応づかない観点を検索キーワード入力欄の下に，欠損観点としてユーザに提示する．対応があった観点については，欠損観点と同様に結果に含まれる観点として列挙する，その上で，検索結果のタイトル横に観点ラベルを付与する．

3　実行例と考察

3.1　観点抽出の例

提案手法を用いた実行結果の例を示す．まず,「初めて美容院に行きます．選び方を教えてください.」という自然言語による入力を行なった場合に，生成された観点を表 1 に示す．また,「冷蔵庫」のように商品名のみを入力とした場合に生成された観点を表 2 に示す．この 2 つの例では，生成された観点は，それぞれの目的を達成するために必要な検索の観点であると考えられる．また，それぞれ独立した観点であると考えられ，冗長にならない生成が行えている．観点の生成数を増やした場合の挙動については今後の課題とする．

3.2　検索結果の欠損の可視化の例

「美容院」を用いた結果を表 3 示す．類似度の閾値 α については，0.78 を用いた．検索結果は 10 件取得し，表中には検索結果のタイトルのみを示しているが，判定には各結果のスニペットも含めている．この結果では，10 件の検索結果の内 7 件はそれぞれ「評判」「施術内容」「スタイリスト」の内のどれかと対応づいた．それぞれのタイトル・スニペットを確認したところ，この観点の内容が含まれていたため，対応関係としては妥当であると考える．また，対応しなかった「価格」「清潔さ」に関しても，各ページのスニペットからはその内容が含まれるか判断はできず，対応づかないことが妥当といえる．しかし,「良い美容室の選び方 5 選！悪いサロンの見分け方とは？現役 ...」というページのスニペットには,「美容師」という単語が使われており，本来であれば「スタイリスト」と対応づくという結果が妥当であると考えるが，今回は対応づかないという結果になった．この結果を見て，システムに改善すべき点があると考える．

表 2　冷蔵庫の観点

整理	冷蔵庫内を整理して収納スペースを最大限に活用しましょう。期限切れの食品や使わなくなった食材を捨ててスペースを確保しましょう。
温度設定	冷蔵庫の温度設定を適切に行いましょう。食材に合った温度で保管することで鮮度を保ちます。
清掃	定期的に冷蔵庫内を清掃しましょう。食品のこぼれなどを取り除き、衛生的な状態を保ちます。
食品の配置	食品の配置に工夫を凝らしましょう。同じカテゴリーの食品をまとめて収納することで見やすく、取り出しやすくなります。
匂い対策	冷蔵庫内の匂いを防ぐために、消臭剤やコーヒーの粉などを使って匂いを吸収する工夫をしましょう。

表 3　美容院の検索結果に含まれる観点

タイトル	観点
美容院の上手な探し方！美容師が教える 1 番失敗しない美容室 ...	スタイリスト
【美容室の選び方】美容師目線でチェックするポイントを教え ...	評判，スタイリスト
良い美容室の選び方 5 選！悪いサロンの見分け方とは？現役 ...	
【後悔しない！】よい美容院の選び方 4 つのコツと見つける ...	
初めての美容院で失敗しない選び方&行きつけを見つける方法	施術内容, スタイリスト
美容室の選び方について教えてください。ホットペッパー	評判，スタイリスト
どんな美容室に通いたい？ 美容室選びで重視することは…3 位 ...	
【美容院の選び方】ヘアライター直伝。失敗しない美容院選び ...	施術内容，スタイリスト
初めて美容院に行くメンズも安心！絶対に失敗しない	スタイリスト
現役美容室経営者が間違えない【美容室の選び方】を教えます！	

4 ま と め

我々は，大規模言語モデルを用いて検索要求の達成に必要な観点とその説明を生成し，検索結果に含まれる観点と欠損している観点を判定し可視化する手法を提案した．本稿では，その実行例を示し，観点の判定が意図した結果になっていない部分があり，改善が必要であることが確認された．

今後の課題としては，大規模言語モデルにより生成する観点の妥当性と網羅性に関して定量的に評価する必要がある．これは，検索要求に対して正解となる観点を人手で用意し，適合率，再現率，F 値などの指標で評価することを予定している．また，観点と検索結果の対応の妥当性についての評価や観点の個数指定を自動制御する手法の検討，並びに Web 検索の観点として「アドバイスの見出し」が適切かどうか検証などに取り組む必要がある．他にも，実ユーザを用いて提案手法により検索行動に改善がみられることを評価する必要があると考えている．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 梅本和俊, 山本岳洋, 田中克己. 網羅性指向タスクにおける未閲覧情報量の提示. 人工知能学会論文誌, Vol. 32, No. 1, pp. WII–G_1–12, 2017.

[2] 若宮翔子, 河合由起子, 熊本忠彦, 張建偉, 白石優旗. 話題に対する多様な感情に基づく web ページ検索システム. 情報処理学会論文誌, Vol. 57, No. 1, pp. 366–378, jan 2016.

大規模言語モデルを用いたカテゴリ説明文付与による ニュース推薦の性能向上

矢田　宙生† 　山名　早人†

† 早稲田大学院基幹理工学研究科　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †{yadayuki,yamana}@yama.info.waseda.ac.jp

あらまし　オンラインニュースプラットフォームにおいて，ユーザが大量のニュース記事の中から関心のある記事を効率的に発見するためには，パーソナライズされたニュース推薦が重要であり，活発に研究がされている．近年の研究では，ニュースのコンテンツ情報およびユーザの嗜好情報を深層ニューラルネットで同じ次元のベクトルに変換し，推薦候補となるニュースをユーザのベクトルとの類似度に基づいて並び替える手法が高い性能を示している．ニュースのコンテンツ情報を理解する上で重要な情報源の一つとして,「*finance-real-estate*」や「*newspolitics*」といったニュースカテゴリが挙げられる．しかし，ニュースのカテゴリ名単体は，極めて簡潔であり，しばしば情報量に乏しいという課題がある．本研究では，この課題を解決するために，カテゴリの説明文を大規模言語モデルで生成し，推薦モデルの推論の際，ニュースに追加情報として付与することで，推薦性能の向上を試みる．検証には，ニュース推薦において高い性能を示す NRMS-BERT を使用し，MIND データセットを用いて訓練および評価を行った．比較対象となるベースラインは次の 2 つである: 1) タイトル文のみを推薦モデルへの入力とする手法，2) タイトル文に加え，事前に定義したテンプレート (「*The news category is* {*category*}」) にカテゴリ名を埋め込んだ文章を追加情報として付与する手法．実験結果から，提案手法は，ベースライン手法と比較し，検証した全ての指標に対して性能が向上することを確認した．

キーワード　ニュース推薦，大規模言語モデル，GPT-4，推薦システム，深層学習

1 はじめに

近年，Microsoft News[1]，Google News[2]や日経電子版 [3]などのオンラインニュースプラットフォームが広く普及し，インターネットからニュースを読む行為は人々の日常生活の一部となっている．これらのプラットフォームは日々膨大な量のニュース記事を配信しており，ユーザが自身の興味に合ったニュース記事を見つけることはますます困難になっている．以上の背景から，パーソナライズされたニュース推薦への需要が高まっている．

ニュース推薦に関する研究は活発に行われており，近年では深層学習を用いた手法が高い性能を示していることが知られている [1]．深層学習を用いたニュース推薦の概要図を図 1 に示す．図に示す通り，深層学習ベースのニュース推薦モデルは，一般的に以下の 3 つの主要なモジュールから構成される:

- *News Encoder*: ニュースタイトルやカテゴリなどの記事情報からニュースベクトル h を生成する
- *User Encoder*: ユーザの過去の閲覧履歴からユーザの嗜好を反映したユーザベクトル u を生成する
- *Similarity Calculator*: ニュースベクトル h とユーザベクトル u の類似度を計算し，高い類似度を持つニュースほどユーザの好みに合致すると推定する

先行研究 [2] [3] [4] では，News Encoder および User Encoder がそれぞれ異なるニューラルネットを使用して構築されている．例えば，ユーザの興味やニュースコンテンツの表現の獲得に Multihead Attention を活用する NRMS [2]，LSTM を活用してユーザの興味を獲得する LSTUR [3]，CNN や Attention 機構によりユーザの興味やニュースコンテンツの表現を獲得する NPA [4] などが挙げられる．このような中，先行研究により，BERT [5] のような事前学習済み言語モデルを News Encoder に活用し，ニュースコンテンツの表現を獲得するニュース推薦の手法が，特に高い性能を出すことが知られている [6] [7] [8] [1].

ニュースコンテンツを理解する上で重要な情報源の一つとして， ニュースカテゴリが挙げられる．例えば，ニュース推薦分野で頻繁に用いられるデータセットである MIND [9] には「*tv-golden-globes*」のようなカテゴリが存在する．推薦モデルへのニュースカテゴリ入力の手法としては，事前に定義されたテンプレート（「*The news category is* {*category*}」など）にカテゴリ名を当てはめ， 他のニュースコンテンツ情報と結合して用いる手法がある [1]．しかし，このテンプレートは全てのニュースカテゴリに適用可能であるほど一般的であるため，情報量が不十分であるという課題がある．各ニュースカテゴリに対して詳細な説明文を構築し，それを推薦モデルに入力することは，推薦モデルがニュースカテゴリをより正確に認識できるようになることが期待される．

しかし，説明文を手動で作成することはコストが高く，新た

1：https://news.microsoft.com
2：https://news.google.com
3：https://nikkei.com

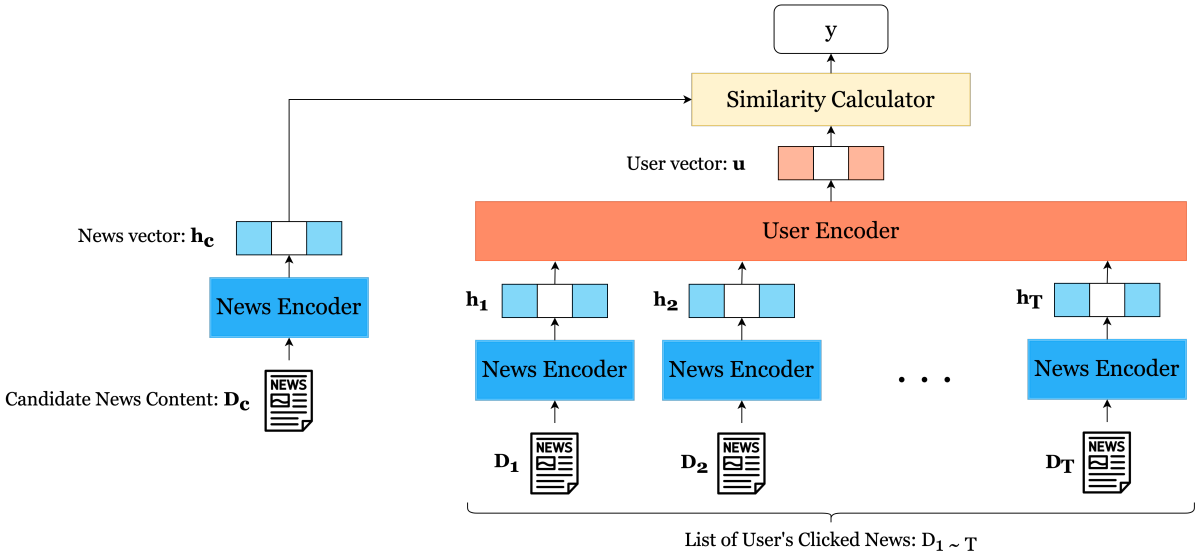


図 1　深層学習を用いたニュース推薦の概要

に定義されるニュースカテゴリに対する拡張性も低い．この課題に対し，本研究では大規模言語モデルを用いてニュースカテゴリの説明文を自動生成することで解決する．GPT-4 [10] のような大規模言語モデルは豊富なテキストデータで事前学習され，多様なトピックに関する広範な知識を有している．したがって，大規模言語モデルが生成する「*finance-real-estate*」や「*newspolitics*」といったニュースカテゴリ名に対する詳細な説明文は，推薦モデルがニュースカテゴリに関する深い理解を得る上で有用であると期待される．本研究は，大規模言語モデルを活用してニュースカテゴリの説明文を生成し，ニュース推薦モデルに追加情報として入力する．具体的には，ニュース記事を図 1 に示す News Encoder に入力する際に，1）ニュースのタイトル文，および 2）大規模言語モデルにより生成されたニュースカテゴリの説明文，を結合した文章を使用する．提案手法は，ニュース推薦の分野で高性能を誇る NRMS-BERT [2] [6] を Microsoft News Dataset（MIND）[9] でファインチューニングすることにより検証した．

本論文の構成は以下の通りである．2 節では本研究における関連研究について説明する．3 節で，提案手法の詳細について説明する．4 節で提案手法の有効性を評価実験の設定について述べる．5 節で評価実験の結果を説明する．最後に，6 節で結論を示す．

2　関 連 研 究

本節では，ニュース推薦および大規模言語モデルにより生成されたテキストの他タスクへの応用に関する関連研究について述べる．

2.1　ニュース推薦

2.1.1　深層学習によるニュース推薦

先行研究により，複数のニュース推薦手法が提案されており，深層学習モデルを活用するアプローチが特に高い性能を示していることが明らかになっている．

2019 年，Wu ら [2] は，ニュースベクトルとユーザベクトルの獲得に Attention 機構を用いる NRMS という手法を提案した．この手法では，BERT [5] や LLaMA [11] といった言語モデルや，GloVe [12] のような単語分散表現を用いて得られる単語ベクトルの配列に Multihead Attention を適用し，ニュースベクトルを生成する．また，ユーザが過去に閲覧したニュースのニュースベクトルも同様に Multihead Attention に入力し，ユーザベクトルを得る．当該研究では，GloVe に基づく単語ベクトルを使用して NRMS の推薦性能を検証し，2019 年時点で最高の結果得た．その後の研究において，NRMS は広く利用され，事前学習済み言語モデルから得られた単語ベクトルを NRMS に適用する手法は現在でも高い性能を出すことで知られている [13] [6].

2023 年，Wang ら [14] は，クリックログに含まれるノイズ（例えば，誤ってクリックされた場合など）の影響を軽減する目的で，対照学習に基づくユーザの嗜好ベクトル獲得手法 MCCM を提案した．MIND データセットを用いたオフライン検証を行い，既存の手法と比較して優れた性能を得た．

2.1.2　事前学習済み言語モデルを活用したニュース推薦

深層学習を用いたニュース推薦の中でも，事前学習済みの言語モデルを活用してニュースベクトルを獲得する手法は特に高い性能を上げている．

2021 年，Wu ら [6] は，News Encoder に BERT [5] や RoBERTa [15] などの事前学習済み言語モデルを用いたニュース推薦モデルの性能を，オフラインおよびオンラインの両方の実験において網羅的に検証した．オフライン実験では，Microsoft News の過去のユーザログとニュース記事データから構築された Microsoft News Dataset (MIND) [9] が使用された．実験結果から，News Encoder に事前学習済みモデルを適用することで性能が向上することが確認された．さらに，Wu らは，Microsoft News 上での推薦に事前学習済み言語モデルを用い

たニュース推薦モデルを適用し，オンライン検証も行った．その結果，事前学習済み言語モデルを使用しない既存の推薦モデルと比較して，8.53 %のクリック率向上が得られた．

2022 年，Shivaram ら [16] は，政治的立場 (リベラルか，保守か) の偏りという観点でのフィルターバブル解消を目的としたニュース推薦手法の提案を行った．ニュース記事内の政治的立場を決定づけるような単語の影響が小さくなるように目的関数を設計した STANPP, ニュースのトピック固有の単語の影響が大きくなるようなマルチタスク学習 MTAN, 二つを組み合わせた MTANPP を提案した．彼らは，政治的立場がラベル付けされたニュース記事データセットを用いて検証を行い，提案手法 MTANPP はベースラインと比較し最も高い性能を得た．

図 1 に示される構造と異なるアプローチを用いてニュース推薦を実現する手法も存在する [8] [17]．2021 年には，Zhang ら [8] が UNBERT を提案した．この手法では，ユーザが過去に閲覧したニュース記事のリストと推薦候補のニュース記事を結合し，単一の事前学習済み言語モデルに入力し，その出力を直接利用して，クリック率の予測を行う．2023 年には，Zhang ら [17] が，BERT などの事前学習済み言語モデルの言語知識を活用する Prompt4NR を提案した．この手法では，ニュース推薦をマスク穴埋め問題（Masked Language Model）として捉える．具体的には,「<CANDIDATE> is [MASK] to <USER>」というテンプレートを用意し，推薦候補となるニュース記事を「<CANDIDATE>」，ユーザの過去の閲覧ニュースを「<USER>」と位置づけ,「[MASK]」に入る単語を予測することでニュース推薦を行う．

2023 年，Liu ら [13] は，ONCE という大規模言語モデルを活用したニュース推薦フレームワークを提案した．このフレームワークでは，News Encoder に BERT-base と比較して数十倍のパラメータを持つ大規模言語モデルである LLaMA [11] を使用している．一般的に，言語モデルはパラメータ数が多いほど高い性能を示すとされている．ONCE は，MIND を用いた実験において，BERT-base のニュース推薦手法を上回る性能を報告している．

2.1.3 マルチモーダルニュース推薦

テキストに加えて，画像などを含むマルチモーダルな要素を入力とするニュース推薦に関する研究も行われている．

2022 年には，Wu ら [7] がニュースタイトル文だけでなく，サムネイル画像も特徴として取り入れたマルチモーダルニュース推薦手法 MM-Rec を提案した．この手法では，まず学習済みの mask R-CNN [18] を用いてサムネイル画像からの region-of-interests（ROIs）の抽出を行い，その後，これらの ROIs とニュースタイトルを VilBERT [19] に入力してニュースベクトルを生成する．実際のサービスから得られたデータセットにおいて行われたオフライン検証では，テキスト情報のみを使用した手法 [6] と比較して，MM-Rec はより高い性能を示した．

2.2 大規模言語モデルにより生成されたテキストの他タスクへの応用

先行研究により，汎用的な知識を事前学習により獲得した大規模言語モデルから生成されたテキストは，様々なタスクの性能向上に寄与することが知られている．

2021 年，Yoo ら [20] は，GPT-3 [21] を用いたテキスト分類のためのデータ拡張手法 GPT3Mix を提案した．この手法では，テキスト分類モデルの学習に用いる訓練データから複数のテキストをサンプリングし，それらを大規模言語モデル GPT-3 に入力して訓練データを拡張する．拡張されたデータを用いて BERT によるテキスト分類器を学習した結果，未拡張のデータを用いて学習されたモデルと比較して高い性能が得られたことが報告されている．

2022 年，Liu ら [22] は質問応答タスクにおいて，大規模言語モデルによる知識拡張を応用した．彼らの提案手法では，まず GPT-3 を使用して，質問に関連する知識を事前に生成する．次に，大規模言語モデルによって生成された知識と質問文を結合した文章を質問応答モデルへの入力として使用し，回答を生成する．実験の結果，質問文のみを使用するベースライン手法と比較して，このアプローチは高い性能を示した．

2023 年，Pratt ら [23] は，CLIP [24] を用いたゼロショット画像分類において，大規模言語モデルにより生成された文章を応用する Customized Prompts via Language models（CuPL）という手法を提案した．CLIP では通常,「*a photo of a* {}」のような事前定義されたテンプレートにカテゴリ名を挿入したキャプションのテキストと，画像データを同次元のベクトルに変換し，その類似度に基づいて画像を分類する．CuPL では，カテゴリごとの画像キャプションを GPT-3 を使用して自動生成し，ゼロショット分類に適用した．実験の結果，CLIP によるゼロショット分類において，通常のテンプレートを用いるベースライン手法と比較し，CuPL が高い性能を示した．

2.3 関連研究のまとめ

関連研究から，ニュース推薦は，深層学習を用いた手法，特に News Encoder に事前学習済み言語モデルを活用した手法が優れた性能を示している．様々な構造のモデルが提案されているが，ニュースカテゴリ名の情報不足という課題は未だ解決されていない．

また，大規模言語モデルにより生成されたテキストの活用は，画像やテキストの分類，質問応答など多様なタスクにおいて有効であることが示されている．

以上の背景を踏まえ，本研究では，ニュースカテゴリ名の詳細な説明文を大規模言語モデルで生成し，これをニュース推薦モデルに組み込むことで，推薦性能のさらなる向上を目指す．本アプローチは，ニュースカテゴリの深い理解を推薦システムに与え，より適切なニュースの推薦に寄与すると考えられる．

3 提案手法

提案手法では大規模言語モデルを用いて自動生成したニュースカテゴリの説明文を，ニュース推薦モデルの推論時に追加情報と付与する．提案手法の概要を図 2 に示す．提案するニュース推薦手法は次の 2 つの手順に分かれる: 1) 大規模言語モデル

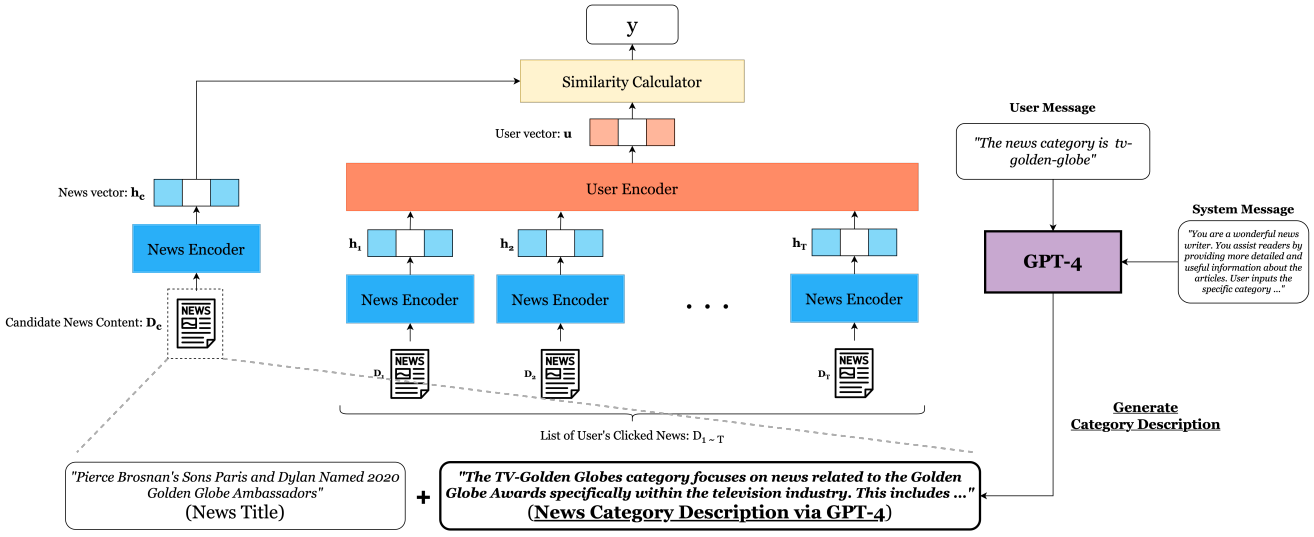


図 2　提案手法の概要

を用いてニュースカテゴリ説明文の自動生成（3.1 項），2) 自動生成したカテゴリ説明文を推薦モデルへと入力する（3.2 項）．

3.1　ニュースカテゴリ説明文の生成

本手順の目標は，人間による作業なく，ニュースカテゴリに応じた多様で詳細な説明文を自動生成することである．

本研究では，カテゴリ説明文を自動生成するための大規模言語モデルとして，OpenAI の GPT-4 [10] を用いた．GPT-4 は，様々な領域のタスクにおいて高い性能を出すことが知られている [25]．

カテゴリ説明文の生成の際に大規模言語モデルに入力したプロンプトについて述べる．本手法では，「*finance-real-estate*」や「*newspolitics*」といった単一の単語で表されるニュースカテゴリ名から，そのカテゴリに関する詳細な説明文を生成する必要がある．ニュースカテゴリ説明文を生成するための入力プロンプトを図 3 に示す．システムメッセージにおいて，言語モデル自身が担うべき役割やカテゴリ説明文の生成タスクに関する記述を行った．具体的には「あなたは素晴らしい報道官であること」(役割),「入力されるニュースカテゴリの詳細な説明文を出力すること」(タスクの概要) や「出力文字数の目安」(タスクの制約) 等の内容を入力している．一方でユーザメッセージは「*The news category is* {*category*}」というテンプレートを事前に定義し，テンプレートにカテゴリ名を当てはめた文章 ({*category*} の箇所にニュースカテゴリ名が入る) を入力するのみにとどめている．

本研究では，検証のためのデータセットとして MIND [9] を用いている．MIND 内のすべてのニュースカテゴリに対し，大規模言語モデルを用いた説明文の自動生成を行なった．生成したニュースカテゴリの説明文の具体例を図 4 に示す．図に示す例は「*tv-golden-globes*」というカテゴリに対して，説明文の生成を行なっている．さらに，表 1 に MIND 内の全てのカテゴリに対し，生成したニュースカテゴリの説明文の統計情報を示す．

3.2　カテゴリ説明文の推薦モデルへの入力

大規模言語モデルを用いて，すべてのニュースカテゴリに対し説明文の自動生成を行なった後，生成した説明文を活用してニュース推薦モデルの学習を行った．先行研究で，様々なニュース推薦モデルが提案されている中，特に高い性能を示すことが報告されている NRMS-BERT [2] [6] を検証に用いた．

システムメッセージ

*You are a wonderful news writer. You assist readers by providing more detailed and useful information about the articles. User inputs the specific category for a news article using the format 'The news category is **{category}**'. Please provide a detailed explanation, in about 50 words, **in English**, on the category of the articles entered (such as politics, economics, sports, etc.). Please avoid using symbols such as double quotes (") and single quotes ('), asterisks (*), and similar for emphasis as much as possible.*

ユーザメッセージ

The news category is tv-golden-globes

図 3　カテゴリ説明の生成に使用するプロンプト．本図の例では，*tv-golden-globes* というニュースカテゴリを生成対象とする．

出力例

The TV-Golden Globes category focuses on news related to the Golden Globe Awards specifically within the television industry. This includes nominations, winners, notable moments from the ceremony, reactions, and any controversies or highlights surrounding the event, which honors excellence in TV.

図 4　GPT-4 により生成されたニュースカテゴリ説明文の具体例．(*tv-golden-globes*)

表 1　MIND に対して生成したニュースカテゴリ説明文の統計

項目	値
カテゴリ数	270
説明文の平均単語数	57.7
説明文の最小単語数	48
説明文の最大単語数	74

生成したカテゴリ説明文の推薦モデルへの入力方法について述べる．ここで，ニュースタイトル文を D_{title}，生成したカテゴリの説明文を D_{desc} とする．推薦モデルの推論時，D_{title} と D_{desc} を BERT の特殊トークンである「*SEP*」トークンで結合したテキストを News Encoder へと入力した．(図 5)

4　評 価 実 験

本節では，3 節で述べた提案手法の有効性を検証するための評価実験について説明する．具体的には，検証に用いるデータセット，ニュースカテゴリ説明文の生成に関する実装の詳細，評価指標，ハイパーパラメータ，およびベースラインとした手法について述べる．

4.1　データセット

本研究ではニュース推薦分野で広く用いられている Microsoft News Dataset(MIND) [9] により実験を行った．MIND は，2019 年 10 月 12 日 ～ 11 月 22 日の間に実サービスである Microsoft News[4] 上で収集されたデータにより構築されている．MIND はサービス上のニュース記事とユーザの閲覧履歴データの 2 種類のデータで構成されている．表 3 および表 4 に格納されている情報の詳細を示す．

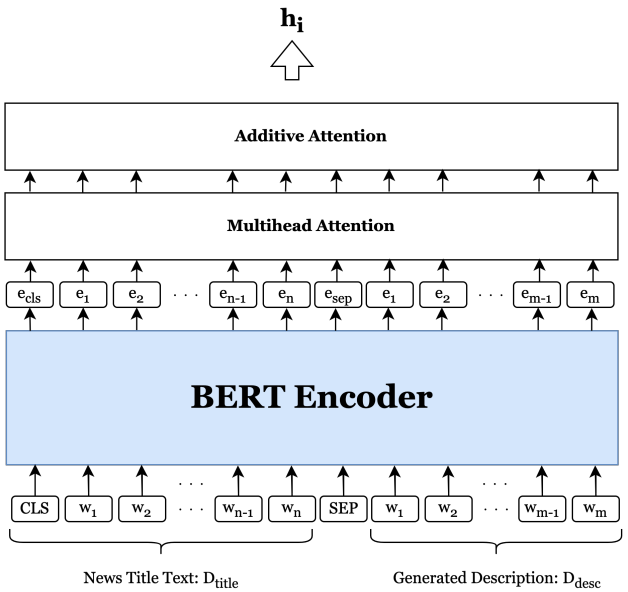


図 5　News Encoder への入力形式．D_{title} と D_{desc} を *SEP* トークンで結合する

4：https://news.microsoft.com

MIND には，元の MIND から，一部ユーザのみをランダムにサンプリングし，データ量を抑えた MIND-small という別バージョンが公開されている．本研究では，MIND-small を用いた．MIND-small の統計情報を表 2 に示す．

4.2　ニュースカテゴリ説明文の生成

GPT-4 を用いたニュースカテゴリ説明文生成の実装に関する詳細を述べる．説明文の自動生成の実装においては，OpenAI が提供する「*gpt-4-1106-preview*」の API を利用した．

さらに，生成の際に GPT-4 に入力したニュースカテゴリとして，表 3 に示す *Subcategory* に格納されたニュースカテゴリ情報を用いた．これは，*Category* に格納されているカテゴリ情報と比較して，*Subcategory* に格納されたカテゴリ名の方が，より厳密で具体的なカテゴリを示しているためである．

4.3　評 価 指 標

評価指標について述べる．本研究では，先行研究 [1] [6] [17] [2] に倣い，評価指標として，Area Under Curve (AUC)，Mean Reciprocal Rank (MRR) および，normalized Discounted Cummulative Gain(nDCG@k，k=5，10) を用いた．

4.4　ベースライン手法

本研究で提案手法との比較対象として用いたベースライン手法である *title only* および *title + template-based* について説明する．

4.4.1　*title only*

title only では，ニュース推薦モデルの News Encoder にタイトル文のみを入力する．ニュースカテゴリに関する情報は一切入力に含めず，図 5 において D_{title} のみを入力し，推薦モデルの推論を行う手法である．

4.4.2　*title + template-based*

title + template-based は事前に定義した「*The news category is {category}*」というテンプレートにカテゴリ名を当てはめた文章を追加情報として付与する手法である．具体的には「タイトル文」と「該当するニュースカテゴリ名をテンプレートに当てはめた文章 (例．*The news category is tv-golden-globes*)」を SEP トークンで結合したテキストを News Encoder に入力する．

4.5　実 験 条 件

推薦モデルとして NRMS-BERT を用いて提案手法の検証を行った．検証に用いたニュース推薦モデルの News Encoder 部分で，単語ベクトル列の獲得に用いる事前学習済みモデルとしては DistilBERT-base(distilbert-base-uncased) [26] および

表 2　データセットの統計

項目	総数
ユーザ	94,057
ニュース	65,238
インプレッション	230,117
クリック	347,727

表 3　MIND [9] におけるニュースデータ

カラム名	説明	具体例
News ID	ニュース ID	*N37378*
Category	カテゴリ	*sports*
Subcategory	サブカテゴリ	*golf*
Title	タイトル	*PGA Tour winners*
Abstract	要約	*A gallery of recent winners on the PGA Tour.*
URL	URL	*https://www. msn. com/en-us/sports/golf/pga-tour-winners/ss-AAjnQjj?ocid=chopendata*

表 4　MIND [9] におけるユーザ閲覧履歴データ

カラム名	説明	具体例
Impression ID	閲覧履歴 ID	*123*
User ID	ユーザ ID	*U131*
Time	閲覧した時間.	*11/13/2019 8:36:57 AM*
History	ユーザーのニュースクリック履歴（ニュース ID）	*N11 N21 N103*
Impressions	表示されたニュースリストおよび各ニュースに対するクリック有無	*N4-1 N34-1 N156-0 N207-0 N198-0*

表 5　評 価 結 果

推薦モデル	手法	AUC	MRR	nDCG@5	nDCG@10
NRMS + DistilBERT	*title only*	0.674	0.297	0.322	0.387
	title + template-based	0.675	0.311	0.341	0.400
	title + generate-description(ours)	**0.707**	**0.324**	**0.359**	**0.422**
NRMS + BERT	*title only*	0.689	0.306	0.336	0.400
	title + template-based	0.667	0.301	0.329	0.389
	title + generate-description(ours)	**0.706**	**0.320**	**0.355**	**0.418**

BERT-base(bert-base-uncased) [5] を用いた．モデルは HuggingFace[5]の Transformers [27] および PyTorch [28] を用いて実装した．

最適化関数としては AdamW [29] を用いている．学習率は 1e-4，バッチサイズは 128，エポック数は 3 と設定した さらに，ユーザの嗜好情報を獲得するために用いる直近で閲覧した記事の本数の上限は 50 本とし，学習の際のネガティブサンプリング数は 4 とした．いずれも先行研究に倣っている [17] [2] [9] [14]．NVIDIA V100(1 台) 上でモデルの学習を行った．

5 実 験 結 果

ベースライン手法および提案手法の評価結果を表 5 に示す．*title + generate-description* が提案手法の評価結果である．

結果から提案手法は， ベースライン手法の *title + only* および *title + template-based* を上回る性能を示していることがわかる．News Encoder に BERT-base を用いたモデルと DistilBERT-base を用いたモデル両方において， 提案手法が全ての評価指標において，最も高い性能を示した．

6 終 わ り に

本研究では，ニュース推薦モデルがニュースカテゴリ情報に対し，より深い理解を得ることを目的に， 大規模言語モデルを用いてニュースカテゴリの説明文を自動で生成し，生成した説明文を追加情報としてニュース推薦モデルに入力する手法の提案を行った．実サービスのニュース記事データおよびユーザの閲覧履歴データにより構築された MIND を用いて，NRMS-BERT をファインチューニングすることにより， 検証を行った．実験の結果，提案手法は， 全ての評価指標において，ベースラインと比較し，最も高い性能を得た．

5：https://huggingface.co

文　　献

[1] C. Wu, F. Wu, Y. Huang, and X. Xie, “Personalized News Recommendation: Methods and Challenges,” *ACM Trans. Inf. Syst.*, vol. 41, pp. 1–50, 2023.

[2] C. Wu *et al.*, “Neural News Recommendation with Multi-Head Self-Attention,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6389–6394.

[3] M. An *et al.*, “Neural News Recommendation with Long- and Short-term User Representations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 336–345.

[4] C. Wu *et al.*, “NPA: Neural News Recommendation with Personalized Attention,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 2576–2584.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[6] C. Wu, F. Wu, T. Qi, and Y. Huang, “Empowering News Recommendation with Pre-Trained Language Models,” in

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 2021, pp. 1652–1656.

[7] C. Wu *et al.*, "MM-Rec: Visiolinguistic Model Empowered Multimodal News Recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, 2022, pp. 2560–2564.

[8] Q. Zhang *et al.*, "UNBERT: User-News Matching BERT for News Recommendation," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Virtual, 2021, pp. 3356–3362.

[9] F. Wu *et al.*, "MIND: A Large-scale Dataset for News Recommendation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Virtual, 2020, pp. 3597–3606.

[10] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[11] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[12] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[13] Q. Liu, N. Chen, T. Sakai, and X.-M. Wu, "ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models," *arXiv preprint arXiv:2305.06566*, 2023.

[14] J. Wang, Y. Jiang, H. Li, and W. Zhao, "Improving News Recommendation with Channel-Wise Dynamic Representations and Contrastive User Modeling," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, Singapore, 2023, pp. 562–570.

[15] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[16] K. Shivaram *et al.*, "Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation," in *Proceedings of the 16th ACM Conference on Recommender Systems*, Seattle, WA, USA, 2022, pp. 220–228.

[17] Z. Zhang and B. Wang, "Prompt Learning for News Recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Taipei, Taiwan, 2023, pp. 227–237.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[19] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019, pp. 13–23.

[20] K. M. Yoo *et al.*, "GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, 2021, pp. 2225–2239.

[21] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, BC, Canada: Curran Associates Inc., 2020, pp. 1877–1901.

[22] J. Liu *et al.*, "Generated knowledge prompting for commonsense reasoning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022, pp. 3154–3169.

[23] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? Generating customized prompts for zero-shot image classification," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 15 645–15 655.

[24] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[25] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint arXiv:2303.12712*, 2023.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2020.

[27] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Virtual, 2020, pp. 38–45.

[28] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019, pp. 8024–8035.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

LLMベース推薦システムにおけるユーザのネガティブ嗜好の活用

徐　雲帆†　牛尼　剛聡††

† 九州大学大学院芸術工学府　〒 815–8540 福岡県福岡市南区塩原 4 丁目 9-1
†† 九州大学大学院芸術工学研究院　〒 815–8540 福岡県福岡市南区塩原 4 丁目 9-1
E-mail: †xu.yunfan.518@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年，大規模言語モデル (large language model：LLM) の発展に伴い，LLM を活用した推薦システムに対する注目が高まっている．人が行う推薦では，まず，推薦者がユーザの好みに基づき推薦アイテムと説明を提示し，それから，ユーザが自分の嗜好に基づいて推薦者へフィードバックをする場合がある．このような推薦で，推薦者はユーザからのフィードバックを受けることで，高精度の推薦が可能になる．この絞り込みの手法を推薦システムで実現するには，ユーザの嗜好と推薦アイテムの差異を強調するため，ユーザのポジティブ嗜好だけではなく，ネガティブ嗜好の理解も必要であると考えられる．本論文では，LLM を活用したユーザのネガティブ嗜好の分析による推薦アイテムの選別効果に着目した推薦手法を提案する．LLaMA2 と MovieLens 1M のデータセットを用いた評価実験を行い，既存手法と性能を比較することにより，提案手法の有効性を検証する．

キーワード　推薦システム，大規模言語モデル，ネガティブ嗜好

1 初　め　に

推薦システムは人工知能の応用において，最も重要なアプリケーションの一つである．推薦システムは膨大な情報から，ユーザにとって有用な情報を予測してユーザに提供する．推薦システムは，主にユーザのインタラクション履歴を分析し，ユーザの嗜好を予測し，ユーザにパーソナライズされた推薦アイテムを提供する．例えば，映画推薦タスクの場合，推薦システムはユーザの映画鑑賞履歴やレビューなどの情報に基づいてユーザの嗜好を推測する．そして，推測された嗜好からユーザが興味を持ちそうな新しい映画をユーザに推薦する．

近年，大規模言語モデル (large language model：LLM) に注目が集まっている．特に LLM が有する強い言語理解能力と各下流タスクにおける汎化能力は研究者たちに強い印象を与えている．LLM ベースの手法は，様々な自然言語処理の分野において高い性能を示し，既に主流な手段の一つとなって幅広く活用されている．その中の一つとして，推薦システムに関連する文書の分析と理解，情報検索や対話的インタラクションなどに LLM の能力を適用できるため，LLM ベースの推薦システムに関する研究が注目されている [1].

LLM ベースの推薦システムにおいては主に以下の問題を解決する必要がある：

1. 膨大な知識を持つ LLM を特定領域の推薦タスクに適用することが必要である．Fine-tuning が行われていない LLM は大規模な言語知識を備える一般的なモデルで，ベースモデルでは特定のタスクに関する知識を活性化できていない．従って，ベースモデルは特定のタスクを高精度で実行するのは困難であり，推薦アイテムに関しても社会における一般的な評価に影響されやすい．この問題点を解決するには，Prompt Engineering と Instruction Tuning が多く使われている [2]．Prompt Engineering とはプロンプトで LLM にタスクの実施手順を教え込むことで出力の精度を向上させる手法である．例えば，タスク目標を Chain-of-Thought (CoT) [3] の形式で複数の段階に分解し，モデルの知識を特定のタスクに限定させる．一方，Instruction Tuning は特定の Instruction データセットにより Fine-tuning で LLM において特定のタスクに関する知識を誘出する．現在一般的に利用されている ChatGPT [4] などの LLM ベースのチャットモデルは，ほとんどベースの LLM から，汎用的な Instruction データセット (Alpaca [5] など) で Fine-tuning を行いプロンプトに対する理解能力を高め，高精度の対話能力を獲得したものである．現在，多くの LLM ベースのタスクはこの二つの手段を併用して高い精度を実現している．

2. LLM の出力には不完全なところがあり，ハルシネーション [6] が現れることがある．ハルシネーションが発生する原因は，LLM の出力には巨大な言語知識に基づき，人の言葉を模倣し正確そうな文書を生成するが，知識が不足している場合，モデルは正しそうな回答を作り上げてしまうためである．前述した Fine-tuning を行うことにより追加された知識があれば，この問題をある程度改善させることが期待できる．なお，LLM の Fine-tuning においては，LLM はモデル規模が非常に大きいため，伝統的な言語モデルに使われた全パラメータによる Fine-tuning の手法ではコストが大きすぎる．そこで，LoRA [7] など Fine-tuning のコストを軽減する手法が使われることが一般的である．

3. 一般的に LLM が処理できるプロンプトの長さは限定されている．最近では LLM が処理できるトークン列は長くなっているが，長いプロンプトに対する出力性能はあまり良くない [8]．その原因は，プロンプトが長ければその中に含むノイズも多くなり，出力の性能に悪影響を及ぼすためである．一方，長いプロンプトは処理時間も長くなり，モデル全体の効率にも下がることがある．故に，プロンプトの長さのコントロールも LLM

ベースの推薦システムにおいて重要なポイントの一つである．

以上のことにより，本研究ではこれらの問題点に注意しつつ，人が行う推薦のシチュエーションを参考にして推薦処理の流れをデザインし，LLM ベースの推薦システムを提案する．推薦処理の流れについては以下で述べる．



図 1　人が行う推薦例

人が推薦をする際に，図 1 に示すような流れが多く見られる．まず，推薦者がユーザの好みに関して尋ねる．次に，ユーザの好みに基づいて推薦アイテムと説明をユーザに提示する．ユーザは説明を参考にしながら自分の嗜好を踏まえて推薦者へフィードバックをする．推薦者はフィードバックを受け，それに応じて推薦アイテムを更に絞り込み，高精度の推薦を行う．

従来の手法では，このようなフィードバックにより推薦アイテムの絞り込みを行っている研究は少ない．その理由としては，この絞り込みの手法では，ユーザの嗜好と推薦アイテムの差異を強調するため，ユーザのポジティブ嗜好だけではなく，ネガティブ嗜好もとらえる必要がある．推薦タスクにおいて，ネガティブサンプルとポジティブサンプルのバランスをとることが困難である．実際にユーザが低い評価を付与したサンプルをハードネガティブサンプルと呼ばれている [9]．ハードネガティブサンプルの数は少ないため，多くの従来手法ではユーザ履歴にないアイテムからランダム抽出によりネガティブサンプルとして扱っている．その場合，ネガティブサンプルにノイズがあり，ネガティブ嗜好の推測精度に悪影響を及ぼす．また，ハードネガティブサンプルだけを利用し嗜好分析を行う場合，ポジティブ嗜好を扱うモデルや元の推薦モデルとの規模が異なるため，モデルの出力をまとめて処理することが困難である．

我々は LLM の強みを生かせば，これらの問題点を解決することが可能になると考えた．まず，LLM は特に zero-shot や few-shot などサンプルが少ない場合に性能が優秀であり，cold-start の推薦システムに強みを発揮できている [10] ため，比較的に少ないハードネガティブサンプルを対象としても適していると考えられる．図 2 の示すように，ChatGPT を使用し 10 件ユーザの映画鑑賞履歴からユーザのネガティブ嗜好推定を行った結果，少ないネガティブサンプルに基づき LLM はユーザのネガティブ嗜好をまとめることができた．また，LLM ベースの手法では，すべてのデータは自然言語の形で処理されるため，複数のモデル出力をまとめて統一な処理を行うことが可能になる．

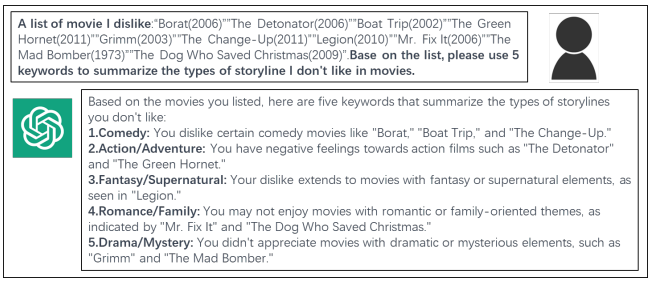


図 2　LLM によるネガティブ嗜好の推定

従って，本稿では推薦におけるフィードバックによる推薦アイテムの絞り込み手法に着目し，特にその中のネガティブ嗜好の推定とそれを活かす推薦手法の可能性を検討する．LLM ベースの推薦システムの問題点を含む，提案手法の流れを以下に示す．

（1） 推薦タスクを三つのステップに分け，ステップごとに処理を行う．

（2） 推薦タスクのデータセットを用いてポジティブ嗜好推定モデルとネガティブ嗜好推定モデルという 2 種類のモデルを構築する．そのためにデータセットを Instruction Tuning 用のデータ仕様に作り直し，それぞれ学習データとして LLM に Fine-tuning を行う．ポジティブモデルの入力に履歴のポジティブサンプルを使用し，ネガティブモデルの入力にはネガティブサンプルを使用する．

（3） テストデータのユーザ履歴から訓練された LLM のプロンプトを作る．推薦候補の生成には深層学習ベースの推薦モデルを Retrieval model として使用する．LLM のプロンプトで few-shot prompting を使用しポジティブ嗜好リストとネガティブ嗜好リストを出力する．

（4） 2 種類の嗜好推定モデル出力と Retrieval model の出力を合わせて重みづけ計算を行い，嗜好推定により調整後の推薦出力結果を従来手法と比較する．

本研究では推薦タスクのドメインとして映画推薦を選択した．使用するデータセットは MovieLens 1M [11] で，使われた LLM は LLaMA2-7B [12] である．提案システムの評価は従来の映画推薦モデルと性能比較を行った．

本論文の構成は以下の通りである．2 章では関連研究として LLM ベースの推薦モデルについて述べる．3 章では提案手法について述べる．4 章では評価実験の設定と結果について述べる．5 章では実験結果に関する考察について述べる．6 章では本研究のまとめと今後の展望について述べる．

2 関連研究

近年では，生成型 AI を利用した推薦システムが流行している．例えば，典型的な手法として P5 モデル [13] を利用した手法が提案されている．この手法では P5 言語モデルに対して推薦に関する複数の種類のタスクにそれぞれプロンプトを提案し，データセットを自然言語に統一し事前学習を行い，性能評価を行った．この研究は各推薦タスクに関して生成型推薦システムの基本的な手法をまとめたものである．

Chen ら [14] の研究では，ユーザのポジティブ嗜好を推定するように Instruction データを作り，LLaMA-7B に基づいて全パラメータで fine-tuning を行い，推薦候補から映画推薦を提供する．推薦候補の生成には従来手法による Retrieval model を使用している．提案手法と伝統的な推薦モデルの性能を比較することで LLM ベースの推薦モデルの有効性を検証している．この研究では主にポジティブサンプルを中心とした LLM を利用する手法を実現した．

Bao ら [15] の研究では LoRA を使って LLaMA の Fine-tuning を行い，複数の流行している推薦タスクのデータセットを用いて性能評価を行っている．この研究では Instruction の仕様では，入力はユーザの履歴アイテムはポジティブアイテムとネガティブアイテム一つずつであり，出力は一つのテストサンプルについてユーザは好きかどうかの二値分類である．この研究はネガティブサンプルを入力として扱えたが，ポジティブサンプルとネガティブサンプルの数はどちらも少ない．また，この研究の目標は特定のテストサンプルに関する好き嫌いの二値分類で，嗜好による推薦リストの絞り込みに関して検討していない．

その他，LLM を応用する推薦システムの手法として，LLM の出力を推薦モデルの知識拡張として使用する手法 [16] などが挙げられる．

これらの手法では，主にポジティブサンプルを中心に推薦精度の向上を目的として検討している．本稿で提案するネガティブ嗜好の推測とそれを用いる推薦リストの絞り込み手法についてはほとんど触れていない．ネガティブサンプルの効果は過去の推薦システムにおいて確認されているが，ハードネガティブサンプルの数が少ない問題点がそれを利用するネックとなっている [9]．本研究では汎化能力の高い LLM をネガティブサンプルの対応手段として提案し，ポジティブとネガティブ嗜好両方を考慮する推薦システムの構築を目指す．

3 提案手法

3.1 提案手法のフレームワーク

提案手法のフレームワークを図 3 に示す．まず，ユーザ履歴に基づき，Retrieval model により候補リストを出力する．次に，履歴からポジティブサンプルとネガティブサンプルをそれぞれ取りだし，Fine-tuning を行った二種類の嗜好推定モデルにより嗜好推測を行い，候補リストからポジティブリストとネガティブリストを生成する．最後に，候補リスト，ポジティブリストとネガティブリストにおいて各推薦アイテムに正規化を行い，重み付け計算により提案手法の推薦リストを出力する．

Retrieval model を導入する目的を以下に述べる．一つ目の目的は，ベースとなる推薦モジュールとして使用することである．人が行う推薦の流れでは，最初はベースの推薦リストを出し，次に嗜好推定による絞り込みを行うため，Retrieval model はベースの推薦リストを出すために使用する．二つ目の目的はプロンプトの長さ問題の解決である．すべてのアイテムを候補として扱うにはプロンプトが長すぎるため，Retrieval model で推薦候補を限定することにより，プロンプトを適切な長さにすることができる．本研究では SASREC [17] を Retrieval model として使用する．

Fine-tuning に使用する LLM は LLaMA2-7B に汎用 Instruction データによる Instruction Tuning で構築した LLaMA2-7B-chat である．事前テストを行った結果から LLaMA2-7B-chat は LLaMA2-7B より良い性能が見られた．その理由は，LLaMA2-7B-chat は LLaMA2-7B よりプロンプトに対する理解能力が高いと考えられる．実際，関連研究の Bao ら [15] の研究にもベースモデルを汎用 Instruction データで Instruction Tuning した LLaMA を使用した．

Fine-tuning に使用する Instruction データは訓練データから取り出している．Fine-tuning と重み付け計算の詳細は次の節で述べる．

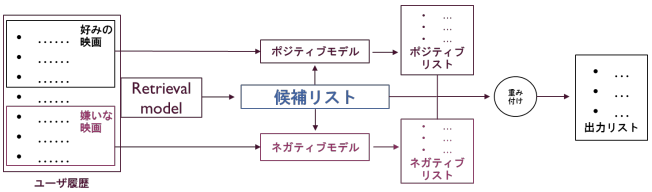


図 3 提案手法全体の framework

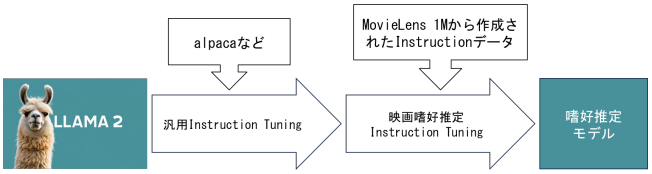


図 4 LLM の Fine-tuning の流れ

3.2 fine-tuning

Fine-tuning の流れを図 4 に示す．ベースの LLM に対する Tuning として，まず汎用 Instruction Tuning で LLM の指示文理解能力を高める．次に，MovieLens 1M のデータから Instruction Tuning 用に Instruction，Input，Output の文書をそれぞれ作成し，データセットを整理する．整理されたデータセットを使用し LLM を嗜好推定モデルに Fine-tuning する．

MovieLens 1M データセットから，2 種類の嗜好推測モデルの instruction データは以下のように整理する．

ポジティブモデルの Instruction データは図 5 で示す．具体的な指示は『ユーザの鑑賞履歴に基づいて，候補リストから 10 件他の映画を推薦せよ．推薦度順で並べる』である．入力は 10 件のユーザ鑑賞履歴，推薦候補はその 10 件の履歴以外の履歴 20 件で，出力はその 20 件の履歴におけるそのユーザが付与したレーティング上位 10 件の映画である．

```
Few-shot Example/instruction(positive):
Instruction:recommend 10 other movies and rank them by recommendation level based on user's watching history from the candidate list , sorted by recommendation level.
input:User watched movies {watching history}.the candidates are {candidate list}.
output:{recommendation list}.
```

図 5　ポジティブモデルの instruction

ネガティブモデルの Instruction データを図 6 に示す．具体的な指示は『ユーザの鑑賞履歴に基づいて，ユーザが低いレーティングをつける他の映画 10 件を予測せよ．1 番目は最もレーティングが低いアイテム』というものである．入力はユーザが低いレーティングを付与した映画 5 件，推薦候補はその 5 件の履歴以外の履歴 20 件で，出力はその 20 件の履歴におけるそのユーザが付与したレーティング下位 10 件の映画である．データセット全体のネガティブレーティングの数はポジティブアイテムの半分以下のため，ネガティブモデルの履歴アイテム数はポジティブモデルの半分にした．

```
Few-shot Example/instruction (negative):
Instruction:Predict 10 other movies that users will give low ratings to based on user's watching history from the candidate list and rank them, for 1 is the worst one.
input:User gave low ratings to movies {watching history}.the candidates are {candidate list}.
output:{recommendation list}.
```

図 6　ネガティブモデルの instruction

LLM の Fine-tuning は膨大な計算量を要するため，本稿では Fine-tuning の手法として Qlora [18] を使用した．また，LLM は優れた汎化能力を持つため，Instruction Tuning に使用したデータはすべての訓練データではなく，20%の訓練データを使用した．

3.3　重みづけ計算

図 3 のように，モデルの出力は 3 つの順位リストである．(1) Retrieval model の出力の推薦，(2) ポジティブ嗜好推定モデルからのポジティブリスト，(3) ネガティブ嗜好推定モデルによるネガティブリスト．提案手法では，嗜好推定モデルはベースの推薦リストを絞り込むように処理するため，推薦リストにおいて各アイテムのスコアはベースの順位スコア $Score_{rec}$ と嗜好スコア $Score_{pref}$ の 2 つの部分に分けられる．2 つのスコアは重み α をかけて足す．具体的には以下の式で表す：

$$Score_i = \alpha * Score_{rec} + (1 - \alpha) * Score_{pref} \tag{1}$$

ベースの順位スコアは以下の式で表す：

$$Score_{rec} = \frac{1}{R_{base}} \tag{2}$$

R_{base} はベースの推薦リストの順位である．嗜好推定スコアはそれぞれポジティブリストとネガティブリスト順位の逆数の差を取る．具体的には以下の式で表す：

$$Score_{pref} = \frac{1}{R_{pos}} - \frac{1}{R_{neg}} \tag{3}$$

3 つの出力リストからベースの推薦リストの各アイテムのスコア $Score_i$ を計算し，スコアから再び順位リストを作り，提案手法の出力リストとして扱う．

本稿では，LLM が一般評価に影響されやすい問題点を考慮して，α の値は 0.6 に設定した．

4　評 価 実 験

4.1　実 験 設 定

4.1.1　Dataset

データセットの設定はベースラインの手法と同じ，インタラクション記録があるアイテムは 1 で，それ以外のアイテムは 0 として扱う．各ユーザの最後のインタラクションアイテムを正解アイテムとして，それより前のアイテムはすべて履歴として扱う．モデルの学習データは各ユーザの最後の二つのインタラクション記録を除くデータで，最後のアイテム記録はテストデータとして使用し，学習の検証データはその前のアイテム記録である．MovieLens 1M データセットの詳細は表 1 で示す．

Users	item	interactions
6040	3883	1000209

表 1　MovieLens 1M の基本情報

LLM の Instruction Tuning 用データは学習データから構築した．LLM のプロンプトの長さをコントロールするため，ポジティブ嗜好推測モデルにはアイテム履歴のレーティング上位 10 件を履歴として使用し，ネガティブ嗜好推測モデルにはレーティング下位 5 件をの履歴として使用する．

4.1.2　評 価 指 標

ネガティブ嗜好の絞り込み効果を検証するため，正解データの順位を重視する NDCG を評価指標として使用した．出力リストの長さは 10 件に限定する．

4.1.3　ベースライン

GRU4REC [19]：RNN ベースのシーケンシャル推薦システム，ユーザのインタラクション履歴を GRU で処理する手法である．

SASREC [17]：セルフアテンション機構を用いる手法で，ユーザの好みの変化を考慮する推薦モデルである．

few-shot：Fine-tuning を行っていない LLaMA2-7B-chat の出力をベースラインの一つとして使用する．提案手法の Instruction データと同じフォーマットの few-shot prompting による出力する．重み付け計算を行わずに出力結果で評価を行った．

4.2 性能比較

Model	NDCG@10
GRU4REC	0.5055
SASREC	0.5806
LLaMA2-chat+few-shot	0.2771
Retrieval-positive	0.5877
Retrieval-positive+negative	**0.5950**

表 2　性能評価表

評価実験の結果を表 2 に示す．提案手法の性能はベースラインの推薦モデルより上回り，有効性が確認された．ネガティブ嗜好推定を加える推薦モデルの評価は最も良い結果であり，ネガティブ嗜好推定の絞り込み効果が見せられている．

Fine-tuning を行っていない LLM の出力結果に重のみ付け計算を行わず，直接評価を行った場合性能が低いことが示された．その原因として考えられるのは以下 3 つである．まず，LLM の推薦は推薦モデルより一般評価に影響される傾向が強い．次に，プロンプトの長さや情報不足の原因でハルシネーションや生成失敗のサンプルが多く見られた．最後に，Top-N 推薦タスクでは履歴にあるアイテムがポジティブサンプルで，他のアイテムはネガティブサンプルとして扱い．レーティング情報は使用していないため，正解データはユーザが低いレーティングを付けた映画の場合もある．その場合，LLM ではユーザの嗜好推定を行ったため，逆に評価の精度に悪影響を及ぼすことになる．

提案手法では，一般評価の影響はある程度 Fine-tuning でカバーすることが可能である．また，一般評価が多いアイテムでは，高い評価と低い評価の両方が多くなるため，ポジティブ嗜好推定とネガティブ嗜好推定両方のリストにおいて，共に高い順位になることがある．二つの出力を組み合わせ，重み付け計算を導入することで一般評価の影響を軽減することが可能になる．Fine-tuning した LLM の出力を確認することにより，ハルシネーションと生成失敗のサンプルの数は明らかに減少した．正解サンプルについては重みを調整することである程度影響を減らすことができる．この問題点について次の章で詳しく説明する．

5 考　　察

前述のように，Top-N 推薦のタスクにおいて，推薦モデルはユーザの鑑賞履歴を分析し，ユーザが興味を持ちそうな新しいアイテムを推測する．入力に使用するデータでは，鑑賞した映画はすべてポジティブサンプルとして扱い，鑑賞していない映画はネガティブサンプルとして分析を行った．モデルの学習にレーティング情報は利用していない．故に，テストデータの正解アイテムはユーザが低いレーティングを付けた場合もある．一方，本研究の嗜好推定モデルではユーザの嗜好を考慮する推薦を行うため，元々正解アイテムはユーザが低いレーティングを付けられた場合ノイズになり，推薦の精度に悪影響を及ぼすことがある．これは提案手法の性能評価が提案手法より大きく改善されていない原因の一つであると考えられる．

以上の要素の影響を確認するために，100 件の出力を取り出し，その中の提案モデルにより順位調整の成功例（正解アイテムの順位が上がった例）と失敗例（正解アイテムの順位が下がった例）をそれぞれ抽出し，そのレーティングを確認した．その結果は以下の図で示している．

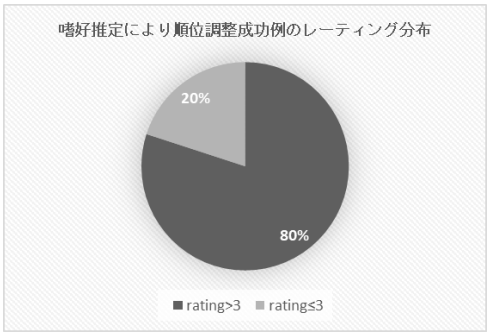


図 7　嗜好推定により順位調整成功例の正解サンプルレーティング分布

嗜好推定モデルの結果により推薦リストの精度を向上させた例において，正解サンプルのレーティング分布は図 7 に示している．順位調整により精度を向上させたサンプルにおいて 80%の正解サンプルはユーザが 4 点以上のレーティングを付けたサンプルである．正しくユーザの嗜好を表すサンプルに，嗜好推定により推薦リストの調整効果が確認された．

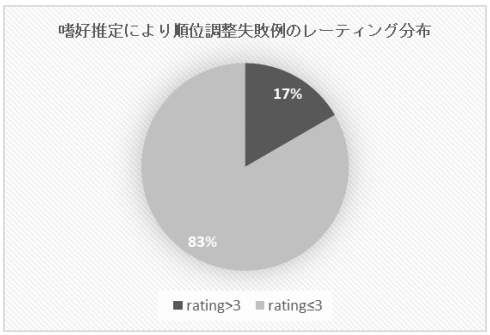


図 8　嗜好推定により順位調整失敗例の正解サンプルレーティング分布

図 8 に示すように，嗜好推定モデルにより順位調整が失敗した 83%の正解サンプルはユーザが 3 点以下のレーティングを付けたサンプルであった．推定モデルはユーザの嗜好を推定することで，指定アイテムはユーザの好みではないと判断し，アイテムの順位をリストの下位にランキングした．この場合，実際のところ嗜好推定の結果は正確であるが，評価実験の観点から見ると失敗例として扱われることになる．

以上の結果から見ると，正解サンプルに対するユーザ嗜好は評価結果に影響を及ぼすことが確認された．一方，この実験の結果から，嗜好推定モデルの嗜好推定の有効性も示された．この問題点の原因は，タスク目標とモデル出力の差異である．より適切なランキングタスクにおいて，提案手法の評価性能は更に向上させられる可能性があると考えられる．

また，図 9 は提案手法による推薦生成例の一つである．映画タイトルの後のスコアは社会の一般評価によるレーティングで

ある．全体的に一般評価が高い映画が上位で，LLM による推薦は一般評価に影響される傾向は見られるが，一部の一般評価が低い映画の順位は上位になっているため，Fine-tuning によるパーソナライズの効果が確認できる．

Output:
1. Pulp Fiction (1994) **(9.3/10)**
2. 10 Things I Hate About You (1999) **(6.9/10)**
3. Being John Malkovich (1999) **(7.1/10)**
4. Ghostbusters (1984) **(5.8/10)**
5. Clerks (1994) **(8.1/10)**
6. Magnolia (1999) **(7.8/10)**
7. Office Space (1999) **(8.0/10)**
8. Shakespeare in Love (1998) **(7.1/10)**
9. X-Men (2000) **(7.4/10)**
10. Superman (1978) **(5.3/10)**

図 9　提案手法による推薦例

推薦生成例全体から見ると，汎用 Instruction データによる Instruction Tuning を行った LLM の出力は，ベースの LLM より安定していない．教え込むフォーマットと若干異なるフォーマットで生成される出力が多く現れた．また，プロンプトの長さはある程度制限されたが，鑑賞履歴と候補リストをプロンプトに入れないといけないため，依然として長さは長めである．これらの原因で生成にハルシネーションと重複などの失敗例はまだある．これからは複数プロンプトの出力チェインによりプロンプトの長さを更に縮め，生成精度の向上を目指したい．

6 ま と め

本研究では，人の推薦で行うユーザ嗜好の推定による推薦リストの絞り込みを参考にし，LLM を利用した推薦手法を提案した．絞り込みは差異の比較であるため，嗜好推定はポジティブ嗜好だけでなく，ネガティブ嗜好の効果も考慮する必要がある．嗜好推定モデルには汎化能力が強い，言語理解能力を持つ LLM を使用した．LLM の Tuning では，Fine-tuning することでモデルのパーソナライズ化をし，Retrieval model を使用することでプロンプトの長さを制限させた．ポジティブ嗜好とネガティブ嗜好の推定結果を合わせて重み付け計算を行い，推薦リストの絞り込みを実現させた．MovieLens 1M データセットを用いて評価実験を行った．評価実験の結果において提案手法の性能が最も良いため，提案手法の有効性を検証した．

今後の展望として，前述の推薦流れをフレームワークとして LLM ベースの対話型推薦システムを実現させたい．対話型推薦システムは大きく分けて『請求』『推薦』『説明』『応答』四つのタスクがある [20]．本稿ではその中のフィードバックおよび推薦アイテム順位調整の流れについて検証した．今後は他のタスクも LLM の言語理解能力を活用し，それぞれプロンプトにより提案し，モデルの間の出力を繋げ，高精度の対話型推薦システムの構築を目指したい．

文　献

[1] Fan, W., Zhao, Z., Li, J., Liu, Y., Mei, X., Wang, Y., ... Li, Q. (2023). Recommender systems in the era of large language models (llms). arXiv preprint arXiv:2307.02046.
[2] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., ... Chen, E. (2023). A Survey on Large Language Models for Recommendation. arXiv preprint arXiv:2305.19860.
[3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.
[4] https://chat.openai.com
[5] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model.
[6] Rawte, V., Sheth, A., Das, A. (2023). A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922.
[7] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
[8] Li, L., Zhang, Y., Chen, L. (2023, October). Prompt distillation for efficient llm-based recommendation. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 1348-1357).
[9] Ma, H., Xie, R., Meng, L., Chen, X., Zhang, X., Lin, L., Zhou, J. (2023, September). Exploring false hard negative sample in cross-domain recommendation. In Proceedings of the 17th ACM Conference on Recommender Systems (pp. 502-514).
[10] Sanner, S., Balog, K., Radlinski, F., Wedin, B., Dixon, L. (2023, September). Large language models are competitive near cold-start recommenders for language-and item-based preferences. In Proceedings of the 17th ACM conference on recommender systems (pp. 890-896).
[11] https://grouplens.org/datasets/movielens/1m/
[12] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
[13] Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y. (2022, September). Recommendation as language processing (rlp): A unified pretrain, personalized prompt predict paradigm (p5). In Proceedings of the 16th ACM Conference on Recommender Systems (pp. 299-315).
[14] Chen, Z. (2023). PALR: Personalization Aware LLMs for Recommendation. arXiv preprint arXiv:2305.07622.
[15] Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., He, X. (2023). Tallrec: An effective and efficient tuning framework to align large language model with recommendation. arXiv preprint arXiv:2305.00447.
[16] Lyu, H., Jiang, S., Zeng, H., Xia, Y., Luo, J. (2023). Llm-rec: Personalized recommendation via prompting large language models. arXiv preprint arXiv:2307.15780.
[17] Kang, W. C., McAuley, J. (2018, November). Self-attentive sequential recommendation. In 2018 IEEE international conference on data mining (ICDM) (pp. 197-206). IEEE.
[18] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.
[19] Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D. (2015). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.
[20] Jannach, D., Manzoor, A., Cai, W., Chen, L. (2021). A survey on conversational recommender systems. ACM Computing Surveys (CSUR), 54(5), 1-36.

大規模言語モデルに基づく検索手法の日本語文書検索への適用

阿部　健也†　　薄羽　皐太†　　加藤　　誠††

† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †{s2321672,s2221635}@u.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　大規模言語モデル (LLM) に基づいた文書検索は近年研究が進展している一方で，日本語に注目した文書検索に関する研究の例は少ない．LLM を利用した日本語の文書検索では，エンコーダとしての言語モデルや学習データをどう用いるのが適切であるのか，モデルサイズや検索モデルの違いによってどのような影響が生じるのかなどを網羅的に調査した研究は存在しない．本研究では LLM に基づいた検索手法を日本語に適用した実験を行うことでこれらの問題を明らかにする．さらに，実験結果やデータの分析から翻訳データに学習に寄与していないクエリが存在することを考察し，翻訳データの中から質の悪いクエリを取り除くために検索結果を利用した基準を提案した．提案したフィルタリング基準により，学習結果を改善できるケースがあること，学習に有効ではない翻訳クエリが存在することを明らかにした．

キーワード　情報検索、検索モデル、LLM、言語モデル

1　はじめに

文書検索はユーザーが与える情報要求を表すクエリによって文書集合から適切な文書を返すタスクであり，Web 検索，質問応答システム，推薦システムなど多くの技術に応用することができる．文書検索タスクは BERT [5] や RoBERTa [17] などの大規模言語モデル（LLM）の登場や MS MARCO [22] と呼ばれる英語の検索タスクの大規模なデータセットが公開されたことにより研究が進み，LLM に基づいた検索モデルの多くが従来の手法を大きく上回った [7, 13, 14, 27].

LLM に基づいた検索モデルを英語以外の言語へ適用する研究はいくつか存在するが，日本語に注目した検索に関する研究の例は少ない．MIRACL [34] や Mr.TyDi [32] などでは中国語検索，アラビア語検索等の様々な言語の単言語検索に取り組むためのテストコレクションが作成されている. [33] では学習データや学習手法，エンコーダとして用いる言語モデルの違いによって生じる様々な言語における精度への影響を Mr.TyDi で評価し，多言語検索における最善策を検証した．これらのデータセット，実験には日本語も含まれているがあくまで多言語の検索に着目しているため，言語の 1 つとして扱われるに留まり，個別の詳細な分析は行われていない．LLM が登場する以前には情報検索のワークショップ NTCIR において日本語の文書検索タスク [12] が存在したが，2022 年まで同タスクは開催されていなかった [1]．このように英語以外の検索という方向での研究は進められているが，特に日本語に注目した検索に関する研究の例は少ない．

英語における文書検索タスクでは，構造や学習方法に関して多くの知見が [35] が蓄積されている．例えば，検索モデルの学習に関する分析 [30, 31] が，現在における検索モデルを利用する際の指針や精度改善の基盤となっている．しかし，日本語ではこのような分析は行われておらず，検索モデルをどのように用いるのが適切であるのかは分かっていない．そのため，本研究では日本語の文書検索に関して，以下のリサーチクエスチョン（RQ）の検証を行う．

- RQ1：エンコーダとしての言語モデルとそのための学習データとして適切な組み合わせは何か
- RQ2：言語モデルのサイズと検索モデルの構造によって検索精度はどのように変化するのか
- RQ3：機械翻訳したデータを学習に用いる場合，データをフィルタリングすることで学習結果を改善できるか

はじめに RQ1，RQ2 に関する実験を行う．日本語の検索においては検索モデルの学習に関して以下の 3 つのアプローチが考えられる.（1）エンコーダとして日本語の言語モデルを用いて日本語のデータで学習する.（2）多言語モデルを日本語のデータで学習する.（3）多言語モデルを英語のデータで学習する．英語以外の言語ではモデルの学習に十分なデータを用意することができないため,（1）と（2）で利用する日本語のデータは英語のデータを翻訳することで生成する．英語の文書検索において用いられることの多い 3 つ検索モデルを基本的な構造として，上記 3 つのアプローチやその他の設定で実験を行うことで，日本語文書検索における最適な設定に関して調査を行う．

1 つ目の実験結果や分析結果から翻訳したデータの翻訳精度が学習に寄与していないクエリが存在すると考えられ，RQ3「機械翻訳したデータを学習に用いる場合，データをフィルタリングすることで学習結果を改善できるか」という新たに生じた RQ に関する実験を行う．2 つ目に行った実験では，検索結果に基づいてクエリをフィルタリングした学習データで検索モデルを学習することでフィルタリングの効果を検証した．また，学習に有効でないクエリの存在を確認するために，基準を

1：我々は 2023 年に開催された NTCIR-17 Transfer タスク [11] に参加し，日本語文書検索に取り組んでおり [1]，本研究はこの実験を拡張したものである．

満たさない翻訳クエリのみを残す逆フィルタリングによる実験を行った．

2 つの実験結果より次のことがわかった．

- RQ1：同じモデルサイズであれば日本語モデルに日本語翻訳データを用いることが最適である．
- RQ2：基盤モデルのサイズが大きくすることで精度を改善することができ，最も精度が高い検索モデルの構造は ColBERT であった．
- RQ3：翻訳したデータを学習に用いる際，フィルタリングにより学習結果を改善できるケースがある．学習に貢献していないクエリが存在する可能性があることも明らかになった．

本論文の構成は以下の通りである．2 節では LLM に基づいた検索モデル，多言語検索における LLM ベースの検索モデルに関する既存研究について述べる．3 節では実験に使用するモデルや検証する手法について述べる．4 節ではデータセットと実験の設定，実験結果を示し，5 節では今後の課題と共に本論文の結論を述べる．

2 関 連 研 究

本節では，LLM に基づいた検索モデルに関する既存研究について述べたのち，LLM ベースの検索モデルの多言語への適用について述べる．

2.1 LLM に基づいた検索モデル

LLM に基づいた検索モデルに関する関連研究について述べる．一般的に検索モデルではクエリと文書をベクトルで表現し，2 つのベクトルの内積等の計算をすることにより類似度スコアを出力する．以前のアドホック検索では，BM25 などの単語一致を利用した手法が長い間主流となっていた．BM25 などではクエリや文書を単語の出現に基づいた高次元（語彙数の次元）でほとんどの次元がゼロである疎なベクトルで表す．このようにベクトル表現の多くの要素がゼロであるような検索手法を Sparse Retrieval（疎検索）と呼ぶ．従来の疎検索では，同義語や言い換え表現などをマッチしない単語として見逃してしまうことが課題である．この課題を解決するために，BERT [5] をはじめとする LLM によりクエリや文書を低次元なベクトルで表現する Dense Retrieval（密検索）[35] が提案され，この方法は単語の意味まで考慮することができる．密検索は多くのケースにおいて単語一致の手法よりも有効であることが分かっているため，近年では密検索手法が主流となってきている．Karpukhin らが提案した Dense Passage Retrieval (DPR) [13] は，クエリと文書を別々に BERT に入力してそれぞれを 1 つのベクトルに変換し，その 2 つのベクトルの内積により類似度を計算する手法である．このようなモデルをシングルベクトルのデュアルエンコーダと呼び，計算コストが比較的低いため，最も広く用いられる構造である [8, 9, 16, 27]．Khattab らが提案した ColBERT [14] では，BERT によってクエリと文書をトークン毎にベクトルに変換して類似度を計算すること（マルチベクトル）で，1 つのベクトルにまとめて計算するよりも精度の高い検索が可能になった．密検索の研究が進められる一方で LLM を利用して，教師ありの疎検索モデルも提案されている．[4, 7] Dai らはクエリと文書に登場する単語の重みを BERT で予測する DeepCT [4] を提案した．Formal らが提案した SPLADE [7] ではクエリや文書に登場しない単語に対しても重みを予測を行う．本実験ではこれらの中でも代表的なモデルである DPR, ColBERT，SPLADE を検索モデルとして利用した．

LLM に基づいた検索モデルは数多く提案され，それぞれのモデルの詳細な分析や改善手法に関しても多くの研究が行われている [15, 23, 26, 27]．Wang ら [26] は ColBERT に着目し，エンコーダとして用いる言語モデルの違いによってどのような違いがあるのか，ColBERT の類似度計算の挙動についての分析を行っている．mackenzie ら [18] は SPLADE の表現力に関する実験を行った．この実験では SPLADE の語彙を制限することで疎ベクトルをより小さい次元で表現した際の影響について分析を行っている．一方で日本語においては検索モデルに関する分析は非常に少ない．本研究ではこれらの研究が行なっているような各検索モデルの挙動に関する詳細な分析ではなく，いくつかの検索モデルを日本語文書検索に適用する際の最適な設定に関する傾向を明らかにすることを目的としている．

2.2 多言語検索における LLM ベースの検索モデルの利用

多言語検索における LLM ベースの検索モデルに関する研究について述べる．中国語検索，アラビア語検索など様々な言語の単言語検索をまとめた検索を多言語検索と呼んでいて，多言語検索では Mr.TyDi，MIRACL [32, 34] などのデータセットが公開されている．Zhang らの研究 [33] では学習データや学習手法，基盤モデルの違いによって生じる様々な言語における精度への影響を Mr.TyDi で評価し，多言語検索における最善策を検証した．実験には日本語も含まれているがあくまで多言語の検索に着目しているため，言語の 1 つとして扱われるに留まり，個別の詳細な分析は行われていない．また本研究では比較に日本語特化の言語モデルの追加している点や翻訳された学習データに関する実験，分析を行っている点が異なる．

Nair らは，英語のクエリで他の言語の文書を検索するタスク（言語横断情報検索）において ColBERT を多言語に応用した ColBERT-X を提案した [19]．ColBERT-X ではエンコーダとしての言語モデルを BERT から多言語モデルである XLM-RoBERTa に置き換えている．多言語モデルでは事前学習のデータに英語以外のデータが含まれているため，ファインチューニングの際には目的の言語のデータが含まれていなくても目的のタスクに適応させることができる．そのため，学習時には英語のデータで学習を行う Zero-Shot（ZS）と翻訳したデータを用いる Translate-Train（TT）の 2 つが提案されている．Nair らの実験から，翻訳して得られたデータで学習することで英語のデータで学習したモデルを上回るケースがあることが明らかとなった．本研究は言語横断情報検索ではなく日本語の単言語検索に取り組んでいる点，エンコーダとして多言語モデルのみでなく日本語特化の言語モデルでも実験を行っている点が異なる．

3 手　　法

3.1 問題設定

アドホック検索は文書集合を D としてクエリ $q \in Q$ が与えられたとき，q に適合する順に文書をランク付けする問題である．特に，本論文では日本語の文書集合 D 中の $d \in D$ に対して，q との適合度スコア $s_{q,d}$ を推定し，スコアを降順に並べることで文書をランク付けを行う．

3.2 DPR

近年，アドホック検索では，BERT をエンコーダとしてクエリと文書をベクトルに変換して類似度を計算する手法が有効である．その手法の 1 つである DPR [13] は，クエリと文書をそれぞれベクトルに変換してその類似度を計算することで検索を行う．検索時には近似最近傍探索 [10] を利用して大量の文書から上位の文書になりうる文書を検索する．

DPR の学習ではあるクエリ q，そのクエリに対する適合文書 d^+，非適合文書の集合 $K = \{d_1^-, d_2^-, ...d_{|B|}^-\}$ を用意する．この時，B は非適合文書の集合である．DPR においては事前に用意した 1 個の非適合文書に加えて，バッチ内の他のクエリに対する適合文書を追加の非適合文書としている．つまり，バッチサイズが $|B|$ である場合には，事前に用意した 1 個の非適合文書とバッチ内の適合文書を $(|B|-1)$ 個用いることになる．クエリ，適合文書，非適合文書をそれぞれエンコーダに与えベクトルに変換した後，クエリ q_i と適合文書 d_i^+，クエリ q_i と適合文書 d_i^- の 2 つのペアの適合度（内積）$s_{q,d}$ は式 1 である，この時，損失関数は以下の式 2 のように計算する．

$$s_{q,d} = \eta(q_i)\eta(d_j)^T \tag{1}$$

$$L = -\frac{\exp(s_{q,d^+})}{\exp(s_{q,d^+}) + \sum_{i=1}^{B} \exp(s_{q,d_i^-})} \tag{2}$$

3.3 ColBERT

ColBERT [14] は，クエリと文書をそれぞれトークン毎にベクトルに変換してその類似度を計算することで検索を行う．より具体的には，クエリの各トークンに対して文書の各トークンとの類似度を計算して，最も高い類似度の和を取る．3 の式はある文書 d のクエリ q に対する類似度を計算する式である．クエリ q，文書 d が与えられた場合，クエリをトークナイザーによってトークンに分割することで $q = (q_1, q_2, \ldots, q_{|q|})$, 文書も同様に分割することで $d = (d_1, d_2, \ldots, d_{|d|})$ とする．これらのトークンは BERT を使用したエンコーダ η によってクエリと文書のベクトル $\eta(q_i)$，$\eta(d_i)$ に変換される．

$$s_{q,d} = \sum_{i=1}^{|q|} \max_{j=1,\ldots,|d|} \eta(q_i)\eta(d_j)^T \tag{3}$$

式 3 ではクエリの各トークンのベクトルに対して，文書の各トークンのベクトルとのコサイン類似度を計算し，クエリのトークン毎にそのトークンのベクトルと最も類似度の高い文書のトークンのベクトルの類似度を加算していき，その合計をクエリ q に対する文書 d の適合度 $s_{q,d}$ とする．

$$L = -\frac{\exp(s_{q,d^+})}{\exp(s_{q,d^+}) + \exp(s_{q,d_i^-})} \tag{4}$$

ColBERT の損失関数では非適合文書は 1 つであり，バッチ内の他の文書を利用せずに損失を計算する．今回の実験ではエンコーダとして XLM-RoBERTa，日本語 BERT を使用した．

3.4 SPLADE

SPLADE [7] は BERT に語彙次元 $|V|$ の疎な表現をさせながら，同義語や言い換え表現のミスマッチを防げるようにすることを目的としていて，BERT が事前学習時に用いている MLM head をエンコーダの最後の層に接続することで BERT の tokenizer が持つ語彙次元の出力をさせている．全ての語彙に対して重みを予測させることで，BM25 のように疎な表現を持ちながら同義語や言い換え表現のミスマッチを防ぐことができる．クエリと文書が q，d であり，それぞれがトークナイザーによって $t = \{t_1, t_2, ...t_{|t|}\}$ のように分割されるとするとエンコーダによってトークンは $\{h_1, h_2, ...h_{|t|}\}$ のような埋め込みに変換される．そこで，トークン i に対する語彙 j の重要度 $w_{i,j}$ を以下のように計算する：

$$w_{i,j} = \phi(h_i)^T E_j + b_j \quad j \in \{1, ...|V|\} \tag{5}$$

E_j は BERT の入力埋め込み層のトークン j を入力した際の出力であり，GeLU 関数と LayerNorm 層を含む線形層 ϕ を通した埋め込み h_i との内積を計算し，トークンバイアス b_j を加えることで，トークン i に対する語彙 j の重要度 $w_{i,j}$ を計算する．シーケンスのトークン毎に全語彙に対する重要度を計算することでシーケンス長 $|t|*$ 語彙サイズ $|V|$ の行列が出力される．最終的な出力を語彙サイズの疎なベクトルとするために ReLU 関数で全ての値調節した後に語彙次元で Max プーリングを適用する：

$$w_j = \max_i \log(1 + \mathrm{ReLU}(w_{i,j})) \tag{6}$$

式 5，6 を用いてクエリ q と文書 d はある語彙 j に対する重要度を w_j^q，w_j^d のように持つ語彙次元の疎なベクトルに変換することができ，SPLADE のスコアは次のように計算される．

$$s_{q,d} = \sum_i w_j^q w_j^d \tag{7}$$

SPLADE では DPR と同様に式 2 の損失を用いる．モデルが出力するベクトルが疎であることを保証するために [6, 7, 29] などで様々な方法が調査されているが，今回は [29] を参考に出力の上位 topK 以外を 0 とする方法を選択した．今回の実験ではエンコーダとして日本語 BERT と bert-base-multilingual-cased（mBERT）を用いた．mBERT は多言語モデルであり，語彙が大きくなり過ぎてしまう問題に対処するため，mBERT を SPLADE のエンコーダとして用いている既存の研究 [20, 21] を参考に，出力する語彙を日本語のみに制限する方法を選択した．

4 実　　験

4.1 実験の概要

実験の概要について説明する．本研究では1つ目の実験で**RQ1**：「適切な基盤モデルと学習データの組み合わせは何か」を明らかにするために実験を行う．そのために（1）目的の言語の言語モデルを日本語翻訳データで学習する．（2）多言語モデル [3] を日本語翻訳データで学習する．（3）多言語モデルを英語のデータで学習する．という 3 つの学習に関するアプローチを適用する．さらに 1 つ目の実験では，**RQ2**：「基盤モデルのサイズと検索モデルの構造によって検索精度はどのように変化するのか」についても検証する．RQ2 については DPR, ColBERT，SPLADE といった検索モデルのエンコーダを変更して実験することで確認している．**RQ3**：「機械翻訳したデータを学習に用いる場合，データをフィルタリングすることで学習結果を改善できるか」については実験 2 で検証した．実験 2 では学習データのクエリに基準を設けてフィルタリングを行うことで学習データの質の改善を図り，基準を満たさないクエリのみを残す逆フィルタリングによって学習に貢献していないクエリが存在する可能性を示す．

表 1　テストコレクションの統計情報

	文書数	クエリ数	適合文書
NTCIR2	74 万	49	5093
Mr.Tydi	700 万	720	923
MIRACL	695 万	860	1790

4.2 実 験 設 定

我々は NTCIR2 [12]，Mr.TyDi [32]，MIRACL [34] をデータセットとして検索手法の評価を行った．NTCIR2 は日本語の論文検索のデータセットで，Mr.TyDi と MIRACL は Wikipedia をコーパスとしたデータセットである．Mr.TyDi と MIRACL は類似したデータセットであるが，Mr.TyDi は Google 検索，MIRACL では既存の検索手法に基づいて検索を行い，適合性判定を行っている点が異なる．モデルを学習するためのデータとしては MS MARCO と mMARCO を用いた．MS MARCO [22] はマイクロソフトが公開した英語のデータセットで 50 万のクエリと 880 万の文書でクエリ 1 つあたり適合文書が 1 つ以上与えられている．mmarco は MS MARCO を機械翻訳モデルで機械翻訳したものであり，日本語の学習データとして mmarco の日本語データを利用する．

本研究においてのベースライン手法と実験の設定について述べていく．我々は本節で紹介したモデルに加えて比較する手法として，単語一致を利用した手法である BM25 を用いた．評価指標は Recall@1000，nDCG@1000，MAP@100 を用いて各モデルを評価した．学習では microsoft が提供している triples.train.small[2]を学習データとして使用し，フィルタリングを行う場合を除いてデータは統一している．また，推論時に用いるコーパスやクエリによる学習は行わず，本研究ではこれを Zero-Shot 設定と呼ぶ．日本語の言語モデルを用いたエンコーダには東北大が提供する日本語 BERT（J-BERT）[3] した．多言語モデルには mBERT[4]，XLM-RoBERTa[5]を用いた．学習率は既存の研究 [1, 2, 14, 19, 21, 28, 33] を参考にエンコーダ毎に決定した．日本語 BERT と mBERT では 5×10^{-6}，XLM-RoBERTa-large には 3×10^{-6} を用いた．また，すべてのモデルはバッチサイズ 128 で 10 万 steps の学習を行った．クエリエンコーダの入力長は 32 トークン，文書エンコーダの入力長は 180 トークンとした．

また，比較手法として多くの研究で利用されている検索，埋め込みモデルを利用する．mContriever [9]：mContriever は mBERT に追加で検索特化の事前学習と hard negative mining による学習行ったデュアルエンコーダのモデルであり，実験では検索特化の事前学習の効果を確認するために利用することができる．multilingual-e5-large（E5）[25]：E5 は XLM-RoBERTa をエンコーダとしたデュアルエンコーダで，検索特化の事前学習，強力なモデルによる hard negative mining，知識蒸留を行っている．さらに，実験で用いる Mr.TyDi や MIRACL の学習データも学習しているため Zero-Shot 設定ではない．そのため，E5 は多言語モデルを用いた，シングルベクトルデュアルエンコーダの精度の上限として解釈することができる．text-embedding-3-large[6]：参考手法の 1 つとして OpenAI が提供する embedding 用の API を用いた．

4.3 実験 1：日本語文書検索における最適な設定の調査

最初に，**RQ1**：「適切な基盤モデルと学習データの組み合わせは何か」，**RQ2**：「基盤モデルのサイズと検索モデルの構造によって検索精度はどのように変化するのか」についての検証する実験を行った．表 2 は学習データや基盤モデルが異なるケースにおける実験結果である．下線は (2)-(4)，(7)-(9)，(12)-(14) それぞれの結果の中で最も高い値であることを表し，**太字**は全体の中で最も高い値を表す検索モデルの構造に SPLADE を用いて mBERT をエンコーダとしたケースにおいて結果が著しく低いため，以降の議論ではこの結果は考慮しない．この結果の原因として mBERT の日本語語彙では単語一致検索が難しいこと，英語の学習データにより日本語語彙の単語一致検索を学習するのが難しいことが考えられる．言語横断情報検索において，SPLADE のエンコーダとして mBERT を用いて語彙を制限する手法を提案している論文 [20] においては教師モデルからの蒸留を行ったとしても期待される精度を達成できておらず，多言語モデルで SPLADE を実装するためにはこれらの課題を解決するような工夫が必要である．

RQ1 について，すべての検索モデルにおいて日本語モデル（J-BERT）を日本語翻訳データ（J-MARCO）で学習する方法

2：https://microsoft.github.io/msmarco/Datasets.html

3：https://huggingface.co/tohoku-nlp/bert-base-japanese-v2
4：https://huggingface.co/bert-base-multilingual-cased
5：https://huggingface.co/xlm-roberta-large
6：https://platform.openai.com/docs/guides/embeddings

表 2　学習データや言語モデルが異なるケースにおける実験結果

		NTCIR2		Mr.TyDi		MIRACL	
	パラメータ数	NDCG@10	Recall@1000	NDCG@10	Recall@1000	NDCG@10	Recall@1000
(1) BM25		0.516	0.533	0.225	0.857	0.369	0.931
DPR							
(2) J-BERT + J-MARCO	111M	0.433	0.392	0.401	0.934	0.476	0.955
(3) mBERT + MS MARCO	179M	0.340	0.320	0.316	0.867	0.422	0.913
(4) mBERT + J-MARCO	179M	0.309	0.325	0.379	0.894	0.441	0.928
(5) XLM-R + MS MARCO	559M	0.451	0.426	0.390	0.917	0.460	0.957
(6) XLM-R + J-MARCO	559M	0.435	0.419	0.401	0.920	0.489	0.951
ColBERT							
(7) J-BERT + J-MARCO	111M	0.508	0.437	0.497	0.943	0.594	0.974
(8) mBERT + MS MARCO	179M	0.486	0.414	0.418	0.911	0.521	0.972
(9) mBERT + J-MARCO	179M	0.455	0.379	0.455	0.908	0.560	0.958
(10) XLM-R + MS MARCO	559M	**0.563**	0.497	**0.536**	**0.958**	**0.630**	**0.990**
(11) XLM-R + J-MARCO	559M	0.506	0.461	0.520	0.941	0.607	0.981
SPLADE							
(12) J-BERT + J-MARCO	111M	0.477	0.429	0.340	0.889	0.405	0.931
(13) mBERT + MS MARCO	179M	0.173	0.248	0.098	0.591	0.113	0.585
(14) mBERT + J-MARCO	179M	0.369	0.337	0.205	0.780	0.241	0.863
参考（）内はエンコーダとして用いる言語モデル							
(15) mContriever(mBERT)	179M	0.479	0.440	0.357	0.932	0.428	0.971
(16) multilingual-e5-large(XLM-R)	559M	0.498	0.489	0.549⋆	0.896⋆	0.694⋆	0.982⋆
(17) text-embedding-3-large		0.478	**0.557**	-	-	-	-

が一貫して最も良い結果となった．これは実験結果の (2)-(4), (7)-(9), (12)-(14) を比較することで確認できる．多言語モデル（mBERT）をエンコーダとする場合には，NTCIR2 では MS MARCO，それ以外のデータセットでは J-MARCO を学習したモデルが良い結果であることが分かる．これらのことから RQ1 に対する回答は次のように考えられる．日本語文書検索においてはエンコーダとして日本語の言語モデルを用いて，日本語翻訳データを学習する方法が最も最適な組み合わせである．一方で，仮に多言語モデルをエンコーダとして利用する場合には英語データ MS MARCO と J-MARCO のどちらを用いるかはデータセットや検索モデルなどによって異なる．

RQ2 について，モデルのサイズ（パラメータ数）や検索モデル毎の実験結果を比較することで考察を述べる．パラメータ数が小さいモデル（111M，179M）と大きいモデル（559M）に関して比較を行うと，全体として結果が良くなっていることが確認できる．検索モデルでは ColBERT がすべてのデータセットで一貫して最も良い結果となっている．また，参考に記載している (15)(16)(17) のモデルと比較しても比較的高い精度である．ColBERT は実験で用いた検索モデルの中で唯一マルチベクトルで類似度計算を行っているモデルであり，計算コストが高い一方で非常に強力であることがわかる．SPLADE は論文を検索対象としている特殊なドメインの NTCIR2 においてのみ DPR を上回る結果となった．特殊なドメインにおいて疎検索が密検索（DPR）を上回るというのは英語においても同様な結果が確認できる [24]．同様に論文検索のような偏りのあるドメインにおいて強力である BM25 を上回るのは難しいこともわかる．その一方で，NTCIR2 のデータセットは LLM による検索が登場する前に作成されたテストコレクションであるために，LLM を用いた検索手法が上位にランキングする文書に適合性判定がされていないことでモデルが過小評価されている可能性があることも考えられる．

学習の設定に関して既存の研究との比較を述べる．(2)(3)(4) と (15) を比較すると mBERT をエンコーダとして学習した場合が mContriever を上回るケースがあることが確認できる．これは Zhang らの研究 [33] においても同様の結果が示されており，この結果を支持する結果である．さらに，日本語 BERT を用いることで同様に mContriever を上回るケースがあることも明らかとなったがこれは新たな発見である．mContriever は検索特化の事前学習を解いた上で MS MARCO による学習をしているにも関わらずこの結果となるのは注目すべき結果であると言える．一方で NTCIR2 では mContriever が DPR を大きく上回っており，事前学習の効果が確認できる．mContriever [9] が Zero-Shot な設定での検索精度を向上させるための手法であるため，偏ったドメインの NTCIR2 で有効に作用したと考えられる．

4.4　実験 2：翻訳データのフィルタリングにより学習を改善できるか

実験 2 では **RQ3**：「機械翻訳したデータを学習に用いる場合，データをフィルタリングすることで学習結果を改善できるか」についての実験を行う．4.3 の実験結果では日本語のモデルを用いて，日本語翻訳データを学習する方法が一貫して高い

表 3　学習データをフィルタリングした場合の結果

	NTCIR2		Mr.TyDi		MIRACL	
	NDCG@10	Recall@1000	NDCG@10	Recall@1000	NDCG@10	Recall@1000
DPR						
(a) J-BERT + JMARCO	0.433	0.392	0.401	0.934	0.476	0.955
(b) + JMARCO-filter	**0.494**	**0.409**	**0.417**	0.929	**0.507**	0.950
(c) + JMARCO-逆 filter	0.390	0.382	0.397	0.926	0.464	0.953
(d) mBERT + JMARCO	0.309	0.325	0.379	0.894	0.441	0.928
(e) + JMARCO-filter	**0.351**	0.323	0.359	0.882	0.434	**0.932**
(f) + JMARCO-逆 filter	0.333	0.309	0.356	0.882	0.427	0.923
ColBERT						
(g) J-BERT + JMARCO	0.508	0.437	0.497	0.943	0.594	0.974
(h) + JMARCO-filter	**0.553**	**0.442**	0.487	**0.947**	**0.604**	**0.980**
(i) + JMARCO-逆 filter	0.461	0.393	0.440	0.919	0.531	0.948
(j) mBERT + JMARCO	0.455	0.379	0.455	0.908	0.560	0.958
(k) + JMARCO-filter	**0.477**	**0.393**	0.443	**0.924**	**0.572**	**0.966**
(l) + JMARCO-逆 filter	0.388	0.307	0.381	0.886	0.472	0.931

クエリ	パッセージ
一般的なものは何ですか（what is the general）	将軍は米国陸軍の 28 位であり、中将より上で、元帥のすぐ下にランクされています...（General is the 28th rank in the United States Army , ranking above Lieutenant General and directly below General of the Army...）
何が影響するのか（what does affect mean）	最も一般的な使用法では、感情は動詞です。何かに影響を与えることは、何かに影響を与えることです... （In its most common usage, affect is a verb; to affect something is to make an impact on something...）

表 4　J-MARCO における翻訳の失敗例（）内は翻訳前の文章

表 5　翻訳前後での検索結果への影響

	MRR@10	Recall@1000
MS MARCO	0.187	0.857
J-MARCO	0.141	0.711
検索結果の重複割合の平均	33.8%	

結果であり，多言語モデルの場合にはどちらのデータを用いるのかが設定によって異なることがわかった．直感的には日本語の文書検索を行うためは日本語データで学習する場合の方が他の言語のデータで学習するよりも有効であると推測できるが，そうでないケースが存在するのは日本語翻訳データに翻訳の質が悪いデータが存在していて，学習を妨げている可能性が考えられる．この推測を定量的に観察するために，翻訳前のデータ（MS MARCO）と翻訳後のデータ（J-MARCO）で検索実験を行う．仮にクエリと文書の適切な翻訳が行われていれば，検索結果や精度には大きな影響が見られないはずである．表 5 はデータセット翻訳前後での BM25 による検索結果である．検索結果の重複割合には式 8 を用いている．表から，検索の精度が下がっているだけでなく本来検索される文書自体も 66%が検索できていないことがわかり，翻訳の失敗により元の文の意味が保持できていない可能性が示唆される．次に検索に失敗したクエリ，適合文書の翻訳の失敗例と元の文を表 4 に示す．1 つ目の例ではクエリ内の単語 “general”の翻訳に失敗していることが検索に影響している．検索クエリは一般的に短く，十分な文脈が与えられることは少ないため，このような事例が多く起きると考えられる．2 つ目の例ではクエリのみでなく，文書側の翻訳にも問題があり，元の文の意味が保持されていないことが分かる．

そこで我々は翻訳されたデータセットから，翻訳が成功している可能性の高いクエリをフィルタリングする手法を提案する．文書のフィルタリングはクエリの翻訳がうまくいっていることが前提であり，学習データとしての文書の選択方法は数多く研究されているため，今回は特にクエリのフィルタリングに注目する．今回の実験では学習データとして用いた triples.train.small からクエリのフィルタリングを行うために 2 つの基準を提案する．

- BM25 によって適合文書が topK 以内に検索できる
- 翻訳前と翻訳後の BM25 での検索結果が類似している

1 つ目の基準は適合文書が最低限の翻訳を保っていること，クエリが適合文書を検索できるようなクエリであることを保証するために設定する．2 つ目の基準は適合文書以外の文書に関しても翻訳前と同様に検索できていれば，クエリがより正確に翻訳ができている可能性が高いと考えて設定する．検索結果を基準としたフィルタリングの効果を確認するために，基準に基づいてフィルタリングしたデータで検索モデルを学習し，精度を比較する．フィルタリングの基準の topK は $K = 200$ とする．また，検索結果の類似度には翻訳前におけるあるクエリ q，$\hat{q}$ に対する検索結果に含まれる文書 ID の集合を R_q，$R_{\hat{q}}$ とする時，そ

の集合を $R_q = \{d_{q1}, d_{q2}, ...d_{q1000}\}$，$R_{\hat{q}} = \{d_{\hat{q}1}, d_{\hat{q}2}, ...d_{\hat{q}1000}\}$ とすると類似度を以下のように表す：

$$sim(R_q, R_{\hat{q}}) = \frac{|R_q \cap R_{\hat{q}}|}{\max(|R_q|, |R_{\hat{q}}|)} \quad (8)$$

式 8 は 2 つのランキングの文書 ID が順序に関わらずどれだけ重複しているかを表す．また，ランキングには 1000 文書を検索していないケースが存在するため，割合を算出するためにどちらかの集合の大きさの最大値で割る．この時，類似度の閾値を 0.35 とする．表 3 に実験結果を示す．図の (b)(e)(h)(k) はフィルタリングしたデータに基づいた学習結果であり，改善したものを**太字**で表している．多くのケースで性能の改善が見られ，特に大きい改善はデータセットは実験 1 において mBERT の学習データとして MS MARCO の方が効果が高かった NTCIR2 で確認できる．さらに追加の実験として，基準を満たさないデータのみを残す逆フィルタリングを行ったデータで学習を行った結果が (c)(f)(i)(l) である．ほとんどのケースにおいて結果が下がっていることが確認できる．これら 2 つの結果から翻訳したデータには学習において有効ではないクエリが含まれている可能性があると示唆され，我々の提案したフィルタリング基準がそのクエリを取り除くことができると考えられる．提案した基準は非常に単純な方法であるため，より詳細な基準によりクエリをフィルタリングした後で文書についても基準を設けてフィルタリングを行うことでさらに効果的なデータが生成できると考えられる．

5 ま と め

本研究では LLM を利用した日本語の文書検索において，基盤モデルや学習データをどう用いるのが適切であるのか，モデルサイズや検索モデルの違いによってどのような影響が生じるのかをいくつかのテストコレクション評価することで網羅的に調査した．実験より，LLM に基づく検索においては日本語モデルをエンコーダとして日本語翻訳データを学習する方法が最も効果的であることがわかった．文章をマルチベクトルで表現しているため計算コストやインデクシング時のメモリの増大があることに注意する必要があるが，検索モデルの構造としては ColBERT が有効であることがわかった．特に ColBERT のエンコーダとして XLM-RoBERTa-large を用いた場合には，Zero-Shot な設定にも関わらず既存の最先端なモデルに匹敵する精度を達成した．また，実験結果と実データの分析から日本語翻訳データの中に学習に寄与していないクエリがある可能性を考察した．そこで，学習に使用する翻訳データについて検索結果に基づいたフィルタリングを行うことで質の悪いクエリを取り除く手法を提案し，学習結果を改善できるケースがあること，学習に寄与していないクエリがある可能性があることを明らかにした．え今後の課題としては日本語翻訳データのより詳細なフィルタリング基準，手法の検討，日本語で検索特化の事前学習を行うモデルの検討，データセットの種類を増やして様々なドメインに関する実験を行うことが挙げられる．

謝辞 本研究は JSPS 科研費 JP22H03905，JP23H03400 の助成を受けたものです．ここに記して謝意を表します．

文 献

[1] Kenya Abe, Kota Usuha, and Makoto P Kato. Kasys at the ntcir-17 transfer task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, 2023.

[2] Benjamin Clavié. Jacolbert and hard negatives, towards better japanese-first embeddings for retrieval: Early technical report. *arXiv preprint arXiv:2312.16144*, 2023.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[4] Zhuyun Dai and Jamie Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*, 2019.

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

[6] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021.

[7] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

[8] Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253*, 2021.

[9] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[11] Hideo Joho, Atsushi Keyaki, and Yuki Oba. Overview of the ntcir-17 transfer task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, 2023.

[12] Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. Overview of japanese and english imformation retrieval tasks (jeir) at the second ntcir workshop. In *NTCIR*, 2001.

[13] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[14] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[15] Carlos Lassance, Maroua Maachou, Joohee Park, and Stéphane Clinchant. Learned token pruning in contextualized late interaction over bert (colbert). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2232–2236, 2022.

[16] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, 2021.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[18] Joel Mackenzie, Shengyao Zhuang, and Guido Zuccon. Exploring the representation power of splade models. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 143–147, 2023.

[19] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer, 2022.

[20] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W Oard. Learning a sparse representation model for neural clir. *Design of Experimental Search and Information REtrieval Systems (DESIRES)*, 2022.

[21] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W Oard. Blade: Combining vocabulary pruning and intermediate pretraining for scaleable neural clir. 2023.

[22] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*, 2016.

[23] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*, 2021.

[24] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[25] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[26] Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. Reproducibility, replicability, and insights into dense multi-representation retrieval models: from colbert to col. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2552–2561, 2023.

[27] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

[28] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W Oard. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2512, 2022.

[29] Jheng-Hong Yang, Xueguang Ma, and Jimmy Lin. Sparsifying sparse representations for passage retrieval by top-k masking. *arXiv preprint arXiv:2112.09628*, 2021.

[30] Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. Neural retrieval for question answering with cross-attention supervised data augmentation. *arXiv preprint arXiv:2009.13815*, 2020.

[31] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512, 2021.

[32] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*, 2021.

[33] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward best practices for training multilingual dense retrieval models. *ACM Transactions on Information Systems*, 42(2):1–33, 2023.

[34] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023.

[35] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*, 2022.

文書への追加情報に対するニューラル検索モデルの挙動分析

薄羽　皐太† 　加藤　誠†† 　藤田　澄男†††

† 筑波大学 大学院人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
††† LINE ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3
E-mail: †k-ush@klis.tsukuba.ac.jp, ††mpkato@acm.org, †††sufujita@lycorp.co.jp

あらまし　我々は大規模言語モデルに基づく情報検索モデルの振る舞いに関する分析を行った．追加情報に対するスコア付けに焦点を当て，文書に一文を加えることが文書のランクを不当に低下させる現象を発見した．この現象を OPEX とし，複数の大規模言語モデルに基づく情報検索モデルを評価した．その結果，ほとんどの大規模言語モデルに基づく情報検索モデルで OPEX が観察され，特に追加された文が長く，文書の始めの方にあり，文書の他の文と類似している場合に多く発生することが判明した．OPEX を軽減するために，文書をセグメントに分割し，そのセグメントのスコアを集計して全体の文書スコアを計算するウィンドウベースのスコアリングアプローチを適応し，効果を検証した．検証したウィンドウベースのスコアリングアプローチは OPEX を軽減し，場合によっては検索性能を向上させることが判明した．

キーワード　検索モデル, LLM, Beyond Accuracy

1　はじめに

文書検索タスクは，ユーザから与えられたクエリに対して，適合度が高い順に文書を並び替えるタスクである．近年の深層学習の発展を受け，情報検索分野でも深層学習モデルを用いた情報検索モデルが発展し，大規模言語モデルをベースとした検索モデルが登場しており，高い性能であることが報告されている．

検索モデルの実アプリケーションへの適応を考える上で，検索モデルの振る舞いが妥当であるか検証することは重要である．伝統的な lexical なマッチングをする検索モデルと比較して，大規模言語モデルに基づく検索モデルはそのパラメータ数の膨大さから振る舞いを予測することが難しい．大規模言語モデルに基づく検索モデルの振る舞いに関してさまざま分析が進められている [1–4]．例えば，BERT をベースとした検索モデルを対象に，情報検索の公理を用いて振る舞いの分析を行った研究 [5] や，クエリに誤字が含まれていた場合の，大規模言語モデルに基づく検索モデルのスコア計算の振る舞いを調査した研究が存在する．しかし，文書に対するどのような操作が文書のスコアに影響を及ぼすかは完全には分かっていない．もし文書の内容への多少の変更が検索結果に大きく影響してしまうと，検索結果の予測可能性を損ない，ユーザや検索サービスの運用者に取って好ましくない振る舞いとなる可能性が考えられる．

我々も，大規模言語モデルに基づく検索モデルのスコアリングに関する分析を行った．その結果，大規模言語モデルに基づく検索モデルでは，文書に自然な 1 文を追加した際に，その文書の順位が大きく降下してしまうケースが存在することを発見した．大規模言語モデルに基づく検索モデルでは，自然な 1 文の追加であっても，BM25 と比較して，大きい順位降下がしやすくなっている．文書長が大きくなることでスコアが減衰することは，文書長の正規化の観点や IR の伝統的な axiom から受け入れられる [6]．1 文伸びた分の文書長に対してペナルティがつくことは想定通りであり，実際に BM25 では多少の順位の低下が多くなっていることが確認された．1 文を文書に追加することは，ユーザが web サイトに情報を加筆したり，ニュース記事において続報が書き込まれた時など，現実的に起こり得るが，それによって文書の順位が大きく下がってしまうことは over-penalization であると言える．そのため，短いテキストの追加は，元々の文書の情報を損なわないため，大規模言語モデルに基づく検索モデルでは，大きな順位低下が BM25 より多く見られることは，検索モデルとしては期待しない振る舞いであると言える．このような，通常想定される文書長に応じたスコアの減衰以上のスコアの低下が引き起こされている現象を **O**ver **P**enalization for **EX**tra Information (OPEX) とする．

本論文では，この結果を踏まえて，大規模言語モデルに基づく検索モデルにおいて，文書に小さな変更を加えた際に，大きく順位を落とすような現象を調査する．文書に 1 文だけを追加し，追加された文書の順位付けを再度行って順位の変化を観測する．この時，加えた 1 文がクエリに対して適合していれば，文書の順位は上昇すると思われる．そうでなければ，元の文書の情報は変わっていないため，文書のスコアには大きく影響しないことが期待される．ただし，1 文増加した分の文書長に対するペナルティによって，スコアの多少の減少は許容される．検索モデルとして，近年提案された大規模言語モデルに基づく検索モデルを取り上げ，メジャーな文書検索データセットを用いて評価した．その結果，大規模言語モデルに基づく検索モデルでは，BM25 と比較して，1 文追加によって大きく順位が変化する割合が高いことがわかった．

この OPEX を軽減する手法として，文ごとにスコア計算を行う方法が考えられる．OPEX が 1 文の追加による入力テキストの変化によって生じていると考えると，文のスコア計算を独立して行うことで，理論的に OPEX を無くせることが示せる．そのため，1 文ごとにクエリとのスコアを計算し，得られた各スコアの最大値を文書全体のスコアとして計算する．これにより，文書全体のスコアは，文書の局所的な類似度に強く影響されるようになり，追加される 1 文の影響を受けづらくなると考えられる．我々は，単一の文ごとに分割してスコアを計算する極端なケースでは，OPEX が完全に抑制されることを理論的に示し，実験により効果を検証した．OPEX は抑制されたが，1 文ごとのスコア計算は OPEX を抑制し，さらに文書長の増加に対してスコアが反応しなくなる結果を引き起こした．文書長の増加に対して反応しないことは，スコア計算において長文の文書がスコアリングで有利になる可能性が考えられる．1 文の追加に対して敏感すぎず，鈍感すぎない範囲を探るために，我々は Window を持って複数の文でまとめてスコアを計算する方法を用いて同様の実験を行った．Window サイズが 1 の場合は 1 文ごとのスコアリングと同等になり，Window サイズが大きくなるほど，文書全体を入力する通常の手法に近づく．その結果，Window サイズを大きくするごとに，OPEX が見られるようになり，中間的な Window サイズでは，中間的な OPEX の抑制と，1 文追加によるスコアへの影響が得られた．また，検索性能についても改善が見られ，一部の Window サイズでは，元の手法を上回る検索性能が観測された．

我々は，上記の調査を遂行する上で，以下の 4 つのリサーチクエスチョン（RQ）を設定した．

RQ1　どの検索モデルで OPEX が見られるか

RQ2　どのようなケースだと OPEX が起こりやすいか

RQ3　1 文ごとに適合度を計算することで OPEX は改善されるか

RQ4　Window を増やした時に OPEX にどのような影響があるか

本論文の貢献は以下のとおりである．

（1） 近年の大規模言語モデルに基づく検索モデルにおいて，通常想定される文書長に応じたスコアの減衰以上のスコアの低下が引き起こされている現象（OPEX）を発見した (Section 3).

（2） OPEX を抑制するシンプルな手法を考案し (Section 4)，実験によりその有効性を確認した．その結果，この手法によって強力に OPEX が抑制できることが明らかとなった (Section 5).

（3） 考案した手法を，OPEX の抑制と文書長への敏感さと検索性能を両立させる手法に拡張した (Section 4)．拡張した手法では，OPEX の抑制と文書長の増加に対する感度のトレードオフが成り立ち，一部では性能の低下も改善が得られた (Section 5).

2 関 連 研 究

最初に，検索モデルのスコアリングの振る舞いを分析するという観点で，深層学習による検索モデルの振る舞いを調査した先行研究を紹介する．次に，文書にテキストを追加して順位の変化を観測するという観点で，LLM ベースの検索モデルに対する敵対的攻撃を行った研究を紹介する．

2.1 検索モデルの振る舞い

大規模言語モデル（Large Language Model, LLM）に基づく検索モデルは，情報検索タスクで BM25 に代表される伝統的な検索手法を上回る性能を発揮し，近年注目を浴びている．

一方で，このような LLM ベースの検索モデルのスコアリングの振る舞いは十分にわかっているとは言えず，LLM ベースの検索モデルの振る舞いの調査が行われている [2–4, 7, 8]．このような研究として，初期には，BERT の Attention に着目して調査を行った研究がある [2, 8]．Qiao らは，既存の検索モデルとのスコア分布の違いや，BERT レイヤーごとに入力トークンへの Attention を分析するなど，BERT のスコアリングに関する振る舞いを調査した．

Diagnosing Dataset を用いた，情報検索の公理と検索モデルの関連も調査されている [1, 5, 6, 9]．情報検索では，クエリ語を加えた文書はスコアが上昇するべきである等の経験則が知られており，BM25 などの伝統的な検索モデルはこのような経験則に基づいている．Rennings らは，よく知られた経験則を公理して定式化し，深層学習ベースの検索モデルがこのような経験則に基づいているかどうか調査した [9]．また，情報検索タスクのデータセットでファインチューニングした BERT を対象に，Diagnosing Dataset を用いて振る舞いの分析を行った研究も存在する [5]．彼らも同様に定式化された公理に基づいた Diagnosing Dataset を作成し，伝統的な公理では BERT の振る舞いを説明できないことを明らかにした．これを受け，伝統的な公理だけでなく，多様な観点で LLM ベースの検索モデルに焦点を当て分析を行った研究も存在する [1].

これらの研究では，LLM ベースの検索モデルについて，クエリの変更に対する振る舞いや，文書の属性や伝統的な公理に焦点を当てた分析が行われている．しかし，文書に自然な変更を加えた時の，大きな順位変化に着目した研究は存在しない．文書にテキストを加えたことで，文書長の増加分に対するスコアリングのペナルティは妥当である [6]．そのため，自然で小さなテキストの追加であれば，文書が内包する情報に損失がないため，順位が大きく降下することは望ましくない．我々は文書に自然な 1 文を加えた際に，順位降下が起きるケースに焦点を当て，LLM ベースの検索モデルの分析を行う．

2.2 検索モデルに対する敵対的な攻撃

深層学習モデルに対する敵対的攻撃は，画像処理の分野から広まり，自然言語処理分野でも研究が進められている．情報検索モデルに対する攻撃を行った研究もあり，文書になるべく小さく自然な摂動を加えることで，文書の順位を操作する試みが

行われている [10,10–12]．Wang らは，アドホック検索モデルを対象に，文書の順位を向上させる攻撃を行い，BERT ベースの検索モデルの脆弱性を示した [13]．Liu らは流暢になるように制約を加えた敵対的サンプル生成手法を提案し，自然な単語列の追加によって既存の攻撃手法より効果的に順位を上昇させることが可能であることを示した [11]．Wu らは文書中からスコア計算において重要な単語を特定し，その単語を類似する別の単語に置き換えることによって，文書の順位を上昇させる手法提案した [10]．

これらの研究では，LLM ベース検索モデルに対して，テキストの追加や置き換えによって文書スコアを上昇もしくは降下させる攻撃を行っている．我々の研究は，テキストを追加した時に，特に大きく順位が降下するケースに着目するという観点で，これらの研究と異なる．追加したテキストは，追加した後の文書が自然であるようにするために，自然な 1 文に限定し，1 文を追加することによって大きく順位が降下する現象を調査する．

3 分　　析

この節では，大規模言語モデルに基づく検索モデルにおいて，文書に小さな変更を加えた際に，大きく順位を落とすような現象を調査する方法を説明する．

3.1 分析設定

3.1.1 データセット

我々は，文書中の 1 文を削除もしくは追加することで分析を行う．この時，1 文追加の影響が文書に対して大きな変更になりすぎないために，コーパス中の文書の文書長が長い文書検索タスク向けのデータセット Robust04 [14] と MS MARCO Document [15]，TREC Deep Learning 2019 and 2020 [16,17] を用いる．

3.1.2 検索モデル

実験に用いる検索モデルは，伝統的な lexical model として BM25 と，近年注目を浴びている大規模言語モデルに基づく検索モデルを用いる．検索モデルとして，Dense Retriever (DR) と Learned Sparse Retriever (LSR) の 2 種類のモデルを比較する．DR として ANCE と ColBERT を用いる [18]．ANCE は文書とクエリを 1 つの密ベクトルにエンコードし，それぞれのベクトルの内積から類似度を計算する [18]．ANCE の学習では，通常の学習中のモデルによる負例のサンプリングを行った対照学習でモデルのファインチューニングを行っている．ColBERT では，文書とクエリの各単語に着目し，単語の埋め込み表現の類似度を集約することで，文書のスコアを計算する [19]．LSR ではクエリと文書を語彙数次元の疎ベクトルにエンコードし，スコアの計算を行う．LSR として，DeepCT と SPLADE を用いる．DeepCT は，文書中に含まれる語の重み付けを行い，term frequency に mapping することで従来の lexical matching 手法を用いた検索と両立している [20]．SPLADE は，BERT の MLM ヘッドを用いて文書拡張と同時に重みを推定し，転置インデックスに格納する [21]．検索時にはクエリの拡張を行い，term frequency に基づいてスコア計算を行う．我々は DPR を用いた実験も行ったが，ANCE と類似したアーキテクチャを持つ検索モデルであり，性能が ANCE より低かったため省略する [22]．これらの検索モデルは MS MARCO Passage Dataset の訓練データを用いて学習され，広く用いられている FirstP を用いる [23]．

3.2 分析方法

1 文の追加の実験を行う上で，文書に対する変化はなるべく自然にしたい．そのため，ここでは 1 文を単位として変化を加える．情報検索タスクでは，クエリ q とコレクション D が与えられた時，情報検索モデル f を用いて各文書 $d_i \in D$ の q に対する適合度を示すスコア $s_i = f(q, d_i)$ を推定する．スコアは，文書の順位付けに用いられ，各文書に順位が割り当てられる．D 内の文書がスコアの降順に並べられた時の d_i の順位を $r(d_i, D)$ と定義する．また，L を上位の文書，すなわち，$L = \{d \mid d \in D \wedge r(d_i, D) \leqq N\}$ とする．ここで N は cutoff である．情報検索の一般的なアプリケーションでは，L が検索結果として返される．

先行研究に従い，我々は文書 d_i に小さな変更が加えられた後の L での d_i の順位変化に基づいて情報検索モデル f を評価する [11]．d_i' を内容が変更された文書 d_i とするとき，順位変化を $\mathrm{RS}(d_i, d_i') = r(d_i', L') - r(d_i, L)$ と定義する．ここで $L' = L \setminus \{d_i\} \cup \{d_i'\}$ である．

文が d_i に挿入されて d_i' が得られたと仮定する．順位変化が負の場合，情報検索モデル f は d_i' が与えられたクエリ q に対して d_i よりも適合していると推定する．これは，追加された文が与えられたクエリ q に適合しており，我々の研究ではこの場合は扱わない．ポジティブで小さい順位変化（例えば，1 から 50）は，追加された文がクエリに適合しない場合に BM25 でよく観察される．これは，BM25 が明示的に文書の長さに基づいてスコアを正規化し，長い文書にペナルティを課していることに起因する．一方で，ポジティブで大きな順位変化（例えば，$\geqq 50$）は過度で不合理である．追加された文が無関係であっても，それは文書の冗長性を増すだけであり，ペナルティはその冗長性の増加分のみに限定されるべきである．そのため，順位変化は BM25 のそれよりも大きくなることは期待されない．

BM25 は提案されてから広く用いられており，文書長による正規化を明示的に行なっており，文書長の長い文書に対してはペナルティが課される [24]．この性質は IR における文書長に関する伝統的な axiom と合致しており， BM25 を今回の OPEX の基準として定める [6]．順位変化の頻度を測定するために，RS_t，すなわち $\mathrm{RS}(d_i, d_i') \geqq t$ が観察されるケースの割合を指標とする．例えば，RS_{100} は順位変化が 100 以上である情報検索モデルのケースの割合を示す．したがって，OPEX を測る指標を，BM25 よりも高い頻度で大きな順位変化が起きる現象として定義し，BM25 と各大規模言語モデルベースの検索モデルの RS_t を計算し，比較する．順位の変化は上昇と降下の両方において発生し得る．そのため 1 文の追加による順位の上

昇と降下の両方を観測するために，データセット内の各クエリに対して，Sentence Deletion Analysis では 100 から 300 位までの，Addition Analysis では 1 位から 200 位までの各文書に対して順位変化 $\mathrm{RS}(d_i, d_i')$ を計算する．

BM25 の文書長の正規化はそのハイパーパラメータに大きく依存している．そのため，我々の分析では，他の研究とは異なり，全てのテストクエリ（MS MARCO Document の dev クエリ）を使用してハイパーパラメータを調整し，理想的な文書長の正規化を得る．この設定は BM25 の検索性能を過大評価する可能性があるが，この研究の目的は検索性能の面で BM25 を上回ることではないため，我々の結論には影響しない．

以下では，文書に加えられた二種類の小さな変更について説明し，その利点と欠点について議論する．１つめの分析では，１文を削除する Sentence Deletion Analysis を行う．この分析では，1 文の削除によって自然な 1 文の追加をシミュレートできるが，1 文を削除した後の文書が自然であるとは限らないという限界がある．２つめの分析では，単純に 1 文を文書に追加して，文書の順位への影響を調査する．一方で，追加する 1 文が文書に対して自然であることは稀であるため，自然な 1 文の追加にはなりづらい．

3.3 Sentence Deletion Analysis

Sentence Deletion Analysis では，文書から一様分布にランダムに 1 文を抜き，1 文抜いた文書を d とし，元の文書を d' として分析を行う．1 文追加の影響を調査する上で，追加する 1 文はなるべく文書の内容に対して自然な 1 文にしたい．不自然な文の追加では，1 文を追加して順位が降下したとしても，それは検索モデルが無関係なテキストであることに反応していたことに由来している可能性があり，ノイズとなってしまうことが考えられる．そこで，分析対象の文書に 1 文を追加する代わりに，文書から 1 文だけ抜いて，順位を再計算する．もし，この 1 文抜いた文書に再度スコア付けを行い，その結果順位が 1 文抜いた文書より高い順位を得た場合，1 文を削除したためにその文書は順位が上昇したと考えられる．裏を返すと，元の文書は，その削除した一文を付け加えることによって順位が降下していたと考えることができる．つまり，1 文削除した文書を d だと考えると，削除した 1 文は a に対応し，元々の文書は d' であるとみなすことができる．そして，追加する 1 文 a は，元々この文書に含まれていたものであるから，d に対してとても自然な 1 文であると言える．このようにして，自然な 1 文の追加をシミュレートし，Sentence Deletion Analysis では，元々の文書を d' とし，1 文削除した文書を d として扱う．元々の文書を d' とするため，1 文の削除によって上昇した文書に着目する．

この分析方法は，文書に含まれるような自然な文の追加をシミュレートすることができるという利点がある．一方で，1 文を削除した後の文書が常に自然であるとは限らないという欠点がある．そこで，我々は 1 文を単純に追加する方法の分析も行った．

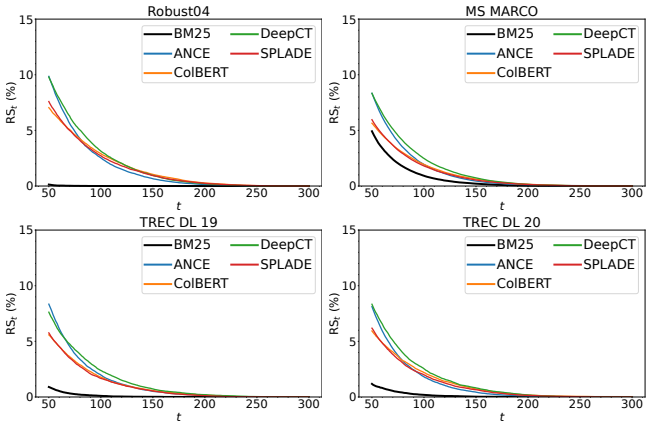


図 1　Sentence Deletion Analysis の結果．横軸が順位変化を表し，縦軸がその順位変化までの累積度数を表している．

3.4 Sentence Addition Analysis

Sentence Deletion Analysis では，1 文の追加を 1 文の削除によってシミュレートするが，単純に 1 文をサンプリングして文書に追加する方法は手続きとしてより自然である．Sentence Intact Analysis では，1 文をサンプリングして，文書中のランダムな位置に 1 文を挿入し，d' を形成する．1 文のサンプリング元はコーパス全体からと同じランキング l の中からサンプリングする方法が考えられる．同じランキング中から 1 文をサンプルする方が，より元の文書に類似した文をサンプリングしやすいと考えられるが，追加する文が自然であるケースは Sentence Deletion Analysis で分析するため，ここではコーパス全体から 1 様分布に 1 文をサンプリングする．

Sentence Deletion Analysis と比較して，1 文を追加して順位の降下を確認するという分析を行ううえで，この分析はこの手続きを直接的に従った方法であると言える．また，追加される前の文書に変更を加えないという利点もある．一方で，1 文はコーパス中から 1 様分布にサンプリングされるため，追加されるテキストが元の文書と関連することはごく稀であると言える．そのため，Sentence Addition Analysis では，無関係な 1 文の追加による影響を調査可能である．

3.5 Analysis Result

3.5.1 RQ1： どの検索モデルで OPEX が見られるか

Sentence Deletion Analysis の結果を Figure 1 に示す．それぞれのデータセットにおいて，$50 \leqq t$ にフォーカスした図であり，横軸が t を表し，縦軸が RS_t を表している．50 以上の順位変化に着目すると，全てのデータセットで BM25 が線グラフ上で最も下に来ている．これは，BM25 が最も 50 以上の順位変化の割合が小さく，大規模言語モデルベースの検索モデルでは BM25 より大きい高い順位変化の割合があったことを示している．大規模言語モデルベースの検索モデルの中では，ColBERT と SPLADE が似た傾向にあり，相対的に大きな順位降下の割合が小さい．

この結果は，表 1 でも確認できる．全ての大規模言語モデルをベースとした検索モデルにおいて，$t = 50, 100, 150$ のい

表 1　Sentence Deletion Analysis の RS_{50}, RS_{100}, RS_{150} (%) の結果．$*$ は，Holm 補正を用いたカイ 2 乗検定において統計的に有意な差があったことを示している．

	t	BM25	ANCE	ColBERT	DeepCT	SPLADE
Robust04	50	0.130	9.85*	7.04*	9.79*	7.59*
	100	0.002	2.53*	2.92*	3.12*	2.72*
	150	0.000	0.656*	1.04*	0.955*	0.942*
MS MARCO	50	4.92	8.33*	5.65*	8.33*	5.96*
	100	0.922	1.94*	1.94*	2.48*	1.81*
	150	0.209	0.448*	0.626*	0.731*	0.539*
DL19	50	0.899	8.34*	5.56*	7.62*	5.76*
	100	0.0964	1.98*	1.78*	2.38*	1.67*
	150	0.0156	0.506*	0.580*	0.726*	0.541*
DL20	50	1.18	8.12*	5.93*	8.33*	6.18*
	100	0.192	1.81*	2.19*	2.54*	1.99*
	150	0.0285	0.436*	0.718*	0.816*	0.651*

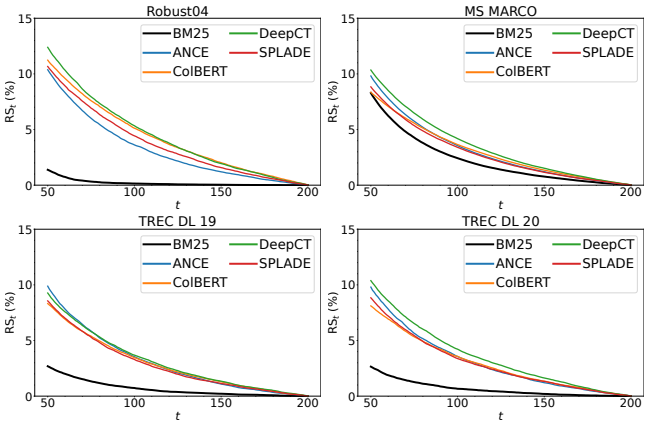


図 2　Sentence Addition Analysis の結果．横軸が順位変化を表し，縦軸がその順位変化までの累積度数を表している．つまり，全体の中でその順位変化までが占める割合を意味している．(top) Robust04 における結果．(bottom) MS MARCO での結果．

ずれにおいても統計的に有意に BM25 より大きい RS_t であったことがわかる．特に Robust04 では，100 位以上の順位変化は，BM25 では 0.002%以下であるが，大規模言語モデルに基づく検索モデルでは最も少ない ANCE でも 2.53%以上の割合で 100 以上の順位変化が起きている．これは，ANCE では削除された 1 文を再度文書に追加することで，2%の文書で 100 位以上の順位の降下が起きることを意味している．

Sentence Addition Analysis の結果を Figure 2 に示す．Sentence Deletion Analysis の結果と同様に，それぞれのデータセットにおいて，$50 \leqq t$ にフォーカスした図であり，横軸が t を表し，縦軸が RS_t を表している．100 以上の順位変化の割合は，常に大規模言語モデルに基づく検索モデルが多くなっており，特に Robust04 で顕著である．また，大規模言語モデルベースの検索モデルの中では，ColBERT と SPLADE が似た傾向にあり，相対的に大きな順位降下の割合が小さい．これらの傾向は Sentence Deletion Analysis での結果と一致している．

MS MARCO Document では，$t = 50$ の時は BM25 が線グラフ上で大規模言語モデルベースの検索モデルと同等の大きさになっている．これは BM25 で 50 以上の順位変化が起きた割合が大規模言語モデルベースの検索モデルでの割合と同程度であったことを示している，

表 2　Sentence Deletion Analysis で，文書とそこから削除した 1 文の分析結果．$*$ は unpaired の両側 t 検定において統計的に有意な差があったことを示している（$p < 0.05$）．

Dataset	RS_t	Length	Position	Jaccard	SBERT
Robust04	$RS_{t<100}$	14.5	6.27	0.0494	0.464
	$RS_{t\geqq100}$	18.9*	3.04*	0.0659*	0.552*
MS MARCO	$RS_{t<100}$	13.5	8.03	0.0461	0.461
	$RS_{t\geqq100}$	19.5*	3.52*	0.0629*	0.585*
DL19	$RS_{t<100}$	12.2	8.04	0.0462	0.476
	$RS_{t\geqq100}$	17.2*	3.60*	0.0639*	0.581*
DL20	$RS_{t<100}$	12.1	8.16	0.0452	0.475
	$RS_{t\geqq100}$	16.9*	3.53*	0.0615*	0.583*

これらの結果から，1 文を追加することによって，全体として，検証した大規模言語モデルに基づく検索モデルらでは BM25 以上の順位降下が起きやすくなっていたことがわかる．また，Deletion Analysis で検証した通り，追加される 1 文が，元から文書に含まれていたような自然な文であったとしても，順位降下が起きることが判明した．この 1 文追加による順位降下の傾向は Robust04 で特に顕著であり，MS MARCO ではその傾向は比較的弱かった．

3.5.2　RQ2：どのようなケースで OPEX が起こりやすいか

このセクションでは，これらの順位降下がどのようなケースだと起こりやすくなるのか調査する．

まず，文書に追加された 1 文の長さを調査する．追加される 1 文が長いほど，その文書における変更は大きくなるため，文の追加がスコアリングに与える影響は大きくなると考えられる．追加した文が文書中において，無関係な文であったかどうかを調査するために，追加した文と文書の類似度を計算する．語彙的な類似度には，bag-of-words の積集合の大きさと （Word overwrap) と Jaccard similarity coefficient を用いる．よりセマンティックな類似度として，Sentence-BERT を用い，追加された文と文章をエンコードし，得られたベクトルの類似度を計算する [25]．BERT ベースのモデルでは，文書の先頭に摂動を加えることで，スコアリングへ与える影響が比較的大きいことが報告されている [11, 13]．このことから，1 文が削除された位置が順位降下に影響していた可能性が考えられるため，1 文を削除した位置も調査する．Sentence Deletion Analysis でこの分析を行い，100 以上の順位変化が合ったケースとそれ以外のケースを比較する．

表 2 に，Sentence Deletion Analysis での結果を示す．この表から，両方のデータセットで，追加した 1 文は大きい順位降下の方が長くなっていることがわかり，1 文が長い方が順位の低下に大きな影響を与えることがわかる．また，Sentece position では，大きい順位降下の方が値が小さくなっている．これは，文書中の先頭に近い位置に 1 文を削除する方が，順位が降下しやすくなっていたことを示しており，既存の敵対的攻撃の研究での結果と一貫している [11, 13]．1 文と文書の similarity では，

大きい順位降下の方で Word overwrap と Jaccard similarity coefficient が大きくなっていることがわかる．word overwrap が多くなることは，1 文の文長が長くなっていることに起因している可能性がある．そのため，追加して順位が大きく低下した文では，追加した文書本体と語彙的な一致が多かったと言える．同様に，Sentence-BERT のスコアも大きい順位降下の方が大きくなっており，意味的な類似度でも大きい順位降下の方が大きくなっていたことがわかり，語彙的にも意味的にも，追加する文は文書本体と類似している方が順位の降下に貢献したことを示している．この結果は，Doc2Query を LLM ベースの検索モデルで用いたときにスコアが減少することを発見した先行研究の結果と，文書に関連する情報を追加してスコアが減少するという点で一致している．これらの傾向は，Sentence Addition Analysis でも同様であった．

4 手　　法

この節では，節 3 で確認した OPEX を抑制する方法を提案する．

4.1 Window-based Scoring

我々は OPEX を軽減する手法として，文書をセグメントごとに独立にスコアを計算する方法の適応を提案する．この方法では，文書を一定のセグメントに切り出してセグメントごとにスコアを計算し，各セグメントのスコアの最大値を文書のスコアとする．各セグメントのスコアが独立に計算され，スコアの最大値を取ることにより，文の追加がスコア計算に与える影響を限定することができる．

具体的には，Window-based scoring 手法では，文書 d_i のスコアを次のように計算する．

$$f_{\text{win}}(q, d_i) = \max_j f(q, \Delta_{i,j}) \tag{1}$$

ここで，$\Delta_{i,j}$ は 文書 d_i 中の j 番目のセグメントであり，d_{ij} を d_i 中の j 番目の文とすると，$\Delta_{i,j} = d_{i,j}d_{i,j+1}\dots d_{i,j+w-1}$ で⊠⊠ある．このように，文書を分割して各部分のスコアの最大値を利用する手法は，すでにいくつかの手法で行われている [2, 18, 26, 27]．w を大きくしていくと，スコアの計算において，1 文追加が文書中の他のセグメントに影響するようになり，$w = |d_i|$ の時にベースの検索モデルと同一になる．

この手法を用いることにより，$w = 1$ の時に OPEX が完全に抑制できることを証明できる．この場合，$\Delta_{i,j}$ は d_i の j 番目の文に対応している．つまり，$\Delta_{i,j} = d_{i,j}$ である．文書 d_i の文の数を n とする．文書の最後に余分な文が挿入されていると仮定し，この余分な文を $d_{i,n+1}$ で表す．余分な文が追加された文書 d'_i のスコアが d_i のスコア以上になることは次のように証明できる：

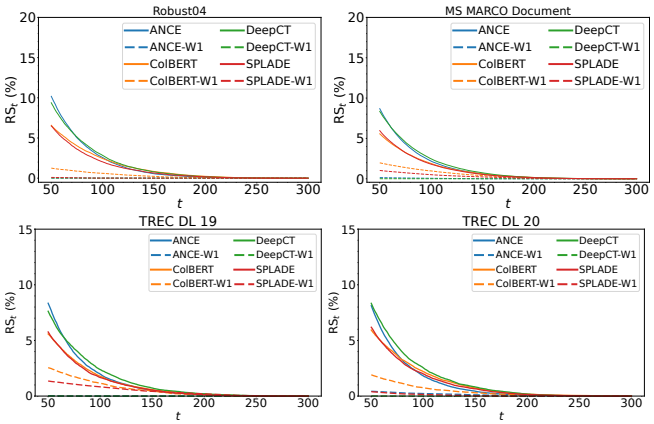


図 3　$w = 1$ で Window-based scoring 手法を適応した大規模言語モデルに基づく検索モデルの Sentence Deletion Analysis の結果．

$$\begin{aligned} f_{\text{win}}(q, d'_i) &= \max_j f(q, \Delta_{i,j}) \\ &= \max\{f(q, d_{i,1}), f(q, d_{i,2}), \dots, f(q, d_{i,n+1})\} \\ &= \max\{\max\{f(q, d_{i,1}), f(q, d_{i,2}), \dots, f(q, d_{i,n})\}, \\ &\quad f(q, d_{i,n+1})\} \\ &= \max\{f_{\text{win}}(q, d_i), f(q, d_{i,n+1})\} \geqq f_{\text{win}}(q, d_i). \end{aligned}$$

したがって，文書に文を追加しても，文書 d_i のランクが低下することはないため，OPEX は起こり得ない．

Window-based scoring では $w = 1$ の場合に OPEX は発生しないが，これは文書のスコアがただ一つの文のみによって決定され，文書長の正規化が行われない極端な設定である．$w = n$ の時は，$f(q, d_i) = f_{\text{win}}(q, d_i)$ となり，$f(q, d_i)$ は $f_{\text{win}}(q, d_i)$ のまた極端なケースであると言える．そのため，我々は文書スコアを計算する際に文脈を考慮し，適切な文書長の正規化を有する適切なウィンドウサイズを 1 から n の間で探索する．ここでは，データセット中の文数を踏まえ， 5, 10, 20 の 3 つを W として設定し，Section 3 で用いた各検索手法に適応する (Figure 3. 1. 1).

5 実 験 結 果

5.1 RQ3: 1 文ごとに適合度を計算することで OPEX は改善されるか

$w = 1$ の時の Window-based scoring を対象に Sentence Deletion Analysis を行った結果を Figure 3 に示す．$W1$ の手法が $w = 1$ の時の Window-based scoring を行った手法であることを示す．この図から，全てのデータセットで，W1 の大規模言語モデルに基づく検索モデルは，元の手法と比較して，大きな順位降下の割合が小さくなっていることがわかる．この結果から，高い順位上昇が全体的に起こりにくくなったことが読み取れる．特に，Robust04 では，ColBERT を除き，Sent によるスコア計算を行った検索モデルは，BM25 と同様に 100 より大きい順位上昇はほとんど見られなくなっている．ANCE-Sent や DeepCT-Sent では，全てのデータセットで 100 より大きい順位上昇がほぼ 0 となっているが，SPLADE-Sent,

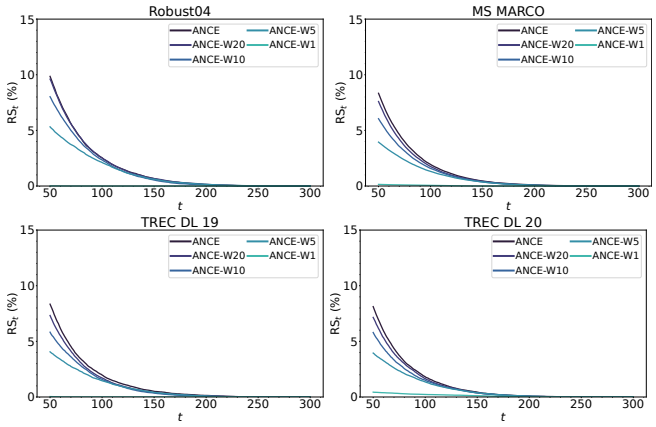


図 4　$w = 5, 10, 20$ のそれぞれで Window-based scoring を行った ANCE の Sentence Deletion Analysis の結果.

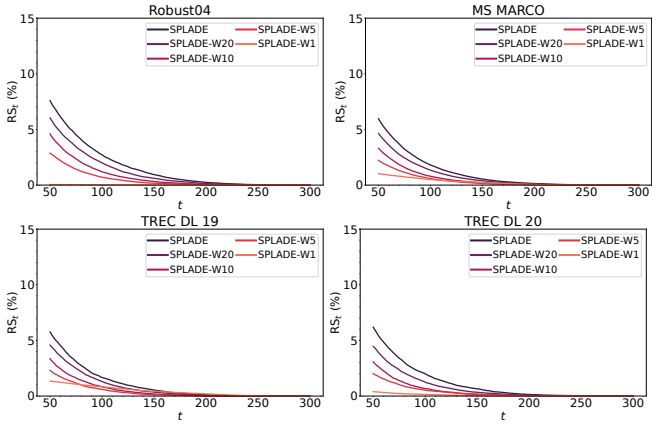


図 5　$w = 5, 10, 20$ のそれぞれで Window-based scoring を行った SPLADER の Sentence Deletion Analysis の結果.

表 3　Sentence Deletion Analysis における Window-based scoring 手法の RS_{100}. $*$ は Holm 補正を行った χ^2 検定 ($p < 0.05$) において，統計的に有意な差があったことを示している.

ANCE	W1	W5	W10	W20	Vanilla	BM25
Robust04	0.002*	2.12*	2.35	2.52	2.53	0.002
MS MARCO	0.056*	1.31*	1.58*	1.76*	1.94	0.922
DL19	0.008*	1.46*	1.58*	1.71*	1.98	0.0964
DL20	0.228*	1.36*	1.51*	1.65	1.81	0.192
SPLADE	W1	W5	W10	W20	Vanilla	BM25
Robust04	0.022*	0.708*	1.19*	1.99*	2.72	0.002
MSMARCO	0.497*	0.589*	0.793*	1.34*	1.81	0.922
DL19	0.829*	0.587*	0.831*	1.30*	1.67	0.0964
DL20	0.148*	0.522*	0.668*	1.26*	1.99	0.192

ColBERT-Sent では僅かに 100 より大きい順位変化が見られる．これは，ベクトルのマッチングに使用される ANN の誤差や文書への操作時に生じた僅かな摂動が影響した可能性が考えられる.

5.2　RQ4: Window を増やした時に OPEX にどのような影響があるか

我々はさらに w を大きくしたときの影響を調査した.

w を 5,10,20 に設定した Window-based scoring を行った手法をそれぞれ W5,W10,W20 とする．複数のウィンドウサイズを用いた実験には膨大な計算コストがかかるため，LSR と DR の検索モデルを 1 つづつ用いて，この実験を行った．SPLADE は今回実験に用いた 4 つの大規模言語モデルに基づく検索手法の中で最も高い性能であったため用いた．DR には ANCE と ColBRET があるが，複数のウィンドウサイズの実験に多くの時間を要したため，ANCE を用いた．ANCE と SPLADE に Window-based scoring 手法を適応し，それぞれのウィンドウサイズでの Sentence Deletion Analysis の結果をそれぞれ Figure 4 と Figure 5 に示す.

表 4　Window-based scoring を適応した時の nDCG@100. $*$ は Holm 補正を行った paired t 検定 ($p < 0.05$) において，統計的に有意な差があったことを示している.

ANCE	W1	W5	W10	W20	Vanilla
Robust04	0.326*	**0.334***	0.322*	0.316	0.317
MS MARCO	0.408*	0.446	0.452	**0.454**	0.452
DL19	0.397	**0.415**	0.414	0.405	0.404
DL20	0.456	**0.480**	0.479	0.476	0.477
SPLADE	W1	W5	W10	W20	Vanilla
Robust04	0.372*	**0.417***	0.412*	0.389	0.407
MS MARCO	0.432*	0.500	0.508	0.512	**0.516**
DL19	0.438*	0.517	**0.522**	0.504	0.504
DL20	0.505*	**0.605***	0.600*	0.578	0.571

ANCE と SPLADE の両方において，全体的な傾向として，ウィンドウサイズが大きくなるほどベースの検索モデルの結果に近づいていることがわかる．Robust04 では，W20 でベースの検索モデルでの結果と概ね重なっており，これ以上 Window を大きくしても ANCE と変化が見られないと考えられる．MS MARCO では W20 でも ANCE-W20 と ANCE とは離れているが，MS MARCO では，20 以下のセンテンス数である文書は約 75%であり，さらなるウィンドウサイズの増加で ANCE に近付いていくと思われる.

5.2.1　検索性能と OPEX の軽減

Window-based Scoring 手法を適応が大規模言語モデルに基づく検索モデルの検索性能へ与える影響は自明ではない．検索モデルの検索性能として nDCG@100 を表 4 に示す．この結果から，一部統計的に有意ではないが，MS MARCO を除きウィンドウサイズ 5 と 10 のモデルでは，ANCE と SPLADE の両方でベースの検索モデルをすべて上回っていることがわかる．ウィンドウサイズが 20 のようにさらに大きくなると，ベースの検索モデルの性能に近付いていき， ウィンドウサイズ 20 ではほぼベースの検索モデルと同等の性能となった．W1 の時では全体としてパフォーマンスの低下が見られ，表 3 から W10 や W20 では OPEX の抑制する効果が弱まっていくことを踏まえると，W5 が適切なウィンドウサイズであると考えられる.

MS MARCO では，特に SPLADE を用いた際に，検索性能の低下する傾向が見られた．これはデータセットの適合ラベルが影響している可能性が考えられる．Robust04 や TREC DL19，20 では適合ラベルが多段階あり，かつ 1 つのクエリに複数個の適合性ラベルがあるが，MS MARCO の適合ラベルには 1 段階のラベルがクエリにつき 1 つである．MS MARCO 以外では W5 や 10 で性能の低下が見られなかった結果を踏ま

えると，Window-based scoring 手法では，文書中の一部が適合であれば文書が適合と判定されることでより細かい適合シグナルを拾えるようになり，広い順位での検索性能の改善に貢献していた可能性が考えられる．

5.2.2 検索速度

インデックスを作成する際，1 文書を処理するのに，通常の検索モデルでは通常 1 度検索モデルに入力すれば計算可能である．Window-based scoring では $\max(1, n-w+1)$ 回検索モデルに文書を入力する必要があり，保存するベクトル数も同じ数だけ増加する．そのため，ウィンドウサイズが十分に大きくない限り，文書のセンテンス数の増加に比例して線形にインデックスにかかる時間とストレージコストが増加する．検索時には，Window ごとに分割された文書を検索することになるため，最終的な検索結果を形成するために，重複した文書 ID は最大値を取得する操作が必要となる．大規模なコレクションでの max を計算する操作は計算コストが大きく難しい課題であるが，この問題は ColBERT で使用されている上位 k 件の文書を検索するための，二段階によるスコアリング手順によって，軽減することが可能である [19]．

また，Window-based scoring のスコアリングの振る舞いを蒸留することができれば，1 度の検索モデルへの入力で Window-based scoring と同じようなスコア計算ができるようになる可能性が考えられるが，本論文では扱わず今後の課題とする [28,29]．

6 まとめ

本論文では，大規模言語モデルに基づく検索モデルの追加情報に対する過剰ペナルティ（OPEX）に関する分析を行った．その結果，多くの大規模言語モデルに基づく検索モデルで OPEX が観測され，特に追加された文が長く，文書の先頭に近く，文書中の他の文と類似している場合に発生することが示された．また，ウィンドウベースのスコアリング手法により，OPEX が軽減され，場合によっては検索効果も向上することが示された．今後の課題として，ウィンドウベースのスコアリング手法の検索速度の改善が考えられる．

謝辞 本研究は JSPS 科研費 JP21H03775，JP22H03905 の助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. ABNIRML: Analyzing the Behavior of Neural IR Models. *Transactions of the Association for Computational Linguistics*, 10:224–239, 2022.

[2] Zhuyun Dai and Jamie Callan. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *SIGIR*, pages 985–988, 2019.

[3] Gustavo Penha, Arthur Câmara, and Claudia Hauff. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *ECIR*, pages 397–412, 2022.

[4] Shengyao Zhuang and Guido Zuccon. CharacterBERT and Self-Teaching for Improving the Robustness of Dense Retrievers on Queries with Typos. In *SIGIR*, pages 1444–1454, 2022.

[5] Arthur Câmara and Claudia Hauff. Diagnosing BERT with Retrieval Heuristics. In *ECIR*, pages 605–618. 2020.

[6] Hui Fang, Tao Tao, and ChengXiang Zhai. A Formal Study of Information Retrieval Heuristics. In *SIGIR*, pages 49–56, 2004.

[7] Negar Arabzadeh, Bhaskar Mitra, and Ebrahim Bagheri. MS MARCO Chameleons: Challenging the MS MARCO Leaderboard with Extremely Obstinate Queries. In *CIKM*, pages 4426–4435, 2021.

[8] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the Behaviors of BERT in Ranking. *arXiv preprint arXiv:1904.07531*, 2019.

[9] Daniël Rennings, Felipe Moraes, and Claudia Hauff. An Axiomatic Approach to Diagnosing Neural IR Models. In *ECIR*, pages 489–503, 2019.

[10] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. PRADA: Practical Black-Box Adversarial Attacks against Neural Ranking Models. *ACM Transactions on Information Systems*, 41(4):1–27, 2023.

[11] Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. Order-Disorder: Imitation Adversarial Attacks for Black-Box Neural Ranking Models. In *CCS*, pages 2025–2039, 2022.

[12] Nisarg Raval and Manisha Verma. One word at a time: adversarial attacks on retrieval models. *arXiv preprint arXiv:2008.02197*, 2020.

[13] Yumeng Wang, Lijun Lyu, and Avishek Anand. BERT Rankers are Brittle: a Study using Adversarial Document Perturbations. In *ICTIR*, pages 115–120, 2022.

[14] Ellen Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *TREC-7*, 2005.

[15] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[16] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2019 deep learning track, 2020.

[17] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track, 2021.

[18] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *ICLR*, 2021.

[19] Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*, pages 39–48, 2020.

[20] Zhuyun Dai and Jamie Callan. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. *arXiv preprint arXiv:1910.10687*, 2019.

[21] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In *SIGIR*, pages 2353–2359, 2022.

[22] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[23] Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. Re-

sources for brewing beir: Reproducible reference models and an official leaderboard, 2023.

[24] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, 1995.

[25] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[26] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. pages 3490–3496. Association for Computational Linguistics, November 2019.

[27] Xinyu Zhang, Andrew Yates, and Jimmy Lin. Comparing Score Aggregation Approaches for Document Retrieval with Pretrained Transformers. In *ECIR*, pages 150–163, 2021.

[28] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

[29] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, pages 163–173, August 2021.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

🏛2024年3月1日(金) 13:30 ~ 15:40 | 🏛T3-B オンライン（Zoom Events)

ドメイン指向推薦 (2)

座長:國島 丈生(追手門学院大学)
コメンテータ:梅本 和俊(東京大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T3-B-8-01] 周辺環境分析に基づく賃貸物件のPersonalized Scoring方式の開発

*須山 瑠万[1]、栗 達[2]、河合 由起子[3]、中島 伸介[3] (1. 京都産業大学大学院、2. 福岡大学、3. 京都産業大学)

14:00 ~ 14:25
[T3-B-8-02] 現代美術業界のモデル化と作家への美術機関推薦システム

*松野 仁志[1]、メンドンサ イスラエル[2]、有次 正義[2] (1. 熊本大学 工学部 情報電気工学科 有次研究室、2. 熊本大学 大学院 先端科学研究部 有次研究室)

14:25 ~ 14:50
[T3-B-8-03] 繰り返し聴取による興味の変化を考慮した楽曲推薦

*重富 竜ノ介[1]、西田 紘子[2]、澤井 賢一[2]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府牛尼研究室、2. 九州大学大学院芸術工学研究院)

14:50 ~ 15:15
[T3-B-8-04] 旅行者の想起プロセスを反映した観光場面と条件に基づく旅程生成方式

*箕輪 真衣[1]、伴 匠生[1]、林 康弘[1]、清木 康[1] (1. 武蔵野大学データサイエンス学科)

15:15 ~ 15:40
[T3-B-8-05] 【技術報告】DMM.comのレコメンドに関する取り組み

森 雅也、土屋 裕太郎 (合同会社DMM.com)

DEIM 2024

周辺環境分析に基づく賃貸物件の Personalized Scoring 方式の開発

須山 瑠万† 栗 達†† 河合由起子††† 中島 伸介†††

† 京都産業大学大学院先端情報学研究科 〒 603–8555 京都府京都市北区上賀茂本山
†† 福岡大学工学部電子情報工学科 〒 814–0180 福岡県福岡市城南区七隈 8 丁目 19 番 1 号
††† 京都産業大学情報理工学部情報理工学科 〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †i2286220@cc.kyoto-su.ac.jp, ††lida@fukuoka-u.ac.jp, †††{kawai,nakajima}@cc.kyoto-su.ac.jp

あらまし 賃貸物件探しをする際，多くのユーザが不動産・住宅情報検索サイトを利用している．しかし，カフェやスポーツジムといったスポット情報や，治安や自然の多さ等の周辺状況を考慮した物件検索は困難である．本研究では，ユーザの年齢・性別・趣味嗜好等に合致する周辺環境を考慮した Personalized Scoring による賃貸物件推薦手法の開発を行う．本稿では，ユーザの趣味や嗜好に関連するスポットを発見し，関連度や距離を基に物件をスコアリングする．また，口コミ分析により地域ごとの周辺環境を可視化する．

キーワード パーソナライゼーション，推薦システム，不動産

1 はじめに

賃貸物件探しをする際，多くのユーザが不動産・住宅情報検索サイトを利用している．既存の不動産・住宅情報検索サイトでは，賃料や専有面積などの条件で物件に対する絞り込み検索が可能だが，その結果については賃料が安い順などの単純な条件でしか並べ替えられないという問題がある．ユーザは賃料の安さや，駅やカフェまでの近さなど，自分が求める複数の条件を考慮して物件を探すが，単純な並べ替えではバランスの良い物件を見つけにくいと考えられる．このことから，ユーザの年齢・性別・趣味嗜好に合う物件の周辺環境を考慮したスコアに基づくランキングを作成することで，ユーザが自分にとってバランスの良い物件を見つけやすくなると考えた．

本研究は，周辺環境を考慮し，対象ユーザの年齢・性別・趣味嗜好などに合致する物件を適切に評価することが可能な Personalized Scoring 方式の開発を目的としている．本稿では，物件の賃料や専有面積に加え，交通や買い物の利便性，ユーザの趣味に関する周辺施設を評価項目とした Personalized Scoring によるランキングを作成し，従来の不動産・住宅情報検索サイトにおける単純な並べ替えによるランキングと比較を行ったので報告する．周辺環境としては，駅・バス停・コンビニ・スーパー・ユーザの趣味に関連するスポットを考慮した．

本稿の構成は以下の通りである．2 章では関連研究を紹介する．3 章では本稿で開発するシステムの概要を解説する．4 章では周辺環境を考慮した Personalized Scoring 方式によるランキングの有効性の検証実験について述べる．最後に 5 章でまとめを記述する．

2 関連研究

パーソナライズされた情報推薦に関する研究として，Xu ら [1] は Pinterest において，ユーザの意図を長期的興味と短期的意図に分解し，それぞれ長期的な将来の行動と即時の次の行動に対して最適化を行なった．ユーザが 2 週間の間にとった行動とリアルタイムログを分析し，長期的興味と短期的意図の両方を捉えた結果，Pinterest でのパーソナライズ体験が大幅に改善した．

また，物件検索に関する研究として，諏訪ら [2] は IoT デバイスを用いて騒音や日当たりを定量的かつ直感的に比較・評価できる指標を構築した．この研究では，物件を探索する際，場所，賃料，広さなどの検索条件のみを用いると，移住後に騒音や日当たりなどの問題に直面することを指摘している．そして，物件探索者が騒音や日当たりなどについても定量的に評価できるように，IoT デバイスを用いて，静穏性 (屋内・屋外)，防音性，採光性，断熱性の 5 指標を構築し，防音性を除く 4 指標において，適切な指標化がなされたことを確認している．しかし，周辺環境の指標化や検索結果の個人化は行えていない．

本村ら [3] は，賃貸物件検索において，ユーザの評価特性をもとに物件検索とランキングを行う手法を提案した．この研究では，ユーザの検索感覚を受容度として表現することで，それを反映した物件検索とランキングを可能にした．また，受容度を導入することで検索システムを改善できる可能性があることを示した．しかし，周辺環境とユーザ特性とのマッチングは考慮されていない．

生活者から見た住み心地をビッグデータから定量的に算出する先駆的な取り組みとしては，米国のスタートアップが展開している Walk Score [4] という指標がある．この指標は，徒歩圏内での生活の快適度を，周辺のアメニティをもとに 100 点満点で算出するものである．

Zhou ら [5] は熊本市内にある 215,309 件の住宅を対象に，Walk Score を算出し，Walk Score と近隣環境の安全要素との相関を求めた．その結果，ウォーカビリティの測定に Walk Score を用いることには限界があることが明らかになり，施設のアクセス性を高めながらウォーカビリティ環境の安全性を向上させることが極めて重要であることを警告した．このことから，物件を評価する際に周辺の施設だけではなく，治安などに

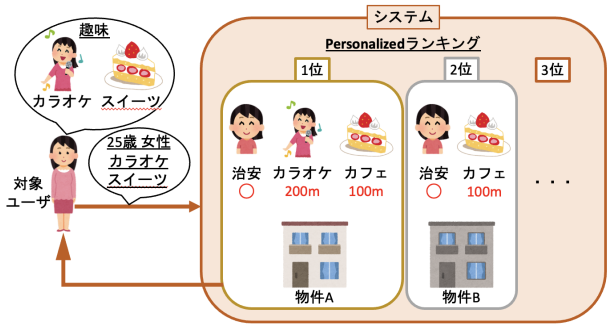


図 1　Personalized Scoring システムの概要 1

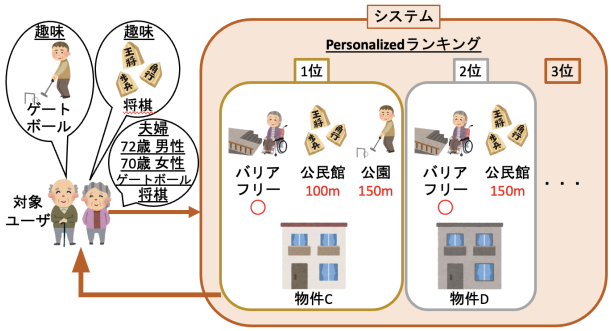


図 2　Personalized Scoring システムの概要 2

ついても考慮すべきだと考えられる．

3　賃貸物件の Personalized Scoring 方式

3.1　提案システムの概要

図 1，2 に提案システムの概要を示す．例えば対象ユーザが 25 歳の女性（カラオケ・スイーツ好き）の場合，周辺の治安が良く，カラオケやカフェが近くにある物件が上位になるような Personalized ランキングによる推薦が可能である．また，対象ユーザが 70 代の夫婦（ゲートボール・将棋好き）の場合は，周辺のバリアフリーが整備されていて，公民館や公園が近くにある物件の推薦が可能な Personalized Score に基づくランキングを作成できる．

3.2　賃貸物件の Personalized Scoring 方式の実装

本研究では，Personalized Scoring の実装のため主に以下の 4 項目の課題に取り組む．

- 各種データの収集

賃貸物件公開データ，地図情報および店舗やスポットのレビューデータ（Google Maps），治安情報（各自治体，SNS），バリアフリー情報（国土交通省，SNS）等を収集し，位置情報と併せて DB に格納する．

- 対象ユーザの基本情報および趣味嗜好に基づく重要評価項目の判定

対象ユーザの年齢・性別および趣味嗜好から重要評価項目を判定する．このとき，学習済みの Word2Vec モデル [1]を用いて，収集したデータから趣味のベクトルと周辺施設のベクトルを生成し，類似度を算出する．例えば，25 歳の女性（カラオケ・スイーツ好き）の場合には，周辺環境としての治安や，周辺施設としてカラオケ店やカフェとの距離が重要評価項目となる（図 1 参照）．

- 家族構成を考慮した重要評価項目の判定

家族で同居する場合には，一人ひとりの重要評価項目を合算した上でスコアリングを実施する．ただし，治安やバリアフリーなど安全に関わる評価項目に対しては，平均的に合算するのではなく，重要度を高めることも検討する．

- Personalized Scoring 方式の実装

上述の重要評価項目毎に賃貸物件のスコアを算出し，各評価項目の重要度を考慮してスコアの統合を行う．

3.3　各評価項目スコアの算出

- 賃料

ユーザの賃料上限値以下の物件の賃料に関する評価項目スコアを 100 とし，上限値を上回る場合には，その状況によって 0〜100 の範囲で設定（上限値+30%以上でスコア 0）（図 3 参照）

- 専有面積

ユーザの専有面積下限値以上の物件の専有面積に関する評価項目スコアを 100 とし，下限値を下回る場合には，その状況によって 0〜100 の範囲で設定（$0m^2$ でスコア 0）（図 4 参照）

- 交通の利便性

最寄りの駅 / バス停までの徒歩距離に応じて，評価項目スコアをそれぞれ 0〜50 の範囲で設定
（1000m(駅) / 500m(バス停) でスコア 0）（図 5 参照）

- 買い物の利便性

最寄りのコンビニ / スーパーまでの徒歩距離に応じて，評価項目スコアをそれぞれ 0〜50 の範囲で設定
（500m(コンビニ) / 1000m(スーパー) でスコア 0）（図 6 参照）

- 趣味に関する周辺施設

ユーザの趣味と最も類似度の高いスポットが 1km 圏内にあるかを判定し，存在した場合にはその類似度に 100 を掛けた値を評価項目スコアとして 0〜100 の範囲で設定（スポットは「ジム」や「カフェ」など，趣味に関連するものに限定）
類似度の算出については，Word2Vec でユーザがフリーワードで入力した趣味 1 つとスポットのカテゴリ名との関連性を分析する．（図 7 参照）

これら 5 つの評価項目スコアの和を取り，賃貸物件の Personalized Score とする．

3.4　口コミから類推可能な地域の暮らしに関するスコアリング及び可視化

株式会社 LIFULL が運営する LIFULL HOME'S の「まちむすび」[2]の口コミデータを用いて，実際に各まちに暮らす住民のまちに対する意見から，住環境をワードクラウドによって可視化し，住みやすさの評価項目別スコアリングも行う．

- まちむすびの口コミデータ

1：https://github.com/WorksApplications/chiVe

2：https://www.homes.co.jp/machimusubi/

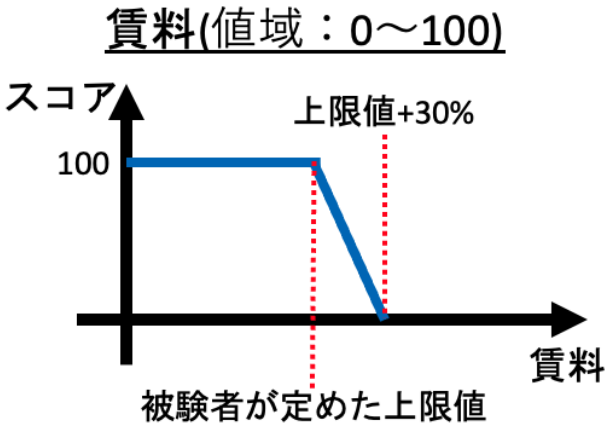


図 3　評価項目「賃料」のスコア算出方法

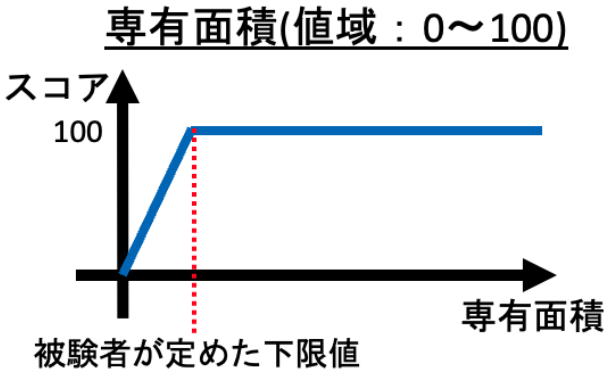


図 4　評価項目「専有面積」のスコア算出方法

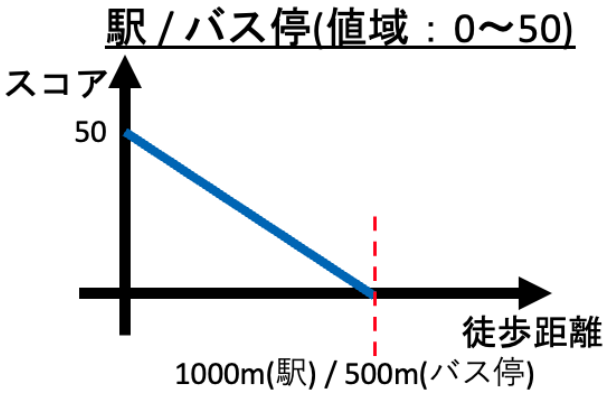


図 5　評価項目「交通の利便性」のスコア算出方法

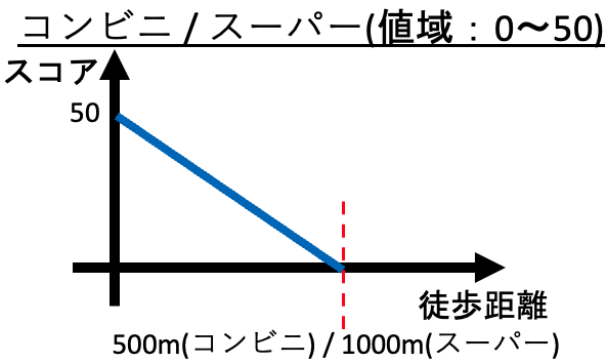


図 6　評価項目「買い物の利便性」のスコア算出方法

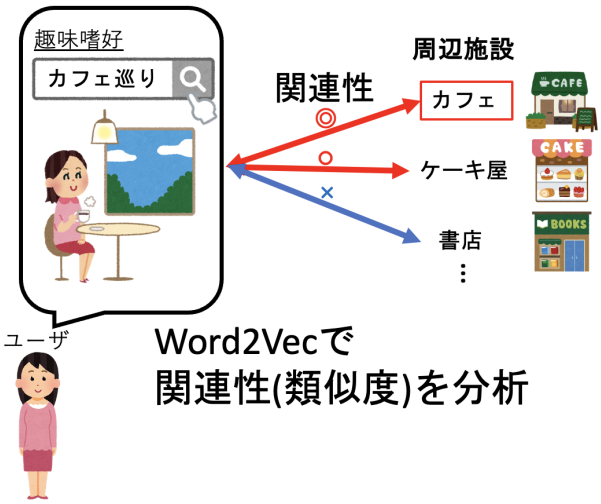


図 7　趣味とスポットとの関連性分析



図 8　ある地域の住環境を可視化したワードクラウド

本研究で用いるまちむすびの口コミデータ数は 330,338 件である．また，各口コミデータには口コミ内容以外に回答者の住まいの郵便番号も含まれている．

- ワードクラウドによる住環境の可視化

ワードクラウドを用いて，口コミに頻出する単語を郵便番号エリアごとに可視化する（図 8 参照）．このとき，名詞・形容詞以外の単語および「こと」などの不要な単語は取り除く．

- 住みやすさの評価項目別スコアリング

5 つの評価項目「買い物のしやすさ」,「交通の利便性」,「子育てのしやすさ」,「治安の良さ」,「自然の多さ」を用いる．fastText によって各評価項目に関連するポジティブな言及をしている口コミと，ネガティブな言及をしている口コミを評価項目ごとにポジティブ: ネガティブ=1:1 となるように収集し学習データとすることで分類器を作成する．そして，郵便番号エリアごとの口コミを各評価項目の分類器へ通し，ポジティブ分類確率の平均を評価項目スコアとする．

4　評 価 実 験

従来の単純なランキングと，提案手法である Personalized Scoring 方式によるランキングを比較し，提案手法の有効性の検証を行う．

4.1　実験の概要

- 対象とする賃貸物件データ

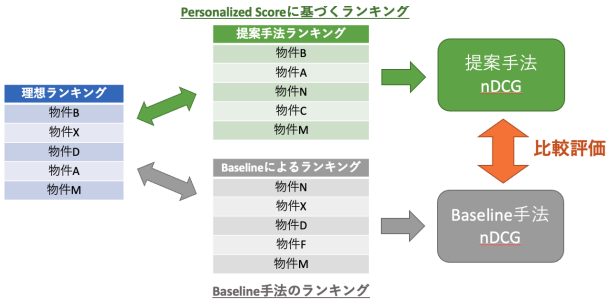


図 9　評価方法の概要

賃貸物件（1 K，バス・トイレ付）10 件

- 被験者

大学生 13 名

- 提案手法で利用する評価項目

賃料，専有面積，交通の利便性，買い物の利便性，
趣味に関する周辺施設

4.2　評 価 方 法

図 9 に評価方法の概要を示す．提案手法の Personalized Score に基づくランキングと，Baseline 手法の単純なランキングの精度を比較するために理想ランキングを用意する．このランキングは各被験者に全ての賃貸物件の情報を提示し，住みたいと思う順に並べ替えてもらうことで作成する．そして，理想ランキングと提案手法のランキングから提案手法の nDCG 値，理想ランキングと Baseline 手法のランキングから Baseline 手法の nDCG 値を求め比較評価を行う．

4.3　事前アンケート

提案手法で利用する評価項目に対するスコアを算出するために，被験者が賃貸物件に求める条件や自身の趣味について回答してもらう．事前アンケートの質問項目は以下の 3 つである．なお，今回の評価実験では大学生向けの物件を対象とするため，回答の選択肢についても大学生向けに制限している．

- 許容できる専有面積（広さ）の下限

$\{15m^2, 20m^2, 25m^2, 30m^2\}$

- 許容できる賃料（家賃）の上限

{4 万円，5 万円，6 万円，7 万円，8 万円 }

- 趣味（1 つ）

記述式

4.4　理想ランキングの作成

提案手法の Personalized Score に基づくランキングと，Baseline 手法の単純なランキングを作成した後，各被験者に賃貸物件 (A〜J) の 10 件（図 10）を住みたいと思う順にランク付けしてもらい，これを Ground Truth とする．図 10 の趣味の欄には，各被験者の趣味に最も類似するスポットが入る．

4.5　実 験 結 果

nDCG@5 に基づく提案手法と Baseline 手法（賃料，専有面積の単純ソート）のランキング精度を図 11 に示す．Baseline 手法である賃料や専有面積の単純ソートよりも，提案手法である Personalized Score に基づくランキングのほうが nDCG@5 の値が大きくなった．

物件名	賃料（万円）	専有面積（m²）	徒歩距離（単位：m）				半径1km圏内
			交通		買い物		趣味
			駅	バス停	コンビニ	スーパー	？？？
A	7.20	33.25	250	250	240	400	ある
B	4.00	17.30	750	110	120	550	ない
C	3.50	22.82	350	32	250	39	ある
D	9.00	30.70	450	400	300	550	ない
E	5.50	30.00	500	260	300	450	ある
F	6.10	23.74	1,200	400	170	550	ある
G	5.40	22.97	1,000	280	350	700	ない
H	8.20	25.02	300	160	300	150	ある
I	3.80	15.00	260	350	300	140	ない
J	8.70	34.71	140	73	83	450	ない

図 10　評価実験で対象とする賃貸物件

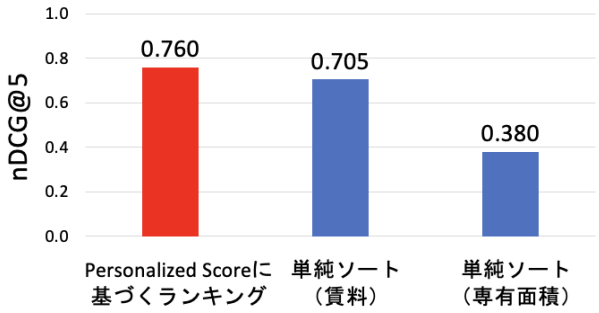


図 11　提案手法と Baseline 手法（賃料，専有面積の単純ソート）のランキング精度

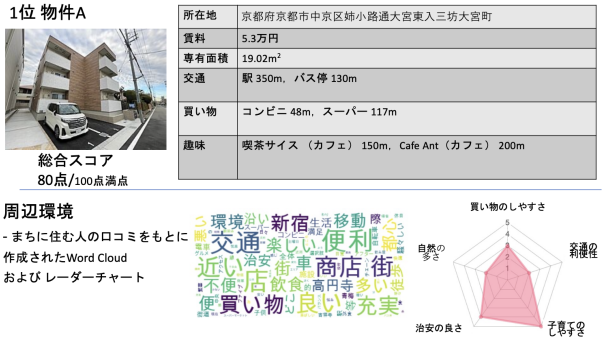


図 12　提案システムのインタフェース

4.6　提案システムのインタフェースに関する検討

提案システムのインタフェースについて，物件の検索結果として図 12 に示すような「趣味」,「総合スコア」,「Word Cloud」,「レーダーチャート」が役に立つと思うかを尋ねた結果を図 13 に示す．

5　お わ り に

周辺環境分析に基づく賃貸物件の Personalized Scoring 方式の開発を行った．Personalized Scoring 方式（提案手法）は，Baseline 手法よりも優れた結果を示した．このことから，複数の評価項目（趣味や周辺の関連施設など）を考慮した Personalised Scoring 手法の有効性が実証された．今後は．新たな評価項目の追加や各評価項目の算出方法の改善を行う予定である．

	5段階評価
評価項目「趣味」	4.0
総合スコア	4.4
Word Cloud	3.5
レーダチャート	4.7

図 13　インタフェースの評価結果

謝　　辞

本研究を遂行するにあたり，株式会社 LIFULL 様に「LIFULL HOME'S まちむすびデータセット」をご提供いただいた．また，本研究の一部は，JSPS 科研費 (課題番号：20H04293, 22H03700, 22K19837) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] Jiajing Xu, Andrew Zhai, Charles Rosenberg, Rethinking Personalized Ranking at Pinterest: An End-to-End Approach, RecSys ‘22: Proceedings of the 16th ACM Conference on Recommender Systems, Pages 502-505, 2022.
[2] 諏訪 博彦, 大坪 敦, 中村 優吾, 野口 真史, IoT センシングによる賃貸物件快適度指標の構築と評価，JSAI2020，2P6-GS-13-02, 2020.
[3] 本村 駿乃介，高木 英行，受容度を用いた賃貸物件データベース検索に関する研究，FSS2018，WG1-3，2018.
[4] Walk Score
https://www.walkscore.com/
[5] Qiaohui Zhou, Riken Homma, Estimating Walk Score and Examining Its Association with Safety Factors of Neighborhood Environment in Kumamoto, Japan, International review for spatial planning and sustainable development, Vol 10 No.3 (2022), 4-15.

現代美術業界のモデル化と作家への美術機関推薦システム

松野　仁志†　　メンドンサ ドス サントス イスラエル††　　有次　正義††

† 熊本大学工学部情報電気工学科　〒 860–8555 熊本市中央区黒髪 2 丁目 39 番 1 号
†† 熊本大学大学院先端科学研究部　〒 860–8555 熊本市中央区黒髪 2 丁目 39 番 1 号
E-mail: †c9248@st.cs.kumamoto-u.ac.jp, ††{israel,aritsugi}@cs.kumamoto-u.ac.jp

あらまし　現代美術の評価においては作品の視覚的な特徴とともに，その作品が生み出された文脈のような知識的特徴が重要な要素である．これまでに，美術作家やその作品，現代美術業界のメカニズムをより深く分析し理解するために，知識的な特徴に基づいた研究が行われてきた．最新の研究では，作家の展覧会歴を利用し作家自身や美術機関の評価値をモデル化するものなどがあるが，具体的に作家のキャリアの戦略に生かすことができる研究は少ない．本研究では，現代美術における美術機関の評価値作成や作家のコミュニティ分析を通して現代美術業界のモデル化を行うと共に，GNN(Graph Neural Network) を用いて美術機関の展覧会の遷移と美術作家の展覧会歴を学習し，美術作家に対して美術機関を推薦するシステムを実装した．現代美術業界のモデル化では，作家の展覧会への参加に基づいたグラフ構造が，作家の活動領域や出生年代に起因する活動年代の特徴を内包していることを示した．GNN の推薦システムについては，エッジの重みとノード表現の比較実験を行い，ウィンドウサイズの変更による性能への影響と，ノード表現学習の有無による性能への影響を示した．

キーワード　現代美術，モデル化，作家コミュニティ，展覧会，知識グラフ，GNN

1　はじめに

美術作家の多くは，作品の制作と作品展示の開催によって新たな評価を築いている．作品展示は，作家自身やその作品に対する評価を築き，作品の販売や仕事を得る重要な機会である．しかし，現代美術業界においてキャリアを構築する際，どのような美術機関で展覧会を行い，どのように人脈をつくり評価を得るかによって，その後の作家のキャリアは大きく異なる．現代美術は作品の媒体に縛りがなく，その定義も曖昧であることから，業界におけるキャリア構築の仕方は多様的である．そのためコミュニティや権威も多様的であり，作家が思い描くキャリアを実現するためには，どのようなコミュニティを選択して活動を行うかを意識する必要がある．

作家の評価やその後のキャリアにおいて，人脈やコミュニティは重要な要素であり，現代美術業界の環境をモデル化する際には，作品の視覚的特徴だけでなく，その作家が作ってきた文脈などの知識的特徴を考慮した分析を行う必要がある．これまで，美術作家やその作品，現代美術業界のメカニズムをより深く分析し理解するために，知識的特徴に基づいた研究が行われてきた．その中には，作家の展覧会歴を利用して，作家の評価の遷移や美術機関の権威をモデル化する研究 [1] や，作品同士の文脈的関連性を利用し GNN(Graph Neural Network) を用いて分類や予測を行う研究 [2] がある．これらの研究により，現代美術の分析や定量化において知識的特徴を用いることの有効性が示されている．特に Fraiberger ら [1] の研究では，作家の展覧会歴から将来の展覧会における評価値の予測を行っている．しかしこの研究においては，展覧会を開催する際の場所となる美術機関がどれくらいの評価値であるかという数値の予測にとどまっており，文献内では美術機関を特定して予測できないことが課題として挙げられていた．

本研究では，美術作家に美術機関を推薦し，作家の戦略の参考となることを目的に，以下の二つを実施した．

（1）　美術機関の評価値作成と作家コミュニティの分析による現代美術業界のモデル化

（2）　GNN のリンク予測を用いた美術機関推薦システムの実装と実験

その結果によって明らかとなった項目は，以下の三つである．

（1）　美術作家のコミュニティ分析において，作家の展覧会への参加に基づいたグラフ構造の構築とグラフベースのクラスタリングを行い，展覧会に基づくコミュニティが，作家の活動領域や出生年代に起因する活動年代の特徴を内包していることを示した．

（2）　GNN の推薦タスク実験にて，エッジの重みに作家の過去評価の移動平均を用いて実験を行い，そのウィンドウサイズを拡大することでリンク予測の精度が向上したが，推薦結果には悪影響を与えていることが示された．

（3）　GNN の推薦タスク実験にて，GraphSAGE によるノード表現の事前学習を行う場合とノード表現学習を行わなかった場合を比較し，ノード表現学習を行うとリンク予測の精度が向上したが，推薦結果を確認すると，表現学習を行わない場合が実際の展覧会をより的中できていることが示された．

本稿の構成は以下の通りである．2 章では関連研究を紹介し，3 章で美術機関の評価値作成と作家のコミュニティ分析によって現代美術業界のモデル化を行った後，4 章で推薦システムの実装と実験を行う．その結果をもとに 5 章で今回の研究内容をまとめ，今後の展望を述べる．

2 関連研究

本章では美術業界の定量化を行う関連研究と，美術作品を視覚的特徴のみ，または視覚的特徴と知識的特徴を組合せて分析する関連研究を紹介する．

2.1 美術業界の定量化に関する研究

前章で述べた通り，美術作家やその作品，現代美術業界のメカニズムをより深く分析し理解するために，視覚的特徴に加えて知識的特徴に基づいた研究が行われてきた．

Fraiberger ら [1] は，客観的な評価が難しい美術業界において，作家の展覧会歴を用いて美術機関の評価値を計算することで，美術的権威を定量化し，作家のキャリア遷移をモデル化している．作家が行った展覧会歴をもとに，各美術機関の間に遷移ネットワークを構築し，固有ベクトル中心性 [3] や PageRank [4] などのネットワーク中心性指標を用いた計算を行った．これによって美術作家のキャリアにおけるネットワークを表現し，その権威レベルを定量化できることを示した．また，作家が過去に行った展覧会の開催機関の平均評価値をもとに，将来その作家が展覧会を行う美術機関の評価値を予測するマルコフモデルの構築を行い，過去の展覧会歴の評価が将来の評価に影響を与えることを示した．しかし Fraiberger ら [1] の研究では，美術機関の評価値のみの予測にとどまっており，美術機関を予測するまでには至っていない．本研究では Fraiberger ら [1] が行った，ネットワークベースの美術業界のモデル化手法を利用しつつ，美術機関を推薦することを目的としている．

2.2 作品の視覚的特徴に基づく研究

美術作品の視覚的な分析に関しては，WikiArt [5] や OmniArt [6] などの美術作品の大規模データセットが公開され，美術を取り巻くデータの整備が進歩している．それにより，コンピュータビジョンの技術を用いて，作品の色や絵画の筆致などの視覚的特徴に基づいた分析を行う研究が進められてきた．Strezoski ら [7] は，作品の色に基づいたフィルタリングによって，OmniArt データセット内の作品を探索するインターフェースである ArtSight を提案している．

2.3 作品の視覚的特徴と知識的特徴を組合せた研究

現代美術作品や作家を評価する際には，その作品が生み出された文脈のような知識的特徴が重要な要素であり，どのようなコンセプトで作品を制作するかが作品や作家自身の価値に大きく影響している．Efthymiou ら [2] は，視覚的特徴と知識的特徴をそれぞれ CNN と GNN を用いてモデル化するマルチモーダルな手法である ArtSAGENet を提案し，視覚芸術作品の分析を行った．美術作品をノードとして扱い，作品のメディアや流派，作家が活動した場所などの意味特性に基づいたエッジを張ることで，作品間の関係性を表すネットワークを構成する．大規模なグラフの計算に適応できる GraphSAGE [8] を用いてその知識グラフを学習し，作品のスタイル分類や制作年代推定などのタスクを行い，VGG や ResNet などの CNN，GraphSAGE や GraphSAINT などの GNN 単体を用いた手法との比較実験を行った．この実験では作品のノード表現を作品画像の視覚特徴から取得し，ArtSAGENet で関係性を学習する組合せが，他の CNN や GNN を単体で用いた時よりも特に優れた性能を発揮した．これによって，美術作品の視覚的特徴に加えて，その関係性などの知識的特徴を学習することの有効性が示されている．本研究では，作家の展覧会歴という知識的特徴に焦点を当て，Efthymiou ら [2] と同じく GraphSAGE を用いて美術機関の推薦を実施する．

3 現代美術業界のモデル化

本章では，4 章で紹介する推薦システムを実装するために，本研究で用いたデータセットの内容を紹介し，美術機関の評価値作成と作家コミュニティの分析によって，現代美術業界のモデル化を行う．

3.1 データセット

本研究のデータセットとして，国立アートリサーチセンターが主催する日本近現代アートのリサーチポータルである Art Platform Japan に掲載されている，日本アーティスト事典（ベータ版）[9] と日本現代美術展調査 [10] から取得したデータのうち，2520 名の作家，698 件の美術機関，2354 件の展覧会データを利用する．作家の情報には生没年月日や，20 に分割された活動領域などの情報が含まれており，展覧会の情報には展覧会を開催した美術機関名や参加した作家，展覧会の会期が含まれている．なお，1 人の作家が複数の活動領域を持つこともある．

本研究では，このデータセットから，特定の作家が参加した展覧会やそれを開催した美術機関などの，データリレーションを円滑に取得するために，Art Platform Japan から取得したデータをグラフデータベースに保存し，データを再構成している．グラフデータベースは，グラフにおけるノードとエッジをそのままデータモデルとして表現し保存できるため，直感的にデータリレーションを作成し取得できる点で，グラフデータの利用に適している．図 1 に構築したグラフデータベースのデータモデルを示す．美術作家を Artist ノード，展覧会を Exhibition ノード，美術機関を Institution ノードとして，展覧会への参加や開催に基づき各ノードの間にエッジを張る．具体的には，作家が参加した展覧会に対して JOINED エッジ，美術機関が開催した展覧会に対して HELD エッジを張っている．また，作家が有する活動領域に関して Field ノードとして表現し，各作家の活動領域に対して BELONGED エッジを張る．

3.2 美術機関の評価値作成

美術機関の評価値を定量化するために，Fraiberger ら [1] の研究における美術機関の権威の定量化手法と同様に評価値の計算を行う．Fraiberger ら [1] の手法では，各作家の展覧会の変遷を美術機関の遷移として捉えたネットワークを構築し，固有ベクトル中心性や PageRank などのネットワーク中心性指標を用いて評価値を計算している．作家の展覧会の変遷を美術機関

の遷移として捉えたグラフである TRANSITIONED における，エッジの例を図 2 に示す．特定の作家の展覧会以前に，その作家の展覧会を開催した美術機関から，当該の展覧会を開催した美術機関へのエッジを張り，全ての作家の展示遷移において同じ手順でエッジを張る．例えばある作家が美術機関 A，B，C，D の順で展覧会を行うと，図 2 に示すように機関 D へのエッジが最も多く張られる．この美術機関同士に張られたエッジを集計することによって，遷移回数をエッジの重みとした美術機関の遷移ネットワークを構築することができる．Fraiberger ら [1] の研究では，特定の展覧会以前の何回前までの展覧会の美術機関から当該の美術機関にエッジを張るかという，展覧会のウィンドウサイズを比較した実験が行われているが，本研究では Fraiberger ら [1] の研究で最終的に採用されている，過去の展覧会のウィンドウサイズ τ を $\tau = \infty$ とするネットワークを用いる．この遷移エッジを，グラフデータベースでは図 1 の TRANSITIONED エッジとして保存している．

上記で構築したグラフ構造に対し，ネットワーク中心性の指標である PageRank [4] を用いて計算を行い，各美術機関の評価値を算出した．PageRank の計算結果の上位 20 の機関を表 1 に示す．国内の展覧会を中心に構成されたデータセットであるため，PageRank の計算結果も国内の機関が中心となっている．現代美術を扱う関東圏の公立美術館が上位の多くを占めており，9 位には世界での権威が高いポンピドゥー・センター・メッスが入っている．

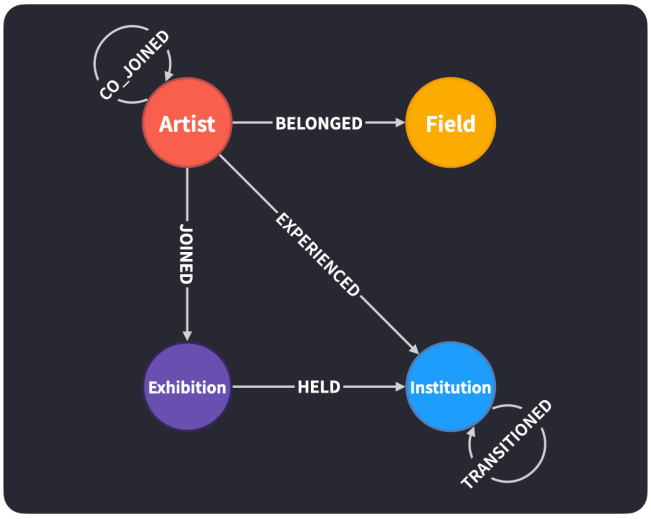


図 1　グラフデータベースのデータモデル

3.3　作家のコミュニティ分析

推薦システム実装に用いるノード表現を作成するために，美術作家の関連性について分析を行う．ここでは，作家の関連性をネットワークとして表現するために，同じ展覧会への参加に基づいたグラフ，CO_JOINED を構築する．CO_JOINED エッジの例を図 3 に示す．同じ展覧会に作品が出展されている作家 A，B，C 同士をエッジで結び，その回数を重みとして付加することで作家同士の無向グラフを構築している．このエッジを，グラフデータベースでは図 1 の CO_JOINED エッジとし

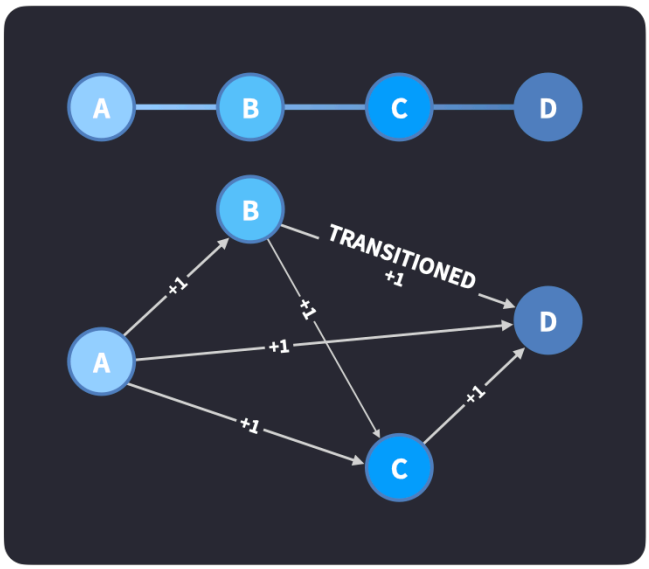


図 2　美術機関の遷移ネットワーク TRANSITIONED

表 1　PageRank の上位 20 機関

順位	美術機関名	PageRank
1	国立国際美術館	17.0742
2	東京国立近代美術館（新館）	14.0209
3	千葉市美術館	13.7927
4	高松市美術館	12.1386
5	東京都現代美術館	11.1339
6	宮城県美術館	9.6781
7	サンタゴスティーノ美術館	7.6313
8	群馬県立近代美術館	7.3854
9	ポンピドゥー・センター・メッス	6.9230
10	名古屋市美術館	6.8918
11	静岡県立美術館	6.8600
12	板橋区立美術館	6.8022
13	世田谷美術館	6.6926
14	神奈川県立近代美術館葉山館	6.5548
15	練馬区立美術館	6.5252
16	埼玉県立近代美術館	5.9736
17	京都国立近代美術館（新館）	5.7768
18	金沢 21 世紀美術館	5.7361
19	横浜美術館	5.6033
20	鳥取県立博物館	5.5866

て保存している．現代美術の展覧会においては，美術キュレーターが美術作家や作品の選定を行うことが多く，同じ展覧会に出展しているということは，キュレーターによってそれらの作家や作品の間に何らかの意味付けがなされたと捉えることができる．よって，今回作成した展覧会参加に基づくネットワークは，現代美術における文脈などの知識的特徴を反映できると考えられる．

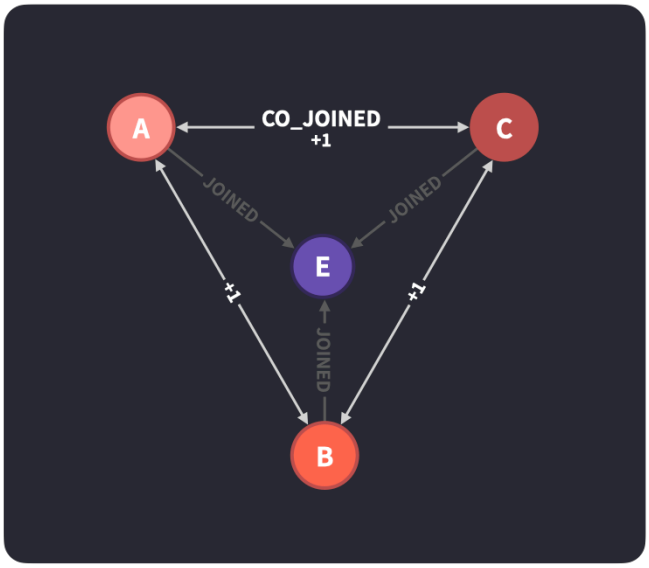


図 3　作家の共同参加ネットワーク CO_JOINED

このグラフに対してクラスタリングを行うことで，美術作家のコミュニティを検出し，展覧会に基づく関連性に影響を与える作家の情報を抽出する．グラフのクラスタリング手法としては Louvain 法 [11] を用いる．Louvain 法は，クラスタ内の疎密の指標である Modularity [12] を用いたクラスタリングの一種であり，Modularity が最も高くなるように隣接ノードを選択し，ノードを集約する手順を繰り返すことにより，クラスタリングを行う手法である．隣接ノード同士のみに Modularity の計算を適用することにより，大規模なグラフに対して高速にクラスタリングを行うことができる．Louvain 法によるクラスタリング結果から，属する作家が 10 名以上のコミュニティの活動領域の割合を図 4 に示す．縦軸に活動領域，横軸にコミュニティをとり，コミュニティごとに属する作家の活動領域を集計した後，全体を 1 とする割合を算出した．各コミュニティの中での最大値を赤字で表記している．作家が 10 名以上のコミュニティは 8 つあり，各コミュニティ内の作家の活動領域の割合から，それぞれのコミュニティには活動領域に基づいた特徴があることが分かる．

また，図 4 より，コミュニティ6，7，8 は絵画領域の作家が多く属するコミュニティであることが分かる．作家の活動領域以外でクラスタリングが行われた要因を特定するため，コミュニティ6，7，8 に対して，さらに作家の出生年での比較を行った．比較結果のヒストグラムを図 5 に示す．各コミュニティの最頻値は異なり，特にコミュニティ6 とコミュニティ7 を比較すると作家の分布がほとんど異なるため，これらのコミュニティ

	1	2	3	4	5	6	7	8
絵画	0.03	0.16	0.14	0.01	0.25	**0.70**	**0.21**	**0.44**
彫刻	0.00	0.01	0.09	0.04	0.00	0.06	0.16	0.21
版画	0.00	0.04	0.04	0.02	**0.64**	0.14	0.03	0.12
映像	0.03	0.01	0.07	0.01	0.00	0.00	0.12	0.01
マンガ	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
建築	0.00	**0.44**	0.02	0.00	0.00	0.00	0.01	0.01
イラストレーション	0.00	0.01	0.09	0.00	0.00	0.01	0.02	0.01
コンセプチュアルアート	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.03
工芸	0.07	0.07	0.04	**0.87**	0.00	0.05	0.01	0.01
デザイン	0.00	0.15	**0.36**	0.02	0.00	0.01	0.01	0.00
サウンドアート	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.01
写真	**0.85**	0.03	0.04	0.01	0.11	0.01	0.12	0.04
パフォーマンスアート	0.00	0.01	0.02	0.00	0.00	0.00	0.06	0.02
舞台芸術	0.01	0.02	0.04	0.00	0.00	0.01	0.01	0.01
メディアアート	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.01
ファッション	0.00	0.03	0.00	0.01	0.00	0.00	0.01	0.00
書	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01
インスタレーション	0.00	0.00	0.02	0.02	0.00	0.00	0.12	0.03
詩	0.01	0.03	0.00	0.00	0.00	0.00	0.01	0.00
いけばな	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

図 4　CO_JOINED グラフのクラスタリング結果

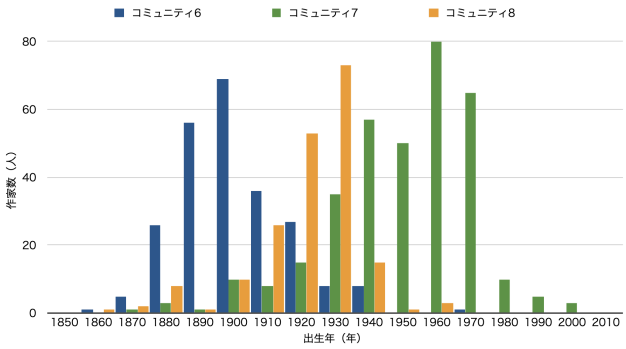


図 5　コミュニティ6，7，8 の作家の出生年ごとの比較

は出生年代によって特徴付けられていると言える．これは，作家が生まれた年代によって活動年代も限定されいることが原因であると考えられる．現代美術においては社会的背景と展示のテーマが深く関わっていることにより，同じ時代に活動した作家同士が展覧会で扱われやすいため，年代によるコミュニティが形成されていると考えられる．

4　推薦システムの実装と実験

提案する美術機関推薦システムのパイプラインを図 6 に示す．本提案手法では，まず美術機関の遷移と作家コミュニティの特徴を美術機関と作家ノードの特徴量に反映するため，美術機関の遷移を表す TRANSITIONED グラフと，美術作家同士の共同展示を表す CO_JOINED グラフから GraphSAGE を用いてノード表現学習を行った後，それを用いてリンク予測のための学習を行う．リンク予測のための学習では，作家が展覧会を経験した美術機関に EXPERIENCED エッジを張ることによって，作家の展覧会経験を表現する．作家と美術機関を EXPERIENCED エッジで結んだ 2 部グラフから，GraphSAGE を用いて学習を行い，最終的にエッジの存在予測とエッジの重み回帰の 2 つのタスクを行う．4.1 節で作家と美術機関のノード表現の作成の詳細を説明し，4.2 節で推薦システムの予測タスクの詳細を説

明する．4.3 節で実際に推薦タスクの実験と考察を行う．

4.1 作家と美術機関の特徴量の作成

モデルのリンク予測に用いるために，各作家と美術機関のノード表現の作成を行う．美術機関の遷移を定式化して埋め込むために，前章で作成した美術機関の遷移ネットワークに基づいた表現学習を行った．美術機関のノードの初期特徴量としては，各美術機関が関わった活動領域の割合を用いている．具体的には，その美術機関が開催した展覧会に参加した全ての作家が有する活動領域を集計し，全体を 1 とした各領域の割合を算出した 20 次元の配列を用いる．表現の学習には GraphSAGE [8] を用い，32 次元の特徴量を作成する．GNN はノードの表現を得るために，隣接するノードの特徴量ベクトルを集約しながら学習を行うが，GraphSAGE は特定ノードの隣接ノードを固定数でランダムにサンプリングすることで，大規模なグラフでの計算に対応できる GNN モデルである．

美術作家のノード表現を得るために，作家同士の関係性を捉えたネットワークを作成し表現学習を行った．前章で作成した作家の展示参加ネットワークに基づいた表現学習を行うことで，作家のコミュニティをノードの特徴として反映する．作家の初期特徴量としては，各作家が有する活動領域を 20 次元の二値配列化したものを用い，美術機関のノード表現の作成時と同様に GraphSAGE を用いて，32 次元の特徴量を得る．

4.2 エッジの存在予測と重みの回帰

作成したノード表現を用いて，作家に対し美術機関を推薦するための予測器を作成する．本研究では，エッジの存在予測とエッジの重みの回帰の 2 つのタスクを行うことで，美術機関と作家の間に関係性が存在しうるか，また作家がどれくらいの評価を獲得した場合にその美術機関に遷移しうるかを予測する．重みの回帰タスクの際に用いる作家の評価値については，Fraiberger ら [1] の研究で用いられた，作家が過去に行った展覧会の評価値の算出方法を参考にする．Fraiberger ら [1] は，将来作家が展覧会を行う美術機関の評価値を予測するマルコフモデルを構築する際，作家が過去に行った展覧会の開催機関の平均評価値を現在の状態として用いている．ある遷移時における作家の直前の評価値 m_τ は，作家の過去 τ 回の各展覧会の評価 π の平均値であり，式 1 で表される．

$$m_\tau = \frac{1}{\tau}\sum_{k=1}^{\tau}\pi_{\tau-k+1} \tag{1}$$

この評価値を，図 6 の EXPERIENCED エッジの重み m_τ として用いる．学習に利用する際は評価値を最小値 0，最大値 1 に正規化した上で，0 から 10 までの 11 個の整数値に変換している．なお，作家が同じ美術機関での展覧会を複数回経験している場合は，最初に経験した際の評価値をエッジの重みとしている．また，エッジの存在予測の際には EXPERIENCED エッジの重みとして，作家がその美術機関の展覧会を行った回数 t_e を用いて学習を行う．

4.3 推薦タスクの実験

本研究では，エッジの存在予測と，エッジの重みの回帰の両タスクとも GNN に GraphSAGE を利用した．プログラミング言語として Python を使用し，GNN を実装するために PyTorch Geometric ライブラリ [13] を使用している．両タスクともミニバッチによる学習を行い，エッジの存在予測に関しては評価指標として AUC を用い，エッジの重みの回帰に関しては RMSE を用いて計測を行った．データセットの分割はエッジに基づきそれぞれ，80%，10%，10%の割合で train，validation，test に分割を行い，各エポックごとに validation データを用いて予測を行い評価指標での計測を行う．各タスクでの学習に使用したパラメータを表 2 に示す．学習率は，エッジの存在予測タスクで 0.0002，エッジ重みの回帰タスクで 0.0003 とした．エポック数としては，エッジの存在予測タスクで 90，エッジ重みの回帰タスクで 70 までの学習で予測精度が上限に達した．エッジの重み回帰タスクにおける重み m_τ や各ノードに入力するノード表現を変化させた実験を行い，その結果を考察する．

表 2　各タスクに用いたパラメータ

	設定値	
パラメータ	存在予測タスク	回帰タスク
バッチサイズ	256	256
エポック数	90	70
学習率	0.0002	0.0003

4.3.1 エッジの重み m_τ のウィンドウサイズによる比較

エッジの重み回帰タスクにおいて評価値 m_τ をエッジの重みとして用いているが，ここでは過去 τ 回の展覧会ウィンドウのサイズを変化させることで，結果を比較する．$\tau = 3$ から $\tau = 50$ まで変化させた validation の精度の結果を図 7 に示す．なお，validation の計測は学習を終了する 5 回前までのエポックの各 validation の結果を平均して比較している．$\tau = 3$ からウィンドウサイズを大きくするほど RMSE の値が低下する結果となった．しかし，$\tau = 20$ 以上のウィンドウサイズではほとんど精度に変化がなかった．

また，現代美術家である草間彌生を例に，推薦内容の出力を行った．両タスクにおける結果をもとに，存在予測の確率が 0.5 以上の美術機関を取得し，評価値の予測値が高い順に上位 10 機関を推薦した．その内容を表 3，表 4 に示す．美術の Web メディアである Tokyo Art Beat に掲載されている，日本全国の展覧会とアートイベント情報 [14] を参照し，2018 年以降に草間彌生の作品が出展された展覧会の開催機関を赤字で表記している．Art Platform Japan のデータセット内で，草間彌生の展覧会は 2019 年までが登録されており，2018 年以降の展覧会歴が Tokyo Art Beat のものと重複しなかったため，2018 年以降の展覧会歴と推薦結果を比較した．海外の美術機関に関しては Tokyo Art Beat に展覧会歴としてほとんど登録されていないため，それぞれの美術機関の Web ページから 2018 年以降に草間彌生の作品展示を行ったかどうかを確認した．τ の変化によって，推薦結果の上位の内容にばらつきがあるが，現代美術

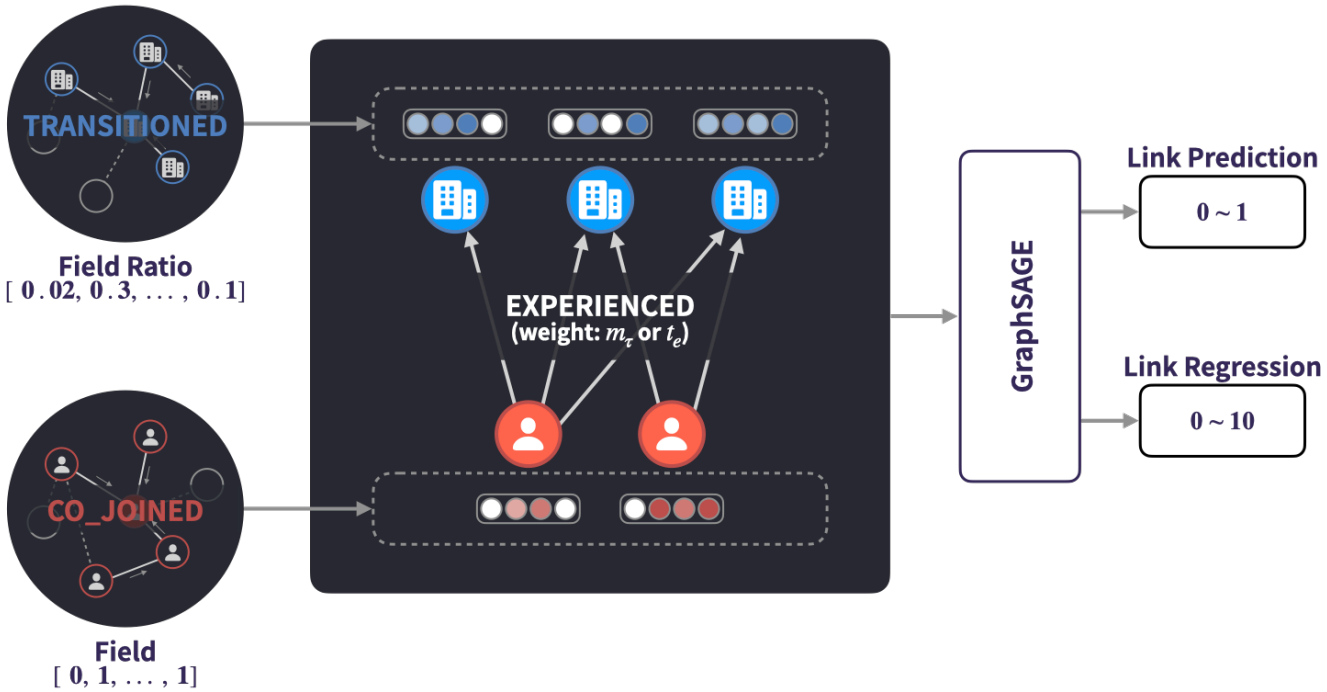


図 6　美術機関推薦モデルのパイプライン

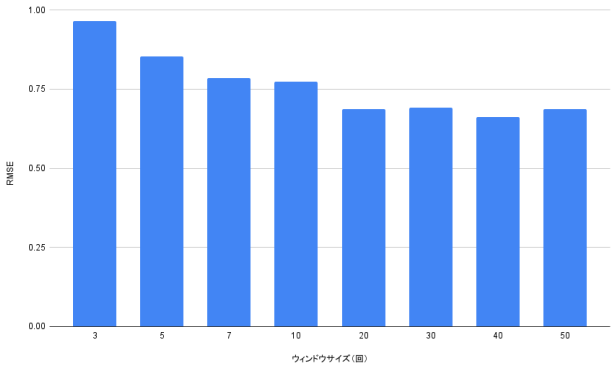


図 7　エッジの重みの変化による精度の結果

の美術館の中でも権威のあるポンピドゥ・センター・メッスが $\tau = 5$ と $\tau = 10$ にて上位に入っている．実際に開催された美術機関を推薦している数としては，$\tau = 10$ の場合が最も多く，3 つの機関で展覧会が行われた．$\tau = 20$ の場合の推薦結果は，予測評価値が全て 0 となり美術機関を推薦することができていない．

実験結果を考察する．ウィンドウサイズを変化させた際の重みの分布を表 5 に示す．重みの値ごとに，各ウィンドウサイズの中で最もエッジの数が多いものを赤字で表記している．ウィンドウサイズが大きくなるほど 3 から 7 までの中間層の評価値をもつエッジが少なくなり，エッジの重みのばらつきが少なく，より低い評価値に分布していることが分かる．ウィンドウサイズを大きくするほど RMSE の評価指標による予測誤差が下がり，さらに予測評価値が小さくなっていたのは，エッジの重みのばらつきがなくなり，低い値に偏っていることが原因であると考えられる．

表 3　提案手法 $\tau = 5, 7$ における草間彌生への推薦結果

$\tau = 5$		$\tau = 7$	
美術機関名	評価値	美術機関名	評価値
ポンピドゥー・センター・メッス	1.03	高知県立郷土文化会館	0.93
京都国立近代美術館（新館）	0.22	田川市美術館	0.88
金沢 21 世紀美術館	0.12	板橋区立美術館	0.87
世田谷美術館	0.12	ツェリエ現代美術館	0.86
東京オペラシティアートギャラリー	0.10	練馬区立美術館	0.84
横浜美術館	0	青森県立美術館	0.74
藤沢市民ギャラリー	0	岡山県立美術館	0.72
メキシコ・ティファナ文化センター	0	入善町下山芸術の森発電所美術館	0.67
デュッセルドルフ市立博物館	0	周南市美術博物館（徳山市美術博物館）	0.67
フォルクローレ・ミュージアム	0	ヘルシンキ市立美術館	0.63

表 4　提案手法 $\tau = 10, 20$ における草間彌生への推薦結果

$\tau = 10$		$\tau = 20$	
美術機関名	評価値	美術機関名	評価値
東京国立近代美術館工芸館	0.97	ジャパン・ソサエティー・ギャラリー	0
ポンピドゥー・センター・メッス	0.81	サンディエゴ・ポートコミッションギャラリー	0
名古屋市美術館	0.81	秋田総合生活文化会館アトリオン美術展示ホール	0
愛知県美術館	0.79	坂出市民美術館	0
シティ・スクエア	0.72	よねざわ市民ギャラリー・ドリーム	0
世田谷美術館	0.69	藤沢市民ギャラリー	0
ニューヨーク近代美術館	0.63	マッケンジー・アート・ギャラリー	0
群馬県立近代美術館	0.60	デュッセルドルフ市立博物館	0
松本市美術館	0.59	ローマ市立フォルクローレ美術館	0
ザ・ライト・ファクトリー・フォト・アーツ・センター	0.58	ニュー・サウス・ウェールズ州立美術館	0

表 5　ウィンドウサイズごとのエッジ数の分布

	エッジの重み									
τ	9	8	7	6	5	4	3	2	1	0
3	9	25	50	161	411	962	1938	2436	3306	7835
5	10	25	6	48	209	646	1905	2788	3496	8000
7	9	22	3	26	134	648	1910	3249	3757	7375
10	9	22	3	20	82	453	1837	3325	4042	7340
20	9	22	3	19	55	217	1465	3296	4571	7476

4.3.2　ノード表現による比較

この実験では，モデルの学習時のノード表現を変化させた

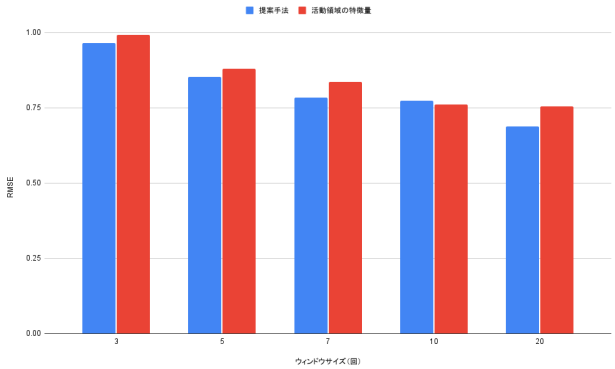


図 8　ノード表現の変化による回帰タスクの結果

上で結果を比較する．提案手法の中で用いているノード表現は 4.1 節で作成しているが，その他に，美術機関には各機関が関わった活動領域の割合を，美術作家には各作家が有する活動領域を二値配列化したものをノード表現として用いて比較する．この比較により，提案手法においてノード表現を得るために，美術機関の遷移グラフや美術作家同士の共同展示グラフを GraphSAGE で学習することの有効性を確認できる．エッジの重み回帰タスクにおける validation の精度の結果を図 8 に示す．$\tau = 10$ 以外では提案手法の方が，活動領域をノード表現として用いるよりも RMSE の値が小さく，より精度が高い結果となった．また，エッジの存在予測タスクにおける実験では，提案手法での精度の結果では AUC が 0.918，活動領域を入力した場合では AUC が 0.883 となり，提案手法での精度が高い結果となった．

この結果の差の有意性を示すため，エッジの重み回帰タスクにおける各ノード表現を用いた結果の差異を，表 6 に示す．提案手法の予測誤差を $RMSE_{a_\tau}$，活動領域をノード表現とした場合の予測誤差を $RMSE_{b_\tau}$ とすると，両結果の差異の割合 x_τ は式 2 で表される．表 6 から，差異の割合 x_τ の平均値は 4.68%と算出できる．

$$x_\tau = \frac{|RMSE_{a_\tau} - RMSE_{b_\tau}|}{\frac{1}{2}(RMSE_{a_\tau} + RMSE_{b_\tau})} \times 100 \qquad (2)$$

また，各ウィンドウサイズでの実験ごとに $RMSE_{a_\tau}$ のばらつきを調べるため，それぞれもう 1 セットづつ実験を行い，$RMSE'_{a_\tau}$ と $RMSE'_{a_\tau}$ の差異の割合 y_τ を式 3 で算出する．表 6 から，差異の割合 y_τ の平均値は 2.39%と算出できる．よって，提案手法と活動領域をノード表現とした場合の予測誤差の差異は，実験ごとに生じる予測誤差の差異の範囲より大きく，それぞれ予測精度に差があると言える．

$$y_\tau = \frac{|RMSE_{a_\tau} - RMSE'_{a_\tau}|}{\frac{1}{2}(RMSE_{a_\tau} + RMSE'_{a_\tau})} \times 100 \qquad (3)$$

推薦された内容を表 7，表 8 に示す．この実験でも，提案手法での実験同様に，Tokyo Art Beat の展覧会歴を参照し，実際に展覧会が行われた美術機関を赤字で表記している．提案手法での結果と比較して，国内の美術館の割合が高い結果となり，

表 6　各ノード表現を用いた場合の RMSE の差異

τ	$RMSE_{a_\tau}$	$RMSE_{b_\tau}$	$RMSE'_{a_\tau}$	x_τ[%]	y_τ[%]
3	0.964	0.993	0.953	2.96	1.15
5	0.853	0.879	0.837	3.00	1.89
7	0.784	0.835	0.828	6.30	5.46
10	0.773	0.760	0.750	1.70	3.02
20	0.687	0.755	0.684	9.43	0.44

表 7　活動領域の特徴量を用いた $\tau = 5, 7$ の推薦結果

$\tau = 5$		$\tau = 7$	
美術機関名	評価値	美術機関名	評価値
金沢 21 世紀美術館	1.12	京都国立近代美術館（新館）	1.73
千葉市美術館	1.04	東京国立近代美術館工芸館	1.73
ポンピドゥー・センター・メッス	0.84	森美術館	1.57
豊田市美術館	0.75	神奈川県立近代美術館葉山館	1.52
ナビオ美術館	0.72	練馬区立美術館	1.34
国立新美術館	0.68	千葉市美術館	1.31
神奈川県立近代美術館葉山館	0.67	国立国際美術館	1.26
徳島県立近代美術館	0.64	国立新美術館	1.17
宮城県美術館	0.62	金沢 21 世紀美術館	1.14
世田谷美術館	0.62	愛知県美術館	1.13

表 8　活動領域の特徴量を用いた $\tau = 10, 20$ の推薦結果

$\tau = 10$		$\tau = 20$	
美術機関名	評価値	美術機関名	評価値
森美術館	0.17	宮城県美術館	0.16
東京都美術館（新館）	0.16	神奈川県立近代美術館葉山館	0.12
東京国立近代美術館工芸館	0.15	石川県立美術館	0.10
高松市美術館	0.15	世田谷美術館	0.08
富山県立近代美術館	0.14	横浜市民ギャラリーあざみ野	0.05
東京都写真美術館	0.14	高松市美術館	0.04
愛知県美術館	0.14	山口県立美術館	0.03
ザ・ライブラリー・アンド・ギャラリー	0.12	シドニー現代美術館	0.01
山梨県立美術館	0.12	広島市現代美術館	0.01
いわき市立美術館	0.12	熊本市現代美術館	0.00

国内の現代美術で権威の高い金沢 21 世紀美術館や森美術館，1945 年以降の現代美術を取り扱いの方針として掲げている千葉市美術館が，複数回上位に入っている．$\tau = 7$ の場合に，実際に展覧会が開催された美術機関が最も多くなり，5 つの機関が含まれた．なお，この実験結果でも提案手法での実験同様，$\tau = 20$ の場合の評価値が最も低い値となった．

表 3，4，7，8 より，ウィンドウサイズを変化させた 4 つの推薦結果の的中数は，提案手法で平均が 1.75，活動領域をノード表現に用いた場合で平均が 2.5 であるため，活動領域を用いた場合の方が実際の展覧会を的中できていると言える．本研究で提案手法においてノード表現学習の有効性が示されなかった原因の一つとして，初期特徴量の内容が挙げられる．提案手法においてノード表現を作成する際，作家ノードには活動領域の二値配列，美術機関ノードには美術機関が関わった活動領域の割合の配列を初期特徴量としている．この手法では活動領域を元にした情報しか与えていないため，推薦タスクにおいて活動領域をノード表現とした場合と比べて，有効なノード表現が作成できていないと考えられる．

5　ま　と　め

本研究では展覧会歴をベースにした知識グラフ構造を構築し，その分析を行うとともに，知識的特徴を GNN を用いて学習することで，美術作家に対して美術機関を推薦するシステムを実装した．現代美術業界のモデル化では，同じ展覧会への参加に基づいたネットワークの構築とクラスタリングを行い，展覧

会に基づくコミュニティが，作家の活動領域や出生年代に起因する活動年代的特徴を内包していることを示した．推薦タスクの実験では，エッジの重みとして用いる過去の展覧会のウィンドウサイズを大きくすることで予測精度が向上したが，ウィンドウサイズを $\tau = 10$ 以上大きくすることで，重みの値に偏りが生まれ，推薦結果には悪影響を与えることを示した．また，ノード表現を変化させた場合では，提案手法の方が，作家の活動領域の特徴量を用いた場合よりも予測誤差が小さくなったが，推薦結果の内容を確認すると作家の活動領域の特徴量を用いる方が，実際の展覧会を的中できていることを示した．

今後の展望として，美術機関や作家， 展覧会のさらに異なる情報をもとに特徴量を作成することによって，どのような情報がモデルの性能に影響を与えるかを検証する必要がある．特に今回の提案手法においては，ノード表現作成時に作家の活動領域の情報のみを用いていたため，今後は活動年代や出身など多様な情報を用いてノード表現の作成を行い，モデルの検証を行う必要がある．本研究で用いたデータセットは，現代美術業界を網羅できる作家や展覧会の数を有していないため，今後はより広範囲な作家や展覧会データを用いてモデル化とシステム実装を行っていきたい．また，今回は展覧会歴をベースにした知識的特徴を用いたが，現代美術業界において展覧会におけるキャリア以外にもオークションなどの異なるキャリアが存在するため，異なる情報をモデル化して検討を行う必要があると考えられる．現代美術領域に対してより広範囲な情報を検討し，適切なモデル化を行っていくことによって，作家のより具体的な戦略に役立てるための実装に取り組んでいきたい．

文　　献

[1] Samuel P. Fraiberger, Roberta Sinatra, Magnus Resch, Christoph Riedl, and Albert-László Barabási. Quantifying reputation and success in art. *Science (New York, N.Y.)*, Vol. 362 (6416), pp. 825-829, 2018.

[2] Athanasios Efthymiou, Stevan Rudinac, Monika Kackovic, Marcel Worring and Nachoem Wijnberg. Graph Neural Networks for Knowledge Enhanced Visual Representation of Paintings. *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3710-3719, 2021.

[3] Phillip Bonacich. Power and Centrality: A Family of Measures. *American Journal of Sociology*, Vol. 92, No. 5, pp. 1170-1182, 1987.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *The Computer Networks and ISDN Systems*, Vol. 30, pp. 107-117, 1998.

[5] WikiArt. VisualArtEncyclopedia. `https://www.wikiart.org/`, (accessed 2024-01-09).

[6] Gjorgji Strezoski and Marcel Worring. OmniArt: A Large-Scale Artistic Benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 14, No. 4, pp. 1-21, 2018.

[7] Gjorgji Strezoski, Inske Groenen, Jurriaan Besenbruch, and Marcel Worring. ArtSight: An artistic data exploration engine. *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1240-1241, 2018.

[8] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 1025-1035, 2017.

[9] Art Platform Japan. 日本アーティスト事典（ベータ版）. `https://artplatform.go.jp/ja/resources/collections/artists`, (accessed 2024-01-06).

[10] Art Platform Japan. 日本現代美術展調査. `https://artplatform.go.jp/ja/resources/research-projects/exhibitions`, (accessed 2024-01-06).

[11] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, p. 10008, 2008.

[12] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E69(026113)*, pp. 56-68, 2004.

[13] Matthias Fey and Jan E. Lenssen. Fast Graph Representation Learning with PyTorch Geometric. *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[14] Tokyo Art Beat. 日本全国の展覧会＆アートイベント最新情報一覧. `https://www.tokyoartbeat.com/events`, (accessed 2024-01-30).

繰り返し聴取による興味の変化を考慮した楽曲推薦

重富竜ノ介† 西田 紘子†† 澤井 賢一†† 牛尼 剛聡††

† 九州大学大学院芸術工学府 〒 815-8540 福岡市南区塩原 4-9-1
†† 九州大学大学院芸術工学研究院 〒 815-8540 福岡市南区塩原 4-9-1
E-mail: †shigetomi.ryunosuke.408@s.kyushu-u.ac.jp, ††{nishida,ken1,ushiama}@design.kyushu-u.ac.jp

あらまし 音楽は繰り返し消費されることの多いコンテンツであり，ユーザが過去に聴取した楽曲をどのような回数・頻度で推薦するかは重要な問題である．また, 音楽を含む刺激一般に対するヒトの快感情は，繰り返しによって逆 U 字型の曲線を描くことが知られている．このことは，ユーザの楽曲に対する興味が，繰り返し聴取によって動的に変化することを示唆している．しかし，既存の音楽推薦システムでは，このような繰り返し聴取による個々の楽曲への興味の変化は取り入れられていない．本研究では，楽曲の繰り返し回数から繰り返し埋め込み表現を作成し，楽曲埋め込み表現とセルフアテンション層内で関連付けることで楽曲の複雑な繰り返しパターンを学習する ReSAN を提案する．評価のために，音楽ストリーミングサービス Last.fm のユーザ聴取履歴を用いて実験を行った結果，ReSANs は他のセルフアテンションベースのシーケンシャル推薦システムを凌駕する性能をみせた．また，結果の分析によって，ReSANs は楽曲・順序だけでなく繰り返しの情報を考慮して楽曲推薦を行っていることが確認された．

キーワード 音楽推薦，シーケンシャル推薦，繰り返し推薦，心理学に基づく推薦システム

1 はじめに

近年の音楽ストリーミングサービスの普及により，ユーザには膨大な数のコンテンツが提供されるようになった．このような状況において，ユーザが自身の好みに適合する楽曲を発見するのは困難である．そこで，良質な楽曲推薦を行いユーザ満足度を向上させるために，様々な楽曲推薦手法が提案されている．代表的な楽曲推薦の手法には，協調フィルタリングを用いた手法 [1] や深層学習を活用した手法 [2], [3] がある．

また近年では，音楽聴取が最近の聴取とコンテクストの影響を強く受けることを利用して，ユーザの間近の聴取履歴をモデルの入力として, ユーザが次に聴取する可能性が高い楽曲を予測するシーケンシャル推薦システムが提案されている [4], [5], [6]．それらの手法では，再帰的ニューラルネットワーク（RNN）やセルフアテンションネットワーク（SANs）を利用して楽曲聴取の逐次的なパターンやアイテム間の関係を学習することで潜在表現を獲得し，適切な楽曲を推薦する．

また，楽曲は繰り返し消費される割合が高いコンテンツであり，ユーザの聴取履歴の 6 割から 7 割程度が繰り返し聴取であることが知られている [7], [8]．そのため，これまでにもユーザの繰り返し聴取に関する分析や推薦システムへの応用に関する研究が行われてきた．繰り返し消費とは，ユーザが過去に消費したアイテムを再び消費することである [9]．例えば，ある楽曲を聴取した後，数日に渡って同じ楽曲を何度も繰り返して聴取するような行動は代表的な繰り返し消費である．

そして，繰り返しに関する行動は，ヒトの認知特性と関連があることが知られている [8], [10], [11]．例えば，音楽を含む一般的な刺激に対するヒトの快感情は，繰り返して刺激を提示されることで，逆 U 字型に変化することが知られている [10]．この特徴を考慮すると，初めは快感情が低い楽曲であっても，繰り返して聴取することで快感情が増加し，ユーザにとっての価値が向上する可能性がある．あるいは，快感情が高く頻繁に聴取している楽曲は，繰り返し回数とともに徐々に飽きていき，ユーザにとっての価値が減少していく可能性がある．このようなヒトの楽曲に対する認知特性を考慮した推薦システムの重要性が主張されている [8], [11]．

Ren ら [12] は忘却曲線に従い，過去の聴取履歴ほど現在の嗜好への影響が少ないことを利用した重み付き協調フィルタリングを提案している．また，Moscati ら [13] や Sguerra ら [11] はヒトの記憶に関する認知特性をモデル化した Adaptive Control of Thought–Rational（ACT-R）[14] を利用して認知特性を考慮した繰り返し推薦手法を提案している．しかし，これらの手法では，コンテクストや複数の楽曲との相互作用により影響を受ける，ユーザの複雑な繰り返し聴取パターンを考慮できていないという問題点がある．

そこで本研究では，楽曲聴取シーケンスと楽曲の繰り返し回数をシーケンシャル推薦システムの入力として, ユーザの複雑な繰り返し聴取パターンを学習し, 楽曲の興味変化に沿った推薦を行うことを目的とする ReSANs（Repeat-aware Self-Attention networks）を提案する．通常の SANs で用いられるアイテム埋め込み表現の Query, Key ベクトルに対して繰り返し埋め込み表現の Query,Key ベクトルを関連付けることで，繰り返し聴取パターンを学習し，繰り返しによって変化するユーザにとっての楽曲価値をモデル化する．提案手法の有効性を評価するために，音楽ストリーミングサービスの Last.fm のデータセット [15] を用いて，ユーザの 1 週間分の聴取履歴における，最後のセッションの 2 曲目以降の楽曲を予測する実験を行い手法の有効性を確認した．ここで，セッションとは聴取間隔が 30 分

以内の連続した楽曲聴取の集合である．

本論文では，2 章で先行研究と本研究の関連性について述べた後，3 章で提案手法について説明する．次に，4 章で実験方法について説明し，5 章で結果と考察について述べる．最後に 6 章でまとめと今後の展望について言及する．

2 関 連 研 究

これまでにもシーケンシャル推薦システムや繰り返しとヒトの認知特性の関係と楽曲聴取への影響について研究が行われてきた．2.1 節では，シーケンシャル推薦システムと楽曲推薦への応用について，2.2 節では繰り返し聴取とヒトの認知特性の関係，及びその推薦への応用について先行研究を紹介し，本研究との関係を述べる．

2.1 シーケンシャル推薦システムと楽曲推薦へ応用

ユーザの消費傾向は間近の消費から強い影響を受けることが知られている [4], [7], [9]．そのため，音楽を含めた一般的な推薦タスクにおいて，ユーザの聴取履歴の系列を RNN やセルフアテンションネットワークの入力として，アイテム間の逐次的なパターンや関係性を学習し，ユーザが次に消費するアイテムを予測する取り組みが行われてきた．

GRU4Rec [16] は，RNN を用いたシーケンシャル推薦システムであり，ユーザの消費シーケンスから RNN によって隠れ状態を獲得し，次の消費アイテムを予測する．

RepeatNet は [17] は RNN を用いたシーケンシャル推薦システムを応用し，繰り返し消費を明示的に扱うことができる手法である．この手法では，まず入力系列を RNN を用いてエンコードする．その後，*repeat mode* と *explore mode* によって新規コンテンツの推薦と過去に消費したことのあるコンテンツの予測をそれぞれ行う．そして，*repeat-explore mechanism* によって *repeat mode* と *explore mode* の実行確率を求めることで，新規コンテンツと過去に消費したことのあるアイテムを明示的に考慮した推薦を実現している．

セルフアテンションネットワークを利用した代表的なシーケンシャル推薦システムには，SASRec [18] と LightSANs [19] がある．SASRec [18] は，SANs を用いて入力系列中のアイテム間の関係性を学習し，最終的な系列の状態を獲得する．この手法は Transformer [20] に用いられるセルフアテンション機構を推薦に応用し，推薦タスクにおいても優れた性能を達成した．LightSANs [19] は，SASRec をベースとして，セルフアテンション層の Key,Value ベクトルをユーザの興味を表す低ランク表現に変換したモデルであり，入力系列中のアイテム間の関係性をより効果的かつ効率的に学習できる．

シーケンシャル推薦システムを楽曲推薦に応用した代表的な手法には，CoSeRNN [4] がある．セッションと呼ばれる連続した聴取の集合を定義し，あるセッションに含まれる楽曲集合をセッション埋め込み表現とする．このセッション埋め込み表現とともに聴取時間帯やデバイスなどのコンテキスト情報を LSTM（Long Short-Term Memory）に入力し，次のセッションに含まれる楽曲を予測する．

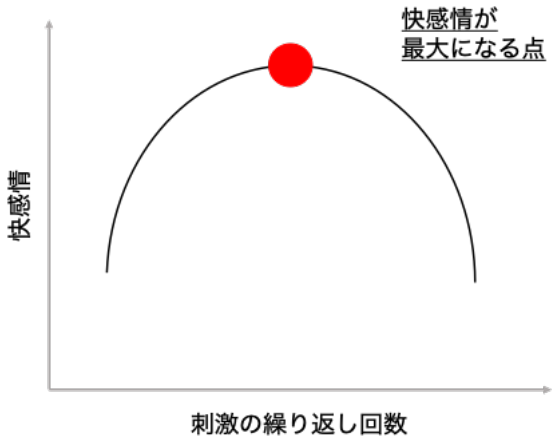


図 1　繰り返し刺激による快感情の逆 U 字型変化

これらのモデルは，繰り返しによる楽曲の興味の変化を考慮しておらず，与えられた系列においてユーザにとって興味の高い楽曲，低い楽曲を考慮した推薦を行うことに焦点を当てていない．それに対して，本研究では繰り返し回数から獲得した繰り返し埋め込み表現を楽曲埋め込みに関連づけることでシーケンシャル推薦システムに繰り返しパターンを統合し，ユーザの楽曲に対する興味が変化することを考慮した推薦手法を提案する．

2.2 繰り返し聴取とヒトの認知特性の関係，及びその推薦への応用

音楽は繰り返し消費されることは上で述べたが，インターネットコマースや特定の目的地などでも繰り返し消費の存在が確認されている [9], [7]．Anderson ら [9] はソーシャルメディアのログイン，Wikipedia の閲覧，YouTube の視聴など，様々な領域での繰り返し消費を分析した．その結果，最後の消費からの経過時間が繰り返し消費の最も強い予測因子であることを明らかにした．Benson ら [7] は，繰り返し消費するアイテムには飽きと寿命があることを明らかにした．繰り返し消費するアイテムでも，その消費間隔は徐々に大きくなり最後の消費では最初の間隔に比べて 3〜5 倍の差があることを確認した．

また，音楽を含めた刺激に対する一般的な認知特性として，繰り返し刺激を提示されると，その刺激に対する快感情が逆 U 字型に変化することが知られている [10]．図 1 のように刺激を何度も繰り返し提示することでその刺激に対する快感情が増加していく．さらに刺激を繰り返すことである閾値を迎えると快感情は最大値を迎え，その後は減少していく．この曲線は刺激の不確定さによって変化し，一般的には，より不確定さが大きな刺激ほど曲線の最大値へ達するまでに時間を要する [21]．

Sguerra ら [11] は，この快感情の逆 U 字型変化の影響が，音楽ストリーミングサービスのユーザ聴取履歴に現れていることを示した．ユーザがサービスプラットフォーム上で初めて聴取する楽曲を対象とした時，楽曲を総再生時間の 8 割以上聴取する確率は聴取回数に対して逆 U 字型に変化することを明らか

にした．

そこで，このようなユーザの心理学的な側面に焦点を当てた推薦を行うことの重要性が指摘されている [8], [11].

L.Ren [12] が提案した，time enhanced collaborative filtering では，協調フィルタリングのユーザ間類似度及びアイテム間類似度の計算時，コンテンツの最後の消費からの経過時間によってバイアスをかけることで，忘却曲線を活用した推薦システムを提案している．例えば，あるユーザから高評価を得ているコンテンツでも，その消費から現在までの経過時間が長くなるほど強いバイアスがかけられ，低い評価値として計算に利用される．

Haas ら [8] が提案した認知アーキテクチャ ACT-R を用いた推薦システムは，ヒトの記憶に関する認知特性をモデル化した推薦システムである．この手法では，「間近に頻繁に聴取している」，「間近の聴取楽曲と共起している」，「間近の聴取楽曲と類似度が高い」等の特徴の楽曲が繰り返し聴取されやすいことを利用している．しかし，繰り返し楽曲のみを考慮しているため，推薦において重要な新規楽曲の推薦を行うことができない．

Moscati ら [13] は，ACT-R を協調フィルタリングベースの推薦モデルに統合し，認知特性を考慮した繰り返し推薦と新規楽曲の推薦を同時に可能な手法を提案した．楽曲 k の次の聴取を予測するために，協調フィルタリングによって計算された楽曲 i, k の類似度に対して，ACT-R に基づいて算出された楽曲 i の活性化スコアを加算することで，認知特性を考慮した推薦スコアを算出している．

また，Sguerra ら [11] は，繰り返しによる快感情の逆 U 字型変化を考慮して，ユーザと楽曲のベクトル表現の距離を繰り返し回数と頻度によって変化させる Ex2Vec を提案している．ユーザ及び楽曲の埋め込み表現を基本距離として，ACT-R の Base-level 活性化 $B_{u,i}(t) = \sum_{j=1}^{n}(t-t_j+c)^{-d}$ に従って時刻 t の距離を調整する．ここで，n はユーザが楽曲 i を繰り返し聴取した回数，d は減衰パラメータ，c は学習可能なパラメータである．そして，調整された距離 $d_{u,i}(t)$ を $\alpha d_{u,i}(t) + \beta d_{u,i}(t)^2 + \gamma + b$ の 2 次関数従うユーザの興味のスコアに変換した後，sigmoid 関数を通し，時刻 t におけるユーザの楽曲 i の聴取確率を得る．ここで，α, β, b は学習可能なパラメータである．

これらの手法では，ヒトの認知特性を明示的にモデル化し，繰り返しによる楽曲の興味の変化を考慮した推薦を行っている．しかし，コンテクストや楽曲間の複雑な相互作用を学習していないため，その性能は従来の推薦手法を下回っている．本研究では，忘却曲線や逆 U 字型曲線を明示的にモデル化するのではなく，セルフアテンションネットワークの構成要素として学習可能な Repeat-Query, Repeat-Key ベクトルを作成することで，コンテクストや楽曲間の相互作用を考慮した複雑な繰り返し聴取パターンを学習する ReSANs を提案する．

3 提案手法

提案手法の概要図を図 2 に示す．繰り返し回数から Repeat-Query, Repeat-Key ベクトルを作成し Item-Query, Item-key

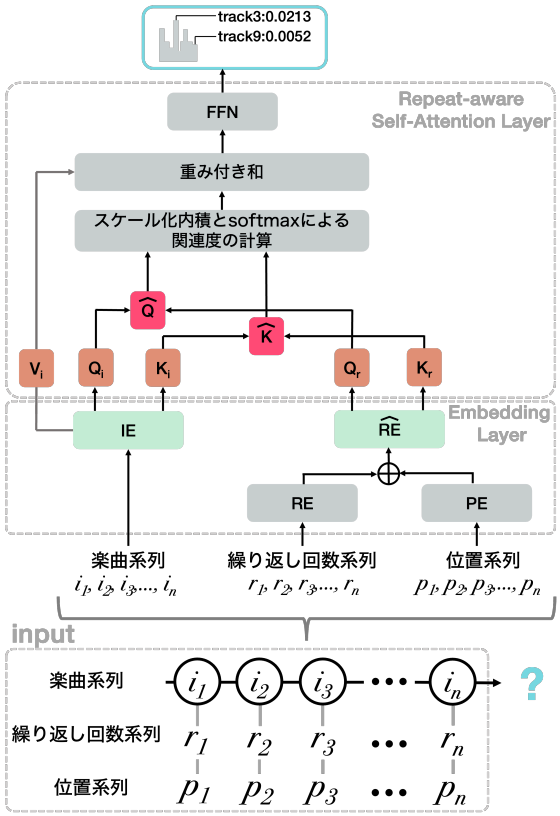


図 2　提案手法の概略図．Q_i, K_i, Q_r, K_r がそれぞれ $Item-Query, Item-Key, Repeat-Query, Repeat-Key$ を表す．

ベクトルと関連付けることでユーザにとっての各楽曲の繰り返し回数を考慮したアテンションスコアを計算する．

学習時には，入力をユーザの 1 週間の聴取履歴 $\{i_1, i_2, i_3, ..., i_{t-1}\}$ とし，各時刻の楽曲の次の楽曲 $\{i_2, i_3, i_4, ..., i_t\}$ を予測することでパラメータの更新を行う．ここで，i は楽曲である．

3.1.1 及び 3.1.2 節で，提案手法の主な貢献である通常のアイテム埋め込みに加えて繰り返し埋め込みを導入すること，及びそれらをセルフアテンション層内で関連づける方法を説明する．3.1.3 節で，セルフアテンション機構で獲得した知識について蓄積・抽出を行う Feed forward 層を，3.2 節でモデルが予測を行う方法と損失関数について説明する．

3.1 繰り返し埋め込みを取り入れた self-attention

3.1.1 楽曲・繰り返し・位置の埋め込み

本手法では SASRec [18] と同様に，モデルに入力するシーケンス ss は固定長 n であり，1 週間分のユーザ聴取履歴 SS の長さが n よりも大きい場合，最も直近の n 件の聴取履歴を入力とする．SS の長さが n よりも小さい場合，0 を用いてパディングを行う．

ss から d 次元の楽曲埋め込み系列 $\mathbf{IE} \in R^{n \times d}$ を作成する．ここで $\mathbf{IE}$ は，全楽曲の楽曲埋め込み表現が格納された行列 $\mathbf{IM} \in R^{I \times d}$ から，入力シーケンス ss に含まれる楽曲 id を検索して得られたものである．

次に，アイテム埋め込み表現と同様の方法で，繰り返し埋め込み表現系列 $\mathbf{RE} \in R^{n \times d}$ を作成する．シーケンス内の位置情報を考慮するために，学習可能な d 次元の位置埋め込み表現系列 $\mathbf{PE} \in R^{n \times d}$ を作成し，(1) 式によって繰り返しの位置を表現するベクトル系列 $\widehat{\mathbf{RE}}$ を取得する．

$$\widehat{\mathbf{RE}} = \mathbf{RE} + \mathbf{PE} \tag{1}$$

得られた 2 つの埋め込み表現系列 $\mathbf{IE}$ と $\widehat{\mathbf{RE}}$ を次の repeat-aware self-attentnion 層の入力とする．

3.1.2 繰り返しを考慮したセルフアテンション層

$\mathbf{IE}$ 及び $\widehat{\mathbf{RE}}$ を線形変換を通して，Query, Key, Value ベクトルに変換する．

$$\mathbf{Q}_i = \mathbf{IE} \cdot \mathbf{W}^{Q_i}, \tag{2}$$
$$\mathbf{K}_i = \mathbf{IE} \cdot \mathbf{W}^{K_i}, \tag{3}$$
$$\mathbf{V}_i = \mathbf{IE} \cdot \mathbf{W}^{V}, \tag{4}$$
$$\mathbf{Q}_r = \widehat{\mathbf{RE}} \cdot \mathbf{W}^{Q_r}, \tag{5}$$
$$\mathbf{K}_r = \widehat{\mathbf{RE}} \cdot \mathbf{W}^{K_r} \tag{6}$$

ここで，$\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}, \mathbf{W}^{Q_r}, \mathbf{W}^{K_r} \in R^{d \times d}$ である．そして，楽曲の繰り返し聴取パターンを考慮した入力系列間の関係性を学習するために，以下のように $\mathbf{IE}$ と $\widehat{\mathbf{RE}}$ のそれぞれの Query, Key ベクトルの和を計算し，その後セルフアテンションの計算を行う．

$$\mathbf{S} = SA(\mathbf{IE}, \widehat{\mathbf{RE}}) = \text{softmax}\left(\frac{(\mathbf{Q_i} + \mathbf{Q_r})(\mathbf{K_i} + \mathbf{K_r})^T}{\sqrt{d/h}}\right)\mathbf{V_i} \tag{7}$$

従来のセルフアテンションベースの手法において，アイテム i とアイテム j の間の関連度は，入力系列の楽曲・順序に依存して決定される．一方提案手法では，繰り返しの頻度を考慮することで，i が 5 回聴かれた場合の j との関連度は，i が 1 回聴かれた場合の j との関連度は異なるものとして計算可能である．さらに，アテンション行列は動的であり，繰り返し回数に基づいて適応する柔軟性の高い推薦システムを実現する．例えば，アイテム i が文脈的に最も適切な推薦であるように見えても，アイテム i が対象ユーザによって過剰に繰り返し聴取され，興味が飽和していることを考慮することで，別のアイテム j の推薦スコアを上昇させることができる．

また，提案手法の ReSANs では Value ベクトルのみ楽曲の Value ベクトルをそのまま使用している．これは Liu らの提案した NOVA [22] と同様に，最終的な推薦スコアを得る際，モデルの出力 F_t と楽曲埋め込み IE の内積を取るためである．もし Value ベクトルに繰り返しの情報を統合した場合，異なる潜在空間上で F_t と IE_i の内積を計算することになる．これを防ぎ，同じ潜在空間上でベクトルの内積を取るため，楽曲のみから作成された純粋な Value ベクトルが用いられる．

3.1.3 フィードフォワードネットワークとセルフアテンション層の積み重ね

モデルに非線形性を持たせ，各トークンの最終的な出力を得るために，全結合フィードフォワードネックワーク層を追加する．

$$\mathbf{F} = \text{FFN}(\mathbf{S}) = \text{ReLU}\left(\mathbf{S}\mathbf{W}_{(1)} + \mathbf{b}_{(1)}\right)\mathbf{W}_{(2)} + \mathbf{b}_{(2)} \tag{8}$$

ここで，$\mathbf{W}_{(1)}, \mathbf{W}_{(2)} \in R^{d \times d}$ であり，b_1, b_2 は d 次元ベクトルである．フィードフォワードネックワークはセルフアテンションで獲得したトークン間の関連性に関する知識を蓄積する場所と考えられており [23]，この場合繰り返しを考慮したトークン間の関係性を知識として蓄積・抽出することを目的としている．

$\mathbf{F}$ は Transformer [20] と同様に残差結合，レイヤー正規化及びドロップアウト層を通過して，次のブロックに渡される．b 個目のブロックは，

$$\begin{aligned}\mathbf{S}_b &= \text{SA}\left(\mathbf{F}^{(b-1)}\right), \\ \mathbf{F}_b &= \text{FFN}\left(\mathbf{S_b}\right)\end{aligned} \tag{9}$$

となる．

3.2 次聴取楽曲の予測と損失関数

時刻 t+1 の楽曲 i_{t+1} を予測するため，時刻 t までの楽曲を入力系列とした時の最後の出力層 $\mathbf{F_t}$ を用いる．ある楽曲 i に対するモデルの予測値は，$\mathbf{F_t}$ と i の埋め込み表現 $\mathbf{IE_i}$ を用いて，

$$\Pr\left(i_{t+1} \mid \{i_1, \ldots, i_t\}\right) = \langle \mathbf{F}_t, \mathbf{E}_i \rangle \tag{10}$$

で計算される．

学習の際は，バイナリクロスエントロピーを損失関数として用いて学習を行う．

$$\mathcal{L} = r_{t+1,g} \log\left(\sigma\left(r_{t+1}\right)\right) + \left(1 - r_{t+1,g}\right) \log\left(1 - \sigma\left(r_{t+1}\right)\right) \tag{11}$$

$r_{t+1,g}$ は時刻 $t+1$ におけるグランドトゥルースの楽曲である．r_{t+1} は時刻 $t+1$ におけるモデルの予測であり，

$$r_{t+1} = \arg\max \sum_{i=1}^{I} \langle \mathbf{F}_{t+1}, \mathbf{E}_i \rangle \tag{12}$$

で計算される．

4 実　　験

4.1 データセットと前処理

提案手法の有効性を評価するため，音楽ストリーミングサービス LAST.FM のユーザの聴取履歴からなる LFM-2b Dataset[1] [15] を用いて実験を行った．このデータセットは，

1：http://www.cp.jku.at/datasets/LFM-2b/

120,322 人のユーザの 2005 年 2 月 14 日から 2020 年 3 月 20 日までの音楽聴取履歴で構成されており，ユーザの年齢・性別・国などの情報が含まれている．

本研究では，2020 年 1 月 1 日から 2020 年 3 月 20 日までのユーザ聴取履歴を利用し，データセットには userId， trackId, timestamp, ユーザ u の楽曲 i の繰り返し回数を利用する．

データの偏りを防ぐため，前処理としてユーザと楽曲のフィルタリングを行った．まず，総聴取回数が全体の 99 %以上であるユーザを除外した．また，ユーザ u の楽曲 i に対する総繰り返し回数が全体の 99 %以上であるレコードを除外した．さらに，極端にインタラクションの少ないユーザや楽曲をフィルタリングするために，ユーザについては楽曲へのインタラクションが 100 以上を対象とし，楽曲ついてはユーザからのインタラクションが 100 以上を対象とした．

そして，ユーザが聴取した楽曲のみをモデルの入力及び予測の対象とするために，楽曲聴取の timestamp から楽曲聴取時間を計算し，ユーザが 30 秒以上聴取している楽曲のみを用いた．

次に，ユーザの聴取シーケンスのうち，楽曲間の聴取時間間隔が 30 分以内の楽曲の集合をセッション s として定義し，セッションに含まれる楽曲数が 1 つのみであればそのセッションを除外した．

3 節で述べたように，モデルへの入力はユーザの 1 週間分の聴取履歴である．データセット内でユーザが最後の楽曲とのインタラクションを行った timestamp を基準に 1 週間ごとのセッションの集合であるセッションセット ss とした．セッションセット ss，セッション s，楽曲 i の関係は以下のようになる．

$$ss_t = \{s_1, s_2, \ldots, s_n\}, \tag{13}$$

$$s_n = \{i_1, i_2, \ldots, i_m\} \tag{14}$$

以上の前処理を行った結果，対象とするデータはユーザ数 13,171 人，楽曲数 33,797 曲，レコード数 7,250,512 であった．

4.2 評　　価

検証，及びテストではユーザの 1 週間分の聴取履歴 ss のうち最後のセッション s_{last} の 1 曲目までをモデルに入力し，s_{last} の 2 曲目以降を予測する．s_{last} の 3 曲目以降を予測するときには，2 曲目の予測スコアが高い順に推薦する方法をとった．

4.2.1 評 価 指 標

本実験で用いる評価指標について以下で説明する．

- **R-Precision:** 予測シーケンスのアイテム集合がが正解のシーケンスのアイテム集合にどの程度適合しているかを表す [8], [24].

$$R\text{-}prec = \frac{1}{|SS|}\sum_{ss}\frac{|R \cap G|}{|G|} \tag{15}$$

ここで SS は全セッションセット，G が ss の正解データ集合，R が正解データ集合と同じ長さを持つ（$|S| = |G|$）予測アイテム集合である．R は推薦タスクの評価でよく用いられる topK の推薦アイテムとは異なり，正解シーケンスの各アイテムに対して予測を行った結果に相当する．

- **Recall@K:** topK の予測アイテム集合が正解データ集合にどの程度適合しているかを表す．

$$Recall@K = \frac{1}{|SS|}\sum_{ss}\frac{|R_k \cap G|}{|G|} \tag{16}$$

ここで，R_k は topK の予測アイテム集合である．

- **MRR@K:** 予測アイテムが正解データ集合に含まれるアイテムをどれだけ高い順位にランキングできたかを表す．

$$MRR@K = \frac{1}{|SS|}\sum_{ss}\frac{1}{|k|} \tag{17}$$

ここで k は，はじめて適合したアイテムの topK の予測アイテム集合における順位である．

上記のすべての指標において，値が高いほど推薦の性能が高いことを意味する．

4.2.2 比 較 手 法

提案手法の有効性を確認するため，以下に示す先行研究のモデルとの比較を行った．

- **SASRec [18]:** Self-Attention を利用したシーケンシャル推薦システムであり，本研究はこのモデルをベースとしている．
- **SASRec+R:** SASRec のアイテム埋め込みに対して，単純に繰り返し埋め込みを加算した手法である．
- **LightSANs:** SASRec を改良し，低ランク分解自己注意を導入することで，ユーザの消費したアイテムを少数の潜在興味空間に投影し，アイテムと興味の相互作用を利用する手法である．ハイパーパラメータである潜在興味の数は，著者実装 [2] を参考に 10 として実験を行った．

これらの手法は，self-attention 機構をベースにしており，入力系列の楽曲・順序に適した好ましい楽曲を推薦することが可能である．実験によって，ReSANs において，楽曲・順序に加えて繰り返しの情報を考慮することの有効性を確認する．

4.3 学習及びパラメータ

全ユーザについて，最後の 1 週間の聴取履歴 ss_{last} をテスト，ss_{last-1} を検証データ，それ以前の聴取データを学習データとしてモデルを学習した．

提案手法を含めたすべてのモデルでバッチサイズ=128, エポック数=200 で学習を行った．埋め込み表現の潜在因子数は $d = 64$ であり，FFN 層の中間層は $d_{inner} = 256$ である．入力長の長さは 100 とした．また，アテンションヘッド数とブロック数はともに 2 とした．ネットワークの最適化には Adam [25] を用いた．

5 結果と考察

5.1 結　　果

表 1 に実験結果を示す．全ての評価指標において，提案手法である ReSANs は，比較手法を上回る性能を示した．そして，評価指標が 2 番目に高いモデルであった SASRec に比べて

2：https://github.com/BELIEVEfxy/LightSANs

表 1　各モデルの推薦性能の評価．各評価指標において，太字が最も高い性能で下線が 2 番目に高い性能を示す．

	R-Precision	Rcall@10	Rcall@20	MRR@10	MRR@20
SASRec	0.0820	0.134	0.187	0.153	0.159
SASRec+R	0.0771	0.130	0.184	0.147	0.153
LightSANs	0.0391	0.0668	0.0947	0.0826	0.0864
ReSANS	**0.111**	**0.182**	**0.233**	**0.200**	**0.205**
Improvement vs SASRec+R	35.0%	35.4%	24.5%	30.8%	28.8%

20-30 %ほど高い性能向上が確認された．

SASRec+R については，3 番目に高い性能となり，すべての指標において SASRec とほぼ同値となっている．これは，セルフアテンション層より前の段階で繰り返し埋め込みが導入された場合，モデルが効果的に繰り返し埋め込み情報を学習できない可能性を示唆している．

ReSANs は，楽曲埋め込みと繰り返し埋め込みの統合をセルフアテンション層の内部で行うことでこの問題を解決している．楽曲の Query,Key ベクトルに加えて，Repeat-Query,Repeat-Key を用いることによるモデルの表現力が強化され繰り返しパターンを考慮した学習が行われたと考えられる．

次に，Recall@K の SASRec からの改善率について，K=10 が K=20 よりも高く，Recall@10 においては 35.4 %の改善率を見せた．これは，提案手法がユーザの好む楽曲集合をより高順位に予測できていることを示しており，ランキング指標である MRR@k の改善率が高いことにもその影響が現れている．

また, topK の予測楽曲ではなく, ユーザの実際の聴取シーケンスと同じ長さだけ推薦を行なった場合の性能を示す R-Precision において，35.0 %程度の性能向上が確認できる．このことから，提案手法はあるセッションにおいて，2 曲目以降を自動継続するという現実的なタスクにおいても，従来手法より高い性能を示すことが確認された．

5.2　アブレーションスタディ

ReSANs の各構成要素が性能にどの程度寄与しているか考察するために，特定の要素をモデルから除外した場合の性能を調査した．ReSANs において重要な，繰り返し埋め込み表現と位置埋め込み表現について，モデルの埋め込み表現の次元数 $d = 64, 512$ の 2 パターンで調査を行った．

表 2，3 に結果を示す．全体的に，埋め込み表現の次元数が $d = 64$ の場合より，$d = 512$ の場合の方が繰り返し・位置ともに埋め込み表現の除外の影響を強く受けた．これは，次元が大きくなったことによりモデルの表現力が向上し，楽曲・位置・繰り返しのそれぞれに対して十分にパターン学習が行われた結果，各埋め込み表現のモデルへの寄与が大きくなったためだと考えられる．

$d = 64$ においては，ほぼすべての指標について繰り返し埋め込みの除外と位置埋め込みの除外ともに性能が向上している．5.1 節において，SASRec と比べて大幅に性能が向上していることから，楽曲・位置・繰り返しのパターンについてある程度効果的に学習が行われていることは確認されている．しかし，$d = 64$ では次元数が小さいため，データとモデル構造に対して次元数が十分ではなく，トークン間の複雑な相互作用を学習しモデル内に蓄積することが難しかったことが考えられる．そのため，繰り返し埋め込み・位置埋め込みのどちらかが除外されることで，モデルがよりシンプルになって効率的に学習が進み，性能が向上したと考えられる．

一方，$d = 512$ の場合，すべての指標について繰り返し・位置埋め込みの除外ともに性能が低下している．これは，モデルの表現力が向上し，楽曲・位置・繰り返しのそれぞれに対して十分にパターン学習と知識の抽出が行われた結果だと考えられる．性能の低下率は位置埋め込みの除外のほうが大きく，ReSANs が予測を行う際，入力シーケンスの位置・順序情報が重要であると考えられる．同時に，繰り返し埋め込みの除外についても性能の低下が確認でき，ReSANs は入力シーケンスの繰り返し情報にも注意を向けて予測を行っていることがわかる．この結果から，ReSANs は，位置埋め込みと繰り返し埋め込みの両方を用いることによって最も性能を発揮している．そのため，位置埋め込みから得られる順序情報と繰り返し埋め込みから得られる繰り返し回数を統合することで，繰り返しに関する回数・頻度などの複雑なパターンを効果的に学習し，楽曲情報と合わせて次聴取楽曲を予測していることが示唆された．

5.3　埋め込み表現の次元数の性能への影響

5.2 節において，ReSANs の性能が次元数に大きく影響を受けることが示された．本節では，埋め込み表現の次元数の性能への影響について ReSANs と SASRec を比較することで，提案手法の特性を考察する．

具体的には，埋め込み表現の次元数を変化させた場合の R-Precision の変化を分析し，その影響を調査する．R-Precision は正解データと同じ長さの推薦リストを作成しその性能を評価するものである．そのため，ユーザがセッションのはじめに 1 曲のみ楽曲を選択した後，再生を自動継続するなどの現実的なタスクの評価として妥当であり，本提案手法の評価を代表する指標として最も適切である．なお，提案手法のベースとなる self-attention 機構には FFN 層と呼ばれる層があり，この層の次元数はアイテム埋め込みの 4 倍とされることが慣例である．そのため，今回の分析でも埋め込み表現の次元の増加に伴って，FFN 層の次元数も 4 倍ずつ増加させた．

図 3 に実験結果を示す．ReSANs と SASRec はともに，次元

$d = 64$

	R-Precision	Rcall@10	Rcall@20	MRR@10	MRR@20
ReSANs	0.111	0.181	0.233	0.200	0.205
繰り返し埋め込みの除外	0.115 (+4.21%)	0.178 (**-1.70%**)	0.233 (+0.08%)	0.198 (**-1.21%**)	0.203 (**-0.96%**)
位置埋め込みの除外	0.114 (+3.13%)	0.186 (+2.75%)	0.241 (+3.25%)	0.203 (+1.62%)	0.208 (+1.69%)

表 2　$d = 64$ における，繰り返し埋め込み及び位置埋め込みに対するアブレーションスタディの結果

$d = 512$

	R-Precision	Rcall@10	Rcall@20	MRR@10	MRR@20
ReSANs	0.164	0.276	0.335	0.258	0.262
繰り返し埋め込みの除外	0.150 (**-8.37%**)	0.248 (**-9.98%**)	0.309 (**-7.81%**)	0.248 (**-3.80%**)	0.252 (**-3.67%**)
位置埋め込みの除外	0.123 (**-24.7%**)	0.199 (**-27.9%**)	0.251 (**-25.3%**)	0.208 (**-19.4%**)	0.212 (**-19.2%**)

表 3　$d = 512$ における，繰り返し埋め込み及び位置埋め込みに対するアブレーションスタディの結果

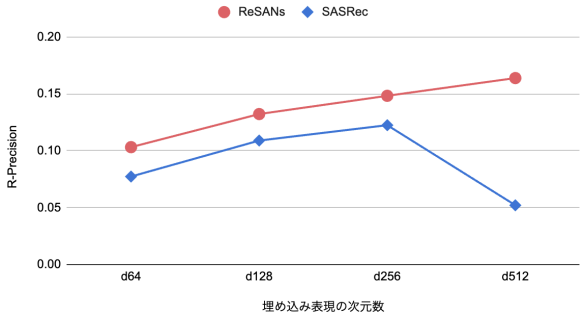


図 3　次元数に対する R-Precision の変化

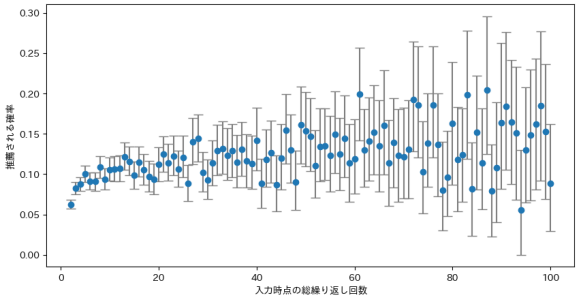


図 4　入力シーケンスの繰り返し回数に対する推薦確率の変化

数が 64 から 256 の間，次元数の変化に対してほぼ比例して性能が向上した．しかし，$d = 512$ においては ReSANs の性能のみが向上し，SASRec の性能は低下した．この実験結果により，512 次元の比較的高次元の埋め込み表現であっても，ReSANs の性能が SASRec に比べて高いことが確認された．

SASRec については，アイテム埋め込みと位置埋め込みから入力シーケンスのパターンを学習しているが，繰り返しの情報が存在しないため 512 次元の埋め込み表現は大きすぎ，モデルが過学習を起こしている可能性が考えられる．

一方 ReSANs は，楽曲・繰り返し・位置埋め込みの合計 3 つの埋め込みを作成し，それらをセルフアテンション層で統合するという特殊な構造をもっていることから，次元数の増加に伴ってモデルの表現力を向上させ，入力シーケンスのパターンについての学習を深めることができたと考えられる．

5.4　推薦リストの特徴と繰り返し聴取パターンの学習

本節では，提案手法によって出力された推薦リストを分析することで，モデルが繰り返し聴取パターンをどのように学習しているかについて分析を行う．対象とするモデルは，本研究を通して最も高い性能を示した 512 次元の埋め込み表現を持つ ReSANs である．

まず，推薦リストの繰り返し楽曲の割合は 36.3%であった．推薦リストの 7 割弱の楽曲が新規楽曲であり，ReSANs が繰り返しを考慮しつつもユーザの音楽聴取体験にとって重要な新規楽曲の推薦を行えていることを示している．

次に，入力された楽曲の繰り返し回数が推薦に影響を与えるか調べるため，入力シーケンス中の楽曲がモデルによって再度推薦される確率が，繰り返し回数ごとにどのように変化するのかを調査した．入力シーケンスに繰り返しが含まれる楽曲のみを対象とし，ある楽曲について入力に含まれる最大の繰り返し回数だけを計算に使用した．また，推薦リストの長さは R-Precision と同じく正解データ集合と同じ長さとした．結果を図 4 に示す．図中のバーは 95%信頼区間を表している．

この結果から，特に繰り返し回数が 2 回から 40 回付近までにおいて，繰り返し回数が増えるほど，繰り返し推薦の確率が上昇していることがわかる．これは，入力シーケンス中に含まれる楽曲の繰り返し回数が大きいほど，モデルが同じ楽曲を繰り返して推薦しやすいことを示している．

また 40 回の繰り返し以降は，推薦される確率のばらつきが大きくなっており，モデルが入力シーケンスの他の要素である楽曲や位置から得られる頻度の情報を考慮して推薦スコアを調整している可能性が示唆された．

5.5　提案手法の分析結果のまとめ

本節では，アブレーションスタディ，次元数の影響の調査，推薦リストの分析によって ReSANs の特徴を分析した．

結果として，5.2 節と 5.3 節により，埋め込み表現の次元数が比較的大きい 512 次元において，楽曲・繰り返し・位置のパターンを効果的に学習していることが確認された．また，5.4 節の推薦リストの分析から，入力シーケンスである 1 週間のユーザ聴取履歴中で，繰り返して聴取されている楽曲は，その繰り返し回数が多いほどモデルによって再度推薦されやすいことが明らかになった．

6 おわりに

本研究では，繰り返し聴取のパターンを学習し，ユーザにとっての楽曲の興味変化をモデル化することを目的とした Repeat-aware Self-Attention Networks（ReSANs）を提案した．音楽ストリーミングサービスの大規模なデータセットを用いて評価実験を行ったところ，他のセルフアテンションベースのシーケンシャル推薦モデルに比べて 20-30 %程度の性能向上が認められた．モデルの特徴を分析した結果，提案手法は入力シーケンスの楽曲・繰り返し・順序の関係性を学習し，次聴取楽曲を推薦していることが確認された．

本報告の限界として，ReSANs の繰り返しによる興味の細かい変化を分析できていない点が挙げられる．推薦リストを分析したところ，入力シーケンスの繰り返し回数が増えるほど繰り返して推薦しやすい傾向が確認された．しかし，ユーザの楽曲関する興味の変化はより複雑であり，例えば，ある楽曲に熱中する期間があればと飽きる期間も存在する．また，長期間聴取していなかったにも関わらず，何らかのきっかけで興味が回復することもある．今後の研究により，そのような細かい興味の変化をモデルが学習しているかについて，分析を行う必要がある．

将来的には，提案手法にヒトの心理学的な特性を統合することで，より効果的に興味の変化を学習した推薦システムを構築できる可能性がある．このようなシステムが実際の音楽ストリーミングサービスなどに導入されることによって，ヒトの心理学的な特性に寄り添った，より人間中心的な技術の普及が期待される．

文　献

[1] M. Schedl, H. Zamani, C. Chen, Y. Deldjoo, M. Elahi, "Current challenges and visions in music recommender systems research," International Journal of Multimedia Information Retrieval, vol. 7, pp. 95 – 116, 2018.
[2] A. van den Oord, S. Dieleman, B. Schrauwen, "Deep content-based music recommendation," in Proc. of the 27th Annual Conf. on Neural Information Processing Systems 2013 (NIPS 2013), 2013.
[3] S. Oramas, O. Nieto, M. Sordo, X. Serra, "A deep multimodal approach for cold-start music recommendation," in Proc. of the 2nd Workshop on Deep Learning for Recommender Systems (DLRS@RecSys 2017), 2017.
[4] C. Hansen et al., "Contextual and Sequential User Embeddings for Large-Scale Music Recommendation," in Proc. of the Fourteenth ACM Conf. on Recommender Systems (RecSys '20), 2020.
[5] G. Fazelnia et al., "Variational User Modeling with Slow and Fast Features," in Proc. of the Fifteenth ACM Int. Conf. on Web Search and Data Mining (WSDM '22), 2022
[6] W. Bendada et al., "Track Mix Generation on Music Streaming Services using Transformers," in Proc. of the Seventeenth ACM Conf. on Recommender Systems (RecSys '23), 2023.
[7] A. R. Benson, R. Kumar, A. Tomkins, "Modeling user consumption sequences," in Proc. of the 25th Int. Conf. on World Wide Web (WWW' 16), 2016.
[8] M. Haas et al., "Predicting Music Relistening Behavior Using the ACT-R Framework," in Proc. of the Fifteenth ACM Conf. on Recommender Systems (RecSys ' 21), 2021.
[9] A. Anderson et al., "The dynamics of repeat consumption," in Proc. of the 23rd Int. Conf. on World Wide Web (WWW' 14), 2014.
[10] D. Berlyne, "Aesthetics and psychobiology," J. of Aesthetics and Art Criticism, pp. 553-553, 1971.
[11] B. Sguerra, V.-A. Tran, R. Hennequin, "Ex2Vec: Characterizing Users and Items from the Mere Exposure Effect," in Proc. of the 17th ACM Conf. on Recommender Systems (RecSys '23), 2023.
[12] L. Ren, "A Time-Enhanced Collaborative Filtering Approach," 2015 4th Int. Conf. on Next Generation Computer and Information Technology (NGCIT), pp. 7-10, 2015.
[13] M. Moscati et al., "Integrating the ACT-R Framework with Collaborative Filtering for Explainable Sequential Music Recommendation," in Proc. of the 17th ACM Conf. on Recommender Systems (RecSys '23), 2023.
[14] J. R. Anderson et al., "An integrated theory of the mind," Psychological Review, vol. 111, no. 4, pp. 1036, 2004.
[15] A. B. Melchiorre et al., "Investigating Gender Fairness of Recommendation Algorithms in the Music Domain," Information Processing Management, vol. 58, 2021.
[16] Y. K. Tan, X. Xu, Y. Liu, "Improved Recurrent Neural Networks for Session-based Recommendations," in Proc. of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016), 2016.
[17] P. Ren et al., "RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-Based Recommendation," Assoc. for the Advancement of Artificial Intelligence (AAAI), pp. 4806-4813, 2019.
[18] W. -C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," 2018 IEEE Int. Conf. on Data Mining (ICDM), Singapore, 2018.
[19] X. Fan et al., "Lighter and Better: Low-Rank Decomposed Self-Attention Networks for Next-Item Recommendation," in Proc. of the 44th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '21), 2021.
[20] Vaswani, A. et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.
[21] D. Berlyne, "Novelty, complexity, and hedonic value," Perception and Psychophys, pp. 279 – 286, 1970.
[22] Liu, Chang, et al. "Noninvasive self-attention for side information fusion in sequential recommendation." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 5. 2021.
[23] G. Geva et al., "Transformer feed-forward layers are key-value memories," in Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing, Nov. 7 – 11, 2021.
[24] H. Zamani et al., "An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation," ACM Trans. Intell. Syst. Technol., vol. 10, no. 5, Article 57, Sep. 2019.
[25] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

旅行者の想起プロセスを反映した観光場面と条件に基づく旅程生成方式

箕輪 真衣† 伴 匠生† 林 康弘† 清木 康†

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

E-mail: {s2022031，s2122084}@stu.musashino-u.ac.jp，{yhayashi，y-kiyoki}@musashino-u.ac.jp

あらまし 本研究は観光を対象として，旅行者が想起する場面（要求・景観・感性）と条件（時間・予算）に基づいて旅程を生成する方式について提案する．本方式は旅行者により入力される観光都市，観光地区を絞り込んだ上で，地区内の観光地において体験可能なアクティビティ，景観，感性と，予算と時間を満たす一連の距離計算とスコア統合により，旅程を生成する．本方式の有効性と実現可能性について示す．

キーワード 意図理解，文脈情報，時空間情報，旅程生成，観光

1. はじめに

検索エンジンを用いた情報検索では，ユーザの意図した印象語と一致する結果が得られないことがしばしばある．菊原らによれば，観光スポット情報は発信源が多岐にわたり，ユーザが自身の状況や意図，感性に合致した情報を見つけ出すことは非常に困難であるとされている[1]．

旅工房の「旅行計画」に関する意識調査によると現代のインターネット社会では，旅行の計画を立てる際に「インターネット」を利用する人が 85.4%であり，旅行の計画を手間がかかる，難しいと感じたことがある人がおおよそ 7 割ほどと示されている[2]．

従来の旅行では，ある特定の観光都市に旅行に行きたいと考えてから，更に観光地区を決めて，そこから複数の観光地について調べて自身の旅行条件と比較しながらルートを考案しなければならないものであった．

本研究の目的は，旅行者にとって煩雑な観光地区に関する情報収集と観光地選定を行うプロセスを簡略化し，旅行者が気軽に旅行に赴けるようにすることである．

これを実現するために，観光都市の選定，観光地区のリサーチ，訪問する観光地の選択，予算と時間を考慮した観光計画の立案が含まれる．「場面」と「条件」を考慮した多面的な観光計画の提案を行う．

本システムにより旅行者にとって煩雑な観光地区に関する情報収集と観光地選定を行うプロセスを簡略化し，旅行者のニーズに合った観光行程の提案が可能となる．

2. 関連研究

本稿では，要求と景観を用いた観光地区選択と，感性を用いた観光地の絞り込み，予算と時間を考慮した観光地の組み合わせ生成を行った後最適なルートの生成を行う．

本節では，本研究を遂行する上で発想の原点となった先行研究について述べる．これまでも観光地の検索・推薦システムに関する研究は数多く行われてきた．中でも本研究で参考にした先行研究は以下のとおりである．

2.1. 意味の数学モデル

言葉の意味を，数値を用いて表現し演算処理を行う手法として意味の数学モデルがある．これは，キーワードとキーワードの文脈を説明する文脈語を与えることで，意味的な連想検索と情報抽出を実現する計算モデルである[3]．

川島らによれば，これを旅行者の状況や意図に合わせた移動手段を発見するための計量機構として応用したシステムであり，旅行者の感性コンテキストに基づく最適化された提案を実現するとされている[4]．

本研究ではこの考え方を参考として，各観光地区と観光地に対する場面に対して数値的な意味的情報を持たせることで感性語から各観光地区，観光地とユーザの入力との距離計算を行う．

2.2. セマンティックオーダリング

ベクトル空間を用いた観光地とユーザコンテキスト場面の相関量計量及び順位付けを行う意味的スポット探索モデルである[1]．

本システムでは，観光地区の選定で場面の要求と景観を用いた計算で用いる．また，観光地の類似度による並び替えと絞り込みの後，観光地の組み合わせ生成に用いる．

2.3. シンボリックフィルタリング

ユーザの状況と意図に応じ，スポットのメタ情報の取捨選択と計量空間選択を行う[1]．

本システムでは，観光地の感性とユーザの提示する感性の類似度による並び替えを行う．

2.4. サイバーフィジカルシステム

現代社会には，物理空間と情報空間が共存しており，それらを対象とした統合系は，サイバーフィジカルシステム(CPS)と称される．サイバー・フィジカル連動のサイクルとして，“物理区間”において発生する事象に関するデータは“情報空間”に写像され，それらは“情報空間”において記憶，共存，統合，検索，分析され，それらの結果は“物理空間”へ発信される．このサイクルを対象とした新たな知識共有・分析・発信や，様々なビジネスチャンスの創造が期待される[5].

2.5. EPA 構造分類

感性の階層構造は，原因と結果の因果関係を前提していると述べられており，感性を引き起こす原因として刺激の物理的特徴が起点となっている[6].

大まかな分類は知覚・認知や対象に対する心理的反応を表す「印象層」対象の総合評価を表す「価値層」の区別が共通してみられる[7].

EPA 構造感性モデルのイメージ図を図 1 に示す.

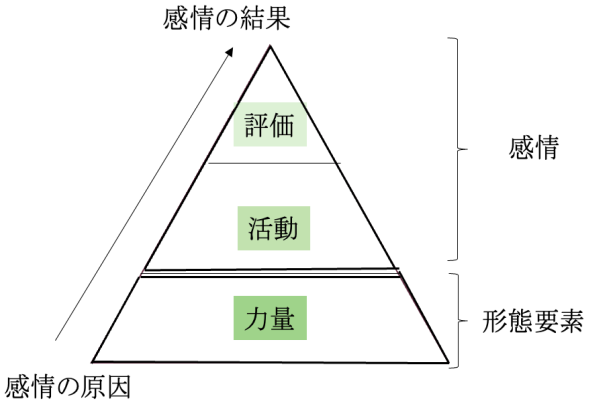


図 1 EPA 構造感性モデル

EPA 構造モデルに分類される形容詞の対義語の例

- 評価性因子
  - 「好き - 嫌い」
  - 「良い - 悪い」
  - 「美しい - 醜い」
- 活動因子
  - 「騒がしい - 静かな」
  - 「動的な - 静的な」
  - 「派手な - 地味な」
- 力量因子
  - 「硬い - 柔らかい」
  - 「鋭い - 鈍い」
  - 「緊張した - 緩んだ」

3. 実現方式

本章では，提案方式である，旅行者の想起プロセスを反映した観光場面と条件に基づく旅程生成方式について示す．本実装は Python 言語[8]を用いて行われた．本実装のユーザインターフェース部分は Web アプリケーション技術が用いられた．本実装のユーザインターフェースとなる Web アプリケーションを実装するにあたり，フレームワークとして Flask[9], HTML[10], JavaScript[11]を利用した．出力された観光行程の MAP 表示には folium[12]を用いた．観光都市データとユーザの入力データを蓄えるものには PostgreSQL［13］を用いた．PostgreSQL を扱う SQL サーバは ElephantSQL[14]を用いた．

本方式におけるシステム構成図を図 2 に示す.

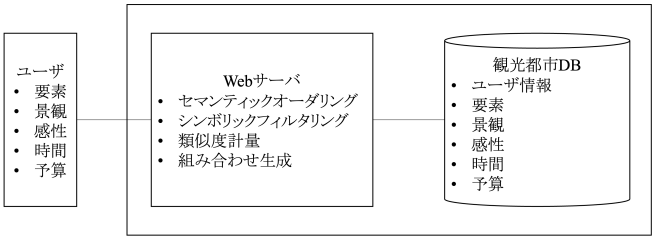


図 2　システム構成図

3.1. 提案手法の概要

本節では，提案手法の全体像について述べる．提案システムの全体像を図 3 に示す．

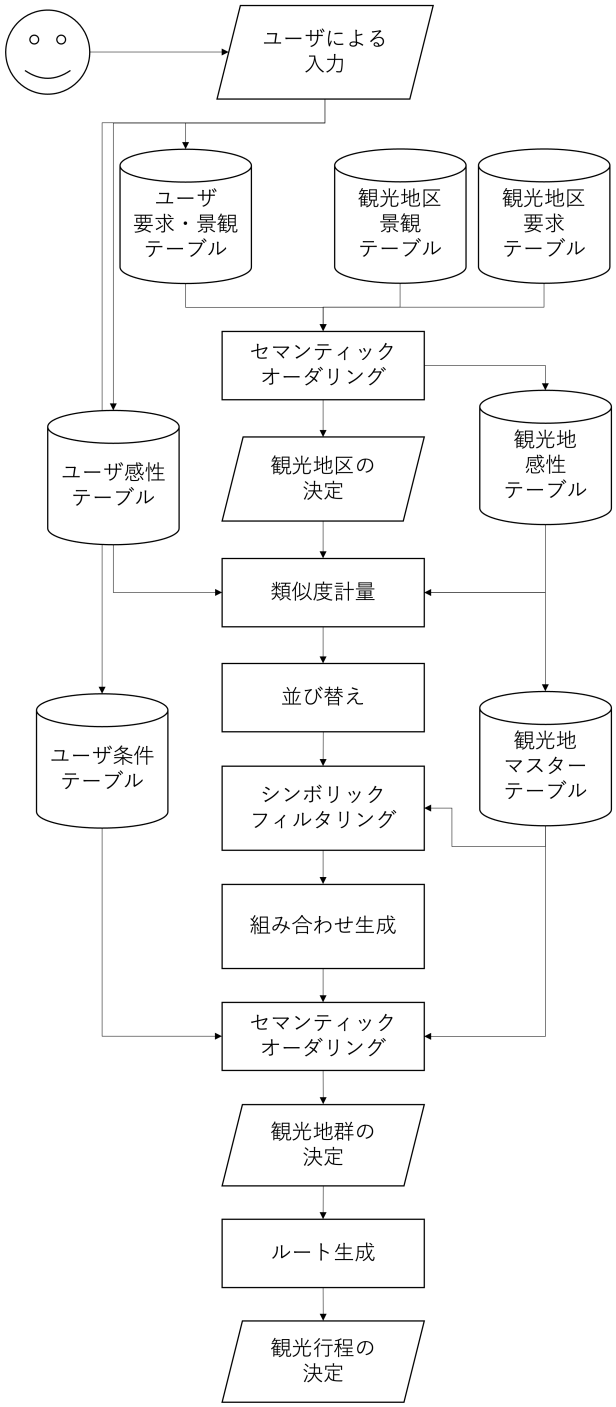


図 3　提案システムの全体像

本方式では，入力データとしてユーザが想定するショートトリップの場面である要求，景観，感性，所要時間と予算を与え，要求・景観の重み付けにより観光地区の選定を行い，観光地の組み合わせを生成する．生成された観光地の組み合わせから，観光地間の距離を計算し，最も移動距離が短くなるルートで観光行程を出力する．

本方式は，観光地区選定機能，観光地選定機能，観光地組み合わせ生成機能，移動ルート最適化機能からなる．

3.2. 観光地区選定機能

本節では，ユーザが入力した要求・景観から観光地区を選定する方法について述べる．観光都市 DB の観光地区の場面を示す観光地区要求テーブルと観光地区景観マトリクは共に 8 カラムから成る．

ユーザは観光地区要求・景観マトリクスと同様の 8 カラムから旅行に求める要求・景観を第一希望・第二希望・第三希望まで選択する．

ユーザ要求ベクトルとユーザ景観ベクトルに対してユーザの入力で第一希望から第三希望まで選択されたカラムへそれぞれの重みとなる値を付与して，ユーザ要求ベクトル p，ユーザ景観ベクトル q を作成する．観光地区要求マトリクスから抽出された観光地区の要求ベクトル e ，観光地区景観マトリクスから抽出された景観ベクトル f とする．各観光地区について，p ,e と q, f の内積をそれぞれ計算する．これら 2 つの内積の和をすべての観光地区について計算し，その和が最大となる観光地区 D を算出する(式 1).

$$D = argmax_i(p \cdot e_i + q \cdot f_i) \quad (式\ 1)$$

「要求」

本研究で定義される要求とは，旅行者が旅行に求めるもの，その観光地でできること（行動，物理）である．本研究での要求は以下に示される 8 単語で構成される．

表 1 要求の表現語

学習	文化	アウトドア・スポーツ	買い物
食事	温泉	遊歩道	自然

「景観」

本研究で定義される景観とは，その観光地で体感することができる景色のことである．本研究での景観は以下に示される 8 単語で構成される．

表 2 景観の表現語

山	川	滝	ダム
鍾乳洞	湖	桜	紅葉

3.3. 観光地選定機能

本節では，観光地区選定機能により選定された観光地区内に存在する観光地を，感性項目の類似度計量によって選定する方法について述べる．本機能はユーザが入力したユーザ感性ベクトルoと,観光地を表現する観光地感性マトリクスから抽出される観光地感性ベクトルj間で cos 類似度による類似度計量を行う(式 2).

$$cos(o,j) = \frac{o \cdot j}{|o||j|} = \frac{\sum_{i=1}^{n} o_i j_i}{\sqrt{\sum_{i=1}^{n} o_i^2}\sqrt{\sum_{i=1}^{n} j_i^2}} \quad (式\ 2)$$

本研究で用いられる感性語は，旅行と感性に関する先行研究である山田ら[15], 小島ら[16]により用いられた感性項目に，自然資源の観光地を表現するための独自の印象語を加えた 15 単語からなる．感性項目は複数の観点から観光地を表現するため，EPA 構造分類に該当する各因子から 5 単語ずつ用いられている．

本研究での感性は以下に示される 3 因子 15 単語で構成される.(表 3)

表 3　本研究で用いられる感性項目(15 項目)

評価因子

わくわく	気持ちいい	おもしろい	ロマンチック	癒やされる

活動因子

気が晴れる	きれい	落ち着く	にぎやか	びっくり

力量因子

神秘的	有名	歴史的	豪華	文化的

3.4. 観光地組み合わせ生成機能

本節では，観光地選定機能により選定された複数の観光地から，ユーザが入力した時間・予算の制約を満たす観光地の組み合わせを生成する方法について述べる．本機能は，観光地選定機能により選定された観光地による全通りの組み合わせとなる観光地群を作成する．作成された観光地郡の中から，ユーザが入力した時間・予算の制約を満たす観光地群のみを抽出する．抽出された観光地群における，各観光地に設定された観光目安時間，観光目安予算の合計値とユーザが入力した時間，予算それぞれの差を取る．本機能では，計算した差の絶対値が最も小さくなる観光地群を出力とする．

3.5. 移動ルート最適化機能

本節では，観光地組み合わせ生成機能により生成された観光地群に含まれる観光地間の距離から観光地間の移動ルートを最適化する方法について述べる．前述の方式により選択された各観光地の緯度・経度から全ての観光地間の距離を計算する．地球は球体であるため，平面上の距離を測る三平方の定理を用いた距離計算では，厳密な距離は測れない．そのため観光地間の距離計算には，球面上の距離である大円距離の計算を採用する．本システムでの始点と終点は奥多摩駅と想定されているため，奥多摩駅を含めた観光地群に含まれる観光地間の距離を全通り求めた後，全ての観光地を訪問し，最も移動距離の短くなるように観光地の訪問順を設定する．また，これによる移動ルートの最適化を行う．

4. 実験

本システムの有効性を検証するために，３つの実験を行った．実験 1 では，本計算方式とデータセットの有効性を示すため，筆者自身が想定する入力に対して想定できる出力となる観光行程の提案が行われるかについて検証を行った．実験 2 では，本計算方式の正当性を示すため,データセットを他のデータベース(以下 DB)に差し替えて正しい出力が得られるかについて実験を行った．実験 3 では，本システムの有用性を示すため，実験参加者が本システムに入力を行い，実地にて出力された観光行程を巡ることによりシステムの有用性と観光行程の満足度を調査した．

4.1. 実験 1

4.1.1. 予測出力の精度検証

構築したシステムの有効性を確認するために，理論的出力と実際の出力を対比して検証を行った．

本実験では，予め定めた場面と条件を入力として，想定される観光地区および観光地の出力を定義した．具体的には，観光地区の選定，観光地の並び替え，絞り込み，および組み合わせ計算手法の有効性を評価することを要求としている．本実験では観光行程のルート生成は対象外とした．

［観光地区］

入力となる景観と要求に基づき，観光地区の出力は奥多摩湖を想定し，以下の優先順位を用いた．本実験における入力画面を図 4 に示す．

- 要求：1 位「学習」，2 位「遊歩道」，3 位「自然」
- 景観：1 位「ダム」，2 位「湖」，3 位「紅葉」



図 4 実験 1 の入力画面

［観光地］

観光地区として奥多摩湖が選定されたと仮定し，観光地の出力は奥多摩湖に属する以下の観光地 ID を設定した：

- 19
- 20
- 23
- 22
- 25
- 26

実験 1 における入力とする感性項目を表 4 に示す.

表 4 観光地を求めるための入力

評価因子	わくわく	おもしろい
印象因子	気が晴れる	にぎやか
力量因子	有名	文化的

4.1.2. 実験 1 の実験結果

観光地区の出力：奥多摩湖

観光地の出力：20，23，24，34，40

本実験における出力画面を図 5 に示す.

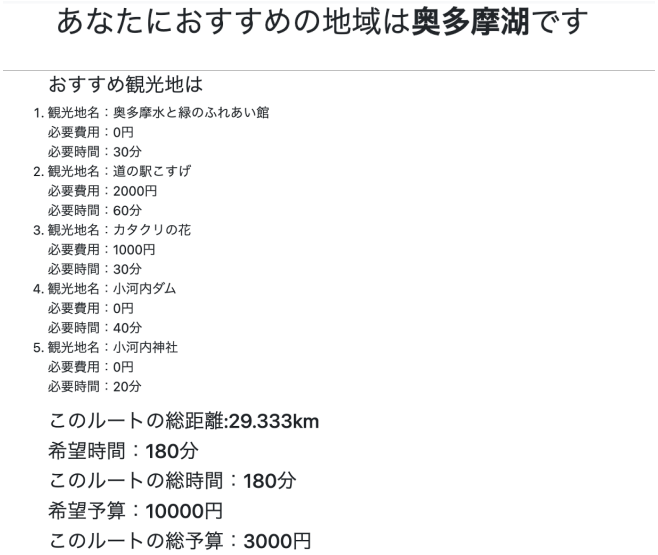


図 5 実験 1 の出力画面

実験の結果，入力された条件に基づいて奥多摩湖が観光地区として選定された．観光地に関しては，適合率 57%および再現率 80%になった．

観光地区の選定については，ダムや湖といった景観場面が高い値を有しているため，奥多摩湖が選ばれる確率が高い．また，奥多摩湖内に学習施設が存在し，これも選定に寄与する場面である．

観光地の選定においては，ID19，23，25 が複数の評価場面に対して高い値を持つため，選択される可能性が高いと予測された．ID20，24 は「わくわく」,「気が晴れる」,「文化的」,「有名」というカラムで高い値を有し，ID26 は「わくわく」,「気が晴れる」,「おもしろい」のカラムにおいて高い値を示した．実際に出力された観光地は ID20，23，24 であり，予測された観光地のうち 57%が実際の出力に一致した．選択されなかった観光地については，予測された ID19，25，26 が含まれず，代わりに ID34，40 が提案された．これは，予測には含まれなかったものの，入力された値に基づいて類似度が高いと判定されたためである．

4.2. 実験 2

4.2.1. 対象データを変更しての適用範囲の検証

本システムは，対象データセットを自然や景勝地が中心となる観光都市として想定されており，本研究の場合は観光都市を奥多摩と定めて研究を進めている．本節では，本システムの適用範囲の検証を行うため，自然や景勝地などが中心であり，同等の観光都市である秩父 DB を作成し実験を行った．

秩父 DB は奥多摩 DB と同条件で作成され，登録されているデータ数は観光地区が 6 件，観光地が 170 件である．ルート提案の始点と終点はともに西武秩父駅に設定されている．入力は以下の通りである．

- 要求：1 位「文化」，2 位「温泉」，3 位「食事」
- 景観：1 位「川」，2 位「山」，3 位「ダム」
- 感性：
  評価因子：わくわく, ロマンチック
  印象因子：気が晴れる, 落ち着く
  力量因子：神秘的, 文化的
- 活動可能時間：300 分
- 予算：8,000 円

4.2.2. 実験 2 の実験結果

秩父 DB を用いてシステムの操作を行ったところ

観光地区：大滝

観光地： 142,148,159,161

観光行程：159→161→142→148

予算：3,850 円

時間：295 分

上記の観光行程が出力された．本実験の観光行程の出

力画面を図 6，観光行程を地図上に可視化した画面を図 7 に示す．

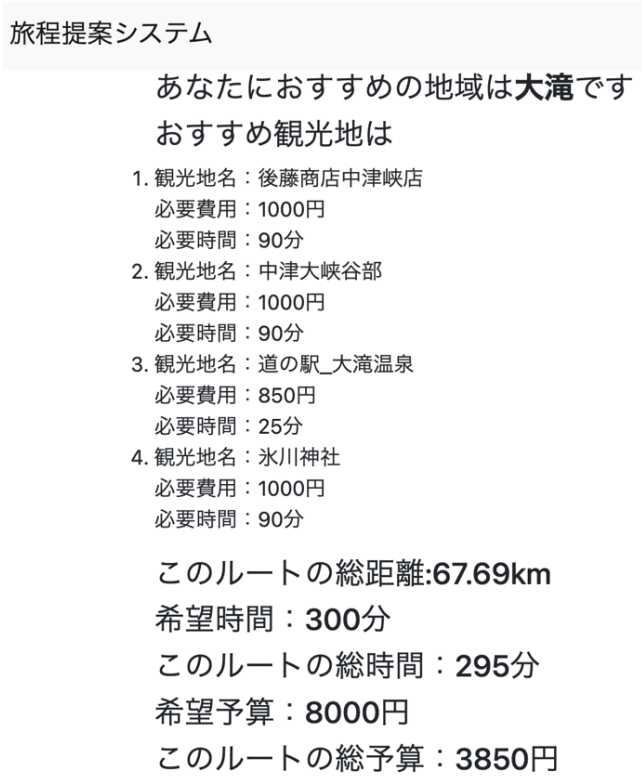


図 6 実験 2 の出力画面

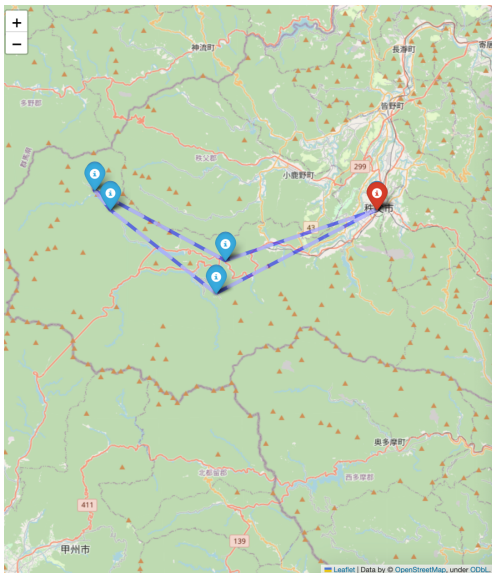


図 7 実験 2 の出力画面地図表示

要求で選択されている，文化＝氷川神社，食事＝商店，温泉＝大滝温泉が反映されている出力結果となった．希望時間が 300 分に対して，出力結果の必要時間が 295 分，総予算が 3,850 円であり，いずれもユーザが入力した条件を満たしている．

4.3. 実験 3

4.3.1. 実地調査による実証実験

本研究では，提案された観光行程生成システムの有用性を評価するために，実際のユーザによるテストを行った．このテストでは，参加者に本システムを使用してもらい，提案された観光行程を巡り旅程に対する満足度，使いやすさ，およびシステムの応答時間を評価した．また，参加者には，システムを介して観光行程を計画した経験と，従来の手法との比較についてもコメントを求めた．

ユーザの入力は以下の通りである．

- 要求：1 位「自然」，2 位「遊歩道」，3 位「文化」
- 景観：1 位「紅葉」，2 位「湖」，3 位「山」
- 感性：
  評価因子：気持ちいい，癒やされる
  印象因子：気が晴れる，きれい
  力量因子：豪華
- 時間：240 分
- 予算：10,000 円

4.3.2. 実験結果及びユーザフィードバックの分析

ユーザの入力より出力された実験結果と，ユーザからのヒアリング結果について論述する．各観光地の満足度と観光行程の満足度は 1–5 の 5 段階で評価されている．

観光行程
観光地区：奥多摩湖
観光地：
1. 八方岩展望台
2. 山のふるさと村
3. 小河内神社
予算：2500 円
時間：210 分

上記の観光行程が出力された．観光行程の出力画面を図 8，観光行程を地図上に可視化した画面を図 9 に示す．

あなたにおすすめの地域は**奥多摩湖**です

おすすめ観光地は

1. 観光地名：八方岩展望台
   必要費用：0円
   必要時間：40分
2. 観光地名：山のふるさと村
   必要費用：2500円
   必要時間：150分
3. 観光地名：小河内神社
   必要費用：0円
   必要時間：20分

このルートの総距離: 7.17km
希望時間：240分
このルートの総時間：210分
希望予算：10000円
このルートの総予算：2500円

図 8 実験 3 の出力結果

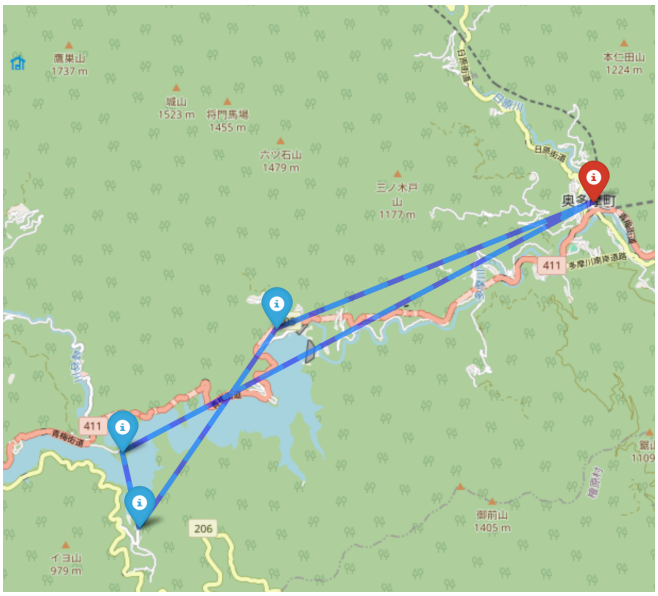


図 9 実験 3 の出力画面地図表示

1. 八方岩展望台
   観光地の満足度：4
   この観光地を訪れてのユーザ評価：
   ユーザの感性評価は，概ね肯定的であり，特に「気持ちいい」と「きれい」に関する評価が高かった．ユーザは，時間帯によって感性の評価が変わる可能性があることを指摘し，アクセスの良さと季節的な要因が感性に影響を与えると述べた．

2. 山のふるさと村
   観光地の満足度：3
   この観光地を訪れてのユーザ評価：
   ユーザは特定の感性に対する強い印象はないものの，多岐にわたる感性に触れることができると評価している．この観光地が情報提供のハブとして機能する可能性を示唆している．

3. 小河内神社
   観光地の満足度：3
   この観光地を訪れてのユーザ評価：
   ユーザは神社を中心とした文化的な要素に注目しているが，他の観光スポットと比較して，特に際立った魅力は感じられなかったとコメントしている．

観光行程に対する評価
観光行程の満足度：3.5-4
観光行程に対してのユーザ評価：
出力された観光行程の見やすさを肯定的に評価した．ただし，より現実的な経路表示や交通アクセス情報の追加が望まれるとも指摘された．

本システムに関して：
ユーザからは季節や時間帯，交通アクセス方法を含めた感性評価の調整に関する提案がなされた．本研究で使用されたシステムを季節限定で特化させるか，利用時の日付から季節を推定し，観光地の感性属性を季節バージョン(例：春バージョン，秋バージョン)で調整することで,ユーザ体験の向上が期待される.また，車だけでなく，バスなどの公共交通手段によるアクセス情報の組み込みも，システムの有用性を高めると提案されている．

5. 考察

本方式により，旅行者の想起プロセスを反映した観光場面と条件に基づく旅程生成方式を実現した．3 つの実験より，本方式による観光行程生成の実現可能性を示した．

実験 1 では，構築したシステムの有効性を確認するために，理論的出力と実際の出力を対比して検証を行った.実験 1 の結果より,観光地区の選定については，ダムや湖といった景観場面が高い値を有しているため，奥多摩湖が選ばれる確率が高い．また，奥多摩湖内に学習施設が存在し,これも選定に寄与する場面である.これにより本方式で構築したシステムの有効性を確認できた.

実験 2 では，本システムの適用範囲の検証を行うため，観光資源が同等である秩父 DB を作成し実験を行った．実験 2 の結果より，大滝地区はダムがあり水産

資源も豊富な観光地区であるので，観光地区の選択は正しい出力だといえる．観光地に関しては自然の景勝地が多く，感性項目の印象因子の一つである「落ち着く」が大きく反映されていると考えられる．本実験の結果より本システムの適用範囲が確認できた．

実験 3 では，提案された観光行程生成システムの有用性を評価するために，実際のユーザによるテストを行った．本実験で提案された観光行程は観光地区が奥多摩湖となり，八方岩展望台，山のふるさと村，小河内神社をめぐる観光行程が提案された．この観光行程はユーザの求める観光場面に該当する観光地であり，ユーザの提案する条件を満たしているものであった．実際に現地に赴いてルートをたどることにより本システムの有用性を実証できた．

6. 結論と今後の展望

本研究で開発された観光行程生成システムは，ユーザの場面(要求，景観，感性)と条件(時間，予算)を考慮に入れた上で，個別化された観光行程提案を行うことができる．本システムは，複数の完結な入力に基づいて迅速に情報を処理し，観光地の選定からルートの計画まで，一連のプロセスを簡略化する．それにより旅行のプランニングという面倒な行程の簡略化を行い，ショートトリップへの促進を行う．

今回，奥多摩にある観光地 82 件を対象とした検索実験システムを構築し，評価実験を行うことにより本システムの実現可能性を示した．実験 2 により，本システムに汎用性が確認できたので他の舞台となる観光都市データセットを作成することにより，本システムを用いた自然資源が豊かな観光地へ赴くきっかけの一因となると考える．

参 考 文 献

[1] 菊原魁斗，清木康．“個人適合学習機構を有する検索者の状況・感性に応じた スポット探索方式の実現” DEIM Forum 2018

[2] ＜最新！旅行意識調査＞7 割が旅行の「計画疲れ」に悩む！ 気づくといつも同じ「マンネリ旅」を 6 割以上が経験！ ～プロがあなたに代わって旅行計画！旅工房の「コンシェル旅」が本格始動~
https://about.ta-bikobo.com/news/press/2018/06/180605/

[3] Yasushi Kiyoki , Takashi Kitagawa , Takanari Hayama. ”A meta-database system for semantic image search by a mathematical model of meaning, ”ACM SIGMOD Record, Vol. 23 Issue4, December 1994.

[4] 川島宏一郎，林 康弘，清木 康，三田 哲也，移動経路における感性コンテキストに対応する移動・活動手段の組合せ形 成システム －Emotional MaaS－ 日本データベース学会

[5] 清木 康，厳 網林，IoT×GIS のシステム環境：5D WORLD MAP SYSTEM -グローバル環境知識共有・検索・分析可視化システムの実現-
https://www.jstage.jst.go.jp/article/tha-gis/27/2/27_107/_pdf/-char/ja GIS – 理論と応用 Theory and Application of GIS, 2019, Vol.27, No.2, pp.41-48

[6] 片平建史，武藤和仁，橋本翔，飛谷謙介，長田典子, SD 法を用いた感性の測定における評価の階層性 -EPA 構造の評価性因子の多義性に注目して- 日本感性工学会 2018

[7] 宮井彩希，片平建史，杉本匡史，長田典子(関西学院大学) 3 次元形状の感性指標化における階層構造 第 13 回日本感性工学会春季大会 2018

[8] Python https://www.python.org/

[9] Flask https://flask.palletsprojects.com/en/3.0.x/

[10] HTML https://html.spec.whatwg.org/multipage/

[11] JavaScript
https://developer.mozilla.org/ja/docs/Web/JavaScript/Reference

[12] Folium
https://python-visualization.github.io/folium/latest/

[13] PostgreSQL https://wwwpostgresql.org/

[14] ElephantSQL https://www.elephantsql.com/

[15] 山田厚子，林康弘，清木康，三田哲也,「感性」に基づく訪日外国人向け最適経路計量および short trip 可視化システム

[16] 小島隆次，仲村哲明 観光情報における感性語と色彩印象の結びつき 日本感性工学会論文誌 Vol. 14

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

ドメイン指向推薦 (2)

座長:國島 丈生(追手門学院大学)
コメンテータ:梅本 和俊(東京大学)
2024年3月1日(金) 13:30 〜 15:40 T3-B (オンライン（Zoom Events))

15:15 〜 15:40

[T3-B-8-05]【技術報告】 DMM.comのレコメンドに関する取り組み

森 雅也、土屋 裕太郎 (合同会社DMM.com)

本発表では、DMM.comのサービス概要やデータサイエンスグループが掲げるミッション、そこに属するレコメンドGrowthチームの取り組みやアーキテクチャについて紹介する。具体的に、ユーザー情報からアイテムを推薦するUser-to-Itemモデルや、アイテム情報からアイテムを推薦するItem-to-Itemモデルの概要と構築、またレコメンドに関連する施策提案から効果検証までのサイクルについて発表する。

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

🏛2024年3月1日(金) 13:30 ~ 15:40 | 🏛T3-C オンライン（Zoom Events)

情報探索支援

座長:倉島 健(NTT人間情報研究所)
コメンテータ:鬼塚 真(大阪大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T3-C-8-01] 代替行動推薦のためのアチーブメント関係にある行動の発見

*平野 司[1]、荘司 慶行[2]、山本 岳洋[3]、大原 剛三[1] (1. 青山学院大学、2. 静岡大学、3. 兵庫県立大学)

14:00 ~ 14:25
[T3-C-8-02] 芸能人への感想を表すマイクロブログ・ポストに対する大規模言語モデルを用いた理由集約

*横山 響[1]、宇津呂 武仁[1] (1. 筑波大学)

14:25 ~ 14:50
[T3-C-8-03] 論文調査の支援に向けた研究
（最新の読むべき論文の判別方法）

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

14:50 ~ 15:15
[T3-C-8-04] 手続き的知識源学習によるタスク検索における方法と目的の生成

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

15:15 ~ 15:40
[T3-C-8-05] 【技術報告】会社訪問アプリ「Wantedly Visit」における新規ユーザーの行動量に基づいた推薦方策の選択

林悠大 (ウォンテッドリー株式会社)

代替行動推薦のためのアチーブメント関係にある行動の発見

平野　司† 荘司　慶行†† 山本　岳洋††† 大原　剛三†

† 青山学院大学 理工学部　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
†† 静岡大学 情報学部　〒 430 – 0807 静岡県 浜松市 中区 城北
††† 兵庫県立大学 情報科学研究科　〒 651 – 2197 兵庫県 神戸市 西区 学園西町
E-mail: †hirano@sw.it.aoyama.ac.jp, ††{shoji,ohara}@it.aoyama.ac.jp, †††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本論文では，商品レビューデータと Web 検索データを用い，本来の目的を達成可能な別の行動を発見する方法を提案する.「講義に集中する」という目的を達成したいと考えた際に，始めに「眠気を防ぐ」という行動を考え付いたとき，それを達成可能な「カフェインをとる」「ガムを噛む」などの行動のみを探す場合がある．しかし,「睡眠の質を高める」「講義を変更する」など，元の目的を達成可能な行動は他に多く存在する．そこで，本研究では，商品レビューデータと Web 検索データから抽出した行動を用い，達成される目的行動と，目的を達成する行動という関係を表すアチーブメント関係グラフを構築した．この関係を用いることで，本来の目的を達成可能な代替行動の発見を可能にした.

キーワード　レビュー, 多様化, 行動抽出, グラフ解析

1　はじめに

人は日々，何かの目的を達成するために，様々な行動をとっている．たとえば「痩せた状態を保つ」ために「炭水化物を減らす」ことは，一般的である．こういったように，目的のために行動をとる際に，人々は 2 種類の視野狭窄に陥りがちである．一つ目は，手段の視野狭窄である．たとえば,「痩せた状態を保つ」方法として,「炭水化物を減らす」ことにこだわると，健康に悪い場合がある．本来であれば,「ジョギングする」「筋トレする」などの，ほかの行動も選択肢に入れるべきであるが，それらが思いつかなくなる場合がある．また同様に，目的にも視野狭窄は発生する．この場合,「痩せた状態を保つ」ことにこだわると，健康に悪い場合がある．痩せた状態を保つ理由が，モテたいという理由であれば，服やメイクを変えたり，出会いを増やすといった別の手段がある．お気に入りの服を着続けたいという理由であれば，リメイクをすること，同じデザインの大きい服を購入するなどという手段もある．しかし，視野狭窄に陥った人にとって，これらの代替行動を思いつくことは困難だと考えられる.

一般的な推薦手法として，従来の推薦システム [1] では協調フィルタリングが用いられている．協調フィルタリングでは，選択した商品と同じ種類のもの，その商品を購入した別のユーザが他に購入したもの，商品と併用して使われるものなどが推薦される．そのため，たとえば「ダイエット食品」をユーザが選択した場合，推薦される商品は別の種類のダイエット食品，異なるメーカの似たようなダイエット食品，合わせて利用されるサプリメントなどが推薦される．この時ユーザが「痩せた状態を保つ」という目的を持っている場合，これらの推薦では食品に関する解決方法のみに，より注目してしまう．実際には，筋トレグッズ，ランニングシューズなどの「ランニングをする」「筋トレをする」などの別の手段で解決可能な商品が存在する．また，そもそも「痩せた状態を保つ」理由がモテたいという理由であれば,「化粧品」,「ワックス」など身だしなみを整えるという別の目的を解決可能な商品が，本来の目的を解決する場合もある.

近年の推薦手法ではセレンディピティ性を保った推薦についても考慮されつつある．セレンディピティ性を保った商品推薦の定義として，Kotkov ら [2] は「関連性」「新規性」「意外性」の 3 つを備えたものであるとしている．そのため，これらの推薦手法では従来の推薦システムでは推薦することの出来ない意外性のある商品を推薦可能である．これらの意外性のある商品の推薦はユーザの視野を広げることができるかもしれないが，目的と行動を理解した推薦を行っている訳ではないため，本研究の課題を解決することは困難である．目的と行動を考慮した人々の視野を広げることのできる推薦を行うためには，行動間の達成と被達成の関係を表すアチーブメント関係を用いる必要がある．しかし，現状それらの関係を表すコーパスは存在しない.

そこで，本研究ではある行動名を与えると，それに関係する行動名を収集し，アチーブメント関係に基づく部分グラフを生成する方法を提案する．アチーブメント関係を表すグラフの例を図 1 に示す．具体的には,「摂取カロリーを減らす」という行動名を与えると，その兄弟行動として,「ランニングをする」「筋トレをする」などが出力される．また，親行動として入力された行動で達成可能である「痩せた状態を保つ」「異性に好かれる」が出力される．このとき，親行動はその子行動で達成可能であるため,「アチーブドバイ」という関係で結ばれる．このアチーブメント関係は，一般的に目的の達成と呼ばれ，親ノードは相対的に「目的」，子ノードは相対的に「手段」と呼ばれるものである.

このような部分グラフを作成するにあたり，まず兄弟関係を

構築するために商品レビューデータに着目した．ショッピングサービスではユーザが商品に対するレビューを投稿可能である．これらの商品レビューには，ユーザが商品をどの様な行動に用いたか，という情報を含む場合がある．同じ商品につく複数のレビューには，同じ目的を達成する行動が含まれることが多い．また，同じ目的を持つ商品は，類似している商品である可能性が高い．そのため，商品-目的の関係から兄弟関係を計算することが可能である．しかし，行動名はルールベースでの抽出が難しく，また同じ意味を持つ行動名でも表記ゆれによって異なる行動として抽出されてしまう．そのため，行動を含むか判定を行う分類器と，行動名の表記ゆれを解消する言語モデルを作成した．これらを用いて商品レビューから商品の使用目的を抽出し，表記揺れを解消することで，正確な行動名を取得した．取得した行動名と商品の関係を表すグラフを構築し，グラフ計算を行うことで兄弟関係を抽出した．

次に，親子関係を構築するためにウェブ検索データに注目した．ウェブ検索で，たとえば「ランニングをする」という行動名を用いて「ためにランニングをする」等のクエリで検索を行うと，「痩せるためにランニングをする～」といった検索結果を得ることが出来る場合がある．そのため，収集した行動名を使ってクエリを作成し，得られたウェブ検索を用いることで，親子関係にある行動を抽出できる．これらを言語モデルによって行動名に直したうえで，出現頻度順にランキングすることで，入力された行動名の親，子，兄弟からなるアチーブメント関係の部分グラフを作成した．

提案手法の有効性を示すために，実データを用いた評価実験を行う予定である．行動名クエリを入力とし，提案手法による処理によって入力行動の兄弟行動，親子行動を出力する．出力された行動が，それぞれ行動名として成立しているか，関連性のある行動か，兄弟または親子関係であるかという評価を，被験者によるラベル付けによって行う．ラベル付けによって得られたデータを用い，ベースライン手法との精度比較によって提案手法の有用性を示す．

本論文の構成は以下の通りである．第 2 節では，関連研究を紹介し，本研究の位置づけを示す．第 3 節では，本研究における提案手法について説明し，第 4 節で提案手法とベースライン手法を比較する被験者実験について述べる．

2 関 連 研 究

本研究は，ある行動に対する代替行動，達成される目的，達成する手段を発見する研究である．そのために，達成される目的と達成する手段の関係であるアチーブメント関係に基づくグラフを構築する．よって本研究は代替達成案の研究と含意関係認識の研究と関連する．

2.1 代替達成案

本研究は，ある行動に対して代替可能な行動を，達成される目的と達成する手段という関係で構築されたアチーブメント関係を用いることで抽出する．代替可能な行動の抽出について，

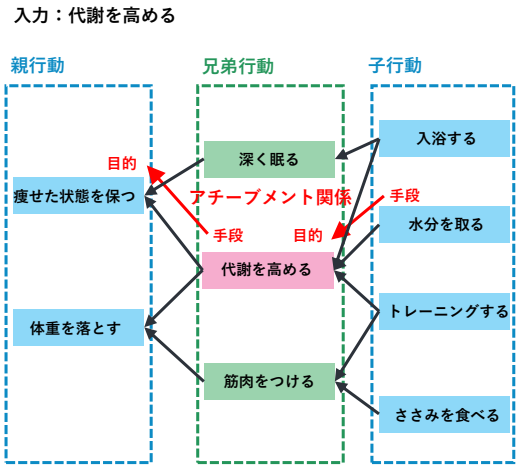


図 1　アチーブメント関係を表すグラフ

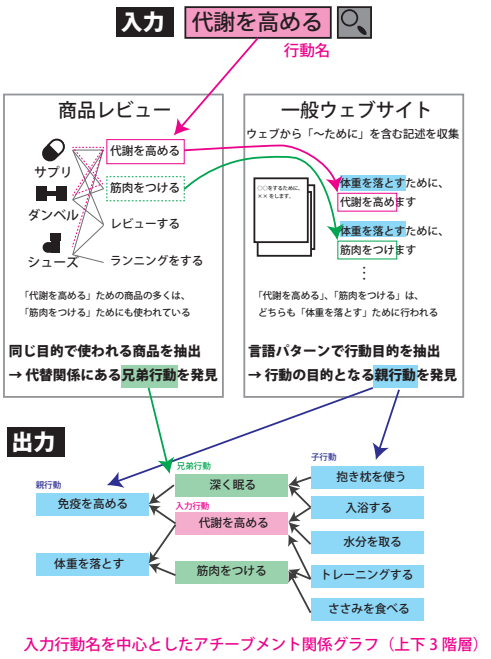


図 2　手法の概要

Pothirattanachaikul ら [3] は CQA サイトから，目的を達成出来る代替案を抽出する手法を提案している．例として，「睡眠問題の解決」という目的に対し，「睡眠薬を飲む」という行動に注目した場合，「就寝前に散歩する」という代替行動を提示する．そのために，行動と目的の関係性を CQA サイトの質問と解答情報を 2 部グラフ化し，類似度を順位付けすることで，代替行動を発見している．

また，Yamamoto ら [4] と Yang ら [5] は主目的と副目的の関係から，代替案を求める手法についての研究を行っている．Yamamoto らは，「体重を減らす」という主目標を達成するための，「運動をする」「ダイエット薬を飲む」などを副目標と定義し，スポンサードサーチデータを利用することで，副目標をクエリに合わせたクラスタリングを行う手法を提案している．Yang らは，「会議を開催する」という主タスクの達成に向けて

行う必要がある,「ホテルの選択」「ボランティアの募集」などを副タスクと定義し，クエリと wikiHow で述べられているタスクとの関係を結ぶ手法を提案している．本研究では，これらの代替案の定義に基づき，代替可能な行動を発見可能なアチーブメント関係を表すグラフを構築する．

2.2 含意関係

本研究は，達成される目的と達成する手段という行動の関係を構築する．この二種類の行動の関係は上位・下位の関係として捉えることができ，含意関係の研究と関連する．含意関係とは，二つの文のうち前提である文が真であるとき仮説である文が真であると推論される関係のことであり，上位・下位概念としてとらえることができる．含意関係タスクのアプローチとして，含意関係データセットを用いたディープラーニングモデルによる学習 [6]，語彙，論理などをベースにした含意認識モデルの構築などの手法 [7] が存在する．Haggag ら [7] は，含意認識モデルを文字列の比較による語彙ベース，グラフで表現された構文を利予数る構文ベース，テキストと仮説の表現を比較する意味ベース，公理，推論規則を利用した論理ベースの方法に分類し，それぞれの性能について分析を行っている．

論理ベースのアプローチを用いている研究として，Punyakanok ら [8] は，記述論理言語を用いることで対象になる文章を階層関係に帰納させ，文章をより表層的な表現に抽象化しており，前提の文が仮説である文の表現を含んでいるかを求めるアプローチを提案している．また，Martinez ら [9] は，含意関係の公理を用いた上で，WordNet などの外部言語知識データを用いることで，含意関係に必要な背景知識を補足したアプローチを提案している．本研究は，商品と行動をグラフ化し，類似行動を推論したうえで親子行動を抽出するという手法をとる．そのため，本研究は特に推論規則を用いる論理ベースの含意関係の研究と関連する．

また，行動名の含意関係に注目した研究も行われている．Hasimoto ら [10] は,「離婚した」という文章から「(それ以前に）結婚した」という推論,「勝訴する」という文章から「告訴する」という推論を獲得するため，同じ文脈に出現する単語は同じ意味を持ちやすいという仮説を定義し，含意関係を持つ動詞のペアの類似度尺度を提案している．本研究は目的と手段という上位・下位の関係を求めており，加えて兄弟関係にある行動を求めるという点で異なる．

3 手　　法

ある行動名が入力された場合，その行動で達成される目的，その行動と同じ目的を達成可能な行動，その行動を目的とした場合に手段になる行動のアチーブメント関係を表すグラフを構築する手法について述べる．本研究では，このようなアルゴリズムを実現するために，商品のレビューデータから言語モデルを用いた行動名の抽出と抽象化を行う．得られた行動名を用い，兄弟関係になる行動を商品データから，親子関係になる行動を web 検索データから抽出する．これらの抽出した行動を親子，兄弟関係でリンクを与えることで，アチーブメント関係を表すグラフを構築する．

3.1 レビュー中の目的となる行動を含む文の抽出

本研究では，行動名を得られるデータとして，商品のレビュー情報を利用する．これは，商品レビューには「○○をするために，この商品を購入した」というような，アチーブメント関係が多く含まれていると考えたためである．一方で，レビュー全体の中で，商品を使ってどのような目的達成したかを書いてある部分はごく一部に過ぎない．そのため，レビュー全体をそのまま行動名として使うことはできない．そこで，レビューを複数の文に分割し，1 文ごとに，目的となる行動を含んでいるかどうか判定を行う．

ある文が与えられた際に，その文が購入目的となる行動名を含んでいるかを，2 値分類で判定する．判定には Delvin ら [11] が提案した言語モデルである BERT（Bidirectional Encoder Representations from Transformers）を用いる．既存の BERT のモデルを目的判定に使用可能にするために，ファインチューニングを行う．BERT の学習方法と目的判定の概要を図 3 に示す．

ファインチューニングのために，クラウドソーシングで正解データセットを作成した．具体的なタスクとして，複数の文からなるレビューデータを提示し，商品をどのように利用するかという目的を含む文章を抽出させ，簡単に言い換えた表現を入力させた．50 人のクラウドワーカーが，一人あたり 40 件のレビューについてタスクを行った．こうして，合計で 2,162 件の目的を含む文と，言い換えた表現が得られた．ここで得られた目的を含む文を本節で行う抽出器のトレーニングに用い，言い換えを行動名の要約（次節で説明）に用いる．

クラウドソーシングで得られた目的を含む文を正例，抽出されなかった文を負例とし，BERT を二値分類タスクでファインチューニングする．こうすることで，任意の文を入力すると，その文に目的となる行動が含まれているかを 0 か 1 かで分類可能になる．こうしてファインチューニングした分類機に，実際に抽出対象となる文を入力して，目的となる行動を含む文だけを残す．具体的には，レビューデータを区切り文字「。」「！」「？」で切り離して文に切り分け，すべての文を分類機にかける．こうすることで，レビュー全体から，目的となる行動を含む文だけを抽出することができる．

3.2 行動名の要約

本研究では，最終的に行動名をノードとしたグラフを作成することを目的とする．現時点で，レビュー全体から，目的行動を含む文だけを抽出できた．これらの文は，長く自由な表現を持つため，グラフのノードとして扱うことができない．文をそのままノードにした場合，完全一致する文はほとんど存在しないため，疎なグラフになる．そこで，目的を含む文を簡単に言い換えた，行動名を生成する．

生成的言語モデルを用いて，任意の行動を含む文を入力すると，それを簡潔な単語に変換する，行動名の生成を行う．生成

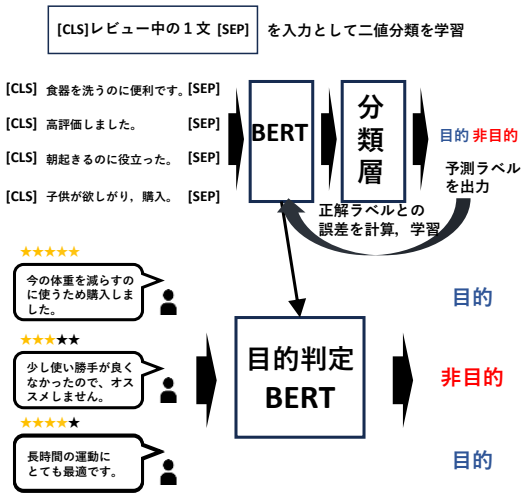


図 3　BERT を用いた目的を含む文の判定

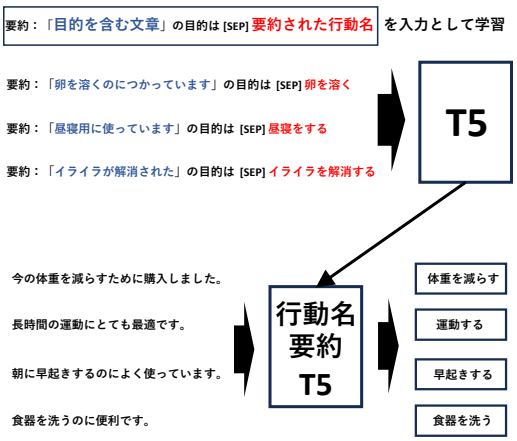


図 4　T5 を用いた行動名要約

のために，事前学習済み生成的言語モデルである T5（Text-to-Text Transfer Transformer）を用いる．T5 はテキストからテキストへの変換に適した言語モデルであるが，特定の目的に使うためには，ファインチューニングが必要である．今回の場合，文から行動名を生成ができるするため，入力された文を簡単な表現に変換するというタスクでファインチューニングを行った．

T5 の学習方法と，行動名生成の概要を図 4 に示す．ファインチューニングには，第 3.1 節で行ったクラウドソーシングで集めた，目的行動を含む文と，それを簡単に言い換えた表現を用いた．行動を含む文を与えると，その文を要約した行動名を生成するというタスクで，モデルをファインチューニングした．

具体的な学習手順として，まず，T5 は，prefix でタスク名を指定する形式のファインチューニングに適しているので，prefix を用いたプロンプト込みで追加学習を行った．具体的には，目的を含む文の前に，「**要約：「** 」という prefix を置き，文のあとに「**」の目的は**」というプロンプトを加えた．そして，セパレータトークン（[**SEP**]）を挿入したのち，クラウドソーシングで得られた行動名を書き加えた．たとえば，「昼寝用に購入しました.」という文と「昼寝する」という行動名からなる学習データは，実際の学習時には「**要約：「昼寝用に購入しました.」の目的は [SEP] 昼寝する**」となる．このようなタスクで追加学習した T5 を用いて，3.1 節で得た目的を含む文章を入力とし，要約された行動名を生成する．

3.3　行動名の表記ゆれ解消

本研究は，入力として受け取った行動名のアチーブメント関係を表すグラフを構築する．兄弟関係を抽出する前に，入力された行動名の表記ゆれを解消する．たとえば，「睡眠をとる」という行動名が入力された場合，そのままグラフ計算を行うと，「睡眠をする」「睡眠する」などの行動が兄弟行動として抽出される可能性がある．しかし，これらの行動は同じ意味を持っているため，兄弟行動や親行動ではなく，同一行動である．そのため，同じ意味を持つ行動についての表記ゆれを解消する必要がある．

表記ゆれの解消のために，それぞれの行動名をベクトル化し，類似度比較することで，あまりにも似すぎたノード同士をまとめる．類似度を計算するため，文脈の意味を持つベクトル化を行うことが可能な Sentence BERT [12] を用いる．グラフ内のノードとなるすべての行動名をベクトル化し，総当たりでコサイン類似度を計算した．ペアのうち，類似度が閾値（今回の実験時は 0.7 とした）を超え，なおかつ元クエリの先頭の単語を含む場合，入力された行動名と同一の行動名とみなし，ノードを結合した．

3.4　商品行動グラフによる兄弟行動の発見

表記ゆれを解消した行動名を用い，兄弟関係を持つ行動を抽出する．ここで，多くの共通の商品で達成可能な行動同士は，お互いに代替可能，すなわち兄弟関係にあるという仮定を立てた．たとえば，目覚まし時計が「早起きをする」と「目を覚ます」という行動名を持つ場合，どちらも朝起きることに関する目的を共通して解決する．そのため，入力された行動を持つ商品の他の行動名を抽出することで，兄弟行動を発見できる．

一方で，商品へのレビューに含まれたすべての行動名が，兄弟行動の対象になるとは限らない．第一に，クレンジングが必要である．たとえば，行動名には「レビューする」「購入する」など，特定の商品の購入目的とは関係のないものが存在する．加えて，一度でも同じ商品で達成された行動同士が，すべて兄弟関係にあるとは限らない．

そこで，この課題を解決するために，商品と行動をグラフとして扱い，グラフ処理によって真に兄弟関係にある行動名を抽出することを考えた．商品と，その商品へのレビューに含まれる行動名を用いて，商品と行動ノードからなる二部グラフを構築する．ここで，同じ商品で達成可能な行動名同士は類似しており，またある行動を共通して達成可能な商品同士は類似しているという仮定を置く．このような仮定は，二部グラフ上でのグラフ処理で計算可能である．グラフ計算を用いることで，入力された行動名と関係のある，特徴的な行動名を発見可能にし，無関係な行動を除外可能にする．

実際に，商品と行動名からなる二部グラフを作成する．計算用のグラフの構築のために，クレンジングとして，ストップワードを含む行動ノードの除去と，過剰なリンク数を持つノードの除去を行う．商品レビューサイトには，特定の製品に関係なく，ほぼすべての商品で達成可能な行動が書かれる場合が多い．例えば,「注文する」「送料を減らす」「ポイントを付ける」などの行動名は，商品の種類とは関係なく，すべての商品レビューに登場しがちである．そのため，まず，多くのカテゴリをまたいで登場する頻出語を集め，人手でクレンジング用のストップワード辞書を作成した．このストップワード辞書には,「購入」「注文」「ポイント」「送料」などの単語が含まれる．それぞれのノードのうち，行動名にストップワード中の語が 1 語でも含まれた場合，そのノードはグラフから取り除いた．

また，本研究では，商品は単体の商品で，たかだかいくつかの目的で使用されることを前提に仮定を考えた．一方で，実際の商品の中には，福袋や，様々な商品の詰め合わせセットなど，複数の種類の商品を混ぜて売っているものが多く含まれた．これらに対するレビューは，全く関係ない多くの行動が含まれる．こういった一部の商品は，グラフ化した際に，あまりにも多くの行動とリンクを持ち，計算に悪影響を与える．そのため，一定以上の数の行動名ノードへのリンクをもつ商品ノードは，グラフから取り除いた．今回の実験時には，経験的に 15 件より多くの行動を持つ商品を除去した．また，行動についても同様に，どの商品でも達成可能な行動名ノードも除去した．今回の実験では，経験的に全体の商品数の 1 %より多く商品とリンクを持つ行動について計算用グラフから除外した．

こうして作成したグラフについて，実際にグラフ計算で，任意の行動名ノードに対して，兄弟行動である度合いの高いノードを発見する方法について考える．ここでは，ランダムウォークによって計算される滞在確率に基づいて，グラフ中の行動名ノードをスコア付けする．作成したグラフを，隣接行列を G で表す．商品ノードの集合を I，行動名ノードの集合を A とすると，G は $(|I|+|A|) \times (|I|+|A|)$ の次元からなる行列となる．この行列では，行は行動から商品へのリンクの有無を，列は商品から行動へのリンクを表す．

ここで，i_n とリンクを持つ A の集合を $L_a(i_n)$，p_n とリンクを持つ I の集合を $L_i(a_m)$ とする．行列の右上の場合，i_n が $L_i(a_m)$ に含まれる場合，G_{nm} の値は 1 であり，それ以外は 0 になる．左下の場合も，a_n が $L_a(i_m)$ に含まれる場合，G_{nm} の値は 1 であり，それ以外は 0 になる．このように，ノード間のエッジが繋がっているかどうかという関係をもとに，二部グラフを隣接行列として定義する．

次に，隣接行列 A を，遷移確率行列に変換する．遷移確率行列の変換には，ノードから出発するエッジの重みの和で各列を割る正規化によって行った．G を正規化した遷移確率行列を G' とすると，正規化の式は

$$G'_{ij} = \frac{G_{ij}}{\sum_{k=1}^{|I|+|A|} G_{kj}} \tag{1}$$

で定義される．

ここで，この遷移確率行列で計算を行った場合，クレンジングによって除外しきれていない，多くのリンクを持つ行動が高い類似度を持つ．そこで，多くの商品とリンクを持つ行動名について遷移確率を下げるため，正規化した遷移確率をリンク数で割ることで，重みを調整した．すなわち，重みづけは

$$G''_{i<|A|j>|I|} = \frac{G'_{ij}}{\sum_{k>|I|}^{|A|} G'_{kj}} \tag{2}$$

という計算で行われる．

構築されたグラフを用いて，RWR（Random Walk with Restart）を計算する．RWR では，ある任意のノードから出発したサーファーが，リンクで結ばれた隣のノードに移動する（Random Walk）が，一定確率で出発地点のノードにジャンプする（Restart）条件を置く．そして，ある瞬間にそれぞれのノードにサーファーが滞在している確率を，そのノードのスコアとして扱う．今回の場合，入力された行動名を持つノードを出発ノードとして RWR を計算することで，グラフ内の各行動名ノードが，入力行動と兄弟である度合いを計算する．

入力行動名ノードが与えられた際に，そのノードの次元だけを 1 として，残りの次元を 0 とした $|A|$ 次元のワンホットベクトルを q と置く．ここで，c をジャンプ率とすると，RWR の n 回目の計算における入力行動名から出発したサーファーの滞在確率 p_n は，

$$p_n = (1-c)G'' p_{n-1} + cq \tag{3}$$

と表すことができる．

最終的な各ノードごとのサーファーの滞在確率からなるベクトル p_∞ は，べき乗法による収束計算で求められる．p の初期値として q を与え，結果が収束するまでこの計算を繰り返す．この際，式 (2) で重み付けを行っているため，そのまま繰り返し計算した場合，1 回の計算ごとにグラフ全体の滞在確率の総和が減少する．そのため，実際の計算の際には，1 回の計算ごとに，p を確率が 1 になるように正規化した．こうして求めた類似度ベクトル p を用い，すべての行動名集合 A について順位付けを行い，上位 15 件の行動を兄弟行動とする．

3.5 言語パターンを用いた親子行動の発見

次に，得られた兄弟行動を用いた，親子行動の抽出を行う．複数の行動名が与えられた際に，その行動名のすべてに対して共通して目的である行動名を親行動と呼ぶ．このような行動を発見するために，一般文書中での行動名の記述のされ方に注目する．行動 A と行動 B があった際に,「C するために A する」と,「C するために B する」という記述が世の中に多く存在した場合，行動 C は行動 A と行動 B の親行動とみなすことができる．このような一般文書中の記述を用いて行動間の親子関係を判定するために，言語パターンを用いて Web 文書から行動名の書かれた文書を収集する．

実際の手順として，世界中の「ために」を含むすべての文書をあらかじめ用意することは困難であるため，Web 検索エンジンを用いて 2 つの行動名とその関係を含む文を収集する．具体

的には，言語パターンを含む検索クエリを，入力行動とその兄弟行動を用いて作成する．Web 検索クエリには，一般的なウェブ検索で使われるワイルドカード記号である “*” を用いる．親行動を発見したい場合は，“*ために【行動名】*”，子行動を発見したい場合は “*【行動名】ために*” とした．

実装上の注意事項として，親行動のクエリについて「昼寝する」「寒さを防ぐ」などの動詞は,「昼寝します」「寒さを防ぎ」など，活用形に応じて別の表現がある場合がある．そのため，サ変活用についてはそれ以前の名詞以降をワイルドカードにし，それ以外の活用については動詞の一文字目以降をワイルドカードにした．たとえば,「昼寝をする」という行動名は「*ために昼寝*」と「*昼寝をするために*」というクエリに変換され,「寒さを防ぐ」という行動名は「*ために寒さを防*」と「*寒さを防ぐために*」というクエリにそれぞれ変換される．

得られた検索結果から，スニペット内の該当箇所を用いて親子行動の抽出を行う．親行動の場合,「ために」以前の文章を抽出し，子行動の場合「ために」以降の文章を抽出する．抽出した文章を区切り文字「、」「。」「！」「？」で分割し，3.1 節で学習した分類器によって目的を含む文を判定した．こうして得られた目的を含む文を用いて，3.2 節で学習した T5 によって行動名を生成する．得られた行動名のうち，共通して得られた行動の頻度で順位付けを行い，上位のものを親行動，子行動とみなす．今回の実験では，共通した行動が 3 件未満であった場合，共通した行動のみを親子行動とした．

3.6 アチーブメント関係を表すグラフの構築

得られた兄弟行動，親子行動を用いてグラフを構築する．入力された行動名を基準に，兄弟行動，親子行動でリンクを持たせることで，目的と手段の関係を表すアチーブメント関係を表すグラフを構築する．グラフの細かい構造については現在検討中である．構築されたグラフを最終的に出力とする．このグラフを入力とした行動を基準に確かめることで，本来達成すべき行動が何か，別の行動は何かという判断を可能にする．

4 評　　価

本研究では，実際に楽天市場のデータセットを用いて，被験者にラベル付けをさせる評価実験を行い，手法の有用性について評価を行った．実験では行動名クエリに対する兄弟行動，親子行動についての評価を行った．評価にはあらかじめ用意した 15 個の行動名クエリについて，提案手法を含めた 4 手法で兄弟行動，親子行動を求め，兄弟行動について 4 つの項目の評価，親子行動について 5 つの項目の評価を行った．

4.1 データセット

実験には，楽天市場の商品レビューデータ [13] を利用した．レビューデータにはレビュー文章以外に商品名，カテゴリデータなどが含まれており，これらのデータを利用したグラフの作成が可能である．しかし，レビューデータ中には本研究に適さないデータが存在するため，絞り込みを行った．今回は，カテゴリデータによる除外を行った．楽天市場のデータには 39 種類のルートカテゴリが存在し，商品はそれぞれ「ファッション」「家電」「食品」などの大きなグループに分かれる．これらのデータを利用し，こちらで指定したカテゴリを持つ商品のレビューデータを除外した．除外したカテゴリの例として,「不動産」「チケット」などの本研究で求める利用目的を持たない可能性の高いものがある．この際，例外処理として，カテゴリを持たないデータはすべて排除した．また，今回の研究では言語モデルによる分類，生成タスクを必要とし，全体のデータを対象にした処理は現実的ではないため，一部の範囲のデータのみを利用した．最終的に，607,096 件の商品を対象に実装を行った．

4.2 実　　装

獲得した楽天市場のレビューデータについて，目的を含む文章の抽出，行動名の生成を行った．この時，3.4 節で説明した，ストップワードによる行動ノードの除外と，一定のリンク数以上を持つ商品，行動ノードの除外に加えて，リンク数が 1 以下のノードについての除外を行った．また，行動ノードに関して，動詞で終わらない場合が複数存在した．そのため，形態素解析を行い動詞で終わらない行動ノードについての除外を行った．最終的に，27,881 件の商品ノードと 14,414 件の行動ノードを得た．

次に，兄弟関係の発見に利用するグラフの作成を行った．事前に用意した行動名クエリを用い，3.3 節の方法で同一ノードに類似した行動をまとめた．なお，用意した行動名クエリは，必ず二つ以上の商品にリンクを持つ行動名が，表記ゆれ解消によって得られた行動を含めて存在するものを利用した．用意した行動名クエリを表 1 に示す．表記ゆれ解消を行ったノードを始点とし，Random Walk with Restart による計算を行い，行動の関連度を計算した．この時，始点に戻る確率を 0.80 とし，再帰計算は結果が収束するまで行った．計算の結果得られた上位 5 件の行動を兄弟行動とした．

最後に，親子行動発見のための Web 検索を行った．検索には Google 社が提供する Custom Search JSON API[1] を用いた．各行動名クエリに対して，グラフ計算によって得られた兄弟行動の上位 15 件と行動名クエリを加えた 16 件の行動を用い，検索結果のうち 3 ページ分までの結果を得た．検索結果の概要文から行動名を生成し，頻度ベースで上位 3 件の行動をそれぞれ親子行動とした．

4.3 比 較 手 法

提案手法の有用性を評価するために，

（1） **提案手法**：
重み付けと表記ゆれ解消を用いたグラフの関連度計算を行う手法，

（2） **重み付けを用いたグラフ計算**：
リンクの重み付けをのみを用いたグラフ計算を行う手法，

（3） **表記ゆれ解消を用いたグラフ計算**：

1：「Custom Search JSON API」：
https://developers.google.com/custom-search/v1/introduction?hl=ja

表 1　実験に利用する行動名クエリ

行動名クエリ	表記ゆれ解消による同一行動の種類
加湿する	2
乾燥を防ぐ	12
代謝を高める	1
筋肉をつける	6
ダイエットする	2
布団を干す	95
風邪を予防する	9
寒さを防ぐ	7
睡眠の質を上げる	8
貴重品の紛失を防ぐ	3
カビを防ぐ	15
液晶を掃除する	77
健康を保つ	9
肩こりをなくす	6
ストレスを減らす	6

表記ゆれの解消のみを用いたグラフ計算を行う手法,

（4） **共起頻度**：
同一商品に付く行動名の頻度が高い行動を兄弟行動とする手法,

の 4 つの手法を使い，兄弟行動，親子行動について抽出を行った．

4.4　兄弟行動の評価実験

用意した行動名をクエリとし，上位 5 件の行動名に対してラベル付けを行った．評価項目として，

- **行動名として適切か**：
  純粋に行動として成立しているかどうかのみの評価,
- **関連する行動か**：
  兄弟，親子関係にかかわらず関連性を持つ行動であるかどうかの評価,
- **兄弟行動か**：
  クエリ行動と共通の目的を解決できる行動であるかどうかの評価,
- **ただの言い換えではないか**：
  ほぼ同じ意味の行動でないかどうかの評価,

の 4 つを用意し，それぞれ 4 段階による評価を行った．最も当てはまる場合は 4 とし，当てはまらない場合は 1 とする．

4.5　親子行動の評価実験

行動名クエリを用いて得られた上位 15 件の兄弟行動によって得られた，上位 3 件の親行動，子行動に対してラベル付けを行った．評価項目として，

- **行動名として適切か**：
  純粋に行動として成立しているかどうかのみの評価,
- **関連する行動か**：
  兄弟，親子関係にかかわらず関連性を持つ行動であるかどうかの評価,
- **兄弟行動か**：
  クエリ行動と共通の目的を解決できる行動であるかどうかの評価,
- **親行動か**：
  クエリ行動を用いて対象の行動を目的として達成可能であるかどうかの評価,

表 2　兄弟行動における手法，項目ごとの平均評定 (5 段階中)，上位 5 件の精度と平均逆順位．

手法名	適切な行動名	関連行動	兄弟行動	言い換えではない	p@5	mrr@5
提案手法	**4.71**	**2.11**	**1.81**	4.86	**0.20**	**0.38**
重み付け	**4.71**	2.01	1.74	4.85	0.17	0.37
表記ゆれ解消	4.67	**2.03**	1.74	**4.91**	0.19	0.35
行動名の頻度	4.51	1.95	1.71	4.83	0.16	0.30

- **子行動か**：
  対象の行動を用いてクエリ行動を目的として達成可能であるかどうかの評価,

の 5 つを用意し，それぞれ 4 段階による評価を行った．

4.6　評 価 尺 度

兄弟行動の評価方法について，各項目の評価値をそれぞれのクエリで得た結果の全体の結果で評価した．また，兄弟行動として適切な行動であるかどうか，という判定については，行動名として適切か，兄弟行動かの二つの項目の評価値が 3 以上の場合，適切な行動であるとする．適切な行動の結果を用いて，上位 5 件の精度，平均逆順位を求めた．親子行動の評価方法についても同様であり，各項目の評価値をそれぞれのクエリで得た結果の全体の結果で評価した．また，適切な行動の判定のために親行動を求めるクエリで得た行動と，子行動を求めるクエリで得た行動に分割し，それぞれ行動名として適切か，親（子）行動かの二つの項目の評価値が 3 以上の場合，適切な行動であるとした．適切な行動の結果を用いて，上位 3 件の精度，平均逆順位を求めた．

4.7　実 験 結 果

表 2 に兄弟行動の評価実験によって得られた回答の平均値と上位 5 件の精度，平均逆順位を示す．全クエリの平均結果として，提案手法の結果がほぼすべての項目において高い評価となった．適切な兄弟行動の発見についても提案手法が最も高い評価となり，ベースライン手法の精度を上回った．

表 3 に親行動の評価実験によって得られた回答の平均値と上位 3 件の精度，平均逆順位を示す．全クエリの平均結果として，提案手法の結果がほぼすべての項目において高い評価となった．適切な親行動の発見については，精度は提案手法が最も高い結果となり，ベースライン手法の精度を上回った．また，平均逆順位では重み付けのみを用いた手法が最も高い結果となった．

表 4 に子行動の評価実験によって得られた回答の平均値と上位 3 件の精度，平均逆順位を示す．全クエリの平均結果として全体を通して高い精度を出した手法は無く，またそれぞれの結果に大きな差は見られなかった．適切な子行動の発見についてもグラフ計算手法はベースライン手法を上回っているものの，各工夫点による精度の違いは見られなかった．

5　考　　察

まず，兄弟行動と親子行動の発見について提案手法の全体的な評価を行う．兄弟行動と親行動ではほぼ全ての項目と適切な行動の発見精度について，ベースライン手法を上回ったが，子

表 3　親行動における手法，項目ごとの平均評定 (5 段階中)，上位 3 件の精度と平均逆順位.

手法名	適切な行動名	関連行動	兄弟行動	親行動	子行動	p@3	mrr@3
提案手法	**4.94**	**3.24**	**2.92**	**2.46**	**3.24**	**0.40**	0.45
重み付け	4.58	2.63	2.18	2.18	2.61	0.38	**0.53**
表記ゆれ解消	4.73	2.85	2.11	2.11	2.89	0.30	0.43
行動名の頻度	4.46	1.95	1.59	1.59	1.86	0.23	0.32

表 4　子行動における手法，項目ごとの平均評定 (5 段階中)，上位 3 件の精度と平均逆順位.

手法名	適切な行動名	関連行動	兄弟行動	親行動	子行動	p@3	mrr@3
提案手法	4.53	1.87	**1.77**	1.38	1.97	**0.24**	**0.37**
重み付け	4.44	1.80	1.58	1.43	1.93	**0.24**	0.36
表記ゆれ解消	4.49	**2.01**	1.69	**1.46**	**2.04**	**0.24**	**0.37**
行動名の頻度	**4.56**	1.84	1.54	1.37	2.00	0.20	0.33

行動ではベースライン手法との大きな精度の差は見られなかった．これらの結果から，グラフ計算手法を用いた手法は単純な頻度を用いた手法と比較すると，ある行動と同じ目的を達成可能な行動を発見することが可能であり，それらを利用した web マイニング手法によって適切な親行動を発見することも可能であることが分かった．一方で，子行動の発見については現状の web マイニング手法では不十分であることが分かった．

提案手法が有効に働いた行動名として,「寒さを防ぐ 加湿する」などがあった．これらの行動名では同一の意味を持つ行動名の表記ゆれを解消する際に,「寒さ 加湿」といった単語が含まれる行動名は多くの場合において一つの単語である程度行動の内容が一致しやすく，同じ目的の意味を持つ傾向が高い．そのため表記ゆれ解消のタスクがうまく働いたことが要因であると考えられる．

重み付けのみを行った手法について,「筋肉をつける 筋肉を鍛える」などの同じ意味を持つ行動名が兄弟行動として上位の結果となったことで，一部行動名では精度が低下した．一方で，重み付けが有効に働いたと見られる結果も見られた.「乾燥を防ぐ」という行動名において，表記ゆれ解消のみの手法の場合,「加湿する」という行動名は上位 5 件中 3 番目という結果であり,「寝室で使う 着ける」という行動名が上位の結果となった．重み付けのみの手法では「加湿する」は 1 番目という結果となり，表記ゆれ解消のみの手法では見られた行動名はより低い順位となった．この結果から，重みのペナルティによって「使う着ける」などのより多くの商品で使われるであろう一般的な表現が，下位の結果となることに作用したと考えられる．

親子行動の web マイニング手法について，親行動と子行動で同様の手法を用いたが，親行動に対して子行動の精度は大幅に低い結果となった．親行動は「*ために行動名*」というクエリを用いたため,「ために」につながる行動名を含むページが少なく，そもそもの該当検索結果数が少ない傾向があった．一方で子行動は「*行動名ために*」というクエリを用いたため，該当検索結果数が多い傾向があった．そのため，結果が少ない親行動の場合は検索に用いた行動名と関連しない行動名などのノイズが少なく，結果が多い子行動の場合は行動名と無関係の行動名がより多く存在していたことが考えられる．今回の手法では単純な頻度を用いたランキング手法であるため，ノイズを持つ結果が上位になりやすいことが要因であると考えられる．

6　まとめと今後の課題

本論文では，商品レビューデータと Web 検索データを用い，本来の目的を達成可能な別の行動を発見する方法を提案した．ある目的を達成したいと考えた際に，達成可能な行動は多く存在する．そこで，本研究では，商品レビューデータと Web 検索データから抽出した行動を用い，達成される目的行動と，目的を達成する行動という関係を表すアチーブメント関係グラフを構築した．

構築したグラフを用いて被験者による 2 つの評価実験を行った結果，提案手法は兄弟関係の発見について一般的な手法の精度を上回り，兄弟行動を発見可能であることを示した．一方で親子行動の発見については子行動の場合，一般的な手法とほぼ精度が変わらないという結果となり課題が見られた．

今後の課題として，親子行動の発見における web マイニング手法，ランキング手法を改善する必要がある．現状の手法では検索結果から得られた行動名を頻度順としてランキングを行っているが，この方法ではノイズである行動名を除外することはできない．そのため，検索に用いた兄弟行動のランキングを用いた新たなランキング方法を考えている．具体的には，より上位の兄弟行動で得られた親子行動については高いスコアを付与し，逆に下位の場合は低いスコアを付与することで，新たなランキングの指標を用意する．また，表記ゆれ解消の方法についても，全ての行動名が類似した行動名の表記ゆれを解消しているわけではない．そのため，同じ意味を持つ文章の類似度を計算する手法についても今後調べていく必要がある．

謝　　辞

本研究の一部は JSPS 科研費 21H03775，21H03774，22H03905 による助成を受けたものです．本研究では，楽天グループ株式会社が国立情報学研究所情報学研究データリポジトリを通じて提供している楽天市場データ [13] を使用しています．ここに記して謝意を表します．

文　　献

[1] Smith B. Linden, G. and J York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, Vol. 7, No. 1, pp. 76–80, (2003).

[2] Shuaiqiang W. Denis, K. and V Jari. A survey of serendipity in recommender systems. *Knowledge-Based Systems*, Vol. 111, pp. 180–192, (2016).

[3] Yamamoto T. Fujita S. Tajima A. Pothirattanachaikul, S. and K Tanaka. Mining alternative actions from community q&a corpus for task-oriented web search. Web Intelligence, p. 607–614, (2017).

[4] Sakai T. Iwata M. Yu C. Wen J. Yamamoto, T. and K. Tanaka. The wisdom of advertisers: Mining subgoals via query clustering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, p. 505–514, (2012).

[5] Z. Yang and E. Nyberg. Leveraging procedural knowledge

for task-oriented search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, p. 513–522, (2015).

[6] A. Mishra and P Bhattacharyya. Deep learning techniques in textual entailment. *Survey Paper, Center For Indian Language Technolgy*, No. January, pp. 273–282, (2018).

[7] ELFattah Marwa MA. Haggag, Mohamed H. and Ahmed M. Ahmed. Different models and approaches of textual entailment recognition. *International Journal of Computer Applications*, Vol. 142, No. 1, pp. 32–39, (2016).

[8] de Salvo Braz R. Girju R. Roth D. Punyakanok, V. and M. Sammons. An inference model for semantic entailment in natural language. In *Machine Learning Challenges Workshop, MLCW*, pp. 261–286. Springer, (2006).

[9] Mineshima K. Miyao Y. Martínez-Gómez, P. and D. Bekki. On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 710–720, (2017).

[10] Torisawa K. Kuroda K. De Saeger S. Murata M. Hashimoto, C., et al. Large-scale verb entailment acquisition from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1172–1181, (2009).

[11] Chang M. Lee K. Devlin, J. and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, (2019).

[12] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, (2019).

[13] 楽天グループ株式会社. 楽天市場データ, (2020). https://doi.org/10.32130/idr.2.1, https://rit.rakuten.com/data_release/.

芸能人への感想を表すマイクロブログ・ポストに対する大規模言語モデルを用いた理由集約

Aggregating Reasons for Microblog Posts Expressing Opinions on Celebrities by Large Language Models

横山　響†　宇津呂武仁††

† 筑波大学大学院 システム情報工学研究群 知能機能システム学位プログラム
〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1

あらまし　本論文では，芸能人のファンが，芸能人に関連する事柄についての批評や関連情報の探索を行いやすくすることを目的とする．そのために，まず，X(旧 Twitter) において特定の芸能人に関連する事柄について述べているポストを収集し，その芸能人に関するメジャーな感想を収集する．次に，収集した芸能人に関する感想を利用して，ウェブページを検索し，最大上位 30 件にヒットしたページの内容を，感想の理由を探すためのコンテキストとして収集する．一通り，ウェブページの内容の収集が完了すると，それぞれのページの内容から，芸能人に関する詳細な感想の理由の収集・集約を行う．この部分では，大規模言語モデル (LLM) の一例として ChatGPT を利用し，Retrieval Augmented Generation (RAG) を活用することで，LLM の事前知識のみに依存せず，検索されたウェブページの内容に基づき，芸能人に関する詳細な感想の理由の収集・集約を行う．人手で収集・集約した感想の理由と比較して，評価実験を行った結果，ChatGPT によって高い性能で理由収集・集約が行えていることが明らかとなった．また，理由収集において，既存のモデルを用いた場合との比較も行い，ChatGPT が高い性能を示していることを確認できた．
キーワード　マイクロブログ，LLM，情報抽出，質問応答，集約

1 はじめに

本論文では，芸能人のファンが，芸能人に関連する事柄についての批評や関連情報の探索を行いやすくすることを目的とする．そのために，本論文ではまず，X (旧 Twitter) において特定の芸能人に関連する事柄について述べているポストを収集し，その芸能人に関するメジャーな感想を収集する．具体的には，特定の芸能人の評価対象が含まれるポストを対象として，芸能人の評価対象と形容詞の共起頻度の統計を出す．ここで芸能人の評価対象としては，先行研究 [7] より，「○○ (芸能人名) の×× (名詞)」を対象とする．芸能人の評価対象と形容詞が共起している例としては，「山田涼介の演技・凄い」や「二宮和也の Twitter・面白い」などが挙げられる．本論文では，特定の芸能人の評価対象との共起頻度が高い形容詞は，その芸能人の評価対象に対するメジャーな感想を示していると判定し，該当する「芸能人の評価対象と形容詞の組」を詳細情報の収集のためのキーワードとする．具体的なキーワードの選出手順については 2 節で説明する．

キーワードを選定した後は，そのキーワードを用いてウェブページの検索を行い，上位 30 件にヒットしたページの内容を，感想の理由を探すためのコンテキストとして収集する．

一通り，キーワードに関するウェブページの内容の収集が完了すると，それぞれのページの内容から，感想の詳細な理由を収集する．この部分では，大規模言語モデル (LLM) の一例として ChatGPT を利用する．本研究の重要な側面は，この理由収集の部分で，LLM の出力の信頼性を向上させる上で，非常に大きな役割を果たす Retrieval Augmented Generation (RAG) [2] を活用していることである．RAG は，LLM が外部データベースから取得した情報を参照することで，出力の信頼性を向上させるものである．本論文では，RAG を活用して，ウェブページの内容に基づいて感想の理由を収集することで，LLM の事前知識のみに依存せずに，収集結果の信頼性を向上させることを狙っている．具体的には，ChatGPT に与える指示 (以降，プロンプト) を調整し，ウェブページの内容に感想の理由が含まれる場合はその理由を出力し，含まれない場合は「×」のみを出力することで，感想の理由の収集を行う．しかし，この段階では，重複する冗長な理由が多数収集されるため，ユーザが芸能人についての批評や関連情報の一覧を概観することは困難である．

そこで，本論文では，再度，ChatGPT を用いることにより，各芸能人の評価対象，および，評価対象に対する感想ごとに，重複して収集された理由の候補に対して，理由の頻度を考慮してカテゴリ分けし，高頻度なものから順に順位付けする．これ

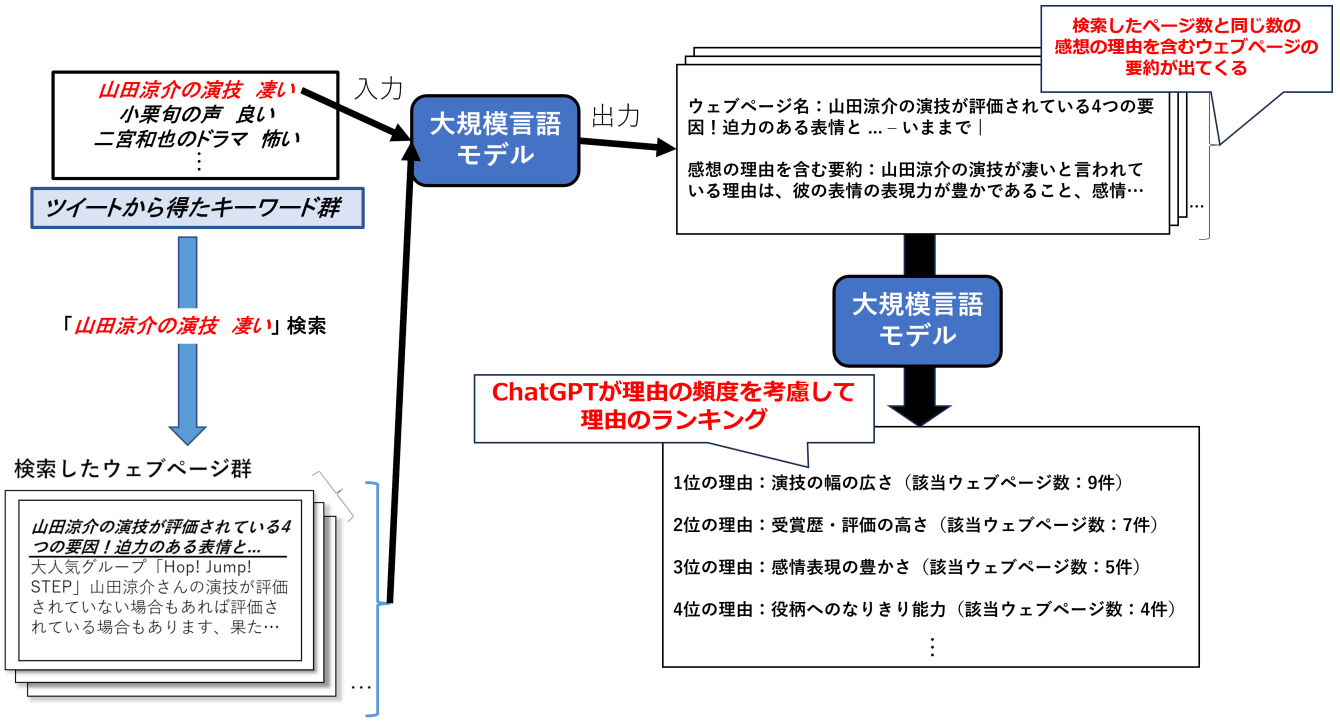


図 1　芸能人に関する感想ポストの理由集約の流れ

によって，ユーザは，各芸能人の評価対象に関する感想の理由を，集約され順位付けされた形式で知ることができ，芸能人に関連する事例についての批評や関連情報の情報探索が容易になることが期待される．

ここまでに説明した一連の理由収集・集約の流れは図 1 に示す通りである．

2　芸能人の評価対象に対するポストからの感想収集

本節では，X のポストから，芸能人の評価対象に関する感想の収集を行う手順について記述する．主に 2.1 節では，ポストの収集，収集対象とする芸能人および，芸能人の評価対象の選出方法について述べる．2.2 節では，芸能人の評価対象と共起する形容詞の組のうち，芸能人の評価対象に関する感想として使用するものの選出方法について述べる．

2.1　ポストの収集期間・対象とする芸能人の評価対象

本論文では，X 上で話題になることが多い芸能人を 10 名を選出し，その芸能人名を検索ワードとして，2022/09/07 から 2023/04/09 までの期間でポストの収集を行った．ポストの収集には，Twitter Search API[1]を用いた．芸能人名ごとに収集した「ポスト数」,「リポスト以外のポスト数」については表 1 に示す通りである．本論文では，このうち「リポスト以外のポスト」のみを用いる．

芸能人の評価対象の選出方法としては，まず，JUMAN++[2]を用いて，表 1 に具体的な収集数を示した「リポスト以外のポスト」に対して，形態素解析を行う．次に，形態素解析の結果を用いて「○○ (芸能人名) の」の直後に「×× (名詞)」が続く頻度を求め，各芸能人の評価対象である「○○ (芸能人名) の×× (名詞)」の出現頻度の統計を算出した．

本論文では，各芸能人の評価対象の出現頻度のうち，上位 3 件の評価対象を使用する．また，芸能人の評価対象から，意味の推測が困難と判断したものは，出現頻度が上位 3 件であっても対象から除外した (例：「二宮和也の 2012」)．結果として，計 24 件の芸能人の評価対象が使用対象となった．

2.2　芸能人の評価対象に対する形容詞の選定

まずは，2.1 節で選出した各芸能人名に対する評価対象について，それぞれポスト中で共起している形容詞の統計を取る．そのうち，共起頻度が上位 3 件の形容詞のみを使用する．次に，その形容詞が各芸能人の評価対象に対するメジャーな感想として成り立っているかを判定する．具体的には，共起頻度が 20 回に満たない形容詞は，その評価対象に対するメジャーな感想ではないと判断して，使用対象から除外する．また，意味の推測が困難な「芸能人の評価対象と形容詞の組」も使用対象から除外する (例：山田涼介の顔・無い)．加えて,「凄い」と「素晴らしい」などの似た意味を表している形容詞は，共起頻度が高い方のみを採用して，共起頻度が低い方は使用対象から除外する．

ここまでの手順を経て，各芸能人の評価対象に関するメジャーな感想の絞り込みが完了した．結果として，表 1 に示す 24 件の「芸能人の評価対象と形容詞の組」を，芸能人の評価対象に関するメジャーな感想として，感想の理由の収集のためのキーワードとする．実際にキーワードとして使用する際は，表 1 に示す形式のままではなく，先頭に対応する「○○ (芸能人名) の」を付け加える．具体的に，表 1 の芸能人名「山田涼介」に

1：https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
2：http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++

表 1　ポスト収集対象の芸能人名および評価対象・形容詞の組

芸能人名	ポスト数	リポスト以外のポスト数	評価対象・形容詞の組および組数	
山田涼介	938,882	213,886	ドラマ・怖い，ドラマ・面白い，演技・凄い，顔・良い	4
二宮和也	851,579	164,130	Twitter・面白い，演技・凄い	2
菊池風磨	1,131,863	185,823	演技・凄い，沼・深い	2
小栗旬	425,188	120,583	演技・凄い，演技・悪い，顔・良い，声・良い	4
綾野剛	272,232	95,890	魅力・多い	1
坂口健太郎	284,622	64,472	彼女・羨ましい，顔・良い	2
竹内涼真	83,368	31,670	演技・凄い，弟・おもろい，妹・面白い	3
有村架純	370,956	105,084	演技・優しい，瀬名・可愛い	2
中村倫也	702,807	206,632	結婚・嬉しい，変身・恰好いい，演技・凄い	3
永野芽郁	246,489	58,636	演技・上手い	1
合計	5,307,986	1,246,806	-	24

関する行の,「評価対象・形容詞の組」に該当する「演技・凄い」については,「山田涼介の演技・凄い」といった形式で使用する．

3　ChatGPT によるウェブページから感想の理由の収集・集約および人手評価

本節では，2.2 節で選出したキーワードを用いて，ChatGPT によって，ウェブページから感想の理由を収集・集約する手順について記述する．主に 3.1 節では，キーワードを用いてウェブページを検索し，ウェブページの内容を収集する方法について述べる．3.2 節では，収集したウェブページの内容から，感想の理由を収集する手順，および人手での評価結果について述べる．3.3 節では，3.2 節で収集した感想の理由に対して，理由の集約を行う手順，および人手での評価結果について述べる．また、3.4 節においては，3.3 節の理由集約においてプロンプト内の例示の有効性を評価した結果について述べる．

3.1　ウェブページの検索

まず 2.2 節で選出したキーワードを用いて，Google 検索エンジン [3]でウェブページの検索を行う．次に，検索にヒットしたウェブページ最大上位 30 件に対して，ウェブページの内容の収集を人手で行う．この一連の操作を全てのキーワードに対して行う．例として，キーワード「山田涼介の演技・凄い」の場合，まずはそのまま「山田涼介の演技・凄い」でウェブページの検索を行い，検索にヒットした最大上位 30 件のウェブページの内容を収集する．この手順で収集されたウェブページは，その内容にキーワードの感想の理由を含んでいると期待される．今回の例だと「山田涼介の演技が凄い理由」が内容に含まれていることを期待する．

3.2　プロンプトチューニングによるウェブページからの感想の理由の収集

次に，前節で収集したウェブページの内容を利用して，それぞれのウェブページごとに，感想の理由を収集する．この部分は図 3 の左側に該当する．ここで感想の理由を収集するための枠組みとして，ChatGPT[4]モデルの `gpt-4-0613` を使用する．具体的に ChatGPT に与えるプロンプト全体は，図 4 上部に示す通りである．ここでは，例としてキーワード「山田涼介の演技・凄い」を対象としたプロンプトを示す．始めに，以下に示すプロンプトを用いて，事前知識は用いずにこれから与えるコンテキスト情報から，感想の理由を探すように指示を与える．また，もしコンテキスト情報に感想の理由が含まれない場合，「×」のみを出力させる．

> 事前知識ではなくコンテキスト情報を活用して、質問文に回答してください。
> 山田涼介の演技が凄いといわれているのは何故ですか？
> コンテキストに理由が記載されていない場合は「×」とだけ出力してください

このように，事前知識ではなくウェブページの内容をもとに感想の理由を探させることで，ChatGPT が実際の情報とは異なる，もしくは存在しない情報を出力してしまう Hallucination を抑えることを期待している．次に，以下に具体的なプロンプトを示すように，ウェブページの内容をそのまま渡し，感想の理由を探させる．この操作を収集したすべてのウェブページの内容に対して行う．

> コンテキスト情報
> ---
> 山田涼介の演技が評価されている 4 つの要因！迫力のある表情と演じた役が重要!?「お姫様に捧ぐ親指」大人気グループ「Hop! Jump! STEP」山田涼介さん･･･
> ---

この手順を経て収集される感想の理由は，最大で収集したウェブページの数と同じになる．

ここで，ChatGPT を利用して収集された感想の理由のうち，キーワード 10 組を対象としたものに対して，人手による感想の理由収集結果と比較することで，評価を行う．当該ウェブページ d に対して ChatGPT が出力した理由の集合 $S(d)$ と，当該

3 : https://www.google.co.jp/

4 : https://platform.openai.com/docs/models/

ウェブページ d に対して人手で用意した理由の集合 $R(d)$ をもとに，再現率・適合率は，以下のように定義される．

$$\text{再現率} = \frac{\sum_d |R(d) \cap S(d)|}{\sum_d |R(d)|}, \qquad \text{適合率} = \frac{\sum_d |R(d) \bigcap S(d)|}{\sum_d |S(d)|}$$

それぞれのキーワードに関連したウェブページごとに評価を行い，そのミクロ平均を，キーワードごとの評価結果とする．キーワードごとの評価結果については，表 2 の「理由収集 (例示なし)」の列に示す．全体の評価結果としては，キーワードごとの評価結果のマクロ平均とする．感想の理由収集における，全体としての評価結果を図 2 に示す．結果として，理由収集においては再現率・適合率，F 値のいずれも，9 割程度の高い性能が得られた．

3.3 プロンプトチューニングによるウェブページからの感想の理由の集約

次に，前節で収集した感想の理由に対して，理由の集約を行う．この部分は図 3 の右側に該当する．ここでは ChatGPT モデルの `gpt-4-1106-preview` を利用する．具体的に ChatGPT に与えるプロンプト全体は，図 4 下部に示す通りである．ここでも，例としてキーワード「山田涼介の演技・凄い」を対象としたプロンプトを示す．まず，以下に示すプロンプトで，一連の指示を終えた後に，特定のキーワードに関するウェブページの要約を渡す旨を示す．また，複数渡すウェブページの要約に対して，それぞれのウェブページの要約の言及内容に基づいて，カテゴリ分けを行うように指示する．これにより，感想の理由の集約を行う．

> すべての指示を伝えた後に、山田涼介の演技が凄い理由に関わる【ウェブページの要約】を渡します。
> まずは、【ウェブページの要約】の言及内容に基づいて、【ウェブページの要約】のカテゴリ分けを行ってください。

今回の例では，キーワード「山田涼介の演技・凄い」に関するウェブページの要約を渡すことを始めに提示している．次に，具体的なカテゴリ例とそのカテゴリに該当する文を提示し，例に基づいてカテゴリ分けを行うように指示をする．また，例で示した以外のカテゴリが存在する場合は，新たなカテゴリを作成するように指示をする．

> カテゴリの例を以下に示します。
> カテゴリ名は【】で囲み、その下に該当文を列挙しています。また、一つの文が複数のカテゴリを持つこともあります。
>
> カテゴリの例
> ----------
> 【ファッションセンスの高さ】
> ・インスタグラムでは美しいドレス姿からレディーススーツまで様々な衣装を着こなしており、どれもイメージにぴったりのすらっとした美しさを最大限に引き出す着こなしとなっています。
> ----------
> これら例を元に、【ウェブページの要約】のカテゴリ分けを行ってください。
> 例で示したカテゴリ以外のカテゴリが存在する場合は、新たにカテゴリを作成してください。

このように，カテゴリの例を事前に示すことで，ChatGPT の出力を安定させることを狙っている．ここで実際に渡しているカテゴリの例とそれに該当する文は，今回のキーワードである「山田涼介の演技・凄い」とは全く関係ないものである．その後の指示としては，こちらの都合の良い形式で結果を出力してもらうために，具体例を用いて出力形式を指示する．具体的には，推定されるカテゴリ名とそれに該当するウェブページの番号の組を，該当ウェブページ数の多い順にランキング形式で出力させる．出力形式としては，json 形式で出力するように指示する．

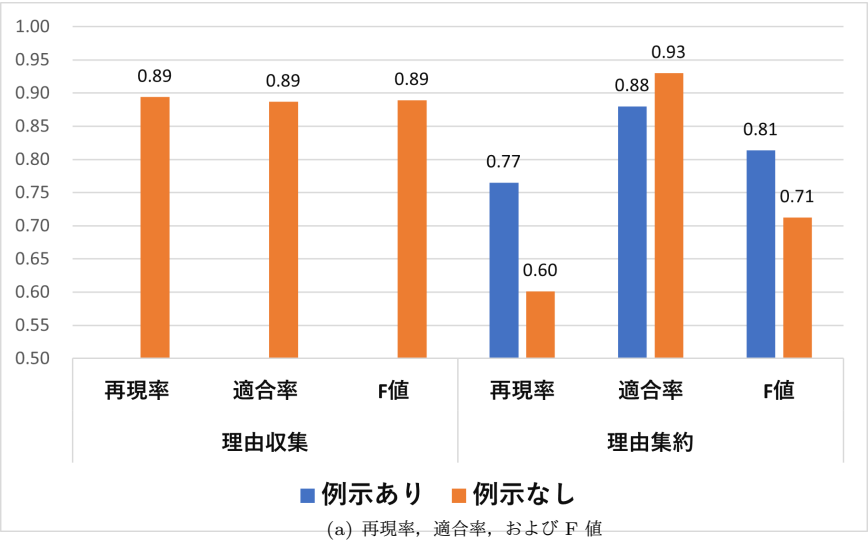
> 最終的な出力の形は以下です。
> 出力例
> ----------
> {
> ”1”: {
> ”カテゴリ”: ”【ファッションセンスの高さ】”,”該当するウェブページの番号”: [1,3,8,12]},
> ----------
> カテゴリ部分にはカテゴリを、該当するウェブページの番号の部分にはそのカテゴリに含まれるウェブページの番号を記述してください。
> カテゴリについては、理由内容が一目で推測できるようなカテゴリ名にすることを心がけてください。(不適切な例 :「その他の○○」、「関連性のない○○」)
> 該当するウェブページの数が多いものから降順にランキング形式で並べて出力してください。出力は json 形式にしてください。
> また、理由が示されていないものは除外してください。
>
> ここまでで一連の指示が完了しました。

このように，理由の集約後に，理由の頻度を考慮してランキングを行うことで，ユーザが感想の理由を確認しやすくなることを期待している．ここまでで，一連の ChatGPT に対する指示が終了し，最後に 3.2 節で作成したウェブページの要約を並べて ChatGPT に与える．プロンプトの具体例を以下に示す．

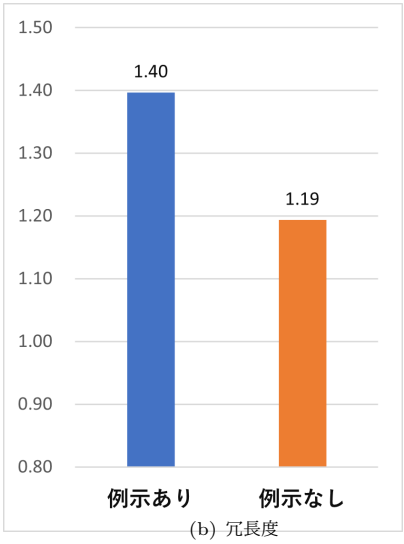
> 山田涼介の演技が凄い理由に関わる【ウェブページの要約】を以下で渡します。
> 【ウェブページ 1 の要約】山田涼介の演技が凄いと言われている理由は、彼の表情の表現力が豊かであ · · ·
> 【ウェブページ 2 の要約】山田涼介の演技が凄いと · · ·
> 【ウェブページ 3 の要約】山田涼介の演技が凄いと · · ·
> · · ·

表 2　キーワードごとの感想理由収集・集約の人手評価結果

芸能人名	キーワード	ページ数	理由収集 (例示なし)			理由集約 (例示あり)				理由集約 (例示なし)			
			再現率	適合率	F 値	再現率	適合率	F 値	冗長度	再現率	適合率	F 値	冗長度
山田涼介	山田涼介の演技・凄い	25	0.88 (=51/58)	0.94 (=51/54)	0.91	0.86 (=6/7)	1.00 (=6/6)	0.92	1.66 (=10/6)	0.57 (=4/7)	0.80 (=4/5)	0.67	1.75 (=7/4)
	山田涼介のドラマ・怖い	28	0.95 (=19/20)	0.86 (=19/22)	0.90	0.83 (=5/6)	1.00 (=5/5)	0.91	1.20 (=6/5)	0.83 (=5/6)	1.00 (=5/5)	0.91	1.20 (=6/5)
	山田涼介のドラマ・面白い	29	0.85 (=34/40)	0.85 (=34/40)	0.85	0.75 (=6/8)	0.86 (=6/7)	0.80	2.00 (=12/6)	0.63 (=5/8)	1.00 (=5/5)	0.77	1.20 (=6/5)
	山田涼介の顔・良い	29	0.95 (=20/21)	0.95 (=20/21)	0.95	0.50 (=4/8)	0.80 (=4/5)	0.62	1.25 (=5/4)	0.63 (=5/8)	0.83 (=5/6)	0.71	1.20 (=6/5)
菊池風磨	菊池風磨の演技・凄い	30	0.85 (=35/41)	0.90 (=35/39)	0.86	0.62 (=8/13)	0.89 (=8/9)	0.73	1.63 (=13/8)	0.54 (=7/13)	1.00 (=7/7)	0.70	1.29 (=9/7)
	菊池風磨の沼・深い	30	0.93 (=13/14)	0.87 (=13/15)	0.90	0.63 (=5/8)	0.63 (=5/8)	0.63	1.00 (=5/5)	0.25 (=2/8)	1.00 (=2/2)	0.40	1.00 (=2/2)
小栗旬	小栗旬の演技・凄い	30	0.82 (=37/45)	0.90 (=37/41)	0.86	0.85 (=11/13)	0.85 (=11/13)	0.85	1.36 (=15/11)	0.77 (=10/13)	1.00 (=10/10)	0.87	1.30 (=13/10)
	小栗旬の演技・悪い	30	0.94 (=15/16)	0.94 (=15/16)	0.94	1.00 (=8/8)	0.89 (=8/9)	0.94	1.00 (=8/8)	0.75 (=6/8)	1.00 (=6/6)	0.86	1.00 (=6/6)
	小栗旬の顔・良い	30	0.89 (=17/19)	0.77 (=17/22)	0.83	0.78 (=7/9)	0.88 (=7/8)	0.83	1.86 (=13/7)	0.22 (=2/9)	0.67 (=2/3)	0.33	1.00 (=2/2)
	小栗旬の声・良い	30	0.89 (=16/18)	0.89 (=16/18)	0.89	0.83 (=5/6)	1.00 (=5/5)	0.91	1.00 (=5/5)	0.83 (=5/6)	1.00 (=5/5)	0.91	1.00 (=5/5)



(a) 再現率，適合率，および F 値



(b) 冗長度

図 2　ChatGPT による感想理由収集・集約の人手評価結果 (表 2 のマクロ平均)

この一連のプロンプトを，全てのキーワードに対して適用する．

ここで，ChatGPT を利用して集約された感想の理由のうち，キーワード 10 組を対象としたものに対して，人手による感想の理由集約結果と比較することで，評価を行う．ここで，以下の 6 つを定義する．

S': ChatGPT が出力した理由の集合
R: 人手で用意した理由なり理由の集合
$S'_r = \{s' \in S' | s'$は対応する $r \in R$ が存在する $\}$
$S'_{\neg r} = S' - S'_r$
$S_r = \{s' \in S'_r | r \in R$ に対応する複数の s'を一つの理由に集約 $\}$
$S = S_r \cup S_{\neg r}$

具体的には，まず，事前に収集した理由をもとに，ChatGPT が集約した理由の集合である S' と，人手で用意した集約後の異なり理由の集合である R を定義する．次に，S' のうち，R 中に対応する理由が存在するものを S'_r と定義する．それに対し，R 中に対応する理由が存在しないものを $S'_{\neg r}$ と定義する．次に，S'_r のうち，R 中の一つの理由に対応する，複数の理由を，一つの理由に集約したものを S_r と定義する．それに対し，R 中に対応する理由が存在しないものを $S_{\neg r}$ と定義する．最後に，S_r と $S_{\neg r}$ の和集合を S と定義する．これを踏まえて，再現率・適合率は，以下のように定義される．

$$\text{再現率} = \frac{|S_r|}{|R|}, \qquad \text{適合率} = \frac{|S_r|}{|S|}$$

また，ChatGPT が理由の集約を行う際に，どれだけ冗長にならずに集約を行えたかを評価するため，冗長度を以下のように定義する．

$$\text{冗長度} = \frac{|S'_r|}{|S_r|}$$

これらの評価を，キーワードごとに行う．キーワードごとの評価結果については，表 2 の「理由集約 (例示あり)」の列に示す．全体の評価結果としては，キーワードごとの評価結果のマクロ平均とする．感想の理由集約における，全体としての評価結果を図 2 の「理由集約」に関するグラフのうち，「例示あり」に対応するもので示す．結果として，例示がある場合の理由集約においては，再現率が 8 割程度，適合率が 9 割程度，F 値が 8 割程度，冗長度が 1.4 となった。

3.4　理由集約におけるプロンプト内の例示の有効性評価

本節では，前節の理由集約において，プロンプト内でのカテゴリの例示が人手の評価結果について与える影響について検証する．具体的には，以下に示すカテゴリの例示に関連した指示を除外したプロンプトを使用し，前節の例示を行った場合の評

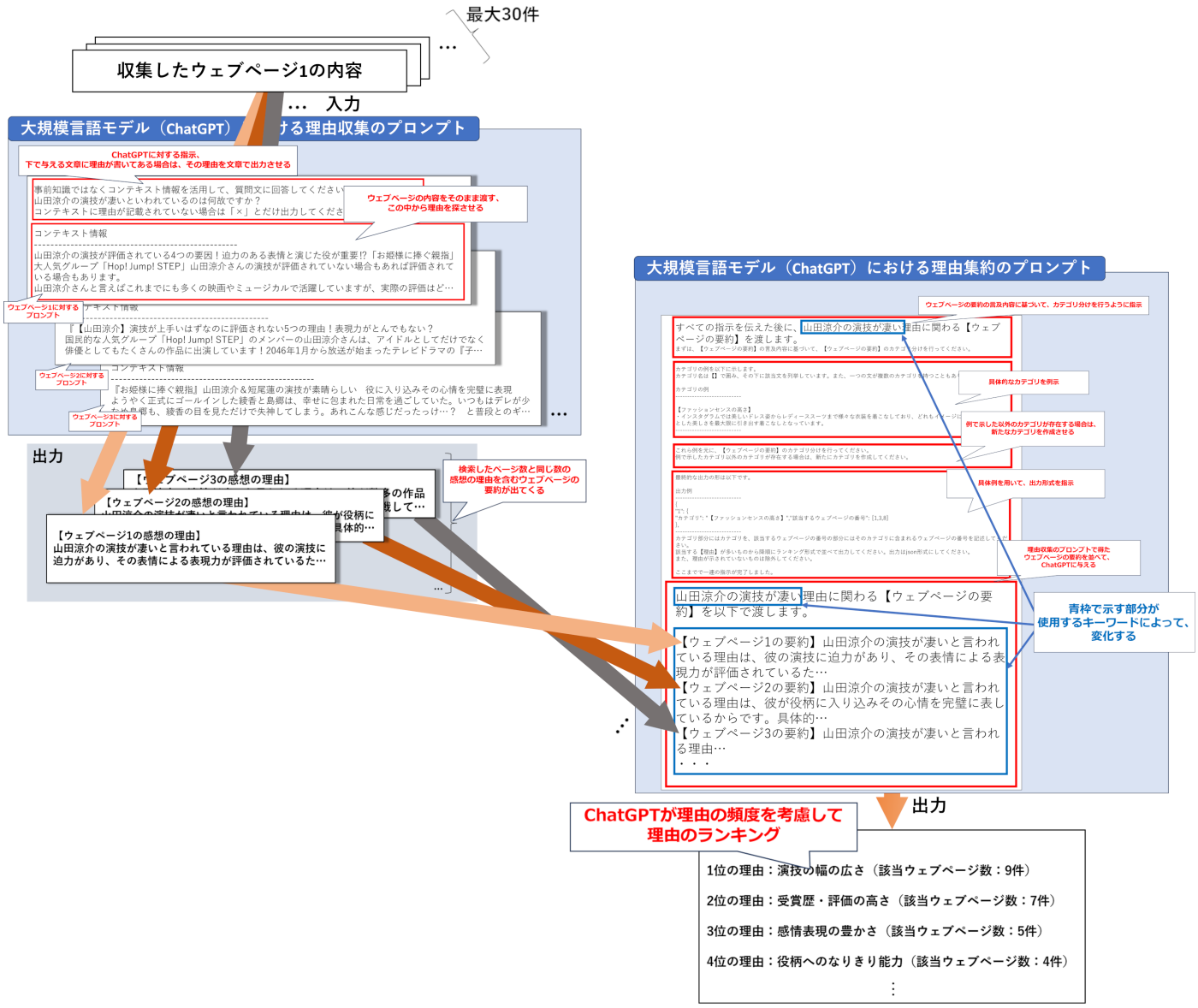


図 3　感想の理由収集・集約の流れとプロンプト

表 3　作成したデータセットのキーワードごとの内訳

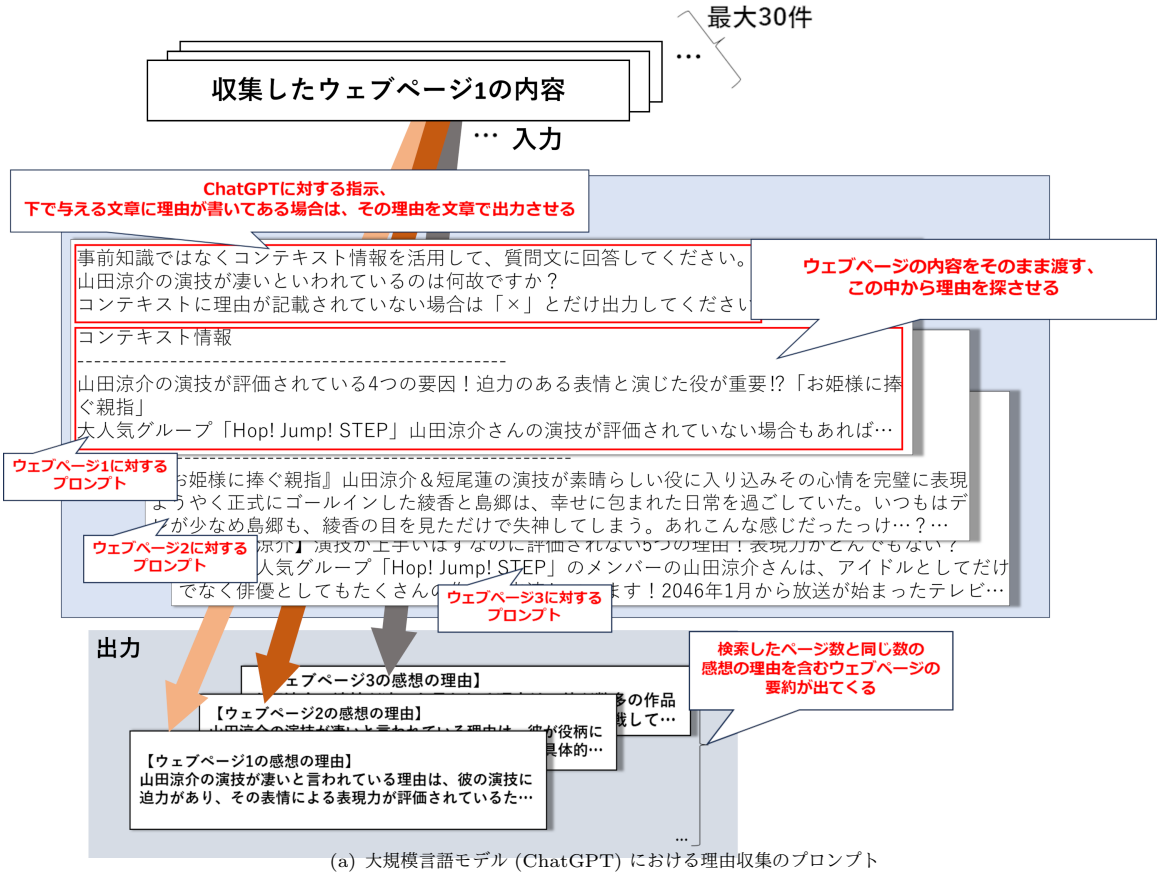
キーワード	データセット		
	理由あり	理由なし	合計
山田涼介の演技・凄い	56	56	112
山田涼介のドラマ・怖い	22	22	44
山田涼介のドラマ・面白い	36	36	72
山田涼介の顔・良い	20	20	40
菊池風磨の演技・凄い	33	33	66
菊池風磨の沼・深い	8	8	16
小栗旬の演技・凄い	42	42	84
小栗旬の演技・悪い	15	15	30
小栗旬の顔・良い	12	12	24
小栗旬の声・良い	13	13	26
合計	257	257	514

表 4　各ウェブページに対する参照情報 (理由あり・なし), および，モデルの予測 (理由あり・なし) の組み合わせに対するウェブページ数

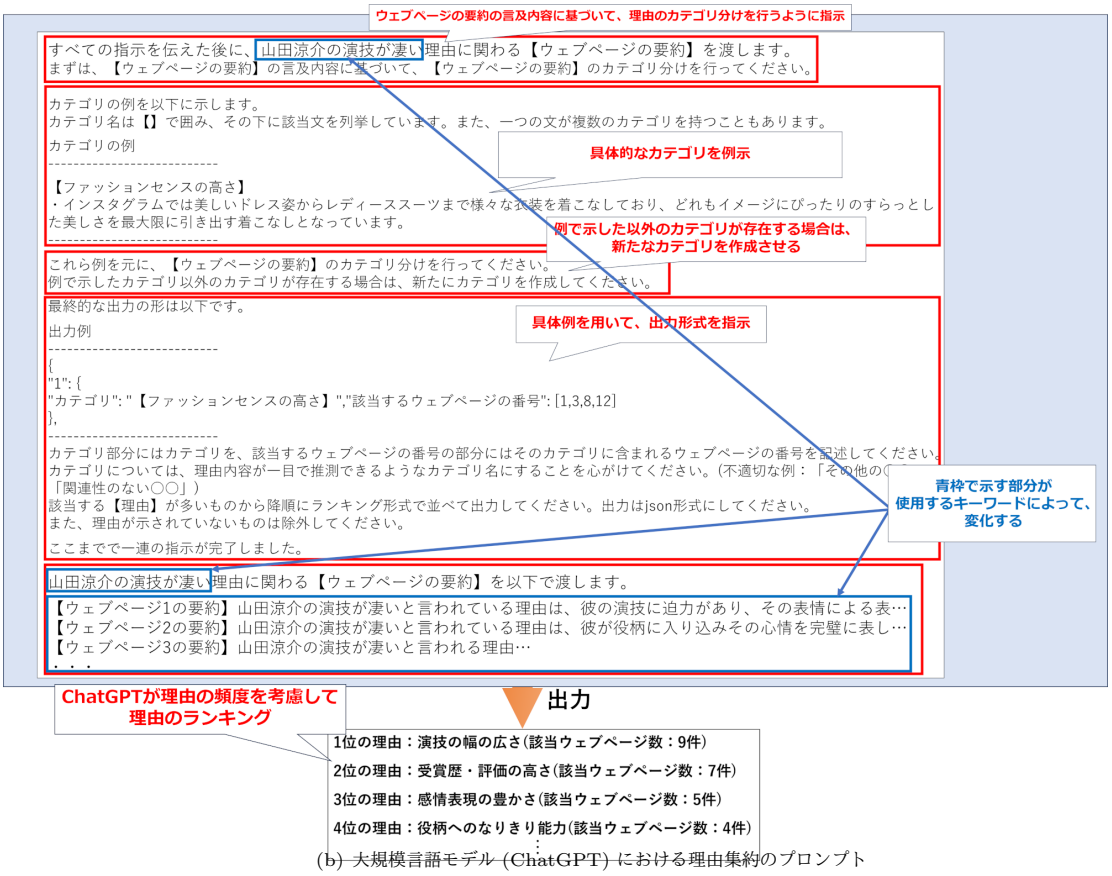
		モデルの予測	
		理由あり	理由なし
参照情報	理由あり	TP	FN
	理由なし	FP	TN

価結果との比較を行う．

カテゴリの例を以下に示します。
カテゴリ名は【】で囲み、その下に該当文を列挙しています。また、一つの文が複数のカテゴリを持つこともあります。



(a) 大規模言語モデル (ChatGPT) における理由収集のプロンプト



(b) 大規模言語モデル (ChatGPT) における理由集約のプロンプト

図 4　プロンプト

表 5　ChatGPT・mT5 による感想理由有無判定タスク・理由収集タスクの自動評価結果

モデル	感想理由有無判定タスク			理由収集タスク
	再現率	適合率	F1	ROUGE-L
ChatGPT	0.96	0.93	0.94	32.31
mT5	0.84	0.47	0.57	9.22

カテゴリの例

【ファッションセンスの高さ】
・インスタグラムでは美しいドレス姿からレディーススーツまで様々な衣装を着こなしており、どれもイメージにぴったりのすらっとした美しさを最大限に引き出す着こなしとなっています。

これら例を元に、【ウェブページの要約】のカテゴリ分けを行ってください。
例で示したカテゴリ以外のカテゴリが存在する場合は、新たにカテゴリを作成してください。

カテゴリの例示に関連した指示を除外したプロンプトを用いて理由集約を行い，その結果を人手で評価した．キーワードごとの評価結果を表 2 の「理由集約 (例示なし)」の列に示す．全体としての評価結果を図 2 の「理由集約」に関するグラフのうち,「例示なし」に対応するもので示す．例示を除外した場合，再現率が 6 割，適合率が 9 割程度，F 値が 7 割程度，冗長度が 1.19 となった。例示を行った場合と比較すると，適合率は 0.04 ポイント上昇しているものの，再現率は 0.17 ポイント低下する結果となった．結果として，適合率と再現率から算出される F 値は 1 ポイント低下した．冗長度については，0.21 ポイント改善されたが，これは再現率が低下したことによる影響だと考えられる．

これらの結果から，カテゴリの例示を除外することで，ChatGPT はより簡潔で的確なカテゴリ分けを行えるようになる一方で，人手で作成したカテゴリを見逃す傾向が強くなることがわかる．つまり，例示を行うことで，ChatGPT は人手のアノテーションにより近い結果を生成できるが，その代償としてより冗長なカテゴリ分けを行ってしまう可能性があることが示唆された．

4　ChatGPT および mT5 によるウェブページ中の理由有無判定・理由収集の自動評価

4.1　評 価 手 順

本節では，ChatGPT および mT5 によって，ウェブページ中の感想理由有無判定，および，感想理由収集結果に対する自動評価手順について述べる．

本論文では，3.2 節で述べたように，感想の理由の収集をウェブページから行っているが，理由に該当する文は一箇所に固まっていることは少なく，複数箇所にまたがっている場合が多い．そのため，コンテキストから理由部分を抽出するタスクよりも，コンテキストを元に理由を生成するタスクを適用する方が適切であるといえる．ここで，ChatGPT との比較対象として，T5 [6] の多言語版である mT5 [8] を用いる．

mT5[5]の fine-tuning 用のデータセットとしては，3.1 節で収集したウェブページを利用して作成を行った．具体的には，まず，ウェブページごとの収集キーワードを元に，質問文の設定を行う．例として,「山田涼介の演技・凄い」に関連して収集されたウェブページの場合，質問文には「山田涼介の演技が凄いと言われているのは何故ですか？」を設定する．次に，質問に回答するためのコンテキストの設定を行う．はじめに，収集したウェブページを句点で文単位に分割し，その後，mT5 の入力トークン数上限に収まるようにチャンクとして連結した．その後，チャンクごとに，感想の理由となる文が含まれる場合は，該当する文を全て人手で抜き出し，それを足し合わせたものを回答とした．感想の理由となる文が含まれない場合は，回答不可能なので,「」(空白) を回答とした．このように回答を作成することで，コンテキストに感想の理由が含まれる場合は，その感想を出力し，コンテキストに感想の理由が含まれない場合は，何も出力しないことを期待している．具体的なデータセットの内訳を表 3 に示す．

mT5 の訓練および評価については，表 3 に示すデータセットを元に，キーワード単位で 5 分割交差検証を行った．具体的には，表 3 に示す 10 個のキーワードのうち，8 つのキーワードに関連したデータセットを訓練用データ，2 つのキーワードに関連したデータセットを評価データとした．評価データに対する回答の生成については，まずは，mT5 の入力トークン数上限に収まるように分割したチャンク単位で回答の生成を行い，その後，生成された回答を，ウェブページごとにまとめたものをコンテキストとして，もう一度回答を生成するといった手順を踏んだ．こうすることで，3.2 節の ChatGPT と同じく，ウェブページごとに回答を生成することが可能となる．

以上の手順のうち，ChatGPT および mT5 による理由収集結果に対する自動評価指標として ROUGE-L を用いる．参照文については，3.2 節で作成した，ウェブページごとに人手で用意した理由の集合を用いる．また,「感想の理由がある場合には，感想の理由を出力し，感想の理由がない場合に，理由がない旨を出力する」感想理由有無判定タスクがどれだけ上手くできているかについても自動評価を行う．ChatGPT については，図 4 に示すプロンプトで，理由がない場合は「×」を出力するように指示を行っているため,「×」のみが出力された場合は理由なしとして扱う．mT5 については，出力が「」(空白) の場合に理由なしとして扱う．また，mT5 の出力には「。」や「?」などの記号のみが出力される事例も見られたため，それらについても理由なしとして扱う．各ウェブページに対する理由有無の参照情報と，モデルが予測した理由有無の組み合わせに対するウェブページ数は，表 4 に示す TP，TN，FP，FN で定義される．これらのウェブページ数 TP，TN，FP，FN に基づ

5 : https://huggingface.co/google/mt5-base

き,「感想の理由有無判定タスク」の再現率・適合率は次式で定義される.

$$\text{再現率} = \frac{TP}{TP+FN}, \qquad \text{適合率} = \frac{TP}{TP+FP}$$

4.2 評価結果

各手法に対する評価結果を表 5 に示す．本論文のタスクにおいて，どの評価結果についても，ChatGPT が mT5 を上回る性能が得られた．

ROUGE-L は，モデルが生成した回答と人手で作成した回答の，最長共通部分列を評価するものである．ROUGE-L の評価結果から，ChatGPT は mT5 と比較して，人手で作成した回答に近い出力が行えていることが伺える．

また，理由の有無に関する評価では，どの指標においても ChatGPT が mT5 の結果を上回っており，特に適合率で大きく差をつける結果となった．mT5 の適合率に対して，再現率は 8 割程度あることから，理由がない場合にも何かしらの文を出力してしまう FP に該当する事例が多いことが伺える．それに対し，ChatGPT は再現率，適合率，F1 ともに 9 割を上回っており，感想の理由がある場合には出力し，理由がない場合には理由がない旨を出力するタスクにおいて，非常に優れた結果となっている．

5 関連研究

芸能人を対象とした関連研究としては，ポスト内で述べられている特定の芸能人と形容詞との間で感想関係の有無の判定をした研究 [3] や，特定の芸能人の評価対象と形容詞の間で感想関係の有無の判定を行った研究 [7] が挙げられる．一方，本論文では，ポスト内の情報を基に得た，芸能人の評価対象に対する感想の理由を，ウェブページから探しているという点が上記の研究と異なる点である．

本研究の重要な側面は，LLM の出力の信頼性を向上させる上で，非常に大きな役割を果たす Retrieval Augmented Generation (RAG) [2] を使用していることである．RAG は，LLM が外部データベースから取得した情報を参照することで，出力の信頼性を向上させるものである．本論文では，RAG を活用して，ウェブページの内容に基づいて感想の理由を収集することで，LLM の事前知識のみに依存する場合と比較して，収集結果の信頼性を向上させることを狙っている．

ChatGPT を利用した関連研究としては，エンティティリンキングに関する研究 [4] や，対話分析に関する研究 [1] が挙げられる．ChatGPT によるテキスト要約では，抽出型要約に焦点を当てた研究 [10] や，要約の出力スタイルの調整に焦点を当てた研究 [5]，反復的な要約プロセスを提案した研究 [9] が挙げられる．一方，本論文では，ChatGPT を感想の理由の収集・集約で利用するといった点が上記の研究と異なる点である．

6 おわりに

本論文では，芸能人のファンが，芸能人に関連する事柄についての批評や関連情報の探索を行いやすくするための手法を提案した．提案手法で得られた結果に対して，人手で収集・集約した感想の理由と比較して，評価実験を行った．また，ChatGPT を用いた場合と mT5 を用いた場合での，感想の理由の収集についても評価を行い，ChatGPT の方が高い性能を示すことを確認した．今後の展望としては，今回は評価の対象としなかった，残り 14 件のキーワードに対しても，評価実験を進めていく．

謝　辞

本研究は科研費 21H00901 の助成を受けたものである．

文　献

[1] S. E. Finch, E. S. Paek, and J. D. Choi. Leveraging Large Language Models for Automated Dialogue Analysis. In *Proc. 24th SIGDIAL*, pp. 202–215, 2023.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proc. NeurIPS*, pp. 483–498, 2020.

[3] Y. Nozaki, K. Sugawara, Y. Zenimoto, and T. Utsuro. Tweet Review Mining focusing on Celebrities by Machine Reading Comprehension based on BERT. In *Proc. 36th PACLIC*, pp. 757–766, 2022.

[4] R. Peeters and C. Bizer. Using ChatGPT for Entity Matching. *arXiv preprint arXiv:2305.03423*, 2023.

[5] D. Pu and V. Demberg. ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer. In *Proc. 61st ACL-SRW*, pp. 1–18, 2023.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P J. Liu. Exploring the Limits of Transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*, pp. 1–67, 2020.

[7] 菅原琴絵, 宇津呂武仁. ツイートにおける芸能人の評価対象に対する感想のマイニング. 第 15 回 DEIM フォーラム論文集, 2023.

[8] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proc. NAACL*, pp. 483–498, 2021.

[9] H. Zhang, X. Liu, and J. Zhang. Extractive Summarization via ChatGPT for Faithful Summary Generation. In *Findings of EMNLP*, pp. 3270–3278, 2023.

[10] H. Zhang, X. Liu, and J. Zhang. SummIt: Iterative Text Summarization via ChatGPT. In *Findings of EMNLP*, pp. 10644–10657, 2023.

論文調査の支援に向けた研究
－最新の読むべき論文の判別方法－

佐々木 陸† 清水 敏之‡ 冨浦 洋一‡

†九州大学大学院統合新領域学府ライブラリーサイエンス専攻 〒819-0395 福岡県福岡市西区元岡 744

E-mail: †sasaki.riku.162@s.kyushu-u.ac.jp, ‡shimizu.toshiyuki.457@m.kyushu-u.ac.jp,

‡tomiura.yoichi.802@m.kyushu-u.ac.jp

あらまし 新たな研究着手の際，その研究テーマの扱う課題，手法などの周辺トピックを調査する．その際，レビュー論文は有用な情報源となるが，レビュー論文には，最新の論文が含まれていないことも多い．引用データベースで引用関係をたどり，最新の論文を探すことも可能であるが，引用の理由は様々で，全てが同じトピックを扱っているとは限らない．また論文の質を問わず，多数の論文が見つかるため，調査すべき論文を選定するには高い労力が必要となる．本研究では，被引用論文の抄録と引用元論文の抄録，導入，結論，引用文に基づいて，引用関係にある二つの論文が同じトピックを扱っており，かつその論文がそのトピックにおいて代表的かどうかを判定する方法を提案する．

キーワード 引用分類，引用文脈，機械学習，深層学習，BERT

1. はじめに

新たな研究着手の際、その研究テーマに関して，扱っている課題，手法とその性能，理論などの周辺トピックを調査する．その際，レビュー論文に引用されている論文を調査することは一つの有用な手段である．しかし，実際ある研究テーマで用いられている手法やその性能などに関して、最新の動向を論じている論文は，まだレビュー論文によって引用されていない場合がほとんどである．引用データベースを用いてレビュー論文で紹介されていた特定の論文を引用している比較的新しい論文を見つけることは可能だが，引用の理由はさまざまであり，引用元論文すべてが同じトピックを扱っているとは限らない.また論文の質を問わず，多数の論文が見つかるうえに，被引用数といった客観的な評価にもまだ差がないため，読むべきものを絞るのが難しいという問題点もある．そのため，最新の論文のサーベイは一般的な論文のものと比較して，より手間と時間を要する作業となっている．

そこで本研究では被引用論文の抄録，引用元論文の抄録，導入，結論，引用文に基づいて，引用関係にある 2 つ論文において，

- 引用元論文が被引用論文と同じトピックを扱っており(条件①)，かつ
- そのトピックを代表するよい(条件②)

論文を判定する方法を提案する．ここでいうトピックとは研究対象やそれに対する主要なアプローチを指す．この方法を用いて，与えられた研究トピックの最新の論文を見つけることができる．まず，レビュー論文等を用いてその研究トピックにおける代表的な論文を 1 つ見つけ，これをキー論文とする．引用データベースを用いて，キー論文を引用している論文を見つけ，提案法に基づいてキー論文を引用している論文がキー論文と同じトピックを扱っており，かつそのトピックを代表するよい論文かどうかを判定する．これによってキー論文を引用している論文の中で読むべき論文を厳選することができる．厳選された論文の中から新たなキー論文を設定し，それを引用している論文を判別することで上記の 2 つの条件を満たすより新しい論文を見つけることができ，最終的には客観的評価がまだない最新の論文の判別も可能になる．

また，本文中では論文 A が論文 B を引用しているとき，論文 A を引用元論文，論文 B を被引用論文と呼ぶ.

2. 先行研究

Valenzuela ら[1]は引用元論文をまず引用理由による 4 クラスに分類し，そのあとその引用理由によって重要か非重要かの 2 クラスへの分類を行った．この研究は引用を重要と非重要に分類する最初の研究である．分類には SVM とランダムフォレストを用いており，特徴量は引用総数、抄録の類似度といった合計 12 種類を用いていた．分類精度の向上のため，各特徴量の寄与率を求め，その比較，分析を行っていた．Saeed-Ul Hassan ら[2]も [1]と同じ研究を行っており，一般に公開されている [1]のデータセットを用いて実験を行った．ただし，[1]と異なる点として，テキストベースの特徴量を加えたより多くの特徴量と分類器（5 種類の分類器，SVM，ナイーブベイズ，決定木，k-近傍法，ランダムフォレストと既存特徴量に加え，6 種類の新規特徴量，引用ベース，キューワード

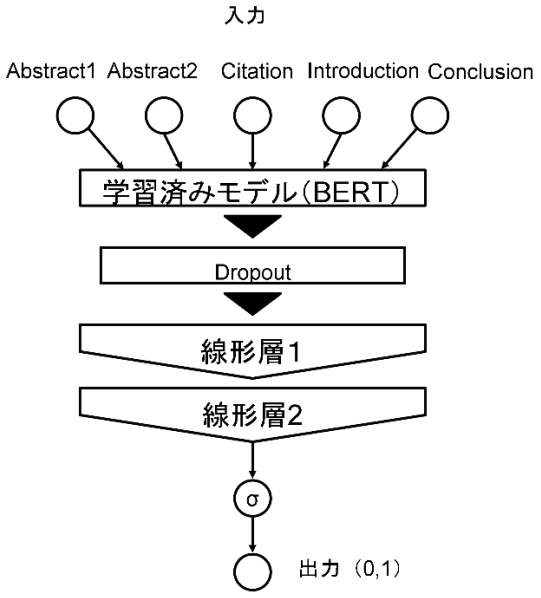


図 1　分類モデル

ベース，テキストベースを含む）を用いて実験を行い，よりよい分類精度を達成した．

Tirthankar Ghosal ら[3]も [1]のデータセットを用いており，引用が重要か文脈的（研究の流れの説明）なものかの分類を行い、その結果を用いて重要引用グラフを作成し，研究の系統を確立する方法を提案した．分類には 4 つの分類器と 30 個の特徴量を用いて実験を行った．これらの研究では，引用元論文が被引用論文を用いている(using the work)場合，あるいは被引用論文を拡張している（extending the work）場合に，被引用論文を引用元論文の重要論文としているが，引用元論文において，どれが重要な被引用論文であるかは読者である研究者によって異なるものである．それに対して，我々の研究では，重要という主観的なものではなく，引用関係にある２つの論文が同じトピックを扱っているか否か，またそのトピックにおいて読む価値のある代表的な論文であるかの判別を扱った．

Muhammad Roman ら[5]の研究では引用意図の自動注釈を提案している．彼らは 3 つ word embedding の手法を用いてクラスタリングを行った．そしてクラスタのサンプルを調査することで 3 つの引用意図（Background, Method, Result）でラベル付けしている．BERT を用いた場合が最もパフォーマンスが高く，Precision は 89%を達成している．これによって我々の研究でも，文章の意味を上手く扱える BERT[4]を用いた embedding を利用している．

3. 判別モデル

本研究では引用関係にある論文ペアのうち下記の 2 つの条件をともに満たす論文を判定する．

- 引用元論文が被引用論文と同じトピックを扱っている．
- そのトピックを代表するよい論文である．

これらの判定を自作したデータセットで学習した分類器モデルを用いて機械的に行い，最終的には判定結果を Recall，Precision で評価する．情報検索の一種で，見逃しが少ないことを評価するために Recall を重視する．

また，図 1 は本研究で用いる判別モデルである．本モデルでは引用関係にある論文ペアにおいて被引用論文の抄録，引用元論文の抄録，導入，結論，引用文（入力テキスト）を入力とし、引用論文が①，②の条件をともに満たしているかどうかの二値判定を行う．まず， BERT[4]を用いて，入力テキストの分散表現を求め，これらを連結したものを 2 値判別器に渡す．このとき BERT は学習率の1/100 倍で fine-tuning を行っている．また Word embedding のサイズを BERT の隠れ層の次元である 768 とする．その後の 2 値判別器は入力段で Dropout を適用する線形層を 2 つ用いて段階的に次元を落とす．1 つ目で 765 次元から 5 次元に落とし，2 つ目で 5 次元から 1 次元に落とする．その後，シグモイド関数を通すことで，最終的な出力は入力が正例である確率を表し，閾値 0.5 で正例・負例に判別する．

4. 実験

4.1. データセット（作成方法）

本研究では，データ作成者（annotator）の主観の影響がでないように，機械的な方法でデータを作成する．レビュー論文はある特定の研究トピックを中心に論文をサーベイしたものであるから，同一のレビュー論文で引用（紹介）されている論文は，同一のトピックを扱っていると考えられる．今回は「Citation Classification」で検索して見つけた 7 つのレビュー論文(付録)を調べ，その論文で引用（紹介）されている論文 X と X を引用している論文 Y のペアでデータセットを作成した．Y のうちレビュー論文 R で引用されている論文を A とすると，

X の Abstract,
A の Abstract,
A の Introduction,
A の Conclusion,
引用文脈,

からなる 5 つ組を作り，これを正例とする．ただし，引用文脈とは箇所の前後最大 4 文（引用を含む文，同じ段落内の前の 1 文（もしあれば），同じ段落内の後 2 文まで）である．もし同一の論文の引用が複数ある場合は，それらすべての引用箇所の引用文脈を連結したものを引用文脈とする．これらの特徴量は私たちが

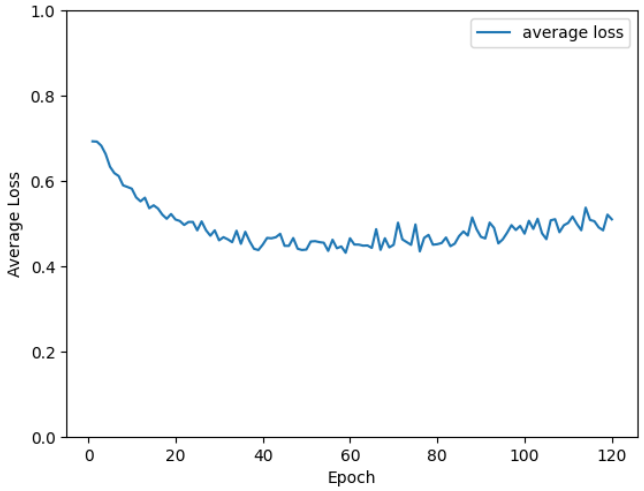


図 2　Average Loss

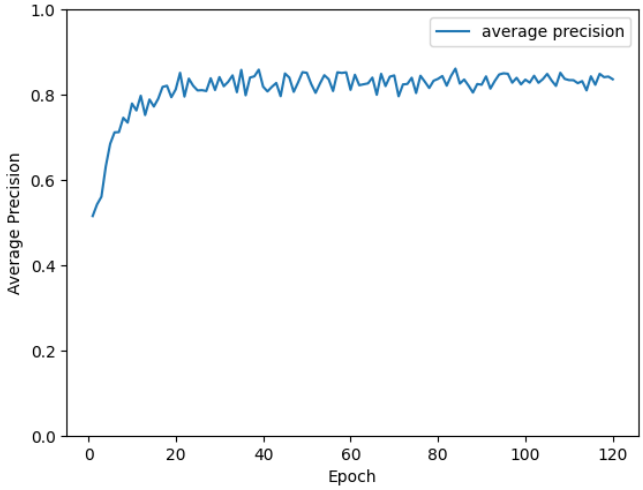


図 3　Average Precision

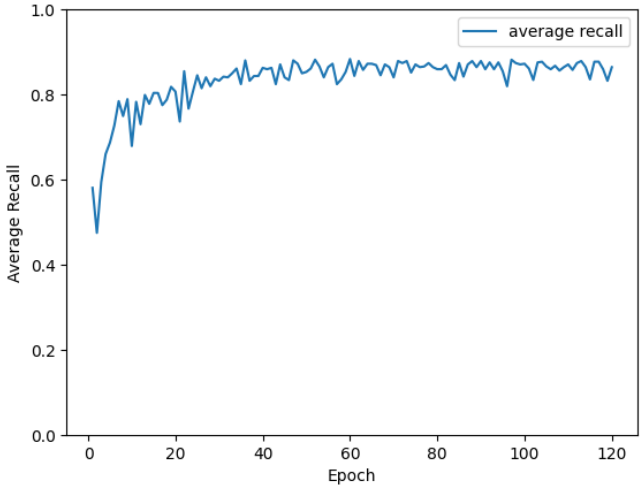


図 4　Average recall

実際に全文を読むべき論文かを判断するために読まれるものである．

また，Y のうちレビュー論文 R では引用されてない論文を B とすると，

X の Abstract，
B の Abstract，
B の Introduction，
B の Conclusion，
引用文脈，

からなる 5 組を負例とする．サンプル数は全体で 1257 件，うち正例が 626 件，負例が 631 件である．確かに，この方法でもレビュー論文の著者の主観は入っているが，彼らはこの作業の専門家であり，ラベル付けが必要な訓練データの作成においては最も信頼できる方法であると考えている．

4.2. 機械学習による分類

本実験での学習のパラメータは 2 値判別器の Dropout 率を 0.2，バッチサイズを 10，エポック数を 120 とする．またオプティマイザーのパラメータは 1 つ目の線形層の学習は行わず，2 つ目の線形層の学習率を 5e-5 と設定する．オプティマイザーのパラメータについては周辺の値を試した中で最もよかったものを設定する．使用した学習済み BERT[4]は transformers ライブラリ（https://pypi.org/project/transformers/）から入手する．

4.3. 評価方法

学習には 5-クロスバリデーションを用いる．各 Fold で得たメトリクス平均をとって評価を行う．その際の全データに対するテストデータの割合は Fold 1，2 が 20.05%，Fold3，4，5 が 19.97%である．

5. 結果

5-クロスバリデーションで行った 5 つの fold のテストの結果を平均した Loss，Precision，Recall をグラフに示す．横軸にエポック数をとり，縦軸に各メトリクスの値をとった．まず，図 1 で Loss がきれいに減少しており，学習がうまく進んだことが期待される．また図 2 から Precision は約 83%，図 3 から Recall は約 87%を達成していることがわかる．

6. 考察

6.1. 議論

結果の分析本実験では条件①，②の判定をまとめて行っている．またデータセットの負例は正例を構成する引用元論文の中からランダムに得ているため，条

表 1　正例・負例の引用元論文の被引用数

正例

0～10	11～50	51～100	101～500	501～
14	167	128	258	59

負例

0～10	11～50	51～100	101～500	501~
277	222	59	67	6

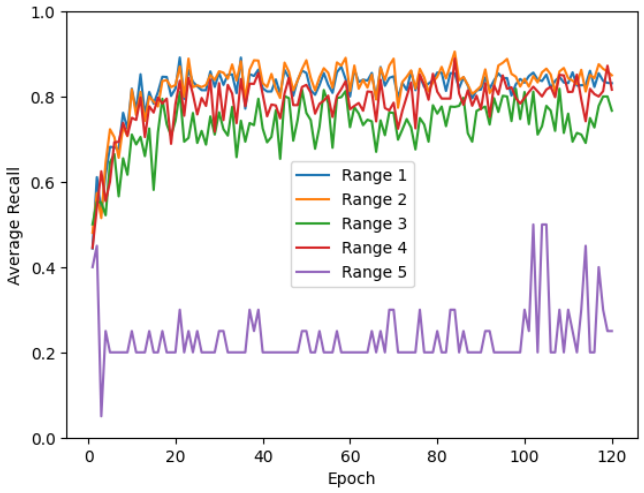


図 5 Range ごとの Average Recall

件②を満たすような論文はあまり含まれていないと推測される．実際，正例と負例それぞれを構成する引用論文の被引用数を調べた結果，表 1 のようになった．

この結果，正例と負例の判別の差が論文の質の差であることが考えられ，判別モデルが条件②の判別を強く学習し，総合的な判定も条件②の判定に強く影響を受けている可能性がある．つまり，判別モデルが引用元論文の被引用数が高いと予想されるものを正と判定していることが疑われる．そこで引用元論文の被引用数と判別精度，特に Recall と関係があるのかを調べる検証を行った．

6.2. 検証

データセットの論文ペアのうち，引用元論文の被引用数を調べ，被引用数について 0 以上 10 以下，11 以上 50 以下，51 以上 100 以下，101 以上 500 以下，501 以上の 5 つの Range を作った．そして，ラベルが負例のものについて，この 5 つの Range ごとに判別モデルが負例であると正しく判別できた割合（負例の Recall）を図 5 に示した．これによると，Range5 はサンプルが少なかったためにグラフの形状はきれいではないが，明らかに Recall が低かった．一方，それ以外の 4 つの Range 間では大きな差異はなかった．これによって判別モデルは被引用数が 501 件以上極端に良いものは，条件②の学習から正例と間違えることがあるが，それ以下のものでは引用元論文の被引用数が高いと予想されるから正例であると判定しいるのではないことが示された．

7. 結論

本研究ではある研究トピックにおいて客観的評価がない最新の論文の検索方法について分散表現を入力とする判別モデルを提案した．この判別モデルでは引用関係にある論文のペアに対して，引用元論文が被引用論文と同じトピックを扱っているか，また引用元論文がそのトピックを代表するよい論文か，という 2 つの条件をともに満たしているものの判定を行った．被引用論文の抄録，引用元論文の抄録，導入，結論，引用文を集めた自作のデータセットで 5-クロスバリデーションを用いて実験を行った結果，Precision は約 83%，Recall は約 87%を達成した．また 2 つの条件の判定をまとめて行っている点について，引用元論文がそのトピックを代表するよい論文であるか，という条件だけを強く学習しているのではないかという疑問点があった．そこで，判別モデルの判定結果と引用元論文の被引用数の関係を調べる検証を行った．その結果，被引用数が 501 件以上の極端によい論文については条件②の学習の影響で，他の Range より正例であると間違えて判定をすることが多かったが，500 件以下のものについては被引用数と判定精度に因果関係はみられなかった．よってこの判定モデルは基本的には論文の質だけで判別を行っているわけではないが，ある一定以上の被引用数を持つ論文については，条件②の影響が強く表れ、間違えて正例と判定されやすいことがわかった．

本研究では最終的に Precision は約 83%，Recall は約 87%という高い判別精度を達成したが，今回は比較実験を行わなかったパラメータも複数存在し，それらを含めてより細かく調整を行うことで，同モデルでもさらなる精度向上が見込まれる．また本実験で用いたテストデータはキー論文を引用しているすべての論文を集めることは数が膨大で不可能だったため，正例と負例が 1 : 1 になるように集めた．しかし，実際には負例が圧倒的に多いため，キー論文を引用しているすべての論文を集めた大規模なテストデータにおいてテストを行い，本研究で達成した Recall を維持すると，Precision はどれほど下がるのかを調べる必要がある．また実際に検索システムに活用するときには，正例と判別された引用 元論文を新たなキー論文とする判別を繰り返すことで，引用関係をたどり，キー論文から 条件①，②をともに満たす最新の論文を検索

できることを理想としている．しかし，この判別モデルでは正例である確率が 50%より大きいもの正例と判定するため，間接的な引用 関係にある論文間ついては直接的な引用関係で達成した判別精度を保てない．現状では引 用関係にある論文間の判別を行い，正例と判別されたものの本文を読み，人の目で条件を満たしていることを確認できたものを新たなキー論文として引用関係をたどる必要がある．また，本研究で作成したデータセットの論文は本研究トピックである引用分類に関するものが多いが，このデータセットで学習した判別モデルが他分野の論文に対しても機能するのかを調べる必要がある．他分野の論文においても同様の精度を達成できれば，より多くの研究者の役に立てるものになると考えている．

8. 付録

本データセットで用いたレビュー論文

- Review of Text Classification Methods on Deep Learning
- Text classification using embeddings: a survey
- Feature selection for text classification: A review
- Machine learning in automated text categorization
- Deep learning for time series classification: a review
- A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures
- A Survey of Modern Authorship Attribution Methods

参 考 文 献

[1] Marco Valenzuela, Vu Ha and Oren Etzioni(2015). Identifying Meaningful Citations .AAAI Workshop: Scholarly Big Data Mar-1-2015, WS-15-13, page 21-26

[2] Saeed-Ul Hassan, Anam Akram, Peter Haddawy(2017). Identifying Important Citations using Contextual Information from Full Text .JCDL '17: Proceedings of the 17th ACM/IEEE Joint Conference on Digital LibrariesJune 2017, Pages 41‑48

[3] Tirthankar Ghosal , Muskaan Singh(2021). Towards Finding a Research Lineage Leveraging on Identification of Significant Citations. Proceedings of the ASIS Annual Meeting 2021,Volume 58, Issue 1, Pages 456-460

[4] Jcob Devlin, Ming-Wei Chang, Kenton Lee,Kristina Toutanova, BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding.in Proc. Of NAACL-HLT 2019, Pages 4171-4186

[5] Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubaa, Lisu Yu. Citation Intent Classification Using Word Embedding. IEEE ACCESS 2021, Volume 9, Pages 9982-9995

手続き的知識源学習によるタスク検索における方法と目的の生成

工家　昂之†　　山本　岳洋†　　荘司　慶行††

† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
†† 静岡大学 情報学部 行動情報学科　〒 422–8529 静岡県静岡市駿河区大谷 836
E-mail: †ad22l024@guh.u-hyogo.ac.jp,t.yamamoto@sis.u-hyogo.ac.jp, ††shojiy@inf.shizuoka.ac.jp

あらまし　本稿では，あるタスクについて，そのタスクを達成に必要なタスク，そのタスク達成と同時に達成されるタスク，あるタスクを共通して達成可能なタスクをユーザに提示する手法を提案する．例えば，タスク達成のために必要なタスクの提示とは,「適度な運動をする」タスクの場合「ランニングを始める」タスクの提示をすることである．提案手法では，文生成モデル T5 を用いて手続き的知識源を学習し，入力に応じたタスクを生成することで提示を行う．比較手法は，先述した 3 種類の提示を 1 つの T5 モデルで行うマルチタスク学習と，1 種類を専用に学習するシングルタスク学習の 2 種類の学習方法を比較する．評価実験では，学習に用いた wikiHow データセットより作成したテストデータセットと，独自に作成した自作データセットによる生成結果を人手で評価した．実験結果より，平均正解種類数，平均逆順位においてマルチタスク学習を施した T5 が優れた値を示した．

キーワード　情報検索，Web 検索，タスク検索

1　はじめに

近年では，検索エンジンの普及で，あらゆる日常的な解決方法や達成方法を調べることができるようになっている．タスク検索とは「あるタスクを達成するために必要なタスクの検索」[15] で，ユーザは日常的にタスク検索を実行している．例えば,「健康を維持する」タスクを達成したいユーザが，その達成方法である「適度な運動をする」というタスク達成のためにタスク検索を実行すると,「ランニングを始める」というタスクが得られる．また,「健康を維持する」タスクの「適度な運動をする」タスク以外の達成方法として,「がんを予防する」というタスクも存在する．このように，日常的に実行されるタスクの達成方法には，複数の達成方法が存在することが多い．

しかし，既存の検索エンジンでタスク検索を実行する際,「適度な運動をする」タスクから「ランニングを始める」タスク,「がんを予防する」タスクの全てを発見することは難しい．例えば,「健康を維持」したいユーザがその達成方法の 1 つである「適度な運動」を既存の検索エンジンで検索する場合を考える．この場合,「適度な運動」というクエリに対して，検索システムより「目安」,「効果」,「何する」などが推薦される．確かにこのクエリ推薦では,「適度な運動」タスクを達成するために必要なタスクを知ることは可能だが,「適度な運動」タスクの達成で達成されるタスク「健康を維持する」についての推薦はできていない．ユーザの「適度な運動」という入力に対して,「食生活に気を付ける」をクエリ推薦できれば，ユーザが気付かない，または，未知の情報獲得ができると考えている．

そこで本研究では，(i) ユーザが入力したクエリを達成するために必要なタスク，(ii) クエリを達成すると同時に達成されるタスク，(iii) クエリとは別のタスクでクエリと共通するタスクを達成可能なタスク，のタスク検索を支援するシステムを提案する．

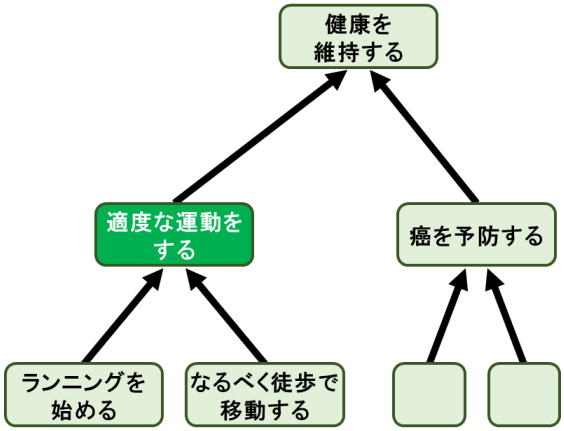


図 1　タスク検索におけるクエリ推薦.

図 1 には「適度な運動をする」というクエリを入力した場合に提示されるタスクを示している．例えば,「ランニングを始める」は「適度な運動をする」を達成可能であるため，(i) を満たす.「適度な運動をする」を達成すると同時に「健康を維持する」も達成されるため，(ii) を満たす．また，図 1 中の「食生活に気を付ける」を達成すると「健康を維持する」も達成されるため，(iii) を満たす．このタスク検索の実現で従来の検索システムには難しい，ユーザがクエリにできないタスクの検索も可能になる．

タスク検索では，実行するタスクとあるタスクを達成すると同時に達成されるタスクがある．本研究では前者を**方法**，後者を**目的**と呼ぶ．例えば，2 種類のタスク「適度な運動をする」と「ランニングを始める」があるとき，タスク「ランニングを始める」を達成することで，タスク「適度な運動をする」も達成される．このとき図 2 のようにタスク「ランニングを始める」を始点，タスク「適度な運動をする」を終点とした矢印を

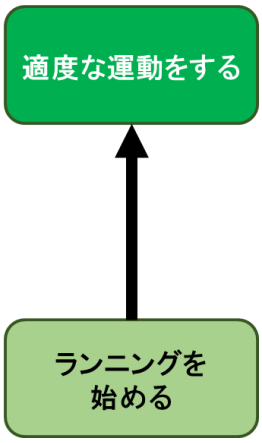


図 2　方法と目的の例．矢印の始点が方法，終点が目的を表す．

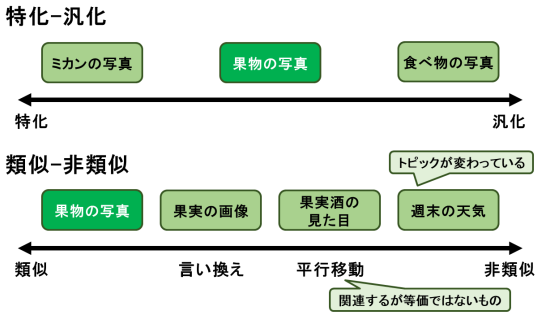


図 3　クエリ修正における特化，汎化，平行移動についてを表す．

書く．より具体的には，この矢印の始点を方法，終点を目的と定義する．

本研究の提案手法は，マルチタスク学習が可能な文生成モデル T5 のファインチューニングにより，ユーザの入力に対して方法，目的を生成することである．つまり，(i) 〜 (iii) のタスク検索を 1 つのモデルで支援するモデルを構築する．ファインチューニングに必要となるデータセットは wikiHow より構築した．wikiHow は手続き的知識源を集めたまとめサイトであり，クローリングによりウェブサイトのデータを収集する．その後，wikiHow の記事タイトルと章見出しのみを抽出しデータセットを構築した．

提案手法の有効性を評価するために，日常で検索されると考えられるタスクを考案し，wikiHow データセットのテストデータと共に評価に用いた．実際の評価では，まず，人手で適合不適合を評価した後，平均正解数，平均正解種類数，平均逆順位を計算し，これをもとにモデルの有効性を評価した．また，T5 のマルチタスク学習性能を検証するために，(i) 〜 (iii) をそれぞれ専用に学習したシングルタスク学習と結果を比較した．その結果，マルチタスク学習を施したモデルが高い平均正解種類数，平均逆順位を達成した．

2　関 連 研 究

本研究と関連が深い，タスク検索および，クエリ推薦，手続き的知識源に関する研究について述べる．

2.1　タスク検索

加藤らは，あるタスクを達成するために必要なタスクの集合を Web から発見する「タスク検索」を提案している．加藤らの提案手法では，実行することでユーザが入力したタスクを達成できるタスクを Web 検索により発見し，一連のタスク集合を形成する．その後，形成したタスク集合について，実行することでタスクを達成できる度合いのランキングを行うというものであると報告している [15]．

ある目的を達成するために必要な方法の検索をタスク検索と呼ぶ．例えば，健康を維持するために運動をしようと考えたユーザは，「適度な運動をする」というクエリで検索をする．しかし，「健康を維持する」という目的を達成する別の解決策として「食生活に気を付ける」方法も存在するが，ユーザはこの解決策の存在に気付くことができない可能性がある．もし，この解決策に気づけない場合，ユーザはこれらの解決策を考慮せずに「適度な運動をする」というクエリより提示される「ランニングを始める」や「筋トレをする」などの解決策を選択する．この問題を解決するために，Pothirattanachaikul らは特定のクエリの代替アクションをコミュニティ Q&A サイトからマイニングする方法について報告している [9]．

Zhang らは手続きに関する知識はタスク検索クエリに応答する際に重要であると報告している [14]．

タスク検索においてあるタスクを完了するために必要なサブタスクが存在する．このサブタスクとは，完了すると目的のタスクが完了するようなタスクである．このサブタスクのマイニングについて Yamamoto らは，検索連動型広告を利用したクエリのクラスタリングによる取り組みについて報告している．彼らはあるタスクとそのサブタスクについて，search goal と subgoal という概念を提案している [12]．本研究において目的が search goal，方法が subgoal と対応している．

Hassan らは，タスクを完了するために必要な手順をユーザに提供するために，タスク間の関連性を自動的に発見する手法を提案している [1]．

2.2　クエリ推薦

Boldi らはクエリ修正の分類について報告している．あるクエリが与えられ，そのクエリが修正されたとき，修正前後の関係が特化であるのか，あるいは汎化であるのかの分類を行っている [3]．例えば，図 3 のように「果物の写真」がクエリとして与えられたとする．特化，汎化の軸ではクエリが「ミカンの写真」へ修正された場合，特化と分類する．また，類似，非類似の軸では「果実の画像」,「果実酒の見た目」のような言い換えや，関連するが等価でない修正を分類する．本研究における達成方法の提示や検索目的の抽出はこの特化，汎化に該当し，同じ目的を達成可能な別の方法の提示が類似に該当すると考えている．

大石らは検索エンジンに与えるクエリを改善するクエリ拡張

を行うために，重要な語の近くに出現する単語は重要であるという考えに基づいた，センテンス間の距離に注目した関連単語抽出アルゴリズムの提案を報告している [17].

Bhatia らはクエリログが少ない状況でクエリ推薦を行う場合，クエリログを使用せずに確率論的メカニズムでクエリ推薦を行い，既存の試みよりも高品質なクエリの提案が行えたと報告している [2].

田麥らは，ユーザの目的推定に関する研究として，サービスの検索において，入力されたクエリに対してどの種類のサービスを利用したいかの推定について報告している．なお，クエリとサービスの種類の結び付けには検索連動型広告を利用している [18].

Zamani らは，クエリを提示する検索システムの提案を行っている．彼らは，ユーザが入力するクエリは短く，抽象的であるという仮説を立て，ユーザ実験を実施した．実験ではクエリの明確化をシステムが行い，それをユーザに提示するというユーザ実験である．実験結果では「検索しやすくなった」や「商品検索においてストレスが軽減された」とアンケート結果が得られたと報告している [13].

Cao らは，ユーザの意図を考慮したクエリ推薦を検索ログから実現するシステムの提案をしている．彼らは，既存システムの推薦精度が低い理由について，ユーザの検索意図がわからないシステムが様々な検索意図の可能性に対応するため，ユーザの意図に沿わない推薦をする必要があると述べている．彼らの意図を考慮する提案手法は，検索ログ中のクリックスルーデータとセッションデータから意図とクエリの対応付けを行い，意図の遷移についてグラフを作成し，クエリ推薦として遷移先の意図に属するクエリを推薦する，という手法である [4].

2.3 手続き的知識源

一般的な知識の多くは特定の目的を達成するための手順や行動の連続として存在している [6]．既存の知識構築システムとして Singh らは，Open Mind Common Sense というウェブサイトを構築し，知識のデータベースを構築したと述べている．このウェブサイトでは，難しい問題をインターネット上の数千人のボランティアに作業を分散させるという手法を取っており，非専門のネットユーザでも共同で知識のデータベースを構築することを可能にした [11]．手続き的知識源とは何かをする際のやり方に関する知識であり，本研究では，方法と目的のペアを手続き的知識源と呼ぶ.

3 提案手法

本節では，方法と目的を生成する手法について説明する．まず，タスク検索の支援における，タスク提示の種類について説明する．その後，それぞれのクエリ推薦を実現するための生成手法について述べる.

3.1 問題定義

本研究における問題定義は，あるタスクの入力に対して，そのタスクを特化，汎化，平行移動させたタスクを出力することである．具体的な説明を図 1 を用いて行う．タスクの特化では，入力タスクに対してその達成方法を出力する．例えば,「適度な運動をする」の達成方法として「ランニングを始める」,「筋トレをする」を出力することである．タスクの汎化では，入力タスクを達成することで達成される目的を出力する．例えば,「適度な運動をする」を達成することで達成される「健康を維持する」を出力することである．タスクの平行移動では，入力タスクを達成することで達成される目的を達成可能な入力タスク以外の方法を出力する．例えば,「適度な運動をする」の達成と同時に達成される「健康を維持する」を達成可能な「食生活に気を付ける」を出力することである．特化，汎化，平行移動のという名前は，図 3 に示す Boldi らの研究を参考にした [3]．なお，図 3 中の類似-非類似について「非類似」と「言い換え」は本研究では扱わないため，同じ目的を達成可能な別の方法の提示を「平行移動」とする.

3.2 文生成モデル T5

本研究では方法と目的の生成に文生成モデル T5（Text-To-Text Transfer Transformer）を用いる．T5 は Encoder-Decoder モデルで，入力した文情報を Encoder で集約し，集約した情報をもとに Decoder で文を生成するモデルである．T5 では事前学習とファインチューニングの 2 種類の学習を行う．まず，事前学習で大規模なデータセットを用いて普遍的な学習を行い，ファインチューニングで小規模なデータセットを用いて必要なタスクに特化させるという学習である．事前学習とファインチューニングのように学習を分けることで，大規模なデータセットを用意せずとも，小規模なデータセットで解きたい問題に対して速やかにモデルを構築できる．また，T5 の事前学習には次の 3 種類のタスクが検討された.

- Prefix language modeling
- BERT-Style
- Deshuffling

Prefix language modeling は文の出だしが表示され，その後にどう続くかを予測するタスクである．BERT-Style は文章中の 15%の語を伏字にした後，その 10%をランダムな語に置き換えて元の文章を復元させるタスクである．また，Deshuffling 語の順番を入れ替えて元の文章を復元するタスクである．上記 3 つのタスクでは BERT-Style のスコアが最良だったことが報告されている [10]．本研究で用いる T5 は事前学習に日本語版 Wikipedia のダンプデータ (2020 年 7 月 6 日時点) を使用し，BERT-Style を事前学習で学習している，sonoisa/t5-base-japanese を用いる [1].

T5 では**マルチタスク学習**と呼ばれる学習が可能で，これは単一モデルが同時に複数のタスクを解けるようにする学習である．本研究では特化，汎化，平行移動タスクを単一モデルに学

1：https://huggingface.co/sonoisa/t5-base-japanese（2022 年 11 月 24 日閲覧）



図 4　wikiHow の記事

習させるために，タスクを識別する特殊な文字列（特殊トークン）を入力側に付与した．各学習タスクと特殊トークンの関係を表 1 に示す．

3.3　wikiHow を用いた訓練データ構築

3.2 節で小規模なデータセットでモデル構築ができると説明したが，小規模とはいえこれを人手で構築するには膨大な時間がかかる．また，本研究で必要な手続き的知識源を集めた既存データセットは存在しない．そこで，自動で訓練データを構築するために，本研究では手続き的知識源を集めたまとめサイト wikiHow を用いる．

図 4[2]に「テレビで YouTube を見る」というタイトルの wikiHow の記事を示す．図 4 より wikiHow では記事タイトル「テレビで YouTube を見る」を達成する方法として「スマートテレビやゲーム機を利用する」,「Chromecast を使用する」「Apple TV を利用する」のように複数の達成方法が章見出しとして示される．

wikiHow の記事タイトルと章見出しをそれぞれ目的と方法の関係に位置付けることで，訓練データの自動構築を実現する．

3.4　データセット

本節では，本研究で用いるデータセットとその作成方法について説明する．本研究では wikiHow データセットと自作データセットの 2 種類のデータセットを構築した．wikiHow データセットでは，文生成モデルのファインチューニングとモデルの性能評価のためのテストデータとして用いる．自作データセットでは，全てを性能評価のためのテストデータとして用いる．

3.4.1　wikiHow データセット

3.1 節で説明した 3 種類のクエリ推薦を実現するために，特化，汎化，平行移動の 3 種類のデータセット構築を行う必要がある．日本語版 wikiHow より 2,560 件の記事をクローリングにより収集した後，各記事を訓練，検証，テストデータに 9 : 0.5 : 0.5 の割合で分割した．続いて共通の前処理として各記事より記事タイトルと章見出しを抽出する．その後，次の a）〜c）に従ってデータセットを構築した．

a）　特化データセット

特化では，あるタスクを達成するためのタスクを提示する．つまり，目的からその方法を提示する．これを図 4 の記事を用いて説明すると，記事タイトル「テレビで YouTube を見る」より章見出し「スマートテレビやゲーム機を利用する」,「Chromecast を使用する」,「Apple TV を使用する」を出力することに置き換えられる．よって，特化データセットでは入力が記事タイトル，出力が章見出しとなるデータセットを作成する．

図 4 の記事にあるように，目的に対して複数種類の達成方法が存在するため，複数種類の生成が可能なモデル構築をする必要がある．本研究で用いるモデルでは，一般的に 1 つの入力に対して複数件の出力を得ることが可能で，この出力の順番は入力に対する正解である確率が高い順に出力される．しかし，先述した順番の出力は語尾が多少変化した程度の意味的に類似した生成が並び，本研究で必要な生成結果は得られない．解決策として，生成時のパラメータ設定などで各出力間に多様性を持たせることが可能であるが，綿密なパラメータ調整が必要になる．また，パラメータ調整で多様化せずに，モデルが出力する順番で生成した後，各出力間の類似度をもとに生成結果を並べ替えることで多様化する試みもある [16]．よって本研究では，多様性を持たせる学習を施すことで，生成時には最低限のパラメータ調整で多様な生成を行う．

出力結果に多様性を持たせるためのデータセット作成方法として，1 種類の記事タイトルについて 3 種類の章見出しを連結したデータセットを作成した．また，章見出しが 3 種類以上存在する場合，その組み合わせで出力を複数種類作成し，記事タイトルと対応付ける．例えば，n 種類の章見出しがある場合その記事からは ${}_nC_3$ 個のデータセットが作成できる．

b）　汎化データセット

汎化では，あるタスクを達成すると同時に達成されるタスクを提示する．つまり，方法から目的を提示する．これを図 4 の記事を用いて説明すると，章見出し「スマートテレビやゲーム機を利用する」,「Chromecast を使用する」,「Apple TV を使用する」それぞれから記事タイトル「テレビで YouTube を見る」を出力することに置き換えられる．よって，汎化データセットでは入力が記事タイトル，出力が章見出しとなるデータセットを作成する．また，汎化データセットでは，複数章見出しに対して 1 種類の記事タイトルの関係であるため，1 入力に対して複数出力を持たせることはできない．

c）　平行移動データセット

平行移動では，共通するあるタスクを達成可能な複数のタスクを提示する．つまり，方法から別の方法を提示する．これを図 4 の記事を用いて説明すると，記事タイトル「テレビで YouTube を見る」を達成可能な章見出し「スマートテレビやゲーム機を利用する」から別の章見出し「Chromecast を使用する」,「Apple TV を使用する」を出力することに置き換えられる．よって，平行移動データセットでは入出力ともに方法からなるデータセットである．特化データセットと同様に出力結

2：「https://www.wikihow.jp/テレビで YouTube を見る」より作成（2023 年 11 月 13 日閲覧）.

表 1　図 4 をもとに作成したデータセット.

種類	特殊トークン	入力	出力
特化	⟨extra_id_2⟩	テレビで YouTube を見る	スマートテレビやゲーム機を利用する,Chromecast を使用する,Apple TV を使用する
汎化	⟨extra_id_1⟩	スマートテレビやゲーム機を利用する	テレビで YouTube を見る
汎化	⟨extra_id_1⟩	Chromecast を使用する	テレビで YouTube を見る
汎化	⟨extra_id_1⟩	Apple TV を使用する	テレビで YouTube を見る
平行移動	⟨extra_id_0⟩	スマートテレビやゲーム機を利用する	Chromecast を使用する,Apple TV を使用する
平行移動	⟨extra_id_0⟩	Apple TV を使用する	スマートテレビやゲーム機を利用する,Chromecast を使用する
平行移動	⟨extra_id_0⟩	Chromecast を使用する	スマートテレビやゲーム機を利用する,Apple TV を使用する

表 2　wikiHow データセットの内訳.

種類	訓練	検証	テスト
特化	5,424	298	363
汎化	7,764	397	517
平行移動	10,104	532	697
合計	23,292	1,227	1,577

表 3　自作データセットの内訳.

種類	テスト
特化	45
汎化	45
平行移動	45
合計	135

果に多様性を持たせるために，1 種類の記事タイトルについて 3 種類の章見出しを連結したデータセットを作成した．また，章見出しが 4 種類以上存在する場合，その記事からは入力として選んだ章見出しを除き，残りの章見出しの組み合わせで出力を複数種類作成する．例えば，n 種類の章見出しがある場合，${}_{n-1}C_3$ 個のデータセットが作成できる．なお，実際に構築したデータセットを表 1 に示し，最終的なデータセットの内訳を表 2 に示す.

d）　特化，汎化，平行移動の度合い

wikiHow データセット特化，汎化，平行移動の度合いについて図 5 を用いて説明する．図 5 は，上部で「健康を維持する」というタイトルの記事が「適度な運動をする」,「癌を予防する」を紹介する場合を表しており，下部では同じタイトルの記事が「ランニングを始める」,「なるべく徒歩で移動する」を紹介する場合について表している．wikiHow の記事では，図 5 上部のように目的から直接接続される方法が紹介されることもあれば，図 5 下部のように本来は直接接続されない方法が紹介されることもある．本節で述べる度合いとはこのような違いを指し，wikiHow では実際に「靴底の修理」の方法として「古くなった靴底を交換する」,「剥がれかけた靴底をくっつける」が紹介されており，これは図 5 上部の関係であると考えられる．また,「お酒の臭いを消す」の方法として「食べ物や飲み物を口にする」,「体を清潔にする」が紹介されており，これは図 5 下部の関係であると考えられる．このように，wikiHow には様々な度合いの記事が存在しているが，本研究では区別せずにデータセットの構築を行った.

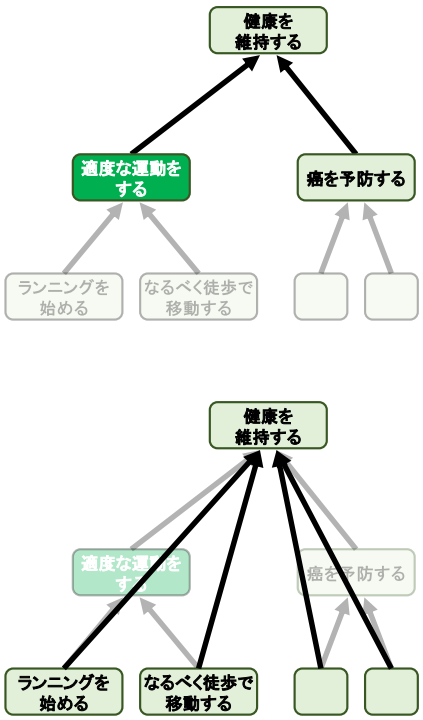


図 5　特化，汎化，平行移動の度合い

3.4.2　自作データセット

自作データセットでは，wikiHow における図 4 の目的と方法のペアを人手で考案する．このとき，作成する方法の数は 2 個以上，3 個以内とした．その後，wikiHow データセットと同様に特化，汎化，平行移動それぞれのデータセットを作成し，モデルの評価に用いる．なお，自作データセットでは目的と方法のペアを 45 件作成し，最終的なデータセットの内訳を表 3 に示す.

4　評価実験

4.1　比較手法

本研究の提案手法は，3.2 節で述べたマルチタスク学習が可

能な文生成モデル T5 による，手続き的知識源のマルチタスク学習である．そこで，T5 のマルチタスク学習をせず，特化，汎化，平行移動をそれぞれ 3 個のモデルに分けて学習する学習方法と比較する．この理由は,「T5 マルチタスク学習では，特化，汎化，平行移動すべてを学習するため，マルチタスク学習の方が方法と目的の関係性を深く学習することができる」と仮説を立てているからである．特に平行移動タスクでは，方法と共にその目的も学習したマルチタスク学習モデルの方が高い精度で生成が可能になると考えている．なお，この手法はマルチタスク学習に倣って**シングルタスク学習**と名前を付ける

4.2 評価手法

モデルの性能を評価するために，wikiHow データセット，自作データセットを用いた生成を行い，生成結果を人手で評価した．

特化，汎化，平行移動の生成タスクの評価を簡単に行うために，それぞれのタスクでテストデータ中の入力，目的，生成モデルより生成された出力を用いて「方法 1」と「方法 2」を定義した．実際の評価では,「「方法 1」を達成することで「方法 2」も達成されるか」という基準で評価を行った．この「方法 1」,「方法 2」は各生成タスクにおいてその定義が異なる．特化タスクでは，入力に対する達成方法を生成するため，出力が「方法 1」，入力が「方法 2」とすることで評価基準を満たす．例えば,「テレビで YouTube を見る」という入力に対して「スマートテレビやゲーム機を利用する,Chromecast を使用する,Apple TV を使用する」が出力された場合,「スマートテレビやゲーム機を利用する」,「Chromecast を使用する」,「Apple TV を使用する」が「方法 1」となり,「テレビで YouTube を見る」が「方法 2」となる．汎化タスクでは，入力に対する目的を生成するため，入力が「方法 1」，出力が「方法 2」とすることで評価基準を満たす．例えば,「Chromecast を使用する」という入力に対して「テレビで YouTube を見る」が出力された場合,「Chromecast を使用する」が「方法 1」となり,「テレビで YouTube を見る」が「方法 2」となる．平行移動タスクでは，入力を達成することで達成可能な目的に対して，達成可能な入力以外の方法を生成する．そのため，出力が「方法 1」，達成可能な目的が「方法 2」とすることで評価基準を満たす．例えば,「スマートテレビやゲーム機を利用する」という入力に対して「Chromecast を使用する」,「Apple TV を使用する」が出力された場合,「Chromecast を使用する」,「Apple TV を使用する」が「方法 1」となり,「スマートテレビやゲーム機を利用する」を達成することで達成される「テレビで YouTube を見る」が「方法 2」となる．この評価手法で評価する際に，不適合と評価された生成結果の不適合理由を調査した．この調査では不適合理由を予め決められた項目から選択する形式で調査し，それぞれ「「方法 2」を満たすことで「方法 1」も満たされる」,「「方法 2」が「方法 1」の言い換えである」,「「方法 1」が日本語として成り立たない」,「「方法 2」が日本語として成り立たない」,「1〜5 に該当せず「方法 1」を満たしても「方法 2」を満たすことができない」のいずれかを選択する．

表 4　ファインチューニングの学習条件.

項目	T5
バッチサイズ	128
学習率	1.0×10^{-5}
最適化関数	AdamWeightDecay [8]

4.3 評価尺度

人手で評価した後，平均正解数，平均正解種類数，平均逆順位を計算し，これをもとにモデルの性能を評価する．平均正解数では，1 入力あたり平均で何件の正解を生成できたかを計算する．平均正解種類数では，1 入力あたり平均で何種類の正解を生成できたかを計算する．平均逆順位では，入力に対する正解が平均で何番目に出現するかを計算する．平均正解種類数と平均逆順位の評価では，3.4 節で説明した生成結果の多様性についての評価である．なお，マルチタスク学習の汎化タスクおよびシングルタスク学習の汎化タスクでは，学習条件の都合で平均正解数と平均正解種類数については計算できないため，これらの項目の評価は行わない．本評価では，3 人の評価者により評価を行い，多数決で適合，不適合を評価した．

4.4 学習条件

シングルタスク学習では，各学習タスクについて表 2 に示すデータ量で学習を行った．また，マルチタスク学習では，それぞれのシングルタスク学習で用いたすべてのデータで学習を行った．なお，マルチタスク学習で用いたデータセットの内訳は，表 2 中の「合計」行に示す．

続いて，ファインチューニングの学習条件について説明する．T5 の事前学習済みモデルとして sonoisa/t5-base-japanese を用いた．学習の終了条件については予め 500 回と学習回数を決定し，既定の学習回数に到達するまで訓練した．その他の学習条件を表 4 に示す．なお，これらの条件はマルチタスク学習，シングルタスク学習の間で同じ条件である．

5 実験結果

本節では，提案手法および比較手法の評価結果について述べる．まず，評価者間の一致度について述べてから，学習方法の違いによる生成結果の違いについて述べる．続いて，4.3 節で述べた評価尺度に従って評価した結果について述べる．

5.1 評価者間の一致度

評価者間の評価の一致度の計算には Krippendoff の α [7] と Fleiss の κ [5] を用いた．その結果 $\alpha = 0.372$，$\kappa = 0.372$ となった．一般的に κ は $0.2 < \kappa \leqq 0.4$ で Fair agreement とされている．

5.2 生成結果

表 5 は各データセットを入力した際の適合と評価された生成結果について示している．表では，各入力に対する特化，汎化，平行移動について生成した結果のうち，適合と評価された生成結果の一部である．表 5 に示したように，どちらの学習方法で

表 5　テストデータセットによる生成結果の成功例．上段はマルチタスク学習で，下段がシングルタスク学習の生成結果．

データセット	種類	入力	生成結果
自作	特化	⟨extra_id_2⟩ パソコンを買う	ノートパソコンのスペックを確認する, ノートパソコンのスペックを確認する
wikiHow	汎化	⟨extra_id_1⟩ 食器用洗剤を使用する	服についたガソリンの臭いを取り除く
自作	平行移動	⟨extra_id_0⟩ 冷房を効率よく使う	涼しい恰好をする, 家を涼しく保つ, 暑さ対策を立てる
自作	特化	⟨extra_id_2⟩RPG を作る	rpg の基本ルールを作る,rpg の基礎ルールを作る
自作	汎化	⟨extra_id_1⟩ 着やせする	太目の男性が格好良く服を着る
wikiHow	平行移動	⟨extra_id_0⟩ 食事制限をする	減量に向かって食事を計画する, 減量に運動を取り入れる, 生活習慣を変える

表 6　テストデータセットによる生成結果の失敗例．上段はマルチタスク学習で，下段がシングルタスク学習の生成結果．

データセット	種類	入力	生成結果
wikiHow	特化	⟨extra_id_2⟩ リーバイスの偽物を見抜く	偽のウェアラブル端末を見分ける, 偽の...
wikiHow	汎化	⟨extra_id_1⟩ ブレンダーで米粉を作る	小麦粉でドリップ・ペインティング用の絵の具を作る
自作	平行移動	⟨extra_id_0⟩ 包丁の切れ味をよくする	コンロで包丁を切る, フライパンで表面を切る, フライパンで...
自作	特化	⟨extra_id_2⟩ データを完全に消去する	パソコンのデータを完全に消去する, パソコンのデータを...
自作	汎化	⟨extra_id_1⟩ 包丁を研ぐ	包丁で包丁を研ぐ
wikiHow	平行移動	⟨extra_id_0⟩iTunes を使用する	youtube から音楽をダウンロードする, 無料の音楽をストリーミングする,...

表 7　テストデータセットによる生成後，1 入力あたり何件の正解が生成されたかの平均．特化，平行移動では最大値が 3，汎化は 1 に近くなるほど精度の高いモデルとなる．

	特化		汎化		平行移動	
データセット	マルチ	シングル	マルチ	シングル	マルチ	シングル
wikiHow	**1.18**	1.11	0.125	**0.275**	**0.825**	0.688
自作	1.22	**1.33**	0.111	**0.133**	1.09	**1.11**

表 8　テストデータセットによる生成後，1 入力あたり何種類の正解を生成できたかを計算．汎化では多様性に関する学習を行っていないので汎化については評価しない．

	特化		平行移動	
データセット	マルチ	シングル	マルチ	シングル
wikiHow	**1**	0.827	**0.704**	0.580
自作	0.933	**0.978**	**1.09**	0.867

表 9　テストデータセットによる生成後，正解を何番目に生成できたかの平均逆順位を計算．

	特化		平行移動	
データセット	マルチ	シングル	マルチ	シングル
wikiHow	**0.477**	0.434	**0.426**	0.358
自作	**0.519**	0.515	**0.496**	0.489

あっても入力に対して意図した生成が可能にであることがわかる．また，生成結果多様化の部分についても，2 種類の学習方法それぞれで期待した出力ができることがわかった．

続いて，失敗例について述べる．表 6 には各データセットを入力した際に不適合と評価された生成結果について示している．不適合と評価された生成結果に共通する部分として，生成結果の文法がおかしい，生成結果が入力と全く同じ，などが共通していた．また，入力に固有名詞が含まれる場合，失敗しやすい傾向にあった．しかし，データセット中に頻出する固有名詞（例えば iPhone など）を入力に含む場合，成功することもあった．

5.3　各評価尺度による評価結果

続いて，各評価尺度による評価結果を示す．

5.3.1　平均正解数

表 7 に 1 入力あたりの平均正解数を示す．この評価では，特化，平行移動では最大値が 3，汎化は 1 に近くなるほど精度の高いモデルとなる．表 7 より，wikiHow データセットでは，マルチタスク学習モデルの方が良い結果になることもあったが，自作データセットではシングルタスク学習モデルの方が平均正解数が多くなる結果となった．

5.3.2　平均正解種類数

表 8 に 1 入力あたりの平均正解種類数を示す．この評価では，3.4 節で述べた生成結果の多様性について評価する．汎化では，多様性について学習していないのでこの評価は行わない．表 8 より，自作データセット特化ではシングルタスク学習モデルの平均種類数が多くなったが，その他ではマルチタスク学習モデルの平均種類数の方が多くなった．

5.3.3　平均逆順位

表 8 に平均で何番目に正解を生成できたかについて評価した．この評価でも，多様性について学習していない汎化の評価は行わない．表 8 より，全ての結果でマルチタスク学習モデルの値が高くなった．

5.4　エラー分析

4.2 節に従って評価した結果，不適合と評価された結果につ

いての不適合理由について調査した．この調査では 5 段階で不適合理由を調査しており，多数決により不適合理由を決定した．3 者それぞれが異なる理由を選択した場合，不適合理由は決定しない．

続いて，不適合理由の調査結果について説明する．不適合理由として，入力文がそのまま出力として生成される結果が両学習方法に共通して見られたが，特化シングルタスク学習では，マルチタスク学習を含む他の学習方法と比較して多くなった．この理由について，データ量の差に原因があると考えている．まず，特化シングルタスクと特化マルチタスク学習の差について考察する．表 2 に示したように，特化マルチタスク学習では汎化，平行移動を含めた 23,292 件のデータを用いて学習しているため，特化シングルタスク学習で用いたデータ量 5,424 件と比較すると圧倒的に多い．また，入出力関係が類似している特化シングルタスク学習と平行移動シングルタスク学習の比較においても，そのデータ量には約 2 倍の差がある．よって，特化データセットを何らかの方法で増加するか，平行移動データセットを減らすことで，特化シングルタスク学習と平行移動シングルタスク学習の結果は近づくと考えている．

6 まとめと今後の課題

本研究では，タスク検索を支援するための手法について述べた．これは，タスク検索においてユーザの入力に対する達成方法や，入力を達成することで達成される目的，共通する目的を達成可能な入力以外の方法を提示することである．本問題に取り組むために，マルチタスク学習が可能な文生成モデル T5 のファインチューニングにより，入力に対する特化，汎化，平行移動を提示するモデルを構築した．比較手法として，特化，汎化，平行移動をそれぞれ専用に学習させるシングルタスク学習と比較する．ファインチューニングに際して，データセットが必要になるが，本研究で必要とする大規模な手続き的知識源データセットは存在しない．そこで，あるタスクとそれを達成可能なタスクを集めたまとめサイト wikiHow に注目し，これを手続き的知識源とした．提案手法の評価には，生成結果を人手で評価した後，平均正解数，平均正解種類数，平均逆順位を用いた評価を行った．平均正解数では，シングルタスク学習がマルチタスク学習を上回る結果となった．平均正解種類数では，一部の結果を除きマルチタスク学習がシングルタスク学習を上回る結果となった．平均逆順位では，全ての結果でマルチタスク学習がシングルタスク学習を上回る結果となった．

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

文　献

[1] Hassan Ahmed and White Ryen, W. Task tours: helping users tackle complex search tasks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1885–1889, 2012.

[2] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 795–804, 2011.

[3] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. From “dango” to “japanese cakes”: Query reformulation models and patterns. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 183–190, 2009.

[4] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 875–883, 2008.

[5] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, p. 378, 1971.

[6] Michael P Georgeff and Amy L Lansky. Procedural knowledge. *Proceedings of the IEEE*, Vol. 74, No. 10, pp. 1383–1398, 1986.

[7] Klaus Krippendorff. Computing krippendorff’s alpha-reliability. 2011.

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[9] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, and Katsumi Tanaka. Mining alternative actions from community Q&A corpus for task-oriented web search. In *Proceeding of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 607–614, 2017.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67, 2020.

[11] Push Singh. The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, 2002.

[12] Takehiro Yamamoto, Tetsuya Sakai, Mayu Iwata, Chen Yu, Ji-Rong Wen, and Katsumi Tanaka. The wisdom of advertisers: mining subgoals via query clustering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 505–514, 2012.

[13] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pp. 418–428, 2020.

[14] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. Query understanding via intent description generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1823–1832, 2020.

[15] 加藤龍, 大島裕明, 山本岳洋, 加藤誠, 田中克己. タスクの汎化と特化に着目した web からのタスク検索. 第 6 回データ工学と情報マネジメントに関するフォーラム, C1–6, 2014.

[16] 工家昂之, 山本岳洋, 荘司慶行. 手続き的知識源を用いた方法を表すクエリからの目的抽出. 第 15 回データ工学と情報マネジメントに関するフォーラム, 3a–2–2, 2023.

[17] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. 関連単語抽出アルゴリズムを用いた web 検索クエリの生成. 電子情報通信学会論文誌 D, Vol. 92, No. 3, pp. 281–292, 2009.

[18] 田麥節治, 赤星祐平, 是津耕司, 木俵豊, 田中克己. Web サービスを対象とした検索意図推定とその応用. 第 4 回データ工学と情報マネジメントに関するフォーラム, F4–3, 2012.

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

情報探索支援

座長:倉島 健(NTT人間情報研究所)

コメンテータ:鬼塚 真(大阪大学)

2024年3月1日(金) 13:30 〜 15:40 T3-C (オンライン（Zoom Events))

15:15 〜 15:40

[T3-C-8-05]【技術報告】会社訪問アプリ「 Wantedly Visit」における新規ユーザーの行動量に基づいた推薦方策の選択

林悠大 (ウォンテッドリー株式会社)

©一般社団法人日本データベース学会

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

2024年3月1日(金) 13:30 ~ 15:40 | T4-A オンライン（Zoom Events)

認証

座長:金 京淑(産業技術総合研究所)
コメンテータ:横山 大作(明治大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T4-A-8-01] 全天球カメラを用いた人物認証と人物追跡

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

14:00 ~ 14:25
[T4-A-8-02] 2台の全天球カメラを用いた同一人物判定

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

14:25 ~ 14:50
[T4-A-8-03] 複数カメラ映像に跨ったエピソード記憶

*桑田 若菜[1]、山崎 智史[2]、劉 健全[2] (1. 兵庫県立大学、2. 日本電気株式会社)

14:50 ~ 15:15
[T4-A-8-04] 不均衡データを活用した認証手法の提案−タッチベース認証を対象として−

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

15:15 ~ 15:40
[T4-A-8-05] eCommTouch：ECサイトを対象としたタッチベース認証のための新たなデータセットの提案

*工藤 雅士[1]、高橋 翼[2]、山名 早人[1] (1. 早稲田大学、2. LINEヤフー株式会社)

全天球カメラを用いた人物認証と人物追跡

矢野　健斗†　　横山　昌平††,†††,††††

† 東京都立大学 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所　〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †yano-kento@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　全方位を一度に撮影することができる小型の全天球カメラが普及している．建設現場の施工管理やドライブレコーダーなどに利用されており，画像内の人物検出や人物追跡が課題となっている．しかし，全天球カメラで撮影された画像は正距円筒図法で投影されるため，画像内の右端と左端は実空間上では繋がっているが，画面上では離れて見える．そのため，既存の物体追跡アルゴリズムをそのまま適用すると，画面の右端と左端を跨いだ追跡ができない．本論文では，正距円筒図法の経度方向を拡張することにより，画面の左端と右端を跨いで移動する人物を追跡する手法を提案する．実験を行い，全天球カメラの周りを周回する人物を正確に追跡できることを示す．

キーワード　全天球カメラ，物体検出，物体追跡，姿勢推定

1　はじめに

近年，RICOH[1]をはじめ，Insta360[2]，GoPro[3]といった様々なメーカーから，360 度全方向を一度に撮影することができる小型の全天球カメラが発売されている．撮影不動産の物件紹介や建設現場の現場記録，Vitual Reality（VR）コンテンツの撮影など，幅広いシーンで活用されている．人々の外出が制限されたコロナ禍では，VR 動画を利用したバーチャルツアーが不動産業界や観光業界で流行した．

また，撮影された画像から周囲の状況を一度に把握できるため，ドライブレコーダーや防犯カメラといった防犯対策としての利用や，小売店における顧客の属性や動線の分析といったマーケティングツールとしての利用も進んでおり，動画内の人物検出や人物追跡技術が必要とされている．しかし，全天球カメラで撮影された動画に対応した人物追跡の手法には課題が多い．それは，全天球カメラで撮影された画像は正距円筒図法という投影法で全天球パノラマ画像に変換されて保存されることが原因である．正距円筒図法の性質上，3 次元の球面を 2 次元の平面に変換しているため，全天球パノラマ画像の右端と左端は実空間では繋がっているが，画像上では離れて見える．そのため，方位角 0 度を跨いで移動すると，画面の端に消えた人物が逆端から表れる．実際に方位角 0 度を跨いで移動する前後を図 1，図 2 に示し，その様子を上から見たイメージ図を図 3 に示す．

図 1，図 2 に映る人物の実空間での位置関係は図 3 のように隣り合っているが，全天球パノラマ画像で見ると離れて見えるため，既存の平面動画用の人物追跡アルゴリズムをそのまま適用すると正しく追跡が行えず，別の人間として認識されてしまう．



図 1　方位角 0 度に向かう人



図 2　方位角 0 度を跨ぎ，方位角 360 度からでてくる人

この課題を解決するために，全天球カメラで撮影された全天球パノラマ動画に対する人物追跡アルゴリズムの手法はいくつか存在する．八木らは，全天球パノラマ動画上の座標を全天球パノラマ画像上の位置座標 (x, y) を世界座標系 (x', y', z') に変換してから Deep SORT で人物追跡を行うことで，方位角 0 度を跨いだトラッキングを行った [5]．

Liliana らは，全天球パノラマ動画で検出した人物の足元の

1：https://theta360.com/ja/etc/technology.html
2：https://www.insta360.com/jp/
3：https://gopro.com/ja/jp/

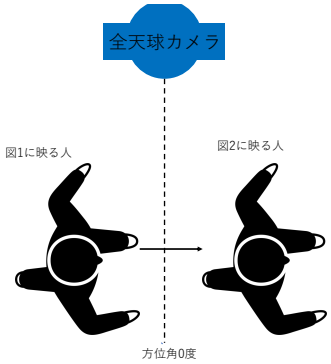


図 3　方位角 0 度を跨ぐ人を上から見た図

座標から，全天球カメラと検出された人物の距離を求め，カメラを中心とした地表座標上に人物をマッピングし追跡を行った [6].

これらの人物追跡アルゴリズムでは，全天球パノラマ画像に対して人物検出を行った後，座標を変換してから人物追跡を行う必要があるため，各フレームでの処理が多くなってしまうという課題がある，また，正距円筒図法の動画専用の人物追跡アルゴリズムの開発は普及していない.

そこで本論文では，既存の物体追跡アルゴリズムを応用することで，全天球パノラマ動画において，方位角 0 度を跨いだ人物を追跡する手法を提案する．本研究で提案する手法は，平面動画を対象とした既存の物体追跡アルゴリズムの Deep OC-SORT [1] を応用した．Deep OC-SORT とは，Observation-Centric SORT（OC-SORT）[2] と呼ばれる非線形的な動きとオクルージョンに頑強なモーションベースのトラッキング手法に対し，フレーム間の物体の関連付けに外観情報を取り込んだ，動画内の複数の物体をそれぞれ区別して追跡する MOT(Multi Object Tracking) アルゴリズムである．Deep OC-SORT をベースに，方位角 0 度を跨いだ人物に関しては，追跡中の人物との重なり具合をスコア化した Intersection over Union（IoU）を計算し，一番スコアが高い人物の ID を割り当てることで，360 度の人物追跡を可能にした.

また，本論文では指定のジェスチャーをした人物を認証し，方位角 0 度を跨いでも正しく認識できることを実験で示す．骨格検出を行うことで対象のジェスチャーを認識し，人物認証を行う．実験で使用する動画では，挙手をした人物を認証対象とし，対象人物が固定された全天球カメラの周りを周回し，方位角 0 度を跨いでも認証できていることを確認した.

本論文の構成は以下の通りである．2 章では，関連研究について述べる．3 章では，提案手法を述べる．4 章では，実験結果と考察について述べる．5 章では，まとめと今後の課題について述べる.

2　関 連 研 究

全天球カメラを用いた人物検出や人物追跡の研究はいくつか存在している．林田らは，物体検出アルゴリズムの You Only Look Once version3（YOLOv3）を用いて，全天球パノラマ画像の高緯度領域で歪んで映った人物を ”人 ” のクラスとして再学習させることで，リアルタイムの人物検出を可能にした [3].

井上らは，You Only Look Once version2（YOLOv2）と MOT アルゴリズムの Simple Online and Realtime Tracking（SORT）を用いた全天球カメラでの人物追跡手法を提案した [4]．撮影点を中心として水平方向が 8 枚のキューブマップに対してそれぞれ物体検出を行い，”person ” を抽出して統合，正距円筒図法の画像上で人物追跡を行った．物体検出と人物追跡を異なる投影法で行ったため，処理速度が遅いことが改善点として挙げられた.

八木らは，MOT アルゴリズムの Deep SORT を応用した全天球カメラでの人物追跡手法と，QR コードを利用した人物認証の手法を提案した [5]．全天球パノラマ画像上の位置座標 (x, y) を全天球カメラを中心とした球面座標系 (r, θ, ϕ) に変換し，さらに世界座標系 (x', y', z') に変換することで，全天球パノラマ画像内の平面座標が実空間に即した座標になり，方位角 0 度を跨いだトラッキングが可能になった．また，QR コードを全天球カメラに読み込ませ，QR コードとトラック間のコサイン類似度を求めることで，最も近いトラックと QR コードを認証させた．この手法を用いて，ユーザがスマートフォン上の QR コードを全天球カメラに読み込ませることで，スマートフォンとユーザを紐づけしながら追跡するシステムを提案した.

Liliana らは，You Only Look Once version5（YOLOv5）で検出した人物を SORT に類似した手法を用いて追跡する手法を提案した [6]．カメラの高さと検出された人物の足元の座標から，カメラから人物までの距離を測定することで，カメラを中心とした地表座標上での位置を特定．地表座標上で追跡を行った．また，カメラを中心としたヒートマップを作成し，歩行者が関心を示している領域を可視化した.

3　提 案 手 法

本章では，平面動画に用いる MOT アルゴリズムの Deep OC-SORT と姿勢推定アルゴリズムを応用することで，全天球パノラマ動画の方位角 0 度を跨いだ人物追跡と人物認証を実現する手法を提案する．また本実験では RICOH の THETA Z1[4] を使用する.



図 4　両端を 1.1 倍ずつ拡張したフレーム

4：https://www.ricoh360.com/ja/theta/

3.1 人物検出

本項では，人物検出の手法の説明を行う．はじめに，人物追跡を行うために各フレームに対して左右を 1.1 倍ずつ拡張してから物体検出を行う．図 4 に全天球パノラマ動画の両端を左右 1.1 倍ずつ拡張したフレームを示す．全天球パノラマ画像を左右に引き延ばすことで，方位角 0 度付近にいる人物の全体像を検出することができるようになる．本手法では，物体検出と姿勢推定を行うために，You Only Look Once version8（YOLOv8）[5]を用いる．

YOLOv8 とは，2023 年に Ultralytics によって開発され公開された 2024 年 1 月現在で最新バージョンの YOLO である．物体検出だけでなく，セグメンテーションや姿勢推定といった様々なタスクがこなせるモデルであり，今回は物体検出と姿勢推定を利用する．

物体検出，姿勢推定を行うために，COCO dataset を用いて事前学習済みのモデルの YOLOv8n-pose を使用する．本モデルを用いることで，検出した人物のバウンディングボックスの座標と 17 種類の関節の座標を取得することができる．

検出した人物のバウンディングボックスの左上の x 座標，y 座標をそれぞれ x_{min}，y_{min}，右下の x 座標，y 座標をそれぞれ x_{max}，y_{max}，確信度スコアを $conf$，クラスを $class$ とし，(1) 式の配列に保存する．

$$detection = \left[x_{min}, y_{min}, x_{max}, y_{max}, conf, class\right] \quad (1)$$

各フレームでは複数生成される人物情報 $detection$ を (2) 式で表される配列 $detections$ に保存する．

$$detections = \left[detection_1, detection_2, ..., detection_n\right] \quad (2)$$

本モデルは平面画像に対応した物体検出モデルであるため，正距円筒図法で投影され，高緯度領域が歪んでいる全天球パノラマ画像では正しく物体を検出できない可能性がある．そのため，歪みの小さい低緯度領域に人物が映るように全天球カメラを設置する．

3.2 人物追跡

本項では，人物追跡の手法の説明を行う．提案手法では，平面動画に対応した MOT アルゴリズムの Deep OC-SORT をベースに，方位角 0 度付近の人物に対しては IoU による関連付けを行うことで，方位角 0 度を跨いだ人物追跡を実現した．本手法で用いる Deep OC-SORT では，追跡中の人物の外観特徴量や ID，最終観測地点などの情報をまとめたトラッカーと，各フレームで検出された物体の外観情報や位置情報等を用いて，追跡中の物体と検出物体のマッチングを行う．同一物体とみなされればトラッカーを検出物体で更新し，同じ ID が割り当てられ，未追跡物体とみなされれば新たな ID が割り当てられる．

本手法では，各フレームで Deep OC-SORT を用いて，検出した人物のバウンディングボックスの座標から人物の外観特徴量等を取得し，現在追跡している人物と照らし合わせることで，マッチングを行う．

方位角 0 度を跨いだトラッキングには IoU による関連付けを用いる．IoU とは 2 つの領域の重なり具合をスコア化したものであり，(3) 式で求めることができる．

$$IoU = \frac{\text{Intersection}}{\text{Union}} = \frac{\text{共通面積}}{\text{合計面積}} \quad (3)$$

比較したい 2 矩形の座標を以下の $detA$，$detB$ とする．

$$detA = \left[x_{Amin}, y_{Amin}, x_{Amax}, y_{Amax}\right] \quad (4)$$

$$detB = \left[x_{Bmin}, y_{Bmin}, x_{Bmax}, y_{Bmax}\right] \quad (5)$$

(3) 式の Intersection について，共通面積の幅 w と高さ h は (6) 式，(7) 式のように表される．

$$w = max(0, min(x_{Amax}, x_{Bmax}) - max(x_{Amin}, x_{Bmin})) \quad (6)$$

$$h = max(0, min(y_{Amax}, y_{Bmax}) - max(y_{Amin}, y_{Bmin})) \quad (7)$$

$detA$，$detB$ との面積をそれぞれ S_A，S_B とすると Intersection と Union はそれぞれ (8) 式，(9) 式に表されるため，IoU は (10) 式で表される．

$$Intersection = w \times y \quad (8)$$

$$Union = (S_A + S_B) - w \times h \quad (9)$$

$$IoU = \frac{w \times h}{(S_A + S_B) - w \times h} \quad (10)$$

方位角 0 度を跨ぐとき，全天球パノラマ画像上では，画面の端に消えた人物が逆端から出現することになる．通常，端に消えた人物と逆端から出現した人物は，同一人物とはみなされず，異なる ID が割り当てられてしまう．しかし全天球パノラマ画像の特性上，図 4 から見てわかるように，拡張したフレーム内で検出された人物は同じ人物がフレーム内に映っている．そのため，拡張したフレーム内で検出された人物に関して座標を方位角 360 度分増減したうえで IoU を計算すれば同一人物を判断することができると考えた．

そこで，本論文で提案する手法の概要を図 5 に，各フレームの処理を図 6 に示す．拡張されたフレーム内で検出され，Deep OC-SORT の関連付けで追跡中の人物情報 (トラッカー) と不一致と見なされた人物に対して，検出された座標に応じて座標を方位角 360 度分増減し，追跡している人物の予測位置との IoU を計算する．そして，スコアが一番高い人物の ID を割り当てることで方位角 0 度を跨いだ人物追跡を実現した．本手法を用いることで，全天球カメラで撮影された全天球パノラマ動画に対して，通常の xy 平面動画の MOT アルゴリズムをそのまま適用しながらも，全方位の追跡が可能になる．

5：https://docs.ultralytics.com/ja/

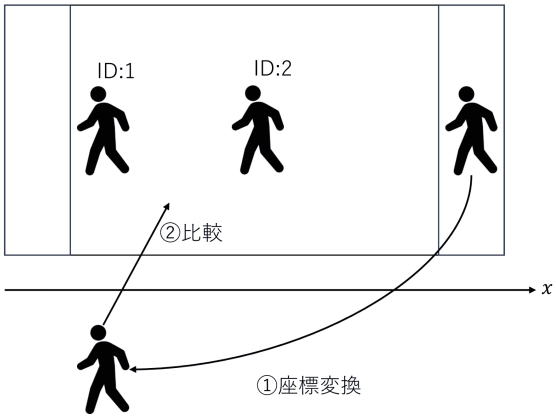


図 5　方位角 0 度を跨いだ関連付けの概要

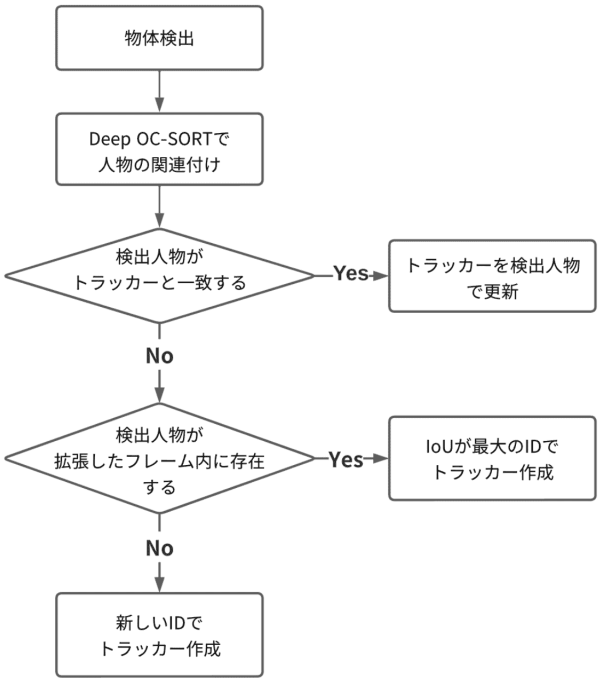


図 6　各フレームの流れ

3.3　人 物 認 証

本項では，人物認証の手法の説明を行う．本論文では，挙手を行った人物を認証し，認証された人物は緑色のバウンディングボックスで表示されるように変更する．挙手の判定は YOLOv8 を用いて姿勢推定を行い，右手首と鼻の座標から判定を行った．

各フレームで YOLOv8 で人物検出を行い，検出された人物ごとに挙手の動作の有無を示す挙手フラグを生成する．また，追跡中の人物の情報をまとめたトラッカーにも挙手フラグを追加する．

Deep OC-SORT を用いてトラッカーを *detections* で更新する際に，*detection* ごとに挙手フラグも更新し，一度でも True で更新されればトラッカーの挙手フラグも True になるように設定した．方位角 0 度を跨いだ移動の際には，引き継がれた ID の挙手フラグも同時に引き継ぐことで，画面の両端を跨いでも認証を行えるようにした．

4　実 験 結 果

本章では，人物追跡，認証の実験を行い，提案手法の評価を行う．実験は屋内で，背景情報が少なく，比較的広い部屋 1(約 $6\times10m^2$) と，背景情報が多く，比較的狭い部屋 2(約 $4\times4m^2$) の 2 種類の部屋の中で行った．実験を行った部屋の画像を図 7 と図 8 に示す．

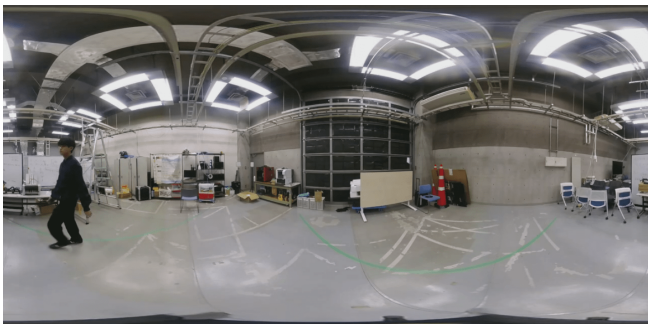


図 7　実験を行った部屋 1 の様子



図 8　実験を行った部屋 2 の様子

実験では，カメラから約 0.5 から 5m の範囲内を約 1 分間自由に歩き回る．挙手をした人物を認証し，画面上のバウンディングボックスの色を赤色から緑色に変化させることで認証状況を可視化する．

4.1　追跡人物が 1 人の場合

まず，追跡人物が 1 人の場合の評価を行う．実験結果を表 1 にまとめた．

対象人物が挙手をしたあと，カメラの周りを周回した．動画開始 3 秒後に方位角 0 度を跨いだ．方位角 0 度を跨ぐ前後のフレームを図 9，図 10 に示す．

結果は，方位角 0 度を跨いだ追跡は 2 つの実験を通して 100 % の精度で行うことができた．また，挙手を正しく認識し，画面上で緑色に表示することも成功した．途中で追跡が失敗してしまい ID が切り替わった（ロストしてしまった）タイミングは，方位角 90 度付近と 270 度付近に存在するカメラのつなぎ目の部分を跨いだ移動の時や，比較的薄暗い場所で移動しているときに多い傾向があった．



図 9　動画開始 2 秒後



図 10　動画開始 3 秒後

表 1　追跡人物が 1 人の時の実験結果

背景情報	フレーム数	ロスト回数	0 度を跨いだ成功率	誤認証回数
小	1655	2	7/7	0
大	1792	1	5/5	0

4.2　追跡人物が 2 人の場合

次に，追跡人物が 2 人の場合の評価を行う．実験結果を表 2 にまとめた．

また，部屋 1 で行った実験の様子を図 11 に，部屋 2 で行った実験の様子を図 12 示す．

結果は，方位角 0 度を跨いだ追跡は 2 つの実験を通して 100 % の精度で行うことができた．また，挙手を正しく認識し，画面上で緑色に表示することも成功した．途中で追跡が失敗してしまい ID が切り替わった（ロストしてしまった）タイミングは，1 人で実験を行ったときと同様に，カメラのつなぎ目の部分を跨いだ移動の時や，比較的薄暗い場所で移動しているときに多い傾向があった．また，動画内で 3 回すれ違いが発生したが，全て正しく追跡することができた．

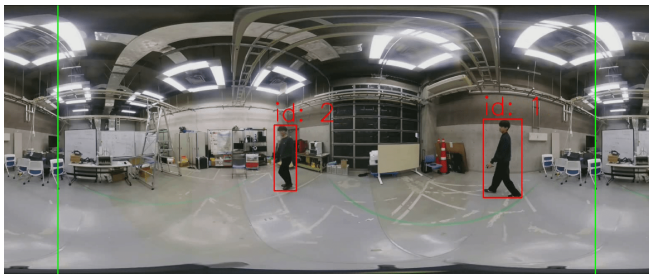


図 11　部屋 1 で行った実験の様子

4.3　追跡人物が 3 人の場合

最後に，追跡人物が 3 人の場合の評価を行う．実験結果を表 3 にまとめた．



図 12　部屋 2 で行った実験の様子

表 2　追跡人物が 2 人の時の実験結果

背景情報	フレーム数	ロスト回数	0 度を跨いだ成功率	誤認証回数
小	1383	2	7/7	0
大	1764	3	6/6	0

また，部屋 1 で行った実験の様子を図 13 に，部屋 2 で行った実験の様子を図 14 示す．

結果は，方位角 0 度を跨いだ追跡は 2 つの実験を通して約 92 %の精度で行うことができた．しかし，部屋 2 で行った実験では，挙手をしていない人物が，画面上で緑色に表示されてしまった．これは，背景情報が多いかつ，3 人の人間が重なり合ったタイミングで発生してしまったため，正しく骨格が認識されなかったことが原因だと考えられる．また，1，2 人で行った実験結果よりも，ロスト回数が 10 回程度多くなってしまった．これは，人が増えたことにより，すれ違いによるオクルージョンが多く発生してしまったことが原因だと考えられる．

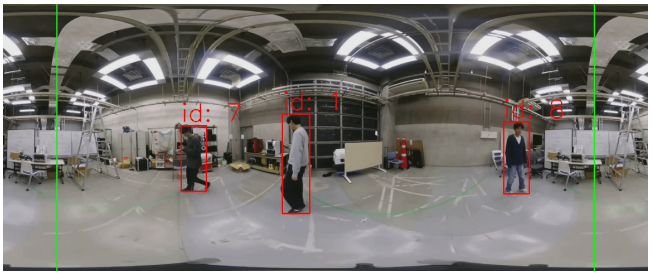


図 13　部屋 1 で行った実験の様子

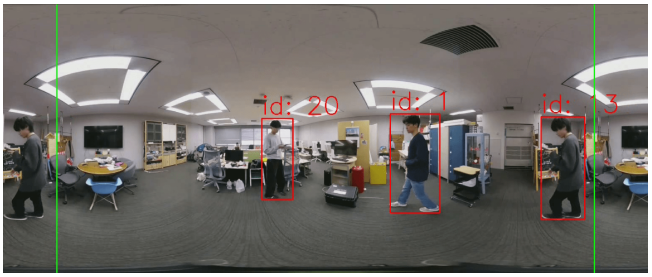


図 14　部屋 2 で行った実験の様子

表 3　追跡人物が 3 人の時の実験結果

背景情報	フレーム数	ロスト回数	0 度を跨いだ成功率	誤認証回数
小	1795	9	6/6	0
大	1445	15	5/7	2

4.4 評　　価

実験結果から，全天球パノラマ動画に対して提案手法を用いることで方位角 0 度を跨いだ追跡は，高精度で行えることが分かった．また，本実験で使用した Deep OC-SORT では，比較的明るい室内かつ，人数は 1，2 人程度ならば，高精度で 360 度の追跡が行えることが分かった．

5　まとめと今後の課題

本論文では，全天球カメラで撮影した全天球パノラマ動画に対する人物追跡手法と人物認証手法の提案を行った．提案手法では，左右に 1.1 倍ずつ拡張した全天球パノラマ動画に対して YOLOv8 を用いて人物検出，姿勢推定を行い，Deep OC-SORT をベースに，拡張したフレーム内に表れた人物は IoU によって追跡中の人物との関連付けをすることで，方位角 0 度を跨いだ人物追跡を実現した．また，挙手をした人物のトラッカーの挙手フラグを True にすることで追跡している人物の認証を実現した．

今後の展望としては，方位角 0 度を跨いだ人物の関連付けの改善が挙げられる．IoU を用いて関連付けを行うと，同時に 0 度を跨いだ場合に追跡の精度が下がってしまうため，骨格情報を用いたり，外観特徴量を用いたりして関連付けを行いたいと考えている．本研究では人物追跡の手法として Deep OC-SORT を用いたが，照明や背景色，人数などの様々な要因によってトラッキングの精度が落ちてしまい，全天球パノラマ動画に対しては高精度なトラッキングが行えるとはいいがたい結果だった．他の MOT アルゴリズムの利用や，カメラ台数を増やして 3 次元座標に変換をしたうえで追跡を行うなど，人物追跡手法の精度の改善を検討していきたい．また，本研究は録画した動画に対して人物追跡を実行した．研究背景や利用目的を考えると，リアルタイムで実行できるように実行速度の改善していきたい．

文　　献

[1] Gerard Maggiolino,Adnan Ahmad,Jinkun Cao,Kris Kitani, “Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification” arXiv:2302.11813

[2] Jinkun Cao,Jiangmiao Pang,Xinshuo Weng,Rawal Khirodkar,Kris Kitani, “Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking” arXiv:2203.14360

[3] 林田 和磨, 横山 昌平 “全天球カメラにより配信される正距円筒図法動画からのリアルタイム人物検出”, DEIM2020 C6-2(day2p19)

[4] 井上 慶彦, 岩村 雅一, 黄瀬 浩一 “全方位カメラを用いた物体検出とトラッキング—視覚障害者支援システムの実現に向けて—”, Technical Report 20, 大阪府立大学, may 2018

[5] 八木 優成, 横山 昌平 “全天球カメラを用いた単一カメラシェアリングシステムの開発”, DEIM2022 A33-1

[6] Liliana Lo Presti,Giuseppe Mazzola,Marco La Cascia, “Activity Monitoring Made Easier by Smart 360-degree Cameras” ECCV 2022: Computer Vision – ECCV 2022 Workshops pp 270–285, 2023.

2台の全天球カメラを用いた同一人物判定

織田　大幹†　　横山　昌平††,†††,††††

† 東京都立大学 システムデザイン学部　〒 191–0065 東京都日野市旭が丘 6-6
†† 東京都立大学大学院 システムデザイン研究科　〒 191–0065 東京都日野市旭が丘 6-6
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
†††† 東京大学生産技術研究所　〒 153–8535 東京都目黒区駒場 4-6-1
E-mail: †20140094@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　本論文では，2 台の全天球カメラで撮影した全天球パノラマ動画上に映る人物のカメラ間での同一人物判定手法を提案する．これまで，複数の監視カメラをあらかじめバスの設計をもとに決まった場所に死角が発生しないように配置し，それぞれのカメラから得た情報をカメラ間で共有し合うことで同一人物の判定を行う手法が提案されている．また，ロボット等に 2 台以上のカメラを積み，人や物を立体的に認知する技術を使用して，観察者が物体を認識する手法も提案されている．筆者の提案手法では，全天球カメラと呼ばれる 360° 全方向の映像をまとめて撮影可能なカメラを 2 台使用し，カメラを空間中のランダムな場所から動かしながら撮影を行い，撮った映像の中に映った人を検出，それらが同一の人物であることを判別する．提案手法のアイデアは，撮影された 360° 動画から正距円筒図法に投影された全天球パノラマ動画上における被写体の赤道上での位置情報とカメラの設置された高さの情報を利用することで，設置場所や積むマシンによらずカメラを動かした状態で撮影した映像から機械的な同一物の判定を行う．
キーワード　全天球カメラ，物体検出，映像解析

1　はじめに

現代社会においてカメラで得た動画像の解析の重要性は年々増しており，人物照合，位置推定など，その手法と実用方法もさまざまである．この動画像の解析をするうえで全天球カメラと呼ばれる死角の無いカメラを使用する手法も増えている．その例として，1 台の全天球カメラを机上に設置し，検出した物体の実平面上での位置推定を行うという手法が提案されている [1]．また，地面に設置した 2 台の全天球カメラを使用して，カメラの周りを周回する 2 輪車の運動解析をするといった，カメラの台数を増やすことでより広い視野での物体解析手法が提案されている [2].

一方でカメラを固定するのではなく，被写体である機械や人物そのものにカメラを設置し，カメラが動いている状態で，カメラに映った事象の解析や，それに伴う自己解析などの手法も近年増えている．その例として，2 台の車載カメラを使用して映った景色と映像マップをもとに自車位置推定を行うという手法が提案されている [3]．しかし，2 つのカメラは動いているが，同じ車に設置され，撮影範囲を広げるためであるため，1 台のカメラを動かしながら映像解析をしているという状況に近い．

また，人物の判定の技術として，対象物体にマーカーを貼り付け，そのマーカーを検出することで，同じマーカーの持ち主を判別する手法が提案されてきた [4]．しかし，この手法では，人物に対してマーカーを付与しなければならないという点で，一般人にマーカーを普段から貼り付けることはできないため，監視カメラ的利用での同一人物判定は難しいえる．

そこで，本研究では 2 台の全天球カメラを使用し，それぞれのカメラを撮影者の頭上に設置し，撮影者は空間上で自由に動き回る．このとき，設置位置や場所によらない人物照合を行うことを目的にして，全天球カメラが動いている状態であっても人間が物理的に確認する必要のない同一人物の判定手法を提案する．本研究では，検出対象を人間に限定し，対象物体にマーカ等を貼り付けずに行う．全天球カメラは広範囲な撮影が可能で，映したい人間にカメラを向けるという手順を踏まないために今回の実験に利用することを決定した．通常の単眼カメラで実験を行った場合，前述したように，カメラに人物を写すためにカメラの向きを変える必要があり，映すことが可能な範囲を比較しても，全天球カメラのほうがこの実験に適していると考えた．

全天球カメラで撮影された動画像は 2 つの魚眼レンズから撮影された 2 枚の動画像から構成される．この動画像を正距円筒図法という投影法で平面に投影することで一般的なカメラで撮影された動画像と同じ形式で閲覧することができる全天球パノラマ動画像として保存することが可能になる．全天球パノラマ画像では，一般的な動画像とは違い，歪みが発生しており，1 枚の写真で全方位を閲覧することが可能である．本実験ではこの形式に変換された全天球パノラマ動画像上での検出人物の情報を利用して，異なるカメラ間での同一人物の判定を行う．

本論文の構成は以下の通りである．2 章では本研究を行う上で使用した手法とそれらを活用した関連研究について述べる．3 章では，本研究の提案手法について述べる．4 章では実験内容とその結果，及び考察について述べる．5 章では本研究のまとめと今後の課題について述べる．

2 関連研究

本章では本研究の関連研究について述べる．まず，全天球カメラで撮影した物体の位置推定手法として増田ら [1] は，1 台の全天球カメラで撮影した写真の情報から学習を行ったマーカーの位置推定を行った．マーカーが全天球パノラマ画像における歪みの中に入ってしまった場合でも正常に検出が行われるように，様々な位置でのマーカーの写真を学習させたモデルを使用した．

また，笠原ら [6] は，全天球カメラで撮影した写真に映った人物の向きに着目することによって，人物の状態を検知することを可能にした．向きを検知するために，検出した人物の骨格検知を行い，検出した左肩と右腕の関節の位置とカメラと物体との推定距離を使用した．

同一人物の判定手法として，西浦ら [7] は，バスの入り口と出口に設置された単眼カメラでそれぞれ得た映像に映った人物の特徴量を抽出し保存することで，乗車区間別の乗降者数の解析に使用する手法を考案した．結果として，実際の乗降者数と推定数との誤差が 1 人ずつ，という高精度な推定を行うことができた．

また，本研究と同様に 2 台の全天球カメラを使用して物体の検出を行っている研究がいくつかあったので紹介する．

平澤ら [2] は，2 台の全天球カメラを使用し，カメラの周囲を走る 2 輪車の運動計測を行った．しかし，結果としては正確に検出できたものが運転手のヘルメットのみであり，その中心点を人の視認で動画のフレーム毎に収める必要がある．そのため，適切なマーカー等を使用しなければ正確な実験が行えないと考えた．

池井ら [5] は，2 台の全天球カメラを用いた立体視システムを考案した．移動ロボットに観察する人間の両眼に見立てるように 2 つの全天球カメラを搭載し，そこから得た映像を観察者の両眼として対応させることで，遠隔体験を可能にさせた．また，視覚だけでなくロボットの運動に対応するように観察者に身体的な刺激を同時に与えることにも成功している．

最後に，本研究で使用した物体検出の際に使用される代表的なアルゴリズムである YOLOv8 を使用した基礎的な物体検出に関する研究を紹介する．黒岩ら [8] は，単眼カメラから得た動画内の人物追従システムを考案した．物体検出の手法として YOLOv8 を使用し，物体追跡を行うモデルとして SORT を使用した．しかし，カメラが動くことによって，追跡の精度が低下してしまうという問題が発生した．

3 提案手法

提案手法の概要を示す．まず，図 1 のような形で全天球カメラを頭上に設置した 2 人の人間に，進む方向を固定したまま動いてもらい，その周りをほかの人物が周回する映像を集める．その後，編集した全天球パノラマ動画から YOLOv8 を用いて人物検出を行う．人物検出で得られた Bunding Box（以下,BBox と略す）の全天球パノラマ動画上の座標を取得し，もう一方のカメラを持つ人物の判定を行い，そのほかの人物の実空間上での位置をそれぞれのカメラ映像をもとに算出し，同一人物だと判定したのち，各人物に ID をふった BBox をつけた映像を出力するという流れで実験を行った．



図 1　カメラを設置した撮影者の様子

3.1 座標処理

まず，全天球カメラで撮影した全天球パノラマ動画に物体検出を行い，その結果からもう一方のカメラの持ち主の判別ととそのほかの検出人物の実空間上での推定位置を算出する．本節では，実空間での座標と全天球パノラマ動画像内での検出座標との関係性について述べる．

3.1.1 実空間上での検出人物

まず，一方のカメラから見たもう一方のカメラの持ち主検出の手法について説明する．全天球カメラでは 360° 全方位の映像が撮影できるため，今回の実験での 2 人の撮影者，すなわち 2 つのカメラは同じ向きを向きながら動いて撮影を行うものとする．

全天球カメラと検出人物との位置関係を上から簡易的に図にしたものが図 2 である．ここでは，全天球カメラの正面をどちらの面としているかがわかりやすいように，正面としている片方の魚眼レンズのみを図に示した．図 2 のように撮影者の立つ平面を XY 平面と定義している．ここでは，全天球カメラから地面へとおろした垂線と地面との交点を XY 平面における各全天球カメラの原点として位置推定を行った．全天球カメラの正面レンズ向いている方向を Y 軸の正の向きとし，レンズの右側を X 軸の正の向きとした．次に，全天球カメラから検出された人物の XY 平面上での極座標について説明する．ここでは Y 軸負の方向に伸ばした直線を極座標の始点として原点から人間までの距離を r [cm]，始線から人物までのなす角を θ とした．

全天球カメラと検出人物との位置関係を横から簡易的に図にしたものが図 3 である．検出人物の足先，ここでは BBox の四角形の最も下の辺の中心を足先としている．カメラレンズの中

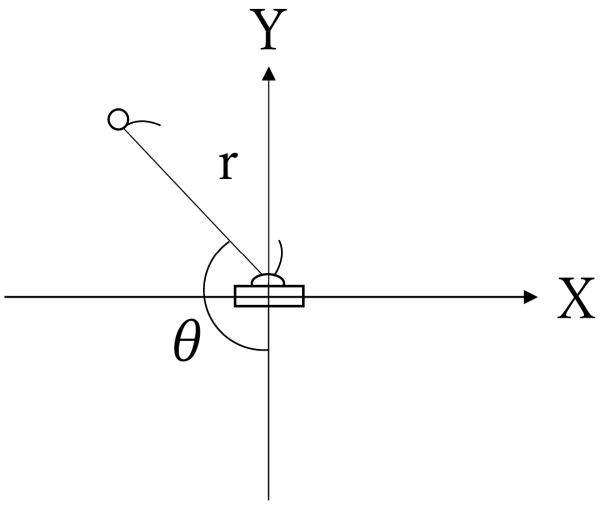


図 2　カメラと検出人物との位置関係を上からみた場合の簡易図

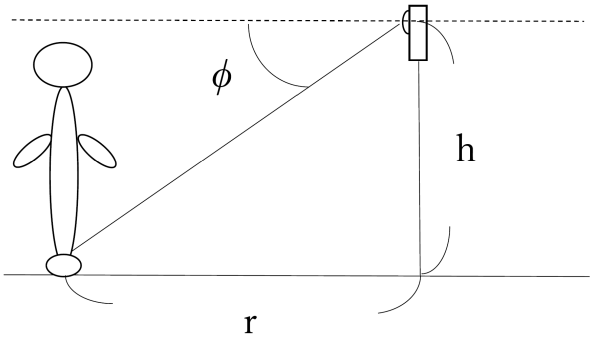


図 3　カメラと検出人物の位置関係を横から見た場合の簡易図

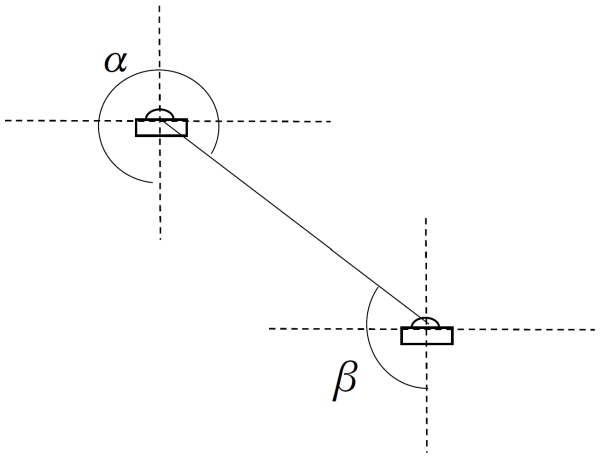


図 4　2 つのカメラ間の向きの関係性を示した図

心から，地面までの距離を h [cm]，カメラレンズの中心から地面から検出人物の足先を結ぶ直線と，カメラレンズの中心から伸ばした地面と平行な直線とのなす角を ϕ とする．

3.1.2　もう一方の撮影者との関係性

カメラ同士の位置関係を上から簡易的に図にしたものが図 4 である．図 4 のように，カメラ 1 から検出されたカメラ 2 をもつ人物に対して極座標上でのなす角を α，同様に，カメラ 2 からのなす角を β とすると，(1) 式が成り立つことがわかる．

$$|\alpha - \beta| = \pi \tag{1}$$

しかし，ほかの検出人物に対しては，それぞれのカメラとの極座標のなす角に関して常に成り立つものはない．この関係性を利用して，もう一方の全天球カメラをもつ検出人物がどれであるかを判別する．

3.2　全天球パノラマ動画像からの位置推定

位置推定では，YOLOv8 に搭載されている検出物体の BBox の画像中での位置座標を出力する機能を使用した．撮影した動画を Insta360 Studio 2023 というアプリケーションを用いて全天球パノラマ動画へと変換し，毎フレームに対してこの位置座標を出力した．30fps の動画を使用したため，1 秒の動画につき 30 個の位置座標が書かれたファイルが出力される，各ファイルには，その画像中で検出された人物の数だけの行数ぶんの BBox 情報が示されている．また，アプリケーション上で向きを固定する機能を使用し，カメラの回転に対する全天球パノラマ動画の表示方向を固定した．

図 5 に，簡易的な全天球パノラマ動画上での 1 シーンを切り取った場合の簡易的な図を示す．図 5 のように全天球パノラマ画像上では左上を原点として座標軸を X ´ 軸，Y ´ 軸とする．またそれぞれの座標の最大値は 1 である．YOLOv8 であらかじめ学習されているモデルを用いて人間のみが検出されるように指定し，人物の BBox 検出結果としては，BBox の中心の x ´，y ´ 座標と BBox の幅 w，高さ l が出力される．

次に，BBox の座標から検出人物の実平面上の XY 平面における推定座標を算出する．図 2，図 3，図 5 で示したように，極座標における原点から人間までの距離を r [cm]，始線から人物までのなす角を θ，カメラレンズの中心から，地面までの距離を h [cm]，カメラレンズの中心から検出人物の足先を結ぶ直線と，カメラレンズの中心から伸ばした地面と平行な直線とのなす角を ϕ，BBox の中心の座標 x ´，y ´，幅 w，高さ l，とする．

検出人物の足先の座標 (a, b) は BBox の情報より，a は BBox の x 座標と等しく，b は次式で求められる

$$b = y\,' + \frac{l}{2} \tag{2}$$

式 (2) で求めた a，b より，なす角 θ，ϕ を求める．

$$\theta = 2 \times a \times \pi \tag{3}$$

$$\phi = (\frac{1}{2} - b) \times \pi \tag{4}$$

(4) 式の結果と高さ h を用いて距離 r を求める．

$$r = |\frac{h}{tan\phi}| \tag{5}$$

(3)，(5) 式より推定座標 (x, y) が次式のように求められる．

$$x = -r \times \sin\theta \tag{6}$$

$$y = -r \times \cos\theta \tag{7}$$

これらの式を用いてテキストファイルとして出力された BBox

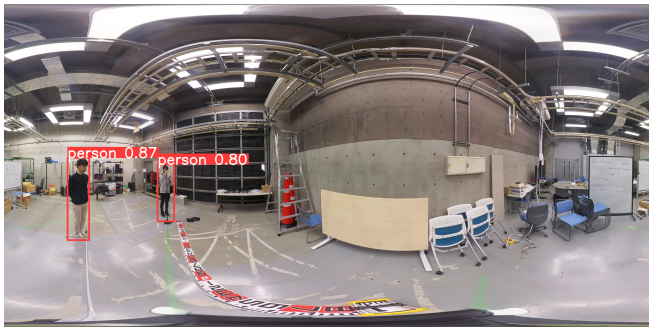


図 5　全天球パノラマ画像において BBox が出力された画像

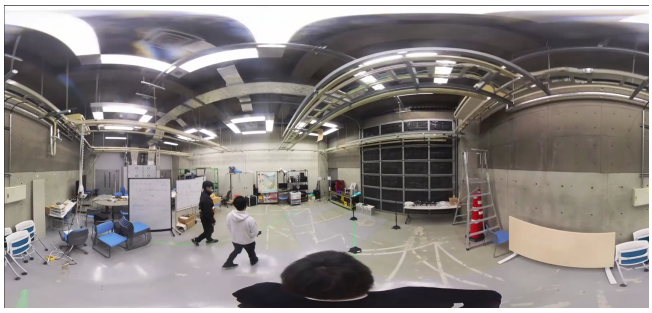


図 6　同時刻に切り取られた 2 つのカメラからの画像 1 枚目

情報をそれぞれの人物の座標情報として (極座標における始線からなす角，推定 x 座標，推定 y 座標, BBox の中心 x 座標, BBox の中心 y 座標, BBox の幅, BBox の高さ) = (θ_1, x'_1, y'_1, w_1, l_1) として変換し，json ファイル形式で保存する．

3.3　検出人物の ID 処理

同一人物であると異なるカメラから検出する方法として，3.2 節で求められた各人物の座標情報をもとにまず毎フレームごとにもう一方のカメラの持ち主を判別する．図 4 のように，ここでは，カメラ 1 の極座標における始線からなす角を α，カメラ 2 におけるなす角を β とすると (1) 式が動いていても常に成り立つことがわかる (1) 式が成り立つ人物の座標を各フレームで取得し，3.2 節で作成した json ファイルにおける座標情報に”camera”という撮影者を認識するためのラベル名を新たな情報として追加する．

もう一方の撮影者の座標情報より，各カメラ映像における検出人物の XY 平面上での位置を照らし合わせ，それぞれのカメラ間での x 座標誤差，y 座標誤差を考慮し，設定したパラメータ以内の範囲にいる人物を同一人物とみなし 3.2 節で作成した json ファイルに保存されている座標情報に人物の ID 番号を新たな情報として追加し，再保存する．ここで設定したパラメータの説明に関しては 4.1 節の実験部分で行う．

最後に，編集された json ファイルを読み取り，保存されている座標情報をもとに，撮影した動画から生成された全天球パノラマ動画上に OpenCV を用いて ID が振られた新たな BBox を描画し，新たな全天球パノラマ動画を出力する．図 6，図 7 にに出力した 2 つの動画で同じ時刻で切り取った画像を示す．

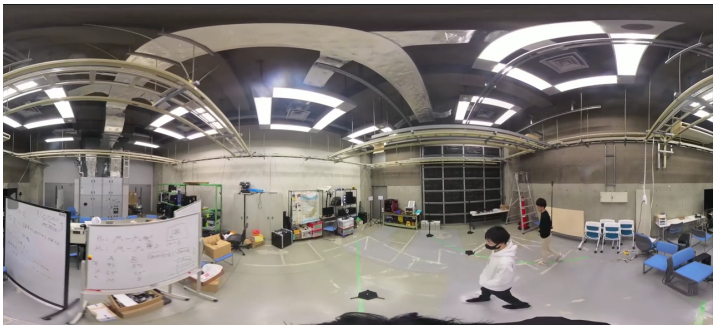


図 7　同時刻に切り取られた 2 つのカメラからの画像 2 枚目

4　実　　験

本実験では Insta360 社の Insta360 X3 を 2 台使用し，東京都立大学日野キャンパス内の実験室と屋外で実験を行った．実験は，両方のカメラを動かさずに人間の頭上に設置して，撮影を行い推定座標の精度を確認するための実験と，両カメラを人間の頭上に設置して動かしながら撮影を行い，同じ人間に対して各カメラの映像に同じ ID 番号を振る実験の 2 種類を行った．

4.1 節では実験 1 として，パノラマ画像から得た推定座標と実座標との誤差を計測する実験，4.2 節には実験 2 として，撮影者同士の相対的位置の関係性と検出人物の推定座標を利用した同一人物判定を行う実験について述べる．

それぞれの実験において全天球カメラの向きは手動で決めており，全天球カメラについている 2 つの魚眼レンズのうち操作画面がない面のレンズ中心を正面としている．

4.1　実平面での推定座標

今回の実験ではカメラから得た画像から導いた推定座標と実座標との誤差を把握し，どの程度の範囲に入っている人物を同一人物とみなすかのパラメーターを設定するため行う．2 つの全天球カメラは実験 2 での環境と近づけるため，カメラは三脚などで固定を行わず，人間の頭上に設置し，人間がその場に直立する形をとる．図 8 に実験の簡易的なカメラの配置とその向き，そして検出する人物を配置する位置を示す．撮影をしていない検出人物は 1 人であり，図 8 のように 8 つの地点に配置する．各地点でカメラから写真を撮り，計 8 枚の画像を用意する．これらの画像で推定位置を検出し，カメラから見た検出物体の推定座標と実座標の，X 座標，Y 座標のそれぞれにおける

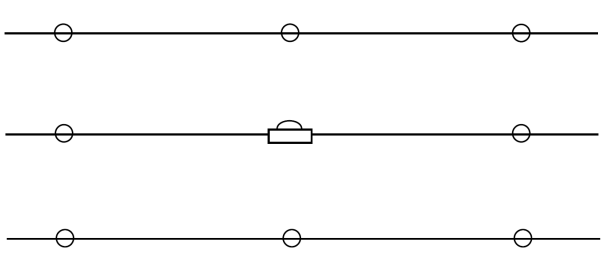


図 8　カメラの位置と検出人物が立つ 8 つの位置

単位：cm

	推定x座標	推定y座表	想定したx座標との差	想定したy座標との差
Point1	219.8331955	-201.9754067	-19.83319546	-1.975406686
Point2	224.3414328	4.797527765	-24.34143281	-4.797527765
Point3	213.5147626	210.2979253	-13.51476263	-10.29792533
Point4	0.534382373	189.842347	-0.534382373	10.157653
Point5	-8.95301636	-203.9005083	-8.95301636	-3.900508278
Point6	-211.3480403	182.8962568	-11.34804026	17.10374318
Point7	-205.693421	-12.37810224	-5.693421001	-12.37810224
Point8	-185.4002748	-199.3427518	14.59972519	0.657248167

図 9　推定座標と実座標との誤差

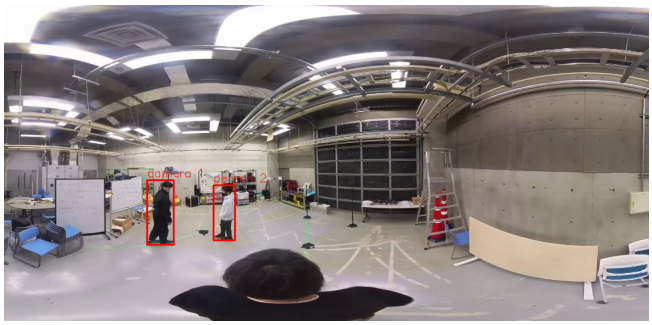


図 10　異なるカメラ間で同時刻に切り取った同じ ID が振られた画像 1 枚目

誤差がどの程度か実験する．図 9 に得られた推定座標と実座標との差をまとめた表を図示する．図 9 でわかる通り，X 座標の誤差が最大でおよそ 24cm，Y 座標の差が最大でおよそ 17cm となっている，本実験ではそれぞれのカメラから検出された人物にこの数値以上の差がない場合に同じ人物であると判別するようなパラメータを設定した．

4.2　動いているカメラ間での同一人物判定

本実験では 2 つのカメラ間での撮影時刻をそろえるために両カメラを撮影状態にした上で，1 人の人間がパンッと手を叩いて大きな音を収録することで，撮影終了後に映像中の音の波形の大きな位置を利用して撮影動画を切り抜くことで，撮影開始時刻と終了時刻がそろえることで時間軸をそろえた 2 つの動画を書き出すようにしている．

また，(1) 式で求めたなす角の差を毎秒出力した結果，もう一歩のカメラとの関係性が最低でも $|\alpha - \beta| = 0.096$ 誤差以内であるため，これをもう一方の撮影者であると判別する上でのパラメータに設定した．

図 10，図 11 にそれぞれのカメラで撮影した映像において検出人物の BBox に同じ ID を振った出力結果を示す．図 10，図 11 のようにもう一方の撮影者に”camera”のラベルを付け，もう一方の検出人物に対しては同様の”person”のラベルと ID が振れていることがわかる．また，ここで peron の ID が 2 から始まっているのは初めに検出人物全員に person の番号をいったん振ってからもう一方のカメラの持ち主を判別しているためである．

5　まとめと今後の課題

本研究では動く 2 台の全天球カメラ間での同一人物判定手法

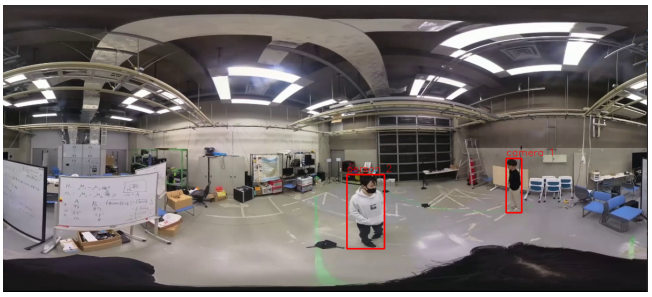


図 11　異なるカメラ間で同時刻に切り取った同じ ID が振られた画像 2 枚目

を提案した．カメラを設置する場所によらない同一人物判定において，部屋情報や積むマシンによらず，動き回る人間に設置した 2 つのカメラ間で，映った人物の機械的同一人物判定が可能となった．全天球カメラで撮影された動画から変換された全天球パノラマ動画とカメラの向きとの関係性，そしてもう一方の撮影者との位置関係を利用して，検出人物の XY 実平面上での実座標を導き，判定を行った．

今回の実験では最も被写体としての出現頻度が高いという点で，被写体を人間のみに設定したが，学習済みのモデルではなく新たに検出したいものを学習させたモデルを使用すれば，被写体が動いているかどうかに関わらず今回の手法が応用できると考える．人間以外の物体，マシンなどが撮影した場合でも同様の手法が応用できる．

今後の課題として，本実験ではカメラの向きを固定するという制限をかけなければいけなかったが，カメラから方角の情報を抽出することができれば，2 つのカメラ間での向きの情報が得られるため，より撮影の自由度が増すと考えられる．

また，本実験ではあらかじめ撮影したものを処理するという実験方法であるが，最終的にこれらの処理をリアルタイムで行うことを可能にしたいと考えている．そのためにカメラとサーバー間での通信や，異なるカメラ間での撮影人の差異の問題などを中心に解決に取り組んもうと考えている．

最後に，本実験よりも被写体がより遠い距離での同一人物判定を行えるようにするため，使用するモデルの改善や，撮影者である人間のブレの問題を解決する必要がある．

文　　献

[1] 増田純也，横山昌平”全天球カメラを用いた机上物体に対する位置推定”DEIM2020 C6-3(day2 p20)
[2] 平澤順治”二輪車の位置計測のための全天球カメラの活用”，自動車技術会論文集，2023
[3] 内山寛之，出口大輔，高橋友和，井手一郎，村瀬洋”市街地映像マップと 2 台の車載カメラを用いた自車位置推定”,「画像の認識・理解シンポジウム (MIRU2009)」，2009
[4] 井上晴可，梅原喜政，今井龍一，神谷大介，田中成典，中畑光貴，島野寛己”オブジェクト追跡技術を用いた人物識別の補正に関する研究”，土木学会論文集 F3（土木情報学），Vol 78，No 2，I-122-I-130，2022
[5] 八木龍之介，藤江柊生，雨宮智浩，北崎充晃，ヤェム ヴィボル，池井 寧”遠隔体験の臨場感向上手法に関する研究 ”，第 24 回日本バーチャルリアリティ学会大会論文集，2019
[6] 笠原悠樹，横山昌平”全天球カメラを用いた三角測量による人物

ポジショニング推定”，DEIM Forum 2023 5C-4
[7] 西浦 翼，横山 想一郎，山下 倫央，川村 秀憲，佐藤 好美，長谷川 怜，平澤 幸”人物追跡手法を用いた路線バスにおける乗客 OD 推定手法の開発”，人工知能学会第二種研究会資料社会における AI 研究会 SIG-SAI-044-10，2019
[8] 黒岩大雅，倪侃，姜雄文，川口貴弘，橋本誠司，張浩浩，磯貝孝夫，曽我裕史”単眼カメラを用いた人物追従アルゴリズムの検討”，自動制御連合講演会講演論文集 第 66 回自動制御連合講演会, 2023

複数カメラ映像に跨ったエピソード記憶

桑田　若菜† 　山崎　智史†† 　劉　健全††

† 兵庫県立大学大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† NEC ビジュアルインテリジェンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: †ad23g022@guh.u-hyogo.ac.jp, ††{s-yamazaki31,jqliu}@nec.com

あらまし　本論文では，複数カメラに跨る人物行動映像を用いた，複数カメラ間エピソード記憶タスクの提案を行う．従来のエピソード記憶タスクは，単一カメラ映像からユーザ興味の行動が発生した時刻の特定を行う．提案タスクは従来の単一カメラ映像を対象としたタスクから，複数カメラ映像を対象としたタスクに拡張することで，カメラ間に跨った行動の認識に基づく人物特定と行動時刻の特定を要求するタスクである．このタスクでは，同期間の複数カメラ映像と，ある特定の人物に対する複数カメラに跨った 2 つの行動が含まれるクエリが与えられたときに，1 つ目の行動から人物を特定し，2 つ目の行動から行動時刻の特定を要求する．本論文では，提案タスクとその正解アノテーションの詳細について述べ，事前学習済みモデルである Contrastive Language-Image Pre-training (CLIP) を用いた実験を行う．

キーワード　映像シーン検索，人物行動，エピソード記憶

1 はじめに

私たちの日常生活は様々なイベントの連続であり，防犯カメラ等の映像として蓄積されている．しかしながら膨大な蓄積映像の中からユーザーの興味に合う映像シーンを人手で抽出することは手間がかかる．例えば犯罪の事後捜査において，防犯カメラなどの映像を目視確認し，犯行時刻を割り出す作業には時間とコストを消費する．もし抽出したいイベントをテキストで与えるだけで，該当する映像シーンを自動抽出できるのであれば，作業時間とコストの削減につながる．

Ego4D [4] ではエピソード記憶タスクとして，目的の映像シーンの抽出に取り組まれている．Ego4D ではアクションカメラなどで撮影された一人称視点の日常生活映像データセットを提供し，そのデータセットを用いた 5 つのタスクを提案している．この 5 つのタスクの内 1 つがエピソード記憶タスクである．

エピソード記憶タスクとは，1 つの長時間映像とその映像終了時点から過去を質問するクエリが入力されたときに，そのクエリの回答に関する映像の開始時刻と終了時刻を出力するタスクである．例えば,「What did I put in the drawer?」というクエリが与えられたときは，カメラ着用者が引き出しに鍵を置いた映像シーンの開始時刻と終了時刻を出力する．

Ego4D のエピソード記憶タスクは興味のイベントを含む映像シーンの抽出タスクを実現しているが，抽出対象が単体のカメラで捉えた狭域の映像に限られている．より広域でのイベントを捉えるのであれば，複数カメラ映像に跨った映像シーン抽出が求められる．しかしながら従来のエピソード記憶タスクを行う技術 [2, 5, 6, 8, 9, 14] は複数カメラを対象に適用できない．

そこで本論文では複数カメラから撮影した日常生活の長時間映像を提供しているデータセット Narative Dataset [13] に基づいた新規エピソード記憶タスクを提案する．Narrative Dataset は 3 視点分の撮影映像を 1 つの映像セットとして出現人物と物体，それらの関係性をアノテーションしている．各映像セットの中では，各出現人物が複数カメラに跨って様々な行動を行っている．本論文で提案する複数カメラ映像間エピソード記憶タスクは，ある特定の人物に対する複数カメラに跨った 2 つの行動記述を含むクエリ文が与えられたときに，1 つ目の行動記述から人物を特定し，2 つ目の行動記述から行動時刻の特定を要求する．

2 関連研究

2.1 Natural Language Video Localization

Natural Language Video Localization (NLVL) とは，トリミングされていない映像に対して，入力された自然言語クエリからクエリに対応する映像の開始時刻から終了時刻を出力するタスクであり，多くの研究で取り組まれている [1, 3, 16, 17]. このタスクに初めて取り組んだのは Gao ら [3] と Hendricks ら [1] の研究である．これら 2 つの研究では映像内の出来事をクリップレベルで直接的に説明する文章と，それに対応する開始時刻と終了時刻がアノテーションされたデータセットを用いている．

本論文で提案する複数カメラ映像に跨ったエピソード記憶タスクも NLVL の一種であり，入力されたクエリに該当する映像シーンの開始時刻と終了時刻を出力する．

2.2 エピソード記憶

エピソード記憶タスクは Ego4D [4] で初めて提案されたタスクである．エピソード記憶には様々な入力クエリタイプがあり，Ego4D では Natural language queries (NLQ)，Visual queries (VQ)，Moments queries (MQ) の 3 つのクエリタイプを提案している．さらに Yang ら [15] は Activity-Query,

Object-Query，Time-Query という新たなクエリタイプを提案している．これらのクエリタイプのうち，NLQ タスクはこれまで多くの研究で取り組まれている [2, 5, 6, 8, 9, 14]．エピソード記憶 NLQ タスクは NLVL タスクの一種であると言える．2.1 節で挙げた他の研究との違いは，入力される自然言語クエリが，映像内の出来事を直接的に説明したものではなく，映像終了時点から過去を問う質問である点であり，より難しい問題となっている．

本論文で提案する，複数カメラ映像に跨ったエピソード記憶タスクの入力クエリも過去を問う質問の形式である．

2.3 映像データセット

自然言語クエリとそれに対応する時間的アノテーションがされた映像データセットはこれまでにいくつか作成されている．TACoS [10] は，MPII Cooking Composite activity ビデオコーパス [11] をもとにテキスト記述を追加したデータセットである．MPII Cooking Composite activity ビデオコーパスは 1 分から 23 分の調理中の映像を提供している．調理法に関するアクティビティラベル，オブジェクト，アクティビティの開始時刻と終了時刻がアノテーションされている．TACoS ではアクティビティラベルに加えて映像内の出来事を説明する文章が追加でアノテーションされている．Gao らの研究 [3] で提供された Charades-STA は，映像に対する説明文と人の動作ラベルが付いたデータセットである Charades [12] をもとに，半自動で文章と時間的アノテーションを追加している．Ego4D [4] は，931 人のカメラ着用者が撮影した一人称視点の日常生活映像データセットを提供している．5 つのタスクに対し映像にアノテーションを行い，それぞれベースラインモデルを開発し評価を行っている．既存のデータセットより大規模で多様性のあるデータセットであり，数百万のアノテーションも前例のない規模と深さを備えている．

複数視点のカメラ映像を提供する映像データセットとしては Narrative Dataset [13] がある．このデータセットの映像内には商品購入，バイク窃盗，ストーカーなどのイベントが複数含まれている．複数イベントが含まれる複数の視点映像に基づく人物行動データセットは今までになかった．しかしながら Narrative Dataset は自然言語クエリとそれに対応する時間的アノテーションがされていない．そこで本論文は追加でそれらの時間的アノテーションを行うことで複数カメラ映像に跨ったエピソード記憶タスクのためのデータセットを構築する．

3 複数カメラ映像に跨ったエピソード記憶タスク

3.1 問題定義

本論文における入出力を図 1 に示す．ある 1 つの環境に対する 3 視点分の映像と，その映像に対するクエリを入力とする．このクエリに対応する映像中の開始時刻と終了時刻を出力とする．

今回入力されるクエリは表 1 に示すクエリテンプレートに従って作成される．表 1 中の X と Y はイベント，Z は人の属

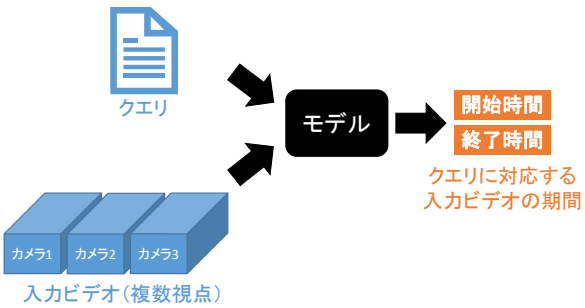


図 1: 入出力

表 1: クエリテンプレート

	テンプレート
When	Z did X. When did the person Y?
Where	Z did X. Where did the person Y?
How long	Z did X. How long did the person Y?
Who	Z did X. Who did the person Y?
Which	Z did X. Which O did the person Y?

性，O は物体である．クエリにはイベント X とイベント Y の 2 つのイベントが含まれている．イベントとは人が見たときに意味があると判断できる行動のことを指す．例えば，「商品を購入する」という行動や「本を読む」という行動などがイベントとして挙げられる．そして，クエリに含まれるイベント X とイベント Y は属性 Z を持つ同一人物が行ったイベントである．

表 1 のクエリテンプレートに従うクエリが入力されたときに，イベント Y に相当する開始時刻と終了時刻を出力することが今回解く問題となる．イベント X とイベント Y は同一カメラ映像に含まれる場合もあれば，別々のカメラ映像に含まれる場合もある．そしてイベント Y のみでは人物を特定できない場合がある．したがって，この問題を解くためには，クエリに含まれるイベント X や属性 Z の情報を利用して，イベント Y の開始時刻と終了時刻を特定することが必要となる．

3.2 データセット

Narrative Dataset [13] を利用する．Narrative Dataset にはシナリオと呼ばれる環境設定の映像が 5 シナリオ分ある．シナリオにはコンビニエンスストア環境やバー店舗環境などがあり，1 シナリオに対し 3 視点分の固定カメラ映像がある．Narrative Dataset には，各フレームにおける出現オブジェクトやそのバウンディングボックス，人物のアクション，オブジェクト同士の関係などがすでにアノテーションされている．人物のアクションには standing，walking などの動作ラベルが付与されている．オブジェクト同士の関係では，2 つのオブジェクトに対して touching，at side of，speaking to などのラベルが付与されている．本論文では，これらのアノテーションに加えて，追加で時間的アノテーションを行う．

3.3 追加アノテーション

Narrative Dataset に対して，クエリテンプレートに従うクエリと，そのクエリに対応する開始時刻と終了時刻をアノテー

ションする方法を示す．今回はできる限り自動でアノテーションができるように，大規模言語モデル（LLM）を活用してクエリを生成することとした．追加アノテーションは以下の手順で行う．

（1） 人物の属性とイベント一覧を取得
（2） 各人物に対してイベントペアを作成
（3） 属性とイベントペアから正解入出力を作成

最終的に（3）で取得できたクエリと開始時刻と終了時刻を，複数カメラ映像に跨ったエピソード記憶タスクに用いる．

3.3.1 人物の属性とイベント一覧を取得

クエリテンプレートにおけるイベント X，イベント Y，属性 Z に当てはまるイベントと属性を最初に取得する．

イベントとは人が見たときに意味があると判断できるものを指す．例えば，本を手にして読んでいる状態は read a book イベントであり，配達物を持ちながら歩いている状態は carry a delivery item イベントである．ただ歩いているだけの状態や，立っているだけの状態はイベントには含めないこととする．アノテーション対象のイベント種別の一覧を表 2 に示す．映像中の出現人物には表 2 にあるイベントが該当する期間の開始時刻と終了時刻がアノテーションされていれる．表 2 の多くのイベントに対するアノテーションは既に Narrative Dataset にある関係性アノテーションを流用し，不足しているイベントアノテーションに関しては新たに人手で付与することとした．また Narrative Dataset には属性に関するアノテーションはないため，各人物に対して性別，年代，服の上下，髪についても人手で付与する．

3.3.2 各人物のイベントペアを作成

3.3.1 節で取得した各イベントを用いて，クエリテンプレートにおけるイベント X とイベント Y に当てはめるためのイベントペアを作成する．ペアにするイベントは同一人物のイベントである必要があるため，各登場人物ごとにその人物のイベント同士を組み合わせる操作を行う．もし 1 人の人物のイベントが n 個取得できているとすると，${}_nP_2$ 通りのイベントペアを作成する．全ての登場人物に対してこの操作をそれぞれ行う．

3.3.3 属性とイベントペアから正解入出力を作成

3.3.1 節で取得した人物の属性と，3.3.2 節で取得したイベントペアから，LLM を用いて入力クエリ文を自動で生成する．なお LLM としては GPT-3.5-turbo [1]を利用した．LLM を利用した入力クエリ文生成のための問い合わせ文の例を表 3 に示す．文中の「...」には表 1 のクエリテンプレートを埋める．Z, X, Y には人物の属性とイベントペアの情報を埋めて LLM への問い合わせ文を作成する．

そして，入力クエリ文の元となったイベントペアのうちイベント Y に相当するイベントの，3.3.1 節で取得した開始時刻と終了時刻をクエリ文の回答となる期間の開始時刻と終了時刻として付与する．

1：https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

4 人物・イベント関連検出

本論文で提案する複数カメラ映像に跨ったエピソード記憶タスクの最終的な出力はイベント Y の開始時刻と終了時刻である．しかし，イベント Y のみでは人物を特定することができない場合もあり，開始時刻と終了時刻を特定することができない．よって，イベント X や属性 Z の情報を利用して人物を特定する必要がある．そこで，イベント X と関連性の高い人物を検出することができるかどうか試すこととした．

イベント関連人物検出の流れを図 2 に示す．ここでは，入力クエリと入力映像からイベント特徴と人物特徴をそれぞれ抽出し，イベント特徴と人物特徴を用いて人物・イベント関連スコアを算出する．

4.1 イベント特徴抽出

入力されたクエリから，1 文目と 2 文目を切り分ける．例えば，「A person rode a bicycle. When did the person purchase a store item?」というクエリであれば，「A person rode a bicycle.」と「When did the person purchase a store item?」に分ける．1 文目をイベント X についてのテキストとみなし，文頭に「a photo of」を付けて「a photo of A person rode a bicycle.」というテキストにする．バックグラウンドテキストとして「a photo of person」を用意する．事前学習済みモデルである CLIP [7] の Text Encoder を用いて，イベント X テキストとバックグラウンドテキストのテキスト特徴をそれぞれ抽出する．

4.2 人物画像特徴抽出

まず，入力された映像から人物画像列の取得を行う．人物画像列とは，ある 1 人の登場人物にフォーカスしたときに，その人の全ての登場画像を並べたものとする．

人物画像列を取得するために，カメラ映像に対して物体追跡を行う．今回は物体追跡が完全にできると仮定し，Narrative Dataset [13] のオブジェクトに対するバウンディングボックスのアノテーションを利用する．ある 1 人の人物の登場フレームに対し，その人物のバウンディングボックスから短辺が長辺と同じサイズになるよう正方形に切り抜く．この切り抜いた画像を並べたものを人物画像列とする．入力映像の全ての登場人物に対して同様に人物画像列を取得する．さらに事前学習済みモデルである CLIP [7] の Image Encoder を用いて，得られた人物画像列から人物画像特徴を抽出する．

4.3 人物・イベント関連スコア

各人物の人物特徴列の長さは，登場フレーム枚数と対応している．各特徴量に対してイベント X 特徴とバックグラウンド特徴とのコサイン類似度を算出する．イベント X 特徴とのスコアとバックグラウンド特徴とのスコアで Softmax をかけることでイベント X に該当する人物である確率とそうでない確率が出力できる．縦軸をイベント X に該当する人物である確率，横軸を全体のフレーム数としてプロットすることで，どのフレー

表 2: イベント一覧

イベント		
ride a bicycle	read a book	carry a delivery item
take a shopping basket	call on a cell phone while walking	talk
put a handbag	put a sidebag	the first appearance at the location
the last appearance at the location	purchase a store item	enter a building
exit a building	enter a room	exit a room
enter a shop	exit a shop	check inside a hand bag
snatch a hand bag	be snatched a hand bag	steal a hand bag
be stolen a hand bag	look at a shelf	look for a handbag
look around restlessly	sway a stroller	wave a hand
put a bag on the table	change a seat	

表 3: 入力クエリ文生成のための LLM への問い合わせ文の例.

LLM への入力例：
これからある人物の 2 件の行動とその人物の属性を提示します．この 2 件の行動を X と Y とし，この人物の属性を Z とします． X と Y と Z を組み合わせて，Y が行われた時間を問う文章を作ってください． このとき,「Z did X. When did the person Y?」というテンプレートに沿って英語の文章を作ってください． X: purchase a store item Y: the last appearance at the location Z: a person 提示した行動 X と行動 Y と属性 Z から英語の文章を作ってください．
LLM からの出力例：
A person purchased a store item. When was the person's last appearance at the location?

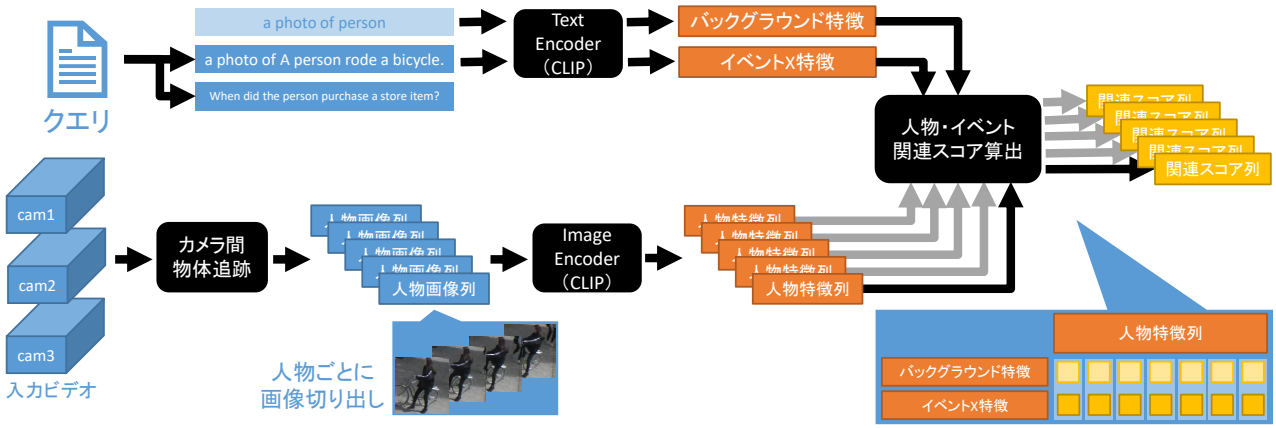


図 2: イベント特徴と人物画像特徴の抽出と人物・イベント関連スコアの算出

ムでイベント X に近い動作を行っているかを確認する．

5 実　　験

イベント X になりうる ride a bicycle イベントと carry a delivery item イベントについて，そのイベントを行った人物とそうでない人物に関して検出することができるかどうか実験を行った．

5.1 ride a bicycle イベント

コンビニエンスストアシナリオにおける自転車に乗って来た人物の人物画像列と，コンビニエンスストア店員の人物画像列に対して人物特徴を抽出し，イベント X のテキストとして「A person ride a bicycle.」というテキストとバックグラウンドテキストのテキスト特徴を抽出し，人物・イベント関連スコアの算出を行った．

算出したスコアから，縦軸をイベント X に該当する人物である確率，横軸を全体のフレーム数としてプロットしたものをそれぞれ図 3，図 4 に示す．cam1 はコンビニエンスストアの外の映像，cam2 と cam3 はコンビニエンスストアの店内の映像である．自転車に乗って来た人物で特にスコアが高いフレームのサンプルを図 5 に，コンビニエンスストア店員で特にスコアが高いフレームのサンプルを図 6 に示す．自転車に乗って来た人物に関しては，図 5a のような自転車に乗っている様子が映っているフレーム番号 16000 付近のスコアが高くなっている．ま

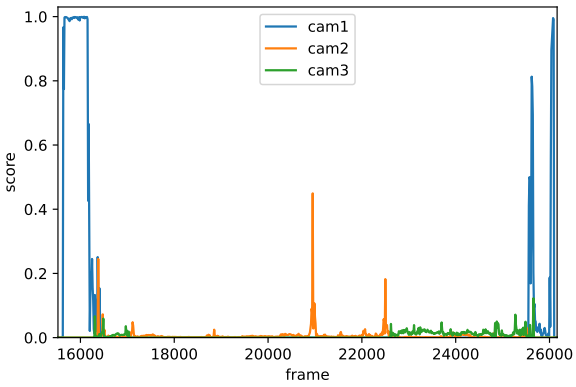


図 3: 自転車に乗って来た人物に対する関連スコア

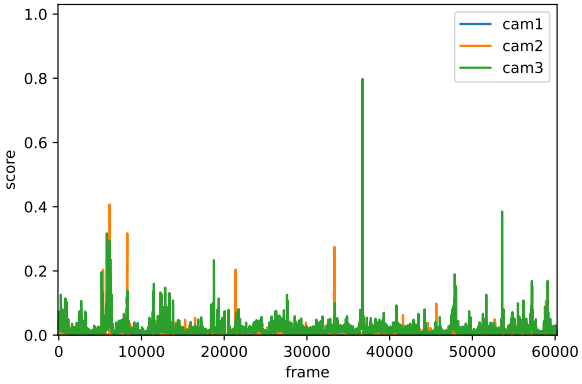


図 4: コンビニエンスストア店員に対する関連スコア



(a) フレーム番号 16004　　(b) フレーム番号 26050

図 5: 自転車に乗って来た人物のサンプルフレーム



(a) フレーム番号 6138　　(b) フレーム番号 36696

図 6: コンビニエンスストア店員のサンプルフレーム

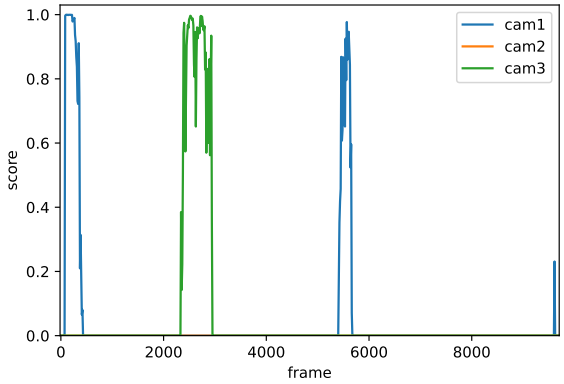


図 7: 配達物を持って来た人物に対する関連スコア

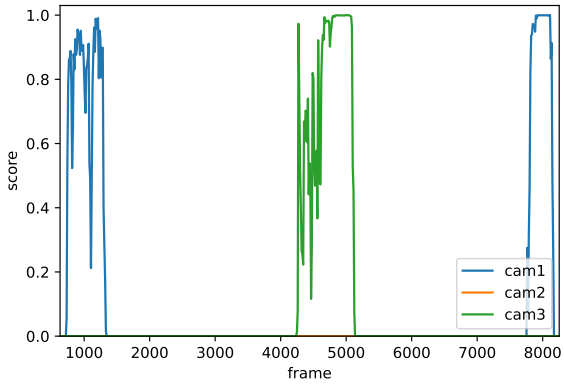


図 8: 配達物を盗んだ人物に対する関連スコア

た，図 5b のような乗って来た自転車とは違う自転車が映った瞬間もスコアが高くなっている．一方でコンビニエンスストア店員に対しては，図 6a や図 6b などのフレームで瞬間的に高いスコアが出ていることがあるものの，全体的には低いスコアが出力されている．

以上より，自転車自体に関してはかなり検出できていることが分かった．自転車に乗っている場合，連続的にスコアが高くなるため,「自転車に乗る」というイベントを検出するためには複数フレームにわたってスコアが高い部分があるかどうかという点を考慮する必要があると言える．

5.2 carry a delivery item イベント

バー店舗シナリオにおける配達物を持って来た人物の人物画像列と，配達物を盗んだ人物の人物画像列と，配達物とは関係がない人物の人物画像列に対して人物特徴を抽出し，イベント X のテキストとして「A person carry a delivery item.」というテキストとバックグラウンドテキストのテキスト特徴を抽出し，人物・イベント関連スコアの算出を行った．

算出したスコアから，縦軸をイベント X に該当する人物である確率，横軸を全体のフレーム数としてプロットしたものをそれぞれ図 7，図 8，図 9 に示す．cam1 は建物の外の映像，cam2 は建物の中にあるバーの映像，cam3 は建物の中にある配達先

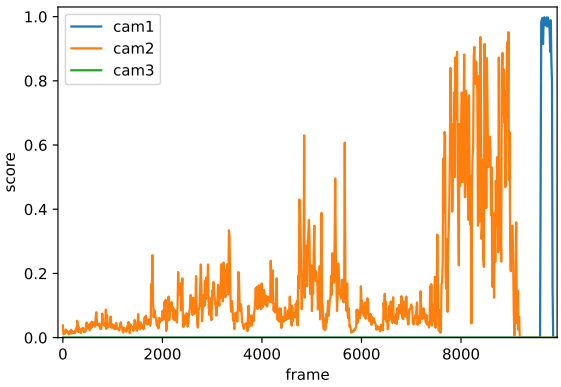


図 9: 配達物とは関係がない人物に対する関連スコア



(a) フレーム番号 108　　(b) フレーム番号 2518

図 10: 配達物を持って来た人物のサンプルフレーム



(a) フレーム番号 4999　　(b) フレーム番号 8044

図 11: 配達物を盗んだ人物のサンプルフレーム

の映像である．配達物を持って来た人物で特にスコアが高いフレームのサンプルを図 10 に，配達物を盗んだ人物で特にスコアが高いフレームのサンプルを図 11 に，配達物とは関係がない人物で特にスコアが高いフレームのサンプルを図 12 に示す．配達物を持って来た人物は，図 10a のように外から配達物を運んでいるフレーム番号 100 から 400 付近と，図 10b のように配達物を置いているフレーム番号 2300 から 3000 付近でのスコアが連続的に高くなっている．配達物を置いた後のフレーム番号 5400 から 5600 に関しては局所的にスコアが高くなっているのみにとどまっている．スコアが連続的に高くなっている部分はほとんど 1.0 に近い高さで推移しているため，しきい値を設けることで局所的に高くなっている部分と区別できる可能性がある．配達物を盗んだ人物も同様に，図 11a や図 11b のよ



(a) フレーム番号 8380　　(b) フレーム番号 9663

図 12: 配達物とは関係がない人物のサンプルフレーム

うに配達物を持ち運んでいる部分のスコアが連続的に高くなっている．ただ，配達物を持っていく前のフレーム番号 700 から 1300 において，局所的にスコアが高くなっている部分があり，配達物を持って来た人物とは逆のようなスコアの推移をしている．配達物とは関係がない人物は，cam2 のバーで長時間座っており，図 12a や図 12b のように外に出るタイミングでかばんを持っていくため，その部分のスコアが少し高くなっている．

以上より，配達物を運ぶイベントは配達物ではなく何かを運んでいるイベントを検出してしまうことがわかった．ただ，配達物を持って来た人物の物を運んできて置くという動作と，配達物を盗んだ人物の置いてある物を持っていくという動作における関連スコアの推移に違いが見られるため，実際に配達している人物とそうでない人物の関連スコアの違いを学習することで，配達物を運ぶイベントを検出できる可能性がある．

6　まとめと今後の課題

本論文では複数カメラ映像に跨ったエピソード記憶タスクを提案した．そして，提案タスクの詳細とその正解アノテーションの詳細を述べ，事前学習済みモデルである CLIP を用いた実験の結果を示した．今回はクエリとフレーム毎の関連スコアの算出までを行ったが，イベントが起きた時刻を特定するまでには至っていない．今後はこの関連スコアを用いて開始時間と終了時間との関連を学習することによって，クエリに対応する映像の開始時間と終了時間を特定できるようなモデルを作成する予定である．

参 考 文 献

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. “Localizing Moments in Video With Natural Language”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV’17)*. 2017, pp. 5803–5812.

[2] Pierfrancesco Bruni, Alex Falcon, and Petia Radeva. “Time-Aware Circulant Matrices for Question-Based

Temporal Localization". In: *Proceedings of the 22nd International Conference on Image Analysis and Processing (ICIAP'23)*. 2023, pp. 182–195.

[3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. "TALL: Temporal Activity Localization via Language Query". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 2017, pp. 5267–5275.

[4] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: *Proceedings of the 2022 IEEE/CVF Computer Vision and Pattern Recognition (CVPR'22)*. 2022, pp. 18995–19012.

[5] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, et al. "Egocentric Video-Language Pretraining". In: *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS'22)*. 2022, pp. 7575–7586.

[6] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. "Collaborative Static and Dynamic Vision-Language Streams for Spatio-Temporal Video Grounding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*. 2023, pp. 23100–23109.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*. 2021, pp. 8748–8763.

[8] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. "NaQ: Leveraging Narrations As Queries To Supervise Episodic Memory". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR'23)*. 2023, pp. 6694–6703.

[9] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. "SpotEM: Efficient Video Search for Episodic Memory". In: *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. 2023, pp. 28618–28636.

[10] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. "Grounding Action Descriptions in Videos". In: *Transactions of the Association for Computational Linguistics* 1 (2013), pp. 25–36.

[11] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. "Script Data for Attribute-Based Recognition of Composite Activities". In: *Proceedings of the 12th European Conference on Computer Vision (ECCV'12)*. 2012, pp. 144–157.

[12] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding". In: *Proceedings of the 14th European Conference on Computer Vision (ECCV'16)*. 2016, pp. 510–526.

[13] Karen Stephen, Rishabh Sheoran, and Satoshi Yamazaki. "Narrative Dataset: Towards Goal-Driven Narrative Generation". In: *Proceedings of the 1st Workshop on User-Centric Narrative Summarization of Long Videos (NarSUM'22)*. 2022, pp. 7–12.

[14] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. "UnLoc: A Unified Framework for Video Localization Tasks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision 2023 (ICCV'23)*. 2023, pp. 13623–13633.

[15] Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. "Relational Space-Time Query in Long-Form Videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR'23)*. 2023, pp. 6398–6408.

[16] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. "Span-based Localizing Network for Natural Language Video Localization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*. 2020, pp. 6543–6554.

[17] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. "Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language". In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*. 2020, pp. 12870–12877.

不均衡データを活用した認証手法の提案
ータッチベース認証を対象としてー

古川 大樹† 工藤 雅士‡ 宮﨑 公彦‡ 山名 早人§

†早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§ 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1
E-mail: †‡§ {taikifurukawa, kudoma34, loup, yamana}@yama.info.waseda.ac.jp

あらまし 認証システムの学習データは多くの場合，本人データよりも他人データが多い不均衡なデータである．しかし不均衡なデータの学習は，多数派に偏った予測を招く可能性がある．本稿では，近年スマートフォンの新たな認証手法として期待されているタッチベース認証を題材として，学習データが不均衡データであることを前提とした新たな認証手法を提案する．具体的には，「アンダーサンプリングを適用したデータで訓練した分類器」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」をスコアレベルで統合した手法を提案する．評価実験では，統合前の２つの分類器をベースラインとし，ベースラインと提案手法間で精度比較を行った.「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」と提案手法の精度比較の結果，Random Forest を使用した検証で最大 0.23%（14.68%から 14.45%）の EER 改善を確認した．

キーワード 認証，タッチベース認証，スマートフォン，不均衡データ，アンダーサンプリング，オーバーサンプリング

1. はじめに

近年，スマートフォンのセキュリティに対する関心が高まる中，新たな認証システムの開発や既存の認証システムの精度向上が進められている．認証システムの構築に使われるデータセットは多くの場合，本人のデータに対して他人のデータが豊富にある不均衡データとなっている．これは 1 人 n 個のデータを収集した際，本人のデータは n 個であるのに対して他人のデータは n×人数分のデータが存在するためである．不均衡データのまま認証システムに使用する分類器の訓練を行った場合，データ数の偏りによって分類器の精度が著しく低下する可能性がある．その対策としては主に，本人と他人のデータ数を揃えて学習するデータレベルの対策[1-10]や，分類器の訓練方法を工夫するアルゴリズムレベルの対策[11][12]がある．本稿では 2 種類の対策のうち，様々な分類器と組み合わせることが容易であるデータレベルの対策として新たな手法を提案し,データレベルの対策のさらなる精度向上を図る．

本稿ではタッチベース認証を題材として，本人の特徴をよく捉えた分類器と他人の特徴をよく捉えた分類器を統合するというアイデアのもと,「アンダーサンプリングを適用したデータで訓練した分類器」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」をスコアレベルで統合する認証方式を提案する．提案手法は，現在すでに提案されている他の手法と組み合わせることが容易であり，様々な分類器と組み合わせることでさらなる精度向上が期待できる．

本稿では次の構成をとる．まず 2 節で背景知識と関連研究について述べ，3 節で提案手法について説明する．4 節では評価実験の説明，5 節では実験の結果の説明と考察を行い，6 節で本稿をまとめる．

2. 背景知識と関連研究

本節では，不均衡データとその対策方法，及び，タッチベース認証について説明する．

2.1. 不均衡データ

複数のクラスを持つデータを扱う場合，クラス間でデータ量の偏りが生じることは少なくない．このような不均衡データでは一般的に少数派のクラスの分類が難しい場合が多く，その解決方法としては大きく分けて 2 通りの方法が存在する[13][14].

1 つ目の解決方法は，データレベルで解決する方法である．データレベルで解決する方法では，多数派のデータを削減するアンダーサンプリングを適用する手法[1-3]，少数派のデータをかさ増しするオーバーサンプリングを適用する手法[4-7]，そしてこの 2 つをどちらも適用する手法が提案されている[8-10].

2 つ目の解決方法は，アルゴリズムレベルで解決する方法である．アルゴリズムレベルで解決する方法としては，学習アルゴリズムや学習の前処理・後処理にコスト（重み）を導入する手法（cost-sensitive learning）[11]や，1 つのクラスのみを用いて学習する手法（one-class learning）[12]が存在する．

本稿では，様々な分類器への適用のしやすさの観点からデータレベルで解決する方法に着目しているため，

以降ではアンダーサンプリング，オーバーサンプリング，そしてこれらを両方適用する方法について説明する．

2.1.1. アンダーサンプリング

アンダーサンプリングは，多数派のデータを削減することでデータの不均衡を解消する手法である．アンダーサンプリングの方法としては，１）RUS (Random UnderSampling)[13]と２）informed undersampling[1-3]の2種類がある．RUSでは，データをランダムに削減する．一方でinformed undersamplingでは，少数派のデータと多数派のデータの距離をもとにクラスの境界に近い多数派のデータのみ抽出するなど，計算をもとにデータを削減する．アンダーサンプリングの利点としては，本物のデータのみを学習に使用する点や計算量が削減できる点が挙げられる．また，informed undersamplingの手法によってはノイズを除去するなどデータを整備した上で学習に使用することも可能である．一方で，アンダーサンプリングは少数派のデータ量に合わせて多数派のデータを削減するため，多数派の重要な特徴を捉えきれない可能性がある[13].

2.1.2. オーバーサンプリング

オーバーサンプリングは，少数派のデータをかさ増しして少数派と多数派のデータの不均衡を解消する手法である．オーバーサンプリングの手法としては，元のデータをランダムに複製するrandom oversampling [13]や，元のデータから擬似的なデータを作り出す手法があり,擬似的なデータを作り出す手法ではSMOTE (Synthetic Minority Over-sampling Technique)[4]とその派生手法[5][6][7]が広く知られている．オーバーサンプリングの利点としては，訓練に使用できるデータ数を増やすことができる点が挙げられる．一方で，元のデータに基づいて擬似的なデータを生成するため，元のデータと同じようなデータの学習量が増えてしまい，過学習に繋がりやすい点が欠点となる[13].

2.1.3. アンダーサンプリングとオーバーサンプリングの同時適用

前述したアンダーサンプリングとオーバーサンプリングのそれぞれの欠点を補い合うことを目的に，両方の手法を適用する研究も存在する[8-10]．具体的には，多数派のデータと少数派のデータをある特定のデータ量で均衡化するために，多数派のデータをアンダーサンプリングで減らしつつ，少数派のデータはオーバーサンプリングで増やしている．

2.2. タッチベース認証

現代のスマートフォンは主にパスコード，パターン認証，指紋認証，顔認証などの認証方式で保護されているが，これらの認証方式はショルダーサーフィン[15]やスマッジ攻撃[16]，生体複製[17]などで破ることができる．また，一度破られると以降はデバイスを保護することができないという問題がある．このような背景から，ユーザが本人であるかどうかを継続的に監視し続ける継続認証としてタッチベース認証が注目されている．

本項では，タッチベース認証の流れ，利点，不均衡データへの対策，課題，最新の研究を説明する．

2.2.1. タッチベース認証の流れ

タッチベース認証は，データ収集，前処理，特徴量抽出，マッチング，判定の5つのステップに分けることができる．まずデータ収集では，タッチセンサからの時間情報と位置や圧力などのタッチ情報を含む生データを収集する．次に前処理で外れ値の除外やデータの標準化を行う．そして前処理をしたデータからストロークの持続時間や速度・加速度などの認証に使用する特徴量を抽出する．マッチングでは，特徴量が抽出されたストロークを使用して事前に学習した本人のストロークとどれだけ一致しているか計算する．そして最後に計算結果をもとに本人か偽者かの判定を行う．

2.2.2. タッチベース認証の利点

タッチベース認証は，ユーザによる追加操作無しで認証ができる点と，認証に特別なセンサを必要としない点が利点として挙げられる.タッチベース認証では，ユーザが普段通りにスマートフォンを操作している状態のタッチ情報を取得して認証を行うため，現代のスマートフォンの認証に使われているパスコード，パターン認証，指紋認証や顔認証などと違い，ユーザが認証のための操作をする必要がない．ただし，タッチベース認証は単体で使用することを想定しておらず，すでに搭載されている認証システムと併せて利用することで,セキュリティの向上を目指す認証である.また，指紋認証や顔認証では専用のセンサをスマートフォンに搭載する必要があるのに対して，タッチベース認証で使用するタッチセンサはほとんどのスマートフォンに標準で搭載されているため，新たなセンサを必要としない利点がある．

2.2.3. タッチベース認証における不均衡データへの対策

タッチベース認証の分野における不均衡データへの対策としては，オーバーサンプリングとしてSMOTE[4]を適用しているもの[18]やSMOTEの派生手法であるADASYN[6]を適用しているもの[19][20]も存在したが，本人のデータ数に合わせて他人のデータを使用するアンダーサンプリングを適用している場合が主であった．

2.2.4. タッチベース認証の課題

タッチベース認証の現状の課題としては，認証の精度が挙げられる．Frankらの研究[21]で示されているよ

うに，タッチストロークにはユーザ間で十分な違いがあり認証に使用できる可能性を持っているものの，同じユーザでも周りの環境やユーザの状況などに応じて時間的変化が大きいことも知られている[21]. また，スマートフォン上で常に監視し続けるため，システムとしての計算量や電力消費量も小さく抑える必要がある．その制約の中で認証システムとして求められる精度を達成することが現状の課題となっている．

2.2.5. 最新のタッチベース認証研究

本項では，認証精度の高いタッチベース認証の最新の研究を紹介する．

Zaidi ら[22]は 2022 年に，分類器を 6 種類（SVM（Support Vector Machine），NB（Naïve Bayes），DT（Decision Tree），kNN（k-Nearest Neighbor），LR（Logistic Regression），MLP（Multiple Layer Perceptron））用意し，その中から 1 ストロークごとに最適な分類器を 1 つまたは複数動的に選択し，多数決のアンサンブルで判定を行う手法を提案した．Zaidi らは，動的な分類器の選択手法（DES: Dynamic Ensemble Selection）として 12 種類の手法を検証し，RRC（Randomized Reference Classifier，以下 DES-RRC）が最も安定的に高精度を実現できると報告している．論文で提案した手法の最高精度としては Frank ら[21]のデータセットのセッション内水平ストロークを用いたシナリオにおいて EER 0.55%を達成している．ただし，アンサンブル学習では複数の分類器での推論が必要であり，計算コストが大きくなる点が欠点として挙げられる．不均衡データへの対応としては，アンダーサンプリングのみを行っている．

3. 提案手法

本節では，提案手法について詳しく説明する．

3.1. 提案概要

不均衡データへの対策であるデータレベルの対策とアルゴリズムレベルの対策のうち，データレベルの対策は様々な分類器と組み合わせることが容易という利点がある．そこで，データレベルの対策のさらなる精度向上を目的に，本人の特徴をよく捉えた分類器と他人の特徴をよく捉えた分類器を統合するというアイデアのもと,「アンダーサンプリングを適用したデータで訓練した分類器」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」をスコアレベルで統合する手法を提案する．

2 節で述べたように，アンダーサンプリングを適用すると本物のデータのみで訓練することができるが，多数派のデータは一部失われてしまう．反対に，オーバーサンプリングを適用すると多数派のデータをすべて利用することができるが，少数派のデータは本物のデータに基づいて擬似的に作り出したデータを追加して学習するため，過学習になりやすい．

一般的に認証システムでは，本人は 1 ユーザのみでそれ以外のユーザは全て他人として扱われるため，本人のデータが少数派，他人のデータが多数派となる．他人のデータに対してアンダーサンプリングを適用した場合，本人のデータは全て本物のデータを使用することができるが,他人のデータは一部が失われるため,多様な特徴を捉えきることができない．一方で，本人のデータに対してオーバーサンプリングを適用した場合，他人の多様な特徴を捉えることができるが，本人に関しては本物のデータに基づいて生成した疑似的なデータを使用することになる．オーバーサンプリングの欠点である過学習を防ぐために，データにアンダーサンプリングとオーバーサンプリングの両方を適用する手法も存在する[8-10]が，このような手法においても本人に関しては疑似的なデータを使用することになる．

以上をまとめると,「アンダーサンプリングを適用したデータで訓練した分類器」は本人の正しい特徴を学習しているものの，他人の判定に弱い．そして,「オーバーサンプリングを適用したデータまたはアンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」は他人の判定に強いものの，本人については過学習している可能性がある．したがって，前者は本人の特徴をよく捉えていて，後者は他人の特徴をよく捉えていると解釈することができる．

3.2. システムの構築方法

本項では，システムの構築方法を，１）データの成形，２）分類器の訓練，３）分類器の予測結果の統合の 3 つの段階に分けて説明する．

3.2.1. データ成形

認証システムの構築において，データはスマートフォンの所有者である本人のデータとその他の他人のデータに分けられる．多くの場合，他人のデータは複数人から集められたデータであるため，データが不均衡となってしまう．

提案手法では，訓練の際に「アンダーサンプリングを適用したデータ」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータ」の 2 つが必要である．「アンダーサンプリングを適用したデータ」は，少数派のデータである本人のデータの数に合わせて多数派である他人のデータをアンダーサンプリングすることで抽出する.「アンダーサンプリングとオーバーサンプリングを両方適用したデータ」は，少数派のデータである本人のデータを規定の数(実験時に設定)までオーバーサンプリングし，その数に合わせて多数派である他人のデータをアンダーサンプリングしてデータを抽出する．

以上のようにして得られた本人のデータと他人のデータをそれぞれ訓練に使用する．本項で説明したデータ成形の部分を含め，提案手法の流れを図 1 に示す．

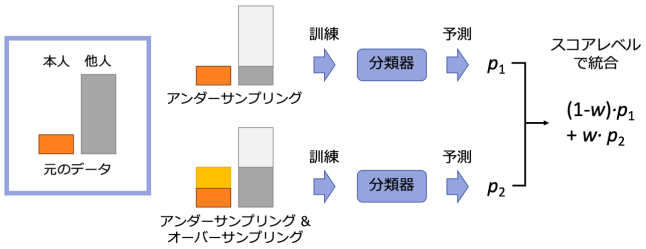


図 1 提案手法の流れ

3.2.2. 分類器の訓練

3.2.1 項で生成した，「アンダーサンプリングを適用したデータ」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータ」をそれぞれ使用して分類器を訓練する．訓練の段階では判定結果を統合することは考えず，それぞれ別々に訓練を行う．

3.2.3. 分類器の予測結果の統合

3.2.2 項で訓練した 2 つの分類器の統合においては，2 つの分類器の判定結果である本人確率を，重み付き和を用いて統合する．重み付き和は，本人の特徴をよく捉えた分類器と他人の特徴をよく捉えた分類器の予測結果の影響を，重みを変えることで調整することができる．「アンダーサンプリングを適用したデータで訓練した分類器」が出力した本人確率をp_1，「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」が出力した本人確率をp_2とすると，統合した本人確率pは以下の式(1)で表される．

$$p = (1-w)\cdot p_1 + w\cdot p_2 \quad (0 \le w \le 1) \qquad (1)$$

4. 評価実験

本節では，提案手法の有用性を評価するために実施した評価実験について説明する．

4.1. 認証シナリオ

評価実験は，以下の表 1 にまとめた 4 つのシナリオについて実施した．

表 1 各シナリオとその説明

シナリオ	説明
v-inter	上下方向のセッション間の検証
h-inter	左右方向のセッション間の検証
v-intra	上下方向のセッション内の検証
h-intra	左右方向のセッション内の検証

タッチストロークは時間に応じて特性が変化することから，セッション内での認証精度に比較してセッション間での認証精度が低下する[21]ため，セッション間とセッション内の検証に分け評価を行う．なお，セッションとは連続して収集されたデータの集合を言う．

4.2. データセット

評価実験には，独自に収集したデータセット eCommTouch[23]を使用した．独自のデータセットを使用した理由は，現在公開されていて利用可能なデータセット[21][24-26]はユーザ数が少なく，ストロークの種類ごとのデータ数にばらつきも大きいためである．

本データセットは，105 人の参加者から収集したスマートフォンにおける上下左右のタッチストロークから構成される，e コマースを対象としたデータセットである．上下方向 3 セッションと左右方向 3 セッションの合計 6 セッションに渡って，ストローク数が均等になるようにデータを収集した．実験には iPhone XS と iPhone 8 Plus が使用され，参加者は座った状態で右手の親指でスマートフォンを操作するよう指示された．1 つのストロークに含まれる特徴量として，Frank ら[21]によるデータセットと同じ 27 種類抽出した．抽出した特徴量の説明を表 2 に示す．

表 2 抽出した特徴量の説明

特徴量	説明
Inter-stroke time	ストローク間の時間間隔[ms]
Stroke duration	ストロークの持続時間[ms]
Start x, Start y	開始地点の x，y 座標[px]
Stop x, Stop y	終了地点の x，y 座標[px]
Direct end-to-end distance	始点と終点の距離[px]
Mean resultant length	ストロークの直線性[27]
Direction	方向が上下左右のどれか
20%, 50%, 80% pairwise velocity	20%，50%，80%地点での速度[px/ms]
20%, 50%, 80% pairwise velocity	20%，50%，80%地点での加速度[px/ms^2]
Median vel at last 3 pts	サンプリングの終了 3 点における速度中央値[px/ms]
Largest deviation from end-to-end line	始点終点を結んだ直線から一番離れた点への距離[px]
20%, 50%, 80% dev from end-to-end line	20%，50%，80%地点でのストロークの始点と終点を結んだ直線からの距離[px]
Average direction	平均角度[rad]
Ratio end-to-end dist and length of trajectory	（軌跡の長さ）/（始点と終点の距離）
Median acc at first 5 pts	サンプリングの開始 5 点における加速度中央値[px/ms^2]
Mid-stroke pressure	50%地点における圧力（0～1）
Mid-stroke area covered	50%地点での接触面積[px^2]
Mid-stroke finger orientation	50%地点における指の向き
Phone orientation	スマートフォンの向き

本稿では，iPhone XS を使用して収集した 78 人のデータから次に説明する方法でデータを成形した．

1 データセットは，全ユーザの内，1 ユーザを本人（正例），1 ユーザを偽者（テスト用の負例），残りのユーザを他人（訓練用の負例）としたデータから構成される．その上で，本人と偽者の組み合わせ分のデータセットを生成（$_{78}P_2$ 通り）する（5 節で詳述）．ここ

で，本人はスマートフォンの所有者，偽者は「本人のスマートフォン」を盗用しようとしているユーザである．データ成形の概要を図 2 に示す．

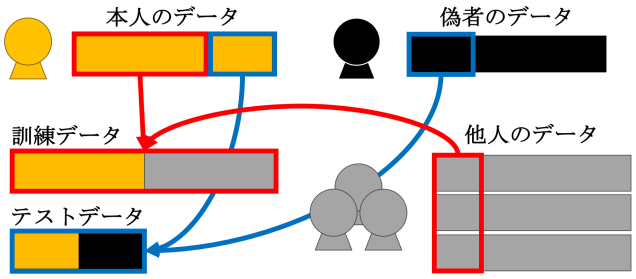


図 2 データ成形の概要

分類器から見て本人のデータを正例データ，本人でないデータを負例データとしてまとめると，訓練では本人のデータの一部を正例データ，他人のデータの一部を負例データとして使用する．テストでは本人のデータの一部を正例データ，偽者のデータの一部を負例データとして使用する．訓練，テストの際は正例データと負例データの数を均等に揃える．

提案手法では訓練データとして，「アンダーサンプリングを適用したデータ」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータ」を生成する．「アンダーサンプリングを適用したデータ」に関しては，正例データの数に応じて負例データを他人のデータからランダムに抽出する（RUS）．「アンダーサンプリングとオーバーサンプリングを両方適用したデータ」に関しては，正例データを元のデータ数の 10 倍まで SMOTE[4]でオーバーサンプリングし，その数に合わせて負例データを他人のデータからランダムに抽出する（RUS）．10 倍という数は，予備実験で{5, 10, 15, 20, 25, 30, 40, 76}倍の場合を比較し，最も精度が良かった値として 10 を選択した．訓練の負例に使用する他人は 76 人いるため，76 倍が最大値となる．

正例データは，セッション内の検証では，1 つのセッションで訓練およびテストを行い，セッション間の検証では，3 つあるセッションのうち最初の 2 つのセッションのデータで訓練をし，3 つ目のセッションでテストを行った．負例データは，セッションに関係なく訓練とテストに使用した．

4.3. 検証内容

本稿では，提案手法の有用性を確認するために 2 つの評価実験を実施した．

4.3.1. ベースライン手法

ベースラインとして統合前の 2 つの分類器である「アンダーサンプリングを適用したデータで訓練した分類器」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」を用意する．以降では，前者を BaselineU，後者を BaselineU&O と表記して説明する．

評価実験では，ベースラインとして BaselineU と BaselineU&O を用意し，提案手法である BaselineU と BaselineU&O をスコアレベルで統合する手法を評価した．「アンダーサンプリングを適用したデータで訓練した分類器」（BaselineU）をベースラインに設定しているにも関わらず「オーバーサンプリングを適用したデータで訓練した分類器」をベースラインとして設定していない理由としては，4.2 項で説明した予備実験の段階で「オーバーサンプリングを適用したデータで訓練した分類器」は倍率が 76 倍の場合に対応し，今回選んだ 10 倍の場合よりも精度が劣っていたためである．

3.2.3 項で示した予測結果を統合するための式(1)を用いると，$w = 0, 1$のときが 2 つのベースラインを表す．そして，$0 < w < 1$の範囲が 2 つの分類器を統合した提案手法となる．本稿では，提案手法において予測結果を統合する際の重み付けwを訓練後にユーザごとに$0 \leq w \leq 1$の範囲で最適な値を求め，テストを行った．データ量が限られていることから，本稿では，バリデーションのデータセットを別に設けるのではなく，訓練で使ったデータを再度用いて最適な重みwを求めた．分類アルゴリズムとしては，タッチベース認証の研究に多く使われている[28] RF（Random Forest），SVM，kNN の 3 つを使用した．これらの分類器のパラメータは scikit-learn ライブラリのデフォルトの値を使用した．

また，タッチベース認証をはじめとする継続認証は，認証に使用するデータが時系列的に連続して得られるため，1 つのストロークだけで判定するのではなく複数のストロークの予測をまとめて最終的な判定を行うことができる．判定結果をまとめる方法としては，本人確率の平均を取る方法，多数決を取る方法，最終判定を下すための新たな変数を作成し，予測結果に応じてその値を上下させる方法などが存在する[29]．今回はその中で最も主流となっている本人確率の平均を取る方法[29]を採用した．複数ストロークの判定結果を連続してまとめていく方法としては，スライディングウィンドウ形式を採用した．また，本稿では表 1 に示す各シナリオにおいて，スライディングウィンドウのウィンドウサイズが 1 の場合，5 の場合，10 の場合についてそれぞれ検証を行った．

4.3.2. DES-RRC との組み合わせ

提案手法は他の手法と組み合わせることが容易であることも利点の 1 つである．そのためベースラインとの比較に加えて，現在の state-of-the-art である Zaidi らの手法[22]と組み合わせた検証も実施した．本稿では，Zaidi らによる検証で最も汎用的で性能が良いとされた DES-RRC を組み合わせる対象とした．具体的には，DES-RRC にベースライン（BaselineU と

BaselineU&O）を適用した手法と DES-RRC に提案手法を適用した手法の精度比較を行った．

本項の検証で利用した DES-RRC の詳細については，Zaidi らの論文[22]で示してあったものに揃えた．最初に訓練する分類器としては SVM, NB, DT, kNN, LR, MLP を使用し，ハイパーパラメータに関しても表 3 に示すものの中から，ユーザごとに最適なパラメータを探索して使用した．シナリオとウィンドウサイズについてはベースラインとの比較と同様に検証を行った．

表 3 ハイパーパラメータの候補一覧
（[22]の Table 4 より引用）

モデル	ハイパーパラメータ	値または範囲
SVM	C	[0.001, 0.01, 0.1, 1, 10, 25, 50, 100, 1000]
	Kernel	RBF
	Kernel coefficient	1/number of features
	Tolerance of stopping criterion	1e-3
NB	-	-
DT	Maximum depth of the tree	[none, 5, 10, 15, 20, 25, 30]
	Minimum number of samples to split	[2, 4, 6, 8, 10]
	Minimum number of samples at a leaf	[1, 2, 3, 4, 5]
kNN	Number of neighbors	[1, ..., 10]
	Algorithm	KD tree
	Distance metric used for finding neighbors	Euclidean
LR	-	-
MLP	Number of hidden layers	1
	Number of hidden nodes	50

4.4. 評価方法

本項では，評価実験で使用した評価指標について説明する．タッチベース認証のような認証システムでは，本人（正）か偽者（負）の 2 パターンのストロークを本人（正）か偽者（負）の 2 パターンに分類するため，合計 4 つの分類パターンが考えられる．4 つの分類パターンを表 4 に示す．

表 4 混同行列

		予測したクラス	
		正	負
実際のクラス	正	TP (True Positive)	FN (False Negative)
	負	FP (False Positive)	TN (True Negative)

- **FAR（False Acceptance Rate）**

FAR は，実際は偽者のデータを誤って本人であると予測した割合を表し，式(2)で定義される．

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2}$$

FAR は認証システムにおいて偽者のユーザをどれだけ正しく拒否できるかを表し，セキュリティの信頼度に直結する指標であるため，認証システム評価において重要な指標である．

- **FRR（False Rejection Rate）**

FRR は，実際は本人であるデータを誤って偽者と予測した割合を表し，式(3)で定義される．

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

FRR は認証システムにおいて本人を拒否する確率を表すため，FAR と同様に認証システム評価において重要な指標である．

- **EER（Equal Error Rate）**

認証システムのような 2 値分類を扱う分類モデルは多くの場合，正である確率と負である確率を出力し，その確率をあらかじめ設定した閾値と比較することで最終的な予測を行う．すなわち閾値に依存して異なる混同行列が得られる．

EER は FAR と FRR の値が等しくなるように閾値を決定した際のエラー率を表す．EER は FAR と FRR の 2 つの指標に基づく評価指標であるため，1 つの値でセキュリティに関する観点とユーザの使用感に関する観点を同時に評価することができ，認証システムの分類性能を表す際に有用である．本稿ではこの EER を評価指標として採用した．

5. 評価結果と考察

4.2 項で説明したデータセットを使用し，ベースラインとの比較を行った．独自のデータセットに含まれる 78 人のデータから，1 人を本人，もう 1 人をテストの偽者，その他を訓練の偽者として使用する組み合わせは 6,006（$={}_{78}P_2$）通り存在するため，その平均 EER を 4 種類の認証シナリオ，3 種類のウィンドウサイズ，3 種類の分類手法，3 種類の分類アルゴリズムにおいてそれぞれ算出した．

ベースラインとの比較に加え，現状 state-of-the-art である Zaidi ら[22]によって提案された DES-RRC と組み合わせた検証も行った．DES-RRC にベースラインを適用した手法と DES-RRC に提案手法を適用した手法について，すべてのユーザの組み合わせの平均 EER を各シナリオ，ウィンドウサイズで算出し，比較を行った．

5.1. ベースラインとの比較

ベースラインとの比較の結果を表 5 にまとめる．セッション内のシナリオに関しては，3 セッションそれぞれの平均 EER の平均を取っている．ベースラインとの比較で各分類アルゴリズムの各シナリオにおいて最も性能の良かった手法を下線で強調し，すべての手法の比較で各シナリオにおいて最も性能の良かった手法を太字にしている．2 つのベースライン BaselineU と BaselineU&O それぞれと提案手法に対して対応のある t 検定を行い，有意水準 $p < 0.05$ において有意差が確

表 5 各シナリオ，ウィンドウサイズ，分類手法における平均 EER[%]

Scenario	Window	RF			SVM			kNN		
		BaselineU $w=0$	Proposed method $0 \le w \le 1$	BaselineU&O $w=1$	BaselineU $w=0$	Proposed method $0 \le w \le 1$	BaselineU&O $w=1$	BaselineU $w=0$	Proposed method $0 \le w \le 1$	BaselineU&O $w=1$
v-inter	1	15.24	***14.30**	14.48	30.76	19.61	*19.17	21.47	*18.13	18.18
h-inter	1	15.34	***14.45**	14.68	32.47	20.03	*19.76	23.16	19.46	19.46
v-intra	1	7.92	***6.61**	6.65	26.67	14.07	*13.84	15.52	*11.33	11.35
h-intra	1	8.15	***6.92**	6.99	28.08	15.17	*14.88	17.79	*13.17	13.19
v-inter	5	4.64	***3.95**	4.04	21.94	8.19	*7.74	8.10	*5.88	5.97
h-inter	5	5.19	***4.44**	4.49	21.68	8.61	*8.33	9.40	*6.80	6.86
v-intra	5	0.79	**0.42**	***0.42**	17.91	4.33	*4.05	3.87	*1.89	1.92
h-intra	5	0.94	***0.53**	**0.53**	16.44	4.70	*4.48	5.08	*2.55	2.57
v-inter	10	2.73	**2.21**	2.24	18.27	5.25	*4.97	4.50	*2.86	2.90
h-inter	10	3.64	2.81	**2.79**	17.74	6.06	*5.78	5.75	3.61	3.61
v-intra	10	0.24	**0.10**	***0.10**	14.69	2.50	*2.30	1.55	*0.52	0.53
h-intra	10	0.30	0.13	***0.12**	12.50	2.65	*2.48	2.21	*0.78	0.79

* $p < 0.05$ (Proposed method と BaselineU&O の検定)

認できた結果に＊を付与した（表 5）．なお，提案手法は，BaselineU に対して全てのパターンで有意に EER が低いことを確認しており，＊は BaselineU&O との比較での有意差の有無を示す．

表 5 から読み取れる結果を以下に示す．

- RF と kNN では提案手法が 24 種類の検証のうち，17 種類の検証で 2 つのベースラインよりも低い EER を達成している．
- SVM では RF，kNN と比較して，アンダーサンプリングよりもオーバーサンプリングを適用した方が精度の改善が大きく，提案手法よりも BaselineU&O の方が EER が低くなった．

5.2. DES-RRC との組み合わせ

DES-RRC にベースライン（BaselineU と BaselineU&O）を適用した手法と DES-RRC に提案手法を適用した手法の比較結果を表 6 に示す．ここで Zaidi らの論文[28]では本人ストロークから N ストローク，他人として(N/10)人を選んでそれぞれ 10 ストローク抽出することでデータ数を揃えている(N=40)．これは抽出方法が異なるものの，アンダーサンプリングを適用しているため，BaselineU と同等であると言える．よって本検証では DES-RRC に BaselineU を適用した手法は DES-RRC そのものとして捉えることができる．

表 6 から分かるように，すべてのシナリオおよびウィンドウサイズで BaselineU&O が最も低い EER を達成した．対応のある t 検定の結果，提案手法は両方のベースラインに対して，すべてのシナリオおよびウィンドウサイズで有意水準 $p < 0.05$ において有意差が確認できた．結果として，DES-RRC に関しては提案手法よりもアンダーサンプリングとオーバーサンプリングを両方適用した手法の方が効果的であった．これは，「本人の特徴をよく捉えた分類器と他人の特徴をよく捉えた分類器を統合する」という提案手法のアイデアが，DES-RRC による複数の分類器のアンサンブルの効果に含まれていたからであると考えられる．

表 6 DES-RRC との組み合わせの平均 EER[%]

Scenario	Window size	Baseline U $w=0$	Proposed method $0 \le w \le 1$	Baseline U&O $w=1$
v-inter	1	16.91	15.40	***15.21**
h-inter	1	16.77	14.92	***14.64**
v-intra	1	10.42	8.58	***8.41**
h-intra	1	10.65	8.85	***8.70**
v-inter	5	5.30	4.32	***4.20**
h-inter	5	5.62	4.48	***4.24**
v-intra	5	1.80	1.00	***0.87**
h-intra	5	1.83	1.07	***0.98**
v-inter	10	2.94	2.29	***2.20**
h-inter	10	3.65	2.89	***2.66**
v-intra	10	0.72	0.31	***0.26**
h-intra	10	0.71	0.35	***0.31**

* $p < 0.05$

6. おわりに

本研究では，スマートフォンにおけるタッチベース認証を題材として，学習データが不均衡データであることを前提とした新たな認証手法を提案した．具体的には,「アンダーサンプリングを適用したデータで訓練した分類器」と「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」をスコアレベルで統合する手法を提案した．独自のデータセットを用いた評価実験では，ベースラインとの比較と，DES-RRC に提案手法を組み合わせた検証を行った．ベースラインとの比較では，2 つのベースラインのうち，精度が良かった「アンダーサンプリングとオーバーサンプリングを両方適用したデータで訓練した分類器」とのウィンドウサイズ 1 における比較の結果，RF を使用した検証では最大 0.23%（14.68%から 14.45%），kNN を使用した検証では最大 0.05%（18.18%から 18.13%）の EER 改善を確認した．SVM を使用した検証では精度改善が見られなかった．DES-RRC に提案手法を組み合わせた検証でも，ベースラインと組み合わせた手法と比較して精度改善は見られなかった．

今後の課題としては，１）他のアンダー・オーバーサンプリング手法を使用した検証，２）様々な分類器と組み合わせた検証，３）ユーザごとの重みの決定方法の改善が挙げられる．

参 考 文 献

[1] D. L. Wilson, “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,” IEEE Trans. Systems, Man, and Cybernetics, vol. 2, no. 3, pp. 408-421, 1972.

[2] I. Tomek, “Two Modifications of CNN,” IEEE Trans. Systems, Man, and Cybernetics, vol. 6, no. 11, pp. 769-772, 1976.

[3] J. Zhang and I. Mani, “kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction,” Proc. of ICML ’03, Workshop on Learning from Imbalanced Datasets II, pp. 1-7, 2003.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic Minority Over-sampling Technique,” J. of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[5] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” Proc. of ICIC ’05, pp. 878-887, 2005.

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” Proc. of IJCNN ’08, pp. 1322-1328, 2008.

[7] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem,” Proc. of PAKDD ’09, pp. 475-482, 2009.

[8] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” Proc. of ICICS ’20, pp. 243-248, 2020.

[9] H. Shamsudin, U. K. Yusof, A. Jayalakshmi, and M. N. Akmal Khalid, “Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset,” Proc. of ICCA ’20, pp. 803-808, 2020.

[10] S. Choirunnisa and J. Lianto, “Hybrid Method of Undersampling and Oversampling for Handling Imbalanced Data,” Proc. of ISRITI ’18, pp.276-280, 2018.

[11] C. Elkan, “The Foundations of Cost-Sensitive Learning,” Proc. of IJCAI ’01, pp. 1-6, 2001.

[12] B. Schölkopf, J. C. Platt, J. S. Taylor, A. J. Smola, and R. C. Williamson, “Estimating the Support of a High-Dimensional Distribution,” Neural Computation 2001; vol. 13, no. 7, pp. 1443-1471, 2001.

[13] H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Trans. on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[14] M. Lv, Y. Ren, and Y. Chen, “Research on imbalanced data: Based on SMOTE-AdaBoost algorithm,” Proc. of EITCE ’19, pp. 1165-1170, 2019.

[15] F. Tari, A. A. Ozok, and S. H. Holden, “A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords,” Proc. of SOUPS ’06, pp. 56-66, 2006.

[16] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith, “Smudge attacks on smartphone touch screens,” Proc. of WOOT ’10, pp. 1-10, 2010.

[17] T. Ogane and I. Echizen, “BiometricJammer: Use of pseudo fingerprint to prevent fingerprint extraction from camera images without inconveniencing users,” Proc. of SMC ’18, pp. 2825-2831, 2018.

[18] O. D. Incel, S. Gunay, Y. Akan, Y. Barlas, O. E. Basar, G. I. Alptekin, and M. Isbilen, “Dakota: Sensor and touch screen-based continuous authentication on a mobile banking application,” IEEE Access, vol. 9, pp. 38943-38960, 2021.

[19] M. Agrawal, P. Mehrotra, R. Kumar, and R. R. Shah, “Defending touch-based continuous authentication systems from active adversaries using generative adversarial networks,” Proc. of IJCB ’21, pp. 1-8, 2021.

[20] M. Agrawal, P. Mehrotra, R. Kumar, and R. R. Shah, “GANTouch: An attack-resilient framework for touch-based continuous authentication system,” IEEE Trans. on Biometrics, Behavior, and Identity Science, vol. 4, no. 4, pp. 533-543, 2022.

[21] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, “Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication,” IEEE Trans. on Information Forensics and Security, vol. 8, no. 1, pp. 136-148, 2013.

[22] A. Z. Zaidi, C. Y. Chong, R. Parthiban, and A. S. Sadiq, “A framework of dynamic selection method for user classification in touch-based continuous mobile device authentication,” J. of Information Security and Applications, vol. 67, no. 103217, pp. 1-19, 2022.

[23] 工藤雅士，高橋翼，山名早人，“eCommTouch : EC サイトを対象としたタッチベース認証のための新たなデータセットの提案”，DEIM2024, T4-A-8-05, 2024.

[24] Z. Sitova, J. Sedenka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K. S. Balagani, “HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users,” IEEE Trans. on Information Forensics and Security, vol. 11, no. 5, pp. 877-892, 2016.

[25] M. Antal, Z. Bokor, and L. Z. Szabo, "Information revealed from scrolling interactions on mobile devices," Pattern Recognition Letters, vol. 56, pp. 7-13, 2015.

[26] M. Georgiev, S. Eberz, H. Turner, G. Lovisotto, and I. Martinovic, “Common Evaluation Pitfalls in Touch-Based Authentication Systems,” Proc. of Asia CCS ’22, pp. 1049-1063, 2022.

[27] K. V. Mardia and P. E. Jupp, “Directional Statistics,” Wiley Series in Probability and Statistics, Wiley, 1999.

[28] 古川大樹，宮﨑公彦，工藤雅士，山名早人，“タッチストロークを用いた継続認証におけるトレンドと問題点に関するサーベイ”，CSS2023, 4E2-2, pp. 1-8, 2023.

[29] M. Georgiev, S. Eberz and I. Martinovic, “Techniques for Continuous Touch-Based Authentication,” Proc. of ISPEC ’22, pp. 409-431, 2022.

eCommTouch：EC サイトを対象としたタッチベース認証のための新たなデータセットの提案

工藤 雅士† 高橋 翼‡ 山名 早人§

†早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
‡LINE ヤフー株式会社 〒102-8282 東京都千代田区紀尾井町 1-3 東京ガーデンテラス紀尾井町紀尾井タワー
§早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: †§ {kudoma34, yamana}@yama.info.waseda.ac.jp, ‡ tsubasa.takahashi@lycorp.co.jp

あらまし 近年，スマートフォンを使用した EC サイトの利用が拡大している．EC サイトを含め，スマートフォンで個人情報を扱う場面が増加しており，ログイン認証後にユーザのタッチ特性を利用したタッチベース認証の導入が検討されている．これまでに様々なタッチベース認証のためのデータセットが提案されてきたが，従来のデータセットは EC サイトとは特徴が異なる利用場面が想定されているため，EC サイトを再現した評価が困難であった．そこで本研究では，EC サイトの操作を再現したデータセットとして eCommTouch を提案する．105 ユーザのデータを含む本データセットを使用した評価実験では，eCommTouch のベンチマークを提供するとともに，これまで検証が困難であったストローク方向別に分類器を構築する手法を検証し，EER 低下の可能性を確認した．

キーワード タッチベース認証，バイオメトリクス，スマートフォン，データセット

1. はじめに

近年，スマートフォンの爆発的な普及に伴い，スマートフォンの利用場面が増加した．中でも EC（e-commerce）サイトの利用は年々増加傾向にあり，2022 年のモバイルデバイスの利用用途トップ 10 にランクインしている[1]．さらに，2023 年における EC サイトへのアクセスは 70%以上がスマートフォンであると報告されている[2]．EC サイトでは，家電や服など様々な商品を手軽に購入することができる．一方で，EC サイトを利用する場合，クレジットカード番号や住所などの個人情報を扱う必要がある．スマートフォンには個人情報を保護するために指紋認証や顔認証といったログイン認証が標準的に導入されているが，ログイン認証は突破されるリスク[3][4]が報告されている．

タッチベース認証は，スマートフォンログイン後のパッシブ認証として近年導入が検討されている認証方式である．タッチベース認証では，スマートフォンの操作を継続的に監視し，座標や圧力，速度などのタッチ特性に基づいて認証を行う．タッチベース認証はユーザに追加の操作を要求しないため，スマートフォンとの親和性が高い認証方式としてこれまでに様々な手法が提案されている[5][6][7][8][9][10][11][12][13]．また，タッチベース認証の認証性能向上のために，認証手法の提案と共にデータセットの公開も行われている．

2024 年現在，タッチベース認証で利用可能な公開データセット[5][6][14][15][16]では，テキストを読む場面[5][6][14][15]や SNS（Social Networking Service）を利用する場面[16]などを対象にデータ収集が実施されている．このような利用場面は，1) ユーザ毎にストローク数に大きなばらつきが生じる，2) 特定のストローク方向のデータ数が極端に少ないという特徴を持っている．スマートフォンで利用される EC サイトは，商品がリスト形式で閲覧できる画面設計となっており，ユーザは複数の商品を比較しながら商品の選択を行う．そのため，EC サイトはテキストを読む場面や SNS を利用する場面と比較して，様々なストローク方向で連続したストローク操作が多く発生する．したがって，既存のデータセットでは EC サイトの利用場面を再現したタッチベース認証の評価を行うことができない．

こうした背景から，本稿ではタッチベース認証で利用可能な新たなデータセットとして eCommTouch（e-commerce touch）を公開する．既存のデータセットと eCommTouch の詳細を表 1 にまとめる．本データセットは，EC サイトを再現したアプリケーションから収集されたタッチデータで構成されており，既存のデータセットと比較して次のような特徴を持っている．

- セッション（連続したストローク操作のグループ）毎のデータ数が多く，セッション間でデータ数のばらつきが少ない
- 垂直（水平）ストロークで，上方向（左方向）と（右方向）下方向のデータ数の偏りが少ない

eCommTouch を公開することで，これまで利用場面として想定されていなかった EC サイトにおけるタッチベース認証の評価を可能にする．また，本データセットを使用した評価実験を実施し，eCommTouch のベンチマークを提供するとともに，既存のデータセットでは検証が困難であったストローク方向毎に分類器を構築する手法[7][8]の有用性を検証する．

表 1　タッチベース認証で利用可能な公開データセット

年	データセット	OS	ユーザ数	利用場面	操作方向	セッション内のストローク数	
						平均	標準偏差
2013	Touchalytics[1] [5]	Android	41	Image comparison	水平	89.2	33.4
				Text reading	垂直	92.7	61.4
2014	HMOG[2] [6][14]	Android	100	Text reading	垂直/水平	51.4	35.6
				Text typing	-	-	-
				Map navigation	垂直/水平	186.2	81.8
2015	BioIdent[3] [15]	Android	71	Text reading	垂直	44.5	33.4
				Image gallery	水平	177.3	92.8
2022	CEP[4] [16]	iOS	470	Social media	垂直	70.2	36.1
				Image gallery	水平	123.8	27.5
2024	**eCommTouch[5] (ours)**	**iOS**	**105**	**EC website**	**垂直**	**210.8**	**16.4**
				EC website	**水平**	**207.6**	**11.8**

本稿は次の構成をとる．2 節で既存の公開データセットを説明し，3 節で eCommTouch の詳細を述べる．4 節で評価実験について説明し．5 節では eCommTouch の限界について述べ，6 節で本稿をまとめる．

2. 既存の公開データセット

本節では，2024 年現在もタッチベース認証で利用可能な既存の公開データセットについて説明する．

2.1. Touchalytics

2013 年，Frank らはタッチベース認証に利用可能なデータセットとして Touchalytics を公開した[5]．Touchalytics は 41 人の大学生から取得したタッチデータを含み，2 つのタスクを実施した際のタッチデータで構成される．1 つ目のタスク（image comparison）では，画面に表示された 2 枚の画像の違いを見つける操作を行い，水平方向のストロークを収集した．2 つ目のタスク（text reading）では，画面に表示された Wikipedia の記事を読む操作を行い，記事をスクロールする際の垂直方向のストロークを収集した．Touchalytics を使用した評価実験では，タッチデータから抽出した 27 種類のタッチ特徴量を使用して k-NN および SVM で分類器を構築し，EER 0–4%を達成した．

2.2. HMOG

2014 年，Yang らは HMOG と名付けられたマルチモーダルなデータセットを公開した[6][14]．HMOG は Android デバイスを使用して 100 ユーザから取得したタッチセンサ，加速度センサ，ジャイロセンサ，地磁気センサのデータで構成される．HMOG を構築するために，各ユーザは 1) text reading，2) text typing，3) map navigation の 3 つのタスクを実施した．なお，text typing のタスクではストロークデータは収集されていない．本データセットを使用した評価実験では，11 のタッチ特徴量と他のセンサデータを組み合わせたタッチベース認証が検証された．評価の結果，分類器に SVM を採用した場合に EER 7–10%を達成した．

2.3. BioIdent

2015 年，Antal らは 71 人のユーザから取得したタッチデータを BioIdent と名付けて公開した[15]．BioIdent は Android デバイスに実装された 2 つのタスク（text reading と image gallery）を実施した際のタッチデータで構成される．本データセットは，画面サイズの異なるタブレットとスマートフォンの両方のデータが混在している．Antal らはスマートフォンユーザの特性（性別やスマートフォンの習熟度など）を分析するために本データセットを構築したため，タッチベースの認証に関する評価は実施されていない．

2.4. CEP

2022 年，Georgiev らはクラウドソーシングを通じて 470 人のユーザから取得したタッチセンサ，加速度センサ，地磁気センサのデータを含むデータセット（CEP）を公開した[16]．データの収集は，2 つのタスク（social media と image gallery）を iOS アプリケーションとして実装し，圧力センサを搭載した 9 種類の iOS デバイスを対象に実施した．本データセットは，タッチベースの認証における性能評価のベストプラクティスを構築するために使用された．

3. eCommTouch データセット

現在，タッチベース認証で利用可能な公開データセットは 4 種類に限定されており，これらのデータセットではカバーできない新たな利用場面が発生している．特に EC サイトの利用は年々増加傾向にあり，既存の公開データセットではこの重要な利用場面をカバーできていないことが大きな課題となっている．本稿では，EC サイトを対象とした新たなデータセットである eCommTouch を公開することにより，タッチベース認

1 https://www.mariofrank.net/touchalytics/index.html
2 https://www.cs.wm.edu/~qyang/hmog.html
3 https://www.ms.sapientia.ro/~manyi/bioident.html
4 https://github.com/ssloxford/evaluation-pitfalls-touch
5 https://github.com/yamanalab/eCommTouch

証の更なる精度向上と実用化の推進を目指す.

また，eCommTouch は既存の公開データセットと比較して，1 セッションあたりのストローク数が多く，上下および左右でストローク数の偏りが小さいという特徴を持っている（3.4 項で詳述）．そのため，eCommTouch の公開により，ストローク方向別に分類器を構築する手法など，既存の公開データセットでは困難であった新たな認証手法に関する議論が今後展開されることを期待する.

3.1. 提供データ詳細

eCommTouch には，タッチベース認証で広く使用されているタッチ特徴量[13]が抽出可能なセンサデータが含まれている.提供データの詳細を表 2 にまとめる.

表 2　eCommTouch のデータ詳細

属性	説明
USER_ID	ユーザ ID
SESSION_ID	セッション番号
VER_HOR	方向フラグ
GAME_NUM	ゲーム番号
YMD	実験日付
STROKE_NUM	ストローク番号
DATA_POINTS	ストローク中のデータポイント数
POINTS_IDX	データポイントのインデックス
ACTION	タッチアクション
TIMESTAMP	タイムスタンプ
STROKE_IDX	ストロークのインデックス
X_COOR	x 座標 [px]
Y_COOR	y 座標 [px]
FORCE	タッチ圧力
X_RADIUS	接触楕円の水平半径 [px]
Y_RADIUS	接触楕円の垂直半径 [px]

3.2. データ収集

3.2.1. デバイス

データ収集には，iPhone XS と iPhone 8 Plus を使用した．これらのデバイスは 2024 年現在，それぞれの画面サイズにおいて正確な圧力が取得可能な圧力センサを搭載した最新の機種であるため採用した．iPhone XS と iPhone 8 Plus の仕様を表 3 にまとめる.

表 3　デバイスの仕様

デバイス	画面サイズ [inch]	解像度 [px]	ピクセル密度 [ppi]	周波数 [Hz]	OS
XS	5.8	2436x1125	458	60	14.2
8 Plus	5.5	1920x1080	401	60	13.6

3.2.2. アプリケーション

eCommTouch は，EC サイトで商品を購入する場面を再現したデータセットである．本データセットでは，スマートフォンからアクセスされる EC サイトは，次のような特徴を持つことを前提としている.

- 商品が縦並びまたは横並びで配置されている
- 各商品の情報（商品名，画像，価格，星の数など）は，限られた範囲内に集約されている

また，EC サイトで商品の購入を行うユーザは，次のような特徴を持つことを前提としている.

- 欲しい商品について，抽象的な基準（値段が安い，評価が高い，お得感があるなど）を持つ
- 複数の商品を比較して購入する商品を決める

EC サイトおよびユーザに関するこれらの特徴に基づいて，タッチデータ収集のための独自アプリケーションの設計を行った．本アプリケーションは，EC サイトにおける売り上げのトップシェアを占めるファッションカテゴリを対象に画面設計を行った[2]．本アプリケーションの動作の流れを以下に示す.

1. アプリ：商品条件を表示する
2. ユーザ：商品条件を確認し，「OK」をタップする
3. アプリ：10 個の商品で構成される商品リスト（縦並びと横並びの 2 パターンが存在）を表示する
4. ユーザ：商品リストをスワイプして 10 個の商品を閲覧し，条件に合う商品を 1 つ選択する
5. アプリ：商品条件に一致する場合は「Correct」，一致しない場合は「Incorrect」を表示する
6. ステップ 1〜5 を 1 ゲームとし，規定数のタッチデータが収集されるまでゲームを繰り返す

画面に表示される商品には，1) 商品画像，2) 商品名，3) 星の数，4) レビュー数，5) 価格，6) 割引率の 6 つの情報が含まれている．商品条件は商品リストに含まれる 10 個の商品を比較することによって一意に決まるように設定されている．また，商品条件は 90 通り用意されており，ゲームが切り替わる度に決まった順番で新たな商品条件が表示される．データ収集アプリケーションの例を図 1 に示す.



図 1　データ収集アプリケーションの例

3.2.3. データ収集手順

データ収集実験は，本大学の学生を対象に 3 期間に分けて実施し，各期間で複数回実験を行った．実験への参加は各期間とも一人一回に制限されたが，期間を跨いでの複数回の参加は可能とした．本実験は各回 1 時間で構成され，以下の流れで実施した．

- データ収集用アプリケーションをインストールしたデバイスの配布
- 実験手順の説明（10 分）
- アプリケーションの操作練習（5 分）
- アプリケーションの実施（40 分）
- 事件に関するアンケートの実施（5 分）

本実験では,データ収集用アプリケーションにおける，商品リストをスワイプする操作（3.2.2 項，step.4）をデータ収集の対象とした．同一の条件でデータ収集を行うために，実験参加者には座った姿勢で右手の親指でアプリケーションを操作するように指示を出した．指が画面に触れてから離れるまでのスワイプ操作を 1 ストロークとし，200 ストローク収集されるまで参加者はゲームを繰り返した．なお，200 ストローク以上収集されるまで行われた一連のストローク操作を，本データセットにおいて 1 セッションとした．タッチデータは 1 ストロークの間に最大周波数 60Hz で複数回取得され，取得されたデータは JSON 形式でウェブサーバに送信された後，CSV 形式でサーバ内に保存された．実験参加者はセッション間で 1 分以上の休憩を取り，6 セッションを終えた時点で実験終了とした．商品リストの配置は，垂直方向および水平方向のストロークを取得するために，セッション毎に縦並びと横並びを交互に変化させた．デバイス別のユーザ数の内訳を表 4 にまとめる．

表 4　デバイス別のユーザ数

デバイス	ユーザ総数	各ユーザの実験参加回数		
		1 回	2 回	3 回
XS	78	50	22	6
8 Plus	27	24	2	1

3.3. データ整理

eCommTouch に含まれるタッチデータは，3.2 項で詳述したデータ収集で取得されたデータに対して，タップデータおよび異常データの除去を行ったものである．1 ストローク内のデータポイントが 5 つ未満，またはストロークの移動が 5px 未満のデータはタップデータとして削除した[16]．また，ストローク時間が 3,000ms 以上のデータと，ストローク間隔時間が 0ms のデータは異常データとして削除した．さらに，ストローク間隔時間が 10,000ms 以上のデータは，セッション間の移行データとして削除した．これらの処理により，一部セッションでデータ数の減少が発生する．

3.4. eCommTouch の特徴分析

3.4.1. ストロークデータの統計分析

eCommTouch を構成するタッチデータを，ストローク毎に統計的に分析した結果を表 5 に示す．本データセットに含まれるストロークは，水平方向よりも垂直方向の軌跡が長く，ストローク間隔時間も垂直方向が長かった．一方で，ストローク時間は垂直方向と水平方向は共に約 100ms であった．各デバイスを操作している際のストローク軌跡を，セッション毎に可視化した例を図 2 と図 3 に示す．

3.4.2. セッションデータの分析

eCommTouch と既存の公開データセットについて，各セッションに含まれるストローク数をヒストグラムにまとめたものを図 4 に示す．図 4 より，eCommTouch に含まれる全てのセッションは，最低 150 ストローク以上のデータを含んでいる．また，表 1 より，既存のデータセットと比較して，本データセットはセッションに含まれるストローク数のばらつきが少ない．

3.4.3. ストローク方向の分析

eCommTouch と既存の公開データセットについて，垂直セッションおよび水平セッションに含まれるストローク数をヒストグラムにまとめたものをそれぞれ図 5，図 6 に示す．図 5 および図 6 より，既存の公開データセットでは，上方向（左方向）か下方向（右方向）のいずれかしか含まれないセッションが存在した．これに対して，eCommTouch の各セッションには，上方向（左方向）および下方向（右方向）のストロークが少なくとも 30 ストローク以上含まれている．

表 5　eCommTouch に含まれるストロークデータの統計情報

デバイス	ストローク方向	ストローク時間 [ms]		間隔時間 [ms]		軌跡の長さ [px]		端点間距離 [px]	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
XS	垂直	107.7	122.7	717.2	1002.5	397.4	128.1	409.2	138.8
	水平	104.0	90.9	632.0	949.2	269.6	92.1	275.2	96.4
8 Plus	垂直	100.7	110.9	777.8	1108.1	388.9	144.7	400.3	155.8
	水平	95.0	88.9	679.1	1023.0	288.5	109.1	294.9	115.0

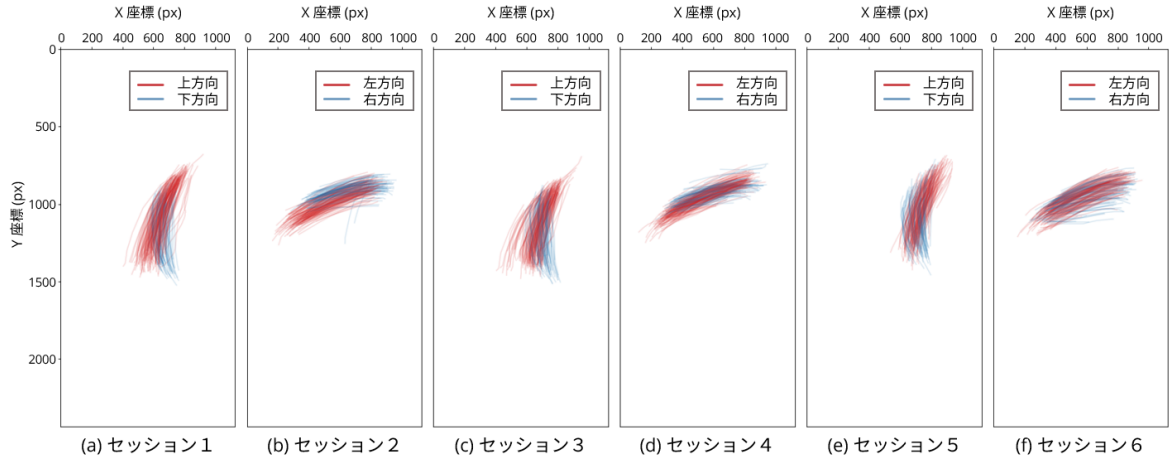


図 2　iPhone XS のストローク軌跡の例（USER_ID: u231）

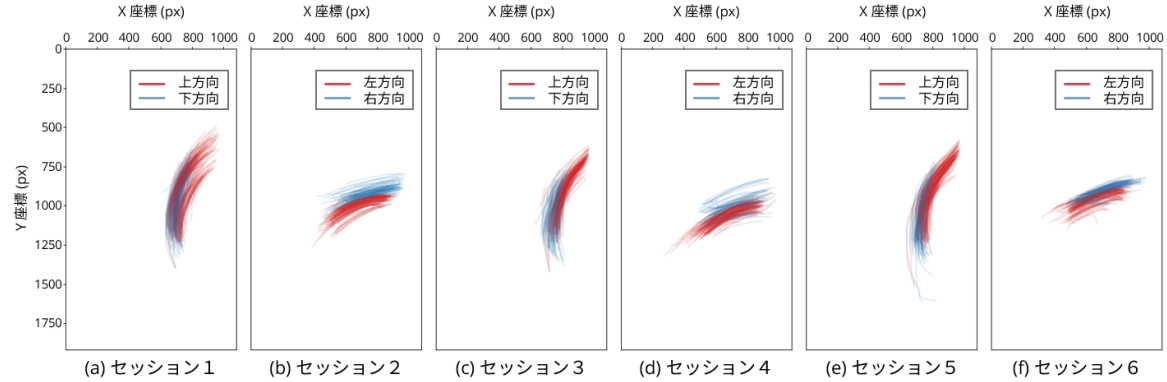


図 3　iPhone 8 Plus のストローク軌跡の例（USER_ID: u152）

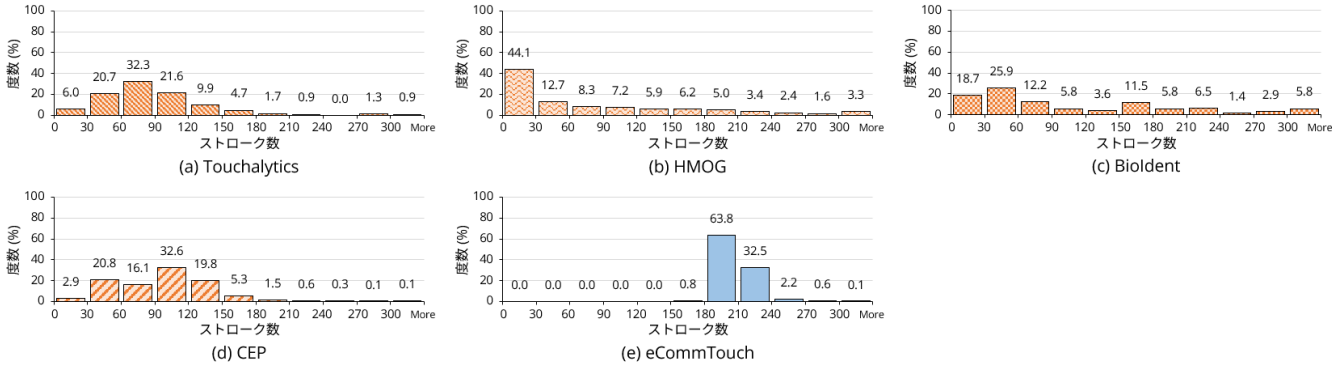


図 4　eCommTouch と公開データセットにおけるセッション毎のストローク数の分布

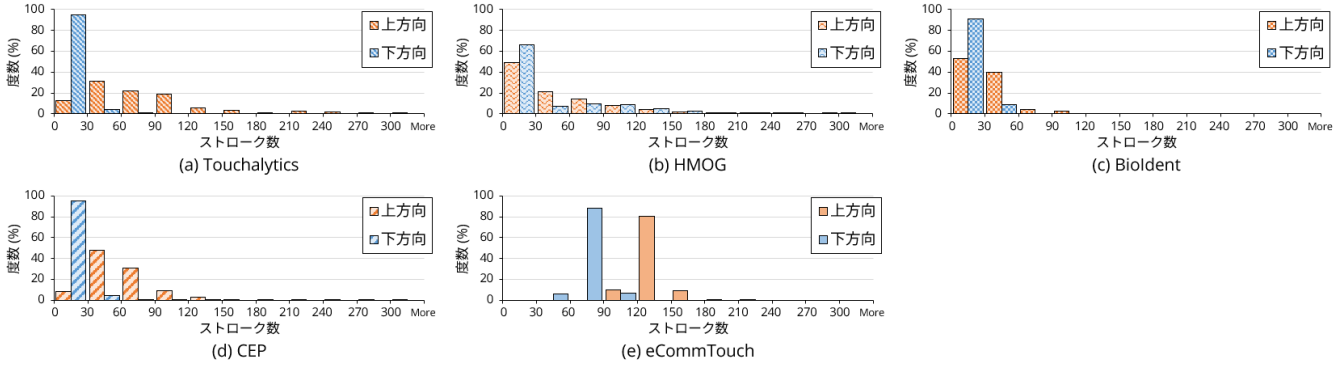


図 5　eCommTouch と公開データセットにおける垂直セッション毎のストローク数の分布

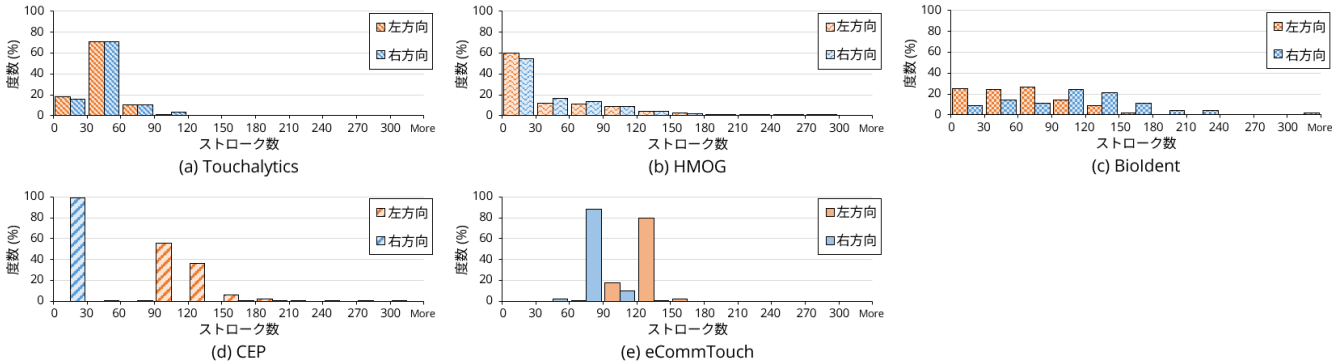


図 6　eCommTouch と公開データセットにおける水平セッション毎のストローク数の分布

4. 評価実験

本節では，eCommTouch を使用して実施した 2 つの評価実験について説明する．第一の実験では，垂直および水平方向のシナリオにおいて，単一の分類器を構築する手法を検証する．本手法は，タッチベース認証において一般的に採用されている分類器構築手法[5][6][10][11][12][13]であり，eCommTouch におけるベンチマークを提供する．第二の実験では，垂直および水平方向のシナリオそれぞれにおいて，ストローク方向（上下または左右）毎に 2 つの分類器を構築する手法を検証する．本手法は，既存の公開データセットではストローク数の制約により評価が困難であるため，先行研究が限られている[7][8]. 第一の評価実験の結果と比較することにより，本手法の有用性を確認する．

両実験では，タッチベース認証[9][10][16][17]の評価において一般的に採用されている，Touchalytics [5]で提案された 27 のタッチ特徴量を使用する．

4.1. 実験設計

4.1.1. 認証シナリオ

タッチベース認証の性能は様々な外的要因によってブレが生じる[11][12][16]．そのため，デバイス，セッション，ストローク方向の違いで 8 つの認証シナリオを構築した．各認証シナリオの詳細を表 6 に示す．

表 6　認証シナリオ

デバイス	シナリオ	説明
XS	$XS_{v\text{-}intra}$	セッション内垂直（XS）
	$XS_{h\text{-}intra}$	セッション内水平（XS）
	$XS_{v\text{-}inter}$	セッション間垂直（XS）
	$XS_{h\text{-}inter}$	セッション間水平（XS）
8 Plus	$8P_{v\text{-}intra}$	セッション内垂直（8 Plus）
	$8P_{h\text{-}intra}$	セッション内水平（8 Plus）
	$8P_{v\text{-}inter}$	セッション間垂直（8 Plus）
	$8P_{h\text{-}inter}$	セッション間水平（8 Plus）

4.1.2. 分類モデル

本稿の評価実験では，タッチベース認証の包括的なベースライン評価を行った Georgiev らの研究[13]において，単一の分類器で性能が良いとされた RF（Random Forest），SVM（Support Vector Machine），NN（Neural Network）を分類モデルとして採用した．評価実験は Python 3.7.8 を使用して実施し，scikit-learn 0.24.0 と PyTorch 1.7.0 のパッケージを使用して分類モデルの構築を行った．各分類器のパラメータは Georgiev らの評価実験[13]で採用された値を使用し，値の指定がないパラメータについてはデフォルトの値を使用した．

4.1.3. 評価データセットの構築

表 6 に示す認証シナリオを再現するために，デバイスとストロークの方向を条件に eCommTouch からデータを抽出した．なお，データ収集実験への参加回数によるバイアスを無くすために，各ユーザの初回参加時のデータのみを抽出対象とした．評価データセットの構築は著者らの先行研究[17]をもとに，ユーザおよびセッションの選択を改良した以下の手順で実施した．

- **ユーザの選択**：抽出したデータに含まれるユーザの中から正例ユーザを 1 名選択する．その後，正例ユーザを除くユーザの中から，テスト用負例ユーザをランダムに 4 ユーザ選択する．残りのユーザを訓練用負例ユーザとして選択する．
- **セッションの選択**：各ユーザは，垂直方向と水平方向のストロークデータをそれぞれ 3 セッション分持っている．垂直方向のシナリオでは垂直方向の 3 セッションを選択し，水平方向のシナリオでは水平方向の 3 セッションを選択する．
- **訓練・テストデータセットの構築（正例）**：セッション間シナリオでは，最も古い 2 セッションが正例訓練データセット（PD_{train}）の構築に使用され，最も新しいセッションが正例テストデータセット（PD_{test}）の構築に使用される．セッション内シナリオでは，3 セッションからランダムに 1 セッションを選択し，選択されたセッションから，最も古い 3 分の 2 のデータを抽出し，正例訓練データセット（PD_{train}）を構築する．選択されたセッションの残りのデータで正例テス

トデータセット（PD_{test}）を構築する．

- **訓練データセットの構築（負例）**：セッションに関係なく，各訓練用負例ユーザから同数のストロークをランダムに抽出する．抽出されたストロークを集約し，PD_{train} と同数の負例訓練データセット（ND_{train}）を構築する．
- **テストデータセットの構築（負例）**：セッションに関係なく，テスト用負例ユーザから同数のストロークをランダムに抽出する．抽出されたストロークを集約し，PD_{test} と同数の負例テストデータセット（ND_{test}）を構築する．

4.1.4. 訓練と評価

分類器の訓練およびテストは min-max 正規化を適用した PD_{train}，ND_{train}，PD_{test}，ND_{test} を使用する．

第一の評価実験では，ストローク方向を区別せずに単一の分類器の訓練およびテストを実施する．

第二の評価実験では，ストローク方向毎に分類器の訓練を行い，テストでは入力データのストローク方向を判別し，対応する分類器を選択してユーザ判定を行う．なお，ストローク方向の判別は，ストローク方向を表すタッチ特徴量を使用して行う．

分類結果の集約にはスライディングウィンドウを採用し，最終的な分類結果はスライディングウィンドウ内の値を平均することで算出する．分類性能は，タッチベース認証で一般的に採用されている EER（Equal Error Rate）に基づいて評価を行う．タッチ特性はユーザ毎に異なるため，分類性能は選出するユーザの組み合わせに依存する．そのため，分類性能は正例ユーザを選出する全パターンにおいて EER を算出し，その平均値を報告する．また，スライディングウィンドウでは，ウィンドウサイズ 1 と 10 の場合を評価する．

4.2. 評価結果

ウィンドウサイズ 1 と 10 の場合において，単一の分類器を構築する手法およびストローク方向毎に分類器を構築する手法の評価結果を表 7 および表 8 に示す．

表 7 および表 8 の EER マクロ平均を比較すると，NN を除き，方向別の分類器の EER は単一の分類器よりも低くなる傾向にある．また，XS を用いたシナリオでは，方向別の分類器が単一の分類器よりも性能が良い傾向にあることが確認された．単一の分類器と方向別の分類器の差は，ウィンドウサイズ 1 よりもウィンドウサイズ 10 の方が明確になり，ウィンドウサイズ 10 での評価では，24 の検証のうち，17 の検証で方向別の分類器の平均 EER の低下が確認された．一方で，性能のばらつきが大きいため，統計的な有意性が確認できない検証が多く見られた．正例ユーザ毎に評価結果を分析したところ，特定のユーザを正例として設定した検証で EER が著しく高くなる傾向が見られた[17]．このようなユーザの存在が，統計的検定による分類性能の比較を困難にしていると考えられる．

表 7　スライディングウィンドウサイズ 1 における EER（%）（カッコ内は標準偏差）

シナリオ	単一分類器の EER（%）						ストローク方向別分類器の EER（%）					
	SVM		RF		NN		SVM		RF		NN	
$XS_{v\text{-}intra}$	22.13	(11.01)	11.41	(7.01)	**11.17**	**(6.76)**	**18.65***	**(9.22)**	**11.10**	**(6.78)**	11.64	(6.73)
$XS_{h\text{-}intra}$	22.15	(8.86)	11.72	(7.57)	11.86	(6.57)	**21.48**	**(8.70)**	**11.47**	**(6.98)**	**11.84**	**(6.38)**
$XS_{v\text{-}inter}$	19.04	(11.30)	**13.75**	**(10.34)**	**14.69**	**(9.80)**	**18.48***	**(10.86)**	13.78	(10.67)	15.33	(9.82)
$XS_{h\text{-}inter}$	**20.43**	**(11.69)**	14.80	(11.20)	**15.33**	**(10.04)**	20.72	(11.46)	**14.42**	**(10.67)**	15.79	(9.55)
$8P_{v\text{-}intra}$	**19.34**	**(9.55)**	**8.87**	**(5.35)**	**9.45**	**(5.47)**	19.82	(10.08)	8.87	(5.33)	9.65	(5.03)
$8P_{h\text{-}intra}$	**17.19**	**(8.53)**	9.93	(6.64)	**10.05***	**(6.76)**	17.76	(8.70)	**9.38**	**(5.82)**	10.84	(6.35)
$8P_{v\text{-}inter}$	**15.06***	**(7.92)**	**9.70**	**(6.38)**	**8.95***	**(6.21)**	16.04	(8.20)	10.29	(6.95)	10.56	(6.48)
$8P_{h\text{-}inter}$	17.21	(11.47)	**11.75**	**(9.36)**	**12.47***	**(9.63)**	**17.16**	**(11.29)**	11.76	(8.49)	13.56	(10.00)
マクロ平均	19.07		11.49		**11.75**		**18.76**		**11.38**		12.40	

太字は 2 手法のうち EER が低い手法，網掛けは更に有意差が生じた手法を示す（*: $p<0.05$，対応のある t 検定）

表 8　スライディングウィンドウサイズ 10 における EER（%）（カッコ内は標準偏差）

シナリオ	単一分類器の EER（%）						ストローク方向別分類器の EER（%）					
	SVM		RF		NN		SVM		RF		NN	
$XS_{v\text{-}intra}$	9.73	(9.76)	2.74	(4.58)	2.70	(4.69)	**6.43***	**(7.22)**	**2.70**	**(4.50)**	**2.43**	**(4.46)**
$XS_{h\text{-}intra}$	8.59	(8.35)	3.37	(6.89)	2.02	(4.53)	**8.31**	**(8.46)**	**2.70**	**(5.53)**	**1.74**	**(4.13)**
$XS_{v\text{-}inter}$	9.72	(11.68)	**5.89**	**(9.31)**	6.16	(10.80)	**9.15**	**(11.59)**	6.01	(10.29)	**5.97**	**(10.58)**
$XS_{h\text{-}inter}$	10.68	(14.12)	7.22	(13.17)	5.62	(11.41)	**10.48**	**(13.88)**	**6.68**	**(12.56)**	**5.28**	**(10.93)**
$8P_{v\text{-}intra}$	**7.94**	**(8.68)**	1.27	(2.45)	1.46	(2.76)	8.87	(10.16)	**1.15**	**(3.02)**	**1.22**	**(2.30)**
$8P_{h\text{-}intra}$	6.03	(7.11)	3.42	(6.37)	**1.85**	**(3.35)**	**5.60**	**(6.62)**	**2.30***	**(5.15)**	1.86	(3.67)
$8P_{v\text{-}inter}$	**5.99**	**(7.68)**	**2.36**	**(4.90)**	**1.71**	**(3.55)**	6.43	(8.49)	2.55	(5.01)	2.84	(4.56)
$8P_{h\text{-}inter}$	7.75	(12.14)	5.21	(10.90)	**4.96**	**(11.26)**	**7.66**	**(12.28)**	**4.36***	**(9.36)**	5.01	(12.53)
マクロ平均	8.26		3.93		3.31		**7.87**		**3.56**		**3.29**	

太字は 2 手法のうち EER が低い手法，網掛けは更に有意差が生じた手法を示す（*: $p<0.05$，対応のある t 検定）

以上の結果から，方向別に分類器を構築する手法は認証シナリオと分類モデルに依存して性能のばらつきが見られる一方で，EER を低下させる可能性があることを確認した．本アプローチは既存のデータセットでは検証が困難であり，先行研究が限られている[7][8]．そのため，今後のさらなる発展が期待される．

5. eCommTouch の限界

本稿で提案した eCommTouch は特定の利用場面，すなわちECサイト利用時のタッチ操作を再現している．スマートフォンは多様な用途で使用されるため，タッチベースの認証の評価を行う場合は eCommTouch を含め，多くのデータセットを使用して多角的な観点から評価を行うことが望まれる．また，eCommTouch に含まれるデータはユーザに対して座った状態で右手の親指を使うように指示をして収集したデータである．そのため，ユーザが自然にデバイスを操作している場面と比較して，潜在的なバイアスを含んでいる可能性がある点に注意が必要である．

6. まとめ

本稿では，タッチベースの認証に利用可能な新たな公開データセットとして，eCommTouch を提案した．eCommTouch は EC サイトで一般的に見られるタッチ操作を再現しており，上下方向および左右方向のストロークを含んでいる．既存のデータセットと比較して，eCommTouch は次のような特徴を持っている：1）セッションあたりのストローク数が多く，ばらつきが少ない，2）上方向（左方向）と下方向（右方向）のストローク数の偏りが少ない．eCommTouch を使用した評価実験では，eCommTouch のベンチマークを提供するとともに，データの制約から既存のデータセットでは困難であった，ストローク方向を区別した分類器を検証した．評価の結果，ストローク方向別に分類器を構築することで，EER を低減できる可能性を確認した．さらに，RF を使用した方向別分類器は，平均 EER が 1.15%であり，本評価実験で最高性能を達成した．我々は，eCommTouch がタッチベース認証手法，特に最近の EC サイトにおけるタッチ操作に焦点を当てた認証手法の進歩に貢献することを期待している．

参 考 文 献

[1] K. Fu, “Top 10 smartphone uses: New consumer report reveals why we’re at the point of no return [EmpowerQ]”, Qualcomm, 2023, https://www.qualcomm.com/news/onq/2023/04/top-10-smartphone-uses-new-consumer-report-reveals-why-were-at-the-point-of-no-return (accessed Jan. 5, 2024).

[2] K.V. Gelder, “E-commerce worldwide - statistics & facts”, statista, 2023, https://www.statista.com/topics/871/online-shopping (accessed Jan. 5, 2024).

[3] I. Echizen, and T. Ogane, “BiometricJammer: Method to Prevent Acquisition of 40 Biometric Information by Surreptitious Photography on Fingerprints”, IEICE Trans. Inf. Syst., Vol. E101D (1), pp. 2–12, 2018.

[4] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial Attacks Against Face Recognition: A Comprehensive Study", IEEE Access, Vol. 9, pp. 92735–92756, 2021.

[5] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, “Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication”, IEEE Trans. Inform. Forensic Secur., Vol. 8 (1), pp. 136–148, 2013.

[6] Z. Sitova, J. Sedenka, Q. Yang, G. Peng, G. Zhou, P. Gasti, and K.S. Balagani, “HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users”, IEEE Trans. Inform. Forensic Secur., Vol. 11 (5), pp. 877–892, 2016.

[7] M. Kudo and H. Yamana, “Active Authentication on Smartphone using Touch Pressure”, Adjunct Proc. of the 31st Annual ACM Symp. on User Interface Softw. and Technol., pp. 96–98, 2018.

[8] J. Chao, M.S. Hossain, and L. Lancor, “Swipe gestures for user authentication in smartphones”, J. of Inf. Secur. and Appl., Vol. 74, pp. 1–13, 2023.

[9] A.Z. Zaidi, C.Y. Chong, R. Parthiban, and A.S. Sadiq, “A framework of dynamic selection method for user classification in touch-based continuous mobile device authentication”, J. of Inf. Secur. and Appl., Vol. 67, pp. 1–19, 2022.

[10] I. Chang, C. Low, S. Choi, and A.B. Teoh, “Kernel Deep Regression Network for Touch-Stroke Dynamics Authentication”, IEEE Signal Process. Lett., Vol. 25 (7), pp. 1109–1113, 2018.

[11] C. Shen, Y. Li, Y. Chen, X. Guan, and R.A. Maxion, “Performance Analysis of Multi-Motion Sensor Behavior for Active Smartphone Authentication”, IEEE Trans. Inform. Forensic Secur., Vol. 13 (1), pp. 48–62, 2017.

[12] Z. Syed, J. Helmick, S. Banerjee, and B. Cukic, “Touch gesture-based authentication on mobile devices: The effects of user posture, device size, configuration, and inter-session variability”, J. of Syst. Softw., Vol. 149, pp. 158–173, 2018.

[13] M. Georgiev, S. Eberz, and I. Martinovic, “Techniques for Continuous Touch-Based Authentication”, Inf. Secur. Practice and Experience, Vol. 13620, pp. 409–431, 2022.

[14] Q. Yang, G. Peng, D.T. Nguyen, X. Qi, G. Zhou, Z. Sitova, P. Gasti, and K.S. Balagani, “A multimodal data set for evaluating continuous authentication performance in smartphones”, Proc. of the 12th ACM Conf. on Embedded Netw. Sensor Syst., pp.358–359, 2014.

[15] M. Antal, Z. Bokor, and L.Z. Szabo, “Information revealed from scrolling interactions on mobile devices”, Pattern Recognit. Lett., Vol. 56, pp. 7–13, 2015.

[16] M. Georgiev, S. Eberz, H. Turner, G. Lovisotto, and I. Martinovic, “Common Evaluation Pitfalls in Touch-Based Authentication Systems”, Proc. of the 2022 ACM on Asia Conf. on Comput. and Commun. Secur., pp. 1049–1063, 2022.

[17] M. Kudo, T. Takahashi, S. Ushiyama, and H. Yamana, “Fair and Robust Metric for Evaluating Touch-based Continuous Mobile Device Authentication”, 28th Int. Conf. on Intell. User Interfaces, pp. 141–144, 2023.

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

2024年3月1日(金) 13:30 ~ 15:40 | T5-A オンライン（Zoom Events)

ライフスタイル

座長:王 元元(山口大学)
コメンテータ:吉川 正俊(京都大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 14:00
[T5-A-8-01] 肌質や肌悩みを考慮した類似ユーザ判定に基づくコスメアイテム推薦手法

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

14:00 ~ 14:25
[T5-A-8-02] 大規模言語モデルを用いた他者の観点に基づくファッションアドバイザ

*吉倉 健人[1]、関 峰[1]、北山 大輔[1] (1. 工学院大学)

14:25 ~ 14:50
[T5-A-8-03] 中長期的影響を考慮した株式の取引支援

*福田 京司[1]、馬 強[1] (1. 京都工芸繊維大学)

14:50 ~ 15:15
[T5-A-8-04] 食品ロス削減を目的とした小売店舗の割引情報・料理レシピ提示システムの実現

*永浦 実樹[1]、阿部 広夢[1]、小林 千沙希[1]、石橋 直樹[1]、佐々木 史織[1] (1. 武蔵野大学データサイエンス学部データサイエンス学科)

15:15 ~ 15:40
[T5-A-8-05] 【技術報告】都市・ビル向けサービスの実装を容易化する空間デジタルツイン基盤

水品圭汰 (株式会社日立製作所)

DEIM 2024 T5-A-8-01

肌質や肌悩みを考慮した類似ユーザ判定に基づくコスメアイテム推薦手法

松岡　天音†　　山口　史弥††　　上田真由美†††,††††　　中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都産業大学 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
††† 流通科学大学 経済学部　〒 651–2188 兵庫県神戸市西区学園西町 3-1
†††† 大阪大学 サイバーメディアセンター　〒 560–0043 大阪府豊中市待兼山町 1 番 32 号
E-mail: †{g2054263,nakajima}@cc.kyoto-su.ac.jp, ††i2286211@cc.kyoto-su.ac.jp,
†††Mayumi_Ueda@red.umds.ac.jp

あらまし　協調フィルタリングに基づくコスメアイテム推薦を実現するためには，精度の高い類似ユーザ判定が重要である．ユーザ毎の肌質や肌悩みは多岐にわたり，単に肌タイプによる分類では不十分である．そこで本研究では，ユーザに部位毎の肌質や肌悩みに関するアンケートに回答してもらうと共に，このユーザが過去に投稿したレビュー文書を分析することで対象ユーザの肌質や肌悩みに関する特徴ベクトルを生成する．このアンケートで得た情報とレビュー文書から得られた情報を用いて，部位ごとの肌質や肌悩みを考慮したコスメアイテムの利用における高度な類似度判定する．この類似ユーザ判定結果を基にコスメアイテムを推薦する．

キーワード　コスメアイテム推薦，類似ユーザ抽出，協調フィルタリング，情報抽出，クチコミ分析

1　はじめに

近年，各種商品やサービスを紹介する Web サイトでは，ユーザのクチコミを掲載しているケースが非常に多くなっている．我々の研究対象であるコスメアイテムに関しては，@cosme[1]などがこれに該当し，多くのユーザが商品の購入を検討する際に利用している．特にコスメアイテムは直接肌に付けるものであり，高価な商品も多いため，ユーザ自身の肌の状態や好みに合うアイテムを選びたいという傾向が顕著であるものの，ユーザ毎の好みや使用感が大きく異なるため，自分にとって真に適切なアイテムを発見・購入することは容易でない．したがって，類似ユーザを用いた情報推薦方式として広く活用されている協調フィルタリングに対して，この推薦精度を向上させるためには，個々のユーザの好みだけでなく，肌の状態や肌悩みについても考慮した高精度の類似ユーザ判定手法を実現する必要がある．そこで本研究では，コスメアイテム推薦のための肌特徴や肌悩みを考慮した高精度な類似ユーザ判定手法の提案を目的としている．

ここで，一般的な 4 つの肌タイプの分類を図 1 に示す．図 1 に示すように，一般的な肌タイプは，普通肌，脂性肌，乾燥肌，混合肌の 4 タイプに分類されるが，それぞれの肌タイプの中でも，肌の特徴や悩みはユーザ毎に多岐に渡る．適切な類似ユーザ判定を行うためには，これら 4 つの肌タイプ分類だけでは十分とは言えない．そこで本研究では，各ユーザの肌特徴や肌悩みに状況と，このユーザが過去に投稿したクチコミの内容の両方を考慮した特徴ベクトルを抽出することで，図 2 に示すユーザ 1 とユーザ 2 の関係のように，単なる肌タイプ分類だけでなく，より詳細な類似ユーザ判定の実現を目指している．

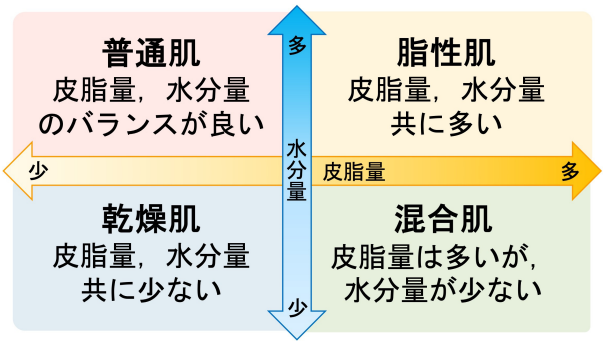


図 1　一般的に知られる肌タイプ分類

本稿の構成は以下の通りである．2 章では関連研究を紹介する．3 章では提案手法について詳細を説明する．4 章では評価実験計画，最後に 5 章ではまとめを記述する．

2　関連研究

以下に，コスメに対するクチコミの分析および類似ユーザ判定に関する関連研究をあげ，本研究との差異を示す．先行研究では，まず，松波らは，化粧水に関する評価表現辞書を構築し，評価項目別スコアを活用することで任意のユーザに対する類似ユーザ判定を提案している [1]．また，奥田らは，レビュー情報のみで得た評価項目別スコアを用いたコンテンツベースによるコスメアイテム推薦を提案している [2]．本研究では，部位ごとの肌特徴や肌悩みが類似したユーザの発見を可能にすることを目的としている．

1：@cosme,https://www.cosme.net/

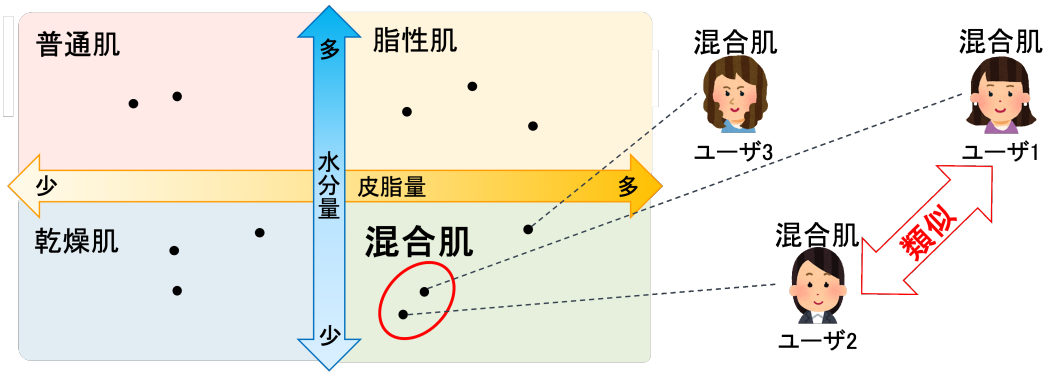


図 2　同肌タイプ内での高度な類似ユーザ発見の概要

中辻らは，ユーザが投稿したアイテムに対するレビュー上での，アイテムにかかる形容表現の頻度を分析し，評価アイテムに係る形容表現の傾向も考慮し，ユーザがアイテムに対しどのような感性を持って評価を下しているかを解析している [3]．本研究では，コスメアイテムを対象としており，アイテム自体の評価によりユーザの好みに合うか否かを推定するのみでなく，アイテムとユーザとの肌の相性がユーザ自身にとって適切か否かについても推定したうえで，類似ユーザの判定を行う．

Kavyashree らは，ディープラーニングを使用し，各ユーザの肌タイプに適した商品を予測し選択する手法を提案している [4]．人によって違う肌タイプとなる上に，製品によって合う肌状態が異なるため，最適な商品を選択することは非常に複雑化している．AI アルゴリズムを扱うことは，このような大量の非構造化データを評価することが可能であるため，最適であるとしている．本研究では，部位毎の肌質を考慮するとともに，肌質がより詳細に類似したユーザを発見すること，よりユーザの肌質に合ったコスメアイテム推薦の実現を目指しており，研究のアプローチが異なる．

山脇らは，関係性マイニングの手法を用いて 2 個のアイテム間での評価値の差が大きい選好パタンを明示的に選定し，ユーザ間の選好パタンの類似度を用いることで，ユーザベース協調フィルタリングを改良した推薦手法を提案している [5]．本研究では，総合評価ではなく，評価項目別の評価を使用し，これに基づいたユーザのより詳細な肌質や肌悩みから類似ユーザを発見することを目指しており，研究のアプローチが異なる．

以上のように，レビューやクチコミ分析に基づくコスメアイテム推薦，類似ユーザ判定のための取り組みは盛んに行われている．しかし，ユーザが過去に投稿したクチコミデータに加えて部位ごとの肌特徴や肌悩みを考慮したコスメアイテム推薦のための類似ユーザ発見に関する取り組みは十分に行われているとは言えない．このような背景から，本研究では，部位ごとの肌特徴や肌悩みを考慮することで，コスメアイテムの利用における高度な類似度判定手法を提案する．

3　コスメアイテム推薦のための肌特徴や肌悩みを考慮した類似ユーザ判定手法

本研究における，部位ごとの肌特徴や肌悩みを詳細に考慮した類似ユーザの発見は以下の手順で行う．

3.1　概　　要

ユーザはプロフィールに事前に肌の特徴や肌悩みを入力する．このユーザが入力した情報をもとにユーザ毎に肌悩みに関する明示的な特徴ベクトルを生成する．次に，ユーザが過去に投稿したクチコミから部位ごとの肌特徴や肌悩みの情報を抽出し，ベクトルを生成する．最後に，クチコミから抽出した情報とユーザ入力から生成した特徴ベクトルを組み合わせて，部位ごとの肌特徴と肌悩みを詳細に考慮した特徴ベクトルを生成する．特徴ベクトル生成方法としては，両ベクトルを連結させる連結ベクトル方式，両ベクトルそれぞれで類似度度を算出し，類似度を重みを付けて足し合わせる類似度加重平均方式の二つを挙げている．連結ベクトル方式では，このユーザ毎の特徴ベクトルの類似度を比較し，類似度の高いユーザを類似ユーザとする．類似度加重平均方式では，足し合わせた類似度が高いユーザを類似ユーザとする．この類似度が高いユーザが良い評価を付けたコスメアイテムを対象ユーザに提示させる協調フィルタリングを行う．

3.2　ユーザの事前入力情報から，特徴ベクトル生成

ユーザには事前に肌の悩みに対する気になる部位を選択するアンケート形式で答えてもらう．ユーザは事前に肌の特徴や肌悩み毎に気になる部位を，図 3 のようににチェックする．ユーザには特に気になる部位複数，他に気になる部位複数をチェックする．これを全ての肌の特徴や肌悩みに対して行う．この際表示される部位は，肌の特徴や肌悩みによって異なる．部位毎の項目は表 1 に示す．この入力情報から，図 4 のように特徴ベクトルを生成する．一番気になる悩みを 1 として，その他の悩みを 0.5 にする．また，チェックを付けていない箇所に関しては，全て 0 とする．この特徴ベクトルの次元数は，ユーザ毎に表 1 に示した全項目数とする．この表では全項目数は 32 個あるため，ユーザ毎に 31 次元のベクトルが用意される．

3.3　ユーザの過去のクチコミに基づく特徴ベクトル生成手法

次に，ユーザが過去に投稿した現在使用しているコスメアイテムに対するクチコミから肌質や使用感に関する情報を抽出し，特徴ベクトルを生成する．これは，3.2 節にて述べたユー



図 3　肌特徴や肌悩み毎に気になる部位を選択する入力インタフェース

表 1　肌特徴や肌悩み毎の部位の項目

肌特徴や肌悩み	項目
乾燥	全体，額，鼻，目元，口元，頬，こめかみ
テカリ	全体，額，鼻，頬，顎
ニキビ	全体，額，鼻，口元，頬，こめかみ，顎
毛穴汚れ	全体，額，鼻，口元，頬，顎
肌のハリの低下	全体，額，鼻，目元，口元，頬



図 4　ユーザの事前入力情報に基づく特徴ベクトル生成手法

ザが気になる部位にチェックをつける明示的な特徴ベクトルの生成に対して，暗黙的な特徴ベクトルの生成である．主に，対象ユーザが使用しているコスメアイテムと似たコスメアイテムに対して，同じようなクチコミをしているユーザを発見する目的で用いる．対象ユーザがあるコスメアイテム A に対して過去にクチコミを投稿していた場合，コスメアイテム A に対してクチコミを投稿していた全ユーザの評価項目別スコアを算出する．また，対象ユーザのコスメアイテム A に対する評価項目別スコアを算出する．この際，先行研究を用いて，各アイテムの評価項目別スコアを算出している．先行研究による評価項目は，香り，うるおい／浸透，美白／ UV，毛穴・角質ケア／高クレンジング，爽快感／顔のテカリ防止，サラサラ～トロトロ，低刺激，肌荒れ対策，エイジングケア，コストパフォーマンスの 10 個である．

3.4　評価項目別辞書に基づく自動スコアリング方式

本節では，我々の先行研究である，評価表現辞書を用いた評価項目別自動スコアリング方式について述べる．また，先行研究を用いて，各アイテムの評価項目別スコアを算出する方法を説明する．図 5 に評価表現辞書を用いた化粧品に対するクチコミの採点方法を示す．このシステムは，化粧品に対するクチコミのテキストを読み込み，クチコミ中に存在する評価表現を特定し，その評価表現を用いてクチコミの採点を行う．その後，評価表現辞書に基づいて各評価表現にスコアを付与する．1 つの評価項目について複数の評価表現が含まれている場合には，それぞれの評価表現に付与されたスコアの平均値をその評価項目に対するスコアとする．図 5 では，クチコミ中の「かなりしっとり」の部分に「7」,「よく潤い」の部分に「5」のスコアが付与されるため，このクチコミの評価項目「うるおい/浸透」に対するスコアは，(7+5)/2=「6」となる．先行研究のシステムは，1 つのクチコミにスコアを付与するものであるが，本研究では各アイテムに対する評価項目別スコアを使用したい．そこで，自動スコアリングシステムを用いてレビューごとのスコアを算出し，同じアイテムに対するスコアの平均値を評価項目ごとに算出することで，各アイテムの評価項目別スコアとする．

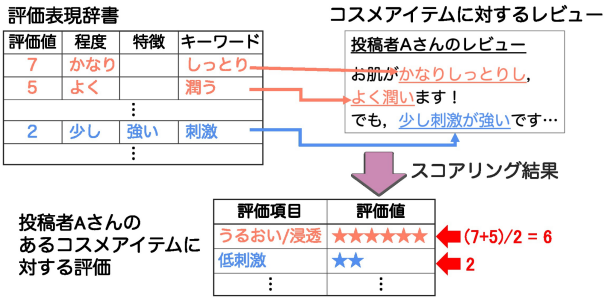


図 5　化粧品レビュー自動スコアリング方式の例

このコスメアイテムに対する評価項目別スコアをユーザ毎のクチコミ対するベクトルとする．

3.5　各ユーザの肌特徴・肌悩みに関する特徴ベクトルの生成

3.2 節でユーザが事前に入力した情報から生成したベクトルと 3.3 節でユーザが過去に投稿したクチコミから生成したベクトルの活用方法について説明する．

3.5.1　連結ベクトル方式

連結ベクトル方式については，図 6 のように後ろに付け加える．これを各ユーザの肌特徴・肌悩みに関する特徴ベクトルとする．このユーザ毎の特徴ベクトルの類似度を比較し，類似度の高いユーザを類似ユーザとする．

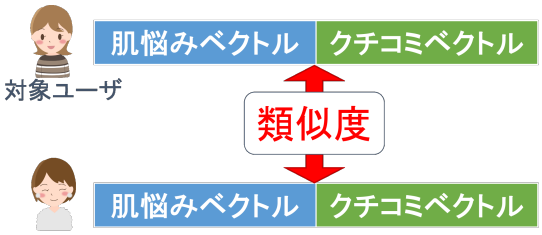


図 6　連結ベクトル方式

3.5.2 類似度加重平均方式

類似度加重平均方式については，図 7 のようにそれぞれ別で類似度を算出し重みを付けて足し合わせる．これを対象ユーザと共通アイテムを使用している各ユーザとの類似度を算出し，類似度の高いユーザを類似ユーザとする．

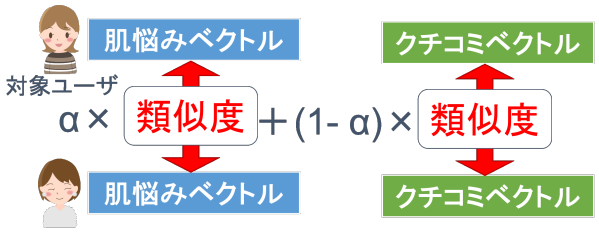


図 7 類似度加重平均方式

以上の手順によって生成された，ユーザ固有の肌特徴と肌悩みを部位毎に詳細に考慮した特徴ベクトルを生成することができる．この特徴ベクトルを用いることで，同じ肌タイプに分類されたユーザ同士でも，部位ごとの肌特徴や肌悩みを詳細に考慮した類似ユーザ発見を行うことができ，より高精度な肌タイプに基づくコスメアイテム推薦が可能になる．また連結ベクトル方式，類似度加重平均方式どちらの場合でも共通コスメアイテムに対するクチコミを用いているが，共通アイテムに対するクチコミが類似しているということは，真に類似していると言える．全ての手法の場合で類似度を算出する際，コサイン類似度を適用し，それぞれの類似度を算出する．コサイン類似度は以下の式で求めることができる．

$$cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|} \quad (1)$$

4 評価実験

提案手法であるユーザの事前入力情報と評価項目別スコアを組み合わせたコスメアイテム推薦方式の有効性を確認するための評価実験を実施した． 評価実験の実施前に，推薦者 30 名に対しサンプル化粧水に使用といった，事前準備を行った．

4.1 実験事前準備

本節では，評価実験を行うにおいての事前準備について述べる．被験者はまず以下のことを実施するよう指示した．

- 肌の悩みに対して特に気になる部位を複数，他にも気になる部位あれば複数選択するアンケート
  - 内容は表 1 で示したものと同様
- 今まで購入したことのある化粧水の中で良かった化粧水の複数回答
  - 回答する化粧水の個数は 1 個は必須でそれ以上の回答は任意
  - 今まで購入して良かった化粧水は推薦時に活用
- サンプル化粧水の使用とクチコミの記入
  - サンプル化粧水は 1 日分 (朝・夜の 2 回分) を 6 個

サンプル化粧水の使用とクチコミの記入に関しては，被験者には，以下の手順 (図 9 参照) に沿うよう指示した．また，被験者にはサンプル化粧水使用時には，図 8 のように容器に貼付している番号順①～⑥で使用することも併せて指示している．

（1） 夜にサンプル化粧水を使用
（2） 翌日朝に同じサンプル化粧水を使用
（3） 使用後時間を空けずにクチコミを記入するアンケートへの回答
（4） 次の番号のサンプル化粧水に対して手順 1 を実施しサンプル化粧水⑥まで繰り返し



図 8 被験者に実際に渡したサンプル化粧水

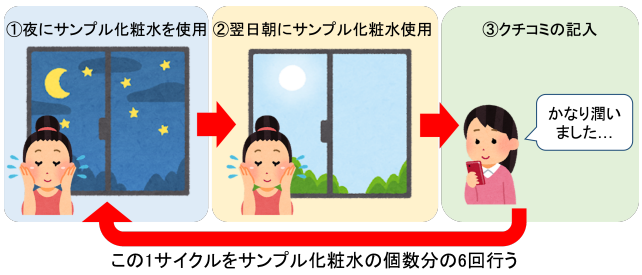


図 9 サンプル化粧水使用とクチコミ記入の流れ

この事前準備より，推薦者 30 名間で提案手法の肌悩みに関するアンケート，共通アイテムに対するクチコミの記入を擬似的に再現している．これにより，提案手法を基に各推薦者毎にベクトルを生成する．評価実験において，推薦者の中から被験者を選び，被験者の類似ユーザを被験者を除いた推薦者から発見する．

4.2 類似度加重平均方式における重みの検証実験

図 7 の類似度加重平均方式において適切な重み (図中の α) の検証を実施した．この実験は，類似度加重平均方式の重み α を調整し，被験者にとっての理想順位を最も適切に再現できる重み α の発見するために実施した．

4.2.1 実験手順

被験者数は推薦者 30 名の内の 5 名である．図??のように，被験者 5 名それぞれに対して，対象被験者を除いた推薦者全員分の今まで購入して一番良かった化粧水を，購入してみたい化粧水上位 20 位まで選択していただいた．

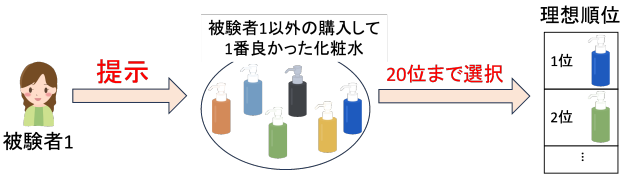


図 10　類似度加重平均方式における重みの検証実験の概要

対象被験者を除いた推薦者全員分の今まで購入して一番良かった化粧水を，購入してみたい化粧水上位 20 位まで選択した順位を理想順位とし，図 11 のように類似度の重みを調整して，提案手法の化粧水の順位≒理想順位になるような重みを見つける．化粧水の情報として，被験者に提示する内容は，図 12 のように，ブランド名，商品名，商品説明，価格，容量，既存サイトより商品トップページに記載されているクチコミ 5 つを被験者に提示した．

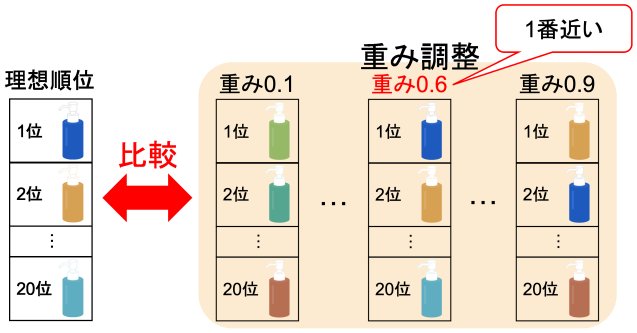


図 11　類似度加重平均方式における重みの検証実験の概要

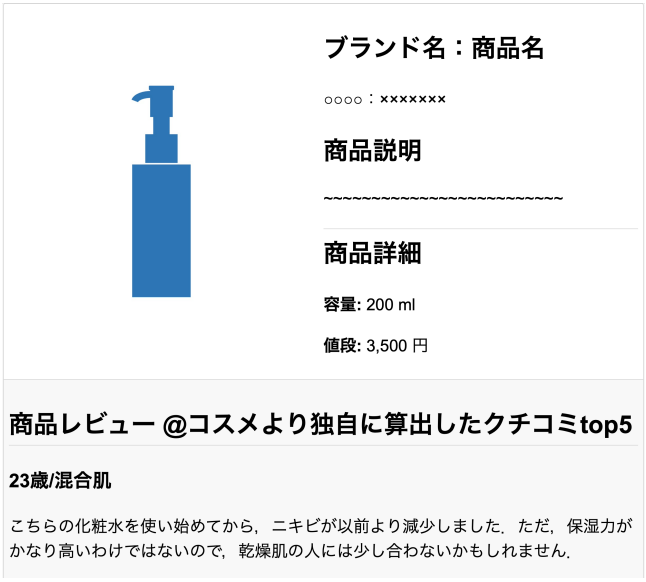


図 12　被験者に実際に提示する予定の内容

4.2.2　評 価 方 法

評価指標にはランキングの評価指標の一つである nDCG を用いた．nDCG は，情報検索や推薦システムなどで利用される評価指標の一つである，主にランキングの品質を評価するために使用され，0 から 1 の範囲の値を取り，1 に近いほどランキングの品質が高いことを示す．理想順位と重み α を 0.1 刻みで変更した類似度加重平均方式をそれぞれ比較し，nDCG@10 より評価を行う．

4.2.3　実 験 結 果

類似度加重平均方式の肌悩みベクトルの重み α(クチコミベクトルの重みは 1-α) を 0 1 で 0.1 刻みで変更させた，nDCG@10 の遷移を図 13 に示す．

図 13 より，肌ベクトルの重みが 0.8 の時 nDCG@10 が約 0.4 と最大になったことがわかる．この結果より肌悩みベクトルの重みを 0.8 にすることが妥当ではないかと考える．また，0.8 で最大かどうかを確認するため，0.7〜0.9 で 0.05 単位で重みを遷移させた図を図 14 に示す．この図 14 からも，0.8 で最大になっていることが分かる．以上より，類似度加重平均方式の重みを 0.8 と設定する．また，被験者ごとの nDCG@10 の結果を表 2 に示す．表 2 より，被験者 5 名中 4 名が nDCG@10 が 0.7 から 0.8 になるにかけて増加の傾向が見られた．しかし，重みが 1 の時 (肌悩みベクトルのみの時) が最大になった被験者も 2 名見られたため，肌悩みベクトルとクチコミベクトルの精度の差も比較する必要もあると考える．

4.3　提案手法とベースラインとの比較，最適な特徴ベクトル生成方法の検証

提案手法とベースラインとの比較と，連結ベクトル方式，類似度加重平均方式 (α=0.8) のどちらが適切かの検証を同時に実施する予定である．

4.3.1　実 験 手 順

被験者は，20 50 代の 11 名である．被験者に 10 位までの化粧水を提示し，購入したい化粧水を任意で選択していただいた．ここで提示した 10 位までの化粧水について説明する．

連結ベクトル方式による対象被験者との他の類似被験者上位 5 名の今まで購入して良かった化粧水 5 個，類似度加重平均方式による対象被験者との他の類似被験者上位 5 名の今まで購入して良かった化粧水 5 個，ベースラインの既存サイトによる対象ユーザと同じ肌タイプの化粧水ランキングの上位 5 個の化粧水の合計 3 つの化粧水のランキングをマルチリービングにより一つのランキングに統合する．

ここでマルチリービングについて説明する．ランキング評価手法であるマルチリービング [7] は，わずかなサンプルサイズでも複数のランキングアルゴリズムを効果的に評価できることが報告されている．マルチリービングは，入力として複数のランキング集合 $I = I_1, I_2, \ldots, I_n$ を受け取り，これらを混合して新しいランキング集合 $O = O_1, O_2, \ldots, O_n$ を生成し，複数のランキングを評価する手法である [7]．マルチリービングによる評価により，どのランキングが支持されたかが分かる．

4.3.2　実 験 結 果

マルチリービングの結果に基づき，ベースラインと類似度加重平均方式 (α=0.8) と連結ベクトル方式の平均順位を算出した表を以下に示す．

まず，ベースラインと重みが 0.8 の時の類似度加重平均方式 (α=0.8) とを比較する．比較類似度加重平均方式が 1.82 と比

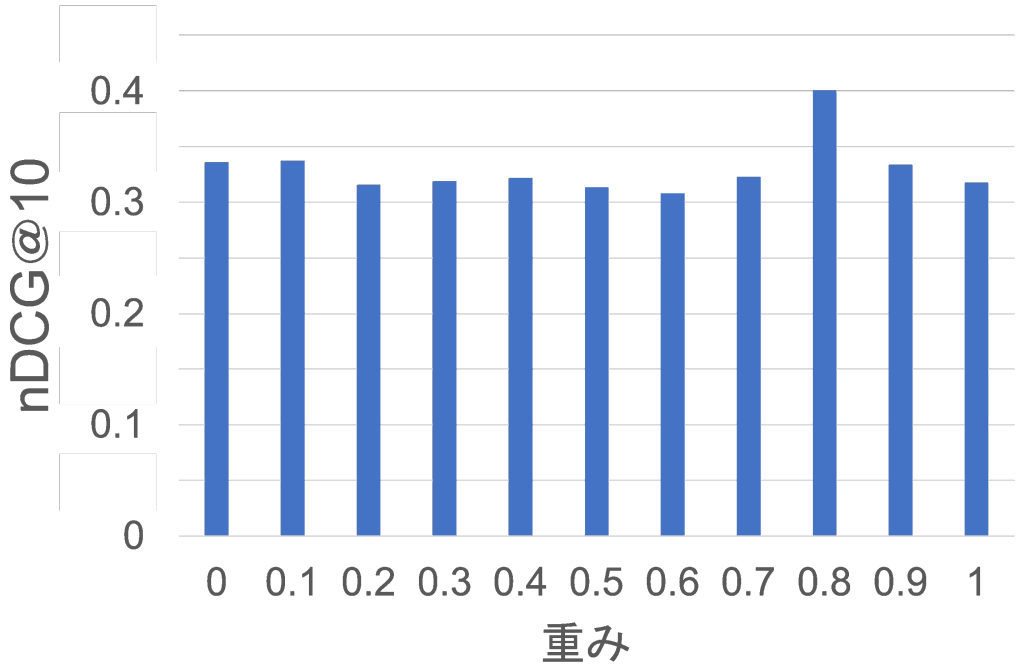


図 13　類似度ごとの nDCG@10 の遷移

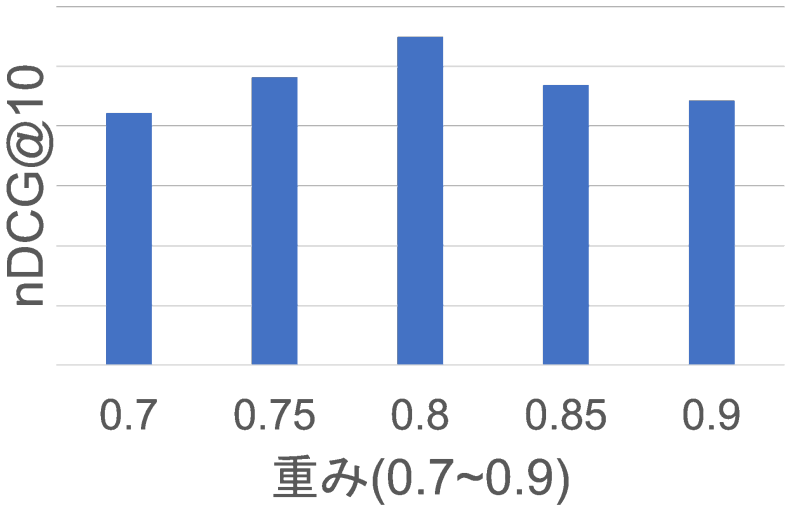


図 14　類似度ごとの nDCG@10 の遷移 (重み 0.7 0.9 を 0.05 刻みで表示)

表 2　被験者 5 名の nDCG@10

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
被験者 1	0.396	0.396	0.359	0.359	0.359	0.359	0.359	0.343	0.306	0.306	0.226
被験者 2	0.333	0.336	0.319	0.331	0.316	0.264	0.239	0.205	0.265	0.244	0.343
被験者 3	0.555	0.558	0.506	0.506	0.530	0.538	0.558	0.650	0.675	0.557	0.370
被験者 4	0.195	0.195	0.198	0.202	0.210	0.213	0.191	0.222	0.347	0.337	0.428
被験者 5	0.201	0.201	0.197	0.197	0.195	0.195	0.194	0.194	0.409	0.226	0.221

表 3　ベースラインと類似度加重平均方式 (α=0.8) と連結ベクトル方式の平均順位

	ベースライン	類似度加重平均方式 (α=0.8)	連結ベクトル方式
平均順位	2.27	1.82	1.91

較的上位であることが分かる．また，ベースラインと連結ベクトル方式の比較では，連結ベクトル方式が 1.91 と比較的上位であることが分かる．以上の結果をもとに．ベースラインに比べ提案手法の方が支持されたと言える．

次に，類似度加重平均方式 (α=0.8) と連結ベクトル方式を比較した際の支持された割合を示した図を以下に示す．

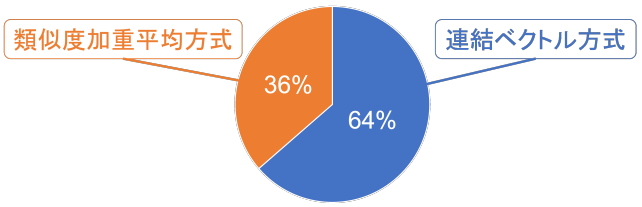


図 15　類似度加重平均方式 (α=0.8) と連結ベクトル方式を比較した際の支持された割合

この結果より，連結ベクトル方式が 64 %と高いことが分かる．これにより，連結ベクトル方式が類似度加重平均方式と比較すると，有効であると言えた．

5　おわりに

本稿では，ユーザに部位毎の肌特徴や肌悩みに関するアンケートに回答してもらうと共に，このユーザが過去に投稿したクチコミを分析することで対象ユーザの肌特徴や肌悩みに関する特徴を抽出することで，部位ごとの肌特徴や肌悩みを考慮するコスメアイテムの利用における高度な類似度判定手法を提案した．特徴ベクトル生成方式として，連結ベクトル方式と類似度加重平均方式を挙げた．評価実験として，まず，類似度の重みを検証する実験を実施し，0.8 が妥当であると結論づけた．また，提案手法とベースラインとの比較，最適な特徴ベクトル生成方式の検証を目的とした実験も実施し，提案手法の有効性，両方式のうち，連結ベクトル方式の方が比較的有効であることを示した．今後は，季節による肌質の変化等も考慮した肌特徴や肌悩みを考慮したコスメアイテム推薦の実現へ向ける．

謝　　辞

本研究を遂行するにあたり，株式会社アイスタイル様にコスメアイテムに関するレビューデータをご提供いただいた．また，本研究の一部は，JSPS 科研費 (課題番号：19K12243，20H04293，22K12281) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 松波 友稀ら，コスメアイテムに関する評価項目別スコアを用いた類似ユーザ判定手法,DEIM Forum 2019 G4-4，2019.
[2] 奥田　麻美ら，評価項目別スコアを用いたコスメアイテム推薦手法,DEIM2020 B1-3(day1 p10).
[3] 中辻真, 近藤光正, 田中明通, 内山匡, アイテムに係る形容表現を用いたユーザ類似度測定, 人工知能学会全国大会論文集, 3C4-02, 2010.
[4] Kavyashree N, Rama Satish K V, Prasanna Rajaram Rasal, Dr.Rahul Jalindar Jadhav, Dr. M.Saidireddy, K G Kharade, Artificial Intelligence based Smart Cosmetic Suggestion System based on Skin Condition, ICACRS 2022.
[5] 山脇 淳一, 工藤 康生, 村井 哲也, 関係性マイニングと協調フィルタリングを用いた情報推薦手法, 日本感性工学会論文誌, Vol.17 No.4 pp.481-488.
[6] 伊藤寛明, 吉川大弘, 古橋武, アイテム推薦のためのアソシエーションルールを用いた類似ユーザの抽出に関する基礎的検討, 情報処理学会第 75 回全国大会, 2P-8, 2013.
[7] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. 2014. Mul- tileaved Comparisons for Fast Online Evaluation. In CIKM. 70-80.

大規模言語モデルを用いた他者の観点に基づくファッションアドバイザ

吉倉　健人† 　関　　峰†† 　北山　大輔†

† 工学院大学 情報学部 システム数理学科　〒 163-8677 東京都新宿区西新宿 1-24-2
†† 工学院大学大学院 工学研究科 情報学専攻　〒 163-8677 東京都新宿区西新宿 1-24-2
E-mail: †{j320306,em22006}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　毎日の服選びは生活において欠かせない行動である．また，よりよい服の組み合わせを考えてファッションコーディネートを楽しむ者もいる．この時，雑誌や SNS に掲載されている他者のコーディネートを参考にする場合がある．しかし，その要素をうまく自分のコーディネートに取り入れることは困難である．本研究では，ユーザのコーディネートと参考にしたいコーディネートを入力とし，ユーザコーデを参考コーデに近づけるためのアドバイスを大規模言語モデルを用いて生成する手法を提案する．また，プロトタイプシステムを構築し，生成されたアドバイスの妥当性を評価する．

キーワード　大規模言語モデル，SNS データ，ファッション，アドバイス生成

1　はじめに

衣服は，我々の生活の三大要素であり，衣食住に含まれるほど，生活において重要な要素である．そのため毎日の服選びは，外出する際に必ず行われる行為であり，服を選ぶ際，その日の天気や訪れる場所，自分自身の服の好み，選んだ服の組み合わせなど，さまざまな要素を考慮する必要がある．また，その日に会う予定の人がいる場合には，その人との関係性，その人と以前に会ったとき，自分がどのような服を着ていたかなどを考慮して，服を選ぶことがある [1]．また，よりよい服の組み合わせを考えて，ファッションコーディネートを楽しむ者もいる．その際に，自分だけの力で組むと，どうすればおしゃれなコーディネートになるのか悩むことがある．そのため，雑誌やファッション投稿型 SNS「WEAR[1]」を用いて，さまざまなコーディネートを参考にすることがある．しかし，参考になるコーディネートを見つけたとしても，その要素をうまく自分のコーディネートに取り入れることは容易ではない．このとき，現在の自分のコーディネートに対し，参考にしたいコーディネートに近づけるためのアドバイスをもらえるファッションアドバイザのような存在が身近にいるのが望ましい．

そこで本研究ではユーザのコーディネートをよりよくするために，他者観点を利用した，ファッションにおけるアドバイス生成手法を提案する．アドバイス生成には大規模言語モデルである gpt-3.5-turbo[2]を用いる．自然言語処理とファッションを掛け合わせた研究として，有川ら [2] はキーワードをクエリとして，クエリの意味に適合し組合せの良さを考慮したコーディネートの検索手法を提案した．本研究はユーザにわかりやすい文章を提示することによって，普段のファッションコーディネートをよりおしゃれにするための指針を手軽に得ることで，誰でもファッションコーディネートを楽しく行えるようになることを目的とする．

1：https://wear.jp/
2：https://platform.openai.com/docs/api-reference/chat

2　関連研究

2.1　ファッションの UI による支援

ファッションの分野に関する先行研究は過去にいくつも行われている．これらの研究では，コーディネートを作るためだったり，アドバイスに基づいてコーディネートを試すときに使われる．

池田ら [3] はユーザの評価情報を基にして k-means によってクラスタリングを行い，それを個々のユーザの嗜好として CNN によって学習することで，個々のユーザの嗜好を考慮したコーディネート推薦方式を提案した．服飾コーディネートに対する評価情報を基準にしてクラスタリングさせたグループと，クラスタリングを行わずにさまざまな嗜好を混合させたグループで予測精度を比較し，前者の方が精度がいいことを検証することで有意性を示した．

大野ら [4] は特定の SNS からではなく，複数のソーシャルメディアサイトのデータやファッション通販サイト，Web ファッション情報誌サイトを活用することで，ファッションブランド探索支援に向けたデータセットの構築，ファッションブランド間の関係性の可視化，およびファッションブランドの社会性を加味したファッションブランド探索支援システムの構築例について議論した．

小野ら [5] はハンガーにかかった服を掛けると，それに合わせたコーディネートをその服に合わせて表示し，推薦するコンテンツを作成した．上着をスクリーンの前に掛けると，それに合わせたパンツの映像をちょうど来た時のように上着の下に表示し，パンツを掛けると上着をその上部に表示する．このようなインタラクティブなシステムを用意することで，多くの組み合わせを提示することができる上に，同じ服に対しても，他の条件などから異なった組み合わせを勧めることも可能であると

述べている．

佐藤ら [6] はユーザが過去に着用した服を鏡越しに撮影し，その姿をトップスとボトムスとの境界線で 2 枚に切り分け，組み合わせることでコーデのシミュレーションを実現するシステム”suGATALOG”を提案し，その試作と評価実験，試験運用の結果について考察した．服を着なくてもコーディネートを想定できることがシステムの大きな特徴である．これに関して評価実験では，実際の服を手にもってコーディネートする場合，服のみの写真を PC 上でコーディネートする場合，この手法によって姿の写真を用いてコーディネートする場合の 3 つをアンケートによって比較した結果，この手法を利用した場合が確信をもってコーディネートができるという結果となり，有用性を示すことができた．

本研究では，完成されたコーディネートを入力として用いるため，このようなコーディネートや組み合わせるアイテムの推薦により，入力を容易にすることができる．

2.2 言語モデルを用いたアドバイス・解説文生成

言語モデルによってアドバイス文やレビュー文を生成する手法は過去にも提案されている．それは，さまざまな分野で行われている．

教育の分野において，廣瀬ら [7] は，大規模言語モデルである GPT-3 を用いたアドバイスの自動生成によって，学生の学習計画と振り返り内容に対する支援を行った．アドバイス自動生成のための再学習のトレーニングデータに，ルーブリックを用いた評価と個別の記述内容に対して，教員がアドバイスしたテキストを組み合わせた手法を提案した．そして，自動生成されたアドバイスの内容について性能評価を行った．事前に定めたルーブリックによって学習の記述の問題点を判定することで，学生個別の内容に応じたアドバイスの生成ができるか，品質の揃ったアドバイスを生成することがなどを検証した．その結果，教員作文に近いバランスのアドバイスが生成できたことが示唆された．

ゲームプログラミングの分野において，亀甲ら [8] は，局面や指し手についてコンピュータ将棋プログラムが示す評価を，自然言語による解説として提供することができれば，観戦の際により有益な情報となることが期待できるとして，2 段階の機械学習による将棋解説文の自動生成手法を提案した．与えられた局面を表現する特徴的な単語を予測し，その単語の情報と言語モデルを組み合わせ，解説文を生成する．そして解説文を戦型に言及するものに限定して実験を行い，提案システムを用いることで有用な解説文を生成できる局面が存在することを確認した．

本研究は，ファッションコーディネートの分野で言語モデルを用いたアドバイス生成を行う．

2.3 ファッションドメインにおける言語モデルの応用

ファッションドメインにおいて言語モデルを活用した先行研究は数々存在する．

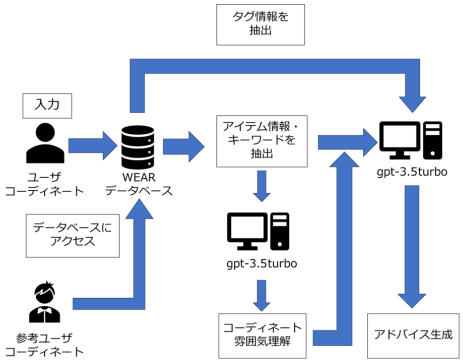


図 1　提案手法の概要図

青葉ら [9] は，パーソナルスタイリングにおいて，2 つのファッションアイテムをコーディネートに推薦する際の，コメントを生成するモデルを構築した．具体的には Adapter を用いることで言語モデルをマルチモーダルな入力に対応させた手法である MAGMA を利用し，2 枚の服画像を結合した画像とプロンプトを入力として，コーディネートコメント文を生成する手法である．実験では人手によって作成されたコメントや，機械学習をベースとした従来手法との定量・定性比較を行い，提案モデルが十分な能力を持つことを確認した．

学習者による教授者の感性の理解は，教授者の感性を推測しながら作品を作成する活動によって促進されると考えられる．そこで福村ら [10] は，作成された作品の是非を知ることは，自身が推測している教授者の感性の正誤の理解につながると考え，学習者に教授者の感性の推測に基づいた作品を作成できる環境を構築した．この研究ではファッションを対象として，教授者の作品を正解データとして格納でき，学習者にその正解データを回答させるとともに，正誤を判定して必要に応じてヒントを導出するようなシステムを考案した．

本研究では，入力には 2 つの完成されたコーディネートを用い，片方を他方のコーディネートの基準として近づけるための助言を，大規模言語モデルを用いることで生成する．

3 大規模言語モデルを用いたアドバイス生成

3.1 概　　要

本研究のファッションアドバイス生成手法について説明する．概要図を図 1 に示す．まずユーザは自分のコーディネートと参考ユーザのコーディネートを入力する．コーディネートの入力は WEAR に投稿されたコーディネートの URL によって行う．次に，それぞれのコーディネートに使われているアイテム情報やタグ情報を抽出する．これは，アイテムの名前だけで実行すると，生成されるアドバイス文に，アイテムの特徴が反映されず，適切なアドバイスにならないからである．アイテム情報は，WEAR からスクレイピングによって取り出すことで収集したデータベースにアクセスすることで抽出した．このデータベースは 2022 年 10 月から 2023 年 6 月上旬に投稿されたデー

表 1　コーディネートの雰囲気理解のためのプロンプト設定

Role	Prompt
user (アイテム数分繰り返し)	ユーザのコーディネートに使われているアイテムの情報と特徴語は以下です。 このコーディネートはどのような印象を与えてますか。簡潔かつ具体的に説明してください。 [アイテム名] は [ブランド] の [カテゴリー] です。 [カラー] 色で、[サイズ] です。 このアイテムには以下のような特徴語が得られました。* 「○○」、「○○」、「○○」、

表 2　アイテム情報を用いたアドバイス生成のためのプロンプト設定

Role	Prompt
user	ユーザのコーディネートは以下の通りです。 (ユーザコーディネートに基づく表 1 のプロンプト)
assitant	ユーザのコーディネートは次のような印象を与えます。 (上記のプロンプトに基づく出力)
user	参考ユーザのコーディネートは以下の通りです。 (参考ユーザコーディネートに基づく表 1 のプロンプト)
assistant	参考ユーザのコーディネートは次のような印象を与えます。 (上記のプロンプトに基づく出力)
user	ユーザのコーディネートに対して、参考ユーザのコーディネートに近づけるための助言を 200 文字程度で生成してください。

タが収集されている．本研究で抽出したアイテム情報はアイテム名，ブランド，カテゴリー，カラー，サイズである．以下に UNIQLO のエアリズムコットンオーバーサイズ T シャツ [3]のホワイトの XL サイズの例を示す．

- アイテム名: エアリズムコットンオーバーサイズ T シャツ
- ブランド:UNIQLO
- カテゴリー:T シャツ・カットソー
- カラー: ホワイト系
- サイズ:XL

本稿では特徴語として，アイテムに付与されているキーワードとコーディネートに付与されているタグ情報を採用した．キーワードは，アイテムの登録者が入力しているもので，アイテムの公式な提供者が登録した場合は適切なものが入力されている．さらに，プロンプトを以下のように三段階に分けた．このように段階を踏むことで，コーディネートの特徴・雰囲気を GPT に適切に理解させ，特徴や雰囲気に基づくアドバイスを生成できると考えた．

（1） ユーザコーディネートの雰囲気理解
（2） 参考ユーザコーディネートの雰囲気理解
（3） (1),(2) を GPT に入力し, 参考ユーザを基にしたアドバイス生成

3：https://wear.jp/item/46247251/

表 3　タグ情報を用いたアドバイス生成のためのプロンプト

Role	Prompt
user	ユーザのコーディネートに付与されているタグ情報は以下の通りです。 [タグがカンマ区切りで列挙される]
user	参考ユーザのコーディネートに付与されているタグ情報は以下の通りです。 [タグがカンマ区切りで列挙される]
user	ユーザのコーディネートに対して、参考ユーザのコーディネートに近づけるための助言を 200 文字程度で生成してください。

3.2　各コーディネートの雰囲気理解とアドバイス生成

3.2.1　アイテム情報に基づく手法

アイテム情報に基づく手法については，入力されたコーディネートに使われている各々のアイテム情報と特徴語を抽出したのち，これらをプロンプトに埋め込んで GPT に入力し，そのコーディネートの雰囲気を文書として生成する．本稿では，アイテムとしては，トップス，ボトムス，シューズのカテゴリに属するものだけを使用し，アクセサリなどは対象としない．ここでは gpt-3.5-turbo-1106 を用い，出力するテキストの文字数は 200 文字程度とし，出力する単語の創造性を 0 から 1 で表す temperature は 0.0 とした．そのときのプロンプトを表 1 に示す．アイテム情報の箇所はアイテムの個数分だけ繰り返して追加される．また * で示したアイテムの特徴に関する文は，特徴語がないアイテムの場合は用いない．参考ユーザのコーディネートについても同様の手順とプロンプトでコーディネートの雰囲気に関する説明文を生成する．

アドバイス生成では，入力したコーディネートのアイテム情報と特徴語，および生成されたコーディネートの雰囲気に関する説明文をプロンプトに入力する．アドバイス生成でも gpt-3.5-turbo-1106 を用い，出力するテキストの文字数は 200 文字程度とし，出力する単語の創造性を 0 から 1 で表す temperature は 0.0 とした．そのときのプロンプトを表 2 に示す．1 つ目と 2 つ目の user は，各ユーザのコーディネートに使われているアイテム情報と特徴語を表示させるプロンプトである．1 つ目と 2 つ目の assistant は，各ユーザのコーディネートの雰囲気を表示させるプロンプトである．最後の user は，ユーザのコーディネートに対して参考ユーザを基にしたアドバイスを生成させるプロンプトである．

3.2.2　タグ情報に基づく手法

タグ情報に基づく手法については，入力されたコーディネートに付与されているタグ情報をすべてプロンプトに用いる．そのときのプロンプトを表 3 に示す．[] 内は変数であり，コーディネートのタグが列挙されている．

3.2.3　選別したタグ情報に基づく手法

選定したタグ情報に基づく手法については，3.2.1 節で得られるコーディネートの雰囲気の文書を MeCab による形態素解析を行い，抽出した名詞と，コーディネートに付与されている

表 4　予備検討に用いたコーディネートのアイテム名とキーワード

アイテム名	キーワード
ボタンデニムパンツ	ストレート (パンツ ひざ下) ハイウエスト (股上) フルレングス (パンツ丈) ワイド・ルーズ (パンツ シルエット) ヴィンテージ (デニム加工) 無地 (柄)
厚底ダッドシューズ	あり (紐) クッション性 (機能性) ローカット (スニーカー 型) ワンポイント (柄・デザイン (シューズ)) 無地 (柄・デザイン (シューズ)) 花柄 (柄・デザイン (シューズ)) 軽量 (機能性)
カラーカーディガン	ショート丈 (着丈) ワンポイント (柄) 無地 (柄) 長袖 (袖丈)

表 5　キーワードなしの場合の出力結果

このコーディネートは、ブルー系のボタンデニムパンツとホワイト系の厚底ダッドシューズ、そしてピンク系のカラーカーディガンを組み合わせたものです。ブルーのデニムパンツはカジュアルでありながらも、ボタンデザインがアクセントとなり、少し個性的な印象を与えます。ホワイトの厚底ダッドシューズはトレンド感があり、スポーティで元気な印象を与えます。そして、ピンクのカラーカーディガンは女性らしさを引き立て、柔らかな印象を与えます。全体的には、カジュアルでありながらも個性的で元気な印象を与えるコーディネートです。

表 6　キーワードありの場合の出力結果

このコーディネートは、ボタンデニムパンツと厚底ダッドシューズを組み合わせたカジュアルな印象を与えます。ボタンデニムパンツはストレートでハイウエストのデザインであり、ヴィンテージ感があります。厚底ダッドシューズはホワイト系色で、クッション性と軽量さがあり、ワンポイントの花柄がアクセントになっています。カラーカーディガンはピンク系色で、ショート丈であり、ワンポイントの柄があります。全体的には、リラックス感のあるカジュアルスタイルで、女性らしさも演出されています。

タグ情報の類似度を計算し，0.4 以上のタグ情報を採用してプロンプトに用いる．類似度を 0.4 以上にしたのは，3 個から 4 個ほどのタグ情報が選別できるので，数として適切であると考えたからである．類似度の計算には fasttext を用いた．プロンプトは表 3 と同じものを用いる．この手法によって，コーディネートの雰囲気と無関係なタグを除くことができる．

3.3　雰囲気理解のプロンプトの予備検討

本節ではコーディネートの雰囲気理解において，アイテムの持つキーワードを入力に含めることの影響についての実験を行った．システムにあるコーディネート [4]を入力し，アイテム

4：https://wear.jp/can06/22164362/

表 7　評価実験のアンケート

Q1	システムが作成したアドバイスは、参考になりましたか。 1 - 全く参考にならなかった 2 - 参考にならなかった 3 - わからない 4 - 参考になった 5 - とても参考になった
Q2	Q1 の選択理由を教えてください。

キーワードを用いなかった場合と，用いた場合のコーディネート雰囲気理解の説明文生成を行った．用いたコーディネートのアイテム名とキーワードについては表 4 の通りである．ここで，コーディネートに関しては，コーディネートの雰囲気理解にキーワードが介入しているかを確認しやすくするために，アイテムの特徴が推測しにくいようなアイテム名をコーディネートに取り入れているものを使用した．キーワードなしの場合の出力結果を表 5 に，キーワードありの場合の出力結果を表 6 に示す．キーワードがない場合とある場合で，コーディネート全体の雰囲気説明に大きな差異は生まれず，結論として同じような文書生成をしている．たとえば,「カジュアル」という言葉は両方の出力で使用されており，差異としては，キーワードありの方では最後に「女性らしさ」という言葉に対し，キーワードなしの方では「個性的で元気」という言葉を使っているのみである．また，キーワードの役割についても，コーディネートに使われているアイテムの特徴に関する説明に，そのキーワードを使っているか否かの変化のみに過ぎず，効果としても不十分である．これは，アイテムに付与されているキーワード自体が特徴語として弱かったことに加え，付与されたアイテム情報で大まかなアイテムの特徴を GPT が捉えることができているからであると考えられる．そのため，特徴語が得られないアイテムについてはとくに情報の追加を行う必要はないと判断した．また，特徴語が原因で出力が悪くなることもないため，特徴語がある場合はそのまま用いるのが妥当であると判断した．

4　評 価 実 験

4.1　概　　要

評価実験の方法について説明する．被験者に提示した入力もしくは参考とするコーディネート (以下，入力コーデ，参考コーデと記す) は，WEAR のいいね数によって選定し，男性ユーザから上位 20 件，女性ユーザから上位 20 件，計 40 件を用いた．ここで，提示するコーディネートを投稿しているユーザにはそれぞれの嗜好があるとして，それを分散させるために，20 件の中で同一ユーザのコーデは 2 つまでとした．3 章で説明した 3 タイプを入力コーデと参考コーデにそれぞれ用いることから，計 9 パターンのアドバイスについて比較してもらった．本研究では以下のように手法を名付けた．

- タイプ A: アイテム情報に基づく手法
- タイプ B: タグ情報に基づく手法

表 8　集 計 結 果

入力コーデ	A	A	A	B	B	B	C	C	C
参考コーデ	A	B	C	A	B	C	A	B	C
平均値	3.62	3.44	3.66	3.37	3.32	3.73	3.25	3.47	3.55
回答者数	71	73	74	73	77	75	69	75	76

表 9　使用タイプごとの平均値

	A	B	C
入力コーデ	3.57	3.48	3.42
参考コーデ	3.41	3.41	3.65

- タイプ C: 選別したタグ情報に基づく手法

実験はクラウドソーシング [5]によって集めた被験者 80 人に，任意に入力および参考コーデを選んでもらい，それを用いて生成された 9 パターンのアドバイスについてすべて評価してもらった．評価項目としては表 7 の通りである．生成された各アドバイスについて，被験者に評価してもらう．この実験によって，アドバイス生成に用いる情報としてふさわしいコーディネートの情報は何かということと，生成されたアドバイスは妥当なものかを確かめる．仮説は，入力コーデと参考コーデには同じ手法を用いることで，評価の高いアドバイスが生成されるとする．

4.2 結　　果

表 7 の Q1 の評価を，各アドバイスについて集計した結果を表 8 に示す．本研究では表 8 で，何かしらのエラーによってアドバイスが出力できなかった場合，そのアドバイスを出力した被験者は全体の回答者数に数えず平均値を算出した．平均値は小数第三位以下を四捨五入したものである．

実験の結果から，入力コーデにタイプ B，参考コーデにタイプ C を用いたときの平均値が高いことがわかる．このアドバイスに対して Q1 で 5 や 4 を与えた被験者の多くは Q2 では具体性の高さについて高く評価していた．とくに,「カジュアルさをどのように出せばいいか参考になった。」や,「自分が目指すスタイルと合致しており、何を組み合わせることでどういう効果があるか、具体的に示してくれている。」など，アイテムの組み合わせ方に着目していることに対する意見が述べられており，具体性の高さに定評があったといえる．しかし，このアドバイスについて Q2 で 1 や 2 を与えた被験者は，生成されたアドバイスの抽象性を指摘していた．例として,「『自分らしいアレンジ』というのが抽象的で、自分で考える要素が多い」という意見が述べられていた．

また，それぞれのコーディネートを各タイプに用いたときの評価の平均値をまとめた表 9 を見ると，入力コーデとしてはタイプ A を，参考コーデとしてはタイプ C を用いたときがもっとも評価が高く，表 8 においても，その平均値は 2 番目に高い評価を得ている．このアドバイスに対して Q1 で 5 や 4 を与えた被験者も，アドバイスの具体性の高さを評価していた．中には,「自分では考えつかなかったカラーがアドバイスされていたから。」という記述がされており，気づき支援につながっている場合もあり得ることが確認された．

5：https://crowdworks.jp/

表 10　良いアドバイス例

参考ユーザのシンプルでカジュアルなスタイルに近づけるためには、テーラードジャケットやシャツスタイルのアイテムを取り入れると良いでしょう。例えば、カーキのテーラードジャケットをオーバーサイズのシャツとレイヤードすることで、シンプルでありながらおしゃれな印象を与えることができます。また、カーゴパンツやワイドパンツなどのゆるいシルエットのボトムスを選ぶことで、参考ユーザのようなカジュアルな雰囲気を演出できます。さらに、ボディバッグなどの小物を取り入れることで、トレンド感もプラスされます。GU などのプチプラブランドを活用しながら、シンプルでありながらおしゃれな大人カジュアルスタイルを目指してみてはいかがでしょうか。

実験結果に対して,「アドバイス生成にはどの情報を用いても，有意な差は得られない」という帰無仮説を立て，有意水準 0.05 で一元配置分散分析を行った．その結果，全体の評価では p 値が 0.183 と有意水準を上回ったため，帰無仮説は棄却されない．表 9 についても，入力コーデについても，p 値は 0.415 と有意水準を上回ったため，ここでも帰無仮説は棄却されない．しかし，参考コーデの評価では p 値は 0.047 と有意水準を下回ったため，帰無仮説は棄却され，有意な差があることが確認された．そこで，参考コーデで集約した結果について，各タイプにおいて有意水準 0.05 で t 検定を行った．両側検定の結果，タイプ A とタイプ B の p 値 0.971 と有意水準を上回ったので，帰無仮説は棄却されない．しかし，タイプ A とタイプ C は p 値 0.032，タイプ B とタイプ C は p 値 0.028 と，どちらも有意水準を下回ったので，帰無仮説は棄却され，有意な差があることが確認できた．

表 8 から，平均値がもっとも低かったのは，ユーザのコーディネートにタイプ C，参考ユーザのコーディネートにタイプ A を用いた場合であった．このアドバイスに対して Q1 で 1 や 2 を与えた被験者の 1 人は Q2 では「ニットのセーターなのに、T シャツを勧められてもわからない。」という出力がされていたことを指摘していた．その被験者の入力を見ると，参考ユーザのコーディネートが，ニットのインナーにシャツを用いているものであった．また，表 9 から，参考ユーザの入力にタイプ A を用いている場合は比較的評価は低く，その理由としては,「専門用語が多い。」という意見があった．

4.3 考　　察

実験結果から，多くの被験者は，参考にしたいユーザの情報として，そのユーザのアイテム情報から得られるコーディネートの雰囲気を基に選別されたタグ情報を用いることで，ユーザにとって具体性の高く，参考になるアドバイス生成ができることが確認できた．入力コーデにタイプ B，参考コーデにタイプ C から生成したアドバイスである表 10 のように，参考にしたいコーディネートにするための雰囲気の出し方や，アイテムの

表 11　悪いアドバイス例 1

参考ユーザのコーディネートに近づけるための助言として、以下のようなアドバイスが考えられます。 まず、シンプルながらも洗練された印象を出すために、ワイドパンツを選んでみてはいかがでしょうか？UNIQLO などでワイドパンツを探してみてください。また、大人カジュアルの雰囲気を出すために、シンプルなトップスと合わせることでバランスが取れます。例えば、無地のシンプルな T シャツやシャツが良いでしょう。足元には、スニーカーやローファーなどのカジュアルシューズを合わせると、大人カジュアルな印象が強調されます。色合いにもこだわり、モノトーンや落ち着いたカラーを選ぶと、洗練された雰囲気が演出できます。これらのアイテムを組み合わせることで、参考ユーザのコーディネートに近づけることができるでしょう。

表 12　悪いアドバイス例 2

ユーザのコーディネートを参考ユーザのコーディネートに近づけるためには、ワイドバルーンデニムパンツに代わるアイテムとして、同様のワイドシルエットのデニムパンツや、ワイドレッグパンツを検討してみてはいかがでしょうか。また、CONVERSE のスニーカーに代わるアイテムとして、同じくカジュアルで洗練された印象を持つスニーカーや、カラフルなスニーカーを選ぶことで、コーディネートにアクセントを加えることができます。さらに、ウール混ブークレヤーンのオーバーサイズクルーネックプルオーバーニットに代わるアイテムとして、同様の暖かみのあるニットや、オーバーサイズのトップスを選ぶことで、モダンでスタイリッシュな印象を演出できます。全体的には、カジュアルながらも洗練された要素を取り入れることで、参考ユーザのコーディネートに近づけることができます。

組み合わせ方を丁寧に出力していた．これは参考ユーザの情報の中にアイテム情報をある程度踏まえていることで，タグ情報が選別され，ムダのないアドバイスができているからであると考えられる．

また，参考ユーザのコーディネートの情報にアイテム情報を取り入れた場合のアドバイスは，アイテム情報に寄ってしまう．そのため，ニットの下にシャツや T シャツを用いているようなコーディネートの場合は，入力コーデにタイプ C，参考コーデにタイプ A から生成した表 11 のように，プロンプトに加わってしまい，アドバイスでもアイテムの推薦として出力してしまうという場合や，入力コーデにタイプ B，参考コーデにタイプ A から生成した表 12 のように,「ウール混ブークレヤーン」といった，アイテム名から抜き取られた専門用語が含まれることが原因で，ユーザにとって納得されにくいアドバイスとなってしまうと考えられる．

以上のことから，参考コーデにはアイテム情報から得られるコーディネートの雰囲気を基に選別されたタグ情報を用いるのが，より具体性のあるアドバイス生成ができることが確認できた．入力コーデについては，どの情報を用いても，アドバイスの評価に有意な差は得られなかった．本稿ではアイテム情報を用いた方がいいという結果にはなったが，検定では有意差は表れなかった．したがって，さらに被験者を増やして，より多くのデータを得ることが必要である．

5 ま と め

本研究では大規模言語モデルを用いた他者の観点に基づくファッションアドバイザについて提案した．GPT のプロンプトに用いるコーディネートの情報の違いによって，生成されるアドバイスが被験者にとってどれくらい参考になったかを 5 段階で評価してもらった後，その評価の理由を記入してもらった．その結果，参考コーデにそのコーデのアイテム情報から得られるコーディネートの雰囲気を基に選別されたタグ情報を用いることが，ユーザにとって具体性が高く，もっとも参考になるアドバイス生成となることが確認された．今後の課題として，まず，ユーザの体型やファッションの経験値などを加味することで，個々のユーザのニーズに合わせられるようなアドバイス生成を検討している．また，ユーザの入力コーデの情報としてふさわしい情報は何かを，被験者を増やして明らかにすることや，ROUGE 等を用いて，生成されたアドバイスと，人手で生成したアドバイスとの整合性を評価することも検討している．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 佐竹雪乃, 山田篤志, 築舘多藍, 小林稔. 服と思い出：人間関係に基づくコーディネート支援の提案. Technical Report 35, 明治大学総合数理学部先端メディアサイエンス学科, 明治大学大学院先端数理科学研究科先端メディアサイエンス専攻, 明治大学総合数理学部先端メディアサイエンス学科, 明治大学総合数理学部先端メディアサイエンス学科, jan 2021.

[2] 有川魁人, 加藤誠, 吉川正俊. キーワードに基づく組み合わせの良さを考慮したファッションコーディネート検索. 第 12 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2020), pp. B2–4, 2020.

[3] 池田雅隆, 中村喜宏. ユーザクラスタリングと CNN を用いたコーディネート推薦システムの検討. Technical Report 29, 日本大学大学院生産工学研究科, 日本大学大学院生産工学研究科, jan 2023.

[4] 大野航生, 濱崎雅弘, 中野倫靖. SNS データと Web データを活用したファッションブランド間の関係性の分析. Technical Report 8, 筑波大学, 産業技術総合研究所, 産業技術総合研究所, may 2023.

[5] 小野圭介, 太田高志. 衣服 のコーディネートを推薦するインタラクティブなシステム. 第 78 回全国大会講演論文集, 第 2016 巻, pp. 307–308, mar 2016.

[6] 佐藤彩夏, 渡邊恵太, 安村通晃. 姿を利用したファッションコーディネート支援システム suGATALOG の提案と評価. 情報処理学会論文誌, Vol. 53, No. 4, pp. 1277–1284, apr 2012.

[7] 廣瀬伸行, 白松俊, 奥原俊. 学習者の振り返りと計画を支援するアドバイスの gpt-3 による自動生成. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 1N4GS1002–1N4GS1002, 2023.

[8] 亀甲博貴, 三輪誠, 鶴岡慶雅, 森信介, 近山隆. 対数線形言語モデルを用いた将棋解説文の自動生成. 情報処理学会論文誌, Vol. 55, No. 11, pp. 2431–2440, nov 2014.

[9] 青葉紗矢香, 佐藤勇元, 益川良藏, 佐藤真, 櫨翔佑, 松井太我, 石川桂太, 高木友博. MAGMA を利用したファッションコーディ

ネートのためのコメント生成．人工知能学会全国大会論文集, Vol. JSAI2023, pp. 2K5GS204–2K5GS204, 2023.
[10] 福村望, 小尻智子．ファッションコーディネートの作成を介した教授者の感性理解支援システム．第 80 回全国大会講演論文集, 第 2018 巻, pp. 743–744, mar 2018.

DEIM Forum 2024 02-16

中長期的影響を考慮した株式の取引支援

福田　京司†　　馬　　強†

† 京都工芸繊維大学　〒 606-0951 京都府京都市左京区松ケ崎橋上町
E-mail: †b0122048@edu.kit.ac.jp, ††qiang@kit.ac.jp

あらまし　老後の資産形成のニーズや NISA の普及に伴い株式投資人口は増加傾向にある．しかしながらこうした人々の多くは投資において初心者であり，株式取引支援の重要性は高まっている．強化学習を用いた株取引を行うシステムが開発されているが，SNS 投稿やニュースなどをトリガーとしており，業績報告書等株価に中長期的に株価に影響を及ぼす情報が十分に活用されていない．そこで，本研究では，短期的に影響を与える文章に加えて決算報告書等中長期的に影響を与える文章も併用した株式取引支援の強化学習手法を提案する．本手法ではこれら 2 種類の文章から特徴量を得るネットワークを更新し，株式取引支援のための方策を獲得する．提案手法を用いて株式取引の学習，テストを実行した結果，既存手法やインデックス投資と比較して収益率が高くなることが分かった．

キーワード　投資情報学, FinTech, 強化学習, 取引支援

1　はじめに

老後の資産形成のニーズや NISA の普及に伴い株式投資人口は増加傾向にある．東証「株式分布状況調査」によると個人株主数（延べ人数）は 2014 年-2022 年の 9 年間で連続して増加し 2022 年には 6982 万人となっている [21]．また日本政府も株式投資を支援しており，2024 年には新たな NISA 制度が施行され，非課税投資枠の大幅な拡大及び制度の恒久化等が行われており，株式投資人口は今後も増加していくことが考えられる．しかしこういった個人投資家の多くは株式投資に対し十分な知識を持っておらず，効率良く投資を行うことが難しい．日本証券業協会の 2023 年の調査によると，証券投資に関する教育を受けた経験がない人の割合が 85.5%である．また金融に関する知識に関する設問,「金利が上がったら、通常、債券価格は下がる」の正答率は 48.9%であり半分を下回っている [22]．投信協会の 2020 年の調査でも投資信託の非購入理由は「投資の知識がない」が 39.5%と最も高いことが報告されている [20]．

株式取引において初心者をはじめとした投資家を支援するシステムは重要であり，様々な株式取引支援システムが研究されている．株式取引支援の分野としてはアルゴリズム取引やポートフォリオ管理や注文執行等が挙げられる [8]．特に近年では強化学習を用いた株式取引支援システムの研究・開発が活発である [17]．これらのシステムの入力としてニュースや SNS 等株価に短期的に影響を及ぼすと考えられる文章が多く用いられているが，既存研究では業績報告書等株価に中長期的に株価に影響を及ぼす情報が十分に活用されておらず，株価に短期的に影響を与える出来事と中長期的に影響を与える出来事を同時に考慮することができていない．Ramit ら [16] の手法ではツイートやニュースしか対象としていないため，中長期の影響を十分に考慮できないうえ，ニュースやツイートの少ない銘柄に対する意思決定支援にはさらなる工夫が必要である．

そこで，本研究では株価に短期的に影響を及ぼす文章（短期的文章）と株価に中長期的に株価に影響を及ぼす文章（中長期的文章）の双方を併用して株式取引の強化学習を行う手法を提案する．図 1 に提案手法の概要を示す．短期的文章と中長期的文章をそれぞれエンコーディングして，DDPG [7] における環境を構成する特徴量として利用し，株式取引を行う方策を学習する．

本研究の貢献は次のとおりである

- 短期的文章，長期的文章を併用した投資支援するための強化学習手法を提案している．これらの異種テキストを SentenceBERT [13] を用いてエンコーディングし，強化学習の環境に取り込むことで，銘柄に関する短期，中長期，双方の影響要因を考慮した方策を学習する（4 節）．
- 公開データセットに加えて業績報告書などのデータを追加収集して提案手法の検証を行なっている（5 節）．
- 既存研究の提案手法に加えて，インデックス型投資手法

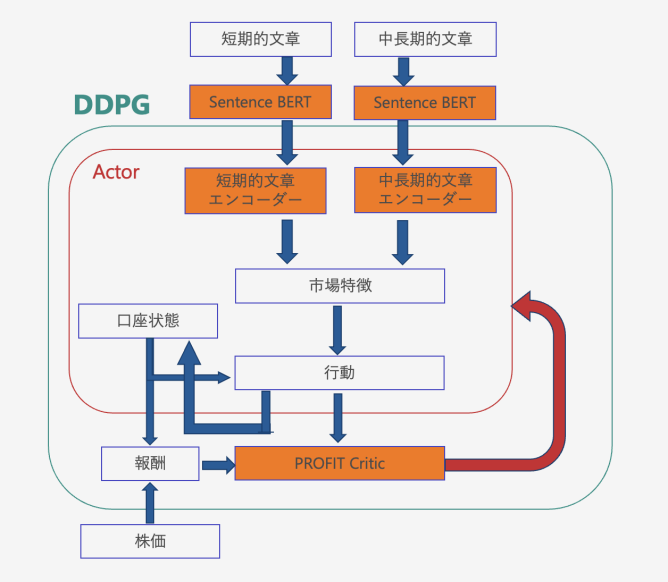


図 1　提案手法の概略図

との比較分析を行い，提案手法の実用性についても検証している（5 節）．

本論文の構成は次のとおりである．2 節で関連研究を紹介する．3 節にて用語説明，問題定義と先行研究の手法説明を行う．4 節にて本研究の提案手法を説明する．5 節にて実験の方法，結果及び考察を述べる．6 節で結言を述べる．

2 関 連 研 究

株式の取引を支援するシステムの研究は幅広く行われている．これらの研究の中には手法として強化学習を用いるものとそうでないもの，また株価やチャート等に着目したテクニカル手法と企業や国の財務，経済状況等に着目したファンダメンタル手法が存在する．

2.1 強化学習を用いない株式取引支援

2.1.1 テクニカル手法

桐畑ら [18] はトレンドの長さの閾値を用いてトレンドを検知し，金融市場の状況推定を手助けする手法を提案している．Patel らは [12] 株価変動を用いた教師あり学習を用いて株価を予測する手法を提案している．竹田ら [19] はソーシャルトレーディングサービスにおいて，取引履歴を用いた行列因子分解モデルによるトレーダーの推薦方法を提案している．また Lee ら [6] や Liu ら [9] はソーシャルトレーディングサービスにおいてポートフォリオ理論に基づきエキスパートトレーダーを発見する手法を提案している．

2.1.2 ファンダメンタル手法

坂路ら [15] は経済に関する新聞記事の構文のパターンに着目し，機械学習を用いて，因果関係を抽出する手法を提案している．酒井ら [14] は決算短信文章から業績を抽出する手法を提案している．また馬場ら [23] は数値データである株価とテキストデータであるニュース記事を観測データとして用いるヘテロトピックモデルを提案し，企業間の関係分析と株価の変動推定を行っている．

2.2 強化学習を用いた株式取引支援

2.2.1 テクニカル手法

Filos ら [4] の研究ではポートフォリオ管理における株価変動を用いた強化学習の手法を提案している．Cong ら [10] の研究では直近の株価変動と長期的なマーケットトレンドの双方を特徴量とした強化学習による株取引モデルを考案している．

2.2.2 ファンダメンタル手法

Ramit ら [16] は入力としてツイート，ニュースを用いて取引モデルの強化学習を行う手法を提案している．Gupta ら [5] も同様に入力としてツイート，ニュースを用いて市場の傾向を分析し，ポートフォリオ管理を行う強化学習の手法を提案している．

2.3 本研究について

既存研究では様々な強化学習を用いた株式取引支援が行われており，ファンダメンタル的手法ではテキストデータを用いた手法が多い．しかしこれらの多くは用いるデータがニュースや SNS 等，株価に短期的に影響を及ぼすと考えられる文章であり，中長期的な影響を十分に考慮できないことが考えられる．故に本研究ではニュースや SNS 等，株価に短期的に影響を及ぼすと考えられる文章に加え，業績報告書等，株価に中長期的に影響を及ぼすと考えられる文章の双方を用いることで，より多様な株価への影響を考慮した株式取引システムの強化学習手法を確立することを目指す．

3 予　　備

本節では，用語の説明，問題定義をしてから，本研究のベースとなる既存研究について説明する．

3.1 用　　語

本論文で用いる用語を説明する．

短期的文章　株価に短期的に影響を及ぼすと考えられる文章のことであり，ツイートなどの SNS 文章やニュースなどを指す．

中長期的文章　株価に中長期的に影響を及ぼすと考えられる文章のことであり，企業の年次業績報告書や四半期業績報告書などを指す．

文章状態 (O_m)　短期的文章，及び中長期的文章を指す．

市場特徴 (p)　短期的文章及び中長期的文章から得られる各銘柄の特徴を指す．

短期的市場特徴 (p'_s)　短期的文章から得られる各銘柄の特徴を指す．

中長期的市場特徴 (p'_l)　中長期的文章から得られる各銘柄の特徴を指す．

購入　株式取引において対象の銘柄を購入することを指す．

売却　株式取引において対象の銘柄を売却することを指す．

ホールド　株式取引において購入，売却のどちらも行わないことを指す．

行動　購入，売却，ホールドのいずれかを実行することを指す．

手数料　購入，売却時に生じる取引手数料や約定手数料等の総称を指す．

時間ステップ (τ)　取引モデルが行動を行い，その行動からネットワークが更新される一連の流れの時間単位を指す．

口座状態 (O_τ)　時間ステップ τ における取引口座の残高（現金），及び各銘柄の所持数を指す．

全体状態 (s)　口座状態と市場特徴を統合したものを指す．

初期資産　株式取引開始時における資産を指す．株取引開始時は株式を保有していないと仮定し，初期資産は取引口座の残高（現金）で構成される．

最終資産　株取引終了時における資産を指し，その時点の取引口座の残高（現金）及び保有株式の合計で構成される．

3.2 問 題 定 義

本節では本研究のシステムの入出力，及び強化学習問題について説明する．

3.2.1 システムの入出力

システムの入力はツイートやニュースなどの短期的文章，業

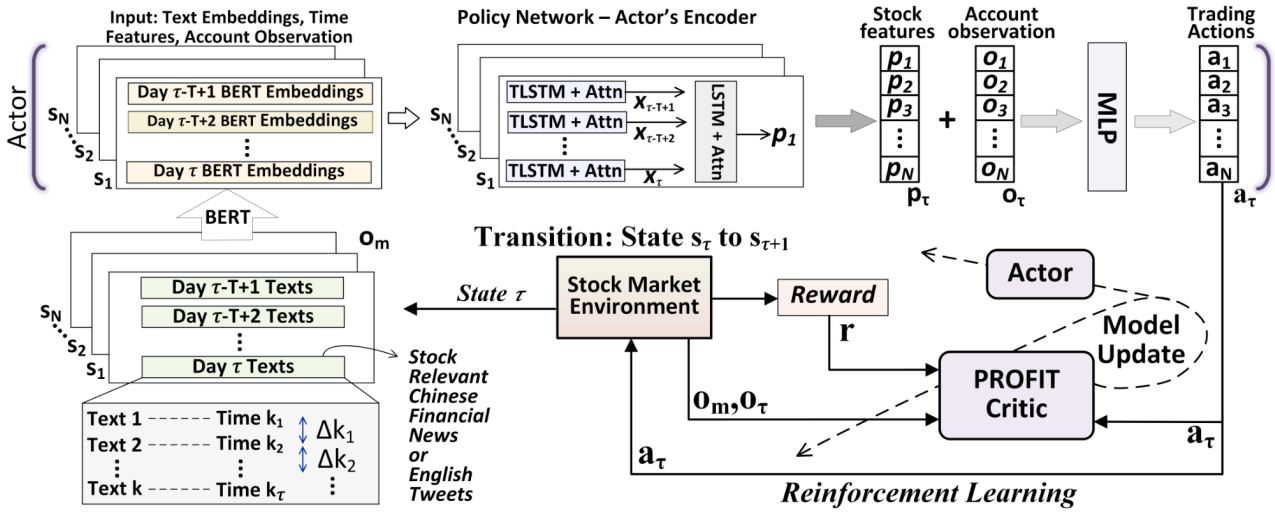


図 2　Ramit 手法のモデル概略図 ([16] から引用)

績報告書等の中長期的文章，及び株価である．取引システムは各時間ステップにおいて短期的文章，中長期的文章を元に行動を選択し，その行動の報酬が株価によって決定する．強化学習を行うことで入力から行動を出力するネットワークは更新され，次の時間ステップに進む．そしてこの過程で獲得した株式取引支援の方策が出力される．

3.2.2　強化学習問題の設定

本研究では株式取引を Ramit ら [16] の手法を元に，強化学習問題として定式化している．状態，行動，報酬は以下のように定義される．

状態 (s)　時間ステップ τ において状態 s_τ は口座状態 O_τ と文章状態 O_m から構成される．口座状態 O_τ は時間ステップ τ における口座残高と各銘柄の所有数によって構成される．文章状態 O_m は時間ステップ τ から遡って T 日間に発表された各銘柄に関連する短期的文章及び，時間ステップ τ において最も新しい N 個の中長期的文章から構成される．

行動 (a)　エージェントは時間ステップ τ において各銘柄に対し購入，売却，ホールドの 3 つの選択肢から 1 つの行動をとる．

報酬 (r)　報酬は状態 s_τ において行動をとり，新たな状態 $s_{\tau+1}$ となった時の値の変化として表される．報酬を r, 時間ステップ τ における行動を a_τ, 口座残高を b_τ, 銘柄の株価ベクトルを p_τ, 銘柄の所持数を h_τ, 取引手数料を c_τ, とすると報酬 $r(s_\tau, a_\tau, s_{\tau+1})$ を以下のように定義している．

$$r(s_\tau, a_\tau, s_{\tau+1}) = (b_{\tau+1} + p_{\tau+1}^T h_{\tau+1}) - (b_\tau + p_\tau^T h_\tau) - c_\tau \quad (1)$$

3.3　既存手法

本節では提案手法のベースとなる Ramit ら [16] の手法について説明する. なお以降本文章では Ramit らの手法を Ramit 手法と表記する．

3.3.1　モデル

Ramit 手法で提案されている取引モデルを図 3 に示す．

まずツイート等のテキストは単語単位でトークン化した後，BERT [3] に入力され，各トークンの最終隠れ層を集約しエンコードされた埋め込み表現を得る．各企業，各日付のツイート群の投稿間隔は不規則でありながら，市場に大きな影響を与えることが O ' Hara ら [11] に指摘されている．そのため Ramit 手法では，time-aware LSTM（TLSTM）[2] を使用してツイートの不規則性を捕捉する．また各ツイートが同様に有益であるとは限らずこれらが株価に多様な影響を与える可能性があることが Barber ら [1] に指摘されている．ゆえに Ramit 手法では，アテンション層を導入して，より株価に大きい影響を与える文章を強調する．こうして得られた出力に再度 TLSTM，アテンション層を用いて，遡った期間 T 日間のツイートからなる市場特徴ベクトル p_τ を得る．そして市場特徴ベクトル p_τ と口座状態ベクトル O_τ を連結し，全体状態ベクトル s_τ=$[O_\tau, p_\tau]$ を形成する．この全体状態ベクトル s_τ を順伝播型ニューラルネットワークに入力し tanh 活性化関数を用いることで，時間ステップ τ において各銘柄に対する行動 a_τ を出力する．

3.3.2　強化学習

強化学習の手法として Deep Deterministic Policy Gradient (DDPG) [7] が用いられる．Ramit 手法では行動とその行動の評価を Actor と PROFIT Critic に分離する．Actor は取引モデルの埋め込状態 s_τ を入力とし、行動 a_τ を出力する．PROFIT Critic は状態 s_τ と行動 a_τ を入力として，行動 a_τ を評価するスカラー値 $Q(s_\tau, a_\tau)$ を出力する．各状態 s_τ で Actor は a_τ を実行し，報酬 r_τ を受け取り，次の状態 $s_{\tau+1}$ に到達する．これらの遷移は $(s_\tau,\ a_\tau\ ,\ s_{\tau+1},\ r_\tau)$ としてバッファD に格納される．その後，モデルを更新するために，N 個の遷移からなるミニバッチ B が D からサンプリングされる．各バッチ B について，DDPG は ϕ に関して以下の損失 L を最小化し，PROFIT Critic を次のように更新する．

$$y_\tau = r_\tau + \gamma Q^{\phi'}(s_{\tau+1}, \mu^{\theta'}(s_{\tau+1})) \quad (2)$$

$$L = E[(y_\tau - Q^{\phi}(s_\tau, a_\tau))^2] \quad (3)$$

ここで，y_τ は更新された Q 値，γ は割引係数，θ と θ'，μ は

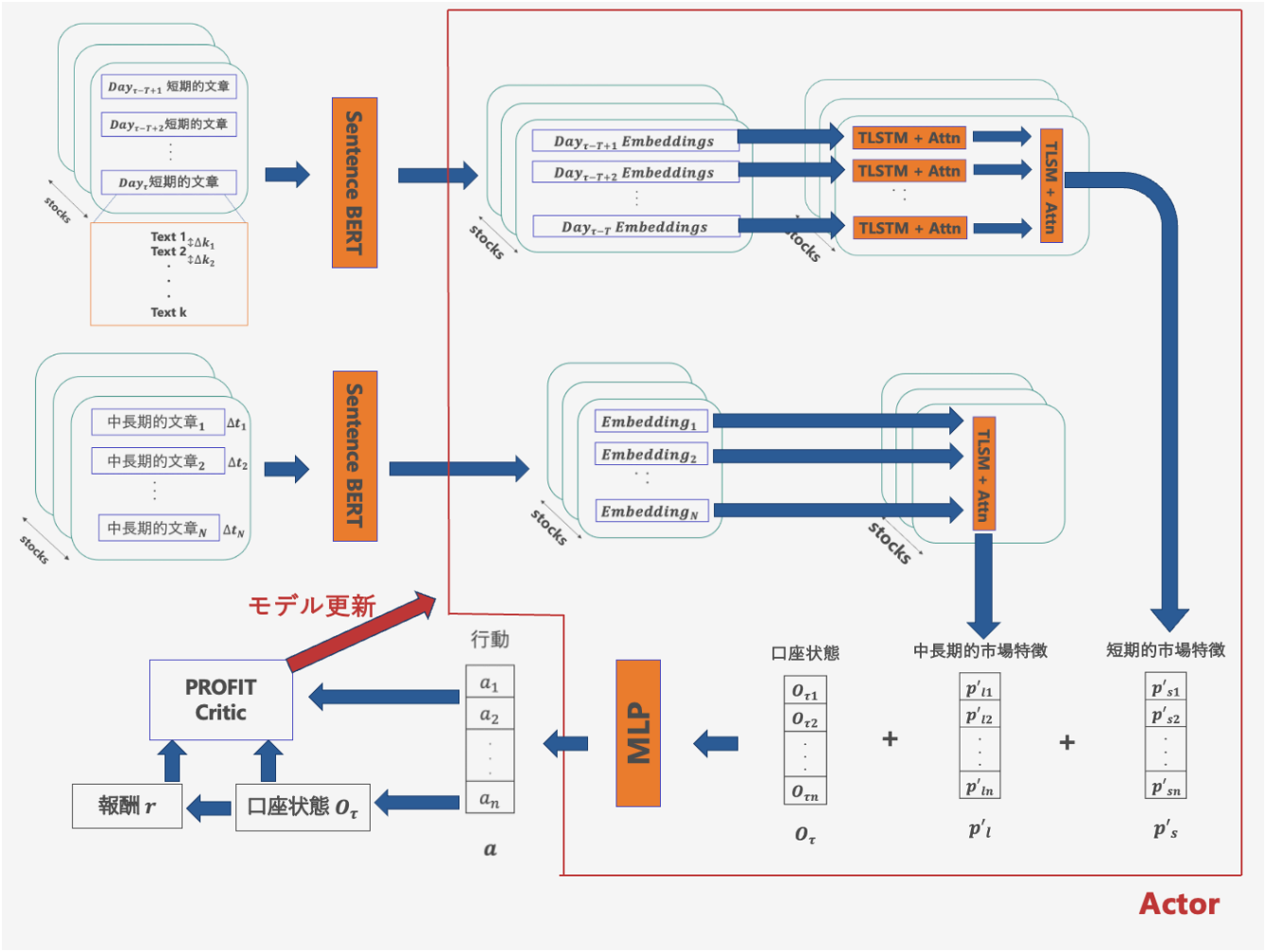


図 3　本研究の提案手法

Actor, Q は PROFIT Critic，ϕ と ϕ' はそれぞれ μ と Q の 2 つのコピーパラメータである．またポリシーの勾配 $\nabla_\theta J$ を用いて，Actor を次のように更新する．

$$\nabla_\theta J = E[\nabla\ aQ^\phi(s_\tau, \mu\theta(s_\tau))\ \nabla\theta\mu\theta(s_\tau)] \tag{4}$$

4　提 案 手 法

本節では本研究の提案手法について述べる．本研究では既存研究で用いられたツイート等株価に短期的に影響を及ぼす文章に加え，業績報告書等株価に中長期的に影響を及ぼす文章を入力に追加し，その双方を用いて株取引の強化学習を行うモデルを提案する．本研究の提案手法のモデルを図**??**に示す．

4.1　エンコーディング

入力のテキストをエンコーディングして埋め込み表現を得る．Ramit 手法ではツイートを単語単位でトークン化した後 BERT [3] を用いていたが，業績報告書のように分量の多い文章では単語単位のトークン化では文章間のつながりが反映されない．よってこの文章間のつながりを表現するため，本研究では SentenceBERT [13] を用い，文章全体を SentenceBERT に入力しその最終隠れ層から埋め込み表現（Embedding）を得る．短期的文章，中長期的文章でエンコーディングの方法を SentenceBERT に統一している．

4.2　特徴量の生成

埋め込み表現から短期的市場特徴 p'_s，中長期的市場特徴 p'_l を得る手法を説明する．

短期的文章　取引日から T 日間遡った短期的文章は Ramit 手法と同様に，TSLTM とアテンションを用いて，Ramit 手法の p_τ に対応するものを短期的市場特徴 p'_s として得る．

中長期的文章　中長期的文章は短期的文章と異なり文章の発表間隔が長いため，取引日によって文章発表から経過した時間差 Δt が生じる．この差による株価への影響を考慮するため，TLSTM [2] を使用して，取引日に応じた中長期的文章の影響を捕捉する．また中長期的文章においても短期的文章と同様，遡った N 個の各報告書が同様に有益であるとは限らないことによる株価への影響が考えられる [1]．そのため N 個の各報告書間でアテンション層を導入して株価により大きい影響を与える文章を強調し，遡った N 個の業績報告書から中長期的市場特徴 p'_l を得る．

4.3　行動の出力

短期的市場特徴 p'_s，中長期的市場特徴 p'_l，口座状態 O_τ を連

結し，全体状態 z_τ=[O_τ, p'_s, p'_l] を形成した後，MLP と活性化関数 tanh を経て、時間ステップ τ において各銘柄 s に対する行動 a_τ を出力する．

4.4 強化学習

本研究では Ramit 手法と同様に DDPG を用いて強化学習を行う．PROFIT Critic, Actor はそれぞれ (2) 式，(3) 式で更新する．

5 実　　験

米国の株式市場である NASDAQ において取引量の多い 20 銘柄に対し，提案手法の学習，テストを行い，その結果を分析する．本節では初めに研究目的を述べ，データセット，実験設定，評価指標と比較手法を説明した後，実験結果を示し，そこから得られた考察を述べる．

5.1 目　　的

提案手法が既存手法や他の手法等と比較し，どのような特徴を持つのかを調査し，株式取引における優位性の有無等を判断する．また同一の提案手法に対し取引期間や初期資産額を変更して結果を比較し，どのような状況下において本提案手法が有効に機能するかについて調査する．

5.2 データセット

本実験で入力データとして用いた短期的文章，中長期的文章，株価について説明する．

短期的文章　本実験では短期的文章として Xu ら [1]によって収集された，旧 Twitter の 2014 年 12 月から 2015 年 12 月の各銘柄に関するツイートを使用する．ツイートは全て英語であり，各銘柄，各日付ごとに最大 5 つのツイートが含まれる．ツイートのうち 2014 年 12 月から 2015 年 9 月までを訓練用，2015 年 10 月から 2015 年 12 月までをテスト用として用いる．

中長期的文章　本実験では中長期的文章として米国証券取引委員会 [2]で公開されている，2014 年 6 月から 2015 年 12 月の各銘柄の年次業績報告書 (以下 10-K)，及び四半期業績報告書 (以下 10-Q) を使用する．これらの文章には Legal Proceedings 章など株価に影響を与えないと考えられる箇所が存在するため，本実験では 10-Q, 10-K から Management ' s Discussion and Analysis of Financial Condition and Results of Operations 章など特定の章を入力として用いた．10-K, 10-Q のうち 2014 年 6 月から 2015 年 9 月までを訓練用，2015 年 4 月から 2015 年 12 月までをテスト用として用いる．

株価　本実験では株価として各銘柄，各日付の終値を使用する．学習，テストにおける株式の売買は終値で行われる．

5.3 実験設定

本実験に際して設定した項目を以下で説明する．

1：https://github.com/yumoxu/stocknet-dataset/tree/master
2：https://www.sec.gov

5.3.1 取引条件

本実験では米国株式市場 NASDAQ において取引量の多い 20 銘柄を取引の対象とした．対象とした銘柄の一覧を表**??**に示す．株式取引における手数料はブローカーによって異なるが，本実験においては楽天証券 [3]を参考に，取引時に生じる手数料を取引額の 0.495%と仮定する．また株式取引においてレバリッジをかける等の場合口座残高がマイナスになるケースが存在するが，本実験は口座残高がマイナスにならないことを前提とする．取引において購入額が口座残高を上回る場合，その購入を中止し，行動はホールドとする．

5.3.2 ハイパーパラメータ

文章状態 O_m において短期的文章を遡る日数 T を Ramit 手法 [16] と同じく 7 日，また遡る中長期的文章の数 N を 1 とする．強化学習においてモデルを更新するためのミニバッチのサイズを 1，バッファのサイズを 25 とする．訓練のステップ数を 27200 ステップ（100 エポック）とする．2015 年 1 月 1 日から 2015 年 9 月 30 日までを訓練期間，2015 年 10 月 1 日から 2015 年 12 月 31 日までをテスト期間とする．訓練における初期資産額を 100000 米ドルとする．

5.3.3 テストの種類

各手法に対し以下の取引期間，初期資産額でテストを行う．

取引期間　1 週間，1ヶ月，2ヶ月，3ヶ月の取引期間でそれぞれテストを行う．各取引期間において，取引開始日を 2015 年 10 月 1 日を起点として，10 月 8 日，10 月 15 日，と 1 週間単位でずらしていき，各開始日ごとにテストを行い，それらの結果の平均をその取引期間におけるテストの結果とする．

初期資産額　10000, 100000, 1000000, 10000000 米ドルでテストを行う．

5.3.4 評価指標

テストの結果を評価するために以下の評価指標を用いる．

収益率

収益率は初期資産に対して取引終了後の資産の変化率を示す．初期資産を b_0，最終資産を b_f として収益率 ROI は次のように表される．

$$ROI = \frac{b_f - b_0}{b_0} \tag{5}$$

シャープレシオ

シャープレシオは運用で取ったリスクに見合うリターンを上げたかどうかを測る指標であり，この数値が大きいほど投資効率が高いことを示す。手法による取引利益を R_a，無リスク資産による利益を R_f としてシャープレシオ SR は次のように表される．なお本実験では R_f の値を 0 としている．

$$SR = \frac{E[R_a - R_f]}{std[R_a - R_f]} \tag{6}$$

最大ドローダウン

最大ドローダウンは取引期間における資産の最大の下落率を示し，この値が小さいほど取引におけるリスクが低いことを示す．取引期間における資産の最大額を b_{max}，最小額を b_{min} と

3：https://www.rakuten-sec.co.jp

して最大ドローダウン MDD は次のように表される．

$$MDD = \frac{b_{max} - b_{min}}{b_{min}} \quad (7)$$

5.4 比較手法

実験にあたって複数の手法で得られた方策を用いてテストを行う．その手法は以下の通りである．

（1） T 手法：短期的文章のみを用いる Ramit 手法

（2） S 手法：中長期的文章のみを用いる提案手法

（3） TSt 手法：短期的文章，中長期的文章の双方を用い，取引日と中長期的文章発表日の時間差を考慮する提案手法

（4） TS 手法：短期的文章，中長期的文章の双方を用い，取引日と中長期的文章発表日の時間差を考慮しない提案手法

5.5 実験結果

5.4 節の各手法に対し，5.3.3 節で示した各期間，各初期資産額でテストを実行した結果を本節で示す．

5.5.1 手法間の比較

取引期間 1ヶ月，初期資産額 100000 米ドルで手法 T，手法 S，手法 TSt，手法 TS に対しそれぞれテストを実行した．この結果を表 1 に示す．収益率は TSt, TS, S, T の順に高く，提案手法が既存手法や他の手法と比較して高い収益を上げていることが確認できる．一方最大ドローダウンは S, T, TSt, TS の順に小さく，提案手法が最もリスクをとっていることが確認できる．そしてシャープレシオは TSt1, TS, SR, T の順に高く，提案手法が既存手法や他の手法と比較してリスクに対して大きなリターンを得ていることが確認できる．

表 1　手法間の比較結果

手法	ROI	SR	MDD
T	-0.0067 ± 0.0018	-0.24 ± 0.15	-3.28 ± 0.08
S	-0.0015 ± 0.0023	0.05 ± 0.17	-3.03 ± 0.08
TSt	0.0296 ± 0.0043	1.96 ± 0.25	-3.68 ± 0.11
TS	0.0275 ± 0.0039	1.95 ± 0.23	-3.37 ± 0.09

5.5.2 取引期間に応じた比較

初期資産額 100000 米ドルで手法 TSt に対し，取引期間を 1 週間，1ヶ月，2ヶ月，3ヶ月にしてそれぞれテストを実行した．この結果を表 2 に示す．収益率は 3ヶ月，2ヶ月，1ヶ月，1 週間の順に大きく，取引期間が長いほど収益率が高くなることが確認できる．シャープレシオは 1 週間，3ヶ月，2ヶ月，1ヶ月の順に大きい．最大ドローダウンは 1 週間，1ヶ月，2ヶ月，3ヶ月の順に小さく，取引期間が長いほどリスクが大きいことが確認できる．

表 2　取引期間に応じた比較結果（TSt 手法）

取引期間	ROI	SR	MDD
1 週間	0.0071 ± 0.0009	2.91 ± 0.32	-0.35 ± 0.02
1 ヶ月	0.0296 ± 0.0043	1.95 ± 0.25	-3.68 ± 0.11
2 ヶ月	0.0756 ± 0.0463	2.64 ± 1.29	-4.38 ± 0.54
3 ヶ月	0.1364	2.87	-5.32

5.5.3 初期資産額に応じた比較

取引期間 1ヶ月で手法 TSt に対し，初期資産額を 10000, 100000, 1000000, 10000000, 米ドルにしてそれぞれテストを実行した．この結果を表 3 に示す．収益率は 10000, 100000, 1000000, 10000000, 米ドルの順に大きく，初期資産額が小さいほど収益率が高くなることが確認できる．一方最大ドローダウンは 10000000, 1000000, 100000, 10000, 米ドルの順に小さく，初期資産額が小さいほどリスクが大きいことが確認できる．そしてシャープレシオも同様に 10000, 100000, 1000000, 10000000, 米ドルの順に大きく，初期資産額が小さいほどリスクに対して大きなリターンを得ていることが確認できる．

表 3　初期資産額に応じた比較結果

初期資産額	ROI	SR	MDD
10000.0	0.0457 ± 0.0059	2.37 ± 0.28	-4.23 ± 0.12
100000.0	0.0296 ± 0.0043	1.95 ± 0.25	-3.68 ± 0.11
1000000.0	0.0056 ± 0.0011	1.88 ± 0.26	-0.95 ± 0.05
10000000.0	0.0006 ± 0.0001	1.87 ± 0.26	-0.1 ± 0.0

5.6 考　　察

5.5.1 節において TSt 手法, TS 手法が T 手法, S 手法よりも収益率が高いことは，短期的文章，中長期的文章の双方が株価に影響を与えていることを示唆している．このことから，株価予測において短期的文章，中長期的文章の双方を用いることが，どちらか一方を用いるよりも，精度向上につながることが考えられる．

5.5.1 節において TSt 手法が TS 手法よりも収益率が高いことは，中長期的文章の発表日からの時間経過が株価に影響を与えることを示唆している．このことから，株価予測において中長期的文章を用いる場合は取引日に対する発表日からの時間経過も同時に考慮することが精度向上につながると考えられる．

5.5.2 節において取引期間が長いほど収益率が高い結果は，中長期的文章を用いることが株価の中長期的値動きの捕捉に寄与することを示唆している．比較として初期資産額 100000 米ドルで T 手法に対し，取引期間を 1 週間，1ヶ月，2ヶ月，3ヶ月にしてテストを実行した結果を表 4 に示す．表 4 節より T 手法においては取引期間による収益率の差は最大で 0.0077 であり，また取引期間と収益率に相関関係は見られない．このことからツイートのような短期的文章は数日単位でしか株価に影響を与えず，数ヶ月単位の値動きを予測することに寄与していないことが考えられる．対して表 2 節より TSt 手法においては取引期間が 3ヶ月の場合と 1 週間の場合を比較すると収益率は 0.13 以上大きく，また取引期間と収益率に正の相関がある．このことから中長期的文章は株価に数ヶ月単位で影響を与えることが考えられ，この期間の取引において特に有効であると言える．

5.5.3 節において初期資産額が小さいほど収益率が高いことから，本提案手法は企業投資家よりも個人投資家に適していると言える．しかしこの結果は訓練における初期投資額が 100000 米ドルであったことが寄与している可能性が考えられる．

株式投資初心者に人気のある投資オプションとしてインデッ

表 4　取引期間に応じた比較結果（T 手法）

取引期間	ROI	SR	MDD
1 週間	0.001 ± 0.0004	1.57 ± 0.3	-0.15 ± 0.01
1 ヶ月	-0.0067 ± 0.0018	-0.24 ± 0.15	-3.28 ± 0.08
2 ヶ月	-0.0038 ± 0.0139	0.22 ± 0.41	-4.5 ± 0.09
3 ヶ月	-0.0004	0.3	-6.27

クス投資が挙げられる．インデックス投資は複数銘柄や市場全体の指数と連動するように投資を行う方法で，銘柄の選定が不要であるため，初心者でも利用がしやすい．本手法の取引支援としての有意性を考察するため，TSt1 手法，初期資産額 10000 米ドル，取引期間 3ヶ月における収益率推移と、テストで用いた 20 銘柄の平均株価の取引開始日（2015 年 10 月 1 日）を 0 としたときの変化率推移を図 4 に示す．図 4 より，取引終了時点において提案手法の収益率がテストで用いた 20 銘柄の平均株価の変化率を上回っている．このことから提案手法はインデックス投資と比較してより収益を得ることが可能な手法であると言える．

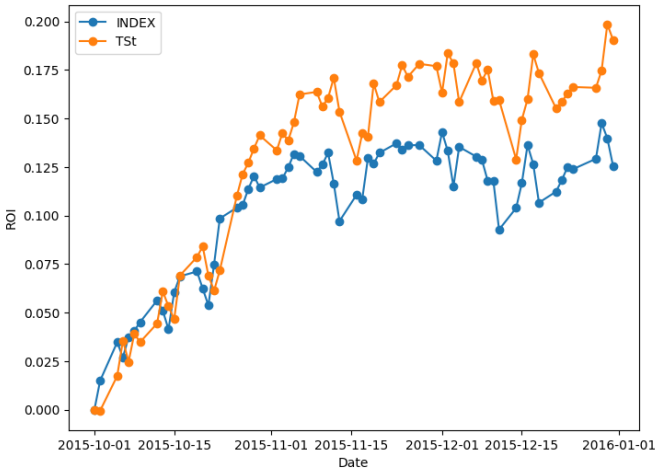


図 4　TSt 手法の収益率推移及び平均株価の変化率推移

6　ま　と　め

本研究ではニュースや SNS 等の短期的文章に加え，業績報告書等の中長期的文章の双方を用いた取引モデルの強化学習手法を提案している．提案手法では株価に短期的文章，中長期的文章のそれぞれで特徴量を出力し，それらを結合することで双方の影響を考慮できるようにしている．提案手法を用いることで既存手法やインデックス投資と比較し，高い収益を得ることができた．また提案手法は数ヶ月間の取引期間において特に高い収益を得ることができることが分かった．

本研究の課題と展望は以下の通りである．

- 本研究の実験において対象とした銘柄の平均株価は期間内で上げ相場であったが（図 4），下げ相場や値動きが小さい相場において提案手法がどのような結果を示すかについて考察していく
- 文章状態 O_m において遡る中長期的文章数 N が 2 つ以上の場合や，3ヶ月より長い取引期間の場合，また訓練時の初期資産が 100000 米ドルでない場合において提案手法がどのような結果を示すか検討していく
- 本研究の実験において対象とした銘柄は取引量が多く，各日付に複数のその銘柄に関する短期的文章が存在したが，取引量が少なく，その銘柄に関する短期的文章が少ないような銘柄に対し提案手法がどのような結果を示すか検討していく

謝　　辞

本研究の一部は科研費（23H03404）による．

文　　献

[1] B. M. Barber and T. Odean. The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2):pp.785–818, 12 2007.

[2] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages pp.65–74, 8 2017.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 10 2018.

[4] A. Filos. Reinforcement learning for portfolio management. *CoRR*, abs/1909.09571, 2019.

[5] P. Koratamaddi, K. Wadhwani, M. Gupta, and S. G. Sanjeevi. Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Engineering Science and Technology, an International Journal*, 24(4):pp.848–859, 8 2021.

[6] W. Lee and Q. Ma. Whom to follow on social trading services? A system to support discovering expert traders. In *proceedings of Tenth International Conference on Digital Information Management*, pages pp.188–193, 10 2015.

[7] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *proceedings of 4th International Conference on Learning Representations*, pages pp.159–169, 7 2016.

[8] Y. Liu, Q. Liu, H. Zhao, Z. Pan, and C. Liu. Adaptive quantitative trading: An imitative deep reinforcement learning approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(2):pp.2128–2135, 3 2020.

[9] Z. Liu and Q. Ma. Portfolio-based ranking of traders for social trading. *IEEE Access*, 8:pp.145363–145371, 8 2020.

[10] C. Ma, J. Zhang, J. Liu, L. Ji, and F. Gao. A parallel multi-module deep reinforcement learning algorithm for stock trading. *Neurocomputing*, 449:pp.290–302, 8 2021.

[11] M. O'Hara. High frequency market microstructure. *Journal of Financial Economics*, 116(2):pp.257–270, 5 2015.

[12] J. Patel, S. Shah, P. Thakkar, and K. Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Syst. Appl.*, 42(1):pp.259–268, 9 2015.

[13] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages pp.3980–3990, 3 2019.

[14] H. SAKAI and S. MASUYAMA. Assigning polarity to causal information in financial articles on business performance of companies. *IEICE Transactions on Information*

and Systems, E92.D(12):pp.2341–2350, 4 2009.

[15] M. Sakaji, Sekine. Extracting causal knowledge us- ing clue phrases and syntactic patterns, international conference on prac- tical aspects of knowledge management. *proceedings of International Conference on Practical Aspects of Knowledge Management*, pages pp.111–122, 2 2008.

[16] R. Sawhney, A. Wadhwa, S. Agarwal, and R. R. Shah. Quantitative day trading from natural language using reinforcement learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages pp.4018–4030, 6 2021.

[17] S. Sun, R. Wang, and B. An. Reinforcement learning for quantitative trading. *ACM Transactions on Intelligent Systems and Technology*, 14(3):pp.1–29, 6 2023.

[18] 桐畑誠 and 馬強. トレンドを考慮した投資信託商品の要因分析に関する研究. *DEIM Forum 2019 D8-3*, 3 2019.

[19] 竹田創 and 馬強. ソーシャルトレーディングサービスにおけるトレーダの特徴分析. *DEIM Forum 2017 C7-2*, 3 2017.

[20] 投資信託協会. 投資信託に関するアンケート調査 調査結果サマリー. *https://www.toushin.or.jp/statistics/report/research2021/*, 32, 2021.

[21] 日本証券業協会. 株式分布状況調査. *https://www.jsda.or.jp/shiryoshitsu/toukei/2022kozinkabunusidoukou.pdf*, 16, 2023.

[22] 日本証券業協会. 個人投資家の証券投資に関する意識調査. *https://www.jsda.or.jp/ shiryoshitsu/toukei/2023kozintoushika.pdf*, 43, 2023.

[23] 馬場慧 and 馬強. 株価とニュースの統合分析のためのヘテロトピックモデル. *DEIM Forum 2017 F5-3*, 3 2017.

食品ロス削減を目的とした小売店舗の割引情報・料理レシピ提示システムの実現

永浦 実樹† 阿部 広夢† 小林 千沙希† 石橋 直樹† 佐々木 史織†

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

E-mail: †{s2122082, s2122072, s2122025}@stu.musashino-ac.jp, {n-ishi, ssasaki}@musashino-u.ac.jp

あらまし 本稿では，ユーザが投稿した小売店舗のタイムセール食材の割引商品情報をもとに，ユーザの現在地に近い店舗の割引商品情報を提案すると共に割引食材を使用して調理できる可能性の高い料理レシピを推薦するシステムの実現方法を示す．本システムのユーザは，割引商品情報を投稿・共有・検索する機能を用いて，リアルタイムでのタイムセール食材情報へアクセスし，提案された近隣店舗の食材の調達が可能となる．本システムの特徴は，従来の店舗情報提示システムや料理レシピ推薦システムとは異なり，ユーザのニーズに合致した食材とレシピの組み合わせの提案を可能とすることにより，現在地から近い店舗のタイムセール食材を購入すると共に，割引商品情報をユーザ間で共有することが可能となる．

キーワード 食品ロス，持続可能性，リコメンドシステム，位置情報，データベース

1. はじめに

近年，世界中で食品ロスが問題となっている．食品ロスとは，本来食べられるにも関わらず捨てられてしまう食品のことである[1].

日本では令和 3 年度における食品ロスが約 523 万トン発生したと推計されている．このうち，事業系食品ロス量は約 275 万トン，家庭系食品ロス量は約 247 万トンという内訳になっている．一方，世界では，同年度における栽培・生産された全食品のうち約 40%となる 25 億トンもの食品が年間で廃棄されている．

食品ロスは，平成 27 年 9 月に国際連合にて採択された「持続可能な開発のための 2030 アジェンダ」で定められている，「持続可能な開発目標」 (Sustainable Development Goals : SDGs)のターゲットの 1 つに，2030 年までに小売・消費レベルにおける世界全体の一人当たりの食品廃棄物を半減させることが盛り込まれており，国際的な食品ロス削減の促進が近年高まっている．

これらの食品ロスの削減に対する取り組みとして ecoeat と呼ばれるショップが存在する[2]．この ecoeat では，廃棄予定の飲料や食品の買い取り，または，無償で引き取り，その中から賞味期限の残存の有無に関わらず，安全且つ，美味しく食べられる食品のみを店舗で陳列する取り組みを行なっている．また，販売だけでなく，ecoeat が販売している商品を購入することによって，生活に困っている人々に対して食糧無償支援を行っている．

これらの取り組みによって，本来であれば廃棄となってしまう食品を販売することによって，食品ロスを削減することが可能になる．しかし，ecoeat の運営店舗は全国で 20 店舗しか展開されておらず，日本全体で普及しているとは言えない状況である．

また，食品ロス削減に向けた取り組みとして，家庭内で余っている食品を持ち寄り，福祉団体や施設に提供するフードドライブという活動が行われている[3].

フードドライブでは，持ち込むことが可能な食品の種類や条件があり，それらを満たした食品のみを寄付することが可能である．東京都江東区では現在，17 箇所の常設回収窓口での回収が行われている．また，窓口での回収だけでなく，江東区内で行われるイベントでの実施も行われている．フードドライブの取り組みは東京都江東区だけでなく，他の道府県の市区町村でも実施されている．

これらの紹介した 2 つの取り組みは家庭内で捨てられてしまう可能性のある食品を持ち寄り，廃棄予定の食品を廃棄することなく，食品を求める人々のもとに届けることによって，食品ロス削減に繋げることが可能であるという目標を掲げている．2 つの取り組みの共通点として，商業施設で購入した食品が廃棄されてしまうこと，つまり，家庭系食品ロス量の削減を目標としている．

上記の 2 つの取り組みのように，家庭系食品ロス量に着目している取り組みが多く存在しているが，日本の現状は前述で示したように，食品ロスのうち事業系食品ロス量は家庭系食品ロス量を上回る全体の約 52%を占めている．

そこで，本研究では，消費者目線ではなく販売者目線から考え，店舗で販売されている食品の中でも当日に売り切る必要がある生鮮食品や賞味期限が近い商品

の販売を促進するシステムを提案する．具体的には，ユーザが訪れた店舗に販売されている割引商品をデータベースに登録することによって，それを閲覧した他のユーザが商品を購入することを促すことが可能であり，かつ，登録された複数の商品から調理することの出来る料理をリコメンドするシステムの構築を実現した．

本システムのユーザは，ユーザが登録した食品，または他のユーザが登録した食品をリアルタイムで閲覧することが可能である．閲覧する際に表示される食品は，登録された時間のタイムスタンプとともに表示されるようになっており，そこから 2 時間以上が経過すると自動的にそのデータは削除される仕組みになっている．

また，登録された食品から作ることの出来る料理をリコメンドする際は，必ずその食品が使用されている料理が表示される仕組みとなっている．

本稿は，次の構成からなる．2 節では，関連研究について示す．3 節では，本方式である小売店舗の割引商品の提案及び料理リコメンドシステムについて示す．4 節では本方式を実現する実験システムを構築し，5 節では，システム評価実験の概要とその結果について示す．6 節では考察，7 節では結論と展望について示す．

2. 関連研究

2.1 食品ロス削減に向けた取り組みに関する研究

本節では，食品ロス削減に向けた取り組みに関する研究について 2 つ述べる．

はじめに，スーパーマーケットにおける食品ロス削減に向けた需要予測に関する研究である．

桐生ら[4]は，スーパーマーケットで販売される食品のうち，販売数量の日変動が大きく，売上予測が難しいとされるおにぎりに着目し，深層学習を用いて販売数量の予測精度向上に取り組んだところ，需要予測により，約 1 割程度の誤差で販売数量が推定することが可能であると示した．この結果から，深層学習を適用することにより，販売商品の余剰分を廃棄する量が減少し，食品ロス削減に貢献できる可能性があると示唆した．

続いて，購買履歴に基づいた位置情報つき食品レコメンドシステムに関する研究である．

本節では，ユーザの購買履歴と位置情報をもとに，パーソナライズされた値引き食品のレコメンドを行うシステムの研究について示す．

川田ら[5]は，購買履歴と位置情報をもとに，食品の購買履歴と位置情報をもとに，食品の種類別の購入周期を計算し，ユーザの現在地と現在時刻に近い店舗で販売されている食品の中で，ユーザが最も購入しそうな割引商品をリコメンドすることによって，賞味期限切れにより，廃棄されてしまう食品を削減することを目指している．

2.2 ユーザの現在地情報とデータベースを使用したシステムに関する研究

本節では，ユーザの現在地情報とデータベースを使用したシステムに関する研究について 2 つ述べる．

はじめに，乗車中の列車の各駅到着時間を自動表示するスマートフォンアプリケーションに関する研究である．

井上ら[6]は，列車に乗車しているユーザのスマートフォン端末を用いて，自動的に乗っている列車を推定し，ユーザに到着時間情報を提示するアプリケーションの開発および乗り物の推定手法の提案を行った．この研究では，列車に乗っているユーザは，アプリケーションを開くだけで，何も操作せずに列車の情報と到着時間情報を得ることが可能となっている．

続いて，災害弱者を考慮した避難共助アプリケーションに関する研究である．

佐々木ら[7]は，災害における犠牲者数を抑制するため，地域住民がスマートフォン端末を使用することで，Map 上で災害弱者の位置を確認し，SNS 上で援護者が災害弱者を救助に向かい，共助することが可能なシステムの開発を行った．SNS 上で災害時にユーザがコメントの投稿をすることによって，援助者がそのコメントとともに，GPS で取得したユーザの位置情報を閲覧することが可能となる仕組みである．

2.3 本研究の位置付け

本方式では，ユーザが日常的に利用する店舗で販売されている食品のうち割引商品を対象とした割引商品情報を登録することにより，全ユーザが登録した食品の閲覧が可能なシステムを実現し，それらの情報からユーザの現在地に近い小売店舗の食品を組み合わせることによる，料理リコメンドシステムを実現する．

2.1 では，スーパーマーケットにおける売上予測が難しいとされているおにぎりの需要予測によって，販売数量を適切に設定し，廃棄を減らすことが可能であることを示した．また，ユーザの食品購入頻度と位置情報を用いて，パーソナライズされた店舗の値引き食品の推薦を実現している．本研究では，ユーザの履歴情報は一切使用せず，リアルタイムで販売されている割引商品を対象に料理を推薦することによって，ユーザ，つまり消費者だけでなく，販売者にとっても食品ロス削減が可能なシステムの実現を目指す．

2.2 ではスマートフォン端末の GPS 機能を用いて位置情報を取得することによって，乗っている列車の到

着時間情報を推定するシステムと，災害時において，災害弱者のコメント投稿に基づき地域住民が援助を行うシステムについて示した．本研究では，ユーザの位置情報に基づき，店舗の割引商品を推薦するだけでなく，現在地から近い小売店舗の割引商品を使って調理できる料理推薦の実現を目指す．

3. 料理リコメンドシステムの実現方式

本節では，提案方式である，料理リコメンドシステムの実現方式について提示する．

本システムは，ユーザ自身が割引商品情報を入れることによって，現在時刻から直近で 2 時間以内に登録された情報をもとに，現在置から近い位置にある複数の店舗で販売されている割引商品を使用して調理できる料理を，複数推薦する．

本節では，本システムで使用しているデータベースの概要と本研究における提案手法の概要を述べる．提案システムの概要と機能群を図 1 に示す．

本システムでは，ユーザの現在地，現在時刻，食品販売店舗の位置情報，割引商品の商品情報，料理に使用する食品のデータを用いる．

また，本システムでは，1) 本システムの概要，2) 本システムで使用するデータベースの概要，3) 距離計算による店舗選択機能，4) 料理推薦機能で構成される

以下に各機能の詳細について述べる．

3.1. 本システムの概要

本システムの導入は図 1 に示したとおり，ユーザの現在地を取得することから始まる．システム上でユーザの緯度経度の取得が可能である．

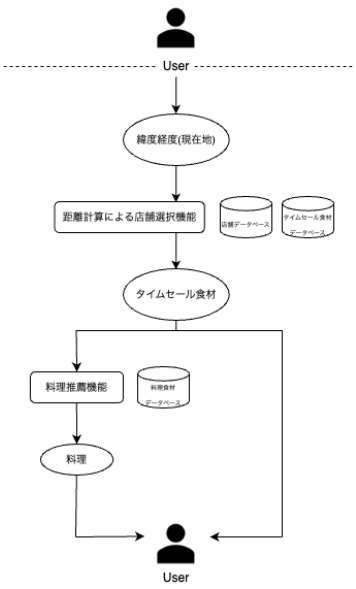


図 1 本システムの概要

そこから，距離計算による店舗選択機能により，緯度経度が近似している近隣店舗の情報と，その店舗で販売されているタイムセール食材が表示される．また，料理推薦機能により，表示されたタイムセール食材と料理食材データベースに格納されている料理とその料理に使用される食材に基づき，調理できる可能性が高い料理が表示され，ユーザにその結果が提案される仕組みになっている．

3.2. 本システムで使用するデータベースの概要

本実現方式では，PostgreSQL を用いてデータベースを作成する．作成するデータベースは 3 つあり，1) 店舗データベース，2) 割引商品データベース，3) 料理食材データベースである．

はじめに，店舗データベースのデータ内容を図 2 に示す．

店舗データベースには，店舗 ID，店舗名，緯度，経度が格納されている．

次に，タイムセール食材データベースの内容を図 3 に示す．タイムセール食材データベースには，割引商品 ID，ユーザ ID，食品 ID，店舗 ID，食品名，店舗名，食品の値段，割引率，割引適用後の値段，店舗の緯度経度のデータが格納されている．図 3 で示した内容は，割引商品データベースから一部抜粋したものである．

次に，料理食材データベースのデータ内容を図 4 に示す．

sid	shop_name	lat	lng
1	イオン東雲店	35.6488209894	139.80215484606
2	業務スーパー立川錦町店	35.688100089616	139.41822143638
3	いなげや花小金井駅前店	35.7270499	139.5134478
4	イオンスタイル有明ガーデン店	35.638421496626	139.79295554416

図 2 店舗データベース

tid	uid	fid	sid	food	shop	price	discount	real_price	time
51	13	26	3	えび	いなげや花小金井駅前店	200	5	190	2023-11-06 22:06:26.980128
52	13	79	3	みりん	いなげや花小金井駅前店	350	5	332	2023-11-06 22:06:50.446983
53	13	10	3	豚レバー	いなげや花小金井駅前店	200	5	190	2023-11-06 22:07:43.22633

(3 行)

図 3 タイムセール食材データベース

fid	food_name
1	豚肉
2	牛肉
3	合いびき肉
4	鶏肉
5	牛スジ肉
6	豚ひき肉
7	ラム肉
8	鶏ひき肉
9	鶏ささみ
10	豚レバー
11	ハム
12	さんま
13	きんき
14	いか
15	いわし
16	あさり
17	ぶり
18	あじ
19	鯛
20	たら
21	鮭
22	鰹
23	まぐろ
24	いくら
25	たけのこ
26	えび
27	カニカマ

図 4 料理食材データベース

図 4 で示した内容は料理食材データベースから一部抜粋したものである．料理食材データベースには，食品 ID，食品名が格納されている． 料理食材データベースに格納されている料理とそれに対応する食材のデータは，AJINOMOTO PARK のサイトに掲載されているレシピから抜粋したものである[8].

3.3. 距離計算による店舗選択機能

距離計算による店舗選択機能とは，ユーザの現在地から，近い店舗を選択する機能である．

この機能は，ユーザの現在地の情報を用いて，割引商品が販売されている小売店舗の中からユーザの現在地と店舗の位置情報から，調理することが可能な料理を推薦するために，近い位置にある店舗を選択する機能である．この機能によって，ユーザの緯度経度から割引商品のデータが登録されている店舗までの距離を計算し，近い位置にある店舗の選択し，遠い位置にある店舗で販売されている食品を組み合わせた料理のリコメンドを避けることが実現可能である．

3.4. 料理推薦機能

料理推薦機能とは，距離計算による店舗選択機能によって選択された，現在地から近い小売店舗で販売されている割引商品を用いて，調理することが可能な料理を複数推薦する機能である．

この機能は，データベースに登録された小売店舗の割引商品から調理できる可能性の高い料理を複数推薦することが可能である．

推薦される料理は，本システム内の料理食材データベースに格納されている料理と使用する食品のデータに基づいて推薦される．

推薦される料理の表示される順番は Jaccard 係数の値によって位置付らけれる．Jaccard 係数の計算によって，0.1 以上の値が出力された料理のみを降順で表示する．本研究ではデータベースに格納されているデータ数が少ないため，値を 0.1 よりも大きく設定した場合，最後に料理レシピを表示する際に該当する料理が存在せず，表示されない場合があるため，値を 0.1 と低く設定している．

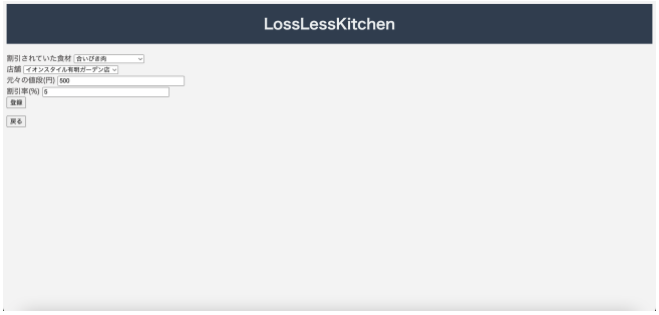


図 5 割引商品情報登録画面



図 6 料理推薦画面

4. プロトタイプ実装

3 節で述べた実現方式により，プロトタイプシステムを実装した．

本プロトタイプシステムの割引商品情報登録画面を図 5 に示す．割引商品情報登録画面は，ユーザ自身で割引商品の名前・販売店舗・値段・割引率の 4 項目を入力することにより，割引商品情報が料理食材データベースに格納される．

本プロトタイプシステムの料理推薦画面を図 6 に示す．料理推薦画面は，GPS 機能によって Web ブラウザ上から取得したユーザの現在地・割引商品情報登録画面で入力したタイムセール食材・タイムセール食材を用いて調理することが可能な料理レシピを表示する．

5. システム評価実験

本節では，4 節で実装したプロトタイプシステムを用いてその実現可能性と有効性を検証するための実験を行った．

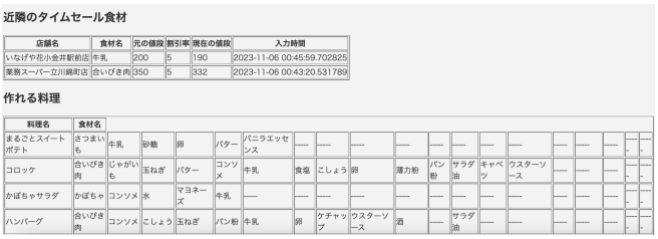
実験では，本システムを用いた被験者実験を行うことで，有効性を検証する．

5.1. 実験環境

本実験は事前に 3 つのシチュエーションを想定し，そのシチュエーションのもと行った．

本システムはタイムセール食材の情報が 2 時間を超えると自動的に非表示になる仕組みとなっているため，同時間に他のユーザが情報を登録したとしても，その店舗がユーザの現在地から直径 15km 以内でないと表示されないようになっている．被験者 1 人ずつ，本システムを用いて実際の店舗のタイムセール食材の割引商品情報を入力する作業を委託した．

シチュエーションは，1) 東京都北多摩地域西部に在住するユーザ，2) 東京都北多摩地域北部に在住するユーザ，3) 東京都区部東部に在住するユーザと設定した．それぞれ異なる地域に住むユーザのニーズに応えた割引商品・料理レシピが提示されるかを検証するためである．


近隣のタイムセール食材

店舗名	食材名	元の値段	割引率	現在の値段	入力時間
いなげや花小金井駅前店	牛乳	200	5	190	2023-11-06 00:45:59.702825
業務スーパー立川錦町店	合いびき肉	350	5	332	2023-11-06 00:43:20.531789

作れる料理

料理名	食材名																					
まるごとスイートポテト	さつまいも	牛乳	砂糖	卵	バター	バニラエッセンス	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
コロッケ	合いびき肉	じゃがいも	玉ねぎ	バター	コンソメ	牛乳	食塩	こしょう	卵	薄力粉	パン粉	サラダ油	キャベツ	ウスターソース	—	—	—	—	—	—	—	—
かぼちゃサラダ	かぼちゃ	コンソメ	水	マヨネーズ	牛乳	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
ハンバーグ	合いびき肉	コンソメ	こしょう	玉ねぎ	パン粉	牛乳	卵	ケチャップ	ウスターソース	酒	—	サラダ油	—	—	—	—	—	—	—	—	—	—

図 7 シチュエーション 1 を想定した際の料理推薦画面


近隣のタイムセール食材

店舗名	食材名	元の値段	割引率	現在の値段	入力時間
イオンスタイル有明ガーデン店	大根	140	5	133	2023-11-06 18:41:49.565560
イオンスタイル有明ガーデン店	なす	140	10	126	2023-11-06 18:42:01.791897
イオン東雲店	白菜	140	5	133	2023-11-06 18:41:27.033491
イオン東雲店	鶏肉	300	5	285	2023-11-06 18:42:16.500450

作れる料理

料理名	食材名																					
ほうとう	ほうとう	白菜	にんじん	大根	しいたけ	ねぎ	油揚げ	かぼちゃ	ほんだし	水	みそ	七味唐がらし	—	—	—	—	—	—	—	—	—	—
さんまの塩焼き	さんま	食塩	大根	すだち	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
寄せ鍋	たら	鶏肉	えび	木綿豆腐	白菜	春菊	にんじん	ねぎ	えのきだけ	しいたけ	水	みりん	酒	しょうゆ	ほんだし	—	—	—	—	—	—	—

戻る

図 9 シチュエーション 3 を想定した際の料理推薦画面

5.2. 実験結果

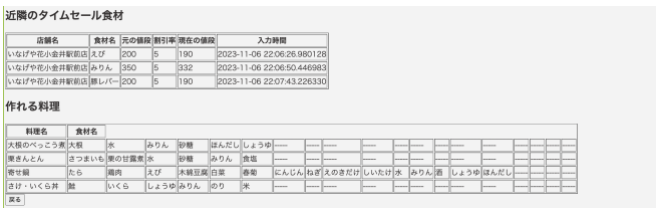
本システムのプロトタイプを使用して実施した後に回収したアンケートの結果を図に示す.

シチュエーション 1) 東京都北多摩地域西部に在住するユーザの実験結果を図 7 に示す.

図 7 は東京都立川市と小平市にあるスーパーマーケットの割引商品情報を登録したのちに，最終的にその食品を使用して調理することが可能な料理が Jaccard 係数の降順に表示された画面である．立川市のスーパーマケットの割引商品情報は合いびき肉・業務スーパー立川錦町店・350・5 を登録した．小平市の割引商品情報は，牛乳・いなげや花小金井駅前店・200・5 である．結果は，いなげや花小金井駅前店で販売されている牛乳を使用した料理が最上部に表示されたが，実用性の点から考えると，ユーザの現在地である東京都北多摩地域西部に近い業務スーパー立川錦町店で販売されている合いびき肉を使用した料理が一番上に表示されることが望ましい.

次に，シチュエーション 2) 東京都北多摩地域北部に在住するユーザの実験結果を図 8 に示す.

図 8 は東京都小平市にあるスーパーマーケットの割引商品情報を登録したのちに，その情報から調理できる可能性が高い料理が推薦されている．入力情報は全てで 4 つあり，いなげや花小金井駅前店の食品である．食品の詳細はえび・200・5，みりん・350・5，豚レバー・200・5 となっている．残り 1 つがイオン東雲店の玉ねぎ・400・3 である．結果は料理推薦された 4 つの料理のうち，3 つがみりんを使用した料理となった．また，ユーザの居住地域である東京都北多摩地域西部から直径 15km 圏外に位置するイオン東雲店の割引情報と，それを使用した料理は表示されないという正しい結果となった.


近隣のタイムセール食材

店舗名	食材名	元の値段	割引率	現在の値段	入力時間
いなげや花小金井駅前店	えび	200	5	190	2023-11-06 22:06:26.980128
いなげや花小金井駅前店	みりん	350	5	332	2023-11-06 22:06:50.446983
いなげや花小金井駅前店	豚レバー	200	5	190	2023-11-06 22:07:43.226330

作れる料理

料理名	食材名																					
大根のべっこう煮	大根	水	みりん	砂糖	ほんだし	しょうゆ	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
栗きんとん	さつまいも	栗の甘露煮	水	砂糖	みりん	食塩	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
寄せ鍋	たら	鶏肉	えび	木綿豆腐	白菜	春菊	にんじん	ねぎ	えのきだけ	しいたけ	水	みりん	酒	しょうゆ	ほんだし	—	—	—	—	—	—	—
さけ・いくら丼	鮭	いくら	しょうゆ	みりん	のり	米	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

戻る

図 8 シチュエーション 2 を想定した際の料理推薦画面

次に，シチュエーション 3) 東京都区部東部に在住するユーザの実験結果を図 9 に示す.

図 9 は，東京都江東区にある 2 つのスーパーマーケットの割引商品情報が登録されたのちに，その情報から調理できる可能性の高い料理が推薦されている．入力情報はイオンスタイル有明ガーデン店とイオン東雲店の食品であり，大根・140・5，なす・140・10，白菜・140・5，鶏肉・300・5 となっている．結果は，白菜と大根を使用したほうとうの次に，大根を使用したさんまの塩焼き，最後に鶏肉と白菜を使用した寄せ鍋の順に表示された.

この全てのシチュエーションより，距離計算による店舗選択機能に関して，全てのシチュエーションで直径 15km 以内にある店舗のみが表示される結果となった.

また，料理推薦機能に関して全てのシチュエーションで，ユーザによって登録された割引商品を使用した料理レシピが表示される結果となった.

6. 考察

本システムの実験より，距離計算による店舗選択機能によって，登録された割引情報がユーザの現在地から直径 15km 以内に位置する店舗のみを表示し，それらを使用した料理レシピが正しく提示されることが証明された.

また，リコメンドされる料理レシピは全てその時の割引商品が一部として使用されているものが表示された．しかし，リコメンドされる料理レシピは登録された割引商品の中の一部のみを使用したものが多く，登録された割引商品が多い場合，全ての割引商品を使用した料理が表示されない結果となり，調理することが可能な料理の食材に偏りが生じる結果となった.

また，最上部に表示される料理レシピに使用される割引商品はユーザの現在地との距離を考慮せずに，料理に何の食材が使用されているかに重きを置く

Jaccard 係数の値を基準として考慮した結果となった.

7. 結論と展望

本稿では，店舗で販売されている食品の廃棄量を減少させること，そして食品ロス削減を目的とした小売店舗の割引商品の提案及び料理リコメンドシステムの実現方式について示した.

本方式は，ユーザ自身がリアルタイムで販売されているタイムセール食材の割引商品情報をシステム上に登録することによって，それらを使用した料理のリコメンドや，他のユーザがその情報の閲覧を可能とするシステムを実現した．また，提案方式を実装し，有効性を検証するための実験を行った.

今後の課題としては，実用性をさらに高めていくために，登録された割引商品のみを使用して調理できる料理レシピだけでなく，ユーザの冷蔵庫に保存されている食材と割引商品を組み合わせて調理できる料理レシピをリコメンドする機能の追加が必要であると考えられる．また，ユーザの行動パターンの分析を行い，個々人に最適化されたリコメンドが可能な機能の開発などが考えられる．個々人に最適なシステムに改善するために，リコメンドされる料理の難易度や，使用される食材の個数などについても考慮する必要がある.

さらに，現在のシステムではユーザの現在地に基づき直径 15km 以内に位置する店舗の選択を行っているが，複数店舗に移動することを考慮した場合，店舗と店舗の距離を考慮したリコメンドが可能な店舗選択機能の計算の改善が必要である．そして本システムの利用ユーザとしてスーパーマーケットなど小売店舗の利用者を考えているが，食品ロス削減を目的としたシステムを実現させるため，より多くの人々の目に留まるようにするためには購入者視点だけでなく，従業員視点でも使用可能なシステムに改善する必要がある．ユーザインターフェースの使いやすさを向上させることによって，より多くのユーザに利用してもらえるような環境を整備することが重要である.

本システムによって，消費者であるユーザにとっても，販売者である店舗の従業員にとってもメリットのある食品ロス削減に向けたシステムを実現することを目指す.

参　考　文　献

[1] 消費者庁，”食品ロスについて知る・学ぶ”
(https://www.caa.go.jp/policies/policy/consumer_policy/information/food_loss/education/)

[2] NPO 法人日本もったいない食品センター，食品ロス削減ショップ ecoeat（エコイート），
https://www.mottainai-shokuhin-center.org/store/

[3] 江東区，フードドライブ，
https://www.city.edogawa.tokyo.jp/e025/kurashi/gomi_recycle/gomigenryo/tabekirisuisin/fu-dodoraibu/index.html

[4] 桐生侑恵，松井康弘，太田光一，岡本朋樹，スーパーマーケットにおける食品ロス削減に向けた需要予測に関する一考察，第 30 回廃棄物資源循環学会研究発表会，2019

[5] 川田明香利，佐々木史織，”食品ロス防止のための購買履歴に基づく位置情報つき値引き食品レコメンドシステムの実現”，DEIM forum 2023，2023

[6] 井上晴稀，梶克彦，”乗車中の列車の各駅到着時間を自動表示するスマートフォンアプリケーション”，「マルチメディア，分散，強調とモバイル(DICOMO2019)シンポジウム」，2019

[7] 佐々木健，橋本浩二，柴田義考，”災害弱者を考慮した避難共助アプリケーション”，情報処理学会第 78 回全国大会，2016

[8] AJINOMOTO PARK，https://park.ajinomoto.co.jp/

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

ライフスタイル

座長:王 元元(山口大学)
コメンテータ:吉川 正俊(京都大学)
2024年3月1日(金) 13:30 〜 15:40 T5-A (オンライン（Zoom Events))

15:15 〜 15:40

[T5-A-8-05]【技術報告】都市・ビル向けサービスの実装を容易化する空間デジタルツイン基盤

水品圭汰 (株式会社日立製作所)

日立製作所ではサイバー空間と現実空間をシームレスにつなぐ空間情報プラットフォームを開発している．中でもインフラ管理やモビリティの分野で都市スケールの3Dモデルの活用が注目されている。しかし、3D都市モデルや建物情報モデルのBIM、点群データなど屋内外がシームレスにつながるデジタルツインの構築には、複数のジオメトリデータの統合が必要であるため，構築にかかる時間や精度の担保が課題となっている。そこでジオメトリデータから特徴量を抽出することによって種類の異なるジオメトリデータの位置合わせを自動化することを可能とする技術を開発した。本手法により、デジタルツイン空間構築に係る工数を81%削減する効果が得られた。

©一般社団法人日本データベース学会

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

2024年3月1日(金) 13:30 ~ 15:40 | T5-B オンライン（Zoom Events)

環境科学

座長:加藤 誠(筑波大学)
コメンテータ:伊藤 貴之(お茶の水女子大学)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T5-B-8-01] 灯油消費に関する多変量時系列予測方式

*横須賀 祐太[1]、神長 宏侑[1]、齋藤 瑞紀[1]、林 康弘[2]、清木 康[2] (1. 武蔵野大学アジアAI研究所　客員研究員、2. 武蔵野大学　データサイエンス学部)

13:50 ~ 14:15
[T5-B-8-02] メタ学習を用いた物理条件の異なるハミルトニアンの学習

*坂本 州[1]、岩田 具治[2]、田中 佑典[2] (1. 奈良先端科学技術大学院大学、2. NTTコミュニケーション科学基礎研究所)

14:15 ~ 14:40
[T5-B-8-03] ブラックボックス関数が与えられた予算制約付組合せ最適設備投資問題に対するモンテカルロ木探索法の提案

*行木 渉真[1]、中村 篤祥[2]、安ヶ平 裕介[3]、内垣内 洋[3]、平松 道晶[3]、竹本 享史[3] (1. 北海道大学 情報エレクトロニクス学科 情報理工学コース アルゴリズム研究室、2. 北海道大学 情報科学研究院 情報理工学部門 知識ソフトウェア科学分野、3. （株）日立製作所)

14:40 ~ 15:05
[T5-B-8-04] 微生物学的水質予測のための符号制約付き学習アルゴリズム

*高田 裕也[1]、小野 晴生[1]、門屋 俊祐[2]、佐野 大輔[3]、加藤 毅[4] (1. 群馬大学大学院理工学府 、2. 東京大学大学院工学系研究科、3. 東北大学大学院工学研究科、4. 群馬大学情報学部)

15:05 ~ 15:30
[T5-B-8-05] SABM：大規模言語モデルに基づくエージェントベース実世界シミュレーション

*呉 増青[1,2]、彭 潤[3]、韓 勗[4]、鄭 舒元[1]、肖 川[1,5] (1. 大阪大学、2. 京都大学、3. ミシガン大学、4. フォーダム大学、5. 名古屋大学)

灯油消費に関する多変量時系列予測方式

横須賀祐太† 神長宏侑† 齋藤瑞紀† 林康弘‡ 清木康‡
†武蔵野大学アジア AI 研究所 〒135-8181 東京都江東区有明３丁目 ３番３号
‡武蔵野大学データサイエンス学部 〒135-8181 東京都江東区有明３丁目 ３番３号
E-mail: †yuta.yokosuka@ds.musashino-u.ac.jp, ‡{yhayashi, y-kiyoki}@musashino-u.ac.jp

あらまし 本稿は, 灯油の出荷予測における多変量時系列予測モデルの提案である. 本予測モデルは, 前処理として, 灯油の出荷における大量購入や外れ値への対処を行い, 灯油販売の実績データ(販売量)と天気データ(気温)との時系列相関量計算を行う. 本モデルは予測したい日の曜日と予報された天気を条件として実績データにおける近似値を予測値として算出する. 今回, 私たちは提案するモデルの有効性を実データとの検証を通じて評価する.

キーワード 時系列データ処理, 相関量計算, 灯油消費量, 資源・エネルギー

1. はじめに

灯油の出荷・灯油販売において, 販売店への卸売りである BtoB の関係と消費者へ直接販売する BtoC の関係が存在する. 販売店に対する灯油販売では, 1 日に運べる灯油の量が決まっている為, 各店舗への配送は, 店舗の在庫が無くなる前に, 適切な量を配送する必要がある. 参考文献[1]の調査結果から, 灯油消費の用途として, 暖房用に使用する割合が 89.7%以上となり, 冬における販売が売上の大半を占める. また, 店舗型経営において, 売行きは休日・祝日が高くなる傾向がある. 本研究の目的は, 灯油を販売する立場から取得可能な情報である, 天候や過去の出荷量を元に, 灯油消費における予測システムを確立し, 効率の良い配送を行えることである.

2. 本研究の方針

我々は, 2022 年 12 月 1 日から 2023 年 11 月 30 日までに取得した店舗の出荷量を対象に, 出荷量に関係する要素を絞り込み, 要素に応じた場合分けをした後にユークリッド距離による計算を行う. 売上予測の方式として重回帰分析や SARIMA モデル等が存在するが, 本研究は曜日や月毎により変動する特徴を考慮したフィルタリング, 気象庁が提供する過去の天気データ[2]との距離計算により, 近似する売上を取得し, 予測することが主体である. 従来方式との優位性として, 重回帰分析では時間に依存する特有のパターンやトレンドがある為, 各説明変数が変動する可能性があり, SARIMA モデルでは単一変数を仮定している為, 多変量に直接の対応はしていない.

本研究では, 季節に依存する特有のパターンやトレンドを踏まえ, 気象データとの相関を計量するモデルをシステムに導入する. 重回帰分析, SARIMA モデルで表現した予測結果と 2023 年 12 月の実際の出荷量を比較し, 提案手法の長所と短所について検討を行う.

3. 関連研究

売上予測はさまざまな産業で実施されており, その手法も幅広い. Fukushima ら[3]は過去の販売履歴データを基にしたニューラルネットワークモデルの学習を行うことで予測モデルの構築を試み, Chang ら[4]は, 非定常性の時系列データとしてアパレル業の売上データを使用し, SARIMA モデルの構築とその有用性を検証している. Nakamura ら[5]は, 相関分析と分散分析を適用することで日々の売上高と気象要因との関係について解析し, 回帰分析と比較したファジィ推論モデルを構築した.

灯油の消費量の要因分析については, Adachi ら[6]は, 新潟市を対象に, 夏季, 冬季の住まい方の実態及びエネルギー消費量原単位を把握し, エネルギー消費量の要因分析を行い, 住宅用エネルギー消費構造の解明を行っている. Adachi らの調査結果では灯油消費量が 4 月から 10 月においては激減し, エネルギー消費量に大きく影響する要因が「気候」という地域差によるところが大きいという結論を Adachi らは支持している.

4. 提案手法

4.1. データ構造

本研究では, 消費者に対する直接の販売ではなく, 販売店から販売店への灯油の販売を想定しているため, 一般的に売上には, 広告活動の有無や店頭販売価格との相関が認められるが, 毎日の営業終了時の灯油用オイルタンクの残量, 入荷量, 出荷量のみを取得可能な情報として定義する. また, 灯油の使用用途は, 89.7%[1]が暖房用であるとされ, 灯油販売における出荷量と該当店舗の平均気温を比較すると, 負の相関関係が確認できる(図 1)が, 単月の出荷量と平均気温からは, 曜日によって売上に差が生じ(図 2), 降水量との比較においては負の相関が示されている (図 3).

雨天の際に出荷量が少ない傾向として, 一般的に灯

油缶は屋外に保管されており，店頭まで雨の中持ち運ぶ必要がある為，買い控えが起きていると推測される．また，業界内では，気温が 7 度以下の場合，灯油が売れる傾向があると言われている．図 2 からも平均気温が 7 度以下を下回った場合に売上が上がる傾向があることから，気象庁より，時間単位の気温を取得し，1 日の中で 7 度を下回った回数を計測した．なお，Mae らの研究[7]より，就寝中とみられる深夜においては暖房の使用率が著しく下がることから，7 度を下回った回数をカウントする時間帯は 6 時から 24 時までに限定する．本研究では，1 日辺りの出荷量，平均気温，降水量，時間帯辺りの 7 度を下回った回数を対象に，計算を実施する．

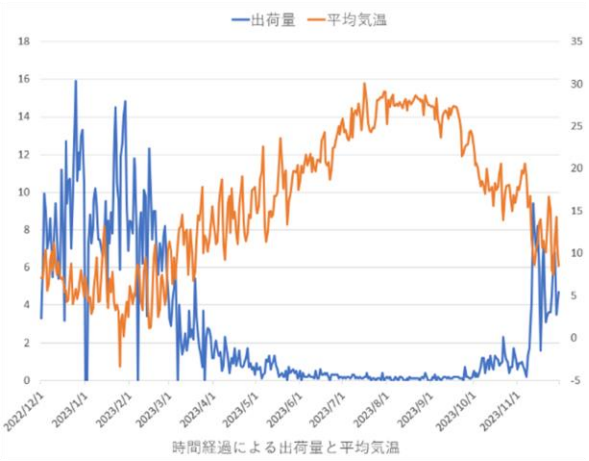


図 1 ：時間経過による出荷量と平均気温

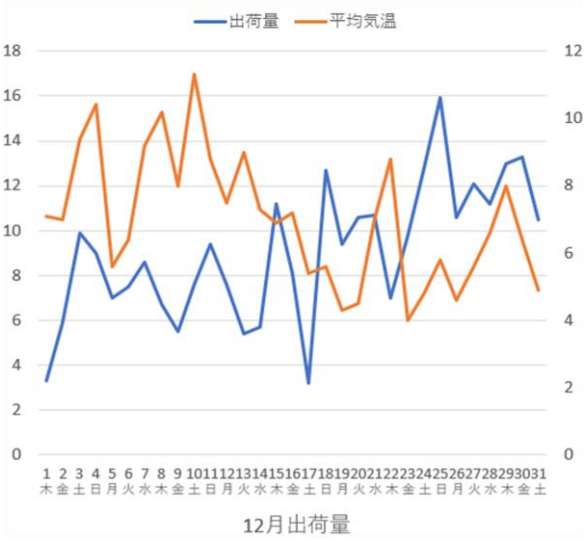


図 2 ：12 月　出荷量と平均気温

4.2. 重回帰分析モデル

重回帰分析は，複数の説明変数が目的変数にどのように影響しているか理解する為の統計的手法である．



図 3 ： 12 月　出荷量と降水量

目的変数は灯油の出荷量である．説明変数は，平均気温，時間帯辺りの 7 度以下回数，降水量，土日・祝日である．曜日については，過学習や共線形性の発生を防ぐ為，土曜日または日曜日，国民の祝日の場合のみ 1 とし，他は 0 とする.トレンド，季節性，周期性が影響していることを考慮し,12 月のみを対象とした場合と，出荷量が多い 11 月から 3 月に限定した場合で比較を行う．

4.3. SARIMA モデル

SARIMA モデルは，時系列データの予測に使用されるモデリング手法であり，自己回帰成分，積分成分，移動平均成分を持つ ARIMA モデルに季節要素を追加したモデルである．季節成分も考慮できるため，トレンドが非定常性を示すデータにも適している．本研究では，対象データをトレーニングする．ディッキー-フラー検定を実施して，帰無仮説が棄却できるかどうかを判定し，定常性の有無を確認する．モデルの各パラメータ（自己回帰，階差，移動平均，季節性の自己回帰，季節性の階差，季節性の移動平均）は探索的に組み合わせを試し，RMSE を指標として最適なパラメータを決定する．周期は 12 とする．

4.4. 本方式の数学モデル

本方式の数学モデルとして 3 次元のユークリッド距離計算を採用する．本方式の特徴は，現在の状況と最も類似する過去の事象を，距離計算により発見し，その過去事象において発生した次時点の状況を予測として抽出する点にある．平均気温，時間帯辺りの 7 度以下回数，降水量を計算次元として設定し，それぞれの単位を指数に変換するため，蓄積された過去分と予測日，それぞれのデータに対して標準化を実施し，予測

したい日付の曜日と合致する曜日だけをフィルタリングした上で，予測日と過去の平均気温，時間帯辺りの 7 度以下回数，降水量を対象に距離計算を行う．距離計算の結果，予測をしたい日と最も近い過去の日付を求め，その出荷量を予測値として算出する．本方式における標準化と距離計算の式を以下に示す．

標準化 ： $Z = \frac{X-\mu}{\sigma}$

（X：元のデータの値，μ：データの平均値，σ：データの標準偏差）

3 次元の距離 ： $d_i = \sqrt{(X - X_i)^2 + (Y - Y_i)^2 + (Z - Z_i)^2}$
$\{i = 1,2,\dots,N\}$

（X：予測日の平均気温，Y：予測日の 1 時間単位での 7 度以下回数，Z：予測日の降水量，X_i：過去の平均気温，Y_i：過去の 1 時間単位での 7 度以下回数，Z_i：過去の降水量，N：データ数）

距離計算の結果は，距離が最も近かった 1 件の出荷量と，上位 3 件の出荷量の平均値を求め，比較を行う．本研究では，灯油が多く出荷される時期(11 月から 3 月)とそれ以外(4 月から 10 月)に分離し，参照する過去データにおいても 11 月から 3 月に限定した．店舗によっては，稀に業者による大量購入や，残量の計測を怠ったことによる出荷量の不明，休業日により出荷量が 0 の場合も存在する為，これらに該当する場合は，計算結果への影響を考慮し，参照するデータからは除外した．

5. 実験結果

各モデルにおける予測結果は表 1 の通りである．また，予測期間内の実測値との差の合計と，実測値を基準とした差の絶対合計は表 2 に示される．

予測に関して，11 月から 3 月までの期間を対象とした，ユークリッド距離計算が高い精度で予測できていることが分かった．また，距離が最も近かった 1 件よりも，上位 3 件の平均値の方がより高い精度での判定ができている．他方，重回帰分析モデルでは，12 月だけの過去データから求めた回帰式では，比較的高い精度で予測できているが，11 月から 3 月の期間を含めた回帰式では，実測値から大幅に上回る予測値が生じた．SARIMA モデルでは，絶対合計においては高い精度で予測できているが，合計においてはマイナスの結果となった．また，ディッキー-フラー検定の結果(図 4)は，p 値が 0.7123 となり，有意水準である 0.05 を大きく上回り，帰無仮説を棄却できないことが分かり，出荷量の非定常性を示した．

表 1：実測値と各モデル予想結果

日付	曜日	実測値	重回帰分析 2022年12月	重回帰分析 11月～3月	SARIMA 2022年12月～2023年11月	ﾕｰｸﾘｯﾄﾞ 11月～3月	ﾕｰｸﾘｯﾄﾞ (上位3平均) 11月～3月
2023/12/1	金	5.9	10	20.5	4.212884	4.2	7.8
2023/12/2	土	7.8	11.5	22.1	4.70753	7.6	7.4
2023/12/3	日	9.2	10.7	21.2	4.988309	6.6	7.4
2023/12/4	月	6.7	7.1	19	5.297221	7.9	7.8
2023/12/5	火	6.3	6.5	18.4	5.532859	3.1	4.5
2023/12/6	水	3.4	5.8	17.3	6.520527	8.2	6.5
2023/12/7	木	4.4	3.3	16.9	6.060472	2.3	2.3
2023/12/8	金	4.4	5.3	17.9	5.634725	6.6	4.9
2023/12/9	土	6.2	4.7	18.5	6.0763	7.6	4.5
2023/12/10	日	6.2	2.1	17.6	5.904972	2.5	6.3
2023/12/11	月	4.6	3.3	16.9	4.950101	2.3	4.2
2023/12/12	火	2.1	4	16.3	5.635608	6.3	5.3
2023/12/13	水	3.6	6.9	19	5.823884	8.6	8.3
2023/12/14	木	4.3	8	18.9	5.53247	6.7	4.8
2023/12/15	金	6.9	5.3	17.6	5.221046	6.6	4.6
2023/12/16	土	6.2	0.6	16.9	5.554693	7.6	6
2023/12/17	日	7	7.8	20.7	5.753498	10.2	8.5
2023/12/18	月	7.4	10.4	21.4	6.472958	9.4	9.8
2023/12/19	火	8.1	8.9	20.7	6.382235	7.3	8.4
2023/12/20	水	7.1	7.8	19.3	6.64237	3.6	7.6
2023/12/21	木	9.8	8.8	19.9	7.018344	7	8.3
2023/12/22	金	8.1	9.7	21.5	5.997406	9.8	8.6

表 2：合計と絶対合計

	重回帰分析 2022年12月	重回帰分析 11月～3月	SARIMA	ﾕｰｸﾘｯﾄﾞ 11月～3月	ﾕｰｸﾘｯﾄﾞ (上位3位の平均) 11月～3月
合計	12.8	282.8	-9.779588	6.3	8.1
絶対合計	45.2	282.8	36.495162	52.7	32.5

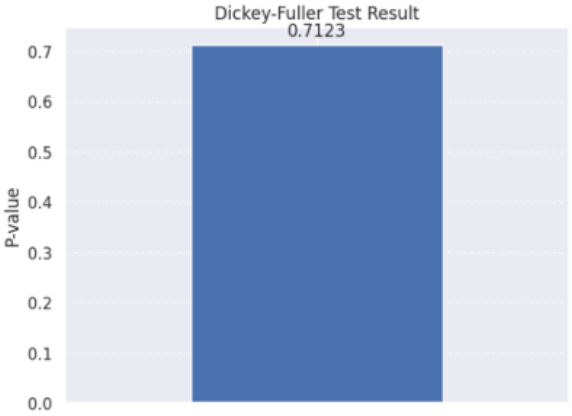


図 4：ディッキー-フラー検定結果

5.1. 実験結果に関する考察

重回帰分析モデルにおいて，参照するデータの期間を変更することによって大きく差が生じた理由としては，データが(図 4)からも，非定常性の時系列データになっていた為，トレンドや季節性を考慮できず，各説明変数が変動した(表 3，4)ことが原因と推察される．

SARIMA モデルの合計値がマイナスになった理由としては，図 1 のようにデータの周期性が安定していな

かった為，少ない出荷量で算出されることが多かったと考えられる．

表 3： 重回帰分析　比較

	12月のみ	11月から3月
補正 R2	0.324381848	0.574615172
有意 F	0.006075214	0.006075214

表 4： 重回帰分析　係数・t・p 値

		12月のみ	11月から3月
係数	切片	24.41657846	24.41657846
	休日・祝日	2.069690726	2.069690726
	平均気温	-1.690018818	-1.690018818
	7度以下回数	-0.51613877	-0.51613877
	降水量の合計(mm)	-0.041157842	-0.041157842
t	切片	5.619256495	5.619256495
	休日・祝日	2.008049418	2.008049418
	平均気温	-3.690546377	-3.690546377
	7度以下回数	-2.964101904	-2.964101904
	降水量の合計(mm)	-0.324571363	-0.324571363
P-値	切片	6.61044E-06	6.61044E-06
	休日・祝日	0.055133259	0.055133259
	平均気温	0.001042187	0.001042187
	7度以下回数	0.006420249	0.006420249
	降水量の合計(mm)	0.748102521	0.748102521

ユークリッド距離モデルの精度が高かった理由としては，曜日や平均気温，降水量，暖房が良く使われると思われる外気温の回数の特徴を捉えることが出来た為と思われる．

6. まとめ

本稿では，灯油を販売する立場から取得可能な情報として，天候や過去の出荷量を対象とし，灯油消費における予測システムを提案し，高い精度の予測結果を示した．さらに本研究を発展させるためには，出荷量予測と連携した，配送における効率的なシフト形成が必要になる．また，実際の天気と天気予報においては誤差が生じる為，誤差を考慮したパラメータの調整を行う必要がある[8][9]．今後の課題としては，より多くのデータセットを用いてモデルを訓練することによりさらなる精度向上を図ることが挙げられる．

参　考　文　献

[1] 経済産業省資源エネルギー庁 (財)日本エネルギー経済研究所石油情報センター．平成 18 年度灯油消費実態調査, 2005.

[2] 国土交通省気象庁, https://www.data.jma.go.jp

[3] 福島航平 (兵庫県立大学), 大島裕明 (兵庫県立大学), チラシ媒体通販における商品の販売数予測, DEIM2022, 2022.

[4] 張瀚天, 村井諒, 佐々木健佑, 片山茜, 善甫啓一, 中林紀彦．アパレル業の売上予測における時系列モデルの適応条件の検討．人工知能学会全国大会, pp. 1-4, 2016.

[5] 中村俊輔, 古殿幸雄, 気象要因との関連におけるファジィ推論需要予測モデルの一考察, バイオメディカル・ファジィ・システム学会誌, 10 巻, 1 号, p. 26-32, 2008.

[6] 足立直之, 赤林伸一, 吉野博, 真保聡裕, 坊垣和明, 澤地孝男, 住宅におけるエネルギー消費量と住まい方に関する実態調査 : 新潟市におけるエネルギー需要構造に関する研究 その1, 日本建築学会計画系論文集, 59 巻, 465 号, p. 49-59, 1994.

[7] 前真之, 鍋島美奈子, 二宮秀與, 永村一雄, 長井達夫, 鎌田元康．集合住宅における暖房の使用状況に関する研究, 空気調和 ・衛生 工学会論文集, No ． 89, Apr. 2003.

[8] Yasushi Kiyoki and Saeko Ishihara: “A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases,” Information Modeling and Knowledge Bases (IOS Press), Vol. 14, pp.86-103, May 2002.

[9] Kiyoki, Y., Chen, X., Veesommai, C., Wijitdechakul, J., Sasaki, S., Koopipat, C., & Chawakitchareon, P.: "A semantic-associative computing system with multi-dimensional world map for ocean-environment analysis", Information Modelling and Knowledge Bases XXX, pp. 147-168.

メタ学習を用いた物理条件の異なるハミルトニアンの学習

坂本　州† 岩田　具治†† 田中　佑典††

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916 番地 5 号
†† NTT コミュニケーション科学基礎研究所　〒 619–0237 京都府相楽郡精華町光台 2 丁目 4 号
E-mail: †sakamoto.shu.sn2@is.naist.jp, ††{tomoharu.iwata,ysk.tanaka}@ntt.com

あらまし　近年，物理現象を観測したデータから機械学習を用いて物理ダイナミクスを推定する手法が注目されている．その 1 つである Hamiltonian Neural Network (HNN) は，ハミルトン力学に基づいて設計され，エネルギー保存則を満たす出力が得られるように力学系を学習できる．しかし，HNN は 1 つの系のみを学習することを前提としており，物理パラメータの異なる系を推定するためには再学習を行う必要がある．本研究では，物理パラメータを特徴量に加えることでその違いを考慮できるように HNN を拡張し，異なる系のハミルトニアンを予測できる手法を提案する．さらに，このモデルに既存のメタ学習手法を適用し，複数の系のデータを学習することで，学習に使用していない系に対して少量のデータでも高い性能で適合する手法を提案する．物理パラメータを特徴量に加えることの有効性を示す実験と，少量の学習に使用しない系のデータに対する適合性能を示す実験を行い，既存の手法に比べ，それぞれの提案手法が高い性能を示した．

キーワード　物理モデル，ハミルトン力学，物理シミュレーション，メタ学習，Physical Models，Hamiltonian Mechanics，Physical Simulation，Meta-Learning

1 はじめに

物理現象の多くは，これまでの物理学の発展によって明らかになった微分方程式によって記述される．これらの方程式を解くことで，様々な条件での物理現象のシミュレーションが可能となる．物理シミュレーションは多くの自然科学分野や工学分野などにおいて活用されてきた．しかし，複雑な物理現象のモデル化には限界があることや，モデル化に高度な専門知識が必要な場合があることなどの課題があった．一方で，Neural Ordinary Differential Equation (Neural ODE) [1] のような，観測データから機械学習を用いて物理ダイナミクスを推定する手法が近年研究されている．機械学習によって，複雑な物理モデルを高度な専門知識を必要とせずに物理シミュレーションを可能とすることが期待されている．しかし，データのみから学習を行うブラックボックスな手法では，物理法則を満たさないモデルが学習されてしまう可能性がある．そこで，物理法則を満たす出力が得られるようにモデルに制約を与える手法が研究されている．その 1 つが Hamiltonian Neural Network (HNN) [2] である．HNN は，ハミルトン力学に基づいて設計されたニューラルネットワークであり，エネルギー保存則を満たす出力が得られるように力学系を学習できる．しかし，HNN は 1 つの系のみを学習することを前提としているため，物理パラメータの異なる系を推定するためには再学習を行う必要がある．

本論文ではこの問題を解決するため，系の物理パラメータを入力できるよう HNN を拡張し，系ごとの違いを学習できるようにするアプローチを提案する．例として振り子のデータを学習する場合について考える．HNN はひもの角度と角運動量のみを学習に用いるが，本論文の提案手法では，ひもの長さや物体の質量という物理パラメータを特徴量として追加することで系ごとの違いを学習し，異なる系の間で共通する性質を学習する．これにより，学習していない系に対しても，別の系のデータを学習することで高い性能でモデル化できるようになることが期待できる．HNN への入力にこれらの物理パラメータを特徴量として加えたデータを用いつつ，HNN と同様に，エネルギー保存則を満たすように力学モデルを学習するのが本論文の提案手法である．このアプローチは，観測が難しい条件下の系やデータを手に入れるのに費用や手間がかかる系について，データが用意できなかった際に，関連する他の系のデータを学習に使用することで高い性能でシミュレーションを行うことを可能にする．

また，少量のデータだけが得られる場合，メタ学習を用いたアプローチが考えられる．メタ学習とは，関連する別のタスクのデータから学習の仕方を学習し，目的のタスクにおいて少量のデータからでも性能を高めることを目的とした手法である．複数の異なる系のデータによりメタ学習を行い，系を推定する方法を学習することで，学習していない未知の系を推定する性能が高まることが期待できる．

1 つ目の実験では，複数の系についてのデータを学習に用いて，学習に使用していない系についてのデータを用いて性能を評価することで，物理パラメータを特徴量に追加することの有効性を確認する．2 つ目の実験では，提案手法をメタ学習に拡張した際の性能の評価をする．学習に使用していないデータを用いて系に適合した際の性能を比較し，それぞれのメタ学習の手法に性能の差がどのように現れるかを確認する．

2 関連研究

力学系の時間発展は一般的に微分方程式により表現されるが，Neural Ordinary Differential Equations (Neural ODE) [1] やその発展 [3, 4] では Neural Network (NN) によって観測データから系のダイナミクスをモデル化する．Hamiltonian Neural Network (HNN) [2] やその変形 [5,6] は NN によって系のエネルギーを表す関数をモデル化し，ハミルトンの運動方程式に従ってダイナミクスを導出することにより，エネルギー保存則に従う予測値を得ることができる．近年では，HNN を拡張しエネルギー散逸を考慮する研究 [7,8] や，制御可能な入力を考慮する研究 [9] が進められている．また，HNN は状態変数の時間微分を教師データとして利用するが，ODE ソルバを用いて状態変数の軌道を学習に利用する研究 [10,11] が行われている．本論文に近い位置づけの研究としては，系の物理パラメータを入力するよう Neural ODE を拡張した研究 [12] や，メタ学習の手法を HNN に適用することで異なる系の間で共通する性質を学習し，データが少量な系でも高い性能で適合するモデルの研究 [13] が行われている．本論文では，系の物理パラメータを特徴量として加えた HNN の拡張モデルを提案する．また，そのモデルにメタ学習を適用し，データが少量な系でも高い性能で適合する手法を提案する．

本論文で使用するメタ学習手法は Model-Agnostic Meta-Learning (MAML) [14] と Conditional Neural Processes (CNP) [15] である．近年，勾配法を用いて汎化性能の良いモデルの初期パラメータを学習するメタ学習手法の研究 [14,16,17] が行われており，MAML は代表的な手法の 1 つである．他方では，ブラックボックスな NN によってタスクへの適合をモデル化するメタ学習手法の研究 [15,18–20] が進んでいる．その手法の 1 つが CNP であり，タスク別の特徴ベクトルを得ることでタスクへの適合をモデル化する．また，本論文で提案する手法のように，タスクの情報をメタ学習時に特徴量に追加する研究 [21] が行われている．

3 ハミルトニアンニューラルネットワーク

ハミルトン力学では，自由度 N の系を考えるとき，一般化座標 $\boldsymbol{q} = (q_1, q_2, ..., q_N)$ と一般化運動量 $\boldsymbol{p} = (p_1, p_2, ..., p_N)$ を用いて系が記述される．ハミルトン力学における運動方程式は，これらの状態変数 $\boldsymbol{x} = (\boldsymbol{q}, \boldsymbol{p})^\top$ とその時間微分 $\dot{\boldsymbol{x}}$ を用いて以下のように表される．

$$\dot{\boldsymbol{x}} = \Omega \nabla_{\boldsymbol{x}} H(\boldsymbol{x}) \tag{1}$$

ここで，$\Omega = \begin{bmatrix} 0 & I_N \\ -I_N & 0 \end{bmatrix}$ (I_N は N 次元の単位行列) は $2N \times 2N$ の交代行列であり，$H(\boldsymbol{x}) : \mathbb{R}^{2N} \to \mathbb{R}$ はハミルトニアンと呼ばれるスカラー関数である．図 1 に概要を示す．ハミルトニアン H は系の全エネルギー量を示す関数であり，物理現象ごとに異なる式によって表現される．また，式 (1) よりハミルトニアン H は時間 t に依存しない保存量であることを導くことができる．すなわち，式 (1) によって記述できる系はエネルギー保存則を満たす．互いに独立な変数 $q_1, q_2, ..., q_N, p_1, p_2, ..., p_N$ からなる $2N$ 次元空間を考えると，系の状態は空間の 1 点で表され，系の時間発展は空間上の軌道として捉えることができる．このような空間は相空間と呼ばれる．相空間の概要を図 2 に示す．式 (1) より得られる状態変数の時間微分 $\dot{\boldsymbol{x}}$ は相空間上のベクトル場と見なすことができ，軌道の初期値 $\boldsymbol{x}_0$ が与えられると，

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \int_{t_0}^{t_1} \dot{\boldsymbol{x}} dt \tag{2}$$

のように，時間 t について積分を行うことで系の時間発展を表す軌道を計算することができる．エネルギー保存則を満たす系では，図 2 の相空間上に示す青い軌道のように，状態変数は同じ軌道を周回し続ける．

Hamiltonian Neural Network (HNN) は，式 (1) に基づきニューラルネットワークを用いてハミルトニアンを学習する手法であり，損失関数は次のように表される．

$$\mathcal{L} = \| \dot{\boldsymbol{x}} - \Omega \nabla_{\boldsymbol{x}} H_\theta(\boldsymbol{x}) \|_2^2 \tag{3}$$

ここで，$H_\theta : \mathbb{R}^{2N} \to \mathbb{R}$ はパラメータ θ をもつニューラルネットワークである．図 3 に示すように，図 1 においてハミルトニアンを計算していた部分をニューラルネットワークに置き換えたのが HNN である．式 (1) を利用した損失関数を使用しており，エネルギー保存則を満たす力学系を学習することができる．ある物理パラメータを持つ 1 種類の系についての観測データの集合 $\{(\boldsymbol{x}, \dot{\boldsymbol{x}})\}$ を用いて学習を行う．計算過程に含まれる勾配の計算は，自動微分ライブラリを使用して簡単に行うことができる．

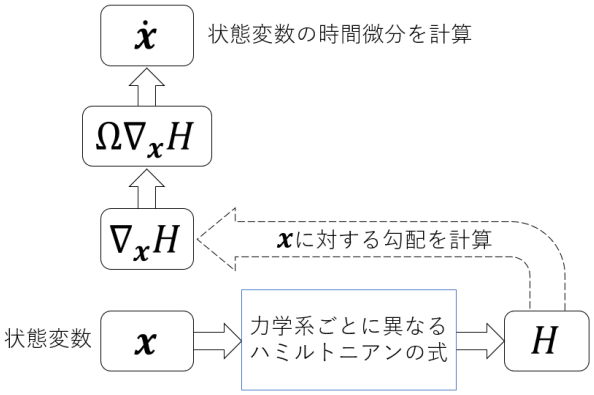


図 1 式 (1) の概要

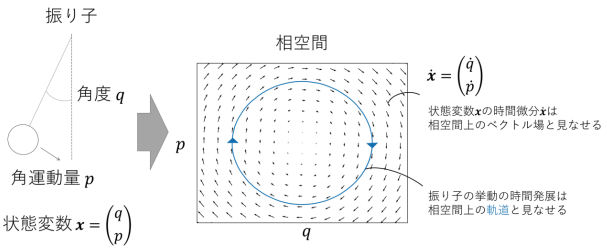


図 2 相 空 間

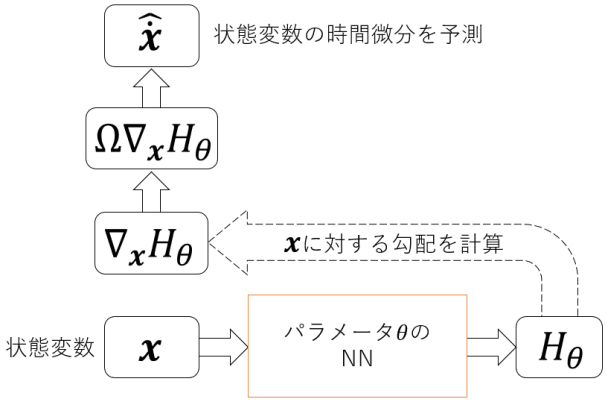


図 3　HNN

4　物理パラメータに依存するハミルトニアンニューラルネットワーク

4.1　問題設定

まず初めに，タスクを定義する．タスクはある物理パラメータを持つ力学系の観測データを用いてその系をモデル化するという 1 つの問題を表す．本章で提案する手法の目的は，複数のタスクのデータを利用して，データのない未知のタスクの系をモデル化することである．このアプローチは，観測データを用意するのが難しい系に対して，関連する他の系のデータを学習に使用することで高い性能でシミュレーションを行うことを可能にすると期待できる．

上記の問題設定を数学的に表記すると以下のようになる．D をタスクの集合，$d \in D$ をタスクとして，データ集合 $\{\mathcal{D}_d\}_{d\in D}$ と物理パラメータ $\{\boldsymbol{c}_d\}_{d\in D}$ が与えられる．$\mathcal{D}_d = \left\{\{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit})\}_{t=1}^{T}\right\}_{i=1}^{I}$ はタスク d に対応する系における T 時刻間の軌道 I 個のデータ集合であり，$\boldsymbol{x}_{dit}$ は i 番目の軌道の t 時刻の状態変数，$\dot{\boldsymbol{x}}_{dit}$ は i 番目の軌道の t 時刻の状態変数の時間微分を表す．$\boldsymbol{c}_d = (c_{d1}, c_{d2}, ..., c_{dB})$ はタスク d に対応する系の物理パラメータを表す．$\{\mathcal{D}_d\}_{d\in D}$ と $\{\boldsymbol{c}_d\}_{d\in D}$ を用いてモデルの学習を行い，未知のタスク $d^* \notin D$ に対応する系の状態変数 x_{d^*} と物理パラメータ $\boldsymbol{c}_{d^*}$ によって状態変数の時間微分 $\dot{\boldsymbol{x}}_{d^*}$ を推定することを目的とする．問題設定を図 4 に示す．

4.2　モデル

提案手法では，物理パラメータに依存するハミルトニアンをパラメータ θ をもつニューラルネットワーク $H_\theta^P : \mathbb{R}^{2N} \times \mathbb{R}^B \to \mathbb{R}$ によってモデル化する．状態変数 $\boldsymbol{x} = (\boldsymbol{q}, \boldsymbol{p})^\top$ と物理パラメータ $\boldsymbol{c}$ を入力し，状態変数の時間微分 $\dot{\boldsymbol{x}}$ を出力するモデルである．以下に式を示す．

$$\hat{\dot{\boldsymbol{x}}} = \Omega\nabla_{\boldsymbol{x}} H_\theta^P(\boldsymbol{x}, \boldsymbol{c}) \tag{10}$$

このモデルの概要図を図 5 に示す．

4.3　学習方法

提案手法において最適なパラメータ θ を推定するために解くべき最適化問題はデータ集合 $\{\mathcal{D}_d\}_{d\in D}$ と物理パラメータ $\{\boldsymbol{c}_d\}_{d\in D}$ を用いて

$$\hat{\theta} = \arg\min_\theta \sum_{d\sim D} \mathcal{L}\left(\mathcal{D}_d; \theta\right) \tag{11}$$

と計算する．目的関数 $\mathcal{L}(\mathcal{D}_d; \theta)$ は式 (10) より

$$\mathcal{L}(\mathcal{D}_d; \theta) = \frac{1}{IT} \sum_{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit})\in\mathcal{D}_d} \left\| \dot{\boldsymbol{x}}_{dit} - \Omega\nabla_{\boldsymbol{x}} H_\theta^P(\boldsymbol{x}_{dit}, \boldsymbol{c}_d) \right\|_2^2 \tag{12}$$

とする．各タスクのデータに対する目的関数を計算し，合計とそれに対する勾配を計算してパラメータ θ を更新する．

4.4　実験設定

4.4.1　データセット

実験には調和振動子，単振り子の 2 種類の物理現象についてのデータセットを使用した．データに含まれる各軌道は以下に示すハミルトニアンの式を用いて，式 (1) より得られる微分方程式と式 (2) による積分によって生成した．訓練データは物理パラメータがそれぞれ異なる 5000 個の系からなり，テストデータとバリデーションデータは 100 個の系からなる．各系のデータは 1s 間の 20 個の軌道の集合と系の物理パラメータにより構成される．各軌道のサンプリング周波数は 10Hz である．また，各サンプルは各時刻における状態変数とその時間微分により構成される．それぞれの力学系を表現するハミルトニアンの式と，系を決定する際に使用した物理パラメータの分布を以下に説明する．

a）　調和振動子

調和振動子のハミルトニアンは以下の式で表される．摩擦がなく，保存力のみが働く系を考える．

$$H = \frac{k\boldsymbol{q}^2}{2} + \frac{\boldsymbol{p}^2}{2m} \tag{9}$$

ここで，$\boldsymbol{q}$ は座標，$\boldsymbol{p}$ は運動量であり，k はばね定数，m は質量を表す．それぞれの物理パラメータは $(m, k) \in ([0.5, 2], [0.5, 2])$ の範囲における一様乱数より決定した．また，状態変数の軌道の初期値は $(q, p) \in ([-10, 10]^2)$ の範囲における一様乱数より決定した．

b）　単振り子

単振り子のハミルトニアンは以下の式で表される．摩擦がなく，保存力のみが働く系を考える．

$$H = mgl(1 - \cos\boldsymbol{q}) + \frac{\boldsymbol{p}^2}{2ml^2} \tag{10}$$

ここで，$\boldsymbol{q}$ は鉛直を基準とした振り子の角度，$\boldsymbol{p}$ は運動量であり，l は振り子の長さ，m は質量を表す．g は重力加速度を表し，本研究では $g = 1$ とする．それぞれの物理パラメータは $(m, l) \in ([0.5, 2], [0.5, 1])$ の範囲における一様乱数より決定した．また，各状態変数の初期値について，座標は $q = 0$，運動量は $p \in [-\pi, \pi]$ の範囲における一様乱数より決定した．

4.4.2　比較するモデル

比較するモデルは以下の 4 つである．

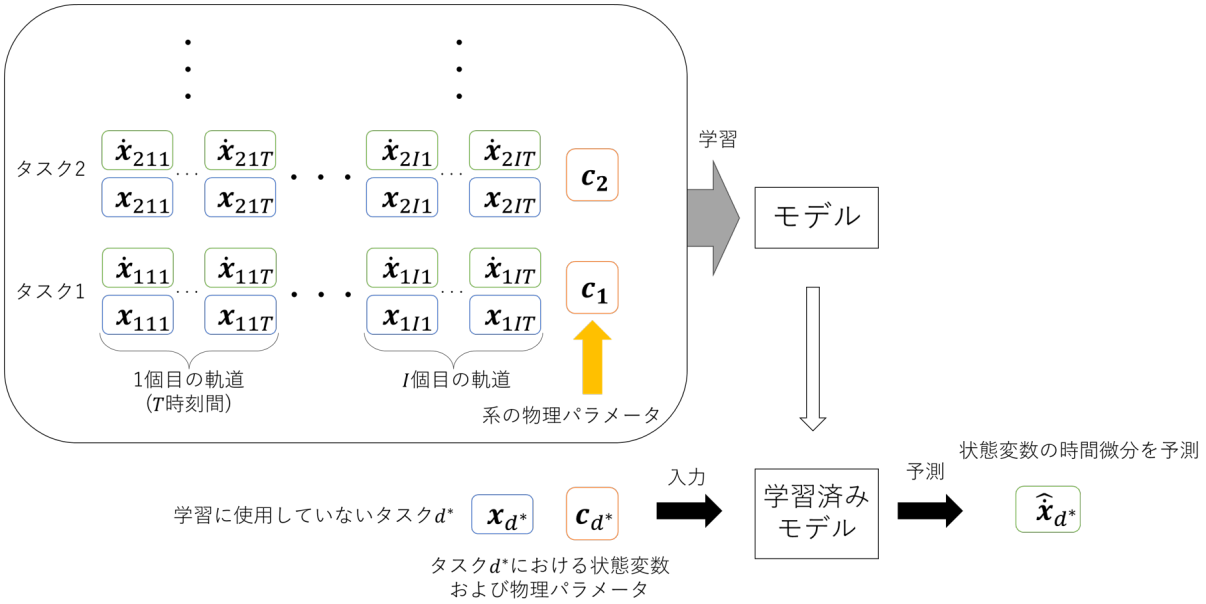


図 4　問 題 設 定

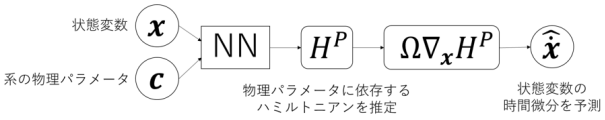


図 5　提案モデル

(a) 状態変数 $\boldsymbol{x}$ のみを特徴量とする MLP
(b) 状態変数 $\boldsymbol{x}$ と系の物理パラメータを特徴量とする MLP
(c) HNN
(d) 提案モデル

いずれのモデルも 64 ノードからなる 3 層の隠れ層からなり，各層の活性化関数には Softplus を用いた．勾配法は Adam を使用して最適化を行う．バッチサイズは 10 系分のデータ (10 系× 20 軌道× 50 時刻) として学習率 0.002 で学習を行った．バリデーションデータは学習の早期終了に利用し，学習するエポック数の最大値は 1000 とした．簡単のため，以下では (a)，(b) を MLP-X，MLP-XC と表記することにする．

4.4.3　評 価 方 法

テストデータにおける状態変数の時間微分 $\dot{\boldsymbol{x}}$ の予測値と真値の MSE を評価に使用する．また，テストデータを構成する 100 個の系において，式 (2) より各学習済みモデルを用いて 100 時刻分の軌道とエネルギーを積分計算し，真の軌道とのエネルギーに対する MSE を評価に用いる．

4.5　実 験 結 果

テスト誤差と生成した状態変数の軌道のエネルギーの MSE を表 1 に示す．また，生成した状態変数の軌道の 1 つとそれに対応するエネルギー，および各時刻のそれぞれの値の MSE を図 6，7 に示す．黒い破線は真値の軌道を示しており，時刻が進むと状態変数はこの軌道上を周回する．図 8 は訓練データとテストデータに対する学習中の各エポックの損失のグラフであ

表 1　テストデータに対する誤差と予測した軌道に対応するエネルギーの MSE (MSE と標準偏差を示す．最も精度の良い値を太字で示す．)

力学系	学習モデル	テスト誤差	エネルギーの MSE
調和振動子	MLP-X	11.652 ± 13.705	110.198 ± 124.561
	MLP-XC	0.037 ± 0.047	70.161 ± 90.865
	HNN	9.560 ± 12.588	1.233 ± 2.521
	提案モデル	**0.021 ± 0.034**	**0.021 ± 0.038**
単振り子	MLP-X	3.432 ± 5.200	18.742 ± 28.917
	MLP-XC	0.056 ± 0.109	3.757 ± 5.933
	HNN	2.931 ± 6.235	0.089 ± 0.141
	提案モデル	**0.035 ± 0.048**	**0.011 ± 0.030**

る．提案モデルはテスト誤差とエネルギーの MSE のいずれも他のモデルより高い精度を示した．MLP-XC はエネルギー保存則を事前知識として考慮していないため，提案モデルと比べ系を学習する性能が低く，図 6，7 の軌道を見ると，シミュレーションが進むごとに正しい軌道から離れており，エネルギーの誤差も徐々に増加しているのがわかる．HNN は 1 つの系のみを学習することを前提としているため，他の系との違いや共通する性質を学習することはできず，未知の系を推定する性能は低い．提案モデルは物理パラメータを特徴量として入力しているため，物理パラメータに依存して変化するハミルトニアンを学習しており，各系において高い精度でシミュレーションを行うことができる．

5　提案手法へのメタ学習の適用

5.1　問 題 設 定

本章では，複数のタスクのデータを利用して，学習に使用していない未知のタスクのデータからその系を推定することを目的とする．学習フェーズにて，前章と同様にタスク集合 D に

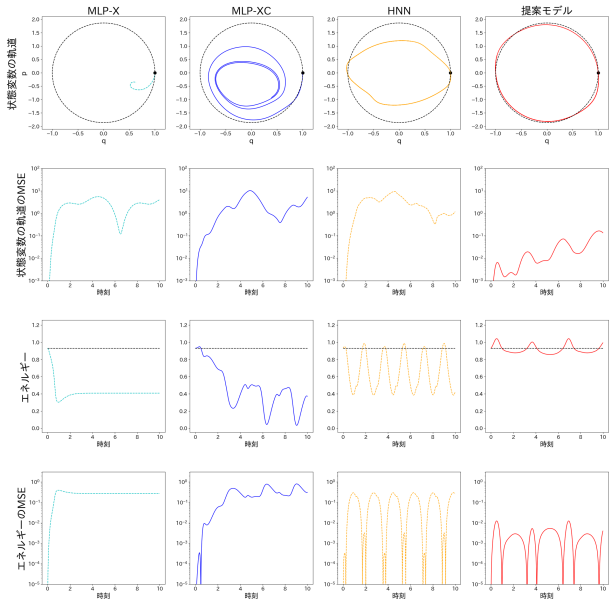


図 6　調和振動子における各モデルの初期値 $(q, p) = (1, 0)$ に対する状態変数の軌道の予測とそれに対応するエネルギー，および各時刻でのそれぞれの値の MSE を示す．黒い破線は軌道の真値，黒い点は軌道の初期値である．軌道の MSE とエネルギーの MSE は対数スケールで示している．

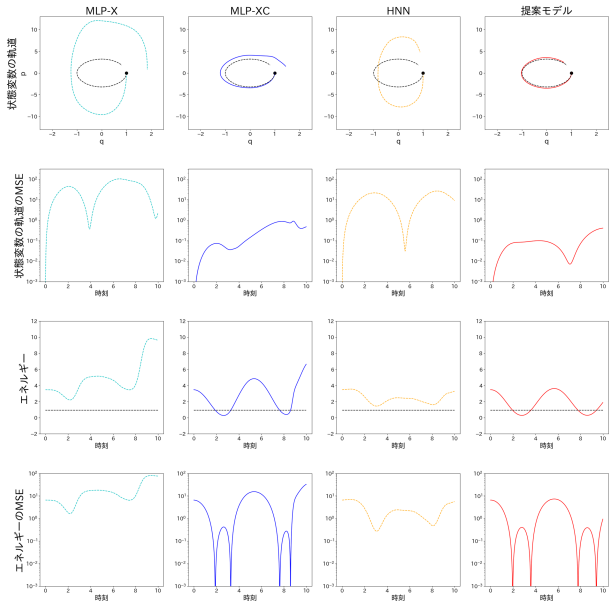


図 7　単振り子における各モデルの初期値 $(q, p) = (1, 0)$ に対する状態変数の軌道の予測とそれに対応するエネルギー，および各時刻でのそれぞれの値の MSE を示す．黒い破線は軌道の真値，黒い点は軌道の初期値である．軌道の MSE とエネルギーの MSE は対数スケールで示している．

ついてのデータ $\{\mathcal{D}_d\}_{d\in D}$ と物理パラメータ $\{\boldsymbol{c}_d\}_{d\in D}$ を用いてモデルの学習を行う．その後予測フェーズにて，学習に使用していない未知のタスク $d^* \notin D$ の系の軌道 I_{d^*} 個の集合であるデータ $\mathcal{D}_{d^*} = \left\{\{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit})\}_{t=1}^{T}\right\}_{i=1}^{I_{d^*}}$ と物理パラメータ $\boldsymbol{c}_{d^*}$ を用いて d^* の系を推定し，d^* の系において観測された状態変数 $\boldsymbol{x}_{d^*}$ と $\boldsymbol{c}_{d^*}$ を用いて状態変数の時間微分 $\dot{\boldsymbol{x}}_{d^*}$ を予測する．

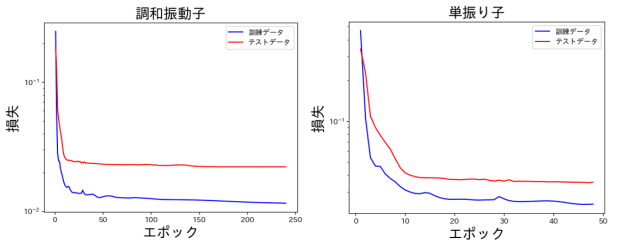


図 8　学習中の訓練データとテストデータに対する損失の変化を示す．左は調和振動子，右は単振り子の学習中についてのグラフである．損失は対数スケールで示している．

5.2　Model-Agnostic Meta-Learning の適用

Algorithm 1 提案モデルに MAML を適用した手法の学習アルゴリズム

Require: データ集合 $\{\mathcal{D}_d\}_{d\in D}$
Ensure: 学習後のパラメータ θ
1: タスク共通パラメータ θ を初期化
2: D からタスク d をサンプリング
3: **for** $d \in D$ **do**
4:　$\mathcal{D}_d$ よりサポート集合 $\mathcal{S}$ とクエリ集合 $\mathcal{Q}$ をサンプリング
5:　タスク専用パラメータ θ_d にタスク共通パラメータ θ を代入
6:　**for** L 回 **do**
7:　　目的関数 (13)，(14) とその勾配を計算
8:　　パラメータ θ_d を目的関数と勾配から更新
9:　**end for**
10:　目的関数 (15)，(16) とその勾配を計算
11:　パラメータ θ を目的関数と勾配から更新
12: **end for**

上記の問題に対し，提案手法に Model-Agnostic Meta-Learning (MAML) [14] を適用したモデルを使用する．MAML とは，複数のタスクを用いた学習により，新しいタスクに適合した際に高い性能を示すようなタスク共通の初期パラメータを探索するためのアルゴリズムである．状態変数の時間微分 $\dot{\boldsymbol{x}}$ の推定には式 (10) を用いる．MAML では Algorithm 1 のように学習を進める．サポート集合 $\mathcal{S}$ とクエリ集合 $\mathcal{Q}$ は，データ $\mathcal{D}_d$ に含まれる I 個の軌道よりそれぞれ $I_{\mathcal{S}}$ 個，$I_{\mathcal{Q}}$ 個の軌道をサンプリングして決定する．2 重ループで行われる学習のうち，内側のループではタスク d に適応したタスク専用パラメータを推定するため目的関数を

$$\hat{\theta}_d = \arg\min_{\theta} \mathcal{L}\left(\mathcal{S}; \theta\right) \tag{13}$$

$$\mathcal{L}_d\left(\mathcal{S}; \theta\right) = \frac{1}{I_{\mathcal{S}} T} \sum_{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit}) \in \mathcal{S}} \left\| \dot{\boldsymbol{x}}_{dit} - \Omega \nabla_{\boldsymbol{x}} H_\theta\left(\boldsymbol{x}_{dit}, \boldsymbol{c}_d\right) \right\|_2^2 \tag{14}$$

として勾配降下法を用いてパラメータの更新を行う．本研究では，内側のループ回数を $L = 5$ とした．また，外側のループで用いる目的関数は

$$\hat{\theta} = \arg\min_{\theta} \sum_{d \in D} \mathcal{L}\left(\mathcal{Q}; \theta_d\right) \tag{15}$$

$$\mathcal{L}_d\left(\mathcal{Q}; \theta_d\right) = \frac{1}{I_{\mathcal{Q}} T} \sum_{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit}) \in \mathcal{Q}} \left\| \dot{\boldsymbol{x}}_{dit} - \Omega \nabla_{\boldsymbol{x}} H_{\theta_d}\left(\boldsymbol{x}_{dit}, \boldsymbol{c}_d\right) \right\|_2^2 \tag{16}$$

として確率的勾配降下法を用いてパラメータ更新を行う．

予測フェーズでは，未知のタスク $d^* \notin D$ についてのデータ $\mathcal{D}_{d^*}$ をサポート集合 $\mathcal{S}$ と見なし，式 (14) を用いてタスク d^* に適合したパラメータ θ_{d^*} を得る．

5.3 Conditional Neural Processes の適用

5.3.1 モ デ ル

提案手法に Conditional Neural Processes (CNP) [15] を適用したモデルについて説明する．CNP とは各タスクのデータからエンコーダによってタスク別の特徴ベクトルを得ることでタスクへの適合をモデル化するメタ学習の手法である．あるタスク d に関する T 時刻間の $I_{\mathcal{S}}$ 個の軌道についてのサポート集合 $\mathcal{S} = \left\{ \{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit})\}_{t=1}^{T} \right\}_{i=1}^{I_{\mathcal{S}}}$ と状態変数 $\boldsymbol{x}_d$，および物理パラメータ $\boldsymbol{c}_d$ が与えられた場合を考える．まず，$(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit}) \in \mathcal{S}$ を用いて，ニューラルネットワーク $f : \mathbb{R}^{2N} \times \mathbb{R}^{2N} \to \mathbb{R}^K$ により特徴ベクトル $\boldsymbol{r}_{dit} \in \mathbb{R}^K$ を得る．式は以下の通りである．

$$\boldsymbol{r}_{dit} = f\left(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit}\right) \tag{17}$$

その後，以下の式のように $\{\{\boldsymbol{r}_{dit}\}_{t=1}^{T}\}_{i=1}^{I_{\mathcal{S}}}$ を平均することで，特徴ベクトル $\boldsymbol{r}_d$ を得る．

$$\boldsymbol{r}_d = \frac{1}{I_{\mathcal{S}}} \sum_{i=1}^{I_{\mathcal{S}}} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{r}_{dit} \tag{18}$$

そして特徴ベクトル $\boldsymbol{r}_d$ と物理パラメータ $\boldsymbol{c}_d$ より，ニューラルネットワーク $a : \mathbb{R}^K \times \mathbb{R}^B \to \mathbb{R}^K$ を用いて以下の式のようにタスク表現 $\boldsymbol{r}'_d$ を計算する．

$$\boldsymbol{r}'_d = a\left(\boldsymbol{r}_d, \boldsymbol{c}_d\right) \tag{19}$$

その後，タスク表現 $\boldsymbol{r}'_d$ と状態変数 $\boldsymbol{x}_d$ より状態変数の時間微分 $\hat{\dot{\boldsymbol{x}}}_d$ を推定する．

$$\hat{\dot{\boldsymbol{x}}}_d = \Omega \nabla_{\boldsymbol{x}} H^{CNP}\left(\boldsymbol{x}_d, \boldsymbol{r}'_d\right) \tag{20}$$

ここで，$H^{CNP} : \mathbb{R}^{2N} \times \mathbb{R}^K \to \mathbb{R}$ はタスク表現 $\boldsymbol{r}'_d$ に依存するハミルトニアンをモデル化したニューラルネットワークである．このモデルの概要を図 9 に示す．学習アルゴリズムの説明は次節で行う．

5.3.2 学 習 方 法

ニューラルネットワーク f，a，H^{CNP} のパラメータを $\boldsymbol{\Phi}$ とする．学習アルゴリズムを Algorithm 2 に示す．最適なパラメータ $\boldsymbol{\Phi}$ を推定するための目的関数は

$$\hat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} \mathbb{E}_{d \sim D}\left[\mathbb{E}_{(\mathcal{S}, \mathcal{Q}) \sim \mathcal{D}_d}\left[\mathcal{L}\left(\mathcal{Q} | \mathcal{S}\right)\right]\right] \tag{21}$$

$$\mathcal{L}\left(\mathcal{Q} | \mathcal{S}\right) = \frac{1}{I_{\mathcal{Q}} T} \sum_{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit}) \in \mathcal{Q}} \left\| \dot{\boldsymbol{x}}_{dit} - \Omega \nabla_{\boldsymbol{x}} H_{\theta}^{CNP}\left(\boldsymbol{x}_{dit}, \boldsymbol{r}'_d\right) \right\|_2^2 \tag{22}$$

Algorithm 2 提案モデルに CNP を適用した手法の学習アルゴリズム

Require: データ集合 $\{\mathcal{D}_d\}_{d \in D}$
Ensure: 学習後のパラメータ $\boldsymbol{\Phi}$
1: D からタスク d をサンプリング
2: **for** $d \in \mathcal{D}$ **do**
3: $\mathcal{D}_d$ よりサポート集合 $\mathcal{S}$ とクエリ集合 $\mathcal{Q}$ をサンプリング
4: 目的関数 (21)，(22) とその勾配を計算
5: パラメータ $\boldsymbol{\Phi}$ を目的関数と勾配から更新
6: **end for**

となり，これらの式を用いてパラメータの更新を行う．$\mathbb{E}[\cdot]$ は期待値を表す．

5.4 実 験 設 定

調和振動子，単振り子の 2 種類の力学モデルについて実験を行う．学習には前章での実験と同じデータ $\{\mathcal{D}_d\}_{d \in D}$ と物理パラメータ $\{\boldsymbol{c}_d\}_{d \in D}$ を用いる．$\mathcal{D}_d = \left\{ \{(\boldsymbol{x}_{dit}, \dot{\boldsymbol{x}}_{dit})\}_{t=1}^{T} \right\}_{i=1}^{I}$ は 20 個の軌道より構成されており，学習フェーズにてサンプリングするサンプル集合 $\mathcal{S}$ とクエリ集合 $\mathcal{Q}$ の軌道の数 $I_{\mathcal{S}}$ と $I_{\mathcal{Q}}$ はいずれも 10 とする．未知のタスク $d^* \notin D$ についてのデータ $\mathcal{D}_{d^*}$ は $\mathcal{D}_d$ とは異なる物理パラメータの制約範囲より生成されており，10 個の軌道より構成される．本実験では，タスク d^* に対する適合性能を評価するため 10 個の軌道の集合であるテストデータ $\mathcal{D}_{d^*}^{test}$ を用意した．$\mathcal{D}_{d^*}, \mathcal{D}_{d^*}^{test}$ はそれぞれ 100 セット用いて評価をする．タスク $d^* \notin D$ の系を決定する物理パラメータの分布を以下に説明する．

5.4.1 調和振動子

物理パラメータは $(m, k) \in ([2.5, 4], [2.5, 4])$ の範囲における一様乱数より決定した．また，状態変数の軌道の初期値は $(q, p) \in \left([-5, 5]^2\right)$ の範囲における一様乱数より決定した．

5.4.2 単 振 り 子

物理パラメータは $(m, k) \in ([2.5, 4], [1.5, 2])$ の範囲における一様乱数より決定した．また，座標の初期値は $q = 0$，運動量の初期値は $p \in [-\pi, \pi]$ の範囲における一様乱数より決定した．

5.5 比較するモデル

比較するモデルは以下の 4 つである．

(a) MLP-XC
状態変数 $\boldsymbol{x}$ と系の物理パラメータを特徴量とする MLP で，64 ノードからなる 3 層の隠れ層によって構成される．学習率 0.002 のもと 1000 エポックの学習を行った．

(b) HNN
64 ノードからなる 3 層の隠れ層によって構成される．学習率 0.002 のもと 1000 エポックの学習を行った．

(c) 提案モデル
前章で説明している提案手法である．64 ノードからなる 3 層の隠れ層によって構成される．学習率 0.002 のもと 1000 エポッ

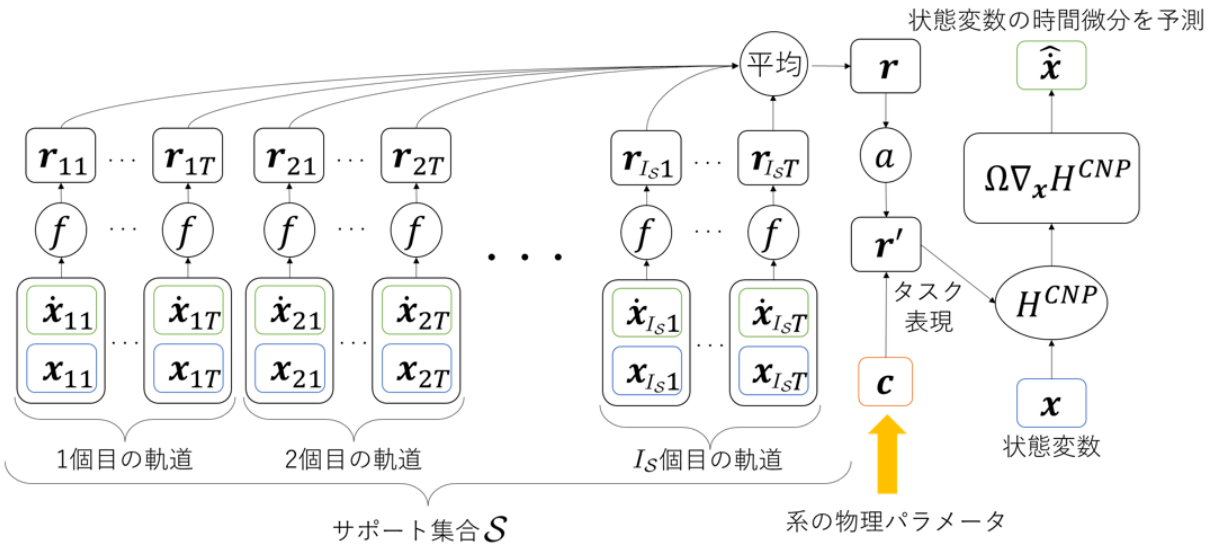


図 9　提案モデルに CNP を適用した手法.（タスクを表す添え字 d は簡単のため省略した.）

クの学習を行った.

(d) 提案モデル + MAML
(b) と同様，64 ノードからなる 3 層の隠れ層によって構成される．内側のループでは学習率 0.001 のもと 5 エポックのパラメータ更新を行い，外側のループでは学習率 0.002 のもと 200 エポックのパラメータ更新を行った.

(e) 提案モデル + CNP
f_θ，H_θ^{CNP} はそれぞれ 64 ノードからなる 3 層の隠れ層によって，g_θ は 128 ノードからなる 1 層の隠れ層によって構成される．特徴ベクトル $\boldsymbol{r}_{dit}$ および $\boldsymbol{r}'_d$ は 128 次元とした．学習率 0.002 のもと 1000 エポックの学習を行った.

全モデルについて，各層の活性化関数には Softplus を用いた．(d) 以外のモデルの最適化には Adam を使用し，バッチサイズは 10 系分のデータ (10 系× 20 軌道× 50 時刻) とした.

5.6 評価方法

MLP-XC，HNN，提案モデル，提案モデル+MAML では未知のタスク $\mathcal{D}_{d^*}$ により 5 エポックの学習を行うことでタスク d^* に適合させ，テストデータ $\mathcal{D}_{d^*}^{test}$ と物理パラメータ $\boldsymbol{c}_{d^*}$ によってテスト誤差を計算する．提案モデル + CNP では，$\mathcal{D}_{d^*}$ によってタスク表現 $\boldsymbol{r}'_{d^*}$ を計算し，$\boldsymbol{r}'_{d^*}$ と $\mathcal{D}_{d^*}^{test}$，物理パラメータ $\boldsymbol{c}_{d^*}$ を用いてテスト誤差を計算する．また，タスク適合後のモデルを用いて式 (2) より 100 時刻分の軌道とエネルギーを積分計算し，真値となる軌道とのエネルギー MSE を評価に用いる.

6 実験結果

テスト誤差と生成した状態変数の軌道のエネルギーの MSE を表 2 に示す．また，生成した状態変数の軌道の 1 つとそれに対応するエネルギー，および各時刻のそれぞれの値の MSE を図 10，11 に示す．調和振動子については提案モデル+CNP がテスト誤差とエネルギーの MSE ともに高い性能を示した．単振り子については提案モデル+MAML が高い性能を示した．前章の実験にも使用したモデルの精度が前章に比べて悪いのは，各モデルの性能の差を顕著に示すため，適合したい未知のタスク d^* の系を決定する物理パラメータの生成範囲を訓練データのタスク $d \in D$ の生成範囲と異なるものにして系の推定を難しくしたためである．MLP-XC はエネルギー保存則を事前知識として考慮していないため，提案モデルを用いた手法と比較すると未知の系に適合する性能は低い．HNN は 1 つの系のみを学習することを前提としているため，他の系との違いや共通する性質を学習することはできず，未知のタスク d^* の系に適合するには d^* についてのデータ数が少なかったため，性能が低いと考えられる．MAML や CNP を適用したモデルは提案モデル以上にタスクに適合する性能が高いことがわかる.

7 結　　論

本研究では，力学系の物理パラメータを特徴量に加えるよう HNN を拡張し，物理パラメータに依存するハミルトニアンを学習するモデルを提案した．実験を通して，異なる物理パラメータを持つ系を学習し，学習していない系を推定する性能が既存の手法より高いことを確認した．また，既存のメタ学習手法を提案モデルに適用した手法を提案し，メタ学習を用いない手法に比べ，学習していない系のデータを与えられた際に高い性能で適合することを示した.

今後の課題としては，本研究に使用した調和振動子や単振り子に限らず，様々な力学系についてのデータを使用することで，提案手法に対する理解を深めることが挙げられる．また，エネルギー保存則を満たす条件下での力学系に限らず，エネルギー散逸や外部入力のある，複雑な系に対しても同様のアプローチが有効であることを示す実験などが考えられる.

表 2　5 エポックのタスク適合後のテスト誤差と予測した軌道に対応するエネルギー MSE (MSE と標準偏差を示す．最も精度の良い値を太字で示す．)

力学系	学習モデル	テスト誤差	エネルギーの MSE
調和振動子	MLP-XC	1.798 ± 2.521	67.410 ± 155.907
	HNN	57.673 ± 78.275	1.548 ± 2.904
	提案モデル	0.541 ± 1.516	0.192 ± 0.410
	提案モデル + MAML	0.473 ± 1.135	0.081 ± 0.192
	提案モデル + CNP	**0.462 ± 0.859**	**0.040 ± 0.054**
単振り子	MLP-XC	2.349 ± 5.069	989.447 ± 1943.090
	HNN	28.798 ± 70.523	1.584 ± 2.970
	提案モデル	1.436 ± 3.065	0.077 ± 0.172
	提案モデル + MAML	**0.895 ± 1.559**	**0.066 ± 0.230**
	提案モデル + CNP	1.026 ± 2.426	0.072 ± 0.167

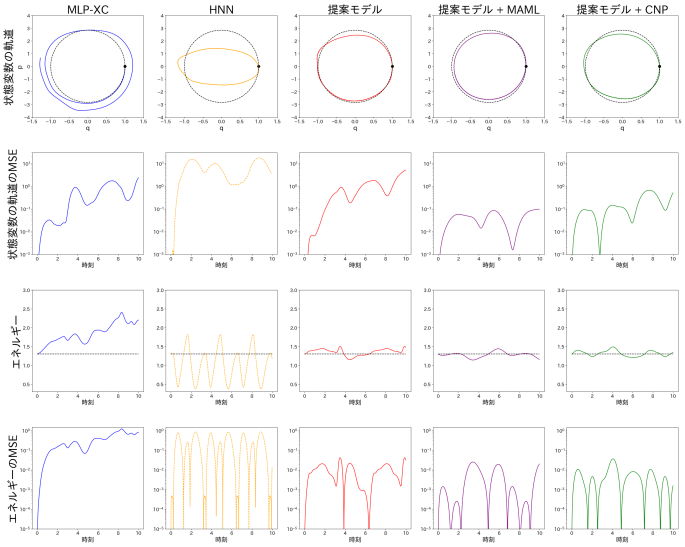


図 10　調和振動子における各モデルの初期値 $(q, p) = (1, 0)$ に対する状態変数の軌道の予測とそれに対応するエネルギー，および各時刻でのそれぞれの値の MSE を示す．黒い破線は正解の軌道，黒い点は軌道の初期値である．軌道の MSE とエネルギーの MSE は対数スケールで示している．

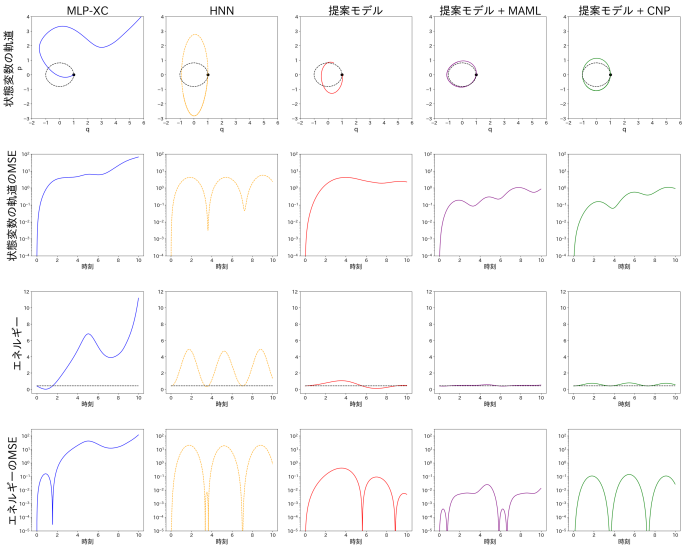


図 11　単振り子における各モデルの初期値 $(q, p) = (1, 0)$ に対する状態変数の軌道の予測とそれに対応するエネルギー，および各時刻でのそれぞれの値の MSE を示す．黒い破線は正解の軌道，黒い点は軌道の初期値である．軌道の MSE とエネルギーの MSE は対数スケールで示している．

文　献

[1] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, Vol. 31, , 2018.

[2] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, Vol. 32, , 2019.

[3] Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Learning neural event functions for ordinary differential equations. *International Conference on Learning Representations*, 2021.

[4] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6696–6707. Curran Associates, Inc., 2020.

[5] Zhengdao Chen, Jianyu Zhang, Martín Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[6] Marc Finzi, Ke Alexander Wang, and Andrew Gordon Wilson. Simplifying hamiltonian and lagrangian neural networks via explicit constraints. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[7] Andrew Sosanya and Sam Greydanus. Dissipative hamiltonian neural networks: Learning dissipative and conservative dynamics separately. *CoRR*, Vol. abs/2201.10085, , 2022.

[8] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Dissipative symoden: Encoding hamiltonian dynamics with dissipation and control into deep learning. *CoRR*, Vol. abs/2002.08860, , 2020.

[9] Shaan A Desai, Marios Mattheakis, David Sondak, Pavlos Protopapas, and Stephen J Roberts. Port-hamiltonian neural networks for learning explicit time-dependent dynamical systems. *Physical Review E*, Vol. 104, No. 3, p. 034312, 2021.

[10] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa,*

Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[11] Zhengdao Chen, Jianyu Zhang, Martín Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[12] Kookjin Lee and Eric J. Parish. Parameterized neural ordinary differential equations: Applications to computational physics problems. In *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 477, 2021.

[13] Haesang Yang Seungjun Lee and Woojae Song. Identifying physical law of hamiltonian systems via meta-learning. *International Conference on Learning Representations*, 2021.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

[15] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018.

[16] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, Vol. abs/1707.09835, , 2017.

[17] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[18] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2017.

[19] Adam Santoro, Sergey Bartunov, Matthew M. Botvinick, Daan Wierstra, and Timothy P. Lillicrap. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, 2016.

[20] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

[21] Jean Kaddour, Steindor Saemundsson, and Marc Deisenroth (he/him). Probabilistic active meta-learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 20813–20822. Curran Associates, Inc., 2020.

ブラックボックス関数が与えられた予算制約付組合せ最適設備投資問題に対するモンテカルロ木探索法の提案

行木　渉真† 　中村　篤祥†† 　安ヶ平裕介††† 　内垣内　洋††† 　平松　道晶†††
竹本　享史†††

† 北海道大学工学部情報エレクトロニクス学科　〒 060–8628 北海道札幌市北区北 13 条西 8 丁目
†† 北海道大学大学院情報科学研究院　〒 060–0814 北海道札幌市北区北 14 条西 9 丁目
††† 株式会社日立製作所 日立北大ラボ　〒 001–0021 北海道札幌市北区北 21 条西 11 丁目
E-mail: †nameki@ist.hokudai.ac.jp, ††atsu@ist.hokudai.ac.jp,
†††{yusuke.yasugahira.yc,hiroshi.uchigaito.nx,michiaki.hiramatsu.nq,takashi.takemoto.tj}@hitachi.com

あらまし　モンテカルロ木探索 (Monte Carlo tree search; MCTS) は、木で表された探索空間の部分空間をランダムサンプリングによって評価し、有望な部分空間を優先して探索する手法である。本論文では、ブラックボックス関数が与えられた予算制約付き組合せ最適設備投資問題に対して MCTS を用いる方法を提案する。具体的には、特徴選択問題に対して MCTS を適用する手法である FUSE アルゴリズムをベースとして、予算制約の下に無向グラフの頂点上に複数の施設を配置し、それぞれの施設に割り当てる多種類の装置の数量を最適化する問題に適用できるように MCTS アルゴリズムを設計する。また、実際の地図・人口・気候に基づくシミュレーションによる評価値をブラックボックス関数として用い、複数ナノグリッド最適設備投資設計へ適用し有効性を検証する。

キーワード　地理データ活用，スマートシティー，資源・エネルギー，設備投資最適化，ブラックボックス最適化，モンテカルロ木探索

1　はじめに

近年の持続可能な地域社会への関心の高まりを受け、地域産業の低炭素化や防災機能強化を目標とし、ナノグリッドと呼ばれる発電・蓄電設備や EV によって構成された、地産地消・自立型地域エネルギーシステムの構築のための取り組みが進められている [1]。

しかし実際の都市上へのナノグリッドによるエネルギーシステムの構築は非常に大規模な取り組みとなることから、使用可能な予算を効率的に活用するため、施設の設置数と各施設の設置位置、各施設内の複数の装置の種類・数量の可能な限り全てを最適化することが望ましい。目的関数が明示的に与えられていれば最急降下法や Adam 法等の一般の最適化手法による解決が望めるが、本問題における施設配置の評価はナノグリッドの電力収支や配置した EV による輸送需要への対応状況、発電時の環境負荷や災害時の対処能力など多様な観点による複雑な評価及び天候や人的要因による変動の考慮を必要とするため、目的関数は明示的に与えられず勾配などを用いた最適化は困難である。

そこで、本研究では複数ナノグリッドの配置とそれぞれのナノグリッドに設置する設備と EV の初期位置等をシミュレーションにより評価し、予算内で最適化する問題を予算制約付組合せ最適設備投資問題として定式化し、この問題の解法アルゴリズムを提案する。具体的には、モンテカルロ木探索 (Monte Carlo tree search; MCTS) [2] と呼ばれる木で表された探索空間の部分空間をランダムサンプリングによって評価し、有望な部分空間を優先して探索する手法を用い、制約なしの 0-1 最適化問題である特徴選択問題に対して開発された FUSE アルゴリズム [3] をベースとして、予算制約付組合せ最適設備投資問題に適用できるように MCTS アルゴリズムを設計し、複数ナノグリッド最適設備投資設計への適用によってその有効性を検証する。

2　MCTS を用いた特徴選択アルゴリズム

モンテカルロ木探索 (Monte Carlo tree search; MCTS) [4] は、木で表された探索空間の部分空間 (部分木) をランダムサンプリングによって評価し、有望な部分空間を優先して探索する手法であり、碁 [5, 6] やオセロ [7] などのボードゲームにおける最適手の探索や、特定の吸収波長を持つ分子の設計 [8]、機械学習における特徴選択の最適化 [3] などに適用されている。

MCTS における木の探索は、根ノードから開始して葉ノードに到達するまである基準に従って子ノードを選択することを繰り返す**選択**、到達した葉ノードが終端ノードでなければ、そのノードから新たな子ノードを生成する**展開**、展開された子ノードから生成しうる終端ノードからランダムに抽出を行い、これについて評価を行う**シミュレーション**、シミュレーション結果に基づいて、探索経路上の各ノードの評価値を更新する**バックプロパゲーション**の 4 つの操作を繰り返すことで行われる (図 1) [2]。選択操作における子ノードの選択基準としては、UCB1(Upper Cnfidence Bound 1) 方策 [9]((1) 式) や EXP3

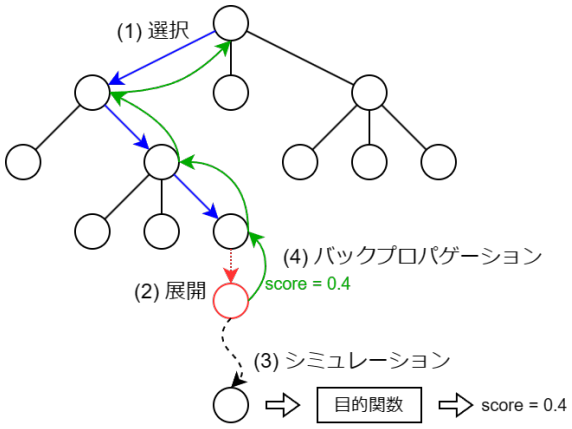


図 1　MCTS による探索のイメージ

方策 (Exponential-weight algorithm for Exploration and Exploitation) [10] に基づくものなど様々な手法が提案されており [2]、特に UCB 方策が適用された MCTS は、UCT(Upper Confidence Tree) [11] と呼ばれる。

Gaudel らの提案した FUSE アルゴリズム [3] は、UCT を 0-1 最適化問題である特徴選択問題に対して適用したものである。より詳細には、One-Player Game アプローチによって特徴選択問題を強化学習として定式化することで理論上最適な特徴選択方策を導出し、UCT による特徴集合の集合束上の探索によってこれを近似したものが、FUSE アルゴリズムである。MCTS を碁 [6] やオセロ [7] などのゲーム AI に適用した研究において、MCTS が扱う探索空間として、ゲーム木と呼ばれる各ノードに現在の盤面、各エッジに取りうる手、各子ノードに次の盤面を対応付けた木を用いるように、FUSE アルゴリズムにおいても、根ノードに空集合、エッジに新たな特徴の追加あるいは追加の終了を対応付ける。このとき、集合への追加操作はその追加の順序によらないことから、複数の探索経路から同一の状態に到達する場面が多く出現するため、FUSE アルゴリズムは探索空間として純粋な木でなく、feature lattice と呼ばれる束状の空間を扱う。

FUSE アルゴリズムでは、選択操作における指標として UCB1((1) 式) でなく、よりタイトな指標である UCB-tuned [9]((2) 式) の適用や、MoGo [5] などで使用される AMAF(All Moves As First) ヒューリスティックの改良であって、探索経路に依存しない各特徴の評価である RAVE(Rapid Action Value Estimation) スコア [3] による暫定的な腕の評価の活用によって、探索の効率化が図られている。

本論文では、UCT を特徴選択問題に対して適用する手法である、FUSE アルゴリズム [3] をベースとして、MCTS を予算制約付組合せ最適設備投資設計へと適用するアルゴリズムを提案する。

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left(\mu_a + \sqrt{\frac{2 \ln N}{n_a}} \right) \tag{1}$$

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left(\mu_a + \sqrt{\frac{2 \ln N}{n_a} \min \left(\frac{1}{4}, \sigma_a^2 + \sqrt{\frac{2 \ln N}{n_a}} \right)} \right) \tag{2}$$

a^*　選択する行動
$\mathcal{A}$　取りうる行動の集合
μ_a　行動 a をとった時に得られた報酬の標本平均
N　総選択回数
n_a　行動 a の選択回数
σ_a^2　行動 a をとった時に得られた報酬の標本分散

3　問題設定

本論文では、予算制約のもとに無向グラフの任意個数の頂点上に施設を配置し、与えられたノイズ入りブラックボックス関数上で最大の期待評価値が得られる設備投資設計を探索する問題を扱う。

この問題は以下のように定式化される。

与えられたグラフを $G = (V, E)$ とする。施設の種類の集合を $\{1, 2, \ldots, K\}$、種類 i の施設の数を n_i とする。種類 i の h 番目の施設の設計は m_i 次元のフィールド変数ベクトル $\mathbf{w}_{i,h} = (\mathrm{w}_{i,h,1}, \mathrm{w}_{i,h,2}, \ldots, \mathrm{w}_{i,h,m_i})$ を設定することで決定する。変数 $w_{i,h,j}$ の取りうる値の集合は、何番目であるかには依存せず $R_{i,j}$ とする。与えられたノイズ入りブラックボックス関数 F は、$((n_1, \mathbf{w}_{1,1}, \ldots, \mathbf{w}_{1,n_1}), (n_2, \mathbf{w}_{2,1}, \ldots, \mathbf{w}_{2,n_2}), \ldots, (n_K, \mathbf{w}_{K,1}, \ldots, \mathbf{w}_{K,n_K}))$ を入力すると、実数評価値を出力するものとする。与えられた予算を B とし、変数 $\mathrm{w}_{i,h,j}$ に対するコスト関数 $c_{i,j}$ が設定され、$\mathrm{w}_{i,h,j}$ の値に依存して $c_{i,j}(\mathrm{w}_{i,h,j})$ のコストがかかるものとする。このとき、以下の予算制約付組合せ最適設備投資問題を考える。

$$\max \mathbb{E}\left[F((n_1, \mathbf{w}_{1,1}, \ldots, \mathbf{w}_{1,n_1}), (n_2, \mathbf{w}_{2,1}, \ldots, \mathbf{w}_{2,n_2}), \ldots, (n_K, \mathbf{w}_{K,1}, \ldots, \mathbf{w}_{K,n_K}))\right]$$

$$\text{s.t.} \quad \sum_{i=1}^{K} \sum_{h=1}^{n_i} \sum_{j=1}^{m_i} c_{i,j}(\mathrm{w}_{i,h,j}) \leqq B$$

$$n_i \in \mathbb{N} \quad (i = 1, 2, \ldots, K)$$

$$\mathrm{w}_{i,h,j} \in R_{i,j}$$

$$(i = 1, 2, \ldots, K,\ j = 1, 2, \ldots, m_i,\ h = 1, 2, \ldots, n_i)$$

各フィールド変数 $\mathrm{w}_{i,h,j}$ は数値型、位置型、カテゴリ型のいずれかであり、それぞれ以下の条件を満たすものとする。

- 数値型の変数は実数あるいは整数の変数であり、変数に依存する有限の範囲 $[a, b]$ の値をとる。
- 位置型の変数は位置情報を表す変数であり、与えられた無向グラフ上の頂点番号を値にとる。
- カテゴリ型の変数は $\{1, 2, \ldots, k\}$ の中の 1 つの値をとる。ただし、k は変数に依存する整数とする。

〈補足〉 上述の問題設定ではグラフの接続関係が全く使われておらず無関係のようにみえる。実際には施設はある程度分散して配置した方が評価関数 F の値が大きくなることを仮定しており、接続関係は施設の最適配置に大きく影響する。本稿では位置型の値の階層構造をグラフの近接関係を用いて定義することにより与えられたグラフにおける (辺の重みや頂点の重みまで考慮した) 接続関係を取り込んでいる (4.1 節の位置変数に対する値域分割木の構築法の説明参照)。

4 提案アルゴリズム

本節では、本論文で提案する、予算制約付組合せ最適設備投資設計のアルゴリズムについての詳細を述べる。

4.1 探索状態空間

問題の探索空間 S は以下のように表現できる。種類 i の h 番目の施設はフィールド変数ベクトル $\mathbf{w}_{i,h} = (\mathrm{w}_{i,h,1}, \mathrm{w}_{i,h,2}, \ldots, \mathrm{w}_{i,h,m_i})$ で設計され、$\mathrm{w}_{i,h,j} \in R_{i,j}$ であるから、$\mathbf{w}_{i,h} \in \prod_{j=1}^{m_i} R_{i,j}$ である。$S_i = \prod_{j=1}^{m_i} R_{i,j}$ とする。種類 i の施設を n_i 個ずつ配置した場合の取りうる状態の集合を $S_{(n_1,\ldots,n_K)}$ とすると、$S_{(n_1,\ldots,n_K)} = \prod_{k=1}^{K} S_k^{n_k}$ となる。各施設は任意の個数を配置できるが、本論文で扱う問題は予算制約付きのものであるため、ある十分大きい値 M をとり、$P = \{0, \ldots, M\}^K$ を用いて得られる $S = \bigcup_{p \in P} S_p$ によって、解が取りうる全ての状態を含む集合 S が得られる。

値を徐々に絞り込む探索を行うために、変数 $\mathrm{w}_{i,h,j}$ は探索空間 S の値のみでなく、値の範囲をとることができるようにする。数値型フィールド変数 $\mathrm{w}_{i,h,j}$ に対しては、値域 $R_{i,j} = [a, b]$ を階層的に分割した**値域分割木**を構築し、各ノードに対応する値の範囲を $\mathrm{w}_{i,h,j}$ の値として設定できるようにする。位置型フィールド変数 $\mathrm{w}_{i,h,j}$ に対しては、与えられたグラフ $G = (V, E)$ を階層的に連結領域分割した場合の頂点集合 V に対応する値域分割木を構築し、各ノードに対応する頂点集合を値として設定可能とする。カテゴリ変数に関しては、全ての値の集合に対応する根ノードと、各々の値に対応する葉ノードからなる高さ 1 の木を値域分割木とする。各種類 i の施設の数を表す変数 n_i に関しても、予算制約から範囲が $R_i = [0, M]$ に定まるため数値型変数として値域分割木を構築し、ノードに対応する範囲としての値を設定して探索する。

各 $R_{i,j}, R_i$ の値域分割木の各ノードに対応する値の集合をそれぞれ $\bar{R}_{i,j}, \bar{R}_i$ とする。探索中は、変数 $\mathrm{w}_{i,h,j} \in \bar{R}_{i,j}$ $(i = 1, 2, \ldots, K,\ j = 1, 2, \ldots, m_i, h = 1, 2, \ldots, n_i)$、$n_i \in \bar{R}_i$ $(i = 1, 2, \ldots, K)$ の値で状態を表す。探索時にとる行動としては「種類 i の h 番目の施設の変数 j を値 a に変更する」$(\mathrm{w}_{i,h,j} \leftarrow a)$、または「種類 i の施設の数 n_i を値 a に変更する」$(n_i \leftarrow a)$ の 2 つを考える。

数値型の値域分割木の例としては、根ノードに範囲 $[a, b]$ を、その子ノードにはそれぞれ $[a, (a+b)/3), [(a+b)/3, 2(a+b)/3), [2(a+b)/3, b]$ を、といったように対応付けていき、区間の幅がある値 ϵ 以下になるまで同様に分割を繰り返す、といった手順で構築された木が考えられる。純粋な 3 等分に限らず、対数的な値域の分割や、あるいは問題に対する事前知識を活用した、局所的に詳細な分割の実行など、応用に適した構築法を用いるのが効率的であろう。

また位置型の値域分割木の例としては、根ノードには与えられたグラフ $G = (V, E)$ の全頂点の集合 V を、その子ノードには (V, E) 上での 3-median による頂点のクラスタリング結果 V_1, V_2, V_3 を、といったように対応付けていき、頂点集合のサイズが 1 になるまで同様に分割を繰り返す、といった階層的な 3-median 操作の繰り返しによって木を構成する方法が考えられる。なおこの例では、3-median の計算のために問題情報としてグラフの各辺の重み (距離) とグラフの各頂点の重み (人口情報など) が与えられることを前提としている。

4.2 選択操作

選択操作の段階における子ノード選択の指標は、UCB-tuned((2) 式) をベースに、活用項 (報酬の標本平均) を RAVE スコア [3] によって補正したものを用いる。

ここで、RAVE スコアは、以下のように定義される指標である。

g-RAVE$_{i,j,x}$　種類 i の施設の j 番目の変数の値が x である解で行われたシミュレーション全ての評価値の平均値

ℓ-RAVE$_{v,i,j,x}$　頂点 v から生成可能な、種類 i の施設の j 番目の変数の値が x である解で行われたシミュレーション全ての評価値の平均値

ここで各 i, j, x は、

i　施設の種類インデックス

j　指定された種類の施設が持つ変数のインデックス

x　変数の値

を表す。また、アクション a が種類 i の施設の j 番目の変数の値を x とする行動であるとき、それぞれ g-RAVE$_a$ および ℓ-RAVE$_{v,a}$ とも表記する。

本来の MCTS においては、シミュレーションによって得られた評価値はそのイテレーションにおける探索経路上のノードのみに反映されるが、RAVE スコアの導入によって、探索経路外のノードに対してもある程度探索結果を共有することが可能となる。これによって、新たな層の探索を開始した直後など腕の本数に対して探索回数が不足している場面において、未知の腕の評価値を大まかに推定することが可能となるため、選択の効率化を図れる。

(3) 式が、提案アルゴリズムが選択操作の段階で用いる、UCB-tuned と RAVE スコアを組み合わせた指標である。

指標の活用項 $(X_{v,a})$ は、重み α, β による子ノードの平均報酬 μ_a, ℓ-RAVE, g-RAVE の加重平均で構成されている。各値の算出に使用されるサンプルの個数は μ_a, ℓ-RAVE, g-RAVE の順に少ないため、μ_a の方が局所的な依存性や相互関係をよく反映するが値は不安定で、g-RAVE の方が値は安定しているが局所的な依存性や相互関係はあまり反映されない。そのため、n_a を子ノード a の訪問回数、n'_a を子ノード a から生成可能な終端ノード全ての訪問回数の総和とすれば、n_a が十分大きい場

合は平均評価値 μ_a を、n_a が少なく、n'_a が十分大きい場合は ℓ-RAVE を、両方少ない場合は g-RAVE を重視することで、探索回数の不足状況に合わせた評価値の補正を実現している。このために、重み係数 α, β は、それぞれ n_a と n'_a の増加に伴って減少するように構成されている。

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left(X_{v,a} + Y_{v,a} \right) \tag{3}$$

$$X_{v,a} = (1-\alpha)\mu_a + \alpha((1-\beta) \cdot \ell\text{-RAVE}_{v,a} + \beta \cdot \text{g-RAVE}_a)$$

$$Y_{v,a} = \sqrt{\frac{c_e \ln n_v}{n_a} \min\left(\frac{1}{4}, \sigma_a^2 + \sqrt{\frac{2\ln n_v}{n_a}}\right)}$$

$$\alpha = \frac{c}{c + n_a}, \quad \beta = \frac{c_\ell}{c_\ell + n'_a}$$

- v 現在選択操作を行っているノード
- a^* 選択する子ノード
- $\mathcal{A}$ 選択可能な子ノードの集合
- μ_a 子ノード a における平均報酬
- σ_a^2 子ノード a における報酬の標本分散
- n_v 現在選択操作を行っているノードの訪問回数
- n_a 子ノード a の訪問回数
- n'_a 子ノード a から生成可能な終端ノード全ての訪問回数の総和
- c_e 探索項 $(Y_{v,a})$ の強さを指定するパラメータ
- c RAVE スコアの強さを指定するパラメータ
- c_ℓ g-RAVE の強さを指定するパラメータ

4.3 探索手順

MCTS におけるサンプルの抽出操作は、探索木の根から葉までのある選択指標 (UCB スコア等) による探索と、探索木の葉から新しいノードを生成し探索木を拡大する展開および、展開されたノードから生成しうる探索木の部分木からのランダムな抽出の 3 段階によって行われる。つまり、探索木に展開済みのノードであれば探索する子ノードは選択指標によって決定され、未展開のノードであれば探索はランダムに行われる。また未展開のノードを新たに選択指標の適用対象とする操作が展開である。

本提案手法では明示的な探索木の構築は行わず、「探索木に展開する」代わりに探索束上の各ノードに「(指標による) 選択操作の対象である」かそうでないかの状態を持たせることで、疑似的な探索木の構築を行う。具体的には、根ノードから探索を開始し、ノードが選択操作の対象であれば (3) 式に従って、ノードが選択操作の対象でなければランダムに探索束上を探索することで、探索空間からのサンプルの抽出を実行する。このとき、(ノードの訪問回数) $> \tau_{visit}$ を満たすノードを選択操作の対象として扱うものとする。ここで、τ_{visit} は閾値を指定するパラメータである。ただし、通常の MCTS においてノードの訪問回数や平均評価値の記録が行われるのは探索木上に展開済みのノードのみであるため、本手法においても、それらの記録は既に選択操作の対象であるノードと、その子ノードのみで行う。

また、通常の MCTS では探索木上で親ノードを遡ることで評価値のバックプロパゲーションを実現しているが、本提案手法においては、探索束上のノードは複数の親ノードを持つ場合があるため、バックプロパゲーションの対象となる探索経路上のノードの集合を V_{path} として別に保管しておくことで、評価値の伝播を行う。

4.4 アルゴリズム

アルゴリズム 1 は、本論文で提案する予算制約付組合せ最適設備投資設計のアルゴリズムである。

本アルゴリズムは、解の抽出 (5–19 行目)、解の評価 (20 行目)、各ノードの平均スコアおよび RAVE スコアの更新 (21–24 行目) を T 回繰り返すことで最適解の探索を行う。

解の抽出操作は、「すべての変数の値が未決定である」解を初期状態として、「いずれかの変数の値を決定する」操作を繰り返すことで実行される。これは、アルゴリズム 5 行目の現在のノード v を「すべての変数の値が未決定である」解に対応するノード v_{root} で初期化する操作と、7–19 行目の while 内での探索束上の遷移を繰り返す操作に対応する。ここで、8 行目の関数 `get_child_list`(v) が、現在のノードから遷移可能な子ノードの一覧 $\mathcal{C}$ を出力する関数であり、12 行目でノードが選択操作の対象であるかを確認したのち、13,16 行目で式 (3) あるいは一様な乱数によって $\mathcal{C}$ からノードを選択し v を更新することで、探索束上の遷移を実行している。またこの探索束上の遷移は「すべての変数の値が決定済みである」終端ノード ($\mathcal{C}$ が空であるノード) に到達するまで繰り返される (9 行目)。

5 実　　験

本章では、本論文の提案する予算制約付組合せ最適設備投資設計アルゴリズムの有効性を検証するため、2021 年に AtCoder で開催された「日立北大ラボ×北大コンテスト 2021「未来の自律分散型まちづくり」」 [12] で使用されたテストデータおよび設備配置評価シミュレータを用いて実験を行う。

5.1 データセットおよびシミュレータ詳細

本実験では、問題設定およびシミュレーションで使用する環境データとして、岩見沢市における実際の地図、天候、人口等から作成されたデータを用いる。地図情報は 6401 頂点からなる無向グラフで表現され、各辺には頂点間の距離、各頂点には人口、面積、電力需要量、土地面積当たりの価格の値が設定されている。またシミュレーションには、予算額、利用可能な施設の種類・コスト、天候情報、電力需要量、輸送需要量、評価日数 (1 シミュレーションで処理するシミュレータ内部のシミュレーション日数) の設定を変更した複数のテストケースを用いる (表 1)。

地図上に配置可能な施設は、発電・蓄電・給電設備であるナノグリッドと、人や電力の輸送に利用可能な EV の 2 種類である。ナノグリッドと EV の配置数および、各ナノグリッドの位置、初期充電量および 4 カテゴリの装置の種類・数量と、各 EV の初期位置、初期充電量および車種を最適化し、シミュレー

表 1　使用する問題設定

	小規模テストケース	大規模テストケース
予算額	47058	247058
評価日数	4 日	6 日
天候情報	実データをもとにそれぞれランダムに生成	
電力・輸送需要量	実データをもとにそれぞれランダムに生成	
利用可能な設備の設定	実設備データをもとにそれぞれランダムに設定	

Algorithm 1 提案アルゴリズム

施設構造体の持つ各フィールド変数について、値域分割木を構築
探索束 (hash map) を初期化
$v_{root} \leftarrow$ 探索束の根ノード
for $t = 1 \dots T$ **do**
  $v \leftarrow v_{root}$
  $V_{path} \leftarrow \{v_{root}\}$
  while true **do**
    $\mathcal{C} \leftarrow$ `get_child_list`(v)
    if $\mathcal{C} = \phi$ **then**
      break
    else
      if v の訪問回数 $> \tau_{visit}$ **then**
        $v \leftarrow \mathrm{argmax}_{c \in \mathcal{C}}$(式 (3))
        $V_{path} \leftarrow V_{path} \cup \{v\}$
      else
        $v \leftarrow$ ($\mathcal{C}$ から一様にランダムに選択)
      end if
    end if
  end while
  `score` $\leftarrow$ (v のノイズ入りブラックボックス関数による評価値)
  for $v \in V_{path}$ **do**
    v の平均スコアと訪問回数を更新
  end for
  g-RAVE、ℓ-RAVE を更新
end for
最も高いスコアを出した解 v^* を出力

関数 `get_child_list`(v)

$\mathcal{C} \leftarrow \phi$
$X \leftarrow \{n_1, \dots, n_K, \mathrm{w}_{1,1,1}, \dots, \mathrm{w}_{K,n_K,m_K}\}$
for all $x \in X$ **do**
  $i \leftarrow v$ の変数 x の値 (値域分割木のインデックス)
  for 値域分割木のノード i の各子ノード a **do**
    $c \leftarrow$「変数 x の値を a にする」操作によって得られる状態
    if c が制約を満たす **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
    end if
  end for
end for
return $\mathcal{C}$

ションにより計算される期待評価値を最大化することが目標である。

シミュレータは入力された設備配置設定について運用シミュレーションを行い、ナノグリッドの電力収支や EV による輸送需要への対応状況、発電時の環境負荷や災害発生時の対処能力等を評価し、それらを総合した評価値を出力する。(付録参照)

5.2 比較手法

実験では以下の既存手法と比較を行い、提案アルゴリズムの有効性を検証する。

SOO 方策 [13]　探索空間の階層構造を 3 分木で表し、各層 (深さ) で葉ノードに対応する部分空間の中点で関数値が最大のノードを 3 分割して展開することを繰り返すことで探索を行う手法。

optuna [14]　ハイパーパラメータ探索のためのブラックボックス最適化フレームワークで、ベイズ最適化の一種である TPE(Tree-structured Parzen Estimator) 手法による探索が可能。サンプラーは TPESampler、その他のオプションはデフォルトとした。

5.3 実験 1: 小規模テストケースにおける探索

実験 1 では、岩見沢市の地図情報から作成された予算額 47058 の比較的小規模なテストケースについて最適な施設配置の探索を行う。図 2 は、各シミュレータ呼び出し回数における評価値の推移を表したグラフである。各プロットは、各シミュレータ呼び出し回数時点までに得られた上位 10 個の解についてその平均評価値を算出したものであり、またその標準偏差を上下の半透明の領域の幅としている。

提案手法と SOO 方策を用いた実験では、概ね同等の、高い平均評価値を持つ解が得られた。本実験は予算額の少ないテストケースに対する最適化であったため、同等の平均評価値で頭打ちになったと考えられる。

本テストケースは予算額が比較的少額であり、また特定の変数 (EV の台数) の影響が支配的であった。提案手法が早期に平均評価値の高い解を発見できているのは、本提案手法の探索木は、「どの変数の値から決定するか」も含めて動的に探索可能であるため、UCB 等の選択指標の働きによって、このような重要度の高い変数を決定する選択肢を優先的に探索することが可能であったためであると考える。また、SOO 方策には「どの変数の値から決定するか」を動的に決定する機能は無いため、事前にランダムに設定した変数の決定順序に沿って探索を実行している。このため、SOO 方策による 1e8 スコアを超える解の発見は提案手法に比較して僅かに遅れたと考えられる。

optuna は他 2 手法と比較してうまく探索できていないことが確認できる。ある特定の要素に予算のほとんどを投入する場合、そのほかの変数が取りうる値は極端に制限されるため、そ

のような解からなる領域は、解空間全体に対して非常に小さく、かつ実行可能領域にない要素に囲まれた領域となる。optuna がその主な用途とするハイパーパラメータ最適化は、各ハイパーパラメータの純粋な直積空間上を探索する問題であるため、このように複雑な領域を探索する本テストケースに対しては、その性能を十分に発揮できなかったと考えられる。

図 4 は、提案手法によって得られた、高平均評価値の解の一例である。図 4 下図のような特定地域への重点的な施設配置や、図 4 上図のような広域への EV の分散配備など、設備配置として一定の妥当性がある解が生成されたと考える。

5.4 実験 2: 大規模テストケースにおける探索

実験 2 では、岩見沢市の地図情報から作成された予算額 247000 の比較的大規模なテストケースについて最適な施設配置の探索を行う。図 3 は、各シミュレータ呼び出し回数における評価値の推移を表したグラフである。実験 1 同様、各プロットは、各シミュレータ呼び出し回数時点までに得られた上位 10 個の解についてその平均評価値を算出したものであり、またその標準偏差を上下の半透明の領域の幅としている。

実験 1 とは異なり、提案手法によって SOO よりも明確に平均評価値の高い解を得られた。予算額が多い分多くの施設が配置可能であり、またその分設定すべき変数の個数も増加するため、探索過程における変数の決定順序も最適化可能である、提案手法の有利性がより強く表れたと考える。

また、探索空間や実行可能領域は実験 1 と同様に複雑で、かつそのサイズも拡大したため、optuna によって得られた解は実験 1 同様低い平均評価値の解であった。

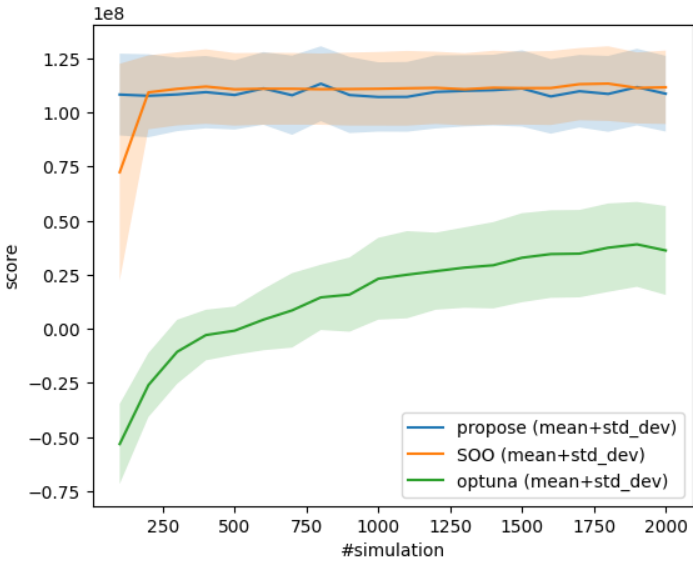


図 2　小規模テストケースにおける評価値の推移

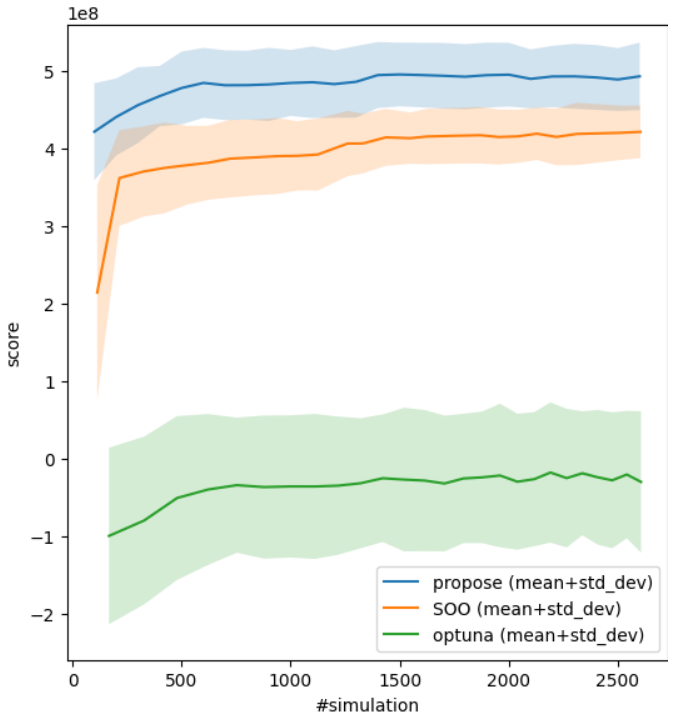


図 3　大規模テストケースにおける評価値の推移

6 おわりに

本稿では、ブラックボックス関数が与えられた予算制約付き組合せ最適設備投資問題を定式化し、MCTS アルゴリズムによる最適解の探索手法を提案した。

また実データから生成された問題を対象とした実験においては、他既存手法と比較して良い解を発見できることを確認した。

今後の課題としては、アルゴリズム (特に MCTS における選択・展開操作) の改良や、計算の並列化による探索の高速化を考えている。

文　献

[1] 株式会社日立製作所公式 Web サイト「研究開発」.「地域産業の低炭素化と防災機能強化を両立する地産地消・自立型地域エネルギーシステムを開発」. https://www.hitachi.co.jp/rd/news/topics/2021/2111_ing.html, 2020.

[2] Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 4, No. 1, pp. 1–43, 2012.

[3] Romaric Gaudel and Michele Sebag. Feature selection as a one-player game. In *Proceedings of the 27th International Conference on International Conference on Machine*

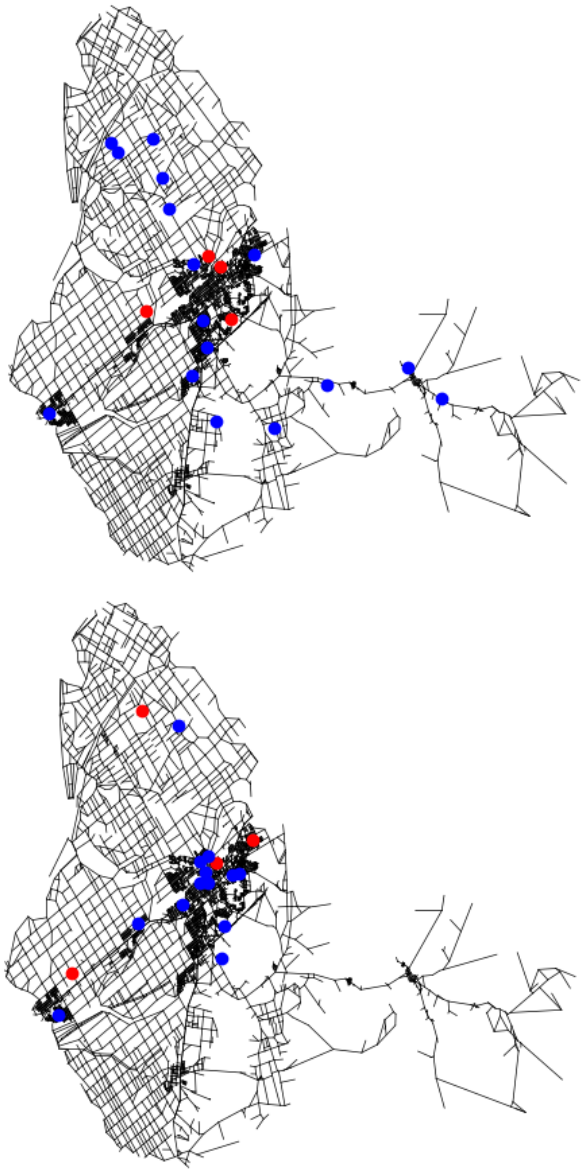


図 4　提案手法によって得られる解の一例 (赤丸がナノグリッド、青丸が EV の配置位置)

Learning, ICML'10, pp. 359–366, Madison, WI, USA, 2010. Omnipress.

[4] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *Proceedings of the 5th International Conference on Computers and Games*, CG'06, pp. 72–83, Berlin, Heidelberg, 2006. Springer-Verlag.

[5] Sylvain Gelly and David Silver. Combining online and offline knowledge in uct. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 273–280, New York, NY, USA, 2007. Association for Computing Machinery.

[6] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, Vol. 529, pp. 484–489, 01 2016.

[7] P. Hingston and M. Masek. Experiments with monte carlo othello. In *2007 IEEE Congress on Evolutionary Computation*, pp. 4059–4064, 2007.

[8] Masato Sumita, Xiufeng Yang, Shinsuke Ishihara, and Ryo Tamura. Hunting for organic molecules with artificial intelligence: Molecules optimized for desired excitation energies. Vol. 4, pp. 1126–1133, 08 2018.

[9] Djallel Bouneffouf. Finite-time analysis of the multi-armed bandit problem with known trend. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 2543–2549, 2016.

[10] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, Vol. 32, No. 1, pp. 48–77, 2002.

[11] Levente Kocsis and Csaba Szepesvari. Bandit based monte-carlo planning. In *European Conference on Machine Learning*, 2006.

[12] Hitachi hokudai lab. & hokkaido university contest 2021. `https://atcoder.jp/contests/hokudai-hitachi2021`, 2021.

[13] Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc., 2011.

[14] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

付　　録

1　評価実験 シミュレータ詳細

提案手法の評価実験に使用したシミュレータについて、その詳細を記述する。

まず、本設備投資評価シミュレータにおいて、地図上に配置可能な施設は、発電・蓄電・給電設備である**ナノグリッド**と、人や電力の輸送に使用される **EV** の 2 種類である。地図上に配置された各ナノグリッドは、その内部設定として、その配置位置、太陽光発電設備の機種・数量、燃料発電機の機種、蓄電池の機種・数量・初期蓄電量、EV チャージャーの機種の設定を必要とする。同様に、地図上に配置された各 EV は、その内部設定として、その初期配置位置と車種、初期蓄電量の設定を必要とする。

また、本設備投資評価シミュレータには、問題情報として以下の情報が与えられる。

シミュレーション時間情報　シミュレーションを行う日数や、一日分のシミュレーションにかけるステップ数を表す整数値
スコア情報　評価指標となる各種スコアの比重を表す実数値
地図情報　設備の配置可能位置やその位置関係を表す無向グラフで、各頂点には座標、人口、利用可能面積、地価の値、また各エッジにはその距離を表す重みが設定されている
電力需要情報　電力需要の設定された頂点番号のリストと、その各頂点の各シミュレーション時刻における電力需要量の期待値および分散
日射量情報　地図上の各頂点の各シミュレーション時刻における日射量の期待値
輸送需要情報　各シミュレーション時刻における輸送需要の発生頻度
避難所情報　避難所が存在する頂点番号のリストと、各避難所の収容人数および避難時の電力需要量
利用可能な設備の情報　利用可能な設備のパラメータを記述したリストで、各設備種類につきそれぞれ 2 から 4 機種程度
各設備のパラメータ

- 太陽光発電設備：1 台あたりの面積、価格
- 燃料発電機：最大出力量、価格
- 蓄電池：1 台あたりの蓄電容量、価格
- EV チャージャー：最大入力電力量、最大出力電力量、価格
- EV：充電容量、乗車可能人数、充電速度、放電速度、移動距離当たりの消費電力、価格

本設備投資評価シミュレータは、与えられた環境・問題情報のもと、入力された設備投資設計についてその運用をシミュレーションし、その評価値を返す。このとき評価は、輸送スコア、電力スコア、環境スコア、災害対応スコア、の 4 つのスコアの線形和によって算出される。

各スコアの詳細は以下の通りである。
輸送スコア　輸送依頼の達成度合いを表すスコアで、シミュレーション終了までに完了した輸送依頼の数と、時間内に完了できなかった輸送依頼の数から算出される。このとき輸送依頼は、シミュレーション中の各時刻に、各頂点の人口に応じた確率でランダムに発生する。
電力スコア　施設群全体の電力収支を表すスコアで、各ナノグリッドおよび EV の初期蓄電量とシミュレーション終了時の蓄電量の差と、シミュレーション中に系統から購入した電力量から算出される。
環境スコア　環境負荷の発生する行動に対するペナルティを表すスコアで、シミュレーション中に燃料発電機を使用した量に応じて算出される。
災害対応スコア　災害による停電への対処能力を表すスコアで、系統電力からの電力購入を禁止した状態での無停電運用日数によって算出される。

微生物学的水質予測のための符号制約付き学習アルゴリズム

高田　裕也† 　小野　晴生† 　門屋　俊祐†† 　佐野　大輔††† 　加藤　毅††††

† 群馬大学大学院理工学府　〒 376–8515 群馬県桐生市天神町 1–5–1
†† 東京大学大学院工学系研究科　〒 113–8656 東京都文京区本郷 7–3–1
††† 東北大学大学院工学研究科　〒 980–8579 仙台市荒巻字青葉 6–6–06
†††† 群馬大学情報学部　〒 371–8510 群馬県前橋市荒牧町 4–2
E-mail: †{t221d047,t221d023,katotsu.cs}@gunma-u.ac.jp, ††kadoya@env.t.u-tokyo.ac.jp,
†††daisuke.sano.e1@tohoku.ac.jp

あらまし　病原微生物の暴露に由来する水質汚染の公衆衛生上のリスクを管理することは，世界的な課題となっている．病原微生物による暴露のリスクを正しく評価するためには，物理化学的水質データから病原微生物を高精度に予測する機械学習アルゴリズムを確立する必要がある．本研究では，水質工学で蓄積されたドメイン固有の事前知識を符号制約で表し，その制約下で効率的に学習問題を解くアルゴリズムを開発した．このアルゴリズムは，確率的双対座標上昇法（SDCA）の枠組みに基づいており，最適解への線形収束が保証されている．微生物学的水質予測の実データを用いた数値実験を通して，符号制約を用いた学習の有効性を報告する．

キーワード　符号制約，凸最適化，確率的双対座標上昇法 (SDCA)，経験リスク最小化，微生物学的水質予測．

1　はじめに

水は生命にとって最も重要な資源の一つであるが，その汚染によって公衆衛生の安全が脅かされている．一見飲めるように見える透明な水であっても，実際は微生物によって汚染されており，重大な健康リスクを引き起こす可能性がある．水が病原微生物にさらされると，深刻な下痢性疾患を引き起こしており，毎年 485,000 人が死亡している [1]．国家経済にも被害を及ぼすほどの事態に発展するリスクもあることから [2]，病原微生物の暴露に由来する水質汚染の公衆衛生上のリスクを管理することは，世界的な課題となっている．

規制当局は，水質ガイドラインを策定することで公衆衛生を保全しようと努めている [1], [3]．水質ガイドラインは，最小限のモニタリングコストで微生物学的安全性を確保することを目的としている．水中の病原微生物を直接測定することはしばしばコストがかかりすぎるため，現実的ではない．その代わりに，測定が容易な物理化学的水質データによって水質管理が行われている．そのため，物理化学的水質データから病原微生物を高精度に予測できれば，適切に水質管理を行うことが可能となる．

本研究では，物理化学的水質データから病原微生物を予測するために，以下のような**経験リスク最小化**問題を考える．

$$\begin{aligned} &\min \quad P(\boldsymbol{w}) \quad \text{wrt } \boldsymbol{w} \in \mathbb{R}^d, \\ &\text{where} \quad P(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle). \end{aligned} \tag{1}$$

ただし，入力ベクトルを $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$，凸な損失関数を $\phi_1, \ldots, \phi_n : \mathbb{R} \to \mathbb{R}$，正則化定数を $\lambda > 0$ とする．

サポートベクトルマシン (SVM) は，損失関数を**ヒンジ損失** $\phi_i(s) = \max(0, 1 - y_i s)$ と設定することで表現できる．ただし，$y_i \in \{\pm 1\}$ はクラスラベルである．損失関数を **2 値交差エントロピー損失** $\phi_i(s) = \log(1 + \exp(-y_i s))$ とすると，ロジスティック回帰になる．y_i を連続値のラベルとし，損失関数を**二乗誤差損失** $\phi_i(s) = \frac{1}{2}(y_i - s)^2$ とすると，リッジ回帰になる．

線形予測器 $\langle \boldsymbol{w}, \cdot \rangle$ の重みベクトル $\boldsymbol{w} := [w_1, \ldots, w_d]^\top \in \mathbb{R}^d$ の学習タスクにおいて，高い汎化性能を得るためには，十分な個数の訓練用例題が必要である．しかし，微生物学的水質予測に対しては，データの収集に多大なコストがかかるため，訓練用例題の規模は限られる．訓練用例題の不足による学習性能の低下を抑制するには，事前知識の活用が有効である．

水質工学の分野では，一部の物理化学的水質データと病原微生物との間に正の相関または負の相関があることが知られている [19]．この事前知識は**ドメイン知識**と呼ばれる．例えば，ドメイン知識として，h 番目の特徴量 x_h がクラスラベルと正の相関があることが分かっている場合，対応する重み w_h は正になるべきである．逆に，負の相関があることが分かっている場合，w_h は負になるべきである．しかし，訓練用例題数が十分でなく，相関が弱い場合は，対応する重みは，既知の相関の符号と逆の符号に学習されてしまうことがしばしば起こる．その結果，学習された重みが正しい予測を妨げてしまう．

最近では，Tajima et al [33] が線形 SVM の重みベクトル $\boldsymbol{w}$ に**符号制約**を課し，生物学的配列分類への応用において，その有効性を示した．符号制約は，重みベクトル $\boldsymbol{w} = [w_1, \ldots, w_d]^\top$ のいくつかの要素に与える．このとき，クラスラベルと正の相関が既知の特徴の添え字集合 $\mathcal{I}_{\geq} \subseteq [d] (:= \{1, \ldots, d\})$ に対して，$h \in \mathcal{I}_{\geq}$ となる w_h に非負制約 $w_h \geq 0$ を課し，クラスラベルと負の相関が既知の特徴の添え字集合 $\mathcal{I}_{\leq} \subseteq ([d] \setminus \mathcal{I}_{\geq})$ に対して，$h \in \mathcal{I}_{\leq}$ となる w_h に非正制約 $w_h \leq 0$ を課す．このような符号制約を課すことで，特徴とクラスラベルの相関方向に関する事前知識との矛盾を明示的に回避することができる．

本研究では，符号制約の下で経験リスク最小化する問題に対する新しい最適化アルゴリズムを開発した．提案するアルゴリズムは，主目的関数値を直接最小化する代わりに双対問題を解く．すると，主目的関数値と双対目的関数値の差を計算することができるため，明確な停止条件が得られるという長所がある．主目的関数値と双対目的関数値の差が閾値以下の場合，目的誤差は閾値より小さいことが保証される．Tajima et al [33] は，損失関数を古典的な非平滑ヒンジ損失関数に限定したうえで，フランクウルフアルゴリズム [16] を採用した．本研究で提案するアルゴリズムは，**確率的双対座標上昇法（SDCA）** [30] に基づいており，**平滑**な損失関数で定式化された符号制約付き経験リスク最小化問題を解く．ここで，平滑とは勾配がリプシッツ連続であることを指す．提案するアルゴリズムの長所は，ϵ-**最適解**（目的関数の最小値との差が ϵ 以下の解）に到達するための反復回数を上限とする**線形収束** [27] を理論的に保証できるところである．

微生物学的水質データセットを用いた数値実験により，符号制約を課すことで予測性能が顕著に向上することを確認した．さらに，公開データセットを用いた数値実験により，提案するアルゴリズムが従来のアルゴリズムより高速に収束することも確認した．

本稿の構成は以下の通りである．関連する研究は次節で述べる．3 節では，符号制約付き学習問題を定式化し，その双対問題を示す．4 節で一般的な SDCA の枠組みを紹介し，5 節で符号制約付き学習問題に対する SDCA 反復処理の実装を示す．6 節で微生物学的水質予測への応用に関する実験結果を報告し，最後に本稿の結論を述べる．なお，本稿に記述した理論や実験結果の多くは，採択済みの査読付き論文 [32] に掲載したものと同一である．

2 関 連 研 究

符号制約は，回帰タスクや識別タスクなど多様な学習問題に広く用いられてきた．符号制約付き回帰は非負行列分解の重要な構成要素の 1 つである [8], [11], [21], [24], [34]．他には，符号制約付き最小二乗推定は，非負画像復元 [14], [22], [31], [35]，顔表現 [13], [17]，微生物分析 [6]，病原性水質分析 [19]，画像の超解像 [9]，スペクトル分析 [37]，断層イメージング [26]，音源定位 [25] などの多様な分野で応用されている．識別タスクに関しては，Tajima et al [33] が符号制約付き SVM を開発した．

二乗誤差損失に関しては，高速かつ安定した最適化アルゴリズムが利用可能である [5], [20], [23]．古典的なヒンジ損失に関しては，Tajima et al がフランクウルフ最適化アルゴリズムを開発した [33]．一方，符号制約を課さない一般的な経験リスク最小化であれば，安定した最適化アルゴリズムが多数存在する [7], [18], [28], [29], [36]．しかし，筆者らの知る限り，符号制約の下で一般的な損失関数を用いて最適化するアルゴリズムは，これまでにほとんど研究されていない．

3 主問題と双対問題

本研究の目的は，以下の制約付き経験リスク最小化問題に対する最適化アルゴリズムを開発することである:

$$\begin{aligned} \min \quad & P(\boldsymbol{w}) \quad \text{wrt } \boldsymbol{w} \in \mathbb{R}^d, \\ \text{subject to} \quad & \forall h \in \mathcal{I}_{\geq}, \quad w_h \geq 0, \\ & \forall h \in \mathcal{I}_{\leq}, \quad w_h \leq 0, \end{aligned} \tag{2}$$

添え字集合 $\mathcal{I}_{\geq}$ と $\mathcal{I}_{\leq}$ は以下を満たすように設定する．

$$\mathcal{I}_{\geq} \cup \mathcal{I}_{\leq} \subseteq [d] \quad \text{and} \quad \mathcal{I}_{\geq} \cap \mathcal{I}_{\leq} = \emptyset. \tag{3}$$

残りの添え字集合 $\mathcal{I}_0 := [d] \setminus (\mathcal{I}_{\geq} \cup \mathcal{I}_{\leq})$ は空でなくてもよい．通常，バイアス項の添え字は $\mathcal{I}_0$ に含まれる．1 節で前述したように，符号制約は事前知識に基づいて設定する．もし，h 番目の特徴量 x_h がクラスラベルと正の相関があるという事前知識があれば，w_h に非負制約を適用することができ，逆の関係が成り立つ場合は，w_h に非正制約を適用することができる．

あらかじめ $h \in \mathcal{I}_{\leq}$ の特徴量 $x_{h,i}$ の符号を逆転しておく (すなわち，$x_{h,i} \leftarrow -x_{h,i}$) ことで，非正制約を非負制約に変換することができる．そのため，以降のアルゴリズムの説明においては，$\mathcal{I}_{\leq} = \emptyset$ として，議論を進める．$x_{h,i}$ は $\boldsymbol{x}_i \in \mathbb{R}^d$ の h 番目の要素を表す．ただし，$\boldsymbol{x}_i \in \mathbb{R}^d$ は，訓練に用いられる i 番目の入力ベクトルである．訓練後，対応する重み w_h の符号を逆転することで，$w_h \leq 0$ を満たす重みを得ることができる．

実行可能領域を $\mathcal{S}$ とすると，(2) の制約問題は以下のように書き換えることができる：

$$\min \quad P(\boldsymbol{w}) \quad \text{wrt } \boldsymbol{w} \in \mathcal{S} \subseteq \mathbb{R}^d. \tag{4}$$

$\boldsymbol{\sigma} \in \{1, 0\}^d$ を次のようにおく：その h 番目の要素が $h \in \mathcal{I}_{\geq}$ のときは，$\sigma_h = 1$ とし，$h \in \mathcal{I}_0$ のときは，$\sigma_h = 0$ とする．すると，実行可能領域 $\mathcal{S}$ を次のように簡潔に表すことができる：

$$\mathcal{S} := \{\boldsymbol{w} \in \mathbb{R}^d \,|\, \forall h \in [d],\, \sigma_h w_h \geq 0\}. \tag{5}$$

開発した最適化アルゴリズムは，フェンシェル双対 [4] と呼ばれる双対関数を最大化する SDCA の枠組みに基づいている．双対関数を $D : \mathbb{R}^n \to \bar{\mathbb{R}}$ とおく．ただし，$\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ とする．このとき，双対関数 D は以下のように表される：

$$D(\boldsymbol{\alpha}) := -\frac{1}{2\lambda n^2} \left\| \boldsymbol{\pi}\left(\sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i \right) \right\|^2 - \frac{1}{n} \sum_{i=1}^{n} \phi_i^*(-\alpha_i), \tag{6}$$

ただし，$\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_n]^\top \in \mathbb{R}^n$ は双対変数ベクトル，$\phi_i^* : \mathbb{R} \to \bar{\mathbb{R}}$ は ϕ_i の**凸共役**，ベクトル値関数 $\boldsymbol{\pi} : \mathbb{R}^d \to \mathbb{R}^d$ は $\mathcal{S}$ への射影演算子であり，$\boldsymbol{\pi}(\boldsymbol{v}) := \boldsymbol{v} + \max(\boldsymbol{0}, -\boldsymbol{\sigma} \odot \boldsymbol{v})$ である．また，演算子 $\odot$ はアダマール積を表す．$\boldsymbol{\pi}(\boldsymbol{v})$ の h 番目の要素を $\pi_h(\boldsymbol{v})$ とすると，$h \in \mathcal{I}_{\geq}$ かつ $v_h \leq 0$ のとき，$\pi_h(\boldsymbol{v}) = 0$ が成り立つ．そうでなければ，$\pi_h(\boldsymbol{v}) = v_h$ となる．

双対関数 $D(\boldsymbol{\alpha})$ を最大にする解 $\boldsymbol{\alpha}_\star := [\alpha_1^\star, \ldots, \alpha_n^\star]^\top$ が見つかれば，次のように主問題 (4) の最適解を復元することがで

きる：

$$\boldsymbol{w}_\star = \frac{1}{\lambda n}\boldsymbol{\pi}\left[\sum_{i=1}^{n}\alpha_i^\star \boldsymbol{x}_i\right]. \tag{7}$$

損失関数 ϕ_i は，$1/\gamma$-**平滑**であると仮定する (すなわち，$\nabla\phi_i$ は $1/\gamma$-**リプシッツ連続**である)．例えば，**2 値交差エントロピー損失**は 0.25-平滑である．**2 次ヒンジ損失**

$$\phi_i(s) := \frac{1}{2}(\max\{0, 1-y_i s\})^2 \tag{8}$$

および，**平滑化ヒンジ損失**

$$\phi_i(s) := \begin{cases} \frac{1-2y_i s}{2} & \text{for } s < 0, \\ \frac{1}{2}(\max\{0, 1-y_i s\})^2 & \text{for } s \geq 0 \end{cases} \tag{9}$$

は両方とも，1-平滑である．$(1/\gamma)$-平滑な凸関数の凸共役は，γ-強凸関数 [15] である．これは，$\forall \eta \in [0,1]$ に対して，以下を満たす：

$$\begin{aligned} &\eta\phi_i^*(-u) + (1-\eta)\phi_i^*(-\alpha) \\ &\geq \phi_i^*(-\eta u - (1-\eta)\alpha) + \frac{\gamma}{2}(u-\alpha)^2(1-\eta)\eta. \end{aligned} \tag{10}$$

SDCA の枠組みでは，上記の不等式を次のように式変形したものを使用する：

$$\begin{aligned} &\phi_i^*(-\alpha) - \phi_i^*(-\alpha - \eta(u-\alpha)) \\ &\geq (\phi_i^*(-\alpha) - \phi_i^*(-u))\eta + \frac{\gamma}{2}(u-\alpha)^2(1-\eta)\eta. \end{aligned} \tag{11}$$

4 SDCA の枠組み

SDCA は，各反復ごとに双対変数ベクトル $\boldsymbol{\alpha}$ の中から無作為に要素を一つ選択し，更新する．$\boldsymbol{\alpha}$ の中から選択された要素の添え字を i とおく．無作為に選択された要素の 1 反復前の値との差を $\Delta\alpha$ とすると，$\alpha_i^{(t)} := \alpha_i^{(t-1)} + \Delta\alpha$ と表せる．$i' \in [n]\backslash\{i\}$ のとき，双対変数の値は変化しない (つまり，$\alpha_{i'}^{(t)} := \alpha_{i'}^{(t-1)}$)．

$$\begin{aligned} \bar{\boldsymbol{w}}^{(t)} &:= \frac{1}{\lambda n}\sum_{i'=1}^{n}\alpha_{i'}^{(t)}\boldsymbol{x}_{i'}. \quad \text{and} \\ \boldsymbol{w}^{(t)} &:= \boldsymbol{\pi}\left[\bar{\boldsymbol{w}}^{(t)}\right]. \end{aligned} \tag{12}$$

とおく．各反復で $\Delta\alpha$ が決まれば，このベクトル $\bar{\boldsymbol{w}}^{(t)}$ を次のように $O(d)$ のコストで更新できる：

$$\begin{aligned} \bar{\boldsymbol{w}}^{(t)} &= \frac{\alpha_i^{(t-1)} + \Delta\alpha}{\lambda n}\boldsymbol{x}_i + \frac{1}{\lambda n}\sum_{i'\in[n]\backslash\{i\}}^{n}\alpha_{i'}^{(t-1)}\boldsymbol{x}_{i'} \\ &= \bar{\boldsymbol{w}}^{(t-1)} + \frac{\Delta\alpha}{\lambda n}\boldsymbol{x}_i. \end{aligned} \tag{13}$$

表記を簡単にするために，ここでは上付き文字 $(t-1)$ を省略して次のように書く：

$$\boldsymbol{w} := \boldsymbol{w}^{(t-1)}, \quad \boldsymbol{v}_0 := \bar{\boldsymbol{w}}^{(t-1)}, \quad \text{and} \quad \boldsymbol{\alpha} := \boldsymbol{\alpha}^{(t-1)}. \tag{14}$$

$\Delta\alpha$ は，次の関数を最大化する値を選ぶのが理想的である：

Algorithm 1: SDCA algorithm for maximizing $D(\boldsymbol{\alpha})$.

```
begin
    Choose α^(0) s.t. α^(0) ∈ dom(−D);
    for t := 1 to T do
        Pick i randomly from {1, ..., n};
        η_t ∈ argmax_{η∈[0,1]} J_t^1(η);
        α^(t) := α^(t−1) − (∇φ(⟨w^(t−1), x_i⟩) + α_i^(t−1)) η_t e_i;
        Compute w̄^(t) and w^(t);
    end
end
```

$$\begin{aligned} J_t^0(\Delta\alpha) &:= D(\boldsymbol{\alpha} + \Delta\alpha\boldsymbol{e}_i) - D(\boldsymbol{\alpha}) \\ &= \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0\right]\right\|^2 - \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0 + \frac{\Delta\alpha}{\lambda n}\boldsymbol{x}_i\right]\right\|^2 \\ &\quad + \frac{1}{n}\left(\phi_i^*(-\alpha_i) - \phi_i^*(-\alpha_i - \Delta\alpha)\right) \end{aligned} \tag{15}$$

ただし，$\boldsymbol{e}_i$ は i 番目の要素が 1 の単位ベクトルである．$\Delta\alpha$ は，ϕ_i^* の引数のままなので，最適な $\Delta\alpha$ を見つけるのは一般的に難しい．閉形式の更新則を得るために，$\Delta\alpha$ の範囲を以下のように制限する．$\alpha_i + \nabla\phi(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle) \neq 0$ のとき，

$$\eta := -\frac{\Delta\alpha}{\alpha_i + \nabla\phi(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle)} \in [0,1] \tag{16}$$

それ以外のとき，$\Delta\alpha := 0$ とする．以下では，$u := -\nabla\phi(\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle) \neq \alpha_i$ が非自明な場合のみを議論する．ただし，α_i は $\boldsymbol{\alpha}^{(t-1)}$ の i 番目の要素である．このとき，$q := u - \alpha_i$ とすると，$\Delta\alpha = q\eta$ となる．

$$\boldsymbol{v}_q := \frac{q}{\lambda n}\boldsymbol{x}_i \tag{17}$$

とする．不等式 (11) を適用すると，$J_t^0(q\eta)$ は次のように下から抑えられる：

$$\begin{aligned} J_t^0(q\eta) &= \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0\right]\right\|^2 - \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0 + \eta\boldsymbol{v}_q\right]\right\|^2 \\ &\quad + \frac{1}{n}\left(\phi_i^*(-\alpha_i) - \phi_i^*(-\alpha_i - q\eta)\right) \\ &\geq \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0\right]\right\|^2 - \frac{\lambda}{2}\left\|\boldsymbol{\pi}\left[\boldsymbol{v}_0 + \eta\boldsymbol{v}_q\right]\right\|^2 \\ &\quad + a_{\text{offs}}\eta^2 + b_{\text{offs}}\eta =: J_t^1(\eta). \end{aligned} \tag{18}$$

ただし，

$$\begin{aligned} a_{\text{offs}} &:= -\frac{q^2\gamma}{2n}, \quad \text{and} \\ b_{\text{offs}} &:= \frac{\phi_i^*(-\alpha_i) - \phi_i^*(-u) + 0.5q^2\gamma}{n}. \end{aligned} \tag{19}$$

下限 J_t^1 では，$\Delta\alpha$ が ϕ_i^* の引数に含まれていないので，J_t^1 は J_t^0 よりも扱いやすい．以上の議論により，符号制約の下で学習する SDCA は，Algorithm 1 のようにまとめることができる．J_t^0 の代わりに J_t^1 を最大化しても，SDCA の線形収束は保証される：

定理 1. $R := \max_{i\in[n]}\|\boldsymbol{x}_i\|$, $h_P^{(t)} := P(\boldsymbol{w}^{(t)}) - P(\boldsymbol{w}_\star)$, $h_D^{(t)} := D(\boldsymbol{\alpha}_\star) - D(\boldsymbol{\alpha}^{(t)})$ とする．任意の $\epsilon_P > 0$ に対して，*Algorithm 1* が以下の反復数に達したとき，$\mathbb{E}\left[h_P^{(t)}\right] \leq \epsilon_P$ が成り立つ：

$$t \geq \frac{\lambda n\gamma + R^2}{\lambda\gamma} \log\left(\frac{h_D^{(0)}}{\epsilon_P}\frac{\lambda n\gamma + R^2}{\lambda\gamma}\right). \tag{20}$$

この定理の証明は，付録 1 を参照されたい．従来の符号制約のない学習を行う SDCA アルゴリズムが要する反復数の下限は，基本的に定理 1 で与えた下限と変わらない．[30] で示された従来 SDCA の反復数の下限では，$h_D^{(0)}$ を $\boldsymbol{\alpha}^{(0)} = \boldsymbol{0}, \phi_i(0) \leq 1$ と仮定して，1 に置き換えている．

次節では，Algorithm 1 の第 5 行の実装方法について説明する．

5 SDCA 反復処理の実装

本節では，$J_t^1(\eta)$ の最大値を求めるアルゴリズムを示す．本研究の主要な発見は，J_t^1 が凹な区分 2 次関数で表されるという性質である．この発見により，更新則の効率的なアルゴリズムを開発することができた．以下では，まず $J_t^1(\eta)$ が区分 2 次関数の形式で表すことができることを示し (5.1 節)，そのうえで $J_t^1(\eta)$ の最大値を求める 2 種類のアルゴリズムについて説明する (5.2 節と 5.3 節).

5.1 区分 2 次関数

$v_{h,0}$ と $v_{h,q}$ を，それぞれ $\boldsymbol{v}_0$ と $\boldsymbol{v}_q$ の h 番目の要素とする．$\mathcal{I}_0 := \{h \in [d] \,|\, \sigma_h = 0\}$ とする．$0 = \theta_1 < \cdots < \theta_{d_t+1} = 1$ となるような $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_{d_t}, \theta_{d_t+1}]^\top$ を定義する．ただし，$\theta_1, \ldots, \theta_{d_t}, \theta_{d_t+1}$ は，以下で定義される $\mathrm{Card}[\Theta] = d_t + 1$ の離散集合 $\Theta \subset \mathbb{R}$ の元である：

$$\begin{aligned}\Theta :=& \{0,1\} \cup \{\theta \in (0,1) \\ & \,|\, \exists h \in \mathcal{I}_{\geq} \text{ s.t. } v_{h,0} = -\theta v_{h,q} \neq 0\}.\end{aligned} \tag{21}$$

$k \in \{2, \ldots, d_t\}$ における要素 θ_k は，ある $h \in \mathcal{I}_{\geq}$ に対して，アファイン関数 $\eta \mapsto v_{h,0} + \eta v_{h,q}$ が 0 の水平線を横切る位置である．図 1 は，$d = 3, \mathcal{I}_{\geq} = \{1,2,3\}$, $\boldsymbol{v}_0 = [0.5, 0.75, -0.5]^\top$, $\boldsymbol{v}_q = [0.5, -1, 1]^\top$ のときのアファイン関数の数値例である．Θ の定義より，$d_t = 3$, $\theta_1 = 0$, $\theta_2 = 0.5$, $\theta_3 = 0.75$, $\theta_4 = 1$ となる．アファイン関数 $\eta \mapsto v_{3,0} + \eta v_{3,q}$ は，$\eta = \theta_2$ で 0 の水平線を横切り，$\eta \mapsto v_{2,0} + \eta v_{2,q}$ は，$\eta = \theta_3$ で 0 の水平線を横切る．

$k \in [d_t]$ に対して，次のような添え字集合を定義する：

$$\begin{aligned}\mathcal{H}_k := \mathcal{I}_0 \cup \{h \in \mathcal{I}_{\geq} \,| \\ 2v_{h,0} + (\theta_k + \theta_{k+1})v_{h,q} > 0\}.\end{aligned} \tag{22}$$

図 1 の例の場合，定義 (22) より，添え字集合は
$\mathcal{H}_1 = \{1,2\}$, $\mathcal{H}_2 = \{1,2,3\}$, $\mathcal{H}_3 = \{1,3\}$ となる．

$h \in \mathcal{H}_k \cap \mathcal{I}_{\geq}$ に対して，アファイン関数 $\eta \mapsto v_{h,0} + \eta v_{h,q}$ は 0 の水平線を超える．すなわち，

$$\forall \eta \in (\theta_k, \theta_{k+1}), \quad \forall h \in \mathcal{H}_k, \quad v_{h,0} + \eta v_{h,q} > 0, \tag{23}$$

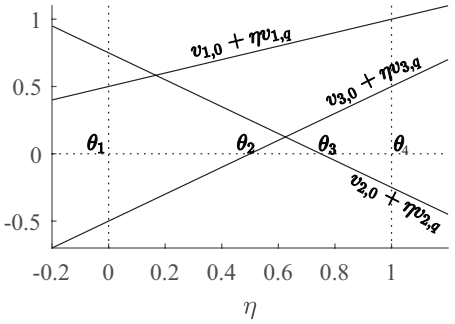


図 1: アファイン関数 $\eta \mapsto v_{h,0} + \eta v_{h,q}$ の数値例.

が成り立ち，$\forall \eta \in (\theta_k, \theta_{k+1})$ に対して，

$$[\boldsymbol{\pi}(\boldsymbol{v}_0 + \eta\boldsymbol{v}_q)]_h = \begin{cases} v_{h,0} + \eta v_{h,q} & \text{for } h \in \mathcal{H}_k, \\ 0 & \text{for } h \notin \mathcal{H}_k \end{cases} \tag{24}$$

が導ける．ただし，$[\boldsymbol{\pi}(\boldsymbol{v}_0 + \eta\boldsymbol{v}_q)]_h$ は，d 次元ベクトル $\boldsymbol{\pi}(\boldsymbol{v}_0 + \eta\boldsymbol{v}_q)$ の h 番目の要素である．ベクトル $\boldsymbol{\theta}$ と $k \in [d_t]$ に関する集合 $\mathcal{H}_k$ より，関数 J_t^1 は以下の区分 2 次関数で表すことができる：

$$\forall \eta \in [\theta_k, \theta_{k+1}], \quad J_t^1(\eta) = a_k\eta^2 + b_k\eta \tag{25}$$

ただし，a_k と b_k は以下のように与えられる．

$$\begin{aligned} a_k &= a_{\text{offs}} - \frac{\lambda}{2}\sum_{h\in\mathcal{H}_k} v_{h,q}^2, \quad \text{and} \\ b_k &= b_{\text{offs}} - \lambda\sum_{h\in\mathcal{H}_k} v_{h,q}v_{h,0}. \end{aligned} \tag{26}$$

5.2 $O(d^2)$ の実装

J_t^1 の凹性と微分可能性から，$\eta_\star$ で表される $J_t^1(\eta)$ の最大値の一つは，次のように求められる：

- $\nabla J_t^1(0) = b_1 \leq 0$ のとき，$\eta_\star = 0$;
- $\nabla J_t^1(1) = 2a_{d_t+1} + b_{d_t+1} \geq 0$ のとき，$\eta_\star = 1$;
- それ以外のとき，区間 $[\theta_{k_\star}, \theta_{k_\star+1}]$ に最大値 $\eta_\star = -0.5b_{k_\star}/a_{k_\star}$ を含むような $k_\star \in [d_t]$ が存在する．

第 3 の場合（つまり，$2a_{d_t+1} + b_{d_t+1} < 0 < b_1$) の区間番号 $k_\star$ は，全ての区間を調べることで求めることができる．なぜなら，J_t^1 の微分可能性より，$\nabla J_t^1(\theta_{k_\star}) \geq 0 \geq \nabla J_t^1(\theta_{k_\star+1})$ が成り立つからである．この議論と前述の発見を組み合わせると，SDCA の各反復は以下のように実装することができる：

1. $i \in [n]$ を無作為に選ぶ; $O(1)$.
2. $\boldsymbol{v}_0$ および $\boldsymbol{v}_q$ を計算する; $O(d)$.
3. Θ を計算する; $O(d)$.
4. Θ の要素を昇順に整列する; $O(d\log d)$.
5. $k \in [d_t]$ に対して，$\mathcal{H}_k$ を計算する; $O(d^2)$.
6. $k \in [d_t]$ に対して，係数 (a_k, b_k) を計算する; $O(d^2)$.
7. 最大値 $\eta_\star$ を求める; $O(d)$.
8. $\Delta\alpha = q\eta_\star$; $O(1)$.
9. $\bar{\boldsymbol{w}}^{(t)} = \bar{\boldsymbol{w}}^{(t-1)} + \frac{\Delta\alpha}{\lambda n}\boldsymbol{x}_i$ を計算する; $O(d)$.

この実装より，各反復は $O(d^2)$ の計算コストで実行できる．この実装で最も重いステップは，添え字集合 $\mathcal{H}_k$ を計算するステップ (ステップ 5) と係数 a_k, b_k を計算するステップ (ステップ 6) であり，どちらも $O(d^2)$ のコストがかかる．これらの時間複雑度は次のように導かれる．区分 2 次関数の個数は $d_t \leq \mathrm{Card}(\mathcal{I}_{\geq}) + 1 \leq d + 1 = O(d)$ で抑えられる．このとき，$k \in [d_t]$ に対して，各 $\mathcal{H}_k$ は，$\mathcal{H}_k \subseteq [d]$ より，$O(d)$ 時間で計算できる．したがって，ステップ 5 の時間複雑度は $O(d^2)$ であることが証明される．$\mathrm{Card}(\mathcal{H}_k) = O(d)$ より，$2d_t (= O(d))$ 個の係数 $a_1, b_1, \ldots, a_{d_t}, b_{d_t}$ を (26) を用いて計算すると，全部で $O(d^2)$ のコストがかかる．

一方，筆者らは，時間複雑度を線形コストに削減できるアルゴリズムを発見した．その線形時間アルゴリズムを以下に示す．

5.3 $O(d \log d)$ の実装

ここでは，$J_t^1(\eta)$ を $\eta \in [0, 1]$ に関して厳密に最大化するための線形時間のアルゴリズムを示す．5.2 節で与えられたアルゴリズムの理論的な時間複雑度は $O(d^2)$ であるが，以下に示すアルゴリズムの時間複雑度は，$O(d \log d)$ に減少する：

$$\mathcal{H}_{k,\mathrm{in}} := \mathcal{H}_k \setminus \mathcal{H}_{k-1} \quad \text{and} \quad \mathcal{H}_{k,\mathrm{out}} := \mathcal{H}_{k-1} \setminus \mathcal{H}_k \tag{27}$$

を定義すると，区分 2 次関数の係数は，$\forall k \geq 2$ に対して，

$$\begin{aligned} a_k &:= a_{k-1} - \frac{\lambda}{2} \sum_{h \in \mathcal{H}_{k,\mathrm{out}}} v_{h,q}^2 + \frac{\lambda}{2} \sum_{h \in \mathcal{H}_{k,\mathrm{in}}} v_{h,q}^2, \\ b_k &:= b_{k-1} - \lambda \sum_{h \in \mathcal{H}_{k,\mathrm{out}}} v_{h,q} v_{h,0} + \lambda \sum_{h \in \mathcal{H}_{k,\mathrm{in}}} v_{h,q} v_{h,0}, \end{aligned} \tag{28}$$

として再帰的に表現できる．(28) を使って，a_k と b_k を計算するには，集合 $\mathcal{H}_{k,\mathrm{in}}$ と $\mathcal{H}_{k,\mathrm{out}}$ だけでなく，$\mathcal{H}_1$ もあらかじめ必要になる．集合 $\mathcal{H}_1$ は，以下の 3 つの条件の内 1 つが満たされているかどうかを調べることで，$O(d)$ 以内で求めることができる：

$$\begin{aligned} &\text{(i) } h \in \mathcal{I}_0; \quad \text{(ii) } v_{h,0} > 0; \\ &\text{(iii) } v_{h,0} = 0 \quad \text{かつ} \quad v_{h,q} > 0. \end{aligned} \tag{29}$$

すなわち，$h \in [d]$ が上記の 3 つの条件の内 1 つを満たすなら，$h \in \mathcal{H}_1$ となり，満たさないなら，$h \notin \mathcal{H}_1$ となる．ここで，$\mathcal{H}_{k,\mathrm{in}}$ と $\mathcal{H}_{k,\mathrm{out}}$ の計算方法について述べる．そのために，まず，$h \in \mathcal{I}_{\geq}$ に対して，$\tilde{\theta}_h^\circ := -\frac{v_{h,0}}{v_{h,q}}$ を計算する．次に，$(d_t - 1)$ 個の端点 $\theta_2, \ldots, \theta_{d_t}$ を，$\tilde{\theta}_h^\circ$ の値を昇順に並べ替え，開区間 $(0, 1)$ の外の値と重複する値を取り除くことで求める．このとき，$\boldsymbol{\theta}$ を計算する過程で，$k \in \{2, \ldots, d_t\}$ に対して，集合 $\mathcal{H}_{k,\mathrm{in}}$ および $\mathcal{H}_{k,\mathrm{out}}$ は，次のように同時に計算できる：

$$\begin{aligned} \mathcal{H}_{k,\mathrm{in}} &= \left\{ h \in \mathcal{I}_{\geq} \,\middle|\, \theta_k = \tilde{\theta}_h^\circ,\, v_{h,q} > 0 \right\}, \quad \text{and} \\ \mathcal{H}_{k,\mathrm{out}} &= \left\{ h \in \mathcal{I}_{\geq} \,\middle|\, \theta_k = \tilde{\theta}_h^\circ,\, v_{h,q} < 0 \right\}. \end{aligned} \tag{30}$$

これらの議論から，5.2 節で与えられた $O(d^2)$ の実装は以下のように修正することができる．

1. $i \in [n]$ を無作為に選ぶ; $O(1)$.
2. $\boldsymbol{v}_0$ および $\boldsymbol{v}_q$ を計算する; $O(d)$.
3. $\mathcal{H}_1$ を計算する; $O(d)$.
4. $h \in \mathcal{I}_{\geq}$ に対して，$\tilde{\theta}_h^\circ$ を計算する; $O(d)$.
5. $k \in [d_t]$ に対して，$(\mathcal{H}_{k,\mathrm{in}}, \mathcal{H}_{k,\mathrm{out}}), \theta_k$ を計算する; $O(d \log d)$.
6. $k \in [d_t]$ に対して，係数 (a_k, b_k) を計算する; $O(d)$.
7. 最大値 $\eta_\star$ を求める; $O(d)$.
8. $\Delta\alpha = q\eta_\star$; $O(1)$.
9. $\bar{\boldsymbol{w}}^{(t)} = \bar{\boldsymbol{w}}^{(t-1)} + \frac{\Delta\alpha}{\lambda n} \boldsymbol{x}_i$ を計算する; $O(d)$.

ステップ 5 は，並び替える値の数が $\mathrm{Card}(\mathcal{I}_{\geq}) = O(d)$ であるため，$O(d \log d)$ のコストで $\tilde{\theta}_h^\circ$ を並べ替えることができる．ステップ 6 の計算量は，$O(d)$ である．なぜならば，関係式

$$\bigcup_{k=2}^{d_t} \mathcal{H}_{k,\mathrm{in}} \subseteq \mathcal{I}_+ \subseteq [d] \quad \text{and} \quad \bigcup_{k=2}^{d_t} \mathcal{H}_{k,\mathrm{out}} \subseteq \mathcal{I}_+ \subseteq [d] \tag{31}$$

より，$k \in \{2, \ldots, d_t\}$ に対して，(28) の項の総数の上限は，$4d$ であるからである．したがって，SDCA の各反復は，$O(d \log d)$ の計算コストで済むことが示された．

6 数 値 実 験

6.1 パターン認識性能

微生物学的水質予測において，符号制約がパターン認識性能に及ぼす効果を確認するために数値実験を行った．使用した微生物学的水質データセットは，Sapporo, NY top, NY bottom, Indian の 4 つである．Sapporo データセットは, [19] で提供されたもので，$n_{\mathrm{tot}} := 177$ 個の例題からなり，各例題は，目的変数 E.coli と 9 個の特徴変数 WT, pH, EC, DO, SS, BOD, TN, TP, FR から構成されている．このタスクは，大腸菌が 500 MPN を超えるかどうかを予測することである．このデータセットは，88 個の正例と 89 個の負例からなる．NY top, NY bottom, Indian は，kaggle.com で提供されていたものである．この 3 つのデータセットには，それぞれ，534 個，523 個，896 個の例題が含まれている．各例題は目的変数 FC と 5 種類の物理化学的水質データからなる．これらに対し，FC が中央値を超えれば陽性，下回れば陰性とする 2 クラス分類問題を構成した．

符号制約は，表 1 の通りである．3 種類の損失関数，平滑化ヒンジ損失，2 次ヒンジ損失，2 値交差エントロピー損失を試した．各損失関数に対して，従来の学習と符号制約学習を行った．その結果，6 種類の予測結果が得られた．性能基準には，正解率（すなわち，評価用例題数に対する真陽性と真陰性の合計）を用いた．訓練用例題数 n は，5 から 100 まで変化させた．ただし，n 個の例題は，各データセットから無作為に抽出した．この n 個の例題を 6 種類の学習手法に与え，6 個の予測器を得た．残りの $(n_{\mathrm{tot}} - n)$ 個の例題は，評価に用いた．この手順を 200 回繰り返した．

図 2，図 3，図 4，図 5 は，それぞれ 4 つの水質データセットから得られた 200 個の正解率の平均をプロットしたものであ

表 1: 4 つのデータセットの特徴と符号制約．チェックがついている特徴は，対応する各データセットに含まれている．符号制約は，一番右の列に表す．≥ 0 と ≤ 0 はそれぞれ非負と非正の制約を意味する．

特徴	Sapporo	NY top	NY bottom	Indian	符号制約
WT	✓				≥ 0
pH	✓	✓	✓		≤ 0
EC	✓	✓	✓	✓	
DO	✓	✓	✓	✓	≤ 0
SS	✓				≥ 0
BOD	✓	✓	✓	✓	
TN	✓				≥ 0
TP	✓				≥ 0
FR	✓				≤ 0
TC		✓	✓	✓	≥ 0
Nitro				✓	≥ 0

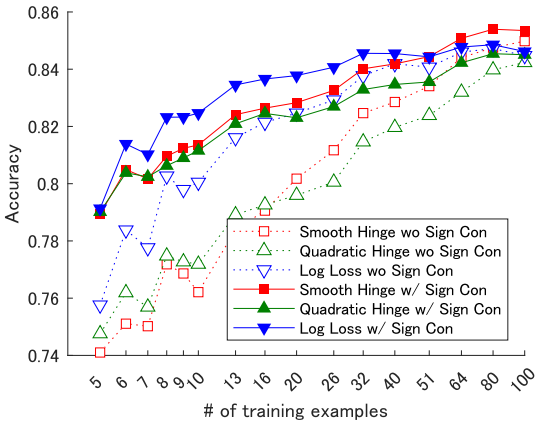


図 2: Sapporo データセットでの予測性能．実線の曲線は符号制約の下で最適化したときの予測性能を示し，破線の曲線は符号制約を用いない場合の予測性能を示す．正方形，上向きの三角形，下向きの三角形の印は，それぞれ，平滑化ヒンジ損失，2 次ヒンジ損失，2 値交差エントロピー損失を用いた場合を表す．

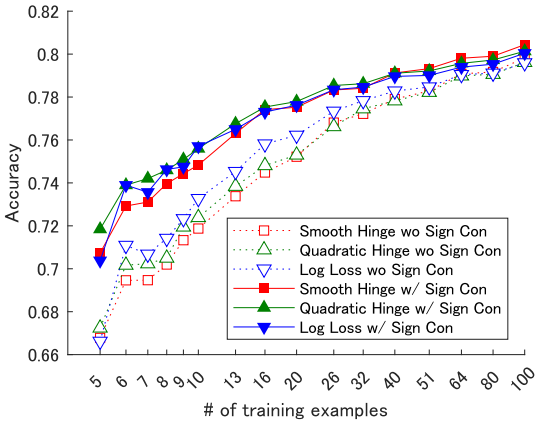


図 3: NY top データセットでの予測性能．

る．横軸は訓練用例題数 n を表す．4 つのデータセットと 3 個の損失関数の全てにおいて，符号制約は予測性能を向上させた．

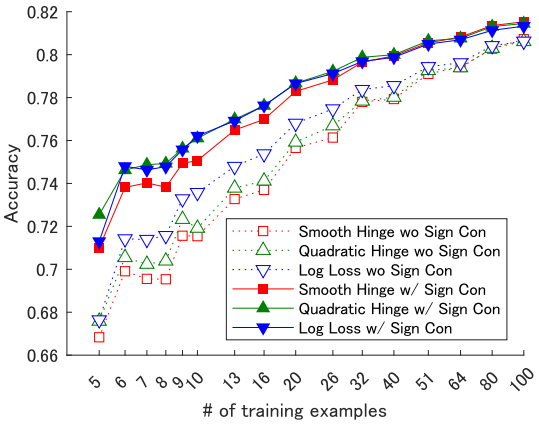


図 4: NY bottom データセットでの予測性能．

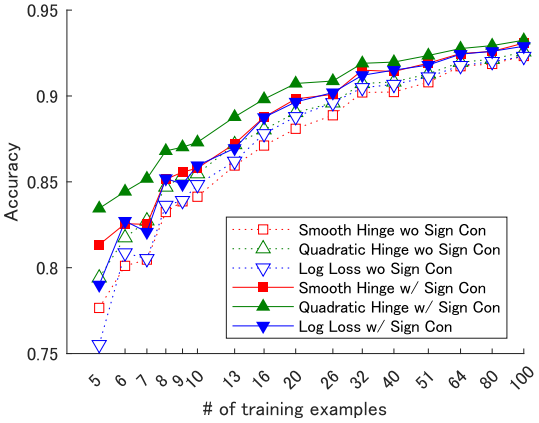


図 5: Indian データセットでの予測性能．

特に，訓練用例題が少ない程，符号制約の効果は顕著であった．符号制約は，ドメイン固有の事前知識の一種であり，学習が事前知識に矛盾することを明示的に防ぐことができる．符号制約がないと，訓練用例題数が少ない場合，重み w_h の符号が特徴とクラスラベル間の相関の真の符号から反転してしまうという現象がしばしば起こる．重みの符号反転を回避する符号制約が効果的に働き，汎化性能を向上させた．

6.2 収束性能

提案する SDCA アルゴリズムの収束の速さを評価するために，3 つの異なるデータセット (`Magic04`, `Segment`, `Waveform`) で収束実験を行った．この実験では，提案アルゴリズムと射影確率的勾配降下法 (PSGD) の収束性能を比較した．PSGD は，$\boldsymbol{w}^{(t+1)} := \boldsymbol{\pi}(\boldsymbol{w}^{(t)} - \eta \nabla P(\boldsymbol{w}^{(t)}))$ という式で解を更新する．ただし，η はステップサイズである．

データセットの特徴は，以下の通りである．`Magic04` は，19,024 個の例題と 10 個の特徴を持ち，`Segment` は，2,310 個の例題と 19 個の特徴を持ち，`Waveform` は，5,000 個の例題と 21 個の特徴を持つ．PSGD のステップサイズは，$\eta = 10^0$, $\eta = 10^{-1}$, $\eta = 10^{-2}$, $\eta = 10^{-3}$ とした．正則化定数は，$\lambda = 1/n$ とした．最適化には，2 値交差エントロピー損失 $\phi_i(s) = \log(1 + \exp(-y_i s))$ を用いた．無作為に選択した特徴の半分に非負制約を課し，残りの半分に非正制約を課し

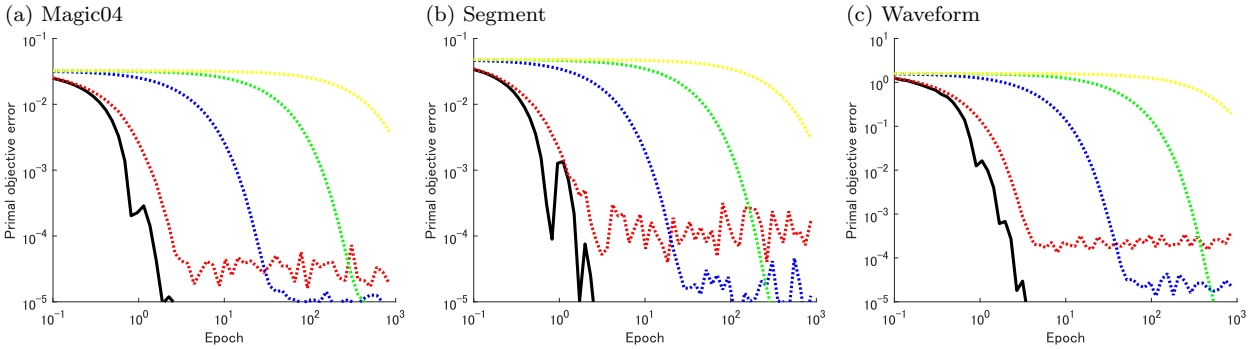


図 6: 3 つのデータセットにおける収束のふるまい．黒の実線が提案アルゴリズム，破線が PSGD の収束のふるまいを表す．赤，青，緑，黄の破線はそれぞれ，ステップサイズを 10^0, 10^{-1}, 10^{-2}, 10^{-3} としたときの PSGD の収束のふるまいを表す．

た．このとき，各反復で目的誤差を観察した．しかし，$P(\boldsymbol{w}_\star)$ は未知のため，真の目的誤差 $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star)$ は直接評価できない．そこで，この誤差を近似するために，提案アルゴリズムを実行し，$T' := 1,000$ 回目の反復における双対目的関数値 $D(\boldsymbol{\alpha}^{(T')})$ を用いて評価した．$D(\boldsymbol{\alpha}^{(T')})$ は，真の双対目的関数値の最大値より若干小さくなる可能性があるが，十分最大値に近い値とみなせる．そこで，$P(\boldsymbol{w}) - D(\boldsymbol{\alpha}^{(T')})$ を目的誤差の評価値とした．いずれのデータセットでも，提案アルゴリズムを用いた場合，$P(\boldsymbol{w}^{(T')}) - D(\boldsymbol{\alpha}^{(T')})$ が 10^{-7} 以下となり，目的誤差は 10^{-7} に収まることが確認できた．

図 6 では，3 つのデータセット `Magic04`, `Segment`, `Waveform` でのエポック数に対する目的誤差の変化を，それぞれ，(a), (b), (c) の 3 個のパネルに示す．提案アルゴリズムは，各データセットに対して，約 1.9, 2.7, 3.7 エポック後に，10^{-5} の誤差に達した．一方，PSGD は，どのデータセットでも，ステップサイズが $\eta = 10^0$ と $\eta = 10^{-3}$ のとき，10^{-5} に達しなかった．これは，ステップサイズが大きすぎたり，小さすぎたりしたためである．$\eta = 10^{-2}$ のステップサイズでは，最終的に 10^{-5} に達したが，かなりの数のエポックを経た後だった．すなわち，3 つのデータセットでそれぞれ，429, 306, 536 エポックを要して，目的誤差が 10^{-5} に達した．これらのエポック数は，提案アルゴリズムの 100 倍以上である．また，$\eta = 10^{-1}$ のステップサイズでは，PSGD は，`Magic04` と `Segment` の各データセットで，79 エポック, 109 エポックで 10^{-5} に達した．しかし，データセット `Waveform` では，PSGD は 10^3 エポック経っても，10^{-5} に達しなかった．このように，提案アルゴリズムが PSGD よりも有意に速く最適解に収束した．

筆者らの実装では，3 つのデータセットにおける提案アルゴリズムの 1 エポック当たりの平均実行時間はそれぞれ，0.488 秒，0.0556 秒，0.146 秒であった．一方，PSGD の 1 エポック当たりの平均実行時間はそれぞれ，0.141 秒，0.0231 秒，0.0428 秒であった．提案アルゴリズムの 1 エポック当たりの実行時間は PSGD を上回るが，高精度な解を得るために必要なエポック数を大幅に減少させることができる．この効率性により，PSGD に比べて，高精度な解に達するまでにかかる全体的な計算時間は短縮される．

7 ま と め

本稿では，符号制約下での経験リスク最小化のための新しいアルゴリズムを提案した．本研究で開発したアルゴリズムは，任意の平滑で凸な損失関数を適用できる．この最適化アルゴリズムは，SDCA の枠組みに基づいており，最適解への線形収束が保証されている．パターン認識における符号制約の効果を，微生物学的水質予測の実データを用いた計算機実験により実証した．

筆者らは，未発見の符号制約に適した事前知識が，多くの応用分野で眠っていると考える．今後の課題として，符号制約に適した他の応用を模索していくことが考えられる．

謝　　辞

本研究の一部は，独立行政法人環境保全機構の環境研究総合推進費 (5-2006) により，実施された．また，JSPS 科研費 22K04372 の助成を受けた．

文　　献

[1] World Health Organization. Guidelines for drinking-water quality: Fourth edition incorporating the first and second addenda. https://www.who.int/publications/i/item/9789240045064, 2022.

[2] Helen Bridle. Waterborne Pathogens: Detection Methods and Applications. *Academic Press*, 2020.

[3] T. J. Wade and N. Pai and J. N. Eisenberg and J. r. Colford JM. Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environ Health Perspect*, 111(8):1102-9, Jun 2003. doi: 10.1289/ehp.6241.

[4] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[5] M. Bierlaire, Ph.L. Toint, and D. Tuyttens. On iterative algorithms for linear least squares problems with bound constraints. *Linear Algebra and its Applications*, 143:111–143, jan 1991. doi: 10.1016/0024-3795(91)90009-l.

[6] Y. Cai, H. Gu, and T. Kenney. Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. *Microbiome*, 5(1):110, Aug 2017.

[7] Aaron Defazio, Francis Bach, and Simon Lacoste-julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[8] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD06*. ACM Press, 2006. doi:10.1145/1150402.1150420.

[9] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and G. Li. Hyperspectral image super-resolution via non-negative structured sparse representation. *IEEE Trans Image Process*, 25(5):2337–52, May 2016.

[10] Kelwin Fernandes and Jaime S. Cardoso. Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, nov 2017. doi:10.1007/s00521-017-3281-4.

[11] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, sep 2011. doi:10.1162/neco_a_00168.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[13] R. He, W. S. Zheng, B. G. Hu, and X. W. Kong. Two-stage nonnegative sparse representation for large-scale face recognition. *IEEE Trans Neural Netw Learn Syst*, 24(1):35–46, Jan 2013.

[14] Simon Henrot, Said Moussaoui, Charles Soussen, and David Brie. Edge-preserving nonnegative hyperspectral image restoration. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, may 2013. doi: 10.1109/icassp.2013.6637926.

[15] Jean-Baptiste Hiriart-Urruty. *Fundamentals of Convex Analysis*. Springer, 2001.

[16] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[17] Yangfeng Ji, Tong Lin, and Hongbin Zha. Mahalanobis distance based non-negative sparse representation for face recognition. In *2009 International Conference on Machine Learning and Applications*. IEEE, dec 2009. doi:10.1109/icmla.2009.50.

[18] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 315–323, 2013.

[19] T. Kato, A. Kobayashi, W. Oishi, S. S. Kadoya, S. Okabe, N. Ohta, M. Amarasiri, and D. Sano. Sign-constrained linear regression for prediction of microbe concentration based on water quality datasets. *J Water Health*, 17(3):404–415, Jun 2019.

[20] Dongmin Kim, Suvrit Sra, and Inderjit S. Dhillon. Tackling box-constrained optimization via a new projected quasi-newton approach. *SIAM Journal on Scientific Computing*, 32(6):3548–3563, jan 2010. doi:10.1137/08073812x.

[21] Keigo Kimura, Mineichi Kudo, and Yuzuru Tanaka. A column-wise update algorithm for nonnegative matrix factorization in bregman divergence with an orthogonal constraint. *Machine Learning*, 103(2):285–306, mar 2016. doi:10.1007/s10994-016-5553-0.

[22] Germana Landi and Elena Loli Piccolomini. NPTool: a Matlab software for nonnegative image restoration with Newton projection methods. *Numerical Algorithms*, 62(3):487–504, jun 2012. doi: 10.1007/s11075-012-9602-x.

[23] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, jan 1995. doi:10.1137/1.9781611971217.

[24] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[25] Yuanqing Lin, D.D. Lee, and L.K. Saul. Nonnegative deconvolution for time of arrival estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004. doi:10.1109/icassp.2004.1326273.

[26] Jun Ma. Algorithms for non-negatively constrained maximum penalized likelihood reconstruction in tomographic imaging. *Algorithms*, 6(1):136–160, mar 2013. doi: 10.3390/a6010136.

[27] Yurii E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2003.

[28] Nicolas L. Roux, Mark Schmidt, and Francis R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc., 2012.

[29] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, jun 2016. doi:10.1007/s10107-016-1030-6.

[30] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, February 2013.

[31] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning - ICML'05*. ACM Press, 2005. doi:10.1145/1102351.1102451.

[32] Yuya Takada, Rikuto Mochida, Miya Nakajima, Syun suke Kadoya, Daisuke Sano, and Tsuyoshi Kato. Stochastic dual coordinate ascent for learning sign constrained linear predictors. *IEICE Transactions on Information & Systems*, Exxx-D:–, xxx 2024. accepted.

[33] Kenya Tajima, Kohei Tsuchida, Esmeraldo Ronnie R. Zara, Naoya Ohta, and Tsuyoshi Kato. Learning sign-constrained support vector machines. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, January 2021. doi:10.1109/icpr48806.2021.9412786.

[34] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang. Diverse non-negative matrix factorization for multiview data representation. *IEEE Trans Cybern*, -(-):–, Sep 2017.

[35] Yanfei Wang and Shiqian Ma. Projected barzilai–borwein method for large-scale nonnegative image restoration. *Inverse Problems in Science and Engineering*, 15(6):559–583, sep 2007. doi: 10.1080/17415970600881897.

[36] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, jan 2014. doi:10.1137/140961791.

[37] Qiang Zhang, Han Wang, Robert Plemmons, and V. Paul Pauca. Spectral unmixing using nonnegative tensor factorization. In *Proceedings of the 45th annual southeast regional conference on ACM-SE 45*. ACM Press, 2007. doi: 10.1145/1233341.1233449.

付　　録

1　定理 1 の証明

ここで示す定理 1 の証明は，主に [30] の従来の SDCA で用いられた証明技法に従う．以下の補助定理がこの証明において重要である．

補助定理 2. 任意の $\boldsymbol{v}_0, \boldsymbol{v}_q \in \mathbb{R}^d$ と $\eta \in [0,1]$ に対して，以下が成り立つ．

$$\begin{aligned}&\|\boldsymbol{\pi}[\boldsymbol{v}_0]\|^2 - \|\boldsymbol{\pi}[\boldsymbol{v}_0 + \eta\boldsymbol{v}_q]\|^2 \\ &\quad + 2\eta\left\langle \boldsymbol{\pi}[\boldsymbol{v}_0], \boldsymbol{v}_q\right\rangle + \eta^2\|\boldsymbol{v}_q\|^2 \geq 0.\end{aligned} \tag{A·1}$$

補助定理 3. $F_i^{(t-1)} := \phi_i(z_i^{(t-1)}) + \phi_i^*(-\alpha_i^{(t-1)}) + z_i^{(t-1)}\alpha_i^{(t-1)}$. とする．このとき，以下が成り立つ．

$$h_P^{(t-1)} + h_D^{(t-1)} \leq \frac{1}{n}\sum_{j=1}^{n} F_j^{(t-1)}. \tag{A·2}$$

補助定理 2 の証明: これは，$\forall h \in [n]$ に対して，

$$\begin{aligned}&\pi_h[v_{h,0}]^2 - \pi_h[v_{h,0} + \eta v_{h,q}]^2 \\ &\quad + 2\eta\pi[v_{h,0}]v_{h,q} + \eta^2 v_{h,q}^2 \geq 0.\end{aligned} \tag{A·3}$$

を証明すれば十分である．このとき，$v_h \in \mathbb{R}$ に対して，$\pi_h[v_h] := v_h + \max\{0, -\sigma_h v_h\}$ となる．これは，$h \in [n]$ 上で，(A·3) の左辺の和は，(A·1) の左辺と等しくなるためである．

$v_{h,0} + \eta v_{h,q} \geq 0$ かつ $v_{h,0} \geq 0$ となるような $h \in \mathcal{I}_0$ および $h \in \mathcal{I}_{\geq}$ に対して，(A·3) の左辺は以下のように式変形できる．

$$v_{h,0}^2 - (v_{h,0} + \eta v_{h,q})^2 + 2\eta v_{h,0}v_{h,q} + \eta^2 v_{h,q}^2 = 0. \tag{A·4}$$

$v_{h,0} + \eta v_{h,q} \leq 0$ かつ $v_{h,0} \geq 0$ となるような $h \in \mathcal{I}_{\geq}$ に対して，(A·3) の左辺は以下のように式変形できる．

$$v_{h,0}^2 - 0 + 2\eta v_{h,0}v_{h,q} + \eta^2 v_{h,q}^2 = (v_{h,0} + \eta v_{h,q})^2 \geq 0. \tag{A·5}$$

$v_{h,0} + \eta v_{h,q} \geq 0$ かつ $v_{h,0} \leq 0$ となるような $h \in \mathcal{I}_{\geq}$ に対して，(A·3) の左辺は以下のように式変形できる．

$$0 - (v_{h,0} + \eta v_{h,q})^2 + 0 + \eta^2 v_{h,q}^2 \geq 0. \tag{A·6}$$

これは，$0 \leq v_{h,0} + \eta v_{h,q} \leq \eta v_{h,q}$ より導かれる．

q.e.d. of 補助定理 *2*

補助定理 3 の証明: $\forall \boldsymbol{v} := [v_1, \ldots, v_d]^\top \in \mathbb{R}^d$ に対して，

$$\begin{aligned}\|\boldsymbol{\pi}[\boldsymbol{v}]\|^2 &= \sum_{h\in\mathcal{I}_0} v_h^2 + \sum_{h\in\mathcal{I}_{\geq}} \pi_h[v_h]^2 \\ &\leq \sum_{h\in\mathcal{I}_0} v_h^2 + \sum_{h\in\mathcal{I}_{\geq}} v_h^2 = \|\boldsymbol{v}\|^2,\end{aligned} \tag{A·7}$$

$$\begin{aligned}\|\bar{\boldsymbol{w}}^{(t-1)}\|^2 &= \left\langle \bar{\boldsymbol{w}}^{(t-1)}, \frac{1}{\lambda n}\sum_{j=1}^{n} \boldsymbol{x}_j\alpha_j^{(t-1)} \right\rangle \\ &= \frac{1}{\lambda n}\sum_{j=1}^{n} z_j^{(t-1)}\alpha_j^{(t-1)}.\end{aligned} \tag{A·8}$$

に着目する．これらの不等式は以下を導く．

$$\begin{aligned}&h_{\mathrm{P}}^{(t-1)} + h_{\mathrm{D}}^{(t-1)} = P(\boldsymbol{\pi}[\boldsymbol{\alpha}^{(t-1)}]) - D(\boldsymbol{\alpha}^{(t-1)}) \\ &= \lambda\|\boldsymbol{\pi}[\bar{\boldsymbol{w}}^{(t-1)}]\|^2 + \frac{1}{n}\sum_{j=1}^{n}\phi_j(z_j^{(t-1)}) + \phi_j^*(-\alpha_j^{(t-1)}) \\ &\leq \lambda\|\bar{\boldsymbol{w}}^{(t-1)}\|^2 + \frac{1}{n}\sum_{j=1}^{n}\phi_j(z_j^{(t-1)}) + \phi_j^*(-\alpha_j^{(t-1)}) \\ &= \frac{1}{n}\sum_{j=1}^{n} F_j^{(t-1)}.\end{aligned} \tag{A·9}$$

q.e.d. of 補助定理 *3*

ここまでで定理 1 を証明する準備ができた．$z_i^{(t-1)} := \left\langle \boldsymbol{w}^{(t-1)}, \boldsymbol{x}_i \right\rangle$, $u_i^{(t-1)} := -\nabla\phi_i(z_i^{(t-1)})$, $q_i^{(t-1)} := u_i^{(t-1)} - \alpha_i^{(t-1)}$, $\bar{\beta} := \lambda\gamma/(\lambda n\gamma + R^2)$ とする．このとき，

$$\begin{aligned}&h_{\mathrm{D}}^{(t-1)} - h_{\mathrm{D}}^{(t)} \geq J_t^1(\eta_t) \geq J_t^1(n\bar{\beta}) \\ &\geq -2\lambda n\bar{\beta}\left\langle \boldsymbol{w}^{(t-1)}, \frac{\boldsymbol{x}_i q_i^{(t-1)}}{\lambda n}\right\rangle - \lambda n^2\bar{\beta}^2\left\|\frac{\boldsymbol{x}_i q_i^{(t-1)}}{\lambda n}\right\|^2 \\ &\quad + a_{\mathrm{offs}}n^2\bar{\beta}^2 + b_{\mathrm{offs}}n\bar{\beta} \\ &= \bar{\beta}F_i^{(t-1)} + \frac{\gamma\bar{\beta}(q_i^{(t-1)})^2}{2}\left(1 - \frac{(\lambda n\gamma + \|\boldsymbol{x}_i\|^2)\bar{\beta}}{\lambda\gamma}\right) \\ &\geq \bar{\beta}F_i^{(t-1)} + \frac{\gamma\bar{\beta}(q_i^{(t-1)})^2}{2}\left(1 - \frac{(\lambda n\gamma + R^2)\bar{\beta}}{\lambda\gamma}\right) = \bar{\beta}F_i^{(t-1)}\end{aligned} \tag{A·10}$$

が成り立つ．ここで，3 番目の不等式は補助定理 2 より得られた．t 回目の反復で $i \in [n]$ を選択するためのランダム性に関する期待値をとると，以下が得られる．

$$\begin{aligned}&h_{\mathrm{D}}^{(t-1)} - \mathbb{E}\left[h_{\mathrm{D}}^{(t)}\right] \geq \frac{\bar{\beta}}{n}\sum_{i=1}^{n} F_i^{(t-1)} \\ &\geq \bar{\beta}\cdot\left(h_{\mathrm{P}}^{(t-1)} + h_{\mathrm{D}}^{(t-1)}\right) \geq \bar{\beta}\cdot\max\left\{h_{\mathrm{P}}^{(t-1)}, h_{\mathrm{D}}^{(t-1)}\right\}\end{aligned} \tag{A·11}$$

ここで，2 番目の不等式は補助定理 3 より成り立つ．これは，以前の反復におけるランダム性に関して，期待される目的誤差の下限を導く：

$$\begin{aligned}\mathbb{E}[h_{\mathrm{P}}^{(t)}] &\leq \bar{\beta}^{-1}\mathbb{E}\left[h_{\mathrm{D}}^{(t)} - h_{\mathrm{D}}^{(t+1)}\right] \leq \bar{\beta}^{-1}\mathbb{E}\left[h_{\mathrm{D}}^{(t)}\right] \\ &\leq \bar{\beta}^{-1}\mathbb{E}\left[h_{\mathrm{D}}^{(t-1)}\right](1-\bar{\beta}) \\ &\leq \bar{\beta}^{-1}h_{\mathrm{D}}^{(0)}\cdot(1-\bar{\beta})^t \leq \bar{\beta}^{-1}h_{\mathrm{D}}^{(0)}\exp\left(-\bar{\beta}t\right).\end{aligned} \tag{A·12}$$

よって，任意の $\epsilon_{\mathrm{P}} > 0$ に対して，$\mathbb{E}[h_{\mathrm{P}}^{(t)}] \leq \bar{\beta}^{-1}h_{\mathrm{D}}^{(0)}\exp\left(-\bar{\beta}t\right) \leq \epsilon_{\mathrm{P}}$ が成り立つとき，以下を得る．

$$t \geq \frac{1}{\bar{\beta}}\log\left(\frac{h_{\mathrm{D}}^{(0)}}{\epsilon_{\mathrm{P}}\bar{\beta}}\right). \tag{A·13}$$

q.e.d. of 定理 *1*

SABM：大規模言語モデルに基づくエージェントベース実世界シミュレーション

呉　増青[†,††]　彭　潤[†††]　韓　勗[††††]　鄭　舒元[†]　肖　川[†,†††††]

† 大阪大学情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5
†† 京都大学情報学研究科　〒 606–8501 京都市左京区吉田本町 36–1
††† ミシガン大学コンピューターサイエンス工学部　2260 Hayward Street, Ann Arbor, MI 48109, USA
†††† フォーダム大学ガベリビジネススクール　140 W 62nd St, New York, NY 10023, USA
††††† 名古屋大学情報学研究科　〒 464–8601 名古屋市千種区不老町
E-mail: †wuzengqing@outlook.com, ††roihn@umich.edu, †††xhan44@fordham.edu,
††††zheng@ist.osaka-u.ac.jp, †††††chuanx@nagoya-u.jp

あらまし　エージェントベースのモデリング（ABM）は，コンピュータシミュレーションの分野において特に影響力のあるアプローチである．それは，個々のエージェント間の相互作用を利用して複雑なシステムダイナミクスをシミュレートするものである．本研究では，従来の ABM 方法が非定型的な行動パターン，主観的判断，常識，コミュニケーションのシミュレーションにおいて不足している問題に対処し，新たなスマート ABM フレームワーク（SABM）を構築した．このフレームワークでは，大規模言語モデルに基づくエージェントを使用して，エージェント間の相互作用をより現実的に設定し，実際の現象を模倣し解釈することができる．また，SABM に基づくケーススタディを提示し，現実世界のシステムモデリングにおけるこのシミュレーション方法の有効性を示すとともに，複雑なシステムをより深く理解する上でのその潜在的な可能性について議論する．

キーワード　エージェントベースのモデリング, 大規模言語モデル, 実世界シミュレーション

1　はじめに

コンピュータシミュレーションの出現は，様々な分野における問題解決へのアプローチを革新的に変革し，複雑なシステムを理解するための強力なツールを提供した．シミュレーション手法の中でも，エージェントベースモデリング（ABM）は，単純なルールによって特定された個々のエージェントの相互作用を通じて複雑なシステムをモデル化する独自の能力により，特に有力なツールとして登場している [5]．このボトムアップ構築アプローチにより，ABM はこれらの相互作用から生じる複雑な現象を観察し理解することを研究者に可能にし，理論やモデルの開発を促進する．ABM は多様な分野での応用が見られ，疫学における病気の蔓延モデル [24]，経済学における市場ダイナミクスのシミュレーション [26]，生態学における生態系の振る舞い理解 [16] など，これらの分野における複雑なシステムの理解を進展させる上で重要な役割を果たしてきた．これらの研究は，ABM が取り組む問題の幅広さを示し，コンピュータシミュレーションの領域におけるその重要性を強調している．

エージェントベースモデリング（ABM）は，複雑なシステムのシミュレーションにおいて強力なツールであるが，その限界も存在する．例えば，自然言語の記述や常識を含む高度な人間行動のモデリングにおいて，これらを形式言語で表現することは困難であり，ABM 手法ではこれらの自然言語要素を無視するか，発見則や学習規則によって簡略化することが一般的である．これらの行動の高い複雑さのため，このようなモデリングはしばしば過度の単純化をもたらし，その特性の一部のみを捉えるに留まり，その精度を制限し，実世界のシナリオへの関連性を低減する [5]．さらに，ABM 手法はエージェントの行動を定義するために予め決定されたパラメータに大きく依存する．このアプローチは特定のシナリオでは効果的だが，パラメータ設定が適切に校正されていない場合には，研究者の偏見やモデル出力の高い不確実性を引き起こす可能性がある [16]．

このような限界に対処するために，生成的事前学習トランスフォーマー（GPT）などの大規模言語モデル（LLM）の出現が新たな可能性を提供している．自然言語の処理と生成が可能な LLM は，前述の高度な人間行動を模倣するために使用できる．プロンプトエンジニアリング [28] を LLM 上で行うことにより，研究者は自然言語でモデルを設計し，エージェントの行動を定義するためのパラメータ選択において労力を軽減できる．さらに，LLM は強力な推論能力を持っており [3]，これを活用することでモデルの解釈可能性を向上させることができる．LLM を ABM に組み込むことにより，研究者は定められた規則に従うだけでなく，手作りの例 [6] や推論指示 [17] から学習することにより意思決定を行うエージェントを作成することができる．これによりシミュレーションのリアリズムが高まり，システムの振る舞いを駆動するメカニズムに関するより深い洞察が得られる．LLM の学習能力を活用することで，コンピュータシミュレーションはシステムの変化に対してより適応的となり，より堅牢なシミュレーションが可能となる．LLM はさま

ざまな役割を模倣することもでき [25]，実世界のシステムで見られるエージェント行動の多様性を捉える．したがって，LLM を ABM に組み込むことは，シミュレーションにおける ABM 手法のパフォーマンスを向上させる可能性を秘めている．

本論文の主要な目的は，エージェントベースのアプローチの能力を向上させ，理論，仮説，説明を自然言語に基づくコンピュータシミュレーションの枠組みで構築することである．本論文では，実世界のシステムをモデル化するために，特に大規模言語モデル（LLM）を活用する最新の AI モデルの力を利用した革新的なエージェントベースのアプローチである**スマートエージェントベースモデリング（SABM）**のモデリングフレームワークを提案し，この枠組みに基づくコンピュータシミュレーションの方法論的構築を紹介する．

スマートエージェントの概念は，将来の組織の文脈において，Carley によって提案された [9]. [9] において，スマートエージェントは，知的で，適応的で，計算可能な実体として定義され，人間は典型的なスマートエージェントである．現代の AI モデル，特に LLM の出現により，人間の行動を模倣する驚異的な言語能力と推論能力を備えたこれらのモデルを用いて，ABM の領域におけるスマートエージェントの概念を拡張する．我々は，LLM によって駆動されるこれらのスマートエージェントが，より繊細で現実的な方法で現実世界の問題をシミュレートすることにより，複雑なシステムの理解を深めることができると考えている．

2 関 連 研 究

2.1 ABM が対象とする問題の本質

現在，複雑なシステムをシミュレートするために，さまざまな発展した方法が使用されている．これらの方法には，マイクロシミュレーション，システムダイナミクス，ABM が含まれる．マイクロシミュレーションは，実際のサンプルから全体的な予想状況を予測する方法であり，個々のデータが利用可能な状況で特に有用である．これは個々の特性と転移確率の数に基づいて統計結果を提供するが，主な限界はエージェント間の相互作用を考慮していないことである [13, pp. 13, 58]．一方，システムダイナミクスは，マクロスケールの変数セットに焦点を当てた方法である．これは多重フィードバックループを含む大規模システムを扱うことができ，システムの変数の時間経過による動的な因果関係を捉えることができる [2]．しかし，システムダイナミクスではシステム全体が唯一のエージェントとしてシミュレートされるため，マイクロレベルからのモデル対象の違いを区別することが難しく，個々の行動や適応性を実行することが困難である [13, pp. 28–30]．ABM は，システムの挙動が個々のエージェントの相互作用から生じる状況で特に有用である．しかし，現実に合致するエージェントと相互作用を確立することの難しさも抱えている．これらのアプローチの比較は表 1 に示されている．これらの方法はそれぞれ適用範囲を持ち，複雑なシステムの研究において互いに補完し合っている．一般的に言えば，個々の間の動的で複雑な相互作用がシステム全体

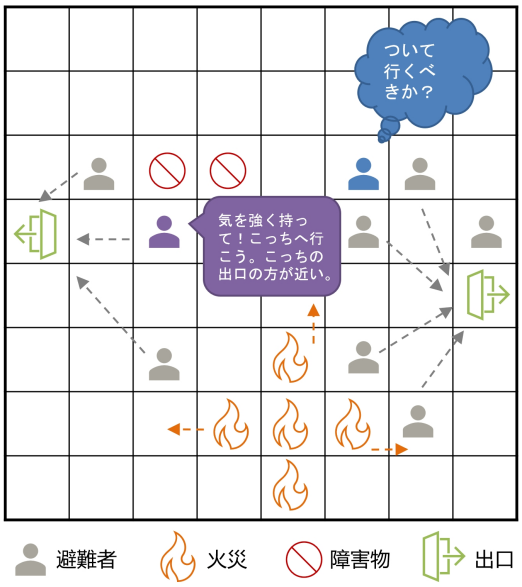


図 1 ABM ではモデル化が困難な要素を含む火災避難シミュレーション．

のパフォーマンスにどのように影響するかを理解したい場合，ABM を使用することが良い選択である [11]．

2.2 現在の ABM 手法の欠点

ABM は理論上，上記の特性を持つ問題を扱うことが可能であるが，従来のモデルや計算能力の制約により，実際の運用においては依然として限界が存在する．議論は，以下の例から始める．

例として，図 1 は，避難者の行動をモデル化することに焦点を当てた火災避難のシミュレーションを示している．避難中の情報共有の実世界のダイナミクスを捉えるために，エージェント間のコミュニケーションをシミュレーションに含める．一部のエージェントは自然に情報の伝達者となり，近くのエージェントに自然言語のメッセージを伝える．これらのエージェントは，メッセージおよび出口までの距離や混雑のレベルなどの他の要因に基づいて，これらのメッセージの提案に従うかどうかを決定する．コミュニケーション内容のシミュレーションプロセスは自然言語生成を要求し，膨大な数のアクション（すなわち，すべての可能なメッセージを出力として）を生み出す．一方，コミュニケーションを考慮に入れたフォロワーの意思決定をシミュレートするためには，自然言語理解が必要であり，膨大な数の状態（すなわち，すべての可能なメッセージを入力として）を処理する必要がある．これらの複雑さは，意思決定ルールと学習アルゴリズムに依存する現在の ABM 方法の限界を超えている．

上記の例は行動科学のシナリオを強調しているが，同様の状況は他の分野でも見られる．たとえば，独立した市場プレイヤーが共謀して利益を増やすカルテルの場合を考えてみよう．現在の市場競争のシミュレーション研究は，人間のボランティ

表 1　複雑なシステムを研究するためのシミュレーション方法の比較.

シミュレーション方法	研究スケール	個体行動	システム状態
マイクロシミュレーション	ミクロ	単純（統計）	静的
システムダイナミクス	マクロ	なし（システムレベルの行動）	動的
ABM	ミクロ	複雑	動的

アに依存する [1] か，暗黙の共謀に焦点を当てて，企業間のコミュニケーションを迂回する [8]．これらの例では，自然言語と常識に基づいてコミュニケーションをシミュレートすることが，問題のモデリングと理解に不可欠である．リアルな状況におけるコミュニケーションの範囲が特定の範囲に限定されていないため，これらのエージェント間のコミュニケーションの影響をその行動に単純にマッピングすることはできない．ABM アプローチは特定のタスクのために事前学習された言語モデルを使用することがあるが [31]，言語はエージェントの入出力の一部としてほとんど使用されない．ABM の限界により，これらの複雑な行動を含む問題をモデル化することは困難である．

3 提案手法

ABM のシミュレーションにおける基本的な側面は，その個々の構成要素であるエージェント，環境，相互作用，そしてこれらの要素を特徴づけるパラメータとデータから成る [19]．シミュレーションの品質と関連性は，これらの要素の正確な表現に大きく依存しており，モデルの全体的なパフォーマンスに影響を与える．関連研究で要約された制約を踏まえ，これらの制約に対処し，コンピュータシミュレーションの適用範囲を拡大する方法が求められている．自然言語理解と人間行動の模倣において優れた能力を持つ LLM [7] を利用することにより，ABM の構成要素に LLM を統合することが改善のための潜在的な道である．我々は，LLM を使用し，その構成要素を自然言語で定式化することにより，ABM からモデリングアプローチを拡張するスマートエージェントベースモデリング（SABM）を提案する．この新しいモデリングアプローチの名前には，Carley によって定義された，知的で，適応的で，計算可能な実体としてのスマートエージェント [9] という用語を借用している．これらの特性は，LLM を活用することにより新しいモデリングアプローチが具備するものである．

図 2 は，SABM の特徴を示し，それを ABM の特徴と比較している．続く議論では，言語能力，モデリングパラダイム，適応性，解釈可能性などの観点から SABM の特徴を探求する．これらの側面を分析することにより，SABM がコンピュータシミュレーションにもたらす可能性のある改善と，SABM の潜在的な適用範囲を判断することを目指す.

3.1 エージェントに自然言語能力を持たせることの再考察

SABM は，自然言語を入力と出力の両方で取り扱うことができるエージェントを生み出し，より繊細でリアリスティックなシミュレーションを実現し，その適用範囲を拡大する．自然言語を取り扱う効果は，モデリングと理解能力の観点から説明されるであろう.

3.1.1 自然言語におけるモデリング

LLM を核として，エージェントは自然言語を入力と出力の両方で処理する能力を持つ．エージェントの属性を定義するために経験則や数学的な公式に依存する代わりに，自然言語で与えられたテキスト記述が主要な参照となる.

例：アリの行動のモデリング ゴードンは ABM を利用してアリの行動をモデル化した [14]．この研究では，個々のアリの行動は，観察されたアリの行動を反映するように設計・微調整された数学的ルールに基づいてモデル化された.

数学的アプローチとは対照的に，SABM はアリの決定プロセスに関する質的な記述に基づいてアリの行動をモデル化する可能性を提示する．研究者はアリがどのようにして採餌場所を決定し，仲間のアリと食糧源の場所をどのように伝達し，環境の変化にどのように適応するかについての自然言語の記述を提供することができる．これらの自然言語の記述は，数学的表現よりも直感的に理解しやすく，アリの行動をモデル化するためのアクセスしやすい方法を提供する．また，LLM はこれらの自然言語の記述を利用して，モデル内のアリエージェントに適切な行動を生成することができる．たとえば，モデルはアリが他のアリが食糧を見つけた場所でより多く採餌する可能性が高いことを示すかもしれない．この情報は，シミュレーション内のアリエージェントの行動を指示するために利用される.

さらに，ルールを表現するためのパラメータの使用を避けるこのアプローチは，シミュレーション実験の再現性と比較可能性にも寄与する：同じタイプの背景を持つ研究問題に対して，同様のプロンプトを使って背景設定を記述し，その上に問題固有の要素を追加することができる.

3.1.2 複雑な行動のモデリング

ABM から自然言語を省略することは，形式言語で表現するのが難しい記述を正確にモデル化し，自然言語が重要なシステムをシミュレートする際に情報交換の詳細を失うことを意味する [21], [22]．LLM が自然言語入力を理解し，文脈に関連した一貫性のある出力を作成する能力を活用することにより，SABM はこの問題に対する解決策を提供する.

例：公衆衛生 公衆衛生の分野を考えると，ABM は病気の蔓延を予測し，さまざまな介入の影響を評価するために利用されている．マーシャルらは，複雑なシステムアプローチを使用して高リスク集団における HIV 伝播をモデル化し，統合予防介入の潜在的な集団的利益について貴重な洞察を得た [20]．このようなモデルでは，エージェントは個々の行動や相互作用が定められたルールに基づいている個人を表す．しかし，通信や意思決定プロセスなど，人間の行動の複雑さやニュアンスを取り込むことにしばしば苦労している．これらは通常，自然言語で説明される.

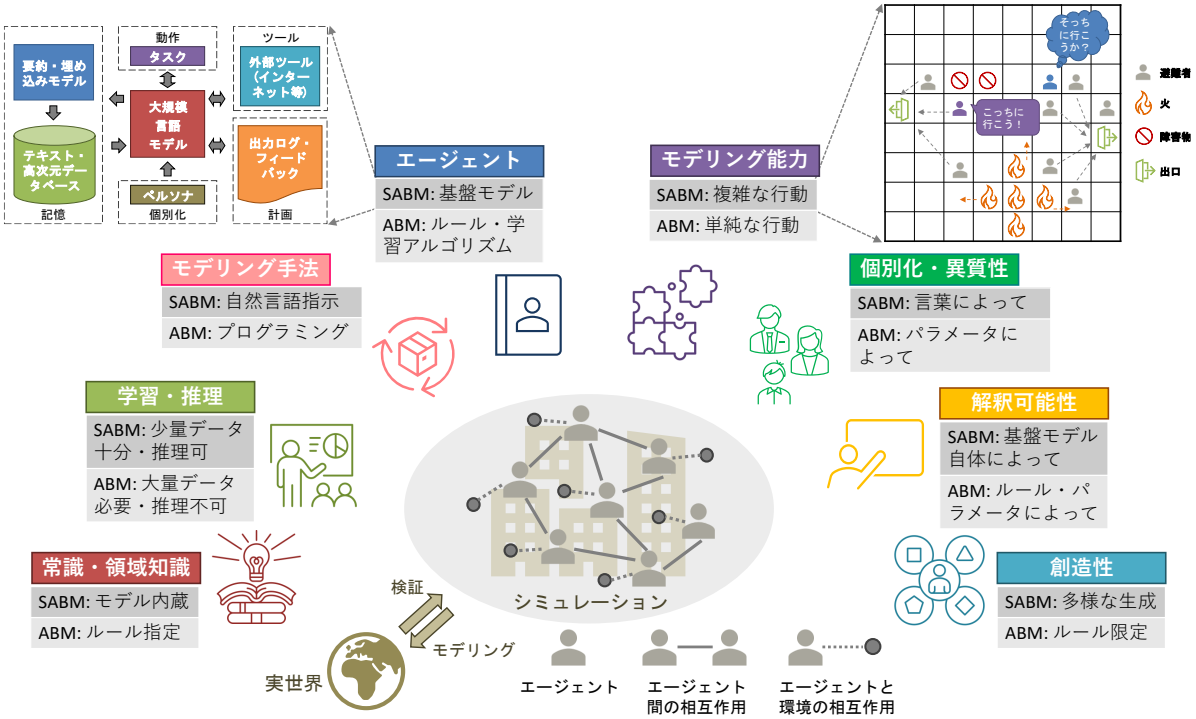


図 2　SABM と ABM の特徴の比較.

SABM は，エージェントが自然言語による指示（LLM の文脈では**プロンプト**と呼ぶ）を処理することを可能にする．エージェントは自然言語で記述された複雑な振る舞いや相互作用をシミュレートできる．例えば，エージェントは自身の健康状態について議論したり，疾病予防情報を共有したり，これらの議論に基づいて検査や治療を求める決定をすることができる．これにより，コミュニティ内での情報伝播とその行動への影響をよりリアルにシミュレートすることが可能となる．

3.2　事前モデリングのためのアラインメント

多くの LLM は，人間の価値観と知識とのアラインメントを目指し，RLHF を用いて微調整されている [23]．これにより，モデルのパフォーマンスが向上するだけでなく，その反応の倫理的・道徳的なアラインメントも増すことが期待される [12], [18]．これは SABM にとって特に有益であり，人間の相互作用により近いシミュレーションを可能にし，シミュレーション結果の精度を高める．

Carley の定義によれば，人間は典型的な賢明なエージェントであるが [9]，GPT-4 や他の LLM は，少なくとも短い会話においては多くの人を欺くことができ，おそらくチューリングテストに合格するとされている [4]．評価によると，GPT-3.5/4 は感情認識 [10]，数学，コーディングなどの面で人間と同等かそれ以上のパフォーマンスを発揮するという．LLM がこれらの能力を持ち，人間の価値観や行動とアラインメントされていることから，SABM 内で事前モデリングパラダイムを採用することが可能となる．逆に，ABM では，行動規則やパラメータが観察された人間の行動から導き出されたり推測されたりする事後モデリングパラダイムを採用する．

これら二つのパラダイムの違いを明確にするために，消費者行動の領域を考える．ABM では，研究者は実際の消費パターンを調査し，それを数学的表現や公式に落とし込むことで，消費者が商品やサービスを選ぶ際の細かなニュアンスを表現する．しかし，実際には，個人の購買決定は，必要性，個人的好み，認識された価値を評価するといったより本能的な推論や常識によって支配されていることが多い．SABM の枠組み内では，このような消費行動を，人間とアラインメントされた LLM の本能的な常識から出発点として模倣することができる．これにより，そのような行動を包括するための規則やパラメータが不要となる．さらに，LLM は少数例から学習することができるため，行動を人間のデータにさらに調整することが可能である．

人間の価値観と知識とのアラインメントのおかげで，SABM は事前モデリングの地平を広げる．事前モデリングは従来，実データに依存しない形式科学や理論研究（例えば，初等セルオートマトン [29, pp. 23–50]）に限定されていた．さらに，SABM では，モデルを二つのコンポーネントに分けることができる．「エンジン」は LLM によって表され，「シャーシ」はモデルの残りの部分を包括する．AI 研究者がエンジンの強化に取り組むことで，事前モデリングの基盤（上記の仮定）を固めると同時に，SABM を使用する人々はシャーシの作成に集中し，特定の取り組みに最も適したエンジンを使用することができる．この実践はモデリングの難易度を大幅に低減し，研究者が迅速にモデリングのメダワーゾーンを見つけるのに役立つ [16]．より高度な LLM が利用可能になった場合，研究者はエンジンを交換するだけでパフォーマンスを向上させることができる．一方，ABM では，人間の行動を模倣するエージェントのパフォーマンス向上とシミュレーションモデルの構築が密接に結びついており，

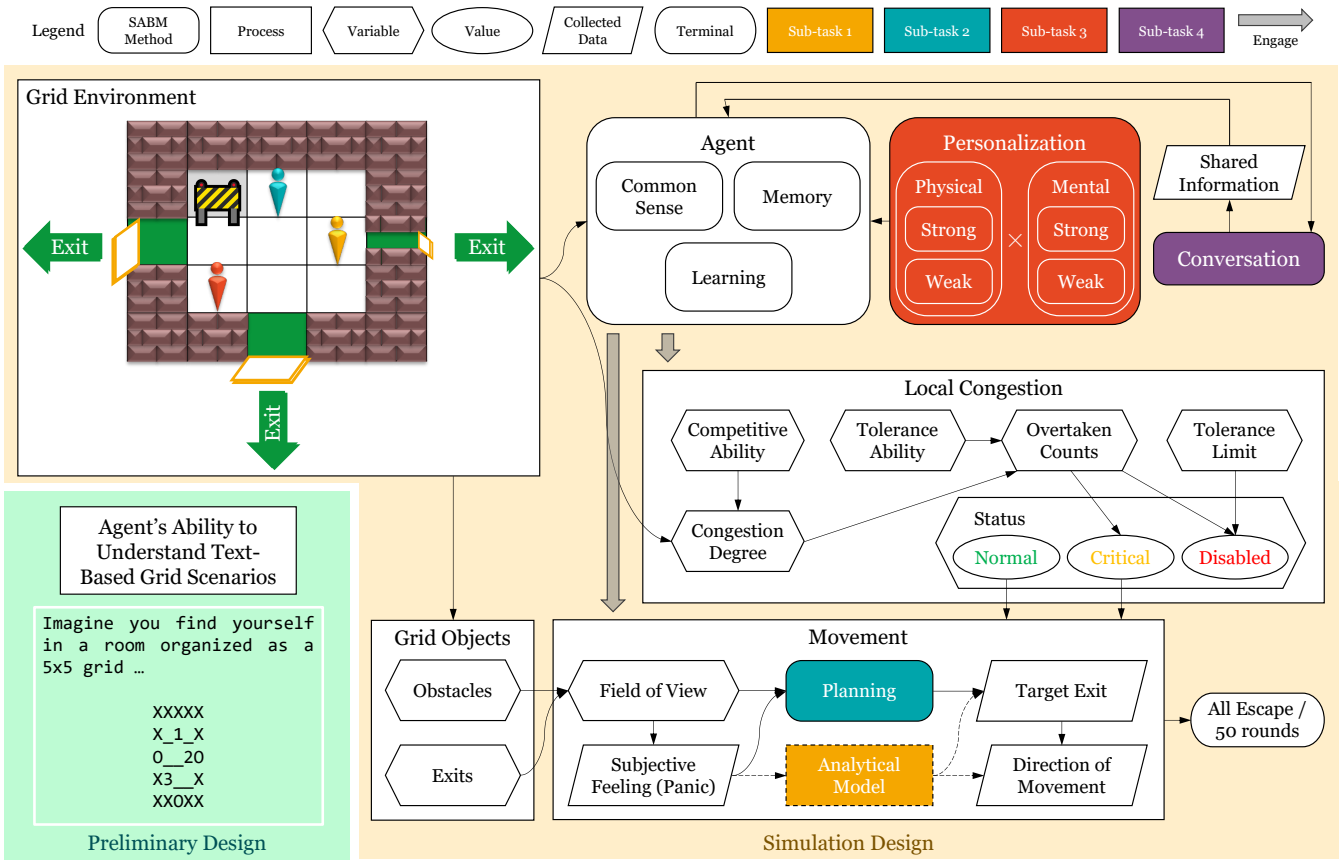


図 3　緊急避難のモデル設計.

研究者はこれらの両方の手続きを管理する必要があり，それは必然的により多くの専門知識とモデリング努力を要求する．

3.2.1　共 通 認 識

先に議論されたように，LLM の注目すべき特徴の一つは，共通認識推論能力である [32]．この能力は，幅広いトピックや日常知識を含む広範かつ多様な訓練データから派生する．共通認識推論は，SABM におけるモデル設定を単純化し，複雑な構成の必要性を減少させることができる．例えば，エージェントはこの特性を利用して，環境を理解し，他のエージェントと相互作用し，文脈に基づいて行動を適応させることができる．これらの能力を明示的にプログラムする必要はない．

3.2.2　内 蔵 知 識

ABM はしばしば，エージェントに明示的にコーディングされたドメイン知識によって制約され，その行動の深さと多様性が限定される．これに対して，SABM は，多様なインターネットテキストによる訓練を受けた LLM の持つ広大なドメイン知識のリソースにアクセスすることが可能である．LLM に内在するドメイン知識は，SABM を単純化し豊かにし，潜在的な応用の範囲を広げ，これらのシミュレーションの複雑なシステム理解への有用性を高める大きな機会を提供する．

3.2.3　個　人　化

人間とのアラインメントにより生じる SABM の重要な特徴は，エージェントの個人化である [25], [27]．つまり，エージェントは，シミュレートされるシナリオに関連する異なる特性，例えば異なる性格やスタイルを持つように個人化することができる．この異質性は，シミュレーション内の現実感を高め，モデルの忠実度を向上させる．

例：歩行者の避難 このような応用の一例として，ABM を使用した歩行者避難シミュレーションがある [30]．この研究では，個人化された空間認識道路ネットワーク（PSCRN）モデルが導入されている．PSCRN モデルは，各歩行者が空間認識に基づいて特定の避難経路を認識する確率が異なるという仮定に基づいている．この個人化は統計的確率に基づいており，理論的およびシミュレーション的に健全であるが，個々の経験や特性を完全に包括することはできないかもしれない．これは元の研究の限界ではないが，LLM はシミュレーションと方法論をさらに発展させる可能性がある．

SABM は，エージェントの特徴付けに自然言語をツールとして利用することで，個人化の度合いを高めることができる．各エージェントは，知識，経験，視点を反映した特有の言語入力を用いて個人化される．例えば，歩行者避難シミュレーションでは，エージェントに環境に関する多様な知識を割り当てることができる．これは，彼らの個人的な経験や特性を反映している．高齢者を表すエージェントは，異なる知識を持ち，動きが遅い可能性がある一方で，若いアスリートを象徴するエージェントとは区別される．このような個人化の度合いにより，より現実的で微妙なシミュレーションが可能となる．重要なことに，複雑なパラメータ設定を回避し，自然言語の手がかりを用いてエージェントの個人化プロセスを単純化することができる．

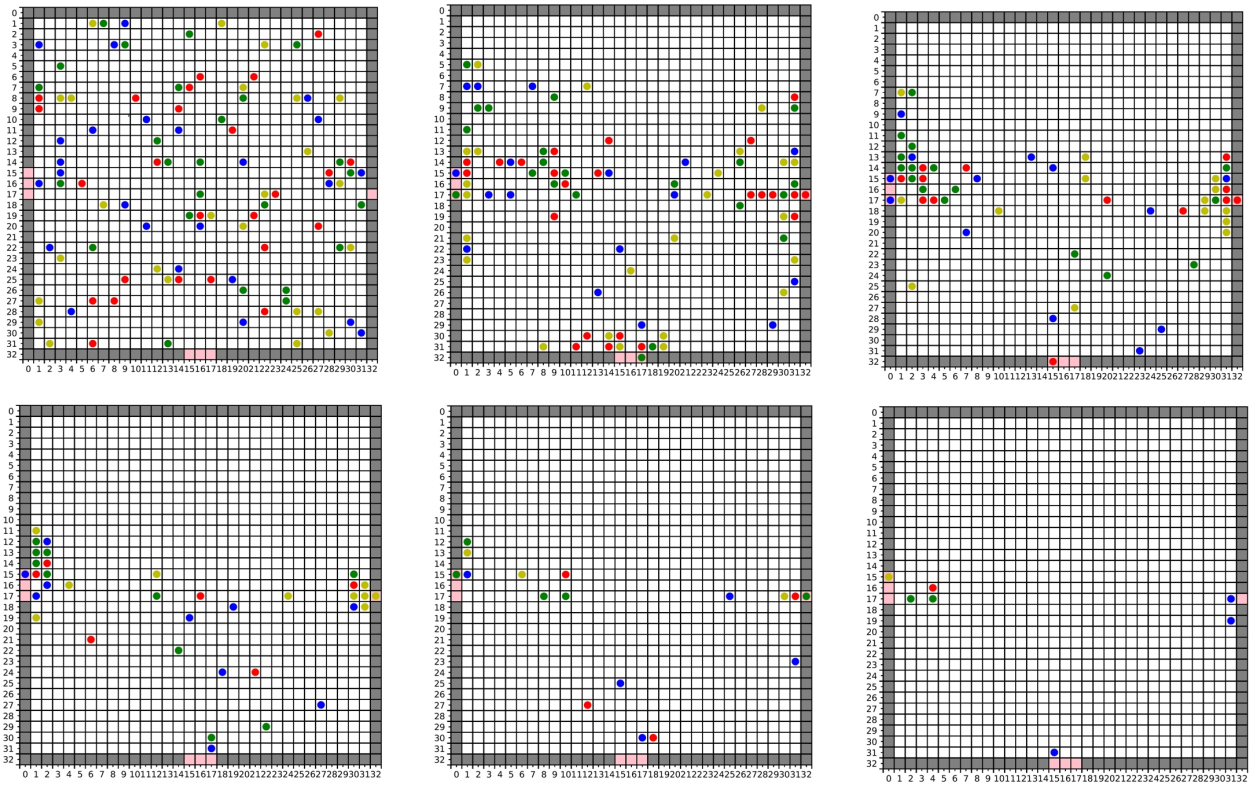


図 4　緊急避難の実行スナップショット.

3.3　より適応的で解釈可能な相互作用

静的で規定されたルールとパラメータを持つ ABM は，動的で複雑なシステムを正確に表現するために必要な柔軟性と適応性を欠いている．ABM に LLM を導入することで，硬直的なルールベースの相互作用から脱却し，よりダイナミックで適応的なエージェント間の相互作用を可能にする．

3.3.1　学習と推論

LLM の固有の学習と推論能力は，少ないデータサンプル [6] やデータサンプルがない場合でも [17]，様々なタスクを効果的に行うことがすでに証明されている．SABM では，この学習と推論能力により，事前のトレーニングプロセスや学習アルゴリズムの必要性を回避する．その代わりに，現実に密接に模倣する適応的な学習の側面を導入し，実世界のシステムのダイナミックな性質を捉えることがしばしば困難な規定されたルールからの重要な進歩を示している．

3.3.2　モデルの解釈可能性

LLM によるエージェントの解釈能力とテキスト生成能力を組み合わせることで，SABM の解釈可能性を向上させる独自の機会を提供する．ABM におけるエージェント間の相互作用では，エージェントの行動の背後にある動機がしばしば曖昧である．これらの行動を複雑なパラメータやルールの複雑な網から解き明かすことは困難な作業であり，モデルのダイナミクスを包括的に理解することを妨げることがある．一方で，LLM の解釈的な性質を活用することにより，エージェントの行動や相互作用をより直感的で人間に理解しやすい表現で抽出することができる．規定されたルールに従って相互作用するのではなく，エージェントは自分の「意図」や「感情」を自分のパフォーマンスと他のエージェントとの相互作用に基づいて解釈し，それに応じて行動を調整することができる．これは，意味が文字通りの言語自体ではなく，言語的手がかりの解釈からしばしば導かれる人間のコミュニケーションの解釈的な性質を反映している [15]．この文脈では，エージェント相互作用の基礎となるテキスト自体が，シミュレーションを解釈する能力を持つ．

上記の SABM の特徴についての議論から，SABM は実際の状況の複雑さ，常識，適応性，およびその他の特性を再現するためのより広い適用範囲を持ち，問題解決に適切なリアリズムのレベルを保証し，ABM よりも優れた解釈可能性を持つことが期待できる．モデル化されたシステムのシミュレーション度の改善は，複雑でリアルなシミュレーション設定の説明能力を維持しながら，シミュレーション結果と現場調査結果とのギャップを狭めることができる．

4　実　　験

実験では，緊急避難の複雑なパターンをシミュレートするために，100〜400 の異質なエージェントをモデル化する（図 3）．これらのエージェントはそれぞれ避難者を表し，共通感覚推論，記憶，学習，計画立案，会話などの方法を通じて複雑な行動をシミュレートする．モデルの設計には，ABM を用いた緊急避難のシミュレーションに関する研究を基盤とし，そのコンポーネントを SABM で段階的に置き換える．様々な人口密度のシ

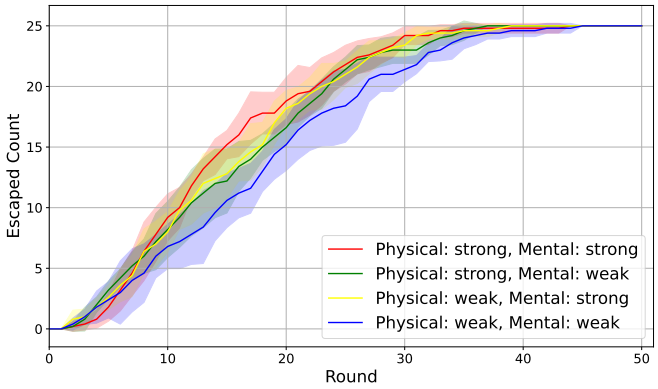


図 5　緊急避難の実行結果.

ナリオをシミュレートし，緊急避難のダイナミクスを検証する．主に，個々の特性（身体的な強さ，精神状態，位置関連要因）がエージェントの意思決定や動きにどのように影響するか，そして彼らが同僚の行動にどう反応するかを観察する．特に，混雑の存在がエージェントに元の計画された避難経路を放棄させ，より近い出口ではないかもしれないが混雑が少ない代替ルートを探すようにさせるかどうかに注目する．

この研究は，大規模言語モデルを用いたエージェントベースモデリングが，緊急避難シミュレーションのような複雑なシナリオを効果的にシミュレートするための有望なアプローチであることを示唆している（図 4，図 5）．今後の研究では，さらに多様な状況下でのモデルの適用性と精度を評価する必要がある．

5　まとめと今後の課題

本論文では，GPT のような LLM を ABM に組み込むことで，SABM フレームワークを提案した．LLM が強化されたエージェントが，複雑な行動をモデル化し，実世界の状況をより本物らしく再現する方法について検討した．ABM の詳細なレビューを提供し，SABM の能力と方法論を説明し，SABM の方法論と実世界システムのエミュレーションにおけるその有用性を示すケーススタディを提供する．さらに，SABM の将来の方向性について概説し，その使用の広範な範囲を予見する．

LLM の統合を超えて，SABM は，自然言語でのモデリングを容易にし，ABM にはない 2 つの側面である事前モデリングパラダイムを採用することで，ABM とは区別される．行動の複雑さを測定するための基礎理論を提唱し，これにより SABM を ABM とさらに対比させるのに役立つ．また，LLM が人間の行動をどの程度模倣できるか，および SABM の適用範囲の境界を定義する上でも助けとなる．技術的な観点からは，より進んだプロンプトエンジニアリングが適用され，シミュレーションの性能を向上させる可能性がある．

6　謝　　辞

本研究の一部は，JSPS 科研費 JP22H03903，JP23H03406，JP23K17456，JP21K19767，JP22KJ1721，JST CREST JPMJCR22M2 の助成を受けたものである．ここに記して謝意を表す．本研究を遂行するにあたり，資金面および設備面での支援をしてくださった鬼塚真先生および佐々木勇和先生に感謝する．

文　　献

[1] Maximilian Andres, Lisa Bruttel, and Jana Friedrichsen. How communication makes the difference between a cartel and tacit collusion: A machine learning approach. *European Economic Review*, 152:104331, 2023.

[2] Ahmad Taher Azar. System dynamics as a useful technique for complex systems. *International Journal of Industrial and Systems Engineering*, 10(4):377–410, 2012.

[3] Jacob Bayless. It's not just statistics: GPT-4 does reason. `https://jbconsulting.substack.com/p/its-not-just-statistics-gpt-4-does`, 2023.

[4] Celeste Biever. ChatGPT broke the turing test-the race is on for new ways to assess AI. *Nature*, 619(7971):686–689, 2023.

[5] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287, 2002.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

[9] Kathleen M Carley. Smart agents and organizations of the future. *The handbook of new media*, 12:206–220, 2002.

[10] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058, 2023.

[11] Joël Foramitti. Agentpy: A package for agent-based modeling in python. *Journal of Open Source Software*, 6(62):3065, 2021.

[12] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

[13] Nigel Gilbert and Klaus Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[14] Deborah M Gordon. Ant encounters. In *Ant Encounters*. Princeton University Press, 2010.

[15] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[16] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M Mooij, Steven F Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L DeAngelis. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, 310(5750):987–991, 2005.

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[18] Gabrielle Kaili-May Liu. Perspectives on the social impacts of reinforcement learning with human feedback. *arXiv preprint arXiv:2303.02891*, 2023.

[19] Charles M Macal and Michael J North. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14–pp. IEEE, 2005.

[20] Brandon DL Marshall, Magdalena M Paczkowski, Lars Seemann, Barbara Tempalski, Enrique R Pouget, Sandro Galea, and Samuel R Friedman. A complex systems approach to evaluate hiv prevention in metropolitan areas: preliminary implications for combination intervention strategies. 2012.

[21] Birgit Müller, Stefano Balbi, Carsten M Buchmann, Luís De Sousa, Gunnar Dressler, Jürgen Groeneveld, Christian J Klassert, Quang Bao Le, James DA Millington, Henning Nolzen, et al. Standardised and transparent model descriptions for agent-based models: Current status and prospects. *Environmental Modelling & Software*, 55:156–163, 2014.

[22] Mark E Nissen and Raymond E Levitt. Agent-based modeling of knowledge dynamics. *Knowledge Management Research & Practice*, 2(3):169–183, 2004.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[24] Liliana Perez and Suzana Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International journal of health geographics*, 8(1):1–17, 2009.

[25] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*, 2023.

[26] Leigh Tesfatsion. Agent-based computational economics: Growing economies from the bottom up. *Artificial life*, 8(1):55–82, 2002.

[27] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023.

[28] Lilian Weng. Prompt engineering. `https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/`, 2023.

[29] Stephen Wolfram et al. *A new kind of science*, volume 5. Wolfram media Champaign, IL, 2002.

[30] Lei Wu and Hui Lin. A personalized spatial cognitive road network for agent-based modeling of pedestrian evacuation simulation: A case study in Hong Kong. *Annals of GIS*, 18(2):109–119, 2012.

[31] Guangping Zhang, Dongsheng Li, Hansu Gu, Tun Lu, Li Shang, and Ning Gu. Simulating news recommendation ecosystem for fun and profit. *arXiv preprint arXiv:2305.14103*, 2023.

[32] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

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

2024年3月1日(金) 13:30 ~ 15:40 | T5-C オンライン（Zoom Events)

エンタメ・メディア (2)

座長:上田 真由美(流通科学大学)
コメンテータ:佃 洸摂(産業技術総合研究所)

13:30 ~ 13:35
イントロダクション

13:35 ~ 13:50
[T5-C-8-01] 生成AIによる小説挿絵の自動生成のためのアセスメント機構

*甲斐田 哲之佐[1,3]、牛尼 剛聡[2,3] (1. 長崎県立 口加高等学校、2. 九州大学 大学院芸術工学研究院、3. 九州大学 未来創生科学者育成プロジェクト)

13:50 ~ 14:05
[T5-C-8-02] ニュース記事とサッカー選手特徴を用いた地域的な嗜好の分析と可視化

*松葉 龍基[1]、王 元元[1] (1. 山口大学)

14:30 ~ 14:55
[T5-C-8-04] ライブ配信サイト Twitch における人気配信者の要因の分析

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

14:55 ~ 15:20
[T5-C-8-05] 実際のタワーディフェンスゲームのステージを模したシミュレータでのステージ攻略AI開発

*野村 大輔[1]、秋岡 明香[1] (1. 明治大学総合数理学部ネットワークデザイン学科　秋岡研究室)

生成 AI による小説挿絵の自動生成のためのアセスメント機構

甲斐田　哲之佐†§　　牛尼　剛聡‡§

†長崎県立 口加高等学校 〒859-2502 南島原市口之津町甲 3272
‡九州大学大学院芸術工学研究院 〒819-8540 福岡市南区塩原 4-9-1
§九州大学 未来創生科学者育成プロジェクト 〒819-0395 福岡市西区元岡 744
E-mail: †kaidanagasaki@gmail.com, ‡ushiama@design.kyushu-u.ac.jp

あらまし　挿絵は読書を魅力的にし、豊かな読書体験を構成するための重要な要素である。近年の生成 AI の飛躍的な発達により、挿絵がない文字だけの書籍に対して挿絵を自動的に生成する手法が提案されている。しかし、従来手法では、生成された画像が、対象とする書籍に適合していない場合が少なくない。個の問題を解決するために、本論文では、生成 AI を用いて、小説から挿絵を自動生成し、その挿絵の適切性を評価・修正するアセスメント機構を利用した手法を提案する。本手法では、ChatGPT を用いて小説の内容を分析し、場面分けを行い、画像生成 AI のためのプロンプトを時代や場所などの要素を含んで生成し、Stable Diffusion で画像を生成することで小説から挿絵を生成する。そして、生成された画像の妥当性を判断するために、GPT-4V を用いて生成した挿絵と小説を読み込み、小説の内容を要約する評価・修正を行うアセスメント機構により、小説の内容を理解して、生成した挿絵がどの場面に対応付けられるかを推定し、時代や場所などの観点から挿絵を生成するプロンプトを修正することで、小説の内容に適した挿絵の生成を実現する。

キーワード　挿絵, 生成 AI, ChatGPT, プロンプト, Stable Diffusion

1. 研究の背景

近年、高校生を始めとする若年層において本を読まない人の数（不読率）が増加している。平成 28 年度に行われた文部科学省の調査(図 1)によると平成 26 年度に比べ 28 年度では約 10%不読率が増加している。また、平成 30 年度の文化庁の全国の 16 歳以上の男女 3950 人に行った調査では、1 ヶ月中に本を読まないと回答したものが 47.3%存在していた。

しかし、2023 年にインプレス総合研究所が行った電子書籍ビジネス調査報告書によると、電子書籍の利用者は増加しており、電子書籍の売り上げも増加している。しかし、増加分のほとんどがコミックであり、小説などの文字ものの売り上げの増加割合は低くなっている。これらの調査から、文字だけで表現される文字ものよりも、絵と文字で表現されるコミックは魅力が高いと考えられる。そのため、文字ものの小説にイラストや画像などを入れることで、普段小説を読まない読者が小説を読むきっかけになる可能性がある。また、日頃から小説を読んでいる読者であっても、より小説の世界に没入しやすく、読書体験を向上できる可能性がある。

また、現在では生成 AI の能力が飛躍的に上昇している。例として、ChatGPT のように文書を読み込み、それに関する質問を行うと、その文書の内容に基づいて適切な回答を返すことができるようになった。また、Stable Diffusion を用いることで線画や猫、明るい雰囲気などの画像の内容や特徴を入力するだけで自分の考えている画像を誰でも手軽に高品質の画像を作ることができるようになった。

しかし、世界中には膨大な数の存在する書籍が存在し、それらのすべての小説に人手で挿絵を作成することはコストが大きく実現が難しい。その問題を解決するために、生成 AI を利用して自動的に挿絵を生成する研究が行われている。久永ら[1]は、Text2Image モデルと BERT を用いて小説から挿絵を生成する手法を提案している。この手法では、巨大言語モデルを利用して小説を解釈して、画像を生成する手法を提案している。しかし、生成した画像の中には、小説の内容を踏まえて適切とは判断できないものが少なからず存在してしまうという問題が存在している。

本論文では、不適切な挿絵を生成してしまう問題を解決するために、ChatGPT を用いて小説から挿絵を作り出し、さらにその挿絵を評価・改善するアセスメント機構を導入することで、より小説に適した小説を作る手法を提案する。

OpenAI 社提供している GPT-4V では、従来はテキストしか利用できなかった ChatGPT で、画像を扱うことができるようになった。また従来では困難だった長い文章を含むプロンプトも利用可能になった。そこで、本手法では、まず、小説本文を分析して、挿絵の対象とする場面を同定して、挿絵を生成するために適切なプロンプトを生成し、生成したプロンプトに基づいて挿絵の候補となる画像を生成する。次に、生成された挿絵候補の画像を GPT-4V を用いることで挿絵を分析し、対象とする場面の推測を行い、その場面と生成した挿絵を比較し、適切であるかどうかを分析（アセスメン

ト）を行い、不適切な部分を修正してより良い画像を生成するプロンプトを生成する。

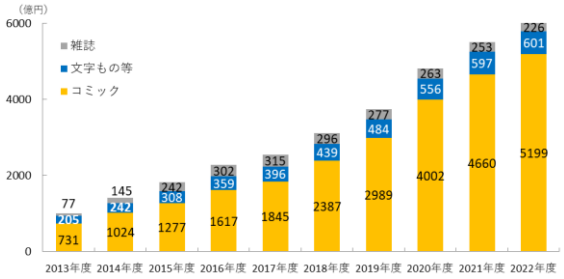


図 1: 電子書籍市場規模のジャンル別内訳 [2]

2. 提案手法

2.1 提案手法の概要

本節では提案手法について説明する。提案手法では、読書の魅力を高めるために場面を表した挿絵画像の生成を自動的に行うことを目的とする。また、挿絵画像の質を向上させ更に小説に適した挿絵画像を生成するために一度生成した挿絵画像を評価し、修正を行う。

2.2 小説本文からの画像生成

一般的に、小説から挿絵を自動的に生成する手法として、ChatGPT を用いて、小説本文から画像生成用のプロンプトを生成し、そのプロンプトを使って Stable Diffusion で画像を生成する手法が考えられる。しかし、ChatGPT に入力できるトークン数には限界があり、対象とする小説の本文をすべて入力することは困難である場合が多い。小説本文をすべて入力できない場合、本文の内容に合わせた画像を生成するためのプロンプトの内容を適切に生成することができない。そこで、本手法では、小説から画像を生成するための特別なプロセスに基づいた処理を行う。

本手法では、ChatGPT が長文を読み込むことができないという問題点を解決するために、まず最初に、小説本文を分割し、ベクターストアである Chroma に保存し、ChatGPT にベクターストアから文章を読み込ませることで文章全体を読み込ませる。



次に、ChatGPT に対して、入力した小説の内容に基づいて小説を均等に指定された数の場面に分けるようにプロンプトを利用して指示し、場面分けを行う。基本的に挿絵は一つの小説に対して複数存在するが、ChatGPT にプロンプトの生成を行わせても一つだけしか生成できないため、このような処理を行うことで、対象とする小説の場面分けを行うことで、小説に対して複数の挿絵を生成することができる。

次に、それぞれの場面に対して挿絵画像を生成するためのプロンプトを生成することを考える。Stable Diffusion 等の画像生成 AI で画像生成を行う場合、プロンプトの内容に描きたい画像の特徴を適切に指定することが重要である。小説の世界観に合わせるため舞台となっている場所や時代、そしてその場面を構成する要素を 7 つ以上含ませてプロンプトの生成を行う。以下に太宰治の小説「走れメロス」での生成例を示す。

```json
{
    "scene1": "メロスが激しい濁流と闘いながら峠を駆け下りる。",
    "scene2": "メロスが山賊に襲われるが、勇敢に戦い勝つ。",
    "scene3": "メロスが疲労困憊し、峠の草原で倒れる。",
    "scene4": "メロスが友人のセリヌンティウスを思い、感謝と悔いを抱える。",
    "prompt1": "A brave man fights against a raging torrent on a mountain pass.",
    "prompt2": "A man bravely defeats a group of bandits who attack him.",
    "prompt3": "Exhausted, the man collapses on a grassy meadow on the mountain pass.",
    "prompt4": "The man reflects on his actions and feels gratitude and regret towards his friend."
}
```

図 2: 場面分けとプロンプトの生成例

次に、上で生成した画像生成用のプロンプトに基づいて、Stable Diffusion を用いて挿絵の生成を行う。この際、ChatGPT で生成したプロンプトをそのまま用いて生成すると、生成した挿絵のスタイルが異なり、一貫性がない挿絵を出力してしまう。そこで、一貫性を持たせるために画像のスタイルを明示的に指定する。例として「走れメロス」の挿絵を白黒かつ線画で生成する生成した例を図 3 に示す。



図 2: 「走れメロス」に対して生成しされた挿絵画像

2.3 アセスメント機構

図 3 に示した生成例を見てもわかるように、これらの中には、小説「走れメロス」の内容に即していないと考えられる画像が生成されてしまっている。つまり、画像生成の際に与えるプロンプトには、修正、改善を行う必要がある。そこで、一度生成した画像に対して、それらが、より小説の内容に合った画像となるように、AI を利用して生成した画像を評価し、修正を行うアセスメント機構を導入する。

提案するアセスメント機構について述べる。本機構による修正は 2 段階で行う。

1 段階目の修正は、Stable Diffusion で挿絵画像を生成する前に、ChatGPT で生成した挿絵画像生成用プロンプトが、小説の内容に対して、適切かどうかを評価しするものである。そして、適切ではない内容がプロンプトに含まれていたら修正を行うというものである。これらの処理は ChatGPT を用いて行う。

2 段階目の修正では、1 段階目の修正が行われた挿絵画像生成用プロンプトを Stable Diffusion に入力して生成した挿絵画像を分析し、それに対して評価、修正を行う。ここでの挿絵画像の分析には GPT4-V を使用し、挿絵の画像データと小説のテキストデータの 2 つの入力を総合的に判断して、挿絵画像生成用プロンプトを修正する。

挿絵画像に対しての評価、修正の手法について具体的に説明する。まず GPT4-V に小説の文章全体を読み込ませ、その内容を要約することで内容の理解をさせる。次に GPT4-V に生成した挿絵画像を読み込み、その画像がどの場面に対して生成されたものなのかを小説の文章全体から推測する。図 3 に示した「走れメロス」で生成した画像に対して、それぞれの画像が表す場面を予測した例を以下の図 3 に示す。

場面1	場面2	場面3	場面4
この画像は、牧人のメロスが妹の結婚式の後で約束の時間内に王城に戻ろうと急ぐ途中で山賊に襲われるシーンから創られたものです。シーンは森の中で設定されています。メロスは山賊との戦闘に巻き込まれながらも、信義と勇気をもって困難に立ち向かい、最終的に約束の時間までに王城へと帰還します。	この画像は、小説「走れメロス」の中でメロスが荒れ狂う川を泳ぎ切って山の峠を駆け下りるシーンを表しています。小説の中では、メロスが誓いを守るために濁流を勇敢に泳ぎ切り、困難な自然の中を力強く進んでいく様子が描かれています。凸凹のある険しいランドスケープがその冒険の雰囲気をより引き立てており、メロスの決意と勇気を感じさせる場面です。	この画像は、疲労困憊し動けなくなったメロスが、灼熱の太陽の下で草原に倒れ込む小説の一場面を描いたものです。この場面は、メロスが妹の結婚式を行い、友人セリヌンティウスの助けるために、予定の時間までにシラクスの市に戻ろうとする途中で発生します。彼は体力の限界に達し、希望を失いかけたその・・・	この画像は、メロスが友人セリヌンティウスと平和な村で再会するシーンから作成されたものです。実際の小説の具体的な部分に該当するシーンは存在しませんが、メロスとセリヌンティウスの友情や絆を象徴するシーンとして描かれていると考えられます。画像は二人が互いに信頼と友情を確認する穏やかな時を・・・

図 3 ：小説画像の場面の推測

ぞれぞれの挿絵が表している小説中の場面を予測した後、GPT-4V を利用して、推測した場面の挿絵画像において時代や場所、風景などの要素が適しているのかを評価し、1 段階目で利用した挿絵画像生成ようプロンプトに対して修正を行う。

場面1	場面2	場面3	場面4
A man named Melos defends himself against bandits in a mountain pass on his way back to the royal castle after attending his sister's wedding.	Melos overcomes a swollen river at dawn and descends a steep ravine amidst the rugged wilderness."	Melos, exhausted and on the verge of collapse, seeks solace under a leafy tree on a lush meadow, as the scorching sun gives way to a gentle, golden dusk.	Melos heartily welcomes his friend Serinuntius back to the village after the resolution of their ordeal, set against the backdrop of a rustic hamlet in the warm light of early evening.

図 4 ：修正を行ったプロンプトの例

最後に、修正したプロンプトをもとに Stable Diffusion を用いて、同じ画風を指定して、再度挿絵を生成する。図 3 に示した挿絵画像に関して、アセスメント機構を利用して修正を行った結果生成された画像の例を図 4 に示す。



図 4 ：「走れメロス」で生成した挿絵の修正後

3. まとめ

本論文では生成 AI を利用して、小説のテキストのみから適切な挿絵を生成するために、生成した画像の自動評価、修正を行うアセスメント機構を有する小説の挿絵生成手法を提案した。提案手法では、ChatGPT によって小説を均等に 4 つの場面に分け、時代や場所、その情景に含まれる要素を 7 個以上推測し、その内容を含んだ Stable Diffusion 用のプロンプトを生成する。そして、そのプロンプトの内容に加え、絵のスタイルを白黒かつ線画に統一して一般の挿絵と似た絵を生成する。そして、アセスメント機構により、生成した絵が小説の内容に適しているかを評価し、プロンプトの内容で適していない部分を修正、不足した部分を補うこともできた。最後に修正したプロンプトを用いて、

挿絵をもう一度生成する。低藩主法では、従来手法より、小説の内容に適した挿絵を生成することができるようになった。

しかし、実際にこの挿絵を用いて評価実験を行うことができていないため、現時点では、客観的に有用性を評価することができていない。また、挿絵を生成することはできているが、小説のどこに挿絵を入れるかを指定できていないことが今後の課題である。

参 考 文 献

[1] 久永 龍星, 牛尼 剛聡：プロンプト・プログラミングによる小説挿絵の自動生成, DEIM2023 論文集、2023
[2] 文化庁：平成 30 年度「国語に関する世論調査」の結果について (2018).
[3] インプレス総合研究所：電子書籍ビジネス調査報告書 2023(2023).
[4] OpenAI Documentation
https://platform.openai.com
[5] Chroma
https://docs.trychroma.com/

ニュース記事とサッカー選手特徴を用いた地域的な嗜好の分析と可視化

松葉　龍基† 　王　元元†

† 山口大学工学部知能情報工学科　〒 755–8611 山口県宇部市常盤台 2-16-1
E-mail: †{a072ffu, y.wang}@yamaguchi-u.ac.jp

あらまし　近年，多くのメディアでサッカーの試合が生中継され，国内リーグへの興味も高まり，サッカーへの関心を維持・向上させるためには，個々の嗜好に合致した選手の存在が重要である．しかし，すべての選手情報を網羅的に把握し，自分の嗜好に合った選手を見つけることは困難である．そこで本研究では，地域の特徴と選手の特徴に着目し，ニュース記事からそれぞれの特徴語を抽出し，その後，特徴語間の類似度を算出する．この結果を基に，地域との類似度が高い選手を特定し，その選手のアイコンおよび詳細情報を地図上に可視化する手法を提案する．これにより，個人は地域嗜好を考慮に入れた選手の推薦が可能となり，自治体や企業は地域に影響力を持つ選手を認識できる．本論文では，地方の新聞サイトに掲載された地域ニュースと選手名を用いて Google ニュースでの検索結果を収集し，各地域と各選手の特徴語を抽出する．そして，これらの特徴語間の類似度を算出することで，各地域の嗜好に合致したサッカー選手を地図上に可視化する．最後に，サッカー選手の情報を活用した地図上の可視化インタフェースを構築し，被験者による地域別嗜好選手の可視化の有用性を評価する．

キーワード　ニュース記事，サッカー，地域嗜好，選手特徴，可視化

1 はじめに

近年，サッカー観戦は実際に試合会場に足を運んで観戦する人だけでなく，地上波放送やインターネット上の動画配信サービス SPOTV NOW[1]や YouTube[2]などの動画配信サービスを利用することで，サッカー観戦をメディアを通じてリアルタイムやオンデマンドで楽しむ人も多くいる．また，日本人選手の活躍なども相まって，その観戦対象は日本国内だけでなく海外にも及んでおり，現在サッカー観戦は老若男女、性別を問わず娯楽として親しまれている．

このように幅広い層から支持を受けているサッカー観戦であるが，その興味や関心が世間の風潮から始まることも珍しくない．たとえば，サッカー日本代表の試合があると，ニュースメディアや SNS などで情報が拡散され，これによって試合があることを知り，試合を観るという流れが形成される．しかし，この場合のサッカーへの興味は一時的なものに過ぎないという問題点が存在する．確かに，一時的な興味や関心であってもサッカー業界は盛り上がるが，この興味と関心を維持させたり，向上させることができればより一層サッカー業界に良い影響となる．そのために，本研究では漠然としたサッカーへの興味や関心を特定の選手に向けることによって，ある特定の選手を観たい，その選手の所属しているクラブチームの試合を観たい，その選手のプライベートを知りたいとった一時的な興味や関心が長期的なものへと変化すると考える．

一方で，選手に興味や関心を向けるためにはいくつかの問題が存在する．一つは選手数が多く，どのような選手が存在しているのか把握するのが難しいという問題である．選手の情報が多すぎるため，全ての選手情報を網羅的に調べるということは困難である．また，選手への好みは基本的に個人の嗜好に影響されるという問題もある．そのため，興味や関心をもった選手が，個々の嗜好に合わない，期待はずれとなるという状況に陥ることも少なくない．しかし，嗜好は人によって完全に異なるというものでもなく，地域的な嗜好も存在する．いわゆる「県民性」というもので，各都道府県ごとに特徴付けることができる傾向や考え方，文化や価値観の差異などを指すものである．

そこで本研究では，多くのサッカー選手の中から地域的な嗜好を利用して個々の嗜好に合致した選手を発見する．各地方のニュース記事と選手名検索を用いた Google ニュース記事を利用し，文書中の単語の TF-IDF 値の算出を行う．その結果から TF-IDF 値が上位である単語を地域の特徴語および選手の特徴語として，各地域の特徴語と各選手の特徴語間で類似度を計算し，類似度の高い選手を特定する．最後に嗜好に合致した選手を地図上に可視化するとともに，その選手情報を掲載する可視化インタフェースを構築する．これにより，地域嗜好を反映した選手が一目で分かるので，個人では地域的な嗜好を考慮に入れた選手の推薦が可能となり，自治体や企業は地域に影響力のある選手を認識することが可能になる．これによりサッカー業界の発展に貢献できると考える．

本論文の構成は以下のとおりである．2 章では，ニュース記事や SNS データの分析などの関連研究を紹介し，本研究の位置付けを示す．3 章では，ニュース記事から地域とサッカー選手の特徴語を抽出する方法について説明する．4 章では，地域別の嗜好選手の特定と可視化について説明する．5 章では，構築した選手可視化インタフェースの評価実験について述べる．最後に，6 章では本研究のまとめと今後の課題について述べる．

1：https://www.spotvnow.jp
2：https://www.youtube.com

2 関連研究

近年，ニュース記事や SNS データを基にした表現の抽出や分散表現の活用，情報推薦や可視化に関するさまざまな研究・システムの開発が盛んである．

本研究と同様にサッカー業界とニュース記事を組み合わせた研究として，加藤ら [1] はサッカー選手の移籍を題材にした記事の報道元と報道時期の予測精度についての研究を行っている．また，西山と安藤 [2] は文間類似度の精度向上として，名詞・動詞・形容詞といった文に含まれる単語の実質的な内容を含む語が重要であることを示した．本研究では，ニュース記事から特徴語として一般名詞，形容詞，副詞を用いる．

さらに，分散表現を利用した研究として，小林と松下 [3] はキャラクタの構成要素の性格と外見に着目し，分散表現を利用してこれらの要素の加減算を行い，キャラクタを検索するシステムを提案している．高野ら [4] は音楽ファンの Twitter プロフィール文を用いてコーパスを作成し，音楽アーティストの類似性について分析・評価を行っている．そして，宮本と北山 [5] は書籍レビューから書籍特徴ベクトルを作成し，推薦候補書籍との類似度を算出することで書籍推薦を行うシステムを提案している．本研究でも，特徴語のベクトルから類似度を算出し，地域とサッカー選手の紐付けを行う．

ここで，地理情報と単語の紐付けの研究として，陰山と乾 [6] はカテゴリごとにエンティティの特徴を抽出し，ジオロケーションを特定し，地域と単語の対応付けの研究を行っている．本研究では，地域から特徴語を抽出することで紐付けとしている．本研究でも行う個人の嗜好に合致するものの推薦として，河合ら [7] は個人の選好に基づいた記事の閲覧システムを提案している．関と安藤 [8] は小学生が新聞を用いた学習に利用しやすい新聞記事の推薦をするシステムを提案している．本研究では，地域嗜好に基づいて選手を推薦する．

地図上への可視化として，Wang [9] は方言ツイートの投稿場所の可視化，藤崎ら [10] はニュース記事の言及地域の関連度からニュース記事をマッピングするシステムを提案している．本研究では各地域に最も嗜好が合致している選手のマッピングを行う．

3 ニュース記事の特徴語抽出

本章では，ニュース記事から地域とサッカー選手の特徴語の抽出ついて説明する．使用したデータについて，地方の特徴語抽出には中国地方の 5 県（山口県・広島県・岡山県・島根県・鳥取県）を対象として，山口県の特徴語には山口新聞電子版 [11]，広島県の特徴語には中国新聞デジタル [12]，岡山県の特徴語には山陽新聞デジタル [13]，鳥取・島根県の特徴語には山陰中央新報デジタル [14]，これらの Web サイトの地域ニュースの欄から収集可能なニュース記事直近 50 件（収集期間は 2023 年 11 月 20 日から 12 月 20 日まで）を対象としてその本文を利用する．また，サッカー選手の特徴語抽出には，12 月時点で dmenu スポーツの注目選手 [15] に掲載されている 50 名の選手を対象として，これらの選手名を用いた Google ニュース [16] での検索結果上位 10 件のニュース記事本文を利用する．

3.1 地域の特徴語抽出

地域の特徴語抽出は次の通りに行う．各県のニュース記事 50 件の文書を 1 つの文書にまとめ，形態素解析を行う．そして各文書中に出現した単語について TF 値，IDF 値，TF-IDF 値を算出し，TF-IDF 値が上位の単語をその地域の特徴語とする．

形態素解析を行う際には，形態素解析エンジン Mecab[3] および辞書 IPAdic[4] を用いた．形態素解析エンジン Mecab は辞書に基づいて文を単語単位に分かち書きを行う．そして分割された単語の品詞や活用形，単語の読み方などを得ることができる．この Mecab を用いて，まとめられたニュース記事の形態素解析を行い，品詞が一般名詞・形容詞・副詞である単語を対象にその TF 値，IDF 値，TF-IDF 値を求める．

TF (Term Frequency) 値は文書内でのある単語の出現頻度である．すなわち，文書内のすべての単語の出現回数のうち，その単語の出現回数が占める割合を表す．次の式 (1) から TF 値は求められ，出現回数が多いほど TF 値は高くなり，出現回数が少ないほど TF 値は低くなる．

$$\mathrm{TF} = \frac{\text{あるニュース記事におけるある単語の出現回数}}{\text{あるニュース記事内の総単語数}} \quad (1)$$

IDF (Inverse Document Frequency) 値は，文書集合の中のある単語が含まれる文書の割合の逆数を表す．次の式 (2) から IDF 値を求めることができ，単語がほかの文書に多く出現しているほど IDF 値は低くなり，単語がほかの文書にあまり出現していないほど IDF 値は高くなる．なお，ゼロ除算を防ぐため分母に 1 を加えている．

$$\mathrm{IDF} = \log \frac{\text{総ニュース記事数}}{\text{ある単語が出現するニュース記事数} + 1} \quad (2)$$

式 (1) と式 (2) で求めた TF 値，IDF 値を用いて，TF-IDF 値はその積で求められる．これにより，ある文書内での出現回数が多いがほかの文書には出現していない単語の TF-IDF 値は高くなり，それ以外の単語については TF-IDF 値は相対的に低くなる．この TF-IDF 値を算出し，TF-IDF 値の高い単語を地域の特徴語とする．結果の例として，広島県の TF-IDF 値が上位 10 位である単語とその TF-IDF 値を表 1 に示す．このようにして 5 県に対して特徴語を各 10 単語特定する．ただし，4 章で用いる学習済み Word2Vec モデルに登録されていない単語は無視し，無視した単語より下位の単語を繰り上げて特徴語とした．

3.2 サッカー選手の特徴語抽出

サッカー選手の特徴語抽出は，地域の特徴語抽出と同様に各選手のニュース記事 10 件の文書を 1 つの文書にまとめ，形態素解析を行う．各文書中に出現した単語について全文書との間で TF-IDF 値を算出し，TF-IDF 値上位の単語をその選手の特徴語とする．結果の例として，藤井陽也選手の TF-IDF 値が上

3 : https://taku910.github.io/mecab/
4 : https://ja.osdn.net/projects/ipadic/

表 1　広島県の TF-IDF 値が高い単語（上位 10 単語）

順位	広島県の特徴語	TF-IDF 値
1	東部	0.00514
2	市場	0.00430
3	お好み焼き	0.00386
4	ホーム	0.00386
5	JR	0.00386
6	サミット	0.00386
7	バラ	0.00386
8	餅	0.00386
9	先進	0.00386
10	公印	0.00257

位 10 位である単語とその TF-IDF 値を表 2 に示す．このようにして選手 50 名に対して特徴語を各 10 単語特定する．ただし，4 章で用いる学習済み Word2Vec モデルに登録されていない単語および選手名，クラブチーム名は無視し，無視した単語より下位の単語を繰り上げて特徴語とした．

表 2　藤井陽也選手の TF-IDF 値が高い単語（上位 10 単語）

順位	藤井陽也選手の特徴語	TF-IDF 値
1	店舗	0.03524
2	ユーロ	0.02765
3	パン	0.02765
4	お客	0.02054
5	列	0.01762
6	商品	0.01762
7	神社	0.01762
8	長蛇	0.01762
9	打点	0.01762
10	センター	0.01538

4　地域的な嗜好分析と可視化

本章では，抽出された特徴語から地域的な嗜好に合致した選手の特定とその選手を地図上で可視化する手法について述べる．

4.1　地域別嗜好選手の特定

地域別嗜好選手の特定は，地域とサッカー選手それぞれのニュース記事から抽出された特徴語を用いて，特徴語間で類似度を計算し，平均類似度を算出する．各地域の特徴語に対して平均類似度が最も高い特徴語を持つ選手をその地域的な嗜好に合致した選手として特定する．

本研究では，特徴語間での類似度を計算する際に，東北大学 乾・岡崎研究室の日本語 Wikipedia エンティティベクトル [21] を使用した学習済み Word2Vec モデルを利用した．Word2Vec [18] [19] は，文章中の単語を数値ベクトルに変換してその意味を把握する自然言語処理の手法である．また，日本語 Wikipedia エンティティベクトルとは，日本語版 Wikipedia の本文全文から学習した単語、および Wikipedia で記事となっているエンティティの分散表現ベクトルである．この Word2Vec と 3.1 節で特定した地域の特徴語および 3.2 節で特定したサッカー選手の特徴語のコサイン類似度を計算する．コサイン類似度は，次の式（3）で計算される．

$$cos(a,b) = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\|\|\vec{b}\|} \tag{3}$$

実際に表 1 中の 3 位の特徴語「お好み焼き」と表 2 中の特徴語でコサイン類似度を計算する例を表 3 に示す．

表 3　広島県の特徴語「お好み焼き」と藤井陽也選手の特徴語のコサイン類似度

順位	類似度を計算する単語	コサイン類似度
1	(お好み焼き, パン)	0.532
2	(お好み焼き, お客)	0.422
3	(お好み焼き, 店舗)	0.375
4	(お好み焼き, 商品)	0.372
5	(お好み焼き, 神社)	0.227
6	(お好み焼き, 長蛇)	0.139
7	(お好み焼き, センター)	0.092
8	(お好み焼き, ユーロ)	0.079
9	(お好み焼き, 列)	0.050
10	(お好み焼き, 打点)	0.034

表 3 のように，各地域の特徴語 10 単語と各選手の特徴語 10 単語の類似度を計算し，選手別に平均類似度を算出する．10 単語同士の計算となるため，100 回のコサイン類似度計算を行い，その和を 100 で割ることで平均類似度を算出する．最後に，各地域で平均類似度が最も高い選手をその地域の嗜好に合致した選手として選出する．各地域とその地域の嗜好に合致した選手，平均類似度を表 4 に示す．

表 4　各地域とその地域の嗜好に合致した選手および平均類似度

地域	地域嗜好に合致した選手	平均類似度
山口県	脇坂泰斗	0.44
広島県	藤井陽也	0.46
岡山県	森下龍矢	0.36
島根県	川崎颯太	0.30
鳥取県	山本悠樹	0.28

4.2　サッカー選手の可視化

本研究のインタフェースは，OpenStreetMap に Leaflet を利用して各地域に，前章で特定したその地域の嗜好に合致した選手のアイコンを表示し，ポップアップでその選手の情報を掲載し，加えてその選手の特徴語から生成したワードクラウドを表示する選手可視化インタフェースを実装する．選手アイコンの画像は J リーグ公式サイトの選手名鑑 [22] から，選手情報は選手の Wikipedia ページ [23] [24] [25] [26] [27] から利用する．ワードクラウドは各選手の特徴語の TF-IDF 値が大きい特徴語を大きく表示し，TF-IDF 値が小さいと文字が小さく表示される．このとき，各選手の TF-IDF 値が上位 40 位に含まれる特徴語を用いた．可視化インタフェースを図 1，図 2 および図 3 に示す．



図 1　地域嗜好に合致した選手の可視化



図 2　地域嗜好に合致した選手の可視化と選手情報



図 3　地域嗜好に合致した選手の可視化と選手のワードクラウド

5　評価実験

5.1　実験概要

生成した選手可視化インタフェースの有用性を評価するために SUS (System Usability Scale) [28] を用いる．SUS では，被験者に実際に生成した選手可視化インタフェースを閲覧してもらい，表 5 の設問について 5 段階で評価するアンケートを実施し，スコアを算出する．スコアは奇数項目は回答番号から 1 を引き，偶数項目は 5 から回答番号を引いた後，すべての項目を 0 から 4 で評価し、各項目の合計数値を 2.5 倍して最小値 0，最大値 100 のスケールへ変換することでスコアを求める．算出されたスコアは次の表 6 の通りにグレード分類される．

表 5　SUS の設問

番号	設問
Q1	このインタフェースをしばしば利用したいと思う
Q2	このインタフェースを利用するには説明が必要となるほど複雑であると感じた
Q3	このインタフェースは容易に使いこなすことができると思った
Q4	このインタフェースを利用するのに専門家のサポートが必要だと感じる
Q5	このインタフェースにあるコンテンツやナビゲーションは十分に統一感があると感じた
Q6	このインタフェースでは一貫性のないところが多々あったと感じた
Q7	たいていの人は、このインタフェースの利用方法をすぐに理解すると思う
Q8	このインタフェースはとても操作しづらいと感じた
Q9	このインタフェースを利用できる自信がある
Q10	このインタフェースを利用し始める前に知っておくべきことが多くあると思う

表 6　SUS スコアのグレード

SUS スコア	グレード	評価
>80.3	A	Excellent
68 – 80.3	B	Good
68	C	Okay
51 – 68	D	Poor
<51	E	Awful

5.2　実験結果と考察

被験者を 10 人としたアンケート評価を実施した．表 5 の各設問に対する平均スコアと SUS スコアを表 7 に示す．

表 7　各設問の評価と SUS score

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS スコア
70	75	92.5	82.5	77.5	80	90	87.5	92.5	60	80.75

評価実験の結果，選手可視化インタフェースの SUS スコアは 80.75 となった．表 6 より選手可視化インタフェースの評価は A (Excellent) である．また，各設問の評価に関して，Q10 での評価が低い結果になった．Q10 は，このインタフェースの概要説明や利用の方法についての情報が不足していたため，評価が低くなったと考えられる．被験者の意見として，操作説明のページが欲しいといった意見もあったので，修正の余地があると考えられる．Q10 以外の設問では高い評価を得たため，全体としてはシステムユーザビリティの高いインタフェースを実現できたと考えられる．今後の研究において，選手可視化インタフェースの機能追加などが考えられるが，システムユーザビリティの維持向上に努める．

6　おわりに

本研究では，地方の新聞サイトに掲載された地域ニュースと選手名を用いた Google ニュースでの検索結果を収集し，各地域と各選手の特徴語を抽出した．そして，これらの特徴語間の

類似度を算出することで，各地域の嗜好に合致したサッカー選手を地図上に可視化するインタフェースを構築し，被験者による地域別嗜好選手の可視化の有用性を評価した．

今後の課題として，対象地域や対象とするサッカー選手の範囲を拡大し，インタフェースの操作性を改善していく予定である．

文　献

[1] 加藤 郁之, 吉川 正俊, 加藤 誠: 報道元と報道時期を利用した Web 記事による将来予測. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2019), B2-1, 2019 年.
[2] 西山 朋貴, 安藤 一秋: 子供向け日本語換言コーパス構築のための文間類似度計算法の評価. 人工知能学会全国大会論文集, Vol.4J3-GS-6f-04, 2021 年.
[3] 小林 達哉, 松下 光範: 性格要素と外見要素の加減算による類似キャラクタの検索. 第 16 回 WI2 研究会論文集, No. 16, WI2-2020-22, 2020 年.
[4] 高野 翔吾, 北 研二, 吉田 稔, 松本 和幸: 文書分散表現を用いた音楽アーティストの特徴分類に関する研究. 第 82 回全国大会講演論文集, pp. 477-478, 2020 年 2 月 20 日.
[5] 宮本 達矢, 北山 大輔: 分散表現を用いたユーザの書籍整列順序に基づく書籍推薦手法. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2019), G4-3, 2019 年.
[6] 陰山 宗一, 乾 孝司: 位置属性を有しない事物に対する地理的特定性の分析. 言語処理学会第 29 回年次大会，2023 年 3 月.
[7] 河合 由起子, 官上 大輔, 田中 克己: 個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム. 情報処理学会論文誌データベース（TOD26）, Vol. 46, No. SIG 8, pp. 14-25, 2005 年 6 月 15 日.
[8] 関 伸也, 安藤 一秋: 小学校での地域学習 NIE に適した Web ニュース記事推薦システム. 第 82 回全国大会講演論文集, Vol. 2020, No. 1, pp. 361-362, 2020 年 2 月 20 日.
[9] Yuanyuan Wang: Visualizing Dialectal Variation in Japan through Geo-tagged Tweets. in Proceedings of 2019 IEEE 5th International Conference on Computer and Communications (ICCC 2019), pp. 1317-1321, December 2019.
[10] 藤崎 日和, 荘司 慶行, 田中 克己: ニュース記事の地域度分析に基づく地図への自動分類. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2023), 3a-6-3, 2023 年.
[11] 山口新聞電子版: https://yama.minato-yamaguchi.co.jp/ (参照: 2023 年 12 月 22 日).
[12] 中国新聞デジタル: https://www.chugoku-np.co.jp/ (参照: 2023 年 12 月 21 日).
[13] 山陽新聞デジタル: https://www.sanyonews.jp/ (参照: 2023 年 12 月 20 日).
[14] 山陰中央新報: https://www.sanin-chuo.co.jp/ (参照: 2023 年 12 月 20 日).
[15] dmenu スポーツ注目選手: https://soccer.sports.smt.docomo.ne.jp/jleague/player/ (参照: 2023 年 12 月 15 日).
[16] Google ニュース: https://news.google.com/.
[17] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis. in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 230-237, July 2004.
[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean: Distributed Representations of Words and Phrases and their Compositionality. in Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013), Vol. 2, pp. 3111–3119, December 2013.
[19] Yoav Goldberg, and Omer Levy: word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. ArXiv, Vol. abs/1402.3722, 2014.
[20] 鈴木 正敏, 松田 耕史, 関 根聡, 岡崎 直観, 乾 健太郎: Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会, 2016 年 3 月.
[21] 日本語 Wikipedia エンティティベクトル: https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/.
[22] J リーグ選手名鑑 2023: https://www.jleague.jp/player/ (参照: 2024 年 1 月 5 日)
[23] 脇坂 泰斗: https://ja.wikipedia.org/wiki/%E8%84%87%E5%9D%82%E6%B3%B0%E6%96%97 (最終更新: 2023 年 12 月 10 日).
[24] 藤井 陽也: https://ja.wikipedia.org/wiki/%E8%97%A4%E4%BA%95%E9%99%BD%E4%B9%9F (最終更新: 2024 年 1 月 3 日 18:04).
[25] 森下 龍矢: https://ja.wikipedia.org/wiki/%E6%A3%AE%E4%B8%8B%E9%BE%8D%E7%9F%A2 (最終更新: 2024 年 1 月 1 日 21:07).
[26] 川崎 颯太: https://ja.wikipedia.org/wiki/%E5%B7%9D%EF%A8%91%E9%A2%AF%E5%A4%AA (最終更新: 2023 年 11 月 22 日 11:55).
[27] 山本 悠樹: https://ja.wikipedia.org/wiki/%E5%B1%B1%E6%9C%AC%E6%82%A0%E6%A8%B9 (最終更新: 2023 年 12 月 26 日 13:32).
[28] John Brooke: SUS: A Quick and Dirty Usability Scale. Usability Eval. Ind., Vol. 189, 1995.

ライブ配信サイト Twitch における人気配信者の要因の分析

佐藤　大樹† 　清　雄一† 　田原　康之† 　大須賀昭彦†

† 電気通信大学　〒 182–8585　東京都調布市調布ケ丘一丁目 5 番地 1
E-mail: †sato.hiroki@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　ゲームのライブ配信は今日急速に成長しているエンターテインメント分野である. Twitch はライブ配信の市場シェアの大部分を占めており, ゲーム特化型配信サイトとして, ゲーム大会の公式配信やプロゲーマーの配信などに利用されている. しかし, ゲーム配信の人気に伴い, 配信者の増加やゲームタイトルの多様化が起こり, 人気配信者の共通特徴が不明瞭になってきている. 本研究の目的は, ライブ配信サイト Twitch における配信ごとの視聴者数を用いて, どのようなゲーム, あるいはどのような配信者が人気を獲得しているのか要因を分析する. 配信者が分析結果を参考にすることで, 視聴者数の増加, そして配信者の利益増加に繋がり, また, 人気ゲームの特徴は, ゲーム開発者とってゲームの改善や新しい戦略を立てるための有益な情報となる.

キーワード　ソーシャルメディア, ゲーム・娯楽

1 はじめに

ゲームのライブ配信は今日急速に成長しているエンターテインメント分野である. ゲームのライブ配信は, 配信者がゲームをプレイしながら, 同時にその様子をリアルタイムでインターネットを通じて配信することを指す. 視聴者は配信者のプレイを見たり, コメントを通して配信者とコミュニケーションを取ったりして, その配信やゲームを楽しむことができる.

Twitch [1] はライブ配信の市場シェアの大部分を占めており, ゲーム特化型配信サイトとして, ゲーム大会の公式配信やプロゲーマーの配信などに利用されている.

ニュース記事 [2] によると, 世界中で毎月 700 万人が配信をしており, 1 日のサイト訪問者数は平均 3500 万人にのぼる. また, 2022 年の視聴時間数は約 1 兆 3000 億分, 配信言語は 35 か国語と巨大な市場規模であることがわかる [2]. その中でも日本は最も成長している地域の 1 つで, 視聴者数は 22 年に前年比で 51 %上昇, 配信時間は前年比で 50 %上昇, 配信者数も 63 %上昇となっている [2].

数年前はゲームのライブ配信が今よりも一般的ではなく, 配信技術やコミュニティが未発達で, コンテンツや個性も幅広くはなかった. しかし, 現在はゲーム配信の人気が高まり, 配信者の急激な増加やゲームタイトルの多様化によってそれらが発達・幅広くなってきている. 一方, 複雑化している中で人気配信者の共通特徴というのはゲームが上手いや話が面白いなどと一言で表すことができない状況にある.

2 研究目的

本研究の目的は, ライブ配信サイト Twitch における配信ごとの視聴者数を用いて, どのようなゲーム, あるいはどのような配信者が人気を獲得しているのか要因を分析する. 配信者が分析結果を参考にすることで, 視聴者数の増加, そして配信者の利益増加に繋がり, また, 人気ゲームの特徴は, ゲーム開発者とってゲームの改善や新しい戦略を立てるための有益な情報となる.

3 関連研究

3.1 関連研究

Andrew Dybka ら [3] は, Twitch 配信者の成長を促進するための意思決定を支援する分析を行った. Twitch の統計および分析サービスの sullygnome.com [4] から 4 か月間の配信者, 配信, ゲームのデータを収集し, 獲得フォロワー数を目的変数として予測を行った. 結果は,「配信時間」,「平均視聴者数」,「フォロワー数」が重要度の高い特徴量であったこと, 全体平均による欠損値補完が有効でない点などが挙げられた.

3.2 関連研究の問題点

3.2.1 使用したデータについて

データ取得に使用したサイトには, スクレイピング禁止の記述がある (2024 年 2 月 7 日現在). そのため, 獲得フォロワー数を含むすべてのデータは 1 つ 1 つ手動で csv 形式のファイルをダウンロードしなければならず, 多くの時間と労力を要する.

本研究ではスクレイピング禁止サイトを使用しないで取得できる配信者の成長指標として, Twitch API から取得できる配信の視聴者数を用いる. これは無料の Twitch アカウントをもっていれば誰でも利用可能である. また, ゲーム特化型配信サイトであるため, ゲームデータの重要度は高い. このデータは IGDB API から取得する. IGDB は Twitch のゲームカテゴリを集約したデータベースで, Twitch で利用可能なゲームカテゴリは常にこの IGDB 上にあるデータが参照されている.

3.2.2 重要度の高い特徴量について

重要度の高い特徴量として,「平均視聴者数」,「フォロワー数」を挙げている. しかし, 両方が増えると獲得フォロワー数が増えることはごく自然であり, また, 配信者自身では制御できない要素である. 本研究では, このようなリークとなる情報や配信者自身では制御不可能な要素を時系列や相関関係を基に削除

する.

3.2.3 人気向上の要因について

関連研究には, 人気向上の要因について具体的な言及がされていない. したがって, 人気がある配信者の特徴が不明なままである. 本研究では, 説明可能な AI 手法を使用する. 説明可能な AI 手法を適用することで, 解釈性や透明性, 信頼性を加え, 意思決定プロセスを配信者やゲーム開発者が理解できるようにする.

4 アプローチ

視聴者数を目的変数とした高精度な予測モデルを作成し, このモデルに対して説明可能な AI 手法を適用することで, モデルがどの特徴量が視聴者数増加に寄与しているのかを把握することができる.

4.1 視聴者数予測

4.1.1 データセット

初めに, Twitch API から言語が日本語という条件で 5 分毎に配信中の配信データを 2023 年 11 月 1 日から 1 か月間取得した. 5 分毎に取得するため, 同じ配信で, 視聴者数の異なるデータが複数得られる. この場合は視聴者数の一番多いデータのみを残した. 続いて, Twitch API のゲーム ID 検索でゲームごとの IGDB ID を取得し, IGDB API でそのゲームの詳細情報を取得する. 雑談や旅行配信などゲームとは関係ない配信は除外した. 以上のデータを結合し, 配信情報とゲーム情報を持つデータセットを作成した. この時点で配信に関する特徴量は 15 個, ゲームに関する特徴量は 52 個であった.

4.1.2 前 処 理

初めに, 配信開始時間やゲーム発売日など時間情報はすべて UNIX 時間に変換し, 一部のデータに対して One-Hot エンコーディングの適用や要素数変換を行った. また, 配信時間が短い配信は外部のノイズやランダムなイベントがデータに影響を与えやすく, 配信者やゲームが与える影響はかなり小さくなり, 本来の傾向やパターンがわかりにくくなる. 一定以上の配信時間を確保することで, ノイズの影響を軽減し, より信頼性の高いデータを得ることができるため, 本研究では 1 時間以上配信している配信に絞った. また, 配信者自身が制御不可能な配信 ID や配信者 ID などの特徴量は削除した.

この時点でデータ数は約 78 万件であったが, 視聴者数が 10 人未満の配信が 60%以上を占めるなど, データの偏りが生じた. データの偏り解消や計算効率の向上のために視聴者数が 10 人未満のデータを削除し, 100 人未満のデータに関しては 2 万件にランダムサンプリングを行った. 最終的に 112 個の特徴量を持つ約 4 万件のデータセットが完成した.

4.1.3 視聴者数予測

視聴者数を目的変数とし, その他を説明変数として予測を行う. 評価指標には平均絶対誤差（MAE）を使用するが, 視聴者数が大きいものに強く影響されてしまうため, 目的変数に対して 1 を足したのちに対数変換を行い, 影響を抑える. モデルには XGBoost [5] を使い, ハイパーパラメータの最適化には Optuna [6] のベイズ最適法を用いた. 図 1 に視聴者数予測のアプローチを示す.

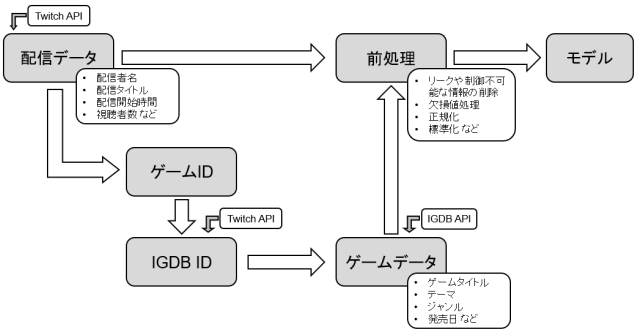


図 1 視聴者数予測のアプローチ

4.2 要 因 分 析

説明可能な AI 手法によって, モデルがどの特徴量の影響で視聴者数の多寡を判断しているのかを明らかにする. 説明可能な AI 手法は機械学習のブラックボックス化を解消する手法として注目されており, A. Adadi [7] らの研究によると, 説明可能な AI 手法の関連論文が 2016 年以降に急激に増加していることが報告されている. 手法の代表例として, SHAP [8] や LIME [9], Grad-CAM [10] が挙げられる. Grad-CAM は画像認識タスクに限定される. LIME は「個々の予測結果の判断理由を理解すること」を目的とした局所的解釈を与えるのに対し, SHAP は局所的解釈だけでなく, 「モデルの全体的な振る舞いを理解すること」を目的とした大域的解釈も与えることができる. そういった理由から, 本研究では SHAP を用いて要因分析を行った.

4.2.1 SHAP

SHAP（SHapley Additive exPlanations）は 2017 年に提案された機械学習モデルの予測結果を説明するための Python ベースの統合アプローチである. 協力ゲーム理論における個々のプレイヤーの寄与を算出する Shapley 値を用いて予測モデルを近似して解釈可能なモデルを構築する. 近似したモデルは以下の 3 つの性質を持つ.

性質 1. 局所的な精度：元のモデルと近似したモデルの予測値が一致していること

性質 2. 欠落性：存在しない特徴量は影響しない

性質 3. 一貫性：任意の特徴量がモデルに与える影響が大きくなるようにモデルを変更した際に, その特徴量の寄与度が減少しない

一般に Shapley 値を正確に算出することは難しいが, SHAP の計算アルゴリズムを適用することで近似することができる. 算出された値を SHAP 値と言う. 図 2 は SHAP の概要図である. SHAP を用いた研究には, M. Kuzlu らの研究 [11] や T. Chen らの研究 [12], Julio C. S. Reis らの研究 [13] がある.

5 実験結果と考察

5.1 視聴者数予測

38,433 行の配信データを使用し, 表 1 には対数変換した視聴

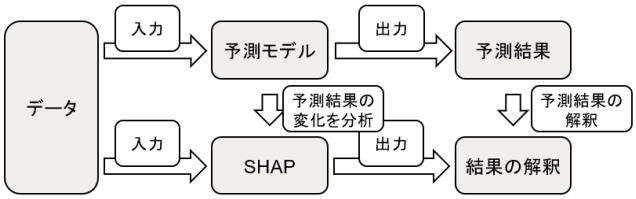


図 2　SHAP の概要図

者数の平均値などの要約情報を示す.

図 3 は横軸が実測値, 縦軸が予測値の二値比較の結果, 表 2 には決定係数と平均絶対誤差の結果を示す. 黒線上では実測値と予測値は同じ値を持つ.

表 1　対数変換後の視聴者数の要約情報

平均値	4.466
標準偏差	1.765
最小値	2.398
最大値	10.915

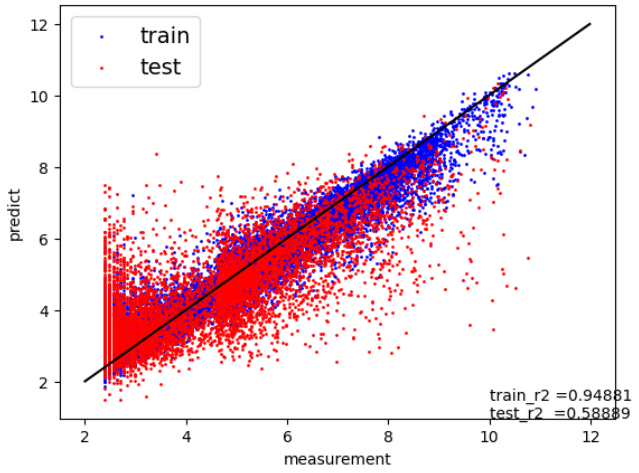


図 3　実測値と予測値の比較

表 2　決定係数と平均絶対誤差の結果

R^2	0.589
MAE	0.819

表 2 を見ると, 決定係数が 0.589 とやや当てはまりの良い予測モデルを構築することができた. また, 平均絶対誤差が 0.819 であった. 標準偏差が 1.765 で, 最小値から最大値までの範囲も広い. このようなデータの中で 0.819 の MAE は, そのばらつきを考慮すると, まずまず優れた予測精度を示していると考える. しかし, 図 3 を見ると, 実測値よりもかなり小さく予測されてしまっているケースが複数見られる. また, 実測値がかなり小さいデータでは逆に大きい値に予測されている.

この原因を考えると, 特徴量生成が挙げられる.

まず配信データでは配信者名と配信タイトルは文字数として変換を行った. しかし, 配信者名と配信タイトルは Twitch 公式サイトのトップで各配信者の配信概要として掲載される重要な要素である. 特に配信タイトルはハッシュタグやゲームの進捗状況などを載せている配信者が一定数見られ, 他配信者との差別化を図れる. 単純に文字数として処理するのではなく, 自然言語処理の知見を取り入れて複数の特徴量として処理できれば, より個々の配信の違いを明確にし, 精度を高めることができると考える.

次にゲームデータについて, 特徴量「player_perspectives」や「game_modes」, 「themes」, 「genres」に One-Hot エンコーディングを適用した. 例えば, 「genres」では 23 ジャンルがゲームごとに複数個割り当てられており, 23 個の特徴量を新たに作成し, そのジャンルに割り当てられるか否かで 01 に変換した. このように 23 ジャンルといった比較的要素数の少ないものに関しては適用することができたが, 「game_tags」や「platforms」など多くの特徴量は数百または数千以上の要素数を持ち, One-Hot エンコーディングを適用するのは現実的ではなかったため, ゲームごとに割り当てられている要素の数として要素数変換をした. ゲームのプラットフォームやサポート言語といった重要な特徴量を個々に見ることができればより改善できると考える.

5.2　SHAP

図 4 は SHAP 値を用いた重要度が高い上位 20 個の特徴量を示している.

次に図 5 は重要度が高い 20 個の特徴量において, データセット全体での各特徴量の影響を可視化したものである. 各特徴量の色はその特徴量の値の大小を表しており, 赤が大きく, 青が小さいことを意味する. 各特徴量において 1 点がデータの 1 行に該当し, ある 1 点の SHAP 値が負であるとその行の中でその特徴量は予測値を小さくする影響を与えている. SHAP 値が正であれば予測値を大きくする影響を与えている.

また, 図 6 と図 7 はテストデータ内の最も目的変数が小さいまたは大きいデータにおける各特徴量の影響を示している.

図 8 や図 9, 図 10, 図 11 は横軸に特徴量の値, 縦軸にその特徴量の SHAP 値を取っており, 大まかな SHAP 値と予測値への影響の関係を示したものである. また, 各図の右の特徴量は SHAP ライブラリによって選択された交互作用が最も大きい特徴量である. 特徴量が 112 個と多いため, 傾向が比較的顕著な特徴量のみに絞って掲載する.

図 4 を見ると, 特徴量「tags」が最も重要度が高いことがわかる. 特徴量「tags」は配信者が設定できる配信のタグの個数を示している. 図 5 を見ると, タグの個数が多いものは予測値を小さくする影響があり, 逆に少ないものは予測値を大きくする影響があることがわかる. 実際に図 6 と図 7 を見ると, それぞれその影響を見ることができる. 図 8 を見ると, 特に 1, 2 の値を持つ場合はプラスの影響を持つ傾向があり, 個数が増えると次第にマイナスの影響を持つ傾向が高まることがわかる. また, 特徴量「tags」が 1, 2 かつ特徴量「themes_18」が 1, つまり「ミステリー」というテーマをもつゲームを配信すれば, よりプラスの影響を与える傾向があることがわかる.

次に重要度の高い特徴量「follows」について, この特徴量は

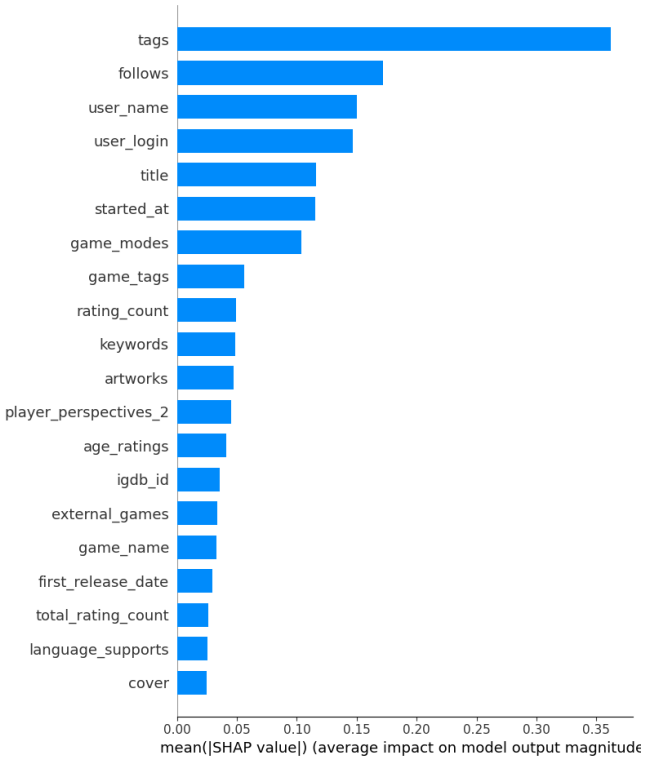


図 4　SHAP 値を用いた重要度が高い上位 20 個の特徴量

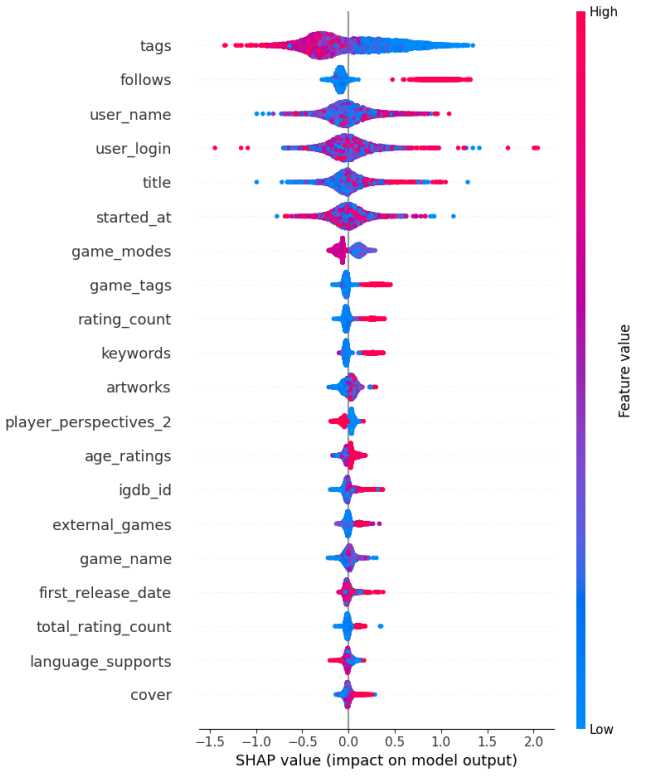


図 5　データセット全体における各特徴量の影響

IGDB の公式サイトでそのゲームをフォローしているユーザー数を意味する. 図 5 からある基準値を超えるフォロー数を持っていれば予測値に大きなプラスの影響を与え, 基準値以下であればマイナスの影響を与えることがわかる. 図 9 を見ると 1500 あたりが基準値となっていることがわかる. しかし, 特徴

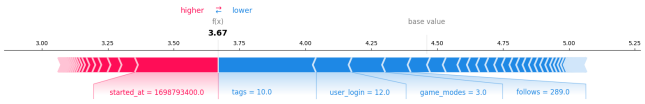


図 6　最も目的変数が小さいデータにおける各特徴量の影響

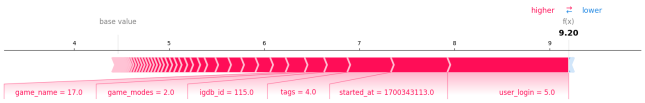


図 7　最も目的変数が大きいデータにおける各特徴量の影響

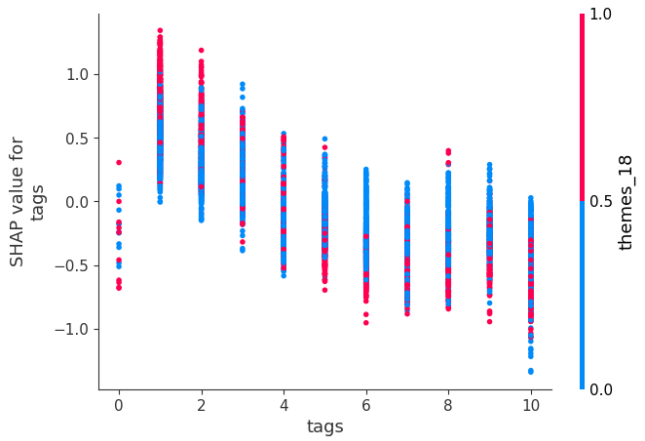


図 8　特徴量「tags」が与える影響の散布図

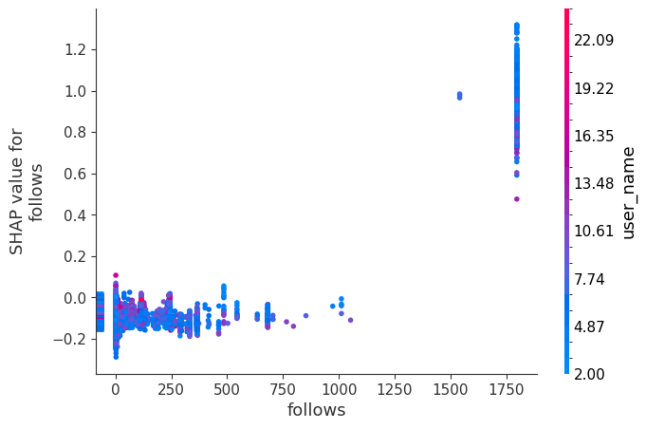


図 9　特徴量「follows」が与える影響の散布図

量「user_name」との明確な関係は見られず, フォロー数が多いゲームほど視聴者が集まってきやすいという自然な結果のみ得られた.

他の特徴量を見てみると, 図 10 では IGDB 上でゲーム情報が初めて作られた時間を意味する「created_at」が最新であればあるほどプラスの影響を与える傾向があることがわかる. また図 11 ではシングルプレイヤーやマルチプレイヤーなどを意味するゲームモードの要素数を表す特徴量「game_modes」が 1 か 2 でかつ「教育」のテーマを持つゲームであればよりプラスの影響を与える傾向があることがわかる.

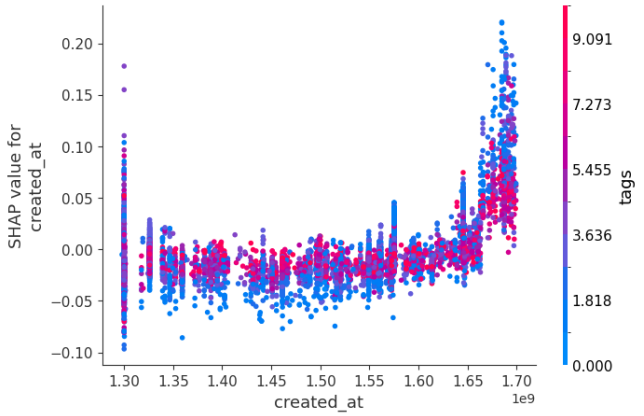


図 10　特徴量「created_at」が与える影響の散布図

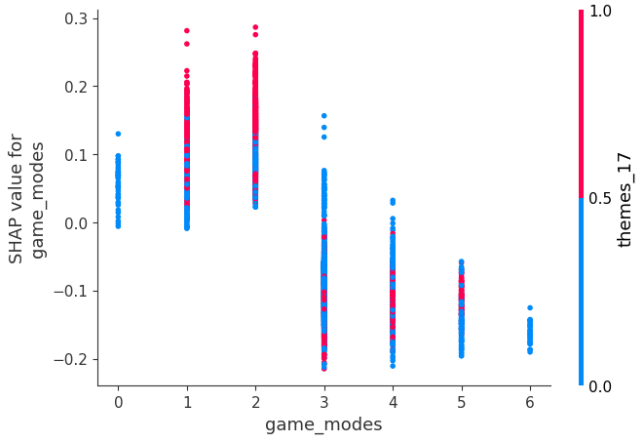


図 11　特徴量「game_modes」が与える影響の散布図

6 ま と め

本研究では Twitch API と IGDB API を用いたデータセットを用いて XGBoost で視聴者数予測を行った. 精度はまずまずであり, 特徴量生成の面で改善点が挙げられた. SHAP を用いた要因分析では以下のことが分析された.

- 配信タグが 1, 2 個でかつ, 「ミステリー」のテーマを持つゲームを配信すると視聴者数が増える傾向がある
- IGDB 上でフォロワー数が 1500 人以上のゲームであれば視聴者数が増える傾向がある
- 最新のゲームほど視聴者数が増える傾向がある
- ゲームモードが 1, 2 個でかつ, 「教育」のテーマを持つゲームを配信すると視聴者数が増える傾向がある

今後の展望として, 3 つ挙げられる. 1 つ目はデータ収集について, 期間は 2023 年 11 月の 1 か月間であった. しかしゲームの人気には流行があり, それはゲームコミュニティのイベントやプロゲーマーの大会などに大きく左右されると考える. 年単位のデータが集められれば様々な影響を考慮した研究ができると考えている. 2 つ目は特徴量生成についてだ. 節 5.1 でも言及したが, 自然言語処理や複数の要素を持つ特徴量の最適な処理に取り組んでいきたい. 3 つ目は要因分析の手法について, 今回は大域的分析と局所的分析の両方ができるという理由で SHAP を用いて分析を行った. 他にも様々な手法が開発されているため, 様々な手法を比較・検討を行っていきたい.

謝辞　本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです. また, 本研究は, 電気通信大学人工知能先端研究センター (AIX) の計算機を利用して実施したものです.

文　献

[1] Twitch. https://www.twitch.tv/.
[2] 稲垣 宗彦. ”毎月約 700 万人が配信　Twitch, CCO が語るライブ配信市場の進化”. 日経クロストレンド. 2023-11-09. https://xtrend.nikkei.com/atcl/contents/watch/00013/02343/. (参照 2023-12-24).
[3] A. Dybka, D. Kocjan and S. A. Ajila, ”A Semisupervised Approach to Predicting a Twitch Streamer ’s Growth based on the Game Streamed,” 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 4110-4119, doi: 10.1109/BigData55660.2022.10020838.
[4] SullyGnome.com. https://sullygnome.com/.
[5] Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016. Crossref, https://doi.org/10.1145/2939672.2939785.
[6] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD ’19). Association for Computing Machinery, New York, NY, USA, 2623–2631. https://doi.org/10.1145/3292500.3330701
[7] A. Adadi and M. Berrada, ”Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
[8] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.
[9] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. “why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics.
[10] Selvaraju, Ramprasaath R., et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” International Journal of Computer Vision, vol. 128, no. 2, Oct. 2019, pp. 336–59. Crossref, https://doi.org/10.1007/s11263-019-01228-7.
[11] M. Kuzlu, U. Cali, V. Sharma and Ö. Güler, ”Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools,” in IEEE Access, vol. 8, pp. 187814-187823, 2020, doi: 10.1109/ACCESS.2020.3031477.
[12] T. Chen et al., ”Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine,” in IEEE Access, vol. 7, pp. 150960-150968, 2019, doi: 10.1109/ACCESS.2019.2946980.
[13] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable Machine Learning for Fake News Detection. In Proceedings of the 10th ACM Conference on Web Science (WebSci ’19). Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/3292522.3326027

実際のタワーディフェンスゲームのステージを模したシミュレータでのステージ攻略 AI 開発

野村 大輔　秋岡 明香

明治大学総合数理学部 〒168-8525 東京都中野区中野 4-21-1

E-mail: {ev201082, akioka}@meiji.ac.jp

あらまし タワーディフェンスゲームのうち，キャラクター（タワー）の編成が必要でかつ，スキルを持つゲームを対象に，実際のステージをシミュレータに再現し，ゲームを攻略するプレイヤーAI の開発を目的に ML-Agents を用いて学習を行った．学習は 1000 万ステップで収束し，モデルが学習したステージをランダムな編成で攻略できることが示された．また，エピソード中に得ている報酬量はほとんど変化しなくなったが，学習したステップ量によって推論時にステージを攻略できる能力に差が存在した．また，入力に画像を用いることで，さらなる性能の向上を期待したが，画像の入力の有無ではモデルの性能に有意な影響がないことが分かった．さらに，学習に用いていないステージに関しては，完璧に攻略することはできなかったが，人間のプレイヤーに近い置き方を確認することもできた．

キーワード ゲーム・娯楽，ゲーム AI，タワーディフェンスゲーム

1. はじめに

タワーディフェンスゲーム（TD）は，リアルタイムストラテジーゲームの一種で，敵拠点から自拠点に向かって侵攻してくる敵キャラクターが自拠点に到達することを防ぐゲームである．プレイヤーは，敵を攻撃できるタワーを資金（コスト）を使って設置と強化を繰り返す．TD は複数の機械学習の研究対象となっており，Warcraft III World Editor[1]で作成した TD で，タワーの配置を遺伝的アルゴリズムで最適化したプレイヤーAI の開発[1]や，Kingdom Rush: Frontiers[2]をベースに，TD のマップ（敵の出現位置やタワーを設置できる場所，敵の出現位置と拠点の位置，敵が通る通路の位置を含む）と敵の出現するタイミングを，強化学習を用いた AI で生成し，AI でプレイテストすることによって，生成された TD マップの難易度を評価する研究[2]が存在する．

しかし，現在の TD は Arknigts[3]のように，キャラクター（タワー）[4]の種類が非常に多くなり，マップに持ち込むキャラクターを所持しているキャラクターから編成したり，ゲーム中にキャラクターがスキルを発動できたり，マップを攻略する要素が追加されたり，キャラクターがゲーム中に使えるそれぞれ異なるスキルを持ち，よりプレイヤーがゲームに介入できるようになった．そこで本研究では Arknights を対象に，どのような編成でも適切な手順でゲームを攻略できる AI の開発を目的に，実際のゲームを模したシミュレータを Unity に開発し，ML-Agents[3]でランダムな編成でゲームの攻略ができることを目的として，機械学習を行った．また，カメラセンサーを用いてマップのリアルタイム画像を入力に使用した際に，モデルがより多くの情報を利用することで，性能が向上するかどうかを確認するために，カメラセンサーの有無でニューラルネットワークモデルを 2 種類開発し，シミュレータ上で比較を行った．

[1] https://warcraft3.blizzard.com/en-us
[2] https://www.ironhidegames.com/Games/kingdom-rush-frontiers
[3] https://www.arknights.global
[4] 本来 TD ではタワーと呼ぶが，Arknights ではタワーと呼ぶのは不自然なため，キャラクターと表現する．

2. 関連研究

Arknights を対象に，手続き型コンテンツ生成と呼ばれる，アルゴリズムや AI によってゲームのコンテンツを生成，改善を行う研究として，プレイヤーの実力に合わせ，マップの難易度を調整することでより高い難易度の TD に対応できるプレイヤーAI の開発を行う研究[4]が存在する．この研究では，敵対的強化学習をベースにプレイヤーAI とマップの生成を行う AI の二つの AI をそれぞれ交互に学習させることで，適切なマップ難易度でのプレイヤーAI を学習させる．しかし，この研究では，使用するキャラクターを 12 体最初から決めており，学習中に編成が変化することが無い．そこで，本研究ではより多くのキャラクターを再現し，編成をランダムにすることによって，より実際のゲームの条件に近い AI 開発を行うことで，様々な編成に対応できるプレイヤーAI を開発する．

3. 提案手法

本研究では，Unity で開発した TD のシミュレータ上で，ML-Agents を用いて強化学習を行い，TD を攻

略できるニューラルネットワークを開発する．

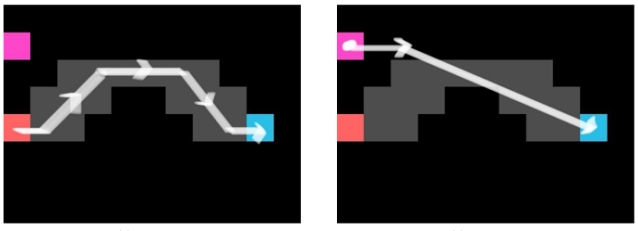
3.1 シミュレータの概要

表 1 は学習に用いたステージの実際のゲーム内のステージ名と，地上空中それぞれから進行する敵の数である．図 1 はステージ 0-4 における地上の敵と空中の敵の侵攻ルートを矢印で示したルートであり，地上の敵は，通路上を通って進行するのに対し，空中の敵は通路や壁を無視し直接拠点に向かって侵攻する．

学習の対象となるエピソード[5]は次のように構成される．エピソードが開始されると，表 1 の 22 ステージからランダムに 1 つマップが生成される．マップが生成されると決められたタイミングで敵キャラクターが生成され，敵キャラクターごとに決められた通路を通って敵が拠点に向かって侵攻してくる．敵が 3 体拠点に到達するか，すべての敵が流れ終わる（撃破される，もしくは拠点に到達する）と，エピソードは終了し次のエピソードが始まる．

表 1　学習に用いるマップ（ステージ）の一覧

実際のゲームでのステージ名	地上の敵の数	空中の敵の数
0-1	11	0
0-2	14	0
0-3	21	2
0-4	19	5
0-5	21	3
0-6	28	3
0-7	20	0
0-8	40	0
0-9	64	0
0-10	33	2
0-11	37	0
1-1	33	0
1-2	41	0
1-3	36	0
1-4	36	3
1-5	47	10
1-7	41	0
1-8	38	8
1-9	35	0
1-10	49	0
1-11	45	0
1-12	30	0



地上の敵の侵攻ルート　　空中の敵の侵攻ルート

図 1　地上の敵と空中の敵の侵攻ルートの例

[5] 学習の対象となるタスクの開始から終了までの期間．将棋なら初期盤面から投了するまでの期間．

表 2　キャラクターの一覧

シミュレータでの ID	置ける場所	HP	攻撃力	ブロック可能数	攻撃の種別	設置コスト
0	高台	1673	540	1	物理	14
1	高台	1060	375	1	物理	9
2	高台	1080	365	1	物理	9
3	高台	1370	465	1	物理	12
4	高台	1632	478	1	物理	12
5	高台	1320	475	1	物理	12
6	高台	1230	485	1	物理	12
7	高台	1565	977	1	物理	20
8	高台	1576	735	1	物理	18
9	高台	1280	528	1	物理	12
10	地上	3150	485	3	物理	22
11	地上	3105	365	3	物理	21
12	地上	2550	435	3	物理	20
13	地上	2035	295	3	物理	18
14	地上	1950	500	2	物理	13
15	地上	1325	325	2	物理	11
16	地上	1565	520	1	物理	10
17	地上	1835	546	1	物理	12
18	地上	2484	586	1	物理	13
19	地上	1845	558	1	物理	11
20	地上	3550	1064	1	物理	19
21	地上	2745	738	1	物理	15
22	地上	2745	587	1	物理	11
23	地上	2916	672	1	術	21
24	地上	2821	765	3	物理	24
25	地上	3850	430	3	物理	23
26	地上	1645	558	1	物理	10
27	高台	1135	362	1	回復	17
28	高台	1220	345	1	回復	17
29	高台	1520	505	1	回復	19
30	高台	1345	488	1	回復	18
31	高台	1689	534	1	物理	14
32	高台	1926	861	1	物理	29
33	高台	1680	870	1	術	34
34	地上	2560	713	2	物理	20

3.2 ゲームのルール

エピソードが開始する際に，表 2 の 35 キャラクターからランダムに 12 体選択される．この操作をランダムに 12 体編成されると呼ぶ．表 2 に示された体力，攻撃力，ブロック可能数，設置コストのパラメータのほかに，物理防御力，術防御力，攻撃速度のパラメータが割り振られており，キャラクターを設置する際はそれぞれ設定された設置コストを支払い，キャラクターを設置する．また、キャラクターによって地上に設置できるキャラクターと高台に設置できるキャラクターが存在する．地上に設置できるキャラクターは，一部の場所を除き敵の通る侵攻ルート上に設置することができ，キャラクターそれぞれに設定されたブロック可能数まで，地上の敵をキャラクターの場所で押しと

どめることができる．対して，高台のキャラクターは敵をブロックできない代わりに，遠距離の敵や，空中の敵を攻撃することができる．他に，各キャラクターの性質は物理攻撃，術攻撃，回復の 3 種類に分類され，物理攻撃のキャラクターは敵キャラクターの物理防御のパラメータを参照してダメージを与えるのに対し，術攻撃のキャラクターは敵の術耐性を参照してダメージを与える．また，回復のキャラクターは敵にダメージを与えられない代わりに味方キャラクターを回復することができる．

エピソード中プレイヤー（モデル）は編成された 12 体からコストを払って，同時に 8 体までキャラクターを設置することができる（ただし同じ種類のキャラクターを同時に設置することはできない）．コストは毎秒 1 ずつ入手するのに加え後述するスキルでも入手することができる．プレイヤーは設置したキャラクターを入れ替えたり，他の場所に設置したければ，キャラクターを一度撤退させることがでる．撤退したキャラクターは，それぞれに設定されたクールタイムを待てば再度設置することができる．

さらに重要な要素として，プレイヤーは必要に応じて各キャラクターに設定されたスキルを発動することができる．スキルはキャラクターによってさまざまな設定をされており，例として ID0 のキャラクターは 15 秒間攻撃回数が 5 倍になったり，ID14 のキャラクターは 12 コストを入手しつつ周囲一定範囲の敵に術ダメージと 3 秒のスタン（一定時間移動不能と攻撃不可）を与えることができる．スキルはキャラクターごとに発動可能になるまでの時間や再使用までの時間が決まっているため，ある程度予期して設置し，使用するタイミングもしっかりと判断しなければならない．

3.3 ニューラルネットワークの構成

ML-Agents はカメラセンサーと呼ばれる，入力にゲーム状態の画像を直接用いることのできる機能がある，カメラセンサーの有無で 2 種類のニューラルネットワークモデルを開発した．どちらのモデルも入力層，中間層，出力層から構成されるモデルであり，カメラセンサー無しのモデルの入力を表 3 に示す．グリッドマップ情報は，マップ上の設置可能マス，設置不可マス，高台，壁，敵出現位置（地上，空中，両方），拠点をそれぞれ 0 から 7 に割り振って入力する．また，地上と空中それぞれの敵の位置情報は，それぞれのマスに敵キャラクターが何体含まれるかを入力する．編成されたキャラクターの情報は編成された 12 キャラクターが現在設置されているか，クールタイム中かどうか，そのキャラクターが高台と地上のどちらに設置できるか，攻撃キャラか回復キャラか，そのキャラクターは正面にどこまで攻撃が届くかを入力し，キャラクターの状況は，設置されていない場合は固定の値が入力され，設置されている場合は，置かれている場所と向き，スキルが使用できる状況かどうか，今の体力（HP），攻撃力，攻撃速度が入力される．カメラセンサー有のモデルは入力に，表 3 と共通の長さが 421 の入力にマップの画像データ（120×50）を 36×36 の CNN で処理した入力を追加して用いる．中間層はどちらも 3 層で構成され一層あたり 256 ユニットを持つ全結合のニューラルネットワークである．出力層も共通で，内訳を表 4 に示す．

表 3　ニューラルネットワークの入力

		サイズ	型	範囲
グリッドマップ情報		88	int	0-7
敵の位置情報（地上）		88	int	0-20
敵の位置情報（空中）		88	int	0-20
所持コスト		1	int	0-99
編成されたキャラクターの情報	ID	12	int	0-34
	使用状況	12	bool	
	使用状況 2	12	bool	
	置ける場所	12	int	0-2
	攻撃可能か	12	bool	
	コスト回収	12	bool	
	攻撃範囲	12	int	0-4
キャラクターの状況	場所	12	int	0-88
	方向	12	int	0-4
	スキル待ち	12	bool	
	HP	12	int	0-10000
	攻撃力	12	int	0-3000
	攻撃速度	12	float	0-3.00

表 4　ニューラルネットワークの出力

Branch	操作対象		範囲
0	すでに設置されているキャラクター		0-2
1			0-2
2			0-2
3			0-2
4			0-2
5			0-2
6			0-2
7			0-2
8			0-2
9			0-2
10			0-2
11			0-2
12	キャラクターの設置	場所	0-88
13		キャラクター	0-12
14		方向	0-4

出力は 15 ユニット存在し，前半の 12 のユニットは編成したキャラクターの 1-12 番目のキャラクターがすでに配置されている場合 1 でスキルの発動，2 でキャラクターの撤退を行う．また設置されていない場合は 0 のみを出力する．後半の 3 ユニットは，マップ内の配置数が上限の 10 に達していないときに，そ

れぞれのユニットで場所とキャラクターと方向を指定し，そのキャラクターが配置可能のマスを指定できていればキャラクターを配置する．

3.4 報酬体系

ニューラルネットワークの学習に用いるアルゴリズムは PPO[5]であり，報酬はエピソード終了時に分配される．表 5 に報酬体系を示す．なお撃破率 t は（倒した敵の数）/（ステージの敵の総数）である．

表 5　報酬体系

	報酬量	
撃破率 t に対する報酬	t=1	+0.9
	拠点侵入が 3 体未満	+0.8
	拠点侵入が 3 体以上	+0.7*exp(-5t)
ゲーム中の行動に関する報酬	キャラクターの撤退	-0.05
	キャラクターが敵をブロック	+0.005（1 回につき）
	キャラクターの設置位置（地上）	隣接 4 マスに拠点，敵出現位置が存在するとき +0.04
	キャラクターの設置位置（高台）	隣接 4 マスのうち，タワーの正面が通路なら +0.04 左右それぞれが通路で +0.01

エピソードが終了時に，すべての敵が撃破できていれば+0.9，3 体未満の敵が拠点に到達してしまったが，すべての敵が流れ終わった場合は+0.8，拠点に 3 体以上敵が到達し，敵が全て流れ終わる前にエピソードが終了した場合+0.7*exp(-5t)の報酬が与えられる．これは，エピソード序盤の報酬が撃破率に影響されやすくすることで，エピソード序盤の敵を倒すことを優先するためである．また，エピソード中にキャラクターを自主的に撤退した回数（ニューラルネットワークの branch0-11 が 2 を出力した回数）に応じて 1 回につき報酬が-0.05 される．撤退時の報酬は，TD の報酬体系の考察[6]によると，序盤の敵を倒すことを優先した場合，撤退時の報酬が小さいと正常に学習できなくなってしまうことから，負に大きい-0.05 とした．また，キャラクターが敵をブロックすることは，敵の侵攻ルート上にキャラクターを置くことができていることを意味するため，キャラクターが敵をブロックするたびに+0.005 の報酬を与える．さらに，TD の経験上，拠点の隣や敵が出現する場所の隣にキャラクターを置くと，マップの攻略が簡単になる．そこで，地上に設置できるキャラクターは設置したときに，隣接 4 マスに拠点か敵出現位置があった場合+0.04 の報酬を与える．また，試行錯誤の段階で学習が収束してもモデルが敵の通る通路に背を向けて高台のキャラクターを設置して，キャラクターの設置が無駄になってしまうことがあったため，それを防ぐために正面に通路があった場合+0.04，左右に通路があった場合+0.01 が与えられる．すなわち，最大 3 方向（正面と左右）すべてが通路だった場合，+0.06 の報酬が与えられる．

3.5 シミュレータ開発環境と学習パラメータ

表 6 に学習を行ったシミュレータに用いた Unity と ML-Agents，Python，CUDA のバージョンを示す．また，表 7 に ML-Agents に用いた学習のパラメータを記す．

表 6　学習環境のバージョン

名称	バージョン
Unity	2021.3.6.f1
ML-Agents	0.30.0
Python	3.9.13
CUDA	10.2

表 7　学習パラメータ

パラメータ名称	値
batch_size	100
buffer_size	1E5
learning_rate	1E-4
beta	0.001
epsilon	0.2
lambd	0.91
num_epoch	3

batch_size は勾配降下 1 度に使用する経験[6]の数であり，buffer_size はモデルの方策を更新するまでに収集する経験の数，learning_rate は勾配降下の更新ステップの大きさであり，学習ステップ数が増えるにつれて徐々に線形に減少する．また，beta はどれだけランダムに行動をするかを指示する値であり，epsilon は方策を更新する際に，更新できる幅の最大値を与え，lambd は Generalized Advantage Estimate（GAE）の計算に用いられる正則化パラメータである．num_epoch は方策更新の際に学習データを何回学習するかを示す．

4. 実験結果

最初に，どれくらいの学習ステップ数で報酬量が収束するか確認し，その後収束したモデルの性能を評価する．しかし，実際にどのくらい攻略できるかの性能は，報酬量から評価するのは難しい．そこで，学習させたモデルをシミュレータ上で一定時間推論させることで評価を行う．

図 2 はカメラセンサーの有無による 2 種類のモデ

[6]経験の数=(状態，行動，報酬)の組み合わせの数

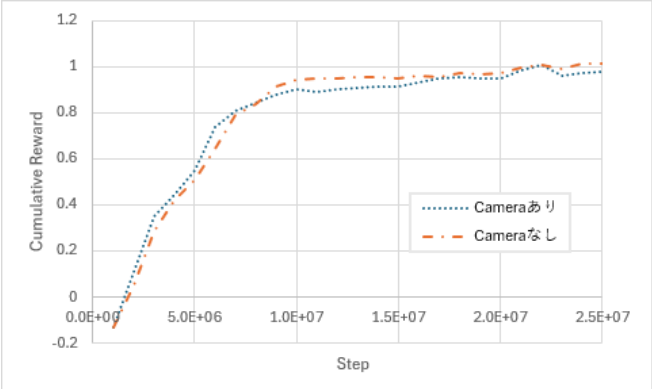


図 2　ステップ数と報酬量のグラフ

ルの学習時の報酬量に関する学習曲線である．また，表 8，表 9 はカメラセンサーを持つモデルと持たないモデルの性能を，1000 万ステップを学習した時点のモデルから 300 万ステップごとに，それぞれで一定時間推論することで，90%以上の敵を撃破できたエピソード数と，撃破できた敵の数が 10%に満たないエピソードの数をそれぞれで比較した表である．すなわち，t≧0.9 のエピソード数の割合が最も高く，かつ t<0.1 のエピソード数の割合が最も低いモデルが性能の良いモデルとなる．図 2 によると，どちらの学習も 1000 万ステップ程度である程度収束し，それ以降は得られる報酬量がほとんど変化しない．そこで，1000 万，1300 万，1600 万，1900 万，2200 万，2500 万ステップの学習段階でモデルの性能を比較する．その結果，表 8 によるとカメラセンサー有では 1600 万ステップ時点でのモデルが最も性能が良く，2500 万ステップ時点でのモデルの性能とは t≧0.9 では最大 14%の差がついてしまっていた．また表 9 の結果から，カメラセンサー無のモデルに関しては 1300 万ステップのモデルが，他のモデルに対して特に劣ってしまっていることがわかる．以上から，学習が収束した後も，各モデルを実際に推論するまで，最適なステップ数がわからないことが分かる．

一方で，ステージごとの攻略性能も比較した．表 10 は各モデルが，各ステージを推論した際に，ステージのすべての敵を撃破できる確率と，拠点にたどり着く敵が 3 体未満で，すべての敵が流れ終わる確率（表中の括弧内）を求めた表である．なお，敵をすべて倒せたエピソードの割合は（そのステージで敵を全て倒せたエピソード数）／（そのステージが選ばれたエピソード数）で与えられ，敵が侵入した数が 3 体未満のエピソードの割合は，（敵が全て流れ終わったエピソード数）／（そのステージが選ばれたエピソード数）で与えられる．また白抜き文字は，各ステージにおいて敵が一体も侵入しなかった割合がそれぞれ最も高いモデルをカメラセンサーの有無でそれぞれ示し，背景が灰色のセルは，敵をすべて倒せたエピソードの割合と異なる場合にすべての敵が流れ終わる確率が最も高いモデルを表す．実際に，カメラセンサーありのモデルについて，表 8 で最も性能が良かった 1600 万ステップのモデルは多くのステージで黒または灰色のセルであり，ステージごとの観点からも最も良い結果を示していた．一方で，表 8 では他のモデルと比較して性能が悪かった 2200 万ステップのモデルもいくつかのステージで，良い結果を示していた．同様のことがカメラセンサー無のモデルでも起きており，表 9 で最も性能の良かった 2500 万ステップのモデルは多くのステージで最も良い結果を示してはいたが，2500 万ステップのモデルに性能で劣る 1600 万ステップのモデルが最も良い結果を示すステージがいくつか存在した．このことから，得られる報酬量があまり変化しなくなってから学習する際に，いくつかのステージでの攻略を犠牲にしていると推測できる．特にカメラセンサー無しの 0-10 や 1-2 は，学習ステップが進むとかなり攻略する能力が落ちてしまっている．この点については，今後どのように一部のステージを攻略する能力が劣化してしまうことを防いでいくかを検討する必要がある．

表 8　カメラセンサー有のモデルの推論時性能

モデルのステップ数	t:撃破率		エピソード総数
	t<0.1 のエピソード数(割合[%])	t≧0.9 のエピソード数(割合[%])	
1000 万	2662(15.90)	11537(68.89)	16746
1300 万	1743(11.23)	12046(77.62)	15520
1600 万	1237(8.15)	12472(82.21)	15170
1900 万	1563(10.11)	12243(79.21)	15456
2200 万	2140(13.54)	12245(77.45)	15810
2500 万	2767(16.60)	11500(68.98)	16672

表 9　カメラセンサー無のモデルの推論時の撃破率

モデルのステップ数	t:撃破率		エピソード総数
	t<0.1 のエピソード数(割合[%])	0.9≦t のエピソード数(割合[%])	
1000 万	3185(19.06)	11865(71.01)	16710
1300 万	4545(25.42)	11728(65.60)	17877
1600 万	3267(19.39)	12071(71.64)	16850
1900 万	2327(14.60)	12157(76.27)	15939
2200 万	2363(14.87)	12091(76.10)	15899
2500 万	1898(12.02)	12238(77.49)	15794

表 10　モデルの推論時のステージごとの性能比較

モデル	入力	カメラセンサーの有						カメラセンサー無					
	学習したステップ数[百万]	10	13	16	19	22	25	10	13	16	19	22	25
ステージ		敵を全部倒せたエピソードの割合[%] (敵が侵入した数が 3 体未満のエピソードの割合[%])											
0-1		82.6 (85.6)	87.0 (90.2)	94.7 (95.5)	89.0 (93.4)	86.8 (89.7)	85.8 (87.5)	92.1 (94.8)	80.7 (84.2)	95.6 (96.8)	98.2 (98.9)	99.1 (99.6)	97.8 (98.9)
0-2		93.9 (97.9)	98.3 (99.3)	98.4 (99.0)	94.7 (97.2)	85.4 (89.8)	88.6 (92.0)	91.5 (94.4)	92.7 (95.2)	88.7 (90.8)	86.7 (89.6)	93.3 (95.3)	96.6 (97.7)
0-3		86.5 (94.7)	93.3 (98.0)	90.7 (95.4)	89.7 (93.6)	87.4 (90.2)	77.1 (83.6)	91.0 (98.1)	92.8 (97.2)	97.5 (99.8)	89.8 (94.4)	94.3 (97.9)	81.8 (85.1)
0-4		71.7 (86.4)	80.8 (89.4)	94.1 (97.8)	90.2 (93.6)	83.4 (85.4)	75.4 (79.7)	80.2 (87.3)	72.2 (75.4)	87.0 (89.9)	94.3 (98.8)	88.1 (92.1)	96.1 (99.2)
0-5		47.8 (85.9)	36.2 (74.8)	53.5 (87.6)	58.0 (87.4)	48.8 (81.2)	31.9 (71.5)	28.9 (87.1)	17.3 (73.1)	24.9 (77.2)	23.8 (88.9)	30.4 (86.2)	45.1 (93.7)
0-6		19.6 (67.8)	21.8 (74.1)	34.6 (87.9)	34.4 (83.4)	35.8 (79.7)	28.9 (73.6)	13.9 (59.2)	6.2 (41.3)	9.8 (47.1)	14.9 (52.0)	12.0 (37.5)	13.4 (47.4)
0-7		69.7 (79.0)	82.6 (88.0)	88.3 (94.9)	82.9 (91.6)	76.5 (85.2)	65.2 (80.0)	75.0 (81.2)	63.6 (70.3)	68.2 (76.9)	80.0 (84.9)	81.4 (85.5)	74.0 (86.0)
0-8		60.4 (81.4)	75.6 (86.3)	78.2 (91.8)	76.2 (84.9)	71.0 (83.9)	54.6 (74.1)	79.8 (91.1)	77.0 (89.1)	79.5 (91.0)	87.8 (95.0)	87.7 (94.8)	86.8 (94.8)
0-9		45.6 (69.6)	62.1 (81.9)	61.3 (86.5)	63.3 (81.8)	61.9 (77.5)	48.6 (67.4)	61.2 (82.6)	54.0 (76.1)	65.4 (84.5)	65.5 (86.2)	69.4 (89.4)	69.8 (86.0)
0-10		27.5 (56.7)	60.0 (83.0)	67.7 (83.9)	63.3 (81.6)	50.5 (75.4)	44.8 (71.4)	34.4 (56.9)	23.7 (47.8)	50.8 (75.0)	37.1 (68.3)	49.7 (76.4)	36.5 (66.7)
0-11		45.5 (72.3)	65.3 (87.2)	70.5 (91.2)	66.2 (90.5)	64.6 (91.7)	45.0 (75.7)	51.4 (73.9)	47.6 (65.6)	45.6 (64.3)	52.8 (72.7)	60.8 (81.2)	72.9 (91.0)
1-1		36.7 (81.7)	35.3 (87.2)	42.1 (87.8)	44.4 (84.5)	51.5 (89.0)	44.7 (88.3)	28.2 (72.3)	28.4 (63.2)	44.4 (72.6)	37.4 (80.2)	37.0 (78.9)	41.1 (74.9)
1-2		36.3 (63.6)	36.2 (63.1)	38.1 (62.9)	26.2 (54.4)	21.1 (46.3)	16.6 (39.4)	37.9 (62.1)	43.6 (66.4)	49.5 (71.8)	51.2 (71.0)	48.5 (72.9)	28.5 (54.2)
1-3		22.7 (43.0)	33.3 (55.0)	34.6 (58.3)	29.4 (50.1)	30.1 (50.7)	21.0 (42.5)	29.2 (50.9)	30.1 (50.6)	27.9 (47.8)	43.4 (64.2)	54.6 (70.4)	49.8 (70.8)
1-4		27.0 (47.3)	44.0 (63.5)	51.1 (73.6)	62.9 (80.5)	68.0 (88.7)	66.8 (87.3)	32.7 (52.8)	32.6 (53.6)	34.4 (59.3)	40.4 (64.5)	36.9 (60.4)	44.9 (67.7)
1-5		14.6 (33.1)	17.4 (40.3)	26.9 (49.2)	30.8 (59.1)	32.0 (62.2)	21.7 (51.6)	9.84 (31.0)	11.0 (28.8)	12.5 (32.6)	18.6 (41.0)	21.7 (46.2)	18.3 (42.1)
1-7		59.4 (80.2)	73.1 (85.8)	81.2 (93.4)	70.1 (83.0)	86.9 (93.8)	72.5 (85.0)	75.0 (85.2)	80.0 (89.6)	81.5 (91.4)	73.6 (82.9)	80.7 (89.0)	70.4 (80.7)
1-8		12.5 (29.9)	27.9 (48.4)	34.0 (55.9)	26.7 (48.6)	24.9 (39.8)	12.6 (30.9)	21.3 (35.3)	16.6 (28.6)	24.8 (37.8)	36.7 (54.2)	31.0 (49.8)	50.4 (63.6)
1-9		50.8 (81.0)	62.2 (89.7)	63.1 (89.0)	55.6 (85.6)	54.0 (90.4)	31.4 (77.0)	49.6 (73.5)	41.7 (64.2)	42.1 (69.2)	50.6 (80.3)	53.8 (75.3)	65.7 (87.5)
1-10		44.9 (69.5)	57.6 (80.6)	59.2 (81.3)	54.8 (78.6)	48.5 (81.9)	31.1 (71.8)	48.8 (66.8)	41.4 (59.7)	38.2 (61.2)	47.5 (70.4)	56.0 (70.9)	61.9 (81.0)
1-11		22.9 (39.1)	40.3 (60.8)	44.8 (65.5)	43.1 (61.9)	47.6 (69.6)	20.1 (34.9)	33.3 (48.7)	35.4 (51.8)	35.1 (55.8)	37.6 (57.7)	40.2 (60.4)	31.2 (51.0)
1-12		28.5 (63.1)	38.0 (77.4)	42.0 (74.2)	34.5 (69.8)	21.2 (57.6)	14.8 (47.3)	46.3 (77.2)	40.8 (67.4)	44.2 (77.2)	47.2 (76.2)	37.1 (69.6)	45.0 (75.5)

表 11 カメラセンサーの有無で
大きく性能の差が出たステージ

ステージ	カメラセンサー有	カメラセンサー無
0-6	35.8	14.9
0-10	67.7	50.8
1-2	38.1	51.2
1-3	34.6	54.6
1-4	68.0	44.9
1-8	34.0	50.4

表 11 は各ステージで最も良い性能（敵を全部倒せたエピソードの割合が一番高い）のモデルをカメラセンサーの有無で比較した際に，大きく差のついたステージを表 10 から抜粋したものである．図 3 で，各ステージの大まかな形状と敵の侵攻ルートの例を示す．図 3 から，カメラセンサー有で性能の良かった 0-6，0-10，1-4 の共通点と，カメラセンサー無で性能の良かった 1-2，1-3，1-8 の共通点は，明確なものが存在しないと考える．したがって，今回，カメラセンサーの有無での性能差は，機械学習のランダム性によっ

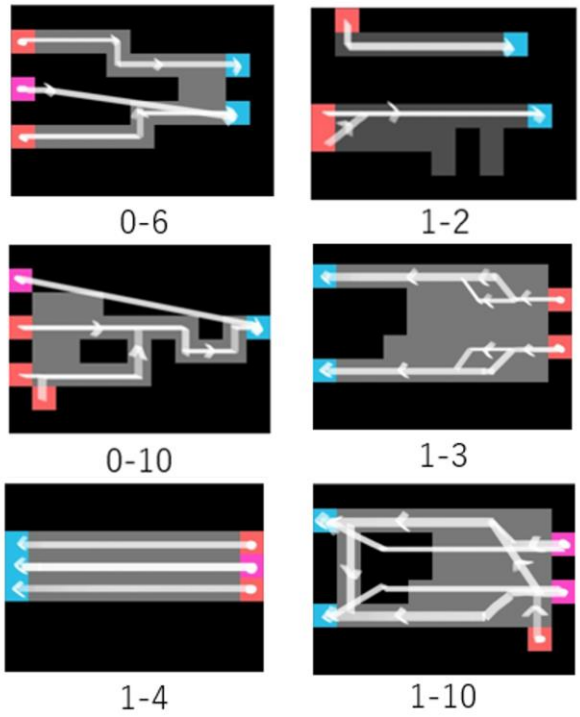


図 3　表 1 の各ステージの形状と敵の侵攻ルートの例

表 12　推論に用いるステージ

実際のゲームでのステージ名	地上の敵の数	空中の敵の数
2-1	50	0
2-2	51	0
2-5	40	0

て起きたと考えるのが自然であり，現状の入力系と報酬系でニューラルネットワークモデルを開発した際に，カメラセンサーの有無は学習が終了したモデルの性能に影響を与えるとは言えない．最後に，学習に用いていないステージで開発したモデルを比較する．表 12 は推論に用いた学習に用いていないステージの一覧である．また，ステージの形状と敵の侵攻ルートを図 4 に示す．表 13 は表 10 と同様の計算方法で算出した，ステージのすべての敵を撃破できる確率と，拠点にたどり着く敵が 3 体未満で，すべての敵が流れ終わる確率の表である．さらに，図 5 はモデルが推論する際に，行ったキャラクターの配置の例である．この結果によると，3 ステージともに難易度が学習したステージとあまり差が無いのにも関わらず，すべてのモデルで全ての敵を倒せる確率が 30%を切ってしまっており，ステージ 2-2 の結果に至っては 3%に満たない結果となってしまったため，モデルが攻略できるステージは学習に用いたステージのみという結果にとどまった．これはステージが 22 ステージしかないために用いるステージを増やして検証する必要がある．マップに過適合してしまっている可能性があり，学習に用

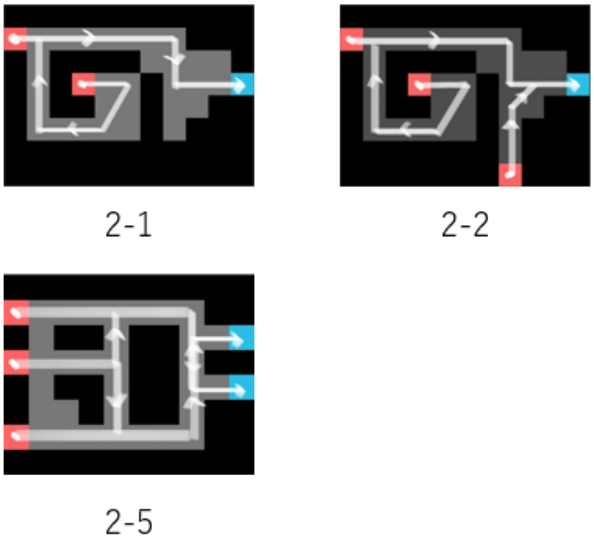


図 4 ステージの形状と侵攻ルートの例

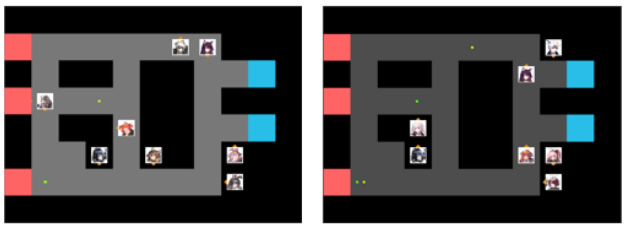


図 5　未知のステージに対する推論中の配置の例

いるステージを増やして検証する必要がある．また，2-1 と 2-2 に関しては，敵の出現位置が一つ増え，キャラクターの設置できる位置は変化していないにもかかわらず，かなりの差がついてしまっている．そのため，ステージの変化が微小であっても，モデルの推論に大きく影響を与える可能性があると考える．ただし，図 5 のように人間がステージを攻略する際と同様のキャラクターの配置も見られたため，TD の基本的な部分については学習できていると思われる．

以上の結果から，今後はより多くのステージを学習に用いる必要があるが，本研究のように実際のステージのみを用いる場合，用意できるのは高々100 ステージ程度である．さらに，実際のゲームの特性上，高い難易度のステージをシミュレータに実装する場合に，ステージギミックをシミュレータで再現することが難しい．そのため，実際のステージのみを用いて AI を開発することは限界がある．そこで，学習に用いるステージを自動で生成することで，学習に用いるステージを増やすことが必要となる．この点については，2-1 と 2-2 の結果から，少ないマップの変化で推論に影響を与えるならば，少ないマップの変化で学習に影響を与えられるのではないかと考えているため，今後は既存のマップを 1～2 マス変化させることで，学習に用いるマップを増やして研究を進めたいと考えている．

表 13　学習に用いていないステージでのモデルの性能

モデル	入力	カメラセンサーの有						カメラセンサー無					
	学習したステップ数[百万]	10	13	16	19	22	25	10	13	16	19	22	25
ステージ		敵を全部倒せたエピソードの割合[%] （敵が侵入した数が 3 体未満のエピソードの割合[%]）											
2-1		6.1 (31.3)	4.3 (22.3)	56 (33.3)	3.8 (13.3)	5.1 (19.0)	4.6 (22.0)	7.5 (17.8)	9.2 (20.7)	2.8 (16.5)	5.2 (19.8)	6.5 (19.4)	5.9 (21.2)
2-2		2.4 (7.2)	1.0 (2.1)	0.0 (4.5)	0.0 (1.3)	1.0 (2.0)	0.0 (2.4)	2.6 (2.6)	0.0 (2.8)	2.5 (3.3)	0.0 (7.6)	1.2 (6.0)	0.0 (3.1)
2-5		9.9 (46.9)	13.1 (42.4)	20.8 (51.9)	27.7 (64.9)	23.7 (64.5)	9.4 (44.7)	18.6 (48.8)	10.4 (39.0)	10.9 (22.7)	14.9 (48.6)	9.3 (34.9)	7.8 (30.1)

5. まとめと今後の課題

本稿では，TD の攻略 AI の開発を目的に，入力にカメラセンサーを用いるか否かの比較を行った．今回の比較において，カメラセンサーの有無による，ニューラルネットワークモデルの性能差は機械学習のランダム性の範囲でしかみられなかった．したがって，Arknights では，学習する際に入力の量が少なくなくなるカメラセンサー無のモデルで十分だと考える．また，モデルの学習がある程度収束してからも，学習の長さが異なるモデルごとにかなり推論の性能に差が出てしまうため，いくつかのステップ数で比較する必要があることもわかった．さらに，学習に用いたステージに関しては，編成がランダムに変化しても対応できていたといえる．しかし，学習に用いていないステージに関しては，モデルが完璧には適応できなかった点から，ステージへの過適合が疑われる．したがって，今後の研究においては，未知のステージへの対応力を対象とし，もっと多くのステージを用いて学習を行うべきといえる．

参　考　文　献

[1] Tan Tse Guan, Yong Yung Nan, Chin Kim On, Jason Teo, and Rayner Alfred, “Automated Evaluation for AI Controllers in Tower Defense Game Using Genetic Algorithm”, Communications in Computer and Information Science (2013)

[2] Simon Liu, Li Chaoran, Li Yue, Ma Heng, Hou Xiao, Shen Yiming, Wang Licong, Chen Ze, Guo Xianghao, Lu Hengtong, Du Yu, and Tang Qinting, “Automatic generation of tower defense levels using PCG”, FDG '19: Proceedings of the 14th International Conference on the Foundations of Digital Games(2019)

[3] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange, “Unity: A general platform for intelligent agents”, https://arxiv.org/abs/1809.02627 (2020,2024/01/10 取得)

[4] Yueming Xu, and Tetsuro Tanaka, “ Procedural content generation for tower defense games:a preliminary experiment with reinforcement learning”, 第 26 回ゲームプログラミングワークショップ (GPW-21) (2021),

[5] Jhon Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, “Proximal Policy Optimization Algorithms”, https://arxiv.org/abs/1707.06347 (2017,2024/01/10 取得)

[6] 野村大輔，秋岡明香,“タワーディフェンスゲームの攻略 AI 開発における報酬体系の考察”,電子情報通信学会　総合大会（2024）

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年3月1日(金) 16:00 ~ 18:10 | T1-A オンライン（Zoom Events)

トピックモデリング

座長:伏見 卓恭(東京工科大学)
コメンテータ:白井 匡人(島根大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T1-A-9-01] DQNを用いたトピックモデルと予測モデルの相互学習
*中村 礼音[1]、伊藤 寛祥[1]、森嶋 厚行[1] (1. 筑波大学)

16:30 ~ 16:55
[T1-A-9-02] Human-in-the-Loop ニューラルトピックモデリング
*米谷 亜夕[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学 情報学群、2. 筑波大学 図書館情報メディア系)

16:55 ~ 17:20
[T1-A-9-03] Pragmatic Approach to Automatic Topic Labeling with Zero-shot Classification
*周 子楷[1]、若林 啓[1] (1. 筑波大学)

17:20 ~ 17:45
[T1-A-9-04] BERTopic文書分類器の分類安定性に関する評価手法の提案
*櫻井 勇気[1]、小林 亜樹[2] (1. 工学院大学 工学研究科修士課程 電気・電子工学専攻 メディア情報システム研究室、2. 工学院大学 工学研究科 電気・電子工学専攻 メディア情報システム研究室)

17:45 ~ 18:10
[T1-A-9-05] 【技術報告】文書のバージョン管理の提案
楢木 悠士 (株式会社LegalOn Technologies)

DEIM Forum 2024 T1-A-9-01

多目的強化学習に基づくトピックモデルと予測モデルの同時最適化

中村　礼音†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学人間総合科学研究群情報学学位プログラム　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †s2321708@u.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp

あらまし　新興技術のトピックやトレンドを理解することは，企業や研究機関が技術開発の意思決定をする際に有益である．BERTopic は，Transforme を活用したトピックモデリング手法であり，高い精度でトピックの推定を行うことが可能である．本研究では，多目的強化学習の枠組みでトピックモデルとトピックのトレンド予測モデルを同時最適化する手法を提案する．具体的には，トピックの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）を状態（State）として扱い，報酬（Reward）の最大化を目指す．Coherence と Topic Diversity をトピックモデルの報酬として，予測値と実測値の乖離度（Euclidean Distance）をトピックのトレンド予測モデルの報酬として扱うことで，トピックモデルとトピックのトレンド予測モデル双方を考慮した同時最適化が可能となる．建築分野の特許文書を用いた実験の結果，提案手法を使用することで，比較手法よりも精度が高いトピックの抽出と将来のトピックのトレンド予測が可能であることが確認できた．

キーワード　トピックモデル，テキストマイニング，自然言語処理応用，LLM

1　序　　論

新興技術のトピックやトレンドを理解することは，企業や研究機関が技術開発の意思決定をする際に有益である．一方，科学論文や特許などの文献情報から，新興技術のトピックやトレンドを理解することは非常に困難で，多大な努力と費用が必要である [1].

トピックモデルは，文書集合からトピックを自動で推定する手法であり，「文書分類」「テキスト要約」「検索」「推薦」など，幅広い領域で利用されている [2] [3]．代表的なトピックモデルとして，潜在ディリクレ配分法（LDA）[4] や，ニューラルトピックモデル（NTM）[5] [6] が存在し，大規模言語モデル（LLM）の登場以降は，文脈表現をトピックモデルに導入する試みが盛んになっている．

BERTopic [7] は，Transformer を活用したトピックモデルであり，高い精度でトピックの推定を行うことが可能である．また，BERTopic はさまざまな拡張機能をサポートしている．例えば，merge_topics は類似するトピック同士を結合することができる．さらに，BERTopic の拡張機能を利用することで，時間経過に伴うトピックの変化を分析することも可能である．これにより各トピック出現頻度が時刻ごとにプロットされるため，トピックのトレンドがこれまでどのように変化したきたかを確認することができる．

本研究では，BERTopic と予測モデルを使用することで，トピックのトレンド予測を行う．さらに，多目的強化学習の枠組みでトピックモデルとトピックのトレンド予測モデルを同時最適化する手法を提案する．具体的には，トピックの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）を状態（State）として扱い，Coherence と Topic Diversity をトピックモデルの報酬として，予測値と実測値の乖離度（Euclidean Distance）をトピックのトレンド予測モデルの報酬として扱うことで，トピックモデルとトピックのトレンド予測モデル双方を考慮した同時最適化が可能となる．

実験では，はじめに DQN の枠組みでモデル学習が適切に行われているか否かを確認した．次に，トピックモデルとトピックのトレンド予測モデルの同時最適化を行うことで，両モデルを考慮した学習が行えているか否かを確認した．前者に関しては，DQN 学習前後の Coherence， Topic Diversity，予測値と実測値の乖離度（Euclidean Distance）を比較することでトピックモデルとトピックのトレンド予測モデルの性能がそれぞれ向上するかを検証した．後者に関しては，トピックモデルのみで最適化を行った場合と，トピックモデルとトピックのトレンド予測モデルの同時最適化を行った場合で実験結果を比較するこで，両モデルを考慮した学習が行えているか否かを検証した．

建築分野の特許文書を用いた実験の結果，提案手法を使用することで，比較手法よりも高いトピックの抽出と将来のトピックのトレンド予測が可能であることが確認できた．

本研究の貢献は次の通りである．

（1） 多目的強化学習に基づき，トピックモデルとトピックのトレンド予測モデルを同時最適化する新しい手法を提案した．

（2） 提案手法により，トピックモデルとトレンド予測モデルの精度が向上することを実験的に示した．

2　関 連 研 究

トピックモデルは，文書が生成される過程を，確率を用いてモデル化した確率的生成モデルであり，代表的な手法として，潜在ディリクレ配分法（LDA） [4] やニューラルトピックモデル

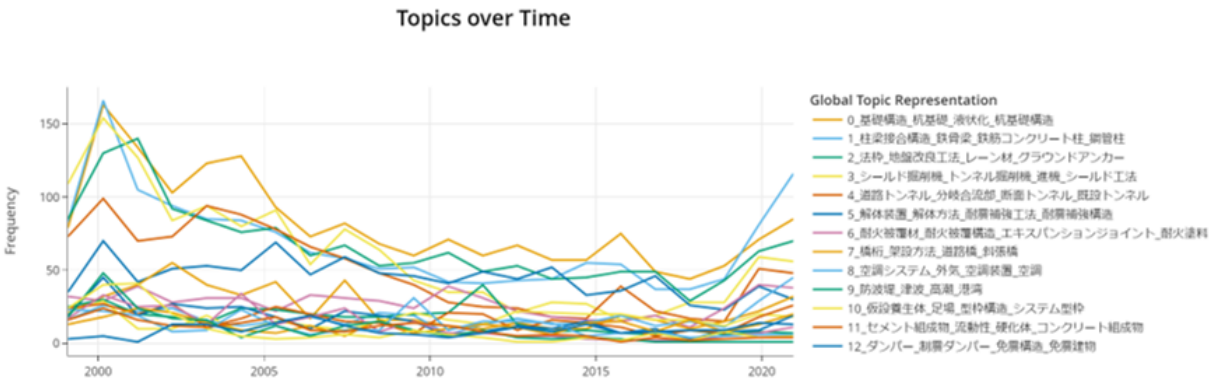
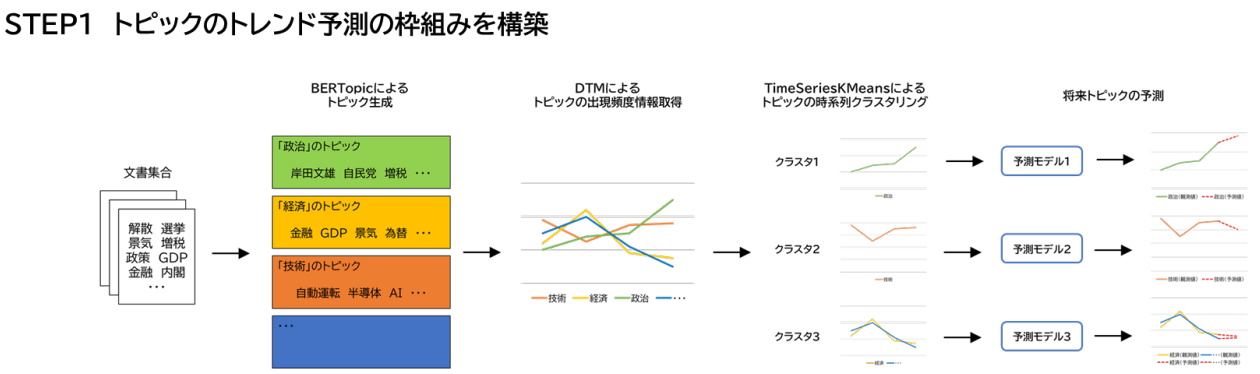


図 1: BERTopic の拡張機能である Dynamic Topic Modeling の結果



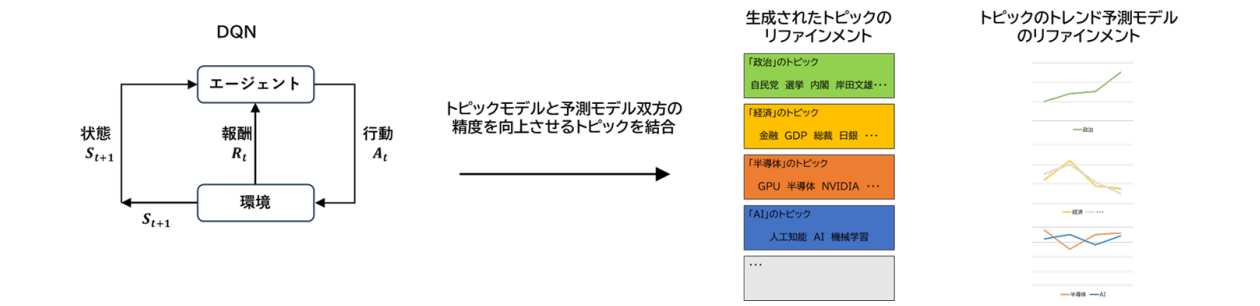


図 2: 提案手法のアーキテクチャ

（NTM）[5] [6] が存在する．また，トピックモデルの派生モデルも存在し，Blei らの Dynamic Topic Model（DTM）[8] は時間変化を扱った代表的なトピックモデルである．しかし，DTM は将来のトピックのトレンド予測を目的としておらず，筆者の調べではトピックの時間変化と将来のトピックのトレンド予測双方を学習としたモデルは存在しない．１，２節では，本研究に関連したトピックモデル手法を概説し，提案手法との関係を述べる．３節では本研究に使用する予測モデルを紹介する．

2.1 BERTopic: Neural topic modeling with a class-based TF-IDF procedure

LDA [4] や Non-Negative Matrix Factorization (NMF) [9] といった従来のモデルでは，文書を Bag of Words（BoW）として記述しており，トピックの生成にトークンの出現頻度のみしか考慮していない．そのため BoW の入力では，単語間の関係や文脈を考慮しないため，トピック抽出の際に単語間の関係や文脈を考慮することができないという問題がある．一方 LLM 登場以降、LLM のテキストエンベーディング技術をトピックモデルに導入する試みが盛んになっており，文脈を考慮したトピックの生成において大きな成果を示している [10]．Top2Vec [11] や HDBSCAN [12] では，単語とクラスタのエンベーディングを作成し，クラスタのセントロイド（重心）からトピック表現を抽出する手法であり，LLM のテキストエンベーディング技術を導入したトピックモデルの学習の有効性が確認された．一方セントロイドベースのクラスタリングアプローチを利用したトピックモデルは，クラスタの重心に近い単語がそ

のクラスタを最も代表すると仮定している．しかし，実際はクラスタの中心がクラスタの重心となるという仮定が成り立たない場合も多く，クラスタにおける単語の重要度は考慮されていないという問題が存在する．

BERTopic [7] は，Transformer とクラスベースの TF-IDF を活用して，クラスタにおける単語の重要度を考慮したトピックモデルであり，高い精度でトピックの学習を行うことが可能である．また，BERTopic の拡張機能である Dynamic Topic Modeling を利用することで，各タームにおけるトピックの出現頻度シーケンスを手に入れることができ，トピックのトレンドがこれまでどのように変化したきたかを確認することができる．図 1 は Dynamic Topic Modeling の結果である．

本研究では，トピックの学習精度の高さと時系列コーパスを扱うことが可能であるという理由から，トピックモデルの学習に BERTopic を採用する．

2.2 強化学習に基づくトピックモデルの学習

Reinforcement Learning for Topic Models [13] は，強化学習の枠組みを利用してトピックの学習を行うトピックモデル手法である．LDA [4] などの従来のトピックモデルは，元の文書の再構成に焦点を当てており，トピックの質を向上させるという本質の問題に取り組んでいない [14] [15] [16] [17].

Reinforcement Learning for Topic Models では，強化学習の枠組みを利用して，Coherence を最大化するようトピックを反復的に学習する．アーキテクチャの詳細は以下の通りである．はじめに，行動（Action）は高い Coherence スコアを持つ単語を文書中からサンプリングし，サンプリングされなかった単語を文書からマスクする．次に，単語間の潜在的なトピックの分布を状態（State）とし，トピックを学習する．最後に，報酬（Reward）に，トピックの一貫性を図る Coherence を利用することで，より高い Coherence スコアを持つ単語により高い重みを割り当てることが可能となる．

Reinforcement Learning for Topic Models はより高度なトピック抽出とトピックの一貫性の測定の両方において優れたパフォーマンスを示している．

本研究では，Reinforcement Learning for Topic Models の枠組みを応用して，トピックモデルとトピックのトレンド予測モデル双方のリファインメント手法を構築する．

2.3 時系列モデル

本研究では，BERTopic と予測モデルを使用することで，トピックのトレンド予測を行う．一方，トピックのトレンドはさまざまな因子によって決定づけられるため，トピックの出現頻度シーケンスをそのまま予測モデルで学習することは不適切な学習結果を招いてしまう恐れがある．例えば，異なるトレンドを持つトピックを同一の予測モデルで学習することは不適切な結果を招いてしまう可能性が高く，予測結果の信頼度も低いものとなる．

複雑なシステムをモデル化する際には，システムを部分的に分析する手法が一般的である．この手法は複数モデル学習（multiple model learning）と呼ばれる [18]．交通状況の予測などの分野では，多くの場合で時系列シーケンスのクラスタリングを事前ツールとして使用している例が存在する [19] [20].

本研究もこのアイデアを借用し，トピックのトレンド予測のプロセスをクラスタリング部分と予測部分に分割して実行する．はじめに BERTopic の拡張機能である Dynamic Topic Modeling から取得したトピックの出現頻度シーケンスデータを自然なクラスタに分割する．次にクラスタごとの予測モデルを利用することで，トピックのトレンド予測を実行する．クラスタはシステムをより正確に表すことができるため，予測精度と予測結果の信頼度を向上させることが期待できる．

本研究では，時系列クラスタリングモデルとして TimeSeriesKMeans [21]，予測モデルとして Prophet [22] を使用することで，トピックのトレンド予測を実施する．

3 提案手法

本研究では，BERTopic と予測モデルを使用することで，トピックのトレンド予測を行う．さらに，多目的強化学習の枠組みでトピックモデルとトピックのトレンド予測モデルを同時最適化する手法を提案する．具体的には，類似するトピックを結合することでトピックの更新（リファインメント）を実行する．多目的強化学習の枠組みでは，トピックペアの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）を状態（State）として扱い，状態（State）を受け取ったエージェントが報酬（Reward）が最大となるようトピックの結合判定を行う．この際，トピックモデルの報酬として Coherence と Topic Diversity を，トピックのトレンド予測モデルの報酬として予測値と実測値の乖離度（Euclidean Distance）を扱うことで，トピックモデルとトピックのトレンド予測モデル双方を考慮した同時最適化が可能となる．

提案手法の概要を示した図が図 2 である．提案手法は 2 つのステップで構築される．STEP1 では，トピックのトレンド予測の枠組みを構築する．STEP2 では，トピックモデルとトピックのトレンド予測モデルの同時最適化を行う．

3.1 トピックのトレンド予測の枠組みを構築

STEP1 では，トピックのトレンド予測の枠組みを構築する．BERTopic [7] は，Transformer を利用したトピックモデル手法であり，高い精度でトピックの推定を行うことが可能である．次に BERTopic の拡張機能である Dynamic Topic Modeling を使用し，各タームにおけるトピックの出現頻度シーケンスを取得する．

トピックのトレンド予測モデルは，2.3 節の説明の通り，時系列クラスタリングモデルの TimeSeriesKMeans [21] と予測モデルの Prophet [22] を使用して行う．具体的には，Dynamic Topic Modeling から取得したトピックの出現頻度シーケンスを，TimeSeriesKMeans を利用しクラスタに分割し，Prophet を利用することでクラスタごとに予測を行う．以上のプロセスを経て，トピックのトレンド予測の枠組みを構築する．

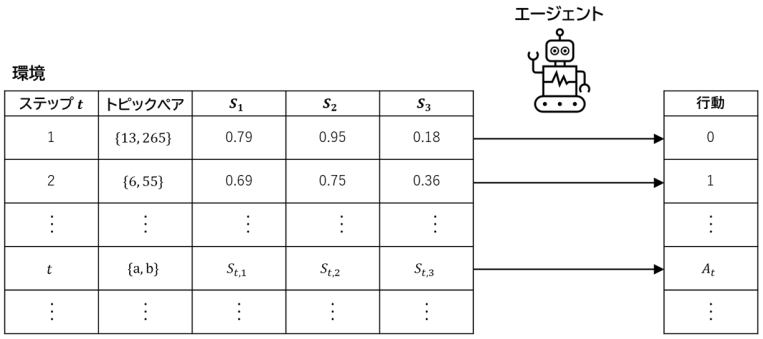


図 3: 環境とエージェントの関係

3.2 トピックモデルとトピックのトレンド予測モデルの同時最適化

STEP2 では，類似するトピックを結合することでトピックの更新（リファインメント）を実行する．トピックペアの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）を状態（State）として扱い，状態（State）を受け取ったエージェントが報酬（Reward）が最大となるようトピックの結合判定を行う．トピックモデルの報酬として Coherence と Topic Diversity を，トピックのトレンド予測モデルの報酬として予測値と実測値の乖離度（Euclidean Distance）を扱うことで，トピックモデルとトピックのトレンド予測モデル双方を考慮した同時最適化が可能となる．

以下のサブセクションにて，DQN による多目的強化学習の枠組みの詳細を説明する．

3.2.1 状態 (State)

状態 (State) は，3つのトピックペアの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）で構成される．Topic Diversity を $S_{t,1}$，Embeddings Similarity を $S_{t,2}$，Trend Divergence を $S_{t,3}$ で表すと状態 S_t は以下のようにあらわされる．なお，t は各学習のステップ値である．

$$S_t = \{S_{t,1}, S_{t,2}, S_{t,3}\} \tag{1}$$

Topic Diversity は，トピック同士の構成単語（上位 10 語）に同一単語がどれだけ含まれるかを測るための指標である．同一単語が複数含まれる場合 Topic Diversity は 0 に近づき，同一単語が一つも含まれない場合 Topic Diversity は 1 となる．Topic Diversity が 0 に近いほど類似しているトピックペアである．一般に Topic Diversity はトピックモデルから生成されるすべてのトピックを対象に計算されるが，本件研究における強化学習の環境ではトピックのペアに対して Topic Diversity を計算する．

Embeddings Similarity は，トピックペアの Embedding の cosine 類似度である．BERTopic は sentence-transformers を利用してトピックを生成しているため，各トピックは Embedding をもつ．Embedding の cosine 類似度が 0 に近いほど類似しているトピック同士である．

Trend Divergence は，トピックペアのトレンド乖離スコアの平均である．トピックの出現頻度シーケンスを p，TimeSeriesKMeans で時系列クラスタリングの中心線を c とする．なお，c はそれぞれのクラスタに所属するトピックの出現頻度シーケンスを p の平均で構成される．p と q のユークリッド距離を $d(p, q)$ とすると Trend Divergence は以下のような式で表すことができる．

$$\frac{d(p_a, c_a) + d(p_b, c_b)}{2} \tag{2}$$

Trend Divergence が高いほど，トピックのトレンド把握の際にノイズとなり，Trend Divergence が低いほど，トピックのトレンドの中心である．

3.2.2 行動 (Action) とエージェント

行動 (Action) では，トピックのペアを結合するか否か決定する．α_0 をトピックを結合しないという行動，α_1 をトピックを結合するという行動とすると行動空間は以下のようにあらわされる．

$$\mathbb{A} = \{\alpha_0, \alpha_1\} \tag{3}$$

エージェントは，トピックペアの類似度 ($S_{t,1}$, $S_{t,2}$) とトピックのトレンド乖離度（$S_{t,3}$）双方を考慮して行動を決定する．$S_{t,1}$, $S_{t,2}$ に基づいてトピックを結合するか否か判定することで，よりよいトピックの生成が可能となる．また，$S_{t,3}$ に基づいてトピックを結合するか否か判定することで，トピックのトレンド予測精度の向上が期待される．高い報酬が得られる行動 (Action) に，より高い確率を割り当てられ，エージェントはその確率に基づいて行動 (Action) を選択する．エージェント行動は以下のようにあらわされる．ϵ の確率でエージェントがランダムな行動 (Action) をとるようにすることで，探索と活用の双方を考慮できるよう設計する．

$$\begin{cases} \alpha \overset{\mathrm{R}}{\leftarrow} \mathbb{A} & \text{if random} < \epsilon \\ \underset{\alpha}{\operatorname{argmin}} f(\text{State}) & \text{else} \end{cases} \tag{4}$$

3.2.3 環境 (Environment)

環境 (Environment) ではエージェントが結合判定をするトピックペアの順序を提供する．環境 (Environment) は，トピックペア a, b とトピックペアの結合判定指標 $S_t = \{S_{t,1}, S_{t,2}, S_{t,3}\}$ を含んでいる．エージェントは各ステップ t ごとに環境 (Environment) から S_t が与えられ，トピックペア a, b を結合する

か否か判定する．環境 (Environment) がエージェントに与える S_t の順序は，本研究では学習の対象とはせず，あらかじめトピックの類似度等を参考に作成する．エージェントは，環境から与えられた状態 S_t に基づいて，以下のように行動 A_t を決定する．ただし，f_a は DQN ネットワークを指す．

$$A_t = \underset{a}{\operatorname{argmin}} f_a(S_t) \tag{5}$$

環境とエージェントの関係は，図 3 のようになる．

3.2.4 報酬 (Reward)

報酬 R_t は，トピックモデルの報酬 $R_{t,1}$ とトピックのトレンド予測モデルの報酬 $R_{t,2}$ で構成される．トピックモデルの報酬 $R_{t,1}$ はさらに Coherence の報酬 $Q_{t,1}$ と Topic Diversity の報酬 $Q_{t,2}$ で構成される．α は Coherence と Topic Diversity の報酬の調整を担い，β はトピックモデルとトピックのトレンド予測モデルの報酬の調整を担う．結合しないという判定をエージェントがとった場合は，報酬 R_t は 0 となる．以上の定義より，報酬は以下の式となる．

$$\begin{cases} R_t = 0 & \text{if } A_t = 0 \\ R_t = \beta \cdot R_{t,1} + (1-\beta) \cdot R_{t,2} & \text{else} \end{cases} \tag{6}$$

ただし，

$$R_{t,1} = \alpha \cdot Q_{t,1} + (1-\alpha) \cdot Q_{t,2} \tag{7}$$

はじめにトピックモデルの報酬 $R_{t,1}$ の説明をする．Coherence と Topic Diversity をトピックモデルの報酬として扱うことで，トピックモデルのリファインメントが可能となる．$Q_{t,1}$ はトピックペアの結合前後で Coherence が増加したか否かで決定する．$Q_{t,2}$ はトピックペアの結合前後で Topic Diversity が増加したか否かで決定する．以上より，$Q_{t,1}$ と $Q_{t,2}$ は以下のように決定する．

$$\begin{cases} Q_{t,1} = 1 & \text{if Coherence}_t > \text{Coherence}_{t-1} \\ Q_{t,1} = -1 & \text{else} \end{cases} \tag{8}$$

$$\begin{cases} Q_{t,2} = 1 & \text{if Topic Diversity}_t > \text{Topic Diversity}_{t-1} \\ Q_{t,2} = -1 & \text{else} \end{cases} \tag{9}$$

トピックのトレンド予測モデルの報酬 $R_{t,2}$ は，トピックペアの結合前後で，予測値と実測値の乖離度 $Q_{t,3}$ が減少したか否かで決定する．トピックの出現頻度シーケンスを p，トピックの出現頻度シーケンスの予測値を $\hat{p}$，p と q のユークリッド距離を $d(p,q)$ とすると $R_{t,2}$ は以下のような式で表すことができる．

$$\begin{cases} R_{t,2} = 1 & \text{if } Q_{t,3} > 0 \\ R_{t,2} = -1 & \text{else} \end{cases} \tag{10}$$

ただし，

$$Q_{t,3} = \frac{d(p_a, \hat{p}_a) + d(p_b, \hat{p}_b)}{2} - d(p_{ab}, \hat{p}_{ab}) \tag{11}$$

4 実　　験

実験では，はじめに DQN による多目的強化学習の枠組みでモデル学習が適切に行われているか否かを確認した．次に，トピックモデルとトピックのトレンド予測モデルの同時最適化を行うことで，両モデルを考慮した学習が行えているか否かを確認した．前者に関しては，DQN 学習前後の Coherence，Topic Diversity，予測値と実測値の乖離度（Euclidean Distance）を比較することでトピックモデルとトピックのトレンド予測モデルの性能がそれぞれ向上するかを検証した．後者に関しては，トピックモデルのみで最適化を行った場合と，トピックモデルとトピックのトレンド予測モデルの同時最適化を行った場合で実験結果を比較するこで，両モデルを考慮した学習が行えているか否かを検証した．

4.1 データセット

本研究では，日本語の時系列コーパスとして「建築分野に関連した特許文書（1999 年から 2022 年）」を使用した．総文書数は，21352 件である．

4.2 性能評価指標

提案手法は，トピックモデルとトピックのトレンド予測モデル双方の精度を検証する必要がある．トピックモデルの性能評価指標として，Coherence u-mass と Topic Diversity を導入した．また，トピックのトレンド予測モデルの性能評価指標として，Euclidean Distance を導入した．いずれの評価指標も 3.2.4 節で説明したトピックモデルの報酬とトピックのトレンド予測モデルの報酬と同一である．

Coherence u-mass は，トピックの品質・一貫性を評価する指標であり，以下の式で定義される．

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{12}$$

ここで，$V^{(t)} = v_1^{(t)}, ..., v_M^{(t)}$ はトピック t の上位 M 個の単語集合，$D(v)$ は単語の出現文書数，$D(v_1, v_2)$ は単語の共起文書数である．

次に，Topic Diversity は以下のようにあらわされる．Topic Diversity はトピック同士の構成単語 (上位 10 語) に同一単語がどれだけ含まれるかを測るための指標である．同一単語が複数含まれる場合 Topic Diversity は 0 に近づき，同一単語が一つも含まれない場合 Topic Diversity は 1 となる．Topic Diversity が 0 に近いほど冗長なトピックを生成しており，1 に近いほど多様なトピックを生成していることを示す．

Euclidean Distance は，予測値と実測値のユークリッド距離 (乖離度) であり，予測値と実測値が近い場合，スコアは 0 に近づく．本研究では，使用した時系列コーパス「建築分野に関連した特許文書（1999 年から 2022 年）」の最終年 2022 年のトピックの出現頻度をテストデータとして扱った．

4.3 実験設定

報酬で導入した α と β を複数用意して交差検証を行う．実験で使用した Q 関数は，10 の隠れ層サイズをもつネットワーク層 3 層で構成され，活性化関数として ReLU 関数を使用した．学習の最適化は，学習率が 0.0005 の Adam で実行した．そのほかのハイパーパラメータの設定として，学習回数を 50，環境のステップ数を 63，DQN のバッファ数を 50，バッチ数を 16，エージェントがランダムな行動をとる確率 ϵ を 0.1 と設定した．

4.4 実験結果

本節はじめで説明した通り，本研究では二つの検証事項を確認するための実験を行う．検証事項の一つ目は，DQN の枠組みでモデル学習が適切に行われているか否かである．検証事項の二つ目は，トピックモデルとトピックのトレンド予測モデルの同時最適化を行うことで，両モデルを考慮した学習が行えているか否かである．以下，各検証事項の実験結果を紹介する．

4.4.1 検証事項 1

検証事項 1 では，DQN の枠組みでモデル学習が適切に行われているか否かを検証する．方法としては，DQN 学習前後の Coherence， Topic Diversity，予測値と実測値の乖離度（Euclidean Distance）を比較することでトピックモデルとトピックのトレンド予測モデルの性能がそれぞれ向上するかを確認した．

実験の結果は以下の通りである．なお，TC は Topic Coherence u-mass，TD は Topic Diversity，ED は Euclidean Distance を表す．また，以下の結果は α が 0.3，β が 0.5 のものを採用した．

表 1: 検証事項 1 の結果

	TC	TD	ED Train	ED Test
Before	0.65687	0.97	0.02081	0.08918
After	**0.66019**	**0.98116**	**0.01958**	**0.08025**

表 1 を確認すると，Coherence，Topic Diversity，Euclidean Distance いずれのスコアも DQN の学習後に向上していることが確認できる．

図 6，4，5 は，報酬で導入した α と β を複数用意して交差検証を行った結果である．乖離度は結合後にどれだけ減らせたかを加算しているため，上昇すればするほど精度が向上していることを示している．いずれの場合も学習回数を経るほどモデルの精度が向上していることが確認できる．

以上の結果より，提案手法を用いることで，トピックモデルとトピックのトレンド予測モデルの精度がそれぞれ向上することが確認できた．

4.4.2 検証事項 2

検証事項 2 では，トピックモデルとトピックのトレンド予測モデルの同時最適化を行うことで，両モデルを考慮した学習が行えているか否かを検証する．方法としては，トピックモデルのみで最適化を行った場合と，トピックモデルとトピックのトレンド予測モデルの同時最適化を行った場合で実験結果を比較するこで，両モデルを考慮した学習が行えているか否かを確認した．

トピックモデルのみで最適化を行う場合では，3.2.1 節で紹介した 3 つのトピックペアの類似度（Topic Diversity，Embeddings Similarity，Trend Divergence）のうち，トピックのトレンド予測モデル部分に関係する Trend Divergence を除いたトピックペアの類似度（Topic Diversity，Embeddings Similarity）を状態として使用した．また，トピックモデルのみで最適化を行う場合は，トピックのトレンド予測モデル部分に関係する報酬を取り除くため報酬のパラメータ β を 0 とした．

実験の結果は以下の通りである．Single はトピックモデルのみで最適化を行ったモデル，Joint は同時最適化を行ったモデル (提案手法) であり，Joint は表 1 の After と同一である．また，Single の報酬パラメータ α は 0.3，β はトピックのトレンド予測モデル部分に関係する報酬を取り除くため 0 とした．Joint の報酬パラメータはそれぞれ α が 0.3，β が 0.5 である．

表 2: 検証事項 2 の結果

	TC	TD	ED Train	ED Test
Single	**0.66031**	0.97926	0.01994	0.08469
Joint	0.66019	**0.98116**	**0.01958**	**0.08025**

Single モデルと Joint モデル双方のトピックモデルの精度を比較してみると，Coherence は Single モデル，Topic Diversity は Joint モデルの方が高い結果となった．また，Euclidean Distance のスコアは Joint モデルの方がより高い精度を持つことが確認できた．Single モデルと Joint モデルで結合されたトピックのうち 5 件を表示したものが図 7 と図 8 である．

4.5 実験の考察

実験結果より，提案手法を用いることでトピックモデルとトピックのトレンド予測モデルの精度がそれぞれ向上することが確認できた．また，トピックモデルとトピックのトレンド予測モデルの同時最適化の枠組みを用いることで，両モデルを考慮したトピックのリファインメントが行われていることが定量的には確認できた．

一方で，図 7 と図 8 を確認してみると，Single モデルと Joint モデル双方で直観的には類似していないがトピックが結合されている事例が確認できる．この理由としては，2 つ考えられる．一つ目が，トピック数が減ること自体がトピックモデルの報酬となってしまい適切な報酬の設計がされていない点である．二つ目が，人間がトピック同士が直観的に類似しているか判断する基準と Topic Diversity との間にズレが生じている点である．例えば，Topic Diversity では「養生」と「養生方法」,「空調システム」と「空調」は全く別の単語として認識されてしまうといった問題が生じる．これらの問題に対処するため，再度報酬の設計と Topic Diversity の調整を行う必要がある．

5 考察と結論

本研究では，BERTopic と予測モデルを使用することで，ト

ピックのトレンド予測を行い，さらに，DQN による多目的強化学習の枠組みでトピックモデルとトピックのトレンド予測モデルを同時最適化する手法を提案した．実験を通して，提案手法を用いることでトピックモデルとトピックのトレンド予測モデルの精度がそれぞれ向上することが確認できた．また，トピックモデルとトピックのトレンド予測モデルの同時最適化の有効性が定量的には確認できた．

一方で，図 7 と図 8 を確認してみると，直観的には類似していないがトピックが結合されている事例が確認できる．この理由としては，２つ考えられる．一つ目が，トピック数が減ること自体がトピックモデルの報酬となってしまい適切な報酬の設計がされていない点である．二つ目が，人間がトピック同士が直観的に類似しているか判断する基準と Topic Diversity との間にズレが生じている点である．今後の課題として，報酬の設計と Topic Diversity の調整を行うことでこれらの問題を解決し，さらなる実験と考察を行う必要があると考えられる．

謝　　辞

本研究の一部は　JST CREST(JPMJCR22M2)，JSPS 科研費 (22H00508，22K17944)　および株式会社熊谷組の支援による．

文　　献

[1] Shuo Xu, Liyuan Hao, Guancan Yang, Kun Lu, and Xin An. A topic models based framework for detecting and forecasting emerging technologies. *Technological Forecasting and Social Change*, Vol. 162, p. 120366, 2021.

[2] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, pp. 29–41, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[3] Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol.*, Vol. 3, No. 2, feb 2012.

[4] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, Vol. 14. MIT Press, 2001.

[5] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference, 2017.

[6] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[8] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 113–120, New York, NY, USA, 2006. Association for Computing Machinery.

[9] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence, 2011.

[10] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and optimizing topic models is simple! In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 263–270, Online, April 2021. Association for Computational Linguistics.

[11] Dimo Angelov. Top2vec: Distributed representations of topics, 2020.

[12] Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, September 2020.

[13] Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. Neural topic model with reinforcement learning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3478–3483, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[15] Dallas Card, Chenhao Tan, and Noah A. Smith. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[16] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

[17] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations, 2017.

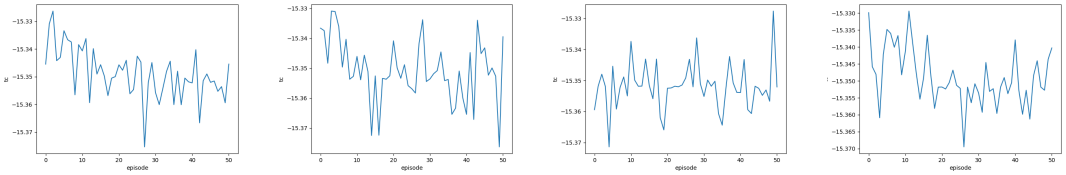
[18] Nikolas Julio, Ricardo Giesen, and Pedro Lizana. Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. *Research in Transportation Economics*, Vol. 59, pp. 250–257, 2016. Competition and Ownership in Land Passenger Transport (selected papers from the Thredbo 14 conference).

[19] Dongjoo Park and Laurence R. Rilett. Forecasting multiple-period freeway link travel times using modular neural networks. *Transportation Research Record*, Vol. 1617, No. 1, pp. 163–170, 1998.

[20] Andres Ladino, Alain Kibangou, Hassen Fourati, and Carlos Canudas de Wit. Travel time forecasting from clustered time series via optimal fusion strategy. In *2016 European Control Conference (ECC)*, pp. 2234–2239, 2016.

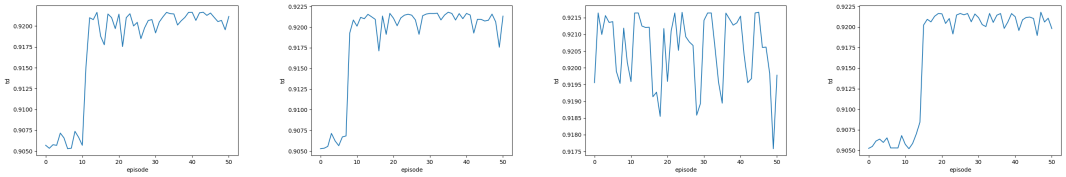
[21] Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, and Shaokai Wang. Time series k-means. *Inf. Sci.*, Vol. 367, No. C, p. 1–13, nov 2016.

[22] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, Vol. 72, No. 1, pp. 37–45, 2018.



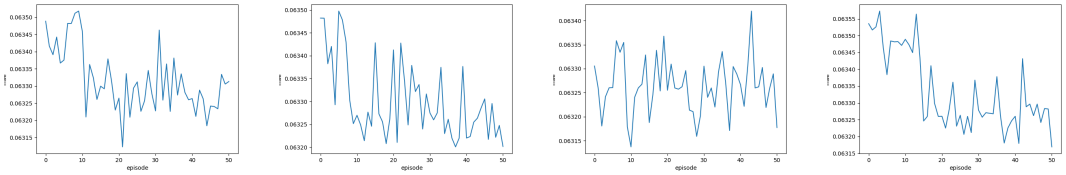
(a) α 0.3, β 0.3　(b) α 0.3, β 0.7　(c) α 0.7, β 0.3　(d) α:0.7, β:0.7

図 4: Coherence の比較



(a) α 0.3, β 0.3　(b) α 0.3, β 0.7　(c) α 0.7, β 0.3　(d) α:0.7, β:0.7

図 5: Topic Diversity の比較



(a) α 0.3, β 0.3　(b) α 0.3, β 0.7　(c) α 0.7, β 0.3　(d) α:0.7, β:0.7

図 6: Euclidean Distance の比較

	Topic 1	Topic 2
26	空調システム_空調_室内_空調装置_外気	強度コンクリート_圧縮強度_剥離_放電破砕装置_コンクリート構造物
14	コンクリート杭_鋼管杭_杭頭処理方法_杭頭処理工法_杭頭処理	コンクリート体_コンクリート製防護壁_サイトｐｃａ化工法_アン...
35	風力発電装置_風力_発電_ブレード_発電装置	マイクログリッド_電力システム_考え方_小型分散型電源_電気供給系
29	養生方法_養生装置_養生_養生シート_コンクリート養生方法	養生方法_養生_コンクリート用型枠パネル_コンクリート養生型枠...
54	廃棄物処分場_水構造_濁水_水シート_脱水ケーキ	耐震補強構造_耐震性_耐震補強工法_耐震補強方法_耐震性能

図 7: Single モデルで結合されたトピック

	Topic 1	Topic 2
2	本明細書_設済_新設コンクリート_上床版_空気圧送時	機器据付架台_ペデスタル_プレキャストコンクリートスラブ_配筋...
38	鉄骨梁_鉄筋コンクリート柱_接合部構造_梁床部材接合構造_柱梁...	空隙_岩盤_注入構造_注入_亀裂
25	ジョイント_エキスパンション_エキスパンションジョイント構造_...	梁部材同士_プレキャスト鉄筋コンクリート梁部材同士_ｈ型鋼梁_...
21	廃棄体_緩衝材_放射性廃棄物_廃棄体パッケージ_緩衝体	構面内_耐震補強構造_耐震補強壁_複合梁架構_耐震補強要素
43	煙突_解体装置_塔状構造物_屋根_屋根面	建設機械_制限表示ゲート_運転席_作業員_警報装置

図 8: Joint モデルで結合されたトピック

Human-in-the-Loop ニューラルトピックモデリング

米谷　亜夕†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学 情報学群　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †s2213614@u.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp

あらまし　トピックモデルは文書集合に含まれる潜在的なトピックを抽出するモデルであり，テキストマイニングの手法の一つとして知られている．しかしながらトピックモデルで抽出されるトピックは， 人間にとって必ずしも好ましい分析結果になるとは限らない．本研究では，ニューラルトピックモデルに対して， Human-in-the-Loop によって人間からのフィードバックをモデルに反映する手法を提案する．本論文では特に，トピックへの単語の追加・削除について検討する．ニューラルトピックモデルに対して Human-in-the-Loop を適用することで，勾配法に基づくモデルの最適化が可能になり，モデルの拡張が容易になるという利点が存在する．実験では，ベースラインとするトピックモデルと本研究で提案する Human-in-the-Loop ニューラルトピックモデルを Perplexity と人間のフィードバックの反映率を評価指標として用いて比較実験する．結果として，ベースラインとした LDA と比較して提案手法はより良い Perplexity を獲得し，反映率は 0.9 以上のスコアを獲得することができた．

キーワード　トピックモデル，Human-in-the-Loop，テキストマイニング, 情報抽出

1　はじめに

トピックモデルは潜在的なトピックを抽出するモデルでテキストマイニング手法の一つとして知られていおり，様々な領域に応用されている．例えば，文書分類 [1] やテキスト分析 [2]，推薦システム [3] のような分野では必要不可欠な技術である．トピックモデルはこれまでいくつか提案されてきた．潜在意味解析（Latent Semantic Indexing: LSI）[4] や確率的潜在意味解析（Probabilistic Latent Semantic Indexing: PLSI）[5]，潜在ディリクレ配分法（Latent Dirichlet Allocation: LDA）[6] は特に代表的なトピックモデルである．

ここでは，本研究と関連の深いトピックモデルである LDA について紹介する．LDA は，文書の確率的生成モデルとして提案された，よく用いられるトピックモデルの 1 つである．LDA では，各文書中に潜在的なトピックがあると仮定し，統計的に共起しやすい単語集合を潜在的なトピックとする確率変数を用いて表現する．これにより，データセット中に存在する潜在的なトピックを抽出でき，データセット全体を要約する表現が得られる．また，LDA は文書の形式に限らず，共起情報があるようなデータ全般に適応することができるモデルである．そのため，自然言語処理のみならず，画像処理や音声処理，情報検索などのさまざまな分野で応用されている．

LDA に対して，ニューラルトピックモデル（Neural Topic Model: NTM）[7] [8] [9] は，変分オートエンコーダ（Variational Auto Encoder: VAE）[10] によって潜在的なトピックを推定することができるモデルである．VAE は変分ベイズ推論とニューラルネットワークを組み合わせたモデルであり，深層学習による生成モデル 1 つである．VAE によってより複雑な推論を行うことが可能な上，確率的勾配法でモデルの最適化が行えることやニューラルなモデルとの組み合わせなどの拡張が容易である．しかし， LDA や NTM を用いたトピックの抽出においては，人間にとって必ずしも好ましい分析結果になるとは限らない．

先行研究において，トピックモデルに Human-in-the-Loop を組み込んだ手法である Human-in-the-Loop トピックモデル（Human-in-the-Loop Topic Model: HL-TM）が注目を集めている [11] [12] [13]．Human-in-the-Loop とは，人間を何らかのシステムに組み込むこのによって，人間にとってより好ましい結果に近づける方法である．トピックモデルに Human-in-the-Loop を取り入れる場合，抽出されたトピックが人間にとって好ましくないものにならないように，トピックに単語を追加・削除する，といった操作を人間が加えることによってシステムを補助する．既存研究では，確率的生成モデルに基づくトピックモデルをベースにしたものが提案されているが， NTM に基づく Human-in-the-Loop トピックモデルは提案されていない．

そこで，本研究では NTM の学習のプロセスに Human-in-the-Loop 組み込んだ Human-in-the-Loop ニューラルトピックモデリング（Human-in-the-Loop Neural Topic Modeling: HL-NTM）を提案する．本研究では特に，Human-in-the-Loop によって，人間が必要と考える単語をトピックに追加する操作と，人間が不必要と考える単語をトピックから削除する方法について検討する．HL-NTM は人間のフィードバックに対する重みのパラメータをもち，このパラメータを調整することで，人間のフィードバックをどれだけ取り入れるかを調整可能である．さらに， NTM に基づく手法を構成することで，トピックモデルを用いたニューラルネットワークに基づく手法に対して本手法を適用することが可能になる．

実験では Livedoor ニュースコーパスを用い，トピックモデルの予測精度を示す指標である Perplexity と，モデルに反映

されたフィードバック数と反映されなかったフィードバック数に基づいて算出される反映率を用いて評価を行った．実験の結果， HL-NTM が既存の手法と比較してモデルの予測精度を維持しながら，優れた反映率のスコアを得られることが示された．

本研究の貢献をまとめると，以下の通りである．

（1） 本論文では NTM に Human-in-the-Loop 組み合わせた構造を持つ， Human-in-the-Loop ニューラルトピックモデリングを提案する．

（2） ベースラインとした LDA と比較して，より優れた Perplexity のスコアを保ちながら Human-in-the-Loop の構造を組み込むことが可能であることを，ランダムな操作を行うユーザを想定した実験によって示した．

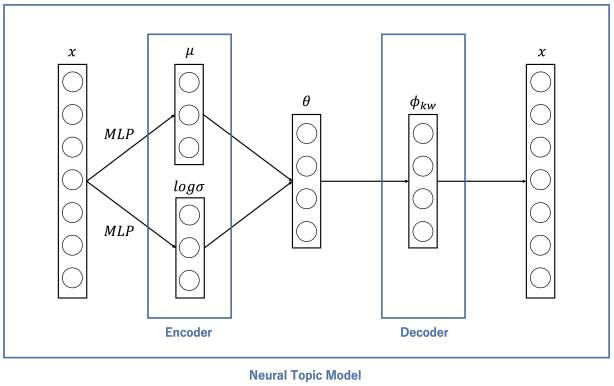


図 1: NTM のアーキテクチャ図．

2 関 連 研 究

本節では，トピックモデルに関する内容の中でも，特に本研究に関係が深い LDA と NTM について説明する．加えて，Human-in-the-Loop とトピックモデルを組みわせたモデルの提案をしている論文をいくつか挙げ，その内容について触れる．また，各項で紹介するトピックモデルに対して，本研究との関連性と新規性について改めて述べる．

2.1 トピックモデル

トピックモデルとは，自然言語処理の分野で用いられる潜在意味解析手法のひとつで文書集合のトピックを分析する手法である [4] [5] [6]．トピックモデルでは，文章を複数の単語の集合であると捉え，それらの単語の共起性に着目して文章をいくつかのクラスに分類する．トピックモデルは，文章が複数の潜在的なトピックから構成され，それらは確率的に生成されると仮定し，単語がそのトピック分布に従って出現するという構造になっている．

最も代表的なトピックモデルとしては，潜在ディリクレ配分法（LDA）[6] が知られている．LDA は，これまで拡張した手法がさまざま提案されてきている．例えば，トピック間の相関をモデル化した Correlated Topic Model（CTM）[14] や文書中の単語の順序も考慮できるように系列データを扱える隠れマルコフモデル（Hidden Markov Model: HMM）と LDA を組み合わせた HMM-LDA [15] などが挙げられる．

しかし， LDA のような統計的手法に基づくトピックモデルに新しく Human-in-the-Loop を組み合わせて推論をするためには，モデルをもう一度構築し直さなければならない．

ニューラルトピックモデル（NTM）は， VAE を元に構築されたトピックモデルである． LDA に対して NTM は，モデル構築の度に最適化アルゴリズムの導出を行う必要がなく，勾配法に基づいてモデルの最適化を行うことが可能である．また，ニューラルなモデルであり，他のニューラルな手法での拡張が容易なため， LDA の問題点を解決した． NTM に関しての詳細な説明は，3.1 項で述べる．

2.2 Human-in-the-Loop トピックモデル

Human-in-the-Loop トピックモデル（HL-TM）において，既存の研究で提案されている HL-TM についてここでは 3 つ紹介する．

HL-TM のシステム評価に関する研究として， Varun Kumar 氏らによって提案された HL-TM システム [11] では，ユーザの想定に近い出力になったかどうかを評価するための指標を導入している． Khan Muhammad Haseeb Ur Rehman 氏らによって提案されたキーフレーズによる文書分類を可能とした HL-TM [12] では，リファインメント操作が実際の応用上重要なドキュメント-トピックの関連にほとんど影響を与えないことを指摘し，ドキュメント-トピックの効果的な改善を意図したキーフレーズベースのリファインメント関数を提案している．また， Zheng Fang 氏らは，ユーザが全てのステップを比較・記録できるユーザフレンドリーなインターフェースと，ユーザがフィードバックを反映させるための新しいトピック提案機能を備えた，新しいインタラクティブな HL-TM を開発している [13]．

2.3 本研究の位置付け

これらの研究では，これまで提案されてきた統計的手法を用いたトピックモデルを元に Human-in-the-Loop を組み込んだトピックモデルとなっている．したがって，本研究の位置付けとしては， HL-NTM の基本的な構築とその有用性を示すことを目的とする．

3 事 前 知 識

本節では，前提知識となる NTM について詳しく説明する．NTM は本研究において， Human-in-the-Loop を組み込む上で重要な基礎知識である．

3.1 ニューラルトピックモデル

NTM は， 変分自己符号化器 （Variational Auto Encoder: VAE） を用いて潜在トピックを抽出するニューラルなトピックモデルである．VAE は，次のような仕組みでデータ x を生成する．まず，潜在変数 z が分布 $p(z)$ に従ってランダムに生成されるとする．この生成された z を条件付き分布 $p(x|z)$ に与えると，x が生成されるという仕組みである．また，条件付き分布 $q(z|x)$ を考え，条件付き分布 $p(z|x)$ との 2 つの分布の距離を

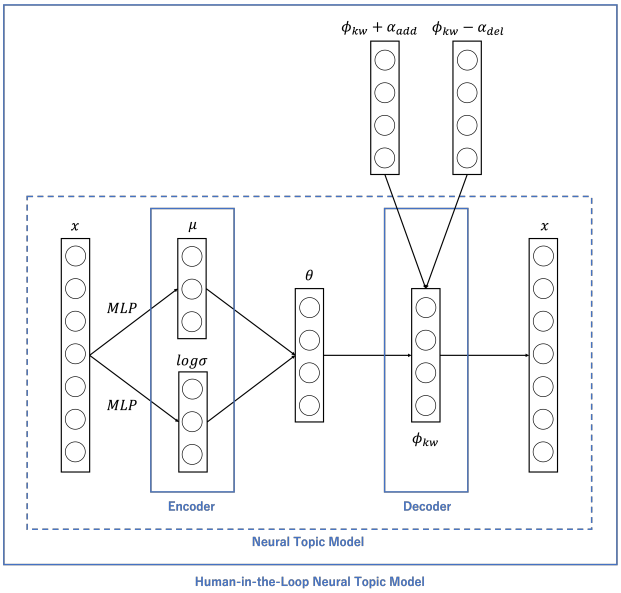


図 2: HL-NTM のアーキテクチャ図.

測る KL ダイバージェンスを考える．$q(z|x)$ はニューラルネットワークで表現する．損失関数は式（1）のように定義される．

$$\mathcal{L}_{NTM} = D_{KL}[q(z|x)||p(z)] - \mathbb{E}_{q(z|x)}[\log p(x|z)] \quad (1)$$

基本的に VAE と同じ形であり，KL ダイバージェンス損失とクロスエントロピー誤差の損失を組み合わせた関数である．条件付き分布 $q(z|x)$ から z を生成するプロセスはエンコーダ，条件付き分布 $p(x|z)$ から x を生成するプロセスはデコーダである．

ここで， NTM のアーキテクチャを図 1 に示す. アーキテクチャ図を元に， NTM の構造について詳しく説明する．文書集合のそれぞれの文書 x から BoW ベクトル $x_{bow} \in \mathbb{R}^V$ を生成し，これを入力とする．V は文書集合に含まれる単語の種類数（語彙数）である．NTM のエンコーダについて説明する．まず，$p(z|x)$ が標準正規分布 $q(z)$ に従うと仮定する．すると，$p(z|x)$ のパラメータ μ と $\log\sigma$ は，x_{bow} を入力として学習される．μ と $\log\sigma$ は式（2）である．

$$\mu = f_\mu(f_e(x_{bow})), \quad \log\sigma = f_\sigma(f_e(x_{bow})) \quad (2)$$

$f_*(\cdot)$ の全結合層は ReLU を活性化関数とする．

Decoder は，最終的に x_{bow} を再構成する．潜在変数 z は式（3）のようにサンプリングされる．

$$z \sim \mathcal{N}(\mu, \sigma^2) \quad (3)$$

トピック分布を式（4），x_{bow} の再構成は式（5）に示す．

$$\theta = softmax(f_\theta(z)) \quad (4)$$

$$x_{bow} \sim softmax(f_\phi(\theta)) \quad (5)$$

式（4）は潜在変数 z によって導出されるトピック分布である．（5）のパラメータ ϕ は θ を元に出力される，トピックごと単語分布となる．エンコーダと同様，$f_*(\cdot)$ の全結合層は ReLU を活性化関数とする．

4 提 案 手 法

本研究では，ニューラルトピックモデルに Human-in-the-Loop の構造を組み合わせた Human-in-the-Loop ニューラルトピックモデル（HL-NTM）を提案する．本節では， HL-NTM のアーキテクチャ， Human-in-the-Loop として導入する操作，モデルの学習方法について説明する．

4.1 HL-NTM のアーキテクチャ

本項では， HL-NTM のアーキテクチャについて説明する．HL-NTM のアーキテクチャは図 2 に示す．基本的には NTM と同じような構造をとる． NTM との違いは Human-in-the-Loop の機能として，デコーダ側に人間のフィードバックの反映に対する損失関数 $\phi_{kw}+\alpha_{add}$， $\phi_{kw}-\alpha_{del}$ が追加される点である．HL-NTM は事前学習済みの NTM をもとに，デコーダ側で人間のフィードバックを反映するようにファインチューニングを行う．デコーダ側に追加されるハイパーパラメータ等に関しては，4.2 項で解説する．

4.2 Human-in-the-Loop について

本研究では， HL-NTM における Human-in-the-Loop の操作として，ユーザによるトピックに対する単語の追加・削除を想定する．それぞれの操作についてより具体的に述べる．

（1） 単語の追加　　ユーザによって指定された，k 番目のトピックから削除する単語の集合を $\mathcal{W}_{add}^{(k)} = \{w_1, w_2, \cdots, w_n\}$ とする．

（2） 単語の削除　　ユーザによって指定された，k 番目のトピックに追加する単語の集合を $\mathcal{W}_{del}^{(k)} = \{w_1, w_2, \cdots, w_n\}$ とする．

4.3 学　　習

本項では， HL-NTM の学習方法について述べる．学習に用

いる損失関数は式（6）のように定義する．

$$\mathcal{L}_{HitL} = \mathcal{L}_{NTM} + \mathcal{L}_{add} + \mathcal{L}_{del} \tag{6}$$

ここで，$\mathcal{L}_{NTM}$ は NTM の損失関数である．また，$\mathcal{L}_{add}$ と $\mathcal{L}_{del}$ はそれぞれ単語の追加，単語の削除のための損失関数であり，式（7）,（8）のように定義される．

$$\mathcal{L}_{add} = \sum_{k=1}^{K} \sum_{w \in \mathcal{W}_{add}^{(k)}} \|\phi_{kw} + \alpha_{add}\| \tag{7}$$

$$\mathcal{L}_{del} = \sum_{k=1}^{K} \sum_{w \in \mathcal{W}_{del}^{(k)}} \|\phi_{kw} - \alpha_{del}\| \tag{8}$$

ここで，ϕ_{kw} は k 番目のトピックの w に対するデコーダのパラメータであり，α_{add}，α_{del} はハイパーパラメータである．ハイパーパラメータの数を大きくすればするほど，操作する単語に加わる重みが大きくなる．

よって，学習に用いる損失関数である式（6）は， NTM の損失，単語の追加を行った分の損失，単語の削除を行った時の損失を加算した式で構成されている．

5 実　　験

実験は，提案する HL-NTM と今回ベースラインとする LDA と NTM のスコアを比較して，モデルの予測精度を落とさずに Human-in-the-Loop の操作が反映可能か検証する．データセットには， livedoor ニュースコーパスを使用した．モデルの評価指標としては，予測精度を評価する Perplexity とモデルに反映されたフィードバック数と反映されなかったフィードバック数によって算出する反映率を用いることによって， HL-NTM の有用性を確かめる．

5.1 データセット

データセットは， livedoor ニュースコーパスを使用する．livedoor ニュースコーパスは， NHN Japan 株式会社が運営する livedoor ニュース記事を収集して作られ， 可能な限り HTML タグを取り除かれているものである． ニュース記事は，トピックニュース， Sports Watch， IT ライフハック， 家電チャンネル， MOVIE ENTER， 独女通信， エスマックス， livedoor HOMME， Peachy である． 収集時期はいずれも 2012 年 9 月上旬である．livedoor ニュースコーパスに対して janome Tokenizer を用いて， URL の除去， ストップワードの除去， 品詞の抽出の処理を行った． janome Tokenizer の処理によって除去した単語数は 928 語であり，今回学習に使用した総単語数は 20220 語である．

5.2 評 価 指 標

評価指標として，モデルの予測精度を測る Perplexity と，モデルに反映されたフィードバック数と反映されなかったフィードバック数によって算出する反映率を用いる．Perplexity は，式（9）のように定義される．

$$Perplexity = \exp\left\{-\frac{\sum_{d \in \mathcal{D}} \sum_{w \in d} \log p(w \mid \mathcal{M})}{\sum_{d \in \mathcal{D}} n_d}\right\} \tag{9}$$

ここで，$\mathcal{D}$ は文書集合，w は文書 d に含まれる単語，$p(w \mid \mathcal{M})$ は，モデル $\mathcal{M}$ による単語 w の生成確率である． Perplexity は小さいほど良い性能を示す指標となっている．ランダムな単語を返すモデルでは文書の語彙数，最小では 1 をとる． Perplexity が 1 の場合は，正しい単語の予測確率が 100% であるということになるため，単語は 1 択に絞られていると解釈することができる．

人間によるフィードバックのうち，出力されたトピック内の単語に対してフィードバックが反映された割合を 反映率（Reflection Rate）で評価する． 定義は（10）の通りである．

$$ReflectionRate = \frac{TP}{TP + FN} \tag{10}$$

ここで， TP はモデルに反映されたフィードバックの数， FN はモデルに反映されなかったフィードバックの数である．反映率は，実際の正解のうち，モデルによる予測がどれだけの割合で positive であったかを表現した指標である． 0.0（0%）-1.0（100%） の範囲で値をとり，大きいほど良い性能を示すことになる指標となっている．単語 1 つの操作を反映するフィードバックを例とすると，反映された場合は， $TP = 1$， $FN = 0$ であるため，反映率 $= 1$ となるため，単語が反映されたということがわかる．

5.3 実 験 設 定

実験比較には，提案手法の HL-NTM に対して，ベースラインとするトピックモデルは LDA と NTM である．これらのトピックモデルは前述の通り， Human-in-the-Loop を組み合わせていないシンプルなトピックモデルである．よって，Human-in-the-Loop の操作を組み込んでいないこれらのトピックモデルと比較して HL-NTM の Perplexity が悪化しないこと確認する．また， Perplexity を悪化させないように設定した状態のときのモデルに Human-in-the-Loop を組み込んだ場合のフィードバックの反映度合いを反映率によって示す．

各モデルに関する実験設定等について説明する．LDA はトピック抽出の実験を 10 回行う． LDA の Perplexity は，10 回の実験における平均値を結果とする． Perplexity はテストデータによって算出されたスコアである．また，今回の LDA による予測では，20 のトピックと各トピック内で出現する上位 10 単語が挙げられる．

NTM のパラメータに関して，隠れ層は 1000，トピック数は 20，トピック内で挙げられる単語数を 10 とする．学習に関するパラメータに関して，バッチサイズは 32，学習率は 0.001，エポック数は 150， spasity は 0.23 としている．それぞれ，学習のプロセスには，式（5）の f_ϕ の重みに対する L1 ペナルティが追加されている． L1 ペナルティを追加することによって，各トピック内の重みがゼロに近い単語は単語の情報量を削減し，学習の途中で f_ϕ の重みをスパースにしている．スパースにすることによって，重要度の低い単語や高い単語の重みをより極端な値にして，単語の情報を 2 分することができる．損失関数， Perplexity， L1 ペナルティに関しては，学習を行った結果のうち，平均的な学習のプロセスの出力結果を示す．

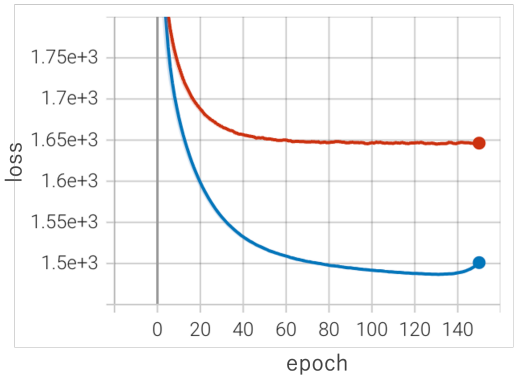


図 3: NTM による学習中の損失関数の推移．青は訓練データ，赤は検証データ．

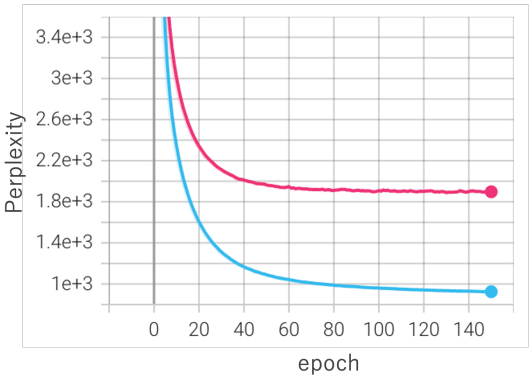


図 4: NTM による学習中の Perplexity の推移．青は訓練データ，赤は検証データ．

	Perplexity	
	LDA	NTM
平均	1841.9	1720.3

表 1: LDA と NTM の Perplexity 平均スコア．

NTM は，トピック抽出の実験を 10 回行う． NTM の Perplexity は，10 回の実験における平均値を結果とする． Perplexity はテストデータによって算出されたスコアである．また，今回の NTM による予測では，20 のトピックと各トピック内で出現する上位 10 単語が挙げられる．

HL-NTM は，事前学習を行なった NTM をベースに，デコーダの更新を行うように学習を行うことで Human-in-the-Loop の操作を導入する．事前学習を行なった NTM のパラメータは，前述したベースラインとする NTM の学習で設定したものとパラメータと同様に設定している．本実験では，単語の追加操作と単語の削除操作を追加する操作を Human-in-the-Loop とする．それぞれ，ランダムに選ばれたトピックにランダムに選ばれた単語を追加する操作とランダムに選ばれたトピックにランダムから選ばれた単語を削除する操作を分けて行う．ただし，ランダムに選出された単語がすでに選択されていた場合は単語の再選出を行う．単語の操作を行うのはそれぞれのトピックに出現する上位 10 単語に対してである．α_{add} のハイパーパラメータは 10，20，50，100，200 とし，α_{del} のハイパーパラメータは，0.1，0.2，0.5，1，3 とし，単語の操作数は 1，10，50，100，150 とする．トピックと単語の抽出の実験はそれぞれ 10 回行う． HL-NTM の Perplexity は，ハイパーパラメータ別に出力する． Perplexity はテストデータによって算出されたスコアである．反映率も同じようにして，ハイパーパラメータと単語の操作数を設定してハイパーパラメータ別に出力を行う．

5.4 実験結果

本項では， LDA， NTM， HL-NTM の各トピックモデルで行った実験結果を示し，その結果の特徴について述べる．

LDA， NTM で行なった 10 回の実験で出力された Perplexity を表 1 に示す． Perplexity は小数点第 2 位を四捨五入している．

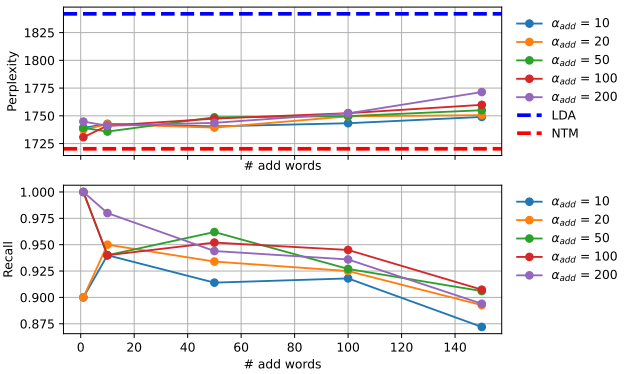


図 5: HL-NTM で単語の追加操作を行った時の Perplexity と反映率ののスコア．

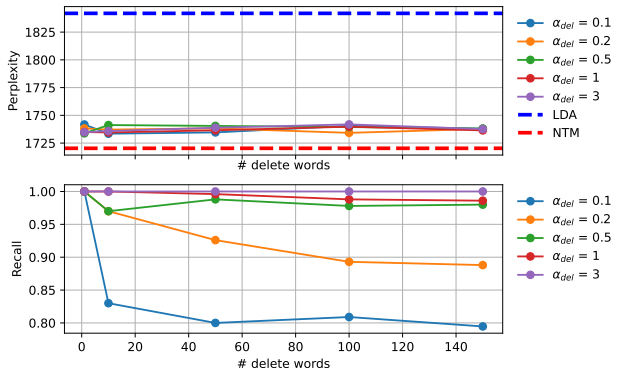


図 6: HL-NTM で単語の削除操作を行った時の Perplexity と反映率ののスコア．

NTM の学習プロセスの出力結果を，図 3，4 に示す．それぞれ，青は訓練データ，赤は検証データを入力とする学習のプロセスである． NTM の学習過程の loss の値を示す図 3 では，150 エポック付近になると損失関数の収束の傾向が見られる．図 4 も学習が進むにつれて同じように収束している．

HL-NTM の実験結果について，単語の追加操作と単語の削除操作を行った結果をそれぞれ示す．HL-NTM で単語の追加操作を行った実験結果として出力された Perplexity と反映率のスコアを図 5 に示す．それぞれ縦軸が Perplexity のスコアと反映率のスコア，横軸は単語の追加操作数である．今回設定し

たハイパーパラメータ数は 5 つのため，5 パターンの結果を出力している． Perplexity に関して， ベースラインとする LDA と NTM の Perplexity と HL-NTM の Perplexity を比較するために，グラフ上にそれぞれ破線で示している．

単語の追加操作を行った場合の Perplexity について，単語の操作数を増やすほどスコアが悪化する傾向が見られた．しかし，ハイパーパラメータの違いによってスコアの悪化度合いが大きく変化することは無かった．3 つのトピックモデルの Perplexity を比較すると， LDA の Perplexity が最も大きくなった． NTM の Perplexity が最も小さいため，最も良いスコアであるということがわかる．また， NTM と HL-NTM の Perplexity のスコアは近い値となり， HL-NTM はわずかに高くなった．反映率について，単語の操作数を増やすほどスコアが悪化する傾向が見られた．ハイパーパラメータを大きくするほど，反映率のスコアは 1 に近づき，加えた操作が出力された単語の結果に反映されていることがわかる．

HL-NTM で単語の削除操作を行った実験結果として出力された Perplexity と反映率のスコアを図 6 に示す． それぞれ縦軸が Perplexity のスコアと反映率のスコア， 横軸は単語の削除操作数である．単語の追加操作と同じように，今回設定したハイパーパラメータ数は 5 つのため，5 パターンの結果を出力している． Perplexity に関しても，ベースラインとする LDA と NTM の Perplexity と HL-NTM の Perplexity を比較するために，グラフ上にそれぞれ破線で示している．

単語の削除操作を行った場合の Perplexity について，単語の操作数を増やしてもスコアが悪化する傾向は見られなかった．加えて，ハイパーパラメータの違いによってもスコアの悪化度合いが大きく変化することは無かった．3 つのトピックモデルの Perplexity を比較すると， LDA の Perplexity が最も大きくなった． NTM の Perplexity が最も小さいため，良いスコアであるということがわかる．また， NTM と HL-NTM の Perplexity のスコアは近い値となり、こちらも単語の追加操作と同様に HL-NTM の Perplexity はわずかに高くなった．反映率について，単語の操作数を増やすほどスコアが悪化する傾向が見られた．ハイパーパラメータを大きくするほど，反映率のスコアは 1 に近づき，加えた操作が出力された単語の結果に反映されていることがわかる．

5.5 考　　察

本項では，5.4 項で述べた実験結果について考察する．

LDA の Perplexity と比較した場合， HL-NTM の Perplexity はスコアが大幅に下回った．よって， HL-NTM はモデルの予測精度を落とさずにフィードバック操作を反映させることができた． NTM と HL-NTM を比較すると， HL-NTM の Perplexity は NTM の Perplexity を僅かに上回る結果となった．このような結果になった理由としては， 事前学習済みの NTM からランダムに操作を加えて出力される単語に変更を加えたため，結果としては予測精度を落とすような操作を加えているということになるためであると考えられる．また，HL-NTM の Perplexity は NTM のスコアを大きく上回ってはいないため，予測精度をある程度保ちながら HL-NTM を構築することができたと考えることができる．

単語の追加操作の Perplexity は単語の削除操作の Perplexity と比較して， 単語の操作数が増えると悪化の傾向が見られた．今回は，ランダムなトピックとランダムな単語を選出して追加操作を行っているため，トピックに関係のない単語が追加されている可能性が高く，悪化しやすかったのではないかと考えられる．反対に，単語の削除操作の Perplexity が横ばいになったのは，そのトピックに関連のない単語が無理に追加されることはなく，ランダムに選択されたトピック内の上位に挙げられた単語が消えるだけであるためであると考えられる．下位に挙げられている比較的関連のある単語が上位に繰り上がることによって， Perplexity に影響を大きく与えることなく操作を加えることができた．

また，今回の実験では， Perplexity と反映率を評価指標として用いて比較実験を行った．今回の実験では， HL-NTM 以外のトピックモデルと Perplexity を比較することによってモデルの予測精度が比較的悪化していないことを示し，反映率によって Human-in-the-Loop がうまく機能していることを示した．しかし，これらの評価指標のみでは，実際に人間にとって好ましいスコアを得ることができたと断定することは難しい．なぜならば，どちらも人間にとって好ましい結果を出力できているかを直接，定量的に評価するための指標ではないからである．例えば，選出した単語がそのトピックに占める割合を考慮して学習を行うことでより好ましい結果が得られるような HL-NTM を構築することができるのではないかと考えられる．したがって，人間が求める結果に適応できたということを正確に評価するために，より適したモデルの構築や評価指標を模索する必要がある．

6 まとめと今後の展望

本研究は， NTM と Human-in-the-Loop を組み合わせた Human-in-the-Loop ニューラルトピックモデルを構築し，既存のトピックモデルと比較実験を行い， HL-NTM の有用性が確認できることを目指した．実験では，評価指標として Perplexity と反映率を用いて，比較評価実験を行った．結果として， LDA よりも優れた Perplexity であり， NTM と比較しても Perplexity を大きく悪化させず， Human-in-the-Loop として単語の追加操作と単語の削除操作をランダムで行う仕組みを NTM に加えることが可能であることを示した．

今後の HL-NTM で必要となる操作としては，語順の入れ替え，トピックの結合，トピックの分割，トピックの作成，文書の削除，文書の追加のような操作である．ユーザの操作選択肢を増やし，より人間の求めるトピックが抽出されるようにするには，これらの追加実装と実験が必要不可欠である．また，これらの操作を増やすにあたって，操作の組み合わせや操作の順番によってスコアが変化する可能性も新たに考えられる．

本研究ではランダムなユーザを想定した実験を行った．これに関して，本研究に実装可能である既存手法のシミュレーショ

ン実験について述べる．既存の HL-TM の手法には，Human-in-the-Loop を組み込む構造として，ランダムな操作を行うユーザだけではなく，人間にとって好ましいと考えられる操作を加えるように傾向を与えた「グッドユーザ」を定義してシミュレーション実験を行っているものが存在する [11]．よって，実際に人間による Human-in-the-Loop の操作を加えた実験を行う前実験として，このようなシミュレーション実験を行った後に，人間によるフィードバックを交えた実験を行うことがより好ましいと考えられる．

また，人間による Human-in-the-Loop の操作を加えた実験を行う場合についての考えを述べる．今回の実験では，ランダムに操作を行うユーザを想定した実験を行ったが，今後は，HL-NTM の実用性を視野に入れると，自然言語処理系の専門家を対象のユーザとして HL-NTM に操作をした場合との比較実験も行う必要がある．人間が操作するにあたって，ユーザフレンドリーな UI を設計することが，ヒューマンエラーによる誤操作を削減して，より正確なフィードバックを得ることができる可能性が考えられる．そのため，[13] で紹介されている UI ように，モデルがどのような出力して学習しているのかが理解しやすく，そのプロセスを追跡・変更できるような柔軟な機能と UI が求められる．

このように，シミュレーション実験と実際に人間による Human-in-the-Loop の操作を加えた実験のスコアを比較して実験を行うことによって，性能を評価する 1 つの指標になると考えられる．

謝　　辞

本研究の一部は JSPS 科研費 (JP22H00508，JP23H03405)，JST CREST(JPMJCR22M2) の支援を受けたものである．ここに謝意を示す．

文　　献

[1] D. Magatti, M. Faini, and F. Stella. A software system for topic extraction and document classification. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 1, pp. 283–286, Los Alamitos, CA, USA, sep 2009. IEEE Computer Society.

[2] M. Wang and P. Mengoni. How pandemic spread in news: Text analysis using topic model. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 764–770, Los Alamitos, CA, USA, dec 2020. IEEE Computer Society.

[3] T. Chen, J. Liu, B. Cao, Z. Peng, Y. Wen, and R. Li. Web service recommendation based on word embedding and topic model. In *2018 IEEE Intl Conf on Parallel amp; Distributed Processing with Applications, Ubiquitous Computing amp; Communications, Big Data amp; Cloud Computing, Social Computing amp; Networking, Sustainable Computing amp; Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, pp. 903–910, Los Alamitos, CA, USA, dec 2018. IEEE Computer Society.

[4] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, Vol. 41, No. 6, pp. 391–407, 1990.

[5] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[7] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 2410–2419. PMLR, 06–11 Aug 2017.

[8] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

[9] Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, and Shuming Shi. Topic-aware neural keyphrase generation for social media language, 2019.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[11] Varun Kumar, Alison Smith-Renner, Leah Findlater, Kevin Seppi, and Jordan Boyd-Graber. Why didn't you listen to me? comparing user control of human-in-the-loop topic models, 2019.

[12] Khan Muhammad Haseeb Ur Rehman and Kei Wakabayashi. Keyphrase-based refinement functions for efficient improvement on document-topic association in human-in-the-loop topic models. *Journal of Information Processing*, Vol. 31, pp. 353–364, 2023.

[13] Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. A user-centered, interactive, human-in-the-loop topic modelling system. *arXiv preprint arXiv:2304.01774*, 2023.

[14] John Lafferty and David Blei. Correlated topic models. *Advances in neural information processing systems*, Vol. 18, , 2005.

[15] Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. Integrating topics and syntax. *Advances in neural information processing systems*, Vol. 17, , 2004.

Pragmatic Approach to Automatic Topic Labeling with Zero-shot Classification

Zikai ZHOU† and Kei WAKABAYASHI††

† Graduate School of Comprehensive Human Science, University of Tsukuba
1–2 Kasuga, Tsukuba-shi, Ibaraki 355–8550, Japan
†† Faculty of Library, Information and Media Science, University of Tsukuba
1–2 Kasuga, Tsukuba-shi, Ibaraki 355–8550, Japan
E-mail: †s1813024@klis.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

Abstract In natural language processing, topic models are proposed to discern latent semantic structures within a text corpus. Once a topic model is applied, the extraction of topic terms becomes essential for interpreting the corpus's latent structure. The cognitive challenges of understanding a list of top terms lead to an interest in automatic topic labeling, which involves generating meaningful labels for topics and enhancing their interpretability. However, existing topic labeling methods often focus on individual topics, neglecting the overall corpus context and topic relationships. To address this, we propose a pragmatic approach that evaluates the ambiguity degree of topic labels through a listener model, which is instantiated by a zero-shot classification method. This model serves as a metric to assess the quality of label sets in the context of the entire corpus. Additionally, we introduce an automatic topic labeling pipeline employing combinatorial optimization to refine label sets for a given topic distribution. Experimental results demonstrate the correlation between our pragmatic approach and human evaluation, showcasing the effectiveness of our pipeline in automatic topic labeling. Our pragmatic approach emphasizes the importance of considering the holistic corpus context for more accurate and interpretable topic labeling.

Key words automatic topic labeling, pragmatic analysis

1 Introduction

In natural language processing (NLP), a topic typically refers to a specific subject or theme that is discussed in a context, which could be also considered as a latent semantic group of text. Topic models are proposed to find those latent semantic groups by analyzing a collection of documents, and automatically finding groups of words that frequently co-occur in the same document as topics. Understanding topics in text can be valuable for various applications, such as document summarization, sentiment analysis, information retrieval, and more. Traditional topic models such as Latent Semantic Analysis (PLSA) [1] and Latent Dirichlet Allocation (LDA) [2] are based on the Bayesian probabilistic model. Modern topic models such as Neural Variational Document Model (NVDM) [3], Bidirectional Adversarial Topic (BAT) model [4], and BERTopic [5] leverage modern neural networks and even large language models (LLM) to generate expressive topics.

In practice, topic terms are presented after a topic model is applied to the corpus. Topic terms are words or phrases defining and characterizing topics generated by topic modeling algorithms. Generally, these terms are representative of the underlying themes or subjects of topics and could help people interpret the topics and understand the latent structure of the corpus. For example, a topic related to computer science may have topic terms such as "computer", "programming", "algorithm", "code", etc.. In Latent Dirichlet Allocation, topic terms are the words with the highest probability of being associated with a particular topic. However, the cognitive load of understanding a whole list of (in practice, top 10 or top 20) terms is considered high compared to that of understanding an expressive label. Previous work [6] showed that in the task of information retrieval, humans can identify more documents when textual labels are presented than when keywords are presented. This nature has led to interest in automatic topic labeling, which is the task of outputting meaningful words or phrases as labels for topics.

A topic label is a short text that precisely describes the semantics of what the topic represents. Topic labels are expected to be comprehensible for humans to understand the topics. In definition, the topic labels can be just a concatenation of top words (e.g., "computer, programming, algorithm, code"), while a more concise and precise term that represents

the topic (e.g., , "Computer Science") is preferred.

There have been many automatic topic labeling researches. The main task of automatic topic labeling is associated with the task of finding proper label candidates and ranking them. Mei et al. [7] first introduced automatic topic labeling to LDA topics by extracting bigrams and ranking them with KL-divergence. Many automatic topic labeling methods follow this idea, such as Lau et al. [8] using the title of English Wikipedia pages as label candidates and then applying the ranking method with lexical association feature to the corpus. On the other hand, with the advancement of computing power and the development of language models based on transformers, direct approaches to label generation have been studied. Alokaili et al. [9] proposed a sequence-to-sequence model that takes topic terms as input and output corresponding labels, and Popa et al. [10] fine-tuned a pre-trained BART [11] model to generate labels from topics.

However, the ambiguity of topics generated by topic models has caused trouble for automatic topic labeling. For example, consider two different topics {*Twitter, ChatGPT, 5G, Apple watch*} and {*Northern lights, SpaceX, NASA, Gravitational waves*}. While it is reasonable to label them as "Technology" and "Science" respectively, a more distinctive label set would be "Information Technology" and "Space and Astronomy". Automatic topic labeling methods often focus solely on a single topic while ignoring the integrity of the whole corpus and topic relationship, which could bring problems to humans interpreting the topics. This problem drives us to handle topic labeling in a pragmatic manner, that labels should be generated in the context of the entire corpus, and could bring the least misconception among topics. To achieve this goal, we propose a pragmatic automatic topic labeling method by leveraging zero-shot classification of text as the evaluation of the distinctness degree of the topic labels. The overview of our proposal is shown in Figure 1. The listener model could be used to compute a metric for evaluating the distinctness of the topic label sets for a given topic distribution. Furthermore, we construct an automatic topic labeling pipeline that takes label candidates as input and applies combinatorial optimization to find the proper topic label set to the corpus.

The contributions of this paper are summarized as follows:

(1) We propose a pragmatic listener model by leveraging zero-shot classification of text as the evaluation of the distinctness degree of the topic labels.

(2) We validate the distinctness of label sets generated by existing automatic topic labeling methods by conducting crowd-sourcing tasks on Amazon Mechanical Turk. The evaluation of our pragmatic listener model on these label sets shows a correlation with the crowd-sourcing results.

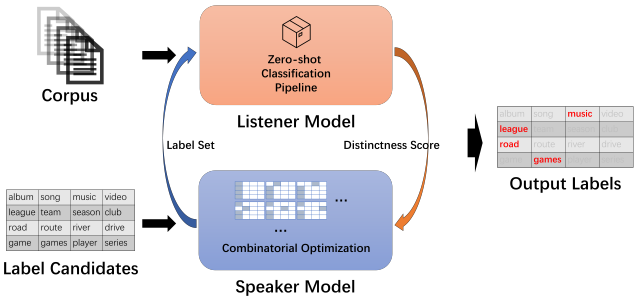


Figure 1 Overview: a pragmatic listener-speaker architecture leveraging combinatorial optimization and zero-shot classification of text for automatic topic labeling.

(3) We propose a pragmatic automatic topic labeling pipeline in the listener-speaker architecture. The experiments show that our pipeline has improved the distinctness of the topic label set.

2 Related Work

2.1 Topic Modeling

Latent Dirichlet Allocation (LDA) [2] is a probabilistic topic model that assumes documents are mixtures of topics and topics are mixtures of words. It adds a prior distribution to the document-topic and topic-word distributions on the previous work of Latent Semantic Analysis (PLSA). By applying the idea of topic distributions, other probabilistic topic models have been proposed, such as the Correlated Topic Model (CTM) [12] and the Structured Topic Model (STM) [13]. Some other approaches, such as NVDM [3] and DocNADE [14], use modern neural networks to generate similar distributions. In recent years, neural topic models have been cooperating with other methods such as graph neural networks [15], reinforcement learning [16], contrastive learning [17], and generative adversarial networks [18] to improve their performance.

Aside from probabilistic topic models, some other approaches interpret topics as clusters of documents. Top2Vec [19] leverages joint document and word semantic embeddings to find topic vectors. Top2Vec employs a Doc2Vec [20] model for embeddings and applies a clustering algorithm to the document embeddings to generate topic vectors. Sia et al. [21] introduced a similar framework that leverages pre-trained embeddings. This framework provides flexibility in the model choice on embedding models, dimension reduction methods, and topic term extraction. BERTopic [5] leverages embeddings provided by BERT [22], a pre-trained transformer-based language model, while using a class-based variation of the TF-IDF to generate coherent topic representations instead of cluster centroids.

Our proposed automatic topic labeling model has the ben-

efit of not restricting a specific type of topic model, since the labeling process is separate from the topic modeling process and does not require any special model structure.

2.2 Automatic Topic Labeling

Automatic topic labeling has experienced significant progress over the years, evolving from unsupervised to supervised approaches and incorporating the latest advancements in natural language processing (NLP). Pioneering work by Mei et al. [7] established a foundation by introducing methods to extract bigrams and rank them using KL-divergence, framing the practice of label generation from within the original corpus itself. Subsequent research such as Lau et al. [8] and Bhatia et al. [23] expanded upon this idea by employing external corpora like Wikipedia and leveraging supervised learning techniques to rank candidates.

Some other approaches leverage automatic summarization technology for labeling topic models, aiming to enhance the informativeness and meaningfulness of topic labels. Unlike other approaches, these methods mainly produce topic labels in sentence form and focus on more informative expressions of topics rather than common phrase-level labels. Cano et al. [24] introduced an independent summarization framework for identifying dominant words, which only relies on the identification of dominant terms in documents related to the latent topic. Following the idea, other approaches [25] [26] [27] [28] leverage different summarization and ranking methods and achieve better results.

As the field advanced, word embedding techniques such as Word2Vec [29] and Doc2Vec [20] were employed, enhancing the process of computing semantic similarity between topics and potential labels. The introduction of transformers [30] marked a paradigm shift in the field. The encoder-decoder architecture and the pre-training on vast corpora of text brought about a generation of transformer-based models [22] [31] [11] that excel in transfer learning, enabling unprecedented performance on a wide array of NLP tasks, including topic labeling.

Most recently, direct labeling techniques exploiting the power of transformer models have been explored. Alokaili et al. [9] introduced a sequence-to-sequence model for generating labels based on topic terms. Popa et al. [10] demonstrated the effectiveness of fine-tuning a pre-trained BART [11] model for extracting labels directly from topics. These approaches mark a shift towards leveraging pre-existing language models, which have been trained on general language understanding, to tailor them to the specific task of automatic topic labeling. Through fine-tuning these models on targeted datasets, research has enabled more accurate and contextually appropriate labels, even with smaller and domain-specific data, pushing the boundaries of what's possible in automatic topic labeling.

Among all the research mentioned above, no research focuses on inter-topic relationships when assigning labels to topics. When similar topics are presented to an existing automatic topic labeling method, one may output ambiguous labels, which could bring problems to humans interpreting the topics. Our proposal on the other hand takes the entire topic distribution as input and focuses on assigning pragmatic labels to topics in the holistic corpus context.

2.3 Pragmatic Analysis

A fundamental concept in computational pragmatics is that speaker and listener agents function within a cooperative framework, while both gain advantages by engaging in reasoning about the intentions and actions of others within this context. William et al. [32] firstly used pragmatic reasoning alone with weighted inference to address ambiguity and generate clarification requests in a human-robot dialog task. Daniel et al. [33] introduced a pragmatic listener-speaker model to reason why speakers produce certain instructions, and how listeners will react to them, and showed that such pragmatic inference improves both the listener and speaker models in diverse settings. Further work [34] extended the idea of pragmatics in other NLP systems, examining how task goals, environmental contexts, and communicative affordances in each study enhance linguistic meaning, and offering suggestions for future grounded task design aimed at organically eliciting pragmatic phenomena.

Our proposal involves the idea of pragmatic analysis by constructing a listener-speaker architecture that works cooperatively to generate distinctive topic labels for each topic.

3 Proposal

3.1 Speaker Model

Consider a topic distribution of $T = \{t_1, t_2, \ldots, t_m\}$ where m is the number of topics. For i-th topic t_i, the label candidates of the topic is $L_i = \{l_{i,1}, l_{i,2}, \ldots, l_{i,n_i}\}$ where n_i is the number of label candidates of topic t_i. The input of the speaker model will be $S = (L_1, L_2, \ldots, L_m)$, which is the set of all label candidates. The set of all possible combinations could be presented as $\Phi = L_1 \times L_2 \times \cdots \times L_m$.

A possible output label assignment $\phi \in \Phi$ for topic distribution T will be $\phi = (l_1, l_2, \ldots, l_i)$ where $l_i \in L_i$. The total number of possible label sets could be calculated as

$$|\Phi| = \prod_{i=1}^{m} n_i, \tag{1}$$

which will be a very large number when m and n_i grow. The speaker model should address the issue of exploring the large combinatorial space in terms of computational efficiency. In Section 3.3, an efficiency-aware method based on a combi-

natorial optimization technique is presented.

3.2 Listener Model

The listener model will evaluate each ϕ with a distinctness score function $f(D, \phi)$ by classifying the corpus D on ϕ. For document $d \in D$, the gold label $y_{true}(d)$ is the index of the topic assigned to the document by the topic model. The classification result $y_{pred}(d, \phi)$ will be granted by passing d and ϕ to a zero-shot text classification model, which is the corresponding index of the topic of the classification result label l_{pred}. The classification metric would consider the prediction correct if and only if $y_{true}(d) = y_{pred}(d, \phi)$.

The distinctness score could then be calculated as:

$$f(D, \phi) = \frac{1}{|D|} \sum_{i=1}^{m} n(D, i) F1(D, \phi, i), \tag{2}$$

$$n(D, i) = |\{d \in D \mid y_{true}(d) = i\}| \tag{3}$$

$$F1(D, \phi, i) = \frac{2 \cdot precision(i) \cdot recall(i)}{precision(i) + recall(i)}, \tag{4}$$

$$precision(i) = \frac{TP(i)}{TP(i) + FP(i)}, \tag{5}$$

$$recall(i) = \frac{TP(i)}{TP(i) + FN(i)}, \tag{6}$$

$$TP(i) = |\{d \in D \mid y_{pred}(d, \phi) = i, y_{true} = i\}|, \tag{7}$$

$$FP(i) = |\{d \in D \mid y_{pred}(d, \phi) = i, y_{true} \neq i\}|, \tag{8}$$

$$FN(i) = |\{d \in D \mid y_{pred}(d, \phi) \neq i, y_{true} = i\}|. \tag{9}$$

which is the weighted F1-score of the classification on corpus D. The distinctness score represents how well the given label assignment ϕ induces the correct classification from the pre-trained zero-shot classification model. The zero-shot classification model is assumed to represent a common sense association of a document with a topic label. A higher distinctness score is expected to indicate that the given topic labels ϕ are more natural for labels to classify documents.

3.3 Combinatorial Optimization

The goal of the proposed pragmatic listener-speaker method is to find the label set with the highest distinctness score $\phi_{best} = \arg\max f(D, \phi)$. To brute force through $\prod_i^m n_i$ combinations is impossible when m and n_i are large. In order to find the ϕ_{best}, we leverage local search [35] for combinatorial optimization.

Let (Φ, f) be the combinatorial optimization problem, where the score function f is a mapping from the set of solutions Φ to $\mathbb{R}$. We define a neighborhood function $\mathcal{N}(\phi)$ by changing one label l_i in label set ϕ to another label candidate l_i' for that topic:

$$\mathcal{N}(\phi = (l_1, \ldots, l_m)) = \bigcup_{i \in [1..m]} \{(l_1', \ldots, l_m') \mid l_i' \in L_i \wedge \forall j \in [1..m] \backslash i (l_j' = l_j)\}. \tag{10}$$

A local optimum is a solution $\hat{\phi}$ where the score of it is no worse than its neighbors:

$$f(\hat{\phi}) \geqq f(\phi) \quad \forall \phi \in \mathcal{N}(\hat{\phi}). \tag{11}$$

We use the simple local search algorithm by iterating through neighbors for each topic label until a local optimum is found. The pseudo-code is shown in Algorithm 1.

Algorithm 1 Local search for topic labels.

```
function LOCAL_SEARCH(S, D, m, f)
    φ̂ ← RandomSampling(S)
    BestScore ← f(D, φ̂)
    Checked ← [False for x in range(0, m)]
    repeat
        for i in range(0, m) do φ ← φ̂
            for candidate in S[i] do φ[i] ← candidate
                if f(D, φ) > BestScore then
                    φ̂ ← φ
                    BestScore ← f(D, φ)
                    Checked ← [False for x in range(0, m)]
                end if
            end for
            Checked[i] ← True
        end for
    until False not in Checked
    return φ̂
end function
```

3.4 Pragmatic Topic Label Pipeline

The pragmatic topic label pipeline could leverage different automatic topic labeling methods in an ensemble learning manner. After applying topic modeling to corpus D and obtaining topic distribution T, they will be passed into off-the-shelf automatic labeling methods. The label candidates $ATL_j(D, T) = (s_{j,1}, \ldots, s_{j,m})$ will be obtained, and the sets from different automatic topic labeling methods will then be concatenated to generate the input for the listener-speaker model,

$$S = (\bigcup_j s_{j,1}, \ldots, \bigcup_j s_{j,m}) \tag{12}$$

where ATL is an off-the-shelf automatic labeling method and j is the number of different automatic topic labeling methods used.

The pipeline uses the zero-shot classification of text [36] for the distinctness score calculation. Zero-shot classification involves predicting a class that the model has not encountered during training. This approach could be thought of as transfer learning by utilizing a pre-trained language model. The model is provided with a prompt and a natural language text sequence that describes the task, classification target, and label set. In practice, it is considered reasonable to utilize language models that are pre-trained on natural language inference (NLI) tasks.

After the combinatorial optimization process through the

pragmatic listener-speaker model, the optimized label set would then be outputted as the final label set for the topic.

4 Experiments and Discussion

4.1 Experiment Settings

4.1.1 Corpus

We use the WikiText-2 dataset [37] for evaluation. The WikiText-2 dataset is extracted from the set of verified English articles on Wikipedia. It was introduced in 2016 and contains approximately 2.5 million tokens in total. The articles are in their raw form which retains numbers, cases, and punctuations. Since topic modeling is an unsupervised machine learning method, all three splits (test, train, validation) are used in topic modeling. In experiments, each paragraph is treated as a document in the training step since the articles are too long to fit in the language model used in the topic model and the zero-shot classification pipeline.

4.1.2 Topic Modeling

We choose BERTopic [5] as our topic modeling method. BERTopic leverages transformers and c-TF-IDF to create dense clusters, allowing for easily interpretable topics whilst keeping important words in the topic descriptions. It has the benefits of being one of the state-of-the-art topic modeling methods on various corpus, competence in different settings, and user-friendly APIs.

The BERTopic model generates topic representations through the following steps.

（1） **Embedding representation**: Each document is converted to its embedding representation using a pre-trained language model. We used the `sentence-transformers/all-mpnet-base-v2` model from Huggingface [38], which is a pre-trained Sentence-BERT [31] model based on MPNet [39].

（2） **Dimension reduction**: The dimensionality of the resulting document embeddings is reduced using UMAP [40] to optimize the clustering process.

（3） **Document clustering**: The reduced embeddings are clustered using HDBSCAN [41]. We apply hierarchical clustering to set the number of topics to 40. Although HDBSCAN is a soft-clustering approach allowing noise to be modeled as outliers, in order to avoid too many outliers affecting the consistency of the model and match the topic representation of traditional topic models, we apply outlier reduction using firstly class TF-IDF (c-TF-IDF) representations with threshold = 0.05 and then the topic distributions.

（4） **Topic representation**: The c-TF-IDF score for each term is calculated by treating each document cluster as a big document. This procedure models the importance of words in clusters and generates topic-word distributions for each cluster of documents.

4.1.3 Zero-shot Classification

We use the Huggingface zero-shot classification pipeline [42] in our listener model. We choose the `valhalla/distil bart-mnli-12-1` model as the pre-trained language model, which is the distilled version of `bart-large-mnli` BART [11] model. The batch number is set to 1.

4.1.4 Automatic Topic Labeling Baselines

To evaluate our proposed pragmatic model and generate label candidates for the automatic topic labeling pipeline, we choose the following automatic topic labeling methods as baselines.

- **Top terms**: The topic representations generated by the BERTopic model are used as simple topic labels. The top 4 words of each topic are concatenated into one label.
- **KPE**: Following the idea of utilizing automatic summarization technology for labeling topic models while trying to restrict the lexical length of the labels, we apply YAKE [43], an off-the-shelf key-phrase extraction (KPE) method on top 50 documents of each topic. The max lexical length of the generated labels is set to 3.
- **NETL**: We choose Neural Embeddings Automatic Labeling (NETL) [23] as the baseline which stands for the methods that employ external corpora like Wikipedia and leverage supervised learning techniques to rank candidates. The top 10 words of each topic are passed to the model to generate label candidates.
- **LLM**: Regarding the advance of large language models, we leverage ChatGPT [44] as automatic topic labeling methods by simply passing the top 10 words to the `gpt3.5-turbo-1106` model and prompting to ask the model to generate topic labels based on them.
- **Human labels**: An expert in topic modeling is asked to assign a topic label after reading the top 10 words and top 20 documents for each topic.

4.2 Listener Model Evaluation

To evaluate the zero-shot distinctness score in the proposed pragmatic listener model, we conduct crowd-sourcing topic labeling tasks on Amazon Mechanical Turk (AMT). Participants in the AMT tasks are provided with a document in the corpus and five topic labels. The topic labels are the labels of the five closest topics to the document (including the corresponding topic), generated by one of the baselines. Participants are asked to select the most relevant label for the text. After filtering out short documents, 30 documents are sampled from each of the 40 topics to generate a batch of 1200 tasks. For each baseline method, one batch is published. A demo of the AMT tasks is shown in Figure 2. The result is gathered after the tasks are completed in AMT crowd-sourcing. After the result is gathered, we calculate the F1-score in crowd-sourcing of each topic among each la-

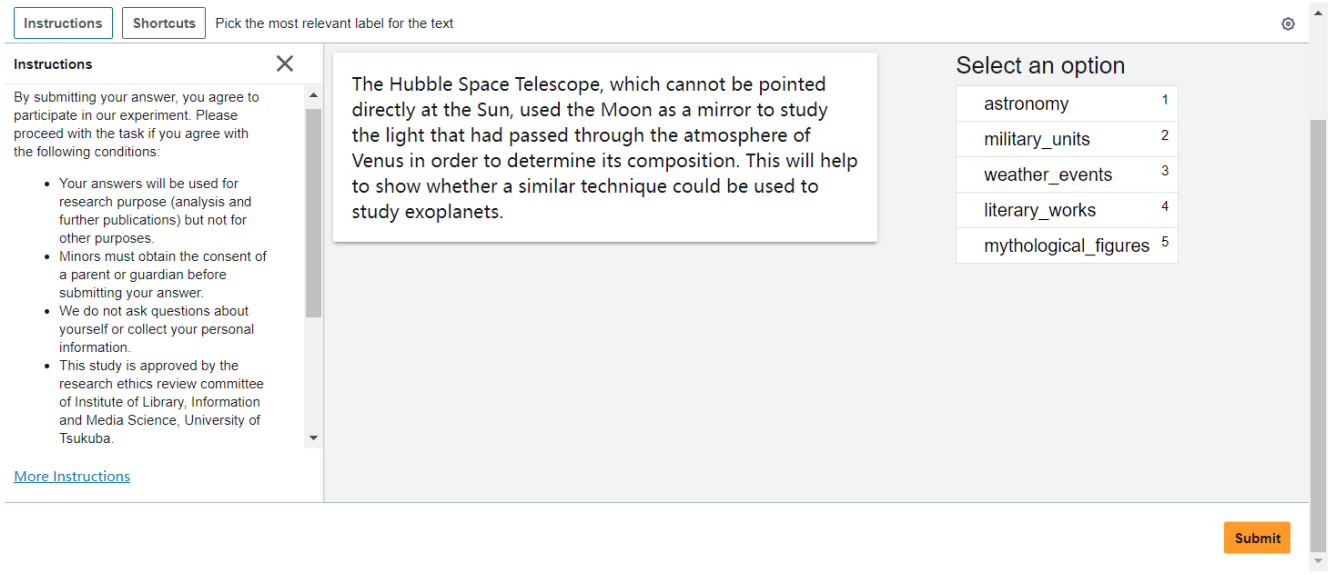


Figure 2 A demo of the Amazon Mechanical Turk crowd-sourcing tasks. Participants are asked to select the most relevant label to the text among five choices.

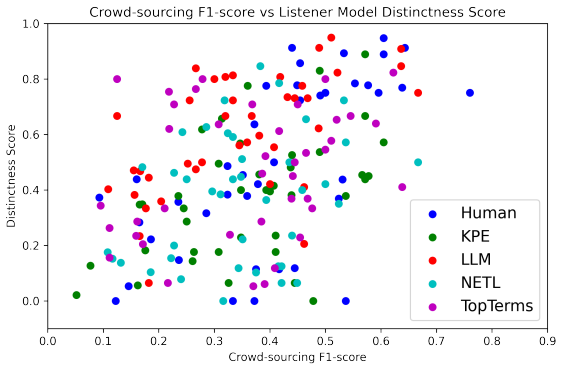


Figure 3 A scatter plot of topic labels evaluation on each topic labeling baseline. Each point on the graph stands for a topic.

bel set, as long as the weighted overall F1-score of each label set.

The same sampled corpus and label sets are then passed into the proposed listener model for evaluation. The distinctness score of each topic among each label set and the overall distinctness score are gathered. We conduct a correlation analysis between the crowd-sourcing result and the listener model result. The result is shown in Figure 3 and Table 1. The correlation between the crowd-sourcing F1-score and distinctness score is $r = 0.4299$, with $p < 0.0001$, consistent with our hypothesis that there is a positive correlation between the crowd-sourcing F1-score and distinctness score.

4. 3 Automatic Labeling Pipeline Evaluation

To evaluate our proposed automatic labeling pipeline, we utilize our proposed pipeline to generate the label set for the corpus. Four out of five of the baselines (**Top Terms**, **KPE**, **NETL**, **LLM**) are used to generate label candidates for the automatic labeling pipeline. We run two different settings on our proposed pipeline.

Table 1 The overall score (weighted-average) of each topic labeling baseline compared between crowd-sourcing F1-score and distinctness score.

	Crowd-sourcing F1-score	**Distinctness Score**
Human Labels	0.4092	0.4903
KPE	0.3653	0.3782
LLM	0.3475	0.6061
NETL	0.3466	0.3722
Top Terms	0.3597	0.4658

- **ppl_4**: We use the exact labels outputted by baselines in Section 4. 2, comprising 4 candidates for each topic. 30 documents from each of the 40 topics are sampled from the corpus to generate a subcorpus of 1200 documents.(注1)
- **ppl_10**: Top 3 candidate labels outputted by each of **KPE**, **NETL**, **LLM** and the label generated by **Top Terms** comprise 10 candidates for each topic. The same subcorpus used in **ppl_4** is used.

After the two label sets are generated by the proposed pipeline, we evaluate them by conducting the same AMT crowd-sourcing tasks in Section 4. 2. To remove the effect of uninterpretable topics, we filter out topics which answer accuracy is lower than 0.4 on **Human labels** label sets in crowd-sourcing tasks. The crowd-sourcing F1-score results on baselines and proposed methods after filtering are shown

(注1) : We use hyperparameter `random_seed` to make sure the documents are sampled differently to the subcorpus in Section 4. 2

in Table 2.

Table 2 Crowd-sourcing F1-score on filtered topics.

	Crowd-sourcing F1-score
Baselines	
Human	0.4792
KPE	0.3508
LLM	0.3645
NETL	0.3654
Top Terms	0.3398
Proposed Method	
ppl_4	0.3172
ppl_10	0.3762

4.4 Discussion

According to the result in Table 1, in terms of crowd-sourcing F1-score, **Human labels** performs the best, obtaining a score of 0.4092. **KPE** and **Top Terms** perform a bit weaker, while **LLM** and **NETL** perform the worst. This may be due to the fact that both **KPE** and **Top Terms** labels are generated from the corpus, while **LLM** and **NETL** labels are generated from the top terms. The labels of **LLM** and **NETL** are more general terms of subjects, while the labels of **KPE** and **Top Terms** tend to consist of specific objects. Additionally, the labels of **KPE** and **Top Terms** contain more words on average compared to other baselines. This nature helps participants choose the correct labels once the words in labels appear in the document, even if the document is weakly related to the topic.

In terms of the distinctness score outputted by the proposed listener model, the labels of **LLM** score the highest of 0.6061, followed by **Human Labels** of 0.4903 and **Top Terms** of 0.4658. This may be because while being the most similar label with **Human Labels**, **LLM** produces multiple concepts in a single label (e.g., "Historical architecture" compared to "Archaeology, Historical Sites"), which helps the zero-shot classification in matching the documents. This could also be proved by the score of **Top Terms** since each word could be interpreted as an independent concept. On the other hand, the zero-shot classification method tends to give higher scores to common-sense labels (e.g., "Historical architecture" compared to "sun temple"), making the score of **KPE** and **NETL** relatively low. This may be due to the choice of the language model or the prompt used in the zero-shot classification pipeline, which required further investigation.

Nonetheless, despite the different aspects between the crowd-sourcing F1-score and the distinctness score assigned by the proposed method, the correlation analysis shows a strong belief that there is a positive correlation between the crowd-sourcing F1-score and distinctness score, addresses that the proposed listener model could serve as a metric to assess the quality of label sets in the context of the entire corpus.

During the listener model evaluation, we noticed that some of the **Human labels** have exceptionally low accuracy in crowd-sourcing tasks, some of them even lower than the expected accuracy of random choices. This phenomenon warns us of the possibility that the topic model could output improper topics, that even humans could not assign proper labels to them. We consider this as a problem of topic modeling and therefore exclude these topics by filtering out them with crowd-sourcing accuracy on **Human labels** during the automatic labeling pipeline evaluation.

Before we discuss the result of the automatic labeling pipeline evaluation, two deficiencies in the proposed automatic topic labeling pipeline need to be addressed. First, the problem of the local optimum. Since we chose local search for combinatorial optimization, there is a possibility that the outputted label set is a local optimum, rather than the global optimum with the highest distinctness score. Approaches such as simulated annealing, tabu search, and genetic algorithms are designed to solve this problem. However, all of those approaches require branching for worse neighbors or finding multiple optimums in order to overcome the problem of local optimum. This brings us to the second problem of the computational cost. As mentioned in Section 3.4, zero-shot classification involves querying modern language models with prompts. The computational cost of such querying is extremely high compared to traditional classification methods. At the same time, combinatorial optimization executes a huge amount of such classification in order to compare the scores between neighbors. This nature makes the computational cost of the proposed automatic topic labeling pipeline significantly high, which is the reason for sampling a sub-corpus for the pipeline, restricting the number of label candidates, and using a distilled version of the language model during the experiments. As for now, there is no proper solution other than increasing the computational power or reconstructing the off-the-shelf zero-shot classification method for the proposed pipeline.

After uninterpretable topics are filtered out, according to Table 2, the best score is 0.4792 of **Human labels**, followed by 0.3762 of the proposed method **ppl_10**. **ppl_4** obtain the worst score of 0.3172. This may indicate that the number of candidates provided to the pipeline affects the performance by a large margin, even if those candidates are not been chosen as the best choice according to the off-the-shelf automatic topic labeling methods. This phenomenon shows that in this specific problem setting of combinatorial optimization, a better solution is more likely to be found with the

naive local search method in a larger combination set. On the other hand, the proposed method **ppl_10** outperforms other baselines, showcasing the effectiveness of our proposed automatic topic labeling pipeline.

5 Conclusion

In this research, we proposed a pragmatic automatic topic labeling approach that evaluates the ambiguity degree of topic labels in the holistic corpus. The pragmatic automatic topic labeling approach consists of two parts, the listener model and the speaker model, leverage zero-shot classification and combinatorial optimization respectively, to output the pragmatically proper label set among the input label candidates. Our crowd-sourcing experiments demonstrated the correlation between our pragmatic approach and human evaluation, proving that the listener model serves as a metric to assess the quality of label set in the aspect of distinctness. Our automatic topic labeling pipeline experiments showed that the proposed method is effective in improving the quality of the topic label set under certain circumstances.

The main drawback of this work is that the proposed method relies on the zero-shot classification of text, whose computational cost weighs on the efficiency of the entire pipeline. As this research explores the idea of the pragmatic approach in automatic topic labeling, future work should focus on exploring other methods or training a lightweight designated machine learning model to achieve the same or better performance in this task.

Acknowledgment

This work was supported by JSPS KAKENHI #21H03552, #22K12039 and JST CREST #JPMJCR22M2.

References

[1] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pp. 50–57, New York, NY, USA, August 1999. Association for Computing Machinery.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[3] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, p. 1727–1736. JMLR.org, 2016.

[4] Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural topic modeling with bidirectional adversarial training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 340–350, Online, July 2020. Association for Computational Linguistics.

[5] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[6] Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Representing topics labels for exploring digital libraries. In *IEEE/ACM Joint Conference on Digital Libraries*, pp. 239–248. IEEE, September 2014.

[7] Qiaozhu Mei, Xuehua Shen, and Chengxiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pp. 490–499, New York, NY, USA, August 2007. Association for Computing Machinery.

[8] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1536–1545, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[9] Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. Automatic generation of topic labels. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 1965–1968, New York, NY, USA, July 2020. Association for Computing Machinery.

[10] Cristian Popa and Traian Rebedea. BART-TL: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[12] David Blei and John Lafferty. Correlated topic models. *Adv. Neural Inf. Process. Syst.*, Vol. 18, p. 147, 2006.

[13] Lan Du, Wray Buntine, and Mark Johnson. Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200. aclweb.org, 2013.

[14] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc., 2012.

[15] Qingqing Long, Yilun Jin, Guojie Song, Yi Li, and Wei Lin. Graph structural-topic neural network. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1065–1073, New York, NY, USA, August 2020. Association for Computing Machinery.

[16] Jeremy Costello and Marek Reformat. Reinforcement learning for topic models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4332–4351, July 2023.

[17] Thong Thanh Nguyen and Anh Tuan Luu. Contrastive learning for neural topic model. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, December 2021.

[18] Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural topic modeling with bidirectional adversarial training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors,

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 340–350, Online, July 2020. Association for Computational Linguistics.

[19] Dimo Angelov. Top2Vec: Distributed representations of topics. August 2020.

[20] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. May 2014.

[21] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! pp. 1728–1736, November 2020.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[23] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Automatic labelling of topics with neural embeddings. December 2016.

[24] Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from Twitter by summarisation. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 618–624, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[25] Xiaojun Wan and Tianming Wang. Automatic labeling of topic models using text summaries. In Katrin Erk and Noah A Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2297–2305, Berlin, Germany, August 2016. Association for Computational Linguistics.

[26] Mohamad Hardyman Barawi, Chenghua Lin, and Advaith Siddharthan. Automatically labelling Sentiment-Bearing topics with descriptive sentence labels. In *International Conference on Applications of Natural Language to Information Systems*, pp. 299–312. unknown, June 2017.

[27] Dongbin He, Minjuan Wang, Abdul Mateen Khattak, Li Zhang, and Wanlin Gao. Automatic labeling of topic models using graph-based ranking. *IEEE Access*, Vol. 7, pp. 131593–131608, 2019.

[28] Domenic Rosati. Moving beyond word lists: towards abstractive topic labels for human-like topics of scientific documents. October 2022.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. January 2013.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. June 2017.

[31] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[32] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 1387–1393. AAAI Press, January 2015.

[33] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *North American Chapter of the Association for Computational Linguistics*, 2017.

[34] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12619–12640, Singapore, December 2023. Association for Computational Linguistics.

[35] Aarts Emile H L. and Jan Karel Lenstra. *Local search in combinatorial optimization*. Princeton Univ. Press, 2003.

[36] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics.

[37] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. 2016.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M Rush. HuggingFace's transformers: State-of-the-art natural language processing. October 2019.

[39] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, No. Article 1414 in NIPS'20, pp. 16857–16867, Red Hook, NY, USA, December 2020. Curran Associates Inc.

[40] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, Vol. 3, No. 29, p. 861, 2018.

[41] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, Vol. 2, No. 11, p. 205, 2017.

[42] Hugging Face. What is zero-shot classification? - hugging face, 2023.

[43] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! keyword extraction from single documents using multiple local features. *Inf. Sci.*, Vol. 509, pp. 257–289, January 2020.

[44] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Others. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*, Vol. 33, pp. 1877–1901, 2020.

BERTopic 文書分類器の分類安定性に関する評価手法の提案

櫻井　勇気†　　小林　亜樹††

† 工学院大学大学院工学研究科 電気・電子工学専攻　〒 163–8677　東京都新宿区西新宿 1-24-2
†† 工学院大学情報学部情報通信工学科　〒 163–8677　東京都新宿区西新宿 1-24-2
E-mail: †cm23034@g.kogakuin.jp, ††aki@cc.kogakuin.ac.jp

あらまし　BERTopic は 2020 年に提案された，埋め込み表現と分類，次元圧縮，可視化手法の組み合わせを自由に設定するトピックモデリング手法であり，Transformer ベースの学習済み言語モデルと任意の分類アルゴリズムを用いた分類器を作成することが可能である．BERTopic によって作成される分類器は LDA など従来の分類手法よりも高い性能を持つとされている．しかし，予備実験の結果，分類処理毎に異なる分類クラスターを出力し，このときの分類結果の一致性は必ずしも高くない．そこで本稿では，分類器の分類結果出力の安定性を評価する枠組みを提案し，対象文書の種類などによる分類結果の安定性の違いを表現できることを目指す．提案手法は，異なる 2 文書の文書対を単位データと考え，同一の対象文書集合に対する独立した相異なる分類処理結果において，文書対の分類結果が同一となるか否かの比率として分類の安定性を表すものである．いくつかの評価対象文書集合において提案評価手法に関する評価を行い，指標としての有効性などを検証する．

キーワード　テキスト分類，LLM，言語モデル，文書要約，情報抽出

1　はじめに

大規模言語モデルで注目を浴びる Transformer を文書分類に応用した手法として，2020 年に BERTopic [1] が提案されている．BERTopic は自由に文書分類器を作成する手法であり，文書を分類するにあたって必要な埋め込み表現の獲得や，次元圧縮，分類，可視化に用いる手法を自由に組み替えることが可能である．Abeer らは Transformer をベースとした複数の自然言語処理モデルと密度ベースのクラスタリングアルゴリズムである HDBSCAN [2] を組み合わせた分類器による文書分類を行い，LDA や NMF といった他の文書分類手法による分類結果と比較して良好な結果が得られたと報告している [3]．HDBSCAN はクラスタの数や形状を柔軟に変更できるが，k-means のような従来のアルゴリズムは事前のクラスタ数の決定を必要とする。そのため、教師無し学習を用いて未知の文書を分類していく場合，HDBSCAN は適したアルゴリズムであると考えられる．しかし HDBSCAN を採用した分類器を用いる場合分類を行うたびにクラスターが変化するため，複数回の分類で内容の一貫した出力を得ることは難しい．本論文では BERTopic を用いて，Sentence-bert [4]，USE [5]，Flair [6]，SpaCy [7] [8] の 4 つの異なる埋め込み表現を持つ分類器を作成し，これらを用いて同一データセットに対し分類を行うことで BERTopic が文書集合に対してどれほど安定した結果をもたらすのか調査を行い，結果を考察する．

1.1　BERTopic

本章では BERTopic による分類器の作成について説明する。BERTopic によって作成される分類器の概要を図 1[1]に示す．BERTopic の具体的な仕組みとして，最初に文書を埋め込み表現に変換する．この時文書が多くの語彙を含むほど埋め込み表現に変換された場合には高次元の表現となる．そのため次に次元圧縮を行い，埋め込み表現の次元を減らすことでより低次元化した埋め込み表現を得る．こうして得た低次元の埋め込み表現を分類（クラスタリング）することで，似た特徴を持つ文書による文書集合（クラスタ）を複数得る．最後にクラスタを一つの文書と見做してトークン化したうえで，ラベルの付与などを経て人間が理解できる表現に可視化される．このように，BERTopic は埋め込み表現，次元圧縮，分類手法，可視化手法を組み合わせて文書分類器を作成する手法の一つであるが，その特徴は各手順で用いる具体的な手段を作成者が任意に変更できる点にある．各手順は独立しているため，文書の特徴や分類に必要なコストに応じて各手順で用いる手段を変更することで，作成者が望むような分類器を作成することが出来る．

1：https://maartengr.github.io/BERTopic/algorithm/algorithm.html

2　用語定義

本章では本稿で用いる独自の用語について説明する．

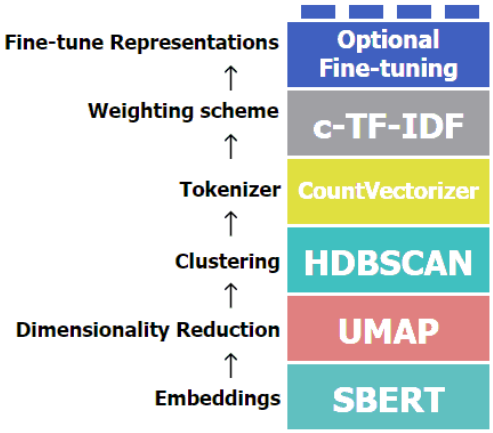
2.1　分類放棄文書

分類器は一般に複数のクラスタへ文書分類を行う．このとき，これらのクラスタのいずれにも属さないと判定された文書を分類放棄文書と呼ぶ．本稿で扱う分類器では，HDBSCAN の密度閾値未満の文書が該当する．

2.2　文書対

分類対象文書集合に含まれる相異なる 2 文書のことを文書対と呼ぶ．

. width.



width

図 1　BERTopic 分類器の概要

表 1　作成分類器

手法	種類
埋め込み表現	Sentence-bert，USE，Flair，Spacy
次元圧縮	UMAP
分類	HDBSCAN
トークン化	CountVectrizer
ラベル付与	c-tf-idf

2.3　分類一致文書対

一般に複数回の分類において，文書対に含まれる両文書の分類結果が一致する文書対のこと．各回の分類が完全に安定している場合，分類一致文書対によって全ての文書が被覆される．

3　予 備 実 験

本章では，評価手法の提案に際して行われた予備実験の結果から，BERTopic による分類器の特性を示す．

処理対象とする文書集合として，scikit-learn で配布されている 20 グループからなる英ニュース記事によるデータセット [9] と，Google Colaboratory で配布されているドナルド・トランプにまつわる英 tweet によるデータセット [10] の 2 種類を用いた．本論文では以降，英ニュース記事を分類対象とした場合は N，英 tweet を分類対象とした場合は T と文字を割り振り，埋め込み表現とともに表記している．

予備実験では表 1 に示したように，BERTopic を用いて Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成し，それぞれの分類器を用いてデータセットに対して 6 回の分類を行った．1 回目の分類結果と 2-6 回目の分類結果を比較し，同一の文書が比較において同一のクラスタに分類された回数を測定し，すべての比較回数における割合を測った．

1 回目の分類結果と 2-6 回目の分類結果の比較は図 2 に示した．また 1 回目の分類結果と 2-6 回目の分類結果から得られる割合の平均を表 2 に示した．

図 2 および表 2 に示したように，BERTopic 分類器による文書分類結果は，用いられた埋め込み表現や分類の対象となった

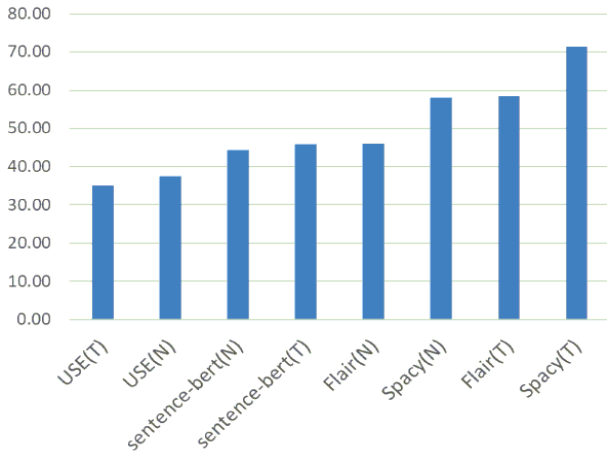


図 2　予備実験結果 (割合平均)

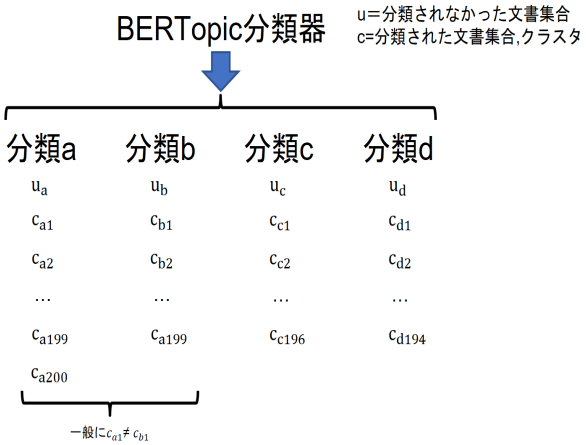


図 3　BERTopic 分類器の仕様

データセットにより変化し，その一致割合は必ずしも高くない．この仕様の原因となる BERTopic 分類器の挙動は図 2 に示したように，BERTopic 分類器は同一の文書集合を分類対象とする場合であっても，分類 a，b，c，d と分別されるような複数回の分類において分類処理毎に独立したベクトル表現や次元圧縮結果，分類クラスタを獲得するためである．よって分類ごとに得られるクラスタの総数や，各クラスタの形状は異なっており，複数回の分類において付与される番号やラベルが同一なクラスタが得られたとしても実際には異なるクラスタが生成されている．

本論ではこのように複数回の分類において異なる出力が得られる場合，それら出力の差はなるべく小さいほうが望ましいと考える．よって後述するように，文書対を用いた分類の安定性を評価する指標を提案する．

4　提 案 手 法

本稿で提案する文書分類器の安定性指標について説明する．BERTopic 文書分類器は内部で乱数要素を持つため分類結果は毎回異なる．分類器の性能評価には分類精度のほかこの分類結果の一貫性，安定性も重要であると考え，分類安定性指標を提案する．

複数の分類回の結果が完全に一致している場合，ある分類回で同一クラスタに分類された文書対は，他の分類回においても同一クラスタに分類されることとなる．そこで全文書対中，2回の分類それぞれにおいていずれも同一クラスタに分類された文書対の比率を用いて分類安定性指標とする手法を提案する．本指標は同一クラスタ数への分類において，値が大きいほど安定した（各回での結果の一致度合いが高い）分類を行っていることを示す．

まず，i 個の文書 d を含む文書集合 D を定義する．D は分類器によって処理される文書集合を想定している．

$$D = \{d_1, d_2, \ldots, d_i, \ldots d_j, \ldots\} \tag{1}$$

ある文書 d_i の独立した $r(r = \{0, 1\})$ 回目の分類クラスタ番号を

$$c(d_i, r) \tag{2}$$

で表す．

文書集合 D に属する相異なる 2 文書からなる文書対 p_k は，式 3 のように定義される．ここで，k は，文書対を識別する番号で，文書対集合 P の要素数 $|P|$ を用いて $1 \leqq k \leqq |P|$ である．$|P|$ は D を用いて式 4 で計算される．

$$P = \{p_k \mid p_k = (d_i, d_j) \in D^2, \quad i < j\} \tag{3}$$

$$|P| = \frac{1}{2}|D|(|D| - 1) \tag{4}$$

0 回目の分類において文書 d_i，d_j が同じクラスタに分類された場合，それらの文書からなる文書対 p_k は式 5 のようになり，同様に 1 回目の分類においても文書 d_i，d_j が同じクラスタに分類された場合，それらの文書からなる文書対 p_k は式 6 のようになる．

$$\{p_k \mid p_k \in P, c(d_i, 0) = c(d_j, 0), \quad i < j\} \tag{5}$$

$$\{p_k \mid p_k \in P, c(d_i, 1) = c(d_j, 1), \quad i < j\} \tag{6}$$

ここで，0 回目および 1 回目の両分類において文書 d_i，d_j が同じクラスタに分類された場合，文書 d_i，d_j からなる文書対 p_k を分類一致文書対と定義する．分類一致文書対の集合である分類一致文書対集合 T は式 7 のように示せる．

$$T = \{p_k \mid c(d_i, 0) = c(d_j, 0) \wedge c(d_i, 1) = c(d_j, 1), \quad i < j\} \tag{7}$$

以上から，文書対集合 P の要素数における分類一致文書対集合 T の要素数の割合を，分類器における安定性 s として式 8 のように定義し，分類器の分類結果出力の安定性を評価する手法として提案する．

$$s = \frac{|T|}{|P|} \tag{8}$$

5 分類器評価

分類器の評価実験にあたり，3 章と同様に，Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成した．それぞれの分類器を用いて英ニュース記事および英 tweet の 2 つのデータセットに対して 10 回の分類を行い．そこから 5 つの文書対集合 P_1-P_5 を得た．さらに P_1-P_5 から 5 つの分類一致文書対集合 T_1-T_5 を得て，ここから求められる 5 通りの安定性 s の平均を最終的な評価値として，表 3 の最右列および図 4 に示している．また英ニュース記事を分類対象とした場合は N，英 tweet を分類対象とした場合は T と文字を割り振り，用いられた埋め込み表現とともに表記している．

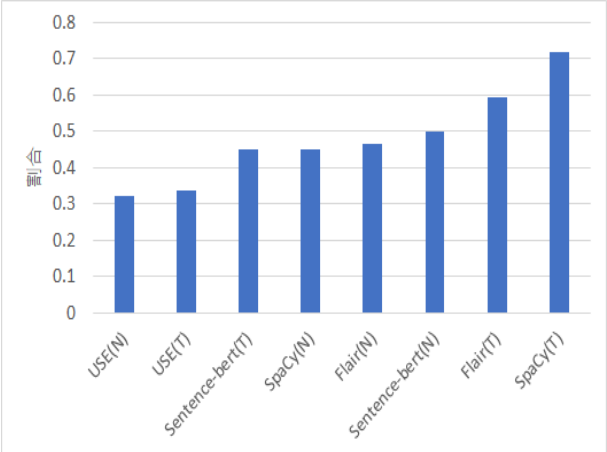


図 4　安定性評価

5.1 結　　果

評価手法を用いた評価結果を図 1 に示す。埋め込み表現に Sentence-BERT を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.50、英 tweet を対象とした分類の安定性が 0.46 となり、分類対象文書による安定性の差は 0.06 となった。埋め込み表現に USE を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.32、英 tweet を対象とした分類の安定性が 0.34 となり、分類対象文書による安定性の差は 0.02 となった。また埋め込み表現に SpaCy を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.45、英 tweet を対象とした分類の安定性が 0.72 となり、分類対象文書による安定性の差は 0.27 となった。埋め込み表現に Flair を用いた分類器は英ニュース記事を対象とした分類の安定性が 0.46、英 tweet を対象とした分類の安定性が 0.60 となり、分類対象文書による安定性の差は 0.14 となった。

5.2 考　　察

Sentence-BERT 以外の埋め込み表現を用いた場合，英 tweet を対象とした分類のほうが英ニュース記事を対象とした分類よりも安定性が高くなった．これは Sentence-BERT の有する文章単位でトークンを獲得する特性から，含まれる語数の多い news 記事の分類においてより高い安定性を発揮したためと考えられる．Sentence-BERT と同様に文章単位でトークンを獲得する特性を持つ USE は英 tweet を対象とした分類の安定性が高く表れているが、分類対象文書による安定性の差は 0.02 と他分類器と比較して低いため、Sentence-BERT と同様

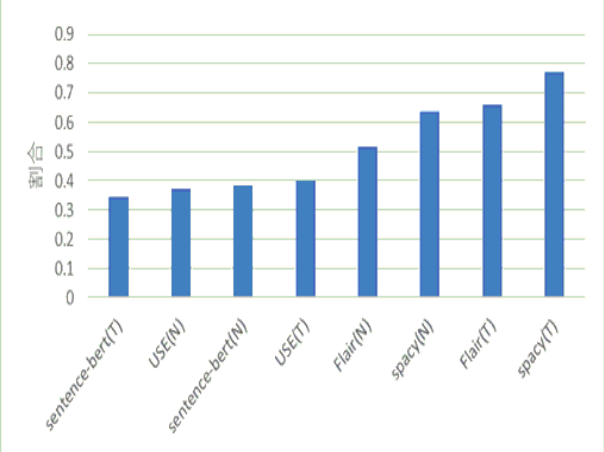


図 5　BERTopic 分類器における分類放棄文書割合

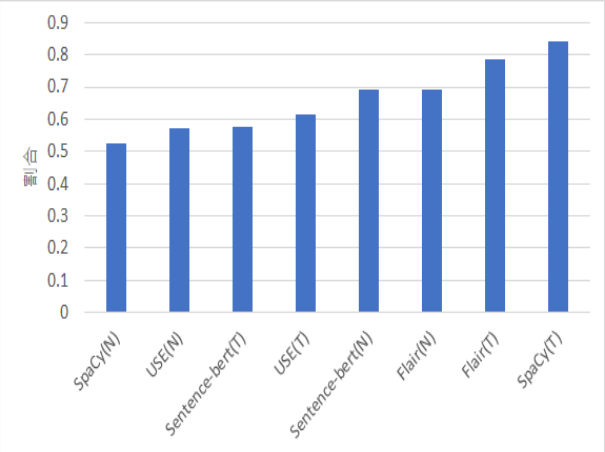


図 6　追加実験結果（分類放棄文書なし）

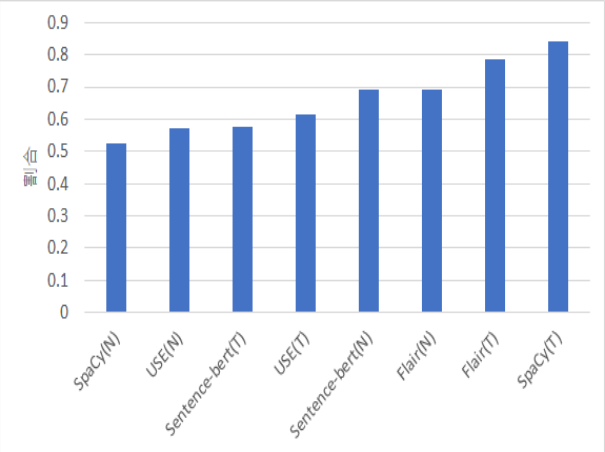


図 7　追加実験結果（分類放棄文書あり）

に含まれる語数の多い文書において高い安定性を発揮したと考えられる。一方で SpaCy，Flair を利用した分類器の場合，英 tweet を分類対象とした場合に安定性 s が 0.6-0.7 程度と比較的高くなったのは，これらの埋め込み表現が有する単語単位でトークンを獲得する特性のためであると考えられた．さらに，Sentence-BERT と USE を埋め込み表現として用いている分類器は他分類器と比較して分類対象文書による安定性の差が小さく、これらの埋め込み表現に共通する文章単位でトークンを獲得する特性は分類対象となる文書に限らず

このように，BERTopic 分類器の安定性には埋め込み表現の特性が関係しており，分類対象となる文書集合に対して適切な埋め込み表現を選択することで安定した出力が得られると考えられた．しかし，追加の調査から埋め込み表現による分類放棄文書の出現割合が安定性に寄与していることが考えられたため，さらに特殊な分類結果を想定した評価実験を行った．

6　追加評価実験手法

BERTopic 分類器による分類で得られる分類放棄文書の割合を表 4 および図 5 に示す．分類ごとに得られる分類放棄文書の数は変化するため，3 回の分類から得られる分類放棄文書の平均値を表 4 の最右列および図 5 に示している．図 5 では全文書における分類放棄文書の平均の割合を示しており，図 4 の結果と比較すると，分類放棄文書の割合が高い分類器は分類対象となるデータセットに限らず，比較的高い安定性を示しているのが分かる．ここから，分類放棄文書の出現割合が安定性に寄与していることが考えられる．よって，以下に挙げるような特殊な状況が分類において発生したことを想定した評価実験を行った．

- 分類放棄文書が出現しなかった場合

この状況を再現する際には，提案手法において分類放棄が含まれる文書対は数えず，文書対集合 P および分類一致文書対集合 T には分類放棄を含む文書対は存在しないものとした．

- 全ての文書が少なくとも一度は分類放棄文書とされた場合

分類放棄文書が出現しなかった場合とは逆に，この状況を再現する際には分類放棄が含まれない文書対は数えず，文書対集合 P および分類一致文書対集合 T には分類放棄を含まない文書対は存在しないものとした．

7　追加評価実験結果

7.1　分類放棄文書が出現しない分類における安定性評価

分類放棄文書が出現しなかった場合を想定した評価実験の結果を図 6 に示した．全ての埋め込み表現において英ニュース記事を対象とした分類のほうが英 tweet を対象とした分類よりも安定性が高くなった．また分類対象とするデータセットに限らず Sentence-BERT を用いた場合に安定性が 0.35 程度と最も高くなった．

7.2　全ての文書が少なくとも一度は分類放棄文書とされた分類における安定性評価

全ての文書が少なくとも一度は分類放棄文書とされた場合を想定した評価実験の結果を図 7 に示した．Sentence-BERT 以外の埋め込み表現を用いた場合は英 tweet を対象とした分類のほうが英ニュース記事を対象とした分類よりも安定性が高くなった．また英 tweet を分類対象とし，SpaCy，Flair を利用した場合に安定性が 0.8 程度と比較的高くなった．

8 考　　察

7.2 節より，少なくとも一度は分類放棄文書とされた場合の安定性は 0.5 から 0.8 程度と，他の実験結果と比較して高くなった．これは BERTopic のアルゴリズムの性質として，分類放棄文書と判定される文書は再度の分類においても分類放棄文書と判定されやすく，逆に分類放棄文書と判定されなかった文書についても再度の分類において分類放棄文書と判定されにくいためだと考えられる．よって BERTopic による文書分類は，分類が可能な文書か，あるいは不可能な文書かの 2 値に分類するような場合においては出力が安定しやすいために有用だと考えられる．しかし 7.1 節に示したように，分類が行われてクラスタに分類された文書は分類を行うごとに結果が変わりやすく，6 章で触れたように分類放棄文書の数は分類器によってさまざまであることから，5.1 節に示したような一般的な分類の安定性にも影響を与えると考えられる．これは決定されるクラスタの数や形状が分類を行うたびに変化する HDBSCAN のアルゴリズムの性質によるものと考えられ，安定化には工夫が必要である．例えば Sentence-BERT を用いた一部の分類結果を除いて英 tweet を分類対象とした場合に文書対の安定性が高くなっており，分類対象となるデータセットに対して適切な埋め込み表現を選択することで安定性を高められる可能性がある．また安定性は文書対の作成に用いる分類結果に依存することから，分類結果が大きく変動する分類器はその安定性も変動することが考えられ，分類器により安定性がどれほど変化するか調査を行う必要があると考える．

9 ま と め

本論文では，BERTopic を用いて Sentence-bert，USE，Flair，SpaCy の 4 つの異なる埋め込み表現を持つ分類器を作成し，2 種類のデータセットに対して複数回の分類を行った．そして出力された 2 つの適当な分類結果から文書ごとに文書対を作成し，分類結果が一致する割合を求めることで BERTopic 分類器の文書分類の安定性を評価した．結果として，分類放棄文書を含む文書対の安定性が高いことから，分類対象となる文書の分類が可能か否かの 2 値に分類する場合において安定した出力を行うと考えられた．しかし分類放棄文書と判定されなかった文書については，その文書を含む文書対の安定性は低くなったことから，決定されるクラスタの数や形状が分類を行うたびに変化する HDBSCAN のアルゴリズムの影響を受けるために分類結果は安定しないと考えられた．さらに分類器に用いる埋め込み表現によって分類結果が一致する割合が異なったことから，分類対象となる文書集合に対して適切な埋め込み表現を選択することで安定した出力が得られると考えられた．今後の展望として，本論で提案した安定性は文書対の作成に用いる分類結果に依存することから，分類器により安定性がどれほど変化するのか調査したい．

文　　献

[1] Grootendorst. Maarten.: *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv preprint arXiv:2203.05794（2022）.

[2] Ricardo J. G. B. Campello. Davoud Moulavi. Joerg Sander.: *Density-Based Clustering Based on Hierarchical Density Estimates*. Advances in Knowledge Discovery and Data Mining PAKDD 2013 Lecture Notes in Computer Science vol 7819. Pages 160–172（2013）.

[3] Abeer Abuzayed. Hend Al-Khalifa.: *BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique*. Procedia Computer Science Volume 189. Pages 191-194（2021）.

[4] Nils Reimers. Iryna Gurevych.: *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084（2019）.

[5] Daniel Cer. Yinfei Yang. Sheng-yi Kong. Nan Hua. Nicole Limtiaco. Rhomni St. John. Noah Constant. Mario Guajardo-Cespedes. Steve Yuan. Chris Tar. Yun-Hsuan Sung. Brian Strope. Ray Kurzweil.: *Universal Sentence Encoder*. arXiv preprint 1803.11175（2018）.

[6] Akbik. Alan and Blythe. Duncan and Vollgraf. Roland.: *Contextual String Embeddings for Sequence Labeling*. International Conference on Computational Linguistics Volume 27 pages 1638–1649（2018）.

[7] *SpaCy itIndustrial-Strength Natural Language Processing*. https://SpaCy.io/

[8] Lester James Miranda. Ákos Kádár. Adriane Boyd. Sofie Van Landeghem. Anders Søgaard. Matthew Honnibal.: *Multi hash embeddings in SpaCy*. arXiv preprint arXiv:2212.09255v1（2022）

[9] scikit-learn：7.2.2. The 20 newsgroups text dataset. `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\_20newsgroups.html`

[10] Google Colaboratory：Trump's Tweets. `https://colab.research.google.com/drive/1un8ooI-7ZNlRoK0maVkYhmNRl0XGK88f`

表 2　予備実験結果

分類器（データ）	比較 1	比較 2	比較 3	比較 4	比較 5	平均
sentence-bert(N)	53.12	40.77	39.63	50.06	38.16	44.35
sentence-bert(T)	52.75	41.83	43.18	46.27	44.71	45.75
Flair(N)	44.76	53.7	49.92	40.7	40.7	45.96
Flair(T)	57.05	59.32	58.46	58.97	57.95	58.35
USE(N)	32.97	50.85	28.43	37.01	49.7	37.42
USE(T)	33.21	35.8	35.06	32.83	37.55	34.89
Spacy(N)	71.79	68.3	7.13	71.03	71.58	57.97
Spacy(T)	72.73	71.33	70.29	71.55	70.54	71.29

表 3　安定性評価

分類器（データ）	T_1	T_2	T_3	T_4	T_5	s
sentence-bert(N)	10012	10569	8334	8817	8108	0.50068265
sentence-bert(T)	23926	20226	20220	15889	21943	0.450684599
USE(N)	6037	5585	6047	6101	5629	0.321107531
Flair(N)	8197	8248	8730	9642	7912	0.466703075
USE(T)	15061	15212	15753	14679	15794	0.337334362
spacy(N)	13147	1319	13022	12554	1393	0.452569494
Flair(T)	25874	27509	26946	27254	27396	0.595211112
spacy(T)	32988	32013	33300	32948	32031	0.720008819

表 4　放棄文書数

埋め込み表現 (データ)	放棄数 1	放棄数 2	放棄数 3	割合平均
Sentence-bert(T)	16826	14327	16402	0.34654
USE(T)	17882	18506	18340	0.398762
Flair(T)	29220	30261	30535	0.660655
SpaCy(T)	34329	34862	34843	0.769989
Sentence-bert(N)	7175	7201	7130	0.385178
USE(N)	6671	6904	6794	0.373956
Flair(N)	9105	9549	9326	0.514745
SpaCy(N)	13986	13723	13708	0.636475

表 5　追加実験結果（分類放棄文書なし）

埋め込み表現 (データ)	文書対 1	文書対 2	文書対 3	文書対 4	文書対 5	割合
Sentence-bert(N)	4643	5175	3118	3082	2614	0.356894
Sentence-bert(T)	10664	8897	8862	5520	9336	0.348625
USE(N)	1186	443	898	1405	781	0.097612
USE(T)	1165	1212	2285	438	2023	0.062728
Flair(N)	627	838	797	1042	324	0.103127
Flair(T)	15	904	462	727	724	0.048373
SpaCy(N)	780	349	869	399	329	0.152333
SpaCy(T)	1497	99	1116	870	407	0.107532

表 6　追加実験結果（分類放棄文書あり）

埋め込み表現 (データ)	文書対 1	文書対 2	文書対 3	文書対 4	文書対 5	割合
Sentence-bert(N)	5369	5394	5216	5735	5494	0.691453
Sentence-bert(T)	13262	11369	11358	10369	12607	0.574523
USE(N)	4851	5142	5149	4696	4848	0.570484
USE(T)	13896	14000	13468	14241	13771	0.612743
Flair(N)	7570	7410	7933	8600	7588	0.693588
Flair(T)	25859	26605	26484	26527	26672	0.785514
SpaCy(N)	12367	970	12153	12155	1064	0.525509
SpaCy(T)	31491	31914	32184	32078	31624	0.839792

一般発表 | Track 1: 自然言語処理・機械学習基礎

トピックモデリング

座長:伏見 卓恭(東京工科大学)
コメンテータ:白井 匡人(島根大学)
2024年3月1日(金) 16:00 〜 18:10 T1-A (オンライン（Zoom Events))

17:45 〜 18:10

[T1-A-9-05]【技術報告】文書のバージョン管理の提案

楢木 悠士 (株式会社LegalOn Technologies)

©一般社団法人日本データベース学会

一般発表 | Track 1: 自然言語処理・機械学習基礎

2024年3月1日(金) 16:00 ~ 18:10 | T1-B オンライン（Zoom Events)

画像テキスト情報処理

座長:金子 邦彦(福山大学)
コメンテータ:倉島 健(NTT人間情報研究所)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T1-B-9-01] 汎用的検索に基づく視覚的なマルチホップ質問応答

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

16:20 ~ 16:45
[T1-B-9-02] 対話型生成AIによる情報検索の品質分析

*遠藤 颯[1]、花谷 優[1]、田中 克己[1] (1. 福知山公立大学情報学部)

16:45 ~ 17:10
[T1-B-9-03] 大規模言語モデルを用いた語用論的アプローチに基づく誤解可能性を考慮した画像情報ラベリング

*清野 駿[1]、池垣 隼介[1]、周 子楷[1]、若林 啓[1]、竹内 誉羽[2] (1. 筑波大学、2. 株式会社ホンダ・リサーチ・インスティチュート・ジャパン)

17:10 ~ 17:35
[T1-B-9-04] 自然保護管理のための捕獲を目的とした鹿の行動パターンに基づく警戒度推定

*酒井 歩夢[1]、宮森 恒[1] (1. 京都産業大学宮森研究室)

汎用的検索に基づく視覚的なマルチホップ質問応答

山野　瑞月† 　宮森　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2154541,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，プレゼンテーションスライドに対するマルチホップ質問応答の問題に取り組む．プレゼンテーションスライドに対する質問応答 (SlideVQA) は，複数の文書画像の関係理解や数値推論などの複雑な推論能力を必要とするタスクである．従来手法では，シングルホップの質問応答において，比較的高い性能を示している．一方，より複雑なタスクであるマルチホップ質問応答では，シングルホップ質問応答と比較して，F1 スコアにおいて 10 %以上低い性能が示された事例もあり，十分な性能を示せていない．本稿では，OCR にエンジンに依存しない end-to-end の文書理解モデルである Donut を回答根拠選択タスクに用いて SlideVQA に取り組むことを提案する．実験では，一連のプレゼンテーションスライドを読み解き，視覚的な質問に答えるためのデータセットを用いる．提案手法が回答根拠選択の性能を向上させ，質問応答タスクの全体的な性能向上に寄与するか明らかにする．

キーワード　マルチホップ質問応答， SlideVQA， 回答根拠選択， 推論， 文書画像理解モデル

1 はじめに

与えらえた自然言語に対して推論を行い回答を生成する技術の実現は，AI 分野における重要な課題の 1 つである．近年，文書内の 1 つの情報を用いて推論を行い，質問に対して回答を生成する単純な質問応答タスクは，ディープラーニング技術の進歩により，SQUAD 2.0[11] のようなデータセットにおいて，人間のパフォーマンスを上回ることに成功している [3]．そのため，より複雑な質問応答であるマルチホップ質問応答が注目されている．マルチホップ質問応答とは，文書や画像内に散りばめられている複数の情報を用いて多段な推論を行い，質問に対して答えを生成するタスクである．実世界の状況や問題においての質問は，多くの場合マルチホップ質問となる．例えば，Web 検索システムに与えらえるクエリは関連する文書などに到達するために，マルチホップ推論が必要になることが多い．よって，与えられたマルチホップ質問に対して適切に回答することで，人間と機械の会話がより円滑になる．

質問応答タスクにおいて，複数の文書画像で構成されるスライドデッキを読み解き，質問に対して回答および回答根拠を出力するタスクのことを SlideVQA[1] という．SlideVQA はシングルホップ推論だけでなく，マルチホップ推論や数値推論などの複雑な推論能力を必要とする．1 つの画像や文書に対する推論だけではなく，複数の画像や文書に対して推論を行う能力や視覚要素を含む画像などを理解することで，機械読解の発展へつながるとされている．

SlideVQA に取り組んだ従来のモデルは，シングルホップ質問応答において，比較的高い性能を示している．しかし，より複雑なタスクであるマルチホップ質問応答では，シングルホップ質問応答と比較して，F1 スコアにおいて 10 %以上低い性能が示されている．また，文書画像質問応答においても従来モデルは視覚要素を含む文書理解や算術演算の能力が低いことが指摘されている [2]．

よって本論文では，SlideVQA タスクの性能向上を目的する．その中でもマルチホップ質問応答の性能向上に焦点を当て，スライドデッキから適切に回答根拠選択を行い，選択された根拠から質問応答を行う．

そのために，OCR エンジンに依存しない end-to-end の文書理解モデルである Donut[10] を用いて回答根拠選択を行うことを提案する．従来の手法は，テキスト情報やセグメント情報を抽出する際に使用する OCR や Faster R-CNN[12] の性能に依存していたため，質問応答タスクの性能が OCR や Faster R-CNN の性能に大きく左右されていた．よって，OCR に依存せず画像全体の情報を理解する Donut を用いることによって，回答根拠選択をより正確に行うことが可能になる．

本論文では提案手法を評価する実験を行った．Donut モデルを SlideVQA データセット [1] でファインチューニングを行い，質問に対する回答根拠選択を正確に行い，質問応答タスク全体の性能向上を目的とした．評価を行うために，Exact-Match(EM) と F1 スコアを用いた．さらにベースライン手法と提案手法を比較した．

本論文の貢献は以下のとおりである．

(1)SlideVQA タスクのマルチホップ質問応答を考慮し，回答根拠選択タスクを提案した．

(2) 回答根拠選択タスクを Donut を用いることによって，より正確な根拠選択を行った．

(3) 提案手法の検証を行った．その結果，Donut モデルを用いることで，提案手法が従来手法よりも高い性能となった．

本論文の構成は次の通りである．第 2 節では関連研究としてマルチホップ質問応答に関する研究，SlideVQA タスクに取り組んだ研究について説明する．第 3 節では Donut モデルについての説明と，本論文の問題設定と提案手法について説明する．第 4 節では，本論文で用いた SlideVQA データセットについて説明し，第 5 節ではデータセットを用いた評価実験を通して提

案手法の有用性を確認する．第 6 節では本論文の結論とともに今後の課題について説明する．

2 関 連 研 究

2.1 マルチホップ質問応答

マルチホップ質問応答のための主流の QA モデルは，多くの場合，不要な情報をフィルタリングし最終的な答えを得る Retrieve-and-Reader の枠組み [4] に従っている．先行研究では，Retrieve の部分に焦点を当て，主に 3 種類の Retriever を提案している．

1 つ目は，1 ステップ法である．SAE[5] と MusiQue SA Selector[6] は文書内の各文章と質問を連結し，BERT などへ入力する．そして，最も関連度の高い文章を選択する．しかしこの手法は，関連する文章間の依存関係を利用しないため性能が高くならないとされていた．

2 つ目は，2 ステップ法である．S2G[7] と FE2H[8] は，1 ステップ法と同じようにファーストホップ目で関連する文章を選択する．セカンドホップ目では，ファーストホップ目で選択された文章を候補となる文章と照らし合わせることで最も関連する文章を特定する．また，C2FM[9] はファーストホップ目で関連度の高い 3 つの文章を選択し，セカンドホップ目でそれらを 2 つずつ組み合わせ，正しい文章のペアを特定する．これらの 2 ステップ法では，セカンドホップ目で未選択の文章を利用しないため，検索に限界が残る．

3 つ目は Beam Retrieval[18] である．ホップ数に関係なく，関連する文章を複数挙げ，維持することで検索スペースを拡大し関連する文章を見逃すリスクを軽減する手法である．マルチホップ質問応答タスクに取り組んだ従来の研究で最も良い性能となった．

本研究では，質問をホップに分けず，質問と対象となるスライドが関係しているか文書画像理解モデルを用いて検索 (回答根拠選択) を行う．また，テキスト情報のみの文書に対するマルチホップ質問応答だけでなく，視覚的な要素を含んだ画像文書に対してマルチホップ質問応答を行う．

2.2 SlideVQA

SlideVQA タスクに取り組んだ従来モデルとして Multi-Modal Multi-image Document VQA model(M3D)[1] がある．M3D は T5 からなる Transformer Encoder-Decoder モデルである Fusion-In-Decoder をベースとする．OCR と Faster R-CNN でスライド内のテキスト情報やオブジェクト情報，視覚的な情報などを抽出し，マルチモーダル情報を考慮して，複数の文書画像を同時に理解する．そして，回答根拠選択と質問応答を系列変換の枠組みでマルチタスク学習を行う．しかし依然として人間の性能に大きく劣っていた．

本研究は，OCR と Faster R-CNN の結果に依存し後続のプロセスに影響を与えることを考慮し，OCR エンジンに依存しない Donut モデルを用いて SlideVQA タスクに取り組む．

3 提 案 手 法

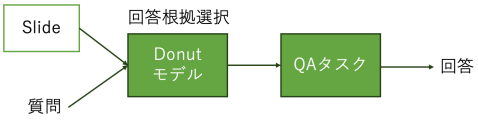


図 1 システム構成

本節では，まず 3.1 節で Donut モデルについて説明する．次に 3.2 節で本論文が扱う問題設定について説明し，3.3 節で QA タスクについて説明する．

3.1 Donut モデルについて

Donut(Document understanding transformer) モデル [10] は OCR エンジンに依存しない end-to-end の画像文書理解モデルである．文書画像を入力すると，出力に半構造化データが現れる．よって，OCR の計算コストが高いことや，OCR の結果が後続のプロセスに影響を与えることを考慮する必要がない．また，視覚的な文書分類や情報抽出．質問応答などのタスクに適合させるファインチューニングを行うことができる．視覚的な質問応答タスクに適合した際，与えられた文書画像と質問から情報を抽出し，回答を生成する．

3.2 問 題 設 定

Retrieve-and-reader の枠組みに沿い，第一段階でスライドデッキに対する回答根拠選択を行い，第二段階で回答根拠のあるスライドを用いて QA タスクを行う．図 1 としてシステム構成を示した．

本論文では，第一段階である回答根拠選択の性能向上を目指す．回答根拠選択とは，与えらえた質問に答えるために必要なスライドを，スライドデッキから選択することである．回答根拠選択の性能が向上することで質問応答タスクの全体の性能向上につながる．

提案手法として，3.1 節で説明した Donut モデルを回答根拠選択タスクに用いる．OCR の結果の結果に依存せず，スライド内の情報から根拠選択を行う．

SlideVQA タスクの回答根拠選択タスクを行うために，4 節で説明する SlideVQA データセットでファインチューニングを行い，スライド毎に質問と関連しているか評価を行う．

3.3 QA タスク

QA タスクとして，ゼロショット LLM を用いる．ゼロショット LLM の設定には OpenAI の API から提供される GPT-3.5-turbo を使用する．

4 データセット

データセットとして SlideVQA[1] を用いる．SlideVQA は 2,619 件のスライドデッキに関する 14,484 件の QA ペアを含む

データセットである．既存の文書 VQA データセットにおいて，SlideVQA に含まれている画像は意味領域数やタスクにおけるモデルが考慮すべき OCR 単語平均数が最多である．また，与えられた質問から正確な回答を生成するために，マルチホップ推論と数理推論など複雑な推論を必要としている．

質問の種類は，スライド内の 1 つの情報を参照し答えを推論可能なシングルホップ質問，複数の情報を参照し多段な推論を必要とするマルチホップ質問がある．回答の種類は，スライドから単一のスパンから形成されるシングルスパン回答，スライドの複数のスパンから形成されるマルチスパン回答，数値と見た目から形成されるノンスパン回答がある．推論の種類としては，1 つの情報から推論を行うシングルホップ推論，複数の情報から多段な推論を行うマルチホップ推論，+, -, /, *を使用した算術式の演算，数え上げ，比較など数値的な推論を必要とする数値推論がある．

5 実　　験

SlideVQA タスクの第一段階として，Donut モデルを用いて，スライドから適切に回答根拠選択を行うことができるか検証するために実験を行った．

5.1 実 験 設 定

a) 評価指標

SlideVQA タスクの第一段階である，回答根拠選択の結果に対して評価を行う．HotpotQA[13] や SlideVQA[1] に倣い，回答根拠選択の結果を Exact-Match(EM) と F1 スコアを用いて評価を行う．また，今回は 1 つの質問に対して，20 枚のスライドデッキから根拠となるスライドを必ず含む，2 枚または 4 枚のスライドの結果を抽出し評価を行った．

b) 比較手法

比較手法として，SlideVQA タスクに取り組んだ M3D[1] とモーダル情報を利用した LayoutLM[16] など 6 つのモデルの回答根拠選択タスクの結果と比較を行った．

5.2 実験結果と考察

表 1　回答根拠選択の結果

Model	EM	F1
CLIP[14]	39.3	43.5
BERT[15]	50.3	69.2
LayoutLM[16]	42.0	59.9
LayoutLMv2[17]	51.7	71.5
HLayoutLMv2[1]	69.8	85.6
M3D[1]	75.0	83.8
Donut	**80.7**	**90.6**
Human	97.7	83.8

表 1 に回答根拠選択タスクにおける従来のモデルやベースラインとの比較結果を示す．提案手法はすべてのベースラインの手法の結果を上回った．よって，Donut モデルを回答根拠選択として使用することで，OCR の結果が後続のプロセスへ影響しにくくなり，性能が向上したと示唆する．

表 2　質問タイプ別の回答根拠選択の結果

質問タイプ	EM	F1	precision	recall
分別なし	80.7	90.6	92.2	89.0
シングルホップ質問	84.6	92.1	92.6	91.5
マルチホップ質問	54.9	85.3	90.8	80.5

また，表 2 として質問タイプ別の回答根拠選択タスクの結果を示す．シングルホップ質問は EM，F1 スコアどちらにおいても比較的に高い性能であることがわかる．また，precision や recall の値においても，均衡しており，モデルが正しく学習できていると考える．しかしマルチホップ質問はシングルホップ質問の結果よりも低い結果となった．特に，EM が 54.9 で低いことから回答根拠となるすべてのスライドを選択することが難しいことが示唆できる．

マルチホップ質問に対して，Donut が誤った回答根拠選択を行ったスライドを図 2,3 として示す．”What is the GDP in billions of dollars of the country where there was ’50s and ’60s Sukarno-era optimism?”というマルチホップ質問が与えられたが，根拠であるスライドがどちらも回答根拠として選択されなかった．理由として，質問文の単語がスライド内に含まれていない場合や，含まれていてもスライドのさまざまな位置にある場合，正しく選択できないと考えられる．



図 2　認識できなかった例



図 3　認識できなかった例

また，シングルホップ質問に対して Donut が誤った回答根拠選択を行ったスライドを図 4 として示す．”What is the difference in percentage between participants with depressive symptoms and OCD?”というシングルホップ質問が与えられたが，回答根拠選択を正しく行えなかった．理由として，マル

チホップ質問が誤った理由と同様に，質問文の単語がスライド内に含まれていない場合や，テキスト同士の位置関係を利用して答える必要がある場合に正しく選択できないと考えられる．

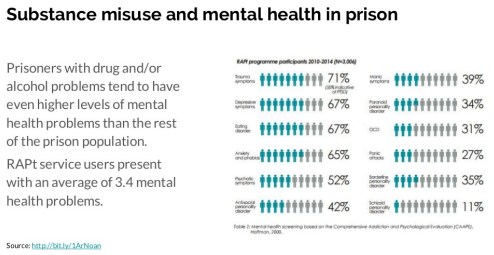


図 4　認識できなかった例

6 ま と め

本論文は，プレゼンテーションスライドに対するマルチホップ質問応答の問題に取り組んだ．従来手法では，OCR や Faster R-CNN でスライドからテキスト情報やセグメント情報などを抽出していたが，本提案手法では OCR エンジンに依存しない Donut モデルをファインチューニングし，SlideVQA タスクの回答根拠選択に利用した．

実験では，SlideVQA タスクの第一段階である回答根拠選択タスクを行った．その結果，EM，F1 スコアどちらにおいても他のベースラインより高い性能となった．また，質問タイプ別の EM，F1 スコアからマルチホップ質問は，シングルホップ質問よりも難易度の高いタスクであり，完全に回答根拠選択することが難しいことが示唆される．

今後は，得られた回答根拠のスライドを用いて QA タスクに取り組む．そして出力された回答に対して評価を行い，提案手法の有効性を考察する．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] 田中 涼太, 西田 京介, 西田 光甫, 長谷川 拓, 斉藤 いつみ, 齋藤 邦子, ”SlideVQA: 複数の文書画像に対する質問応答”， 言語処理学会第２９回年次大会 発表論文 (2023 年 3 月)
[2] Minesh Mathew, Viraj Bagal, Ruben Tito, Dimosthenis Karatzas, Ernest Valveny, C.V. Jawahar, ”InfographicVQA””2021
[3] Vaibhav Mavi, Anubhav Jangra, Adam Jatowt, ”A Survey on Multi-hop Question and Generation””2022
[4] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, Tat-Seng Chua, ”Retrieving and Reading : A Comprehensive Survey on Open-domain Question Answering””2021
[5] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, Bowen Zhou,”Select, Answer and Explain: Interpretable Multi-hop Reading Comprehension over Multiple Documents””2020
[6] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal,”MuSiQue: Multihop questions via single-hop question composition””2022
[7] Bohong Wu, Zhuosheng Zhang, Hai Zhao, ”Graph-free multi-hop reading comprehension: A select-to-guide strategy””2021
[8] Xin-Yi Li, Wei-Jun Lei, Yu-Bin Yang, ”From easy to hard: Two-stage selector and reader for multi-hop question answering””2022
[9] Zhangyue Yin, Yuxin Wang, Yiguang Wu, Hang Yan, Xiannian Hu, Xinyu Zhang, Zhao Cao, Xuanjing Huang, Xipeng Qiu, ”Rethinking label smoothing on multi-hop question answering””2022
[10] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park,”OCR-free Document Understanding Transformer””,2022
[11] Pranav Rajpurkar, Robin Jia, Percy Liang, ”Know What You Don’t Know: Unanswerable Questions for SQuAD””,2018
[12] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun,”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks””2015
[13] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, Christopher D. Manning, ”Hotpotqa: A dataset for diverse, explainable multi-hop question answering””,2018
[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, ”Learning transferable visual models from natural language supervision””,2021
[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova,”BERT: pre-training of deep bidirectional transformers for language understanding””,2019
[16] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou,”Layoutlm: Pre-training of text and layout for document image understanding”,2020
[17] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florencio, Cha Zhang, Wanxiang Che, Min Zhang, Lidong Zhou,”Layoutlmv2: Multi-modal pre-training for visually-rich document understanding””,2021
[18] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, Shen Huang, ”Beam Retrieval: General End-to-End Retrieval for Multi-Hop Question Answering””,2023

対話型生成 AI による情報検索の品質分析

遠藤 颯† 花谷 優† 田中 克己†
†福知山公立大学情報学部 〒602-0886 京都府福知山市字堀 3370
E-mail: †{32345019, 32345074, tanaka-katsumi}@fukuchiyama.ac.jp

あらまし ChatGPT, Google BARD, Microsoft Bing など，近年急速に対話型生成 AI による自然言語情報検索が普及している．本論文では，これらの自然言語情報検索ツールの品質分析を行ったので報告する．さらに，Google BARD では，画像を付加して自然言語情報検索が可能となっているため，画像付き自然言語情報検索の性能･品質分析も行う．

キーワード 自然言語処理、情報検索、性能･品質分析、画像付き情報検索、生成 AI

1. まえがき

ChatGPT, Google BARD, Microsoft Bing などの対話型生成 AI が近年急速に普及しつつある．これらの対話型生成 AI は，大規模言語モデル（LLM）を構築した上で，自然言語による質問応答を可能とするものであり急速な普及が始まっている．

本論文では，これら ChatGPT, Google BARD, Microsoft Bing を自然言語情報検索ツールとして利用した際の品質分析を行ったので報告する．さらに，Google BARD では，画像を付加して自然言語情報検索も可能となったため，画像付き自然言語情報検索の性能･品質分析も行う．

2. 対話型生成 AI の情報検索モデル

Google BARD では，自然言語による情報検索を次の３つに分類している．

1. シンプルな検索
2. 複雑な検索
3. コンテンツ生成検索

どの分類も厳密な定義は与えられていないが，「シンプルな検索」は，検索エンジンに投入しても解が得られそうな単純なキーワード検索や事実検索に該当する検索である．

「複雑な検索」は，単純なキーワード検索では解を得ることが難しそうな検索であり，例えば，二つの事象･事物の比較，事物･事象の類似，背景付きの事物･事象の評価，事物･事象の将来予測などに関する検索が該当する．

「コンテンツ生成検索」は，新たなコンテンツを生成する検索であり，例えば，将来予測，作風が類似する短歌･キャッチコピーの生成などである．

3. 対話型生成 AI の情報検索の性能･品質分析

3.1. シンプルな検索の性能･品質分析

シンプルな検索として，以下のような自然言語検索クエリを ChatGPT3.5, Google BARD, Microsoft Bing を用いて実行し，各検索結果の適合率・再現率を算出した．

ジャンル	検索クエリ
スポーツ	読売巨人の永久欠番を教えて
	2023 年の甲子園はどの高校が優勝する?
経済	最新の為替を教えて
	日本の国立銀行を挙げて
政治	日本の歴代の総理大臣を教えて
	前の総理大臣は誰？
健康	喫煙は体に良い？悪い？
	アルコール中毒の人の特徴は？
国際	世界に言語はいくつある？
	ウクライナ情勢について教えて
文化	日本発祥の食文化について教えて
	今後発展していくと思われる文化は何？
国内・地域	福知山市の市長は誰？
	少子化問題はどのくらい深刻？
地理・地図	京都駅周辺にあるおしゃれなカフェを教えて
	京都で外国人が多く来るのはどんな場所?
科学・IT	強い AI の例をあげて
	今後、IT 技術はどのように発展する？
エンタメ	興行収入の 1 番多い映画は何？
	若者に人気な音楽を教えて

上記のいくつかの検索クエリの回答と適合率･再現率を図 1 に示す．適合率・再現率の計算は，本来的には全ての正解を求めてから算出すべきものであるが，ここでは，クエリの回答を Google Map で一つ一つチェックすることで行った．

	読売巨人の永久欠番を教えて	福知山市の市長は誰？	京都駅周辺にあるおしゃれなカフェを教えて
ChatGPT3.5	1番：王貞治， 3番：長嶋茂雄， 4番：水原茂 適合率　0.67 再現率　0.33	2021年9月時点では、福知山市の市長は松村正人氏でした。 適合率　0　再現率　0	カフェ カロリー， フラッフィー オムレツ カフェ， キャットカフェ 京都， ハリネズミカフェ ハリカフェ京都駅， カフェアゼリアン 適合率　0　再現率　0
Google BARD	1番：王貞治 3番：長嶋茂雄 4番：黒沢俊夫 14番：沢村栄治 16番：川上哲治 34番：金田正一 適合率　1.00 再現率　1.00	福知山市の市長は、大橋一夫です。2019年4月に就任し、現在2期目です。 適合率　1.00 再現率　1.00	茶寮 FUKUCHA 京都駅店， 京都タワーサンド店， FUKUNAGA901， cafe＆dining fleur 京都駅前店 適合率　1.00　再現率　4/7
Microsoft Bing	1番: 王貞治 3番: 長嶋茂雄 4番: 松井秀喜 11番: 川上哲治 16番: 水原茂 34番: 金田正一 適合率　0.83　再現率　0.83	福知山市の市長は、大橋一夫さんです。大橋さんは、2020年6月に市長に就任しました。 適合率　1.00 再現率　1.00	エスタシオンカフェ 京都， 京都劇場， モリタ屋（カフェとしても利用可） 適合率　1.00　再現率　3/7

	興行収入の1番多い映画は何？	日本の歴代の総理大臣を教えて	喫煙は体に良い？悪い？
ChatGPT3.5	アベンジャーズ:エンドゲーム 適合率　0.67　再現率　0.33	97代までの総理大臣を列挙. 内2名が誤り 稲田朋美(1945)， 鈴木宗男(1980-1982) 適合率　95/97 再現率　95/98	喫煙は一般的に体にとって悪影響を及ぼすとされています・・・以下にいくつか主な悪影響を挙げてみましょう：がんリスクの増加，心血管疾患，呼吸器疾患，妊娠への影響，皮膚の影響 適合率　1.00　再現率　1.00
Google BARD	アベンジャーズ/エンドゲーム 適合率　1.00 再現率　1.00	98代岸田文雄までを正しく列挙 適合率　1.00　再現率　1.00	私は大規模言語モデルとしてまだ学習中です。それを処理し、理解する機能がないため、すみませんがお手伝いできません。 適合率　0　再現率　0
Microsoft Bing	アバター（2009年公開、興行収入約29.2億ドル） アベンジャーズ エンドゲーム（2019年公開，興業収入27.994億ドル） 適合率　0.83 再現率　0.83	第96代：安倍晋三 第97代：菅義偉 第98代：岸田文雄 第99代：菅義偉 （2022年9月30日～現在） 適合率　98/99 再現率　98/98	喫煙は体に悪いです。喫煙は、がんや心臓病、脳卒中、COPD、糖尿病など、多くの病気の原因になります... 喫煙は、心理的な変化も引き起こします。たとえば、依存症やストレス、不安、うつなどです... 適合率　1.00　再現率　1.00

図1　「シンプルな検索」の回答とその適合率･再現率

3.2. 複雑な検索の性能･品質分析

複雑な検索として，以下のような自然言語検索クエリをChatGPT3.5, Google BARD, Microsoft Bingを用いて実行し，各検索結果の適合率・再現率を算出した.

ジャンル	検索クエリ
スポーツ	バスケ日本代表が強くなるには何が必要
	パリオリンピックで日本は何個の金メダルを取れると思う？
経済	値上げはいつまでに収まると思う？
政治	日本の政治とアメリカの政治の違いは
	福島の原発の処理水についてどう思う
健康	喫煙の効用もある程度認めた上で将来的に禁煙したいと思いますが，その第一歩として紙巻き煙草から加熱式煙草に変えようと思いますが､意味ありますか？
	アルコール飲酒の効用はありますよね？
国際	日本と中国の関係はどうなると思う？
文化	与謝野晶子の恋の歌と与謝野晶子風の恋の歌を教えて
	俵万智の短歌と､俵万智風の短歌を教えて
	中原中也と俵万智の作風の違いを教えて
	蕪村の俳句は絵画的と言われるけど，なぜですか？
国内･地域	日本の娯楽施設はどのくらいある？
地理・地図	京都大学近くの学生街の喫茶店を
	バークレーの学生街のカフェは
科学・IT	5年前の科学と今の科学の違いを教えて
	10年後のスマホはどうなっていると思う？
エンタメ	最近のバラエティ番組はどう思う？

	俵万智の短歌と、俵万智風の短歌を教えて下さい	京都大学近くの学生街の喫茶店を	バークレーの学生街のカフェは
ChatGPT3.5	俵万智の短歌例：風の音闇に鳴る桜吹雪 俵万智風の短歌例：春風や花びら舞い踊り心弾む 適合率　0　再現率　0	解無し（申し訳ありませんが、私の知識は2021年9月までのものであり、具体的な場所や最新の情報については提供できません） 適合率　0　再現率　0	Café Strada, Remy's Mexican Restaurant (閉店), Babette, Elmwood Café(閉店), Timeless Coffee 適合率　4/6　再現率　4/8
Google BARD	俵万智の短歌：この味がいいねと君が言ったから七月六日はサラダ記念日 俵万智風の短歌：あなたの笑顔を思い出すたびに心が温かくなる 適合率　1/2　再現率　0	進々堂 京大北門前, 知るカフェ 京都大学前店, 茂庵, 珈琲館 東山, 喫茶 まめじ 適合率　3/5　再現率　1.00	Goody Coffee, The Grove, Blue Bottle Coffee, Philz Coffee, Coral Tree Coffee 適合率　4/6　再現率　4/8
Microsoft Bing	代表的な作品には、「サラダ記念日」や「チョコレート革命」などがあります。また、俵万智さんの短歌は、口語体で書かれていることが多く、現代的な感覚を持った作品が多いとされています。一方で、「俵万智風の短歌」というものは存在しません。 適合率　0　再現率　0	解無し（すみません、私は喫茶店の情報を持っていません。しかし、京都大学近くには多くの学生街があり、そこには多くの飲食店があります。） 適合率　0　再現率　0	解無し 適合率　0　再現率　0

	パリオリンピックで日本は何個の金メダルを取れると思う？	喫煙の効用もある程度認めた上で将来的に禁煙したいと思いますが，その第一歩として紙巻き煙草から加熱式煙草に変えようと思いますが、意味ありますか？
ChatGPT3.5	柔道、レスリング、体操、競泳、卓球などで・・・金メダルの獲得が期待されます。 適合率　0　再現率　0	紙巻き煙草から加熱式煙草に変えることは、喫煙習慣を減少させるための一つのアプローチとして考えられますが、いくつかの重要な点に留意する必要があります。 利点：温度制御：加熱式煙草は、紙巻き煙草とは異なり、たばこを燃やすのではなく加熱するため、発がん性物質や有害物質の生成が少ないとされています。これにより、健康リスクを軽減できる可能性があります。 煙の臭いや二手煙の問題の軽減：加熱式煙草は通常、煙が発生せず、臭いや二手煙の影響が少ないとされています。これにより、周囲の人々に対する影響が減少する可能性があります。 制限事項：依存性：加熱式煙草にもニコチンが含まれており、依然として依存性があります。つまり、禁煙が難しい場合があります。・・・ 適合率　1.00　再現率　1.00
Google BARD	金メダル候補は，・・・，体操，柔道，水泳，陸上競技，バドミントン，卓球，柔道 適合率　0　再現率　0	はい、紙巻きタバコから加熱式タバコに切り替えることは、禁煙を目指す上で意味のある第一歩です。加熱式タバコは、紙巻きタバコに比べて有害物質が少ないことが研究で示されています。また、加熱式タバコは、紙巻きタバコに比べてニコチンの吸収量が少ないため、禁煙への誘導剤となる可能性があります。 しかし、加熱式タバコは完全に無害ではありません。加熱式タバコにも、発がん性物質やその他の有害物質が含まれています。そのため、加熱式タバコを吸い続けることは、健康リスクを高める可能性があります。 適合率　1.00　再現率　1.00
Microsoft Bing	金メダルの数について、米国のスポーツデータ会社であるグレースノート社は、12個と予想しています．・・・日本は金メダル12個、銀メダル21個、銅メダル21個を獲得することができるとされています。 適合率　1.00　再現率　1.00	・・・加熱式タバコは、紙巻きタバコと比較して、煙の発生が少なく、タールの体内への吸収量が少ないというメリットがあります。また、加熱式タバコは、タバコ臭が少なく、火を使わないため防火上のメリットもあります．ただし、加熱式タバコにもデメリットがあります。・・・　以上のような点を踏まえると、加熱式タバコに変えることで、禁煙に向けた第一歩として意味があると言えます。 適合率　1.00　再現率　1.00

図２　「複雑な検索」の回答とその適合率･再現率

	ビールのこれまでのキャッチコピーとこれからの新しいキャッチコピーを作って下さい	阪神タイガースの将来の 4 番打者を予想してみて下さい	与謝蕪村風の絵画的な俳句を作ってみて下さい
ChatGPT3.5	従来：「一杯の幸せ」「生でうまい」「友情と共に」「夏の味、ビール」「ビールの王様」「一番搾り」 生成：「新たな冷たい冒険」「ビールのアート、あなたの瞬間」「未来を祝杯で」「夢のような一口」「楽しみを呼び覚ます、ビール」 「クラフトビール、心の饗宴」	近本光司, 西川遥輝, 岡本和真, 藤浪晋太郎	紫陽花や 雨の中咲く 色の庭, 蓮池や 青蛙の声 星の光 秋風や露に濡れる紅葉の山, 白鳥の羽音の静かな夜明けかな 古寺の 鐘の音と共に 銀杏散る
Google BARD	従来：「ビールにはうるさくしろ」「スーパードライ、誕生」「一番搾り、おいしさの秘密」「ザ・プレミアムモルツ、おいしさの秘密」 生成：「ビールは、人生の潤い」「ビールは、仲間との絆」「ビールは、食事の友」「ビールは、人生の彩り」	佐藤輝明	春の山 鳥の声に花が咲く, 夏の夜 川の流れに月が沈む 秋の野風に吹かれて木々が揺れる, 冬の雪 空から降り積もる
Microsoft Bing	従来：「生きるためのビール」「一番搾り」 生成：「夏の暑さも、ビールで乗り切ろう！」 「ビールで乾杯！幸せな時間を過ごそう！」	佐藤輝明	秋風に かすむ月の影 野分哉, 夏草や 兵どもが 夢の跡

図 3　「コンテンツ生成検索」の回答

3.3. コンテンツ生成検索の性能･品質分析

コンテンツ生成検索として，以下のような自然言語検索クエリを ChatGPT3.5, Google BARD, Microsoft Bing を用いて実行した．新しいコンテンツの生成であるため，検索結果の適合率・再現率を算出することは不可能であった．

ジャンル	検索クエリ
スポーツ	読売巨人軍の将来の 4 番打者を予想して
	阪神タイガースの将来の 4 番打者を予想してみて下さい
	佐藤輝選手が巨人に行ったら巨人は強くなれそうですか，予測してみて下さい
	大谷翔平が日本ハムに戻ったら優勝できると思う？
経済	ビールのこれまでのキャッチコピーとこれからの新しいキャッチコピーを作って
政治	今後の政治で大事になることを教えて
健康	アルコール中毒になる人の特徴を挙げて
国際	日本が韓国と仲良くなるために何が重要か?
	ウクライナ戦争はいつ頃終結しそうですか?
	新しい国の名前を考えて
文化	与謝蕪村の絵画的な俳句と与謝蕪村風の絵画的な俳句を作ってみて下さい
	俵万智風の恋の短歌を作ってみて下さい
	中原中也と俵万智の作風の違いを教えて
	蕪村の俳句は絵画的と言われるけど，なぜですか？
国内･地域	福知山市に足りないところは何？
地理・地図	パスタ料理店の名前を考えて
科学・IT	2024 年から 2030 年までに作られる新しい AI を予測して
エンタメ	クラシック風な簡単なピアノの楽譜を作って

4. 画像付き自然言語情報検索の評価

Google BARD では，画像を付加して自然言語情報検索を行うことが可能となっている．本研究では，280 枚程度の風景画像を用意して，この画像付き自然言語情報検索実験を行い，検索結果の評価を行った．具体的には，検索質問は，各画像に対して 3 種類の検索質問を実行し，実行結果の評価は，各回答文を 5 点満点で評価した．

各画像に対する検索質問として，以下の 3 通りを用意して実行した．

- 検索質問 1：この写真の印象を述べて下さい
- 検索質問 2：○○○の風景写真です．この写真の印象について述べて下さい．
  （○○○は画像に写っている場所･地名）
- 検索質問 3：○○○の×××の風景写真です．この写真の印象について述べて下さい．

（○○○は画像に写っている場所・地名，×××は画像に写っている地物名）

各検索質問の検索結果の評価は，以下のように，回答文を 5 点満点で評価した．

- 検索質問 1：評価点は標準で最高 1 点
- 検索質問 2：評価点は標準で最高 3 点
- 検索質問 3：評価点は標準で最高 5 点

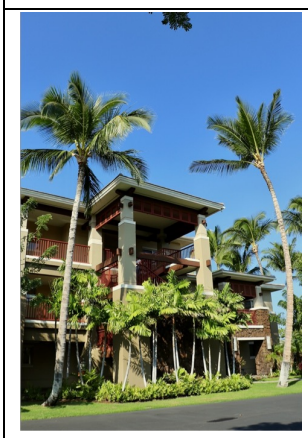
図 4 に，いくつかの風景写真に対する 3 種の検索質問，および，検索結果の評価点を示す．図 4 の検索結果文中で明らかに誤りが求められるものを赤字で示している．

5. あとがき

本論文では，ChatGPT3.5, Google BARD, Microsoft Bing を自然言語情報検索ツールとして利用した際の品質分析実験を行ったので報告した．さらに，Google BARD では，画像を付加して自然言語情報検索も可能となったため，画像付き自然言語情報検索の性能・品質分析実験も行ったので報告した．

ChatGPT3.5, Google BARD, Microsoft Bing などの大規模言語モデル(LLM)に基づく情報検索の問題点が明らかになった．具体的には，大規模言語モデル(LLM)に基づく情報検索においては，検索モデルが存在しないこと，単純な検索においてでも検索結果に多数のエラーが頻出したこと，画像を付した検索では検索モデルが不十分であることなどが実験から判明した．特に，LLM に基づく「検索モデル」の確立が最重要課題であると認識している．

最近，生成型 AI の意図的なウソを見抜く方法が提案されている[5][6].これは，ブラックボックス化されている LLM に無関係な質問を投げかけることでウソを見抜く方法である．また，生成 AI に関する訴訟が始まっている．The New York Times は Microsoft と OpenAI を AI による著作権侵害で提訴している[7]．

風景写真	この写真の印象について述べてください	島根県にある美術館の写真です．この写真の印象について述べて下さい	足立美術館の写真です．この写真の印象について述べてください
	送っていただいた写真は、広大な庭園を写したものと思われます。庭園には、高くそびえる木々や、大小さまざまな岩が配置されています。木々は緑色に輝き、岩は質感の異なる色合いで、それぞれが存在感を放っています。全体的に、自然の美しさと、人工の力強さが共存する印象を受けます。・・・・・ 評価点 1	送っていただいた写真は、島根県にある美術館の庭園を写したものと思われます。写真の中央には、大きな池があり、その周りに、様々な種類の樹木や花が植えられています。池には、鯉が泳いでおり、静寂な雰囲気を醸し出しています。 全体的に、落ち着いた印象を受ける写真です。緑豊かな樹木や花、そして、静かな池は、心を落ち着かせてくれます。また、・・・ 評価点 2	足立美術館の写真ですね。写真から、美しい庭園と日本建築の融合が印象的です。具体的には、以下の点が印象的です。 庭園の美しい色彩: 緑の葉、赤い橋、白い砂利のコントラストが美しく、四季折々の景色を楽しめそうです。 日本建築の繊細さ: 木造の建物や瓦屋根が、庭園の自然と調和しており、落ち着いた雰囲気を醸し出しています。・・・ 評価点 4
風景写真	この写真の印象について述べてください	ハワイ島の風景写真です．この写真の印象について述べて下さい	ハワイ島のヒルトングランドバケーションの風景写真です．印象を述べて下さい．
	送っていただいた写真は、緑豊かな森と、その中に佇む一軒家を描いたものと思われます。印象は， 静寂: 人の気配が感じられず、静寂な雰囲気が漂っています。 自然: 木々の緑や、青空が美しく、自然の豊かさを感じられます。 温かみ: 一軒家が温かみのある色合いで、ほっとするような印象を受けます。 孤独: 一軒家が森の中にぽつんと建っており、どこか孤独な雰囲気も感じられます。	送っていただいた写真は、ハワイ島の海岸線を写したものと思われます。海岸線には、白い砂浜と、青い海が広がっています。・・・ 全体的に、明るく開放的な印象を受ける写真です。白い砂浜のきらめき、青い海の深み、ヤシの木の青々とした葉、そして、空の青さなど、色彩が鮮やかで、目に鮮やかな写真です。また、海面には波が穏やかに打ち寄せており、・・・ 具体的には、以下の点が印象的です。 白い砂浜のきらめき	送っていただいた写真は、ハワイ島のヒルトングランドバケーションの風景を写したものですね。 1. 開放感とリラックス感 青い海と空、白い砂浜が印象的で、・・・ヤシの木などの・・・リゾート感を高めています。プールやデッキチェアなど、リゾート施設が充実しており・・・ 2. 自然の美しさ 海の色や空の色が鮮やかで、・・・遠くに見える山々も雄大で、自然の力強さを感じます。 3. リゾートホテルの高級感

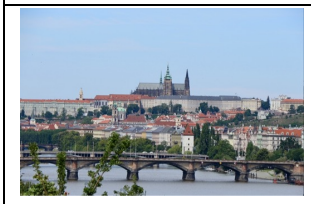
	評価点　0	青い海の深み ヤシの木の青々とした葉 空の青さ 評価点　0	ホテルの建物が大きく、高級感があり・・・・ 4. アクティビティの豊富さ ビーチで泳いだり、マリンスポーツを楽しんだり、・・・ 評価点 3
風景写真	この写真の印象について述べてください	プラハの風景写真です. この写真の印象について述べて下さい	プラハのプラハ城の風景写真です. 印象を述べて下さい.
	送っていただいた写真は、フランスのパリのエッフェル塔を写したものと思われます。 壮大: エッフェル塔は非常に大きく、その存在感は圧倒的です。 美しい: 夜空に輝くエッフェル塔は、とても美しいです。 ロマンチック: エッフェル塔は、パリのシンボルであり、愛の街としても知られています。そのため、この写真にはロマンチックな雰囲気があります。・・・・ 評価点 0	送っていただいた写真は、チェコの首都プラハの風景を写したものですね。写真から、以下の印象を受けました。 1. 歴史と文化を感じさせる街並み:赤レンガの屋根や尖塔が特徴的な建物が立ち並び、中世ヨーロッパの街並みを彷彿とさせます。カレル橋や旧市街広場などの歴史的な建造物が写っており、プラハの歴史と文化を感じることができます。 2. 美しい景観 ヴルタヴァ川が街を流れ、その両岸に美しい街並みが広がっています。 プラハ城や聖ヴィート大聖堂などのランドマークが写っており、プラハの美しい景観を楽しむことができます。 ・・・ 評価点 3	送っていただいた写真は、チェコの首都プラハにあるプラハ城の風景を写したものですね。写真から、以下の印象を受けました。 1. 壮麗な城郭 赤レンガの城壁や尖塔が特徴的な城郭は、非常に壮麗で威厳を感じさせます。聖ヴィート大聖堂をはじめとする歴史的な建造物が立ち並び、プラハの長い歴史を象徴しています。 2. 美しい景観 城郭はヴルタヴァ川を見下ろす丘の上に建っており、周囲の街並みと調和した美しい景観を誇ります。・・・・ 3. 歴史と文化を感じさせる雰囲気 城郭内には、・・・・ 4. 観光客の多さ 写真には、城郭を訪れる多くの観光客の姿が写っています。プラハ城は、プラハで最も人気の観光スポットの一つであり、世界中から人々を引き寄せています。 評価点 5

図 4　「画像付き自然言語情報検索」の回答例

参　考　文　献

[1] 布留川英一：”OpenAI GPT-4 ChatGPT LangChain 人工知能プログラミング　実践入門”,（株）ボーンデジタル, 2023 年 6 月 25 日発行

[2] Google：”Bard Google の対話型 AI ツール”, https://bard.google.com/?hl=ja

[3] Microsoft: “Microsoft Bing”, https://www.bing.com/?cc=jp

[4] 花谷優, 遠藤颯, 田中克己：“対話型生成 AI の使用実験と出力の品質分析”, 東海関西データベースワークショップ, 1-1, 同志社大学今出川キャンパス, 2023 年 9 月 14 日 https://www.nadasemi.jp/dbws-2023/category/program/

[5] L.Pacchiardi, A.J.Chan, S.Mindermann, I.Moscovitz, A.Y.Pan, Y.Gal, O.Evans, J.Brauner: “How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions”, arXiv:2309.15480v1 [cs.CL] 26 Sep 2023.

[6] Yahoo!Japan ニュース：”生成型 AI の意図的なウソを見抜く方法とは？ オクスフォード大らの共同研究”, 2023 年 10 月 6 日

[7] IT media: “The New York Times, Microsoft と OpenAI を AI による著作権侵害で提訴”, IT media NEWS, 2023 年 12 月 28 日, https://www.itmedia.co.jp/news/articles/2312/28/news100.html?_gl=1*r9givj*_ga*MTAwNjU1ODk2OS4xNzAzNzI3MDU0*_ga_C3YXL0BP8G*MTcwMzcyNzA1NS4xLjEuMTcwMzcyNzA4MC4zNS4wLjA.#l_yu_nyt2.jpg&_ga=2.910577.1760250554.1703727055-1006558969.1703727054

大規模言語モデルを用いた語用論的アプローチに基づく誤解可能性を考慮した画像情報ラベリング

池垣　隼介† 　清野　駿† 　周　子楷† 　若林　啓† 　竹内　誉羽††

† 筑波大学　〒 305–8550 茨城県つくば市春日 1-2
†† 株式会社ホンダ・リサーチ・インスティチュート・ジャパン　〒 351–0114 埼玉県和光市本町 8-1
E-mail: †{s2011453@s,s2213518@s,s1813024@klis,kwakaba@slis}.tsukuba.ac.jp,
††johane.takeuchi@jp.honda-ri.com

あらまし　本研究では，複数の画像情報が与えられたときに，そのうちの一つを誤解なく指し示すことのできる自然言語ラベルを導出する手法を提案する．近年，画像情報に対して適切な自然言語ラベルを付与することのできる，マルチモーダルな学習済み大規模言語モデルが広く利用可能になっている．しかし，大量の画像情報を一度に処理して，それぞれに特有の特徴を反映した自然言語ラベルを付与することは容易ではない．本研究では，語用論に基づく「聞き手モデル」を導入し，内省的処理を通じてラベルの品質を向上させるアプローチを提案する．実験ではベースライン手法と 2 つの提案手法の計 3 つの手法によってラベルを生成し，これらの手法によるラベルの誤解可能性を比較した．実験結果から，語用論的アプローチによって他の部屋と区別するのに必要な特徴表現を選択できることが明らかになったものの，人間の誤解を低減させるためには特徴の認知しやすさを考慮することが重要であるという示唆が得られた．

キーワード　画像情報ラベリング，語用論的アプローチ，大規模言語モデル

1 はじめに

画像情報に対して自然言語のテキストを生成する技術は，様々な産業的な応用への活用が期待される [11]．特に，複数の画像が与えられたときに，それらを誤解なく指し示すことのできるテキスト表現を生成することは，ロボットによる場所のナビゲーションや，人間に対する情報提供を正確に行う上で，重要なタスクである．本稿では，与えられた複数の画像のそれぞれに対して，それらを互いに区別可能な自然言語ラベルを生成する問題を，画像情報ラベリングと呼ぶ．

画像情報を説明する自然言語テキストを生成する手法は，画像キャプション生成の分野で研究されている [5], [7], [18]．特に近年では，与えられた画像情報を説明する自然言語テキストの生成や，画像情報に関する質問に回答することのできる，マルチモーダルな事前学習済み大規模言語モデルが広く利用可能になっている [20]．深層学習モデルは，モデルの規模（パラメータ数）と訓練データの規模を増加させることで品質が向上し続ける傾向があり [6]，画像キャプション生成においても有効性が示されている．このことから，可能な限り大きなスケールで事前学習した共有の大規模言語モデルを，再学習することなく活用する技術の検討が重要になってきている．

しかし，大規模言語モデルは，大量のデータをそのまま入力することはできないことから，多数のデータの関係を考慮した処理には適していない．このため，複数の画像に対して，それらを区別する自然言語ラベルを生成する，画像情報ラベリングの問題設定における性能は十分に検証されていない．似た画像と区別できるようなキャプションを生成する手法の研究は行われている [1], [4], [14], [17] が，これらの研究ではキャプション生成モデルの学習を伴うことから，大規模言語モデルの性能を活用することができないという課題がある．

本研究では，画像情報ラベリングタスクにおいて，大規模言語モデルを用いたテキスト生成手法の性能を評価すると共に，語用論的アプローチに基づいて性能の向上ができるかどうかを検討する．語用論的アプローチは，聞き手の解釈を考慮して，誤解を避けるような言語を生成するアプローチであり，複数のテキスト生成タスクにおける品質の向上が報告されている．本研究では，2 つの手法を通して，画像情報ラベリングにおける語用論的アプローチの有効性を検証する．1 つ目の手法では，大規模言語モデルに生成させた画像情報のラベルの中から，誤解の起きやすいラベルを検出し，再び大規模言語モデルに修正させる．2 つ目の手法では，大規模言語モデルに直接ラベルを生成させるのではなく，多数の特徴をテキストで挙げさせて，そのうちのいくつかを選択してラベルを構成する．この選択においては，誤解の起きやすさを評価する聞き手モデルを内部的に持つことによって，誤解可能性を最小化するような特徴の組み合わせを発見する最適化手法を提案する．

また，手法の性能を評価するためのデータセットを構築し，クラウドソーシングを用いた評価実験を行った．構築したデータセットは，複数のベッドルームのパノラマ画像（図 1 に例を示す）から構成されている．実験では，特定の部屋を指し示すことを意図して生成された自然言語ラベルを，いくつかの部屋の画像と共にクラウドワーカに提示し，自然言語ラベルがどの部屋を指し示しているかを質問した．実験の結果から，語用論的アプローチによって他の部屋と区別するのに必要な特徴表現を選択できることが明らかになったものの，クラウドワーカの



図 1: 部屋のパノラマ画像

正答率はベースラインを上回ることはなかった．結果の分析から，本タスクにおいて人間の誤解を低減させるためには，人間にとっての特徴の認知しやすさを考慮することが重要である示唆が得られたことを示す．

2 関連研究

本章では，本研究で扱う画像情報ラベリングタスクに関連する研究として，画像に基づくテキスト生成技術について 2.1 節で述べる．また，提案手法の着想に関連する研究として，語用論的アプローチに基づく機械学習手法について 2.2 節で述べる．

2.1 画像に基づくテキスト生成

画像内容に基づくテキスト生成技術の発展は，画像キャプション生成の研究によって牽引されてきた [7]．特に，深層学習を用いた画像キャプション生成では，ニューラルネットワークに画像の特徴を抽出し，得られた特徴に基づいて再帰型ニューラルネットワークでテキストを生成するモデルの枠組みが用いられる [18]．この枠組みにおいて Xu ら [19] は，注意機構を用いて画像内の注目すべき箇所の情報を抽出することで，生成されるキャプションの品質が向上することを示した．

一方近年，タスクによらない画像やテキストの表現学習によって，深層学習の性能が大幅に向上することが明らかになってきた [6]．この知見を発展させて，画像とテキストのマルチモーダルモデルを構成し，これを大規模な画像とテキストのデータセットを用いて特徴量表現を学習する手法が研究されている [11], [20]．事前学習済みの大規模マルチモーダルモデルを用いることで，画像キャプション生成の品質が向上する [5] だけでなく，画像のゼロショット分類や，画像質問応答 [16] などのタスクの精度も向上することも示されている．

さらに近年では，モデル規模と訓練データをスケールアップすることで，テキストでの指示（プロンプト）に応じた画像に基づくテキスト生成ができることが明らかになった．Li ら [13] は，事前訓練済みの画像特徴量抽出器とテキスト生成器の間を橋渡しする Querying Transformer を訓練する手法によって，画像に関する自然な会話や，柔軟な指示に応じた説明の生成ができることを示した．GPT-4 [2] は，入力した画像を描画するプログラムを出力したり，グラフの解釈もできることを示している．例えば，「この画像の特徴を 10 個挙げてください」というプロンプトと共に画像を入力することで，自然言語テキストで画像の特徴を得ることができる．

これらの深層学習に基づく手法はブラックボックスであり，どこまでのタスクを実現できるかは未知数であるが，基本的には 1 枚あるいは少数の画像とテキストの間を結びつける処理を想定している．このため，多数の画像に対して，それらを区別する自然言語ラベルを生成する問題設定における性能は十分に検証されていない．本研究では，画像情報ラベリングタスクにおいて，これらのテキスト生成手法の性能を評価すると共に，画像情報ラベリングタスクを分割して再構成することによるタスク性能の向上を検討する．

2.2 語用論的アプローチに基づく機械学習

コミュニケーションにおいて，話し手が聞き手の解釈を考慮し誤解を避けるために適切な語を選ぶことは重要である [10]．このアプローチは計算語用論として知られ，その中の Rational Speech Acts (RSA) フレームワークは，話し手と聞き手の互いの解釈や目的について再帰的に推論するプロセスを計算する代表的なモデルである [8], [12]．このフレームワークを基に，テキスト生成や画像キャプション生成の品質向上に向けた研究が行われている [9]．例えば，室内ナビゲーションタスクや画像の参照ゲームにおける語用論的アプローチが提案され，ナビゲーション成功率の向上や聞き手の誤解を低減するキャプション生成が可能とされている．本研究では，大規模事前学習済みモデルを再学習せずに利用する語用論的アプローチを通じて，誤解可能性を考慮した画像情報ラベリングを実現する手法を検討する．

3 提案手法

3.1 問題設定

n 枚の画像（それぞれが概念に対応する）からなる集合を $X = \{x_1, x_2, \ldots, x_n\}$ とする．画像情報ラベリングは，X が入力されたときに，X のそれぞれの画像に対応する自然言語ラベル $Y = \{y_1, y_2, \ldots, y_n\}$ を生成するタスクとする．ただし，画像 x_i に対して生成される自然言語ラベル y_i は，X の中で x_i にのみユニークに対応するようなラベルでなければならない．例えば，図 1 の画像において「絵が飾ってある部屋」といったラベルを付与した場合，他に X の中に絵が飾ってある部屋がある場合，これは当該画像にユニークに対応するラベルとはいえない．

このため，画像情報ラベリングにおいては，x_i に対する適切なラベルは，x_i には当てはまるが X に含まれる他の画像には当てはまらない特徴表現を抽出する必要がある．ラベル y_i の誤解可能性とは，X に含まれる x_i 以外の画像に当てはまる可能性の大きさである．本研究では，ラベルの誤解可能性を，人間がラベル y_i および X を見たときに，y_i が指し示している画像が x_i 以外の画像であると判断する割合によって評価する．

3.2 ベースライン（ChatGPT）手法

ChatGPT を用いたベースライン手法は，ChatGPT に画像 1 枚を与えて，その画像を表現するラベルを付与する手法である．この手法では，X に含まれる他の画像を考慮しない．画像と共に以下のプロンプトを与えて，画像に対するラベルを出力する．

Generate a one-sentence description that captures only three features of the room provided.

ベースライン手法によって生成されるラベルの例を以下に示す.

Bright bedroom with large windows, modern decor, scenic view.

3.3 Pragmatic ChatGPT 手法

本研究では，語用論的アプローチに基づいて，ラベルを改善する手法を提案する．語用論的アプローチでは，話し手モデルが生成したラベルに対して，聞き手モデルがそのラベルの誤解可能性を評価して，その評価を話し手モデルにフィードバックすることでラベルを改善する．

Pragmatic ChatGPT (PCG) 手法は，話し手モデルと聞き手モデルの両方に ChatGPT を導入した語用論的アプローチに基づく手法である．PCG におけるラベル改善処理のプロセスは，以下の通りである．

（1） **初期ラベル生成**：全ての画像 X を ChatGPT に与える．このとき，それぞれの画像の部屋の特徴を捉えた文を生成する指示をプロンプトとして与える．結果として得られる，それぞれの画像に対応する文を初期ラベル Y とする．

（2） **聞き手モデル（ヒントの生成）**：全ての画像 X と，現時点でのラベル Y を ChatGPT に与える．このとき，誤解する可能性の高い 2 つのラベルの組み合わせを「ヒント」として生成する旨のプロンプトを与える．この結果として，ヒントのテキストデータ h を得る．出力されるヒントの例を以下に示す．

hint 1. Caption 3 and Caption 10: Both mention contemporary elements with large windows and a spacious layout, which could refer to modern architecture with a minimalist theme.

（3） **ラベル改善**：全ての画像 X，現時点でのラベル Y，聞き手モデルから得たラベル改善のヒントのテキスト h を ChatGPT に与える．このとき，ラベル改善のヒントをもとにして，既に生成されているラベルを改善したラベルを，それぞれの画像について出力する旨の指示をプロンプトとして与える．結果として得られる，それぞれの画像に対応する文で，ラベル Y を更新する．

（4） 更新されたラベル Y を用いて，再び聞き手モデルの処理を行い，得られたヒントを用いてラベル改善の処理を繰り返す．この処理を，あらかじめ定めた回数だけ繰り返す．

最終的に得られたラベル Y を出力とする．繰り返しの改善処理によって，誤解する可能性の高いラベルが特定され，より誤解可能性の小さいラベルが得られることが期待される．ChatGPT は，自らの出力を改めて入力することで出力の質が向上する現象があることが知られている．本研究では，このような質向上の効果が画像情報ラベリングにおいて確認できるかどうかを検証する．

表 1: 特徴表現の集合の例

$z_{i,1} =$	Tray ceiling Recessed lighting
$z_{i,2} =$	Beige wall paint
$z_{i,3} =$	White ceiling
$\vdots$	$\vdots$
$z_{i,49} =$	White orchids
$z_{i,50} =$	Distressed wood mirror frame

表 2: ある画像に対する特徴表現の類似性スコアの例

特徴表現	類似性スコア
Tray ceiling Recessed lighting	28.051437
Two-tone wood flooring	25.21447
White ceiling	25.036436
…	
Mountainous landscape view	17.63242
Glass vase with white flowers	17.470116
Striped throw on the bed	16.694729

3.4 特徴表現に基づく組み合わせ最適化手法

特徴表現に基づく組合せ最適化手法（Feature-based Combinatorial Optimization; FCO）では，ChatGPT に直接ラベルを出力させるのではなく，それぞれの部屋の特徴を表す特徴表現を列挙させ，その特徴表現の最適な組み合わせを求める．話し手モデルは，それぞれの部屋について候補となる特徴表現の中から K 個を選択してラベルを構成する（ただし，K はあらかじめ定められた定数）．聞き手モデルは，画像とテキストの類似度を計算する CLIP モデル [15] に基づいて，誤解可能性の高いラベルのペアを推定する．

FCO のプロセスは以下の通りである．

（1） **特徴表現の列挙**：画像 x_i を 1 枚ずつ ChatGPT に与える．このとき，画像の部屋の特徴を 50 個列挙する指示をプロンプトとして与える．結果として得られる，画像 x_i に対応する 50 個の特徴表現テキストからなる集合を，$Z_i = \{z_{i,1}, \ldots, z_{i,50}\}$ とする．この処理を全ての画像に対して行い，$Z = (Z_1, \ldots, Z_n)$ を得る．例えば，図 1 の部屋を与えた場合，特徴表現として $z_{i,1}$ = “Tray ceiling Recessed lighting”，$z_{i,2}$ = “Beige wall paint” などが得られる（表 1）．

（2） **初期ラベルの構成**：それぞれの画像 x_i に対して，Z_i からランダムに K 個の特徴表現を選択し，それらをカンマ区切りで並べたテキストを，対応するラベル y_i とする．ラベル y_i を構成する特徴表現を $z(y_i) = (z(y_i)_1, \ldots, z(y_i)_K)$ と表記する．この処理を全ての画像に対して行い，初期ラベル Y を得る．

（3） **聞き手モデル（ラベルのゼロショット分類）**：以下の 2 つの処理と出力を行う．

- **誤解可能性スコアの計算**：画像 x_i とラベル集合 Y を CLIP モデルに与えて，x_i がラベル y_j に分類されるゼロショット分類確率 $p(y_j \mid x_i)$ を求める．ラベル y_j の適合率は $\frac{p(y_j|x_j)}{\sum_i (p(y_i|x_j))}$，

再現率は $p(y_j \mid x_j)$，F1 スコアは適合率と再現率の調和平均として計算できる．全てのラベル $y_1, \ldots, y_n$ の F1 スコアの平均を，現時点のラベル Y の誤解可能性スコアとして記録する．

- **誤解可能性の高いラベルの検出**：ラベル y_j に対して，最も誤解しやすいラベルのインデックスを $distractor_Y(j) = \arg\max_{j' \neq j} p(y_{j'} \mid x_j)$ とする．この誤解しやすさが最大であるラベル $j^* = \arg\max_j p(distractor_Y(j) \mid x_j)$ を求めて，最も誤解しやすいと考えられるラベルペア $(j^*, distractor_Y(j^*))$ をラベル改善のヒントとして出力する．

（4） **改善ラベルの構成**：ラベル聞き手モデルから得られたヒント $(j^*, distractor_Y(j^*))$ を用いて，以下の 4 つの処理を行い，ラベル y_{j^*} を改善対象として更新する．ここでは，改善対象ラベル y_{j^*} を構成する K 個の特徴表現 $z(y_{j^*})_1, \ldots, z(y_{j^*})_K$ のうちいくつかを，候補特徴表現 $Z_{y_{j^*}}$ の別の要素と入れ替えることで，誤解可能性の改善を目指す．前処理として，改善対象に対応する画像 x_{j^*} と，候補特徴表現 $z_{j^*,1}, \ldots, z_{j^*,50}$ それぞれを CLIP モデルに与えて，類似性スコア $sim(x_{j^*}, z_{j^*,k})$ を求める（計算される類似性スコアの例を表 2 に示す）．また，改善対象と誤解されやすい画像 $x_{distractor_Y(j^*)}$ と，候補特徴表現 $z_{j^*,1}, \ldots, z_{j^*,50}$ それぞれを CLIP モデルに与えて，類似性スコア $sim(x_{distractor_Y(j^*)}, z_{j^*,k})$ も求める．

- **対応画像との類似性が低い特徴表現の削除**：y_{j^*} を構成する特徴表現 $z(y_{j^*})$ の中から，対応する画像 x_{j^*} との類似性スコアの逆数にソフトマックス関数を適用して得られた確率分布に従ってランダムに 1 つの特徴表現を選択し，y_{j^*} を構成する特徴表現から削除する．
- **対応画像との類似性が高い特徴表現の追加**：Z_{j^*} の中から，対応する画像 x_{j^*} との類似性スコアにソフトマックス関数を適用して得られた確率分布に従ってランダムに 1 つの特徴表現を選択し，y_{j^*} を構成する特徴表現に加える．
- **誤解しやすい画像との類似度が高い特徴表現の削除**：y_{j^*} を構成する特徴表現 $z(y_{j^*})$ の中から，誤解しやすい画像 $x_{distractor_Y(j^*)}$ との類似性スコアにソフトマックス関数を適用して得られた確率分布に従ってランダムに 1 つの特徴表現を選択し，y_{j^*} を構成する特徴表現に加える．
- **誤解しやすい画像との類似度が低い特徴表現の追加**：Z_{j^*} の中から，誤解しやすい画像 $x_{distractor_Y(j^*)}$ との類似性スコアの逆数にソフトマックス関数を適用して得られた確率分布に従ってランダムに 1 つの特徴表現を選択し，y_{j^*} を構成する特徴表現に加える．

（5） 更新されたラベル Y を用いて，再び聞き手モデルの処理を行い，得られたヒントを用いてラベル改善の処理を繰り返す．この処理を，あらかじめ定めた回数だけ繰り返す．

聞き手モデルが記録する誤解可能性スコアが最も大きくなった繰り返し試行におけるラベルを，最終的な出力とする．例えば，繰り返し試行における誤解可能性スコアが図 2 の場合，赤丸の時のラベルを最終的な出力とする．

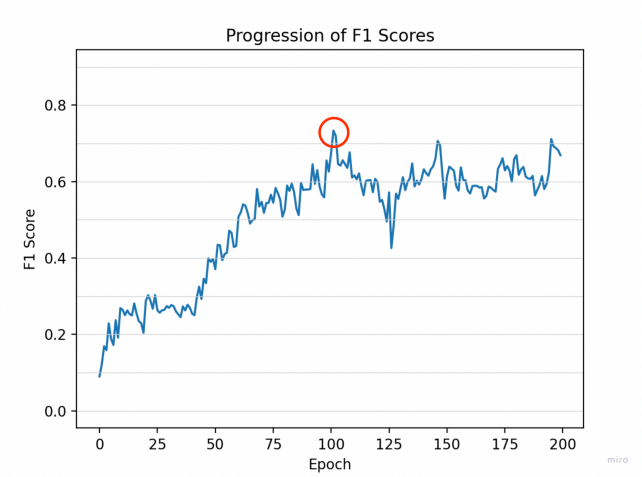


図 2: 200 回試行内の誤解可能性スコアの推移

4 評価実験

本実験では，3 節で述べた 3 つの手法（ChatGPT 手法，Pragmatic ChatGPT 手法，FCO 手法）によって生成されるラベルを，人の認識における誤解可能性によって評価する．また，ラベルの長さの制約や，区別する必要のある概念の数が，誤解可能性に与える影響について確認する．

4.1 評価手法

クラウドソーシングプラットフォームの Amazon Mechanical Turk（AMT）を用いる．AMT タスクの例を図 3 に示す．ワーカには，1 つの自然言語ラベルと，4 枚の部屋の画像（図 3 には 1 枚しか写っていないが，スクロールすることで 4 枚の画像が表示される）が提示される．ワーカは，ラベルが指し示していると考えられる部屋の画像を 1 枚選択する．選択された部屋の画像が，提示したラベルの生成時に意図した部屋の画像である場合，誤解なく指し示すことができたと評価できる．

特定の条件において生成されたラベル集合 Y を評価するごとに，800 件のタスクを実施した．n 個の概念ラベルを評価するため，それぞれの概念に対して $\frac{800}{n}$ 個のタスクを割り当てた．概念ラベル y_i を評価するためのタスクでは，提示する自然言語ラベルは y_i，提示される部屋の画像のうち 1 枚は対応する画像情報 x_i，残りの 3 枚は $\frac{800}{n}$ 個のタスクでそれぞれランダムに X の中から選択された x_i 以外の画像情報である．y_i を評価するためのタスクにおいて，ワーカが x_i の画像を選択すれば正答，x_i 以外の画像を選択すれば誤答とする．800 件のタスクのうち，正答だったタスクの割合を正答率とする．

4.1.1 データセット

Matterport3D データセット [3] の画像を利用して，評価用のデータセットを構築した．Matterport3D データセットは，90 の建物の内部の画像 194,400 枚からなるデータセットである．それぞれの建物内部に複数の撮影地点が設定され，各撮影地点につき 6 方向，3 角度（上方，水平，下方）で撮影した 18 枚の画像が収録されている．各撮影地点には，部屋のタイプ（寝室，

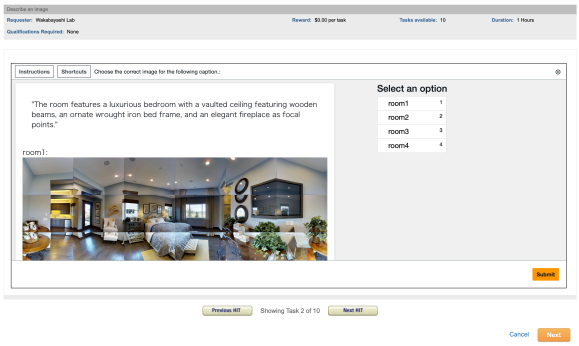


図 3: 被験者に表示される AMT タスクの例



図 4: 同一地点で撮影した画像 18 枚を結合したパノラマ画像

バスルーム，リビング，廊下など）がタグづけされている．

本研究では，寝室とタグづけされている撮影地点から，ランダムに 30 の撮影地点をそれぞれ異なる部屋になるように抽出した．それぞれの撮影地点に対応する 18 枚の画像を適切な位置で結合して，パノラマ画像を生成した．生成したパノラマ画像の例を図 4 に示す．ラベリング対象とする画像をパノラマ画像とするのは，画像に含まれる特徴が，18 枚のうちのどれを選んだかに依存するのを避けるためである．それぞれの部屋（撮影地点）を概念，それに紐づいたパノラマ画像を画像情報として，概念ラベリング（画像情報ラベリング）の評価用データセットとして用いる．

4.2 ラベル長比較実験

本節では，3 つの手法の性能を比較するとともに，ラベルの長さの制約が誤解可能性に与える影響を評価した結果を示す．ラベルが短いほど，含めることのできる情報量が制限されるため，正答率は減少する傾向にあると予想できる．実際の応用においては，簡潔でありながら誤解のないラベルが有用と考えられるため，短いラベルにおける性能評価は重要である．本実験では，データセットの中から 10 枚の画像情報をランダムに選んで用いた．

ChatGPT 手法と Pragmatic ChatGPT 手法では，以下のプロンプトをラベル出力処理時に追加し，ラベル生成に 10 単語以内の制約を設けた場合を短出力条件，以下のプロンプトを含めない場合を長出力条件とする．

> Sentences must be output within 10 words.

また，FCO では，使用する特徴の数を $K = 3$ に設定した場合を短出力条件，$K = 7$ に設定した場合を長出力条件とする．

手法	正答率 (%)	平均単語数
ChatGPT 長出力条件	37.125	26.9
Pragmatic ChatGPT 長出力条件	31.125	20.7
FCO 長出力条件	36.75	20.3
ChatGPT 短出力条件	34.625	7.5
Pragmatic ChatGPT 短出力条件	29.625	8.1
FCO 短出力条件	30.75	9.3

表 3: ラベル長比較実験における条件ごとの AMT タスクの正答率と平均単語数

手法	正答率 (%)
ChatGPT 短出力条件 $n = 10$	34.625
ChatGPT 短出力条件 $n = 20$	32.375
ChatGPT 短出力条件 $n = 30$	35.125
FCO 短出力条件 $n = 10$	30.75
FCO 短出力条件 $n = 20$	32.375
FCO 短出力条件 $n = 30$	33.125

表 4: 区別枚数比較実験における概念の数を変化させたときの正答率

実験では，3 手法それぞれで短出力と長出力のラベルを生成し，合計 6 種類の条件で生成されたラベルの評価を AMT を用いて行う．Pragmatic ChatGPT 手法のラベル改善サイクルは 10 回とした．

表 3 に，条件ごとの AMT タスクの正答率と，タスクに使用したラベルの平均単語数を示す．長出力条件と短出力条件のどちらにおいても，語用論的アプローチを用いた Pragmatic ChatGPT 手法や FCO 手法よりも，ベースラインである ChatGPT 手法の方が正答率が高かった．Pragmatic ChatGPT 手法は最も正答率が低く，ChatGPT で生成したヒントを繰り返し反映することで，正答率はむしろ減少することが示唆される．

4.3 区別枚数比較実験

本節では，概念の数 n が正答率に及ぼす影響を評価するため，区別する概念の数を 10，20，30 に変化させて実験を行う．区別する必要のある概念の数が多いほど，誤解可能性の排除が難しくなっていくため，正答率は減少する傾向にあると予想できる．本実験では，短出力条件の ChatGPT 手法と FCO 手法を用いた．

表 4 に，条件ごとの AMT タスクの正答率を示す．いずれの手法においても，概念の数が増えるにつれての性能が低下する傾向は見られなかった．FCO 手法においても正答率は低下することなく向上した．どの枚数の場合でも，ChatGPT 手法は FCO 手法と比べて高いか同等の正答率を示した．

4.4 考　　察

実験結果は，語用論的アプローチを用いたラベル改善手法が，ベースラインである ChatGPT 手法と比べて正答率をむしろ低下させることを示唆している．この原因を考察する上で，手法が出力するラベルに含まれる特徴表現の抽象性が重要な観点であると考えられる．



図 5: 実験に用いた画像の例．ChatGPT 手法の出力ラベルは “elegant interior design, panoramic window view, cozy ambiance”, FCO 手法の出力ラベルは “beige upholstered bench, coordinated picture frames, sliding closet doors”.

図 5 に，実験に用いた画像の例を示す．この画像に対して，ChatGPT 手法は “elegant interior design, panoramic window view, cozy ambiance”，FCO 手法は “beige upholstered bench, coordinated picture frames, sliding closet doors” というラベルを出力した．FCO 手法は，画像の特徴において，具体的な要素（例：「coordinated picture frames」や「beige upholstered bench」）に焦点を当てたラベルを生成している．これらのラベルは，空間の個別のアイテムや構造的なディテールを表現しており，識別性の高い情報を含んでいる．しかし，このような具体性すぎる特徴は，人間ワーカが画像を選択する際に見逃される可能性が高く，このことが原因で正答率が悪化する場合があると考えられる．

一方で，ChatGPT 手法は,「elegant interior design」や「traditional elegance」といった，空間全体の雰囲気やデザインコンセプトを表す抽象的な特徴表現を含んだラベルを生成している．これらの抽象的な記述は，人間ワーカが画像を選択する際に見逃されることがより少ないと考えられる．Pragmatic ChatGPT 手法や FCO 手法は，ラベル改善を行う際に識別性の高い特徴表現に着目することを強いるため，このような抽象的な特徴表現は排除され，具体的な特徴表現になる傾向にある．具体的な特徴表現は見逃されて正答につながらない場合があることを踏まえると，ChatGPT 手法の抽象的な特徴表現を多く含む性質が，AMT タスクの正答率を高める要因であったと考えられる．このことから，画像情報ラベリングにおいて人間の誤解を低減させるためには，人間にとっての特徴の認知しやすさを考慮することが重要である示唆が得られたといえる．

また，区別枚数比較実験の結果からは，概念の数が増加しても，正答率が顕著に低下する傾向が見られなかった．これは，両手法が画像の枚数の増加に対して頑健であることを示している可能性があるが，評価方法による限界も影響していると考えられる．本実験では，AMT タスクの都合で，n 枚の画像の中から 4 枚をランダムに選択して提示している．n 枚の画像の中には，正答の部屋と似ている（誤解しやすい）部屋が含まれている場合が多いが，ランダムに選択した画像の中に，似ている部屋が選ばれない場合もある．この場合，似ている部屋との区別がつかないラベルが生成されていて，実際の応用では問題になるようなラベルだとしても，AMT タスク上ではワーカは正答できてしまう．このような場合は，区別する必要のある概念の数が増えた条件においてより頻繁に起きる可能性があることから，概念ラベリングの評価手法の改善を検討する必要があると考えられる．

5 おわりに

本研究では，概念を誤解なく指し示すことのできるテキスト表現を生成する概念ラベリングの問題，とりわけ画像情報で表現された概念を対象とした画像情報ラベリングの問題を解決する手法の検討を行った．特に，マルチモーダルな事前学習済み大規模言語モデルを再学習することなく活用する手法を検討し，さらに語用論的アプローチに基づいて性能の向上ができるかどうかを検討した．画像情報ラベリングの手法の性能を評価するためのデータセットを構築し，クラウドソーシングを用いた評価実験を行った．実験結果から，語用論的アプローチに基づく性能の向上は確認できなかったが，人間の誤解を低減させるためには，特徴表現の識別性だけでなく，人間にとっての特徴の認知しやすさを考慮することが重要であるという知見が得られた．また，特に区別する必要のある概念の数が増えた条件において，概念ラベリングの評価手法を改善するための示唆が得られた．

今後の課題としては，人間の情報処理のプロセスや傾向を考慮して，これに適合する概念ラベリング手法を開発することが挙げられる．特徴を認知する可能性を聞き手モデルに導入し，特徴表現を見逃すという要因を考慮することで，誤解可能性を適切に推定する手法が検討できると考えられる．

謝　辞

本研究の一部は，JST CREST (#JPMJCR22M2) および JSPS 科研費 21H03552 の助成によって行われた．

文　献

[1] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, 2016.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv [cs.CL] 2303.12712*, 2023.

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[4] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with Character-Level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 439–443, 2018.

[5] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in Vision-Language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*,

pages 2136–2148, 2023.

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

[7] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV*, pages 15–29, 2010.

[8] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998, 2012.

[9] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *North American Chapter of the Association for Computational Linguistics*, pages 1951–1963, 2017.

[10] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, 2023.

[11] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-Language pre-training: Basics, recent advances, and future trends. *arXiv [cs.CV] 2210.09263*, 2022.

[12] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.*, 20(11):818–829, 2016.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image pre-training with frozen image encoders and large language models. 2023.

[14] Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic Issue-Sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, 2020.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, 139:8748–8763, 2021.

[16] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are Few-Shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, 2022.

[17] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. 2017.

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):652–663, 2017.

[19] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of Machine Learning Research*, 37:2048–2057, 2015.

[20] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv [cs.CV] 2306.13549*, 2023.

自然保護管理のための捕獲を目的とした鹿の行動パターンに基づく警戒度推定

酒井　歩夢†　　宮森　　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2053471,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，自然保護管理のための捕獲を目的とした鹿の行動パターンに基づく警戒度推定手法を提案する．近年，鹿や猪による山林や農産物への損害が増加しており，これに対処するために箱罠やくくり罠を用いた捕獲が一般的に行われている．しかし，箱罠においては動物の警戒心が低くなるのを待ちストッパーを外し扉を閉める必要があり，特に初心者にとってはそのタイミングを見極めるのが難しいという課題がある．先行研究では，カメラ映像から害獣の頭数を自動計測し，ストッパーを外すタイミングを算出していた．しかし，警戒度が高い個体においては扉が閉まる前に脱出されてしまう可能性が高いという問題があった．本研究では，箱罠のストッパーを外す適切なタイミングを算出するため，害獣の頭数だけでなく，警戒度を推定することを提案する．これにより，警戒度が高い個体が脱出する可能性を軽減する判断が可能となり，捕獲率が向上することが期待できる．具体的には，害獣として鹿を対象とし，山林で撮影した動画に対して警戒度をラベル付けしたデータセットを構築する．これ基づき訓練した警戒度推定器を用いて，箱罠のストッパーを外すタイミングを算出する．プロトタイプを用いた実験により，タイミングの算出がより適切に行えるか検証する．

キーワード　害獣捕獲，箱罠，行動認識，物体検出，推論

1　はじめに

1.1　背　　景

近年，シカやイノシシなどの害獣による山林や農産物への損害が増加しており，農林水産省によると，令和 4 年度については全国の農作物被害金額が約 156 億円と報告されており問題となっている. これに対処するために箱罠やくくり罠などの罠を用いた捕獲がされている.

箱罠は箱状の構造物内に餌などの動物が引き寄せられる仕掛けがある. 箱の中には餌が配置され，動物がその餌を求めて箱内に進入する. 仕掛けには扉や門を保持するストッパーがあり，餌を食べたりすることでストッパーが外れ扉が閉まる仕組みである.

1.2　現状の問題点

箱罠においては動物の警戒心が低くなるのを待つ必要があり，警戒心が低なるまでは扉が閉じないように固定し箱罠に慣れさせる. しかし，初心者にとっては警戒心の見分けが難しかったり，ストッパーの調整が上手くいかない場合が多くノウハウを得るのに時間がかかる点が問題点として挙げられる.

既製品として，カメラの映像をキャリア通信を使用しリアルタイムで監視を行い扉を遠隔で閉めるシステムが販売されているが，鹿や猪などの害獣は夜間も活動しておりリアルタイムでの監視が難しいうえに，獣が箱罠に入るまでに時間がかかる場合があり時間がとられる場合がある. また，キャリア通信の電波が届かない場所などでは使用することができないなどの問題点がある.

1.3　目的・取り組む問題

1.3.1　目　　的

害獣の箱罠をカメラによって自動制御を行う. 自動制御に必要な情報として警戒度の見分けが必要であり，本実験では鹿の行動パターンから警戒度を推測するモデルを作成する.

1.3.2　問　　題

鹿の映像から警戒度合いの推定を行うモデルを作成する.

1.4　提案手法

山林において撮影された鹿の動画から鹿の位置を抽出し，バウンディングボックスを用いて鹿の行動を時系列データとして取得した. このデータを入力として， 組み込まれた CNN と LSTM から構成されるネットワークを用いて，鹿の行動パターンの学習を行った. 最終的な出力層は，多クラス分類として構築し，鹿の警戒行動パターンを識別する.

従来実験では，箱罠内の鹿の頭数を測定する従来の手法が提案されたが，この手法は箱内での鹿の位置や行動に関する情報が不足していると考える. 鹿の警戒度が高い場合には，箱の扉が閉まる前に脱出される可能性が高い. 捕獲率の向上において警戒度が鍵となると考えた.

高い警戒度を示す個体は，餌に夢中になることなく，箱の奥に進まず，入口でのみ餌を摂取する姿勢や，首を伸ばして時折少し後退りをしながら餌を食べる行動を示す. これらの行動パターンは，個体が常に逃走を視野に入れながら行動していると考えられる. 対照的に， 捕獲が成功した場合においては，個体

は餌に夢中になり，後退りのような行動は観察されなかった．

長期間の餌付けが警戒度を低下させる可能性がある一方で，環境への悪影響や本来そこにいなかった個体の誘引などが考えられるため，捕獲を適切なタイミングで行う必要があると考えられる．

コンピュータを用いて鹿の警戒度を推定することで，適切なタイミングでの捕獲が可能となり，捕獲率の向上に寄与すると考えられる．

1.5 実験方法

実験では，動物の行動パターンから警戒度の抽出と評価を行う．

京都市内の山林にカメラを設置してデータを収集した．撮影された映像を，YOLOv8 を使用し物体検出を行いバウンディングボックスのみの画像を OpenCV を使用しバウンディングボックス内のみを抽出した．動画の映像から警戒行動は，先行研究 [5] より定義しラベル付を行った．フレームごとの画像を時系列データとして扱い，ネットワークに入力した．

学習は Python3.8 と Tensorflow2.12 を使用した．学習を行うネットワークは，時系列の鹿の画像を CNN 層に通し畳み込みを行った．その後 LSTM 層に通し時系列の変化の学習を行った．出力層は警戒しているか，していないかの 2 クラス分類とした．

1.6 実験結果

実験を行った結果に loss: 0.5907，accuracy:0.7248，val_loss: 0.5923，val_accuracy:0.7227 となり，図 6 のように推移した．この結果より提案したモデルでは正しく識別することはできなった．

1.7 論文の貢献

動物の行動パターンから警戒度の抽出と評価を行う．

2 関連研究

2.1 檻周辺における害獣の頭数測定 [1]

2.1.1 従来手法

・背景差分による動体領域抽出を行い鹿を検出する．
・カルマンフィルタを用いて鹿の数を数え規定数以上が箱罠に入っていれば扉を閉じる．

2.1.2 提案手法との違い

警戒度などは考えていない為，箱罠に対してどの程度警戒しているかどうかの見極めが難しい．警戒している場合においては，扉が閉まる前に逃げられてしまうなどの可能性がある．

2.2 ICT を活用した大型捕獲檻によるイノシシの捕獲実証 [2]

2.2.1 従来手法

大型捕獲檻と遠隔監視型捕獲システムを活用した IoT 罠．
大型の檻を利用することで群ごと罠に入ったことを確認して遠隔で折の扉を閉じる．

2.2.2 提案手法との違い

大型の檻を使用するため通常の箱罠に比べ警戒していても扉が閉まるまでに逃げられる可能性は低減されると考えられる．しかし，大型の罠は可搬性が悪く，費用が高いことが問題点として挙げられる．
また，遠隔であるために夜間の監視や，長時間罠に入らない場合においては負担になることが考えられる．

2.3 Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model [3]

2.3.1 従来手法

CNN-LSTM を用いて文章から感情の分類を行なっている．

2.3.2 提案手法との違い

CNN-LSTM を文章データではなく画像データから警戒度の推測を行う．

2.4 Human action recognition using attention based LSTM network with dilated CNN features [4]

2.4.1 従来手法

拡張畳み込みと残差ブロックを使用し情報の保持を行う．
長期的な依存関係を学習するために LSTM を使用する．

2.4.2 提案手法との違い

効率的な人間の動作認識のための軽量な CNN アーキテクチャを提案している．
拡張畳み込みやスキップ接続を備えた残差ブロックが含まれ，BiLSTM を用いた注意メカニズムも使用している．
ソフトマックスを備えた中心損失関数を使用して性能向上している．

3 提案手法

3.1 データ収集と前処理

京都市内の山林に設置されたカメラでデータを収集し，ヘイキューブと呼ばれる牧草を使用して鹿を誘引をおこなった．

警戒行動の定義は，頭部を肩と並行あるいは肩より高く保持して立ち，視線の方向に耳を向けて静止する．時折耳を動かし，周囲を窺う行動を伴い，餌を頬張ったまま口を動かしたり，反芻したりする行動を伴わないとした．

YOLOv8 を使用して物体検出を行い，OpenCV を用いてバウンディングボックス内のみの画像を抽出．フレームごとの画像を時系列データとして扱い，ネットワークに入力した．

画像サイズは全て縦横 180px にリサイズをした．

3.2 ネットワーク構築

学習には Python3.8 と Tensorflow2.12 を使用した．ネットワークは，時系列の画像を CNN 層に通し畳み込みを行い，その後 LSTM 層に通して時系列の変化を学習している．
出力層は 2 クラス分類で，警戒しているかどうかを判定している．学習データは，警戒していないラベルは 1374 本の動画 (123,660 枚)，警戒しているラベルは 488 本の動画 (43,920 枚)

の画像から 70%を学習データとして学習を行い，残りの 20%を検証データ，10%をテストデータとし，バッチサイズは 16 とした．オプティマイザは Adam を使用し，学習率は 0.001，エポック数は 50 とした．

3.3 従来手法の課題

従来手法の研究では，箱罠内の鹿の頭数を測定する手法が提案されたが，箱内での鹿の位置や行動に関する情報が不足していると考えた．特に，鹿の警戒度が高い場合には箱の扉が閉まる前に脱出される可能性が考えられる．

3.4 警戒度と捕獲率の関連

高い警戒度を示す個体は，通常は箱の奥に進まず，入口でのみ餌を摂取する姿勢や，逃走を視野に入れつつ餌を食べる行動を示す．捕獲が成功した場合には，個体は餌に夢中になり，後退りのような行動は観察されない．

3.5 捕獲のタイミング

長期間の餌付けが警戒度を低下させる可能性があるが，環境への悪影響や本来そこにいなかった個体の誘引などが懸念される．[6] コンピュータを用いて鹿の警戒度を推定し，適切なタイミングでの捕獲が可能となり，捕獲率の向上に寄与すると考えられる．

4 実　　験

4.1 研究課題

害獣の箱罠をカメラによって自動制御するために，鹿の映像から警戒度合いの推定を行うモデルを作成する．

4.2 実験設定

4.2.1 実験方法

実験では，動物の行動パターンから警戒度の抽出と評価を行う．京都市内の山林にカメラを設置してデータを収集した．カメラの前にはヘイキューブと呼ばれる牧草を撒き鹿を誘引した．

撮影された映像を，YOLOv8 を使用し物体検出を行いバウンディングボックスのみの画像を OpenCV を使用しバウンディングボックス内のみを抽出した．



図 1　YOLOv8 によって抽出された画像

動画の映像から警戒行動について， 先行研究 [5] より頭部を肩と並行あるいは肩より高く保持して立ち，視線の方向に耳を向けて静止する．時折耳を動かし, 周囲を窺う行動を伴う．餌を頬張ったまま口を動かしたり，反芻 (はんすう) したりする行動を伴わないと定義を行いラベル付を行った．



図 2　頭部を肩より高く保持して立ち視線の方向に耳を向けて静止する行動

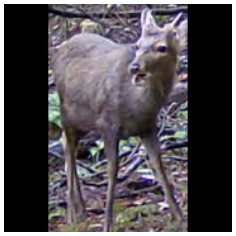


図 3　餌を頬張ったまま口を動かす行動

フレームごとの画像を時系列データとして扱い，ネットワークに入力した．1 つの時系列データは 90 フレーム（3 秒間）とした．画像のサイズはすべて縦横 180px にリサイズした．リサイズはアスペクト比を固定せずに行い，空白部分は全てのピクセルを 0 で埋めた．

学習は Python3.8 と Tensorflow2.12 を使用した．学習を行うネットワークは，時系列の鹿の画像を CNN 層に通し畳み込みを行った．その後 LSTM 層に通し時系列の変化の学習を行った．出力層は警戒しているか，していないかの 2 クラス分類とした．

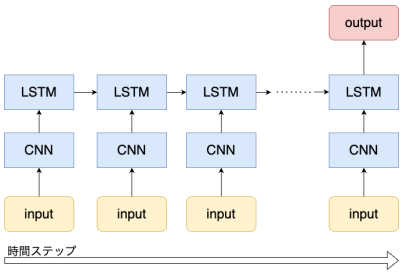


図 4　ネットワーク構成

学習データは，警戒していないラベルと警戒しているラベルの画像から 70%を学習データとして学習を行い，残りの 20%を検証データ，10%をテストデータとし，バッチサイズは 16 とした．オプティマイザは Adam を使用し，学習率は 0.001，エポック数は 50 とした．

評価指標: ネットワークを構成を作成し，性能は検証データを用いて評価をおこなう．

評価項目: 警戒しているかどうかの二値分類における精度や損失関数の値で評価を行う．また，先行研究のモデルを参考に動物においても行動から特徴を抽出することができないか評価を行う．

4.2.2 ネットワーク

入力層: 鹿の時系列画像データを入力として使用.
畳み込み層 (CNN): 入力された時系列画像から特徴を抽出し，位置や形状の変化を学習を行う.
LSTM: CNN からの特徴マップを時系列データとして LSTM 層に入力する. 時系列データの長期的な依存関係を学習し，動的なパターンや変化を捉える. また，行動に対するパターンを理解し，その情報を保持をする.
出力層: 2 クラス分類のため，出力ユニットは 2 つとする.

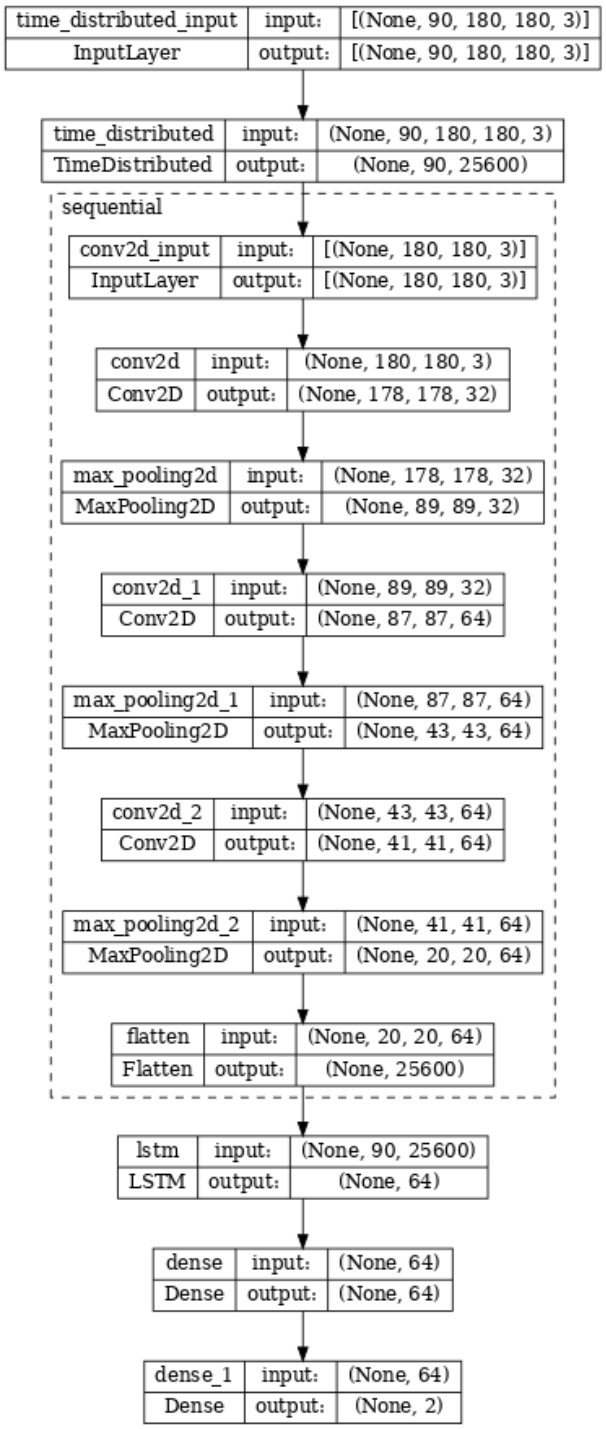


図 5　ネットワーク構成

学習: Adam を使用する. クロスエントロピー損失関数を使用して，出力と正解ラベルの差異を最小化するように学習をおこなった. 学習データは警戒していないラベルと警戒しているラベルからなり，70%が学習，20%が検証，10%をテストに使用した.

実際のネットワーク構成は図 5 のようになった.
CNN は，画像データから特徴を抽出し，画像内の異なる部分のパターンや特徴を検出する. 畳み込み層とプーリング層が交互に配置され，画像の特徴が抽出される. その後，これらの特徴がフラット化され，時系列データとして処理される LSTM に渡される. 活性化関数は relu を用いた.
LSTM は，CNN からの特徴データを受ける. データ内の時間的な関係を考慮しながら情報を処理をおこなう. フレームごとの特徴データの系列を処理し，これらのデータの長期的な依存関係を学習する.
最後に，全結合層が LSTM からの出力を受け取り，最終的な分類を行う. 活性化関数は sigmoid を用いた.

5　実験結果

実験の結果，ネットワークの学習を行ったが訓練データおよび検証データにおいて損失と正確度にあまり変化が見られなかった. 訓練データの正解率は 0.7248，検証データも 0.7227 で一定であり，学習の進展がなかった. 損失関数についても， 同様にほぼ一定で学習の進展がなかった.

この結果から，モデルは訓練データと検証データの両方に対して同様の性能を示しており，正しくネットワークが学習できていないと考えた.

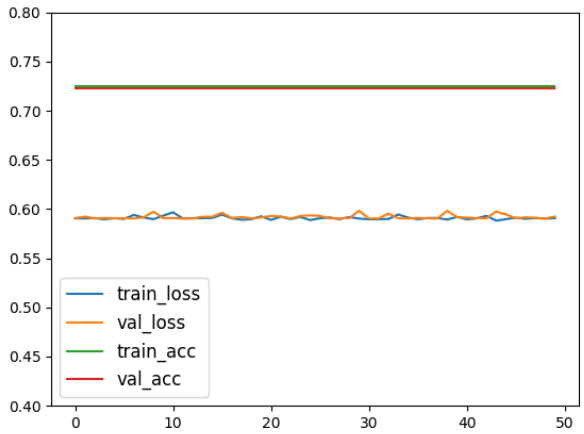


図 6　学習結果 (正解率, 損失関数) のグラフ

図 6 を確認すると訓練データでは学習時の loss が 0.59 あたりを横ばいに推移しており，accuracy は 0.7248 のまま変化しなかった. 検証データでも同様に loss は 0.59 あたりを横ばいに推移しており，accuracy は 0.7227 のまま変化しなかった. 損失関数を Adam から変更した場合においても，結果は同様となり大きな変化は見られなかった.

accuracy が学習を進めても変わらないことから，提案したモデルでは分類することができなかった．

6 ま と め

6.1 取り組んだ問題，提案手法

この研究では，鹿などの害獣の箱罠を効率的にに制御を行うために，カメラによる自動制御を行おうと考えた．問題点として，鹿の警戒度合いを推定し，それに基づいて適切なタイミングで箱罠を制御することであった．

提案手法は，山林で撮影された鹿の動画から鹿の位置を抽出し，その時系列データを用いて鹿の行動パターンを学習するネットワークを構築した．このモデルでは，鹿の警戒行動パターンを 2 クラス分類として識別することを目的としている．

6.2 実験目的，実験方法，データセット，実験結果

実験の目的は，提案手法を用いて鹿の映像から警戒度合いを推定することであった．実験では，山林に設置されたカメラで撮影された動画データを使用し，ヘイキューブと呼ばれる牧草を用いて鹿の誘引をおこなった．警戒行動として，頭部の位置や視線方向，耳の動きなどで判断を行いラベルづけを行った．

モデルは，YOLOv8 による物体検出と，CNN-LSTM ネットワークによる時系列データの学習を組み合わせて構築を行った．学習データは，警戒していないラベルと警戒しているラベルの動画から取得された画像を使った．

実験結果では，50 エポックの学習が行ったが，訓練データと検証データの損失や正解率に変化は見られなかった．学習が進んでもモデルの性能が向上しておらず，ネットワークがうまく適合できなかった．

6.3 今後の課題

今回の手法では十分な性能が示されなかった．考えられる理由として，ネットワーク構成が適切ではない，用意したデータ量が少なすぎる，データ自体に問題ある点が考えられた．また，現状の問題点としては計算量が多すぎる点がある．

ネットワーク構成については，計算量をもっと減らす必要がある．このモデルは箱罠のストッパーを外す適切なタイミングを算出するため，警戒度を推定するためのものであり，山林内で Jetson Nano のような小型のコンピュータを用いて計算を行う必要があるからである．計算量をへらしつつ警戒度を算出することが課題である．

データ量については，2 クラスで 1862 本 (1 本 90 フレーム) の動画を用意し学習を行ったが，事前学習モデルを用いずに 1 から学習する場合においてはデータセットの量が少なすぎるのではないかと考えた．課題として，事前学習したモデルをファインチューニングを行うか，データ自体をさらに集める必要がある．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　　献

[1] 稲田 樹† 矢倉 章恵† 山本 陽祐† 濱口 沙織† 中井 一文† 江崎 修央† 山端 直人‡ 高橋 完, “檻周辺における害獣の頭数測定” 映像情報メディア学会技術報告 ITE Technical Report Vol.41,No.5

[2] 市川剛士 “ICT を活用した大型捕獲檻によるイノシシの捕獲実証” 新近畿中国四国農業研究 286 － 89,（2019）

[3] Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang1 “Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)

[4] Khan Muhammad ,Mustaqeem , Amin Ullah , Ali Shariq Imran , Muhammad Sajjad , Mustafa Servet Kiran , Giovanna Sannino , Victor Hugo C. “Human action recognition using attention based LSTM network with dilated CNN features” Future Generation Computer Systems Volume 125, December 2021, Pages 820-830

[5] 池田　敬, 野瀬　紹未 , 淺野　玄, 岡本　卓也 , 鈴木　正嗣 “くくりわな捕獲がニホンジカの警戒行動や捕獲場所の利用状況に及ぼす影響” 哺乳類科学　62(2)：141-149，2022　DOI: 10.11238/mammalianscience.62.141

[6] 遠藤 晃, 松隈 聖子, 井上 渚, 土肥 昭夫 “霧島山, えびの高原における観光客によるニホンジカへの餌付けの現状” 哺乳類科学 46（1）：21 - 28，2006

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

2024年3月1日(金) 16:00 ~ 18:10 | T2-A オンライン（Zoom Events)

大規模データ処理応用

座長:鈴木 伸崇(筑波大学)
コメンテータ:塩川 浩昭(筑波大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T2-A-9-01] 犯罪種類間の相関を考慮したニューラルネットワークによる犯罪予測

*杉浦 光彩[1]、白井 匡人[1] (1. 島根大学)

16:30 ~ 16:55
[T2-A-9-02] 1 細胞単位の遺伝子発現の時系列データに対する遺伝子の疾患の進行による関係性の変化の抽出

*増田 智也[1]、浅野 泰仁[1]、小川 達郎[2]、七野 成之[2]、上羽 悟史[2]、松島 綱治[3]、小倉 淳[4] (1. 東洋大学情報連携学部情報連携学科浅野泰仁研究室、2. 東京理科大学大学院生命科学研究科、3. 東京理科大学生命医科学研究所炎症・免疫難病制御部門、4. 長浜バイオ大学アニマルバイオサイエンス学科)

16:55 ~ 17:20
[T2-A-9-03] Telegramにおけるグループ名と投稿メッセージの分析によるグループ探索

*伊藤 純菜[1]、趙 智賢[2]、長田 繁幸[2]、中川 直樹[2]、小口 正人[1] (1. お茶の水女子大学、2. 株式会社日本総合研究所)

17:20 ~ 17:45
[T2-A-9-04] 子宮収縮圧データのための潜在ステージ進行推定

*田村 優香[1,2]、松原 靖子[2]、木村 輔[2]、町村 栄聡[3]、遠藤 誠之[3]、木村 正[3]、櫻井 保志[2] (1. 大阪大学情報科学研究科、2. 大阪大学産業科学研究所、3. 大阪大学大学院医学系研究科)

17:45 ~ 18:10
[T2-A-9-05] SNSにおけるサイバー犯罪対策: 交グラフを活用した犯罪グループ探索アルゴリズム

*大原 望乃[1]、伊藤 純菜[2]、趙 智賢[3]、長田 繁幸[3]、中川 直樹[3]、小口 正人[1] (1. お茶の水女子大学大学院、2. お茶の水女子大学、3. 株式会社 日本総合研究所)

犯罪種類間の相関を考慮したニューラルネットワークによる犯罪予測

杉浦　光彩†　　白井　匡人††

† 島根大学 自然科学研究科　〒 690-0823 島根県松江市西川津町 1060
†† 島根大学 学術研究院理工学系　〒 690-0823 島根県松江市西川津町 1060
E-mail: †n23m104@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp

あらまし　近年，地域の類似性や相互作用に関する相関を利用して犯罪数の時系列予測を行う研究が進められており，その有効性が高まっている．本論文では過去の犯罪の発生場所と種類ごとの犯罪数を基に，その犯罪の発生数を地域ごとに予測する手法を提案する．具体的には，地域間や種類間の相関を考慮した Transformer に基づく予測手法を提案する．提案手法は犯罪が発生した地域の相関に加え犯罪の種類の類似性や相互作用に関する相関を利用する．それにより，異なる地域同士や異なる種類同士の依存関係を学習することができる．そして地域ごと且つ種類ごとに分割された少ないデータの中でも犯罪件数の予測を可能にする．実験では実際のニューヨークの犯罪データを使用し，犯罪種類間の相関の効果を確認することで，提案手法の有効性を示した．

キーワード　時系列予測，犯罪予測，Transformer，CCRF

1　前　書　き

犯罪の発生は生活の安全性や経済の安定性を脅かし社会的な問題となっている．犯罪の抑止はこれらの問題を軽減し社会の発展と大きく関係する．そのため，近年では機械学習を用いた犯罪の予測が多く進められている．既存の手法として，長期的な依存関係を学習することのできる Long Short-Term Memory(LSTM) [1] を使用した手法が多く提案されている．その中でも特に，犯罪の発生した犯罪間の相関を追加で学習し，それぞれの地域で将来発生する犯罪の件数を予測する手法の有用性が高い．これらの手法は犯罪間の相関を学習することで異なる犯罪間や地域間にも犯罪の関連性や発生パターンが存在するか否かを分析できる．例えば既存の手法は，犯罪が発生した特定の地区同士で相関を考慮して犯罪件数を予測している．その結果，犯罪間の関連性を学習し，予測の精度が向上している [2]．そのため，犯罪間の相関を考慮することは予測精度の向上に重要だと考えられる．また，近年時系列予測の分野でも注目されている Transformer [3] を使用した手法も提案されている．既存研究では，将来のイベントの発生の有無を各地域毎に時系列で予測する研究が行われている [10]．この手法の中に Transformer を使用することで高精度の予測を可能にしている．この研究から，Transformer で学習することにより時系列の依存関係を効率よく学習することができると考えられる．

犯罪の抑止に繋がる予測を行うためには，犯罪の種類まで予測する必要がある．しかし全体の犯罪件数のデータから種類ごとの犯罪件数に分割するとデータの数が少なくなる．そのため種類ごとの犯罪数を入力として各種類の犯罪数を予測すると精度が低下するという問題がある．

本稿では図 1 に示すように地域を複数の地区に分割し，各地域で発生する犯罪の件数を種類ごとに，より高い精度で予測する手法を提案する．使用するデータは，実際にニューヨークで発生した犯罪の履歴データを基に犯罪の件数を地区ごとに収集したものである．具体的な手法としては，まず各種類・各地区の犯罪件数を入力し地区ごとに犯罪件数を予測する．そしてニューラルネットワークを用いて異なる地域同士の空間相関を考慮することに加え，異なる種類間の相関も考慮して発生パターンを学習する．これにより種類ごとに分割した数の少ないデータを入力として用いても予測の精度を向上させることが可能である．さらに時系列予測をする際 LSTM の他に Transformer を使用して比較実験を行う．その結果，より精度の高い結果を出力できる手法を調査する．

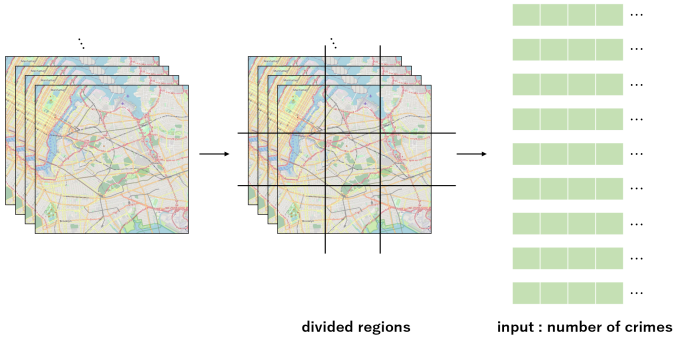


図 1　各地区に分割

2　関 連 研 究

2.1　イベント予測

Wu ら [5] は Transformer を用いてイベントの発生を予測する手法を提案している．Transformer は Attention のみで予測をするため，複雑で長期的な依存関係も効率よく学習できる [10]．元々Transformer は自然言語処理を目的として開発されたモデルであるが，時系列予測においても汎用性が高く，近年時系列予測

の分野でも研究が盛んに行われている．Transformer を使用してイベントの発生する確率の高い地域を予測する手法は，既存の手法である ARIMA(Auto Regressive Integrated Moving Average) や SVR(Support Vector Regression) と比較して高精度である．そのことから，時系列予測における Transformer の有用性が示されている．

2.2 時系列予測のためのモデル

Taylor ら [6] は，トレンド性・周期性・イベント性を分析して SNS のイベント投稿数について時系列予測を行う prophet モデルを提案している．トレンド性の分析にはロジスティック成長モデルを使用し，長期的な成長率を予測している．周期性にはフーリエ級数を使用することで，週ごとや月ごとの発生のパターンを分析する．更に，休日の一覧をモデルに組み込むことで，休日とイベントの発生の関連性を分析しイベント性を学習している．これらの分析によりイベントが発生する日時のパターンを細かく分析し精度の高い予測を可能にしている．Bhatnagar ら [7] は予測と異常検知のための複数のモデルを組み合わせたアンサンブルモデルである Merlion を提案している．このモデルは自動でハイパーパラメータや使用するモデルを選択している．その結果，精度の高い予測や異常検知が可能である．Ke ら [8] は勾配ブースティング決定木を用いた LightGBM モデルを提案している．これはデータ数を減らす手法と特徴量を減らす手法を組み合わせて学習することで計算コストを下げている．これにより，勾配ブースティング決定木を精度を落とさずに高速化している．

2.3 NN-CCRF

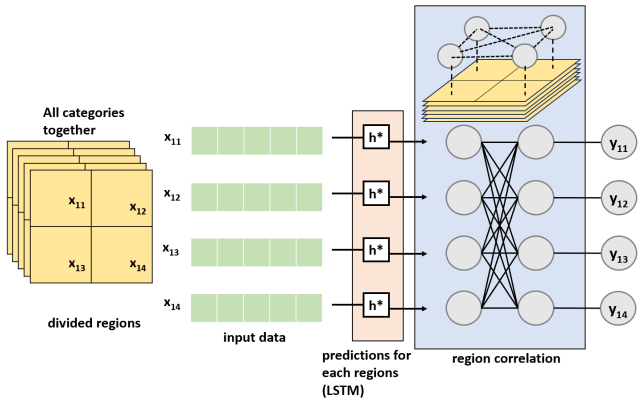


図 2　NN-CCRF モデル

犯罪間の相関を分析することは発生した犯罪同士の依存関係の解析を可能にする．Yi ら [2] は地区ごとに犯罪の予測をすることに加え，地区間の相関を学習することの有効性を示した NN-CCRF モデルを提案している．NN-CCRF モデルは相互依存関係を効率よく学習可能な，ニューラルネットワークに基づいた Continuous Conditional Random Field (CCRF) [9] を使用している．その結果，地区間の類似性や依存関係を学習している．提案された手法は二段階で構成されている．一段階目では地区別に予測を行い，各入力と出力間の関係をモデル化する．二段階目では計算された相関行列に基づいて出力を制約する．NN-CCRF モデルを図 2 に示す．

2.3.1 地区別の予測

$\mathbf{x}$ は入力値である．各要素 x_i は過去に i 番目の地域での発生した犯罪件数である．また，$\mathbf{y}$ は予測した犯罪件数である．各要素 y_i を対象の i 番目の地域の犯罪件数の予測値とする．一段階目では地区ごとに犯罪件数を予測を行う．具体的に地区別の予測は，隠し状態 $\mathbf{h}^*$ を持つ LSTM を使用して，以下のように定式化する．

$$\psi_u(\mathbf{y}, \mathbf{x}, \mathbf{h}^*) = -(\mathbf{y} - R(\mathbf{x}, \mathbf{h}^*))^2,$$
$$R(\mathbf{x}, \mathbf{h}^*) = \sigma(W_h \mathbf{h}_t^* + b_h | \mathbf{x}, W_*, U_*, b_*)$$

ここで，$\psi_u(\mathbf{y}, \mathbf{x}, \mathbf{h}^*)$ は一段階目の予測を表しており，$R(\mathbf{x}, \mathbf{h}^*)$ は空間相関を考慮しない予測値である．$\sigma(\cdot)$ はシグモイド関数，W_*, U_* は重み行列である b_* はバイアスベクトルである．これらは以下の式のように LSTM の中で隠し状態 $\mathbf{h}_t^*$ を更新する際に指定される．

$$f_t = \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1}^* + b_f),$$
$$i_t = \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1}^* + b_i),$$
$$o_t = \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1}^* + b_g),$$
$$g_t = tanh(W_o \mathbf{x}_t + U_g \mathbf{h}_{t-1}^* + b_g)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t,$$
$$\mathbf{h}_t^* = o_t \odot tanh(c_t)$$

ここで f_t は忘却ゲートベクトル，I_t は入力ゲートベクトル，o_t は出力ゲートベクトル，c_t は長期記憶の変数である．

2.3.2 地区間の相関

さらに二段階目では，破損した入力値を除去しながら学習することのできる Stacked Denoising AutoEncoder (SDAE) [4] を使用して犯罪間の相関を調査する．そして各地域で発生し得る犯罪の件数を予測を出力する．具体的には以下の式のように表される．

$$\psi_p(y_i, y_j, \mathbf{K}^*) = -\mathbf{K}_{i,j}^* (y_i - y_j)^2$$

ここで，$\psi_p(y_i, y_j, \mathbf{K}^*)$ は二段階目の予測を表す．$\mathbf{K}^*$ は y_i と y_j の地域間の依存相関を示し，これにより全体的な精度の向上させるために仮推定値を平滑化する．具体的にこの構造は，SDAE を使用して入力データをエンコード・デコードし，破損した入力値と正解を照合するフィードフォワードニューラルネットワークである．破損した入力値 $\hat{\mathbf{y}}$ と正解 $\mathbf{y}$ において最初に $\hat{\mathbf{y}}$ を $\mathbf{z}$ にエンコードする．次に $\mathbf{z}$ をデコードして次のように $\mathbf{y}$ の予測として $\mathbf{y}'$ 生成する．

$$\mathbf{z} = \sigma(W_z \hat{\mathbf{y}} + b_z)$$
$$\mathbf{y}' = \sigma(W_y \mathbf{z} + b_y)$$

SDAE の目的関数は次のように示される．

$$\min_{\{\mathbf{N}*\}} ||\mathbf{y} - \mathbf{y}'||_F^2 + \lambda ||\mathbf{W}||_F^2$$

ここで，W は重み行列であり，y_i と y_j の地区間の相関を学習する．SDAE 層を複数回使用する場合，最終的に以下の目的関数を最小化することによって地区間の相関行列 $\mathbf{K}^*$ を学習する．

$$\min_{\{\mathbf{N}*\}} ||\mathbf{y} - \sigma_n(\mathbf{N}^* R(\mathbf{x}, \mathbf{h}^*))||_F^2 + \lambda ||\mathbf{N}^*||_F^2$$

ここで，$\sigma_n(\mathbf{K}^* R(\mathbf{x}, \mathbf{h}^*))$ は単項予測した結果を n 個の SDAE 層に従って n 回エンコード・デコードした結果である．これにより犯罪が発生した地域間の依存関係を学習することが可能である。

ニューラルネットワークを基本とした CCRF モデル (NN-CCRF) により，犯罪が発生した地域間の関連性や犯罪の発生パターンを効率よく学習し高精度の予測を可能にしている．

さらに詳細に犯罪の予測をするには，各地域で発生する種類ごとの犯罪件数を予測する必要がある．しかし種類ごとに分割した犯罪件数のデータは数が少なくまばらになる．そのため地区間の相関の学習だけでは依存関係の学習が困難になる．そして，予測の精度が低下するという問題がある．

3 提 案 手 法

本章では，NN-CCRF モデルを基に地区間の関連性に加えて犯罪間の関連性も考慮して犯罪件数を予測する手法を述べる．提案手法は地区間の相関と犯罪の種類間の相関を学習する．それにより，各地域と各犯罪種類同士の発生パターンを同時に分析し，より詳細な予測が可能になる．また，一段階目で各地区の犯罪件数を予測する手法として Transformer を使用する．Transformer は自己アテンションによって長期的な時系列の依存関係を効率よく学習できると推測できる．提案する手法を図 3 に示す．

3.1 設　　定

犯罪件数の入力を $\mathbf{x}$ とする．各要素 x_i を種類 i が発生した犯罪件数，x_k を k 番目の地域で発生した犯罪件数とする．$x_{i,k}$ は k 番目の地域で発生した種類 i の犯罪件数を表す．犯罪件数を予測した結果を $\mathbf{y}$ とする．各要素 $y_{i,k}$ は $x_{i,k}$ に対応する k 番目の地域で発生すると考えられる種類 i の犯罪件数の予測値を示す．提案手法のモデルの式を次のように定義する．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(x)} exp\{\psi_u(\mathbf{y}, \mathbf{x}, \mathbf{h}^*) + \sum_{i,j} \psi_p(y_i, y_j, \mathbf{M}^*) + \sum_{k,l} \psi_p(y_k, y_l, \mathbf{N}^*)\}$$

ここで Z(x) は正規化項であり，以下のように定義する

$$Z(x) = \int_y \{\psi_u(\mathbf{y}, \mathbf{x}, \mathbf{h}^*) + \sum_{i,j} \psi_p(y_i, y_j, \mathbf{M}^*) + \sum_{k,l} \psi_p(y_k, y_l, \mathbf{N}^*)\}$$

各項については以下のサブセクションで説明する．

3.2 Transformer による各地区の犯罪件数予測

$\psi_u(\mathbf{y}, \mathbf{x}, \mathbf{h}^*)$ は一段階目での地区別の予測を示す．提案手法は NN-CCRF モデルの地区別の予測で使用された LSTM に代わり Transformer を使用して各地区の犯罪件数を予測する．

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{D_k}})V$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

$$FFN(\mathbf{x}) = max(0, \mathbf{x}W_1 + b_1)W_2 + b_2$$

$$\mathbf{h}' = LayerNorm(SelfAttn(\mathbf{x}) + \mathbf{x})$$

$$\mathbf{h}^* = LayerNorm(FFN(\mathbf{h}') + \mathbf{h}')$$

Q, K, V はクエリ，キー，バリューの長さを示す．D_k は Q, K の次元を示す．W_i^Q, W_i^K, W_i^V, W^O はそれぞれ学習可能な重みであり，b_* はバイアスベクトルである．$SelfAttn(.)$ は自己アテンションモジュールを示し，$LayerNorm(.)$ はレイヤー正規化を示す．$FFN(.)$ はフィードフォワードネットワークを表す．Transformer を使用することで長期間の依存関係を効率よく学習できると考えられる．そして，地区ごとの予測の段階で予測精度を向上させることを目指す．地区ごとの予測精度が向上することで，その後の地区間や種類間の依存関係を考慮した予測も精度が向上すると考えられる．

3.3 地区間の依存関係

NN-CCRF モデルと同様，二段階目では以下のように空間依存関係を学習する．それにより，地域間の犯罪数の類似性や関連性を分析する．

$$\psi_p(y_k, y_l, \mathbf{N}^*) = -\mathbf{N}_{k,l}^*(y_k - y_l)^2$$

ここで，$\mathbf{N}^*$ は y_k と y_l の地域間の依存相関を示す．これにより全体的な精度の向上させるために推定値を平滑化する．近隣の地区など，犯罪の発生パターンが類似する地区は多く見られる．そのため，地区間の依存関係の学習は予測をするうえで重要である．

3.4 種類間の依存関係

ここでは NN-CCRF モデルの空間依存関係に加え種類間の依存関係を学習する．学習は地区間の依存関係の学習と同様に SDAE を使用する．以下のように種類間の犯罪数の類似性を学習する．

$$\psi_p(y_i, y_j, \mathbf{M}^*) = -\mathbf{M}_{i,j}^*(y_i - y_j)^2$$

$\mathbf{M}^*$ は y_i と y_j の種類間の依存相関を示す．

$$\min_{\{\mathbf{M}*\}} ||\mathbf{y} - \sigma_n(\mathbf{M}^* R(\mathbf{x}, \mathbf{h}^*))||_F^2 + \lambda ||\mathbf{M}^*||_F^2$$

ここで，$\mathbf{M}^*$ は重みであり，y_i と y_j の種類間の相関行列を示す．他の種類の犯罪との依存関係を学習することで，より多くの犯罪の発生パターンの学習が可能になる．その結果，少ないデータの中でも効率よく発生する犯罪の件数を予測できると考えられる．

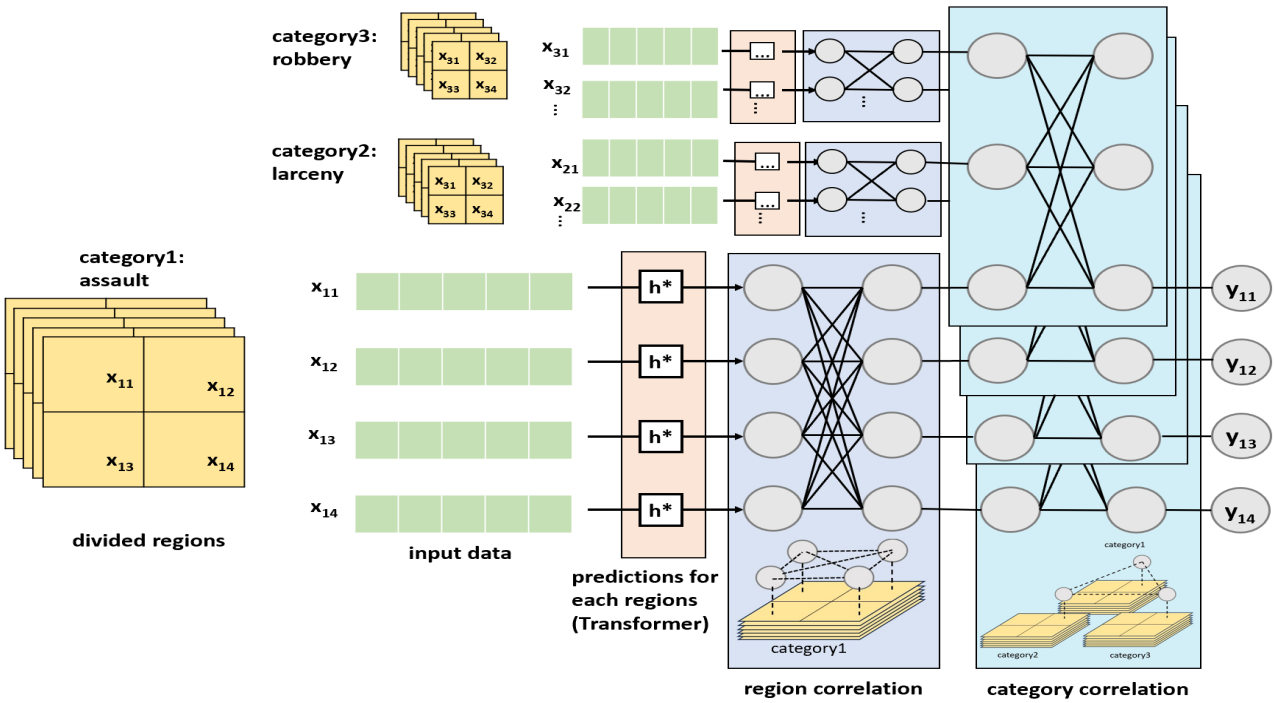


図 3　提案モデル

以上のフレームワークをまとめ，提案手法を以下の式として定める．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(x)} exp\{-(y - \sigma(W_h \mathbf{h}_t^* + b_h|\mathbf{x}, W_*, b_*))^2 - \sum_{i,j} \mathbf{M}_{i,j}^* (y_i - y_j)^2 - \sum_{k,l} \mathbf{N}_{k,l}^* (y_k - y_l)^2\}$$

4 実　　験

4.1 データセット

本論文では実際のニューヨークの 2017 年 1 月から 2019 年 12 月の 47,831 件の犯罪履歴データを使用して実験を行う．この内，学習には 39,280 件のデータを使用する．それぞれの犯罪履歴には発生した日時，場所 (緯度・経度)，犯罪の種類が記されている．緯度と経度を基にニューヨークの一部の地域を複数の地区に分割する．今回は各地区の大きさを 1km × 1km とし，35 地区 (7 × 5) 用意する．時間ステップを 1 日とし，1 日ごとの各地域の犯罪発生を種類ごとに予測をする．

4.2 実 験 設 定

予測した結果を二乗平均平方根誤差 (RMSE) を使用して評価し，他の手法と予測の誤差を比較する．予測する犯罪の種類は，暴行・薬物・強盗・窃盗・交通違反・武器の使用の 6 種類である．予測精度の比較に使用するモデルとして，地区間の相関のみを考慮する LSTM を使用した従来の NN-CCRF(以下，「NN-CCRF」とする) と Transformer を使用した NN-CCRF(以下,「TNN-CCRF」とする)・地区間に加えて種類間の相関を追加で学習するよう改変した LSTM を使用の NN-CCRF(以下，「typeNN-CCRF」とする) で実験を行う．そして提案手法である，地区間と種類間の相関を学習する Transformer を使用した NN-CCRF と比較を行う．さらに時系列予測で汎用性の高いモデルである LSTM・Transformer・Merlion・Prophet・LightGBM との比較も行う．

4.3 実 験 結 果

それぞれのモデルと提案手法を比較した結果を表 1 に示す．TNN-CCRF と提案手法のモデルを比較すると，後者のモデルの精度が 2017 年から順にそれぞれ平均で 0.94%，3.96%，1.98% 上昇している．また，NN-CCRF と typeNN-CCRF モデルの比較においても，後者がそれぞれ順に平均 1.41%，0.27%，2.30%上昇している．NN-CCRF モデルと TNN-CCRF モデルを比較すると，後者の精度が 2017 年から順に平均で 4.41%，5.35%，6.58% 上昇している．さらに，他の時系列予測のために提案された手法との差は 2017 年から 2019 年を平均して，Merlion は 3.22%，Prophet は 27.8%，LightGBM は 2.63%となっている．いずれも Transformer を使用した種類間の相関を学習する NN-CCRF モデルの精度が高くなっている．

4.4 考　　察

地域間の相関関係に加えて種類間の相関関係を考慮したモデルでは，地区間・種類間の依存関係をどちらも効率よく学習できており，精度が上昇したと考えられる．さらに LSTM を使用したモデルよりも Transformer を使用したモデルの精度が上昇している．LSTM のみで実験したものと Transformer のみで実験したものを比較しても Transformer の方が予測精度が高くなっている．そのため，Transformer は犯罪の時系列予測に使用するモデルとして適していると考えられる．

5 ま と め

本研究では犯罪件数を犯罪の種類ごとに予測する手法を提案した．提案手法では既存手法で学習していた，犯罪が発生した地区間の相関関係に加えて，犯罪の種類間の相関関係も学習する方法を示した．この相関の分析により予測の精度が向上したことを確認した．さらに各地区の犯罪件数を予測する手法として，LSTM と Transformer を比較した．結果として Transformer を使用したものがより高い精度で予測できることが示された．

文　　献

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. ”Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.
[2] Yi, F., Yu, Z., Zhuang, F., Guo, B. (2019, August). Neural Network based Continuous Conditional Random Field for Fine-grained Crime Prediction. In IJCAI.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[4] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(12).
[5] Wu, X., Huang, C., Zhang, C., Chawla, N. V. (2020, April). Hierarchically structured transformer networks for fine-grained spatial event forecasting. In Proceedings of The Web Conference 2020.
[6] Taylor, S. J., Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.
[7] Bhatnagar, A., Kassianik, P., Liu, C., Lan, T., Yang, W., Cassius, R., ... Wang, H. (2021). Merlion: A machine learning library for time series. arXiv preprint arXiv:2109.09265.
[8] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
[9] Tao Qin, Tie-Yan Liu, Xu-Dong Zhang,De-Sheng Wang, and Hang Li. Global ranking using continuous conditional random fields. In Advances in neural information processing systems, pages 1281–1288, 2009.
[10] Wu, N., Green, B., Ben, X., O’Banion, S. (2020). Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv:2001.08317.

表 1　各モデルとの比較結果 (上から順に 2017 年，2018 年，2019 年の結果)

	assault	drug	robbery	larceny	vehicle	weapon	平均
LSTM	0.4730	1.1515	0.3543	0.5117	0.6005	0.4264	0.5862
Transformer	0.3922	0.8615	0.3723	0.4859	0.4452	0.3865	0.4909
NN-CCRF(種類間の相関無)	0.3265	0.8080	0.3277	0.4852	0.4489	0.3647	0.4601
NN-CCRF(種類間の相関有)	0.3190	0.7991	0.3166	0.4844	0.4468	0.3559	0.4536
Transformer NN-CCRF(種類間の相関無)	0.3080	0.7847	0.3245	**0.4220**	0.4327	0.3543	0.4377
Transformer NN-CCRF(種類間の相関有)	**0.3065**	**0.7729**	**0.3119**	0.4223	0.4299	**0.3517**	**0.4336**
Merlion	0.3202	0.7963	0.3255	0.4616	0.4107	0.3776	0.4487
Prophet	0.4389	1.0566	0.4165	0.5948	0.4823	0.4355	0.5708
LightGBM	0.3176	0.7791	0.3236	0.4376	**0.4103**	0.3661	0.4390

	assault	drug	robbery	larceny	vehicle	weapon	平均
LSTM	0.5023	1.0077	0.4322	0.5176	0.4530	0.4098	0.4843
Transformer	0.3622	0.6213	0.3755	0.4527	0.3600	0.3312	0.4174
NN-CCRF(種類間の相関無)	0.3510	0.6101	0.3222	0.4231	0.3470	0.3196	0.3955
NN-CCRF(種類間の相関有)	0.3360	0.5808	0.4143	**0.4091**	0.3262	0.3005	0.3944
Transformer NN-CCRF(種類間の相関無)	0.3338	0.5869	**0.3006**	0.4800	0.3291	0.2938	0.3887
Transformer NN-CCRF(種類間の相関有)	**0.3304**	**0.5776**	0.3040	0.4098	**0.3254**	**0.2929**	**0.3733**
Merlion	0.3421	0.6431	0.3166	0.4255	0.3538	0.3144	0.3792
Prophet	0.4643	0.8181	0.5105	0.5637	0.4632	0.4104	0.5384
LightGBM	0.3387	0.6093	0.3124	0.3987	0.3462	0.3078	0.3855

	assault	drug	robbery	larceny	vehicle	weapon	平均
LSTM	0.4978	0.6385	0.4254	0.5040	0.2897	0.3428	0.4388
Transformer	0.3843	0.4646	0.3408	0.4235	0.2227	0.3270	0.3604
NN-CCRF(種類間の相関無)	0.3207	0.4851	0.2903	0.4522	0.2192	0.3504	0.3529
NN-CCRF(種類間の相関有)	0.3223	0.4638	0.3137	**0.4018**	0.2163	0.3511	0.3448
Transformer NN-CCRF(種類間の相関無)	0.3236	0.4316	0.2894	0.4177	0.2113	0.2985	0.3286
Transformer NN-CCRF(種類間の相関有)	**0.3177**	**0.4193**	0.2808	0.4072	**0.2077**	**0.2896**	**0.3221**
Merlion	0.3354	0.4565	**0.2796**	0.4350	0.2239	0.2979	0.3381
Prophet	0.4471	0.5549	0.4480	0.5817	0.2992	0.3748	0.4510
LightGBM	0.3343	0.4277	0.2909	0.4167	0.2134	0.3030	0.3310

1 細胞単位の遺伝子発現の時系列データに対する遺伝子の疾患の進行による関係性の変化の抽出

増田 智也† 浅野 泰仁‡ 小川 達郎‡ 七野 成之‡ 上羽 悟史‡ 松島 綱治‡ 小倉 淳‡

†東洋大学情報連携学部 〒115-8650 東京都北区赤羽台 1-7-11

E-mail: †tomomasu1113@gmail.com

あらまし 高齢化が進行している社会を持続可能にするためには慢性に関連すると言われている生活習慣病やがんなどを予防又は早期診断することが望ましいとされている．そのために疾患進行の仕組みを解明することが求められている．本研究は 1 細胞単位の遺伝子発 現の時系列データを使用して疾患進行の仕組みを解明することの一助となる分析を行う．具体的には疾患進行による遺伝子の発現量に関する相関の違いを利用したネットワークを構築し，このネットワークに対してグラフ理論を用いた分析を行う．この分析により疾患の進行による重要な遺伝子対を捉え，仕組みの解明の発展に寄与することが目的である．実験では肺線維症誘導マウスから採取したデータを使用し実験を行った．その結果，肺線維症誘導マウスに関連する可能性がある遺伝子群を発見した．

キーワード 空間・時空間・時系列データ 処理, グラフデータ処理, 医療・ヘルスケア

1. はじめに

日本国では急速に高齢化が進んでおり，また平均寿命が延びていることにより生活習慣病，肺線維症，がんなどの基盤となる慢性炎症によって苦しむ可能性がある人数や年数が増加している．慢性炎症になる前の状態は病気を患っていない状態(正常状態)，自覚症状を伴わない状態(未病状態)の 2 つの状態がある．この未病状態での検出は現在の技術では困難であるがこの時点での早期対処が求められている．そのため，検出し早期対処を行うため未だに解明されていない未病状態のメカニズム解明が求められている．

昔は細胞群単位でしか遺伝子を採取する技術しかなかったが，昨今では 1 細胞単位での採取が可能となった．細胞単位での解析は個々の細胞集団，さらには組織微小環境全体の多様性を解析でき疾病における細胞間相互作用，特異的マーカー遺伝子の探索，薬物治療抵抗性の制御機構などの研究にとって非常に有用である[1]．本研究では，1 細胞単位の遺伝子発現の時系列データを使用し未病状態の解明を行う．

1 細胞単位の遺伝子発現の時系列データから遺伝子相関ネットワーク[2]を作成して未病状態の解明することは以前から行われてきた．遺伝子相関ネットワークを使用することによって疾病の進行に伴い負の辺からなる遺伝子相関ネットワークは有意に変化することが観察された[2]．また遺伝子相関ネットワークを用いてクラスタリングを行うことにより似た特徴を持つ遺伝子群の発見した研究などが行われている[3]．

本研究の目的は疾患の進行につれ，弱い相関から強い相関へと変化した遺伝子対が慢性炎症の仕組みの解明に繋がると仮説を立て病状の進行により変化する相関によって現れる遺伝子群の特徴を求めることである．そのために肺線維症に誘導したマウスの 1 細胞単位の遺伝子発現の時系列データを使用し正常状態では弱い相関であるが疾患進行時には強い相関を持つ遺伝子対の集合を辺集合，辺集合に現れる遺伝子の集合を点集合とするネットワークを構築する．このネットワークに対してグラフ理論を用いて分析することにより炎症によって新たに現れる遺伝子群の特定と特徴を捉える．

本研究の結果，肺にある筋肉に関連する遺伝子群が確認されたため肺の筋肉に変化が生じている可能性があると思われる．

2. 関連研究

本節では本論文で使用する研究を紹介する．本論文では 単一細胞から遺伝子を抽出するときに発生するノイズを除去する MAGIC，MAGIC を用いた遺伝子相関ネットワークの構築，作成した遺伝子相関ネットワークから似た遺伝子群の抽出を目的とする誘導部分グラフの構築，構築した誘導部分グラフの点集合に含まれる遺伝子の重要度ランキングの作成，作成したネットワークの可視化が行われている．

2.1. MAGIC

MAGIC(Markov Affinity-based Graph Imputation

of Cells)とは Dijk ら[4]によって提唱された単一細胞から遺伝子を抽出するときに発生するノイズを拡散マップ[5]を用いて除去する手法である．このノイズは発現している遺伝子の数が少ない場合に「ドロップアウト」と呼ばれる発現している遺伝子が検出されない現象が起きる可能性があるため発生する．この「ドロップアウト」が発生しているとデータ内の遺伝子間の関係性が失われてしまい研究を行うことが困難になるためノイズを除去する必要がある．MAGIC は拡散マップを使用しこのノイズの除去を可能にした．具体的には，行を遺伝子，列を細胞，要素は細胞における遺伝子の発現量である行列データを行を細胞，列は遺伝子，要素は特徴量である行列データにすることによりノイズ除去をする．本研究では MAGIC の提案者の一人である Krishnaswamy の研究室が公開している実装の Python 版[*1]を用いている

2.2. 遺伝子相関ネットワークの作成と可視化

本研究では遺伝子相関ネットワークの作成と可視化は遺伝子相関ネットワークの提案者である浅野教授らが公開している実装[*2]を用いている．

2.2.1. 遺伝子相関ネットワーク

遺伝子相関ネットワークとは浅野ら[2]が提唱した遺伝子間の関係を分析するためのものである．このネットワークは疾患の進行につれて遺伝子の発現量に変化が観測されるため関連性があるのではないかと仮説に基づいて提案された．ネットワークを構成する頂点集合は遺伝子の集合であり辺集合は遺伝子対の集合である．提唱された遺伝子相関ネットワークは強い相関を持つ遺伝子対間に辺が存在する．本論文では遺伝子相関ネットワークの解釈を広げ，任意の区間内に相関がある遺伝子対間に辺が遺伝子相関ネットワークに対して可視化を行う場合，遺伝子を表す点の位置を固定することにより異なる可視化結果同士を比べることを容易にする．図 1 は，遺伝子相関ネットワークを可視化した結果の一例である．点の Y 座標を遺伝が存在する染色体の番号に対応し X 座標はその染色体内での遺伝子の位置に対応する．また遺伝子を結ぶ重み付きの辺は重みである相関係数が非負の時は青，負の時は赤とした[2]．肺線維症誘導マウスの 1 細胞単位の遺伝子発現の時系列データに対して遺伝子相関ネットワークを作成し可視化した結果，負の相関が一過性の増加していた．負の相関の一過性の増加は病気の進行で細胞機能の転換が起きている可能性があると示唆された．

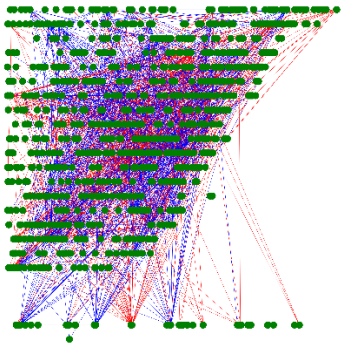


図 1　遺伝子相関ネットワークの可視化の例

2.3. 遺伝子群の抽出と遺伝子の重要度のランキング

本節では，遺伝子群の抽出によりおおむね似た働きの遺伝子で形成されるコミュニティの抽出した研究を紹介する．遠藤[3]は本研究と同様に肺線維症に誘導したマウスの 1 細胞単位の遺伝子発現の時系列データから遺伝子群の抽出と遺伝子の重要度のランキングを作成した．遺伝子群の抽出は Louvain 法を使用し，重要度のランキングは PageRank を使用した．その結果，リボソームプロテインに関係する遺伝子や小胞体ストレス関連のネットワークが確認され遺伝子機能に何らかの異常が起きている可能性が示唆された．

2.3.1. Louvain 法

Blondel ら[6]が提案した Louvain 法はデータを一定の条件下で複数個のグループに分類するクラスタリング手法である．この Louvain 法はクラスタリング手法の中でも比較的優れている．この手法はヒューリスティックでありクラスタリングの手法の精度を示す Newman ら[7]が提案した Modularity を必ずしも最大化するものではない．Modularity と実行時間はトレードオフの関係である．しかし Louvain 法は Modularity と実行時間の両方を高い水準でクラスタリングできる比較的優れているクラスタリング手法である．

2.3.2. PageRank

Brin と Page[8]が提案した PageRank とはネットワーク上のどの点が重要であるかをランキング化する手法の 1 つである．この PageRank はインターネット検索に使われており．多くの点と繋がっている点は重要であるという思想のもとに開発された．本論文では辺の重みを考慮する Xing と Ghorbani [9]が提案した WPR(Weighted PageRank)を使用する．

3. 研究手法

本節では，先行研究に則り遺伝子発現データから

*1 https://krishnaswamylab.org/projects/magic
*2 https://github.com/gnladdy/gene_correlation_network

遺伝子相関行列を作成する．次に本研究で提案する弱い相関から強い相関へと変化した遺伝子対を辺集合とする遺伝子相関ネットワークを構築する．このネットワークを変化した遺伝子相関ネットワークと呼称する．最後に作成した遺伝子相関ネットワークを先行研究に従い分析を行う．なお，データを取得した日時の系列を $(t_1, t_2, .. , t_d)$と表す(d はデータの個数)．ある日時t_iにおける各細胞の遺伝子発現データを$M(t_i)$とする．

3.1. 前処理とノイズ除去と遺伝子相関行列の作成

今回使用するデータにはどの細胞にもほとんど発現していない遺伝子がある．その遺伝子を含めて実験を行う場合，後で説明する遺伝子相関行列を作成時にとても強い正の相関が現れる．しかしこの相関関係には因果関係がない確率が高いと思われるため除去する必要があるとされている．そのため今回は 1500 個程度の遺伝子が残るようにした．また単一細胞から遺伝子を抽出時にノイズが発生するため除去する必要がある．そのため MAGIC を適用した．次にノイズを除去した後の各時点のデータから遺伝対の相関を要素とした遺伝子相関行列を作成する．作成した遺伝子相関行列を$C(t_i)$と表す．

3.2. 変化した遺伝子相関ネットワークの作成

まず作成した遺伝子相関行列$C(t_i)$から 2 種類の辺集合を作成する．1 つ目は，強い相関を持つ遺伝子対の集合$X(t_i)$を作成する．2 つ目は，弱い相関を持つ遺伝子対の集合$Y(t_i)$を作成する．次に変化した遺伝子相関ネットワークの辺集合を作成する．この辺集合はt_1では相関が弱かったがt_i(i は 2 以上)では相関が強い遺伝子対の集合である．この集合を$Z(t_i) \equiv (Y(t_i) \cap X(t_1))$と表す．作成した辺集合$X(t_i), Y(t_i), Z(t_i)$と遺伝子の点集合を使用し遺伝子相関ネットワークを作成する．辺集合$X(t_i), Y(t_i)$で作成した遺伝子相関ネットワークをそれぞれ遺伝子強相関ネットワークと遺伝子弱相関ネットワークと呼称する．

3.3. クラスタリングとランキングと可視化

変化した遺伝子相関ネットワークに対して Louvain 法を適用しクラスタリングを行う．クラスタリングにより似た特徴を持つ遺伝子群を得られることが期待される．また遺伝子群に対して PageRank を使用する遺伝子の重要度のランキング作成手法を適用する．この手法は重みである相関係数は絶対値をとることにより正負による違いが出ないにする．重要度のランキングは遺伝子群内の重要度を表しどの遺伝子が他の遺伝子に対して影響力を持つかを推測することの一助となることを期待されている．遺伝子強相関ネットワークと遺伝子弱相関ネットワークと変化した遺伝子相関ネットワークに対して遺伝子相関ネットワークに可視化を適用する．

4. 実験結果と考察

4.1. 遺伝子相関ネットワークの作成結果と考察

本節では 3.2 で作成した遺伝子相関ネットワークの結果と考察を行う．

4.1.1. 遺伝子相関ネットワークの作成結果

表 1 は遺伝子強相関ネットワークの辺の数をまとめたものである．どの時点においても相関が正である辺と比べて負である辺が多い結果となった．また 7 日目と 21 日目以外の区間[0.95, 1.00]を除いてすべての時点ですべての区間において辺の数は 0 日目，7 日目，21 日目，3 日目の順に多いという結果となった．

表 2 は遺伝子弱相関ネットワークの辺の数をまとめたものである．どの時点においても相関が正と負である辺の数はほぼ同じであった．

表 3 は変化した遺伝子相関ネットワークである．どの時点においても区間[0.95, 1.00]にある辺は約 10 本程度であった．

表 4 は疾患の進行により相関が強くなる変化があった遺伝子対の割合を示したものであある．つまり$Z(t_i)$の辺の数を$X(t_i)$の辺の数で割った値である．区間の集合{[-1.00, -0.60], [0.95, 1.00]}に含まれる遺伝子対は約 15％であった．区間[-1.00, -0.60] に含まれる遺伝子対は約 20％共通していたが区間[0.95, 1.00]に含まれる割合は 1%にも満たない結果となった．

図 2 は表 3 で表した変化した遺伝子相関ネットワークを可視化したものである．遺伝子を緑色の点，遺伝子対の相関が負の区間に含まれている場合は赤い線，非負の区間に含まれている場合は青い線で結ばれている．3 日目，7 日目，21 日目には赤い線，つまり負の相関が大半を占めていることがわかる．

表 1　遺伝子強相関ネットワークの辺の数(本)

相関係数の区間	0 日目	3 日目	7 日目	21 日目
{[-1.00, -0.60], [0.95, 1.00]}	42,768	36,221	42,512	35,606
[-1.00, -0.60]	34,738	30,048	36,021	27,743
[0.95, 1.00]	8,030	6,173	6,491	7,863

表 2　遺伝子弱相関ネットワークの辺の数(本)

相関係数の区間	0 日目	3 日目	7 日目	21 日目
(-0.30, 0.30)	432,757	416,726	529,203	412,257
(-0.30, 0.00)	214,903	230,354	284,717	223,020
(0,00, 0.30)	217,848	186,362	244,475	189,232

表 3　変化した遺伝子相関ネットワーク

相関係数の区間	3 日目	7 日目	21 日目
{[-1.00, -0.60], [0.95, 1.00]}	6,447	6842	4861
[-1.00, -0.60]	6433	6835	4848
[0.95, 1.00]	14	7	13

表 4　相関が変化した遺伝子対の割合(%)

相関係数の区間	3 日目	7 日目	21 日目
{[-1.00, -0.60], [0.95, 1.00]}	17.8	16.1	13.7
[-1.00, -0.60]	21.4	19.0	17.5
[0.95, 1.00]	0.0237	0.0108	0.0165

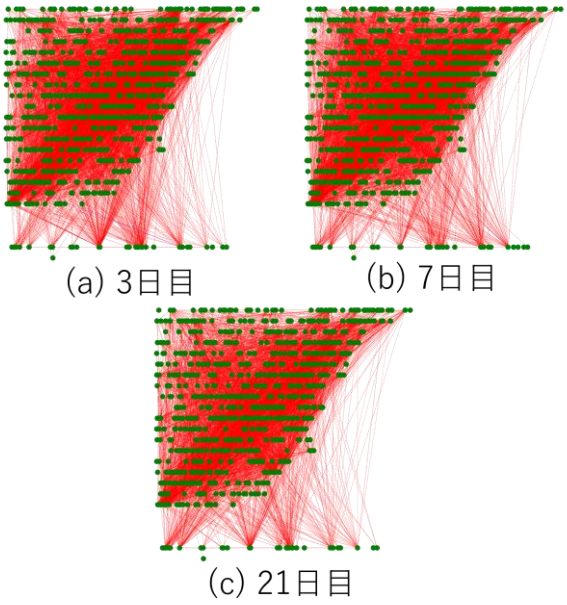


図 2　変化した遺伝子相関ネットワーク

4.1.2. 変化した遺伝子相関ネットワークの考察

相関関係が変化した遺伝子対の割合が約 15%であった．そのため最大で約 15%の遺伝子対に因果関係は疾患の進行により新たに現れた可能性があることがわかる．

4.2. コミュニティ抽出結果と考察

本節では 3.3 で示した手法により遺伝子群のコミュニティを抽出した結果と考察を行う．

4.2.1. コミュニティ抽出結果

図 1 で示したそれぞれの変化した遺伝子相関ネットワークに対して Louvain 法を使用する．表 5 は 3 日目，7 日目，21 日目で作成されたコミュニティの数と遺伝子の数を示す．クラスタは所属している遺伝子の数が多い順に番号が振られている．

表 6 は表 5 で表したコミュニティに対して PageRank を使用し遺伝子の重要度ランキングを出したものである．

図 3，4，5 はそれぞれ 3 日目，7 日目，21 日目にコミュニティ抽出を可視化したものである．

表 5　変化した遺伝子ネットワークをクラスタリングした結果

クラスタ	3 日目	7 日目	21 日目
1	253	185	163
2	180	161	148
3	169	149	148
4	148	121	146
5	129	115	143
6	106	109	120
7		74	37
8		16	
9		2	
10		2	

表 6　重要度上位 10 個の遺伝子

採取日	重要度	クラスタ1	クラスタ2	クラスタ3	クラスタ4	クラスタ5
3日目	1	Galnt2	Zc3h11a	Ncor1	Phlda1	Gpx1
	2	Anxa6	Zfp207	Rorb	Tmed2	Cox8a
	3	Psmb4	Dnaja1	Tnrc6b	Lmo4	Tuba1c
	4	Gstp1	Ppm1k	Ube2s	Ssr4	Selenos
	5	Psmb1	Malat1	Actg1	Apoe	Atp6v1g1
	6	Cope	Cox6a1	Vdac2	Cd63	Ptp4a1
	7	Swi5	Gm12174	Atox1	Lgmn	Bri3
	8	Gstp2	Rpl22	Psma2	Cox7b	Reep5
	9	Pdap1	Rps13-ps4	Eln	P4hb	Dnajb14
	10	Magt1	Gm9794	Arpp19	Sec61b	Ybx3
7日目	1	Map1lc3b	Tmed2	Ghitm	Cox7c	Mia3
	2	Dcaf10	Tmem258	Irf2bp2	Atp5h	Kcnq1ot1
	3	Atp6v1e1	Bag1	Ubxn1	Gm12338	Zfp12
	4	Fth1	Tln1	Ppp3ca	Yy1	Gm19950
	5	Ptp4a1	Cast	Galnt2	Ndufb11	Zeb2
	6	Sdcbp	Ccdc82	Gnb2	Uqcrb	Atp5g2
	7	Cd24a	Prkar1a	Rock1	Bhlhe40	Mrgpra6
	8	Prdx5	Btg2	D8Ertd738e	Prdx2	Lrrc7
	9	Riok3	Ndufa13	Wnk1	Ndufa4	Calr
	10	Atp6v0c	Atp5d	Vdac2	Fosl2	Tbca
21日目	1	Dad1	Ndufa13	Pabpn1	Dag1	Map1lc3b
	2	Grcc10	Tmem258	Ppp3ca	Ndufb11	Selenos
	3	Fkbp8	Xist	Klhl24	Gstp2	Prdx5
	4	Anxa6	Hmgn1	Opa1	Rrbp1	Fth1
	5	Tm9sf3	Atp5d	Dnaja2	Wls	S100a11
	6	Atp5o	Ndufa3	Txnip	Gstp1	Gm11100
	7	Bcap31	Zc3h11a	Tmed2	Tmem30a	Atp6v1e1
	8	Sarnp	Nisch	Fis1	Dynlt3	Zfand5
	9	Cd24a	Mrgpra6	Rbm8a	Mettl7a1	Cd63
	10	Atp5h	Son	Ewsr	Os9	Ctsb

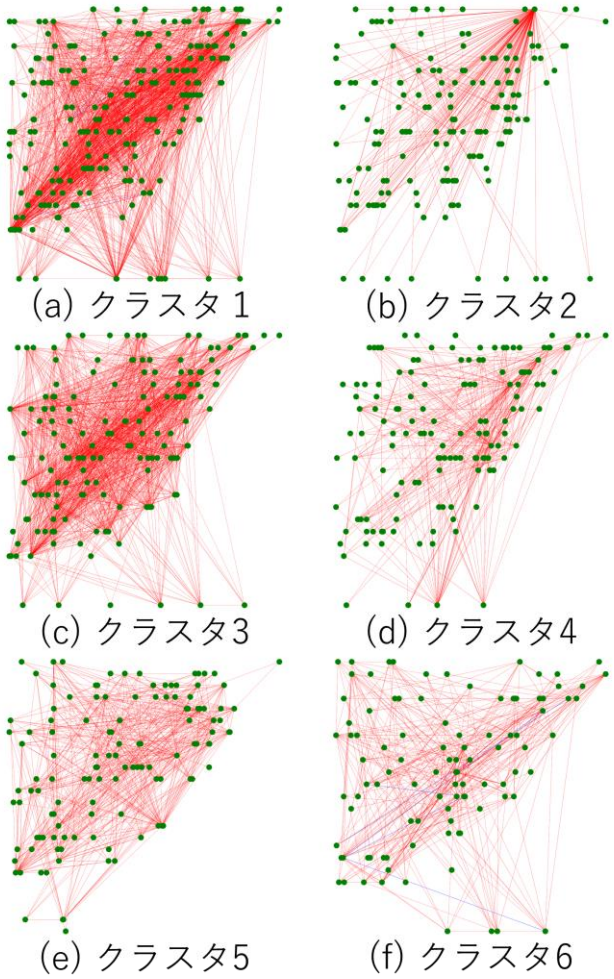


図 3　3 日目のクラスタリング結果

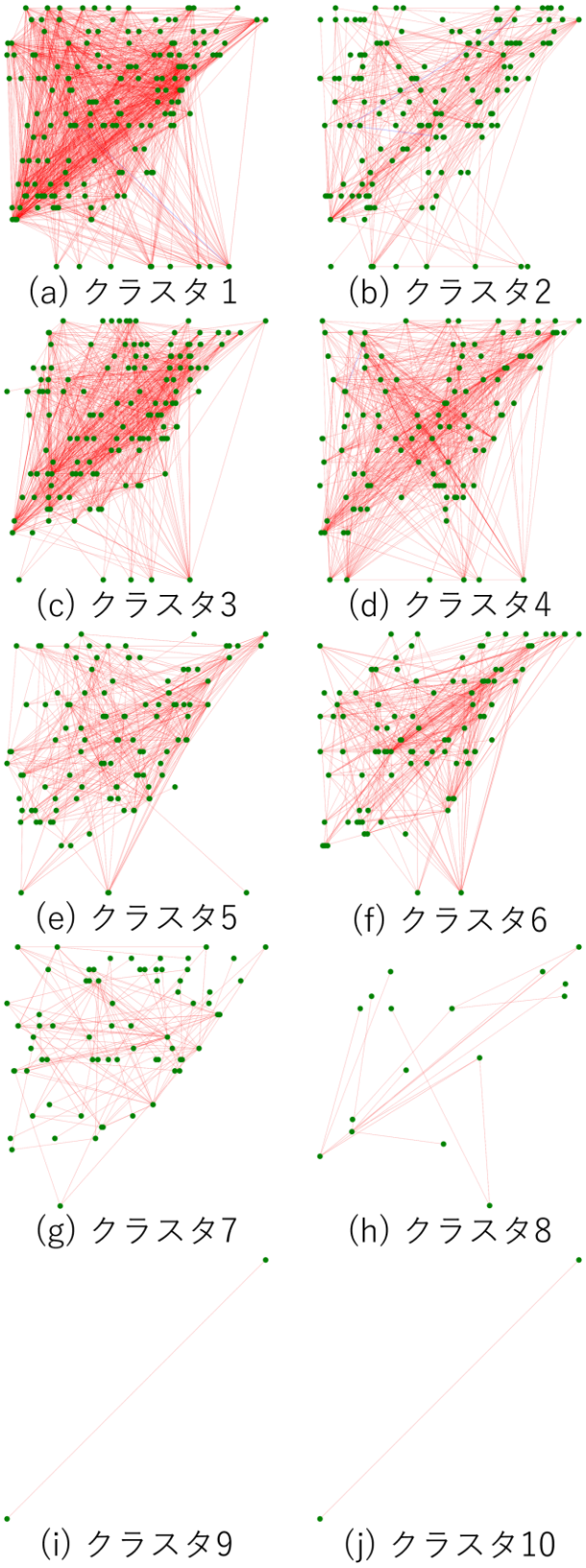


図 4　7 日目のクラスタリング結果

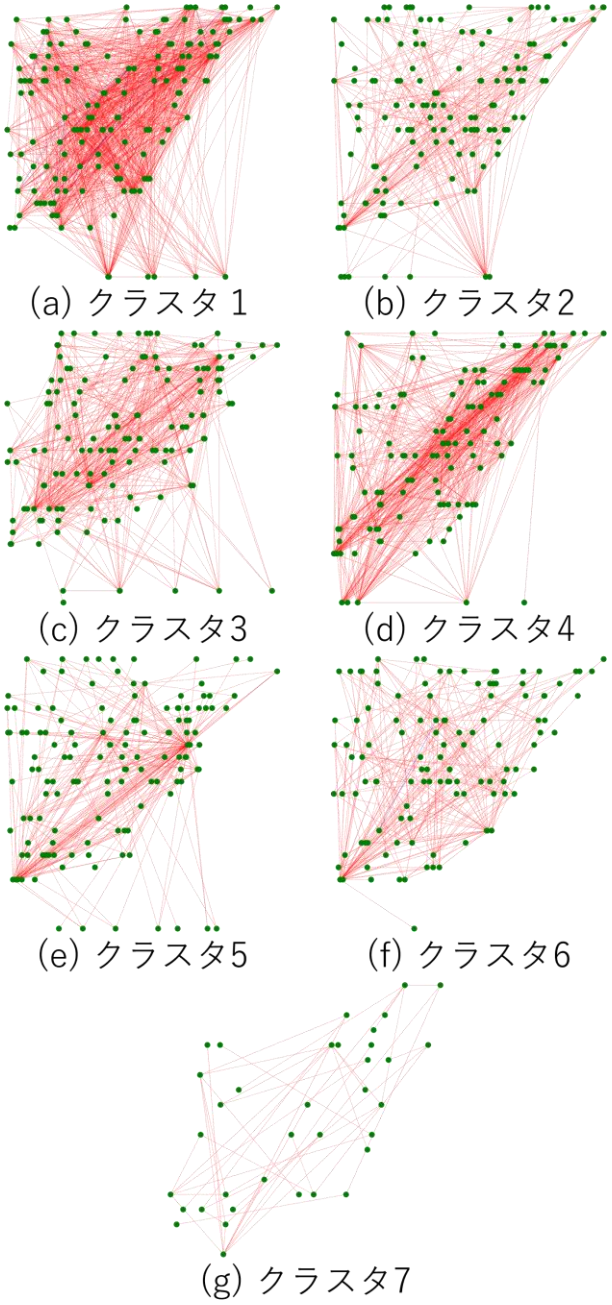


図 5　21 日目のクラスタリング結果

4.2.2. クラスタに関する考察

g:Profiler と THE HUMAN PROTEIN ATLAS を使用してクラスタに対して考察を行った．g:Profiler とは BIIT[10]によって開発されたエンリッチメント解析を行うことができるツールである．エンリッチメント解析とは，ある遺伝子またはタンパク質のセットにはどのような特徴があるかを調べる解析手法である．THE HUMAN PROTEIN ATLAS とは Mathias Uhlén ら[11]が主導している特定の遺伝子またはタンパク質を検索できるツールである．

どの時点においても得られたクラスタ 1 からクラスタ 6 までのクラスタは 100 個以上の遺伝子が所属しているため考察を行える十分な大きさのコミュニティを抽出できたと思われる．100 個以上の遺伝子が存在するクラスタに対して考察を行った．以下のクラスタに関する考察によりクラスタ内の遺伝子には関連性があると思われる．

3 日目のクラスタ 1 はこの 10 個の遺伝子は平滑筋の制御に関連する可能性がある．平滑筋とは内臓や血管，気管の内側に存在する筋肉である[12]．筋肉が弛緩・収縮することによって内臓や血管の働きを維持する役割を持つ．肺にも平滑筋は存在するため肺線維症によって平滑筋に関する制御に影響が出たのではないかと思われる．

3 日目のクラスタ 2 と 7 日目のクラスタ 1 には神経系の機能に関連している可能性がある．肺線維症により神経系に変化をもたらす[13]と報告されているが今回使用したデータは肺細胞から採取した遺伝子であるため関連性は疑わしいと思われる．

3 日目のクラスタ 3 と 7 日目のクラスタ 4 と 21 日目のクラスタ 3，5 には代謝プロセスに関連するクラスタであると思われる．代謝の変化は，線維症の発症機序の特徴であるだけでなく，その機序において影響力のある役割を果たす可能性もあることが示唆されている[14]．

3 日目のクラスタ 4 と 7 日目のクラスタ 2 には循環器系や血管に関連している可能性がある．肺と血管は肺で血管中の血液との間でガス交換が行われるため肺線維症と関連性はあると思われる．

3 日目のクラスタ 5 には血管や有毒物質や解毒に関連する可能性がある．実験に用いられたネズミはシリカ粒子と呼ばれる有害物質を用いて肺線維症を発症させたため関連性があると思われる．

3 日目のクラスタ 6 と 7 日目のクラスタ 3 にはタンパク質の輸送に関する可能性がある．しかし現状のところ肺線維症との関連は不明である．

7 日目のクラスタ 5 と 21 日目のクラスタ 1，2 はヌクレオチド 3 リン酸の合成に関するクラスタである可能性がある．しかし現状のところ肺線維症との関連は不明である．

7 日目のクラスタ 6 はアクチンと呼ばれるたんぱく質に関連する可能性がある．アクチンは細胞運動，細胞質分裂，細胞の形態や細胞膜構造の維持，細胞内輸送，形態形成，貪食，さらには細胞核内構造の維持や転写制御など極めて多様な現象で中心的な役割を果たしていることが判明したタンパク質である[15]．しかし現状のところ肺線維症との関連は不明である．

21 日目のクラスタ 4，6 は酵素やタンパク質などの物質の結合に関連する可能性がある．しかし現状のところ肺線維症との関連は不明である．

使用したデータセットは 1 セットのみであったため検証が不十分であると思われるが肺線維症に関連すると思われる遺伝子群を抽出できたため提案手法は成功したと思われる．

5. おわりに

本研究では 1 細胞単位の遺伝子発現の時系列データを使用し正常状態の相関が弱いが未病状態や慢性炎症や線維化した状態では相関が強くなった遺伝子対に辺を持つ遺伝子相関ネットワークの作成手法を提案した．その結果，提案は成功したと思われる．今後の展望としては正常状態では強い相関を持つが疾患の進行により相関が弱くなる遺伝子対に辺を持つ遺伝子相関ネットワークを作成すると良いと思われる．これによって正常状態にはあった機能に関する遺伝子群の発見に繋がると期待している．また本研究では疾患の進行により相関が変化した遺伝子対に辺を持つ遺伝子相関ネットワークを抽出したが相関が変化した辺の集合を遺伝子強相関ネットワークにおける重みを強くすることにより遺伝子強相関ネットワークの遺伝子群の関係性を失わずに抽出できると思われる．はたまた，本研究では 0 日目と 3 日目，7 日目，21 日目の組合せで比較したが 3 日目と 7 日目で相関が変化した辺を探してもよいだろう．

本研究で扱った予防を科学する炎症細胞社会学は未だ解明されていない点があり解明の余地はあると思う．今後の研究によって誰でも病気を早期治療することが可能になることを祈る

謝　　辞

本研究は JSPS 科研費 22K12043，22H05064 および AMED の課題番号 JP21gm6210025 の助成を受けたものです．感謝いたします．

参　考　文　献

[1] 和歌山県立医科大学. “【総説】シングルセル遺伝子発現解析と細胞集団”. 和歌山県立医科大学. （n.d.）. https://www.wakayama-med.ac.jp/dept/igakubu/160928/sousetsu-01.html/, (参照 2023-12-20).

[2] 浅野泰仁，小倉淳，七野成之，上羽悟史，松島綱治. 1 細胞遺伝子発現時系列データの遺伝子相関ネットワークによる分析. 第 39 回医療情報学連合大会，2019.

[3] 遠藤美空. 東洋大学情報連携学部卒業論文

[4] Dijk D. et al. Recovering gene interactions from single-cell data using data diffusion. Cell, 2018

[5] Ronald R Coifman and Stéphane Lafon. Diffusion maps. Applied and computationalharmonic analysis, Vol. 21, No.

1, pp. 5–30, 2006.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, Vol. 2008, No. 10, p.p10008, 2008.

[7] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical review E, Vol. 69, No. 2, p. 026113, 2004.

[8] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, Vol. 30, No. 1-7, pp. 107-117, 1998.

[9] W. Xing and A. Ghorbani. "Weighted PageRank algorithm," Proceedings. Second Annual Conference on Communication Networks and Services Research, pp. 305-314, 2004.

[10] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, Hedi Peterson: g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update) Nucleic Acids Research, May 2023; doi:10.1093/nar/gkad347 [PDF].

[11] The Human Protein Atlas, https://www.proteinatlas.org/. (参照 2024-01-04).

[12] 公益財団法人長寿科学振興財団. “筋肉の種類とその特徴”. 健康長寿ネット. 2023-08-16. https://www.tyojyu.or.jp/net/kenkou-tyoju/undou-kiso/kinniku-kinou.html. (参照 2024-01-04).

[13] 須藤守夫. 呼吸器疾患における気管支壁神経系の電子顕微鏡的研究. 日本胸部疾患学会雑誌, Vol. 17, No. 6, pp.343-347, 1979.

[14] B. Selvarajah, I. Azuelos, D. Anastasiou, R. C. Chambers, Fibrometabolism-An emerging therapeutic frontier in pulmonary fibrosis. Sci. Signal. 14, eaay1027 (2021).

[15] [上田太郎. “アクチン”. 生物物理について. (n.d.). https://www.biophys.jp/highschool/A-04.html, (参照 2023-12-20).

Telegramにおけるグループ名と投稿メッセージの分析によるグループ探索

伊藤　純菜† 　趙　智賢†† 　長田　繁幸†† 　中川　直樹†† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2 丁目 1 番 1 号
†† 株式会社 日本総合研究所　〒 141–0022 東京都品川区東五反田 2 丁目 18 番 1 号
E-mail: †{g2020507,oguchi}@is.ocha.ac.jp, ††{zhao.zhixian,osada.shigeyuki,nakagawa.naoki}@jri.co.jp

あらまし　近年のデジタル化に伴い，フィッシング攻撃によるクレジットカード（以下，単にカードと呼ぶ.）情報の窃取及び不正利用被害は拡大している．窃取されたカード情報は SNS を介して広く売買されており，先行研究によって SNS に投稿されたメッセージを監視することが犯罪抑止方法として有効であることが判明している．監視をより効率的に行うためには監視対象となる犯罪グループを効率的に発見する手法が必要である．本稿では，従来の投稿されたメッセージに含まれる他のグループを示す URL から繋がりのあるグループを発見することで新たにグループを発見する手法に対して，既知の犯罪グループのグループ名から抽出したキーワードをグループ名に対する共起頻度を基準に組み合わせて指定することで新たにグループを発見する手法を提案し，有効性を確認する．

キーワード　グループ探索，Telegram，サイバー犯罪対策，共起頻度

1 はじめに

近年，インターネットやデジタルテクノロジーの普及により，SNS が我々の生活と深く結びつくようになった．そのようななか，新たな方法で犯罪を行う機会が増えたことでサイバー空間上での犯罪が増加している．

一般社団法人日本クレジット協会 [1] によると，クレジットカード不正利用被害は年々増加しており，2022 年通年の不正利用被害額は 436.7 億円にも上る．これは統計を取り始めた平成 9 年以降，過去最悪である．また，クレジットカード不正利用被害額の内訳は 3 種類に分けられるが，そのうち番号盗用被害額は 411.7 億円であり，全体の約 9.5 割を占めている．このことから，番号盗用被害額の被害額に占める構成比は増大傾向にあると言える．

番号盗用の手口の 1 つとしてフィッシング攻撃が挙げられる．フィッシング攻撃とは，偽のウェブサイトや電子メールを使用してカード情報などの個人情報を窃取する攻撃である．フィッシング対策協議会 [2] によると，フィッシング情報の届出件数について，2022 年は前年と比較して著しく増加している．また，2023 年上半期のフィッシング報告件数は 53 万 804 件（前年同期比で 8 万 722 件増加）で，右肩上がりで増加している．そこでこれらのサイバー犯罪への対策を講じるため，フィッシング攻撃に対しては様々な対策が検討されてきた．

趙ら [3] は，投稿メッセージの収集を行い，それらを分析し，行為別に犯罪者のグループ分けを行った．その結果として，カード情報の窃取と盗用のプロセスについてのモデルが得られた．この犯罪モデルによると，カードの不正利用は，カード情報の窃取から現金化まで大きく 3 つに分業化されており，それぞれの行為は別の人物あるいはグループによって行われている．また，窃取されたカード情報は SNS 上で売買されており，その際カード情報の売り手は買い手からの信用度を上げるために，カード情報の一部をサンプルとして投稿する傾向にあることを明らかにしている．趙らは，この点に着目し番号盗用の抑止方法として投稿メッセージのモニタリングを提案し，その有効性を確認している．しかし，このモニタリングツールには効率の面で改善の余地がある．そこで，本稿ではグループ名と投稿メッセージを分析することによるグループ発見手法を提案する．

本稿の構成は以下の通りである．第 2 章では，先行研究と先行研究で観察対象に選択されている SNS である Telegram についての概要を紹介する．第 3 章では，モニタリングの対象となる犯罪グループを効率的に発見する手法を提案し，第 4 章では，提案手法の有効性を確認するための実験の設計を述べる．第 5 章で実験結果とそれに対しての考察を行い，第 6 章でまとめる．

2 先行研究

2.1 Telegram

趙らの研究では，カード情報の売買が行われている SNS として Telegram（テレグラム）[4] が観察対象に選択されている．Telegram はテキスト，写真，ビデオなどのメッセージの送信や，音声電話，ビデオ電話などができるメッセンジャーアプリである．月間アクティブユーザ数は 7 億人を超え，世界で最もダウンロードされた 10 のアプリの 1 つである [5]．Telegram の特徴は，高度な暗号化機能によって，投稿内容を盗み見ることができないようになっていることや，投稿メッセージが自動削除される機能によって，画面上だけでなく，運営会社にも内容が残らないようにできることなど秘匿性の高さにある．このように，セキュリティ性能が高い反面，その秘匿性の高さから犯罪に利用されやすいという一面もある．加えて Telegram に

は，最大 20 万人が参加することのできるチャットルームを作成することができる機能が存在し，この機能を使用したカード情報の売買事例が観察されている．このようなチャット機能を利用して，カード情報の売買を行う集団を犯罪グループと呼ぶことにする．

2.2 カード情報の窃取と盗用の手口のモデル化

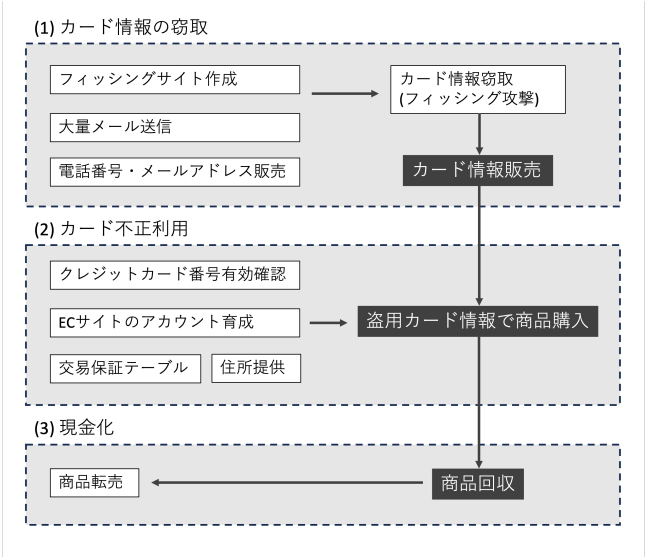


図 1　カード情報の窃取と盗用の一連の流れ

趙らはカード情報の窃取と盗用の全体の流れについて図 1 のようにモデル化をしている．

（1）まず，カード情報窃取の事前準備として，誘導先のフィッシングサイトの作成や，誘導するための大量のメールの送信，それによって得られた電話番号やメールアドレスなどの個人情報の販売を行う．フィッシングサイトや売買によって入手したカード情報は主に SNS を介して販売される．

（2）次に，盗用の手口として，EC サイトを利用して換金性の高い物品の購入を試みる．近年では，多くの EC サイトではこのような行為への対策として，新規作成された直後のアカウントが高額で換金性の高い物品を購入しようとすると拒否する場合もあるため，アカウントの育成や販売が行われている．また，ユーザ間で共有される Web サイトを使用しての入手したカード情報が実在するものであるかの確認や，リスクを低減するために，第三者が売買を仲介することによる取引の保証，番号盗用で購入した物品を受け取るときに使用する住所の提供などの行為も存在する．

（3）最後に購入した物品の現金化が行われる．これは提供された住所に届いた物品の回収と，回収した物品の転売という行為に分類され，留学生やアルバイトを雇って代行させるケースが多い．

このように，番号盗用によるクレジットカードの不正利用には犯罪エコシステムが存在している．一般的な商取引では法律に基づく契約により確実に行われることが保証されている一方で，SNS を介しての番号盗用を狙った取引では違法性が高いために法的な保証を持つことができない．そのためこの犯罪エコシステムは弱い信頼関係の上に成立している．

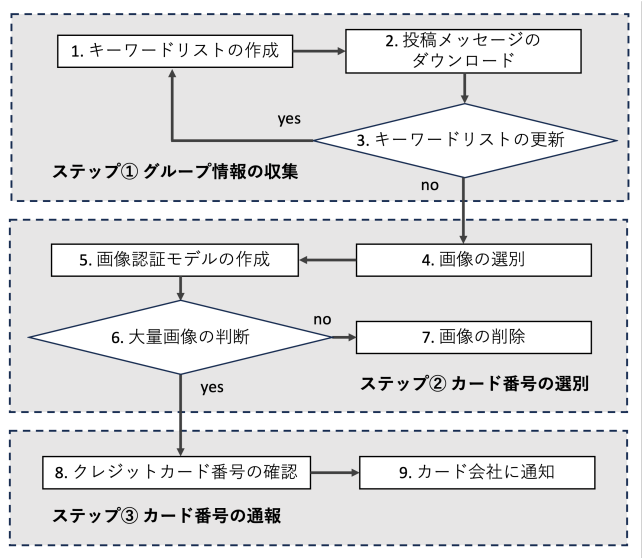


図 2　モニタリングツールの設計

2.3 モニタリングツールの設計とその効果

趙らは，番号盗用の抑止を目的に，前述した弱い信頼関係をさらに弱化させる方法を提案している．SNS での取引では，売り手と買い手は互いに身分を隠しているため，信頼関係を結びにくい．そのため，売り手は買い手からの信用度を上げるためにカード情報の一部をサンプルとして投稿する傾向にある．この情報をクレジットカード会社が入手できれば，カード会社はそのカードの利用を停止する等の対応をとることができる．この対応は番号盗用被害を防ぐことに加えて，買い手から見た売り手への信用度を低下させることができるため，買い手のカード情報入手に関する行動の妨害にも繋がる．このように SNS を介して行われる番号盗用の犯罪エコシステムを弱化させることができ，その結果として番号盗用の抑止に繋がると言える．

上述の抑止方法の有効性を確認するために，図 2 のようなモニタリングツールを実装し，Telegram の複数のグループに対して実際にモニタリングを実施している．モニタリングツールは大きく分けて 3 つのステップに分けられる．

（1）ステップ①では，カード情報の売買を行なっているグループの情報を収集する．そのために，まず，経験的に得られている特徴的な単語を設定し，検索キーワードリストを作成する．次に，作成した検索キーワードリストを用いて，Telegram が提供する API である Telethon[6] でグループの検索を行い，カード情報の売買を行なっているグループを特定する．このとき，特定したグループの投稿メッセージをダウンロードし，ここで得られたテキスト情報から，番号盗用に関係する単語を特定し，検索キーワードリストを更新する．

（2）ステップ②では，ダウンロードした投稿メッセージから得た画像ファイルの中で，カード番号が含まれるファイルを判別する．ここでは，機械学習を使って，カード番号を含む画像か否かを判別するモデルを作成し，このモデルを活用する．カード番号を含まない画像については以降の処理をせず，単に削除する．

（3）ステップ③では，ステップ②で含まれると判断された画像ファイルを目視で確認し，関係するカード会社へ連絡を行う．

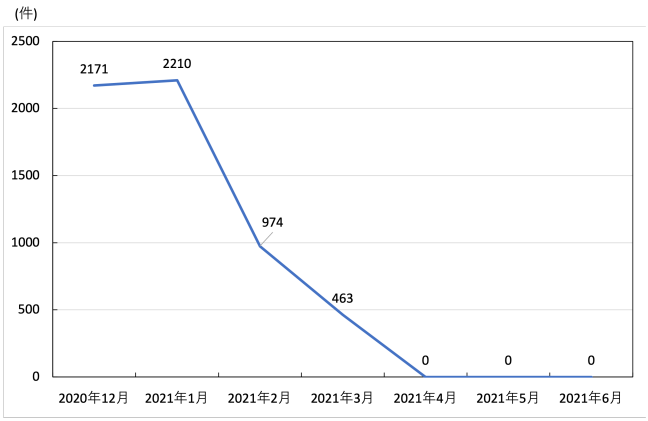


図 3　投稿されたカード番号の数

図 3 は実験対象としたグループの 1 つについての投稿されたカード番号の数の推移の様子である．このグループは元々1ヶ月間に 2171 件のカード番号が投稿されていたが，4ヶ月にわたって対策を継続したところカード番号の投稿がなくなった．他の 2 つのグループに対しても，数ヶ月でカード番号の投稿が 0 になったことが確認されたことから，モニタリングには抑止方法として一定の効果が認められている．

2.4　投稿メッセージによるグループ発見

Kloo ら [7] は，2022 年のロシア・ウクライナ情勢の悪化の中で，Telegram 上で行われた情報活動を分析している．分析にあたっては，Kloo らは，Telegram でグループ検索する前に，X（旧 Twitter）に投稿されたメッセージの中で，Telegram に関係する URL を含むものを検索する．Telegram では，投稿の内容で検索を行うことができないため，内容で検索することができる X で最初のグループを得るようにしている．この検索結果を起点に，Telegram でスノーボールサンプリングを行う．ここで，スノーボールサンプリングとは，ある既知の要素から始め，その要素と直接繋がっている新たな要素を収集する方法のことであり，このグループ探索においては，既知のグループから言及のあったグループを新規のグループとして収集している．このようにして，既知のグループから未知のグループを発見する．

3　提　　案

3.1　従来法の課題と方針

前節で述べた投稿メッセージを監視する方法の有効性を向上させるには，監視対象となるカード情報の売買に関係のあるグループを効率的に発見する手法が必要である．監視対象とするグループが多くなればなるほど監視することができる投稿メッセージ数も増やすことができ，その結果としてカード情報の売買を抑止できる機会も増加するためである．Telegram ではグループを探索するとき，キーワード検索のみが可能であり，また Telegram のキーワード検索では，ある単語を指定して検索を行った時，その単語が含まれるユーザやグループを検出することはできるが，その単語が含まれるメッセージを検出することはできない．そのため，今まではカード情報の売買を試みるグループがそのグループ名をつける時に頻繁に使用するキーワードとして既に得られている“CVV”という単語と数字やアルファベットを組み合わせて検索を行ってきた．“CVV”とはカードの裏のセキュリティコードを示す単語である．しかしこの手法には改良の余地がある．例えば，キーワード検索に使用する単語間の関連度を把握することで，探索の精度を向上させられる可能性があると考える．また，Kloo らの投稿メッセージを分析することでグループを発見する手法によって，キーワード検索で得られるグループとは違う新たなグループの発見に繋がる可能性もある．そこで本研究では，投稿メッセージを分析することでグループを発見する手法とグループ名を分析することでグループを発見する手法を組み合わせる．

3.2　提案手法の処理フロー

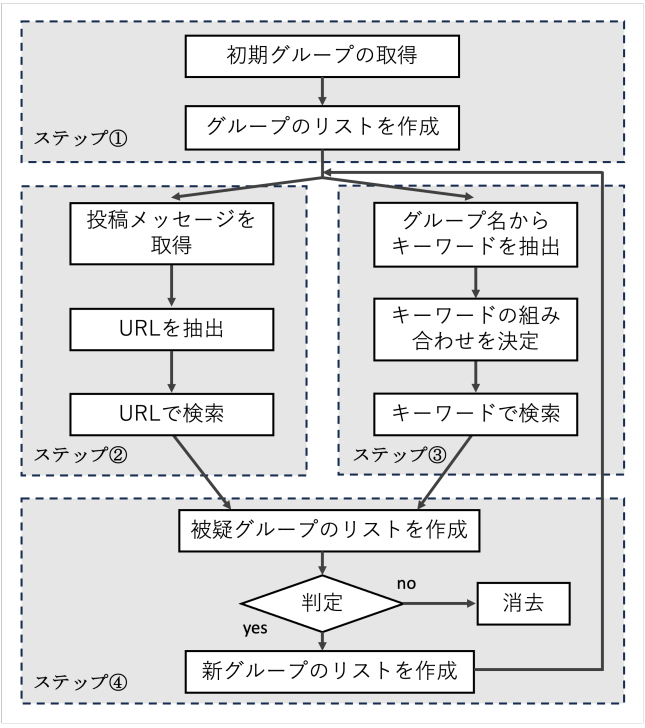


図 4　処理フロー

図 4 に示すように，提案手法は，投稿メッセージの分析によるグループ探索とグループ名の分析によるグループ探索との 2 つを組み合わせた処理フローで構成する．

（1）まずステップ①では，初期グループの取得を行う．初期グループの取得は，Kloo らの手法と同様に，X のキーワード検索機能を使用して，Telegram のグループの URL を含むポストから取得する．特殊詐欺や窃盗等の犯罪行為の実行役の募集は，発見されやすい一般的な SNS で募集され，応募が完了すると匿名性の高い SNS をインストールさせ，以降のやりとりは匿名性の高い SNS で行うことが多い [8]．同様に，X のような多くの人の目に触れるような SNS では，カード情報を載せて直接売買を行うような投稿を行うことはないが，実際のやりとりを行う誘導先である Telegram の URL を載せた宣伝目的での投稿は存在する．このような投稿から誘導先のグループを取得することになる．並行して，趙らが行なっている，Telegram のキーワード検索機能を使用して，既知の単語を指定した検索結果からも取得する．取得したグループからグループのリストを作成する．このリストにはグループ名，グループ ID，所属人数，グループの URL（例えば，https://t.me/GroupName）などのグループ情報が含まれていることとする．

（2）ステップ②では，投稿メッセージの分析による探索を行う．初めに，既知のグループの投稿メッセージを取得する．続いて，取得した投稿メッセージに含まれるテキスト情報から Telegram のグループの URL を抽出し，抽出した URL が示すグループの名前や所属人数などのグループ情報を取得する．

（3）ステップ③では，グループ名の分析による探索を行う．まず，既知のグループ名に対して形態素解析を行う．次に，辞書を用いてストップワードを除去し，キーワードとなり得る単語とその出現回数 f_a を取得する．Telegram では，3 文字以下のキーワードでは検索することができず，得られる単語は 3 文字以下であることが多いため，検索の際には 2 つの単語を組み合わせることで検索キーワードとする．続いて，出現回数があらかじめ定めた閾値を超えた単語について，グループ名に対する共起頻度 f_c を計算する．ここで，グループ名に対する共起頻度とは，グループ名の中で 2 つの単語が同時に出現する回数を計測したものとする．例えば，4 つのグループ名 N_1 から N_4 がそれぞれ図 5 のようにいくつかの単語で構成されていたとする．この時，W_A と W_B の組み合わせは、N_1 のグループ名では 2 回出現しており、N_2 のグループ名では 1 回出現している．また，N_1 のグループ名のときのように，W_A と W_B 間の距離に関係なくそれぞれ回数を数えることとする．N_3 や N_4 のグループ名ではそれぞれ W_A のみや、W_B のみでは出現しているが、W_A と W_B が同時に出現していないため回数は数えない．したがってグループ名の中で同時に出現している回数は合計 3 回であり、グループ名 N_1 から N_4 に対する W_A と W_B の組み合わせの共起頻度は 3 となる．

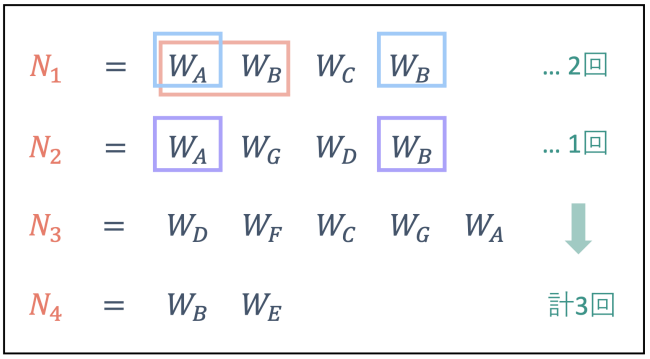


図 5　共起頻度計測の例
グループ名 N_1〜N_4 に対する W_A と W_B の組み合わせの共起頻度 $f_c(W_A, W_B) = 3$

共起頻度があらかじめ定めた閾値を超えた単語の組を検索キーワードとして，Telegram のキーワード検索機能に指定し，検索を行う．Telegram では単語間の空白の有無で検索結果が異なるため，同じ組み合わせでも空白の有無を考慮する．

（4）ステップ④では，ステップ②の URL を使った検索結果と，ステップ③のキーワードを使った検索結果で得られた両方のグループを，番号盗用の疑いがある被疑グループとしてリスト化する．その後，有識者によって関係性の有無を目視で確認することで判定し，番号盗用に関係のあるグループをリストに追加する．
ステップ④終了後，得られたリストを用いて，ステップ②〜ステップ④を繰り返す．

4　実　　験

X のキーワード検索機能を使用して取得したグループ 6 個と Telegram のキーワード検索機能を使用して取得したグループ 89 個の計 95 個のグループに対して提案手法を実行する．ただし，この 95 個のグループは有識者によって番号盗用に関係のあるグループだとあらかじめ判定されたものである．提案手法の有効性を確認するために，以下の 3 つの点について確認を行う．

- Telegram のキーワード検索機能に指定する単語の組み合わせを共起頻度によって決定することで検索の精度を向上させることができるのか
- 投稿メッセージの分析による探索によって取得したグループとグループ名の分析による探索によって取得したグループとにはどの程度違いがあるのか
- 新たに発見したグループから新規のキーワードを得ることができるのか

4.1　共起頻度の閾値ごとの比較

提案する手法では，共起頻度があらかじめ定めた閾値以上の単語の組み合わせのみキーワード検索機能に指定して検索することとしている．そこで，共起頻度の閾値を 0 から 3 まで変更して検索し，検索結果を比較する．閾値が 0 のときとは，共起

頻度に関係なく，抽出した単語の全ての組み合わせで検索を行う場合を指す．ただし，グループ名から単語を抽出する際，出現回数が多い方から 30 番目の単語の出現回数を閾値として閾値以上の出現回数の単語を抽出することとする．

4.2 ２種類の探索手法で発見できるグループの比較

提案する手法では，投稿メッセージの分析による探索とグループ名の分析による探索を組み合わせることとしている．もし，この 2 種類の探索によって発見できるグループに差がなく共通しているものが多いのであれば，2 種類の探索を組み合わせる意味が失われる．そこで，同じグループに対して，投稿メッセージの分析による探索を行なった結果とグループ名の分析による探索を行った結果を比較する．ただし，投稿メッセージの分析による探索のとき，3ヶ月分の投稿メッセージが 10000 件を超えない場合は 3ヶ月分取得し，10000 件を超える場合は最新の方から 10000 件取得することとする．

4.3 グループ名から抽出したキーワードの比較

提案する手法では，初期グループから新たに発見したグループに対しても同じ処理を繰り返すこととしている．そこで，初期のグループ名を分析することで抽出したキーワードと新たに発見したグループ名を分析することで抽出したキーワードの比較を行う．

5 実験結果

5.1 共起頻度の閾値ごとの比較結果

95 個のグループに対して形態素解析を行い，ストップワードを除去した結果，172 個の単語が得られた．これらの単語の出現回数を計測し，出現回数の多い方から順に並べたところ，30 番目の単語の出現回数は 3 であったため，出現回数の閾値を 3 とした．閾値以上の単語 35 個の全ての組み合わせ 595 個に対して，共起頻度を計測した．その結果，共起頻度の閾値を 1 とする場合の検索に用いる単語の組み合わせは 171 個，2 とする場合の検索に用いる単語の組み合わせは 68 個，3 とする場合の検索に用いる単語の組み合わせは 37 個であった．ただし，検索のときには空白の有無を考慮するため，単語間に空白を入れる場合と入れない場合とで，それぞれの検索回数は組み合わせの個数の 2 倍の回数となる．

共起頻度の閾値を変え，検索を行った結果を図 6 に示す．横軸は検索に指定する共起頻度の閾値であり，縦軸は発見グループ数である．また，発見グループを番号盗用に関係のあるグループと，番号盗用に関係のないグループの 2 つに分けた．まず，1 回の検索に対する発見グループ数の平均をみると，共起頻度の閾値を 0 とした場合は約 1.5 個，1 とした場合は約 2.4 個，2 とした場合は約 3.3 個，3 とした場合は約 3.8 個であった．このことから，番号盗用に関係するかに関わらず，共起頻度の閾値を上げる方が効率良くグループを発見できている．また，発見グループ数に対する番号盗用に関係のあるグループ数の割合をみると，共起頻度の閾値を 0 とした場合は 58%，1 とした場合は 71%，2 とした場合は 79%，3 とした場合は 81%であった．このことから，共起頻度の閾値を上げるにつれて，有識者対応の効率性を左右する精度が向上していると言える．ただし，閾値を 0 とした方が検索回数が多くなるため，番号盗用に関係のあるグループの発見数自体は閾値を大きくした時よりも多くなっている．そのため，精度と量のバランスを考えて閾値を決定する必要があるが，複数回探索を繰り返すことを考えるとある程度閾値を設定する方が良いと言える．

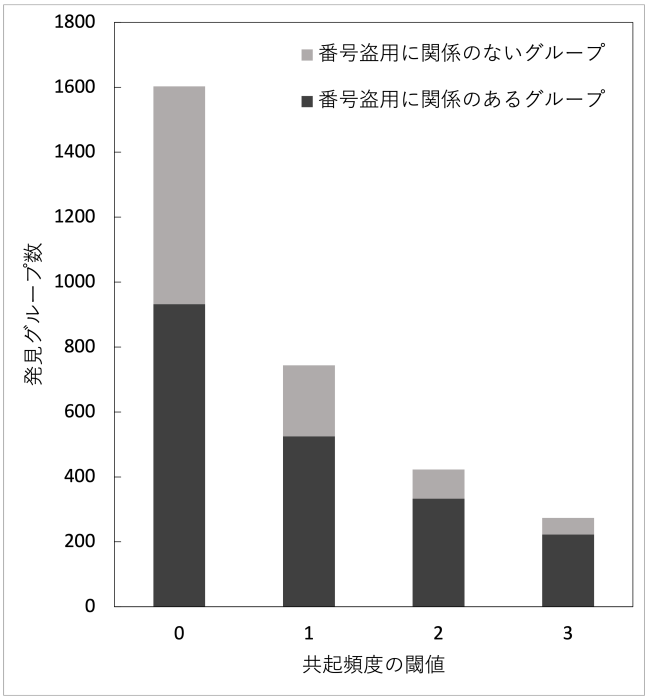


図 6　共起頻度の閾値ごとの検索結果

5.2 ２種類の探索手法で発見できるグループの比較結果

95 個のグループに対して，グループ名の分析による探索を行った結果は前述した通りである．同様のグループに対して，投稿メッセージの分析による探索を行った結果，2830 個のグループを発見できた．そのうち番号盗用に関係のあるグループは 577 個であった．

この 577 個のグループとグループ名の分析による探索で発見したグループとで共通するグループの数を調べまとめた結果を図 7 に示す．横軸はグループ名の分析による探索のときに指定した共起頻度の閾値であり，縦軸は番号盗用に関係があると判定された発見グループ数である．共起頻度の閾値を 0 とし取得したグループ 932 個と共通しているグループは 57 個でありグループ数全体に占める割合は約 3.9%，閾値を 1 とし取得したグループ 525 個と共通しているグループは 50 個でありグループ数全体に占める割合は約 4.8%，閾値を 2 とし取得したグループ 333 個と共通しているグループは 42 個でありグループ数全体に占める割合は約 4.8%，閾値を 3 とし取得したグループ 223 個と共通しているグループは 36 個でありグループ数全体に占める割合は約 4.7%であった．閾値がいずれの場合も共通しているグループの割合は 5%以下となっており，このことから投稿メッセージの分析による探索によって発見できるグループと

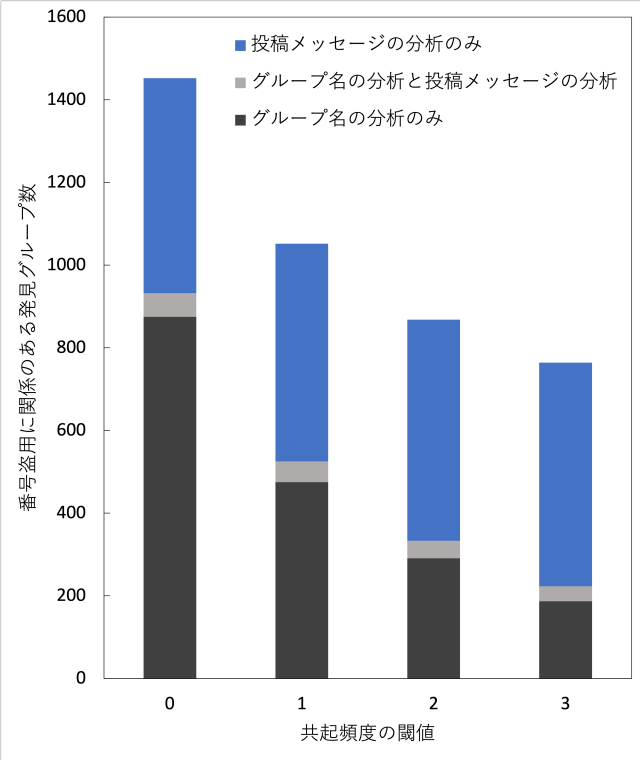


図 7　投稿メッセージの分析による探索とグループ名の分析による探索を合わせた結果

グループ名の分析による探索によって発見できるグループには違いがあることがわかる．よって 2 種類の探索を並列して実行することで，より多くのグループを発見することができると期待できる．

5.3　グループ名から抽出したキーワードの比較結果

投稿メッセージの分析による探索によって取得したグループとグループ名の分析による探索によって取得したグループのグループ名に対して形態素解析を行い，単語の出現回数を計測した．ただし，投稿メッセージの分析による探索によって取得したグループと同数程度取得することができているため，グループ名の分析による探索によって取得したグループとは，共起頻度の閾値を 1 として取得したグループ 525 個とした．次に，初期グループから抽出した単語数が 35 個であったため，それに合わせて，出現回数が多い方から 35 番目までの単語を抽出した．抽出した単語間の共起頻度を計測し，共起頻度を重みとしてエッジを張り，グラフを作成した．同様に，初期のグループ名からもグラフを作成し，このグラフと共通するエッジをそれぞれ削除した．その結果として作成されたグラフが図 8 と図 9 である．図 8 は投稿メッセージの分析による探索で取得したグループ名から作成したグラフのうち，初期のグループ名から作成したグラフと共通するエッジを削除したグラフであり，図 9 はグループ名の分析による探索で取得したグループ名から作成したグラフのうち，初期のグループ名から作成したグラフと共通するエッジを削除したグラフである．ノードはグループ名から抽出した単語であり，エッジは太く，青の濃度が濃くなっているほど重みが大きい，すなわち単語間の共起頻度が大きいこと示している．図 8 と図 9 を見ると，どちらのグラフにもある程度の重みを持つエッジがあり，初期のグループ名から取得した検索ワードとは別の検索ワードを取得できていることがわかる．また，新しく発見したグループから得られた検索ワードの中には，初期グループの時点で組み合わせとしては存在していたが，閾値を超えることができなかった組み合わせも見られた．このことから，閾値を設けることで発見することができなくなってしまったグループも，2 周目以降で得られる見込みがあると言える．例えば，今回の結果で言うと，共起頻度の閾値を 1 とした場合と閾値を設けない場合とでは，番号盗用に関係のある発見グループ数の差は 407 個であったが，この 407 個のグループを発見するために検索に用いられた単語の組み合わせは，新たに発見するグループが増えるにつれ，いずれかのグループ名で共起する可能性が高いと思われる．

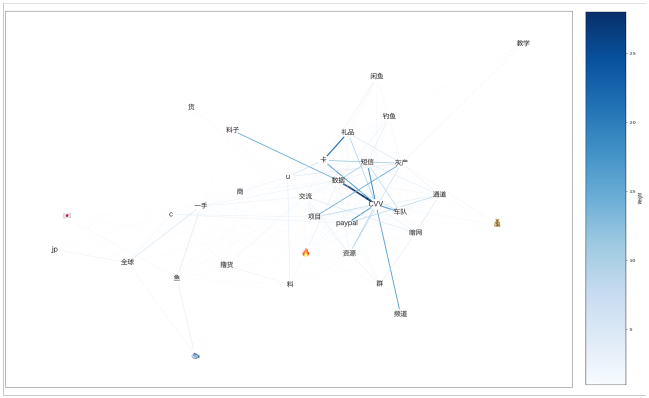


図 8　投稿メッセージの分析による探索で取得したグループ名から作成したグラフ

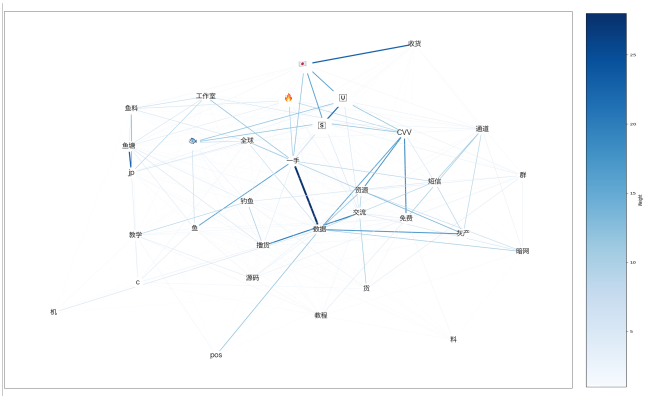


図 9　グループ名の分析による探索で取得したグループ名から作成したグラフ

6　おわりに

本研究では，Telegram のグループを効率的に発見する手法について，投稿メッセージの分析による手法とグループ名の分析による手法を組み合わせた手法を提案した．実験では，共起頻度の閾値を変えて検索を行った結果の比較を行った．その結果，共起頻度の閾値をあらかじめ定めておくことで精度が向上

することが確認できた．さらに，投稿メッセージの分析による手法によって発見できるグループとグループ名の分析による手法で発見できるグループとの比較を行ったところ，発見できるグループには違いがあり，並行して実行することの有効性も確認できた．また，新たに発見したグループからも初期グループから取得した検索ワードとは別の検索ワードを得ることができていることも確認できた．今後の課題としては，現在では取得したグループが番号盗用に関係のあるグループであるかを有識者に判定してもらっているので，自動で判定するプログラムを実装し，判定部分での効率化を図ることを考えている．

文　　献

[1] 一般社団法人日本クレジット協会，クレジットカード不正利用被害額の発生状況，https://www.j-credit.or.jp/information/statistics/download/statistics_domestic_2022.pdf （2024 年 1 月参照）
[2] フィッシング対策協議会，フィッシング報告状況（月次報告書），https://www.antiphishing.jp/report/monthly/ （2023 年 1 月参照）
[3] 趙 智賢，長田 繁幸，SNS を経由するクレジットカード不正利用のモデル化と抑止方法の検討，信学技報，セキュリティ心理学とトラスト（SPT），2022-SPT-48，No.25，pp.1-7，2022.
[4] Telegram, https://telegram.org/
[5] Telegram FAQ, https://telegram.org/faq#q-what-is-telegram-what-do-i-do-here
[6] Telethon, https://tl.telethon.dev
[7] lan Kloo, kathleen M.Carley, “Social Cybersecurity Analysis of the Telegram Information Environment During the 2022 Invasion of Ukraine,” Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2023. Lecture Notes in Computer Science, vol 14161. Springer, Cham, 2023.
[8] 警察庁，犯罪実行者募集の実態，https://www.npa.go.jp/bureau/safetylife/yamibaito/yamibaitojirei.pdf （2024 年 2 月参照）

子宮収縮圧データのための潜在ステージ進行推定

田村　優香[†,††]　松原　靖子[†]　木村　輔[†]　町村　栄聡[†††]　遠藤　誠之[†††]
木村　正[†††]　櫻井　保志[†]

† 大阪大学産業科学研究所　〒 567–0047 大阪府茨木市
†† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市
††† 大阪大学大学院医学系研究科　〒 565–0871 大阪府吹田市

あらまし　本論文では、Uterine contraction (UC) 信号などの特徴的な波形を持つ，大規模な医療生体データの新しい解析モデルである UCSCAN を提案する．この手法には，信号の時系列的な文脈を捉え効果的に処理するための波形検出手法 WAVEDETECTION とフィッティングアルゴリズム UC-FIT を含む．提案手法は次のような特徴を持つ．(a) 各シーケンスのデータダイナミクスを「ステージ」の進行として捉え，効果的にモデル化する．(b) 与えられたすべてのシーケンスから共通のパターンを解釈可能な「コンポーネント」として集約して抽出する．(c) WAVEDETECTION は指定されたデータを要約して表現し、UC-FIT は入力サイズに対して線形である．実際の UC 信号に関する広範な実験により、UCSCAN がデータのダイナミクスと集約的なパターンを捉えることが実証された．さらに，最新の既存手法と比較を行い提案手法が大幅な精度，性能向上を達成していることを明らかにした．

キーワード　パターン発見，セグメンテーション，時系列クラスタリング

1　はじめに

分娩は母親と赤ちゃんの命が危険にさらされる出産における最も注意が必要な過程の一つである．例えば, 産科出血や早産は, 母体および乳児の死亡リスクを増加させる．出産後の大量の産科出血（PPH）は, 世界的に母体死亡の主要な原因である．毎年, 約 1400 万人の女性が PPH に苦しんでおり [1, 2], 世界中で約 7 万人の母体死亡を引き起こしている．早産とその関連する合併症は, すべての周産期死亡の約 85 %を占める, 最も重要な世界的な健康課題の一つである．早産児はすべての出生の約 10 %を占め,2020 年には 1340 万人の赤ちゃんが早産で生まれたと推定されている．出産に関連するこれらのリスクを軽減することは, 人類の安定した生存のために不可欠である．安全な出産を実現するために, 母子の生理状態を電子的にモニタリングし, 分娩のプロセスを観察することは, 母子の命をリスクから保護する方法の一つである．

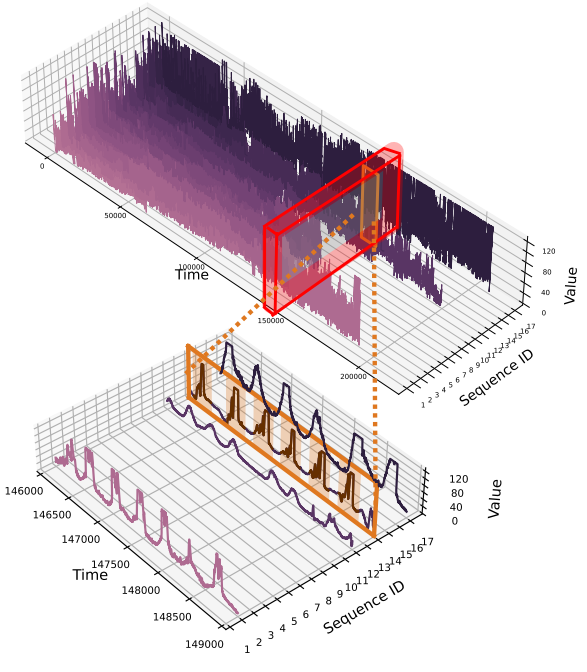
分娩のプロセスは, いくつかの段階で表現される．フリードマンは 500 例の妊婦を対象に研究を行い, 子宮口の開大の度合いと時間の経過との関係を示すフリードマン曲線を開発し，この曲線に基づいて分娩プロセスを数段階に分類した [3]．分娩の正常な進行からの逸脱は, 内診によって子宮口の開大を測定し，この曲線と比較することで判断される．したがって, 妊婦が分娩のどの段階にあるかを判断することは重要である．一方で, 内診の回数が増えるにつれて, 胎児感染のリスクが増加することが知られており，特に破水後はそのリスクが高まる [4]．そこで，内診の必要なく分娩の進行をモニタリングするため, 本研究では出産時の子宮収縮によって加わる圧力を記録する UC 信号に焦点を当てる．UC 信号は分娩時における胎児の健康状態をモニタリングするために胎児心拍と合わせて記録される信号であり， 陣痛時の子宮収縮が山形に現れる．図 1（a）は実際の UC 信号を示している．UC 信号には，陣痛の持続時間や陣痛の発生間隔の変化に分娩の進行が現れ，分娩が進行するごとに陣痛の持続時間は長く，陣痛の発生間隔は短くなる．本研究では，このような UC 信号から複数の分娩事例に共通する特徴的なパターンを抽出し, 分娩の進行を自動的に推定することを目的とする．具体的には, 複数時系列から集約されたパターン「コンポーネント」をパターンに関する事前知識なしに抽出し，各時系列において分娩の進行を表すパターン「ステージ」を推定することを目指す．

具体例　図 1 (a) は, オリジナルのシーケンスと,Sequence ID #14 に対する提案アルゴリズム WAVEDETECTION の結果を示している．データには, 長さが異なる長いシーケンスが含まれており, オレンジ色で着色された部分は,WAVEDETECTION が検出した陣痛部分である．

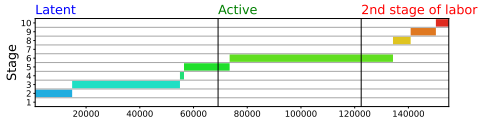
図 1 (b) は, 各データポイントに割り当てられたステージラベル (1, · · · , 10) を示している．出産プロセスは, 潜在期, 活動期, および分娩第二期に分けることができ，黒い垂直線は, それらの境界を示している．ステージの割り当ては UC 信号のダイナミクスをステージ遷移として捉えたものであり, 境界でのステージ遷移は UC 信号がこれらの 3 つの段階の変化を反映していることを示唆している．

図 1 (c) は, 各ステージに割り当てられたオリジナルのシーケンスの一部 (約 4 分間分) と, そのステージにおけるコンポーネントを示している．着色された部分は提案手法により陣痛波形として検出された部分であり, 色は割り当てられたステージに対応している．UCSCAN は， 分娩の進行による陣痛波形の変化を捉えており， その特徴は各ステージにおける コンポーネントに反映されている．

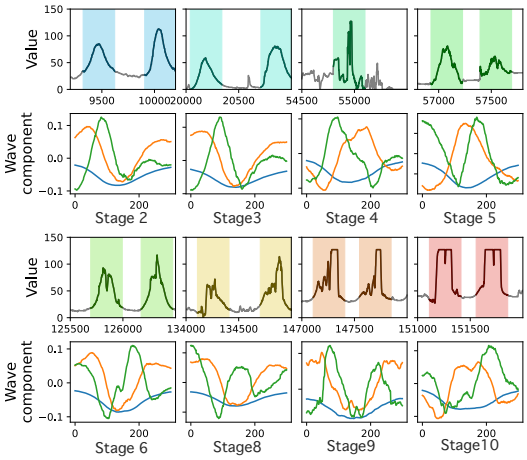
本論文の貢献　本論文では UC 信号のための解析モデル UCSCAN



(a) オリジナルシーケンス (上段) and 提案による出力 (下段)



(b) ID #14 のデータに対するステージ割り当て



(c) 各ステージにおける ID #14 のオリジナルデータとコンポーネント

図 1: UCScan を OUH データセットに適用した結果の例

を提案する．これには, 陣痛部分の波形を自動的に検出するアルゴリズム WaveDetection とフィッティングアルゴリズム UC-fit が含まれる．主な貢献は以下の通りである．

- 特徴的な波形を持つ信号を効果的に処理するための新しいデータ観測スキームを導入する．さらに,UCScan は各シーケンスのダイナミクスをステージの遷移として捉える
- UCScan は, シーケンス全体で集約されたパターンをコンポーネントとして抽出する．
- WaveDetection は与えられたデータを要約された形で表現し，UC-fit は入力サイズに対して線形に動作する．

2 関 連 研 究

医療およびヘルスケア分野における時系列データ解析は, 非常に注目されている重要なトピックである [5–10]．特に，女性に特有のヘルスケア解析として，生理や妊娠に関する研究も行われている [11–13]．これらはモバイルアプリケーションによって追跡されたデータを使用した, 月経周期や妊娠の予測に関するものである．しかし, 分娩中に測定される UC 信号のようなセンサーデータの時系列解析に関する研究は存在しない．

本研究で使用される統計モデルは, 隠れマルコフモデル（HMM）に関連しており, 多くのデータ解析手法が HMM に基づいている [14–16]．HMM は, データのダイナミクスが潜在的な状態遷移によって生じると仮定している．しかし,HMM は潜在的な状態間の任意の遷移が可能であると仮定する一方で, 私たちは遷移に特定の構造を強制している． [17] は, 同様に特定の構造を強制する戦略を使用する方法である．ただし, この方法は離散的なイベントシーケンスを対象としており,UC 信号の波形的なパターンを捉えることはできない．

近年, 時系列データのパターン発見に関する関心が高まっており, 多くの研究が行われている [18–21]．TICC [22] は, 各時系列データポイントをその直前のデータとの文脈で扱うことを目指しており, 提案手法と同様に時系列を短いサブシーケンス単位で扱い，これをクラスタリングする．このサブシーケンスは時系列的文脈を捉えるための有用な手段であり, 実世界のデータに対して優れた性能を発揮する．ただし,TICC はクラスタが共に発展する複数時系列の相関ネットワークによって定義されるため, 単一次元データに適用することはできない．

3 提案モデル

本章では, 提案モデルについて述べる．

表 1 に本論文で使用する記号と定義を示す．本研究では，複数の UC 信号のセット $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{I}$ が与えられたとき，それぞれのシーケンスにおいて時間発展するパターンと，シーケンス間に共通するパターンを抽出することを目指す．ここで $\mathbf{x}_i$ は i 番目のシーケンスであり, 時刻 t での $\mathbf{x}_i$ の信号値を $x_{i,t}$ で表す（$t = 1, 2, \cdots, T_i$）．各レコードの開始時刻は異なるため, シーケンス $\mathbf{x}_i$ の時刻 t は, 別のシーケンス $\mathbf{x}_{i'}$ の時刻 t とは異なる．図 1 (a) は与えられた信号 $\mathcal{X}$ （上段）と信号の拡大図（下段）を示しており,“Sequence ID” はシーケンスインデックス i に,“Time” は時間点 t に対応する．図 1 が示すように, 各シーケンスの長さは異なる．これは，それぞれ分娩の所要時間やセンサーの取り付け時刻が異なるためである．提案手法ではこのような長さが不規則な時系列を同時に処理する．

知見 ここでは,UC 信号の観察とドメイン知識から得られる, 提案モデルを設計するための知見を述べる．最初に, 信号の構造に関する知見を述べる．信号は, 山形の波形部分とその間を繋

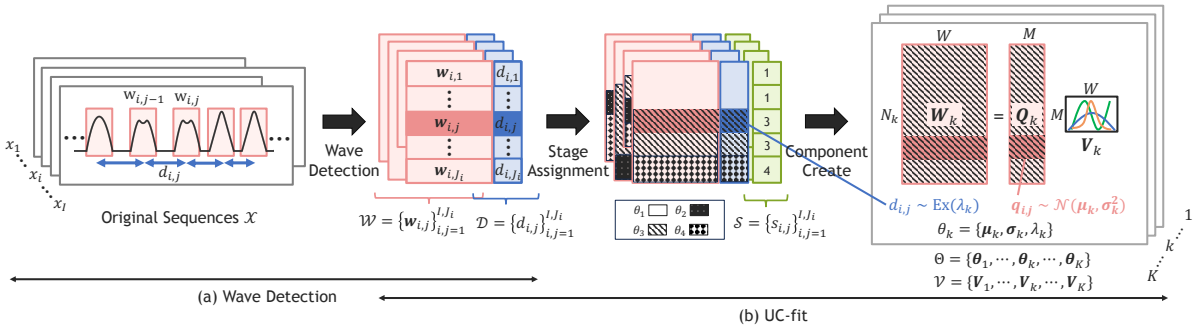


図 2: 提案モデルの概要

ぐ平らな部分との繰り返し構造からなる．山形の波形部分は陣痛が発生していることを意味する重要な部分である．

［知見 1］（UC 信号の周期的構造） 信号では陣痛時の情報を含む山形の波形部分とその間隔が交互に繰り返される．

次に, 陣痛波形の形状に焦点を当てる．図 1 (a) 下図を見ると, 陣痛波形にはいくつかの形状の種類があり, 異なるシーケンスがパターンを共有してることがわかる．さらに, 図 1 (c) は同一シーケンス内で時間発展に伴いその形状の変化を示している．すなわち，データには二つの種類のパターンがある．

［知見 2］（陣痛波形の共通パターン） 陣痛波形の形状には, 異なるシーケンス間で共有されるパターンがある．

［知見 3］（時間経過に伴うパターン） 陣痛波形のパターンは時間とともに変化する．

最後に, ドメイン知識から得られる重要な知見を述べる．フリードマンは，子宮口開大と時間の関係を示すフリードマン曲線を開発した [3]．フリードマン曲線はシグモイド曲線状に単調増加する．すなわち子宮口開大は時間と共に単調に増加し，分娩の進行は逆戻りしない．

［知見 4］（分娩進行の不可逆性） 分娩の進行は不可逆的である．

まとめ 本論文では, これらの以下のすべての特性を取り入れた新しいモデル UCscan を提案する．

(P1) 信号の周期構造
(P2) シーケンス間に共通するパターン
(P3) シーケンス内で時間変化するパターン
(P4) 進行の不可逆性

我々の目標は, 時間変化するパターン（ステージ）を抽出し，シーケンス全体で共通のパターン（コンポーネント）を集約することである．以下では，ステージとコンポーネントを効果的に捉えるためのモデルの詳細に述べる．

3.1 UCscan モデル

生体データは, 個々の値というよりも, 時系列的な文脈に重要な特徴を持つことがある．例えば心電図 （ECG: Electrocardiogram） を使用して患者の容態を確認する場合，単に現在の値ではなく波形に焦点を当てて患者の容態を確認する．そこで, データの連続的な特徴を捉えるために, 本論文では個々のデータ点をそのまま扱うのではなく, 新しい観測方法を導入する．この方法では, サイズ $W \ll |\mathbf{x}_i|$ の陣痛波形を表すサブシーケンス

表 1: 記号と定義.

記号	定義
$\mathbf{x}_i$	i 番目のオリジナルシーケンス *i.e.*, $\mathbf{x}_i = [x_{i,1}, \cdots, x_{i,t}, \cdots, x_{i,T_i}]$
$\mathcal{X}$	オリジナルシーケンスの集合 *i.e.*, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{I}$
W	波形のサイズ
M	コンポーネントの数
K	ステージの数
$\mathbf{w}_{i,j}$	$\mathbf{x}_i$ から抽出された j 番目の波形
$d_{i,j}$	波形 $\mathbf{w}_{i,j-1}$ と $\mathbf{w}_{i,j}$ との間の間隔
$s_{i,j}$	波形 $\mathbf{w}_{i,j}$ と間隔 $d_{i,j}$ に割り当てられるステージラベル
$\boldsymbol{q}_{i,j}$	$\mathbf{w}_{i,j}$ から変換された M 次元重みベクトル
N_k	ステージ k に割り当てられた波形の数
$\mathbf{W}_k$	ステージ k に割り当てられた波形による行列 $(N_k \times W)$
$\mathbf{V}_k$	ステージ k のコンポーネント行列 $(M \times W)$
J_i	シーケンス $\mathbf{x}_i$ から抽出された波形の数
$\mathcal{W}$	全ての波形の集合 *i.e.*, $\mathcal{W} = \{\mathbf{w}_{i,j}\}_{i,j=1}^{I,J_i}$
$\mathcal{D}$	全ての間隔の集合 *i.e.*, $\mathcal{D} = \{d_{i,j}\}_{i,j=1}^{I,J_i}$
$\mathcal{Q}$	全ての重みベクトルの集合 *i.e.*, $\mathcal{Q} = \{\boldsymbol{q}_{i,j}\}_{i,j=1}^{I,J_i}$
θ_k	ステージ k におけるパラメータ集合 *i.e.*, $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, \lambda_k\}$
$\boldsymbol{\Theta}$	全ステージパラメータの集合 *i.e.*, $\boldsymbol{\Theta} = \{\theta_k\}_{k=1}^{K}$
$\mathcal{S}$	ステージ割り当ての集合 *i.e.*, $\mathcal{S} = \{s_{i,j}\}_{i,j=1}^{I,J_i}$
$\mathcal{V}$	全ステージのコンポーネントの集合 *i.e.*, $\mathcal{V} = \{\mathbf{V}_k\}_{k=1}^{K}$

を抽出し,$\mathbf{x}_i$ の j 番目の部分系列を W 次元の行ベクトル $\mathbf{w}_{i,j}$ で表す（オリジナルシーケンスから適切なピーク部の部分系列 $\mathbf{w}_{i,j}$ を抜き出す方法については，4.1 節で説明する）．提案モデルではこのサブシーケンスを波形と呼び，データ要素として使用する．

オリジナルシーケンスから抽出されたサブシーケンスだけを使用する場合, 波形として選択されなかったデータ点の情報を破棄することになる．再度 ECG の例を考える．ECG の波形は不整脈などの異常を示すが, 心拍部分以外の値にはあまり情報がない．しかし，波形の間隔からは心拍数が分かる．このような情報を利用するため, 間隔 $d_{i,j}$ を,$\mathbf{w}_{i,j-1}$ と $\mathbf{w}_{i,j}$ の時間間隔として定義し使用する．

［定義 1］（データ要素 $\mathcal{W}$ および $\mathcal{D}$） $\mathcal{W}$ と $\mathcal{D}$ を,$\mathcal{X}$ の要約されたデータ要素とする．

- 波形の集合：$\mathcal{W} = \{\mathbf{w}_{i,j}\}_{i,j=1}^{I,J_i}$
- 間隔の集合：$\mathcal{D} = \{d_{i,j}\}_{i,j=1}^{I,J_i}$

ここで,J_i は $\mathbf{x}_i$ から抽出された波形の数を示す．

次に，モデルの重要な概念である「ステージ」について説明する． 各波形 $\mathbf{w}_{i,j}$ と間隔 $d_{i,j}$ について, ステージラベル

$s_{i,j} \in \{1, \cdots, K\}$ を定義する. ここで,K はステージの数を示す.

[定義 2]（ステージ割り当て） データ要素 $\mathcal{W}$ および $\mathcal{D}$ に対するステージ割り当て $\mathcal{S}$ を $\mathcal{S} = \{s_{i,j}\}_{i,j=1}^{I,J_i}$ と定義する.

提案モデルでは, ステージが波形と間隔の振る舞いを指定する潜在的な状態であると仮定する. つまり,(P3) シーケンス内で時間変化するパターン をステージの遷移として捉える. また,(P4) 進行の不可逆性 を反映する特定の構造を進行ステージとして定義する.

[定義 3]（進行ステージ） 進行ステージ $\mathcal{S} \nearrow_j$ において, 任意の系列 $\mathbf{x}_i$ に関して, ステージラベル $\{s_{i,1}, \cdots, s_{i,J_i}\}$ は次の制約を満たす.

$$\forall i, j, j' \ j \leqq j' \Rightarrow \ s_{i,j} \leqq s_{i,j'}. \tag{1}$$

つまり，観察された系列が「逆戻り」しないと仮定する.

最後に, データ要素 $\mathbf{w}_{i,j}$ および $d_{i,j}$ が潜在的な進行ステージからどのように生成されるかを述べ，(P2) シーケンス間に共通するパターン を捉える方法について説明する.

本モデルではピーク部分の連続的な特徴を捉える方法として, 波形 $\mathbf{w}_{i,j}$ を利用する方法を提案した. ここではどのように W 次元のベクトル $\mathbf{w}_{i,j}$ を生成し, 波形の時系列的な文脈を保持しながらデータを表現するかを説明する. 本論文ではサブシーケンスの時系列的な特徴を保持しながらデータを表現するために, 行列分解を導入する. 波形集合 $\mathcal{W}$ およびステージ割り当て $\mathcal{S}$ が与えられたとき, ステージ k に属するすべての波形を積み重ねた行列として $\mathbf{W}_k$ を定義する. すなわち条件 $s_{i,j} = k$ を満たすすべての W 次元行ベクトル $\mathbf{w}_{i,j}$ を積み重ねた行列として定義する. $\mathbf{W}_k$ の各行はステージ k における 波形 $\mathbf{w}_{i,j}$ であり，$\mathbf{W}_k$ は $\{\mathbf{w}_{i,j} \in \mathcal{W} | s_{i,j} = k\}$ によって構成される. $s_{i,j} = k$ を満たす 波形の数を N_k とすると, 行列 $\mathbf{W}_k$ は $N_k \times W$ 行列となる. ステージ行列 $\mathbf{W}_k$ を特異値分解を用いて重み行列 $\mathbf{Q}_k \in \mathbb{R}^{N_k \times M}$ とコンポーネント行列 $\mathbf{V}_k \in \mathbb{R}^{M \times W}$ に分解する.

$$\mathbf{W}_k \simeq \mathbf{Q}_k \mathbf{V}_k \tag{2}$$

ここでは, 上位 M 個の特異値と特異ベクトルのみを使用して低ランク近似を行う. この分解により, コンポーネント $\mathbf{V}_k$ としてステージ k に共通して現れる波形の連続的な特徴を保持できる. ステージ行列 $\mathbf{W}_k$ はステージ k に属するすべての波形を含んでいる. そのため,$\mathbf{W}_k$ のコンポーネントは, 与えられた複数シーケンス全体からステージ k に共通のパターンを捉える.

[定義 4]（コンポーネント） ステージごとのコンポーネントの集合 $\mathcal{V}$ を次のように定義する.

$$\mathcal{V} = \{\mathbf{V}_k\}_{k=1}^{K} \tag{3}$$

重み行列 $\mathbf{Q}_k$ は, コンポーネント $\mathbf{V}_k$ を基底として $\mathbf{W}_k$ を表現するための重みを表す. この分解により, 行列 $\mathbf{W}_k$ の行, すなわち波形 $\mathbf{w}_{i,j}$ を, 重み行列 $\mathbf{Q}_k$ の行, すなわち $\boldsymbol{q}_{i,j}$ に変換する. ステージ k に属する波形は類似した形状を持つことを仮定し，元波形から変換された重みはステージ k の平均的な重みとその分散で特徴づけられる. すなわちガウス分布から次のように引かれると仮定する.

$$\boldsymbol{q}_{i,j} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) \ (s_{i,j} = k) \tag{4}$$

間隔 $d_{i,j}$ については, 指数分布を使用する. 指数分布は, 連続的かつ独立に発生するイベント間の間隔の確率分布である. 各陣痛は連続的かつ独立に発生すると仮定し, ステージ k における間隔 $d_{i,j}$ はパラメータ λ_k で以下のように表される.

$$d_{i,j} \sim Ex(\lambda_k) \ (s_{i,j} = k) \tag{5}$$

ここで,$Ex(\cdot)$ は指数分布の分布関数を表す.

以上をまとめると, データ要素 $\mathbf{w}_{i,j}$ および $d_{i,j}$ に対するステージラベル $s_{i,j}$ が与えられたとき，各ステージはパラメータ $\boldsymbol{\mu}_k$,$\boldsymbol{\sigma}_k$, および λ_k で定義される.

[定義 5]（ステージパラメータ θ_k およびステージモデル $\boldsymbol{\Theta}$）
ステージパラメータ θ_k は, ステージ k のパラメータのセットであり, すなわち $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k, \lambda_k\}$ である. ステージモデル $\boldsymbol{\Theta}$ は, 全てのステージのステージパラメータセットであり，以下で表される.

$$\boldsymbol{\Theta} = \{\theta_k\}_{k=1}^{K} \tag{6}$$

Algorithm 1 UCscan ($\mathcal{X}$)

Input: Sets of sequences $\mathcal{X}$, waveform size W,
the number of stages K and the number of components M
Output: Model $\boldsymbol{\Theta}$ and assignment$\mathcal{S}$ components $\mathcal{V}$,

```
for i = 1 : I do
    w_i, d_i = WaveDetection(x_i)
end for
W = {w_i}_{i=1}^I
D = {d_i}_{i=1}^I
S, V, Θ = UC-fit(W, D)
return S, V, Θ
```

4 UCscan

3 章では,UC 信号の波形パターンとステージ進行をどのように表現するかを述べた. 本章では，このモデルに基づいた問題を求めるためのアルゴリズムを述べる. 本論文で解くべき問題は以下の通りである.

[問題 1] 与えられたシーケンスの集合 $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{I}$ に対し,

- データ要素 $\mathcal{W}$ と $\mathcal{D}$ を抽出する.
- 時間発展に伴うステージの推移, および時系列に共通するコンポーネント，それらを表現するモデルを決定する. すなわち $\{\mathcal{S}, \mathcal{V}, \boldsymbol{\Theta}\}$ を決定する.

この問題を解くため, 本研究ではアルゴリズム UCscan を提案する. これは二つのサブアルゴリズム,WaveDetection と UC-fit からなる. Algorithm 1 は提案アルゴリズムの手順を示し,Figure 2 は提案アルゴリズムの概要を示す. WaveDetection は各シーケンス $\mathbf{x}_i$ から陣痛部の波形とその間隔を抽出し,UC-fit は進行的なデータパターンを捉える. 以下では, 各サブアルゴリズムを詳しく説明する.

4.1 自動波形検出

本論文では，UC 信号の陣痛部分に着目し，サブシーケンス $\mathbf{w}_{i,j}$ とその 間隔 $d_{i,j}$ をデータ要素として使用することを提案する. 本節では与えられたシーケンスから陣痛部をどのように検出するかについて述べる. すなわち，シーケンス $\mathbf{x}_i$ が与えられたとき，陣痛部分の波形の集合 $\mathbf{w}_i = \{\mathbf{w}_{i,j}\}_{j=1}^{J_i}$ およ

Algorithm 2 WaveDetection ($\mathbf{x}_i$)

Input: Original sequence $\mathbf{x}_i$, a window size W
Output: Set of waveforms $\mathbf{w}_i$ and set of intervals $\mathbf{d}_i$

$\mathbf{z}$ = HMM-infer($\mathbf{x}_i$)
$\{c_j\}_{j=1}^{J_i}$ = FindCenterPoint($\mathbf{z}$)
/* Extract waveforms and intervals by center points */
for $j = 1 : J_i$ **do**
  $\mathbf{w}_{i,j} = [x_{i,c_j - W/2+1}, \cdots, x_{i,c_j+W/2}]$
  if j=1 **then**
    $d_{i,j} = 0$
  else
    $d_{i,j} = c_j - c_{j-1}$
  end if
end for
$\mathbf{w}_i = \{w_{i,1}, \cdots, \mathbf{w}_{i,j}, \cdots, w_{i,J_i}\}$
$\mathbf{d}_i = \{d_{i,1}, \cdots, d_{i,j}, \cdots, d_{i,J_i}\}$
return $\mathbf{w}_i, \mathbf{d}_i$

Algorithm 3 UC-fit ($\mathcal{W}, \mathcal{D}$)

Input: Sets of data elements $\mathcal{W}$ and $\mathcal{D}$,
the number of stages K and the number of components M
Output: Stage model $\mathbf{\Theta}$, assignment$\mathcal{S}$ and wave components $\mathcal{V}$,

$\mathcal{S}$ = Initialize($\mathcal{W}, K, M$)
repeat
  for $k = 1 : K$ **do**
    $\theta_k, \mathbf{V}_k$ = UpdateStageModel($k, \mathcal{W}, \mathcal{D}, \mathcal{S}, M$);
    // Eq.(2),(14), (15) and (16)
  end for
  $\mathbf{\Theta} = \{\theta_k\}_{k=1}^{K}$
  $\mathcal{V} = \{\mathbf{V}_k\}_{k=1}^{K}$
  for $i = 1 : I$ **do**
    $\mathcal{S}$ = UpdateStageAssignment($\mathbf{w}_i, \mathbf{d}_i, \mathbf{\Theta}, K, \mathcal{V}$);
    // Eq.(17)
  end for
until convergence.
return $\mathcal{S}, \mathcal{V}, \mathbf{\Theta}$

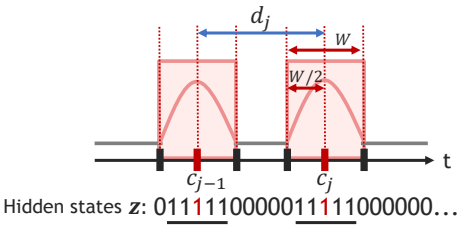


図 3: WaveDetection による自動波形検出の概要

びその間隔の集合 $\mathbf{d}_i = \{d_{i,j}\}_{j=1}^{J_i}$ を決定したい．そこで，本研究では Hidden Markov Model (HMM) に基づいた波形検出アルゴリズム WaveDetection を提案する．Algorithm 2 は WaveDetection の概要を示す．このアルゴリズムは次の 3 ステップからなる．

（1） **HMM-infer:** HMM に基づき，シーケンス $\mathbf{x}_i$ に対して隠れ状態 $z_t \in \{1, 0\}$ (陣痛部かそれ以外の部分) を推測する．

（2） **FindCenterPoint:** 隠れ状態系列から 波形 $\mathbf{w}_{i,j}$ を定めるための波形の中心点 c_j を決定する．

（3） **波形 $\mathbf{w}_{i,j}$ と 間隔 $d_{i,j}$ の抽出.**

以降では各ステップの詳細を述べる．

HMM-infer 本手法で利用する HMM は，2 つの隠れ状態をもつ．1 つは 陣痛部を生成し (State 1)，もう 1 つは陣痛部以外の部分を生成する (State 0)．詳細な HMM パラメータについては付録 8.1 に述べる．この HMM をシーケンス $\mathbf{x}_i = [x_{i,1}, \cdots, x_{i,t}, \cdots, x_{i,T_i}]$ にフィットさせ，最も尤度が高くなる隠れ状態 $\mathbf{z} = [z_1, \cdots, z_t, \cdots, z_{T_i}]$ を推定する．このステップは，陣痛部として $z_t = 1$ である部分を探す．

FindCenterPoint 前のステップで，隠れ状態系列 $\mathbf{z}$ を得る．HMM の隠れ状態は 2 状態のみであると仮定しているため，隠れ状態系列 $\mathbf{z}$ は Figure 3 に示すように連続する 0 と 1 の繰り返しとなっている．隠れ状態系列 $\mathbf{z}$ において $z_t = 1$ となる部分は陣痛部の中心付近を示す．そこで j 番目の中心点 c_j を，j 番目の連続する 1 の中心を指すインデックスとして定義する．

波形と間隔の抽出 最後に，前ステップで得た中心点により波形と間隔を抽出する．Figure 3 は波形と間隔の抽出の概要を示す．中心点 c_j によってサイズが W のサブシーケンス $\mathbf{w}_{i,j}$ を以下のように抽出する．

$$\mathbf{w}_{i,j} = [x_{i,c_j - W/2+1}, \cdots, x_{i,c_j+W/2}]. \tag{7}$$

また，その間隔を次のように定義する：

$$d_{i,j} = \begin{cases} 0 & \text{if } j = 0 \\ c_j - c_{j-1} & \text{otherwise.} \end{cases} \tag{8}$$

各中心点についてこの抽出を行い，シーケンス $\mathbf{x}_i$ に対してデータ要素 $\mathbf{w}_i$ と $\mathbf{d}_i$ を得る．

4.2 UC-fit

本節では，ステージの波形の特徴を表すコンポーネント $\mathcal{V}$ を抽出しつつ，最適なモデルパラメータ $\mathbf{\Theta}$ とステージ割り当て $\mathcal{S}$ とを決定するためのアルゴリズムの詳細を述べる．

本手法では WaveDetection を利用して, 与えられたデータを 波形 $\mathbf{w}_{i,j}$ と 間隔 $d_{i,j}$ で表現する．得られたデータ要素の集合 $\mathcal{W}$ と $\mathcal{D}$ に対して, モデルパラメータ $\mathbf{\Theta}$, ステージ割り当て $\mathcal{S}$, および コンポーネント $\mathcal{V}$ を得たい．提案モデルでは, 各 $\mathbf{w}_{i,j}$ が行列分解によって $\boldsymbol{q}_{i,j}$ に変換される．したがって，すべてのデータは重みベクトルの集合 $\mathcal{Q} = \{\boldsymbol{q}_{i,j}\}_{i,j=1}^{I,J_i}$ と間隔のセット $\mathcal{D}$ から成る．本手法では各ステージをガウス分布と指数分布のパラメータ θ_k によって定義する．したがって, 我々が解く最適化問題を以下のように定義する

$$\underset{\mathbf{\Theta}, \mathcal{S} \nearrow_j}{\arg\max} \sum_{k=1}^{K} \Big[\underbrace{\mathbb{L}_g(\mathcal{Q}|\theta_k, \mathcal{S})}_{\substack{\text{Gaussian} \\ \text{likelihood}}} + \alpha \underbrace{\mathbb{L}_e(\mathcal{D}|\theta_k, \mathcal{S})}_{\substack{\text{Exponential} \\ \text{likelihood}}} \Big] \tag{9}$$

ここで,$\mathcal{S} \nearrow_j$ は式 1 の単調性制約を表す．さらに, 係数 α[1] $\geqq 0$ は, 波形と間隔によるモデリングの影響力を調整するパラメータである．ここで, 目的関数の第 1 項は以下のように表される．

$$\mathbb{L}_g(\mathcal{Q}|\theta_k, \mathcal{S}) = \sum_{s_{i,j}=k} \ell\ell_g(\boldsymbol{q_{i,j}}|\boldsymbol{\mu_k}, \boldsymbol{\sigma_k}) \tag{10}$$

$\ell\ell_g(\boldsymbol{q_{i,j}}|\boldsymbol{\mu_k}, \boldsymbol{\sigma_k})$ は $\boldsymbol{q}_{i,j}$ がステージ k から得られる対数尤度を示す．各重みベクトル $\boldsymbol{q}_{i,j}$ は M 次元ベクトルであるため, この対数尤度は次のようになる．

1：係数 α はクロスバリデーションを用いて決定できる．

$$\ell\ell_g(\boldsymbol{q}_{i,j}|\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k) = \sum_{m=1}^{M} \ell\ell_g(q_{i,j,m}|\mu_{k,m},\sigma_{k,m})$$
$$= \sum_{m=1}^{M} -\frac{1}{2}\left\{\frac{(q_{i,j,m}-\mu_{k,m})^2}{\sigma_{k,m}} + \ln\sigma_{k,m}^2 + \ln 2\pi\right\} \quad (11)$$

次に, 目的関数の第 2 項は以下のように表される.

$$\mathbb{L}_e(\mathcal{D}|\theta_k,\mathcal{S}) = \sum_{s_{i,j}=k} \ell\ell_e(d_{i,j}|\lambda_k) \quad (12)$$

$\ell\ell_e(d_{i,j}|\lambda_k)$ は $d_{i,j}$ がステージ k の指数分布から得られる対数尤度を示す. この対数尤度は次のようになる.

$$\ell\ell_e(d_{i,j}|\lambda_k) = N_k \ln\lambda_k - \lambda_k \sum_{s_{i,j}=k} d_{i,j} \quad (13)$$

式 9 の問題において, モデルパラメータ $\boldsymbol{\Theta}$ とステージ割り当て $\mathcal{S}$ の 2 つの変数を決定する必要があるため, 大域解を求める簡単な方法はない. そのため, 本研究では目的関数を最大化しながらモデルパラメータとステージ割り当てを交互に推定するアルゴリズム UC-FIT を提案する. アルゴリズム 3 は UC-FIT の概要を示す. 本アルゴリズムでは, まず, 各波形 $\mathbf{w}_{i,j}$ のステージ割り当て $\mathcal{S}$ を K-means 法で初期化する. 次に, ステージ割り当て $\mathcal{S}$ を固定してステージモデル $\boldsymbol{\Theta}$ を更新し, ステージモデル $\boldsymbol{\Theta}$ を固定してステージ割り当て $\mathcal{S}$ を更新する. これら 2 つのステップ, すなわち, **ステージモデルの更新** と **ステージ割り当ての更新** を交互に繰り返す. 各ステップの詳細を以下に説明する.

ステージモデルの更新 最新のステージ割り当て $\mathcal{S}$ を与えられたとき, モデルパラメータ $\boldsymbol{\Theta}$ を更新する. 目的関数はステージごとの和を計算するため, 計算はステージごとに行うことができる. 各ステージでは, ガウス分布に関する尤度 $\mathbb{L}_g(\mathcal{Q}|\theta_k,\mathcal{S})$ と指数分布に関する尤度 $\mathbb{L}_e(\mathcal{D}|\theta_k,\mathcal{S})$ の合計を最大化する必要がある. まず, ガウス分布の更新方法を述べる. ステージ割り当て $\mathcal{S}$ を与えられたとき, 現在のステージ行列 $\mathbf{W}_k$ を定義し, 重み $\mathbf{Q}_k$ と コンポーネント $\mathbf{V}_k$ に分解する (式 2). その後, 対数尤度 $\ell\ell_g(\boldsymbol{q}_{i,j}|\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k)$ を最大化するための値 $\{\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k\}$ を見つける. ベクトル $\boldsymbol{q}_{i,j}$ は M 次元のベクトルであり, $\boldsymbol{q}_{i,j} = [q_{i,j,1},\cdots,q_{i,j,m},\cdots,q_{i,j,M}]$ と表せる. 各 $q_{i,j,m}$ は各ガウス分布から引かれる. したがって, 最尤推定によりパラメータは以下のように決まる.

$$\mu_{k,m} = \sum_{s_{i,j}=k} \frac{q_{i,j,m}}{N_k} \quad (14)$$

$$\sigma_{k,m} = \sqrt{\frac{1}{N_k}\sum_{s_{i,j}=k}(q_{i,j,m}-\mu_{k,m})^2} \quad (15)$$

次に, 指数分布を更新する. 対数尤度 $\ell\ell_e(d_{i,j}|\lambda_k)$ を最大化する値 λ_k を定める. 間隔 $d_{i,j}$ は指数分布から引かれることを仮定しており, 最尤推定により, パラメータは次のように決まる.

$$\lambda_k = \frac{N_k}{\sum_{s_{i,j}=k} d_{i,j}} \quad (16)$$

ステージ割り当ての更新 次に, 最新のステージモデル $\boldsymbol{\Theta}$ が与えられたとき,$\mathcal{S}$ を推定する. ステージの遷移は各シーケンスのダイナミクスを表すため, 他のシーケンスから独立に決定される. したがって, このステップでは各データ要素 $\{\mathbf{w}_{i,j},d_{i,j}\}$ に対してステージラベル $s_{i,j}$ を, 式 1 の単調性制約のもとで, シーケンスごとのデータ要素 $\{\mathbf{w}_{i,j},d_{i,j}\}_{i=1}^{I}$ に対する尤度が最大化されるように割り当てる. 各シーケンス $\mathbf{x}_i$ に対し, 次の最適化問題を解く

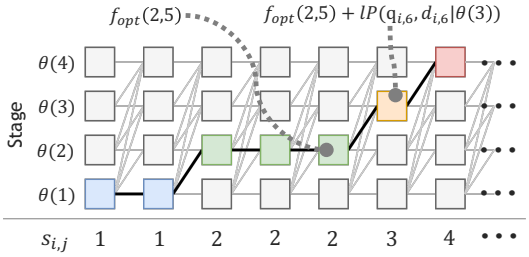


図 4: 単調性制約の下での動的計画法

$$\arg\max_{\{s_{i,j}\}\nearrow} \sum_{j=1}^{J_i} lP(\boldsymbol{q}_{i,j},d_{i,j}|\theta_{s_{i,j}}) \quad (17)$$

$$lP(\boldsymbol{q}_{i,j},d_{i,j}|\theta_{s_{i,j}}) = \ell\ell_g(\tilde{\boldsymbol{q}_{i,j}}|\mu_{s_{i,j}},\sigma_{s_{i,j}}) + \alpha\ell\ell_e(d_{i,j}|\lambda_{s_{i,j}}) \quad (18)$$

ここでは, 直前のステップで得られた コンポーネント $\mathbf{V}_{s_{i,j}}$ を用いて重み $\tilde{\boldsymbol{q}_{i,j}}$ を再構成したものを用いる.

$$\tilde{\mathbf{Q}} = \mathbf{W}\mathbf{V}_{s_{i,j}}^{-1} \quad (19)$$

単調性制約の下でのステージ割り当て $\mathcal{S}$ の最適化は, 動的計画法を使用して効率的に解くことができる. 図 4 は,i 番目のシーケンスに対する動的計画法を示す. 各行はステージを表し, 各列はインデックス j を表す. 一番下の行は割り当てられたステージラベルを表す. 再帰的な計算により, 以下のようにして, ステージ k の j 番目の要素に到達するための最適なコスト $f_{opt}(k,j)$ を計算できる.

$$f_{opt}(k,j) = \max_{1\leqq k'\leqq k}\{f_{opt}(k',j-1)\} + lP(\boldsymbol{q}_{i,j},d_{i,j}|\theta_k) \quad (20)$$

すべての k と j について $f_{opt}(k,j)$ を計算し, そのパスを記録する. 最終コスト $f_{opt}(k,J_i)$ が最大となるパスが, 最適なステージ遷移を表す最短パスとなる.

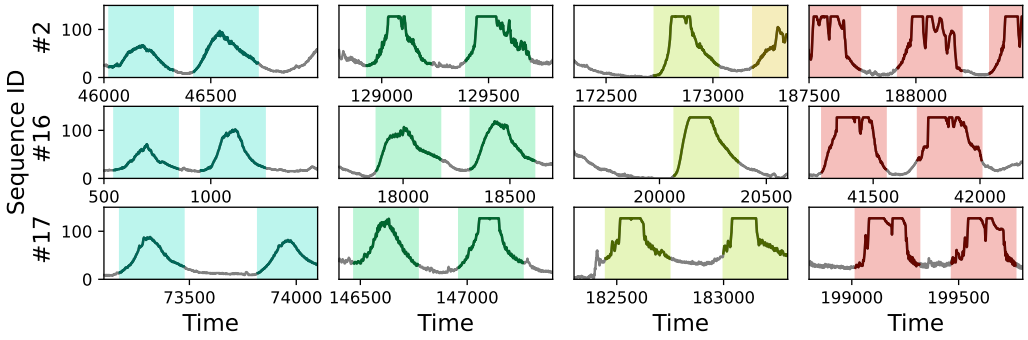
[定理 1] (UC-FIT の計算速度) UC-FIT にかかる計算速度はたかだか $\mathcal{O}(NK^2)$ である.

[証明 1] 証明は付録 8.2 に示す.

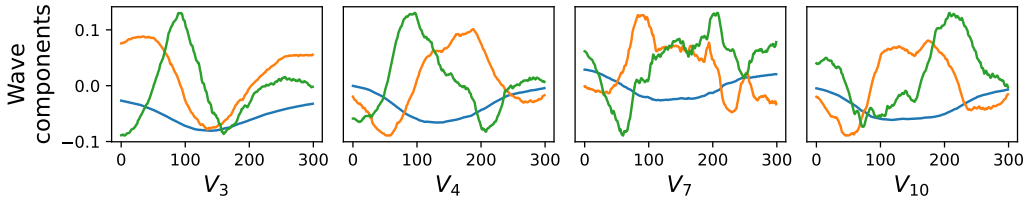
5 評価実験

このセクションでは, 実際の UC 信号を用いて UCSCAN を評価する.

データセット 2 つの実データセットを使用し, そのうちの 1 つ



(a) 各ステージに割り当てられた波形



(b) 各ステージのコンポーネント

図 5: UCScan により OUH データセットから抽出されたパターン

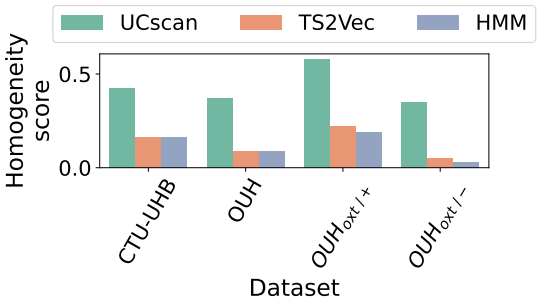


図 6: UCScan によるクラスタリング精度

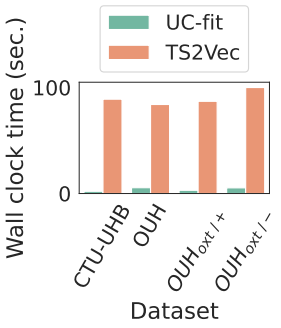


図 7: 各データセットにおける提案手法の計算速度

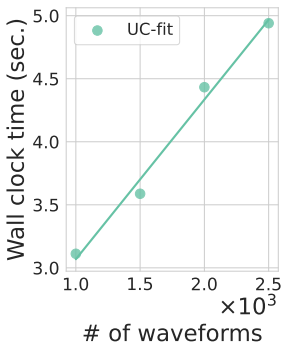


図 8: 入力数を変化させた時の提案手法の計算速度

からは分娩の変化を変える可能性のある属性情報を使用して 2 つのデータセットを作成した．以下は, 使用する 4 つのデータセットである．

- CTU-UHB [23]: 2010 年から 2012 年までの期間にプラハのチェコ工科大学（CTU）とブルノの大学病院（UHB）で収集された公開データセット．このデータセットには,UC 信号といくつかの医療データが含まれている．この実験では,UC 信号と分娩第二期の開始点データを使用し，精度実験に使用する正解ラベルを作成した．
- OUH: 2021 年に大阪大学病院から収集されたデータセット．このデータセットには,UC 信号といくつかの医療データが含まれている．この実験では,UC 信号とパルトグラムにおける子宮口開大度のデータを使用する．精度実験では, 補間された子宮口開大データを使用して正解ラベルを作成する．0〜6 cm のデータ点を潜在期に,6〜10 cm を活動期に,10 cm 以上を分娩第二期に割り当てる．
- $OUH_{oxt/+}$: OUH データセットのうち分娩誘発剤の投与を受けた妊婦のデータ．
- $OUH_{oxt/-}$: OUH データセットのうち分娩誘発剤の投与を受けなかった妊婦の OUH データ．

5.1 有 効 性

UCScan の出力は図 1 にも示した．ここでは,UCScan のモデリング能力について, 潜在的なステージの割り当てとデータからのパターン抽出について詳しく説明する．実験では, ステージの数を $K = 10$, コンポーネントの数を $M = 3$ と設定した．提案手法は, パターンの事前知識なしに, 時間的に変化するパターンをステージとして捕捉し, シーケンス全体の典型的なパターンをコンポーネントとして集約する．図 5 は,3 つの異なるシーケンスから抽出された元のシーケンス（上 3 段）とそのステージのコンポーネント（下段）を示している．横方向に同一シーケンスから得られた波形を比べると，波形が形状変化を示している．最初に, 波形は Stage 3 で丸く,Stage 4 と 7 で徐々に角ばり, 最後に Stage 10 では波形にいくつかのピークが見られる．縦方向に見ると, 同じ段階に割り当てられた波形とコンポーネントを比較することができる．提案モデルでは, 類似した形状の波形を同じ段階に割り当てる．また, コンポーネント

はその段階の波を反映しており，Stage 3 では, オレンジの線は丸く,Stage 4 では徐々に鋭くなる．Stage 7 では, 緑とオレンジの線が角ばった形状をしており,Stage 10 ではオレンジの線は元の波と同様に 2 つのピークを持つ.

5.2 精　　度

次に, クラスタリングにおける UCScan の精度を評価する．本実験では，次の 2 つの手法と比較して,UCScan の精度を評価する．TS2Vec [24] は, 時系列の表現を学習する手法であり,Hidden Markov Model（HMM）は確率的生成モデルである．TS2Vec による実験では，特徴ベクトルを取得し, その特徴ベクトルを用いて K-means クラスタリングを行った．また, HMM による実験ではデータをフィットさせて潜在状態を取得し, 潜在状態をクラスターとして扱った．オリジナルシーケンスは一次元のデータであり, また比較手法への入力としてデータ数が多すぎるため, 比較手法の入力として我々の波検出手法から抽出された W 次元のベクトルを使用した．性能評価は,Homogeneity score を使用した. これは, クラスターにおける正解ラベルの均一性を評価するクラスタリングの尺度である. より高い値は, より良くクラスターが分離されていることを示す.

図 6 は,UCScan を TS2Vec や HMM と比較したものである．我々の手法は, データの進行性を考慮し, 各段階での波形と間隔の特徴を捉えることができるため, 一貫して競合手法よりも優れた性能を発揮しする．TS2Vec は一部のパターンの変化を捉えることができるが, データの進行を捉えられないため, そのクラスターは断片的なものとなった．HMM は, データの形状の文脈でデータを捉えることができない.

5.3 計 算 時 間

大規模データの解析に重要となる計算時間の評価を行った. 図 7 では, 各データセットごとに UC-fit と TS2Vec を比較している．UC-fit は TS2Vec よりも最大で 40 倍速い計算速度を示した．注意しておきたいのは，TS2Vec への入力は, 我々の WaveDetection によって数千の波形のみであり, 実際の総データ長は最大約 180 万点であるため，実際には提案手法はより速度の面において優れている．WaveDetection が大規模データをコンパクトに表現し,UC-fit が入力に対し線形に動作するため,UCScan はこのような大規模データを効果的に処理できる．図 8 は, 入力波形のサイズを変化させた場合の UC-fit の計算時間を示している．UC-fit は定理 1 に示したように $\mathcal{O}(NK^2)$ で動作し，実験において実際に計算速度が波形の数に対して線形であることが示された.

6　アプリケーション

本章では，UCScan の実用的なアプリケーションとして，欠損値の補間に焦点を当てる．UC 信号の測定では, センサーの接触不良や他の測定のための一時的なセンサーの取り外しなどにより, 連続的な欠損が生じることがある. このような状況を想定し,OUH データセットのシーケンスに連続的な欠損を作成し,

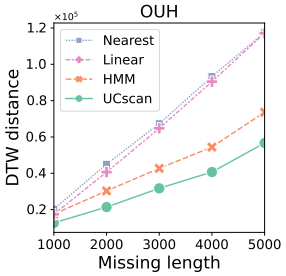


図 9: 欠損長を変化させた時の，実際のシーケンスと提案手法による再構成シーケンスとの類似度

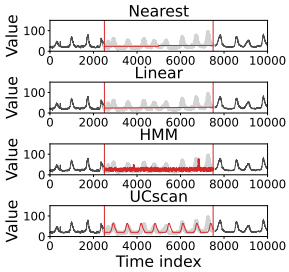


図 10: 提案手法により再構成された欠損値

それに対する UCScan の再構成効果を評価した.

本実験においては与えられたシーケンス全体が特定のステージに属するという仮定に基づき実験を行った. 具体的には,10000 点のサブシーケンスを入力とし，そのシーケンスから波形と間隔を抽出し, 尤度が最も高くなるステージをシーケンスが属するステージとした. そして, 欠損部分は, 決定された段階のパラメータを用いて再構成した. 実験では, 線形補間や最近傍補間といった他の補間方法, および提案手法と同様に確率的な生成モデルである HMM との比較も行った. 復元品質の評価には, 医師が UC 信号を目視で使用することを考慮しシーケンスの類似性を測定する DTW 距離を性能指標として用いた. これは値が大きいほどシーケンスがよく類似していることを示す.

図 9 には, 欠損長を 1000 から 5000 まで変化させた際の平均 DTW スコアが示されている. この結果から,UCScan が 4 つの方法の中で最も優れた性能を発揮することが示されている. また, 図 10 では, 再構成により欠損が補間された信号が示されている. 線形および最近傍補間では山型の信号を適切に表現することができず，HMM は多少改善されているが, 依然として実際の値からは程遠い. 一方,UCScan は信号の波形と間隔の両方を捉え, 欠損値の再構成を達成している.

7　む　す　び

本論文では，大規模生体データのための統合解析モデルとして UCScan を提案した．UCScan は，生体時系列データにおいて，重要な意味を持つ波形部分を自動的に抽出し，さらにそれらの波形を用いて，シーケンス内におけるダイナミクスのパターンをステージとして捉え，またシーケンス間にまたがる特徴的な波形のパターンをコンポーネントとして集約できる．実データを用いた実験では，最新の時系列解析手法と比べてより高い精度と性能を持つことを示した.

謝辞. 本研究の一部は JSPS 科研費，JP21H03446，国立研究開発法人情報通信研究機構委託研究（NICT）委託研究 JPJ012368C03501，JSTAIP 加速課題 JPMJCR21U4，JST CREST JPMJCR23M3, の助成を受けたものです.

文　献

[1] Carla AbouZahr. Global burden of maternal death and disability. *British Medical Bulletin*, 67(1):1–11, 12 2003.
[2] WHO. Who postpartum haemorrhage (pph) summit. 2022.
[3] Emanuel A. Friedman. The graphic analysis of labor. *American Journal of Obstetrics and Gynecology*, 68(6):1568–1575, 1954.
[4] Characterization and control of intraamniotic infection in an urban teaching hospital. *American Journal of Obstetrics and Gynecology*, 175(2):304–310, 1996.
[5] Ian Davidson, Sean Gilpin, Owen Carmichael, and Peter Walker. Network discovery via constrained tensor analysis of fmri data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 194–202, New York, NY, USA, 2013. Association for Computing Machinery.
[6] Kristjan Greenewald, Seyoung Park, Shuheng Zhou, and Alexander Giessing. Time-dependent spatially varying graphical models, with application to brain fmri data analysis. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
[7] Yasuko Matsubara, Yasushi Sakurai, Willem G. van Panhuis, and Christos Faloutsos. Funnel: automatic mining of spatially coevolving epidemics. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 105–114, New York, NY, USA, 2014. Association for Computing Machinery.
[8] Tasuku Kimura, Yasuko Matsubara, Koki Kawabata, and Yasushi Sakurai. Fast mining and forecasting of co-evolving epidemiological data streams. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3157–3167, New York, NY, USA, 2022. Association for Computing Machinery.
[9] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.
[10] Yakir Yehuda, Daniel Freedman, and Kira Radinsky. Self-supervised classification of clinical multivariate time series using time series dynamics. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5416–5427, New York, NY, USA, 2023. Association for Computing Machinery.
[11] Emma Pierson, Tim Althoff, and Jure Leskovec. Modeling individual cyclic variation in human behavior. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 107–116, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
[12] Jonathan R Bull, Simon P Rowland, Elina Berglund Scherwitzl, Raoul Scherwitzl, Kristina Gemzell Danielsson, and Joyce Harper. Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *NPJ digital medicine*, 2(1):83, 2019.
[13] Bo Liu, Shuyang Shi, Yongshang Wu, Daniel Thomas, Laura Symul, Emma Pierson, and Jure Leskovec. Predicting pregnancy using large-scale data from a women's health tracking mobile application. In *The World Wide Web Conference*, WWW '19, page 2999–3005, New York, NY, USA, 2019. Association for Computing Machinery.
[14] Peng Wang, Haixun Wang, and Wei Wang. Finding semantics in time series. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, page 385–396, New York, NY, USA, 2011. Association for Computing Machinery.
[15] Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos. Autoplait: automatic mining of co-evolving time sequences. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, page 193–204, New York, NY, USA, 2014. Association for Computing Machinery.
[16] Yasuko Matsubara, Yasushi Sakurai, Naonori Ueda, and Masatoshi Yoshikawa. Fast and exact monitoring of co-evolving data streams. In *2014 IEEE International Conference on Data Mining*, pages 390–399. IEEE, 2014.
[17] Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 783–794, New York, NY, USA, 2014. Association for Computing Machinery.
[18] Spiros Papadimitriou, Anthony Brockwell, and Christos Faloutsos. Adaptive, hands-off stream mining. In Johann Christoph Freytag, Peter C. Lockemann, Serge Abiteboul, Michael J. Carey, Patricia G. Selinger, and Andreas Heuer, editors, *Proceedings of 29th International Conference on Very Large Data Bases, VLDB 2003, Berlin, Germany, September 9-12, 2003*, pages 560–571. Morgan Kaufmann, 2003.
[19] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos. Streaming pattern discovery in multiple time-series. In Klemens Böhm, Christian S. Jensen, Laura M. Haas, Martin L. Kersten, Per-Åke Larson, and Beng Chin Ooi, editors, *Proceedings of the 31st International Conference on Very Large Data Bases, Trondheim, Norway, August 30 - September 2, 2005*, pages 697–708. ACM, 2005.
[20] Deepayan Chakrabarti, Spiros Papadimitriou, Dharmendra S. Modha, and Christos Faloutsos. Fully automatic cross-associations. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 79–88. ACM, 2004.
[21] Spiros Papadimitriou and Philip S. Yu. Optimal multi-scale patterns in time series streams. In Surajit Chaudhuri, Vagelis Hristidis, and Neoklis Polyzotis, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 27-29, 2006*, pages 647–658. ACM, 2006.
[22] David Hallac, Sagar Vare, Stephen P. Boyd, and Jure Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 215–223. ACM, 2017.
[23] . `https://physionet.org/content/ctu-uhb-ctgdb/1.0.0/`.
[24] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987, 2022.

8 付　録

8.1 自動波形検出における HMM パラメータ

陣痛部分を検出するために, 本論文では HMM を使用する. HMM は, 陣痛部分（State 1）とその他の部分（State 0）の 2 つの隠れ状態から構成される.HMM は初期状態確率 π, 状態遷

移確率 $\mathbf{A}$, 出力確率 $\mathbf{B}$ から構成される．以下ではその詳細を述べる

まず初期確率 π を, データ点が陣痛の間に記録される確率 p によって定義する. 陣痛の持続時間を D, 周期を T とすると, データ点が陣痛の間に記録される確率は $p = D/T$ と表される. パラメータ D と T はドメイン知識から設定することができる. 我々の実験では,$D = 70$，$T = 250$ とした.State 1 は陣痛部分を生成し,State 0 は別の部分を生成するため,$\pi = \{1-p, p\}$ となる. 同様に, 遷移確率 $\mathbf{A}$ は次のようになる.

$$\mathbf{A} = \begin{bmatrix} 1-p & p \\ 1-p & p \end{bmatrix}$$

実際には, 遷移確率はこのパラメータ設定よりも複雑な場合がある. 例えば, 陣痛の期間中に持続する可能性がより高い場合（すなわち,State 1 から State 1 への遷移が State 0 から State 1 への遷移よりもより頻繁に発生する）などである. しかし,WaveDetection がこのパラメータ設定で効果的に動作したため, 実験ではこのパラメータ設定を採用した.

最後に, 出力確率 $\mathbf{B} = \{b_0(\cdot), b_1(\cdot)\}$ について説明する. ここで $b_r(\cdot)$ は状態 r の出力確率関数である. この HMM 手法では,State 1 から生成されたと思われるデータ点の値ではなく, その近くの波形が State 1 に属している中心点を探索したいということに注意されたい. そのため, 出力確率関数 $b_r(\cdot)$ には単一の点 $x_{i,t}$ ではなく, サイズが D の部分列 $\mathbf{X}_{i,t}$ を入力データとして使用する. 部分列 $\mathbf{X}_{i,t}$ はデータ点 $x_{i,t}$ を中心とする部分列であり, 次のように表される: $\mathbf{X}_{i,t} = [x_{i,t-D/2+1}, \cdots, x_{i,t}, \cdots, x_{i,t+D/2}]$, これは D 次元の行ベクトルである. したがって, シーケンス $\mathbf{x}_i$ が与えられた場合, 新しいシーケンスに対する隠れシーケンス $[z_1, \cdots, z_t, \cdots, z_{T_i}]$ を推論する. ここで, 部分列 $\mathbf{X}_{i,t}$ は状態 r で平均波形 [2] $[u_1^{(r)}, \cdots, u_D^{(r)}]$ から生成されたと仮定する. 分散パラメータ ρ を用いて, 出力確率関数は次のようになる:

$$b_r(\mathbf{X}_i) = \left\{ \mathcal{N}(x_{i,t-D/2+d} | u_d^{(r)}, \rho^2) \right\}_{d=1}^{D}.$$

8.2 定理 1 の証明

ステージモデル θ_k を更新する際には, 各ステージにおいて, M 次元の $\boldsymbol{\mu}_k$ と $\boldsymbol{\sigma}_k$, そして 1 次元の λ_k を計算する．各パラメータの計算には, ステージ k の波形の数 N_k に比例した計算量が必要であり, したがって, 各ステージの計算コストは $\mathcal{O}(N_k \cdot (M+1))$ となる. ステージごとの更新コストは並列に計算することができるため，従って,$N_k{}_{k=1}^{K}$ の最大値を $max(N_k)$ とすると, 更新コストは $\mathcal{O}(max(N_k) \cdot (M+1))$ となる.

次に,$\mathbf{w}_i$ と $\mathbf{d}_i$ のステージ割り当て s_i を更新する. 提案手法における動的計画法アプローチは, 標準的な動的計画法よりも少ないパスで計算できるため（図 4）, この計算量は標準的な動的計画法の計算量 $\mathcal{O}\big(K^2 \cdot J_i\big)$ よりも少ない．ここで J_i はシーケンス $\mathbf{x}_i$ の波形の数である. モデルの更新と同様に, 各シーケンスの割り当てコストを並列に計算することができ, ステージの割り当てコストは最大で $\mathcal{O}\big(K^2 \cdot max(J_i)\big)$ である.

したがって, 合計のコストは最大で $\mathcal{O}\big(\#iter \cdot (max(N_k) \cdot (M+1) + K^2 \cdot max(J_i))\big)$ である. ここで,$\#iter$ と $M+1$ は小さな無視できる定数値であり，また,$max(N_k)$ と $max(J_i)$ は最大でも波形の総数 N である. したがって, 合計の計算量は $\mathcal{O}\big(NK^2\big)$ である.

2：実験では,State 0 では長さ D のゼロシーケンス,State 1 では 陣痛シグナルにスケーリングされた半周期正弦波を使用した.

DEIM Forum 2024 T2-A-9-05

SNS におけるサイバー犯罪対策 : 交グラフを活用した犯罪グループ探索アルゴリズム

大原　望乃† 　伊藤　純菜†† 　趙　智賢††† 　長田　繁幸††† 　中川　直樹†††
小口　正人†

† お茶の水女子大学大学院　〒 112–8610 東京都文京区大塚 2 丁目 1 番 1 号
†† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2 丁目 1 番 1 号
††† 株式会社 日本総合研究所　〒 141–0022 東京都品川区東五反田 2 丁目 18 番 1 号
E-mail: †{g1920513,oguchi}@is.ocha.ac.jp, ††g2020507@is.ocha.ac.jp ,
†††{zhao.zhixian,osada.shigeyuki,nakagawa.naoki}@jri.co.jp

あらまし　近年，フィッシング詐欺等で盗まれたクレジットカード情報が SNS 上で売買されている問題が深刻化している．本論文では，このようなサイバー犯罪の抑止を目的とした新たな SNS ネットワーク分析手法として，グラフ理論を駆使した手法を提案し，そのアルゴリズムを実装する．提案手法では，犯罪コミュニティのデータから複雑ネットワークグラフを生成し，これを交グラフに変換することで，犯罪コミュニティ同士の結びつきや中心的なグループを特定することを試みる．また，提案手法の有効性を検証するために，実装されたアルゴリズムを既存の犯罪コミュニティデータに適用し，その結果を考察する．

キーワード　グラフデータ処理，ソーシャルメディア，セキュリティ，交グラフ

1　はじめに

近年，インターネットの普及と共に，SNS が我々の日常生活に深く結びついてきている．SNS は情報発信や収集などに便利な反面，犯罪手段として悪用される事例も増加している．例えば，フィッシング攻撃を利用したクレジットカード（以下，単にカードと呼ぶ.）情報の窃取やカード番号の不正利用被害は年々増加し [1]，窃取された情報が SNS 上で取引されている例も見受けられる．このように，犯罪者たちは SNS の匿名性を利用して犯罪ビジネスを拡大させている．筆者らはこれらのサイバー犯罪に対応すべく，SNS の公開情報を収集してきた．犯罪者らが参加する公開グループの会話履歴を分析し，犯罪者らをグループ分けした結果，犯罪ビジネス俯瞰図が導かれた [2]．彼らは犯罪ビジネスとして長期間個々で関わりを持っており，SNS を通じて「知り合い」や「集団」という関係が発生していると考えられる．本稿では，犯罪者間の関連性を発見することを目的として，収集した犯罪者グループのデータから交グラフを利用した犯罪者ネットワークの分析手法を提案する．さらに，提案法の有効性を検証するために，収集した既存の犯罪コミュニティデータに提案法を適用し，その結果を考察する．また，先行研究においてデータ収集を行ったとき，監視者の個人情報の漏洩が原因で攻撃を受けたことから，本稿ではこれを防ぐために，より安全なデータ収集方法 [3] を利用してデータを収集するものとする．

本稿の構成は以下の通りである．第 2 章では SNS を利用したサイバー犯罪の動向と先行研究を紹介する．第 3 章では交グラフを利用したコミュニティ発見手法を提案する．第 4 章では実際のデータを使った実験を行うにあたって，具体的な実装とテストケース，実験概要を述べる．第 5 章では実験結果とその考察を述べ，第 6 章でまとめる．

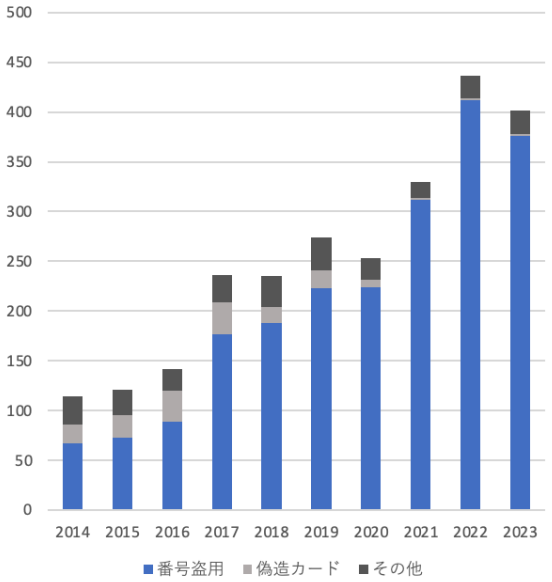


図 1　クレジットカードの不正利用被害額の推移

2　先行研究

2.1　フィッシング攻撃と SNS

サイバー犯罪及びフィッシング攻撃に関する多岐にわたる研究が進められている．Konradt ら [4] は，サイバー犯罪とフィッ

シング攻撃について経済的な観点から分析を行い，サイバー犯罪に係るコストと得られる利益の関係について統計的なモデルを作成している．また，Singhs[5] は，カードの不正利用は犯罪エコシステムが形成されていると主張して，フィッシング攻撃でカード情報を集めるところから現金化するまでの一連の流れを分業によって行うモデルを作成している．国内においては，一般社団法人日本クレジット協会 [1] の報告によると，図 1 に示すように 2022 年に発生したカード不正利用被害額 436.7 億円のうち，94.3%に当たる 411.7 億円がフィッシング攻撃などに起因する番号盗用被害であったと報告されている．

趙ら [2] は，国内のカード不正利用被害において，カード情報の窃取と盗用のプロセスについてモデル化を行っている．また，モデルに基づき犯罪グループ内の弱点が犯罪者同士の脆弱な信頼関係にある可能性を指摘し，盗用被害の抑止策として，弱い信頼関係を更に弱化する方法，すなわち犯罪グループの監視を提案している．SNS 上でのカード情報の売買は，相互の身元を隠した状態で行われるため，売買成立に必要な信頼関係を構築しにくい．そのため，売り手は信頼度を高めるためにカード情報の一部（例えばカード番号）をサンプルとして投稿する傾向がある．このとき，この情報をカード会社が入手できれば，その時点でカードの利用停止などの対処をすることで盗用被害を防ぐことが可能である．趙らは，実際に監視を行った結果を示して，その有効性を確認している．また，研究対象の SNS としては Telegram[6] が採用されている．

2.2 安全なデータ収集

Telegram を安全に利用する手法については，様々な観点から検討がなされている．Luo ら [7] は，SNS 上の個人情報の漏洩に繋がる行為を，脈絡のない情報開示，ネットワーク内情報集約，複数ネットワークを跨いだ情報集約の 3 つに分類している．Luo らは，個人情報の漏洩につながる行為が発生していないことを確認するためのユーザのプライバシを観察するという機構を提案し，これを“プライバシサンドボックス”と呼んでいる．

筆者らは，Luo らのプライバシサンドボックスを，SNS 上の犯罪グループを監視する際のプライバシ保護のために応用することを提案している [3], [8], [9]．これらの論文では，Telegram で活動する犯罪グループの監視を目的としたアカウント（以下，監視者）の情報が，犯罪グループに参加するアカウント（以下，監視対象者）に意図せず開示されていたことが原因で発生したサイバー攻撃事案を踏まえ，監視者の情報が第三者にどのように見えるか確認するための機能を実装し，評価している．また，Telegram はマルチクライアント利用が許可されており，複数のクライアントで異なるプライバシ設定がなされたときに，意図しない情報開示が発生する可能性がある．これを防ぐために，プライバシサンドボックスで常時外部から情報開示状況を確認し，もし望ましくない設定が確認されれば，外部から強制的に安全な設定に変更するような機構を提案して，その有効性を確認している．

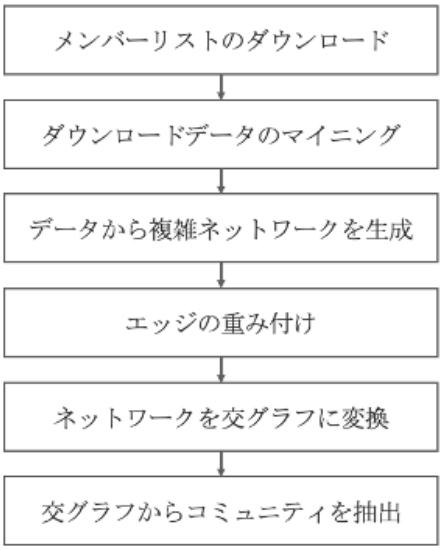


図 2　提案法の処理フロー

2.3 交グラフを用いたコミュニティ発見手法

倉持ら [10] は，様々な複雑ネットワークに対して交グラフとネットワーク上に存在する意味的な情報の解析を利用したグループ発見手法を提案している．この提案手法は，入力されたグラフ $\mathbf{G} = (\mathbf{V}, \mathbf{E})$（$\mathbf{V}$ はノードの集合，$\mathbf{E}$ はエッジの集合を示す）に対し，1. 密な部分グラフの列挙，2. 交グラフへの変換，3. エッジの重みの算出，4. モジュール性に基づくクラスタリングの 4 つのステップからなる処理フローを行うことでグループを抽出する．部分グラフを交グラフへ変換する処理は Everett ら [11] の提案手法に基づいており，ある人物が複数のグループに属しているような場合もグループ間の重複を抽出することが可能である．さらに，エッジに対して集合の重なり度合いと意味的情報の類似度を用いて重みを算出し，エッジの重み付けをすることで，課題であったネットワークの不均一性の反映を可能にしている．

また，著者らは，過去に収集した実際の犯罪者グループのデータを分析し，犯罪者らの中に中心となるようなユーザや集団が存在するかどうか検証した [12]．検証の結果，犯罪者らの中に中心と予想されるユーザや集団の存在が確認できた．このことから，本稿の提案は，犯罪者らの中に中心となるようなユーザや集団が存在することを前提として進める．

3 提　　案

3.1 提 案 法

提案法は，図 2 に示すような 6 つの処理で構成する．なお，この手法を適用する SNS としては，先行研究と同じく，サイバー犯罪の事例が多く見られる Telegram を対象とする．

Step 1. メンバーリストのダウンロード

過去の研究で収集したデータからなる犯罪者グループのデータベースから，グループごとに所属するユーザの一覧（以下，

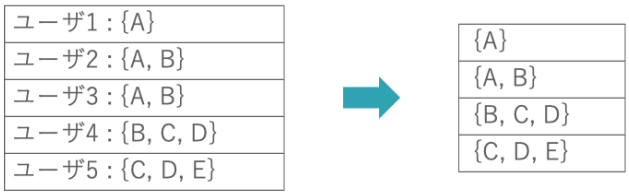


図 3　Step 2 データのマイニング

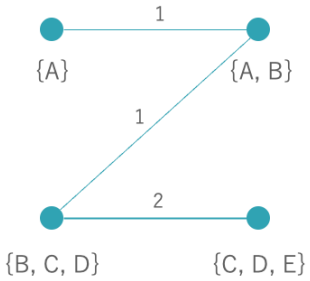


図 4　Step 3, 4 グラフ生成

メンバーリストと呼ぶ.）をダウンロードする.

Step 2. ダウンロードデータのマイニング

ダウンロードしたデータに対し，グラフ変換に必要な形になるよう処理を施し，新たなデータ構造で出力する．具体的には，Step 1 で取得したメンバーリストから，リスト内のユーザが所属するグループを全て求め，所属するグループ群を一覧として出力する．例として，図 3 のように，メンバーリストに 5 人のユーザが存在し，それぞれ所属するグループが {A}，{A, B}，{A, B}，{B, C, D}，{C, D, E} だったとすると，出力するデータは {A}，{A, B}，{B, C, D}，{C, D, E} の 4 つである.

Step 3. データから複雑ネットワークを生成

マイニングしたデータから，複雑ネットワークのグラフを生成する．具体的な定義としては，先ほど得られたデータを入力データとして，所属するグループ群それぞれをノードとし，グループ群に同じグループが含まれる場合は対応するノード間にエッジを引くことでグラフを描く.

Step 4. エッジの重み付け

設定した定義に従い重みを計算し，エッジに重みを付ける．今回はグループの重複回数を重みとして計算する．先ほどの例の場合は，Step 3 の処理と合わせて，図 4 のようなグラフが描かれる.

Step 5. ネットワークを交グラフに変換

重み付けを完了したグラフを交グラフに変換する．変換処理の実装としては，倉持らと同じく，Everett らの提案手法を採用する．入力グラフから極大クリークを列挙し，各クリークを 1 つのノードとするグラフを生成すると，交グラフが得られる.

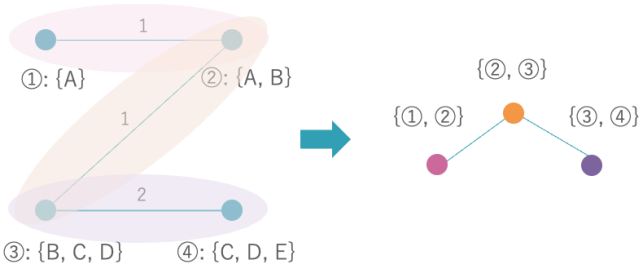


図 5　Step 5 交グラフへの変換

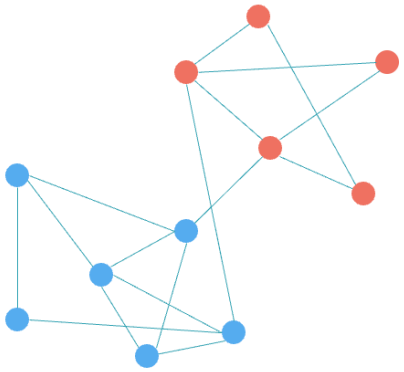


図 6　交グラフのクラスタリング

図 4 のグラフを交グラフに変換したものが図 5 である.

Step 6. 交グラフからコミュニティを抽出

交グラフをクラスタリング分析することで，中心と予想される集団やユーザを抽出する.

3.2　グラフの彩色

提案法の Step 6 ではさまざまなコミュニティの抽出方法が挙げられるが，本稿では，交グラフのノードを彩色し元のグラフに戻すことでコミュニティを抽出する．具体的には，まず Step 3 でグラフを生成した際，ノードのデータベースを作成しておく．Step 5 で交グラフに変換されたのち，Step 6 でクラスタリングを行う際，図 6 のように，クラスタ数に合わせてノードを彩色する．この例の場合はクラスタ数を 2 に設定したため，赤と青の 2 色でノードを彩色する．さらに，図 7 のように，先ほど作成したノードのデータベースとクラスタリングによる彩色の情報を合わせ，変換前のグラフに対してもノードの彩色を行う．このとき，複数のクラスタに属するノードが存在する可能性があり，そのようなノードは紫で彩色を行う．この紫で彩色されたノードは，コミュニティの重なりの中心となっており，本稿で特定を試みていた犯罪の中心となる集団やユーザである可能性が高い．紫で彩色されたノードについて，データベースよりグループや所属するユーザを抽出することで，中心と予想される集団やユーザを抽出することが可能である.

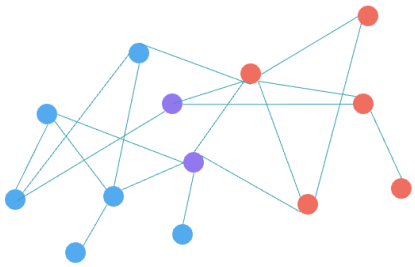


図 7　元に戻したグラフ

単一のグループだけに所属するユーザ数	90856
複数のグループに所属するユーザ数	21625
合計ユーザ数	112481

表 8　テストケースのユーザ数

4　実　　験

提案法を過去に収集した実際の犯罪コミュニティのデータに対して適用する実験を行った．以下では具体的な実装，用いたテストケースの選出方法，一について述べる．

4.1　実　　装

実装には Python，NetworkX，Matplotlib を利用した．また，比較のために，グラフの描画はすべてノード間の反発係数 $k = 0.5$，ランダムシード $seed = 1$ を固定値とする．Step 6 のクラスタリング手法としては k-means 法を採用した．

4.2　テストケース

実験のテストケースとして，犯罪に関与するグループをまとめたデータベース（2023 年 8 月時点のもの）内の 50 人以上のユーザを有するグループの中から，100 個のグループを選出し，英語だけを使用してコミュニケーションを行うグループ 5 個と，グループ名にキーワード”CVV”（クレジットカードのセキュリティコードを指す．カード情報を売買しているグループのタイトルに使用されることが多い．）を持つグループ 95 個とを，それぞれ無作為に選び、これら計 100 個のグループのいずれかに所属するユーザの情報を収集した． 結果，収集したデータの内訳は表 8 のようになった．合計ユーザ数は 112481，単一のグループだけに所属するユーザ数は 90856，複数のグループに所属するユーザ数は 21625 である．提案法は，犯罪者間の繋がりや，中心となる集団を見つけることを目的としているため，このうち 2 つ以上のグループに所属するユーザのみ実験対象とし，この 21625 人からなるメンバーリストをテストケースとして実験に適用する．

4.3　ノードとエッジの削減

Step 1 にあたる作業では 21625 人のユーザからなるメンバーリストのダウンロードを行うことになる．さらに，Step 2 ではこのリスト内の 21625 人のユーザが所属するグループを全て求め，所属するグループ群を一覧として出力することになるが，この結果，出力されたグループ群の総数は 7780 個であった．提案法の処理フローに従うと，Step 3 では，このグループ群それぞれをノードとしたグラフ，すなわちノード数が 7780 個のグラフを生成することとなるが，これはデータ数が多すぎるため，現実的な値ではない．そのため，このデータの中から重要度の低いと思われるノードとエッジを削減する必要がある．

そこで，本稿では実験を以下の 2 段階に分けて行った．

実験 1 閾値によるノード・エッジ削減の比較

Step 1 〜 Step 3 までの処理に加え，条件を変更しながらノードとエッジを削減し，生成したグラフを比較する．削減処理のステップとしては以下である．

処理 1.　グループ群の中で，グループ群 A の要素の集合 $\subset$ グループ群 B の要素の集合となるような場合，グループ群 A はグループ群 B に吸収されるものとして削除する．例えば，グループ群の中に {A}，{A, B} が存在するとき，{A} は削除する．

処理 2.　エッジの重みに閾値を設け，重みが閾値を下回るエッジは削除する．

処理 3.　処理 2 を行った結果，接続されているエッジの数が 0 本になったノードを削除する．

実験 2 コミュニティの抽出

Step 5〜Step 6 の処理を行う．実験 1 で得られたグラフからノード数とエッジ数が適しているものを選び，交グラフに変換したのち，クラスタリングと彩色を行う．さらに元のグラフに戻し，コミュニティを抽出する．

5　結果・考察

5.1　実験結果 1

まず，削減処理を行う前のデータは，ノードが 7780 個，エッジが 21861892 本であった．削減処理のうち，処理 1. を実行したところ，ノード数は 1857 個まで減少した．さらに，処理 2. で重み $weight$ に対する閾値 n を 0〜20 まで変化させた際の，処理 3. 終了時点でのノード数とエッジ数をまとめたものが図 9，表 10 である．

さらに，これをそれぞれ力データとしてグラフを生成した．図 11 は $n = 5, 8, 11, 14$ を抜粋したものである．ノードを赤で，エッジを青で濃淡をつけながら彩色した．ここで，ノードは赤の濃度が濃いほど繋がっているエッジの本数が多いことを，エッジは青の濃度が濃いほど重みが大きいことを示している．$n = 5$ のときのグラフは全体的に円形になっており，繋がりが目立つような集合は見られないが，$n = 8, 11$ と大きくなるにつれ，グラフに偏りが生じ，$n = 14$ のときには，グラフには 3 つの集合が存在することが読み取れる．

5.2　実験結果 2

実験 1 の結果より，今回は $n = 11$ のときのデータを用いて

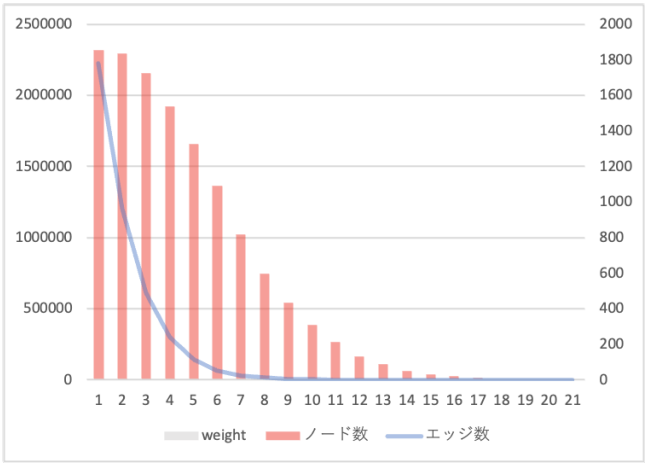


図 9　実験 1 削減処理によるノードとエッジの個数推移

weight >n	ノード数	エッジ数
0	1857	2226250
1	1835	1204576
2	1723	613732
3	1536	300802
4	1327	142678
5	1092	65756
6	819	30590
7	596	14604
8	432	7036
9	308	3300
10	214	1514
11	132	698
12	89	368
13	48	180
14	32	104
15	21	56
16	13	28
17	5	8
18	5	8
19	2	2
20	0	0

表 10　ノードとエッジの個数

実験 2 を進めた．まず，Step 4 の段階で生成されたグラフを，ノードの色の濃淡をつけず一定にし，エッジの色をグレースケールに変更し，各ノードに A1〜ラベルをつけたものが図 12 である．このラベルはコミュニティを抽出する際にわかりやすくするためのものである．また，後ほど各ラベルとグループの対応を照会するので，ここでノード情報のデータベースも作成した．

次に，このグラフを交グラフに変換し，$k = 3$ でクラスタリングを行った．各クラスタのノードをそれぞれ赤，青，緑で彩色し，B1〜ラベルをつけたものが図 13 である．なお，再現性を保つため，クラスタリングの乱数は固定している．

最後に，ノード情報のデータベースをもとに，交グラフで

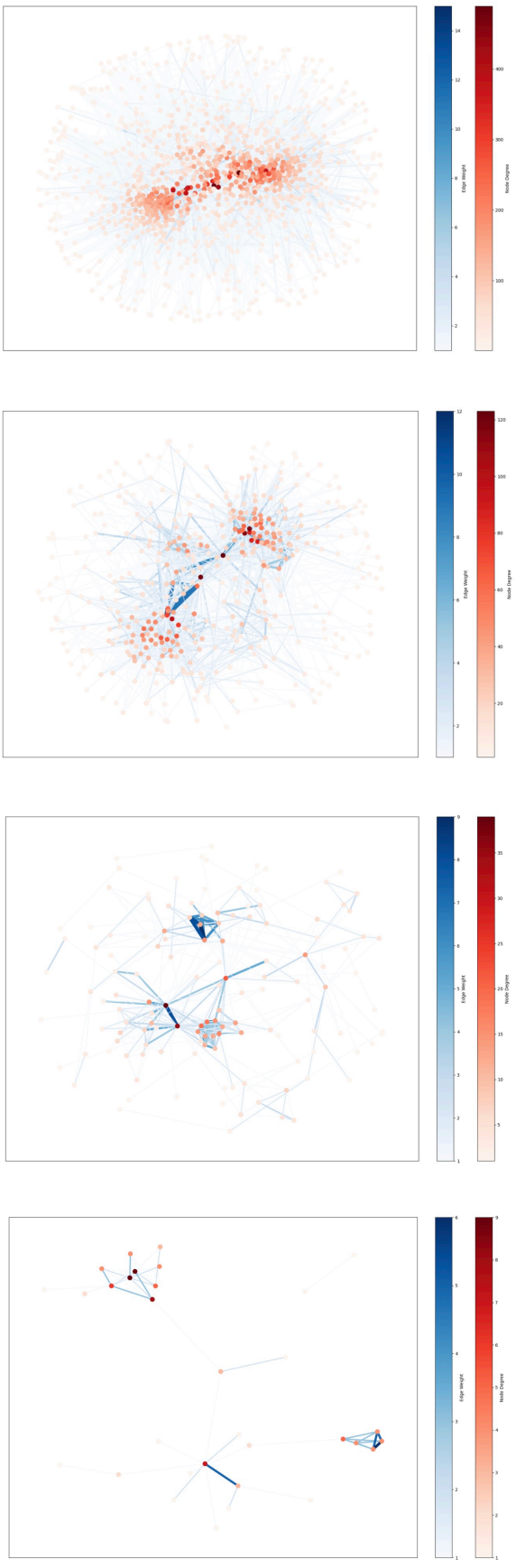


図 11　$n = 5, 8, 11, 14$ のグラフ

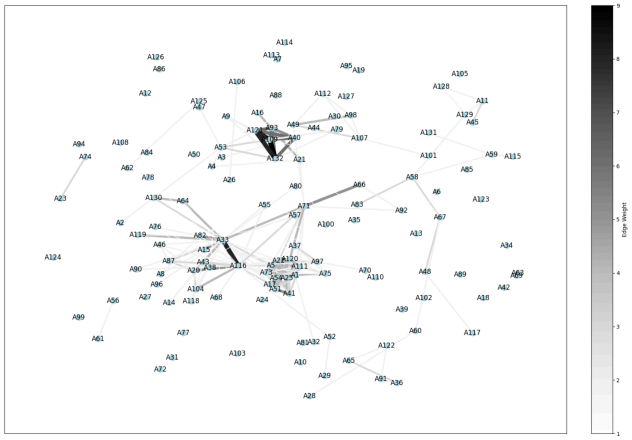


図 12　実験 2-1 Step4 のグラフ

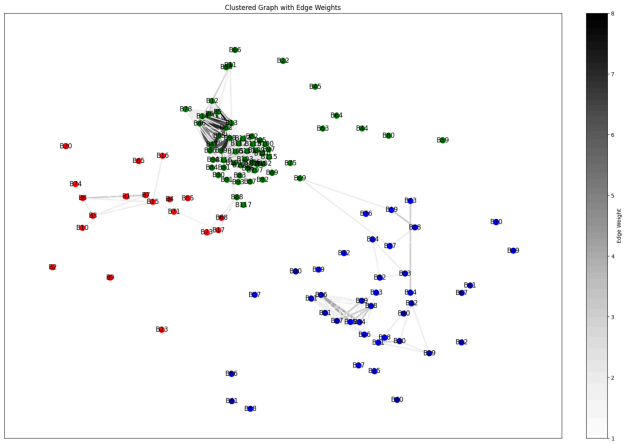


図 13　実験 2-2 交グラフ

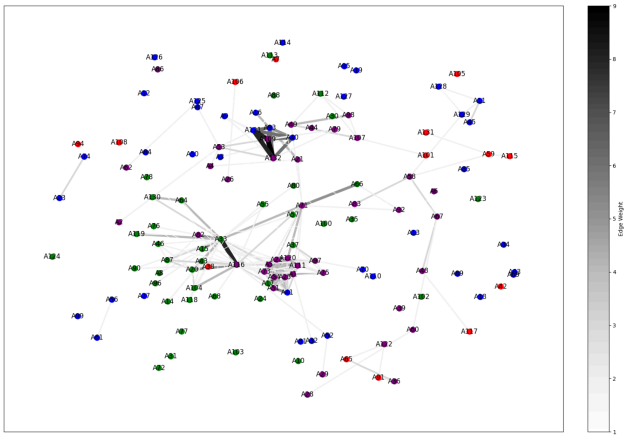


図 14　実験 2-3 元に戻したグラフ

の彩色を元のグラフで再現した．ここで，彩色ルールとして，交グラフで 2 つ以上のクラスタに属していたノードは紫で彩色する．結果が図 14 のグラフである．詳しくノードを見てみると，例えば，右上の赤いノード A105 は，データベースよりノード B74 に対応している．図 13 で B74 を確認すると，B74 は赤で彩色されており，これが正しいことがわかる．次に，右下の紫のノード A36 を見てみると，これはノード B15 と B33 に対応している．それぞれ確認すると，B15 は赤，B33 は青で彩色されており，2 つ以上のクラスタに属しているのでこの彩色も正しいことがわかる．さらに，中央の緑のノード A37 を確認すると，ノード B12 と B113 に対応している．それぞれ確認すると，B12 も B13 も緑で彩色されており，クラスタとしては 1 つのクラスタにしか属していないので，これも正しいことがわかる．

5.3　考　　察

まず，実験 1 の結果に関して，グラフの定義に立ち返ると，他のグループ群と所属するグループが被ることの多いグループ群が赤く可視化され，共通するグループが多い繋がりが青く可視化されている．すなわち，ここで可視化されている集合が所属するグループは，同じようなユーザが所属することが多く，さらにこれらの集合に属するユーザ同士は濃い繋がりを持つ可能性が高いと推測される．また，閾値 $n = 20$ でノードとエッジの数が 0 になったことから，このデータの中には，最高で 20 個のグループが被っているグループ群が存在することがわかる．

次に，実験 2 の結果に関して，交グラフをクラスタリングすることは新たなコミュニティの発見につながる．また，紫で彩色されたノードは複数のコミュニティの重なりであるため，このノードに属するグループやユーザは，今回特定を試みていた犯罪の中心となる集団やユーザである可能性が高い．

さらに，図 11 の $weight = 11$ のグラフと図 12 を比べてみると，赤が濃く可視化されたノードと紫のノードは被るものが非常に多い．このことにより，ここまでの推論の妥当性が高まったといえる．

6　おわりに

SNS を利用したサイバー犯罪抑止を目的に，交グラフを用いた新たなネットワーク分析手法を提案した．提案法では，犯罪コミュニティのデータから複雑ネットワークグラフを生成し，これを交グラフに変換，彩色することで犯罪コミュニティの中心である可能性の高い集団やユーザを特定することを試みた．さらに，提案法を実際の犯罪コミュニティのデータに適用し，ノードとエッジの削減の試行とコミュニティの抽出を行った．

今後の課題としては，抽出したコミュニティが実際に犯罪の中心となっているかどうか確認するための中心である判定をどのように行うかの提案や，クラスタリング時の手法やクラスタ数による結果の変化の考察などが挙げられる．

文　　献

[1] 一般社団法人日本クレジット協会, クレジットカード不正利用被害額の発生状況, https://www.j-credit.or.jp/information/statistics/download/toukei_03_g.pdf
（2024 年 2 月参照）
[2] 趙 智賢，長田 繁幸，SNS を経由するクレジットカード不正利用のモデル化と抑止方法の検討，研究報告セキュリティ心理学とトラスト（SPT），2022-SPT-48，No.25，pp.1-7，2022.
[3] 大原望乃，趙智賢，長田繁幸，中川直樹，小口正人，“SNS コ

ミュニケーション分析手法を用いた金融犯罪情報の早期検出に関する研究”，日本データベース学会，第 15 回データ工学と情報マネジメントに関するフォーラム（DEIM2023），5b-5-2，2023 年 3 月．
[4] Christian Konradt, Andreas Schilling and Brigitte Werners, “Phishing: An economic analysis of cybercrime perpetrators,” Elsevier, Computers & Security, Vol. 58, pp. 39-46, May 2016.’
[5] Abhinav Singh, “The Underground Ecosystem Of Credit Card Frauds,” BlackHat, Asia 2015, March 2015.
[6] Telegram, https://telegram.org/
[7] B. Luo and D. Lee, “On Protecting Private Information in Social Networks: A Proposal,” 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, pp. 1603-1606, doi: 10.1109/ICDE.2009.158., April 2009.
[8] 大原望乃，趙智賢，長田繁幸，中川直樹，小口正人，“金融産業を標的にしたサイバー犯罪抑止のための SNS コミュニケーション分析手法の研究”，情報処理学会第 85 会全国大会，2023 年 3 月．
[9] 大原望乃，趙智賢，長田繁幸，中川直樹，小口正人，“ SNS 上の犯罪グループを監視する際のプライバシ保護に関する研究 ’’，マルチメディア，分散，協調とモバイル DICOMO2023 シンポジウム，pp. 1443-1450，2023 年 7 月．
[10] 倉持俊也，岡田直樹，谷川恭平，土方嘉徳，西田正吾，“複雑ネットワークにおける交グラフと意味的解析を用いたグループ発見手法 ”，日本知能情報ファジィ学会誌，情報と知能，Vol.25, No.1，pp.540-555, 2023 年．
[11] Everett, M.G. and Borgatti, S.P., “Analyzing Clique Overlap, ’’ Connections, Vol.21, No.1, pp.49-61, 1998.
[12] 大原望乃，伊藤純菜，趙智賢，長田繁幸，中川直樹，小口正人，SNS サイバー犯罪対策に関する検討 〜交グラフを用いた犯罪グループ発見手法〜，

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年3月1日(金) 16:00 ~ 18:10 | T3-A オンライン（Zoom Events)

大規模言語モデル (2)

座長:梅本 和俊(東京大学)
コメンテータ:服部 峻(滋賀県立大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-A-9-01] Cross-lingual Relevance Estimation with Soft and Hard Prompts
*Fang Hanpei[1]、酒井 哲也[2] (1. 早稲田大学基幹理工研究科酒井研究室、2. 早稲田大学)

16:30 ~ 16:55
[T3-A-9-02] A Study on Automatic Nugget Weight Generation with LLMs
*Chang Kai-Xin[1]、酒井 哲也[2] (1. 早稲田大学基幹理工研究科酒井研究室、2. 早稲田大学)

16:55 ~ 17:20
[T3-A-9-03] 大規模言語モデルを用いたユーザの目的に基づく説明性を有するPOI推薦手法
*片山 一[1]、牛尼 剛聡[2] (1. 九州大学大学院芸術工学府　牛尼研究室、2. 九州大学大学院芸術工学研究院)

17:20 ~ 17:45
[T3-A-9-04] 特定ドメインの問合せを対象とした大規模言語モデルの性能計測手法の検討
*神田 智也[1]、伊藤 寛祥[2]、森嶋 厚行[2] (1. 筑波大学融合知能デザイン研究室、2. 筑波大学図書館情報メディア系)

17:45 ~ 18:10
[T3-A-9-05] 【技術報告】リクルートの横断検索基盤における取り組み紹介
新田洸平 (株式会社リクルート)

Cross-lingual Relevance Estimation with Soft and Hard Prompts

Hanpei FANG† and Tetsuya SAKAI†

† School of Fundamental Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
E-mail: †hanpeifang@ruri.waseda.jp, ††tetsuyasakai@acm.org

Abstract The demand for Cross-Lingual Information Access is growing rapidly. However, due to the ambiguity in translation, estimating the relevance between the query and the documents is challenging. We propose a neural-based binary relevance estimation architecture using both soft and hard prompts with limited samples based on a pre-trained multilingual BERT model. Our results show the possibility of building a cross-lingual relevance estimation system with a small model and a small training dataset.

Key words Prompt Engineering, Cross-lingual Relevance Estimation, Information Retrieval

1 Introduction

A Cross-Lingual Information Retrieval (CLIR) task is an Information Retrieval (IR) task where a query is given in one language but the system should return the documents in another language. With the exponentially growing amount of information on the Internet, the demand for cross-lingual information needs is rapidly growing around the world.

A traditional CLIR system consists of two components: machine translation and monolingual information retrieval [1] [2], aiming to translate the query and the documents into the same language and convert the CLIR problem to a monolingual IR problem. However, the quality of the translation system is limited by the lexical ambiguity [3] between languages, which constrains the performance of translation-based CLIR systems.

How to estimate the relevance between query and document is an important topic in all IR tasks, however, it is particularly challenging in cross-lingual cases. In CLIR tasks, the relevance can be estimated through translation approaches, including a dictionary, a machine translation system, or other similar resources. Some researchers propose a probabilistic method [4] which does not rely on either query translation or document translation two decades ago. Recently, the research community has been focusing on constructing cross-lingual relevance matching models by fine-tuning pre-trained language models, such as BERT [5], which outperforms traditional statistical methods.

However, fine-tuning methods typically rely on large parallel datasets. For example, in Jiang et al.'s work [5], they extract about 54 million training samples from the parallel data to fine-tune a pretrained BERT model. Constructing such a dataset is resource-consuming. Furthermore, fine-tuning a pre-trained model is not only a difficult task [6] but also not very energy- and space-efficient.

To solve this problem, we propose a neural-based method to automatically estimate the relevance between query and documents in a cross-lingual scenario by tuning prompts for a pre-trained multilingual BERT (mBERT) model [7], a relatively small model, with only a limited number of samples. Both soft and hard prompts are involved in our architecture. The soft prompts are trainable, while the hard prompts are human-designed. These two kinds of prompts are used for different purposes. The soft prompt instructs the model on what should be done, while the hard prompts indicate the positions of the query and document. We call them **main prompt** (soft) and **assisting prompt** (hard), respectively.

2 Related Work

2.1 Cross-lingual Relevance Estimation

In monolingual IR tasks, the relevance can be estimated by probabilistic methods such as BM25 [8]. In CLIR tasks, probabilistic methodologies can be applied as well by introducing a probabilistic translation dictionary [9] or neural network translation models [10].

In recent years, multilingual contextual pre-trained language models, such as mBERT [7], XLM [11], mT5 [12], have proved useful for many multi-lingual tasks, including cross-lingual information retrieval [5] [13]. Estimating the relevance with a pre-trained multi-lingual language model directly is another possible way.

Jiang et al. [5] fine-tuned a pre-trained multilingual BERT model with a large homemade multilingual parallel dataset. They used the query and the sentence in different languages as the input, and the output embedding of the [CLS] token is fed into a single-layer feed-forward neural network to predict

the relevance. Wang et al. [14] encoded the query and document by a mBERT model respectively and calculated their cosine similarity to predict the relevance. Faggioli et al. [15] proposed a method to judge the relevance by prompting a Large Language Model (LLM) prompt in a monolingual scenario [16], but they have not examined the effectiveness in cross-lingual tasks and LLMs can be computationally heavy and energy-consuming.

2.2 Parameter-Efficient Tuning

As mentioned in the previous section, current natural language processing (NLP) research often makes use of a technique called transfer learning [17], where a model is pre-trained on a data-rich task before being fine-tuned on a downstream task. These pre-trained models have been widely used in various tasks, including text classification, named entity recognition, question answering, and information retrieval [5] [13] [18].

However, fine-tuning a pre-trained model is a challenging task [6], particularly with a small dataset. When fine-tuning a pre-trained language model, a phenomenon called catastrophic forgetting could happen if the dataset used is apparently smaller than the pre-training dataset, which means that the model could forget what it learned from the huge pre-training dataset. This phenomenon causes a serious problem in that the model cannot utilize the knowledge it learned from the pre-training dataset anymore, and dramatically lowers its performance. Some other problems, such as overfitting and insufficient adaptation, can reduce performance and make training difficult as well.

Furthermore, fine-tuning a pre-trained model is not very energy- and space-efficient, especially when we need to fine-tune the model for every single case. For example, in cross-lingual tasks, every single pair of languages probably needs a unique model.

Therefore, how to use the models efficiently has been a crucial topic. Houlsby et al. [19] proposed transfer with adapter modules. In their method, rather than tuning the entire model, they inserted adapter modules into the transformer layers of the pre-trained BERT model. In their results, the performance is close to the vanilla fine-tuned BERT, while the number of trainable parameters decreases to only around 3%.

Brown et al. [16] showed that prompting a frozen GPT-3 model is very effective. In their work, prompts are the human-designed text instruction (discrete prompts) of a task description and/or several canonical examples. Li and Liang [20] proposed a method called prefix tuning. Unlike vanilla fine-tuning, they only optimize the "prefix" of each layer as continuous word embeddings. Lester et al. [21] proposed prompt tuning, which can be seen as a simplified version of prefix tuning by only optimizing the input continuous prompts (soft prompts).

All of these approaches have the advantage that it is unnecessary to store the entire fine-tuned model for every subtask while reducing the training cost, which dramatically improves time and space efficiency. These methods are particularly suitable for small datasets since they prevent the occurrence of catastrophic forgetting.

3 Approach and Experiments

3.1 Approach

We propose a method of cross-lingual relevance estimation by using the pre-trained multilingual BERT model. Instead of vanilla fine-tuning, we use prompt tuning [21] in our method. For each language pair, we tune our prompts with only limited samples. The numbers of relevant and irrelevant samples are equal in our experiments. Furthermore, to fully utilize the knowledge learned from the pre-training stage, we instruct the model to predict a [MASK] token and feed the output into a single-layer feed-forward neural network to estimate the relevance.

Instead of using only soft prompts, we combine both soft and hard prompts. We call them **main prompt** and **assisting prompt**. We initialize our **main prompt** from a handcrafted hard prompt ("**(true or false)**" in our experiments) and pad it to a length of 10 by adding additional ones. We add this **main prompt** between the [CLS] and [MASK] tokens. We use the **assisting prompts** to instruct the model on the positions of the **Query** and the **Document**. Therefore, we add them before the query and document tokens. The details of our architecture can be seen in Figure 1.

The model is instructed to predict the masked token. We only consider the probabilities of the two tokens, **true** or **false**, in our approach, which is the reason why we initialized the **main prompt** from **(true or false)**. The output where the probability of **true** is greater than the probability of **false** indicates that the document is relevant to the query, and vice versa. Figure 2 shows a sample of our **main prompt**. The left hand side is the initial prompt, which is "**(true or false)**" (5 tokens) and 5 adding additional ones (tensors). The right hand side is the prompt after tuning. In a BERT model [7], each word is converted to a 768-dimensions tensor, so each token of our prompt has 768 dimensions.

3.2 Dataset

The dataset used in the experiments is the MULTI-8 dataset in CLIRMatrix collection [22], which is a multilingual dataset of queries and documents jointly aligned in 8 different languages. In the experiments, English is our query language, while French, Spanish, German, Chinese, Japanese,

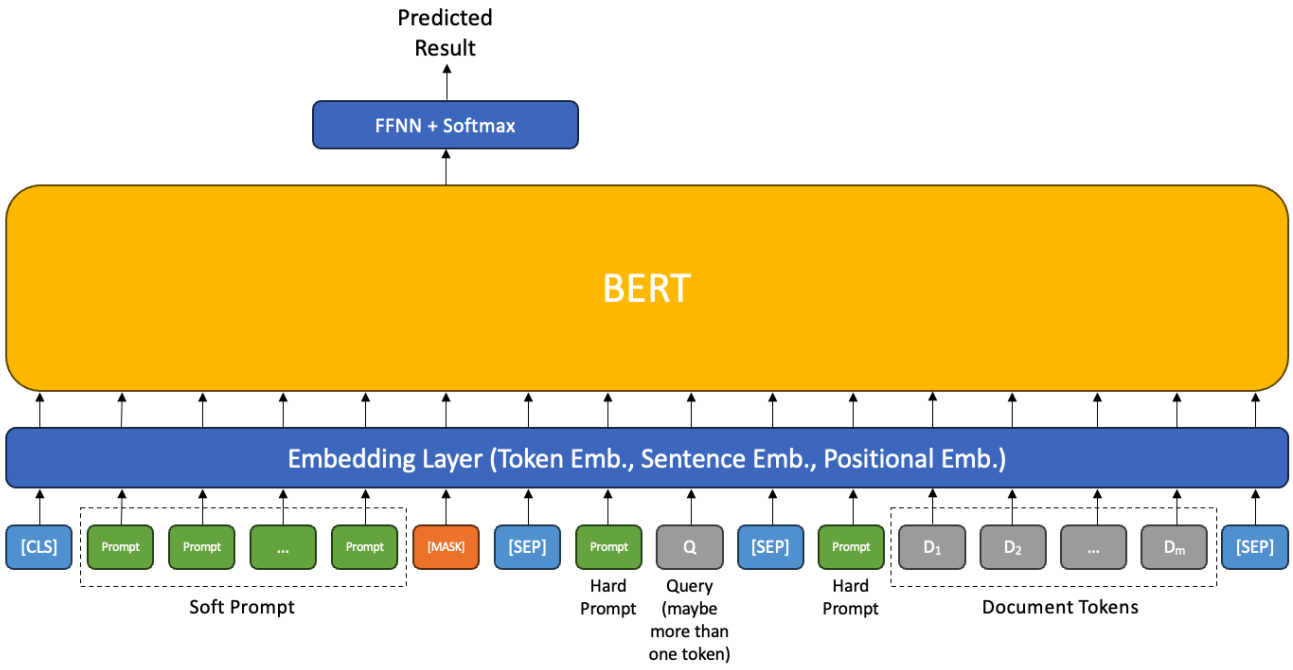


Figure 1 Prompt tuning cross-lingual relevance estimation model architecture.

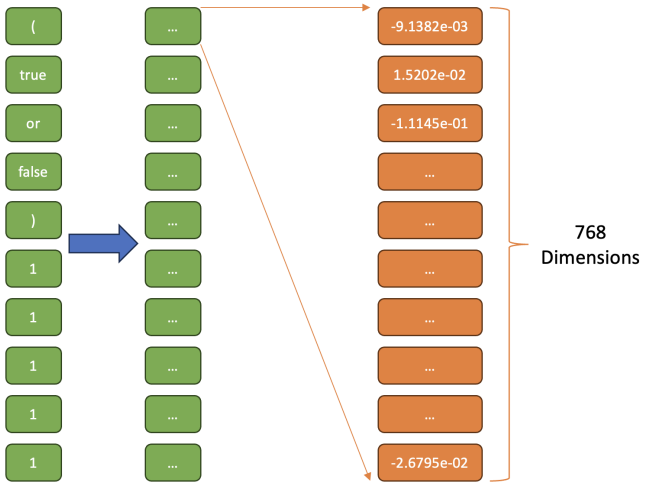


Figure 2 A sample of main prompt.

Table 1 Sample data of English query and Japanese documents. The query language is English, and the document language is Japanese.

Query	Relevance	Document
Finland	Relevant (6)	フィンランド共和国 Suomen tasavalta（フィンランド語） Republiken Finland（スウェーデン語） 国の標語：なし 国歌：Maamme（フィンランド語） Vårt Land（スウェーデン語） 我等の地 ...
	Irrelevant (0)	国民体育大会バレーボール競技（だい 67 かい こくみんたいいくたいかい バレーボールきょうぎ）は、2012 年 10 月に岐阜県で開催された第 67 回国民体育大会のバレーボール競技大会である...

and Arabic are used as document languages.

In the Multi-8 dataset, the relevance score between query and document is rated discretely between **0** and **6**. In our experiments, we only consider **0** and **6** for simplicity. We consider **6** as *relevant*, while **0** as *irrelevant* in our experiments. For those queries whose search engine results page (SERP) do not contain **0** and/or **6** results, we excluded them to ensure consistency. A sample of the dataset is shown in Table 1.

3.3 Experiment Settings

We used a single Nvidia RTX 3090 GPU is used for tuning the prompt in our experiments. For each pair of query and document languages, the model is tuned in 5, 10, and 15 shots, respectively. Adam optimizer [23] is used in our training process. The batch size is 16 in all experiments. Our loss function is Cross Entropy Loss [24], which only considers the [MASK] token, and only the **main prompt** is trainable. All random seeds are fixed to 0 to make sure the results are reproducible.

3.4 Evaluation Metrics

We described the details of our evaluation metrics in this section.

In our experiments, we used **Accuracy** and **Macro F1 Scores** [25] [26] for the evaluation. The **Accuracy** is the percentage of correct results. We collected the **Overall**, **Relevant** (for relevant documents), and **Irrelevant** (for irrelevant documents) accuracy for each of the experiments, so

that we can examine bias in performance towards different types of documents. Macro F1 scores [25] [26] is the harmonic mean of F1 score of different classes, which can be used to evaluate a binary or multi-class system. There are two definitions of Macro F1 Scores in Opitz and Burst's literature [25], $F1^M$, and Harmonic mean of Macro-averaged Precision and Macro-averaged Recall (*HMPR*) [26].

The formulae of $F1^M$ and *HMPR* are given below. In the formulae, C is the set of ordinal classes, i denotes a single class, and c_i is the number of items in class i. Let $C^+ = \{i \in C | c_i > 0\}$. $Prec_i$ and Rec_i denote the precision and recall of class i, and $Prec^M$ and Rec^M denote macro precision and macro recall, which are the average precision and recall across all classes C^+. That is, $F1^M$ averages class-wise F1 scores over the classes, while *HMPR* is the F1 score based on Macro-averaged Precision and Recall.

$$F1^M = \frac{1}{|C^+|} \sum_{i \in C^+} f1(Prec_i, Rec_i) \quad (1)$$

$$HMPR = f1(Prec^M, Rec^M) \quad (2)$$

$$Prec^M = \frac{1}{|C^+|} \sum_{i \in C^+} \frac{True\ Positive_i}{True\ Positive_i + False\ Positive_i} \quad (3)$$

$$Rec^M = \frac{1}{|C^+|} \sum_{i \in C^+} \frac{True\ Positive_i}{True\ Positive_i + False\ Negative_i} \quad (4)$$

4 Results and Discussions

The accuracy, $F1^M$ and *HMPR* of each language pair can be seen in Table 2. Each language pair is tuned with 5, 10, and 15 shots, respectively. In our experiments, *n-shots* means n relevant samples and n irrelevant samples are used during tuning. The results without tuning can be seen in Table 3, which is the *zero-shot* case. In this case, we use the same experiment settings without tuning the **main prompt**.

From the experiment results in Table 2, we can see that in most cases, the performance is improved as more instances are used in tuning. However, the increase in the number of sample instances may lead to a decrease in performance in some cases. Particularly, the performance and bias of the English-Spanish pair drop as more instances are used. We suspect that the quality and/or bias of the tuning instances could be a possible factor.

We also tested the *0-shot* case and the results are shown in Table 3. By comparing the results, we can see that the performance of each language pair has been largely improved after tuning. Without tuning, the model gives a quite biased result in some cases, such as Arabic and Japanese, which substantially affects the $F1^M$ and *HMPR* scores.

Furthermore, it is intuitively expected that when the Query Language and the Document Language are similar, the system tends to have a better performance. For example, Spanish achieves 0.8 for both $F1^M$ and *HMPR* scores in the case of 5-shots, which outperforms its Chinese counterpart. However, intriguingly, Arabic surpasses German in performance, despite both German and English stemming from the Germanic language family. We suppose the composition of the pre-training corpus influencing performance could be a potential factor.

Table 2 Accuracy and Macro F1 Scores for each language (Document) with different numbers of sample shots. The query language is English. The results are rounded to four decimal places. All languages are labeled with ISO 639-1 code.

Lang.	Shots	Accuracy			$F1^M$	*HMPR*
		Overall	Relevant	Irrelevant		
ar	5	0.7402	0.7177	0.7626	0.7400	0.7404
	10	0.7377	0.7248	0.7505	0.7376	0.7377
	15	0.7683	0.7532	0.7833	0.7682	0.7684
de	5	0.6933	0.7188	0.6678	0.6931	0.6935
	10	0.7107	0.7278	0.6936	0.7106	0.7108
	15	0.7578	0.7106	0.7108	0.7577	0.7580
es	5	0.8001	0.8100	0.7901	0.8000	0.8001
	10	0.7705	0.8683	0.6727	0.7683	0.7758
	15	0.6794	0.7805	0.5782	0.6760	0.6832
fr	5	0.7867	0.7631	0.8103	0.7866	0.7870
	10	0.7899	0.7979	0.7820	0.7899	0.7900
	15	0.7912	0.8614	0.7211	0.7902	0.7941
ja	5	0.7393	0.7380	0.7407	0.7393	0.7393
	10	0.7083	0.7428	0.6739	0.7080	0.7088
	15	0.7372	0.8454	0.6290	0.7341	0.7430
ru	5	0.7214	0.8797	0.5631	0.7142	0.7335
	10	0.7515	0.8059	0.6971	0.7508	0.7530
	15	0.7576	0.8708	0.6443	0.7544	0.7645
zh	5	0.6652	0.6758	0.6545	0.6651	0.6652
	10	0.6927	0.7165	0.6688	0.6925	0.6929
	15	0.7368	0.7694	0.7041	0.7365	0.7373

Table 3 Accuracy and Macro F1 Scores for each language (Document) without tuning. The query language is English. The results are rounded to four decimal places. All languages are labeled with ISO 639-1 code.

Lang.	Accuracy			$F1^M$	*HMPR*
	Overall	Relevant	Irrelevant		
ar	0.5188	0.1560	0.8816	0.4459	0.5290
de	0.5391	0.6714	0.4069	0.5309	0.5406
es	0.5437	0.6753	0.4120	0.5356	0.5453
fr	0.5721	0.7468	0.3974	0.5586	0.5771
ja	0.4703	0.7417	0.1989	0.4282	0.4640
ru	0.6018	0.4150	0.7886	0.5874	0.6099
zh	0.5546	0.8044	0.3048	0.5250	0.5635

5 Ablation Experiments

For ablation experiments, we removed the **assisting prompt** (hard) and tested the performance with the modified structure shown in Figure 3. In this modified structure, the absence of hard prompts is the only difference, and other parts remain unchanged. The experiment settings are consistent with our previous experiments, and the results are shown in Table 4.

Table 4 Accuracy and Macro F1 Scores for each language (Document) with different numbers of sample shots without hard prompts. The query language is English. The results are rounded to four decimal places. All languages are labeled with ISO 639-1 code.

Lang.	Shots	Accuracy			$F1^M$	*HMPR*
		Overall	Relevant	Irrelevant		
ar	5	0.7256	0.7068	0.7444	0.7255	0.7258
	10	0.7239	0.8036	0.6441	0.7221	0.7268
	15	0.7647	0.7149	0.8145	0.7641	0.7660
de	5	0.6532	0.8118	0.4946	0.6442	0.6617
	10	0.7119	0.7274	0.6965	0.7119	0.7120
	15	0.7471	0.7747	0.7195	0.7469	0.7475
es	5	0.6607	0.8434	0.4780	0.6490	0.6729
	10	0.6846	0.8162	0.5530	0.6790	0.6914
	15	0.6792	0.7811	0.5772	0.6758	0.6830
fr	5	0.7322	0.7435	0.7210	0.7322	0.7323
	10	0.7917	0.7972	0.7862	0.7917	0.7917
	15	0.7916	0.8593	0.7239	0.7906	0.7943
ja	5	0.7357	0.7435	0.7279	0.7357	0.7357
	10	0.7063	0.7544	0.6583	0.7056	0.7073
	15	0.7353	0.8355	0.6351	0.7326	0.7402
ru	5	0.7214	0.8797	0.5631	0.7142	0.7335
	10	0.7515	0.8059	0.6971	0.7508	0.7530
	15	0.7575	0.8672	0.6478	0.7546	0.7640
zh	5	0.6629	0.6522	0.6737	0.6629	0.6630
	10	0.6979	0.7388	0.6570	0.6974	0.6986
	15	0.7447	0.7802	0.7091	0.7443	0.7453

According to Bella et al.'s dataset [27], we can find the lexical similarity between English and the other seven languages, which is shown in Table 5.

Table 5 Lexical similarity between languages.

Language 1	Language 2	Similarity
English	Arabic	0.189
English	German	4.720
English	Spanish	7.907
English	French	9.665
English	Japanese	1.786
English	Russian	2.025
English	Chinese	0.012

By comparison between the results in Table 2 and Table 4, we can find that some language pairs suffer a large drop in both accuracy and **Macro F1** scores (German, Spanish, and French), while other languages did not show apparent differences, especially in the case of 5-shots. With the data shown in Table 5, we can make a reasonable assumption that languages similar to English, German (4.720), Spanish (7.907), and French (9.665), tend to be affected greatly by the absence of hard prompts since the hard prompts are written in English, which shows that the language similarity does affect the performance of neural language models, especially when the training data is insufficient.

6 Multilingual Case

We tested a multilingual case as well, to discover the possibility of constructing a multilingual information retrieval system with pre-trained language models. In this experiment, we used English as the query language and used the documents in all other seven languages simultaneously. We used 4 samples, including relevant and irrelevant cases, for each language pair, so there were 28 training samples in total. As shown in Table 6, the accuracy of all language pairs is more biased compared with their counterparts with hard prompts.

Table 6 Multilingual Case: Accuracy and Macro F1 Scores for each language (Document) with 4 shots in each language. The query language is English. The results are rounded to four decimal places. All languages are labeled with ISO 639-1 code.

Lang.	Accuracy			$F1^M$	*HMPR*
	Overall	Relevant	Irrelevant		
ar	0.5188	0.1560	0.8816	0.5525	0.5578
de	0.5391	0.6714	0.4069	0.7439	0.7471
es	0.5437	0.6753	0.4120	0.6596	0.6601
fr	0.5721	0.7468	0.3974	0.6279	0.6281
ja	0.4703	0.7417	0.1989	0.6939	0.6974
ru	0.6018	0.4150	0.7886	0.7224	0.7247
zh	0.5546	0.8044	0.3048	0.6925	0.6960

7 Conclusions and Future Work

In conclusion, we explored the possibility of leveraging knowledge from the pre-trained multilingual BERT model to establish a cross-lingual relationship between queries and documents with a limited number of samples, even though the mBERT is not pre-trained on an exactly parallel corpus [7]. We conducted experiments on the CLIRMatrix Multi-8 dataset. We find more training samples usually lead to better performance, and we find the language similarity does affect the performance of neural language models, especially when the training data is insufficient.

One limitation of the present study is that our experiments

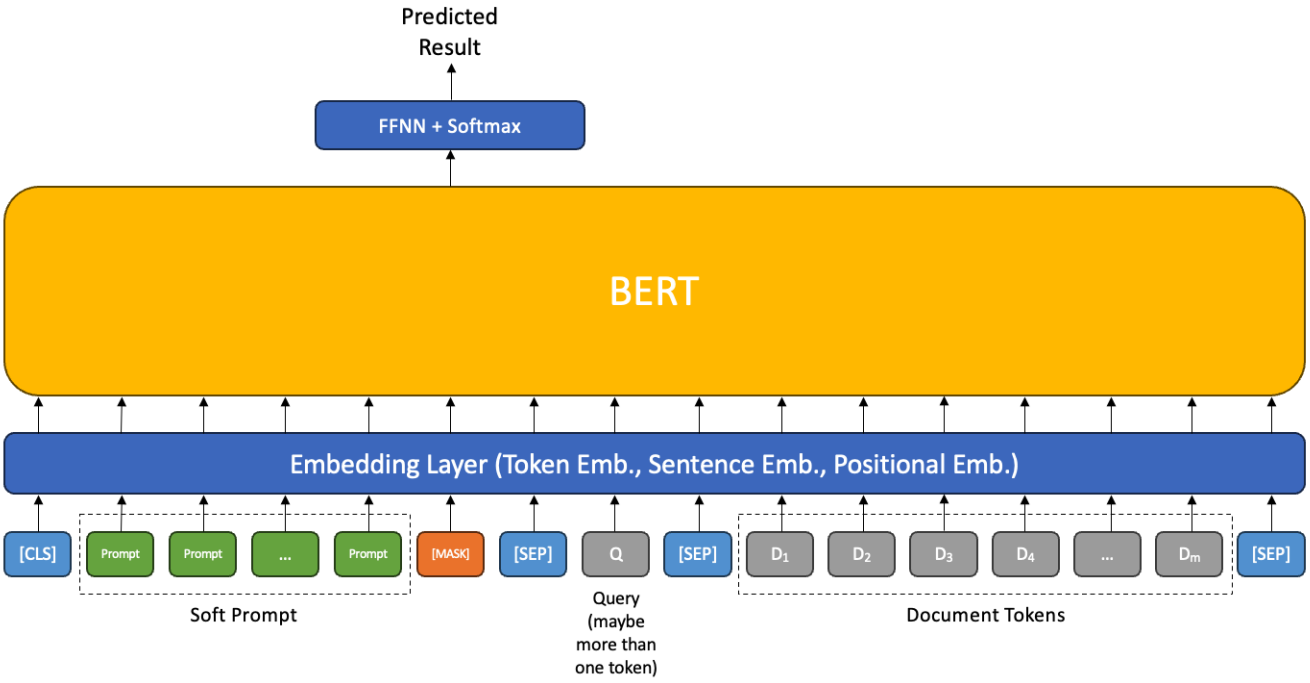


Figure 3 Prompt tuning cross-lingual relevance estimation model architecture without hard prompts.

lack an established baseline, as there is no prior art that tackles our task on the dataset to the best of our knowledge. Our ablation experiments show that the hard prompt contributes more to the language pairs which are similar to the written language of the hard prompt itself, especially with insufficient training data.

Given the increasing demand for multilingual applications, many researchers have turned their focus to enhancing the cross-lingual transferability of pre-trained language models. Pan et al. [28] align embeddings in different languages on both word and sentence levels by using parallel data. Cattan et al. [29] proposed a method for improving the cross-lingual transferability of low-resource languages by using prototypical representations. By applying these approaches, we expect the performance in cross-lingual relevance estimation to be improved, but further investigation is required.

While binary cross-lingual relevance estimation represents the initial phase of our research, we aim to extend our approach to graded cross-lingual relevance estimation and use it as a re-ranker in the next stage. Our goal is to build a CLIR system that operates efficiently without relying heavily on vast amounts of parallel training data or the need for Large Language Models.

Acknowledgement

This research was undertaken with the support from The Real Sakai Laboratory, Waseda University. I would like to express my thanks of gratitude to the members in The Real Sakai Laboratory, who helped me and gave me many useful pieces of advice for continuing the research. Also, I would like to thank Professor Tetsuya Sakai, who gave me this great opportunity and lots of support to research this topic.

References

[1] Jian-Yun Nie. *Cross-language Information Retrieval*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers, 2010.

[2] Kazuaki Kishida and Kuang-hua Chen. *Experiments on Cross-Language Information Retrieval Using Comparable Corpora of Chinese, Japanese, and Korean Languages*, pages 21–37. Springer Singapore, Singapore, 2021.

[3] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. Translation techniques in cross-language information retrieval. *ACM Comput. Surv.*, 45(1), dec 2012.

[4] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[5] Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France, May 2020. European Language Resources Association.

[6] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the As-*

sociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009.

[9] Jinxi Xu and Ralph Weischedel. Cross-lingual information retrieval using hidden Markov models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 95–103, Hong Kong, China, October 2000. Association for Computational Linguistics.

[10] Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, et al. Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654, 2019.

[11] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[12] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[13] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019.

[14] Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. Adversarial domain adaptation for cross-lingual information retrieval with multilingual bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3498–3502, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment, 2023.

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[17] Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[22] Shuo Sun and Kevin Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, November 2020. Association for Computational Linguistics.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[24] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.

[25] Juri Opitz and Sebastian Burst. Macro f1 and macro f1, 2021.

[26] Tetsuya Sakai. Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769, Online, August 2021. Association for Computational Linguistics.

[27] Gábor Bella, Khuyagbaatar Batsuren, and Fausto Giunchiglia. A database and visualization of the similarity of contemporary lexicons. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, page 95–104, Berlin, Heidelberg, 2021. Springer-Verlag.

[28] Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online, June 2021. Association for Computational Linguistics.

[29] Oralie Cattan, Sophie Rosset, and Christophe Servan. On the cross-lingual transferability of multilingual prototypical models across NLU tasks. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 36–43, Online, August 2021. Association for Computational Linguistics.

A Study on Automatic Nugget Weight Generation with LLMs

Kai-Xin CHANG† and Tetsuya SAKAI†

† School of Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169–0072, Japan
E-mail: †victorchang@toki.waseda.jp, ††tetsuyasakai@acm.org

Abstract The automation of nugget weight assessment can be considered a crucial part of implementing nugget-based assessment systems, such as the Schematised Weighted Average Nugget (SWAN) framework. In this work, we study the possibility of leveraging large language models (LLMs) to generate nugget scores automatically. Our experiment results showed that there is merely none to slight or fair inter-rater agreement between GPT-4/GPT-3.5 and human assessors.

Key words LLM, conversational system, chatbots, conversational search, evaluation, evaluation measures, nuggets, nugget weight generation, prompt engineering

1 Introduction

In the era of rapidly evolving large language models, we are facing an ongoing revolution in information access; while most of us are familiar with traditional search engines, we are now provided a novel way of accessing information: AI chatbots powered by pre-trained state-of-the-art large language models (LLMs). Such chatbots are often textual conversational systems that allow natural language interaction between the user and the powering language model. Instead of a conventional search engine result page (SERP), users with information needs can now receive a textual response regarding their query of interest.

The evaluation of textual conversational systems requires distinct methodologies compared to those employed for conventional search engines and the corresponding search engine results pages (SERPs) generated by traditional information retrieval systems. In this context, a prospective evaluation approach for contemporary AI chatbots could involve the adoption of a nugget-based framework, which accounts for their aptness in terms of both textual and conversational characteristics.

The automation of nugget weight assessment can be considered a crucial part of implementing nugget-based assessment systems, such as the Schematised Weighted Average Nugget (SWAN) framework [1]. In this work, we study the possibility of leveraging large language models (LLMs) to generate nugget scores automatically.

2 Related Work

2.1 Nugget-based Test Collection

Past NTCIR tasks such as One Click Access (1CLICK) [2, 3] and MobileClick [4, 5] were designed to assess concise and informative informational retrieval systems. The pilot NTCIR-9 1CLICK [2] task aims to provide users with a textual output instead of a conventional search engine result page (SERP), allowing direct and immediate information access. NTCIR-9 1CLICK included Japanese queries and nuggets exclusively, where nuggets were initially defined as a 5-tuple containing information including nugget ID, weight, semantics, vital string, and URL. The primary assessment for 1CLICK tasks was the S-measure [6], which mainly factors in nugget weight assessed by Japanese task organizers and participants as well as the nugget's offset position in the system output. In the NTCIR-10 1CLICK-2 [3] task, iUnits were introduced to solve the ambiguity in length and dependency of nuggets. Nuggets can be broken into iUnits that are relevant, atomic, and dependent pieces of information. Please note that in this work the terms "nugget" and "iUnit" are used interchangeably and can be considered the same concept, as we concentrate on the automation of weight generation based on queries and nugget/iUnit context.

As a sequel to the 1CLICK tasks, the NTCIR-11 MobileClick tasks contain two subtasks, iUnit retrieval and summarization subtasks, and participating systems are expected to output a concise summary of information relevant to a given query and to provide immediate and direct information access for mobile users [4]. Similar to the definition in NTCIR-10 1CLICK-2, iUnits are defined as relevant, atomic, and dependent pieces of information; for the English task, the weight of each English iUnit was assigned by the original extractor of the iUnit.

In 2023, Sakai proposed the Schematised Weighted Average Nugget (SWAN) framework [1], a generic framework for

auditing textual conversation systems. A SWAN (Schematised Weighted Average Nugget) is computed based on different criteria, such as fluency, correctness, conciseness, and so on, where a unique score can be calculated for each nugget and each criterion, providing a versatile evaluation of the conversational behavior of a given system. It is also worth mentioning that the list of 21 potential criteria for evaluating textual conversational systems presented in the paper is not yet final.

In our previous technical report, we evaluated LLM-based textual conversational systems for direct and immediate information access [7], with S-measure [6] as our main evaluation measure for conciseness as well as weighted recall to reflect the characteristics of an information retrieval system. Our main goal was to examine two of the criteria in SWAN, correctness and conciseness, by investigating the outputs of textual conversational systems powered by LLMs, such as OpenAI's ChatGPT. Among the four query categories, CELEBRITY, LOCAL, DEFINITION, and QA, results showed that ChatGPT (注1) answered CELEBRITY queries most concisely with the highest S-measure metric score while providing the most informative answers for QA queries with the highest weighted recall. Despite conversational AI bots' overwhelming popularity nowadays, we also observed conversational behavior that may hinder the users' direct and immediate information needs, such as providing incorrect nuggets, showing apologies, or hedging. As an early attempt to evaluate modern textual conversational systems with nugget-based measures, there remain unfinished tasks such as the implementation of automatic or semi-automatic nugget extraction, or designing evaluation methodologies for other criteria.

2.2 Inter-rater Agreement

Often symbolized by the Greek letter κ (kappa), Cohen's kappa is one of the most employed measures for evaluating inter-rater agreement [8]. As a coefficient of agreement for nominal scales, κ takes into consideration both the expected results and observed agreement from two independent sources and calculates the degree of agreement by subtracting the possibility of disagreement. Cohen's kappa is originally designed to evaluate the degree of agreement between two judges for tasks using nominal categories.

To generalize this statistic to cope with tasks with multiple ordinal categories, Cohen introduced weighed kappa [9] and extended Cohen's kappa by assigning pre-defined weights to the confusion matrix, providing a ratio scaling of agreement and disagreement. Cohen's weighted kappa reflects larger differences between inconsistent ordinal scales, penalizing more serious disagreements. Possible ways of introducing weights to the confusion matrix include linear weight assignment, where the distance between each ordinal category is evenly spaced, or quadratic weight assignment, where the penalty for more significant disagreement between two raters is magnified exponentially. Cohen's weighted kappa is a natural extension of Cohen's (unweighted) kappa and can be simplified to κ when all entries in the confusion matrix are given the same weight.

In 2021, Sakai compared and discussed different evaluation measures for ordinal classification (OC) and ordinal quantification (OQ) tasks [10]. Measures that consider the ordinal nature of OC tasks, such as Mean Absolute Error (MAE) [11], Closeness Evaluation Measure (CEM) [12], Cohen's weighted κ [9], and Krippendorff's α [13], are often used to assess the quality of OC systems, among which Sakai recommended the use of Cohen's weighted κ as the primary evaluation measure under ordinary conditions.

2.3 Automated Relevance Labelling

Thomas et al. presented in their recent work the possibility of leveraging large language models for generating relevance labels for information retrieval tasks at a large scale [14].

The authors designed the prompts with different variations, where each prompt variation can be generated based on a basic template. The prompt features five variable parts: role, description, narrative, aspects, and "multiple" judges. A total of 32 combinations of prompts were generated from the five prompt features, and the authors compared the performances between each combination mainly based on mean absolute error (MAE) and Cohen's κ.

The results showed that some prompt variations led to performances that compare favorably to those done by crowd workers in prior research by Damessie et al. in 2017 [15], where some of the prompts outperformed two groups of trained human assessors in Cormack et al's paper in 1998 [16].

However, the authors noted that there remain uncertainties when leveraging LLMs for relevance label assessment. For example, the same prompt structure with slight changes in wording may lead to substantial changes in Cohen's κ. Therefore, it is sensible to explore new variations by rephrasing and other alternative modifications to fine-tune the original prompt.

3 Experiment

As the first step of this study, we conducted a semi-automatic trial to test the effectiveness of ChatGPT. The first trial is not fully automated because we manually extract the nugget scores from the responses, while the large language model generates the answers automatically. The

(注1)：GPT-3.5; August 2023

complete process requires an in-depth design of the prompts as well as a stable connection with the API, followed by well-written source code to collect, analyze, and evaluate the results obtained from the responses. To understand the core of this automation, which is the quality of the answers and the performance of the language models, we created a prompt for each of the 50 queries and used the ChatGPT web application to generate an answer for each prompt. Inspired by Thomas et al's attempt to generate relevance labels with large language models [14], we also plan to create prompts with several variations and illustrate the impact of different prompt features on nugget weight generation.

In this current work, our prompt is based on a fixed template for each query-nugget list pair. Our intuitive prompt includes an instruction with initialization of the LLM's role, a description of the query, and an exhaustive list of all nuggets followed by a specification of the output format. Listing 1 shows the instruction part of the prompt we used.

Listing 1 Prompt: initial instruction

```
You are a nugget weight assessor. Nuggets are
    defined as relevant, atomic, and dependent
    pieces of information.
Given a query and a list of relevant nuggets,
    you must provide a weight to each nugget on
     an integer scale of 1 to 3 with the
    following meanings:
3 = very important and helpful
2 = partly important and helpful
1 = not important or helpful
```

Following the initial instruction is the description of the query, where [QUERY] is replaced by the actual query string from the test collection, such as "java vs python text processing" (MC-E-0001). The nugget list of each query is then presented below the query description. Despite the varying number of nuggets in the list of each query, we present the exhaustive list of all nuggets in our first intuitive prompt, as shown in Listing 2.

The exact number of nuggets for each query varies from 13 to 255 in the test collection. Listing 3 shows the remaining part of the prompt, which is a short description of the desired output format with an example, followed by the leading square bracket showing a more explicit indication.

In our work, we utilized the test collection from the legacy NTCIR-11 MobileClick task [4]. The original dataset consisted of queries, iUnits, and documents; we only make use of the original queries and iUnits. There are 50 Japanese queries and English queries respectively. As our current experiments mainly focus on assessing the ChatGPT ' s capability of generating weights for nuggets written in English, we only use the 50 English queries.

Listing 2 Prompt: query description and nugget list

```
Query: A person has typed "[QUERY]" into a
    search engine.
Consider the following list of nuggets:

--- START OF NUGGETS ---
1. Nugget 1
2. Nugget 2
3. Nugget 3
...
--- END OF NUGGETS ---
```

Listing 3 Prompt: output specification

```
Produce an array that contains the scores of
    each nugget in the list. Example: [2, 1, 3,
     2, 1, ...
Results: [
```

We have experimented with two versions of OpenAI's GPT model provided on ChatGPT, GPT-3.5 (free tier) and GPT-4 (paid subscription). A new chat is created for every new prompt to ensure the independency between each response. We manually examined the responses and extracted the array with the weights of the provided nugget list in the prompt. Afterward, we compared the obtained weights with the gold-standard weights assigned by the original assessor of the task. Our experiment mainly focuses on illustrating the degree of agreement between GPT and the gold assessor; therefore, we selected Cohen's weighted κ [9] as our main evaluation measure as suggested by Sakai [10].

4 Results

4.1 Inter-rater Agreement

We follow Sakai's recommendation [10] and select Cohen's weighted κ as our primary measure, as our experiments center upon evaluating the degree of agreement between two raters, namely gold standard labelers (original task assessors) and test subjects (LLM; GPT-3.5 or GPT-4).

To our disappointment, our first attempt using GPT-3.5 did not yield satisfactory results; among the 50 responses we received using the prompts generated from 50 queries and nuggets lists, we were only able to extract 28 successful answers that matched the correct number of nuggets in the list, failing to retrieve correct results from the remaining 22 responses. Our next attempt with GPT-4 yielded 31 successful results; despite a slight increase from GPT-3.5, we could not claim that our prompt is sufficient for generating nugget scores in general.

Moreover, Cohen's weighted kappa [9] calculated from the results and the gold nugget weights showed low agreement between the original assessors and both LLMs. Figure 1 shows the histogram of the distribution of the results; we interpret κ as follows: negative values indicating no agreement, $0 < \kappa \leqq 0.20$ indicating none to slight agreement, $0.20 < \kappa \leqq 0.40$ as fair agreement, $0.40 < \kappa \leqq 0.60$ as moderate agreement, $0.60 < \kappa \leqq 0.80$ as substantial agreement, and $0.80 < \kappa \leqq 1.00$ as almost perfect agreement, as shown in the following table [17].

Table 1 κ and degree of agreement

Range	Agreement
$\kappa \leqq 0$	no agreement
$0 < \kappa \leqq 0.20$	none to slight
$0.20 < \kappa \leqq 0.40$	fair
$0.40 < \kappa \leqq 0.60$	moderate
$0.60 < \kappa \leqq 0.80$	substantial
$0.80 < \kappa \leqq 1.00$	almost perfect

The Cohen's weighted kappa computed from our obtained results mainly fell between -0.2 and 0.4, mainly showing none to slight or fair agreement between GPT-4/GPT-3.5 and human assessors. Only one list of weights generated by GPT-3.5 showed moderate agreement. Figure 1 shows the distribution of κ computed from results generated by both models.

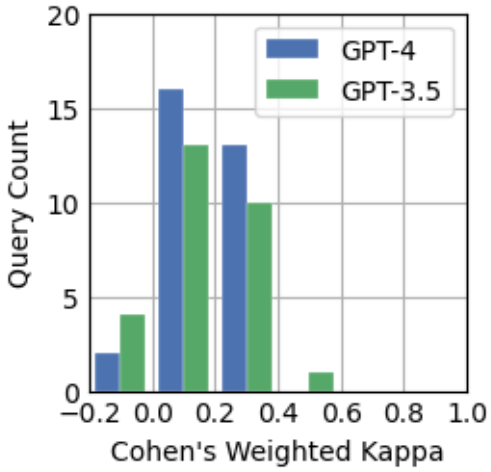


Figure 1 Distribution of Cohen's weighted kappa computed from results generated by GPT-4 and GPT-3.5

Overall low weighted kappa indicates that it is not desirable to fully automate the nugget weight assessment step as the received results disagree with human-generated weights. Furthermore, although we expected better results from the superior GPT-4 model, no improvements can be observed when the model being tested is switched from GPT-3.5 to GPT-4; the only observable enhancement was the number of successful responses, from 28 to 31.

4.2 Failure Analysis

We observed both the weaknesses of GPT-3.5 and GPT-4 as dealing with longer prompts that contain a greater number of nuggets in the nugget list. Tables 2 and 3 show the relationship between nugget count in the prompt and successful/failed attempts respectively; the mean metric shows the average number of nuggets in the prompts and the range shows the most and least case for each category.

Table 2 Relationship between #nugget and successful attempts

Model	Mean	Range
GPT-4	35.71	[13,66]
GPT-3.5	37.69	[13,101]

Table 3 Relationship between #nugget and failed attempts

Model	Mean	Range
GPT-4	142.74	[60,268]
GPT-3.5	124.77	[29,268]

From the tables, we observed a wider tolerance range for GPT-3.5, as it successfully produced a weight array for as many as 101 nuggets, while GPT-4 topped out at 66 nuggets. However, GPT-3.5 also failed to generate results for a prompt containing as few as 39 nuggets, while the least number of nuggets GPT-4 failed to handle was 60. On the other hand, there is a more explicit threshold for GPT-4, which lands approximately between 60 and 66 nuggets.

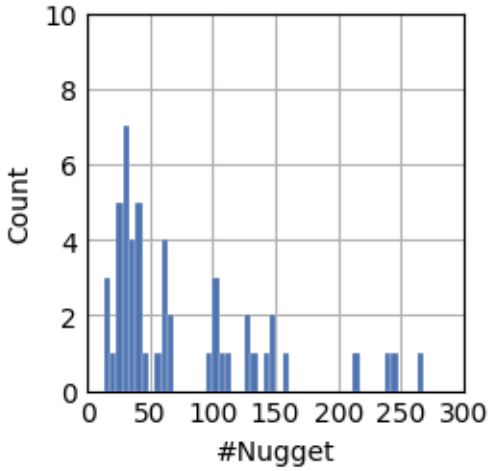


Figure 2 Distribution of nugget count in each nugget list

Figure 2 shows the distribution of nugget count in each nugget list for all 50 queries and nugget lists. In the test collection, 22 lists contain more than 60 nuggets, among which 16 lists exceed 100 nuggets. It is necessary that we consider alternate approaches to deal with queries paired with large numbers of nuggets to increase the success rate of generating weights with GPT-4 and GPT-3.5.

5 Conclusions and Future Work

In this work, we experimented with GPT-4 and GPT-3.5 and studied the potential of leveraging LLMs for automated nugget weight assessment. In our initial attempt, we created a baseline prompt consisting of task instruction, query description, and an exhaustive list of nuggets that belong to the query. However, the baseline approach has its shortcomings as the prompt naturally becomes lengthy when a query has a large number of nuggets, causing a decrease in the quality of the LLMs' answers. The nugget weight array returned from the language model is manually extracted from the response. To evaluate the quality of successful extractions, weights generated by the LLMs are used to compute Cohen's kappa to evaluate the degree of agreement with gold-standard weights. Results showed that there is merely *none to slight* or *fair* inter-rater agreement between both LLMs and human assessors; as a result, we could not conclude that it is wise to replace human assessors with LLMs until the performance is improved.

Our ongoing research includes the implementation of OpenAI's API to reach a higher degree of automation. As our results generated by the initial prompt were not satisfactory, we also plan to re-design the prompts and seek to discover variations that yield better outputs. We also consider point-wise assessment, which evaluates each nugget independently, as an alternative to our current approach of providing the complete list all at once. Nonetheless, it is deducible that point-wise assessment may lead to inefficient use of prompts as each nugget contains only a single piece of atomic data.

In the long run, we plan to build a fully automated nugget extraction system that extracts nuggets from the textual response in each conversational turn, reaching the ultimate objective of complete automation of nugget-based evaluation frameworks. With automatically extracted nuggets, the implementation of nugget-based assessment no longer needs to rely on human extractors and the time and cost can be reduced. It is also imperative to broaden our research to encompass additional aspects for a comprehensive assessment of LLM-based textual conversational systems and explore evaluation criteria for each criterion as suggested in the SWAN framework.

References

[1] Tetsuya Sakai. SWAN: A generic framework for auditing textual conversational systems. *arXiv preprint arXiv:2305.08290*, 2023.

[2] Tetsuya Sakai, Makoto P Kato, and Young-In Song. Overview of NTCIR-9 1click. In *NTCIR*, 2011.

[3] Makoto P Kato, Matthew Ekstrand-Abueg, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Overview of the NTCIR-10 1click-2 task. In *NTCIR*. Citeseer, 2013.

[4] Makoto P. Kato, Virgil Pavlu Matthew Ekstrand-Abueg, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Overview of the ntcir-11 mobileclick task. *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014*, 2014.

[5] Makoto P Kato, Tetsuya Sakai, Takehiro Yamamoto, Virgil Pavlu, Hajime Morita, and Sumio Fujita. Overview of the ntcir-12 mobileclick-2 task. In *NTCIR*, 2016.

[6] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 621–630, New York, NY, USA, 2011. Association for Computing Machinery.

[7] Kai-Xin Chang. Evaluating LLM-based textual conversational systems for direct and immediate information access, IPSJ SIG Technical Report, Vol.2023-IFAT-152 No.15, pp.1–4. Technical report, 2023.

[8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[9] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[10] Tetsuya Sakai. Evaluating evaluation measures for ordinal classification and ordinal quantification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769, Online, August 2021. Association for Computational Linguistics.

[11] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Evaluation measures for ordinal regression. In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*, ISDA '09, page 283–287, USA, 2009. IEEE Computer Society.

[12] Enrique Amigo, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. An effectiveness metric for ordinal classification: Formal properties and experimental results. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3949, Online, July 2020. Association for Computational Linguistics.

[13] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[14] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. September 2023.

[15] Tadele T Damessie, Thao P Nghiem, Falk Scholer, and J Shane Culpepper. Gauging the quality of relevance assessments using inter-rater agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1092, 2017.

[16] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 282–289, New York, NY, USA, 1998. Association for Computing Machinery.

[17] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

DEIM Forum 2023 XX-Y

大規模言語モデルを用いたユーザの目的に基づく説明性を有するPOI推薦手法

片山　一†　牛尼　剛聡††

† 九州大学大学院 芸術工学府　〒 815–8540 福岡県福岡市南区塩原 4 丁目 9-1
†† 九州大学大学院 芸術工学研究院　〒 815–8540 福岡県福岡市南区塩原 4 丁目 9-1
E-mail: †katayama.hajime.498@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　本研究では「デートにおすすめな場所」「ドライブにおすすめな場所」などユーザーが Point of Interest（POI）を探す際の目的を直接クエリとして入力することで，目的に適した POI とその POI が目的に適している理由を提示する手法を提案する．提案手法では事前学習済みモデル BERT に POI に関するレビューを用いて追加事前学習を行った後，レビューテキストからそのレビューが対象とする POI を予測するタスクでファインチューニングを行ったモデルを POI 推薦に利用する．推薦理由提示手法では，BERT の事前学習の際に用いられるタスクの一つである Next Sentence Prediction タスクと POI 推薦で用いた POI 予測タスクを併用して推薦対象の POI がクエリに対し適合すると考えられる理由をレビューから抽出する．また本論文では各手法に対して被験者実験を行い，提案手法の有効性を検証する．

キーワード　POI 推薦、レビュー、観光、情報抽出、BERT

1　はじめに

近年，人々が旅行やレジャーのために訪問する地点 (Point of Interest: POI）を探す際，地図アプリなどで POI 検索を行うことが一般化した．一方，ユーザが POI を訪問するのは,「デート」や「ドライブ」等の目的が存在することが一般的であり，ユーザの目的に適した POI を探したいという要求が存在する．しかし，一般的な POI 検索では,「レストラン」や「コンビニエンスストア」等，対象とする POI のカテゴリをクエリとして指定すれば適切な候補が検索できるが,「デート」や「ドライブ」といった目的を直接クエリとしても期待されるような検索結果が得られない場合が多い．そのため，ユーザは，自分の目的に適した POI のカテゴリを予測して POI の検索を行い，検索結果として得られた個々の POI に関するオンラインレビュー等を参照して，それぞれの POI がユーザの目的に適しているかを一つ一つ確認する必要がある．しかし，対象とする POI に多数のレビューが投稿されている場合には，全てのレビューを一つずつ確認することはユーザにとって負担であり現実的でない．そこで，本研究では，ユーザが POI を探したい時に，ユーザの目的を直接クエリとして入力するだけで，システムがユーザの目的に適した POI を効率的に発見可能なシステムを開発することを目的とする．

ユーザが目的をクエリとして入力した場合，ユーザの目的に適した POI は複数存在する場合が多い．その場合，ユーザは候補となる POI からユーザの目的に適したものを選別することが一般的である．このとき，ユーザはなぜ，それぞれの POI が候補として推薦されたのかを効率的に知ることができれば，ユーザの選別に関する負担を軽減することが期待できる．

本論文では，ユーザが特定の目的に基づいて POI 検索を行う際に，ユーザが指定した目的に適した POI を検索し，それぞれの POI がユーザが指定した目的に合致する理由を提示する手法を提案する．提案手法を利用した POI 検索の例を図 1 に示す．本手法を利用することで，ユーザがクエリとして指定した目的に応じた POI を理由と共にするようなサービスを実現できる．例えば，ユーザがデートを目的として「デートにおすすめ」と入力した場合，対象地域でデートに適したスポットがその理由とともに提示される．一方，ユーザが休日に家族で出かける場所を目的として「家族連れにおすすめ」と入力した場合には，対象地域で家族連れで訪問するのに適したスポットがその理由とともに提示される．たとえこれらの 2 つのクエリで同一の POI が推薦された場合にも，それぞれのクエリに適した別の理由が提示される．このようにユーザの目的に応じて適した POI を理由とともに提示することで，ユーザはそれぞれの POI に対して多くのレビューを読むことなく，効率的に適切な POI を選別可能となる．

ユーザの目的に適した POI を検索し，その理由を提示するために，本研究では POI に関して投稿されたユーザレビューを利用する．Trip Adviser[1]やじゃらん net[2]等の旅行情報サイトには，それぞれの地域における代表的な POI とそれらに関するユーザレビューが投稿されている．ユーザレビューにはそれぞれの POI を訪問した人々がそこを訪問した目的やその結果が記述されていることがある．そこで，本研究では，POI に投稿されているユーザレビューに対して自然言語処理を行うことにより，それぞれの POI がユーザの目的にどの程度適して

1：https://www.tripadvisor.jp/
2：https://www.jalan.net/

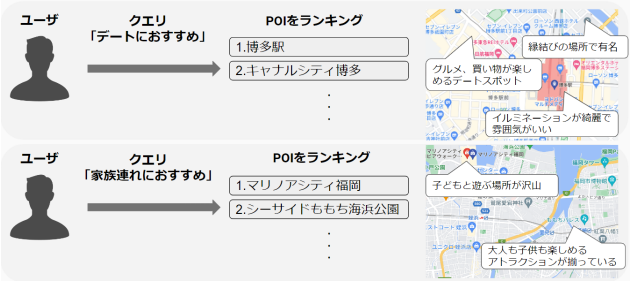


図 1　提案手法の実現例

いるか，またその理由は何かを予測するアプローチを採る．

近年，ニューラルネットワークを利用した自然言語処理技術が飛躍的に発達した．特に，Transformer を利用して膨大なコーパスを利用して事前学習を行った大規模言語モデル（Large Language Model: LLM）を利用することで，自然言語文の意味を高精度に理解して様々な目的に応用することが可能となった．本論文では，代表的な LLM の一つである BERT を利用して，ユーザが指定した目的に対する POI の適合度を予測し，その理由を表現している文を提示する手法を提案する．

本論文の貢献は以下の通りである

- ユーザの目的に基づいて POI を検索し，検索結果となるそれぞれの POI がユーザの目的に適する理由を示すことにより効率的に POI を選別するフレームワークを示した．
- POI に関して登録されたユーザレビューを利用して，それぞれの POI のユーザの目的に対する適合度を予測する手法を開発した．
- それぞれの POI がユーザの目的に対して適切である理由を表す文を提示する手法を開発した．
- POI 以外のアイテム群に関しても提案手法が有効であるかを検証した．
- 提案手法の有効性を被験者実験により評価した．

本論文の構成は以下のとおりである．2 章では，関連研究について述べる．3 章では提案手法について述べる．4 章では評価実験について述べる．5 章では提案手法の応用について述べる．最後に 6 章でまとめと今後の課題について述べる

2　関 連 研 究

2.1　POI 推薦

これまでにも，POI を推薦するために様々な手法が提案されている [1]．Hu ら [2] は，グラフ埋め込み法による POI 推薦手法を提案している．この手法では，ユーザ，POI，POI の特徴間にある高次の相互関係をペアワイズによる相互作用関係の集合に分解して効率的なモデル化を実現している．Yang ら [3] は，協調フィルタリングと半教師あり学習を組み合わせて POI に対するユーザの嗜好をモデル化する手法を提案している．Ma ら [4] は，地理的情報，カテゴリー情報，他ユーザに関する情報，スポットの人気度という複数の情報を組み込んだ POI 推薦フレームワークを提案している．Zhao ら [5] は時間間隔，距離感覚，近傍の POI を POI に関するコンテキストとし，コンテキストとユーザの逐次的な訪問パターンを同時にモデル化し，次にユーザがどこに訪れるべきかを推薦する手法を提案している．本研究では，ユーザが目的をクエリとして直接入力し，目的に適した POI 推薦を行う点がこれらの研究とは異なる．

2.2　説明可能性

Zhang ら [6] は，e コマースでのアイテム推薦において，各アイテムの特徴に基づいて推薦理由を説明することでユーザの購買行動に大きな差が出ることを示した．そのため，情報推薦における説明可能性に関する研究は，推薦システムに対する信頼性やユーザの満足度を高めるうえで重要なトピックの一つとなっている．

Li ら [7] は，ユーザとアイテムの関係を利用して，Transformer を用いた個別化された推薦と説明生成を行うモデルを提案している．Geng ら [8] は，個別化されたプロンプトのテンプレート集を用いて学習を行うことで，説明可能推薦など様々な推薦タスクを一つのフレームワークに統合する手法を提案している．また，Ge ら [9] は，推薦システムにおける公平性という観点から，説明可能な公平性という考え方を導入している．Li ら [10] は，説明可能な推薦システムによって生成された説明文を標準的なランキング手法で定量的に評価するためのベンチマークデータセットを提供している．これらのように，ユーザに対してアイテムを推薦する理由に基づいて高度な推薦システムを実現するために，説明可能な推薦に関する研究は広く行われている．本研究では，ユーザがクエリとして与えた目的に対応する POI を推薦する際に，ユーザは推薦された理由が不明であれば POI を選別できないという問題に着目し，ユーザが与えたクエリに基づいた推薦理由をクエリと対応付けて提示することでユーザの効率的な POI 選別を支援することを目的とする．

2.3　BERT による文書検索

本研究では，ユーザの目的に適した POI 推薦に代表的なニューラル言語モデルの一つである BERT [11] を用いる．BERT を利用することで，情報推薦タスクや検索タスクの性能の向上を目指す様々な手法が提案されている．Yang ら [12] や MacAvaney ら [13] は，アドホック文書検索において BERT を用いることで精度が向上することを報告している．Yilmaz ら [14] は，ユーザのクエリに類似した文書の中から，文書を文単位に分割した情報をもとに BERT を用いて再ランキングを行う手法を提案している．Zheng ら [15] は，ユーザのクエリに対して BERT を用いてクエリ拡張を行うことで文書の再ランキング化を行う手法を提案している．また，BERT を用いたオープンドメインの質疑応答タスクに関する手法も提案され，Lee ら [16] は Inverse Close Task という教師無し事前学習を用いることでユーザの入力した質問文に対し，Wikipedia などの外部知識から適切な文書を検索する機構と，検索した文書から質問に対する適切な回答を抽出する機構の両方を学習させる手法を提案している．また，Guu [17] らは，BERT の事前学習タスクの一つである

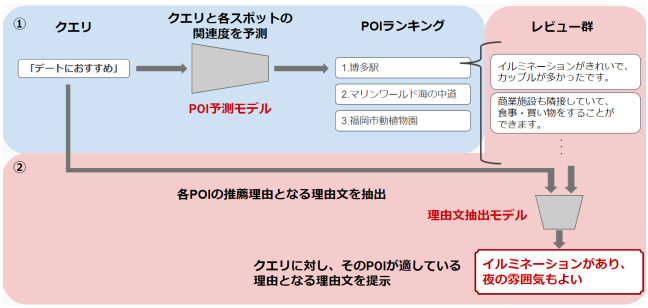


図 2　提 案 手 法

Masked Language Modeling（MLM）を拡張することで，外部知識を用いた MLM による事前学習を行うことでオープンドメインの質問応答タスクの精度向上させる手法を提案している．

本手法では，レビューから対象の POI を予測させるタスクを利用して BERT をファインチューニングすることで，ユーザが指定した目的に合致する POI 推薦を行う．

3　提 案 手 法

3.1　提案手法の概要

本研究では，ユーザが指定した目的に，それぞれの POI がどの程度適合しており，それが目的に対して適合していると考えられる理由を示すために，POI に関して投稿されたユーザレビューに対して自然言語処理を適用するアプローチを取る．本研究では，上記の処理のために BERT [11] を用いる．BERT は，一般的な言語パターンを学習するために，大規模コーパスを用いた事前学習が行われた代表的なニューラル言語モデルの一つである．本論文で提案する手法の概要を図 2 に示す．

本手法は，①ユーザの目的を表すクエリを POI 予測モデルに入力し，出力された分類スコアの高い順に POI をランキングし，②前の処理でランキングが上位の POI をクエリに対応する POI とし，それぞれの POI に関するレビュー群からクエリの内容に関係するレビューを抽出する，という 2 ステップから構成される．それぞれの処理でユーザが指定したクエリを用いているため，POI 推薦と特徴文生成，それぞれの処理の出力がクエリと関連付けられる．

3.2　目的に適した POI 予測

ユーザが与えたクエリと POI を予測するために，機械学習を利用した多クラス分類モデルを利用する．具体的には，図 3 に，提案手法で利用する機械学習モデルの概要を示す．

提案手法を形式的に定義する．いま，対象とする POI 集合 $PT = \{pt_1, pt_2, \cdots, pt_{|PT|}\}$ と，POI に含まれる POI に関して記述されたレビュー集合 $RV = \{rv_1, rv_2, ..., rv_{|RV|}\}$ が与えられているとする．ここで，それぞれのレビュー rv は，それが対象とする POIpt とレビューテキスト T のペア $rv = (pt, T)$ として以下のように定義される．

$$rv = (pt, T) \tag{1}$$

として定義される．ここで，$pt(rv)$ はレビュー rv の対象とする POI を表し，$text(rv)$ はレビュー rv のレビューテキストを表す．ここで，レビューテキスト T は，文 s_i の列として以下のように定義される．

$$T = \{s_1, s_2, ..., s_{|T|}\} \tag{2}$$

あるレビュー文 s が与えられたときに，そのレビュー文 s が POI_{pt} について述べられたと考えられる確率を以下の条件付き確率として以下の式で表す．

$$POI(pt, s) = P(pt|s) \tag{3}$$

本手法では，ユーザのクエリ q がテキストとして与えられたとき，確率 $P(pt|q)$ をクエリテキスト q に対する pt の適合度ととらえる．本手法では，$P(pt|q)$ を求めるために機械学習モデルを利用する．本研究では以下の式で定義されるモデルを考える．

$$\boldsymbol{y} = f(\mathbf{x}; \theta) \tag{4}$$

ここで，$\mathbf{x}$ は，対象とするテキストの埋め込みベクトルである．また，θ はパラメータであり，$\boldsymbol{y} = (y_1, y_2, ..., y_{|PT|})$ は予測値である．ここで，y_i はスポット pt_i に対する確率を表し，$\sum_j y_j = 1$ とする．

上記の機械学習モデルを学習させるために，以下の教師データ D を利用する．

$$D = \{(\mathbf{x}, \mathbf{y}) | \mathbf{x} = emb(s), \mathbf{y} = one_hot(pt), s \in text(rv), pt = pt(rv), rv \in RV\} \tag{5}$$

ここで，$emb(s)$ は文 s の埋め込み表現であり，$one_hot(pt_i)$ はスポット i 番目の要素のみが 1 であり，ほかの要素が全て 0 となる one-hot ベクトルである．

提案手法では，上記のモデルを訓練するための教師データとして，旅行サイトに投稿された POI に関するレビューを利用する．具体的には，対象とする POI に投稿されたレビューを文に分割し，それぞれの文の埋め込み表現と POI の one-hot ベクトルのペアを教師データとする．つまり，レビュー中の文を入力とし，そのレビューが対象とする POI を予測するというタスクで機械学習モデルの訓練を行う．

上記の手法で訓練した機械学習モデルは，入力したテキストの内容に適した POI ほど分類確率が大きく出力することが期待できる．そこで本手法では，上記の手法で構築した POI 予測モデルに対して，ユーザの目的をクエリとして入力し，POI を分類確率が高い順にランキングを行い，上位のスポットをユーザの目的に適した POI としてユーザに推薦する．

具体的には，POI 予測モデルの学習の際に，入力データをレビュー全文としたモデルと，レビューをセンテンスで区切り，各センテンスに対象のスポットを正解ラベルとして学習させた二つのモデルを構築する．本論文ではそれぞれのモデルを「BERT-full」と「BERT-sentence」と呼び，各モデルの精度を比較する．日本語事前学習済みモデル BERT に対し，レビューの内容から正解ラベルのスポットを予測させるタスクで

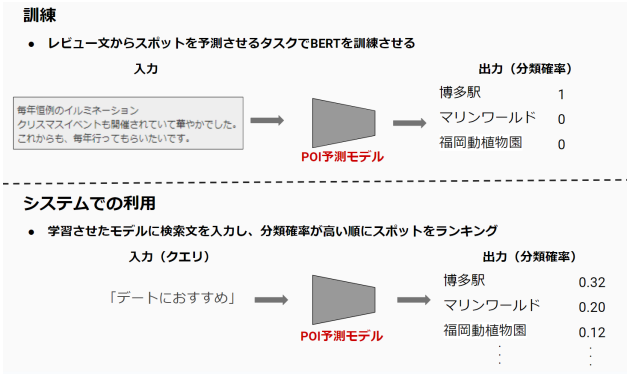


図 3　POI 予測モデル

fine-tuning を行い，POI 予測モデルを学習させる．

Lee ら [18] や仁木ら [19] らは一般的な文章（Wikipedia など）をコーパスとして事前学習を行ったモデルにある特定領域に関するコーパスを用いて追加事前学習を行うことで，特定領域でのテキストマイニングタスクの精度向上を報告している．筆者らの研究 [20] でもオンラインレビューをコーパスとした事前学習を行うことで，オンラインレビューに関するタスク精度の向上が見られた．そこで本研究では東北大学の乾研究室が公開している日本語事前学習済みモデル [3] の BERT に，「じゃらん net」[4] に記載されているレビューをコーパスとして追加事前学習を行わせたモデルを用いる．

3.3　理由文の推定

本研究では，ユーザが指定した目的に適していると予想される POI に対して，その POI がユーザの目的に適していると考えられる理由を提示する手法を開発することを目的とする．前節において，ユーザが指定した目的に対して，それぞれの POI がどの程度適しているかを予測する手法について述べた．本節では，それぞれの POI が適していると考えられる理由を提示する手法を述べる．

本研究では，対象とする POI に関して投稿されたレビューに含まれる文の中から，ユーザが指定した目的に POI が適していると考えられる文を抽出し，その文を提示することで，POI が適していると考えられる理由を提示するアプローチを採る．本論文では，ユーザが指定した目的に POI が適していると考えられる文を，「理由文」と呼ぶ．

本研究では，理由文を抽出する方法として，Next Sentence Prediction を利用する手法を提案する．Next Sentence Prediction (NSP) は，BERT において事前学習に利用されるタスクの一つである．BERT において NSP を利用する主な目的は，BERT が与えられた文の後に続く文が論理的に連続しているかどうかを判断できるようにすることである．これは，文脈的な理解を深め，言語モデルがより豊かな表現を学習するのに役立つ．具体的には，ある文 A とそれに続く別の文 B のペアを入力し、B 文が A 文に論理的に続くかどうかを予測させる．

3：https://github.com/cl-tohoku/bert-japanese
4：https://www.jalan.net/

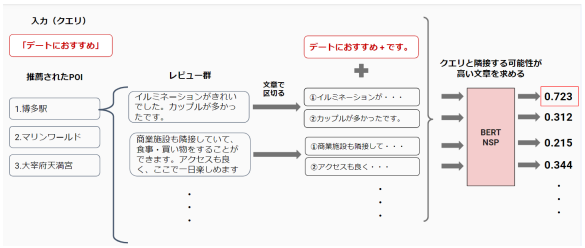


図 4　NSP を用いたレビュー抽出

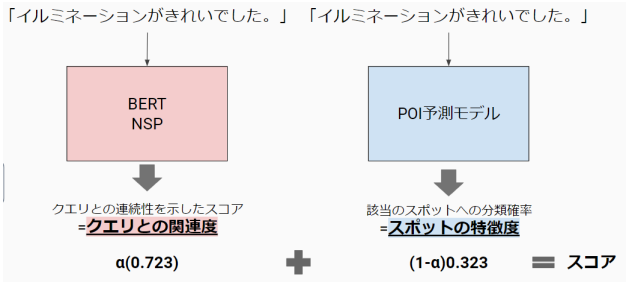


図 5　NSP と POI 予測モデルを用いたレビュー抽出

本手法の概要を図 4 に示す．本手法では，まず，POI として推薦されたスポットのレビューを文に分割する．クエリと各文をペアとして BERT に入力し，NSP タスクにより入力された 2 文が連続した文章であるかどうかを予測する．全ての文の中でクエリと連続する可能性が最も高いと予測された文が，意味的もしくは論理的に整合性がとれている，つまりクエリに対する理由となる文が抽出できると考えられる．

本研究ではクエリと関連しているだけでなく，対象スポットの特徴を表している文かも考慮する．なぜなら，ユーザの目的に適する理由が一般的なものである場合，その分は POI を選別する際の役に立たないと考えられるからである．そこで，理由文抽出の際に NLP タスクに加え，3.2 節で述べた POI 予測モデルを用いる．その概要を図 5 に示す．ある POI のレビュー中の文に対し，NLP によってクエリとの連続性を示したスコアをクエリとの関連度とみなす．また，センテンスを 3.2 節で述べた POI 予測モデルに入力し，出力された該当 POI への分類確率を POI の特徴度とみなす．クエリとの関連度と POI の特徴度の二つのスコアを用いて，POI 推薦理由を示す文の適合度を計算し，スコアが高い文をクエリに対応する POI の理由文とする．形式的には POI 推薦理由スコアを以下の式で定義する．

$$ex_score(q,s) = \alpha NSP(q,s) + (1-\alpha)POI(pt(s),s) \quad (6)$$

ここで q はユーザのクエリ，s は対象とする文，$pt(s)$ は文 s が含まれるレビューの対象となっている POI である．α は文書の連続性と POI の特徴の一致度のどちらをどの程度重要視するかを表すパラメータである．

4　評価実験

4.1　実験手法

4.1.1　データセットと事前学習

提案手法の有効性を評価するために評価実験を行った．本実

験では「じゃらん net」に掲載されているレビューをスクレイピングで収集して実験に利用するデータセットを構築した．予測に利用する BERT モデルを，今回対象とする POI に関するレビューに適したモデルとするために，追加事前学習を行った．ベースとなる BERT モデルとしては，東北大学乾が公開している日本語モデルを利用した．BERT モデルの追加事前学習に用いたデータセットは，各都道府県からレビュー数が 100 件以上存在する POI をランダムに 50 個サンプリングし，それぞれの POI から最新のレビュー 100 件を利用した．また，一部の都道府県では 100 件以上レビューが存在する POI が 50 個未満の場合もあり，その際は 100 件以上のレビュー数が存在する全ての POI をデータセットとして利用した．全てのレビューを文ごとに区切り，合計で 747,318 件の文を対象とした．事前学習のタスクとして，Masked Language Model と Next Sentence Prediction を行った．

提案手法で POI 予測モデルを構成するために用いたデータセットは追加事前学習を行ったデータセットの中から，福岡県のレビューを用いた．具体的には，福岡県に含まれる 50 個の POI に対しそれぞれ 100 件のレビュー，合計で 5,000 件のレビューを利用した．POI 予測モデルに関しては，5,000 件のレビューの内，6 割を学習データ，2 割を検証データ，2 割を POI 予測モデルの精度を確認するためのテストデータに用いた．

4.1.2 POI 予 測

4.1.1 で述べたデータセットを用いてレビューからスポットを予測させるタスクで fine-tuning を行い，テストデータを利用して，POI を正しく予測できるかを評価する．実験の対象としたのは以下の 3 手法である．

- BERT-full（提案手法）

3.2 節で提案した学習の際に，レビュー全文を入力データとして学習を行ったモデル．

- BERT-sentence（提案手法）

3.2 節で提案した学習の際に，レビューをセンテンスごとに区切り各センテンスを入力データとして学習を行ったモデル．

- BM25（比較手法）

情報検索における代表的なランキング手法の一つであり，文書内の単語の出現頻度や文書の長さに基づいてクエリに対する文書集合のランキングを行う.「デート」や「ドライブ」などユーザがスポットを探す際に目的となる名詞の出現頻度が高いスポットは，対象の目的に適したスポットであると考えられる．そこで本研究では各スポットに関する全てのレビューに対して形態素解析を行い名詞のみを抽出する．クエリとして目的となる名詞を入力し，BM25 によるスコアが高い順に順位付けを行う．

データセット中の福岡県の 50 件のスポットに対して,「デート」,「ドライブ」,「家族で出かける」,「写真撮影」の 4 つの目的をクエリとして実験を行った．それぞれの手法で得られた結果に対して，被験者に対して，クエリとして与えた目的にどれほど適したスポットであるかを {1: 適切である (5 点)，2: どちらかといえば適切である (4 点)，3: どちらでもない (3 点)，4: どちらかといえば適切でない (2 点)，5: 適切でない (1 点)} の 5 段階で回答してもらった．また回答者が該当のスポットを知らない場合は {6: 該当のスポットを知らない } と回答してもらった．被験者は，福岡県に 3 年以上在住経験がある 30 名であり，クラウドソーシングサービスであるクラウドワークス上で被験者を募り実験を行ってもらった．

手法の評価方法として nDCG を用いる．提示した目的に対して被験者が回答した各スポットの適切度の平均値を計算し，値が高い順にスポットをランキングしたものを正解データとする．各手法で予測したランキングとの nDCG を計算し提案手法の評価を行った．

4.1.3 理由文抽出

理由文抽出には以下の 2 手法で比較を行う．

- 提案手法

3.3 で提案した手法．NLP タスクにより出力されたスコアに α の重みを掛け，スポット予測モデルにより出力されたスコアに（1-α）の重みを掛ける．本研究では α の値を 0.2，0.4，0.6，0.8，1.0 に変化させてそれぞれの結果を比較する．

- BM25

レビューをセンテンスで区切った後に，BM25 手法によってクエリと類似したセンテンスを抽出する．

本論文では理由文抽出法の比較として，各手法でスコアが上位 3 件のセンテンスを抽出し，各センテンスについて二つの項目で評価する．一つ目は，被験者に，クエリに対して抽出したセンテンスが POI の推薦理由としてどれだけ適切であるかを {1: 適切である (5 点)，2: どちらかといえば適切である (4 点)，3: どちらでもない (3 点)，4: どちらかといえば適切でない (2 点)，5: 適切でない (1 点)} の 5 段階で回答してもらう実験である．二つ目は，被験者に，抽出したセンテンスがどれだけ対象の POI の特徴を表しているかを {1: よく表している (5 点)，2: どちらかといえば表している (4 点)，3: どちらでもない (3 点)，4: どちらかといえば表していない (2 点)，5: 表していない (1 点)} の 5 段階で回答してもらう実験をである．被験者実験の結果に対しウィルコクソンの符号順位検定によって比較手法との優位性を検討する．被験者は，福岡県に 3 年以上在住経験がある 30 名であり，クラウドソーシングサービスであるクラウドワークス上で被験者を募り実験を行ってもらった．

4.2 実 験 結 果

4.2.1 POI 予 測

POI 予測モデルについて被験者実験の結果を表 1 に示す．まず二つの提案手法に注目した場合，両者とも nDCG の値が 0.900 以上となった．特に BERT-full モデルは nDCG@10 の時に最も高い値を記録した．このことから，提案手法によって適切な POI 推薦を行うことができると考えられる．提案手法は，比較手法である BM25 と同程度の精度で POI 予測が実現できている．

4.2.2 理由文抽出

各手法で抽出した理由文に関して，被験者の評価結果を表 2，表 3 に示す．表 2 は「POI 推薦理由の適切性」の評価を，表 3 は「POI の特徴度としての妥当性」の評価を，各手法で抽出し

表 1　各手法による POI 予測の nDCG

	nDCG@3	nDCG@5	nDCG@10
BERT-full	0.913	0.910	0.933
BERT-sentence	0.918	0.929	0.929
BM25	0.897	0.915	0.931

た特徴文に対して，被験者与えた評価値の平均を示している．提案手法の有効性を求める手法としてウィルコクソンの符号順位検定を採用し，優位水準を 0.05 とする．各提案手法と BM25 を比較し，提案手法が優位とされた場合は表の数字にアスタリスクを付けている．

提案手法と比較手法を比べた場合,「推薦理由として適切性」,「POI 特徴度としての妥当性」いずれの場合も提案手法が有意であった．このことから提案手法によってユーザの目的に合わせた POI 推薦を行う際の適切な理由文を抽出することが可能であると考えられる．

各手法で実際に抽出した理由文の例を表 4，表 5 に挙げる．表 4 のように BM25 ではクエリに対して関係のない理由文を抽出しているのに対し，提案手法では推薦理由 POI の特徴の観点で適切なセンテンスを抽出できている．また，表 5 に示す例のように比較手法では抽出された文章から POI がどのような場所か読み取ることが難しいが，提案手法では POI の特徴を上手く表していると考えられる．以上のことから，いずれの質問でも提案手法が高い評価を得られた．

また本論文では，提案手法において POI 予測モデルを導入することの有効性も評価した．具体的には POI 予測モデルを用いた手法 (α=0.4) と NSP タスクのみの手法 (α=1.0) での被験者実験の結果をウィルコクソンの符号順位検定によって評価する．検定の結果,「推薦理由としての適切性」「POI 特徴度としての妥当性」いずれの場合も POI 予測モデルを用いた手法が有意であることが示された．このことから理由文抽出手法には NSP タスクと POI 予測モデルを併用した手法が有効であると考えられる．

5　提案手法の応用

本章では 3 章で述べた提案手法の応用について述べる．提案手法ではクエリから POI を予測するために，機械学習を利用した多クラス分類モデルを利用し,「じゃらん net」に記載されたオンラインレビューをデータセットとして図 3 で示した訓練を行った．また，対象の POI に関するレビューを文に分割し，5 で示した手法によって理由文の抽出を行った．

これら手法はデータセットとして用いるレビューを変えることで別のアイテム群に関する推薦も行える可能性がある．そこで，本章では映画に関するオンラインレビューで訓練した機械学習モデルは，ユーザの入力するクエリから適切な映画を推薦し，対象の映画についての推薦理由文を抽出することが可能であるか追加実験として検証した。

表 2　推薦理由としての適切性

提案手法 (α =0.2)	4.046*
提案手法 (α =0.4)	4.169*
提案手法 (α =0.6)	4.005*
提案手法 (α =0.8)	3.925*
提案手法 (α =1.0)	3.984*
BM25	3.064

表 3　POI 特徴度としての妥当性

提案手法 (α =0.2)	4.163*
提案手法 (α =0.4)	4.207*
提案手法 (α =0.6)	4.024*
提案手法 (α =0.8)	3.962*
提案手法 (α =1.0)	3.750*
BM25	2.640

表 4　クエリ「ドライブにおすすめ」，POI「 関門海峡 (早鞆ノ瀬戸)」

提案手法 (α=0.4)	とても眺めが良く素晴らしい オーシャンビューで癒される景色です。
BM25	関門海峡にかかる大橋を ドライブで通りました。

表 5　クエリ「写真撮影におすすめ」，POI「門司港駅」

提案手法 (α=0.4)	写真をとってもお洒落だし いい雰囲気の場所です。
BM25	おすすめです。

5.1　評 価 実 験

5.1.1　データセットと事前学習

本実験では「Yahoo!検索」[5]の機能のひとつである「映画作品ランキング」に掲載されているレビューをスクレイピングで収集して実験に利用するデータセットを構築した．本実験も 3.2 節で述べた追加事前学習を行った．

追加事前学習に用いたデータセットは 226 作品の映画からそれぞれ 100 件のレビューを利用した．レビューには「面白かった」のみ記載されたものなど作品の感想について詳細に述べていないものも存在する．そこで，文字数が 20 文字以上であるものに限定をしてレビューの収集を行った．全てのレビューを文章ごとに区切り 144574 件の文章を対象とした．ベースとなる BERT モデルは乾研究室が公開している日本語モデルを利用した．事前学習のタスクとして，Masked Language Model と Next Sentence Prediction を行った．

提案手法によって映画予測モデルを構成するために用いたデータセットは追加事前学習と同じく 226 作品の映画からそれぞれ 100 件のレビュー，合計で 22600 件のレビューを利用した．本実験では訓練時の入力データはレビュー全文とし，6 割を学習データ，2 割を検証データ,2 割を映画予測モデルの精度を確認するためのテストデータに用いた．

5.1.2　映画予測モデル

5.1.1 項で述べたデータセットを用いて訓練した機械学習モデルに，筆者が用意したクエリを入力した．今回用いたクエリは以下の二つである．

- 「実話を基にしたストーリー」
- 「子どもと一緒に見たい映画」

各クエリに対して本手法で提案したモデルが正しく推薦を行うことができているか，筆者の主観によって評価を行う．

5：https://search.yahoo.co.jp/

5.1.3 理由文抽出

理由文抽出には以下の 2 手法で比較を行う．

- 提案手法

3.3 で提案した手法．NLP タスクにより出力されたスコアに α の重みを掛け，スポット予測モデルにより出力されたスコアに（1-α）の重みを掛ける．本実験では α の値を 0.2，0.4，0.6，0.7，0.8，1.0 に変化させてそれぞれの結果を比較する．

- BM25

レビューをセンテンスで区切った後に，BM25 手法によってクエリと類似したセンテンスを抽出する．

4.1.3 項と同様に，本実験では推薦理由としての適切性，映画の特徴を説明した文章としての妥当性の二つの項目で評価を行う．一つ目の項目では，各手法でスコアが上位 3 件のセンテンスを抽出し，各センテンスが映画の推薦理由としてどれだけ適切であるかを {1: 適切である (5 点)，2: どちらかといえば適切である (4 点)，3: どちらでもない (3 点)，4: どちらかといえば適切でない (2 点)，5: 適切でない (1 点)} の 5 段階で回答してもらった．二つ目の項目では，抽出したセンテンスがどれだけ対象の映画の特徴を表しているかを {1: よく表している (5 点)，2: どちらかといえば表している (4 点)，3: どちらでもない (3 点)，4: どちらかといえば表していない (2 点)，5: 表していない (1 点)} の 5 段階で回答してもらう実験をである．被験者実験の結果に対しウィルコクソンの符号順位検定によって比較手法との優位性を検討する．被験者は，3 クラウドソーシングサービスであるクラウドワークス上で被験者を募った 30 名に実験を行ってもらった．

5.2 実験結果

5.2.1 映画予測

各クエリに対して推薦された作品の例を表 6，表 7 に示す．

表 6 ではいずれの作品も実話が基となった作品が推薦されている．本実験では訓練の際に作品のあらすじやジャンルなどといったメタデータはデータセットとして利用していないため，表 9 の結果から本手法によってレビューから作品の情報を学習することが可能であることが分かる．

表 7 ではアニメ作品が多く推薦された．これらの作品のレビューには「子どもと一緒に鑑賞しました.」「大人も子供も楽しめます.」といった内容が多く含まれる．これらの結果から，レビューを基にすることで作品のターゲット層も予測することが可能であると考えられる．しかし,「IT/イット ””それ””が見えたら，終わり.」がランキング上位に推薦されているが，対象の作品は R-15 指定を受けており，クエリに適した作品とは言えない．この作品が上位に来た理由として,「子どもなら怖がりそう」といった感想が多い点，演者に子どもが多く出演しており，そこに言及したレビューが存在した点などが，このような結果になったと考えられる．今後このようなクエリに適さない作品をどのように取り除くか検討する必要がある．

これらの結果から，本論文で POI 推薦を行うために提案した手法を，別のデータセットを利用して訓練を行うことで，対象のレビューに関するアイテム群に対しての推薦も行うことが可能であると考えられる．今後は被験者実験によって提案手法による有効性を検討する．

表 6 推薦された作品の例 1

クエリ	実話を基にしたストーリー
作品	アメリカンスナイパー
	ダンケルク
	ハドソン川の奇跡
	英国王のスピーチ
	硫黄島からの手紙

表 7 推薦された作品の例 2

クエリ	子どもと一緒に見たい映画
作品	リメンバー・ミー
	トイストーリー 3
	崖の上のポニョ
	アナと雪の女王
	IT ／イット“ それ ”が見えたら、終わり。

5.2.2 理由文抽出

各手法で抽出した理由文に関して，被験者実験の結果を 8，9 に示す．表 8 は「映画推薦理由の適切性」の評価を，表 3 は「映画の特徴度としての妥当性」の評価を，各手法で抽出した理由文に対して，被験者与えた評価値の平均を示している．提案手法の有効性を求める手法としてウィルコクソンの符号順位検定を採用し，優位水準を 0.05 とする．各提案手法と BM25 を比較し，提案手法が優位とされた場合は表の数字にアスタリスクを付けている．

提案手法と比較手法を比べた場合，α の値が 0.7 の場合は「推薦理由としての適切性」「映画の特徴度としての妥当性」いずれの場合も提案手法が有意であった．また,「推薦理由としての適切性」という評価項目では α の値が大きい，つまり NSP によるスコアに対する重みを大きくした場合の結果が比較手法よりも有意であった．

しかし，4.2.2 項で述べたように，POI 推薦に関する理由文の抽出ではいずれの重みに対しても提案手法が有意であるのに対し，映画推薦に関する理由文の抽出では α の値を小さい場合，つまり映画予測モデルが出力したスコアに対する重みを大きくした場合は比較手法より有意でないという結果になった．

このような結果となった理由として，映画に関するレビューには登場人物名や作中オリジナルの固有名詞が多く登場することが挙げられる．このような固有名詞が含まれるレビュー文を映画予測モデルに入力した場合，対象の映画への分類確率が非常に高く出力される．したがって，理由文抽出に関する提案手法での映画予測モデルで出力されたスコアに対する重みを大きくすると，理由文としては適切でない文章が抽出されたと考えられる．

これらの結果から，データセットの内容によって NSP によるスコアと予測モデルによるスコアに対する重みを調整することで，提案手法によって適切な理由文が抽出することができる

表 8　推薦理由としての適切性

提案手法 (α =0.2)	3.053
提案手法 (α =0.4)	3.422
提案手法 (α =0.6)	3.467
提案手法 (α =0.7)	3.666*
提案手法 (α =0.8)	3.733*
提案手法 (α =1.0)	3.911*
BM25	3.319

表 9　映画の特徴度としての妥当性

提案手法 (α =0.2)	3.498
提案手法 (α =0.4)	3.540
提案手法 (α =0.6)	3.594
提案手法 (α =0.7)	3.715*
提案手法 (α =0.8)	3.599
提案手法 (α =1.0)	3.518
BM25	3.526

と考えられる．

6 ま と め

本論文では，オンラインレビューをユーザの目的クエリに応じて POI を推薦し，その推薦理由となるレビューを抽出する手法を提案した．本手法では，日本語事前学習済み BERT モデルに対し，レビューをコーパスとした追加事前学習させたモデルをレビューに含まれる文から POI を予測する自己教師あり学習により fine-tuning したモデル構築する．そして構築したモデルに対して，ユーザの目的を表すクエリを入力として，POI への分類確率に基づいてそれぞれの POI のクエリへの適合度を予測する．そして，ユーザの目的を表すクエリと，対象とする POI に含まれるレビュー文のペアに対して Next Sentence Prediction を用いることで，クエリに対する推薦理由として適切かつ，推薦されたスポットの特徴を述べたセンテンスを理由文として抽出する．福岡エリアを対象とした４つのクエリを利用した 30 名の被験者による評価実験の結果，提案手法の有効性が示された．

また，映画に関するオンラインレビューを用いて同様の実験を行うことで，別のアイテム群に対しても本論文で提案した手法が応用可能であることが示された．

今後の課題として，ユーザが与えた自由なクエリに基づいた，より多様な地域での評価実験を行うこと，別のデータセットを用いた際の提案手法による結果のさらなる検証を行うことを考えている．

文　献

[1] Md.Ashraful Islam, Mir Mahathir Mohammad, Sarkar Snigdha Sarathi Das, Mohammed Eunus Ali, “A survey on deep learning based Point-of-Interest (POI) recommendations” Neurocomputing, 472, 306-325, 2022

[2] Xiaojiao Hu, Jiajie Xu, Weiqing Wang, Zhixu Li, An “A graph embedding based model for fine-grained POI recommendation” Neurocomputing, 428, 376-384, 2021

[3] Carl Yang, Lanxiao Bai, Chao Zhang, Quan Yuan, Jiawei Han “Bridging Collaborative Filtering and Semi-Supervised Learning:A Neural Approach for POI Recommendation” KDD, 1245–1254, 2017

[4] Yaxue Ma, Jin Mao, Zhichao Ba, Gang Li,“Location recommendation by combining geographical, categorical, and social preferences with location popularity” Information Processing and Management 57, 2020

[5] Pengpeng Zhao, Anjing Luo, Yanchi Liu, Jiajie Xu, Zhixu Li, Fuzhen Zhuang, Victor S. Sheng, Senior Member, IEEE, and Xiaofang Zhou, Fellow, IEEE “Where to Go Next: A Spatio-Temporal Gated Network for Next POI Recommendation” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL 34, NO 5, MAY 2022

[6] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, Shaoping Ma, “Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis” In Proceeding of the 37th international ACM SIGIR conference on Research & development in information retrieval, 83-92, 2014

[7] Lei Li, Yongfeng Zhang, Li Chen,https “Personalized Transformer for Explainable Recommendation” In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4947–4957, 2021

[8] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Yongfeng Zhang, “Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)” arXiv preprint arXiv:2203.13366, 2022

[9] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, Yongfeng Zhang, “Explainable Fairness in Recommendation” arXiv preprint arXiv:2204.11159, 2022

[10] Lei Li, Yongfeng Zhang, Li Chen “EXTRA: Explanation Ranking Datasets for Explainable Recommendation” SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,Association for Computing Machinery, Inc, 2463-2469, 2021

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.4171-4186, 2019

[12] Wei Yang, Haotian Zhang, Jimmy Lin ”Simple Applications of BERT for Ad Hoc Document Retrieval” arXiv preprint arXiv:1903.10972，2019

[13] Sean MacAvaney, Andrew Yates, Arman Cohan, Nazli Goharian ”CEDR: Contextualized Embeddings for Document Ranking” arXiv preprint arXiv:1904.07094，2019

[14] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, Jimmy Lin ”Cross-domain modeling of sentence-level evidence for document retrieval” In EMNLP/IJCNLP (1), pages 3488–3494. 2019 Association for Computational Linguistics.

[15] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, Andrew Yates ”BERT-QE: Contextualized Query Expansion for Document Re-ranking” In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4718–4728

[16] Kenton Lee, Ming-Wei Chang, Kristina Toutanova, ”Latent Retrieval for Weakly Supervised Open Domain Question Answering”, ACL(1)2019.6086-6096, 2019

[17] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang, ”REALM: Retrieval-Augmented Language Model Pre-Training”, ICML, 2020

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang “BioBERT: a pre-trained biomedical language representation model for biomedical text mining” Bioinformatics, Volume 36, No.4, pp.1234–1240, 2020

[19] 仁木裕太, 坂地泰紀, 和泉潔, 松島裕康, “再事前学習した BERT を用いた金融文書中の因果関係知識有無の判別”, 人工知能学会全国大会論文集, 2020.

[20] 片山一, 牛尼剛聡, “オンラインレビューに関する自己教師あり学習を利用した説明性を有する POI 推薦手法”, DEIM Forum 2023 5a-5-4, 2023

特定ドメインの問合せを対象とした大規模言語モデルの性能計測手法の検討

神田　智也†　　伊藤　寛祥††　　森嶋　厚行††

† 筑波大学大学院 人間総合科学学術院 人間総合科学研究群 情報学学位プログラム 博士前期課程
†† 筑波大学図書館情報メディア系
〒 305–8550 茨城県つくば市春日１－２
E-mail: †s1911475@klis.tsukuba.ac.jp, ††{ito,mori}@slis.tsukuba.ac.jp

あらまし　近年の様々な研究によって大規模言語モデルは追加学習なしで様々な下流タスクで高い性能を発揮するようになった. それらの性能は自然言語処理におけるベンチマークデータセット等で計測・比較されることが多い. しかし, 言語モデルの実際の利用においては言語モデルにクエリを入力した後に都度評価を行う必要がある. つまり, 良い結果が得られるまで言語モデルに何度も同じ入力を与えたり, より良い結果を得るために入力プロンプトを変更したりして試行錯誤を行うのである. この試行錯誤には時間的・金銭的なコストがかかる. 我々は言語モデルへの問い合わせを効率化するために, 事前に言語モデルの詳細な性能を予測するための手法を提案/検討する. 詳細な評価指標を得るために, 実タスクにおける述語単位での性能計測クエリを発行する. 本論文では, 述語単位の評価指標を得る手法を複数個比較し, より効率的な計測方法の検討を行った.

キーワード　評価・データセット，LLM

1　はじめに

近年急速に発展する大規模言語モデル (LLM: Large Language Model, 例えば GPT-3 [2], LLaMA [11], PaLM [4] 等) の性能は Big-bench [1] や SuperGLUE [12] 等の複数のベンチマークデータセットにおいて測定され, 様々な自然言語処理におけるタスクで高い性能を示すことが明らかになっている [10]. ベンチマークスコアはモデル間の性能比較や, 大まかな性能指標となる.

しかし，利用者にとって重要な関心は，その利用者が投げる特定の問合せに対して，どれぐらいの品質の結果が期待されるかである．例えば「現在のパリーグのチームに所属する日本のプロ野球選手を列挙せよ」といった問合せに対して，どれぐらいの品質の結果が期待されるかを，これらのベンチマークの結果から推し量ることは困難である．一方，それぞれの問合せ結果に対して個別に品質検証を行うことはスケーラビリティに欠ける．つまり,「ベンチマークによる総合性能指標」と「各試行の後の評価」の２つの間に位置づけられる指標 (図 1) が必要であると考える.

また，ベンチマークデータセットの１つである TruthfulQA の各アイテムをトピックごとに分類して分析した結果，大規模言語モデルは扱うトピックによって異なる性能を示すことが示されている [9]．そのため，ベンチマークデータセットのような多くのトピックを網羅した結果の性能ではなく，扱う問題のトピックごとに性能を推定する必要がある.

本論文は，大規模言語モデルの性能評価の一手法として，特定ドメインの述語レベルでの性能を見積もるというアプローチを提案する．例えば,「現在のパリーグのチームに所属する日本のプロ野球選手を列挙せよ」といったような日本のプロ野球に関する問合せは，次のような方法で記述できる.

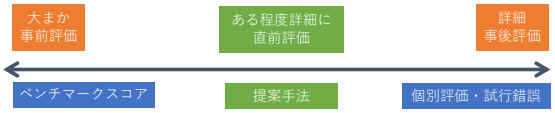


図 1　提案手法の位置づけ

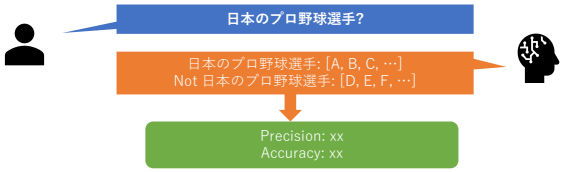


図 2　提案手法概略

- パリーグのチームに所属する日本のプロ野球選手 (p)
- パリーグ (t), 日本のプロ野球選手 (p), 所属 (p,t)

前者は全体を述語化しており，後者は複数の述語に分解している．本論文ではこれらの個々の述語に対応するプロンプトを大規模言語モデルに入力した結果の品質を推測する良い方法を検討する．具体的な述語化の方法について限定するものではないが，後者のように細かく分解しておくと結果を再利用することもできる.

本論文ではまず二値分類述語 $P(x)$ を対象とする．それ以外の形式の述語については今回は直接は扱わないが，類似のアプローチで可能と考えている.

したがって，本研究の学術的問いは,「少ないコストで大規模言語モデルにおけるこれらの述語の性能を見積もるにはどうすればよいか」である．我々の知る限り，本研究はこのような述語レベルの性能推定を行うための初めての研究である.

具体的には，本論文では図 2 に示すように, 大規模言語モデルに問い合わせたい概念 P に対して「$P(x)$ が真となる x」と「$P(x)$ が偽となる x」を問い合わせるプロンプトを作成して大規模言語モデルに列挙させ, その結果から実際の適合率と正解率を推定する枠組みを考える．本論文では，具体的なプロンプトと推定手法をいくつかの検討し，実験を通じてこれらを比較する．

本論文では，次のようなリサーチクエスチョンを設定する．(RQ1) 少数の例の列挙から推定する際に，指示する例の数によって推定性能の正確さはどのように変化するか．(RQ2) 推定手法のバリエーションによって，どのような違いが出るか．(RQ3) 特定ドメインに関する LLM 性能の推定はどの程度現実的か．

本論文では 2 つの実験を通して以下の 3 つの結果を得た．

（1） 列挙する例の数を増やすほど推定は正確になる傾向があるが，20 個程度の列挙個数があれば，性能の高低をある程度判断することは可能である．

（2） 負例列挙の際に上位概念を踏まえた負例を列挙させると，推定性能指標が向上する．

（3） 提案手法は，ある分野について大規模言語モデルが不得意な場合は，それを高確率で判断できる．

本論文の構成は次の通りである．2 章では関連研究について述べる．3 章では提案手法の詳細について述べる.4 章では提案手法の分析のための実験とその結果について述べる.5 章では既存のベンチマークデータセットのスコアと提案手法によって推測したスコアを比較する実験とその結果を示す.6 章で総括と今後の発展について述べる．また, 末尾に今回の実験で大規模言語モデルに対して用いたプロンプトを示す.

2 Related Works

2.1 大規模言語モデルの評価手法

大規模言語モデルの性能評価に関しては様々な指標が提案されている [3]. 大規模言語モデルは多種の下流タスクに応用可能であり, それぞれに適した評価指標は異なる．そのため, 様々な評価タスクと評価指標を備え, 場合によってはそれらに基づき全体のスコアを算出する自然言語処理データセットによって評価されることが多い.

個別の評価指標としても様々なものがある．例えば BertScore [14] は言語モデル Bert [5] を利用した評価手法であり,Bert の出力するトークンを利用して精度, 再現率が定義されており, そこから F1 値を得る．評価対象となる大規模言語モデル自身に対する問い合わせによって性能を測定する手法も提案されている.Neel らの手法 [7] は大規模言語モデルの出力の一貫性をみることで性能を評価している．これら大規模言語モデルの評価に大規模言語モデルを用いる手法はコストが大きい.

2.2 プロンプト戦略

大規模言語モデルに与える入力（プロンプト）を工夫することで出力の品質を向上させることが可能であると知られている．プロンプト戦略には様々なものが提案されている．例えば，大規模言語モデルには与えられたプロンプトによって性能を向上させる特性（文脈内学習, ICL: In-Context Learning [6]）があり, それを活用したプロンプト戦略がいくつかある．それを活用したものとして, 例えば,Chain-of-Thought [13] はプロンプト内に思考過程の例示を含めることで追加学習なしでも性能を向上させることを示した. また,Zero-Shot Chain-of-Thought [8] はプロンプトに “Let’s think step-by-step” と加えるだけで性能向上が確認されている．しかし, 文脈内学習等の考えに基づく手法は多かれ少なかれプロンプトを長くするものである．モデルに入力可能なプロンプトの長さには上限があり, また, 長い入力は推論にコストがかかる.

以上の点を考慮して, 本論文で提案する手法は可能な限り短い, 基本的なプロンプトのみで構成することとする．将来的にはプロンプト長の増加に伴うコストとのトレードオフを考慮しつつ, 上記に挙げたような様々なプロンプト戦略を取り入れることも検討したい.

Algorithm 1 提案手法

Require: A target model M, a topic T, a predicate P, prompts for models to pick up positive and negative items p_p, p_n a number of examples for pick up n
Ensure: Get consistency of confusion matrix, TP, FP, TN, FN
1: $positive_items \leftarrow PickUpItems(M, P, T, p_p, n)$
2: $negative_items \leftarrow PickUpNegativeItems(M, P, T, p_n, n)$
3: $TP \leftarrow VerificateItems(positive_items)$
4: $FP \leftarrow n - TP$
5: $TN \leftarrow VerificateItems(negative_items)$
6: $FN \leftarrow n - TN$
7: **return** TP, FP, TN, FN

3 手　　法

提案手法の流れをアルゴリズム 1 に示す．まず, 大規模言語モデル M に対してある概念 T に当てはまるものと当てはまらないものの例をプロンプト p_p, p_n を用いて列挙させる（1 行目 $PickUpItems$，2 行目 $PickUpNegativeItems$）．次に, 列挙された結果を検証することで, 大規模言語モデルの該当概念に対する性能を測定する．3 行目と 5 行目の $VerificateItems$ で列挙されたアイテムが正しいかどうか判定し，正しいものの個数（3 行目は True Positive，5 行目は True Negative）を得る．得られた True Positive と True Negative から False Positive と False Negative を得る（4 行目，6 行目）.

これら正例と負例列挙，それらの検証の大規模言語モデルへの問い合わせはそれぞれ別の文脈による問い合わせで行う．これは文脈内学習によって各問い合わせが他の問い合わせに影響することを防ぐためである.

3.1 列挙サンプル数

統計値を得るためには大規模言語モデルに「ある概念 x に該当する」という意味の述語 $P(x)$ を満たす複数のサンプルを列挙させる必要があり，それを何個列挙させるかが問題となる．しかし,「A であるものを全て列挙せよ」という指示は全ての概念に対して適用できない．なぜなら, 概念 A に該当するアイテムが大量もしくは無数にある場合, 出力が長大になったり, 大規

模言語モデルの出力の上限によって制限されて全てを出力できない場合があるためである．そこで我々は列挙する個数 n を予め与えて「A に該当するものを n 個列挙せよ」という指示を与える必要がある．加えて，大規模言語モデルの入出力にはその長さに応じて，時間的なコストや場合によっては金銭的なコストが生じる．そのため，コスト面からより少ない列挙アイテム数 n が望ましい．

また，例えば 47 個しかない日本の都道府県を対象とする場合において列挙させる数 n を 100 とすると，そのうち 53 個は存在しないものを列挙させることとなり，正確な性能推定ができない．対象とする概念に該当するアイテムの個数が少ない場合を考えると，より列挙アイテム数が少ないほうが良い．

以上のようにコスト面，正確性の観点から, 列挙する個数 n は小さいほど好ましい．

3.2 負例プロンプトの戦略

大規模言語モデルに対して, ある概念 A に該当するものを列挙させる場合は単純に「A に当てはまるものを n 個列挙してください」というような指示を与えることで実現が可能である．しかし, 負例として「A でないものを挙げさせる」場合は検討すべき点がある．単に「A でないもの」と言った場合, 暗にその母集合として全概念を示していると考えられ,「全概念に対して A でないものを列挙せよ」と大規模言語モデルに対して指示することになる．

しかし, 負例の候補の範囲としてそれは適切ではないと考えられる．より適切な負例候補の範囲，より狭い範囲の負例として「A の上位概念には含まれるが A でないもの」が考えられる．そこで, 大規模言語モデルに対して概念 A の負例を列挙させる際のプロンプト戦略として本論文では以下の 2 つを比較する．

- （特に前置きをせずに,）A でないものを列挙させる
- A の上位概念ではあるが,A でないものを列挙させる

3.3 検 証 手 法

大規模言語モデルが列挙したアイテムの検証は, 大規模言語モデル自体に再度問い合わせることで実施する．すなわち, 概念 A の例として大規模言語モデルが a を挙げた場合,「a は A ですか」と問い合わせることで検証する．

列挙したアイテムの検証を既存のナレッジベース, 例えば Wikipedia 等によって行うことも考えられる．しかし, それらは必ずしも全てのアイテムを網羅しているわけではないため, 適切とは考えられない．本論文では, モデル自身による検証の比較対象として DBpedia に基づいたデータセットによる検証を行う．

3.4 得られる指標

本論文で提案する手法では, 例の正解/不正解から真陽性と偽陽性が得られる．また, 負例の正解/不正解から, 真陰性と偽陰性が得られる．これらから混同行列を作成することができるが, 大規模言語モデルに対して正例と負例を独立に問い合わせており, 独立に検証しているという点に注意が必要である．

そのため, 真陽性と偽陽性から計算できる適合率と, 真陰性と偽陰性から計算できる特異率はそれぞれ得ることができる．列挙させるサンプル数を n, 真陽性の数を TP, 真陰性の数を TN とすると,

$$Precision = TP/n$$

$$Specificity = TN/n$$

ここで, 偽陽性の数 FP と偽陰性の数 FN はそれぞれ,$FP = n - TP$, $FN = n - TN$ であるから正解率の定義より,

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

として, 正解率が求められる．

しかし, 真陽性と偽陰性から得られる再現率をこの混同行列から計算するのは不適切であると考える．例えば,5 つ例を列挙させる場合に 3 つ真陽性と 4 つの偽陰性がある場合を考える．この場合においてもし真陽性と偽陰性に共通して, あるアイテム i が挙げられていたとする．通常の混同行列ではこのようなことは起こり得ないが, 本手法では独立して質問しているために重複が起こりうる．

表 1 実 験 条 件

列挙アイテム検証方法	自身, Dataset
負例プロンプト戦略	通常, 上位概念
列挙サンプル数	5, 10, 15, 20, 25

3.5 プロンプト

本文中では簡便のために日本語でプロンプト例を示しているが, 前述したデータセットを用いた検証のために英語で実験を行う必要がある．そのため, 実際には英語のプロンプトを大規模言語モデルに与えた．実際に与えたプロンプトは末尾の付録に示す．

4 性能推定実験

実際のベンチマークデータセットのスコアとの比較の前に、本手法の特徴を調査するための実験を行った．目的は以下の通りである．

（1） 列挙アイテム数の変化による指標の変化の確認.

（2） 負例を列挙させるプロンプト戦略の比較.

（3） 提案手法が異なるトピックに対して異なる presision, accuracy を示すか確認する．また，その傾向がデータセットによる検証と一致するか.

4.1 実 験 設 定

実験の条件を表 1 に示す.

列挙するアイテムを検証する方法は, 大規模言語モデル自身に問い合わせるものと,DBpedia に基づくデータセットとの照合による検証の 2 つを行う.

対象データセット

検証方法の比較に用いるデータセットには dbpedia_14 [15] を用いた．これは本来はテキスト分類のためのデータセットで

あるが,DBpedia[1]に基づいて概念の包含関係を示しており, それぞれに重複がないため, 検証の比較対象として適切であると考える. 本論文では,dbpedia_14 の 14 個のラベルを対象として実験を行った.

大規模言語モデル設定

大規模言語モデルに列挙させる例の個数が, 対象とする概念が含むアイテムの個数に近づくと, 列挙の難易度が上がると予想される. 例えば日本の都道府県は 47 しかないため, そのうち 40 個を列挙させると, 後半ほど難易度が上がると考えられる. もしくは, 都道府県を 100 個列挙するように指示すると架空の都道府県を挙げたり, 同じ都道府県を挙げたりして正確に性能評価ができない. しかし, 今回は列挙させるサンプル数の上限を 25 としており,dbpedia_14 の訓練用データは各ラベル 40000 個のアイテムを持っているため, その懸念はないものと考える.

今回利用した大規模言語モデルは OpenAI 社の大規模言語モデル gpt-3.5-turbo-instruct である. 広く利用されている大規模言語モデルサービスである ChatGPT のモデルとして知られる GPT-3.5-turbo に近いモデルであり, より現実的な検証ができると考える. 大規模言語モデルのベンチマークを行う際は出力を固定するようにハイパーパラメータを設定することがあるが, 今回は OpenAI API のデフォルトの $temperature = 1.0$ として, 出力にある程度多様性を持たせた. そのかわりに各条件, 各データで 10 回ずつの試行を行った.

4.2 結　　果

実験結果の概要を表 2, 3, 4, 5 に示す. また, サンプル数の変化による各ラベルごとの適合率 (precision), 正解率 (accuracy) の変化を図 3, 4, 5, 6 に示す. 値は各条件の複数回の試行の平均値である. 大規模言語モデルを扱うため, 必ずしも期待される結果が得られるとは限らない. 今回の場合,Python のリスト形式での結果を要求してそれを読み込んでいるが,Python のリストとして不適切な形式の出力が得られる場合もある. その場合には 3 回までリトライをするようにしている. 3 回とも Python のリストとして解釈不能な結果だった場合, そこは欠損値となる.

（1）列挙サンプル数と指標の関係

傾向として, よりサンプル数が少ない条件で, データセットによる検証が高い適合率と正解率を示した. これは各表の下に示した#samples と accuracy, precision の相関係数によっても示されている.

ただし, 提案手法は大規模言語モデルの性能を推定するものであるから, 適合率と正解率が高いことが提案手法の良さを示す訳では無い.

（2）負例プロンプト戦略の比較

表 2 と表 3, 図 3 と図 4 の比較から, 負例プロンプト戦略の違いを分析する. 負例プロンプトにおいて上位概念を訊ねる戦略の方が TN が多くなる結果となっており, Accuracy の値が大きくなる傾向にある. 上位概念を訊ねるプロンプトの方がより適切に負例を列挙できていると考えられる.

1：https://www.dbpedia.org/

#samples	accuracy	precision	TP	TN	FP	FN
5	0.515942	0.704348	3.521739	1.637681	1.478261	3.362319
10	0.471269	0.632836	6.328358	3.097015	3.671642	6.902985
15	0.452840	0.622222	9.333333	4.251852	5.666667	10.748148
20	0.438953	0.588760	11.775194	5.782946	8.224806	14.217054
25	0.413280	0.581440	14.536000	6.128000	10.464000	18.872000

correlation	accuracy	precision
#samples	-0.976562	-0.937000

表 2　検証:Dataset, 負例：通常, データ数:661/700

#samples	accuracy	precision	TP	TN	FP	FN
5	0.610476	0.693333	3.466667	2.638095	1.533333	2.361905
10	0.604545	0.666364	6.663636	5.427273	3.336364	4.572727
15	0.532051	0.594872	8.923077	7.038462	6.076923	7.961538
20	0.506683	0.559406	11.188119	9.079208	8.811881	10.920792
25	0.554607	0.608539	15.213483	12.516854	9.786517	12.483146

correlation	accuracy	precision
#samples	-0.733254	-0.803003

表 3　検証:Dataset, 負例：上位概念, データ数:509/700

#samples	accuracy	precision	TP	TN	FP	FN
5	0.905797	0.828986	4.144928	4.913043	0.855072	0.086957
10	0.839695	0.703053	7.030534	9.763359	2.969466	0.236641
15	0.812745	0.652451	9.786765	14.595588	5.213235	0.404412
20	0.783942	0.597810	11.956204	19.401460	8.043796	0.598540
25	0.785778	0.612444	15.311111	23.977778	9.688889	1.022222

correlation	accuracy	precision
#samples	-0.929964	-0.912541

表 4　検証: 自身, 負例：通常, データ数: 677/700

#samples	accuracy	precision	TP	TN	FP	FN
5	0.671014	0.820290	4.101449	2.608696	0.898551	2.391304
10	0.647810	0.673723	6.737226	6.218978	3.262774	3.781022
15	0.607246	0.647343	9.710145	8.507246	5.289855	6.492754
20	0.614179	0.585448	11.708955	12.858209	8.291045	7.141791
25	0.650382	0.609771	15.244275	17.274809	9.755725	7.725191

correlation	accuracy	precision
#samples	-0.443444	-0.875335

表 5　検証: 自身, 負例：上位概念, データ数：678/700

また, 表 4 と表 5, 図 5 と図 6 を比較すると, 自身での検証においてその差はより顕著になっている.

(RQ2) 推定手法のバリエーションによって, どのような違いが出るか.

(A2) 負例列挙の際に上位概念を踏まえた負例を列挙させると, 推定性能指標が向上する.

（3）トピックごとの分析

図 3, 4, 5, 6 より, ラベルによって性能差がある程度あることが確認できる. 特にトピック “Village”はどの条件においても Accuracy, Presicion ともに他より低い値を示している. また, 全体としてはサンプル数が少ないほど正解率, 精度ともに高くなる傾向にあるが, 各トピックに対しては値の傾向は安定しない. これはサンプル数の間隔が小さすぎることに加えて各試行が 10 回と少ないことが要因として考えられる.

（4）自己検証の妥当性の検証

また, 自己検証による難易度指標が各トピックの難易度を正確に推定できるか検証した. すなわち, 大規模言語モデルが自身の得意不得意なトピックを自身によって推定する能力を検証

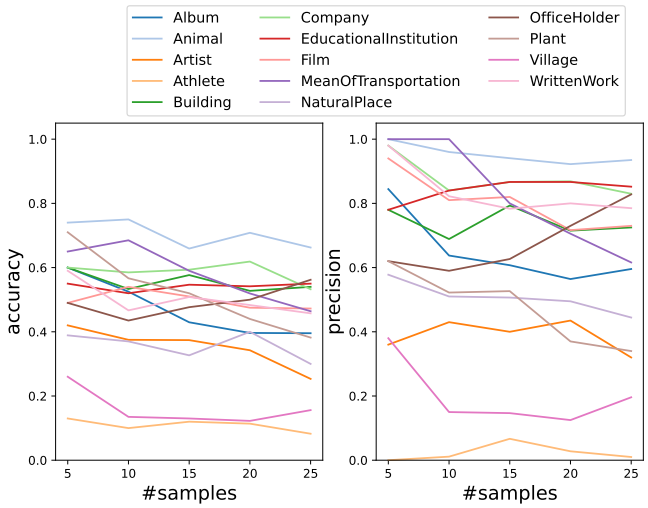


図 3　検証:Dataset, 負例：通常

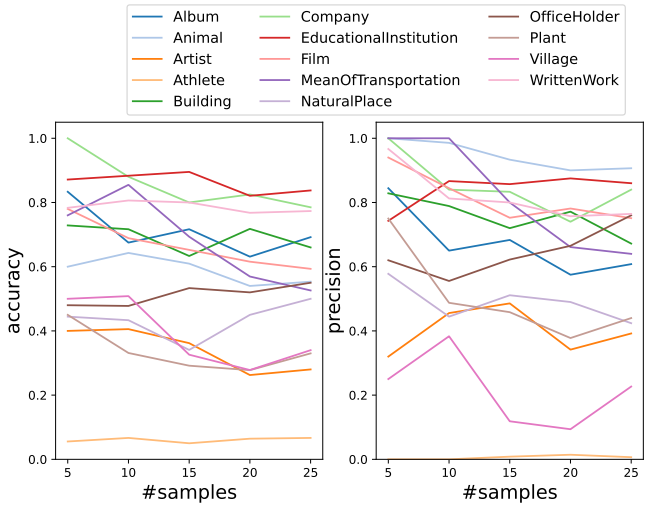


図 4　検証:Dataset, 負例：上位概念

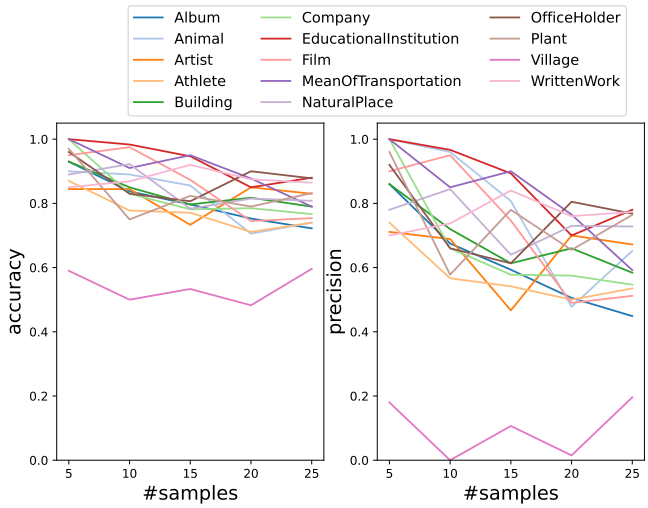


図 5　検証: 自身, 負例：通常

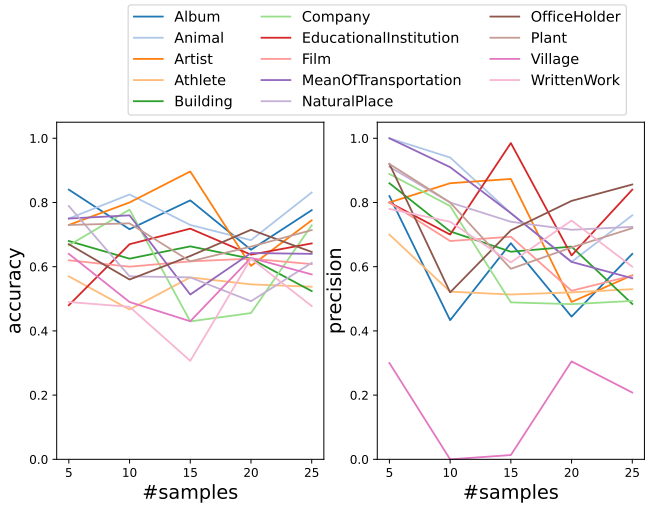


図 6　検証: 自身, 負例：上位概念

した．Dataset による検証の結果をトピックごとに，難易度指標 (presicion または accuracy，対象大規模言語モデルの推定性能) が 0 から 0.5(難しい) クラス，0.5 から 0.75(比較的易しい) クラス，0.75 から 1.0(易しい) クラスの 3 クラスに分類した．Dataset による検証と同じように自己検証が性能を推定し，同じ難易度指標クラスに分類できるか確認した結果を表 6 に示す．列挙サンプル数 $\#sample = 20$ の際の値である．推定 accuracy の範囲への分類問題の混同行列 (表 6 右側) において，混同行列の正答を表す対角線より下の値が 0 であることから，提案手法は「各トピックにおける対象大規模言語モデルの性能を高く推定し過ぎることはない」，すなわち,「不得意なトピックを誤って高い性能であると推定しない」といえる．

(RQ3) 特定ドメインに関する LLM 性能の推定はどの程度現実的か．

(A3) 提案手法は，ある分野について大規模言語モデルが不得意な場合は，それを高確率で判断できる．

表 6　自己検証の性能の検証．トピックの性能を presicion 区分, accuracy 区分のそれぞれの範囲のクラスに分類する問題における混同行列．$\#sample = 20$

負例プロンプト戦略：単純

precision 区分

pred	[0, 0.5]	[0.5, 0.75]	[0.75, 1.0]
true:[0, 0.5]	2	3	0
true:[0.5, 0.75]	1	2	2
true:[0.75, 1.0]	1	2	1

accuracy 区分

pred	[0, 0.5]	[0.5, 0.75]	[0.75, 1.0]
true:[0, 0.5]	1	2	6
true:[0.5, 0.75]	0	1	4
true:[0.75, 1.0]	0	0	0

負例プロンプト戦略：上位概念

precision 区分

pred	[0, 0.5]	[0.5, 0.75]	[0.75, 1.0]
true:[0, 0.5]	1	3	1
true:[0.5, 0.75]	0	1	1
true:[0.75, 1.0]	0	4	3

accuracy 区分

pred	[0, 0.5]	[0.5, 0.75]	[0.75, 1.0]
true:[0, 0.5]	0	3	2
true:[0.5, 0.75]	0	4	5
true:[0.75, 1.0]	0	0	0

5　比 較 実 験

提案手法と実際のベンチマークスコアのスコアを比較するために，ベンチマークデータセットの各アイテムに対して提案手法を用いて評価指標を得る実験を行う．実際の利用例に近い検

証ができるように，質問応答（QA タスク）のベンチマークデータセットとの比較を行う．今回は TruthfulQA [9] を用いた．

TruthfulQA にはいくつかの評価指標があるが，今回は人間によって評価された大規模言語モデルの出力のうち正しいもの割合である $\%true$（を 100 で割った値）と今回の提案手法によって推定した適合率を比較する．適合率の算出は正例の列挙のみで可能であるため，負例の列挙戦略の違いは関係ない．

5.1 実験設定

データセット及びトピック

TruthfulQA の全 817 アイテムに対して実験を行う．TruthfulQA は基本的に質問とその回答の組になっている．提案手法はトピックをもとに大規模言語モデルの性能を推定するものであるが，TruthfulQA からは質問のジャンルやトピックに相当するものは与えられない．大規模言語モデルの実際の利用例においてもトピックは明示的には与えられないことが多いため，今回の実験においては大規模言語モデル自身に付録に示す簡単な質問によって質問のトピックを挙げさせることにする．質問からトピックを尋ねるプロンプトは付録に示す．

このように列挙させたトピックが重複することもあり，全 817 アイテムに対してトピック数は 500 個となった．最も数の多いトピックは “Geography”で，20 アイテムであった．また，トピックあたりのアイテム数の平均は 1.634 であった．トピックごとにアイテム数の偏りがあるため，結果の分析は各トピックのアイテム数で重みをつけた重み付き平均で評価する．

大規模言語モデル設定

いくつかのモデルに関して TruthfulQA のスコアが公開されており [2]，それらの中でも OpenAI 社の提供する GPT モデルは広く利用者を得ているモデルである．そのため，本論文でも GPT モデルとの性能比較を行う．しかし，OpenAI 社のモデルの多くは非公開であり，現在利用可能なモデルの TruthfulQA のスコアは公開されていない．そこで，今回の比較では InstructGPT(175B) の TruthfulQA のスコアと，現在利用可能なモデルである gpt-3.5-turbo-instruct-0914(以後，GPT-3.5-Instruct) に提案手法を適用したものを比較する．

シミュレーション実験とは異なり既存のベンチマークスコアとの比較が目的であるため，大規模言語モデルのハイパーパラメータの設定も $temperature = 0$ として出力を固定し，同条件同データに対する試行回数は 1 回とした．ただし，出力結果の検証においては期待される結果でない場合において訂正が必要であるため，$temperature = 0.7$ として，検証失敗時にリトライするようにした．

5.2 実験結果

図 7 に提案手法と既存のスコア [9] の比較を示す．推定適合率は TruthfulQA における InstructGPT の%true の実際の値よりも高く，列挙サンプル数が多くなるほど実際の値に近づいた．列挙させるサンプル数が多くなるほど，推定コストが大きくなることを踏まえると，$\#sample = 20$ あたりでも列挙個数に対して十分性能を推定できていると考える．(RQ1) 少数の例の列挙から推定する際に，指示する例の数によって推定性能の正確さはどのように変化するか．(A1) 列挙する例の数を増やすほど推定は正確になる傾向があるが，20 個程度の列挙個数があれば，性能の高低をある程度判断することは可能である．

2：https://paperswithcode.com/dataset/truthfulqa

	accuracy (normal)	accuracy (super)	precision
#samples	-0.806969	0.188192	-0.760504

表 7　TruthfulQA における accuracy と precision の列挙サンプル数との相関．accuracy の負例列挙方式：通常 (normal)，上位概念 (super)．precision は負例列挙方式が通常の際の実験の値．

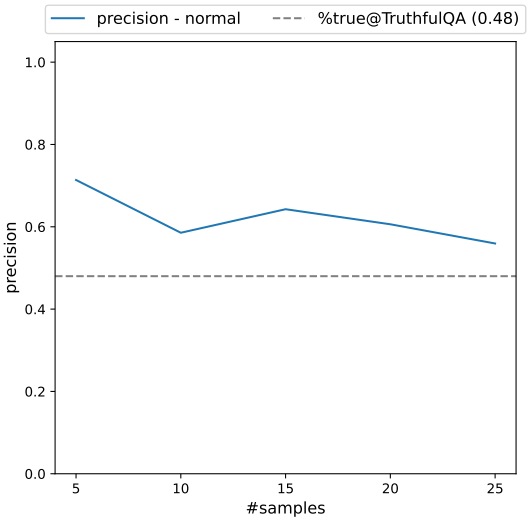


図 7　TruthfulQA における InstructGPT の $\%true/100$(破線) と，GPT-3.5-Instruct に提案手法を用いた際の適合率の比較．横軸は提案手法における列挙サンプル数．

表 7 に列挙サンプル数と推定 Accuracy,Precision の相関を示す．シミュレーション実験でも示されたように，列挙サンプル数が増加するほど適合率の値が低く推定される傾向にあるといえる．しかし，正解率に関しては上位概念を踏まえた負例プロンプト戦略において列挙サンプル数との相関が見られなかった．

6 おわりに

本論文では, 大規模言語モデルの詳細な性能評価を事前に低コストで行う手法の提案を目的として，大規模言語モデルにアイテムの例を列挙させてそれを自身によって検証させる手法を提案した．ベンチマークデータセットのような外部の大規模なデータを必要とせずに，個別具体的なトピックに関する大規模言語モデルの性能を，対象となるモデル自身に問い合わせることによって推定することの可能性を実験の範囲で示した．

また，20 個程度の列挙個数によってある程度性能の推定が可能であることが示されたが，実際にこれがどの程度の低コストであるかは検証できていない．今後，入出力トークン数を含めた分析が必要である．

今後この得られた評価指標：適合率と正解率の妥当性の検証，

すなわち, 得られた指標がどれだけ真の性能指標に近いかに関する検証をより多くの既存ベンチマークと行うことを計画している.

また，本論文の実験は限定的であり，複数の大規模言語モデルに共通する傾向や列挙サンプル数による傾向は十分に得られていない．今後は複数種類のモデルに対する検証や広範なサンプル数での検証が必要であると考えている.

謝　　辞

本研究の一部は JSPS 科研費 (JP22H00508), と JST CREST(Grant Number JPMJCR22M2)，AIP チャレンジの支援を受けたものである．ここに謝意を示す.

文　　献

[1] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models, October 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways, October 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A Survey on In-context Learning, June 2023.

[7] Neel Jain, Khalid Saifullah, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Bring Your Own Data! Self-Supervised Evaluation for Large Language Models, June 2023.

[8] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners, January 2023.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, February 2023.

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[12] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, February 2020.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023.

[14] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, February 2020.

[15] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc., 2015.

付　　録

利用プロンプト

大規模言語モデルに与えたプロンプトを以下に示す．{} で囲まれた部分は変数である．変数の対応を表 8 に示す．

例を列挙させるプロンプト

```
You should answer with the literal of list of Python with all string item.  For example, ["example1", "example's 2", "3 examples"].
Please pick up some examples of {label}.  You need to pick up {n_examples} examples.
```

負例を列挙させるプロンプト

```
You should answer with the literal of list of Python with all string item.  For example, ["example1", "example's 2", "3 examples"].
Please pick up some examples which are not {label}. You need to pick up {n_examples} examples.
```

負例を列挙させるプロンプト：上位概念を聞くパターン

```
You should answer with the literal of list of Python with all string item.  For example, ["example1", "example's 2", "3 examples"].
Please pick up some examples which are the superordinate of {label}, but not {label}.  You need to pick up {n_examples} examples.
```

検証させるプロンプト

```
You should answer with the literal of list of the programming language, Python and its contents should be list of 'bool'.
Even when all items are the same answer, you should answer with the list of bool.
For example, [True, True, False, True, False], [False, False, False, False, False]
The {item} is a kind of {label}?
```

トピック列挙プロンプト

```
Please pick up the topic of the class of the following statements, in a word.  Statements: {question} Topic:
```

表 8　プロンプトの変数

label	対象となる概念
n_examples	列挙させる例の個数
item	列挙されたもの
question	質問文

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

大規模言語モデル (2)

座長:梅本 和俊(東京大学)

コメンテータ:服部 峻(滋賀県立大学)

2024年3月1日(金) 16:00 〜 18:10 T3-A (オンライン（Zoom Events))

17:45 〜 18:10

[T3-A-9-05]【技術報告】リクルートの横断検索基盤における取り組み紹介

新田洸平 (株式会社リクルート)

©一般社団法人日本データベース学会

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

🏛2024年3月1日(金) 16:00 ~ 18:10 | 🏛T3-B オンライン（Zoom Events)

推薦モデル・評価

座長:佐々木 勇和(大阪大学)
コメンテータ:江口 浩二(広島大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-B-9-01] 複数回クラスタリングによる協調フィルタリングを用いた推薦システム
*FU LEI[1]、張 建偉[1] (1. 岩手大学)

16:55 ~ 17:20
[T3-B-9-03] Light distillation for Incremental Graph Convolution Collaborative Filtering
*Fan Xin[1]、Fan Mo[1]、Chen Chongxian[1]、山名 早人[1] (1. 早稲田大学　山名研究室)

17:20 ~ 17:45
[T3-B-9-04] グラフニューラルネットワークによる分離表現学習を用いた推薦の多様化
*杉山 拓都[1]、吉田 壮[1]、棟安 実治[1] (1. 関西大学大学院 理工学研究科)

17:45 ~ 18:10
[T3-B-9-05] Polya–Gamma分布を使ったBayesian Factorization Machines
*北原 洋一[1] (1. 株式会社リブセンス)

複数回クラスタリングによる協調フィルタリングを用いた推薦システム

FU LEI† 張 建偉†

† 岩手大学理工学研究科 〒 020–8551 岩手県盛岡市上田 4-3-5
E-mail: †{s3223006,zhang}@iwate-u.ac.jp

あらまし 協調フィルタリングは、過去数年間で過剰な情報の問題に対処するための効果的な手法として台頭している。しかし、シリング攻撃では、攻撃者が偽の評価を導入して推薦システムを影響することができる。本研究では、データセットに対する複数回クラスタリングを行って、攻撃の影響を軽減し、予測精度を維持または向上させる協調フィルタリング手法を提案している。複数回クラスタリングの効果は、映画に関する大規模なデータセットを用いて実験的に探求されている。提案手法は、実際のユーザ評価と予測評価の間の誤差を測定する方法を通じて評価する。また、攻撃前後の予測シフトや推薦リストの差分を比較することで、協調フィルタリングに対するロバスト性も調査する。研究結果は、複数回クラスタリングが協調フィルタリングに対して推薦精度とロバスト性が上がることを示している。

キーワード 推薦モデル、協調フィルタリング、シリング攻撃

1 はじめに

推薦システムは情報を処理する際に製品やサービスを候補から推薦することで人々をサポートする。協調フィルタリング（CF）は、推薦システムで使用される代表的な手法の一つである [1] [2]。協調フィルタリングがユーザ間の類似度を推定し、そのユーザに似たユーザの評価に基づいて評価しないアイテムの評価値を予測する。協調フィルタリングはインジェクション攻撃に対して脆弱性を抱えている [3]。インジェクション攻撃では、攻撃者が偽のユーザプロファイルを推薦システムに注入し、推薦結果に影響を与える。この欠点を取り除くことは、推薦システムの信頼性向上に重要である。協調フィルタリングは、ターゲットユーザに似たユーザを検索し、そのユーザの好みに基づいてアイテムを推薦する。類似したユーザをクラスタリングすることで、予測精度が向上すると期待される。本研究では、複数回クラスタリングを行い、各クラスター内で評価値を予測する手法を提案している。提案手法を評価するために、実際のユーザ評価と予測評価の誤差が測定される。攻撃に対するロバスト性は、予測シフトと TopN シフトの数値を比較している。

2 関連研究

研究に基づいて推薦システムに対する攻撃に関連する研究課題と手法の概要を提供している [4] [5] [6]。協調フィルタリングの攻撃は、特定の攻撃ユーザと推薦システムのロバスト性向上の、2 つの大きなカテゴリに分類できる [7]。攻撃ユーザを検出する手法には、統計的手法 [8]、攻撃の識別 [9]、およびクラスタリング手法の使用が含まれる。クラスタリングに基づく手法の 1 つは、定期的なクラスタリングを実施し [10]、クラスタの代表点に大きな変化がある場合に攻撃を検知する手法である。またクラスタリングに基づく別の手法としては、攻撃ユーザを検知する手法がある [11]。頑健な推薦システムを開発する手法には、攻撃検出 [12]、SVD ベースの協調フィルタリングアルゴリズム [13]、および段階的なヒントトレーニングなどが含まれる [14]。類似な研究では、クラスタリング手法を採用したモデルベースの協調フィルタリング を使用して攻撃を防御することを示している [15] [16] [17]。本研究の特徴は、2 回クラスタリング手法を導入し、その有効性を示すために比較実験を行っている。

3 提案手法

3.1 推薦システム

一般的な通販サイトなどでは多種多様かつ膨大な数のアイテムを取り扱っているため，サイトを利用するユーザが自力で求めるアイテムを探し出すのは非常に困難である．そのためにユーザが求めるアイテムを推薦するシステムはユーザに対してはもちろん，サイトの利便性を高める点において，サイト運営者に対しても有用である．

推薦システムに用いられる代表的な手法は 2 つあり，内容ベースフィルタリングと協調フィルタリングである．内容ベースフィルタリングは，あらかじめサイト内で扱うアイテムについての特徴をデータとして管理し，利用者から得た嗜好データに合うものを推薦する手法である．一方の協調フィルタリングは,「自分と似た嗜好のユーザが好むアイテムは，自分も好むだろう」という仮定を元に，アイテムに対しての評価から自分と似たユーザを探し出し，それらのユーザが好むアイテムを推薦する手法である．

3.2 協調フィルタリング

協調フィルタリング（CF）は、ユーザに推薦すべきアイテムを、そのユーザに似ているユーザがアイテムに行った評価を基に予測するシステムである。そのため予測の際にはあらかじめ似ているユーザ同士を同一クラスタに分類をしておくことで予

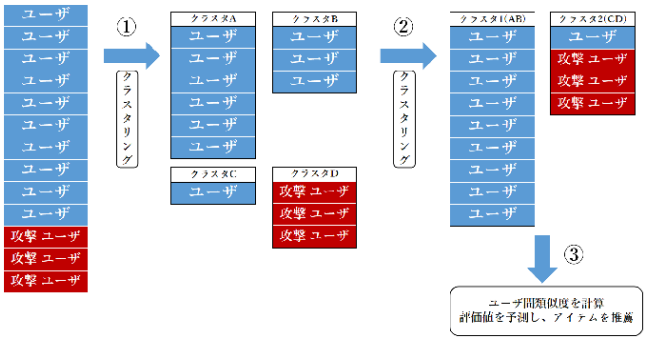


図 1　提案手法の流れ

測精度の向上が期待でき、また、攻撃ユーザが同一のクラスタに分類されれば、そのほかのデータは攻撃による影響を受けずに済む。しかし、攻撃ユーザが多くの一般のユーザに似ている場合、ユーザの分類を行うことでかえって攻撃の影響を増大させることも考えられる。

本研究の目標は、複数回クラスタリングを行い、1 つのクラスタ内のデータ数を増やすことで、推薦精度を維持または向上させ、攻撃の影響を軽減することを目指す。

3.3　提案手法の概要

提案手法は、データセットに対してクラスタリングを行い、ユーザをいくつかのクラスタに分けた後、それぞれのクラスタの中心点を計算し、再度クラスタリングを行うことで、各クラスタを結合する。その後、各クラスタ内のユーザ間の類似度を計算し、その類似度とユーザが付けた評価を用いて評価値を予測する。提案手法の流れを図 1 に示す。

3.4　クラスタリング

クラスタリングでは、データセット内のユーザを類似のクラスタにクラスタリングされる。まず、K-means クラスタリング [18] によって、ユーザが各クラスタに分けられ、次に各クラスタに属するユーザの中心点を使用して再度クラスタリングが実行される。データセット内のクラスタ C に属する各ユーザが $U_i(i = 1, \ldots, m)$ とマークされ、各アイテムが $I_j(j = 1, \ldots, n)$ とマークされ、ユーザ U_i がアイテム I_j に対して与えた各評価が R_{ij} とマークされる場合、クラスタ $C = (c_1, \ldots, c_j, \ldots, c_n)$ 内の各要素 c_j は次のように計算できる。第 1 回クラスタリングではクラスタ数を 2k、2.5k、および 3k に設定し、第 2 回クラスタリングでは k に設定する。

$$c_j = \frac{\sum_{i=1}^{m} R_{ij}}{m} \tag{1}$$

3.5　予測値の計算

協調フィルタリングは、2 つのステップで構成される。Step1 は、ユーザ間の類似度を計算し、似ているユーザを類似グループに形成する。ユーザ間の類似度は、ピアソンの積率相関係数を使用して計算する。ここで、$r(u,i)$ および $r(v,i)$ はユーザ u とユーザ v がアイテム i に付けた評価を表し、$\bar{r_u}$ および $\bar{r_v}$ はそれぞれユーザ u とユーザ v が付けた評価の平均値である。ユーザ u とユーザ v の類似度 $s(u,v)$ は次のようになる。

$$s(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r_u})(r_{v,i} - \bar{r_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r_u})^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r_v})^2}} \tag{2}$$

Step2 は、ユーザ間の類似度と似ているユーザが付けた評価に基づいて、ターゲットユーザの評価値を予測する。ユーザ u が付けた評価の平均値 $\bar{r_u}$ を使用して、ユーザ u のアイテム i への予測評価 $P(u,i)$ が計算される。このとき、最も類似したユーザ $v(v = 1, \ldots, N)$ がアイテム i に付けた評価 r_v, i と、ユーザ u が付けた評価の平均値 $\bar{r_u}$ とユーザ v が付けた評価の平均値 $\bar{r_v}$ の差が、ユーザの類似度 $s(u,v)$ で重みづけされる。計算式は次のようになる。

$$P_{u,i} = \bar{r_u} + \frac{\sum_{v \in N} s(u,v)(r_{v,i} - \bar{r_v})}{\sum_{v \in N} |s(u,v)|} \tag{3}$$

予測値の計算は、クラスタリングが行わない場合はすべてのデータを使用し、クラスタリングが行う場合は同じクラスタ内のデータのみを使用して計算される。

4　評 価 手 法

4.1　評価手法の概要

研究目標は攻撃の影響を軽減しつつ、推薦精度を維持または向上させることである。ユーザが評価したアイテムに対して協調フィルタリングを適用し、ユーザが付けた実際の評価と協調フィルタリングによる予測評価との誤差を測定することで推薦精度を評価する。さらに、研究における攻撃に対するロバスト性は、予測シフトと TopN シフトを構成されている。攻撃に対するロバスト性は、攻撃前と攻撃後の予測値の差を測定することで、予測評価の変化に対応している。同様に、攻撃に対するロバスト性では、攻撃前と攻撃後のユーザの推薦結果の差を測定することで評価される。これは攻撃前と攻撃後の TopN 値の差を計算し、攻撃前と攻撃後の協調フィルタリングが推薦結果の変化を反映している。

4.2　誤 差 測 定

ユーザが付けた評価と予測評価との誤差は、協調フィルタ リングの推薦精度を評価するために使用される。ユーザが評価した各アイテムについて、それを未評価と仮定し、この仮定に基づいてユーザがアイテムの予測評価が計算される。その後、ユーザがつけた評価と予測評価との誤差を測定できる。まず、攻撃前の誤差が図 2 に示されている。攻撃が注入されていない場合で、協調フィルタリングが正確な予測を行うことが証明される必要があるため、攻撃前の誤差測定が重要である。次に、攻撃後の誤差が図 3 に示されている。攻撃前後の予測値は協調フィルタリングのロバスト性を評価する部分で、攻撃前後の予測評価の変化を調査するために使用される。誤差の指標としては MAE（Mean Absolute Error），RMSE（Root Mean Squard Error） を用いる。

$$MAE = \frac{\sum_{k=1}^{N} |P_k - R_k|}{N} \tag{4}$$

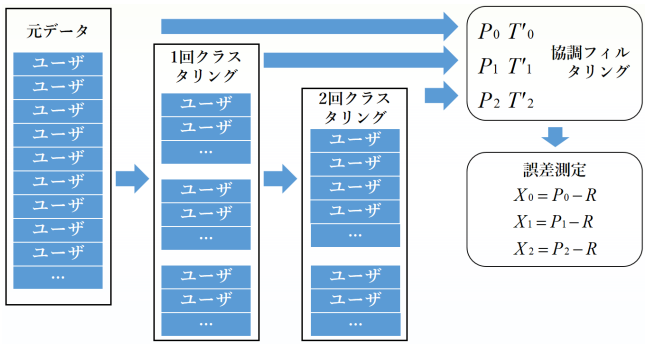


図 2　攻撃前の誤差測定

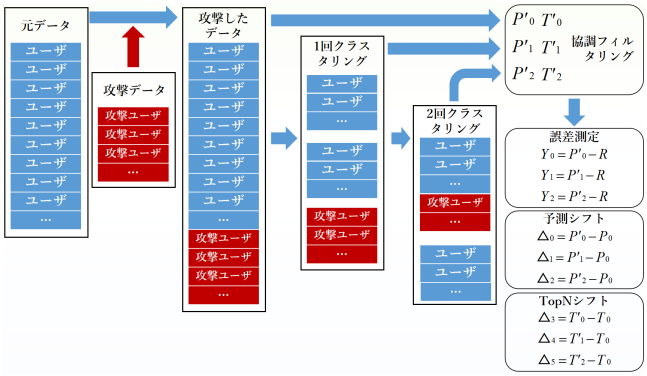


図 3　攻撃後の誤差測定とロバスト性評価

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(P_k - R_k)^2}{N}} \tag{5}$$

ここで、P_k は予測評価、R_k は実際の評価、N はユーザが評価したアイテムの数である。

提案方法では、第 1 回クラスタリングのクラスタ数を 2k、2.5k、および 3k に設定した場合、第 2 回クラスタリングのクラスタ数は k から始まりる。第 2 回クラスタリングでは、異なるクラスタ数を比較するために、2k → k 値、2.5k → k 値、および 3k → k 値の誤差が取られる。

4.3　攻撃に対するロバスト性

4.3.1　シリング攻撃

協調フィルタリングに対する攻撃にはいくつかの種類があり、実験ではランダム攻撃と平均攻撃が使用されている。ランダム攻撃では、各フィラーアイテムに対して評価が正規分布に基づいてランダムな評価値を与える。平均攻撃では、各フィラーアイテムに対して評価が各アイテムの評価の平均値を与える。攻撃の意図に応じて、特定のアイテムの評価を増加させることを目的とする攻撃を「プッシュ攻撃」と呼び、特定のアイテムの人気を減少させることを目的とする攻撃を「ニューク攻撃」と呼ぶ。本研究では、評価を増加させることを意図したプッシュ攻撃に焦点を当てている。プッシュ攻撃は、攻撃ユーザを元のデータに注入することで行われ、ターゲットアイテムがランダムに選択され、最高の評価が与えられる。ターゲットアイテム以外のアイテムはランダムに選択され、対応するアイテムの評価が与えられる。実験では、攻撃ユーザ数（攻撃サイズ）とフィラーアイテム数（フィラーサイズ）を異なる値で変化させ、攻撃の影響と協調フィルタリングのロバスト性を評価している。

4.3.2　予測シフト

攻撃に対するロバスト性は、攻撃前と攻撃後の予測値の差を計算することで評価される。図 3 は、ロバスト性評価（予測シフト）の流れを示している。元のデータに攻撃ユーザを生成して注入する。その後、協調フィルタリングが実行され、ユーザがアイテムの評価を予測する。攻撃後、各アイテムの予測評価は攻撃前の誤差測定と同じ方法で測定される。次に、攻撃前と攻撃後の予測値の差を計算する。攻撃前と攻撃後の予測値の差は、攻撃による予測のシフトを表すため、協調フィルタリングのロバスト性を評価するのに使用する。予測値の差が小さいほど、攻撃に対するロバスト性が高いことを示す。

4.3.3　TopN value

実際のアプリケーションでは、ユーザは推薦結果に注目している。この研究では、ユーザの推薦リストを測定するために TopN 値を導入し、予測値に基づいてユーザにアイテムを推薦する。まず、攻撃前の TopN 値を測定し、攻撃されたアイテムがユーザー推薦リストに入った回数を計算する。次に、攻撃後の TopN 値を測定し、攻撃されたアイテムがユーザー推薦リストに入った回数を計算する。攻撃前と攻撃後の TopN 値は、協調フィルタリングのロバスト性を評価する際に、攻撃前と攻撃後の推薦結果の変化を調査するためにも使用される。すべてのユーザは M でマークされる。各ターゲットアイテムを $j(j = 1, \ldots, N)$ とし、ターゲットアイテム j を評価したユーザの数を m_j、ターゲットアイテム j に未評価のユーザの数を l_j とする。すべてのターゲットアイテムがユーザの推薦リストに現れる回数は、以下のように計算できる。

$$TopN = \sum_{j=1}^{N} \frac{l_j}{M - m_j} \tag{6}$$

4.3.4　TopN shift

攻撃に対するロバスト性は、攻撃前と攻撃後の TopN 値の差を計算することで評価される。図 3 は、ロバスト性評価（TopN シフト）の流れを示している。攻撃前と攻撃後の TopN 値の差を測定し、その差分は攻撃による TopN の変化をしている。さらに、その差分は協調フィルタリングのロバスト性を評価するために使用される。TopN 値の差が小さいほど、攻撃に対するロバスト性が高いことを示す。

5　実　　験

実験では、MovieLens Latest Datasets small[1] を使用している。データセットには、610 人のユーザが 9,742 の映画アイテムに対して与えられた 100,836 件の評価が含まれており、評価は 0.5 から 5 の範囲である。クラスタリングには、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの 3 つのメソッドが設定されている。1 回クラスタリングでは、クラスタの数 k が設定される。2 回クラスタリングでは、第 1 回

1：https://grouplens.org/datasets/movielens/latest

表 1　クラスタ数

1 回 クラスタリング	2 回クラスタリング			
	第 1 回			第 2 回
k	2k	2.5k	3k	k
2	4	5	6	2
6	12	15	18	6
10	20	25	30	10
15	30	37	45	15

表 2　誤 差 測 定

	0 回	1 回	2 回
攻撃なし	$X_0 = P_0 - R$	$X_1 = P_1 - R$	$X_2 = P_2 - R$
攻撃あり	$Y_0 = P'_0 - R$	$Y_1 = P'_1 - R$	$Y_2 = P'_2 - R$

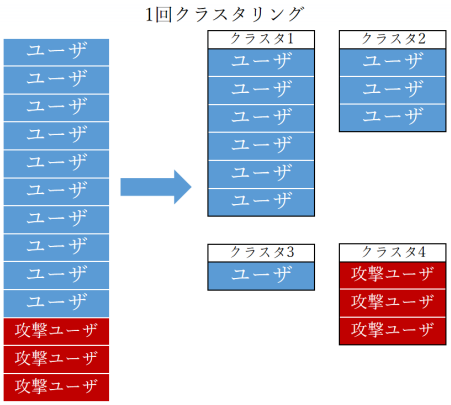


図 4　1 回クラスタリング傾向

クラスタリングには 2k、2.5k、および 3k のクラスタ数が設定され、第 2 回のクラスタリングには k が設定される。クラスタ数の設定は表 1 に示されている。実験結果を公平に比較するために、第 2 回クラスタリングのクラスタ数は第 1 回クラスタリングと同じである。実験ではクラスタ数の設定による 2k → k 値、2.5k → k 値、および 3k → k 値の結果を計算し、それぞれの結果が最小値、中央値、最大値の順に並べて統計する。

攻撃に対するロバスト性を測定する際には、全てのアイテムから 50 個のアイテムがランダムに選択され、これらを攻撃のターゲットアイテムとし、最高の評価を与える。攻撃ユーザは、ターゲットアイテム以外をランダムに選択し、各選択されたアイテムの評価をユーザが与えられた実際の評価に設定することで生成される。

5.1　クラスタリング傾向

5.1.1　1 回クラスタリング傾向

実験では、1 回クラスタリングを行い、クラスタ内データの傾向を収集して分析する。1 回クラスタリング傾向が図 4 に示されている。1 回のクラスタリングした、ユーザと攻撃ユーザは異なるクラスタに分割されて、クラスタ 1、2、3 は攻撃の影響を受けられない。1 回クラスタリングを行い、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 60%を占めている。

5.1.2　2 回クラスタリング傾向 1

2 回クラスタリング傾向 1 は、主にクラスタ数が 6、10、15

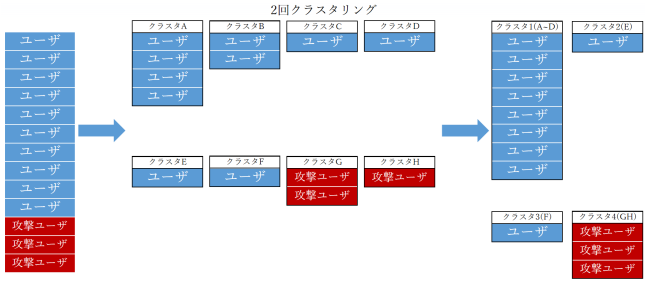


図 5　2 回クラスタリング傾向 1

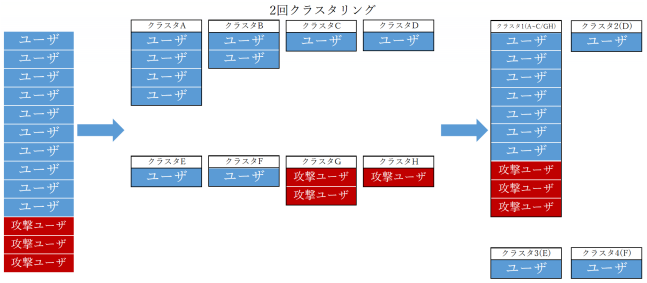


図 6　2 回クラスタリング傾向 2

の場合に発生する。2 回クラスタリング傾向 1 が図 5 に示されている。2 回のクラスタリングした、ユーザと攻撃ユーザは異なるクラスタに分割されて、クラスタ 1、2、3 は攻撃の影響を受けられない。傾向 1 は、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 90%以上を占めている。

5.1.3　2 回クラスタリング傾向 2

2 回クラスタリング傾向 2 は、主にクラスタ数が 2 の場合に発生する。2 回クラスタリング傾向 2 が図 6 に示されている。ユーザと攻撃ユーザは同じクラスタに分割されて、クラスタ 1 は攻撃の影響を大きく受けられる。傾向 2 も同様に、常に 1 つのクラスタ内のユーザ数量が全ユーザ数量の 90%以上を占めている。

5.2　誤 差 測 定

攻撃前と攻撃後の誤差をクラスタリング回数（0 回、1 回、2 回）ごとに比較するため、実験の誤差測定は表 2 に示す 6 つのタイプに分類する。

5.2.1　攻撃前の誤差測定

異なる回数のクラスタリングに対して、攻撃前の予測誤差が測定され、それが表 2 の X_0 から X_2 に対応している。この実験では、クラスタリングなし (0 回)、1 回クラスタリング (1 回)、および 2 回クラスタリング (2 回) の各手法に対して協調フィルタリングの予測が行われる。予測された評価と実際の評価の誤差が各手法ごとに測定される。その後、各手法の誤差値が比較され、各行で最小のものを太字で示す。クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。攻撃前、MAE に基づく結果が表 3 に示されている。結果から、クラスタリングなしに基づく誤差が最小であり、次に 2 回クラスタリングが続き、1 回クラスタリングが最も大きいことが示されている。さらに、2 回クラスタリングの最大誤差が 1 回クラスタリングの誤差よりも一

表 3　誤差測定 MAE (攻撃なし)

0 回	**1 回**		**2 回**		
MAE	クラスタ数	MAE	MAE		
			最小値	中央値	最大値
0.6068	2	0.6188	0.6176	0.6182	0.6182
0.6068	6	0.6514	0.6162	0.6201	0.6250
0.6068	10	0.6599	0.6224	0.6251	0.6353
0.6068	15	0.6689	0.6267	0.6353	0.6357

表 4　誤差測定 RMSE (攻撃なし)

0 回	**1 回**		**2 回**		
RMSE	クラスタ数	RMSE	RMSE		
			最小値	中央値	最大値
0.7992	2	0.8167	0.8137	0.8144	0.8144
0.7992	6	0.8563	0.8109	0.8145	0.8221
0.7992	10	0.8670	0.8174	0.8202	0.8352
0.7992	15	0.8761	0.8224	0.8350	0.8352

貫して小さいことがわかる。これは、2 回クラスタリングが 1 回クラスタリングよりも予測精度の向上に効果的であることが示されている。また、クラスタの数が増加するにつれて予測精度が低下する傾向がある。攻撃前、RMSE に基づく結果が表 4 に示されている。RMSE 結果が示す傾向は MAE 結果と類似している。

5. 2. 2　攻撃後の誤差測定

攻撃ユーザが攻撃を注入したことにより測定された、予測された評価と実際の評価との誤差を、表 2 の Y_0 から Y_2 にに示す。攻撃サイズとフィラーサイズは元のデータの 5 %と 10 %に設定されている。ランダム攻撃後、MAE 結果と RMSE 結果は表 5、6 に示されている。平均攻撃後、MAE 結果と RMSE 結果は表 7、8 に示されている。結果から、クラスタリングなしに基づく誤差が最小であり、次に 2 回クラスタリングが続き、1 回クラスタリングが最も大きいことが示されている。また、2 回クラスタリングの最大誤差が 1 回クラスタリングの誤差よりも一貫して小さいことがわかる。これは、2 回クラスタリングが 1 回クラスタリングよりも精度の向上に効果的であることが示されている。さらに、クラスタの数が増加すると誤差が増加する傾向がある。原因はクラスタ内ユーザの評価数量と関連している可能性がある。

5. 3　ロバスト性

5. 3. 1　予測シフト

攻撃前と攻撃後の予測値の差を測定し、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの各手法について比較する。これは、表 9 の Δ_0 から Δ_2 に対応している。各手法の予測シフトの値を比較し、各行で各行で最小のものを太字で示す。クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。ランダム攻撃前後の予測シフトの結果は表 10 に示されている。平均攻撃前後の予測シフトの結果は表 11 に示されている。ほとんどの結果から 2 回クラスタリングに基づく予測シフトが最小

表 5　誤差測定 MAE (ランダム攻撃)

攻撃サイズ	**フィラーサイズ**	**0 回**	**1 回**		**2 回**		
		MAE	クラスタ数	MAE	MAE		
					最小値	中央値	最大値
30	487	**0.6033**	2	0.6141	0.6034	0.6100	0.6109
30	487	**0.6033**	6	0.6345	0.6110	0.6202	0.6207
30	487	**0.6033**	10	0.6657	0.6217	0.6219	0.6234
30	487	**0.6033**	15	0.6603	0.6255	0.6343	0.6434
30	974	**0.6030**	2	0.6128	0.6091	0.6096	0.6136
30	974	**0.6030**	6	0.6451	0.6123	0.6159	0.6176
30	974	**0.6030**	10	0.6439	0.6156	0.6180	0.6184
30	974	**0.6030**	15	0.6594	0.6223	0.6238	0.6299
60	487	**0.5980**	2	0.6080	0.5977	0.6027	0.6083
60	487	**0.5980**	6	0.6415	0.6088	0.6110	0.6171
60	487	**0.5980**	10	0.6512	0.6181	0.6220	0.6243
60	487	**0.5980**	15	0.6613	0.6267	0.6336	0.6376
60	974	**0.5949**	2	0.6156	0.5951	0.5951	0.5953
60	974	**0.5949**	6	0.6322	0.6026	0.6048	0.6061
60	974	**0.5949**	10	0.6338	0.6173	0.6185	0.6220
60	974	**0.5949**	15	0.6506	0.6165	0.6192	0.6264

表 6　誤差測定 RMSE (ランダム攻撃)

攻撃サイズ	**フィラーサイズ**	**0 回**	**1 回**		**2 回**		
		RMSE	クラスタ数	RMSE	RMSE		
					最小値	中央値	最大値
30	487	**0.7951**	2	0.8107	0.7952	0.8027	0.8048
30	487	**0.7951**	6	0.8349	0.8042	0.8166	0.8174
30	487	**0.7951**	10	0.8738	0.8175	0.8180	0.8191
30	487	**0.7951**	15	0.8673	0.8210	0.8322	0.8443
30	974	**0.7944**	2	0.8088	0.8018	0.8034	0.8084
30	974	**0.7944**	6	0.8476	0.8051	0.8098	0.8114
30	974	**0.7944**	10	0.8461	0.8106	0.8120	0.8121
30	974	**0.7944**	15	0.8671	0.8181	0.8208	0.8283
60	487	**0.7875**	2	0.8024	0.7871	0.7923	0.8017
60	487	**0.7875**	6	0.8435	0.7995	0.8053	0.8098
60	487	**0.7875**	10	0.8560	0.8128	0.8170	0.8238
60	487	**0.7875**	15	0.8667	0.8223	0.8319	0.8375
60	974	**0.7834**	2	0.8123	0.7836	0.7836	0.7839
60	974	**0.7834**	6	0.8336	0.7941	0.7942	0.7956
60	974	**0.7834**	10	0.8324	0.8109	0.8124	0.8169
60	974	**0.7834**	15	0.8552	0.8105	0.8151	0.8235

であり、次にクラスタリングなしが続き、1 回クラスタリングが続くことが示されている。さらに、2 回クラスタリングに対する最大の予測シフト値は、他の手法よりも大きい。この結果は、クラスタリング傾向と関連している。

5. 3. 2　TopN シフト

攻撃前と攻撃後の TopN 値の差を測定し、クラスタリングなし、1 回クラスタリング、および 2 回クラスタリングの各手法について比較する。これは、表 9 の Δ_3 から Δ_5 に対応している。この実験では、TopN の N 値は 50 に設定されており、これはターゲットアイテムの数量と同じである。各手法の TopN シフトの値を比較し、各行で最小のものを太字で示す。クラスタリ

表 7　誤差測定 MAE (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		MAE	クラスタ数	MAE	MAE		
					最小値	中央値	最大値
30	487	**0.5870**	2	0.6086	0.5874	0.5926	0.5926
30	487	**0.5870**	6	0.6341	0.6020	0.6119	0.6146
30	487	**0.5870**	10	0.6549	0.6068	0.6095	0.6153
30	487	**0.5870**	15	0.6668	0.6185	0.6320	0.6344
30	974	**0.5818**	2	0.6060	0.5874	0.5961	0.5969
30	974	**0.5818**	6	0.6438	0.6027	0.6053	0.6186
30	974	**0.5818**	10	0.6497	0.5994	0.6073	0.6243
30	974	**0.5818**	15	0.6618	0.6187	0.6219	0.6224
60	487	**0.5788**	2	0.6022	0.5844	0.5913	0.5922
60	487	**0.5788**	6	0.6410	0.5922	0.5952	0.5970
60	487	**0.5788**	10	0.6624	0.6002	0.6029	0.6074
60	487	**0.5788**	15	0.6565	0.6113	0.6349	0.6359
60	974	**0.5757**	2	0.6042	0.5758	0.5814	0.5845
60	974	**0.5757**	6	0.6378	0.5919	0.5942	0.6154
60	974	**0.5757**	10	0.6457	0.5971	0.6144	0.6160
60	974	**0.5757**	15	0.6582	0.6217	0.6233	0.6281

表 8　誤差測定 RMSE (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		RMSE	クラスタ数	RMSE	RMSE		
					最小値	中央値	最大値
30	487	**0.7740**	2	0.8026	0.7745	0.7796	0.7796
30	487	**0.7740**	6	0.8344	0.7906	0.8057	0.8113
30	487	**0.7740**	10	0.8621	0.7975	0.8000	0.8079
30	487	**0.7740**	15	0.8751	0.8119	0.8302	0.8323
30	974	**0.7665**	2	0.7988	0.7719	0.7838	0.7852
30	974	**0.7665**	6	0.8473	0.7952	0.7959	0.8132
30	974	**0.7665**	10	0.8572	0.7857	0.7997	0.8213
30	974	**0.7665**	15	0.8685	0.8128	0.8168	0.8186
60	487	**0.7628**	2	0.7937	0.7683	0.7789	0.7794
60	487	**0.7628**	6	0.8433	0.7769	0.7837	0.7846
60	487	**0.7628**	10	0.8709	0.7872	0.7927	0.7994
60	487	**0.7628**	15	0.8632	0.8019	0.8333	0.8348
60	974	**0.7585**	2	0.7977	0.7587	0.7640	0.7697
60	974	**0.7585**	6	0.8401	0.7790	0.7818	0.8094
60	974	**0.7585**	10	0.8484	0.7837	0.8078	0.8095
60	974	**0.7585**	15	0.8655	0.8165	0.8182	0.8260

表 9　ロバスト性

	0 回	1 回	2 回
攻撃なし	P_0 , T_0	P_1 , T_1	P_2 , T_2
攻撃あり	P'_0 , T'_0	P'_1 , T'_1	P'_2 , T'_2
予測シフト	$\Delta_0 = P'_0 - P_0$	$\Delta_1 = P'_1 - P_0$	$\Delta_2 = P'_2 - P_0$
TopN シフト	$\Delta_3 = T'_0 - T_0$	$\Delta_4 = T'_1 - T_0$	$\Delta_5 = T'_2 - T_0$

ングなし、1 回クラスタリング、および 2 回クラスタリングの比較における最大値を下線で示す。ランダム攻撃前後の TopN シフトの結果は表 12 に示されている。平均攻撃前後の TopN シフトの結果は表 13 に示されている。結果から、1 回クラスタリングに基づく TopN シフトが最小であり、次に 2 回クラスタリング、クラスタリングなしとなっている（クラスタ数が 2 の場合)。ほとんどの場合、2 回クラスタリングに基づく TopN シフトが最小であり、次に 1 回クラスタリング、クラスタリングなしとなっている（クラスタ数が 6、10、15 の場合)。さらに、2 回クラスタリングに対する最大の TopN シフト値は、1 回クラスタリングの TopN シフトよりも大きい。この結果は、クラスタリング傾向と関連している。

表 10　予測シフト (ランダム攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		予測シフト	クラスタ数	予測シフト	予測シフト		
					最小値	中央値	最大値
30	487	0.0013	2	**0.0006**	0.0008	0.0013	0.0018
30	487	0.0013	6	0.0036	**0.0006**	0.0053	0.0073
30	487	0.0013	10	0.0150	**0.0005**	0.0020	0.0036
30	487	0.0013	15	0.0073	**0.0001**	0.0094	0.0105
30	974	**0.0004**	2	0.0010	0.0036	0.0040	0.0040
30	974	**0.0004**	6	0.0132	0.0008	0.0029	0.0075
30	974	0.0004	10	0.0006	**0.0002**	0.0006	0.0029
30	974	**0.0004**	15	0.0072	0.0020	0.0035	0.0070
60	487	0.0035	2	0.0028	**0.0006**	0.0014	0.0040
60	487	0.0035	6	0.0069	**0.0009**	0.0016	0.0041
60	487	0.0035	10	0.0208	**0.0009**	0.0013	0.0045
60	487	**0.0035**	15	0.0192	0.0059	0.0071	0.0133
60	974	0.0028	2	**0.0019**	0.0022	0.0029	0.0033
60	974	0.0028	6	0.0041	**0.0000**	0.0008	0.0072
60	974	0.0028	10	0.0026	**0.0011**	0.0028	0.0049
60	974	0.0028	15	0.0023	**0.0007**	0.0042	0.0050

表 11　予測シフト (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		予測シフト	クラスタ数	予測シフト	予測シフト		
					最小値	中央値	最大値
30	487	0.0021	2	0.0046	**0.0018**	0.0049	0.0049
30	487	0.0021	6	0.0053	**0.0020**	0.0036	0.0043
30	487	0.0021	10	0.0093	**0.0005**	0.0022	0.0054
30	487	**0.0021**	15	0.0080	0.0022	0.0024	0.0059
30	974	**0.0050**	2	0.0058	0.0064	0.0074	0.0078
30	974	0.0050	6	0.0135	**0.0003**	0.0066	0.0080
30	974	0.0050	10	0.0026	**0.0003**	0.0024	0.0062
30	974	0.0050	15	0.0021	**0.0000**	0.0051	0.0063
60	487	**0.0027**	2	0.0051	0.0052	0.0065	0.0083
60	487	0.0027	6	0.0054	**0.0009**	0.0009	0.0076
60	487	0.0027	10	0.0070	**0.0011**	0.0023	0.0030
60	487	0.0027	15	0.0025	**0.0010**	0.0066	0.0094
60	974	0.0147	2	**0.0037**	0.0145	0.0166	0.0171
60	974	0.0147	6	0.0089	**0.0032**	0.0128	0.0171
60	974	0.0147	10	0.0158	**0.0034**	0.0144	0.0205
60	974	0.0147	15	0.0066	**0.0014**	0.0032	0.0033

表 12　TopN シフト (ランダム攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		TopN	クラス	TopN	TopN シフト		
		シフト	タ数	シフト	最小値	中央値	最大値
30	487	2.4747	2	**0.1836**	0.2051	0.9621	2.4846
30	487	2.4747	6	0.0665	**0.0391**	0.2543	0.3647
30	487	2.4747	10	**0.0132**	0.0232	0.2461	0.2493
30	487	2.4747	15	0.0251	**0.0095**	0.0717	0.1882
30	974	5.7368	2	**0.1044**	2.6409	3.9595	4.4177
30	974	5.7368	6	0.0750	**0.0066**	0.9741	2.7886
30	974	5.7368	10	0.0902	**0.0067**	0.0408	0.4320
30	974	5.7368	15	0.0688	**0.0258**	0.1290	0.1934
60	487	5.3052	2	**0.1225**	1.0826	4.7498	4.9241
60	487	5.3052	6	0.0633	**0.0232**	2.3195	2.3881
60	487	5.3052	10	0.0634	**0.0112**	0.1186	0.1671
60	487	5.3052	15	**0.0015**	0.0670	0.1357	0.1484
60	974	17.430	2	**1.2529**	16.911	16.970	17.041
60	974	17.430	6	**0.0090**	13.717	13.764	14.726
60	974	17.430	10	0.0582	**0.0129**	0.0248	0.0326
60	974	17.430	15	0.0172	**0.0060**	0.0131	3.7134

表 13　TopN シフト (平均攻撃)

攻撃サイズ	フィラーサイズ	0 回	1 回		2 回		
		TopN	クラス	TopN	TopN シフト		
		シフト	タ数	シフト	最小値	中央値	最大値
30	487	0.4975	2	**0.0188**	0.5057	0.5583	0.5583
30	487	0.4975	6	0.1125	**0.0763**	0.1509	0.6687
30	487	0.4975	10	**0.0779**	0.1728	0.3602	0.4405
30	487	0.4975	15	0.0895	**0.0111**	0.0417	0.0758
30	974	0.7537	2	**0.1044**	0.6749	0.7145	0.7718
30	974	0.7537	6	0.1588	**0.0210**	0.2133	0.5767
30	974	0.7537	10	**0.0372**	0.0883	0.3531	1.0087
30	974	0.7537	15	0.0238	**0.0047**	0.0753	0.1026
60	487	0.0540	2	0.0632	**0.0526**	0.0672	0.0837
60	487	0.0540	6	**0.0476**	0.0671	0.1198	0.1936
60	487	0.0540	10	**0.0049**	0.1131	0.1504	0.1919
60	487	0.0540	15	**0.0284**	0.0666	0.0719	0.1752
60	974	0.5041	2	**0.0032**	0.5008	0.5846	0.6650
60	974	0.5041	6	0.0452	**0.0049**	0.5567	0.6492
60	974	0.5041	10	0.0896	**0.0230**	0.1421	0.5999
60	974	0.5041	15	0.0311	**0.0013**	0.0215	0.0448

6　結　　論

本研究では 2 回クラスタリングを行い、クラスタ内でアイテムを推薦することで、頑健な協調フィルタリング手法を提案した。攻撃前後の予測評価の誤差を測定することで、予測評価の手法を紹介した。さらに、攻撃前後の予測値と TopN 値の差を測定することで、ロバスト性評価の手法を導入した。

実験結果によれば、2 回クラスタリングを行う推薦方法は、攻撃の影響を抑制することで一定の効果を表している。2 回クラスタリングの誤差は 1 回クラスタリングよりも小さく、予測精度が向上していることを示している。ロバスト性では、2 回クラスタリングのロバスト性が他のクラスタリング手法よりも優れていることがある。ただし、実験結果では、2 回クラスタリングの安定性は 1 回クラスタリングほど良くないことも示されている。平均攻撃と比較して、協調フィルタリングはランダム攻撃に大きく影響を受けられる。さらに、シリング攻撃が予測シフトに小さい影響を与えるが、実際の推薦結果に大きな影響を与える可能性があることを示す。

今後の課題として、クラスタリング手法については DBSCAN 法を使用し、データセットについては MovieLens 1B Synthetic Dataset を追加し、評価に用いた指標の吟味が挙げられる。

謝　　辞

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

文　　献

[1] J. B. Schafer, D. Frankowski, J. L. Herlocker, and S. Sen. Collaborative filtering recommender systems. *Adaptive Web 2007*, pages 291–324, 2007.

[2] Rui Chen, Qingyi Hua, Yanshuo Chang, Bo Wang, Lei Zhang, and Xiangjie Kong. A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access*, 6:64301–64320, 2018.

[3] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Promoting recommendations: an attack on collaborative filtering. *DEXA 2002*, pages 494–503, 2002.

[4] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. Shilling attacks against recommender systems: a comprehensive survey. *Artif. Intell. Rev.*, 42(4), 2014.

[5] M. Si and Q. Li. Shilling attacks against collaborative recommender systems: a review. *Artif. Intell. Rev.*, pages 1–29, 2018.

[6] Agnideven Palanisamy Sundar, Feng Li, Xukai Zou, Tianchong Gao, and Evan D. Russomanno. Understanding shilling attacks and their detection traits: A comprehensive survey. *IEEE Access*, 8:171703–171715, 2020.

[7] Reda A. Zayed, Lamiaa Fattouh Ibrahim, Hesham A. Hefny, Hesham Abou El Fetouh Salman, and Abdulaziz Almohimeed. Experimental and theoretical study for the popular shilling attacks detection methods in collaborative recommender system. *IEEE Access*, 11:79358–79369, 2023.

[8] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. A regression framework to interpret the robustness of recommender systems against shilling attacks. *IIR*, 2021.

[9] Siyu Wang, Yuanjiang Cao, Xiaocong Chen, Lina Yao, Xianzhi Wang, and Quan Z. Sheng. Adversarial robustness of deep reinforcement learning based dynamic recommender systems. *CoRR*, abs/2112-00973, 2021.

[10] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Collaborative filtering-safe and sound? *ISMIS 2003*, pages 506–510, 2003.

[11] Carlos Pedroso and Aldri Santos. Dissemination control in dynamic data clustering for dense iiot against false data injection attack. *Int. J. Netw. Manag*, 32(5), 2022.

[12] Hongtao Yu, Lijun Sun, and Fuzhi Zhang. A robust bayesian probabilistic matrix factorization model for collaborative filtering recommender systems based on user anomaly rating behavior detection. *KSII Trans. Internet Inf. Syst*, 13(9):4684–4705, 2019.

[13] F. Zhang and S. Xu. Analysis of trust-based e-commerce recommender systems under recommendation attacks. *ISDPE 2007*, pages 385–390, 2007.
[14] Y. Du, M. Fang, C. Xu, J. Cheng, and D. Tao. Enhancing the robustness of neural collaborative filtering systems under malicious attacks. *IEEE Trans. Multimedia*, 21(3):555–565, 2019.
[15] B. Mobasher, R. D. Burke, and J. J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. *AAAI 2006*, pages 1388–1393, 2006.
[16] Lam SK and Riedl JT. Shilling recommender systems for fun and profit. *WWW 2004*, pages 393–402, 2004.
[17] Zhang S, Ouyang Y, Ford J, and Makedon F. Analysis of a low-dimensional linear model under recommendation attacks. *SIGIR 2006*, pages 517–524, 2006.
[18] Renato Cordeiro De Amorim. A survey on feature weighting based k-means algorithms. *CoRR*, abs/1601-03483, 2016.

Light distillation for Incremental Graph Convolution Collaborative Filtering

Xin Fan † Fan Mo ‡ Chongxian Chen ‡ Hayato Yamana §

† ‡ Graduate School of Fundamental Science and Engineering 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
§ Faculty of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

E-mail: † fan_xin@fuji.waseda.jp, ‡ bakubonn@toki.waseda.jp, ‡ chenc@toki.waseda.jp

§ yamana@yama.info.waseda.ac.jp

Abstract Recommender systems presently utilize vast amounts of data and play a pivotal role in enhancing user experiences. Graph Convolution Networks (GCNs) have surfaced as highly efficient models within the realm of recommender systems due to their ability to capture extensive relational information. The continuously expanding volume of data may render the training of GCNs excessively costly. To tackle this problem, incrementally training GCNs as new data blocks come in has become a vital research direction. Knowledge distillation techniques have been explored as a general paradigm to train GCNs incrementally and alleviate the catastrophic forgetting problem that typically occurs in incremental settings. However, we argue that current methods based on knowledge distillation introduce additional parameters and have a high model complexity, which results in unrealistic training time consumption in an incremental setting and thus difficult to actually deploy in the real world. In this work, we propose a light preference-driven distillation method to distill the preference score of a user for an item directly from historical interactions, which reduces the training time consumption in the incremental setting significantly without noticeable loss in performance. The experimental result on two general datasets shows that the proposed method can save training time from 1.5x to 9.5x compared to the existing methods and improves Recall@20 by 5.41% and 10.64 from the fine-tune method.

Keyword GCNs, Collaborative Filtering, Incremental Learning, Knowledge Distillation

1. INTRODUCTION

In this age of information explosion, recommendation systems [1] [2] are playing an increasingly irreplaceable role in several real-world domains, alleviating information overload and improving the user experience. Collaborative filtering [3] is the most used approach in the field of recommender systems that attempts to capture user preferences based on explicit or implicit feedback from users on items in historical data and has been very successful in the real world [4]. In recent years, deep learning has been widely used in several research areas including natural language processing and computer vision due to its immensely powerful representation learning capabilities. Deep learning is also widely used in collaborative filtering to build rich representations of users and items to enhance the recommendation performance [5]. On top of deep collaborative filtering, graph convolutional networks (GCNs) stimulate the potential of collaborative filtering even further through their superiority of capturing higher-order interactive relationships between users and items [6][7].

Collaborative filtering based on graph neural networks has been extensively researched and much progress has been made on recommendation performance on several generic datasets [8][9]. However, in the face of huge historical interaction data, training GCNs needs to bear exponentially increasing time cost due to its high computational complexity, introduced by neighbor sampling and message passing. The unacceptably low training efficiency prevents GCNs from being widely used in practical scenarios because it is not practical to update the model frequently to provide the up-to-date recommendations. The inability to update frequently can result in the model losing its ability to model short-term user preferences. On the other hand, updating the model using only the most recent data can lead to another problem called catastrophic forgetting [10] [11], with the model losing its ability to model users' long-term preferences.

Incremental learning has been explored in several deep learning domains to train models incrementally using the latest data while mitigating the problem of catastrophic forgetting. The two main incremental learning methods are *experience replay* and *knowledge distillation*. Methods based on experience replay [14][15] address the issue of forgetting by selecting representative data from past tasks and then training the model on new tasks using this data in

combination with new data. The *knowledge distillation*-based methods [12][13] aim to preserve the model's efficacy on prior tasks by limiting alterations to its parameters. This is accomplished by incorporating task-specific regularization elements into the loss function. In this work, we focus on the methods based on *knowledge distillation*.

The key to *knowledge distillation*-based methods is how to design matching distillation terms for a specific task or a specific domain. In the field of GCNs-based collaborative filtering, several incremental learning methods based on knowledge distillation have been explored and achieved superior recommendation performance [16]. However, to achieve the desired recommended performance, these methods usually introduce additional trainable parameters or design model structures with high computational complexity leading to excessive training time consumption in the incremental setting.

In this work, we argue that current methods based on knowledge distillation over-consider distilling complex structural information as well as multiple layers of user or item representations in GCNs. The contributions of this study can be summarized as follows:

- We proposed a simple but effective light distillation method that focuses on user preferences for items in historical interactions, discarding structural information as well as representational information about users and items, thereby significantly reducing training time consumption. Specifically, we directly distilled the user's preference information by minimizing the user's preference scores for the same historical items in the previous model as well as in the current model.
- We demonstrated the effectiveness of our proposed novel distillation mothed by conducting the experiment on two commonly used datasets. Experimental results showed that our proposed distillation method can reduce training time consumption significantly and achieve competitive recommendation performance in an incremental environment.

The remainder of this paper is organized as follows: Section II introduces related studies on incremental learning. In Section III, we review the basis of GCN-based collaborative filtering and knowledge distillation. Section IV presents the proposed method. Section V presents the experimental evaluation. Finally, Section VI concludes the paper.

2. RELATED WORK

1. *Incremental Learning for Deep Learning*

Incremental learning is dedicated to exploring methods to periodically update the model as new data arrives. Simply updating the model with the most recent data will lead to catastrophic forgetting [10] [11], that the model loses previously acquired knowledge and shows an excessive inclination towards recent data. The central goal of incremental learning is to mitigate catastrophic forgetting. The two main lines of research on incremental learning are *experience replay* [14] [15] and *knowledge distillation* [12][13].

The *experience replay*-based methods involve sampling the most significant historical data and replaying this data during the learning of new tasks to mitigate the issue of forgetting. As a representative *experience replay*-based method. Prabhu *et al*. [12] proposed GDumb that greedily stores samples in memory as they come and trains a model from scratch using samples only in the memory at test time. The *knowledge distillation*-based methods commonly add penalty terms to the loss function to restrict the model weights from deviating excessively from their adjusted values based on historical data, thereby avoiding forgetting. Hinton *et al.* [17] first proposed transferring knowledge from a large and complex teacher model to a smaller student model without significant loss in performance. In the incremental setting, knowledge distillation is used to migrate knowledge from the previous model to the current model, thereby alleviating forgetfulness.

2. *Incremental Learning for GCNs*

Several incremental methods have been proposed to tailor the GCNs-based collaborative filtering to preserve information from different aspects. As a latest *experience replay*-based method, Fan *et al.* [6] proposed a connectivity-based sampling mechanism to enhance the graph connectivity constructed by the new date. For the *knowledge distillation*-based methods. Xu *et al.* [13] first employed knowledge distillation to preserve information from three aspects; node level, the local, graph level and the global graph level. Wang *et al.* [18] introduced a contrastive distillation mechanism for each layer of GCNs on user-item, user-user, and item-item graphs.

It can be observed that incremental learning methods based on knowledge distillation become more and more complex and cumbersome when they are tailored to collaborative filtering based GCNs. Additional trainable parameters and complex model structures improve recommendation performance to some extent, but also lead to higher training

time costs. In this work, we propose a preference-driven light distillation method to align the current model with the previous model at the level of the final goal, i.e., the user's preference score for the item.

3. PRELIMINARIES

Before introducing the proposed methodology, we briefly review the GCN-based collaborative filtering and knowledge distillation.

1. GCN-based Collaborative Filtering

GCNs aim to capture the high-order collaborative signal from historical interaction data by convolution operation on the user-item bipartite graph constructed by the historical user-item interactions. For the process of forward propagation, given the initial representations $\boldsymbol{e_u}$ and $\boldsymbol{e_i}$ for user u and item i, the GCNs first aggregate the information from neighbors to form the final representations $\boldsymbol{emb_u}$ and $\boldsymbol{emb_i}$ for user u and item i respectively by aggregating the information from neighbors recursively. The $\boldsymbol{emb_u}$ and $\boldsymbol{emb_i}$ then are used to calculate the preference score $Score_{u,i}$ of user u for item i. Finally, the preference scores $Score_u$ of user u for the unknown items is used to form the recommendation list. In the backpropagation step, the parameters of GCNs are optimized by the Bayesian personalized ranking (BPR) loss [19], which is a pairwise loss so that a user's preference score for seen items is higher than that user's preference score for unseen items.

2. Knowledge Distillation

Originally, knowledge distillation aimed to transfer insights from a large, complex (teacher) model to a smaller (student) model by aligning the output logits, thereby lowering computational expenses [17]. Later studies concentrated on enhancing this transfer process and investigating its uses across different fields [12] [13]. Knowledge distillation has also been used as a mechanism to mitigate catastrophic forgetting in the field of incremental learning [10] [11]. To preserve historical learned knowledge in the previous model, the knowledge distillation mechanism minimizes the extent to which the parameters of the current model are shifted from the previous model in two successive incremental intervals by introducing additional distillation losses when training the model incrementally on new data.

TABLE I. Definition of symbols

U, u	U is the set of users in a given dataset, $u \in U$
I, i	I is the set of items in a given dataset, $i \in I$
(u,i)	User–item interaction between user u and item i
S_t	Training interactions in the t-th incremental interval, $S_t = \{(u,i,i') \mid (u,i) \in S_t^+, (u,i') \in S_t^-\}$
S_t^+	Oberved interactions in the t-th incremental interval, $S_t^+ = \{(u,i) \mid u \in U, i \in I\}$
S_t^-	Unoberved interactions in the t-th incremental interval, $S_t^- = \{(u,i') \mid u \in U, i' \in I\}$
MOD_t	Model trained on S_t

4. METHODOLOGY

In this section, we elaborate on our proposed methodology, preference-driven distillation loss for efficiently training GCNs-based collaborative filtering models in an incremental setting. The preference-driven distillation loss mitigates catastrophic forgetting by comparing the difference between a user's preference scores for items from previous interactions in the previous model and in the current model to minimize this difference to preserve the user's historical preferences. User preference scores for items reflect the degree of user preference for items, and in GCNs, users have high preference scores for items they have seen and low preference scores for items they have not seen.

Our proposed method targets to align the user's scores for previously seen items in the previous model and the current model when the training model is in the incremental environment. Since no additional learnable parameters are introduced and the computational complexity of the model is significantly increased, it is possible to significantly reduce the training time consumption. TABLE I shows the symbols used in the study.

A. Incremental Training

To train the GCNs incrementally for collaborative filtering as new data arrive, we first define a set of users and items as U and I, respectively. Second, the interactions between user u and item i, denoted by (u,i), in a dataset are sorted chronologically in an incremental setting. Besides, the interactions of user u_i for an unknown item i' is defined as (u,i'). Third, the interactions are divided into a base interval and a set of incremental intervals to simulate an incremental setting in the real world.

The observed interactions and the unobserved interactions in the t-th incremental interval is defined as $S_t^+ = \{(u,i) \mid u \in U, i \in I\}$ and $S_t^- = \{(u,i') \mid u \in U, i' \in I\}$, where $t \geq 0$. Initially, the base model is trained on S_0. The incremental model is trained on S_t and initialized by the model trained on S_{t-1}, where $t \geq 1$. Fig. 1 shows the *knowledge distillation*-based incremental training procedure. The incremental model shown as MOD_t distills

the historical information from the MOD_{t-1} by introducing the distillation loss term into the optimization objective.

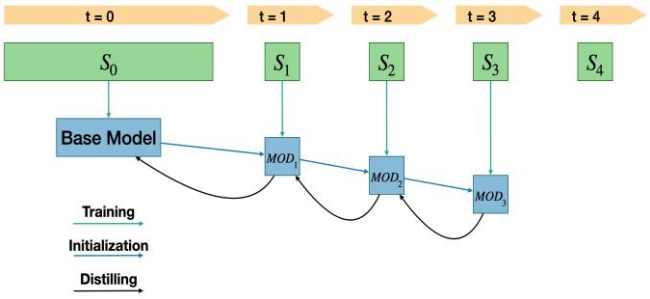


Fig. 1. The procedure of incremental learning

B. *Preference-driven Distillation*

GCNs implement recommendations by learning enriched user and item representation, $\boldsymbol{emb_u}$ and $\boldsymbol{emb_i}$ to accurately predict users' preference scores for items. The proposed preference-driven distillation targets to distil preference information from the previous model to the current model when training GCNs in the incremental setting.

1. *BPR Loss*

Given the final representation $\boldsymbol{emb_u}$ and $\boldsymbol{emb_i}$ for user u and item i derived by forward propagation. To optimize the parameters in a GCNs-based model, the most common choice is the pairwise Bayesian Personalized Ranking (BPR) loss [19], which enforces the prediction of an observed interaction to be scored higher than its unobserved counterparts:

$$L_{BPR} = \frac{1}{|S_t|} \sum_{(u,i,i') \in S_t} -log\, \sigma\left(Score^t_{u,i} - Score^t_{u,i'}\right) \qquad (1)$$

, where $S_t = \{(u,i,i') \mid (u,i) \in S_t^+, (u,i') \in S_t^-\}$ is the training data at the t-th incremental interval. The $Score^t_{u,i}$ is the dot production between $\boldsymbol{emb_u}$ and $\boldsymbol{emb_i}$ representing the preference score at the t-th incremental interval. The σ is the sigmoid function. Minimizing the BPR loss makes the user's preference score for items that have been interacted with greater than the user's preference score for items that have not been interacted with.

2. *Distillation Loss*

To preserve the user's historical preference information simply but efficiently in an incremental setting, we designed the distillation loss based on the user preference. Given the incremental model MOD_{t-1} at the (t-1)-th incremental interval. We can obtain the $Score^{t-1}_{u,i}$ for each $(u,i) \in O^+_{t-1}$ that represents the user preference for the interacted item at the (t-1)-th incremental interval. The distillation loss is defined by minimizing the difference between $Score^{t-1}_{u,i}$ and $Score^t_{u,i}$ for observed interactions at the (t-1)-th incremental interval:

$$L_{KD} = \frac{1}{|S^+_{(t-1)}|} \sum_{(u,i) \in S^+_{(t-1)}} \left(Score^{t-1}_{u,i} - Score^t_{u,i}\right)^2 \qquad (2)$$

, where the $|O^+_{t-1}|$ is the number of observed interactions in the (t-1)-th incremental interval. Minimizing the distillation loss allows the current model to inherit the user's preference information for interacted items from the previous model thus preserving the user's historical preference information in the incremental environment.

C. Model Training

When training the model in an incremental setting, the BPR loss and distillation loss work together to optimize the model parameters in the process of backpropagation. The BPR loss is used to learn new user preferences on the current incremental interactions, and the distillation loss is used to inherit the historical user preferences from the previous model. The final model loss function is defined as follows:

$$L_{loss} = L_{BPR} + \lambda_1 L_{KD} + \lambda_2 \left|\Theta\right|_2 \qquad (3)$$

, where Θ is the set of model parameters in L_{BPR} since L_{KD} introduces no additional parameters. The λ_1 and λ_2 are hyperparameters to control the strengths of distillation and L2 regularization, respectively.

Algorithm I. Preference-driven Distillation

Input:
Training interactions $S_t = \{(u,i,i') \mid (u,i) \in S_t^+, (u,i') \in S_t^-\}$
Oberved interactions $S_t^+ = \{(u,i) \mid u \in U, i \in I\}$
Unoberved interactions $S_t^- = \{(u,i') \mid u \in U, i' \in I\}$
Embedding of u and i $\boldsymbol{emb_u}$, $\boldsymbol{emb_i}$
Output: Final training loss function L_{loss}
Initialize:
Preference Score for i: $Score^t_{u,i} = \boldsymbol{emb_u} \cdot \boldsymbol{emb_i}$
Preference Score for i': $Score^t_{u,i'} = \boldsymbol{emb_u} \cdot \boldsymbol{emb_{i'}}$
/* Definition of BPR Loss */

$$L_{BPR} = \frac{1}{|S_t|} \sum_{(u,i,i') \in S_t} -log\, \sigma\left(Score^t_{u,i} - Score^t_{u,i'}\right)$$

/* Definition of Distillation Loss */

$$L_{KD} = \frac{1}{|S^+_{(t-1)}|} \sum_{(u,i) \in S^+_{(t-1)}} \left(Score^{t-1}_{u,i} - Score^t_{u,i}\right)^2$$

/* Final Training Loss */

$$L_{loss} = L_{BPR} + \lambda_1 L_{KD} + \lambda_2 \left|\Theta\right|_2$$

The backpropagation of the parameters in the model is based on the final training loss which combines the BPR Loss and Distillation Loss

5. EXPERIMENTAL EVALUATION

A. *Datasets*

To evaluate the effectiveness of the proposed method in saving the training time consumption and recommendation performance, we conducted experiments on two common publicly available real-world datasets: *Gowalla* [20] and

TABLE II. Statistical information of datasets

Dataset		#Users	#Items	New Users* (%)	New Items* (%)	#Interactions	Density **(%)	Timespan (month)
Gowalla	Base interval	44,169	18,154			777,256	0.0969	19
	Interval I			19.7	1.6			
	Interval II			18.4	1.0			
	Interval III			16.1	0.7			
	Interval IV			10.4	0.3			
Yelp	Base interval	7,413	13,286			217,262	0.2206	24
	Interval I			7.1	7.4			
	Interval III			5.9	6.5			
	Interval IV			9.8	5.5			
	Interval IV			2.8	3.5			

* Percentage of cold-start users/items in each incremental interval.

** Density of each dataset is defined as (the number of interactions)/(the number of users × the number of items).

Yelp [21]. These datasets differ in several aspects such as the data scale, sparsity, time range, number of interactions, and numbers of new users and items. TABLE II presents the statistical details of the datasets.

- **Gowalla** [20]**:** Gowalla is a location-based social networking dataset collected from user check-in records, including 6,442,890 check-ins of users from February 2009 to October 2010. We filtered out duplicate interactions between users and items with fewer than 10 interaction records.
- **Yelp** [21]**:** Yelp is an all-purpose dataset provided by the Yelp business and includes 6,990,280 user check-in records from February 2005 to January 2022. We utilized data from the period 2017-2018 and applied the same filtering method as that used for the Gowalla dataset.

B. Experimental Setting

Following the temporal order, we divided each dataset into a base interval, containing 60% of the entire dataset, and four incremental intervals, each containing 10% of the entire dataset. The process of incremental training is described in Figure I. The base model is trained on S_0 at t=0 and the incremental model is trained on S_t sequentially. The model trained on S_{t+1} as training data is validated and tested on S_{t+1} as validation and test data.

Base Model: The proposed incremental learning method is built on a base GCNs-based model. In this work, we choose LightGCN [22] as our base model, which is widely used in the field of collaborative filtering-based recommender systems. LightGCN removes the nonlinear activation functions and learnable parameters for linear transformation of traditional GCNs, which significantly simplifies the training burden of GCNs and achieves the desired recommended performance.

Baselines: To compare the effectiveness of our proposed method in reducing training time and its recommended performance. We compare our method with the following knowledge distillation-based baselines in terms of both training time consumption and accuracy:

- *Fine Tuning*: The model trained in S_t initializes the trainable parameters by the parameters of the model has been trained in S_{t-1} with no additional distillation loss. The most naive method which is often used as a benchmark to compare the effectiveness of other incremental learning methods.
- *Graph SAIL* [13]: Information at the node, local and global levels in the GCN is distilled through three customized distillation losses, respectively. Among them, clustering of user and item representations is required for distilling information at the global level.
- *LWC* [18]: The introduction of a comparative distillation mechanism combines BPR losses with distillation losses. The item-item and user-user graphs are introduced on top of the traditional user-item bi-directional graphs and the information is distilled layer by layer.
- *Connectivity Sampling* [6]: The connectivity of graph structure is considered in the incremental setting. The incremental learning method based on experience replay by sampling representative historical

TABLE IV. OVERALL EXPERIMENTAL RESULTS ON THE TOP 20 RECOMMENDED ITEMS

DataSet	Methods	Interval I	Interval II	Interval III	AVG R@20	AVG Improvement from Fine Tune (%)	AVG Runtime (s)
Gowalla	Fine Tune	0.1065	0.1167	0.1206	0.1146	0	**2,428**
	Graph SAIL [13]	0.1089	0.1181	0.1244	0.1171	2.18%	14,345
	LWC [18]	0.1156	0.1218	0.1281	0.1218	6.28%	40,185
	Connectivity Sampling [6]	0.1181	0.1245	0.1309	0.1245	**8.63%**	10,471
	Proposed	0.1149	0.1191	0.1283	0.1208	5.41%	4,503
Yelp	Fine tune	0.0559	0.0589	0.0576	0.0575	0	**952**
	Graph SAIL	0.0603	0.0664	0.0612	0.0626	8.87%	7,997
	LWC [6]	0.0627	0.0661	0.0632	0.0640	11.31%	10,853
	Connectivity Sampling [6]	0.0642	0.0675	0.0635	0.0651	**13.21%**	3,046
	Proposed	0.0611	0.0672	0.0624	0.0636	10.61%	2,047

interactions to enhance the connectivity of the graph structure of the current model.

Hyperparameters: The tuning process followed a greedy strategy. The learning rate was tuned with *fine tuning* in the range of {0.005, 0.001, 0.0005, 0.0001} and then fixed to 0.0005 for the model trained on S_0 and 0.0001 for the model trained on S_t for other baselines. The weights of the different loss terms for Graph Sail [13] were tuned by a grid search in the range of {0.1,1,100} for λ_{self} and {1e4,1e5,1e6,1e7} for both λ_{local} and λ_{global}. The hyperparameter τ for *LWC* [18] was searched in the range of {0.1,0.2,0.5,1}. The hyperparameters λ_1 for our proposed method were tuned in the range of {0.01,0.1,1,10,100}. TABLE III presents the values of the final hyperparameters.

C. Comparison

The proposed distillation method was compared with the baselines in terms of both recommendation performance and training time. Recall on the top 20 recommendation items was used as the metric to evaluate the recommendation performance. For the evaluation of training time, we calculate the time in seconds it takes for the model to reach the best Recall. TABLES IV shows the experimental results for the dataset Gowalla and Yelp, respectively. From the experimental results, we can conclude the following:

Recommendation Performance: The proposed method can comprehensively outperform *Graph SAIL* [13] in each incremental interval evaluated by Recall@20 on the Gowalla dataset. Compared to fine tune method, our proposed method achieved an average improvement of 5.41% over three incremental intervals, which fully demonstrated the effectiveness of the proposed method despite its simplicity. Although there was still a gap compared to *LWC* [18] and *Connectivity Sampling* [6], this is acceptable in the face of the expensive cost of training time at *LWC* which employs a complex distillation mechanism to distill information from each layer of multiple graph structures.

Training Time: The training time mainly depends on the number of parameters of the model, the time complexity of the model, and the number of epochs required for the model to reach convergence. The experimental evaluation for training time was performed on an NVIDIA A100 Tensor Core GPU by calculating the product of the number of epochs required to achieve the best recall@20 and the time required for each epoch. The results on the two datasets show that our proposed method can significantly reduce the time consumed for model training from 1.5x to 9.5x compared to these baselines.

D. Ablation Study

To explore the effect of the degree of distillation on the recommended performance, we conducted experiments on the hyperparameter λ_1 in the range of {0.01,0.1,1,10,100} on both *Gowalla* and *Yelp* datasets. TABLE. V shows the experimental results for the average result of Recall@20 for three incremental intervals. Because the level of distillation does not significantly affect training time, we do not discuss training efficiency here. We can observe that the effect of hyperparameter λ_1 on recommendation performance can reach a maximum of 7% on the Gowalla dataset and 30% on Yelp dataset, which indicates that fine-tuning the hyperparameter λ_1 is of practical importance.

TABLE III. Hyperparameters

	τ	λ_{self}	λ_{local}	λ_{global}	λ_1	λ_2
Gowalla	1	100	1e5	1e7	1	1e-3
Yelp	0.5	0.1	1e3	1e2	0.01	1e-3

TABLE III. Study for Hyperparameters λ_1 on the Recall@20

	0.01	***0.1***	***1***	***10***	***100***
Gowalla	0.1145	0.1128	0.1149	0.1145	0.1074
Yelp	0.0603	0.0563	0.0553	0.0494	0.0462

6. CONCLUSION

In this work, we propose a novel incremental learning method based on knowledge distillation that can efficiently train collaborative filtering models based on GCNs and achieve competitive recommendation performance. To mitigate the catastrophic forgetting problem in incremental environments, we propose methods to distill the user's historical preference information from the previous model to the current model. Experimental results on two publicly available datasets show that the proposed method can reduce the training time from 1.5x to 9.5x compared to the existing methods. With such time consumption, the proposed method can achieve competitive recommendation performance, improving Recall@20 by 5.41% and 10.61 respectively on two datasets.

References

[1] P. Covington, J. Adams, and E. Sargin. “Deep neural networks for youtube recommendations.” Proc. ACM Conf. Recommender System, pp. 191–198. 2016.

[2] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. "Recommendation systems: Principles, methods and evaluation." Egyptian Informatics Journal, Vol. 16(3), pp. 261-273, 2015.

[3] G. Linden, B. Smith, and J. York. "Amazon. Com recommendations: Item-to-item collaborative filtering." IEEE Internet Computing, Vol. 7(1), pp. 76-80, 2003.

[4] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. "Large-scale parallel collaborative filtering for the netflix prize." Algorithmic Aspects in Information and Management (AAIM 2008), LNCS, Vol. 5034, pp. 337- 348, 2008.

[5] R. Mu. "A survey of recommender systems based on deep learning." IEEE Access, Vol. 6, pp. 69009-69022, 2018.

[6] X. Fan, F. Mo, C.X. Chen, C.H. Bai, and H. Yamana. “Connectivity-aware Experience Replay for Graph Convolution Network-based Collaborative Filtering in Incremental Setting”, Proc. of Int. Conf. on Big Data Analytics, in press, 2024.

[7] C.X. Chen, F. Mo, X. Fan, C.H. Bai, and H. Yamana. “MOBARecGCNFP: Champion Recommendation for Multi-Player Online Battle Arena Games Using Graph Convolution Network with Fewer Parameters.” Proc. of Int. Conf. on Big Data Analytics, pp. 147-153, 2023.

[8] F. Mo and H. Yamana. "GN-GCN: Combining Geographical Neighbor Concept with Graph Convolution Network for POI Recommendation." Proc. of Int. Conf. on Information Integration and Web, pp. 153-165, 2022.

[9] F. Mo and H. Yamana. "EPT-GCN: Edge propagation-based time-aware graph convolution network for POI recommendation." Neurocomputing, Vol. 543, pp. 1-15, 2023.

[10] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. "Overcoming catastrophic forgetting in neural networks." J. of the National Academy of Sciences, Vol. 114(13), pp. 3521-3526, 2017.

[11] K. Shmelkov, C. Schmid, and K. Alahari. “Incremental Learning of Object Detectors Without Catastrophic Forgetting.” Proc. of Int. Conf. Computer Vision (ICCV), pp.3400– 3409, 2017.

[12] F. M. Castro, M. J. Mar´ın-Jimenez, N. Guil, C. Schmid and K. Alahari. “End-to-End Incremental Learning.” Proc. European Conf. Computer Vision (ECCV), pp. 241–257, 2018.

[13] Y. Xu, Y. Zhang, W. Guo, H. Guo, R. Tang, and M. “Coates. GraphSAIL: Graph Structure Aware Incremental Learning for Recommender Systems.” In Proc. ACM Int. Conf. Info. & Knowledge Management, pp. 2861–2868, 2020.

[14] A. Prabhu, P. H. Torr, and P. K. Dokania. “GDumb: A simple approach that questions our progress in continual learning.” Proc. European Conf. Computer Vision (ECCV), pp. 524– 540, 2020.

[15] K. Ahrabian, Y. Xu, Y. Zhang, J. Wu, Y. Wang, and M. Coates. “Structure Aware Experience Replay for Incremental Learning in Graph-based Recommender Systems.” Proc. ACM Int. Conf. Info. & Knowledge Management (CIKM), pp. 2832–2836, 2021.

[16] M. Farajtabar, N. Azizan, A. Mott, and A. Li. "Orthogonal gradient descent for continual learning." Proc of the Int. Conf. on Artificial Intelligence and Statistics(PMLR), Vol. 108, pp. 3762-3773, 2022.

[17] G. E. Hinton, O. Vinyals and J. Dean. “Distilling the Knowledge in a Neural Network.” arXiv:1503.02531, pp.1-9, 2015.

[18] Y. Wang, Y. Zhang, and M. Coates. “Graph Structure Aware Contrastive Knowledge Distillation for Incremental Learning in Recommender Systems.” Proc. ACM Int. Conf. Info. & Knowledge Management(CIKM), pp. 3518–3522, 2021.

[19] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. "BPR: Bayesian personalized ranking from implicit feedback." Proc. of the Twenty-Fifth Conf. on Uncertainty in Artificial Intelligence, pp. 452–461, 2009.

[20] E. Cho, S.A. Myers, and J. Leskovec. "Friendship and mobility: user movement in location-based social networks." Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1082-1090, 2011.

[21] Y. Cui. "An evaluation of yelp dataset." arXiv:1512.06915, pp. 1-6, 2015.

[22] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. "Lightgcn: Simplifying and powering graph convolution network for recommendation." Proc. of the 43rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 639-648, 2020.

グラフニューラルネットワークによる分離表現学習を用いた推薦の多様化

杉山　拓都†　　吉田　　壮††　　棟安　実治††

† 関西大学大学院 理工学研究科　〒 564–8680 大阪府吹田市山手町 3–3–35
†† 関西大学 システム理工学部　〒 564–8680 大阪府吹田市山手町 3–3–35
E-mail: †{k226567,sohy,muneyasu}@kansai-u.ac.jp

あらまし　Graph neural network (GNN) は，collaborative filtering における精度を向上させることで，推薦システムを大きく進歩させてきた．しかし，この進歩はしばしば，ユーザ満足度の要因である推薦の多様性を見落としている．このギャップに対処するため，本論文では分離表現 GNN (Disentangled representation GNN: DRGNN) を提案する．DRGNN は 2 つのモジュールを用いて，多様化を推薦候補となるアイテムの生成段階に統合する．1 つ目のモジュールは，アイテムの嗜好とカテゴリの嗜好を分離するために，分離表現学習を用いることで，推薦におけるカテゴリの偏りを緩和する．二つ目のモジュールは，ポジティプサンプルの選択に焦点を当て，カテゴリの偏りをさらに軽減する．このアプローチは，GNN の高次接続性を維持するだけでなく，推薦の多様性を大幅に改善する．DRGNN を代表的な二つの Web サービスのデータセット，Taobao と Amazon Beauty で広範に検証した結果，DRGNN は精度で最先端の手法に匹敵するだけでなく，推薦における精度と多様性のバランスの取れたトレードオフを達成することに優れていることが示された．

キーワード　推薦システム，多様性，グラフニューラルネットワーク，分離表現学習

1　はじめに

インターネットの急激な発展に伴い，情報過多の問題が顕在化している．この課題に対応するため，ユーザのニーズに合わせたアイテムを予測・提示する推薦システムが，Web サービスにおいて広く利用されている [1,2]．これらのシステムは，個々のユーザに対して関心のあるアイテムを推薦することで，ユーザ体験の向上やビジネス成長を促進させるため，パーソナライズされた推薦の重要性が高まっている．

パーソナライズされた推薦システムの中核をなすのは，collaborative filtering (CF) [3] であり，これは類似したアイテム評価を行うユーザ間に共通の好みが存在するという前提に基づいている．従来のアプローチ，例えば matrix factorization (MF) [4] は，ユーザとアイテム間の相互作用から潜在的な要因を学習し，推薦システムにおいて効果的なベンチマークを提供している．さらに，最近では neural network を活用した潜在因子の学習が提案されている [5,6]．特に，graph neural network (GNN) の導入は，この分野に顕著な進化をもたらしている [7,8]．GNN ベースの推薦システムは，ユーザとアイテムの相互作用から構築される二部グラフを利用して高次の接続性を捉え，精度の向上を実現している．

しかし，CF などの推薦システムは，多様性を犠牲にして精度に重点を置く傾向がある．一般的に，アイテムはカテゴリに分類され，多様性は推薦リスト内のカテゴリの種類によって測定されることが多い [9]．本論文では，カテゴリレベルの多様性の強化に焦点を当てる．推薦システムにおける多様性の向上は，フィルターバブルの解消や目新しいアイテムを推薦する機会の増加により，ユーザ満足度を高めることが知られている [10]．これは，優れた推薦システムの開発に不可欠な要素である．しかし，多様性は，精度とのトレードオフを伴うため，精度を損なうことなく多様性を高めることが課題である．

多様化された推薦システムは従来，スコアに基づいて並び替えを行うリランキングなどの後処理に依存してきた [11–13]．しかし，これらの手法は，推薦候補となるアイテムの生成段階での精度と多様性のトレードオフに直接対処していない．この問題に対応するため，推薦するアイテムの候補を生成する段階で GNN に基づいて多様性を向上させる手法が提案されている [14,15]．これらの手法は，観測済みの相互作用であるポジティブサンプルのカテゴリバイアスに対処するものの，特定のケースに限定されることがある．さらに，アイテム嗜好とカテゴリ嗜好の区別が別の重要な課題である．アイテム嗜好は，ユーザの特定のアイテムに対する好みを指し，その特徴や過去の相互作用に基づく．一方で，カテゴリ嗜好は，より幅広いアイテムが属するカテゴリに対するユーザの全体的な関心を関連付ける．推薦モデルがカテゴリ嗜好の影響を強く受ける場合，推薦が特定のカテゴリに偏り，ユーザの個別のアイテム嗜好や過去の行動と一致しないアイテムが推薦される可能性がある．

本稿では，推薦システムにおけるカテゴリレベルの多様性を強化するために，分離表現 GNN (Disentangled representation GNN: DRGNN) を提案する．DRGNN は，学習過程においてアイテム嗜好とカテゴリ嗜好を分離するための損失関数を用いた分離表現学習を実装する．また，DRGNN は，ポジティブサンプル選択における二重レベルのカテゴリバイアスを緩和する新しいサンプリング戦略を特徴とする．この戦略は，1) ユーザとアイテムとの相互作用におけるカテゴリの不均衡を是正し，推薦プロセスにおいて頻繁に相互作用しないカテゴリを考慮に入れること，2) 広範なデータセットレベルのカテゴリの不均衡

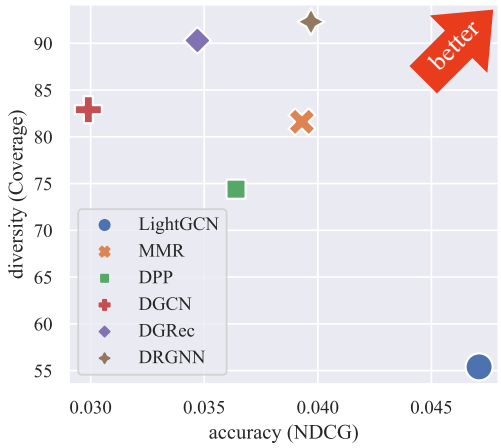


図 1: Performance comparison on Taobao. Our DRGNN addresses the trade-off between diversity and accuracy in recommendations.

に対処し，多様なカテゴリからアイテムが推薦されるようにする，という 2 点で構成される．これらの改善により，DRGNN は，推薦システムにおける精度と多様性を効果的に調整し，バランスの取れた多様な推薦を提供することができる．

本論文の主要な貢献は以下の 3 点である：

（1） 分離表現学習を取り入れた GNN ベースのモデルの導入．このモデルは，アイテム嗜好とカテゴリ嗜好を特徴表現で分離することにより，推薦性能を向上させ，個々のユーザの興味に合わせた推薦を可能にしながらカテゴリバイアスを最小限に抑える．

（2） ユーザとデータセット全体の 2 つのレベルで存在するカテゴリバイアスに対処するポジティブサンプル選択の開発．この戦略により，カテゴリの不均衡を是正し，推薦プロセスにおける多様性を強化する．

（3） 実際の Web サービスのデータセットを用いた DRGNN の包括的な検証．図 1 に示すようにこの検証により，DRGNN が推薦の精度と多様性のバランスを取ることにおいて，既存の最先端 (SOTA) 手法を上回る能力を持つことを示した．

2 関 連 研 究

2.1 Collaborative Filtering

CF [3] は，パーソナライズされた推薦システムの基礎を築いた．特に，行列分解 (MF) [4] を用いたモデルベースの CF は，大規模かつ疎なデータに対しても有効的な結果を得ている．MF では，ユーザとアイテムを表す潜在的な要因を学習し，ユーザのアイテムに対する評価を予測する．深層学習の進歩に伴い，Deep MF [5] や Neural CF [6] は，より複雑な潜在表現を得るためにニューラルネットワークを活用している．

2.2 Graph Neural Network

GNN は，グラフ構造を入力として扱うニューラルネットワークである [16–18]．Neural graph CF (NGCF) [7] は，GNN を用いて複雑なユーザ・アイテムの相互作用をモデル化する．これは，ユーザとアイテムのノードを持つ二部グラフを構築し，非線形変換を取り入れながら，グラフ畳み込み処理によって隣接ノードからの階層的集約を通じて埋め込みを更新する．Light graph convolutinal network (LightGCN) [8] は，NGCF で最も効果的なグラフ畳み込み処理のみを使用し，NGCF よりも高速かつ高精度な推薦を可能としている．近年では，埋め込み表現における整合性と均一性を考慮した，対照損失を統合することに注目が集まっている [19–22]．

2.3 Disentangled Representation Learning

相互作用から学習される埋め込み表現は，さまざまな要因から構成される．これらの要因を明確化し，分離することで，頑健性と解釈可能性をもたらすことができる [23]．Wang ら [24] は，潜在要因をチャンクに分割し，複数の意図として分離することで，有望な結果を得た．Zheng ら [25] は，ユーザの興味と適合性を分離し，人気度バイアスを抑制した．Zhang ら [26] は，人気度バイアスだけでなく，人気度分布の変化など，より一般的な実世界シナリオに対しても頑健な表現を学習する方法を提案した．

2.4 Diversified Recommendation

多様化推薦は，ユーザに様々なアイテムを提供し，興味の幅を広げ，目新しいアイテムを発見する機会を与えることができる．Maximum marginal relevance (MMR) のようなベンチマークとなる多様化推薦手法は，アイテムの類似性を考慮してリランキングを行うことで，多様な推薦を提供する [11]．Chen ら [12] は，推薦リストのリランキングのために Determinantal point process (DPP) を用いた．従来のパラメータのない DPP モデルとは対照的に，Chen らはパラメータを組み込むことで，多様な推薦において重要な機能である精度と多様性のトレードオフを柔軟に制御することを可能にした．一方で，これらの手法は，後処理に基づいており，生成段階でのアイテムに依存している．GNN ベースの手法は，隣接ノードのカテゴリのリバランスなど，生成段階での多様化を自然に統合することを可能にする．Diversified graph convolutional network (DGCN) [14] は，生成段階から多様性を考慮した最初の GNN ベースの手法である．DGCN は，隣接ノードのカテゴリの不均衡を解消，カテゴリ内のユーザ嗜好の学習，埋め込み空間においてカテゴリを遠ざけることによって多様性を向上させる．Diversified GNN-based RecSys (DGRec) [15] は，隣接ノードのカテゴリの不均衡に対処するためのサブモジュラー関数，過剰平滑化に対処するためのレイヤーアテンション．カテゴリ内のアイテム数に応じたサンプルへの再重み付けなどを用いて多様性を促進する SOTA である GNN ベースの多様化推薦法である．

3 提 案 手 法

DRGNN は GNN 技術を活用し，新たにカテゴリ嗜好のための分離表現学習モジュール (Disentangled Representation Learning Module for Category Preference，**DCP**) を導入する．DCP の主な目的は，GNN を通じて得られる嗜好表現か

らカテゴリ嗜好表現を分離し，アイテム嗜好とカテゴリ嗜好に分離して学習することである．DCP に加えて，DRGNN にはカテゴリを考慮したポジティブサンプル選択モジュール (Category-aware Positive Sampling Module，**CPS**) も導入されている．これらのモジュールの組み合わせは，DRGNN の他のいくつかの調整可能なパラメータとともに，DRGNN の効果を高める．

3.1 準　　備

本節では，提案手法の基盤となる LightGCN のアーキテクチャについて説明する．LightGCN は，埋め込み層，埋め込み伝播層，そして予測層の 3 つの層から構成されている．埋め込み層では，ユーザ u 及びアイテム i を d 次元の埋め込み $\mathbf{e}_u^{(0)}$ および $\mathbf{e}_i^{(0)}$ で表現し，次の式に従って埋め込み行列を生成する．

$$\mathbf{E}^{(0)} = [\mathbf{e}_{u_1}^{(0)}, \cdots, \mathbf{e}_{u_M}^{(0)}, \mathbf{e}_{i_1}^{(0)}, \cdots, \mathbf{e}_{i_N}^{(0)}], \tag{1}$$

ここで，M と N はそれぞれユーザとアイテムの総数を表す．埋め込み伝播層では，相互作用グラフを通じて埋め込み行列の伝播が行われる．対象となる相互作用グラフは，ユーザとアイテムをノード，ユーザとアイテムのインタラクションをエッジとする 2 部グラフである．このグラフの隣接行列 $\mathbf{A}$ は，次の式によって表現され，インタラクションが観測済みであれば 1，未観測であれば 0 を要素とする評価行列 $\mathbf{R}$ を利用する．

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{R} \\ \mathbf{R}^\mathsf{T} & \mathbf{0} \end{bmatrix}, \tag{2}$$

K 層のネットワークを考慮すると，k 層目における情報伝播は次の式に従って行われる．

$$\mathbf{E}^{(k+1)} = \left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{E}^{(k)}, \tag{3}$$

ここで，$\mathbf{D} = Diag(D_{11}, D_{22}, \cdots, D_{(M+N)(M+N)})$ であり，D_{ii} はノード i の次数である．予測層では，各層を連結した埋め込みから評価の予測が行われる．最終的な埋め込みは次の式で表される．

$$\mathbf{e}_u = \frac{1}{K+1}\sum_{k=0}^{K}\mathbf{e}_u^{(k)}, \mathbf{e}_i = \frac{1}{K+1}\sum_{k=0}^{K}\mathbf{e}_i^{(k)}. \tag{4}$$

式 (4) で得られるユーザ u およびアイテム i の埋め込み $\mathbf{e}_u$，$\mathbf{e}_i$ を用いて，次の式によりユーザに対するアイテムの評価予測が行われる．

$$\hat{y}_{ui} = \mathbf{e}_u^\mathsf{T}\mathbf{e}_i. \tag{5}$$

ネットワークの学習のため，BPR 損失 [27] を最小化する．これは，既に相互作用が観測されているユーザとアイテムのペアが，観測されていないペアよりも予測評価が高くなるという仮定に基づいて学習が行われる．

$$l(\mathbf{e}_u, \mathbf{e}_i, \mathbf{e}_j) = -\log\sigma(\hat{y}_{ui} - \hat{y}_{uj}), \tag{6}$$

ここで，$\sigma(\cdot)$ はシグモイド関数を表す．最近の研究では，BPR 損失の代わりに対照損失を用いることで良い結果を得ている [20,21]．本研究では，次に定義される InfoNCE 損失を採用する．

$$l(\mathbf{e}_u, \mathbf{e}_i) = -\log\frac{\exp\left(\cos\left(\mathbf{e}_u, \mathbf{e}_i\right)/\tau\right)}{\sum_{j\in I}\exp\left(\cos\left(\mathbf{e}_u, \mathbf{e}_j\right)/\tau\right)}, \tag{7}$$

ここで，I はミニバッチ内のアイテム集合を，τ は温度パラメータを表す．埋め込みは損失の勾配に基づいて最適化される．推論時には同様に情報伝播が行われ，式 (5) により評価予測が行い，値が最大となるアイテムから順に推薦される．

3.2 分離表現学習

本節では，アイテム嗜好とカテゴリ嗜好を区別するための表現学習法について説明する．我々が提案する DCP は，アイテム嗜好の表現を推定するエンコーダ，カテゴリ嗜好表現を推定するエンコーダ，そしてこれらの表現を分離する 2 つの損失関数から構成される．これらの損失関数は，カテゴリに依存しない表現を得ることを目的とし，カテゴリ嗜好をアイテム嗜好から分離するように設計されている．

アイテム嗜好を推定するエンコーダは，前節で述べた LightGCN エンコーダを使用する．ユーザとアイテムの最終的なアイテム嗜好表現 $\mathbf{e}_u, \mathbf{e}_i$ は次の式で表される．

$$\mathbf{e}_u, \mathbf{e}_i = \text{LightGCN}(u, i). \tag{8}$$

カテゴリ嗜好を推定するエンコーダも同様に LightGCN エンコーダを使用するが，アイテム嗜好との区別のために異なる埋め込み行列を用いる．具体的には，ユーザ u とアイテム i のカテゴリ嗜好は d 次元の埋め込み $\bar{\mathbf{e}}_u^{(0)}$，$\bar{\mathbf{e}}_{c(i)}^{(0)}$ として表され，次の式で埋め込み行列が構築される．

$$\mathbf{E}_c^{(0)} = [\bar{\mathbf{e}}_{u1}^{(0)}, \cdots, \bar{\mathbf{e}}_{u_M}^{(0)}, \underbrace{\bar{\mathbf{e}}_{c(i_1)}^{(0)}, \cdots, \bar{\mathbf{e}}_{c(i_N)}^{(0)}}_{\text{Share embeddings in the same category}}], \tag{9}$$

ここで，$c(i)$ はアイテム i のカテゴリを示し，括弧で囲まれた部分の埋め込みにおいて同じカテゴリのアイテムは同じ埋め込みを持つ．式 (9) の埋め込み行列を使用したモデル $\text{LightGCN}_\text{c}(\cdot)$ により，最終的なカテゴリ嗜好表現 $\bar{\mathbf{e}}_u, \bar{\mathbf{e}}_i$ が次の式により計算される．

$$\bar{\mathbf{e}}_u, \bar{\mathbf{e}}_i = \text{LightGCN}_\text{c}(u, i). \tag{10}$$

これらのエンコーダは相互作用を利用して学習されるが，互いの埋め込みに影響を与えないため，アイテム嗜好の表現にカテゴリ嗜好が含まれる可能性がある．したがって，2 つの損失関数を用いることでアイテム嗜好とカテゴリ嗜好の正確な分離を促進する．ユーザとアイテムはそれぞれ 2 つの埋め込み，すなわちアイテム嗜好とカテゴリ嗜好を持ち，これは次のように表される．

$$\mathbf{e}_u^* = \mathbf{e}_u \parallel \bar{\mathbf{e}}_u, \mathbf{e}_i^* = \mathbf{e}_i \parallel \bar{\mathbf{e}}_i, \tag{11}$$

ここで，$\parallel$ はベクトルの連結を示す．カテゴリ嗜好を必要としないアイテム嗜好の表現を得るため，拡張された表現を使用する．具体的には，ユーザとアイテムのカテゴリ嗜好の表現を別のユーザ-アイテムペアのカテゴリ嗜好の表現に置き換えると，次のようになる．

$$\mathbf{e}'_u = \mathbf{e}_u \parallel \bar{\mathbf{e}}_{u'}, \mathbf{e}'_i = \mathbf{e}_i \parallel \bar{\mathbf{e}}_{i'}. \tag{12}$$

埋め込みとその拡張された表現をポジティブサンプルとした拡張損失を最小限に抑えることにより，モデルはカテゴリ嗜好によらない，一貫した嗜好の表現を獲得するように学習される．

$$\mathcal{L}_{\text{aug}} = l(\mathbf{e}^*_u, \mathbf{e}'_i) + l(\mathbf{e}'_u, \mathbf{e}^*_i). \tag{13}$$

さらに，アイテム嗜好の表現とカテゴリ嗜好の表現が独立していることを確保するため，表現を分離するために広く使用される距離相関 [28] を最小化する [25,26].

$$\mathcal{L}_{\text{dis}} = dCor(\mathbf{e}_u, \bar{\mathbf{e}}_u) + dCor(\mathbf{e}_i, \bar{\mathbf{e}}_i), \tag{14}$$

$$dCor(\mathbf{e}, \bar{\mathbf{e}}) = \frac{dCov(\mathbf{e}, \bar{\mathbf{e}})}{\sqrt{dVar(\mathbf{e}) \cdot dVar(\bar{\mathbf{e}})}}, \tag{15}$$

ここで，$dCov(\cdot)$ は距離共分散，$dVar(\cdot)$ は距離分散を表す．

式 (16) における損失関数の最適化により，カテゴリ嗜好を含まないアイテム嗜好の表現を獲得する．推論中には，式 (8) のエンコーダから得られた埋め込みを使用して評価を予測する．

$$\mathcal{L} = \sum_{(u,i)\in\Omega} \left(l(\mathbf{e}_u, \mathbf{e}_i) + \lambda_1 l(\bar{\mathbf{e}}_u, \bar{\mathbf{e}}_i) + \lambda_2 \mathcal{L}\text{aug} + \lambda_3 \mathcal{L}\text{dis} \right), \tag{16}$$

ここで，Ω は学習サンプル集合を表す．λ_1 はカテゴリの好みを推定する影響を制御するパラメータ，λ_2 は拡張損失の影響を調整するパラメータ，λ_3 はアイテム嗜好とカテゴリ嗜好の表現間の独立性の強さを決定する正則化パラメータである．

3.3 カテゴリを考慮したポジティブサンプリング

ポジティブサンプルにおけるカテゴリの偏りは多様性の低下に繋がる．本手法では学習に使うポジティブサンプルをカテゴリを考慮して選択することで多様性の向上を図る．提案する手法を Algorithm 1 に示す．学習データにおける偏りは 2 つに分けられる．1 つめはユーザが接触したカテゴリの偏りである．ユーザが多く見ているカテゴリのアイテムに圧倒され，ユーザが少ししか見ていないカテゴリへの学習がおろそかになる．しかし，観測したカテゴリではあるため，興味のあるカテゴリである可能性が高い．本手法ではこのようなサンプルを用いる確率を高くする．具体的には，ユーザが接触したカテゴリのアイテム数 $|u_c|$ の逆数を用いることで学習への寄与度の向上を図る．α はこの偏りを調節するパラメータである．値を大きくするほど，少ない接触のカテゴリを重要視する．2 つめはデータセット全体におけるカテゴリの偏りである．アイテムを多く持つカテゴリはよく学習されるのに対して，少ししかアイテムを持たないカテゴリはより良い表現を学習することが困難である．この問題を対処するために，アイテムが属するカテゴリのアイテム数 $|c|$ の逆数を用いる．この 2 つの偏りを考慮した確率を各サンプルに割り当て，多様性を向上させる．β はこの偏りを調節するパラメータである．値を大きくするほど，アイテムが少ないカテゴリを重要視する．

Algorithm 1 Category-aware positive sampling

Input: users U; trade-off parameter α, β
Output: training samples Ω

for all $u \leftarrow U$ **do**
  $I \leftarrow$ GetPositiveItems(u) ▷ Get a list of positive items that interacted with the user
  for all $i \leftarrow I$ **do**
    $c \leftarrow$ GetCategory(i) ▷ Get the category of item i
    $r_i \leftarrow \frac{1}{|u_c|^\alpha |c|^\beta}$ ▷ Calculate dual-level category bias assigned to an item
  end for
  for all $i \leftarrow I$ **do**
    $p_i \leftarrow \frac{r_i}{\sum_{i=0}^{|I|} r_i}$ ▷ Convert ratio to probability
    $P \leftarrow P + [p_i]$ ▷ Append probability
  end for
  $i_{pos} \leftarrow$ SamplingItemWithProbability(I, P) ▷ Sample item based on probability assigned to each item
  $\Omega \leftarrow \Omega \cup \{(u, i_{pos})\}$ ▷ Add training sample
end for
return Ω

表 1: Statistics of the dataset.

Dataset	#users	#items	#categories	#interactions
Taobao	82,633	136,710	3,108	4,230,631
Beauty	8,159	5,862	41	98,566

4 実　　験

4.1 データセット

本実験では，カテゴリレベルの多様化推薦システムを評価するために，2 つの実世界のウェブサービスデータセットを用いた．本実験では，最新の文献 [15] で使用されている条件に基づいて比較している．1 つ目のデータセットは，EC サイト Taobao から収集された **Taobao** [29] で，クリックや購入などのユーザの相互作用を暗黙的フィードバックとして扱っている．このデータセットは 10-core 設定が採用され，各ユーザとアイテムは最低 10 回の相互作用を持っている．二つ目のデータセットは，Amazon が提供する **Beauty** [30] で，ユーザとアイテムの両方に少なくとも 5 回の相互作用がある 5-core 設定を採用している．各データセットの詳細な統計は，表 1 の通りである．データセットは訓練，検証，テストデータに分割され，検証データに基づいて早期終了を行っている．

4.2 実 験 条 件

実験に用いた条件は以下の通りである．畳み込み層，埋め込み次元，バッチサイズは既存研究 [14,15] に基づき，1 層，32 次元，2048 を採用した．オプティマイザには Adam [31] を使用する．モデルは PyTorch を用いて実装した．提案手法に用いたデータセットごとのハイパーパラメータは以下の通りである．Taobao では $\tau = 0.15, \lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.01, \alpha = 0.4, \beta = 0.05$ を，Beauty では $\tau = 0.75, \lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 0.01, \alpha =$

表 2: Performance comparison on Taobao.

	Taobao (Top-K=300)					Taobao (Top-K=100)				
	Recall ↑	NDCG ↑	Coverage ↑	Entropy ↑	Gini ↓	Recall ↑	NDCG ↑	Coverage ↑	Entropy ↑	Gini ↓
LightGCN [8]	**0.121**	**0.0471**	55.4	2.77	0.651	**0.0660**	**0.0325**	24.2	2.23	0.563
MMR [11]	0.105	0.0393	81.6	3.37	0.584	0.0544	0.0259	36.7	2.87	0.485
DPP [12]	0.0985	0.0364	74.4	3.34	0.595	0.0478	0.0229	34.2	2.87	0.493
DGCN [14]	0.0835	0.0299	82.9	3.55	0.560	0.0397	0.0180	37.4	3.05	**0.452**
DGRec [15]	0.0918	0.0347	90.3	3.47	0.570	0.0457	0.0222	39.4	2.95	0.469
DRGNN	0.102	0.0397	**92.3**	**3.64**	**0.558**	0.0551	0.0270	**40.0**	**3.07**	0.462

表 3: Performance comparison on Beauty.

	Beauty (Top-K=300)					Beauty (Top-K=100)				
	Recall ↑	NDCG ↑	Coverage ↑	Entropy ↑	Gini ↓	Recall ↑	NDCG ↑	Coverage ↑	Entropy ↑	Gini ↓
LightGCN [8]	**0.420**	**0.120**	24.6	2.29	0.663	**0.264**	**0.0942**	15.6	2.03	0.570
MMR [11]	0.383	0.102	26.6	2.46	0.643	0.215	0.0683	17.3	2.21	0.556
DPP [12]	0.325	0.0836	26.2	2.47	0.636	0.180	0.0577	17.4	2.19	0.572
DGCN [14]	0.356	0.0964	27.2	2.57	0.622	0.215	0.0733	18.9	2.38	0.540
DGRec [15]	0.385	0.101	27.5	2.42	0.663	0.219	0.0730	19.3	2.25	0.577
DRGNN	0.377	0.106	**29.1**	**2.63**	**0.609**	0.231	0.0817	**21.0**	**2.48**	**0.514**

$0.8, \beta = 0.4$ を用いた．

4.3 比較手法

本節では，比較手法について説明する．実験では，提案手法 (DRGNN) を以下に示すベースラインの手法と最先端 (SOTA) とされる多様化推薦手法と比較した．

i) **LightGCN**† [8]：SOTA の GNN ベースの推薦システム．比較のために，オリジナルの LightGCN とは異なり，InfoNCE 損失を損失関数として使用した．

ii) **MMR**‡ [11]：後処理ベースのベンチマークとなる多様化推薦手法．この手法では，冗長性を最小限に抑えるために，貪欲アルゴリズムに基づいてアイテムをリランキングする．

iii) **DPP**‡ [12]：後処理ベースのベンチマーク多様化推薦手法．予測スコアとアイテムの類似度行列から構築されたカーネルを利用して，多様なアイテムの集合を選択する．

iv) **DGCN**† [14]：GNN ベースのベンチマーク多様化推薦手法．

v) **DGRec**† [15]：SOTA の GNN ベースの多様化推薦手法．

† でマークされた手法は，著者によって提供された公式コードを使用している．また，‡ でマークされた手法は，我々独自のコードを使用して実装した．比較評価においては，[15] の実験設定に従い，Top-K を 100 および 300 に設定した．MMR や DPP などの後処理ベースの多様化手法では，LightGCN を使用してアイテム生成段階で Top-3K 候補アイテムを生成し，その後スコアに基づいてリランキングする．その後，Top-K のアイテムを推薦リストとした．DGCN，DGRec，DRGNN などの手法は，Top-K アイテムを含む推薦リストを直接出力する．

4.4 評価指標

我々は，2 つの精度評価指標と 3 つの多様性評価指標を用いた．精度評価には，既存研究で広く使用されている [7, 8]，**Recall**@K と **NDCG**@K を採用した．ここで，カテゴリレベルの多様性を評価するために，Top-K の推薦リストに n 個のカテゴリ $C = \{c_1, \cdots, c_n\}$ が含まれている場合，各カテゴリに対応するアイテムの頻度を $M = [m_1, \cdots, m_n]$ とする．多様性評価指標を以下に示す．

- **Coverage**@K は，Top-K の推薦アイテムにどれだけ多くの異なるカテゴリが表されているかを評価する．これは $Coverage@K = |C|$ として定義される．
- **Entropy**@K は，推薦リスト内のカテゴリ分布を評価する．Entropy は，推薦リストに現れるカテゴリのランダム性や予測不可能性を定量化し，$Entropy@K = -\sum_{i=1}^{n} p_{c_i} \log p_{c_i}$ として計算される．ここで $p_{c_i} = m_i / \sum_{j=1}^{n} m_j$ である．
- **Gini Index**@K は，推薦リスト内のカテゴリの不平等を測定する．Gini Index が低いほど，カテゴリ内での平等性が高いことを示す．これは $Gini@K = \sum_{i=1}^{n} \sum_{j=1}^{n} (|m_i - m_j|)/(2n^2\bar{M})$ として定義される．

4.5 実験結果

4.5.1 結　　果

提案手法である DRGNN と比較手法を評価した．結果は表 2 および 3 に示されており，最高の値は太字，次点は下線で示されている．SOTA 推薦手法である LightGCN は最高の精度を達成したが，多様性を考慮していないため多様性の低下が見られた．MMR や DPP のような後処理ベースの多様化推薦手法は多様性を高めるが，生成された候補のアイテムの多様性に依存するため，大幅な多様性向上には至らなかった．一方，GNN ベースの多様化推薦手法は，高い多様性と精度を示している．DGCN は高い Entropy と Gini Index を有し，カテゴリの不均衡を大幅に改善するが，精度は他の手法に比べて低下した．SOTA の多様化推薦方法である DGRec は，精度のわずかな低下で非常に高い Coverage を達成したが，Entropy と Gini Index は DGCN よりも劣っており，偏ったカテゴリ分布であることが確認できる．DRGNN は，多くの多様性評価指標で最

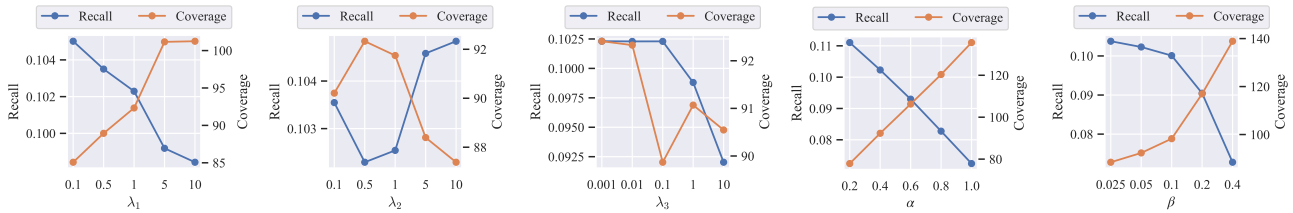


図 2: Parameter sensitivity of λ_i $(i = 1, 2, 3)$, α, and β for the Taobao dataset. λ_i controls the influence of the components in the loss function for the DCP. α and β are used to adjust the strength of the CPS.

表 4: Ablation study on Taobao.

	Recall↑	NDCG↑	Coverage↑	Entropy↑	Gini↓
LightGCN	0.121	0.0471	55.4	2.77	0.651
DRGNN	0.102	0.0397	92.3	3.64	0.558
w/o DCP	0.106	0.0413	80.6	3.46	0.577
w/o CPS	0.118	0.0461	60.7	2.89	0.644

表 5: Ablation study on Beauty.

	Recall↑	NDCG↑	Coverage↑	Entropy↑	Gini↓
LightGCN	0.420	0.120	24.6	2.29	0.663
DRGNN	0.377	0.106	29.1	2.63	0.609
w/o DCP	0.380	0.107	27.6	2.58	0.613
w/o CPS	0.412	0.119	25.0	2.30	0.664

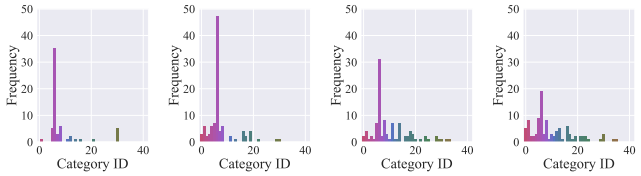
高の値を達成し，精度では LightGCN に次ぐものであった．したがって，DRGNN は多様性向上において優れた手法であるといえる．

4.5.2 パラメータ分析

さまざまな状況で精度と多様性のバランスを取ることは難しく，トレードオフを調整する能力は多様性を意識した推薦システムにおいて重要な側面である．図 2 は，Taobao データセットにおける精度と多様性の間のトレードオフに対するハイパーパラメータの影響を示している．パラメータ分析のために変更されたもの以外のすべてのパラメータは，4.2 節に準じている．カテゴリの好みを推定するエンコーダーの影響を決定するパラメータ λ_1 は，その値が増加するにつれて多様性が向上する．また，このパラメータは，DCP の中で，精度と多様性のバランスを取る上で最も影響力があることが確認できる．拡張損失の影響力を調節するパラメータ λ_2 は，表現がカテゴリ嗜好に影響を受けないことを保証する．しかし，λ_2 の値が過剰に高くなると多様性の低下につながる．これは，損失を最小限に抑えるためにすべての表現がアイテム嗜好表現に含まれてしまう可能性が考えられる．パラメータ λ_3 は，アイテム嗜好表現とカテゴリ嗜好表現との間の独立性を調整する．λ_3 の特定の値は良い精度と多様性を維持するが，過度に高い値は学習能力を低下させることが確認できる．CPS のパラメータ α と β は，ユーザが相互作用したカテゴリのバイアスと全体的なデータセットのバイアスの 2 つのレベルを制御する．α と β は，精度と多様性のトレードオフを調整するために重要であることが確認できる．

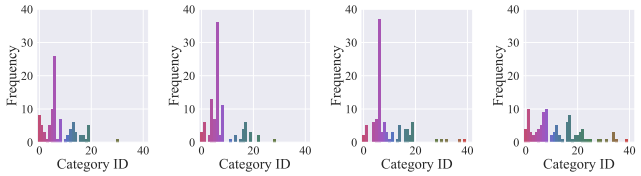
4.5.3 アブレーション研究

提案手法の DCP と CPS の有効性を検証するため，アブレーション研究を行った．LightGCN，DRGNN，DCP を除いた DRGNN (すなわち，式 (16) の第一項のみで訓練されたもの)，および CPS を除いた DRGNN を比較した．アブレーション研究の結果は，表 4 および 5 に示す．DCP は精度をわずかに低下させるが，多様性を大幅に向上させ，特に多くのカテゴリを持つ Taobao データセットでの有効性が顕著であることがわかった．CPS は多様性の向上に顕著に寄与し，相互作用の少ないカテゴリの表現を改善する．さらに，DRGNN は，DCP と CPS の両方を組み合わせることで，最も高い多様性を示した．



(a) Categ. dist. (b) LightGCN (c) DGRec (d) DRGNN

図 3: Comparison of the category distribution for the user with the lowest entropy on Beauty across different methods.



(a) Categ. dist. (b) LightGCN (c) DGRec (d) DRGNN

図 4: Comparison of the category distribution for the user with the highest entropy on Beauty across different methods.

4.5.4 定性的分析

本節では，推薦リストにおけるカテゴリ分布を調査することにより，DRGNN の特性を定性的に分析する．Beauty データセットを使用し，特定のユーザの学習データと，LightGCN，DGRec，DRGNN によって生成された推薦リストのカテゴリ分布と比較した．学習データの中で 50 回以上の相互作用を持つユーザに焦点を当て，特に Entropy が最も低いユーザおよび最も高いユーザについて考察した．この結果を図 3 および図 4 に示す．各図の (a) は，ユーザが相互作用したアイテムのカテゴリ分布を示している．(b) から (d) は，(a) のカテゴリ分布を持つユーザに対して，各手法が生成した推薦リストのカテゴリ分布を示している．

図 3 は最も低い Entropy のユーザに対する結果であり，偏っ

た相互作用を持っていることが図からもわかる．LightGCN は正確な好みを学習するため，学習データに依存して特定のカテゴリに強く偏った推薦を行うことが分かる．DGRec は，LightGCN によって提供されないカテゴリからアイテムを推薦するが，依然として特定のカテゴリが突出している．一方，提案された DRGNN は，特定のカテゴリに大きく影響されることなく，幅広いカテゴリのアイテムを網羅した推薦が可能であることを示す．

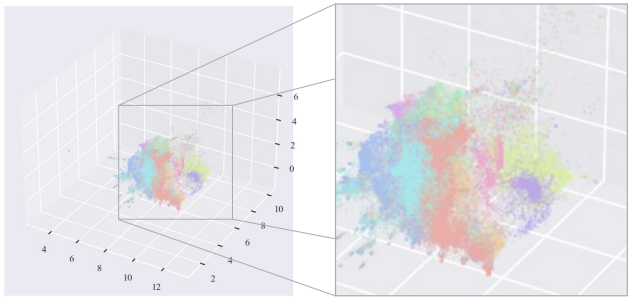
図 4 は，多様なカテゴリとの相互作用を示す高い Entropy のユーザである．このユーザに対しても同様の効果が確認でき，提案手法が学習データに含まれるカテゴリの偏りに関係なく，多様な相互作用パターンを持つさまざまなユーザに対して多様な推薦を提供することを確認できる．

4.5.5 提案手法の限界点

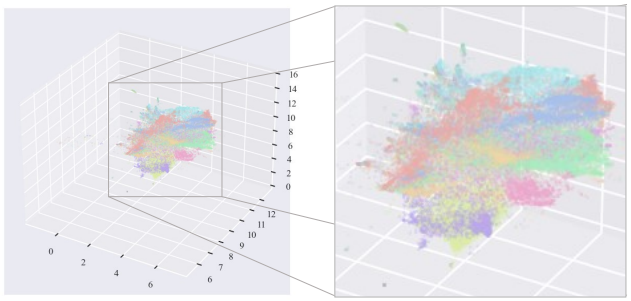
4.5.1 節で示されている通り，DRGNN は，精度をわずかに犠牲にして多様性を向上させる．さらに，DRGNN と CPS なしの DRGNN を比較すると，Taobao データセットでは LightGCN と比べて 9.6%から 14.5%，Beauty データセットでは 1.6%から 5.4%の多様性の増加が見られる (表 4 および 5 参照)．2 つのモジュールの組み合わせが多様性の増加に寄与していることを示しているが，これは，学習サンプルの偏りを緩和することがカテゴリの好みの推定に良好な影響を与えることを示唆している．また，DCP によって達成される 2 つのデータセット間の改善の大きさの相違は，各データセットのカテゴリの数の顕著な違いに起因する可能性がある．これは，カテゴリの数が少ないデータセットで DCP の適用が困難である可能性を示唆している．

4.5.6 埋め込み表現の可視化

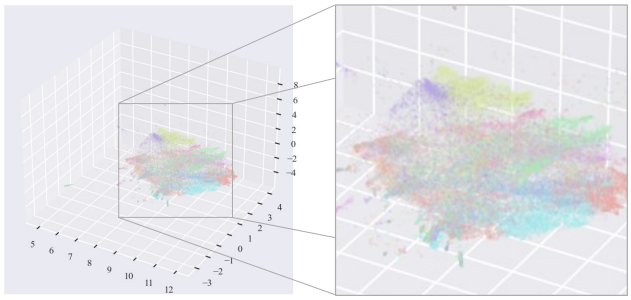
本節では，提案するモジュールの適用による埋め込み表現の変化を確認するため，可視化を行った．予測への影響を確認するため，予測に用いる埋め込み同士を比較する．具体的には，LightGCN はエンコーダの出力の d (= 32) 次元の埋め込み，DCP を用いる際はアイテム嗜好を推定するエンコーダの出力の d (= 32) 次元埋め込みである．まず，それぞれの手法ごとに全てのアイテムの埋め込みを Uniform Manifold Approximation and Projection (UMAP) [32] を用いて 3 次元に次元削減し，そのうちアイテム数が多い上位 10 件のカテゴリのアイテムを描画した．ここで，異なる色は異なるカテゴリであることを示している．LightGCN とパラメータを変化させた 2 種類の DCP の適用 ($\lambda_1 = 1$ と $\lambda_1 = 10$ とした)，そして提案手法の埋め込みの分布の結果を図 5 に示す．(a) より LightGCN によって得られる埋め込みは同じカテゴリのアイテム同士は大まかに固まって分布していることが確認できる．DCP の導入によってカテゴリに対する表現を分離することで，(b) より同カテゴリのアイテムは分布として広がる傾向が見られ，(c) よりパラメータを強くすることで分離が進み，カテゴリごとに固まるのが緩和されていることが確認できる．一方，水色や黄緑色のように固まった表現が残されている点は今後の課題である．また，(d) より提案手法は CPS との組み合わせによって最もカテゴリに影響されない，分離された表現を獲得していることが分かる．



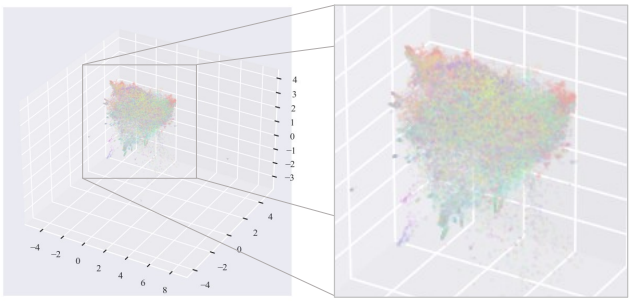
(a) LightGCN



(b) LightGCN+DCP($\lambda_1 = 1$)



(c) LightGCN+DCP($\lambda_1 = 10$)



(d) DRGNN

図 5: Visualization results of embedding representations on Taobao.

5 ま と め

本稿では，GNN ベースの推薦システムの多様性向上のための新しいアプローチである DRGNN を提案した．DRGNN は，DCP および CPS という 2 つのモジュールによって実現される．これらのモジュールは，精度を損なうことなく推薦の多様性を高めるために，GNN アーキテクチャに統合される．DRGNN は，多様性の面で最先端の多様化推薦手法を上回るだけでなく，精度においても精度のみを考慮したモデルに匹敵する．ま

た，精度と多様性の調整が可能なハイパーパラメータにより，DRGNN は多様なユーザのニーズに合わせてトレードオフを調整することができる．

今後の研究では，異なるデータセットにわたる α および β パラメータの決定のための体系的な方法の開発，カテゴリ嗜好表現の類似性を考慮した推論について調査する予定である．

謝辞 本研究の一部は JSPS 科学研究費補助金 (若手研究：22K18007)，および関西大学先端科学推進機構 (近未来 ICT の社会実装研究グループ) の助成を受けて行われた．ここに謝意を表す．

文　献

[1] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” IEEE Internet Comput., vol.7, no.1, pp.76–80, 2003.
[2] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” Proc. of the 10th ACM Conf. on Recomm. Syst., pp.191–198, 2016.
[3] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” Commun. of the ACM, vol.35, no.12, pp.61–70, 1992.
[4] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Comput., vol.42, no.8, pp.30–37, 2009.
[5] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, “Deep matrix factorization models for recommender systems,” Proc. of the 26th Int. Jt. Conf. on Artif. Intell., vol.17, pp.3203–3209, 2017.
[6] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” Proc. of the 26th Int. Conf. on World Wide Web, pp.173–182, 2017.
[7] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, “Neural graph collaborative filtering,” Proc. of the 42nd Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.165–174, 2019.
[8] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, “Lightgcn: Simplifying and powering graph convolution network for recommendation,” Proc. of the 43rd Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.639–648, 2020.
[9] H. Wu, Y. Zhang, C. Ma, F. Lyu, F. Diaz, and X. Liu, “A survey of diversification techniques in search and recommendation,” arXiv prepr. arXiv:2212.14464, 2022.
[10] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E.H. Chi, and J. Gillenwater, “Practical diversified recommendations on youtube with determinantal point processes,” Proceedings of the 27th ACM Int. Conf. on Inf. and Knowl. Manag., pp.2165–2173, 2018.
[11] J. Carbonell and J. Goldstein, “The use of mmr, diversity-based reranking for reordering documents and producing summaries,” Proc. of the 21st Annu. Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.335–336, 1998.
[12] L. Chen, G. Zhang, and E. Zhou, “Fast greedy map inference for determinantal point process to improve recommendation diversity,” Adv. in Neural Inf. Proc. Syst., vol.31, pp.5627–5638, 2018.
[13] A. Ashkan, B. Kveton, S. Berkovsky, and Z. Wen, “Optimal greedy diversity for recommendation.,” Proc. of the 14th Int. Jt. Conf. on Artif. Intell., vol.15, pp.1742–1748, 2015.
[14] Y. Zheng, C. Gao, L. Chen, D. Jin, and Y. Li, “Dgcn: Diversified recommendation with graph convolutional networks,” Proc. of the ACM Web Conf. 2021, pp.401–412, 2021.
[15] L. Yang, S. Wang, Y. Tao, J. Sun, X. Liu, P.S. Yu, and T. Wang, “Dgrec: Graph neural network for recommendation with diversified embedding generation,” Proc. of the 16th ACM Int. Conf. on Web Search and Data Min., pp.661–669, 2023.
[16] T.N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv prepr. arXiv:1609.02907, 2016.
[17] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” Adv. in Neural Inf. Proc. Syst., vol.30, pp.1024–1034, 2017.
[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S.Y. Philip, “A comprehensive survey on graph neural networks,” IEEE Trans. on Neural Netw. and Learn. Syst., vol.32, no.1, pp.4–24, 2020.
[19] C. Wang, Y. Yu, W. Ma, M. Zhang, C. Chen, Y. Liu, and S. Ma, “Towards representation alignment and uniformity in collaborative filtering,” Proc. of the 28th ACM SIGKDD Conf. on Knowl. Discov. and Data Min., pp.1816–1825, 2022.
[20] K. Mao, J. Zhu, J. Wang, Q. Dai, Z. Dong, X. Xiao, and X. He, “Simplex: A simple and strong baseline for collaborative filtering,” Proc. of the 30th ACM Int. Conf. on Inf. & Knowl. Manag., pp.1243–1252, 2021.
[21] H. Tang, G. Zhao, Y. Wu, and X. Qian, “Multisample-based contrastive loss for top-k recommendation,” IEEE Trans. on Multimed., 2021.
[22] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, and Q.V.H. Nguyen, “Are graph augmentations necessary? simple graph contrastive learning for recommendation,” Proc. of the 45th Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.1294–1303, 2022.
[23] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, “Learning disentangled representations for recommendation,” Adv. in Neural Inf. Proc. Syst., vol.32, pp.1–12, 2019.
[24] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, “Disentangled graph collaborative filtering,” Proc. of the 43rd Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.1001–1010, 2020.
[25] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin, “Disentangling user interest and conformity for recommendation with causal embedding,” Proc. of the Web Conf. 2021, pp.2980–2991, 2021.
[26] A. Zhang, J. Zheng, X. Wang, Y. Yuan, and T.-S. Chua, “Invariant collaborative filtering to popularity distribution shift,” Proc. of the ACM Web Conf. 2023, pp.1240–1251, 2023.
[27] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “Bpr: Bayesian personalized ranking from implicit feedback,” arXiv prepr. arXiv:1205.2618, 2012.
[28] G.J. Székely, M.L. Rizzo, and N.K. Bakirov, “Measuring and testing dependence by correlation of distances,” The Ann. of Stat., vol.35, no.6, pp.2769–2794, 2007.
[29] H. Zhu, X. Li, P. Zhang, G. Li, J. He, H. Li, and K. Gai, “Learning tree-based deep model for recommender systems,” Proc. of the 24th ACM SIGKDD Int. Conf. on Knowl. Discov. & Data Min., pp.1079–1088, 2018.
[30] Y. Liang, T. Qian, Q. Li, and H. Yin, “Enhancing domain-level and user-level adaptivity in diversified recommendation,” Proc. of the 44th Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr., pp.747–756, 2021.
[31] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv prepr. arXiv:1412.6980, 2014.
[32] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” arXiv prepr. arXiv:1802.03426, 2018.

Pólya-Gamma分布を使ったBayesian Factorization Machines

北原　洋一†

† 株式会社リブセンス　〒105–7510 東京都港区海岸 1-7-1 東京ポートシティ竹芝 10 階
E-mail: †yoichi.kitahara@livesense.co.jp

あらまし　Factorization Machines は Matrix Factorization を一般化したモデルで、コンテキスト情報を利用可能な協調フィルタリングに使うことができる。評価情報が 2 値の場合ロジスティック回帰タイプのモデルが使われることが多いが、ロジスティック回帰タイプの Factorization Machines に関するギブスサンプリング法は知られていなかった。本研究では Pólya-Gamma 分布を使ったデータ拡大法を利用して、ロジスティック回帰タイプの Bayesian Factorization Machines のパラメータをギブスサンプリングする方法を提案する。

キーワード　協調フィルタリング，行列分解

1 導　　入

レコメンデーションは、大量の情報の中からユーザーの興味や関心があるものを効率的に提示するのに有用で、さまざまな領域で広く使われている。例えば、オンラインショッピングでは顧客に適切な商品を提案するのに使われているし [1]、人材等のマッチングサービスなどでも利用されている。よりよいレコメンデーションを実現するため、レコメンデーション技術のさらなる発展が望まれている。

協調フィルタリングはレコメンデーションにおいて有効な手法の一つである。協調フィルタリングでは、行動パターンが類似した他のユーザーの評価情報に基づいてアイテムを推薦するため、パーソナライズされた推薦ができるという特徴がある。モデルベース協調フィルタリングを使う場合、評価情報がスコアのような連続値であれば回帰、クリックや購入のような離散値であれば分類として扱う。商品やエンターテイメントコンテンツのレビューなどでは回帰として扱われることが多いが、サービスの KPI に関わる機能では分類問題として扱われることが多い。

協調フィルタリングでは行動データがないあるいは少ないユーザーの扱いが困難であるというコールドスタート問題がある。レビューを利用したエンターテイメントコンテンツのレコメンデーションなどでは継続的なサービス利用が多いためコールドスタート問題の影響は比較的小さい。しかし、人材等のマッチングサービスではサービスの質が向上するほどユーザーの利用期間が短くなるためコールドスタート問題の影響を受けやすい。コールドスタート問題に対処する方法としては、コンテキスト情報を利用可能な協調フィルタリング手法が挙げられる。

Factorization Machines(FM) [2] [3] は Matrix Factorization(MF) [4] を一般化したモデルで、　コンテキスト情報を利用可能な協調フィルタリングに使うことができる。多くの MF 派生モデルを表現可能な表現能力の高いモデルであるにもかかわらず、高速な学習が可能であることが知られている [5]。しかしながら、学習データにオーバーフィットしやすいため、モデルパラメータの推定にはオーバーフィットが生じにくいベイズ推定が有効である。

FM のベイズ推定手法はいくつか提案されている。回帰については、ギブスサンプリング [6] や変分ベイズ [7] [8] が提案されている。分類については、プロビット回帰タイプのギブスサンプリング [3] やロジスティック回帰タイプの変分ベイズ [8] が提案されている。しかし、ロジスティック回帰タイプのギブスサンプリングは知られていない。

本研究では、協調フィルタリングでの利用を想定して、ロジスティック回帰タイプの Bayesian FM のパラメータをギブスサンプリングする方法を提案する。ロジットモデルのパラメータの条件付き分布を導出するために、Pólya-Gamma 分布を使ったデータ拡大法 [9] を利用する。また、[3], [5], [6] と同様に、ギブスサンプリングにおいてはモデルパラメータが一つずつサンプリングされるという特徴を活かした計算量の削減を図る。

2 関 連 研 究

FM は、スパースデータの交互作用を表現するのに優れたモデルである。協調フィルタリングでよく用いられる MF だけでなく、その派生モデルである Pairwise Interaction Tensor factorization [10] や SVD++ [11]、FPMC [12]、BPTF [13]、TimeSVD++ [14] といった幅広い行列分解モデルを表現可能である [3]。さらに、学習データにコンテキスト情報を含めることで、コンテキスト情報を利用可能な協調フィルタリングに使うこともできる。

FM のモデルパラメータを Altrnating Least Square(ALS) もしくはギブスサンプリングによって推定する場合は、事前計算を利用することで高速な学習が可能であることが知られている [5] [3]。FM がモデルパラメータについて線形であることと、ALS のパラメータ更新やギブスサンプリングのサンプリングではパラメータを独立させて扱うことを利用して、更新あるいはサンプリング対象となったパラメータに関わる値の差分のみを計算することで大幅な計算量が削減できる。近年では自動

微分を利用した機械学習フレームワークが充実しているため、様々な機械学習モデルや統計モデルのパラメータ推定が容易になったものの、このようなモデル特有の計算高速化手法を適用するのは難しい。なお、[5] で提案された計算量削減方法は、Stochastic Gradient Descent(SGD) のように一度に複数のモデルパラメータを更新する手法では使えない。さらに、内部に同一のパターンを有しているデータについては、ブロック構造をまとめて計算することで高速に計算する方法も提案されている [15]。

FM をベイズ推定する方法もいくつか提案されている。[6] では、モデルパラメータを表す分布に共役事前分布を仮定することで FM のモデルパラメータをギブスサンプリングする Bayesian Factorization Machines(BFM) が提案されている。モデルパラメータを表す分布のパラメータであるハイパーパラメータも推定することで、オーバーフィットを抑制するためのパラメータ調整を容易にしている。また、[5] で提案された事前計算を利用した高速な計算も利用できるという利点もある。[3] では、分類問題に対応した切断正規分布を利用したプロビット回帰タイプの FM のパラメータをギブスサンプリングする方法が紹介されている。[7] では、FM の変分ベイズ推定がバッチ学習とオンライン学習の二つについて提案されている。[8] では、機械学習フレームワークを使った変分ベイズ推定が提案されており、回帰だけでなく分類問題に対応したロジスティック回帰も扱っている。

FM の派生モデルも提案されている。[16] では特徴量の種類を Field として考慮した拡張がなされている。[17] では、より高次の交互作用を含むモデルの効率的な計算アルゴリズムが提案されている。また、Deep learning への応用もなされている [18] [19] [20]。

3 Bayesian Logistic Fctorization Machines

3.1 Factorization Machines

FM は、特徴量数を M、因子数を K とし、目的変数 $y \in \mathbb{R}$、説明変数 $\mathbf{x} \in \mathbb{R}^M$、モデルパラメータ $w_0 \in \mathbb{R}$、$\mathbf{w} \in \mathbb{R}^M$、$\mathbf{V} \in \mathbb{R}^{M \times K}$ を使って

$$\hat{y}(\mathbf{x}) = w_0 + \sum_{i=1}^{M} w_i x_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} \hat{w}_{i,j} x_i x_j \tag{1}$$

$$\hat{w}_{i,j} = \sum_{f=1}^{K} v_{i,f} v_{j,f} \tag{2}$$

と表される。式 (1) の第 3 項はそのまま計算すると $O(KM^2)$ の計算量となるが、変形することで $O(KM)$ で計算可能であることが知られている [2]。因子数 K は調整パラメータで、大きくするほど表現能力は高まるが、学習データにオーバーフィットしやすくなる。そのため、データに応じて適切な因子数を選択する。

2 値の分類問題に対応するため、ロジスティック回帰タイプの FM を考える。式 (1) の左辺を改めて $\hat{\psi}(\mathbf{x})$ とおき、目的変数を $y \in [0, 1]$ とし

$$\hat{\psi}(\mathbf{x}) = w_0 + \sum_{i=1}^{M} w_i x_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} \hat{w}_{i,j} x_i x_j \tag{3}$$

$$\tilde{y}(\mathbf{x}) = \frac{\exp(\hat{\psi}(\mathbf{x}))}{1 + \exp(\hat{\psi}(\mathbf{x}))} \tag{4}$$

としたモデルを本稿では Logistic Factorization Machines(LFM) と呼ぶ。

FM と同様に LFM の式 (3) は、いずれのモデルパラメータについても線形であるため、ある一つのモデルパラメータ $\theta \in \Theta = \{w_0, w_1, ..., w_M, v_{1,1}, ..., v_{M,K}\}$ のみに着目すると

$$\hat{\psi}(\mathbf{x}) = g_\theta(\mathbf{x}) + \theta h_\theta(\mathbf{x}) \tag{5}$$

と表記できる。$h_\theta(\mathbf{x})$ は

$$h_\theta(\mathbf{x}) = \begin{cases} 1 & (\theta = w_0) \\ x_l & (\theta = w_l) \\ x_l \sum_{j=1}^{M} v_{jf} x_j - v_{lf} x_l^2 & (\theta = v_{l,f}) \end{cases} \tag{6}$$

となる。ギブスサンプリングでは一つずつモデルパラメータをサンプリングするので、式 (5) を使うことでモデルパラメータを見通しよく統一的に扱うことができる。そこで、本稿でも式 (5) の表記を利用する。

3.2 Pólya-Gamma 分布を使ったデータ拡大法

ロジットモデルは解析的な扱いが難しいが、[9] にて Pólya-Gamma 分布を使ったデータ拡大法によって効率的なベイズ推論が可能であることが示されている。[9] では、次の関係が証明されている。

$$\frac{\left(e^{\psi}\right)^a}{\left(1 + e^{\psi}\right)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega \tag{7}$$

ここで、$\kappa = a - b/2$ であり、$p(\omega) = PG(\omega|b, 0)$ は Pólya-Gamma 分布である。式 (7) の左辺はロジットモデルに類似した形になっており、右辺は ω を与えたとき ψ について正規分布のカーネルになっていることを利用する。a を 2 値の目的変数 $y \in [0, 1]$ とし、$b = 1$ として、1 サンプルの条件付き分布は

$$P(y|\psi) = \frac{\left(e^{\psi}\right)^y}{1 + e^{\psi}} \tag{8}$$

$$\propto \exp\left(\kappa\psi - \frac{\omega\psi^2}{2}\right) \tag{9}$$

となる。このとき、潜在変数 ω の分布は

$$P(\omega|\psi) = PG(1, \psi) \tag{10}$$

となる。Pólya-Gamma 分布はガンマ分布の無限和となっているが、 $PG(1, \psi)$ から効率的なサンプリングが可能であることが知られている [9]。本稿でも、式 (9)、(10) の関係を利用してギブスサンプリングに必要な条件付き分布を導出する。

ロジットモデルであれば式 (9) の関係を広く利用できるため、多くの応用事例がある。例えば、MF を含む因子モデル [21] や Tensor Factorization [22] にも利用されている。

3.3 Bayesian Logistic Factorization Machines

本節では、Bayesian Logistic Factorization Machines(BLFM) のモデルと、ギブスサンプリングに必要な条件付き分布を示す。モデルパラメータの事前分布は [3] とほぼ同じものを用い、モデルパラメータを表す分布のパラメータであるハイパーパラメータも推定されるようにする。モデルパラメータの数が多いケースを想定し、モデルパラメータ間の相関を考慮しない事前分布となっている。

BLFM の各パラメータの分布は次のように設定する。

$$P(y_i|\mathbf{x}_i,\Theta,\omega) \propto \kappa_i\psi(\mathbf{x}_i,\Theta) - \frac{1}{2}\omega_i\psi(\mathbf{x}_i,\Theta)^2 \quad (11)$$

$$P(\omega_i|\mathbf{x}_i,\Theta) = PG(1,\psi(\mathbf{x}_i,\Theta)) \quad (12)$$

$$P(\theta|\Theta_H) = N(\mu_\theta,\lambda_\theta) \quad (13)$$

$$P(\mu_\theta|\Theta_H,\Theta_0) = N(\mu_\mu,\gamma_\mu\lambda_\theta) \quad (14)$$

$$P(\lambda_\theta|\Theta_0) = Gam(\alpha_\lambda,\beta_\lambda) \quad (15)$$

データサンプル数は N とし、i 番目のデータサンプルを $(y_i,\mathbf{x}_i)$ と表記した。ハイパーパラメータはまとめて $\Theta_H = \{\mu^w,\lambda^w,\mu_1^v,\cdots,\mu_K^v,\lambda_1^v,\cdots,\lambda_K^v\}$ と表した。また、モデルパラメータ θ に関するハイパーパラメータを μ_θ、λ_θ と表した。例えば、$\theta = w_1$ のハイパーパラメータは $\mu_\theta = \mu^w$、$\lambda_\theta = \lambda^w$ であり、$\theta = v_{2,3}$ であれば $\mu_\theta = \mu_3^v$、$\lambda_\theta = \lambda_3^v$ である。ハイパーパラメータの分布のパラメータは定数とし、まとめて $\mathbf{\Theta}_0 = \{\mu_\mu,\lambda_\mu,\alpha_\lambda,\beta_\lambda,\mu_0,\lambda_0\}$ と表記した。$N(\mu,\sigma^2)$ は平均 μ、分散 σ^2 の正規分布、$Gam(\alpha,\beta) \propto x^{\alpha-1}\exp(-\beta x)$ はガンマ分布である。

図 1 に BLFM のプレート表現を示す。

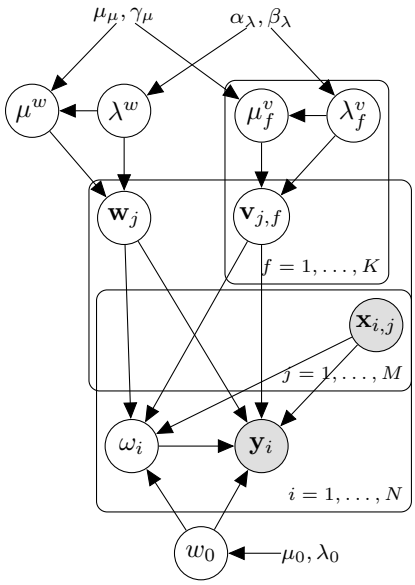


図 1　Bayesian Logistic Factorization Machines のプレート表現

次に、ギブスサンプリングに使う条件付き分布を示す。

式 (5) の表記を利用すると、モデルパラメータの条件付き分布は次のようになる。

$$P(\theta|\mathbf{y},\mathbf{X},\mathbf{\Theta}\setminus\{\theta\},\mathbf{\Theta}_H,\omega) \propto N(\tilde{\mu_\theta},\tilde{\sigma_\theta^2}) \quad (16)$$

$$\tilde{\mu_\theta} = \tilde{\sigma_\theta^2}\big(\mu_\theta\lambda_\theta + \sum_{i=1}^{N}\{\kappa_i - \omega_i g_\theta(\mathbf{x}_i)\}h_\theta(\mathbf{x}_i)\big) \quad (17)$$

$$\tilde{\sigma_\theta^2} = \big(\lambda_\theta + \sum_{i=1}^{N}\omega_i h_\theta(\mathbf{x}_i)^2\big)^{-1} \quad (18)$$

そのまま計算すると $h_\theta(\mathbf{x}_i)$ と $g_\theta(\mathbf{x}_i)$ の計算に時間がかかる。そこで、[5] と同様に 4.2 節で説明する方法で計算量を削減する。

パラメータ θ を表す分布のパラメータであるハイパーパラメータ μ_θ と λ_θ の条件付き分布は

$$P(\mu_\theta|\mathbf{\Theta},\mathbf{\Theta}_H\setminus\{\mu_\theta\},\mathbf{\Theta}_0) \propto N(\tilde{\mu_{\mu_\theta}},\tilde{\sigma_{\mu_\theta}^2}) \quad (19)$$

$$\tilde{\mu_{\mu_\theta}} = \tilde{\sigma_{\mu_\theta}^2}\lambda_\theta(\mu_\mu\gamma_\mu + \sum_{j=1}^{M}\theta_j) \quad (20)$$

$$\tilde{\sigma_{\mu_\theta}^2} = \big(\lambda_\theta(\gamma_\mu + M)\big)^{-1} \quad (21)$$

$$P(\lambda_\theta|\mathbf{\Theta},\mathbf{\Theta}_H\setminus\{\lambda_\theta\},\mathbf{\Theta}_0) \propto Gam(\tilde{\alpha_{\lambda_\theta}},\tilde{\beta_{\lambda_\theta}}) \quad (22)$$

$$\tilde{\alpha_{\lambda_\theta}} = \frac{1}{2}(2\alpha_\lambda + M + 1) \quad (23)$$

$$\tilde{\beta_{\lambda_\theta}} = \frac{1}{2}\Big\{\sum_{j=1}^{M}(\theta_j - \mu_\theta)^2 + \gamma_\mu(\mu_\theta - \mu_\mu)^2\Big\} + \beta_\lambda \quad (24)$$

となる。

また、潜在変数 ω_i の条件付き分布は

$$P(\omega_i|\mathbf{y},\mathbf{X},\mathbf{\Theta}) = PG(1,\hat{\psi}(\mathbf{x}_i,\mathbf{\Theta})) \quad (25)$$

となる。

4 協調フィルタリングにおける実装

4.1 学習データの構造

協調フィルタリングに FM を利用する場合、学習データはユーザー、アイテム、状況といった 3 種類のデータが連結されたものとなる。

ユーザーデータは、ユーザーが決まると一意に定まるデータで、ユーザー ID とユーザー属性から構成される。ユーザー ID は、ユーザー ID に対応するカラムの値が 1、それ以外が 0 となるスパースなデータとして表現される。そのため、ユーザー ID を表すカラムはユーザー数と同数必要になる。ユーザー属性は、ユーザーの特徴を表すデータである。例えば、ユーザーの年齢を表す数値、性別や職業を表すダミー変数などはユーザー属性である。また、過去に閲覧したアイテムなどもユーザー属性である。

アイテムデータは、アイテムが決まると一意に定まるデータで、アイテム ID とアイテム属性から構成される。ユーザー ID と同様にアイテム ID も、アイテム ID に対応するカラムの値が 1 それ以外が 0 となるスパースなデータとして表現される。アイテム属性は、アイテムの特徴を表すデータである。例えば、アイテムのリリース年やジャンルを表すダミー変数などはアイテム属性である。

状況データは、ユーザーやアイテムが決まっても定まらないデータである。例えば、アクセス時間やアクセス媒体などは状

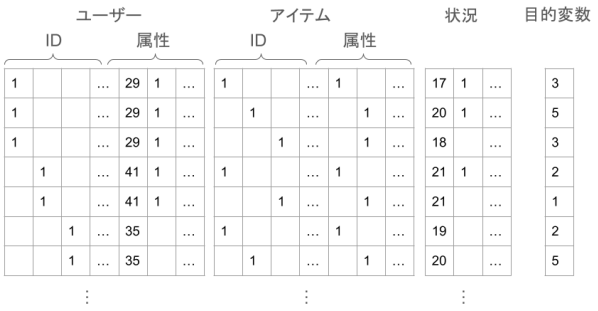


図 2　協調フィルタリングで使う FM のデータ

況データである。

協調フィルタリングで使うデータを図に示すと図 2 となる。例えば、図 2 において、ID データはカラム順に ID が連番で対応しているとし、一番左のユーザー属性は年齢、一番左のアイテム属性は何らかのジャンル A、一番左の状況データは見た時間とすると、図 2 のデータにおいて、29 歳のユーザー ID=1 のユーザーは 17 時にジャンル A に属するアイテム ID=2 のコンテンツを見たときに、評価などの目的変数の値が 3 であったことを示す。図 2 からもわかるように、ID データはカラム数が多い非常にスパースなデータになるため、実装においてはスパースデータ構造を用いるなどの対応が必要になる。

協調フィルタリングにおいては、あるユーザーに対してアイテムと状況のデータが与えられた時のスコアやコンバージョン率などの予測値を計算し、予測値が高いアイテムを予測値が高くなる状況においてレコメンドする。

4.2　計算量の削減

モデルパラメータのサンプリング時と予測値計算時の双方において計算量の削減を図る。

モデルパラメータのサンプリング時の計算量削減は、[5] とほぼ同様の方法を使う。共通計算部分を事前に計算しておくことで繰り返し計算を回避することと、差分のみを計算することで計算量の多い $\psi(\mathbf{x}_i, \Theta)$ や $g_\theta(\mathbf{x}_i)$ の直接計算を回避することで計算量の削減を図る。

まず、$\psi(\mathbf{x}_i, \Theta)$ の計算について考える。更新後のパラメータを θ^* で表すと、パラメータ更新後の $\psi_i(\mathbf{x}_i, \Theta^*)$

$$\psi_i(\mathbf{x}_i, \Theta^*) = \psi_i(\mathbf{x}_i, \Theta) + (\theta^* - \theta) h_\theta(\mathbf{x}_i) \tag{26}$$

となる。そのため、事前に $\psi_i(\mathbf{x}_i, \Theta)$ を計算し記憶しておけば、モデルパラメータ更新時に $O(1)$ で更新できる。

次に、$h_\theta(\mathbf{x}_i)$ の計算について考える。$h_\theta(\mathbf{x}_i)$ は θ が w_0 あるいは w_l の場合は容易であるため、$\theta = v_{l,f}$ のケースについてのみ考える。$\theta = v_{l,f}$ のときの $h_\theta(\mathbf{x}_i)$ は

$$\bar{q}_{i,f} = \sum_{j=1}^{M} v_{j,f} x_{i,j} \tag{27}$$

を事前計算しておくことで、次のように $O(1)$ で計算できる。

$$\begin{aligned} h_\theta(\mathbf{x}_i) &= x_{i,l} \sum_{j=1}^{M} v_{j,f} x_{i,j} - v_{l,f} x_{i,l}^2 \\ &= x_{i,l} \bar{q}_{i,f} - v_{l,f} x_{i,l}^2 \end{aligned} \tag{28}$$

$v_{l,f}$ が更新された時 $\bar{q}_{i,f}$ は次のように $O(1)$ で更新することができる。

$$\bar{q}^*_{i,f} = \bar{q}_{i,f} + (v^*_{l,f} - v_{l,f}) x_{i,j} \tag{29}$$

なお、$g_\theta(\mathbf{x}_i)$ は $\psi(\mathbf{x}_i, \Theta) - \theta h_\theta(\mathbf{x}_i)$ から計算する。$\psi(\mathbf{x}_i, \Theta)$ と $h_\theta(\mathbf{x}_i)$ が計算されていれば、$g_\theta(\mathbf{x}_i)$ を直接計算する必要はない。

以上のようにすることで、1 パラメータの更新を $O(N)$ で行うことができるので、1 エポックあたり $O(NMK)$ で計算が可能になる。これは、回帰タイプの FM と同じ計算量である。

次に、予測値の計算方法について説明する。

[3] では、モデルパラメータのサンプリング時にテストデータに関する予測値を計算するアルゴリズムを提案しているが、評価実験などでは有用でも実運用では利用しにくいという問題がある。評価実験等では予測値計算対象となるデータ量は学習データのデータ量と同等もしくはそれより少ないことが多い。一方、実際に協調フィルタリングで利用する場合は、最大でユーザー数とアイテム数と状況数を掛け合わせた数のデータに対して予測値を計算する。状況データの特徴量数は多くないことが多いので以下では状況データを無視して考える。ユーザーとアイテムのみを考えユーザー数を N^u、アイテム数を N^i、サンプリング数を S としたとき、1 データの予測値を計算するのに $O(SMK)$ の計算量が必要なので、予測計算に $O(SMKN^uN^i)$ の計算が必要になる。学習時と予測時のデータ量の比はデータに強く依存するものの現実のデータでは $N \ll N^uN^i$ であることが多く、レビューデータなど 1 人あたりのサンプルデータが多いデータではおよそ 10 倍以上、人材サービスデータなどの 1 人あたりのサンプルデータが少ないデータではおよそ 100 倍以上、予測時のデータ量が多くなる。　そのため、実運用においては [3] にて提案されている方法を使うのは現実的ではないことがある。

本稿では、事前計算を利用して予測計算全体の計算量を特徴量数に依存しない $O(SKN^uN^i)$ にする方法を考える。大量の計算が必要になるのはユーザーとアイテムのデータを組み合わせた計算のときに生じるため、可能な限りユーザーのみあるいはアイテムのみで計算可能な部分を事前に計算しておくことで計算量の削減を図る。

予測値算出式について、ユーザーの特徴量数を M^u、アイテムの特徴量数を M_i、サンプル数を S とし、サンプル番号を下付き添字 $_s$ で示し、ユーザーデータとアイテムデータを表すのに上付き添字 ^{u} と ^{i} を使って、式 (3) を書き直すと

$$\hat{\psi}(\mathbf{x}) = \frac{1}{S}\sum_{s=1}^{S}\Big[w_{s,0} + \sum_{j=1}^{M^u} w_{s,i}^u x_j^u + \sum_{j=1}^{M^i} w_{s,j}^i x_j^i + \sum_{l=1}^{M^u}\sum_{j=l+1}^{M^u} \hat{w}_{s,l,j}^u x_l^u x_j^u + \sum_{l=1}^{M^i}\sum_{j=l+1}^{M^i} \hat{w}_{s,l,j}^i x_l^i x_j^i + \sum_{l=1}^{M^u}\sum_{j=1}^{M^i} \hat{w}_{s,l,j}^{u,i} x_l^u x_j^i\Big] \quad (30)$$

$$\hat{w}_{s,l,j}^u = \sum_{f=1}^{K} v_{s,l,f}^u v_{s,j,f}^u \quad (31)$$

$$\hat{w}_{s,l,j}^i = \sum_{f=1}^{K} v_{s,l,f}^i v_{s,j,f}^i \quad (32)$$

$$\hat{w}_{s,l,j}^{u,i} = \sum_{f=1}^{K} v_{s,l,f}^u v_{s,j,f}^i \quad (33)$$

となる。

ユーザーのみ、アイテムのみで計算可能な式 (30) の第 2、3、4、5 項は、ユーザーのみあるいはアイテムのみのデータで

$$\bar{r}^u = \frac{1}{S}\sum_{s=1}^{S}\Big[\sum_{j=1}^{M^u} w_{s,j}^u x_j^u + \sum_{l=1}^{M^u}\sum_{j=l+1}^{M^u} \hat{w}_{s,l,j}^u x_l^u x_j^u\Big] \quad (34)$$

$$\bar{r}^i = \frac{1}{S}\sum_{s=1}^{S}\Big[\sum_{j=1}^{M^i} w_{s,j}^i x_j^u + \sum_{l=1}^{M^i}\sum_{j=l+1}^{M^i} \hat{w}_{s,l,j}^i x_l^u x_j^u\Big] \quad (35)$$

として、学習時に計算できる。ユーザーとアイテムそれぞれ $O(SKM^uN^u)$ と $O(SKM^iN^i)$ で計算できるため計算量が問題になることは少ない。

計算量が多くなるのはユーザーとアイテムの組み合わせの項である式 (30) の第 6 項である。ここで、

$$\bar{q}_{s,f}^u = \sum_{j=1}^{M^u} v_{s,j,f}^u x_j^u \quad (36)$$

$$\bar{q}_{s,f}^i = \sum_{j=1}^{M^i} v_{s,j,f}^i x_j^i \quad (37)$$

をサンプリング時に事前計算しておくと、

$$\begin{aligned}
&\frac{1}{S}\sum_{s=1}^{S}\sum_{l=1}^{M^u}\sum_{j=1}^{M^i} \hat{w}_{s,l,j}^{u,i} x_l^u x_j^i \\
&= \frac{1}{S}\sum_{s=1}^{S}\sum_{l=1}^{M^u}\sum_{j=1}^{M^i}\sum_{f=1}^{K} v_{s,l,f}^u v_{s,j,f}^i x_l^u x_j^i \\
&= \frac{1}{S}\sum_{s=1}^{S}\sum_{f=1}^{K}\left(\sum_{l=1}^{M^u} v_{s,l,f}^u x_l^u\right)\left(\sum_{j=1}^{M^i} v_{s,j,f}^i x_j^i\right) \\
&= \frac{1}{S}\sum_{s=1}^{S}\sum_{f=1}^{K} \bar{q}_{s,f}^u \bar{q}_{s,f}^i
\end{aligned} \quad (38)$$

と $O(SK)$ で計算が可能になる。つまり、学習時に $\bar{r}^u$ と $\bar{r}^i$、$\bar{q}_{s,f}^u$、$\bar{q}_{s,f}^i$ を計算しておくことで、式 (30) は $O(SK)$ で計算できるため、予測計算全体の計算量を $O(SKN^uN^i)$ とすることができる。

4.3 アルゴリズム

式 (16) から式 (25) の θ に具体的なモデルパラメータ を当てはめ、4.2 節の計算量削減を行うと、BLFM のギブスサンプリングは Algorithm1 となる。なお、$\mathbf{h} = (h_1, \cdots, h_N)^\top$ など、サンプルデータを一括して扱っているところではベクトル表記を用いている。Algorithm1 ではモデルパラメータの初期値をすべて 0 としているが乱数を使用しても構わない。モデルパラメータの初期値を全て 0 とすると、ψ や $\mathbf{q}$ も 0 になるため初期計算を省略できる利点がある。Algorithm1 では明示していないが、初期のサンプルはバーンインとして予測には使わないようにすることが望ましい。また、協調フィルタリングではデータがスパースなため、実装にはスパースデータ構造を利用するなどスパースデータを効率的に扱うための工夫が必要である。

5 検証実験

本節では、プロビットタイプの FM と比較した検証実験の結果を示す。評価指標には ROC 曲線の AUC(Area Under Curve) と Accuracy を用いた。比較対象モデルとして、libfm[1]のプロビットタイプの FM を用いた。以下、提案手法は lfm、比較対象モデルは libfm と表記する。検証はロジスティックタイプとプロビットタイプの FM で以下の点に違いが生じるかに着目して行われた。

- データ量
- 2 値のバランス
- 属性の有無
- 1 人あたりの評価データ数
- 因子数

データは Movielens[2]の 100K データセットと 1M データセットを利用した。100K データセットはあらかじめ 5 分割されたテストデータが用意されているため、それらをそのまま利用した。1M データセットについてはランダムに 10 分割したデータを用いた交差検証を行った。モデルパラメータの初期値については、libfm は 0 を平均とし 0.1 を標準偏差とした乱数とし、lfm は全て 0 としている。また、定数として与える、ハイパーパラメータを表す分布のパラメータは、$\mu_\mu = \mu_0 = 0$、$\lambda_\mu = \lambda_0 = \alpha_\lambda = \beta_\lambda = 1$ としている。なお、バーンインは 30 としている。

2 値のバランスの影響を調べるため、目的変数については、評価 4 点以上を 1 それ以外を 0 としたデータセットと、評価 5 点以上を 1 それ以外を 0 としたデータセットの 2 種類を用いた。100K データセットでは評価 4 点以上の割合は 55.4%、5 点以上の割合は 21.2%であり、1M データセットでは評価 4 点以上の割合は 57.5%、5 点以上の割合は 22.6%である。2 値のバランスが、評価 4 点以上を 1 としたときは比較的均衡しており、5 点以上を 1 とした場合は不均衡になっているため、前者を均衡データ、後者を不均衡データと呼ぶ。コンバージョンなどを

1：http://www.libfm.org
2：https://grouplens.org/datasets/movielens/

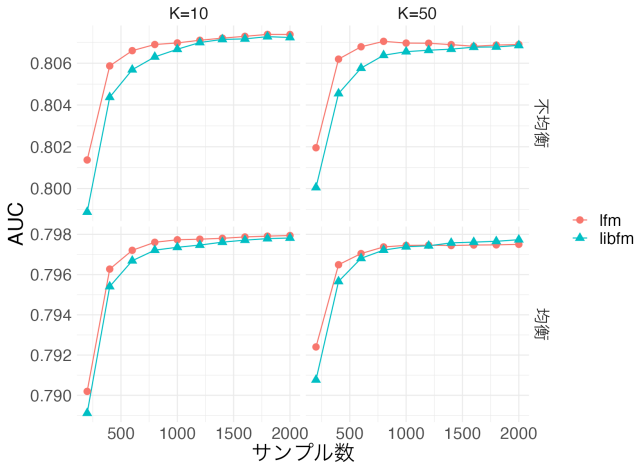


図 3　サンプル数と AUC の関係 (100K データセット)

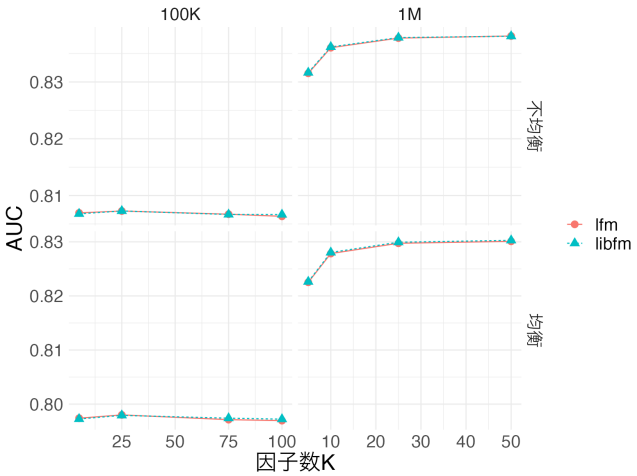


図 4　属性なしでのデータサイズ (100K, 1M) による AUC の違い (左段:100K, 右段:1M、上段: 不均衡データ、下段: 均衡データ)

扱う場合は不均衡データとなることが多いため、不均衡データにおいて高精度なモデルが望ましい。

属性の影響を調べるため、特徴量にユーザーの性別と映画ジャンルを加えたケースについて検証した。いずれの特徴量も 1 もしくは 0 をとるダミー変数とした。

1 人あたりの評価データ数が少なくなったときの影響を調べるため、1M データセットの学習データにおいてユーザーごとに 5%、10%、20%の評価しか学習に利用しないケースについて検証した。なお、人材サービスのレコメンデーションでは、1 人あたりの評価データ数が少くなる傾向にあるため、1 人あたりの評価データ数が少なくなったときでも精度が維持されるモデルが望ましい。

属性情報なしの 100K データセットを使って、均衡、不均衡データについて、因子数 $K = 10, K = 50$ における ROC 曲線の AUC を計算し、サンプル数との関係をプロットしたものを図 3 に示す。概ねサンプル数が増加するほど AUC は上昇する傾向が見られるが、1,000〜1,500 あたりから上昇幅が急激に小さくなる。そこで、本研究では、サンプル数を 1,500 とした。ただし、属性情報なしの 100K データセットではサンプル数 2,000 の結果を示す。なお、提案手法と libfm とでサンプル数が少ない時の AUC に違いが生じている原因としては、モデルの違いに加えて初期値の扱いの違いが考えられる。

属性なしの 100K データセットと 1M データセットそれぞれの不均衡データと均衡データを使って ROC 曲線の AUC を因子数別にプロットしたものを図 4 に示す。提案手法と libfm とでは、データサイズや 2 値のバランスが異なっても AUC にほとんど差が生じないことがわかる。なお、いずれのモデルでも因子数 K の影響を受けることも確認できる。

100K データセットで属性ありのデータとなしのデータを使って ROC 曲線の AUC と Accuracy を計算し因子数別にプロットしたものを図 5 に示す。提案手法と libfm とでは、Accuracy にほとんど差がないことがわかる。属性なしの AUC でも提案手法と libfm にはほとんど差がないものの、属性ありの AUC ではやや違いが見られる。均衡データ、不均衡データいずれにおいても、libfm と比較して提案手法は因子数 K が小さい時に

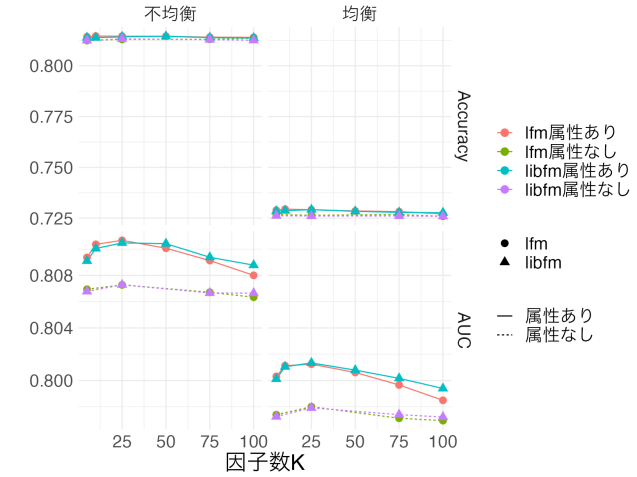


図 5　100K データセットにおける属性情報の有無による AUC と Accuracy の違い (左段: 不均衡データ, 右段: 均衡データ、上段:Accuracy、下段:AUC)

AUC が比較的高く、K が大きくなると AUC が比較的小さくなる傾向が見られる。ただし、AUC が高くなる因子数 K が 5〜25 の範囲ではモデル間で大きな差は生じておらず、モデルの違いより適切な因子数 K の選択のほうが AUC の向上に影響すると考えられる。また、いずれのモデルでも因子数 K の影響を受けることや属性を使うことで AUC が向上することが確認できる。

1M データセットにおいてユーザーごとに 5%、10%、20%の評価のみを学習データとして使って均衡データと不均衡データについて AUC を計算し因子数別にプロットした結果を図 6 に示す。いずれのケースにおいても libfm と比較して提案手法の AUC がわずかに高くなっているものの、大きな差はないことがわかる。評価データ数を 20%、10%、5%に削減してもモデル間の AUC の差はほとんど変わらないため、評価データ数による影響も見られない。

プロビット回帰とロジスティック回帰の予測結果は類似していることが知られている。サンプル数が十分にあり適切な因子

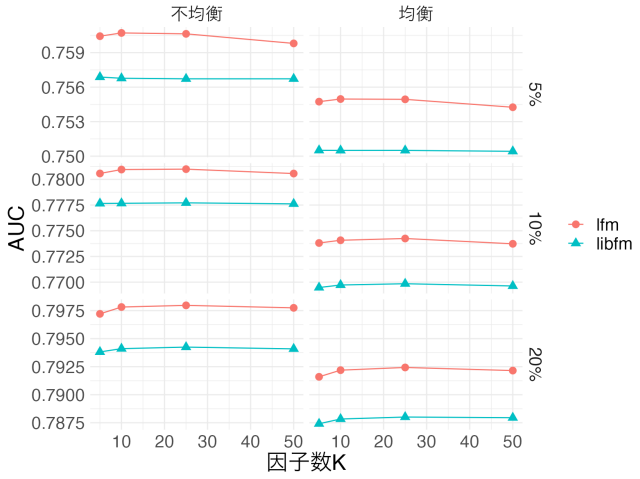


図 6 (1M データセットにおいて一部のデータのみを学習したときの違い (左段: 不均衡データ, 右段: 均衡データ、上段:5%データで学習、中段:10%データで学習、下段:20%データで学習))

数を選択した場合、プロビットタイプの FM とロジスティック回帰タイプの FM の予測精度の差はほとんどないことから、FM の場合でも類似した予測結果になりやすいと考えられる。

6 ま と め

本研究では Pólya-Gamma 分布を使ったデータ拡大法を利用することで、ロジスティック回帰タイプの Bayesian Factorization Machines のパラメータを、近似なしの条件付き分布を使ってギブスサンプリングする方法を提案した。また、学習時および予測時における計算量の削減方法も示した。

検証実験では、データサイズ、2 値のバランス、属性の有無、1 人あたりの評価データ数、因子数によって、ロジスティックタイプの FM とプロビットタイプの FM の精度に差が生じるかについて検証した。その結果、サンプル数が十分にあり適切な因子数が選択された場合、ロジスティックタイプの FM とプロビットタイプの FM はほとんどのケースにおいてほぼ同程度の精度となることが確認された。

文 献

[1] Linden, G., Smith, B. and York, J.: Amazon.Com Recommendations: Item-to-Item Collaborative Filtering, *IEEE Internet Computing*, Vol. 7, No. 1, pp. 76–80 (2003).
[2] Rendle, S.: Factorization Machines, in *2010 IEEE International Conference on Data Mining*, pp. 995–1000 (2010).
[3] Rendle, S.: Factorization Machines with libFM, *ACM Transactions on Intelligent Systems and Technology*, Vol. 3, No. 3, pp. 1–22 (2012).
[4] Srebro, N. and Jaakkola, T.: Weighted low-rank approximations, in *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 720–727 (2003).
[5] Rendle, S., Gantner, Z., Freudenthaler, C. and Schmidt-Thieme, L.: Fast Context-Aware Recommendations with Factorization Machines, in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 635–644 (2011).
[6] Freudenthaler, C., Schmidt-Thieme, L. and Rendle, S.: Bayesian Factorization Machines, in *Proceedings of the NIPS Workshop on Sparse Representation and Low-rank Approximation* (2011).
[7] Saha, A., Misra, R., Acharya, A. and Ravindran, B.: Scalable Variational Bayesian Factorization Machine (2017).
[8] Vie, J.-J., Rigaux, T. and Kashima, H.: Variational Factorization Machines for Preference Elicitation in Large-Scale Recommender Systems, in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 5607–5614 (2022).
[9] Polson, N. G., Scott, J. G. and Windle, J.: Bayesian Inference for Logistic Models Using Pólya – Gamma Latent Variables, *Journal of the American Statistical Association*, Vol. 108, No. 504, pp. 1339–1349 (2013).
[10] Rendle, S. and Schmidt-Thieme, L.: Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 81–90 (2010).
[11] Koren, Y.: Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 426–434 (2008).
[12] Rendle, S., Freudenthaler, C. and Schmidt-Thieme, L.: Factorizing Personalized Markov Chains for Next-Basket Recommendation, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 811–820 (2010).
[13] Xiong, L., Chen, X., Huang, T.-K., Schneider, J. and Carbonell, J. G.: Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization, in *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 211–222 (2010).
[14] Koren, Y.: Collaborative Filtering with Temporal Dynamics, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 447–456 (2009).
[15] Rendle, S.: Scaling Factorization Machines to Relational Data, *Proceedings of the VLDB Endowment*, Vol. 6, No. 5, pp. 337–348 (2013).
[16] Juan, Y., Zhuang, Y., Chin, W.-S. and Lin, C.-J.: Field-Aware Factorization Machines for CTR Prediction, in *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 43–50 (2016).
[17] Blondel, M., Fujino, A., Ueda, N. and Ishihata, M.: Higher-Order Factorization Machines, in *Advances in Neural Information Processing Systems*, Vol. 29 (2016).
[18] Wen, P., Yuan, W., Qin, Q., Sang, S. and Zhang, Z.: Neural Attention Model for Recommendation Based on Factorization Machines, Vol. 51, No. 4, pp. 1829–1844 (2021).
[19] He, X. and Chua, T.-S.: Neural factorization machines for sparse predictive analytics, in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 355–364 (2017).
[20] Li, Z., Wu, S., Cui, Z. and Zhang, X.: GraphFM: Graph Factorization Machines for Feature Interaction Modeling (2022).
[21] Klami, A.: Polya-Gamma Augmentations for Factor Models, in *Proceedings of the Sixth Asian Conference on Machine Learning*, Vol. 39, pp. 112–128 (2015).
[22] Rai, P., Hu, C., Harding, M. and Carin, L.: Scalable Probabilistic Tensor Factorization for Binary and Count Data, in *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 3770–3776 (2015).

Algorithm 1 BLFM のギブスサンプリング

function BLFM_GS($\mathbf{y}$, $\mathbf{x}, K, S$)
  $w_0, \mathbf{w}, \mathbf{v}, \psi, \mathbf{q}, \bar{w_0}, \mathbf{r}^u, \mathbf{r}^i \leftarrow 0,$
  for $s \in \{1, \ldots, S\}$ **do**
    $\mu_{\tilde{\mu^w}} \leftarrow \sigma^2_{\tilde{\mu^w}} \lambda^w \left(\mu_\mu \gamma_\mu + \sum_{j=1}^M w_j\right)$
    $\sigma^2_{\tilde{\mu^w}} \leftarrow \left(\lambda^w(\gamma_\mu + M)\right)^{-1}$
    sample μ^w from $N(\mu_{\tilde{\mu^w}}, \sigma^2_{\tilde{\mu^w}})$
    $\alpha_{\tilde{\lambda^w}} \leftarrow \frac{1}{2}(2\alpha_\lambda + M + 1)$
    $\beta_{\tilde{\lambda^w}} \leftarrow \frac{1}{2}\left\{\sum_{j=1}^M (w_j - \mu^w)^2 + \gamma_\mu(\mu^w - \mu_\mu)^2\right\} + \beta_\lambda$
    sample λ^w from $Gam(\alpha_{\tilde{\lambda^w}}, \beta_{\tilde{\lambda^w}})$
    for $f \in \{1, \ldots, K\}$ **do**
      $\mu_{\tilde{\mu^v_f}} \leftarrow \sigma^2_{\tilde{\mu^v_f}} \lambda^v_f (\mu_\mu \gamma_\mu + \sum_{j=1}^M v_{j,f})$
      $\sigma^2_{\tilde{\mu^v_f}} \leftarrow \left(\lambda^v_f(\gamma_\mu + M)\right)^{-1}$
      sample μ^v_f from $N(\mu_{\tilde{\mu^v_f}}, \sigma^2_{\tilde{\mu^v_f}})$
      $\alpha_{\tilde{\lambda^v_f}} \leftarrow \frac{1}{2}(2\alpha_\lambda + M + 1)$
      $\beta_{\tilde{\lambda^v_f}} \leftarrow \frac{1}{2}\left\{\sum_{j=1}^M (v_{j,f} - \mu^v_f)^2 + \gamma_\mu(\mu^v_f - \mu_\mu)^2\right\} + \beta_\lambda$
      sample λ^v_f from $Gam(\alpha_{\tilde{\lambda^v_f}}, \beta_{\tilde{\lambda^v_f}})$
    end for
    sample ω from $PG(1, \hat{\psi}(\mathbf{x}_i, \mathbf{\Theta}))$
    $\mu_{\tilde{w_0}} \leftarrow \sigma^2_{\tilde{w_0}}\left(\mu_0\lambda_0 + \sum_{i=1}^N \{\kappa_i - \omega_i(\psi(\mathbf{x}_i) - w_0)\}\right)$
    $\sigma^2_{\tilde{w_0}} \leftarrow \left(\lambda_0 + \sum_{i=1}^N \omega_i\right)^{-1}$
    sample w_0^* from $N(\mu_{\tilde{w_0}}, \sigma^2_{\tilde{w_0}})$
    $\psi \leftarrow \psi + (w_0^* - w_0)$
    $w_0 \leftarrow w_0^*$
    for $j \in \{1, \ldots, M\}$ **do**
      $\mu_{\tilde{w_j}} \leftarrow \sigma^2_{\tilde{w_j}}\left(\mu^w\lambda^w + \sum_{i=1}^N \{\kappa_i - \omega_i(\psi(\mathbf{x}_i) - w_j x_{i,j})\} x_{i,j}\right)$
      $\sigma^2_{\tilde{w_j}} \leftarrow \left(\lambda^w + \sum_{i=1}^N \omega_i x_{i,j}^2\right)^{-1}$
      sample w_j^* from $N(\mu_{\tilde{w_j}}, \sigma^2_{\tilde{w_j}})$
      $\psi \leftarrow \psi + (w_j^* - w_j)\mathbf{x}_j$
      $w_j \leftarrow w_j^*$
    end for
    for $f \in \{1, \ldots, K\}$ **do**
      for $j \in \{1, \ldots, M\}$ **do**
        for $i \in \{1, \ldots, N\}$ **do**
          $h_i \leftarrow x_{i,j}\bar{q}_{i,f} - v_{j,f}x_{i,j}^2$
        end for
        $\mu_{\tilde{v_{j,f}}} \leftarrow \sigma^2_{\tilde{v_{j,f}}}\left(\mu^v_f\lambda^v_f + \sum_{i=1}^N \{\kappa_i - \omega_i(\psi(\mathbf{x}_i) - v_{j,f}h_i)\}h_i\right)$
        $\sigma^2_{\tilde{v_{j,f}}} \leftarrow \left(\lambda^\theta_f + \sum_{i=1}^N \omega_i h_i^2\right)^{-1}$
        sample $v^*_{j,f}$ from $N(\mu_{\tilde{v_{j,f}}}, \sigma^2_{\tilde{v_{j,f}}})$
        $\psi \leftarrow \psi + (v^*_{j,f} - v_{j,f})\mathbf{h}$
        $\mathbf{q}_f \leftarrow \mathbf{q}_f + (v^*_{j,f} - v_{j,f})\mathbf{x}_j$
        $v_{j,f} \leftarrow v^*_{j,f}$
      end for
    end for
    $\bar{w_0} \leftarrow \bar{w_0} + w_0$
    for $i \in \{1, \ldots, N^u\}$ **do**
      $r^u_i \leftarrow r^u_i + \sum_{j=1}^{M^u} w^u_i x^u_{i,j} + \sum_{l=1}^{M^u}\sum_{j=l+1}^{M^u} \hat{w}^u_{l,j} x^u_{i,l} x^u_{i,j}$
    end for
    for $i \in \{1, \ldots, N^i\}$ **do**
      $r^i_i \leftarrow r^i_i + \sum_{j=1}^{M^i} w^i_i x^i_{i,j} + \sum_{l=1}^{M^i}\sum_{j=l+1}^{M^i} \hat{w}^i_{l,j} x^i_{i,l} x^i_{i,j}$
    end for
    for $f \in \{1, \ldots, K\}$ **do**
      $\bar{q}^u_{s,f} \leftarrow \sum_{j=1}^{M^u} v^u_{s,j,f} x^u_j, \quad \bar{q}^i_{s,f} \leftarrow \sum_{j=1}^{M^i} v^i_{s,j,f} x^i_j$
    end for
  end for
  $\bar{w_0} \leftarrow \bar{w_0}/S$, $\bar{\mathbf{r}}^u \leftarrow \mathbf{r}^u/S$, $\bar{\mathbf{r}}^i \leftarrow \mathbf{r}^i/S$
  return $\bar{w_0}, \bar{\mathbf{q}}^u, \bar{\mathbf{q}}^i, \bar{\mathbf{r}}^u, \bar{\mathbf{r}}^i$
end function

一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2024年3月1日(金) 16:00 ~ 18:10 | T3-C オンライン（Zoom Events)

グラフ処理

座長:早水 悠登(東京大学)
コメンテータ:加藤 誠(筑波大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T3-C-9-01] 異種グラフ表現学習を用いた悪意のあるSNS投稿の検出
*吉田 燎[1]、吉田 壮[1]、棟安 実治[1] (1. 関西大学大学院 理工学研究科)

16:30 ~ 16:55
[T3-C-9-02] 隣接ノードの情報を利用するランダムウォークベースのグラフ特徴量推定
*長谷川 毅[1]、廣中 詩織[1]、首藤 一幸[1] (1. 京都大学)

16:55 ~ 17:20
[T3-C-9-03] 局所的な密度の指標として機能するハイパーグラフ向けクラスタ係数
*宮下 陸矢[1]、廣中 詩織[2]、首藤 一幸[2] (1. 東京工業大学、2. 京都大学)

17:45 ~ 18:10
[T3-C-9-05] GPUを用いた大規模グラフ処理アルゴリズムの高速化
*岡﨑 大知[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学)

異種グラフ表現学習を用いた悪意のあるSNS投稿の検出

吉田　燎† 吉田　壮†† 棟安　実治††

† 関西大学大学院 理工学研究科　〒564–8680 大阪府吹田市山手町 3–3–35
†† 関西大学 システム理工学部　〒564–8680 大阪府吹田市山手町 3–3–35
E-mail: †{k924073,sohy,muneyasu}@kansai-u.ac.jp

あらまし　SNS の発展に伴い，ユーザは気軽に意見や感情を投稿し，他者と議論を交わせるようになった．しかし，これによりヘイトスピーチや有害情報の拡散など，悪意ある投稿が問題視されている．従って，SNS 上のこれらの行動を効果的に検出する技術開発が求められている．既存の研究では，自然言語処理を用いた投稿の意味情報の抽出や，ユーザ，投稿，その他要素を含む異種グラフにグラフニューラルネットワーク（GNN）を適用することで，投稿間の潜在的な関係やユーザ間のつながりを考慮したモデルが有効であることが示されている．しかし，これらのモデルでは，異種グラフのノード（例えば，ユーザ，投稿，ハッシュタグなど）とエッジ（例えば，ユーザ間のフォロー関係や投稿間の共有関係）の異種性を十分に考慮せずに学習を行っており，グラフの表現力は最大化されていない．本論文では，これらの異種性を考慮した学習手法を採用し，より効果的に悪意ある投稿を検出できるアーキテクチャを提案する．提案手法の有効性は X(旧 Twitter) データを用いた悪意のある投稿を検出するタスクにおいて評価され，各手法との比較の結果，提案手法は従来手法を上回る性能を示した．

キーワード　SNS, グラフニューラルネットワーク, 異種グラフ, 悪意のある投稿検出

1 はじめに

X(旧 Twitter) [1] に代表される SNS の発展にともない，ユーザは SNS を通じて自由な発言と議論を容易に行うことができるようになった．その一方で，ヘイトスピーチ，攻撃的な発言，誤情報の拡散など悪意のある投稿も存在し，大きな社会問題となっている．近年では，SNS における選挙活動において，多くの誹謗中傷やデマが飛び交う「選挙ヘイト」問題 [2] や，COVID-19 のワクチンに関して，虚偽の文書の配布などにより刑事上，民事上問題になる様々な事件が発生 [3] するなど，SNS の投稿を原因に多くの問題が生じている．

このような背景のもと多くの研究者によって，テキスト情報から自然言語処理手法 (NLP) などを用いて悪意のある投稿を検出する研究が行われてきた．こういったテキストの意味情報を用いて悪意を検出する手法は，主に BERT (Bidirectional Encoder Representations from Transformers) [1] などの手法を用いて投稿におけるテキストの特徴量を抽出し，特徴量を MLP などに入力することで悪意のある投稿の検出を行う．しかし，投稿の意味情報のみを考慮する NLP を用いた手法では，例えば悪意のある情報を発信しているにもかかわらず，発言がもっともらしい内容であったり，文面が丁寧であったりすると，NLP のみではその悪質性が判断できないという問題点がある．つまり，テキスト情報だけではわからない投稿のコンテキストを考慮することが，悪意のある投稿の検出精度向上に繋がる．

これに対して，近年，投稿内容の意味情報に加え，投稿におけるコンテキスト，つまり関係情報を考慮した検出手法が有望視されている．このような手法では Jiele らが提案した GraphBERT [2] が代表的であり，この手法では投稿に含まれるハッシュタグやエンティティなどの要素を抽出し，「ユーザ」「投稿」「投稿の要素」の大きく 3 種類のノードを形成し，これらの関係性を表現した異種グラフを構築する．構築した異種グラフにおいて，NLP を用いて各ノードにおける文書や単語の特徴量をグラフの初期特徴量とし，GNN を用いてグラフ構造を学習することで，投稿における意味情報だけでなくコンテキスト情報，つまり関係情報も考慮したアーキテクチャを設計する．

このような手法は，投稿の意味情報のみを考慮する従来の手法に比べ，悪意のある投稿を検出するタスクで高い精度を上げている．しかし，GraphBERT では投稿の表現方法に異種グラフを採用しているが，GNN の学習においてグラフの異種性を考慮できていないという問題点がある．先述したように，投稿における関係情報を表現するグラフを作成する際，投稿に含まれるハッシュタグやエンティティなどの要素を抽出し，「ユーザ」「投稿」「投稿の要素」の 3 種類のノードから異種グラフを構築する．こうすることで，投稿のテキスト情報では表現できない関係情報を補完することができるが，構築した異種グラフに対して同種グラフを想定している GNN を適用しても，その異種性を考慮する効果は薄いと考えられる．具体例を挙げると，⟨ 投稿，ユーザ ⟩ 間のエッジと，⟨ 投稿，投稿の要素 ⟩ 間のエッジにはそれぞれ特有の意味があり，既存の GNN ではそれらを区別なしに学習してしまうため，異種グラフにおける関係情報を効果的に学習できない．

そこで本研究では，先ほどの未検討点に対し，SNS 投稿の関係性をより詳細に表現した異種グラフを構築し，GNN の学習に Ziniu らの提案した HGT(Heterogeneous Graph Trans-

1：https://twitter.com/
2：https://www.fnn.jp/articles/-/417710
3：https://www.pref.tottori.lg.jp/item/1230295.htm

former) [3] を用いて学習を行うことで，投稿の意味情報だけでなくコンテキスト情報も考慮した異種グラフ表現に基づく悪意のある SNS 投稿の検出を提案する．HGT は既存の GNN と異なり，学習の際にノード，エッジの異種性に注目した学習を行う．具体的には，異種グラフ内の関係情報 (メタパス) ごとに特有の表現空間を保持し，それらの表現空間を考慮した学習を行う．また，HGT の適用にあたりグラフの表現力をより高めるため，GraphBERT [2] で提案された異種グラフに，更なるノード，エッジタイプの追加を検討する．このようにして構築された異種グラフに，メタパスを考慮したモデルである HGT を適用することで，従来手法より高精度な検出を可能とするモデルを提案する．また，一意に「悪意のある投稿」といっても，その定義は非常に曖昧で困難のものである．その投稿が「他者への誹謗中傷」「他人や特定団体を陥れたり扇動する内容」なのか,「単なる感想」なのかどうかを判断することが重要である．SNS の魅力は誰でも意見を発信できることであり,「単なる感想」までも排除するのは良策ではない．そのためにも，ある投稿が普段どういった投稿を行っているユーザから発信されたものなのか，どういったユーザから拡散されているのかなどを考慮する本手法は，目標に対しての有効な手法であると考えられる．実験では，実際に収集した X データを用いて，悪意のある投稿を検出するタスクにおいて各手法との比較を行い，提案手法は従来手法を上回る性能を示した．

本論文の貢献は以下の 3 点に整理できる：

（1） **拡張異種グラフの構築**：異種グラフにおけるノードとエッジの種類を増やすことで，SNS 投稿に含まれる多様な情報を表現する包括的な異種グラフの構築を実現する．このアプローチにより，悪意のある SNS 投稿に関連する様々な要素間の微妙な関係を詳細に把握することが可能になる．

（2） **異種性を考慮したグラフ表現学習**：本研究は，GNN の学習において，異種グラフのノードとエッジの多様性をより深く捉えるために，HGT を用いたアプローチを採用する．これにより，SNS 投稿の複雑な関係性をより正確に表現する．

（3） **実データに基づく実験と評価**：実際に収集した X の実データを用いた実験に加え，アブレーション研究を通じて，異種グラフ内の多様なノードとエッジタイプの具体的な貢献を評価した．これにより，各要素が悪意のある投稿検出性能に与える影響を明らかにする．

2 関 連 研 究

悪意のある投稿を検出する手法は様々提案されており，投稿の単純な意味情報だけでなく，その関係情報も考慮に入れた手法が有望視されている．また，異種グラフに有効な GNN についての研究も行われており，グラフの異種性を考慮した手法が提案されている．本章ではそれらの関連研究を示す．

2.1 悪意ある投稿検出に関する従来研究

SNS における悪意のある投稿を検出する手法として，主にヘイトスピーチの検出に関する研究が多く行われている．SNS への投稿において，各ユーザは自由に匿名で自分自身を表現することができるが，過度な表現やヘイトスピーチは，他者に対する憎しみの誘発や事件の引き金となってしまう可能性を含み，一般的な SNS ではヘイトスピーチのコンテンツを削除するポリシーを設けている [4]．そのため，こういったヘイトスピーチを自動的に検出する技術が必要となっている．

こういった検出手法の多くは，投稿のテキスト情報から NLP を用いて検出に有効な埋め込みを取得している．その中でも代表的なものが，BERT [1] を用いた手法である BERT は，双方向性 Transformer エンコーダ [4] に基づき，大規模なコーパスから双方向性表現を学習するように事前に訓練された言語モデルであり，SNS における悪意のある投稿を検出する手法においても有効な結果を示している．Rajput らの研究 [5] では，他の NLP を用いて得られる SNS 投稿の単語埋め込みを，BERT 埋め込みに置き換えまたは統合することによって，ヘイトスピーチ検出の精度が向上することを示している．しかしこれらの NLP を用いた投稿の意味情報のみを考慮する手法では，テキスト情報のみからその悪質性を判断するため，テキストだけでは判断できない投稿のコンテキストを考慮できていないという問題点がある．

2.2 GraphBERT [2]

Jiele によって提案された GraphBERT は，悪意のある投稿を検出する際，投稿におけるテキストの意味情報に加え，投稿におけるコンテキスト，つまり関係情報を考慮した検出手法である．GraphBERT のアーキテクチャは，以下のようになる．まず，SNS における「ユーザ」と「投稿」をグラフのノードとして形成する．ある「ユーザ」が各「投稿」を行っているという関係があれば，それらのノード間にエッジを形成する．こうすることで，各ユーザがどのような投稿を行っているのかを表現したグラフが完成する．次に,「投稿の要素」に注目する.「投稿の要素」とは，投稿内に登場したハッシュタグ，メンション，エンティティ，トピックのことを指す．それらの要素を抽出し,「投稿」内に各「投稿の要素」が含まれていれば，それらのノード間にエッジを形成する．こうすることで，各投稿にどのような要素が含まれているかを表現したグラフが完成する．以上のような手順で SNS 投稿の関係情報を表現した異種グラフを構築し，グラフに対して一般的な GNN モデルである GCN [6] を適用することで，悪意のある投稿を検出するために有効な埋め込みを学習する．

このような特徴を持つ GraphBERT の問題点として，埋め込みを取得するためのグラフ学習方法が挙げられる．GraphBERT では，投稿を表現する際に「投稿」ノードに加え,「ユーザ」や「投稿の要素」など異種ノードを使うことで複雑な関係を表現している．しかし，悪意のある投稿を検出するタスクのグラフ学習の際には，異種グラフに対し GCN など同種グラフ用に設計された従来の GNN を適用している．この方法では，学習の際にノードの異種性などを考慮せず学習を行っていることにな

4：https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

り，グラフの異種性による表現力を最大限に引き出すことはできないと考えられる．

2.3 Heterogeneous Graph Transformer [3]

HGT は，異種グラフ用に設計された GNN であり，グラフの異種性を考慮するための Transformer [4] のアーキテクチャを採用したモデルである．異種グラフは，異なるタイプのノードが様々な方法で互いに相互作用する複雑なシステムの抽象化とモデル化に使用されている．同種グラフと異なり，異種グラフには，注目ノードにおける「メタパス」という概念が存在する．メタパスとは，異種グラフ内の関係情報のことを指し，「注目しているターゲットノード」「ターゲットノードと近傍ノードを結ぶエッジ関係」「注目している近傍ノード」の種類から定義される．例えば，GraphBERT においては，「ユーザ」ノードと「投稿」ノードの関係を表すメタパス，「投稿」ノードと「投稿の要素」ノードの関係を表すメタパスなど，ノードやエッジの種類ごとに複数のメタパスが存在する．HGT では，ノードやエッジの異種性を表現するために，各エッジ上の異種 Attention を特徴づけるためのメタパスに依存するパラメータを設計し，異なるタイプのノードとエッジに対して固有の表現を維持できるようにする．

HGT の具体的なアーキテクチャを説明する．異種グラフは $G = (V, E, A, R)$ として表現され，各ノード $v \in V$ および各エッジ $e \in E$ はそれぞれ，タイプマッピング関数 $\tau(v) : V \to A$ および $\phi(e) : E \to R$ に関連付けられる．また，ソースノード s からターゲットノード t にリンクされたエッジ $e = (s, t)$ のメタ関係は，$\langle \tau(s), \phi(e), \tau(t) \rangle$ として示される．HGT では，t をターゲットノードとし，s_1, s_2 をソースノードとする異種グラフが与えられると，そのエッジ $e_1 = (s_1, t), e_2 = (s_2, t)$ およびそれらに対応するメタ関係 $\langle \tau(s_1), \phi(e_1), \tau(t) \rangle, \langle \tau(s_2), \phi(e_2), \tau(t) \rangle$ を入力として，各ノードに対するコンテキスト化された表現 $H[t]$ を得る．このようなメタ関係に基づいた学習を実現するために，Transformer のアーキテクチャを採用する．

まずは，ソースノードとターゲットノード間の Attention を計算する．Attention を用いた同種 GNN である GAT [7] では，近傍ノードの情報集約の際に，1 つの重み行列 W を用いてソースノードとターゲットノードが同じ特徴分布を持つと仮定している．このような仮定は，注目ノードごとに異なるメタパスが存在する異種グラフでは有効ではない．そこで，Transformer のアーキテクチャを参考に，ターゲットノード t を Query ベクトルに，ソースノード s を Key ベクトルにマッピングし，それぞれの積を Attention として計算する．具体的には，各エッジ $e = (s, t)$ に対する Attention を次のように計算する．

$$\begin{aligned}&\mathbf{Attention}_{HGT}(s, e, t)\\ &= \underset{\forall s \in N(t)}{\mathrm{Softmax}}\Big(\big\|_{i \in [1,h]} \mathit{ATT\text{-}head}^i(s, e, t)\Big) \qquad (1)\end{aligned}$$

$$\begin{aligned}\mathit{ATT\text{-}head}^i(s, e, t) &= (K^i(s) W^{ATT}_{\phi(e)} Q^i(t)^T)\\ &\quad \cdot \frac{\mu_{\langle \tau(s), \phi(e), \tau(t) \rangle}}{\sqrt{d}} \qquad (2)\end{aligned}$$

$$K^i(s) = \text{K-Linear}^i_{\tau(s)}(H^{(l-1)}[s]) \qquad (3)$$

$$Q^i(t) = \text{Q-Linear}^i_{\tau(t)}(H^{(l-1)}[t]) \qquad (4)$$

ここで，$N(t)$ はノード t のすべてのソースノード，$H^{(l-1)}[t]$ は注目ノードにおける $(l-1)$ 層目の HGT 層の出力である．まず，i 番目の Attention ヘッド $\mathit{ATT\text{-}head}^i(s, e, t)$ を得るために，$\tau(s)$ 型のソースノード s を，線形射影 $\text{K-Linear}^i_\tau(s) : \mathbb{R}^d \to \mathbb{R}^{d/h}$ を用いて，i 番目の Key ベクトル $K^i(s)$ に射影する．ここで，h は Attention ヘッドの数であり，d/h はヘッドごとのベクトル次元である．$\text{K-Linear}^i_\tau(s)$ は，ソースノード s のタイプ $\tau(s)$ によって定義される．これにより，各ソースノードタイプごとに固有の表現を保持する．同様に，ターゲットノード t も，線形射影 $\text{Q-Linear}^i_\tau(t)$ を使用して i 番目の Query ベクトルに射影する．次に，各エッジタイプ $\phi(e)$ に対して固有のエッジベース行列 $W^{ATT}_{\phi(e)} \in \mathbb{R}^{d/h \times d/h}$ を保持する．これにより，Attention を各メタパスごとに固有の表現分布に基づいて計算する．さらに，すべての関係が等しくターゲットノードに寄与するわけではないので，テンソル $\mu \in \mathbb{R}^{|\mathcal{T}_V| \times |\mathcal{T}_E| \times |\mathcal{T}_V|}$ を追加し，Attention に対する適応的なスケーリングを行う．最後に，h 個の Attention ヘッドを連結して，各ノードペアの $\mathbf{Attention}_{HGT}(s, e, t)$ を得る．

Attention の計算と並行して，ソースノードからターゲットノードに情報を受け渡す Message の計算を行う．この際にも，Attention の計算と同様にメタ関係を考慮することで，異なるタイプのノードとエッジの分布の違いを緩和する．注目しているソースノード s に対して，Message を次のように計算する．

$$\mathbf{Message}_{HGT}(s, e, t) = \big\|_{i \in [1,h]} \mathit{MSG\text{-}head}^i(s, e, t) \qquad (5)$$

$$\mathit{MSG\text{-}head}^i(s, e, t) = \text{M-Linear}^i_{\tau(s)}(H^{(l-1)}[s]) W^{MSG}_{\phi(e)} \qquad (6)$$

ここで，i 番目の Message ヘッド $\mathit{MSG\text{-}head}^i(s, e, t)$ を得るために，$\tau(s)$ 型のソースノード s を，線形射影 $\text{M-Linear}^i_{\tau(s)} : \mathbb{R}^d \to \mathbb{R}^{d/h}$ を用いて i 番目のメッセージベクトルに射影する．次に，各エッジタイプに対して固有のエッジベース行列 $W^{MSG}_{\phi(e)} \in \mathbb{R}^{d/h \times d/h}$ を保持することで，各メタパスごとに固有の表現分布に基づいて計算を行う．最後に，h 個の Message ヘッドを連結して，各ノードペアの $\mathbf{Message}_{HGT}(s, e, t)$ を取得する．

$\mathbf{Attention}_{HGT}(s, e, t)$ と $\mathbf{Message}_{HGT}(s, e, t)$ を計算した後，それらをソースノードからターゲットノードに集約する．ここでは，ソースノードからの対応する Message を平均し，更新されたベクトル $\tilde{H}^{(l)}[t]$ を次のように得る．

$$\tilde{H}^{(l)}[t] \leftarrow \underset{\forall s \in N(t)}{\mathbf{Aggregate}}(\mathbf{Attention}_{HGT}(s,e,t) \cdot \mathbf{Message}_{HGT}(s,e,t)) \quad (7)$$

$$\mathbf{Aggregate}(\cdot) = \sigma\left(\mathrm{Mean}(\cdot)\right) \quad (8)$$

ここで $\sigma(\cdot)$ は活性化関数を表す．最後に，ターゲットノード t のベクトルを，そのノードタイプ $\tau(t)$ によって固有の分布にマッピングする．そのために，更新されたベクトル $\tilde{H}^{(l)}[t]$ に線形射影 A-Linear$_{\tau(t)}$ を適用し，その後，以下のように残差接続を適用する．

$$H^{(l)}[t] = \text{A-Linear}_{\tau(t)}(\sigma(\tilde{H}^{(l)}[t])) + H^{(l-1)}[t] \quad (9)$$

このようにして，ターゲットノード t に対する l 層目の HGT 層の出力 $H^{(l)}[t]$ を得る．HGT モデルの特徴をまとめると，重み行列を別々にパラメータ化するために，Transformer のアーキテクチャを参考にメタ関係 $\langle \tau(s), \phi(e), \tau(t) \rangle$ に依存したモデルを構築している．こうすることで，異種性を考慮し，異種グラフの表現力を最大限に活用した学習を可能としている．

3 提 案 手 法

本章では，本研究で提案する，投稿の意味情報と関係情報を表現したグラフの異種性を考慮したモデルについて説明する．提案手法では，Jiele の提案した GraphBERT のアーキテクチャを参考に異種グラフを構築し，さらに表現力を高めるためにノードやエッジの追加を行う．次に構築した異種グラフに対して Ziniu らの提案した HGT を適用することで，グラフの異種性を考慮し，関係情報を取得した埋め込みを取得することができる．取得した埋め込みを MLP に入力することで，投稿の意味情報と関係情報から悪意のある投稿の検出を行う．

3.1 異種グラフの構築

構築する異種グラフの概要を図 1 に示す．異種グラフは $G = (V, E, X, \tau, \phi)$ として定義され，V はノードの集合，E はエッジの集合，$X \in \mathbb{R}^{|V| \times d_x}$ は特徴行列を示す．また，$x_v \in \mathbb{R}^{d_x}$ はノード $v \in V$ の特徴ベクトルを示す．各ノード $v \in V$ にはタイプ $\tau(v)$ があり，各エッジ $e \in E$ にはタイプ $\phi(e)$ がある．ここで，異種グラフに登場するノード V の種類は，V_u(ユーザ),V_p(投稿),V_o(投稿の要素) の大きく 3 種類のノードとそれらを結ぶエッジ関係が存在する．

異種グラフのエッジを構築するためにまず,「ユーザ」と「投稿」にエッジを形成する．つまり，ある「ユーザ」がある「投稿」を発信した場合,「ユーザ」と「投稿」間にエッジ E_p を結ぶ．次に，収集した投稿の要素を抽出する．ここでの要素とは，投稿内のメンション，ハッシュタグ，絵文字，エンティティ，およびトピックのことを指す．つまり，V_o を詳しく見ると，V_{me}，V_{ht}，V_{em}，V_{et}，V_{tp} といったように，構築する異種グラフには合計で 7 種類のノードが存在する．投稿に各要素が含まれていた場合,「投稿」と「投稿の要素」間にエッジ E^o_{in} を結ぶ．このような「ユーザ」「投稿」間のエッジ,「投稿」「投稿の要素」間のエッジを形成することは，投稿の意味情報だけでは取得できない情報を補完する役割を果たす．具体的には，特定のユーザが発信，特定のメンションやハッシュタグ，絵文字などを含んだ投稿は悪質性を持つ可能性が高いことなど，投稿に含まれている要素に注目することによって，投稿文中の意味だけでは推測できないコンテキストを表現することが可能となる．

各ノードの具体的な形成方法について説明する．まず，投稿されたテキスト全体から，メンションとハッシュタグ，絵文字を抽出しノードとする．そして Python ライブラリ SpaCy[5]を用いてエンティティを抽出する．ここでは，今回対象としている投稿に関連深いと考えられるエンティティタイプとして，地名，施設名，組織名，名前の 4 つを選択した．これらのノード (メンション，ハッシュタグ，絵文字，エンティティ) に関しては，より影響力が強いものを抽出するために収集した投稿内で 3 回以上登場するもののみに限定し，ノードとして形成している．次に，投稿のトピック分類を行い，抽出したトピックをノードとする．トピック分類のために，各投稿におけるハッシュタグや絵文字などを削除，投稿内の数字を全て 0 に変換するなどクリーニング処理を行い，得られた文書群に対して latent dirichlet allocation (LDA) [8] を使用してトピック分類を行う．LDA は文書群のすべての文書の潜在的な意味構造のクラスタリングを行い，すべての投稿に最も可能性の高いトピックを提供することができる．LDA では生成するトピック数，トピックを構成する単語の数をそれぞれ指定することがでる．今回の実験ではトピック数 30，そしてトピック内は 10 個の単語で構成されるよう設定した．

ノード v からは事前学習済み BERT [9] を用いて 768 次元 $(=d_x)$ の特徴ベクトル x_v を取得している．投稿，ハッシュタグ，エンティティ，トピックに関しては，各文書，単語に対して BERT を適用し，絵文字に関しては Python ライブラリ emoji[6] を使用して絵文字を単語に変換し BERT を適用している．ユーザノードに関しては，各ユーザの投稿の特徴量の平均をユーザ特徴量としている．

次に，先ほど形成した「ユーザ」「投稿」間,「投稿」「投稿の要素」間のエッジに加え,「投稿」「投稿」間のエッジを追加する．「投稿」間のエッジは，各投稿における拡散ユーザの類似度に基づいて結ぶ．つまり，投稿を拡散したユーザが類似している投稿間にはエッジを結ぶということになる．これは，拡散ユーザが類似している投稿同士は，同じような嗜好のユーザが興味を示している投稿として関連性があるという仮定に基づいてエッジを結ぶ手法になる．エッジの形成方法については，鳥海らが提案した類似度計算方法 [10] を用いる．類似度は，各投稿ペアの拡散ユーザを Simpson 係数を用いて測定する．Simpson 係数は 2 つの集合における類似度を計算する手法であり，対象とする要素ペア，つまり今回の実験においては，投稿を拡散したユーザ集合ペアのうち，拡散数の小さい投稿の拡散数を分母とすることで，拡散数の差異の影響を抑えた類似度を算出することができる．投稿セット T に含まれる t_i を拡散したユーザ集

5：https://spacy.io/
6：https://github.com/carpedm20/emoji

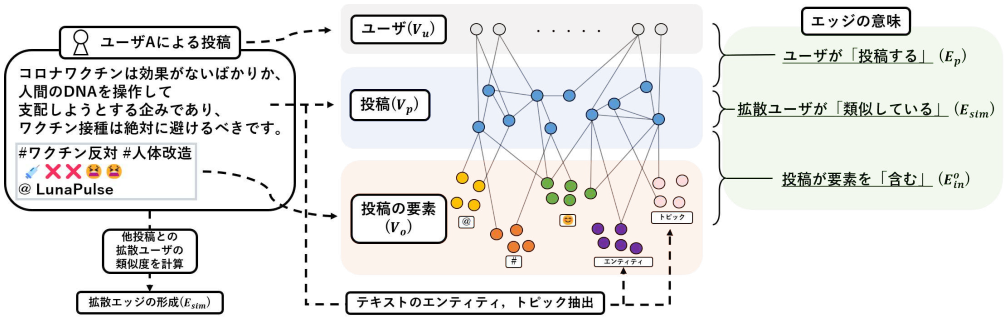


図 1: 異種グラフの概要図

表 1: ノード情報

node_type		counts		
user		23,308		172,110
post		134,950		
other	hashtag	2,262	13,852	
	emoji	753		
	mention	750		
	topic	30		
	entity	10,057		

表 2: エッジ情報

edge_type		counts		
user-post		134,950		1,061,409
post-post		465,986		
other	post-hashtag	41,642	460,473	
	post-emoji	49,987		
	post-mention	10,173		
	post-topic	134,950		
	post-entity	223,721		

合を $\mathbf{r}_i$ としたとき，2 つの投稿 $\mathbf{r}_i, \mathbf{r}_j$ の類似度は，以下のような式で表される．

$$Sim(\mathbf{r}_i, \mathbf{r}_j) = \frac{|\mathbf{r}_i \cdot \mathbf{r}_j|}{min(|\mathbf{r}_i|, |\mathbf{r}_j|)} \tag{10}$$

ここで，$|\mathbf{r}_i \cdot \mathbf{r}_j|$ は，2 つの集合 $\mathbf{r}_i$ と $\mathbf{r}_j$ のうち，重複している要素数を表す．すべての投稿のペアについて，類似度 $Sim(\mathbf{r}_i, \mathbf{r}_j)$ を計算し，類似度が閾値 0.8 を超えた投稿ペアについてエッジを結ぶ．こうすることで，拡散したユーザが特に類似している投稿間の関係を表現することができる．

これらすべての処理を収集した全投稿に施し，投稿の関係情報を表現した大規模な異種グラフを構築する．以上の処理を行うことで，投稿における意味情報に加え，コンテキストを考慮した表現力の高い異種グラフを構築する．これらの条件によって構築されたグラフは，172,110 個のノードと 1,061,409 個のエッジを含む．統計の詳細は，表 1，表 2 に示す．

3.2 HGT を用いた異種グラフ表現学習

構築した異種グラフに対して，異種グラフのメタパス情報を考慮したモデルである HGT を適用して学習を行う．構築した異種グラフには，ノード間ごとに $\langle V_u, E_p, V_p \rangle$ (ユーザ，投稿する，投稿)，$\langle V_p, E^o_{in}, V_o \rangle$(投稿，含む，投稿の要素)，$\langle V_p, E_{sim}, V_p \rangle$(投稿，類似する，投稿) のようにそれぞれのメタ関係を持っている．今回構築した異種グラフのメタパスを具体的にみると，エッジが結ばれたノード間の関係，つまり，ユーザと投稿間では $\langle V_u, E_p, V_p \rangle$，投稿と投稿の要素間では $\langle V_p, E^{me}_{in}, V_{me} \rangle$，$\langle V_p, E^{ht}_{in}, V_{ht} \rangle$，$\langle V_p, E^{em}_{in}, V_{em} \rangle$，$\langle V_p, E^{tp}_{in}, V_{tp} \rangle$，$\langle V_p, E^{et}_{in}, V_{en} \rangle$，そして投稿間では $\langle V_p, E_{sim}, V_p \rangle$ の計 7 種類のメタ関係が定義される．このようにノードやエッジの種類の違いを注視し，それらを考慮した学習を行うことが，グラフ内の複雑な関係性を正確に表現することに繋がる．

HGT を適用することで，これら全てのメタパス情報ごとに特有の表現空間を保持することができる．つまり，異種グラフのグラフ学習における Attention を計算する際，各メタパスごとに保持した特有の表現空間を考慮することにより，異種性を考慮した適切な Attention を計算することができる．そうすることで，従来の同種グラフを想定した GNN と比較し，異種グラフ構造をより表現した埋め込みを算出することができる．今回構築した異種グラフにおいては，ユーザと投稿を結ぶエッジと，投稿と投稿の要素を結ぶエッジ，そのなかでも投稿とメンション，投稿とハッシュタグといったように，それぞれのノード間の関係を考慮した表現空間を保持することで，グラフの異種性の効果を最大限に発揮した学習から，投稿の特徴量の学習を行う．

HGT の学習から得られた埋め込みを MLP に入力することで特徴を抽出し，投稿における悪意の有無の 2 値分類を行う．このように HGT から得られた埋め込みを直接 MLP に入力し学習する手法を，**Proposed method** と表記する．また，投稿本文のテキスト情報をより効果的に維持するため，事前学習済み BERT から得られる投稿の初期特徴 x_{t_i} を，L 層の HGT 学習により得られる埋め込み $H^{(L)}[t_i]$ とスキップ接続する．この方法を **Proposed method**+と表記する．一般に，データセットに含まれる悪意のある投稿は正常な投稿に比べて数が少なく，不均衡である．このため，本研究では損失関数に focal loss (FL) [11] を使用して学習を行う．

$$\mathcal{L}_{\text{GNN-FL}} = -\sum_{v \in V_p} y_i (1-p_v)^{\gamma} \log(p_v) + (1-y_i) {p_v}^{\gamma} \log(1-p_v) \tag{11}$$

$$p_v = \text{Softmax}[\sigma(H^{(L)}[v] \cdot W_1) \cdot W_2] \tag{12}$$

ただし，$H^{(L)}[v]$ は L 層 HGT により得られる埋め込み，y_i は

正解ラベルであり，1 は悪意のある投稿，0 は悪意のない投稿示し，γ はハイパーパラメータ，W_1 および W_2 は MLP の学習可能なパラメータ行列，σ は活性化関数であり，予測は p_v を用いて行われる．FL では，誤分類した学習データの重みを増加させることで，データの不均衡問題に対処している．

3.3 アノテーション

本研究ではアノテーションのために，Google が開発した，文章の攻撃性をスコアで出力できる機械学習 API である Perspective API [7] を用いる．この API は，Wikipedia [8] や The New York Times [9] などのオンラインフォームからのコメントを含む，様々なソースからの何百万ものコメントに基づいて各モデルを学習させている API であり，与えられた文章がどれほど有害であるかを，0〜1 の確率スコアで出力できるものである．取得できるスコア属性としては，TOXICITY（毒性，攻撃性）の他に，INSULT（侮辱性），PROFANITY（冒涜性），THREAT（脅威性），INFLAMMATORY（挑発，または煽る意図である炎上性）などがあり，目的により使用するスコアを使い分けることができる．

本研究ではこの Perspective API における TOXICITY，INSULT のスコアを用いて投稿本文の悪意を分析し，実験で用いる実データのアノテーションの判断材料とする．これらのスコアは，収集した投稿のうちサンプリングされたいくつかの投稿の Perspective API スコアを目視で確認した結果から選定されたものになる．いくつかの投稿のうち目視で悪意があると判断した投稿において，スコアにおける TOXICITY，INSULT の 2 つの値が高い傾向が見られた．また，これらの悪意を判断するのためのスコアは，同じく目視で確認した値を参考に，TOXICITY が 0.35，INSULT が 0.25 を閾値と設定し，どちらか一方以上が閾値を超えたものを悪意のある投稿のアノテーション付けを行った．

ここで，現時点で考えられるアノテーションの問題点について説明する．先述したように，アノテーションにはテキストベースの機械学習 API である Perspective API を用いている．ここで，Perspective API はテキストの有害性を検出するのに有効な手法ではあるが，検出の判断にはテキスト情報のみを使用しており，投稿のコンテキストや関係情報を考慮するという本手法の本質とは異なっているという問題点がある．しかし，本研究の目的は異種性を考慮したアーキテクチャの提案であり，この点においては本研究のような実験計画でも有効性の検証が正しく行われると考えられるため，本実験においては Perspective API を用いたアノテーションを行っている．

4 実　　験

4.1 データセットの詳細

今回実験対象とした X のデータは，日本における COVID-19 のワクチン接種に関する投稿である．COVID-19 に関しては，接種開始時期や緊急事態宣言が発令されていた時期に，COVID-19 やそのワクチン接種に関する投稿が活発に行われていた．その中には，ワクチン接種に対して懐疑的な内容や，接種者や医療機関に対する攻撃的で悪意のある内容のものも多く含まれており，本論文の趣旨である悪意のある投稿を検出するタスクにおいて有効なデータセットであると考えられる．具体的には，「コロナ ワクチン 接種」or「COVID ワクチン 接種」のクエリを含むものを収集し，期間は 2020 年 1 月から 2023 年 4 月のものを対象としている．収集した投稿は約 340 万件に上り，実際に実験の対象としたデータは，他者への影響が強いと考えられるものに限定するために，収集したもののうち拡散数が 10 件以上のものである 134,950 件に絞り込んだ．このデータ規模は，文献 [2] を上回るものであり，本分野における要求仕様は満たしていると考えられる．

4.2 実験の詳細

次に，実験の詳細について説明する．実験で使用する各 GNN では，最適化には Adam を使用し，学習率は 5e-4，重み減衰は 1e-4 である．MLP の活性化関数には Rectified Linear Unit (ReLU) を使用した．各手法は 2 層の GNN からなり，隠れ層の次元数は 64 と 128 の 2 条件で実験を行っており，提案手法のアテンションヘッド数は 8 と設定している．次に，実験において提案手法と比較する従来手法について説明する．**Proposed method** および **Proposed method**+の比較手法として，一般的な GNN モデルである GCN [6], GAT [7], GraphSAGE [12] を用いた手法と比較を行う．構築した異種グラフに対し従来の GNN を適用した手法と比較することで，異種グラフを用いたグラフ学習における異種性考慮の重要性を確認することができる．つまり，提案手法で適用する異種 GNN 学習において「グラフ内要素の異種性の考慮」という側面の有効性を検証することができる．加えて，異種グラフ用に開発された Attention ベースのモデルである HAN [13] を用いた手法との比較も行った．HAN のアルゴリズムには，ノードとそのメタパスに基づく隣接ノードとの間の重要度を学習する「ノードレベル Attention」と，異なるメタパスの重要度を学習する「セマンティックレベル Attention」が存在し，HGT と違い 2 つの階層的 Attention に基づく GNN であることが特徴である．また，テキストベースの比較手法として事前学習済み BERT [9] を用いたものを採用する．この比較手法では，投稿文に BERT を適用することによって抽出される 768 次元の特徴ベクトルを基に MLP を訓練し，悪意のある投稿の検出を行う．これらの比較手法は提案手法と同じく G を入力とし，分類損失は FL を用いる．また，実験条件も提案手法に準拠し，MLP のパラメータ数も同じである．

また，今回対象としたデータセットの問題点として，データの不均衡性が挙げられる．データセットにおける全投稿 134,950 件のうち，正常な投稿が 125,438 件，悪意のある投稿が 9,512 件となっており極めて不均衡なデータセットであることがわかる．本研究においては，不均衡データの影響を軽減するために

7：https://perspectiveapi.com/
8：https://en.wikipedia.org/wiki/Main_Page
9：https://www.nytimes.com/

表 3: 実験結果

Methods	F1_macro_score							
	条件 1（隠れ層の次元数 64，層数 2）				条件 2（隠れ層の次元数 128，層数 2）			
	20% Train	40% Train	50% Train	60% Train	20% Train	40% Train	50% Train	60% Train
BERT+MLP	68.37	68.30	70.61	69.74	**71.30**	**71.81**	71.41	71.44
GCN	59.15	62.93	56.28	63.00	67.80	63.35	67.51	61.60
GAT	63.70	64.17	63.32	64.24	63.84	64.69	64.93	64.47
GraphSAGE	62.28	63.11	62.32	59.41	62.14	64.26	59.97	64.29
HAN	65.76	65.87	65.34	65.45	67.66	67.16	66.82	67.30
Proposed method	68.91	70.12	70.95	70.58	68.44	69.86	70.56	70.63
Proposed method+	**70.51**	**71.37**	**72.03**	**72.17**	70.08	71.47	**71.74**	**71.98**

2 つの対策を行っている．まず，学習に用いる損失として Focal loss [11] を採用する．Focal loss は，画像の物体検知の研究で提案された損失関数であり，対象としているデータの不均衡問題を考慮して提案された損失になる．今回実験に用いる Focal loss のハイパーパラメータ γ は，1 と設定する．次に，実験の評価指標として F1_macro_score を採用する．F1_macro_score は，正常な投稿の識別精度と悪意のある投稿の識別精度の平均で算出され，不均衡なデータセットに対して有効な評価指標とされている．このような対策から，データセットの問題点である不均衡性の影響を軽減し，実験結果の信頼性を高めている．

次に，訓練データとテストデータの分割方法について説明する．実験では，訓練データ 20%，40%，50%，60%の 4 条件でスコアの算出を行っている．ここでの訓練データの割合とは，投稿数の割合のことを指す．つまり，異種グラフ全体のノード数の割合を指定して分割するのではなく，投稿ノード数の割合を指定してグラフデータの分割を行っている．例えば訓練データ 60%では，全投稿 134,950 件の 60%を訓練データ異種グラフの投稿ノードとし，もともとの異種グラフでそれらの投稿ノードとエッジが結ばれていたユーザノード，投稿の要素ノードを，訓練データ異種グラフのノードとして登場させている．テストデータ異種グラフについても，同様の処理を行っている．以上の条件から，悪意のある投稿を検出する学習を行った．なお，実験結果の信頼性を向上させるため，同一条件で 5 回実験を行った平均を実験結果として提示している．

4.3 精度の比較

実験の結果を表 3 に示し，最も精度が高いものを太字で，次に精度が高いものを下線で表す．表 3 を見ればわかるように，提案手法が異種グラフの理解を取り入れることにより，全ての条件下で他の GNN ベースの比較手法を上回る性能を示していることが確認される．条件 1 では，他の GNN と比較して+3〜15%，条件 2 では，他の GNN と比較して+1〜11%のスコア向上が見られた．また，GNN ベースの比較手法の中でも，提案手法に次いで高いスコアを獲得したのが異種 GNN である HAN を採用した手法であり，異種 GNN の優位性を裏付ける結果となっている．さらに，Proposed method+は，条件 2 の低訓練割合 (20%および 40%) を除き，BERT+MLP を含む全ての手法のスコアを上回っている．Proposed method+は，条件 2 における低訓練割合で学習不足に陥り，シンプルな BERT+MLP に劣る結果となったが，他手法と比較し様々な訓練データ割合において一貫して高い性能を示し，限られたデータから効果的に学習できること，および大量のデータを用いた際に過学習に陥りにくいことが示されている．隠れ次元数はモデルが学習可能な特徴の複雑さや量を決定するパラメータであり，次元数を増やすことによる性能向上が計算コストの増大や過学習のリスクと比較して，提案手法は隠れ次元数 64 での運用が最適なバランスであることが示されている．

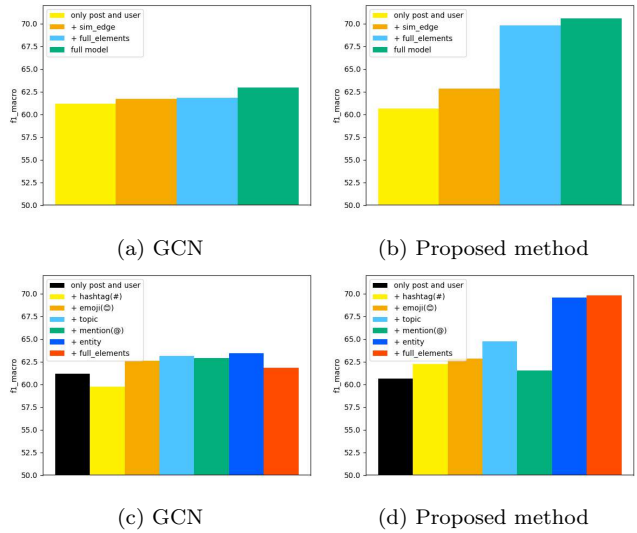
表 4: アブレーション研究

構成要素	F1_macro_score
only user and post	60.64
+ hashtag	62.22
+ emoji	62.85
+ topic	64.73
+ mention	61.50
+ entity	69.57
+ full_elements	69.79
only user and post + sim_edge	62.80
Proposed method	**70.58**

4.4 アブレーション研究

今回提案した異種グラフは，投稿とユーザノードに加え,「投稿の要素」ノード,「拡散ユーザの類似度」エッジなどいくつかの構成要素を統合したグラフになる．グラフにおける各構成要素の有効性を検証するために，特定の要素のみを追加した異種グラフでの学習を行うアブレーション研究を行った．具体的には「ユーザ」と「投稿」のみから構成される異種グラフを構築し，その異種グラフに対し「投稿の要素」ノードと,「拡散ユーザの類似度」エッジのみを加えた異種グラフの構築，また,「投稿の要素」ノードにおける「ハッシュタグ」「絵文字」「トピック」「メンション」「エンティティ」の各ノードを追加した異種グラフの構築を行った．それらを用いたグラフ学習結果から，各要素の重要性の検証を行った．結果を表 4 に示す．

表 4 を見ればわかるように，異種グラフ内の各構成要素の統合が，F1_macro_score の向上に寄与していることがわかる．また,「投稿の要素」ノード内の各要素を追加することがスコア向上に大きく寄与していることがわかる．その中でも特に「エンティティ」ノードの追加や,「トピック」「絵文字」ノードの追加がスコア向上に大きく起因しており，逆に「メンション」や「ハッシュタグ」ノードの追加はほとんど効果が見られない



(a) GCN (b) Proposed method

(c) GCN (d) Proposed method

図 2: アブレーション研究における性能比較

ことがわかった．つまり，投稿を表現した異種グラフを構築する際に重要なノードとして「エンティティ」など文字ベースのものが有効であり,「メンション」などの情報ベースのものは検出には大きく影響を与えないことがわかる．このような結果となった原因のひとつとして考えられるのが，3.3 節で説明したアノテーション方法にある．今回行った実験では，実験の詳細でも記述したようにアノテーションに言語ベースの Perspective API を用いている．先述したように，Perspective API はテキストの有害性を検出するのに有効な手法ではあるが，検出の判断にはテキスト情報のみを使用しており，投稿の関係情報を考慮するという本手法の本質とは異なっているという問題点がある．そのため，テキスト情報のみならず，より詳細に判断されたアノテーションを用いれば，アブレーション研究において「メンション」や「ハッシュタグ」ノードの重要性が向上すると期待される．

また，異種グラフに対して HGT を適用した提案手法と，従来の同種 GNN である GCN を適用した際のアブレーション研究における性能比較を図 2 に示す．

図 2 における (a),(b) は,「投稿の要素」ノードと,「拡散ユーザの類似度」エッジのどちらかのみを追加し学習を行った結果になり，図中の full model が Proposed method となる．図中における (c),(d) は「投稿の要素」ノード内の各要素のうち一つのみを追加し学習を行った結果になり，図中の + full_elements がすべての「投稿の要素」ノードを追加したグラフになる．図 2 より，GCN ではノードタイプおよびエッジタイプを増加させることによる性能向上は限定的であることがわかる．この結果は，Transformer をベースにした HGT を用いる提案手法が，異種グラフの複雑な構造とパターンを効果的に捉え，高精度にコンテキスト化された埋め込みを得られることを示している．

5 ま と め

本論文では，SNS 上の悪意ある投稿を効率的に検出するモデルを提案し，実データを用いた悪意のある投稿を検出するタスクにおいて提案手法の有効性の検証を行った．提案手法で構築した異種グラフには，投稿の意味情報と関係情報を統合するために,「投稿」ノードに加え「投稿の要素」ノードや「拡散ユーザの類似度」エッジなどを追加した．構築した異種グラフに対して HGT を適用することで，グラフの異種性を十分に考慮し，投稿の表現力を最大化した検出手法を提案した．各手法との比較の結果，提案手法は他の従来手法を上回る性能を示し，構築した異種グラフと，適用する GNN の有効性を確認した．将来的には，より大規模なデータセットへの適用や，言語や文化の違いが結果に与える影響を考慮した検討，悪意だけでなくデマ情報や誤った情報を検出するなど，より複雑なタスクへの学習などが期待される．

謝辞 本研究は科研費 (22K18007)，および関西大学 (近未来 ICT の社会実装研究グループ) の助成を受けて行われた．

文 献

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.4171–4186, 2019.

[2] J. Wu, C. Zhang, Z. Liu, E. Zhang, S. Wilson, and C. Zhang, “Graphbert: Bridging graph and text for malicious behavior detection on social media,” Proceedings of the IEEE International Conference on Data Mining, pp.548–557, 2022.

[3] Z. Hu, Y. Dong, K. Wang, and Y. Sun, “Heterogeneous graph transformer,” Proceedings of the International World Wide Web Conference, pp.2704–2710, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, vol.30, pp.6000–6010, 2017.

[5] G. Rajput, N.S. Punn, S.K. Sonbhadra, and S. Agarwal, “Hate speech detection using static bert embeddings,” Proceedings of the International Conference on Big Data Analytics, pp.67–77, 2021.

[6] T.N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” International Conference on Learning Representations, 2017.

[7] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” International Conference on Learning Representations, 2017.

[8] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[9] Tohoku NLP Group, “Bert japanese pretrained model”. `https://github.com/cl-tohoku/bert-japanese`.

[10] 鳥海不二夫，吉田光男，榊 剛史，“ソーシャルメディア上の情報拡散とその理解，” 電子情報通信学会誌，vol.105，no.4，pp.326–332，2022.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.42, no.2, pp.318–327, 2020.

[12] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” Advances in Neural Information Processing Systems, vol.30, 2017.

[13] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P.S. Yu, “Heterogeneous graph attention network,” Proceedings of the International World Wide Web Conference, pp.2022–2032, 2019.

隣接ノードの情報を利用する ランダムウォークベースのグラフ特徴量推定

長谷川　毅†　　廣中　詩織†　　首藤　一幸†

† 京都大学

あらまし　大規模かつ未知なソーシャルネットワークの特徴量を推定するためには，ランダムウォークを用いたサンプリングが有用であり，多くのアルゴリズムが提案されてきた．実際のソーシャルネットワーク上のサンプリングでは，隣接ノードを取得できる API によるクエリを繰り返し使用することになるが，単位時間あたりに使用できるクエリ数の制限や，API の料金の観点から，できるだけ少ない API の呼び出し回数で特徴量を推定することが望ましい．本研究では，隣接ノードを取得する行為をコストと考えたうえで，既存アルゴリズムよりも精度の良い特徴量推定アルゴリズムを提案する．具体的には，隣接ノードの一覧とともに得られる，隣接ノードについての情報，例えば次数情報を活用する．実験では，既知のグラフに対してサンプリングのシミュレーションを行い，提案アルゴリズムが既存アルゴリズムよりも高い精度で特徴量を推定できることを示す．

キーワード　ソーシャルネットワーク，グラフサンプリング，ランダムウォーク

1　はじめに

Online Social Networks (OSNs) をノードとエッジの集合のグラフ構造と捉え，解析することは重要な問題であり，盛んに研究が行われてきた [1–4]．しかし X[1]をはじめとする一般的な OSN はそのデータへのアクセスを制限しており [5–7]，グラフ全体を取得し，解析することは不可能に近い．ゆえに OSN の一部分をサンプリングすることでグラフの特徴を推定することが現実的な手段である．

サンプリングによって OSN の特徴を推定するためにはランダムウォークを用いたサンプリングが有用であり，多くのランダムウォークを用いた特徴量推定アルゴリズムが提案されている [8–15]．多くの OSN はユーザのフォロイーまたはフォロワーのリストの情報，つまり隣接ノードの情報が取得できるアプリケーション・プログラミング・インタフェース（API）を提供しており [5–7]，API から取得した隣接ノードからランダムに一つ選択し遷移することを繰り返すことで，OSN 上でのランダムウォークサンプリングが可能である．またランダムウォークが持つマルコフ性を利用することで，ランダムウォークで取得したサンプル列に適切な重みづけを計算し，OSN の不偏な推定量を求めることができる [3, 9, 16]．ここでノード ID に基づく一様独立サンプリングは，一般にノード ID の分布を知ることができないため適用が難しく [17]，また幅優先サンプリングなどの走査的手法 [18] は，取得したサンプル列の偏りが未知のため不偏な特徴量を得ることができない．

一般的な OSN が提供する API は単位時間あたりに使用できるクエリ数の制限がある．また X など API の使用自体が有料化された OSN も存在する．よって時間的，料金的な観点から，できるだけ少ない API の呼び出し回数で OSN の特徴量を推定することが重要である．岩崎らは API の呼び出し回数をコストとして考え，既存のランダムウォークに基づく特徴量推定アルゴリズムを比較した [19]．我々は，多くの API では，隣接ノードのリストを取得する際に隣接ノードの次数を同時に知ることができることに着目し，この情報を活用してより効率的なアルゴリズムを構築する．提案アルゴリズムでは，次数に限らず，隣接ノードのリストを取得する際に同時に取得できるあらゆる特徴に対し，OSN での不偏特徴量を推定することができる．

本研究では，隣接ノードの情報を利用するランダムウォークによる OSN の特徴量推定アルゴリズムを提案し，シミュレーション実験により提案アルゴリズムが既存アルゴリズムよりも高い精度で OSN の特徴量を推定できることを示す．提案アルゴリズム出は，隣接ノードのリストを取得する際に，同時に取得したそれぞれの隣接ノードの情報を OSN の不偏特徴量推定に活用する．

2　準　　備

2.1　記　　法

本研究では，ソーシャルグラフを有向グラフ $G = (V, E)$ として表す．$V = \{v_1, v_2, ...v_n\}$ はノード（ユーザ）の集合であり，グラフ全体のノード数を $n = |V|$ とする．E は有向エッジ（フォロー関係）の集合である．また各エッジ (v_i, v_j) に対して逆向きのエッジ (v_j, v_i) を追加したエッジの集合を E' とする．有向エッジ (v_i, v_j) が存在するとき，ノード v_j をノード v_i のフレンド，ノード v_i をノード v_j のフォロワーと呼ぶ．ノード $v_i \in V$ のフレンドの集合を $N_{\text{out}}(v_i) = \{v_j \in V : (v_i, v_j) \in E\}$，フォロワーの集合を $N_{\text{in}}(v_i) = \{v_j \in V : (v_j, v_i) \in E\}$ と定義し，$N(v_i) = N_{\text{out}}(v_i) \cup N_{\text{in}}(v_i)$ とする．次にそれぞれのノードの出次数，入次数をそれぞれ $d_{\text{out}}(v_i) = |N_{\text{out}}(v_i)|$, $d_{\text{in}}(v_i) = |N_{\text{in}}(v_i)|$ とする．また出次数，入次数の合計を $d_{\text{sum}}(v_i) = d_{\text{in}}(v_i) + d_{\text{out}}(v_i)$

This is unrefereed paper.

1：https://twitter.com

とし，$d_{\text{in-out}}(v_i) = |N_{\text{out}}(v_i) \cap N_{\text{in}}(v_i)|$ とする．

ノード v_i のプロパティを $a(v_i)$ とする．プロパティ $a(v_i)$ の例としては，ノード v_i の次数や，投稿数，ボットラベルなどの二値ラベルが挙げられる．

2.2 モ デ ル

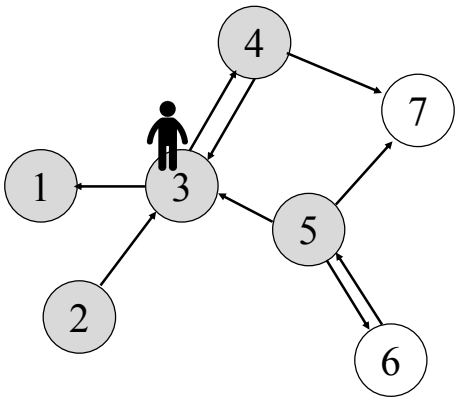


図 1: API が提供する隣接ノードの情報のモデル．人がいるノードが隣接ノードを取得したノード，灰色のノードは次数の情報を取得できるノードを示す．

本研究では，図 1 のように，隣接ノードの情報を取得する際，取得したそれぞれの隣接ノードの次数（フレンド数，フォロワー数）を知ることができる API を対象とする．つまりノード v_i の隣接ノードのリスト $N(v_i)$ を取得した時，同時に $N(v_i)$ に含まれるそれぞれのノードの出次数，入次数が取得できることを前提とする．実際の OSN では，X と Mastodon がこのモデルと一致する API を提供している [5,7]，

また本研究では，隣接ノードの取得回数をコストとし，1 回の隣接ノードの取得で $N_{\text{in}}(v_i)$ と $N_{\text{out}}(v_i)$ の取得が可能であるとする．つまり隣接ノードの数に依らず，コスト 1 ですべての隣接ノードを取得できることとする．

加えてグラフ G は弱連結であり，ランダムウォーク中は静的であるとする．また，一度ノードに遷移すると，遷移したノードのプ次数，プロパティ，取得したそれぞれの隣接ノードの次数，プロパティを保存することとする．

2.3 マルコフ連鎖の基礎

マルコフ連鎖の基礎を述べる．ランダムウォークにより得られたサンプル列に対し，OSN での不偏特徴量を推定するためには，マルコフ連鎖の定常分布に基づく偏りを是正する必要がある．$\mathbf{P} = \{P_{i,j}\}_{i,j \in S}$ を状態空間 S 上のマルコフ連鎖の遷移確率行列とすると，$\mathbf{P}$ の定常分布に対して以下の定理が成り立つ．

定理 1. 分布 $\boldsymbol{\pi} = \{\pi_i\}_{i \in S}$ について，$\pi_j = \sum_{i \in S} \pi_i P_{i,j}$ を満たすならば，分布 $\boldsymbol{\pi}$ は確率遷移行列 $\mathbf{P}$ に基づくマルコフ連鎖の定常分布である．

次に定常分布 $\boldsymbol{\pi}$ に対して定理 2 [20,21] が成り立つ．

定理 2. $\mathbf{P}$ に基づくマルコフ連鎖がエルゴード的（周期性を持たない，任意の状態から任意の状態に到達可能，状態数が有限）であるとき，定常分布 $\boldsymbol{\pi}$ が一意に存在する．

マルコフ連鎖に基づくサンプル列が，推定したい特徴量に収束することを示すために必要な，大数の強法則に基づく以下の定理 3 [11,16] を引用する．

定理 3. $\{X_s\}$ を有限な状態空間 S 状の，エルゴード的なマルコフ連鎖であるとし，$\boldsymbol{\pi}$ をその定常分布とすると，任意の関数 $f : S \to \mathbb{R}$ に対して，$t \to \infty$ のとき，初期状態に関わらず以下が成立する．

$$\frac{1}{t}\sum_{s=1}^{t} f(X_s) \to \sum_{i \in S} \pi_i f(i) \text{ almost surely (a.s.)}$$

2.4 ランダムウォークサンプリング

本研究では，ランダムウォークをベースとした独自サンプリング手法と OSNs の不偏特徴量の推定プロセスを提案する．本節では単純なランダムウォーク (Simple random walk (SRW)) について紹介する．

SRW サンプリングは，適当に選択した初期ノードから隣接ノードの 1 つをランダムに選択し遷移してサンプリングする手法である．SRW のノード v_i からノード v_j への遷移確率を P_{v_i,v_j} とすると，

$$P_{i,j} = \begin{cases} \frac{1}{d_{\text{sum}}(v_i)} & \text{if } v_j \in N(v_i) \\ 0 & \text{otherwise.} \end{cases}$$

と表せる．

ランダムウォークサンプリングでは各ノードの訪問確率は数学的に示されている．具体的には，SRW の t ステップ後の各ノードへの訪問確率の分布 $\boldsymbol{\pi}$ は，事象 A が起きる確率を $Pr[A]$ とすると，$\boldsymbol{\pi} = (Pr[x_t = 1], Pr[x_t = 2], ..., Pr[x_t = n])$ と表すことができる．ここで，$\boldsymbol{\pi}$ の i 番目の要素を $\pi^{(i)}$ とすると，$\pi^{(i)}$ は $d_{\text{sum}}(v_i)/2|E|$ に収束することが知られている [22]．よって SRW において各ノードへの訪問確率は次数に比例する．SRW が持つこの特性を利用することで，グラフ全体の不偏特徴量を得ることができる [3]．

多くのランダムウォークベースのサンプリング手法の研究では，サンプル列のサイズを用いて精度の評価を行った [3,8,16]．しかし本研究では，隣接ノードのリストを取得する行為をコストとして考える．これは岩崎ら [19] と同じアプローチであり，実際の OSNs でのサンプリングでは，隣接ノードのリストを取得する行為がボトルネックとなる場合があることが採用理由である．また 2.2 節で定義したとおり，隣接ノードの取得はコスト 1 で行えることとし，取得した隣接ノードのリストに含まれるそれぞれのノードの出次数，入次数，プロパティが取得できるものとする．これらの情報は SRW では使用しない．

3 提 案 手 法

本章では，隣接ノードの情報を取得する際に，取得したそれ

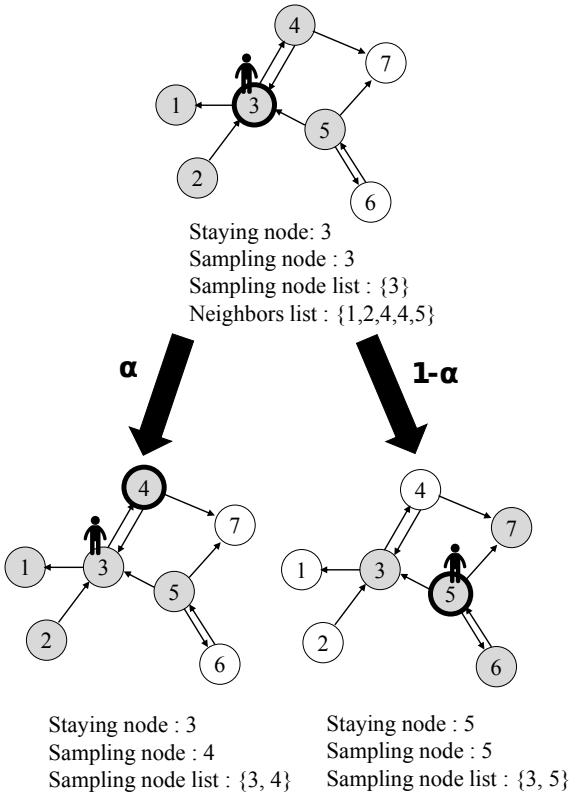


図 2: 提案サンプリングアルゴリズムの概要

Algorithm 1 提案サンプリングアルゴリズム

Input: n_0 : 初期ノード, α : パラメータ $(0 \leqq \alpha < 1)$, b : クエリ数
Output: *sanple_list* : サンプル列

$v_i \leftarrow n_0$
sample_list $\leftarrow \{v_i\}$
stay_node_list $\leftarrow \{v_i\}$
query_num $\leftarrow 1$
while *query_num* $< b$ **do**
  neighbors ←A list obtained by concatenating $N_{\text{out}}(v_i)$ and $N_{\text{in}}(v_i)$
  $p \leftarrow$ Random number generated between 0 and 1
  while $p < \alpha$ **do**
    $v_j \leftarrow$ A node uniformly selected at random from *neighbors*
    Append v_j to *sample_list*
    $p \leftarrow$ Random number generated between 0 and 1
  end while
  $v_i \leftarrow$ A node uniformly selected at random from *neighbors*
  if $v_i \notin$ *stay_node_list* **then**
    query_num $\leftarrow$ *query_num* $+ 1$
  end if
  Append v_i to *sample_list*
  Append v_i to *stay_node_list*
end while
return *sample_list*

ぞれの隣接ノードの情報を利用するランダムウォークベースのサンプリングアルゴリズムを提案する．提案サンプリングアルゴリズムにおけるマルコフ連鎖について議論することで，それらの特徴量の推定アルゴリズムを示す．

本研究で提案する手法で推定可能な OSN 上の特徴量は，隣接ノードを取得する際に同時に取得できるプロパティから得ることができる特徴量である．具体的には，ノード v_i の隣接ノードのリスト $N(v_i)$ を取得したとき，$N(v_i)$ に含まれる任意のノード v_j に対し，プロパティ $a(v_j)$ が得られるとすると，提案アルゴリズムはプロパティ $a(v), (v \in V)$ のグラフ全体での平均値や分布を推定することができる．

3.1 隣接ノードを確率的にサンプル列に加えるランダムウォーク

提案するサンプリングアルゴリズムの概要を図 2，詳細を Algorithm 1 に示す．提案サンプリングアルゴリズムでは，基本的な遷移方法は通常のランダムウォークと同様であるが，ノード v_i に遷移しサンプル列に加え，隣接ノードリストを取得した後，確率 α で遷移せずに取得した隣接ノードからランダムに選んだノードをサンプル列に加えるという点で異なる．α は $0 \leqq \alpha < 1$ を満たすパラメータであり，ノード v_i からノード v_j への遷移は確率 $1-\alpha$ で行われる．本論文では，遷移した先のノードを滞在ノード，サンプル列に加えられたノードをサンプリングノード，確率 α で行われる操作を隣接ノードサンプリング，確率 $1-\alpha$ で行われる操作を遷移サンプリングと呼ぶ．隣接ノードサンプリングでは，サンプリングノードは滞在ノードの隣接ノードからランダムに選ばれ，遷移サンプリングでは，更新された滞在ノードとサンプリングノードが等しくなる．

本研究は OSN を有向グラフとして扱うが，遷移やサンプリングは逆向きのエッジに対しても可能であるとする．これは実際の OSN でユーザ間を遷移する際，フォロワーとフレンドの両方から遷移するユーザを決めることができることに起因する．そのためノード v_i の隣接ノードのリストは $N_{\text{in}}(v_i)$ と $N_{\text{out}}(v_i)$ を結合したものである．このときノード v_i の隣接ノードのリストは重複を許す配列とする．つまりノード v_j について，$v_j \in N_{\text{out}}(v_i) \cap N_{\text{in}}(v_i)$ ならば，ノード v_i の隣接ノードのリストにノード v_j は 2 つ存在することになる．遷移もしくはサンプリングするノードは隣接ノードのリストから一様ランダムに選ばれるため，隣接ノードのリストに 2 つ存在するノードは，1 つしか存在しないノードの 2 倍の確率で選ばれることとなる．

隣接ノードサンプリングはコストがかからない．これは本研究が隣接ノードの取得回数をコストと考えているためである．隣接ノードサンプリングは既に取得している隣接ノードリストからランダムにサンプル列に加える操作であるため，新たにコストは発生しない．逆に遷移サンプリングは隣接ノードのリストからランダムに選び，遷移先のノードの隣接ノードを取得する必要があるためコストが発生する．ただし，本研究では，一度取得した隣接ノードの情報を保存するものとしているため，一度遷移サンプリングされたノードが再び遷移サンプリングさ

れた場合はコストが発生しない．よって隣接ノードの取得回数は，新規の遷移サンプリング数と一致する．

新規の遷移サンプリングにより，隣接ノードのリストを取得する回数をクエリ数と呼ぶ．提案サンプリングアルゴリズムは，隣接ノードの取得回数が，与えられたクエリ数 b に達した際に停止する．クエリ数 b は OSN の API の仕様をもとに，時間的に許容できるクエリ数を設定すればよい．ただしクエリ数 b は大きければ大きいほど推定精度が向上する傾向がある．クエリ数と推定精度の関係については 4 章で実験により示す．

3.2 提案サンプリングアルゴリズムにおけるマルコフ連鎖

提案サンプリングアルゴリズムにおけるマルコフ連鎖について議論するために，以下のような状態空間を定義する．

定義 4.

$$\Omega \stackrel{\text{def}}{=} \{(v_i, v_j) : v_i, v_j \in V \text{ s.t.}((v_i, v_j) \in E') \text{ or } (v_i = v_j)\}$$

t ステップ目の滞在ノードが X_t，サンプリングノードが X'_t であるとき，$Z_t \stackrel{\text{def}}{=} (X_t, X'_t) \in \Omega$ とすると，提案サンプリングアルゴリズムは状態空間 Ω 上のマルコフ連鎖 $\{Z_t \in \Omega : t = 1, 2, ...\}$ であるといえる．簡単に表記するために e_{ij} は状態 $(v_i, v_j) \in \Omega$ を表すこととする．つまり提案手法においては，e_{ij} は $i \neq j$ のとき隣接ノードサンプリングを表し，滞在ノードが v_i，サンプリングノードが v_j となり，$i = j$ のとき遷移サンプリングを表し，滞在ノード，サンプリングノードともに $v_i (= v_j)$ である．

$\mathbf{P} \stackrel{\text{def}}{=} \{p(e_{ij}, e_{lk}) : e_{ij}, e_{lk} \in \Omega\}$ を状態空間 Ω 状のマルコフ連鎖 $\{Z_t \in \Omega : t = 1, 2, ...\}$ の遷移確率行列とすると，以下の定理が成り立つ．

定理 5. $\mathbf{P}$ の定常分布 $\boldsymbol{\pi}$ が一意に存在する．

証明． まず，$|V| < \infty$ より，$|\Omega| < \infty$ である．次に本研究では，対象のグラフを弱連結有向グラフと仮定しているが，逆向きのエッジに対しても遷移可能であるため，任意のノードに到達可能であり，その任意の隣接ノードをサンプリング可能である．このことから Ω 内の任意の状態から任意の状態に到達可能であり，同様の理由から周期性を持たない．よって，$\mathbf{P}$ はエルゴード的であるので，定理 2 より，定常分布 $\boldsymbol{\pi}$ は一意に存在する． □

次に状態 e_{ij} に対して次のような関数 m を定義する．

定義 6.

$$m(e_{ij}) \stackrel{\text{def}}{=} \mathbb{1}_{v_j \in N_{\text{out}}(v_i)} + \mathbb{1}_{v_j \in N_{\text{in}}(v_i)}$$

関数 m はノード v_i とノード v_j 間のエッジの数を示し，$m(e_{ij}) = m(e_{ji})$ である．

これにより，提案サンプリングアルゴリズムにおいて，ノード v_i に滞在しているとき，隣接ノード v_j を隣接ノードサンプリングする確率 $p_{neighbor}$ は以下のように与えられる．

定義 7.

$$p_{neighbor}(v_j; v_i) = \alpha \cdot \frac{m(e_{ij})}{d_{\text{sum}}(v_i)}$$

同様に，ノード v_i に滞在しているとき，隣接ノード v_j を遷移サンプリングする確率 p_{walk} は以下のように与えられる．

定義 8.

$$p_{walk}(v_j; v_i) = (1 - \alpha) \cdot \frac{m(e_{ij})}{d_{\text{sum}}(v_i)}$$

提案サンプリングアルゴリズムの遷移確率行列 $\mathbf{P}$ は以下のように与えられる．

定義 9.

$$P(e_{ij}, e_{lk}) = \begin{cases} p_{neighbor}(v_k; v_i) & i = l \\ p_{walk}(v_k; v_i) & l = k, v_l \in N(v_i) \\ 0 & \text{otherwise.} \end{cases}$$

このとき，提案サンプリングアルゴリズムのマルコフ連鎖 $\{Z_t : t = 1, 2, ...\}$ の定常分布について以下の定理が成り立つ．

定理 10. 状態空間 Ω 状のマルコフ連鎖 $\{Z_t : t = 1, 2, ...\}$ の遷移確率行列が定義 9 のように与えられたとき，そのマルコフ連鎖の定常分布 $\boldsymbol{\pi}$ は，以下のように与えられる．

$$\pi(e_{ij}) = \begin{cases} \alpha \cdot \frac{m(e_{ij})}{2|E|} & i \neq j \\ (1 - \alpha) \cdot \frac{d_{\text{sum}}(v_i)}{2|E|} & i = j. \end{cases}$$

証明． 付録に示す． □

3.3 特徴量推定

提案サンプリングアルゴリズム上での特徴量推定のアルゴリズムを示す．本節の目標は，提案遷移サンプリングアルゴリズムによって得られたサンプル列から，OSN 上での特徴量の期待値を推定するための重みづけプロセスを得ることである．

f を任意の関数 $f : V \to \mathbb{R}$ とし，一様分布を $\mathbf{u} \stackrel{\text{def}}{=} [u(1), u(2), ..., u(n)] = [1/n, 1/n, ..., 1/n]$ とすると，推定したい特徴量の OSN での期待値 $\mathbb{E}_u(f)$ は以下のように与えられる．

定義 11. $\mathbb{E}_u(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{v \in V} f(v)$

関数 f を適切に設定することで，推定したい特徴量の OSN での期待値を得ることができる．例えば，グラフ G の出次数分布 $\mathbb{P}\{D_G = d\}$, $(d = 1, 2, ..., n - 1)$ を推定したい場合は，$f(v) = \mathbb{1}_{d_{\text{out}}(v) = d}$ となる関数 f を選べばよい．

提案サンプリングアルゴリズムによって得られるサンプル列 $\{Z_s\}_{s=1}^t$ に対し，$\mathbb{E}_u(f)$ を得るための再度重み付けプロセスを定義するために，関数 f に対して新たに別の関数 $g : \Omega \to \mathbb{R}$ を次のように定義する．

定義 12. $g(e_{ij}) \stackrel{\text{def}}{=} f(v_j)$

関数 g は，サンプル列 $\{Z_s\}_{s=1}^t$ に対し，サンプリングノードに関数 f を適用するものである．

次に以下のような推定量を定義する．

定義 13. $\hat{\mu}_t(g) \stackrel{\text{def}}{=} \frac{1}{t} \sum_{s=1}^t g(Z_s)$

表 1: データセットの概要

Network	種類	ノード数	エッジ数
ego-Twitter	ソーシャルネットワーク	81306	1768149
soc-Slashdot	ソーシャルネットワーク	82168	948464
Amazon	商品ネットワーク	262111	1234877
DBA モデル	生成ネットワーク	100000	1000000

次に以下のような重みづけ関数を定義する．

定義 14. $w(e_{ij}) \stackrel{\text{def}}{=} \frac{1}{d_{\text{sum}}(v_j)}$

本研究では，サンプリングノードのフレンド数 $d_{\text{out}}(v_j)$，フォロワー数 $d_{\text{in}}(v_j)$ は取得可能であるとしているので，サンプリングノード v_j に対し，$d_{\text{sum}}(v_j)$ を得ることができる．

ここで以下の定理が成り立つ．

定理 15. 提案遷移アルゴリズムによって得られるサンプル列 $\{Z_s\}_{s=1}^t$ に対して，$t \to \infty$ のとき，

$$\frac{\sum_{s=1}^t w(Z_s)g(Z_s)}{\sum_{s=1}^t w(Z_s)} \to \mathbb{E}_u(f) \text{ a.s.}$$

である．

証明. 付録に示す. □

定理 15 より，提案サンプリングアルゴリズムによって得られるサンプル列 $\{Z_s\}_{s=1}^t$ に対し，推定量 $\sum_{s=1}^t w(Z_s)g(Z_s)/\sum_{s=1}^t w(Z_s)$ は，OSN での推定したい特徴量の期待値に収束する．またサンプル Z_t にはサンプリングノード X_t' の情報の他に滞在ノード X_t の情報も含まれるが，特徴量の推定においてはサンプリングノード X_t' の情報のみ保存すればよいこともわかる．このため Algorithm 1 においては，滞在ノード X_t の情報を返していない．

提案手法は，隣接ノードサンプリングの際に取得できる任意のプロパティに関する不偏特徴量を推定することができる．これは，隣接ノードサンプリングによりサンプル列に加えられたサンプル e_{ij} の推定可能な特徴量 $g(e_{ij})$ は，隣接ノードのリストを取得した際に同時に得ることができる，隣接ノード v_j のプロパティ $a(v_j)$ に依存するためである．隣接ノードの取得時に同時に取得できるプロパティに何が含まれるかは OSN の API の仕様によって異なるが，次数を取得できる X [5] においては，平均次数や次数分布，ボットラベルや投稿数も取得可能な Mastodon [7] においては，ボット率や平均投稿数も推定することができる．

以上より，提案サンプリングアルゴリズムによって得られるサンプル列に対し，OSN 上での特徴量の期待値を推定するための重みづけプロセスを得ることができた．

4 実　　験

複数のネットワークに対して提案手法の精度を評価する．実際の OSN でのサンプリングでは，対象となるグラフは未知であるが，ここでは既知のグラフに対してサンプリングのシミュレーションを行う．

4.1 実 験 準 備

データセット: Stanford Large Network Dataset Collection [23] で公開されている 3 つのデータセットに加え，複雑ネットワークの生成モデルの一つである，Directed-Barabasi-Albert モデル（DBA モデル）[24] で生成したグラフを用いて実験を行った．対象とするグラフは有向グラフで，最大弱連結成分を扱った．DBA モデルは Barabasi-Albert モデル [24] を有向グラフに拡張したもので，以下の確率であるノードからあるノードに有向エッジが追加される．

$$\prod(d_{\text{in}}(v_i)) = \frac{d_{\text{in}}(v_i) + A}{\Sigma_j(d_{\text{in}}(v_j) + A)}$$

パラメータ A は 1 を選択した．表 1 は各データセットの概要を示したものである．

シミュレーション: ランダムウォークの初期ノードはグラフ上からランダムに選択する．また実験ごとに独立に選択する．すべてのサンプリングシミュレーションは 1000 回独立に行う．クエリ数はグラフの全ノード数に対する割合を実験に応じて選択する．

クエリ数 b: クエリ数はグラフ中の全ノード数の 0.5%から 5%まで 0.25%刻みで変化させる．

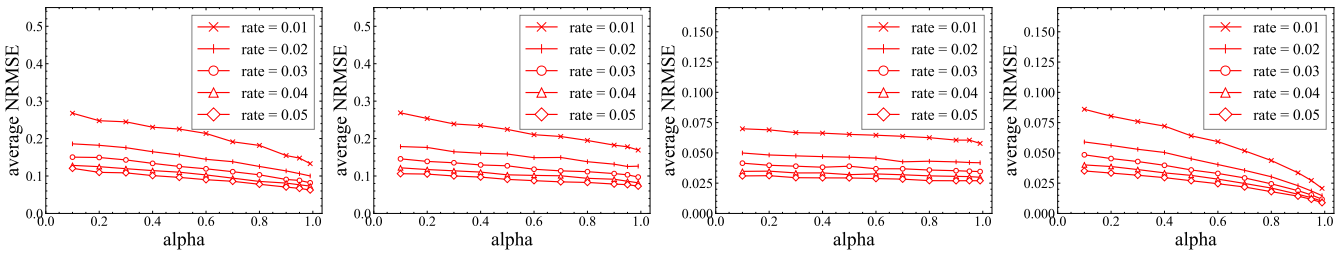
評価指標: 各特徴量推定の評価指標は，正規化平方二乗誤差（Normalized Root Mean Square Error，NRMSE）を用いる．NRMSE は推定値の誤差と分散を評価できる指標であるため値が小さいほど望ましく，多くの関連する研究で用いられている [8,12,16,19]．NRMSE は x を特徴量の真値，$\hat{x_i}$ を i 回目のサンプリングシミュレーションの特徴量の推定値，N をシミュレーション回数とすると，以下のように計算できる．

$$\text{NRMSE} = \frac{1}{x}\sqrt{\frac{1}{n}\sum_{i=1}^{N}(x - \hat{x_i})^2}.$$

特徴量: 実験対象とする特徴量は，平均出次数，ランダムラベル率，高次数ラベル率，低次数ラベル率である．ランダムラベル，高次数ラベル，低次数ラベルは合成的にデータセットに付与した 2 値のラベルで，実際の OSN でのボットラベルなどを想定している．ランダムラベルとは，ランダムに付与されるラベルである．高次数ラベルとは，$\hat{d}(v_i) = \max(d_{\text{in}}(v_i), d_{\text{out}}(v_i))$ としたとき，確率 $\hat{d}(v_i)/\Sigma_{v_i \in V}\hat{d}(v_i)$ で付与されるラベル，低次数ラベルとは，確率 $(1/\hat{d}(v_i))/\Sigma_{v_i \in V}(1/\hat{d}(v_i))$ で付与されるラベルである．各ラベルは，ラベル付けされたノードの割合が全体の 10% に達するまで，対応する確率から選ばれたノードを繰り返しラベル付けする．これらのラベルにより，実際の OSN において，ランダムに付与されるようなラベル，高次数のノードに多く付与されるラベル，低次数のノードに多く付与されるラベルといった特徴を持つラベルが付与されたノードの割合の推定のシミュレーションを行うことができる．ラベルの付与の方法は福田ら [25] による方法を用いた．

4.2 α と推定精度の関係

提案手法において，隣接ノードサンプリングを行う確率であ



(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 3: 各クエリ割合における α による各特徴量の NRMSE の平均の変化

る α を変化させ実験を行い，推定精度との関係を調べた．各データセットに対して，α を 0.1 から 0.9 まで 0.1 刻みに変化させ実験を行った．加えて $\alpha = 0.95$ と 0.99 でも実験した．クエリ数は全ノード数に対して，1%から 1%刻みで変化させた．図 3 は，各 α における，4.1 節で述べた各特徴量の NRMSE の平均をプロットしたものである．クエリ数は全ノードの 1%, 2%, 3%, 4%, 5% の場合の結果を示した．図 3 より，すべてのグラフ，すべてのクエリ数において，α が 1 に近づくにつれて NRMSE が小さくなることがわかる．

4.3 既存手法との比較

提案手法と既存手法の各特徴量の推定精度の比較を行う．比較する既存手法は，代表的なランダムウォークベースの特徴量推定手法である Simple Random Walk with Re-weighting (SRW-rw) [3, 26],Non-backtracking Random Walk with Re-weighting (NBRW-rw) [16],Metropolis-Hastings Random Walk (MHRW) [3, 26] である．これらの既存手法は，本研究と同じコスト設定を行っている岩崎ら [19] が比較した手法と同一である．岩崎らの実験では，隣接ノードの取得をコストとして考えると，NBRW-rw が優れた推定精度を示したが，グラフや推定する特徴量によっては，評価が逆転する可能性も示唆した．

SRW-rw は通常のランダムウォークにより得たサンプル列に対し，重み付けをして特徴量を推定する手法であり，NBRW-rw は一つ前にサンプリングしたノードには遷移しないランダムウォークにより得たサンプル列に対し，重み付けをして特徴量を推定する手法である．MHRW は各ノードの次数に基づき遷移確率を調整することで，定常分布が一様分布になるため，サンプル列の特徴量の期待値がそのまま不偏特徴量となる手法である．本研究では，クエリ数，つまり隣接ノードの情報を取得した回数をそろえて特徴量の推定精度を比較した．

図 4 から図 7 は，各特徴量の推定値の NRMSE について，提案手法のパラメータ $\alpha = 0.5, 0.9$ と，既存手法を比較したものである．横軸は全体のノード数に対する，クエリ数の割合である．クエリ数は，全ノード数に対して，0.25%から 0.25% 刻みで変化させた．DBA モデルの平均出次数の推定結果のみ，MHRW の精度が他と大きく乖離し悪かったため，図から省いている．図 4 から図 7 より，すべてのグラフ，すべての特徴量において，$\alpha = 0.9$ の場合の提案手法は既存手法よりも同等か，精度がよい．また図 3 で示した通り，提案手法の精度は α が 1 に近づくにつれて精度が良くなるが，$\alpha = 0.5$ の場合でも多くのグラフ，特徴量で既存手法よりも優れていることがわかる．またクエリ数 b が大きければ大きいほど推定精度が向上することがわかる．

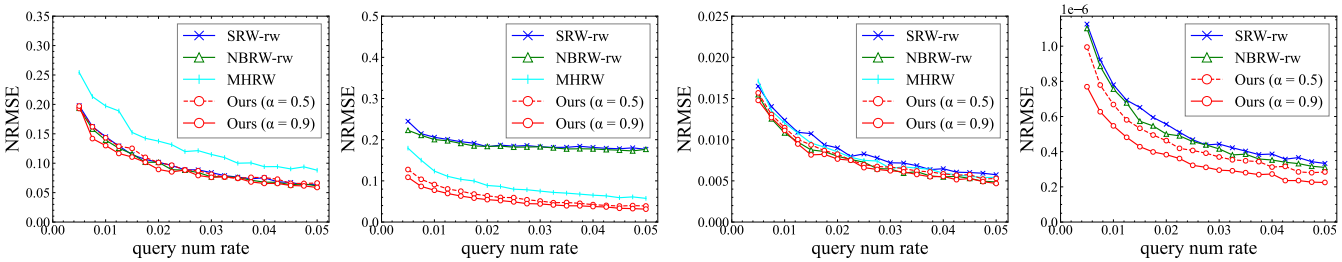
4.4 考　　察

提案手法では，より多くのサンプルを推定に利用できることは推定精度の向上に繋がると考えられる．隣接ノードの取得をコストとして考えると，同コストでの遷移サンプリングの回数は α に依らず一定である．またこの回数は，同コストでの SRW における遷移回数と等しくなる．提案手法では，α を 1 に近づけるにつれて推定に利用できるサンプル数を増やすことが可能である．これにより，図 3 に示したように，α が 1 に近づけるにつれて推定精度が向上したと考えられる．図 8 は提案手法の $\alpha = 0.5, 0.9$ の場合と，既存手法の，クエリ数に対する平均サンプルサイズを示したものである．提案手法は $\alpha = 0.9$ とすると，同コストでの他の手法に比べ最も大きなサンプルサイズとなった．これが図 4-7 に示したように，提案手法が既存手法よりも推定精度が優れていた要因であると考えられる．

サンプルサイズの増大は実用上のボトルネックとはならない．本提案手法で求める推定量 $\sum_{s=1}^{t} w(Z_s)g(Z_s)/\sum_{s=1}^{t} w(Z_s)$ (3.3 章) はサンプリングを行いながら逐次的に計算できる．このことはサンプル列全体を保存しておく必要がないことを意味する．本研究において保存しておく必要がある情報は，2.2 章で述べた通り，一度遷移したノードとその隣接ノードの情報である．したがって，保存しておく情報は隣接ノードサンプリングによるサンプルサイズによって変化しない．

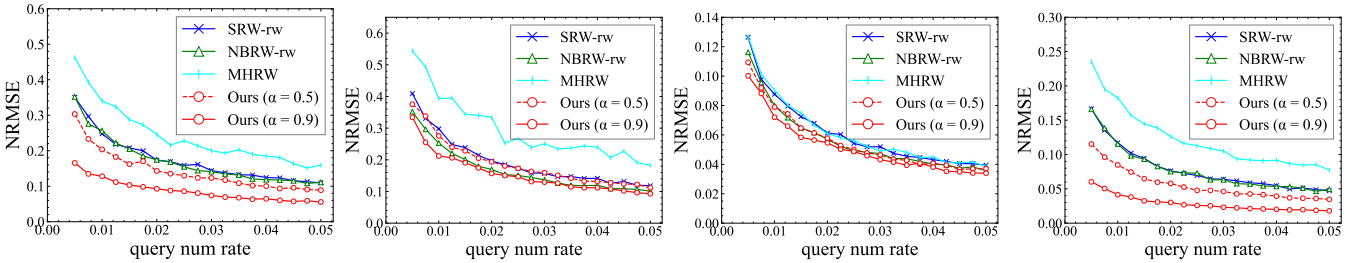
5 関 連 研 究

グラフサンプリングに関する関連研究について述べる．Gjoka ら [3] は，ランダムウォークの定常分布から再度重みづけをすることで不偏な推定値を得る SRW-rw と MHRW との比較を行い，SRW-rw が精度面で優位であることを示した．Lee ら [16] は，1 つ前に遷移したノードへの後戻りをしない NBRW を提案し，NBRW-rw による不偏推定の方法と，NBRW-rw が SRW-rw よりも優位であることを理論面と実験で示した．Ribeiro ら [27] は複数の連結成分が存在する場合に推定精度が向上す



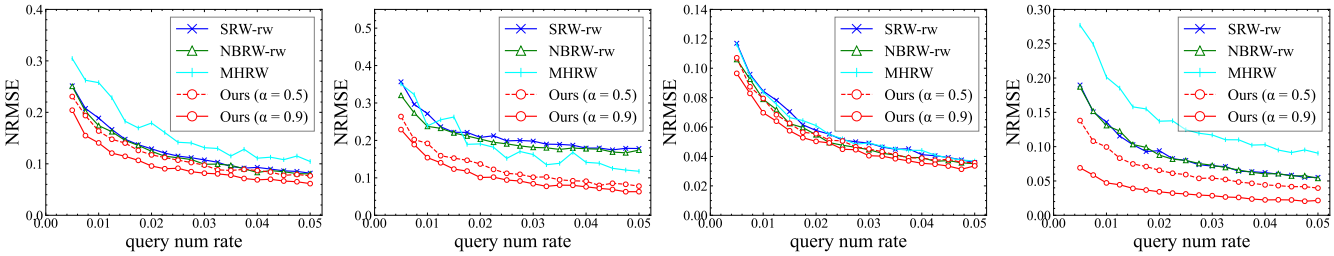
(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 4: 平均出次数の推定値の NRMSE



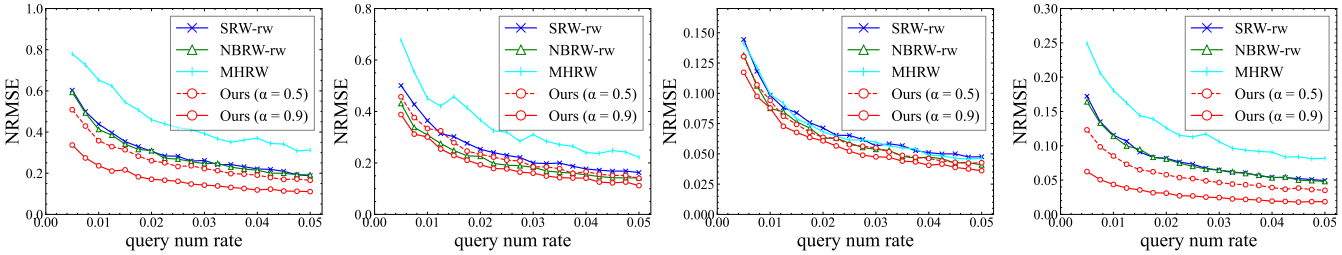
(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 5: ランダムラベルの推定値の NRMSE



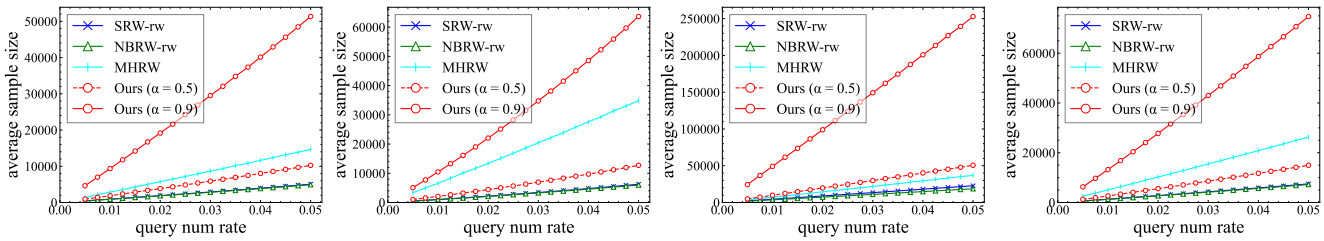
(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 6: 高次数ラベルの推定値の NRMSE



(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 7: 低次数ラベルの推定値の NRMSE



(a) ego-Twitter (b) Slashdot (c) Amazon (d) DBA モデル

図 8: 各クエリ数の割合におけるサンプル列のサイズの平均値

る Multidimensional Random Walk を提案した．本研究では対象とするグラフは単一の弱連結の連結成分からなるものとしているが，実世界で応用する際は，Multidimensional Random Walk の考え方で精度が向上する可能性は考えられる．岩崎ら [19] はクエリ数でサンプリングアルゴリズムを比較することを提案し，特徴量によっては SRW-rw と NBRW-rw の精度の評価が逆転する可能性を示した．

次に隣接ノードの情報を利用するサンプリングに関する関連研究について述べる．Han ら [28] はランダムウォークを行う際，推定するモチーフに応じて隣接ノードをサンプリングしていく手法を提案した．推定するモチーフにより隣接ノードの情報を取得するかを決定するため，隣接ノードの情報を取得する確率，深さが本研究と異なる．またサンプリングしたノード数をコストとしているため，API のクエリ数をコストとし，隣接ノードから得られる任意の特徴量を推定する本研究とは異なる．Illenberger ら [29] は，アクセスしたノードの隣接ノードの情報を集めていくスノーボールサンプリングのサンプル列に補正をかけることで特徴量を推定する手法を提案した．しかし全体のノード数がわかっていることを前提としており，全体のわからない OSN に対し，ランダムウォークを用いることで特徴量を推定する本研究とは異なる．

6 結　　論

本研究では，OSN 上のグラフサンプリングについて，隣接ノードの情報を取得するクエリの回数をコストとして考え，そのうえで隣接ノードの情報を確率的に利用するランダムウォークを提案した．提案手法により推定可能な特徴量は隣接ノードのリストを取得するときに同時に取得できる情報から得られる特徴量であり，X では平均次数や次数分布，Mastodon ではボット率や平均投稿数などが推定可能である [5, 7]．実験により，平均次数推定，ランダムに付与した 2 値ラベルの付与割合推定，高次数のノードに偏らせて付与した 2 値ラベルの付与割合推定，低次数のノードに偏らせて付与した 2 値ラベルの付与割合推定において，提案手法が既存手法よりも精度のよい推定が可能であることを示した．また隣接ノードサンプリングの確率であるパラメータ α については，α が 1 に近づくほど精度が向上することを，複数のグラフにおける実験により示した．

謝　　辞

本研究は JSPS 科研費 JP21H04872 の助成を受けたものです．

文　　献

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web*, pages 835–844, 2007.

[2] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, 2007.

[3] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, 2011.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, 2010.

[5] Twitter API documents. `https://developer.twitter.com/en/docs/twitter-api`. (Accessed on 19/11/2023).

[6] Facebook API documents. `https://developers.facebook.com/docs/graph-api`. (Accessed on 19/11/2023).

[7] Mastodon API documents. `https://docs.joinmastodon.org/api`. (Accessed on 19/11/2023).

[8] Stephen J Hardiman and Liran Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 539–550, 2013.

[9] Rong-Hua Li, Jeffrey Xu Yu, Lu Qin, Rui Mao, and Tan Jin. On random walk based graph sampling. In *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering*, pages 927–938, 2015.

[10] Bruno Ribeiro and Don Towsley. On the estimation accuracy of degree distributions from graph sampling. In *Proceedings of the 51st IEEE Conference on Decision and Control*, pages 5240–5247, 2012.

[11] Xiaowei Chen, Yongkun Li, Pinghui Wang, and John C. S. Lui. A general framework for estimating graphlet statistics via random walk. *Proceedings of the VLDB Endowment*, 10(3):253–264, 2016.

[12] Kenta Iwasaki and Kazuyuki Shudo. Estimating the clustering coefficient of a social network by a non-backtracking random walk. In *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing*, pages 114–118, 2018.

[13] Toshiki Matsumura, Kenta Iwasaki, and Kazuyuki Shudo. Average Path Length Estimation of Social Networks by Random Walk. In *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing*, pages 611–614, 2018.

[14] Kazuki Nakajima, Kenta Iwasaki, Toshiki Matsumura, and Kazuyuki Shudo. Estimating top-k betweenness centrality nodes in online social networks. In *Proceedings of the 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, pages 1128–1135, 2018.

[15] Toshiki Matsumura and Kazuyuki Shudo. Metropolis-hastings random walk with a reduced number of self-loops. In *Proceedings of the 2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*, pages 1556–1563, 2019.

[16] Chul-Ho Lee, Xin Xu, and Do Young Eun. Beyond random walk and metropolis-hastings samplers: Why you should not backtrack for unbiased graph sampling. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, pages 319–330, 2012.

[17] Flavio Chiericetti, Anirban Dasgupta, Ravi Kumar, Silvio Lattanzi, and Tamás Sarlós. On sampling nodes in a network. In *Proceedings of the 25th International Conference on World Wide Web*, pages 471–481, 2016.

[18] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. Towards unbiased bfs sampling. *IEEE Journal on Selected Areas in Communications*, 29(9):1799–1809, 2011.

[19] Kenta Iwasaki and Kazuyuki Shudo. Comparing graph sampling methods based on the number of queries. In *Proceedings of the 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, pages 1136–1143, 2018.

[20] Galin L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1(none):299–320, 2004.

[21] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

[22] David Aldous and James Allen Fill. Reversible markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html`.

[23] Stanford large network dataset collection. `https://snap.stanford.edu/data/index.html`. (Accessed on 19/11/2023).

[24] Márton Pósfai and Albert-Laszlo Barabasi. *Network Science*. Cambridge University Press, 2016.

[25] Mei Fukuda, Kazuki Nakajima, and Kazuyuki Shudo. Estimating the bot population on twitter via random walk based sampling. *IEEE Access*, 10:17201–17211, 2022.

[26] Amir Hassan Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proceedings of the 28th International Conference on Computer Communications*, pages 2701–2705, 2009.

[27] Bruno Ribeiro and Don Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 390–403, 2010.

[28] Guyue Han and Harish Sethu. Waddling random walk: Fast and accurate mining of motif statistics in large graphs. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining*, pages 181–190, 2016.

[29] Johannes Illenberger and Gunnar Flötteröd. Estimating network properties from snowball sampled data. *Social Networks*, 34(4):701–711, 2012.

付　　録

1　提案遷移アルゴリズムの定常分布の証明（定理 10）

まず，以下の補題を示す．

補題 16. $\sum_{v_j \in N(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} = 1$

証明.

$$
\begin{aligned}
&\sum_{j \in N(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} \\
&= \sum_{v_j \in N_{\mathrm{out}}(v_i) \setminus N_{\mathrm{in}}(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} + \sum_{v_j \in N_{\mathrm{in}}(v_i) \setminus N_{\mathrm{out}}(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} \\
&\quad + \sum_{v_j \in N_{\mathrm{out}}(v_i) \cap N_{\mathrm{in}}(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} \\
&= \frac{1}{d_{\mathrm{sum}}(v_i)} \cdot (1 \cdot (d_{\mathrm{out}}(v_i) - d_{\text{in-out}}(v_i)) \\
&\quad + 1 \cdot (d_{\mathrm{in}}(v_i) - d_{\text{in-out}}(v_i)) + 2 \cdot (d_{\text{in-out}}(v_i))) \\
&= 1
\end{aligned}
$$

□

次に定理 10 を示す．定理 5，定理 1 より，$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ であることを示せば必要十分である．ここで，$\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}\mathbf{P}$ とする．

(i) $l \neq k$ のとき

$$
\begin{aligned}
&\tilde{\pi}(e_{lk}) \\
&= \sum_{v_j \in N(v_l)} \pi(e_{lj}) \cdot p_{neighbor}(v_k; v_l) \\
&\quad + \pi(e_{ll}) \cdot p_{neighbor}(v_k; v_l) \\
&= \sum_{v_j \in N(v_l)} \alpha \cdot \frac{m(e_{lj})}{2|E|} \cdot \alpha \cdot \frac{m(e_{lk})}{d_{\mathbf{sum}}(v_l)} \\
&\quad + (1-\alpha) \cdot \frac{d_{\mathbf{sum}}(v_l)}{2|E|} \cdot \alpha \cdot \frac{m(e_{lk})}{d_{\mathbf{sum}}(v_l)} \\
&= \frac{\alpha^2 \cdot m(e_{lk})}{2|E|} \cdot \sum_{v_j \in N(v_l)} \frac{m(e_{lj})}{d_{\mathrm{sum}}(v_l)} + \frac{\alpha(1-\alpha) \cdot m(e_{lk})}{2|E|} \\
&= \frac{m(e_{lk})}{2|E|}(\alpha^2 + \alpha(1-\alpha)) \\
&= \alpha \cdot \frac{m(e_{lk})}{2|E|} \\
&= \pi(e_{lk})
\end{aligned}
$$

1 つ目の等号は定義 9 の確率遷移行列 $\mathbf{P}$ より成り立つ．2 つ目の等号は定義 7 を適用したものである．3 つ目の等号は代数的な式変形により成り立ち，4 つ目の等号は補題 16 を適用することで成り立つ．残りの等号は代数的な式変形により成り立つ．

(ii) $l = k$ のとき

$$
\begin{aligned}
&\tilde{\pi}(e_{lk}) \\
&= \sum_{v_i \in N(v_l)} \pi(e_{i,i}) \cdot p_{walk}(v_l; v_i) \\
&\quad + \sum_{v_i \in N(v_l)} \sum_{v_j \in N(v_i)} \pi(e_{ij}) \cdot p_{walk}(v_l; v_i) \\
&= \sum_{v_i \in N(v_l)} (1-\alpha) \cdot \frac{d_{\mathrm{sum}}(v_i)}{2|E|} \cdot (1-\alpha) \cdot \frac{m(e_{il})}{d_{\mathrm{sum}}(v_i)} \\
&\quad + \sum_{v_i \in N(v_l)} \sum_{v_j \in N(v_i)} \alpha \cdot \frac{m(eij)}{2|E|} \cdot (1-\alpha) \cdot \frac{m(e_{il})}{d_{\mathrm{sum}}(v_i)} \\
&= \sum_{v_i \in N(v_l)} (1-\alpha)^2 \cdot \frac{m(e_{il})}{2|E|} \\
&\quad + \sum_{v_i \in N(v_l)} \alpha(1-\alpha) \cdot \frac{m(e_{il})}{2|E|} \sum_{v_j \in N(v_i)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_i)} \\
&= \sum_{v_i \in N(v_l)} ((1-\alpha)^2 + \alpha(1-\alpha)) \cdot \frac{m(e_{il})}{2|E|} \\
&= (1 - 2\alpha + \alpha^2 + \alpha - \alpha^2) \cdot \frac{d_{\mathrm{sum}}(v_l)}{2|E|} \sum_{v_i \in N(v_l)} \frac{m(e_{li})}{d_{\mathrm{sum}}(v_l)} \\
&= (1-\alpha) \cdot \frac{d_{\mathrm{sum}}(v_l)}{2|E|} \\
&= \pi(e_{lk})
\end{aligned}
$$

1 つ目の等号は定義 9 の確率遷移行列 $\mathbf{P}$ より成り立つ．2 つ目の等号は定義 8 を適用したものである．3 つ目の等号は代数的な式変形により成り立ち，4 つ目の等号は補題 16 を適用することで成り立つ．5 つ目の等式は $m(e_{ij}) = m(e_{ji})$ であることから成り立ち，6 つ目の等式は同じく補題 16 より成り立つ．
(i)，(ii) より任意の $e_{lk} \in \Omega$ において，$\tilde{\pi}(e_{lk}) = \pi(e_{lk})$ となる．
以上より，定理 10 は示された．

2　特徴量推定の証明（定理 15）

まず，定常分布 $\boldsymbol{\pi}$ に関する関数 g の期待値は次のように与えられる．

定義 17. $\mathbb{E}_\pi(g) \stackrel{\mathrm{def}}{=} \sum_{e_{ij} \in \Omega} \pi(e_{ij}) g(e_{ij})$

ここで，定理 3 より，以下の定理が成り立つ．

定理 18. $\{Z_t\}$ が 定常分布が $\boldsymbol{\pi}$ である有限で既約なマルコフ連鎖に従うとき，任意の初期状態から，$t \to \infty$ のとき，

$$\hat{\mu_t}(g) \to \mathbb{E}_\pi(g) \text{ a.s.}$$

定理 15 を示すために，ここで以下の補題を示す．

補題 19. $t \to \infty$ のとき，

$$\frac{2|E|}{n} \cdot \hat{\mu_t}(wg) \to \mathbb{E}_u(f)$$

証明.

$$
\begin{aligned}
&\frac{2|E|}{n} \cdot \hat{\mu_t}(wg) \\
&\to \frac{2|E|}{n} \cdot \mathbb{E}_\pi(wg) \\
&= \frac{2|E|}{n} \cdot \sum_{e_{ij} \in \Omega} \pi(e_{ij}) \cdot w(e_{ij}) \cdot g(e_{ij}) \\
&= \frac{2|E|}{n} \cdot \sum_{v_j \in V} \sum_{v_i \in N(v_j)} \pi(e_{ij}) \cdot w(e_{ij}) \cdot f(v_j) \\
&\quad + \frac{2|E|}{n} \cdot \sum_{v_j \in V} \pi(e_{jj}) \cdot w(e_{jj}) \cdot f(v_j) \\
&= \sum_{v_j \in V} \sum_{v_i \in N(v_j)} \frac{2|E|}{n} \cdot \alpha \cdot \frac{m(e_{ij})}{2|E|} \cdot \frac{1}{d_{\mathrm{sum}}(v_j)} \cdot f(v_j) \\
&\quad + \sum_{v_j \in V} \frac{2|E|}{n} \cdot (1-\alpha) \cdot \frac{d_{\mathrm{sum}}(v_j)}{2|E|} \cdot \frac{1}{d_{\mathrm{sum}}(v_j)} \cdot f(v_j) \\
&= \sum_{v_j \in V} \frac{\alpha}{n} \cdot f(v_j) \sum_{v_i \in N(v_j)} \frac{m(e_{ij})}{d_{\mathrm{sum}}(v_j)} + \sum_{v_j \in V} \frac{1-\alpha}{n} \cdot f(v_j) \\
&= \frac{\alpha}{n} \sum_{v_j \in V} f(v_j) + \frac{1-\alpha}{n} \sum_{v_j \in V} f(v_j) \\
&= \frac{1}{n} \sum_{v_j \in V} f(v_j) \\
&= \mathbb{E}_u(f)
\end{aligned}
$$

最初の式は定理 18 より成り立つ，次の等式は定義 17 を適用したものであり，2 つ目の等式は，定義 4 をもとに展開し，定義 12 を適用したものである．3 つ目の等式は，定理 10 により得られる定常分布 $\boldsymbol{\pi}$ を代入し，定義 14 を適用したものである．4 つ目の等式は，代数的な式変形と，補題 16 を適用したものである．残りの等式は代数的な式変形により成立し，最後の等式は定義 11 により成立する． □

加えて以下の補題が成り立つ．

補題 20. $t \to \infty$ のとき，

$$\frac{2|E|}{n} \cdot \hat{\mu_t}(w) \to 1$$

証明. 任意の $e \in \Omega$ について，$g(e) = 1$ となる関数 g を選び，補題 19 と同様の式変形により成立する． □

補題 19，20 により，$t \to \infty$ のとき，

$$\frac{\sum_{s=1}^{t} w(Z_s) g(Z_s)}{\sum_{s=1}^{t} w(Z_s)} = \frac{\frac{2|E|}{n} \cdot \hat{\mu_t}(wg)}{\frac{2|E|}{n} \cdot \hat{\mu_t}(w)} \to \mathbf{E}_u(f) \ a.s.$$

以上より，定理 15 は示された．

局所的な密度の指標として機能するハイパーグラフ向けクラスタ係数

宮下　陸矢†　　廣中　詩織††　　首藤　一幸††

† 東京工業大学
†† 京都大学

あらまし　本稿では，ハイパーグラフにおけるノードのクラスタ係数の新たな定義を提案する．提案する定義は，局所的な密度の指標として機能するよう，ハイパーエッジ内の全ノードペアを考慮する．クラスタ係数は，3 つのノード間の関係を測るネットワークの主要な統計量の一つである．無向単純グラフだけでなく，集団間の相互作用を捉えるデータ構造であるハイパーグラフに対しても様々なクラスタ係数の定義がなされているが，ハイパーグラフに対する既存の定義はいずれも，同一ハイパーエッジ内で 2 ノードが接続されていてもそれが考慮されない場合がある．本研究では，同一ハイパーエッジ内のノードペアはすべて接続されているとして扱う，局所的な密度の指標として機能する新たなクラスタ係数の定義を提案する．提案した定義を実世界のハイパーグラフに適用した結果，既存の定義では捉えない構造を反映するような特徴を示した．

キーワード　ハイパーグラフ，クラスタ係数，ノードペア関係

1 はじめに

ネットワークは相互作用を表現する構造であり，その記述には構成要素を 1 対 1 で接続する無向単純グラフが広く用いられている．しかし，実世界のネットワークには，構成する要素同士の相互作用が 1 対 1 だけでなく，集団間で起こるものも多く存在する．例えば，電子メールのネットワーク [9] では，1 通のメールに対しての送受信者のアドレス全体は 2 つより多くなり得る．このような集団間の相互作用を持つネットワークを表現するためには，無向単純グラフを一般化した，ハイパーグラフが用いられる．

ノードに対するクラスタ係数 [17] は，3 ノード間の関係を定量化する，無向単純グラフ上の主要な統計量の一つであり，Web のネットワーク [4] など，様々なネットワークの分析に用いられている．また，クラスタ係数は，ネットワークの局所的な繋がりの密度を測るものとして広く知られている．

ハイパーグラフ上でのクラスタ係数の定義も提案されているが，無向単純グラフの定義と矛盾しているものとしていないものとがある．本研究では無向単純グラフの定義と矛盾しないクラスタ係数に着目する．矛盾しないクラスタ係数の定義はいずれも，同一ハイパーエッジ内のノードペアの関係をすべて反映していない．

本研究では，同一のハイパーエッジ内のノードペアの関係を考慮することで，局所的な密度の指標としてより適切な，新たなハイパーグラフのクラスタ係数の定義を設計した．そして，実世界の 3 つのハイパーグラフに対して既存の 2 つのクラスタ係数と提案クラスタ係数を計算し，提案クラスタ係数の評価を行った．さらに，無向単純グラフの定義と比較し，提案クラスタ係数の示す意味を考察した．

2 関連研究

ハイパーグラフや，ハイパーグラフと相互に変換が可能な二部グラフに対して，多くのクラスタ係数が定義されている．Zhou ら [19] は，ハイパーグラフ上の各ノードに対するクラスタ係数と大域的クラスタ係数を，無向単純グラフ上の定義と矛盾しない形で定義している．Gallagher ら [7] は，ハイパーグラフ上の各ノードとノードペアに対するクラスタ係数をいくつか提案し，既存の定義とともにタンパク質の相互作用の観点から物理的な意味を調査している．Aksoy ら [1] は，ハイパーグラフ上のウォークを s-walk により特徴付け，オーダー s に対する局所的クラスタ係数と大域的クラスタ係数を定義している．Kim ら [8] は，ハイパーグラフに対し，一方がもう一方の部分集合でない交差するハイパーエッジのペアであるハイパーウェッジのレベルと，ハイパーグラフのレベルでの望ましい性質を公理として提案し，それらを基にクラスタ係数にあたる推移性を定義している．Opsahl [13] は，二部グラフ上の各ノードに対するクラスタ係数と大域的クラスタ係数を，長さ 4 のパスに基づいて定義している．Latapy ら [10] は，二部グラフ上のノードペアに対して，隣接ノードの重複の度合いを捉えるクラスタ係数を定義した上で，それに基づき各ノードに対するクラスタ係数を提案している．また，二部グラフ上の 4 つのノードのサイクルに着目した定義もいくつかなされている [2], [15].

これらの定義のうち，無向単純グラフのノードに対するクラスタ係数と矛盾しない，Zhou ら [19] の定義と Opsahl [13] の定義はいずれも複数のハイパーエッジによる三項間の関係を捉えており，同一ハイパーエッジ内のノードペアの関係をすべては反映していない．本研究ではこれら 2 つの定義と提案クラスタ係数を比較し，提案クラスタ係数の評価を行う．

This is an unrefereed paper.

3 準　　備

3.1 表　　記

本稿では，ハイパーグラフ G を，$G = (V, E)$ と表す．集合 $V = \{v_1, v_2, \ldots, v_N\}$ はノード集合であり，集合 $E = \{e_1, e_2, \ldots, e_M\}$ はハイパーエッジの集合である．ここで，N はノード数，M はハイパーエッジ数であり，ハイパーエッジ集合 E の各要素 e_i は，ノード集合 V の部分集合である．また，あるハイパーエッジに含まれるノードの個数を，ハイパーエッジのサイズと表現する．

3.2 ハイパーグラフと二部グラフ

ハイパーグラフ $G = (V, E)$ は，二部グラフ $G' = (V, E, \mathcal{E})$ と相互に変換可能である．G' 上の集合 V, E は，二部グラフの 2 つのノード集合であり，集合 $\mathcal{E}$ は二部グラフのエッジ集合である．ここで，ハイパーグラフ G 上で，ノード v_i がハイパーエッジ e_j に含まれることと，二部グラフ G' 上で 2 つのノード v_i, e_j 間を接続するエッジ (v_i, e_j) が存在することは同値である．本研究ではこの変換を用いて，二部グラフのノード $v \in V$ に対する統計量を，ハイパーグラフのノード $v \in V$ に対して適用した．

4 提 案 手 法

本章では，同一ハイパーエッジ内のノードペアの関係を捉えることで，局所的な密度として機能する，ノードに対するハイパーグラフ向けクラスタ係数の定義を提案する．提案クラスタ係数は以下の 3 つの条件を満たすとする．

1. 提案クラスタ係数は $[0, 1]$ 間の値をとる．
2. 提案クラスタ係数は無向単純グラフのクラスタ係数と矛盾しない．
3. 提案クラスタ係数は同一ハイパーエッジ内のノードペアの関係を反映する．

条件 1, 2 は Zhou ら [19] の挙げた条件に従うものであり，これらを満たすことにより，ハイパーグラフと無向単純グラフを統一して扱えるようにする．条件 3 は既存手法と異なるものであり，同一ハイパーエッジ内の関係まで反映することで，隣接ノード間の結び付きをより詳細に反映することを目的としている．

まず，ノードペアの関係をすべて考慮するために，対象とするハイパーグラフを，ノード同士が 1 対 1 で接続された重み付き無向グラフに変換して考える．ここで，2 つのノード間の結びつきは，1 対 1 のとき最も強く，その他のノードをより多く含むハイパーエッジによる結びつきのときはより弱いと仮定する．変換したグラフ上のノード v, w 間のエッジ重み W_{vw} は次の式で定義する．

$$W_{vw} = \begin{cases} \max\limits_{e \in E; v, w \subseteq e} \dfrac{1}{|e| - 1} & \text{if } \{e \in E; v, w \subseteq e\} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

すなわち，2 ノード v, w がハイパーエッジ $e = \{w, w\}$ によって接続されているとき，重み W_{vw} は最大になり，よりサイズが大きいハイパーエッジによる接続しか存在しないとき，重み W_{vw} はより小さい値となる．ただし，2 ノード v, w 間を接続するハイパーエッジが存在しないとき，重み W_{vw} は 0 となる．

次に，対象とするノード v とその 2 つの隣接ノードの組 u, w が形作る三角形に対して重みを設定する．重み付きグラフに対しては様々なクラスタ係数が定義されており [3], [12], [14], [18]，三項間の関係の尺度はいくつか考えられるが，ここでは，3 ノード u, v, w 間を接続するエッジの積 $W_{uv} \cdot W_{vw} \cdot W_{uw}$ と定義する．隣接ノード間の W_{uw} を含めて積をとることにより，いずれか 1 つのエッジの重みが 0 であり三角形の形とならないとき，三角形の重みは 0 となる．

そして，無向単純グラフのクラスタ係数における隣接ノード間にとり得るエッジの数に相当する，基準となる三角形の重みの総和を設定する．基準の取り方はいくつか考えられるが，隣接ノード間の実際の重みの度合いを大きく反映させるために，対象とするノードの隣接ノード間の重みが 1 のときである，$W_{uv} \cdot W_{vw} \cdot 1$ を基準とする．すなわち，ノード v に対する提案クラスタ係数 $C_{\text{proposed}}(v)$ は次の式で表される．

$$C_{\text{proposed}}(v) = \frac{\sum\limits_{i,j \in N(v), i \neq j} W_{iv} \cdot W_{vj} \cdot W_{ij}}{\sum\limits_{i,j \in N(v), i \neq j} W_{iv} \cdot W_{vj} \cdot 1}$$

ただし，$N(v)$ は v の隣接ノードの集合である．すべての重み W_{ab} は $[0, 1]$ 間の値をとるから，提案クラスタ係数は常に $[0, 1]$ 間の値をとる．

また，ハイパーグラフが無向単純グラフとして表せるとき，すなわち，すべてのハイパーエッジのサイズが 2 であるとき，変換後の重み付き無向グラフは，接続されていたノード間のエッジの重みが 1，ハイパーエッジにより接続されていなかったノード間のエッジの重みが 0 となる．したがって，提案クラスタ係数の分子は対象ノード v と 2 つの隣接ノードが実際に形作る三角形の数，分母は対象ノード v と 2 つの隣接ノードが形作り得るすべての三角形の数となるから，無向単純グラフ上のクラスタ係数の定義と一致する．

5 評　　価

本章では，既存のクラスタ係数や統計量との比較を通して，提案するクラスタ係数の性質を評価する．

なお，比較する既存の定義として，提案するクラスタ係数と同様に，無向単純グラフのクラスタ係数と矛盾しない，Opsahl の定義 [13] と，Zhou らの定義 [19] を，それぞれ C_{Opsahl}，C_{Zhou} と表して用いた．ここで，C_{Opsahl} は二部グラフ上で定義されたクラスタ係数であり，あるノード v に対する $C_{\text{Opsahl}}(v)$ は，

$$C_{\text{Opsahl}}(v) = \frac{(\text{ノード } v \text{ を中心とする閉じた長さ 4 のパス})}{(\text{ノード } v \text{ を中心とするすべての長さ 4 のパス})}$$

と表される．また，C_{Zhou} はハイパーグラフ上で定義されたクラスタ係数であり，あるノード v に対する $C_{\text{Zhou}}(v)$ は，

表 1: 3 ノードのハイパーグラフに対するクラスタ係数

ノード v のクラスタ係数	I	II	III	IV-a	IV-b	V	VI
							
$C_{\mathrm{Opsahl}}(v)$ [13]	0	1	0	0	0	1/3	1
$C_{\mathrm{Zhou}}(v)$ [19]	0	1	0	0	0	1/3	1/3
$C_{\mathrm{proposed}}(v)$	0	1	1/2	1/2	1	1/2	1

$$C_{\mathrm{Zhou}}(v) = \begin{cases} \frac{1}{\binom{|M(v)|}{2}} \sum_{e_i, e_j \in M(v)} EO(e_i, e_j) & \text{if} d(v) > 1 \\ 0 & \text{if} d(v) = 1 \end{cases}$$

と表される．ここで，

$$EO(e_i, e_j) = \frac{|N(D_{ij}) \cap D_{ji}| + |N(D_{ji}) \cap D_{ij}|}{|D_{ij}| + |D_{ji}|}$$

であり，$d(v)$ はノード v の次数，すなわち v を含むハイパーエッジの個数，$M(v)$ はノード v を含むハイパーエッジの集合，$N(U)$ は，各ノード $v \in U$ 自身と v と隣接するすべてのノードからなる集合 $N(v)$ の積集合，D_{ij} は差集合 $e_i - e_j$ である．

5.1 提案クラスタ係数の性質

提案クラスタ係数が満たすべき 3 条件のうち，2 と 3 を満たすことを確認するために，オーダー 3 のすべての高次モチーフ [11] を例に，3 ノードのみで構成されるハイパーグラフ上の，あるノード v に対して，既存のクラスタ係数 $C_{\mathrm{Opsahl}}(v), C_{\mathrm{Zhou}}(v)$ と提案クラスタ係数 $C_{\mathrm{proposed}}(v)$ を計算した．それぞれのハイパーグラフの図と計算結果を表 1 に示す．

無向単純グラフとして表せるハイパーグラフ I，II では，提案クラスタ係数を含むいずれの定義でも無向単純グラフ上の定義と等しい値をとっている．このことから，既存の定義と同様に提案クラスタ係数は無向単純グラフのクラスタ係数と矛盾しておらず，条件 2 を満たしている．

既存の定義ではどちらも，v の 2 つの隣接ノード同士が接続されていても，関わるハイパーエッジが 3 つ未満であるハイパーグラフ III, IV-a, IV-b ではクラスタ係数が 0 になっている．また，3 ノードが最も密に繋がっているハイパーグラフ VI において，$C_{\mathrm{Zhou}}(v)$ は 1/3 と小さい値をとっている．これに対して，提案クラスタ係数は，隣接ノード間が接続しているハイパーグラフでは 0 より大きい値をとっており，最も密に繋がっている VI では最大の値である 1 をとっている．したがって，提案クラスタ係数は，既存の定義では考慮されていない隣接ノード間のノードペア関係を捉えており，条件 3 も満たす．これらの結果から，提案クラスタ係数は，局所的な密度として機能するために定めた条件を満たしていることが示された．

5.2 実世界のハイパーグラフのクラスタ係数

本節では，実世界のハイパーグラフのデータセットに対してクラスタ係数を計算し，提案クラスタ係数と既存の 2 つのクラスタ係数の性質の違いを示す．

primary-school ハイパーグラフは人々の接触ネットワーク [5], [16] であり，人をノードとして，同時に接触した人々の集合をハイパーエッジとして表現している．email-Enron ハイパーグラフは電子メールのネットワーク [5], [9] であり，電子メールのアドレスをノードとして，それぞれの電子メールの送受信者のアドレスの集合をハイパーエッジとして表現している．NDC-classes ハイパーグラフは医薬品のネットワーク [5] であり，クラスラベルをノードとして，ある医薬品に適用されているクラスラベルの集合をハイパーエッジとして表現している．なお，用いたデータセットに対して，重複するハイパーエッジを取り除き，最大連結成分を抽出する前処理を行った．表 2 に前処理を行ったハイパーグラフの統計量と，それぞれの定義により算出した平均クラスタ係数を示す．表中の $\bar{C}_{\mathrm{Opsahl}}, \bar{C}_{\mathrm{Zhou}}, \bar{C}_{\mathrm{proposed}}$ はそれぞれ，すべてのノード v に対する $C_{\mathrm{Opsahl}}(v), C_{\mathrm{Zhou}}(v), C_{\mathrm{proposed}}(v)$ の平均である．

表 2 の平均クラスタ係数の結果より，3 つのデータセットにおける平均クラスタ係数の大小関係は 3 つの定義で変わらず同じであるため，どの定義もネットワークの局所的なつながりの密度を測れていると考えられる．

図 1a は，それぞれのデータセットにおける，3 つのクラスタ係数の分布を示すものである．図 1a をみると，提案クラスタ係数は他の 2 つのクラスタ係数と少し異なる分布になっている．特に，NDC-classes ハイパーグラフでは，他の 2 つの定義では 0 に近い偏った値が多いが，提案クラスタ係数はより幅広い値をとっている．

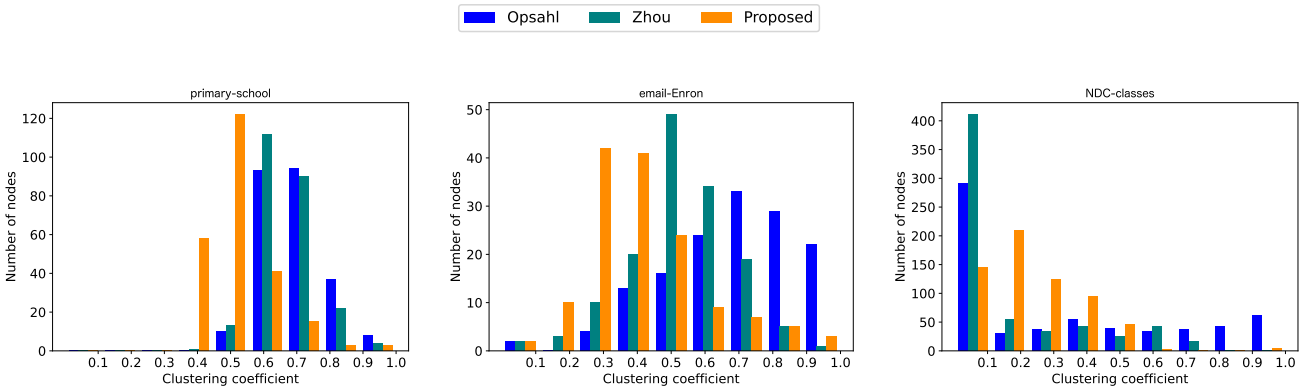
図 1b は提案クラスタ係数の定義と，他の 2 つの定義とで計算される値の大小関係を比べたものである．図 1b より，primary-school ハイパーグラフでは，3 つの定義で各ノードに対して似たような値が計算できているが，残りの 2 つのデータセットでは異なる値が計算されている傾向にあることがわかる．既存のクラスタ係数と提案クラスタ係数の結果の相関係数をみると，primary-school ハイパーグラフと email-Enron ハイパーグラフがどちらの定義に対しても正の値をとるのに対し，NDC-classes ハイパーグラフではどちらの定義に対しても負の値であり，特に大きく異なる結果がみられた．

図 1 の結果から，提案クラスタ係数と既存のクラスタ係数の定義との違いは，データセットごとに差があることがわかった．表 1 より，提案クラスタ係数と既存の定義の違いはオーダー 3 の高次モチーフ [11] の III, IV にあると考えられるため，各データセットに含まれる高次モチーフの数を比較する．3 つのデータセット中のそれぞれの高次モチーフの個数を図 2 に示す．

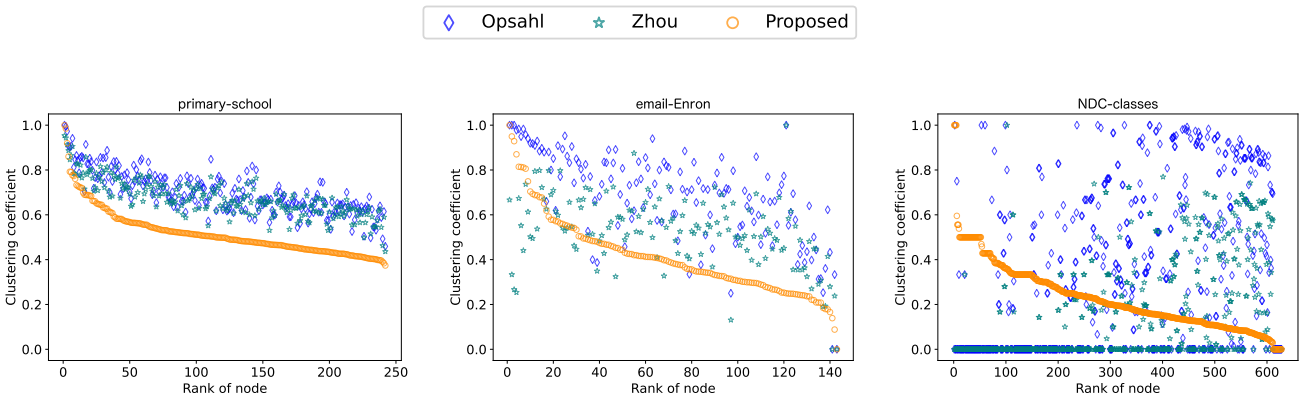
図 2 をみると，図 1 で提案クラスタ係数が比較的大きく幅広い値をとっていた NDC-classes ハイパーグラフは，他の 2 つの

表 2: データセット

データ	ノード数	ハイパーエッジ数	エッジ数	平均次数	平均ハイパーエッジサイズ	$\bar{C}_{\text{Opsahl}}$	$\bar{C}_{\text{Zhou}}$	$\bar{C}_{\text{proposed}}$
primary-school [5], [16]	242	12704	30729	126.98	2.42	**0.70**	**0.67**	**0.51**
email-Enron [5], [9]	143	1512	4550	31.82	3.01	**0.68**	**0.52**	**0.41**
NDC-classes [5]	628	816	5688	9.06	6.97	**0.31**	**0.14**	**0.23**



(a) 3 つのクラスタ係数の値を，0.1 ごとの区間の値をとるノードの個数を縦軸としたヒストグラムにより示した．C_{Opsahl}，C_{Zhou}，C_{Proposed} の計算結果をそれぞれ青，緑，橙で表した．



$\rho_{\text{OP}} = 0.827,\ \rho_{\text{ZP}} = 0.840$　　$\rho_{\text{OP}} = 0.702,\ \rho_{\text{ZP}} = 0.346$　　$\rho_{\text{OP}} = -0.359,\ \rho_{\text{ZP}} = -0.408$

(b) すべてのノードの 3 つのクラスタ係数の値を，提案クラスタ係数の値の降順で示した．横軸のノード番号が大きいほど提案クラスタ係数は小さい値である．$\rho_{\text{OP}}, \rho_{\text{ZP}}$ はそれぞれ，C_{Opsahl} と C_{Proposed} のピアソンの相関係数，C_{Zhou} と C_{Proposed} のピアソンの相関係数である．

図 1: 実世界の 3 つのハイパーグラフに対するクラスタ係数の計算結果

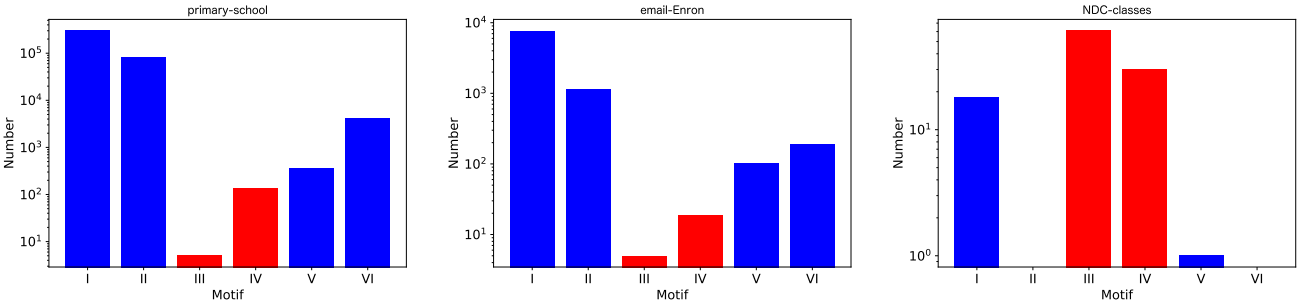


図 2: オーダー 3 の高次モチーフの個数

ハイパーグラフと比較して既存の定義では捉えられないモチーフ III, IV の割合が大きく，モチーフ II，V，VI の割合は小さい．このことから，提案クラスタ係数と既存の定義で大きく値が異なった (図 1b) ことが裏付けられた．また，email-Enron ハイパーグラフで提案クラスタ係数と Zhou らの定義の相関が低くなっているのは，Zhou らの定義で繋がりの密度に対して小さい値となるモチーフ VI が影響していると推察される．

表 3: Davis' Southern Women データセットのクラスタ係数

ノード	無向単純グラフの クラスタ係数	提案クラスタ係数
EVELYN	0.90	0.185
LAURA	0.96	0.204
THERESA	0.90	0.182
BRENDA	0.96	0.200
CHARLOTTE	1.00	0.228
FRANCES	0.96	0.184
ELEANOR	0.96	0.169
PEARL	0.93	0.159
RUTH	0.90	0.156
VERNE	0.90	0.163
MYRNA	0.93	0.180
KATHERINE	0.93	0.190
SYLVIA	0.90	0.176
NORA	0.90	0.147
HELEN	0.90	0.156
DOROTHY	0.93	0.154
OLIVIA	1.00	0.195
FLORA	1.00	0.195

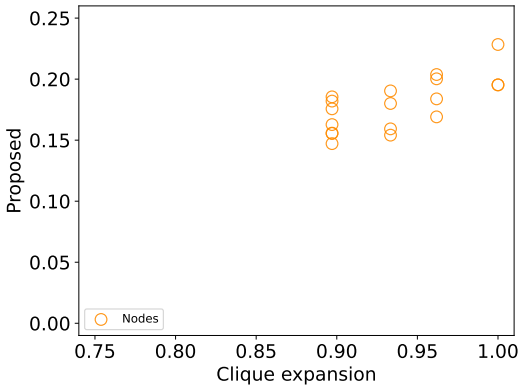


図 3: Davis' Southern Women データセットの各ノードの 2 種類のクラスタ係数の散布図で示した．横軸はクリーク展開したグラフに対するクラスタ係数の値，縦軸は提案クラスタ係数の値を表す．

5.3 クリーク展開で得られる無向単純グラフのクラスタ係数との比較

提案クラスタ係数のようにハイパーグラフの隣接ノードペアの繋がりを考慮する方法として，各ハイパーエッジを無向単純グラフにおける完全グラフとみなして表現するクリーク展開を行い，得られた無向単純グラフに対してクラスタ係数を計算する手法が挙げられる．本節では，ラベル付きのデータセットに対して無向単純グラフのクラスタ係数と提案クラスタ係数を計算して比較することで，実世界のネットワークにおける提案クラスタ係数の示す意味を考察する．

Davis' Southern Women データセット [6] は，18 人の南部の女性の 14 の社交イベントへの参加のネットワークである．ここでは，人物をノードとし，あるイベントに参加した人物の集合をハイパーエッジとしたハイパーグラフとして扱う．

表 3 は，Davis' Southern Women データセットのハイパーグラフの各ノードに対して，提案クラスタ係数と，クリーク展開したときの無向単純グラフのクラスタ係数を示すものである．また，ノードごとの 2 つのクラスタ係数を図 3 に示した．図 3 より，無向単純グラフの定義による値は比較的大きく，等しい値となるノードが多く存在するのに対して，提案クラスタ係数による値は比較的小さく，様々な値をとる．

提案クラスタ係数はその計算の過程で，各エッジが $[0, 1]$ 間の重みをとる重み付き無向単純グラフに変換しているため，重み付けせずにすべての繋がりを表すクリーク展開を用いるときよりも，提案クラスタ係数による値は小さくなる．また，この重み付けにより，クリーク展開では捉えられないハイパーグラフの特徴を反映しているため，提案クラスタ係数はより多様な値をとっていると考えられる．エッジ間の重みはハイパーエッジのサイズが小さいほど大きくなるように定義しているため，提案クラスタ係数はこのネットワークにおいて，対象とするノードの人物と関係する人物同士の関わりの強さを，より関わる人物が少ない方が関わりが強いという仮定の下で，より詳細に捉えていると言えると考えられる．

このような仮定による重み付けは，ハイパーエッジにより表している事象に関わる要素が多くなるほど要素同士の繋がりが希薄になるようなネットワークに対して特に有効であると考えられる．また，分析の目的に応じて重み付けの方法を変化させることで，隣接ノード間の繋がりをより適切に表すことができると予想される．

6 結　　論

本研究では，ハイパーグラフの各ノードのペアに適切な重みを付けることで，同一ハイパーエッジ内のノードペアの関係を反映する，新たなクラスタ係数の定義を提案した．そして，提案クラスタ係数が実際に，既存の定義では捉えていないノードペアの関係を反映することを，3 ノードのみから構成されるハイパーグラフに対する計算結果を用いて示した．

さらに，3 つの実世界のハイパーグラフのデータセットを用いて，提案クラスタ係数と既存のクラスタ係数の値を比較した．提案クラスタ係数は，オーダー 3 の高次モチーフの比率によらず，いずれのデータセットにおいても値が 0 となるノードが少なく幅広い値をとることから，既存の定義では捉えられない構造を反映していると考えられる．

また，ラベル付きネットワークを用いて，提案クラスタ係数が重み付けにより，クリーク展開で得られる無向単純グラフのクラスタ係数では捉えられない特徴を反映していることを示した．

謝　　辞

中嶋一貴氏（東京都立大学）には，議論を通じて多くの示唆を頂きました．

文　献

[1] Sinan G Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science*, 9(1):16, 2020.

[2] Sinan G Aksoy, Tamara G Kolda, and Ali Pinar. Measuring and modeling bipartite graphs with community structure. *Journal of Complex Networks*, 5(4):581–603, 2017.

[3] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101(11):3747–3752, 2004.

[4] Luca Becchetti, Paolo Boldi, Carlos Castillo, and Aristides Gionis. Efficient algorithms for large-scale local triangle counting. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):1–28, 2010.

[5] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018.

[6] A Davis, BB Gardner, and MR Gardner. *Deep South; a social anthropological study of caste and class*. University of Chicago Press, 1941.

[7] Suzanne Renick Gallagher and Debra S. Goldberg. Clustering coefficients in protein interaction hypernetworks. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 552–560. Association for Computing Machinery, 2013.

[8] Sunwoo Kim, Fanchen Bu, Minyoung Choe, Jaemin Yoo, and Kijung Shin. How transitive are real-world group interactions? - measurement and reproduction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 1132–1143. Association for Computing Machinery, 2023.

[9] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer, 2004.

[10] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[11] Quintino Francesco Lotito, Federico Musciotto, Alberto Montresor, and Federico Battiston. Higher-order motif analysis in hypergraphs. *Communications Physics*, 5(1):79, 2022.

[12] Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.

[13] Tore Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social networks*, 35(2):159–167, 2013.

[14] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[15] Garry Robins and Malcolm Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, 10:69–94, 2004.

[16] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.

[17] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.

[18] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):17, 2005.

[19] Wanding Zhou and Luay Nakhleh. Properties of metabolic graphs: biological organization or representation artifacts? *BMC bioinformatics*, 12(1):1–12, 2011.

GPU を用いた大規模グラフ処理アルゴリズムの高速化

岡崎　大知†　　常　　穹†　　宮崎　　純†

† 東京工業大学情報理工学院情報工学系　〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †okazaki@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.ac.jp

あらまし　近年では，ソーシャルネットワークやウェブ検索，交通ネットワークなどさまざまな分野で用いられる実世界グラフデータが日々蓄積され，大規模化している．そのため，大規模グラフの処理を高速化する手法への需要が高まっている．そこで本研究では，CPU メモリの容量を超えるサイズの大規模グラフを処理することのできるシステムである GridGraph のストリーミング処理を，並列計算に優れた処理装置である GPU で実装することにより，大規模グラフ処理の高速化をはかる．評価実験では，幅優先探索と PageRank という二つのアルゴリズムについて GPU を用いて実装し，CPU 実行時間との比較を行った．その結果，幅優先探索では最大 1.45 倍，PageRank では最大 10.6 倍という高速化の効果が得られたが，PageRank において一つの頂点へのアクセスが膨大になるようなグラフに対しては速度が大幅に低下してしまうという問題も発生した．

キーワード　効率性・スケーラビリティ，ソーシャルメディア

1　はじめに

今日のビッグデータ時代において，実世界の多くのデータはグラフとして表現されている．たとえば，ソーシャルネットワークを例にとると，X(旧 Twitter) や Instagram などのフォロワー関係はグラフとして表現されるデータの代表例である．他にも，Web ページのリンク構造，交通ネットワーク，タンパク質の相互作用ネットワークなど，実世界の多分野のデータがグラフとして表現されている．そして，これらのグラフデータは日々蓄積され，その大きさは爆発的に増加している．

こうした実世界グラフには，様々な利用価値がある．例えば，交通ネットワークを例にとると，歩行者への最適な道案内，電車やバスなどの公共交通機関の運行スケジュールの見直しなどにより，道路や公共交通機関の効率を向上させることができ，都市計画においても，道路や他のインフラの最適な配置を検討する際に重要な情報源となる．このように，実世界グラフは新たなプロジェクトの土台として非常に有用である．しかし，それらのデータは大規模であることがほとんどであり，その処理を効率的に行えるかどうかは重要な課題である．

大規模グラフを高速に処理するために有効な計算資源としては，分散システムが挙げられる．分散システムは，複数の独立した計算機がネットワークを介して接続され，一つのタスクを同時に処理することのできるコンピュータシステムである．分散システムを用いることで，一つのコンピュータでは処理することのできないグラフデータについても処理することが可能になるが，新しいハードウェアやソフトウェアの追加，定期的なメンテナンスが不可欠であり，そのコストは無視できない．

そこで近年では大規模グラフの処理に GPU を用いる動きが活発化している．GPU は，画像処理のために設計されたプロセッサであり，多くのコアを搭載している．これにより，同時に多くのタスクを並列処理することが可能であり，その並列演算能力を画像処理以外の目的にも利用する技術は GPGPU(General-purpose computing on GPU) [1] とも呼ばれ，一般的なものとなっている．グラフデータを処理する際に用いられるグラフアルゴリズムは，多くの頂点やエッジの処理が必要な演算であり，GPU の並列性と親和性が高い．

本稿では，ホストメモリ上にデータが収まらないグラフについて，使用頻度の高い幅優先探索 (BFS) や PageRank といったグラフアルゴリズムを，GPU を用いることで高速化する手法について検討する．具体的には，大規模グラフを高速に処理することのできる CPU ベースのシステムである GridGraph [2] の幅優先探索と PageRank アルゴリズムについて，そのストリーミング処理を GPU で実装し，その処理に要する時間と CPU 実装の実行時間との比較を行う．

2　関連研究

2.1　GridGraph

GridGraph は，メモリサイズを超えるような大規模なグラフを処理することのできる CPU ベースのシステムである．その性能は，単一のマシンで大規模なグラフを処理することのできる GraphChi [3] や X-Stream [4] といった他の既存の CPU

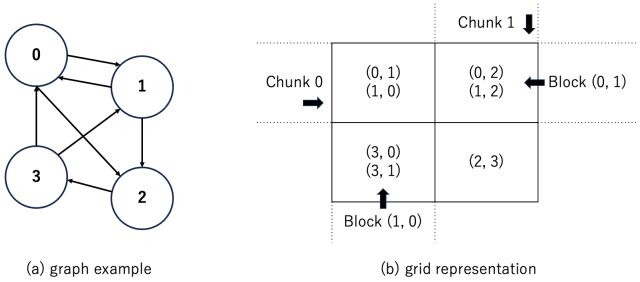


図 1　グラフとそのグリッド表現

ベースのアウトオブコアシステムを凌駕する．GridGraph では前処理として，グラフをグリッド表現という形に分割する．グリッド表現とは，1 次元に分割された頂点チャンクと 2 次元に分割されたエッジブロックによって構成される分割方法で，エッジブロックは，ソース頂点チャンクが行，デスティネーション頂点チャンクが列を表すようにエッジが格納されている．グラフのグリッド表現への変換の例を図 1 に示す．この分割は，グラフデータとして読み込まれたエッジリストから，エッジの順次読み込み，エッジのバッチへの分割，各バッチのタスクバーへのプッシュ，各スレッドによるタスクキューからのフェッチ，バッチ内のエッジが属するブロックの計算，対応するエッジブロックファイルへのエッジの追加，という流れで行われる． 未ソートのエッジリストから対応するブロックへエッジの割り当てが行われるため，小さな前処理オーバーヘッドでグラフを分割することが可能である．

GridGraph は，二階層のパーティションという特徴的な手法を採用している．これは，グリッド表現に分割されたブロックを，そのアクセス順序を調整することで仮想的にグループ化するという方法である．その分割数は，パーティションに含まれる頂点のデータサイズが指定したメモリ量よりも小さくなるような最小の数を選択し，パーティションに必要な頂点データが CPU メモリに収まるようにしている．こうしてエッジブロックへのアクセス順序を調節し，処理に必要な頂点データをメモリにキャッシュすることで，高コストなディスクへの書き込み量を削減することが可能となる．

GridGraph では，エッジのストリーミングと頂点のストリーミングという二種類のストリーミング操作によってグラフの処理を行う．エッジのストリーミングでは，アクティブなエッジを持つエッジブロックを読み込み，エッジブロック内の各エッジに対して，ソース頂点がアクティブであれば演算を行う，という処理を行なっている．エッジのストリーミングの疑似コードを Algorithm1 に示す．このコードにおいて，F_e はエッジを処理するユーザ定義関数であり，F は頂点を引数として，その頂点がアクティブであるかどうかを返す関数として定義されている．この F を用いることによって，更新の可能性がないエッジへのアクセスをスキップすることができ，実行時間を短縮することができる．

Algorithm 1 streaming edges

```
1: function stream_edges(F_e, F)
2:     Sum = 0
3:     for each active block do
4:         for each edge ∈ block do
5:             if F(edge.source) then
6:                 Sum += F_e(edge)
7:             end if
8:         end for
9:     end for
10:    return Sum
11: end function
```

Algorithm 2 streaming vertices

```
1: function stream_vertices(F_v, F)
2:     Sum = 0
3:     for each vertex do
4:         if F(vertex) then
5:             Sum += F_v(vertex)
6:         end if
7:     end for
8:     return Sum
9: end function
```

頂点のストリーミングでは，各頂点に対して，その頂点がアクティブであれば特定の処理を行う．頂点のストリーミングの疑似コードを Algorithm2 に示す．このアルゴリズムにおいて，F_v はユーザ定義の頂点を処理する関数である．頂点のストリーミングにおいても，スキップ関数 F を用いることで更新の必要がない頂点へのアクセスをスキップすることができる．

2.2 GraphChi

GraphChi は大規模グラフを効率的に計算するためのディスクベースのシステムである．このシステムでは，前処理として，頂点を不連続な区間に分割し，各区間にはその区間内にデスティネーション頂点を持つすべてのエッジを格納するシャードが関連づけられる．シャード内のエッジはそのソース頂点のインデックスによってソートされており，区間は各シャード内のエッジ数のバランスが取られるように選択される．そして，区間数はどのシャードも完全にメモリにロードできるような値に設定される．この前処理では，エッジをソース頂点によりソートする必要があるが，そのためにはエッジを何度も通過させる必要があり，GridGraph のグリッド表現への分割と比べると非常に時間のかかる処理になっている．

この手法では，エッジの値を読み込むことで隣接する頂点からデータを収集し，頂点で処理を行い，新しい値をエッジに書き込むという頂点中心の処理モデルが採用されている．そして，ランダムな I/O を削減するために新たな Parallel Sliding Windows(PSW) という方法が提案され，大規模なグラフを効率的に処理することが可能となっている．

2.3 X-Stream

X-Stream は，大規模グラフを単一のマシンで処理することのできるシステムである．このシステムでは，頂点に状態を保持する scatter-gather プログラミングモデルを保持しつつ，辺中心の処理を行い，順序なしエッジリストのストリーミングを行う．scatter フェーズではすべてのエッジをストリーミングし，頂点の状態を近傍に伝播するための更新を生成する．そして gather フェーズでは，すべての更新をストリーミングし，再計算して頂点の状態を更新する．このシステムは，エッジへのアクセスが連続的であるため最大のスループットを達成することができるが，エッジと同じサイズの更新を生成する必要があり，大規模グラフを扱う際には致命的な問題となる可能性がある．

このシステムでは前処理において，各パーティションが高速

メモリ (インメモリーの場合は CPU キャッシュ，アウトオブコアの場合はメインメモリ) に収まるように頂点集合を分割する．これは，エッジをストリーミングするために頂点へのアクセスがランダムになっているためである．そしてエッジについては，ソース頂点が対応するパーティションに含まれるように分割される．この分割方法を用いることで，X-Stream はエッジリストをソートする必要がなく，前処理時間の短縮につながっている．

2.4 WolfGraph

WolfGraph [5] は，GPU のグローバルメモリに収まりきらないグラフを処理することのできる GPU ベースグラフ処理フレームワークである．WolfGraph はエッジに対して繰り返し処理を行うエッジ中心処理モデルを採用し，エッジリストと CEL(Concatenated Edge List) というグラフ表現を使用している．CEL は，各サブグラフに均等に分割されたエッジに加えて，すでに処理が完了したサブグラフから現在のサブグラフ内のエッジのデスティネーション頂点と同じソース頂点を持つエッジを追加したものであり，このグラフ表現を利用することで，すでに処理されたサブグラフを再びロードする必要がなくなり，反復回数，そして全体の処理時間を削減することができる．

WolfGraph では，GPU で処理を行う際，スレッドブロックにエッジを割り当てるためにエッジブロックというリストが構成される．これには，メモリに連続的に格納されたエッジの順序なしリストと，ブロック内のエッジに関連する頂点のセットが含まれる．頂点へのアクセスはランダムであるが，頂点を保持するためのデータ構造は GPU のシェアードメモリ上に置かれており，そのアクセスはグローバルメモリへのアクセスと比べて高速である．エッジブロックを構成するためにはエッジリストを単純にエッジ数によって分割すれば良いため，大規模グラフの総処理時間に占める割合が大きい前処理時間を短縮することが可能である．

3 提案手法

本研究では，一つのマシンで大規模グラフを処理することが可能な CPU ベースのシステムとして比較的高速な GridGraph のストリーミング処理を GPU で実装することによりさらなる高速化を目指す．GridGraph は，ナビゲーションクエリとグローバルグラフ処理を両立するグラフデータベースシステム [6] の基盤にもなっており，その処理の高速化は重要な課題である．また，GridGraph の前処理として行われるグリッド表現への分割はデータの局所性に優れており，GPU の限られた計算資源と相性が良い．

GridGraph のストリーミング処理は，膨大な数のデータに対して単純な処理を行わせるという特性から，GPU の並列演算能力を有効に利用することが可能である．しかし，例えば頂点のストリーミングにおいて，BFS が訪問済みの頂点数を取得するという処理であるのに対し，PageRank では訪問確率値の更新が行われるというように，グラフアルゴリズムによってその処理の内容は大きく異なる．そのため，各グラフアルゴリズムに対して異なる GPU 実装の方法を検討する必要がある．

3.1 幅優先探索の GPU 実装

BFS では，アクティブな頂点数が 0 になるまで，エッジのストリーミングを繰り返し行い，最終的に頂点のストリーミングで訪問済みの頂点数を取得し，結果を出力するという処理が行われる．エッジのストリーミングでは，頂点数と同じサイズで，その頂点に到達するエッジのソース頂点のインデックスを示す領域を *parent* としたとき，読み込まれた各エッジのソース頂点がアクティブであれば，デスティネーション頂点のインデックスが指す *parent* 領域を参照し，未訪問であれば，その領域にソース頂点のインデックスを格納するという操作が行われる．

このアルゴリズムを GPU で実装する際，頂点のストリーミングは一度しか行われない処理であり，その処理の内容としては各頂点に対してその頂点が訪問済みであるかを調べるだけであるため，CPU で実装する方が，ホストとデバイス間のデータ転送も必要となる GPU に対して高速に処理できる可能性が非常に高い．そのため，BFS ではエッジのストリーミングに焦点を絞って GPU を用いた実装を行う．

GridGraph の BFS の疑似コードを以下の Algorithm3 に示す．ここで，ストリーム関数 *stream_edges* と *stream_vertices* はそれぞれ，各エッジ，頂点に対して行う具体的な処理関数 *contribute* と *compute* を引数としてとるように記述している．

そして，Algorithm4 にエッジをストリーミングする関数の疑似コードを示す．本論文中の疑似コードでは，頂点が訪問されたかどうかを参照するための領域である *parent*，頂点がアクティブかどうかを表す領域など，頂点に関する領域はすべて GPU グローバルメモリ上に収まると仮定している．カーネル関数には，それがデバイスで行われる処理であることを表すため，*__global__* 修飾子を記述している．

Algorithm 3 BFS

```
1: function contribute(e)
2:     if parent[e.dst] == −1 then
3:         parent[e.dst] = e.src
4:         return 1
5:     end if
6:     return 0
7: end function
8: function compute(vertex)
9:     return parent[vertex] ! = −1
10: end function
11: parent = {−1, ..., −1}
12: parent[start_vertex] = start_vertex
13: active_vertices = 1
14: memcpy vertices HostToDevice
15: while do active_vertices ! = 0
16:     active_vertices = stream_edges(contribute)
17: end while
18: memcpy vertices DeviceToHost
19: return stream_vertices(compute)
```

Algorithm 4 streaming edges in BFS

```
1: function __global__ process_edges(edges, Sum)
2:     e = edges[threadIdx.x]
3:     if F(e.src) then
4:         value = contribute(e)
5:         atomicAdd(&Sum, value)
6:     end if
7: end function
8: function stream_edges(contribute)
9:     Sum = 0
10:    for each active block do
11:        for edges ⊆ block do
12:            memcpy edges HostToDevice
13:            process_edges(edges, Sum)
14:        end for
15:    end for
16: end function
```

Algorithm4 の 11 行目から 14 行目において，ブロック内の各エッジに対して CPU から GPU にデータを転送し，GPU カーネルで処理を行わせるという記述がなされているが，具体的な実装においては，CPU から GPU へのデータ転送と GPU のカーネル実行の回数を減らすために，特定のエッジ数を超えないようにホスト側で連続するブロック内のエッジをまとめてからデバイスへエッジデータの転送を行う．

3.2 PageRank の GPU 実装

PageRank [7] は Web サイトの重要度を示す訪問確率値を計算する最も一般的なアルゴリズムであり，Google の検索エンジンでも使用されている．頂点 A に対して，A をデスティネーション頂点とするようなエッジを持つ頂点を $T_1, ..., T_n$ と定義すると，A の訪問確率値は，

$$PR(A) = \frac{(1-d)}{N} + d \times (\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

として計算される．ここで，パラメータ d は減衰係数と呼ばれる定数で 0 から 1 の間の値を設定することができ，N はグラフ全体の頂点数，$C(X)$ は頂点 X をソース頂点とするエッジの数として定義される．このアルゴリズムでは各反復においてすべての頂点がアクティブになるため，ほとんどの場合，大きなメモリ帯域幅と浮動小数点演算能力が必要となる．

グラフの頂点数が多いとき，PageRank では各頂点の訪問確率値が非常に小さくなってしまうため，更新の際に単純に加算を行うと大きな誤差を生む可能性がある．そのため，確率値などの小さな値の加算では，その対数をとることにより誤差を軽減し，精度を保つという手法がしばしば用いられる．

対数の真数値の加算には，二つの値 a，b の和の対数 $\log(a+b)$ が，

$$\log(a+b) = \log b + \log(1 + \exp(\log a - \log b))$$

と変形できることを利用する．三つ以上の対数の真数の和は，この演算を繰り返すことにより求めることができる．PageRank アルゴリズムを並列処理する際は，エッジのストリーミングにおいて共通のデスティネーション頂点を持つエッジが同時に値の更新を行う可能性があるため，上の演算についてはアトミックに行うことによりデータ競合を回避することとした．デバイス側のアトミック演算はすべて $atomicCAS$ 関数を用いることにより実装することができ，上の演算を行うアトミック関数については $atomicLogAdd$ として実装した．

各頂点の PageRank 値は，値そのものが用いられることはほとんどなく，主に頂点間で値の比較のために使用される．そのため，実装においては対数値のまま取り扱うこととした．以上の点をまとめた具体的な GridGraph の PageRank に対する GPU 実装の疑似コードを Algorithm5 に示す．

このコードでは，反復の回数はユーザが定義することを仮定している．15 行目のエッジのストリーミング処理では，各頂点から出るエッジ数を求めるという操作を行う．18 行目のエッジのストリーミング処理では，BFS の場合と同様，ファイルからエッジを読み込み，データサイズの不均衡を軽減するため，隣接するエッジ数の小さなブロックをまとめてからデバイスにエッジを転送し，各スレッドが 1 つのエッジに対して $contribute$ 処理を並列的に行う．$contribute$ 処理ではデータ競合が発生する可能性があるため，対数の真数値のアトミックな

Algorithm 5 PageRank

```
1: function logAdd(log a, log b)
2:     return max(log a, log b) + log(1 + exp(min(log a, log b) −
   max(log a, log b)))
3: end function
4: function calc_degree(e)
5:     atomicAdd(&Degree[e.src], 1)
6: end function
7: function contribute(e)
8:     atomicLogAdd(&Sum[e.dst], PR[e.src]−log Degree[e.src])
9: end function
10: function compute(v)
11:     PR[v] = logAdd(log(1 − d) − log N, log d + Sum[v])
12: end function
13: d = 0.85
14: PR = {0, ..., 0}
15: stream_edges(calc_degree)
16: while ¬converged do
17:     Sum = {−∞, ..., −∞}
18:     stream_edges(contribute)
19:     stream_vertices(compute)
20: end while
```

Algorithm 6 streaming vertices in PageRank

```
1: function __global__ process_vertices(vertices)
2:     v = vertices[threadIdx.x]
3:     compute(v)
4: end function
5: function stream_vertices(compute)
6:     process_vertices(vertices)
7: end function
```

加算により Sum のデスティネーション頂点のインデックスが指す領域に各ソース頂点の訪問確率値の対数を集計している．19 行目の頂点のストリーミング処理では，エッジのストリーミングと同様に各頂点に対して対数の真数値の加算を行うことによって新しい訪問確率値の対数を求める．頂点のストリーミングについては，その疑似コードを Algorithm6 に示す．

4 評価実験

4.1 実験準備

評価実験に用いた実行環境について，表 1 に示す．また，比較に用いた Epinions [8]，LiveJournal [9]，Twitter [10] という三つのデータセットについて，その頂点数とエッジ数を表 2 に示す．各データセットは，int 型のソース頂点のインデックスと int 型のデスティネーション頂点のインデックスのペアによってエッジを表現するエッジリスト形式で与えられる．また，各グラフデータセットにおけるエッジのデータ量と頂点のデータ量を計算すると，表 3 のようになる．表 3 から，Epinions，LiveJournal グラフについては BFS，PageRank ともに頂点もエッジも GPU メモリ上に収まるが，Twitter グラフについてはエッジが GPU メモリに収まりきらないため，分割して渡す必要があることがわかる．

4.2 実験内容

本実験では，Epinions，LiveJournal，Twitter というグラフデータセットを用いて幅優先探索 (BFS) と PageRank を実行した際の GridGraph の CPU 実装における実行時間と GPU 実装におけるデータ転送時間と実行時間との比較，そして BFS の GPU 実装について一度にデータ転送するエッジ数を変更した際のデータ転送時間と実行時間の比較を行った．CPU の実行時間の計測には $std::get_time$ 関数を使用した．GPU のデータ転送時間，実行時間については，CUDA のイベントを作成し，二つのイベント間の経過時間を計算する $cudaEventElapsedTime$ 関数によって計測を行った．BFS では各反復でエッジのストリーミングの計算にかかる時間，PageRank では各反復でエッジのストリーミングと頂点のストリーミングの計算にかかる時間を計測した．

CPU との比較において，CPU 側では並列処理は行わず，逐次的な処理を行うように実装した．また，GPU 側ではホストからのエッジデータの転送回数が最小限になるようにした．そして，各データセットの前処理として，Epinions，LiveJournal グラフでは $4 \times 4 = 16$ 個のブロックが生成されるような分割を行い，Twitter グラフでは $16 \times 16 = 256$ 個のブロックが生成されるように分割した．PageRank においては，その反復回数をどのグラフデータセットに対しても 10 回と設定した．

GPU へのデータ転送回数による比較については，GPU に渡す最大のエッジ数を増減させ，そのデータ転送時間と実行時間を計測し，その合計を計算することにより行った．データセットとしては，すべてのデータを GPU 上に載せることのできる LiveJournal グラフを用い，分割数は $4 \times 4 = 16$ と設定した．

4.3 実験結果

4.3.1 BFS 実行時間の CPU との比較

各グラフデータにおける BFS の CPU 実行時間と GPU 実行時間を図 2，図 3，図 4 に示す．LiveJournal，Twitter グラフなど，データサイズの大きいグラフに対して BFS を実行すると，CPU から GPU へのデータ転送時間を含めても，CPU 実装に対して GPU 実装の方がそれぞれ 1.45 倍，1.27 倍高速に処理されることがわかった．Epinions グラフのようにデータサイズが小さいグラフでは，GPU のデータ転送時間を含めると，実行にかかる時間はほとんど同等であった．

表 1　実行環境

OS		Windows 11 Pro 64bit
CPU		Intel Core i7-11700
	最大動作周波数	2.50GHz
GPU		NVIDIA GeForce GTX 1060
	CUDA コア	1152
	最大動作周波数	1.71GHz
	メモリ量	3GB
RAM		16GB
CUDA		CUDA 12.3

表 2　使用したデータセット

	頂点数	エッジ数
Epinions	75879	508837
LiveJournal	4847571	68993773
Twitter	41652230	1468364884

表 3　各データセットにおけるエッジのデータ量と頂点のデータ量

	エッジのデータ量	頂点のデータ量	
		BFS	PageRank
Epinions	3.88MB	315KB	889KB
LiveJournal	526MB	19.7MB	55.5MB
Twitter	10.9GB	169MB	477MB

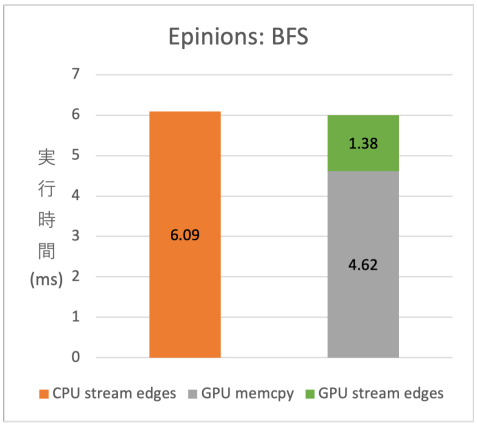


図 2　Epinions における CPU と GPU の BFS 実行時間

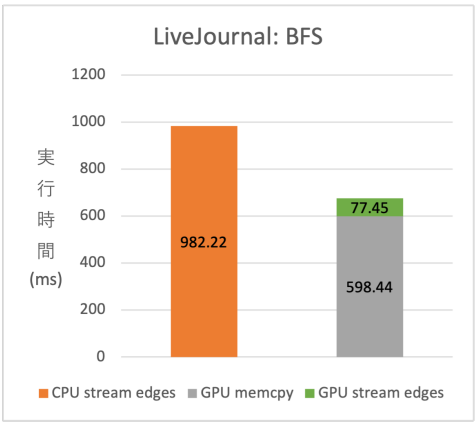


図 3　LiveJournal における CPU と GPU の BFS 実行時間

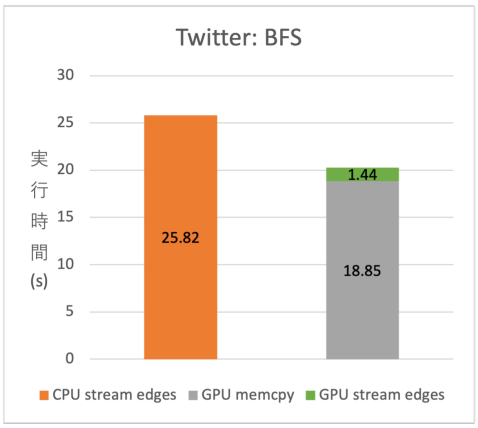


図 4　Twitter における CPU と GPU の BFS 実行時間

4.3.2　PageRank 実行時間の CPU との比較

各グラフデータに対して，CPU と GPU で PageRank を実行した際にかかる時間の比較を図 5，図 6，図 7 に示す．PageRank では，BFS の演算と比べて計算コストの高い浮動小数点演算が行われるため，どのグラフにおいても GPU へのデータ転送より GPU カーネル実行の方が時間がかかっている．図 5 と図 6 から，Epinions と LiveJournal グラフにおいては，ホストからデバイスへのデータ転送時間を含めても，それぞれ 6.85 倍，10.6 倍の高速化の効果が得られることがわかる．しかし，図 7 に示されるように Twitter グラフでは GPU のエッジストリーミングにおいて CPU と比べても 3 倍以上もの時間がかかっており，むしろ速度が低下するという結果になった．

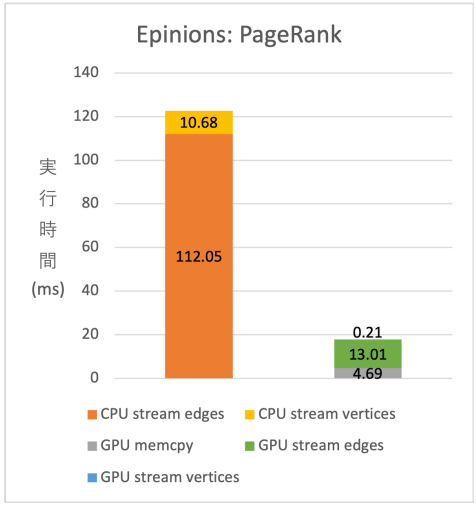


図 5　Epinions における CPU と GPU の PageRank 実行時間

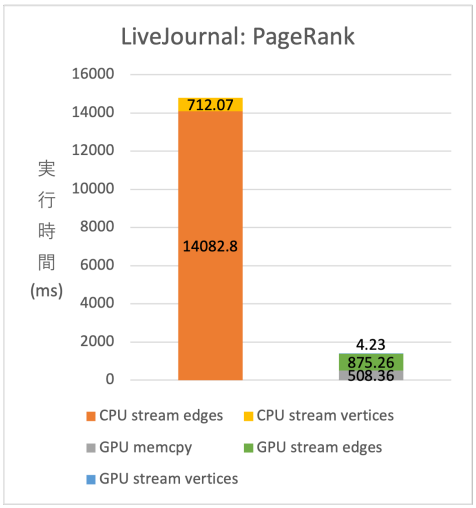


図 6　LiveJournal における CPU と GPU の PageRank 実行時間

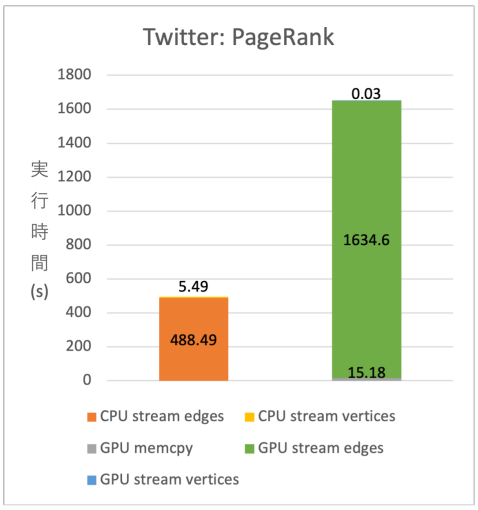


図 7　Twitter における CPU と GPU の PageRank 実行時間

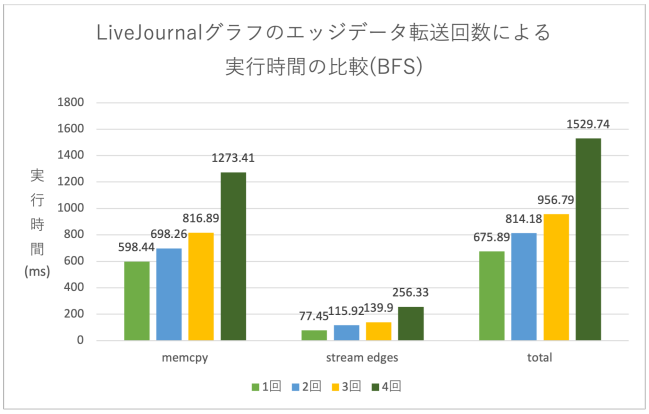


図 8　データ転送回数による実行時間の比較

4.3.3 データ転送回数による実行時間の比較

各反復におけるデータ転送の回数の上限を 1 回から 4 回に増加させたときの LiveJournal グラフの BFS 実行時間の変化を図 8 に示す．この実験結果から，GPU へのデータ転送の回数を増やすと，データ転送時間だけでなくカーネル実行の時間も増加することがわかる．したがって，GPU にデータを渡す際はできるだけその回数を少なくした方が GPU 実装による恩恵を受けやすいと考えられる．

4.4 考　　察

実験結果の GPU の実行時間から読み取れるように，BFS の実行においては GPU の実行よりもデータ転送に多くの時間がかかっている．これは，BFS のエッジストリーミングでは非常に軽い処理しか行われていないことが原因であると考えられる．

また，図 3 の LiveJournal グラフの BFS 実行時間と図 4 の Twitter グラフの BFS 実行時間を比較したとき，LiveJournal グラフの方が GPU 実装における高速化の恩恵が大きいのは，各反復において，Twitter グラフの方が非アクティブなソース頂点を持つエッジを多く含んでいることが原因の一つとして考えられる．アクティブなソース頂点を持たないエッジでは，CPU 実装において処理がスキップされるのに対し，GPU 実装では CPU からのデータ転送前にスキップされることがない．そのため，処理されることのない多くの余分なエッジについてもホストからデバイスへデータ転送する必要があり，そのためにデータ転送時間が膨大になってしまっている可能性が高い．

また，図 7 により示される Twitter グラフの PageRank 実行時間について，GPU の実行時間が CPU と比べても多くかかってしまっているのは，現在の PageRank の実装において，データ競合を回避するために関数をアトミックに実装していることが主な原因であると考えられる．PageRank では，エッジのストリーミングにおいて一つの頂点に複数回アクセスされる可能性があり，その回数が非常に多い場合，アトミックな実装ではその頂点に対する演算が逐次的に行われ，GPU の並列性を十分に発揮することができない．図 4 で Twitter グラフの BFS 実行では PageRank と比べて GPU 実行時間に膨大な時間がかからなかったのは，BFS が一つの頂点に対して最大でも一回のみしかアクセスしないという特徴を持つためであると考えられる．図 9，図 10，図 11 では，横軸を頂点の incoming edge 数，縦軸をその数の incoming edge を持つ頂点の個数として，各グラフデータセットにおける incoming edge 数の特性を示した．ここで，incoming edge とは，ある頂点に対して，その頂点をデスティネーション頂点として持つようなエッジのことである．この結果から，Epinions，LiveJournal と比較して，Twitter グラフでは，100,000 個以上という膨大な incoming edge を持つ頂点が数多く存在していることがわかり，PageRank アルゴリズムの速度低下はアトミック操作の遅延によるものであると考えるのが妥当である．

したがって，PageRank アルゴリズムのエッジストリーミングにおける遅延を解消するためには，アトミック操作に代わってデータ競合を回避しながら総和の計算を行うことのできる他の実装について検討する必要がある．その方策の一つとして，各スレッドが計算した値について，ModernGPU [11] などの GPU 並列処理ライブラリにより提供されている prefix sum と segmented reduction を組み合わせることで計算された値の総和を各デスティネーション頂点で求めるという手法が考えられる．prefix sum と segmented reduction は配列や領域などの要素の集合に対して行われる処理であり，prefix sum では各要素にそれまでの要素の総和を計算した値を格納した配列や領域

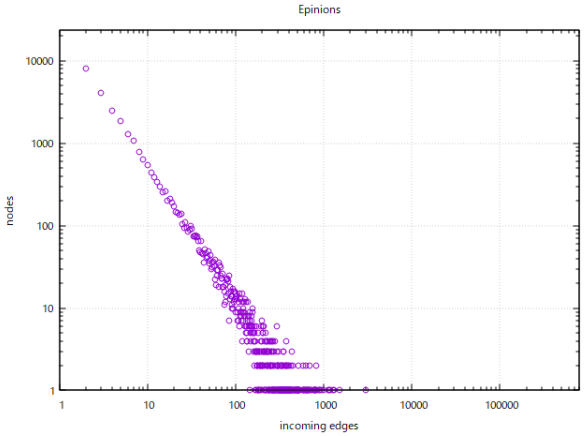


図 9　Epinions における incoming edge 数の分布

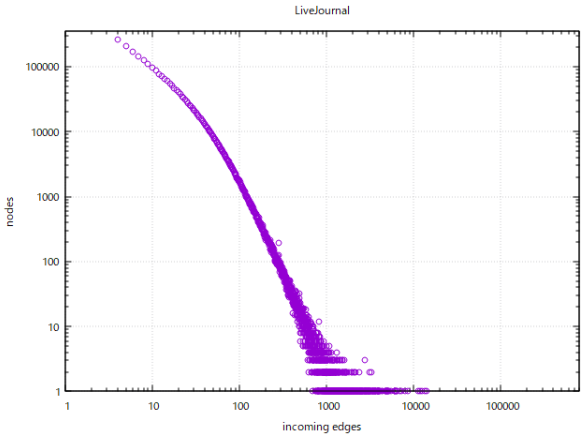


図 10　LiveJournal における incoming edge 数の分布

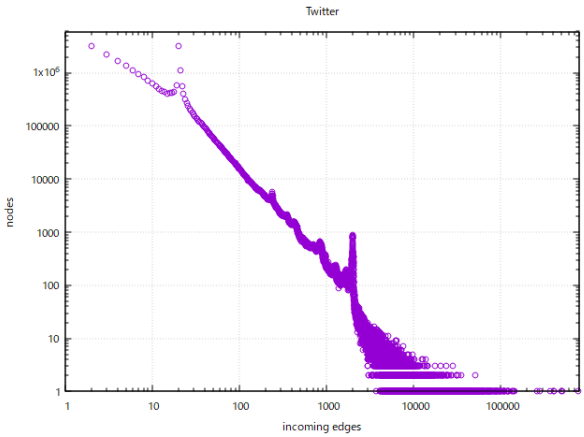


図 11　Twitter における incoming edge 数の分布

を作成し，segmented reduction では配列や領域の各セグメントに対して総和や最大値などを求めるリデュース操作を行う．これらの組み合わせによる総和の計算は，エッジ数分の領域を確保し，各頂点の incoming edge 数から prefix sum によって各デスティネーション頂点が指す領域のオフセットを求め，各エッジで計算された値をオフセットを参照することで領域に格納し，segmented reduction を用いて各デスティネーション頂点に対応する領域に対してその総和を求める，といった流れで行われる．この方法を実装に用いることで，アイドル状態のスレッド数が大幅に減少するため，PageRank アルゴリズムにおいても GPU のリソースを最大限に利用することができ，高速な処理が可能になると考えられる．

5 ま と め

本研究では，大規模グラフデータを処理することのできる GridGraph のストリーミング処理を GPU で実装することにより BFS や PageRank といった基本的なグラフアルゴリズムを高速化する手法について検討した．

提案手法では，GridGraph のグリッド表現により生じるブロック間のエッジ数の不均衡を軽減するため，エッジ数の少ない隣り合うブロックのエッジをホスト側でまとめてからデバイスに転送するように実装した．これにより，ホストからデバイスへのエッジデータの転送回数が減少し，GPU の性能を最大限に発揮することが可能になった．ストリーミング処理については，各グラフアルゴリズムでその内容が大きく異なるため，異なるアプローチで GPU 実装を行った．

評価実験において，BFS のエッジストリーミング処理を GPU で実装したところ，Epinions など小規模なグラフデータにおいてはほとんど実行時間が同じであったが，LiveJournal, Twitter など比較的大規模なグラフでは CPU 実装と比べてそれぞれ 1.45 倍，1.27 倍という高速化の効果が得られた．また，PageRank のストリーミング処理を GPU で実装すると，Epinions，LiveJournal ではそれぞれ 6.85 倍，10.6 倍という高速化の効果が得られたが，Twitter グラフでは高速化することができず，CPU よりも速度が低下してしまった．これは，現在の実装においてエッジストリーミングのデータ競合回避のためにアトミック演算が用いられているため，アクセスされるエッジ数が非常に多い頂点では，膨大な数の処理が逐次的に行われてしまい，GPU の並列処理性能を発揮できていないことが原因であると考えられる．そして，データ転送回数による実行時間の実験では，CPU から GPU にデータを渡す際，できるだけ多くのデータを渡した方がデータ転送時間とカーネル実行時間の双方で有利であるという結果が得られた．

今後の課題としては，まず，現在の実装においては頂点のデータ量が GPU メモリサイズを超えるようなグラフに対して処理を行うことができないため，そのようなグラフでも処理が行えるようにプログラムを修正する必要がある．また，PageRank アルゴリズムにおいて，エッジストリーミングの演算処理をアトミックに実装したとき，一つの頂点にアクセスが集中するソーシャルネットワークなどのグラフにおいて処理速度が大幅に低下してしまうという問題を解決するため，同じ頂点データにアクセスするエッジについて，GPU 並列処理ライブラリなどを利用してリダクションを行い，計算された値の総和を求めるようにするといった工夫が必要がある．さらに，現在では BFS と PageRank の実装しか完了していないため，他の基本的なグラフアルゴリズムについても実装することが求められる．将来的には，CUDA により提供されている CPU と GPU の双方からデータアクセスが可能なメモリである CUDA managed memory などを用いることにより，GPU で演算を行いながら CPU の計算資源も同時に利用するヘテロジニアスな実装についても検討したい．

謝　　辞

本研究は，JST CREST JPMJCR22M2 の支援を受けたものである．

文　　献

[1] Ghorpade, Jayshree, et al. "GPGPU processing in CUDA architecture." arXiv preprint arXiv:1202.4347 (2012).
[2] Zhu, Xiaowei, Wentao Han, and Wenguang Chen. "GridGraph: Large-Scale Graph Processing on a Single Machine Using 2-Level Hierarchical Partitioning." 2015 USENIX Annual Technical Conference (USENIX ATC 15). 2015.
[3] Kyrola, Aapo, Guy Blelloch, and Carlos Guestrin. "GraphChi: Large-Scale Graph Computation on Just a PC." 10th USENIX symposium on operating systems design and implementation (OSDI 12). 2012.
[4] Roy, Amitabha, Ivo Mihailovic, and Willy Zwaenepoel. "X-stream: Edge-centric graph processing using streaming partitions." Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles. 2013.
[5] Zhu, Huanzhou, et al. "WolfGraph: The edge-centric graph processing on GPU." Future Generation Computer Systems 111 (2020): 552-569.
[6] Tarek Aoukar, Jun Miyazaki, Evaluation and Analysis of Graph Navigation query in Edge Streaming Database, The 13th Forum on Data Engineering and Information Management (DEIM 2021), Proc. of DEIM 2021, 1 Mar. 2021.
[7] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 (1998): 107-117.
[8] M. Richardson and R. Agrawal and P. Domingos. Trust Management for the Semantic Web. ISWC, 2003.
[9] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. KDD, 2006.
[10] Haewoon Kwak, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.
[11] Sean Baxter. moderngpu. Retrieved February 1, 2024, from https://github.com/moderngpu/moderngpu

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

2024年3月1日(金) 16:00 ~ 18:10 | T4-A オンライン（Zoom Events)

地理情報

座長:角谷 和俊(関西学院大学)
コメンテータ:横山 昌平(東京都立大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T4-A-9-01] 降水量と河川水位の関係の可視化
～動的時間伸縮法を用いた観測点間類似度の導入

*田中 悠来[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:20 ~ 16:45
[T4-A-9-02] 47都道府県単位のデータ比較可視化によるジェンダーギャップの要因解析

*石戸谷 由梨[1]、伊藤 貴之[1] (1. お茶の水女子大学)

16:45 ~ 17:10
[T4-A-9-03] SNS上の投稿写真の位置情報と意味情報を考慮したピクトリアルマップの自動生成

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

17:10 ~ 17:35
[T4-A-9-04] 地点データのインタラクティブなクラスタリングのための可視化インタフェース

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

17:35 ~ 18:00
[T4-A-9-05] 【技術報告】情報学と交通工学の融合に向けた研究開発 - 万博での交通整流化に向けて

西村拓哉 (株式会社NTTデータグループ)

降水量と河川水位の関係の可視化
～動的時間伸縮法を用いた観測点間類似度の導入

田中悠来† 伊藤貴之‡

†‡お茶の水女子大学 〒112-8610 東京都文京区大塚 2-1-1
E-mail: †g1920522@is.ohca.ac.jp

あらまし 降水量の時間変化の観察は河川水位の予測などの目的で重要である．降水量と河川水位のデータ可視化に関する先行研究では，地図上の多数の観測所にグラフやカラーバーをプロットしていた．本手法では地図上での可視化でなく，河川の支流などの接続構造を模したグラフを構成し，そのノードとなる観測所上に二色塗り分けスパークラインを配置することで，上流での降水量増加から下流での河川水位上昇までの時間差の可視化を実現する．加えて、動的時間伸縮法を適用して算出した，各観測所における時系列データの類似度を活用することで可視化をおこなった情報量を調整する．

キーワード 情報可視化

1. はじめに

降水量の時間変化と河川水位の時間変化は密接に関係している．そのため，降水量の時間変化の観察は河川水位の予測する際に重要な要素となる．この問題に対して我々は，河川の上流で降水量が増加してから下流で水位が上昇するまでの時間差を可視化する一手法を提案している[1]．この手法では，画面の縦方向で河川の上流から下流への流れを，横方向で支流などの接続構造を模したグラフを構成する．そのノードとなる観測所に降水量と河川水位のデータをプロットした二色塗り分けスパークラインを配置することで，両者の数値変化を比較しやすくする．本報告では我々の手法の拡張として，動的時間伸縮法（DTW, Dynamic Time Warping）を利用して，複数の時系列データの類似度を算出し，類似度の高いデータを 1 つにまとめる処理を提案する．この処理により，多数の観測地点におけるデータから，ユーザにとって重要度の高いデータの可視化を実現する．

2. 関連研究

3 種類に分けて関連研究を紹介する．1 つ目は降水量と河川水位のデータを地図上に可視化する手法の研究である．八木ら[2]および Ibrahim ら[3]は，地図上で，各観測地点における詳細情報を適切に配置することで．多数の観測所からなる数値情報を，地理的特徴を保持しつつ，観察することが可能であることを示した．しかし，この手法では地点の位置関係は可視化されるものの，上流から下流への河川の流れや，支流から本流への合流に沿った情報を視覚的に捉えづらい．

2 つ目は二色塗り分けスパークラインに関する研究である．斎藤ら[4]は二色擬似カラー表示を用いて，二色の離散色で上部と下部を塗り分けることで，従来の折れ線グラフ等と比べてより狭いスペースでデータの変化の読み取り精度を向上させることが可能であることを示した．また，星谷ら[5]によって，気象データへの適用も有効であると示されている．擬似カラーのスケールとして，各色が示す値の幅が一定で読み取り精度も一定の線形スケール表示と，少ない色数で広い数値幅に対応できる Log スケールの 2 種類を目的に応じて使いわけている．

3 つ目は動的時間伸縮法 (DTW) に関する研究である．Müller[5]は，2 つの時系列データの各点に対して距離が最小となるように対応づけて，各々の対応点間の距離の合計値を DTW 距離と定義した．DTW 距離の値が小さいほど，2 つの時系列データが類似していることを示す．

3. 提案手法

3.1. 降水量データの補間

一般的に降水量データは気象庁のアメダスによって観測されているのに対して，河川水位データは国土交通省所管の各自治体の整備局の河川部の管理する水位計によって観測されている．その結果，降水量データと河川水位データは異なる位置で計測される．そこで，河川水位観測地点における降水量を補間により求める．具体的には，3 箇所の降水量観測地点p_0，p_1，p_2に囲まれた河川水位観測地点pにおける降水量を，線形補間で求める．

降水量観測地点$p0$，$p1$，$p2$における既知の降水量の各値をC_0，C_1，C_2とおく．それらに囲まれたある点pの局所座標系上での位置座標u_p，v_pを以下の式で求める．

$$p = p_0 + u_p(p_1 - p_0) + v_p(p_2 - p_0)$$

点pにおける降水量Cを，3 角形要素内の線形補間を利用して，以下の式で求める．

$$C = C_0 + u_p(C_1 - C_0) + v_p(C_2 - C_0)$$

3.2. 二色塗り分けスパークライン

河川水位観測地点における河川水位データと，補間で求めた降水量データの表示方法として，本手法では二色擬似カラー表示(図 1)を採用する．連続色による擬似カラーを使用することで，データの時間変化の読み取りが容易になる．柴田ら[7]は，異なる気象データの推移情報に対して，同じ配色を使用すると，異なるデータであることが認識しづらいことを示した．

河川水位データには線形スケール表示を適用する．各色が示す値の幅は一定であるため，読み取り精度も一定となる．それに対して降水量データには，少ない色数で広い数値幅に対応できる Log スケールを適用する．Log スケールを利用することで，稀にしか登場しない値の大きなデータにつぶされて，平常時の小さなデータの変化が読み取りづらくなる，という問題が生じるのを防ぐことができる．なお我々の実装では 5 色で二色擬似カラー表示を実装している．

本手法では，河川水位データの線形スケールは以下のように定義する．河川水位データは観測地点によって，平常時の水位が異なる．そこで，ある地点における 7 日間にわたる河川水位データのうちの最小値をw_{min}，最大値をw_{max}とし，以上の式によって数値区間の境界w_0～w_4を観測地点ごとに設定する．

$$w_o = w_{min}, \quad w_4 = w_{max}$$
$$w_{(i+1)} = w_i + t(i = 1,\ 2,\ 3)$$
$$t = \frac{(w_{max} - w_{min})}{4}$$

また，$[w_m,\ w_{m+1}]$に割り当てられた色をC_mとする．ここで，w_4以上の範囲に割り当てられた色をC_5とする．ある地点の任意の時刻の値wが$[w_m,\ w_{m+1}]$の範囲にあるとき，上部 $\frac{w_{m+1}-w}{t}$を色C_{m+1}で，下部 $\frac{w-w_m}{t}$を色C_mで塗り分ける．

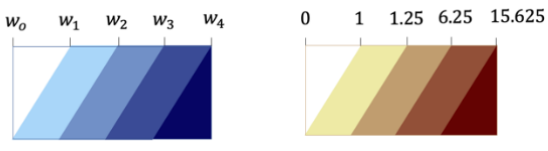


図 1　二色擬似カラー表示

3.3. 動的時間伸縮法

我々自身による先行研究[1]からの進展として本報告では，複数の観測地点における時系列データの類似度の算出方法として，動的時間伸縮法 (DTW) を採用した処理について論じる．DTW は 2 つの時系列データの各点の距離を総当たりで求めて，2 つの時系列が最短となるパスを特定する．そして，2 つの時系列データ間の各対応点間の距離を求め，その合計値を DTW 距離とする．DTW 距離が近い波形は類似度が高く，距離が遠い波形は類似度が低いと考える．

類似度を算出する 2 つの時系列データを$S = s_1, s_2 \ldots s_m$および$T = t_1, t_2 \ldots t_n$ とし，2 つの時系列データにおける各点を結んだワーキングパスを$W = w_1, w_2 \ldots w_k$と設定する．各要素の距離$\delta(i,j) = |x_i - y_j|$と定義すると，DTW 距離は以下の式で求める．

$$DTW(S,T) = \min_W \sum_{k=1}^{P} \delta(w_k)$$

4. 実行結果

4.1 使用したデータ

本章では，気象庁によって新潟県 43 箇所，長野県 46 箇所の計 89 箇所において観測された 1 時間ごとの降水量データおよび，新潟県と長野県の河川部によって信濃川水系において観測された 1 時間ごとの河川水位データを適用した結果を示す．なお，雨の振り方の違いによる降水量と河川水位の可視化結果の違いの比較の例として，2004 年 7 月 10 日から 16 日までと 2022 年 8 月 1 日から 7 日までにかけて，いずれも 7 日間にわたる 2 つの期間におけるデータを利用した．降水量の補間のために，1 時間ごとに降水量を観測している新潟県 43 箇所，長野県 46 箇所の計 89 箇所に対して，各地点における緯度と経度を座標に設定して，ドロネー三角分割（図 2）を適用した．三角形を構成する降水量観測地点p_0，p_1，p_2に囲まれた河川水位観測地点pにおける降水量を，線形補間で求めた．

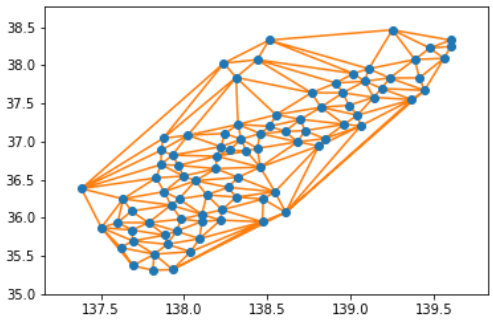


図 2　観測点を頂点としたドロネー三角分割

4.2 二色塗り分けスパークライン

例として，2022 年 8 月 1 日から 7 日にかけて，新潟県の信濃川水系内の臼井橋で観測されたデータを二色塗り分けスパークラインで表示する．上部が降水量データ，下部が河川水位データを示している（図 3）．色が濃くなっている段階において，データの数値が極端に上昇していることがわかる．このことから，二色塗

り分けスパークラインより，臼井橋において，降水量が増加してから水位が上昇するまでに時間差が発生していることが読み取れる．

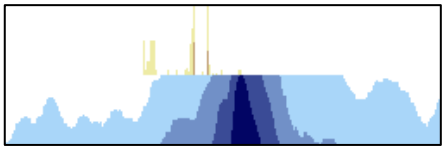


図 3　臼井橋における二色塗り分けスパークライン
（上部：降水量データ　下部：水位データ）

4.3 動的時間伸縮法

各観測所において，水位の時系列データから河川局によって定められた平常時水位の値を差し引いて求めた時系列データを作成する．2 つの観測所における同期間の作成した時系列データ間の類似度を，動的時間伸縮法を利用することで求める．

例として，2022 年 8 月 1 日から 7 日にかけて，新潟県の信濃川水系内の 2 つの水位観測所，長岡と保明新田において観測された河川水位データに対して，動的時間伸縮法を適用する．この処理により，2 つの時系列データの各点におけるパスの距離の合計が最短となるパスを見つける（図 4）．2 つの時系列データにおいて，データのピーク時のような特徴的な波形が対応していることがわかる．

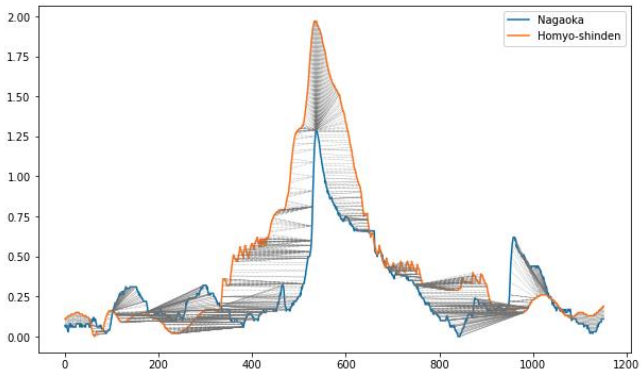


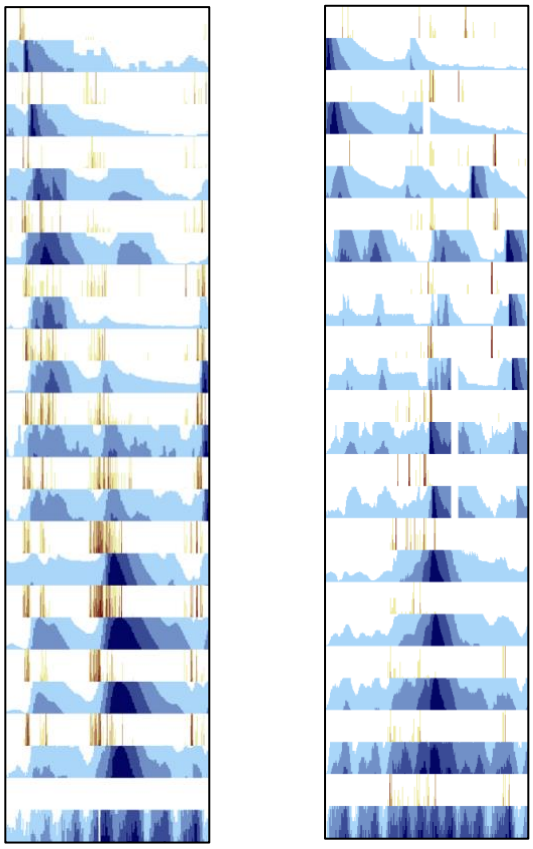
図 4　長岡と保明新田の 2 か所の水位データへの
動的時間伸縮法の適用

4.4 スパークラインの配置

計 28 箇所の降水量データと河川水位データを描画した二色塗り分けスパークラインを，河川の流れに沿って配置する．ここで，

- 河川本流に設置された観測所におけるデータのみを表示する(図 5).
- 全観測所におけるデータを表示する（図 6）．
- 全観測所におけるデータの中から，類似度の高いデータ（2 カ所の水位データに関する DTW 距離の合計が 100 以下）を 1 つにまとめたデータを表示する（図 7).

の 3 つのモードをユーザが選択可能とする．これにより表示する情報量を調整することで，ユーザが求める情報を適切に読み取ることができる．



2004 年 7 月
10 日～18 日

2022 年 8 月
1 日～7 日

図 5 河川本柱に設置された観測所のデータ表示

今回の信濃川における降水量と河川水位の関係を可視化し，全観測所におけるデータを表示することで，以下の情報を読み取ることができた．

- 下流になるにつれて，一旦河川水位が上昇した後，水位が低下するまでに時間がかかる傾向にある．
- 信濃川本流においては，降水量が増加してから，河川水位が上昇するまでに時間差が発生する傾向にあるものの，支流である魚野川（登川）と三国川においては，その時間差が殆ど確認されない．

続いて，2 つの期間における可視化結果を比較する．2004 年 7 月 12 日夜から 13 日にかけて，停滞する梅雨前線の活動が活発化して，新潟県において一時的に非常に激しい雨が降った．　また，2022 年 8 月 3 日から 4 日にかけて，下越において，線状降水帯による非常に激しい雨が同じ場所で長時間降り続いた．

- 長時間雨が降り続いたとき，水位が低下するまで

に時間がかかる．

・ 一時的に非常に激しい雨が降ったとき，その地点から離れた下流の観測地点まで水位増加の影響を受けやすい．

別の要因で雨が降った 2 つの期間の可視化結果より，これらの情報を読み取ることができた．

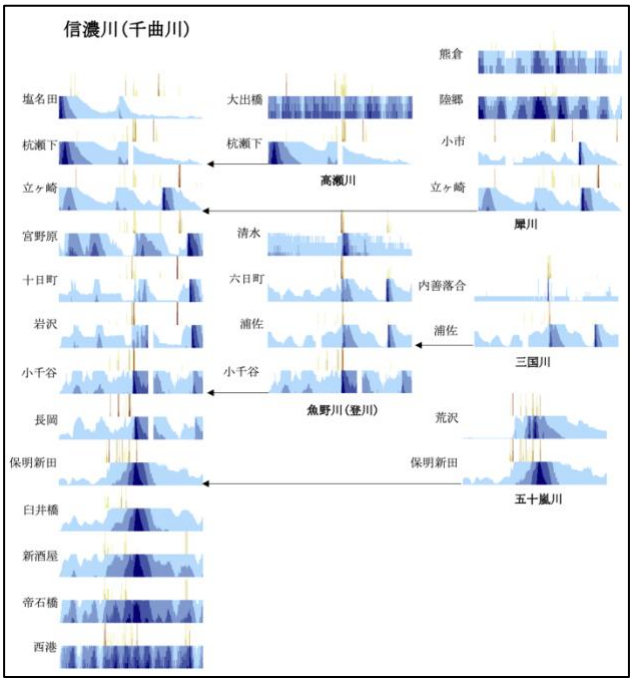


図 6 全観測所におけるデータ表示

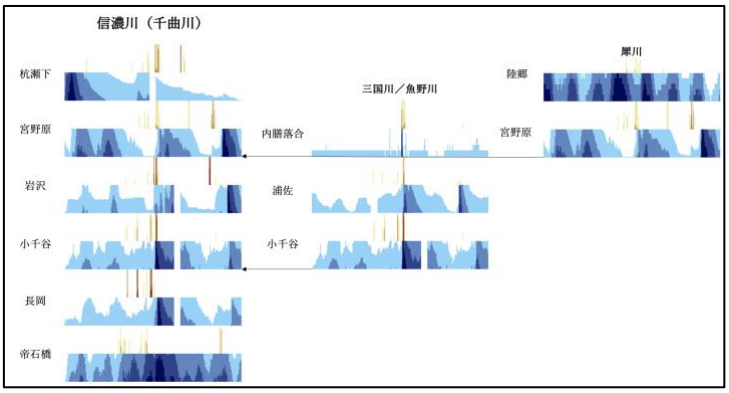


図 7 図 6 のデータ表示から
類似度の高いデータを削減したデータ表示

5. まとめと今後の課題

本研究では，河川水位の観測地点における降水量を補間により求め，それらの地点における河川水位と降水量を二色塗り分けスパークラインで描画し，河川の流れに沿って配置する可視化手法を開発している．この手法により，多数の観測所における河川水位と降水量の時間変化の関係を観察しやすくする．また，我々自身による先行研究からの進展として，本報告では，動的時間伸縮法を適用することで複数の時系列データにおける類似度を求めて，近隣観測所間で数値の類似度が高い場合は，そのうちの 1 箇所のみの表示する処理を提案した．これにより，ユーザが一度に可視化するデータの情報量を調整可能することが可能になった．

2 色塗り分けスパークラインは，連続的なデータである河川水位データの推移の可視化に適している．一方で不連続なデータである降水量データの推移の可視化において，一時的に非常に強い雨が降った時間における 2 色塗り分けスパークラインの幅が狭く，変化が読み取りづらいという問題点が残っている．このような欠点を解消するために，生データではなく差分をとったデータを活用したり，散布図などの他の可視化手法を検討したりするなどの余地がある．

本手法の最終的な目標として，観測された過去のデータをもとにして，雨がおさまっても水位が下がるのに時間がかかる地点を示すなどの形で，洪水の可能性予測を支援する可視化システムの開発を検討している．

参 考 文 献

[1] 田中悠来，伊藤貴之，降水量と河川水位の関係の可視化，可視化情報シンポジウム 2023, OS12-5, 2023.

[2] 八木佐也香，伊藤貴之，黒川真由美，伊豆裕一，米山貴久，小原隆志，“地図上の水防災データ可視化における画面配置と詳細度制御”，第 25 回人工知能学会全国大会，2011.

[3] Ibrahim Demir, Witold F. Krajewski, “Towards an integrated Flood Information System: Centralized data access, analysis, and visualization”, Environmental Modelling & Software, Vol. 50, Pages 77-84, 2013.

[4] Takafumi Saito, et al., “Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data”, IEEE Symposium on Information Visualization., 2005.

[5] 星谷有香，山本光義，斉藤博貴，斎藤隆文，宮本（中村）浩子，“二色塗り分け法による大規模気象観測データの可視化”，第 67 回情報処理学会全国大会，217-218，2005.

[6] M Müller, “Dynamic time warping”, Information Retrieval for Music and Motion, 69-84, 2007.

[7] 柴田直道，北直樹，斎藤隆文，“気象データ可視化における二色塗り分け擬似カラー表示の配色による考察”，第 84 回情報処理学会全国大会，237-238，2022.

47都道府県単位のデータ比較可視化による ジェンダーギャップの要因解析

石戸谷由梨† 伊藤 貴之†

† お茶の水女子大学理学部情報科学科 〒 112–8610 東京都文京区大塚 2–1–1
E-mail: †{g2120505,itot}@is.ocha.ac.jp

あらまし 本研究は，文化的・社会的な性差による影響やその原因を都道府県ごとのデータを用いて分析した．既存研究では，都道府県版ジェンダー・ギャップ指数が算出されているが，そこでは不十分であった地域の特色を加味した分析を行った．結果，男女共同参画関連の政策は地方ごとの傾向があり，東京近郊の就業している女性が多い地方においては，九州や四国と比べ育児などのライフイベントに合わせた政策ではなく，企業内での地位的な差の是正の傾向が強いように思われた．また，教育に関しては地方ごとの産業と進学先の関連性があったが，産業以外の項目とでは都道府県ごとに特徴にばらつきがあり，都道府県ごとの特徴が出るデータとの相関を調べること でより詳細な要因を探ることができると考えた．

キーワード 情報可視化，ジェンダーギャップ，都道府県別，地域性

1 はじめに

若い女性の地方から都会への流出が近年拡大しており，その原因としてジェンダーギャップがあげられている．具体的な要因としては，都会の会社の方が女性にとってキャリアが積みやすい点が多くあげられており，それを受けて地方の企業が働き方の改革に取り組んでいる．その一方でもう 1 つの要因として,「夫は仕事，妻は家庭」という従来の役割分担から，共働きの夫婦間で女性の負担が増えているという問題が挙げられている [1].

そこで，後者が地方のジェンダーギャップにおける本質的な問題であると仮定し，47 都道府県のジェンダーギャップ [2] に起因すると予測されるデータの地域差を可視化することで，ジェンダーギャップ解消の手段を検討できるのではないかと考えた．

本報告では，伊藤らが開発した可視化手法 Hidden [3] を用いて，47 都道府県の基本統計，労働・経済，政治，教育，の各データにおける地域差を可視化した結果を示すとともに，そこから導かれるジェンダーギャップ解消の手段を議論する．

2 データ収集と可視化手法

今回の調査では，伊藤らが開発した高次元データ可視化手法 Hidden を用いて，47 都道府県分の計 30 項目のデータを分析した．以下，データ収集方法と可視化手法について論じる．

2.1 データ収集

本研究で使用した主要なデータは以下の通りである．各データは，研究の目的に沿った詳細な分析と理解を深めるために選択された．

政治的指標：知事・市区町村長の女性首長数，都道府県議会の女性議員数，市区議会の女性議員数，町村議会の女性議員数，地方議会の女性議員数，女性がいる議会数の割合，SDGs を推進している自治体の割合，自治体内部における SDGs の普及啓発活動を推進していない自治体の割合. [4], [5].

人口と地理的データ：日本の人口，女性人口総数，男性人口総数，女性の人口の割合，女性の 65 歳以上の割合，他都道府県への転出数，他都道府県への転出数に占める女性の割合，東京都庁から各都道府県庁までの距離，市数，町数，村数. [6], [7], [8], [9].

労働市場関連指標：高齢化率，労働力人口，就業者，女性の労働力率（15〜64 歳），女性の労働力率（25〜44 歳），職業従事者総数，管理的職業従事者数，専門的・技術的職業従事者数，事務従事者数，販売従事者数，サービス職業従事者数，農林漁業従事者数，生産工程従事者数，輸送・機械運転従事者数，建設・採掘従事者数，運搬・清掃・包装等従事者数，分類不能の職業数. [10], [11], [12].

教育関連指標：女性の 4 年生大学へ進学した割合，女性の短期大学へ進学した割合，短大・通信・高校専門進学者の女子率，女性の小学校の校長の割合，女性の中の校長の割合，女性の教育委員会事務局の管理職の割合，4 年制大学入学者に占める女子の理学分野入学者割合，4 年制大学入学者に占める女子の工学分野入学者割合. [13], [14].

その他の指標：魅力度ランキング，託児所数，利用交通手段の総数，利用交通手段が 1 種類，利用交通手段が 2 種類，食品・飲料・たばこ製造業，繊維・衣服・革製品製造業，非金属鉱物製品・金属製品・機械器具製造業，電気機械器具・電子機器・輸送機器製造業，その他の製造業. [15], [16], [17], [18].

これらのデータは，日本におけるジェンダー格差の現状を客観的に評価するために不可欠である．特に，女性の進路選択や，理系進学者の少なさ，賃金格差の要因を解析する指標として，上記の統計は重要な役割を果たすため今回の分析データに導入した．

なお，このデータにおいて・日本の地方区分と東京からの距離を，表 1,2 に示すカテゴリ変数で示している．また，右図は表 1，左図は表 2 を使用して，データをカテゴリに分類して色

分けしている．

表 1: カテゴリ変数 1: 地方区分

1	全国	全都道府県の平均
2	北海道	北海道
3	東北	青森県，岩手県，秋田県，宮城県，山形県，福島県
4	関東	茨城県，栃木県，群馬県，埼玉県，千葉県，東京都，神奈川県
5	中部	新潟県，富山県，石川県，福井県，山梨県，長野県，岐阜県，静岡県，愛知県
6	近畿	三重県，滋賀県，京都府，大阪府，兵庫県，奈良県，和歌山県
7	中国	鳥取県，島根県，岡山県，広島県，山口県
8	四国	徳島県，香川県，愛媛県，高知県
9	九州	福岡県，佐賀県，長崎県，熊本県，大分県，宮崎県，鹿児島県
10	沖縄	沖縄県

表 2: カテゴリ変数 2: 東京都庁からの距離

1	距離が 0
2	距離が 250km 未満
3	距離が 250km 以上 500km 未満
4	距離が 500km 以上 1000km 未満
5	距離が 1000km 以上

2.2 可視化手法

本研究では，伊藤らが開発した高次元データ可視化手法 Hidden [3] を採用する．Hidden は高次元データを構成する各次元の中から，高い相関を有する次元グループを抽出し，各グループを平行座標プロットにより可視化する．本研究では上述のデータを構成する各変数の中から相関の高い組み合わせを発見し，それをもとにしてジェンダーギャップの要因を議論するという目的から，この可視化手法を分析ツールとして採用している．

なお，次章以降で示す可視化結果は，あくまでも変数間の相関を示すものであり，両変数間に因果関係があることを必ずしも示すものではない．言い換えれば，次章以降で示す可視化結果の中には，疑似相関と思われる結果が混じっている可能性がある点に注意されたい．

3 論点 1: 基本統計

本章では，人口統計などの基本的な統計データを用いて，性差に関して包括的に議論する．具体的には，これらのデータセット間の相関を探り，回帰分析を通じて性差のパターンを可視化し，それらが社会的，経済的な文脈でどのように現れているかを探求する．

3.1 女性人口

図 1 に示される色分けされた線に関して，カテゴリー 1 の各地域ごとに類似した傾向が観察される．具体的には，東京から距離の離れた地域（都道府県）ほど，各都道府県の人口に占める女性と高齢女性の割合の両方が低くなることがわかる．カテゴリー 2 の図からも，東京都周辺の女性人口の割合の低さが顕著に表れている．しかし地域ごとのデータセットを詳細に分析すると，弱い負の相関があり，各地域内でも都道府県ごとに女性人口全体に占める高齢女性の割合に差があることがわかる．具体的には，九州地方（右図，ピンクの線）において，この傾向が顕著である．

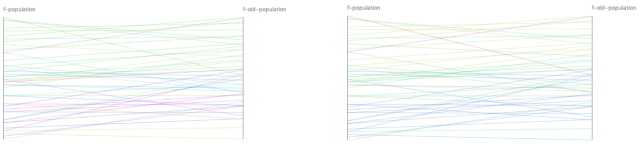


図 1: 女性人口の割合と 65 歳以上の女性の割合

3.1.1 女性人口と産業

図 2 より，女性人口比率と電気機械器具・電子機器・輸送機器製造業の数は負の相関関係を示している．これは，電気機械器具，電子機器，輸送機器製造業の数が少ないほど，人口に占める女性の割合が高くなり，九州や四国地方で特にこの傾向が顕著に現れていることが確認される．電気機械器具・電子機器・輸送機器製造業が東京近郊に集まっており，そのような職業に就こうとする男性が流入していったことが要因として考えられる．

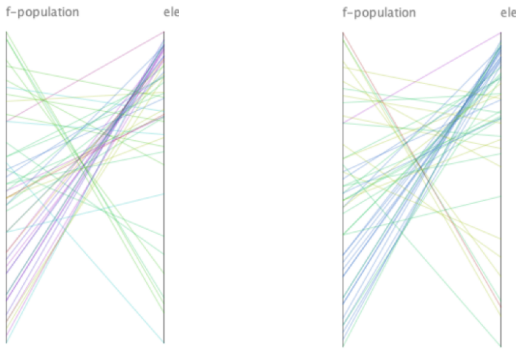


図 2: 女性人口の割合，電気機械器具・電子機器・輸送機器製造業数

3.2 女性の人口転出

図 3 より，男性の他都道府県への転出数と女性の他都道府県への転出数は正の相関関係を示している．ここから，男性の他都道府県への転出数の増加が増加するにつれて，女性の他都道府県への転出数も同様に増加する傾向が観察される．また，グラフの線が比較的平行に並んでいることから，転出数における性差が少ないことが確認される．市数，交通手段の総数，託児所数，女性の市議会議員の数もそれぞれ他都道府県への転出数と正の相関関係があることが確認される．これより，人口が大

きく，多様な交通手段を確保することができる都市では，託児所の確保による育児サポートや女性の参政がより進んでいる傾向がみられる．また，この傾向はカテゴリー 2 で分類されたグラフをみると，東京近郊の都道府県に多く確認される傾向であることがわかる．

図 4 より，市数と他都道府県への転出数に占める女性の割合には弱い負の相関関係が確認される．これは，市数が少ないほど，他都道府県への転出数に占める女性の割合が低くなることを示しており，この傾向は東京から距離の離れた九州や四国地方で現れていることがわかる．これより，人口が大きく，多様な交通手段を確保することができる都市は，女性の転出が少なくなる傾向があることが確認される．

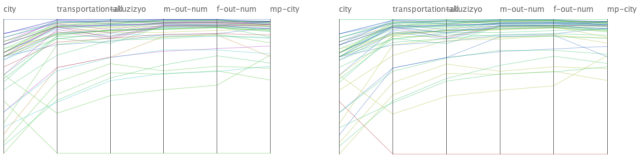


図 3: 市数，交通手段の総数，託児所数，男性の他都道府県への転出数，女性の他都道府県への転出数，女性の市議会議員の数

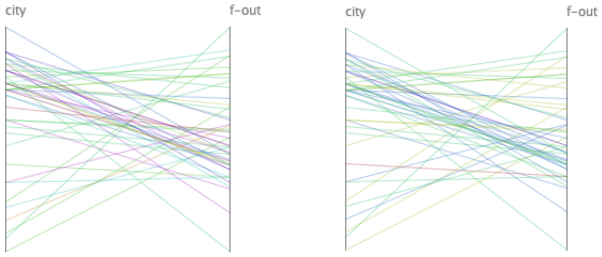


図 4: 市数，他都道府県への転出数に占める女性の割合

3.3 自治体数

図 5 の左図より，市数と町数は弱い負の相関関係を示している．また，右図より，村数と交通手段の総数も弱い負の相関関係を示しており，3.2 節で示した，市数の多さが育児サポートや女性の政治参画に多くの影響を与えていることが示唆される．



図 5: カテゴリー 1 の左図は SDGs の推進自治体の割合，市数，町数，右図は 65 歳以上の女性の割合，女性人口の割合，SDGs の推進自治体の割合，村数，交通手段の総数

4 論点 2: 労働・経済

本章では，各都道府県における女性の労働力率や職業区分別就業者割合を分析する．この分析では，可視化結果から相関のある要因を導き出し，労働や経済活動におけるジェンダーギャップの要因と女性のキャリア選択への影響を明らかにする．最終的に，女性の雇用に関する現状の理解を深め，性差を縮小する方法について検討する．

4.1 女性の労働力率

労働力率とは 15 歳以上の人口において，就業者と完全失業者が占める割合を示している．図 6 より，関東や中部地方の東京周辺の都道府県に関して，特に女性の労働力率（15〜64 歳）が高い値を示している．

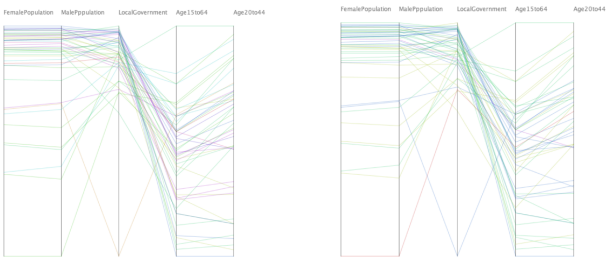


図 6: 女性人口総数，男性人口総数，自治体数，15〜64 歳の女性の労働力率，20〜40 歳の女性の労働力率

4.1.1 女性の労働力率と都市統計

図 7 より，都道府県ごとの市数と交通手段の数との間に正の相関が見られる．これは，市数の増加が交通手段の数の増加と関連していることを示している．また，交通手段の数と 15〜64 歳の女性の労働力率との間に負の相関が観測される．これは，都道府県内の交通手段が増加すると，女性の労働力率が低下する傾向があることを示している．色分けされた線に注目すると，地域ごとのまとまりや東京からの距離ごとのまとまりはなく，市数，交通手段の数，労働力率には地域性がないことが観測される．

図 8 より，交通手段の数と女性人口の割合の間に負の相関が見られた．交通手段の数が多いほど，女性の人口の割合が低いことを意味する．前述の「都道府県内の交通手段が増加すると，女性の労働力率が低下する」という傾向は，女性人口が少なくなることが作用しているとわかる．

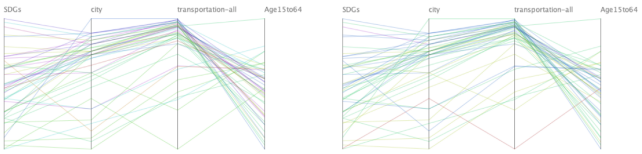


図 7: SDGs の推進自治体の割合，市数，交通手段の数，15〜64 歳の女性の労働力率

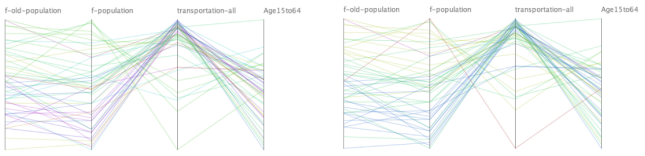


図 8: 女性人口の割合，65 歳以上の女性の割合，交通手段の数，15〜64 歳の女性の労働力率

4.2 職業区分別就業者の女性率

職業区分別就業者割合とは，各職業の就業者数が全職業の合計就業者数に占める割合を示しており，今回は就業者数に占める女性の割合を職業区分別に求めたデータを使用している．図 9 の分析結果によれば，管理的職業従事者に占める女性の割合と販売従事者に占める女性の割合の間には弱い負の相関が見られる．これは販売職の女性の就業者数が増加するにつれて，管理職の就業者数がわずかに減少する傾向があることを示唆している．

また，通常女性の人口の割合が多いほど，管理的職業従事者に占める女性の割合も多くなると考えられるが，図 9 からはそのような結果は見受けられない．特に目立つ結果として，中部地方は販売従事者に占める女性の割合は軸の中央に分布していたのに対し，管理的職業従事者に占める女性の割合では下部に分布している．

図 10 より，色分けされた線からは，全職業の就業者割合に関して，地域ごとのまとまりや東京から距離と比例していることが観測される．これは，人口に占める女性の割合も同様の傾向を示すことより，女性の人口が女性の就業率に大きく作用していることがわかる．この結果は，先ほどの管理的職業従事者に占める女性の割合や販売従事者に占める女性の割合とは異なった結果を示している．

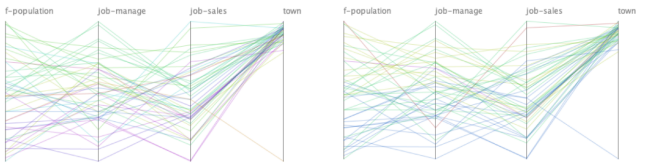


図 9: 女性人口の割合，管理的職業従事者に占める女性の割合，販売従事者に占める女性の割合，町の数

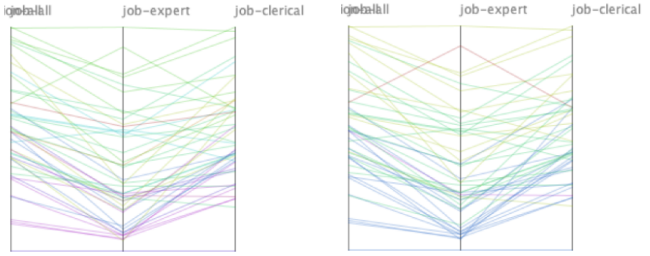


図 10: 全職業の女性就業者割合，専門的・技術的職業従事者に占める女性の割合，事務従事者に占める女性の割合

4.2.1 女性の職業区分別就業者割合と都市統計

図 11 より，交通手段の数と専門的・技術的職業従事者に占める女性の割合は負の相関を示していることがわかる．これは交通手段の数が減少すると，女性の専門的・技術的職業従事者に占める女性の割合が増加することを示している．これは 3.1.1 節で示した，電気機械器具・電子機器・輸送機器製造業が東京近郊に集まっており，そのような職業に就こうとする男性が交通手段が比較的多い東京近郊に流入している可能性が考えられる．

図 12 より，交通手段の数と女性の全職業の就業者に占める割合はどちらも負の相関関係にあることも観測される．3.2 節にて，交通手段の数と女性の他都道府県への転出数が正の相関関係を有することを述べているが，このことから転出が少ない都道府県の方が就業率が高いことがわかる．

また，東京近郊の地域や中部地方において，女性の就業率は中程度の値であった都道府県に関しては，女性の労働力率が特に高い結果となっている．これは，労働力率に就業以外に家事の傍ら仕事をしている者，通学のかたわらで仕事をしている者，休業者，完全失業者なども数値に含められており，女性の労働力率が特に高い都道府県は主婦や休業者，失業者が多いことが要因として考えられる．また，4.2 節からも中部地方などの東京近郊の地域は女性の販売職が多く，管理職が少ない特徴があり，夫となる男性は東京などの大都市に働きに出かけ，妻となる女性は自宅近くなどのシフトに融通がきく販売職についている傾向がある可能性がある．

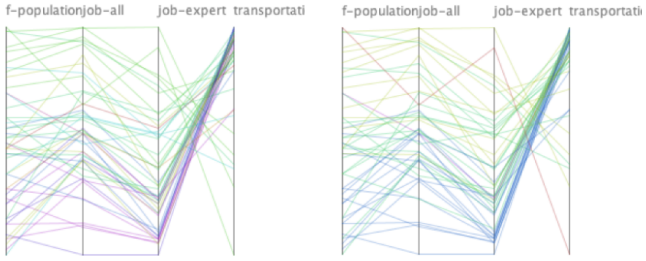


図 11: 女性人口の割合，全職業の就業者割合，専門的・技術的職業従事者に占める女性の割合，交通手段の数

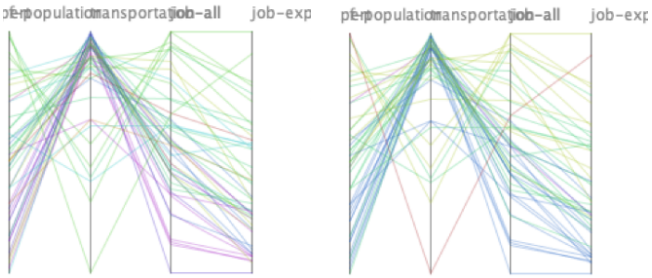


図 12: 全職業の女性就業者割合，専門的・技術的職業従事者に占める女性の割合，事務従事者に占める女性の割合

4.3 考　　察

女性の就業者数に関して，管理職に関しては，その他の職業とは異なり，人口に占める女性の割合に比例しないことが確認される．また，対照的に販売職は人口との相関や地域的な傾向が見られない．これは，女性にとって就業することよりも，就業後に組織内で管理職に昇進することの難しさに地域差があることを示している．特に中部地方など，首都圏に隣接する地域においてその傾向が強い．さらに，これらの地域においては，労働力率は高いが，就業率は低いという傾向もあり，非正規雇用者が多いことが伺える．

山極氏 [19] によると,「女性活躍を阻害する要因を明らかにするために，阻害要因を整理してみると，下記の 5 つの個別要素が浮上する．現在においても日本的雇用慣行が恒常的な長時間労働と固定的性別役割分担との対構造が常態化・常識化しているから，女性の多くは，仕事と家事・育児の二重負担や時間制約を抱えることになるのである」と述べている．その要素の 1 つとして「総合職」と「一般職」をベースに,「中間職」「専門職」「現業職」という雇用区分を多層化していることがあげられている．コース別雇用管理制度における，総合職採用者に占める女性割合は 3 割にも満たない．つまり，採用の時点で性差が大きく，女性の総合職への雇用が進まない．また，派遣業や非正規雇用などの雇用形態の多様化が進み，それらの昇進の見込めない職種は女性が多くを占めている．それ以外にも夫の昇進・昇格を優先する家族の意識や行動なども要素の 1 つとしてあげられており，従来の女性の役割 [1] が根強いことが示唆される．さらにこの現状が他都道府県への流出を助長していると考えられる．この現状に対して，性差を縮小する方法としては，女性が雇用において不利な状況に陥りやすい日本特有の雇用形態と，従来の役割分担を意識した職業選択が続いていることを自覚し，家事育児も考慮したワーク・ライフ・バランスが取れている雇用形態に変化することがあげられる．

5　論点 3: 政治

本章では，各都道府県における女性議員数や SDGs 推進率など，ダイバーシティに関連する政策にもとづくデータを分析する．この分析では，可視化結果から相関のある要因を導き出し，女性の政治参画が与える影響について洞察する．最終的に，女性の政治参画に関する現状の理解を深め，性差を縮小する方法について検討する．

5.1　女性の議員数

図 13 より，地方議会の女性議員数，市区議会の女性議員数，町村議会の女性議員数はそれぞれ，東京都庁から各都道府県庁までの距離と弱い負の相関があることがわかる．これは，東京から距離が遠いほど，女性の議員数は少ないということを示している．また，東京都庁から各都道府県庁までの距離を示す軸との関係を見ると，地方議会の女性議員数や市区議会の女性議員数，町村議会の女性議員数は同じような分布をしている．このことから，各地域における女性の政治参画が進んでいないことで，どの役職に関しても女性議員数が相対的に少なくなることがわかる．特に九州地方や四国地方では，人口に占める女性の割合が多いにもかかわらず，この傾向が顕著である．

また，3.2 節でも触れたが，女性の市議会議員の数は市数，交通手段の総数，託児所数，男女の他都道府県への転出数と正の相関関係があった．市などの規模が大きい自治体数が多い都道府県では，女性の政治参画が他の都道府県と比較して積極的であることがわかる．対照的に，九州地方や四国地方においては，市などの規模が大きい自治体数が少なく，女性人口が多く，転出数も少ない．にもかかわらず，女性の政治参画が進んでいないことを踏まえると，規模の小さい自治体，つまり議席数の少ない自治体が多い地域において女性の政治参画が進んでいないことが考えられる．

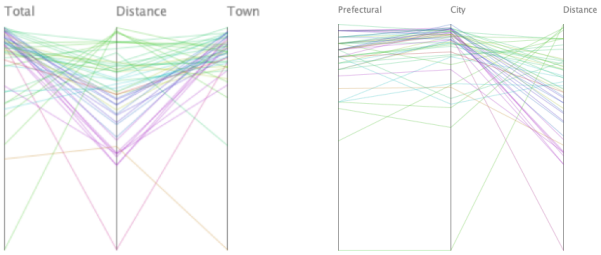


図 13: カテゴリー 1 の左図は地方議会の女性議員数 (合計)，東京都庁から各都道府県庁までの距離，町村議会の女性議員数，右図は都道府県議会の女性議員数，市区議会の女性議員数，東京都庁から各都道府県庁までの距離

5.2　SDGs の推進率

図 14 より，女性人口の割合と SDGs の推進自治体の割合には負の相関関係にあることも観測される．これは，女性人口の割合が多いほど，SDGs の推進率が低くなることを示しており，女性人口の割合が東京都からの距離に比例することを踏まえると，地方にかけて保守的な傾向が強くなることが言える．

また，SDGs の推進自治体の割合と自治体内部における SDGs 普及啓発活動を推進していない自治体の割合にも負の相関関係が認められる．これは，SDGs の推進自治体の割合が高いほど，自治体内部における SDGs 普及啓発活動を推進していない自治体の割合が低くなることを示している．さらに，自治体内部における SDGs 普及啓発活動を推進していない自治体の割合には，色分けされた線からは地域性があることが観測される．九州や四国などの地域では，SDGs の推進率も自治体内部における SDGs 普及啓発活動をしている割合も軸の中央から上に分布している．対照的に関東地方は，SDGs の推進率が特に高い上に，自治体内部での SDGs 普及啓発活動をしていない自治体の割合は特に低い結果となっている．

5.2.1　SDGs の推進率と就業率

図 15 より，SDGs の推進自治体の割合は全職業の女性就業者割合と負の相関関係にあることも観測される．また，図 16 からは，SDGs の推進自治体の割合は管理的職業従事者の割合

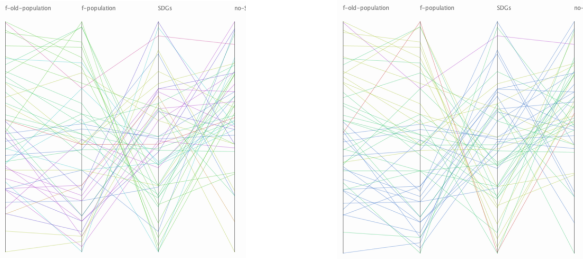


図 14: 65 歳以上の女性の割合，女性人口の割合，SDGs の推進自治体の割合，自治体内部における SDGs 普及啓発活動を推進していない自治体の割合

とも負の相関関係にあることがわかった．さらに色分けされた線より，地域性が顕著に現れている．特に，東京からの距離が 250km 未満の関東地方の都道府県に関しては，女性人口の割合が少なく，SDGs の推進自治体の割合は高いが，女性の全職業の就業者割合は際立って低い現象が現れている．図 14 も踏まえると，関東地方の都道府県では，SDGs を推進する政策には積極的であるが，意識的な改革やジェンダーに関連する政策の取り組みがあまり進んでいない可能性がある．

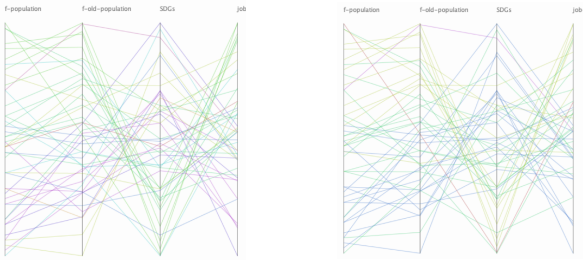


図 15: 女性人口の割合，65 歳以上の女性の割合，SDGs の推進自治体の割合，全職業の女性就業者割合

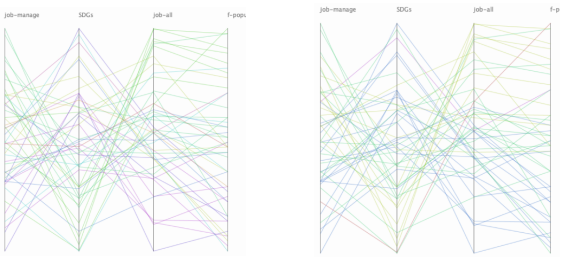


図 16: 管理的職業従事者に占める女性の割合，SDGs の推進自治体の割合，全職業の女性就業者割合，女性人口の割合

5.3 考　　察

女性の地方議会における議員数においては，そもそも議員数が少ないため，地域ごとの格差が激しい結果となった．特に，規模の小さい，議席数の少ない自治体が多い地域には，女性の政治参画が進んでいない傾向がみられる．これは，女性地方議員が少ない原因と考えられる理由として指摘されている「政治は男性のものという意識が九州地方や四国地方で強い [20]」という傾向につながっている可能性がある．さらに，議席数の少なさが，そのような傾向を助長している可能性がある．

また，これらの都道府県は交通手段や託児所が少ないという傾向があった．この傾向は，立候補から選挙期間中における課題（家事・育児・介護との両立が難しい等）や，現在の議員活動における課題（移動負担が大きい等）と深く関連していることが原因として考えられる．女性議員数を増やすためには，規模の小さい自治体における閉鎖感を解消することや，地域における住み心地をより良くしていくことが鍵と考えられる．また，地方自治体間のジェンダーに対する意識改革も推進していくことで，規模の小さい自治体の女性の参政を促進させる必要があると考えられる．

SDGs の推進率は人口に占める女性の割合や女性の就業率などと負の相関関係を示していた．これは，SDGs への取り組みに，ジェンダーギャップへの政策や，女性の声が反映されていないことを示唆するような結果である．内閣府地方創生推進室が作成した「SDGs 未来都市・自治体 SDGs モデル事業 事例集」[21] では，モデル事業が SDGs のどの目標に対する政策課が書かれており，その中でも 5 番目の目標であるジェンダー平等を掲げている事業は，2018 年度には，全 29 事業のうち 6 事業，2019 年度は 31 事業中 9 事業，2020 年度は 33 事業中 15 事業，2021 年度は再び 31 事業中 9 事業，2022 年度は再び 30 事業中 9 事業あった．5 年間を通して，事業数の増減は特になく，ジェンダーギャップ指数 [22] も 2018 年から 2022 年の間は常に 100 位以下である．以上により，日本における SDGs 政策において，ジェンダー平等への施策は，あまり重要視されていないか，あるいは効果が十分でないことがわかる．

6　論点 4: 教育

本章では，各都道府県における女性の校長の割合や 4 年生大学への進学者，理系進路選択の割合などを分析する．この分析では，可視化結果から相関のある要因を導き出し，教育におけるジェンダーギャップの要因と女性の進路選択への影響を明らかにする．最終的に，女性の教育に関する現状の理解を深め，性差を縮小する方法について検討する．

6.1　女性校長の割合

図 17 より，女性人口の割合と小学校の校長の女性率には負の相関がある．これは，女性人口の割合が増加すると，小学校の校長の女性率は低くなる傾向を示している．具体的には，関東や東京近辺の地域においては女性の校長の割合が高くなるのに対し，九州や四国においては著しく低くなっていることが観測される．中学校における校長の女性率には，特に地域ごとの特徴は観測されていない．加えて，図 10 などから，九州や四国などは，教師を含んでいる専門的・技術的職業従事者の女性比率は他の都道府県と比較して高い．これに対して追加で調査したところ，九州地方は学生数が少ないのにも関わらず，女性の中学校と高校の教師の数が他県と比べて多くなっていることがわかった．この背景として，高度な教育を受けた女性が就職先として都会に出たり，転勤などの可能性がある大きな企業等

を選ぶのではなく，比較的地元の近くで働ける教員という選択肢が積極的に取られている可能性が考えられる．さらに中学校と高校の女性教員が多い要因として，教員免許を取る際に，小学校の教員免許は取得単位数が多く，教育学部の教員養成課程での取得が一般的なため，教育学部以外の学部に進学して教員免許を取る学生の多くは中学校・高校の免許を選ぶ可能性が考えられる．またこの結果，小学校教育における男女格差が目立つ結果となっていることがわかる．

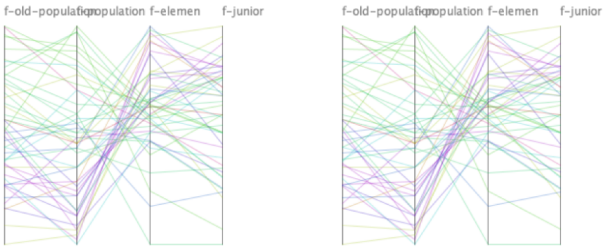


図 17: 女性人口の割合，65 歳以上の女性の割合，女性の小学校の校長の割合，女性の中の校長の割合

6.2 女性の 4 年制大学・短期大学進学率

図 18 より，4 年生大学に進学する女子率と小学校の校長の女子率にはわずかに相関関係が確認される．具体的に，九州地方に関しては，女性校長の割合が他都道府県と比べて少ないが，女性の 4 年生大学へ進学した割を見ると他都道府県と比べて高い値を示している都道府県が多い．しかし，九州地方のいくつかの都道府県は全都道府県と比較しても進学率が著しく低いなど，九州地方内での格差が大きく開いている．関東や中部地方に関しては，女性校長の割合が多いのにも関わらず，進学率が中程度やそれ以下の都道府県がある．これは，小学校の女性校長の割合が女性の 4 年生大学への進学率に与える影響がそれほどないことを示している．ただし，九州地方に関しては，女子の進学率における地域内格差が特に目立つ形となり，進学における何らかの障壁があることが予測される．また，東北地方に関しては，4 年生大学に進学する女子率も小学校の校長の女子率もどちらも低くなっている．

図 19 より，短大・通信・高校専門進学者の女子率と人口に占める女性の割合には弱い負の相関が確認される．女性の人口比率が少ない，東京に近い都道府県に注目すると，他都道府県と比べて短大・通信・高校専門進学者の女子率が高いことがわかる．さらに，4 年生大学に進学する女子率も比較的高く，女性の全体的な進学率が高い傾向にある．中部地方に注目すると，女性人口の割合も 4 年生大学に進学する女子率も中程度だが，短大・通信・高校専門進学者の女子率は低いという特徴が観測されている．また，東北地方に注目すると，先ほどの結果も踏まえ，全体的に女子の進学率が低い結果である．九州地方の都道府県は基本的に同じような傾向を示していたが，福岡県だけは九州地方の特徴から外れる値を示している．九州は全体的に女性の人口比は高い値を示しているが，福岡県だけは低く，SDGs の推進率も著しく低い．

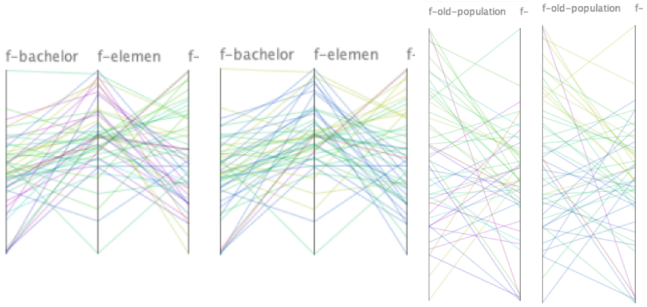


図 18: 左図は女性の 4 年生大学へ進学した割合，女性の小学校の校長の割合，65 歳以上の女性の割合，右図は 65 歳以上の女性の割合，女性の 4 年生大学へ進学した割合

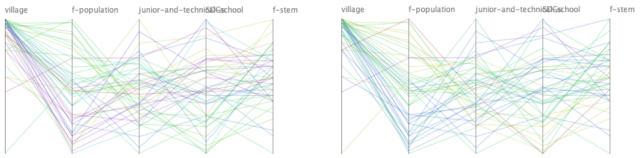


図 19: 村数，女性人口の割合，短大・通信・高校専門進学者の女子率，SGDs を推進している自治体の割合，4 年制大学入学者に占める女子の理学分野入学者割合

6.2.1 理学部進学率

図 20 より，4 年制大学入学者に占める女子の理学分野入学者割合と高齢女性比率は弱い負の相関関係を示している．これは，4 年制大学入学者に占める女子の理学分野入学者割合が高いと女性人口に占める高齢女性比率が低くなることを示している．また，カテゴリー 2 で色分けされている左図を見ると，高齢女性比率は東京から離れるほど大きく，女子の理学分野入学者割合は逆に小さくなることがわかる．

図 21 より，4 年制大学入学者に占める女子の理学分野入学者割合と町数は負の相関関係を示している．これは，女子の理学分野入学者割合が大きいほど，町の数が少なくなることを示している．また，カテゴリー 2 の左図を見ると，SDGs は東京からの距離が大きくなるほど推進率が上がるが，女子の理学分野入学者割合に関してはそのような特徴は見られない．



図 20: 自治体数，65 歳以上の女性の割合，4 年制大学入学者に占める女子の理学分野入学者割合

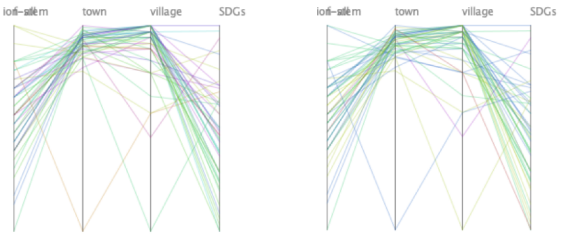


図 21: 4 年制大学入学者に占める女子の理学分野入学者割合，町数，村数，SDGs の推進自治体の割合

6.2.2 工学部進学率

図 22 より，女性人口比率と 4 年制大学入学者に占める女子の工学分野入学者割合は目立った相関は見られない．しかし，地域ごとに着目すると，四国地方に関しては，人口における女性比率が特に高いのにも関わらず，4 年制大学入学者に占める女子の工学分野入学者割合が他の都道府県と比べると最も低くなっている．その他の地域に関しては，全体的に軸の中央によっているため，都道府県間の差の方が目立っており，四国地方における 4 年制大学入学者に占める女子の工学分野入学者割合の低さが目立つ結果となっている．四国地方などの東京から特に距離が離れた地域以外の地域を見ると，東京からの距離が大きいほど女子の工学分野入学者割合がわずかに高い傾向にある．これは 4 年制大学入学者に占める女子の理学分野入学者割合とは対照的な傾向である．この傾向において九州地方は他都道府県と比べて女子の工学分野入学者割合が全体的に高くなっているが，その中でも福岡県が著しく低い値を示している．

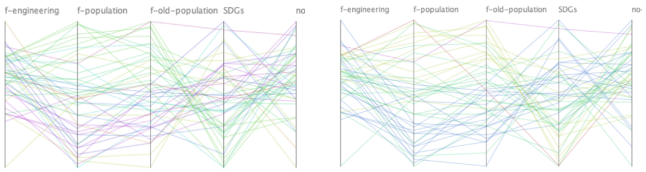


図 22: 4 年制大学入学者に占める女子の工学分野入学者割合，女性人口の割合，65 歳以上の女性の割合，SGDs を推進している自治体の割合，自治体内部における普及啓発活動を推進していない自治体の割合

6.3 考　　察

教育分野に関しては，地域ごとの格差が特に顕著であった．女性校長の割合に関しては，4.2.1 節より，専門的・技術的職業従事者に占める女性の割合は人口に比例する結果となった．それに対して校長の女性の割合は，人口に占める女性の割合と負の相関関係を示しており，教育分野における性差が大きいことがわかった．特に，九州地方や四国地方は，女性人口の比率が高いのに伴い，進学率における格差は地域内では激しいものの，多くの都道府県が 4 年生大学への女子の高い進学率を持っていたり，専門職に就く女性の割合も高かったりしているが，それでも校長の女性の割合に関しては著しく低い．飯島氏 [23] は，女性が管理職を志向しない理由として，家庭生活との両立の困難，自己評価の低さ，労働時間の長さ，根強い性別役割分担意識などをあげている．九州地方や四国地方にはこれらの要因が他の地域の都道府県と比べて深刻である可能性が示唆される．5.3 節でも述べている通り，これらの地域は県内の交通手段が少なかったり，託児所が少ないなどの課題がある．この点も踏まえると，地域内での環境を整えることや，子育てや家事を女性だけの生業とせずに新しい働き方に改革することなどが，重要な課題であると思われる．

進学率に関しては，進学者全体に占める女性の割合を使用したところ，地域ごとの大きな特徴は見られなかった．しかし，その中でも九州地方には著しく 4 年生大学に進学する女子率が低い地域がある．この傾向を分析するには，九州地方内でのより深い分析が必要と思われる．また，4 年制大学入学者に占める女子の理学分野入学者の割合は，東京から距離が離れるほど低くなり，対照的に，4 年制大学入学者に占める女子の工学分野入学者の割合は，四国地方を除いた地域に関して，東京からの距離が離れるほど高くなった．これらの結果に関しては，各地域における理学部や工学部の定員を比較するなど，より深い分析が必要と考える．

7 ま と め

本研究では，47 都道府県単位のデータを比較可視化することで，ジェンダーギャップの要因やそれに対する解消の手段を，労働・経済，政治，教育の 3 つの分野に対して議論した．どの分野にも共通して，日本の経済の中心地である東京からの距離と，それに伴った地域の住みやすさを支える県内移動のためのインフラの整備などは，女性の経済，政治，教育において地域ごとの格差を生んでいるということが示唆される．それを解消するためには，整備するための資金や，現在の各地域の状態が女性のケア労働を前提とした作りになっていることを自覚し，変えていくことが重要である．また，管理職や議員への立候補時にも，性別的役割分担が大きな影響を与えている点を自覚することが望ましいと考える．特に，規模の小さい議席数の少ない自治体でその傾向が強いのにも注意しなくてはならない．また，一都三県の周辺の都道府県に関しては，女性の経済進出が他都道府県と比べて少ない．これは，主に男性が東京などの少し離れた地域に出勤することで，主に女性が自宅に留まる生活を選ばざるを得ない可能性が考えられる．言い換えれば，必ずしも東京からの距離だけで性差を強くする要因を説明できるとも限らないことがわかる．

データを可視化する際に，特に教育分野に表れていた傾向として，地域ごとではなく，都道府県ごとの性質が顕著に表れていた．この現象に関して，地域ごとにより深い分析を進めて，さらなる性差の要因を分析していく予定である．

文　　献

[1] 十六総合研究所, “十六総合研究所 提言書 2022「女子」に選ばれる地方”, 2022. URL: http://www.16souken.co.jp/assets/202203/b58f6efc40c575f1aada1842f23d734611ff4fb7.pdf.

[2] 三浦まり，竹内明香，“都道府県版ジェンダー・ギャップ指数（2022 年版）：算出方法と結果分析”，2022，上智大学．URL: https://dept.sophia.ac.jp/econ/econ_cms/

wp-content/uploads/2022/12/DPNo.J22-2.pdf.
[3] T. Itoh, A. Kumar, K. Klein, J. Kim, High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots, Journal of Visual Languages and Computing, 43, pp. 1-13, 2017.
[4] 内閣部男女共同参画局, “女性活躍・男女共同参画における現状と課題”, 2023. URL: https://dept.sophia.ac.jp/econ/econ_cms/wp-content/uploads/2022/12/DPNo.J22-2.pdf.
[5] 内閣官房デジタル田園都市国家構想実現会議事務局 内閣府地方創生推進事務局, “SDGsに関する全国アンケート調査”, 2023, 自治体 SDGs 推進評価・調査検討会. URL: https://www.chisou.go.jp/tiiki/kankyo/kaigi/sdgs_enquete_chousa.html.
[6] 総務省統計局, “人口推計”, 2021. URL: https://www.stat.go.jp/data/jinsui/2021np/index.html.
[7] 総務省統計局, “住民基本台帳人口移動報告 2022 年（令和４年）結果の概要”, 2022. URL: https://www.stat.go.jp/data/idou/2022np/jissu/pdf/gaiyou.pdf.
[8] 国土交通省国土地理院, “都道府県庁間の距離”, 2018. URL: https://www.gsi.go.jp/KOKUJYOHO/kenchokan.html.
[9] 地方公共団体情報システム機構, “都道府県別市区町村数一覧”, 2018. URL: https://www.j-lis.go.jp/spd/code-address/kenbetsu-inspection/cms_11914151.html.
[10] 内閣府, “令和 4 年版高齢社会白書”, 2022. URL: https://www8.cao.go.jp/kourei/whitepaper/w-2022/zenbun/04pdf_index.html.
[11] 総務省統計局, “労働力調査”, 2023. URL: https://www.stat.go.jp/data/roudou/pref/index.html.
[12] 総務省統計局, “国勢調査 平成 27 年国勢調査 抽出詳細集計（就業者の産業（小分類）・職業（小分類）など）”, 2015. URL: https://www.e-stat.go.jp/dbview?sid=0003209865.
[13] 総務省統計局, “令和 3 年度　学校基本調査”, 2021. URL: https://www.e-stat.go.jp/stat-search/files?page=1&toukei=00400001&metadata=1&data=1.
[14] 三菱 UFJ リサーチ&コンサルティング, “女子生徒等の理工系分野への進路選択における地域性についての調査研究”, 2021, 令和３年度内閣府委託 調査. URL: https://www.gender.go.jp/research/kenkyu/pdf/riko_sentaku_research_r03.pdf.
[15] 株式会社ブランド総合研究所, “地域ブランド調査 2023　都道府県の魅力度等調査結果”, 2023. URL: https://news.tiiki.jp/articles/4854.
[16] 厚生労働省, “令和元年社会福祉施設等調査の概況”, 2019. URL: https://www.mhlw.go.jp/toukei/saikin/hw/fukushi/19/index.html.
[17] 総務省統計局, “国勢調査 令和２年国勢調査 従業地・通学地による人口・就業状態等集計”, 2020. URL: https://www.e-stat.go.jp/dbview?sid=0003454515.
[18] 総務省統計局, “工業統計調査”, 2019. URL: https://www.e-stat.go.jp/stat-search/database?page=1&layout=dataset&toukei=00550010&kikan=00550&statdisp_id=0003432912.
[19] 山極清子, “企業における女性活躍の阻害要因とその解決への道筋”, 社会デザイン学会学会誌 12 (0), 12-23, 2021. URL: https://www.jstage.jst.go.jp/article/socialdesign/12/0/12_12/_pdf/-char/ja.
[20] 内閣府男女共同参画局, 有限責任監査法人トーマツ, “政治分野における男女共同参画の推進に向けた地方議会議員に関する調査研究報告書”, 2018. URL: https://www.gender.go.jp/research/kenkyu/pdf/local-councilor_h29.pdf.
[21] 内閣府地方創生推進室, “SDGs 未来都市・自治体 SDGs モデル事業 事例集”, 2021. URL: https://www.chisou.go.jp/tiiki/kankyo/pdf/bessatsu1.pdf.
[22] 世界経済フォーラム, “Global Gender Gap Report 2023”, 2021. URL: https://jp.weforum.org/publications/global-gender-gap-report-2023/.
[23] 飯島絵理, “女性校長はなぜ少ないのか、少ないことはなぜ問題か”, 国立女性教育会館リポジトリ, 2020. URL: https://nwec.repo.nii.ac.jp/record/18871/files/%E7%AC%AC12%E7%AB%A0.pdf

SNS 上の投稿写真の位置情報と意味情報を考慮したピクトリアルマップの自動生成

周　文トウ†　　牛尼　剛聡††

† 九州大学大学院芸術工学府　〒 815–8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学院芸術工学研究院　〒 815–8540 福岡県福岡市南区塩原 4-9-1
E-mail: †zhou.wentong.904@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　本論文では SNS 上の投稿写真から地域の特徴を表すピクトリアルマップを自動的に生成する手法を提案する．提案手法では，ピクトリアルマップの品質を向上するため，従来手法との違いとして、主に２つの特徴を有する．（１）投稿写真の意味空間と地理空間で密度ベースのクラスタリングによるより人間にとって意味がある POI を抽出する．（２）クラスター内に写真間の意味的類似度を考慮して各 POI に地域の魅力を表す代表的な画像集を選出する．この２つ改善点により，生成されたピクトリアルマップはユーザーの多様な興味を反映される POI を抽出でき，地域の特徴を表す画像の適切さや一致性も向上することができる．

キーワード　画像認識，POI，Semantic Clustering，DBSCAN

1　はじめに

近年，スマートフォンの普及や SNS(Social networking service) の発展によって，ユーザーが目の前の風景を見た時，その瞬間を記録するため，写真に撮影し，SNS に投稿することが一般的行われるようになった．このような投稿写真は投稿者がその地点に対して興味や関心を反映されると言える．そこで，膨大な投稿写真から多くの人にとって地域の特徴や魅力を抽出できると考えられる．また，抽出した地域の魅力を直感的に理解・応用するため，可視化手法の１つとしてピクトリアルマップが広く使われている．

SNS の投稿写真を利用して，地域の特徴を反映されるピクトリアルマップを自動的に生成する手法が提案されている．多くの手法は主に２つステップに分けられる．（１）対象地域に代表的な地点を決定すること．（２）代表的な地点における地域の特徴を表す適切な写真を抽出すること．まず，（１）について，高ら [1]，また著者ら [2] は写真の位置情報を用いて密度ベースのクラスタリング手法 DBSCAN(Density-Based Spatial Clustering of Applications with Noise) を利用して投稿数が多いエリアを発見し POI(Point of Interests) [3] として抽出している．しかし，これらの手法では，写真の位置情報しか利用していなかったため，密度が高いかつ多様性のある場所に対応できないと考えられる．例えば，都市の駅周辺地区にはグルメスポットや撮り鉄スポットなど，複数のスポットが近接して存在し写真の密度が高い場合，位置情報のみを考慮すると，POI を一つしか抽出できない場合がある．この場合，一つ地域の多様な魅力をマップ上に反映できず，異なるユーザの興味も対応できないと考えられる．

上述の問題点に対して，本研究では，SNS の投稿写真の地理情報と意味情報の両方を考慮して構築された意味空間と地理空間でクラスタリングを行うことで，意味的な類似度が近いかつ地理的な距離が近い写真を同じクラスタに分ける．これにより，写真の密度が高い地域でもユーザの様々な興味を反映される異なる POI を抽出でき，生成されたピクトリアルマップの面白さや応用性を向上することができる．

次に，(2) について，我々のこれまでの研究では [2]，写真の内容を画像認識 API を用いて認識し，識別された視覚属性を表すタグの逆文書頻度 [4] を算出し特徴ベクトルに変換して，POI の特徴ベクトルをクラスター内に全て写真の特徴ベクトルの平均したものから算出している．そして，写真と POI 間の類似度を算出しランキングして，上位の写真を最も代表的な写真として抽出している．しかし，この手法では，認識されたタグの差別度を考慮し重みを付ける一方，写真の共通の特徴要素の重みを下げる．変換した特徴ベクトルは写真が撮られたシーンの全体の情報を伝えられないため，抽出された代表的な写真は適切ではない場合がある．さらに，クラスター内に写真の内容のバラツキが大きい場合，ランキングした写真間の一致性もない．



図 1　同じ商店街 POI に対して，従来手法（左）と提案手法（右）を用いて選出された画像集の比較

そこで，本研究では，同じく画像認識でタグ情報を取得したが，単独なタグの差別度を注目するではなく，画像認識 API で

統一的な基準で付けられた全てタグとタグの確信度を基にエンベディング処理を行い，特徴ベクトルに変換する．そして，写真の特徴ベクトル間の類似度を算出しそれぞれ POI の意味空間を構築する．それぞれの POI の意味空間内に写真間の意味的類似性を算出しランキングする．そうすることで，ランキングした上位写真グループ内に写真間の一致性を持つことができ，選出された最も代表的な写真の適切性も確保することができる．図 1 は同じ POI（札幌狸小路商店街）に対して，従来手法（左の図）と提案手法（右の図），それぞれを用いて選出された画像集の例を示す．

本論文の構成は以下の通りである．２章では SNS の画像を用いて地域の特徴を表すピクトリアルマップの自動生成についての先行研究を紹介する．３章では提案手法について述べる．4 章は提案手法に基づく生成されたピクトリアルマップの例について述べる．5 章は評価実験について述べる．6 章はまとめと今後の課題について述べる．

2 関 連 研 究

本章では，SNS 上の位置情報付き画像を用いてピクトリアルマップの自動生成に関する先行研究を紹介する．

Chen ら [5] は SNS 上の位置情報付き投稿写真を用いて地域の代表的なランドマークアイコンが付いている観光地図を自動的に生成する手法を提案している．まず，画像のメタデータの緯度経度と場所を表すタグの TF-IDF 値 (term frequency inverse document frequency) [6] を用いて K-means クラスタリングを繰り返し行うことで重要な POI を抽出する．次に，画像特徴量を基に canonical image selection algorithm を使ってそれぞれの抽出された POI に最適なランドマークビューを特定する．最後に，ランドマークビューの画像を基にアイコンを生成・レンダリングし地図に配置する．

高ら [1] は SNS の位置情報付き投稿写真を利用して，地域の特徴を図を用いて分かりやすく示したピクトリアルマップを自動的に生成する手法を提案する．まず，写真の位置情報に基づいて DBSCAN クラスタリングを行い，投稿数が密集するところを特徴的な地点として抽出する．次に，写真の色情報を利用して写真間の類似度を求める．そして求められた類似度に基づき，DBSCAN クラスタリングを行って，特徴的な地点を表現する画像を絞り込んで代表的な写真を決定する．

著者ら [2] はこれまでもピクトリアルマップを自動的に生成する手法を提案している．POI を抽出する手法は上述の高らの手法と同じく DBSCAN クラスタリングアルゴリズムを利用する．また，POI に代表的な写真を抽出する際に，従来手法のように画像の色特徴や局所特徴などを利用するではなく，画像認識で識別された人間が理解可能な視覚属性の情報を利用し分析する．

本研究では，まずより内容豊富な POI を抽出するため，画像の位置情報だけではなく，意味情報を用いて高次元の意味空間を構築する．意味空間上に DBSCAN クラスタリングをベースに拡張した Semantic-DBSCAN を提案する．Semantic-DBSCAN を用いて高次元空間上にユーザーの多様なニーズを捉える POI を抽出することを目指す．次に，地図に提示する地域の特徴を表す写真を決定するため，写真間の意味的な類似性を考慮し計算を行うことで，それぞれの POI に一致性がある適切な写真コレクションを選出することを目指す．

3 提 案 手 法

提案手法は「写真の意味情報と位置情報を考慮した POI の抽出」と「意味的類似する写真の絞り込む」の２つステップに分けられる．本章ではそれぞれについて述べる．

3.1 写真の意味情報と位置情報を考慮した POI 抽出

POI を抽出するため，SNS に投稿された画像の意味情報と地理情報から構成された高次元潜在空間内に Semantic-DBSCAN を用いてクラスタリングを行う．図 2 は POI を抽出するため手法の流れを示す．本節では手法の詳細を順に述べる．

3.1.1 画像認識による意味情報の取得

近年，膨大な画像の内容を効率的に分析するため，機械学習の多ラベル分類器（Multi-label classification）を利用することが一般的である．そして，画像データを高精度で多ラベル分類アルゴリズムが WebAPI として利用できるようになっている．本研究では，代表的な画像認識 API の一つである Image Analysis API [7] を利用して，人間が理解可能な画像の視覚属性を識別し，取得したマルチラベル情報を画像の意味情報として利用する．具体的に，Image Analysis API の視覚的特徴のタグ付け機能 (Image tagging feature) を利用する．この機能は，数千個の常識的認識可能なオブジェクトを識別してタグ付け，識別されたタグの確信度（confidence）も提示する．

- タグ付けの対象は，前景の被写体に限らず，背景なども含まれる．また，タグは分類として整理されず, 継承階層は存在しない．
- タグの確信度（confidence）はそれぞれクラスに属しているどの程度確信しているかを表し，0 から 1 までの範囲を満たす．

そこで，類似する写真は重複しているタグのセットを取得でき，類似する視覚属性を同程度の確信度を取得できると考えられる．タグ付けの例を図 3 に示す．左の写真はユーザーが撮った札幌テレビ塔のイルミネーションの写真である．Image Analysis API により，背景の「sky」,「night」から，被写体の「tower」や「building」, 都市風景を表す「illumination」, シーズンを表す「christmas」,「winter」まで識別された．多くのタグはユーザーが写真を通じで伝えたい意味情報と一致していると考えられる．本研究はタグと確信度の両方を利用し特徴ベクトルに変換する．次の節は特徴ベクトル化の手法について説明する．

3.1.2 画像の意味情報のベクトル化・類似度の計算

我々の以前の研究 [2] では，画像認識で得たタグの逆文書頻度を計算しベクトルに変換する．本研究ではタグの確信度を使ってベクトルに変換する．いま対象となる地域に二つ写真があるとし，取得したタグが合計７個 {“chritmas tree”，“illumination”，

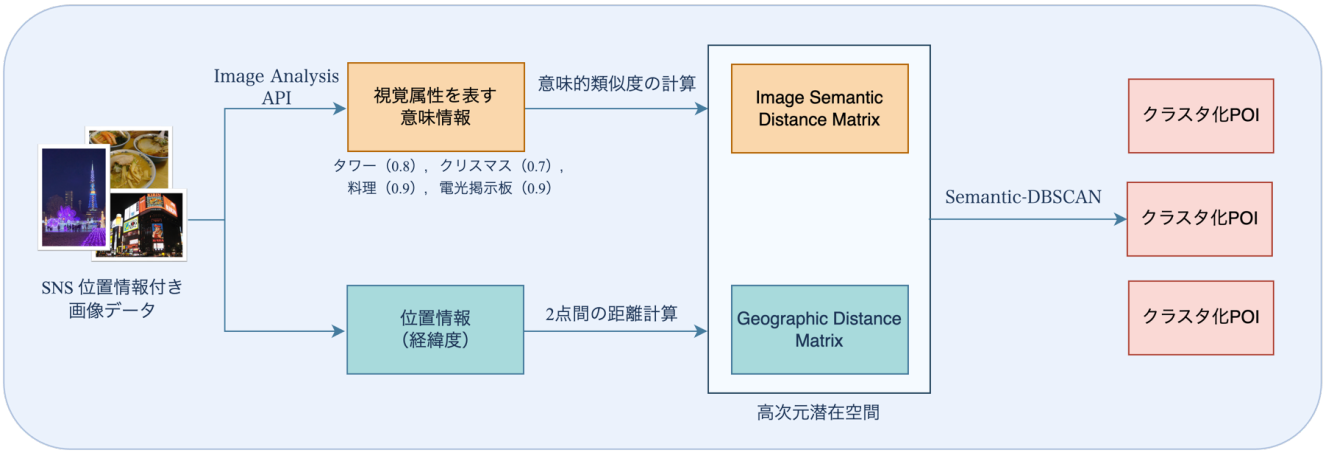


図 2　POI 抽出の流れ



"tags":
[{ "name": "sky", "confidence": 0.9833276271820068 },
{ "name": "christmas tree", "confidence": 0.9190281629562378 },
{ "name": "tree", "confidence": 0.906639575958252 },
{ "name": "skyscraper", "confidence": 0.8968532085418701 },
{ "name": "night", "confidence": 0.8859416842460632 },
{ "name": "building", "confidence": 0.879662275314331 },
{ "name": "outdoor", "confidence": 0.8638666868209839 },
{ "name": "tower", "confidence": 0.8498963117599487 },
{ "name": "illumination", "confidence": 0.8291453123092651 },
{ "name": "christmas", "confidence": 0.731196403503418 },
{ "name": "purple", "confidence": 0.6964696645736694 },
{ "name": "light", "confidence": 0.6486214399337769 },
{ "name": "winter", "confidence": 0.4304569959640503 }]

図 3　Image Analysis API による取得した結果の例

“landmark”, “light, “purple”, “sky”, “skyscraper”} であった場合を例に考える．図 4 を二つ写真 A と B，それぞれ取得したタグと確信度を示す．



タグ	確信度
sky	0.9
christmas tree	0.9
skyscraper	0.8
illumination	0.8
purple	0.6



タグ	確信度
sky	0.9
landmark	0.9
skyscraper	0.8
light	0.7

図 4　画像 A（左）と画像 B（右）のタグ情報の例

そして，タグの逆文書頻度を用いてベクトル化した結果と本研究でタグの確信度を用いてベクトル化した結果を図 5 に示す．確信度を用いてベクトル化手法のほうが類似度が大きくなる．

2 つ画像間の距離を cosine distance を用いて計算する．A と B は画像のベクトルを表し，画像間の距離 $dist(a,b)$ は以下の式を用いて求める．

$$dist(a,b) = 1 - cosine\ similarity(a,b) = 1 - \cos(\theta) = 1 - \frac{A \cdot B}{\|A\|\,\|B\|} \quad (1)$$

$dist(a,b)$ は $0 \leqq dist(a,b) \leqq 1$ を満たし，0 に近いほど，画像 a と画像 b は類似する．これにより，意味的類似度マトリックスを作成できる．また，画像間の位置情報を用いて地理的距離マトリックスを作成できる．意味的類似度マトリックスと地理的距離マトリックスから高次元潜在空間を構築できる．次の節は高次元潜在空間内に行う Semantic-DBSCAN クラスタリングの詳細について説明する．

	chritmas tree	illumination	landmark	light	purple	sky	skyscraper
画像 A:	[0.49,	0.49,	0,	0,	0.49,	0.35,	0.35]
画像 B:	[0,	0,	0.57,	0.57,	0,	0.40,	0.40]

	chritmas tree	illumination	landmark	light	purple	sky	skyscraper
画像 A:	[0.9,	0.8,	0,	0,	0.6,	0.9,	0.8]
画像 B:	[0,	0,	0.9,	0.7,	0,	0.9,	0.8]

図 5　タグの逆文書頻度を用いてベクトル化した結果（上）とタグの確信度を用いてベクトル化した結果（下）

3.1.3　Semantic-DBSCAN アルゴリズム

我々の以前の研究 [2] では地理空間内に投稿写真の密度が高いエリアを抽出するため，DBSCAN を利用している．DBSCAN は，クラスタ間の距離の閾値 Eps とクラスタを構成する最小データ数の閾値 $MinPts$ との 2 つの閾値を持つ．ある点 x から，距離 Eps 内にある点集合を近傍 $N_Epx(x)$ と定義し，以下の接続関係を満たす時，同じクラスタに分類すると考えられる．

$$y \in N_Epx(x) \quad (2)$$

$$|N_Epx(x)| \geq MinPts(\text{コア点}) \quad (3)$$

そこで，地理的な距離が近いだけではなく，意味的な類似度が近いエリアを POI として抽出するため，DBSCAN を拡張して Semantic-DBSCAN を提案する．Semantic-DBSCAN は意味的類似性を考慮した $Eps2$ を追加して三つパラメータ $Eps1, Eps2, MinPts$ を有する．$Eps1$ は二つ写真の位置情報から計算した地理的な距離の閾値であり，$Eps2$ は二つ特徴ベクトル間の類似度の閾値であり，$MinPts$ はクラスタを構成する最小データ数の閾値である．関数 $Retrieve\ Neighbors(o, D, Eps_1, Eps_2)$ を以下のように定義する．

アルゴリズム 1 Semantic-DBSCAN

Input: $Eps1, Eps2, MinPts, D = \{o_1, o_2, \ldots, o_n\}$
Output: $C = \{C_1, C_2, \ldots, C_k\}$

```
Cluster_Label = 0
for i=1 to n do
  if o_i is not in a cluster then
    X=Retrieve_Neighbors(o_i,D,Eps1,Eps2)
    if |X| <MinPts then
      Mark o_i as noise
    else
      Cluster_Label = Cluster_Label + 1
      for j=1 to |X| do
        Mark all objects in X with current Cluster_Label
      end for
      Push(all objects in X)
      while not IsEmpty() do
        CurrentObj = Pop()
        Y = Retrieve_Neighbors(CurrentObj,D,Eps1,Eps2)
        if |Y| >MinPts then
          for All objects o in Y do
            if o is not marked as noise or it is not in a cluster then
              Mark o with current Cluster_Label  Push(o)
            end if
          end for
        end if
      end while
    end if
  end if
end for
```

$$Retrieve\ Neighbors(o, D, Eps_1, Eps_2) = \{0 | x \in D, dist_{geo}(o, x) \leqq Eps_1, \quad dist_{sem}(o, x) \leqq Eps_2\} \tag{4}$$

ここで $dist_{geo}(o_1, o_2)$ は o_1 と o_2 の地理的な距離を表し，$dist_{sem}(o_1, o_2)$ は o_1 と o_2 の意味的な距離を表す．Semantic-DBSCAN のアルゴリズムをアルゴリズム 1 に示す．ただし，実際に応用する際に，地域の特性や生成したいピクトリアルマップの基準により，複数回クラスタリングを行う．

3.2 意味的類似する写真の絞り込む・代表的な写真の選出

本節では写真間の意味的類似性を基に代表写真を絞り込む手法について述べる．

POI のクラスター C におけるある任意の写真 a は，a 以外にほか写真間との意味的距離を計算し合計する．クラスター C における写真 a の意味的距離を以下の式で定義する．cosine distance の計算式は式 (1) である．

$$dist(a|C) = \sum_{x \in C} cosine\ distance(a, x) \tag{5}$$

値の小さい順にランキングすることで，上位一枚の写真を POI に最も代表的な写真として抽出し，二位以降ランキング上の写真を写真集として地図に提示する．写真間の意味的類似性を注目することで，クラスター内に写真の内容のバラツキが大きい場合でも，ランキングした写真間の一致性を保持することができる．

4 ピクトリアルマップ生成

本研究では，従来システムと提案システムに従って二種類ピクトリアルマップを生成する．図 6 は従来システムの流れと提案システムの流れを比較する．今回は札幌駅周辺，天神駅周辺と横浜駅周辺を対象にし，それぞれ二種類ピクトリアルマップを生成された．図 7 は提案システムによる生成された横浜駅周辺のピクトリアルマップの例である．各クラスタ化 POI にその POI の最も代表的な画像をアイコンとして表示する．アイコンをクリックすることで，右側に代表的な画像集が表示される．画像の左上の番号はランキング順を表す．番号「1」が付いている画像はその POI の最も代表的な画像であり，アイコンとしても使用される．

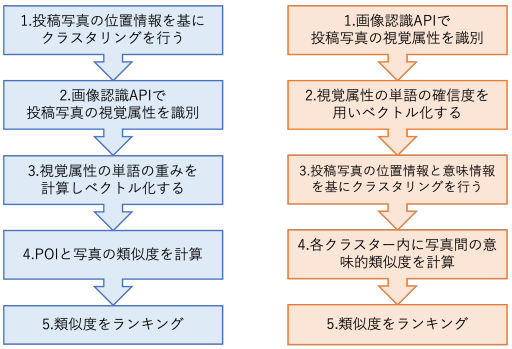


図 6　従来システムの流れ（左）と提案システムの流れ（右）

5 評価実験

本章では，従来手法で構築されたピクトリアルマップと，提案手法で構築されたピクトリアルマップを比較した際の性能評価を行う．

5.1 実験内容

本実験では写真共有サイト Flickr にアップロードされた画像から構成された写真のデータセット YFCC100M [8] を使用した．YFCC100M は，世界中で撮影された約 1 億枚の写真及び動画を含んでおり，その中の約 5,000 万件は位置情報付き写真である．本実験では，写真の位置情報（軽緯度）を用い，福岡市の天神駅周辺，札幌市の札幌駅周辺と横浜市の横浜駅周辺，この三つ地域の写真をフィルタリングした．各地域で収集した写真の数は，札幌駅を中心にして 16 平方キロメートルの範囲に 8625 枚の写真であり，福岡市の天神駅を中心にして 36 平方キロメートルの範囲 8947 枚の写真であり，横浜市の横浜駅を中心にして 16 平方キロメートルの範囲 13426 枚の写真である．

まず，各写真に対して，Image Analysis API を用いて分析を行い，タグと確信度を求めた．次に，意味情報と地理情報の両



図 7　横浜駅周辺のピクトリアルマップの例

方を用いて高次元潜在空間を作成した．そして，多次元空間内に提案した Semantic-DBSCAN を用いて POI の抽出を行った．表 1, 表 2, 表 3 は各地域に対して使用したパラメータと抽出した POI の個数に示す．それぞれ地点に対して，2 回クラスタリングを行ったので，表に示すステージ 2 の POI 結果を採用した．最後に，各クラスター内に写真間の意味的類似度を計算しランキングした．上位一枚の写真を POI に最も代表的な写真として抽出しアイコンとしてマッピングする．二位以降ランキング上の写真を画像集として地図に提示する．

生成された 3 つ地域の 2 種類ピクトリアルマップ（提案システムより生成したマップと提案システムより生成したマップ）の有効性を検証するため，被験者実験を行った. 被験者はピクトリアルマップのインターフェイスを操作してもらい，以下の質問に回答してもらった．

Q1. マップ上のこの画像が表す地点を知っていますか？ 被験者は Q1 の質問に対して「1．はい」,「2．いいえ」で回答する．

Q2. このアイコンと右側の画像集が表している地点は，この地域の特徴的な地点だと思いますか？ この質問について 5 段階（「全くそう思わない-1」,「あまりそう思わない-2」,「どちらとも言えない-3」,「ややそう思う-4」,「とてもそう思う-5」）の尺度を用いた回答欄を設ける．

表 1　Semantic-DBSCAN による天神駅周辺の POI の結果

段階	パラメータ	POI の個数
ステージ 1	Eps1 = 0.1km, Eps2 = 0.75, MinPts = 15	21
ステージ 2	Eps1 = 0.05km, Eps2 = 0.5, MinPts = 3	20

表 2　Semantic-DBSCAN による札幌駅周辺の POI の結果

段階	パラメータ	POI の個数
ステージ 1	Eps1 = 0.05km, Eps2 = 0.75, MinPts = 10	17
ステージ 2	Eps1 = 0.03km, Eps2 = 0.5, MinPts = 3	31

表 3　Semantic-DBSCAN による横浜駅周辺の POI の結果

段階	パラメータ	POI の個数
ステージ 1	Eps1 = 0.08km, Eps2 = 0.65, MinPts = 30	19
ステージ 2	Eps1 = 0.05km, Eps2 = 0.35, MinPts = 5	34

Q3. この地点を表す画像集に含まれる画像のうち，不適切だと思われる画像の番号をチェックしてください．

Q4. この画像集の中で，「1」の番号が付けられた画像は，この地点を表す代表的な写真であると思いますか？ Q2 と同じく 5 段階（「全くそう思わない-1」,「あまりそう思わない-2」,「どちらとも言えない-3」,「ややそう思う-4」,「とてもそう思う-5」）の尺度を用いた回答欄を設ける．

クラウドソーシングサービスであるクラウドワークス [1]でアンケートを実施した. さらに，2 種類ピクトリアルマップに対応する二つアンケートを答える順番によって二つグループに分けて実施した．一つ目のグループはまず提案システムによる生成したマップについての質問を答えてもらい，次は従来システムよる生成したマップについての質問を答えてもらった．二つ目のグループは答える順番は逆にした．一つ目のグループの被験者数のべ 62 名（天神駅周辺 18 名, 札幌駅周辺 19 名, 横浜駅周辺 25 名）から回答を得た．正確に比較するために，もう一つグループも同じく被験者数のべ 62 名（天神駅周辺 18 名, 札幌駅周辺 19 名, 横浜駅周辺 25 名）から回答を得た．

5.2　POI の抽出手法の有効性の評価

アンケート項目 Q1 と Q2 は抽出された POI の有効性を評価する. まず，Q1「マップ上のこの画像が表す地点を知っていますか」との質問における，評価の結果を図 8 に表示する. 従来手法と比較する際に，「はい」と答えた POI は全体の 84%に占め, 一方，従来手法では，75%であった．提案手法の方が被験者が知っている POI の数が多かった．

次に，Q2「このアイコンと右側の画像集が表している地点は，この地域の特徴的な地点だと思いますか」との質問における，評価の結果を図 9 と図 10 に表示する. 二つグループの結果をそ

1：https://crowdworks.jp/

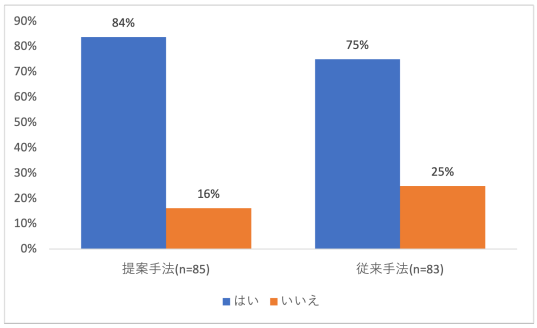


図 8　抽出された POI を知っているかどうかについての質問 (n:POI の個数)

れぞれに述べる．まず，提案手法→従来手法の順番で，図 9 を見ると，従来手法より提案手法の方は「とてもそう思う-5」と「ややそう思う-4」の辺りにデータが集中していることが分かった．もう一つ，従来手法→提案手法の順番で，図 10 を見ても，同じく提案手法の方が「とてもそう思う-5」と「ややそう思う 4」の辺りにデータが集中していることが分かった．

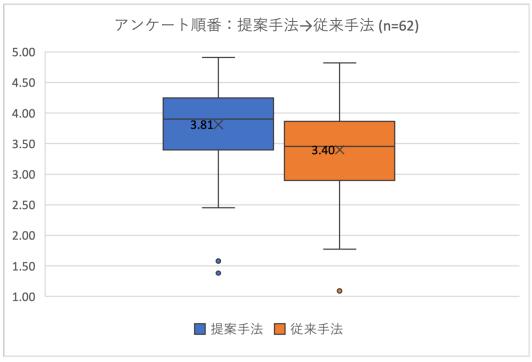


図 9　抽出された POI はこの地域の特徴的な地点だと思うかどうかについての質問 (n: 被験者数)

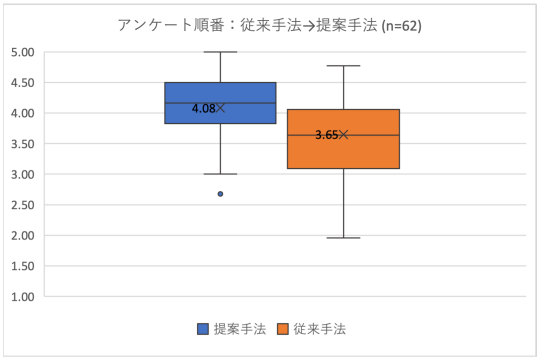


図 10　抽出された POI を知っているかどうかについての質問 (n: 被験者数)

5.3　代表的な画像集の有効性の評価

Q3 は抽出される代表的な画像集の有効性を評価する. それぞれ選出された POI に代表となる画像集に対して適合率を算出する. 適合率は, ピクトリアルマップに表示する画像集中の画像の適切さに関連する指標である. 以下の式を用いて算出する.

$$\text{適合率} = \frac{\text{被験者が適切と判断した画像枚数}}{\text{代表画像集の総数}} \qquad (6)$$

それぞれ POI の適切さを算出して平均した結果を図 11 に表示する. 提案手法の方は適合率は 87%であった．一方，従来手法の方は 79%であった．従来手法より適合率が 8%程度向上したことが分かった．

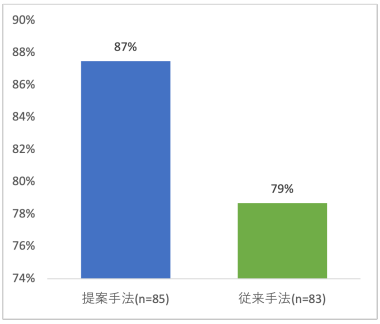


図 11　POI の適切さについての質問（n:POI の数）

5.4　最も代表的な画像の有効性の評価

最後に，Q4 は最も代表的な画像の一枚の有効性を評価する.. Q4「この画像集の中で，「1」の番号が付けられた画像は，この地点を表す代表的な写真であると思いますか？」との質問における，評価の結果を図 12 と図 13 に表示する.

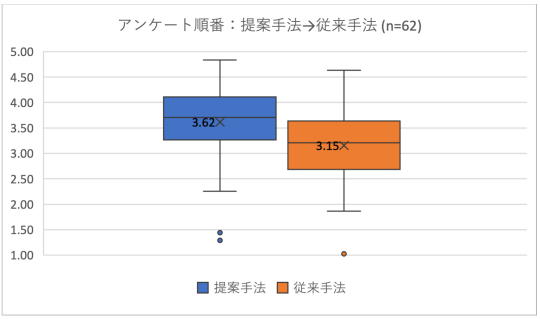


図 12　選出された最も代表的な画像は，この地点を表す代表的な写真だと思うかどうかについての質問 (n: 被験者数)

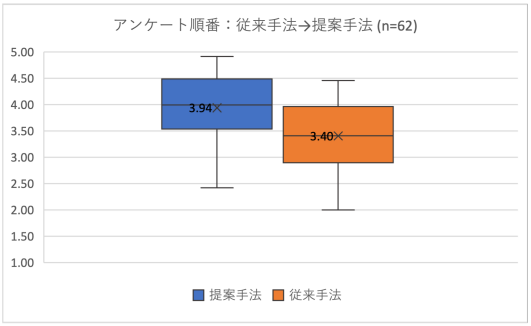


図 13　選出された最も代表的な画像は，この地点を表す代表的な写真だと思うかどうかについての質問 (n: 被験者数)

図 12 と図 13 の両方を見ると，従来手法より提案手法の方が

「とてもそう思う-5」と「ややそう思う-4」の辺にデータが集中していることが分かった．

6 まとめと今後の展望

本論文では，SNS に投稿された写真の意味情報と地理情報を利用してピクトリアルマップを自動的に生成する手法を提案した．評価実験の結果より，提案手法は従来手法より以下の二点に関して，優れた性能を示している．

(1) 地域の特徴的な地点を表す POI を抽出する

(2) マップ上に適切な代表的な画像集をユーザーに提示する

本研究では SNS の投稿写真だけを利用して分析を行う．今後，SNS のテキストデータにも加えて，地域の魅力をより多くの面から分析することで，ユーザーの多様なニーズに満たすピクトリアルマップを生成することを考えている．

7 付　　録

図 14 は提案システムによる生成された天神駅周辺のピクトリアルマップの例である．図 15 は提案システムによる生成された札幌駅周辺のピクトリアルマップの例である．

文　　献

[1] 高尚暉, 牛尼剛聡, “SNS を利用したピクトリアルマップの自動生成,” 第 10 回データ工学と情報マネジメントに関するフォーラム, DEIM2017, D7-4, 2017.

[2] Wentong Zhou, Taketoshi Ushiama, “Automatic Generation of Pictorial Maps from Photos on Social Media to Represent Regional Features,” 2020 14th International Conference on Ubiquitous Information Management and Communication(IMCOM), pp. 1–5, 2020.

[3] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, “A density-Based Algorithm for Discovering Clusters,” KDD-96, pp. 226–231, 1996.

[4] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval,” Cambridge: Cambridge University Press, 2008.

[5] Wei-Chao Chen, Agathe Battestini, Natasha Gelfand, Vidya Setlur, “Visual summaries of popular landmarks from community photo collections,” 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, pp. 1248–1255, 2009.

[6] G. Salton, M. J. McGill, “Introduction to Modern Information Retrieval,” McGraw-Hill, Inc., 1986.

[7] Image Analysis API: https://azure.microsoft.com/ja-jp/products/ai-services/ai-vision/

[8] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K.Ni, D. Poland, D. Borth, and L. J. Li, “YFCC100M: The new data in multimedia research,” Communications of the ACM, Vol.59, No.2, pp. 64–73, 2016.



図 14　天神駅周辺のピクトリアルマップの例



図 15　札幌駅周辺のピクトリアルマップの例

地点データのインタラクティブなクラスタリングのための可視化インタフェース

渡邉　慧汰† 藤田　秀之†† 大森　匡†† 新谷　隆彦††

† 電気通信大学 大学院情報理工学研究科　〒 182–8585 東京都調布市調布市調布ヶ丘 1-5-1
†† 電気通信大学 大学院情報理工学研究科　〒 182–8585 東京都調布市調布市調布ヶ丘 1-5-1
E-mail: †w2231164@edu.cc.uec.ac.jp, ††{fujita,omori,shintani}@is.uec.ac.jp

あらまし　OPTICS は，密度ベースのクラスタリング手法であり，地点データの分析にも用いられる．OPTICS の特徴に，クラスタリング構造を示す図 (到達性プロット) を，OPTICS において重要な，距離パラメータ調整に利用できる点がある．しかし，データが大規模になると，到達性プロットの可読性は著しく低下する．加えて，大規模な地点データは，広域におよび，場所ごとに密度が大きく異なることがあるため，注目する場所ごとに距離パラメータを調整することが自然である．そこで，本研究では，ユーザが，地図のズーム・スクロール操作により対象地域を指定し，対象地域に応じた集約表示した到達性プロットと，地図上のクラスタリング結果とを確認しながら，距離パラメータを調整する可視化インタフェースを提案する．加えて，複数の距離パラメータを推薦し，到達性プロット上に示すことで，ユーザのパラメータ調整を支援する．自身の 5 年分の地点データを対象に，提案システム上で，OPTICS が自然な操作で可能なことを確認した．また，提案した集約手法とパラメータ推薦手法の有効性を定量的に確認した．

キーワード　情報可視化，インタラクティブシステム，地理的情報可視化，インタラクティブ・クラスタリング，OPTICS

1 背　　景

スマートフォンを代表として，個人の行動データを記録するデバイスが普及し，自身のデータを自身で分析するセルフ・トラッキング [11] と呼ばれる活動も注目されており，専門家だけではなく，個人ユーザのためのデータ分析アプリケーションの必要性が主張されている [12]．中でも，Google Location History のような地点データは，自身が訪問した地点の履歴データであり，自身の行動を分析するために有用なデータである．

OPTICS は，データの密度に基づく，地点データの代表的なクラスタリング手法である．OPTICS の特徴に，クラスタリング構造を示す図 (到達性プロット) を，距離パラメータ調整に利用できる点，距離パラメータ変更時のクラスタリング結果の更新が高速な点がある．しかし，データが大規模になると，到達性プロットの可読性は著しく低下する．加えて，大規模な地点データは，しばしば広域におよび，場所ごとに密度が大きく異なるため，注目する場所ごとに距離パラメータを調整することが自然であり，全データでの到達性プロットを見て調整することは望ましくない．ここで，4 章に詳述するとおり，単純にビューポート内のデータに絞り込んでからクラスタリングすると，同一の距離パラメータに対し，クラスタリング結果の一貫性が保たれない．加えて，到達性プロットから単純にビューポート外の点群を除くと，クラスタリング構造を正しく表現できない．

本研究では，ユーザが，地図のズーム・スクロール操作により対象地域を指定し，全データでのクラスタリング構造を維持しながら，対象地域に応じて集約する手法と，地図上のクラスタリング結果と到達性プロットを確認しながら，距離パラメータを調整する可視化インタフェースを提案する．加えて，注目地域における安定性の高いクラスタが出現する距離パラメータを複数推薦する手法を提案し，パラメータ調整を支援する．

本稿の貢献は次のとおりである．

- 大規模データに対する OPTICS によるインタラクティブ・クラスタリングを，地図と到達性プロット上での自然な操作で実現する
- 大規模データに対する OPTICS 到達性プロットの可読性を改善する
- 安定性の高いクラスタの距離パラメータを推薦することで，ユーザのパラメータ調整を支援する
- 提案手法を実データに適用し，定性的，定量的に評価して有用性を確認する

2 関 連 研 究

本章では，本研究の関連研究として，分析を支援するインタラクティブ・クラスタリング，そして，密度ベースクラスタリング手法の 1 つである OPTICS を拡張した研究についてまとめ，本研究の位置づけを述べる．

インタラクティブ・クラスタリングは，ユーザにとって適切なクラスタリング結果を得るために，ユーザが処理過程に介入するクラスタリング手法である．Bae らによるサーベイ [5] では，インタラクティブ・クラスタリングに関する 105 本の論文を対象に，4 つの目的と 3 つのインタラクション方法を挙げている．インタラクティブ・クラスタリングの目的は，(i) クラスタリングが完全に自動化できない前提において，ユーザとアル

ゴリズムが対話的にクラスタリングを行い，結果の質を向上させること [1]，(ii) ユーザがクラスタリング結果を理解しやすくすること [1] [15]，(iii) 特定のデータを強調し，興味深いデータを発見すること [2]，(iv) ユーザの目的とする分析の支援すること [3] である．また，インタラクション方法については，クラスタ数・類似度などのパラメータを明示的に操作を行う，パラメータへのインタラクションが挙げられている [1]．

本研究では，インタラクティブ・クラスタリングの目的を踏まえた上で，パラメータによるインタラクションを行う，可視化インタフェースを提案する．

次に，可視化分析のために，OPTICS を用いる提案や，拡張する提案についてまとめる．Richter らは，大規模データのクラスタリングにおいて，複数の点を集約した，重み付きの点を対象とすることで，処理を高速化しつつ，粗い到達性プロットを提示する手法を提案している [13]．Rinzivillo らは，軌跡データのインタラクティブなクラスタリングのために，距離関数に複数の指標をユーザが順に加え，クラスタリング結果を更新する過程を繰り返す，漸次的なクラスタリング手法を提案している [14]．Brecheisen らは，同一のデータセットに対する，異なる距離関数による異なるクラスタリング構造間の関係を直感的に理解するために，到達性プロットを 2 つ並べたインタラクティブな可視化インタフェースを提案している [6]．Andrienko らは，大規模データのクラスタリングにおける，ユーザによる，距離関数の選択やパラメータの調整の効率化するために，サンプリングしたデータでクラスタリングを行い，パラメータ調整等を行う手法を提案している [3]．Guo は，多次元データ分析における，ユーザによる特徴量の選択とクラスタリングの効率化を目的とする可視化分析環境において，各次元の関係をグラフ構造で表現し，そのクラスタリング構造を表現する到達性プロットを提案している [9]．

以上のように，到達性プロットの描画方法の改善を行う取り組みも見られるが，本研究の特徴である，地図のズームに応じた到達性プロットの可読性の改善を行う研究は，筆者らの知る限り存在しない．

3 OPTICS [4]

本稿の提案に用いる OPTICS について説明する．DBSCAN [8] は，密度の高い領域の点群をクラスタとするクラスタリング手法である．後述のとおり，パラメータとして，クラスタ内の点間の最長距離 (距離パラメータ) と，最小点数を指定する．ユーザによるパラメータの調整が重要であり，しばしば，データの分布と目的に応じて，結果の確認と値の変更を繰り返すことになるが，パラメータ変更時の再計算の処理時間が，インタラクティブに動作させるうえで問題となる．

3.1 用語の定義とアルゴリズム

OPTICS では，指定された距離パラメータ ϵ，最小点数 $MinPts$ に対し，対象とする点データセット D 内の点を，いずれかのクラスタ，または，クラスタに属さないノイズに分類する．OPTICS のクラスタは，ある点から，直接密度到達可能な点を推移的にたどり到達可能な極大集合と定義される．ここで，指定された ϵ と $MinPts$ に対して，以下を満たすとき，$p \in D$ から $q \in D$ を直接密度到達可能と呼ぶ．

$$q \in N_\epsilon(p), |N_\epsilon(p)| \geqq MinPts \tag{1}$$

ここで，$N_\epsilon(p)$ とは，p から半径 ϵ 以内に存在する，p を含む点の集合である．例として，図 1 では，橙点および緑点が 1 つのクラスタを構成し，青点はノイズとなる．半径 ϵ 以内の点の数が，B では $MinPts$ 未満だが，A では $MinPts$ 以上であり，A から B が直接密度到達可能となり，B もクラスタに含まれる．

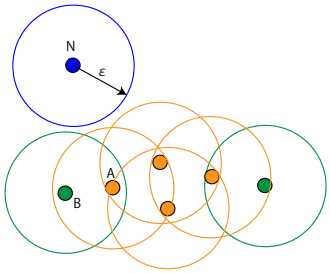


図 1: OPTICS で抽出されるクラスタ ($MinPts = 4$)．橙色・緑色で示した点が 1 つのクラスタを構成．青色の点はノイズ点

続いて，指定された ϵ と $MinPts$ に対して，p のコア距離 $core_distance$ を，p を中心とする円内の点の数が $MinPts$ 以上となる最小の半径として，式 2 で定義する．ただし，半径 ϵ 以内の点の数が $MinPts$ 未満の場合，$Undefined$ とする．

$$core_distance_{\epsilon,MinPts}(p) = \begin{cases} Undefined & (|N_\epsilon(p)| < MinPts) \\ \mathrm{distance}(p, q_{MinPts}) \end{cases} \tag{2}$$

ここで，q_{MinPts} は点 p の $MinPts$ 番目の近傍点である．また，$\mathrm{distance}(p, q)$ は，点 pq 間のユークリッド距離である．

続いて，指定された ϵ と $MinPts$ に対して，点 $o \in D$ の p からの到達性距離 $reachability_distance$ を，op 間の距離と o のコア距離のうち大きい方として，式 3 で定義する．

$$reachability_distance_{\epsilon,MinPts}(p) = \begin{cases} Undefined & (|N_\epsilon(p)| < MinPts) \\ \max(core_distance(o), \mathrm{distance}(o, p)) \end{cases} \tag{3}$$

ここで，$\mathrm{distance}(o, p)$ は op 間の距離である．

以上の定義に基づき，OPTICS は次の処理を行う．アルゴリズムは省略する．まず，OPTICS の事前処理は，D，$MinPts$，ϵ_{max} を入力とし，D 内の各点のクラスタリング順序と到達性距離を出力する．ここで，クラスタリング順序は D 内の点の順序であり，次の性質を持つ．

- 指定された $\epsilon \leqq \epsilon_{max}$ に対し，到達性距離 ϵ 以下，かつ，クラスタリング順序が連続する各極大列が，それぞれクラスタとなる

また，各点の到達性距離は，クラスタリング順序が前の点からの，$\epsilon = \epsilon_{max}$ のときの値である．本稿では，D 内の各点に上記の到達性距離の値を持たせ，クラスタリング順序で並べたリストを，到達性リストと呼ぶ．

続いて，OPTICS における各点へのクラスタラベルの割り当て，すなわちクラスタリング結果の取得は，指定された ϵ と，到達性リストを入力として行う．データ数 N に対して，計算量 $O(N)$ の高速な処理であり，同一の到達性リストを用い，ϵ を変えて実行できる．

3.2 到達性プロット

到達性プロットは，横軸にクラスタリング順序，縦軸に到達性距離をとり，到達性リストを可視化する棒グラフである．到達性プロットは，以降で示すように，クラスタの階層関係を読み取れるため，階層的クラスタリングにおける樹形図に類する性質を持つと言える．到達性プロットを，指定された高さ ϵ の横線により区切ってできる「谷」それぞれがクラスタとなる．例として，図 2 は平面上の点分布 (上) と，その到達性プロット (下) を示している．青線で示した ϵ_1 では，到達性プロットに 1 つの「谷」ができ，青丸で囲んだ 1 つのクラスタが得られる．橙線で示した，より小さい ϵ_2 では，到達性プロットに 2 つの「谷」ができ，橙丸で囲んだ，より高密度な 2 つのクラスタが得られる．

図 2 でも確認できるとおり，ϵ を減少させるとき，到達性プロット上の極大値を超える毎にクラスタは分割され，極小値を超えるとクラスタはノイズとなって消滅する．つまり，クラスタ数が変化する ϵ は，到達性プロットの各極値である．したがって，到達性プロットの極値の可読性は，ユーザによる ϵ の調整のために重要である．

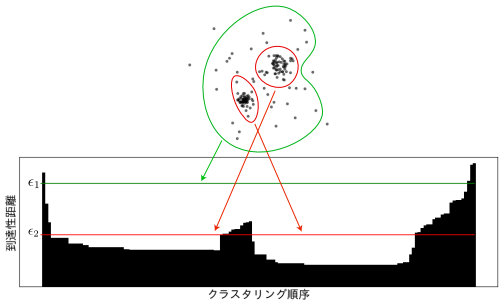


図 2: クラスタと到達性プロットの関係．到達性プロット上の「谷」がクラスタに対応するため，ϵ_1 では，より広範囲な 1 つのクラスタ，ϵ_2 では，より高密度な 2 つのクラスタが得られる

4 提案手法

OPTICS によるインタラクティブなクラスタリングの問題と方針をまとめ，到達性プロットの集約描画手法と，距離パラメータ ϵ の推薦手法，そしてそれらを利用した OPTICS の可視化インタフェースを提案する．

4.1 問題と方針

大規模データにおける OPTICS の問題点をまとめ，提案の方針を示す．

4.1.1 大規模データによる到達性プロットの可読性低下

OPTICS の到達性プロットには，データが大規模になると，可読性が著しく低下する問題がある．具体的には，図 3 に示すとおり，データ数が画面横幅の解像度を越えると，到達性プロットの縦棒の幅が 1 ピクセル以下となり，正しく描画できない．加えて，到達性プロットの縦軸最大値が，極端に大きな到達性距離で決まるため，到達性距離が 0 でない多くの点で読み取りが困難になる．

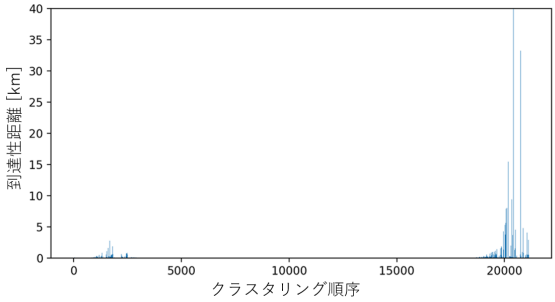


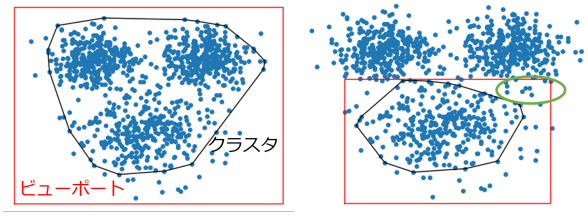
図 3: 画面横幅 (約 1,900 ピクセル) よりデータ数 (約 21,000 点) が多い到達性プロット．画面右の値の大きな縦棒を観察すると，描画されている点とされていない点があり，ウィンドウサイズのわずかな変更で，それらは入れ替わる (Python Matplotlib や Google Chrome の SVG で確認)．図中央部も大多数はゼロではないが，高さが低く視認できない

4.1.2 ビューポートに対するクラスタリング構造の一貫性

前述のとおり，大規模データでは，ビューポート内の点データに対して，クラスタリングの距離パラメータを調整することが自然である．まず単純に，ビューポート内の点に絞り込んでからクラスタリングする手法を考える．しかし，この手法では，ビューポート毎にクラスタリング構造が変化し，一貫性が保たれない．言い換えれば，同一の距離パラメータでも，ズームのようなビューポート操作で，クラスタリング結果が変化する．図 4 に例を示す．赤枠がビューポートであり，黒線はクラスタの凸包である．図 4a から図 4b にズームを行うと，同一の距離パラメータでも，図 4b のビューポート右上の緑で囲った点がクラスタから外れ，クラスタリング結果が変化してしまう．これは，ビューポート外になった点群のクラスタリング構造が失われることで発生する．

以上より提案の方針をまとめると，以下の通りである．

- ビューポート内のクラスタリング構造のみを示すように，到達性プロットを集約表示する
- クラスタリングの一貫性を保つため．クラスタリングはデータ全体で行い，到達性プロットのみを集約表示する
- 到達性プロットから単純にビューポート外の点群を除くと，クラスタリング構造を正しく表現できない問題 (4.2 節) を考慮する

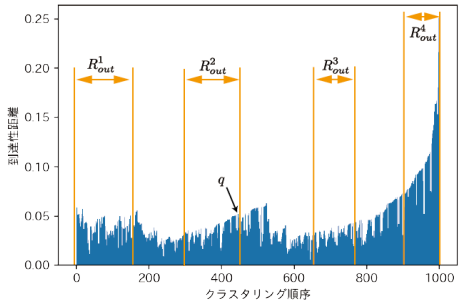


(a) ビューポート変更前　(b) ビューポート変更後

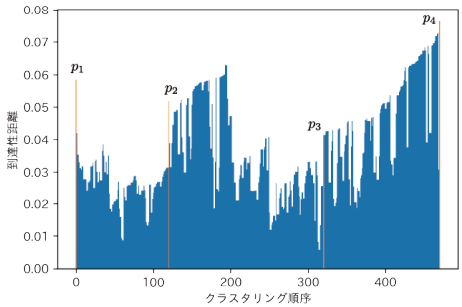
図 4: ビューポート内でのクラスタリングの問題．距離パラメータ一定でも，ズームによりビューポートが変わっただけで，緑で囲んだ点群がクラスタから外れてしまう

4.2 到達性プロットの集約描画手法の提案

ビューポートに応じて，到達性プロットを集約描画する手法を検討する．点をクラスタリング順序により辿ると，図 4 で示したようなビューポートへの出入りを，しばしば複数回繰り返す．ビューポートを出てから再び入るひとつ前までの点列を，ビューポート外点列と呼ぶ．図 4 の緑の点列や，図 5 上の橙色の範囲の R^1_{out}，R^2_{out}，R^3_{out}，R^4_{out} はそれぞれビューポート外点列である．ビューポート外点列を削除すると，到達性プロットはクラスタリング構造を正しく表現しない．例えば，図 5a のビューポート外点列を削除し，$\epsilon = 0.5$ における谷を観察すると，R^2_{out} 内の高さ 0.5 を超える極大値 q がないため，本来は q の左右で 2 つに分かれていた谷が，1 つの谷として表現される．つまり，2 つに分かれるべきクラスタが 1 つに見えてしまう．



(a) 到達性プロットの集約前



(b) 到達性プロットの集約後

図 5: 到達性プロットの集約前後．ビューポート外点列 R^1_{out}，R^2_{out}，R^3_{out}，R^4_{out} を，ビューポート外点列をそれぞれビューポート外仮想点 p_1，p_2，p_3，p_4 に置き換えた

そこで，提案する到達性プロットの集約処理では，ビューポート外点列を，その代表点であるビューポート外仮想点に置き換えることにより，クラスタリング構造を正しく表現する．ビューポート外仮想点の高さは，対象とするビューポート外点列の到達性距離の最大値と，ビューポート内の点の到達性距離の最大値の小さい方である．これにより，縦軸最大値がビューポート外点列の極端な値によって決まる問題を回避できる．

全データ D の到達性リストを R として，ビューポート V のビューポート内点列 $R_{in}(V)$ を，V に座標が含まれる点のみを選んだ R の部分列と定義する．V のビューポート外点列 $R_{out}(V)$ は，V に座標が含まれない点のみを選んだ部分列であり，その中の極大な連続部分列の数を n として，それぞれを $R^i_{out}(V)(1 \leqq i \leqq n)$ とする．$R_{in}(V)$ に各 $R^i_{out}(V)$ を挿入した列が R となる．各 $R^i_{out}(V)$ に対するビューポート外仮想点を，式 4 で算出する到達性距離 $r^i_{out}(V)$ を持ち，地図上の座標値を持たない点と定義する．

$$r^i_{out}(V) = \min(\max(R_{in}(V)), \max(R^i_{out}(V))) \qquad (4)$$

V に対して集約した到達性リスト $R_{agg}(V)$ を，R 中の各 $R^i_{out}(V)$ を，それぞれビューポート外仮想点に置き換えたリストと定義する．

図 5b は，提案する上記の集約処理を行った到達性プロットである．図 5a のビューポート外点列 R^1_{out}，R^2_{out}，R^3_{out}，R^4_{out} を，それぞれ橙線のビューポート外仮想点 p_1，p_2，p_3，p_4 に置き換えている．点の数が削減され，$\epsilon = 0.5$ において，p_2 の左右でクラスタが分割されることを正しく読み取れる．加えて，縦軸最大値を抑え，全体として縦棒を拡大している．

到達性プロットの集約アルゴリズムは，以上の定義に基づき，R，V を入力とし，$R_{agg}(V)$ を出力する処理である．手順を Algorithm1 に示す．ここで，inMap(p, V) は，点 $p \in R$ の座標が V に含まれるとき真，それ以外で偽を返す関数，$v$createVp(r) は到達性距離が r のビューポート外仮想点を返す関数である．

4.3 提案する距離パラメータ ϵ の推薦手法

ユーザは，到達性プロット上の「谷」に基づいて距離パラメータ ϵ を選択する．通常，到達性プロット上で「谷」は複数あり，ユーザに適切な ϵ を選ぶことは難しい．特に，データ数が大きいデータではこの問題が顕著になる．加えて，ユーザにとって適切なパラメータを定めるためには，複数の候補があることが望ましい．クラスタは距離パラメータの変化に頑健で，点数が大きいほど安定している [7]．そこで，本研究では，以下の特徴を持つ指定された件数の ϵ を推薦する．

- 推薦するクラスタの安定性が高い
- 推薦する ϵ の多様性が高い

この特徴を満たす，到達性リストを利用したクラスタの安定性を新たに提案する．

ϵ の減少に伴い，クラスタは分割される．分割前を親クラスタ，後を子クラスタと呼ぶ．まず，集約した到達性リストを利用し，ビューポート内の全てのクラスタの集合 C と，各 $c \in C$ の子クラスタと，c が出現，および，消滅する ϵ である，ϵ^{appear}_c，および，$\epsilon^{disappear}_c$ を求める．続いて，安定性 $stability(c)$ の上

Algorithm 1 到達性プロット集約アルゴリズム

Input: R, V
Output: $R_{agg}(V)$

$R_{agg}(V) \leftarrow$ empty list
$r_{in}, r_{out} \leftarrow 0$
for $p \in R$ **do**
  if inMap(p, V) **then**
    if $r_{out} \neq 0$ **then**
      $R_{agg}(V)$.push(createVp(r_{out}))
      $r_{out} \leftarrow 0$
    end if
    $R_{agg}(V).push(p)$
    $r_{in} \leftarrow \max(r_{in}, p.reachability)$
  else
    $r_{out} \leftarrow \max(r_{out}, p.reachability)$
  end if
end for
if inMap$(R[R.length-1], V) == False$ **then**
  $R_{agg}(V)$.push(r_{out})
end if
for $p_{agg} \in R_{agg}(V)$ **do**
  $p_{agg}.reachability \leftarrow \min(r_{in}, p_{agg}.reachability)$
end for
return $R_{agg}(V)$

位 k 個となる c の ϵ_c^{appear} を推薦する．c の安定性 $stability(c)$ は次式で算出する．

$$stability(c) = \sum_{p \in c} r(p) \tag{5}$$

$$\begin{cases} r(p) = p.reachability & \text{if } p.reachability \geqq \epsilon_c^{disappear} \\ 0 & otherwise \end{cases} \tag{6}$$

ここで $p.reachability$ は点 p の到達性距離である．$\epsilon_c^{disappear}$ は消滅する ϵ の値であり，c が子クラスタが持たない場合 0，それ以外の場合 c が分割される ϵ である．

各クラスタ c の安定性は，c の出現から消滅までに，クラスタから外れる点の到達性距離の和にあたる．このため，出現時と消滅時の ϵ の差が大きく，多数の点を含むクラスタほど，安定性が高くなると期待できる．図 6 に例を示す．

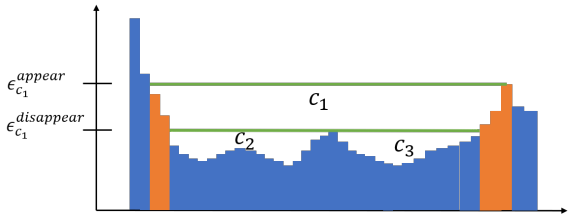


図 6: クラスタの出現と消滅．ϵ を減少させるとき，$\epsilon_{c_1}^{appear}$ でクラスタ c_1 は出現し，$\epsilon_{c_1}^{disappear}$ である c_2，c_3 に分割されて消滅．c_1 の安定性は，橙の点の到達性距離の和

4.2 節で提案した集約した到達性リストは，全データでのクラスタリング構造を失わないため，ビューポート内においては全データの到達性リストと同様に扱うことができる．よって，ユーザの指定する推薦件数を k として，集約した到達性リストから安定性の高い上位 k 個の ϵ_c^{appear} を推薦することで，ビューポートに応じた複数の ϵ を推薦できる．

4.4 OPTICS の可視化インタフェースの提案

事前に計算した到達性リストを利用し，地図や距離パラメータのユーザ操作に応じて，到達性プロットを集約し，クラスタリング結果の更新を即時に行う，可視化インタフェースを提案する．加えて，集約した到達性プロットでは推薦した距離パラメータ ϵ を示す．

可視化インタフェースの画面構成を図 7 に示す．上部は地図であり，地点データと，各クラスタを囲む凸包を描画する．地図は，ズーム，スクロール，回転のビューポート操作が可能である．下部は，集約した到達性プロットであり，ビューポート内のクラスタリング構造を可視化する．推薦する ϵ は，その安定性を計算したクラスタのクラスタリング順序の範囲を緑線で示す．加えて，現在の ϵ は，到達性プロット上の横長の操作バーで表され，ドラッグ操作が可能である．ϵ 操作バーを上下することで，クラスタリング結果を更新する．

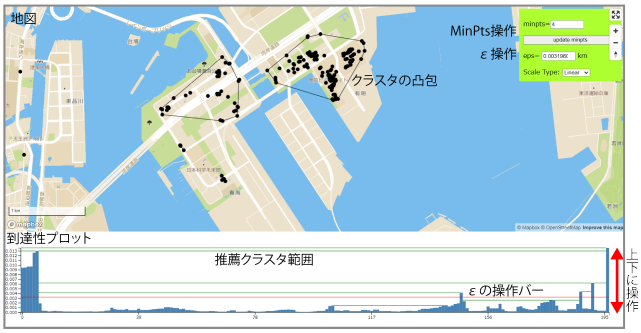


図 7: 可視化インタフェースの構成．地図操作によってビューポートを操作し，ϵ バーの操作によってクラスタリング結果を更新する

可視化インタフェースの動作手順を図 8 に示す．提案インタフェースでは，ユーザに対する応答性を上げるため，一部処理をサーバで行い，クライアント側での応答性を向上させる．ユーザは，操作とクラスタリング結果の確認を交互に繰り返して分析を進める．

サーバ側の事前処理として，D から，ユーザが指定可能な複数の $MinPts$ に対する各到達性リストを生成する．以降は，クライアント側のオンライン処理である．まず，ユーザの指定した $MinPts$ の到達性リストをサーバから取得し，ϵ 初期値によるクラスタラベルを割り当て，地図上への点群とクラスタ凸包の描画，到達性プロットの描画を行う．ユーザによる操作は，地図操作，到達性プロット上での ϵ 操作，$MinPts$ 操作である．地図操作後は，ビューポートが変化するため，到達性リストの集約を行う．ϵ 操作後は，クラスタラベルの割り当てを行う．$MinPts$ 操作後は，到達性リストの取得から再処理する．

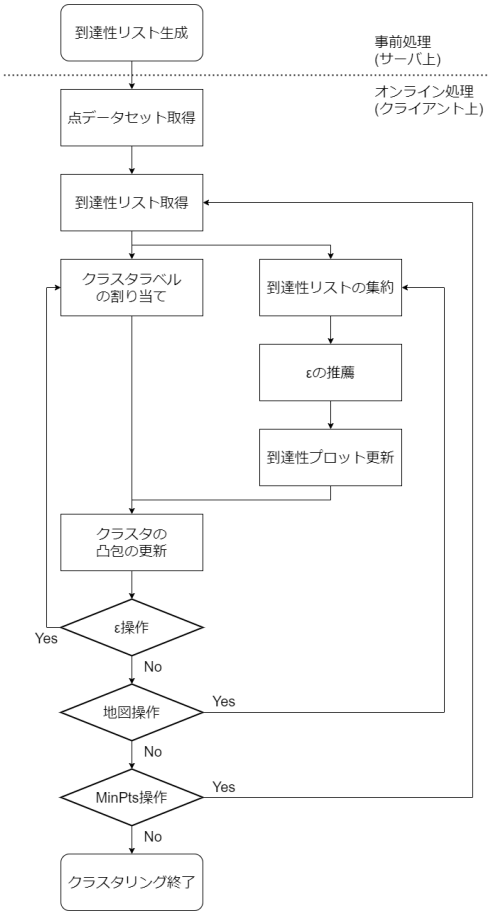


図 8: 可視化インタフェースの動作手順

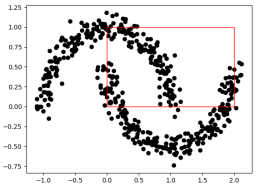
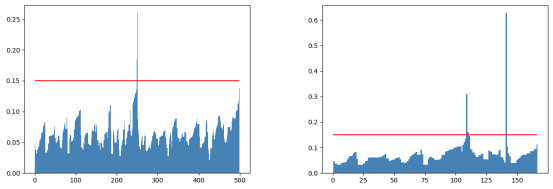
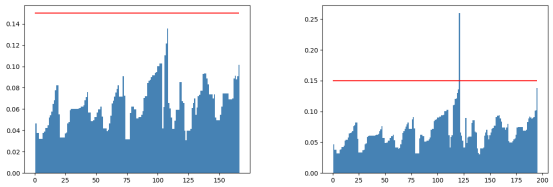


図 9: 点データ (黒点) とビューポート (赤枠)



(a) データ全体を対象に生成した到達性プロット　(b) ビューポート内の点を対象に生成した到達性プロット



(c) 図 10a からビューポート内の点に絞った到達性プロット　(d) ビューポート外仮想点を使った到達性プロット

図 10: それぞれの集約方法での到達性プロット．$\epsilon = 0.15$(赤線) において，図 10a は 2 つ，図 10b は 3 つ，図 10c は 1 つ，図 10d は 2 つ確認できる

5　適 用 結 果

5.1　ビューポート外仮想点の観察

データ全体からズームインするときの集約方法ごとの到達性プロットを比較し，ビューポート外仮想点の効果を観察する．

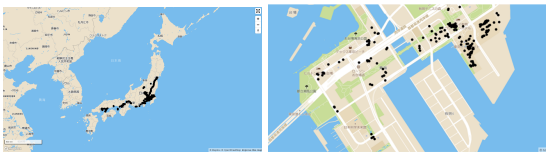
図 9 に，点データとズームするビューポートの外枠を示す．点データは，2 つの比較的高密度な点群を含む．図 10a は，点データ全体を対象に生成した到達性プロットである．$\epsilon = 0.15$ では,「谷」を 2 つ観察することができ，点データの 2 つの高密度な点群に対応する．ここから，ビューポート内の点のみを対象として生成した到達性プロットが図 10b である．ビューポート外の点が辿れなくなることで，クラスタリング構造は変化した．よって，縦軸最大値が大きくなり，加えて，$\epsilon = 0.15$ において,「谷」は 3 つ観察されてしまう．また，データ全体の到達性プロットからビューポート内の点のみの到達性プロットにすると，ビューポート外の点の情報が失われ，$\epsilon = 0.15$ において,「谷」は 1 つと見えてしまう．(図 10c) 提案手法である，ビューポート外仮想点を使って集約した到達性プロットは，ビューポート外仮想点がデータ全体のクラスタリング構造を示すことで,「谷」が 2 つと正しく認識することできた．

5.2　提案インタフェースの利用例

提案インタフェースを利用して，地点データが探索的に分析できることを観察する．

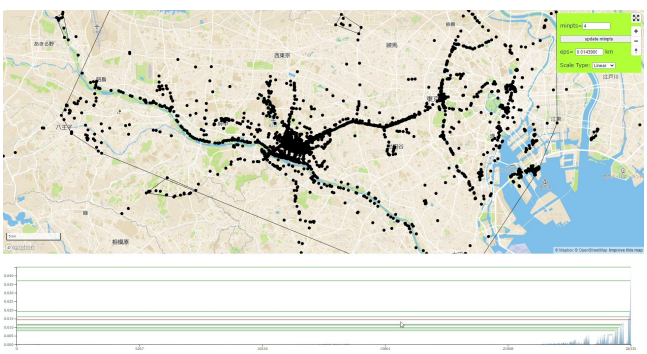
提案する可視化インタフェースを，ウェブブラウザで動作するクライアント・サーバ型のウェブアプリケーションとして実装した．サーバ側では，データベースに MongoDB を用い，地点データと到達性リストを提供するデータ API の実装に flask を用いた．クライアント側では，到達性プロットの描画に D3.js，地図描画に Mapbox API を用いた．

地点データとして，自身の 2017 年 4 月から 2022 年 7 月の約 5 年分の Google Location History(GLH) データ [10] を使用した．GLH データは，スマートフォンの位置情報を，複数の方法により，バックグラウンドで継続的に取得し，記録されたデータである．GLH データから地点データとして利用できる 30,359 点を抽出して利用した．図 11 に例を示す．

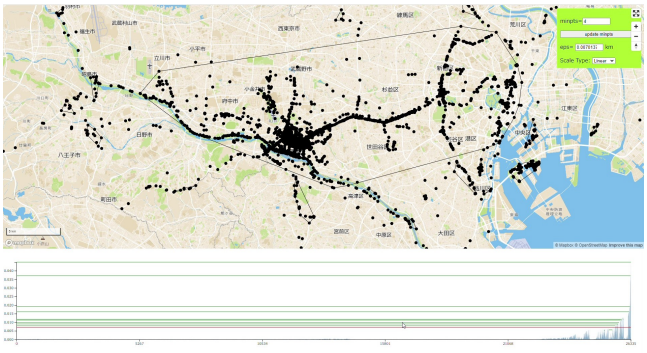


(a) 日本全体　(b) 青海付近

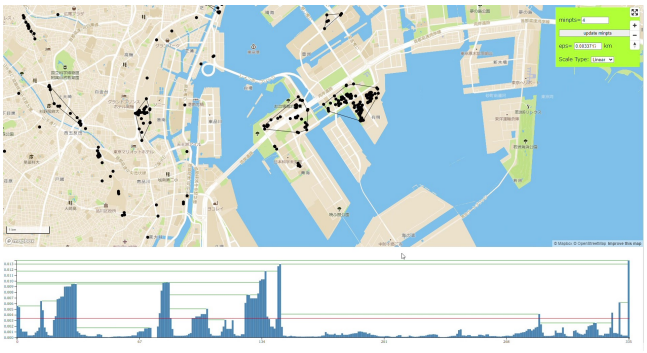
図 11: GLH の地点データの地図へのプロット



(a) 初期状態．ϵ が比較的大きく，範囲の大きなクラスタ



(b) 図 12a から ϵ を減少



(c) 青海付近にズームし，ϵ を調整



(d) 国際展示場付近にズームし，ϵ を調整．クラスタリング終了

図 12: 提案インタフェースを使った GLH データの分析．($k = 20$) 推薦された ϵ を参考に，ϵ を調整して，対象地域を絞ることと，ズーム操作を反復して，探索的に分析

前述の GLH データを提案インタフェースを用いて，期間を空けて複数回訪れるような場所について，より詳細な高頻度に訪れている場所を分析する．

初期状態は図 12a であり，比較的大きい ϵ に設定しているため，ビューポートに対し，比較的範囲の大きいクラスタであった．推薦された ϵ のうち，より小さい値に調節した状態が図 12b である．どちらの図でも，ビューポート内の点が多いため，到達性プロットの正確な読み取りは難しいにも関わらず，ϵ の推薦によって，ϵ の調整を行うことができた．図 12c は，青海付近のクラスタにズームした状態である．ズームしたことで，到達性プロットの集約により，点数と縦軸最大値は減少した．さらに，ビッグサイト付近へズームを行い，ϵ を調節した状態が図 12d である．ビッグサイト付近では，会場前の待機列，東 4・5 ホールの移動，西・南ホールの移動が多いことが分かった．このうち，ホール間の移動が会場前と同等の密度で訪れていることは，意外な結果であった．

6 評　　価

提案した到達性プロットの集約描画手法による可読性向上の性能を，複数の観点で定量的に評価する．評価において，5.2 のデータを用い，ビューポートは東京都江東区青海付近を中心座標とした．

6.1 到達性プロットの集約の評価

ビューポート外仮想点による到達性プロットの集約の効果を評価する．各ズームレベルにおける到達性プロットの点数を計算した．図 13 にズームレベルと到達性プロットの関係を示す．黒の破線は集約前，青の実線が集約後の到達性プロットである．この図から，ズームインするほど，提案した集約処理により，到達性プロットの点数は減少することが分かる．また，ズームレベル 11 から大きく減少していることから，ビューポート内の点が少ないほど集約の効果が大きくなることが分かる．

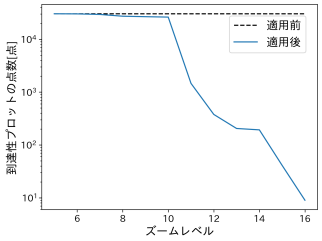
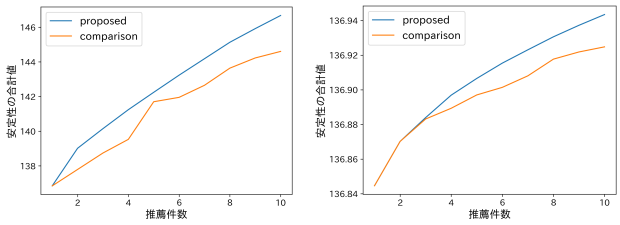


図 13: ズームレベルと描画点数

6.2 パラメータ推薦に用いられるクラスタの安定性の評価

推薦した ϵ がより大きな安定性に基づいていることを評価する．2 つのズームレベルにおいて，提案手法と比較手法の安定性の合計値を推薦件数 k を変えて比較する．ここで，比較手法は，クラスタの樹形図の根から順に k 個のクラスタを選択する

ことである．図 14 に，推薦件数と，クラスタの安定性の合計値の関係を示す．青線が提案手法，橙線が比較手法である．提案手法は，ズームレベルと推薦件数によらず，安定性の合計値がより高くなると分かる．



(a) ズームレベル 5　　(b) ズームレベル 14

図 14: 推薦件数と安定性の合計値

7 ま と め

本稿では，大規模な地点データを対象に，場所ごとの密度の違いを考慮した，OPTICS によるインタラクティブ・クラスタリングを，地図と到達性プロット上での自然な操作により実現する可視化インタフェースを提案した．あわせて，到達性プロットの集約表示方法と距離パラメータの推薦手法を提案した．到達性プロットの集約表示手法では，ビューポート外の点列を，仮想点に置き換えることで，全データでのクラスタリング構造を維持しながら，到達性プロットの点数を削減し，縦軸最大値を減少させた．距離パラメータの推薦手法では，クラスタの安定性を定義し，安定性の大きいクラスタの距離パラメータを複数推薦した．提案システムを長期間にわたる自身の地点データを用いて，ズーム前後における到達性プロットと，推薦された距離パラメータでのクラスタリング結果から，定性的に確認した．加えて，ズームレベルと到達性プロットの点数，推薦件数と安定性の合計値から，提案手法の有効性を定量的に確認した．

提案手法において，クラスタの安定性はビューポートごとに集約した到達性リストで計算を行うため，ビューポートの変化に対し，重複して計算している部分がある．今後の展望として，より高速に動作させるため，この重複部分を減らす手法の検討が挙げられる．

文　献

[1] Gennady Andrienko and Natalia Andrienko. Interactive cluster analysis of diverse types of spatiotemporal data. *Acm Sigkdd Explorations Newsletter*, 11(2):19–28, 2010.

[2] Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, and Dino Pedreschi. A visual analytics toolkit for cluster-based classification of mobility data. In *Advances in Spatial and Temporal Databases: 11th International Symposium, SSTD 2009 Aalborg, Denmark, July 8-10, 2009 Proceedings 11*, pages 432–435. Springer, 2009.

[3] Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, Dino Pedreschi, and Fosca Giannotti. Interactive visual clustering of large collections of trajectories. In *2009 IEEE Symposium on visual analytics science and technology*, pages 3–10. IEEE, 2009.

[4] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proc. SIGMOD Conference 1999*, pages 49–60, 1999.

[5] Juhee Bae, Tove Helldin, Maria Riveiro, Sławomir Nowaczyk, Mohamed-Rafik Bouguelia, and Göran Falkman. Interactive clustering: A comprehensive review. *ACM Computing Surveys (CSUR)*, 53(1):1–39, 2020.

[6] Stefan Brecheisen, Hans-Peter Kriegel, Matthias Schubert, and Michael Gruber. Vico: Visualizing connected object orderings. In *Advances in Database Technology-EDBT 2006: 10th International Conference on Extending Database Technology, Munich, Germany, March 26-31, 2006 10*, pages 1151–1154. Springer, 2006.

[7] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proc. PAKDD*, pages 160–172, 2013.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[9] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2:232–246, 12 2003.

[10] Andrew Kirmse, Tushar Udeshi, Pablo Bellver, and Jim Shuma. Extracting patterns from location history. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 397–400, 2011.

[11] Deborah Lupton. *The quantified self*. John Wiley & Sons, 2016.

[12] Jonathan K. Nelson. Cartography & geovisual analytics in personal contexts: designing for the data creator. *International Journal of Cartography*, 0(0):1–21, 2023.

[13] Florian Richter, Yifeng Lu, Daniyal Kazempour, and Thomas Seidl. “show me the crowds!” revealing cluster structures through amtics. *Data Science and Engineering*, 5:360–374, 2020.

[14] Salvatore Rinzivillo, Dino Pedreschi, Mirco Nanni, Fosca Giannotti, Natalia Andrienko, and Gennady Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3-4):225–239, 2008.

[15] Robert St. Amant and Paul R Cohen. Interaction with a mixed-initiative system for exploratory data analysis. In *Proceedings of the 2nd international conference on Intelligent user interfaces*, pages 15–22, 1997.

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

地理情報

座長:角谷 和俊(関西学院大学)
コメンテータ:横山 昌平(東京都立大学)
2024年3月1日(金) 16:00 〜 18:10 T4-A (オンライン（Zoom Events))

17:35 〜 18:00

[T4-A-9-05]【技術報告】情報学と交通工学の融合に向けた研究開発 - 万博での交通整流化に向けて

西村拓哉 (株式会社NTTデータグループ)

NTTグループは主に2025年大阪万博で発生すると目されている交通渋滞の解消・提言をターゲットとして、阪神高速道路株式会社と共同検討を実施中である。
本技術報告ではその共同検討の中で取り組んでいる技術開発について紹介する。
具体的にはシミュレーションや施策評価のための交通デジタルツインの開発、ユーザに対する働きかけの最適化、全体のシステム構築などが挙げられる。
データドリブンで精度重視の情報学とモデルドリブンで説明性重視の交通工学の協業によって見えてきた課題や今後の展望についても述べる。

©一般社団法人日本データベース学会

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

2024年3月1日(金) 16:00 ~ 18:10 | T5-A オンライン（Zoom Events)

スポーツ

座長:太田 学(岡山大学)
コメンテータ:木村 昌臣(芝浦工業大学)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T5-A-9-01] 共通テスト用プログラム表記で記述された運動イメージと身体動作の差異認識アプリケーションの試作

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

16:20 ~ 16:35
[T5-A-9-02] 卓球の得点予測における重要要素の分析

*楊 明哲[1,2]、橋本 敦史[2]、馬 家昕[2]、本田 秀仁[3]、田中 一敏[2] (1. 東京大学、2. オムロンサイニックエックス、3. 追手門学院大学)

16:35 ~ 17:00
[T5-A-9-03] バドミントンの試合データを用いたショットの成功確率予測

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

17:00 ~ 17:25
[T5-A-9-04] ボール奪取と攻撃の防御に基づくアクションによるサッカー選手の攻守評価

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

17:25 ~ 17:50
[T5-A-9-05] 音響型ARオンラインマラソンシステムにおける空間認識率向上に向けた取り組み

*小西 侑樹[1]、Siriaraya Panote[2]、栗 達[3]、田中 克己[4]、河合 由起子[5]、中島 伸介[5] (1. 京都産業大学大学院、2. 京都工芸繊維大学、3. 福岡大学、4. 福知山大学、5. 京都産業大学)

共通テスト用プログラム表記で記述された運動イメージと身体動作の差異認識アプリケーションの試作

角 颯真†, 伊藤 一成†

† 青山学院大学社会情報学部 〒252-5258 神奈川県相模原市中央区淵野辺 5 丁目 10-1

あらまし 高等学校では，2022 年度から「情報 I」が必履修科目として運用されている．大学入学共通テストでは，2025 年度から教科「情報」が導入される．試験範囲にプログラミング分野を含むが，そこでは Python 言語をベースとした共通テスト用プログラム表記が擬似コード仕様として用いられる．人型ピクトグラムを利用して，共通テスト用プログラム表記の書式で記述された擬似コードで与えられた運動イメージと，利用者が再現した身体動作の差異の認識を体感できるアプリケーションを試作したので報告する．

キーワード 共通テスト用プログラム表記，プログラミング，アルゴリズム，人型ピクトグラム，身体動作

1. はじめに

高等学校では，2022 年度から情報科において「情報 I」が必履修科目として運用されている[1]．情報 I の学習単元の一つである「コンピュータとプログラミング」では，Python, JavaScript, VBA, Scratch の利用が主に想定されている[2]．さらに，大学入学共通テストでは，2025 年度から情報 I の範囲を対象とする教科「情報」が導入される予定である．プログラミングの問題では，Python 言語をベースとした共通テスト用プログラム表記が擬似コード仕様として用いられる予定であり，すでにサンプル問題や試作問題が公開されている[3][4]．そこで共通テスト用プログラム表記で記述された擬似コードを用いて，実際の身体動作と運動イメージの差異を理解させることにより，プログラムの理解を促進することを目的としたアプリケーション「ヒューマンピクトグラミング」を設計・試作したので報告する．

コンテンツの記述には人型ピクトグラムを採用する．ピクトグラムとは，日本語で絵記号，図記号と呼ばれるグラフィックシンボルであり，意味するものの形状を使ってその意味概念を理解させる記号である[5]．ピクトグラムはコンテンツ表現の抽象度の高さから，それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果があると言われている．有名な「非常口」ピクトグラムのデザインの策定に関わった太田は“走る人型を囲む空間が見る人を包む空間とつながって走る人は見る人の投影になる”と述べ[5]，実際に避難中の人が，出口へ向かって走る人型ピクトグラムを見たときに，いかに自身をピクトグラムに投影し，同一視するか，という観点にデザインの労力が払われた．この特性を利用して我々の研究グループでは人型ピクトグラムを活用した各種学習アプリケーションを開発し，インターネット上で公開している[6][7][8]．

以下 2 章で試作したアプリケーションについて説明し，3 章で先行研究も踏まえて考察する．4 章でまとめと今後の展望について述べる．

2. ヒューマンピクトグラミング

2.1 概要

今回試作したヒューマンピクトグラミングは，共通テスト用プログラム表記の書式で記述された身体動作の指示を利用者に提示し，それをもとに身体動作を再現してもらい，再現された身体動作に対してのフィードバックを与えるウェブアプリケーションである．コンピュータに搭載された単眼カメラで撮影した人間の姿勢の情報を取得することで，利用者が指示された身体動作を再現し比較する．ヒューマンピクトグラミングには 3 つの進行状態がある．それぞれ順に，(1)身体動作理解段階・(2)身体動作再現段階・(3)結果提示段階である．以下，2.2 節から 2.4 節までそれぞれの段階について順に解説する．

2.2 (1)身体動作理解段階

身体動作理解段階において，利用者は共通テスト用プログラム表記の書式で記述された身体動作の指示を理解する必要がある．身体動作理解段階時のアプリケーションのスクリーンショットを図 1 に示す．図 1 において，画面左側に位置している擬似コードエリアに身体動作の指示を共通テスト用プログラム表記の書式で記述した擬似コードで提示する．



図 1 身体動作理解段階時のアプリケーションのスクリーンショット

画面右側のカメラエリアにはデバイスに搭載されたカメラから取得した映像が表示される．利用者が提示された擬似コードを理解した後，擬似コードエリア下部の「START ボタン」を押下すると次の身体動作再現段階に移行する．身体動作再現段階に移行すると擬似コードエリアが表示されなくなる．これは擬似コードによって指示された身体動作を利用者が運動イメージに変換してから再現を行ってもらうためである．

図 2 は提案アプリケーション内で擬似コードを用いて指示される身体動作の例である．

```
1  part = ""
2  n = 2
3  clockwise = 0
4  もし n % 2 == 0 ならば:
5  ｜ clockwise = -1
6  ｜ part = "左肩"
7  そうでなければ:
8  ｜ clockwise = 1
9  ｜ part = "右肩"
10 x を 0 から 3 まで 1 ずつ増やしながら繰り返す:
11 ｜ 回転して待つ(part, clockwise * 145, 0.5)
12 ｜ 回転して待つ(part, clockwise * -145, 0.5)
```

図 2 指示される身体動作の表示例
（行番号は説明上付与したもので，実際には表示されない）

変数や条件分岐，繰り返しなどに関しては共通テスト用プログラム表記の書式に準じている．図 2 の 1 行目から 3 行目では変数に値を代入している．4 行目から 9 行目では条件式の評価結果に基づいて条件分岐し変数の値の書き換えを行っている．10 行目から 12 行目までは 4 回スコープ内の命令を繰り返す．組込み関数として「回転する()」と「回転して待つ()」の 2 つを用意した．これらの関数は第一引数に動作させる体の部位，第二引数に回転させる角度，第三引数に動作させる時間を与える．11 行目は，変数 part の代入されている部位を反時計回りに変数 clockwise と 145 の積の値分の角度を，0.5 秒間かけて回転する．ただし，「回転して待つ」という関数を使用しており，これは命令の実行が終了するまで，次の命令の実行を開始しないことを示す．一方「待つ」がないと次の命令と同時に実行する．「待つ」の有無により，逐次処理と並列処理を組み合わせることで，様々な動きを実現できる．図 2 で示した擬似コード例では，n の値に応じて，左または右の腕を振り上げ，振り下ろしを行うという命令である．

正面方向のピクトグラムを図 3 に示す．体と頭を組み合わせた部分が 1 つと，肩，肘，股，膝が左右それぞれ 1 つの計 9 種の部品から構成される．

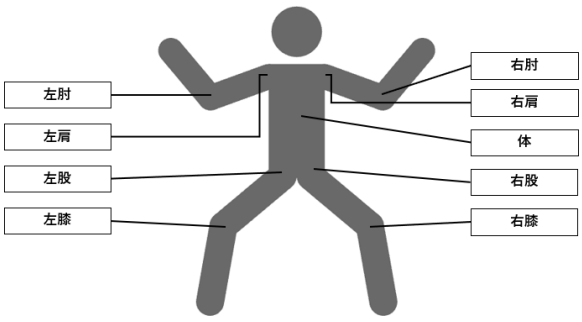


図 3 人型ピクトグラムの部位を回転する際の支点

身体動作の指示の擬似コードは擬似コードエリア上部にある選択メニューから選択することができる．選択できる擬似コードにはレベルとして 1 から 7 までのいずれかの整数値が割り振られている．レベルの値はその擬似コードの難易度によって割り振られる．表 1 は擬似コードに割り振られる各値と擬似コードの内容を表したものである．レベルは，そのレベルの要素を含み，そのレベルより高いレベルの要素は含まない．また，そのレベルより低いレベルの要素は含まれていてもいなくても良い．

表 1 擬似コードの各レベルと含まれるプログラミングの諸概念

レベル	擬似コードの内容
1	単一の身体部位を動かす身体動作．
2	2 つ以上の身体部位を逐次，または並列で動かす身体動作．
3	繰り返し表現の使用．
4	変数の使用．
5	配列変数の使用．
6	算術演算の使用．
7	条件分岐の使用．

擬似コードは Google スプレッドシート上で管理される．試作アプリケーションが管理する擬似コードのデータ属性を表 2 に示す．

表 2 擬似コードデータの属性

属性名	定義
id	擬似コードを識別する文字列．
title	擬似コードに設定された表題を示す文字列．
code	共通テスト用プログラム表記で記述された擬似コード．
time	擬似コードに記載された身体動作にかかる時間を表す文字列．
simpleCode	結果提示段階で人型ピクトグラムが実際に行う動作を記述したもの．
level	擬似コードの内容に基づいて割り振られた難易度を示す文字列．その値は 1 から 7 までの整数値のいずれかである．

第三者がスプレッドシート上のデータを改変することで，問題バンクを更新することができる．

2.3 (2)身体動作再現段階

身体動作再現段階では利用者に指示された動きの再現を行ってもらう．身体動作再現段階時のアプリケーションのスクリーンショットを図 4，図 5 に示す．身体動作再現段階に移行すると，はじめに図 4 に示すように画面左側に 10 秒間のカウントダウンが表示される．カウントダウンが終了するまでに利用者は一度直立の姿勢をとる．



図 4 身体動作再現段階時のアプリケーションのスクリーンショット(動作再現開始前状態時)

カウントダウンが終了すると図 5 のように画面が変更される．この状態になったら，身体動作理解段階で指示された身体動作の再現を行う．



図 5 身体動作再現段階時のアプリケーションのスクリーンショット(動作再現時)

擬似コードで指示された身体動作を行うのにかかる時間が経過すると，2.4 節で示す，結果提示段階に移行する．

2.4 (3)結果提示段階

結果提示段階では利用者が行った身体動作の再現のフィードバックを返す．結果提示段階時のアプリケーションのスクリーンショットを図 6 に示す．



図 6 結果提示段階時のアプリケーションのスクリーンショット

結果提示段階時では，擬似コードエリアが再び画面左側に表示される．また，カメラエリアに代わり録画映像エリアと正解動作再現ピクトグラムが表示される．録画映像エリアには利用者が身体動作再現段階で再現した身体動作の録画映像が表示される．正解動作再現ピクトグラムは擬似コードエリアで提示された擬似コードを忠実に再現してくれる．利用者はこの 2 つを比較することで身体動作の再現が正しく行えていたかどうかを振り返ることができる．録画映像エリアの下部にある「PLAY ボタン」を押下すると，もう一度正しい身体動作を行う人型ピクトグラムと利用者の再現した身体動作の録画を視聴できる．擬似コードエリアの下にある「END ボタン」を押下すると再び身体動作理解段階に戻る．

3. 関連研究・考察

本章では試作したアプリケーションについて，関連研究を踏まえて考察する．

ヒューマンピクトグラミングでは擬似コードで記述された身体動作を利用者自身の身体で再現する．擬似コードの内容を正しく理解できないと正しい身体動作を再現することはできない．身体動作の再現が，擬似コードを理解できているかどうかの指標になると考えられる．

また，擬似コードを理解して実際に身体を動かす過程で，コードとその動作の対応関係が明確となると考えられる．松田らは，動作とプログラムコード，動作とプログラム仕様の対応関係について説明することが重要であるとし，再帰プログラムの教育支援システムの開発を行った[9]．岡本らは，プログラミングの概念理解に関して，出力の動作が物理的現象として見えるだけではなく学習者がそれを認識していなければ学習上の効果が期待できないと述べ，視覚的に顕在化することの重要性を指摘している．そして視認性，判別性，予測可能性，独立性の 4 つの状態から評価し有効性を示し，プログラミング教材作成の際の指針として提案している[10]．ヒューマンピクトグラミングでは利用者の再現した身体動作が擬似コードの出力となる．自身の身体動作が出力となるため，視認性，判別性，予測可能性が高い．擬

似コードに関して,1 行に多くて 1 つの身体部位に対する命令が記述されるため独立性も高いと言える.

ヒューマンピクトグラミングは身体を動かしながら行う.運動による認知機能向上の効果を得ることが期待できる.Schmidt らはランニングマシン上で歩行しながら単語の暗記学習を行ったときの方が,椅子に座りながら学習した時よりも成績が良かったことを報告している[11].Oppezzo らは歩くことで,その最中および直後に創造的なアイデアが促進されることを実証している[12].運動によって認知機能が向上し,コード理解の一助となる可能性がある.

擬似コードの書式として共通テスト用プログラム表記を採用している.共通テスト用プログラム表記に慣れる目的での利用が考えられる.

一方で,試作アプリケーションの利用でプログラミング技能や実践力が軽視されてしまう可能性がある.ヒューマンピクトグラミングではコードの入力作業を必要としない.また,身体動作の指示に擬似コードを利用している.実際にプログラミングを行う経験が大切であり,試作アプリケーション単体での利用は慎重を期す必要がある.

コンテンツの記述には人型ピクトグラムを利用している.正しい身体動作を提示する際に余計な情報を含まずに身体動作に関する情報を利用者に伝達することができる.また,ピクトグラムの,それを見た人物が自分自身や本人に関わる人物事物などを想起させる効果[5]を利用できる.正しい身体動作を行う人型ピクトグラムに自分自身を投射することで,自分自身が正しい身体動作を行う場面を想起させることが可能である.

4. まとめと今後の展望

本稿では,実際の身体動作と運動イメージの差異を理解させることにより,プログラムの理解を促進することを目的としたアプリケーションの試作を行った.今後はアプリケーションの評価を進めつつ機能や要件を再構成していく予定である.

謝辞 本研究は JSPS 科研費 21H03560 の助成を受けたものです.

参 考 文 献

[1] 平成 30 年告示高等学校学習指導要領 http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2018/07/11/1384661_6_1_2.pdf（2023 年 11 月 10 日閲覧）

[2] 高等学校情報科「情報 I」教員研修用教材(本編) https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1416756.htm（2023 年 11 月 10 日閲覧）

[3] 大学入学共通テスト「情報」サンプル問題 https://www.mext.go.jp/content/20211014-mxt_daigakuc02-000018441_9.pdf（2023 年 11 月 10 日閲覧）

[4] 大学入学共通テスト試作問題「情報 I」 https://www.dnc.ac.jp/albums/abm.php?d=511&f=abm00003277.pdf&n=6-2-1_%E8%A9%A6%E4%BD%9C%E5%95%8F%E9%A1%8C%E3%80%8E%E6%83%85%E5%A0%B1%E2%85%A0%E3%80%8F%E2%80%BB%E4%BB%A4%E5%92%8C4%E5%B9%B412%E6%9C%8823%E6%97%A5%E4%B8%80%E9%83%A8%E4%BF%AE%E6%AD%A3.pdf （2023 年 11 月 10 日閲覧）

[5] 太田幸夫: 国際安全標識のピクトグラムでデザインの研究, http://www.tamabi.ac.jp/soumu/gai/hojo/seika/2003/kyoudou-ota1.pdf （2023 年 11 月 10 日閲覧）

[6] 伊藤一成: 複数のプログラミング言語で記述可能なピクトグラムコンテンツ作成環境の提案と実装, 情報処理学会論文誌 TCE, Vol.7, No.3, pp.1-11 (2021)

[7] 渡辺大智, 御家雄一, 伊藤一成：人型ピクトグラムを用いたソートアルゴリズムを学ぶ アプリケーション「人型ピクソートグラム」の拡張と評価, 第 15 回データ工学と情報マネジメントに関するフォーラム DEIM2023 (2023)

[8] 高橋伶奈, 御家雄一, 伊藤一成：ピクトグラフィングーデータ活用の学習を目的としたピクトグラフ生成アプリケーションの実装ー, 第 15 回データ工学と情報マネジメントに関するフォーラム DEIM2023 (2023)

[9] 松田憲幸, 柏原昭博, 平嶋宗, 豊田順一: プログラムの振舞いに基づく再帰プログラミングの教育支援, 電子情報通信学会論文誌 D, 80(1), 326-335. (1997)

[10] 岡本雅子, 村上正行, 吉川直人, 喜多一:「視覚的顕在化」に着目したプログラミング学習教材の開発と評価, 日本教育工学会論文誌, Vol.37, pp.35-45 (2013)

[11] Schmidt-Kassow, M., Zink, N., Mock, J., Thiel, C., Vogt, L., Abel, C. Kaiser, J: Treadmill walking during vocabulary encoding improves verbal long-term memory, Behavioral and Brain Functions, Vol. 10, No. 1, p. 24 (2014)

[12] Oppezzo, M., Schwartz, D. L: Give your ideas some legs: the positive effect of walking on creative thinking. Journal of experimental psychology: learning, memory, and cognition, 40(4), 1142 (2014)

卓球の得点予測における重要要素の分析

楊　明哲[†,††]　橋本　敦史[††]　Jiaxin　Ma[††]　本田　秀仁[†††]　田中　一敏[††]

† 東京大学
†† オムロンサイニックエックス
††† 追手門学院大学
E-mail: †mingzhe-yang@g.ecc.u-tokyo.ac.jp

あらまし　スポーツのデータ分析において、機械学習モデルを用いた勝敗予測や戦術解析が盛んに行われている。サッカーなどを題材とする既存研究では、試合全体を考慮した戦略に注目が集まっている。一方で、卓球のような一定の得点獲得を目指す競技では１つ１つのプレーに焦点を当てた分析が求められている。そこで本研究では、試合データをもとに各得点における重要な要素を推定することを試みた。具体的には、外部から観測可能な得点状況などの試合データをもとに機械学習モデルによる重要特徴量の推定を行い分析する。さらに本研究では、選手の行動意図を反映した攻守ラベルを卓球経験者によって作成し、得点獲得にどのように寄与するか分析を行った。

キーワード

1　はじめに

スポーツのデータ分析において、機械学習を用いることが盛んに行われている。テニスやサッカーなどを対象とした既存研究では、試合における勝敗予測や重要なイベントの推定が行われてきた [2], [3]。これらは外部から観測可能な情報であるゲームや試合における得点やサーブ権を特徴量として、勝敗予測を行なった。これらの研究では、勝敗予測の重要な要素として、得点数の多さやレシーバーによる得点率の高かさなどを挙げた [3]。これらのスポーツにおける重要要素の推定は、いずれも試合全体に対する戦術アドバイスであった。卓球のように一定の得点の獲得を目指す競技では、一打一打のプレーに対する戦術的アドバイスが重要である。一打一打に注目した分析を行うために、卓球競技における既存研究では、飛行特性、回転特性や配球特性など物理的な球に対する試みは行われている [4]。卓球競技において外部から観察できる情報に基づいた分析は行われているが、外部から直接観測ができない情報を用いた分析は行われていない。そこで本研究では、各プレーに注目した特徴量として、選手の内部状態に注目した攻守ラベル [5] を用いる。攻守ラベルとは、選手がそのプレーを行う際に、得点するために攻撃的なプレーをしたかどうかである。本研究では、卓球経験者による選手の行動に対して、付与した攻守ラベルを用いることで、各プレーにおける重要要素の推定を行った。

2　データセット

2.1　競技の選定と試合状況のデータセット

本研究では、対象の競技として卓球を採用した。卓球は、一定の得点獲得を目指すであるため一打一打のプレーに対して、戦術的アドバイスが重要である。また、卓球は、サーブによってプレーの流れが変化する競技である。このような特徴を持つ

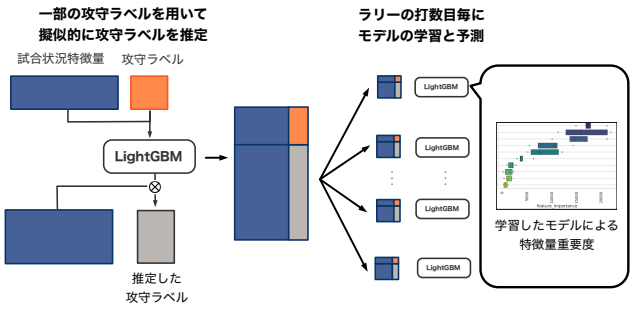


図 1: 本実験における全体の流れ。初めに一部の攻守ラベルを教師として、分類器の学習を行う。その後学習された分類器によって擬似的に攻守ラベルを作成する。次にデータセットを打球目について分割し分類器の学習を行う。

卓球を対象とすることで、各プレーに対する重要要素の推定を行うことができる。本研究では、平成 27 年度および 28 年度の全日本卓球選手権ジュニア男子，全日本選手権男子，女子のシングルス，合計 45 試合に対してデータセット作成を行った。データセット作成には、上記の試合映像をもとに、試合におけるサーブ権、現在の得点状況などの特徴量を人間の手作業によって収集した。表 1 に本研究で用いた特徴量についてまとめる。各特徴量は、すべてサービスを行う選手を起点としたものに特徴量変換を行っている。これによって、分析する際にサーバー、レシーバーの視点から重要要素の分析が可能になる。

2.2　攻守ラベルの付与

本研究では、得点に向けたプレーの重要要素の分析を行うために、試合状況の特徴量に追加して、選手のプレーに対する攻守ラベルの利用を考える。本研究では攻守ラベルを、選手のプレーの意図が得点のために攻めているか否かと定義した。この攻守ラベルの判断は、卓球のルールだけでなく、戦術に関する深い知識が必要だと考え、ラベル作成には卓球経験者を用いた。

表 1: 作成した特徴量一覧。サーバーを S、レシーバーを R とする。

特徴量	内容	変数タイプ
Set	試合中の何ゲーム目か	数値
Play	同ゲーム内で何本目のラリーか	数値
NRallyOrdered	現在の打球がラリー中の何打目か	数値
ScoreStatusFromServer	S の現在の得点	数値
ScoreStatusFromReciever	R の現在の得点	数値
Server	現在の打球が S のものか否かの 2 種類	カテゴリカル
ScoreStatusDiffBinaryFromServer	S と R の得点差があるかどうか。1） S のほうが得点が高い 2）R のほうが得点が高い 3）同点の 3 種類	カテゴリカル
AttackLabelsFromServer	S に攻撃意図があるか否か、R に攻撃意図があるか否かの 4 種類	カテゴリカル

アノテータとしては中学生から現在まで選手として卓球に携わっている大学生 2 名 (男女 1 名ずつ) に依頼した。アノテータには、利用した試合データのうち、5 本の動画に対してアノテーション付与を行った。アノテーション対象動画を 5 本のみにした理由としては、今回の作成したラベルは選手の一打ごとに攻守ラベルを付与するものであり、すべての動画に対してラベル作成するのは非常にコストがかかるためである。アノテータが付与したラベルの品質を確認するために、アノテータ同士のラベル一致率を計算した。アノテータによるラベル一致率は、0.934 ± 0.04 であった。アノテータのラベルの不一致は、アノテータの作業精度や、アノテータのこれまで経験に依存しているために発生している問題であると考える。しかしながら、ラベルの一致度が 0.9 を超えていることから、少なくとも人間のによる攻守ラベルは高い精度で達成可能であると考える。

本研究では、アノテーションコストの観点から、データセットで用いる一部の試合のみにしか攻守ラベルが存在していない。そこですでに付与されている攻守ラベルを用いて、擬似的に攻守特徴量の作成を行なった。このアプローチは半教師あり学習に類似している。攻守ラベルが付与されている 5 試合のデータセットを訓練データとして、LightGBM [1] による攻守ラベルの擬似特徴量の生成を行った。LightGBM は、勾配ブースティング木を用いた機械学習モデルであり、高い精度で予測を行うことができる。

3 実　　験

3.1 分類器の学習と評価方法

本実験における全体の流れを図 1 に示す。Section2 で作成した特徴量を用いて、各プレーにおける得点を予測し、得点獲得に重要な要素の分析を行う。さらに本研究にて新たに作成した攻守ラベルはプレーの重要度にどのような影響を与え、戦術分析に有用であるかを検証する。分類器として LightGBM を用いた。LightGBM を用いた理由は、LightGBM は、勾配ブースティング木を用いた機械学習モデルであり、特徴量の重要度を計算することができるためである。この特徴量重要度は、各特徴量が予測に対してどの程度寄与しているかを表す。

作成したデータセットを試合単位で分割し、試合単位で交差検証を行った。試合単位で交差検証を行うことで、試合単位での汎化性能を評価することができる。また試合単位での交差検証を行うことで、試合単位での特徴量の重要度を計算することができる。本研究では、選手の一打一打に注目して各プレーにおける重要要素の分析を行うため、我々はサーブから打数に応じてデータセットを分けて、それぞれ異なるモデルの学習を行った。本実験では 1 ラリー内の打数に応じて、6 種類に分割した。分割したデータセットはそれぞれ 1) 1 打目（サーブ）、2) 2 打目（レシーバ）、3) 3 打目（三球目攻撃）、4) 4 打目、5）5 打以上の奇数打（サーバーの打球）、6) 6 打以上の偶数打（レシーバーの打球）とした。本研究の正解ラベルは、1 つのラリーにおいて、サーバーとレシーバーのどちらが得点するかの二値ラベルを設定した。サーバーが得点した時は 0 で、レシーバーが得点した時は 1 となる。分類器には試合状況のデータセットが与えられて、そのラリーにおいてサーバーが得点するかレシーバーが得点するかを予測を行う。モデル性能を評価する指標として二値ラベルの正解率を用いた。

表 2: 攻守ラベルありの場合となしの場合の精度比較

訓練データセット	攻守ラベルなし	攻守ラベルあり
1 打球目（サーブ）	0.660	**0.731**
2 打球目（レシーブ）	0.621	**0.733**
3 打球目（三球目攻撃）	0.625	**0.675**
4 打球目	0.635	**0.695**
5 打球目以降の奇数打目	0.489	**0.633**
6 打球目以降の偶数打目	**0.670**	0.637

4 結　　果

Section2. 2 で推定した攻守ラベルの正解率について報告する。事前に人間の手で作成した攻守ラベルを正解ラベルとし、交差検証で汎化性能を評価したところ、予測精度は 0.31 であった。

表 2 に、攻守ラベルを用いた場合と用いない場合の各打球における予測精度を示す。攻守ラベルを用いた場合は用いない場合と比べて予測精度が向上していることがわかる。このことから、機械学習モデルで得点を予測する際に攻守ラベルを用いることは精度向上に寄与することが明らかになった。

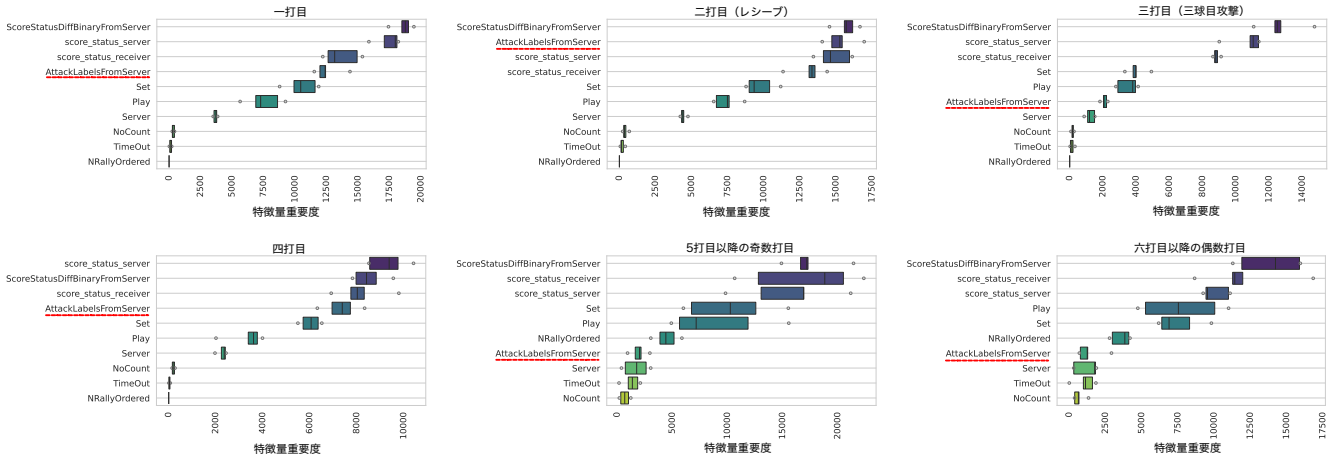


図 2: 分割したデータセット毎に学習した分類器の特徴量重要度。赤点線は、本研究で作成した攻守ラベルを指している。

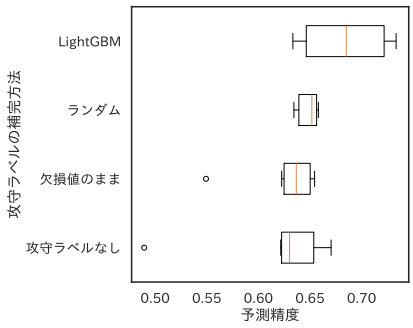


図 3: 攻守ラベルの作成方法の違いによる分類精度の比較

図 2 に、LightGBM での特徴量重要度を示す。分類器として重要な特徴量として、サーバー、レシーバーの得点状況（ScoreStatusFromServer, ScoreStatusFromResiver）、そして両者の得点差（ScoreStatusDiffBinaryFromServer）が挙げられていた。これらの特徴量は既存研究でも重要な要素として考えられており、本研究でも同様の結果が示された [3]。攻守ラベルについては、特に 2 打目や 4 打目において、分類器の予測において重要要素になっていることが確認できた。この 2 打目、4 打目はレシーバーによる打球であり、攻守ラベルが重要要素になっていることから、レシーバーによる得点獲得において、この得点獲得のために攻撃する意図が重要であると考えられる。卓球競技をはじめとするネット型競技の性質上、サーブは他者の影響を受けないためサーバーは得点する上で非常に有利である。また卓球は 3 球目攻撃と呼ばれるほど、サーバーはサーブから三球目攻撃の展開を作ることが有用な戦術である。レシーバーはこの展開を阻止するためにレシーブを行う。攻守ラベルが重要要素になっていることから、不利な状況になりやすいレシーバーは、得点獲得のために攻撃的なプレーを行うことが重要であると考えられる。

図 3 に、擬似攻守ラベルの生成方法の違いによる予測精度の変化を示す。比較対象として、攻守ラベルを欠損値として扱う方法とランダムに割り当てる方法を用いた。欠損値とランダム割り当てのいずれも分類器を用いて推定した本手法に精度が及ばなかった。本実験において分類器によって付与された攻守ラベルは、予測精度が 0.31 であり、チャンスレートの 0.25 と比べて僅かに高いものであった。ランダムや欠損値と比較すると、分類器による割り当てを行うことで、予測精度の向上が見られた。このことから、さらに精度の高い攻守ラベルを分類器に与えることで、より高い予測精度が期待できる。

5 おわりに

本研究では、卓球のプレーに対する得点予測を行うときに、選手の攻撃意図を表す攻守ラベルを用いることで、分類器の予測精度向上に寄与することを示した。また、分類器の特徴量重要度を見ることで、不利な状況になりやすいレシーバーが得点するためには、選手の攻撃意図が重要であることを示唆した。今後は、既存の選手の配球戦略と攻守ラベルを組み合わせることで、より有用な戦術分析を行うことができると考える。

謝　辞

本研究は、JST、さきがけ、 JPMJPR22C6 の支援を受けたものである。

文　献

[1] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 2017 International Conference on Neural Information Processing Systems*, 2017.

[2] Michal Sipko and William Knottenbelt. Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*, Vol. 2, , 2015.

[3] Chen Yang, Tian Yubo, and Zhong Yi. Final project report: Real time tennis match prediction using machine learning. 2017.

[4] 塩入彬允, 牛山幸彦, 佐藤悠樹. 卓球競技における得点に影響する配球の分析方法の検討. スポーツパフォーマンス研究, Vol. 14, pp. 234–242, 2022.

[5] 楊明哲, 橋本敦史, Ma Jiaxin, 本田秀仁, 田中一敏. 卓球映像からの打球の攻守推定. In *Proceedings of the 2023 Meeting on Image Recognition and Understanding*, 2023.

バドミントンの試合データを用いたショットの成功確率予測

美濃岡知樹† 清 雄一† 田原 康之† 大須賀昭彦†

† 電気通信大学 〒 182–8585 東京都調布市調布ケ丘一丁目 5 番地 1
E-mail: †minoka.tomoki@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし 近年バドミントンの世界では, 機械学習技術を用いたスポーツ分析が注目を集めている. その例として, バドミントンの試合映像を基にしたストロークの分類や, 試合映像をデータ化してラリーの結果予測などに活用する研究などが挙げられる. これらの多くの研究は選手のプレースタイルや試合の傾向を分析することを一つの目的としているが, 選手の弱点やミスの傾向に着目した研究はまだ行われていない. そこで本研究では, 選手の弱点やミスの傾向を分析するために, バドミントンの試合データを用いてショットの成功確率の予測を行うモデルを提案する. 結果は, 失敗するショットを予測するという点で提案モデルが他のシステムよりも高い精度での予測を行った. 失敗しやすいショットなどの選手固有の傾向を分析することができれば, 戦術の決定や選手への指導などへ活用されることが期待できる.

キーワード スポーツデータ活用, スポーツ分析, 深層学習, バドミントン, ラリー予測

1 はじめに

1.1 背景

近年, バドミントンの世界では機械学習技術を用いたスポーツ分析が注目を集めている. スポーツ分析とは, 自分または対戦相手の選手やチームの動きをデータ化して分析することである. 選手やコーチは分析結果をもとに, 次の戦術の決定や, トレーニングの強化を行う. 機械学習技術を用いたバドミントンのスポーツ分析の例として, Chu らはバドミントンの試合映像を基に, 機械学習技術を用いてプレイヤの検出やストロークと戦術の分類を行い, バドミントンの試合を可視化することで選手や試合の特徴理解を促すシステムを開発した [1]. また Weeratunga らは, 選手の位置を検出して戦術的な動きを抽出, 分類し, 戦術分析を行った [2]. さらに近年では, 試合映像をデータ化して, そのデータを基にラリーの結果予測を行う研究も, 様々な手法を用いて多数行われている [3] [4] [5]. これらの多くの研究はバドミントン選手のプレースタイルや試合の傾向を分析することを一つの目的としている. しかし選手の弱点やミスの傾向に着目し, これらの内容を分析する研究はまだ行われていない.

1.2 目的

本研究では, 選手の弱点やミスの傾向を分析するために, バドミントンの試合データを用いてショットの成功確率の予測とショットの位置の予測を行うことを目的とする.

各選手における失敗の確率の高いショットやラリーの流れが分かれば, 選手の弱点を見つけ出すことにつながると考えられる. そして選手固有の傾向を分析することができれば, 戦術の決定や, 選手への指導などへ活用されることが期待できる.

1.3 構成

本論文は, 第 1 章から第 6 章で構成される. 第 2 章では関連研究, 第 3 章では本研究の提案手法の紹介を行う. 第 4 章では提案手法を用いた実験とその結果, 第 5 章ではその結果を受けての考察を示し, 第 6 章では本研究のまとめと今後の展望について述べる.

2 関連研究

2.1 試合映像のデータ化

Wang らは, バドミントンの試合データをもとに, ショット単位でのアノテーションがされた公開バドミントンデータセット「ShuttleSet」を作成した [6]. これには, 2018 年から 2021 年にかけて行われた 27 人のトップランキングの男子シングルス及び女子シングルスの選手による 44 試合, 104 セット, 3685 ラリー, 36492 ストロークが含まれている. 試合のアノテーションは, バドミントンの専門家たちによって, S2 ラベリングツール [7] を使用し手動で行われた.

バドミントンの試合の構成を図 1 に示す. バドミントンの試合は, 1 試合 2 セットまたは 3 セットで構成されていて, 各セットは複数のラリー, 各ラリーは複数のショットで構成されている. ShuttleSet データセットには, 各ショットに図 2 のようにいくつかのショット情報が含まれていて, そのアノテーションの形式は, BLSR フォーマット [3] から拡張された以下の 4 つのカテゴリに分類される. ラリーカテゴリーのラリー勝者や勝敗決定の理由の情報は, 各ラリーの最後のショットにのみ含まれる.

- ラリーカテゴリー (両プレイヤのスコア, ラリー勝者, 勝敗決定の理由)
- 時間カテゴリー (ショットの打撃時間, フレーム)
- 空間カテゴリー (ショット時の両プレイヤの座標, シャトルの座標, ショットの高さ)
- 打撃カテゴリー (打撃ステータス, バックハンド, 頭の周り, ショットの種類)

このデータセットの大まかな目標は, 幅広い選手の戦術的なショットやパターンを分析できる機械学習モデルを実装することである. このデータは同著者によるストローク予測やプレイヤの動きの予測タスクなどの研究に用いられている.

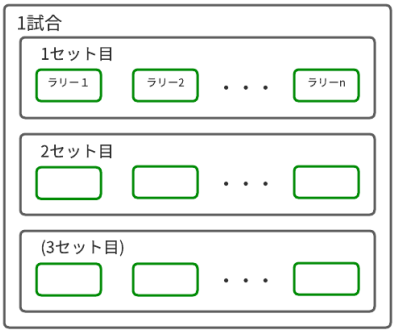


図 1　バドミントンの試合の構成

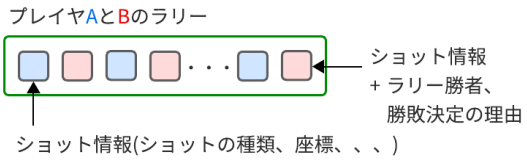


図 2　ShuttleSet の構成

2.2　シーケンス予測

ターンベーススポーツであるバドミントンにおいて, 次のショットの成功確率を予測することはシーケンス予測タスクにあたる. Wang らは, 過去のラリーデータを基に次のショットのタイプと位置を予測するモデル「ShuttleNet」を提案した [8]. このモデルは, 複数回のショットの種類と位置, プレイヤのシーケンスデータが与えられた時, その次のショットの種類と位置を予測するモデルである.

ShuttleNet は, バドミントンのラリーのようなターンベースのシーケンスデータからラリーの進捗をモデル化し, プレイヤのスタイルを抽出するための 2 つのエンコーダデコーダ抽出器と, 各ストロークでのラリーとプレイヤの依存関係を考慮するためのフュージョンネットワークから構成されている. 1 つ目のエンコーダデコーダ抽出器である TRE ではラリーの進捗状況を捉え, 提案されたマルチヘッドタイプエリアアテンション層で, ショットの種類と座標の重要性を別々に特徴づけ, 対応するスコアを最終的な注意スコアとして集約する. 2 つ目のエンコーダデコーダ抽出器である TPE では各プレイヤの情報を分離してそれぞれのプレイヤのコンテキストを生成する. 各エンコーダデコーダ抽出器は, Transformer [9] に基づいた設計となっている. さらに Position-aware Gated Fusion Network(PGFN) でラリーとプレイヤの文脈を融合し各ストロークのコンテキストを学習することで, 各ストロークの情報の重要性を考慮する. 最後に予測層では, ショットの種類予測を学習するためにクロスエントロピー誤差を, ショットの位置予測を学習するために負の対数尤度損失をそれぞれ最小化し, その和をモデル全体の損失とすることで学習を行っている.

このように ShuttleNet は, 他のシーケンス予測が可能なモデルと比較して, バドミントンのストローク予測タスクに特化したモデル構成となっている. このモデルを実世界のバドミントンの試合データに適用すると, ShuttleNet は他の手法を上回る予測性能であった.

3　提 案 手 法

3.1　アプローチの概要

本研究は, バドミントンの複数回のショットから, 次のショットの成功または失敗の確率及び, 次のショットの位置の予測を行うことを目標としている. そこで本研究では, バドミントンの特性に合わせたシーケンス予測を行うために, 複数回のショットから次のショットのタイプと位置を予測するモデルである ShuttleNet を参考にモデル開発を行った. ShuttleNet 開発に使用されたコードは GitHub の CoachAI-Projects [10] から取得した. 本研究の提案モデルは, 埋め込み層と予測層, および損失関数は ShuttleNet で使用されたものを参考に開発し, それを使用した LSTM ベースのモデルである. アプローチの概要を図 3 に示す. エンコーダには, 複数回のショットの情報とプレイヤのシーケンスを入力する. デコーダではエンコーディングコンテキストを取り込んで, 次のショットの成功確率と座標を予測する. エンコーダデコーダは共に LSTM を基にしている. また出力層では, 予測座標と成功失敗を予測する際に算出された成功失敗の確率分布から, 予測した次のショットの位置と成功確率をバドミントンコートにプロットしている.

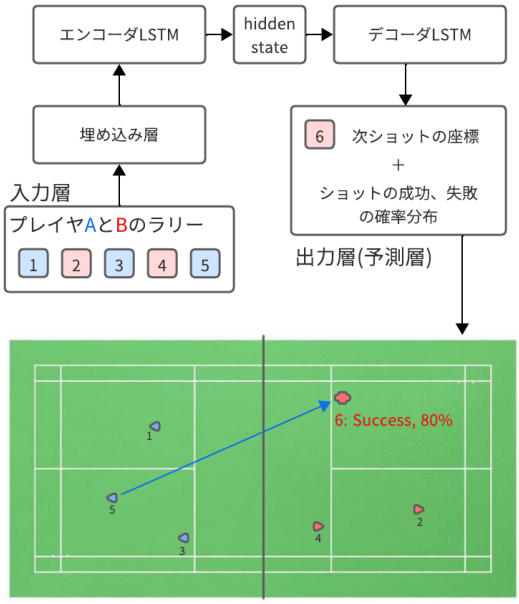


図 3　アプローチの概要

3.2　問題の定式化

バドミントンの試合を $R = (S_r, T_r, P_r)_{r=1}^{|R|}$ と表す. このとき r 番目のラリーについて, $S_r = (\langle s_1, a_1 \rangle, ..., \langle s_{|S_r|}, a_{|S_r|} \rangle)$ はショットのシーケンス, $T_r = (\langle t_1, h_1 \rangle, ..., \langle t_{|S_r|}, h_{|S_r|} \rangle)$ はショットの種類のシーケンス, $P_r = (p_1, ..., p_{|S_r|})$ はプレイヤのシーケンスを示す. ここで各変数は次のような意味を持つ.

- s_i：ショットが成功したか失敗したかを表す変数.
- a_i：シャトルの目的地の xy 座標 $(a_i = (x_i, y_i) \in \mathbb{R}^2)$
- t_i：ショットの種類 (スマッシュ, ロブなど)
- h_i：そのショットを打った手の向きを表す変数.
- p_i：シャトルを打ったプレイヤ

3.3 エンコーダ及びデコーダ

エンコーダには, τ 個のショット $S^e = (s_i, a_i)_{i=1}^{\tau}$ とショットの種類 $T^e = (t_i, h_i)_{i=1}^{\tau}$, プレイヤ $P^e = (p_i)_{i=1}^{\tau}$ のシーケンスを入力する. 入力された各情報に対して埋め込み層を適用することで, 各情報のインデックスに対応する埋め込みベクトル s^E, a^E, t^E, h^E, p^E を生成する. 各埋め込み層では一般的な nn.Embedding 層を適用し, また位置の埋め込み a^E に対しては, 埋め込み層の適用後に ReLU 活性化関数を適用し負の値を除去することで非線形性を導入する. その後さらに各埋め込み情報に対して, プレイヤの埋め込み情報 p^E を加算する.

その後各埋め込み情報を nn.linear による線形変換を適用して統合し, その埋め込み情報 st^E を LSTM に入力することで, 隠れ状態とセル状態を取得する.

$$st^E = \mathrm{Linear}(s^E, a^E, t^E, h^E) \tag{1}$$

$$\mathrm{output}, (\mathrm{hidden}, \mathrm{cell}) = \mathrm{nn.LSTM}(st^E, (\mathrm{hidden}, \mathrm{cell})) \tag{2}$$

デコーダには, エンコーダからの情報と各時間ステップごとの入力, また予測対象のプレイヤ $P^d = (p_i)_{i=\tau+1}^{\tau+n}$ の情報を入力する. エンコーダと同様の埋め込みや線形変換による統合を行い, LSTM の出力に対して追加の線形変換を行う.

$$\mathrm{output} = \mathrm{Linear2}(\mathrm{output}) \tag{3}$$

その結果と予測対象のプレイやの情報をもとに, 予測層で次の n ステップのショット $S^d = (s_i, a_i)_{i=\tau+1}^{\tau+n}$ を自己回帰的に予測する.

3.4 予 測 層

予測層では, エンコーダとデコーダの入出力により, 次のショットが成功するかどうかと, ショットの座標のそれぞれの予測を行う. モデルは損失関数を通じてこれら予測値と真値との誤差を最小化し, 学習を行う.

ショットの成功失敗の予測では, ショットの成功失敗の真値である s_i と予測した成功失敗の確率分布 $\hat{s_i}$ からクロスエントロピー誤差 L_{shot} を導出し, 最小化することで学習する.

$$L_{\mathrm{shot}} = -\sum_{r=1}^{|R|}\sum_{i=\tau+1}^{|S_r|} s_i \log(\hat{s_i}) \tag{4}$$

またエリア座標は, ショットを返す際の不確実性と潜在的な多様性を考慮して, 二変量ガウス分布 (平均値 $\mu = (\mu_x, \mu_y)$, 標準偏差 $\sigma = (\sigma_x, \sigma_y)$, 相関係数 ρ) に従うと仮定する. このときエリアの予測座標 V_{pred} と真の座標 $V_{\mathrm{trgt}} = (x, y)$ の誤差を, 負の対数尤度損失 L_{area} を最小化することで学習する.

$$L_{\mathrm{area}} = -\sum_{r=1}^{|R|}\sum_{i=\tau+1}^{|S_r|} \log\left(P(x_i, y_i|\mu_i, \sigma_i, \rho_i)\right) \tag{5}$$

ここで関数 $\log(P(x, y|\mu, \sigma, \rho))$ は以下のように計算する.

$$\log\left(P(x, y|\mu, \sigma, \rho)\right) = -\log\frac{\exp\left(-\frac{z}{2(1-\rho^2)}\right)}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \tag{6}$$

また z は以下のように計算する.

$$z = \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \tag{7}$$

モデル全体の合計損失 L は, ショット予測の損失関数とエリア予測の損失関数の和で求める.

$$L = L_{\mathrm{shot}} + L_{\mathrm{area}} \tag{8}$$

4 実　　験

4.1 データセットと前処理

4.1.1 データセット

データセットには ShuttleSet [6] を使用した. 各ショットには 10 個のショットの種類 (ネットショット, クリア, プッシュ/ラッシュ, スマッシュ, ディフェンシブショット, ドライブ, ロブ, ドロップ, ショートサービス, ロングサービス) がラベリングされている. 各ショットタイプの詳細を表 1 に示す. またそのショットをフォアハンドで打ったか, バックハンドで打ったかの情報も含まれている. 本研究ではさらに各ショットに対して, 目的変数として成功または失敗のラベルと, 失敗したショットに対して, 本来狙っていた位置を推定しその位置情報も追加した. 各ラベリング方法については次の章で述べる.

また成功失敗のラベリングをした結果, 失敗したショットに対して成功したショットが圧倒的に多く, データに偏りが生じてしまった. そこでシーケンスの長さが長いラリーに対しては, 後半の 8 ラリーのみを残して前半のラリーを削除し, 予測結果の比較を行うショットのラベルが, 成功と失敗の比率が約 4:1 になるように調整した.

4.1.2 成功失敗のラベリング

本研究ではショットの成功失敗を予測することが目的であるため, ショットの成功失敗のラベリングを行う必要がある. データセットには各ラリーの最後のショットに対して, ラリーが終了した理由 (片方のプレイヤが敗北した理由) が, 「lose reason」としてそのショットのデータとして含まれている. ラリーが終了した理由の種類とその詳細を, 「lose reason」のラベルがついているショットを打った選手が敗北したか勝利したかに分けて, それぞれ表 2, 3 に示す. 本研究では, 最後にショットを打った選手が敗北となる表 2 の, 敗北した理由が「ネットを超えない」, 「ネットにかかる」, 「アウト」のショットを失敗とし, それ以外のショットを成功と定義した.

4.1.3 失敗したショットが本来狙っていた位置の推定

失敗したショットはネットやアウトしたショットであるため, その座標はデータセット全体でみると離れた位置である場合が多く, モデルの学習に影響が出ると考えられる. そこで失敗したショットに対して, そのショットが本来狙っていた位置を推定し, 失敗したショットの位置と置き換えることで, 置き換える前の予測結果との比較を行った.

本来狙っていた位置座標 (X, Y) の推定には, 失敗したショットの座標 (x_1, y_1) と一つ前のショットの座標 (x_0, y_0), および全

データから算出した各ショットの種類ごとの平均 y 座標 y_m を用いた．まず失敗したショットが本来狙っていた位置は，失敗したショットとその一つ前のショットの直線上にあると仮定し，2点 $(x_0, y_0), (x_1, y_1)$ を通る直線の式を求めた．次に推定 Y 座標に，そのショットの種類の平均 y_m 座標を設定し，直線の式に代入することで，推定 X 座標を算出した．最後に，算出した X 座標がコート外に出るようなショットであった場合，x 座標がコート内に収まるように閾値に基づく条件処理を行うことで，推定 XY 座標を求めた．

表 1　バドミントンのショットタイプとその説明

ショットタイプ	説明
ネットショット	ネット近くからシャトルをわずかに白帯上を狙って落とすショット．
クリア	コート奥から相手コートの奥まで上げるオーバーヘッドのショット．
プッシュ/ラッシュ	ネット前に高く浮いてきたシャトルを下降軌道で強く突き刺すショット．
スマッシュ	シャトルを下降軌道で強く叩きつける攻撃的なオーバヘッドのショット．
ディフェンシブショット	相手の攻撃的なショットを受けて相手コート内に返す守備的なショット．
ドライブ	ラケットヘッドが体の中央付近から繰り出される，コートに平行した軌道となる強いショット．
ロブ	ネット前から相手コートの奥まで高く上げるアンダーハンドのショット．
ドロップ	コート奥から相手コート手前までシャトルを急な角度で滑らかに打つオーバーヘッドのショット．
ショートサービス	ラリーを始めるショットで，ネットすれすれを狙い相手コートのサービスラインに落とすサーブ．
ロングサービス	ラリーを始めるショットで，高く打ち上げて相手コートのエンドライン付近に落とすサーブ．

表 2　ラリー終了の理由とその説明 (最後にショットを打った選手が敗北する場合)

ラリー終了の理由	説明
ネットを超えない	打ったシャトルがネットを超えずに自分のコートに落ちること．
ネットにかかる	打ったシャトルがネットにかかり自分のコートに落ちること．
アウト	シャトルが相手コート外に落ちる，またはサーブがサービスラインより前に落ちること．

表 3　ラリー終了の理由とその説明 (最後にショットを打った選手が勝利する場合)

ラリー終了の理由	説明
ミスジャッジ	選手がシャトルの落下地点を誤って判断しシャトルを打たないこと．
シャトルを打てない	選手がシャトルの落下地点を正しく判断するが，ラケットがシャトルに届かずシャトルを打てないこと．
フォルト	選手やラケットがネットに触れたり，相手選手のプレイを妨害したりするなどの反則．

4.2 実験結果

4.2.1 定量評価

バッチサイズを 16，トレーニングエポック数を 150，学習率を 0.0001 に設定して全ての実験を行った．また観測されたショット数を $\tau = 2$ に設定した．モデルの学習及び評価は全て Google Colaboratory を用いて行った．全プレイヤの情報をモデルに含めるために，各試合の最初の 70%のラリーをトレーニングデータとして使用し，残りの 30%のラリーをテストに使用した．学習と評価は seed 値を変えそれぞれ 10 回行い，各評価値は 10 回の平均から算出した．

ショットの成功と失敗の予測結果を評価するために，不確実性の測定に広く使用されるクロスエントロピー誤差 (CE) を用いた．CE は確率分布の違いを測定する際に効果的であり，成功と失敗の確率の予測に対する不確実性を考慮することができる．また位置の予測結果を評価するために，平均二乗誤差 (MSE) および平均絶対誤差 (MAE) を用いた．各評価値の結果を表 4 に示す．また最もショットの CE が小さくなった seed 値での，学習のロスの推移は図 4 のようになった．ショットの種類と位置を予測する既存研究のモデル ShuttleNet と比較すると，ショットの位置の誤差 MSE, MAE は共に，提案モデルの方が予測値の誤差が小さいという結果になった．

表 4　ショットの成功失敗と位置予測のロス値

	CE	MSE	MAE
ShuttleNet		1.5761	1.3747
提案モデル	0.3040	**1.1050**	**1.1672**

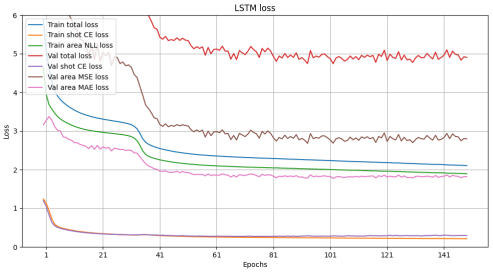


図 4　学習のロスの推移

また提案モデルを用いてショットの成功失敗の予測を行い，評

価指標として失敗ラベルの Recall, Precision, F1 score を用いて実際の値と比較した. 結果を表 5 に示す. ここで「ShuttleNet System」は, ShuttleNet を用いてショットの位置の予測を行い, そのショット位置が相手コート内に入っていれば成功, 入っていなければ失敗と判定するシステムである. 表 5 から, 提案モデルの評価値はすべての評価指標において, ShuttleNet System を大幅に上回る値を示した.

表 5　モデルの性能評価

	Recall	Precision	F1 score
ShuttleNet System	0.184	0.266	0.216
提案モデル	**0.835**	**0.620**	**0.711**

4.2.2　アブレーションスタディ

モデル開発の評価を行うために, 特徴量としてショットの種類と腕の向きを表す変数を埋め込んだ本提案モデルと, それぞれの特徴量を埋め込まないモデルとの比較を行った. 結果は表 6 のようになり, ショットの成功失敗の CE は特徴量 backhand を削減した時のモデルが最も小さく, 位置のロス MSE, MAE は, 特徴量 type と backhand どちらも削減した時のモデルが最も小さいという結果になった.

また各モデルを用いて成功失敗の予測を行い, その性能の比較も行った. 結果は表 7 のようになった. なお表の各列の太字が最も高い値で, 下線を引いている値が 2 番目に高い値である. すべての評価指標において, 特徴量を削減しない提案モデルが最も高い評価値を示した. さらにすべての評価指標において, 腕の向きの特徴量を削減したときの評価値が 2 番目に高いを示した.

表 6　特徴量を削減したロス値比較

	CE	MSE	MAE
w/o (type + backhand)	0.3157	**1.0336**	**1.1277**
w/o type	0.3089	1.0659	1.1474
w/o backhand	**0.2991**	1.0911	1.1575
提案モデル	0.3040	1.1050	1.1672

表 7　特徴量を削減した予測性能比較

	Recall	Precision	F1 score
w/o (type + backhand)	0.769	0.564	0.651
w/o type	0.800	0.579	0.671
w/o backhand	<u>0.806</u>	<u>0.607</u>	<u>0.692</u>
提案モデル	**0.835**	**0.620**	**0.711**

各特徴量を削減したときのモデルのロスの推移を図 5, 6, 7 に示す.

次に, 失敗したショットについて, 失敗したショットの位置のデータをそのまま用いて学習した結果と本来狙っていた位置を差し替えて学習を行った結果を比較した. 比較した CE, MSE, MAE および, 予測結果と真値との比較結果, 位置座標を差し替える前のモデルのロスの推移をそれぞれ表 8, 9, 図 8 に示す. 表 8 から, ショットの成功失敗の CE と位置のロスはどちらも推定位置に差し替えた提案モデルの方が良い結果となった. また表 9 から, Precision は位置データの差し替え前のモデルのほうが高い値となり, Recall と F1 score は失敗したショットの位置を本来狙っていた位置に差し替えて学習した方が高い値を示した.

表 8　失敗ショットの座標を差し替える前とのロス値比較

	CE	MSE	MAE
correct area	0.3113	1.1862	1.2119
提案モデル	**0.3040**	**1.1050**	**1.1672**

表 9　失敗ショットの座標を差し替える前との予測性能比較

	Recall	Precision	F1 score
correct area	0.799	**0.682**	0.705
提案モデル	**0.835**	0.620	**0.711**

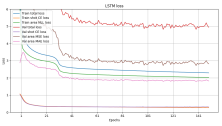


図 5　ロスの推移 (w/o t)

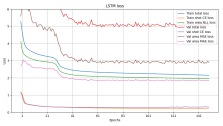


図 6　ロスの推移 (w/o b)

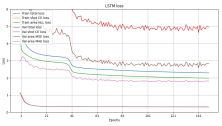


図 7　ロスの推移 (w/o t+b)

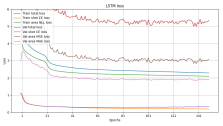


図 8　ロスの推移 (correct area)

4.2.3　ショットの成功確率予測

提案モデルを用いて, データセットのテストラリーに対して, ショットの成功失敗の予測と, 位置の予測を行った. 予測を行った結果のラリーの例を, 図 9, 図 10, 図 11 に示す. 図に示された線はバドミントンコートを表していて, 各点はショットを表している. また一番外側の黒い線がバドミントンのシングルスのコートであり, それより外の領域 (灰色の線も) はアウトとなり失敗ショット扱いとなる. 1 5 のラベルがついている灰色の三角の点が実際のショットであり, 青または赤の三角の点が予測したショットである. 成功したとショット及び成功したと予測したショットに対しては Success, 失敗したショット及び失敗したと予測したショットに対しては Fail とラベルが付けられている. また予測ショットの各数値は, 成功確率または失敗確率を表す. この数値はクロスエントロピー誤差を用いて成功失敗の予測を行う際に算出した各ショットの成功失敗の確率分布から導出している.

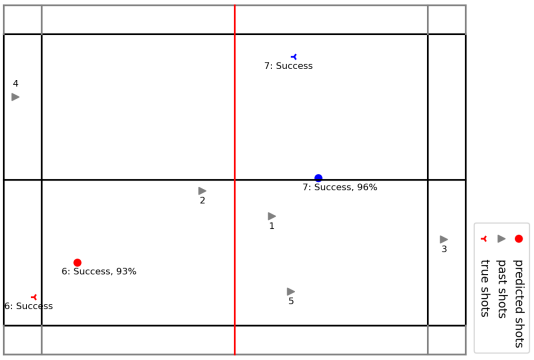


図 9　ショット予測例：ラリー 1

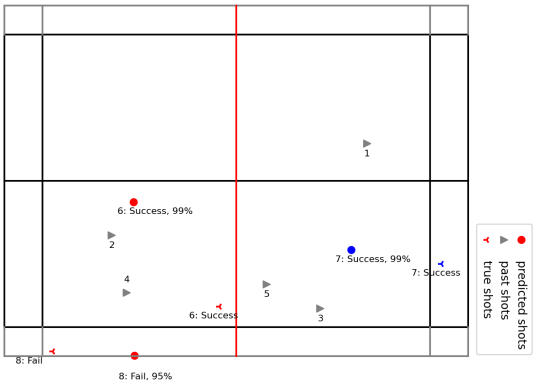


図 10　ショット予測例：ラリー 2

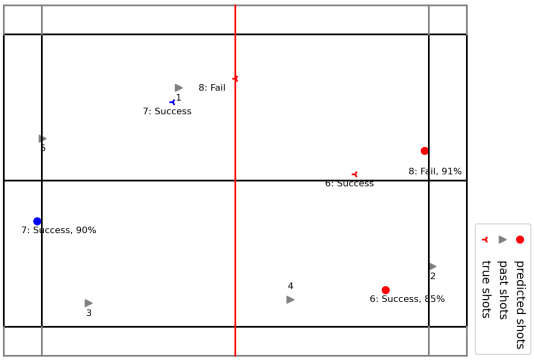


図 11　ショット予測例：ラリー 3

5　分析・考察

5.1　定量評価

バドミントンにおける次のショットの位置の予測は, 提案モデルにより従来の手法に比べて高い精度での予測が可能となった. 従来の手法である ShuttleNet は次のショットの種類と位置を予測するモデルで, 学習の際のモデルの合計損失が, 種類の損失と位置の損失の和になっていた. 本研究の提案モデルでは, ショットの種類ではなく成功確率を予測するモデルであるため, ショットの種類の 10 種分類と比較して位置予測に要する学習量が多いため, 提案モデルの方が位置の予測の精度が高いと考えられる. またショットの成功失敗の予測を行う従来のモデルを見つけることが出来なかったため, 表 5 のように簡単な予測を行うシステムを比較対象として導入した. 本研究では, 次のショットの成功失敗の確率を予測することで, 選手が苦手なショットやコースを分析することを最終的な目的としている. そのため失敗ショットを正しく失敗と予測できるかどうかが最も重要な評価指標である. その結果を示す指標が Recall や F1 score であり, 表 5 から, 提案モデルは ShuttleNet System 世比較して大幅に高い F1 score となっている. このことから, 本研究の提案モデルは失敗するショットを予測するという点において有用であると考察できる.

またアブレーションスタディでは, 特徴量として埋め込んでいるショットの種類「type」と, 打球時の腕の向き「backhand」を削減した時の予測性能の比較を行った. 表 7 から, 特徴量を削減することで全ての評価指標の値が減少していることから, どちらの特徴量もモデルの学習に必要であることが分かった. 特にショットの種類「type」を削減した時の F1 score の値が大幅に減少していることから, ショットの種類の特徴量はより重要であることが分かった. 確かにショットの位置と種類は深く関係していることが多く, 実際失敗したショットの本来狙っていた位置の推定にもショットの種類を使用したため, ショットの種類特徴量を削減したことでモデルの性能が低下することは納得できる結果となった.

さらに失敗したショットに対して本来狙っていた位置を推定し失敗ショットの位置と差し替えた時の比較結果では, 位置データを差し替えた方が MSE, MAE の値が小さくなっている. これは, アウトやネットなどの失敗ショットを位置推定によりコート内に収めたことで, 外れ値が減少した結果であると考察できる. また表 9 から, 提案モデルの方が高い Recall および F1 score となっているため, 失敗ショットの本来狙っていた位置を推定してデータに加えることは, ショットの成功失敗の予測において有用であると分かった.

5.2　ショットの成功確率予測

本研究では, モデルの性能評価と, 提案モデルを用いて特定のラリーに対して次のショットの成功確率を予測して, その結果をバドミントンのコートにプロットすることも行った. 図 9 では, 7 ショットで構成された実際のラリーの, 6 ショット目と 7 ショット目の予測を行っている. どちらのショットも正しく成功と予測できていて, 6 ショット目は位置の予測も他ショットと比較して近い位置の予測となっている. 図 10 では, 別のラリーの 8 ショット目の予測も行っている. このラリーでは 8 ショット目がアウトになり, 失敗ショット扱いとなっていて, 8 ショット目の予測ショットもアウトしていて正しく失敗と予測できている. 図 11 では, 図 10 と同様に 8 ショットのラリーの予測を行っていて, 8 ショット目の失敗を正しく予測できていることが分かる. しかし予測位置はコート内に収まっていて, 図だけで見ると予測ショットは失敗したとは言い難い.

このようにショットの成功失敗の予測と位置予測がどちらも比較的良く予測できているものと, 成功失敗の予測が間違っているが予測位置は近いもの, 成功失敗の予測は合っているが予測位置が大きく離れているものなど様々なラリーのショット予測の結果が得られた. 本研究の目的である選手の失敗の傾向を

分析する上では, 成功失敗の予測とその確率分布が正しく得られていればいいが, さらにショットの位置も予測して試合展開を分析するためには, 今後の研究で位置と成功失敗の矛盾を解決する必要があるだろう.

6 おわりに

6.1 本研究のまとめ

本研究では, バドミントンの試合データを基に, ショットの成功確率の予測と, ショットの位置の予測を行う LSTM ベースのモデルを開発した. 結果は, ショットの位置予測において提案モデルが従来モデルの性能を上回り, またショットの成功失敗の予測においては, 単純な予測システムの性能と比較して大幅に上回る性能を示した. またショットの種類, 打球時の腕の向きの特徴量及び, 失敗したショットに対して本来狙っていた位置を推定しデータを差し替えることは, モデルの精度向上に大きく貢献していることが分かった. さらに本提案モデルを用いて, ショットの予測をバドミントンのコート上に示し, 予測の可視化も行った.

6.2 今後の展望

今後の展望として, 提案モデルを実際のバドミントンの世界で活用するための工夫を 2 点挙げる.

1 点目は, モデルを用いてショットの予測をする際に, モデルに与えるラリーを実際のラリーではなく新しく作成した仮想ラリーにすることで, ラリーのシミュレーションを行うことである. バドミントンの選手やコーチがプレイヤの情報やショットの情報 (ショットの位置や種類など) を簡単にモデルに渡せるように改良することで, 簡単に各選手におけるラリーのシミュレーションを行える. この改良により, どのようなラリー展開にすれば自分または相手選手が失敗しやすいかなどを簡単に分析できるようになると考えられる.

2 点目は, 全体のショットの予測結果を基に, 各選手における失敗しやすい傾向のあるラリーを分析し出力することである. これにより, 各選手は自分の過去の試合データを基に, どのようなラリーやショットが苦手であるかを瞬時に分析できるようになると考えられる.

また上記のような実際のバドミントンの世界で活用できるようなモデルのレベルに到達するために, 提案手法で示した特徴量以外にも有用な特徴量を見つけたり, 別のモデルベースによる手法を検討したりして, モデルの性能をさらに向上させる必要があるだろう.

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです.

文　献

[1] Wei-Ta Chu, and Samuel Situmeang, “Badminton Video Analysis based on Spatiotemporal and Stroke Features.” Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp.448-451, 2017.

[2] Kokum Weeratunga, Anuja Dharmaratne, Khoo Boon How, “Application of Computer Vision and Vector Space Model for Tactical Movement Classification in Badminton.” 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.4219-4227, 2017.

[3] Wei-Yao Wang, Teng-Fong Chan, Hui-Kuo Yang, Chih-Chuan Wang, Yao-Chung Fan, and Wen-Chih Peng, “Exploring the Long Short-Term Dependencies to Infer Shot Influence in Badminton Matches.” 2021 IEEE International Conference on Data Mining (ICDM), pp.1397-1402, 2021.

[4] Karweng Ban, John See, Junaidi Abdullah, Yuenpeng Loh, “BadmintonDB: A Badminton Dataset for Player-specific Match Analysis and Prediction.” MMSports ’22: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, pp47-54, 2022.

[5] Yong En Tan, John See, Junaidi Abdullah, Lai Kuan Wang, Kar Weng Ban, “A deep learning based framework for badminton rally outcome prediction.” 2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2022.

[6] Wei-Yao Wang, Yung-Chang Huang, Tsi-Ui Ik, Wen-Chih Peng, “ShuttleSet: A Human-Annotated Stroke-Level Singles Dataset for Badminton Tactical Analysis.” arXiv.2306.04948, 2023.

[7] Yu-Hsien Huang, Yung-Chang Huang, Hao Syuan Lee, Tsì-Uí Ik, and Chih-Chuan Wang, “S2-Labeling: Shot-By-Shot Microscopic Badminton Singles Tactical Dataset.” 2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS), IEEE, pp.1-6, 2022.

[8] Wei-Yao Wang, Hong-Han Shuai, Kai-Shiang Chang, and Wen-Chih Peng, “ShuttleNet: Position-Aware Fusion of Rally Progress and Player Styles for Stroke Forecasting in Badminton.” Proceedings of the 36th AAAI Conference on Artifical Intelligence, pp.4219-4227, 2022.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, llia Gomez, Łukasz Kaiser, Polosukhin,, “Attention is all you need.” In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 5998–6008. 2017.

[10] Wei-Yao Wang. [onsite] https://github.com/wywyWang/CoachAI-Projects

[11] J SPORTS. 「バドミントン：基本的なショットの種類」. https://www.jsports.co.jp/badminton/about/shot/ (Accessed on 01/11/2024).

ボール奪取と攻撃の防御に基づくアクションによるサッカー選手の攻守評価

前島　涼弥† 　清　雄一† 　田原　康之† 　大須賀昭彦†

† 電気通信大学　〒 182–8585 東京都調布市調布ケ丘一丁目 5 番地 1
E-mail: †maejima.ryoya@ohsuga.lab.uec.ac.jp, ††{seiuny,tahara,ohsuga}@uec.ac.jp

あらまし　サッカーにおいて, パスやシュートなどのアクションが試合結果にどのような影響を与えるかを分析することは重要な課題である. 近年では, 機械学習を用いて選手が行ったアクションの価値を定量的に評価する研究が増えており, その中には算出されたアクションの価値を基に選手の試合への貢献度を評価する研究もある. しかし, アクションデータを用いた選手の評価には 2 つの課題がある. 1 つ目は, ゴールやアシスト以外のアクションの影響が過小評価されていること, 2 つ目は, 守備的な選手を正確に評価することが難しいことである. 本研究では, サッカーの選手評価における課題を克服し, 選手を攻守両面から総合的に評価できる新しいモデルを提案する. 結果として, 既存モデルの守備面の評価指標を改善することで, 既存モデルよりも高い精度で選手を評価できることを示した. また, 提案モデルによる攻守の評価値を基に行った選手のプレースタイル分類では, 選手の特徴を客観的に分析することを可能にした.

キーワード　スポーツデータ活用, スポーツ分析, サッカー

1　はじめに

1.1　研究背景

現代のサッカー界では, データを積極的に活用して試合分析が行われている. 特に, パスやシュートなどの試合中に発生するアクションが試合結果にどのような影響を与えるか分析することは非常に重要な課題とされている. 近年では, 選手が行ったアクションから得られたアクションデータを活用し, 機械学習によって試合を分析する研究が増えている. 例えば, Beal らは, アクションデータから試合結果を予測し, 算出した確率分布を基にどのような戦術が試合の勝率を高めるかを提案するモデルを開発し, 高い精度を達成した [1]. さらに, この研究を発展させて, 1 試合ではなく 1 シーズンの結果を予測し, 最適な戦術を提案するモデルを構築し, シミュレーション内で高い精度を示した [2]. その他にも, Transformer を用いた試合イベントの予測や, 試合状況とボールリカバリーの関係の考察, 逆強化学習を用いた戦術分析など, アクションデータの活用法は多岐にわたる [3] [4] [5].

アクションデータは, 試合結果の予測だけでなく, 選手の評価にも活用されている. 例えば, Decroos らは, アクションデータを SPADL と呼ばれる言語に変換し, SPADL といくつかの特徴量を組み合わせて, 選手が試合にどれだけ貢献したかを計算する選手評価モデルを提案した [6]. さらに, この研究を発展させて, 選手のプレースタイル分類を行い, 選手の特徴を明らかにした [7] [8]. また, Toda らは, アクションデータの中でも特に守備的なアクションに焦点を当て, これらの守備アクションがチームの守備にどれだけ寄与しているかを定量化することでチームの守備力を評価した [9].

ただし, アクションデータを用いた選手の評価には 2 つの課題が存在している. 1 つ目は, ゴールやアシスト以外のアクションが過小評価されていることである. サッカーは他のスポーツと異なり, 得点が発生する頻度が非常に低いため, ゴールやアシストが試合結果に与える影響が非常に大きくなる. このため, 得点に関与した選手が際立ってしまい, 得点に直接関与していなくても勝利に貢献した選手を適切に評価することが難しくなる. 2 つ目は, 守備的な選手を正確に評価することが難しいことである. 攻撃的な選手の場合, ゴール数やアシスト数, ボールのタッチ数など, 基本的な数値からその選手の試合への貢献度をある程度予測することが可能である. しかし, 守備的な選手の場合, このような基本的な数値だけではその選手の試合への貢献度を正確に評価することが困難である. また, チーム全体の守備に関しては, 単に失点の少ないチームの守備が優れていると評価できるが, 選手個々の守備への貢献度を具体的に定量化することは非常に難しい.

1.2　研究目的

本研究の目的は, アクションデータを活用して, サッカーの選手評価における課題を克服し, 選手を攻守両面から総合的に評価できる新しいモデルを提案することである. ここで, 本研究におけるアクションデータとは, フィールドプレイヤーが行うアクションを指し, ゴールキーパーが行うアクションについては評価の対象から除外する.

このようなモデルは, 若手の有望な選手を見つけ出したり, チームに必要なプレースタイルに適した選手がいるかどうかを確認する際に有益である. 従来, これらの活動は主に人がスカウティングすることで実施されてきた. しかし, 人が試合を観戦する場合, シュートやドリブルなどの目立つプレーに注目が集まりがちである. そのため, 機械を使用して客観的に評価を行うことで, 注目度に左右されずに選手を公平かつ総合的に評価することが可能になる.

表 1　アクションの種類の一部

アクション	用語の説明
Pass	プレーヤーがチームメイトに向かってボールを蹴ること
Cross	プレーヤーがボックス内にいるチームメイトに向かってボールを蹴ること
Tackle	プレーヤーが相手からボールを奪おうとすること
Interception	相手のボールが目標のプレーヤーに渡るのを阻止すること
Shot	プレーヤーがゴールに向かってボールを蹴ること
Dribble	プレーヤーが一定の距離ボールを運ぶこと

1.3　サッカーにおけるアクション

本研究で用いる全 21 種類のアクションうち, 一部を表 1 に示す. なお, 用語の説明は Soccer Coaching Pro [10] の用語集を参考にした.

1.4　論文の構成

本論文は, 第 1 章から第 6 章で構成されている. 第 2 章では関連研究に焦点を当て, 第 3 章では本研究で提案された手法を紹介する. 第 4 章では提案手法を用いた実験とその結果について述べ, 第 5 章では得られた結果に基づく考察を示す. 最後に, 第 6 章では本研究のまとめと今後の展望について述べる.

2　関 連 研 究

2.1　VAEP

アクションデータから選手を評価するモデルの代表的な例として, Decroos らによる VAEP(Valuing Actions by Estimating Probabilities) が挙げられる [6]. この研究では, まずアクションデータを SPADL と呼ばれる言語に変換し, その後に SPADL といくつかの特徴量を組み合わせることで, 選手の試合への貢献度を計算する.

2.1.1　SPADL

試合のアクションを表記したデータはイベントストリームデータと呼ばれ, これらのデータは複数の企業によって提供されている. ただし, それぞれの企業が使用するイベントの用語や定義は異なる. これらの多様なデータを人間が理解しやすく, かつシンプルで完全な言語に変換するための手法が SPADL(Soccer Player Action Description Language) である. SPADL では, 試合は m 個のアクションのシーケンス $[a_1, a_2, \cdots, a_m]$ で表現され, 各アクションは以下の 9 つの属性を持つ.

- StartTime：アクションの開始時間
- EndTime：アクションの終了時間
- StartLoc：アクションが開始した場所 (x, y)
- EndLoc：アクションが終了した場所 (x, y)
- Player：アクションを行った選手
- Team：選手の所属するチーム
- ActionType：アクションの種類
- BodyPart：アクションに使用した体の部位
- Result：アクションの結果

2.1.2　VAEP の算出

SPADL に変換されたアクションデータを基に, 選手が行ったアクションの価値を確率の推定を通じて計算するフレームワークが VAEP である. この研究では, アクションの価値を得点確率 P_{scores} と失点確率 $P_{concedes}$ の推定によって求めており, その計算式は以下の通りである.

$$V(a_i, x) = \Delta P_{scores}(a_i, x) + (-\Delta P_{concedes}(a_i, x)) \quad (1)$$

ここで, a_i はアクションを, x はチームを表している. VAEP は攻撃評価と守備評価に分かれており, 攻撃評価は得点確率 P_{scores} の変化量によって求められ, 守備評価は失点確率 $P_{concedes}$ の変化量によって求められる. 言い換えれば, 1 つのアクションが発生したことで, 得点する確率がどれだけ増加するか, 失点する確率がどれだけ減少するかによってそのアクションの価値を評価していると言える.

得点確率と失点確率の予測は, 二値確率的分類問題として行われる. 入力はゲーム状態 $S_i = a_1, a_2, \cdots, a_i$ であり, 出力は得点確率 P_{scores} と失点確率 $P_{concedes}$ である, 確率の学習のための準備として, 各ゲーム状態にラベル付けをし, 特徴量を構築する. 具体的には, 以下の方法でラベル付けが行われる.

得点確率 P_{scores} の学習
k 回の一連のアクションの中で得点すればラベル 1 を,
得点しなければラベル 0 を付与する.
失点確率 $P_{concedes}$ の学習
k 回の一連のアクションの中で失点すればラベル 1 を,
失点しなければラベル 0 を付与する.

ここで, k の値は予備実験の結果, $k = 10$ としている. 特徴量は以下の要素を中心に構築される.

- 2.1.1 項で説明した SPADL の属性
- 連続する 2 つのアクションの間の移動距離と経過時間
- アクション中のボールの持ち替えの有無
- スコアによる影響

このモデルは攻撃評価において高い精度を達成したが, 守備評価の精度に課題が生じた. この問題の原因は次のように考えられる. 守備評価は失点確率に基づいており, 大きなピンチなど失点に近いアクションの影響が大きくなる傾向がある. その結果, ピンチ自体が多いチームの選手が高い評価を受け, ピンチが少ないチームの選手が正当な評価を受けにくくなっていたためである.

本研究では, 上記のような問題に対処するために, 失点確率だ

けでなく, ボール奪取や有効な攻撃を防ぐことなど他のイベントが発生する確率も予測し, これらの値を基に守備を評価する. そのために, 以下の評価指標を導入する.

2.2 VDEP

Toda らは, チームの守備を評価するためのフレームワークとして, VDEP(Valuing Defense by Estimating Probabilities) を提案した [9]. この研究では, チームの守備をボールを奪う確率 $P_{recoveries}$ と有効な攻撃をされる確率 $P_{attacked}$ の推定によって求めており, その計算式は以下の通りである.

$$V_{vdep}(S_i) = \Delta P_{recoveries}(S_i) - C * \Delta P_{attacked}(S_i) \quad (2)$$

ここで, S_i はゲーム状態を表し, C は調整用のパラメータである. VDEP は, ボールを奪う確率 $P_{recoveries}$ の変化量から, 有効な攻撃をされる確率 $P_{attacked}$ の変化量を引くことで求められる. 言い換えれば, ゲーム状態が変化する際に, ボールを奪う確率がどれだけ増加するか, 有効な攻撃をされる確率がどれだけ減少するかによって, その間のチームの守備を評価していると言える. このモデルは, チームの守備の評価において高い精度を記録した.

3 提案手法

本研究では, VAEP の守備評価における問題点を改善し, 選手を攻守両面から総合的に評価できる新しいモデルを提案する.

モデルの概要図を図 1 に示す.

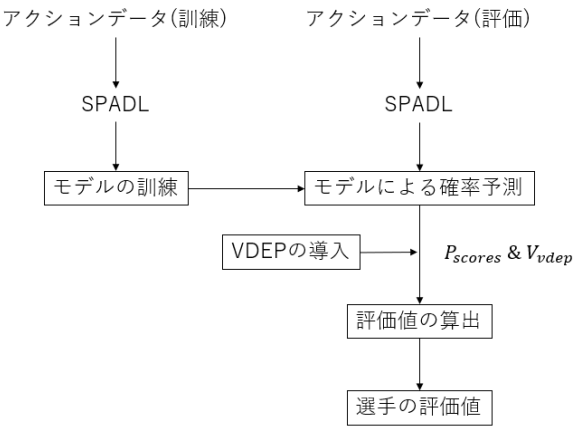


図 1　モデルの概要図

まず, アクションデータを SPADL に変換する. この変換は, 2.1.1 項で詳細に説明している. 次に, 変換されたアクションデータを使用して, モデルを訓練し, 確率予測を行うことで, 評価値を算出する. VAEP では, 攻撃評価は得点確率の変化量, 守備評価は失点確率の変化量によって求めていたが, 守備面の指標に VDEP を導入することで守備評価の改善が図られる. 最後に, 選手ごとのアクションの価値を合計し, その合計値によって選手の総合的な評価値が算出される. このようなモデルを構築することで, 選手の攻守両面での貢献度が総合的に評価される.

3.1 VDEP の導入

本研究では, チームの評価指標である VDEP を選手個人の評価に変更するために, 式 (2) を修正する. 修正した式は以下の通りである.

$$V_{vdep}(a_i, x) = \Delta P_{recoveries}(a_i, x) - C * \Delta P_{attacked}(a_i, x) \quad (3)$$

ここで, C は式 (2) にも登場した調整用のパラメータである. 本研究では再調整の結果, $C = 0.1$ としている. 既存の VDEP では, ゲーム状態に対する確率から評価を導き出していたが, 提案モデルでは VAEP と同様に 1 つのアクションに対する確率を求めている. この変更により, チーム全体の守備ではなく, アクションを行った個々の選手の守備を評価できるようになる.

ボールを奪う確率と有効な攻撃をされる確率の予測は, VAEP と同様に二値確率的分類問題として行われる. 入力はゲーム状態 $S_i = [a_1, a_2, \cdots, a_i]$ であり, 出力はボールを奪う確率 $P_{recoveries}$ と有効な攻撃をされる確率 $P_{attacked}$ である. 確率の学習のための準備として, 各ゲーム状態にラベル付けをし, 特徴量を構築する. 具体的には、以下の方法でラベル付けが行われる.

ボールを奪う確率 $P_{recoveries}$ の学習
k 回の一連のアクションの中で, 「Interception」か「Tackle」が成功すればラベル 1 を付与し, それ以外ならばラベル 0 を付与する.
有効な攻撃をされる確率 $P_{attacked}$ の学習
k 回の一連のアクションの中で, 対戦相手の「Cross」か「Shot」が成功すればラベル 1 を付与し, それ以外ならばラベル 0 を付与する.

ここで, k の値は予備実験の結果, $k = 10$ としている. また, 特徴量は VAEP と同じものを使用している.

3.2 VAEP の改良

式 (3) で求めた VDEP を用いて, VAEP に修正を加える. 修正を加えた式は以下の通りである.

$$V(a_i, x) = \Delta P_{scores}(a_i, x) + \gamma * V_{vdep}(a_i, x) \quad (4)$$

ここで, γ は補正値である. VAEP では守備評価は失点確率の変化量によって求めていたが, 本研究では失点確率の代わりに式 (3) で求めた VDEP の値を使用している. この変更により, 失点するイベントだけでなく, ボールを奪うことや有効な攻撃をされることといったイベントの情報も考慮され, 守備評価の精度が向上する. 本研究では, この修正を加えた式を用いて選手のアクションを評価する. また, 補正値 γ を導入した理由は, サッカーにおいて得点を入れるイベントの回数よりも, ボールを奪ったり有効な攻撃をされたりするイベントの回数の方が遥かに多いため, γ を導入しなければ VDEP の値の方が得点確率よりも大きくなってしまうからである. なお, γ の値は予備実験の結果, $\gamma = 0.1$ としている.

4 実　　験

4.1 データセット

データセットには, StatsBomb [11] が提供している「Fifa World Cup 2022」の全 64 試合のアクションデータを用いた. アクションデータの中には, 試合中に行われた全アクションの位置とその種類（表 1 参照）, アクションを行った選手の名前などが含まれている. なお, このデータセットに登録されている選手の総数は 832 人である. さらに, 提案モデルを評価するうえで必要となる選手の情報を 2 つ追加した. 1 つは, 選手のポジション, もう 1 つは選手の市場価値である. なお, これらの情報は Transfermarkt [12] から引用した. 選手のポジションの詳細を表 2 に示す.

表 2　ポジションの詳細

ポジション	用語の説明
CF	センターフォワード
SS	セカンドトップ
RW	右ウイング
LW	左ウイング
AM	攻撃的ミッドフィルダー
RM	右ミッドフィルダー
LM	左ミッドフィルダー
CM	セントラルミッドフィルダー
DM	守備的ミッドフィルダー
RB	右サイドバック
LB	左サイドバック
CB	センターバック
GK	ゴールキーパー

なお, 1.2 項にあるように, 本研究ではゴールキーパーは評価の対象外であるため, 選手リストから除外した. また, セカンドトップ, 右ミッドフィルダー, 左ミッドフィルダーに関しては, 登録されている選手数が少ないため, それぞれ役割の近いセンターフォワード, 右ウイング, 左ウイングと合同とすることにした.

4.2 実験 1: アクションの価値の直感的理解

提案モデルの評価値がどのように算出されているかを理解するために, 「Fifa World Cup 2022」の決勝, 「アルゼンチン対フランス」の一場面を例に挙げて説明する. 図 2 は守備的なアクションが行われた場面であり, 上の表にはボールを奪う確率 $P_{recoveries}$, 有効な攻撃をされる確率 $P_{attacked}$, 守備の評価値 $defensive_value$ が示されている.

(1) では, Lionel Messi が中央へのパスをミスした. このアクションの評価値は 0.033 となる. (2) では, Raphaël Varane が右にドリブルをした. この評価値は 0.039 となる. (3) では, Raphaël Varane がキーパーの Hugo Lloris にバックパスを出した. このアクションの評価値は-0.012 となる. (4) では, Hugo Lloris が前にパスを出した. この評価値は 0.067 となる. (5) では, Hugo Lloris からの縦パスを Nicolás Tagliafico がインターセプトした. この評価値は 0.991 となる.

	time	type_name	player_name	team_name	recoveries	attacked	defensive_value
1	48m19s	pass	Lionel Messi	Argentina	0.068	0.007	0.033
2	48m20s	dribble	Raphaël Varane	France	0.040	0.011	0.039
3	48m22s	pass	Raphaël Varane	France	0.027	0.009	-0.012
4	48m23s	pass	Hugo Lloris	France	0.095	0.021	0.067
5	48m26s	interception	Nicolás Tagliafico	Argentina	0.992	0.006	0.991

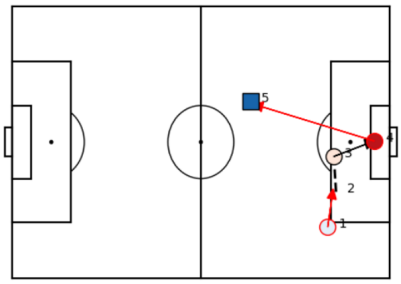


図 2　各アクションの守備評価

このように, 各アクションにはボールを奪う確率 $P_{recoveries}$, 有効な攻撃をされる確率 $P_{recoveries}$, 守備評価の評価値 $defensive_value$ が付与され, これらの値によって選手のアクションが客観的に評価される. 例えば, Nicolás Tagliafico の「Interception」が高い評価値を得ていることから, この守備アクションは非常に成功していたことがわかる.

4.3 実験 2: 市場価値との比較

提案モデルの性能を検証するために, 提案モデルによって形成された選手の評価値と市場価値を比較した. 結果を図 3 に示す.

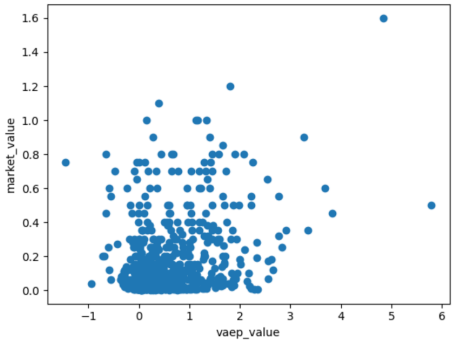


図 3　選手の評価値と市場価値 (Million) の比較

提案モデルによる選手の評価値と市場価値の相関係数を計算した. 相関係数には, ピアソンの相関係数 p とスピアマンの順位相関係数 R_s の 2 つを用いた. 結果, ピアソンの相関係数は $p = 0.314$, スピアマンの相関係数は $R_s = 0.246$ となった.

さらに, 既存の VAEP と提案モデルの性能を比較するために, 同条件下で 2 つのモデルを実行し, 各モデルにおける評価値と市場価値の相関係数をポジションごとに調べた. 結果を表 3 に示す.

表の上段の全選手による相関係数を見ると, 提案モデルの方が相関係数が大きく, 市場価値との相関が高いことがわかる. また, ポジションごとの比較においても, いくつかの例外を除いて, ほとんどが提案モデルの方が市場価値との相関が高いこと

表 3　VAEP と提案モデルの性能比較

	VAEP		提案モデル	
ポジション	p	R_s	p	R_s
全選手	0.251	0.167	**0.314**	**0.246**
CF	-0.002	-0.034	**0.100**	**0.118**
RW	0.262	0.178	**0.346**	**0.240**
LW	0.510	0.187	**0.571**	**0.257**
AM	0.364	0.205	**0.365**	**0.359**
CM	**0.480**	0.350	0.417	**0.352**
DM	0.306	0.179	**0.437**	**0.335**
RB	0.089	0.192	**0.187**	**0.221**
LB	0.227	0.200	**0.290**	**0.304**
CB	0.260	0.268	**0.362**	**0.361**

がわかる．よって，既存の VAEP よりも，提案モデルの方がより高い精度で選手を評価できていることが示されている．

4.4　実験 3: 攻守の評価による選手のプレースタイル分類

提案モデルによって形成された選手の評価値を基に，選手のプレースタイル分類を行った．なお，実験 1,2 では，選手の評価値は式 (4) のように攻守の評価値の和によって算出していたが，本実験では攻撃と守備それぞれの評価値を算出し，それらを x 軸，y 軸に置いたグラフを作成し，クラスター分析を行うことで選手のプレースタイルを分類した．また，プレースタイル分類は全選手ではなくポジションごとに行い，各ポジションのクラスター数は SSE（Sum of Squared Errors）とシルエットスコアを考慮して決定した．プレースタイル分類の結果は図 4 ～ 12 に，各番号がどの選手を表しているかは表 4 ～ 12 に示されている．なお，本論文では，ページ数の都合により，全選手ではなく一部の選手の名前を載せている．

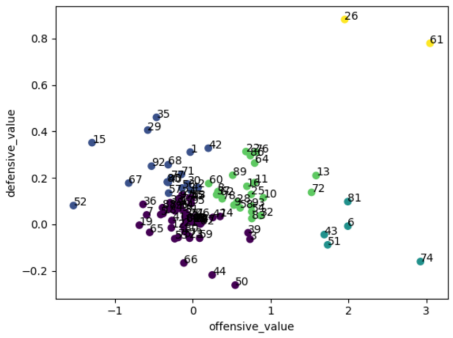


図 4　攻守評価によるプレースタイル分類 (CF)

表 4　番号と選手名の対応 (CF)

番号	選手名
13	Olivier Giroud
26	Antoine Griezmann
29	Thomas Müller
50	Harry Kane
74	Julián Álvarez

センターフォワードの選手は 5 つのクラスターで分類されている．特に，Antoine Griezmann が含まれているクラスターの選手はグラフ右上の離れた位置にある．よって，これらの選手は攻守においてとびぬけて高い評価を得ていることがわかる．

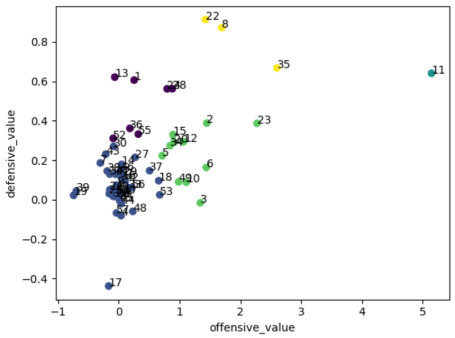


図 5　攻守評価によるプレースタイル分類 (RW)

表 5　番号と選手名の対応 (RW)

番号	選手名
2	Ángel Di María
11	Lionel Messi
17	Iñaki Williams
22	Denzel Dumfries
36	Junya Ito

右ウイングの選手は 5 つのクラスターで分類されている．特に，Lionel Messi はグラフ右上の離れた位置にあり，単体でクラスターを形成している．よって，Lionel Messi は攻守においてとびぬけて高い評価を得ていることがわかる．

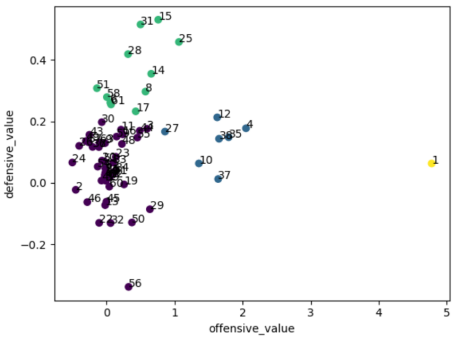


図 6　攻守評価によるプレースタイル分類 (LW)

表 6　番号と選手名の対応 (LW)

番号	選手名
1	Kylian Mbappé
4	Marcus Rashford
15	Mathew Leckie
56	Akram Afif

左ウイングの選手は 4 つのクラスターで分類されている．特に，Kylian Mbappé はグラフ右の離れた位置にあり，単体でクラスターを形成している．よって，Kylian Mbappé は攻撃においてとびぬけて高い評価を得ていることがわかる．

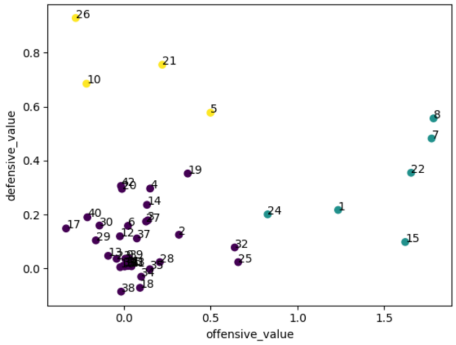


図 7　攻守評価によるプレースタイル分類 (AM)

表 7　番号と選手名の対応 (AM)

番号	選手名
8	Giorgian De Arrascaeta
17	Daichi Kamada
26	Selim Amallah

攻撃的ミッドフィルダーの選手は 3 つのクラスターで分類されている．特に, Giorgian De Arrascaeta が含まれているクラスターの選手はグラフ右上の位置にある．よって, これらの選手は攻守において高い評価を得ていることがわかる．

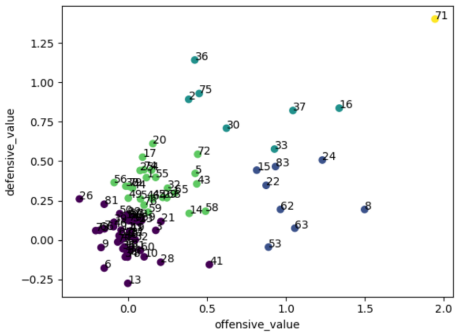


図 8　攻守評価によるプレースタイル分類 (CM)

表 8　番号と選手名の対応 (CM)

番号	選手名
13	Matías Vecino
16	Luka Modrić
20	Héctor Herrera
63	Jude Bellingham
71	Enzo Fernandez

セントラルミッドフィルダーの選手は 5 つのクラスターで分類されている．特に, Enzo Fernandez はグラフ右上の離れた位置にあり, 単体でクラスターを形成している．よって, Enzo Fernandez は攻守においてとびぬけて高い評価を得ていることがわかる．

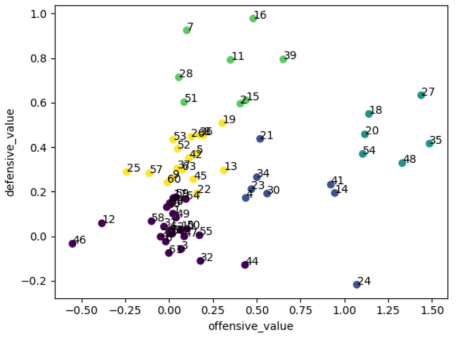


図 9　攻守評価によるプレースタイル分類 (DM)

表 9　番号と選手名の対応 (DM)

番号	選手名
13	William Carvalho
16	Marcelo Brozović
35	Aurélien Tchouaméni
41	Leandro Paredes
44	Wataru Endo

守備的ミッドフィルダーの選手は 5 つのクラスターで分類されている．特に, Aurélien Tchouaméni が含まれているクラスターの選手はグラフ右上の位置にある．よって, これらの選手は攻守において高い評価を得ていることがわかる．

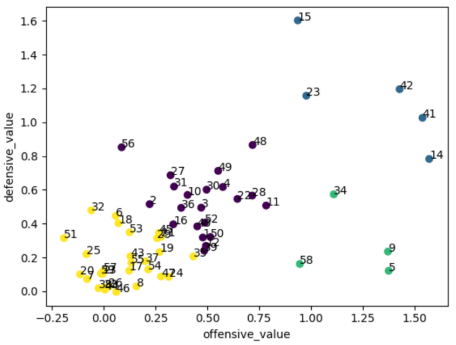


図 10　攻守評価によるプレースタイル分類 (RB)

表 10　番号と選手名の対応 (RB)

番号	選手名
9	César Azpilicueta
15	Achraf Hakimi
48	Angelo Preciado
51	Josip Stanišić

右サイドバックの選手は 4 つのクラスターで分類されている．特に, Achraf Hakimi が含まれているクラスターの選手はグラフ右上の位置にある．よって, これらの選手は攻守において高い評価を得ていることがわかる．

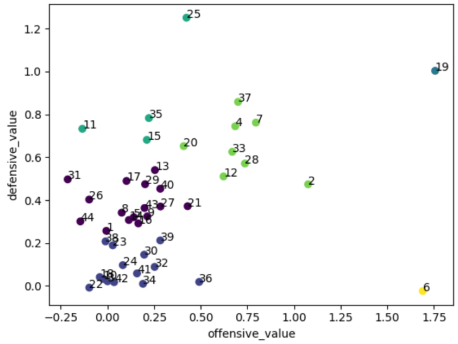


図 11　攻守評価によるプレースタイル分類 (LB)

表 11　番号と選手名の対応 (LB)

番号	選手名
2	Luke Shaw
6	Raphaël Guerreiro
19	Theo Hernández
25	David Raum
31	Marcos Acuña
36	Yahia Attiyat allah

左サイドバックの選手は 6 つのクラスターで分類されている．特に, Theo Hernández はグラフ右上の離れた位置にあり, 単体でクラスターを形成している．よって, Theo Hernández は攻守においてとびぬけて高い評価を得ていることがわかる．

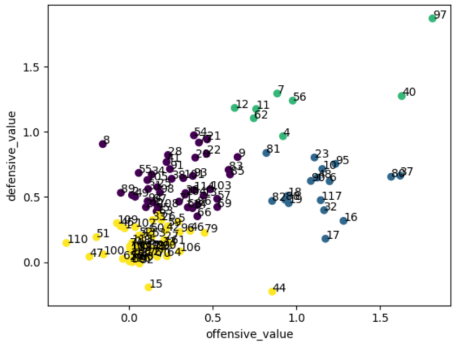


図 12　攻守評価によるプレースタイル分類 (CB)

表 12　番号と選手名の対応 (CB)

番号	選手名
8	Maya Yoshida
44	Niklas Süle
80	Pepe
97	Joško Gvardiol

センターバックの選手は 4 つのクラスターで分類されている．特に, Joško Gvardiol が含まれているクラスターの選手はグラフ右上の位置にある．よって, これらの選手は攻守において高い評価を得ていることがわかる．

5　考　　察

5.1　実験 1: アクションの価値の直感的理解

実験 1 で扱った場面は, データセットの中でも守備的な評価が非常に高いアクションである．既存の VAEP で実験を行った場合, 失点確率を基に評価を行うために, 評価の高いアクションの多くは相手の決定的なチャンスを防ぐようなアクションであった．しかし, この評価方法では, 相手の決定的なチャンス自体が多いチームの選手の評価が高くなってしまうので, この現象を防ぐために本研究の評価方法を用いた．提案モデルで実験を行った場合, 実験 1 にあるように, 評価の高いアクションの多くは相手からボールを奪うアクション, すなわち「Interception」や「Tackle」が成功したアクションであった．したがって, 基本的には「Interception」や「Tackle」をより多く成功させた選手の評価が高くなる．しかし, 有効な攻撃をされる確率 $P_{attacked}$ の影響があることによって, 同じボールを奪うアクションでも, 相手の脅威を取り除いたり, 相手の陣地深くまで侵入する起点となったアクションにはより高い評価がつくようになっている．

5.2　実験 2: 市場価値との比較

全選手に対する提案モデルの評価と市場価値には, 正の相関がみられることが分かった．しかし, その相関係数は高い値とはならなかった．その理由は主に 3 つ考えられる．1 つは, データセットには記載されていない不確定な要素があることである．例えば, 選手のコンディションや知名度が挙げられる．市場価値の高い選手でも, コンディションが悪いために大会中にはなかなか活躍できなかったり, 市場価値が低くまだ世界的な認知度の無い隠れた有能選手が大会で大活躍をした結果, 市場価値との相関が低くなったと考えられる．残りの 2 つは, 市場価値の性質によるものである．1 つは, 市場価値はクラブでの活躍によって変動するものなので, ワールドカップのデータを活用した本研究では, アクションデータと市場価値の間に乖離が生じてしまったこと, もう 1 つは, 市場価値は選手の年齢による影響が大きいために, 年齢が高い選手の活躍度と市場価値の相関が低くなってしまったことである．

既存の VAEP との性能比較では, VAEP の評価値よりも提案モデルの評価値の方が市場価値との相関が強いことを示した．攻撃面の評価値に関してはどちらのモデルも同じであるから, 両モデルの差は守備評価の性能によって生まれたと言える．しかし, ポジションがセントラルミッドフィルダーの選手に関しては VAEP の性能が上回った．この理由は次のように考えられる．本研究では, 攻守の両面から選手を評価しているが, 一部のポジションでは, 攻撃面だけで評価した方が市場価値との相関が高くなる場合があり, セントラルミッドフィルダーもその一つである．VAEP によってセントラルミッドフィルダーの選手を評価した場合, 守備の評価値が全体的に同じような値になるという傾向がみられたために, 結果的に守備評価の差がなくなり, 攻撃評価だけで差がついたので VAEP の方が一見性能が良いようにみられたのである．つまり, 守備評価だけで比較した

場合, 提案モデルの性能が上回っていると考えられる. 実際の値を表 13 に示す.

表 13 VAEP と提案モデルの守備性能比較 (CM)

	VAEP		提案モデル	
ポジション	p	R_s	p	R_s
CM	0.050	0.069	**0.243**	**0.249**

表 13 より, 提案モデルの方が市場価値との相関係数が高いことが分かる. よって, 予想が正しかったことが示された.

5.3 実験 3: 攻守の評価による選手のプレースタイル分類

各ポジションに対する攻守のプレースタイル分類の結果から, 2 つの発見が得られた. 1 つ目は, 同じポジション内でも, こなす役割によって評価が違うということである. 例えば, ポジションがセンターフォワードの選手の中でも, Antoine Griezmann や Thomas Müller のようなセカンドストライカーの働きもこなす選手は他の選手よりも守備評価が高くなっている. このような選手たちは, ストライカーよりも後ろのポジションをこなすこともあるために, その分守備のタスクが大きくなっているので, 守備評価が高くなっていると考えられる. つまり, ポジションがフォワードの選手において, グラフ右寄りのクラスターの選手はストライカーのプレースタイルであり, 中央より上のクラスターの選手はセカンドストライカーなど守備のタスクもこなすプレースタイルであると言える. 2 つ目は, ポジションがディフェンスの選手は, フォワードの選手よりも攻撃評価と守備評価に強い正の相関が見られるということである. この原因としては, フォワードの選手においては, 攻撃に特化した選手や攻守に良い働きを見せる選手など様々な選手がいるが, ディフェンスの選手においては, 良い守備のできる選手は良い攻撃もでき, どちらかがうまくいかないと両方うまくいかなくなってしまうという特徴を示しているためだと考えられる. つまり, フォワードの選手には多様なプレースタイルがあるが, ディフェンスの選手には守備に偏ったプレースタイルの選手が少なく, 攻守両面をこなすプレースタイルの選手が多いということである.

6 おわりに

本研究では, サッカーの試合中に起こるアクションデータを活用して, 選手を評価する上での既存モデルの課題を克服し, 選手を攻守両面から総合的に評価できる新しいモデルを提案した. 結果として, 既存モデルの守備面の評価指標を改善することで, 既存モデルよりも高い精度で選手を評価できることを示した. また, 提案モデルによる攻守の評価値を基に行った選手のプレースタイル分類では, 選手の特徴を客観的に分析することを可能にした. しかし, サッカーには, ボールの無いところでの動きなど, アクションデータだけでは再現できない重要な動きが多くあり, 完全に選手の試合への貢献度を評価することはできなかった.

今後の展望としては, より正確に選手の評価を行うために, アクションデータからアクションデータには無い動きを予測することで, アクションデータに無い情報を再現できるモデルを作成していきたい. また, 今回の実験では, 提案モデルの活用法がプレースタイル分類のみだったので, 実際のサッカーに役立つような提案モデルの活用を行っていきたい. 例えば, 選手のプレースタイルや攻守の評価値から, この選手がどのようなクラブチームに行けば活躍できるか, 反対にこのクラブチームにはどのような選手が必要かを提案するといった活用法が挙げられる. このような方法を用いて提案モデルを実際のサッカーに役立てることで, 提案モデルの重要性を示していきたい.

謝　辞

本研究は JSPS 科研費 JP21H03496, JP22K12157, JP23H03688 及び公益財団法人住友電工グループ社会貢献基金の助成を受けたものです.

文　献

[1] Ryan Beal, Georgios Chalkiadakis, Timothy J. Norman, and Sarvapali D. Ramchurn. Optimising Game Tactics for Football. AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, 2020
[2] Ryan Beal, Georgios Chalkiadakis, Timothy J. Norman, and Sarvapali D. Ramchurn. Optimising Long-Term Outcomes using Real-World Fluent Objectives: An Application to Football. AAMAS '21: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 2021
[3] Ian Simpson, Ryan J. Beal, Duncan Locke, and Timothy J. Norman. Seq2Event: Learning the Language of Soccer using Transformer-based Match Event Prediction. KDD '22: Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2022
[4] 相澤 宏樹, 清 雄一, 田原 康之, 大須賀 昭彦. プロサッカーチームにおける試合状況とボールリカバリーの関係の考察. 情報処理学会研究報告, ICS, 2020
[5] Pegah Rahimian and Laszlo Toka. Inferring the Strategy of Offensive and Defensive Play in Soccer with Inverse Reinforcement Learning. MLSA 2021: Machine Learning and Data Mining for Sports Analytics, 2021
[6] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Jon Davis. Actions Speak Louder than Goals: Valuing Player Actions in Soccer. KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019
[7] Tom Decroos and Jesse Davis. Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, 2019
[8] Tom Decroos, Maaike Van Roy, and Jesse Davis. SoccerMix: Representing Soccer Actions with Mixture Models Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, 2020
[9] Toda K, Teranishi M, Kushiro K, and Fujii K. Evaluation of soccer team defense based on prediction models of ball recovery and being attacked: A pilot study. PLoS ONE 17(1): e0263051, 2022
[10] Coach Watson. 350+ Soccer Terms Everyone Must Know (A – Z Guide) – https://www.soccercoachingpro.com/, 2018
[11] All data provided by StatsBomb – https://statsbomb.com/
[12] According to transfermarkt.com – https://www.transfermarkt.com/

音響型 AR オンラインマラソンシステムにおける空間認識率向上に向けた取り組み

小西　侑樹† Panote Siriaraya†† 栗　達††† 田中　克己††††
河合由起子† 中島　伸介†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
†† 京都工芸繊維大学 情報工学・人間科学系　〒 606–8585 京都市左京区松ヶ崎橋上町
††† 福岡大学 工学部　〒 814–0180 福岡県福岡市城南区七隈 8 丁目 19 番 1 号
†††† 福知山大学 情報学部　〒 620-0886 京都府福知山市堀 3370
E-mail: †{i2386067，kawai，nakajima}@cc.kyoto-su.ac.jp，††spanote@kit.ac.jp，†††lida@fukuoka-u.ac.jp，††††tanaka-katsumi@fukuchiyama.ac.jp

あらまし　近年の健康志向の高まりに伴い，健康改善や体力向上の目的でランニングを実施する人口が増加している．しかし，ランニングのモチベーション維持は容易ではないため継続しないケースも多い．そのためランニング支援システムを開発する意義は大きいと考えている．我々は先行研究において，音響型拡張現実空間内にて過去に記録した走行データを用いて作成した仮想ランナーとの伴走を可能にするランニング支援システムとそれを用いた不特定多数のユーザとの競走ができるオンラインマラソンアプリを開発した．また，アプリを用いた仮想ランナーの認識のしやすい適切な人数についての検証を行った．本稿では人数の検証のみではなく，競走中の仮想ランナーとの前後の位置関係を容易に認識できるような音声の組み合わせおよび演出について調査を行った．

キーワード　ランニング支援システム，仮想ランナー，音響型 AR，GPS

1. はじめに

厚生労働省による健康活動促進に向けた啓発方策の後押しもあり，健康改善や体力向上を目的としたランニングに取り組む人は近年増加傾向にある [1]．スポーツ庁の実施した「スポーツの実施状況等に関する世論調査」によると，ランニングは 1 年間に実施した種目，今後始めてみたいスポーツの両項目において上位に入っている [2]．しかしながら，ランニングに対するモチベーションの維持は困難であり，40〜65 %ほどの人は運動を開始しても 3〜6ヶ月以内に辞めてしまうと予測されるほど継続が困難なスポーツである [3] [4] [5]．そのため，ランニングを楽しくかつ体力向上を支援するシステムを開発する意義は大きいと考える．

近年ではスマートフォンやスマートウォッチなどを活用してランニングを支援するアプリが普及しており，代表的なものとして Nike Run Club [6] や Strava [7]，ASICS Runkeeper [8] などがある．これらのランニング支援システムは距離やペース，カロリー計算などのワークアウトの管理やランニングコースの推薦，プランの作成などを行えることが特徴である．また，ランニング支援の別のアプローチとして Zombies,Run [9] などのゲーミフィケーションを活用したアプリも増加している．ゲーミフィケーションとは，ゲームの要素をゲーム以外の分野に導入する手法であり，人を没頭させ，継続的にわくわく感を提供することによって内的動機づけを促す手法である [10]. ゲーミフィケーションを活用したランニング支援システムでは他のユーザとの記録の共有やゲーム内によるバッジなどの報酬によってユーザのランニングのモチベーションを上げるものが一般的である．しかし，これらの既存のランニング支援システムは実際にランニングを行う際は 1 人で走る場合が多く，ランニング自体を楽しませるというアプローチに欠けている．そのため，既存のランニング支援システムによるモチベーションの維持は不十分であると考える．

そこで我々は先行研究において，音響型拡張現実空間内にて過去に記録した走行データを用いて作成した仮想ランナーとの伴走を可能にするランニング支援システムとプロトタイプアプリの開発，身体能力やモチベーションに対する影響の検証を行った [11] [12]．また，その検証結果をもとに，近年開催されることが多くなっているオンラインマラソンを参考にした大規模かつ長距離の競走が可能なアプリの開発とそれに伴うアプリのユーザビリティ評価，身体的影響及び走行時に出現する仮想ランナーの適切な人数についての検証を行った [13].

本稿では AR 空間における認識率向上に向けた取り組みとして行なった競走中の仮想ランナーとの前後の位置関係を容易に認識できるような音声の組み合わせと認識の補助をするような演出に関する調査の結果を報告する．

2. 関連研究

ランニングは継続して行うことが難しく，モチベーションを維持するための「動機付け」が必要である．なぜなら，ランニングは一般的には辛い，つまらない，苦手といった印象を抱くものだからである．

実際にこのようなランニングの問題点を解決するためのラ

ンニング支援システムは数多く存在する．Nike Run Club や STRAVA などのランニング支援アプリケーションが有名である．また，ランニングのためのゲーム的要素が強い Zombies，Run! や厚生労働省とスポーツ庁がリリースした FUN+WALK [14] といったものも存在する．

Nike Run Club では，消費カロリーや距離，速度，心拍数と言った日々のワークアウトの管理や facebook を通じて記録のシェアができる．また，各ユーザの運動レベルや目標に応じてトレーニングプランの推薦を行うことが特徴である．しかし，ランニングに対して高い意識を持っているユーザには日々の身体管理や記録の管理が行いやすいが，ランニングに対して苦手意識が強い健康を目的としたユーザにとっては動機付けを行うには難しい．

STRAVA はアスリート向けのランニングに限らず，サイクリングや他の様々なスポーツのアクティビティを記録，分析を行う．このアプリケーションの特徴は全ユーザが走ったランニングコースを記録し，ユーザに対してルートの推薦を行ってくれることである．また，走ったルートの写真や記録を他のユーザとシェアでき，ユーザのモチベーションを高めることができるが，アスリート向けの要素が多く，健康を目的としたユーザにはモチベーション維持が難しい．

Fitbit [15] では歩数，距離，心拍数などの様々なデータを管理でき,「アドベンチャー」と呼ばれるゲーミフィケーションを用いた機能がある．この機能では，歩数目標を達成することによって世界の様々なランドマークやトレイルの景色を画像として見ることができる．また，ユーザ間で歩いた歩数を競い合うことでモチベーションを上げる．

FUN+WALK では歩数計を利用し，ユーザの歩数に応じてアプリケーション内のキャラクターが変身したり，実際に様々な店舗で使用できる割引クーポンが手に入る．このアプリケーションはユーザの収集欲を利用したゲーミフィケーションの一つである．

また近年では GPS アプリを使用したマラソンの大会などが開催されることが多くなっている．多くの大会で使用されているアプリとしては TATTA [16] が有名である．TATTA はランニングポータルサイト RUNNET と連携する唯一の GPS ランニングアプリであり，開催されている多くの GPS マラソンイベントに参加ができ，自身で計測した記録の管理ができることが特徴である．しかし，このような GPS マラソンに参加して記録を行う際は一人で計測することが多くモチベーション維持が難しい．

また，近年 VR 技術を使用した運動支援に関する研究も行われている．Soumya C. Barathi [17] らは VR ヘッドセットを活用した自転車でのゴーストとの競走を行うことで，VR 運動ゲームによるパフォーマンスの向上と内発的モチベーションへの有効性の検証を行った．その結果インタラクティブ・フィードフォーワード法 [18] でのユーザのパフォーマンスと モチベーションの向上を示した．また，自己競走とのゴーストの競走下においてもパフォーマンス の向上及びモチベーションの向上を確認している．坂口ら [19] は，VR 映像を提示することによる

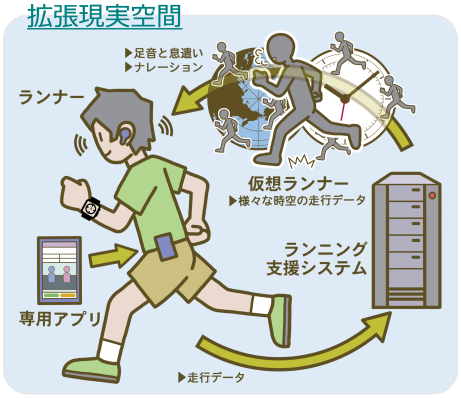


図 1: 音響型 AR ランニング支援システムの概要



図 2: 音響型 AR ランニング支援システムの使用イメージ

運動への影響を調査を行った．その結果 VR 映像を用いることで，使用者の無自覚的な運動量の増加や，運動意識の向上が得られることを示している．Blanchfield [20] らは擬似競技を用いてのセルフトークが持久力パフォーマンスに及ぼす影響を調査した．その結果，疲労を感じた時にセルフトークを行うことで自身とモチベーションを維持し，運動パフォーマンスの向上につながることを示している．

しかし，これまでの VR 技術を使用した運動支援に関する研究では VR 映像や音楽・効果音を用いたものが主であった．そこで本研究では，ランニングの擬似競技において，足音やナビゲーションなどの音響的な手がかりを使用することによるユーザのモチベーション及びパフォーマンスの向上を補助するようなシステムの開発を目指している．

3. 音響型 AR を用いたランニング支援システム

3.1 立体音響を用いたランニング支援

先行研究にて作成したシステム概要を図 1, 使用イメージを図 2 に示す．ユーザは拡張現実空間内にて生成された仮想ランナーと音響型 AR を用いて競走や伴走が可能である．仮想ランナーとは過去に記録した走行データ (走行距離，速度) を基に作成された擬似的な競走相手である．実際にユーザがこのシステムを使用する際はイヤホンを装着し，HRTF アルゴリズム [21]



(a) ホーム画面　(b) 設定画面　(c) マラソン選択画面　(d) ラン設定画面　(e) 走行画面

図 3: オンラインマラソンアプリの画面推移

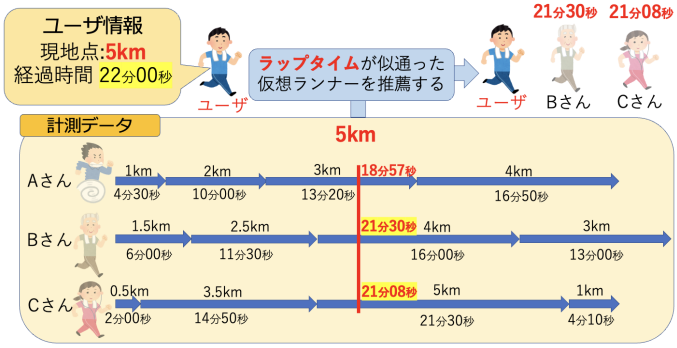


図 4: 仮想ランナーの推薦方式

を利用した 3D 立体音響で再現される仮想ランナーの足音や息遣いの音声を聞きながら競走を行う．

3.1.1　走行距離の算出方法

走行距離はスマートフォンの GPS を用いて 1 秒間隔で取得した緯度・経度の座標から Haversine 式を用いて算出した距離である．Haversine 式は以下の (1) の通りとなっており，r は地球の半径（6371 キロメートル）を表し，l_1 と lo_1 は GPS の取得した場所 n の緯度と経度を表し，l_2 と lo_2 は GPS の取得した場所 n+1 の緯度と経度を表している．

$$d = 2r\sqrt{sin(\frac{l_2 - l_1}{2})^2 + cos(l_1)cos(l_2)sin(\frac{lo_2 - lo_1}{2})^2} \quad (1)$$

3.1.2　速度の算出方法

ランニング支援システムでは仮想ランナーとユーザの速度の算出方法は異なる．ユーザの競走時の速度は仮想ランナーと競走を行うため，リアルタイムで算出する必要がある．そのためランニング時に取得した側近同士の緯度経度（p_k，p_{k-1}）の差分の距離 d_k を元に速度を算出する．それに対して仮想ランナーの速度は過去にランニングを行った際の距離とかかった時間から算出した平均速度を用いる．

3.2　オンラインマラソンシステム

3.2.1　概　　要

本節では我々が開発したオンラインマラソンアプリについて説明する．このアプリは，オンラインマラソンのようにユーザが競走する際の競走距離や走行回数の選択を自由にでき，不特定多数のユーザと競走することが可能となっている．オンラインマラソンとは近年開催されることが増えているマラソンの形式であり，スマートフォンやスマートウォッチを活用することで遠隔地からの参加が可能な点が最大の特徴である．運営側が提示した期限内に走った総距離で目標距離を目指す形式であり，自分の好きな時間，好きな距離，好きな回数に分けて行える自由度の高さも特徴の一つである．

開発したオンラインマラソンアプリの画面推移図を図 3 に示す．機能としてはあらかじめ定めた複数人の仮想ランナーとの競走が可能なフリーランモードとサーバから他ユーザの走行データを取得して作成した仮想ランナーとの競走が可能なオンラインマラソンモードの二種類がある (3a)．本節ではオンラインマラソンモードの機能について紹介する．オンラインマラソンモードで走行を行う際は，マラソン選択画面で参加するマラソンの種類を選択する (3c)．選択可能なマラソンについては以下の通りである．

- 3km マラソン
- 5km マラソン
- 10km マラソン
- 20km マラソン
- ハーフマラソン (21.098km)
- 30km マラソン
- フルマラソン (42.195km)
- 50km マラソン
- ウルトラマラソン (100km)

マラソン選択画面の中央にある「仮想ランナーはグループ内で探す」のチェックを入れると，同じマラソンに参加している且つ設定画面 (3b) にて作成, 参加したグループに所属しているユーザ間でのみの競走が可能となる．マラソンの種類を選択した後，ユーザはランニングを行う距離を選択する (3d)．距離の選択は 100m 刻みでの自由な選択が可能である．その後, 走行



図 5: 音声パターンの調査概要

画面 (3e) にて「Start Run」ボタンを押すと 3 秒のカウントの後, ランニングが開始される.

3.2.2 サーバーとの連携

オンラインマラソンモードでは走行を行うごとにサーバに走行データが保存される. サーバが保管する情報は, 個人 ID, マラソン ID, グループ ID, 緯度・経度及び算出した走行距離, 走行時間である. ユーザが仮想ランナーとの競走を行う際は, サーバに記録された他ユーザの走行データから推薦を行う.

3.2.3 仮想ランナーの推薦方法

仮想ランナーはマラソン選択画面にて選択したマラソンに参加している他のユーザの記録タイムと総走行距離をもとに算出したラップタイムの近い上位 5 名を推薦する (図 4). また, マラソンという長距離の競走では仮想ランナーと距離が離れすぎて足音や息遣いの音声が聞こえずに競走感を失ってしまう恐れがあるため, 100m ごとにその時点でラップタイムが近い 5 名へと更新を行う.

3.3 UX 向上に向けたシステム改良

以前に行った検証実験では, 開発したオンラインマラソンアプリにおいて, 走行中に出現する仮想ランナーの人数がどの程度なら聞き分けやすいかの検証を行った [13]. 検証の結果, 人数が減るごとに聞き分けのしやすさは上昇したが, 人数に関わらず前後関係が分かりにくい, また足音以外の音声が欲しいなど位置関係を認識する際の問題点が浮き彫りになった. 今後, よりユーザ体験の高い競走を実現可能なオンラインマラソンアプリを開発するためにはこれらの問題を解決する必要がある. そのため我々は, これらの問題点を対処する上で必要な調査を行った. 4 章にて実施した内容の詳細を述べる.

4. 認識率向上に向けた調査アンケート

3.3 で述べた UX 向上に向けた試みとして, 我々は音声パターンの違いによる認識への影響の調査とエフェクトによる認識への影響の調査を実施した. この章では実施したアンケートの概要及び調査結果について述べる.

4.1 音声パターンの違いによる認識への影響の調査

4.1.1 音声パターン調査の概要

この調査は複数の仮想ランナーとの競走時において各仮想ランナーが発する個々の音声の識別が容易となる組み合わせを調べることを目的とし, 大学生 17 名を対象に行った. 実施した音声パターン調査の概要を図 5 に示す. アンケート回答者には 4 パターンの音声組み合わせ動画を視聴してもらい,「仮想ランナーの聞き分けやすさ」と「使用したいと感じたもの」の各項目について順位をつけてもらった. 音声組み合わせ動画は開発したオンラインマラソンアプリの走行時に表示される画面であり, 仮想ランナーの人数は 5 人としている. 各パターンにおける 5 人の仮想ランナーそれぞれが発する音声については以下の条件を考慮した音声ファイルをコンテンツ素材サイト Storyblocks [22] にて独自に選定して割り当てた.

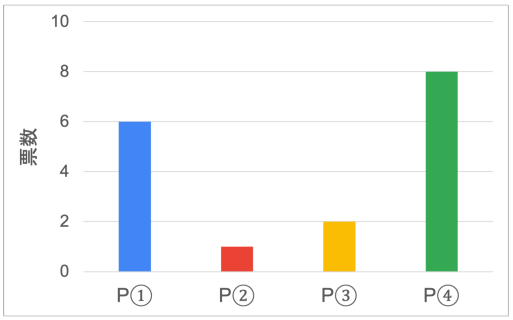


図 6: 「仮想ランナーの聞き分けやすさ」の 1 位の票数

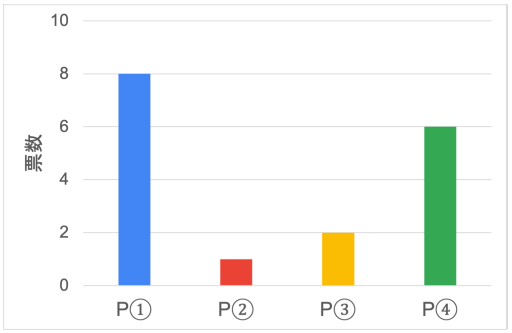


図 7: 「使用したいと感じたもの」の 1 位の票数

パターン 1 (P①)
既存のシステムで使用していた音声の組み合わせ

パターン 2 (P②)
息遣いが含まれる足音を増やした音声の組み合わせ

パターン 3 (P③)
道路状態が大きく異なる音声の組み合わせ

パターン 4 (P④)
息遣いを含まない足音だけの音声の組み合わせ

4.1.2 結　　果

表 1: 「仮想ランナーの聞き分けやすさ」の平均順位

	P①	P②	P③	P④
順位の平均値	2.06 位	2.88 位	3.12 位	1.94 位

表 1 は「仮想ランナーの聞き分けやすさ」という質問に対する被験者の各音声パターンに対する平均順位を示しており, 図 6 は各パターンにおける 1 位の票数を示している. これらを確認すると, 票数, 平均順位と共に P④が一番順位が高く, 次いで P①が高い結果となっており, 下位 2 パターン (②,③) と差がついているのが確認できる.

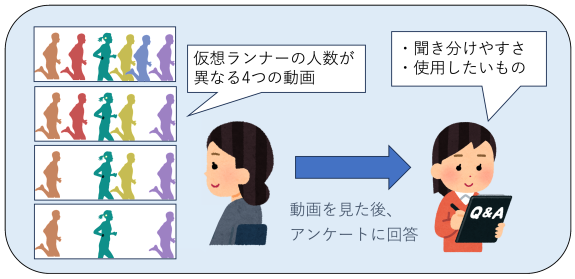


図 8: 仮想ランナーの人数調査概要

表 2: 「使用したいと感じたもの」の平均順位

	P①	P②	P③	P④
順位の平均値	1.88 位	3.18 位	3.12 位	1.82 位

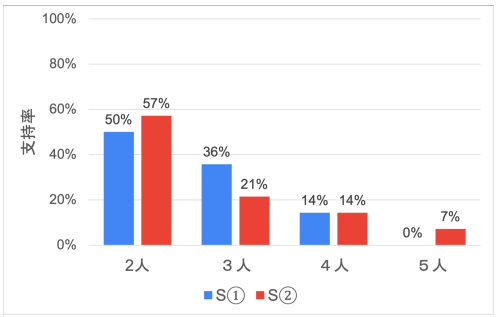


図 9: 「仮想ランナーの聞き分けやすさ」の支持率

表 2 は「使用したいと感じたもの」という質問に対する被験者の各音声パターンに対する平均順位を示しており，図 7 は各パターンにおける 1 位の票数を示している．これらを確認しても同様に票数，平均順位と共に P④と P①が下位 2 パターンと比べて大きく差があり，高い結果となっている．

4.1.3 考　　察

調査の結果を確認すると，どちらの項目においても P④と P①が高く，P③と P②が低い結果となっていた．これらより，走行中に音のみで競走相手を聞き分ける際には足音以外の音声は極力抑えて耳に入る音の種類を少なくシンプルにすることが効果的だと考えられる．

4.1.4 調査したパターンでの人数調査

4.1.1 にて実施したアンケート結果から息遣いを含まない足音だけの音声の組み合わせ (以下 S①) と既存のシステムで使用していた音声の組み合わせ (以下 S②) の音声が概ね高い結果を示していた．そのため，上記の 2 パターンの音声を使用して音声の違いによって適切な仮想ランナーの人数に違いが生じるかを調査することを目的としたアンケートを大学生 14 名を対象に行った．実施した人数調査の概要を図 8 に示す．アンケート回答者には S①と S②のそれぞれの音声を使用した仮想ランナーの人数の異なる 4 つの動画 (2〜5 人) を視聴してもらい，「仮想ランナーの聞き分けやすさ」と「使用したいと感じたもの」の各項目について順位をつけてもらった．

4.1.5 結　　果

表 3 と表 4 はそれぞれの音声パターンについて「仮想ランナーの聞き分けやすさ」という質問に対する被験者の各仮想ランナーの人数に対する平均順位を示しており，図 9 は各人数における 1 位の票数の割合 (支持率) を示している．これらを確認すると，両音声パターン共に票数，平均順位は人数が少ない方が高い結果となっているのが確認できる．

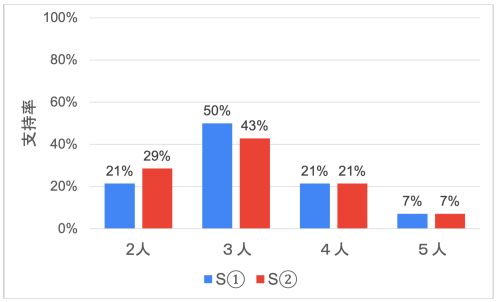


図 10: 「使用したいと感じたもの」の支持率

表 3: S①:「仮想ランナーの聞き分けやすさ」の平均順位

	2 人	3 人	4 人	5 人
順位の平均値	1.64 位	2.00 位	2.57 位	3.78 位

表 4: S②:「仮想ランナーの聞き分けやすさ」の平均順位

	2 人	3 人	4 人	5 人
順位の平均値	1.71 位	2.21 位	2.57 位	3.50 位

表 5: S①:「使用したいと感じたもの」の平均順位

	2 人	3 人	4 人	5 人
順位の平均値	2.21 位	1.79 位	2.36 位	3.64 位

表 6: S②:「使用したいと感じたもの」の平均順位

	2 人	3 人	4 人	5 人
順位の平均値	2.07 位	2.00 位	2.43 位	3.50 位

表 5 と表 6 は両音声パターンにおける「使用したいと感じたもの」という質問に対する被験者の各仮想ランナーの人数に対する平均順位を示しており，図 10 は各人数における 1 位の票数の割合 (支持率) を示している．これらを確認すると，票数，平均順位と共に 3 人，2 人，4 人，5 人の順で高い結果となっていることが確認できる．

4.1.6 考　　察

2 つの音声パターンを使用した人数調査の結果，両パターンとも「仮想ランナーの聞き分けやすさ」の項目において人数が少ない方が高い結果が得られ,「使用したいと感じたもの」に関しても 3 人が多かったものの概ねの傾向は少ない人数の方が高いことがわかった．今後実際にアプリの改良を施す際にはこれらの音声パターンの調査結果を参考に設定しようと考えている．

4.2 エフェクトによる認識への影響の調査

4.2.1 調 査 概 要

実施したエフェクト調査の概要を図 11 に示す．この調査は仮想ランナーとの競走時において，エフェクトを追加すること

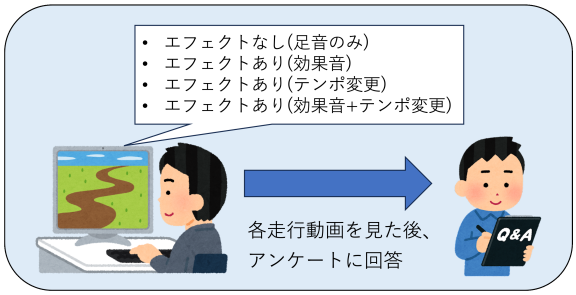


図 11: エフェクトによる認識への影響調査の概要

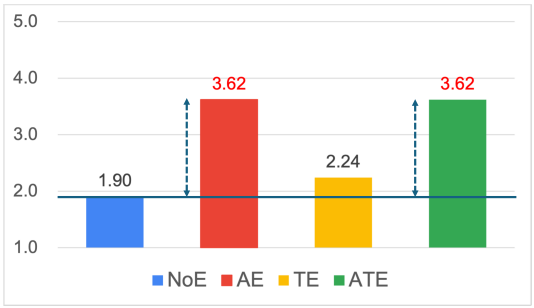


図 12: 追い抜いたケース：追い抜いた表現がわかりやすいか

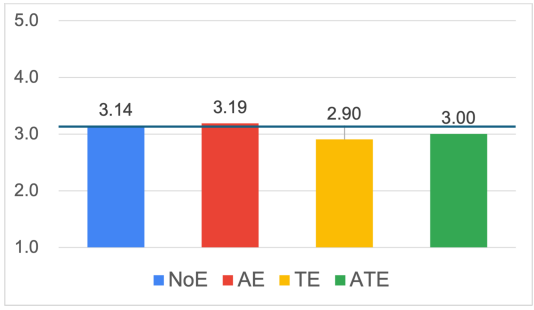


図 13: 追い抜いたケース：臨場感はあると思うか

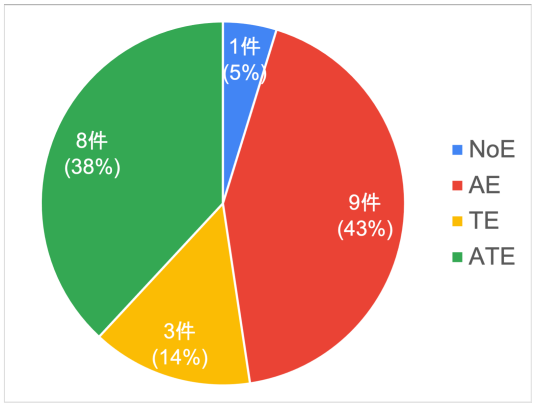


図 14: 追い抜いたケース：どの演出が一番いいと感じたか

による前後の位置関係に関する認識への影響を調べることを目的としている．この調査は，本学及び他大学の学生, 教職員を含めた 21 名を対象に行った．アンケート回答者には仮想ランナーとの 1 対 1 での競走における追い抜いたケースと追い抜かれたケースの動画をそれぞれ 4 パターンずつ試聴してもらい，独自アンケートに回答してもらった．独自アンケートでは「追い抜いた (追い抜かれた) 表現が分かりやすいか」,「臨場感はあると思うか」という 5 段階のリッカート尺度を用いた質問を各動画を視聴した後に回答してもらった．また，4 パターンの動画を視聴した後に「どの演出が一番いいと感じたか」という質問に回答してもらった．調査で使用した動画は実際に仮想ランナー 1 人と競走しているユーザの一人称視点の動画である．各パターンの動画の条件は以下の通りである．

仮想ランナーを追い抜いたケース

・エフェクトなし (NoE)
・エフェクトあり：追い抜いた際に効果音 (AE)
・エフェクトあり：足音のテンポ変更 (TE)
・エフェクトあり
: 追い抜いた際に効果音＋足音のテンポ変更 (ATE)

仮想ランナーに追い抜かれたケース

・エフェクトなし (BNoE)
・エフェクトあり：追い抜かれた際に効果音 (BAE)
・エフェクトあり：足音のテンポ変更 (BTE)
・エフェクトあり
: 追い抜かれた際に効果音＋足音のテンポ変更 (BATE)

また，使用した効果音についてはフリー効果音サイトにて選定した．足音のテンポに関しては追い抜いたケースは遅く，追い抜かれたケースは追い抜かれるまで徐々に速くしていった．

4. 2. 2 結　　果

図 12 は，仮想ランナーを追い抜いたケースにおける各パターンに対する「追い抜いた表現が分かりやすいか」という質問において，5 段階のリッカート尺度を使用して測定された値の平均を比較したものである．NoE の平均値は 1.90(SD=1.04)，AE の平均値は 3.62(SD=1.16)，TE の平均値は 2.24(SD=1.04),ATE の平均値は 3.62(SD=0.97) となった．また，NoE と AE，NoE と TE,NoE と ATE の差をウィルコクソンの符号順位和検定 ($p<.05$) を用いて評価した結果，NoE と AE,NoE と ATE において $p<.001$ となり，有意な差が見られ,NoE と TE においては $p=.33$ となり，有意な差が見られなかった．

図 13 は, 仮想ランナーを追い抜いたケースにおける各パターンに対する「臨場感はあると思うか」という質問において，5 段階のリッカート尺度を使用して測定された値の平均を比較したものを示している．NoE の平均値は 3.14(SD=1.24)，AE の平均値は 3.19(SD=1.17)，TE の平均値は 2.90(SD=1.34),ATE の平均値は 3.00(SD=1.26) となった．また，NoE と AE，NoE と TE,NoE と ATE の差をウィルコクソンの符号順位和検定 ($p<.05$) を用いて評価した結果，NoE と AE においては $p=.87$，NoE と TE においては $p=.31$，NoE と ATE においては $p=.79$ となり，有意な差が見られなかった．

図 14 は仮想ランナーを追い抜いたケースの 4 パターンの動画を視聴した後に回答してもらった「どの演出が一番いいと感じたか」という質問の解答割合である．NoE は 1 件で 4.8 %，AE は 9 件で 42.9 %，TE は 3 件で 14.3 %,ATE は 8 件で 38.1 % となった．

図 15 は，仮想ランナーに追い抜かれたケースにおける各

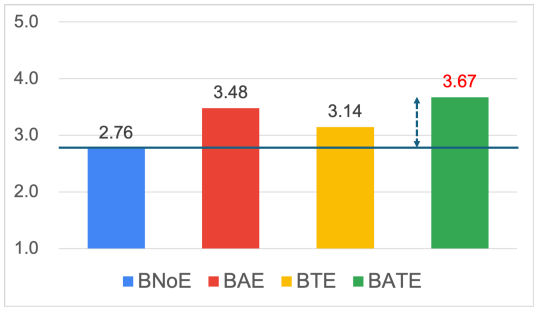


図 15: 追い抜かれたケース：追い抜かれた表現がわかりやすいか

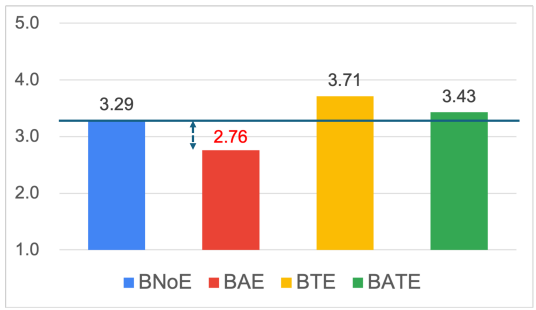


図 16: 追い抜かれたケース：臨場感はあると思うか

パターンに対する「追い抜かれた表現が分かりやすいか」という質問において，5 段階のリッカート尺度を使用して測定された値の平均を比較したものである．BNoE の平均値は 2.76(SD=1.37)，BAE の平均値は 3.48(SD=1.25)，BTE の平均値は 3.14(SD=1.28),BATE の平均値は 3.67(SD=1.11) となった．また，BNoE と BAE，BNoE と BTE,BNoE と BATE の差をウィルコクソンの符号順位和検定 ($p<.05$) を用いて評価した結果，BNoE と BATE においては $p=.02$ となり，有意な差が見られ，BNoE と BAE においては $p=.06$，BNoE と BTE においては $p=.17$，となり，有意な差が見られなかった．

図 16 は, 仮想ランナーに追い抜かれたケースにおける各パターンに対する「臨場感はあると思うか」という質問において，5 段階のリッカート尺度を使用して測定された値の平均を比較したものを示している．BNoE の平均値は 3.29(SD=1.19)，BAE の平均値は 2.76(SD=1.09)，BTE の平均値は 3.71(SD=1.15),BATE の平均値は 3.43(SD=1.33) となった．また，BNoE と BAE，BNoE と BTE,BNoE と BATE の差をウィルコクソンの符号順位和検定 ($p<.05$) を用いて評価した結果，BNoE と BAE においては $p=.04$ となり，有意な差が見られ，BNoE と BTE においては $p=.16$，BNoE と BATE においては $p=.56$ となり，有意な差が見られなかった．

図 17 は仮想ランナーに追い抜かれたケースの 4 パターンの動画を視聴した後に回答してもらった「どの演出が一番いいと感じたか」という質問の解答割合である．BNoE は 2 件で 9.5 %，BAE は 4 件で 19 %，BTE は 6 件で 28.6 %,BATE は 9 件で 42.9 %となった．

4.2.3 考　　察

調査の結果を確認すると，追い抜いたケースにおける各パ

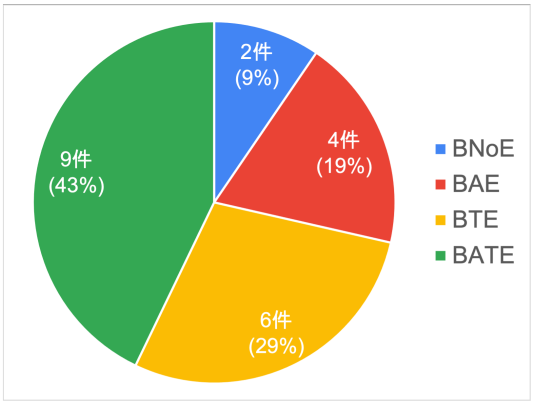


図 17: 追い抜かれたケース：どの演出が一番いいと感じたか

ターンに対する「追い抜いた表現が分かりやすいか」という質問については AE と ATE の値が高く，NoE と比較して有意差があることがわかった．追い抜かれたケースにおける各パターンに対する「追い抜かれた表現が分かりやすいか」という質問についても BATE の値が高く，NoE と比較して有意差があることが確認でき，有意差はないものの BAE においても値が高いことが確認できた．そのため，前後の入れ替わり時に効果音をつけることは前後関係の認識を向上するのに有効であると考えられる．追い抜かれたケースにおいては有意差はないものの BTE の値も高いことが確認できる．これは迫ってくる際に音声のテンポを上げることによって実際の競走のように緊迫感が生じたためだと考えている．両ケースにおける「どの演出が一番いいと感じたか」という質問もこの傾向に従っているため，効果音は両ケースの場合で有用であり，追い抜かれる表現においてはテンポをあげる設計をするのがよいと考えられる．

5.　まとめと今後の展望

以前に実施したオンラインマラソンアプリを使用した検証実験にて人数に関わらず前後関係が分かりにくい，また足音以外の音声が欲しいなど位置関係を認識する際の問題点が浮き彫りになった．そのため本稿では，上記の問題に対処するためのアプローチとして音声パターンの違いによる認識への影響の調査とエフェクトによる認識への影響の調査を実施した．音声パターンの違いによる認識への影響の調査を行った結果として，P④と P①が聞き分けやすさと実際に使用してみたいと感じたかの両項目において高い結果となり，P②や P③に関しては低い結果となった．これらのことから走行中に音のみで競走相手を聞き分ける際には足音以外の音声は極力抑えた音声を選択する方が良いと考えられる．また，P④と P①を使用した仮想ランナーの人数調査を行った結果，3 人の場合が聞き分けやすさと実際に使用してみたいと感じたかの両項目とも高い結果となった．これらの音声調査の結果をアプリの改良の参考にしようと考えている．また，エフェクトによる認識への影響を調査した結果としては，両ケースにおいて順位の変動がある際に効果音をつけた条件の値が高くなり，追い抜かれるケースにおいては BTE の値も高くなった．このことから効果音は位置関係の認識に対して良い効果が期待でき，追い抜かれる表現におい

てはテンポをあげる設計つつ，組み込むことがよいと考える．今後は，これら 2 つの調査結果を踏まえてオンラインマラソンアプリに適応したのち，実際に使用した際の影響について検証することを考えている．

謝　　辞

本研究の一部は，科研費 (課題番号：20H04293，22H03700，22K19837) および京都産業大学先端科学技術研究所（ヒューマン・マシン・データ共生科学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 厚生労働省．健康づくりのための身体活動基準． https://www.mhlw.go.jp/content/10904750/001171393.pdf.
[2] スポーツ庁．令和 4 年度「スポーツの実施状況に関する世論調査」について. https://www.mext.go.jp/sports/content/20230420-spt_kensport01-000028572-01.pdf.
[3] James Annesi. Effects of a cognitive behavioral treatment package on exercise attendance and drop out in fitness centers. *European Journal of Sport Science*, Vol. 3, No. 2, pp. 1–16, 2003.
[4] James J Annesi. Effects of computer feedback on adherence to exercise. *Perceptual and Motor Skills*, Vol. 87, No. 2, pp. 723–730, 1998.
[5] Rod K Dishman. *Exercise Adherence: Its Impact on Public Health*. Human Kinetics, 1988.
[6] 最高のランニングパートナー，NIKE+RUN CLUB アプリ，Nike．com(JP). https://www.nike.com/jp/nrc-app.
[7] Strava の機能 — GPS による追跡，地図，分析，チャレンジ，友達，トップランニングやライドの検索. https://www.strava.com/features.
[8] ランニングを記録しよう - ASICS Runkeeper ランニング記録アプリ. https://runkeeper.com/ja/.
[9] Emma Witkowski. Running from zombies. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, pp. 1–8, 2013.
[10] 山崎和子. ゲーミフィケーション（gamification）. 知能と情報, Vol. 25, No. 3, pp. 93–93, 2013.
[11] 桐生拓海, Mohit Mittal, Panote Siriaraya, 河合由起子, 中島伸介. 音響型 ar ランニング支援システムによるモチベーションの向上と身体的影響の検証. In *DEIM Forum 2020*, pp. I2–1, 2020.
[12] 小西侑樹, 有澤隆生, Panote Siriaraya, 栗達, 河合由起子, 中島伸介. 音響型 ar ランニング支援システムのライバル推薦における複数仮想ランナーの影響に関する検証. In *DEIM Forum 2022*, pp. A31–3, 2022.
[13] Yuki Konishi, Panote Siriaraya, Da Li, Katsumi Tanaka, Yukiko Kawai, and Shinsuke Nakajima. Development of an online marathon system using acoustic ar. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9388–9389, 2023.
[14] FUN+WALK アプリ | SPECIAL — FUN + WALK PROJECT ポータル. https://funpluswalk.jp/special/app/.
[15] Fitbit アドベンチャー. https://www.fitbit.com/jp/challenges/adventures.
[16] TATTA powered by RUNNET. https://tatta.runnet.jp/about/.
[17] Soumya C Barathi, Daniel J Finnegan, Matthew Farrow, Alexander Whaley, Pippa Heath, Jude Buckley, Peter W Dowrick, Burkhard C Wuensche, James LJ Bilzon, Eamonn O'Neill, et al. Interactive feedforward for improving performance and maintaining intrinsic motivation in vr exergaming. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.
[18] P. W. Dowrick. A review of self modeling and related interventions. *Applied and Preventive Psychology*, Vol. 8, No. 1, pp. 23–39, 1999.
[19] 坂口正道, 田山滉士. サドル型走行デバイスを用いたバーチャルランニングにおける vr 映像の効果. 日本バーチャルリアリティ学会大会論文集, 9 月 2021.
[20] Anthony W Blanchfield, James Hardy, Helma M De Morree, Walter Staiano, and Samuele M Marcora. Talking yourself out of exhaustion: the effects of self-talk on endurance performance. *Med Sci Sports Exerc*, Vol. 46, No. 5, pp. 998–1007, 2014.
[21] Corey I Cheng and Gregory H Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. In *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.
[22] Storyblocks. https://www.storyblocks.com/.

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

2024年3月1日(金) 16:00 ~ 18:10 | T5-B オンライン（Zoom Events)

セマンティクス

座長:岩井原 瑞穂(早稲田大学)
コメンテータ:北島 信哉(富士通株式会社)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:20
[T5-B-9-01] 気象データを活用した省電力灌水システムの提案

*真鍋 武丸[1]、大塚 真吾[1] (1. 神奈川工科大学)

16:20 ~ 16:45
[T5-B-9-02] 機械学習と地理情報検索に基づく地物の意味発見

羽田野 唯人[1]、*渡邊 二葉[1]、田中 克己[1] (1. 福知山公立大学情報学部)

16:45 ~ 17:10
[T5-B-9-03] 交通計測地点の増加に対応する知識転移を用いた時空間グラフモデリング

*川端 智紀[1]、戸田 浩之[1] (1. 横浜市立大学)

17:10 ~ 17:35
[T5-B-9-04] 地理オブジェクトに関する事前学習を利用した地域特性推定

*福地 湧[1]、牛尼 剛聡[2]、田中 克己[3]、角谷 和俊[1] (1. 関西学院大学 社会情報デザイン研究室、2. 九州大学、3. 福知山公立大学)

17:35 ~ 18:00
[T5-B-9-05] チャット付き地図における会話・検索履歴に基づく地図検索手法

*小柳 慶真[1]、小竹 神[1]、北山 大輔[1] (1. 工学院大学)

気象データを活用した省電力灌水システムの提案

真鍋　武丸† 　大塚　真吾†

† 神奈川工科大学情報学部情報工学科　〒 243–0292 神奈川県厚木市下荻野 1030
E-mail: †s2021021@cco.kanagawa-it.ac.jp, ††otsuka@ic.kanagawa-it.ac.jp

あらまし　農業分野では担い手の減少や高齢化にともない，労働人口不足が深刻な問題となっている．2015 年には 175.7 万人いた基幹的農業従事者が 2022 年では 122.6 万人にまで減少している．また，新規就農者や後継者の数も減っており，2015 年では 6.5 万人だった新規就農者が 2022 年では 4.58 万人にまで減少している．そのため，これまで培われてきた経験やノウハウ，営農技術などを継承していくことが難しくなっている．そこで，本研究では新規参入者の障壁となっている初期費用などのコストを抑えることを考慮しつつ，農業に必要である気象データや環境データの取得や保存を行うシステムを提案する．さらに，ノウハウや経験が重要である農作物への水やりについても，取得したデータをもとに自動散布可能なシステムを提案する．実験では，ソーラパネルやデジタルタイマーを用いてシステムの低電力化を試み，その結果，ほとんど電力を消費することなく 1 日稼働し続けることを確認した．また，システムの消費電力以外にも，待機電力等も考慮し，ソーラパネルを利用することで数ヵ月稼働し続けるシステムについても考察を行った．

キーワード　農業支援, IoT, センサーデータ

1 はじめに

農業は現在担い手の減少や高齢化に伴い労働人口の不足が深刻な問題になっている．農林水産省の調べによると，2015 年では 175.7 万人いた基幹的農業従事者が 2022 年では 122.6 万人にまで減少している．労働人口が減少している要因として，新規参入者や後継者が少ないことが挙げられ，2015 年の新規就農者が 6.5 万人に対して，2022 年は 4.58 万人まで減少している [1]．その理由として,「令和 3 年度の新規就農者の就農実態に関する調査結果 [2]」などにおいて，就農 1 年目にかかる営農費用は平均 755 万円であり，そのうち機械・設備等に 561 万円かかるという結果が出ている [3]．したがって，農業に携わりたいと考えている人にとって，膨大な初期費用が参入障壁となっている状況である．さらに，農業は農家のこれまでの経験や勘などが重要となるため，技術面や経験などに不安を感じてしまう人が多くいることも原因だと考えられる．調査報告 [2] によると,「離農する農家の有形資産と無形資産をまとめて引き継いだ」という回答が 4.9%,「離農する農家の経営資産をまとめて引き継いでいない」という回答が 87.7%となっており，技術や経験，ノウハウなどをほとんど継承できていないのが現状である．

そこで，本稿では初期費用などのコスト面を考慮しつつ，農業に必要である環境データ等の取得や保存を行うシステムを提案する．さらに，農家のノウハウや経験が重要である農作物への水やりについても，取得したデータをもとに自動散布可能なシステムを提案する．これにより高齢化が進みノウハウや経験，技術等の継承が難しいという問題解決の手助けとなり，新規参入者や後継者不足の解決に役立つと考えている．

2 関連研究

スマート農業を活用することで省力・軽労化，さらに栽培技術等の継承が期待されている．また，農園に気象センサーを設置し，情報収集を行う研究は数多く行われている [4–9]．しかし，市場規模が拡大する一方で，システムが高額であるためなかなか導入にまで至らないのが現状である．本節では関連研究や製品についての概要を述べる．

2.1 e-kakashi

「e-kakashi [10]」は農業を科学的に支援するサービスであり，ソフトバンクが運営している．このサービスでは，温湿度，日射量，土壌湿度，CO2 濃度などの環境データをはじめ，水やりや施肥などの農作業記録，生育記録，収穫量といった営農に必要なデータや農家の方々が培ってきた経験・技術などの栽培ノウハウをインプットすることができ，後継者育成や技術継承に役立てることが可能である．また，センサーは 10 分毎に計測を行い，1 時間ごとに 3G/4G 無線通信を用いて集約したデータをクラウドへアップロードする．アップロードされたデータはすべてウェブアプリやスマホアプリから確認することが可能である．価格については，端末代が 1 台で 10 万 9,780 円，端末 1 台当たりの月額接続料が 1,078 円，1 契約当たりの月額クラウド使用料が 4,378 円である [11].

2.2 営農再開後における通い農業の支援を行うシステム

東北大震災後に営農再開地域で通い農業を行っている生産者や，耕地やハウスが分散している生産者を対象に，水稲育苗だけでなく園芸作物を生産する際に生じる作業の判断や見回りの省力化を目的としたシステムを提案している．拡張性の高い市販プロトタイピング・キットとメッセージアプリの Web API

を連携させて生産者のスマートフォンにハウス環境情報を通知する安価かつ簡便なハウス遠隔監視システム「通い農業支援システム」を構築した．導入コストに関しては，１棟１年費用が8,890 円であり，ランニングコストであるデータ SIM 通信量を除くと約 20,000 円である．電気代以外のランニングコストは通信料 990 円/月である [12].

3　システムの概要

スマート農業の問題点として，システムの構築やランニングコストにかかる費用の高騰が挙げられる．そのため，本システムでは，比較的安価な Raspberry Pi と Arduino UNO を用いて，図 1 のようなシステムの構築を行う．Raspberry Pi には土壌センサー，給水ポンプ，Weather Shield を繋げた Weather meters kit をそれぞれ接続し，Arduino には温湿度センサーを接続することでデータをそれぞれ取得する．また，温湿度センサーと土壌水分量のデータをもとに自動灌水を行うことが可能である．Arduino UNO と Raspberry Pi の通信はデータ送信用と受信用の XBee を使用する．センサーから取得したデータは，公衆回線を通してクラウドサーバーへとアップロードすることで，Web ブラウザ等から温湿度，土壌水分量のデータをグラフで閲覧可能となる．さらに，取得したデータは LINE などのメッセージアプリにも送信可能なため，農家がいつでも簡単に圃場の状況を確認可能である．データの送信等に必要なネットワーク環境は SORACOM の SIM を使用したモバイル Wi-Fi を用いており，各機器への電力供給はソーラパネルを接続したバッテリーを使用する．今回用いたバッテリーは「Jackery ポータブル電源 400(112,200mAh/400Wh)」であり，ソーラパネルは「Jackery SolarSaga 100W」を使用した．天気の良い日中では，約 6 時間でポータブル電源をフル充電することが可能である．また，デジタルタイマーを用いてシステムの電源を制御することで，システムの低電力化を図る．

4　実 証 実 験

この節では我々が構築したプロトタイプシステムの実証実験について延べる．

4.1　バッテリー消費量に関する実験

構築したシステムで用いる機器において，電力消費量が多いと考えられるものは，Raspberry Pi，Arduino，給水ポンプ，モバイル Wi-Fi である．モバイル Wi-Fi は電源管理を自動的にできないため，予備実験において電力消費量の確認を行った．その結果，モバイル Wi-Fi は通信を行っていない時の待機電力が低いことが分かったため，今回は常時起動させて実験を行った．今回使用したバッテリーは，通常の AC コンセントと USB コンセントの 2 種類の出力があるため，モバイル Wi-Fi は常時 USB コンセントに繋ぎ，その他の機器については通常の AC コンセントや USB コンセントに繋いで実験を行った，システムを稼働させるバッテリーの消費量を調べるため，以下の 3 つの方式で実験を行った．

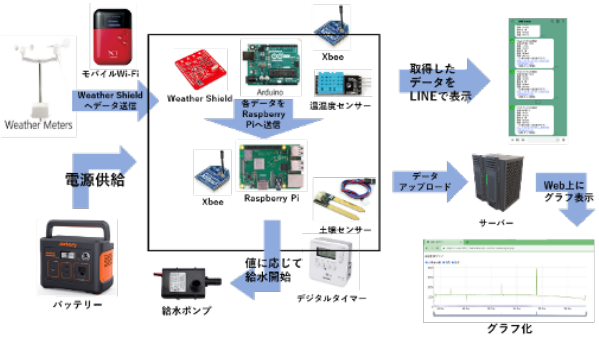


図 1　システム構成図

全ての機器に常時電源供給

Raspberry Pi，Arduino，給水ポンプ，モバイル Wi-Fi など，全ての機器はデジタルタイマーを介さずに AC コンセントや USB コンセントから電源供給を受ける．

Raspberry Pi のみ電源制御

AC コンセントの 1 つにデジタルタイマーを設置し，そこに Raspberry Pi のみを接続する．それ以外の機器は，デジタルタイマーを介さずに AC コンセントや USB コンセントから電源供給を受ける．

主な機器を電源制御

AC コンセントにデジタルタイマーを設置し，そこに Raspberry Pi，Arduino，給水ポンプを接続する．それ以外の機器は，デジタルタイマーを介さずに AC コンセントや USB コンセントから電源供給を受ける．

バッテリー残量の確認はタイムラプスカメラを使用し，システム稼働時間はすべて 20 時間とした [1]．それぞれの方式の結果を以下の図 2,3,4 に示す．

図 2 より，全ての機器に常時電源供給を行った場合のバッテリー残量は 56%であり，Raspberry Pi のみ電源制御した場合のバッテリー残量は 73%であった．また，主な機器を電源制御した場合のバッテリー残量は 78%であった．これらの結果より，全ての機器に常時電源供給を行った場合は，20 時間が経過した時点でバッテリー容量の半分近くの電力を消費してしまうことがわかる．一方でデジタルタイマーを使用して電源の制御を行うことで，バッテリーに直接つないだ場合よりも消費電力を抑えられることもわかる．さらに，Raspberry Pi のみ電源を制御するよりも主要な機器を電源制御する方がバッテリーの残量は多いことがわかる．これらの実験から，デジタルタイマーで各機器の電源を制御することで消費電力が抑えられ，長時間にわたってシステムを稼働し続けられることがわかる．また，本実験ではバッテリーの仕様により通電ボタンが押されている状態では待機電力が発生しており，この影響によりバッテリーの残量が予想以上に大きいという結果であった．対策としては，コンセントを差しても待機電力が流れないバッテリーを使用することで，本実験以上に消費電力を抑えることができると考えている．

1：デジタルタイマーの仕様上，1 日のタイマー設定回数が 14 回のため，1 時間 42 分に 1 回電源の ON/OFF を行った．

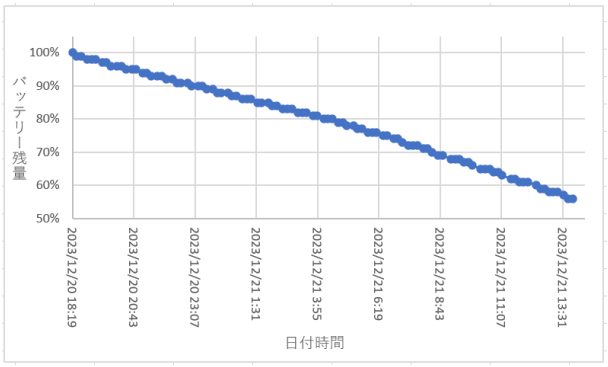


図 2　全ての機器に常時電源供給

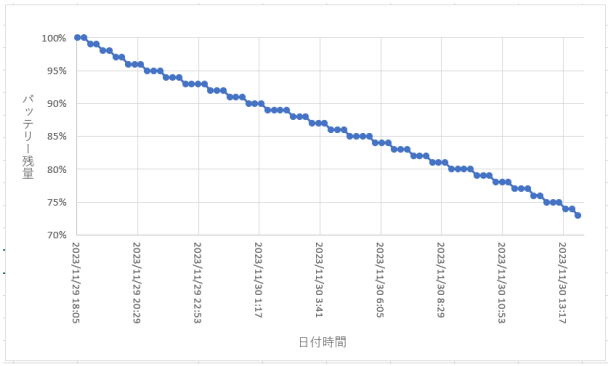


図 3　Raspberry Pi のみ電源制御

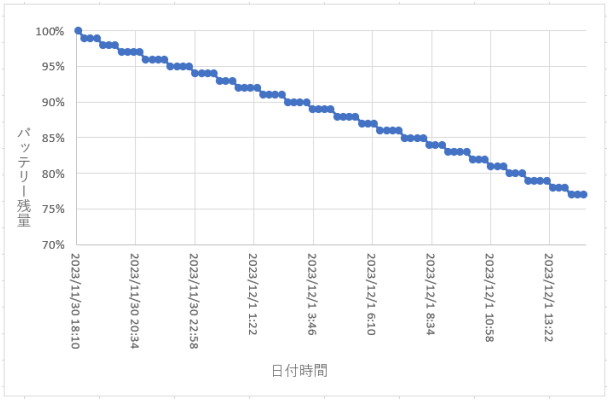


図 4　主な機器を電源制御

4.2　ソーラパネルを用いたバッテリー充電

バッテリーに接続することで充電可能なソーラパネルを用いてバッテリー充電実験を 2 日間行った．1 日目は，夕方ごろに行ったため陽が落ち始めてしまい 40 分ほどで実験を終了した．2 日目は午前中から 2 時間ほど行った．充電開始前のバッテリー残量は 75%であった．実験結果を図 5,6 に示す．

図 5 より，約 40 分の充電で 75%から 82%まで充電された．太陽光がソーラパネルに当たっている間は，しっかりと充電されていたが，陽が沈み始め日陰に入ってしまうと全く充電されなくなった．また，2 日目は 1 時間ほどの充電で 100%まで充電することができた．さらに，ソーラパネルでバッテリーを充電しながら各機器を接続したところ，ソーラパネルで充電しながら電力供給も可能であることが確認できた．これらの結果から，日中はソーラパネルを繋ぎ，充電することでバッテリー残量を気にせずにシステムを稼働させ続けられると考えられる．一方で夕方に実験を行ったことで少しでも日陰になると充電速度が落ちることを確認した．そのため，夜間などの陽の光が当たらない時間は充電されていない状態でシステムを稼働させることになる．しかし，本システムでは将来的には 1 時間に 1 回数分間システムが通電する仕様にしているため，日中にソーラパネルを用いてバッテリーを充電しておくことで夜間など太陽光が得られない時間帯での電力供給も可能であると考えている．

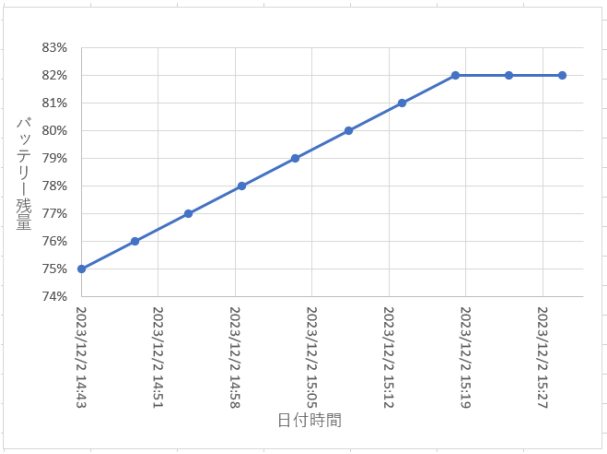


図 5　太陽光による充電 (1 日目)

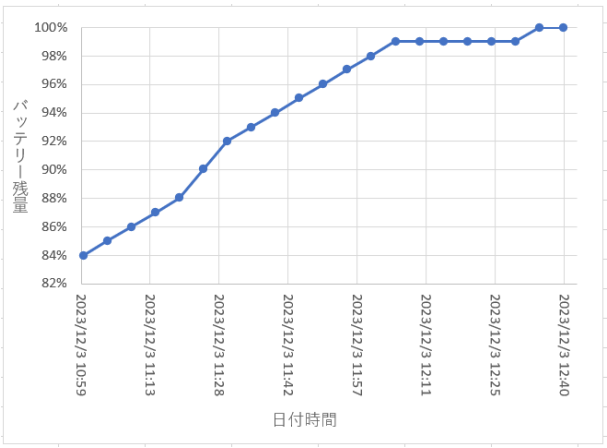


図 6　太陽光による充電 (2 日目)

4.3　システムの稼働とバッテリー消費量

提案システムが 1 日にどの程度バッテリーを消費するかを検証するために，デジタルタイマーを使用せずに各機器をバッテリーに直接接続して実験を行った．まず，バッテリーの通電ボタンを ON にし，データを取得したらシステムをシャットダウンさせ，バッテリーの通電を OFF にする．これを 24 回繰り返す実験を行った．この実験により 1 時間に 1 回の計 24 回システムを動かすことで，1 日のバッテリー減少量を擬似的に検証することが可能である．実験結果を以下の図 7 に示す．

24 回システムの ON/OFF を繰り返した結果，バッテリー残量は 100%から 98%に減り，わずか 2%しか電力を消費しなかった．この結果から 24 時間つまり 1 日システムを動かして

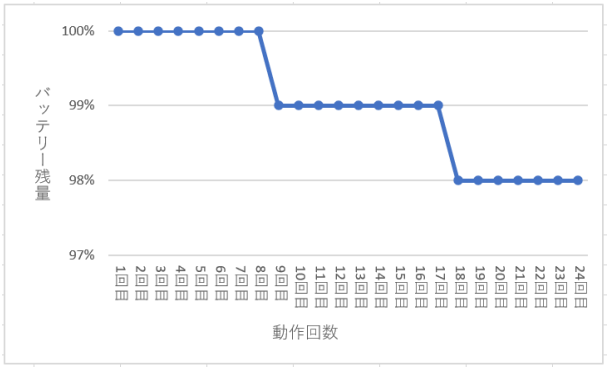


図 7　システムの起動を 24 回繰り返して行った結果

も 2%しかバッテリーを消費しないという計算になる．4.1 節での実験に比べるとかなりの消費電力を抑えられていることがわかる．今回の実験では手動でバッテリーの通電制御を行ったため，待機電力によるバッテリーの電力消費がなくシステムの稼働による消費電力のみであるため，低電力で稼働することができたと考えられる．

5　おわりに

現在，農業では新規就農者や後継者不足，労働人口の減少，ノウハウや技術の継承が難しいなど様々な問題を抱えている，本研究ではコスト面を考慮に入れながら，農業に必要な気象データ等の取得や保存を行うためのシステム構築，農作物への水やりについても，取得したデータをもとに自動散布可能なシステム構築を行った．しかし，本研究で構築したシステムの価格を計算するとおよそ 45,000 円程度となり，さらに，モバイル Wi-Fi などの通信費やバッテリー代，ソーラパネルの費用なども考慮する必要がある．しかし，モバイル Wi-Fi に必要な通信コストは 1 日当たり 10 円程度と比較的安価な SIM を使用していることや，システム全体をデジタルタイマーによる電源制御で消費電力を抑えられているため，通信費や電気代の面では従来の商用システムよりの安価なシステムが構築できたと考えている．また，システム面では LINE 上で取得したデータからグラフ化したデータを確認することができるため，LINE 上で圃場の状態や環境データ，グラフまで一括で確認できる点は，直感的で分かりやすく誰でも簡単に扱うことができるものになっている．本研究では，土壌水分量をはじめとした 6 つの気象・環境データの取得を行った．しかし，これらのデータはごく一部であり，その他にも農業をするにあたって様々なデータを取得する必要がある．それらのデータが誰でも簡単かつ安価に取得できるようになることでスマート農業は今後さらに発展し，新規就農者の増加や農業後継者問題の解決へと繋げていくことができるのではないかと考える．

文　献

[1] 農林水産省，「農業労働力に関する統計」，https://www.maff.go.jp/j/tokei/sihyo/data/08.html，2023.
[2] 一般社団法人全国農業会議所 全国新規就農相談センター，「新規就農者の就農実態に関する調査結果 - 令和 3 年度 -」，https://www.be-farmer.jp/uploads/statistics/YV447s7CQjwBYJ3OtEht202203231858.pdf，2022.
[3] BASF ジャパン株式会社，「新規就農者の資金はいくら必要？内訳や平均，調達方法を解説」，https://minorasu.basf.co.jp/80446，2022.
[4] 藤井宏次朗，渡邊修平，村上幸一，圃場管理のためのフィールドセンサー情報のグラフ化とアラート機能の開発，電子情報通信学会技術研究報告，LOIS，ライフインテリジェンスとオフィス情報システム 112(466)，pp.1-4，2013.
[5] 松野智明，増井崇裕，安部恵一，峰野博史，大須賀隆司，水野忠則，無線センサネットワークを利用した農業支援環境の見える化の実現と評価，第 73 回全国大会講演論文集 2011(1)，pp.167-168，2011.
[6] 中野達彦，増井崇裕，安部恵一，峰野博史，大須賀隆司，水野忠徳，農業疎密無線センサネットワークにおける Data MULE 型データ通信を利用するハイブリッドエナジーハーベスティングセンサノードの開発と評価，研究報告マルチメディア通信と分サイン処理 (DPS) 2013-DPS-155(11)，pp.1-6，2013.
[7] 田尻久幸，内尾文隆，張勇，松田憲幸，瀧寛和，井口信和，亀岡孝治，果樹栽培のための知的センサネットワーク，農業機械学会誌 64(Supplement)，pp.493-494，2002.
[8] 井口信和，谷口祐一，内尾文隆，瀧寛和，亀岡孝治，農場ネットワークのための優先度と電力を考慮したＩＥＥＥ８０２．１１ｅによるアドホック通信方式，農業情報研究 16(13)，pp.81-90，2007.
[9] 大塚真吾，「農家におけるセンサデータ取得の現状」，知能と情報 (日本知能情報ファジィ学会誌)，Vol.30，No.1，pp.35-41，2018.
[10] ソフトバンク株式会社，「e-kakashi」，https://www.e-kakashi.com/.
[11] ソフトバンク株式会社，「農業 AI ブレーン「e-kakashi」を大幅リニューアルして機能拡充と低価格化を実現」，https://www.softbank.jp/corp/news/press/sbkk/2021/20211004_01/，2021.
[12] 山下善道，内藤裕貴，稲葉修武，根本知明，金井源太，星典宏，「営農再開後の通い農業を支援するハウス遠隔監視システムの開発とその展開」，農研機構研究報告，Vol.2021，No.8，pp.211–230，2021.

機械学習と地理情報検索に基づく地物の意味発見

羽田野 唯人† 渡邊 二葉† 田中 克己†
†福知山公立大学情報学部 〒602-0886 京都府福知山市字堀 3370
E-mail: †{32045068, 32145103, tanaka-katsumi}@fukuchiyama.ac.jp

あらまし 地理情報システム(GIS）における地物の意味を発見する手法を提案する．地物の意味は，対象地物の種別と対象地物の周辺領域にある地物群で定義される．地物の意味を発見するために 2 つの方法を提案する．1 つ目の方法は，Word2Vec を 2 次元的に拡張した GeoObject2Vec という機械学習法に基づくもので，これによって地物の意味の分散表現を得ることができる，2 つ目の方法は，地理情報検索に基づいて地物の意味の言語表現を得るものである．これら 2 つの手法を相補的に適用することで，地物の意味を発見することを目指している．

キーワード 地理データ活用 地物の意味発見 機械学習 地理情報検索

1. まえがき

近年，機械学習やニューラルネットワークに基づく深層学習の技術が飛躍的に発展しつつある．特に，テキストや画像の深層学習に基づいて多くの応用が開発されつつある．一方，地理情報システム（GIS）分野でも，機械学習の利用も急速に開始されており，この分野における新しい研究領域として GeoAI が注目を集めている[Lavallin and Downs, 2021] [Li, 2021]. GeoAI は，AI，地理情報ビッグデータ，並列処理を組み合わせた新しい研究領域として認知されている [Usery et al, 2021].

本論文では，地図データからの地物の意味獲得を目的とした研究について報告する．我々は，地物の意味獲得を行うために 2 つの独立した方法を開発した．1 つ目の方法は，地物の意味の分散表現を機械学習で得るものであり，2 つ目の方法は地物の言語的な意味を地理情報検索によって求めるものである.最終的には，この２つの独立した手法を相補的に実行することで地物の意味を発見することを目指している.

地物の意味の分散表現を機械学習で求めるために，Word2Vec を拡張した GeoObject2Vec アルゴリズムを提案する.このアルゴリズムは直観的には Word2Vec の 2 次元的な拡張を行ったものであり，これによって，地物（例えば，あるカフェ）の意味の分散表現を，この地物の周辺にある地物群（例えば，大学の施設･建物や学生アパートなど）を学習することで得ることができる．GeoObject2Vec で機械学習するデータとして OpenStreetMap（OSM）データを用いることとし，日本全国の地物データを PostgreSQL を用いてデータベース化したのでこれについても報告する.

次に，地物の「言語的な」意味を，地理情報検索で獲得するために，地物検索クエリ群を定義しクエリの実行結果の評価を行った.この地理情報検索によって，地物（例えば，あるカフェ）の言語的な意味（例えば，「学生街のカフェ」）を得られている.

最後に，機械学習による意味発見と地理情報検索による意味発見を相補的に適用することで，地物の意味発見性能を改善できる見込みを得たので報告する.

本論文の構成は，以下の通りである．2 節では，地物の意味について定義･説明を行う．3 節では，地物の意味の分散表現を求める GeoObject2Vec 手法と学習用の OpenStreetMap データベースについて説明する.4 節では，地物の「言語的な」意味を，地理情報検索で獲得する手法と実験結果などについて説明する．5 節は機械学習による意味発見と地物情報検索による意味発見の相補的な利用について説明する．6 節は本論文の結論である.

2. 地物の意味

2.1. 語の意味

図 1 に示すように，語の意味はその語の周辺に現れる単語群によって規定されるという Harris の分散仮説が著名である [Harris, 1954] ．この仮説に基づいて Mikolov らは語の意味の分散表現を求めるニューラルネットワークアルゴリズム Word2Vec を提唱した [Mikolov et al., 2013]. 例えば，図 1 においては，単語 iPod の意味はその単語の周辺に現れる単語群{Apple, popular, portable, music, players}や{audio, recording, downloaded, portable, listening}で規定される.

Mikolov らは，この Word2Vec を拡張して，文や文書の意味の分散表現を求めるアルゴリズムを提唱した [Le and Mikolov, 2014]. Yating らは，語の意味は時間や空間によって変化することに着目して，時空間を超えて語の意味を求める手法を提案している[Zhang et al., 2016], [Zhang et al., 2017].



図 1　Harris の分散仮説による語の意味

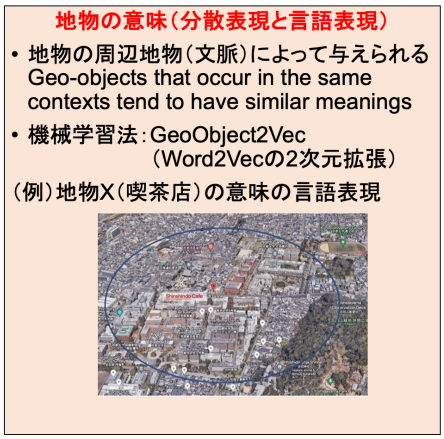


図 2　地物の意味：京大周辺にある喫茶店「進々堂」

2.2. 地物の意味

地物（geographic objects または features）とは，直観的には，地図に現れる任意の地物オブジェクトであり，建物，施設，行政区域，鉄道駅，バス停留所 are any type，河川，湖沼などである．GIS(Geographic Information System)では，地物は次の 3 種類に類別して扱われている．

- 点オブジェクト（point-type objects）
  住居，建物，鉄道駅，バス停留所など
- 線オブジェクト（line-type objects）
  道路，河川，鉄道など
- 多角形オブジェクト(polygon-type objects)
  行政区域，湖など

前節で述べた Harris の語の意味[Harris, 1954]を拡張して，地物の意味を次のように定義する．

「地物の意味は，その地物の周辺の地物群によって規定する」

例えば，京都大学近辺にある喫茶店「進々堂」の意味を考えよう．図 2 に示すように，この喫茶店は京都大学の様々な施設・建物に囲まれたエリアにある．従って，この喫茶店の「意味」は，周辺の大学施設・建物群によって規定されていると考えると，「（京都大学周辺の）学生街にある喫茶店」がその意味になるものと考えられる．すなわち，この例では，地物の意味は，その周辺にある地物群（大学関連の建物・施設）と対象地物の種別（喫茶店）によって規定されると考えることができる．

3. 機械学習による地物の意味獲得

3.1. GeoObject2Vec: 周辺地物の機械学習による地物の意味獲得

Mikolov らは，語の意味を機械学習で計算するアルゴリズム Word2Vec を提案した [Mikolov et al., 2013]．本論文では，地物の意味を計算するために，この Word2Vec アルゴリズムを拡張した GeoObject2Vec アルゴリズムを提案する．

Word2Vec では各語に対してその周辺にある語群を学習するものであったが，GeoObject2Vec では各地物に対してその 2 次元的な周辺にある地物群を学習するものである．この意味では，GeoObejct2Vec は Word2Vec を単純に 2 次元的に拡張したものということができるが，「語」と「地物」との相違を考慮して，さらに以下に述べるような変更を行ったものである．

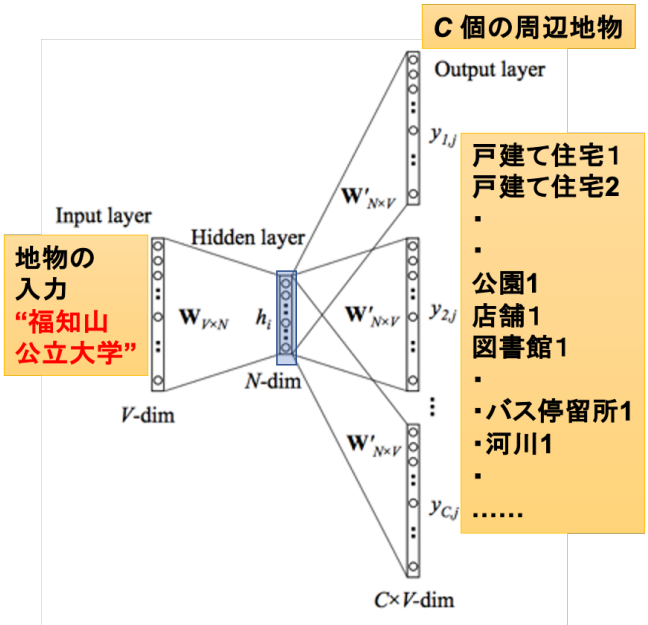


図 3　GeoObject2Vec による地物の意味（分散表現）の獲得

GeoObject2Vec と Word2Vec の主な相違点は以下の通りである．

① **地物の抽象化と機械学習**

地物の意味を求める際，その地物の 2 次元的周辺地物名群を機械学習してしまうと，対象とする地物の意味が，具体的な周辺地物名群で表せてしま

い，どの地物の意味もその周辺にある具体的な地物名群になってしまい，地物の意味類似検索が実質上不可能になってしまう．そこで，GeoObejct2Vec の入力層は各地物に対応するニューロンを配置するが，出力層は地物名ではなく地物の型に対応するニューロンを配置する．

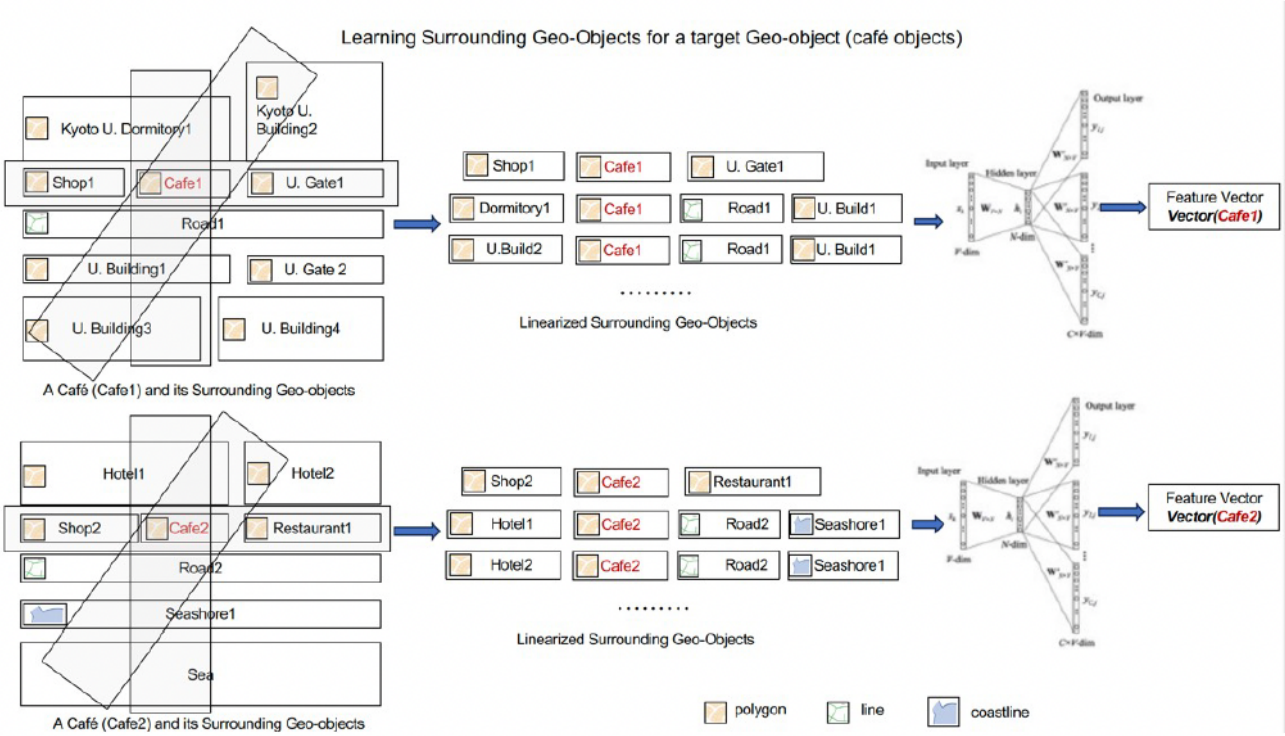


図 4　GeoObject2Vec による周辺地物の機械学習: Cafe1 と Cafe2

例えば，図 3 に示すように，地物「福知山公立大学」の GeoObject2vec による機械学習を考えてみよう．入力層は個々の地物（名）に対応するニューロンが配置されている．一方，出力層は，個々の地物（名）に対応するニューロンではなく，地物の「型」に対応するニューロンが配置されている．このような形で学習が進むと，地物「福知山公立大学」の意味は，出力層で発火しているニューロン群（周辺地物の型の集合）で表現されたものが意味となる．

②　2 次元周辺地物群の線形化

Word2Vec の学習データは，一次元のテキストデータ，すなわち，ある単語の周辺に出現する単語列である．一方，GeoObject2Vec の学習データは 2 次元的に配置された周辺地物群である．このため，GeoObject2Vec では，学習に用いるデータ，すなわち，2 次元的に配置された周辺地物群のデータを「線形化」して複数の学習データとしている．

図 4 は，周辺地物の機械学習を GeoObject2Vec アルゴリズムで行った例である．2 つのカフェ Cafe1 と Cafe2 を対象として，各々のカフェの周辺地物群は 2 次元に広がっているため，周辺地物群を線形化して学習している．Cafe1 の周辺は大学関係の建物・施設が密集しており， Cafe2 の周辺は海岸沿いのホテルやショップが密集している．

③　多種類の周辺地物の学習

GIS では，地物はそのタイプに応じて 3 種類の地物（点オブジェクト，線オブジェクト，多角形オブジェクト）に類別されている．従って，GeoObject2Vec では，このような 3 種類の地物を周辺オブジェクトとして学習する必要がある．

図 4 に Cafe1, Cafe2 の周辺地物を示したが，この中で，例えば，大学の建物・施設やホテル・レストラン・店舗は点オブジェクトであり，道路は線オブジェクトであり，海岸は多角形オブジェクトであることに注意を要する．

④　周辺領域の決定（範囲やサイズ等）

地物の「周辺領域」を決めるには，その範囲や大きさを決定する必要があるが，範囲や大きさは対象地物の種類に依存することに注意を要する．

例えば，対象地物がカフェである場合と城である場合，これらの地物のサイズが異なるため，それらの「周辺

領域」の範囲・サイズも異なるものに設定した方が良いと考えられる（本論文で提案する GeoObject2Vec アルゴリズムでは,「周辺領域」の範囲・サイズは固定したものを用いているが，これは今後の課題であると認識している.）

3.2. OSM データの機械学習

OpenStreetMap (OSM) はオープンで自由にアクセスできる地物データ（geographical data）である[OSMF]. OSM データは，地物名，座標，種々の属性データからなる．図 5 に OSM の属性データの例を示す.

我々は，日本全国の地物の OSM データをダウンロードして PostgreSQL DBMS にテーブルとして保存した．この PostgreSQL DBMS に保存された OSM データを機械学習に用いるために，周辺領域にある地物をアクセスするための Python プログラムなどを開発した．図 6 に，開発した Python プログラムの 1 例を示す．このプログラムは，周辺領域内にあるすべての点オブジェクトを求めるものであり，データベースへのアクセスを行う部分は PostGIS SQL 質問[PostGIS]を埋め込んだ形で実装を行っている.

	true_east_longitude	true_north_latitude	osm_id	amenity	name	way
0	141.937433	45.522049	9906880680	restaurant	てっぺん市場	POINT (15800402.733 5704085.243)
1	141.937549	45.522008	9906880679	restaurant	最北端	POINT (15800415.690 5704078.729)
2	141.940905	45.520457	9009885855	police	稚内市警察署宗谷岬駐在所	POINT (15800789.290 5703832.272)
3	141.879652	45.487013	2669723413	townhall	稚内市役所 宗谷支所	POINT (15793970.592 5698520.387)
4	141.880270	45.486571	2669723412	police	駐在所	POINT (15794039.399 5698450.109)
5	141.880903	45.486276	2669723411	post_office	宗谷郵便局	POINT (15794109.853 5698403.347)
6	141.877106	45.464042	8401628253	fuel	安田石油	POINT (15793687.195 5694873.571)
7	141.879244	45.460585	8401628240	fuel	木村石油ss	POINT (15793925.152 5694324.955)
8	141.752234	45.403851	2615825252	place_of_worship	聲間神社	POINT (15779786.497 5685325.153)
9	141.814568	45.375091	2681531579	restaurant	さるぼぼ	POINT (15786725.541 5680766.402)

図 5 OpenStreetMap の属性データ

```
def search_nearest_point(self, point, count=1, radius=100):
    sql = """
    SELECT
      ST_X(ST_ClosestPoint(ST_Transform(r.way, 4326), p.geom)) as x,
      ST_Y(ST_ClosestPoint(ST_Transform(r.way, 4326), p.geom)) as y,
      ST_Distance(ST_Transform(r.way, 4326), geography(p.geom)) as distance,
      ST_ClosestPoint(ST_Transform(r.way, 4326), p.geom) as point,
      r.*
    FROM
      planet_osm_point r,
      (Select ST_SetSRID(ST_MakePoint(%f, %f), 4326) as geom) p
    WHERE
      ST_DWithin(r.way, ST_Transform(p.geom, 3857), %d)
    ORDER BY distance ASC
    LIMIT %d
    """
    gdf = self._conn.get_gdf(sql % (point[0], point[1], radius, count))
    return gdf
```

図 6 周辺地物を求める PostGIS SQL 質問

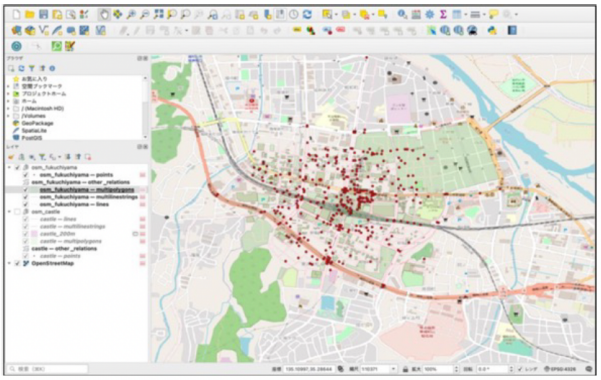


図 7 JR 福知山駅周辺地物データの QGIS による表示

我々の OSM データベース(PostgreSQL テーブル) は QGIS という GIS に接続されている．図 7 は,「JR 福知山駅」周辺の地物を GQIS で表示したものである.

与えられた地物 g の意味（分散表現）を獲得するために，まず，地物 g の周辺の地物群 surrounding(g)を OSM データベースから抽出し，これをスキップグラムニューラルネット GeoObject2Vec により学習している．図 8 は，地物「福知山城」から 200m 以内にある地物を周辺地物群とし,これらを点オブジェクト(赤い点)と線オブジェクト（赤い線）として表示したものである．地物「福知山城」の周辺地物群を GeoObject2Vec で機械学習し，結果として「福知山城」の意味の分散表現を中間層のニューロン群の値から特徴ベクトルとして獲得している（図 9a 参照）． さらに，図 9 b は，出力層のニューロン群の値に基づいて「福知山城」の周辺地物の型情報を出力したものである．図 9c は,「福知山城」という地物と意味類似する地物を検索した結果である．意味類似度計算にはコサイン類似度を用いた．検索結果として,「福知山城」という地物と意味類似性が高い地物として,「盛岡城」や「勝竜寺城」などが得られている．図 9d には,「福知山城」と「盛岡城」の周辺領域を地図で示している.



図 8 福知山城の周囲 200m 以内にある周辺地物

```
array([  8.45028387, -21.13814466,   6.99195111,  18.43155993,
        -6.19986398,   9.31108859,   1.36405781,  -8.2373122 ,
        13.02728328, -13.22325764,   9.20188256, -12.12637051,
        -1.24983991,  -2.98366649,  -2.77713985,   5.05415158,
        -2.48321692,   5.68050077,   9.87510547,  17.48195665,
        16.32950259, -14.51119441,  -0.03137228,   4.90272109,
         6.14963845,   2.82771183,  -6.41255527, -20.17267944,
        -1.97539554,   2.78932904,  -5.4148039 ,  -5.11709611,
       -10.35781185, -20.16748463,  -5.83101174,  -4.21146201,
         1.02811328,  -7.98243418,   6.62720716,   1.2392528 ,
         0.64635427,  -3.86657805,  10.11649007,   0.10953928,
        -4.1317976 , -10.79041161,  24.69978435,  -5.5879894 ,
       -21.97352476,   1.97990154,   1.68202317,  -3.04898263,
        23.93431716,  -1.65810417, -23.03438559,  17.40852135,
        -1.44518629,  -5.27317333, -13.00001908,  -4.09303104,
        -3.30299426,   4.71787828,  -4.22420044,   1.86497441,
        18.55928731, -11.52573497,  -6.14882185, -10.82826594,
       -10.54949354, -10.9092413 ,  10.4565337 , -13.58708818,
        19.18154747,   6.51453107,  20.63472465,   2.0898572 ,
        -0.54682947, -11.12133011,  10.15021182,  10.04874766,
        17.44821983, -11.19583907, -15.08238195, -19.70231583,
        -2.30769606,  20.55874455,  20.93087021, -13.57425616,
       -14.07007781,   8.81251796,  -4.16368885,  -1.26061814,
        -5.62064252,  21.62905189,  -9.84528257, -12.16809983,
        11.19391325,   8.69500426,   3.42264825,   4.23945933])
```

図 9a　「福知山城」の分散表現

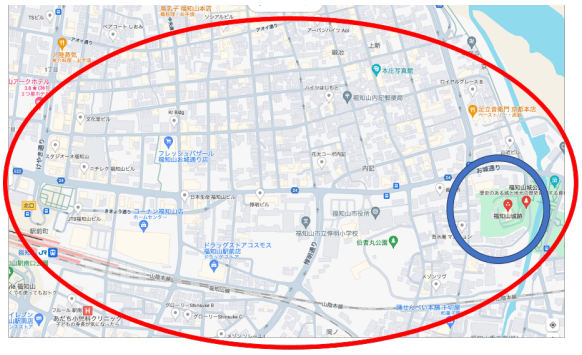
```
['historic=castle', 'man_made=water_tap', 'natural=tree', 'historic=memorial', 'man_made=water_well', 'natural=tree', 'natural=tree', 'natural=tree', 'tourism=information', 'natural=tree', 'amenity=bench', 'historic=memorial', 'tourism=information', 'tourism=information', 'historic=memorial', 'natural=tree', 'tourism=information', 'man_made=water_well', 'tourism=information', 'natural=tree', 'historic=memorial', 'man_made=water_tap', 'historic=memorial', 'historic=memorial', 'historic=memorial', 'tourism=information', 'natural=tree', 'historic=memorial', 'amenity=bench', 'natural=tree', 'amenity=bench', 'amenity=bench', 'amenity=vending_machine', 'amenity=bench', 'historic=memorial', 'historic=memorial', 'natural=tree', 'amenity=bench', 'tourism=information', 'amenity=bench', 'tourism=information', 'natural=tree', 'historic=memorial', 'tourism=information', 'natural=tree', 'tourism=information', 'tourism=information', 'amenity=bench', 'amenity=bench', 'tourism=information', 'amenity=bench', 'amenity=bench', 'barrier=cycle_barrier', 'bicycle=no', 'tourism=museum', 'highway=crossing', 'highway=traffic_signals', 'highway=traffic_signals', 'amenity=courthouse', 'amenity=toilets', 'historic=memorial', 'amenity=courthouse', 'historic=memorial', 'amenity=kindergarten', 'amenity=place_of_worship', 'amenity=restaurant', 'highway=traffic_signals', 'amenity=courthouse', 'amenity=cafe', 'historic=memorial', 'amenity=restaurant']
```

図 9b　「福知山城」の周辺地物群の型情報

```
[array([1.]),
 array([0.96879956]),
 array([0.94048475]),
 array([0.94048475]),
 array([0.93496157]),
 array([0.92960452]),
 array([0.92924144]),
 array([0.92170747]),
 array([0.92076402]),
 array([0.91972403])]
```

```
[0    福知山城
Name: name, dtype: object, 0    盛岡城 (Morioka-jō)
Name: name, dtype: object, 0    勝竜寺城
Name: name, dtype: object, 0    勝竜寺城公園
Name: name, dtype: object, 0    岡崎城
Name: name, dtype: object, 0    江戸城
Name: name, dtype: object, 0    丸岡城
Name: name, dtype: object, 0    笠間城
Name: name, dtype: object, 0    福井城
Name: name, dtype: object, 0    和歌山城
Name: name, dtype: object]
```

図 9c　「福知山城」と意味類似する地物の検索



福知山城と周辺地物



盛岡城と周辺地物

図 9d　福知山城/盛岡城とその周辺地物群

OSM データベースを用いた地物の周辺地物群の機械学習，および，意味類似する地物の検索は，一般的には，次の手順で行っている．

① **周辺地物群データの抽出**
与えられた地物 g に対して，3 種類の周辺地物群データ（点オブジェクト，線オブジェクト，多角形オブジェクト）を OSM データベースから検索する．検索された OSM データは，地物名，座標，種々の属性データから構成される．

② **GeoObject2Vec アルゴリズムによる機械学習**
スキップグラムニューラルネットアルゴリズム GeoObject2Vec を用いて，各地物の意味（分散表現）を得る．このニューラルネットワークの出力層の各ニューロンは周辺地物の 694 個の「型」に相当するものである．

③ **地物 g の意味（分散表現）の獲得**
個々の地物 g の意味（分散表現）は GeoObject2Vec ネットワークの中間層にあるニューロンの値のベクトル（特徴ベクトル）として得ることができる．

④ **意味類似する地物の検索**
対象地物の分散表現と他の地物の分散表現のコサイン類似度を求め，コサイン類似度が大きな地物を意味類似する地物として出力する．

4. 地理情報検索による地物の意味獲得

4.1. 地物の意味の言語表現

前節で述べた地物の意味獲得では，地物の「意味」は，中間層のニューロンの値からなる特徴ベクトルで表されるものであった．すなわち，地物の「意味」の分散表現を得たわけであるが，このような分散表現は人間にとっては理解しにくいものであった．そこで，本節では，地物の「意味」の言語表現を地理情報検索で得る手法を提案する．例えば，大学近辺にあるカフェに対して，提案する手法を用いると，そのカフェの意味の言語表現として，「学生街にあるカフェ」といった言語表現を得ることができる．

地物 g に対して，その地物の型を t，その地物の周辺領域名 r とすると，地物 g の意味の言語的表現は以下の 3 つ組（g, t, r）で表すことができる：

(g, t, r) = (地物 g，地物 g の型名 t，
地物 g の周辺地物を含む領域名 r).

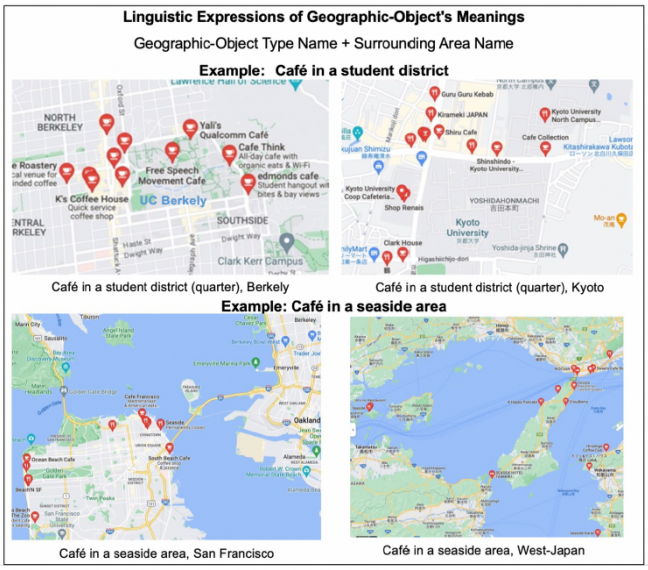


図 10 地物の意味の言語表現例
「学生街のカフェ」と「海辺のカフェ」

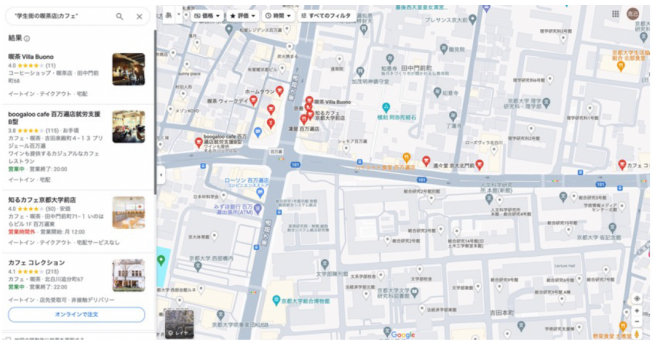


図 11 「学生街のカフェ」の GoogleMap 検索結果

図 10 は，地物の意味の言語表現の例である．「学生街のカフェ（“café in a student district”）」，「海辺のカフェ（“café in seaside area”）」という言語表現が，GoogleMap 検索で得られている． 例えば，カフェ g が京都大学周辺にある「進々堂（Shinshindo）」であるとすると， g の言語表現として， (“Shinshindo”, Cafe, “student quarter”), すなわち，「学生街のカフェ（”a café in a student quarter”）が得られている．同様に，「海辺のカフェ g’」をバークレー周辺で検索で求めることができている．すなわち，地物 g’（”Bayside Café “）の言語表現として，

(“Bayside Café”, Cafe, “seaside area”)

を得ることができる．

図 11 は，「学生街のカフェ」を GoogleMap 検索で求めようとしている画面を示している．



図 12 「商店街にある神社，かつ，駅近の神社」
（京都，錦天満宮）

もう少し複雑な検索例として，京都市内にある神社「錦天満宮」を考えよう．この神社は，「商店街にある神社」という意味と，「鉄道駅に近いエリアにある神社」という意味を併せて持っている．図 12 に，この「錦天満宮」がある商店街（新京極通り）と最寄りの鉄道駅（阪急四条河原町駅）を示す．この例のような場合は，神社「錦天満宮」の言語的な意味は，以下のようにすることが適当と考えられる．

(“錦天満宮”, 神社, “商店街”)

(“錦天満宮”, 神社, ”鉄道駅近く”)

すなわち，錦天満宮の言語的な意味は，「商店街にあり，鉄道駅近くにある神社」とすることが妥当と考えられる．

4.2. 地理情報検索による地物の意味の言語表現の獲得

地物の意味の言語表現を地理情報検索で求めるために，言語表現の構文を次のように設計した．すなわち，地物の意味の言語表現の構文を以下のように定義した．

```
<地物意味> ::=      <周辺領域> <地物種別>
<周辺領域> ::=      <周辺領域名> | <領域修飾句> <周辺領域名>
<周辺領域名> ::=    <街区名> | <近接領域名>
<街区名> ::=        学生街|商店街|繁華街|住宅街|オフィス街|電気街|
                    中華街|温泉街|ビジネス街|田園地帯|過疎地|…
<近接領域名>::=     海辺|川沿い|川縁|湖畔|駅前|駅近| 路地裏| …
<地物種別>::=       <地物種別名>|<地物修飾句><地物種別名>
<地物種別名>::=     喫茶店|カフェ|レストラン|食堂|スーパーマーケット|
                    マンション|アパート|公園|学校|クリニック| …
<領域修飾句>::=     閑静な|賑やかな|緑豊かな|野中の| …
<地物修飾句>::=     おしゃれな| 隠れ家|小さな| 本が読める|
                    音楽が聴ける| …
```

地物の意味の言語表現は，上記の構文で定義されるクエリを実行することで得ることができる．一般的には以下の手順で言語表現を得る．

① 検索エリアを決定し，地物の意味の言語表現をクエリとして実行

② 検索結果の各 Web ページが該当する言語表現を含んでいるものを正解とし，地物オブジェクトの言語表現とする．

5. 機械学習と地理情報検索の相補的利用

前節までに，地物の意味を獲得するための 2 つの独立

した方法を提案した．一つ目の方法は，GeoObject2Vec アルゴリズムによって地物の意味の分散表現（特徴ベクトル）を求めるものであり，二つ目の方法は，地理情報検索によって，地物の言語的な意味を求めるものであった．これらの独立した方法には，それぞれ以下のような長所･欠点が存在する．

- GeoObject2Vec による方法
  GeoObject2Vec では，対象とする地物の周辺にある地物群を学習しているため，周辺領域の意味も正確に計算できているところが長所である．一方，この方法で日本国内の全ての地物の意味を計算するには長い計算時間を要すること，および，計算された意味が分散表現であるため人間にとっては理解しにくいものとなってしまうことが欠点である．
- 地理情報検索による方法
  地理情報検索によって地物の意味の言語表現を直接得ることができる点が最大の長所である．欠点は，地理情報検索の精度（特に再現率）である

と考えられる．例えばある地物は，言語的な意味としては「海辺のカフェ」であった場合でも，その意味記述がその地物のホームページ等に陽に記載されていない場合もある．このような場合は本手法の再現率が低下することとなり今後の問題となる．

両手法の課題については上記したとおりであるが，これらの課題を改善していく方法として，図 13 に示すような，両手法を相補的に実行し精度を改善する方法が考えられる．この方法は，図 13 に示すように，すでに言語表現を持ち，特徴ベクトルが類似している地物の言語表現を継承させるものであり，もし継承させるべき地物が見つからない場合は，周辺地物の定義を修正し再学習を行うものである．この相補的利用の手順を以下に示す．

① すべての地物に対して意味の分散表現を GeoObject2Vec を用いて求める．
② いくつかの地物に対して，その意味の言語表現を地理情報検索で求める．

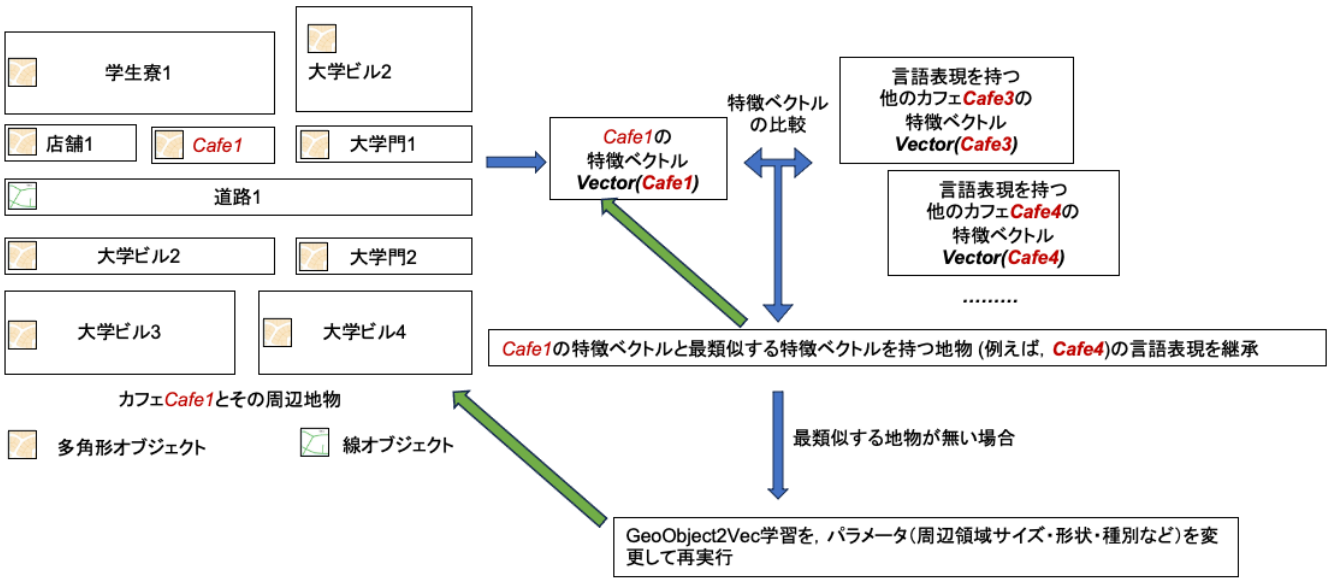


図 13 機械学習と地理情報検索の相補的利用

③ 対象地物の意味（分散表現）と言語表現を持つ地物の意味（分散表現）を比較し最類似の地物が見つかればその地物の意味の言語表現を対象地物に継承させる． 言語表現を持つ最類似の地物が見つからない場合は，GeoObject2Vec アルゴリズムの「周辺領域」の定義（サイズ，形状，種別）を変更して再学習を行う．

6. まとめ

本論文では，地理情報システム(GIS）における地物の意味を発見する手法を提案した．地物の意味は，対象地物の種別と対象地物の周辺領域にある地物群で定義される．この地物の意味を発見するために 2 つの独立した方法を提案した．1 つ目の方法は，Word2Vec を 2 次元的に拡張した GeoObject2Vec という機械学習法に基づくもので，これによって地物の意味の分散表現を得ることができる．2 つ目の方法は，地理情報検索に基づいて地物の意味の言語表現を得るものである．こ

れら 2 つの手法を相補的に適用することで，地物の意味を発見することを目指した．

参 考 文 献

[1] [Harris, 1954] Harris, Z. S.: Distributional structure. WORD, 10(2-3):146–162, 1954.

[2] [Lavallin and Downs, 2021] Lavallin, A. and Downs, J. A.: Machine learning in geography – past, present, and future. Geography Compass, 15(5) April 2021.

[3] [Le and Mikolov, 2014] Le, Q. and Mikolov,T.: Distributed representations of sen-tences and documents. In Proc. of 31st Int ’l. Conf. on Machine Learning, pages 1188–1196, 2014.

[4] [Li, 2021] Li, W.: GeoAI and deep learning. International Encyclopedia of Geogra-phy: People, the Earth, Environment and Technology, September 2021.

[5] [Mai et al., 2022] Mai, G., Hu, Y., Gao, S., Cai, L., Martins, B., Scholz, J., Gao, J., and Janowicz, K. Symbolic and subsymbolic geoai: Geos-patial knowledge graphs and spatially explicit machine learn-ing. Transactions in GIS, December 2022.

[6] [Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and De-an, J.: Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, Advances in Neural In-formation Processing Systems 26 (NIPS 2013), volume 26. Curran Associates, Inc. 2013.

[7] [Usery et al., 2021] Usery, E. L., Arundel, S. T., Shavers, E., Stanislawski, L., Thiem, P., and Varanka, D.: GeoAI in the us geological survey for topographic map-ping. Transactions in GIS, 26(1), August 2021.

[8] [Zhang et al., 2016] Zhang, Y., Jatowt, A., Bhowmick, S. S., and Tanaka, K.: The past is not a foreign country: Detecting semantically similar terms across time. IEEE Transactions on Knowledge and Data Engineering, 28(10):2793–2807, 2016.

[9] [Zhang et al., 2017] Zhang, Y., Jatowt, A., and Tanaka, K.: Is tofu the cheese of asia? searching for corresponding objects across geographical areas. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 1033–1042, 2017.

[10] [OSMF] OpenStreetMap Foundation: OpenStreetMap Licenced Under the Open Data Commons Open Database License (ODbL)

[11] [PostGIS] PostGIS PSC & OSGeo: About PostGIS, http://postgis.net

交通計測地点の増加に対応する知識転移を用いた時空間グラフモデリング

川端　智紀†　　戸田　浩之†

† 横浜市立大学データサイエンス学部データサイエンス学科　〒 236–0027 神奈川県横浜市金沢区瀬戸 22–2
E-mail: †{d204018d,toda.hir.xg}@yokohama-cu.ac.jp

あらまし　過去の交通状態（例：交通量・交通速度・移動時間）に基づいて未来の交通状態を予測する交通予測は，交通管理や交通資源配分の効率化に有用である．交通データは，ネットワーク構造をもつ道路上の計測地点で観測される．さらに，交通状態は複雑な時空間依存関係をもつ．そのため，交通予測に，グラフニューラルネットワークと時系列学習手法を組み合わせた時空間グラフニューラルネットワークモデルを用いる研究が多く存在し，優れた結果を示している．しかし，他の深層学習モデルと同様，学習には多くのデータを必要とする．また，計測地点の位置や数は時間とともに変化するが，既存のモデルは計測地点の増加には対処できず再学習が必要となり，さらに増加直後は十分にデータが得られないという問題がある．そこで我々は，データから多くの知識を抽出・活用できるモデルを構築し，計測地点増加前のデータで学習したモデルのパラメータを増加後のデータで学習する際に事前知識として用いることで，この問題を克服する．我々の知る限り，同一都市間の転移に注目した研究は本研究が初めてである．計測地点が増加する実世界交通データセットを用いて，提案モデルの評価を行う．

キーワード　モビリティ，空間・時空間・時系列データ処理，交通予測，グラフニューラルネットワーク，転移学習

1　序　　論

現在，道路上に設置される交通センサや GPS 搭載機器，交通系 IC カードなどの普及により，人や車両の移動に関する交通データの収集が可能となっている．収集されたデータを用い，交通事故検出や交通需要予測，人流・交通流の予測など多様な研究が行われている [1]．これらの研究の方向性の 1 つに，過去の交通状態（例：交通量・交通速度・移動時間）に基づいて未来の交通状態を予測する交通予測がある [1]．交通予測は，交通管理の効率化やマップアプリケーションのサービス向上に有用である．例えば，正確な交通量予測は交通信号制御や渋滞情報の提供等を通じて混雑の緩和に役立てることができ，正確な移動時間予測はルートナビゲーションサービスにおいて効率的な経路の提案等を通じてユーザー体験を向上させることができる．

都市における複数の計測地点の交通状態を予測する場合，交通データを観測する計測地点は，道路ネットワークや鉄道路線等により相互に接続されており，計測地点間の相互関係をグラフ構造として表現することも可能である．さらに，交通データは，複雑な時間依存関係および空間依存関係を持つ．

上記の交通データが持つ特性のため，都市における交通予測に関する研究分野では，時空間グラフニューラルネットワーク（時空間 GNN）というアーキテクチャを用いる研究が多く存在する．時空間 GNN とは，リカレントニューラルネットワーク（RNN）や時間畳み込みネットワーク（TCN）などの時系列学習手法とグラフニューラルネットワーク（GNN）を組み合わせたアーキテクチャである．時系列学習手法が複雑な時間依存関係を捉えることができ，GNN が計測地点間の関係を考慮した複雑な空間依存関係を捉えることができるため，時空間 GNN を用いたモデルは優れた予測精度を示している [1].

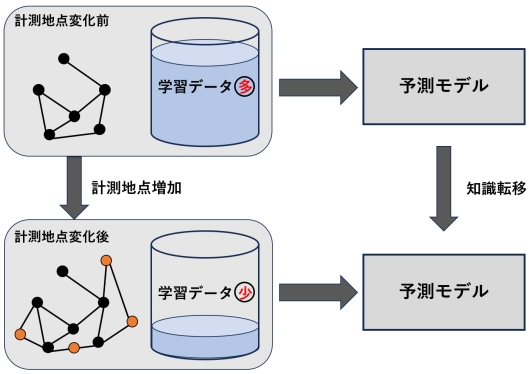


図 1: 計測地点数の変化と転移学習

一方で，都市における交通計測地点数は時間とともに変化するが，変化前のデータを用いて学習した時空間 GNN モデルを，変化後のデータの予測にそのまま活用することは難しい．何故ならば，計測地点数の変化に伴い，時空間 GNN で考慮する計測地点間の関係を表すグラフ構造も変化するからである．グラフ構造の変化により GNN の出力も変化するため，GNN の出力に影響されるモデルパラメータは再学習することが望ましい．

また，時空間 GNN を用いたモデルも，他の深層学習モデルと同様に，学習には多くのデータを必要とする．学習に用いるデータ量が十分でない場合，モデルの学習が十分に行えず高い予測精度を実現することが困難となる．

そのため，計測地点増加した場合，再学習が必要になるが，十分な量の学習データが得られず，予測モデルの学習が十分に行えないという問題に直面すると考えられる．

本研究の目的は，上記の問題の対処に適したモデルを構築し，転移学習を用いることで，上記の問題を克服することである．

具体的には，計測地点増加前のデータで学習したモデルのパラメータを，増加後のデータで学習する際に事前知識として転移しパラメータの初期値として用いる．

本研究の主な貢献は，以下の点に要約される．

- 交通予測問題において，学習データのサイズと時空間 GNN モデルの予測精度との関係を明らかにした点．
- データから多くの知識を抽出・活用できる新たな時空間 GNN モデルを提案した点．提案モデルは，計測地点増加前のデータで学習することで，増加後のデータの学習に有用な知識を得られる．得られた知識を活用することで，計測地点増加後の再学習コストの低減，学習データ量の不足による予測精度劣化の問題の克服することができる．
- 計測地点数の変化を含む交通予測問題を評価するための，新しい実世界交通データセットを構築した点．さらに，構築したデータセットを用いて，最先端手法と提案モデルを評価し，提案モデルの有効性を明らかにした点．

以下に，本論文の構成を示す．第 2 節で，関連研究について説明する．第 3 節で，用語と数学的表記を説明した後，問題定義について述べる．第 4 節で，本研究で用いるデータセット，ベースライン，評価手法について説明する．第 5 節で，予備分析として学習データ量がモデルの予測精度に与える影響を調べる．第 6 節で，提案手法について説明する．第 7 節で，実験を行い提案手法を評価する．最後に，第 8 節で結論を述べる．

2 関 連 研 究

2.1 グラフニューラルネットワークを用いた交通予測

GNN は，交通分野で頻繁に用いられる非ユークリッド構造データに隠された空間情報を捉える強力な学習能力を持っており [2]，交通ネットワークをグリッドとしてモデリングした研究と比較して，グラフとしてモデリングした研究は空間情報を十分に活用できている [1]．そのため，GNN と時系列学習手法を組み合わせた時空間 GNN を交通予測に適用した研究は多く存在する [1] [2] [3].

交通予測に用いられる時空間 GNN において，扱うグラフ構造は隣接行列で表現され，交通ネットワークのトポロジー構造や計測地点間の距離をベースとして事前に定義されたものを用いる場合 [4] [5] と，学習可能なパラメータやモデルの隠れ状態を用いてデータから学習されたものを用いる場合 [6] [7] [8] がある．グラフ構造をデータから学習することをグラフ構造学習といい，予測を容易にする最適なグラフ構造を見つけることを目的として行われる．GWN（Graph WaveNet）[6] は，ランダムに初期された学習可能な行列を介して，自己適応型隣接行列を構築する．MegaCRN（Meta-Graph Convolutional Recurrent Network）[7] は，memory network [9] のアイデアを活用することで，グラフ構造学習を強化する．また，計測地点数が変化したとき，事前に定義したものを用いる場合では再定義する必要があり，データから学習したものを用いる場合では再学習する必要がある．

2.2 転 移 学 習

転移学習は，すでに学習された関連するタスクから知識の転移を通じて，解決したいタスクの学習を改善する方法であり [10]，機械学習において学習データ不足という問題を解決するための重要な手法である [11]．転移学習の中でも，モデルのパラメータやアーキテクチャを転移する手法は，モデルベース転移学習と呼ばれる [10]．深層学習分野におけるモデルベース転移学習に関する研究の例に，ImageNet データセットで学習した畳み込みニューラルネットワークの層を，他の視覚認識タスクで再利用する研究がある [12].

また，交通分野における時空間 GNN に関するサーベイ論文 [1] [2] [3] においても，将来の研究の方向性の一つとして，GNN と転移学習の組み合わせが述べられている．交通予測問題において，時空間 GNN と転移学習を組み合わせた手法についての既存研究 [13] [14] は，異なる都市間の知識転移に注目している．一方で，我々の知る限り，ある一つの都市において変化する交通ネットワーク上の交通状態を予測するために，過去のデータを効果的に活用することを目的として同一都市間の知識転移に注目した研究は存在しない．

3 問 題 設 定

3.1 用語と数学的表記

タイムステップ t における交通状態は，$\mathbf{X}_t \in \mathbb{R}^{N\times C}$ と表記する．ここで，N は交通状態を観測する計測地点の数，C は関心のある交通状態の数である．

交通状態 $\mathbf{X}_t$ が計測される交通ネットワークを，グラフ $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ と表記する．ここで，$\mathcal{V}$ は計測地点を表す $N = |\mathcal{V}|$ 個のノードの集合，$\mathcal{E}$ はノード間の接続性を表すエッジの集合，$\mathbf{A} \in \mathbb{R}^{N\times N}$ は隣接行列である．本研究では，事前に定義する隣接行列として，多くの先行研究 [4] [5] [6] で用いられている距離ベースの行列

$$A_{ij} = \begin{cases} \exp\left(-\frac{\mathrm{dist}(v_i,v_j)^2}{\sigma^2}\right), & \text{if } \mathrm{dist}(v_i,v_j) < \kappa \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

を用いる．ここで，$\mathrm{dist}(v_i, v_j)$ は交通ネットワーク上で測定される距離，σ と κ はハイパーパラメータである．

解決したいタスクをターゲットタスク $\mathcal{T}_{\text{target}}$，補助的に用いるタスクをソースタスク $\mathcal{T}_{\text{source}}$ と呼ぶ．また，それぞれのタスクが存在するドメインをターゲットドメイン $\mathcal{D}_{\text{target}}$，ソースドメイン $\mathcal{D}_{\text{source}}$ と呼ぶ．

3.2 問 題 定 義

本研究で扱う交通予測問題を，α 個の過去のタイムステップの交通状態から β 個の未来のタイムステップの交通状態を予測する問題と定義する．この問題は，以下の式 (2) のように，パラメータ $\boldsymbol{\Theta}$ を持つ関数 $f(\cdot)$ の学習として定式化できる

$$[\mathbf{X}_{t-(\alpha-1)}, \ldots, \mathbf{X}_t] \xrightarrow[\boldsymbol{\Theta}]{f(\cdot)} [\mathbf{X}_{t+1}, \ldots, \mathbf{X}_{t+\beta}]. \tag{2}$$

事前に定義したグラフ構造を予測に用いる場合，グラフ $\mathcal{G}$ も入

力となるため，以下の式 (3) となる

$$\left[\mathbf{X}_{t-(\alpha-1)},\ldots,\mathbf{X}_t;\mathcal{G}\right] \xrightarrow[\boldsymbol{\Theta}]{f(\cdot)} \left[\mathbf{X}_{t+1},\ldots,\mathbf{X}_{t+\beta}\right]. \quad (3)$$

また，転移学習は，$\mathcal{D}_{\text{source}}, \mathcal{T}_{\text{source}}, \mathcal{D}_{\text{target}}, \mathcal{T}_{\text{target}}$ が与えられたとき，$\mathcal{D}_{\text{source}}$ と $\mathcal{T}_{\text{source}}$ の知識を用いて，$\mathcal{D}_{\text{target}}$ のターゲット予測関数 $f_{\text{target}}(\cdot)$ の学習を改善することを目的としたものとして定義される [15]．そのため，モデルベース転移学習と交通予測タスクの組み合わせは，

$$\left[\mathbf{X}^{\text{target}}_{t-(\alpha-1)},\ldots,\mathbf{X}^{\text{target}}_t;\boldsymbol{\Theta}'_{\text{source}}\right] \xrightarrow[\boldsymbol{\Theta}_{\text{target}}]{f_{\text{target}}(\cdot)} \left[\mathbf{X}^{\text{target}}_{t+1},\ldots,\mathbf{X}^{\text{target}}_{t+\beta}\right] \quad (4)$$

として定式化できる．ここで，$\mathbf{X}^{\text{target}}_t$ はターゲットドメイン $\mathcal{D}_{\text{target}}$ におけるタイムステップ t の交通状態，$\boldsymbol{\Theta}_{\text{target}}$ は $\mathcal{D}_{\text{target}}$ で学習される関数 $f_{\text{target}}(\cdot)$ のパラメータである．$\boldsymbol{\Theta}'_{\text{source}}$ は，ソースドメインで学習した関数 $f_{\text{source}}(\cdot)$ のパラメータ $\boldsymbol{\Theta}_{\text{source}}$ の内，$\mathcal{D}_{\text{target}}$ における学習において事前知識として用いるパラメータであり，$\boldsymbol{\Theta}'_{\text{source}} \subset \boldsymbol{\Theta}_{\text{source}}$ である．$\boldsymbol{\Theta}_{\text{source}}$ は，

$$\left[\mathbf{X}^{\text{source}}_{t-(\alpha-1)},\ldots,\mathbf{X}^{\text{source}}_t\right] \xrightarrow[\boldsymbol{\Theta}_{\text{source}}]{f_{\text{source}}(\cdot)} \left[\mathbf{X}^{\text{source}}_{t+1},\ldots,\mathbf{X}^{\text{source}}_{t+\beta}\right] \quad (5)$$

で学習される．予測に事前に定義したグラフ構造を用いる場合，式 (3) と同様に，式 (4)，式 (5) に，それぞれのドメインにおけるグラフ構造 $\mathcal{G}^{\text{target}}$ と $\mathcal{G}^{\text{source}}$ を付け加え，定式化できる．

4 準　　備

4.1 データセット及び前処理

交通状態に関するデータとして，東京の交通量データとロサンゼルスの交通速度データを用いる．東京の交通量データとして，公益財団法人日本道路交通情報センター提供の断面交通量情報 [1]を用いる．元のデータは計測地点で収集された交通量データを 5 分間隔で集計したものであるが，本研究では欠損値を線形補間した後，15 分間隔に再集計し Z スコア正規化をして用いる．使用するデータの範囲は，東京都千代田区・中央区・港区とし，2019 年 5 月 1 日から 2019 年 10 月 31 日までのソースデータセットと 2019 年 11 月 1 日から 2020 年 2 月 7 日までのターゲットデータセットを構築する．データセットを構築する基準として，期間内の各月の欠損値が 3 割未満の計測地点のデータと定義する．この定義に従うと，計測地点の数はソースデータセットでは 151，ターゲットデータセットでは 343 と大きく変化する．ソースデータセットでは，訓練データ，検証データ，テストデータの比を 7 : 1 : 2 に設定する．ターゲットデータセットでは，テストデータを 2020 年 2 月 1 日から 2 月 7 日の 1 週間に固定し，訓練データと検証データの合計を 2019 年 11 月 1 日から 1 週間，2 週間，1 か月，2 か月，3 か月としたデータセットを構築し，訓練データと検証データの比を 7 : 1 に設定する．ロサンゼルスの交通速度データとして，交通予測のベンチマークデータセット METR-LA[2]を用いる．Z スコア正規化をし，訓練データ，検証データ，テストデータの比を 7 : 1 : 2 に設定する．

使用するデータセットの概要について，表 1 にまとめた．

表 1: データセット概要

データセット名	計測地点数	訓練・検証データ量	テストデータ量
source	151	14126	3531
1week	343	669	668
2week	343	1341	668
1month	343	2877	668
2month	343	5853	668
3month	343	8829	668
METR-LA	207	27399	6850

また，事前定義する隣接行列を作成するために，(公財) 日本交通管理技術協会提供の交通量計測地点位置情報データ [3]と，OpenStreetMap[4]の道路ネットワークデータを用いる．

4.2 ベースライン

GWN [6]． 拡散グラフ畳み込みと WaveNet を組み合わせたモデル．事前定義したグラフ構造を表現した隣接行列に加えて，データから未知のグラフ構造を発見するために自己適応型隣接行列を構築し，グラフ畳み込みに用いる．

GWN$-\mathcal{G}$． GWN から事前定義した隣接行列を除き，グラフ畳み込みに自己適応型隣接行列を用いるモデル．

MegaCRN [7]． メモリネットワーク [9] のアイデアを活用したメモリを備えたメタグラフ学習器と，グラフ畳み込みリカレントユニットエンコーダ・デコーダを組み合わせた，時空間メタ学習のためのフレームワーク．メモリでノードレベルのプロトタイプを発見し，メモリに合わせたノード埋め込みを構築することで，時空間グラフ学習を補強する．

4.3 評 価 手 法

本研究では既存研究 [4] [5] [6] [7] [16] に倣い，未来の 1 時間分の交通状態を予測対象とし，その中の 15 分先，30 分先，60 分先の値の予測値でモデルの評価を行う．すなわち，本研究では 15 分間隔のデータを用いるため，$t+1$，$t+2$，$t+4$ の予測値でモデルの評価を行う．また，評価指標として，平均絶対誤差（MAE），二乗平均平方根誤差（RMSE），平均絶対パーセント誤差（MAPE）を用いる．

5 予 備 分 析

本節では，第 4 節で説明したデータセット 1week，2week，1month，2month，3month・ベースライン・評価手法を用いて，学習データ量が時空間 GNN 予測モデルの精度に与える影響について分析する．用いるデータセットは，テストデータの期間・量，データセットに含まれる計測地点が同一で，異なるのは学習データの期間・量のみである．結果を，表 2 に示す．

結果より，すべてのベースラインで学習データ量の増加に伴い，基本的に予測精度が改善していることが分かる．また，予測精度の改善量は，基本的に 15 分先，30 分先，60 分先の順に多くなる．以上より，学習データ量は予測精度に影響し，特に予測対象が時間的に離れているほど影響は大きいことが分かる．

次に，MAE に着目し，事前定義したグラフ構造の利用の有無のみが異なる GWN と GWN$-\mathcal{G}$ の比較と，事前定義したグ

1：https://www.jartic.or.jp/
2：https://paperswithcode.com/dataset/metr-la
3：https://www.tmt.or.jp/research/index9.html
4：https://www.openstreetmap.org

表 2: 予備分析結果

データセット	モデル	15 分先予測			30 分先予測			60 分先予測		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
1week	GWN	15.2197	23.5599	13.08%	16.5380	25.8319	14.58%	**19.1328**	**29.8699**	16.69%
	GWN−$\mathcal{G}$	15.2011	**23.3706**	13.06%	16.4603	**25.7494**	14.21%	19.2052	29.9540	16.49%
	MegaCRN	**15.0620**	27.6001	**12.88%**	**16.4436**	26.1918	**14.03%**	19.4867	31.2840	**16.06%**
2week	GWN	**14.8859**	23.3854	12.69%	**15.9146**	**25.4068**	13.75%	17.9987	**28.8577**	15.97%
	GWN−$\mathcal{G}$	14.9265	23.5014	**12.42%**	15.9691	25.7726	**13.28%**	**17.9352**	29.3508	**14.78%**
	MegaCRN	14.8954	**23.3478**	12.78%	16.0344	25.9041	13.90%	18.3591	29.9284	15.76%
1month	GWN	**14.4891**	22.4655	12.22%	15.4161	24.5596	**12.99%**	**16.9271**	27.6057	**14.23%**
	GWN−$\mathcal{G}$	14.4906	**22.3486**	**12.21%**	**15.3432**	**24.2195**	13.21%	17.0127	**27.3574**	14.32%
	MegaCRN	14.5059	22.6002	12.32%	15.4590	24.8554	13.31%	17.1354	27.9721	14.34%
2month	GWN	14.1769	**21.9515**	12.43%	**14.7838**	**23.5547**	12.75%	**15.9192**	**25.8450**	13.94%
	GWN−$\mathcal{G}$	14.2591	22.2209	**11.97%**	14.9490	23.8476	12.74%	16.1898	26.2680	13.87%
	MegaCRN	**14.0979**	21.9805	**11.97%**	14.8913	23.8727	**12.70%**	16.2728	26.2617	**13.79%**
3month	GWN	13.9784	**21.6694**	12.03%	14.5514	**23.2113**	12.63%	15.4376	**25.1651**	13.26%
	GWN−$\mathcal{G}$	14.1533	21.9186	12.02%	14.7882	25.5301	12.69%	15.8909	25.7274	13.57%
	MegaCRN	**13.9208**	21.7602	**11.86%**	**14.5402**	23.4044	**12.46%**	**15.5142**	25.5363	**13.23%**

ラフ構造を用いない GWN−$\mathcal{G}$ と MegaCRN の比較を行う．

GWN と GWN−$\mathcal{G}$ の比較 データセット 1week から 2week，2week から 1month では，GWN−$\mathcal{G}$ は事前定義したグラフ構造を用いていないにもかかわらず，GWN と類似した性能の向上を示す．一方で，データセット 1month から 2month，2month から 3month では，GWN−$\mathcal{G}$ の性能の向上は，GWN よりも鈍化する．これは，GWN が事前定義したグラフ構造をモデルの入力に用いるため，GWN−$\mathcal{G}$ と比べて学習可能なパラメータ数が多いことが理由だと考える．そのため，事前定義したグラフ構造を入力に用いる場合，入力に用いない場合と比較して，より学習データの量が重要であると考える．

GWN−$\mathcal{G}$ と MegaCRN の比較 データセット 1week から 2week，2week から 1month では，GWN−$\mathcal{G}$ は，MegaCRN と類似した性能の向上を示す．データセット 1month から 2month において，15 分予測・30 分予測では GWN−$\mathcal{G}$ の性能の向上は MegaCRN よりも鈍化するが，60 分先予測では GWN−$\mathcal{G}$ は MegaCRN と類似した性能の向上を示す．データセット 1month から 2month において，すべての予測で GWN−$\mathcal{G}$ の性能の向上は MegaCRN よりも鈍化する．データセットが増えるほど，MegaCRN の方が性能が良くなるのは，学習データからグラフ構造を構築するのに必要なパラメータ数が多いことが原因であると考える．そのため，MegaCRN の方が，GWN−$\mathcal{G}$ と比較して，より学習データ量が重要であると考える．また，データセット 1month から 2month で 60 分先予測のみ GWN−$\mathcal{G}$ の方が性能が良いのは，GWN−$\mathcal{G}$ で用いられてる WaveNet が時間的に離れている予測を容易にしているからであると考える．

6 提 案 手 法

本節では，本研究で提案する時空間 GNN モデルについて説明する．提案モデルが，高い予測精度を実現し，転移可能な知識を多く学習するために，持つと望ましい特性は，(1) 時空間依存関係に加え，長期の時系列情報を考慮することができること; (2) データから計測地点間の関係を表すグラフ構造を学習することができること; (3) グラフ構造の学習に，計測地点数に次元が依存しないパラメータが用いられていることである．

特性 (1) で述べた長期の時系列情報は，ノイズに強く，より正確な時間依存関係を捉えられるため，時空間 GNN モデルの精度向上に役立つ [16]．特性 (2) により，事前に定義したグラフ構造を用いずに，計測地点間の関係を捉えることが可能になる．事前に定義したグラフ構造を用いる場合，計測地点数が変化するたびにグラフ構造を再定義する必要があり，コストがかかる．特性 (3) により，計測地点数変化後の学習に，グラフ構造学習に関する知識を転移することが可能になる．計測地点数に次元が依存するパラメータは転移することができないため，次元が依存するパラメータのみではグラフ構造学習に関する知識を転移することができない．

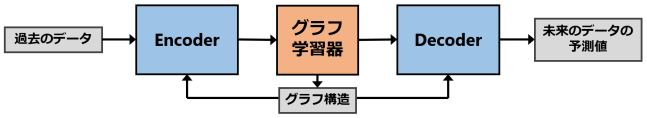


図 2: 提案モデルの全体図

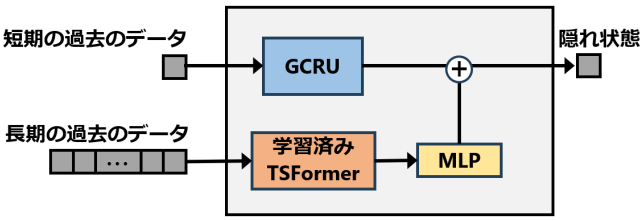


図 3: エンコーダ

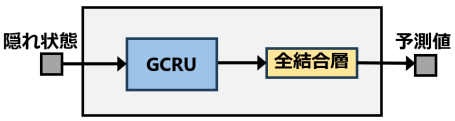


図 4: デコーダ

提案モデルのアーキテクチャは，図 2，図 3，図 4 のように設計する．上述した特性を実現するために，GCRU，TSFormer，グラフ学習器が重要な役割を担っている．本節の以降の節で，これらのモジュールについて説明する．また，提案モデルの学習の流れと，学習済み提案モデルを転移する方法について説明する．

6.1 重要なモジュール

6.1.1 GCRU

GCRU（Graph Convolution Recurrent Unit）は，GRU（Gated Recurrent Unit）にグラフ畳み込みを埋め込むことで，グラフ構造を考慮した時系列データを処理することができるモジュールである．

グラフ畳み込み演算（$\star_{\mathcal{G}}$）は，

$$\mathbf{H}_{\mathrm{GC}} = \sigma\left(\mathbf{X} \star_{\mathcal{G}} \mathbf{\Theta}_{\mathrm{GC}}\right) = \sigma\left(\sum_{k=0}^{K} \mathcal{P}^{(k)} \mathbf{X} \mathbf{W}_{\mathrm{GC}}^{(k)}\right) \tag{6}$$

と定義する．ここで，$\mathbf{X} \in \mathbb{R}^{N \times C}$ と $\mathbf{H}_{\mathrm{GC}} \in \mathbb{R}^{N \times d_{\mathrm{GC}}}$ はグラフ畳み込み演算の入力と出力，$\mathbf{\Theta} \in \mathbb{R}^{K \times C \times d_{\mathrm{GC}}}$ と $\mathbf{W}_{\mathrm{GC}}^{(k)} \in \mathbb{R}^{C \times d_{\mathrm{GC}}}$ は学習可能なグラフ畳み込みフィルタである．また，$\boldsymbol{\mathcal{P}}^{(k)} \in \mathbb{R}^{N \times N}$ はチェビシェフ多項式 [17]

$$\begin{aligned} &T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad n = 2, 3, \ldots \\ &T_0(x) = 1, \quad T_1(x) = x \end{aligned} \tag{7}$$

に対して，$T_0(\boldsymbol{\mathcal{P}}) = \mathbf{I}_N, T_1(\boldsymbol{\mathcal{P}}) = \boldsymbol{\mathcal{P}}$ とすることで得られる $T_k(\boldsymbol{\mathcal{P}})$ である．

上記のグラフ畳み込み演算（$\star_{\mathcal{G}}$）を用いて，GCRU は

$$\begin{cases} \mathbf{u}_t &= \sigma\left(\mathrm{Concat}\left[\mathbf{X}_t, \mathbf{H}_{t-1}\right] \star_{\mathcal{G}} \mathbf{\Theta}_u + \mathbf{b}_u\right) \\ \mathbf{r}_t &= \sigma\left(\mathrm{Concat}\left[\mathbf{X}_t, \mathbf{H}_{t-1}\right] \star_{\mathcal{G}} \mathbf{\Theta}_r + \mathbf{b}_r\right) \\ \tilde{\mathbf{H}}_t &= \tanh\left(\mathrm{Concat}\left[\mathbf{X}_t, (\mathbf{r}_t \odot \mathbf{H}_{t-1})\right] \star_{\mathcal{G}} \mathbf{\Theta}_{\tilde{H}} + \mathbf{b}_{\tilde{H}}\right) \\ \mathbf{H}_t &= \mathbf{u}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{u}_t) \odot \tilde{\mathbf{H}}_t \end{cases} \tag{8}$$

のように定式化される．ここで，$\mathbf{u}$ は更新ゲート，$\mathbf{r}$ はリセットゲート，$\tilde{\mathbf{H}}$ は候補状態であり，$\mathbf{\Theta}_u, \mathbf{\Theta}_r, \mathbf{\Theta}_{\tilde{H}}$ は対応する畳み込みフィルタである．

6.1.2 グラフ学習器

本研究では，グラフ学習器として，MegaCRN [7] で用いられている時空間メタグラフ学習器を用いる．このモジュールは，エンコーダとデコーダの間に挿入され，グラフ構造の学習と隠れ状態の拡張を行う．また，メモリ $\mathbf{\Phi} \in \mathbb{R}^{\phi \times d_{\mathrm{mem}}}$ を持つことで，計測地点レベルの時空間情報を記憶することができる．

グラフ構造の学習 メモリ $\mathbf{\Phi}$ と学習可能なパラメータ $\mathbf{W}_{E_1}, \mathbf{W}_{E_2} \in \mathbb{R}^{N \times \phi}$ を用いて，グラフ構造を学習する．まず，メモリ $\mathbf{\Phi}$ と $\mathbf{W}_{E_1}, \mathbf{W}_{E_2}$ を用いて，ノード埋め込み $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N \times d_{\mathrm{mem}}}$ を生成する

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{W}_{E_1} \mathbf{\Phi} \\ \mathbf{E}_2 &= \mathbf{W}_{E_2} \mathbf{\Phi}. \end{aligned} \tag{9}$$

次に，$\mathbf{E}_1, \mathbf{E}_2$ を用いて，グラフ $\mathbf{A}_1, \mathbf{A}_2$ を構築する

$$\begin{aligned} \mathbf{A}_1 &= \mathrm{softmax}\left(\mathrm{ReLU}\left(\mathbf{E}_1 \mathbf{E}_2^{\top}\right)\right) \\ \mathbf{A}_2 &= \mathrm{softmax}\left(\mathrm{ReLU}\left(\mathbf{E}_2 \mathbf{E}_1^{\top}\right)\right). \end{aligned} \tag{10}$$

式 (10) で構築されるグラフ構造 $\mathbf{A}_1, \mathbf{A}_2$ は，エンコーダとデコーダで用いられる．また，$\mathbf{W}_{E_1}, \mathbf{W}_{E_2}$ の次元は計測地点数に依存するが，メモリ $\mathbf{\Phi}$ の次元は計測地点数に依存しない．そのため，計測地点数変化後の学習に，メモリ $\mathbf{\Phi}$ は転移可能であり，グラフ構造の学習に関する知識を転移することができる．

隠れ状態の拡張 時空間メタグラフ学習器は，エンコーダで得られた隠れ状態 $\mathbf{H}_t \in \mathbb{R}^{N \times d_{\mathrm{GC}}}$ に対して作用し，隠れ状態を拡張する．まず，全結合層（FC パラメータ $\mathbf{W}_Q \in \mathbb{R}^{d_{\mathrm{GC}} \times d_{\mathrm{mem}}}$）を用いて，隠れ状態 $\mathbf{H}_t$ を局所的なクエリ $\mathbf{Q}_t \in \mathbb{R}^{N \times d_{\mathrm{mem}}}$ に射影する

$$\mathbf{Q}_t = \mathbf{H}_t \mathbf{W}_Q + \mathbf{b}_Q. \tag{11}$$

次に，$\mathbf{Q}_t^{(i)} \in \mathbb{R}^{d_{\mathrm{mem}}}$ とメモリ $\mathbf{\Phi}$ を用いて，メタノードベクトル $\mathbf{M}_t^{(i)} \in \mathbb{R}^{d_{\mathrm{mem}}}$ を構築する

$$\begin{cases} a_j^{(i)} &= \frac{\exp\left(\mathbf{Q}_t^{(i)} \mathbf{\Phi}^{\top}[j]\right)}{\sum_{j=1}^{\phi} \exp\left(\mathbf{Q}_t^{(i)} \mathbf{\Phi}^{\top}[j]\right)} \\ \mathbf{M}_t^{(i)} &= \sum_{j=1}^{\phi} a_j^{(i)} \mathbf{\Phi}[j]. \end{cases} \tag{12}$$

式 (12) では，$\mathbf{Q}_t^{(i)}$ とメモリ項目 $\mathbf{\Phi}[j]$ の類似度を計算した後，得られた類似度に従い各メモリ項目を足し合わせ，メタノードベクトル $\mathbf{M}_t^{(i)}$ を構築している．最後に，再構築して得られた $\mathbf{M}_t$ を，隠れ状態 $\mathbf{H}_t$ に連結し拡張する

$$\mathbf{H}_t' = \mathrm{Concat}\left[\mathbf{H}_t, \mathbf{M}_t\right]. \tag{13}$$

式 (13) で得られた拡張隠れ状態 $\mathbf{H}_t' \in \mathbb{R}^{N \times (d_{\mathrm{GC}} + d_{\mathrm{mem}})}$ は，デコーダの入力となる．

6.1.3 TSFormer

STEP [16] で用いられる TSFormer は，1 週間のような長期の過去のデータから豊富なコンテキスト情報が含まれる表現を生成することを目的としたモジュールである．時空間 GNN モデルでは，入力時系列の長さに応じて計算量が線形または二次関数的に増加するため，長期のデータから直接学習することは困難であるが，TSFormer を用いることで，時空間 GNN モデルに長期の時系列情報を含めることができる．

TSFormer は，エンコーダ・デコーダモデルであり，エンコーダは input embeding 層と学習可能な positional encoding 層，複数の Transformer [18] ブロックで構成され，デコーダは 1 つの Transformer ブロックと多層パーセプトロン（MLP）で構成される．エンコーダとデコーダの両方で用いられている Transformer [18] は，逐次データにおける長距離依存関係や相互作用に対して優れたモデリング能力を持ち [19]，複数の Transformer ブロックを連続して適用することで，豊かな内部表現を学習できるネットワークを構築できる [20]．そのため，TSFormer は長期のデータを効率的に学習でき，エンコーダで得られる表現は豊富なコンテキスト情報を含むことができる．

TSFormer は，教師なし事前学習モデルであり，入力時系列の一部をマスクし再構成するという masked autoencoding 手法によって学習される．TSFormer の学習は，時空間 GNN モデルの学習より前に行われ，時空間 GNN モデルの学習の際には学習済み TSFormer のエンコーダ部分のみが用いられる．

6.2 モデルの学習

提案モデルの学習は，以下の手順に従い行われ，損失関数を最小化するようにパラメータが調整される．損失関数は，MegaCRN [7] と同じ損失関数を用いる．

Step.1 TSFormer を事前学習する．

Step.2 長期の過去のデータに対し，事前学習済み TSFormer エンコーダを適用し，最後のパッチに関する隠れ表現 $\mathbf{H}_{\mathrm{TS}}$ を得る．それと同時に，短期の過去のデータに対し，GCRU を適用し，隠れ表現 $\mathbf{H}_t$ を得る．

Step.3 以下の式 (14) のように，隠れ表現 $\mathbf{H}_{\mathrm{TS}}$ に対して MLP を適用し，隠れ表現 $\mathbf{H}_t$ の空間に射影し足し合わせる

$$\mathbf{H}_t' = \mathbf{H}_t + \mathrm{MLP}\left(\mathbf{H}_{\mathrm{TS}}\right). \tag{14}$$

Step.4 Step.3 で得られた隠れ表現 $\mathbf{H}_t'$ に対して，グラフ学習

表 3: 学習済みモデルの転移するパラメータの設定に関する表，固定するものを ×，ファインチューニングするものを ◯ で表す．

モデル	TSFormer	MLP	GCRU-E	$\mathbf{W}_Q$	$\mathbf{\Phi}$	GCRU-D	$\mathbf{W}_P$	パラメータ数
提案モデル 1	×	◯	◯	◯	◯	◯	◯	476697
提案モデル 2	×	×	◯	◯	◯	◯	◯	394201
提案モデル 3	×	×	×	◯	◯	◯	◯	319129
提案モデル 4.1	×	×	×	×	◯	◯	◯	315033
提案モデル 4.2	×	×	×	×	×	◯	◯	313753
提案モデル 5.1	×	×	×	×	◯	×	◯	15129
提案モデル 5.2	×	×	×	×	×	×	◯	13849
提案モデル 6.1	×	×	×	×	◯	×	×	15000
提案モデル 6.2	×	×	×	×	×	×	×	13720

パラメータ	概要
TSFormer	TSFormer に含まれているすべてのパラメータ
MLP	TSFormer の出力として得られる隠れ状態に対して適用する MLP に関するパラメータ
GCRU-E	エンコーダの GCRU に関するパラメータ
$\mathbf{W}_Q$	エンコーダの出力として得られる隠れ状態に対して適用する全結合層に関するパラメータ
$\mathbf{\Phi}$	メモリに関するパラメータ
GCRU-D	デコーダの GCRU に関するパラメータ
$\mathbf{W}_P$	デコーダの GCRU の出力として得られる隠れ状態に対して適用する全結合層に関するパラメータ

表 4: 実験 1 結果

データセット	モデル	15 分先予測			30 分先予測			60 分先予測		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
source	GWN	13.3647	20.7606	12.29%	13.9345	22.2114	13.08%	14.8129	24.0078	**13.96%**
	GWN−$\mathcal{G}$	13.4316	20.8397	12.82%	13.9945	22.3797	13.51%	14.8768	24.1690	14.48%
	MegaCRN	13.3714	20.8560	**12.29%**	13.9667	22.4667	**13.05%**	14.8589	24.4190	14.02%
	提案モデル	**13.2733**	**20.6219**	12.35%	**13.8585**	**22.2099**	13.10%	**14.7142**	**23.9873**	14.15%
METR-LA	GWN	2.6962	5.1548	6.99%	3.0718	6.1918	8.51%	3.5066	7.2705	10.29%
	GWN−$\mathcal{G}$	2.8119	5.4527	7.36%	3.2060	6.4824	8.96%	3.6195	7.4651	10.48%
	MegaCRN	2.6317	6.0970	7.97%	3.0231	**6.1290**	**8.09%**	3.4802	7.3220	**9.75%**
	提案モデル	**2.5928**	**5.0010**	**6.68%**	**2.9890**	6.1296	8.22%	**3.4010**	**7.2585**	9.91%

器を適用し，メモリ・グラフの学習，隠れ表現の拡張を行う．

Step.5 Step.4 で拡張された隠れ表現に対して，GCRU と全結合層を適用し，予測値を算出する．

6.3 学習済みモデルの転移

計測地点増加前のデータで学習した提案モデルのパラメータを，増加後のデータで学習する際に事前知識として転移する方法について説明する．

学習済みモデルのパラメータを事前知識として用いる際，ノード数（計測地点数）によって次元が決まるパラメータは転移することができないが，提案モデルではグラフ学習器の $\mathbf{W}_{E_1}$ と $\mathbf{W}_{E_2}$ のみである．そのため，$\mathbf{W}_{E_1}$ と $\mathbf{W}_{E_2}$ 以外のモデルパラメータは転移することができる．また，$\mathbf{W}_{E_1}$ と $\mathbf{W}_{E_2}$ は初期化する必要がある．

計測地点数が増加し，計測地点増加前のデータで学習した提案モデルが存在する場合，増加後のデータを用いた学習は以下の手順に従い行われる．

Step.1 学習済みモデルのパラメータ $\mathbf{\Theta}_{\text{source}}$ の中から，事前知識として用いる転移可能なパラメータ $\mathbf{\Theta}'_{\text{source}}$ を選ぶ．

Step.2 パラメータ $\mathbf{\Theta}'_{\text{source}}$ を，増加後のデータを用い学習するモデルに転移する．すなわち，増加後のデータを用い学習するモデルのパラメータの初期値に設定する．

Step.3 転移したパラメータについて，増加後のデータを用いてファインチューニングするか否か設定する．ファインチューニングを行わないパラメータは，転移した値に固定する．

Step.4 増加後のデータを用い，モデルの学習を行う．

7 実　　験

本節では，十分な学習データがある場合と，交通計測地点が増加した直後で学習データが少ない場合について考え，提案モデルの有効性と限界に関して調査する．

7.1 実 験 方 法

提案モデルの有効性と限界に関して調査するために，2 つの実験「実験 1」と「実験 2」を行う．実験の目的は，それぞれ以下の通りである．

実験 1　十分な量のデータで学習した場合の，提案モデルとベースラインの予測精度を検証すること．

実験 2　計測地点が増加した直後の十分な量の学習データが得られない状況下で，増加前のデータで学習したモデルが存在し事前知識として用いた場合の，提案モデルとベースラインの予測精度を検証すること．

また，実験方法は，それぞれ以下の通りである．

実験 1　データセット source，METR-LA を用いて提案モデルとベースラインを学習・テストし，予測精度を検証する．ここで，エンコーダの GCRU の入力は過去 1 時間，TSFormer の入力は過去 1 週間とする．

実験 2　実験 1 で学習した提案モデル・ベースラインと，データセット 1week，2week，1month を用いて検証する．具体的には，学習済みモデルのパラメータの一部を，データセット 1week，2week，1month で学習する際に事前知識として転移しモデルの初期値に設定する．本研究では，提案モデル・ベースラインともに，ノード数（計測地点数）に次元が決まるパラメータ以外のすべてのパラメータを転移する．転移するパラメータは，ベースラインではすべてファインチューニングをする．一方，提案モデルでは表 3 のように固定するものとファインチューニングをするものに複数の設定で分ける．提案モデル 4.1 と 4.2，5.1 と 5.2，6.1 と 6.2 の違いは，メモリに関するパラメータをファインチューニングするかしないかである．また，提案モデル 1 から提案モデル 6 になるほど，入力に近い層から

表 5: 実験 2 結果 1week

転移の有無	モデル	15 分先予測			30 分先予測			60 分先予測		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
転移なし	GWN	15.2197	23.5590	13.08%	16.5380	25.8319	14.58%	**19.1328**	**29.8699**	16.69%
	GWN−$\mathcal{G}$	15.2011	**23.3706**	13.06%	16.4603	**25.7494**	14.21%	19.2052	29.9540	16.49%
	MegaCRN	**15.0620**	27.6001	**12.88%**	**16.4436**	26.1918	**14.03%**	19.4867	31.2840	**16.06%**
転移あり	GWN	**14.8407**	23.4370	**12.43%**	**15.6464**	25.2990	**13.19%**	**17.1286**	**28.0982**	**14.40%**
	GWN−$\mathcal{G}$	14.8580	**23.1309**	12.74%	15.7358	**25.0828**	13.54%	17.4679	28.0988	14.95%
	MegaCRN	15.2009	23.8500	12.95%	16.2696	26.0590	13.94%	18.5019	29.7794	15.67%
	提案モデル 1	15.3937	23.9963	13.67%	16.8778	26.8114	15.00%	20.2673	31.9472	18.47%
	提案モデル 2	16.3005	25.0200	14.58%	18.1199	28.4270	16.05%	21.6225	33.8042	18.90%
	提案モデル 3	16.0449	24.5960	14.40%	17.5463	27.3311	15.63%	20.4489	31.9385	18.16%
	提案モデル 4.1	15.8733	24.4441	14.57%	17.1084	27.0541	15.26%	20.0433	31.2858	18.03%
	提案モデル 4.2	16.6285	25.9901	15.03%	17.9414	28.5116	16.24%	20.7632	32.5077	19.45%
	提案モデル 5.1	16.0641	24.9459	13.98%	16.6991	26.8382	14.71%	18.6566	30.0582	16.55%
	提案モデル 5.2	15.9233	24.7687	13.96%	16.6332	26.6665	14.90%	18.6271	29.9709	16.95%
	提案モデル 6.1	15.0758	23.5313	13.77%	15.9400	25.4892	14.68%	17.8130	29.0050	16.36%
	提案モデル 6.2	15.0678	23.5050	13.44%	16.0729	25.7016	14.57%	18.0480	29.3175	16.53%

表 6: 実験 2 結果 2week

転移の有無	モデル	15 分先予測			30 分先予測			60 分先予測		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
転移なし	GWN	**14.8859**	23.3854	12.69%	**15.9146**	**25.4068**	13.75%	17.9987	**28.8577**	15.97%
	GWN−$\mathcal{G}$	14.9265	23.5014	**12.42%**	15.9691	25.7726	**13.28%**	**17.9352**	29.3508	**14.78%**
	MegaCRN	14.8954	**23.3478**	12.78%	16.0344	25.9041	13.90%	18.3591	29.9284	15.76%
転移あり	GWN	14.5749	23.0098	12.35%	15.2425	24.6735	13.00%	16.5021	27.1699	14.09%
	GWN−$\mathcal{G}$	14.6824	23.2572	12.45%	15.4402	25.1253	13.00%	16.7719	27.6267	14.70%
	MegaCRN	14.9770	23.3359	12.83%	15.9538	25.5662	13.82%	17.8069	28.6616	15.36%
	提案モデル 1	14.4356	22.8531	12.15%	15.1641	24.9443	12.81%	16.3737	27.3831	13.80%
	提案モデル 2	14.4083	22.5544	12.11%	15.1162	24.5486	12.73%	16.3054	**26.9741**	**13.69%**
	提案モデル 3	14.3909	**22.4820**	12.12%	15.0755	24.4431	12.75%	16.3370	27.1094	13.72%
	提案モデル 4.1	**14.3538**	22.4847	**12.10%**	**15.0470**	**24.4404**	**12.72%**	**16.2967**	26.9770	**13.69%**
	提案モデル 4.2	14.5189	22.7840	12.25%	15.2064	24.6751	12.88%	16.5322	27.4567	13.92%
	提案モデル 5.1	14.7835	23.1736	12.51%	15.4316	25.0250	13.22%	16.8611	27.8212	14.60%
	提案モデル 5.2	14.7945	23.2286	12.46%	15.4206	25.0675	13.13%	16.8808	27.7959	14.50%
	提案モデル 6.1	14.7518	22.9646	12.58%	15.4181	24.8181	13.25%	16.9863	28.1243	14.52%
	提案モデル 6.2	14.8858	23.1012	12.66%	15.6588	25.0958	13.48%	17.2106	28.2905	14.95%

表 7: 実験 2 結果 1month

転移の有無	モデル	15 分先予測			30 分先予測			60 分先予測		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
転移なし	GWN	**14.1891**	22.4655	12.22%	15.4161	24.5596	**12.99%**	**16.9271**	27.6057	**14.23%**
	GWN−$\mathcal{G}$	14.4906	**22.3486**	**12.21%**	**15.3432**	**24.2195**	13.21%	17.0127	**27.3574**	14.32%
	MegaCRN	14.5059	22.6002	12.32%	15.4590	24.8554	13.31%	17.1354	27.9721	14.34%
転移あり	GWN	14.2846	22.2622	12.04%	14.8968	**23.8898**	12.71%	15.9846	26.2425	13.58%
	GWN−$\mathcal{G}$	14.3299	22.3487	12.37%	15.0131	24.0115	13.13%	16.1151	**26.2260**	14.20%
	MegaCRN	14.6009	22.8401	12.46%	15.4426	24.8916	13.22%	16.8866	27.6094	14.36%
	提案モデル 1	**14.0832**	**22.1766**	12.00%	**14.7069**	23.9625	12.56%	**15.7068**	26.5522	**13.37%**
	提案モデル 2	14.1656	22.4667	11.96%	14.7640	24.1314	**12.55%**	15.7799	26.5801	13.41%
	提案モデル 3	14.1955	22.3270	12.04%	14.8213	24.2676	12.64%	15.9079	26.5909	13.50%
	提案モデル 4.1	14.1528	22.2837	**11.97%**	14.7931	24.1955	12.65%	15.9133	26.7000	13.55%
	提案モデル 4.2	14.2495	22.4831	12.00%	14.8160	24.1833	12.60%	15.9229	26.6696	13.52%
	提案モデル 5.1	14.2753	22.3023	12.09%	14.8522	24.0280	12.72%	15.9818	26.3856	13.73%
	提案モデル 5.2	14.3611	22.4785	12.14%	14.9321	24.2080	12.74%	16.1123	26.7546	13.79%
	提案モデル 6.1	14.3480	22.5194	12.24%	14.9597	24.3329	12.84%	16.1264	26.9655	13.83%
	提案モデル 6.2	14.5487	22.7138	12.42%	15.2000	24.6253	13.07%	16.4656	27.4208	14.22%

パラメータが固定されファインチューニングするパラメータ数は減少する．

7.2 実験結果

実験 1 実験 1 の結果を，表 4 に示す．表 4 より，多くの指標で提案モデルの予測精度はベースラインを上回る．この結果は，提案モデルに含まれる TSFormer が長期の過去データの情報を含む予測に有用な表現を生成し，グラフ学習器が適切なグラフ構造を発見することに成功しているからだと考える．このことより，提案モデルは，十分な量の学習データある状況下で，学習データから予測に有用な知識を抽出・活用することで，入力に事前定義したグラフ構造を用いないにもかかわらず，事前に定義したグラフ構造を用いるモデルを含めたベースラインよりも優れた予測モデルとなる．

実験 2 実験 2 のデータセット 1week，2week，1month における結果を，それぞれ表 5，表 6，表 7 に示す．これらの表には，比較のために学習済みモデルのパラメータを転移しなかった場合，すなわち表 2 に示した予備分析の結果を併記する．実験 2 の結果より，以下のことが分かる．

- 交通計測地点が増加した直後の十分な量の学習データが得られない状況下で，学習済みモデルが存在する場合，学習済みモデルのパラメータを事前知識として転移することは，予測精度の改善に有効である．また，予測対象が時間的に離れているほど，大きな改善が見込める．
- 提案モデルは，計測地点増加直後の十分な量の学習データが得られない状況下においても，優れた予測精度を示し，データセット 2week，1month では最も優れたモデルとなる．このことより，提案モデルは，長期の時系列情報を含む表現の生成やグラフ構造の学習などに関する知識を転移し事前知識として用いることで，計測地点増加前の多くの知識を活用できていると考えられる．
- 一方で，データセット 1week では，提案モデルは最も優れ

たモデルにはならない．これは，TSFormer の入力の長さが原因だと考える．具体的には，TSFormer は，過去 1 週間のデータを入力とするが，過去 1 週間のデータが存在しない場合 0 で代替されるため，適切な学習が行われなかったと考える．

- また，交通計測地点増加後の学習データ量によって，最適なファインチューニングするパラメータの選択は異なる．増加後のデータが増えるほど，ファインチューニングするパラメータ数を増やした方が予測精度の改善が見込める．
- メモリに関するパラメータはファインチューニングすることで，基本的に予測精度の改善が見込める．これは，メモリはグラフ構造の構築に関わっているため，計測地点増加後のより適切なグラフ構造を発見するのに，ファインチューニングする必要があるためだと考える．より適切なグラフ構造を発見することは，GCRU の性能を強化するために必要なことである．一方で，メモリに関するパラメータをファインチューニングしない場合も，大きく予測精度は悪化しない．このことより，メモリが記憶している計測地点レベルの時空間情報は，計測地点数の増加後においても有用であることが分かる．

8 結　　論

本研究では，交通計測地点の増加直後は再学習が必要になるが，学習データが十分に得られず優れた予測モデルの学習が困難という問題を克服することを目的として，新しい時空間 GNN モデルを提案した．そして，増加前のデータで学習した提案モデルのパラメータの一部を，増加後のデータで学習する際に事前知識として転移することで上記の問題に対処した．また，提案モデルを評価するための実世界交通データセットを構築し，学習データ量と時空間 GNN モデルの予測精度の関係を明らかにした．

構築したデータセットを用いた実験の結果，学習データ量が時空間 GNN モデルの予測精度の向上に大きな役割を持つこと，計測地点増加前のデータを用いて学習した時空間 GNN モデルのパラメータの一部を転移することで，増加後の学習データ少量時における予測精度の改善に有用であることを示した．さらに，提案モデルは，モデルが持つ特性により転移に適したモデルであり，問題の対処に適したモデルであることを示した．具体的には，十分な量の学習データが得られない場合でも，学習済みモデルのパラメータの一部を転移することによって，2 倍の学習データを用いた場合と同等以上の予測精度を示すことを明らかにした．

また，今後の研究の方向性として，時空間 GNN モデルに長期の時系列情報を含める手法の改良を挙げる．提案モデルでは，長期の時系列を考慮するために過去 1 週間のデータを入力とする場合，交通計測地点増加後 1 週間経つまで高品質な学習データが得られず，適切な学習ができない．長期の時系列情報を含める手法について改良することは，増加直後のデータが極わずかな状況下でも優れた予測を行うために必要であると考える．

文　　献

[1] J. Ye, J. Zhao, K. Ye, and C. Xu. How to build a graph-based deep learning architecture in traffic domain: A survey. *IEEE Trans. on Intelligent Transportation Systems*, Vol. 23, No. 5, pp. 3904–3924, 2020.
[2] W. Jiang and J. Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, Vol. 207, p. 117921, 2022.
[3] W. Jiang, J. Luo, M. He, and W. Gu. Graph neural network for traffic forecasting: The research progress. *ISPRS International Journal of Geo-Information*, Vol. 12, No. 3, p. 100, 2023.
[4] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv:1707.01926*, 2017.
[5] C. Zheng, X. Fan, C. Wang, and J. Qi. Gman: A graph multi-attention network for traffic prediction. In *Proc. of AAAI2020*, Vol. 34, pp. 1234–1241, 2020.
[6] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv:1906.00121*, 2019.
[7] R. Jiang, Z. Wang, J. Yong, P. Jeph, Q. Chen, Y. Kobayashi, X. Song, S. Fukushima, and T. Suzumura. Spatio-temporal meta-graph learning for traffic forecasting. In *Proc. of AAAI2023*, Vol. 37, pp. 8078–8086, 2023.
[8] Q. Zhang, J. Chang, G. Meng, S. Xiang, and C. Pan. Spatio-temporal graph structure learning for traffic forecasting. In *Proc. of AAAI2020*, Vol. 34, pp. 1177–1185, 2020.
[9] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv:1410.3916*, 2014.
[10] A. Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, Vol. 8, No. 1, p. 24, 2021.
[11] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. *arXiv:1808.01974*, 2018.
[12] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. of CVPR2014*, pp. 1717–1724, 2014.
[13] B. Lu, X. Gan, W. Zhang, H. Yao, L. Fu, and X. Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proc. of KDD2022*, pp. 1162–1172, 2022.
[14] Z. Yao, S. Xia, Y. Li, G. Wu, and L. Zuo. Transfer learning with spatial–temporal graph convolutional network for traffic prediction. *IEEE Trans. on Intelligent Transportation Systems*, 2023.
[15] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on knowledge and data engineering*, Vol. 22, No. 10, pp. 1345–1359, 2009.
[16] Z. Shao, Z. Zhang, F. Wang, and Y. Xu. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proc. of KDD2022*, pp. 1567–1577, 2022.
[17] J.C Mason and D. C Handscomb. *Chebyshev polynomials*. CRC press, 2002.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.
[19] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun. Transformers in time series: A survey. *arXiv:2202.07125*, 2022.
[20] C. M. Bishop and H. Bishop. *Deep Learning: Foundations and Concepts*. Springer, 2024.

地理オブジェクトに関する事前学習を利用した地域特性推定

福地　湧†　牛尼　剛聡††　田中　克己†††　角谷　和俊†

† 関西学院大学　〒 669-1330 兵庫県三田市学園上ケ原１－番
††† 福知山公立大学　〒 620-0886 京都府福知山市堀３３７０
†† 九州大学　〒 815-8540 福岡県福岡市南区塩原４丁目９－１
E-mail: †{hel00521,sumiya}@kwansei.ac.jp, ††ushiama@design.kyushu-u.ac.jp,
†††tanaka-katsumi@fukuchiyama.ac.jp

あらまし　近年, 引っ越し先や宿泊先をオンラインサイトから選択する例が増えている. 引っ越しや旅行のように馴染みの無い地域の特性を把握することは難しい. 不動産サイトなどに掲載されたユーザレビューから地域特性を把握することも考えられるが, 投稿されたレビューが多い場合は, 全てのレビューを読むことはユーザの負担になる一方で, レビューが少ない場合には, 地域特性を正確に把握できないという問題がある. 本稿では, 機械学習を利用して地図から自動的に地域特性を予測する手法を提案する. 地域特性は対象とする地域に存在する地物のカテゴリとその位置関係に強く関係すると考えられる. 本手法では, 対象とする地域の地物をカテゴリごとに異なる色でプロットした地図画像を利用して, 畳み込みニューラルネットワークにより地図画像から地域特性を予測する. ここで, 特定の地物のみをプロットした地図からその地物のカテゴリを予測するタスクと対象地域において駅の有無を予測するタスクを事前学習として行うことで予測精度の向上を目指す. 地域の特性として「子育てのし易さ」を地図画像から予測する評価実験を行った結果, 提案手法は比較手法に比べて有効性が示された.

キーワード　機械学習, 自然言語処理, 事前学習, 地物カテゴリ, 地域特性

1　はじめに

近年, SUMMO[1]や LIFULL HOME'S[2], 楽天トラベル [3]などを用いて物件や宿泊先をオンライン上で選択することが一般化している. 引っ越しや旅行の様に馴染みの無い地域ではその地域の特性を把握したい場合が考えられる. 例えば, 子連れの家族が引っ越しを検討している場合は, 引っ越し先の子育てのしやすさや買い物のしやすさを知りたいと考えられる. 他にも, 毎日外食をしたいユーザであれば, 観光先として自然が多い地域よりも繁華街を好むと考えられる. このように, 地域特性を把握することは馴染みの無い地域においてユーザの有意義な地物検索につながると考えられる.

しかし, 馴染みの無い地域の特性を把握することは困難な場合がある. 地域特性を把握するために不動産サイトなどが公開するユーザレビューを参照する場合が考えられるが, レビューが多い場合, ユーザが全てのレビューを読むことは負担になる. また, 人口の少ない地域の場合, 住民から投稿されているレビューが少なく地域特性を正確に把握することは難しい場合がある. さらに, レビュー中で言及されている地域特性に偏りがあることも考えられる. 投稿されているレビューの数が多い地域であったとしても, その大半が買い物のしやすさに言及している場合, 教育施設の充実などの観点から子育ての利便性を知りたいユーザにとっては参考にならない. このようにレビューは地域によって差がある一方で, Google マップ [4]などの普及により, 地図はどのような地域であっても入手が容易である.

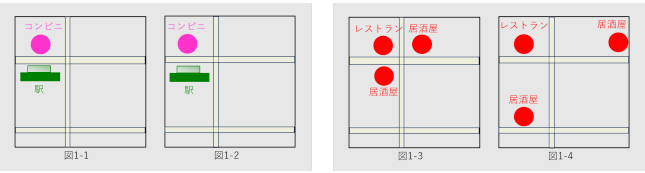


図 1　地物カテゴリの分布が地域特性に影響する例

本稿では機械学習を利用して地図から地域特性を予測する手法を提案する. レビューは地域によって得られる情報量に差があるが,Google マップ [5]などの普及により地図はどのような地域であっても入手が容易であるため, 本手法により, ユーザが任意に選んだ地域の特性を推定できると考える. 本研究では地域特性が地理オブジェクトのカテゴリやその位置関係によって異なることに着目する. 例えば, 図 1-1 と図 1-2 の地域を比較した場合, 図 1-1 の方が駅とスーパーが対面していることから, 買い物がしやすく, 子育て向いている地域である可能性が高い. また, 図 1-3 と図 1-4 を比較したとき, ともに飲食店が 3 件あるが, その密集度合いによって図 1-3 の地域の方が繁華街であるという印象を受けやすい. このように, 地理オブジェクトのカテゴリとその分布は地域特性に影響すると考えられる.

本手法では地域特性を予測するために, 対象とする地域の地

1：https://suumo.jp/
2：https://www.homes.co.jp/
3：https://travel.rakuten.co.jp/
4：https://www.google.com/maps
5：https://www.google.com/maps

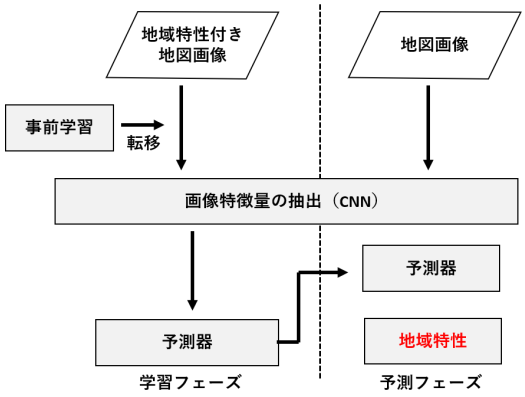


図 2　地域特性予測の処理手順

理オブジェクトをカテゴリごとに異なる色でプロットした地図画像 (*Colored-Category Map Image*. 以降では *CCMI* と呼称する.) を畳み込みニューラルネットワークを用いて学習する. 地域特性を予測する処理の手順を図 2 に示す. 図 2 に示すように本手法は学習フェーズと予測フェーズで構成される. まず, 学習フェーズでは予め地域特性を付与された *CCMI* から畳み込みニューラルネットワークを用いて, 画像の特徴量を抽出し, 地域特性の予測器を構築する. 予測フェーズでは, 学習フェーズで構築した予測器を用いてラベルが付与されていない *CCMI* から地域特性を予測する. 学習フェーズについて詳細を説明する. 学習データには地理オブジェクトを異なる色でプロットした *CCMI* を用いる. また, 学習データに対する正解ラベルは「子育てのしやすさ」や「買い物のしやすさ」のように地域特性の強さを示すスコアを用いる. また, 基本的な地理的特徴を理解した上で地域特性を予測するために, 本手法では地理オブジェクトのカテゴリの性質に関連した自己教師あり学習を事前学習として行う. 本稿で実施する自己教師あり学習は 2 つである.1 つは地理オブジェクトのカテゴリを予測するタスクであり, 地図上に単一のカテゴリの地物のみ描画し, その描画された地物を予測する. 各カテゴリが持つ地物量や位置関係の特徴を学習する狙いがある. ２つ目は駅の有無判定タスクであり, 駅がある地域と駅が無い地域の地図画像をそれぞれ学習する. 周辺地物のカテゴリや密集度合いといった駅とその他の地物カテゴリ間の関係性を学習する狙いがある. 本手法の学習データは地図をベースとした画像であるため画像処理において一般的な事前学習済みモデルをチューニングするのではなく地理的な特徴に関する自己教師あり学習を事前学習として行うことで予測精度の向上を目指す.

評価実験では, 以下の 2 点に関する実験を行うことで, 本手法の有効性を検証した.

（1） *CCMI* に描画する地理オブジェクトの数とカテゴリについて。

（2） 事前学習の有無や組み合わせによる汎化性能の変化について。

本評価実験では, 予測する地域特性の例題を「子育てのしやすさ」として実験を行った.

本論文の構成は次のとおりである. 2 章では関連研究と本論文との関連性について述べる. 3 章では *CCMI* から地域特性を予測する提案手法ついて述べる. 4 章では評価実験を通した手法の考察について述べる. 最後に, 5 章ではまとめと今後の検討について述べる.

2　関 連 研 究

2.1　地域特性の予測

地物検索において従来の検索検索方式を拡張することでユーザの潜在的な検索要求を満たす手法が多く提案されている [10] [2]. 特に, 本研究は現在の検索方式を拡張して, 地域特性に基づいて地物を推薦する手法と深く関連する. Maekawa ら [5] は目的をクエリとする地物検索システムを提案している. 例えば, 「ギターの練習ができる場所」というクエリに対して, 音楽スタジオや公園, カラオケボックスなどを推薦する. 一般的にはギターの練習ができる場所というクエリに対して, 公園やカラオケボックスを推薦することは難しいが, 地物に投稿されたレビューデータからその場所で実行可能な行動を分散表現として取得することで, カラオケボックスや公園なども「ギターの練習ができる場所」の検索結果として表示する. また, Shoji ら [7] は word2vec を地理情報に拡張した Location2Vec を提案した. Location2Vec は周辺地物に対するオンライン投稿から対象地物の周辺環境を推定する. これらの手法はユーザレビューなどのテキストデータから対象地物の特性や対象地物周辺の特性を表す分散表現を取得することで地域の特性を予測する. ユーザレビューを用いることでその地域を訪れた人が感じた地域特性の学習は可能であるが, テキストデータのみの場合, 対象地物とその周辺地物の空間的な分布関係を正確に表現することはできない. 本研究では地域特性は周辺地物との空間的な分布関係と強く関係するという仮説から, テキストデータではなく, 地図画像データを学習することで, 地理オブジェクトの分布関係に基づいた地域特性の予測を目指す.

羽田野ら [9] は周辺の地物集合（地物のカテゴリ）を機械学習して,「海辺のカフェ」や「繁華街のカフェ」などの地物の意味を表す分散表現を獲得する手法を提案した. 地物のカテゴリを機械学習する点は類似するが, 羽田野らは地物の意味を予測する一方で, 本研究は地域の意味を予測する点で異なる.

2.2　CNN を用いた地理画像の解析

畳み込みニューラルネットワーク（CNN）をはじめとした画像解析技術が高度化しており, 地理情報においても例外ではない.maggiori ら [6] により, 都市圏の空撮画像から地物を抽出するためのデータセットを構築した. 国内でも, 大野ら [14] により測量モデルを評価するための空撮画像評価用データセットが構築された. また, Hamaguchi [3] らは建物の有無だけでなく, その大きさや道路などに基づいた建物検出の手法を提案し, Yicong [8] らは 位置情報が付与されていない地域の画像からその位置情報を予測する手法を提案した. このように畳み込みニューラルネットワークは地理情報において高度な汎用性が見られている.

地理画像から地域特性を予測する手法は少ないが,

Kitabayashi [4] は間取り図からアパートの魅力度を算出し, その魅力度から地域の魅力度を算出する手法を提案した. しかし, この手法ではアパートのみの観点から地域の魅力度を予測しているため, 他の地理オブジェクトが地域に与える影響を考慮できていない. 本研究は特定の地物カテゴリに絞らず, 複数の地理オブジェクトとそれらの位置関係から地域特性を予測する手法を提案する.

2.3 事前学習の応用

畳み込みニューラルネットワークを用いた画像認識は高い精度を持っているが, より良好な結果を得るためには多くのデータが必要であるという課題がある. そこで, 古井ら [13] は物体認識において事前学習を行うことで予測精度が向上することを証明した. また, 一つのモデルの中で複数のタスクを同時に学習することで汎化性能が向上することで知られているマルチタスクモデル [1] に関する研究も盛んである. 例えば, 北村ら [15] は文章分類モデルというメインタスクにおいて品詞情報を学習するサブタスクを提案した. 日本語の品詞は機能語と内容語から構成されているため分類が難しかったが, 品詞情報をサブタスクとして学習することでメインタスクである文章の分類精度も向上することが分かった. また, 吉田ら [11] はサブタスクをメインタスクにどの層で学習するべきか提案した. 画像分析においてもマルチタスクモデルは有効であると考えらえている. 例えば, 宮崎ら [12] は料理写真の魅力度を算出するために料理の種類分類や画像の回転角の推定など魅力度を推定するために必要な基本知識をサブタスクとして学習させた. 本手法も画像解析であるため, 事前学習やマルチタスクは有効であると考えられる. しかし, 本手法で用いる学習データは従来の画像解析とは異なり地図をベースとした画像であるため, 従来の画像解析で利用されている様な訓練済みモデルは有効で無いと考える. そこで本稿では地理オブジェクトに関する基本的な知識を獲得することに特化した自己教師あり学習を事前学習として行う. マルチタスクについても検討する必要はあるが, まずは事前学習として地理オブジェクトに関するタスクを行った.

3 *CCMI* を用いた地域特性の推定手法

3.1 手法の概要

本稿は畳み込みニューラルネットワークモデルに地図画像を入力して, その地域の特性を出力する手法を提案する. 地理オブジェクトのカテゴリがその地域の特性に強く関連するという仮説から, 地理オブジェクトのカテゴリに基づいて地域特性を予測する. ここでの「地理オブジェクト」とは, 地図で表現される施設（駅, 公園, 小学校, 中学校, 高等学校など）, 自然（川, 海, 森など）, 土地（田, 畑, 果樹園, 工業地区, 河川敷など）を指す. しかし, 一般的な地図上には表現されている情報量が多く, 地理オブジェクトのカテゴリが強調されていないため, 一般的な地図を学習データに用いた場合, 地理オブジェクトのカテゴリを特徴量として地域特性を予測することは難しいと考えられる. また, 地図から地域特性を予測する場合, 地図に対する正解ラベルが必要となる. さらに, 提案手法は一般的な画像解析とは異なり地理情報を付随した画像を学習データに用いるため, 画像解析自体が難しくなる可能性が高い. そこで, 本章では下記の３点について示す.

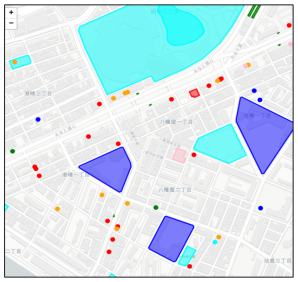



図 3　一般的な地図画像と *CCMI* の比較

- 地物のカテゴリを表現した地図画像の作成
- 地図画像に対する正解ラベル
- 地域特性の予測精度向上のための事前学習

3.2 *CCMI*

3.2.1 地理オブジェクトのカテゴリのプロット

本手法では畳み込みニューラルネットワークを利用して地図画像から地域特性を予測する. 地図はテキストデータと異なり, 地物間の位置関係を表現することができ, さらに, どのような地域であっても容易に入手が可能であるため本稿では地図を学習データに用いる. しかし, 図 3 の左側に示すような一般的な地図画像では地理オブジェクトのカテゴリが強調されていないためその地域に存在する地理オブジェクトのカテゴリを読み取るとは困難である. そこで, 本手法では, 図 3 の右図に示すような地図画像（*CCMI*）を学習データに用いる. *CCMI* 上では地理オブジェクトをカテゴリごとに異なる色でプロットすることにより, 色で地理オブジェクトのカテゴリを表現する. 畳み込みニューラルネットワークは画措値を基に特徴量を抽出するため, 地理オブジェクトのカテゴリを画措置として理解できると考える.

CCMI 上にプロットする候補の地物とそのカテゴリは下記の手順で選定する.

（1） Open Street Map[6]において出現数が高い地物をプロットの候補地物として抽出する.

（2） Yahoo!ロコ [7]のカテゴリ階層を基に地物をカテゴリに分類する.

まず, プロットの候補地物を選ぶために,Open Street Map(OSM) において出現数が高い地物を抽出する. これにより, 一般的でない地物（例: テレビ局, 発電所, 研究所など ）を *CCMI* にプロットせず, 学習が複雑化することを防ぐ. 次に, カテゴリ分類については Yahoo！ロコが公開する地物のカテゴリ情報を基に分類する. 全ての地物に異なる色を割り当てるのではなく, 地物をカテゴリに分類し, カテゴリごとに色を区別して

6：https://openstreetmap.jp/
7：https://loco.yahoo.co.jp/search/advanced/

地図上にプロットする.

3.2.2 *CCMI* の作成

CCMI を作成するために OSM を用いる. 具体的な生成手順は下記の通りである.

（1） 対象地域内に候補地物が存在する場合は地物の位置情報を OSM から取得

（2） 1 で取得した地物をカテゴリごとに可視化

まず, 対象地域内に存在する候補地物の位置情報を OSM から抽出する. 次に, それらの地物をカテゴリごとに区別して異なる色をプロットする. 地図の描画には Python ライブラリの一つである Folium を使用する. また, OSM において地物の位置情報は node と way という 2 種類の方法で表現されている. node は点情報を表しており, コンビニや保育園, 飲食店などは node で表現されている. 一方で, way は node を繋ぎ合わせて表現されており, 学校やスーパー, 公園などの領域を持つ地物は way で表現されている. 本稿においても, way で登録されている地物の位置情報は領域としてプロットする.

3.3 正解ラベル

3.3.1 地域レビューから抽出する地域特性の強さ

本手法の目的は地図から地域特性の強さを予測することである. 地域特性は人口や気候, 風土, 歴史など多様であるが, 本稿における地域特性は社会活動における地域の価値を示す. 具体的には, 「子育てのしやすさ」や「買い物のしやすさ」, 「交通の利便性」, 「地域のにぎやかさ」などを地域特性とする. 本稿ではこれらの地域特性の強さを示す指標を *CCMI* に対する正解ラベルとする. そして, 地域特性の強さを示す指標を抽出するために, 本稿では住民が地域に対して投稿したレビューを用いる. 具体的には地域特性を示すキーワードの出現率を地域特性の強さと仮定する. キーワードの出現率を使用する意図はキーワードの出現率と地域特性の強さには関連があると考えたためである. 例えば, キーワード「子育て」がレビュー中に頻出する地域であれば子育てしやすい地域であったり, キーワード「静か」が頻出する地域であれば閑静な地域であったりすると考えられる. キーワードの出現率が地域特性（社会活動における地域の価値）と直接関連するかに関しては今後, 検討の余地があるが, 本稿では地域特性を示すキーワードの出現率を地域特性の強さと仮定する.

3.3.2 地域特性の強さの算出アルゴリズム

レビューから地域特性の強さを算出するアルゴリズムの概要を図 4 に示す. まず, 対象地域に投稿されているセンテンス N 件を全て取得する. 次に, 対象とする地域特性に関係する単語を含むセンテンス M 件を取得する. 例えば, 子育てのしやすさを予測する場合には, 子育てに関連する単語を含むセンテンスを取得する. 対象地域に投稿されている全てのセンテンス N 件に対して, 地域特性を示す単語を含むセンテンス M 件の割合を算出し, 地域特性の強さを表すスコアとして *CCMI* に対する正解ラベルに用いる.

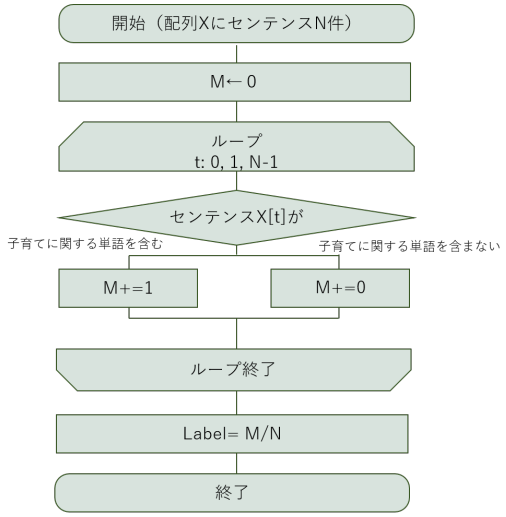


図 4 地域特性の強さの算出アルゴリズム

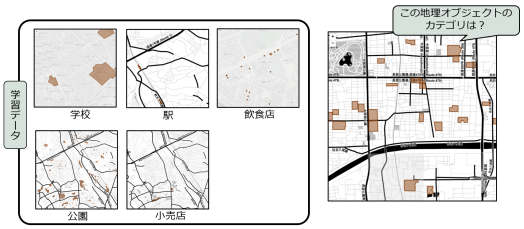


図 5 カテゴリ推定タスクの概要

3.4 地理オブジェクトに関する自己教師あり学習

3.4.1 カテゴリ推定タスク

本稿では *CCMI* から地域特性の強さを予測するために 2 つの自己教師あり学習をメインタスク (*CCMI* から地域特性を予測するタスク) の事前学習として行う. 事前学習の目的は地物カテゴリが持つ基本的な地理的特徴に基づいて地域特性を予測するためである.

1 つ目の自己教師あり学習はカテゴリ推定タスクである. 概要は図 5 に示すとおりである. まず, 単一のカテゴリのみをプロットした *CCMI* を学習する. 本稿では, 学校, 駅, 飲食店, 小売店, 公園のみをそれぞれプロットした *CCMI* を学習する. 次に入力された *CCMI* にプロットされている地物がどのカテゴリに属するか分類する. この分類モデルを構築する目的は各地物カテゴリが持つ地理的な特徴を学習することである. ここでの地理的な特徴とは地物量や密集度合いなどを指しており, 例えば, 特定の範囲における学校の密集度合いや駅の数などを学習することで, 描画されている地理オブジェクトのカテゴリを分類する. データ数は学校, 駅, 飲食店, 小売店, 公園それぞれ 200 枚ずつの合計 1,000 枚である. そのうち, 検証用データ（validation data）は 200 枚である. 学習の推移を図 6 に示す. ここでは過学習を防ぐために Early Stopping=100 として, 損失関数が最も小さく, validation data に対する accuracy が 0.9752 で epoch=220 のときのパラメータを保存する.

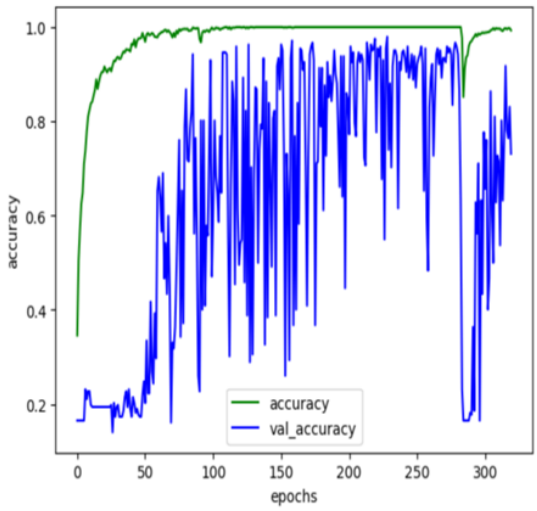


図 6　カテゴリ推定タスクの学習推移

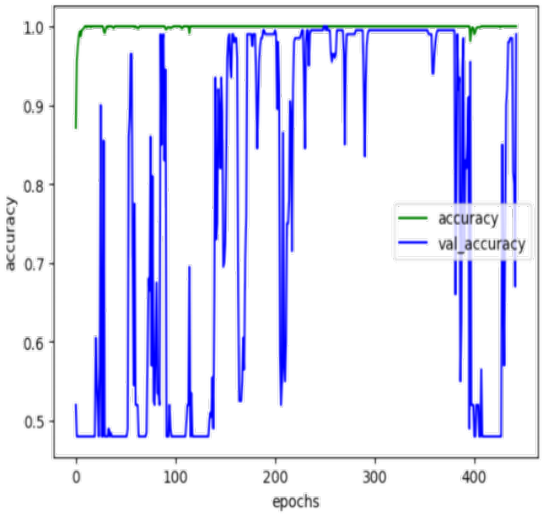


図 8　駅の有無判定タスクの学習推移

3.4.2　駅の有無判定タスク

2 つ目の自己教師あり学習は駅の有無判定タスクである. 概要を図 7 に示す. まず「駅が存在する地域」と「駅が存在しない地域」の *CCMI* を学習する. このとき, 両データ共に駅自体は描画せずに, 周辺地物のカテゴリから駅の有無を予測する. ここで学習した特徴量に基づいて, 入力画像における駅の有無を予測する. 一般的に, 駅の周辺は比較的栄えていたり, 飲食店や小売店が多い傾向にある一方で, 駅が無い地域の周辺は地物量が少なく栄えていない傾向にある. そのため, このタスクを通して, 駅とその周辺地物カテゴリの関係性を学習することが目的である. 学習データは駅がある地域と無い地域それぞれ 500 枚ずつで合計 1,000 枚である. そのうち, 検証用データは 200 枚である. 学習の推移を図 8 に示す. ここでも過学習を防ぐために Early Stopping=100 として,validation data に対する損失関数が 0.0043 で最も小さく, validation data に対する accuracy が 0.9950 で epoch=343 のときのパラメータを保存した.

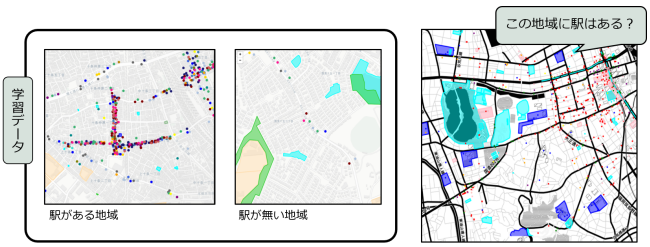


図 7　駅の有無判定タスクの概要

4　地域特性の推定手法に関する考察

4.1　評価実験の概要

地理オブジェクトのカテゴリに関する自己教師あり学習を用いて *CCMI* から地域特性を予測する提案手法の有効性を検証するために 2 つの評価実験を行った. また, 本評価実験では予測する地域特性の例題を「子育てのしやすさ」として実験を行った.

- 評価実験 1: *CCMI* のプロットの対象とする地物数とカテゴリの数について。
- 評価実験 2: 事前学習の有無と組み合わせによる予測精度の変化について。

4.2　評価実験 1

4.2.1　実験の概要

評価実験 1 の目的は *CCMI* にプロットする地物数とそのカテゴリを選定することである. プロットする地理オブジェクトが少ない場合,*CCMI* に表現される情報量が少なく十分に地域特性を読み取ることができない可能性がある. 一方で, 地理オブジェクトが多い場合は表現される情報が多すぎて地域の特性を読み取ることが困難になる可能性がある. また, プロットする地理オブジェクトのカテゴリ分類も予測に影響を与える可能性がある. カテゴリを細かく分類する場合, 各カテゴリを適切に表現できるとも考えられるが, 色が増えた結果, 複雑になり過ぎることも懸念される. 一方で, カテゴリ数が少ない場合, カテゴリを適切に表現できないが, 画像自体が単純になり学習が容易になるとも考えられる.

そこで, 表 1 に示す 4 パターンで *CCMI* を作成し, 比較実験を行った. *CCMI*-1 では OSM において出現数が上位 100 の地理オブジェクトを Yahoo!ロコの第 2 階層に基づいて 35 カテゴリに分類した. *CCMI*-2 は *CCMI*-1 と同じく出現数が上位 100 の地理オブジェクトをプロットの対象とするが, カテゴリ分類には Yahoo!ロコの第 1 階層を用いて 5 カテゴリに分類した. つまり, *CCMI* には最大 5 色がプロットされる. 表 1 の例からも分かるように, カテゴリ数を減らすことで, プロットされている地理オブジェクトの数は同じであるとしても少ない色で地域を表現する. *CCMI*-3 はプロットの対象地物数を出現数が高い上位 50 の地物に減らした. 上位 100 地物から 50 地物に減らすことで, 例えば, 大学や市役所, ショッピングモールなどはプロッ

トの対象から外れた. *CCMI*-4 はさらに対象地物を減らし上位 30 の地物をプロットの対象とした. これによって銀行や警察署, 総合病院などもプロットの対象から外れた.

表 1　評価実験 1 における比較対象

	CCMI-1	*CCMI*-2	*CCMI*-3	*CCMI*-4
地物数	出現数上位**100**地物	出現数上位**100**地物	出現数上位**50**地物	出現数上位**30**地物
カテゴリ	Yahoo!ロコ**第 2 階層**	Yahoo!ロコ**第 1 階層**	Yahoo!ロコ**第 2 階層**	Yahoo!ロコ**第 2 階層**
例				

使用したデータは合計で 1,000 枚であり, そのうち, 訓練データが 720 枚, 検証データ (validation) が 180 枚, テストデータを 100 枚とした. ここでの画像の数は地域の数を指している. 今回使用した地図画像は全て, 関西地域を対象とし, 画像サイズは 256 ピクセル× 256 ピクセルである. 地図画像に対する正解ラベルは LIFULL HOME'S[8]が提供する「まちむすびデータセット」のレビューから抽出した. 本データセットには約 41,250 の地域に対する約 330,000 件のレビューが含まれており, その一例を表 2 に示す. 本評価実験では「子育てのしやすさ」を予測するため, 各地域のレビューにおいて子育てに関連する 30 単語 (子育て, 子ども, 安心, 安全, 豊か, ファミリー, ショッピング, 学生など) が出現する割合をその地域の *CCMI* に対する正解ラベルに用いた. レビューは郵便番号ごとに投稿されているため, 各 *CCMI* は郵便番号を単位として作成されている. また, 学習に使用した畳み込みニューラルネットワークの構造は 3 つの畳み込み層と 1 つのプーリング層, そして全結合層という単純なモデル構造である. この構造は, 大野ら [14] の研究を参考にした. 損失関数は MSE, Optimezer は Adam とする. epoch 数は両手法共に 100 であるが, 過学習を防ぐために early stopping=50 とする.

4. 2. 2　結果と考察

評価実験 1 の結果を表 3 に示す. 本稿では, 予測と正解ラベルの相関の強さを示す相関係数と予測と正解ラベルの誤差を示す平均二乗誤差 (MSE) を評価指標とした. 表 3 には相関係数とその散布図, MSE を示した. 表 3 から 50 地物を候補地物として 27 のカテゴリに分類した *CCMI*-3 を訓練に用いた場合の予測精度が高くなることが分かった. 一方で, プロットの候補地物数が多い *CCMI*-1 と *CCMI*-2 を学習に用いた場合は予測の精度が低下することが分かった. 以上の結果より, 以降では 50 地物をプロットの候補として, そのカテゴリを 27 色で可視化した *CCMI*-3 を学習データに用いる. また,*CCMI*-3 のプロットの候補地物とそのカテゴリ, 色分けの例を表 4 に示す.

4. 3　評価実験 2

4. 3. 1　実験の概要

評価実験 2 の目的は事前学習として行う自己教師あり学習の有無や組み合わせによる予測精度の変化を検証することである. 本稿では地理オブジェクトのカテゴリに関する 2 つの自己教師あり学習を提案した. そこで, これらの自己教師あり学習の有無による予測精度の変化や最適な組み合わせを検証するために以下の 4 パターンで比較実験を行った.

- p-1: 事前学習を行わずに *CCMI* から子育てのしやすさを予測
- p-2: 駅の有無判定のみ事前学習を行い, *CCMI* から子育

8：https://www.homes.co.jp/

表 2　LIFULL HOME'S まちむすびデータの一例

郵便番号	レビュー
〒 1020083	不便なところもあるが、治安がよく安心して暮らせる。
〒 1920083	交通の便も良く、新宿や東京まで一本で行ける。
〒 4650093	駅近で 24 時間スーパーがある。坂道が多いのが難点
〒 7800825	路面電車の本数も多く近場は困らない。

表 3　評価実験 1 の結果

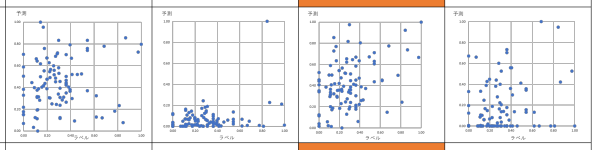
	CCMI-1	*CCMI*-2	*CCMI*-3	*CCMI*-4
相関係数	0.167	0.284	0.399	0.268
散布図				
MSE	0.102	0.097	0.072	0.088

表 4　地理オブジェクトとカテゴリ分類

カテゴリ	地物	カラー
交通	parking, station	# 696969
レストラン	restaurant	# b0c4de
ファストフード	fast food	# 4682b4
カフェ	bakery, cafe	# 4169e1
居酒屋	pub, bar	# 191970
小売店	convenience, supermarket	# 0000ff
レジャー	park, swimming pool, garden	# 00ffff
育児施設	kindergarten, childcare	# 008b8b
学校	school	# ffdead
病院	doctors, dentist, hospital, clinic	# 3cb371
公共	community center, police, fire station	# 66cdaa
寺院, 神社	place of worship	# 8fbc8f
ドラッグストア	pharmacy	# 98fb98
スポーツ	pitch, playground, sports center	# 556b2f
介護	social facility	# 808000
葬祭	grave yard, cemetery	# deb887
郵便	post office	# ffff00
金融	bank, atm	# ffd700
自動車	car	# b8860b
生活用品	variety store	# 8b4513
美容	hairdresser, beauty, massage	# 800000
ファッション	clothes,	# cd5c5c
生活サービス	fuel, dry cleaning	# 800080
食品	confectionery	# ff1493
農地	vineyard, orchard, forest	# 32cd32
工場	industrial	# bc8f8f
避難所	shelter	# 8a2be2

てのしやすさを予測

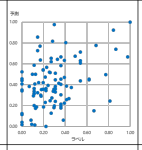
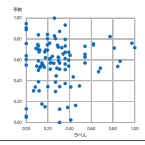
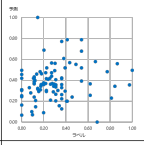
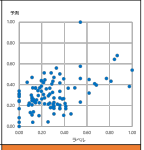
- p-3: カテゴリ推定のみ事前学習を行い, *CCMI* から子育てのしやすさを予測
- p-4: カテゴリ推定と駅の有無判定を事前学習として行い, *CCMI* から子育てのしやすさを予測

使用したデータ数や地域, モデルの構造は評価実験 1 と同じである. *CCMI* には評価実験 1 で精度が最も良かった *CCMI*-3 を採用した. また, ここでも正解ラベルは各地域のレビューにおいて子育てに関連する 30 単語 (子育て, 子ども, 安心, 安全, 豊か, ファミリー, ショッピング, 学生など) が出現する割合を使用した.

4.3.2 結果と考察

評価実験 2 の結果を表 5 に示す. 表 5 の MSE と相関係数から 2 つの事前学習を行うことで地域特性の予測精度は向上したと分かる. さらに, 事前学習を行っていない場合と比較すると相関の強さは向上していることから, 事前学習を行うことで *CCMI* と正解ラベルの関係性を理解した上で予測できていると読み取れる. しかし, 事前学習を行ったとしてもカテゴリ推定のみを行った p-2 で予測の精度が低下するという結果になった. このことから, カテゴリ推定の事前学習のみでは事前学習に有効性が無いとも読み取れる.

表 5　評価実験 2 の結果

	p-1 事前学習無し	p-2 カテゴリ推定のみ	p-3 駅の有無判定のみ	p-4 2 つの事前学習
相関係数	0.399	0.088	0.178	0.464
散布図				
MSE	0.072	0.164	0.068	0.040

4.3.3 自己教師あり学習の有効性

CCMI から地域特性を予測するにあたって, 自己教師あり学習が有効であった地域について考察する. 図 9 には自己教師あり学習により予測精度が向上した地域とその地域に対する各手法の予測と正解ラベルの誤差を示す. この地域で予測の精度が向上した要因として, 駅の周辺に地理オブジェクトが密集しているためだと考察する. 右上部に地物が密集している箇所は駅が存在するエリアであり, その影響によりコンビニやスーパーなどが密集していたため, 駅の有無を判断する「駅の有無判定タスク」とコンビニやスーパーなどの地理オブジェクトの分布特徴を学習する「カテゴリ推定タスク」がそれぞれ効果的に働いたと考える.

次に, 自己教師ありにより地域特性の予測精度が向上しなかった地域を図 10 に示す. 全ての手法において, この地域の予測が向上しなかった要因として地理オブジェクトの数が少なく, 散らばっているためだと考える. 地理オブジェクトが少ない地域であった場合,「閑静で子育てしやすい」と考えられる場合もあれば,「田舎で子育てしにくい」と考えられる場合もあるため事前学習を行ったとしても,*CCMI* から子育てのしやすさを予測することが難しかったと考えられる. しかし, たとえ地理オブジェクトの数が少なくてもコンビニやスーパーが分布していれば閑静で子育てしやすい地域とも考えられるため, それらの地理オブジェクトを重みづけできるような自己教師あり学習の検討が必要である.

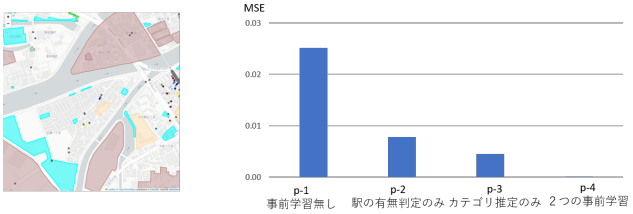


図 9　自己教師あり学習が有効的だった地域

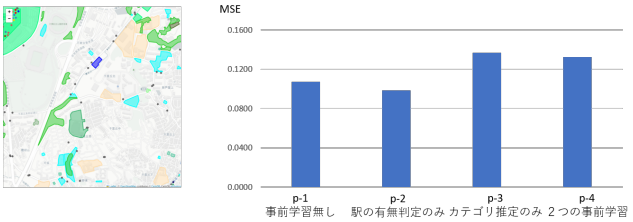


図 10　事前学習の有効性が見られなかった地域

5 おわりに

本研究は地図画像を用いて地域特性を推定するために地理オブジェクトとその位置関係に着目した手法を提案した. 本稿では, 学習データには, 地理オブジェクトのカテゴリごとにマッピングした *CCMI* を使用し, 地域特性の強さを示す正解ラベルには「地域特性を示す単語の出現率」を用いた. また, 一般的な画像処理とは異なり地図をベースとした画像を学習させる本手法の性質を考慮して, 地理オブジェクトに関する自己教師あり学習を事前学習として行うことを提案した. 評価実験では, 提案手法の汎化性能を検証するために地域特性の中でも子育てのしやすさを予測した. 評価実験の結果, *CCMI* 上にプロットする候補地物は 50 地物で 27 のカテゴリに分類し, 本稿で提案した 2 つの自己教師あり学習を共に行ったときの汎化性能が高い傾向にあることが分かった.

今後の課題は主に 2 点である. 1 点目には単語の出現率が地域特性（社会活動における地域の価値）と強い関連性があるか検証することが挙げられる. 単語の出現率が高ければその地域特性が強いと仮定して, 単語の出現率を正解ラベルに用いたが, 実際に単語の出現率が地域特性を示すことを検証する必要がある. 2 点目は地理オブジェクトに対する重みづけである. 本稿では, 出現数が高い地理オブジェクトが人間の生活において重要な役割を果たしていると仮定して,OSM において頻出数が高い地理オブジェクトを全て平等にプロットの対象とした. その結果, 大学やショッピングモール, 市役所などはプロットの対象から外れた. しかし, 市役所などは公共施設として市民の生活にお

いて重要な役割を果たしているとも考えられる. そのため, 頻出数だけでなく, 地理オブジェクトが我々の生活に与える影響度合いを考慮して地理オブジェクトのプロット対象を決める必要がある. また, 田舎や閑静の様に類似した地域特性に対しても, その違いを判別できるように, 特定の地理オブジェクトには重みを付けて処理できるような自己教師あり学習を検討する必要がある.

謝　　辞

本研究では株式会社 LIFULL 様からご提供いただいた「 LIFULL HOME'S まちむすびデータセット」を利用した. また, 本研究の一部は科研費基礎研究（ B ）（課題番号 22H03905 ）の助成を受けたものである.

文　　献

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, Vol. 17, No. 11, pp. 1949–1959, 2015.

[2] Ramesh Baral, XiaoLong Zhu, SS Iyengar, and Tao Li. Reel: Review aware explanation of location recommendation. In *Proceedings of the 26th conference on user modeling, adaptation and personalization*, pp. 23–32, 2018.

[3] Ryuhei Hamaguchi and Shuhei Hikosaka. Building detection from satellite imagery using ensemble of size-specific detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 187–191, 2018.

[4] Ryota Kitabayashi, Taro Narahara, and Toshihiko Yamasaki. Graph neural network based living comfort prediction using real estate floor plan images. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, pp. 1–5, 2022.

[5] Yui Maekawa, Yoshiyuki Shoji, and Martin J. Dürst. eographic entity retrieval for finding places suitable for certain purposes by using relevance graphs on places and reviews. *Journal of Data Intelligence (JDI), Vol.3, No.2, pp. 401 – 420, 2022*, 2022.

[6] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3226–3229. IEEE, 2017.

[7] Yoshiyuki Shoji, Katsurou Takahashi, Martin J Dürst, Yusuke Yamamoto, and Hiroaki Ohshima. Location2vec: Generating distributed representation of location by using geo-tagged microblog posts. In *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25-28, 2018, Proceedings, Part II 10*, pp. 261–270. Springer, 2018.

[8] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3616, 2017.

[9] 羽田野唯人, 渡邊二葉, 三林亮太, 荘司慶行, 山本岳洋, 大島裕, 田中克己. 周辺地物の機械学習による地物の意味の発見と言語化. データ工学と情報マネジメントに関するフォーラム,1a-5-2, 2023.

[10] 加藤誠, 大島裕明, 小山聡, 田中克己. 地域コンテキストを考慮した動的な特徴空間に基づく地理情報例示検索. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3448–3460, 2011.

[11] 鈴木優吉田朋也. マルチタスク学習モデルにおける有効な補助タスクの選定. 第 15 回データ工学と情報マネジメントに関するフォーラム, pp. 1a–3–2, 2023.

[12] 宮崎光明, 道満恵介, 井手一郎, 目加田慶人. 料理写真の魅力度推定に適したマルチタスク学習の検討. 電子情報通信学会技術研究報告; 信学技報, Vol. 121, No. 423, pp. 306–311, 2022.

[13] 古井秀弥, 黒木修隆, 廣瀬哲也, 沼昌宏ほか. 事前教師あり学習を適用した network in network による画像認識の高精度化. 情報科学技術フォーラム講演論文集, Vol. 14, No. 3, pp. 179–180, 2015.

[14] 大野裕幸, 遠藤涼, 中埜貴元, 篠田昌子. 空中写真を用いた地物抽出性能評価用データセットの構築. 人工知能学会全国大会論文集 第 33 回 (2019), pp. 3Q3J1301–3Q3J1301. 一般社団法人 人工知能学会, 2019.

[15] 鈴木優北村拓斗. Bert を用いたマルチタスク学習における 補助タスクの学習に適した層の分析. 第 15 回データ工学と情報マネジメントに関するフォーラム, pp. 1a–3–3, 2023.

チャット付き地図における会話・検索履歴に基づく地図検索手法

小柳　慶真†　　小竹　　神††　　北山　大輔†

† 工学院大学情報学部システム数理学科　〒 163-8677 東京都新宿区西新宿 1 丁目 24-2
†† 工学院大学工学研究科情報学専攻　〒 163-8677 東京都新宿区西新宿 1 丁目 24-2
E-mail: †{jx20099,em22009}@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　地図検索では多様な検索意図が存在する．代表的なものとして，地物検索や経路検索が存在する．既存の検索システムではそれらの意図に対応する個別の UI が用意される．そのため，意図を横断するような一連の検索行動を行う場合，UI を切り替えながら操作することとなる．そこで，我々は，これまでに地図にチャット欄を併設した単一の入力欄を持つ地図検索 UI を提案した．チャット欄にユーザが入力を行うことで，入力に検索意図が含まれる場合，検索意図に応じてチャットによる返信と地図に検索結果が反映される．しかしながら，チャットによる連続的な意図を反映できておらず，それぞれを入力を独立した意図として扱っているという問題があった．本研究では，チャットにおける会話・検索履歴と現在の入力内容から，検索意図の一部を引き継ぐモデルを作成し，システムを用いて有用性を検証する．

キーワード　地理データ活用，地図検索，検索意図抽出，チャット bot

1　はじめに

地図検索では多様な検索意図が存在する．代表的なものには，経路検索や地物検索が存在する．既存の地図検索システムではそれらの意図に対応する個別の UI が用意される．地物検索では，1 つのテキストボックスがある検索 UI が一般的であり，場合によってはジャンルや価格といったファセットを選択する UI が用意される．経路検索では，出発地点と目的地点の 2 つのテキストボックスがある検索 UI が一般的であり，経由地点の数だけテキストボックスが追加されることとなる．このような環境においては，ユーザは意図を横断するような検索を行う場合に，UI を切り替えながら操作を行う．我々は，これまでに地図にチャット欄を併設した UI を提案した [1]．図 1 に提案手法の概要を示す．この研究では，地図の横にチャットが接続しているインタフェースを想定している．チャット欄にユーザが入力を行うことで，入力に検索意図が含まれる場合，検索意図に応じてチャットによる返信と地図に検索結果が反映される． 具体的には，チャット欄に対する入力内容から大規模言語モデルを用いて検索意図とキーワードを抽出し，それに基づいて，地物検索や経路検索を行う．この UI を用いた実験により，ユーザは地図に対する複数種類の検索意図を一連のチャットの中で組み合わせて検索することを確認し，この UI により意図に応じて UI を切り替えることなくシームレスな地図検索体験を提供できる可能性があることを確認した．しかしながら，チャットによる連続的な意図を反映できておらず，それぞれの入力を独立した意図として扱っているという問題がある．チャットによる連続的な意図とは，旅行計画を立てている時に「温泉のあるホテルはどこ?」という検索を行い，候補を決定した後に「郷土料理のお店を探して」と検索をした場合に表示エリア全域からの検索ではなく，候補のホテル付近のお店を暗黙的に要求していることを指す．また,「東京駅付近に歴史的建造物はありますか?」という検索を行い,「池袋ではどうですか?」という検索をした時に「池袋の歴史的建造物」を暗黙的に要求している場合も，連続的な意図を含む検索である．

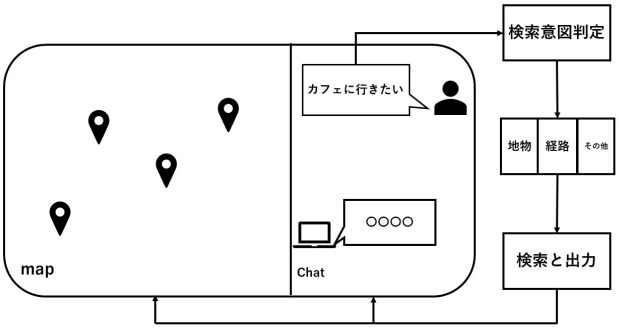


図 1　提案手法の概要

本研究では，チャットにおける会話・検索履歴と現在の入力内容から，検索意図の一部を引き継ぐモデルを作成し，システムを用いて有用性を検証する．

2　関連研究

2.1　地図を用いた検索 UI に関する研究

坂田ら [2] は，地図検索の多様な意図に対応するために，検索キーワード間の空白などを用いて，距離や範囲，方向等の空間演算を行える検索方式を提案している．しかし，扱える検索意図は演算体系の範囲内に収まり，未知の意図を汲み取ることは難しい．したがって，本研究では，多様な検索意図に対応できる枠組みを考える．具体的に，地図検索では 1 つの，経路検索では 2 つ以上の入力窓が用意される．また，飲食店検索やホテル検索では，UI としてさらに価格やジャンルの入力窓が用

意され，システムが人数の入力欄を設けている場合は人数を入力すると考えられる．しかし，そうでない場合，人数の条件は入力されることはない．そこで，検索意図を解釈し地図に反映するチャットボットにより，入力窓の制限を受けない単一の UI を提案する．

倉田 [3] は,「訪日外国人に対する旅行サービスの高度化」を目指した研究プロジェクトの一環で開発された旅行プラン作成支援プログラム「CT-Planner3」について報告している．このシステムでは，地図上に表示された旅行プランを見ながら，訪れたい観光資源を指定したり，プラン全体の性格を指示したり，滞在時間や起終点を調整したりして，探索的かつ能率的に各自の好みにあった旅行プランを作成していくことができる．CT-Plannner3 は外国人個人旅行者をターゲットに，訪日の際のプラン作成を支援するのみならず，訪日検討段階で具体的な旅行イメージを想起させ，旅行意欲を喚起することをねらいとしている．しかし，旅行プランのデザインを旅行者の心情を思い描きながら，メリハリやバラエティに気を配り，全体として大きな感動が得られるようにプランを構築していくことは，コンピュータ上で実現するには解決すべき課題が多いという結論である．本研究では，入力窓の制限を受けない単一の UI を利用することで，ユーザー自身が旅行計画を立てる支援をすることも狙いとしている．

山本ら [4] は，Focus+Context 型 Fisheye View マップの処理速度や通信速度においての課題を解決するために，Focus+Glue+Context 型 Fisheye View マップを提案した．Focus+Context 型 Fisheye View マップは，広域地図 (Context) と詳細地図 (Focus) を魚眼レンズのように部分的に拡大することで，全体の位置関係と注目地域を 1 枚の地図で把握することが可能であり，従来の Web ディジタルマップの広域地図と詳細地図を切り替えて利用しなければならないという問題を解決するためのマップである．Focus+Glue+Context 地図とは，ユーザが注目する縮尺が大きい地図 (Focus) と，全体的な位置関係を鳥瞰できる縮尺の小さい地図 (Context) と，縮尺の異なる Focus と Context の差を吸収し，両領域を結ぶ経路をもつ領域 (Glue) をもつ．これにより，Glue のみを動的に計算し Web ブラウザ側で Focus と Context と Glue を合成し，画面全体を更新する必要がある既存の Fisheye View マップよりも高速で負荷が少ない利点がある．本研究では，単一の検索窓で多様な検索意図を判定し地図に反映することで，ディジタルマップの切り替えの問題に対処している．

平元ら [5] らは，地図操作を解析することによりユーザの意図を抽出し，その意図に合った Web 情報を提示するシステムを提案した．ユーザがズームイン操作とセンタリング操作を続けた場合，最後のセンタリング操作で中央に配置した地点に関する情報を探していたと考えられるように，ユーザが何らかの意図をもち地図操作を行う場合，ある程度決まった操作列が検出できると考えられる，このように意図が抽出できる操作列を操作チャンクと定義し，ユーザの地図操作と定義された操作チャンクと比較しその意図を決定する．次に意図にあったクエリを生成し，Web ページを検索・提示する．本研究では，地図のピン状態によって検索するキーワードの処理を変えるモデルを作成することで，地図操作からユーザの意図を抽出し，検索に反映させている，

藤本ら [6] は，今までのカーナビでは処理が困難であった自然な発話に対して，適切なカーナビの処理を実行するための判定モデルを生成する過程を示し，ユーザの発話を正しく判定できることを示した．NTT ドコモは，ユーザの自然な発話によりカーナビの操作が可能なサービス「ドコモドライブネット（DDN）」の提供を開始した．DDN が実現する意図解釈機能の根幹となる技術をタスク判定と呼び，タスク判定は発話として入力された文章をあらかじめ定められたいずれかのタスクに振り分ける．この判定モデルは実験環境において 92 %，実サービス環境において 97 %以上の精度を示した．本研究では併設されたチャット欄に自然な文を入力された文章を，地物検索，経路検索，その他の 3 つの検索意図に振り分けて処理を行っている．

数原ら [7] は，キーワードに加えて地理的要求と時間的要求を入力する地理日時検索に地理日時検索支援機能を加えたアプリケーション開発を通じてユーザビリティの高い地理日時検索サービスについて検討を行い，システム開発を行った．ユーザが地理日時検索サービスに入力する文字列で表現されるクエリを検索キーワード，地理的要求を表すクエリを地理条件，時間的要求を表すクエリを日時条件と呼ぶ．数原らは検索キーワード，地理条件，日時条件のいずれかが欠けた情報要求をもつユーザに対して，欠けたクエリの入力を支援する機能を加えた．本研究では，エリアとキーワードを単一の検索窓に入力することで地理条件と検索キーワードを取得することができる．検索履歴に基づく検索を行うことで，キーワード入力が欠けていた時に一定の条件下で過去のキーワードを再利用するという処理で欠けたクエリを支援することができる．

2.2 会話型検索システムによる効果を検証する研究

奥瀬ら [8] は，ウェブ検索を通して意見形成を行う際，ユーザが能動的なウェブ検索を行い，意見形成に必要となる知識を習得した状態で意見形成を行うための問いかけ文提示インタラクションを提案した．提案システムは，ウェブ検索サイト上で動作する問いかけボットとして画面右側に問いかけボットによる問いかけインタラクションを設計し，画面左側でウェブ検索を行う．問いかけボットはその観点についての理解を深めるための問いかけ文を提示する．200 名の実験協力者を募集し，オンラインユーザ実験を行った．実験の結果，提案手法を使用した実験協力者は既存のウェブ検索 UI を模した実験協力者と比較して，ウェブ検索タスク所要時間，詳細ページ閲覧回数，クエリ発行回数が優位に増加することを示した．本研究で提案している地図検索 UI では，地図の右側にチャット欄を併設している．問いかけるような動作をすることはないが，検索意図を解釈し，地図に反映することで地図による検索で必要となる情報の取得を支援している．

Andrea ら [9] は，自然言語で検索できる 4 つの異なるデザインバリエーションのインタフェースを提案し，どのようにユー

Place search function

```
{ "name": "place_search",
  "description": "入力文中に地物検索の意図を含む場合に使用する関数．地物検索は地図上で地物を検索することであり，地物の詳細情報の取得の意図は該当しません．",
  "parameters": { "properties": {
    "location": {
        "description": "入力文中に含まれるエリア．例えば「東京」や「東京駅」,「山梨県」「池袋」など．抽出できない場合は「hoge」を入力してください．", "type": "string", },
    "keyword": {
        "description": "入力文中から考えられるエリアを除く keyword を抽出してください．抽出できない場合は「hoge」を入力してください．", "type": "string", },
}, "type": "object", }, },
```

図 2　Place search function

ザに詳細な情報の開示を促すかを長座した．その結果，チャットボットから着想を得た検索インタフェースはより多くの製品の属性を表す語である「キーファクト」が引き出された．本研究の地図検索においても地図検索 UI でチャットボットを併設することは，検索作業において効果的であると推測される．

細川ら [10] は，ユーザが提示した条件に適した情報を提供する対話システムにおいて，未知表現に対しても適切に対応できる手法を提案した．具体的には，ユーザとのインタラクションを通じて，表現と意図との関係をオンラインで適応的に管理する枠組みを提案している．これは，あらかじめシステム内に固定的に組み込んでおくことが多く，学習されていない未知表現を受けたときに，適切に対応することが困難であるという問題を解決するためのものである．提案手法では，システムの提案に対してユーザが示す支持的あるいは拒否的態度の明示性・暗示性に応じて，表現と意図との結び付きの関係を動的に変化させる枠組みが実現されている．本研究では，検索意図の判定を GPT を用いて判定することで自然言語による情報検索要求を可能にし，未知表現に対応することができる．

3　単一の入力文に対する検索意図と地図検索

基本的には [1] と同様の方法であるが，キーワードの抽出方法など一部変更を行っている．

3.1　検索意図の判定

本研究では大規模言語モデルとして OpenAI が提供する GPT(GPT-3.5-turbo-1106)[1]を用いる．入力された自然言語文から GPT により，検索意図を「経路検索」,「地物検索」,「その他」の中から判定する．判定には system プロンプトという挙動を指示する命令文として「入力される文章に含まれる地図検索の意図が経路検索もしくは地物検索を実行してください．その他の場合には other を実行してください.」を与えている．さらに，GPT の Function calling[2]というインスタンスを実装する．検索意図に応じた関数が呼び出されることで任意の語の抽出を行うことができる．本研究では，地物検索を行う

Route search function

```
{ "name": "route_search",
  "description": "入力文中に経路検索の意図を含む場合に使用する関数．経路検索は目的地までの行き方を検索する意図であり，観光ルートのように周遊する経路を探す意図は該当しません．",
  "parameters": { "properties": {
    "start_location": {
        "description": "ユーザの出発地点．例えば「東京」や「東京駅」など．", "type": "string", },
    "end_location": {
        "description": "ユーザの目的地点",
        "type": "string", },
    "waypoints": {
        "description": "ユーザが求める経由地点", "type": "string", },
    "travel_mode": {
        "description": "移動手段: driving (車),
        walking (歩き), transit (公共交通),
        bicycling (自転車)",
        "enum": ["driving", "walking",
        "transit", "bicycling"],
        "type": "string", },
    }, "type": "object",
    "required": ["end_location"], }, },
```

図 3　Route search function

other function

```
{ "name": "other",
  "description": "route_search や place_search 以外の時に実行する関数．",
  "parameters": { "properties": {
    "comment": {
        "description": "Chat による返信．",
        "type": "string", },
}, "type": "object", "required": ["comment"],
}, }
```

図 4　Other function

1：https://platform.openai.com/docs/api-reference/chat

2：https://platform.openai.com/docs/guides/function-calling

ために用いる「place search」，経路検索を行うために用いる「route search」，地物検索と経路検索以外であると判定された時に用いる「other」の 3 つの関数を設計した．設計した関数の Functions を図 2，3，4 に示す．文書生成のトリガとなる user プロンプトはチャットの入力文をそのまま用いる．例えば，ユーザが「おいしいケーキがあるカフェを教えて」という入力をしたときに system プロンプトに適応され，「地物検索」と判定される．また，ユーザが「東京駅までの経路を教えて」という入力をしたときには「経路検索」と判定される．

3.2 地物検索

入力された検索文が「地物検索」と判定された時，図 2 に示す Functions の「place search」を呼び出すことで，「location」，「keyword」を抽出し，エリアを検索する場所，キーワードを検索する内容として，Google places API[3]での検索に用いる．例えば，「東京駅のレストラン」と入力すると「東京駅」と「レストラン」が抽出され，Google places API で検索することで，地図上に地物が表示される．

3.3 経路検索

入力された検索文が「経路検索」と入力された時，図 3 に示す Functions の「route search 関数」を呼び出すことで，「start location」，「waypoints」，「end location」，「travel mode」を抽出する．抽出された語を用いて目的地までの経路と時間を検索することが出来る API である Google Map Direction API[4]で検索を行い，地図上に経路を示す．移動手段については，Google Map Direction API が日本の公共交通機関に対応していないため徒歩と車での実装となる．

3.4 その他の検索意図

入力された文が「その他」と判定された時，図 3 に示す Functions の「other 関数」を呼び出すことで，「comment」を得てその内容により，入力文に対してチャットとして返信を行う．

4 会話・検索履歴に基づく地物検索

現在の入力内容に対する直前の検索意図とキーワード，および地図上の検索ピンの状態を用いて，現在の入力内容を補完することで，連続した入力に対する意図を扱う．なお，3.4 節で説明したその他の意図については直前の検索意図には用いない．すなわち，1. 地物検索，2. その他，3. 地物検索の順で入力されたとき，3 番目の地物検索の直前の検索意図は 1 番目の地物検索となる．

4.1 地図上のピンのパターン

地図上のピンには，検索ピンと保存ピンがある．検索ピンとは，地物を検索した時に表示されているピンのことである．次の検索行った際には，表示されているピンは消え，次の検索ピンが表示される．保存ピンとは，ユーザが地物を保存した時に生じる表示され続けるピンのことである．次の検索が行われたとしても保存ピンは消えずに，表示され続ける．

表 1　ピン状態に応じたランキングの基本の処理

状態	説明	ランキング
P1	どちらもある	保存ピンの近く
P2	検索ピンのみ	なし
P3	保存ピンのみ	保存ピンの近く
P4	どちらもない	なし



図 5　1 回目の検索



図 6　ピンの状態

ピンの状態は 4 つ存在する．図 5 に 1 回目の検索時に表示されているマップを示す．黒いピンが保存ピンで，赤いピンが検索ピンである．この状態から地図操作を行い，2 回目の検索を行う場所に表示されているピンをピンの状態として採用する．各状態を図 6 に示す．1 回目の検索から右下を拡大すると図 6 の P1 のような画面になり，検索ピンと保存ピンがどちらも表示されている状態となる．また，左上を拡大すると P2 の画面になり，検索ピンだけが表示されている状態となる．右下に大きく拡大をすると P3 のように保存ピンだけが表示されている状態となる．表示領域外に移動すると P4 のように保存ピンも検索ピンも表示されていないという状態になる．ランキングには反映されないが，P1 と P2 は以前の検索に関係あるエリアで検索を行っている状態で検索意図が継続する可能性が高いと考える．P3 は過去に保存したピンが見えている状態であり，保存した時の検索意図と関係していると考える．P4 は以前のどのエリアとも異なる検索であるため，独立した検索である可能性が高いと考える．

以上のピンの状態に応じてランキングの基本の処理を決定す

3：https://developers.google.com/maps/documentation/places/web-service?hl=ja
4：https://developers.google.com/maps/documentation/directions/overview?hl=ja

る．ランキングの基本の処理を表 1 に示す．地図の表示領域内に保存ピンがある時は，ランキングを「保存ピンと近い地物を推薦する」に決定する．地図の表示領域内に保存ピンが無い時はランキングのオプションはなしである．これは，現地に行くことを前提に設計した．したがって，地図の表示領域内に保存した地物がある場合，その近くに地物が推薦されることが好ましくなるということを想定をしている．例えば，神社やお寺などの観光地を検索し，その観光地に行くことを前提に保存ピンを立てる．次にそのエリアで飲食店検索を行ったときは，保存ピンを立てた観光地の近くに飲食店が推薦されるとユーザにとって好ましい結果となると考える．

4.2 検索キーワードのパターン

キーワードの状態に応じたキーワードの基本の処理を表 2 に示す．キーワードによる基本の処理では，K1：「今回のキーワードなし」，K2：「直前と重複がある」，K3：「保存ピンと重複がある」，K4：「重複なし」の 4 状態がある．この 4 つの状態に応じて基本の処理を反映することで過去のキーワードを利用して検索を行うことができる．

K1 の基本の処理は，ユーザは場所を変更して同じキーワードのものを知りたいという検索意図でエリアのみの検索をしたと想定し，直前のキーワードを検索に用いる．K2 の基本の処理は，検索文が直前のキーワードと重複がある時，検索意図が横断していると想定し，直前のキーワードとの和集合を検索に用いる．K3 の基本の処理は，保存した地物と重複する検索文を入力した時，その地物と似た地物が推薦されることが好ましいと想定し，保存ピンとの和集合を検索に用いる．K4 の基本の処理は，重複がない検索文は検索意図の横断がなく，独立した検索であると想定し，今のキーワードを検索に用いる．

各処理の例を述べる．K1 の処理は「服屋を探してください」という検索をして，次に「池袋のは?」という検索がある時，2 回目の検索はエリアのみの検索でキーワードはなしであるため，直前の検索のキーワードである服屋を用いて，池袋の「服屋」を検索する．K2 の処理は「辛いラーメン屋はありますか?」という検索をして，次に「他にラーメン屋はありますか？」と検索がある時，2 回目の検索は直前の検索と「ラーメン屋」という重複があるため,「辛い+ラーメン屋」という和集合を 2 回目の検索のキーワードに用いる．K3 の処理は「高級レストラン」と検索していた時の地物に保存ピンを立て，表示領域内に存在していて「レストランを探してください」と検索がある時,「レストラン」という重複があるため,「高級+レストラン」という和集合を検索のキーワードに用いる．K4 の処理は「服屋を探してください」という検索をして，次に「カラオケ屋を探してください」という検索がある時，K1 から K3 のどれにも当てはまらず，独立した検索であるため，そのまま「カラオケ屋」をキーワードとして用いて検索する．

4.3 パターンの組み合わせによる処理

ピンとキーワードのパターンの組み合わせによって処理を変更をするものについて説明する．1 つ目は，直前の検索が地物検索であり,「検索ピンが見えない状態（P3，P4)」である時は,「直前のキーワードと重複がある状態（K2)」であっても,「今の検索キーワード」を用いるというものである．これは，検索ピンが見えない領域まで移動しているため，検索意図は直前と横断しているとは言えないからである．2 つ目は，直前の検索が経路検索であり,「検索ピンが見えている状態（P2)」である時，ランクを検索ピンと近いものにするというものである．

表 2　キーワードの状態に応じたキーワードの基本の処理

状態	キーワード	基本の処理
K1	今回のキーワードなし	直前のキーワード
K2	直前と重複あり	直前と今のキーワードの和集合
K3	保存ピンと重複あり	保存ピンと今のキーワードの和集合
K4	重複なし	今のキーワード

表 3　直前の検索が地物検索である時の経路検索の処理

状態	出発地点	経由地点
P1	目的地から一番遠い保存ピン	間の保存ピン
P2	マップの中心	なし
P3	目的地から一番遠い保存ピン	間の保存ピン
P4	マップの中心	なし

表 4　直前の検索が経路検索である時の経路検索の処理

状態	出発地点	経由地点
前の出発地点と 前の目的地点がある	目的地点から遠い方	目的地点から近い方
前の出発地点のみがある	前の出発地点	間の保存ピン
前の目的地点のみがある	前の目的地点	間の保存ピン
前の経由地点のみがある	前の経由地点	間の保存ピン

5 会話・検索履歴に基づく経路検索

過去の検索を利用したモデルは検索文を入力した時の検索ピンと保存ピンの状態により,「出発地点」,「経由地点」,「目的地点」を決定する．地物検索同様に，直前の検索内容を考慮して検索を支援するモデルである．直前の検索が地物検索か経路検索かで基本の処理が変わる．

5.1 直前の検索が地物検索である時の経路検索

直前の検索が地物検索である時の処理を表 3 に示す．このモデルでは，検索時に保存ピンが検索領域内にある時に，検索された目的地から一番遠い保存ピンを出発地点として扱い，出発地点と目的地点の間にある保存ピンを経由地点として扱うようにするものである．出発地点をユーザが明記している場合はその出発地点が採用される．これによって，過去に保存したピンの情報を利用し，経路検索をすることができる．

例えば，地物検索をした後に「東京駅までの行き方を教えて」と検索した時，表 3 の処理に従い，検索領域内に 1 つ保存ピンがあれば，その保存ピンを出発地点とする．複数保存ピンがあれば，一番遠い保存ピンを出発地点として間の保存ピンを経由地点とする．ユーザが「新宿駅から東京駅までの行き方は？」という様に出発地点を指定した入力を行ったときは，指定された出発地点を用いる．これにより，過去の保存ピンを利用した

表 5　提案手法と比較手法の処理 1

	比較手法	提案手法
P4K1	キーワードなしで検索	直前のキーワードで検索

表 6　P4K1 の検索内容

検索回数	pattern1	pattern2	pattern3
1 回目	服屋を探してください。	京都に書店はありますか?	新宿駅のカラオケ探して
2 回目	池袋のは？	大阪のは？	渋谷にはありますか？

表 7　提案手法と比較手法の処理 2

	比較手法	提案手法
P2K2	今のキーワードで検索	今と直前のキーワードの和集合で検索
P4K2	今と直前のキーワードの和集合で検索	今のキーワードで検索

表 8　P2K2 と P4K2 の検索内容

検索回数	pattern1	pattern2	pattern3
1 回目	辛いラーメン屋はありますか？	おでんの居酒屋を探してください	高級レストランを探してください
2 回目	他にラーメン屋はありますか？	他に居酒屋はありますか？	他にレストランはありますか？

経路検索を行うことができる．

5.2　直前の検索が経路検索である時の経路検索

直前の検索が経路検索であるときの基本の処理を表 4 に示す．検索時に検索領域内に直前の目的地点と直前の出発地点の両方がある時は，目的地点より遠い方を出発地点にし，近い方を経由地点とする．直前の出発地点または直前の目的地点，直前の経由地点のどれかが検索領域内にある時は，検索領域内にあるものを出発地点とし，経由地点は間の保存ピンとする．

例えば,「新宿駅から東京駅までの行き方教えて」という経路検索をした後に，地図操作をして,「渋谷駅までの行き方教えて」という経路検索をしたとき，新宿駅が検索領域内にあるとする．その場合，表 4 の処理に従い，新宿駅を出発地点として目的地点を渋谷駅にする．これにより、出発地点の入力が無い時に直前の経路検索の検索文を利用して検索を行うことができる．

6　実験：直前のキーワードを用いる妥当性の検証

6.1　実 験 方 法

クラウドソーシング [5]で募集した被験者を用いて，会話履歴を扱う検索の妥当性について検証した．具体的には,「どちらもなし，キーワードがなし（P4K1)」の状態で直前のキーワードを用いるという処理が妥当であるかの検証である．表 5 に示す提案手法と比較手法の出力を行った．

実験は 3 パターン用意し，1 パターンにつき 5 名ずつ計 15 名に実施した．最初に地図にチャット欄が併設されたシステムの操作動画を見る．操作動画では，2 回検索が行われ，2 回目の検索をしたところで動画は終わる．次に，2 回目の検索の結果として，現在の入力のみによって生成したマップと提案手法で生成したマップ，処理の変更をして生成したマップと処理の変更を適応せず生成したマップのいずれかの地図を見る．最後に，どちらの検索結果が適切であるか，理由とともに回答する．

5：本稿ではクラウドワークスを用いた．https://crowdworks.jp/

表 9　P4K1 の結果

	比較手法	提案手法
pattern1	0	5
pattern2	1	4
pattern3	1	4
ALL	2	13

操作動画で検索した内容はに表 6 に示す．

6.2　結果と考察

P4K1 の実験結果を表 9 に示す．P4K1 では，比較手法による検索が 2 人，提案手法による検索が 13 人という結果になった．選択された理由として，pattern1，pattern2 において「池袋周辺のみの服屋を検索して欲しいというユーザの意図を入力情報から理解して，池袋周辺のみを教えてくれているため.」など表示された地物の内容を考慮し，提案手法を選択した人が過半数であった．これは，2 回目の検索がエリアの指定のみ行われているめ，検索意図を直前のキーワードの内容であると想定する人が多かったためであると考える．この結果から，P4K1 の状態で直前のキーワードを検索に用いるという処理は妥当であるといえる．

7　実験：今と直前のキーワードの和集合を用いる妥当性の検証

7.1　実 験 方 法

クラウドソーシングで募集した被験者を用いて，会話履歴を扱う検索の妥当性について検証した．具体的には，検索ピンがあり，直前のキーワードと重複がある（P2K2)」の状態と「検索ピンがなく，直前のキーワードと重複がある（P4K2)」の状態で今と直前のキーワードの和集合を用いるという処理と組み合わせによる処理の変更が妥当であるかの検証である．表 7 に示す提案手法と比較手法の出力を行った．

実験は 3 パターン用意し，1 パターンにつき 5 名ずつ計 30

表 10　P2K2 の結果

	比較手法	提案手法
pattern1	1	4
pattern2	2	3
pattern3	2	3
All	5	10

表 11　P4K2 の結果

	比較手法	提案手法
pattern1	4	1
pattern2	5	0
pattern3	5	0
All	14	1

名に実施した．最初に地図にチャット欄が併設されたシステムの操作動画を見る．操作動画では，2 回検索が行われ，2 回目の検索をしたところで動画は終わる．次に，2 回目の検索の結果として，現在の入力のみによって生成したマップと提案手法で生成したマップ，組み合わせによる処理の変更をして生成したマップと処理の変更をせず生成したマップのいずれかの地図を見る．最後に，どちらの検索結果が適切であるか，根拠となる店舗名を必ず 1 つ以上挙げ，理由とともに回答する．操作動画で検索した内容はどちらの状態も統一している．表 8 に検索内容を示す．P4K2 の状態では，1 回目の検索ピンがないところまで移動するという地図操作を 1 回目の検索と 2 回目の検索の間に行う．

7.2　結果と考察

P2K2 の実験結果を表 10 に示す．P2K2 では，比較手法が 5 人，提案手法が 10 人という結果になった．提案手法が選択された理由として，1 回目の検索と同じジャンルの店を探しているという意図を想定している被験者がほとんどであった．つまり，直前の検索の検索ピンが見える状態で，直前のキーワードと重複がある検索をする場合は検索意図が横断していると考える人が多いということだ．したがって，P2K2 の今と直前のキーワードの和集合を検索に用いる処理は妥当であるといえる．

P4K2 の結果を表 11 に示す．P4K2 では，比較手法が 14 人，提案手法が 1 人という結果になった．比較手法が選択された理由として，P2K2 と同様に 1 回目の検索と同じジャンルの店を探しているという意図を想定している被験者がほとんどであった．つまり，直前の検索の検索ピンが見えない状態でも，直前のキーワードと重複がある検索をする場合は検索意図が横断していると考える人が多いということだ．したがって，組み合わせにより，処理を変更させ，検索ピンが見えない状態では，直前のキーワードと重複がある場合でも，今の検索キーワードで検索をするという処理は妥当ではないといえる．

本実験において検索意図の横断が想定されることが顕著に出た理由として，検索内容が大きく寄与していると考えられる．したがって，他の検索内容でも複数のパターン試行することで，妥当性の検証を進めるべきであると考察する．

8　ま　と　め

本研究では，チャットにおける会話・検索履歴と現在の入力内容から，検索意図の一部を引き継ぐモデルを作成し，有用性の検証をした．履歴を扱うパターンの妥当性の検証により，履歴を扱う検索が妥当であるという事がわかった．また，組み合わせにより，処理を変更することは妥当ではないことがわかった．しかし，今と直前のキーワードの和集合を用いる妥当性の検証において，検索意図の横断が想定されるという結果が顕著に出た．これは，検索内容のシナリオが大きく寄与していると考えるため，複数の多様なシナリオで検索意図の横断が見られるかを確認し，妥当性を検証する必要がある．

また，本稿のすべての実験において，地物検索が連続するときのみの実験であり，保存ピンの状態も考慮していない．したがって，経路検索と地物検索の連続するパターンを網羅的に検証し，保存ピンの状態についても考慮して実験を行うという事が今後の課題になる．

謝　　辞

本研究の一部は，2023 年度科研費基盤研究 (C)(課題番号：21K12147) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 小柳慶真, 関峰, 北山大輔. Gpt を用いた自然言語による地図検索手法の検討. 情報科学技術フォーラム講演論文集 (FIT), No. 23, pp. 257–258, 08 2023.
[2] 阪田晴香, パノットシリアーラヤ, 王元元, 河合由起子. 検索ワード間の空間演算の提案と地図検索への応用. 情報処理学会論文誌データベース（TOD）, Vol. 12, No. 1, pp. 6–10, jan 2019.
[3] ヨウヘイクラタ, Yohei Kurata, 陽平倉田. Ct-planer 3 : Web 上での対話的な旅行プラン作成支援. 観光科学研究, No. 5, pp. 159–165, 03 2012.
[4] 山本大介, 小関正太郎, 高橋直久. 道なり道路選別手法に基づく fisheye view マップ総描手法. 電子情報通信学会論文誌 D, No. 1, pp. 1914–1925, 10 2010.
[5] 平元綾子, 角谷和俊. オンライン地図におけるユーザ操作を用いた web 検索方式. 電子情報通信学会論文誌 D, No. 01, pp. 257–268, 02 2007.
[6] 藤本拓, 原隆浩, 西尾章治郎. 自然な発話により操作可能なカーナビゲーションシステムの開発. 電子情報通信学会論文誌 D, No. 01, pp. 2815–1824, 11 2013.
[7] 数原良彦, 藤田尚樹, 廣嶋伸章, 片渕典史, 片岡良治. 地域特有の話題発見を支援するスマートフォン向けマップ型検索システム：発見探地図エリアダス. 電子情報通信学会論文誌 D, No. 01, pp. 1300–1312, 05 2013.
[8] 奥瀬雄哉, 山本祐輔. ウェブ検索におけるトピック理解への内省を促す問いかけボット. 第 15 回データ工学と情報マネジメントに関するフォーラム, Vol. 2023, pp. 1–8, mars 2023.
[9] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. Starting conversations with search engines - interfaces that elicit natural language queries. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, p. 261–265, New York, NY, USA, 2021. Association for Computing Machinery.
[10] 細川健一郎, 藤江真也, 小林哲則, 片岡良治. 対話型情報検索システムにおける提案要求表現の適応的理解. 電子情報通信学会論文誌 D, No. 01, pp. 619–628, 03 2005.

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

2024年3月1日(金) 16:00 ~ 18:10 | T5-C オンライン（Zoom Events)

メディアと社会

座長:若宮 翔子(奈良先端科学技術大学院大学)
コメンテータ:藤原 靖宏(日本電信電話株式会社)

16:00 ~ 16:05
イントロダクション

16:05 ~ 16:30
[T5-C-9-01] メディアの報道に対するソーシャルメディアユーザの感情の可視化

*西上 貴雅[1]、風間 一洋[1]、吉田 光男[2]、土方 嘉徳[3] (1. 和歌山大学、2. 筑波大学、3. 関西学院大学)

16:30 ~ 16:55
[T5-C-9-02] 記事の書き手に着目したコメント生成によるフェイクニュース検出

*渡辺 春希[1]、森 康真[1]、田村 慶一[1] (1. 広島市立大学大学院・情報科学研究科)

16:55 ~ 17:20
[T5-C-9-03] ニュースの構成要素に着目したフェイクニュースのフェイク部分の抽出・訂正手法

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

17:20 ~ 17:45
[T5-C-9-04] メディアの報道に対するソーシャルメディアユーザの反応の時系列変化の分析手法

*井上 大成[1]、風間 一洋[1]、吉田 光男[2]、土方 嘉徳[3] (1. 和歌山大学、2. 筑波大学、3. 関西学院大学)

17:45 ~ 18:10
[T5-C-9-05] Twitterを用いたCOVID-19ワクチン接種の継続性に関係する情報共有行動の分析

*北林 遼大[1]、張 翔[1]、豊田 正史[2]、吉永 直樹[2]、梅本 和俊[2] (1. 東京大学大学院情報理工学系研究科、2. 東京大学生産技術研究所)

メディアの報道に対する ソーシャルメディアユーザの感情の可視化手法

西上　貴雅† 　風間　一洋† 　吉田　光男†† 　土方　嘉徳†††

† 和歌山大学システム工学部　〒 640-8441 和歌山県和歌山市栄谷 930
†† 筑波大学ビジネスサイエンス系　〒 112-0012 東京都文京区大塚 3-29-1
††† 関西学院大学商学部　〒 662-8501 兵庫県西宮市上ケ原一番町 1-155
E-mail: †takamasa.nishigami@gmail.com, ††kazama@ingrid.org, †††mitsuo@gssm.otsuka.tsukuba.ac.jp,
††††contact@soc-research.org

あらまし　メディアが報じるニュースのトピックによっては，ソーシャルメディア上でユーザの意見の分極化が発生するが，その現象を直感的に捉えることは難しい．本論文では，各種メディアの報道に対するソーシャルメディア上のユーザの感情の傾向を，色の分布として捉えるマクロな視点と，個々のユーザの割合として観察するミクロな視点の両方から可視化・分析する手法を提案する．まず，特定のトピックを表すクエリで検索したニュース記事とそれを引用・言及したポスト（旧：ツイート）の共引用関係を利用し，注目しているニュース記事が類似するユーザが近くに集まるように配置する．そして，Louvain 法で分割したコミュニティごとにノードの外側の色を変え，難波らの T5 感情検出器で抽出した感情を統計処理して，ノードの内側を感情ごとに異なる色の円グラフとして表示する．これにより，個々のユーザに焦点を当てると感情の割合が分かり，一方でコミュニティや全体を俯瞰する際には感情の色が視覚混合され単一の色としてユーザの分布が把握できる．さらに，実際のデータに適用して複数のトピックを分析することで，提案手法の有効性を示す．その結果，ユーザが言及したニュース記事が似ているコミュニティ間でユーザの感情に偏りがあることを明らかにした．また，提案した可視化手法は，コミュニティ全体の感情を直感的に捉えるだけでなく，個々のユーザに注目することでユーザが抱く感情を理解でき，グラフ構造の可視化方法において感情の分布を視覚的に理解する手段として有用であることが確認できた．

キーワード　ソーシャルメディア，マイクロブログ，ニュース・マスコミ，情報可視化，ユーザ理解

1　はじめに

1.1　研究の背景

メディアのニュース報道は，そのトピックの内容によっては X（旧：Twitter）などのソーシャルメディア上で広く拡散されて，多くのユーザに大きな影響を与えることが知られている [1]．その影響力の強さゆえに，ユーザの意見を操作するために事実とは異なるフェイクニュースや釣り記事が生み出されたり，政府の報道機関への介入や報道機関の利害関係者への忖度により都合が悪い情報が隠蔽される事例が見られるようになった．さらに，ユーザの個人情報を学習するアルゴリズムによって異なる意見や価値観の情報が隠されるフィルターバブル [2] や，自分と類似したユーザとばかり繋がることで情報の多様性が失われてしまうエコーチェンバー現象 [3] により，ユーザの意見の分極化を招くようになった．

そこで，中立的な視点から多様性のある情報を理解するためには，異なる視点からの報道の全体構造を俯瞰して，ソーシャルメディア空間における意見の分布や偏りを把握することが必要になる．

1.2　研究の目的

本論文では，あるトピックに関する各種メディアのニュース報道に対してソーシャルメディア上のユーザが持つ感情の分布や偏りを，色の分布として捉えるマクロな視点と，個々のユーザの感情の割合として観察するミクロな視点の両方から可視化することで，直感的に理解できるようにする手法を提案する．

まず，特定のトピックを表すクエリで検索したニュース記事とそれを引用・言及したポスト（旧：ツイート）の共引用関係を利用し，注目しているメディアやニュース記事が類似するユーザが近くに集まるように配置する．そして，Louvain 法で分割したコミュニティごとにノードの外側の色を変え，難波らの T5 感情検出器で抽出した感情を統計処理して，ノードの内側を感情ごとに異なる色の円グラフとして表示する．これにより，個々のユーザに焦点を当てると感情の割合が分かり，一方でコミュニティや全体を俯瞰する際には感情の色が視覚混合 [4] され単一の色としてユーザの分布が把握できるようになる．さらに，実際のデータに適用して複数のトピックを分析することで，提案手法の有効性を示す．

2　関連研究

ソーシャルメディア分析では，ユーザの感情を知るために，

テキストから怒りや喜びといった感情 (emotion) を分析する感情分析器や検出された感情がポジティブかネガティブかニュートラルかという極性 (sentiment) を分析する極性分析器が用いられることが多い．これらは，辞書ベースの手法と機械学習ベースの手法に大きく分類できる．以下では特に日本語に限定して説明する．

辞書ベースの感情分析器として，ML-Ask（eMotive eLement and Expression Analysis system）がよく用いられる [5] [6]．これは中村の 10 種類の感情分類 [7] に基づいた感情語辞書を用いて，マッチした感情表現の数から感情成分を推定するシステムであり，CVS（Contexual Valence Shifters）[8] に基づいて否定表現も考慮できる．辞書ベースの極性分析には，日本語評価極性辞書がよく用いられる [9] [10]．これは名詞に評価極性情報を人手で付与した辞書である．辞書ベースの手法は簡単に利用できるが，検出結果が辞書に大きく左右されるだけでなく，感情語が含まれない場合や複雑な状況の正確な判定が難しいなどの問題が存在する．

機械学習ベースの感情・極性分析器としては，福田らによる BERT を用いた感情分析器が存在する [11]．これは事前学習済み日本語 BERT モデルを Kajiwara らの日本語感情分析データセット WRIME [12] [1]でファインチューイングした感情分析器である．WRIME は，SNS 上の投稿に対して，Plutchik の基本 8 感情 [13]（喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼）を 4 段階（0, 1, 2, 3），極性を 5 段階（-2, -1, 0, 1, 2）で人手でラベル付けしたデータセットである．また，難波らはその後継として T5 を用いた感情分析器を提案した [14]．これは約 100GB の日本語コーパスで事前学習した T5 モデルを，さらに WRIME でファインチューニングした感情分析器であり，特に感情別に T5 モデルをファインチューニングした場合に，BERT モデルや全感情を一度にファインチューニングした T5 モデルよりも高い分類性能を示した．

実際に，感情分析器をソーシャルメディア分析に用いた事例として，鳥海らはソーシャルメディア上の新型コロナウイルスに関する投稿を収集し，ML-Ask を用いて各感情成分の時系列変化と日毎の感情成分の相関を分析した [15]．笹原らは，Twitter のコロナ禍における転売現象に関するポストに対して，感情辞書 ML-Ask に加えて，日本語版の心理学的カテゴリ辞書 J-LIWC と道徳基板辞書 J-MFD を用いて，消費者心理・行動を感情成分と動因カテゴリ成分，モラル成分の時系列変化を分析した [16]．福田らは BERT を用いた感情分析器を用いて，Twitter Search API で「ワクチン」をクエリとして収集したポストデータセットに対して，各感情ラベルが付与されたポスト割合の時系列変化を分析した [11]．難波らは，感情分析器を用いて「恐怖」または「不安」が 0 から 3 の 4 段階のうち 3 と判定された不安ポストをさらに 10 カテゴリーに分類し，各カテゴリのポスト数を時系列で可視化することで，技術の社会的受容性を分析した [14]．すなわち，今までは感情分析器の結果は，統計的または時系列的な分析に用いられることが多かった．

1：https://github.com/ids-cv/wrime

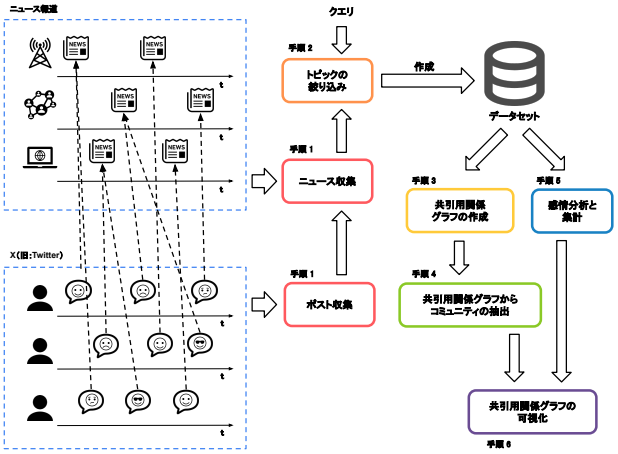


図 1　分析手法の概要

3　提 案 手 法

3.1　ソーシャルメディア空間の感情の分析

今まで感情分析器がソーシャルメディアの情報空間的な分析にほとんど用いられてこなかった理由は，大きく分けて 2 つある．まず，今まで用いられていた ML-Ask のような辞書ベースの感情分析器は，辞書に登録されている単語（感情語）しか判別できず，単独の単語ではなく文章として表される感情の検出もできないために，感情検出率が低い．さらに，各感情の検出は，辞書への感情語の登録状況に大きく左右される．例えば「給料が振り込まれた！」という文章は明らかに「喜び」の感情を含むが,「喜び」を表現する単語が文章中に存在しないため ML-Ask による感情分析結果は「無感情」となる．これらの問題から，全体的な傾向の把握程度にしか利用できず，情報空間の全体的な可視化は難しかった．

しかし，BERT や T5 のような Transformer ベースの言語モデルと感情ラベルが付与されたデータセットの登場により，現在は感情検出率が高い感情分析器が構築できるようになった．先ほどの例において，難波らの手法を用いると「喜び」が 3,「極性」が 2 と検出される．

本論文では，あるトピックに関する各種メディアのニュース報道に対してソーシャルメディア上のユーザが持つ感情の情報空間における分布を可視化することで，全般的な感情の傾向や偏りを分析する手法を提案する．提案手法の手順を以下に，概要を図 3.1 に示す．図中の手順番号は以下の手順番号と一致する．

- 手順 1：ニュース記事とポストの抽出
- 手順 2：トピックの絞り込み
- 手順 3：ユーザの共引用関係グラフの構築
- 手順 4：共引用関係グラフからのコミュニティ抽出
- 手順 5：ポストの感情分析と集計
- 手順 6：共引用関係グラフの可視化

3.2　ニュースとポストの抽出

ある期間に投稿されたニュース記事とポスト（リポストを含

む）を収集し，Elasticsearch に登録する．

3.3 トピックの絞り込み

藤兼らの論争度 [17] を用いて，論争度上位 20 件のニュース記事を求める．その後，トピッククエリを人手で選定する．藤兼らの論争度は，論争を招きやすいニュース記事の発見に使用される指標で，あるニュースが言及された期間の逆順位と言及したユーザ数の逆順位の調和平均で求めることができる．論争度は 0 から 1 の値を取り，値が大きければ大きいほど論争を招きやすいことを表す．

続いて，選定したトピッククエリを用いて Elasticsearch を全文検索して関連するニュース記事集合を抽出し，次にそれらのニュース記事の URL を用いてニュース記事を引用したポスト集合を抽出する．

3.4 ユーザの共引用関係グラフの構築とコミュニティ抽出

ユーザの共引用関係の強さはユーザに引用された記事集合の類似度として扱えると仮定して，Jaccard 係数を用いて定量化し，閾値を越える場合にエッジを張ることでユーザの共引用関係グラフを構築する．ただし，1 回しか引用していないユーザやグラフの最大連結成分に含まれないユーザは除外する．この理由は，前者は興味本位で偶然ニュース記事を引用しただけである可能性が高く，後者は，一般的なユーザとは興味・関心のベクトルが異なるために孤立しているからである．

3.5 共引用関係グラフからのコミュニティ抽出

共引用関係グラフにおける類似するユーザ集合を把握するために，Blondel らが提案した Louvain 法 [18] を用いて，共引用関係グラフを複数のコミュニティに分割する．Louvain 法は Modularity に基づくグラフ分割手法の一つであり，局所最適化と同一コミュニティに属するノードの集約により，大規模なグラフ構造を高速に処理でき，高い Modularity が得られる．

3.6 ポストの感情分析と集計

共引用関係グラフのノードであるユーザの引用ポストを感情分析する．ただし，引用ポストにはニュースカテゴリ，タイトル，メディア名，メディアのユーザ名，ハッシュタグなどのニュース記事やメディアに関する情報が同時に埋め込まれていることが多く，ユーザの発言だけを分析したい場合にはノイズとなる．そこで，中北らの手法 [19] を用いて，ユーザ自身が書いたコメント部分だけを抽出した．

さらに，ニュース記事に難波らの T5 を用いた感情分析器を適用し，Plutchik の基本 8 感情を求め，さらにそれを各ユーザごとに集計することで，ユーザがそのトピック全体に抱いた感情の割合を求める．

3.7 共引用関係グラフの可視化

得られた共引用関係グラフを可視化する．この可視化システムは Unity[2]を用いて実装した．Unity を採用した理由は，大量のオブジェクトのレンダリング処理能力が優れているから，インタラクティブなシステムを簡単にプロトタイピング出来るから，20 種類以上のプラットフォームでプログラムを動作させることができるからである．

表 1　感情とカラーコード

感情	カラーコード	サンプル
怒り	#d94638	
期待	#eb8b2f	
喜び	#eec937	
信頼	#72af4c	
恐れ	#1a824e	
驚き	#049fd9	
悲しみ	#375fac	
嫌悪	#7e4699	
無感情	#7f7f7f	

各ユーザはノードとして表す．ノードの位置は，Gephi[3]のレイアウトアルゴリズム ForceAtlas2 [20] を用いて決定する．これにより引用するニュース記事が類似するユーザ同士は近くに，類似しないユーザ同士は遠くになるように配置される．ノードの外枠は，共引用関係グラフにおける類似ユーザ集合を容易に把握できるように，Louvain 法で抽出されたコミュニティごとに異なる色で着色する．ノードの内側は，そのユーザが抱く Plutchik の基本 8 感情の割合を表 1 に示す色を用いて円グラフとして可視化する．これは Plutchik が論文で用いた配色と同じである．円グラフとして可視化することで，個々のユーザに焦点を当てると感情の割合が分かり，一方でコミュニティや全体を俯瞰する際には感情の色が視覚混合され，単一の色としてユーザの分布が把握できる．視覚混合とは，隣接する 2 色以上の色彩が，遠くから見ると混じり合ってひとつの色に見える光学現象であり，クロード・モネやジョルジュ・スーラをはじめとする印象主義の画家たちによって絵画に応用された．

実際に提案手法の可視化結果として，グラフ全体を俯瞰できるように縮小表示した例を図 2(a) に，また個々のユーザに注目できるように拡大表示した例を図 2(b) と図 2(c) に示す．図 2(a) の右下のノード群は，喜びの黄色と期待の橙色の割合が多く，明るい暖色系のノードとして見ることができる．また，左上のノード群は，嫌悪の紫色と驚きの水色，悲しみの紺色の割合が多く，寒色系のノードとして見ることができる．ここで，各ノード群に注目して拡大表示をすると，ユーザの感情の割合を個々に理解することができる．

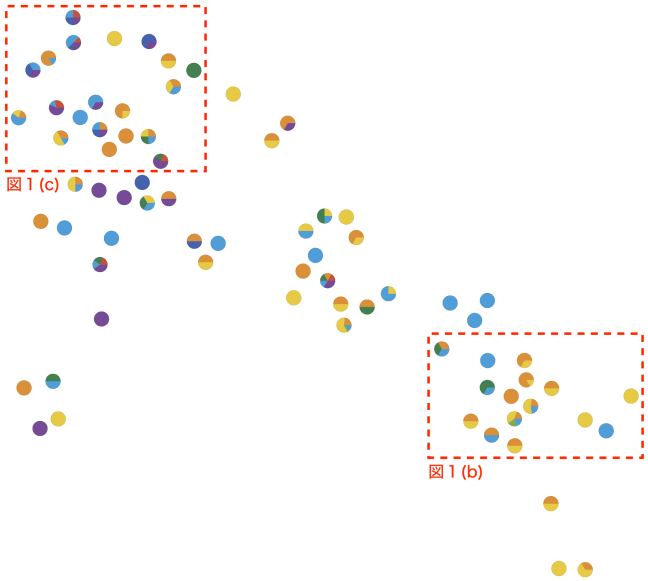
4 分　　析

4.1 データセット

ニュースに言及したポストデータセットとして，Twitter API で取得した 2022 年 1 月 1 日から 2022 年 12 月 31 日までの 1,489,667 件のニュース記事と，それらのニュース記事の URL が含まれる 3,466,472 件のポストとリポスト（旧：リツイート）を収集して分析に使用した．また，以降の章で感情分

2：https://unity.com/ja

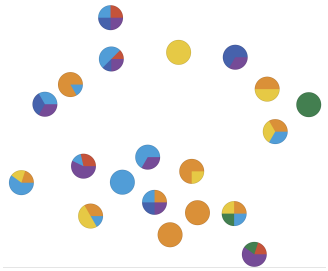
3：https://gephi.org/



(a) 縮小表示



(b) 図 2(a) の右下のノード群の拡大表示



(c) 図 2(a) の左上のノード群の拡大表示

図 2　拡大・縮小表示の例

表 2　感情分析に用いたポスト数

トピック番号	トピックを表すクエリ	ポスト数
1	太地町	232 件
2	コロナ予備費 使途	196 件
3	岸田首相 留学生	596 件
	合計	1,024 件

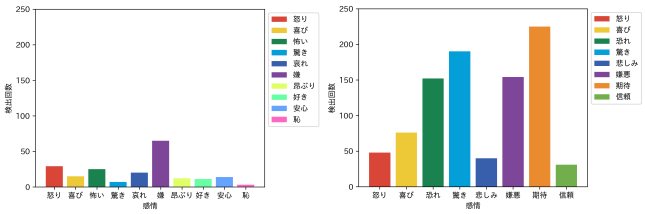
析を行ったニューストピックのポスト数を表 2 に示す.

4.2　トピックの選定

論争度上位 20 件のニュース記事を表 3 に示す．これらの記事の中から 3 件の記事を選定し，そのトピックを表す次のクエリを人手で作成した.

- トピック 1: 太地町
- トピック 2: コロナ予備費 使途
- トピック 3: 岸田首相 留学生

トピック 1 は表 3 の 5 位，トピック 2 は 1 位，トピック 3 は 12 位のニュース記事に該当する.



(a) ML-Ask　　(b) 難波らの手法

図 3　感情別の検出総数

4.3　ML-Ask と難波らの手法の比較

本論文で提案する可視化手法は，ポストからの感情検出率が高いことを前提にしている．難波らの手法の性能は論文 [14] で示されているので，本論文では辞書ベースの ML-Ask 感情検出機と機械学習ベースの難波らの手法で感情検出率がどの程度異なるかを評価した.

4.3.1　感情検出率の分析

全体的な感情検出率について分析した．ML-Ask と難波らの手法の感情検出率のトピック別と全体の結果を表 4 に示す.「感情あり」はいずれかの感情が検出されたポスト数を表すことに注意されたい．この結果を見ると，どの場合も ML-Ask よりも難波らの手法のほうが感情検出率が明らかに高いことがわかる.

4.3.2　感情別の検出総数の分析

今回選定した 3 つのトピックの全ポストから ML-Ask と難波らの手法で検出された感情の総数を比較した．感情別の検出総数を 図 3 に示す．ただし，それぞれ異なる感情分類を用いており，一致する感情は怒り，喜び，恐れ（怖い），驚き，悲しみ（哀れ），嫌悪（嫌）であり，ML-Ask では昂り，好き，安心，恥，難波らの手法では期待と信頼は一致しないので，一致する感情を左側に配置していることに注意すること．また中北らの手法 [19] を適用後にテキストが残らなかったポストは対象外とした.

まず検出総数に関しては，難波らの手法の方がどの感情においても明らかに多かった．次に感情ごとのばらつきに関しては，ML-Ask は嫌だけが突出しているのに対して，難波らの手法は嫌悪以外に恐れ，驚き，期待も同程度に検出されており，特に嫌悪よりも，ポジティブな感情である期待や Ml-Ask ではほとんど検出されていない驚きの方が多く検出されていることがわかる．これらの結果から，難波らの手法は単に感情検出率が高いだけではなく，より多くの感情をあまり偏らずに検出できていることがわかる.

4.3.3　ポストの感情割合の分析

ポストの感情割合を分析した．個々のポストの検出された感情の割合を，積み立て棒グラフを用いて，図 4，図 5，図 6 に示す．横軸はポストを時系列に従って並べた時の投稿番号を表す．X はリアルタイム性が高いので，投稿順に並べることで，同一ニュースを引用しているポスト同士は近くに配置されると共に，時系列に伴う変化が判別しやすくなると期待できる．縦軸はポストの感情割合である．比較しやすいように，一致する

表 3　2022 年の論争度上位 15 件のニュースタイトル

	タイトル	論争度
1	コロナ予備費 12 兆円、使途 9 割追えず 透明性課題	0.5000
2	日本会議系に統一教会系…安倍新内閣はまるで“カルト内閣”	0.1250
3	宮城県が水道運営権を 10 億円で売却 全国初	0.0454
4	国内死亡数が急増、1〜3 月 3.8 万人増 コロナ感染死の 4 倍	0.0425
5	「水族館の人気者」イルカが迎える恐ろしい結末 太地町のイルカ猟を毎日記録してわかったこと —...	0.0416
6	高額医療費負担、財務省「廃止を」 省庁の無駄、予算執行調査	0.0400
7	政府、マイナンバー「全口座ひも付け」義務化検討 来年の法改正目指す	0.0392
8	旧統一教会側と自民党、改憲案が「一致」 緊急事態条項、家族条項…濃厚な関係が影響？	0.0392
9	ワクチン 2 回の陽性率、半数世代で未接種上回る 厚労省再集計で判明	0.0377
10	安倍元首相銃撃の山上容疑者の背後に 2 つの”反アベ団体”か 捜査当局が重大関心	0.0338
11	水道橋博士を訴訟恫喝し顰蹙、松井市長が引き起こしたトンデモ事態！ 保健所逼迫で感染者入力を外...	0.0176
12	岸田首相「留学生 30 万人」見直し さらに増やす計画策定を指示	0.0175
13	SNS 上ひぼう中傷対策強化 侮辱罪に懲役刑導入 刑法改正案決定	0.0149
14	30 代半ば〜50 代半ばの世帯所得 20 年余前と比べ 100 万円超減少	0.0136
15	田崎史郎、橋下徹、東国原英夫、古市憲寿、三浦瑠麗…安倍元首相と統一教会の“関係”をなかったこ...	0.0136

表 4　感情検出率

トピック番号	感情検出器	感情あり	感情検出率
1	ML-Ask	28 件	12.06 %
	難波らの手法	146 件	62.93 %
2	ML-Ask	32 件	16.32 %
	難波らの手法	101 件	51.53 %
3	ML-Ask	94 件	15.77 %
	難波らの手法	330 件	55.36 %
全体	ML-Ask	154 件	15.04 %
	難波らの手法	577 件	56.35 %

感情は下から同じ順で積み上げ，その上に一致しない感情を配置した．

この結果から，感情が検出されるポストの数は難波らの手法の方が明らかに多い．また検出された場合の感情の対応を見ると，ML-Ask と同じ感情が難波らの手法でも検出されているが，難波らの手法の方が検出される感情の種類が多い．これらが感情検出率が高くなる理由だと推測できる．

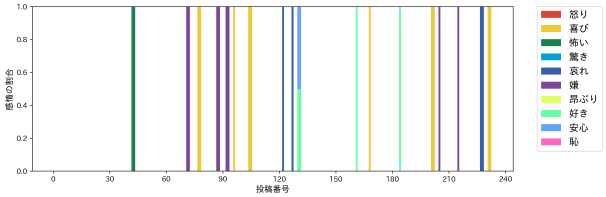
さらに時系列的な特徴に着目すると，難波らの手法は連続するポストで感情の種類や割合が類似する傾向が高いことから，棒グラフが太く表示される傾向がある．同一ニュースを引用しているポストが連続しているという仮定が正しいならば，異なるテキストのポストであっても感情が安定して検出されていると推測できる．また，ML-Ask では長い感情検出不能期間が存在するが，難波らの手法では短く，時系列的にも網羅的に検出できることがわかる．

4.4　可視化結果の分析

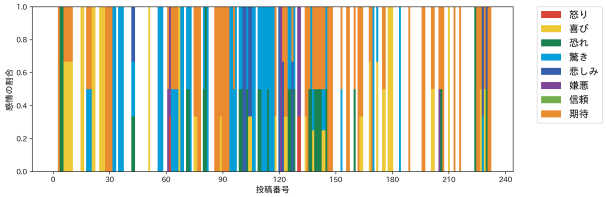
提案手法で 3 つのトピックをグラフ構造，コミュニティ分布，コミュニティ別の感情割合，感情分布の 4 種類の観点から可視化して，その結果を分析した．

4.4.1　トピック 1: 太地町

トピック 1 のグラフ構造を図 7(a) に示す．抽出されたノード数は 20，エッジ数は 105 であった．

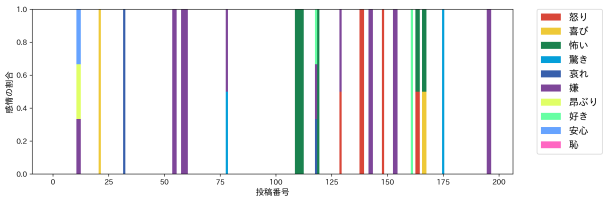


(a) ML-Ask

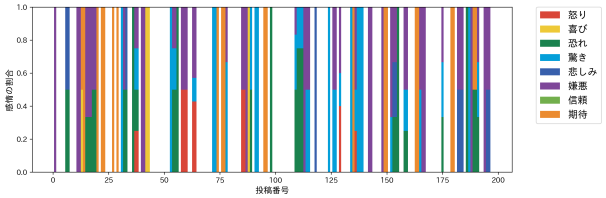


(b) 難波らの手法

図 4　トピック 1 の感情検出結果



(a) ML-Ask

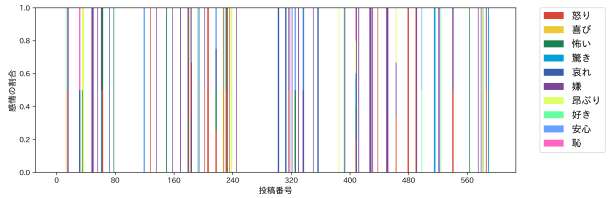


(b) 難波らの手法

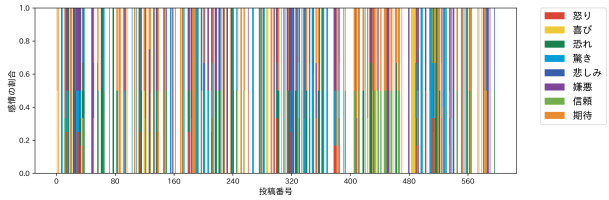
図 5　トピック 2 の感情検出結果

次にコミュニティ分布を図 7(b) に示す．なおコミュニティの違いが明確になるようにエッジは描画しない．抽出されたコミュニティ数は 2 であり，明確に分離していた．

次にコミュニティ別の感情の割合を図 7(c) に示す．横軸はコ



(a) ML-Ask



(b) 難波らの手法

図 6　トピック 3 の感情検出結果

ミュニティ番号を，縦軸はクラスタ内の感情の割合を表す．なお無感情は，テキストは存在しても感情分析の結果が無感情であったことを意味する．このトピックは，主に捕鯨文化やその周辺問題についてのポストであった．コミュニティ 0 では，喜び・希望の感情が相対的に多く検出された．捕鯨の豊漁を願うポストや太地町立くじらの博物館で行われた鯨との触れ合いを通じた教育報告のポストがあり，捕鯨文化に関して肯定的な意見が多くみられた．コミュニティ 1 では，驚き・悲しみ・嫌悪の感情が相対的に多く検出された．クジラやイルカの生態販売を問題視するポストがあり，捕鯨文化への否定的な意見が多くみられた．また，コミュニティ 1 で検出された喜び・期待の感情は，捕鯨問題に言及するニュース記事が発表されたことに対する喜びや，状態が改善されることを期待するポストだった．

最後に感情分布を図 7(d) に示す．左下のコミュニティ 0 は期待・喜びの感情を持っているために暖色系に，右上のコミュニティ 1 の右下は驚きの感情を持っているために寒色系に感じられ，コミュニティごと，またはコミュニティ内で類似した感情を持つノードが近くに配置されていることがわかる．

4.4.2　トピック 2: コロナ予備費 使途

トピック 2 のグラフ構造を図 8(a) に示す．抽出されたノード数は 37，エッジ数は 290 であった．

次にコミュニティ分布を図 8(b) に示す．抽出されたコミュニティ数は 3 であり，2 つの大きなコミュニティと 2 個のノードからなる小さなコミュニティであった．

次にコミュニティ別の感情の割合を図 8(c) に示す．このトピックは，主に予備費の使途が不透明なことに対して疑問を投げかけるポストであった．コミュニティ 0 では，「怒り・嫌悪」の感情が相対的に多く検出され，予備費の使途を管理できなかった与党（自民党）への怒りや，飽きれに近い嫌悪を含むポストがあった．コミュニティ 2 では，「恐れ」の感情が相対的に多く検出され，予備費の使途を追跡できない，管理できない政府の対応への不透明感の恐れを含むポストがあった．

最後に感情分布を図 8(d) に示す．コミュニティ 0 の右側に怒りの感情を，コミュニティ 1 の左側には驚きの感情を，コミュニティ 2 とその近辺には恐れの感情を，コミュニティ 0 とコミュニティ 1 の境界付近には喜びの感情を持つノードが集中的に分布しており，グラフの部分領域ごとに色の傾向が異なるように感じられることがわかる．

4.4.3　トピック 3: 岸田首相 留学生

トピック 3 のグラフ構造を図 9(a) に示す．抽出されたノード数は 94，エッジ数は 1,155 であった．

次にコミュニティ分布を図 9(b) に示す．抽出されたコミュニティ数は 4 であり，3 つの大きなコミュニティと 1 つの小さなコミュニティであった．

次にコミュニティ別の感情の割合を図 9(c) に示す．このトピックは，岸田首相が留学生への支援を増やすことを検討したことに対するポストであった．コミュニティ 1 と 2 は，怒りの感情が検出されると共に，嫌悪の感情の割合が他よりも多かった．これは，自国民の学生を優先するべきといった，自民党の対応についての嫌悪感であった．また，コミュニティ 0 はコロナ水際対策の緩和によるニュースに多く言及されており，期待・恐れが相対的に多く検出された．期待の感情は，コロナ水際対策の緩和による留学生の増加への期待であった．恐れの感情は，コロナ水際対策の緩和による海外からのコロナ感染者の流入への恐れであった．

最後に感情分布を図 9(d) に示す．右側のコミュニティ 0 では，恐れの感情が相対的に多く，緑色っぽく見える．また，コミュニティ 2 では，嫌悪の感情が相対的に多く出現しており，紫色っぽく見える．特に注目したいのは怒りの感情で，コミュニティ 0, 1, 2 の境界部分に固まって赤っぽく見えることがわかる．

5　おわりに

本論文では，ニュース記事の共引用関係に基づいて注目するメディアやニュース記事が類似するユーザを近くに配置し，難波らの T5 感情検出器で抽出した感情の割合を円グラフ表示することで，あるトピックに関する各種メディアのニュース報道に対してソーシャルメディア上のユーザが持つ感情の分布や偏りを可視化する手法を提案した．

まず難波らの T5 感情検出器を辞書ベースの ML-Ask と比較し，提案手法で前提とする高い感情検出率が得られることを示した．さらに実際に 3 つのトピックのニュース記事群に対するポストデータを用いて，Louvain 法で抽出されたコミュニティ間の違いやコミュニティ内の局所的な偏り，さらには複数のコミュニティにまたがる類似を判別できることを示した．

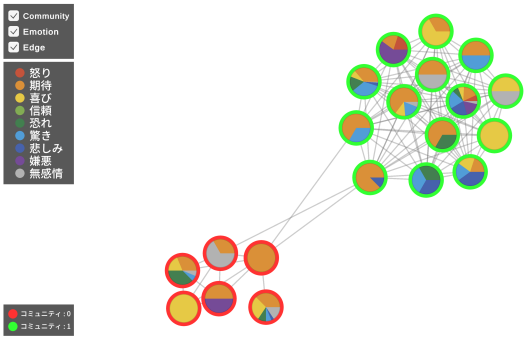
今回は感情の分布や偏りを直感的に把握できるように可視化しただけだが，今後の課題として感情の偏りの定量化指標と，感情的な類似性を考慮した新しいコミュニティ抽出法を検討する予定である．

謝　辞

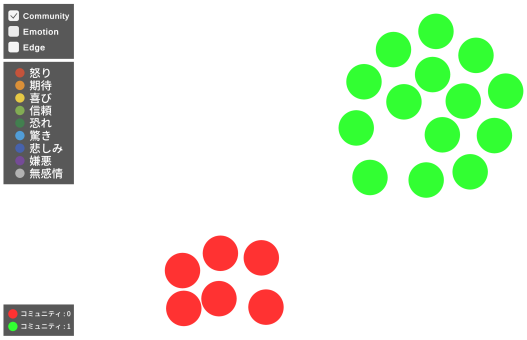
本研究は JSPS 科研費 21H03557 の助成を受けた．

文　献

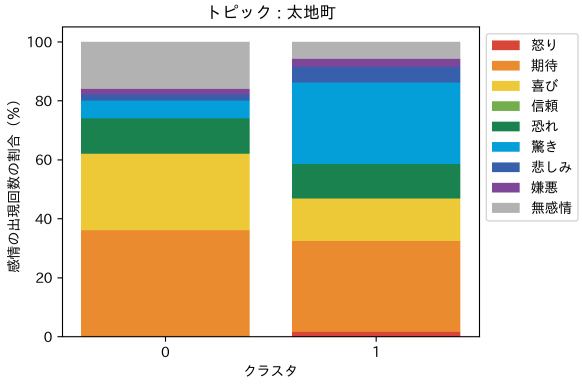
[1]　Roja Bandari, Sitaram Asur, and Bernardo Huberman. The



(a) グラフ構造



(b) コミュニティ分布

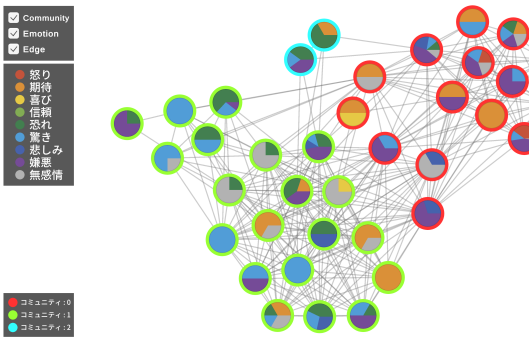


(c) コミュニティ別の感情の割合

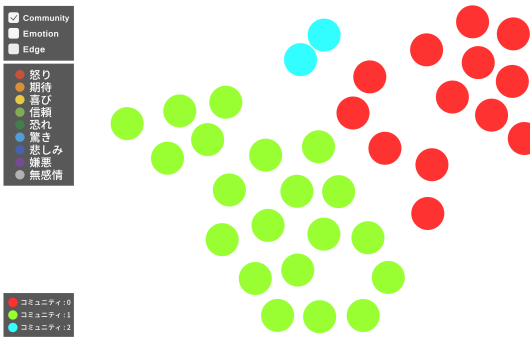


(d) 感情分布

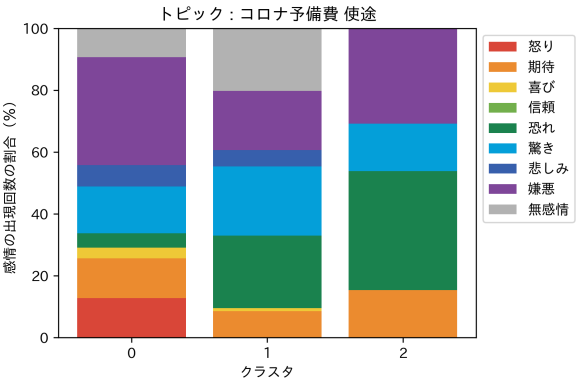
図 7　トピック 1 の可視化結果



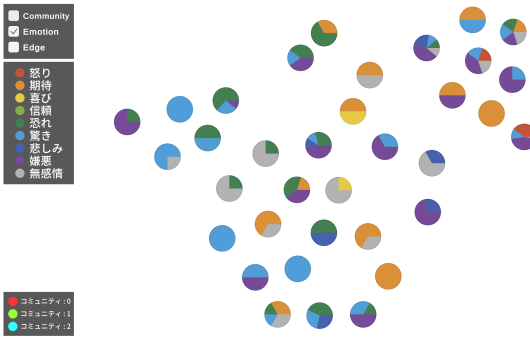
(a) グラフ構造



(b) コミュニティ分布



(c) コミュニティ別の感情の割合

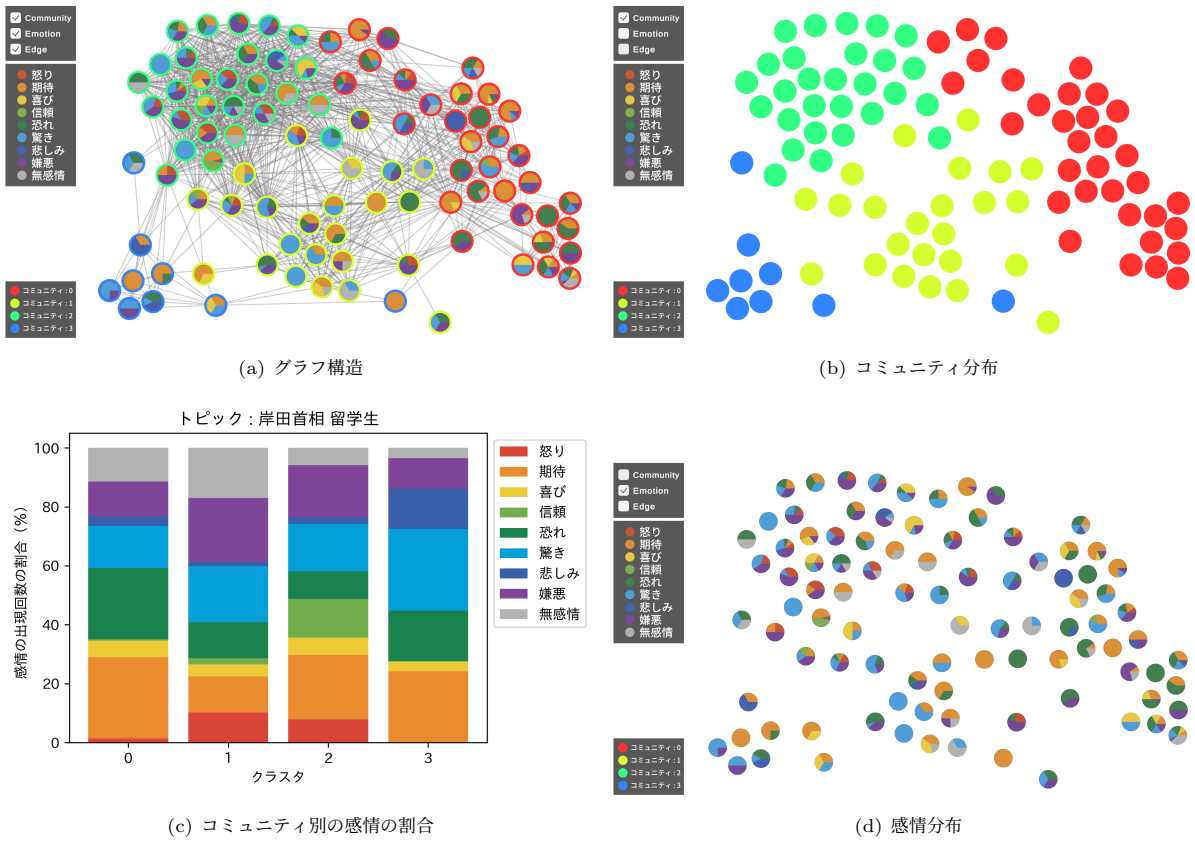


(d) 感情分布

図 8　トピック 2 の可視化結果

pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6, pp. 26–33, 2012.

[2] Eli Pariser. *The Filter Bubble: How the New Personal-*



(a) グラフ構造　(b) コミュニティ分布

(c) コミュニティ別の感情の割合　(d) 感情分布

図 9　トピック 3 の可視化結果

ized Web Is Changing What We Read and How We Think. Penguin Books, 2011.

[3] 笹原和俊. フェイクニュースを科学する: 拡散するデマ、陰謀論、プロパガンダのしくみ. 株式会社化学同人, 2019.

[4] ハーヨ・デュヒティング. ジョルジュ・スーラ：1859-1891：点に要約された絵画. タッシェンジャパン, 2000.

[5] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. A System for Affect Analysis of Utterances in Japanese Supported with Web Mining. 知能と情報, Vol. 21, No. 2, pp. 194–213, 2009.

[6] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Fumito Masui. ML-Ask: Open source affect analysis software for textual input in Japanese. *Journal of Open Research Software*, Vol. 5, No. 1, p. 16, 2017.

[7] 中村明. 感情表現辞典. 東京堂出版, 1993.

[8] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: theory and applications*, pp. 1–10. Springer, 2006.

[9] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[10] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. 584–587, 2008.

[11] 福田悟志, 難波英嗣, 庄司裕子. コロナ禍におけるワクチンに対する人々の感情変化とその要因の分析. 知能と情報, Vol. 34, No. 3, pp. 592–600, 2022.

[12] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, pp. 2095–2104, 2021.

[13] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, Vol. 89, No. 4, pp. 344–350, 2001.

[14] Hidetsugu Nanba, Katsuya Yamamoto, Satoshi Fukuda, Hiroko Shoji, Masayoshi Tanishita, Yasushi Kyutoku, and Mitsuru Yamashina. Modeling the social acceptability of technologies using Twitter data. In *Proceedings of the 2023 IEEE Conference on Systems*, 2023.

[15] 鳥海不二夫, 榊剛史, 吉田光男. ソーシャルメディアを用いた新型コロナ禍における感情変化の分析. 人工知能学会論文誌, Vol. 35, No. 4, pp. F–K45_1–7, 2020.

[16] 笹原和俊, 奥田慎平, 五十嵐祐. テキストマイニングによるコロナ禍の消費者心理・行動の定量化. 第 35 回人工知能学会全国大会 (JSAI2021), 1D3-OS-3b-04, 2021.

[17] 藤兼由生, 風間一洋, 吉田光男, 土方嘉徳. ツイートからの論争を招きやすいニュース記事の発見手法. 第 35 回人工知能学会全国大会 (JSAI2021), 1D3-OS-3b-02, 2021.

[18] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008, October 2008.

[19] 中北雄大, 風間一洋, 吉田光男, 土方嘉徳. 感情とトピックに注目したメディアの報道姿勢の分析. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2022), B24-2, 2022.

[20] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One 9(6): e98679*, 2014.

記事の書き手に着目したコメント生成による フェイクニュース検出

渡辺 春希† 森 康真† 田村 慶一†

†広島市立大学大学院情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

E-mail: †{mh67026@e. , mori@ , ktamura@}hiroshima-cu.ac.jp

あらまし ソーシャルメディアの普及によりフェイクニュースと呼ばれる事実と異なる情報が広がり，多くの人を混乱に巻き込んでいる．このようなフェイクニュースの拡散に対し，機械学習や自然言語処理の技術を用いて自動でテキストの真偽を分類するフェイクニュース検出と呼ばれる研究が行われている．フェイクニュース検出においてニュース記事だけではなく，コメントも活用することで分類精度が向上することが確認されている．しかしコメントはニュースの初期ではあまり存在しない．そのため記事に対するコメントを生成し，検出を補助する手法が存在する．既存の手法では生成の際にニュース記事のテキストのみを使用しており，書き手の情報を用いていない．そこで本研究ではニュース記事のみならず書き手の情報も用いてコメントの生成をおこない，精度向上に影響があるか調査をおこなう．

キーワード ソーシャルメディア，フェイクニュース検出，テキスト分類，ニュース・マスコミ，文章生成

1. はじめに

SNS の普及により多くの人が手軽に情報を発信し，様々な情報を取得することが可能となった．その一方で信憑性の欠けた情報であるフェイクニュースも拡散されるようになった．フェイクニュースにも様々な定義があり，ユネスコが作成した“Journalism, ‘Fake News’ & Disinformation”ではフェイクニュースに関する情報分類として misinformation，disinformation，malinformation の 3 つを提案している[1]．フェイクニュースの例として 2016 年のアメリカ大統領選挙におけるピザゲート事件が挙げられる．これはトランプ前大統領の対抗馬であったヒラリー・クリントン候補陣営の関係者が人身売買や児童性的虐待に関与し，ワシントンのピザ店がその拠点となっているという噂が広がり，疑惑を信じた男性が銃撃を起こした事件である．この背景の原因は掲示板のユーザーによる噂，ヒラリーの私用のメールアドレスの流出だと挙げている[2]．

このようなフェイクニュースの被害を抑えるためにファクトチェックサイトの事実検証が挙げられる．ファクトチェックサイトの例として海外では Politifact[1] や Snopes[2]が存在し，日本では日本ファクトチェックセンター[3]などが挙げられる．ただしファクトチェックは専門知識と多大な労力を要するため，リアルタイムで処理することが難しいことが懸念点として挙げられる．そのため自然言語処理や機械学習の技術を用いてフェイクニュースを分類する研究，フェイクニュース検出が行われている．

フェイクニュース検出の困難な点としてフェイクニュースが巧妙なつくりをしていることが挙げられる．そのためニュース記事のテキストのみならず記事に対するソーシャルコンテキスト(コメント，リプライ，いいね)も用いることで分類精度の向上を示している[3-5]．ただしこれらのソーシャルコンテキストはユーザーの拡散によって生まれる情報のためニュースの初期段階ではあまり存在せず，早期検出に向かない点が挙げられる．そこでそのようなコメントなどのソーシャルコンテキストを生成してフェイクニュース検出をおこなう手法が存在する．

本研究では記事の書き手に着目した文章生成モデルを用いてニュース記事からコメント生成を行う．さらに書き手のユーザー情報の使用の有無によって，どちらのコメントを使えば精度が優れているかについて実験を行う．

2. 関連研究

Feng らは畳み込みニューラルネットワークと CVAE を組み合わせた TCNN-URG[6]というフェイクニュース検出モデルを提案した．これは畳み込みニューラルネットワークでニュース記事の情報抽出を行い，抽出した特徴から CVAE を用いてニュース記事に対するコメントとして尤もらしい単語を生成し，ニュース記事と生成単語を用いてフェイクニュース検出を行う．Weibo などの SNS のデータを用いてフェイクニュース検出を行ったところ，生成コメントを用いることでベースラインよりも高い正解率を示した．またフェイクニュースに対する上位 20 件の反応として “?”，“!”，“fake” など否定的な単語が生成されたことが確認された．

柳らは記事の残り部分から歯抜け部分を予測することで架空のニュース記事を作成する言語モデルの

[1] Politifact：https://www.politifact.com/
[2] Snopes：https://www.snopes.com/
[3] 日本ファクトチェックセンター：https://factcheckcenter.jp/

Grover[7]を用いてコメント生成によるフェイクニュース検出を行った[8]. 柳らはこのモデルを用いてニュース記事と実際に寄せられたコメントから新たにコメントを生成した. FakeNewsNet[9]のデータセットを用いて生成コメントの追加が分類精度に影響があるか実験を行ったところ, 実際のコメントとニュース記事のみを用いた場合の結果と比較して高い再現率を示した. またフェイクニュースに対する生成コメントの単語で“?”, “!”, “fake”のほかに“breaking:”も生成されたことが確認された.

3. 提案手法

本章ではニュース記事とニュース記事の書き手の情報を用いたコメント生成を用いたフェイクニュース検出について説明する.

3.1 生成モデル

先行研究では生成コメントをもちいることでベースラインよりも高い正解率を示した. ただしこれらの手法ではニュース記事のテキストのみを使用しており, それ以外の情報を生成の際に用いていない.

フェイクニュース検出においてユーザー情報を用いた手法がいくつか存在する. Raidipa らはニュース記事の出版社の信頼性スコアとニュースを共有するユーザーの信頼性スコアを用いた CSM[10]モデルを提案しベースラインよりも高い正解率と再現率を示した. また Ali らはニュース記事, 出版社の過去のニュース発表実績, ユーザーの影響度の強さを組み合わせた FR-Detect[11]モデルを提案しベースラインよりも高い正解率と F 値を示した. また Kai らはフェイクニュースを信頼, 共有するユーザーの特徴を分析し分類精度向上に繋がった[12,13]. 分析の結果 SNS のアカウントの登録日数, ユーザーの性格, フォロー数とフォロワー数の比が影響していることが確認された.

このようにフェイクニュースを投稿したユーザーや拡散したユーザーの情報を検出モデルに組み込むことで, テキストでは捉えられなかった特徴を捉え高い分類精度を示している. そこで提案手法としてニュース記事のテキストのみならずニュース記事のメディアや著者の情報も与えてコメント生成を行う, Seq2Seq with attention[14]と CVAE[15]を組み合わせた文章生成モデルを提案する. 更にその生成コメントをフェイクニュース検出モデルに用いることで, ニュース記事のみを用いて生成されたコメントを用いるよりも高い分類精度を示すのではないかと考えた. コメント生成モデルの概要を図 1 に示す. コメント生成の流れは以下のとおりである.

① ニュース記事を Seq2Seq モデルの Encoder である GRU[16]に入力して, 固定長のベクトルに変換する. また初期のコメントを別の GRU に入力してこちらも同様に固定長のベクトに変換する.

② ①で変換されたベクトル, 更にニュース記事の書き手のユーザー情報を CVAE に渡し, 固定長のベクトルを出力する. ここでユーザー情報は One-hot ベクトルで表現される.

③ ②で出力したベクトルと Attention によって作成されたc_tを用いて 1 単語ずつ単語を出力する. 出力された単語を順番に組み合わせて 1 つのコメントにする.

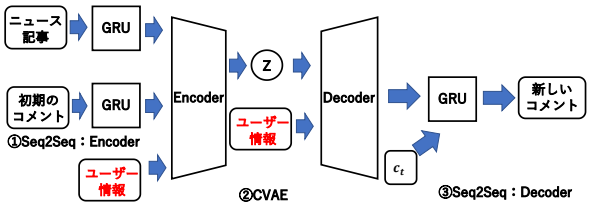


図 1 ：コメント生成モデルの概要

3.2 分類モデル

生成コメントを用いたフェイクニュース検出モデルの概要を図 2 に示す. 生成コメントを用いたフェイクニュース検出の流れは以下の通りである.

① ニュース記事と生成コメントをそれぞれの GRU の入力に渡し固定長のベクトルに変換して, 特徴抽出を行う.

② ①によって抽出されたニュース記事と生成コメントの特徴を結合して 1 つのベクトルに 変換する. その後ニューラルネットワークを用いて計算を行う.

③ ソフトマックス関数を用いて対象のニュース記事がフェイクニュースかどうか分類を行う.

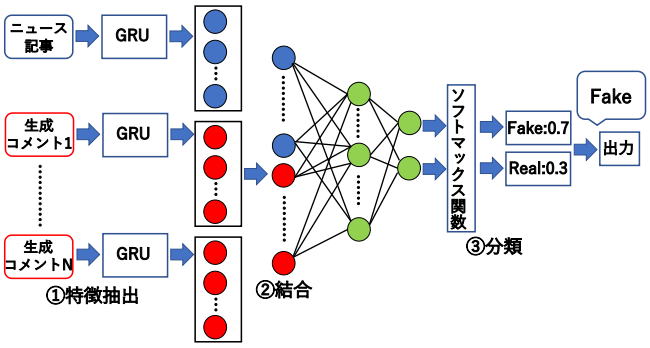


図 2：分類モデルの概要

3.3 処理手順

ここでは生成モデルと分類モデルを用いたフェイクニュース検出の処理手順について説明する. 処理手順をまとめたものを図 3 に示す. ここでは 1 つのニュース記事に対して 3 つのコメントを生成することを想定する.

① データセットの中から 3 件以上のコメントを含む

ニュース記事を抽出する．次に対象のニュース記事のコメントから古い順に 3 つ選び，残りのコメントを切り捨てる．

② ニュース記事とコメントを fastText[17]によって分散表現に変換する．その後生成モデルの訓練を行い，コメント生成モデルを訓練する．訓練したモデルを用いて，ニュース記事に対するコメントを 3 つ生成する．

③ ②でコメント生成が完了したら，②と同様に fastText を用いてニュース記事と生成コメントを分散表現に変換する．その後分類モデルの訓練を行い，分類モデルを訓練する．訓練したモデルを用いて，ニュース記事と生成コメントからフェイクニュースかどうか分類を行う．

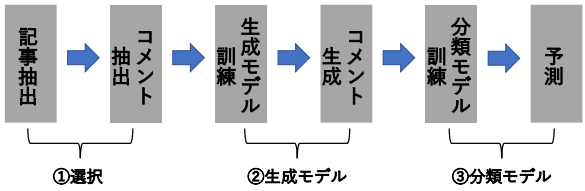


図 3：フェイクニュース検出の処理手順

4. 評価実験

本章では提案手法によって生成されたコメントを用いて分類精度の評価実験を行う．

4.1 使用データ

本研究ではコロナウイルスに関するニュース記事をまとめたデータセット ReCOVery[18]を用いる．ReCOVery は 2020 年の 1 月から 5 月までに発表されたコロナウイルスに関するニュース記事 2029 件とそれに反応した 140820 件のツイートから構成されたデータセットである．ニュース記事のラベルの内訳は 1364 件のリアルニュースと 665 件のフェイクニュースから構成されている．ReCOVery はニュースを発信したメディアの信頼性に基づいてフェイクニュースかどうかのラベル付けが行われている．本研究ではニュース記事の書き手に着目してコメント生成を行うため，ReCOVery でフェイクニュースとリアルニュースの違いを示す特徴の調査をおこなった．調査の結果，著者数と政治的偏見がフェイクニュースとリアルニュースの違いが見られた．政治的偏見には “Extreme left”, “Left”, “Left-center”, “Center”, “Right-center”, “Right”, “Extreme right”の 7 種類が存在する．調査の結果を図 4 から図 6 に示す．

図 4 を見ると著者数が明確な場合，著者数の数が増加するにつれリアルニュースの割合が増加することが見られる．また著者数が不明の場合，フェイクニュースのほうが割合が多いことが分かる．図 5 から“Left-center”や“Center”がほとんどを占めていることから，中立やリベラル派のメディアが ReCOVery データセット内で信頼性の高いメディアであることが分かる．逆に図 6 から“Right”や“Extreme right”がほとんどを占めていることから，保守派のメディアが ReCOVery データセット内で信頼性の低いメディアであることが分かる．

これらの結果からコメント生成のモデルにおいてユーザー情報は政治的偏見と著者数を One-hot ベクトルに変換して，コメント生成をおこなう．

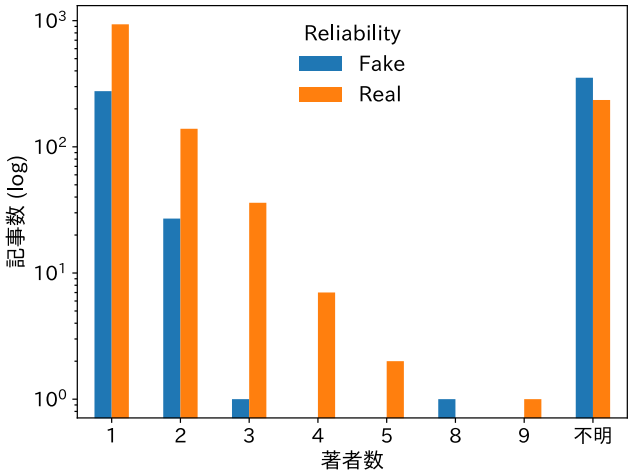


図 4：著者数ごとのフェイクニュースとリアルニュースの分布

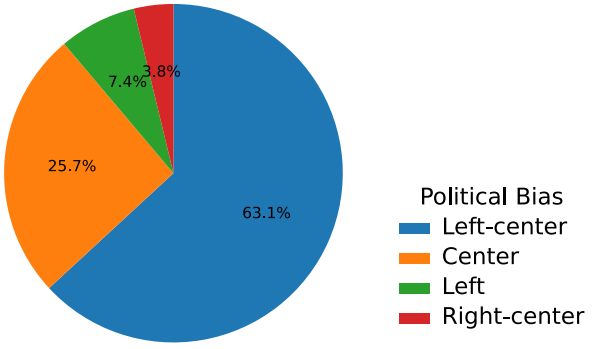


図 5：リアルニュースにおける政治的偏見の割合

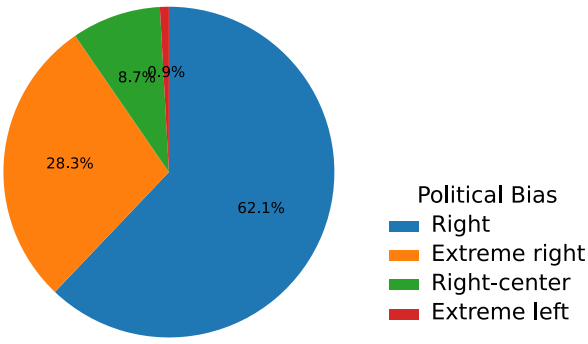


図 6：フェイクニュースにおける政治的偏見の割合

4.2 実験設定

本研究における実験設定について説明する．ユーザー情報の使用の有無によって，どちらのモデルから生成されたコメントを使えば精度が優れているか比較を行う．また実際のコメントも使用して，生成コメントを用いた場合との比較を行う．用いるユーザー情報は著者数のみ，政治的偏見のみ，著者数と政治的偏見の両方である．

4.2.1 ユーザー情報の区分

著者数をユーザー情報に用いる場合“不明”,“1 人”,“複数”のように人数ごとに区分をおこない，3 次元の One-hot ベクトルで表現する．

政治的偏見に関してはそれぞれのラベルごとに区分して 7 次元の One-hot ベクトルで表現する方法と，“リベラル”, “中立”, “保守”のように区分をおこない 3 次元の One-hot ベクトルで表現する 2 つの方法を用いる．ただし 3 次元の One-hot ベクトルの場合，“Left-center” と “Right-center” は両方の性質を持つためどこに区分されるかが曖昧となる．そのため 3 次元の One-hot ベクトルにおいては以下のように区分する．

・“Left-center”と“Right-center”を中立に区分する方法．以降“CtoC”と表現する．
・“Left-center”をリベラルに区分し，“Right-center”を保守に区分する方法．以降“CtoRL”と表現する．

政治的偏見と著者数の両方を用いる場合はそれぞれの One-hot ベクトルを連結してユーザー情報として表現する．

4.2.2 評価指標

分類精度の評価指標にはフェイクニュースを陽性とした場合の再現率と F 値を用いる．また本研究ではコメント数を変化させながら分類精度の変化を確認していく．例えばコメント数を 1 件から 5 件まで変化させる場合，図 2 のフェイクニュース検出モデルにおいて使用するコメントを増加させ対象のニュース記事がフェイクニュースかどうか分類する．そのため図 3 のフェイクニュース検出の処理手順から，使用するコメント数によってフェイクニュースとリアルニュースのラベル数が異なる．本研究ではコメント数を 4 件，10 件まで変化させる．表 1 に著者数をユーザー情報として用いる場合のコメント数ごとのラベル数を示す．

表 1：コメント数ごとのラベル数 (著者数)

	4 件	10 件
Fake	205	76
Real	926	607

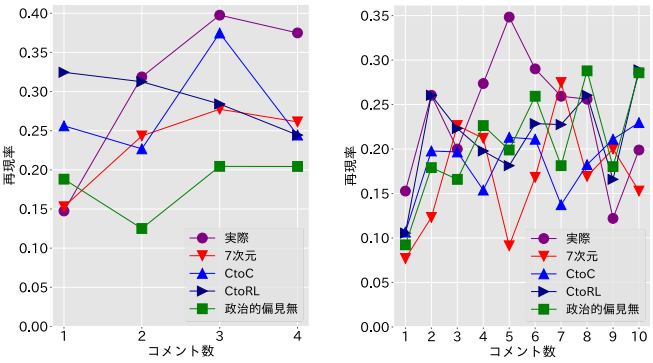
また表 2 に政治的偏見をユーザー情報として用いる場合のコメント数ごとのラベル数を示す．政治的偏見と著者数の両方をユーザー情報として用いる場合のラベル数は表 2 と同じである．

表 2：コメント数ごとのラベル数(政治的偏見)

	4 件	10 件
Fake	176	66
Real	927	607

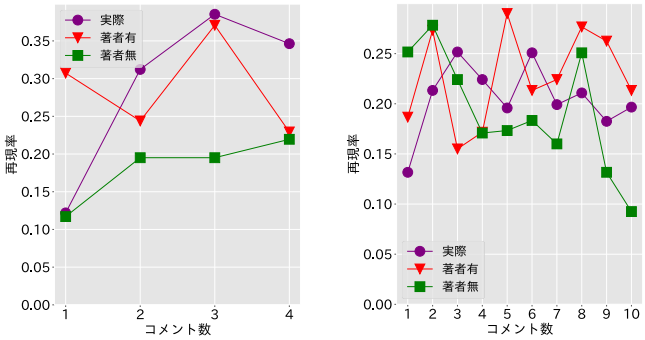
4.3 実験結果

コメント数を変化させた場合の再現率の変化を示す．ユーザー情報に著者数のみを用いた結果を図 7，ユーザー情報に政治的偏見のみを用いた結果を図 8，ユーザー情報に政治的偏見と著者数の両方を用いた結果を図 9 に示す．



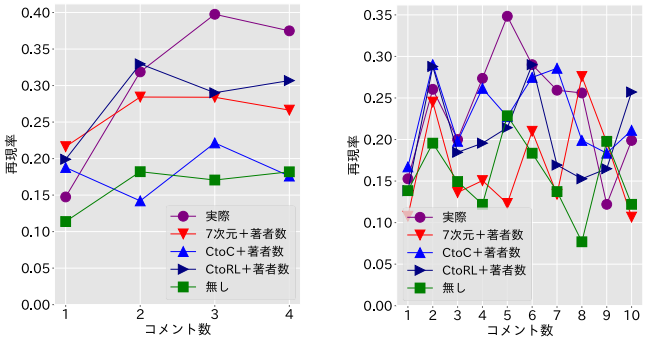
(a)4 件　　(b)10 件

図 7：再現率の変化(政治的偏見のみ)



(a)4 件　　(b)10 件

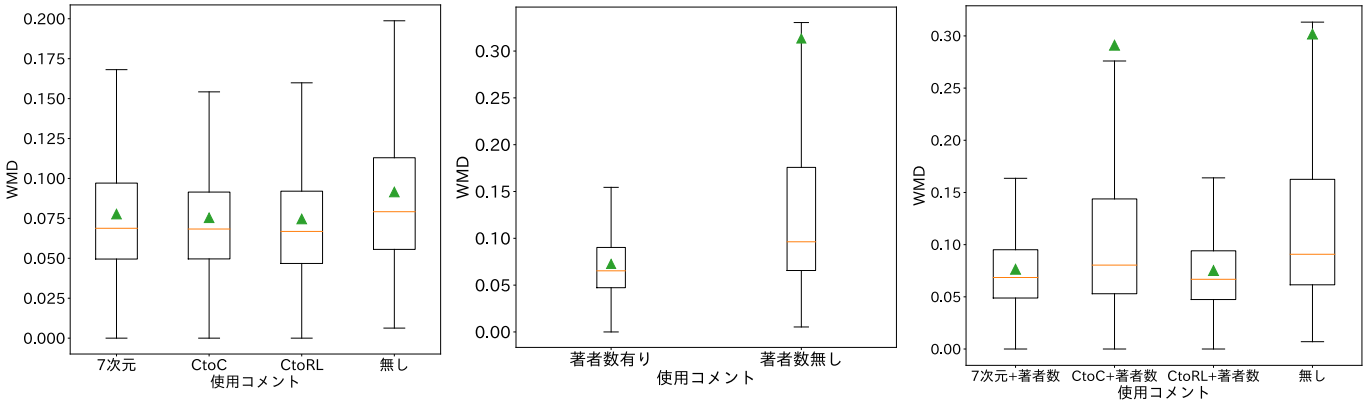
図 8：再現率の変化(著者数のみ)



(a)4 件　　(b)10 件

図 9：再現率の変化(政治的偏見と著者数)

コメント数が 4 件の場合，ほとんどの結果において



(a)政治的偏見のみ　(b)著者数　(c)政治的偏見と著者数

図 10：生成コメントごとの箱ひげ図

表 3：使用情報ごとの WMD の平均値と p 値

使用情報	区分	平均	p 値
政治的偏見のみ	無し	0.0915	-
	CtoC	0.0754	1.18×10^{-54}
	CtoRL	0.0746	9.17×10^{-59}
	7 次元	0.0777	1.88×10^{-39}
著者数のみ	無し	0.313	-
	有り	0.0726	1.16×10^{-256}
政治的偏見と著者数	無し	0.302	-
	CtoC	0.291	0.12
	CtoRL	0.0752	1.12×10^{-237}
	7 次元	0.0764	2.23×10^{-235}

ユーザー情報を用いた生成コメントを用いることでユーザー情報を用いなかった生成コメントよりも高い分類精度を示した.またコメント数が 10 件の場合は 4 件と同じような結果が得られなかった．これはフェイクニュースの数が少なく，評価指標が変化しやすいからと考えられる．F 値の変化は付録に示す．

コメント数が 4 件の場合の生成コメントと実際のコメントの類似度を WMD(Word Mover's Distance)[19]を使用して求めた．これは分散表現を用いて 2 つの文書間の距離を求める．そのためこの値が小さいほど実際のコメントと類似度が高いことを示す．それぞれの箱ひげ図を図 10 に示す．またユーザー情報の有無で WMD の平均に有意差があるか片側検定を行った．帰無仮説と対立仮説を以下に示す．

帰無仮説：ユーザー情報を用いたモデルから生成されたコメントの WMD の平均値はユーザー情法を用いなかったモデルから生成されたコメントの WMD の平均値と同じ．

対立仮説：ユーザー情報を用いたモデルから生成されたコメントの WMD の平均値はユーザー情法を用いなかったモデルから生成されたコメントの WMD の平均値とよりも小さい．

今回は大標本のため中心極限定理により平均の分布が正規分布に従うと仮定し，有意水準 5%で Z 検定を行った．p 値が 0.05 を下回った場合に帰無仮説が棄却され，WMD の平均に有意な差があるといえる．使用したユーザー情報ごとの WMD の平均値と p 値を表 3 に示す．政治的偏見と著者数を用いた場合の CtoC を除くと Z 検定を行った際の p 値が 0.05 を下回っていた．そのためほとんどのユーザー情報を用いた生成コメントはユーザー情報を用いなかった生成コメントと WMD の平均に有意な差があるといえる．

また生成コメントを抽出しユーザー情報の有無による違いを確認した．生成コメントの例を図 11 に示す．これはワシントン州の受刑者がコロナウイルスに感染したニュースに対するコメントである．

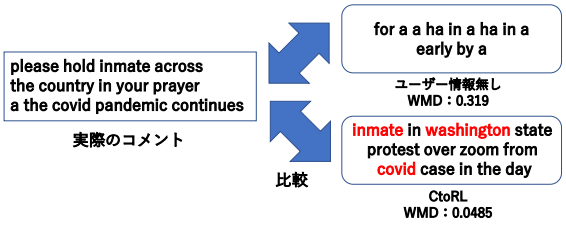


図 11：生成コメントの例

ユーザー情報を用いなかったモデルから生成されたコメントは特定の単語が偏って出力されている．一方ユーザー情報を用いたモデルから生成されたコメン

トは“inmate”，“washington”，“covid”などニュースに関連する単語が出力されていることが確認できる．

5. 考察

4.3 節のコメント数が 4 件の場合の実験結果から実際のコメントを用いることで高い分類精度を示した．そのため表 3 の結果より，実際のコメントと類似度が高いユーザー情報を用いた生成コメントのほうがユーザー情報を用いなかった生成コメントよりも分類精度が高かったと考えられる．また実際のコメントに近いコメントを生成できたのはユーザー情報を用いることで，コメント生成の際に CVAE の部分で取得したいコメントの状態を指定したからだと考えられる．

次に政治的偏見や著者数がフェイクニュースに影響した理由を考える．SNS に流れるフェイクニュースは保守層の人間は信じやすく拡散しやすいという傾向がある[20,21]．また保守層の人々は保守寄りのメディアに接触する傾向がある[22]．これらのことから政治的偏見が保守寄りのメディアから流れたニュース記事はフェイクニュースが拡散されやすいため，フェイクニュースが多かったと考えられる．Niraj らは FakeNewsNet のデータを用いて共著者ネットワークを構築し，著者数とフェイクニュースの関係を調査した[23]．調査の結果ニュース記事の著者数と信頼性は相関関係があることが分かり，複数の著者が存在するニュース記事は信頼性が高く，著者の名前が記されていないニュース記事は信頼性が低いことを示した．これは著者関係の同類性からリアルニュースに従事している著者は多くの人と協力し，フェイクニュースに従事している著者と協力する可能性が低いからだと考えている[24]．これらのことから政治的偏見と著者数がフェイクニュースに影響したと考えられる．

6. おわりに

本研究では Seq2Seq with Attention と CVAE を組み合わせた記事の書き手に着目したコメント生成によるフェイクニュース検出を提案した．コメント数が 4 件の場合，ユーザー情報を用いたコメントはユーザー情報を用いなかったコメントよりも高い分類精度を示した．これは生成モデルの CVAE の部分で取得したいコメントの状態を指定したため，実際のコメントに近いコメントを生成できたからだと考えられる．今後の課題として 異なるユーザー情報の利用，One-hot 表現以外の数値変換などが挙げられる．

参 考 文 献

[1] Claire Wardle et al. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Marking. Council of Europe report，2017.

[2] 藤代裕之．フェイクニュースの生態系．青引社，2021.

[3] Eugenio Tacchini et al. Some Like it Hoax:Automated Fake News Detection in Social Networks. arXiv preprint arXiv:1704.07506，2017.

[4] 古川凌也 et al. SNS ユーザの自己紹介文にあらわれる偏りに着目したフェイクニュース検知．FIT2021(第 20 回情報科学技術フォーラム), pp.35-42，2021.

[5] Li Dun et al. Unsupervised Fake News Detection Based on Autoencoder. IEEE Access, Vol.9, pp.29356-29365，2021.

[6] Feng Qian et al. Neural User Response Generator Fake News Detection with Collective User Intelligence. IJCAI-18，pp.3834-3840，2018.

[7] Rowan Zellers et al. Defending Against Neural Fake News. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp.9054-9065，2019.

[8] Yuta Yanagi et al. Fake News Detection with Comments for News Articles. IEEE 24th International Conference on Intelligent Engineering Systems，pp. 85-90，2020.

[9] Kai Shu et al. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. arXiv preprint arXiv:1809.01286，2018.

[10] Rajdipa Chowdhury et al. Joint Estimation of User And Publisher Credibility for Fake News Detection. CIKM'20:Proceedings of the 29th ACM International Conference on Information & Knowledge Management，pp.1993-1996，2020.

[11] Ali Jarrahi et al. FR-Detect: A Multi-Modal Framework for Early Fake News Detection on Social Media Using Publishers Features. arXiv preprint arXiv : 2019.04835，2021.

[12] Kai Shu et al. The Role of User Profiles for Fake News Detection. ASONAM'19:Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining，pp.436-439，2019.

[13] Kai Shu et al. Understanding User Profiles on Social Media for Fake News Detection. IEEE Conference on Multimedia Information Processing and Retrieval，pp.430-435，2018.

[14] Thang Luong et al. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing，pp.1412-1421，2015.

[15] Kihyuk Sohn et al. Learning Structured Output Representation using Deep Conditional Generative Models. In Advances in Neural Information Processing Systems，volume 28，2015.

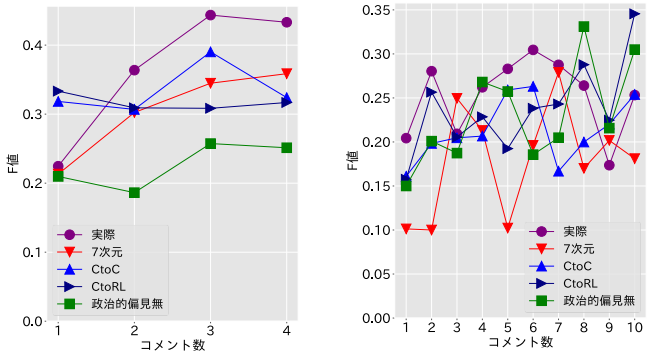
[16] Kyunghyun Cho et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing，pp.1724-1734，2014.

[17] Piotr Bojanowski et al. Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606，2016.

[18] Xinyi et al. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research. CIKM'20 : Proceedings of the 29th ACM International

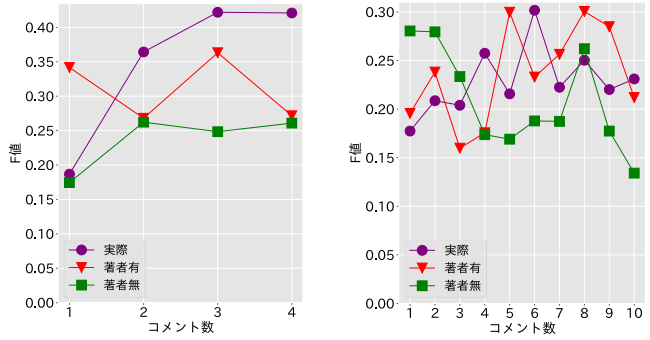
Conference on Information & Knowledge Management, pp.3205-3212, 2020.
[19] Matt Kusner et al. From Word Embeddings To Document Distances. PMLR Vol.37, pp.957-966, 2015.
[20] Baptista et al. Understanding fake news consumption: A Review. Social Sciences, 9(10) : 185, 2020.
[21] Baptista et al. Who believes in fake news? Identification of political(A) symmetries. Social Sciences 11(10) : 460, 2022.
[22] 山口真一. ソーシャルメディア解体全書, 勁草書房, 2022.
[23] Niraj Sitaula et al. Credibility-based Fake News Detection. arXiv preprint arXiv : 1911.00643, 2019.
[24] Reza Zafarani et al. Social media mining:An introduction. Cambridge University Press, 2014.

付録



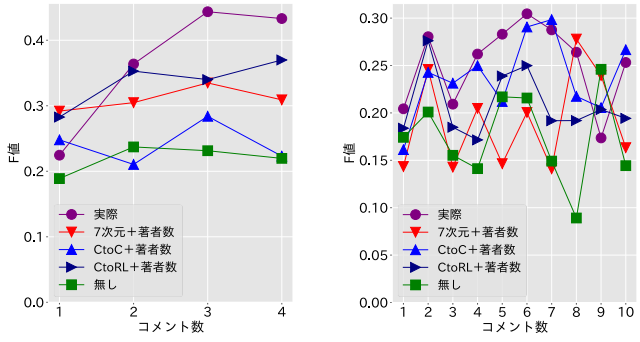
(a)4 件　　(b)10 件

付録 A：F 値の変化(政治的偏見のみ)



(a)4 件　　(b)10 件

付録 B：F 値の変化(著者数のみ)



(a)4 件　　(b)10 件

付録 C：F 値の変化(政治的偏見と著者数)

ニュースの構成要素に着目したフェイクニュースのフェイク部分の抽出・訂正手法

末吉　将也†　　灘本　明代††

† 甲南大学大学院自然科学研究科知能情報学専攻　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
†† 甲南大学知能情報学部知能情報学科　〒 658–8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: †m2324003@s.konan-u.ac.jp, ††nadamoto@konan-u.ac.jp

あらまし　現在インターネット上には様々なフェイクニュースがある．フェイクニュースは全く内容が異なるフェイクニュースと，部分的にフェイク部分があるフェイクニュースの 2 種類ある．本論文では，後者の部分的にフェイク部分が含まれているフェイクニュースを対象とし，そのフェイク部分の抽出と訂正手法を提案する．具体的には，ニュースの主体となる単語（以下，主人公と呼ぶ）とその主人公に対する「いつ」,「どこで」,「何をしたか」(以下，ニュース構成要素) をそれぞれルールベース及び BERT を用いて抽出する．次に，その主人公及びニュース構成要素を用いて対象のフェイクニュースに対する正しいニュース群を抽出し，フェイクニュースと比較する．そしてフェイクニュースのニュース構成要素と正しいニュースのニュース構成要素を比較することにより，フェイクニュースのフェイク部分とその訂正内容の抽出を行う．

キーワード　ニュース・マスコミ，ソーシャルメディア

1　はじめに

インターネット上には，毎日種々のニュースが多数発信されている．そのニュースの中には，フェイクニュースである事実とは異なる内容を含むニュースが存在する．このフェイクニュースにはニュースの内容全てが間違いであるフェイクニュースと，部分的に間違いであるフェイクニュースがある．例えば,「イタリアでは，ワクチンが全ての牛を殺した」は全てが間違いであるフェイクニュースである．一方,「DEIM2024 は岐阜県で 3 月に開催される」は部分的に間違いであるフェイクニュースである．全てが間違いの内容であるフェイクニュースは比較的に間違いであることに気づくことができやすいが，部分的に間違いであるフェイクニュースは間違いに気づくことが難しいと考えられる．そのため，何気なく部分的に間違いであるフェイクニュースを読み，その内容について詳しく調べずにそのフェイクニュースを信じる恐れがある．また，そのフェイクニュースを SNS などで発信することにより，そのフェイクニュースが拡散されてしまう恐れもある．これらの問題点の解決策として，フェイクニュースに対してファクトチェックを行い，間違いの内容を訂正することが行われ，ファクトチェックサイト [1]にて公開されている．しかしながら，ニュースは日々発信されていることからその数は膨大となっている．そのため，全てのニュースに対して人手でファクトチェックを行い，フェイク部分を見つけその訂正を行うことは困難であると考える．そこで，本研究では自動でフェイク部分を抽出し訂正する手法を提案する．

本研究では部分的に間違っているフェイクニュースを対象とする．そして，ニュース記事を作る際に重要な要素である「誰が（何が）」,「いつ」,「どこで」,「何をしたか」[2]に着目する．ここで我々は，ニュース記事において「誰が（何が）」が特に重要であると考える．本研究では，ニュース中の「誰が（何が）」を主人公と呼び，そのニュースの主人公に着目する．さらに，その主人公に対する「いつ」,「どこで」,「何をしたか」にも着目し，フェイクニュースのフェイク部分の抽出とその訂正を行う．本論文では，この主人公に対する「いつ」,「どこで」,「何をしたか」をニュース構成要素と呼ぶ．これまで我々はルールベースを用いて，フェイクニュースを正しい内容にすることを目的とし，フェイクニュースのフェイク部分の抽出をし，その訂正を行っている. [1] ここでは,「どこで」「何をしたか」の精度が悪いことがわかった．そこで本論文では，この「どこで」と「何をしたか」の抽出精度の向上を目指して，新たにディープラーニングを用いてフェイクニュースのフェイク部分を抽出し，その訂正を行う手法の提案を行う．

以下と図 1 にフェイクニュースのフェイク部分の抽出とその訂正の手順を示す．

（1）　対象のフェイクニュースの主人公を抽出する．

（2）　対象のフェイクニュースのニュース構成要素を抽出する．

（3）　(1) で抽出した対象のフェイクニュースの主人公と同じ主人公を持つニュースをニュースサイトから抽出する．

（4）　(3) で抽出したニュース群から (2) で抽出したニュース構成要素の一部を含むニュースを抽出し，それを正しいニュースとする．

（5）　(4) で抽出した正しいニュースからニュース構成要素を抽出する．

（6）　(2) と (5) のニュース構成要素を比較することにより，

1：https://infact.press

2：http://www.at-s.com/blogs/nie/study/howto.html

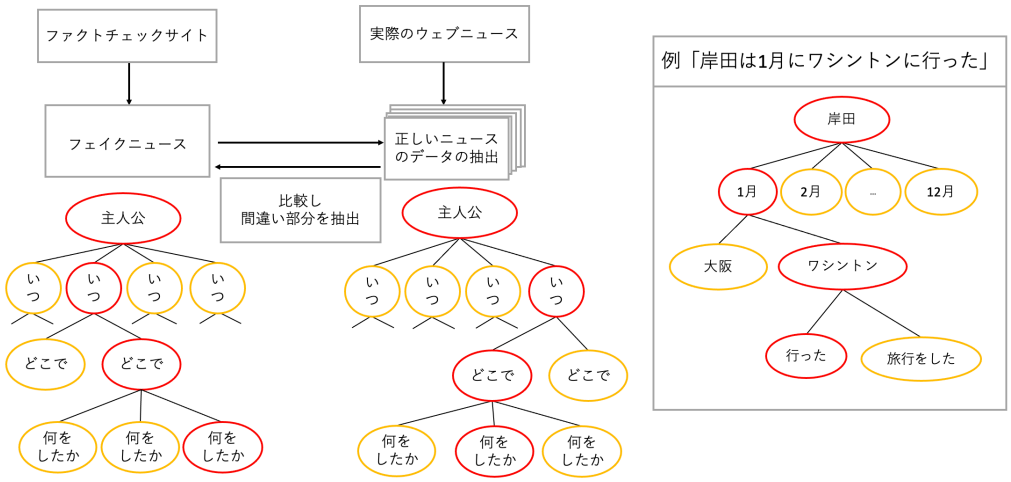


図 1　フェイクニュース抽出のイメージ図

フェイクニュースのフェイク部分とその訂正内容を抽出する．以下，2 章で関連研究について述べ，3 章で本研究での対象のフェイクニュースの定義を行う．4 章では提案手法によるフェイクニュース記事のフェイク部分とその訂正内容の抽出について述べる．5 章では提案手法でのフェイク部分とその訂正内容の抽出に対する評価を行い，結果とその考察について述べる．最後に，6 章で本論文のまとめと今後の課題について述べる．

2　関 連 研 究

フェイクニュースに対して種々の研究が行われている．

渡邉 [3] らは，地震や台風などの出来事に対して誤情報に多く使われているキーワードからツイートを収集し，その誤情報の支持・拡散ツイート，誤情報の反論・訂正ツイートに教師あり学習を用いて分類している．中原 [4] らは，Twitter においてデマかどうかをユーザが判断することを支援するための基準として，非公式リツイートに付加されたコメントをリツイートしたツイートの意見として抽出し，抽出したコメントに含まれる単語によって「危険度」と「訂正ツイート」の提示を提案している．梅島 [5] らは，間違った情報 (デマ) の拡散を防止するため， Twitter 上におけるデマツイートとデマの訂正ツイートの拡散に関してポジティブ，ネガティブなどの評価属性をもとに評価し，分析を行なっている．岡山 [6] らは，フェイクニュースと政治のニュースから学習を行ったフェイクニュース分類器を作成し，Yelp[3]のレビューに対してフェイクかどうかの分析をしている．これらの研究ではフェイクニュースに対して反論や訂正ツイートの抽出や評価属性や言語的な特徴をもとに分析を行なっているのに対し，本研究ではフェイクニュースに対してフェイク部分を抽出することを行なっている点が異なっている．

また，Web 上の情報から日時や場所を表す語の抽出に関する研究も種々行われている．斎藤ら [7] は，機械学習として CRF を用いてタグ付けされた語句を学習することにより，イベントについて記載された記事からイベント名や開催日時などの情報の抽出を行っている．また，大沼ら [8] は，固有表現抽出の技術を用いることで，イベント名や開催日時，開催地などの属性の抽出を行っている．それらの研究に対し，本研究では深層学習である BERT のマスク言語モデルを用いることで，ニュース中の要素を抽出している点が異なる．

3　フェイクニュースの定義

フェイクニュースは様々な研究によって種々定義されている．Lazer [9] らは「フェイクニュースとはニュースメディアのコンテンツを形式的には模倣するが，組織的な考えや意図は模倣しない物である」と定義している．Zhang [10] らは「フェイクニュースとは主にインターネット上で公開・配信され，金融，政治，その他の利益のために意図的に読者を誤解させ，騙し，誘い込む為のあらゆる種類の情報やストーリーのこと」と定義している．本研究を行うにあたり，我々はファクトチェックの推進や普及に取り組む団体である FIJ[4]によりニュースをレーティングされたものを参考に，本研究でのフェイクニュースを定義する．本研究では以下に示すように FIJ のレーティングの「ほぼ正確」と「誤り」,「虚偽」,「不正確」をフェイクニュース」とする．

- ほぼ正確

一部は不正確だが，主要な部分・根幹に誤りはない．

- 誤り

全て，もしくは根幹部分に事実の誤りが存在する．

- 虚偽

全て，もしくは根幹部分に事実の誤りが存在し，事実でないと知りながら伝えた疑いが濃厚である．

- 不正確

正確な部分と不正確な部分が混じっていて，全体として正確性が欠如している．

例えば,「大谷翔平は 2023 年にカブスに移籍した」や「WBC 第 5 回大会は 2022 年にフランスで開催された」などのニュー

3：https://www.yelp.co.jp/

4：https://fij.info/introduction/guideline

スは全体として正確性が欠如されているため「不正確」に分類される．

つまりは本論文では，捏造されたニュースや真実のニュースを改ざんしているフェイクニュースに書き換えたもの，単に不正確なニュースなども含めて，悪意の有無に関わらず，誤字や脱字のような間違いを含むニュース全てをフェイクニュースと定義する．

4 フェイクニュースの訂正手法

フェイクニュースの訂正を行うために，まず初めにフェイクニュースからそのニュースの主人公を抽出する．そして，そのフェイクニュースの主人公に対するニュース構成要素を抽出する．そして，抽出した主人公とニュース構成要素からフェイクニュースに関連する正しいニュースを抽出する．また，抽出した正しいニュースからニュース構成要素を抽出する．最後に，フェイクニュースのニュース構成要素と正しいニュースのニュース構成要素を比較することにより，フェイクニュースのフェイク部分とその訂正内容を抽出する．

4.1 対象のフェイクニュースの主人公の決定方法

まず初めに，フェイクニュースからニュースの主人公の抽出を行う．ここで，ニュースの主人公には一般的に人だけでなく，組織や物事である場合もあるため，人名だけでなく，名詞全般を対象とする．また本研究では，ニュースの構成上，第一段落では最も重要な事柄が書かれており，段落が進むにつれて重要度が落ちてゆくことから，ニュースの第一段落に着目する．[5] 主人公の抽出手順を以下に示す．

（1） タイトルに人物名がある場合，その人物名をそのニュースの「主人公候補」とする．

（2） 「主人公候補」が一人の場合，その「主人公候補」を「主人公」とする．

（3） 「主人公候補」が複数ある場合,「主人公候補」でかつそのニュースの第一段落で最も出現数の多い人物名を「主人公」とする．

（4） 「主人公候補」がない場合，タイトルにある名詞を「主人公候補」とする．

（5） 「主人公候補」が一つの場合，その「主人公候補」を「主人公」とする．

（6） 「主人公候補」が複数ある場合,「主人公候補」でかつそのニュースの第一段落で最も多い名詞を「主人公」とする．

4.2 対象のフェイクニュースのニュース構成要素の決定方法

次に，抽出した主人公に対するニュース構成要素を抽出する．ニュース構成要素の抽出は，BERT のマスク言語モデルを用いる．ここで，これまでの我々の研究 [2] で，マスク言語モデルにて「いつ」の抽出精度が低かったため,「いつ」に関してはルールベースを用いて抽出する．

ニュース構成要素「いつ」の抽出

抽出した文それぞれに対して Mecab[6]にて形態素解析を行い，ニュース構成要素である「いつ」をそれぞれの文から抽出する．「いつ」は時間を表しているため,「数」＋「年」や「数」＋「月」,「数」＋「日」を「いつ」として抽出する．

ニュース構成要素「どこで」と「何をしたか」の抽出

ニュース構成要素である「どこで」と「何をしたか」は BERT を用いて抽出する．本論文では BERT の日本語の事前学習モデルの一つである bert-base-japanese-whole-word-masking[7]を用いる．また，本研究では，ニュース記事を対象としている．そのため，2018 年から 2023 年までの毎日新聞 [8]からニュース記事をランダムに 10,000 件抽出し，この記事を用いて，ファインチューニングを行う．また，BERT の各種パラメータはグリッドサーチにて決定しており，バッチサイズは 32，ラーニングレートは 0.00002，オプティマイザは Adam を用いている．また，学習する際に「どこで」の部分のみを人手にてマスクをして学習したモデルを「どこでモデル」,「何をしたか」の部分のみを人手にてマスクをして学習したモデルを「何をしたかモデル」の 2 つのモデルを生成する．

4.3 対象のフェイクニュースに対する正しいニュースの抽出方法

本研究では，毎日新聞と朝日新聞 [9]と読売新聞 [10]に掲載されているニュースは正しいニュースであるとみなし，これら 3 社に掲載されているニュースを正しいニュースとして用いる．しかしながら，これらのニュースにはまれにフェイクニュースが紛れ込んでしまっている場合がある．そのため本研究では，これら 3 社の新聞記事に対してニュース構成要素を抽出し，このニュース構成要素の比較を行い，3 社すべてにおいて同じニュース構成要素を正しいニュース構成要素とする．この正しいニュース構成要素を持つニュースを正しいニュースとする．具体的な手順を以下に示す．

（1） 対象のフェイクニュースの主人公とニュース構成要素を抽出する．

（2） 毎日新聞と朝日新聞，読売新聞から，対象のフェイクニュースと同じ主人公とニュース構成要素のうちの一つを含むニュース記事を抽出する．

（3） 抽出した 3 社の新聞記事のニュース構成要素の比較し，3 社すべてにおいて同じニュース構成要素を持つニュースを正しいニュースとする．

例えば，フェイクニュースの主人公が「大谷翔平」であり，ニュース構成要素の 1 組が「2023 年 12 月に」,「オーストラリアで」,「ホームランを打った」の時には，毎日新聞と朝日新聞，読売新聞から「大谷翔平」と「2023 年 12 月に」もしくは「大谷翔平」と「オーストラリアで」,「大谷翔平」と「ホームラン

5：http://www.at-s.com/blogs/nie/study/howto.html

6：https://taku910.github.io/mecab/

7：https://huggingface.co/cl-tohoku/ bert-base-japanese-whole-word-masking

8：https://mainichi.jp

9：https://www.asahi.com

10：https://www.yomiuri.co.jp

表 1　ニュース構成要素「いつ」の実験結果

番号	ニュース概要	抽出したいつ	正解のいつ
(1)	3 月 15 日に参議院予算委員会が開かれた。	3 月 15 日	3 月 15 日
(2)	2021 年の衆議院選挙では、メディアの選挙予測が外れたことが話題になった。	2021 年	2021 年
(3)	照井は「2 年分の走りができた」と胸を張った。	2 年	NULL
(4)	WBC2023 で日本は 14 年ぶりに優勝した。	14 年	2023

表 2　ニュース構成要素「どこで」の実験結果

番号	ニュース概要	抽出したどこで	正解のどこで
(1)	愛媛県で高病原性鳥インフルエンザに感染した 疑いのある鶏が見つかったと発表した	愛媛県で	愛媛県で
(2)	音楽業界ではストリーミング配信の利用拡大で楽曲所有権の価値が上昇しているとされる。	音楽業界で	音楽業界で
(3)	会見で，吉村知事は，予備選が知名度を反映した 「人気投票」になる恐れがある点について問われた。	点に	会見で
(4)	大谷翔平投手が 16 日に 2 度目となるア・リーグ MVP に、満票で輝いた。	満票で	NULL

表 3　ニュース構成要素「何したか」の実験結果

番号	ニュース概要	抽出した何をしたか	正解の何をしたか
(1)	沖縄県は 5 日、新たに 623 人の新型コロナウイルスの感染者を確認した。	感染者を確認した	感染者を確認した
(2)	チームは「スーパームーンの影響を考慮した海岸管理が必要だ」と指摘する。	指摘する	指摘する
(3)	五輪出場を目指す高谷惣亮が昨年 12 月のレスリング 全日本選手権男子フリースタイル 92 キロ級で優勝した。	優勝した	五輪出場を目指す 優勝した
(4)	河瀬直美はオペラ演出に初挑戦をして、１２月に沖縄で千秋楽を迎えた。	千秋楽を迎えた	初挑戦をして 千秋楽を迎えた

を打った」を含むニュースをそれぞれ抽出する．その後，抽出したニュースに対して，フェイクニュースと同様にニュース構成要素を抽出する．最後に，抽出したニュース構成要素をそれぞれ比較することで正しい内容のニュースを抽出する．

4.4　対象のフェイクニュースに対するフェイク部分と訂正の抽出方法

対象のフェイクニュースのニュース構成要素と正しいニュースのニュース構成要素を比較することにより，フェイクニュースのフェイク部分と訂正内容の抽出を行う．

- ニュース構成要素が全て異なる場合

フェイクニュースと正しいニュースにて全て異なるニュース構成要素を持つ場合には，比較した対象のフェイクニュースと正しいニュースでは，異なるトピックが書かれていることを示す．

- ニュース構成要素が部分的に異なる場合

フェイクニュースと正しいニュースのニュース構成要素を比較する．このとき，フェイクニュースのうち正しいニュースのニュース構成要素の異なる箇所がフェイク部分である．また，この時の正しいニュースのニュース構成要素が訂正の内容とする．

例えば,「主人公」が大谷翔平でフェイクニュースのニュース構成要素が {2020 年に，WBC で，MVP を取得した } であり，正しいニュースのニュース構成要素が {2023 年に，WBC で，MVP を取得した } であるとき，フェイク部分は「2020 年に」となり，訂正の内容が「2023 年に」となる．

このようにして，フェイクニュースのフェイク部分とその訂正内容を抽出する．

5　実　　験

ニュース構成要素の抽出手法およびフェイク部分の訂正手法の有用性を示す 2 つの評価実験を行った．

5.1　ニュース構成要素の抽出における実験

実験方法

ニュース構成要素が抽出できているかの実験には，実際のニュースを用いる．2018 年から 2023 年までの毎日新聞よりニュース記事をランダムに抽出した 10,000 件のニュース記事に対して，抽出したニュース構成要素それぞれの適合率，再現率，F 値を求める．

結果と考察

結果は表 4 に示す．

表 4　実験の結果

	適合率	再現率	F 値
いつ	0.852	0.723	0.782
どこで	0.855	0.832	0.843
何をしたか	0.846	0.866	0.856

また，表 1，2，3 に抽出したニュース構成要素のそれぞれの結果を示す．

表 1 の (1)，(2) に示すように,「いつ」の実験では，日付が記入されている箇所に関しては抽出することができた．しかしながら，表 1 の (3) や (4) などの「2 年分」や「14 年ぶり」といった相対的な年数や「WBC2023」といった年数を間接的に

表 5　フェイク部分抽出の実験結果

番号	フェイクニュース	抽出したフェイク部分	抽出した訂正の内容	正解のフェイク部分	正解の訂正の内容
(1)	大谷翔平は 2019 年、第 5 回ワールド・ベースボール・クラシック（WBC）の最優秀選手（MVP）を受賞した。	2019 年	2023 年	2019 年	2023 年
(2)	大谷翔平は宮城県の花巻東高校で挑んだ甲子園に一度も勝てなかった。	宮城県	岩手県	宮城県	岩手県
(3)	大谷翔平は 7 月のホームランダービーで優勝した。	優勝した	出場した	優勝した	出場した
(4)	大谷翔平は 2021 年にホームラン王のタイトルを獲得した。	2021 年	2023 年	2021 年	2023 年
(5)	大谷翔平は 2018 年に 2 回目のトミージョン手術を受けた。	NULL	NULL	2018 年	2023 年
(6)	大谷翔平は 6 月 28 日のホワイトソックス戦でホームランを 2 本放った	6 月 28 日	6 月 27 日	NULL	NULL

示した内容は抽出できなかった．この点は，今後の課題である．

表 2 の (1)，(2) に示すように,「どこで」の実験では，地域名などの固有名詞の箇所に関しては抽出することができた．また，(2) のような地域名以外の場所や範囲を表す語，例えば「SNS で」や「国会で」などの語を抽出することもできた．しかしながら，表 2 の (3)，(4) の場合では「どこで」を抽出するモデルの学習の際に「名詞+で」や「名詞+に」を一つの語として学習しているため,「名詞＋で」や「名詞+に」となっている語を全てを「どこで」として誤って抽出した．この点は今後の課題である．

表 3 の (1)，(2) に示すように,「何をしたか」の実験では，表 3 の (1)，(2) のように文中に主人公に対する「何をしたか」が一つのみある場合には正しく抽出することができた．しかしながら，表 3 の (3) や (4) のように「主人公」に対する「何をしたか」が二つ以上あるときには一つのみ抽出でき，他の「何をしたか」が抽出できなかった．この点は今後の課題である．

5.2　フェイクニュースの間違い部分訂正における実験

実験の目的として，提案手法にて正しくフェイクニュースのフェイク部分の抽出ができているかを調べるためのものである．

実験方法

ファクトチェックにてフェイクと診断されたフェイクニュースが少ない為，ChatGPT4 を用いて実験用にフェイクニュースを生成した．生成したフェイクニュースは 6,000 件である．このフェイクニュースに対して，提案手法で得たフェイクニュースのフェイク部分の抽出の適合率，再現率，F 値を求める．またこのとき，ChatGPT4 に入力したプロンプトは 200 種類である．その一部の例を表 6 に示す．

また，正しいニュースとして，毎日新聞，朝日新聞，読売新聞の 2018 年から 2023 年までの 412,342 件のニュース記事を使用する．

結果と考察

結果を表 7 に示す．また，フェイク部分の抽出の実験結果を表 5 に示す．(1)(2)(3)(4) では，正しくフェイク部分や訂正の内容の抽出ができた．しかしながら，(3) ではフェイクでないと誤って判断された部分が,「いつ」の部分が誤りであるフェイクニュースである．これは，大谷翔平がトミージョン手術を受けたのは 2018 年と 2023 年であるが,「2 回目」という状況が本研究では考慮されていないことが原因である．この点は今後の課題である．また，(4) では，フェイクでない一文に対して，フェイクであると抽出された．これは，海外時間で書かれたものや日本時間で書かれたものがあることから，生じた．その他表記揺れに関しても対応することが今後の課題である．

表 6　入力したプロンプト例

番号	入力したプロンプト
(1)	以下は、ニュースの内容を説明する指示です。 要求を適切にみたすニュースをかきなさい。 ###指示: 主人公：大谷翔平 日時：2023 年 場所：アメリカ ホームランを放った
(2)	以下は、ニュースの内容を説明する指示です。 要求を適切にみたすニュースをかきなさい。 ###指示: 主人公：大谷翔平 日時:2023 年 場所:WBC MVP を受賞した
(3)	以下は、ニュースの内容を説明する指示です。 要求を適切にみたすニュースをかきなさい。 ###指示: 主人公: 大谷翔平 日時:7 月 場所: デンバー ホームランダービーに出場した

表 7　実験の結果

	適合率	再現率	F 値
提案手法	0.780	0.752	0.766

6　まとめと今後の課題

本研究は，ニュース構成要素に基づいて，フェイク部分を抽出し，その訂正を行うことである．本論文では，BERT のマスク言語モデルを用いて，ニュース構成要素を抽出する方法を提案した．提案したモデルにて，フェイクニュースのフェイク部分とその訂正の内容の抽出を行った．今後の課題として，表記揺れに対応することや固有表現抽出や係り受け解析を行うことにより，文脈の理解を深め，精度を向上することである．

謝　　辞

論文の一部は私学助成金（大学間連携研究助成金）の助成によるものである．ここに記して謹んで感謝の意を表する．

文　　献

[1] 末吉将也, 灘本明代, “ フェイクニュース記事の誤り箇所抽出手法の提案 ”, 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM2023), pp. 1-7,2023.
[2] ”Extraction News Components Methods for Fake News Correction ”, The 25th International Conference on Information Integration and Web Intelligence (iiWAS2023), 2023
[3] 渡邉研斗, 鍋島啓太, 岡崎直観, 乾健太郎, “ Twitter 上での誤情報と訂正情報の自動分類. ” 言語処理学会第 19 回年次大会, 2013.
[4] 中原英美, 冨永一成, 牛尼剛聡 “ リツート構造を用いたデマ拡散防止支援手法 ” 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), pp. 1-7, 2012.
[5] 梅島彩奈, 宮部真衣, 荒牧英治, 灘本明代, “ 災害時 Twitter におけるデマとデマ訂正 RT の傾向. ” 研究報告情報基礎とアクセス技術, Vol. 2011, No.4, pp. 1-6, 2011.
[6] 岡山光平, 石川博, 廣田雅春, “ フェイクニュース分類機を用いた口コミサイトのレビューの分析. ” 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), pp. 1-5, 2018.
[7] 斉藤隆太, 石野亜耶, 難波英嗣, 竹澤寿幸. (2012). ”新聞記事と Web からのイベント情報の自動抽出”. 第 5 回 Web とデータベースに関するフォーラム (WebDB Forum).
[8] 大沼宏行, 松平正樹, 渕上正睦, 森田幸伯. (2003). ”Web コンテンツの分析に基づくオントロジ構築および属性抽出の試み”. 情報処理学会研究報告情報学基礎 (FI), 2003(98 (2003-FI-072)), 49-54.
[9] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. ”The science of fake news.” Science 359, 6380 (2018), pp. 1094-1096.
[10] XichenZhang and AliAGhorbani, ”An overview of online fake news : Characterization,detection,and discussion.”, Information Processing & Management Volume 57, Issue 2 (2020).

メディアの報道に対するソーシャルメディアユーザの反応の時系列変化の分析手法

井上　大成† 　風間　一洋†† 　吉田　光男††† 　土方　嘉徳††††

† 和歌山大学大学院システム工学研究科　〒 640–8510 和歌山県和歌山市栄谷 930
†† 和歌山大学システム工学部　〒 640–8510 和歌山県和歌山市栄谷 930
††† 筑波大学ビジネスサイエンス系　〒 112–0012 東京都文京区大塚 3–29–1
†††† 関西学院大学商学部　〒 662–8501 兵庫県西宮市上ケ原一番町 1–155
E-mail: †s246023@wakayama-u.ac.jp, ††kazama@ingrid.org, †††mitsuo@gssm.otsuka.tsukuba.ac.jp,
††††contact@soc-research.org

あらまし　クラスタリングや LDA などで得られた文書集合の内容を推測する簡単な手法として，頻度や重要度が高い単語を抽出する方法がよく用いられる．ただし，そのような単語は抽象度が高いことが多く，より詳しい状況を知るために係り受け関係にある単語も同時に抽出する方法が用いられることがある．しかし，係り受け解析器はテキストの形式や構造に過敏であり，ツイートのような口語で記号類が多い文章や並列構造を正しく解析できないなどの問題が存在する．本論文では，ツイートのような文章に対しても，形態素列にパターンマッチングを適用することで，名詞と係り受け関係にある形態素を抽出し，文章の極性と共にわかりやすく可視化する手法を提案する．実際に，いくつかのトピックに対するメディアの報道に特に強く反応した期間のツイート群を変化点検出で取り出し，提案手法を用いることでユーザの反応の詳細と，時期による違いがわかりやすく表示できるかを検証する．

キーワード　ソーシャルメディア，ニュース・マスコミ，テキストマイニング，自然言語処理応用，情報可視化

1 はじめに

Twitter（現 X）などのソーシャルメディアの普及に伴い，ユーザが投稿した大量の投稿の内容を分析する研究が盛んに行われている．内容を知るための簡単な手法として，クラスタリングや LDA などで絞り込んだ後に，出現頻度や TF-IDF，LDA のトピック出現確率などを用いて，重要な単語集合を抽出する手法がよく用いられる．しかし，これらの手法で抽出される単語は抽象度が高い傾向があり，内容を詳しく推測することは難しい．そこで，単語 n-gram や係り受け関係にある単語を表示することで，内容理解を支援する研究が行われている．特に後者の係り受け解析器はテキストの形式や構造に過敏であり，ツイートのような口語で記号類が多い文章を正しく解析できないことも多く，例えば日本語係り受け解析器としてよく用いられる CaboCha [1] には，名詞や文節の並列構造を含む文章を正しく係り受け解析できないという問題が存在する．

本論文では，重要な単語と修飾・被修飾の関係や主語・述語の関係にある係り受け関係に着目して，幾つかの品詞パターンを用意することで，ツイートのような文章に対しても，名詞と係り受け関係にある形態素を抽出し，文章の極性と共にわかりやすく可視化する手法を提案する．

実際には，メディアの報道に対するソーシャルメディアのユーザの反応を分析するため，いくつかのトピックに対するメディアの報道に特に強く反応した期間のツイート群を変化点検出で取り出し，提案手法を用いることでユーザの反応の詳細と，時期による違いがわかりやすく表示できるかを検証する．

2 関連研究

膨大なテキストデータの内容理解を支援するために，単語の組を提示する様々な研究が行われている．

単語 n-gram を用いた研究として，白田らは「シドアルジョ泥流」に関する記事や論文から抽出した 1-gram と 2-gram に LDA を用いてトピック抽出を行った [2]．また Wang らは，n-gram のトピックモデルを提案し，実際に機械学習のトップカンファレンスである NIPS のデータから 1-gram，2-gram のトピックを抽出して，その有効性を示した [3].

係り受け関係にある単語を用いた研究として，赤尾らは LDA で抽出したトピックの特徴語に加えて，CaboCha で特徴語と係り受け関係を持つ単語を同時に抽出する手法を提案した [4]．なお，事前に係り受け解析されやすいように文章を修正すると共に，CaboCha で並列構造を正しく解析できない問題に対処するために解析結果を修正することで解析誤りを低減していた．また，福田らはコロナウイルスワクチンに対して持つ感情とその感情が現れる要因を分析するため，ツイートテキストから抽出した感情の時系列変化において，バーストが検出された係り受け関係を抽出する手法を提案した [5].

ただし，これらの研究では，単語の組をランキングするだけであり，上位のほんの一部の結果しか表示できない．久保らは，単語の関係性を示す単語 2-gram に対して LDA を適用し，各トピック上位の 2-gram をトピックグラフとして再構築することで，元の文章で共通して使われる重要な部分とそれらに対する行動や意見を簡潔に表現する手法を提案した [6]．グラフ化す

ることで多くの単語の組を簡潔に表示できるが，一覧性に劣るなどの欠点も存在する．本論文では修飾関係を踏まえて，その関係が存在する量や文書の極性も同時に可視化し，メディアの報道に対するユーザの反応の詳細を簡潔に表現する手法を提案する．

3 提案手法

3.1 提案手法の概要

提案手法は任意のテキスト集合に対して適用可能であり，基本的には，LDA の各トピックの上位の名詞など，任意の名詞集合を指定して，その名詞集合の係り受け関係を分析できる．本論文では，TF-IDF の値が高い名詞集合を，入力に用いた文書集合において特に重要で注目すべきとみなし，実際にメディアのニュース報道に対するソーシャルメディアのユーザの反応を可視化することで，反応の詳細と時期による違いがわかりやすく表示できるかを検証する．

具体的には，井上らの手法 [7] を用いてメディアのニュース報道に対する引用ツイート数の時系列変化に変化点検出を適用し，変化点直後の 3 日間の引用ツイートを特に反応が強く表れている入力データとして，次の手順で処理する．

（1） ユーザの発言部分の抽出
（2） パターンマッチングによる係り受け関係の抽出
（3） 重要単語の抽出
（4） 発言の極性の推定
（5） 係り受け関係の可視化

また，引用ツイートで変化点がなくニュースでのみ変化点があるような場合でも，何らかの事件が起こって報道が活発に行われており，そのような場合でもユーザがどのような感情を抱き，どのような意見をしていたのか分析するため，ニュース記事数の時系列変化に対しても変化点検出を適用し，引用ツイート数の変化点直後 3 日間と重ならない期間の引用ツイートも入力データに加える．

3.2 ユーザの発言部分の抽出

本論文で用いる引用ツイートは，ニュース記事の URL に加えて，ニュースのタイトルやメディア名なども含まれることが多いが，そのような部分はユーザの意見を反映しておらず，また自動的・慣例的に挿入されるために頻度も比較的高く，重要な単語や発言の極性を検出する上でノイズとなることが多い．そこで中北らの手法 [8] を用いて，ユーザの発言部分だけを抽出する．

3.3 パターンマッチングによる係り受け関係の抽出

次に，名詞の詳細を知ることができるように，引用ツイートの発言部分から，名詞と係り受け関係にある形容詞または動詞を抽出する．このような場合は，係り受け解析を用いる方法が一般的であるが，形態素解析よりも入力テキストの形式や構造に過敏であり，特にツイートのような口語で記号類が多い文章はうまく処理できないことも多い．

そこで，提案手法では名詞との係り受け関係だけに注目するという特徴を利用して，MeCab で日本語形態素解析して得られる形態素列にパターンマッチングを適用することで，入力データに対して比較的頑健で高速な係り受け関係抽出を実現する．

表 1　使用パターンと実例

	使用パターン	実例
1	形容詞／名詞	軽い／症状
2	名詞／助詞／形容詞	感染者／が／多い
3	名詞／助詞／動詞	攻撃／が／始まる
4	名詞／助詞／名詞（サ変接続）／動詞	侵攻／を／非難／する

使用するパターンは，表 1 に示すように，修飾語と被修飾語の 1 パターンと主語と述語の 3 パターンの計 4 通りである．

この結果，上の 2 パターンで名詞と形容詞の組を，下の 2 パターンで名詞と動詞（サ変接続の名詞を含む）の組としてまとめて，ツイートと対応づける．ただし，抽出した形容詞や動詞の中には，ユーザの反応の詳細を知る上で有益ではない単語も含まれるため，4.4.1 節で述べるようにストップワードリストを作成して処理対象から除外する．

3.4 注目する名詞集合の抽出

抽出した係り受け関係から注目すべき重要単語を含んでいるものに限定するため，TF-IDF を適用して，注目する名詞集合を抽出する．

TF-IDF は情報探索やテキストマイニングなどの分野で利用される，文書中に出現した単語がどのくらい特徴的であるかを識別するための指標である [9]．全文書数 D に含まれる全単語数 K のうち，i 番目のある単語 w_i の TF-IDF の値 $TF\text{-}IDF(w_i)$ は，ある文書に単語が出現する度合いを表す TF（Term Frequency）の値 $TF(w_i)$ と，単語が文書全体に出現する度合いの逆数である IDF（Inverse Document Frequency）の値 $IDF(w_i)$ の積で与えられる．本論文では，ツイートを文書として扱う．

$$TF(w_i) = \frac{n(w_i)}{\Sigma_{k=1}^{K} n(w_k)} \tag{1}$$

$$IDF(w_i) = \log \frac{D}{d(w_i)} \tag{2}$$

$$TF\text{-}IDF(w_i) = TF(w_i) \times IDF(w_i) \tag{3}$$

ここで $n(w_i)$ は単語 w_i の出現頻度を表し，$d(w_i)$ は単語 w_i を含む文書数を表す．

実際には，ツイートの発言部分を MeCab で日本語形態素解析し，一般・自立・固有名詞・サ変接続・形容動詞語幹である名詞だけを抽出し，TF-IDF の値が上位 20 件の名詞集合を抽出する．

3.5 発言の極性の推定

さらに，ユーザがどのような感情を持ってニュース報道に言及したかに関する情報も利用するために，引用ツイートから極性を抽出する．

極性の抽出には，Nanba らの T5 を使った感情分析器 [10] を使用した．これは約 100GB の日本語コーパスで事前学習した T5 モデルを，梶原らの主観と客観の感情分析データセット

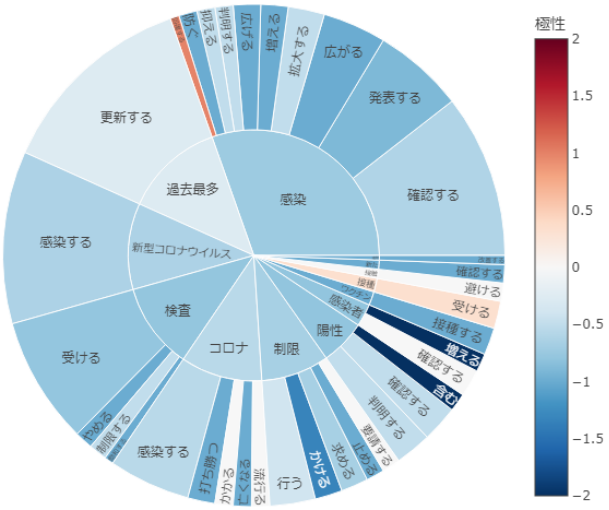


図 1　二重ドーナツグラフ可視化の例

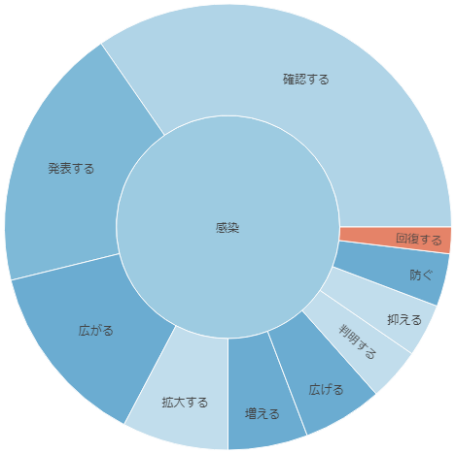


図 2 「感染」に注目して拡大表示した例

WRIME [11][1]を用いてファインチューニングした感情分析器である．WRIME は，日本語の感情分析の研究のために，SNS 上の投稿に対して，Plutchik の基本 8 感情（喜び，悲しみ，期待，驚き，怒り，恐れ，嫌悪，信頼）を 4 段階（0, 1, 2, 3），感情極性を 5 段階（-2, -1, 0, 1, 2）で人手でラベル付けしたデータセットである．このような深層学習ベースのアプローチは，学習用データセットがあれば比較的容易に感情分析器を構築でき，検出率も比較的高い利点がある．

実際には，ツイートの発言部分に感情分類器を適用し，その極性の値（-2〜2）を抽出して，引用ツイートと対応づける．

3.6 可　視　化

本論文では，得られた係り受け関係をそのままランキングするのではなく，名詞と係り受け関係にある形容詞や動詞の関係を，二重ドーナツグラフを用いて可視化する．なお，形容詞は名詞の性質や状態を表し，動詞は名詞の動作や作用を表すと考えて，実際には表 1 の各パターンでそれぞれ出現頻度上位 N 件の係り受け関係を求め，パターン 1 とパターン 2 を形容詞に，パターン 3 とパターン 4 を動詞にまとめて可視化する．動詞に関して，$N = 20$ で二重ドーナツグラフとして可視化した例を，図 1 に示す．内側のドーナツグラフには注目する名詞を，外側のドーナツグラフにはそれと係り受け関係にある形容詞または動詞を，それらの出現頻度に応じた面積で表示する．

これだけでは単語の区別が難しいことと，表示された単語がどのような意図で用いられていたかを示すために，極性の値を用いて着色する．まず，各形容詞や動詞に関して，それらが出現した引用ツイートの極性値の平均を計算する．次に，名詞に関して，それらと関係がある形容詞または動詞の値の重み付き平均を計算する．最後に，形容詞・動詞および名詞を，ポジティブ（2）からネガティブ（-2）まで赤から青に変化するように着色する．

本可視化システムは，Python と Web アプリケーションのフレームワークである Streamlit を用いて Web ブラウザ上で対話的に操作できるシステムとして実装した．トピックと対象期間を指定すれば図 1 のような二重ドーナツグラフ可視化が可能であり，さらに内側の名詞から，詳しく調べたい単語をクリックすることで図 2 に示すように，その名詞だけに限定して拡大表示することもできる．

4　評価・分析

4.1 データセット

ニュースに引用したツイートデータセットとして，Twitter API で取得した 2022 年 1 月 1 日から 2022 年 11 月 30 日までのツイートのうち，ニュース記事の URL が含まれる 4,491,671 件のツイートを利用した．

また，ニュースデータセットとして，ニュースポータルサイト Ceek.jp News[2]で扱われているニュース記事のうち，その URL がツイートデータセット内のツイートに含まれているニュース記事 3,466,472 件を利用した．

4.2 分析対象トピック

データセットの期間内における，引用数上位 10 件のニュース記事から，時系列的な特性が異なる以下の 2 個のトピックを人手で選定して，それぞれ「コロナウイルス」，「ロシア or ウクライナ」を Elasticsearch のクエリとして検索して，ニュース記事と，さらにそれらのニュースを引用しているツイートを抽出した．

（1） コロナウイルス

2019 年に発生してから世界中で感染が確認されているコロナウイルスに関する報道．感染者が急激に増えた時期として，2022 年 2 月ごろをピークとする第 6 波，7 月ごろをピークとする第 7 波がデータセットの期間内に存在している．

（2） ロシアによるウクライナ侵攻

2022 年 2 月 24 日にロシアが開始したウクライナへの軍事侵攻に関する報道．一週間前の 2022 年 2 月 17 日ごろには米当局の情報で侵攻の可能性があることが知らされており，侵攻前

1：https://github.com/ids-cv/wrime

2：http://news.ceek.jp/

表 2　CaboCha とパターンマッチング法の対応

	CaboCha	パターンマッチング法
1	形容詞／名詞	名詞／助詞／形容詞
2	名詞／形容詞	名詞／助詞／形容詞
3	名詞／動詞	名詞／助詞／動詞 名詞／助詞／名詞（サ変接続）／動詞

表 3　係り受け抽出性能

	適合率	再現率	F 値
パターンマッチング法	0.891	0.475	0.619
CaboCha	0.738	0.870	0.799

日の 23 日にはウクライナ全土に非常事態宣言が出された．

4.3　係り受け抽出性能の評価

パターンマッチング法はツイートのような厳密に解析しにくいテキストに対する頑健性と処理速度を重視している．しかし，当然離れた係り受け関係は検出されないため，ある程度の検出漏れがあるはずである．そこで係り受け解析器である CaboCha と比較し，提案手法がどの程度係り受け関係を抽出できているかを評価した．

評価用テキストデータとしては，トピック「コロナウイルス」の 1 つ目の変化点である 1 月 5 日直後 3 日間に報道されたニュース記事の文字数が 30 文字以上の引用ツイートからランダムに 100 件を抽出し，抽出率を向上させるための前処理・後処理は行わずに，URL とハッシュタグだけを除去して使用した．正解データは，人手で表 1 に示す品詞の係り受けパターンを判定して作成した．なお，評価の際の CaboCha とパターンマッチング法で対象とする係り受け関係の対応を表 2 に示す．

まず，評価用テキストデータを CaboCha とパターンマッチング法で 100 回処理した時間を 10 回計測して平均値を求めた．パターンマッチング法は 9.71 秒に対して CaboCha は 44.03 秒で，22%の時間で処理できることがわかった．

パターンマッチング法と CaboCha の解析結果を適合率，再現率，F 値で評価した結果を表 3 に示す．この結果を見ると，CaboCha の方が再現率，F 値共に高く，予想通りパターンマッチング法は検出漏れがあった．しかし，適合率に関してはパターンマッチング法の方が高かった．この原因を調べると，ツイートでは助詞や句点の脱落が頻繁に出現し，そのような場合は CaboCha は解析誤りを引き起こすが，パターンマッチング法ではそもそも抽出対象にはならないからであった．

これらの結果から，パターンマッチング法は，ツイートのテキスト分析のように，大量のテキストデータを対象に統計処理を行う場合のように，多少の検出漏れは構わないが，解析誤りによるノイズは低減させたい用途には向いていると思われる．

4.4　ソーシャルメディアユーザの反応の分析

次に，2 つのトピックに対してソーシャルメディアユーザの反応を提案手法を用いて可視化し，その結果を分析した．また，可視化する出現頻度上位の係り受け関係の件数は $N = 20$ とした．

表 4　ストップワードリスト

形容詞	ない，やすい
動詞	する，いる，れる，なる，ある，てる，できる（出来る），思う，られる，いう（言う），くる，せる，見る，やる，出る，考える，知る

4.4.1　ストップワードの選定

ツイートデータセットを MeCab で形態素解析して，形容詞と動詞の出現頻度の上位 20 件の単語を目視で確認し，そのうち内容を補足するための情報量が少なく有益とは思わない単語をストップワードとした．ストップワードリストを表 4 に示す．

4.4.2　トピック「コロナウイルス」

ニュース記事数と引用ツイート数の時系列変化から変化点検出を行った結果，トピック（1）ではニュース側で 1 月 4 日と 7 月 22 日に，ツイート側で 1 月 5 日に変化点を検出した．そこで引用ツイート数の変化点である 1 月 5 日に加えて，ニュース記事数でのみ検出した 7 月 22 日を分析対象とした．提案手法を適用した結果を図 3 に示す．左側は形容詞の可視化結果，右側は動詞の可視化結果である．

図 3(a) は，1 つ目の変化点である 1 月 5 日直後の可視化結果である．名詞・形容詞では，第 6 波の感染拡大時期であったことから，「接種・酷い」や「接種・重い」などワクチン接種後の副作用を示唆する表現や「米軍・甘い」や「米軍・悪い」など米軍基地におけるクラスターの発生を示唆する表現が多かった．また，「オミクロン・強い」などオミクロン株の感染が注目されている時期でもあったと思われる．名詞・動詞では，「感染・確認する」や「新型コロナウイルス・感染する」などコロナウイルスの感染を示唆する表現が多く抽出された．極性に関してはどちらも負の極性を持つ文章がほとんどであり，正の極性を持つ単語も一部あるが，よく見ると必ずしも正とは言えなかった．

図 3(b) は，2 つ目の変化点である 7 月 22 日直後の可視化結果である．名詞・形容詞では，第 7 波の感染拡大時期を反映して，「コロナ・憎い」や「コロナ・恐ろしい」など，コロナウイルスに対する感情を示唆する表現や，「症状・軽い」や「症状・重い」など感染後の症状を示唆する表現が多かった．特に後者に関しては，「重い」，「しんどい」に対して，「軽い」，「無い」という逆の意味の単語が同時に抽出されており，提案手法ではそれらの割合をある程度推定することも可能だと思われる．名詞・動詞では，1 つ目の変化点と同様に「感染・確認する」など，コロナウイルスの感染に関する表現が多いが，「感染・拡大する」や「過去最多・更新する」，「感染者・増える」など，感染者数の増加に関するさまざまな表現が使われている．極性に関しては，この期間もネガティブなものが多い．

4.4.3　トピック「ロシアによるウクライナ侵攻」

トピック（2）ではニュース側で 2 月 18 日と 2 月 24 日に，ツイート側で 2 月 23 日に変化点を検出した．そこで 2 月 23 日に加えて 2 月 18 日を分析対象とした．提案手法を適用した結果を図 4 に示す．左側は形容詞の可視化結果，右側は動詞の可視化結果である．

図 4(a) は，1 つ目の変化点である 2 月 18 日直後の可視化結

果である．

名詞・形容詞では，ロシアのウクライナ侵攻の直前なので，「批判・強い」や「侵攻・無い」のような戦争を否定する表現や，「ロシア・怖い」，「プーチン・強い」，「ウクライナ・弱い」などのそれぞれの情勢を表す表現が同時に抽出されている．名詞・動詞では，「ロシア・侵攻する」や「攻撃・受ける」，「侵攻・決断する」などロシアによるウクライナ侵攻を示唆する表現が多かった．極性に関しては，どちらもネガティブなものが多いが，「外交・粘り強い」など一縷の希望を示唆する表現も抽出されている．

図 4(b) は，2 つ目の変化点である 2 月 23 日直後の可視化結果である．名詞・形容詞では，開戦直後ということもあり，制裁という名詞が多く使われ，それに対して「厳しい」，「強い」，「甘い」，「軽い」，「重い」などの対極の意味の形容詞が組み合わされた表現が抽出され，それらの割合から厳しい制裁を望むユーザが多かったことが推測できる．名詞・動詞では「戦争・始まる」や「侵攻・受ける」，「攻撃・受ける」などのロシアの侵攻開始を示唆する表現の出現に加えて，「戦争・反対する」，「ロシア・非難する」などロシアを批判する表現も多かった．極性に関しては，負の極性を持つ表現がほとんどである．なお，「良い・戦争」という表現が正の極性として抽出されているが，ツイートデータを調べたところ，確かにそのような用語が使用されていたが，それは否定的な文脈，または否定表現を伴っていて，提案手法では極性が反転する場合には対応できていないことがわかった．

4.5 考　　察

提案手法の妥当性について考察する．

まず，提案手法は実際に名詞と係り受け関係にある形容詞や動詞を単純にリスト表示すると最大で 80 行になる内容を，コンパクトに表示できることがわかった．全体の可視化結果では小さく表示されてしまう場合もあるが，特定の単語に限定すれば詳細に見ることができるために，全体的な俯瞰と詳細な分析を両立できていた．

また，今回は同一トピックで 2 つの時点を分析対象としたが，提案手法ではそれらの違いがわかりやすく，また名詞も形容詞や動詞と組み合わせることでその文脈を推定したり，同じ意味を持つ組み合わせの判断が容易になることがわかった．

さらに，同一の名詞に対して意味が反対の形容詞が同時に抽出される現象も確認でき，その場合はそれらの割合から意見の多様性と多数派の判断ができる可能性も示された．

極性に関しては，ポジティブな単語とネガティブな単語の区別が容易であったが，一部に妥当と思われない極性もあった．これは提案手法は，それらの係り受けが出現したツイート全体の極性を集計して用いているために齟齬が生じることがあることと，否定的な文脈または否定表現が伴っていても，それを反映できていないことがわかった．なお，今回用いたトピックの可視化結果ではほとんどがネガティブなものばかりであった．これが選択したトピックによるものか，ソーシャルメディアではネガティブな発言が多い傾向にあるのか，システム的なバイアスなのかは不明であり，今後多彩なトピックに関して分析を行う必要がある．

5 おわりに

本論文では，いくつかのトピックに関するメディアのニュース記事と，それに言及したツイートデータに対して提案手法を用いて，メディアの報道に特に強く反応した期間のツイート群を変化点検出で取り出し，ユーザの反応の詳細と，時期による違いを分析した．その結果，提案手法では多くの情報をコンパクトにわかりやすく表示できること，同一トピックでも異なる時期の違いが判別しやすいこと，ある事柄に対する多彩な意見と多数派・少数派の識別ができる可能性が示された．ただし，ツイートの極性を統計処理して単語に適用する手法に関しては，一部の単語で正しい極性が得られない問題も確認した．

今後の課題としては，より多くのトピックで分析を行うことで，トピック単位ではなく，ソーシャルメディア全体の性質や傾向について調査すると共に，可視化システムについて，抽出元のツイートテキストも表示することで，1 つの単語に着目した時により詳細な分析ができるように改良し，可視化結果の評価を行う．

謝　　辞

本研究は JSPS 科研費 21H03557 の助成を受けた．

文　　献

[1] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 42, No. 6, pp. 1834–1842, 2002.

[2] Yussanti Nur Fajrina, Yukari Shirota, and Riri Fitri Sari. Topic extraction analysis for sidoardjo mudflow impacts. 学習院大学 経済論集, Vol. 53, No. 3, pp. 101–114, 2016.

[3] Xuerui Wang, Andrew MSCallum, and Xing Wei. Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 697–702, 2007.

[4] 赤尾貴仁, 風間一洋. 係り受け関係を用いた LDA のトピックの内容理解の支援. WebDB 夏のワークショップ 2022, B3-1, 2022.

[5] 福田悟志, 難波英嗣, 庄司裕子. コロナ禍におけるワクチンに対する人々の感情変化とその要因の分析. 知能と情報, Vol. 34, No. 3, pp. 592–600, 2022.

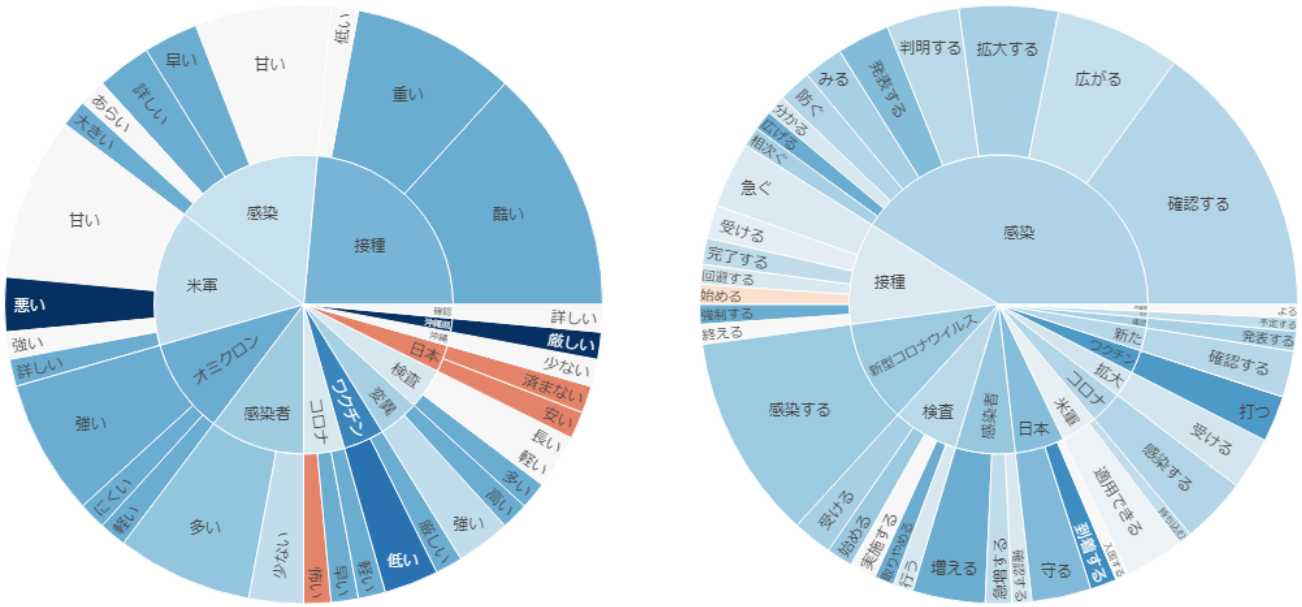
[6] 久保侑哉, 風間一洋, 榊剛史. 単語 2-gram に基づくマルチトピックグラフの可視化. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2018), B1-5, 2018.

[7] 井上大成, 中北雄大, 風間一洋, 吉田光男, 土方嘉徳. メディアの報道とソーシャルメディアの反応の時系列分析. 第 15 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2023), 5b-5-1, 2023.

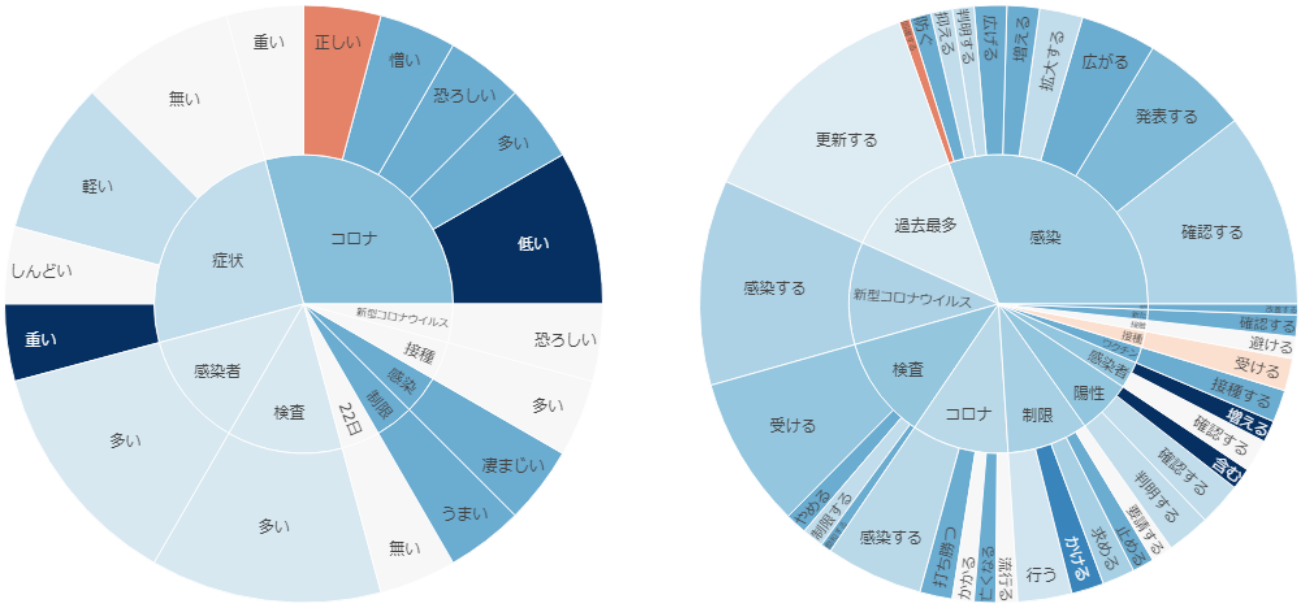
[8] 中北雄大, 風間一洋, 吉田光男, 土方嘉徳. 感情とトピックに注目したメディアの報道姿勢の分析. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2022), B24-2, 2022.

[9] Karen Spärck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

[10] Hidetsugu Nanba, Katsuya Yamamoto, Satoshi Fukuda, Hi-



(a) 1 月 5 日〜 1 月 7 日

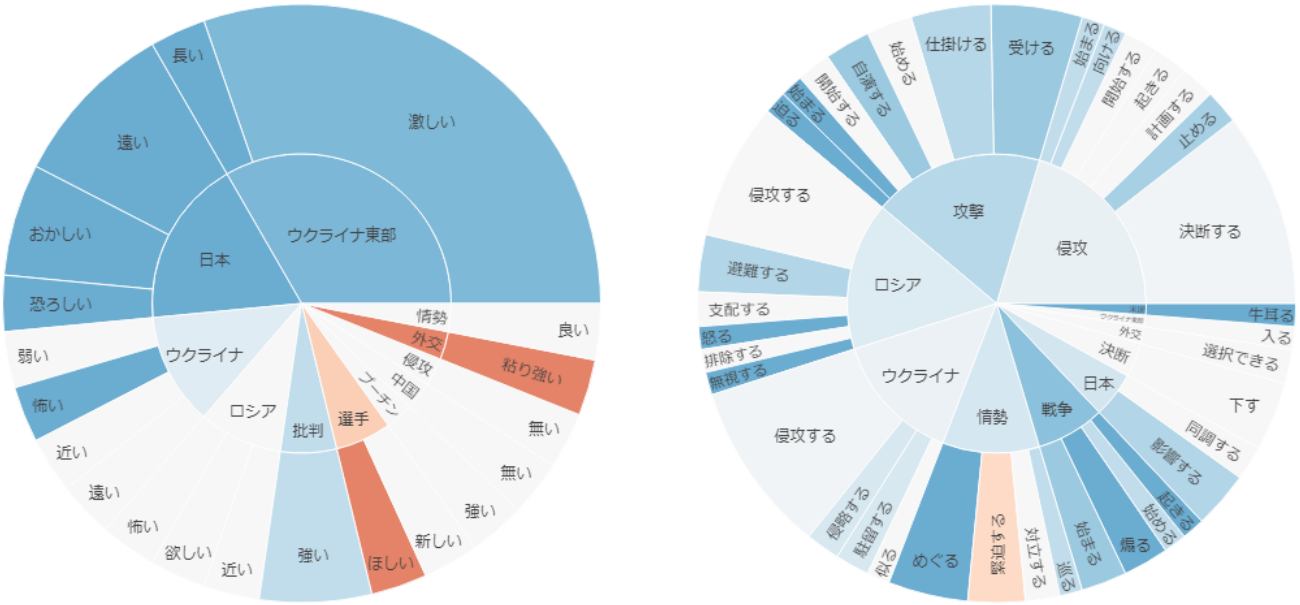


(b) 7 月 22 日〜 7 月 24 日

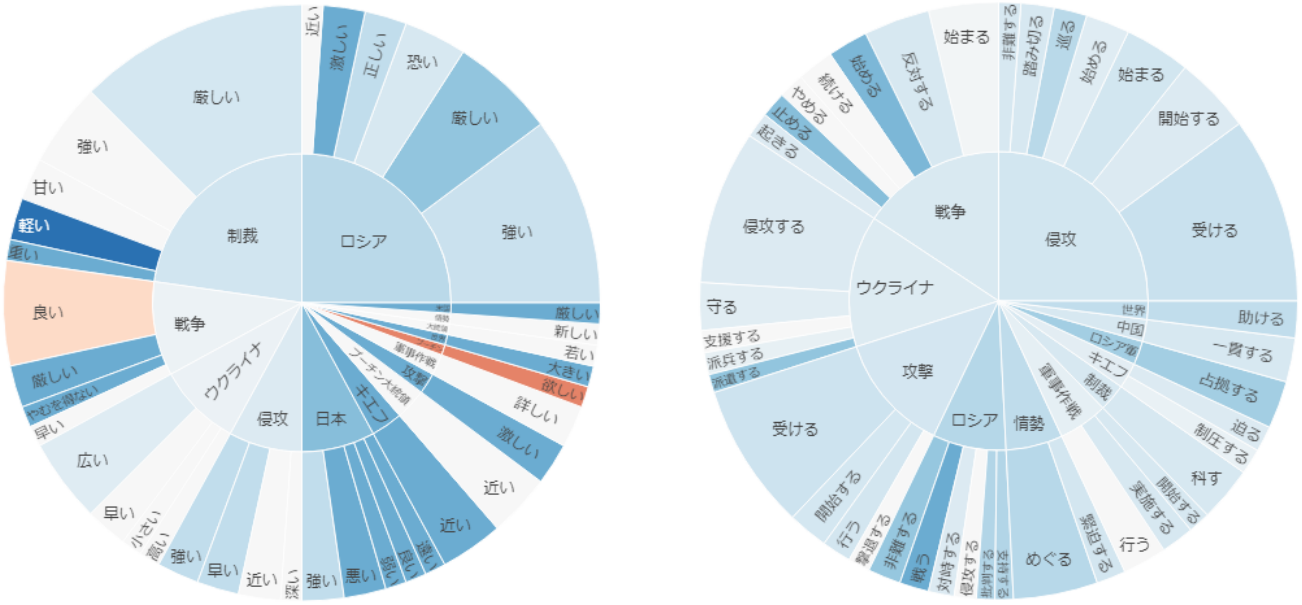
図 3　トピック「コロナウイルス」の可視化結果

roko Shoji, Masayoshi Tanishita, Yasushi Kyutoku, and Mitsuru Yamashina. Modeling the social acceptability of technologies using Twitter data. In *Proceedings of the 2023 IEEE Conference on Systems, Man, and Cybernetics (IEEE SMC 2023)*, 2023.

[11] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, pp. 2095–2104, 2021.



(a) 2 月 18 日～ 2 月 20 日



(b) 2 月 23 日～ 2 月 25 日

図 4　トピック「ロシアによるウクライナ侵攻」の可視化結果

Twitterを用いたCOVID-19ワクチン接種の継続性に関係する情報共有行動の分析

北林　遼大† 　張　翔† 　豊田　正史†† 　吉永　直樹†† 　梅本　和俊††

† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7–3–1
†† 東京大学生産技術研究所　〒153–8505 東京都目黒区駒場 4–6–1
E-mail: †{kitabaya,cs,toyoda,umemoto}@tkl.iis.u-tokyo.ac.jp, ††ynaga@iis.u-tokyo.ac.jp

あらまし　COVID-19 ワクチンは感染予防等の観点から接種が推奨され，国民全体の接種率向上が感染症収束のために求められてきた．日本では，2021 年 6 月より 18 歳以上への接種が始まり，同年 12 月からは追加接種（3 回目）も提供されてきた．2022 年 12 月 28 日現在，18 歳以上の国民に対する初回接種（1,2 回目）は順調に進行したのに対して，追加接種には伸び悩みが見られている．本研究では，国内における COVID-19 ワクチン接種行動の変化及びその背景について明らかにすることを目的とする．日本語の全量 Twitter 投稿から抽出したワクチン関連投稿を用いて，自身の接種を報告したユーザ数と 65 歳未満の接種者数との間に強い相関があることを確認した．さらに，初回接種を報告したユーザの年代を自動推定し，ワクチン接種を継続したユーザ群とそうでないユーザ群の特徴の違いを分析することで，医師や政府等の発信に対する関心やワクチンに関する不正確な情報への共有行動の傾向の違いを明らかにした．
キーワード　新型コロナウイルス, ワクチン接種, 追加接種, ソーシャルメディア (SNS), Twitter (X)

1　はじめに

ワクチン接種は感染症予防に最も有効な対策の一つと考えられており，感染の予防や症状の軽減などの観点から接種が推奨されている．特に，COVID-19 のようなパンデミックを起こす感染症に対しては、拡大防止および収束の観点から国民全体の接種率を高めることが求められる [1]．

日本では 2021 年 6 月 17 日以降，18 歳以上の全国民に対して COVID-19 ワクチン接種の機会が提供されている [2]．また，初回接種（1,2 回目）を完了した人々に対して，3 回目以降の追加接種の機会も提供されている．接種により得られた感染予防効果等は時間の経過に伴って低下していく一方，追加接種により COVID-19 の発症や重症化の予防効果を高めることができることが複数の研究で報告されている [1, 2].

国内の COVID-19 ワクチン接種状況 [3]を見ると，18 歳以上の国民に対する初回接種は順調に進行したのに対して，追加接種には伸び悩みが見られている．2022 年 12 月 28 日時点において全人口の 8 割が 2 回目の接種を完了しており，年齢階級別に見ても，20 代以上の全ての年齢階級において接種率は 8 割以上を達成した．一方で，3 回目以降の追加接種率に関しては全人口の 67.7%にとどまっている．特に 30 代以下ではその割合が 5–6 割を下回っている．こうした国内の接種状況の変化やその要因について把握することは，感染拡大防止や収束に向けて重要である．加えて，人々の接種行動およびその背景について分析することで得られた情報は，今後未知の感染症が出現した際に，ワクチンを活用して感染症被害を最小限に抑えるための対策を実施する上でも重要な知見となりうると考えられる．

本研究では，国内における COVID-19 ワクチン接種行動の推移や，追加接種数が伸び悩んだ背景について明らかにすることを目的として，Twitter ユーザが投稿したワクチン関連投稿や，ワクチンに関連した情報に対する情報共有行動について分析する．本分析における Research Questions (RQs) は以下のとおりである：

RQ1:　国内の Twitter ユーザの接種行動は，日本全体の接種状況を代表しているといえるか？

RQ2:　接種を継続したユーザとそれ以外のユーザ間の違いは，Twitter 上での情報共有行動に現れているか？

RQ3:　ワクチン接種に関する情報共有行動の違いは年代により異なるか？

以上の RQs を明らかとするために，ワクチンに関連した Twitter 上の日本語投稿を収集し，そこからユーザ自身がワクチン接種を受けたことを報告する旨の投稿（接種報告）を抽出した．これを基に Twitter 上のワクチン関連投稿と実際の国内の接種者数の相関係数の計測を通して，国内における接種行動の推移の分析に Twitter データを用いることの妥当性を検証することで，RQ1 に対する回答を試みた．続いて，初回接種を報告したユーザ群を接種行動の継続性に基づいて分類し，参照した情報源等について年代ごとに比較を行うことで，RQ2・RQ3 に対する回答を試みた．本研究の主な貢献は以下の通りである：

- Twitter 上で接種報告を行ったユーザ数と，国内の 65 歳未満の接種者数に強い相関があることを示した．
- 2021 年に初回接種を報告したユーザのうち，2022 年に

1：https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines
2：https://www.niid.go.jp/niid/ja/diseases/ka/corona-virus/2019-ncov/2484-idsc/10569-covid19-53.html
3：https://www.kantei.go.jp/jp/headline/kansensho/vaccine.html

も接種報告を行ったユーザは約 33%，2022 年にワクチン関連投稿を行ったが接種報告はしていないユーザは約 5%，2022 年には一度もワクチン関連投稿を行わなかったユーザは約 60%であったことを示した．

- 2021 年には接種報告を行ったが 2022 年にワクチン関連投稿を行わなかったユーザ群のうち，特に 30 歳未満において，医者や政府などの信頼性の高い情報源からの発信に対する関心が接種開始当初から比較的低い傾向にあったことを示した．
- 2021 年には接種報告を行ったが 2022 年に接種報告を行わなかったユーザ群のうち，特に 30 歳以上において，不正確な情報や接種に対する不安に繋がるような情報を SNS 上で比較的多く参照，共有する傾向にあったことを示した．

2 関連研究

COVID-19 感染拡大防止の観点から国民に対して接種を推奨する上で，ワクチンに対する国民の意識や感情等を把握することは重要である．このような背景から，SNS を活用することにより，これらを把握することを目的とした研究が国内外問わず行われている [3–9]．本章ではまず，初回接種が開始された前後を分析期間とした先行研究について紹介する．次に，追加接種に対する人々の意識や感情などに着目した先行研究について紹介する．最後に，既存研究で明らかになっていない課題を整理し，本研究の位置付けについて説明する．

Hisamitsu ら [3] は，日本国民の接種の是非を決断するに至った要因となる情報源を明らかにすることを目的として，接種に対するスタンス変化と情報共有行動の関係について分析した結果，中立派から賛成派に変化したユーザは医師や政府による投稿を共有する傾向にある一方，反対派に変化するユーザは医師や政府系アカウントではない個人の投稿を共有する傾向にあることを示した．Poddar ら [4] は，2018 年 1 月から 2021 年 3 月の間の英語投稿を用いて，COVID-19 前後に投稿されたワクチン関連投稿の話題をスタンスごとに分析するとともに，スタンスの変化やその要因について分析を行った．その結果，2020 年に賛成派から反対派に移行したユーザは主にワクチン開発を行う製薬会社に対する非難を示した一方で，2021 年に賛成派から反対派に移行したユーザは主にワクチンの強制的な接種に対する非難や有効性への懸念を示したことを明らかにした．

また，追加接種の提供開始にともない，当該接種に対する人々の意識や感情などにも研究の関心が集まっている [10–13]．例えば Praveen ら [10] は，追加接種に関するインドの人々の感情や懸念の把握を目的に，2022 年 3 月から 9 月に投稿された追加接種に関する投稿を用いて，感情分析および Latent Dirichlet Allocation [14] によるトピックモデリングを行った結果，追加接種の安全性や有効性，および若い人の追加接種の必要性などが主な懸念点であったことを明らかにした．Delcea ら [11] は，イスラエルで追加接種が承認された 2021 年 7 月 12 日以降の 1ヶ月間の英語投稿を収集し，接種に対するスタンスの推移や，ハッシュタグ，N-gram などから追加接種を躊躇する要因の分析を行った．その結果，初回接種が開始された時期を対象として同様の分析を行った先行研究 [6] と比較して，副作用に対する懸念や安全性に対する懸念など，接種を躊躇する主な要因はほとんど変化していなかったことを明らかにした．

以上の先行研究では，初回接種の開始前後や追加接種の承認以降などの時期を対象に，SNS ユーザの接種に対する感情や賛成/反対のスタンス等に着目し，その変化や要因等の分析が行われた．一方，ユーザの「接種行動」，すなわちユーザが実際に接種に至ったかどうかについては着目されておらず，SNS 上のワクチン関連投稿と実際の接種状況との関係を明らかになっていない．また，国内では追加接種数が伸び悩むなど接種行動の変化が見られたが，国内のユーザの接種行動の変化およびその要因を分析した先行研究は我々の知る限りでは存在しない．

本研究では，2021 年 6 月から 2022 年 12 月の期間に投稿された日本語投稿を分析することを通して，国内の接種行動の変化及びその背景について明らかにすることを目的とする．具体的には，接種行動に結びつく投稿に焦点を当ててユーザの接種行動の推移を分析することで，接種の継続性に関係する投稿内容や情報共有行動等を年代ごとに明らかにした．

3 分析手法

1 章に示した RQs に回答するには，Twitter 上でワクチンに関連した投稿を行ったユーザが実際に接種に至ったかどうかを推定すること，および，これらの投稿を行ったユーザの年代を推定することが必要である．本章ではまず，今回収集したワクチン関連投稿とこれに対して行ったアノテーション作業について述べる．次に，本研究で構築した，ワクチン関連投稿から接種報告を抽出するための分類器，および投稿内容等からユーザの年代を推定する分類器について説明する．

3.1 ワクチン関連投稿の収集

日本では，2021 年 6 月 17 日から 18 歳以上への COVID-19 ワクチン接種が本格的に開始され，2022 年 12 月末時点で 8 割以上の人が 2 回目の接種を完了した [4]．これを踏まえて本研究では，2021 年 6 月から 2022 年 12 月までの期間を分析対象として，この期間に投稿された日本語の全量ツイート [5]から「ワクチン」という語を含むツイート（リツイート，引用リツイート，メンションを含まない）を抽出し，5,684,911 人のユーザによる 30,150,061 件のワクチン関連投稿を取得した．また，上記に対して発生したすべてのユーザ間リアクションデータ（リツイート，引用リツイート，メンション）についても取得した．

3.2 ワクチン関連投稿に対するアノテーション

Twitter ユーザの接種行動の分析を行うには，投稿を行ったユーザが接種に至ったかどうかを判別する必要がある．しかし，今回収集した 3,000 万件以上のすべてのワクチン関連投稿を人手で調べることは非現実的である．そこで，ユーザが接種に至ったかどうか推測するための深層学習モデルを構築すること

4：https://www.niid.go.jp/niid/ja/diseases/ka/corona-virus/2019-ncov/2484-idsc/10569-covid19-53.html
5：Twitter データは NTT データより提供を受けたものである．

表 1: ワクチン関連投稿のアノテーション基準の概要.

クラス名	基準
接種報告	ワクチンを打った（これから打つ）ことの表明 自身の接種経験の報告（副反応など）
その他	ワクチンを打たない意思の表明 他者が接種するよう推奨・他者が打たないように注意喚起 客観的事実（接種者数など），公的機関等の報道のシェア ワクチン接種の是非が直接的な話題の中心にないもの

で，全ての投稿に機械的にラベルを付与した.

本研究では，投稿内容からそのユーザが実際にワクチン接種に至ったことがわかるものを「**接種報告**」クラスに，それ以外のワクチン関連投稿を「**その他**」クラスに分類を行う．表 1 に，今回設定したアノテーション基準の概要を示している.

この基準に従って，4 名のアノテータがワクチン関連投稿に対してラベリングを行った．まず，アノテータ間でのラベリングの一致度合いを測るために，4 名のアノテータが 500 件の同一の投稿に対してラベリングを行った．このデータに対する Fleiss's kappa 係数 [15] は 0.886 であり，4 人のアノテータ間のラベリングの一致度合いが高いことを確認した.

続いて，4 名のアノテータが分担して，2021 年 6 月から 10 月の間に投稿されたワクチン関連投稿からランダムに抽出されたデータに対してラベリングを行った．その結果，合計 7,000 件のラベル付けされた投稿を取得した．これらのデータを分類モデルの学習に使用した．また，本研究の分析対象期間が 2021 年 6 月から 2022 年 12 月であることを考慮し，4 名のアノテータのうちの 1 人が，ランダムに抽出された 2021 年の投稿 660 件および 2022 年の投稿 203 件に対して別途ラベル付けを行うことで，分類モデルの性能評価用のテストデータを作成した.

3.3 接種報告の分類

前節で述べたアノテーションデータ付きのワクチン関連投稿を用いて，本研究ではユーザが投稿したワクチン関連投稿が「接種報告」であるかどうかを分類するための深層学習モデルを構築した．ここでは，構築した分類器の構造と，分類器の学習および性能評価について述べる.

3.3.1 接種報告分類器の構造

本分類器は，先行研究 [3] と同様に，ワクチン関連投稿およびその投稿を行ったユーザが直近で行ったリツイート等のリアクションの情報を入力として，その投稿が「接種報告」か，あるいは「その他」であるかを判別する．まず，各投稿に対して全角半角変換，大文字小文字変換，各種記号の除去などの基本的な前処理を行った後，BERT [16] に入力することで，投稿内容に基づいた言語特徴量を作成した．なお，本研究では，国立研究開発法人情報通信研究機構 (NICT) の公開する日本語 BERT 事前学習モデル [6]に対して，日本語ワクチン関連投稿を用いて Domain-Adaptive Pretraining (DAPT) [17] を行った BERT を事前学習モデルとして使用した.

続いて，性能向上のために，当該ユーザがワクチン関連投稿を行った前後における，他ユーザとの間のリアクション関係の情報（リツイート，引用リツイート，リプライ）を表す特徴量（以下，RA ベクトルと呼ぶ）を作成した．具体的にはまず，自身のワクチン関連投稿に対する被リアクション数の多いユーザ（情報発信者），もしくは他ユーザのワクチン関連投稿へのリアクション数の多いユーザ（情報拡散者）を収集した．次に，分類の対象であるワクチン関連投稿の投稿日の直近 30 日間に，情報発信者にリアクションを行った，および情報拡散者にリアクションされた回数を要素とするベクトルを作成した．これを全結合層に入力することにより，RA ベクトルを作成した.

このようにして言語特徴量および RA ベクトル作成し，これらを結合し 1 つの特徴量ベクトルとした後，この特徴量ベクトルを全結合層に 2 度通すことで，その投稿が「接種報告」か「その他」のどちらに属するか予測を行う.

表 2: 接種報告分類器の性能（Precision, Recall, F 値）. ただし，“Tweets in 2021” は 2021 年の投稿からなるテストデータ，“Tweets in 2022” は 2022 年の投稿からなるテストデータに対する分類精度を示している.

	model	Prec.	Rec.	F_1
Tweets in 2021	BERT	0.852	0.917	0.883
	BERT+RAvec	0.873	0.917	**0.894**
Tweets in 2022	BERT	0.863	0.953	0.906
	BERT+RAvec	0.857	0.977	**0.913**

3.3.2 接種報告分類器の学習および評価

分類器の学習には，3.2 節で得られた 7,000 件のデータを使用した．この 7,000 件のデータを，学習データ 5,600 件と検証データ 1,400 件に分割した上で分類器の構築を行った．投稿の文字数制限を考慮し，ハイパーパラメタの一つである最大トークン数は 160 と設定した．また，学習の際には，BERT の各隠れ層を含めすべてのパラメタを更新した.

本研究で提案したモデル（BERT+RAvec）の有効性を確かめるために，RA ベクトルを用いないモデル（BERT）との比較を行った．また，2021 年と 2022 年の両方のワクチン関連投稿に対して本分類器が有効であるかどうかを検証するために，3.2 節で作成したテストデータを用いて，2021 年と 2022 年のそれぞれのワクチン関連投稿に対する分類精度の比較を行った.

正例を「接種報告」，負例を「その他」とした時の，テストデータに対するモデルの分類性能の評価指標（Precision, Recall, F 値）を表 2 に示す．まず，提案したモデル（BERT+RAvec）を用いた際の F 値は，RAvec を用いないモデルを用いた場合よりも上回っている．このことから，RA ベクトルが精度向上に貢献していることがわかる．また，提案手法の精度を見ると，2021 年と 2022 年の両方のテストデータに対しても F 値は約 0.9 であり，高精度で分類することに成功した．今回構築したモデルが，2021 年・2022 年の両方のワクチン関連投稿に対し，「接種報告」であるかどうかを高精度で分類することが可能であることが示された.

6：https://alaginrc.nict.go.jp/nict-bert/index.html

3.4 Twitter ユーザの年代推定

ユーザの接種行動を年代という観点から分析を行うためには，ユーザの年代を適当な方法で推定することが必要である．そこで本研究では，追加接種率が 50%を下回る年代 [7]でユーザを分割し，ユーザの年代が 30 歳未満であるか，あるいは 30 歳以上であるかの推定を行うモデルを構築した．

3.4.1 ユーザの年代推定用のデータの収集

まず，2021 年 6 月から 12 月の間に少なくとも 1 件以上の投稿をしたユーザ 20,168,637 件を抽出し，彼らが設定したプロフィール文を取得した．次に，取得したプロフィール文のパターンマッチング（「60 代半ばです」「28 歳.」など）により，年代を取得できたユーザ 25,720 件（上記の約 0.13%．30 歳未満：17,032 件，30 歳以上 8,688 件）を抽出した．続いて，抽出したユーザが先述した期間に投稿した全ての投稿から，ランダムにユーザごとに最大 10 件取得した．また，ユーザ情報として，アカウントができた時点の情報，他のユーザに対して行ったリアクション（リツイート，引用リツイート，メンション）に関する情報をそれぞれ取得した．

3.4.2 年代推定モデルの構造

今回構築したモデルは，投稿の内容とユーザの直近のリツイート等のリアクションに関するデータ，およびアカウントができた時点に関する情報を基に，そのユーザの年代（30 歳未満/30 歳以上）であるかを予測する．まず，ユーザが投稿したそれぞれの投稿を最大 10 件ランダムに抽出し，これらを結合したもの対して全角半角変換，大文字小文字変換，各種記号の除去などの基本的な前処理を行った後，NICT が公開する事前学習済み BERT [16] に入力することで，投稿内容に基づいた言語特徴量ベクトルを作成した．続いて，性能向上のため，被リアクション数が多かったアカウントに対してリアクションを行った回数を元に作成したユーザのリアクション情報を表す特徴量（以下，RA ベクトルと呼ぶ），およびユーザアカウントが作成された年に関する特徴量（以下，Created_at と呼ぶ）を作成した．言語特徴量，RA ベクトル，および Created_at 特徴量を作成した後，これらを結合し 1 つの特徴量ベクトルとした．そして，これを全結合層に 2 度通し，ユーザが 30 歳未満か 30 歳以上のどちらであるか予測する．

3.4.3 年代推定モデルの学習および評価

分類器の学習には，プロフィール文から年代を推定することができた 25,720 件のユーザのデータを使用した．この 25,720 件のデータを 8 : 1 : 1 の比率で，学習データ 20,576 件と検証データ 2,572 件，およびテストデータ 2,572 件にランダムに分割した上で，分類器の構築を行った．また，学習の際には，BERT の各隠れ層を含めすべてのパラメタを更新した．

本モデルの有効性を確認するために，先述した 3 つ全ての特徴量を用いる提案手法（+RAvec, Created_at）を含む 4 つのモデルとの性能比較を行った．ベースモデルは，投稿の言語特徴量のみを用いるモデル（BERT），言語特徴量と RA ベクトルを用いるモデル（+RAvec），および言語特徴量と Created_at 特徴量を用いるモデル（+Created_at）の 3 つとした．

表 3: Twitter ユーザの年代推定モデルの精度.

model	30 歳未満			30 歳以上		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
BERT	0.792	**0.874**	0.831	0.667	0.524	0.587
+RAvec	0.797	0.873	0.833	**0.672**	0.539	0.598
+Created_at	0.805	0.862	0.833	0.666	0.569	0.614
+RAvec, Created_at	**0.807**	0.862	**0.834**	0.668	**0.574**	**0.617**

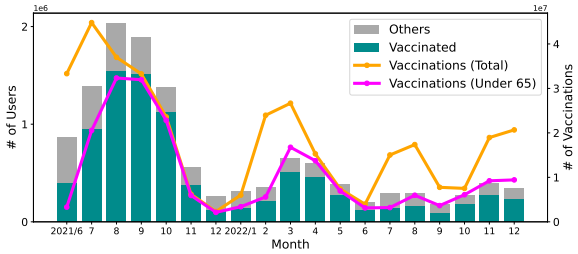


図 1: 2021 年 6 月から 2022 年 12 月における，ワクチン関連投稿を行ったユーザ数の分布と，COVID-19 ワクチン接種者数の推移.“Vaccinated” はその月に 1 回以上「接種報告」を投稿したユーザ数，“Others” は接種報告以外の「その他」のみを 1 回以上投稿したユーザー数を，“Vaccinations (Total)” は COVID-19 ワクチン接種者数，“Vaccinations(Under 65)” はそのうち 65 歳未満の接種者数を示している．

各手法のモデルの分類性能の評価指標（Precision, Recall, F 値）を表 3 に示す．各手法の F 値を比較すると，RA ベクトルと Created_at 特徴量の両方を使用する提案手法が最も高いスコアを示した．また，特に 30 歳以上の分類性能に注目すると，言語特徴量のみを使用する手法（BERT）と比べて，RA ベクトルまたは Created_at 特徴量を用いる手法，および提案手法では Recall の値が改善されたことがわかる．このことから，ユーザの年代を推定するにあたり，ユーザのアクション（リツイート，引用リツイート，メンション）に関する情報，およびアカウント作成時点に関する情報は，特に 30 歳以上のユーザの判別に対して効果があることが示された．

4 ワクチン関連投稿と国内の接種者数の関係の分析

RQ1（国内の Twitter ユーザの接種行動は，日本全体の接種状況を代表しているといえるか？）に回答するために，3.3 節で構築した接種報告分類器を用いてラベル付けしたワクチン関連投稿を用いて，ワクチン関連投稿と実際の国内の接種行動の関係を分析した．具体的には，それぞれのクラスに属するユーザ数および各月の接種回数の推移を基に，Twitter 上のワクチン関連投稿と接種回数との関係について分析を行った．

ワクチン関連投稿を行ったユーザ数および国内の COVID-19 ワクチン接種者数 [8]の推移を図 1 に示す．ワクチン関連投稿を行ったユーザ数は 2021 年 8 月や 2022 年 3 月に大きく増加し

7：首相官邸「年齢階級別接種実績」および総務省統計局「人口推計」をもとに算出（12 歳以上 30 歳未満：約 48%，12 歳以上 40 歳未満：約 52%）

8：https://data.vrs.digital.go.jp/vaccination/opendata/2023-01-01/prefecture.ndjson

ており，特に接種報告を行ったユーザ数に関してはその増加が顕著である．実際の接種者数の推移と照らし合わせると，上記の時期に初回接種や追加接種を受けた人が多く，これに伴い自身などの接種を報告する等の投稿が増加したと考えられる．一方，初回接種が主に行われていた 2021 年 6 月から 12 月にワクチン関連投稿を投稿したユーザ数と比較して，2022 年にワクチン関連投稿を行ったユーザ数は減少しているが，これもワクチン接種者数の減少と同様の傾向にある．

また，国内の 65 歳未満の人に対する接種回数と，接種報告を行ったユーザ数の推移についてみると，分析対象期間を通して増減の傾向が一致していることがわかる．実際，それぞれのクラスに属する各月のユーザ数と，65 歳未満のワクチン接種者数との間の相関係数を計算すると,「その他」クラスのユーザ数との相関係数は 0.649,「接種報告」クラスのユーザ数との相関係数は 0.966 であった．なお，高齢者を含めた全人口に対する接種回数よりも，そのうち 65 歳未満の人に対する接種回数との一致度合いの方が大きいのは，Twitter の利用年代が若年層に偏っている [9]ためであると考えられる．

通常，Twitter ユーザの分布には大きな偏りがあると考えられている [18] が，このことを踏まえると，Twitter ユーザでかつ接種に関連した投稿を行ったユーザ数（特に接種報告を行ったユーザ数）の推移が，実際の接種回数の推移と極めて一致しているということは注目に値する事象であるといえる．以上の結果は，国内の人々のワクチン接種行動と Twitter 上でのワクチン関連投稿，特に接種報告の間に強い相関関係があり，接種報告を行ったユーザ数の推移が実際の接種者数の推移とよく一致しているということを示している．

5　接種の継続性と情報共有行動の関係の分析

RQ2（接種を継続したユーザとそれ以外のユーザ間の違いは，Twitter 上での情報共有行動に現れているか？）および RQ3（ワクチン接種に関する情報共有行動の違いは年代により異なるか？）に回答するために，初回接種を報告したユーザ群の接種行動の継続性に着目した分析を行った．具体的には，2021 年に初回接種を報告したユーザ群を，追加接種が本格的に開始された 2022 年に接種報告を継続したかどうかに基づいて分類し，各ユーザ群の特徴の違いについて，参照・共有した情報源等に基づいて年代ごとに分析を行った．

5.1　ワクチン接種の継続性に基づいたユーザの分類

接種の継続性の分析にあたり，初回接種に至ったユーザ群のうち追加接種に至ったユーザ群，および至っていないユーザ群を推定する必要がある．そこで，初回接種に至ったユーザ群を接種の継続性に基づいて分類を行った．

具体的にはまず，2021 年に接種報告を行ったユーザ群のうち，2022 年に少なくとも 1 件以上の投稿を行ったユーザ 3,553,693 件を抽出した．実際，2021 年には多くの人々が 1,2 回目の接種を受け，かつ，3 回目の追加接種を受けた人はほとんど存在しないこと，および人々の接種行動と Twitter 上に投稿された接種報告の間の相関が強いこと（4 章）を踏まえると，このように 2021 年に接種報告を行ったユーザ群を抽出することで 1,2 回目の接種に至ったユーザ群を推定することが可能であると考えられる．

表 4: 2022 年のワクチン関連投稿の内容およびユーザの年代に基づいた，2021 年に接種報告を投稿したユーザ群の分類．

	30 歳未満	30 歳以上	合計
2022 年に接種報告あり	826,859	344,746	1,171,605
2022 年にワクチン関連投稿なし	1,747,979	462,931	2,210,910
2022 年に接種報告なし	103,210	67,968	171,178

続いて，2022 年のワクチン関連投稿の有無等を基に，抽出した 3,553,693 件のユーザを以下に示す 3 グループに年代ごとに分類した（表 4）．なお，2021 年に接種報告を行ったユーザの年代推定には，3.4 節で構築した年代推定モデルを使用した．

- 2022 年にも接種報告を行ったユーザ群
- 2022 年にワクチン関連投稿を行ったが，接種報告はしていないユーザ群
- 2022 年には接種報告を含めて 1 度もワクチン関連投稿を行わなかったユーザ群

2022 年にも接種報告を行ったユーザ数は，2021 年に接種報告を投稿したユーザの約 33%であった．このユーザ群は，2022 年の間に追加接種も受けて，その際の経験等を Twitter 上に投稿したと考えることができる．

一方で，2022 年にワクチン関連投稿を行ったが接種報告はしていないユーザ数は，2021 年に接種報告を投稿したユーザ群の約 5%であった．このユーザ群は，初回接種には至ったが追加接種を受けない決断をした，追加接種を受けるか迷っている，もしくは，追加接種を今後受ける予定ではあるものの 2022 年の間に接種に至らなかったユーザ群であると考えられる．

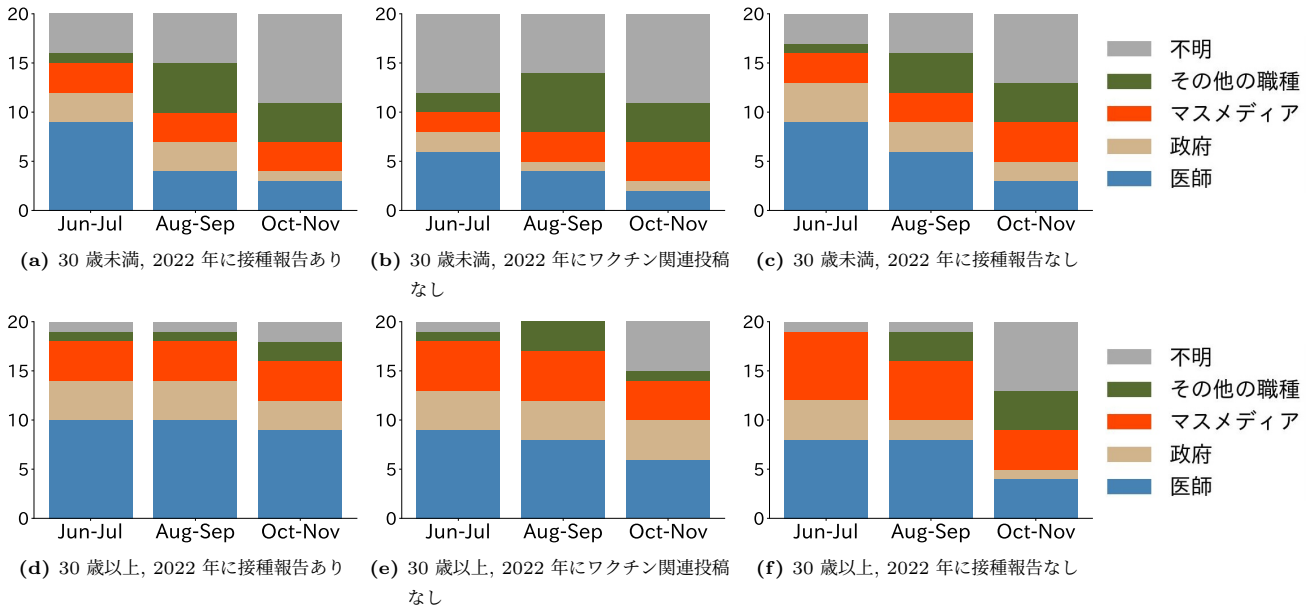
残りの約 60%のユーザ群は，2022 年には 1 度もワクチン関連投稿を行わなかったユーザ群である．表 4 より，2021 年に接種報告を行った 30 歳未満のユーザのうち，2022 年にワクチン関連投稿を行わなかったユーザ数は，2022 年に接種報告を行ったユーザ数の 2 倍以上であり，このことから，30 歳未満のユーザは 30 歳以上のユーザと比べて 2022 年にワクチン関連投稿を行わなかったユーザの割合が大きいことがわかる．追加接種を受けていない人々の多くが若年層であること，および日本国内の Twitter の利用年代が若年層に比較的偏っている [10]ことを踏まえると，このユーザ群の中にも追加接種に至っていない人が含まれていると考えられる．ただし，2022 年の間に追加接種には至ったものの，ワクチン関連投稿を 2022 年に行っていないユーザもこのユーザ群の中に含まれることに注意されたい．

5.2　ワクチン関連投稿を行った他ユーザに対するリアクションの比較

Twitter では，自身のツイートを発信するだけでなく，他ユー

9：https://www.soumu.go.jp/main_content/000887660.pdf

10：https://www.soumu.go.jp/main_content/000887660.pdf



(a) 30 歳未満, 2022 年に接種報告あり　(b) 30 歳未満, 2022 年にワクチン関連投稿なし　(c) 30 歳未満, 2022 年に接種報告なし

(d) 30 歳以上, 2022 年に接種報告あり　(e) 30 歳以上, 2022 年にワクチン関連投稿なし　(f) 30 歳以上, 2022 年に接種報告なし

図 2: 2021 年 6–11 月の期間において，2021 年に接種報告を行ったユーザ群に参照されたユーザ上位 20 件の推移.

ザの投稿に対してリツイート，引用リツイート，及びメンションの 3 種類のリアクションをすることができるが，いずれもユーザがその投稿に関心を持つことを示すといわれている [19]．そこで本研究では，3.1 節で収集したワクチン関連投稿に対するユーザ間アクションデータを用いて，2021 年のワクチン関連投稿に対する情報共有行動についてグループごとに比較を行った．

図 2 に，各グループのユーザが 2021 年 6 月から 11 月の間にリツイート，引用リツイート，メンションをしていたユーザ上位 20 件の職種の推移について示す．なお，ユーザの職種は，著者がプロフィール文を個別に確認し割り当てたものである．

5.2.1 2022 年に接種報告ありユーザ群と 2022 年にワクチン関連投稿なしユーザ群の比較

まず，30 歳未満のユーザ群について比較する．18 歳以上の全国民に対して接種が開始された 6 月から 7 月の期間を見ると，2022 年に接種報告を行ったユーザ群に参照されたユーザ上位 20 件の中には医師やメディア，政府系の公式アカウントが多く含まれていたのに対して，2022 年にワクチン関連投稿を行わなかったユーザ群ではその数は半数であることがわかる．この時期には，政府系アカウントによる公式見解や医師による専門的な解説，国内の接種状況等に関する報道などが多く発信されていたが，2022 年にワクチン関連投稿を行わなかったユーザ群は，こうした信頼性の高い情報源からの発信に対する関心が低い傾向にあったと考えられる．その後は，時間の経過とともに 2 つのグループの間での参照されるユーザ自体の違いは小さくなった．実際，10 月から 11 月の間に参照されたユーザ上位 20 件のうち 18 件が共通している．

各グループのユーザ群の関心が具体的にどのような情報に集まっていたのかを理解するために，実際にどのような内容の投稿が 30 歳未満のユーザ群に多く共有されていたのかについても確認した．まず，どちらのユーザ群からも多く参照されたアカウントの投稿を見ると，「接種の予約が取れない」「若年層向けの接種会場で行列ができている」など，接種希望者がスムーズに接種を受けることができていない状況に関する報道を発信していたメディアアカウントによる投稿や，こうした状況の中ですぐに接種できる会場等の情報を発信するなど，主に若者に対して接種を呼びかける医師による投稿が多く共有されていた．また，2022 年に接種報告を行ったユーザ群の方で比較的共有数が多かったアカウントを見ると，ワクチンの安全性，有効性，副反応に関する情報を繰り返し投稿していた政府系公式アカウントが含まれていた．一方で，2022 年にワクチン関連投稿を行わなかったユーザ群の方で比較的共有数が多かったアカウントを見ると，若者人気の高いアーティストを起用するなど，若者向けのキャンペーンを積極的に発信していた地方自治体の公式アカウントが含まれていた．上記のキャンペーンについて公式ホームページを別途確認すると，この自治体における若年層の初回接種率の低さを受けて 2021 年 10 月以降に始められたものであった．このように若年層をターゲットとしてキャンペーンを展開したことにより，一定数の若年層のユーザの関心を集めることができていたと考えられる．以上の結果は，若年層が関心を示すように適切な対策を講じつつワクチン接種を呼びかけたり，ワクチンに関する正しい情報を効果的な時期に発信することの重要性を示唆している．

続いて，2021 年に接種報告を行った 30 歳以上のユーザ群について比較すると，彼らに参照されたユーザ上位 20 件のうち多くのアカウントが両グループで共通していた．また，30 歳未満のユーザ群と比較して，医師やメディア，政府系の公式アカウントの投稿に対する共有が多く見られており，こうした情報に対する関心が若年層のユーザよりも高かったと考えられる．一方で，2021 年 6–7 月と 10–11 月の期間を比較すると，2022 年にワクチン関連投稿を行わなかったユーザ群の方では医者のアカウントに対する参照が減少し，職種不明ユーザに対する参照が増加していることがわかる．このことは，医者が発信する

ワクチンに関連した情報に対する彼らの関心が時間の経過とともに薄れていったことを示唆している．

5.2.2 2022 年に接種報告ありユーザ群と 2022 年に接種報告なしユーザ群の比較

まず，30 歳未満のユーザ群について比較すると，彼らに参照されたユーザ上位 20 件のうち，医師やメディア，政府系の公式アカウントを含めて，多くのアカウントが共通していた．このことは，医師やメディア，政府系の公式アカウント等の信頼性の高い情報源からの発信に対する関心が 2 つのユーザ群において同程度であったことを示唆している．

一方，30 歳以上のユーザ群を比較すると，初回接種開始当初の 2021 年 6–7 月には，医師やメディア，政府系の公式アカウントの投稿に対する共有が両グループで多くみられたのに対して，10-11 月以降には，2022 年に接種報告を行わなかったユーザ群においてこれらのアカウントに対する参照が減少し，職種不明ユーザに対する参照が増加していることがわかる．これらの職種不明ユーザの内訳を確認すると，2022 年に接種報告を行わなかったユーザ群に参照されたユーザ上位 20 件の中には，ワクチンに関する不正確な情報や接種に対する不安を過剰に煽る投稿を積極的に発信していたアカウントが複数含まれていた．このことから，2022 年に接種報告を行わなかったユーザ群はこのような情報に触れる機会が比較的多かったと考えられる．

SNS 上の COVID-19 ワクチンに関する情報，特にデマやフェイクニュースなどの誤った情報への接触が接種のためらいに繋がることが複数の研究で報告されている [20, 21] ことを踏まえると，このユーザ群の中には，初回接種には至ったものの，SNS 上でワクチンに関する様々な情報に触れるうちに，接種に対するスタンスが変化してしまった，つまり，追加接種を受けないという決断に至ったユーザが一定数存在する可能性があると考えられる．以上の結果は，デマやフェイクニュースなどといったワクチンに関する SNS 上の誤った情報に対して，不適切な投稿の削除などの適切な対応を行う，政府等の信頼できる情報源による正確な情報に触れる機会を増やすような仕組みを導入するなどの対策を施すことの重要性を示唆している．

5.3 ワクチン関連投稿内で共有された外部サイトの比較

前節では，どのようなユーザのワクチン関連投稿を参照していたのかについて分析した．一方で，Twitter ユーザは他のユーザからだけでなく，外部サイトから情報を獲得する可能性がある．そこで，2021 年に接種報告を行ったユーザ群が参照した外部サイトに関する分析を行った．

各ユーザ群が 2021 年のワクチン関連投稿内で共有した外部サイト上位 20 件の推移を図 3 に示す．なお，1 ユーザが同じリンクを複数回共有したものに関しては一度のみカウントした．

5.3.1 2022 年に接種報告ありユーザ群と 2022 年にワクチン関連投稿なしユーザ群の比較

まず，30 歳未満のユーザ群について比較すると，どちらのグループにおいても，よく共有されたサイト上位 20 件のうちテレビや Web ニュースが減少し，他の SNS や動画共有サイト等の割合が増加するなどの変化が見られること，特に 2022 年にワクチン関連投稿を行わなかったユーザの方ではこの傾向が顕著に見られることがわかる．2022 年に接種報告を行ったユーザ群の方では，テレビや Web ニュースに対する参照だけでなく，掲示板や個人ブログに対する参照も期間を通して一定数見られることも特徴である．一方で，2022 年にワクチン関連投稿を行わなかったユーザ群の方に着目すると，他の SNS や動画共有サイトの共有が 8–9 月以降目立つようになり，特に 10–11 月の期間ではよく共有されたサイト上位 20 件のうちの大半が他の SNS もしくは動画共有サイトであった．実際にこれらの SNS や動画共有サイトの共有内容について確認してみると，動画配信者などの影響力のあるインフルエンサーによる接種を報告する内容が多く含まれていた．このようにワクチン接種の感想などをインフルエンサーが共有することは，ファンの接種に対する意識に影響を与える可能性があると考えられる．また，このように SNS や動画共有サイトに対する参照が大部分であることは，彼らの意識が接種の安全性や有効性といった，接種自体への関心から離れていったことを示唆している．

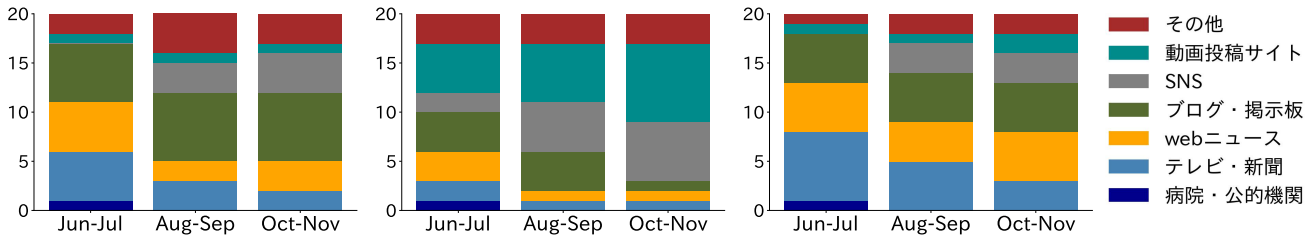
次に，30 歳以上のユーザ群について比較する．2022 年に接種報告を行ったユーザ群の方では，テレビや Web ニュースがどの期間においても半数以上を占めており，最新の情報に敏感に反応していることがわかる．一方で，2022 年にワクチン関連投稿を行わなかったユーザ群の方では，掲示板やブログへの参照が期間を通して比較的目立つ．また，10–11 月には，30 歳未満のユーザ群に見られた傾向と同様に，他の SNS や動画共有サイトに対する参照も目立っており，彼らの情報源が時間の経過とともに多様化していることがわかる．

以上の結果をまとめると，2022 年にワクチン関連投稿をしなくなったユーザ群の外部サイトからの情報源は時間の経過とともに多様化し，掲示板や個人ブログ，他の SNS や動画投稿サイトなども多く参照されていたといえる．このことは，この結果はテレビやニュースメディアだけでなく多様なメディアにおいて，COVID-19 ワクチンに関する信頼性の高い情報を提供することが必要であったことを示唆している．

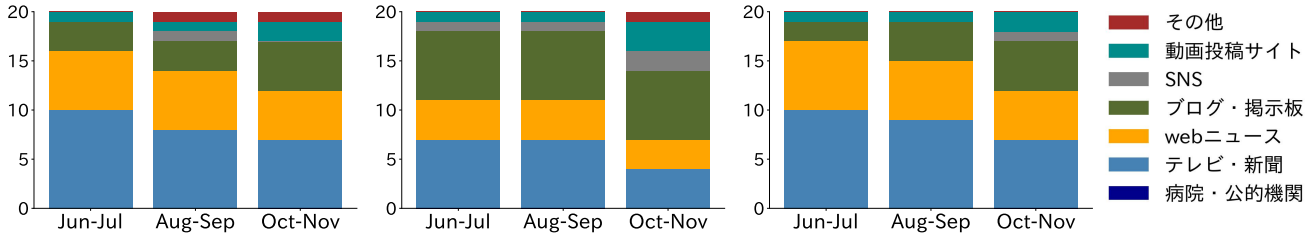
5.3.2 2022 年に接種報告ありユーザ群と 2022 年に接種報告なしユーザ群の比較

まず，30 歳未満のユーザ群について比較すると，接種開始当初である 2021 年 6–7 月において，テレビや Web ニュースに対する参照は 2022 年に接種報告を行わなかったユーザ群の方で比較的多かった．一方で，このユーザ群によって参照された外部サイト上位 20 件のうちのテレビや Web ニュースの割合は時間の経過とともに減少する傾向にあり，2022 年 10–11 月の期間では，参照された外部サイト上位 20 件のうちこれらの割合は 2022 年に接種報告を行ったユーザ群と同程度であった．

また，各ユーザ群の関心が具体的にどのような情報に集まっていたのかを理解するために，多く共有された外部サイトのタイトル・概要についても別途確認を行ったが，多く共有されたサイトの傾向も期間を通して同じであった．これらの結果と前節の他ユーザに対するリアクションを比較した結果と併せると，これら 2 つのユーザ群に関しては Twitter 上でのワクチンに関連した情報共有行動において大きな違いはなかったと言える．



(a) 30 歳未満, 2022 年に接種報告あり　**(b)** 30 歳未満, 2022 年にワクチン関連投稿なし　**(c)** 30 歳未満, 2022 年に接種報告なし



(d) 30 歳以上, 2022 年に接種報告あり　**(e)** 30 歳以上, 2022 年にワクチン関連投稿なし　**(f)** 30 歳以上, 2022 年に接種報告なし

図 3: 2021 年 6–11 月の期間において，2021 年に接種報告を行ったユーザ群に参照された外部サイト上位 20 件の推移.

続いて，30 歳以上のユーザ群について比較すると，参照された外部サイト上位 20 件に含まれているサイト種別に関しては，期間を通して大きな違いはなかった．一方で，多く共有された外部サイトのタイトル・概要についても別途確認を行うと，2022 年に接種報告を行ったユーザ群では，追加接種などの今後の接種スケジュールや，国内の接種率などの接種状況に関するニュースが比較的多く共有される傾向があったのに対して，2022 年に接種報告を行わなかったユーザ群では，ワクチンによる重篤な副作用やワクチン後遺症に関する情報に関するニュース記事および動画など，ワクチン接種に対する過剰な不安に繋がるような情報が比較的多く共有されていたことを確認した．

前節の他ユーザに対するリアクションを比較した結果と併せると，2022 年に接種報告を行わなかったユーザ群は，COVID-19 ワクチンに関する不正確な情報やワクチン接種に対する不安を引き起こすような情報を SNS 上で参照，および共有する回数が他のユーザ群よりも多く，このことが一因となって初回接種後に接種スタンス自体が変化してしまったユーザがこのユーザ群に一定数存在する可能性があると考えられる．以上の結果は，テレビやニュースメディアなど多様なメディアにおいて，厚生労働省などの信頼できる情報源からの COVID-19 ワクチンに関する正しい情報を発信していくこと，および人々がワクチンに関する情報を収集する際に，信頼できる情報源からの正確な情報をすぐ参照できるようにすることの重要性を示唆している．

6 おわりに

本研究の目的は，Twitter データの分析を通して，国内の COVID-19 ワクチン接種行動の推移や，追加接種数が伸び悩んだ背景について明らかにすることである．この目的を達成するために，2021 年 6 月から 2022 年 12 月に投稿された日本語全量ツイートのうちワクチンに関連した投稿を収集するとともに，投稿を行ったユーザが実際に接種に至ったかどうかを推定する分類器，およびユーザの年代を推定する分類器の 2 種類の深層学習モデルを開発した．これらを用いることで，自身の接種を報告したユーザ数と実際の接種者数との間の関係について分析した．また，追加接種が本格的に開始された 2022 年に接種報告を継続して行ったかどうかに基づいて初回接種を報告したユーザ群を分類し，これら 2 つのユーザ群の特徴の違いについて，参照・共有した情報源等に基づいて年代ごとに分析を行った．

ワクチン関連投稿と接種者数との関係についての分析（4 章）では，接種回数が多い月にワクチン関連投稿数が増加している一方，初回接種が行われていた 2021 年のワクチン関連投稿数と比較して，2022 年の投稿数は接種者数と同様に減少したことを示した．また，接種報告を行ったユーザ数と国内の 65 歳未満の接種者数の相関が非常に高く，接種報告を行ったユーザ数の推移が接種者数の推移とよく一致していることを示した．

また，接種の継続性と情報共有行動等の関係の分析（5 章）では，2021 年に接種報告を行ったユーザ群を接種行動の継続性に基づいて分類し，これらのユーザ群の特徴の違いについて，参照した情報源等を基に年代ごとに比較を行った．その結果，2022 年にワクチン関連投稿を行わなかったユーザ群のうち，特に 30 歳未満のユーザ群において，医者やメディア，政府アカウントなどの信頼性の高い情報源からの発信に対する関心が接種開始当初から比較的低い傾向にあったことが分かった．また，2022 年に接種報告を行わなかったユーザ群のうち，特に 30 歳以上のユーザ群において，ワクチンに関する不正確な情報やワクチン接種に対する不安に繋がるような情報を SNS 上で比較的多く参照，共有する傾向にあったことが分かった．以上の結果から，特に若年層がワクチンに対する関心を示すように対策を講じつつ，多様なメディアにおいてワクチンに関する正しい情報を発信していくこと，およびワクチンに関する不正確な情報に接触した際に，信頼できる情報源からの正確な情報をすぐ参照できるようにすることが重要であると考えられる．

今後の展望として，まず，初回接種に至った人々が継続して接種に至るかどうか以外の様々なワクチン接種行動の変化を対象に分析を行い，国内の接種行動について SNS を通してより詳細に分析することが挙げられる．本研究では，2021 年には接種に至らなかったが 2022 年に接種を決断したユーザや，2021 年から一貫して接種行動に至らなかったユーザに関しては分析できていない．これらの接種行動の変化およびその要因について今後さらに分析することで，国内におけるワクチン接種行動の推移の全体像を把握することが可能になると考えられる．また，本研究では国内の接種行動と Twitter 上の投稿の関係について分析したが，接種行動だけでなく，感染状況や政府の施策の変化等に応じた人々の行動の変化について分析することも今後の課題であると考えている．2023 年には新型コロナウイルスの感染症法上の位置づけが 5 類へ移行し，これに応じて人流の増加や「脱マスク」などの個人の感染対策の変化が見られている．こうした行動の変化と強く結びつく SNS 上の投稿をモニタリングすることを通して，今回の感染症下における人々の種々の行動の変化とその要因を把握することは，今後の感染症対策において重要であると考えられる．

謝　　辞

本研究は，三菱総合研究所が内閣官房の委託を受けて推進している COVID-19 AI・シミュレーションプロジェクトの一環として実施したものです．本研究で使用している分析手法は，JST CREST JPMJCR19A4，および，JSPS 科研費 JP21H03445 の支援を受けて開発したものです．

文　　献

[1] N. Barda, N. Dagan, C. Cohen, M. A. Hernán, M. Lipsitch, I. S. Kohane, B. Y. Reis, and R. D. Balicer. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. *The Lancet*, Vol. 398, No. 10316, pp. 2093–2100, 2021.

[2] Y. M. Bar-On, Y. Goldberg, M. Mandel, O. Bodenheimer, L. Freedman, S. Alroy-Preis, N. Ash, A. Huppert, and R. Milo. Protection against Covid-19 by BNT162b2 booster across age groups. *New England Journal of Medicine*, Vol. 385, No. 26, pp. 2421–2430, 2021.

[3] S. Hisamitsu, S. Cho, H. Jin, M. Toyoda, and N. Yoshinaga. Diachronic Analysis of Users' Stances on COVID-19 Vaccination in Japan using Twitter. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp. 237–241, 2022.

[4] S. Poddar, M. Mondal, J. Misra, N. Ganguly, and S. Ghosh. Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter users. In *Proceedings of the International Conference on Web and Social Media*, Vol. 16, pp. 782–793, 2022.

[5] R. Kobayashi, Y. Takedomi, Y. Nakayama, T. Suda, T. Uno, T. Hashimoto, M. Toyoda, N. Yoshinaga, M. Kitsuregawa, and L. E. C. Rocha. Evolution of Public Opinion on COVID-19 Vaccination in Japan: Large-Scale Twitter Data Analysis. *Journal of Medical Internet Research*, Vol. 24, No. 12, p. e41928, 2022.

[6] L.-A. Cotfas, C. Delcea, and R. Gherai. COVID-19 vaccine hesitancy in the month following the start of the vaccination process. *International Journal of Environmental Research and Public Health*, Vol. 18, No. 19, p. 10438, 2021.

[7] P. J. Ramos, K. Ferawati, K. Liew, E. Aramaki, and S. Wakamiya. Emotion Analysis of Writers and Readers of Japanese Tweets on Vaccinations. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pp. 95–103, May 2022.

[8] M. Weinzierl and S. Harabagiu. Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse. In *Proceedings of the Web Conference*, pp. 3196–3205, 2022.

[9] S. M. Driedger, G. Capurro, J. Tustin, and C. G. Jardine. "I won't be a guinea pig": Rethinking public health communication and vaccine hesitancy in the context of COVID-19. *Vaccine*, Vol. 41, No. 1, pp. 1–4, 2023.

[10] S. Praveen, J. M. Lorenz, R. Ittamalla, K. Dhama, C. Chakraborty, D. V. S. Kumar, and T. Mohan. Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding COVID-19 booster vaccine shots in India: Crucial to expanding vaccination coverage. *Vaccines*, Vol. 10, No. 11, p. 1929, 2022.

[11] C. Delcea, L.-A. Cotfas, L. Crăciun, and A. G. Molănescu. New Wave of COVID-19 Vaccine Opinions in the Month the 3rd Booster Dose Arrived. *Vaccines*, Vol. 10, No. 6, 2022.

[12] S.-Q. Ong, M. B. M. Pauzi, and K. H. Gan. Text mining and determinants of sentiments towards the COVID-19 vaccine booster of twitter users in Malaysia. In *Healthcare*, Vol. 10, p. 994, 2022.

[13] T. Daghriri, M. Proctor, S. Matthews, and A. H. Bashiri. Modeling Behavior and Vaccine Hesitancy Using Twitter-Derived US Population Sentiment during the COVID-19 Pandemic to Predict Daily Vaccination Inoculations. *Vaccines*, Vol. 11, No. 3, p. 709, 2023.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[15] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, Vol. 33, No. 1, p. 159, 1977.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[17] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

[18] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist. Understanding the demographics of Twitter users. In *Proceedings of the International Conference on Web and Social Media*, Vol. 5, pp. 554–557, 2011.

[19] P. Metaxas, E. Mustafaraj, K. Wong, L. Zeng, M. O'Keefe, and S. Finn. What do retweets indicate? Results from user survey and meta-review of research. In *Proceedings of the International Conference on Web and Social Media*, Vol. 9, pp. 658–661, 2015.

[20] S. R. Neely, C. Eldredge, R. Ersing, and C. Remington. Vaccine hesitancy and exposure to misinformation: a survey analysis. *Journal of General Internal Medicine*, pp. 1–9, 2022.

[21] P. L. Liu, X. Zhao, and B. Wan. COVID-19 information exposure and vaccine hesitancy: The influence of trust in government and vaccine confidence. *Psychology, Health & Medicine*, Vol. 28, No. 1, pp. 27–36, 2023.

オープニング・クロージング | 全体イベント

2024年3月1日(金) 9:30 ~ 9:50 | オンライン・オープニング会場 オンライン（Zoom Events)

オンライン・オープニング (3)

9:30 ~ 9:50

[3OO-0930-1add] 参加案内

オープニング・クロージング | 全体イベント

オンライン・オープニング (3)

2024年3月1日(金) 09:30 〜 09:50 オンライン・オープニング会場 (オンライン（Zoom Events))

09:30 〜 09:50

[3OO-0930-1add]参加案内

©一般社団法人日本データベース学会