関西大学審査学位論文

マイクロブログを活用した社会事象の抽出と分析技術に関する研究

Research Concerning the Technique for Extracting and Analyzing Social Phenomena Using Microblogs

2021 年　3 月

坂本　一磨

関西大学大学院総合情報学研究科

要旨

要旨

我が国では，「第 5 期科学技術基本計画」にてサイバー空間とフィジカル空間を融合させた新たな社会像である Society 5.0 が提唱され，ネットワーク技術や IoT（Internet of Things），そして AI（Artificial Intelligence）技術を用いて，経済成長に向けた社会課題を解決するための取り組みがなされつつある．その中でも，インターネットから社会の動向やニーズを把握することを目的としたソーシャルセンシング技術の研究が進められており，社会調査やマーケティング，データマイニング等の分野で活用されている．最近では，インターネットからユーザの行動をリアルタイムに分析することで，その時々のユーザの意見や暗黙的な考え，そして本来のニーズをタイムリーに抽出することも注目されている．例えば，東日本大震災では，CGM（Consumer Generated Media）の一つである SNS（Social Networking Service）を用いて，緊急性が高い安否確認や，地域に密着した避難所の設置場所や利用率，そして備蓄食料や資材の情報を集約するために利用された．このことから，SNS は，サイバー空間とフィジカル空間を融合したきめ細やかな情報交換手段として有効であることが確認されている．しかし，投稿記事の確からしさが保証されていないため，非常事態において重要性が高くしかも信頼性も担保された地域密着型の情報を随時正確に抽出することは非常に難しい．

そこで，本研究では，平時の状況を把握することで，異常時や緊急時に素早く対応するためのソーシャルセンシングに着目し，実社会の新たな事象や事変を感度良く見極めることに主眼を置く．具体的には，SNS 上の投稿時間と投稿内容からユーザ毎の日々の生活習慣に応じた「習慣行動」の情報を基に，実世界で生じた事象や動向を適切に抽出するための方法について議論する．一つ目としては，平時の習慣行動と異なる「非習慣行動」を抽出することでユーザの行動を分析する手法を，二つ目として，ユーザの属性単位（性別，年代，職業や地域）による習慣行動の違いから社会事象を抽出する手法を深く検討する．さらに，三つ目として，ソーシャルセンシングで重要となる前述の 4 つのユーザ属性を投稿履歴から推定するためのシステムを開発し，センシング精度を高める技術を考究する．

まず，SNS 上の投稿時間と投稿内容からユーザの習慣行動を取得し，これに基づき非習慣行動を抽出する手法に着目する．SNS には，日常会話のみならず広告目的などの雑多な情報が含まれるため，ノイズを含まずにユーザの状況やニーズを機械的に取捨選択することは難しい．たとえ特定のキーワードを用いて分類する場合でも，それらを選定する時点で恣意的な結果となり，網羅的に世の中の状況を捉えることができない．そこで，本研究では，ユーザの過去の投稿履歴から抽出できる習慣行動に着目し，平時とは異なる行動を非習慣行動として区別しながら，単語の共起確率と出現頻度を考慮したトピックから解釈できる社会事象を抽出する方法を提案した．本手法では，「平時と異なる非習慣行動を引き

要旨

起こすユーザ群を特定することで，その非習慣行動から何らかの大きな社会事象が発生している」という仮説を設定し，実際に多種多様な実世界で起きた事象を発見できるかを試みる．実験結果から，1 年間の平均化した習慣行動を基準として，1 ヶ月毎の習慣行動を比較することから非習慣情報である社会事象を抽出できるかを確認した．さらに，特定キーワードを指定する既存手法よりも多くの社会事象を発見することができる上，同一事象においても複数の関連記事を抽出することができた．したがって，網羅的にしかも詳細に世の中の状況を把握することに成功した．

次に，ユーザの属性毎による習慣行動の違いを考慮した社会事象の抽出手法に着目する．SNS のユーザは，性別，年齢，職業や居住している地域が異なるため，意見やその反応も多様である．そこで，先の方法にユーザ属性を考慮して，「属性毎の重要な記事を発見することができれば，個別のユニークな社会事象を検出できる」との仮定の下，研究を遂行した．具体的には，性別や年齢に関係したトピック記事や，職業に関連するタイムリーな記事，そして地域特性を持った重要記事等を獲得できるかを試みる．ユーザ属性としては，プロフィール欄に記載されているユーザ自身のプロパティ情報を用いた．実験結果からは，各ユーザ属性によって注目する投稿記事が異なること，そして，同一事象でもその内容に差異があることを確認した．さらに，属性毎の日々の生活に直結する行動スケジュールも取得できることがわかった．

最後に，前述のユーザのプロパティ情報を自動的に獲得することを目的に，投稿内容からユーザ属性を推定する技術に着目する．本研究では，性別，年代，職業のユーザ属性を順番に推定する段階的詳細化の考え方を用いる．地域属性に関しては，深層学習の回帰型ニューラルネットワークの一種である BiLSTM（Bidirectional Long Short Term Memory）を用いて地域辞書モデルを構築し，新たな投稿記事から何処の地域のユーザであるかを特定する．実験では，性別ごとの特徴的な語句を用いて性別を判別した後，段階的詳細化により年代と職業を絞り込みながら推定できることがわかった．また，地域辞書モデルを用いることで地方区分レベルの情報を獲得できることも確認した．最終的に，前述の結果と自動的に推定したユーザ属性を用いた結果とを比較し，ほぼ同じ内容の社会事象を抽出できたことも明らかにした．したがって，一連の研究においてソーシャルセンシング技術の有用性を証明することができた．

1)　非習慣行動を用いた社会事象の抽出方法の提案

マイクロブログユーザをソーシャルセンサと捉えて，特定のキーワードの出現数や投稿記事の文脈を解析し，社会事象を検知するソーシャルセンシングに関する研究が注目されてきた．しかし，各事象に合った特定のキーワードを事前に指定する必要があるために抽出できる内容の領域が狭いことや，キーワードの選定に解析者のバイアスがかかり，その結果，内容に偏りが生じることにより網羅的な分析がなされていない．そこで，ユーザの習慣行動の情報を用いて，実世界における社会事象を抽出する新たなソーシャルセンシング技術を考案する必要がある．

本研究では，ユーザの投稿履歴を用いてユーザの習慣行動を解析し，時間毎の単語の出現回数を用いた特徴ベクトルを作成する．そして，1 年間の生活習慣を基準として，その平時の行動から月単位の習慣行動の差を比較できるかを検討する．最終的に，「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「その社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，本提案手法が有用であるかを確認する．

約 300 万件の実験結果から，平時の行動と 1 ヶ月毎の習慣行動を比較することで，平均して約 97 時間の非習慣行動を抽出できることがわかった．また同時に，その非習慣行動に関連するトピックを抽出することができた．それは，事前にキーワードを指定する手法で獲得できなかったものも含まれていることを確認した．本技術は，キーワードを指定する必要がなく，その上，日本語以外へも対応可能である点に新規性があり，また，即時性と網羅的の課題も解消できたことに有用性がある．ただし，ユーザの属性毎のトピック抽出には至っていない．この点には課題が残った．

2)　ユーザ属性を考慮した非習慣行動を用いた社会事象の分析

マイクロブログは，誰もが手軽に記事を投稿できる利便性がある．そのため，ユーザによって，興味や関心のある事象が異なると考えられる．そこで，ユーザ毎にそれらの違いを区別できる可能性があると考え，ユーザ属性を考慮した実世界で生じた社会事象の分析を検討する．

本研究では，まず，ユーザの属性に関しては，Twitter のプロフィール欄と投稿内容を確認してマニュアルで付与することとした．次に，ユーザ属性毎の一年間の習慣行動と月毎の習慣行動を比較することで非習慣行動を抽出する．そして，ユーザの行動傾向に沿った特定の慣習に関する社会事象を抽出できるかを議論する．

前述の約 300 万件の内，属性がわかっている約 200 万件を用いた実験結果から，ユーザ属性毎の感度の高い社会事象を獲得できることがわかった．また，ユーザ属性毎の日々の生活に直結する行動パターンの内容を収集できることから，トピックに明確な違いがあることも確認できた．例えば，同じオリンピックの話題であっても性別，職業（会社員，公務員，自営業，学生，主婦やフリーター）と地域（北海道・東北地方や関東地方など）によって注目するトピックが異なることを示すことができた．したがって，先の大局的なデータセンシングから局所的なデータセンシングの可能性を見出し，属性毎の社会事象の抽出の有効性について明らかにした．課題としては，性別，年代，職業などの属性にはパターン性があるが，地域性に関しては投稿記事の中身を詳細に分析する必要があることがわかった．そのため，行動パターンから読み取れない地域属性に関しては，新たな分析方法が必要であることが明らかになった．

3)　ユーザの基本属性と類語による地域属性の自動獲得による社会事象の最終分析

属性毎の異なる生活習慣を考慮しながら段階的詳細化の考えからユーザの基本属性を推定する．これは，投稿記事からユーザの確度の高い性別の属性から年代の属性，そして確度の低い職業の属性へと段階的に確定できるかを明らかにするものである．

要旨

一方，地域属性に関しては，習慣行動や投稿傾向に顕著な特徴が見受けられないことや，地域に関する特徴となる単語の出現率が低いことから推定は難しい．また，ジオタグの位置情報の付与率も非常に少ない．そこで，同じ意味を表すが地域毎に表現が異なる類語に着目し，その違いから地域属性を推定する手法を提案する．投稿者は類語を無意識に使用する傾向があると考え，類語と地域毎に出現頻度が異なる地域語を組み合わせた「地域辞書」を用いて，派生する関連語や関連記事から地域推定を試みる．具体的には，BiLSTM を用いて地域毎に地域辞書モデルを構築して，投稿記事から地域属性を推定する．

実験では，性別，年代や職業が明らかなユーザの約 200 万件と，地域がわかっているユーザの約 600 万件の投稿記事を対象とした．実験結果より，1）段階的詳細化によってユーザの性別，年代や職業の基本属性は比較的簡単に確定できることがわかった．一方，地域属性に関しては，2）地域辞書モデルを用いることで大まかな地方区分レベルに分類できること，そして3）ユーザの地域語を増加させることで地域辞書が洗練され，地方区分から都道府県レベルへと収斂する傾向があることがわかった．したがって，これまでマニュアルで行ってきたユーザ属性を分類する手間が省け，自動的にユーザ属性を獲得できること，そして，非習慣行動を用いた社会事象の分析が比較的簡単になることがわかった．

さらに，最終確認として，マニュアルで付与したユーザ属性と自動的にユーザ属性を推定した時の非習慣行動によるトピックを比較することにより，同様の社会事象を獲得できることを確認した．

以上のことから，社会事象の抽出においてマイクロブログと習慣行動の情報は有用であることが明確となった．また，ユーザ属性を考慮することで詳細な社会事象を抽出できることもわかった．したがって，平時ではない非習慣の社会動向の情報を検索し，タイムリーに取捨選択するためのソーシャルセンシング技術の高度化について深く議論し，実世界における有益な情報を適切に抽出することを可能とした実践的な研究である．

今後は，ユーザ属性の分類を詳細化し，より汎用性の高いモデルを構築する．また，投稿者の性別，年代，職業，そして地域ごとの平時習慣と特定習慣のタイムスパン（年，月，週，日）の組合せと抽出される社会事象との関係を明らかにし，投稿記事の信頼性を判断しながら平時時と非習慣時のデータセンシング技術の確立を目指す．

目次

目次

目次

第 1 章

緒論

第1章　緒論

1.1　研究の背景

情報処理技術や情報機器の発展により，多様で膨大なデータがインターネット上に日々蓄積されている．我が国では，政府が策定した「第 5 期科学技術基本計画[1]」においてSociety5.0[2]の方針が掲げられ，多種多様な情報の抽出と分析から，領域を超えた交換，連携，共有，再利用することを目指している．特に，サイバー空間とフィジカル空間を融合させ，様々なデータから社会的ニーズの高い情報を抽出することが求められている．また，CGM（Consumer Generated Media）の一つである SNS（Social Networking Service）は，幅広い世代で利用され，あらゆる情報が蓄積されており，サイバー空間からフィジカル空間の情報を抽出することが可能である．具体的には，東日本大震災以降，即時性のある情報交換, 連携, 共有及び利活用の場面で SNS の情報の有用性については非常に注目されている．

内閣府においては，災害時[3]，[4]や消費生活に関する研究会[5]など，SNS の利活用を目指した検討会が発足した．これらの検討会は，大規模災害などでリアルタイムに国民の状況判断や要望に対する決断に即座に対応できなかった現状を改善することが目的である．また，雑多なデータが含まれている上，情報量が莫大であるため，問題に直面している被災者の情報を的確に収集することができない現状を改善する．すなわち，重要な情報を見逃さないことに主眼が置かれており，社会動向，経済動向に加えて，事件，事故，大規模災害などの様々な事象を計測して，重要な情報を獲得するためのソーシャルセンシング技術の確立が期待[6]されている．一方で，最近ではフェイクニュースのような情報が蔓延することから，前述の事項の改善には，ユーザの匿名性が高い SNS などのサイバー空間の情報に対して，匿名性を排除してフィジカル空間へ情報転換することで信頼性の担保が必要である．

信頼性を担保する手段の一つとして，SNS などのサイバー空間の情報から，ユーザの性別，年代（年齢）や職業，そして地域などのプロパティ情報（属性情報）を推定し，尚且つユーザの投稿履歴や URL などから個人の特徴を推測する取り組みが注目されている．これにより，各属性に異なるスポーツイベント情報の提示や，鉄道や交通渋滞の遅延や事故に関する情報共有など，サイバー空間からフィジカル空間への情報転換において信頼性の向上を図り，高度なソーシャルセンシング技術の確立に寄与できると考えられる．

このような背景の下，ソーシャルセンシングに関する研究が盛んであり，マイクロブログのユーザをソーシャルセンサ[7]として捉えてセンシングする幅広い研究が行われている．しかし，解析対象を特定のユーザやコミュニティに限定していることや，分析対象のトピックを特定の商品やサービス，イベントなどに限定していることから，広範囲な視点からの

社会事象の把捉ができていない．そこで，社会動向，経済動向に加え，事件，事故，大規模災害などの多種多様な事象（以下，社会事象）を計測するソーシャルセンシングが注目されている．ソーシャルセンシングでは，ローカルイベントやゲリラ豪雨をユーザの投稿からリアルタイムかつ速やかに検出できる．ソーシャルセンシングに関する研究は，大きく 3 つに分けることができる．1 つ目は，検索エンジンの検索履歴を用いた手法である．既存研究[8]，[9]では，検索エンジンの検索履歴を用いて，インフルエンザの流行を抽出する研究が行われているが，解析対象の検索履歴の入手が困難であるため，多くの研究は行われていない状況である．

2 つ目は, ブログを対象とした手法である. 選挙得票と株式市場を予測する研究[10]では, 特定のキーワードが含まれるブログの記事数や相場の上昇と下落時を表す特徴的な単語を用いて事象を予測している．この研究では，ブログの特性上，イベントからユーザの投稿までのタイムラグが発生するため，即時性の課題があった．

3 つ目は，マイクロブログを用いる手法である．この手法は，即時性と拡散性に優れており，データが入手し易いという特徴があり，ソーシャルセンシング技術を適用した研究に活用されている．マイクロブログを対象とした既存研究では，地震や台風といった災害を検知する研究[11]-[13]やスポーツイベントを検知する研究[14]，[15]，鉄道や交通渋滞などの遅延や事故に関する情報を抽出する研究[16]-[18]，映画の興行収入を予測する研究[19]，経済動向を分析する研究[20]-[22]，インフルエンザの流行を予測する研究[23]，[24]が実施されている．これらのソーシャルセンシングに関する研究では，主に特定のキーワードの出現数や文脈を解析して，社会事象を検知する手法が利用されている．しかし，各社会事象に特定のキーワードを事前に指定する必要があるため，網羅的な分析が困難であることや，キーワード選定に解析者のバイアスがかかり分析に偏りがみられること等の課題が発生する．

属性推定に関する既存研究[25]-[31]では，ユーザの投稿履歴から暗黙的に含まれる生活習慣を抽出することでユーザの職業属性を推定する研究[25]，ユーザが閲覧した新聞記事とそのタイトルの内容からユーザの性別，年齢や職業属性を推定する研究[26]，SNS 上の顔画像から性別を推定する研究[27], 半教師ありトピックモデルを用いた地域を推定する研究[28], 各属性の特徴となる単語の出現頻度に基づいた地域推定に関する研究[29]-[31]など，性別や年代，職業，地域の属性推定に重きを置かれている．しかし性別や年代，職業においては，一定の精度で推定できることがわかっているが，地域に関しては，投稿者の位置情報の付与率が低い点から推定が困難であることもわかっている．

そこで，本研究では，既存研究の課題に対応するため，ユーザのユーザの投稿履歴から暗黙的に含まれる平時の生活習慣を活用した行動（以下，習慣行動）を用いて，実世界における事象を抽出する新たなソーシャルセンシング手法を提案する．本手法では，マイクロブログの一つである Twitter[32]を対象とし，その投稿から「平時と異なる行動（以下，非習慣行動）を起こすユーザ群を特定することで，その非習慣行動から何らかの大きな社会事象が発生している」という仮説を設定し，このユーザの行動の変化を用いて，社会事象の抽出を試みる．これにより，キーワードの出現数や文脈のみに依存せず，多種多様な事象の抽出が可能になると考える．また，各ユーザ属性の社会事象自動抽出技術の実現のための基盤研究として，地域属性を高精度に推定する手法を提案する．まず，同一の事物を表現する語句や各地域によって使用頻度が異なる語句，具体的には「アホ・バカ」などの類語に着目し，各地域性の違いから地域属性を推定する手法を提案する．類語の特徴として，投稿者が無意識に使用する傾向があり，投稿記事内での出現率が高いため，それらの類語をベースとして，そこから派生する関連語や関連文章から逐次学習することで自動拡張したモデルを使用して地域推定を試みる．そして，ソーシャルセンシング技術と属性推定手法の成果を融合することで，ユーザの性別，年代や職業，そして地域のプロパティ情報を獲得した上で，各ユーザの属性の社会事象，平時と非習慣時のソーシャルセンシングを目指すものである．これにより，サイバー空間とフィジカル空間を融合させた社会的ニーズの高い情報を抽出することが可能となる．

1.2 本論文の構成

本論文の構成は，以下のとおりである．

第 2 章では，CGM から実世界における社会事象の収集手法の着眼点と構想を論じる．第 3 章では，実際の投稿記事から社会事象の抽出を行い，提案手法の有効性を検証する．第 4 章では，さらに，ユーザ属性を考慮した社会事象の分析を試み，その有用性を実証する．第 5 章では，半自動で付与したユーザ属性を用いた結果と自動的に獲得したユーザ属性を用いた結果とを比較し，同様の社会事象を抽出できるかを明らかにする．最終的にソーシャルセンシング技術の確立とその汎用性について議論する．第 6 章では，本研究の総括と今後の展開について述べる．

第 2 章

本研究の提案技術及び研究計画

第2章　本研究の提案技術及び研究計画

2.1　研究計画

本研究では，ユーザ属性を考慮し，自動的に社会事象を抽出するための手法を提案することを目的とする．従来研究と本研究の位置付けを図 2.1 に示す.

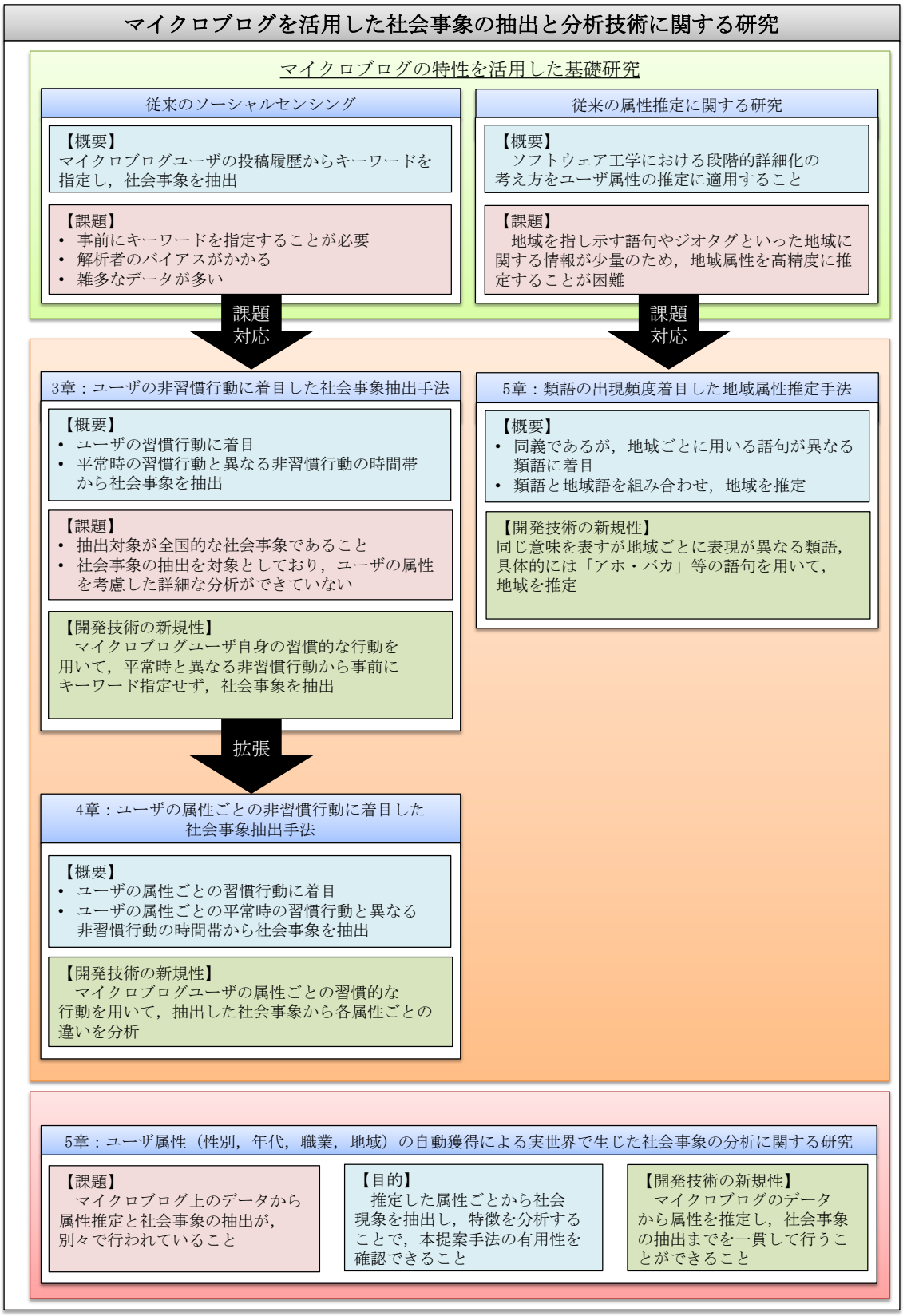


図 2.1 本研究の位置づけ

図 2.1 に示すように従来のソーシャルセンシング技術は，主にマイクロブログユーザの投稿履歴や検索履歴からキーワードを指定して，社会事象を取得していた．しかし，事前にキーワードを指定する必要があり，解析者のバイアスがかかることや分野横断的な分析が困難であった．

第 3 章では，従来のソーシャルセンシングに関する研究の課題に対応するため，マイクロブログユーザの投稿履歴に暗黙的に含まれる投稿パターンと習慣行動に着目し，平時の習慣行動と非習慣行動を比較して，その違いから社会事象を抽出する手法を提案する．本項目の研究成果は，2018 年の情報処理学会論文誌[34]に掲載された．

第 4 章では，前述の研究[34]の課題であった抽出可能な社会事象が全国的な事象であること，社会事象の抽出を対象としていたため，詳細な分析ができていない課題に対応する．そのため，社会事象を抽出する手法を拡張し，ユーザ属性を考慮した習慣行動に着目した社会事象を抽出する手法を提案する．これにより，各ユーザ属性によって同様の社会事象においても反応する時間帯，抽出内容が異なることを示して汎用性を明らかにし，日々増加するビッグデータに対するソーシャルセンシング技術の確立に寄与する．本項目の研究成果は，2020 年日本知能情報ファジイ学会の知能と情報[35]に掲載された．

第 5 章では，地域属性の推定手法に関して，高精度に推定することが困難であった課題に対応する．属性推定に関する研究では，マイクロブログ上の明示的な情報から推定を行われていた．しかし，同じ属性でも多様な習慣行動が存在し，明示的な情報のみではユーザの属性を推定できない課題がある．その課題に対応するため，ソフトウェア工学における段階的詳細化の考え方を適用し，ユーザの属性推定技術[33]を提案した．しかし，ユーザの年代や職業属性は約 5 割から 8 割程度で推定できるが，一部の学生やパート・アルバイト，地域といった性別によって顕著な特徴がみられない属性の推定が難しいことがわかった．また，段階的詳細化手法の特性上，学習データを各属性に分類するとデータ件数が少なくなり，的確な推定モデルを構築できないことが明らかとなった．さらに，地域属性は，地域を示す語句やジオタグといった地域に関する情報が少量のため，高精度に推定することが困難であった. 本項目で提案する研究では, 地域に関する特徴となる語句(以下, 地域語)だけでなく，同じ意味を表すが地域ごとに表現が異なる類語を加増する．そして，Twitter から各地域のユーザを自動的に収集し，そのユーザ群から算出した地域語と類語を組み合わせた語句（以下，地域辞書）を用いて，逐次学習することで自動拡張して学習モデルを更新する新たな手法を提案する．本技術の実現により，ユーザの各属性の社会事象，平時と非習慣時のソーシャルセンシングの高度化に寄与する．

以上のように本研究は，3 つの提案技術をもって，ユーザ属性の自動獲得による実世界で生じた社会事象の分析を可能とする．

2.2　提案技術の概要

マイクロブログを活用した社会事象の自動抽出と分析技術に関する研究に必要な 3 つの技術に関する研究の関係性を表した図を図 2.2 に示す.

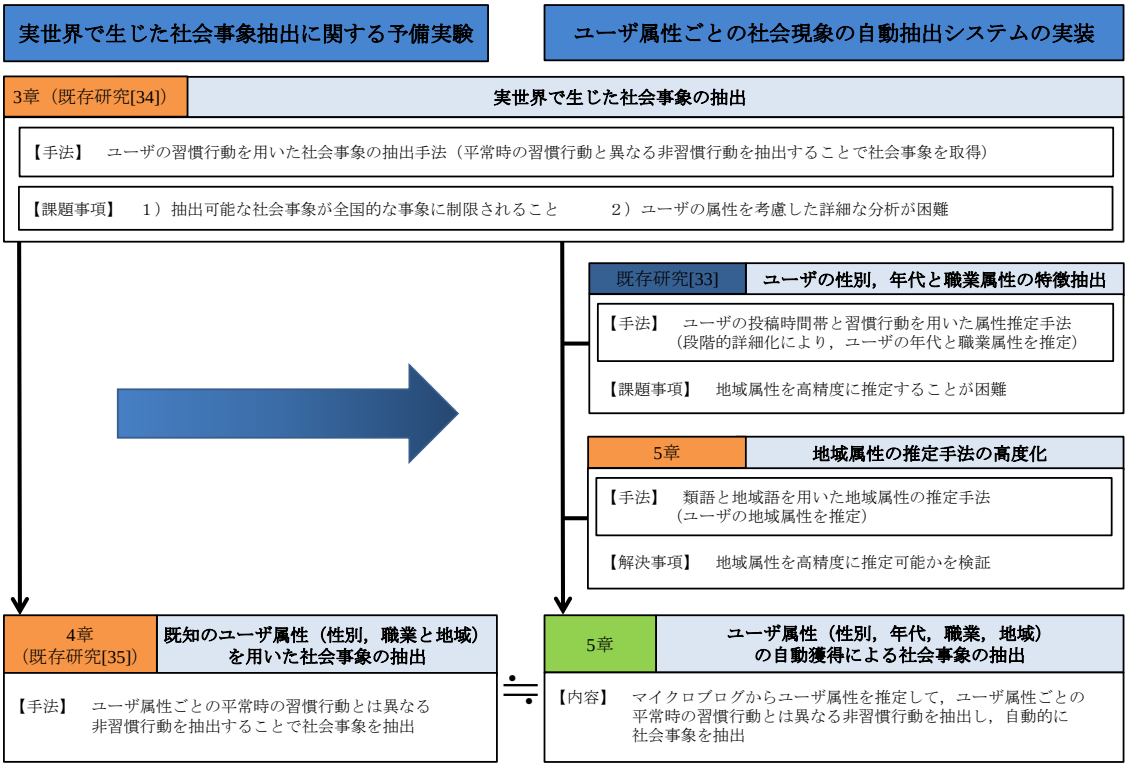


図 2.2　本提案技術の関係性

本研究を遂行するにあたり，実世界における各ユーザ属性の社会事象の自動抽出を目指し，マイクロブログユーザの属性を推定する技術[33]と社会事象を抽出する技術[34]，[35]を提案してきた.

属性推定に関する研究では，各ユーザ属性に推定精度が異なることに着目し，ソフトウェア工学における段階的詳細化に関する考え方を適用した．そして，ユーザの性別，年代と職業といった属性を段階的に推定する手法[33]を提案し，属性推定や習慣行動の予測に対する課題に有用であることを確認した．しかし，この手法では，ユーザの年代や職業属性は約 5 割から 8 割程度で推定できるが，一部の学生やパート・アルバイト，地域といった性別によって顕著な特徴がみられない属性の推定が難しいことがわかった．また，段階的詳細化手法の特性上，学習データを各属性に分類するとデータ件数が少なくなり，的確な推定モデルを構築できないことが明らかとなった.

社会事象を抽出する技術[34]では，ユーザの習慣行動に着目し，非習慣行動を抽出することで社会事象を取得する手法を提案した．そして，基準となる習慣行動と各月の習慣行動

を比較し，取得した非習慣行動を分析することで，従来の事前にキーワードを指定した手法よりも詳細な社会事象が抽出できることを証明した．また，ソーシャルセンシングによる詳細な社会事象の抽出には，性別や職業，地域といった各ユーザ属性の習慣行動と非習慣行動の考慮が必須であることを示唆した．そこで，既存手法[34]を実践的に評価するため，ユーザ属性を考慮して非習慣行動を抽出し，社会事象の取得[35]を試みた．その結果，ユーザ属性を考慮した手法[35]は，既存手法[34]よりも詳細なトピックを抽出できることを明らかにし，ユーザ属性の推定技術が必要であるとの結論を得た．

これらの研究を経て，著者らが行ってきたマイクロブログユーザの属性を推定する技術[33]が必要不可欠であることがわかり，ユーザ属性の自動獲得による社会事象の抽出に大きく寄与することができた．こうした研究経緯を踏まえて，各ユーザ属性の社会事象の自動抽出技術の実現に向けた地域属性の高精度な推定手法を提案する．本技術の実現により，各ユーザの属性の社会事象，平時と非習慣時のソーシャルセンシングの高度化に寄与する．

本研究では，マイクロブログを活用したユーザ属性の自動獲得による社会事象検出を行う．本研究におけるソーシャルセンシング技術の流れを図 2.3 に示す．

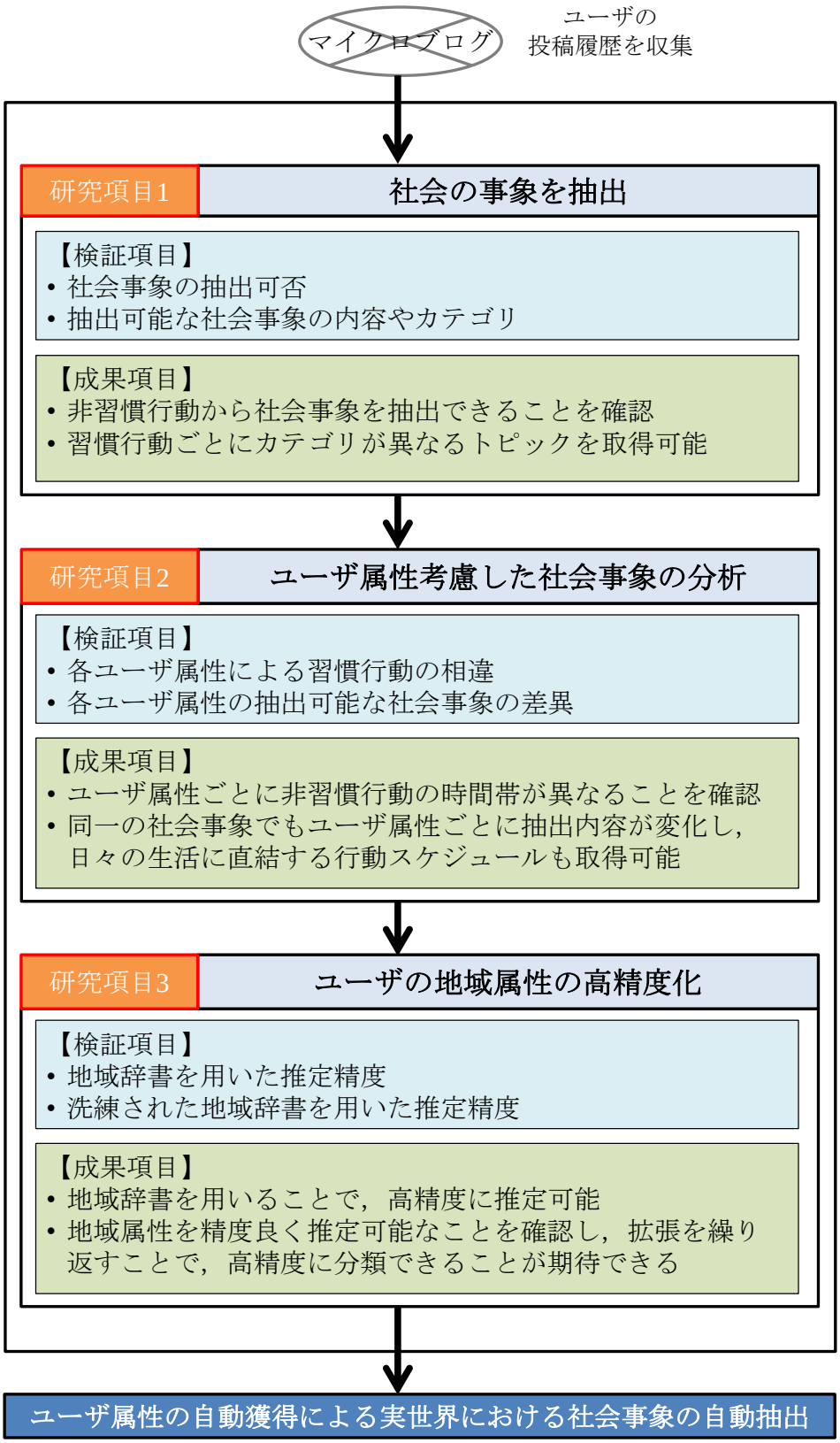


図 2.3　本研究におけるソーシャルセンシング技術の流れ

図 2.3 に示す通り，本研究は 3 つの成果を実現して，ユーザ属性の自動獲得による実世界における社会事象の自動抽出を目指すものである．詳細な技術を以下に示す．

2.2.1 非習慣行動を用いた社会事象の抽出技術

Twitter を対象として，事前にキーワードを指定せずにユーザの習慣行動を使用し，実世界における事象を抽出する新たなソーシャルセンシング手法を提案する．本手法では，習慣行動を分析するために投稿履歴を解析し，各投稿からユーザの行動情報の抽出と，生活習慣に関する単語の各時間の出現回数を示すベクトルを作成する．そして，1 年間の行動を平時の行動とし，各月の習慣行動と比較する．評価実験では，「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，本提案手法のソーシャルセンシング技術が有用であることを検証する．

2.2.2 ユーザ属性を考慮した社会事象の分析技術

ユーザ属性を考慮し，基準となる習慣行動と各属性の各月の習慣行動を比較することで，各属性の非習慣行動を分析する．これにより，ユーザ属性を考慮せずに抽出した際，注目度の大きいイベントや災害などの社会事象に関する話題に集中する課題に対応する．本手法では，ユーザによって投稿履歴に差が生じた場合においても各ユーザ属性の投稿履歴を複数収集して，習慣行動と非習慣行動を算出することにより，各ユーザ属性の傾向を考慮した情報を抽出することができる．

2.2.3 類語に着目したユーザの地域属性の推定技術

Twitter を対象とし，同一の事物を表現する語句や各地域に使用頻度が異なる語句，具体的には地域語とは異なる「アホ・バカ」などの類語に着目し，各地域性の違いから地域属性を推定する．そして，各地域のユーザを自動的に収集し，そのユーザ群から算出した地地域辞書を用いて，逐次学習することで自動拡張した学習モデルを更新する新たな手法を提案する．本研究で用いる類語の定義は，同一の事物を表現する語句や各地域で使用頻度が異なる語句である．また，各地域における特徴を示す地域語と類語を組み合わせることで，ユーザの地域推定の精度向上を試みる．

以上の手法を提案し，マイクロブログ上に日々蓄積されているビッグデータを対象とし，平時と非習慣時のソーシャルセンシング手法とユーザの地域属性に関する推定手法の実現を目的とする．

2.2.4　使用する属性一覧

本研究での章ごとでの使用するユーザ属性を表 2.1 に示す.

表 2.1　章ごとのユーザ属性の関係性

既存研究と章	性別	職業	年代	地域
既存研究[33]	自動	自動	自動	×
3 章	×	×	×	×
4 章	手動	手動	手動	手動
5 章	自動	自動	自動	自動

表 2.1 において，自動は，推定したユーザ属性を用いていることを表し，×は，対象のユーザ属性を考慮していないことを表し，手動は，マニュアルで取得したユーザ属性を表す.

本研究では，上記の表に沿って研究を遂行する.

第 3 章

平時と異なる事象に対する
ソーシャルセンシング技術の提案

第3章　平時と異なる事象に対する ソーシャルセンシング技術の提案

3.1　まえがき

本章では，実世界で生じた社会事象を抽出することを目的とし，マイクロブログユーザの習慣行動を用いて，非習慣行動を取得する手法について検討する．そして，既存研究の課題である「各社会事象に合った特定のキーワードを事前に指定する必要があるため分野横断的な分析が困難である点」と「キーワード選定に解析者のバイアスがかかり分析に偏りがみられる点」に対応する．本研究では，ユーザの投稿履歴に暗黙的に存在する習慣行動に着目し，これらの行動の変化を解析することで，非習慣行動から何らかの大きな社会事象を把握することが可能であると仮説を設定し，新たなソーシャルセンシング手法を開発する．これにより，キーワードに依存せずに，社会事象を抽出することが可能となる．

第 3.2 節では，研究の概要について論じている．第 3.3 節では，提案技術の有用性を検証するための検証項目に関して論じている．第 3.4 節では，習慣行動の相違を用いたソーシャルセンシング技術に関してのアルゴリズムについて論じている．第 3.5 節では，2 つの検証実験に関して論じている．

3.2　研究の概要

3.2.1　研究の位置付け

本研究では，既存研究の課題に対応するため，ユーザ一人ひとりに着目して，投稿傾向や行動の変化といったユーザ特性を用いて実世界における事象を検出する新たなソーシャルセンシング手法を提案する．本手法では，「平時と異なる行動は何らかの大きな社会事象が発生している状況を示している可能性がある」という仮説を設定し，このユーザ行動の変化を用いて，社会事象を抽出する手法を検討する．これにより，キーワードの出現数や文脈のみに依存せず，多種多様な事象の抽出が可能になると考える．本研究の位置付けを図 3.1 に示す．

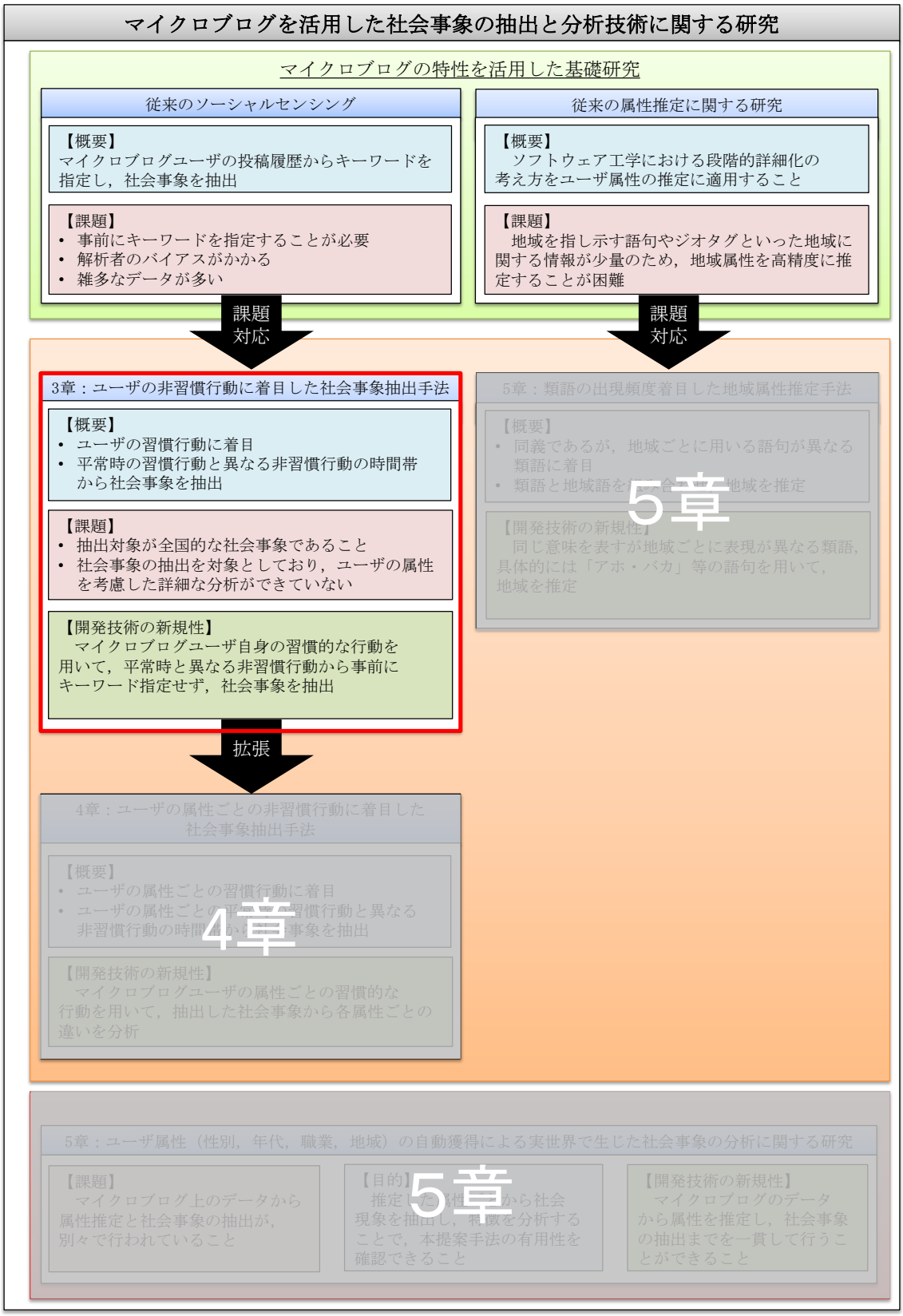


図 3.1　本研究の位置付け

3.3　検証項目の設定

本研究では，次に示す 2 つの検証項目を設定し，これらを明らかにすることで，提案のソーシャルセンサの有用性を確認する．

3.3.1　検証項目1：社会事象の抽出可否

現状のソーシャルセンシングに関する研究では，主に特定のキーワードの出現数や文脈を解析して，事象を検知する手法が利用されている．しかし，これらの手法は，各事象に合った特定のキーワードを事前に指定するため，多種多様な事象を広範囲に把捉することが困難である．既存研究においても，地震やスポーツイベント，交通ネットワークなどの特定の事象を対象に抽出しているものが多い．

本研究では，検証項目として，「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」を設定して，複数のユーザの習慣行動から非習慣行動を抽出することにより，実世界における事象を抽出する手法を提案する．これにより，キーワードの出現数や文脈のみに頼らないソーシャルセンシング手法を開発し，平時と異なる行動を起こすユーザ群を特定してその投稿を解析することで，社会事象が抽出可能かを検証する．

3.3.2　検証項目2：抽出可能な社会事象の内容やカテゴリ

提案手法では，ユーザ生活習慣の変化を用いてキーワードの出現数や，文脈では抽出できない事象の抽出ができると考えられる．

本研究では，検証項目として「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」を設定して，ユーザの習慣行動に着目し，行動の変化を抽出する．そして，平時と異なる行動をした期間の投稿を収集してトピック単位に分類する．その後，実世界で発生した特定の事象を抽出し，平時の事象と比較する．これにより，ソーシャルセンサの特性としてユーザの生活習慣を考慮することで，抽出可能な社会事象の内容やカテゴリが変化するかを検証する．

3.4　習慣行動の相違を用いたソーシャルセンシング技術

ソーシャルセンシングとは，マイクロブログなどのソーシャルメディア上での利用者をソーシャルセンサとして捉え，実世界の事象を観測する方法である．ソーシャルセンサは，抽出可能な事象の範囲が広く，解析データの取得が容易といった特徴がある．本研究では，その中でもリアルタイム性に優れており，投稿される情報量が多い Twitter を採用する．

ソーシャルセンシングに関する既存研究では，社会事象が起こった際の文章やデータを利用しており，社会事象に着目している．本研究では，社会事象時に起きる投稿傾向や行動の変化に着目し，社会事象ではなくユーザに着目して解析する．提案手法の処理として，本研究では投稿傾向を用いた手法と生活習慣を用いた手法の 2 つの手法を検討する．

3.4.1　投稿傾向を考慮する手法

本手法では，社会事象が起きた時の前後のユーザの投稿傾向を確認し，投稿傾向の変化から社会事象を抽出可能であるかを検証する．本手法の処理フローを図 3.2 に示す．

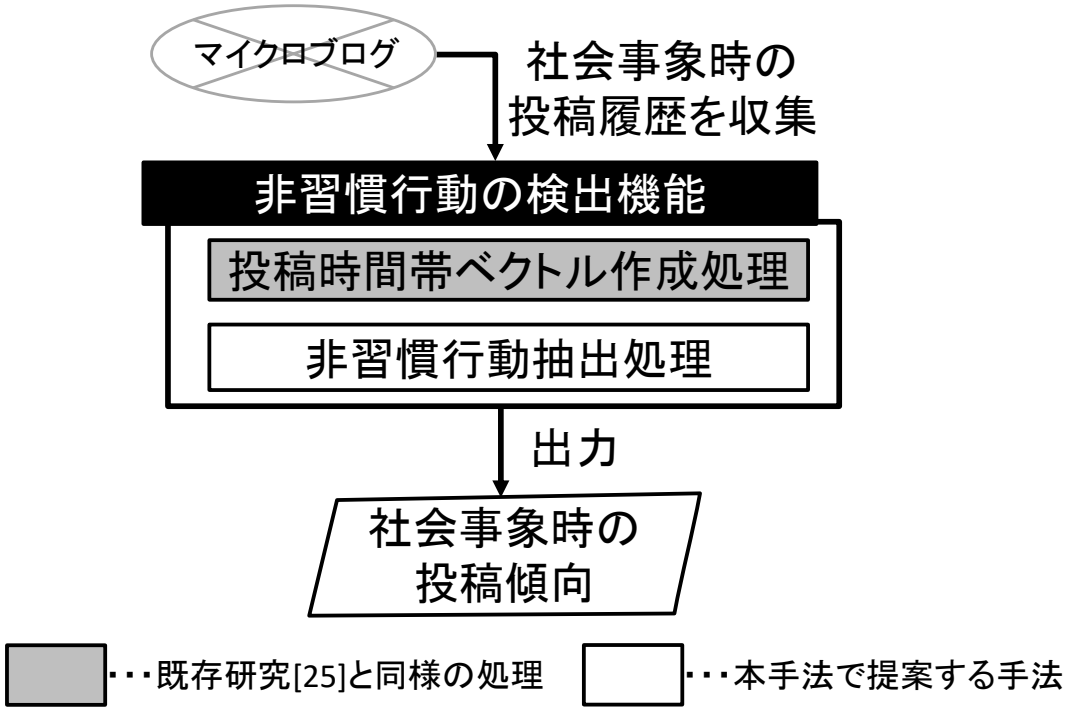


図 3.2　投稿傾向を考慮した処理フロー

図 3.2 に示す通り，本処理フローは，非習慣行動の検出機能で構成される．入力は，社会事象時の投稿履歴であり，出力は，社会事象時の投稿傾向である．

(1)　非習慣行動の検出機能

本機能では, 既存研究[25]を参考に投稿時間帯ベクトルを作成し, そのベクトルを用いて, 非習慣行動の抽出を行う．具体的な処理内容を次に示す．

a)投稿時間帯ベクトル作成処理

本処理では，ユーザの投稿パターンの把握を目的とした処理である．内容として，各曜日の時間帯における投稿数を素性とした投稿時間帯ベクトルを作成する．この投稿時間帯ベクトルは，7 次元（曜日）×24 次元（時間帯）の 168 次元で構成されている．この次元の理由としては，一週間の投稿傾向を把握したかった点，社会事象時の投稿量とその前後の投稿量を比較したいためである．各属性のユーザ *user* における投稿時間帯ベクトル $V_{posttime}(user)$を式 3.1 に示す．

$$\boldsymbol{V_{posttime}(user) = \{Post_{Sun0}(user), Post_{Sun1}(user), \cdots, Post_{Sat23}(user)\}} \qquad \textbf{式 3.1}$$

式 3.1 において，$\boldsymbol{Post_{Sun0}}$は日曜日 0 時台の投稿数を示す．式 3.1 を用いて，投稿時間帯ベクトルを各ユーザで作成後，各曜日の投稿傾向を把握するため，全ユーザの投稿を加算する．作成した時間帯ベクトルの例を図 3.3 に示す．

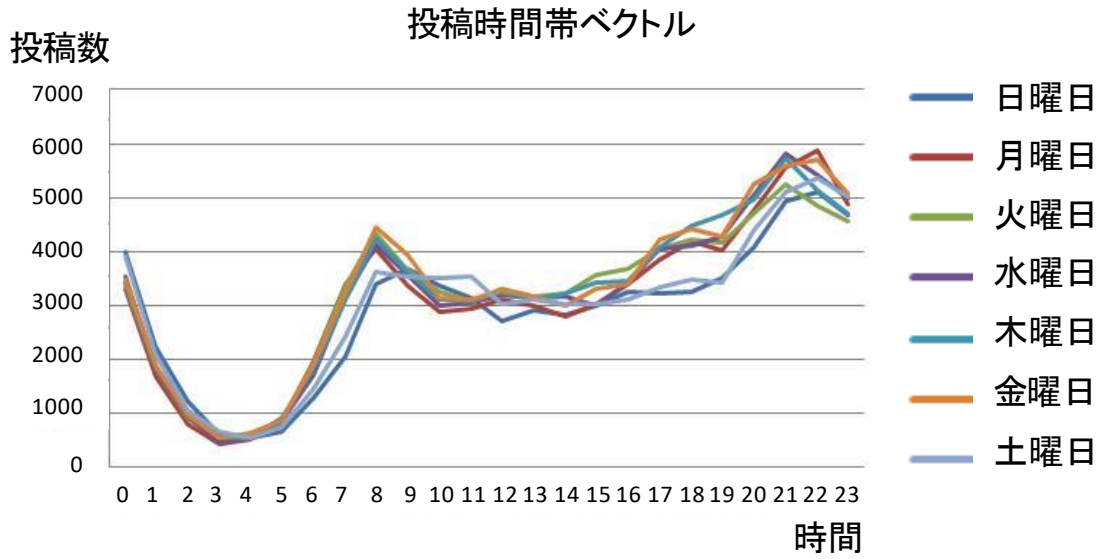


図 3.3 投稿時間帯ベクトル

b)非習慣行動抽出処理

本処理では，指定した社会事象の期間の投稿時間帯ベクトルから他の時間帯と逸脱する投稿傾向を示す時間帯を非習慣行動として，抽出する．処理手順を次に示す．

STEP1： 社会事象を選定する．

STEP2： STEP1 で選定した社会事象の前後の投稿時間帯ベクトルを作成する．

STEP3： STEP2 で作成した投稿時間帯ベクトルで，他の時間帯と逸脱する投稿量を確認し，

抜き出す.

STEP4： 抽出した非習慣行動時の投稿を確認する

以上の処理において，社会事象時の投稿傾向から非習慣行動を抽出する.

3.4.2　生活習慣を考慮する手法

本手法では，平時の生活習慣と特定期間の生活習慣を比較し，平時と逸脱する生活習慣から社会事象を抽出できるかの検証する．本手法の処理フローを図 3.4 に示す.

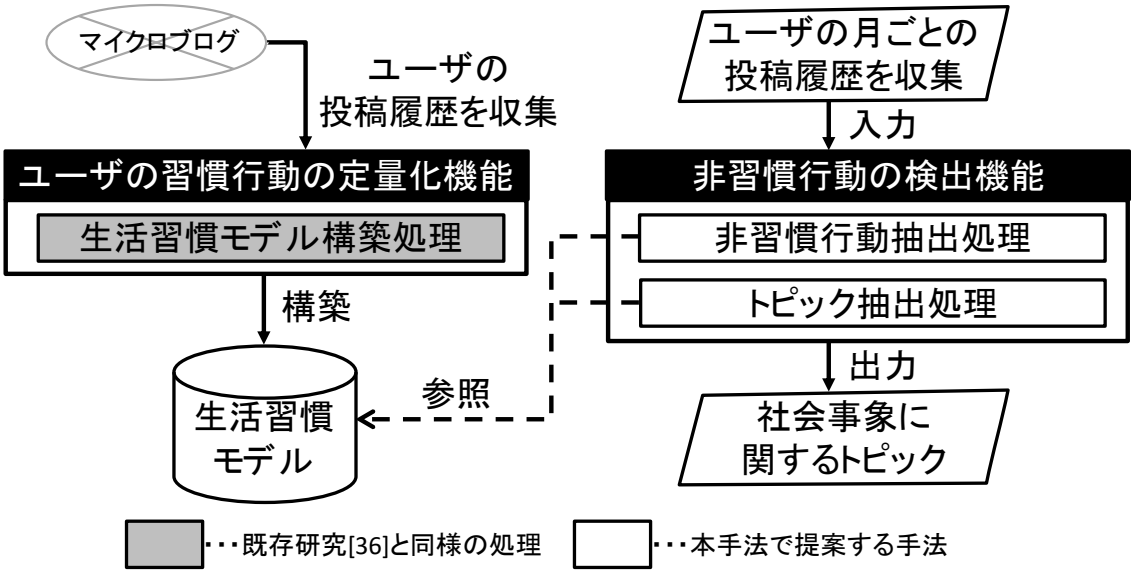


図 3.4　生活習慣を考慮した処理フロー

図 3.4 に示す通り，本処理フローは，ユーザの習慣行動の定量化機能と非習慣行動の検出機能で構成される.

(1)　ユーザの習慣行動の定量化機能

本機能では，既存研究[36]を参考に生活習慣ベクトルを作成し，非習慣行動の抽出と社会事象の分析に用いるための生活習慣モデルを構築する．具体的な処理内容を次に示す.

a)生活習慣モデル構築処理

生活習慣モデル構築処理では，ユーザの投稿履歴から，あらかじめ設定した「起床・就寝」，「在宅」，「外出」と「帰宅」の習慣行動に関する単語の各時間の出現回数を示すベクトルを作成する．行動辞書に登録した用語の例を表 3.1 に示す．

表 3.1　行動辞書に登録した用語の例

行動	用語
起床・就寝	おはよう，寝る，就寝，眠り，おやすみ
在宅	風呂，テレビ，買い物，旅行
外出	出勤，通勤，通学，行ってきます
帰宅	帰宅，帰る，退勤，退社，下校

なお，行動辞書の作成は，日本語語彙体系[37]を参考に作成した．なお，各地域の生活習慣の違いの把握のため，行動辞書の一部に方言も含ませている．

次に，生活習慣の習慣行動の処理手法を図 3.5 に示す．

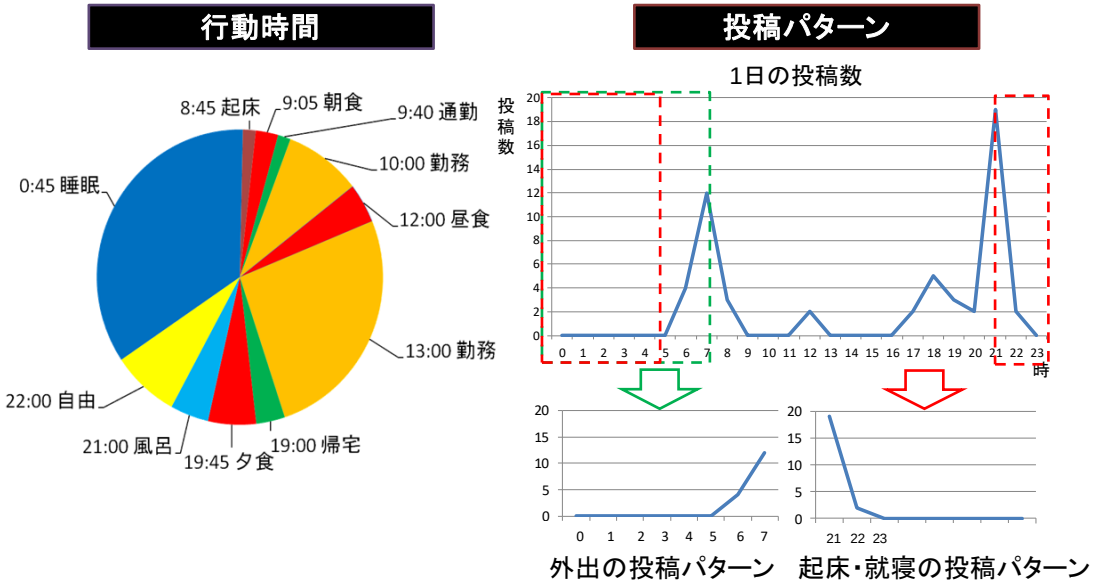


図 3.5　習慣行動の処理手法

図 3.5 において，行動時間は，8 時 45 分起床，9 時 40 分通勤，12 時昼食，19 時帰宅，0 時 45 分睡眠といった何時にどのような行動を行っているかを解析し，時間帯から行動を分析する．

投稿パターンは, 5 時から 7 時にかけて投稿数が増加し始めている投稿パターンにおいて，外出に関する投稿パターンと解釈できる．同様に，21 時から 23 時にかけて投稿量が減少している投稿パターンにおいては，就寝の投稿パターンと解釈できる．このように，ユーザの投稿履歴から習慣行動を解析し，その習慣に基づき任意の時間帯の行動を分析する．この手法を基に生活習慣ベクトルを作成する．

生活習慣ベクトルはユーザの習慣行動を素性としたベクトルであり，4 次元（習慣行動）×7 次元（曜日）×24 次元（時間帯）の 672 次元で構成する．本研究では，平時のユーザの習慣行動を表す生活習慣ベクトル（以下，平時習慣ベクトル），特定の期間の習慣行動を表す生活習慣ベクトル（以下，特定習慣ベクトル）を定義する．本章での平時習慣ベクトルは，年間の投稿から作成した生活習慣ベクトル，特定習慣ベクトルは，各月の投稿から作成した生活習慣ベクトルを表す．年間の行動beh_xにおける各曜日の時間帯hの平時習慣ベクトルと各月の行動beh_xにおける各曜日の時間帯hの特定習慣ベクトルの式を式 3.2 と式 3.3 に示す．

$$\boldsymbol{YearPost(beh_x) = \{YPost_{beh_1,0}, YPost_{beh_1,1}, \cdots, YPost_{beh_x,h}\}} \quad \textbf{式 3.2}$$

$$\boldsymbol{MonthPost(beh_x) = \{MPost_{1,beh_1,0}, MPost_{1,beh_1,1}, \cdots, MPost_{m,beh_x,h}\}} \quad \textbf{式 3.3}$$

式 3.2 において，hは，7 次元（曜日）×24 次元（時間帯）の 1 時間を表す．hが 0 の場合，年間の日曜日の 0 時を示す．$YPost_{beh_x,h}$において$h = 0$の場合，年間の日曜日の 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間に生活習慣beh_xに関連する単語を含む投稿がなされた回数を表す．$h = 167$の場合，土曜日の 23 時 00 分 00 秒から 23 時 59 分 59 秒までを示している．

式 3.3 において，mは月を表し，$MPost_{1,beh_x,0}$の場合，1 月の日曜日の 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間に生活習慣beh_xに関連する単語を含む投稿がなされた回数を表す．

式 3.4 により，生活習慣ベクトルを作成後，各曜日で合計を算出し，正規化を行った．

$$\boldsymbol{YearPost(beh_x)' = \frac{YearPost(beh_x) - MinYearPost(beh_x)}{MaxYearPost(beh_x) - MinYearPost(beh_x)}} \quad \textbf{式 3.4}$$

式 3.4 において，$MaxYearPost(beh_x)$は，日曜日から土曜日までの最大値の投稿数，$MinYearPost(beh_x)$は，日曜日から土曜日までの最低値の投稿数を示す．同様に月ごとでも

生活習慣ベクトルを正規化する．本研究では，この二つのベクトルを生活習慣モデルとする．

(2)　非習慣行動の検出機能

本機能では，平時習慣ベクトルと特定習慣ベクトルとを比較し，各時間帯の差分を抽出する．これにより，平時と異なる行動を起こすユーザが多い曜日・時間帯を明らかにする．そして，その時間帯に発生している非習慣行動からトピック（話題）を抽出し，そのトピックから実世界で起きた社会事象を解釈する．

a)非習慣行動抽出処理

本処理では，平時習慣ベクトルと特定習慣ベクトルから 1 時間ごとの面積で比較し，各時間帯の面積の差分を抽出する．平時習慣ベクトルと特定習慣ベクトルにおける面積の差分 S は，式 3.5 にて算出する．

$$\boldsymbol{S(beh_x, h) = \left| \int_{\alpha-\beta}^{\alpha+\beta} \{YearPost(beh_x, h)\} dh - \int_{\alpha-\beta}^{\alpha+\beta} \{MonthPost(beh_x, h)\} dh \right|} \quad \text{式 3.5}$$

式 3.5 において，αは求める 1 時間の生活習慣ベクトル，βは求める αの前後の生活習慣ベクトルを表す．本研究では，βの値を 3 時間と設定した．そして，式 3.6 において，前後β時間と求めるα時間の計$2\beta + \alpha$時間の平均を算出する．これは，1 日の生活リズムは 0 時～6 時, 6 時～12 時, 12 時～18 時と 18 時～24 時の 6 時間ごとに変化すると考えたためである．また，前後の時間帯を考慮する理由としては，短時間の異常投稿による影響を少なくするためである．これにより，特定の期間に発生した事象を把握できる．なお，差分がマイナス値にならないよう絶対値で計算を行う．α時間ごとの面積の算出を式 3.7 に示す．

$$\boldsymbol{S'(beh_x, h) = \frac{S(beh_x, h)}{2\beta + \alpha}} \quad \text{式 3.6}$$

$$\boldsymbol{S'(beh_x, h) \geqq \frac{\sum_{k=0}^{167} S'(beh_x, k)}{168}} \quad \text{式 3.7}$$

式 3.6 において平時習慣ベクトルと特定習慣ベクトルの差分を算出し，前後それぞれβ時間の計$2\beta + \alpha$時間から算出したα時間ごとの面積から閾値を超える値を非習慣行動として定義する．本研究は，閾値として，α時間ごとの面積が式 3.7 を満たす場合を非習慣行動とした．

この時，全体の 1 時間あたりの差分が閾値，α時間ごとの差分が，$S'(beh_x, h)$を示す．αとβの関係性を図 3.6 に示す．

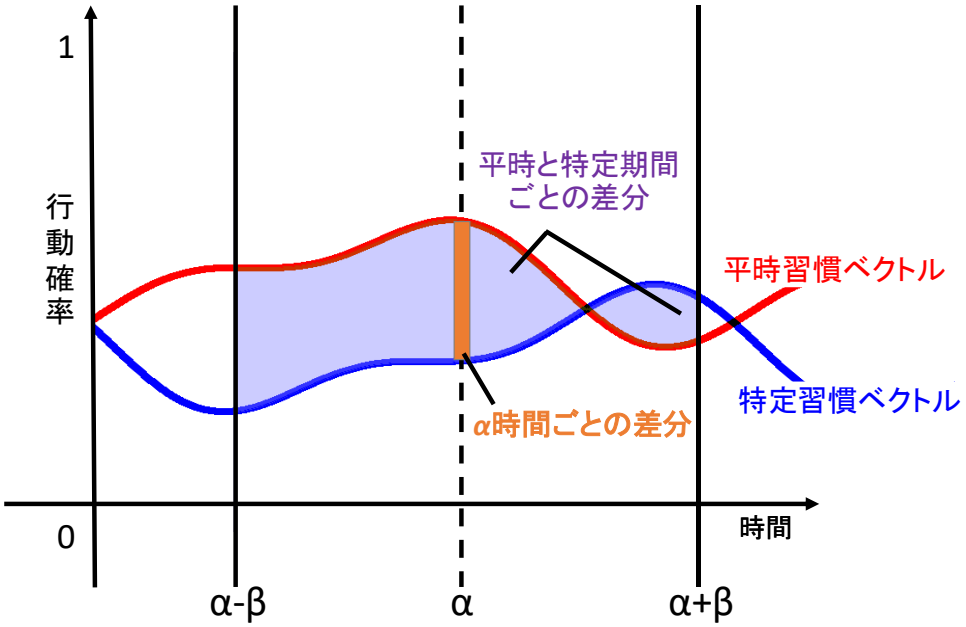


図 3.6　αとβの関係性

b) トピック抽出処理

本処理では，Biei らによって提案された潜在的ディリクレ配分法(LDA：Latent Dirichlet Allocation)[38]を用いて，トピック（話題）を抽出する．LDA とは，文書に出現する単語に存在するトピックの関係を確率的に表したトピックモデルの一つである．抽出した特定の期間と月全体のトピックを比較し，社会事象を抽出する．本研究では，Python のトピックモデルライブラリである gensim[39]を用いて，トピックを抽出する．トピックの生成課程を次に示す．

STEP1： 抽出した各行動の期間と月全体の投稿に対し MeCab を用いて, 形態素解析を行う．

STEP2： STEP 1 で求めた形態素から名詞のみを用いて，gensim を用いて特徴語の辞書を作成する．そこでは記号やアルファベット一文字などは，StopWord を定義し，除外する．

STEP3： STEP2 で作成した辞書を元にトピックを抽出する．

3.5 検証実験

3.5.1 実験概要

本実験では，「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，投稿傾向を考慮する手法と生活習慣を考慮する手法の 2 つの手法を検討し，本提案のソーシャルセンシング技術が有用であることを検証する．

本研究では，次に示す実験を行う．本実験の位置づけを図 3.7 に示す．

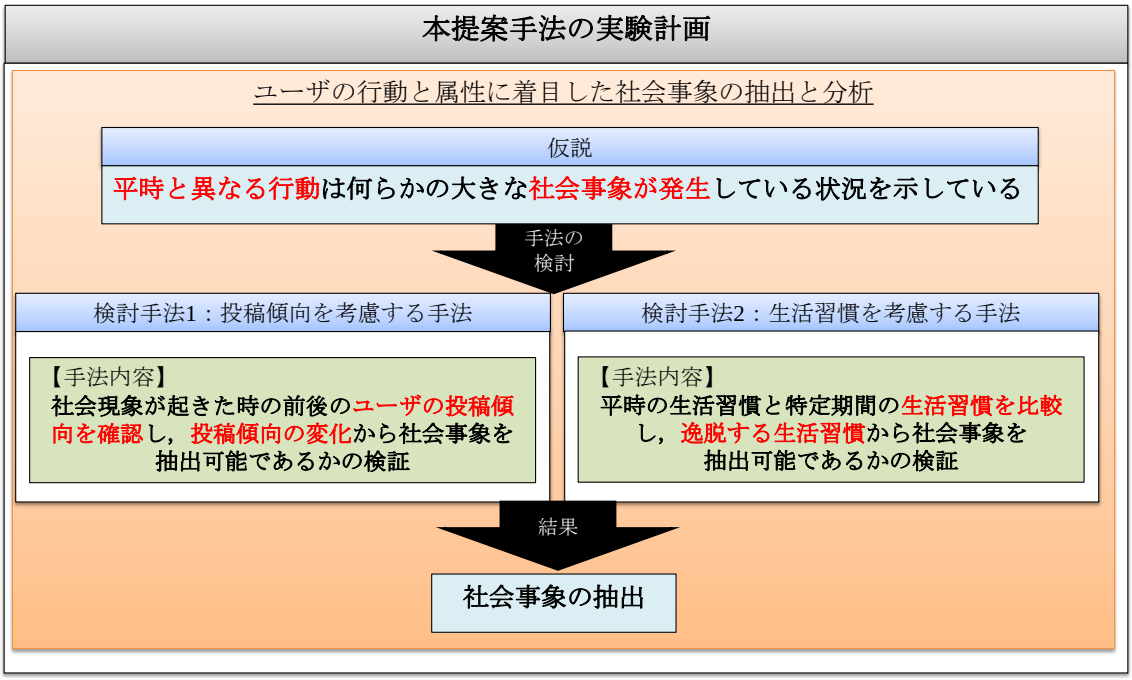


図 3.7 実験の位置づけ

実験 1 では，社会事象が起きた時の前後のユーザの投稿傾向を確認し，投稿傾向の変化から社会事象を抽出可能であるかを検証する．本実験では，実世界で起きた社会事象に対して，その前後の投稿傾向を比較することによって，社会事象時と前後の日時の投稿傾向を把握する．それにより，社会事象時の投稿傾向を分析できる．

実験 2 では，平時の生活習慣と特定期間の生活習慣を比較し，逸脱する生活習慣から社会事象を抽出できるかを検証する．本実験では，年間の全投稿と各月の全投稿から生活習慣ベクトルを作成し，その差分から逸脱する曜日と時間帯を抜き出す．その逸脱する行動を非習慣行動として，各行動のトピックを抽出することで，社会事象を取り出す．

(1)　各月の非習慣行動時間の抽出実験

本実験では，年間の全投稿と各月の全投稿から作成した生活習慣ベクトルを用いて，各習慣行動の逸脱する行動を取り出し，それを非習慣行動として抽出する．この実験において，検証項目 1 を証明する．

(2)　非習慣行動時間のトピック分析実験

本実験では，抽出した各習慣行動の非習慣行動時間のトピックを抽出し，そのトピックから社会事象を見出す．この実験において，検証項目 2 を証明する．

3.5.2　投稿傾向の変化から社会事象を抽出可能であるかの検証実験

(1)　実験内容

本実験では，「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，本提案のソーシャルセンシング技術が有用であることを検証する．実験 1 では，各習慣行動から非習慣行動をとるユーザ群が多い時間帯を明らかにし，実験 2 では，その時間帯のトピックを抽出することで，社会事象を取得可能であるかを検証する．

(2)　実験データ

本実験では，2011 年～2015 年の投稿群を対象とする．実験データの収集方法を次に示す．

STEP1： Twitter のプロフィールから任意の文字列を検索するツイプロ[40]サービスを用いて，Twitter のプロフィール欄や投稿内容に属性名を記載しているユーザを無作為に収集する．

STEP2： STEP1 で収集したユーザの投稿内容を TwitterAPI[41]と各ユーザの投稿履歴をブログ形式で保存する Twilog[42]サービスを用いて収集する．Twilog から取得したユーザが TwitterAPI で取得したユーザと重複している場合は，Twilog のデータを優先して採用する．理由としては，TwitterAPI では，投稿内容の取得数制限のため，最大で 3,200 件であるのに対して，Twilog はユーザの全投稿を抽出できるため，投稿数が多いからである．

STEP3： STEP1 および STEP2 で収集した投稿内容の件数が 1,000 件以上のユーザを実験データとする．ただし，ライフスタイルの解析に最低限必要な 1 週間分の投稿内容を取得できないユーザは，実験データから除外する．

STEP4： STEP3 のユーザを地域属性に分類する．

STEP5： 地域のユーザに関しては，国土交通省が発表している地価平均[43]を参考に地価が 16 万以上を都市，10 万以上を地方，それ以外を田舎とした．

以上の手順にて収集した実験データの詳細を表 3.2 に示す．

表 3.2　3.5 節で用いる実験データ

	分類	ユーザ数	投稿件数
地域	都市	596 件	1,131,963
	地方	435 件	855,096
	田舎	483 件	98,600

表 3.2 の投稿件数は，TwitterAPI および Twilog の仕様に基づき収集したツイートの件数を示している．ただし，TwitterAPI では，投稿内容の取得制限のため，最大で 3,200 件となった．なお，地価以外の地方区分に関しては，八地方に区分[44]して解析した．

本研究では, 対象の社会事象を決定するため 2011 年から 2015 年までで起きた社会事象を調査した．調査した社会事象を表 3.3 に示す．

表 3.3　選定対象の社会事象

年	社会事象	カテゴリ	年	社会事象	カテゴリ
2011	金正日総書記が死去	政治・経済	2014	はやぶさ 2 打ち上げ	政治・経済
	大阪ダブル選挙	政治・経済		ノルディックスキーのW杯ジャンプ	スポーツ
	サッカー女子ワールドカップ	スポーツ		世界体操	スポーツ
	東日本大震災	災害		仁川アジア大会	スポーツ
	計画停電	災害		世界初の iPS 細胞移植手術	政治・経済
2012	北朝鮮人工衛星を搭載したミサイル打ち上げ	政治・経済		8 月豪雨	災害
	山中教授ノーベル賞	政治・経済		世界柔道	スポーツ
	ロンドンオリンピック	スポーツ		FIFA ワールドカップ	スポーツ
	金環日食	その他		集団的自衛権の行使	政治・経済
2013	特定秘密保護法案の廃案デモ	政治・経済		衆議院解散	政治・経済
	2020 年東京オリンピック決定	政治・経済		ソチオリンピック	スポーツ
	プロ野球東北楽天ゴールデンイーグルス日本一	スポーツ		世界卓球	スポーツ
2014	赤崎教授ら 3 人ノーベル賞	政治・経済	2015	世界水泳	スポーツ
	都知事選挙	政治・経済		世界陸上	スポーツ

表 3.3 示す通り，本研究では地域に着目して東日本大震災，プロ野球東北楽天ゴールデンイーグルス日本一，2014 年の 8 月豪雨を対象にした．

社会事象時の投稿時間帯ベクトルの作成手順を次に示す．

STEP1： 社会事象が発生した日を含む 7 日間の投稿群を入力とし，それぞれ投稿時間帯ベクトルと生活習慣ベクトルを作成する．

STEP2： 実験データのユーザを対象に各属性の投稿時間帯ベクトルを作成する．

STEP3： 属性と曜日・時間帯ごとに習慣行動に関する単語の出現数からなる生活習慣ベクトルを作成し，正規化する．

STEP4： STEP3 で作成した 7 日間分のベクトルをグラフ化し，各社会事象の特徴を確認する．

(3)　実験結果

本実験では，社会事象の中でも表 3.3 に示す東日本大震災，プロ野球東北楽天ゴールデンイーグルス日本一と 8 月豪雨の実験結果を以下に示す．

a) 東日本大震災

東日本大震災は，2011 年 3 月に発生した大地震とその影響により発生した津波による大規模な自然災害である．本事象では，各地域によって被害の規模が異なることから，地域によるセンシングが有効であると考えられる．そのため，図 3.8 に示す事象発生期間における各地域の投稿時間帯ベクトルを確認すると次に示す事項が明らかとなった．赤枠は東日本大震災が起きた時間帯，青枠は静岡県で起きた地震の時間帯を示す．

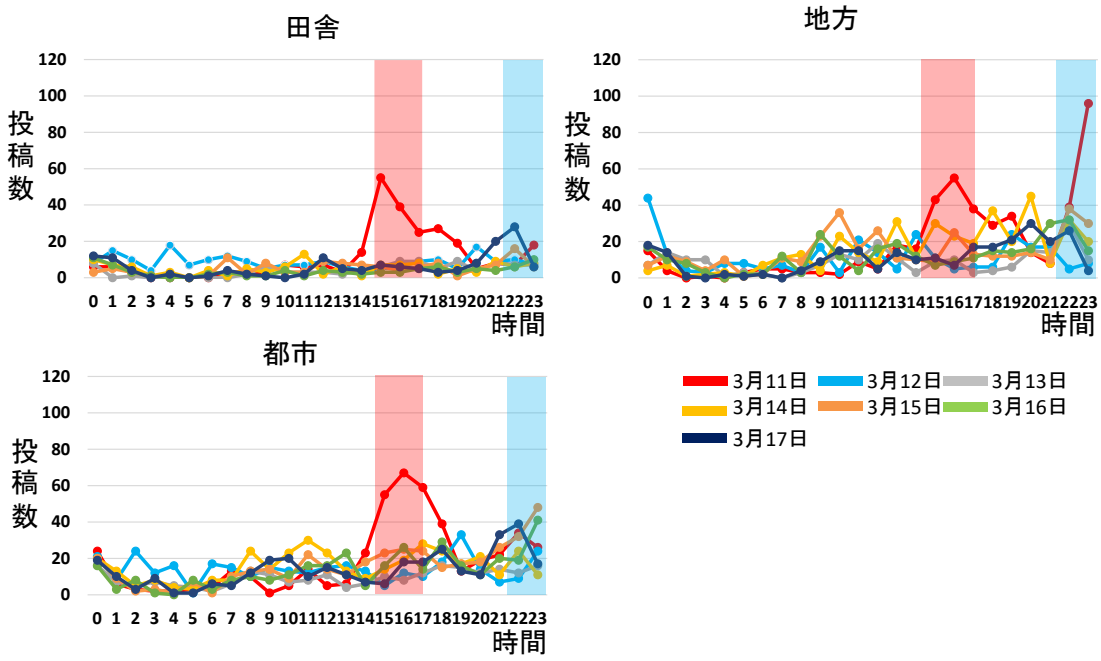


図 3.8　各地域の東日本大震災に関する解析結果

- 全国的な社会事象の発生時に投稿数が大きく変化することがわかる

図 3.8 を確認すると，どの地域も 3 月 11 日 14 時 46 分頃からの投稿数（赤色部分）が急激に増加していることがわかる．また，3 月 15 日には静岡県で震度 6 強の地震が発生（青色部分）しており，その傾向も確認することができる．東日本大震災時に投稿されたツイートを詳細に確認すると，「栃木県南部で外にいたけど立っているのがやっとだった（栃木県のユーザ）」や「大変。おケガなどなければいいのですが（福岡県のユーザ）」などの震源地からそれほど近くないユーザの投稿や震源地から遠いユーザの投稿が多く見られた．そ

のため，さらに詳細に各地域の投稿数の反応を検証するため，各震度の投稿数を図 3.9 に示す．

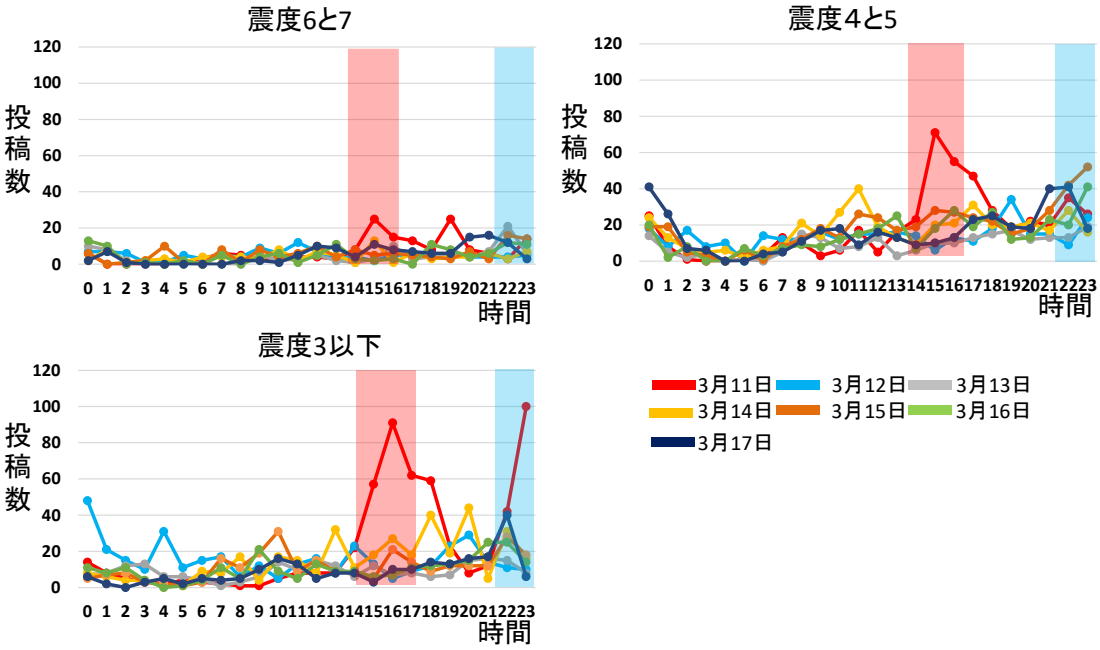


図 3.9　各震度による東日本大震災の解析結果

図 3.9 より震源から近い震度 6 や震度 7 を記録した地域では，投稿数が少なく，震源から遠い震度 3 以下の地域では，投稿数が多い傾向が見られた．これは，大規模自然災害などの深刻な災害において，被害が大きい地域のユーザは投稿する余裕がなかったためであると考えられる．また，静岡県で 3 月 15 日に地震があり，震度 4 と 5 の地域には静岡県が含まれていることから，投稿数が多い傾向が出ている．震度 3 以下の地域では，静岡県の地震にも比較的高い投稿数を示している．これは，3 月 11 日の地震の影響により，社会的に地震発生に敏感なっていることからこのような傾向が見られたものと考えられる．このことから，全国的な社会事象では，全体的な投稿数が増加する傾向があり，都道府県の各地域の投稿数の変化を検出することで社会事象を検出可能であることがわかった．

b)プロ野球東北楽天ゴールデンイーグルス日本一

プロ野球東北楽天ゴールデンイーグルス日本一は，2013 年 11 月 3 日に東北地方の宮城県に本拠地をおく野球チームが日本一になった事象である．各地方の投稿時間帯ベクトルを図 3.10 に示す．赤枠は，東北楽天ゴールデンイーグルス日本一が決まった時間帯を表す．図 3.10 より，次に示すことがわかった．

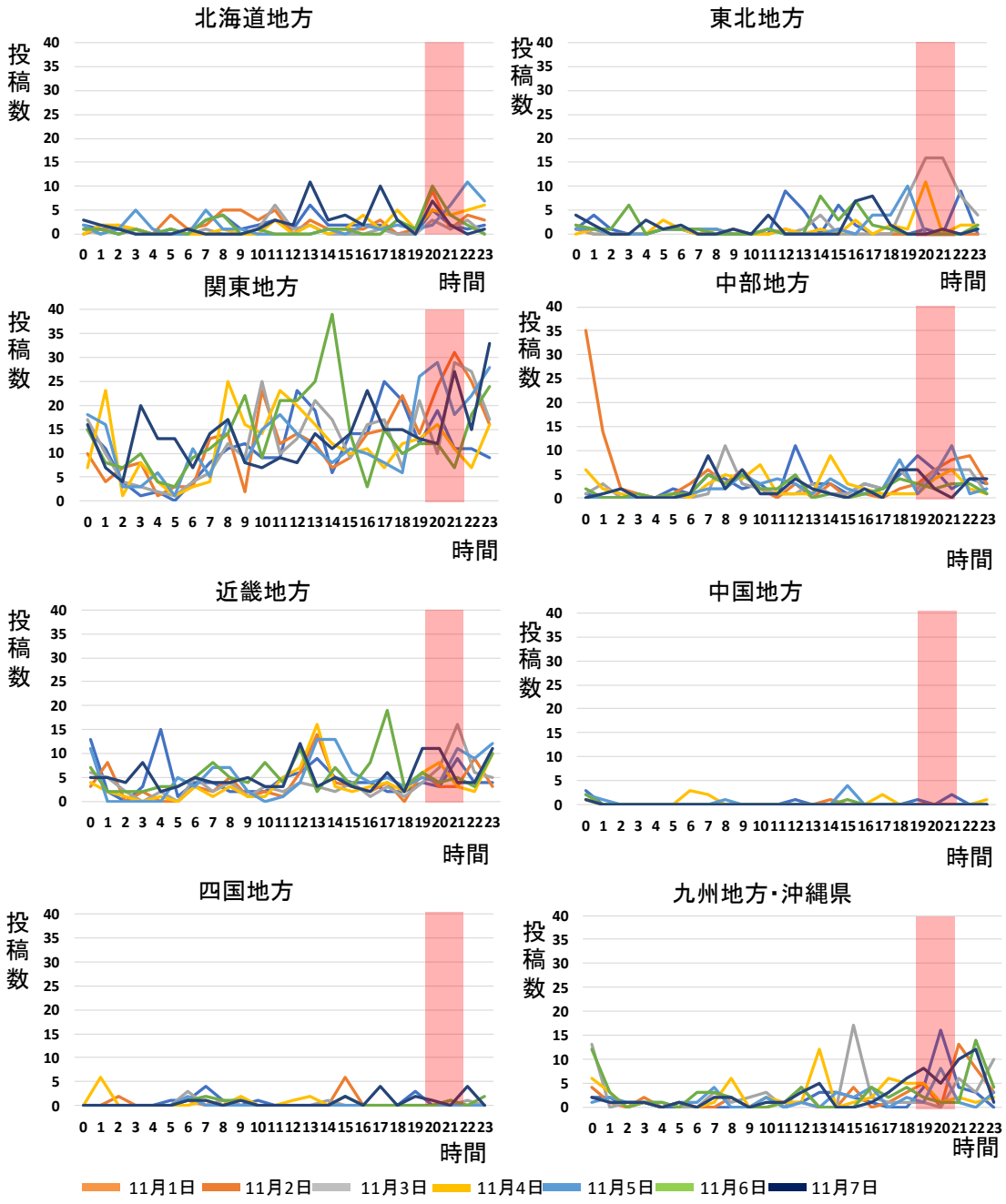


図 3.10　各地域による東北楽天ゴールデンイーグルス日本一の解析結果

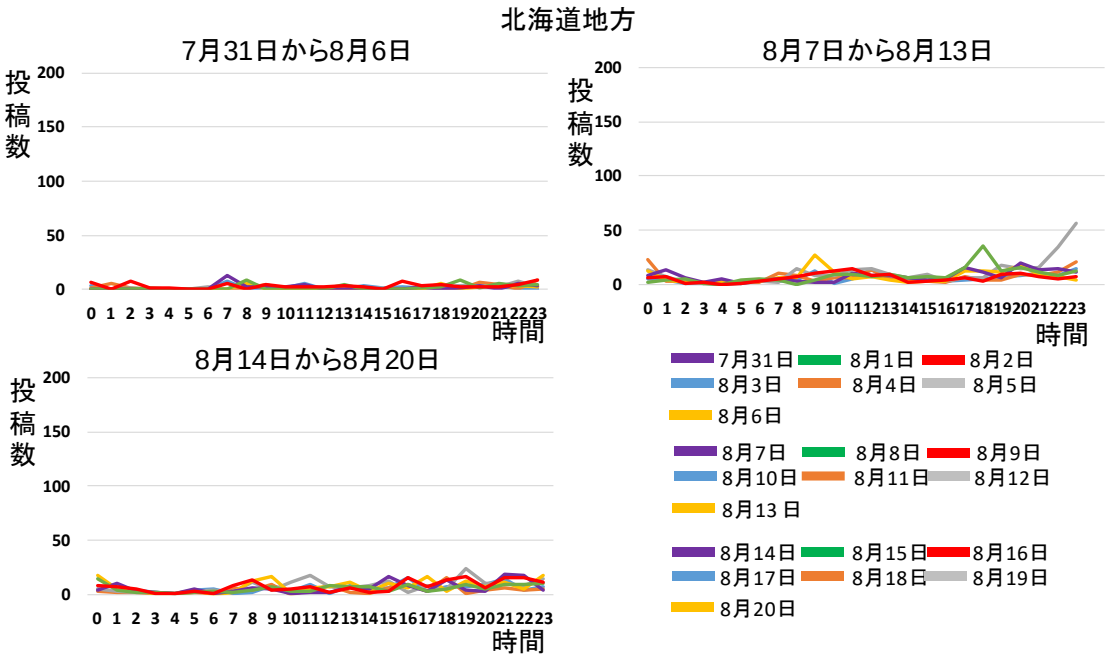
- **東北地方は他の地方に比べ投稿数が多いことがわかった**

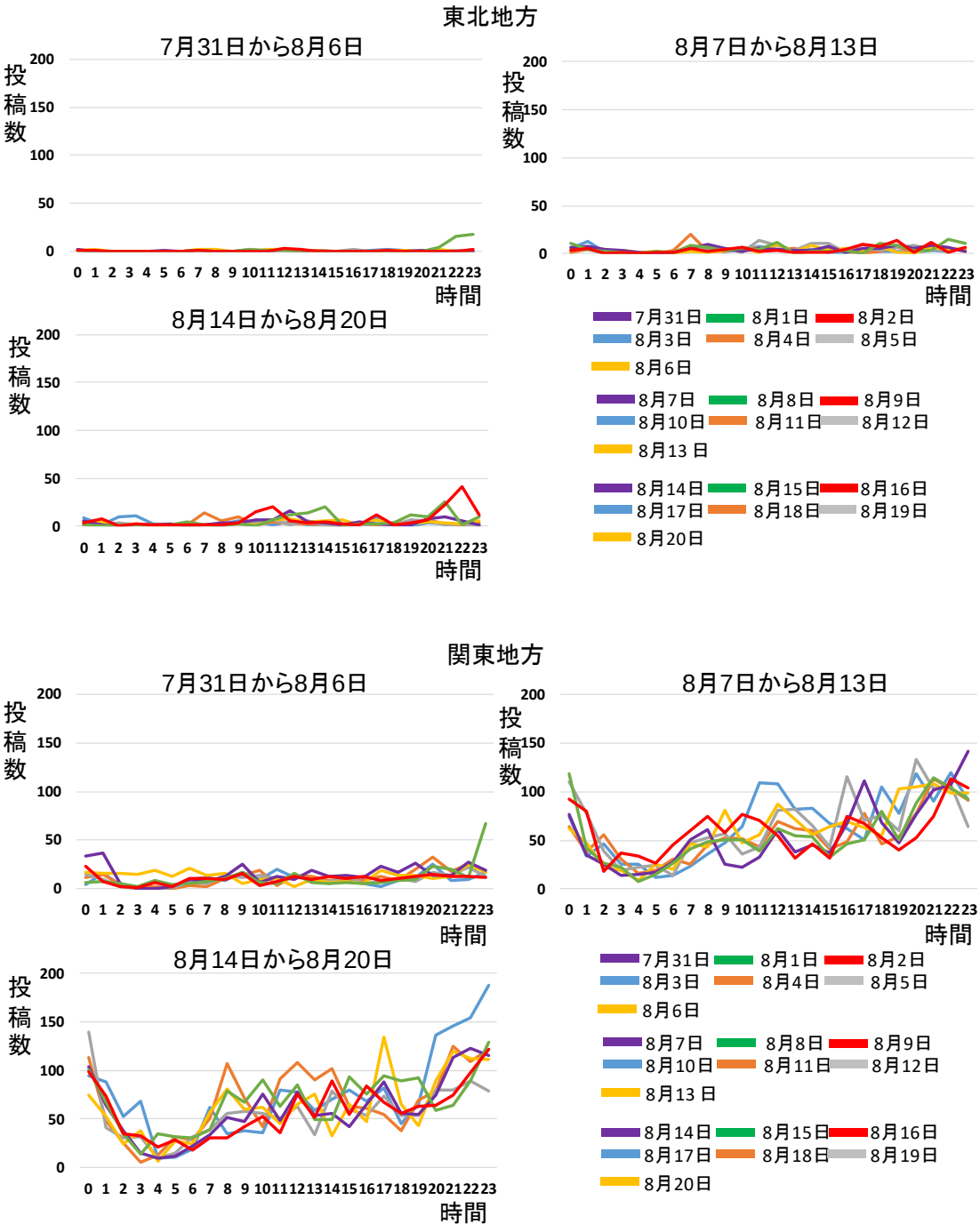
東北地方において 11 月 3 日 18 時以降から急激に投稿数が増加している．これは，東北楽天ゴールデンイーグルスの本拠地が宮城県であり，関心の高いユーザが多いためであると考えられる．また，初回に点数が追加されたのも大きな要因であると考えられる．実際に投稿を確認すると「お祭りだ！気持ちの昂ぶりを抑えられない！」といった内容の投稿が多いことがわかった．また，近畿地方においても，20 時から 21 時の投稿量が急激に増加し

ていることがわかった．これは，近畿地方には，対戦相手の読売ジャインツを敵対するユーザ多く，この傾向が現れたためと考えられる．このことから，各地方の投稿数の変化を検出することで，地方で異なる話題性の違いに関しての社会事象を検出可能であることがわかった．

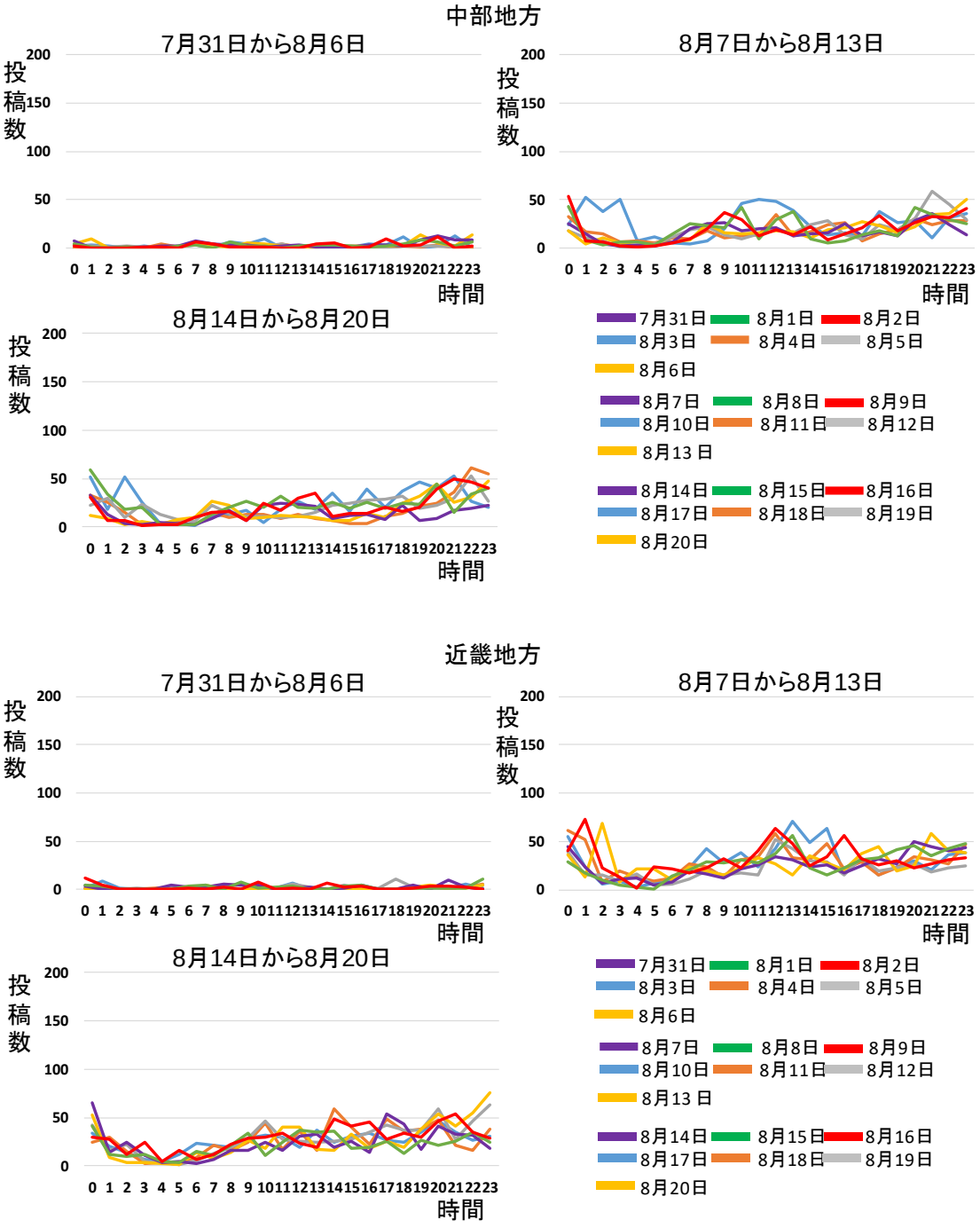
c)8月豪雨

8 月豪雨は，2014 年 7 月 30 日から 8 月 26 日にかけて，台風 11 号と 12 号及び前線と暖湿流により日本の広範囲で発生した豪雨である．各地方の投稿時間帯ベクトルを図 3.11 研究では，気象庁が，8 月豪雨に関して，大きく三段階に分けて示しているため，7 月 31 日から 8 月 20 日の 3 段階の投稿時間帯ベクトルを比較することにより時系列を確認する. 図 3.11 より次に示すことがわかった．




東北地方
7月31日から8月6日
8月7日から8月13日
8月14日から8月20日
投稿数
時間
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
0 50 100 150 200
7月31日 8月1日 8月2日
8月3日 8月4日 8月5日
8月6日
8月7日 8月8日 8月9日
8月10日 8月11日 8月12日
8月13 日
8月14日 8月15日 8月16日
8月17日 8月18日 8月19日
8月20日

関東地方
7月31日から8月6日
8月7日から8月13日
8月14日から8月20日
投稿数
時間
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
0 50 100 150 200
7月31日 8月1日 8月2日
8月3日 8月4日 8月5日
8月6日
8月7日 8月8日 8月9日
8月10日 8月11日 8月12日
8月13 日
8月14日 8月15日 8月16日
8月17日 8月18日 8月19日
8月20日


中部地方
7月31日から8月6日
8月7日から8月13日
8月14日から8月20日
投稿数
時間
0 50 100 150 200
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
7月31日 8月1日 8月2日
8月3日 8月4日 8月5日
8月6日
8月7日 8月8日 8月9日
8月10日 8月11日 8月12日
8月13 日
8月14日 8月15日 8月16日
8月17日 8月18日 8月19日
8月20日

近畿地方
7月31日から8月6日
8月7日から8月13日
8月14日から8月20日
投稿数
時間
0 50 100 150 200
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
7月31日 8月1日 8月2日
8月3日 8月4日 8月5日
8月6日
8月7日 8月8日 8月9日
8月10日 8月11日 8月12日
8月13 日
8月14日 8月15日 8月16日
8月17日 8月18日 8月19日
8月20日

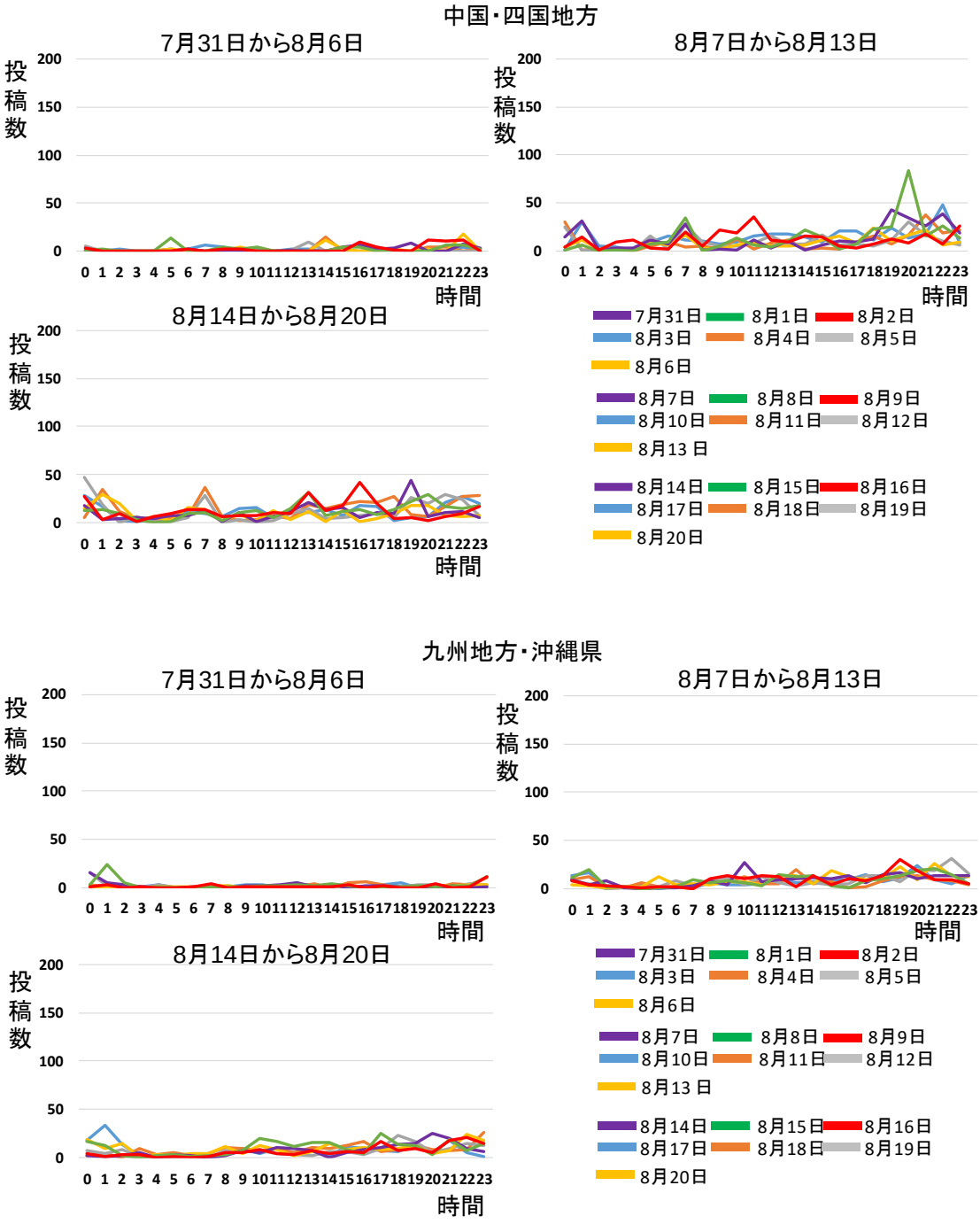


図 3.11　各地方の 8 月豪雨に関する解析結果

- **投稿量の総量が変化することがわかった.**

全体的にグラフを確認すると，関東地方の投稿量が多く，他地方は投稿量が少ない結果となった．これは，関東地方のユーザが多いことも考えられるが，一番の要因として関東

地方は他地方に比べ影響が少なかったからであると考えられる．また，各地方に分散して，被害をもたらしたことも原因の一つとして考えられる．

- **長期的な豪雨などの災害では，投稿傾向に差がみられることがわかった．**

時系列順に確認すると，7 月 31 日から 8 月 6 日は，投稿量が低い結果に対して，8 月 7 日以降の投稿量は全地方で増加する傾向が見られた．これは，通常の台風の豪雨と考えられていたが，長期的な災害となり，ニュースに取り上げられる回数も増加したためであると考えられる．

3.5.3　処理手法の検討

投稿傾向の変化から社会事象を抽出可能であるかの検証実験の結果のまとめを次に示す．

- **投稿傾向の変化から社会事象の抽出が可能であることがわかった**

全国的な話題である東日本大震災において，震度が小さくなるにつれて震災時の投稿量が増加することがわかった．これは，深刻な災害において，震度が大きな地域は被害も甚大であり，投稿する余裕がないユーザが多いことが考えられる．

- **投稿者の属性を考慮することで，地域の社会事象を抽出可能であることがわかった**

東日本大震災の震源地に近いあたりでは，余震が発生すると投稿数に変化がみられた．東日本大震災では，東日本大震災が原因の一つと考えられる震度 6.2 の地震が 3 月 15 日に発生しており，震度 4 と 5，震度 3 以下の地域では，投稿量が増加している．これは，3 月 11 日の地震の影響により，社会的に地震に敏感になっているからであると考えられる．

東北楽天ゴールデンイーグルス日本一の話題は，東北地方において関心が高く，その他の地方では関心が低い傾向がみられた．東北楽天ゴールデンイーグルス日本一の話題では，東北地方の地域では，11 月 3 日に他の日時に比べ投稿量が多く，他の地方では増加していないことがわかった．これは，東北楽天ゴールデンイーグルスの本拠地が宮城県あり，関心の高いユーザが多いためであると考えられる．

- **投稿量の総量が変化することがわかった．**

8 月豪雨に関して，被害が少なかった関東地方の投稿量が変化する点と長期的な災害となったため投稿量が増加する傾向がみられた．これは，通常の台風だと考えられていた台風が，条件が重なり，長期的に豪雨被害をもたらしたことが原因の一つとして考えられる．

上記の結果から課題点として，イベントが投稿量に依存してしまう点，投稿量の総量が変化する点，事象によっては投稿数が少ないため，社会事象の抽出が困難である点がわかった．これにより，投稿傾向のみでは，非習慣行動の抽出が困難であることがわかった．この結果を踏まえ，図 3.7 に示した検討手法 2 を検証する．検討手法 1 の結果を踏まえた対応策を次項で説明する．

(1)　検討手法1の結果を踏まえた対応策

a)社会事象時は，投稿傾向に差がみられるが投稿数のみでは非習慣行動を抽出困難な点

検討手法 1 の結果より，社会事象時には，投稿量の変化と投稿傾向に変化がみられることがわかった．本研究では，この結果を踏まえて社会事象時に投稿傾向が変化することは，ユーザが平時と異なる行動を起こしていると考えられる．その非習慣行動と平時の習慣行動を比較することにより，非習慣行動を抽出できると考えられる．

b)投稿量の総量が変化する点

検討手法 1 の結果より，社会事象時には，投稿量の総量に変化がみられることがわかった．この結果を踏まえて，社会事象時の各習慣行動で正規化を行い，その変化を考慮することで非習慣行動の抽出が可能であると考えられる．

(2)　処理方針

3.5.2 項の結果を踏まえた 3.5.3 項の対応策より，生活習慣を考慮することで，習慣行動の違いから非習慣行動の抽出が可能である考えられる．また，各習慣行動に着目することによって抽出できる社会事象も異なると考えられる．以上の点から検討手法 2 を検証する．

3.5.4　習慣行動の変化から社会事象を抽出可能であるかの検証実験

(1)　実験内容

本実験では，3.3 節で設定した「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，本提案のソーシャルセンシング技術が有用であることを検証する．実験 1 では，各習慣行動から非習慣行動をとるユーザ群が多い時間帯を明ら

かにする．実験 2 では，その時間帯のトピックを抽出することで，社会事象を取得可能であるかを検証する．

(2)　実験データ

本実験では，2014 年 1 月～12 月の投稿群を対象とする．実験データの収集方法を次に示す．

STEP1： TwitterAPI を用いて 2014 年 1 月～12 月に投稿しているユーザおよび投稿内容を取得する．

STEP2： Twilog を解析して，2014 年 1 月～12 月に投稿しているユーザを抽出し，その投稿内容を取得する．Twilog とは，Twitter に投稿された内容を各ユーザの投稿履歴をブログ形式で保存するサービスである．Twilog から取得したユーザが STEP1 のユーザと重複している場合は，Twilog のデータを優先して採用する．

STEP3： STEP 1 および STEP2 で収集した投稿内容の件数が 1,000 件以上のユーザを実験データとする．ただし，ライフスタイルの解析に最低限必要な 1 週間分の投稿内容を取得できないユーザは，実験データから除外する．

以上の手順にて収集した実験データの詳細は，全ユーザ数 1,514 ユーザ，全投稿数 3,331,423 ツイートとなっている．なお，投稿件数は，TwitterAPI および Twilog の仕様に基づき収集したツイートの件数を示している．ただし，TwitterAPI では，投稿内容の取得制限のため，最大で 3,200 件となった．

(3)　パラメータ設定

本実験では，非習慣行動算出に関わるパラメータα，βおよびトピック抽出で用いる潜在的ディリクレ配分手法のトピック数を設定する．パラメータα，βは，非習慣行動抽出処理において，平時習慣ベクトルと特定習慣ベクトルの差分から面積を算出時に用いる値である．

LDA の手法で用いるトピック数は既存研究[45]に倣い，100 トピックとした．なお，習慣行動によって投稿件数が変化するため，月全体の投稿件数を 100 とし，各行動の投稿件数で割合を算出し，投稿量によって抽出できる内容量の変化を軽減する．例えば，月全体の投稿数：10,000 件，起床・就寝：2,000 件，外出：3,500 件の場合，月全体：100 トピック，起床・就寝：20 トピック，外出：35 トピックとする．また，LDA の手法の反復回数は 50 回とした．

3.5.5 実験1：平時と異なる行動の抽出実験

(1) 実験内容

本実験では，3.3 節で挙げた検証項目 1 に関して，平時と異なる行動を起こすユーザ群を特定してその投稿を解析することで，社会事象を抽出可能であるかを検証する．方法として，平時習慣ベクトルと特定習慣ベクトルを元に平時と異なる行動の時間帯を算出し，その非習慣行動から社会事象を抽出する．実験 1 は，次の手順に従って分析を実施する．

STEP1： 平時のユーザの生活習慣を明らかにするため，2014 年 1 月～12 月の全投稿群からリツイートなどの投稿を除去した投稿を対象に平時習慣ベクトルを作成し，各習慣行動で分類する．

STEP2： 2014 年の各月で特定習慣ベクトルを作成し，各習慣行動で分類する．

STEP3： STEP1 で作成した平時習慣ベクトルと STEP2 で作成した 12 ヶ月分の各特定習慣ベクトルを比較し，非習慣行動が発生した時間帯を抽出する．

STEP4： 非習慣行動が発生した時間帯を分析し，その際に発生した社会事象を分析する．

(2) 実験結果

各行動の非習慣行動の抽出時間数を表 3.4 に示す．

表 3.4 各月の非習慣行動数

行動	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	平均	標準偏差
起床・就寝	73	74	62	67	66	**86**	79	73	69	70	75	85	73.3	6.98
在宅	81	84	79	69	**86**	66	78	81	78	76	**86**	70	77.8	6.30
外出	66	**81**	74	59	73	70	75	73	54	65	74	59	68.6	7.70
帰宅	71	67	75	**77**	68	71	74	68	67	68	69	62	69.8	3.94
平均	73	**77**	73	68	73	73	**77**	74	67	70	**76**	69	72.5	3.23
標準偏差	5.40	6.58	6.34	6.40	7.79	7.60	2.06	4.66	8.57	4.02	6.20	10.07		

各行動で最も多い月は太字，低い月は下線で示す．表 3.4 を分析した結果，明らかになった内容を次に示す．また，各月の平均の上位 3 件を太字の網掛けで示す．

- **非習慣行動は，各習慣行動に関連して発生する傾向がみられる**

提案手法では，非習慣行動の時間帯から各月の期間を抽出できることがわかった．表 3.4 より各行動の中で，最も非習慣行動の多い月は起床・就寝で 6 月，在宅で 5 月と 11 月，外出で 2 月，帰宅で 4 月となった．最も低い月は起床・就寝で 3 月，在宅で 6 月，外出で 9 月，帰宅で 12 月となった．また，2014 年の大きなトピックとして，2 月はソチオリンピック，7 月はワールドカップ，集団的自衛権の行使の閣議決定など平時と異なる祭典や関心度の高い事象があったことからこのような結果が出たと考えられる．実際に投稿を確認すると「あー心臓バクバク！凄い緊張ですね！！！神様お願い！羽生くんに金！」，「ああ、残念。ドイツおめでとう。アルゼンチン、本当に惜しかった…」，「安倍内閣の憲法改正発議、集団的自衛権行使の法制には賛成を投じられたら終わり。ここは非自公＋政党助成金目当ての衛星政党でいくしかない。」などの投稿が多いことがわかった．

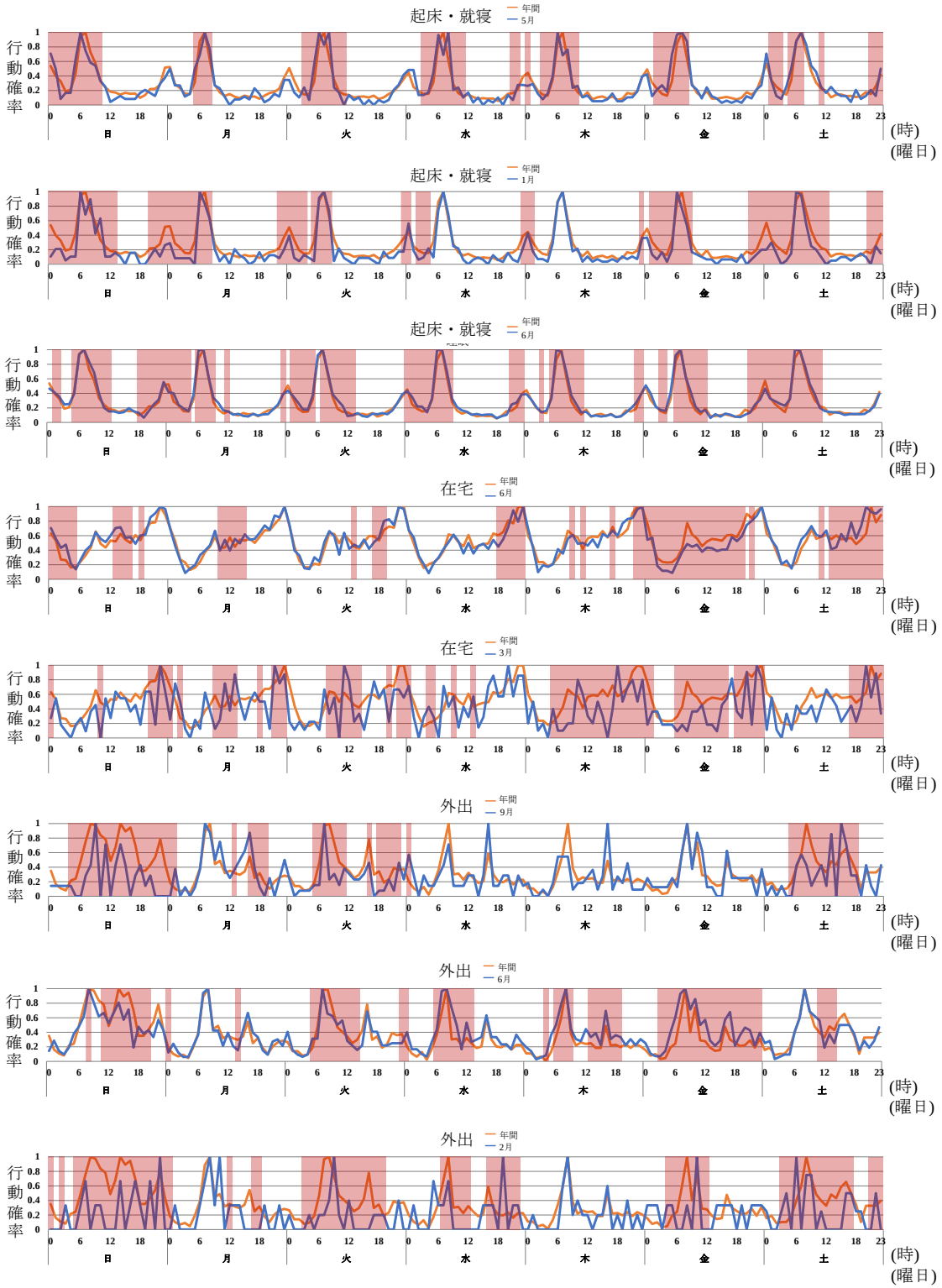
- **各行動の中で外出が最も標準偏差が高いことわかる**

表 3.4 より，どのユーザも基本的に外出時間は変わらず，一定のリズムであると考えられる．しかし，2 月に最も高い 81 時間という値を示していることがわかる．これは，2 月は記録的な大雪に見舞われ公共交通が不安定になったため，非習慣行動が増えたと考えられる．次いで高い時間を示す 7 月は，台風 8 号の影響により初の特別警報があり，土砂災害などの被害をもたらしたためであると考えられる．

各行動の平均が最も高い在宅に関しては，掃除，テレビの視聴や入浴など日常生活において，必ずしも決まった時間に行われない行動であり，外出とは逆に一定のリズムが崩れることが多いためであると考えられる．

- **生活習慣には一定のリズムがある**

生活習慣ベクトルを確認すると，習慣行動には一定パターンがあることがわかった．表 3.4 において，行動の抽出時間が，最小，中央値と最大値の平時習慣ベクトルと特定習慣ベクトルの比較結果を図 3.12 に示す．赤枠は，非習慣行動の検出機能で抽出した非習慣行動の時間帯を示す．


起床・就寝
年間
5月
行動確率
(時)
(曜日)
起床・就寝
年間
1月
行動確率
(時)
(曜日)
起床・就寝
年間
6月
行動確率
(時)
(曜日)
在宅
年間
6月
行動確率
(時)
(曜日)
在宅
年間
3月
行動確率
(時)
(曜日)
外出
年間
9月
行動確率
(時)
(曜日)
外出
年間
6月
行動確率
(時)
(曜日)
外出
年間
2月
行動確率
(時)
(曜日)

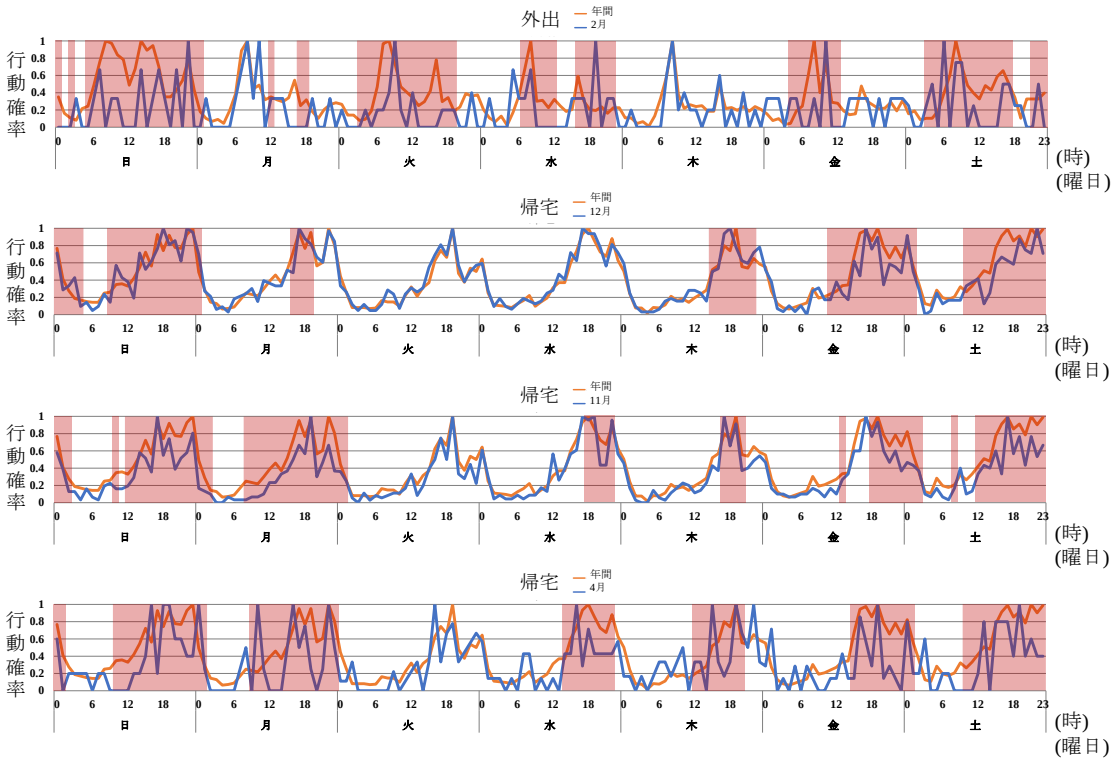


図 3.12　上位と下位の生活習慣解析結果

3.5.6　実験2：非習慣行動に着目したトピック抽出

(1)　実験内容

本実験では，3.3 節で挙げた「ソーシャルセンサの特性としてユーザの生活習慣を考慮することで，抽出可能な社会事象の内容やカテゴリが変化すること」を証明するため，非習慣時と抽出された期間の投稿をトピック分類した結果に基づき解析する．表 3.4 より本実験では，非習慣行動が多くみられた 2 月と 7 月を対象に解析する．

実験では，次の手順に従って分析を実施する．

STEP1：2 月と 7 月の非習慣行動とされた期間の投稿を抽出し，それらの投稿を各月の各習慣行動に分類する．

STEP2：トピック抽出処理にて，分類された投稿を解析し，各月の各習慣行動からトピック（以下，提案手法のトピック）を抽出する．

STEP3：2 月と 7 月の全投稿を抽出し，それらの投稿を各月で分類する．

STEP4：トピック抽出処理にて，分類された投稿を解析し，各月からトピック（以下，既存手法のトピック）を抽出する．

STEP5：各月の既存手法と提案手法のトピックから，それぞれトピック内の単語を取得し，

対応する期間にニュースやイベントがあるかを検索する．検索結果が得られなかった場合，該当トピックをノイズに分類する．

STEP6： 各月の既存手法と提案手法のトピックを構成する単語を比較し，5 単語以上の単語が同一であった場合，一致するトピックに分類する．これは，解釈できたトピック内のキーワード数を平均すると，約 5 単語となったためである．

以上の手順で，一致するトピックと非一致のトピックを分析し，提案手法の特性を考察する．上記の処理手順のフローチャートを図に 3.13 に示す．

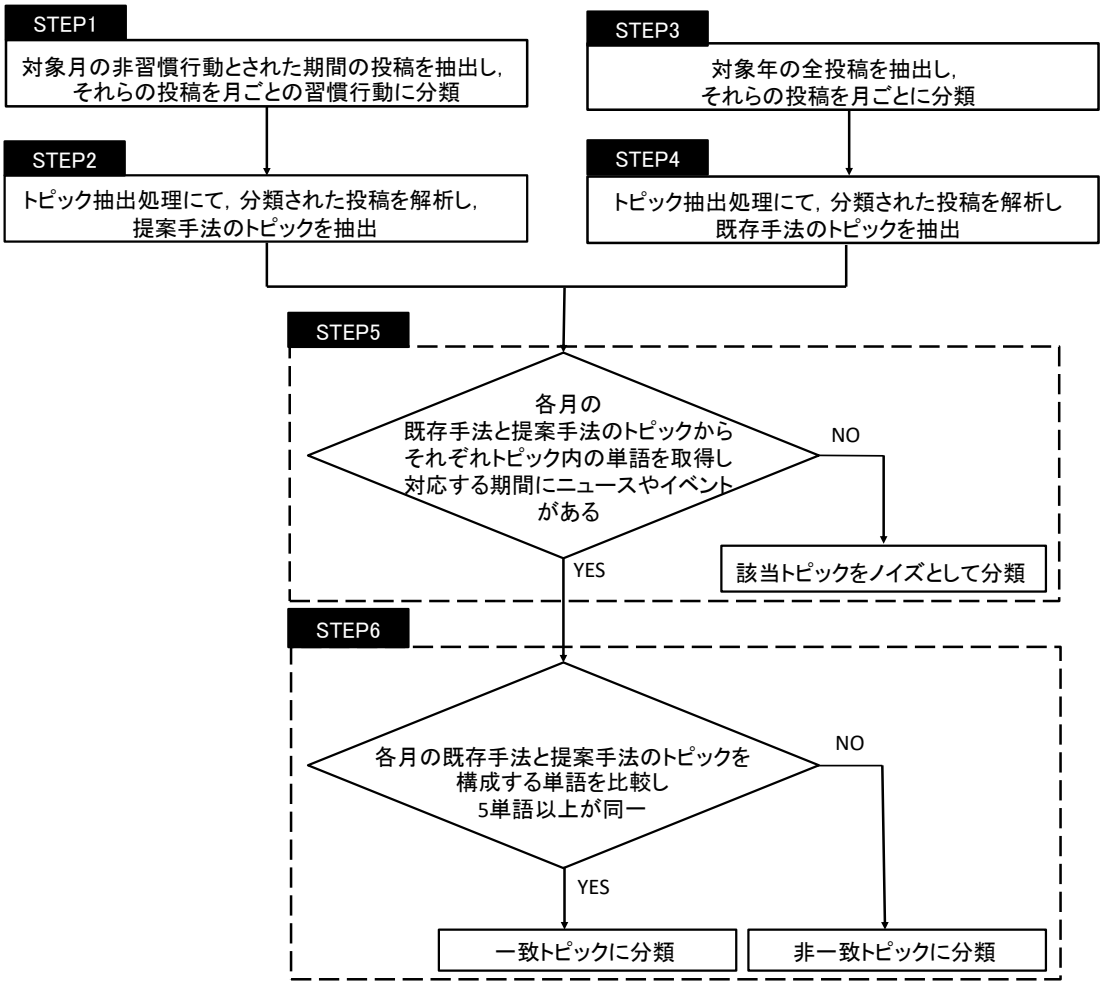


図 3.13　処理手順のフローチャート

(2)　実験結果

各月の各行動の一致率を表 3.5 に示す．本実験では，2014 年の 2 月と 7 月の各行動で算出したトピックから実世界で起きた社会事象を抽出した．2 月の投稿件数は 54,334 件，7 月の投稿件数は 144,920 件である．

表 3.5　各月の非習慣行動数

			既存手法	提案手法				
				起床・就寝	在宅	外出	帰宅	合計
2 月	正常	一致	17(17%)	6(13%)	4(6%)	11(21%)	10(19%)	31(14%)
		非一致	54(54%)	16(33%)	40(63%)	25(48%)	29(56%)	110(51%)
	ノイズ		29(29%)	26(54%)	26(31%)	19(31%)	16(25%)	74(35%)
7 月	正常	一致	11(11%)	6(14%)	6(11%)	6(11%)	3(6%)	21(11%)
		非一致	64(64%)	21(50%)	21(40%)	11(20%)	14(30%)	67(34%)
	ノイズ		25(25%)	15(36%)	26(49%)	37(69%)	30(64%)	108(55%)

- **既存手法と比較して，提案手法は多くのトピックを抽出できる傾向が見られた**

表 3.5 の既存手法の正常トピック数と提案手法の合計の正常トピック数を比較すると，提案手法の方が多くのトピックを抽出できていることがわかった．その一方で，提案手法は，ノイズに分類されるトピックも多くみられ，トピック抽出後のフィルタリング方策を検討する必要があることがわかった．

- **既存手法と提案手法とで抽出できるトピックが異なる傾向が見られた**

表 3.5 のトピック抽出結果を確認すると，既存手法と提案手法のトピックの非一致数が多いことがわかった．これは，それぞれの手法にて，抽出されたトピックが異なることを示している．詳細を分析するため，各手法にて抽出された非一致のトピックの一例を表 3.6 に示す．カテゴリ，キーワードはトピック抽出によって取得したキーワードを基に，トピックを解釈している．

表 3.6　各月の非一致トピック

	月	カテゴリ，キーワード
既存手法	2 月	都知事が議会で報告(都知事，状況，議会，報告) 予算の承認(戦略，予算，年度，承認)
	7 月	政治に関する話題（自民党，選挙，獲得） 経済に関する話題（金，起業，サラリーマン，年収）
提案手法	2 月	安倍政権が調査(安倍，経済，政府，調査) 原発で問題発生(原発，NHK，問題，職員)
	7 月	原発に関するニュース(原発，ニュース，経済，研究，避難) 安倍政権が地方支援を検討(安倍，政権，支援，地方，メンバー)

表 3.6 を確認すると，既存手法は一般的なトピックが抽出されているのに対し，提案手法はイベントやその時事に特化したトピックが抽出されていることが明らかとなった．このことから，既存手法と提案手法とは異なる特性のトピックを抽出することができるため，相互補間的に活用できると考えられる．

- **提案手法では各行動において抽出できるトピックが異なる傾向が見られた**

習慣行動と抽出トピックとの関係を分析するため，抽出トピックをカテゴリ分類した結果を表 3.7 に示す．

表 3.7　トピックのカテゴリ

行動	生活	エンタメ	スポーツ	政治・経済	その他	標準偏差
起床・就寝	15(19%)	**27(35%)**	4(5%)	15(19%)	17(22%)	7.31
在宅	**31(31%)**	21(21%)	5(5%)	24(24%)	18(18%)	9.70
外出	**26(35%)**	12(16%)	3(4%)	22(30%)	11(15%)	8.23
帰宅	22(23%)	23(24%)	**7(7%)**	**25(27%)**	17(18%)	6.46
標準偏差	5.85	6.65	1.64	3.67	3.00	

表 3.7 において，本研究では，生活，エンタメ，スポーツ，政治・経済とその他のカテゴリに分類した．カテゴリ分けに関しては，解析者のバイアスを緩和するため Yahoo!カテゴリを参考に複数人で行った．表 3.7 の数値は，カテゴリと各行動が最大値となる値に下線を引いている．

表 3.7 を確認すると，各行動で抽出できるトピックが変化することがわかった．生活のカテゴリは在宅と外出の行動が高い割合を示している．エンタメのカテゴリは，起床・就寝

の行動で高い割合を示している．これは，分類する際にアニメに関するトピックをエンタメにしたため，深夜のアニメの投稿が目立ったためこのような結果になったと考えられる．スポーツと政治・経済・災害・事故のカテゴリは，帰宅の行動で高い割合を示している．このように，各習慣行動で抽出できるトピックは異なっている傾向が見られた．

- **既存手法と提案手法とで抽出できるトピックの内容が異なる傾向が見られた**

表 3.6 において，既存手法と提案手法の合計の一致トピック数を比較すると，提案手法の件数が多くなっていることがわかる．これは，既存手法の一つのトピックに対して，提案手法では，複数のトピックに分割して抽出していることを示している．各月の調査した主な社会事象を表 3.8 に示す．表 3.8 において，赤字は本提案手法にて抽出できた社会事象を示す．

表 3.8　主な社会事象

月	日	社会事象
2 月	2 日，9 日	長崎知事選，東京都知事選
	5 日	佐村河内氏のゴーストライター問題
	7 日～23 日	ソチオリンピック
	14 日，15 日	関東・甲信地方を中心とした大雪
	22 日	ソニー・コンピュータエンタテインメントが家庭用ゲーム機「プレイステーション 4」を発売
7 月	1 日	集団的自衛権の行使の閣議決定，野々村議員政務活動費不正問題
	1 日～13 日	FIFA ワールドカップ
	2 日	ネイチャーが STAP 細胞に関する 2 本の論文を撤回
	7 日	台風で初となる特別警報を沖縄地方に発令
	13 日	滋賀県知事選挙
	26 日	佐世保女子高生殺害事件が発生

表 3.8 において，幅広いジャンルの社会事象の抽出できていることがわかる．

一致内容の詳細を確認するため，各月の一致したトピックの対応関係をまとめた結果を表 3.9 に示す．

表 3.9　社会事象

月	トピック	指定キーワード	既存手法	提案手法	
			解釈トピック（キーワード）	行動	解釈トピック（キーワード）
2月	ソチオリンピック	指定なし	オリンピック・登場・話題・羽生	起床・就寝	ソチオリンピックを都知事が応援（都知事，応援，オリンピック，ソチ）
		オリンピック，ソチ，羽生，五輪	羽生選手男子フィギュア金メダル（オリンピック，ソチ，羽生，金メダル，男子，フィギュア） 浅田選手や高橋選手の活躍（感動，真央，高橋）	在宅	ソチオリンピック高校生や大学生の活躍（大学，高校生，オリンピック，発見，ソチ）
				外出	浅田真央の活躍（選手，オリンピック，感動，埼玉，彼女，五，ソチ）
	大雪	指定なし	大雪・長期・熊谷 東京・大阪・雪だるま	在宅	静岡県久しぶりの除雪（久しぶり，31，除雪，記録，静岡，配信，視聴）
		積雪，大雪，関東，甲信，仙台	東京，千葉など各地で大雪（大雪，積雪，埼玉，新潟，千葉），（関東，立ち往生，東京），（大雪，大変，山梨）	在宅	千葉で去年の倍の雪（心配，千葉，大雪，NHK，ぶり，去年，倍）
	都知事選	指定なし	田母神・都知事・演説	外出	舛添氏が都知事選に関わった（都知事，東京，舛添）
		東京都知事選，舛添	都知事選舛添氏を安倍総理が応援し当選（都知事，舛添，選挙，安倍，応援，当選） 田母神氏が都知事選に関わった（都知事，選挙，田母神，東京）		
7月	ワールドカップ	指定なし	ワールドカップ・アルゼンチン・過去	在宅	ワールドカップのアルゼンチンの試合が NHK で TV 放送（放送，NHK，アルゼンチン，ワールドカップ，TV）
		ワールドカップ，FIFA	ワールドカップドイツとアルゼンチンの決勝（ワールドカップ，決勝，ドイツ，アルゼンチン） ワールドカップドイツが優勝（ワールドカップ，ドイツ，決勝，優勝）		

7 月	台風	指定なし	雨・台風・傘	在宅	宮崎で台風により大雨 （台風，影響，大雨，発生，宮崎）
		特別警報， 大雨，土石流， 台風 8 号	各地大雨警報で被害 （注意報，警報，大雨，洪水，浸水， 熊谷），（台風，警報，大雨，クラス， 最強，洪水） 沖縄県で特別警報発表 （台風，警報，沖縄，特別，避難， 最強，クラス，発表，暴風雨）		
	野々村議員政務 活動費不正事件	指定なし	議員，野々村	起床・ 就寝	野々村議員号泣 （反対，野々村，説明，号泣，選挙， 行為）
		野々村， 政務活動費， 不正，号泣会見	野々村議員号泣会見 （野々村，議員，耳，手，号泣）， （野々村，議員，西宮，辞職）		
	集団的自衛権の 閣議決定	指定なし	自衛・権・マジ・閣議	在宅	安倍政権が集団的自衛権の行使を 閣議決定 （集団，本日，安倍，政権，閣議， 米国），（自衛，権，集団，的）， （集団，行使，閣議，決定）
		集団的自衛権， 閣議決定， 臨時閣議	安倍政権が集団的自衛権の行使を 閣議決定 （自衛，集団，行使，安倍，首相）， （自衛，集団，アメリカ，日本，容認， 安倍，閣議），（自衛，権，集団， 閣議，決定，行使，安倍）		

表 3.9 を確認すると，提案手法では，既存手法と比較して，該当のトピックを詳細に表していることが明らかとなった．各トピックについて，既存手法と提案手法とを比較して，明らかとなった内容について次に示す．

ソチオリンピックのトピックは，ロシアで行われた冬季オリンピックである．既存研究は羽生選手がオリンピックで活躍したことがわかる．しかし，提案手法では，都知事がオリンピックに応援を行ったこと，高校生や大学生が活躍したことなど既存手法では抽出できなかったことが抽出できていることがわかる．

大雪のトピックは，都心でも記録的な大雪が降った事象である．既存手法からは東京，大阪や熊谷など都市部でも雪が降った事実が確認できる．しかし，提案手法では，千葉で去年の倍雪が降ったことが確認できる．実際に確認すると約 2 倍の雪が降っていることを確認することができた．このように雪が降った事実だけでなく，どれくらい降ったのか詳細を抽出することができた．

都知事選のトピックは，2 月 9 日に執行された東京都知事選挙の事象である．既存手法か

らは田母神氏が都知事選で演説した事実が確認できる．提案手法では，それに加えて，都知事がソチオリンピックを応援したこと，田母神氏以外に舛添氏が都知事選に関わった事実が確認できる．

ワールドカップのトピックは，既存手法ではワールドカップ，アルゼンチンと過去が抽出できた．提案手法では，アルゼンチンの試合が NHK でテレビ放送された事実が確認できた．

台風は，7 月に発生した台風の影響より，全国的に大雨をもたらした事象である．既存手法では台風 8 号に関連するキーワードは台風，雨と傘，各地大雨警報で被害や沖縄県で特別警報発表に関する話題が抽出できた．提案では台風，影響，大雨と宮崎といった被害の大きかった地域も抽出することができた．実際に確認すると，宮崎県のえびの市で 1 時間に 77 ミリといった大雨があったことが確認できた．

野々村議員政務活動費不正事件は，野々村議員が政務活動費を不正に受給していた事象である．既存手法では野々村議員のキーワードは抽出できたものの，野々村議員が何をしたのかが判断できる単語は抽出できなかった．提案手法では，キーワードを指定した場合と同様の内容が抽出でき，野々村議員が号泣した事実が抽出することができた．

提案手法で抽出した抽出対象以外のトピックを表 3.10 に示す．

表 3.10　抽出対象以外の習慣行動のトピック

月	提案手法	
	行動	解釈トピック（キーワード）
2 月	起床・就寝	安倍政権が調査（安倍，経済，政府，調査）
	外出	原発で問題発生（原発，NHK，問題，職員）
	帰宅	横浜と千葉で地震発生（横浜，発生，千葉，地震）
7 月	起床・就寝	原発に関するニュース（原発，ニュース，経済，研究，避難）
	在宅	安倍政権が地方支援を検討（安倍，政権，支援，地方，メンバー）
	外出	平和記念ドラマ放送（ドラマ，平和，７０，ぶり）

表 3.10 に示すとおり，横浜と千葉の地震や平和記念ドラマの放送など既存手法では抽出できないトピックが検出可能であることがわかった．

3.5.7　本手法におけるまとめ

本手法におけるまとめを次に示す．

- **投稿を生活習慣のみで絞ったため，トピック内に判断が難しいキーワードが存在した．**

本研究では，投稿を生活習慣のみで絞り，非習慣行動を抽出したため，トピックの解釈が困難であったことがわかった．解釈が困難なトピックの一部を表 3.11 に示す．

表 3.11　解釈が困難なトピック

月	行動	トピック
2月	起床・就寝	日，1，月，目，2，気，選挙，数，円，万，16，02，3，研究，誕生，24，10，億，記念，投資
		前，家，たち，Feel，幸せ，先生，これ，お母さん，そこ，確認，帰宅，74，58，場所，ヤバイ，こっち，厨，イオン，会見，0315
		報酬，誰，話，gt，ここ，lt，知識，別，見直し，さん，割，勉強，中央，32，堂，借金，惨敗，残念，成果，屋
		省，39，ら，chanrisa，12，末，様，愛，番，or，BS，OPEN，1，to，出番，04，LIVE，笑顔，キー，600
	在宅	者，次，30，心，版，風，amp，人人，活動，2001，団，学生，you，Harry，哲学，on，目標，元気，委員，季節
		東京，名前，友達，大切，無事，ルール，場所，恋，世の中，祭り，歌，54，チェック，法，秋葉原，作曲，面倒，保護，北九州，行政
		期間，限定，大阪，大好き，18，いつ，レポート，日本一，特別，無料，おすすめ，Tatsuya，利用，イン，過去，417，府，新規，メモ，記憶
		好き，今年，夜，やつ，誕生，結果，29，冬，味，ツイッター，黒，旦那，The，06，色，質問，偶，バレンタイン，学会，米国
		最近，普通，犬，トーク，紙，大人，父，向上，高校，ペット，ほんま，dog，注文，おかげ，金額，環境，客，到達，テンション，継承
	外出	８，83418，Leaf，枚，９，大人，行動，komohappy，主婦，中，一言，22，トイレ，企業，LINE，検討，ちゃん，28，天気，イメージ
		二，30，12，さん，正直，13，41，本日，愛，記念，残念，子ども，発売，2，love，物事，3，気分，km，you
		会，交換，気軽，完全，チョコ，案内，２月，いつ，日，時，ぶり，すべて，?。，担当，対応，週間，心配，今日，お待ち，細胞
		動画，何，次，さん，14，意味，女性，今度，大変，31，★★★★★¥u3000，com，僕，大事，気持ち，今日，そう，私，男性，自分

	帰宅	くん，夜，一緒，遊び，参加，かた，トマト，000，綺麗，観，22， 有料，妹，イケメン，正直，mi，大根，相談，匂い，22677873
		13，安，千，こんばんは，無駄，発売，投稿，夫，momo，新聞，汗， 冊，価格，会，今日，後ろ，不正，靴，夫婦，大会
		それ，映画，必要，音，全部，下，大切，kuma，巻，未来，差， 福島，本人，展開，スキル，会場，過去，モデル，高校生，セット
		通り，間，news，区，海外，速報，1028，東京，大学，takapon，shinzan， yu，umai，ホテル，kurage，メッセージ，しま，初日，静岡，外国
		無料，円，オファー，JsnGetaC，回，万，メールアドレス，労働， アフィリエイター，私，性，子，17，可能，23，以降，31， ビジネス，オファーアフィリエイト，比
7 月	起床・就寝	人間，生活，⇒¥u3000，26，まじ，音楽，飯，保護，イメージ， よう，残り，調子，by，紙，組，おかげ，条，容認，on，勇気
		自衛，ダメ，心，決定，場合，すぎ，猫，電話，セット，速報， mxwelsilverhamr，反応，利用，企画，話題，自身，本当，近く， 重要，私
		子供，名，的，参加，絵，氏，者，家族，是非，完全，トーク， 社会，現在，晩，旅行，発言，ゃんこ，bami，ochi，問題
		最近，30，全部，放送，必要，開始，代，屋，内，妊活，主婦， 情報，脳，男性，そば，自分，お客，サラリーマン，瞬間，jhaiku
		時，無料，登録，メール，女性，系，者，サイト，出会い，迷惑， 的，サクラ，意味，最大，フォロワー，活動，24，名前，皆無，50
	在宅	月，日，gt，lt，数，先生，07，うち，作品，お腹，京都，祭， チケット，交換，38，追加，お待ち，news，31，定期
		最近，11，日，なに，名，幸せ，月，娘，公開，枠，さっき，開催， 土，抽選，ひとり，in，詳細，型，日間，以降
		市，県，性，可能，24，ab，個人，通り，34，色，安心，酒，隣， 購入，部分，PC，予想，おじさん，対策，ひま
	在宅	12，mxwelsilverhamr，love，配信，LINE，家族，新聞，理由，逆， 公園，世の中，29，39，福岡，獲得，現代，東，西，マンガ，重要
		好き，今回，週間，席，タイム，via，市民，変更，攻撃，四，DVD， タイムラプス，コマ，札幌，制度，66，ドライブ，見た目，晋， モンハン
	外出	運動，子供，普通，可能，間，性，問題，方法，的，犬，行使，体， 酸素，対応，解決，時間，Google，友人，教師，発言
		台風，次，店，白髪，さん，動画，ツアー，52，ひとり，交換，YouTube，

	21，hatebu，users，評価，コーヒー，予防，ポケモン，セブン，体重
	これ，話，誰，わたし，どこ，言葉，気持ち，名前，方，モンスター， 途中，ドラマ，ポスト，私，徹底，平和，70，放送，ぶり，今
	昨日，力，アルジェリア，くん，顔，想像，頭，公開，中，アン，?！， だめ，さん，小学生，花子，名古屋，内閣，苦手，移動，塩
	め，外，上，全員，ノイアー，ただ，1123，ksk，梅雨，近所， シリーズ，民主，紅，さん，人，東方，定期，機，91，代表
帰宅	情報，夏，性，絶対，可能，肉，行為，残念，話題，書，狼， 自然，卒業，短冊，新聞，祭り，サミタ，綺麗，ほか，劇場
	車，予定，今度，ヶ月，気分，番組，報告，赤，今，わら，メンバー， 章，名古屋，ホームページ，公園，自体，地球，メーカー，差，kingjim
	ブログ，更新，中国，女の子，彼氏，真，本当，反応，89，iPhone， 袋，コメント，状態，逆，miwa，投資，五，形，香港，ホーム
	目，回，先，夜，自衛隊，前，神，とこ，変，意識，歴史，本人， 09，ジャイアンツ，パンツ，了解，UP，科学，当選，サイズ
	俺，rt，20，三，撮影，素敵，京都，桜，料理，イメージ，mitsu， 音楽，失礼，000，The，2000，居酒屋，メニュー，別，チャンネル

表 3.11 より，トピック中のキーワードのみを確認すると社会事象に直結する単語はあるものの，社会事象と解釈するには，他のキーワードが不足していることがわかった．また，it や一文字など，それら単体では意味を把握できないキーワードがあった．これは，一文字でも重要なキーワードは存在するため，辞書の作成を工夫する必要性がある．さらに形態素解析をするのではなく，文章全体を分析することで，ユーザがどのような内容に対して投稿しているのかを把握でき，社会事象との関係を明確に判断できると考えられる．

- **短期的なイベントなど，事象によっては投稿数が少ないため，十分な習慣行動の変化の抽出が困難なことがわかった**

本研究では，非習慣行動から社会事象の抽出を研究目的としたため，社会事象の規模を考慮できていないことからこの課題が発生したと考えられる．また，非習慣行動の抽出においてパラメータ値の設定も原因として考えられる．この課題に対しては，全ユーザの投稿を対象にするのではなく，性別，地域や職業等，各ユーザ属性で絞った非習慣行動を用いて，トピックを抽出することにより，社会事象の把握可能なトピックの割合が増えると考えられる．

3.6 あとがき

本章では，代表的なマイクロブログである Twitter を対象として，事前にキーワードを指定せず，ユーザの習慣行動を使用し，実世界における事象を抽出する新たなソーシャルセンシング手法を提案した．本手法では，習慣行動を分析するためにユーザの投稿履歴を解析し，各投稿からユーザの行動情報の抽出と，生活習慣に関する単語の時間ごとの出現回数を示すベクトルを作成した．そして，1 年間の行動を平時の行動とし，1 ヶ月ごとの習慣行動と比較した．検証実験では，ソーシャルセンサの特性として「ユーザの非習慣行動時の投稿を解析することで社会事象を抽出可能であること」と「ユーザの非習慣行動時の社会事象のカテゴリや内容が変化すること」の 2 つの検証項目に対し，本提案手法のソーシャルセンシング技術が有用であることがわかった．検証結果より，平時の行動と 1 ヶ月ごとの習慣行動を比較することで，非習慣行動を抽出可能であることを確認した．また，その非習慣行動からトピックを抽出し，事前にキーワードを指定する既存手法と異なる内容でトピックを取得できることを確かめた．これによって，カテゴリに絞られない社会事象の抽出が可能となり，網羅的な分析ができるため，本技術の汎用性が明らかとなった．

次章では，ユーザ属性を考慮した社会事象の内容を分析するための方法について，詳述する．

第 4 章

ユーザ属性を考慮した平時と異なる事象に対するソーシャルセンシング技術の提案

第4章　ユーザの属性を考慮した平時と異なる事象に対するソーシャルセンシング技術の提案

4.1　まえがき

本章では，ユーザの属性ごとによる習慣行動の違いから社会事象を抽出することを目的として，マイクロブログユーザの属性を考慮した社会事象抽出手法について検討する．SNS のユーザは，性別，年齢，職業や居住している地域が異なるため，意見や反応も多様である．そこで，前述の方法にユーザ属性を考慮して社会事象を発見することができれば，属性ごとの重要な記事，具体的には，性別や年齢に依存したトピック記事や職業に関連するタイムリーな記事，そして地域特性を持った重要記事等，社会事象でも分析内容や情報の内容が異なるかを検証する．本手法では，マイクロブログの Twitter のプロフィール欄に記載されているユーザ自身のプロパティ情報を基にユーザ属性をマニュアルで付与する．

第 4.2 節では，研究の概要について論じている．第 4.3 節では，提案技術の有用性を検証するための検証項目に関して論じている．第 4.4 節では，各ユーザ属性の習慣行動による相違を用いたソーシャルセンシング技術に関してのアルゴリズムについて論じている．第 4.5 節では，評価実験について論じている．

4.2　研究の概要

4.2.1　研究の位置づけ

本研究では，インターネット上に蓄積されているビッグデータを対象としたソーシャルセンシング手法を適用する．本研究の位置づけを図 4.1 に示す．

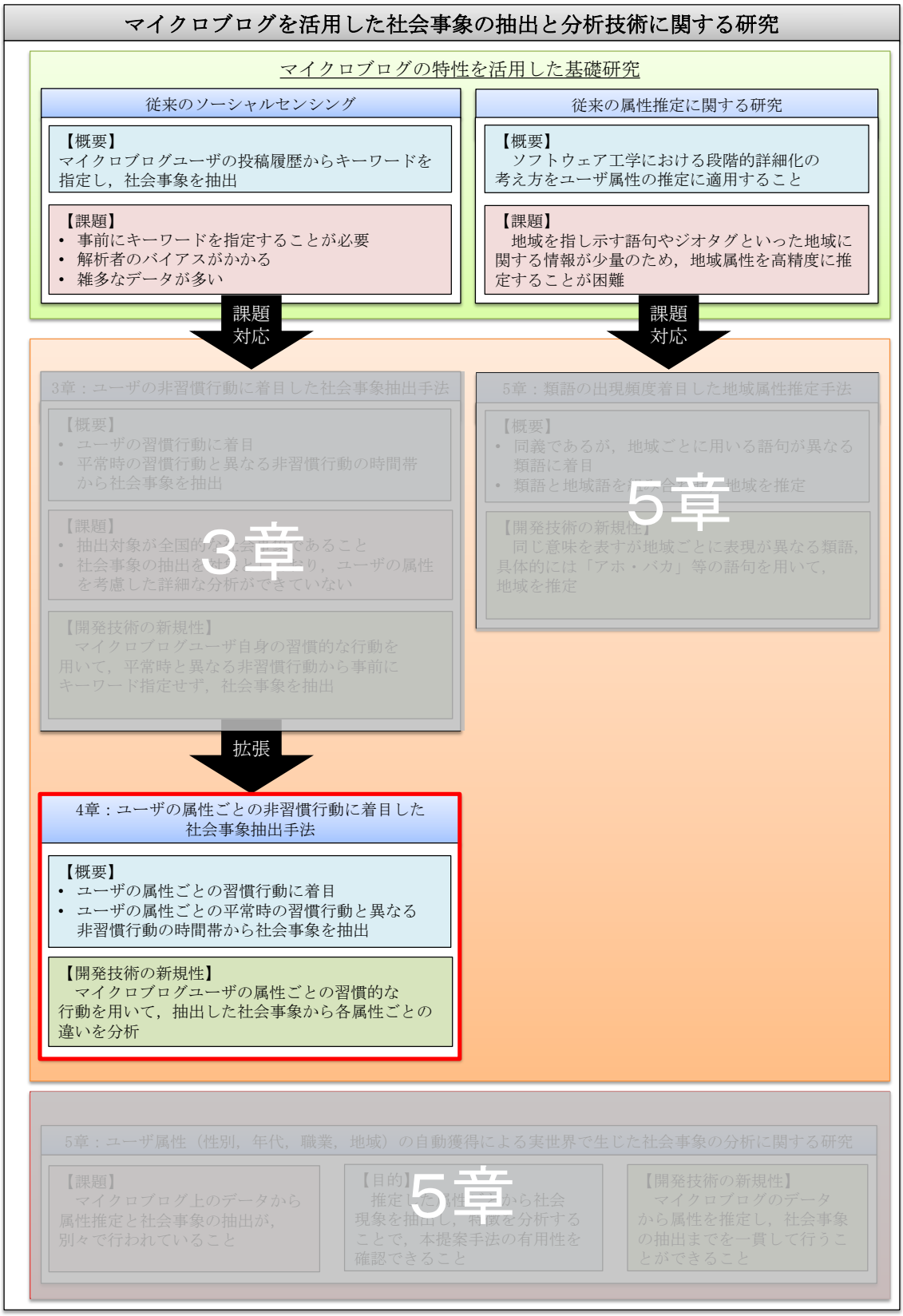


図 4.1　本研究の位置付け

従来のソーシャルセンシングでは，社会事象に関するキーワードや属性全体の情報や投稿内容を対象としている．そのため，指定したキーワードのみの事象や性別，年代といったユーザのパーソナル属性を考慮できずに投稿内容のみで抽出を行っている．これにより，解析者のバイアスがかかることや雑多なデータが多くなる課題がある．著者らが提案した既存手法[34]では，ユーザの習慣行動に着目し，平時の習慣行動の違いから社会事象を抽出する手法を提案した．その手法では，まず，ユーザの習慣行動を「起床・就寝」,「在宅」,「外出」と「帰宅」の 4 行動に分類し，一年間の行動を平時の習慣行動の基準とした．次に，各月の習慣行動を基準と比較して，閾値を超える場合に非習慣行動とした．そして，各非習慣行動の時間帯の投稿を Blei らによって提案された潜在的ディリクレ配分法[38]を用いて解析し，トピック（話題）を検出した．その結果，事前にキーワードを指定する手法で獲得できなかったものも含まれているトピックを取得できることが明らかとなり，提案手法の有用性を証明した．これらの研究により，ユーザ特性を用いたソーシャルセンシングの基盤を構築できた．しかし，トピックが全国的な社会事象のみで，ユーザの性別や職業などの属性を考慮した詳細な分析ができない課題があった．

そこで，本研究では，著者らがこれまで行ってきたユーザの属性や習慣行動を推定する研究[25]，[33]，[36]を応用し，各ユーザ属性の習慣行動に着目して，非習慣行動を抽出する手法[34]を適用する．ユーザの属性や習慣行動を推定する研究では，職業推定に関する手法[25]とユーザの習慣に基づく行動推定手法[36]を応用し, ユーザの性別, 年代と職業といった属性を段階的に推定する手法が提案されている．そして，その属性を考慮した行動推定手法[33]において，属性推定や習慣行動の予測に対する課題に有用であることを証明している. その解析方法を用いて, 既存研究[36]を深く追求することを目的とする. 方法としては，基準の習慣行動と各属性の各月の習慣行動を比較することで，各属性の非習慣行動を分析し，既存手法[36]よりも詳細なトピックを抽出することを目指す．これにより，同じオリンピックの話題であっても性別，職業（社会人，主婦など）と地域（北海道・東北地方，関東地方など）によって注目されるトピックが異なると考えられる．

4.3　検証項目の設定

本研究では，次に示す 2 つの検証項目を設定し，これらを明らかにすることで，提案手法のソーシャルセンサの有用性を確認する．

4.3.1　検証項目3：各ユーザ属性による習慣行動の相違

多くの研究では，各事象に合った特定のキーワードを事前に指定するため，多種多様な事象を広範囲に把握することが困難であった．これに対し，既存研究[34]では，複数のユーザの習慣行動から非習慣行動を取り出すことにより，キーワードに依存せず，実世界にお

ける事象を抽出する手法を提案した．これにより，キーワードの出現数や文脈に頼らないソーシャルセンシング手法[34]を開発し，その有用性を証明した．しかし，ユーザの属性を考慮せずに，社会事象を取得したため，注目度の大きいイベントや災害などの社会事象に関する話題に集中した．これは，ユーザ群の各属性を考慮して，社会事象を抽出することにより，改善できると考えられる．例えば，職業のユーザ属性に着目すると社会人は，朝外出し，夜に帰宅するといった一般的な社会人の特性と，夜に外出して朝に帰宅するなどの夜勤の行動特性がある．学生は，朝から授業があり，夕方はアルバイトを行っていることなどが予測できる．これにより，各属性によって，ニュースや情報収集する時間帯が異なり，社会事象に対して反応する時間帯が相違すると考えられる．

本研究では，検証項目として，「ソーシャルセンサの特性として，ユーザの属性ごとに習慣行動が相違し，非習慣行動の時間帯が異なること」を設定して，属性が既知のユーザから習慣行動を算出し，各属性がどの時間帯で非習慣行動となるかを把握し，検証する．

4.3.2　検証項目4：各ユーザ属性の抽出可能な社会事象の差異

本研究では，各ユーザ属性のトピック分析を行うことにより，属性によって反応する社会事象が変化することや同じ社会事象でも収集できる情報が詳細になることを実証実験から証明する．既存研究[34]では，起床・就寝時には，エンタメ，帰宅時には政治・経済といった傾向があることを確認した．

本研究では，検証項目として，「ユーザ属性ごとに反応や興味を持つ社会事象が異なり，同じ社会事象でも抽出内容が異なること」を設定し，習慣行動と収集したトピックの関係を分析することで，各属性の行動ごとに関心の高い社会事象のカテゴリ（生活，エンタメ，スポーツ，政治・経済など）が変化するかを検証する．また，各ユーザ属性の習慣行動を分析することで，トピックのキーワードが詳細に把握でき，社会事象が的確に抽出できることや取得できる内容自体が変化するかを立証する．

4.4　各ユーザ属性の習慣行動による相違を用いたソーシャルセンシング技術

ソーシャルセンシングとは，マイクロブログなどのソーシャルメディア上での利用者をソーシャルセンサとして捉え，実世界の事象を観測する方法である．ソーシャルセンサは，情報の拡散性と速度が速く，不特定多数のユーザが情報発信しているため抽出可能な事象の範囲が広いことが特徴である．また，解析データの取得が容易といった利便性がある．本研究では，その中でもリアルタイム性に優れており，投稿される情報量が多い Twitter を採用する．

本手法の処理フローを図 4.2 に示す．本手法は，既存研究[34]を参考に「ユーザの習慣行動の定量化機能」と「平時の習慣行動と異なる行動の検出機能」により構成される．

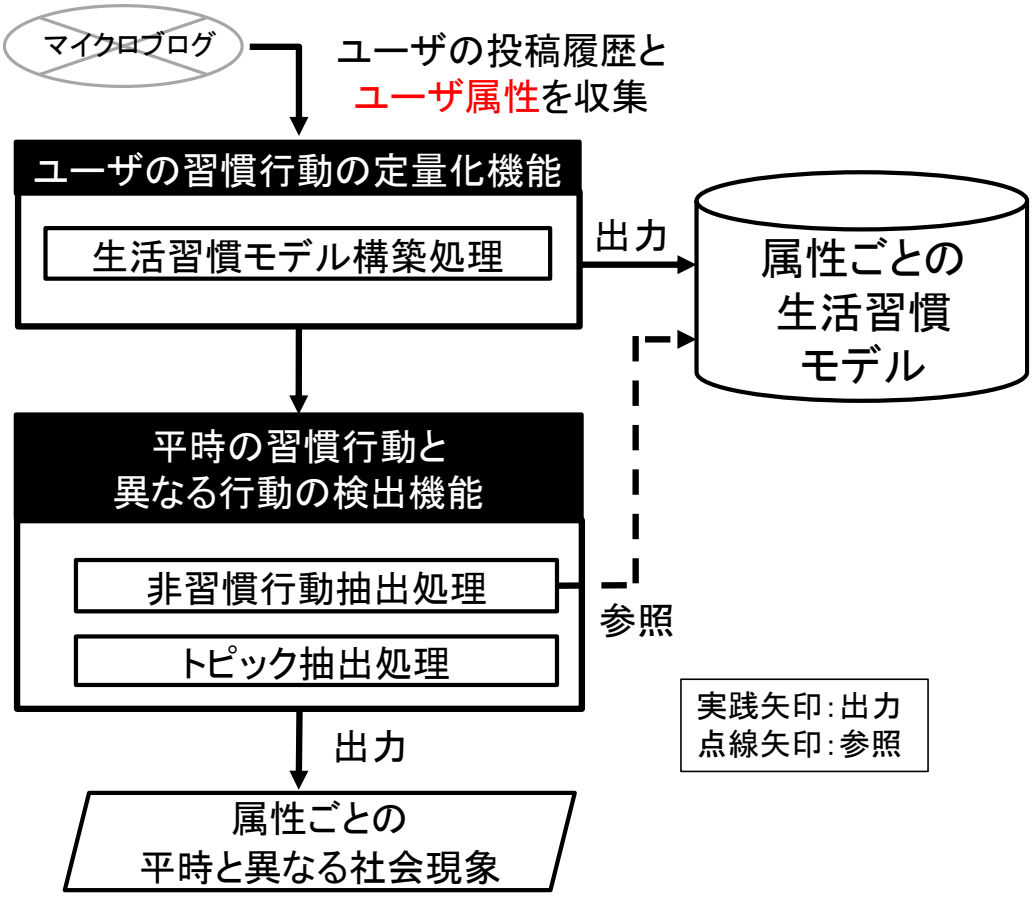


図 4.2　処理フロー

4.4.1　ユーザの習慣行動の定量化機能

本機能では，各ユーザ属性の非習慣行動を抽出するために用いる生活習慣ベクトルを算出する．生活習慣ベクトルは，各ユーザ属性の投稿傾向を分析し，平時の習慣行動を定量化して表現したものである．本研究では，1 年間の習慣行動を平時の習慣行動とし，比較対象として，1 ヶ月ごとの習慣行動を用いた．平時の習慣行動を 1 年間とした理由は，1 年間の習慣行動を平時の行動とすることで，1 年間に起きた社会事象を考慮した行動となるため，各年に特徴があると考えたからである．さらに，社会事象は，日付単位や週単位で変化する可能性があるが，各ユーザ属性の習慣行動と非習慣行動に着目することで，社会事象の内容に変化が生じるかを目的とした．なお，本論文では，各属性の習慣行動の違いを分析し，非習慣行動を用いて，既存手法[34]よりも詳細なトピックを抽出することを目的とする

ため，月単位で比較する．その比較すべき理由は，1 週間ごとにすると，何気ない突発的な行動の差が非習慣行動として分類される可能性があり，1 週間だけでは各ユーザ属性の習慣行動を取り出せないと考えたからである．加えて，ユーザ属性に応じて，平時の生活習慣と特定期間の生活習慣のタイムスパン（年，月，週，日）は変化し，取得できる社会事象も変化する可能性がある．しかし，本研究では，ユーザ属性に応じた平時の生活習慣と特定期間の生活習慣のタイムスパンの関係性が明らかになっていないため今後の議論とする．

(1)　行動抽出処理

本処理では，ユーザの投稿履歴から，既存研究[36]の行動抽出処理で用いた日本語語彙大系[37]を参考に行動辞書を作成し，行動情報を抽出する．そして，習慣行動に関する単語の各時間の出現回数を示す生活習慣ベクトルを作成する．まず，生活習慣ベクトル作成のため，習慣行動の選定を行う．習慣行動として，既存研究[25]，[33]，[36]を参考に「起床・就寝」,「在宅」,「外出」と「帰宅」の 4 種類を習慣行動として採用する．なお，生活習慣ベクトルを構成する素性には，あらかじめ構築した行動辞書に登録されている用語を使用する．行動辞書には，日本語語彙大系を参考にして，手作業で行動に関連する用語を習慣行動ごとに選定したものを登録する．行動辞書に登録した用語は，第 3 章と同様とした．

(2)　生活習慣ベクトル作成処理

本処理では，ユーザの習慣行動を素性とした 4 次元（習慣行動）7 次元（曜日）24 次元（時間帯）の 672 次元で構成される生活習慣ベクトルを作成する．本章での平時習慣ベクトルは，年間の投稿から作成した生活習慣ベクトルとし，特定習慣ベクトルは，月ごとの投稿から作成した生活習慣ベクトルを表す．ユーザ属性$attribute_j$の年間の行動beh_xにおける各曜日の時間帯 h の生活習慣ベクトル$\boldsymbol{YearPost}(\boldsymbol{attribute_j}, \boldsymbol{beh_x})$を式 4.1 に示す．

$$\boldsymbol{YearPost}(\boldsymbol{attribute_j}, \boldsymbol{beh_x}) = \{\boldsymbol{YPost}_{\boldsymbol{attribute_1},\boldsymbol{beh_1},0}, \boldsymbol{YPost}_{\boldsymbol{attribute_1},\boldsymbol{beh_1},1}, \cdots, \boldsymbol{YPost}_{\boldsymbol{attribute_j},\boldsymbol{beh_x},h}\} \quad \textbf{式 4.1}$$

式 4.1 において，h は，7 次元（曜日）×24 次元（時間帯）の 1 時間を表す．$h = 0$の場合，年間の日曜日の 0 時を示す．j は各属性を表しているため，性別，職業，年代と地域であり，最大は 16 である．x は，各習慣行動を表しているため，「起床・就寝」,「在宅」,「外出」と「帰宅」であり，最大は 4 である．$YPost_{attribute_j,beh_x,h}$において$h = 0$の場合，年間の日曜日の 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間にユーザ属性$attribute_j$が習慣行動beh_xに関連する単語を含む投稿がなされた回数を表す．$h = 167$の場合，土曜日の 23 時 00 分 00 秒から 23 時 59 分 59 秒までを示す．生活習慣ベクトルを式 4.1 により作成後，式 4.2 により習慣行動ごとに正規化した．

$$\boldsymbol{YearPost(attribute_j, beh_x)' = \frac{YearPost(attribute_j, beh_x) - MinYearPost(attribute_j, beh_x)}{MaxYearPost(attribute_j, beh_x) - MinYearPost(attribute_j, beh_x)}} \quad \textbf{式 4.2}$$

式 4.2 において，$MaxYearPost(attribute_j, beh_x)$は，日曜日から土曜日までの最大値の投稿数であり，$MinYearPost(attribute_j, beh_x)$は，日曜日から土曜日までの最小値の投稿数のことを指す．また，本研究では，同様に各月でも生活習慣ベクトルを算出する．各ユーザ属性$attribute_j$の各月の行動beh_xにおける各曜日の時間帯 h の生活習慣ベクトル$MonthPost(attribute_j, beh_x)$を式 4.3 に示す．

$$\boldsymbol{MonthPost(attribute_j, beh_x) = \{MPost_{1,attribute_1,beh_1,0}, MPost_{1,attribute_1,beh_1,1}, \dots, MPost_{m,attribute_j,beh_x,h}\}} \quad \textbf{式 4.3}$$

式 4.3 において，m は月を表し，$MPost_{1,attribute_j,beh_x,h}$の場合，1 月の日曜日の 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間にユーザ属性$attribute_j$が習慣行動beh_xに関連する単語を含む投稿がなされた回数を表す．式 4.3 により，生活習慣ベクトルを作成後，式 4.2 と同様に日曜日から土曜日までを算出し，習慣行動ごとに正規化を行った．

4.4.2 平時の習慣行動と異なる習慣行動の抽出機能

本処理では，任意の時間帯における平時習慣ベクトルと特定習慣ベクトルの値とを比較し，各ユーザ属性の時間帯ごとに差分を抽出する．各ユーザ属性の$attribute_j$がbeh_xに関する任意の時間帯hの平時習慣ベクトル$YPost_{attribute_j,beh_x,h}$と特定習慣ベクトル$MPost_{m,attribute_j,beh_x,h}$における差分 S は，式 4.4 にて算出する．

$$\boldsymbol{S(m,h) = \frac{\sum_{k=h-\alpha}^{h+\alpha} |YPost_{attribute_j,beh_x,k} - MPost_{m,attribute_j,beh_x,k}|}{2\alpha+1}} \quad \textbf{式 4.4}$$

式 4.4 において，αは求める時間hの前後の時間数を示す．本研究では，αの値を第 3 章と同様に 3 時間と設定した．前後の時間を考慮する理由としては，短時間の異常投稿による影響を少なくするためである．このことから，任意の時間hにおける差分は，合計した時間数$2\alpha+1$で除算することで平均値とした．なお，差分が，マイナス値にならないよう絶対値で計算を行う．

本研究では，算出した$S(m,h)$が一定の閾値を超える場合を非習慣行動として定義する．非習慣行動の判断式を式 4.5 に示す．

$$S(m,h) \geqq \frac{\sum_{k=0}^{167} S(m,k)}{168} \qquad \text{式 4.5}$$

式 4.5 において，非習慣行動と判断する時の閾値には，全ての時間帯における平時習慣ベクトルと特定習慣ベクトルの差分の平均値を用いる．一定の投稿パターンで行われている平時の習慣行動との差分が大きい箇所は，何らかの社会事象が発生していると考えられることから，習慣行動ごとに式 4.5 により判定することで非習慣行動を抽出する．

(1)　トピック抽出処理

本処理では，既存研究[34]と同様に Blei らによって提案された潜在的ディリクレ配分法[38]を用いて，トピックを抽出する．そして，非習慣時から算出した特定の期間と月全体のトピックを比較し，社会事象を抽出する．本研究では，Python のトピックモデルライブラリである gensim[39]を用いて，トピックを分析する．

4.5　検証実験

4.5.1　実験概要

本実験では，4.3 節で設定した検証項目 3 と検証項目 4 の 2 つの検証項目に対し，本提案のソーシャルセンシング技術が有用であることを検証する．

実験では，検証項目を実証するため，平時習慣ベクトルと各ユーザ属性の特定習慣ベクトルを基に非習慣行動の時間帯を算出する．そして，各ユーザの属性の行動ごとに非習慣行動の時間数が異なることを示し，検証項目 3 を証明する．また，非習慣行動の時間帯の期間の全投稿のトピック，ユーザ属性を指定した非習慣行動の時間帯に行動した習慣行動のトピックを比較し，検証項目 4 を立証する．そして，ユーザ属性ごとに異なる社会事象を取得可能であるかを検証し，提案手法の有効性を証明する．

(1)　実験データ

本実験では，2014 年 1 月～12 月の投稿群を対象とする．実験データの収集方法を次に示す．

STEP1： TwitterAPI[41]を用いて 2014 年 1 月～12 月に投稿しているユーザおよび投稿内容を取得する．

STEP2： Twilog を解析して，2014 年 1 月～12 月に投稿しているユーザを収集し，その投稿内容を取得する．Twilog から取得したユーザが STEP1 のユーザと重複している場合は，Twilog のデータを優先して採用する．理由としては，TwitterAPI では，投稿内

容の取得数制限のため，最大で 3,200 件であるのに対して，Twilog はユーザの全投稿を抽出できるため，投稿数が多いからである．

STEP3： STEP1 および STEP2 で収集した投稿内容の件数が 1,000 件以上のユーザを実験データとする．ユーザにより投稿数に差が出る場合でも習慣行動は取得可能であるため，実験に支障はないものと考えられる．ただし，習慣行動の解析に最低限必要な 1 週間分の投稿内容を取得できないユーザは，実験データから除外する．

STEP4： STEP3 で収集したユーザのプロフィールを確認し，ユーザの属性を人手で付与する．

以上の手順にて収集した実験データの詳細は，ユーザ数が 1,514 ユーザ，対象となる投稿件数は，1,195,237 件である．各属性のユーザ数の構成を表 4.1 に示す．

表 4.1　実験データ

	社会人		学生		主婦	フリーター	
	男性	女性	男性	女性	女性	男性	女性
10 代	0	0	198	102	0	3	11
20 代	52	7	0	101	68	91	42
30 代	85	0	0	0	109	64	25
40 代以上	92	0	0	0	85	27	7
北海道・東北	15	10	13	6	34	14	14
関東	126	34	83	58	175	93	38
中部	39	14	37	16	60	26	13
近畿	43	6	32	1	69	24	16
中国・四国	19	1	14	4	24	11	2
九州・沖縄	15	1	12	8	30	17	4

本実験では，非習慣行動算出に関わるパラメータαとトピック抽出で用いる潜在的ディリクレ配分法[38]でトピック数を設定する．

(2)　パラメータ α

パラメータαは，非習慣行動抽出処理において，平時習慣ベクトルと特定習慣ベクトルの差分を算出する時に用いる値である．式 4.4 において，hは求める対象の時間帯，パラメータαはhを算出する際に考慮する前後の時間を示す．αの値が大きくなると，個々の時間の情報を読み取ることが難しくなる．

本実験では，第 3 章と同様に，1 日のサイクルは 0 時～6 時，6 時～12 時，12 時～18 時と 18 時～24 時の 6 時間ごととし，αの値を 3 に設定した．

(3)　トピック数

最適なトピック数を算出するため，評価指標として Coherence（トピックの品質）[46]と Perplexity（予測性能）[47]を用いた．Coherence と Perplexity のグラフを図 4.3 と図 4.4 に示す．

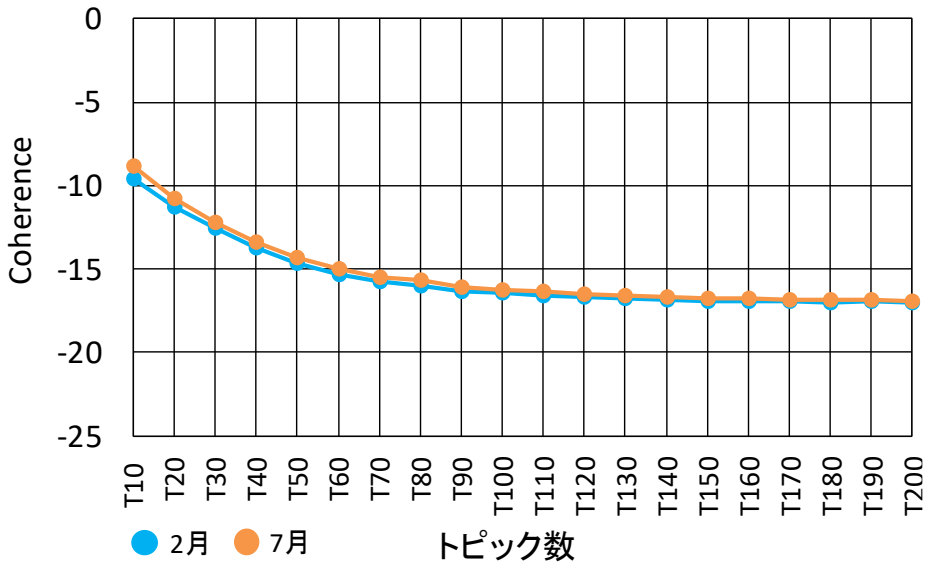


図 4.3　評価指標 Coherence

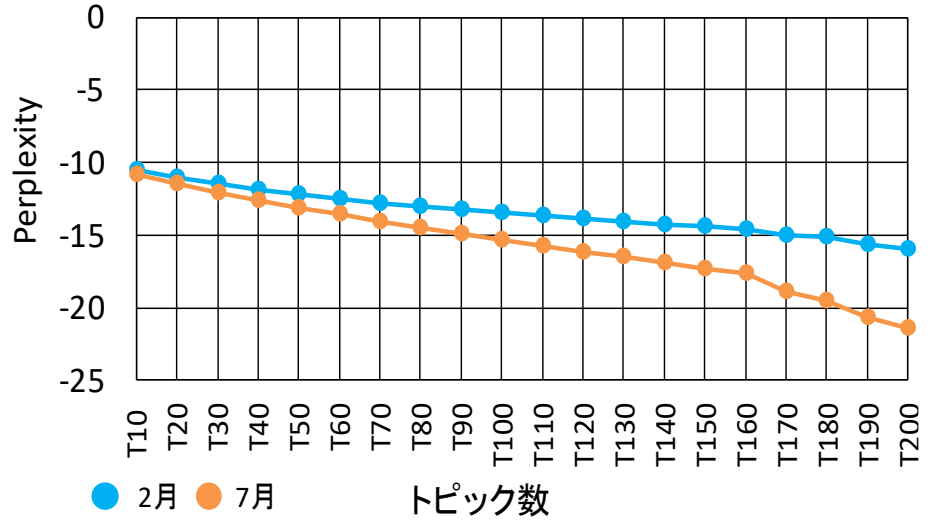


図 4.4　評価指標 Perplexity

縦軸は，Coherence と Perplexity の値を示し，横軸はトピック数である．図 4.3 より，Coherence がトピック数 100 から 160 を境に横ばいになっていることがわかる．図 4.4 の Perplexity のグラフは 7 月に関しては，下がり続けているが，既存研究[48]においても最適

なトピック数は 90 から 140 とされている．そのため，本研究では，潜在的ディリクレ配分法の手法[38]で用いるトピック数は，100 トピックとし，トピック推定の反復回数は 50 回とした．なお，習慣行動によって投稿件数が変化するため，全体の投稿件数を 100 トピックとして，各行動の投稿件数を割合でトピック数を算出し，パラメータとして設定した．例えば，全体の投稿件数が 10,000 件，各行動の投稿件数が 1,000 件の場合，全体では 100 トピックに対し，各行動では 10 トピックとした．これにより，全体と各行動の 1 トピックに対する投稿件数の比率が同様となるため，同様の詳細度でトピックを抽出できる．

4.5.2 実験1：非習慣行動の抽出実験

(1) 実験内容

本実験では，検証項目 3 を検証するため，2014 年 2 月と 7 月の投稿内容を対象に，各ユーザ属性の非習慣行動を算出し，習慣行動が相違していることを示して，立証する．ユーザの属性は，性別，職業（社会人，学生，主婦，フリーター）と地域（北海道・東北地方，関東地方，中部地方，近畿地方，中国・四国地方，九州地方・沖縄県）を対象とする．

本実験では，次の手順に従って分析を実施する．

STEP1： 平時のユーザの習慣行動を明らかにするため，2014 年 1 月～12 月の全投稿群を対象に平時習慣ベクトルを作成し，習慣行動ごとに分類する．

STEP2： 2014 年の 2 月と 7 月で特定習慣ベクトルを作成し，習慣行動ごとに分類する．

STEP3： STEP1 で作成した平時習慣ベクトルと STEP2 で作成した 2 月と 7 月の特定習慣ベクトルを比較し，非習慣行動が発生した時間帯を抽出する．

以上の手順で抽出した各属性の非習慣行動の時間を表 4.2 から表 4.5 に示す．表 4.2 と表 4.3 は，各月の習慣行動の非習慣行動の時間数を属性ごとにクロス集計し，表 4.4 と表 4.5 は地域ごとにクロス集計した値であり，各属性の非習慣行動の時間数を定量的に示した値である．ここでは，一つの属性に着目し，同時間中に他の属性が非習慣行動を行っていない時間帯のことを「重複無し」，一方，同時間中に他の属性が非習慣行動を行った時間がある場合を「重複あり」として扱う．また，非習慣行動の抽出時間の上位と下位の平時習慣ベクトルと特定習慣ベクトルの比較結果を図 4.5 に示す．

表 4.2　2 月における非習慣行動の抽出時間数

分類	行動内容	非習慣行動合計	非習慣行動重複無し	非習慣行動重複あり／分類					
				社会人	主婦	フリーター	学生	男性	女性
社会人	起床・就寝	76	7		49	40	—	60	50
	在宅	91	3		56	55	—	74	69
	外出・帰宅	159	18		97	38	33	115	87
主婦	起床・就寝	74	2	60		41	—	55	67
	在宅	69	3	56		43	—	48	54
	外出・帰宅	153	6	97		48	41	103	112
フリーター	起床・就寝	79	16	40	41		—	44	47
	在宅	85	20	55	43		—	52	50
	外出・帰宅	80	6	38	48		—	36	52
学生	起床・就寝	—	—	—	—	—		—	—
	在宅	—	—	—	—	—		—	—
	外出・帰宅	77	26	33	41	—		44	34
男性	起床・就寝	73	0	60	55	44	—		60
	在宅	80	1	74	48	52	—		62
	外出・帰宅	149	1	115	103	36	44		89
女性	起床・就寝	80	4	50	67	47	—	60	
	在宅	79	1	69	54	50	—	62	
	外出・帰宅	141	3	87	112	52	34	89	

表 4.3　7 月における非習慣行動の抽出時間数

分類	行動内容	非習慣行動合計	非習慣行動重複無し	非習慣行動重複あり／分類					
				社会人	主婦	フリーター	学生	男性	女性
社会人	起床・就寝	69	4		46	24	44	55	56
	在宅	71	3		52	38	35	52	57
	外出・帰宅	153	8		76	84	86	90	94
主婦	起床・就寝	74	2	46		32	35	41	65
	在宅	72	0	52		38	38	44	62
	外出・帰宅	146	6	101		60	94	78	98
フリーター	起床・就寝	79	11	24	32		39	28	37
	在宅	78	5	38	38		48	42	34
	外出・帰宅	149	10	84	60		80	81	83
学生	起床・就寝	70	4	44	35	39		40	40
	在宅	78	6	35	38	48		39	42
	外出・帰宅	166	27	86	89	80		76	85
男性	起床・就寝	68	0	55	41	28	40		58
	在宅	68	4	52	44	42	39		47
	外出・帰宅	135	3	104	78	81	76		81
女性	起床・就寝	85	1	56	65	37	40	58	
	在宅	80	6	57	62	34	42	47	
	外出・帰宅	141	2	101	98	83	85	81	

表 4.4　2 月における地域属性ごとの非習慣行動の抽出時間数

分類	行動内容	非習慣行動合計	非習慣行動 重複無し	非習慣行動重複あり／分類					
				北海道・東北地方	関東地方	中部地方	近畿地方	中国・四国地方	九州地方・沖縄県
北海道・東北地方	起床・就寝	—	—		—	—	—	—	—
	在宅	—	—		—	—	—	—	—
	外出・帰宅	155	27		95	51	128	81	77
関東地方	起床・就寝	80	12	—		—	55	55	63
	在宅	81	13	—		—	55	63	56
	外出・帰宅	152	14	95		51	113	54	46
中部地方	起床・就寝	—	—	—	—		—	—	—
	在宅	—	—	—	—		—	—	—
	外出・帰宅	79	5	51	51		59	—	—
近畿地方	起床・就寝	68	0	—	55	—		58	62
	在宅	80	4	—	55	—		64	66
	外出・帰宅	158	2	99	113	59		74	57
中国・四国地方	起床・就寝	75	8	—	55	—	58		62
	在宅	77	0	—	63	—	64		63
	外出・帰宅	81	0	55	54	—	74		55
九州地方・沖縄県	起床・就寝	89	13	—	63	—	62	62	
	在宅	76	1	—	55	—	66	63	
	外出・帰宅	77	11	45	46	—	57	55	

表 4.5　7 月における地域属性ごとの非習慣行動の抽出時間数

分類	行動内容	非習慣行動合計	非習慣行動 重複無し	非習慣行動重複あり／分類					
				北海道・東北地方	関東地方	中部地方	近畿地方	中国・四国地方	九州地方・沖縄県
北海道・東北地方	起床・就寝	75	3		36	45	55	51	58
	在宅	76	9		45	53	51	55	51
	外出・帰宅	143	12		76	105	102	98	88
関東地方	起床・就寝	76	16	36		48	35	34	48
	在宅	68	3	45		47	51	50	56
	外出・帰宅	142	16	76		86	87	87	95
中部地方	起床・就寝	82	2	45	48		50	51	57
	在宅	71	3	53	47		49	53	53
	外出・帰宅	150	8	105	86		103	106	91
近畿地方	起床・就寝	73	0	55	35	50		65	57
	在宅	73	4	51	51	49		57	57
	外出・帰宅	142	1	102	87	103		122	97
中国・四国地方	起床・就寝	71	0	51	34	51	65		54
	在宅	72	0	55	50	53	57		53
	外出・帰宅	143	3	98	87	106	122		106
九州地方・沖縄県	起床・就寝	85	0	58	48	57	57	54	
	在宅	73	3	51	56	53	57	53	
	外出・帰宅	145	12	88	95	91	97	106	

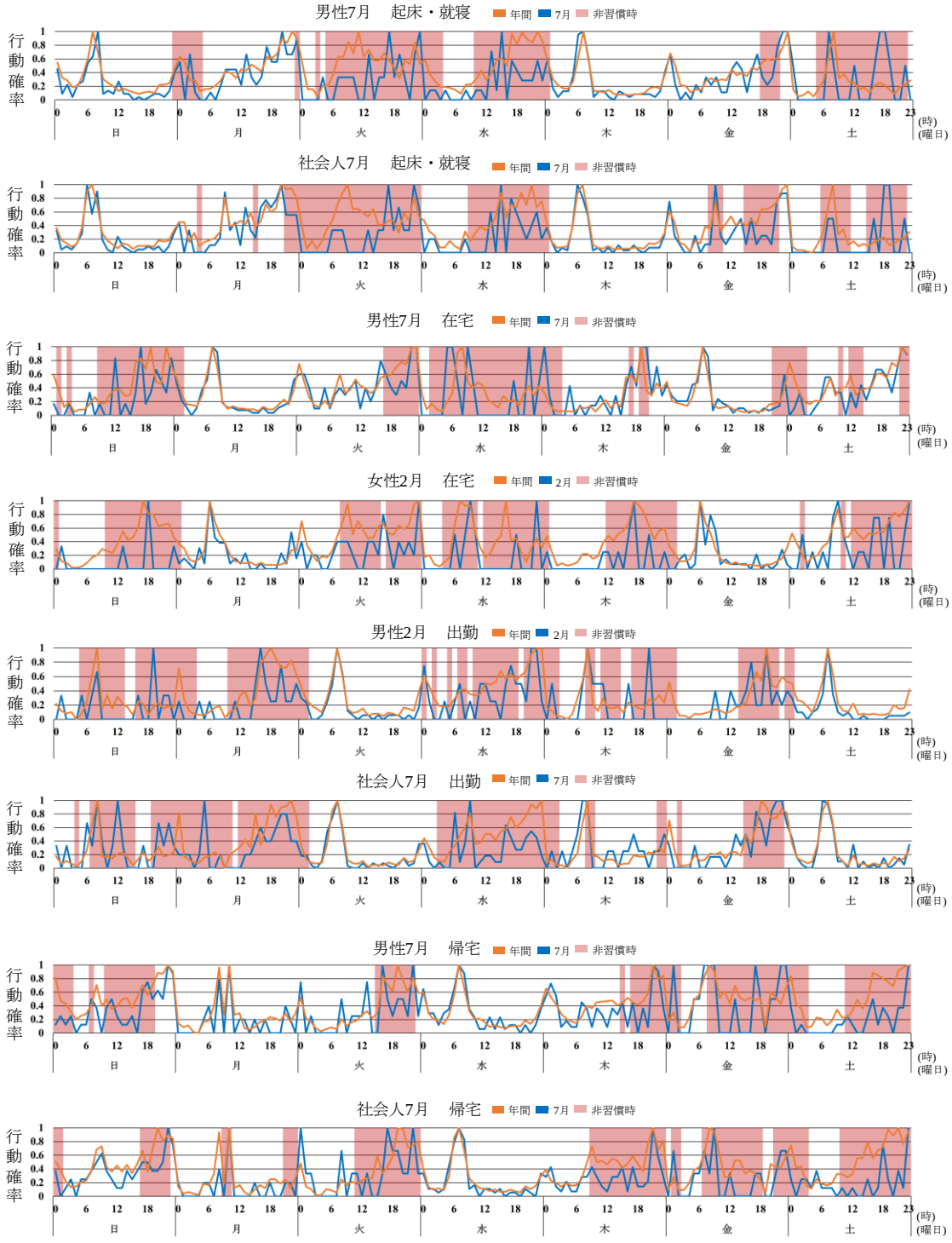


図 4.5　上位と下位の習慣行動解析結果

(2)　実験結果

表 4.2 から表 4.5，図 4.5 より，次に示すことがわかった．

- **各属性の非習慣行動の時間数が異なることがわかった．**

表 4.2 と表 4.3 より，非習慣行動重複無しの時間が 2 月は平均 7.3 時間，7 月は平均 5.6 時間であり，各属性で非習慣行動の時間が異なった．その詳細を確認すると，学生の外出・帰宅の行動の非習慣行動の時間帯が最も多いことがわかった．これは，非習慣行動の時間帯を外出と帰宅の行動で一緒にした要因も推察できるが，中学生と高校生，大学生では授業やアルバイトなどの関係によって，外出や帰宅の時間帯が変化することが原因と考えられる．また，フリーターの起床・就寝の行動に関しても朝勤や夜勤のユーザが同等に存在するからであると推測できる．

表 4.4 と表 4.5 より，非習慣行動の時間数の重複無しの時間数は,，関東地方で多くなっており，他の地方では全体的に少なくなっていることがわかった．これは，関東地方には，首都圏があり，人口密集率も多く多様な生活習慣があることから他の地方と比べ多くなったと考えられる．

- **同様の習慣行動でも属性によって，非習慣行動の時間帯が相違することがわかった．**

図 4.5 より，同じ習慣行動においても属性が異なると，抽出できる非習慣行動の時間帯が相違することがわかった．例えば，起床・就寝の投稿では，男性で抽出されていない金曜日の 7 時から 10 時は，社会人では非習慣行動として算出されている．他にも同様の時間帯も存在するものの各属性によって，非習慣行動の時間が明確に相違することがわかった．

以上より，属性ごとに非習慣行動が異なることを証明でき，非習慣行動の時間帯が異なるため，取得できる投稿内容も変化する．これにより，検証項目 3 を立証できたと考えられる．

4.5.3　実験2：各ユーザの属性における非習慣行動に着目したトピック抽出実験

(1)　実験内容

本実験では，検証項目 4 を検証するため，非習慣時と算出された期間の投稿をトピック分類した結果に基づき検討する．本実験では，実験 1 で抽出した非習慣行動の時間帯を基に 2 月と 7 月を対象に属性ごとに解析する．

実験では，次の手順に従って分析を実施する．

STEP1：2 月と 7 月における非習慣行動とその期間内の投稿を抽出し，それらの投稿を各月の習慣行動ごとに区分する．

STEP2：各ユーザ属性の 2 月と 7 月の全投稿を収集し, それらの投稿を月ごとに選別する.

STEP3：STEP1 と STEP2 で整理した投稿を既存手法のトピックと提案手法のトピックとして分類する.

STEP4：各月の既存手法と提案手法のトピックから，それぞれトピック内の単語を取得し，対応する期間にニュースやイベントがあるかを検索して該当トピックをノイズとして扱う.

STEP5：STEP4 で取得した各月の既存手法と提案手法のトピック内の単語が 5 単語以上一致する場合は，3 章と同様に同一トピックとして取り扱う.

以上の処理手順でトピックを分類する.

(2) 実験結果

既存研究[34]の非習慣行動の時間帯のみの投稿内容から抽出した社会事象に関するトピックの結果を表 4.6 に示す.

表 4.6　既存手法[34]の社会事象に関する代表的なトピックの抽出結果

月	既存手法	
	行動	解釈トピック（キーワード）
2月	起床・就寝	・ソチオリンピックを都知事が応援（都知事，応援，オリンピック，ソチ） ・安倍政権が調査（安倍，経済，政府，調査）
	在宅	・ソチオリンピック高校生や大学生の活躍 （大学，高校生，オリンピック，発見，ソチ） ・静岡県久しぶりの除雪（久しぶり，31，除雪，記録，静岡，配信，視聴） ・千葉で去年の倍の雪（心配，千葉，大雪，NHK，ぶり，去年，倍）
	外出・帰宅	・浅田真央の活躍（選手，オリンピック，感動，埼玉，彼女，五，ソチ） ・舛添氏が都知事選に関わった（都知事，東京，舛添） ・原発で問題発生（原発，NHK，問題，職員） ・横浜と千葉で地震発生（横浜，発生，千葉，地震）
7月	起床・就寝	・野々村議員号泣（反対，野々村，説明，号泣，選挙，行為） ・原発に関するニュース（原発，ニュース，経済，研究，避難）
	在宅	・安倍政権が集団的自衛権の行使の閣議決定 （集団，本日，安倍，政権，閣議，米国），（自衛，権，集団，的）， （集団，行使，閣議，決定） ・宮崎で台風により大雨（台風，影響，大雨，発生，宮崎） ・ワールドカップのアルゼンチンの試合が NHK で TV 放送 （放送，NHK，アルゼンチン，ワールドカップ，TV） ・安倍政権が地方支援を検討（安倍，政権，支援，地方，メンバー）
	外出・帰宅	・平和記念ドラマ放送（ドラマ，平和，７０，ぶり）

また，既存研究[34]と提案手法で，取得できたトピック数を性別と職業は表 4.7，地域ごとは表 4.8 に示す.

表 4.7　既存手法[34]と各ユーザ属性の抽出トピック数

月	既存手法			提案手法					
	行動	抽出トピック数		抽出トピック数					
				社会人	主婦	フリーター	学生	男性	女性
2月	起床・就寝	抽出数	48	45	47	48	－	45	53
		不一致数	43	44	45	46	－	4	52
	在宅	抽出数	55	59	48	55	－	52	20
		不一致数	50	56	46	54	－	49	19
	外出・帰宅	抽出数	100	102	97	52	46	100	33
		不一致数	67	93	88	48	40	88	24
7月	起床・就寝	抽出数	42	43	45	48	46	45	53
		不一致数	37	41	43	46	42	42	50
	在宅	抽出数	53	50	45	43	54	46	54
		不一致数	49	47	44	41	52	42	53
	外出・帰宅	抽出数	103	99	101	94	108	99	88
		不一致数	91	95	97	90	104	96	84

表 4.8　既存手法[34]と地域属性ごとの抽出トピック数

月	既存手法			提案手法					
	行動	抽出トピック数		抽出トピック数					
				北海道・東北地方	関東地方	中部地方	近畿地方	中国・四国地方	九州地方・沖縄県
2月	起床・就寝	抽出数	48	－	52	－	45	49	56
		不一致数	42	－	50	－	44	47	55
	在宅	抽出数	55	－	53	－	53	52	51
		不一致数	50	－	51	－	50	50	50
	外出・帰宅	抽出数	100	97	99	49	108	53	50
		不一致数	72	89	93	45	99	51	45
7月	起床・就寝	抽出数	42	47	52	48	55	43	53
		不一致数	38	44	50	45	52	40	52
	在宅	抽出数	53	47	46	46	50	47	51
		不一致数	49	45	44	44	47	45	49
	外出・帰宅	抽出数	103	86	95	103	88	92	103
		不一致数	89	82	87	97	83	88	100

表 4.7 と表 4.8 は，既存手法[34]で，取得したトピックに対し，性別と職業ごと，地域ごとの非習慣行動の時間帯を考慮して，取り出したトピックの一致，不一致数を定量的に示している．不一致数は，表 4.6 で示したトピック抽出内容と一致していない内容のトピック数を指す．これにより，既存手法と比較して新たなトピックを抽出できていることがわかる．表 4.9 と表 4.10 は，提案手法において，不一致の代表的なトピックをまとめた一覧である．なお，下線のトピックは，既存手法と同様のトピックを示す．表 4.7 と表 4.8 より，各属性で重複しない多くのトピックを抽出できたことがわかった．表 4.7 から表 4.8 より，次に示すことが明らかになった．

- **既存手法と提案手法ともに異なるトピックが抽出できることがわかった．**

表 4.7 と表 4.8 ともに多くのトピックで，トピックが一致しないことがわかった．これは，各属性のトピックの内容が異なり，同様の内容でも興味を示す内容が違うため，トピック内の単語が一致しないからだと考えられる．トピックを詳細に把握するため，提案手法の不一致の代表的なトピックをまとめた表 4.9 と表 4.10 より，次に示すことが明らかとなった．

表 4.9　各ユーザ属性の代表的なトピック抽出結果

月	提案手法						
	行動	代表的なトピック内容					
		社会人	主婦	フリーター	学生	男性	女性
2月	起床・就寝	食事の献立 （今日，食事，献立，味）	恵方巻を旦那と一緒に食べるために買い物 （旦那，一緒，巻き，お腹，帰り，恵方，スーパー，通り）	オリンピック応援に関する内容 （ソチ，オリンピック，修造） 雪の影響により早めの帰宅をした男性ユーザ （雪，帰宅，関東，僕，早め） 毎日昼頃にアルバイトを行うユーザ （今日，毎日，昼，バイト）	－	東京にて大雪に見舞われるニュースと家族の予定 （雪，ニュース，東京，幼稚園，家族，遊び，旅，温泉，延期，明後日） 母親の献立に関する投稿 （食事，献立，ママ，具）	実家での生活を楽しむ女性 （献立，幸せ，ママ，実家，パパ，TV）
	在宅	関東で行われるライブ （昨日，夜，たくさん，ライブ，埼玉，購入，準備，栃木，チケット，ベスト）	子供と車周辺の除雪作業を行う様子 （車，子，運転，除雪） バレンタインを家族で過ごす様子 （チョコ，家族，実家，バレンタイン，ダンナ）	関東地方での大雪 （雪，大変，東京，関東） インフルエンザで病院に行くユーザ （病院，大事，インフルエンザ，年齢）	－	雪かきを行う様子 （自分，家，雪かき，異常，被害） スーパーへ買い物に向かう親子 （車，先，親，薬，スーパー）	ユーザの子供の部活に関する内容 （娘，子供，大変，会，吹奏楽，本，トランペット）
	外出・帰宅	オリンピック出場の上村愛子の活躍 （雪，上村，愛子，準備，夜中） インフルエンザを危惧するユーザ （ホント，病院，インフルエンザ，白湯）	大雪被害に関する報道 （日本，東京，大雪，心配，電話，NHK，交通） 北上温泉を楽しむユーザ （北上，風呂，新幹線，エステ）	自宅でオリンピック応援 （ごはん，オリンピック，感情，酒，スキー，部屋，パソコン） 久々にライブを楽しむユーザ （ライブ，期間，久しぶり，大人，最高）	アルバイト出勤前の学生 （明日，バイト，出勤，朝）	浅田真央の応援 （ちゃん，真央，感動，浅田） 占い結果を確認するユーザ （姜，仕事，運，金，健康，運勢）	ユーザの子供の誕生日 （日，今日，子供，誕生）
7月	起床・就寝	エアコン掃除に関する話題 （設定，エアコン，掃除，対策）	夏休みのため，家事が滅った主婦 （夏休み，幸せ，弁当，高校）	動画サイトでサッカー観戦 （YouTube，2525，サッカー，ツアー，頻度，機能，パソコン）	部活動の野球の試合について （市，試合，今朝，野球）	花火大会に関するニュース （ニュース，花火，本日，大会）	子供の学校の怖い話 （息子，君，学校，花子，女子）
	在宅	ユーザの息子と花火大会 （子供，息子，花火，大会，影響）	梅雨明けでエアコンを起動し，快適に過ごす様子 （エアコン，梅雨，明け，快適）	ネットを楽しむユーザ （わたし，sumo，YouTube，ブログ）	アルバイトを考慮する学生 （店，今年，失敗，バイト，久々，近所，経験，夕方）	地震により被災した地域 （時間，分，震度，地震，県，大変）	花火大会に関する内容 （花火，大会，電車，帰宅，準備，予想）
	外出・帰宅	松坂桃李出演のドラマ （くん，楽しみ，ドラマ，今度，桃李，撮影，共演）	サッカーの試合観戦 （最後，試合，サッカー，初日，上達） 東京に台風が迫る状況 （台風，東京，空，準備）	勉強ではチョコレートが良い （勉強，楽，チョコ，脳）	食事を 1 人で済ます学生 （ごはん，ひとり，食事，状況，コンビニ）	渋谷の大雨，雷に遭うユーザの内容 （周り，雷，渋谷，大雨） 花火を楽しみにするユーザ （楽しみ，花火，福，天気）	子供の熱中症を考慮する様子 （娘，雨，子供，症，熱中，塩，頻度）

表 4.10　地域属性ごとの代表的なトピック抽出結果

月	提案手法						
	行動	代表的なトピック内容					
		北海道・東北地方	関東地方	中部地方	近畿地方	中国・四国地方	九州地方・沖縄県
2 月	起床・就寝	―	東京都知事選挙 （岡田，職員，選挙，舛添，会長，団体，議員） 節分を楽しむユーザ （恵方，電話，巻き，豆）	―	浅田真央に関する話題 （真央，ソチ，オリンピック，五輪，浅田，アスリート） スーパーでカレーを購入するユーザ （カレー，スーパー，屋，残り，味噌汁）	家族で生活する様子 （みんな，一緒，旦那，たくさん，幸せ，息子，電話	昨晩にかけて実家が積雪 （昨日，夜，実家，ビックリ，自然，積雪，深夜，センチ） ラーメンファンによる投稿 （気持ち，腹．ファン，ラーメン，追加，ごちそう）
	在宅	―	ソチオリンピックに関する話題 （雪，2014，今年，生，ソチ，修造，松岡） 夕方にご飯を考慮するユーザ （味噌汁，夕方，生姜，焼き，サラダ，節約）	―	雪によって出勤に影響 （雪，明日，影響，出勤，状況） インフルエンザによる影響 （影響，出勤，インフル，発生，状況，クラス）	大学受験におけるリスニングの話題 （リスニング，どっち，記憶，大学）	大雪により休み （休み，大雪，今秋，雪かき，予報） インフルエンザによる影響を受けるユーザ （昨日，朝，目，インフルエンザ，最悪）
	外出・帰宅	―	夕飯を考慮するユーザ （ごはん，スーパー，子供，鍋，すき，満足）	インフルエンザを危惧するユーザ （元気，子供，病院，インフルエンザ）	ソチオリンピックに関する話題 （ソチ，気温，修造，松岡，高梨） NHK の織田信長に関する番組を視聴するユーザ （織田，信長，放送，NHK，詳細）	羽生選手を応援 （選手，LOVE，羽生，ファン，インタビュー，スケート，ラブ） ビジネス投資を考慮するユーザ （円，ビジネス，お金，億，生活，カード，投資，券）	フィギュアスケート選手の話題 （選手，羽生，高橋） 停電が発生した話題 （大雪，予想，停電，苦手）
7 月	起床・就寝	録画していた番組を見終えたユーザ （終了，番組，録画，昼，アイス，実家，シーン）	主婦が家で過ごす様子 （ニュース，ご飯，生活，旦那，梅雨明け）	大阪にて梅雨明け （大阪，エアコン，梅雨，今朝，明け，快適）	地元のスーパーで買い物するユーザ （外，買い物，スーパー，地元，駿馬，自宅）	ユーザの子供の昼ご飯に関する内容 （娘，ご飯，お昼，昼，実家，我が家）	録画作品に熱中するユーザ （帰宅，録画，作品，熱中，酒）
	在宅	地震の備えについて （自身，震度，水，実家，速報，食）	地震の報道 （地震，放送，震度，防災，速報）	大阪での花火大会に関する話題 （花火，大阪，大会，最終）	次の週にゴジラが放映 （ゴジラ，夏休み，ネット，動画，来週）	昼間に地震発生 （地震，速報，昼，震度，緊急）	ユーザの子供が大阪から帰宅 （明日，息子，大阪，帰宅）
	外出・帰宅	占いに関する話題 （運，勤，健康，仕事，恋愛，運勢，勉強）	仕事前ニュースを確認するユーザ （今日，仕事，ニュース，朝，事件，疲れ，詳細）	バイトを終えたユーザ （お金，終了，バイト，体調）	子供の夏休みの予定 （子供，今回，夏休み，花火，元気，テレビ，大会）	ユーザの娘が花火大会に行く （今日，娘，元気，花火，大会，レベル，光）	仕事が終わり帰宅した子持ちユーザ （息子，会社，帰宅，エアコン）

- **提案手法ではより多くのトピックを抽出していることがわかった.**

既存手法[34]で抽出されている社会事象（ソチオリンピックや大雪被害, サッカー, 台風）に関するトピックを，提案手法においても抽出できていることがわかった．提案手法で新たに出現しているトピックを確認すると, ユーザの詳細がわかるような日々の生活に直結する行動スケジュールの内容が確認できた.

具体的には，主婦は子供を心配しているなど，子供との生活を示す内容が多く，男性は外出と帰宅時に大雪に困っているなどの個人的なことに関連した社会事象が把握できた.

- **行動ごとに各属性で抽出できるトピックが異なることがわかった.**

代表的なトピックを確認すると，ソチオリンピックや大雪，地震などに関するトピックを取り出せることがわかった．ソチオリンピックに関するトピックで社会人や主婦は，外出と帰宅時の行動で算出できているのに対し，フリーターは起床・就寝時に抽出できていることがわかった．これは，朝外出して，夜帰宅するといった規則的な生活を送る社会人は，深夜に行われているオリンピックをリアルタイムに視聴することが困難であるためだと考えられる．そのため，外出や帰宅時に試合結果や状況を確認するユーザが多く，主婦も相手の生活リズムに合わせる傾向があることから，このような結果になったと推察できる．性別，地域や月ごとの特徴的な内容を次に示す.

男性と女性では，男性は雪や地震といったトピックが抽出できていることがわかり，一方，女性では，子供や献立といった家庭内の出来事に関するトピックが確認できる．これは，男性は仕事に影響を与える天候などの情報を確認している傾向があると考えられる．また，女性は主婦のユーザも多いことから，家庭状況を気にかけている様子が多く抽出されている傾向がある.

地域属性を考慮した場合の代表的なトピックを確認すると，職業属性を考慮した場合と同様に，ソチオリンピックや大雪，地震に関するトピックを見出せることがわかった.

2 月は，大雪に関するトピックも多く，近畿地方で，在宅時の行動として，雪によって外出に影響したと推考できる内容が取り出せることがわかった．また，九州地方・沖縄県では，起床・就寝時の行動で，ユーザの実家で積雪したことに関するトピックを確認できた．また，在宅時の行動では，大雪の影響を受けて，休みにするといったトピックも抽出されていることがわかった．これは，当時，近畿地方や九州地方・沖縄県地域の九州地方で大雪に見舞われ，積雪などにより外出等に影響を与えたからだと考えられる.

7 月は，地震に関するトピックが多いことがわかった．在宅時の行動で，地震に関するトピックが出現している傾向があることから，当時，この時間帯に地震が発生したと推測できる．また，主に外出と帰宅時の行動で，花火大会に関するトピックが多く出現していることがわかった．これは，花火大会が夜に開催されるため，帰宅の時間帯で多く投稿される傾向があることによると考えられる.

- **地方に関連したトピックを抽出していることがわかった.**

7 月における中部地方にて, 起床・就寝時の行動で, 大阪にて梅雨明けが発表されたトピックが取得でき，在宅時の行動で，大阪で開催される花火大会に関するトピックが抽出されていることがわかった．これは，中部地方と近畿地方が近しい気候であることから，大阪で梅雨明けした情報にユーザが反応した可能性がある．また，花火大会のトピックを抽出できた理由として，大阪にて天神祭りや淀川花火大会といった全国的に有名な花火大会を示すことからこのような結果になったと考えられる.

同様の方法で，ユーザの年代を考慮した社会事象の抽出結果を表 4.11 に示す.

表 4.11　年代属性ごとの代表的なトピック抽出結果

月	提案手法				
	行動	代表的なトピック内容			
		10 代	20 代	30 代	40 代以上
2 月	起床・就寝	ソチオリンピックに関する話題 （選手，羽生，冬，天才，毎回）	ソチオリンピックに関する話題 （羽生，中盤，一番，今回，弦） 旅行に関する話題 （昨日，旅行，インド，友達，夜行）	大雪の規模を表す話題 （倍，過去，大雪）	除雪に関する話題 （車，不足，最近，除雪，雪，警察） 雪に関する話題 （気温，久しぶり，大阪，週末，雪）
	在宅	ソチオリンピックに関する話題 （記事，キムヨナ，金メダル） ソチオリンピックに関心があるユーザ （世界，メイン，スキー，今度）	女子フィギュアに関する話題 （フィギュア，女子，練習，上手） バレンタインに関する話題 （リア充，予定，バレンタイン）	ソチオリンピックに関する話題 （楽しみ，五輪，応援，メダル） 大雪による被害を懸念しているユーザ （被害，状況，雪，情報，大雪）	積雪により家周辺の状況を危惧するユーザ （雪，家，屋根，サポート，禁止） 体を温めるホットドリンクに関する投稿 （ごま，きなこ，豆乳，ローズヒップティー，生姜，珈琲）
	外出・帰宅	アイドルグループに関する話題 （生田，彼女，本気，乃木坂，ツイン，テール，綺麗）	献立を考慮するユーザ （更新，ブログ，ごはん，ぶり）	ライブを楽しむユーザ （楽しみ，待ち，女性，ライブ，今日）	ソチオリンピックに関する話題 （ビール，五輪，大切，勝利） インフルエンザを危惧するユーザ （時期，危険，インフル，ホンマ）
7 月	起床・就寝	運勢を確認するユーザ （運，わたし，金，今日，仕事，運勢，恋愛，健康）	休日の予定を考えるユーザ （予定，心配，何，休暇，金曜日，相談）	地震速報アプリに関しての話題 （ドコモ，地震，速報，アプリ，信州）	飲食イベントに関する話題 （店，写真，カフェ，イベント，飲食，料理）
	在宅	無料 iOS アプリに関する話題 （アプリ，iOS，無料，App）	花火大会に関する話題 （花火，大会，最高，オススメ，隅田川）	夏祭りの感想を語るユーザ （夏，祭り，2014，年，感想）	ふるさと納税に関する話題 （国，納税，ふるさと，住民）
	外出・帰宅	電車内でゲームを楽しんでいるユーザ （ライブ，ラブ，電車，アイマス，内，スクフェス）	アプリ開発に関する話題 （iPhone，アプリ，制作，受注，Web，Android，可能，即座）	原発に関する話題 （経済，アベノミクス，中国，爆弾，原発，韓国）	台風接近を懸念するユーザ （夜，心配，台風，帰り，夕方）

以上の結果からユーザ属性を考慮し，非習慣行動を算出することで，詳細なトピックを見出すことができる．これにより，検証項目 4 を立証できた．

4.6 あとがき

本章では，ユーザ属性ごとに情報を抽出することで，それらの違いを把握できる可能性を秘めていると考え，ユーザ属性を考慮した社会事象の分析を検討した．方法としては，基準となる習慣行動とユーザ属性ごとの月単位の習慣行動を比較することで，ユーザ属性ごとの非習慣行動を分析した．これにより，ユーザ属性を考慮せずに社会事象を抽出した際，注目度の大きいイベントや災害等の全国的な社会事象に関する話題に集中する課題に対応できた．検証実験の結果より，前述の研究では得ることができなかったユーザ毎の属性を考慮することで，細部に亘った感度の高い社会事象を獲得することを実証した．また，日々の生活に直結する行動パターンの内容を収集できることから，ユーザ属性ごとのトピックに差異があることも確認できた．例えば，同じオリンピックの話題であっても性別，職業（社会人や主婦など）と地域（北海道・東北地方や関東地方など）によって注目するトピックが異なることを示すことができた．

したがって，前述の大局的なデータセンシングから本提案の局所的なデータセンシングの可能性を立証するもので，属性ごとの社会事象の抽出に貢献した．ただし，性別，年代，職業などの属性にはパターン性があるが，地域性に関しては投稿記事の中身を詳細に分析する必要があることがわかった．したがって，行動パターンから読み取れない属性に関しては，一段と掘り下げた詳細な分析方法が必要となることが，判明した．

次章では，類語に着目した地域属性を推定するための方法について，詳述する．

第 4 章　ユーザの属性を考慮した平時と異なる事象に対するソーシャルセンシング技術の提案

第 5 章

マイクロブログユーザの類語に着目した地域属性の推定に関する技術の提案

第5章　マイクロブログユーザの類語に着目した地域属性推定に関する技術の提案

5.1　まえがき

本章では，投稿内容からユーザの属性を推定する技術に着目する．本研究では，段階的詳細化の考え方を基に，性別，年代，職業，地域からなるユーザ属性を複数の段階に分けて推定する手法を提案し，属性推定や習慣行動の予測に有用であることを確認した．しかし，ユーザの年代や職業属性は一定精度で推定できるが，地域属性に関しては習慣行動や投稿傾向に顕著な特徴がみられないという課題がある．そこで，同じ意味を表すが地域ごとに表現が異なる類語，具体的には「アホ・バカ」等の語句に着目し，語句ごとの地域性の違いから地域属性を推定する手法を提案する．

第 5.2 節では，研究の概要について論じている．第 5.3 節では，提案技術の有用性を検証するための検証項目に関して論じている．第 5.4 節では，類語に着目した地域属性の推定技術に関してのアルゴリズムについて論じている．第 5.5 節では，評価実験について論じている．

5.2　研究の概要

5.2.1　本研究の位置付け

本研究では，インターネット上に蓄積されているビッグデータを対象としたソーシャルセンシング手法を適用する．ソーシャルセンシングに関する既存研究[8]-[31]，[33]と本研究との位置づけを図 5.1 に示す．

また，本研究は，マイクロブログ上に日々蓄積されているビッグデータを対象とし，平時と非習慣時のソーシャルセンシングにおいて，課題となっているユーザの地域属性に関する推定手法の実現を目的とする．既存研究[33]-[35]と本研究の関係性を図 5.2 に示す．

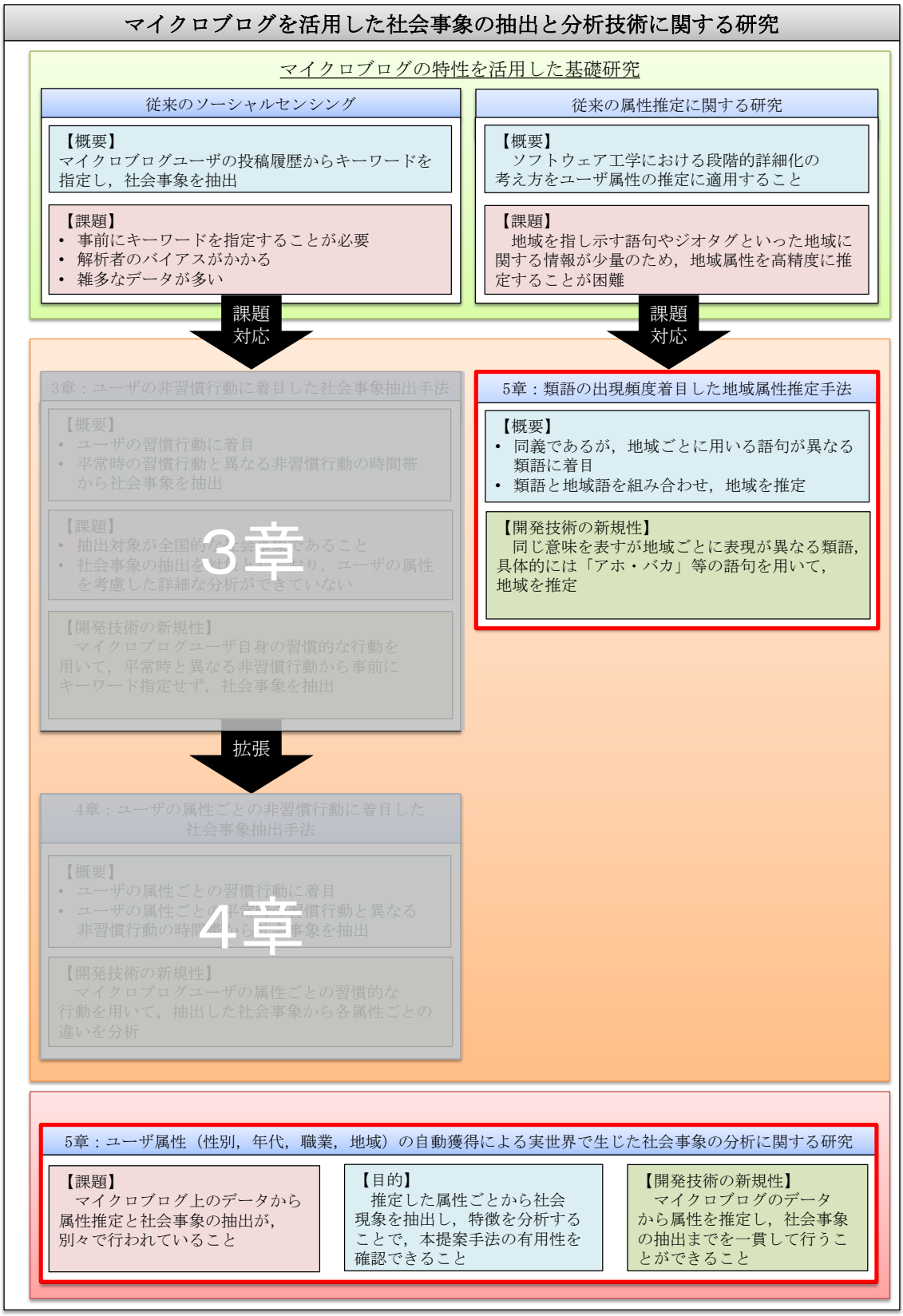


図 5.1　本研究の位置付け

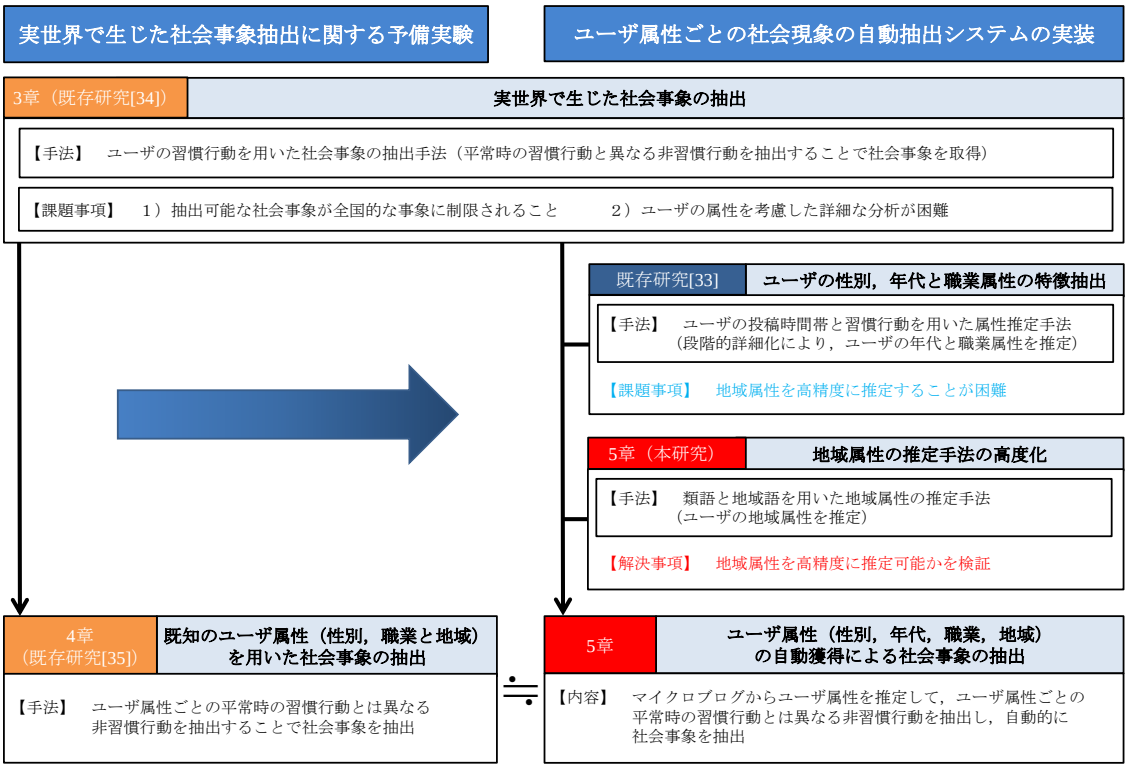


図 5.2 既存研究[33]-[35]と本研究の関係性

著者らは，実世界におけるユーザ属性ごとの社会事象の自動抽出を目指し，マイクロブログユーザの属性を推定する技術[33]と社会事象を抽出する技術[34]を提案してきた．

属性推定に関する研究では，ユーザ属性ごとに推定精度が異なることに着目し，ソフトウェア工学における段階的詳細化に関する考え方を適用した．そして，ユーザの性別，年代と職業といった属性を段階的に推定する手法[33]を提案し，属性推定や習慣行動の予測に対する課題に有用であることを確認した．しかし，この手法では，ユーザの年代や職業属性は約 5 割から 8 割程度で推定できるが，一部の学生やパート・アルバイト，地域といった性別ごとの顕著な特徴がみられない属性の推定が難しいことがわかった．また，段階的詳細化手法の特性上，学習データを各属性に分類するとデータ件数が少なくなり，的確な推定モデルを構築できないことが明らかとなった．

社会事象を抽出する技術[34]では，ユーザの投稿履歴から暗黙的に含まれる平時の生活習慣を活用した習慣行動に着目し，平時と異なる非習慣行動を抽出することで社会事象を取得する手法を提案した．そして，基準となる習慣行動と各月の習慣行動を比較し，取得した非習慣行動を分析することで，従来の事前にキーワードを指定した手法よりも詳細な社会事象が抽出できることを証明した．また，ソーシャルセンシングによる詳細な社会事象の抽出には，性別や職業，地域といった各ユーザ属性の習慣行動と非習慣行動の考慮が必須であることを示唆した．そこで，既存手法[34]を実践的に評価するため，ユーザ属性を考

慮して非習慣行動を抽出し，社会事象の取得[35]を試みた．その結果，ユーザ属性を考慮した手法[35]は，既存手法[34]よりも詳細なトピックを抽出できることを明らかにし，ユーザ属性の推定技術が必要であるとの結論を得た．

これらの研究を経て，著者らが行ってきたマイクロブログユーザの属性を推定する技術[33]が必要不可欠であることがわかり，ユーザ属性の自動獲得による社会事象の抽出に大きく寄与することができた．こうした研究経緯を踏まえて，本研究では，ユーザ属性ごとの社会事象の自動抽出技術の実現に向けた地域属性の高精度な推定手法を提案する．本技術の実現により，各ユーザ属性の社会事象，平時と非習慣時のソーシャルセンシングの高度化に寄与する．

ユーザの属性推定に関する既存研究[25]-[27]には，主にユーザの投稿履歴から暗黙的に含まれる生活習慣を用いた研究[25]，ユーザが既読した新聞記事とそのタイトルの内容を用いた研究[26]，SNS 上の顔画像を用いた研究[27]がある．これらの手法で性別や年代，職業に関しては，習慣行動，顔画像や属性ごとの特徴となる単語から各属性の特徴を予測可能であるが，地域属性の推定は難しい．

ユーザの地域属性を推定する従来手法[28]-[31]，[49]と本手法の関係を図 5.3 に示す．

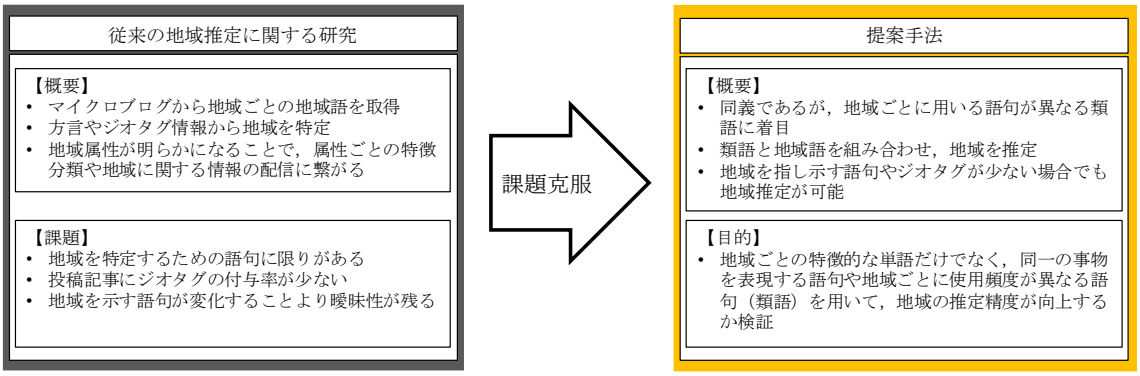


図 5.3　本研究の目的

ユーザの地域属性を推定する既存研究では，地域語を用いた研究[28]-[31]が主に行われている．また，地域に関連する情報として，投稿に付与されているジオタグの情報を使用し，地域属性を推定する研究[31]や移動履歴を推定する手法[49]が主に行われている．これによって，地域語で曖昧であった地域の分類が明確となり，地域ごとの特徴分類や情報の配信に繋がる．しかし，従来手法[28]-[31]，[49]では，次に示す 2 つの課題が考えられる．

1 つ目の課題は，地域語に限りがある点である．約 4,600,000 ユーザを対象に行った伊藤らの調査[50]によると，地域語の記述率は，約 25%と少なく，観光地名に関しては，地域に住むユーザが投稿しているとは限らないのが現状である．また，SNS 上でのプライバシーの問題[51]から，ユーザの個人情報の一種である地域語も減少傾向になると予測される．

2 つ目の課題は，ジオタグなどの地域情報を付与しているユーザは少ない点である．既存研究の調査[33]でもマイクロブログに位置情報の付与率は，全体の 0.42%となっている．また，代表的なマイクロブログである Twitter では，投稿時に位置情報を一緒に共有できる位置情報のタグ付け機能の削除や見直しを公表[52]-[54]しており，今後は取得がより困難になることが予想される．

著者らは，これら 2 つの課題点を解決するため，類語に着目し，「類語は地域ごとに使用頻度が異なり，各類語に地域性が出現する可能性がある」という仮説を立て，地域を推定する研究[55]を行った．その結果，地域ごとに類語の出現頻度が異なることを明らかにし，既存手法[55]の有用性を示した. しかし, 分析対象とした類語群も 6 組のみであることから，地域推定結果の汎用性と信頼性を向上させるには，類語の種類を加増し，地域語と組み合わせる必要があることがわかった．

そこで, 本研究では, 既存研究[55]を拡張し, 地域語とは異なる類語を加増する. そして, Twitter から各地域のユーザを自動的に収集し，そのユーザ群から算出した地域辞書を用いて，逐次学習することで自動拡張した学習モデルを更新する新たな手法を提案する．本研究で用いる類語の定義は，同一の事物を表現する語句や地域ごとに使用頻度が異なる語句である．また，各地域における特徴を示す地域語と類語を組み合わせることで，ユーザの地域推定の精度向上を目指す．

5.3 検証項目の設定

本研究では，次に示す 2 つの検証項目を設定し，提案した手法の有用性を確認する．

5.3.1 検証項目5：地域辞書を用いた推定精度

著者らが行った既存研究[55]では，1 組の類語の出現頻度を比較することで地域性を見出した．例として，大阪に位置するユニバーサル・スタジオ・ジャパンの略称である「USJ」と「ユニバ」の場合,「ユニバ」という語句は近畿地方を中心に出現頻度が高いことがわかった．このことから,「ユニバ」と「USJ」の類語から，近畿地方とそれ以外の地方を分類できることを確認した．同様に，愚かな者を表現する「アホ」,「バカ」も同様の結果を得た．

本研究では，検証項目として,「地域辞書を用いることで，推定精度が向上すること」を設定し, 既存研究[55]を参考に Web ページや書籍等から選定した 250 語の類語リストを用いて，その出現頻度を分析し，ユーザの地域属性を分類する．これにより，本仮説を証明する．

類語の選定方法として，既存研究[55]を参考にして，ユーザが無意識に使用する傾向がある同義語や地域によって抽出数が異なると考えられる語句を選定した．本研究で選定した類語リストは，表 5.1 に示すとおり,「しんどい」,「だるい」や「きつい」などの地域によっ

て呼び方が異なる語句である．また，「絆創膏」，「カットバン」や「バンドエイド」においても同様に，正式名称は絆創膏であるが，地域によっては「カットバン」や「バンドエイド」と呼ぶ地域が存在する．これらの語句は，ユーザの地域によって使用頻度が異なり，一部の地域のみで使用される傾向がある．このことから，意味が類似する語句は，地域によって使用頻度が異なるため，地域性が発現すると考えられる．

表 5.1 類語リストの登録内容（抜粋）

類語（25 語）
（アホ，バカ），（マック，マクド），（USJ，ユニバ），（コーヒー，紅茶），（しんどい，だるい，きつい），（絆創膏，カットバン，バンドエイド），（ラムネ，サイダー），（セブンイレブン，ローソン，ファミリーマート，サークル K サンクス，ミニストップ），（回転焼き，今川焼き，大判焼き，御座候）

5.3.2　検証項目6：洗練された地域辞書を用いた推定精度

地域語を用いた従来の手法では，取集した時期に応じて地域語を選定し，地域属性を推定する必要があるため，時期によって移り変わる地域語に対応することが難しい．また，地域属性を判断するための地域語が増加しない課題があり，取得時期に応じて，地域語が変化し，ユーザの推定精度に差が生じる可能性がある．

そこで，本研究では，検証項目として「収集したユーザの地域語を取得し，そのユーザの地域語を増加させることで，地域辞書が洗練され推定精度が向上すること」を設定し，マイクロブログのプロフィール欄に地域名を記載しているユーザを自動的に加増し，取得した地域語と推定に用いた学習モデルに使用した地域語の特徴を再計算して，学習モデルを自動的に拡張する．これによって，従来の地域語と新たに加増した地域語の地域性を再計算することにより，従来の地域語と新たな地域語にも対応可能である．また，今まで推定が困難であった地域のユーザの推定が可能になると考えられる．さらに，地域を分類するための地域語が増加することにより，地域性をより明確化することができるため，推定精度が向上すると考えられる．これを検証するため，地域を分類する語句を自動的に拡張して構築したモデルを使用し，ユーザの地域属性を推定して，本仮説を検証する．

5.4　類語に着目した地域属性の推定技術

本研究では，地域ごとに分類したユーザの投稿履歴から地域語の抽出を試みる．その際，類語リストを参照することで，地域語と類語の両語句における地域性を考慮した地域辞書

モデルを構築する．そして，地域辞書モデルを使用し，投稿内容からユーザの地域属性を推定する．

本手法の処理フローを図 5.4 に示す．

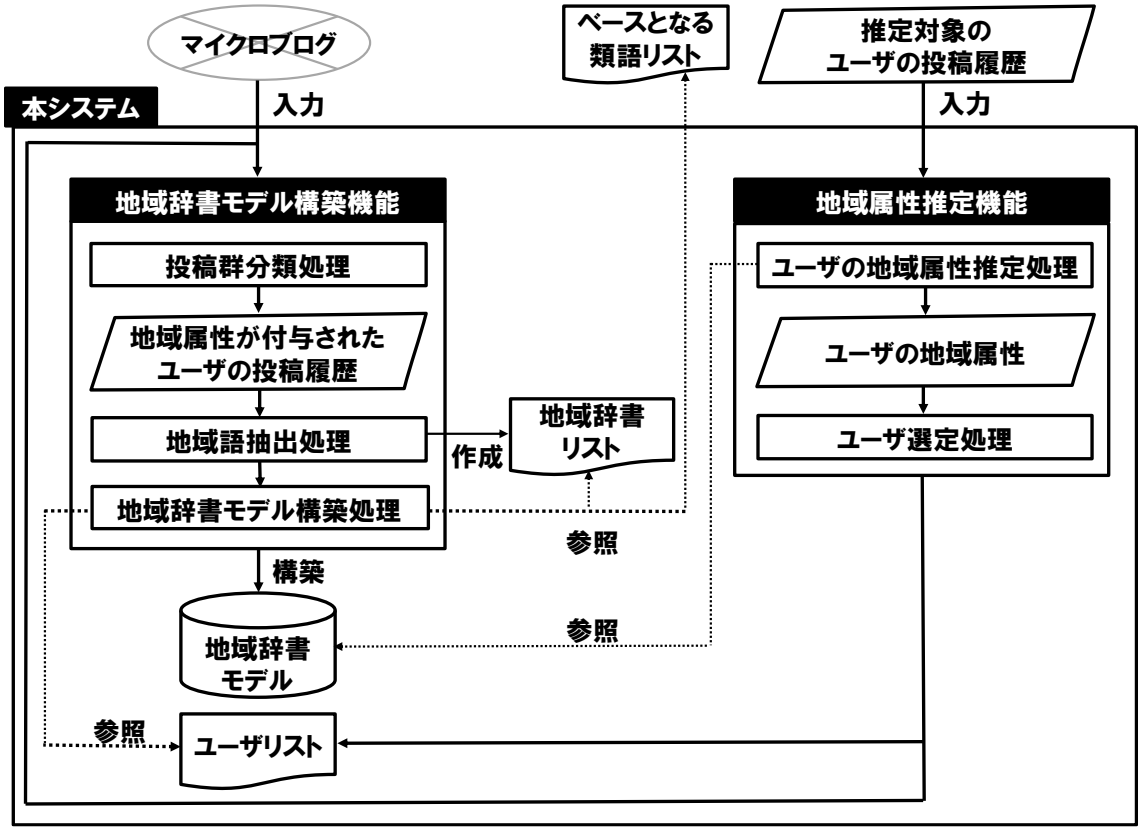


図 5.4　処理フロー

本手法は，「地域辞書モデル構築機能」と「地域属性推定機能」により構成される．

地域辞書モデル構築機能では、入力はマイクロブログユーザの投稿内容と類語リストとし，出力は，地域辞書から学習した各地域の投稿内容を地域辞書モデルとして構築する．

地域属性推定機能では、推定対象のユーザの投稿履歴を入力とし，出力はユーザの地域属性とする．また，この出力したユーザの地域属性の推定結果が既存の地域辞書モデルよりも高精度であった場合，学習に使用したユーザをユーザリストに格納する．これにより，単語を追加し，地域辞書モデルを漸次拡張する．

5.4.1　地域語辞書モデル構築機能

本機能では，所属している地域が明確なユーザの投稿履歴を基に地域辞書の地域性を評価し，地域辞書モデルを構築する．

(1)　投稿群分類処理

本処理では，マイクロブログユーザの投稿履歴を各地域に分類し，形態素解析した品詞から記号以外の品詞を分かち書きする．分類対象の地域は，既存研究[30]のプロフィールの粒度のヒアリング結果に基づき，北海道地方，東北地方，関東地方，中部地方，近畿地方，中国地方，四国地方と九州地方・沖縄県の 8 地域とする．

(2)　地域語抽出処理

本処理では，投稿群分類処理で 8 地域に分類した投稿履歴を用いて，各地域における地域特性が高い語句を地域語として抽出する．地域語の抽出過程を次に示す.

STEP1：分類した各地域の投稿履歴から投稿内容に出現する単語を抽出する．取得する単語は, 顔文字や平仮名 1 字といった語句や形容動詞語幹などの単語を除去した名詞を対象とする．

STEP2：取り出した名詞を評価するため，既存研究[25]を参考に，χ^2 値検定により地域語候補を各地域で取得する．χ^2 値検定によって取得する単語数は，類語の数と同様にするため，上位 250 語を対象とした．地域語の数によって精度が変化すると考えられるが，拡張に応じて単語の地域性が洗練され，単語数も増加するため問題ないと考えられる．

STEP3：取得した地域語候補と地域辞書リストの内容を比較し，重複する語句を除去して，各地域の地域語を抽出する．

以上の処理手順により，各地域の地域語を抽出する．そして，抽出した地域語とベースとなる類語リストを組み合わせて，地域辞書を作成する．

(3)　地域辞書モデル構築処理

本処理では，地域語抽出処理にて取得した地域辞書を用いて，地域ごとに学習し，地域辞書モデルを構築する．まず，地域語抽出処理にて，作成した地域辞書リストから参照した地域辞書を使用している各地域の投稿内容を抽出する．そして，取得した単語を投稿ごとに分かち書きし，単語区切りとなった投稿内容を深層学習の RNN（Recurrent Neural Network）の一種である Bi-LSTM（Bi-directional Long Short-Term Memory）[56]を用いて，学習する．これを 8 地域分全てで，計 8 モデル構築する．本研究では，Bi-LSTM の隠れ層の数は 64 としているが，隠れ層の数によって推定精度が変化する可能性がある．しかし，今回は，単語区切りとなった投稿を学習することで，地域属性が推定できるかの検証が主目的のため，デフォルトの設定である 64 層を用いた．

5.4.2　地域属性推定機能

本機能では，地域辞書モデル構築機能にて構築した地域辞書モデルを参照して，ユーザの地域属性推定処理により，ユーザの地域属性を推定する．

(1)　ユーザの地域属性推定処理

本処理では，まず，推定対象のユーザの投稿履歴を入力とし，形態素解析を行って，単語を取り出す．次に，構築した地域辞書モデルを参照し，ユーザの各投稿内容の地域を 8 モデル全てで推定を行う．そして，各地域のモデルで算出した結果から最も高い地域を推定対象ユーザの地域として推定する．

(2)　ユーザの選定処理

本処理では，地域属性推定機能によって得られた推定結果が拡張前のモデルの結果と比較し，使用するユーザを選定する．ユーザの選定過程を次に示す．

STEP1： 地域属性推定機能により，推定した結果を拡張前のモデルと比較する．

STEP2： STEP1 で比較した結果，精度が低下した場合，該当のユーザ群を加増して構築したモデルとユーザを除去し，再度，別のユーザ群で再学習する．STEP1 で比較した結果，精度が向上した場合，該当のユーザ群を加増して，作成したモデルを新たな地域辞書モデルとして使用する．また，推定精度が変化しない場合，使用したユーザ群は今後，モデルの汎用性を確かめるためのユーザとして保管する．

以上の手順で取得したユーザ群を地域辞書モデル構築機能に再度，入力することで，地域辞書モデルに入力する内容を漸次拡張する．また，再度，地域語抽出処理を行う際は，拡張前の χ^2 値を参考に，使用した各地域の地域語の最も低い値の平均値以上かつ上位 250 語までの地域語を使用する．これにより，地域を分類するユーザを加増して，自動的に拡張することが可能である．

5.5　検証実験

5.5.1　実験概要

提案手法の有用性を検証するため，図 5.5 の実験計画に示すとおり，2 つの評価実験を実施する．また，実験計画とは別に本研究のシステムの有用性を検証する実験を行う．

実験 1 では，各地域によって地域語と地域辞書のそれぞれで学習し，推定結果を比較する．この実験により，検証項目 5 を確認する．

実験 2 では，実験 1 の推定結果を基にモデル構築に使用するユーザを選定し，地域辞書モデルを拡張する．そして，拡張した地域辞書モデル（以下，地域辞書モデル（2 回目））を使用し，推定精度が向上するかを検証する．この実験により，検証項目 6 を確認する．

実験 3 では，自動で推定したユーザ属性とマニュアルで取得したユーザ属性から抽出した社会事象を比較し，同様の社会事象の抽出可能を検証する．この実験により，実世界における社会事象を自動的に取得可能なことを示し，平時ではない非習慣の社会動向の情報を検索し，タイムリーに取捨選択するためのソーシャルセンシングの高度化に寄与する．

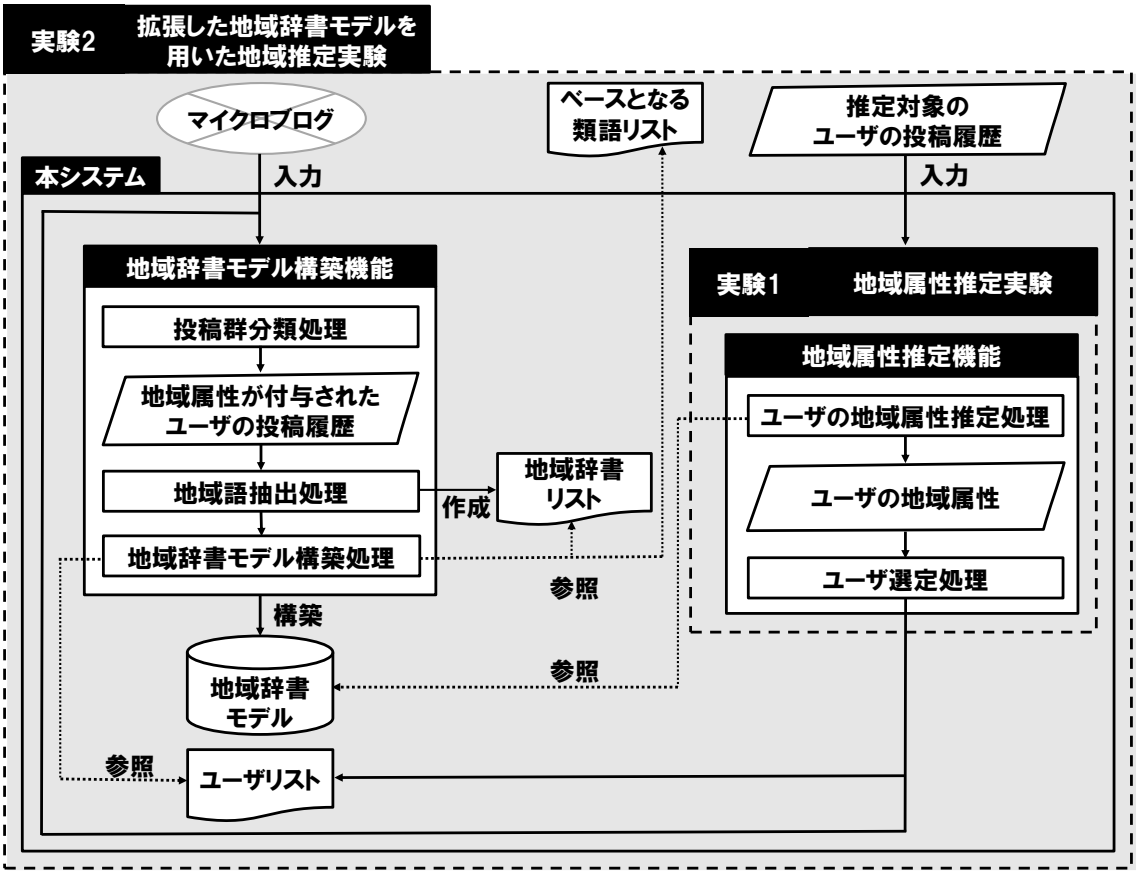


図 5.5　実験計画

本研究では，ユーザ属性ごとの社会事象の自動抽出技術の実現に向けた地域属性の推定手法を目標としている．そのため，災害やオリンピックなどといった社会事象に関しては，近隣地方でも差異は少ないと考えられる．そのため，推定結果を求める際は，各地域の結果と近隣地方を正解とした際の推定結果も一緒に提示する．近隣地方の定義としては，隣り合う地方を対象とし，上位 3 地方に該当の地方が推定されるかで判断する．

(1)　実験データ

本実験では，代表的なマイクロブログサービスである Twitter から以下の手順でデータを収集する．

STEP1： Twitter のプロフィール検索サービスであるツイプロ[40]を用いて，プロフィール欄や投稿内容に地域語を記載しているユーザを無作為に収集する．

STEP2： STEP1 にて収集したユーザの投稿内容を TwitterAPI と Twitter の投稿内容をブログ形式で保存するサービスである Twilog を用いて収集する．Twilog から取得したユーザが TwitterAPI で取得したユーザと重複している場合は，Twilog のデータを優先して採

用する．理由としては，TwitterAPI では，取得数制限があるが，Twilog は，ユーザの全投稿内容を抽出可能であり，投稿数が多いためである．

STEP3： 取得したユーザのプロフィール欄や投稿内容を目視で確認し、地域語の出現や投稿内容を基に，8 つの地域に分類し，地域属性を付与する．

以上の手順で収集したデータを用いて，2 つのデータセットを作成する．各データセットのユーザの内訳を表 5.2 に示す．なお，2 つのデータセットのユーザは重複しないものとする．

表 5.2 実験データ

	学習対象ユーザ	推定対象ユーザ
北海道地方	150 名	49 名
東北地方	150 名	50 名
関東地方	150 名	50 名
中部地方	150 名	50 名
近畿地方	150 名	50 名
中国地方	150 名	50 名
四国地方	150 名	49 名
九州地方・沖縄県	150 名	50 名

(2) パラメータの設定

マイクロブログにおけるユーザの地域属性を推定するため，各地域の投稿履歴より地域語を抽出し，各地域の特徴量を算出する．地域語の特徴量の算出には，既存研究[33]を参考にして，χ^2 値検定による素性の選択による結果を用いる．本研究では，対象とする類語リストの数に合わせて上位 250 語とする．素性の選択を行うにあたり，実験対象の素性数を変更することにより，推定精度に影響があると考えられる．しかし，データ数に応じて，適切な語句数は大きく変化し，拡張に応じて単語の地域性が洗練され，単語数も増加するため今後の議論とする．

5.5.2 実験1：地域辞書モデルを用いた地域属性推定実験

(1) 実験内容

本実験では，検証項目 5 を確認することを目的に，本提案手法で構築した地域辞書モデルを用いて，各ユーザの推定精度を検証する．

評価指標には，適合率，再現率と F 値を算出し，地域語のみで構築したモデルの推定精度と比較する．

(2)　実験結果

実験 1 の結果を表 5.3 から表 5.6 に示す.

表 5.3　地域語のみを活用した地域推定結果

地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方	49	0	0	0.00	0.00	0.00
東北地方	50	0	0	0.00	0.00	0.00
関東地方	50	47	302	0.94	0.13	0.24
中部地方	50	1	35	0.02	0.03	0.02
近畿地方	50	1	12	0.02	0.08	0.03
中国地方	50	0	0	0.00	0.00	0.00
四国地方	49	0	0	0.00	0.00	0.00
九州地方・沖縄県	50	0	0	0.00	0.00	0.00
合計	398	49	349	0.12	0.12	0.12

表 5.4　地域辞書モデルを活用した地域推定結果

地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方	49	12	4	0.24	0.75	0.37
東北地方	50	6	0	0.12	1.00	0.21
関東地方	50	48	270	0.96	0.15	0.26
中部地方	50	9	16	0.18	0.36	0.24
近畿地方	50	4	1	0.08	0.80	0.15
中国地方	50	4	13	0.08	0.24	0.12
四国地方	49	7	0	0.14	1.00	0.25
九州地方・沖縄県	50	4	0	0.08	1.00	0.15
合計	398	94	304	0.24	0.24	0.24

表 5.5　地域語のみを活用した近隣地方を正解とした地域推定結果

地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方 （東北地方）	49	4	45	0.08	0.44	0.14
東北地方 （北海道地方, 関東地方）	50	50	0	1.00	1.00	1.00
関東地方 （東北地方，中部地方）	50	50	0	1.00	0.40	0.57
中部地方 （関東地方，近畿地方）	50	50	0	1.00	1.00	1.00
近畿地方 （中部地方，中国地方， 四国地方）	50	49	1	1.00	0.89	0.94
中国地方 （近畿地方，四国地方， 九州地方・沖縄県）	50	50	0	1.00	1.00	1.00
四国地方 （近畿地方，中国地方， 九州地方・沖縄県）	49	48	1	0.98	0.98	0.98
九州地方・沖縄県 （中国地方，四国地方）	50	3	47	0.06	0.30	0.10
合計	398	304	94	0.76	0.76	0.76

表 5.6　地域辞書モデルを活用した近隣地方を正解とした地域推定結果

地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方 （東北地方）	49	24	25	0.49	0.49	0.49
東北地方 （北海道地方, 関東地方）	50	50	0	1.00	1.00	1.00
関東地方 （東北地方，中部地方）	50	50	0	1.00	1.00	1.00
中部地方 （関東地方，近畿地方）	50	50	0	1.00	1.00	1.00
近畿地方 （中部地方，中国地方， 四国地方）	50	50	0	1.00	1.00	1.00
中国地方 （近畿地方，四国地方， 九州地方・沖縄県）	50	43	7	0.86	0.86	0.86
四国地方 （近畿地方，中国地方， 九州地方・沖縄県）	49	41	8	0.84	0.84	0.84
九州地方・沖縄県 （中国地方，四国地方）	50	40	10	0.80	0.80	0.80
合計	398	348	50	0.87	0.87	0.87

実験結果より，地域語のみを使用した場合，8 地域の推定精度の F 値は，0.12 となり，近隣地方を正解とした場合，0.76 なった．また，地域辞書を使用した場合，8 地域の推定精度の F 値は，0.24 となり，近隣地方を正解とした場合，0.87 となった．表 5.3 から表 5.6 より，次に示す内容が明らかとなった．

- **地域辞書の方が，全体精度が向上することが明らかとなった．**

表 5.3 と表 5.4 を比較すると，全ての地域で精度が向上することがわかった．これは，地域語だけでは曖昧であった地域分類が，類語と組み合わせて用いることにより，地域分類がより明確になったからであると考えられる．また，関東地方のユーザ不正解数が減少していることから，関東地方と分類されていたユーザが，他の地方に分散されたことも要因と考えられる．全体を確認すると，関東地方にユーザが集中することがわかった．これは，他の地域から上京しているユーザが多い点が考えられる．他の地域の推定精度を確認する

と，内容に偏りがあることから，地域差の特徴が現れやすい地域とそうでない地域があることが判明した．

●地方によって精度に偏りがあることが明らかとなった．

表 5.5 と表 5.6 の中国地方と四国地方を確認すると，地域語のみを使用した場合の方が，中国地方は 0.14 ポイント，四国地方は 0.14 ポイント高い結果であることがわかった．これは，類語を用いたことで元々のユーザが分散されたことが考えられる．今後は，類語の出現頻度によって重み付けを変化させるなど工夫する必要がある．

北海道地方と九州地方・沖縄県に関して，地域辞書モデルと比較すると，北海道地方は 0.35 ポイント，九州地方・沖縄県は，0.70 ポイントと精度が大きく向上している．これは，類語を追加することで，地域を分類するための語句が明確となったためであると考えられる．

以上の実験結果より，中国地方と四国地方の地域の精度は低下したが，全体精度は大きく向上したため，検証項目 5 が立証できた．

5.5.3　実験2：地域辞書モデル（2回目）による地域推定実験

(1)　実験内容

本実験では，実験 1 の実験結果を用いて，地域辞書モデルを拡張し，地域属性の推定精度が向上するかを検証する．本実験に用いる学習データは，実験 1 で学習したユーザ（学習対象ユーザ）と無作為に抽出したプロフィール欄に地域を記載している 10 人のユーザを統合したデータとし，実験 1 と同様のユーザを対象に推定する．評価指標には，実験 1 と同様に適合率，再現率と F 値を用いる．

(2)　実験結果

地域辞書モデル（2 回目）を活用した地域属性の推定結果を表 5.7 と表 5.8 に示す．

表 5.7　地域辞書モデルを活用した地域推定結果（2 回目）

地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方	49	13	6	0.68	0.27	0.38
東北地方	50	15	0	1.00	0.30	0.46
関東地方	50	48	248	0.16	0.96	0.28
中部地方	50	7	30	0.19	0.14	0.16
近畿地方	50	3	5	0.38	0.06	0.10
中国地方	50	1	0	1.00	0.02	0.04
四国地方	49	7	13	0.35	0.14	0.20
九州地方・沖縄県	50	2	0	1.00	0.04	0.08
合計	398	96	302	0.24	0.24	0.24

表 5.8 地域辞書モデルを活用した近隣地方を正解とした地域推定結果（2 回目）

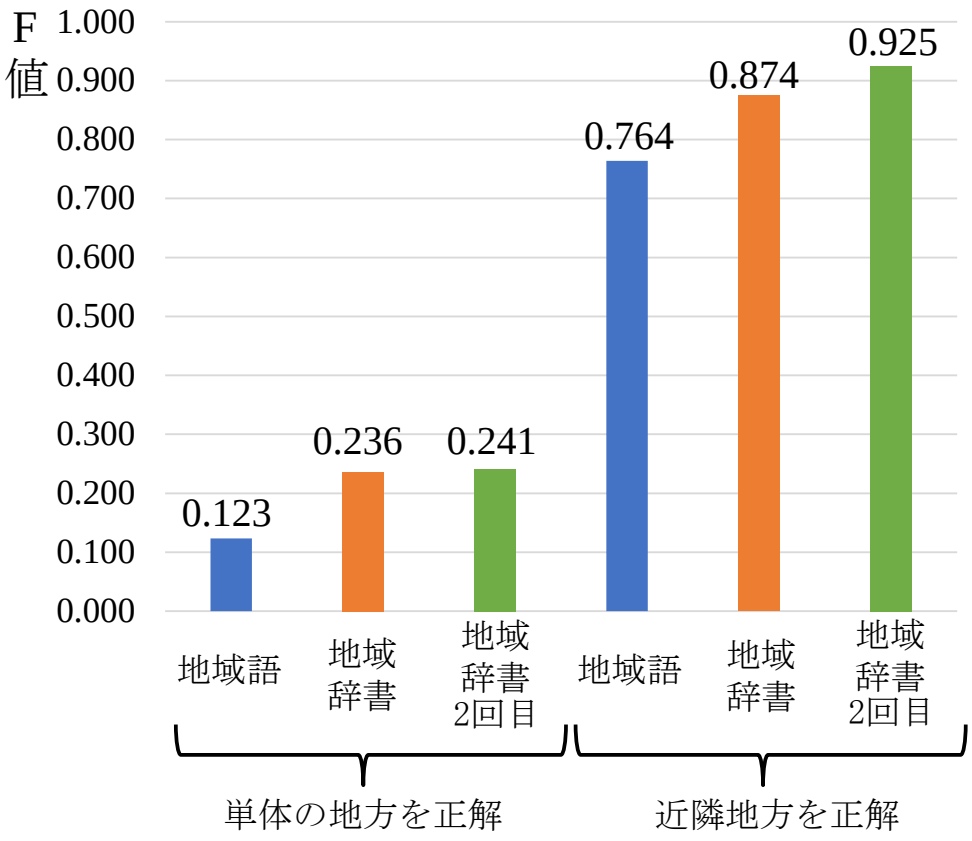
地域	全数	正解数	不正解数	適合率	再現率	F 値
北海道地方（東北地方）	49	41	8	0.98	0.84	0.90
東北地方 （北海道地方，関東地方）	50	40	10	1.00	0.80	0.89
関東地方 （東北地方，中部地方）	50	50	0	0.74	1.00	0.85
中部地方（関東地方，近畿地方）	50	50	0	0.91	1.00	0.95
近畿地方 （中部地方，中国地方，四国地方）	50	50	0	1.00	1.00	1.00
中国地方 （近畿地方，四国地方，九州地方・沖縄県）	50	49	1	1.00	0.98	0.99
四国地方 （近畿地方，中国地方，九州地方・沖縄県）	49	46	3	0.88	0.94	0.91
九州地方・沖縄県 （中国地方，四国地方）	50	42	8	1.00	0.84	0.91
合計	398	368	30	0.92	0.92	0.92

表 5.7 と表 5.8 より，地域辞書モデル（2 回目）を用いた地域属性の推定結果の全体精度が向上することが明らかとなった．この結果より，以下に示すことが明らかとなった．

- **地域辞書モデル（2 回目）を使用すると精度が向上することがわかった.**

表 5.6 と表 5.8 を比較すると，東北地方以外の地域で，精度が向上することがわかった．特に北海道地方は，0.41 ポイントと大幅に精度が向上した．これは，プロフィール欄の情報から自動的にユーザを加増し，地域語の特徴量を再計算することで，地域辞書モデルが成長したためである．これを繰り返し続けることにより，より高精度に分類できるモデルが構築できると考えられる．また，成長するモデルで分類することで，各地方単位でも高精度に分類できることが期待できる．

地域語，地域辞書と地域辞書モデル（2 回目）の全体の F 値を図 5.6 に示す.



	地域語モデル	地域辞書モデル（地域語＋類語）	地域辞書モデル2回目（地域語＋類語）
ユーザ数（各地域）	150名	150名	160名
地域語数（各地域）	250語句	250語句	250語句＋平均27語句追加
類語数（各地域）	0語句	250語句	250語句

図 5.6　各モデルの推定結果

図 5.6 より，徐々に推定結果が向上していることがわかる．特に，地域語から地域辞書の推定結果は 0.1 ポイント向上していることから地域語に加えて，類語を組み合わせる手法の有用性を証明できた．また，地域辞書モデル（2 回目）も向上していることから，繰り返し拡張を行うことで，推定精度の向上が期待できる．

以上の実験結果より，単体地方と近隣地方共に精度が向上しているため，検証項目 6 が立証できた．

5.5.4　実験3：社会事象比較実験

(1)　実験内容

本実験では，自動で推定したユーザ属性とマニュアルで取得したユーザ属性から抽出した社会事象を比較し，同様の内容が取得可能かを検証する．本実験に用いる性別，年代と職業属性のデータは，2014 年とし，地域属性は，2019 年に取得したデータと全期間を比較した投稿を用いる．性別，年代と職業属性の推定は，既存研究[33]の手法を用いている．

(2)　実験結果

マニュアルで取得したユーザ属性と推定したユーザ属性の非習慣行動の時間帯の一致率を表 5.9 から表 5.12 に示す．また，非習慣行動の抽出時間の各属性の上位と下位の平時習慣ベクトルと特定習慣ベクトルの比較結果を図 5.7 に示す．本章での平時習慣ベクトルは，全期間の投稿から作成した生活習慣ベクトルとし，特定習慣ベクトルは，年間の投稿から作成した生活習慣ベクトルを表す．

表 5.9　性別と職業属性に関する非習慣行動の時間帯の一致数

行動	男性				女性				学生				社会人				主婦				フリーター			
	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率
起床・就寝	70	52	50	71%	77	76	57	74%	89	69	61	69%	78	60	58	74%	68	74	57	84%	60	70	41	68%
在宅	72	78	69	96%	75	91	70	93%	62	71	43	69%	61	66	52	85%	81	89	71	88%	68	74	50	74%
外出	70	73	64	91%	80	84	71	89%	78	65	54	69%	57	60	41	72%	66	75	57	86%	70	74	55	79%
帰宅	70	82	66	94%	73	78	52	71%	77	79	67	87%	59	59	47	80%	80	83	59	74%	78	78	55	71%

表 5.10　年代属性に関する非習慣行動の時間帯の一致数

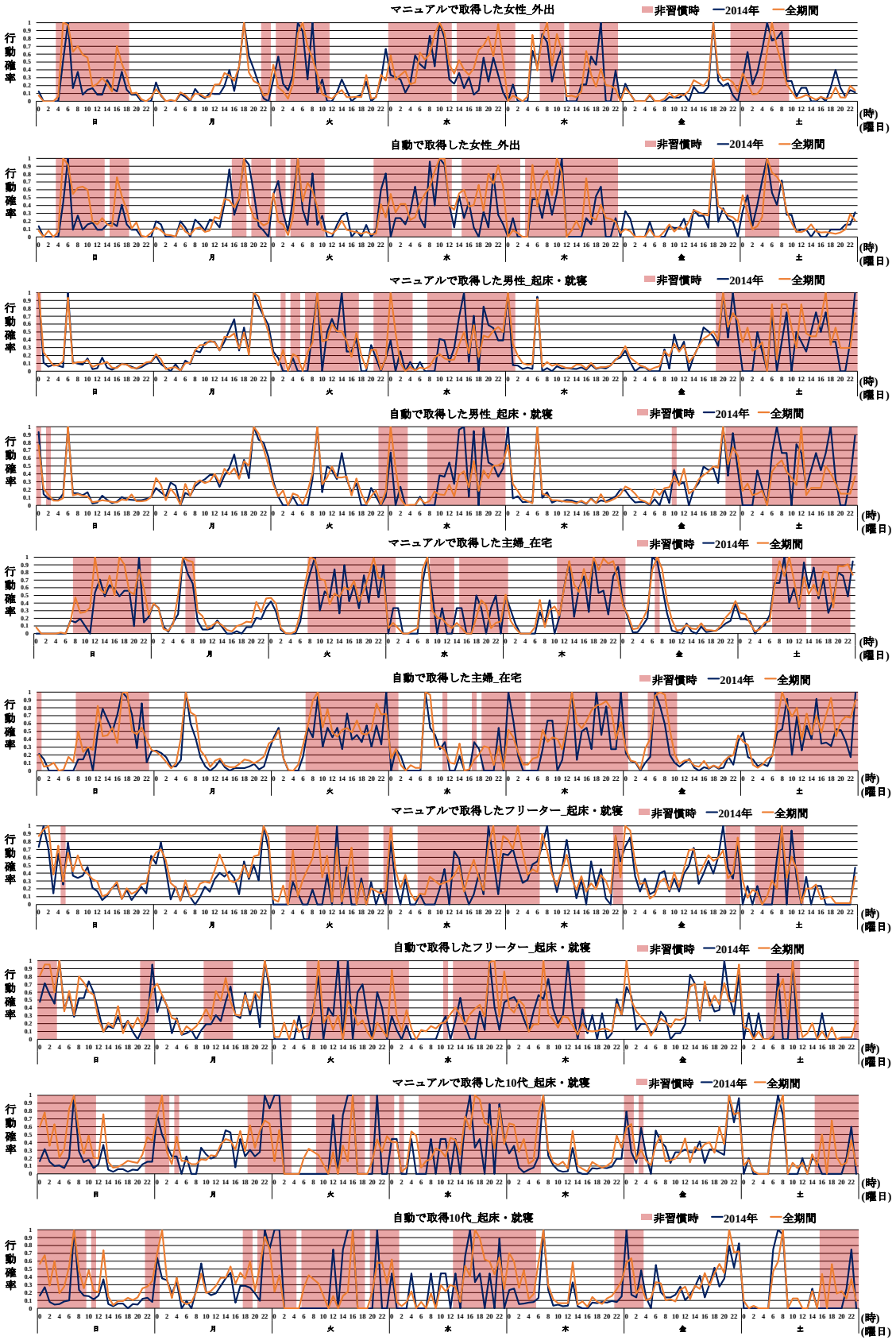
行動	10 代				20 代				30 代				40 代以上			
	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率
起床・就寝	80	74	65	81%	89	79	58	65%	81	90	57	70%	75	80	60	80%
在宅	68	66	49	72%	67	62	36	54%	85	74	58	68%	70	70	52	74%
外出	70	72	61	87%	85	73	43	51%	88	76	64	73%	73	60	47	64%
帰宅	62	75	53	85%	71	79	39	55%	87	80	53	61%	76	84	55	72%

表 5.11　地域属性に関する非習慣行動の時間帯の一致数①

行動	北海道地方（東北地方）				東北地方（北海道地方，関東地方）				関東地方（東北地方，中部地方）				中部地方（関東地方，近畿地方）			
	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率
起床・就寝	71	74	60	85%	67	76	55	82%	79	85	65	82%	79	79	70	89%
在宅	80	72	68	85%	74	83	70	95%	69	66	50	72%	87	77	69	79%
外出	81	80	72	89%	73	78	56	77%	74	63	42	57%	71	68	55	77%
帰宅	76	73	65	86%	73	71	59	81%	77	60	46	60%	73	70	60	82%

表 5.12　地域属性に関する非習慣行動の時間帯の一致数②

行動	近畿地方（中部地方，中国地方，四国地方）				中国地方（近畿地方，四国地方，九州地方・沖縄県）				四国地方（近畿地方，中国地方，九州地方・沖縄県）				九州地方・沖縄県（中国地方，四国地方）			
	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率	手動	自動	一致数	一致率
起床・就寝	79	79	79	100%	70	73	67	96%	69	66	57	83%	73	73	51	70%
在宅	70	70	70	100%	77	75	74	96%	76	75	66	87%	86	72	61	71%
外出	73	73	73	100%	83	80	77	93%	55	69	51	93%	80	69	55	69%
帰宅	60	60	60	100%	74	74	70	95%	76	70	68	89%	80	65	58	73%


マニュアルで取得した女性_外出
非習慣時
2014年
全期間
行動確率
(時)
(曜日)
日
月
火
水
木
金
土
自動で取得した女性_外出
マニュアルで取得した男性_起床・就寝
自動で取得した男性_起床・就寝
マニュアルで取得した主婦_在宅
自動で取得した主婦_在宅
マニュアルで取得したフリーター_起床・就寝
自動で取得したフリーター_起床・就寝
マニュアルで取得した10代_起床・就寝
自動で取得10代_起床・就寝

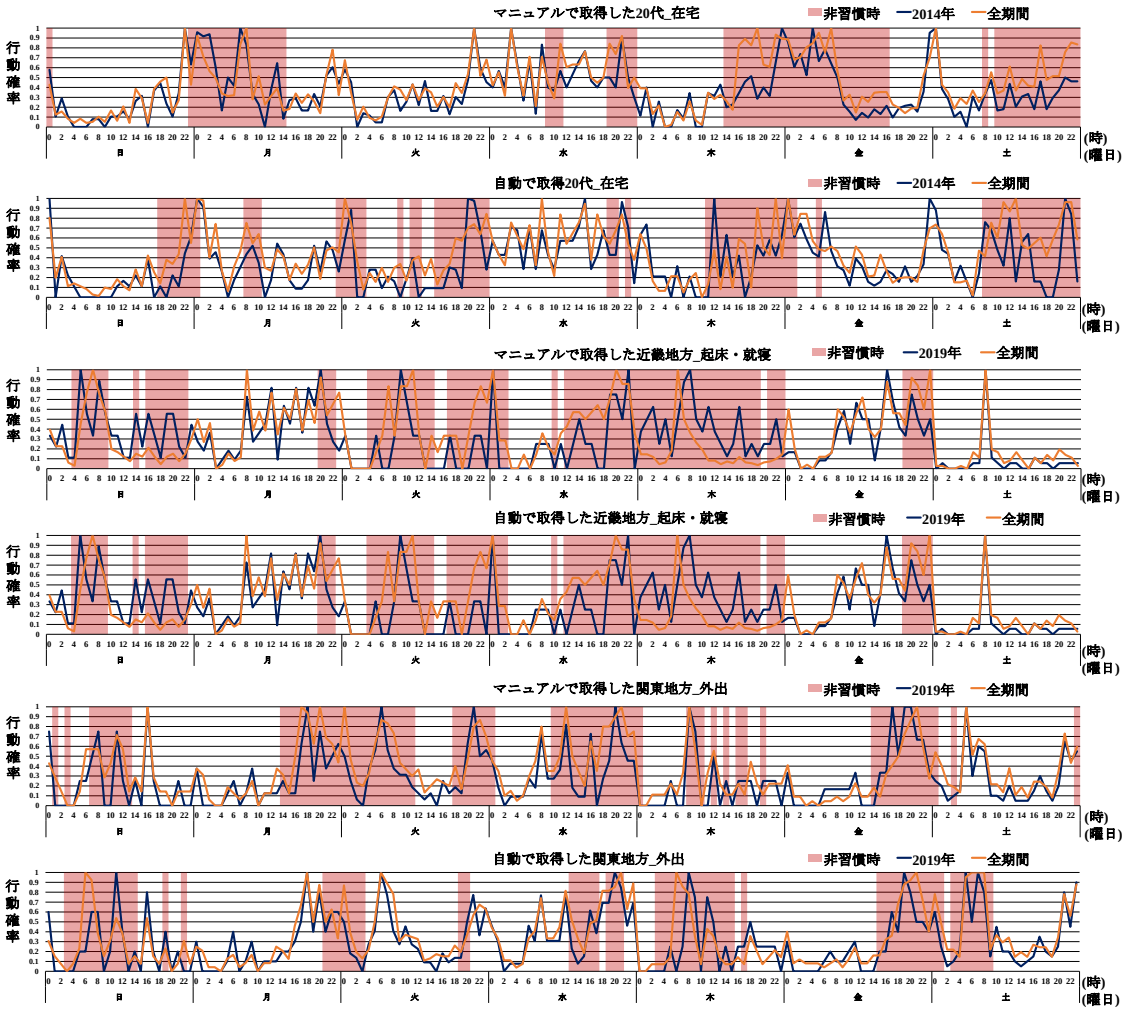


図 5.7　上位と下位の習慣行動解析結果

非習慣行動の時間帯の一致率より，平均して性別 85%，職業 77%，年代 70%，地域 84%と高精度に非習慣行動の時間帯を抽出することができることが明らかとなった.

詳細に把握するため，マニュアルで取得したユーザ属性と推定したユーザ属性から抽出した社会事象の結果を表 5.13 から表 5.17 に示す.

表 5.13　性別属性の代表的なトピック抽出結果

年	行動	マニュアルで取得したユーザ属性		推定で取得したユーザ属性	
		代表的なトピック内容		代表的なトピック内容	
		男性	女性	男性	女性
2014年	起床・就寝	バイトに関する話題 （夜，朝，昼，バイト，コンビニ） ブログに関する話題 （アニメ，タップ，ブログ，更新，Twitter）	最近話題の番組に関する話題 （妖怪，周り，最近，ウォッチ，人気） 朝帰宅でラーメンを食べるユーザ （ラーメン，朝，帰宅，大丈夫） 運勢に関する話題 （今日，運，金，健康，恋愛，運勢，チェック）	バイトに関する話題 （朝，バイト，夜） フジテレビの番組に関する話題 （今日，時間，fujitv，交通，安全）	運勢に関する話題 （運，今日，仕事，金，健康，恋愛，運勢）
	在宅	娯楽に関する話題 （アニメ，好き，萌，ゲーム），（ラジオ，ネット，音楽，放送） 仕事に関する話題 （仕事，明日，紹介，先生）	ある商品に関する話題 （安，激，手，商品，BONNE） 健康に関する話題 （多汗症，肌，プログラム，改善，医療），（健康，今朝，スムージー）	娯楽に関する話題 （ラジオ，更新，アニメ，ネット） 好きなテレビ番組に関する話題 （好き，キン肉マン，時代，戦闘，番組）	年末のテレビ番組に関する話題 （時間，家，録画，年末） 買い物に関する話題 （ポイント，買い物，予定）
	外出・帰宅	外出に関連する話題 （旅，ブログ，更新，文章，ネタ） テレビ番組に関する話題 （tvasahi，消失，ハイジャック），（事故，tvasahi，ハイジャック）	友人との外出に関する話題 （好き，楽しみ，友達，神社，人気） テレビ番組の感想に関する話題 （紅白，時間，夫，最高，息子，家族）	外出に関連する話題 （ゲット，達成，ポケモン），（ブログ，更新，旅） 健康に関する話題 （手，脳，汗，多汗症），（内容，多汗症，改善，プログラム）	テレビ番組の感想に関する話題 （紅白，曲，大好き，気持ち，私） 神社に関する話題 （神社，楽しみ，大事）

表 5.14　職業属性の代表的なトピック抽出結果

年	行動	マニュアルで取得したユーザ属性				推定で取得したユーザ属性			
		代表的なトピック内容				代表的なトピック内容			
		学生	社会人	主婦	フリーター	学生	社会人	主婦	フリーター
2014年	起床・就寝	サイクリングに関する話題（ロード，バイク，サイクリング，多摩湖）	政治に関する話題（安倍，首相，政権，対策，実施），（政治，経済，秘密，保護，法，候補，国民，原発） 集団的自衛権と増税に関する話題（権，的，集団，自衛，消費，税，日本，増税）	懸賞に関する話題（プレゼント，懸賞，キャンペーン）	政治に関する話題（選挙，自民党，選，自民）	サイクリングに関する話題（ロード，バイク，サイクリング，多摩湖） 政治に関する話題（安倍，自民党，アベノミクス，増税，），（政府，増税）	政治に関する話題（安倍，政権，首相，国際，解散，国民，アベノミクス）	占いに関する話題（仕事，運，今日，健康，金，恋愛）	仕事に関する話題（仕事，職場，不安，面接）
	在宅	娯楽に関する話題（ハッカ，ドール，好き，今日，予定，少女），（好き，曲，アニメ，ボカロ，大好き），（好き，アニメ，曲，ライブ，ラブ）	政治に関する話題（安倍，ニュース，政権，田母神，権，演説，自衛，集団，政府），（政策，アベノミクス，安倍）	節約する術に関する話題（節約，方法，保存，カード，冷蔵，冷凍）	政治に関する話題（国，原発，自民党，汚染，朝日新聞，公明党），（安倍，自民党，政権，アベノミクス，増税）	娯楽に関する話題（アイドル，ラブライブ，ベスト，グループ） 政治に関する話題（議員，国会，マスコミ，都知事，政権，省，責任），（権，自衛，集団，解散，改憲）	政治に関する話題（安倍，首相，日本，政権）	地震の募金に関する話題（緊急，支援，東日本，大震災，基金，協賛，募金）	娯楽に関する話題（MH，デレラジ，レギオス，ガチャガチャ），（ロール，マトリョシカ，インビジブル，メルト，ブラックロックシューター）
	外出・帰宅	ロードバイクに関する話題（ロード，バイク，大学生，お金，グレード，サイクリングロード），（バイク，ロード，サイクリング，大学生） 娯楽に関する話題（曲，声，ボカロ，大好き），（好き，ハッカ，アニメ，ドール）	スマートフォンに関する話題（iPhone，ドコモ，種類，docomo，スマートフォン），（iPhone，犬，dog） 仕事に関する話題（感謝，仕事，好き，心，時間）	懸賞に関する話題（プレゼント，懸賞，キャンペーン），（プレゼント，旅行，カード，プレゼント）	感染症に関する話題（エボラ，予測，感染，ミック，パンデ） 政治に関する話題（自民党，批判，アベノミクス，安倍，活動），（安倍，首相，原発，自衛，政権，言論，統制） バイトに関する話題（仕事，金，バイト，時給，文句，早退）	政治に関する話題（自衛，内閣，戦争，憲法），（安倍，解散，首相，原発，自民党，アベノミクス，増税） ロードバイクに関する話題（ロード，バイク，サイクリング）	政治に関する話題（安倍，首相，中国，消費，税） スマートフォンに関する話題（ドコモ，docomo，スマートフォン）	台風に関する話題（雨，窓，台風），（台風，嫌，疲れ，移動，風） 家事に関する話題（子供，物，洗濯，母親） プレゼントに関する話題（プレゼント，キャンペーン，抽選）	プロ野球に関する話題（阪神，勝利，優勝，視聴）

表 5.15　年代属性の代表的なトピック抽出結果

年	行動	マニュアルで取得したユーザ属性				推定で取得したユーザ属性			
		代表的なトピック内容				代表的なトピック内容			
		10 代	20 代	30 代	40 代以上	10 代	20 代	30 代	40 代以上
2014 年	起床・就寝	娯楽に関する話題（野崎，月刊，くん，ファボ），（ライブ，ラブ，マスター，マトリョシカ），（うた，プリ，王子，オオカミ）	ダイエットに関する話題（体，効果，ストレッチ，自転車，運動，エアロビクス，酸素，ダイエット） 友人との予定に関する話題（明日，絶対，友達，休み，予定）	食事に関する話題（年末，ごはん，生活，味，鍋，カロリー）	地震に関する話題（地震，発生，確率，巨大，倍，発表）	娯楽に関する話題（アニメ，少女，好き，アオハライド，王子，オオカミ）（うた，プリ，オタク，Free，野崎，月刊，LOVE）	ダイエットに関する話題（効果，ストレッチ，時間，体重），（運動，時間，酸素，自転車，エアロビクス，健康） 娯楽に関する話題（ライブ，ラブ，アニメ，スクフェス，声優，秋葉原）	大晦日のテレビ番組に関する話題（紅白，曲，終了，大晦日，日本）	旅行の予定に関する話題（花見，登山，アウトドア，下り），（アクセス，車，温泉），（アウトドア，釣り，ツアー，子供，コース，カヌー）
	在宅	娯楽に関する話題（ライブ，ラブ，ゲーム，リクエスト，大好き），（ボカロ，好き，マトリョシカ，インビジブル，）	通勤に関する話題（電車，足，通勤，冬，大変） 占いに関する話題（仕事，運，金，恋愛，明日，ダメ，運勢）	掃除に関する話題（簡単，みんな，風呂，掃除，洗濯）	地震に関する話題（地震，確率，巨大，発生，NHK，倍，去年，発表，千葉，安全，横浜，確保）	娯楽に関する話題（ライブ，ラブ，ゲーム，リクエスト，大好き），（曲，ボカロ，希望，拡散，好き）	娯楽に関する話題（好き，シンデレラガールズ，アイドル，マスター），（ライブ，アニメ，ラブ，放送）	年末に関する話題（楽しみ，雪，冬，そば，実家，天気，ケーキ），（ゲーム，テレビ，最高，大晦日，夜）	アウトドアに関する話題（登山，好き，山，ご飯，スポーツ，温泉，最高，自然，アウトドア）
	外出・帰宅	娯楽に関する話題（大好き，ライブ，ラブ，スクフェス），（パス，黒，うた，プリ，ハイキュー）	通勤に関する話題（電車，足，通勤，大丈夫） 健康に関する話題（運動，時間，毎日，健康，酸素，温泉）	スーパーからの帰宅後にピアノを演奏するユーザ（帰宅，ピアノ，スーパー，帰り，演奏）	台風に関する話題（無理，台風，会社，足）	娯楽に関する話題（少女，野崎，月刊，アオハライド），（バイト，幸せ，みんな，最高，好き）	健康に関する話題（運動，毎日，健康，酸素） バイトメンバーに関する話題（バイト，幸せ，みんな，最高，好き）	地震に関する話題（大変，巨大，NHK，地震，発生，発表，確率），（地震，発生，巨大，確率，発表，安全）	アウトドアに関する話題（アウトドア，専門，店，最高，ブーム，到来），（登山，好き，スポーツ，山），（旅行，アウトドア，アクセス，バス）

表 5.16　地域属性の代表的なトピック抽出結果①

年	行動	マニュアルで取得したユーザ属性				推定で取得したユーザ属性			
		代表的なトピック内容				代表的なトピック内容			
		北海道地方 （東北地方）	東北地方 （北海道地方，関東地方）	関東地方 （東北地方，中部地方）	中部地方 （関東地方，近畿地方）	北海道地方 （東北地方）	東北地方 （北海道地方，関東地方）	関東地方 （東北地方，中部地方）	中部地方 （関東地方，近畿地方）
2019年	起床・就寝	食事に関する話題 （カレー，ラーメン，絶対，室蘭，明日） 女性有名人の書籍の購入に関する話題 （本，購入，櫻子，大原，21，値段，販売）	思い出を投稿するユーザに関する話題 （思い出，投稿，大好き，テンション，チケット）	ライブに関する話題 （たくさん，参加，お金，ライブ，心配）	地域おこしに関する話題 （地域おこし，ニュース，写真，新聞，皆様）	食事に関する話題 （ラーメン，天気，雨，函館，期待），（好き，札幌，北海道，駅，パン），（肉，ご飯，今日，牛），（今日，休み，仕事，久しぶり，，温泉，ビール）	思い出を投稿するユーザに関する話題 （思い出，アンコール，気持ち，映画，，元気） 山形のライブに関する話題 （山形，ライブ，最高，参戦）	映画に関する話題 （最高，沖縄，映画，好き） ライブに関する話題 （好き，曲，セカオワ，ファン，歌），（大好き，ライブ，ワンオク）	地域おこしに関する話題 （地域おこし，ニュース，新聞，地域，新潟）
	在宅	女性有名人に関する話題 （ラーメン，顔，櫻子，大原），（櫻子，大原，2019，LINE） 食事に関する話題 （ブログ，更新，卵，カレー，魚） ライブに関する話題 （ワンオク，セカオワ，ライブ，好き，ファン，動画，アルバム，ツアー）	音楽に関する話題 （aiko，ファン，好き，期待，YouTube，歌），（曲，恋，平成，音）	棋王戦に関する話題 （公式，会，段，将棋，棋王，楽しみ）	娯楽に関する話題 （バンドリ，レベル，無料），（最高，綺麗，スクフェス，，Aqours），（Aqours，ラブライブ，活動，歌）	食事に関する話題 （ブログ，更新，ラーメン，北海道，屋）	音楽に関する話題 （aiko，YouTube，音，演奏，Live），（山形，米津，ファン），（好き，夢，aiko，幕張，フェス）	歌手に関する話題 （好き，曲，セカオワ，ワンオク，動画）	娯楽に関する話題 （Aqours，ラブライブ，ネット，熱），（バンドリ，ガルパ，CVB，alexandros）
	外出・帰宅	北海道のイベントに関する話題 （最高，車，店，釧路，釣り，札幌，期待，イベント） ライブに関する話題 （好き，曲，セカオワ，音，ライブ），（楽しみ，ライブ，ツアー，予約）	東北地方のイベントに関する話題 （写真，岩手，投稿，イベント，盛岡，秋田） ライブに関する話題 （ライブ，曲，一緒，参加） 女性有名歌手に関する話題 （aiko，大好き，毎日）	ライブに関する話題 （子ども，ライブ，動画，元気，イベント），（ライブ，出演，日々，毎日）	娯楽に関する話題 （平成，lovelive，ぷちぐる，日本），（曲，Aqours，ラブライブ，活動，成長） ライブに関する話題 （最高，ライブ，イベント，旦那，息子，子ども）	ライブに関する話題 （ライブ，綺麗，麻衣，白石）	ライブに関する話題 （曲，お願い，ライブ，アンコール），（ライブ，大好き，アンコール，会場，最高，期待，感動），（仙台，宮城，ライブ，米津，チケット） 女性有名歌手に関する話題 （aiko，時間，大好き，女子，愛）	ライブに関する話題 （LiSA，ライブ，曲，紹介，音楽），（ライブ，ワンオク，作品，情報，ファン，セカオワ）	娯楽に関する話題 （ラブライブ，みんな，好き，イベント）

表 5.17　地域属性の代表的なトピック抽出結果②

年	行動	マニュアルで取得したユーザ属性				推定で取得したユーザ属性			
		代表的なトピック内容				代表的なトピック内容			
		近畿地方（中部地方，中国地方，四国地方）	中国地方（近畿地方，四国地方，九州地方・沖縄県）	四国地方（近畿地方，中国地方，九州地方・沖縄県）	九州地方・沖縄県（中国地方，四国地方）	近畿地方（中部地方，中国地方，四国地方）	中国地方（近畿地方，四国地方，九州地方・沖縄県）	四国地方（近畿地方，中国地方，九州地方・沖縄県）	九州地方・沖縄県（中国地方，四国地方）
2019年	起床・就寝	ライブに関する話題（大阪，みんな，神戸，ライブ，最高，曲）	初日の出に関する話題（今年，思い，新年，初日出） 島根県のお酒に関する話題（これ，楽しみ，島根，酒）	徳島での気分転換に関する話題（嫌，全部，徳島，綺麗，これ，元気） ライブに関する話題（ライブ，くん，好き，SexyZone，キンプリ，セクゾ） 四国のラーメン店に関する話題（県，愛媛，店，香川，松山，高松，ラーメン，近く）	沖縄での練習に関する話題（沖縄，演習，時間，今日，帰宅） 食事に関する話題（ランチ，週末，久々，大分）	イルカショーに関する話題（大阪，新春，イルカ，楽しみ，ライブ）	決心を行うユーザ（努力，今日，自分，絶対，成功）	四国のラーメン店に関する話題（大好き，店，愛媛，香川，ラーメン）	練習に関する話題（今日，練習，帰宅） 娯楽に関する話題（好き，俺，幸せ，アニメ）
	在宅	タピオカ店に関する話題（タピオカ，pipipi，屋，綺麗）	バスケットボールチームに関する話題（島根，予定，スサノオマジック，時間）	バイクイベントに関する話題（バイク，ライダー，キャンプ，イベント）	食事に関する話題（家，残念，嫌，ご飯），（子ども，たち，夕飯，準備，疲れ）	タピオカに関する話題（勉強，タピオカ，最高，食い）	仕事に関する話題（ビジネス，努力，価値，提供，成功）	商品購入に関する話題（タイム，セール，情報，Amazon，お買い得，ブランド），（最高，購入，今日，Amazon）	テニスに関する話題（練習，錦織，なおみ，結果） 食事に関する話題（素敵，スイーツ，ランチ，珈琲）
	外出・帰宅	SNS に関する話題（SNS，更新，メンバー毎日）	仕事に関する話題（大丈夫，上司，仕事，心配，無理），（休み，仕事，生理），（仕事，旅行，保険，職場）	ショッピングサイトのタイムセールに関する話題（タイム，Amazon，セール，2019，情報，お買い得） バイクツーリングに関する話題（バイク，ツーリング，ライダー，集い）	cocoa に関する話題（日本，絶対，cocoa，ストレス）	SNS に関する話題（情報，インスタ，DM，更新） 仕事に関する話題（我慢，夜勤，鼻風邪，可哀そう）	バスケットボールチームに関する話題（島根，スサノオマジック，選手）	ショッピングサイトのタイムセールに関する話題（セール，Amazon，情報，販売，発表） ライブに関する話題（配信，サザン，音，ライブ），（最高，ライブ，参加），（歌，曲，感動，限定）	テニス選手に関する話題（ちゃん，なおみ，大阪，最高）

表 5.13 から表 5.17 より, マニュアルで取得したユーザ属性と推定したユーザ属性から抽出した社会事象の結果を同程度に取得可能であることがわかり，本システムの有用性を証明した.

5.6 あとがき

本研究では，同じ意味を表すが地域ごとに表現が異なる「アホ・バカ」等の類語に着目した．そして，Twitter から地域ごとのユーザを自動的に収集し，そのユーザ群から抽出した地域語と類語を組み合わせた「地域辞書」を用いて，学習モデルを拡張する新たな手法を提案し，高精度に地域属性を推定する手法を提案した．RNN の一種である Bi-LSTM を用いて，地域ごとにモデルを構築し，高精度に推定できることを確認した．また，前述の半自動で付与したユーザ属性を用いた結果と本提案で取得した地域属性も用いた結果とを比較し，同様の社会事象を抽出できたことを確認した．最終的に，一連の研究成果の有用性を証明することができた.

第 6 章

総括

第6章　総括

「第5期科学技術基本計画」にてサイバー空間とフィジカル空間を融合させた新たな社会像である Society 5.0 が提唱され，ネットワーク技術や IoT，そして AI 技術を用いて，経済成長に向けた社会課題を解決するための取り組みがなされつつある．その中でも，インターネットから社会の動向やニーズを把握することを目的としたソーシャルセンシング技術の研究が進められており，マーケティングやデータマイニング，社会調査等の分野で活用されている．ソーシャルセンシング技術は，インターネットからユーザの行動を分析することで，ユーザの意見や暗黙的な考え，そしてニーズを抽出することができる．

本研究では，平時の状況を把握することで，異常時や緊急時に素早く対応するソーシャルセンシング技術を開発し，実社会の新たな事象を感度良く見極めることに主眼を置く．具体的には，SNS 上の投稿時間と投稿内容からユーザごとの日々の生活習慣に応じた「習慣行動」を分析し，その情報を基に，社会事象を適切に抽出するための方法について議論する．それは，平時の習慣行動と異なる「非習慣行動」を抽出することで個々のユーザの反応を分析する手法と，性別，年代，職業や地域等のユーザの属性単位での習慣行動の違いから社会事象を抽出する手法によって深く検討する．さらに，ソーシャルセンシングで重要となるユーザの属性，ここでは，性別，年代，職業，地域を投稿内容から推定するためのシステムを実装し，センシング精度を高める技術を考究する．

各章で取り組んだ内容について，それぞれ概説する．

第 1 章では，研究の背景として，政府が策定した「第 5 期科学技術基本計画」にある Society 5.0 と CGM の関係性の現状を整理し，SNS をソーシャルセンサとして活用している取り組みについて述べた．既存研究の解析対象の検索履歴の入手が困難な課題，ブログの特性上，イベントからユーザの投稿までのタイムラグが発生するため，即時性の課題，各社会事象に特定のキーワードを事前に指定する必要があるため，網羅的な分析が困難であることや，キーワード選定に解析者のバイアスがかかり分析に偏りがみられること等の課題に対し，本研究では，平時の状況を把握することで，異常時や緊急時に素早く対応するソーシャルセンシング技術を開発し，実社会の新たな事象を感度良く見極めることに主眼を置き，本研究の目的について概説した．「非習慣行動を用いた社会事象の抽出」では，SNS 上の投稿時間と投稿内容からユーザごとの日々の生活習慣に応じた習慣行動を分析し，その情報を基に，社会事象を適切に抽出するための方法について提示した．それは，平時の習慣行動と異なる非習慣行動を抽出することで個々のユーザの反応を分析する手法と，性別，年代，職業や地域等のユーザの属性単位での習慣行動の違いから社会事象を抽出する手法によって深く検討した．さらに，ソーシャルセンシングで重要となるユーザの属性，ここでは，

性別，年代，職業，地域を投稿内容から推定するためのシステムを実装し，センシング精度を高める技術を提示した．

第 2 章では，マイクロブログ上に日々蓄積されているビッグデータを対象とし，平時と非習慣時のソーシャルセンシングにおいて，課題となっているユーザの地域属性に関する推定手法の実現を目的とした既存研究と本研究の位置づけを提示した．そして，既存研究の課題を整理し，具体的な解決手法を提案した．社会事象を抽出する技術における課題として「キーワードを選定する時点で恣意的な分析方法である上，網羅的に社会事象を捉えることができない」を挙げ，「ユーザの過去の投稿履歴から取得できる習慣行動に着目し，平時とは異なる行動から非習慣行動を取り出す手法」を提案した．また，属性推定における課題として「性別ごとの顕著な特徴がみられない属性の推定が難しい問題」と「地域に関する語句やジオタグといった情報が少ない問題」を挙げ，同じ意味を表すが地域ごとに表現が異なる類語，具体的には「アホ・バカ」などの語句に着目し，語句ごとの地域性の違いから地域属性を推定する手法を提案した．そして，手法内容の有用性を検証するための研究計画を整理した．

第 3 章では，マイクロブログの投稿からユーザ群の非習慣行動を抽出することで，キーワードに依存せずに社会事象を抽出する新たな手法を提案した．提案手法において，既存研究の課題である「社会事象ごとに特定のキーワードを事前に指定する必要があるため網羅的な分析が困難であること」，「キーワード選定に解析者のバイアスがかかり分析に偏りがみられること」の課題に対して，非習慣行動に着目することで，事前にキーワードを指定せずに社会事象を抽出できた．さらに，同一トピックに関しても既存手法と比較して複数の内容が抽出できており，社会事象を詳細に把握できた．これにより，「平時と異なる行動を起こすユーザ群を特定してその投稿を解析することで，社会事象を抽出可能であること」と「ソーシャルセンサの特性としてユーザの生活習慣を考慮することで，抽出可能な社会事象の内容やカテゴリが変化すること」を解消することができた．本研究を通じて，次に示す課題が明らかになった．

課題1：抽出可能なトピックが全国的な社会事象のみであったこと

課題2：投稿を生活習慣のみで絞ったため，トピック内に判断が難しいキーワードがあったこと

課題3：ユーザの属性毎のトピック抽出には至っていないこと

課題4：提案手法で抽出可能な社会現象は，「対象の事象に関わる投稿が Twitter 上でなされていること」と，「各生活パターンにおいて投稿数のずれが生じる内容であること」の 2 つの条件が満たされる必要があること

今後は，上述の 4 つの課題に対応する方策を検討しつつ,「平時習慣と特定習慣のタイムスパン（年，月，週，日）の組合せ」と「抽出される社会事象」との関係を明らかにする予定である．

第 4 章では，ユーザの習慣行動を用いて，その習慣行動の違いから，実世界における事象を抽出するソーシャルセンシング手法を適用した．そして，社会事象との関係性を実践的に評価するために，ユーザ属性ごとにトピックを分類する手法の汎用性を明らかにした．これにより，同一トピックに関してもトピックの抽出内容が変化し，既存研究よりも社会事象を詳細に把握できた．よって,「ソーシャルセンサの特性として，ユーザの属性ごとに習慣行動が相違し，非習慣行動の時間帯が異なること」と「ユーザ属性ごとに反応や興味を持つ社会事象が異なり，同じ社会事象でも抽出内容が異なること」を解消することができた．本研究を通じて，次に示す課題が明らかになった．

課題1：ユーザ属性に応じて，平時習慣ベクトルと特定習慣ベクトルのタイムスパン（年，月，週，日）ごとの変化を考慮できていないこと．

課題2：投稿内容を習慣行動のみで絞ったため，トピック内に判断が難しいキーワードがあったこと．

今後は，上記の 2 つの課題に対応する方策を検討し，社会事象のセンシング技術の実用化に向けて，季節や気象などの他の指標の関係を明らかにする予定である

第 5 章では，初期設定した 250 件の類語をベースとして，収集した投稿記事から地域語を抽出し，類語と組み合わせて地域辞書モデルを生成することに主眼を置いた．そして，その地域辞書モデルを活用して，マイクロブログユーザの地域属性を推測する手法を提案した．実証実験を通じて，１）段階的詳細化によってユーザの性別，年代や職業の基本属性は比較的簡単に確定できることがわかった．一方，地域属性に関しては，２）地域辞書モデルを用いることで大まかな地方区分レベルに分類できること，そして３）ユーザの地域語を増加させることで地域辞書が洗練され，地方区分から都道府県レベルへと収斂する傾向があることがわかった．したがって，これまでマニュアルで行ってきたユーザ属性を分類する手間が省け，自動的にユーザ属性を獲得できること，そして，非習慣行動を用いた社会事象の分析が比較的簡単になることがわかった．また，拡張を繰り返し行うことで，ユーザ数と投稿数を増加し，各地域ではなく，より絞った地域のモデルが構築可能であり，コストの削減にも寄与できると考えられる．

さらに，最終確認として，マニュアルで付与したユーザ属性と自動的にユーザ属性を推定した時の非習慣行動によるトピックを比較することにより，同様の社会事象を獲得できることを確認した．

以上のことから，社会事象の抽出においてマイクロブログと習慣行動の情報は有用であることが明確となった．また，ユーザ属性を考慮することで詳細な社会事象を抽出できることもわかった．し

たがって，平時ではない非習慣の社会動向の情報を検索し，タイムリーに取捨選択するためのソーシャルセンシング技術の高度化について深く議論し，実世界における有益な情報を適切に抽出することを可能であることがわかった．

今後は，地域の分類を詳細化し，より汎用性の高いモデルを構築する．また，投稿者の性別，年代，職業，そして地域を特定することで，投稿記事の信頼性を判断しながら平時と非習慣時のデータセンシング技術の確立を目指す．

参考文献

参考文献

[1] 総務省：第 5 期科学技術基本計画，入手先 < https://www8.cao.go.jp/cstp/kihonkeikaku/5honbun.pdf > (参照 2020-4-6).

[2] 総務省：平成 30 年度版情報通信白書「Society 5.0」，入手先 < https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd102300.html > (参照 2020-4-6).

[3] 総務省消防庁：規模災害時におけるソーシャル・ネットワーキング・サービスによる緊急通報の活用可能性に関する検討会，入手先 < https://www.fdma.go.jp/singi_kento/kento/kento101.html > (参照 2020-4-11).

[4] 首相官邸：防災・減災におけるＳＮＳ等の民間情報の活用等に関する検討会，入手先 < https://www.caa.go.jp/future/meeting_materials/review_meeting_001 > (参照 2020-4-11).

[5] 消費者庁：若者が活用しやすい消費生活相談に関する研究会，入手先 < http://www.kantei.go.jp/jp/singi/it2/senmon_bunka/bousai/dai6/houkokusyo.pdf > (参照 2020-4-11).

[6] 奥村学：マイクロブログマイニングの現在，信学技報，電子情報通信学会，Vol. 111, No. 427，pp. 19-24，2012.

[7] 榊剛史，松尾豊：ソーシャルセンサとしての Twitter：ソーシャルセンサは物理センサを凌駕するか？，人工知能学会誌，人工知能学会，Vol. 27，No. 1，pp. 67-74，2012.

[8] Zhao, Q., Liu, T.Y., Bhowmick, S. and Ma, W.Y.:Event Detection from Evolution of Click-Through Data,Proc. *12th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining*, ACM, pp.484–493, 2006.

[9] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L.,Smolinski, M.S. and Brilliant, L.: Detecting InfluenzaEpidemics Using Search Engine Query Data, *Nature*, nature Vol.457, pp.1012–1014, 2009.

[10] 松尾 豊：ウェブからの実世界の観測と予測，電子情報通信学会論文誌 B，電子情報通信学会，Vol.J96-B，No.12，pp.1309‐1315，2013.

[11] Sakaki, T. Okazaki, M. and Matsuo, Y.. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on World Wide Web*, ACM pp. 851-860, 2010.

[12] 榊剛史，松尾豊，鳥海不二夫，篠田孝祐，栗原聡，風間一洋，野田五十樹：ソーシャルメディアを用いた災害検知及び被災地推定手法の提案，人工知能学会全国大会論文集，人工知能学会，Vol. 26，pp. 1-4，2012.

[13] Dingli, A. Mercieca, L. Spina, R. and Galea, M.. Event Detection Using Social Sensors. *Proceedings of the 2nd International Conference on Information and Communication Technologies for Disaster Management*. 2015, p.35-41.

[14] Zhao, S. Zhong, L. Wickramasuriya, J. and Vasudevan, V.: Human as Real-Time Sensors of Social and Physical Events:A Case Study of Twitter and Sports Games, *Rice University Technical Report*, Rice University, No. TR0620, 2011.

[15] 富田大志，道満恵介，井手一郎，出口大輔，村瀬洋：Twitter を用いたスポーツ試合中のイベント検出に関する検討, HCG シンポジウム 2012 論文集, 電子情報通信学会, pp. 492-498，2012.

[16] 長野伸一：ソーシャルセンサからの情報抽出技術，東芝レビュー，Vol. 69，No. 7, pp. 19-22，2014.

[17] Georgiou, T. Abbadi, A. Yan, X. and Georg e J.: Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application, *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ACM, pp. 330-335, 2015.

[18] Congosto, M. Fuentes, D. and Sanchez, L.: Microbloggers as Sensors for Public Transport Breakdowns, *Proceedings of the IEEE Internet Computing*, IEEE, Vol. 19, No. 6, pp. 18-25, 2015.

[19] Asur, S. and Huberman, B.: Predicting the Future with Social Media, *Proceedings 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, ACM, Vol. 1, pp. 492–499, 2010.

[20] 迫村光秋，和泉潔：twitter テキストマイニングによる経済動向分析，第 9 回人工知能学会ファイナンスにおける人工知能応用研究会資料，人工知能学会，pp. 39-41, 2012.

[21] Bollen, J. Mao, H. and Zeng, X.: Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1-8, 2011.

[22] Ruiz, E. Hristidis, V. Castillo, C. Gionis, A. and Jaimes, A.: Correlating Financial Time Series with Micro-Blogging Activity, *Proceedings of the 5th ACM International Conference on Web search and Data Mining*, ACM, pp. 513-522, 2012.

[23] 荒牧英治，増川佐知子，田瑞樹：Twitter Catches the Flu：事実性判定を用いたインフルエンザ流行予測， 音声言語情報処理研究報告，情報処理学会，Vol. 2011-SLP-86, No. 1，pp. 1-8，2011.

[24] Lampos, V., Bie, T. and Cristianini, N.: Flu Detector – Tracking Epidemics on Twitter. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2010*, Vol. 6323, pp. 599-602, 2010.

[25] 田中成典，中村健二，加藤諒，寺口敏生：マイクロブログの投稿時間に着目したユーザの職業推定に関する研究，情報処理学会論文誌データベース（TOD），情報処理学会，Vol. 6，No. 5，pp. 71-84，2013.

[26] 工藤航，鳥海不二夫：新聞記事のアクセスログを用いたユーザ属性の逐次推定，人工知能学会論文誌，人工知能学会，Vol. 34，No. 5，pp. 1-9，2019.

[27] 山下雄大，森純一郎：深層学習を用いた SNS プロフィール画像からの投稿者属性推定，人工知能学会全国大会論文集，人工知能学会，Vol. 30，pp. 1-4，2016.

[28] 堂前友貴，関洋平：半教師ありトピックモデルにより選択した地域語を用いた Twitter ユーザの生活に関わる地域の推定, 情報処理学会論文誌データベース（TOD），情報処理学会，Vol. 7，No. 3，pp. 1-13，2014.

[29] Chandra, S. Khan, L. and Muhaya, B.: Estimating Twitter User Location Using Social Interactions – A Content Based Approach. *Proceedings of the 3rd IEEE International Conference on Social Computing*, IEEE, pp. 838-843, 2011.

[30] 池田和史，中村健二，服部元，松本一則，小野智弘，東野輝夫：マーケット分析の

ための Twitter 投稿者プロフィール推定手法，情報処理学会論文誌コンシューマ・デバイス＆システム（CDS），Vol. 2，No. 1，pp. 82-93，2012.

[31] 森國泰平，吉田光男，岡部正幸，梅村恭司 : ツイート投稿位置推定のための単語フィルタリング手法，情報処理学会論文誌データベース（TOD），情報処理学会，Vol. 8，No. 4，pp. 16-21，2015.

[32] Twitter : Twitter，入手先<https://twitter.com/> (参照 2016-11-14).

[33] 加藤諒，中村健二，山本雄平，田中成典，坂本一磨 : マイクロブログにおけるユーザの属性と習慣行動の推定に関する研究，情報処理学会論文誌，情報処理学会，Vol. 57，No. 5，pp. 1421-1435，2016.

[34] 坂本一磨，中村健二，山本雄平，田中成典 : 平時と異なる事象に対するソーシャルセンシング技術に関する研究，情報処理学会論文誌，情報処理学会，Vol. 59，No. 10，pp. 1-14，2018.

[35] 坂本一磨，中村健二，山本雄平，田中成典，中村竜也 : ユーザ属性を考慮した平時と異なる事象に対するソーシャルセンシング技術に関する実践研究，知能と情報，日本知能情報ファジイ学会，Vol. 32，No. 1，pp. 556-569，2020.

[36] 田中成典，中村健二，寺口敏生，中本聖也，加藤諒 : マイクロブログから抽出したユーザの習慣に基づく行動推定に関する研究，情報処理学会論文誌データベース（TOD），情報処理学会，Vol. 6，No. 3，pp. 73-89，2013.

[37] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦 : 日本語語彙大系 CD-ROM 版，岩波書店，1999.

[38] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, JMLR, Vol.3, pp.993–1022 2003.

[39] Radim Řehůřek : gensim，入手先<https://radimrehurek.com/gensim/> (参照 2019-4-15).

[40] S21G 社 : ツイプロ，入手先<http://twpro.jp/> (参照 2016-11-14).

[41] Twitter : Twitter Developers，入手先<https://dev.twitter.com/> (参照 2016-11-14).

[42] ropross.net : Twilog，入手先<http://twilog.org/> (参照 2016-11-14).

[43] 国土交通省 : 国土数値情報ダウンロードサービス，入手先<http://nlftp.mlit.go.jp/ksj/>, (参照 2017.2.13)

[44] instant tools : 日本の地域分類，入手先<http://tools.m-bsys.com/data/area_classification.php>，（参照 2017.2.13）

[45] 藤本拓，原隆浩，西尾幸治郎 : 時系列の最適平滑化と動的な語彙集合を考慮した時系列文書に対するトピック解析手法，電子情報通信学会論文誌，電子情報通信学会，Vol. J96-D，No. 5，pp. 1212-1221，2013.

[46] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum.: Optimizing semantic coherence in topic models, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262-272, 2011.

[47] Radim Řehůřek : gensim，入手先< https://radimrehurek.com/gensim/models/ldamodel.html >，（参照 2019.4.15）

[48] Murzintcev Nikita : Select number of topics for LDA model，入手先< https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html >，（参照 2019.4.15）

[49] 斉藤裕樹，高山翼，山上慶，戸辺義人，鉄谷信二：マイクロブログのジオタグと発言コンテキスト解析による行動予測手法，情報処理学会論文誌，情報処理学会，Vol. 55，No. 2，pp. 773-781，2014.

[50] 伊藤淳，西田京介，星出高秀，戸田浩之，内山匡：Twitter と Blog の共通ユーザプロフィールを利用した Twitter ユーザ属性推定，研究報告自然言語処理（NL），Vol. 2013-NL-210，No. 4， pp. 1-8，2013.

[51] 総務省：国民のための情報セキュリティサイト，入手先<https://www.soumu.go.jp/main_sosiki/joho_tsusin/security/enduser/security02/05.html>（参照 2020-2-3).

[52] Twitter：Twitter Support，入手先<https://twitter.com/TwitterSupport/status/1141039841993355264>（参照 2020-2-3).

[53] Twitter：Twitter Japan，入手先<https://twitter.com/TwitterJP/status/1141851959818772481>（参照 2020-2-3).

[54] Twitter：ツイートに位置情報を追加する方法，入手先<https://help.twitter.com/ja/using-twitter/tweet-location>（参照 2020-2-3).

[55] 坂本一磨，山本雄平，中村健二，田中成典，中村竜也：類語の出現頻度に着目した居住地の推定に関する調査研究，第 34 回ファジィシステムシンポジウム講演論文集，日本知能情報ファジイ学会，Vol.34，pp.857-858，2018.

[56] Hayashi, T. Watanabe, S. Toda, T. Hori, T., Roux, J., Takeda, K.: Bidirectional LSTM-HMM Hybrid System for Polyphonic Sound Event Detection, *Detection and Classification of Acoustic Scenes and Events 2016*, No. TR2016-114, 2016.

謝辞

謝辞

本論文を取りまとめるにあたり，関西大学の教員の皆様および田中研究室の学生諸氏より，御多忙の中，終始一貫して，暖かく懇切丁寧な御指導御鞭撻，またすばらしく充実した研究環境を賜りました．中でも，研究全般に渡り，明確な方向付けまでして頂きました関西大学総合情報学部 田中成典教授に心より感謝の意を表しますと共に厚く御礼申し上げます.

本論文の研究を遂行するにあたり，関西大学総合情報学部総合情報学科 伊藤俊秀教授，林勲教授，辻光宏教授，及び吉田宣章教授には，終始多大なご協力と御支援を賜りました．深く感謝する次第であります.

大阪経済大学情報社会学部 中村健二教授，大阪工業大学情報科学部 山本雄平講師，関西大学先端科学技術推進機構 寺口敏生特命助教には，終始一貫した研究への御理解ならびに貴重な御助言を賜りました．深く感謝する次第であります.

法政大学デザイン工学部 今井龍一教授，関西大学環境都市工学部 窪田諭教授，琉球大学工学部 神谷大介准教授，摂南大学経営学部 塚田義典講師，大阪経済大学情報社会学部 井上晴可講師，関西大学先端科学技術推進機構 姜文渊特命准教授，関西大学先端科学技術推進機構 梅原喜政特命助教，関西大学先端科学技術推進機構 田中ちひろ特命助教，関西大学先端科学技術推進機構 中原匡哉特命助教には，研究活動に関する御助言だけでなく，研究者としての心構えについて御指導を賜りました．深く感謝する次第であります.

第 4 章の研究を遂行するにあたり，関西大学大学院総合情報学研究科 中村竜也氏（令和 2 年修士卒）に御協力を賜りました．深く感謝する次第であります.

最後に，勉学への意欲に対して深い理解と協力を頂いた父（坂本学）と母（坂本美智子），そして姉（山本彩）に感謝します.