情報処理学会論文誌　データベース　Vol.6　No.3　73–89 (June 2013)

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

田中 成典[1,a)]　中村 健二[2]　寺口 敏生[3]　中本 聖也[3]　加藤 諒[3]

受付日 2012年12月20日, 採録日 2013年4月11日

概要：携帯端末の普及にともない，ユーザの状況に応じて様々な情報をリアルタイムに提供するサービスに注目が集まっている．そのため，GPS から取得した位置情報や，マイクロブログの投稿内容からユーザの行動を推定する研究が行われている．著者らは，これらに加えて，新たにユーザの習慣的な行動に着目した推定手法について検討を行った．本研究では，マイクロブログにおけるユーザの投稿内容と投稿数の変化から行動のパターンを抽出し，指定した時間帯における習慣的な行動を推定する手法を提案する．この手法により，マイクロブログの投稿内容には行動に関する記述がない場合でも，指定した時間帯におけるユーザの行動を推定できる．実証実験では，投稿内容のみを用いた手法と習慣行動もあわせて考慮する本手法とを比較し，提案手法の有用性について検証した．

キーワード：行動推定，マイクロブログ，習慣行動，ライフログ

Research for Estimating Users' Activities Based on Habitual Behavior in Microblogs

Shigenori Tanaka[1,a)]　Kenji Nakamura[2]　Toshio Teraguchi[3]
Seiya Nakamoto[3]　Ryo Kato[3]

Received: December 20, 2012, Accepted: April 11, 2013

Abstract: Services to provide variety of information in real time with reference to users' situations are receiving attention, as portable terminals have become widespread. Accordingly, some studies are being made to estimate the users' activities from their location information obtained by GPS or from the contents of their microblog posts. In addition to these, the author and his colleagues examined a new estimation approach focused on the habitual behavior of users. The present study proposes an estimation method of users' habitual behavior within designated periods of time by extracting behavioral patterns from the changes in contents and numbers of their posts in microblogs. This method enables estimation of users' behavior within designated time periods without any behavioral description provided in their microblog posts. Our demonstration experiments compare a method of merely using posted contents with this method that also considers habitual behavior as well, and verify its usability.

Keywords: activity estimation, microblog, habitual behavior, life log

1. はじめに

情報技術の発達にともない，個々人が通信端末を携帯することが一般的となった．このため，通信プロバイダやアプリケーションを開発する企業は通信端末に各種センサを搭載し，実空間におけるユーザの行動や移動経路などのライフログ取得を試みている．その一方で，ブログやSNS（Social Networking Service）などに代表される様々

[1] 関西大学総合情報学部
Faculty of Informatics, Kansai University, Takatsuki, Osaka 569–1095, Japan
[2] 大阪経済大学情報社会学部
Faculty of Information Technology and Social Science, Osaka University of Economics, Osaka 533–8533, Japan
[3] 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University, Takatsuki, Osaka 569–1095, Japan
[a)] tanaka@res.kutc.kansai-u.ac.jp

な CGM（Consumer Generated Media）サービスが普及し，ユーザ自身がインターネット上に日記形式でライフログを投稿する事例が増加している．センサや CGM から抽出したライフログを用いてユーザの行動を適切に把握できれば，各ユーザが置かれている状況に応じた広告や情報の配信が可能となる．そのため，取得したライフログを解析し，ユーザの行動を分析，推定する技術や研究 [1] が注目されている．著者らは，特に生活時間に密接に関連した CGM の 1 つであるマイクロブログの情報を利用し，ユーザの行動を推定する手法の開発に取り組んでいる．マイクロブログとは，現在の状況や雑記などを短い文書として投稿するためのストリーム型メディアであり，他の CGM に比べてユーザの生活に密着した情報を取得することが可能である．このため，マイクロブログ上の情報を分析しユーザの行動を把握すると同時に，投稿内容からユーザの趣味嗜好を分析することで，1 人 1 人のユーザに対し，最適な広告や情報を適切なタイミングで配信できるようになる．

一般的に，ユーザの行動は習慣的な行動と非習慣的な行動の 2 種類に分けられる．習慣的な行動とは，睡眠や食事，通勤や勤務など，日常的に繰り返される行動である．一方，非習慣的な行動とは，会議や観光，ショッピングや旅行などの突発的ないしは非日常的な行動である．本論文では，これらの分類の中でも特に習慣的な行動（以下，「習慣行動」と略記）に着目する．ユーザごとに習慣行動のパターンを抽出できれば，それぞれの習慣行動時に適した情報を提供しやすくなる．また，習慣的なパターンとは異なる行動の場合でも，そのことを把握することで，GPS やその他の情報を駆使してユーザの状況を推定し，状況に合致した情報の推薦が可能になると考えられる．本研究では，投稿の記述内容を解析するアプローチの可能性を検討するため，著者らが実際の投稿記事を分析した結果，マイクロブログ上には習慣行動に関する投稿が少ないことが分かった．このため，記述内容のみからでは，適切に習慣行動を抽出することは難しいことが分かった．そこで，本研究では，習慣行動に関するわずかな投稿と前後の投稿数の変化とを関連付け，各ユーザの習慣行動のパターンを抽出する．そして，直近の投稿数の変化と習慣行動のパターンとの対応関係に基づき，指定した時間帯のユーザの行動を推定する手法を提案する．投稿数の変化に着目することにより，推定時において，ユーザ自身の行動についての言及が記述内容にない場合でも，習慣的なパターンに基づき行動推定が可能となる．

2. 関連研究

ライフログを解析しユーザ行動の推定を目的とする既存研究は，解析対象別に分類すると，「携帯端末のセンサ情報を解析する手法 [1], [2], [3], [4]」，「CGM に投稿された記述内容を解析する手法 [5], [6], [7], [8], [9]」と「センサ情報と CGM に投稿された記述内容とを解析する手法 [10], [11]」との 3 種類に分けられる．

携帯端末のセンサ情報を解析する既存研究では，GPS（Global Positioning System）から取得した位置情報を用いてユーザの行動を推定する手法 [1] が提案されている．これは，ユーザの現在地や移動経路に基づき，ユーザの行動やその目的を推定する手法である．しかし，GPS を用いる手法では，トンネル内や地下などにいる場合に位置情報が取得できず，行動を推定できないという問題がある．そこで，GPS 情報の欠損を補完する手法 [2], [3], [4] も研究されている．これらの手法は，主にユーザの現在の行動や直近の未来における行動を推定する際に有用である．

CGM に投稿された記述内容を解析する既存研究では，ブログを対象とする手法 [5], [6] やマイクロブログを対象とする手法 [7], [8], [9] が提案されている．ブログを対象とする手法では，投稿内の形態素間の係り受け関係に基づき，記述内容からユーザの行動を抽出する．しかし，総務省の報告書 [12] によると，9 割以上のブログユーザのブログ更新回数は週 7 回以下と述べている．このことから，ブログ記事の多くは生活時間に密着した情報ではないため，ライフログの取得先として活用することは難しいと考えられる．一方，マイクロブログは携帯端末を介して気軽に情報を投稿できるため，各ユーザの 1 日あたりの投稿件数が他の CGM に比べて多いという特徴がある．そこで，マイクロブログの記述内容を基にユーザの所在地を推定する手法 [8] や，記述内容から地理的特性に関係するトピックを抽出し，ユーザの現在地を絞り込む手法 [9] が提案されている．これらの手法は，非習慣的なイベントとそのイベントの場所とを関連付けてユーザの行動を推定する際に有用であると考えられる．

センサ情報と CGM に投稿された記述内容とを解析する既存研究では，GPS 情報がジオタグとして付加されたマイクロブログの投稿内容を解析する手法 [10] が提案されている．既存研究によると，ジオタグが登録されたマイクロブログの投稿は全体のうち 0.42% [8] であり，特定のキーワードが含まれているものに限れば 0.1% [11] と非常に少ないことが分かっている．しかし，ジオタグが付加された投稿は，記述内容とその時点におけるユーザの位置が関連付けられるため，その時々の行動を把握する際に有用である．

本論文では，これらの既存研究の手法に加えて，新たに「投稿数の変化」を解析対象として扱うことを提案する．既存手法で用いられる GPS 情報や記述情報に，本研究で提案する投稿数の変化の特徴を組み合わせることで，より大きな効果を発揮することが見込まれる．

3. 研究の概要

3.1 研究の目的

本提案手法では，時間ごとの投稿数の変化に基づきユーザの行動を推定する．これは，投稿数の変化がユーザの状

態を表す指標の 1 つとして利用できると考えたためである．投稿数の変化に着目すると，マイクロブログにアクセス可能な時間帯とアクセス不可能な時間帯を把握できる．この変化のパターンとユーザの行動とを関連付けて分析することで，行動ごとに特徴的な投稿数の変化のパターンが明らかとなる．この情報を活用すれば，行動に関する記述が投稿内容に不足する場合でも，投稿数の変化のパターンからユーザの行動を推定できると考えられる．そこで，本研究では，ユーザの習慣行動とマイクロブログへの投稿数の変化のパターンに基づき，指定した時間帯におけるユーザの行動を推定する手法について検討する．

3.2　研究の対象

3.2.1　研究対象の定義

本論文では，習慣行動に焦点を当ててユーザの行動を推定する．既存手法 [1] の行動の分類を調査したところ，「睡眠中」,「出勤中」,「勤務中」,「食事中」と「帰宅中」の 5 種類の行動が習慣的にとられるという結果が得られた．このため，本研究では以上の 5 種類の習慣行動に「その他」を加え，計 6 種類の行動に関する行動推定手法について検討する．これらの行動を推定できれば，出勤中の路線情報の提示や，帰宅中に食事処の情報を推薦するなど，ユーザ中心の新しい広告が実現できる．なお，習慣行動に焦点を当ててユーザの行動を分析する手法の第 1 段階として，本論文では時間間隔を 1 時間単位として解析した場合におけるユーザの行動推定手法について論じ，時間間隔の単位の適正値については，今後の発展研究で検討する．また，本研究では，マイクロブログへの投稿内容とユーザの行動の実態が一致しているという仮定に基づきデータを処理する．これは，オンライン経由でユーザの行動の実態を把握する手段が現実的に存在しないため，求めうる最大の近似値を取得する方法として，長期間にわたる投稿履歴の解析結果からの行動実態把握を試みるためである．

3.3　本研究の意義

本提案手法の必要性を確認するため，マイクロブログの中でもユーザ数の多い Twitter を対象に，習慣行動に関する投稿がどの程度含まれるかを分析する．分析では，ランダムに抽出した Twitter ユーザ 340 人の全投稿 993,528 件中に，3.2 節で研究の対象とした 6 種類の習慣行動に関する語句がどの程度の割合で含まれるかを調査する．習慣行動に関する語句には，本研究で作成した行動辞書（詳細は 4 章で記述）に登録された語句を用いる．習慣行動に関する語句を含む投稿の分析結果を**表 1** に示す．分析結果より，習慣的な行動に関する語句が投稿中に含まれる割合は全体の 13.78%であることが分かった．このことから，投稿中の記述内容に基づく行動抽出手法では，行動情報をわずかしか抽出できないことが明らかとなった．このため，本研究のように，記述内容のみに依存せずユーザの行動を推定する手法には有用性があることを確認できた．なお，各習慣行動の投稿割合について分析した結果，ユーザの中でも社会人は勤務に関する投稿をするユーザが多いことが分かった．このことから，ユーザの職業と投稿内容の間には何らかの関係があることが想定される．一方，食事に関する投稿は，ユーザの職業に依存せずに投稿されやすい傾向があった．これは，食事内容についての記述や画像が投稿されるケースが多かったためと考えられる．

表 1　習慣行動に関する語句を含む投稿の分析
Table 1　Analysis of posts with words concerning habitual behavior.

	抽出件数	割合
睡眠中	14,241	1.43%
出勤中	10,050	1.01%
勤務中	30,781	3.09%
食事中	31,037	3.12%
帰宅中	20,053	2.01%
その他	30,769	3.09%
全体	136,931	13.78%

3.4　提案手法の概要

本提案手法の流れを**図 1** に示す．図 1 に示すとおり，本提案手法は学習部と推定部の 2 つの処理部により構成されている．学習部は，行動確率モデル構築機能と投稿パターンモデル構築機能とで構成されている．推定部は，行動推定機能で構成されている．本提案手法は，推定対象時間と直近の投稿履歴とを入力することで，推定対象の時間帯におけるユーザの行動の推定結果を出力する．処理の流れを次に示す．

学習部の行動確率モデル構築機能では，マイクロブログから収集したユーザの投稿履歴を入力し，投稿の記述内容と投稿数の変化とを関連付けて行動確率モデルを構築する．行動確率モデルには，各曜日の各時間において，ユーザが過去にとった行動の確率（以下，行動確率と略記）が格納されており，投稿パターンモデルの構築時と時間帯に基づくユーザの行動推定時に利用する．

学習部の投稿パターンモデル構築機能では，マイクロブログから収集したユーザの投稿履歴を入力し，投稿パターンモデルを構築する．投稿パターンモデルには，投稿数の変化と 6 種類の行動の確率とを関連付けて構築した投稿パターンが格納されており，投稿パターンに基づくユーザの行動推定時に利用する．

推定部の行動推定機能では，行動を推定したい時間帯と直近の投稿履歴とを入力する．そして，学習部で構築した行動確率モデルと投稿パターンモデルとを参照し，指定さ

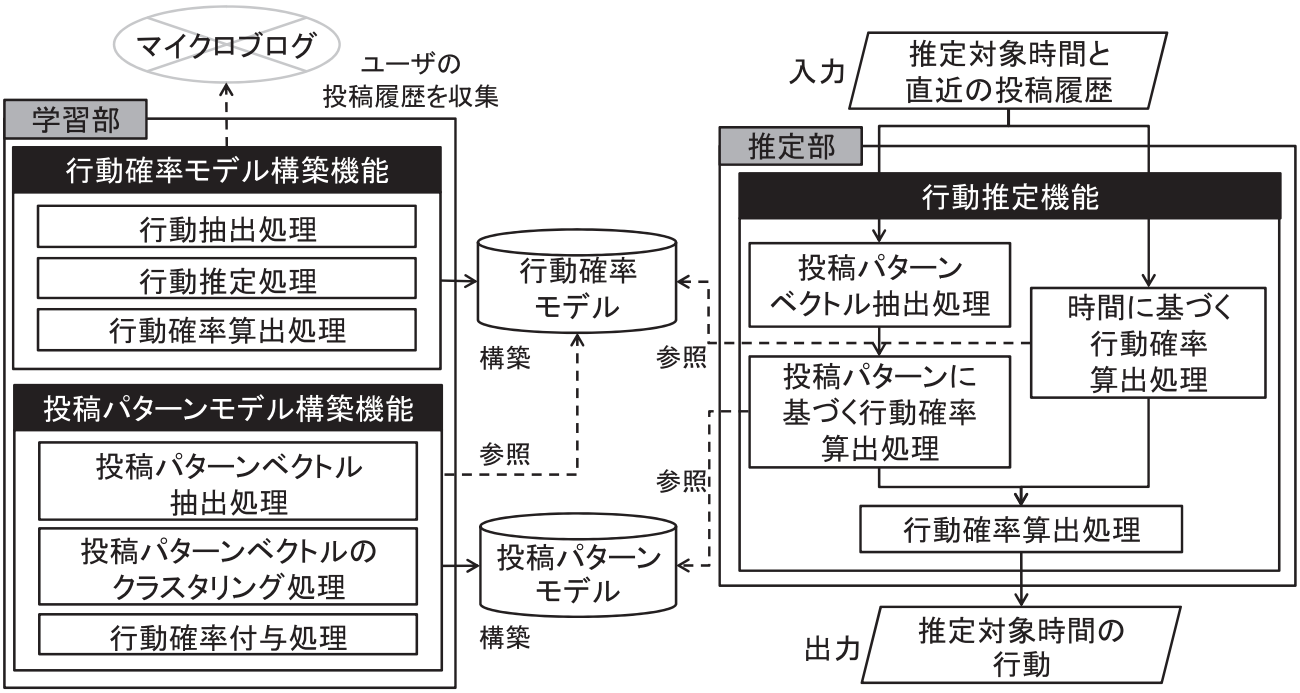

図 1　提案手法の流れ
Fig. 1　Flow of proposal method.

れた時間帯におけるユーザの行動を推定する．なお，本機能では，記述内容は参照せず，各時間帯における投稿数のみを解析対象とする．このため，提案手法による行動推定時には，ユーザの投稿に含まれる語句は活用しない．

4. 行動確率モデル構築アルゴリズム

4.1 行動確率モデル

行動確率モデルは，ユーザの過去の投稿を解析し，各曜日の各時間帯において，どのような行動をとっていたかを確率化したモデルである．しかし，語句のみを分析する手法では，睡眠中や勤務中など，投稿自体が少ない時間帯の行動確率モデルを構築できないと考えられる．そこで，本研究では，ユーザの投稿を分析し，行動情報を補完するアルゴリズムを考案した．本章では，行動確率モデルの構築アルゴリズムについて述べる．なお，各処理の解説ではサンプルユーザの解析結果を具体例にしながら，解説を進めるものとする．サンプルユーザの性別，職業，総ツイート数，1 日の平均投稿数とアカウント作成日からの経過日数を**表 2** に示す．

4.2 行動抽出処理

本処理では，習慣行動を分析するための基礎情報として，投稿履歴を解析し，各投稿からユーザの行動情報を抽出する．既存手法 [5] の行動抽出処理では，NTT コミュニケーション科学基礎研究所が作成した日本語語彙大系 [13] が用いられている．日本語語彙体系は，日本語の語句 30 万語が，品詞や意味，用法などに則り，ツリー構造で 3,000 カテゴリ，深さ 12 階層に体系化されたデータベースである．

表 2　サンプルユーザに関するデータ
Table 2　Data of sample user.

性別	職業	総ツイート数	1 日の平均投稿数	開始日からの経過日数
男性	社会人	35,852 件	40.6 件	883 日

本研究では，日本語語彙体系の生活・行動に関する項目を参考に，「睡眠中」，「出勤中」，「勤務中」，「食事中」，「帰宅中」と「その他」の 6 種類の行動に関する語句を選定すると同時に，必要に応じて手作業で語句を登録し，行動辞書を構築する．手作業で語句を追加したのは，日本語語彙体系に Twitter で用いられるような新しい語句が含まれていない問題や，分類項目が細分化されすぎており分類対象の習慣行動と関連付けがなされていない問題に対処する必要があったためである．行動辞書に登録した語句の一例を**表 3** に示す．

行動辞書中に登録された語句と投稿の記述内容との一致により，行動に関する情報が含まれた投稿を抽出する．この処理を，すべての投稿に対して行い，1 日の各時間帯における 6 種類の行動別の投稿数を抽出する．ここで，1 日の 24 時間を $t_1, t_2, \ldots, t_j, \ldots, t_{24}$，6 種類の習慣行動を $a_1, a_2, \ldots, a_k, \ldots, a_6$ とし，抽出した時間帯 t_j における習慣行動 a_k の投稿数を行動情報 $PostAction(t_j, a_k)$ と定義する．なお，本処理では，ユーザ本人の行動のみを抽出するため，他者に対する投稿である mention や他者の投稿を拡散する retweet などの投稿は抽出対象外とした．

表 3　行動辞書に登録した語句の一例
Table 3　Example of words registered into dictionary about activity.

行動	語句
睡眠中	寝る，就寝，眠り，おやすみ
出勤中	出勤，通勤，通学，行ってきます
勤務中	勤務，仕事，バイト，働く，残業，講義
食事中	食事，昼食，晩御飯，飲み会，食べる，
帰宅中	帰宅，帰る，退勤，退社，下校
その他	風呂，ほかる，テレビ，買い物，旅行

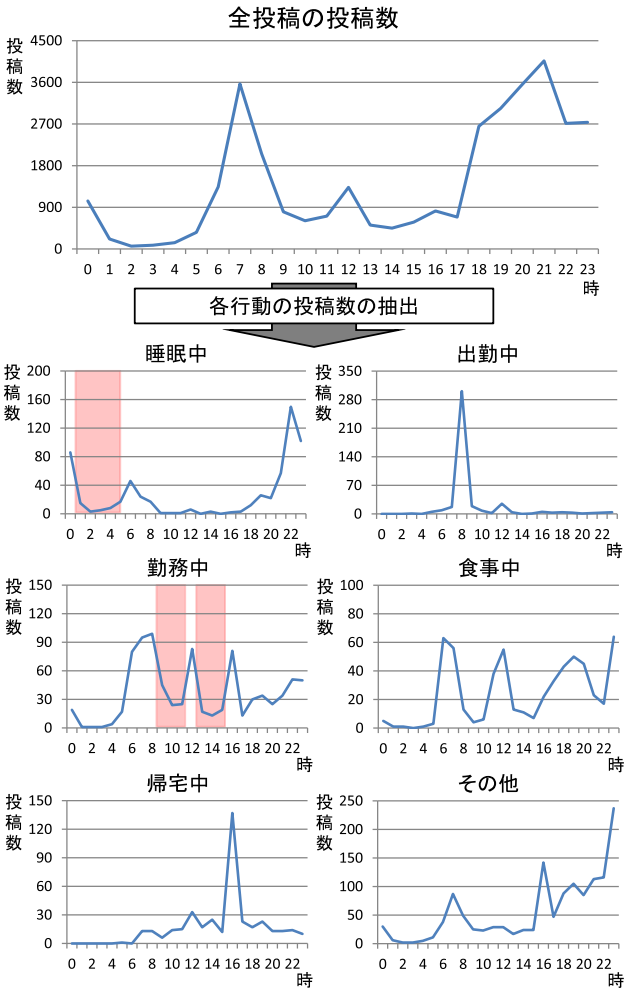

図 2　記述内容から抽出された各行動の投稿数
Fig. 2　Posts of each users' activity extracted from description contents.

4.3　行動推定処理

4.3.1　記述内容に基づき抽出した行動情報

行動抽出処理で記述内容から抽出したサンプルユーザの各行動の投稿数を**図 2** に示す．図 2 において，横軸は 1 日を 1 時間単位で 24 分割したものを表し，縦軸は 1 時間ごとの各行動に関する語句を含む投稿数を表す．また，色付きの箇所は該当する行動がとられたと考えられる時間帯を表す．図 2 に示すように，出勤中，食事中や帰宅中などの行動については，その行動がとられていると考えられる時間の投稿数が増加していることが分かる．このことから，これらの行動については記述内容からの行動抽出が可能であると考えられる．その一方で，色付きで示された時間帯については，睡眠中や勤務中などの行動や状態が継続されていることが想定される．しかし，これらの行動においては，記述内容や時間別の投稿数と行動内容との間に明確な関係はみられなかった．これは，睡眠中や勤務中にはマイクロブログへの投稿が行えないためと考えられる．そこで，本提案手法では，投稿がないことも行動を表す特徴の 1 つと考え，投稿数が少ない時間帯におけるユーザの行動情報を補完することを試みる．

4.3.2　投稿が少ない時間帯の行動の推定

本提案手法では，投稿の少ない時間帯のユーザ行動を推定するとき，投稿数の変化に着目する．これは，マイクロブログへの投稿の多い時間帯は，ユーザ本人が自由に管理できる時間帯であると仮定し，投稿の少ないもしくはない時間帯は，ユーザ本人には管理できない睡眠中や勤務中などの行動であると考えたためである．これらの行動は，複数の時間帯をまたいで継続されるという特徴がある．そこで，投稿数の少ない時間帯には，直前の時間帯の行動情報を付与することで，行動情報の不足を補う手法を試みる．

投稿が少ない時間帯と判断する基準として，本研究では暫定的に，時間帯別投稿数の平均値の 3 分の 2 よりも投稿数が少ない場合を対象とした．この条件に合致する状況が 1 時間以上連続している場合，投稿が少なく，直前の行動を継続していると判定し，$t_s, \ldots, t_f$ として抽出する．ここで，t_s は抽出した最初の時間帯を表し，t_f は抽出した最後の時間帯を表す．そして，抽出した $t_s, \ldots, t_f$ の行動情報 $PostAction(t_s, a_k), \ldots, PostAction(t_f, a_k)$ に対して直前の時間の行動情報 $PostAction(t_{s-1}, a_k)$ を加算する．このとき，投稿数の少なさをそのまま特徴としてとらえずに，直前の行動情報を加算したのは，$PostAction(t_s, a_k)$ を各時間帯で独立した要素として扱うためである．投稿数の増減を扱うためには，投稿を継続的に監視する必要があり，5 章で扱う投稿パターンの分析処理と同様の解析となる．そのため，本研究ではこれらの要素を独立して分析することを考え，$PostAction(t_s, a_k)$ を各時間帯で独立した要素として扱った．以上の処理により，投稿が少ない時間帯に対して行動情報を付与する．なお，付与する行動情報には，投稿が少ない時間帯にとられていると考えられる睡眠中か勤務中のうち，値が大きい行動を用いる．投稿数の変化に基づき補完したサンプルユーザの睡眠中と勤務中の投稿数を**図 3** に示す．図 3 において，色を塗った箇所は行動情報を補完した時間帯を表す．図 3 に示すように，投稿されていない時間帯の投稿数を直前の行動情報に基づき補完することによって，睡眠中と勤務中に関する行動情報が再評価されていることが分かる．

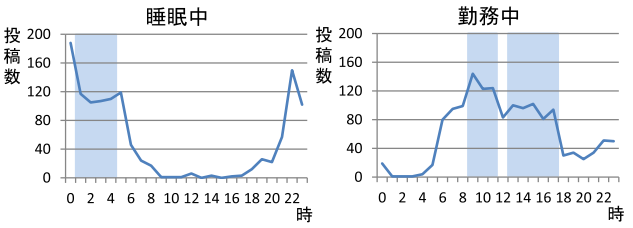

図 3　睡眠中と勤務中の投稿数の補完結果
Fig. 3　Complementary results of posts during sleep and work.

4.4　行動確率算出処理

本処理では，記述内容から抽出した行動と推定した行動の両方の情報を用いて，各時間におけるユーザの行動確率を算出する．このとき，出勤中や帰宅中など，行動の種類によっては投稿数に大きな偏りが発生することが考えられる．このため，投稿数のみに基づく行動確率算出手法では，投稿数の多い行動が相対的に高く評価され，正しく行動確率を評価できない．そこで，本研究では，tf-idf [14] を用いて投稿数を正規化し，出勤中や帰宅中などの 1 日を通じて特徴的な行動の確率を算出する．時間 t_j，行動 a_k の $tf(t_j, a_k)$ の算出式を式 (1)，$idf(t_j, a_k)$ の算出式を式 (2)，$tf\text{-}idf(t_j, a_k)$ の算出式を式 (3) に示す．

$$tf(t_j, a_k) = \frac{PostAction(t_j, a_k)}{\sum_{l=1}^{6} PostAction(t_j, a_l)} \tag{1}$$

$$idf(t_j, a_k) = 1 + \log\left(\frac{\sum_{m=1}^{24} PostAction(t_m, a_k)}{\sum_{o=1}^{24}\sum_{p=1}^{6} PostAction(t_o, a_p)}\right) \tag{2}$$

$$tf\text{-}idf(t_j, a_k) = tf(t_j, a_k) \times idf(t_j, a_k) \tag{3}$$

そして，算出した $tf\text{-}idf(t_j, a_k)$ のスコアを確率化する．時間 t_j，行動 a_k の行動確率 $P(t_j, a_k)$ の算出式を式 (4) に示す．

$$P(t_j, a_k) = \frac{tf\text{-}idf(t_j, a_k)}{\sum_{q=1}^{6} tf\text{-}idf(t_j, a_q)} \tag{4}$$

tf-idf によるサンプルユーザの行動確率 $P(t_j, a_k)$ を**図 4** に示す．図 4 に示すとおり，各時間の各行動の行動確率が正しく算出されており，それぞれの行動に特徴がみられることが分かる．

上記の一連の処理を各曜日に対して行い，各曜日の行動確率を算出する．ここで，各曜日を $w_1, w_2, \ldots, w_i, \ldots, w_7$ と定義し，曜日 w_i，時間 t_j，行動 a_k の行動確率を $P(w_i, t_j, a_k)$ と定義する．これは，現代社会の基本サイクルが 1 週間単位で繰り返されており，曜日ごとに行動が変化すると考えたためである．そして，行動確率 $P(w_i, t_j, a_k)$ を行動確率モデルに格納する．

5.　投稿パターンモデル構築アルゴリズム

5.1　投稿パターンモデル

ユーザの投稿履歴から抽出した投稿数の変化のパターンと行動確率モデルに格納された各曜日，各時間帯における

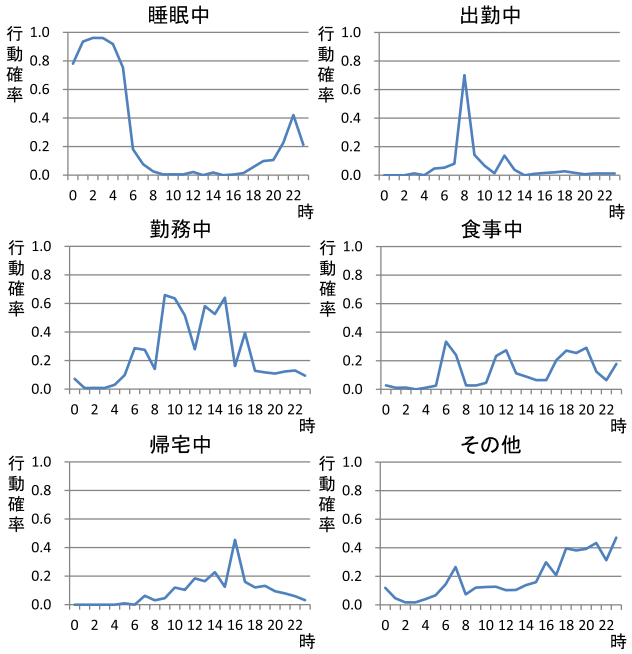

図 4　tf-idf による行動確率の算出結果
Fig. 4　Calculation results of action probability using tf-idf.

行動確率を関連付け，投稿パターンモデルを構築する．投稿パターンモデルの構築には，各時間を次元と見なし，各時間における投稿数を要素と見なす VSM（ベクトル空間モデル：Vector Space Model）[15] を用いる．これは，投稿数の変化のパターンを用いた習慣行動の抽出がユーザの行動推定に有用であることを検証するために，投稿パターンの抽出においては単純な手法を用いることが望ましいと考えたためである．

5.2　投稿パターンベクトル抽出処理

本処理では，ユーザの日々の投稿パターンを学習するために，ユーザの投稿履歴から指定した時間帯の投稿パターンベクトルを抽出する．投稿パターンベクトルの抽出結果を**図 5** に示す．図 5 では，まず，ユーザの投稿履歴から 1 時間単位の投稿数を集計する．そして，集計した各時間帯の投稿数に対して，抽出対象の日時を基準に，過去 n 時間分遡った時間帯までの投稿数を抽出し，投稿パターンとしてベクトル化する．ここで，投稿履歴におけるすべての日付を $d_1, d_2, \ldots, d_h, \ldots, d_z$ と定義し，日付 d_h，曜日 w_i，時間 t_j の投稿パターンベクトル $V_{d_h w_i t_j}$ を式 (5) で表現する．ここで，n は過去に遡る時間数を表す．

$$V_{d_h w_i t_j} = \{Post(d_h, w_i, t_j), Post(d_h, w_i, t_{j-1}), \ldots, Post(d_h, w_i, t_{j-n})\} \tag{5}$$

この投稿パターンベクトル $V_{d_h w_i t_j}$ をユーザの投稿履歴のすべての日時から抽出する．

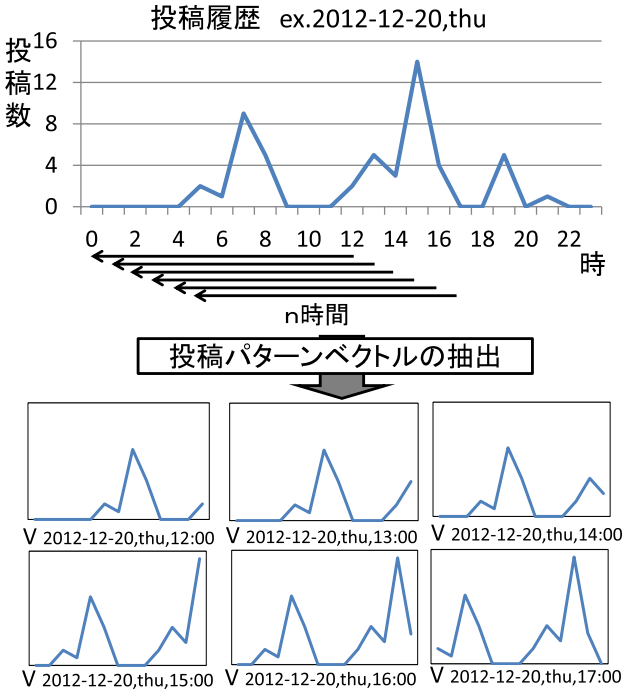

図 5　投稿パターンベクトルの抽出結果

Fig. 5　Extraction results of posted pattern vectors.

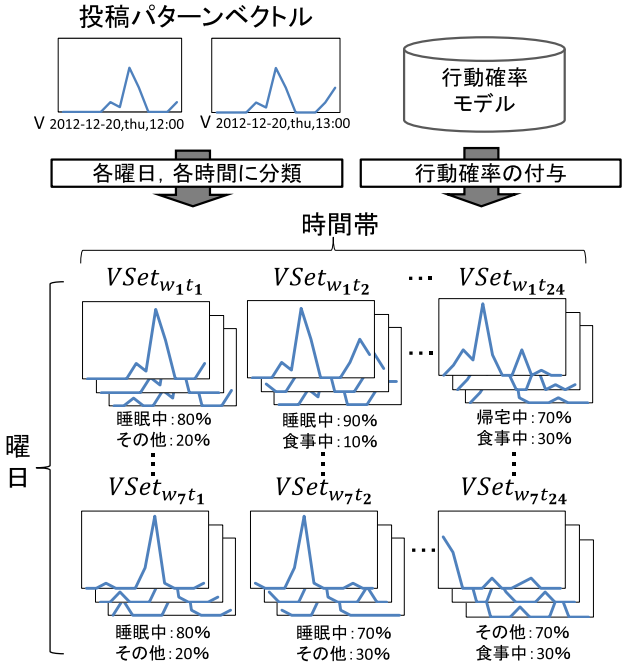

図 6　投稿パターンベクトルに対する行動確率の設定

Fig. 6　Setting action probability to posted pattern vectors.

5.3　行動確率付与処理

本処理では，行動確率モデルを参照し，抽出した投稿パターンベクトルに行動確率を設定する．投稿パターンベクトルへの行動確率の設定を図 **6** に示す．図 6 では，まず，投稿パターンベクトル $V_{d_h w_i t_j}$（図 6 左上）を各曜日，各時間に分類する．曜日 w_i，時間 t_j に分類した投稿パターンベクトル集合 $VSet_{w_i t_j}$（図 6 下）を式 (6) に示す．ここで，z はアカウント開始日からの経過日数を表す．

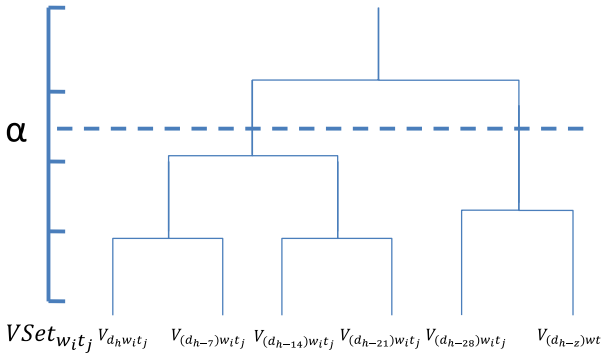

図 7　デンドログラム

Fig. 7　Dendrogram.

$$VSet_{w_i t_j} = \{V_{d_h w_i t_j}, V_{(d_{h-7})w_i t_j}, V_{(d_{h-14})w_i t_j}, \ldots, V_{(d_{h-z})w_i t_j}\} \tag{6}$$

そして，投稿パターンベクトル集合 $VSet_{w_i t_j}$ に行動確率モデル（図 6 右上）の同曜日，同時間における各行動の行動確率を付与する．これにより，投稿パターンベクトルに行動の意味付けを行う．

5.4　投稿パターンベクトルのクラスタリング処理

本処理では，類似する投稿パターンベクトルをクラスタリングにより類型化し，集約することで特徴的なベクトルを抽出する．まず，各曜日，各時間の投稿パターンベクトル集合 $VSet_{w_i t_j}$ を分析し，クラスタリングによって投稿パターンベクトル $V_{d_h w_i t_j}$, $V_{(d_{h-7})w_i t_j}$, $V_{(d_{h-14})w_i t_j}, \ldots,$ $V_{(d_{h-z})w_i t_j}$ を分類する．そして，分類したベクトルの平均値を求めることで，類似するベクトルを集約した特徴的な投稿パターンベクトルを抽出する．クラスタリングにより集約した投稿パターンベクトルをそれぞれ $Vc_1, Vc_2, \ldots, Vc_b, \ldots, Vc_y$ と表す．ここで，y はクラスタの数を表す．

本研究では，クラスタリングの手法として，階層的クラスタリングを採用する．クラスタリング手法には，階層的クラスタリングと非階層的クラスタリングがあるが，非階層的クラスタリングの代表例である k-means [16] では，クラスタリング精度が初期値に依存するという問題や分類するクラスタ数が固定であるなどの問題がある．そこで，投稿パターンの類似度に応じて，任意の数のクラスタを生成可能な階層的クラスタリングを採用した．階層的クラスタリングでは，データの分類が階層的になされ，その結果がデンドログラム（樹形図）で表される．デンドログラムを図 **7** に示す．

デンドログラムを任意の閾値 α で切ることで，分割するクラスタ数を操作する．クラスタ間の距離の算出には，階層的クラスタリングの代表的な手法である Ward 法 [17] を用いる．Ward 法は，クラスタ内のデータの平方和を最小にするように考慮した手法であり，分類感度が高いことが

知られている．Ward 法によりクラスタリングした投稿パターンベクトル Vc_b を投稿パターンモデルに格納する．これらの処理の流れにより，ユーザの習慣行動と投稿パターンベクトルとの関係を学習し，投稿パターンモデルを構築する．

6. 行動推定アルゴリズム

6.1 投稿パターンベクトル抽出処理

構築した行動確率モデルと投稿パターンモデルに基づき，行動推定機能によりユーザの行動を推定する．本処理では，指定された時間帯 x における投稿パターンベクトル $V(x)$ を抽出する．5.2 節の投稿パターンベクトル抽出処理と同様に，直近の投稿履歴に基づき投稿数を 1 時間単位で集計する．そして，集計した各時間帯の投稿数に対して，指定された時間帯から過去 n 時間前まで遡った各時間帯の投稿数を抽出し，推定対象の時間帯の行動を表す投稿パターンとしてベクトル化する．

6.2 投稿パターンに基づく行動確率算出処理

本処理では，抽出した推定対象の時間帯の投稿パターンベクトル $V(x)$ と投稿パターンモデルに格納された投稿パターンベクトル Vc_b を比較し，類似度が最も高いベクトルを抽出する．このとき，比較対象となる投稿パターンは，推定対象の日付と同じ曜日 w_x のベクトル $Vc_b(w_x)$ のみとする．これは，就業している平日と就業していない休日では，投稿パターンに明確な違いがみられたためである．ベクトルの類似度の算出には，コサイン尺度とユークリッド距離を用いた．コサイン尺度は，ベクトルの向きの近さを表す指標であり，ベクトルの向きが類似するほど距離が近くなる．ユークリッド距離は，ベクトルの長さの近さを表す指標であり，ベクトルの長さが類似するほど距離が近くなる．コサイン尺度 $Cos(V(x), Vc_b(w_x))$ の算出式を式 (7)，ユークリッド距離 $Euclid(V(x), Vc_b(w_x))$ の算出式を式 (8)，類似度 $Sim(V(x), Vc_b(w_x))$ の算出式を式 (9) に示す．ここで，n は入力ベクトルの次元数を表す．

$$Cos(V(x), Vc_b(w_x)) = \frac{V(x) \times Vc_b(w_x)}{\sqrt{V(x)^2} \times \sqrt{Vc_b(w_x)^2}} \quad (7)$$

$$Euclid(V(x), Vc_b(w_x)) = \sqrt{\sum_{i=0}^{n} (V(x) \times Vc_b(w_x))^2} \quad (8)$$

$$Sim(V(x), Vc_b(w_x)) = Cos(V(x), Vc_b(w_x)) \times Euclid(V(x), Vc_b(w_x)) \quad (9)$$

上記のように定義した類似度を投稿パターンモデルに格納された投稿パターンベクトルごとに算出し，その中で最も距離が近い投稿パターンベクトルの行動確率を投稿パターンに基づく行動確率 $Ppattern_{a_k}$ として算出する．なお，推定時には，記述内容は参照しない．

6.3 時間に基づく行動確率算出処理

本処理では，行動確率モデルに登録された行動確率 $P(w_i, t_j, a_k)$ に基づき，推定対象の時間帯の行動確率を算出する．投稿パターンのみに基づき行動を推定する場合，投稿が少ない場合に適切に行動確率を算出できないと考えられる．そこで，本提案手法では，過去の投稿履歴のみに基づき構築した行動確率モデルを用いて，推定対象の時間におけるユーザの行動確率を算出する．算出方法としては，行動確率モデルを参照し，推定対象時間の曜日 w_x と時間 t_x の行動確率 $P(w_x, t_x, a_k)$ を時間に基づく行動確率 $Ptime_{a_k}$ として算出する．なお，投稿パターンを用いる場合と同じく，本処理でも記述内容については参照しない．

6.4 行動確率統合処理

行動確率統合処理では，投稿パターンに基づく行動確率 $Ppattern_{a_k}$ と時間に基づく行動確率 $Ptime_{a_k}$ を組み合わせ，推定対象時間における各行動の行動確率 $Paction_{a_k}$ を算出する．行動確率 $Paction_{a_k}$ の算出式を式 (10) に示す．ここで，e は行動確率の組合せの重みを表し，e の値は 0.5 とした．これは投稿パターンと時間の行動確率は同程度に考慮すべきであると考えたためである．

$$Paction_{a_k} = (Ppattern_{a_k} \times e) + (Ptime_{a_k} \times (1-e)) \quad (10)$$

そして，算出した行動確率 $Paction_{a_k}$ のうち，最も確率の高い行動を推定対象時間の行動として出力する．

以上の流れでユーザの行動を推定することで，マイクロブログへの投稿数が少ない場合や投稿パターンのみでは行動が推定できない場合においても，既存の行動パターンをあてはめることにより行動情報を補完できる．

7. 実証実験

7.1 実験の概要

本実験では，本提案手法の有用性を確認するため，**図 8** に示す実験計画に従って，評価実験を行う．図 8 は，評価実験により検証する項目を明確化するため，図 1 と実験内容との対応関係を図に示したものである．それぞれの実験について概説する．

事前実験 1 では，「投稿パターンの時間数 n」の最適値を決定する．時間数 n は，投稿パターンモデル構築機能の投稿パターンベクトル抽出処理で，過去何時間前までを考慮して投稿パターンベクトルを生成するかを決定する値である．投稿履歴を考慮する時間の実験を行うことで，推定対象の時間の行動が，どの程度その前の行動に依存するかを明らかにする．事前実験 2 では，「投稿パターンのクラスタリング閾値 α」の最適値を決定する．閾値 α は，投稿パターンモデル構築機能の投稿パターンベクトルのクラスタリング処理で用いる階層的クラスタリングの閾値である．クラスタリング閾値の実験を行うことで，投稿パターン数

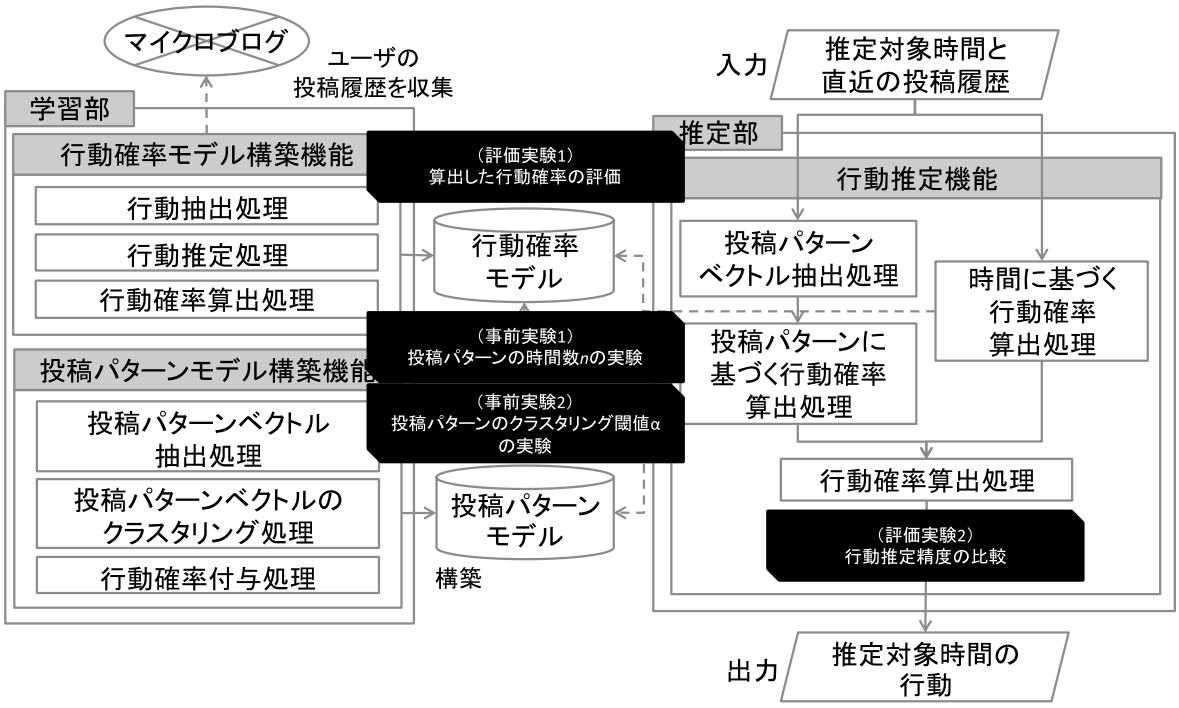

図 8　実験の流れ
Fig. 8　Flow of experiment.

と推定精度との関係を明らかにする．

評価実験 1 では，行動確率モデル構築機能の各処理の有効性を評価するため，それぞれの処理を組み合わせて行動推定精度を算出し，その結果を分析する．評価実験 2 では，推定対象時間と推定対象時間の直前の行動に基づき，適切にユーザ行動の推定が可能であるかを評価する．このとき，投稿パターンのみに基づく手法，過去の学習により取得したユーザの習慣行動に基づく手法とそれらを組み合わせた提案手法とを用いて行動推定精度を算出し，それらを比較することで本提案手法の有用性を評価する．

7.2　実験対象ユーザとテストデータ

本実験では，習慣行動に基づく行動推定手法の有効性を主張するため，習慣的に生活している可能性の高いユーザの代表例として，社会人を実験対象ユーザとした．また，学生が社会人になるなど，生活習慣が大きく変化したユーザの場合，過去の行動パターンを適切に学習できないため，生活習慣が大きく変わらないユーザのみを対象とした．

これに加えて，本研究では各ユーザの投稿履歴をすべて利用するため，Twitter そのものをデータの収集対象とする API を用いたクローリングでは限界がある．そこで，本実験では，ユーザごとに Twitter への投稿内容を保存するサービスである Twilog [18] から，ランダムに 187,178 人のユーザを取得し，それらのユーザの投稿履歴を収集した．この全履歴を解析したところ，ユーザの 1 日の平均投稿件数は 15.8 件であった．しかし，本研究では日々の投稿傾向を用いるため，1 日の平均投稿数が少ないユーザを対象とすることが困難であると考え，1 日の平均投稿件数が 10 件以下のユーザを対象外とした．Twilog から収集したユーザのうち，以上の条件に合致するユーザ 73,508 人の 1 日の平均投稿件数は 36.7 件であった．このことから，本研究では，暫定的に 1 日の平均投稿件数が 30 件以上のユーザを対象に実験を行った．また，1 日の平均投稿数と推定精度の関係を分析するため，1 日の平均投稿数が 30 件以上 150 件以下のユーザの中から，実験対象ユーザを選定した．

これらの条件に一致した 10 人のユーザを本実験の対象として採用する．プライバシ保護の観点から実験対象ユーザを A から J と仮称し，これら 10 人の投稿パターンを分析する．分析対象ユーザの概要を**表 4** に示す．実験対象の各ユーザの全投稿から，「習慣行動の抽出のための学習データ」と「評価実験のための正解データ」とを取得する．

習慣行動の抽出のための学習データは，対象ユーザの全投稿履歴から，評価実験のための正解データを除いたすべての投稿とする．Twitter の開始時期が異なるため，習慣行動の抽出に用いるデータ数は個々人によって大きく異なるが，これらの違いが推定精度にどのように影響するかについても，実験結果より分析する．

評価実験のための正解データは，実験対象ユーザがマイクロブログに投稿した内容を目視で確認し，睡眠中，出勤中，勤務中，食事中，帰宅中とその他の各行動に，正しく分類できたもののみを用いる．各行動の正解データの抽出ルールを次に示す．

・出勤中，食事中，帰宅中とその他

出勤中，食事中，帰宅中とその他の行動の正解データは，

表 4　分析データ
Table 4　Analyzing data.

ユーザ	性別	総投稿件数	1 日の平均投稿件数	開始日からの経過日数	行動に関する情報が含まれる割合
A	男性	29,328	32.2	909	14.77%
B	女性	26,290	32.6	806	7.00%
C	男性	32,089	34.3	933	11.70%
D	男性	35,852	40.6	883	12.66%
E	女性	62,373	58.8	1,059	10.01%
F	女性	62,705	65.7	954	14.79%
G	男性	58,690	66.5	882	8.29%
H	女性	65,892	83.5	789	8.31%
I	女性	87,887	93.6	938	5.17%
J	男性	120,712	115.7	1,043	7.46%

各行動に関する内容が投稿された日時を用いる．行動に関する内容の記述の有無は，4 章で構築した行動辞書に含まれる単語の有無で判断する．また，「今日は 8 時に出勤した」といった内容が 10 時に投稿された場合，投稿時間と内容との乖離が見られ，正確に評価することができないと考えられる．そのため，過去や未来に関する内容の投稿は正解データから除外し，現在の行動について記述していると判断できる投稿のみとした．

・睡眠中と勤務中

睡眠中と勤務中の行動は，マイクロブログにアクセスできず，行動を表現する投稿がない場合があると考えられる．そこで，これらの行動の正解データは，行動していると予測される時間帯を推定して取得する．睡眠中の正解データは，就寝に関する投稿と起床に関する投稿が一対となって存在し，さらにその間の時間帯に投稿が存在しない場合に，その時間帯を対象として取得する．勤務中も同様に，出勤や出社などの仕事の始まりを表す投稿と仕事終わりや帰宅などの仕事の終わりを表す投稿が一対となって存在し，その間に投稿が存在しない場合の時間帯を抽出する．ただし，お昼休憩などの食事中と判断された時間帯は除外する．

本実験では，上記の抽出ルールに該当した正解データとして，1 ユーザにつき約 300 件（約 50 件/行動）を 10 ユーザ分（合計 2,935 件）用意した．

7.3　事前実験 1：投稿パターンの時間数 n の実験

7.3.1　実験内容

本実験では，投稿パターンの時間数 n の最適値について実験する．実験では，投稿パターンの時間数 n を 4 時間間隔で 4 時間から 48 時間前まで変化させる．そして，それぞれの投稿パターンに基づき構築した投稿パターンモデルを用いて，実験対象データを判定することで，精度の違いを評価する．

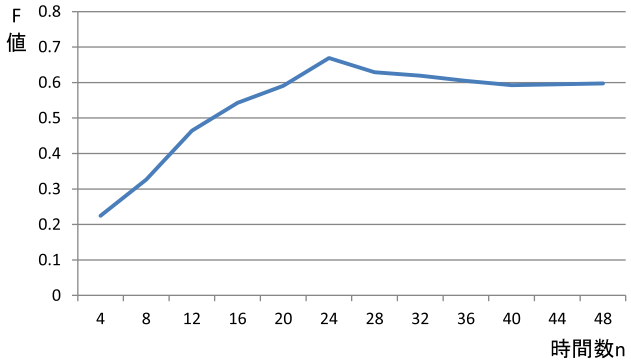

図 9　時間数と F 値との関係
Fig. 9　Relationship between F-measure and number of hours.

テストデータには，実験対象 10 ユーザの判定データ 2,935 件を用いる．なお，パラメータとして，閾値 α には暫定値として 2.0 を設定する．これらの実験データを用いて，次に示す手順により実験を行う．

STEP 1　各ユーザの判定データを除いた全投稿履歴を学習データとして，行動確率モデルと投稿パターンモデルを構築する．このとき，投稿パターンの時間数 n を 4 から 48 まで，4 時間間隔で変化させる．

STEP 2　6.2 節の投稿パターンに基づく行動確率算出処理により，判定データの行動を推定する．

STEP 3　判定データの行動と推定結果が一致する場合に正解とする．なお，本実験では，判定精度を F 値により評価する．

7.3.2　結果と考察

投稿パターンの時間数 n の推定精度の推移を**図 9** に示す．図 9 より，投稿パターンを考慮する時間が長くなるにつれて，習慣的な行動の特徴を反映できるため，推定精度が向上することが分かる．この実験結果から，本提案手法では，投稿パターンの n の時間数を最も精度の高かった 24 時間に設定する．遡る時間が短い場合に精度が低い原因は，睡眠中と勤務中において投稿数が同様に減少するが，それらの違いを考慮できないためであると考えられる．一方，遡る時間が 28 時間以上の場合も推定精度が低下した．これは，過去の習慣行動の影響を強く受けすぎてしまい，突発的な行動の変化に十分に対応できないことが原因と考えられる．

7.4　事前実験 2：投稿パターンのクラスタリング閾値 α の実験

7.4.1　実験内容

本実験では，投稿パターンのクラスタリング閾値 α の最適値について実験する．実験では，クラスタリングしない場合と階層的クラスタリングの閾値 α を 0.5 から 5.0 まで 0.5 刻みで変化させた場合の F 値を比較し，閾値 α と推定精度の関係を評価する．

テストデータには，実験対象 10 ユーザの判定データ

表 5　クラスタリング閾値 α の決定
Table 5　Decision of clustering threshold 'α'.

閾値 α	クラスタ数	F 値
0(クラスタリング無し)	101.1	0.564
0.5	20.8	0.674
1.0	15.7	0.669
1.5	13.6	0.672
2.0	**11.9**	**0.676**
2.5	10.9	0.666
3.0	10.1	0.659
3.5	9.6	0.655
4.0	9.2	0.647
4.5	8.7	0.642
5.0	8.4	0.638

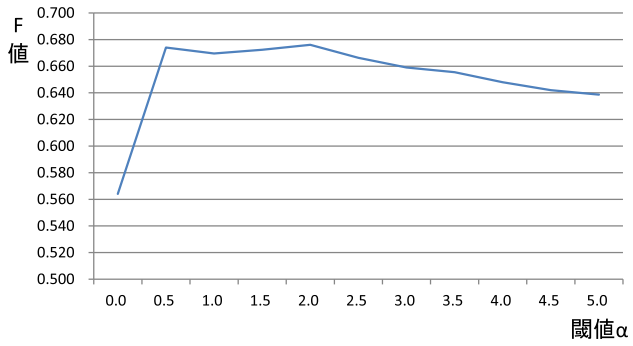

図 10　クラスタリング閾値 α と F 値との関係
Fig. 10　Relationship between F-measure and clustering threshold 'α'.

2,935 件を扱う．なお，パラメータとして，時間数 n には，7.3 節の実験結果に基づき 24 時間を設定する．これらの実験データを用いて，次に示す手順により実験を行う．

STEP 1　各ユーザの判定データを除いた全投稿履歴を学習データとし，行動確率モデルと投稿パターンモデルを構築する．このとき，投稿パターンのクラスタリング処理の閾値 α を 0（クラスタリングなし）から 5.0 まで，0.5 間隔で変化させ，クラスタリング結果を取得する．

STEP 2　6.2 節の投稿パターンに基づく行動確率算出処理により，判定データの行動を推定する．

STEP 3　判定データの行動と推定結果が一致する場合に正解とする．なお，本実験では，判定精度を F 値により評価する．

7.4.2　結果と考察

クラスタリング閾値 α の行動推定の実験結果を**表 5** に示す．表 5 において，クラスタ数は，クラスタリングされた $VSet_{w_i t_j}$ ごとのクラスタ数の平均値を表す．また，クラスタリング閾値 α の推定精度の推移を**図 10** に示す．表 5 と図 10 の結果から，クラスタリングを行わない場合 101.1 であったクラスタ数が，クラスタリングを行うことで類型化され，推定精度の向上につながることが分かった．また，階層的クラスタリングの閾値が 2.0 の場合，クラスタ数が 11.9 となり最も推定精度が良くなっていることが分かった．これは，推定対象時間の投稿パターンと比較する投稿パターンが類型化されていない場合，些細な違いに過敏に反応して，適切な類似ベクトルを取得できないためと考えられる．以上の実験結果から，本提案手法では，クラスタリングの閾値 α を最も精度の高かった 2.0 に設定する．また，各ユーザのクラスタリング結果を個別に確認してみたところ，各ユーザのクラスタ数に違いがみられた．これは，各ユーザの 1 日の平均投稿数や総投稿数の違いから，クラスタリングの分類度合いが変化したためであると考えられる．これにより閾値 α の最適値が変動することが考えられるため，今後は 1 日の平均投稿数や総投稿数を考慮して閾値 α の最適値を算出する手法についても検討する必要がある．

7.5　評価実験 1：算出した行動確率の評価

7.5.1　実験内容

本実験では，記述内容のみから算出した場合の行動確率，記述内容の抜けを推定しユーザ行動を補完した場合の行動確率，そしてユーザの行動確率を補正した場合の行動確率の 3 つの行動確率を算出し，本研究におけるユーザの行動確率算出手法を評価する．これらの実験を通じて，ユーザの行動確率モデルを構築する際の基盤情報であるユーザの行動確率の正しさを検証する．

テストデータには，実験対象 10 ユーザの判定データ 2,935 件を扱う．次に示す手順により実験を行う．

STEP 1　各ユーザの判定データを除いた全投稿履歴を学習データとして，4.2 節の行動抽出処理を用いた行動確率モデル，4.3 節の行動推定処理を用いた行動確率モデルと 4.4 節の行動確率算出処理を用いた行動確率モデルの 3 つの行動確率モデルを構築する．

STEP 2　3 つの行動確率モデルを用いて，6.3 節の時間に基づく行動確率算出処理により，判定データの行動を推定する．

STEP 3　判定データの行動と推定結果が一致する場合に正解とする．そして，正しく判定できた割合について，適合率，再現率と F 値により各手法の行動推定精度を評価する．

7.5.2　結果と考察

10 ユーザ全員の各行動を対象とした各手法の実験結果の平均を**表 6** に示す．表 6 において，下線を付けた箇所は各行動の各評価指標の中で最も精度が高かった手法を表す．実験結果を確認したところ，次に示す 4 項目の知見を得た．

- **行動確率算出処理が最も高精度であった**

実験により求めた F 値を確認したところ，すべての行動

表 6　行動確率の実験結果
Table 6　Experimental results of action probability.

行動	件数	行動抽出処理の結果			行動推定処理の結果			行動確率算出処理の結果 (tf-idf)		
		適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
睡眠中	500 件	0.707	0.474	0.568	0.712	**0.890**	0.791	**0.767**	**0.890**	**0.824**
出勤中	463 件	0.731	0.635	0.680	0.821	0.484	0.609	**0.832**	**0.758**	**0.793**
勤務中	504 件	0.289	0.325	0.306	0.488	**0.817**	0.611	**0.716**	0.754	**0.734**
食事中	494 件	0.511	**0.589**	0.547	0.647	0.567	0.604	**0.733**	**0.589**	**0.653**
帰宅中	489 件	0.460	0.534	0.494	0.540	0.495	0.517	**0.593**	**0.673**	**0.630**
その他	485 件	0.577	0.449	0.505	**0.690**	0.441	0.538	0.655	**0.606**	**0.630**
平均		0.546	0.501	0.517	0.650	0.616	0.612	**0.716**	**0.712**	**0.711**

において行動確率算出処理による結果は，行動抽出処理と行動推定処理により得られた結果よりも高精度に行動推定が可能であることを明らかにした．このことから，投稿内容や時間別の投稿数のみに基づく行動推定手法に比べ，本提案手法の 1 つである行動確率モデルにより行動確率を補正する処理の有用性を確認した．

- **投稿内容の解析に基づく行動抽出処理では，睡眠中と勤務中の推定が難しいことが分かった**

投稿内容の解析に基づく行動抽出処理による結果では，睡眠中の F 値が 0.568，勤務中の F 値が 0.306 となり，行動推定処理や行動確率算出処理により得られた結果と比べて推定精度が低くなった．これは，睡眠や勤務などの行動中には Twitter に投稿できないため，これらの行動を表す記述そのものが少ないことが原因と考えられる．その一方で，出勤中については，関連するキーワードを含む投稿が特定の時間に偏って出現しているため，行動推定時よりも高い精度で行動を抽出することができた．

また，記述内容が少ない場合を推定し行動情報を補完する行動推定処理の場合，平均の F 値は 0.612 となり，記述内容のみに基づく行動抽出処理による結果と比較して高い値を得ることができた．これは，睡眠中や勤務中など，記述内容に現れず，投稿そのものが少ない時間の行動について，前の時間の行動確率を引き継ぐ行動推定手法が効果を発揮したためである．

- **行動確率算出処理で投稿数の偏りを補正することにより推定精度が向上することが分かった**

tf-idf により記述内容の特徴度合いを評価し，それによって投稿数の偏りを補正する行動確率算出処理を実施すると，すべての行動において推定精度を向上させることができた．これは，各時間の発言内容の特徴が強調されて評価され，行動推定時においても良い影響を及ぼしたためであると考えられる．また，行動推定の結果とあわせて行動確率を算出したため，投稿がないという特徴についても適切に評価し，行動推定することが明らかとなった．

- **1 日のうちで複数回繰り返される行動の推定時における問題点と解決方法**

食事中と分類される時間は，行動抽出処理と行動確率算出処理の実験の結果，再現率が 0.589 と同等の精度となった．これは，食事中と判定される時間が「朝」，「昼」，「夜」以外にも多数抽出され，万遍なくどの時間にも確率が分布しているため，tf-idf による出現単語の特徴の補正が十分に効果を発揮しなかったためであると考えられる．この問題の解決方法として，食事を習慣行動として一定時間間隔で区別し，それぞれを別々の行動として学習することが考えられる．

以上の実験結果により，本提案手法による行動確率算出手法がユーザの行動を適切に推定できていることが分かった．このことから，本研究の基盤情報である行動確率モデルに登録されたユーザの行動確率の正しさを検証することができた．

7.6　評価実験 2：行動推定精度の比較

7.6.1　実験内容

事前実験で，投稿パターンモデルの構築に必要な変数の最適値を求めることができた．また，7.5 節で，本提案手法の基盤情報である行動確率の算出手法の適切さを検証した．これらの実験成果を基に，本実験では，過去の投稿傾向から抽出した習慣行動に基づきユーザの行動を推定する手法の有用性を検証する．

本実験では，投稿パターンのみに基づきユーザの行動を推定する手法，過去に抽出した習慣行動の時間情報のみに基づきユーザの行動を推定する手法，そしてそれら 2 つの手法により得られた結果を組み合わせる提案手法，計 3 つの手法の精度を比較する．

テストデータには，実験対象 10 ユーザの判定データ 2,935 件を扱う．なお，パラメータとして，時間数 n には，7.3 節の実験結果に基づき 24 時間を設定し，閾値 α には，7.4 節の実験結果に基づき 2.0 を設定する．これらの実験データを用いて，次に示す手順により実験を行う．

表 7　行動推定の実験結果
Table 7　Experimental results for estimating users' activities.

行動	件数	投稿パターン			時間（tf-idf）			投稿パターンと時間の組合せ		
		適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
睡眠中	500 件	0.719	0.900	0.799	**0.767**	0.890	0.824	0.739	**0.936**	**0.826**
出勤中	463 件	0.800	0.674	0.732	0.832	**0.758**	**0.793**	**0.858**	0.706	0.775
勤務中	504 件	0.684	0.738	0.710	**0.716**	0.754	0.734	0.696	**0.790**	**0.740**
食事中	494 件	0.717	0.545	0.619	0.733	0.589	0.653	**0.785**	**0.599**	**0.680**
帰宅中	489 件	0.577	0.671	0.621	0.593	0.673	0.630	**0.635**	**0.718**	**0.674**
その他	485 件	0.614	0.542	0.576	0.655	0.606	0.630	**0.711**	**0.612**	**0.658**
平均		0.685	0.678	0.676	0.716	0.712	0.711	**0.737**	**0.727**	**0.725**

STEP 1　各ユーザの判定データを除いた全投稿履歴を学習データとして，行動確率モデルと投稿パターンモデルを構築する．

STEP 2　判定データに対して，6.2 節の投稿パターンに基づく行動確率算出処理の手法（投稿パターン），6.3 節の時間に基づく行動確率算出処理の手法（時間），そしてそれら 2 つの手法を組合せる 6.4 節の行動確率統合処理の手法（投稿パターンと時間の組合せ）の 3 つの手法により，判定データの行動を推定する．

STEP 3　判定データの行動と推定結果が一致する場合に正解とする．そして，正しく判定できた結果を用いて，適合率，再現率と F 値により各手法の行動推定精度を評価する．

7.6.2　結果と考察

本実験の結果を**表 7** に示す．表 7 において，下線を付けた箇所は各行動の各評価指標の中で最も精度が高かった手法を表す．なお，表 7 の「時間（tf-idf）」項は，7.4 節の評価実験 1 で評価した行動確率算出処理の結果（tf-idf）と同じ解析であるため，表 6 の該当項目の実験結果と同値である．

表 7 の実験結果を確認したところ，次に示す 4 項目の知見を得た．

- **投稿パターンと時間の両方を組合せた手法が最も高精度に行動を推定できることが分かった**

投稿パターンと時間を組合せてユーザの行動を推定する本提案手法は，推定精度の平均値に着目すると，適合率 0.737，再現率 0.727，F 値 0.725 となり，どちらか片方の情報のみを用いた他の手法と比べて最も精度が良かった．これは，ユーザの日々の習慣的な行動と突発的な非習慣的な行動時間の揺らぎを考慮できたためと考えられる．

以上より，投稿パターンと行動がとられやすい時間を連動させてユーザの行動を推定する本提案手法の有用性を確認することができた．

- **投稿パターンと時間を比較すると，時間に基づく行動推定の精度が高いことが分かった**

時間に基づく手法の推定精度は，F 値の平均が 0.711 であり，投稿パターンのみを用いる手法よりも平均の F 値が 0.035 ポイント高かった．この結果から，人の習慣行動のパターンは一般的に時間に依存していることが分かった．これと同時に，社会人のように習慣的な生活を送る人間の行動は，過去の投稿履歴から抽出した習慣行動時間を用いることで，投稿パターンのみしか用いなかった場合に比べ，出勤中のような定時的な行動を高精度に抽出できることが明らかになった．

- **習慣的な行動の時間がずれるときに投稿パターンによる行動推定が効果的であることが分かった**

時間のみの手法による推定結果と時間と投稿パターンを組み合わせた手法による推定結果は，F 値の平均に着目すると 0.014 の差があった．そこで，時間のみを用いる手法では誤判定となった 834 件を分析したところ，816 件（98.74%）のデータは 2 時間以下の独立した行動であることが分かった．この原因として，時間による手法が効果を発揮するのは，習慣行動の中でも特に定期的な行動であることが関係していると考えられる．定時の始業時間が影響する「出勤」を除く「食事」，「帰宅」や「その他」のような 1 時間から 2 時間で終わる行動，あるいは「勤務中」や「睡眠中」の開始時間や終了時間の揺らぎに対応できず，誤判定が発生したものと考えられる．そして，時間のみでは吸収しきれない行動の揺らぎに投稿パターンによる推定が効果を発揮することで，時間で確率第 2 位だったものが，投稿パターンでの第 1 位との合算で 1 位に繰り上がって正解となり，結果として時間と行動パターンの組合せの精度が向上したと思われる．

その一方で，突発的な休養や出張などの行動変化に際しては，習慣的なパターンと異なるため適切に推定できないことが分かった．この問題については，現状の曜日で行動パターンを学習するのと並行して，投稿パターンが大きく異なる場合は，例外的な行動として判定する処理を追加することで解決可能であると考えられる．

表 8　詳細分析のための他ユーザのデータ
Table 8　Data of other users for detail analysis.

ユーザ	性別	職業	総ツイート数	1 日の平均投稿数	行動に関する情報が含まれる割合
K	女性	社会人	5,738	5.8	26.73%
L	男性	社会人	16,647	14.8	7.08%
M	男性	社会人	21,181	25.3	6.60%

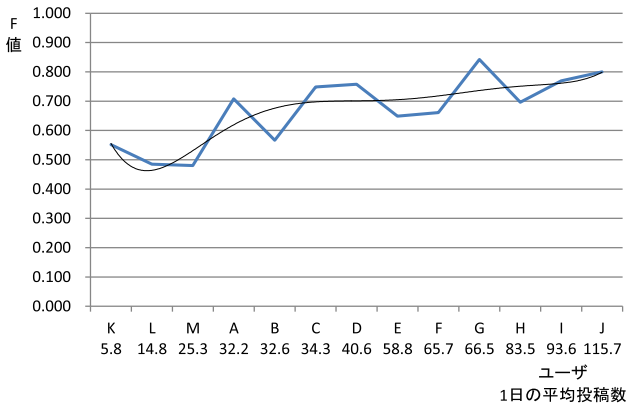

図 11　ユーザの 1 日の平均投稿数と F 値との関係
Fig. 11　Relationship between average of daily posts and F-measure for each user.

- **1 日の平均投稿数の増加にともない行動推定精度も向上する傾向がみられる**

1 日の平均投稿数と行動推定精度との関係を分析するために，全ユーザを対象にそれぞれの関係を分析した．なお，本実験では，習慣行動が取得可能であるユーザを対象とするため，1 日の平均投稿数を 30 件以上と設定していたが，全体の傾向を分析するため**表 8** に示す 30 件以下のユーザ 3 人（K，L，M）も含めて傾向を分析する．

全ユーザの各行動における推定精度を**表 9**，1 日の平均投稿数と行動推定精度の平均との関係を**図 11** に示す．また，図 11 には，各ユーザの行動推定精度を表す折線グラフに加えて，6 次の多項式近似グラフを表示する．図 11 の多項式近似のグラフを確認すると，1 日の平均投稿数の増加にともない行動推定精度も向上する傾向にあることが分かる．また，B のユーザ（1 日の平均投稿数 32.6 件）で多項式近似グラフの傾きが緩やかになっている．これに加えて，1 日の平均投稿数が 30 件以下のユーザは，いずれも投稿パターンによる推定精度が 50%未満となっている．以上の分析結果から，投稿パターンを抽出し正しくユーザの行動を推定するためには，少なくとも 1 日の平均投稿数が 30 件以上であることが望ましいと考えられる．

以上の実験結果により，記述内容を解析することなく投稿パターンのみを用いた場合でも，睡眠中や勤務中などの行動を高精度に取得できていることが分かった．また，帰宅時間が前後しやすいユーザの行動を推定する場合に，時間に基づく手法では正確な行動推定が行えないという問題があるが，投稿パターンを用いる手法により，この問題は解決できることが実験で明らかになった．このように，時間と投稿パターンの推定結果を組み合わせることで，投稿パターンのみを用いる場合や習慣行動時間のみを参照する場合に比べ，高精度にユーザの行動を推定できることから，本提案手法の有用性を実証した．

8.　おわりに

本研究では，各ユーザの投稿履歴を解析し，習慣行動を表す投稿パターンを抽出することによって，指定した時間のユーザの行動を「睡眠中」，「出勤中」，「勤務中」，「食事中」，「帰宅中」と「その他」の 6 種類の中から推定する手法について提案した．そして，実験により提案手法の有用性を検証した．本提案手法を用いることで，CGM や GPS などのライフログ情報が取得できない時間におけるユーザの行動情報を補完する推定技術が実現できると考えられる．しかし，同時に，新たに解決すべき問題として，ユーザの行動分類の詳細化や非習慣的な行動への対策，投稿パターン抽出時に用いるデータの範囲の設定などの項目が明らかとなった．

ユーザの行動分類の詳細化については，今回の実験結果においては，特に「食事中」に分類される行動の推定精度が低い問題への対策として必要である．「食事中」の推定精度が低くなった原因として，「食事中」と分類されている行動については，朝食，昼食，夕食のほかに間食などについての投稿が様々な時間で散見されたことから，キーワードとしての特殊性が強調されず，行動確率に反映されにくかったことが考えられる．また，「勤務中」に分類される行動の適合率が低くなった原因として，時間間隔の粒度が 1 時間単位であり，朝食，昼食や睡眠とで確率が分散してしまったことが考えられる．この問題に対しては，一定間隔で食事時間を分割し，朝食，昼食や夕飯など，行動分類を詳細化することで解決可能と思われる．

非習慣的な行動への対策については，本研究において「その他」に分類した行動の推定精度が低いという問題への対策として必要である．「その他」にはショッピングや風呂，勉強や電話，テレビ視聴など多くの行動が含まれており，時間別の分析や行動辞書による分析では適切な分類が行えていない．この問題に対しては，その他の行動を細分化し適切な行動辞書を自動的に更新する手法についても検討を行うことを考えている．また，「帰宅中」についても，残業などで不規則であることから，行動分類ごとに，投稿パターンの学習期間を調整することや，特徴付けを行うなどの対策が必要であると思われる．

投稿パターン抽出時に用いるデータの範囲の設定については，生活習慣が変わるようなイベントが発生した場合，投稿パターンの抽出に用いるデータを適切に選定する必要

表 9　各ユーザの 1 日の平均投稿数と行動推定精度の関係
Table 9　Relationship between average of daily posts and accuracy of estimating users' activities for each user.

ユーザ	推定手法	行動推定精度（F 値）						
		睡眠中	出勤中	勤務中	食事中	帰宅中	その他	平均
K	投稿パターン	0.783	0.528	0.463	0.240	0.291	0.378	0.447
(1 日の平均	時間	0.786	0.851	0.515	0.254	0.522	0.528	0.576
投稿数:5.8)	組合せ	0.778	0.736	0.585	0.240	0.442	0.529	0.552
L	投稿パターン	0.593	0.316	0.324	0.148	0.365	0.302	0.341
(1 日の平均	時間	0.817	0.561	0.511	0.298	0.508	0.368	0.511
投稿数:14.8)	組合せ	0.733	0.473	0.506	0.295	0.516	0.387	0.485
M	投稿パターン	0.500	0.160	0.389	0.390	0.385	0.280	0.351
(1 日の平均	時間	0.776	0.538	0.485	0.541	0.595	0.588	0.587
投稿数:25.3)	組合せ	0.634	0.190	0.495	0.462	0.515	0.586	0.480
A	投稿パターン	0.831	0.814	0.659	0.627	0.537	0.667	0.689
(1 日の平均	時間	0.868	0.844	0.786	0.698	0.581	0.667	0.741
投稿数:32.3)	組合せ	0.840	0.828	0.691	0.651	0.576	0.660	0.708
B	投稿パターン	0.770	0.500	0.619	0.514	0.509	0.297	0.535
(1 日の平均	時間	0.807	0.571	0.626	0.486	0.569	0.381	0.573
投稿数:32.6)	組合せ	0.748	0.575	0.656	0.514	0.561	0.347	0.567
C	投稿パターン	0.831	0.771	0.811	0.755	0.590	0.395	0.692
(1 日の平均	時間	0.916	0.926	0.878	0.800	0.655	0.636	0.802
投稿数:34.4)	組合せ	0.845	0.813	0.814	0.840	0.673	0.506	0.749
D	投稿パターン	0.796	0.738	0.626	0.725	0.630	0.532	0.675
(1 日の平均	時間	0.839	0.914	0.684	0.711	0.673	0.629	0.742
投稿数:40.6)	組合せ	0.825	0.960	0.677	0.759	0.652	0.674	0.758
E	投稿パターン	0.796	0.568	0.516	0.545	0.479	0.540	0.574
(1 日の平均	時間	0.727	0.854	0.540	0.506	0.549	0.485	0.610
投稿数:58.8)	組合せ	0.905	0.648	0.493	0.633	0.547	0.667	0.649
F	投稿パターン	0.736	0.793	0.769	0.500	0.651	0.480	0.655
(1 日の平均	時間	0.684	0.600	0.685	0.439	0.639	0.506	0.592
投稿数:65.7)	組合せ	0.684	0.732	0.748	0.532	0.724	0.545	0.661
G	投稿パターン	0.854	0.829	0.838	0.659	0.804	0.696	0.780
(1 日の平均	時間	0.883	0.883	0.902	0.769	0.688	0.674	0.800
投稿数:66.5)	組合せ	0.922	0.870	0.907	0.767	0.822	0.764	0.842
H	投稿パターン	0.699	0.500	0.795	0.624	0.642	0.606	0.644
(1 日の平均	時間	0.715	0.571	0.800	0.686	0.610	0.743	0.688
投稿数:83.5)	組合せ	0.709	0.522	0.824	0.667	0.714	0.742	0.696
I	投稿パターン	0.821	0.593	0.727	0.583	0.687	0.725	0.689
(1 日の平均	時間	0.877	0.583	0.851	0.704	0.696	0.716	0.738
投稿数:93.7)	組合せ	0.893	0.522	0.848	0.727	0.779	0.848	0.770
J	投稿パターン	0.929	0.909	0.830	0.630	0.681	0.703	0.780
(1 日の平均	時間	0.943	0.901	0.700	0.659	0.660	0.786	0.775
投稿数:115.7)	組合せ	0.980	0.909	0.860	0.645	0.688	0.716	0.800

があると考えられるためである．この問題に対しては，今後，ユーザの習慣行動を抽出するために必要なデータの範囲を明確化したうえで論じる必要があると思われる．

今後は，本提案手法の有用性を向上させるため，実験により明らかとなった問題点に対処する手法について研究する．また，本提案手法の実用化に際しては，目的に応じて情報推薦を行えるようにするため，本論文では 1 時間単位に固定していた時間間隔の粒度の適正値についても検討する必要があると考えられる．そこで，今後の研究では，分析する時間間隔の単位の最適値や習慣行動の抽出に必要な投稿数と行動の推定精度の関係などについての検証を行う予定である．また，本研究の発展として，複数のユーザ間の投稿傾向の類似性を評価し，共通する習慣行動の抽出手法についても検討することを考えている．これに加えて，本研究ではマイクロブログを対象としたが，提案手法は習慣的な行動を分析可能なあらゆるメディアへの応用可能性を秘めている．このことから，今後の発展として，マイクロブログ以外の新しいメディアへの利活用を含め，提案手法を拡張していくことを計画している．

参考文献

[1] 宮崎雄一郎，山田直治，住谷哲夫，磯谷佳徳：ユーザの行動に合わせたサービス実現のための行動推定技術の開発，NTT DoCoMo テクニカル・ジャーナル，Vol.17, No.3, pp.55–61, NTT ドコモ (2009).

[2] 原木　司，横山昌平，福田直樹，石川　博：GPS ログと Web 情報を用いた移動情報タグの生成，第 3 回データ工学と情報マネジメントに関するフォーラム論文集，日本データベース学会 (2011).

[3] 山田直治，礒田佳徳，南　正輝，森川博之：GPS 搭載携帯電話を用いた移動経路履歴に基づく訪問地・経由地予測システム，電子情報通信学会技術研究報告，Vol.110, No.130, pp.47–54, 電子情報通信学会 (2010).

[4] 山田直治, 礒田佳徳, 南　正輝, 森川博之：屋外行動支援のための GPS 搭載携帯電話を用いた移動経路の逐次的精錬手法，情報処理学会論文誌，Vol.52, No.6, pp.1951–1967, 情報処理学会 (2011).

[5] 倉島　健，藤村　考，奥田英範：大規模テキストからの経験マイニング，電子情報通信学会論文誌 D，Vol.J92-D, No.3, pp.301–310, 電子情報通信学会 (2009).

[6] 池田佳代, 田邊勝義, 奥田英範, 奥　雅博：Blog からの体験情報抽出, 情報処理学会論文誌, Vol.49, No.2, pp.838–847, 情報処理学会 (2008).

[7] グェンミンティ，川村隆浩，中川博之，中山　健，田原康之，大須賀昭彦：条件付確率場と自己教師あり学習を用いた行動属性の自動抽出と評価，人工知能学会論文誌，Vol.26, No.1, pp.166–178, 人工知能学会 (2011).

[8] Cheng, Z., Caverlee, J. and Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, *Proc. 19th ACM International Conference on Information and Knowledge Management*, pp.759–768, ACM (2010).

[9] Eisenstein, J., O'Connor, B., Smith, N. and Xing, E.: A Latent Variable Model for Geographic Lexical Variation, *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, pp.1277–1287, ACM (2010).

[10] 酒巻智宏，岩井将行，瀬崎　薫：マイクロブログのジオタグを用いたユーザの行動パターンの推定に関する研究，電子情報通信学会言語理解とコミュニケーション研究会研究報告，Vol.110, No.400, pp.37–42, 電子情報通信学会 (2011).

[11] 榊　剛史，松尾　豊：ソーシャルメディアからの人物目撃情報抽出システムの試作，第 25 回人工知能学会全国大会論文集，Vol.25, pp.2G1–3, 人工知能学会 (2011).

[12] 総務省：情報通信白書平成 18 年度版, p.42 (2006), 入手先 ⟨http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h18/pdf/i1050000.pdf⟩ （参照 2012-12）.

[13] 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).

[14] Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol.24, No.5, pp.513–523 (1988).

[15] Salton, G. and McGill, M.: Introduction to Modern Information Retrieval, McGraw-Hill College (1983).

[16] Mackay, D.: Information Theory, Inference and Learning Algorithms, Cambridge University Press (2003).

[17] Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol.58, pp.236–244 (1963).

[18] Twilog, 入手先 ⟨http://twilog.org⟩ （参照 2012-12）.

田中 成典　（正会員）

1963 年生．1986 年関西大学工学部土木工学科卒業．1988 年関西大学大学院工学研究科土木工学専攻博士課程前期課程修了．同年（株）東洋情報システム（現在，TIS）に入社．人工知能に関する研究受託開発業務に従事．1994 年関西大学総合情報学部専任講師．1997 年助教授，2004 年教授，2006 年から学生センター副所長，現在に至る．2002 年 8 月から 1 年間，カナダの UBC にて客員助教授．博士（工学）．専門は知識工学と社会基盤情報学．CAD/CG，GIS/GPS，画像処理および Web ソリューションビジネスに関する研究に従事．2000 年（株）関西総合情報研究所を起業，設立当初から現在まで取締役会長．2006～2012 年（株）フォーラムエイトの顧問．建設省土木研究所 CAD 製図基準検討委員会委員長，土木学会土木情報システム委員会幹事長，同委員会土木 CAD 小委員会委員長，ISO/TC184/SC4 国内委員等を歴任．現在，国土交通省日本建設情報総合センター社会基盤情報標準化委員会委員，同委員会 CAD/データ連携小委員会委員長，土木学会土木情報学委員会副委員長．主に，ISO に準拠した CAD 製図基準と CAD データ交換基盤の開発に従事．

中村 健二　（正会員）

1981 年生．2004 年関西大学総合情報学部卒業．2006 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．2009 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．同年関西大学ポスト・ドクトラル・フェロー，2010 年立命館大学情報理工学部助手，2012 年大阪経済大学情報社会学部准教授，現在に至る．博士（情報学）．知識情報処理，Web マイニング，テキストマイニング等の研究に従事．2002 年から（株）関西総合情報研究所で活動．システム設計，データモデル設計等の研究開発にて従事．電子情報通信学会，土木学会，日本データベース学会各会員．

寺口 敏生　（正会員）

1984 年生．2007 年関西大学総合情報学部総合情報学科卒業．2009 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．2012 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．博士（情報学）．

中本 聖也

1988 年生．2011 年関西大学総合情報学部総合情報学科卒業．2013 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．修士（情報学）．Web マイニングの研究に従事．

加藤 諒　（学生会員）

1989 年生．2012 年関西大学総合情報学部総合情報学科卒業．現在，関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程在学中．2012 年（株）関西総合情報研究所入社，現在に至る．システム設計等の研究開発に従事．

（担当編集委員　井上 創造）