情報処理学会論文誌 Vol.57 No.5 1421–1435 (May 2016)

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

加藤 諒[1,a)] 中村 健二[2] 山本 雄平[3] 田中 成典[4] 坂本 一磨[1]

受付日 2015年8月19日, 採録日 2016年2月8日

概要：マイクロブログや SNS（Social Networking Service）の普及により，ビッグデータがインターネット上に蓄積されている．このデータを活用して，投稿者の性別や年代，職業といった属性を把握し，その人々の興味や趣味，嗜好に合った販売促進の戦略がとられている．特に，投稿者の属性を獲得することを目的にしたパーソナルデータの推定に関する研究がさかんに行われている．この推定に関する既存研究では，投稿内容や投稿者のプロフィール情報だけでなく，投稿者のライフスタイルをも加味することで推定精度を上げるための手法が提案されている．しかし，投稿数や投稿記事そのものの量が少ない場合，パーソナルデータの推定精度が低下するという課題がある．そこで，この課題を解決するために抽象的なパーソナルデータを段階的詳細化の手順に基づき具象化する手法を新たに提案する．実証実験では，投稿者のパーソナルデータを推定する既存手法と本提案手法とを比較し，その有用性について検証する．

キーワード：マイクロブログ，Web マイニング，属性推定，行動推定，ライフログ

Research for Reasoning Users' Attributes and Habitual Behavior of Microblog

Ryo Kato[1,a)] Kenji Nakamura[2] Yuhei Yamamoto[3] Shigenori Tanaka[4] Kazuma Sakamoto[1]

Received: August 19, 2015, Accepted: February 8, 2016

Abstract: The spread of microblogs and SNS (social networking services) has made big data to be accumulated on the Internet. These data are utilized to employ sales promotion strategies to meet the interests, hobbies or tastes of the contributors by grasping their characteristics such as gender, age, or occupations. In particular, studies on estimating the contributors' personal data for the purpose of acquiring their characteristics are being actively pursued. Existing studies on such estimation propose some methods for improving the estimation accuracy not only by providing the posted contents and profile information of the contributors but also by taking their lifestyles into account. However, if the number of posts or the amount of the contributed articles themselves is small, there is a problem that the estimation accuracy of personal data decreases. To solve this problem, we propose a new method for specifying abstract personal data based on a stepwise refinement procedure. Demonstration experiments are conducted to verify its usability by comparing the proposed method with the existing method for estimating the contributor's personal data.

Keywords: microblog, Web mining, attribute estimation, behavior estimation, life log

1. はじめに

マイクロブログや SNS（Social Networking Services）の普及にともない，インターネット上に多様で膨大なデジタ

[1] 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
[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] 関西大学先端科学技術推進機構
Organization for Research and Development of Innovative Science and Technology, Kansai University, Suita, Osaka 564–8680, Japan
[4] 関西大学総合情報学部
Faculty of Informatics, Kansai University, Takatsuki, Osaka 569–1095, Japan
[a)] kato@kansai-labo.co.jp

ルデータ（以下，ビッグデータ）が蓄積されている．総務省の調査 [1] によると，2013 年のデータ流通量は，2005 年と比較して約 8.7 倍にまで拡大しており，増加の一途をたどっている．これらビッグデータを活用して，投稿者の性別や年代，職業といったユーザ属性を把握し，販売促進に活かす取り組み [2], [3]（既存サービス）がなされている．特に，マイクロブログを対象とし，投稿者のユーザ属性や習慣行動などのパーソナルデータを獲得する属性推定の研究 [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] がさかんに行われている．

投稿者のユーザ属性を推定する既存研究 [3], [4], [5], [6], [7], [8], [9] では，投稿内容から特性ごとの特徴的な単語の出現率を用いて推定する手法 [3], [4], [5], [6], [7], [8] が提案されてきた．しかし，投稿文字数が制限されるマイクロブログでは，ユーザ属性ごとの特徴を抽出できず，属性推定の精度が十分に確保できない課題がある．一方，ユーザ属性を推定する際に，投稿内容に加えて，マイクロブログの投稿時間などのユーザごとのライフスタイルを考慮する手法 [9] が提案されている．これらの研究 [3], [4], [5], [6], [7], [8], [9] では，属性ごとの推定精度の違いを考慮せず，事前にすべての教師データを用いて生成した各属性の推定モデルを用いて一様に処理をしている．たとえば，性別や職業などのユーザ属性を推定するためのモデルをすべての教師データを用いて生成しており，あらかじめ性別が明らかになっていた場合も，同様の属性推定がなされている状況である．

一方，投稿者の習慣行動を推定する既存研究 [10], [11], [12], [13] では，マイクロブログの投稿内容に加えて位置情報を解析する手法 [10], [11], [12] や投稿数の変化を解析する手法 [13] が提案されている．前者の研究 [10], [11], [12] では，投稿内容と移動経路を行動履歴として蓄積し，その後の移動先を推定する．しかし，マイクロブログに位置情報が付与されている数は全体の 0.42% [14] であり，ユーザの移動履歴を生成することが難しく汎用性が低い．後者の研究 [13] では，投稿内容と前後の投稿数の変化とを関連付けて習慣行動を抽出し，指定した時間帯の行動を推定している．しかし，マイクロブログを利用するユーザの投稿内容は様々であり，さらに投稿数もユーザによって大きな差があることから，行動推定の精度が投稿数や投稿記事の量に依存するという問題がある．

そこで，本論文では，推定対象ユーザの性別や年代，職業といった属性を推定し，その属性ごとのライフスタイルの違いを考慮した習慣行動の推定手法について検討する．ユーザの属性を推定する際は，既存研究の発展として，抽象的なパーソナルデータを段階的詳細化，いわゆる推定確率の高い属性から順に推定する手順に基づき具象化する手法により，ユーザの属性を推定する手法を新たに提案する．段階的詳細化の手法は，同一ユーザに対して複数の属性を推定する場合，かつ，それぞれの属性の推定精度が他の属性の推定結果に依存する場合に有効である．ここで，ユーザ属性を推定する際には，投稿内容に加えて，投稿時間を考慮した方が高精度に推定可能であるため，本研究の組み合わせ対象として既存研究 [9] を採用する．

本論文の構成は，以下のとおりである．2 章では，本研究の位置づけと既存手法の問題点の対応策について説明する．また，段階的詳細化の手法検討のために，投稿内容の解析に適した投稿件数の多いユーザを対象として，各手法の有効性を検証する．3 章では，本提案手法のアルゴリズムに関しての説明と各処理フローに関して述べている．4 章では，本提案手法の有用性を検証するための実験計画に関して述べている．5 章では，2.2 節の属性推定への段階的詳細化の適用方策の検討結果に基づき，実環境下において，段階的詳細化の手法が有効であるかを評価するため，判定ユーザ数を増加させて実験する．6 章では，5 章での属性の推定結果を用いた行動推定の評価実験を行う．7 章では，本研究のまとめを述べている．

2. 研究の概要

2.1 研究の目的と位置付け

本研究の位置付けを**図 1** に示す．著者らは，マイクロブログを用いたパーソナルデータ推定手法の基礎研究として，職業属性の推定に関する研究 [9] と習慣行動の推定に関する研究 [13] に取り組んできた．前者の研究 [9] では，ユーザ属性の推定時に投稿時間の情報を考慮することの有効性を証明した．また，後者の研究 [13] では，マイクロブログの投稿内容および投稿時間に関わる情報を解析することで習慣行動を推定することが可能であることを証明した．本研究では，これらの研究を進める中で明らかとなった「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」と「行動推定の精度が投稿数や投稿記事の量に依存するという問題」を解消する手法を提案することを目的と

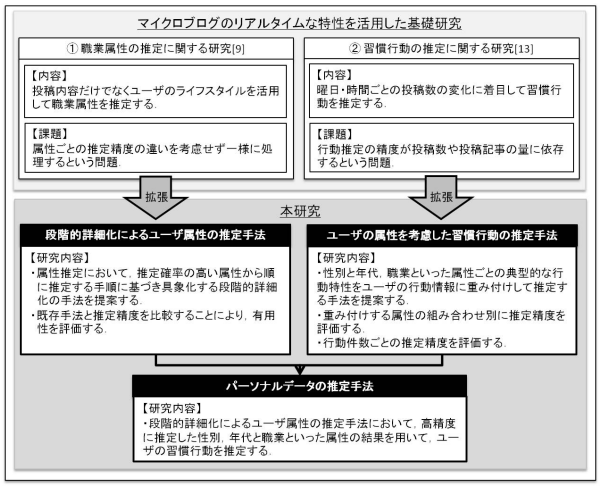

図 1 本研究の位置付け
Fig. 1 Aims to this present research.

設定する．そして，これら 2 つの研究の成果を活用し，高精度なパーソナルデータ推定手法を構築する．本研究における各課題への対応方策を次に示す．

2.1.1 「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」への対応策

ユーザ属性の推定に関する既存研究 [9] では，属性ごとの推定精度の違いを考慮せずに，すべてのユーザ属性の推定処理を一様に処理するという課題がある．たとえば，性別推定は男性と女性の 2 パターン，年代推定は 10 代，20 代，30 代，40 代以上の 4 パターンなど分類対象のパターン数も異なる．そのため，性別推定の精度の方が年代推定の精度よりも高くなる傾向が強い状況 [3] である．

そこで，本研究では，ユーザ属性の種類ごとに推定精度が異なるということに着目し，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たユーザ属性の推定手法の構築を試みる．具体的には，性別などの推定精度の高いユーザ属性を推定したのち，その属性が明らかになっている前提に基づき，他のユーザ属性を推定する手法を提案する．なお，本研究への段階的詳細化の適用方策については，2.2 節で詳述する．

2.1.2 「行動推定の精度が投稿数や投稿記事の量に依存するという問題」への対応策

ユーザの行動推定に関する既存研究 [13] では，1 日平均 30 件以上の投稿を行っているユーザのみを対象として実験を行っており，習慣行動を正しく推定するためには一定数以上の投稿を日常的に行っているユーザであることが前提条件となっている．しかし，日本人 Twitter ユーザ調査 [15] によると，1 日の平均投稿数は 25.7 件である．そのため，既存研究では，平均以上の投稿を行っていなければ習慣行動を推定できない状況である．

そこで，本研究では，この習慣行動推定時の投稿件数の制約を解消するため，同様のユーザ属性を保持するものは，同様の習慣的な行動を行うという仮説を設定し，課題解消を試みる．たとえば，社会人の男性であれば，朝出勤して，夜に帰宅するという一般的な社会人の特性や，夜に出勤して朝に帰宅するなどの夜勤の社会人の特性など，典型的な行動特性が見られると考えられる．そのため，ユーザ自身の行動情報に加えて，性別，年代と職業といったユーザの属性ごとの習慣行動の特性を考慮し，パーソナルデータを推定する手法を提案する．

2.2 属性推定への段階的詳細化の適用方策の検討

2.2.1 適用方針

段階的詳細化とは，抽象的な内容を段階ごとにより具象化する手法である．この手法は，一般的にソフトウェア工学の分野で用いられている手法であり，システム開発において，「決定すべき事象の要素を大まかに抽出」したのちに，「それぞれの要素を詳細化する」という 2 ステップを繰り返すことで，システムの詳細を明らかにする．本節では，この段階的詳細化の考え方を属性推定手法へ適用し，既存手法における「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」を解消可能かの傾向を確認する．段階的詳細化の具体的な適用方策を次に示す．

手法 1. ユーザごとの最も顕著となる属性を推定したのちに，その推定結果をふまえて，他の属性を段階的に推定する手法

手法 2. パーソナルデータの推定において，精度の高い属性（性別など）を明らかにしたうえで，その後，選択肢の多い属性（職業など）を段階的に推定する手法

2.2.2 段階的詳細化の手法検討

(1) 実験概要

本実験では，段階的詳細化の 2 つの適用方策を属性推定に適用した際の有効性を評価し，本研究で採用する手法を選定する．手法 1 の実験では，推定確率の高い属性から順に属性を決定する方策の有効性，手法 2 の実験では，属性の選択パターン数の少ないものから順に推定する方策の有効性を明らかにする．なお，本研究では，マイクロブログの中でも利用者数の多い Twitter を対象として実験を実施する．

実験データは，性別，年代と職業が判明している 1,180 ユーザを Web から収集した．実験データの収集方法を次に示す．

STEP 1：Twitter のプロフィールから任意の文字列を検索するサービスである「ツイプロ [16]」を用いて，Twitter のプロフィール欄や投稿内容に職業名を記載しているユーザを無作為に収集する．

STEP 2：STEP 1 で収集したユーザの投稿内容を TwitterAPI と Twitter に投稿された内容をユーザごとにブログ形式で保存するサービスである Twilog [17] を用いて収集する．なお，投稿内容の収集では，retweet 機能により発信された投稿を対象外とする．

STEP 3：STEP 2 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して収集する．TwitterAPI と Twilog は，ユーザが発信した最新の投稿から取得する仕様である．ここで，1 日の投稿件数が多いユーザは，ライフスタイルの解析に最低限必要な 1 週間分のデータを取得できない場合が考えられる．そのため，本実験の全体を通して，収集したデータの期間が 1 週間未満の場合は実験データから除外した．

STEP 4：ユーザ数が 1,180 件になるまで，STEP 1 から STEP 3 を繰り返し実施する．ただし，社会人の一部のユーザについては，既存研究 [13] と同様のユーザを用いた．収集した実験データの内訳を**表 1** に示す．

STEP 5：収集した職業が明確なユーザに対して，プロフィールや投稿内容を確認して，性別と年代を明らかにした．性別と年代が記載されていないユーザについ

表 1 実験データ
Table 1 Data of experimentation.

	分類	ユーザ数	収集期間
性別	男性	606 件	2012 年 7 月 2 日～2015 年 6 月 30 日
	女性	524 件	2012 年 7 月 5 日～2015 年 7 月 1 日
	不明	50 件	2015 年 2 月 25 日～2015 年 7 月 1 日
年代	10 代	314 件	2015 年 3 月 27 日～2015 年 6 月 30 日
	20 代	276 件	2012 年 8 月 30 日～2015 年 6 月 22 日
	30 代	329 件	2012 年 7 月 2 日～2015 年 7 月 1 日
	40 代以上	210 件	2015 年 2 月 25 日～2015 年 6 月 22 日
	不明	51 件	2015 年 2 月 25 日～2015 年 7 月 1 日
職業	学生	295 件	2015 年 3 月 27 日～2015 年 6 月 30 日
	社会人	295 件	2012 年 7 月 2 日～2015 年 3 月 16 日
	主婦（女性のみ）	295 件	2015 年 2 月 25 日～2015 年 6 月 22 日
	パート・アルバイト	295 件	2015 年 4 月 1 日～2015 年 7 月 1 日
	不明	0 件	—

ては，各判定モデルの構築において対象外とする．

本実験でのデータ件数は，学習データ 1,160 件と，判定データ 20 件とする．本実験では，段階的詳細化の手法検討を目的としているため，投稿内容の解析に適したユーザとして，投稿件数の多い順に採用した．

段階的詳細化の手法では，学習データを多く必要とするため，判定データを少なくして，手法 1 と手法 2 にどのような傾向があるのかを確認した．

【手法 1 の実験手順】

STEP 1：学習データを解析して，投稿される単語や投稿時間などの特徴をベクトル化し，教師あり学習を用いるパターン認識モデルの 1 つである SVM（Support Vector Machine）を用いて性別，年代と職業の推定モデルを構築する．

STEP 2：構築したそれぞれの推定モデルを用いて，判定データの性別，年代と職業を推定した結果と，その際の推定確率を取得する．なお，推定確率は，LibSVM [18] の Predict Probability 機能を用いて得た結果を採用する．

STEP 3：各属性の推定時に得られた推定確率を比較し，推定確率が最良な属性を決定する．

STEP 4：STEP 3 で決定した属性に一致する学習ユーザを取得し，LibSVM を用いて，推定モデルを再構築する．たとえば，STEP 3 で性別が男性と決定した場合，男性の学習データのみを用いて年代と職業を推定するモデルを構築する．

表 2 手法 1 と手法 2 の推定精度
Table 2 Estimation accuracy with method 1 and 2.

			適合率	再現率	F 値
既存手法	性別		0.7917	0.7917	0.7917
	年代		0.6479	0.6964	0.6713
	職業		0.7312	0.7000	0.7153
手法 1	性別		**0.9545**	**0.9375**	**0.9459**
	年代		0.4167	0.4857	0.4485
	職業		0.5542	0.5500	0.5521
手法 2	性別	年代	0.5000	0.4958	0.4979
		職業	**0.8092**	**0.7611**	**0.7844**
	年代	性別	0.7542	0.6958	0.7238
		職業	0.4867	0.3861	0.4306
	職業	性別	0.5722	0.6389	0.6037
		年代	0.3333	0.2130	0.2599

STEP 5：STEP 4 で構築したモデルを用いて，他のユーザ属性を推定する．なお，手法 1 では，属性推定後に最も精度の高い属性を決定し，その後，その結果に基づき他の属性を推定するといった 2 段階の属性推定を行う．3 段階での実験も可能であるが，段階を追うごとに学習データ数が減少して高精度なモデルを構築できないため，本実験では 2 段階までとする．

【手法 2 の実験手順】

STEP 1：属性の組合せごとに推定モデルを構築する．推定モデルは，性別，年代と職業を推定するための 3 モデル，性別を基準とした 4 モデルと年代を基準とした 8 モデル，職業を基準とした 8 モデル，計 23 モデルを構築する．

STEP 2：性別，年代と職業の属性において，1 つの属性のみ正解と仮定した場合にその属性のユーザを用いて構築したモデルを参照することで，他の属性の推定精度を算出する．なお，手法 2 では，1 つの属性が明確な場合に他の属性の推定を行うといった 2 段階の属性推定を行う．2 つの属性（社会人で 30 代の場合に性別を推定するなど）を用いた実験も可能であるが，手法 1 と同様に段階ごとに学習データ数が少なくなるため，本実験では 2 段階での実験を行う．

各手法のアルゴリズムの詳細は，既存研究 [9] を参照されたい．

(2) 実験結果

手法 1 と手法 2 の推定精度を適合率，再現率と F 値により評価した結果を**表 2** に示す．表 2 を確認すると，手法 1 は，性別の推定精度が向上したが，年代と職業の推定精度が低下していることが分かる．この原因としては，20 ユーザ中 4 ユーザにおいて，推定確率が最も高いと判断された属性が誤判定しており，そのデータを用いてモデルを再構

築したため，精度が低下したと考えられる．

手法 2 は，性別を考慮した職業の推定精度が向上した．これは，性別ごとに異なる職業の特徴が獲得できたからであると考えられる．たとえば，社会人では，平成 25 年の内閣府の男女別の職業別就業者調査 [19] を確認すると，特に卸売業・小売業や医療・福祉系は女性が多く，建設業や製造業は男性が多いなど，性別ごとの職種に異なりがあることが分かる．

ただし，その他の条件の場合は精度が低下した．これらは，年代と職業の属性数が多いことにより，学習モデルが的確に構築できなかったことが原因として考えられる．具体的には，性別を考慮した推定では，学習モデルを構築する際に学習データを 2 属性（男性と女性）に分けて構築するが，年代と職業では，それぞれ 4 属性（10 代，20 代，30 代，40 代以上など）に分ける必要があるため，学習データ数が少なくなり，精度が低下したと考えられる．

これらを整理すると，手法 1 では，性別の推定精度が向上し，手法 2 では，職業の推定精度が向上した．ただし，性別の推定は，投稿される単語の特徴だけを用いた手法だけでも高い精度 [3] が確認されている．一方，職業の推定は，本研究では 4 属性を対象としているものの，推定対象となる職業数が数多くあることからも有用性は高いと考えられる．

そのため，本研究の属性推定では，段階的詳細化の手法 2 を用いて，性別を考慮した職業属性の推定を行う．

3. パーソナルデータの推定アルゴリズム

3.1 アルゴリズムの概要

本提案手法の処理フローを**図 2** に示す．本提案手法は，性別，年代と職業といったユーザ属性の推定部とユーザの習慣的な行動の推定部とで構築される．

ユーザ属性の推定部は，属性推定モデル構築部と属性推定部とで構築される．属性推定モデル構築部では，同一属性でも多様なライフスタイルが存在することに対応するためのクラスタリング機能と投稿される単語と生活習慣の特徴を考慮するための特徴ベクトル作成機能，学習モデルを構築するための属性推定モデル構築機能で構成される．

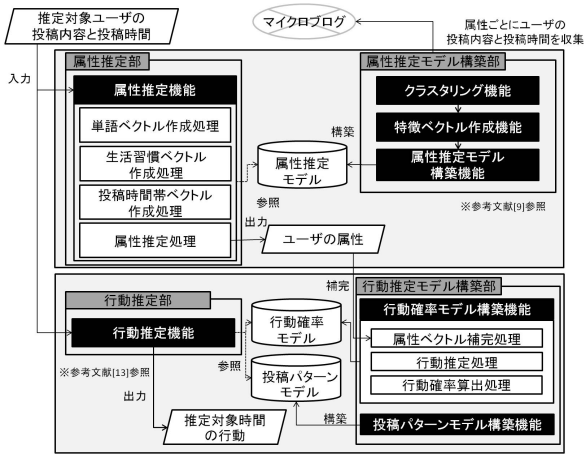

図 2 本提案手法の処理フロー

Fig. 2 Flowchart of process with proposed methods.

クラスタリング機能では，収集した各特性のユーザの投稿内容と投稿時間（以下，投稿履歴）から，曜日ごと（7 曜日）・時間帯ごと（24 時間）の投稿回数を示す投稿時間帯ベクトルを作成し，これらの類似性に基づきユーザ群をクラスタリングする．特徴ベクトル作成機能では，属性ごとの特徴的な単語の出現回数を示す単語ベクトルと生活習慣に関する単語の時間ごとの出現回数を示す生活習慣ベクトルを各クラスタ単位に作成する．属性推定モデル構築機能では，作成した単語ベクトルと生活習慣ベクトルを統合し，学習することで単語・生活習慣モデルを構築する．なお，単語・生活習慣モデルの構築では，高性能で実用的な LibSVM の Predict Probability 機能を用いた．また，クラスタリング機能により分類された各ユーザの投稿時間帯ベクトルの平均を示す中心ベクトルを取得して関連付けた投稿時間帯モデルを構築する．なお，投稿時間帯モデルの構築では，VSM（Vector Space Model）[20] を用いた．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．これらの単語・生活習慣モデルと投稿時間帯モデルを合わせて属性推定モデルと定義する．詳細は，既存研究 [9] を参照されたい．なお，この属性推定モデル構築部では，段階的詳細化の事前実験の結果から，性別推定モデルと年代推定モデル，性別を考慮した職業推定モデルを構築する．

属性推定部では，推定対象ユーザの投稿履歴を入力として，推定する属性ごとに単語ベクトルと投稿時間帯ベクトル，生活習慣ベクトルを作成する．そして，属性推定モデル構築部で構築した各属性推定モデルを参照することで，性別，年代と職業を推定する．これにより，「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」に対応する．詳細は 3.2 節で説明する．

ユーザの習慣的な行動の推定部は，行動推定モデル構築部と行動推定部とで構築される．行動推定モデル構築部では，ユーザの曜日・時間ごとの行動の傾向を示す行動確率モデルと投稿数の変化の特徴を示す投稿パターンモデルを構築する．行動推定モデルは，マイクロブログから収集した投稿内容から各曜日の各時間において，ユーザが過去にとった行動情報に属性推定アルゴリズムで推定した属性のライフスタイルの特徴を加えた確率（以下，行動確率）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．これにより，「行動推定の精度が投稿数や投稿記事の量に依存するという問題」に対応する．投稿パターンモデルは，投稿時間帯ベク

トルと行動確率モデルに格納された各曜日，各時間帯における行動確率を関連付けたものが格納される．詳細は 3.3 節で説明する．

行動推定部では，属性推定部と同様に推定対象ユーザの投稿履歴を入力する．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を参照し，指定された時間帯におけるユーザの行動を推定する．詳細は，既存研究 [13] を参照されたい．

3.2 属性推定部

属性推定部では，属性推定モデル構築部で構築したモデルを参照することでユーザ $target$ の属性を推定する．本処理部では，推定対象のユーザの投稿履歴を入力し，単語ベクトルと生活習慣ベクトル，投稿時間帯ベクトルを作成する．そして，作成したベクトルを使用して，属性推定モデル構築部で構築した属性ごとの推定モデル（性別，年代と職業を推定するための 3 モデル，性別を基準とした 4 モデルと年代を基準とした 8 モデル，職業を基準とした 8 モデル，計 23 モデル）を参照することでユーザの属性を推定する．属性推定部の処理手順を以下に示す．

STEP 1：投稿内容に出現する特徴的な単語 $word$ を素性とした単語ベクトル $V_{word}(target)$ を作成する．単語ベクトル $V_{word}(target)$ は，STEP 1.1 から STEP 1.4 の処理により作成する．

STEP 1.1：マイクロブログから収集したユーザの投稿内容に対して，MeCab を用いて形態素解析を行う．この際，投稿内容に含まれる顔文字や平仮名，片仮名 1 文字といったノイズを取り除くため，形態素が名詞のものを採用する．

STEP 1.2：性別，年代と職業の属性において，特徴的な単語を選定するため，x^2 値を使用して各属性における単語群のランキングを作成する．

STEP 1.3：ランキング上位の単語を素性として抽出する．

STEP 1.4：ユーザの投稿内容から STEP 1.3 で抽出した単語 $word$ を含む投稿がなされた回数を計算する．属性 $attribute$ の判定ユーザ $target$ における単語ベクトル $V_{word}(attribute, target)$ を式 (1) に示す．

$$V_{word}(attribute, target) = \{Post_{word_1}(target), Post_{word_2}(target), \cdots\} \quad (1)$$

式 (1) において，$Post_{word_1}(target)$ は，判定ユーザ $target$ の投稿における単語 $word_1$ の出現回数を表す．

STEP 2：ユーザの習慣行動を素性とした生活習慣ベクトル $V_{lifecycle}(target)$ を作成する．生活習慣ベクトルは，STEP 2.1 から STEP 2.2 の処理により作成する．

STEP 2.1：「睡眠中」，「出勤中」，「勤務中」，「食事中」，「帰宅中」と「その他」の 6 種類の各習慣行動に対して，日本語語彙大系 [21] を参考に手作業で行動に関連する用語を選定することで，行動辞書を作成する．行動辞書に登録した用語の例を**表 3** に示す．

表 3 行動辞書に登録した用語の例
Table 3 Example of terms on behavior dictionary.

行動	用語
睡眠	寝る，就寝，おやすみ，おはよう
出勤	出勤，通勤，通学，行ってきます
勤務	勤務，仕事，働く，残業，バイト，講義
食事	食事，昼食，晩御飯，食べる，飲み会
帰宅	帰宅，帰る，退勤，退社，下校
その他	風呂，テレビ，洗濯，買い物，旅行

STEP 2.2：ユーザの投稿内容から STEP 2.1 で選定した単語を含む投稿がなされた回数を計算する．なお，生活習慣ベクトルは，6 次元（習慣行動 $behavior$）× 24 次元（時間帯）の 144 次元で構成する．属性 $attribute$ の判定ユーザ $target$ における生活習慣ベクトル $V_{lifecycle}(attribute, target)$ を式 (2) に示す．

$$V_{lifecycle}(attribute, target) = \{Post_{behavior_1 0}(target), Post_{behavior_1 1}(target), \cdots, Post_{behavior_6 23}(target)\} \quad (2)$$

式 (2) において，$Post_{behavior_1 0}$ は 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間に判定ユーザ $target$ により生活習慣 $behavior_1$ に関連する単語を含む投稿がなされた回数を表す．

STEP 3：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 $V_{posttime}(target)$ を作成する．投稿時間帯ベクトルは，7 次元（曜日）× 24 次元（時間帯）の 168 次元で構成する．属性 $attribute$ の判定ユーザ $target$ における投稿時間帯ベクトル $V_{posttime}(attribute, target)$ を式 (3) に示す．

$$V_{posttime}(attribute, target) = \{Post_{Sunday0}(target), \cdots, Post_{Sunday23}(target), \cdots, Post_{Saturday23}(target)\} \quad (3)$$

式 (3) において，$Post_{Sunday0}(target)$ は日曜日の 0 時に判定ユーザ $target$ により投稿された件数を表す．

STEP 4：STEP 1 から STEP 3 で作成した単語ベクトル，生活習慣ベクトルと投稿時間帯ベクトルを用いて，属性推定モデル構築部で構築した各属性推定モデルを参照することでユーザの性別，年代と職業を推定する．

3.3 行動推定モデル構築部

行動推定モデル構築部では，属性ごとの特徴を考慮した行動確率モデルを構築する．行動確率モデルの構築手順を以下に示す．

STEP 1：属性推定部で作成したユーザの生活習慣ベクトル $V_{lifecycle}(attribute, target)$ に対して，属性推定部で推定したユーザの属性に対応した，属性の生活習慣ベクトルを補完する．補完手順を STEP 1.1 から STEP 1.3 に示す．

STEP 1.1：属性推定モデル構築部で作成した性別，年代と職業の各生活習慣ベクトルをそれぞれ正規化する．

STEP 1.2：正規化した各生活習慣ベクトルを同次元で加算（以下，属性ベクトル $V_{attribute}$）して，再度正規化する．

STEP 1.3：属性ベクトル $V_{attribute}$ とユーザの生活習慣ベクトル V_{user} を加算する．このときに，属性ベクトルの内容を考慮する割合を示すパラメータ e を用いる．属性ベクトルを加算した属性 $attribute$ の判定ユーザの生活習慣ベクトル $V'_{lifecycle}(attribute, target)$ を式 (4) に示す．

$$V'_{lifecycle}(attribute, target) = V_{lifecycle}(target) \times e + V_{attribute} \times (1 - e) \quad (4)$$

STEP 2：属性の特徴を補完した生活習慣ベクトル $V'_{lifecycle}(target)$ に対して，睡眠中や勤務中の時間といった投稿数の少ない時間帯に直前の行動情報を補完する．

STEP 3：tf-idf [22] を用いて，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重み付けして，行動確率モデルに格納する．

4. 実験計画

本研究で提案したパーソナルデータ推定手法の有用性を検証するため，性別を考慮した属性推定手法と属性を考慮した行動推定手法の有効性について評価実験を行う．本研究の実験計画を**図 3** に示す．図 3 は，評価実験により検証する項目を明確化するため，図 2 と実験内容の対応関係を図示したものである．

5 章では，性別を考慮した属性推定手法の有効性に関する評価実験として，ユーザの職業属性を推定する既存手法 [9] と提案手法とで算出した職業推定の精度を比較する．この実験により，「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」が解消可能であるかを確認する．

6 章では，本研究の行動推定手法の有効性を評価するため，既存手法 [13] との比較実験を 3 つ実施する．1 つ目の実験では，行動推定時に他の属性を考慮することの有効性を評価するため，ユーザ属性（性別，年代と職業）を用いる提案手法と，他のユーザ属性を用いない手法との比較を行う．2 つ目の実験では，投稿数の少ないユーザに対しても高精度に行動を推定可能であるかを評価するため，投稿数別の推定精度を算出して比較する．3 つ目の実験では，提案手法の実運用の可能性の検証と実運用へ適用するための制約条件を明らかにするため，属性推定手法の結果を用いて行動推定を実施した際の精度を評価する．これらの 3 つの実験を通して，既存手法における「行動推定の精度が投稿数や投稿記事の量に依存するという問題」が解消可能であるかを確認する．実験では，データを収集した時期が異なっても提案手法が適用可能なことを証明するため，既存研究 [9] で利用したデータを用いず，新たにデータを収集して実験を行う．

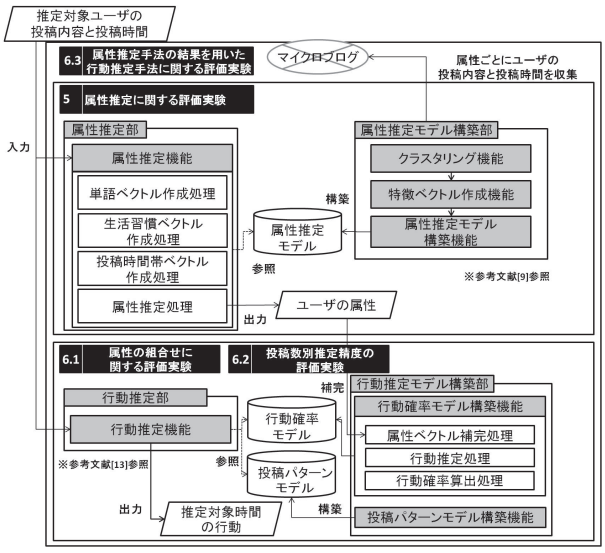

図 3 実験計画
Fig. 3 Plans of experimentation.

5. 属性推定に関する評価実験

属性推定に関する評価実験では，性別を考慮した職業推定の有効性について評価を行う．

5.1 実験概要

属性推定に関する評価実験では，2.2 節の属性推定への段階的詳細化の適用方策の検討の実験結果より，段階的詳細化の手法 2 を用いて，既存手法における「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」が解消できているかを確認する．そのため，提案手法と既存手法とで算出したユーザ属性の推定精度を比較して評価する．

5.2 実験条件

5.2.1 実験データ

実験データには，2.2 節で収集したデータセット（表 1）を用いる．本実験では，実環境下を想定した際の属性推定精度を評価するため，投稿件数は考慮せずに無作為に判定データを決定する．判定データの決定手順を次に示す．

STEP 1：各属性の実験データ 295 件から無作為に 40 件抽出する．

STEP 2：抽出した 40 件の性別と年代を確認する．

STEP 3：STEP 2 で取得したユーザの性別と年代が明ら

表 4 学習データおよび判定データの詳細
Table 4 Details of learning data and judgment data.

	分類	学習データ件数	投稿件数			判定データ件数	投稿件数		
			最大	最小	平均		最大	最小	平均
性別	男性	511 件	161,539 件	1,250 件	4,335 件	95 件	3,200 件	1,354 件	3,126 件
	女性	459 件	123,713 件	1,174 件	4,953 件	65 件	3,200 件	2,429 件	3,167 件
	不明	50 件	3,200 件	3,039 件	3,122 件	0 件	0 件	0 件	0 件
年代	10 代	273 件	123,713 件	1,250 件	4,441 件	41 件	3,200 件	2,383 件	3,138 件
	20 代	243 件	89,408 件	1,367 件	4,221 件	33 件	3,200 件	2,299 件	3,160 件
	30 代	273 件	161,539 件	1,174 件	5,459 件	56 件	3,200 件	1,354 件	3,123 件
	40 代以上	180 件	97,985 件	1,221 件	4,210 件	30 件	3,200 件	2,535 件	3,163 件
	不明	51 件	3,200 件	1,849 件	3,114 件	0 件	0 件	0 件	0 件
職業	学生	255 件	123,713 件	1,250 件	4,535 件	40 件	3,200 件	2,383 件	3,137 件
	社会人	255 件	96,934 件	1,326 件	4,050 件	40 件	3,200 件	1,354 件	3,119 件
	主婦	255 件	117,141 件	1,174 件	4,465 件	40 件	3,200 件	2,695 件	3,179 件
	パート・アルバイト	255 件	161,539 件	1,221 件	5,165 件	40 件	3,200 件	1,911 件	3,134 件

かではない場合は，そのユーザを除き，判定データ数が 40 件になるまで繰り返し実施する．

本実験で用いる学習データおよび判定データの詳細を**表 4** に示す．表 4 に示すとおり，投稿件数が多様なユーザを判定ユーザとして採用しているため，投稿件数に依存せずに精度を検証可能である．

5.2.2 パラメータの設定

パラメータの設定では，単語ベクトルの素性数，クラスタリング機能でのクラスタ数，SVM のカーネル関数および SVM のカーネル関数のパラメータを決定する．

単語ベクトルの素性数とクラスタ数は，既存研究 [9] の結果に基づき，それぞれ 256 件，2 クラスタとする．ただし，職業モデルのクラスタ数は，性別ごとに実験ユーザを分類して構築することでデータ数が減少し，適切なモデルの構築が困難であるため，クラスタリングは行わないものとする．

SVM のカーネル関数は，文書分類で一般的に用いられる線形カーネルを用いる．また，線形カーネルのコストパラメータ C には，職業推定の事前実験において，1〜2,048 間を 2 のべき乗に変更した値を設定して学習モデルを構築し，推定した際の最適値であった 512 を採用した．

5.3 実験手順

実験手順を以下に示す．

STEP 1：既存手法の職業推定の精度を評価するため，学習データ 1,020 件を用いて，職業モデルを構築する．

STEP 2：提案手法の職業推定の精度を評価するため，学習データを性別（男性と女性）に基づき分類する．そして，男性の学習データ 400 件を用いた男性用の職業モデル，女性の学習データ 160 件を用いた女性用の職業モデルを構築する．なお，SVM では各職業で学習データ数を揃えることで適切なモデルを構築することが可能である．今回の学習データでは，女性で社会人のユーザ数（40 件）が少なく，これに合わせて他の職業属性のユーザを学習に用いたため，女性の学習データは少なくなっている．

STEP 3：STEP 1 と STEP 2 で構築した各職業モデルを参照し，判定データの職業推定を実施する．なお，判定データも性別に基づき分類し，男性であれば男性用のモデル，女性であれば女性用のモデルを参照する．

STEP 4：既存手法の推定精度と提案手法の推定精度を算出し，比較する．推定精度については，適合率，再現率，F 値を用いて評価する．

5.4 結果と考察

既存手法と提案手法の推定精度を**表 5** に示す．表 5 より，次に示す内容が明らかとなった．

- **性別を考慮して職業を推定することで属性推定の平均精度が向上することが分かった**

提案手法の属性推定の平均精度を確認すると，提案手法の F 値が 0.7559 となり，既存手法の 0.7375 に比べて 0.0184 ポイント向上している．また，適合率では，0.0311 ポイント，再現率では，0.0063 ポイント向上していることから，本提案手法の有効性を確認できた．これにより，既存研究の課題である「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」に対応できたといえる．

表 5 既存手法と提案手法の推定精度
Table 5 Estimation accuracy with existing method and proposed method.

	職業	適合率	再現率	F 値
既存手法	学生	0.8000	0.9000	0.8471
	社会人	0.7500	0.5250	0.6176
	主婦	0.7333	0.8250	0.7765
	パート・アルバイト	0.6663	0.7000	0.6827
	平均	0.7374	0.7375	0.7375
提案手法	学生	0.7147	**0.9250**	0.8064
	社会人	**0.7680**	**0.7250**	**0.7459**
	主婦	**0.8571**	0.7500	**0.8000**
	パート・アルバイト	**0.7343**	0.5750	0.6450
	平均	**0.7685**	**0.7438**	**0.7559**

●職業ごとに推定精度が異なることが分かった

社会人と主婦において，職業の推定精度が向上した．特に，社会人の推定では，提案手法の F 値が 0.7459 となり，既存手法の 0.6176 に比べて 0.1283 ポイント向上している．これは，社会人の職種によって，使用する単語やライフスタイルが異なる可能性が高く，男性が多い職種と女性が多い職種を切り分けて推定することにより，性別ごとに社会人の異なる特徴を正確に取得できたためと考えられる．しかし，提案手法では，学生の F 値で 0.0407 ポイント，パート・アルバイトの F 値で 0.0377 ポイント推定精度が低下した．これは，学生とパート・アルバイトは，授業やクラブ活動，アルバイトなど男女で同じ内容に取り組むことが多く，性別ごとの違いが顕著に異なるような特徴が取得できなかったためと考えられる．実際に学生の投稿を確認すると男女ともに「部活おわった〜＼(ˆoˆ)／」や「今から部活、頑張りますか。」などの内容が共通して投稿されていた．また，学習データを男性と女性で分けたため，学習データの件数が少なくなり，推定精度が低下したと考えられる．以上のことから，社会人のように性別ごとに異なる特徴を持つ職業においては，提案手法が有効であることを確認した．

6. 行動推定に関する評価実験

行動推定に関する評価実験では，属性の組合せに関する評価実験と投稿数別推定精度の評価実験，属性の推定結果を用いた行動推定の評価実験を行う．

6.1 属性の組合せに関する評価実験

6.1.1 実験概要

属性の組合せに関する評価実験では，行動推定モデル構築部において，性別，年代と職業といった属性からどの属性，およびどの属性の組合せの特徴を用いることが，行動推定に有用であるかを評価する．評価は，属性ごとの特徴を用いない既存手法と提案手法との比較により行う．本実験では，有用な属性の組合せを正しく評価するため，判定ユーザの属性はあらかじめ把握している前提とする．また，属性ベクトルを考慮する割合を示すパラメータ e を投稿数ごとに事前に決定する．

6.1.2 実験条件

(1) 実験データ

実験データには，ライフスタイルが顕著に現れやすい社会人において，投稿数の多い上位 5 ユーザを採用した．学習データ 1 ユーザにつき 30,000 件を用いる．判定データは，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．各行動の判定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判定データは，各行動に関する内容が記述されている投稿日時を用いる．行動に関する内容の記述の有無は，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を対象とした．

・睡眠中と勤務中

睡眠中と勤務中の行動は，行動に関する内容が記述されない場合がある．そこで，これらの行動の判定データは，行動していると予測される時間帯を推定して取得する．睡眠中の判定データは，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判定データとして，1 ユーザにつき約 300 件（約 50 件/行動）を用意した．

(2) パラメータの設定

パラメータには，行動推定モデル構築部において，ユーザ自身の生活習慣ベクトルと属性ベクトルを組み合わせる際の重みに用いるパラメータ e を学習データの投稿数ごとに設定する．これは，ユーザ自身の投稿数が多い場合と少ない場合で属性ベクトルを補完する重みを変更する必要があると考えたためである．パラメータ e が高いと属性ベクトルの重みが低くなり，パラメータ e が低いと属性ベクト

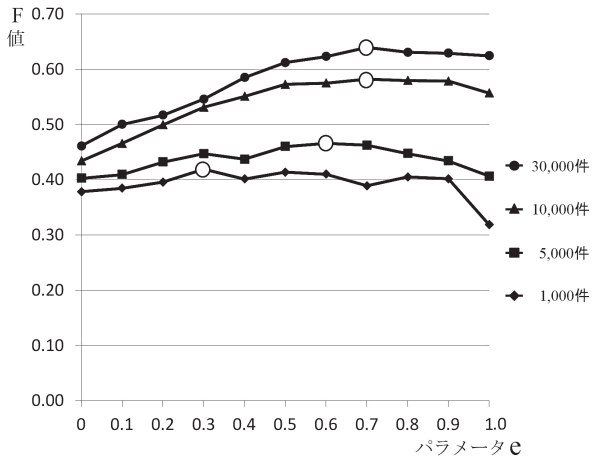

図 4 パラメータ e の F 値
Fig. 4 F-measure by parameter e.

ルの重みが高くなる．パラメータ e の決定手順を次に示す．

STEP 1：ユーザごとに 1,000，5,000，10,000，30,000 件の投稿をそれぞれ無作為に取得する．

STEP 2：属性推定モデル構築部で作成した性別，年代と職業ごとの生活習慣ベクトルを用意し，3.3 節 STEP 1 の処理を行うことで属性ベクトルを作成する．

STEP 3：行動確率モデル構築部の属性ベクトル補完処理において，パラメータ e の数値を変更し，最適値を確認する．なお，パラメータ e は，0 から 1.0 まで 0.1 ごとに変更する．

パラメータ e ごとの推定精度を**図 4** に示す．実験結果より，投稿数 1,000 件の場合，$e = 0.3$，5,000 件の場合，$e = 0.6$，10,000 件と 30,000 件の場合，$e = 0.7$ がパラメータ e の最適値であることが分かった．以上の結果より，30,000 件のデータを用いて行う本実験では，$e = 0.7$ に設定する．

6.1.3 実験手順

実験手順を以下に示す．

STEP 1：ユーザごとに 30,000 件の投稿を無作為に取得する．

STEP 2：属性推定モデル構築部で作成した性別，年代と職業ごとの生活習慣ベクトルを用意し，組合せごとに補完するベクトルを作成する．

STEP 3：行動確率モデル構築部の属性ベクトル補完処理において，属性の組合せごとに精度を算出し，比較する．推定精度については，F 値を用いて評価する．

6.1.4 結果と考察

実験結果を**表 6** に示す．表 6 の F 値より，職業だけの特徴を考慮する実験 C 以外の組合せで既存手法の推定精度が向上したことが分かる．特に性別，年代と職業の全属性を考慮する実験 G が最も精度が高いことから，以降の実験では，性別，年代と職業の 3 つの属性の特徴を考慮して実験を行う．これにより，職業ごとのライフスタイルだけでなく，ユーザの他の属性の特徴を組み合わせて習慣行動を推定する手法の有効性を確認した．

表 6 属性の組合せごとの手法による F 値
Table 6 Results of F-measure by methods combined with attributes.

実験	性別	年代	職業	F 値
既存				0.6224
A	○			0.6421
B		○		0.6395
C			○	0.6149
D	○	○		0.6406
E		○	○	0.6341
F	○		○	0.6349
G	○	○	○	**0.6607**

6.2 投稿数別推定精度の評価実験

6.2.1 実験概要

投稿数別推定精度の評価実験では，既存手法における「行動推定の精度が投稿数や投稿記事の量に依存するという問題」が，属性ごとの典型的な行動特性をユーザの行動情報に補完する提案手法で解消できているかを確認する．既存手法と提案手法において，学習データの件数を変化させて，各件数での行動の推定精度を比較して評価する．本実験では，属性ベクトルを付与すること提案手法の有効性を正確に評価するため，判定ユーザの属性はあらかじめ把握している前提とする．

6.2.2 実験条件

(1) 実験データ

本実験データには，投稿数の多い上位 20 ユーザを採用した．実験データは，1 ユーザにつき，学習データ 1,000，5,000，10,000，30,000 件を用意し，判定データ約 300 件を用いる．ただし，主婦やパート・アルバイトのユーザの中には，出勤中や帰宅中などに関する内容が投稿されていないユーザも含まれる．そのユーザについては，取得が可能であった判定データ数で実験を行う．判定データは，6.1 節と同様に全投稿履歴を人手で解析して設定した．実験では，上記の抽出ルールに該当した判定データとして，1 ユーザにつき約 300 件（約 50 件/行動）を投稿数の多い上位 20 ユーザ分（合計 5,352 件）用意した．

(2) パラメータの設定

パラメータには，パラメータ e とユーザの行動情報に重み付けする属性を設定する．パラメータ e の値は，6.1.2 項の結果より，投稿数 1,000 件の場合，$e = 0.3$，5,000 件の場合，$e = 0.6$，10,000 件と 30,000 件の場合，$e = 0.7$ をパラメータ e の最適値として採用する．なお，既存研究の推定精度は，パラメータ e の値を 1.0 に設定し，属性ベクトルを補完せずに算出する．また，ユーザの行動情報に重み

表 7 行動推定に関する既存手法と提案手法の F 値
Table 7 F-measure with existing method and proposed method concerning behavior estimation.

			睡眠中	出勤中	勤務中	食事中	帰宅中	その他	平均
既存手法	学生	1,000 件	0.3048	0.3206	0.1116	0.1728	0.2167	0.2283	0.2258
		5,000 件	0.3536	0.2745	0.2802	0.2014	0.2730	0.2188	0.2669
		10,000 件	0.5526	0.2195	0.1989	0.2919	0.3382	0.2339	0.3058
		30,000 件	0.6055	0.3621	0.1649	0.3536	0.4221	0.2477	0.3593
	社会人	1,000 件	0.3978	0.4887	0.2512	0.2703	0.2146	0.2072	0.3050
		5,000 件	0.5838	0.5486	0.4193	0.3229	0.3649	0.4154	0.4425
		10,000 件	0.6821	0.6600	0.5846	0.4331	0.4713	0.5092	0.5567
		30,000 件	**0.7502**	0.7363	0.6614	0.4777	0.5484	0.5603	0.6224
	主婦	1,000 件	0.3610	0.5982	**0.5790**	0.3082	0.2864	0.2794	0.3508
		5,000 件	0.6365	0.5901	**0.5291**	0.3462	0.3993	0.2493	0.4260
		10,000 件	0.6799	0.7181	**0.5907**	**0.3961**	0.3881	0.3430	0.4922
		30,000 件	0.7545	**0.7750**	0.6360	0.3778	0.4510	**0.3816**	0.5170
	パート・アルバイト	1,000 件	0.1510	0.1382	0.0881	0.1411	0.1184	0.1913	0.1416
		5,000 件	0.3255	0.1879	0.2251	**0.2245**	0.2235	0.2396	0.2376
		10,000 件	0.4294	**0.2204**	0.1759	0.2489	0.2168	0.2389	0.2589
		30,000 件	0.4980	**0.2919**	0.1436	**0.2808**	0.2964	0.2259	0.2887
提案手法	学生	1,000 件	**0.6754**	**0.3609**	**0.3396**	**0.3383**	**0.3915**	**0.2947**	**0.4001**
		5,000 件	**0.6077**	**0.3105**	**0.2815**	**0.3220**	**0.4208**	**0.2843**	**0.3711**
		10,000 件	**0.6418**	**0.3564**	**0.2597**	**0.3368**	**0.4387**	**0.3008**	**0.3890**
		30,000 件	**0.6607**	**0.3764**	**0.2744**	**0.3657**	**0.4420**	**0.2714**	**0.3984**
	社会人	1,000 件	**0.6185**	**0.6245**	**0.6214**	**0.5659**	**0.5862**	**0.3715**	**0.5647**
		5,000 件	**0.7162**	**0.7393**	**0.6542**	**0.5170**	**0.5973**	**0.5291**	**0.6255**
		10,000 件	**0.7834**	**0.7977**	**0.7006**	**0.5580**	**0.6109**	**0.5388**	**0.6649**
		30,000 件	0.7495	**0.7865**	**0.6778**	**0.5504**	**0.6275**	**0.5725**	**0.6607**
	主婦	1,000 件	**0.7281**	**0.6771**	0.5184	**0.3711**	**0.4027**	**0.3507**	**0.4899**
		5,000 件	**0.7093**	**0.6366**	0.5070	**0.3636**	**0.4497**	**0.3181**	**0.4683**
		10,000 件	**0.7507**	**0.7658**	0.5772	0.3926	**0.4176**	**0.3681**	**0.5168**
		30,000 件	**0.7806**	0.7680	0.6360	**0.4249**	**0.4693**	0.3701	**0.5401**
	パート・アルバイト	1,000 件	**0.5950**	**0.2516**	**0.2451**	**0.1937**	**0.2217**	**0.3007**	**0.3091**
		5,000 件	**0.5280**	**0.2124**	**0.2862**	0.2242	**0.2962**	**0.2704**	**0.3085**
		10,000 件	**0.5263**	0.2098	**0.2236**	**0.2754**	**0.2586**	**0.2567**	**0.2987**
		30,000 件	**0.5393**	0.2861	**0.1918**	0.2514	**0.3136**	**0.2353**	**0.3037**

付けする属性は，6.1 節の結果により，性別，年代と職業を用いて作成した属性ベクトルを用いる．

6.2.3 実験手順

実験手順を以下に示す．

STEP 1：ユーザごとに 1,000，5,000，10,000，30,000 件の学習データを無作為に取得する．

STEP 2：ユーザの行動情報に正しいユーザの性別，年代と職業の属性の特性を補完する．

STEP 3：学習データの件数ごとに習慣行動の推定精度を算出し，比較する．

6.2.4 結果と考察

実験結果を**表 7** に示す．表 7 には，既存手法と提案手法を比較して，行動の推定精度が高い部分に太字の下線を記載している．また，職業ごとの投稿件数別推定精度を可視化したものを**図 5** に示す．表 7 と図 5 により，次に示す内容が明らかとなった．

- **投稿数が少ないユーザで推定精度が向上することが分かった**

職業ごとに学習データ 1,000 件の推定精度の平均を確認すると，既存手法と比較して，学生で 0.1743 ポイント，社会人で 0.2597 ポイント，主婦で 0.1391 ポイント，パート・アルバイトで 0.1675 ポイント精度が向上している．既存手法と提案手法のユーザごとの平均の差が統計的に有意であるかを確かめるために，有意水準 1%で両側検定の t 検

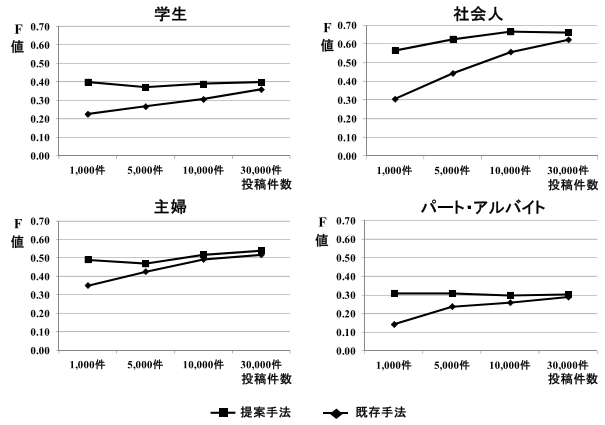

図 5 各属性による投稿数と F 値の関係

Fig. 5 Relation between F-measure and numbers of posts with each attribute.

定を行ったところ，$t(19) = 8.3285$，$p < .01$ となった．このことから，既存手法と提案手法とは有意差があり，提案手法の有効性が明らかとなった．これにより，既存研究の「行動推定の精度が投稿数や投稿記事の量に依存するという問題」に対して，一定の解決策を提示できたといえる．

- **ユーザ属性を考慮することにより行動推定精度が向上することが分かった**

学習データが 30,000 件の推定精度を確認すると，各職業のほぼすべての行動において精度が向上していることが分かる．投稿件数が 10,000 件，および 30,000 件のときにパラメータ e に $e = 0.7$ を採用していることから，ユーザ自身の投稿のみで行動が推定できる場合においても，ユーザ属性を考慮することで精度が向上することが分かった．このことから，ユーザ属性を考慮する提案手法は，行動推定において悪影響を及ぼすものではなく，汎用的に利用できる手法であることが明らかとなった．

- **行動推定の精度が職業ごとに異なることが分かった**

ユーザ属性を考慮して行動を推定する提案手法は，既存手法と比較するとほぼすべての場合においてその精度が向上していることが分かる．しかし，提案手法における職業ごとの平均の推定精度を確認すると，投稿数が 30,000 件の場合でも，学生で 0.3984，社会人で 0.6607，主婦で 0.5401，パート・アルバイトで 0.3037 となっており精度にばらつきが見られた．最も精度が高い社会人の結果では，社会人の多くが同様の行動をとると考えられる出勤中や睡眠中などの行動が最も推定精度が高く，一方で，食事中やその他に分類される旅行などの人により異なる行動では低い傾向にある．また，学生，パート・アルバイトの勤務中の推定精度に着目すると，それぞれ 0.2744，0.1918 となっており，社会人の結果と比較すると大幅に精度が低下している．これらの具体的な行動は授業やアルバイトが主であり，その行動をとるタイミングがユーザごとにまったく異なると考えられる．このことから，社会人や主婦などの一般的に職業ごとに行動が類似すると考えられる範囲に対して提案手法を適用することで，高精度に行動を推定できることが分かった．

6.3 属性推定手法の結果を用いた行動推定手法に関する評価実験

6.3.1 実験概要

属性推定手法の結果を用いた行動推定手法に関する評価実験では，パーソナルデータを高精度に推定できるかを評価するため，提案手法である段階的詳細化の手法を用いた属性推定手法とその属性を考慮した行動推定手法とを組み合わせた場合の行動推定の精度を検証する．本実験により，提案手法の実運用の可能性の検証と実運用へ適用するための制約条件を明らかにする．本実験での行動推定で用いる属性ベクトルは，属性の推定結果を用いる．実験データには，6.2.2 項と同様のデータを用いる．また，パラメータには，6.1.2 項と同様のパラメータを用いる．

6.3.2 実験手順

実験手順を以下に示す．

STEP 1：ユーザごとに 1,000，5,000，10,000，30,000 件の学習データを無作為に取得する．

STEP 2：ユーザの行動情報に属性推定部で出力されたユーザの性別，年代と職業の属性の特性を補完する．

STEP 3：学習データの件数ごとに習慣行動の推定精度を算出する．

6.3.3 結果と考察

実験結果を**表 8** に示す．表 8 は，ユーザごとに提案手法の属性推定を行った結果とその属性を考慮した行動確率モデルを用いて行動推定した結果を示している．表 8 の各ユーザの職業は，A～E が社会人，F～J が学生，K～O が主婦，P～T がパート・アルバイトである．これらのユーザごとに属性推定が正解した場合を「○」とし，投稿件数ごとの F 値をとりまとめた．また，表 8 のすべての属性が正しく推定できなかったユーザにおいて，ユーザ属性がすべて明らかな場合と比較して行動推定の精度が向上した箇所に太字の下線を記載している．表 8 により以下に示す内容が明らかになった．

- **ユーザ属性が明らかな場合と比較して同程度の精度で行動を推定できていることが分かった**

表 7 および表 8 を確認すると，ユーザ属性が明らかである場合（表 7）の全体の推定精度の平均 0.4568 に対し，表 8 の全体の推定精度の平均は 0.4464 であり，0.0104 ポイント差となっている．この 2 手法の差が統計的に有意であるかどうかを評価するため，有意水準 5%で t 検定を行ったところ，$t(19) = 1.2819$，$p > .05$ となり，有意差がないことが分かった．このことから，2 手法の有意差はなく，提案手法は，ユーザ属性が明らかな場合と比較して，同程度の精度で推定できることが分かった．

表 8 属性推定手法の結果を用いた行動推定の提案手法の F 値
Table 8 F-measure of proposed method on behavior estimation using results of attribute estimating method.

ユーザ	推定			投稿数毎の F 値				
	性別	年代	職業	1,000 件	5,000 件	10,000 件	30,000 件	平均
A		○	○	**0.6286**	**0.7344**	0.7645	0.7645	**0.7230**
B	○			**0.5471**	**0.5963**	**0.6231**	**0.6231**	**0.5974**
C	○	○	○	0.4821	0.4240	0.5317	0.5714	0.5023
D	○	○	○	0.5889	0.6888	0.6864	0.6864	0.6626
E	○			0.4624	0.5306	0.5933	0.6948	0.5703
F	○	○	○	0.4837	0.4142	0.3912	0.3779	0.4168
G	○	○	○	0.3773	0.3466	0.3661	0.3548	0.3612
H	○	○		**0.3127**	**0.2645**	0.2858	**0.3371**	**0.3000**
I	○	○	○	0.4628	**0.4690**	**0.4582**	0.4445	**0.4586**
J		○	○	0.3475	0.3364	0.3574	0.4405	0.3705
K	○	○	○	0.5487	0.4943	0.5610	0.6032	0.5518
L	○	○	○	0.5121	0.5031	0.5690	0.5696	0.5385
M	○	○	○	0.4912	0.5374	0.4959	0.5554	0.5200
N	○		○	0.4832	0.3499	**0.5249**	0.2410	0.3998
O	○	○	○	0.4051	0.4330	0.4626	0.4764	0.4443
P	○		○	0.3517	0.3079	0.3744	0.3752	0.3523
Q				0.2425	0.2307	**0.1875**	0.2150	0.2189
R	○		○	0.3486	**0.3046**	0.2906	**0.2901**	**0.3085**
S	○	○		0.3054	0.3947	0.3602	0.3952	0.3639
T				0.2535	**0.2815**	**0.2743**	**0.2593**	0.2672

- **職業が誤判定であったにもかかわらず精度向上する事例があることが分かった**

表 8 を確認すると，ユーザ属性を正しく推定できていないにもかかわらず，推定精度が向上している事例が見られる．この原因を調査するため，投稿件数ごとの推定精度がすべて向上している B のユーザの推定結果を確認すると，年代では 30 代を 40 代，職業では社会人をパート・アルバイトとして誤判定していることが分かった．そこで，このユーザの実際の投稿を確認すると，一般的な社会人の勤務時間とは異なる職業であり，社会人が通常勤務していると考えられる日中の投稿が多く，パート・アルバイトに近い習慣行動であることが分かった．このような，本来の職業とは異なる職業に近い習慣行動をとっているユーザでは，ユーザ属性を正しく推定できていない場合でも，行動推定の精度が向上すると考えられる．このことから，すべての属性が正しく推定できていない状況であったとしても，一部の属性が推定できていることで，推定精度が向上する可能性があることが分かった．

7. おわりに

本研究では，マイクロブログユーザのパーソナルデータを高精度に推定する手法を提案した．属性の推定において，既存研究の課題である「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」に対して，段階的詳細化の考え方を適用したユーザ属性の推定手法を提案した．その手法の有効性を評価する実証実験を行い，社会人など性別ごとに投稿される単語やライフスタイルが異なる特徴が見られる属性に対して，高精度に推定可能であることが分かった．

属性推定の実証実験では，段階的詳細化によるユーザ属性の推定手法により，既存研究の課題である「属性ごとの推定精度の違いを考慮せず一様に処理するという問題」に対応した．

また，行動の推定において，既存研究の課題である「行動推定の精度が投稿数や投稿記事の量に依存するという問題」に対して，ユーザの属性を考慮した推定手法を提案した．その手法の有効性を評価する実証実験において，性別，年代と職業の属性を考慮した場合に特に投稿数が少ないユーザに対して高精度に推定可能であることが分かった．

さらに，属性推定の結果を用いた行動推定を行うことで，提案手法の実運用の可能性の検証と実運用へ適用するための制約条件を明らかにした．

その一方で，以下の 3 つの課題が明らかになった．

課題 1：性別を考慮した職業の推定では，学生やパート・アルバイトといった性別ごとに顕著な特徴が見られない属性の推定が難しい．

課題 2：段階的詳細化の手法により，学習データを属性ごとに分類した場合に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．

課題 3：職業により習慣的な行動を取得しにくいユーザが存在する．

今後は，人間関係や興味・関心の高いトピックなど他のパーソナルデータを考慮した推定手法の検討を行い，データ件数を増やしてモデルを構築することにより課題 1 と課題 2 に対応する予定である．また，課題 3 については，自身の投稿傾向から習慣行動を推定する既存手法 [13] に対して，段階的詳細化の手法を組み合わせることで対応する予定である．

参考文献

[1] 総務省：平成 26 年度版情報通信白書，日経印刷 (2014).
[2] 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，Vol.27, No.1, pp.67–74 (2012).
[3] 池田和史，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，Vol.2, No.1, pp.82–93 (2012).
[4] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying Latent User Attributes in Twitter, *Proc. 2nd International Workshop on Search and Mining User-*

generated Contents, pp.37–44, ACM (2010).
[5] Burger, J., Henderson, J., Kim, G. and Zarrella, G.: Discriminating Gender on Twitter, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.1301–1309, ACL (2011).
[6] Pennacchiotti, M., and Popescu, A.: Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter, *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.430–438, ACM (2011).
[7] Chandra, S., Khan, L. and Muhaya, B.: Estimating Twitter User Location Using Social Interactions – A Content Based Approach, *Proc. 3rd IEEE International Conference on Social Computing*, pp.838–843, IEEE (2011).
[8] 的埜孝宏：Twitter からのユーザ情報抽出およびプロフィール推定，法政大学大学院紀要（情報科学研究科編），No.9, pp.149–154 (2014).
[9] 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，Vol.6, No.5, pp.71–84 (2013).
[10] 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，Vol.2, No.55, pp.773–781 (2014).
[11] 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).
[12] 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 (2011).
[13] 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，Vol.6, No.3, pp.73–89 (2013).
[14] 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).
[15] Facenavi：日本人 Twitter ユーザー調査，入手先 ⟨https://www.facebook.com/facenavi/app_280916948661899⟩（参照 2016-02-16）.
[16] S21G 社：ツイプロ，入手先 ⟨http://twpro.jp/⟩（参照 2016-02-16）.
[17] Twilog, available from ⟨http://twilog.org/⟩ (accessed 2016-02-16).
[18] Chih-Chung, C. and Chih-Jen, L.: LibSVM, available from ⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm/⟩ (accessed 2016-02-16).
[19] 内閣府：平成 26 年度版男女共同参画白書，ウィザップ (2014).
[20] Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
[21] 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).
[22] Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Proc. Information Processing and Management*, Vol.24, No.5, pp.513–523, Pergamon Press (1988).

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．2014 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．現在，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在学中．修士（情報学）．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．

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．2016 年度文部科学大臣表彰 科学技術賞「科学技術振興部門」受賞．

山本 雄平　（正会員）

1986 年生．2009 年関西大学総合情報学部総合情報学科卒業．2011 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．2015 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．同年関西大学先端科学技術推進機構特任助教，現在に至る．博士（情報学）．Web マイニング，自然言語処理，Web ソリューションビジネスに関連する研究に従事．2007 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．

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，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 年から同大学学生センター副所長となり，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．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．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．博士（工学）．2016 年度文部科学大臣表彰 科学技術賞「科学技術振興部門」受賞．

坂本 一磨 （学生会員）

1991 年生．2015 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．システム設計等の研究開発に従事．