情報処理学会論文誌　データベース　Vol.6　No.5　71–84 (Dec. 2013)

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

田中 成典[1,a)]　中村 健二[2]　加藤 諒[3]　寺口 敏生[3]

受付日 2013年6月21日, 採録日 2013年9月29日

概要：マイクロブログから特定の話題に対するユーザの反応を取得する技術が研究されている．マイクロブログをソーシャルセンサとして有効活用するには，ユーザごとの特性を知る必要がある．しかし，マイクロブログでは，ユーザが属性を公開していない場合が多々あるため，ユーザごとの特性を把握できない．このことから，マイクロブログのユーザ属性を推定する研究が注目されている．しかし，既存手法では，主にマイクロブログの投稿内容にのみ着目しており，リアルタイムに発信されるマイクロブログの特性を属性推定に活かせていない．そこで，本研究では，各単位時間の投稿数に基づきユーザをクラスタリングし，投稿内容，生活習慣と投稿時間帯から職業属性を推定する手法を提案する．実証実験では，投稿内容のみを使用して推定する既存手法と，時間的特徴をも考慮する本手法について比較実験を行い，本提案手法の有用性を確認した．

キーワード：マイクロブログ，Web マイニング，ライフスタイル，属性推定

Research for Reasoning Occupation of Users Using Time Posted to Microblogs

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

Received: June 21, 2013, Accepted: September 29, 2013

Abstract: Research is being conducted on technology to get users' reactions to specific topics in microblogs. It is necessary to know the users' characteristics in order to effectively utilize microblogs as social sensors. However, it cannot understand the users' characteristics, because user attributes are not often to the public in microblogs. For this reason, research on estimating user attributes in microblogs has been drawing attention. However, existing methods, which merely focus on the description contents in microblogs, do not take advantage of the characteristics in microblogs that transmit in real time to estimate users' attributes. This research proposes a method for classifying the users according to number of posts per unit time and estimating the occupation attributes by description contents, lifestyle and time zone of posts. Our demonstration experiments verify usability of the proposed method by comparing the existing methods of estimating merely using description contents with the proposed method of estimating using description contents and temporal characteristics.

Keywords: microblog, Web mining, lifestyle, attribute estimation

[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

1. はじめに

インターネット上の情報を分析する研究分野では，解析対象の 1 つとしてマイクロブログが注目されている．代表的なマイクロブログである Twitter には，発言を追跡する follow 関係，投稿を拡散する retweet 機能，投稿内容にタグ付けする hashtag 機能や投稿どうしを関連付ける mention

機能を用いたreply投稿などコミュニケーションを促進する機能が備わっている．このため，マイクロブログ上ではユーザ同士の情報の流通と拡散が促され，リアルタイムな情報が高速に伝搬するという特徴がある．これらの特徴を利用し，マイクロブログを含むCGM上の情報をユーザの意見や評判の傾向を表すソーシャルセンサ [1] として利用する取り組みが行われている．マイクロブログをソーシャルセンサとして活用し，流通する情報を監視・分析することによって，特定の話題，事件やサービスに対する反応をいち早く察知することが可能である．

マイクロブログ上の情報をソーシャルセンサとして活用する際には，年齢や性別，職業といった属性ごとの違いを考慮すると効果的 [2] である．そのため，マイクロブログの分析時には，これらのユーザ属性を抽出することが必要と考えられる．しかし，マイクロブログではプロフィール情報の公開範囲をユーザ自身が自由に決定できるため，プロフィール情報を公開していない多数のユーザの意見を十分に抽出できないという問題がある．そこで，マイクロブログから取得した情報を基に，ユーザの属性を推定する手法が研究されている．ユーザ属性の推定手法に関する既存研究では，主に年齢，性別や居住地などを推定対象の属性として選定している．しかし，ユーザの職業情報はマイクロブログ上にあまり含まれておらず，また推定するための手法も十分には確立されていない．そこで，著者らはマイクロブログユーザの職業を推定する手法の構築に焦点を当てて研究を行っている．本論文では，基礎研究として，ユーザの職業推定時において，マイクロブログへの投稿に暗黙的に含まれるライフスタイル情報を考慮することの有効性を検証する．

2. 既存研究

マイクロブログユーザの属性を推定する既存研究 [3], [4], [5], [6], [7], [8], [9], [10] では，主に過去の投稿内容やプロフィールに記載されている内容に基づき，性別，年齢や居住地域などのユーザ属性を推定する手法が検討されている．また，ブログや掲示板などを対象とした既存研究 [11], [12], [13], [14], [15], [16], [17], [18], [19] でも，投稿内容から年齢や性別などの属性を推定する手法が数多く提案されている．これらの解析手法では，投稿内容から抽出した特徴的な単語を用いてユーザの属性を推定する．しかし，投稿内容に依存する解析手法を適用する場合，マイクロブログの多くが1度に140文字前後のショートテキストしか投稿できない点が問題となる．ブログや掲示板などとは異なり，マイクロブログでは文字数が制限されるため，投稿者の職業属性が異なっていたとしても表現が多様化しにくいと考えられるためである．たとえば，「仕事」に関する投稿でも，学生アルバイタの「仕事おーわり (*^^*)」という投稿と社会人の「さ、今日もお仕事頑張るかー ^_^;」という投稿は，単語的にはほとんど差はない．これに加えて，前後の文章が欠落しているため，投稿内容のみからユーザ属性の差別化に活用可能な特徴を抽出することは難しいと考えられる．また，マイクロブログのプロフィール欄から職業属性を取得する方策も考えられるが，職業を明記しているユーザ数は少なく，全体の13.62% [20] とわずかである．このことから，マイクロブログの投稿内容にのみ依存する手法では，ユーザの属性の推定は難しく，情報の補完手法を考案する必要があることが分かる．

投稿内容に頼らない属性推定手法としては，ソーシャルグラフを用いた解析手法 [21], [22], [23] が研究されている．リンク関係にある近隣ユーザは互いに似た属性を持つと仮定した解析手法では，ユーザの興味関心が高い分野についての情報を取得できる．しかし，インターネット上の人間関係は職業のみに依存せず，趣味や興味などにも影響されるため，多様な人間関係が含まれる．このため，リンク関係やコミュニケーションに基づく人間関係の解析のみでは，ユーザの職業を推定することは難しいと考えられる．

以上の既存手法の問題である投稿内容，人間関係，プロフィールのような「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」に対応するため，著者らは明示的な情報に加えて，マイクロブログへの投稿に暗黙的に含まれるライフスタイルを分析し，その結果に基づきユーザの職業を推定する手法について研究する．職業は，ユーザのライフスタイルに強い影響を持つ要素の1つであると考えられる．そこで，本研究では，マイクロブログへの投稿内容と投稿時間との関係を分析して抽出したライフスタイルに基づきユーザ群を類型化する．そして，類型化結果に基づきユーザの職業を推定する手法を提案する．このとき，同じ職業でも，ライフスタイルによっていくつかのグループに分類されることが想定される．そのため，「同じ職業でも多様なライフスタイルが存在する問題」への対応についてもあわせて検討する．

3. 研究の概要

3.1 研究の目的

本研究では，投稿内容に加え，ライフスタイルを考慮した，マイクロブログユーザの職業属性の推定手法を提案する．そのため，既存手法の課題である投稿内容，人間関係，プロフィールのような「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」と提案手法の検討にあたり課題となる「同じ職業でも多様なライフスタイルが存在する問題」に対応する．

- **「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」への対応方法**

  本課題に対しては，「ライフスタイルに密着した単語が出現する時間帯・曜日ごとの投稿数」を考慮することで対処する．「おはよう」や「おやすみ」などの生

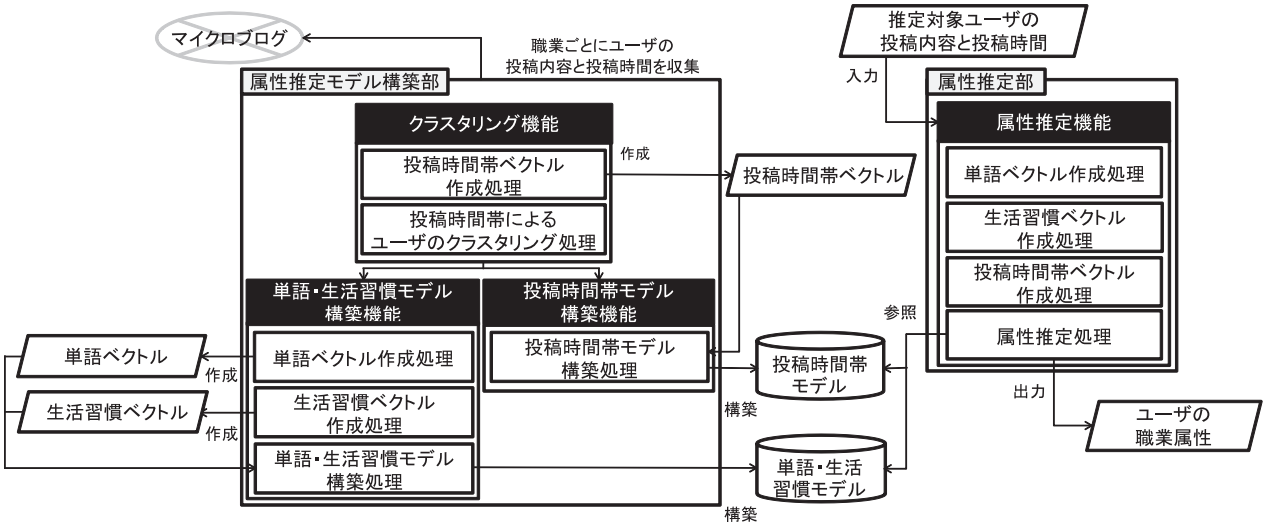

図 1　本提案手法の処理フロー
Fig. 1　Flow of process.

活時間に密着した単語が出現する時間帯を考慮することで，ライフスタイルの特徴を抽出できると考えられる．また，時間帯・曜日ごとの投稿数を考慮することで，週単位，曜日単位のライフスタイルの特徴を抽出でき，職業ごとのライフスタイルの差異を強調することができると考えられる．

- **「同じ職業でも多様なライフスタイルが存在する問題」への対応方法**

  本課題に対しては，同様の職業でもライフスタイルの異なるユーザを整理・分類して判定することで対処する．職業はユーザのライフスタイルを規定する主な要因の 1 つであるが，業務内容や生活態度によって，大小の違いが生じることが想定される．たとえば，同じ大学生でも学業に熱心な学生とアルバイトに熱心な学生では，ライフスタイルには違いがある．このように，分類としては同じ職業であってもライフスタイルが異なる多種多様なユーザが存在し，これらの違いを考慮しなければ，正しくユーザの職業を推定することは難しいと考えられる．そこで，「同じ職業のユーザをマイクロブログへの投稿時間によりクラスタリング」することで，この課題に対応する．

以上のように，マイクロブログに暗黙的に含まれるライフスタイルを考慮することで既存研究の問題点に対応し，マイクロブログユーザの職業属性の推定精度を向上させることを本研究の目的とする．

3.2　提案手法の概要

本研究では，「生活習慣に関わる語句の時間帯ごとの投稿数」をまとめた生活習慣ベクトルと「時間帯・曜日ごとの投稿数」をまとめた投稿時間帯ベクトルとを用いて，ユーザのライフスタイルを表現する．本提案手法の処理フローを図 1 に示す．本提案手法は，属性推定モデル構築部と属性推定部の 2 つの処理部により構成される．

属性推定モデル構築部は，クラスタリング機能，単語・生活習慣モデル構築機能と投稿時間帯モデル構築機能で構成される．クラスタリング機能では，収集した各ユーザの投稿履歴から，曜日ごと（7 曜日）・時間帯（24 時間）ごとの投稿数を抽出し，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．単語・生活習慣モデル構築機能では，各職業に特徴的な単語の出現数を示す単語ベクトルと生活習慣ベクトルを各クラスタ単位に作成する．そして，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．投稿時間帯モデル構築機能では，クラスタリング機能により分類された各クラスタの中心ベクトルを取得して関連付けることで，投稿時間帯モデルを構築する．

属性推定部では，属性推定機能を用いて推定対象ユーザの投稿内容と投稿時間を分析し，それらと単語・生活習慣モデルや投稿時間帯モデルに適用することでユーザの属性を推定する．

4.　属性推定モデル構築アルゴリズム

属性推定モデル構築部では，単語・生活習慣モデルと投稿時間帯モデルを構築する．本処理部では，職業 *job* のユーザ群の投稿内容と投稿時間を入力し，投稿時間帯ベクトル，単語ベクトルと生活習慣ベクトルを作成する．これを推定対象の職業数である m 回分繰り返すことで各職業の特徴を学習したモデルを構築する．本章では，モデル構築の前処理であるクラスタリング機能について述べ，その後，単語・生活習慣モデル構築機能と投稿時間帯モデル構築機能について述べる．

4.1　クラスタリング機能

4.1.1　投稿時間帯ベクトル作成処理

本処理では，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職業 *job* のユーザごとに作成する．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．職業 *job* のユーザ *user* における投稿時間帯ベクトル $V_{posttime}(job, user)$ を式 (1) に示す．

$$\begin{aligned} &V_{posttime}(job, user) \\ &= \{Post_{Sunday0}(job, user), \cdots, Post_{Sunday23}(job, user), \\ &\cdots, Post_{Saturday23}(job, user)\} \end{aligned} \tag{1}$$

式 (1) において，$Post_{Sunday0}(job, user)$ は日曜日の 0 時に職業 *job* のユーザ *user* により投稿された件数を表す．

4.1.2　投稿時間帯によるユーザのクラスタリング処理

本処理では，投稿時間帯ベクトルを使用して，職業 *job* のユーザ群を n 個のクラスタに分類する．分類したクラスタをそれぞれ $\{C_{job_1}, C_{job_2}, \cdots, C_{job_k}, \cdots, C_{job_n}\}$ と表す．クラスタリングには，クラスタ分類で一般的に使用される k-means 法 [24] を採用する．本処理により，同一職業中におけるユーザを投稿時間に基づいて類型化し，同じ職業でも多様なライフスタイルが存在するという問題に対応する．

4.2　単語・生活習慣モデル構築機能

4.2.1　単語ベクトル作成処理

本処理では，投稿内容に出現する特徴的な単語を素性とした単語ベクトルをクラスタに属するユーザ群ごとに作成する．クラスタ C_{job_k} に属するユーザ *user* における単語ベクトル $V_{word}(C_{job_k}, user)$ を式 (2) に示す．

$$\begin{aligned} &V_{word}(C_{job_k}, user) \\ &= \{Post_{word_1}(C_{job_k}, user), Post_{word_2}(C_{job_k}, user), \cdots\} \end{aligned} \tag{2}$$

式 (2) において，$Post_{word_1}(C_{job_k}, user)$ はクラスタ C_{job_k} に属するユーザ *user* による投稿における単語 $word_1$ の出現数を表す．なお，単語ベクトルを構成する素性は次に示す手順により選定する．

STEP 1：各職業において特徴的な単語を抽出する．推定する職業の種類だけ STEP 1.1 から STEP 1.2 の処理を繰り返す．

STEP 1.1：Twitter から収集した職業 *job* のユーザ群の投稿内容に対して形態素解析を行い，単語を取得する．なお，取得する単語は，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：職業 *job* において特徴的な単語を抽出する．単語の評価には，既存研究 [11] にならい，χ^2 値を採用する．χ^2 値を使用することで，職業 *job* に属するユーザ群のクラスタと，ある単語が投稿内容に出現したユーザ群のクラスタがどの程度一致しているかを評価できる．職業 *job* のユーザ群の投稿内容に出現する単語 *word* の χ^2 値の算出方法を式 (3) に示す．

$$\begin{aligned} &\chi^2(job, word) \\ &= \frac{n(U) \times (n(J \cap W) \times n(\overline{J \cup W}) - n(\overline{J} \cap W) \times n(J \cap \overline{W}))}{n(J) \times n(\overline{J}) \times n(W) \times n(\overline{W})} \end{aligned} \tag{3}$$

式 (3) において，U は入力したユーザの集合，J は職業 *job* のユーザの集合，W は投稿内容に単語 *word* が出現するユーザの集合を表す．式 (3) では，職業 *job* において単語 *word* が特徴的な単語であるほど χ^2 値が大きくなる．

STEP 2：STEP 1 で取得した単語群を統合する．複数の職業で同じ単語が出現する場合，χ^2 値の大きな職業で特徴的な単語として使用する．

STEP 3：χ^2 値に基づき，STEP 2 で統合した単語群のランキングを作成する．

STEP 4：ランキング上位の単語を素性として抽出する．

4.2.2　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生活習慣ベクトルをクラスタに属するユーザ群ごとに作成する．本研究では，既存研究 [25] で提案されている分類から「出勤」，「勤務」，「帰宅」，「睡眠」，「食事」と「その他」の 6 種類を習慣行動として採用する．生活習慣ベクトルは，6 次元（習慣行動）× 24 次元（時間帯）の 144 次元で構成する．クラスタ C_{job_k} に属するユーザ *user* における生活習慣ベクトル $V_{lifecycle}(C_{job_k}, user)$ を式 (4) に示す．

$$\begin{aligned} &V_{lifecycle}(C_{job_k}, user) \\ &= \{Post_{behavior_1 0}(C_{job_k}, user), Post_{behavior_1 1}(C_{job_k}, user), \\ &\cdots, Post_{behavior_6 23}(C_{job_k}, user)\} \end{aligned} \tag{4}$$

式 (4) において，$Post_{behavior_1 0}$ は 0 時 00 分 00 秒から 0 時 59 分 59 秒までの間にクラスタ C_{job_k} に属するユーザ *user* 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．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活動辞書に登録されている用語を使用する．活動辞書には，既存研究 [26] にならい，日本語語彙大系 [27] 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．活動辞書に登録した用語の例を**表 1** に示す．

4.2.3　単語・生活習慣モデル構築処理

本処理では，単語ベクトルと生活習慣ベクトルを使用して各職業のクラスタごとに学習することで，単語・生活習慣モデルを構築する．モデルの構築には，人工知能分野で使用される分類手法である SVM（Support Vector

表 1　活動辞書に登録した用語の例
Table 1　Example of terms on dictionary concerning activities.

活動	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

Machine）[28] を採用する．なお，学習には，単語ベクトルと生活習慣ベクトルを結合した単語・生活習慣ベクトルを使用する．クラスタ C_{job_k} に属するユーザ *user* における単語・生活習慣ベクトル $V_{lifestyle}(C_{job_k}, user)$ を式 (5) に示す．

$$\begin{aligned} &V_{lifestyle}(C_{job_k}, user) \\ &= \{Post_{word_1}(C_{job_k}, user), \cdots, Post_{behavior_1 0}(C_{job_k}, user), \\ &\quad \cdots, Post_{behavior_6 23}(C_{job_k}, user)\} \end{aligned} \tag{5}$$

4.3　投稿時間帯モデル構築機能

本処理では，投稿時間帯ベクトルを使用して各職業のクラスタごとに学習することで投稿時間帯モデルを構築する．モデルの構築には，VSM（Vector Space Model）[29] を採用する．なお，学習には，m（推定対象の職業数）× n（各職業のクラスタ数）個のクラスタの中心ベクトルを使用する．

5.　属性推定アルゴリズム

属性推定部では，単語・生活習慣モデルと投稿時間帯モデルに基づき，ユーザの職業を推定する．本処理部では，推定対象のユーザ *target* の投稿内容と投稿時間を入力し，投稿時間帯ベクトル $V_{posttime}(target)$，単語ベクトル $V_{word}(target)$ と生活習慣ベクトル $V_{lifecycle}(target)$ を作成する．そして，作成したベクトルを使用してユーザの職業属性を推定する．属性推定機能のうち，単語ベクトル，生活習慣ベクトルと投稿時間帯ベクトルの作成処理は，属性推定モデル構築部と同様の手順で構築するため，本章では属性推定処理について述べる．

5.1　属性推定機能

本処理では，単語・生活習慣モデルと投稿時間帯モデルによる類似度の算出をそれぞれ行い，これらを組み合わせた統合類似度を用いてユーザ *target* の職業属性を推定する．ユーザの職業属性を推定する手順を次に示す．

STEP 1：単語・生活習慣モデルを用いた各クラスタとの類似度を算出する．算出には STEP 1.1 から STEP 1.3 の処理を行う．

STEP 1.1：単語ベクトル $V_{word}(target)$ と生活習慣ベクトル $V_{lifecycle}(target)$ を結合した単語・生活習慣ベクトル $V_{lifestyle}(target)$ を作成する．

STEP 1.2：単語・生活習慣ベクトル $V_{lifestyle}(target)$ を入力し，各クラスタに分類される確率を算出する．本論文では，SVM の実装である LibSVM [30] の Predict Probability 機能を用いた．

STEP 1.3：推定対象のユーザ *target* とクラスタ C_{job_k} との単語・生活習慣モデル類似度を $P_{lifestyle}(target, C_{job_k})$ と表す．

STEP 2：ユーザと投稿時間帯モデルを用いた各クラスタとの類似度を算出する．算出にはクラスタ数だけ STEP 2.1 から STEP 2.2 の処理を繰り返す．

STEP 3：推定対象のユーザ *target* とクラスタ C_{job_k} との投稿時間帯モデル類似度 $P_{posttime}(target, C_{job_k})$ を算出する．$P_{posttime}(target, C_{job_k})$ の算出方法を式 (6) に示す．

$$\begin{aligned} &P_{posttime}(target, C_{job_k}) \\ &= \frac{Similarity(target, C_{job_k})}{\sum_{l=\mathbf{1}}^{o} \sum_{m=\mathbf{1}}^{n} Similarity(target, C_{lm})} \end{aligned} \tag{6}$$

式 (6) において，o は，推定する職業の総数，$Similarity(target, C_{job_k})$ は，ユーザ *target* の投稿時間帯ベクトル $V_{posttime}(target)$ と投稿時間帯モデルのクラスタ C_{job_k} の中心ベクトル $V_{C_{job_k}}$ との類似度である．類似度 $Similarity(target, C_{job_k})$ は，コサイン類似度 $\mathrm{Cos}(V_{posttime}(target), V_{C_{job_k}})$ とユークリッド距離 $\mathrm{Euclid}(V_{posttime}(target), V_{C_{job_k}})$ とを用いて算出する．類似度の算出方法を式 (7) から式 (9) に示す．

$$\begin{aligned} &Similarity(target, C_{job_k}) \\ &= \frac{1}{Cos(V_{posttime}(target), V_{C_{job_k}}) \times Euclid(V_{posttime}(target), V_{C_{job_k}})} \end{aligned} \tag{7}$$

$$\begin{aligned} &Cos(V_{posttime}(target), V_{C_{job_k}}) \\ &= 1 - \frac{\sum_{l=1}^{168} W(Nt_l) \times W(NC_{job_{kl}})}{\sqrt{\sum_{l=1}^{168} W(Ntarget_l)^2} \times \sqrt{\sum_{l=1}^{168} W(NC_{job_{kl}})^2}} \end{aligned} \tag{8}$$

$$\begin{aligned} &Euclid(V_{posttime}(target), V_{C_{job_k}}) \\ &= \sqrt{\sum_{l=1}^{168} \left(W(Ntarget_l) - W(NC_{job_{kl}})\right)^2} \end{aligned} \tag{9}$$

式 (8) や式 (9) において，$\boldsymbol{W(Ntarget_l)}$ は投稿時間帯ベクトルにおける l 番目の素性の値，$\boldsymbol{W(NC_{job_{kl}})}$ はクラスタ $\boldsymbol{C_{job_k}}$ の中心ベクトルにおける l 番目の素性の値を表す．

式 (8) のコサイン類似度は，ベクトルの方向のみをとらえた距離を示す．式 (9) のユークリッド距離は，ベク

トルの長さも考慮した 2 点間の距離を示す．式 (7) では，コサイン類似度とユークリッド距離は相補的な情報を提供できているため，これら両方を用いて最終的な距離を算出する．ここで，$Similarity(target, C_{job_k})$ は，投稿時間帯ベクトル $V_{posttime}(target)$ とクラスタ C_{job_k} の中心ベクトル $\mathrm{V}_{C_{job_k}}$ が類似するほど大きな値となる．

STEP 2.1：単語・生活習慣モデル類似度 $P_{lifestyle}(target, C_{job_k})$ と投稿時間帯モデル類似度 $P_{posttime}(target, C_{job_k})$ とを統合した統合類似度 $P_{integration}(target, C_{job_k})$ を算出し，それに基づき対象ユーザの属性を推定する．ユーザ $target$ とクラスタ C_{job_k} との統合類似度 $P_{integration}(target, C_{job_k})$ の算出方法を式 (10) に示す．

$$P_{integration}(target, C_{job_k}) = P_{lifestyle}(target, C_{job_k}) \times P_{posttime}(target, C_{job_k}) \quad (10)$$

式 (10) において算出した各クラスタとの統合類似度のうち最も高い類似度のクラスタの職業を推定対象のユーザの職業属性として出力する．

6. 実証実験

6.1 実験の概要

実証実験では，本提案手法であるユーザのライフスタイルを考慮した職業推定手法の有効性を評価する．実証実験では，予備実験と比較実験の 2 つを行う．予備実験では，単語・生活習慣モデル構築機能で使用する SVM の素性数，およびクラスタリング機能で使用する k-means 法のクラスタ数の最適値を決定する．比較実験では，投稿内容に含まれる特徴的な単語に基づきユーザ属性を推定する手法 [11] の推定精度と本提案手法の推定精度とを比較することで，本提案手法の有効性を評価する．

6.2 推定対象の職業属性の決定

本研究の推定対象の職業属性は，既存のアンケートで利用されている職業分類の分析結果に基づき決定する．手順としては，まず，Google 画像検索を用いて Web 上に公開されている様々なアンケートのグラフを上位 50 件収集する．画像検索時には，「職業アンケート　グラフ」の検索クエリを用いる．次に，収集したアンケート中における職業の出現数を集計する．最後に，出現数が多い順に集計結果をソートする．以上の手順で収集したアンケートの内容を確認したところ，スーツの販売や商店街活性化などのマーケティングに関するもの，政治や公共サービスに関するものや平均睡眠時間などの個人の生活に関するものなど，様々な分野のアンケートを収集できたことが分かった．このことから，本調査結果で得られた職業分類はソーシャルセンサとしての役割を果たすことができると考えられる．

表 2　アンケートに用いられる職業属性の集計結果
Table 2　Summary results of occupation attributes using questionnaires.

職業属性	集計結果	職業属性	集計結果
その他	40	会社員	25
主婦	33	公務員	20
学生	31	パート・アルバイト	15
無職	30	会社役員	6
自営業	29	高校生	6

アンケートに用いられる職業属性の集約結果を**表 2** に示す．表 2 より，「その他」以外で出現数が 2 桁以上の職業は，「主婦」「学生」「無職」「自営業」「会社員」「公務員」「パート・アルバイト」の 7 種類であった．そこで，実験においては，マイクロブログの情報から職業推定を始めるにあたり，「主婦」「学生」「無職」「自営業」「会社員」「公務員」「パート・アルバイト」の 7 種類を推定対象の職業として採用した．

6.3 実験データ

実験データは，職業ごとに 330 ユーザ（合計 2,310 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 コミュニティサイト「ツイナビ [31]」から職業を公開しているユーザを無作為に収集する．「ツイナビ」は Twitter ユーザが交流するためのポータルサイトであり，ユーザ自身が地域，性別，年代，血液型や職業などの属性情報を登録するため，信頼度が高いユーザ情報を取得できると考えられる．また，ツイナビに登録されている属性情報は，ユーザが指定した属性情報でありプロフィール欄や投稿内容に依存せずに正解データを抽出可能であると考えられる．

STEP 2：STEP 1 で収集したユーザの投稿内容を TwitterAPI [32] で収集する．投稿内容の収集件数は，TwitterAPI で収集可能な最大件数である 3,200 件までとする．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．retweet 機能は他のユーザの投稿を引用する機能であり，ユーザ自身の発言ではないことから，既存の特徴的な単語に基づく手法において精度が低下すると考えたためである．

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 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：各職業のユーザ数が 330 件になるまで，STEP 1 から STEP 3 を繰り返し実施する．

STEP 5：STEP 4 で収集した実験データのうち，各職業の 250 ユーザを教師データ，80 ユーザを判定データとして使用する．

6.4　予備実験

6.4.1　SVM の素性数の最適値評価実験

(1) 実験内容

SVM の素性数の最適値評価実験では，単語・生活習慣モデル構築機能で使用する SVM の素性数の最適値を決定する．実験対象の素性数は既存研究 [11] を参考し，128 件，256 件，512 件，1,024 件と 2,048 件の 5 種類とする．本実験の手順を次に示す．

STEP 1：単語ベクトル作成処理の STEP 5 で算出した χ^2 値に基づき，ランキングの上位 128 件，256 件，512 件，1,024 件と 2,048 件の単語を素性とした単語ベクトルをそれぞれ作成する．

STEP 2：STEP 1 で作成した単語ベクトルを入力として単語モデルを構築する．学習には教師データ 1,750 ユーザを使用する．また, SVM のライブラリには, LibSVM を利用する．SVM のカーネル関数には，テキストマイニング手法を用いる既存研究 [33], [34], [35], [36] で利用される RBF カーネルを用いる．本実験では，SVM の学習に必要な cost パラメータと，RBF カーネルの使用時に必要な gamma パラメータについて，最適な値の組合せを網羅的に探索する手法であるグリッドサーチを実施する．これにより，素性数 128 件，256 件，512 件，1,024 件と 2,048 件で学習した 5 種類の単語モデルを構築する．

STEP 3：STEP 2 で構築した 5 種類の単語モデルを使用して実験データの推定精度を評価する．推定精度の評価には，職業ごとに算出した F 値の平均を使用する．ここで，推定精度の最も良い SVM の素性数とパラメータを最適値として採用する．

(2) 実験結果

素性数ごとの推定精度を**図 2** に示す．実験結果より，256 件の素性数で学習した場合の推定精度が最も良い値を示していることが分かる．このことから，推定対象の職業数 $m = 7$ の場合は，素性数 256 件を SVM の最適値として採用する．

6.4.2　k-means 法のクラスタ数の決定実験

(1) 実験内容

k-means 法のクラスタ数の決定実験では，クラスタリング機能で使用する k-means 法のクラスタ数の最適値を決定する．実験対象のクラスタ数は，教師データのユーザ数を考慮し，1〜5 クラスタとした．k-means 法の素性数の最適値の決定方法を次に示す．

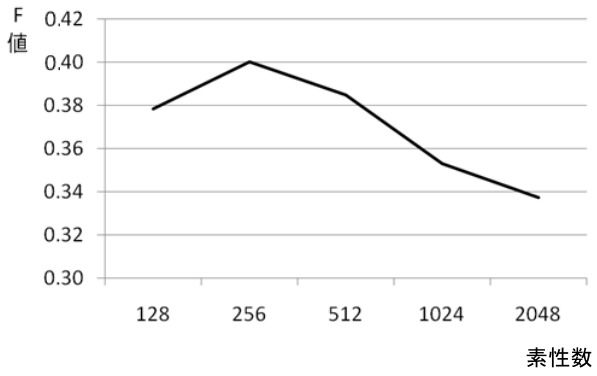

図 2　素性数ごとの推定精度
Fig. 2　Estimation accuracy by each feature.

STEP 1：投稿時間帯ベクトル作成処理で作成した職業ごとの投稿時間帯ベクトルに対して，クラスタ数ごとにクラスタリングを行う．なお，k-means 法は，そのアルゴリズムの特性として，与える中心値の初期値によってクラスタリング結果が変化する．そこで，実験データとは異なるデータセットを対象に k-means 法によるクラスタリングを 10 回試行し，最も高精度にデータセットを分類できた場合の初期値を採用する．ただし，データセットを対象にクラスタリングを行う場合の初期値は，データセットから無作為に選択したユーザを使用する．

STEP 2：STEP1 でクラスタリングした結果を使用して実験データの推定精度を評価する．推定精度の比較には，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考慮した手法で職業ごとに算出した F 値の平均を使用する．

(2) 実験結果

クラスタ数ごとの推定精度を**図 3** に示す．実験結果より，職業ごとの分類として 3 クラスタで学習した場合の推定精度が最も良い値を示していることが分かる．このことから，推定対象の職業数 $m = 7$ の場合は，クラスタ数 3 を最適値として採用する．

6.5　比較実験

6.5.1　実験内容

比較実験では，文献 [11] にならい χ^2 値を用いて特徴的な単語を選定し，SVM でユーザの属性を推定する手法（以下，「既存手法」）の推定精度と，投稿内容に加えてユーザのライフスタイルを考慮する提案手法の推定精度とを比較する．

なお，本比較実験で使用する SVM の素性数は 6.4.1 項

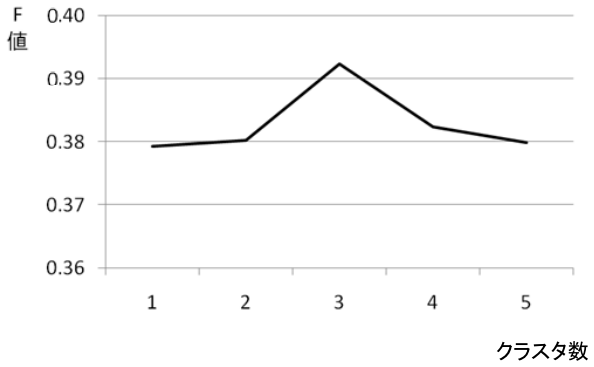

図 3　クラスタ数ごとの推定精度
Fig. 3　Estimation accuracy by each cluster.

表 3　既存手法と提案手法の推定精度
Table 3　Estimation accuracy of existing method and proposed method.

	職業	適合率	再現率	F 値
既存手法	学生	0.619	0.650	0.634
	会社員	0.426	0.288	0.343
	主婦	0.703	0.325	0.444
	パート・アルバイト	0.440	0.413	0.426
	公務員	0.302	0.713	0.424
	自営業	0.360	0.225	0.276
	無職	0.268	0.238	0.252
	平均	0.445	0.407	0.400
提案手法	学生	0.514	0.700	0.593
	会社員	0.358	0.300	0.327
	主婦	0.718	0.350	0.471
	パート・アルバイト	0.509	0.363	0.423
	公務員	0.290	0.500	0.367
	自営業	0.293	0.300	0.296
	無職	0.294	0.250	0.270
	平均	0.425	0.394	0.392

の予備実験の結果に基づき 256 件とする．また，クラスタリング機能で使用するクラスタ数は 6.4.2 項の予備実験の結果に基づき 3 クラスタとする．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．

6.5.2　結果と考察

既存手法と提案手法の推定精度を**表 3** に示す．表 3 より，既存手法と提案手法の両方で精度が低い結果となった．既存手法の推定精度が低いのは，マイクロブログ上には職業を表す特徴的な単語が出てきにくいという予想を裏付ける結果であると考えられる．一方，提案手法の推定精度が低い理由を検証するため，実験結果の誤判定の内訳を確認したところ，特に「会社員」,「パート・アルバイト」,「公務員」,「自営業」や「無職」において，数多くの誤判定が発生したことが分かった．また,「主婦」や「無職」以外の職業では，提案手法の方が推定精度は低い結果となった．これは，ライフスタイルに顕著な差がない職業を判定対象と設定していたため，それがノイズとなって推定精度の低下を引き起こしたと考えられる．このため，以上の実験結果では，属性推定時にライフスタイルを考慮することの有効性の検証ができたとはいい難い．そこで，ライフスタイルが類似し，応用上必ずしも区別する必要がない職業を 1 つにまとめて追試を行う．

6.6　推定対象職業数 4 種類の場合の実証実験

6.6.1　実験の概要

本実験では，6.5 節の実験結果で，提案手法の推定精度が低かった原因は，ライフスタイルに顕著な特徴がみられない職業がノイズになったためであると考え，推定対象の職業を再検討したうえで，改めてライフスタイルの有用性に関する追試を行う．

表 4 の誤判定結果より,「会社員」,「公務員」と「自営業」のライフスタイルは相互に類似しており,「学生」,「主婦」や「パート・アルバイト」とは異なる特徴があると考えられる．このことから,「会社員」,「公務員」や「自営業」のライフスタイルを包括的に内包する職業として,「社会人」を新たな推定対象として選定する．これらの職業をまとめた包括的な「社会人」という考え方は，様々な分野のアンケート調査でも利用されている．具体的なアンケート結果の事例を調査したところ,「健康や体調に関するアンケート [37]」や「映画や娯楽などに関するアンケート [38]」,「大学院や専門学校への進学意思のアンケート [39], [40]」などにおいて，社会人の括りが用いられていた．これらのアンケートでは，社会人の中での職種や業種などの詳細な分類の必要がなく，大まかな傾向を調査する目的で収集されているものが多い．こういった概観を把握するための調査を実施する場合には，社会人として一括りにした職業を用いても問題ないと考えられる．

一方,「無職」のユーザは,「公務員」が最も誤判定として多くみられるが,「会社員」,「公務員」と「自営業」のユーザは,「無職」として誤判定されるケースが少ないことが分かる．また,「無職」のユーザは，その他の職業分類とも強い類似性がみられないことから，本研究では,「無職」として収集したユーザの一部を就労が不安定なユーザであると考え，職業的には同じく短期的な労働形態である「パート・アルバイト」に属するものと仮定して分析する．実際に，内閣府の青少年の社会的自立に関する意識調査 [41] より，無職のユーザのうち 51.6%が以前は非正規雇用者であることからも，上述の仮定には一定の有意性があると判断した．なお，6.5 節に比べ職業数が変化するため，6.4 節の予備実験を含むすべての実験を改めて実施し，提案手法を検討する．

表 4　誤判定の分析結果
Table 4　Analysis results failed in judgement.

		正解の職業						
		学生	会社員	主婦	パート・アルバイト	公務員	自営業	無職
判定結果の職業	学生	-	0.375	0.077	0.353	0.025	0.000	0.150
	会社員	0.125	-	0.192	0.137	0.250	0.125	0.100
	主婦	0.042	0.000	-	0.059	0.100	0.054	0.000
	パート・アルバイト	0.292	0.054	0.154	-	0.000	0.054	0.117
	公務員	0.083	0.304	0.038	0.059	-	0.696	0.583
	自営業	0.167	0.161	0.250	0.118	0.575	-	0.050
	無職	0.292	0.107	0.288	0.275	0.050	0.071	-

6.6.2　実験データ

本実験では，職業が「学生」「社会人」「主婦」「パート・アルバイト」のユーザを実験対象とする．実験データは，6.3 節で収集した職業ごとのデータのうち，「学生」と「主婦」についてはそのままのデータ（各 330 ユーザ）を用いた．一方，「社会人」は「自営業」「公務員」「会社員」から抽出した 330 ユーザを用いた．また，「パート・アルバイト」には，6.5 節の実験で用いた「無職」のユーザを一部含めた 330 ユーザを用いた．

これらの実験データのうち，各職業の 250 ユーザを教師データ，80 ユーザを判定データとして使用することとした．

6.6.3　予備実験

6.4 節と同様の手順で，SVM の素性数と k-means のクラスタ数の最適値を求める予備実験を行った．結果として，素性数 256 件，クラスタ数 2 を最適値として採用する．

6.6.4　比較実験

6.6.3 項で求めたパラメータを用い，職業推定を実施する．比較実験の手順は 6.5 節と同様である．また，本実験においては，既存手法と提案手法の実験に加えて，提案手法のどの情報が職業推定に効果的であるかを分析するため，本提案手法の各項目についての分析を行う．具体的には，「単語の χ^2 値を考慮する既存手法（単語）」，「日常的な語句から取得した生活習慣を考慮する手法（生活習慣）」，「職業に特徴的な語句の時間的特徴を考慮する手法（投稿時間帯）」と「クラスタリング手法（クラスタリング）」という 4 手法のすべての組合せ（14 種類）に対して実験を行い，推定精度を算出する．なお，本実験で使用するパラメータは，6.6.3 項の予備実験の結果に基づき，SVM の素性数は 256 件，クラスタリング機能で使用するクラスタ数は 2 クラスタとする．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．

6.6.5　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．また，4 手法をそれぞれ組み合わせて実験した場合の推定精度を**表 6** に示す．表 6 において，単語列は「単語の χ^2 値を考慮する手法」の適用の有無，生活習慣列は「日常的な語句から取得した生活習慣を考慮する手法」の適用の有無，投稿時間帯列は「職業に特徴的な語句の時間的特徴を考慮する手法」の適用の有無，クラスタリング列は「クラスタリング手法」の適用の有無を表す．表 5 および表 6 より，次に示す 3 つの考察を行った．

表 5　4 属性における既存手法と提案手法の推定精度
Table 5　Estimation accuracy of existing method and proposed method using four attributes.

	職業	適合率	再現率	F 値
既存手法	学生	0.744	0.763	0.753
	社会人	0.681	0.613	0.645
	主婦	0.838	0.713	0.770
	パート・アルバイト	0.673	0.825	0.742
	平均	0.734	0.728	0.727
提案手法	学生	0.813	0.763	0.787
	社会人	0.791	0.663	0.721
	主婦	0.853	0.800	0.826
	パート・アルバイト	0.670	0.863	0.754
	平均	0.782	0.772	0.772

(1)　**職業属性の推定精度に関する考察**

表 5 より，提案手法は，既存手法と比較してすべての職業において F 値の精度が向上していることが確認できる．全職業の推定精度を比較すると，平均して 0.045 ポイント F 値が向上していることが分かる．この差が統計的に有意差であるかどうかを確認するため，t 検定を実施した結果，$t(3) = 2.311$，$p < 0.05$ 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．また，職業属性の差を個別に確認すると，特に社会人においては，0.076 ポイントの違いがみられた．これは，社会人のライフスタイルが規則的であるため，本提案手法の推定精度が高まったと考えられる．

表 6　手法の組合せ別の推定精度
Table 6　Estimation accuracy by combination of methods.

実験	単語	生活習慣	投稿時間帯	クラスタリング	F 値
A	○				0.727
B		○			0.596
C			○		0.355
D	○	○			0.763
E	○		○		0.739
F		○	○		0.593
G	○	○	○		0.769
H	○			○	0.744
I		○		○	0.614
J			○	○	0.340
K	○	○		○	0.768
L	○		○	○	0.770
M		○	○	○	0.550
N	○	○	○	○	0.772

表 6 より，単語の χ^2 値に基づきユーザ属性を推定する既存手法（実験 A）と比較して，既存手法に提案手法を組み合わせた手法（実験 D，E，G，H，K，L，N）の方が，精度が向上していることが分かる．この差が統計的に有意な差であるかを検証するため t 検定を実施したところ，実験 K，L，N については，有意差が認められた．このことから，既存手法（実験 A）に対し，「クラスタリングと生活習慣（実験 K)」，「クラスタリングと投稿時間帯（実験 L)」，「クラスタリング，生活習慣，投稿時間帯（実験 N)」を考慮することで推定精度が向上し，ライフスタイルを反映するための各操作がユーザの職業推定に有効であることが明らかとなった．

これらの結果から，マイクロブログに対して，ユーザのライフスタイルを考慮した職業推定手法を適用することは有効であることが明らかとなった．

(2) **「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」に関する考察**

本提案手法は，3.1 節であげた「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」に対して，「ライフスタイルに密着した単語が出現する時間帯・曜日ごとの投稿数」を考慮することで対応を試みている．提案手法の有用性を検証することを目的として，各職業の生活習慣ベクトルと投稿時間帯の関係を目視で確認することで，各職業の推定に有用な情報を含まれるかどうかを分析した．

本分析では，実験データの教師データ 1,000 ユーザを対象に，生活習慣ベクトルを構成する活動辞書に登録した用語の出現数を時間帯ごとに集計して表示する．行動ごとの生活習慣ベクトルを構成する用語の出現数を**図 4** に示す．職業別に各生活習慣の時間帯を確認すると，特定の職業においてそれぞれ特徴が表れていることが分かる．各生活習慣について分析した結果を次に示す．

・睡眠

7 時に注目すると，いずれの職業でも睡眠に関する投稿が集中していることが分かる．多くの人がこの時間帯に起床することから，投稿に含まれる単語として「おはよう」といった起床に関連するものが多く出現したと考えられる．また，7 時における各職業の投稿数を確認すると，主婦が約 3,500 件，社会人が約 2,500 件，学生やパート・アルバイトが約 1,500 件であることが分かる．これらのことから，睡眠においては 7 時に注目することで主婦や社会人の特徴を得ることができることが明らかとなった．

・出勤

7〜8 時に注目すると，社会人による投稿数が集中しており，その他の職業と異なる特徴を示していることが分かる．これは，社会人が他の職業に比べて早い時間に出勤する傾向があることから，このような特徴の違いが現れたと考えられる．また，16 時の投稿数に注目すると，学生とパート・アルバイトの投稿数が社会人や主婦と比べて増加していることが分かる．この時間帯は，学生やパート・アルバイトがアルバイト勤務のために出勤する時間帯であることから，このような反応を示したと考えられる．これらのことから，出勤においては 7〜8 時に注目することで社会人の特徴を，16 時の投稿に注目することで学生やパート・アルバイトの特徴を得ることができることが明らかとなった．

・勤務

8 時，16 時や 23 時に注目すると，学生による投稿数の増加が顕著に現れていることが分かる．学生は学校，塾，部活やアルバイトなど様々な活動に参加しており，複数の勤務時間が存在するため，このような反応を示したと考えられる．このことから，勤務においては 8 時，16 時や 23 時に注目することで学生の特徴を得ることができることが明らかとなった．

・食事

8 時，12 時や 18 時に注目すると，主婦の投稿数が増加していることが分かる．この時間帯は一般的に食事をとる時間帯であるにもかかわらず，他の職業よりも顕著に投稿数が増加していることから，主婦は他の職業よりも食事に関する投稿を多くする傾向にあると考えられる．また，社会人に注目すると，12 時に特徴的な反応を示していることが分かる．これは社会人が昼食時に食事に関する投稿を他の職業よりも多く投稿する傾向があるためであると考えられる．これらのことから，食事においては 8 時や 18 時に注目することで主婦の特徴を，12 時に注目することで，主婦と社会人の特徴を得ることができることが明らかとなった．

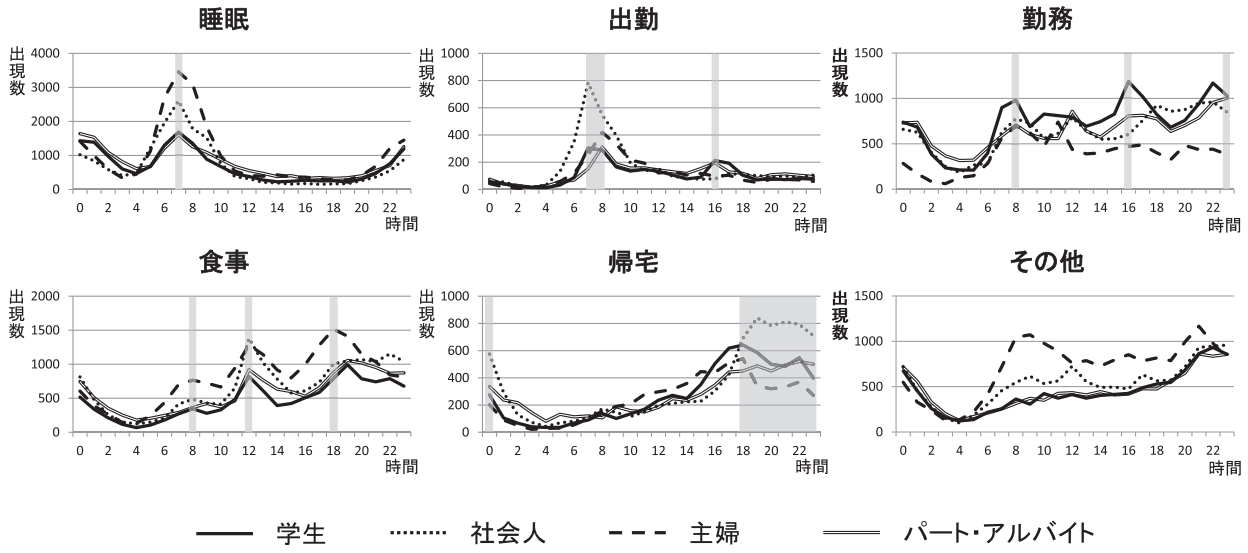

図 4　ユーザの投稿に含まれる活動辞書の用語の出現数

Fig. 4　Appearance numbers of terms of dictionary concerning activities included in users' posts.

・帰宅

18 時以降に注目すると，社会人の投稿数が継続的に増加していることが分かる．これは，社会人は残業などを行い帰宅時間が遅くなるため，継続的に増加したと考えられる．このことから，帰宅においては 18 時以降に注目することで，社会人の特徴を得ることができることが明らかとなった．

・その他

その他を確認すると，主婦の投稿数が全時間帯で増加していることが分かる．これは，主婦は比較的自由な時間を持っており，各時間を娯楽や自由な時間として行動する特徴を持つことから，他の職業に比べて特徴的な反応を示したと考えられる．このことから，その他に注目することで，主婦の特徴を得ることができることが明らかとなった．

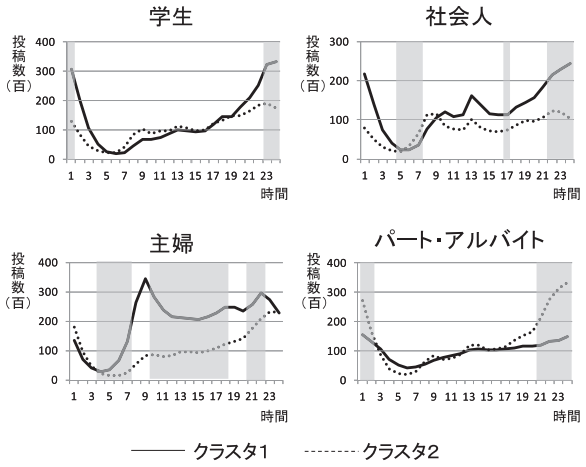

図 5　各職業のクラスタリング結果

Fig. 5　Results of clustering of each occupation.

これらの分析結果から，各職業と各生活習慣には一定の関係を表現する特徴が現れていることを確認した．そのため，いずれの職業で共通して出現する単語であっても，投稿数やその共通の単語を投稿する時間帯を考慮することで，職業ごとの暗黙的な特徴の抽出が可能であることが示唆された．このことから，生活習慣ベクトルは，各職業を特徴付けるために有効な情報源であると考えられる．その一方で，表 3 に示す実験結果で示唆されたように，ライフスタイルには特徴が現れにくい職業があることや，どの職業でも似たような投稿が集中する時間帯があることが確認された．これらの問題に対応するためには，職業ごとに推定時に用いる特徴語とライフスタイルの特徴に重み付けを変更することや，各職業で特徴的な時間のみを抽出し，それらをベクトルとして加えることで，特徴的な時間のみを効果的に学習することなどの方法が考えられる．

(3)「同じ職業でも多様なライフスタイルが存在する問題」に関する考察

本提案手法は，3.1 節であげた「同じ職業でも多様なライフスタイルが存在する問題」に対して，各職業を投稿時間に基づきクラスタリングすることで対応を試みている．

実際に，クラスタリングを行うことが精度向上につながったことを分析するため，同じ職業でも多様なライフスタイルが存在することを確認する．本分析では，実験データの教師データ 1,000 ユーザを対象にクラスタリング後のデータに対して，各時間帯（24 時間）の投稿数を集計してグラフ化する．

各職業のクラスタリング結果を時間単位に集計してグラフ化した結果を**図 5** に示す．たとえば，既存手法に比べて最も本提案手法の推定精度が向上した「社会人」を確認

すると，投稿数が増え始める 7 時頃に起床，17 時頃に帰宅し，投稿数が減り始める 24 時頃に就寝し始める社会人の特徴を示すクラスタ（社会人クラスタ 1）と，投稿数が増え始める 5 時頃に起床し，投稿数が減り始める 22 時頃に就寝し始める社会人の特徴を示すクラスタ（社会人クラスタ 2）を明確に分類できていることが分かる．また，主婦においても，朝早く起床し，昼夜投稿が多く，就寝時間も早い特徴を示すクラスタ（主婦クラスタ 1）と，日中投稿が少なく，深夜投稿が多くなる特徴を示すクラスタ（主婦クラスタ 2）を明確に分類できていることが分かる．これは，クラスタ 1 が専業主婦，クラスタ 2 が兼業主婦として分類ができていると考えられる．一方，学生やパート・アルバイトを確認すると，夜の投稿に多少特徴を示しているものの，どちらのクラスタもよく似た特徴を示しており，多様な特徴をうまく分類できていないことが分かる．そのため，発展手法の検討時には，各属性でクラスタ数を固定するのではなく，属性ごとに分類するクラスタ数を変更するなどの対策が必要であると考えられる．

7. おわりに

本研究では，ユーザの投稿内容に加え，ユーザのライフスタイルを考慮することで，マイクロブログユーザの職業属性を推定する手法を提案した．実証実験では，特徴的なライフスタイルのユーザに対しては，本提案手法を用いることで，既存手法の課題である投稿内容，人間関係，プロフィールのような「マイクロブログ上の明示的な情報だけでは職業を推定できない問題」と提案手法の検討にあたり課題となる「同じ職業でも多様なライフスタイルが存在する問題」の 2 つの課題に対し，一定の解決策を提示できた．また，実証実験の結果から，「日常的な語句から取得した生活習慣を考慮する手法」，「職業に特徴的な語句の時間的特徴を考慮する手法」と「クラスタリング手法」の 3 つの手法は，既存指標と組み合わせることで，推定精度の向上がみられることが明らかとなった．その一方で，ライフスタイルに顕著な特徴を持たない職業（「会社員，公務員，自営業」や「パート・アルバイト，無職」など）については，特徴語やライフスタイルなどでは適切に職業推定することが難しいことが分かった．このことから，発展研究では，単語を考慮した手法以外の推定手法 [6] との組合せや，ブログや Facebook などの情報と関連付ける手法 [20] との連携を試みる必要があると考えられる．

今後は，様々なユーザ属性の推定手法との組合せを試すとともに，ユーザの投稿頻度を考慮することや，社会人に分類されたユーザをより詳細に分類するための方策を検討する．これにより，社会人の職業分類を世論調査のレベルに揃え，マイクロブログのソーシャルセンサとしての利用価値の向上を図る予定である．また，本研究では，実験に用いる各職業のユーザ数を同数にして実験を行ったが，実際のマイクロブログではユーザ数には偏りがあると考えられる．職業ごとのライフスタイルのクラスタ数についても，実験時の教師データ数に大きく影響を受けていることが想定される．このことから，より現実に近い環境を再現した場合の実験もあわせて行い，実用化への検討を進めることを考えている．

参考文献

[1] 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, 人工知能学会 (2012).

[2] 奥村　学：マイクロブログマイニングの現在，言語理解とコミュニケーション研究会技術研究報告，Vol.111, No.427, pp.19–24, 電子情報通信学会 (2012).

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

[4] Eisenstein, J., 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, ACL (2010).

[5] 川中　翔, 西田京介, 倉島　健, 星出高秀, 藤村　考：ソーシャルグラフを利用したユーザ属性の推定による Twitter からのブランド特徴分析，ライフインテリジェンスとオフィス情報システム研究会技術研究報告，Vol.122, No.35, pp.121–126, 電子情報通信学会 (2012).

[6] 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, 情報処理学会 (2012).

[7] 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).

[8] 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).

[9] Pennacchiotti, M. and Popescu, A.: Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter, *Proc. Conference on Knowledge Discovery and Data Mining*, pp.430–438, ACM (2011).

[10] 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).

[11] 大倉　務，清水伸幸，中川裕志：スケーラブルで汎用的なブログ著者属性推定手法，自然言語処理研究会研究報告，Vol.2007, No.94, pp.1–6, 情報処理学会 (2007).

[12] 安田宜仁，平尾　努，鈴木　潤，磯崎秀樹：ブログ作者の居住域の推定，言語処理学会第 12 回年次大会発表論文集，pp.512–515, 言語処理学会 (2006).

[13] 池田大介，南野朋之，奥村　学：blog の著者の性別推定，言語処理学会第 12 回年次大会発表論文集，pp.356–359, 言語処理学会 (2006).

[14] Ikeda, D., Takamura, H. and Okumura, M.: Semi-Supervised Learning for Blog Classification, *Proc. 23rd National Conference on Artificial Intelligence*, Vol.2, pp.1156–1161, AAAI (2008).

[15] Oberlander, J. and Nowson, S.: Whose Thumb is It Anyway?: Classifying Author Personality from Weblog Text, *Proc. COLING/ACM on Main Conference Poster Sessions*, pp.627–634, ACL (2006).
[16] Yasuda, N., Hirao, T., Suzuki, J. and Isozaki, H.: Identifying Bloggers' Residential Areas, *Proc. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp.231–236, AAAI (2006).
[17] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.: Effects of Age and Gender on Blogging, *Proc. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp.199–205, AAAI (2006).
[18] Izumi, M., Miura, T. and Shioya, I.: Entropy-Based Age Estimation of Blog Authors, *Proc. Computer Software and Applications Conference*, pp.795–800, IEEE (2008).
[19] 中村健二，井上健治，小柳　滋：著者属性の推定結果を用いたプロフの出会い目的の書き込み検出のための教師データ自動構築手法，情報処理学会論文誌：データベース，Vol.53, No.3, pp.53–69, 情報処理学会 (2012).
[20] 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，自然言語処理研究会研究報告，Vol.2013-NL-210, No.4, pp.1–8, 情報処理学会 (2013).
[21] Mislove, A., Viswanath, B., Gummadi, K.P. and Druschel, P.: You Are Who You Know: Inferring User Profiles in Online Social Networks, *Proc. 3rd ACM International Conference on Web Search and Data Mining*, pp.251–260, ACM (2010).
[22] Wen, Z. and Lin, C.Y.: On the Quality of Inferring Interests from Social Neighbors, *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.373–382, ACM (2010).
[23] Wen, Z. and Lin, C.Y.: Improving User Interest Inference from Social Neighbors, *Proc. 20th ACM International Conference on Information and Knowledge Management Pages*, pp.1001–1006, ACM (2011).
[24] Mackay, D.: *Information Theory, Inference and Learning Algorithms*, pp.284–292, Cambridge University Press (2003).
[25] 宮崎雄一朗，山田直治，住谷哲夫，磯田佳徳：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 DoCoMo (2009).
[26] 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).
[27] 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).
[28] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol.20, No.3, pp.273–297, Springer (1995).
[29] Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
[30] Chih-Chung, C. and Chih-Jen, L.: LibSVM, available from ⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm/⟩ (accessed 2013-09-01).
[31] 株式会社 CGM マーケティング：ツイナビ，入手先 ⟨http://twinavi.jp/⟩ (accessed 2013-09-01).
[32] Twitter: Twitter Developers, available from ⟨http://dev.twitter.com/⟩ (accessed 2013-09-01).
[33] 松尾　豊，岡崎直観，中村嘉志，西村拓一，橋田浩一，中島秀之：位置履歴からのユーザ属性の推定，情報処理学会論文誌，Vol.48, No.6, pp.2106–2117, 情報処理学会 (2007).
[34] 武吉朋也，帆足啓一郎，松本一則，小野智弘：表層的特徴とテキスト特徴に基づくオンラインディスカッションの健全度定量化手法，情報処理学会論文誌，Vol.53, No.12, pp.2841–2853, 情報処理学会 (2012).
[35] 梅本和俊，中村聡史，山本岳洋，田中克己：Web 検索時の行動情報を用いたクエリ修正タイプの予測，情報処理学会論文誌：データベース，Vol.6, No.3, pp.132–147, 情報処理学会 (2013).
[36] 稲葉通将，鳥海不二夫，石井健一郎：語の共起情報を用いた対話における盛り上がりの自動判定，電子情報通信学会論文誌 D，Vol.J94-D, No.1, pp.59–67, 電子情報通信学会 (2011).
[37] 株式会社エクセレントブレス：口臭に関する「特命リサーチ」結果，入手先 ⟨http://www.excellentbreath.com/breath/2005/07/breath-medical/⟩ (accessed 2013-09-01).
[38] バリアフリー映画鑑賞推進団体シティ・ライツ：映画祭アンケート集計結果，入手先 ⟨http://www.citylights01.org/eigasai/2011.html/⟩ (accessed 2013-09-01).
[39] エン・ジャパン株式会社：アルバイト採用について，入手先 ⟨http://partners.en-japan.com/enquetereport/016/⟩ (accessed 2013-09-01).
[40] 創造社デザイン専門学校：入学前の経歴，入手先 ⟨http://www.sozosha.ac.jp/enjoy/re_entrance/⟩ (accessed 2013-09-01).
[41] 内閣府：平成 16 年度青少年の社会的自立に関する意識調査，入手先 ⟨http://www8.cao.go.jp/youth/kenkyu/syakai/2seishounen5setsu.html/⟩ (accessed 2013-09-01).

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 年同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．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 年同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 年同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．

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，同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．

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 年同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 年同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）．

（担当編集委員　馬 強）