原著論文

ソーシャルメディアユーザの職業推定手法の提案†

榊 剛史*・松尾 豊*

近年、ソーシャルメディア上のデータを分析することで社会調査やマーケティングを実現する試みが増加している。ユーザの投稿内容を分析し、マーケティングに活かすには、調査対象のユーザ属性を取得する必要がある。ユーザ属性によっては、推定しやすいものと推定しにくいものがあり、例えば年齢や性別などの属性の推定は、以前から研究が行われていた。一方、ユーザ属性のなかでも職業は取得することが難しく、これまで研究が行われていない。本研究では、ユーザの職業を判別するために、ユーザの投稿内容に加え、ユーザの自己紹介文およびユーザが他の人から付与されたタグの情報を使うことで精度を向上させる。提案手法では、サポートベクターマシン（SVM）を使い、ユーザの投稿内容や自己紹介文、被リスト名を素性として、会社員の職業推定に関して、適合率 0.85、再現率 0.77 という実用的な精度で職業判別を実現することができた。

キーワード：ソーシャルメディア分析、Twitter、プロフィール推定、職業推定

1. 序論

ソーシャルメディアの普及に伴い、ソーシャルメディア上のデータを分析することで、さまざまな現象を調査・観測する試みが数多く行われている。従来の社会調査やマーケティング手法の代わりに、ソーシャルメディア上のデータを分析することで世論調査を行ったり[2, 4]、商品の評判[10]やテレビ番組の感想[12, 8]を収集することも行われている。このような調査手法はソーシャルメディアマイニングと呼ばれる。

社会科学における調査や、企業におけるマーケティング、情報システムにおける情報推薦の手法などにおいては性別、年齢、居住地域、職業などの調査対象のプロフィール（マーケティングの分野ではデモグラフィクスとも呼ばれる）を取得することは重要な課題である。調査結果の詳細な分析のためには、調査対象のプロフィールの偏りを知る必要があり、また精密な調査分析のためには、層別抽出と線型推定により調査対象の偏りを補正する必要がある[13, 6]。表 1 にあるようにユーザのプロフィールとしては、主に居住地、性別、年齢、職業が用いられており、それらに関するデータや扱いに関する知見が蓄積している[13, 6]。

ソーシャルメディアマイニングにおいても、調査対象となるユーザのプロフィールを収集することが重要である。これまでに、居住地や性別、年齢に関する

ユーザ属性の推定はさまざまな研究において行われてきた。基本的には、ユーザの投稿内容からキーワードを取り出し、それをもとに分類器を作ることでユーザ属性を推定することができる。一方、ユーザ属性のなかでも職業は取得することが難しく、これまで研究が行われていない。それをまとめたものが表 1 である。この理由として、ソーシャルメディア上でのユーザの投稿内容は、ユーザの興味や嗜好に依存する部分が大きく、年齢や性別と比べて職業との関係は強くないためである。ユーザ属性のなかでも、ユーザに内在する性質である性別や年齢、また興味や嗜好といったことはユーザの投稿内容から知ることができるが、社会との関係のなかで構築されるようなユーザの性質や役割、例えば職業や地位といったものは、ユーザの投稿内容だけでなく、ユーザが自分自身を他者にどう紹介しているか、また他者がそのユーザをどう捉えているかという情報から推定することが有用であろう。

そこで、本研究では、ユーザの職業を判別するため

表 1 ユーザプロフィールの活用と取得可能性

プロフィール	既存分野での活用	ソーシャルメディアからの取得可能性
居住地	有り	可能 ^{7, 3)}
性別	有り	可能 ^{5, 11, 14, 16)}
年齢	有り	可能 ^{11, 14, 16)}
職業	有り	未検証
宗教	-	可能 ¹¹⁾
政治的志向	-	可能 ¹¹⁾

† A Proposal for Occupation Estimation Methods for Social Media Users

Takeshi SAKAKI and Yutaka MATSUO

* 東京大学 工学系研究科
The University of Tokyo

に、ユーザの投稿内容に加え、ユーザの自己紹介文およびユーザが他の人から付与されたリストの情報を扱うことで精度を向上させる。自己紹介文は、他者に自分をアピールする情報であり、職業に関する情報も含まれている場合が多い。また、リストの情報は、他のユーザが情報を整理する際に用いるものであり、「同僚」や「ITエンジニア」などといった形で職業に関する情報が含まれる場合がある。提案手法では、ユーザの特徴量に対してサポートベクターマシン(SVM)を用いて職業毎に判別器を生成する。機械学習の特徴量としてはユーザの投稿内容、自己紹介文及び他のユーザから付与されたリスト名を用いる。評価実験において、自己紹介文とリスト名を特徴量として会社員について分類器を生成したところ、適合率0.85、再現率0.77という実用的な精度が得られた。

本研究の貢献は次のようにまとめられる。ユーザの職業推定が可能となることで、マーケティングや社会調査において詳細な分析が可能となる。例えば、会社員や主婦、それぞれの最近の興味を知ること、それらの職業をターゲットとした商品開発が可能となる。また、ある商品に興味を持っているユーザにどの職業が多いかという知識は、その商品の広告や販路の決定に有用である。また、本研究ではTwitterだけに注目しているが、同じような考え方によりFacebookやブログなど他のソーシャルメディアにも適用可能である。

本論文は以下のように構成される。2節ではソーシャルメディアのユーザに関する属性推定の既存研究を紹介する。3節では、提案手法について詳細に説明する。4節では、評価実験を通じて提案手法の実用性について評価を行う。5節では、実験結果及び提案手法についての考察を述べる。6節では本研究の結論について述べる。

2. 既存研究

ソーシャルメディアからユーザのプロフィールを推定する研究はこれまでも行われている。以下にその概要を述べる。

ユーザの居住地推定に関して、Twitterのユーザはそれぞれプロフィールを設定しており、居住地についても記載する欄が設けられている。Hechtらはそれらを用いて、州単位、市単位でユーザの居住地を推定する手法を提案している[9]。これによれば、約4割程度のユーザは州単位で、26%のユーザは市単位で居住地推定が可能であるとしている。Chengらは、ツイート集合からその土地固有のキーワードを抽出し、確率モデルを生成することで、51%のユーザの居住地を推定することができたとしている[7]。またLarsらは、距

離が近いほど友人になりやすいという仮説に基づいて、Facebookの友人関係を用いてユーザの居住地を推定し、70%近い適合率を得ている[3]。

ソーシャルメディアから性別を推定する研究について、Burgerらはツイート中の単語や文字列を手がかりに性別を推定する手法を提案している[5]。彼らによればツイートの量やTwitter ユーザプロフィール内のメタデータ量が推定精度に影響を与えている。性別、年齢、宗教、政治的志向の4つの属性を推定する手法として、Raoらはフォロワー数やツイート内容、RT頻度などを素性として、機械学習を用いる手法を提案している[11]。その他にもソーシャルメディアからユーザの属性を推定する研究は存在する[14, 16]。

3. 提案手法

本節では各ユーザに関してTwitterにより提供される情報及び機械学習を用いて、ユーザの職業推定を行う手法を提案する。

3.1 職業推定手法の全体

ソーシャルメディアユーザの属性推定によく用いられる手法としては機械学習によって分類器を生成するものがある。あらかじめ人手によって属性ごとに分類されたユーザデータから、属性の分類に有効と思われる特徴を機械学習し、分類器を生成し、その分類器を用いてユーザを各属性に分類する手法である[14, 16]。

例えば、池田らの手法では下記に示す手順でTwitter ユーザの性別に関する分類器を生成している。

1. 予め男女に分けられたユーザ集合を用意する。
2. ユーザごとに投稿を収集し、それぞれのユーザの投稿群に出現する単語に重み付けを行う。
3. これらの各ユーザの投稿に出現する語を特徴量とし、その特徴量に対して機械学習手法の一つであるSupport Vector Machine(以下、SVM)を適用することで、分類器を生成する。

この手法により、80%以上の適合率で、男女分類、居住地域推定(日本国内を7地域に分けて分類)、年代分類(10代、20代、30代など)を行うことに成功している[16]。

本論文でもこの手法を用いることで、職業推定を行うこととする。ただし、既存手法で扱われる分類数が限られる属性(性別や年代)と異なり、職業属性については様々な分類、粒度が考えられる。例えば、会社員や医者、教師、公務員といった業種による分類や、正規社員、契約社員といった雇用形態による分類、技術職、事務職、管理職といった職種による分類など、考

えられ得る分類は多岐にわたる。

そこで本稿で提案する手法では、職業ごとに分類器を生成することを目指す。つまり、会社員であるか否か、医者であるか否か、管理職であるか否かなど、目的とする職業であるか否かを判定する二値分類器を生成することとする。本研究では、会社員かどうかの判定器を生成し、その評価を実験により行うが、基本的には任意の職業について同様の手法が適用可能である。

本稿で用いる分類器の生成手順を図2の左半分「分類器生成」に示す。図2では職業の例として会社員を用いている。詳細な手順は下記の通りである。

1. 人手により目的とする職業であるユーザとそうでないユーザに分類したユーザリストを作成し、これを訓練データとする。
2. 各ユーザに関するデータから特徴量を生成する。使用する特徴量については、後述する。
3. 訓練データの特徴量をSVMに入力し学習させることで、目的とする職業であるか否かを判定する分類器を生成する。

また、分類器を適用して、ユーザの職業推定を行う手順を図2の右半分「職業推定」に示す。詳細な手順は下記の通りである。

1. 職業を推定したいユーザリストを準備する。
 2. 各ユーザに関するデータから特徴量を生成する。使用する特徴量については、後述する。
 3. 得られた特徴量を分類器に入力し、正例と判定されたユーザを当該の職業であると推定する。
- 一般にSVMで二値分類を行う際は、正例、負例ともに200件程度あれば、十分な精度が得られることが知られている。実際、blog 著者の男女分類においては、正例と負例を合わせて約350件の学習データを用いて高い精度を実現している[15]。

本研究では、各ユーザに関する特徴量として、ユーザの自己紹介文、および被リスト名を用いる点が新規性である。ここで、Twitterのリスト機能について説明する。

Twitterのリスト機能とは、各ユーザが自分のリンクのあるユーザ(Twitterでは、リンクすることをフォロー、フォローしているユーザをfriendsと呼ぶ)を整理するための機能である。具体的には、自分のフォローしているユーザをいくつかのグループにグルーピングし、各グループに名前をつける機能である。1つ1つのグループを「リスト」、リストにつけた名前を「リスト名」と呼ぶ。

このリスト機能は、ユーザを整理する機能であるが、視点を変えれば、図1で示す通り「自分のfriendsにリスト名でタグ付けをしている」と考えることがで

図1 ユーザによるユーザへのタグ付け

図2 SVMを用いた職業推定手法

きる。つまりユーザによるユーザへのタグ付与機能であると考えられる。この考え方を用いて、ユーザの属性分析を行うサービスも存在している¹。

3.2 機械学習に用いる特徴量

ここでは学習に用いる特徴量について述べる。従来手法では、ユーザ投稿から素性を生成していたのに対し、提案手法では、ユーザ投稿、ユーザの自己紹介文、ユーザが含まれるリスト名を用いる。ソーシャルメディアから得られるユーザに関する特徴量として、これ以外にソーシャルグラフ、すなわちTwitterにおけるフォロー、被フォロー関係が考えられる。しかし、ソーシャルメディアにおけるソーシャルグラフはオフラインでのソーシャルグラフと異なることが知られている[1]²。このため、Twitterを対象とした本研

1 <http://www.mustexist.com/list-tags/>

2 <http://fitzgeraldsteele.wordpress.com/2010/07/06/online-vs-offline-social-networks/>

究ではソーシャルグラフ情報は用いない。

本稿では、データセットとして、無作為に抽出した1,230,331ユーザのTwitterにより提供される情報を用いる。

なお、属性推定のための機械学習による分類器生成において、特徴量として用いる単語の重みによく用いられる下記の tf-idf を用いる。

$$\begin{aligned} tf_{i,j} & \text{ 文書 } d_j \text{ 中での単語 } w_i \text{ の出現頻度} \\ df_j & \text{ 単語 } w_i \text{ のある文書集合 } D \text{ 内でいくつかの文書} \\ & \text{ に出現するか} \\ N & \text{ ある文書集合 } D \text{ に含まれる全文書数} \\ w_{i,j} &= tf_{i,j} \log \left(\frac{N}{df_j} \right) \end{aligned}$$

属性推定の既存研究においては、この tf-idf を単語特徴量の重みとして用いて機械学習を行うことで、高精度な属性推定に成功している [16]。

以下では、ユーザ投稿、自己紹介文、リスト名に対して、具体的に行う処理について説明する。

3.2.1 ユーザ投稿の利用

ソーシャルメディアのユーザ属性を推定する既存手法において、ユーザ投稿を用いる手法はよく用いられている [5, 14, 16]。そこで本研究でもこのアプローチを用いる。ユーザにより投稿された直近 200 投稿を収集し、それらを形態素解析して得られた形態素全てを特徴量として用いる³。なお、6ヶ月以内の投稿数が200件未満のユーザは対象外とした。これは、日常的に投稿を行わないユーザの投稿には職業的な特徴がでない、またそもそも当該ユーザがアクティブユーザでないと考えられるためである。なお、6ヶ月で200件の投稿があった場合、平均して毎日1件以上投稿していることとなるため、日常的に投稿しているユーザであると仮定する。形態素の重み付けは tf-idf を用いる。なお df は、訓練データに含まれるユーザ全体による投稿を文書全体と見なす。この際 df が全文書数の30%を超える形態素については、ストップワードとして用いない。

3.2.2 自己紹介文の利用

Twitter のユーザプロフィールには「bio」という項目がある。これは自己紹介文を記入する項目であり、文

字数は140字以内に制限されている。ユーザによってはこの項目に自分の年齢や性別、職業、趣味などを記入し、公開している。

本研究ではこの自己紹介文を用いてユーザの職業推定を行う。自己紹介文を形態素解析して得られる形態素全てを特徴量として用いる。各語の重みは tf-idf を用いる。なお df は、訓練データに含まれるユーザ全体の自己紹介文を文書全体と見なす。自己紹介文についてストップワードの除去を行わなかった。全文書数の30%を超える df を持つ形態素がなかったためである。

なお、自己紹介文に職業名が含まれているため、それをそのままユーザの職業名として用いることも可能ではないかと考えられる。しかし、自己紹介文で自分の職業を公開しているユーザはごくわずかである。例えば、本データセットにおいて、自己紹介文に「会社員」、「サラリーマン」と記入しているユーザ数は8,439 ユーザであり、全体の0.69%である。同様に「大学生」は19,642 ユーザで1.6%、「主婦」は7,709 ユーザで0.63%と低い値となっており、職業を公開しているユーザは全体と比較してごく一部であることがわかる。そこで、他の手がかりからユーザの職業を推定する手法が必要となる。

3.2.3 リスト機能の利用

他のユーザによってつけられたリスト名(以下、被リスト名と呼ぶ)をユーザに対するタグとみなし、それをユーザの職業推定の特徴量として用いる。また、ユーザ u_i にリスト名 l_m のリストに含まれる回数を以下の式で定義し、これを各リスト名の tf として用いる。

$$tf_{l_m} = N(u_i, l_m)$$

なお、何らかのリストに入れているユーザ数は、534,094 ユーザであり、全体の43.4%にあたるため、ユーザの職業推定の手がかりとしては有効である可能性は低いと考えられる。

3.3 具体例

ある1人のユーザに関して上述の素性を取得すると、表2となる(紙面の都合上、特徴量の一部のみを掲載している)。tw から始まる特徴量は、文書中に出現する形態素を意味する。bio から始まる特徴量は、自己紹介文に出現する形態素を意味する。list から始まる特徴量は、リスト名を意味する。但し、他のユーザからリスト名を付与されていない場合は、null: 1となる。各特徴量の「:」の後の数値は各特徴量の重みを意味する。なお、表2において、tf-idf の重みがユーザによらず各語において共通の値となっている場合が

3 形態素とは言語学の用語で、意味を持つ最小の単位を表す。形態素解析は文を形態素に分割し、各形態素に品詞を付与することである。日本語の文章は単語同士が区切られていないため、計算機で日本語を処理する場合は、形態素解析ツールにて文書を形態素に分解する必要がある

表2 各ユーザの特徴量の具体例

ユーザ	特徴量
ユーザ A	tw 調査:5.630 tw 絶望:3.014 tw 待つ:1.186 tw 自分:0.547 tw 少ない:1.636 tw 連投:5.283 bio 浪板:8.461 bio 毎年:6.669 bio 住む:3.526 bio 泳ぐ:7.074 bio 自分:3.866 bio お祈り:8.461 bio 釣り:4.934 listnull:1
ユーザ B	tw トラフ:6.669 tw 違う:3.025 tw チベット:10.832 tw 四国:3.939 tw 退く:5.325 tw 歓迎:3.219 bio 責任:6.669 bio 人事:6.263 bio 普通:3.076 bio 会社:1.253 bio ある:2.734 bio 社会:3.972 listnull:1
ユーザ C	tw みんな:0.797 tw 違う:1.008 tw 出張:2.234 tw 鳥歌:8.461 tw 行列:3.123 tw 荒い:4.366 tw 出走:5.688 tw テレビ:1.245 tw 記録:2.714 tw gizjp:9.646 tw 暗い:2.534 tw 画像:1.868 bio 会社:1.253 bio 経る:5.203 bio 勤め人:16.921 bio I T:6.158 bio 生活:6.873 bio 産業:7.074 bio 創出:8.461 bio わかる:4.529 bio 営業:9.308 listlivingvision:8.461

多数見受けられる。例えば、ユーザB、ユーザCにおいて「bio会社:1.253」となっている。これは、特に自己紹介文において $tf=1$ となることが多いため、 $w_{i,j} = \log\left(\frac{N}{df_i}\right)$ のように、語の特徴量の重みが語のパラメータ i のみに依存し、ユーザのパラメータ j に依存しない値となってしまうためである。

4. 評価実験

4.1 実験方法

提案手法に基づいて、ユーザの職業推定実験を行った。職業にはさまざまな分類の方法があり、例えば職種と業種を分ける方法や、より簡易に会社員・自営業・主婦・学生・その他といった形で分ける方法もあり、目的に応じて使い分けられる。本稿では、その中でも用いられることが多い分類の一つである「会社員」であるか否かを判定する分類器を生成する。本手法は汎用的なものであり、その精度に違いはあっても基本的にはどの分類に対しても適用可能である。

特徴量としては、ユーザの投稿、ユーザの自己紹介文、ユーザのリスト名を用いた。既存の研究でユーザ属性判定に用いられているのはユーザの投稿であるため、これをベースライン手法とする。

まず、データとしては、会社員1000名、それ以外の職業(主婦、フリーター、学生)1000名の合計2000名のユーザリストを用意した。各ユーザリストの生成は、それぞれ「会社員です」「主婦です」「フリーターです」というフレーズを自己紹介文に含むユーザを収集し、それをリスト化することで実現した。収集したユーザについて、200名ずつ人手で評価したところ、当該職業であるユーザの割合は100%であった。なお評価者は1名であり、評価者がユーザプロフィール及びユーザの日々の投稿から総合的に各ユーザの職業を

判断した。次に、ユーザごとに各特徴量ベクトルを算出する。ユーザ投稿、自己紹介文、被リストそれぞれに対して以下のような処理を行った。

ユーザ投稿 ユーザごとに、2013年1月時点での直近ツイート200件を収集し、形態素解析を行い、各語に tf-idf で重みづけした特徴量ベクトルを生成。df 算出のための文書集合は2000名のツイート全体とする。

自己紹介文 ユーザごとに、2012年10月時点での自己紹介文に対し、形態素解析を行い、各語に tf-idf で重みづけした特徴量ベクトルを生成。df 算出のための文書集合は2000名の自己紹介文全体とする。ただし、ユーザリストの生成に用いた「会社員」「主婦」「フリーター」という職業を表す3つの語は特徴量から自動的に除外した。

被リスト名 ユーザごとに、2013年3月時点での被リスト名を収集し、リスト名ごとに tf-idf で重みづけした特徴量ベクトルを生成。df 算出のための文書集合は2000名の被リスト名全体とする。

そして、データセットを訓練データとテストデータに分け、分類器の精度を算出する。本実験は5分割交差検定にて行う。また機械学習時の各パラメータは、分類器生成のコストを低くするために、利用するSVMツールClassias⁴のデフォルトの値を利用する。それぞれ、損失関数の収束判定の閾値を0.0001、L2正則化係数を1.0損失関数収束判定の試行回数の上限値を1000とした。

4 <http://www.chokkan.org/software/classias/>

表3 Twitter ユーザの職業推定会社員実験結果

特徴量	適合率	再現率	F 値
投稿 (baseline)	0.737	0.732	0.734
自己紹介文	0.840	0.765	0.800
リスト名	0.499	0.972	0.659
自己紹介文 投稿	0.775	0.771	0.773
自己紹介文 リスト名	0.849	0.767	0.805
全て	0.777	0.770	0.773

4.2 結果

この実験環境に基づいて行った提案手法の性能検証結果は表3の通りである。適合率、再現率、F 値、それぞれの値は、5回の試行の平均値である。

表3より、まず個別の特徴量を見ていくと自己紹介文を用いた場合に、最も分類性能がよくなることがわかる。またリスト名は適合率が約0.5、再現率が約1.0であるが、どのベクトルに対しても全て負例と判断してしまうためである。これより、リスト名单独では分類には不向きであると言える。

次にそれぞれの特徴量を組み合わせた際の結果を見ていくと、リストと自己紹介文両方を用いた際に適合率0.849、再現率0.767と最も精度が高くなる。これより、リスト名は単独では職業推定に不向きであるが、他の特徴量と組み合わせることで、職業推定に有効になると考えられる。これはリスト名を付与されているユーザが全体の4割程度で有るため、リスト名が付与されていないユーザが正例、負例ともにいるためと考えられる。

自己紹介文とリスト名の両方を特徴量として、機械学習を用いることで、0.849という高い適合率で会社員であるか否かを判定することができた。これらの特徴量について、自己紹介文は殆ど全てのユーザが持っているものである。またリスト名についても4割近いユーザが有している。これより、提案手法は多くのユーザに適用可能な汎用性のある手法であると言える。

つまり、提案手法で対象とした特徴量のうち、リスト名と自己紹介文を特徴量として機械学習を行い、職業分類器を生成することで、高い精度でユーザが会社員であるかを推定することができた。

5. 考察

実験結果より、ユーザの職業推定には、ソーシャルメディア上の投稿よりも自己紹介文および被リスト名が有効であることが分かる。それぞれの特徴について詳細に比較すると、自己紹介文とは「他者からどのように見られたいか」を表したものである。また被リ

スト名とは、他者からの付与されたタグであると考えられるため、「他者からどう見られているか」を表したものである。すなわち、職業推定に2つの有効な特徴量はいずれも意識的に提示されたユーザの特徴であると考えられる。それに対し、投稿とはユーザの日々の行動や振る舞いが表れたものである。つまり、職業推定に有効で無い特徴量は無意識的に顕在化した特徴であると考えられる。

次にユーザのプロフィールのうち、既存研究で扱われていた性別や年代、居住地と本研究で扱った職業について比較すると、性別や年代、居住地というプロフィールは生得的なものであり、一般的には他者の認識によらない、絶対的な特徴である。それに対し、職業とは社会から与えられる役割であり、他者ないし自身がユーザをどのように認識しているかによって変化する相対的な特徴である。そのため、同じユーザでも「会社員」「正社員」「マネージャ」「IT系」など様々な職業が付与される。

推定に有効な特徴量とプロフィールの特徴について合わせて考えると、既存研究で対象となっていた絶対的なプロフィールの推定には投稿のように無意識的に顕在化した特徴が有効であると推測される。それに対し、本研究で対象とした相対的なプロフィールである職業の推定には、「他者からどう見られているか／見られたいか」のように意識的に提示された特徴が有効であると推測される。

本手法を適用して職業推定を行うためには、職業推定毎にツイート分類器を生成する必要がある。そのため、実際に多様な職業推定を行おうとすると、調査者に多大なコストがかかるように感じられる。しかし、Twitter 社により提供されるSearch API⁵を利用することで、ある検索クエリをプロフィールに含むユーザを容易に収集することができる。例えば「会社員です」「フリーターです」のような定型表現を検索クエリを用いることで、それらの職業に従事しているユーザを収集することができる。実際、「会社員です」「学生です」「主婦です」という検索クエリをユーザプロフィールに含むユーザを200名ずつ収集し人手で評価したところ、収集したユーザが当該職業であった割合は100%であった。このため、機械学習に用いるユーザリストを作成し、さらにその特徴量を収集し、それらに機械学習を適用して職業分類器を生成するという過程を容易に自動化することが可能である。逆にいえば、「会社員です」のような定型表現で十分なユーザ群を収集できない職業については、本手法で職業分類器を生成

5 <https://dev.twitter.com/docs/api/1.1/get/users/search>

することは困難となる。

6. 結論

本論文では、ソーシャルメディアユーザのプロフィール推定のうち、既存研究では実現されていない職業推定に着目し、ユーザの投稿やプロフィール文から職業推定を行う手法を提案した。評価実験を通じて、会社員について適合率0.85、再現率0.77という実用的な精度で職業推定を実現した。評価実験で有効だった特徴量である自己紹介文は殆どのユーザが、被リスト名については半分程度のユーザが有している。このため、同様の手法は他の職業推定にも適用可能であるという意味で提案手法は汎用性を有している。ただし、実用的な精度が得られるかについては、各職業ごとに検証する必要があるだろう。

提案手法を用いてユーザの職業推定が可能となることで、マーケティングや社会調査において詳細な分析や結果の補正が可能となる。またユーザの日常的な行動は職業によって影響されるため、ユーザの行動推定や状態推定といった応用分野が考えられる。本研究が、様々な応用研究につながっていけば幸いである。

参考文献

- [1] Paul Adams. The Real Life Social Net-work. <http://www.slideshare.net/padday/the-real-life-social-network-v2>, 2010.
- [2] Cuneyt Gurcan Akcora, Murat Ali Bayir, Murat Demirbas, and Hakan Ferhatosmanoglu. Identifying Breakpoints in Public Opinion. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp.62–66. ACM Press, 2010.
- [3] Lars Backstrom, Eric Sun, and Cameron Marlow. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp.61–70. ACM Press, 2010.
- [4] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM '11, pp.450–453. AAAI Press, 2011.
- [5] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp.1301–1309. Association for Computational Linguistics, 2011.
- [6] Robin Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, Vol.12, No.4, pp.331–370, 2002.
- [7] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You Are Where You Tweet: A Content-based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp.759–768. ACM Press, 2010.
- [8] Nicholas A. Diakopoulos and David A. Shamma. Characterizing Debate Performance via Aggregated Twitter Sentiment. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, pp.1195–1198. ACM Press, 2010.
- [9] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Just in Bieber's Heart: the Dynamics of the Location Field in User Profiles. In *Proceedings of the 2011 Annual Conference on Human factors in Computing Systems*, CHI '11, pp.237–246. ACM Press, 2011.
- [10] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science*, Vol.60, No.11, pp.2169–2188, 2009.
- [11] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, SMUC '10, pp.37–44. ACM Press, 2010.
- [12] David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? In *Proceedings of the 2010 ACM Conference on Computer-Supported Cooperative Work*, CSCW '10, pp.589–593. ACM Press, 2010.
- [13] グロービス. MBAマーケティング. ダイヤモンド社, 2005.
- [14] 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. 情報処理学会研究報告. 情報学基礎研究会報告, Vol.2013, No.4, pp.1–8, 2013.
- [15] 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, 2006.
- [16] 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 (CDS), Vol.2, No.1, pp.82–93, 2012.
(2013年11月3日 受付)
(2014年4月26日 採録)

[問い合わせ先]

〒113-8654 東京都文京区本郷7-3-1
 東京大学工学系研究科技術経営戦略学専攻
 工学部3号館 坂田・森研究室
 榊 剛史
 TEL: +81-3-5841-8522
 FAX: +81-3-5841-8522
 E-mail: sakaki@ipr-ctr.t.u-tokyo.ac.jp

著者紹介

さかき たけし
榊 剛史 [非会員]

2004年東京大学工学部電子情報工学科卒業。2006年同大学院修士課程修了。電力会社通信部門での勤務を経て2009年月同博士課程入学。2014年博士課程終了。博士(工学)。2014年4月より東京大学特任研究員並びに株式会社ホットリンク研究アドバイザー。専門は、人工知能、自然言語処理、Webマイニング、社会ネットワーク分析。

まつお ゆたか
松尾 豊 [非会員]

1997年東京大学工学部電子情報工学科卒業。2002年同大学院博士課程修了。博士(工学)。同年より、産業技術総合研究所研究員。2005年10月よりスタンフォード大学客員研究員。2007年10月より、東京大学大学院工学系研究科総合研究機構／知の構造化センター／技術経営戦略学専攻准教授。専門は、Webマイニング、人工知能、社会ネットワーク。

A Proposal for Occupation Estimation Methods for Social Media Users

by

Takeshi SAKAKI and Yutaka MATSUO

Abstract :

Recently, more people try to perform social survey or marketing by analyzing data on social media. Such attempts are called "Social Media Mining". Attributes of social media users are not specified. Some attributes of social media users are easy to be estimated. Others are not easy to be estimated. In previous researches, they can succeed in estimating age, gender and habitants of social media users. While, it is difficult to estimate occupation among user attributes of social media. The proposed method use support vector machine to create classifier for occupation estimation. It succeed in estimating occupation with practical accuracy of 0.85 and practical recall of 0.77.

Keywords : social media analysis, twitter, profile estimation, occupation estimation

Contact Address : Takeshi SAKAKI

Dept. of Engineering, the University of Tokyo

Eng. Bld 3rd, room203, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan

TEL : +81-3-5841-8522

FAX : +81-3-5841-8522

E-mail : sakaki@ipr-ctr.t.u-tokyo.ac.jp