Twitterの投稿場所を考慮したユーザ属性推定

武田　直人† 　関　洋平††

† 筑波大学 情報学群 知識情報・図書館学類　〒305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒305–8550 茨城県つくば市春日 1–2
E-mail: †s1413134@u.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

あらまし　Twitterには，商品やサービスに対する評判が多数投稿されており，これらを分析することでマーケティング等への応用が可能である．その際に，ユーザの属性を正しく推定できれば，属性別の意見抽出が可能となる．本研究では，Twitterユーザの属性を推定する際の素性として，ユーザの投稿場所を利用する新たな手法を提案する．提案手法では，Twitterデータから得られるジオタグをクラスタリングすることで，対象ユーザの頻繁に訪れる場所での投稿割合を求め，素性とする．推定する属性の例として,「学生」,「社会人」,「主婦」の3属性を分類する評価実験を行った．その結果,「学生」で0.816のF値が得られ，ベースラインと比較して，有意な向上が見られた．

キーワード　Twitter，属性推定，位置情報，ジオタグ

1. はじめに

1.1 背　景

140字以内の短文を気軽に投稿できるTwitter(注1)は，全世界で主流のSNSとなっており，社会的なインフラとなりつつある．また，Twitterは，それぞれのユーザの考えや行動をリアルタイムに反映しやすいという特徴から，近年研究対象として注目されている．この際，性別，年代，職業といったユーザの属性が明らかとなれば，属性ごとにトレンド分析をすることで,「若年層の女性に人気のブランド」や「社会人の男性が好む車」などの情報を得ることができる．このようなTwitterを利用した意見の抽出は，従来のアンケート調査と比較して，低コストでリアルタイムな集計が期待できる．

しかし，Twitterでは，ユーザが性別，年齢，職業などの属性を明示しない場合が多く，日本で職業属性をプロフィールに記載するユーザは全体の13.62%とわずかである[1]．これらのことから，従来のアンケート調査のような属性ごとの意見抽出が困難であると考える．これらの課題を解決するために，ユーザの属性を推定する研究が数多く行われている[2], [3]．

また，近年ではGoogle Now(注2)やNAVITIME(注3)などの，スマートフォンから得られる位置情報に基づいた情報の提供を行うサービスが登場している．位置情報を利用した研究は，2.1節で述べるように，多岐にわたり，特に投稿内容や投稿時間といった他に得られる情報と組み合わせることによって，新たな知見が得られること，既存の手法よりも高精度な推定モデルを構築できることが示されている．

本論文では，Twitterの利用者について，属性ごとのユーザの位置情報に基づいた投稿場所の違いに着目し，推定を行う．また，各属性の特徴的な単語に着目した手法と各属性の投稿時間帯の差に着目した手法とを組み合わせた素性をベースラインとした評価実験を行い，提案手法の有効性を示す．

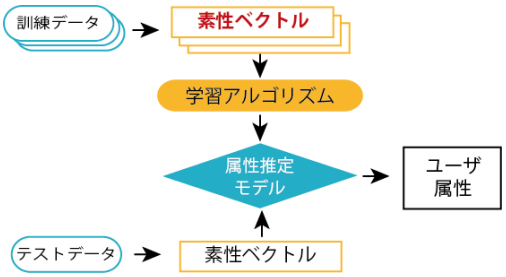

図1　属性推定手法の処理

1.2 本研究の概要

Twitterの投稿には，性別によって投稿する単語が異なること，職業によって投稿場所が異なることなど，利用者に応じて特徴が現れると考える．本研究では，こういった特徴を選択し，機械学習を用いて推定モデルを構築することで，Twitterの利用者の属性推定を行う．図1に手法の概要を示す．

属性の推定は，以下の手順で行う．

（1） ユーザ群を収集し，人手で属性のラベリングを行った訓練データを準備

（2） 訓練データのユーザ群に対して，推定に有効と考えられる素性を選択

（3） 機械学習により，素性を手がかりとした推定モデルを作成

（4） 評価するユーザの素性を，推定モデルに入力し，ユーザの属性を推定

本研究では，以下の3つの素性を採用し，属性推定を行う．

- 1時間ごとの場所別の投稿数の割合
- ツイートに出現するそれぞれの属性に特徴的な単語
- 1時間ごとの投稿数の割合

本研究では，既存研究で提案されていない，ユーザが投稿を行う場所での投稿数の割合を素性とし，属性の推定に有効か検

(注1)：https://twitter.com/
(注2)：https://www.google.com/intl/ja/landing/now/
(注3)：http://www.navitime.co.jp/

証を行う．素性の詳細については，3. 節で説明する．

1.3 論文の構成

2. 節以降の本論文の構成は以下の通りである．2. 節では，ジオタグを活用した研究と，Twitter ユーザの属性推定手法について関連研究を紹介し，本研究の位置づけを述べる．3. 節では，本研究の属性推定手法について詳細を述べる．4. 節では，任意に抽出したアカウント群に対し職業属性のラベリングを行い推定する属性を選択する．また，3. 節で述べる素性選択に必要となる，場所情報 API での場所ラベルを選択し，推定対象の属性に現れる特徴的な単語を選択する．5. 節では，属性推定に関する実験を通して，提案手法の評価を行う．6. 節では，本研究で得られた知見をまとめ，今後の課題を述べる．

2. 関 連 研 究

2.1 Twitter のジオタグを活用した研究

本研究では，属性ごとに訪れる場所に違いがあることに着目し，ユーザ単位でのツイートを集約するとともに，ユーザの投稿したツイートに現れる単語を，属性推定を行うための素性とする．Sakaki ら [4] は，ジオタグが付与されたツイートについて，実生活を観測する手段として利用し，地震や台風といった自然災害の検知を行うシステムを構築した．まず，災害に関するツイートであるか否かを SVM を利用して分類する．次に，カルマンフィルタとパーティクルフィルタを用いてジオタグ付きツイートをフィルタリングすることで，自然災害の発生場所の推定を行う．これにより，気象庁の地震速報よりも迅速な検知システムを構築した．このような実生活のイベントを抽出する研究では，イベントについて言及しているジオタグ付きツイートを収集，集約する際に，位置情報とテキストを利用する．

また，本研究では，曜日ごと，時間ごとの訪れる場所に着目し，属性推定を行う．たとえば,「主婦」属性を持つユーザは平日の昼間に居住地周辺での Twitter への投稿があるが,「学生」や「社会人」といった属性を持つユーザは少ないと考える．このような各曜日，各時間の投稿数の割合を素性とすることで，精度の高い属性推定を実現する．Ye ら [5] は，ツイートの投稿時間と投稿内容を分析することで，時間，曜日ごとに大学やパーティなどの，ツイートの話題が変動することを示した．Gao ら [6] は，こうした投稿時間のパターンを利用し，マルコフ過程に基づく位置情報を利用した手法よりも，位置情報に加えて投稿時間と投稿曜日を組み合わせた手法の方が推定精度が高いことを示した．このように，人の行動は，時間でパターン化できることが明らかとなっている．

2.2 Twitter ユーザの属性推定研究

本研究では，各属性のツイートに現れる特徴的な単語を素性とするベクトルを，提案手法との比較手法である，ベースラインの素性の 1 つとして採用する．池田ら [2] は，ツイートに現れる単語に対し，赤池情報量基準 (AIC) を適用し，各属性に現れる特徴的な単語を素性とすることで，Twitter ユーザの性別，年代，居住地の推定を行い，性別で 88% の精度を得ている．また，Cheng ら [7] は地域に現れる固有の単語を抽出し，素性とすることで 51% の精度で居住地の推定をした．榊ら [8] は，投稿内容，プロフィール文に加え，推定対象ユーザが含まれるリスト名を，他ユーザから付与されたタグ情報とすることで職業属性の推定を行う手法を提案し,「会社員」の推定を行った.「会社員」の推定では，プロフィール文とリスト名を利用した場合が最も推定精度が高く，F 値で 0.805 を得ている．

また，本研究では，投稿時間を素性とするベクトルを，提案手法との比較手法である，ベースラインの素性の 1 つとして採用する．田中ら [3] は，各属性のユーザの投稿時間に基づいてクラスタリングを行った上で，投稿内容と投稿時間を素性とし，職業属性の推定を行った．対象とする職業は学生，社会人，主婦，パート・アルバイトの 4 つで，F 値を用いた評価を行い平均値で 0.772 を得ている．

3. 投稿場所を考慮した属性推定手法

本節では，Twitter ユーザの投稿場所を考慮した属性推定手法において使用する，それぞれの素性と，素性選択に必要な前処理について説明する．

3.1 1 時間ごとの場所別の投稿数の割合

まず，ツイートに付与されたジオタグをクラスタリングすることで，ユーザが普段よく訪れる場所を抽出し，場所別の投稿数の割合を素性とする．本研究では，ジオタグのクラスタリング手法として，Ester ら [9] の DBSCAN (Density Based Spatial Clustering Algorithm with Noise) を用いる．DBSCAN は，データの集合を密度に基づいてクラスタリングする手法であるため，高密度なクラスタを抽出することができる．DBSCAN の特徴として，任意の形状のクラスタの抽出が可能なこと，データに含まれるノイズの影響が少ないこと，抽出するクラスタの数を事前に決める必要がないことなどが挙げられる．これらの特徴から，人の重要な拠点を抽出する位置情報データのクラスタリングに用いることができる．DBSCAN は，クラスタに属するデータ数の閾値である $MinPts$ とデータ間の距離の閾値である Eps の 2 つのパラメータを持つ．クラスタリングの手続きでは，半径 Eps 以内に $MinPts$ 以上のデータ数を含むような点集合をクラスタとし，どのクラスタにも属さないデータはノイズとみなす．

今回は，DBSCAN を Twitter のジオタグに適用できるように変更した以下のアルゴリズムを利用し，ユーザの普段よく訪れる場所を抽出する．

（1） 推定対象ユーザのジオタグ付きツイートを収集

（2） ジオタグの集合から，それぞれのデータ間の地理上の距離を計算し，DBSCAN でクラスタリング

（3） 抽出したクラスタのうち，ツイートを行った日数が 7 日以上ばらつきのあるクラスタを，普段よく訪れる場所として抽出

DBSCAN を利用する際のパラメータは，$MinPts = 5$, $Eps = 100m$ とした．これは，クラスタ中に 5 つ以上のデータが存在し，データはすべて 100m 以内に存在することを表す．

次に，得られたクラスタに対し，場所ラベルを付与する．ラベリングには，Yahoo! JAPAN が提供する YOLP(Yahoo!

Open Local Platform) の場所情報 API[注4] を利用する．場所情報 API は，緯度，経度をリクエストパラメータとして入力すると，付近の主要ランドマーク名やエリア名を返す API である．API のレスポンスの上位には,「六本木」,「東京ミッドタウン」,「外苑東通り」など，人がコミュニケーションの中でよく使う場所の表現が現れる．レスポンスフィールドには，複数の施設が含まれており，施設の種別ごとに設定された重要度と影響範囲によりスコアが計算され，高い順に出力される．たとえば，緯度:35.66521320007564，経度:139.7300114513391（東京都港区赤坂 9 丁目）を入力すると，表 1 のようなレスポンスが得られる．

表 1　場所情報 API のレスポンスの例

Name	Category	...	Score
東京ミッドタウン	ショッピングセンター・モール，複合商業施設	...	87.910
ガレリア	ショッピングセンター・モール，複合商業施設	...	87.366
プレッセプレミアム 東京ミッドタウン店	その他のスーパーマーケット	...	87.015
ユニクロ東京ミッドタウン店	大型専門店 (衣料品)	...	84.992
さわやか信用金庫六本木支店	信用金庫	...	73.524

今回は，場所ラベルとして，レスポンス中の “Score” と “Category” を利用する．“Score” は，その場所に与えられたスコアであり，大きいほどその場所である確率が高いことを表す．“Category” は,「大学・大学院」,「ショッピングセンター・モール，複合商業施設」など，その場所に与えられたカテゴリである．場所情報 API を用いた場所ラベルの付与は，次のように行う．

（1） 抽出されたクラスタの重心を場所情報 API に入力

（2） レスポンスフィールドの “Score” の値が 70 以上のレスポンスのうち，最も高いスコアの施設の “Category” をそのクラスタの場所ラベルと設定

（3） レスポンスフィールドの “Score” の値が最大でも 70 未満であれば，そのクラスタの場所ラベルを「None」と設定

（4） 「None」としたクラスタのうち，属する地点の数が最大であるクラスタの場所ラベルを「居住地周辺」と設定

しかし，場所ラベルの数が多いと属性推定の際にノイズとなる恐れがある．そのため，4.3 節で採用する場所ラベルの選択を行う．また，普段よく訪れる場所は曜日によって異なると考え，得られたクラスタと場所ラベルに対し，各曜日の 1 時間ごとの投稿数の割合を計算し，素性とする．

3.2 ツイートに出現するそれぞれの属性に特徴的な単語

また，推定対象となる属性を持つユーザ群が発信するツイートに現れる特徴的な単語を素性とする．素性として用いる特徴的な単語は，相互情報量によって選択する．相互情報量は，2 つの確率変数の相互依存の尺度を数値化したものである．属性と単語についての相互情報量を求め，その属性に特徴的な単語を抽出する．属性と単語間の相互情報量 $I_{(N)}$ の計算方法を式 (1) に示す．式 (1) は，推定対象の全属性と，訓練データに現れた全単語に対して適用する．

$$I_{(N)} = \frac{N_{11}}{N}\log 2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log 2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log 2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log 2\frac{NN_{00}}{N_{0.}N_{.0}} \quad (1)$$

ここで，N_{11} は訓練データの全ツイートのうち，その属性に該当し，その単語を含むツイート数を示す．N_{10} は訓練データの全ツイートのうち，その属性に該当せず，その単語を含むツイート数を示す．N_{01} は訓練データの全ツイートのうち，その属性に該当し，その単語を含まないツイート数を示す．N_{00} は訓練データの全ツイートのうち，その属性に該当せず，その単語を含まないツイート数を示す．なお，$N_{.1}$ は，$N_{01} + N_{11}$ を示す．式 (1) では，値が大きいほどその属性に偏って現れるような単語が出力される．今回は，計算した相互情報量が上位 2,000 位までの単語をその属性の特徴的な素性とし，推定対象のユーザの全ツイートに対する出現頻度を素性の値とする．また，複数の属性で同じ単語が表れた場合は，スコアの大きい属性の特徴的な単語を素性として採用する．

3.3 1 時間ごとの投稿数の割合

最後に，推定対象となる属性に特徴的な投稿時間を素性とする．具体的には，推定対象のユーザの全ツイートに対して，その投稿時間を抽出し，各曜日の 1 時間ごとの投稿数を計算する．しかし，同じ属性のユーザでも，多数ツイートを行うユーザとあまりツイートを行わないユーザが存在する．そのため，単純な 1 時間ごとの投稿数ではなく，投稿数の割合を素性とする．

4. 推定する属性と素性の選択

4.1 推定する属性の選択

推定する属性を選択するために，任意に抽出した 600 件のアカウント群に対して，著者が職業を手動でラベリングした．プロフィールだけでは職業の推定が難しいアカウントは，過去のツイートをさかのぼることでラベリングを行った．なお，リツイートしか行わないアカウントと，特定の趣味に関する話題しか投稿しないアカウントに関しては，職業属性を「不明」とした．また，職業属性が「無職」,「フリーター」とラベリングされたアカウントは数が少なかったことから,「不明」のユーザとともに「その他」とした．表 2 に結果を示す．この結果に基づき，今回は，投稿場所を考慮した属性推定の例として,「学生」,「社会人」,「主婦」の 3 属性を推定する．なお，企業・団体によるアカウントは，ジオタグ付きツイートの投稿が少なく，推定対象として除外した．

4.2 場所ラベルの選択

投稿場所ごとの投稿数を得るために，場所ラベルの選択を行う．ジオタグを 200 件以上投稿している 300 件のアカウントについて，ジオタグを 3.1 節で述べた手法でクラスタリングし，得られたクラスタの重心を場所情報 API に入力した．その結果，得られた場所カテゴリのうち，上位 15 件を表 3 に示す．

(注4)：http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/placeinfo.html

表 2　人手で付与した職業属性の割合

属性	割合
学生	**0.502**
社会人	**0.290**
企業・団体	0.108
主婦	**0.032**
bot	0.017
その他	0.051

表 3　場所情報 API により付与された場所カテゴリ

カテゴリ名	割合
None	0.331
その他のスーパーマーケット	0.046
ショッピングセンター・モール，複合商業施設	0.044
ホテル	0.028
書店	0.025
マクドナルド	0.023
ドラッグストア	0.022
その他のファミリーレストラン	0.020
駅 (JR 在来線)	0.019
駅 (他社線)	0.019
大学・大学院	0.015
小学	0.015
ファミリーマート	0.017
ローソン	0.014
セブン-イレブン	0.013

このうち,「None」は場所が得られなかった結果であるため，素性として利用しない．ただし,「None」と判断されたクラスタのうち，最も投稿数の多いクラスタに,「居住地周辺」という場所ラベルを付与し，素性として採用する．また，小学校を表す「小学」や「ファミリーマート」,「ローソン」,「セブン-イレブン」などのコンビニエンスストアは，全国各地に点在しており，採用されたクラスタの近くに，それらの施設が存在している場合，場所推定の誤りにつながるため，除外する．このような誤りは，場所情報 API の “Score” の閾値の調整により避けられると考えるが，最適な閾値の検証は今後の課題とする．さらに,「駅」には「駅（地下鉄)」,「駅（JR 在来線)」,「駅（他社線)」などがあるが，これらはすべて「駅」という場所ラベルに統合する．同様に,「マクドナルド」は「その他のファミリーレストラン」と統合する．最後に，下位に出現する「中学」,「高校」のカテゴリは「大学・大学院」と統合し,「学校」という場所ラベルとする．以上の手続きをまとめ，今回は表 3 の場所カテゴリに対し，これらの処理を行った 9 件（「居住地周辺」,「その他のスーパーマーケット」,「ショッピングセンター・モール，複合商業施設」,「ホテル」,「書店」,「ドラッグストア」,「その他のファミリーレストラン」,「駅」,「学校」）を場所ラベルとして採用する．

4.3　特徴的な単語の選択

各属性に特徴的な単語の素性選択実験を行う．今回は,「学生」,「社会人」,「主婦」の 3 属性の分類を行うため，Twitter のプロフィールに「20 歳の大 “学生”」など，それぞれの属性名が記載されている各 100 ユーザ，合計 300 ユーザのツイートを訓練データとして収集した．訓練データのツイートに対しては，MeCab(注5) を用いて形態素解析を行い，訓練データに出現する全単語について，それぞれの属性における相互情報量を計算し，素性として採用する．表 4 に各属性に特徴的な単語の一例を示す．

表 4　各属性に特徴的な単語の一例

学生	社会人	主婦
ポスト	ニュース	おはよう
NAVER	婚活	息子
まとめ	ネット	幸せ
ポケモン	活動	パパ
日本酒	社会人	我が家

5.　投稿場所を考慮した属性推定

5.1　目　　的

本節では,「学生」,「社会人」,「主婦」の 3 属性の推定に，提案した投稿場所を考慮した属性推定が有用であるか検証するために，ベースラインとの比較実験を行う．提案手法は，1 時間ごとの場所別の投稿数の割合を素性としたベクトルと，ベースラインの素性ベクトルとの結合ベクトルとする．ベースラインは，それぞれの属性に特徴的な単語を素性としたベクトルと，1 時間ごとの投稿数の割合を素性としたベクトルとの結合ベクトルとする．

また，実験の結果を踏まえ，各属性において，よく訪れる場所の数と，移動距離の差を調査する．これは，各属性で得られたクラスタ数の平均値と，クラスタの重心間の距離の平均値を求めることにより検証する．以降，実験方法，実験データ，結果について述べ，考察を行う．

5.2　実 験 方 法

属性推定の推定精度の評価尺度には，評価データに対する，全体の正解率と，属性ごとの適合率・再現率・F 値を採用する．また，評価方法は，10 分割交差検定を採用する．

分類には，SVM と Random Forest を利用する．なお，実装にはそれぞれ，LIBSVM(注6) と scikit-learn(注7) を利用した．また，RandomForest は，データをランダムにサンプリングした初期値により，結果が変化するため，10 分割交差検定を 10 回行い，その平均を評価値とする．

5.3　実験データ

実験データは，2014 年 7 月 22 日から 2015 年 7 月 21 日までの 1 年間に投稿された日本語ツイートのうち，日本におけるジオタグ付きツイートを 200 件以上投稿しているユーザとそのツイートを対象とする．収集には，Twitter の Streaming API(注8) を利用した．

(注5)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
(注6)：https://www.csie.ntu.edu.tw/~cjlin/libsvm/
(注7)：http://scikit-learn.org/stable/
(注8)：https://dev.twitter.com/streaming/overview

こうして収集したユーザのうち，今回の評価実験のために，プロフィールに「学生」,「社会人」,「主婦」の単語が現れるユーザを抽出し，人手で属性の正しさを確認した各属性 200 ユーザ，合計 600 ユーザを実験データとする．なお，クラスタリングを行った結果，よく訪れる場所が現れないユーザと，50 件以上のクラスタが抽出されるユーザはノイズとして除外した．

5.4 結　　果

表 5，表 6 にそれぞれの分類器での実験結果を示す．それぞれの分類器について，提案手法で，F 値の平均と，正解率の向上を確認した．また，3 つの職業属性のうち,「学生」,「社会人」において，F 値の向上を確認した．特に SVM を用いた「学生」の再現率と F 値では，t 検定 (有意水準 5% ，両側検定) で有意差が認められた．分類器を比較すると，SVM の方が Random Forest よりも推定精度が高い結果となった．

表 5　SVM を用いた推定精度

手法	属性	適合率	再現率	F 値	F 値平均	正解率
提案手法	学生	**0.822**	**0.810***	**0.816***	**0.800**	**0.813**
	社会人	**0.751**	0.770	**0.760**		
	主婦	0.828	0.820	0.824		
ベースライン	学生	0.754	0.720	0.737	0.789	0.775
	社会人	0.712	0.785	0.749		
	主婦	0.901	0.860	0.882		

* t-検定で有意差あり (有意水準 5%)

表 6　Random Forest を用いた推定精度

手法	属性	適合率	再現率	F 値	F 値平均	正解率
提案手法	学生	**0.758**	**0.858***	**0.797**	**0.760**	**0.768**
	社会人	**0.741**	**0.717**	**0.719**		
	主婦	0.825	0.735	0.765		
ベースライン	学生	0.738	0.846	0.778	0.748	0.758
	社会人	0.726	0.693	0.697		
	主婦	0.826	0.736	0.768		

* t-検定で有意差あり (有意水準 5%)

5.5 考　　察

結果を踏まえて，属性の推定を誤ったユーザを分析すると，ユーザの場所推定が誤っている例があった．特に，採用されたクラスタの重心の近くに，登録された施設が存在しない場合に場所の推定を誤ることがある．たとえば，学校内での投稿ではないが，近くに学校が存在している場合は，場所を「学校」と誤って推定してしまうことがある．そのため，場所情報 API のスコアに対する閾値の調整やノイズとなりそうな場所ラベルの除外が必要であることがわかった．また，それぞれの場所ラベル別に誤りやすい場所があることが考えられる．たとえば，ファミリーレストランは各地に点在しているため,「その他のファミリーレストラン」という場所ラベルは誤推定が発生しやすい．しかし，ファミリーレストラン内でのツイートは属性推定に役に立つ可能性があるため，それぞれの場所ラベルに対して，最適な場所情報 API のスコアに対する閾値を動的に与えることで，推定精度が向上すると考える．

また，誤分類したユーザには「おはよう」,「おやすみ」などの短い文章の投稿が多く見られた．さらに，Swarm(注9) のチェックイン機能を利用したツイートが多くみられた．このようなツイートは，たとえば「I'm at 新宿駅 (Shinjuku Sta.) in 新宿区, 東京都」など，現在の場所のみ，あるいは現在の場所とその場所に対する短いコメントのみが投稿される．これらのツイートが多いユーザは，特徴的な単語を素性としたベクトルが機能せず，精度が下がったと考える．

3 属性の分類結果を，属性別にみると,「学生」に関して特に有意な向上が見られたことから，投稿場所に特徴があることが分かる．これは，他の属性に現れない「学校」という場所ラベルが付与されたクラスタでの投稿があるためである．このことから，それぞれの属性に現れやすい場所を動的に得ることで,「学生」,「社会人」,「主婦」だけでなく様々な属性の推定を行うことができると考える．一方で,「主婦」はベースラインの方が高い精度であった．これは,「主婦」に関しては，他の属性と比較して投稿時間と投稿する単語に強い特徴があるためと考える．また,「社会人」はベースライン，提案手法ともに比較的低い結果となっている．表 7 に正解属性と，SVM を利用した提案手法が予測した属性の件数を示す．表から,「社会人」は「学生」,「主婦」のどちらにも誤分類しやすいことが分かる．これは,「社会人」には年代の幅があり，使用する単語や訪れる場所に共通の特徴が現れなかったためと考える．今回訓練データとして与えた社会人ユーザには 10 代から 50 代の社会人が含まれている．そのため,「学生」や「主婦」との間で誤分類が発生したと考える．

表 7　正解属性と提案手法が予測した属性

予測属性 ＼ 正解属性	学生	社会人	主婦
学生	**162**	24	14
社会人	26	**154**	20
主婦	9	27	**164**

この問題は，それぞれのベクトルに対し異なる重み付けをし，分類を誤った際のペナルティの値を変動することによる改善を検討している．具体的には，推定が難しい「社会人」の誤りに対して，ペナルティを重くすることで，F 値の平均や全体の正解率の向上が望める．

また，属性ごとに現れるクラスタ数と，クラスタ間距離の分析を行った．各属性のユーザのジオタグ付きツイートに対し，クラスタリングを行い，各属性のクラスタ数の平均と，クラスタ間距離の平均を計算した．表 8 に実験結果を示す．今回採用した 3 つの職業属性については，クラスタ間距離に大きく差があることが確認できた．これは，それぞれの属性の移動距離が異なることを示している．この結果から,「主婦」は，比較的狭い範囲で日常的な行動を行っていることが分かった．また,「社会人」は，行動半径が広いことが分かった．これは，居住地周辺での行動だけでなく，居住地から離れた職場周辺での行動が行われるためと考える．このことから，移動距離が属性推定の

(注9)：https://www.swarmapp.com/

際の素性として有効であることが示唆される．

表 8　属性ごとの平均クラスタ数と平均クラスタ間距離

属性	平均クラスタ数	平均クラスタ間距離 (km)
学生	2.8	16.42
社会人	3.0	24.17
主婦	2.5	9.66

6. おわりに

本論文では，実データを用いた実験を通して，投稿場所を手がかりとした提案手法の有効性を確認した．特に，「学生」は訪れる場所に特徴があり，高い F 値を得ることができた．一方で，「社会人」の推定は誤分類が多く，F 値が 0.760 と比較的低かった．これは，「社会人」には年代の幅があり，使用する単語や訪れる場所に共通の特徴が現れなかったためと考える．

今後の課題としては，最適な場所ラベルの選定や，場所推定の精度向上が挙げられる．場所ラベルの選定に関しては，「小学」や「その他のファミリーレストラン」など，各地に存在しているために，推定を誤りやすい場所ラベルの扱いを再検討したい．具体的には，それぞれの場所ラベルに対して，最適な場所情報 API のスコアに対する閾値を動的に与える手法を検討している．また，どの属性にも現れるような場所ラベルは，ノイズとなるため，除外することが望ましい．したがって，それぞれの属性の正解データを増やし，属性に偏って現れる場所ラベルを調査，選定する予定である．また，ジオタグを利用せず，投稿内容から投稿場所を推定する手法を検討している．具体的には，場所ラベルを正解データとして用いることで，その場所に現れる特徴的な単語を素性として選択し，投稿場所を推定する．これにより，ジオタグ付きツイートを投稿しないユーザに対しても提案手法の適用が可能となる．さらに，得られたクラスタ間の距離を考慮し，より推定精度の向上を目指す予定である．

謝　辞

本研究の一部は，筑波大学研究基盤支援プログラム（B タイプ），科学研究費補助金基盤研究 B（課題番号 25280110），萌芽研究（課題番号 25540159）の助成を受けて遂行された．

文　献

[1] 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 情報基礎とアクセス技術研究会 (IFAT), Vol.2003, No.4, 2013.

[2] 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

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

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In Proceedings of the 19th International Conference on World Wide Web (WWW 2010), pp.851-860, Raleigh, NC, USA, Apr 2010.

[5] Mao Ye, Krzysztof Janowicz, Christoph Mülligann, and Wang-Chien Lee. What You Are is When You Are: The Temporal Dimension of Feature Types in Location-based Social Networks. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp.102-111, Chicago, IL, USA, Nov 2011.

[6] Gao Huiji, Tang Jiliang, and Liu Huan. Mobile Location Prediction in Spatio-Temporal Context. Nokia Mobile Data Challenge Workshop, Newcastle, UK, Jun 2012.

[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 2010), pp.759-768, Toronto, ON, Canada, Oct 2010.

[8] 榊剛史, 松尾豊. ソーシャルメディアユーザの職業推定手法の提案. 知能と情報, Vol.26, No.4, pp.773-780, 2014.

[9] Ester Martin, Kriegel Hans-Peter, Sander Joerg, and Xu Xiaowei. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996), pp.226-231, Portland, OR, USA, Aug 1996.