Twitterユーザの投稿場所を考慮した職業属性の推定

武田 直人[†,a]　　関 洋平[††,b]

†筑波大学 情報学群 知識情報・図書館学類　††筑波大学 図書館情報メディア系

a) *s1413134@u.tsukuba.ac.jp*　b) *yohei@slis.tsukuba.ac.jp*

概要　本研究では，Twitter ユーザの職業属性を推定するために，ツイートの投稿場所を考慮した手法を提案する．まず，ユーザの投稿場所を推定するために，ジオタグ付きツイートを利用し，ジオタグ情報をクラスタリングすることでユーザが普段よくツイートを投稿する場所を求める．次に，クラスタに含まれるジオタグの中心を求め，場所情報 API を利用し，学校，駅などの場所のカテゴリを付与し，場所別の投稿割合を素性とする．評価実験では，社会人，学生，主婦の職業属性について，各 100 人ずつのユーザを収集し，投稿内容や投稿時間帯に着目した手法と提案手法とを比較した．実験の結果，投稿場所を手がかりとすることで，各属性の F 値が平均 4.2% 向上し，特に主婦の属性については，有意な向上が見られた．

キーワード　Twitter, 属性推定，ジオタグ

1　はじめに

Twitter の普及に伴い，投稿された内容から商品に対する意見を抽出しマーケティングに活用することが期待されている．Twitter を利用した意見抽出は，従来のアンケート調査に比べ，低コストでリアルタイムな集計が可能である．特に，職業属性をプロフィール文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，Twitter では，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られる．

この課題を解決するために，Twitter ユーザの属性を推定する研究が行われている [2][3]．本研究では，ユーザの投稿場所を考慮した属性推定を行う．ユーザは属性により投稿場所に違いがあると考えられる．たとえば，主婦ユーザは平日の昼間に自宅での投稿があるが，社会人ユーザの場合は少ない．本研究では，ツイートに付与されたジオタグ情報を利用することで，投稿場所を求め，属性推定を行う．また，各属性に現れる特徴的な単語に着目した手法と投稿時間帯の差に着目した手法とを組み合わせたベースラインと，提案手法との比較実験を行い，提案手法の有効性を示す．

2　提案手法

ユーザが普段よくツイートを投稿する場所での時間帯ごとの投稿割合を素性とし，職業属性の推定を行う．普段よくツイートを投稿する場所を求めるために，ツイートに付与されたジオタグを利用し場所推定を行うことで場所別の投稿割合を求め，素性とする．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．クラスタリングには分類精度が良いとされる Ward 法を用いた．次に，クラスタに含まれるツイート数が N 以上 (N=10) のクラスタを普段よく訪れる場所として採用する．得られたクラスタのうち，含まれるツイート数が最大のクラスタを自宅とする．

また，他のクラスタには Yahoo! JAPAN Web API が提供する場所情報 API[4] を用いて，学校，駅などのカテゴリを付与する．場所情報 API は，緯度，経度を入力とし，付近にある施設名や観光地名，駅名を返す API である．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，スコアの高い順に出力される．本研究では，クラスタに含まれるジオタグの中心を場所情報 API の入力とし，レスポンスされた施設のうち，もっともスコアの高い施設のカテゴリを投稿場所とする．スコアには閾値を設定[1]し，閾値以上のスコアの施設のうち，最もスコアの高い施設を投稿場所として採用する．このようにして，採用された投稿場所の時間帯ごとの投稿割合を算出し，素性とする．

3　評価実験

3.1　推定する属性の選択

推定する属性を選択するために，任意に抽出したアカウント群からランダムにピックアップした 600 ユーザの職業を人手でラベリングした．プロフィール文だけでは職業の推定が難しいユーザは，過去のツイートをさかのぼり確認す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，特定の話題しか投稿しないユーザに関しては職業属性を不明とし，数の少ない無職のユーザ，フリーターとともにその他とした．

表 1 に分析結果を示す．分析結果から，本研究では職業属性として学生，社会人，主婦の 3 つの属性を推定する．また，企業・団体アカウントや bot アカウントに関しては今後ツイート頻度などに着目し推定する予定で

[1] 3 章の実験における場所情報 API のスコアに対する閾値は 70 とした．

ある．

表 1 抽出したアカウント群の職業属性の割合

属性	割合 (%)
学生	50.17
社会人	29.00
企業・団体	10.83
主婦	3.16
bot	1.67
その他	5.17

3.2 投稿場所を考慮した属性推定

3.2.1 実験目的

提案手法の有効性を評価するために，評価実験を行った．実験データは，人手で収集した，ジオタグ付きツイートを 200 件以上投稿している学生，社会人，主婦の各 100 ユーザである．また，場所推定の際に付与するカテゴリは，自宅，学校，駅の 3 カテゴリとした．ユーザが普段よく訪れる場所の中心地を場所情報 API に入力し，レスポンスフィールドのカテゴリ名に「学校」「大学」「高校」「中学」「小学」が含まれる場合に学校とした．また,「駅」が含まれる場合にその場所を駅とした．

3.2.2 実験方法

分類器には，LIBSVM[5] を利用した．特徴的な単語を素性としたベクトルと時間帯別の投稿割合を素性としたベクトルを作成し，2 つを結合したベクトルをベースラインとした．ベースラインと，ベースラインに場所別の投稿割合ベクトルを加えたベクトルとを比較することにより提案手法の有効性を示す．特徴的な単語を素性としたベクトルは，各属性のツイートに現れた名詞について，属性と名詞間の相互情報量を計算し，上位 M 語 (M=2000) の登場回数を素性とした．また，時間帯別の投稿割合を素性としたベクトルは，1 時間ごとの投稿数の割合を素性とした．なお，評価方法は，10 分割交差検定を，評価尺度は，全体の分類精度を得るために正解率 (accuracy) を，各属性の分類性能を評価するために F 値を採用した．

3.2.3 実験結果

表 2 に実験結果を示す．3 つの職業属性について提案手法で精度向上ができた．また，正解率においてもベースラインの 0.634 に対し，提案手法では 0.689 となり向上が見られた．この値は，有意水準 5% における t 検定で有意差が認められた．属性別にみると，主婦ユーザは 0.750 の F 値を得ることができ，有意水準 5% における t 検定で有意差が認められた．このことから，主婦ユーザは投稿場所に特徴があることが分かる．主婦の場合は，平日の昼間の自宅での投稿が学生や社会人に比べて特徴的であるため，高い F 値が得られたと考えられる．一方で，学生の F 値がベースライン，提案手法のどちらも低い．これは，学生と若い社会人において使用する単語にあまり差が生じないためだと考えられる．実際に，今回訓練データとして与えた社会人ユーザには，学生より年下の 10 代の社会人がある程度含まれている．そのため，それぞれのベクトルに対し適切な重み付けをすることでより高い精度をあげることができると考えられる．

また，属性の推定を誤ったユーザを分析すると，投稿場所の推定が誤っている例があった．採用されたクラスタに含まれるジオタグの中心の近くに，登録された施設が存在しない場合に場所の推定が誤っていることがわかった．たとえば，社会人や主婦の投稿場所のうち，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い，属性が学生と誤分類されてしまう場合があった．そのため，場所情報 API のスコアに対する閾値の調整が必要であることがわかった．

表 2 提案手法とベースラインの推定精度 (F 値)

	提案手法	ベースライン
学生	0.566	0.550
社会人	0.703	0.659
主婦	0.750*	0.684
平均	0.673	0.631

* t-検定で有意差あり (有意水準 5%)

4 まとめ

本研究では，Twitter を対象とし，ユーザの職業属性について推定を行った．具体的には，投稿場所ごとの投稿割合を考慮し，有効性を示した．今後は，投稿した曜日を考慮し，平日に投稿する場所と休日に投稿する場所を得ることで，推定精度の向上を目指す．また，bot アカウントや企業，団体アカウントが含まれたアカウント群を対象にした実験を予定している．

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．

参考文献

[1] 伊藤淳，西田京介，星出高秀，戸田浩之，内山匡:Twitter と Blog の共通ユーザおよび会話ユーザの同類性に着目した Twitter ユーザ属性推定，DEIM Forum 2013, C5-3, 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，Vol. 6, No. 5, p. 71-84, 2013.

[4] 場所情報 API. http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/placeinfo.html (参照 2015-10-13)

[5] LIBSVM. https://www.csie.ntu.edu.tw/~cjlin/libsvm/ (参照 2015-10-13)