Twitter ユーザの投稿場所を考慮した属性推定

武田 直人

Twitter をはじめとした，短文を気軽に投稿できるマイクロブログサービスは，様々な商品やコンテンツへの意見，感想がリアルタイムに反映される．しかしながら，マイクロブログでは，ユーザが性別，年代，職業といった属性を明示しないことが多く，属性ごとの意見の抽出が困難である．そこで，本研究では，Twitter ユーザを対象とした属性推定を行う．属性推定には，既存研究で提案されていない，ユーザがよく訪れる場所での投稿数の割合を素性としたベクトルを利用する．

提案手法では，ユーザの投稿場所を明らかにするために，ツイートに付与されたジオタグを利用する．ジオタグはクラスタリングし，得られたクラスタに属するツイートの数とその投稿日数とを分析することで，そのクラスタをユーザがよく訪れる場所として採用する．こうして得られたクラスタに対し，場所情報 API を利用して，場所ラベルを付与する．場所ラベルは，「大学・大学院」，「ショッピングセンター・モール，複合商業施設」などの，その場所に与えられたカテゴリを表す．また，普段よく訪れる場所は曜日によって異なると考え，場所ラベルが付与されたクラスタに対し，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．

投稿場所を考慮した属性推定の有効性を検証するために，ベースラインとの比較実験を行った．実験では，「学生」，「社会人」，「主婦」の 3 属性について，各 200 人のユーザを収集し，提案手法とベースラインとを比較した．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した．また，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した．SVM を用いた実験の結果，投稿場所を手がかりとすることで，正解率が 3.8% 向上した．また，属性別には，「学生」で 0.816 の F 値が得られ，有意な向上が見られた．さらに，各属性の得られたクラスタ数とクラスタ間距離の平均値を計算することで，属性によってよく訪れる場所の数と，その移動距離の差を調査した．実験の結果，移動距離に差があることが分かり，「社会人」では，24.17 km と移動範囲が他の 2 つの属性と比較して広いことが分かった．

以上のように，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の確認ができた．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，付与率の低い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．また，bot や企業，団体アカウントを含む，任意に抽出したアカウント群を対象とした評価実験を行う予定である．

（指導教員　関 洋平）