Twitter の複数アカウント所持者を対象とした 投稿分類手法に関する研究

野崎　祐里

マイクロブログサービス Twitter は，アカウントを実名で登録する必要がないため，趣味用や裏アカなど，目的に応じて複数のアカウントを作成して使い分けをしているユーザが存在する．また，企業の Twitter 担当者は，自分の私的アカウントと，公式アカウントの 2 種類のアカウントを管理していることがある．このような複数アカウントを所持しているユーザは，ツイートを投稿するアカウントを間違えることがある．アカウントの選択ミスは，本名を使用しているアカウントと，そうでないアカウントとの関連付けが推定されることや，公式アカウントに私的な内容をつぶやいて信用を失うことがある．このような問題を防ぐために，機械学習を活用してツイートから投稿するアカウントを推定する手法を提案する．

機械学習を利用するためには，ツイートから素性を抽出しベクトルの形に変換しなければならない．本研究では，複数アカウント所持者は特定の目的に応じて使い分けをしていることに着目し，素性を選択する．具体的には，形態素の出現頻度を素性として用いる．形態素の品詞は，内容語の名詞，動詞，形容詞，副詞に加えて記号も素性に使用する．さらに，投稿形態を反映した素性や，単語以外の内容を表す素性を拡張することによってより分類精度の向上を目指す．投稿形態を反映した素性は，ツイートの文字数，リプライの有無，ハッシュタグの有無，リツイートの有無，URL 有無，画像の有無の 6 種類で，単語以外の内容を表す素性は，ツイートの感情極性スコア，ユーザのスクリーンネーム，ハッシュタグのキーワード，URL 中のドメインの 4 種類である．

評価実験を行うために，Twitter の複数アカウント所持者のツイートを収集する．本研究では，Twitter のプロフィール文に同一人物の別のアカウントが記載されているアカウントと，記載された先のアカウントを複数アカウント所持者とする．アカウントの使い分け方として，公式-普段，裏アカ-普段，趣味-普段の 3 種類に人手でラベル付けを行い，それぞれの使い分け方について 20 ペアずつ，計 60 ペアから 1 アカウントにつき 1,000 ツイートを収集した．機械学習の手法は Support Vector Machine(SVM)と Random Forest(RF)の 2 種類を利用し，5 分割交差検定で評価する．結果は著者推定の素性で用いられている素性である文字 3-gram と比較を行う．実験の結果，SVM，RF どちらにおいても形態素解析器で抽出した素性と拡張素性を組み合わせた提案素性の方が，文字 3-gram より高い分類正解率を算出した．特に SVM において，提案素性で 0.828 と最も高い正解率を出し，本手法の有効性を示すことができた．

（指導教員　佐藤哲司）