3M-3

ユーザーツイート解析による人物像推定手法の提案と検討

長浜　祐貴† 遠藤　聡志† 當間　愛晃† 赤嶺　有平† 山田　考治†
Yuuki Nagahama Endo Satoshi Toma Naruaki Yuuhei Akamine Koji Yamada

1. はじめに

現在、日本の SNS 利用者は約 4,965 万人である。その中でも、本名を原則としない SNS の中で最も利用者が多いのが Twitter である。Twitter は最大 140 文字の短文を投稿できる SNS であり、投稿の気軽さと読みやすさを兼ね備えている。その特徴から、利用者は約 1,291 万人にも上り、SNS 全体では Facebook に次ぐ利用者数である。

Twitter に投稿される文章(ツイート)は、ユーザ自身の興味、他ユーザとの交流、周囲の出来事などを口語的な文章で投稿することが多い。そのため、使用する単語、表現、言い回しといった文章特徴の中にユーザ自身の特徴や、性別や出身、趣味等の属性が表れやすいと考えられる。そこで、ユーザの語彙や文章表現を解析することでユーザの人物像を推定出来ると考えられる。そうして得られるユーザの人物像は、個人向けサービスの基礎データ、特定の話題に反応するユーザ属性の調査、それらを応用したマーケティングなど応用範囲が広い。

本稿では人物像推定に関して、ユーザの出身地を推定する手法について検討、提案し、実際の推定結果と、その精度について考察を行う。

2．人物像推定手法

Web 上の文章からユーザの人物像を推定する研究として池田ら(1)は、ブログの著者を性別を「男性」「女性」「性別不明」の 3 クラスに分類する手法を提案した。この手法は、ブログの投稿記事から「一人称代名詞」「機能語」「全形態素」を素性、χ二乗値を素性値として、SVM に学習させることで性別推定の分類器として用いる手法である。学習には著者の性別が明記されているブログを使用している。実際の精度評価実験では、男性には再現率 0.79、精度 0.91、女性には再現率 0.81、精度 0.95 の結果を得ている。

池田らの手法は、長文が多く文語的に書かれるブログを対象に行なった実験である。そこで我々まず、短文が多く、口語的な文章である Twitter でも池田ラの手法が有効か、人物像推定に有用な素性が他に存在するかどうかについて検証、実験した(2)。

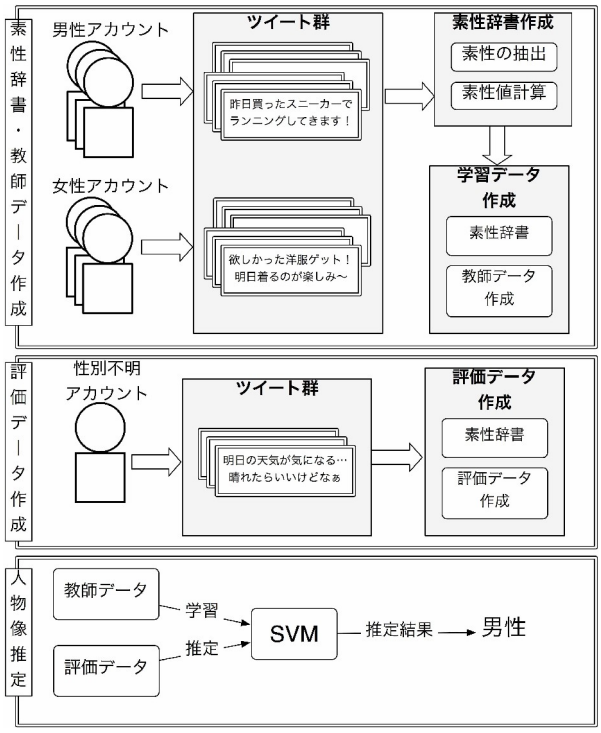

図 1:性別における人物像推定手法

2.1 Twitter における性別推定手法

Twitter の性別推定手法について図 1 に示す。まず、男女が明らかになっているアカウントを教師アカウントとして収集し、それらのツイートを取得する。ツイートから素性を取得し、素性値を計算した結果を素性辞書として保存する。その素性辞書を SVM に学習させる教師アカウントのベクトル化と、性別を推定したいアカウントのベクトル化に使用する。実際に検討した素性と素性値について、表 1 に示す。

表 1:実験を行った素性と素性値

	素性	素性値
実験 1	単語	χ二乗値
実験 2	品詞	出現割合
実験 3	品詞の並び	出現割合
実験 4	品詞の並び	χ二乗値
実験 5	特定の品詞並び直後の単語	χ二乗値

検討した素性は、ツイートの形態素解析によって得られる「単語」と「品詞」、品詞の出現パターンを表す「品詞の並び」、推定に有用な単語の出現条件を考慮する「特定の品詞並び直後の単語」の 4 つである。特に「単語」は池田らの手法で使用した

User profile method by tweet analyses
† University of Ryukyus

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

「形態素」とほぼ同様のものである。また、「特定の品詞並び直後の単語」は、先行研究では用いられていない提案である、詳細は次節で説明する。
素性値としては、「出現割合」と、他のカテゴリと素性の出現頻度を統計的に比較することで得られる値「χ二乗値」の2つを利用した。

2.2 特定の品詞並び直後の単語

「特定の品詞並び直後の単語」は、単語の品詞と単語の出現条件によって、人物像の推定に有用な単語が特定出来ると仮定したものである。
今回は条件として、χ二乗値の高い上位100単語を集計から多かった品詞と、χ二乗値が高い単語の直前2品詞について着目し、頻出した2品詞を集計、それらに続く単語を素性とした。

2.3 性別における人物像推定実験

教師データに男女100人、評価データに男女100人を重複せずに用意し、性別推定を行った結果を表2に示す。最も精度が高かったのは実験5の「特定の品詞並び直後の単語」を素性、「χ二乗値」を素性値とした場合であった。

表 2:性別推定実験の素性数と精度

	使用素性数	精度
実験5:特定条件	194単語	81.3%
実験1:単語χ二乗値	361単語	79.3%
実験2:品詞割合	12品詞	68.7%

このときの特定条件は「品詞が名詞か記号」「χ二乗値が高い単語を導く直前2品詞に続く」の2つの条件である。　以上の結果から、特定条件を満たす単語を素性として用い、男女各50人のアカウントから各50ツイートのデータを基に分類器を作成することで、約8割の精度で性別を推定することが出来ることがわかる。
以上の結果は性別の推定の結果であり、他の属性へ応用が可能か不明である。そこで、対象を出身地にした場合の素性の有効性を確認する。

3. 出身エリア推定実験

出身エリア推定の簡易実験として、エリアを限定した関東エリアと九州エリアの出身推定を行う。関東地方の一都六県を関東エリア、九州の7県を九州エリアとし、教師データ、評価データとしてTwitterの出身地欄に都道府県名が含まれているアカウントを使用する。取得ツイート数は50ツイートとした。

3.1 実験概要

今回は、素性に単語、素性値にχ二乗値を用いた実験1を行った。この時、教師アカウント数を50人、100人、150人の3つのパターンで実験し、同様の実験を3回行うことで、精度の評価を行った。

3.2 実験結果

出身エリア推定実験の結果を表3に示す。

表 3:出身エリア推定実験(単語:χ二乗値)

	1回目	2回目	3回目	平均
50人	56%	54%	53%	54.3%
100人	60%	56%	52%	56.0%
150人	55%	46%	54%	51.6%

最も精度が良かったのは教師データを100人とした場合の56%であったが、精度が良いとは言えなかった。参考として、χ二乗値が高かった名詞の単語を表4に示す。

表 4:χ二乗値の高い単語

関東エリア	χ二乗値	九州エリア	χ二乗値
群馬	3.03	福岡	37.64
午後	2.27	さん	33.68
納得	2.27	今日	31.22
北海道	2.27	佐賀	30.69
初回	1.89	日	28.47

表4の結果から、「群馬」「福岡」「佐賀」といったエリア特有の単語が取得出来ている事が確認できるが、出身エリアとは直接関係無さそうな単語も推定に有用な単語として扱われている。

4. まとめ

出身エリア推定において、単語を素性、χ二乗値を素性値とした実験では精度が低かった。その理由として、出身エリアを推定するのに有用な単語を上手く取得出来なかったためだと考えられる。その原因として、用いた教師アカウントが不適切であった、２つのエリアとの比較だけでは地域性のある単語を上手く取得出来ない、ノイズが多い等の理由が考えられる。
研究の今後として、性別推定で行った実験のうち、出身エリアの実験で行わなかった他の４つの実験を行い、精度の確認を行う。また、性別推定によって得られた属性を出身エリアの推定に応用する手法について考察する。

5. 参考文献

(1) 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(2006).

(2) 長浜祐貴,遠藤聡志,當間愛晃,赤嶺有平,山田考治:”ツイート解析による性別推定に有用な因子の検討”,FIT2013 第12回情報科学フォーラム(2013)

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.