E-007

ツイート解析による性別推定に有用な因子の検討
Estimation of personal features by Tweet analysis

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

Twitter は最大 140 文字の短文を投稿できる SNS である.その 140 文字という制限は,投稿の気軽さと,読みやすい文章量という特徴があり,Facebook と並び普及率が高い.

そのような Twitter に投稿される文章(ツイート)は,周囲の出来事,感情等を口語的な文章で投稿することが多く,その言い回しや単語はユーザの普段の語彙力や話し方の癖が表れると考えられる.そこで,ツイートから性別や出身地等のユーザの属性を推定出来るのではないかと考えた.ツイートを解析することで得られるユーザの属性は,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.

本研究では,ツイートから取得できる因子について有用性を検討し,実際にツイート解析を用いた性別推定を行う.

2．先行研究

性別推定の研究として,池田ら(1)はブログの投稿記事から著者の性別を”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,著者の性別が明記されているブログを元に,「俺」「あたし」といった一人称代名詞,「～ね」等の機能語,全形態素のχ二乗値から値が高い上位一定数の形態素を性別推定の因子として用い,SVM(2)を用いて分類器を作成,精度を判定した.その結果,男性クラスの再現率 0.79 に対し精度 0.91 を,女性クラスでは再現率 0.81 に対し 0.95 の精度を得た.

しかし,池田らの手法は長文記事が多く,整った文章で投稿されるブログを元にした学習であり,短文の集合や口語文章が多い Twitter では精度が悪くなる可能性が考えられる.

そこで,池田らが使用した因子が Twitter にも有用なのかを確かめる.

3．検討する因子の種類

ツイートという文章から得られる因子について考える.文章を形態素解析することで得られる主な結果は,単語とその品詞である.それらを性別推定の因子としてどのように利用するかを考える.今回考察するのは,以下の因子である.複数の因子を検討した.

- 単語のχ二乗値

実際にツイートに含まれている単語そのものを性別推定の因子として利用を検討する.

- ツイートに含まれる品詞の割合

ツイート中に含まれる品詞の出現割合を検討する.

- ツイートに含まれる品詞並びの割合

ツイート中に含まれる品詞並びの出現割合を検討する.

- ツイートに含まれる品詞並びのχ二乗値

ツイート中に含まれる品詞並びのχ二乗値を計算し,その並びと値を因子として検討する.

- 特定の品詞並び直後の単語

性別推定に有用な情報量が高い単語は,特定の品詞並び直後に表れると仮定し,検討する.

今回は,この五つの中の上から三つの因子を用いた性別推定を行った.その精度の比較を行う.

4．提案手法

因子の有用性を確認する手法として,SVM を用いた性別推定を提案する.まず,Twitter 上で性別が判明しているアカウント群を用意する.そのアカウント群のツイートに対し,形態素解析器 MeCab を用いて形態素解析を行う.その文章の解析結果から,検討する因子にあたる部分を用いベクトル化,教師データとして SVM に学習させ,性別推定を行う分類器を作成する.その分類器の性能評価として,教師アカウント群と重複しない評価用アカウント群の性別を推定し,その推定性能結果で因子の性別推定における有用性を評価する.

それぞれのアカウント群をベクトル化する手法は,使用する因子によって異なる.そのベクトル化の方法について説明する.

4.1　単語のχ二乗値

男性と女性では興味のある話題や趣味嗜好が異なるとすれば,ツイートで使用される単語にも男女で違いが表れると考えられる.そこで,男女が使用する単語の違いを性別推定に利用する方法について述べる.

まず,教師アカウントの全ツイートに形態素解析を行う.その結果から,出現した全単語,単語それぞれの出現回数を得る.次に,その得られた全ての単語に対して,男性と女性をカテゴリとしたχ二乗値を計算する.最後に,単語とχ二乗値をそれぞれ素性と素性値として用いベクトルを作成し,教師データアカウントをベクトル化する.

評価用アカウントのベクトル作成にも,教師データアカウントで使用した単語とχ二乗値を用いる.評価アカウントのツイートにその単語が含まれるかでベクトル化を行う.

χ二乗値が高い単語ほど単語の情報量は多い.そこで今回は,全ての単語で推定を行う場合と,自由度 2 におけるχ二乗検定で有意水準 0.05 とした場合のχ二乗統計量である 3.841 以上の単語のみを素性として用いた 2 つについて比較する.今回の単語数は,全単語 14014 単語,χ二乗値 3.841 以上の単語は 665 単語であった.

4.2　ツイートに含まれる品詞の割合

ツイート文は,顔文字が含まれている文章,淡々とした文章など,ユーザによって書き方は異なる.この文章の違いは品詞の出現割合に表れると考えられる.そこで,男女のツイート文は,その文章の品詞割合で推定出来ると予想し,品詞割合を性別推定に用いる方法について述べる.

各アカウント毎に出現した品詞の出現数を計算する.使

† 琉球大学

用した品詞は以下の 12 個である.

名詞,動詞,形容詞,副詞,連体詞,助詞,接頭詞,
助動詞,接続詞,感動詞,記号,フィラー

以上の品詞を素性,全単語数を元にした各品詞の出現数の割合を素性値とした.

4.3 ツイートに含まれる品詞並びの割合

4.2 で述べたように文章の違いが品詞で表れると考えた場合,男女のツイート文は品詞の使用順序の並びにも違いが表れると考えられる.そこで,品詞並びの割合を性別推定に利用する方法について述べる.

まず,各アカウント毎に,出現した品詞並びの出現数を調べる.以下に品詞の並びを得る例を示す.

今日(名詞)-は(助詞)-良い(形容詞)-お(接頭詞)-天気(名詞)-です(助動詞)-ね(助詞)

以上の文章から 3 つの品詞並びを考える場合,以下の五つが得られる.

今日(名詞)-は(助詞)-良い(形容詞)
は(助詞)-良い(形容詞)-お(接頭詞)
良い(形容詞)-お(接頭詞)-天気(名詞)
お(接頭詞)-天気(名詞)-です(助動詞)
天気(名詞)-です(助動詞)-ね(助詞)

アカウント毎に出現した品詞並びを数え,全品詞並びの出現数から算出した各品詞並びの出現割合を性別推定に利用する.

5. 実験方法

以上に示した方法で教師アカウントと評価用アカウントをベクトル化し,SVM に学習,分類させて精度を評価する.

教師データ用のアカウントは,男性 44 人,女性 45 人の計 89 人のアカウントを用意した.それらの 6 月 24 日から 6 月 26 日の 3 日間までのツイート群を取得し,それを利用する.評価用アカウントは,男女各 50 人ずつ,計 100 人のアカウントを教師用アカウントとは重複しないように用意した.それらのツイート群は,24 日,25 日,26 日のそれぞれの直近 50 ツイートである.

5.1 実験結果

実際に性別推定実験を行った結果を表にまとめた.

表 1. 性別推定実験結果

	24 日	25 日	26 日	平均
全単語	73.00%	73.00%	73.00%	73.00%
単語 3.841 以上	74.00%	72.00%	73.00%	73.00%
品詞割合	64.00%	63.00%	63.00%	63.33%
品詞並び	59.00%	55.00%	59.00%	57.67%

今回の実験で最も精度が良かったのは単語を用いた場合であった.χ二乗値が高かった上位 5 単語を男女毎に表 2 に示す.

表 2. χ二乗値が高い上位 5 単語単語

男性単語	χ二乗値	女性単語	χ二乗値
僕	21.75	`	21.87
報告	15.57	ちゃん	19.42
切り	15.57	*	18.94
終了	15.57	∀	17.94
として	14.53	ゃん	17.71

品詞の割合,品詞並びの割合を用いた性別推定の結果は,精度が高くなかった.これは,男女全体で使用される品詞の特徴は表れたが,性別推定に有用ではない品詞まで考慮してしまったためであると考えられる.表 3 に,男女全体の品詞使用割合で最も差があった上位の品詞を示す.

表 3.単語割合

男性品詞	割合	女性品詞	割合
名詞	0.71%	名詞	0.60%
記号	0.04%	記号	0.18%
動詞	0.15%	動詞	0.13%
形容詞	0.03%	形容詞	0.04%
感動詞	0.01%	感動詞	0.01%

表 4 に,男女別でχ二乗値が高かった上位 100 単語の品詞割合を表した.男性と女性ではχ二乗値が高くなりやすい品詞に大きく違いが表れたのがわかる.このことから,品詞情報も性別推定を行うのに有用な可能性がある.

表 4. 男女別,χ二乗値上位 100 単語の品詞割合

男性品詞	割合	女性品詞	割合
名詞	72.00%	記号	51.00%
動詞	10.00%	名詞	27.00%
接頭詞	6.00%	動詞	9.00%
形容詞	3.00%	形容詞	7.00%
助詞	3.00%	助詞	3.00%
副詞	3.00%	感動詞	2.00%

6. まとめ

今回の実験では,因子として単語を利用した場合の精度が最も良い結果になったが,品詞にも男女に特徴が表れたため,因子として検討する余地があると考えられる.

今後は,検討する因子の種類で挙げた候補である,ツイートに含まれる品詞並びのχ二乗値と,特定の品詞並び直後の単語について検討し実験を行う.

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) 高村大也, 松本裕治:“SVM を用いた文書分類と構 成的帰納学習法”, 情報処理学会論文誌 Vol44 No.Sig3,p1-p10.