1T-03

SNS 上への発言の特徴分析に基づくユーザの属性推定

浅妻 佑弥† 山下 晃弘† 松林 勝志†
東京工業高等専門学校 情報工学科†

1. はじめに

SNS(Social Network Service)とは，個人が互いにコミュニケーションを取ることのできるサービスのことを一般的に指す．従来の blog や HP などとは異なり，全てのユーザが相互に情報を発信することができるのが特徴である．2010 年代から瞬く間に普及し，今や米国ではインターネットを利用している大人の 75%超が何らかの SNS を利用している [1].

しかしながら，多くの SNS において，使用するユーザは自分の年齢・性別・職業・名前などの属性を正確に記入する必要がないため，積極的に正確な情報を入力するユーザは少数である．一方で，企業は SNS の情報を用いて自社の製品やサービスの広告戦略を行う動きが，近年広がりを見せている．しかし，現状では企業などが広告を発信する際に，ユーザの属性を自動的に把握することが困難なため，効果のあるユーザに届かないまま意味のない結果に終わってしまうケースも十分考えられる．また，SNS を用いたマーケティング調査なども人手に頼る部分が多くを占めている．

もし容易にユーザの属性を推定することができれば，狙いたいユーザ層に絞って広告を発信することが可能になるため，今後 SNS における企業の広告戦略がより一層豊かなものになると考えられる．また，自社製品への評判や意見などを効率的に抽出することが可能である．

そこで本研究では，ユーザ自身が投稿しているメッセージを利用し，そのユーザがどのような属性の人物であるかを推定する手法について調査・検討する．

2. 先行研究

先行研究として，Claudia Wagnar らが 2013 年に SNS 上でのユーザの職業について多クラス分類を行っている [2]．本論文では，問題を二値分類に特化させることで，分類性能の向上を図る．

3. 実験

3.1. 実験に使用するデータセットの選定

実験では，あるユーザが高専に所属しているかを推定することによって評価を行った．この条件で実験を行う理由として，高専生は比較的プロフィールに所属している学校を書いていることが多い，もしくはリスト等でまとまっていることが多いため，学習における教師ラベルとして活用しやすいと考えられるのと，思想や趣味の似ている人物が多いため，抽出の際にユーザの特徴が取りやすいと考えられるためである．

また，使用する SNS として Twitter[1] を選択した．API を用いたデータの収集が容易であることと，Twitter を使用している高専生が多いと考えられるためである．

3.2. データセットの収集

Twitter 社が提供する API を用いて，高専生と非高専生ユーザの情報とツイートの収集を行った．

（1）高専生ユーザ情報の収集

ある代表高専生ユーザを根，ユーザを葉，フォロワー関係を辺とする木構造に対して，深さ 2 までに含まれるユーザの収集を行った．収集したユーザに対して，高専に関連する単語がプロフィール等に最低1つでも含まれているか正規表現を用いて判定し，満たすユーザに対して手動で高専生であるかの確認を行った．

（2）非高専生ユーザ情報の収集

Streaming API を使用してある一定時期のツイートを取得し，得られるユーザ情報からツイート数などの条件を満たしたユーザについて，高専生ユーザ数と同じだけ抽出した．

（3）ツイートの収集

集めたユーザに対して，過去のツイートを可能な限り収集した．最終的に収集できたユーザ数と総ツイート集を表 1 に示す．

表 1. 収集したユーザとツイートの数

	ユーザ数	ツイート数
高専生	2939	4023804
非高専生	2939	6277300

† Department of Computer Science, National Institute of Technology, Tokyo College
[1] https://twitter.com

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.

3.3. データの前処理

収集したツイートに対して，正規表現を利用してURLやハッシュタグの除去を行い，形態素解析ソフトウェア MeCab[2]を利用して名詞以外の単語の除去を行った．また，実験のため，訓練：テスト＝８：２の割合でデータセットを分割した．

3.4. ユーザ特徴ベクトルの作製

ユーザごとにツイートを一纏めにしたものをBag of Words ベクトル化した後，TF-IDF 値による重み付けを行い，ユーザごとの疎な特徴ベクトルを作製した．さらに，疎な特徴ベクトルに対して，ラベルに対する各特徴量の重要性をLasso 回帰 [3]によって求め，重要度の高い特徴量のみを残した密な特徴ベクトルへの変換を行った．これによって，特徴ベクトルの次元は157259 次元から 1245 次元まで縮小された．

3.5. ユーザの属性推定

ユーザ特徴ベクトルを入力ベクトル，高専生であるかを出力ラベルとして，ロジスティック回帰 [4]とランダムフォレスト [5]を用いた分類器を訓練し，ユーザの属性推定を行った．分類機の各パラメータは，10-fold cross-validation と網羅探索によって最適な値を導出した．パラメータを表２に，それぞれの Precision, Recall, F-score と全体の Accuracy を表 3, 4 に示す．

表 2. 分類機の最適なパラメータ

ロジステック回帰	C	正則化	
	0.01	l2	
ランダムフォレスト	決定木の数	決定木の深さ	特徴量の数
	10000	40	69

表 3. ロジスティック回帰における分類結果

	Precision	Recall	F-score	Accuracy
高専生	0.925	0.852	0.887	0.887
非高専生	0.858	0.928	0.892	
平均	0.892	0.889	0.889	

表 4. ランダムフォレストにおける分類結果

	Precision	Recall	F-score	Accuracy
高専生	0.97	0.946	0.957	0.958
非高専生	0.946	0.969	0.958	
平均	0.958	0.958	0.958	

[2] https://github.com/taku910/mecab

実験より，ロジスティック回帰であってもそれなりに良い性能を得られることが分かった．

また，ランダムフォレストにおいて 95%を超える性能を得られた．ランダムフォレストが使用した特徴量についても調査を行った結果，”高専”・留年”・”寮”といった高専生に馴染みの深い単語や，”物理”・”英語”・”夏休み”などの学生が良く使用すると思われる単語が使用されていた．

4. 結論

本論文で行った手法を用いて，投稿された文章のみを利用してユーザの属性を推定することが十分に可能であることが確かめられた．

今後は，文章以外の特徴量を併せて使用することも検討している．

5. 謝辞

本研究は JSPS 科研費 JP15K16092 の助成を受けたものです．

6. 引用文献

[1] T. Tuygun , A. Ersa, A. Ahmet, C. A. M., K. Umit, G. Bilal and A. Ramazan, ″User Characterization for Online Social Networks,″ Social Network Analysis and Mining, 2016.

[2] C. Wagner, S. Asur , J. Hailpern, “Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter,” proceedings of the 2013 International Conference on Social Computing, SOCIALCOM ’13, 2013.

[3] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society. Series B(Methodlogical) Volume 58, Issue 1, 1996.

[4] D. Cox, “The Regression Analysis of Binary Sequences,” Journal of the Royal Statistical Society. Series B (Methodological) Vol.20, 1958.

[5] L. Breiman, “Random Forests,” Machine Leaning, Volume 45, Issue 1, 2001.

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.