ソーシャルメディアユーザのプロフィール推定手法の提案

米田 康平† 前田 亮‡

†立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1

‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

E-mail: ri0015xe@ed.ritsumei.ac.jp , amaeda@is.ritsumei.ac.jp

あらまし 製品を開発する企業は，マーケティングのためにソーシャルメディアユーザの性別・年代・職業などのプロフィール情報を知りたいという要求がある．製品に関して興味をもっている Twitter ユーザ群の傾向がわかれば，マーケティングに生かすことができる．Twitter ユーザのプロフィール推定を最終目標として，本稿では Twitter ユーザの性別を推定する手法を提案し，評価実験によってその性能を評価した．ツイフィールという外部サービスで「男性」とタグ付けされているユーザは男性，「女性」とタグ付けされているユーザは女性と仮定して教師データを集め，SVM(support vector machine)によって実験を行った結果，73%という精度で男女の判定をすることができた．その後，提案手法を実装した Web アプリケーションを作成し，インターネット上で公開した．

キーワード Twitter, プロフィール推定, SVM

1. はじめに

本研究では，機械学習によって Twitter ユーザの性別・年代・職業などのプロフィール情報を推定することを最終目標に，まずは性別を判定するシステムを作成した．また，提案した男女推定を Web アプリケーションとしてインターネット上で公開し，誰でも Twitter ユーザの男女推定を行えるようにした．

Twitter ユーザのプロフィール推定の利用先は主に 2 つ考えられる．(1)企業におけるマーケティング，(2)社会科学における調査である．

企業が商品を企画する際，どのような性別・年齢の層を主な顧客とするかはマーケティングにおいて重要な要素である．例えば，お酒・車のマーケティングターゲット層は 20 歳～34 歳の男性である．情報収集の方法としては，政府や公共機関などのホームページ，団体の刊行物などを用いることがあるが，Twitter の投稿はユーザの生の声が反映されやすいため，マーケティングにおいて重要視されている．

社会科学の研究分野においても，Twitter のプロフィール推定が役に立つ場合がある．社会科学は，人間の社会の様々な面を科学的に探求する学術分野である．Twitter ではユーザが日常に起きたことを発信していることが多く，社会科学との相性が良い．例えば，アメリカのトランプ大統領に関して興味を示しているのは男性か女性どちらが多いのかを調査したい場合，本研究が役に立つ．Twitter のキーワード検索で「トランプ大統領」と入力すると，トランプ大統領というキーワードをツイートしたユーザのリストが表示される．それぞれのユーザの性別推定をしていけば，トランプ大統領に興味を示しているユーザ群の性別傾向がわかる．あくまで推定であるので，確実な情報であると保証することはできないが，参考値にはなりうる．

本論文は以下のように構成される．2 節では関連研究について紹介する．3 節では提案手法について詳細に説明する．4 節では評価実験を通して提案手法の評価を行う．5 節では提案した男女推定手法にインターネット上からアクセスできるようにした Web アプリケーションを紹介する．6 節で考察を述べ，7 節で本研究をまとめる．

2. 関連研究

池田ら[1]はテキストの中から重要なキーワードを検出することでプロフィール推定を行う手法を提案している．たとえば，コメントに「学校」や「部活」などのキーワードが頻繁にみられるユーザは年代が 10 代である可能性が高くなる．SVM を用いてキーワードの出現傾向を学習し，未知のユーザであってもプロフィールが推定できる．性能評価実験の結果，提案手法の汎用的な推定精度は性別で 88.0%，年代で 68.0%，居住地域で 70.8%であった．

榊ら[2]は，[1]の論文では性別・年代・居住地域のみを推定しているのに対し，榊らははじめて職業を推定する手法を提案した．会社員か否かの二値分類で，適合率 85%，再現率 77%という結果を得ている．

その他にも，Twitter のメンション情報を利用した例[3]や，Twitter の周辺ユーザの情報を使ってプロフィール推定を行った例もある[4].

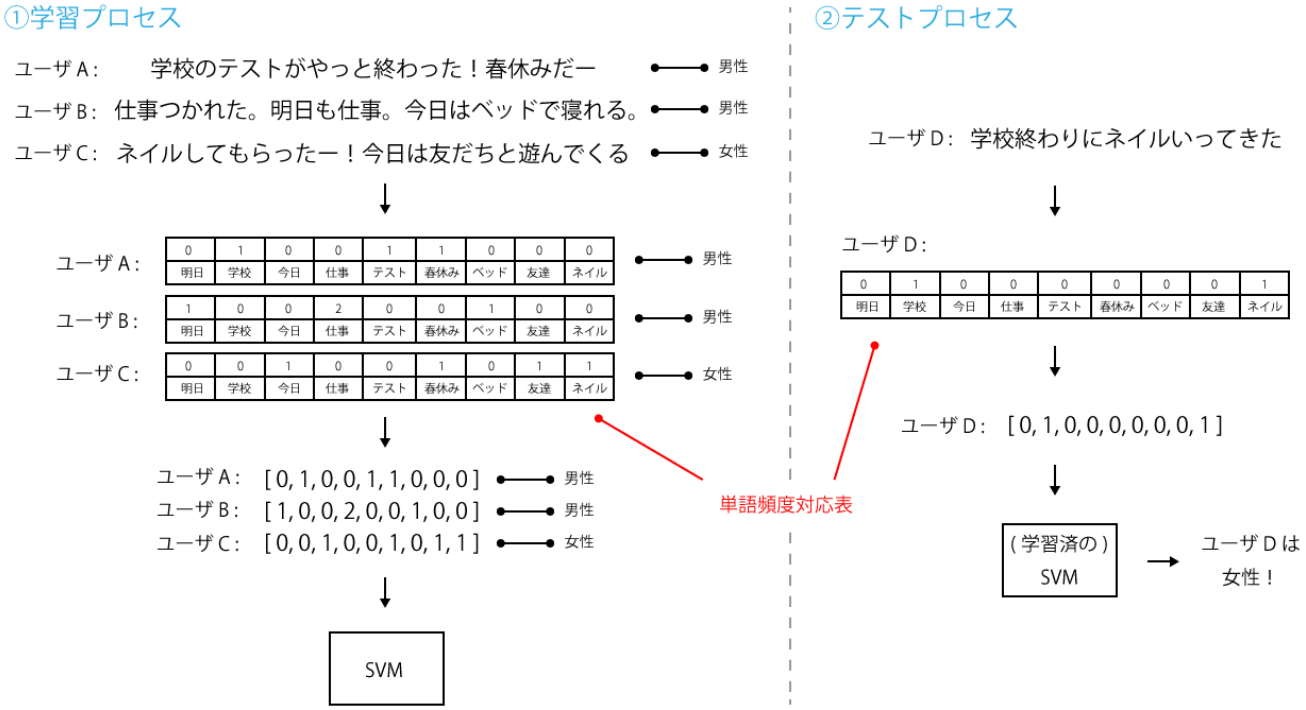

図 1.提案手法の簡単な例

3. 提案手法

提案手法は学習プロセスとテストプロセスに分かれる．学習プロセスでは，まず性別が判明しているユーザの投稿テキストから，単語頻度対応表を作成する．単語頻度対応表とは，文章中に出現した単語とその頻度を表にしたものである．単語頻度対応表の頻度の部分だけを取り出し，特徴ベクトルとする．特徴ベクトルと性別の情報を SVM に学習させる．大量のデータを学習した SVM を使って，テストプロセスでは未知のデータでテストをする．学習プロセスと同様に，ユーザに紐付けられたテキストを特徴ベクトルに変換し，SVM に入力すると，性別が女性であるか男性であるかが判断される．

図 1 に提案手法の簡単な例を示す．①学習プロセスでは，ユーザ A の投稿テキストが「学校のテストがやっと終わった！春休みだ」で性別が男性，ユーザ B の投稿テキストが「仕事つかれた．明日も仕事．今日はベッドで寝れる」で性別が男性，ユーザ C の投稿テキストが「ネイルしてもらったー！今日は友だちと遊んでくる」で性別が女性，といったように 3 つのユーザのデータを SVM に学習させた例を示している．本来であれば，文書はユーザの全ツイートを結合させたものであるのでもっと長くなる上，機械学習するためにユーザの数を十分にそろえなければいけない．簡単に説明するため，ここでは単純な例を示している．まずは文章を単語に分割し，単語頻度対応表を作成する．単語頻度対応表の数値の部分は単語が出現した頻度を示す．それを特徴ベクトル化し SVM に学習させる．その後，②テストプロセスでは，未知のユーザ D を SVM に入力し，性別は男性か女性かを判断する．ユーザ D の投稿テキスト「学校終わりにネイルいってきた」から単語頻度対応表を作成し，特徴ベクトルに変換する．特徴ベクトルを学習済みの SVM に入力すると，性別が判断できる．

提案手法の大まかな流れは以下の通りである．

Ⅰ.Twitter アカウント収集
Ⅱ.データベースの作成
Ⅲ.形態素解析をして単語を取り出す
Ⅳ.特徴ベクトルを作成
Ⅴ.学習とテスト

3.1 Twitter アカウント収集

機械学習をするためには教師データが必要である．男女推定においての教師データとは，既に性別が判明している Twitter アカウントである．Twitter アカウントを 1 つ 1 つ確認し，人手で性別を判断するには限界がある．関連研究[2]においては，Twitter のプロフィール欄に「男性です」や「女性です」というテキストが含まれたユーザを抽出した後，人手で精査している．本研究では多くのデータを収集する目的で以下の方法

をとる．Twitter のプロフィールを拡張できるツイフィール[1]という Web サービスが存在する．Twitter はプロフィールの文章を最大 160 文字までしか入力することができないが，ツイフィールでは最大 10000 文字まで入力することができる．またツイフィールのユーザは自分にタグをつけることができる．例えば，音楽が好きであれば「音楽」，アニメが好きであれば「アニメ」といったタグをつけることができる．ユーザの中には女性タグまたは男性タグをつけている人もいる．自分に女性のタグをつけているユーザは女性，男性のタグをつけているユーザは男性であると仮定し，データを収集する．2017 年 2 月現在，女性タグで検索すると 3473 件，男性タグで検索すると 2078 件のデータが得られた．

3.2 データベースの作成

前節で述べた手法で収集した Twitter アカウントの情報をパラメータにして，TwitterAPI に順次アクセスをし，図 2 のようなデータベースを構成する．１つはユーザのテーブル，もう１つはツイートのテーブルである．ツイートテーブルの userId と，ユーザテーブルの id が対応して紐付いている．このように，ユーザテーブルとツイートテーブルを別の構成でデータベースを作成することで，今後別の実験をする場合にも柔軟に対応できる．例えば最新ツイートを 200 件収集していたところを，100 件のみ収集した場合や，サムネイル画像を用いたアプローチなど，様々なことを試すことができる．本提案手法で実際に使うのはツイートと性別の２つの情報である．必要になった際にはこのデータベースにアクセスし，情報を取得する．

ユーザテーブル

id	screenName	name	description	profileImageUrl	gender
1	@tarou	太郎	こんにちは高校…	https://XXX/AAA.png	0
2	@hanako	花子	私は主婦をして…	https://XXX/BBB.png	1
3	@yamada	山田	暇をしている…	https://XXX/CCC.png	0
4	@nakata	中田	２４歳トレーダ…	https://XXX/DDD.png	0
5	@nanako	奈々子	こんにちは！私…	https://XXX/EEE.png	1

ツイートテーブル

id	userId	screenName	tweet
1	1	@tarou	帰宅！たのしかった。
2	1	@tarou	今日は映画を見に行ってきた。
3	1	@tarou	おもしろい映画ないかなー
4	2	@hanako	旅行に行きたい
5	2	@hanako	仕事つかれた

図 2. Twitter 情報を格納するデータベース

[1] http://twpf.jp/

3.3 形態素解析をして単語を取り出す

前節のデータベースから 1 人のユーザーが投稿したツイートを取り出す．MeCab[2]と NEologd[3][5]を使って文書を形態素解析し，単語を抽出する．MeCab はオープンソースの形態素解析エンジンで，日本語を分割できるツールである．NEologd を使うと，最新の用語にも対応できる．例えば，「中居正広のミになる図書館」というテレビ番組が放送されているが，このキーワードを MeCab で処理をすると「中居/正広/の/ミ/に/なる/図書館」と分割される．NEologd を使うと，「中居正広のミになる図書館」と 1 つのキーワードとして抽出することができる．Twitter 上の文章は，辞書には登録されていない最新のキーワードが含まれている可能性のあることから， NEologd を用いる．

3.4 特徴ベクトルを作成

キーワードの取り出しができたら，キーワードとその頻度の表を作成していく．その後，それぞれのユーザの単語頻度対応表の頻度を特徴ベクトルとして取り出す．

3.5 学習とテスト

前節で述べた手法で取り出した特徴ベクトルを SVM に学習させる．特徴ベクトルと性別情報のセットを次々と SVM に学習させる．学習が完了したら，テストデータでテストをしていく．テストデータの特徴ベクトルを SVM に入力すると，これまでの学習蓄積から，そのテキストを投稿したユーザは男性であるのか女性であるのかを推定する．

4. 評価実験

前節で述べた提案手法の有効性を検証するため評価実験を行う．ツイフィールで男性タグを登録していたアカウント 400 個，女性タグを登録していたアカウント 400 個を収集した．その後，TwitterAPI[4]を用いてユーザごとの最新 200 件のツイート，プロフィール文章を収集した．ツイートすべてとプロフィール文をつなげ，それに加えて 0,1 の性別情報を 2 つのセットにした．これが機械学習に必要なデータである．テキストを形態素解析をして単語を取り出し，単語頻度対応表を作成した．続いて，単語頻度作成表を特徴ベクトルに変換した．特徴ベクトルと性別情報がセットになったものを SVM に学習させていく．男性アカウント 400 個のうち 300 個は学習用で 100 個はテスト用，女性アカウントは 400 個のうち 300 個は学習用で 100 個はテスト用に使った．学習後，テストデータ 200 件を読み込ませたところ，性別が正解していたものは 146 個で，正解率は 73.0%となった．

[2] http://taku910.github.io/mecab/
[3] https://github.com/neologd/mecab-ipadic-neologd
[4] https://dev.twitter.com/docs

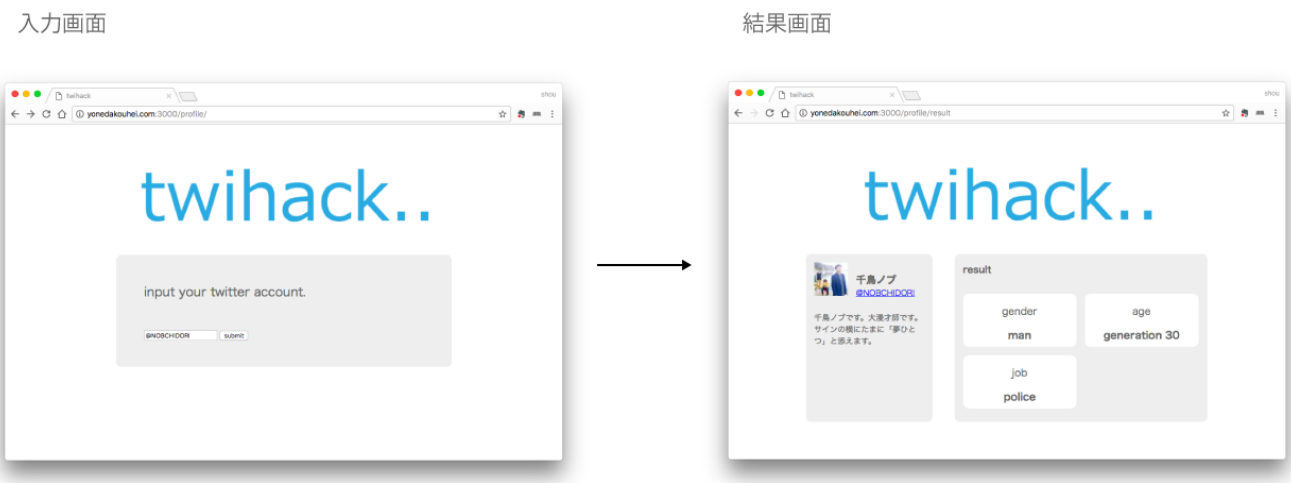

図 3. Web アプリ「twihack」

5. 提案手法の実装システム

本研究で提案した手法をインターネット上でアクセスできるようにした．twihack は，Twitter アカウント名を入力すると，そのユーザの性別・年代・職業を推定する Web アプリケーションである．図 3 に Web アプリケーションが実際に動作している様子を示す．最初の画面で Twitter アカウントを入力し submit ボタンを押すと結果画面に遷移する．結果画面では，左のパネルに調査対象の Twitter アカウント情報，右のパネルに推定結果を表示する．twihack は http://twihack.com からアクセスできる．

6. 考察

精度を上げるために必要なポイントとして(1)データの質を高める，(2)機械学習のチューニングについて説明をする．

精度を上げるため，データの質を高めるというのが非常に重要である．教師データを収集するときにツイフィールという Web サービスを用いて自動で収集した．人の目を通すことなく機械的に処理をしただけなので，本来は収集するべきでなかったものが混ざっている可能性もある．例えば，本当は男性のユーザであるにもかかわらず，女性ユーザであると虚偽の申告をしている場合である．今後は人手によってデータを精査する必要がある．

機械学習のプロセスではチューニングを行うことができる．現在は単語リストをすべて読み込ませているが，出現頻度の非常に多い単語を除去する，もしくは出現頻度の非常に少ない単語を除去するということもできる．機械学習のチューニングによって推定精度が向上する可能性もある．

7. まとめ

本研究では，機械学習を用いて Twitter ユーザの男女判定を行う手法を提案し，73%という精度で見分けることができた．また，提案した男女推定手法に外部からアクセスできるよう，Web アプリケーションを制作して公開した．今後は Web アプリケーションにアンケート機能などをつけて，推定結果が正しかったか間違っていたかを送信できるようにする予定である．多くのの人にアンケートに答えてもらうことができれば，アンケート結果を既存の SVM に追加で学習させ，推定精度を上げることができる．それ以外にもデータの質を高めたり，機械学習のチューニングをすることによって，推定精度の向上を図りたい．

参 考 文 献

[1] 池田和史，服部元，松本一則，小野智弘，東野輝夫“マーケット分析のための Twitter 投稿者プロフィール推定手法”．情報処理学会論文誌 コンシューマ・デバイス＆システム(CDS)，Vol.2，No.1，pp.82-93，2012.

[2] 榊剛史，松尾豊“ソーシャルメディアユーザの職

業推定手法の提案”, 知能と情報 Vo1.26, No.4, p.773-780, 2014

[3] 奥谷貴志, 山名早人 “メンション機能を利用したTwitterユーザプロフィール推定”, DBSJ Japanese Journal, vol. 13-J, No.1, pp.1-6, 2014.

[4] 上里和也, 浅井洋樹, 奥野遼也, 奥野峻弥, 山名早人 “Twitter ユーザを対象とした属性推定の精度向上-周辺ユーザの属性補完を利用して-”, DEIM Forum 2015, D8-5, 2015.

[5] 佐藤敏紀, 橋本泰一, 奥村学 “単語分かち書き用辞書生成システム NEologd の運用-文書分類を例にして-”, 情報処理学会, 研究報告自然言語処理(NL), 2016-NL-229(15), pp1-14, 2016.