クラウドソーシングを用いて作成した教師データによる SNS ユーザーのプロフィール判定

榊茂之　　三浦康秀　　服部圭悟　　坪下幸寛　　大熊智子
富士ゼロックス株式会社　研究技術開発本部　コミュニケーション技術研究所
{sakaki.shigeyuki, yasuhide.miura, keigo.hattori, yukihiro.tsuboshita, ohkuma.tomoko}@fujixerox.co.jp

1. はじめに

ターゲット広告や世論調査を目的として Twitter や Facebook といった Social Networking Service(SNS)を対象としたユーザープロファイリングに関する研究が行われている．性別や年代，職業，居住域といったプロフィール情報を知る最も簡単な方法はプロフィール欄を利用することであるが，プロフィールを記入しないユーザーも多く，また自由記述の場合も多い．そのため投稿からプロフィールを推定する技術が必要である．

プロフィール推定に機械学習を使う問題点として教師データの作成に多大な労力が必要な点が挙げられる．SNS ユーザーのプロフィール判定の先行研究では人手による分類[1]や，ユーザーへのアンケート[2]により教師データを作成しているが，いずれも労力を要している．居住域判定では，投稿に付与される位置情報(ジオタグ）を利用する方法もあるが[3]，ジオタグを利用するユーザーは全体の 1%程度に限られているため[4]，ジオタグを用いる一部のユーザーしか収集することができない点で問題がある．

これらの問題の解決方法として，近年クラウドソーシングの利用が考えられている．クラウドソーシングとは，Web を通じて不特定多数の個人に対して業務を委託する仕組みのことである．現在，Amazon Mechanical Turk[5]や Yahoo!クラウドソーシング[6]などがサービスを提供しており，データ入力やアンケート調査などが委託されている．クラウドソーシングを利用することで，人手でアノテーションしたデータを安価かつ大量に得ることができる．

本研究では，性別，年代，職業，居住域，既婚/未婚，飲酒習慣の有無，喫煙習慣の有無の 7 つの属性について，クラウドソーシングを利用して Twitter ユーザーのアノテーションを行い，ユーザーの過去のツイートとこれらのプロフィール情報が紐付けられた教師データを作成した．以下では，クラウドソーシングによるプロフィール教師データの作成手法，判定器の評価実験，そして精度向上の方法について述べる．

2. 関連研究

Twitter の投稿内容から性別や年代，居住域といったプロフィール情報を推定する研究が行われている．池田らは Twitter において，プロフィール欄に年齢，性別，居住域を記入しているユーザーのツイートを収集し，人手でプロフィールごとに分類して教師データを作成している[1]．そのデータから統計的指標によって抽出した単語を素性として Support Vector Machine(SVM)による判定器を作成している．また，平野らはプロフィール判定において属性間の依存関係を利用する手法を提案している[2]．この研究では，年代が“10 代”ならば職業は“学生”であるはずといったプロフィール属性間の依存関係を用いて精度を向上させることを提案している．

ユーザーの居住域を推定する研究としては，Twitter ユーザーの location 欄を手掛かりにユーザーの過去のツイートと地域を結び付けた教師データを作成し，最尤法による居住域判定を行った Cheng らの研究がある[3]．地域に関連する単語のみを抽出して素性に使用し，抽出による素性のスパースネスをスムージングにより改善することによって居住域判定の精度向上を実現している．

SNS 解析にユーザーのプロフィール情報を利用した研究には，地域情報を利用した宮部らの研究が挙げられる[7]．宮部らは，地震の起きた日時の前後のツイートを被災地域とそれ以外の地域で比較しており，ジオタグとプロフィール情報を利用してユーザーの居住域を決定している．ジオタグは逆ジオコーディングし，プロフィールに対しては文字列マッチングにより地名辞書との一致を見て住所を抽出しユーザーの地域を推定している．

3. クラウドソーシングによる教師データ作成

3.1. クラウドソーシングの作業の設定

教師データの作成には Yahoo!クラウドソーシングを利用した．このサービスは Yahoo! Japan が運営しており，会員に対して，データの収集や入力，新製品やサービスに対するアンケートなどの簡単な仕事を委託し，その対価としてポイントを支払う仕組みとなっている．

作業委託にあたりユーザーごとのツイートデー

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

表 1. プロフィール教師データの内訳 [クラウドソーシングゴールドデータ]

属性	ユーザー数 (不明を除く)	選択肢
性別	3607	男性(1612), 女性(1995), 不明(25)
年代	3607	10 代(1150), 20 代(2277), 30 代(157), 40 代以上(23), 不明(25)
職業	3607	学生(2243), 会社員(1039), 主婦(115), その他(210), 不明(25)
居住域	2239	北海道/東北(196), 関東(1145), 甲信越(77), 東海(170), 関西(412), 中国/四国(97), 九州/沖縄(142), 不明(1363)
既婚/未婚	3400	既婚(137), 未婚(3263), 不明(232)
飲酒習慣の有無	2048	飲む(968), 飲まない(1080), 不明(1584)
喫煙習慣の有無	1487	吸う(107), 吸わない(1380), 不明(2145)

タを以下の手順で準備した．bot ユーザーや 1 ツイートあたりの情報量が少ないヘビーユーザーを除くため，以下の 3 つの条件を設定し，ストリーミングツイートからユーザーを抽出した．

・フレンド，フォロワーが 200 人以下
・携帯または PC からの投稿が 150 件以上
・1 日の平均投稿数が 10 件以下

抽出の結果 3632 人からなるユーザーリストを得た．各々のユーザーについて 100 個前後のツイートを収集し，アノテーションを行うデータとした．

このデータにおいて，各ユーザーのまとまった数のツイートを見て，性別，年代，職業，居住域，既婚/未婚，飲酒習慣の有無，喫煙習慣の有無の 7 つの属性を推定する作業をクラウドソーシングに依頼した．不特定多数の個人に依頼するというクラウドソーシングの特性を考え，データの信頼性を担保するために，各ユーザーのラベルは延べ 10 人のアノテーターによって行われるように設定し，その多数決をとったものをゴールドラベルとして利用することにした．また，人間が読んでも判断できない場合の選択肢として“不明”というラベルを設定した．

3.2. アノテーションの結果

作成したプロフィール教師データの分布を表 1 に示す．10 人のアノテーターの多数決によって決定したそれぞれの属性の個数を括弧付で記載している．“不明”が多数となったユーザーについては教師データから除外した．その結果，性別ではユーザー数は 3632 人から 3607 人に減少し，居住域では 2239 人に減少した．

性別，年代，居住域の 3 つの属性において，ゴールドラベルのアノテーション合意数を横軸に取り，アノテーションされたユーザーの数をプロットしたグラフを図 1 に示す．性別では 10 にピークがあり，アノテーターの意見が高いレベルで一致していることがわかる．このことから，クラウドソーシングのアノテーターの品質が保たれていると判断した．一方，年代や居住域などの属性では 5-6 付近にピークがあるが，これはこうした属性を指し示す言葉がツイートに必ず書かれるものではなく，人間でも判断にも迷うことがあるためと考えられる．

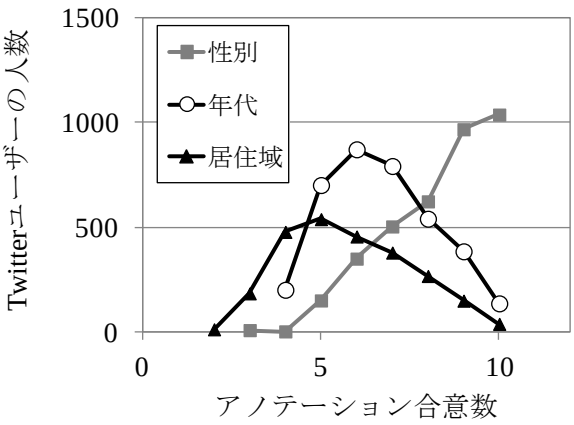

図 1. 各属性のゴールドラベルの
アノテーション合意数の分布
[性別, 年代, 居住域]

3.3. 居住域教師データの自動獲得

クラウドソーシングゴールドデータとの比較を行うために，Cheng らの手法[3]により居住域教師データを作成した．プロフィールの location 欄に居住域を記入しているユーザーを抽出し，都道府県名リストを用いたルールベースの居住域属性のアノテーションを行った．例えば，“熊本”という単語が含まれていれば“九州/沖縄”のラベルを付与した．この方法で 5516 人のユーザーからなる居住域の教師データを得た．そのデータの傾向を表 2 に示す．

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

表 2. プロフィール教師データの内訳 [location から作成した教師データ]

属性	ユーザー数	選択肢
居住域	5516	北海道/東北(354), 関東(3087), 甲信越(169), 東海(438), 関西(909), 中国/四国(205), 九州/沖縄(354)

4. 実験Ⅰ (クラウドソーシングゴールドデータ)

4.1. 実験の設定

ユーザーのプロフィール推定を行う判定器には SVM を用いた．ツイート本文を形態素解析し，Unigram の Bag-of-Words を素性として，SVM(Linear kernel，C=1.0)の学習を行った．形態素解析の実装には kuromoji[8]，SVM の実装には LibSVM[9]を使用した．この SVM の学習と精度の評価を以下の 2 つの設定で行った．

i. クラウドソーシングゴールドデータを入力として，10 回の交差検定で評価
ii. location から作成したデータを入力として，クラウドソーシングゴールドデータによって評価

4.2. 実験結果

実験設定 I - i の結果を表 3 に示す．性別が最も高精度で分類可能であり，Accuracy で 0.805 を達成した．年代，職業，飲酒習慣の有無といった属性においても 0.6 以上の Accuracy を達成した．既婚/未婚，喫煙習慣の有無の精度も非常に高いが，これは教師データが偏っており一方の極性のデータが多いせいであると考えられ，実際の分類精度は低いと考えられる．

居住域判定における設定 I - i，I - ii の結果を表 4 に示す．location から作成した教師データを使用したときの居住域判定の精度は 0.319 となった．location から作成した教師データの方が倍以上多いにも関わらず，クラウドソーシングゴールドデータで学習したときの方が高い精度となった．この原因として無作為なデータを人手でアノテーションしたクラウドソーシングゴールドデータの特徴が挙げられる．今回居住域判定の教師データに使用した 2239 人のユーザーの location 欄を確認したところ 487 人のみが location 欄に都道府県を記入していた．そのため，設定 I - ii のように location 欄に記入しているユーザーだけで学習データを構成すると傾向が異なるデータとなってしまい，高い精度を達成できなかったと考えられる．無作為に収集した今回のクラウドソーシングゴールドデータは実際の適用条件に近いと考えられ，このデータにおいて精度が高いクラウドソーシングゴールドデータの方が優れていると言える．

5. 実験Ⅱ(学習データの併用)

実験 I において最も Accuracy が低かった居住域の地域ごとの精度を参照すると，データ数の多い関東と関西の精度が高いことがわかる．このことから，データを拡充すれば，精度を向上させることができると予想できる．そこで，実験 I で使用した 2 つの教師データ，クラウドソーシングゴールドデータと location から作成した教師データを併用したときの精度を評価した．その結果を表 4 に示す．クラウドソーシングゴールドデータのみで学習したときと比較して Accuracy で 3%向上している．また，地域ごとの精度もクラウドソーシングゴールドデータのみで学習したときと比べて全ての地域で向上している．この結果について，クラウドソーシングゴールドデータのみで学習した場合との paired-T 検定を行い，有意水準 2.226 に対し t=3.05 となり，有意差があることを確認した．location 情報からアノテーションしたデータは単独で使用すると低い精度しか達成できないが，クラウドソーシングゴールドデータと併せて使用することで，さらに高い精度を達成できることが分かった．

6. おわりに

本稿では，ユーザーの過去のツイートとプロフィール属性が紐付いた教師データをクラウドソーシングによって作成することを提案した．この教師データによる判定器は性別，年代，職業，飲酒習慣について 0.6-0.8 の Accuracy を達成した．居住域について，location 情報から作成した学習データとの比較を行ったところ，クラウドソーシングゴールドデータの方が高い精度となることを確認した．また，この location から作成した教師データを併用することで居住域判定の精度を更に向上させることが可能なことを確認した．今後は人手によるアノテーションデータとルールで収集した教師データといった 2 つの異なるデータを効果的に組み合わせる手法を研究していく予定である．

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

表 3. 実験設定 I - i の結果 [クラウドソーシングゴールドデータ]

属性	全体の Accuracy	属性ごとの F 値
性別	0.805	男性(0.783)，女性(0.823)
年代	0.654	10 代(0.528)，20 代(0.744)，30 代(0.157)，40 代以上(0.045)
職業	0.668	学生(0.793)，会社員(0.530)，主婦(0.215)，その他(0.070)
居住域	0.502	北海道/東北(0.113)，関東(0.662)，甲信越(0.017)，東海(0.066)，関西(0.551)，中国/四国(0.186)，九州/沖縄(0.196)
既婚/未婚	0.944	既婚(0.25)，未婚(0.971)
飲酒習慣の有無	0.678	飲む(0.659)，飲まない(0.695)
喫煙習慣の有無	0.898	吸う(0.184)，吸わない(0.946)

表 4. 居住域判定の精度の比較

	全体の Accuracy	地域ごとの F 値						
		北海道/東北	関東	甲信越	東海	関西	中国/四国	九州/沖縄
Ⅰ - ⅰ . Crowd Sourcing	0.502	0.213	0.361	0.048	0.03	0.454	0.073	0.081
Ⅰ - ⅱ . location	0.319	0.113	0.662	0.017	0.066	0.551	0.186	0.196
Ⅱ . Crowd Sourcing + location	0.532	0.134	0.686	0.068	0.101	0.547	0.301	0.245

[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, マルチメディア, 分散, 協調とモバイル（DICMO2011）シンポジウム, pages1308-1315, 京都 (2011)

[2] 平野徹, 牧野俊郎, 松尾義博: Markov Logic を用いたテキストからのユーザ属性推定, 2013 年度人工知能学会年次大会(27 回), 3E3-3, 富山

[3] Zhiyuan Cheng, James Caverlee, Kyumin Lee: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, In *Poc. of the 2010 ACM Conference on Information and Knowledge Management*, pages 759-768, New York, USA(2010)

[4] Bo Han, Paul Cook, Timothy Baldwin: A Stacking-based Approach to Twitter User Geolocation Prediction, In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 7-12, Sofia, Bulgaria(2013)

[5] Amazon Mechanical Turk: https://www.mturk.com/mturk/

[6] Yahoo! クラウドソーシング: http://crowdsourcing.yahoo.co.jp/

[7] 宮部真衣, 荒牧英治, 三浦麻子: 東日本大震災における Twitter の利用傾向の分析, 情報処理学会研究報告, GN-81, No.17 (2011)

[8] kuromoji: http://www.atilika.org

[9] LibSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[商標について]

- Twitter®は，Twitter Incorporated の米国その他諸国における登録商標です.
- Facebook®は，Facebook Incorporated の米国その他諸国における登録商標です.
- Amazon Mechanical Turk™は，Amazon.com Incorporated の米国その他諸国における登録商標です.
- Yahoo!®は，Yahoo! Incorporated の米国その他諸国における登録商標です.
- その他，掲載されている会社名，製品名は，各社の登録商標です.

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.