Vol.2013-IFAT-109 No.4
Vol.2013-NL-210 No.4
2013/1/11

TwitterとBlogの共通ユーザプロフィールを利用したTwitterユーザ属性推定

伊藤 淳[1,a)]　西田 京介[1]　星出 高秀[1]　戸田 浩之[1]　内山 匡[1]

概要：本研究では，Twitter を対象としたユーザ属性推定技術について，ユーザのコンテンツ情報（プロフィール文書とツイート集合）を用いた新たなユーザ属性推定技術を提案する．ユーザ属性を推定する研究は数多く存在するが，本研究は次の 2 点において既存研究と異なる．ひとつめは，Twitter と Blog 両方のアカウントをもつユーザ（共通ユーザ）を発見し，Blog のプロフィールを Twitter の教師ラベルとするラベル伝播学習法によって，人手によるラベリングが不要で高精度な推定器の構築を実現した点である．ふたつめは，ツイート集合に加えてプロフィール文書も利用する 9 種類の手法の検討を通して，ツイート集合のみを用いた推定器よりも高精度な推定器の構築を実現した点である．Twitter ユーザの性別，年齢，職業，興味を推定する評価実験を実施した結果，提案手法は人手ラベリングとツイート集合のみを用いた既存の推定手法よりも高精度であることを示した．

1. 背景

Facebook[*1]や Twitter[*2]に代表されるソーシャルメディアはここ数年で急速な成長を遂げた．Facebook は 2012 年 10 月に月間アクティブユーザ数が 10 億人を超え[*3]，Twitter は 2012 年 3 月にアクティブユーザ数が 1.4 億人を超えた[*4]．また，Facebook は 2012 年 3 月に 1 日あたり 32 億件のコメントや 3 億件の写真投稿があり[*5]，Twitter は 2012 年 3 月に 1 日あたり 3.4 億件のツイート[*6]があった[*4]．

このようにユーザ数や投稿数がかなりの規模に達したことに加え，商品やコンテンツに対する意見や感想が投稿されていることから，ソーシャルメディアをマーケティングに活用することに注目が集まっている．従来主流であったアンケートによるモニタ調査は，モニタ数や質問項目数に応じて費用がかかるため，多くの情報を得ようとするとコストが高くなりがちであった．また，調査開始から集計までに時間がかかるため，リアルタイムに意見や感想を調査することができなかった．一方，ソーシャルメディアを用いたクチコミマーケティングでは，大量の意見や感想をリアルタイムに低コストで調査することができる．こうしたメリットがあることから，国内外問わず多くの企業がソーシャルメディアを用いたクチコミマーケティングに取り組んでおり，様々な分析ツールが提供・販売されている．国内では NTT コミュニケーションズの Buzz Finder[*7]，NTT データのなずきのおと[*8]，NEC BIGLOBE の感゜レポート[*9]，ホットリンクのクチコミ@係長[*10]などがあり，国外では Radian6[*11]，Sysomos[*12]，Forsight[*13]などがある．

ソーシャルメディアを用いたクチコミマーケティングはコスト面やリアルタイム性でメリットがあるものの，クチコミしているユーザがどんな人物であるかわからないデメリットがある．商品やコンテンツに対する意見や感想はユーザの性別，年齢，職業などのデモグラフィック属性や，興味などのサイコグラフィック属性に応じて異なる．そのため，属性の分布傾向を調べたり，属性ごとに意見や感想を集計して分析したりすることがマーケティングで行われている．従来のモニタ調査であれば，質問項目を設けて属性を調査することが可能であったが，ソーシャルメディアにおいては属性が明記されていないことが多く，知ること

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation 1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
[a)] ito.jun@lab.ntt.co.jp

[*1] http://www.facebook.com/
[*2] https://twitter.com/
[*3] http://newsroom.fb.com/Key-Facts
[*4] http://blog.twitter.com/2012/03/twitter-turns-six.html
[*5] http://mashable.com/2012/04/23/facebook-now-has-901-million-users/
[*6] Twitter における投稿記事のこと．140 字以内の制限がある．
[*7] http://www.ntt.com/marketing/bf/
[*8] https://nazuki-oto.com/
[*9] http://kandoreport.jp/
[*10] http://www.hottolink.co.jp/kakaricho
[*11] http://www.radian6.com/
[*12] http://www.sysomos.com/
[*13] http://www.crimsonhexagon.com/

が難しいという課題がある．

この課題を解決するため，本研究では，国内でもユーザ数と投稿数が多く，データがオープンに利用可能なソーシャルメディアである Twitter を対象としたユーザ属性推定技術に取り組んだ．具体的には，ユーザのコンテンツ情報（プロフィール文書とツイート集合）を用いて，性別，年齢，職業，興味を推定する問題を扱った．

ユーザ属性を推定する研究は 2 章に示す通り数多く存在するが，本研究は次の 2 点において既存研究と異なる．ひとつめは，Twitter と Blog 両方のアカウントをもつユーザ（共通ユーザ）を発見し，Blog のプロフィールを Twitter の教師ラベルとするラベル伝播学習法によって，人手によるラベリングが不要で高精度な推定器の構築を実現した点である．なお，共通ユーザを学習データとして利用する際の問題点についても考慮し，評価実験によりその影響と提案手法の有効性を検証した．ふたつめは，ツイート集合に加えてプロフィール文書も利用する 9 種類の手法の検討を通して，ツイート集合のみを用いた推定器よりも高精度な推定器の構築を実現した点である．ツイート集合とプロフィール文書をどのように混合または組み合わせると精度が向上するか，評価実験を通して検証した．

以降，2 章でユーザ属性推定に関する既存研究を紹介し，3 章で Twitter の実データを分析した結果を示しながら，ユーザ属性推定の必要性と教師ラベル作成の難しさについて述べる．4 章で提案手法の詳細について説明し，5 章でその有効性を評価，最後の 6 章でまとめを行う．

2. 関連研究

過去に行われたコンテンツからのユーザ属性推定に関する研究と，ソーシャルグラフからのユーザ属性推定に関する研究について，それぞれ示す．なお，本研究はコンテンツからのユーザ属性推定に位置づけられ，ソーシャルグラフからのユーザ属性推定は今回は扱わず今後の課題とする．

2.1 コンテンツからのユーザ属性推定

Twitter ユーザの属性をプロフィール文書やツイート集合などのコンテンツから推定する研究を示す．池田ら [1] は赤池情報量基準（AIC）を用いて，属性中のクラスごとに特徴語を抽出して素性とし，Support Vector Machine（SVM）で学習・推定する手法を提案している．年齢，性別，地域に関する推定を行い，性別で 88 %という推定精度をあげた．Rao ら [3] は N-gram や SocioLinguistic-feature（社会言語学的な特徴語）を素性とし，SVM で学習・推定する手法を提案している．年齢，性別，地域，政治的志向に関する推定を行い，70〜80 %の推定精度をあげた．さらに予備実験において，フォロワー[*14]数，フレンド[*15]数，フレンド/フォロワー比率，返信率，ツイート数，リツイート[*16]数などが属性ごとに差があるかを検証しており，いずれも素性として利用できるほどの差はなかったことを報告している．Cheng ら [4] は，ツイート集合のみを用いて市レベルでユーザの位置を推定する手法を提案している．ツイート中の各単語と地域との相関をもとにした確率モデルと，ユーザの位置推定を調整するための格子ベースの近隣平滑モデルを提案している．100 件程度のツイートを用いて，51 %のユーザの位置を 100 マイル範囲の誤差で推定することができる．Eisenstein ら [5] は，Cheng らと同様に特定の地域との結びつきの強い単語が存在するというアイデアをもとに，ユーザの位置を推定する手法を提案している．潜在トピックと地域を一緒に推論する生成モデルを用いた手法を提案しているところに違いがある．Burger ら [6] は，単語，文字の N-gram を素性とした教師あり学習による推定器を用いた性別の推定を提案している．ツイート集合，プロフィール文書，スクリーンネーム，名前のすべての素性を用いて 92 %の精度をあげた．比較実験で Amazon Mechanical Turk[*17]を用いて人手で推定したものよりも高い精度であったことを報告している．Pennacchiotti ら [7] は，政治的志向，民族，スターバックスコーヒーへの親近感を推定する手法を示している．プロフィール文書，ツイートの傾向，ツイートに典型的に現れる単語を特徴量とし，ソーシャルグラフを用いてラベル情報を更新させ，Gradient Boosted Decision Trees を用いて推定している．Chu[8] らは，人間と bot[*18]の判別に取り組み，ツイートの仕方，ツイート内容，プロフィールの違いに着目し，線形判別分析（Linear Discriminant Analysis）による判別手法を提案している．Mislove ら [9] は，地理，性別，人種・民族に関して Twitter と現実の人口分布の比較を行い，Twitter のユーザ分布にバイアスが存在することを示している．

2.2 ソーシャルグラフからのユーザ属性推定

ソーシャルグラフ上の近隣ユーザが持つ属性を伝播させることで，ユーザ属性を推定する研究を示す．Mislove ら [10] は，Facebook のソーシャルグラフを用いて，入学年度や学部などのユーザ属性を推定している．同じ属性値をもつノードをシードとして，残りのノードを modularity ベースの評価関数が高くなるように付け加えていくことで推定する手法を提案している．Wen ら [11], [12] は，大規模センシングデータ（メール，インスタントメッセージ，

[*14] ユーザをフォロー（ツイートを購読するために相手を登録）している人
[*15] ユーザがフォローしている人
[*16] 他人のツイートを引用してツイートすること
[*17] https://www.mturk.com/
[*18] プログラムにより自動的に投稿やフォローを行うアカウント

表 1 Twitter プロフィールの構造

項目名	説明
description	自己紹介文を自由記述する項目
location	居住地や所在地を自由記述する項目
statuses_count	現在までの総ツイート数を示す項目
url	外部 URL を自由記述する項目

ソーシャルブックマーク，ファイル共有）からユーザの興味を推定している．コミュニケーション回数によって重み付けがされた伝播モデルを提案している．He ら [13] は，ベイジアンネットワークを用いて homogeneous societies と呼ばれる現実の関係を反映した小規模なグループを作成して，Blog ユーザの属性を推定する手法を提案している．Zheleva ら [14] は，友人情報とグループ情報を用いてユーザ属性が推定できるかを実験している．友人情報よりもグループ情報を用いたほうが推定精度が高いことを報告している．Zamal ら [15] は，フレンド関係にあるユーザ群から得られた特徴量の平均値を算出し，それをユーザ本人の特徴量とどのように組み合わせると推定精度が向上するか，様々な手法を検討している．Lindamood ら [16] は，プライバシー保護の観点から，周囲の公開情報を用いてユーザ属性が推定されないように，ユーザ属性や友人関係を隠蔽することを検討している．

3. Twitter プロフィールの分析

Twitter では，自由記述形式で自己紹介文（プロフィール文書）を記述することができる．プロフィール文書中にユーザ属性が明記されていれば，そもそもユーザ属性を推定する必要はなくなる．したがって，まずはどれくらいのユーザがプロフィール文書を記述しており，その中にユーザ属性を明記しているのかについて，Twitter の実データを用いて分析を行った．

3.1 Twitter プロフィールの構造

Twitter プロフィールの構造を表 1 に示す．表 1 は Twitter プロフィールすべての項目ではなく，本研究で利用しているものだけを掲載して説明したものである．

3.2 分析方法と結果

分析に用いたデータの詳細を表 2 に示す．表 2 に示したユーザにおいて，プロフィール文書の記述があるか，性別，年齢，職業，地域の属性ごとにその属性を示すような単語が含まれているかについて調査した．性別では，男性，女子，♂など性別を示す単語が，年齢では 10 歳，21 才，アラサーなど年齢を示す単語が，職業では主婦，会社員，学生など職業を示す単語が description 項目中に含まれているかを正規表現によるマッチングで調査した．地域は，location 項目中に都道府県名が含まれているかを正規表現によるマッチングで調査した．

表 2 Twitter プロフィールの分析に用いたデータ

対象期間	2012/3/1〜2012/3/31
API レベル	gardenhose（10 %サンプリング）
ツイート数	113,814,861
ユニークユーザ数	4,638,441

表 3 プロフィール文書と属性の記述率

	記述あり	記述率
プロフィール	3,827,885	82.53
性別	353,558	7.62
年齢	154,900	3.34
職業	631,626	13.62
地域	1,158,570	24.98

表 3 に分析結果を示す．表 2 に示したユーザのうち，何らかのプロフィール文書を記述しているユーザは全体の約 83 %と多いが，地域以外の属性の記述率は 14 %未満と低かった．さらに，これら記述率は真に属性を記述している値ではなく，抽出ノイズを含んだ値である．例えば，子供やペットなど本人以外の性別や年齢を記述している場合や，男性声優など対象としている属性（性別）とは異なる属性（興味）を示す単語が含まれる場合が存在した．地域は location 項目という専用の入力欄が存在するため，記述率は約 25 %と比較的高かったが，異なる都道府県が複数記述されているなど，抽出ノイズは含まれていた．

分析結果から，多くのユーザがプロフィール文書を記述しているが，ユーザ属性の記述率は低いことがわかった．このため，ユーザ属性を抽出する方法では多くのユーザにおいて不十分であり，推定が必要であることがわかった．また，プロフィール文書が自由記述形式であるため，正規表現によるマッチングでは抽出ノイズが含まれやすく，人手による目視確認が不可欠であることもわかった．

それから，表 2 に示したユーザの statuses_count 項目を集計し，平均的にどのくらいのツイート数が推定の情報源として見込めるのかについても調査した．その結果，statuses_count の平均値は 68.1，中央値は 553 であった．ツイートは 140 字以内の制限があるため Blog 記事よりも文書長が短く，statuses_count の平均値も 68.1 と少ないため，推定対象のユーザによってはツイート集合だけでは推定に必要な情報が不足し，推定精度が下がる可能性があることがわかった．

4. 提案手法

人手で教師ラベルを作成するのは，難しいうえに手間のかかる作業である．プロフィール文書から正規表現によって教師ラベルを作成する方法を用いても，3.2 章で示した通り抽出ノイズが含まれるため，最終的には人手による目視確認が必要となる．

表 4 プロフィールに含まれる外部メディア

ドメイン	ユーザ数
ameblo.jp	159,768
blog*.fc2.com	20,407
facebook.com	20,237
blog.livedoor.jp	16,500
mixi.jp	16,289
d.hatena.ne.jp	12,348
jugem.jp	11,991
blogspot.com	11,752
exblog.jp	10,706
tumblr.com	10,647

そこで，本研究では Twitter と Blog 両方のアカウントを持つユーザ（共通ユーザ）を発見し，そのユーザが Blog に記述しているプロフィールを Twitter の教師ラベルとする，ラベル伝播学習法を提案する．また，ツイート集合だけでは推定に必要な情報源が不足し，推定精度が下がる可能性がある課題を解決するため，プロフィール文書も利用した推定器による，より高精度な推定器の構築手法も提案する．

4.1 ラベル伝播学習法

Twitter では，表 1 に示すように，url 項目に外部 URL を記述することができる．表 2 に示したユーザにおいて，ドメインごとに url 項目を集計し上位 10 件を抽出したところ，表 4 の結果が得られた．

Blog はプロフィールを属性ごとに記述するようになっている場合が多く，Twitter のような自由記述形式のプロフィール文書と違って，ルールベースでユーザ属性を抽出することが可能である．既存研究においても，Blog プロフィールを教師ラベルとして信頼し，Blog のユーザ属性推定を行なって高精度を上げたものがある [2]．また，人手による目視確認での教師ラベル作成は，コスト面から数千件程度に留まることが多いが，共通ユーザを利用した教師ラベル作成は，表 4 によると数万から数十万件程度得られる可能性がある．一般に，学習データ量を増やすほど推定精度は高まるため，高精度な推定器の構築が期待できる．

以上の理由から，Twitter と Blog 両方のアカウントを持つユーザ（共通ユーザ）を発見し，そのユーザが Blog に記述しているプロフィールを Twitter の教師ラベルとする，ラベル伝播学習法を提案する．共通ユーザは，Twitter プロフィールの url 項目に記述された URL を用いて Blog アカウントと紐付けることによって発見する．Blog プロフィールを教師ラベルとして抽出し，プロフィール文書やツイート集合など，Twitter ドメインで得られる単語出現頻度情報を学習データとして用いて推定器を構築する．特徴量選択手法や推定器の学習方法は任意のものが使用可能であり，本研究では特徴量選択に AIC を，推定器の学習に SVM を用いた．

4.2 プロフィール文書も利用した推定器構築手法

ツイート集合だけでは推定に必要な情報源が不足し，推定精度が下がる可能性がある．そこで，プロフィール文書も利用した推定器による，より高精度な推定器の構築手法を提案する．プロフィール文書はユーザあたり 1 文書しか存在しないが，ユーザ属性が直接記載されることもある質の高い情報源であり，精度の向上が期待できる．ツイート集合を用いて推定器を構築する既存手法は数多く存在するものの，プロフィール文書をどのように混合，または組み合わせると良いのかは自明ではないため，本研究にて様々な手法を提案し，5 章にてその効果を検証する．本研究では，以下の 9 種類の手法を実装して評価した．

MIX　推定器をひとつ構築する．その際，プロフィール文書中の単語とツイート集合中の単語を同じものとしてカウントする．プロフィール文書を 1 ツイートとみなすことに等しい．

JOIN　推定器をひとつ構築する．その際，プロフィール文書中の単語とツイート集合中の単語を別のものとしてカウントする．したがって，特徴量の次元数はツイート集合のみを用いた推定器よりもプロフィール文書の特徴量の分だけ大きくなる．

AVG　プロフィール文書とツイート集合でそれぞれ推定器を構築する．両推定器の出力値の平均値を採用する．

MAX　プロフィール文書とツイート集合でそれぞれ推定器を構築する．両推定器のクラスごとの出力値の中で最大値を出した推定器の出力を採用する．

VAR　プロフィール文書とツイート集合でそれぞれ推定器を構築する．両推定器のクラスごとの出力値に関して分散値を算出し，分散値が大きい推定器の出力を採用する．

DEF　プロフィール文書とツイート集合でそれぞれ推定器を構築する．両推定器のクラスごとの出力値に関して，最大値となるクラスと次点となるクラスに対する惜敗率を算出し，惜敗率が小さい推定器の出力を採用する．

KIND　プロフィール文書とツイート集合でそれぞれ推定器を構築する．両推定結果を (1) 式によって信頼度に応じて重み付けて統合する．推定器構築に用いた特徴量全体のうち，ユーザを推定するためにどのくらいの特徴量を用いたかによって信頼度を定める．信頼度は (4)，(5) 式の通り，使用された特徴量の種類数によって定まる．

$$P(u) = R_p(u)P_p(u_p) + R_t(u)P_t(u_t) \quad (1)$$

$$R_p(u) = \frac{I_t(u_t)}{I_p(u_p) + I_t(u_t)} \quad (2)$$

$$R_t(u) = \frac{I_p(u_p)}{I_p(u_p) + I_t(u_t)} \quad (3)$$

$$I_p(u_p) = -\log\left(\frac{kind(u_p) + \alpha}{|F_p|}\right) \quad (4)$$

$$I_t(u_t) = -\log\left(\frac{kind(u_t)+\alpha}{|F_t|}\right) \quad (5)$$

なお，u はユーザを示し，ユーザのプロフィール文書 u_p およびツイート集合 u_t で構成される．P_p はプロフィール文書からの推定確率，P_t はツイート文書集合からの推定確率であり，それぞれ信頼度 R_p，R_t によって重み付けて統合され，最終的な推定確率 P を得る．R_p と R_t は推定器を構築する際に使用した特徴量のうち，どれだけを利用したかに基づく選択情報量 I_p，I_t によって定められる．I_p，I_t は，文書中に含まれていた特徴量の種類数をカウントする関数 $kind$ によって得られた値および全体の特徴量の種類数 $|F|$ によって定まる．α は対数値が 0 とならないために加える定数であり，今回は 1 を用いている．

AIC　プロフィール文書とツイート集合でそれぞれ推定器を構築する．**KIND** における (4)，(5) 式を，(6)，(7) 式のように使用された特徴量が持つ AIC の値の総和で置き換えたものである．

$$I_p(u_p) = -\log\left(\frac{\sum_{s\in set(u_p)} aic(s)+\alpha}{\sum_{f\in F_p} aic(f)}\right) \quad (6)$$

$$I_t(u_t) = -\log\left(\frac{\sum_{s\in set(u_t)} aic(s)+\alpha}{\sum_{f\in F_t} aic(f)}\right) \quad (7)$$

なお，set は文書中に含まれる特徴量の集合を返す関数であり，aic は特徴量選択時に算出された，特徴量 f の AIC の値を返す関数である．

RANK　プロフィール文書とツイート集合でそれぞれ推定器を構築する．**KIND** における (4)，(5) 式を，(8)，(9) 式のように使用された特徴量が持つランク値の総和で置き換えたものである．

$$I_p(u_p) = -\log\left(\frac{\sum_{s\in set(u_p)} rank(s)+\alpha}{\sum_{f\in F_p} rank(f)}\right) \quad (8)$$

$$I_t(u_t) = -\log\left(\frac{\sum_{s\in set(u_t)} rank(s)+\alpha}{\sum_{f\in F_t} rank(f)}\right) \quad (9)$$

$$rank(f) = \frac{|F|}{index(f)} \quad (10)$$

なお，$rank$ は特徴量 f のランク値を返す関数であり，ランク値は特徴量を AIC の値の降順で整列したときの順位を返す関数 $index$ と特徴量の総数 $|F|$ によって (10) 式の通りに定められる．

5. 評価実験

提案手法の有効性を評価するため，評価実験を行った．評価実験に用いたデータの詳細を表 5 に示す．特徴量選択は池田ら [1] と同様に AIC を，推定器の構築には SVM を用いた．SVM は LIBLINEAR[*19]を使用し，ソルバは L2-regularized logistic regression (primal) を利用した．これは，推定結果を統合するために確率値を出力として得るためである．5.1 章で推定器の最適な特徴量数を求めた後，5.2，5.3，5.4 章で提案手法の有効性を評価した．なお，SVM のパラメータは予備実験によって最適な値を求めて設定した．

[*19] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 5　評価実験データ

		共通ユーザ	Blog ユーザ
ユーザ数	性別[*20]	71,129	49,739
	年齢[*21]	36,234	17,689
	職業[*22]	41,920	37,427
	興味[*23]	20,846	7,417
	全体	86,183	65,873
Blog 記事数		796,583	626,903
ツイート数		15,124,094	-

5.1 特徴量数の設定

推定器を構築する際に利用する特徴量は特徴量選択によって制限されるが，特徴量数をいくつにすれば良いかは自明ではない．そこで，プロフィール文書を用いて構築するプロフィール文書推定器と，ツイート集合を用いて構築するツイート集合推定器のそれぞれについて，最適な特徴量数 F を 10-Fold Cross Validation によって求めた．

表 6，7 に結果を示す．これを見ると，特徴量数を増やすほど精度が向上するわけではなく，一定の特徴量数で精度のピークが存在することがわかる．また，ツイート集合推定器の方がより多くの特徴量を必要することや，属性によって最適な特徴量数は異なることがわかる．

以上の結果を受け，その後の実験を簡単にするため，属性ごとに最適な特徴量数を定めず，プロフィール文書推定器は 5,000，ツイート集合推定器は 30,000 の値に統一した．

5.2 人手によるラベリング手法との比較

ラベル伝播学習法 (DIRECT) と，人手によるラベリングを用いた既存のユーザ属性推定手法との精度を比較するための評価実験を行った．正規表現を用いてプロフィール文書中に属性を示す単語があったものを学習データとする手法 (REGEXP) と，正規表現によるマッチングの後，人手によってプロフィール文書を目視確認し，正しいと判断されたもののみを学習データとする手法 (HUMAN) を比較対象とした．ここで，HUMAN は池田ら [1] の手法に相当する．D1000 は DIRECT における学習データ量を，HUMAN，REGEXP と揃えたものである．性別，年齢の 2 属性を対象とし，属性を構成するクラスごとに 1,000 件の学習デー

[*20] 男性，女性（2 クラス）
[*21] 10 代，20 代，30 代，40 代以上（4 クラス）
[*22] 主婦，会社員，中高生など（8 クラス）
[*23] 音楽，スポーツ，ゲームなど（20 クラス）

表 6 プロフィール文書推定器の F 変化時精度

F	性別	年齢	職業	興味
100	73.31	55.27	48.09	49.32
1,000	79.67	62.27	54.00	55.68
2,500	81.23	63.83	55.53	**56.49**
5,000	**81.80**	**64.17**	56.31	55.75
7,500	81.63	63.37	**56.81**	55.00
10,000	81.10	62.22	56.44	54.53
15,000	80.54	60.77	54.71	52.58

表 7 ツイート集合推定器の F 変化時精度

F	性別	年齢	職業	興味
100	87.69	66.18	55.47	49.00
1,000	92.54	72.33	60.73	54.01
5,000	93.93	74.51	61.63	56.81
10,000	94.19	75.03	61.50	**57.06**
15,000	94.36	75.39	61.67	56.89
30,000	**94.50**	76.27	62.57	55.58
50,000	94.49	**76.79**	63.70	54.35
100,000	94.33	75.15	**63.92**	54.15

表 8 ラベリング手法による精度変化

	REGEXP	HUMAN	D1000	DIRECT
性別	72.59	82.32	89.39	**94.50**
年齢	59.49	61.86	67.72	**76.28**

タを用意した．REGEXP，HUMAN は，DIRECT のデータをテストデータとして評価した．D1000，DIRECT は 5-Fold Cross Validation によって評価した．ただし，D1000 は 5-Fold された際の学習データから各クラス 1000 件ランダムに選択したものを学習データとした．

実験結果を表 8 に示す．REGEXP，HUMAN，D1000，DIRECT の順に精度が向上していくことがわかった．HUMAN が D1000 よりも精度が悪かったのは，プロフィール文書にユーザ属性を記述するような少数派のユーザのみを教師として学習しているためだと考える．DIRECT が最も精度が良かったのは，他の手法と比較して学習データの量が性別で 35 倍，年齢で 9 倍と多いためだと考える．人手による目視確認は時間と労力から多くの学習データを用意することは大変であるが，ラベル伝播学習法では大量の学習データを自動的に収集することが可能である．それにより，高精度な推定器が構築できることを示した．

5.3 Blog ラベル利用の妥当性検証

ラベル伝播学習法では，共通ユーザが Blog で記述したプロフィールを教師ラベルとして Twitter のデータで学習を行う．そのため，Blog のプロフィールで嘘を書いていたり，Blog と Twitter で投稿内容の書き分けを行なっていたりする場合，誤った教師ラベルによって学習を行なってしまう可能性がある．書き分けの例としては，Blog で旅行を趣味としてあげており，実際に Blog では旅行記を投稿しているが，Twitter では日常のことしか記述しておらず，旅行の内容がほとんど記述されていない場合があげられる．

そこで，共通ユーザのツイート集合と Blog 記事の投稿内容に食い違いがないユーザのみを学習データとして採用するフィルタリングを行った．フィルタリングは，大きくわけて 2 つの方法を実験した．ひとつめは，共通ユーザ以外の Blog ユーザをランダムに収集して構築した Blog 文書推定器を用いて，共通ユーザの Blog 記事とツイート集合を推定し，推定結果と共通ユーザのプロフィールが合致するもののみを採用するという方法である．共通ユーザのプロフィール，Blog 記事の推定結果，ツイート集合の推定結果すべてが一致する場合 (BOTH)，ツイート集合の推定結果のみ合致しない場合 (BLOG)，Blog 記事の推定結果のみが合致しない場合 (TWIT) の，これら 3 パターンを評価した．しかし，Blog 文書推定器を用いるこれらの方法では，Blog 文書推定器そのものの精度が問題になってしまう．そこで，ふたつめは，ツイート集合の投稿内容と Blog 記事の投稿内容をそれぞれ bag-of-words の単語ベクトルで表現し，それらのコサイン類似度が 0.8 以上のもののみを採用するという方法 (COS) を評価した．また，これらフィルタリング方法を全く用いず，Blog プロフィールを信頼してそのまま伝播する方法 (DIRECT) と，Blog 文書推定器によって共通ユーザのツイート集合の推定を行う転移学習手法 (TRANS) も合わせて評価した．実験はすべて 5-Fold Cross Validation によって評価した．あらかじめ未フィルタリングのデータを 5 分割し，学習データをフィルタリングした上で学習して，テストデータで評価するという操作を 5 回繰り返している．

実験結果を表 9 に示す．括弧なしの値は Accuracy を示し，括弧付きの値はフィルタリングされた後のデータ数を示している．興味以外の属性ではフィルタリングを行わない DIRECT が最も精度が良く，興味では BLOG と COS の精度が良かった．ただし，DIRECT との差は 0.03 %であり，有意差は無かった．学習とテストでドメインが異なるため，TRANS は良い精度が得られなかった．

この結果から，Blog のプロフィールで嘘をついたり，書き分けを行ったりするユーザの影響は小さく，Blog のプロフィールをそのまま信頼することで，様々な属性に対して安定して高精度な推定ができることがわかった．

5.4 プロフィール文書の利用法ごとの比較

プロフィール文書をどのように用いると推定精度をより向上させることができるか，4.2 章で示した様々な手法について実験を行った．なお，比較対象としてプロフィール文書のみを用いて構築した推定器 (PROF) と，ツイート集合のみを用いて構築した推定器 (TWEET) の精度も合わせて掲載した．実験はすべて 5-Fold Cross Validation によっ

表 9 学習データのフィルタリングによる精度変化

	DIRECT	BOTH	BLOG	TWIT	COS	TRANS
性別	**94.37**	90.58	93.43	91.12	93.24	85.66
	(71,129)	(55,728)	(62,291)	(60,929)	(29,873)	
年齢	**75.82**	68.01	71.60	68.76	70.92	66.14
	(36,234)	(17,661)	(23,569)	(23,964)	(14,751)	
職業	**62.29**	50.66	55.16	52.35	56.69	49.82
	(41,920)	(14,382)	(20,489)	(20,883)	(16,912)	
興味	55.35	49.82	**55.38**	50.96	**55.38**	42.32
	(22,393)	(7,960)	(12,851)	(10,457)	(9,222)	

て評価した．

実験結果を表 10，および，図 1，2，3，4 に示す．表 10 では，手法，属性ごとの Accuracy と，手法の良さを示すために手法ごとの平均順位を掲載した．また，図 1，2，3，4 では，横軸に Coverage，縦軸に Accuracy を取った Accuracy/Coverage Curve を掲載した．Accuracy/Coverage Curve は，推定されたクラスの推定確率が高いものから順にデータを整列したとき，上位 N %（Coverage）のデータにおける正解率（Accuracy）を描いたものである．

表 10 を見ると，JOIN と AIC が平均順位で最も良い値を示している．JOIN は年齢と興味において最も良い値であり，AIC は最も良い精度をあげた属性はないものの，全属性で安定した精度を示した．両手法とも，すべての属性において TWEET と比較して有意水準 5 %における McNemar 検定で有意差が認められた．したがって，プロフィール文書を利用する際には JOIN または AIC の手法が良いと考える．

MIX と JOIN を比較すると，JOIN の方が全属性で良い精度をあげた．このことから，同じ単語であってもプロフィール文書とツイート集合ごとに別物として扱った方が良いということがわかる．

AVG が職業で最も良い精度となったのは，図 3 を見るとわかる通り，PROF の Accuracy が TWEET を低 Coverage 領域で上回っているためである．低 Coverage 領域において，PROF と TWEET の Accuracy に逆転が起きるような属性では，AVG が有効な可能性がある．

6. まとめ

本研究では，Twitter を対象としたユーザ属性推定技術について取り組んだ．具体的には，ユーザのコンテンツ情報（プロフィール文書とツイート集合）を用いて，性別，年齢，職業，興味を推定する問題を扱った．

ユーザ属性を推定する既存研究は数多く存在するが，本研究は次の 2 点において既存研究にはない知見を示した．ひとつめは，Twitter と Blog 両方のアカウントを持つユーザ（共通ユーザ）を発見し，Blog のプロフィールを Twitter の教師ラベルとするラベル伝播学習法によって，人手によるラベリングが不要で高精度な推定器の構築を実現した点である．ラベル伝播学習法によって構築した推定器は，人

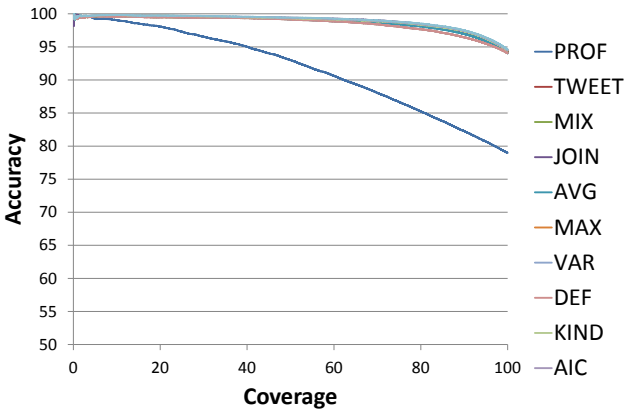

図 1 性別の Accuracy/Coverage Curve

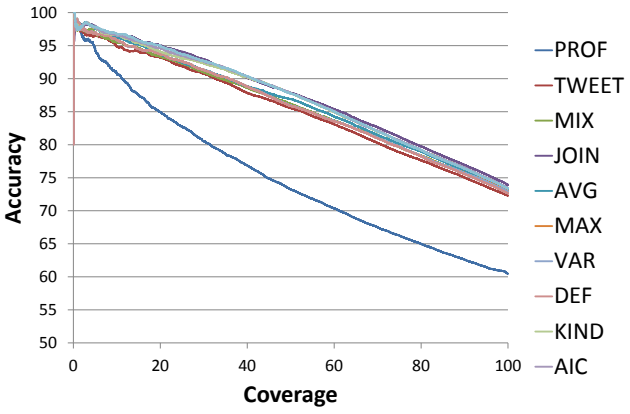

図 2 年齢の Accuracy/Coverage Curve

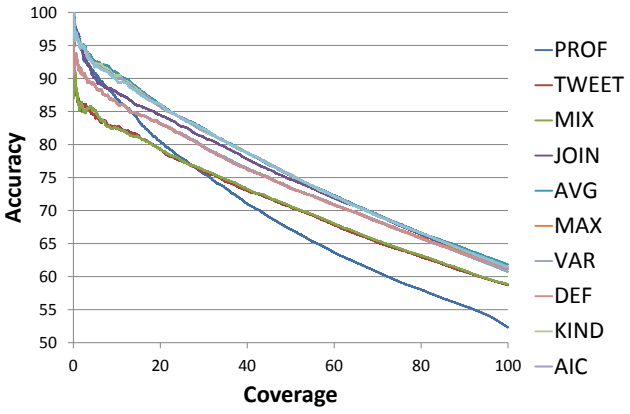

図 3 職業の Accuracy/Coverage Curve

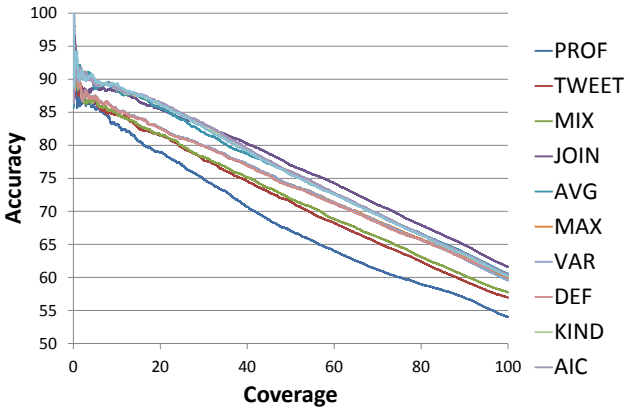

図 4 興味の Accuracy/Coverage Curve

表 10 プロフィール文書の利用法ごとの精度変化

	PROF	TWEET	MIX	JOIN	AVG	MAX	VAR	DEF	KIND	AIC	RANK
性別	78.98	94.20	94.20	94.46	94.03	94.03	94.03	94.03	94.46	94.53	**94.60**
年齢	60.43	72.26	72.90	**73.91**	73.20	72.95	72.89	72.63	73.43	73.45	73.36
職業	52.29	58.71	58.84	61.30	**61.81**	61.13	60.74	61.21	61.49	61.45	61.46
興味	54.00	56.92	57.73	**61.56**	60.51	59.88	59.55	60.23	60.29	60.38	60.18
平均順位	11	9	7.5	**2.75**	4.125	7.125	8.125	7.125	3	**2.75**	3.5

手によるラベリングを用いた従来手法よりも高精度であった．また，ラベル伝播時に書き分けなどの影響を考慮せず，そのまま伝播させることで，様々な属性に対して安定して高精度な推定ができることがわかった．ふたつめは，ツイート集合に加えてプロフィール文書も利用する 9 種類の手法の検討を通して，ツイート集合のみを用いた推定器よりも高精度な推定器の構築を実現した点である．ツイート集合とプロフィール文書をどのように組み合わせて推定器を構築すべきか評価実験を通して検証した結果，本研究で提案した JOIN と AIC が有効であることがわかった．

本研究では，Blog を実例としてラベル伝播学習法を提案したが，ユーザ属性を得ることができれば Facebook など Blog 以外のメディアについても適用可能である．そこで，Blog と Blog 以外のメディアで精度がどのように異なるか今後調査したい．また，本研究ではコンテンツ情報のみを用いてユーザ属性を推定したが，本人のツイートが少ない場合でもソーシャルグラフを利用して周囲のユーザの情報からユーザ属性が推定できるような手法に取り組むことも今後の課題である．

参考文献

[1] 池田和史, 服部元, 松本一則, 小野智弘, and 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] 大倉務, 清水伸幸 and 中川裕志. スケーラブルで汎用的なブログ著者属性推定手法. 情報処理学会研究報告, 自然言語処理研究会報告, vol. 2007, no. 94, pp. 1-5, 2007.

[3] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying Latent User Attributes in Twitter. In SMUC, pp. 37-44, 2010.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In CIKM, pp. 759-768, 2010.

[5] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. In EMNLP, pp. 1277-1287, 2010.

[6] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating Gender on Twitter. In EMNLP, pp. 1301-1309, 2011.

[7] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In KDD, pp. 430-438, 2011.

[8] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In ACSAC, pp. 21-30, 2010.

[9] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the Demographics of Twitter Users. In ICWSM, 2011.

[10] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In WSDM, pp. 251-260, 2010.

[11] Zhen Wen and Ching-Yung Lin. On the Quality of Inferring Interests From Social Neighbors. In KDD, pp. 373-382, 2010.

[12] Zhen Wen and Ching-Yung Lin. Improving User Interest Inference from Social Neighbors. In CIKM, pp. 1001-1006, 2011.

[13] Jianming He, Wesley W. Chu, and Zhenyu Victor Liu. Inferring Privacy Information From Social Networks. In ISI, pp. 154-165, 2006.

[14] Elena Zheleva and Lise Getoor. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In WWW, pp. 531-540, 2009.

[15] Faiyaz Al Zamal, Wendy Liu and Derek Ruths. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In ICWSM, 2012.

[16] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. Inferring Private Information Using Social Network Data. In WWW, pp. 1145-1146, 2009.