TwitterとBlogの共通ユーザおよび会話ユーザの同類性に着目したTwitterユーザ属性推定

伊藤　淳† 西田 京介† 星出 高秀† 戸田 浩之† 内山 匡†

† 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
〒 239–0847 神奈川県横須賀市光の丘 1–1
E-mail: †{ito.jun,nishida.kyosuke,hoshide.takahide,toda.hiroyuki,uchiyama.tadasu}@lab.ntt.co.jp

あらまし 本研究では，Twitter のユーザ属性をコンテンツ（プロフィール文書とツイート集合）とソーシャルグラフ（会話関係）を用いて推定する新たな手法を提案する．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，学習データを大量かつ自動的に収集し高精度な推定器を実現した．また，推定対象ユーザのプロフィール文書や，会話ユーザの同類性に着目することによって，推定精度を向上させた．性別，年齢，職業，興味を推定する評価実験の結果，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手法よりも高精度であることを示した．

キーワード ユーザ属性推定，Twitter，Blog，ソーシャルメディア，同類性

Estimation of Twitter User Attributes by Learning from Users who have both Twitter and Blog Accounts and Utilizing User Homophily

Jun ITO†, Kyosuke NISHIDA†, Takahide HOSHIDE†, Hiroyuki TODA†, and Tadasu UCHIYAMA†

† NTT Service Evolution Laboratories, NTT Corporation
1–1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239–0847 Japan
E-mail: †{ito.jun,nishida.kyosuke,hoshide.takahide,toda.hiroyuki,uchiyama.tadasu}@lab.ntt.co.jp

Abstract We propose a new method for estimating the profile attributes of a Twitter user from the contents (profile document and tweets) generated by the user and the user's social neighbors, i.e. those whom the user conversed with (mentioned). The method first finds the users who have both Blog and Twitter accounts. It then uses the attributes that a user specified in his/her Blog as true labels of the training data in Twitter about the user. It also utilizes the profile documents of the user and the user's social neighbors. Experiments conducted on estimating four attributes (gender, age, occupation, and interests) show that our method achieves higher accuracy than the conventional methods that use manually-labeled tweets.

Key words Profile Estimation, Twitter, Blog, Social Media, Homophily

1. 背　景

Facebook(注1)やTwitter(注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(注3)，Twitter は 2012 年 12 月に月間アクティブユーザ数が 2 億人を超えた(注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(注5)，Twitter は 2012 年 3 月に 1 日あたり 3.4 億件のツイート（投稿文書）があった(注6)．

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

(注1)：http://www.facebook.com/
(注2)：https://twitter.com/
(注3)：http://newsroom.fb.com/Key-Facts
(注4)：https://twitter.com/twitter/status/281051652235087872
(注5)：http://mashable.com/2012/04/23/facebook-now-has-901-million-users/
(注6)：http://blog.twitter.com/2012/03/twitter-turns-six.html

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，NTT データのなずきのおと，NEC BIGLOBE の感゜レポート，ホットリンクのクチコミ@係長などがあり，国外では Radian6，Sysomos，Forsight 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，属性を知ること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（プロフィール文書とツイート集合）とソーシャルグラフ（会話関係）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.** に示す通り数多く存在するが，本研究は次の 3 点において既存研究と異なる．1 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．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．2 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ことで，推定精度が高められることを示した点である．ツイート集合とプロフィール文書の組み合わせ方について網羅的な検証を行い，有効な手法を提案した．3 つめは，推定対象ユーザと会話ユーザの情報量を調節することで，推定精度が高められることを示した点である．ソーシャルグラフでは，類は友を呼ぶ関係が成り立つことが知られている [15] が，どのようなユーザの情報をどの程度まで利用するのが良いか，評価実験を通して検証した．

以降，**2.** でユーザ属性推定に関する既存研究を紹介し，**3.** 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.** で提案手法の詳細について説明し，**5.** でその有効性を評価，最後の **6.** でまとめを行う．

2. 関 連 研 究

コンテンツやソーシャルグラフを用いたユーザ属性推定に関する既存研究について，それぞれ示す．

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(注7)数，フレンド(注8)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(注9)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定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(注10)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，人種，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(注11)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）

(注7)：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
(注8)：ユーザがフォローしている人
(注9)：他人のツイートを引用してツイートすること
(注10)：https://www.mturk.com/
(注11)：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

による判別手法を提案している．Mislove ら [9] は，地理，性別，人種に関して 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．

本研究は汎用的な手法の確立を目指しているため，フォロワー数など Twitter 特有の情報を特徴量として用いず，bag-of-words のみを特徴量としている．また，属性ごとに推定手法を変えない．池田ら [1] の手法はこれらの条件に一致するため，評価実験におけるベースラインとした．

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，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．

本研究は実用的な手法の確立を目指しているため，推定対象ユーザと 1hop の関係にあるユーザのみの情報を用いて推定を行う．Zamal ら [15] の手法はこの条件に一致するため，評価実験におけるベースラインとした．

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

本研究で用いた Twitter プロフィールの構造を表 1 に示す．description および location はユーザ属性が記述される可能性のある項目である．statuses_count は現在までの総ツイート数を示す項目である．url は外部 URL を自由記述する項目である．

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．なお，複数マッチする場合は文頭に近いものを採用した．

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

表 2　Twitter プロフィールの分析に用いたデータ

項目名	値
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 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．なお，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．

さらに，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．

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 4　プロフィールに含まれる外部メディア

順位	ドメイン	ユーザ数	順位	ドメイン	ユーザ数
1	ameblo.jp	159,768	6	d.hatena.ne.jp	12,348
2	blog*.fc2.com	20,407	7	jugem.jp	11,991
3	facebook.com	20,237	8	blogspot.com	11,752
4	blog.livedoor.jp	16,500	9	exblog.jp	10,706
5	mixi.jp	16,289	10	tumblr.com	10,647

4. 提 案 手 法

コンテンツベース手法とソーシャルグラフベース手法の大きく 2 つの手法を提案し，**3.** で明らかになった課題を解決する．

4.1 コンテンツベース手法

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** 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．

そこで，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．

4.2 ラベル伝播学習法

Blog はプロフィールを属性ごとに記述する様式である場合が多く，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．そのため，Blog の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既存研究が存在する [2]．したがって，共通ユーザを発見することができれば，Twitter の教師ラベルとして Blog のプロフィールを利用し，人手を必要とせず自動的に推定器を構築することができる．

共通ユーザがどのくらいいるのか，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．この結果から，最もユーザ数の多いドメインでは約 16 万人の共通ユーザがいることがわかった．また，様々な Blog を扱うとすると，共通ユーザは数万〜数十万人になることがわかった．人手による教師ラベル作成は，コスト面から数千件程度に留まることが多いのに対し，共通ユーザを利用した教師ラベル作成は，その 10〜100 倍のデータ量となる．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．

そこで本研究では，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，推定器にロジスティック回帰を用いた．

4.3 プロフィール文書も利用した推定器構築手法

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，利用による推定精度の向上が期待できる．しかし，ツイート集合に対して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．そこで，本研究では以下に示す 9 種類の手法を提案し，**5.** にてその効果を検証する．

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．

$$P(u) = R_p(u)P_p(u_p) + R_t(u)P_t(u_t) \tag{1}$$

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

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

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

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

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．

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

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

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．

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

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

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

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 と特徴量の種類数 $|F|$ によって (10) 式の通りに定められる．

4.4　ソーシャルグラフベース手法

ソーシャルグラフ上の近隣ユーザは似た属性を持つ傾向があることが知られている [15]．近隣ユーザの情報を利用することで，コンテンツからの情報が少ない場合でも安定したユーザ属性推定が行える可能性がある．

本研究では，Zamal ら [15] の手法を拡張し，推定対象ユーザと近隣ユーザの情報量を調節することで推定精度を高める手法を提案する．プロフィール文書のみ（PR），ツイート集合のみ（TW），両方（TP）という 3 つの情報量制約を考え，推定対象ユーザは TW または TP を，近隣ユーザはすべてを候補とし，それらの組み合わせによる精度変化を評価することで最適な調節方法を検討した．なお，Zamal らの手法は推定対象ユーザと近隣ユーザ共に TW を用いる場合に相当する．また，Zamal らの近隣ユーザの選び方を採用し，すべて（ALL），フォロワーの多い上位 N 人（MOST），フォロワーの少ない上位 N 人（LEAST），会話回数の多い上位 N 人（CLOSEST）の 4 つを比較した．N の値は 10 を用い，近隣ユーザの情報から得る特徴量はそのユーザ数で平均化して用いた．さらに，推定対象ユーザと近隣ユーザの特徴量を平均化する（AVG），連結する（JOIN）ことによる違いも評価した．

表 5　評価実験データ

		共通ユーザ	Blog ユーザ
ユーザ数	性別(注13)	71,129	49,739
	年齢(注14)	36,234	17,689
	職業(注15)	41,920	37,427
	興味(注16)	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	-

なお，ソーシャルグラフは，フレンド/フォロワー関係，会話関係など様々なリンク情報を用いて構築できるが，本研究では会話関係を利用した．フレンド/フォロワー関係は REST API を通して取得できるが，2012 年 12 月現在，350 call/h しか API が利用できないという制約があり，取得は非常に困難である．一方，会話関係は Streaming API を通して取得されるツイートのみから取得が可能であり，REST API を利用しなくても良い．また，会話をするという能動的な行動を元にしたグラフであるため，フレンド/フォロワー関係よりもより親密なグラフが構築できるという特徴がある．

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，推定器は LIBLINEAR(注12) の 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．なお，特徴量数と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．プロフィール文書推定器の特徴量数は 5,000，ツイート集合推定器の特徴量数は 30,000 であった．

5.1　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)

(注12)：http://www.csie.ntu.edu.tw/~cjlin/liblinear/
(注13)：男性，女性（2 クラス）
(注14)：10 代，20 代，30 代，40 代以上（4 クラス）
(注15)：主婦，会社員，中高生など（8 クラス）
(注16)：音楽，スポーツ，ゲームなど（20 クラス）

表 6　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**

表 7　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)	

と，正規表現によるマッチングの後，第一著者がひとりでプロフィール文書を目視確認し，正しいと判断し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 6 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できる．そのため，高精度な推定器が構築できることがわかった．

5.2　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，誤った教師ラベルで学習されてしまう可能性がある．書き分けの例として，Blog では旅行を趣味としており旅行記などを投稿しているが，Twitter では日常のことしか投稿しておらず，旅行の内容がほとんど投稿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，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，上記の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 7 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．これらの結果から，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．

5.3　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向上させることができるか，**4.3** 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 によって評価した．

実験結果を表 8 および図 1〜4 に示す．表 8 では，手法，属性ごとの Accuracy と手法ごとの平均順位を掲載した．また，図 1〜4 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 8 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

表 8　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

表 9　ソーシャルグラフベース手法の比較

	性別	年齢	職業	興味
PROF	65.78	47.90	36.10	40.03
TWEET	85.25	61.38	46.68	47.53
TWEPRO	85.75	61.88	47.51	53.49
NBR-ALL	69.87	**63.28**	42.16	40.69
NBR-MOST	68.57	58.33	37.18	37.01
NBR-LEAST	70.13	61.59	41.43	37.34
NBR-CLOSEST	68.31	59.24	39.58	37.06
AVG-TWTW-ALL	74.25	**64.38**	44.01	44.45
AVG-TWTW-MOST	72.25	60.75	41.26	42.49
AVG-TWTW-LEAST	73.00	**64.13**	44.58	43.22
AVG-TWTW-CLOSEST	71.50	61.38	41.52	43.54
JOIN-TWTW-ALL	83.00	**64.88**	47.45	48.00
JOIN-TWTW-MOST	80.25	**62.13**	45.03	46.98
JOIN-TWTW-LEAST	83.00	**63.63**	47.32	47.01
JOIN-TWTW-CLOSEST	79.25	**64.00**	45.92	47.58
JOIN-TPPR-ALL	**86.75**	**65.00**	**48.09**	**54.45***
JOIN-TPPR-MOST	**86.75**	**63.88**	**48.21**	**54.27**
JOIN-TPPR-LEAST	**86.25**	**64.63**	**48.41**	**53.64**
JOIN-TPPR-CLOSEST	**87.25***	**64.25**	**48.66**	**53.75**
JOIN-TPTW-ALL	83.00	**65.13**	**48.72***	52.81
JOIN-TPTW-MOST	80.50	**62.50**	46.17	51.50
JOIN-TPTW-LEAST	84.25	**63.75**	**48.21**	51.50
JOIN-TPTW-CLOSEST	80.25	**63.50**	**47.64**	52.21
JOIN-TPTP-ALL	82.50	**66.63***	**48.28**	52.73
JOIN-TPTP-MOST	79.75	**62.88**	46.62	51.76
JOIN-TPTP-LEAST	83.25	**64.25**	**47.83**	51.35
JOIN-TPTP-CLOSEST	80.25	**64.00**	47.07	51.87

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．

図 1〜4 では Coverage ごとの詳細を知ることができ，ほぼ単調に右肩下がりであること，低 Coverage 領域で手法の順位変動は多少あるが全体としてほぼ一定であることなどが見て取れる．図 3 のみ PROF の Accuracy が TWEET を低 Coverage 領域で上回っているのは，職業では自信を持って出力されたプロフィール文書推定器の結果が正解しやすいためである．このため，プロフィール文書とツイート集合それぞれの推定器を平等に扱う AVG が職業で最も良い精度をあげた．したがって，属性によってプロフィール文書利用の効果が異なるため，最適な手法も異なることがわかった．

5.4　ソーシャルグラフベース手法

表 5 に示したデータのうち，各クラス最大 200 件に制限をかけたデータを対象として，**4.4** に示したソーシャルグラフベース手法の評価実験を行った結果を表 9 に示す．手法の名前（AVG-TWTW-ALL など）は，[混合/結合方法]-[推定対象ユーザ情報][会話ユーザ情報]-[会話ユーザ利用方法] で表現している．比較対象として，推定対象ユーザのプロフィール文書のみ（PROF），ツイート集合のみ（TWEET），両方（TWEPRO; **4.3** の JOIN に相当）をそれぞれ学習データとしたものと，会話ユーザのツイート集合のみ（NBR）を学習データとしたものも合わせて掲載した．TWEPRO を上回る精度を太字，属性ごとで最高精度のものに*を付与している．実験はすべて 10-Fold Cross Validation によって評価した．

表 9 から，Zamal ら [15] の手法（*-TWTW-*）は年齢のみ精度が向上しているのに対し，提案手法（JOIN-TPPR-*）はすべての属性に対して精度が向上しており，提案手法の方が有意に良かった．また，JOIN-TPPR-*は JOIN-TPTW-*や JOIN-TPTP-*よりも使用している情報量が少ないにも関わらず高精度であった．このことから，会話ユーザの情報はノイズを含みやすいため，安定して高精度な推定を行うためには，会話ユーザの利用はプロフィール文書までに抑えるのが良いことがわかった．また，特徴量の組み合わせ方は AVG よりも JOIN の方が良いことや，属性によって類は友を呼ぶ関係の強さが変わるため，会話ユーザ情報の最適な利用量は異なることもわかった．

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（プロフィール文書とツイート集合）とソーシャルグラフ（会話関係）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 3 点において既存研究にはない知見を示した．1 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．2 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ことで，推

定精度が高められることを示した点である．ツイート集合とプロフィール文書の組み合わせ方について網羅的な検証を行った結果，本研究で提案した JOIN と AIC が有効であることがわかった．3 つめは，推定対象ユーザと会話ユーザの情報量を調節することで推定精度が高められることを示した点である．会話ユーザの情報量をプロフィール文書に抑えるとあらゆるユーザ属性で精度が良く，類は友を呼ぶ関係が強いユーザ属性ではさらに情報量を増やすことで精度が良くなることがわかった．

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．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，どのような会話ユーザが精度向上に寄与するか，ユーザ属性との関係性を鑑みながら今後調査したい．

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] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying Latent User Attributes in Twitter. In SMUC, pp. 37-44, 2010.

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

[5] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In EMNLP, pp. 1277-1287, 2010.

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

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

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

[9] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the Demographics of Twitter Users. In ICWSM, 2011.

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

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

[12] Z. Wen and C.-Y. Lin. Improving User Interest Inference from Social Neighbors. In CIKM, pp. 1001-1006, 2011.

[13] J. He, W. W. Chu, and Z. V. Liu. Inferring Privacy Information From Social Networks. In ISI, pp. 154-165, 2006.

[14] E. Zheleva and L. 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] F. A. Zamal, W. Liu and D. Ruths. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In ICWSM, 2012.

[16] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring Private Information Using Social Network Data. In WWW, pp. 1145-1146, 2009.

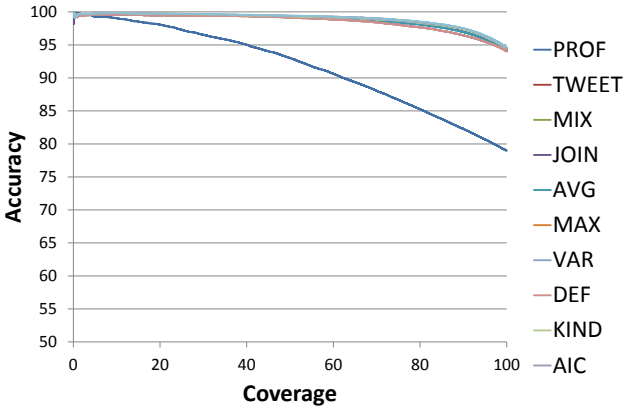

図 1　性別の Accuracy/Coverage Curve

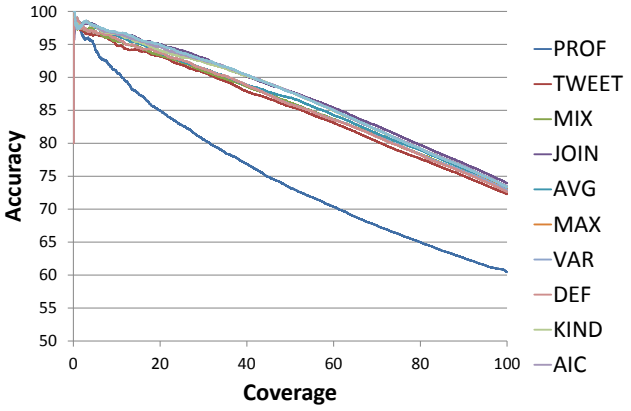

図 2　年齢の Accuracy/Coverage Curve

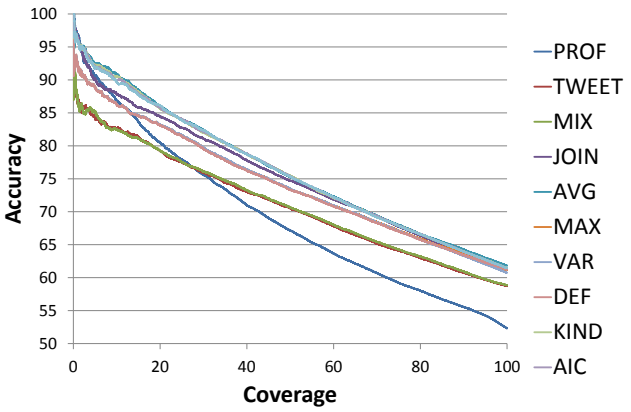

図 3　職業の Accuracy/Coverage Curve

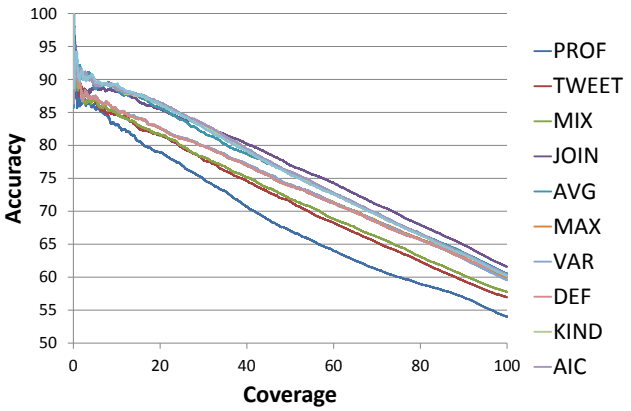

図 4　興味の Accuracy/Coverage Curve