DBSJ Journal, Vol.12, No.1
June 2013

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

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

伊藤 淳 ♥　西田 京介 ♦
星出 高秀 ♠　戸田 浩之 ♣
内山 匡 †

Jun ITO　Kyosuke NISHIDA
Takahide HOSHIDE　Hiroyuki TODA
Tadasu UCHIYAMA

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で，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．

We propose a new method for estimating user attributes of a Twitter user from the user's contents (a profile document and tweets) and social neighbors, i.e. those with whom the user has mentioned. Our method finds Twitter users associated with blog account and uses their profile attributes on blog as true labels of training tweet data. It also utilizes the profile documents of the users and the users' social neighbors. Overall experiments conducted on the estimation of the four attributes (gender, age, occupation, and interests) show that our method achieved higher accuracy than conventional methods that use manually-labeled tweets.

♥ 正会員　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　ito.jun@lab.ntt.co.jp
♦ 正会員　NTT レゾナント株式会社　k-nishi@nttr.co.jp
♠ 非会員　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　hoshide.takahide@lab.ntt.co.jp
♣ 正会員　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　toda.hiroyuki@lab.ntt.co.jp
† 非会員　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　uchiyama.tadasu@lab.ntt.co.jp

1. はじめに

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2013年3月に2億人を超えた[4]．また，Facebookは2012年3月に1日あたり32億件のいいね！やコメント，3億件の写真投稿があり[5]，Twitterは2013年3月に1日あたり4億件のツイート（投稿文書）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 [1]．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Buzz Finder，なずきのおと，感゜レポート，クチコミ@係長などがあり，国外ではRadian6，Sysomos，Forsightなどがある．

一般に，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できたが，ソーシャルメディアで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．我々はTwitterにおける属性の記述率を予備実験として調査した．Twitterではdescription項目中に自己紹介文（プロフィール文書）を自由記述できるようになっている．事前に人手で定義した属性辞書中の単語が含まれるかどうかを，収集した4,638,441ユーザの年齢，性別，職業について調べた．表1を見るとわかる通り，プロフィール文書の記述率は約83％と高いにも関わらず，属性の記述率は14％以下

[1] http://www.facebook.com/
[2] https://twitter.com/
[3] http://newsroom.fb.com/Key-Facts
[4] http://blog.twitter.com/2013/03/celebrating-twitter7.html
[5] http://mashable.com/2012/04/23/facebook-now-has-901-million-users/

表 1　プロフィール文書と属性の記述数と記述率.

	記述数	記述率
プロフィール文書	3,827,885	82.53
性別	353,558	7.62
年齢	154,900	3.34
職業	631,626	13.62

と低かった．これにより，属性を抽出する方法では，商品やコンテンツに対する意見や感想を述べたユーザの一部しか分析できないことがわかった．

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ことにより，プロフィール文書に属性の記述がないユーザの意見や感想も，推定した属性を用いて集計・分析することができるようになる．

2. 関連研究と本研究の貢献

ユーザ属性を推定する研究は Blog が登場した頃から行われている．投稿文書を bag-of-words で表現し，教師あり学習で分類問題を解く手法が一般的である [2]．Twitter でも同様の手法で成果を上げた研究 [3] はあるが，2 つの問題点から Twitter は Blog よりも難しいタスクになっている．

1 つめは，学習に必要な教師ラベルの抽出が困難なことである．Blog は性別や年齢などの属性を項目ごとにプロフィールへ記載することができる．その多くはプルダウンリストから選択する形式になっているため，ルールによって教師ラベルを自動抽出できる [2]．一方，Twitter はプロフィール文書を自由記述するため，ルールによる自動抽出は困難である．例えば，子どもやペットなど本人以外の属性が記述されていたり，“男性声優” のように対象としている属性（性別）とは異なる属性（興味）が抽出されたりして，抽出ノイズが発生する．これらの抽出ノイズをルールによって網羅的に除外するのは難しいため，どうしても人手によるチェックが必要となり，手間がかかる．

2 つめは，ツイートの文書長が短く，また文書数が少ないことである．Twitter は 1 ツイートの文書長が 140 字以内に制限されているうえ，API の制限から 1 人あたり最大 3,200 ツイートしか取得することができない．さらに，1. 章で行った予備実験のユーザにおいて総ツイート数を調査したところ，平均値は 68.1，中央値は 553 であり，多くのユーザが最大取得可能数に達しないことがわかった．比較的まとまった記事が投稿される Blog よりも情報量が少なくなるため，フォロワー数などの言語情報以外の特徴量を利用したり [4, 5]，Facebook など他のメディアの研究 [6, 7, 8, 9] と同様にソーシャルグラフを利用したり [5, 10] して，情報量の少なさを補って推定精度を向上させる研究が行われている．

表 2　プロフィールに含まれる外部ドメイン.

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

本研究では，これら 2 つの問題点を解決するために 3 つの貢献をした．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．Burger ら [11] も同様の手法を提案しているが，本研究ではラベル伝播の信頼性や人手ラベルと比較した有効性を評価している点が異なる．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手法を網羅的に検証し，最適な組み合わせ方法を提案した点である．ツイート集合とプロフィール文書に出現する単語は，同じ単語でも別々に扱った方が良いことも示した．3 つめは，推定対象ユーザに加えて会話ユーザの情報も利用する手法を網羅的に検証し，会話ユーザの最適な利用方法を提案した点である．同類性の強い属性ほど情報量を増やすことで高精度になることも示した．

3. 提案手法

ラベル伝播学習法，プロフィール文書の利用手法，会話ユーザの利用手法という 3 つの手法について本章で述べる．

3.1 ラベル伝播学習法

Blog はユーザ属性を項目ごとにプロフィールへ記載することができるため，Twitter のような自由記述形式のプロフィール文書と違って，ルールを用いて属性を自動抽出することができる．Blog のプロフィールを教師ラベルとして信頼し，Blog ユーザの属性を高精度に推定した研究がある [2]．したがって，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．

共通ユーザがどれくらい存在するか，1. 章で行った予備実験のユーザを調査した．外部 URL が記述できる url 項目をドメインごとに集計し，上位 10 件を抽出した．表 2 の結果から，トップのドメインで約 16 万人，10 位までの Blog を合計すると 24 万人以上の共通ユーザが存在することがわかった．人手による教師ラベル作成は手間がかかるため，学習データは数千件程度になることが多い [3, 4, 5] が，共通ユーザを利用すればその 10 倍以上の学習データを自動で得ることができる．一般に，学習データ量を増やすほど推定精度は向上するため，高精度な推定器の構築が期待できる．

Burger ら [11] も同様の手法を提案しているが，彼らは性別のみについて，ランダムサンプリングした 1,000 ユーザを人手で確認することで，ラベル伝播は信頼して良いとしている．しかし，Blog プロフィールに嘘を書くユーザや，Blog と

Twitter の投稿内容が異なるユーザが存在するため，そのようなユーザを学習データに含めて良いのか疑問が残る．そこで我々は，4.1 節で人手によるラベリング手法と精度比較をしたうえで，4.2 節でそのようなユーザを学習データに含めることの影響を検証し，ラベル伝播学習法の有効性を示す．

3.2 プロフィール文書の利用手法

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，4.3 節にてその効果を比較する．なお，本研究では，ツイート集合やプロフィール文書の bag-of-words から赤池情報量基準（AIC）を用いて選択した特徴量 [3] のみを用いて検証を行い，フォロワー数など Twitter 特有の特徴量は用いない．

MIX　推定器をひとつ構築する．その際，プロフィール文書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．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 (2)，(3) 式の通り，使用された特徴量の種類数によって定まる．

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

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

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

$$I_p(u_p) = -\log\left(\frac{\mathrm{kind}(u_p) + \alpha}{|F_p|}\right), \tag{2}$$

$$I_t(u_t) = -\log\left(\frac{\mathrm{kind}(u_t) + \alpha}{|F_t|}\right). \tag{3}$$

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 (2)，(3) 式を，(4)，(5) 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．

$$I_p(u_p) = -\log\left(\frac{\sum_{s\in\mathrm{set}(u_p)}\mathrm{aic}(s) + \alpha}{\sum_{f\in F_p}\mathrm{aic}(f)}\right), \tag{4}$$

$$I_t(u_t) = -\log\left(\frac{\sum_{s\in\mathrm{set}(u_t)}\mathrm{aic}(s) + \alpha}{\sum_{f\in F_t}\mathrm{aic}(f)}\right). \tag{5}$$

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 (2)，(3) 式を，(6)，(7) 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．

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

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

$$\mathrm{rank}(f) = \frac{|F|}{\mathrm{index}(f)}. \tag{8}$$

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 (8) 式の通りに定められる．ここで，F は F_p または F_t を意味する．

3.3 会話ユーザの利用手法

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 [10]．近隣ユーザの情報を利用することで，コンテンツの情報が少ない場合でも高精度に推定できる可能性がある．

本研究では，Zamal ら [10] 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 4.4 節で評

表 3　評価実験データ.

		共通ユーザ	Blog ユーザ
ユーザ数	性別 [1]	71,129	49,739
	年齢 [2]	36,234	17,689
	職業 [3]	41,920	37,427
	興味 [4]	20,846	7,417
	全体	86,183	65,873
Blog 記事数		796,583	626,903
ツイート数 [5]		15,124,094	-

[1] 男性，女性（2 クラス）.
[2] 10 代，20 代，30 代，40 代以上（4 クラス）.
[3] 主婦，会社員，中高生など（8 クラス）.
[4] 音楽，スポーツ，ゲームなど（20 クラス）.
[5] ユーザごとに最大 200 ツイート（リツイートを除く）.

価し，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）ことによる違いも比較した．

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，2013 年 6 月現在，15 call/15 min しか API を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．

4. 評価実験

表 3 のデータを用いて提案手法の評価実験を行った．特徴量選択は池田ら [3] と同様に赤池情報量基準（AIC）を，推定器は LIBLINEAR[6] 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，特徴量数とコスト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，特徴量数はプロフィール文書とツイート集合でそれぞれ 5,000 と 30,000 を，コストパラメータは 1 を用いた．

4.1 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) の有効性を示すため，性別と年齢について人手によるラベリング手法との比較を行った．人手で定義した属性辞書の単語がプロフィール文書中に含まれているかを正規表現でチェックする手法 (REGEXP) と，そこからさらに第一著者が目視確認を行い，正しいと判断したもののみを用いる手法 (HUMAN) を比較対象とした．こ

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

表 4　ラベリング手法の精度比較.

	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 5　学習データフィルタリング手法の精度比較.

	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)	

こで，HUMAN は池田ら [3] の手法に相当する．REGEXP，HUMAN は属性を構成するクラスごとに 1,000 件の学習データを用いており，表 3 の全データを用いる DIRECT よりも少ない．そこで，DIRECT の学習データ量を HUMAN，REGEXP と揃えた D1000 も比較した．DIRECT のデータを 5 分割し，テストデータに 1 つを，訓練データに DIRECT は残りの 4 つを，D1000 はそこから各クラス 1,000 件ランダムに選択したものを，REGEXP，HUMAN はそれぞれの全データを用いる 5-Fold Cross Validation で評価した．

表 4 の通り，REGEXP，HUMAN，D1000，DIRECT の順で精度が良かった．HUMAN が D1000 に劣ったのは，プロフィール文書にユーザ属性を記述する少数派のユーザ（1. 章の予備実験を参照）を学習データとした偏りによるものだと考える．DIRECT が最も良かったのは，学習データの量が他の手法より多い（性別 35 倍，年齢 9 倍）ためだと考える．

この実験により，ラベル伝播学習法は自動的に大量の学習データを収集できるため，人手によるラベリング手法よりも手間をかけずに高精度な推定器が構築できることを示した．

4.2 Blog ラベル伝播の妥当性検証

Blog プロフィールに嘘を書くユーザや，Blog と Twitter の投稿内容が異なるユーザを学習データに含めることの影響を調査するため，ツイート集合と Blog 記事の投稿内容に食い違いがない共通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．大きく 2 つの手法を比較した．

1 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 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，共通ユーザの Blog プロフィールと推定結果が合致するもののみを採用するという手法である．共通ユーザの Blog 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) を比較した．

しかし，Blog 推定器を用いるこれらの手法では，Blog 推定器そのものの精度が問題になる．そこで 2 つめは，ツイート集合と Blog 記事をそれぞれ bag-of-words の単語出現頻度ベクトルで表現し，それらのコサイン類似度が 0.8 以上のも

表 6　プロフィール文書の利用手法の精度比較.

	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

ののみを採用するという手法 (COS) を比較した.

また，Blog プロフィールを信頼してそのまま伝播する手法 (DIRECT) と，Blog 推定器で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) も合わせて比較した．性別，年齢，職業，興味について，データを 5 分割し，1 つをテストデータ，残りをフィルタリングしてから訓練データとする 5-Fold Cross Validation で評価した.

実験結果は表 5 の通りであり，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す．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 %で有意差は無かった（McNemar 検定）．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.

この実験により，Blog プロフィールの嘘や，Blog と Twitter における投稿内容の書き分けの影響は少なく，Blog ラベルをそのまま伝播させることで様々な属性に対して高精度な推定が行えることを示した.

4.3　プロフィール文書の利用手法の比較

ツイート集合に対して，プロフィール文書をどのように組み合わせて利用すると良いかを明らかにするため，3.2 節で示したプロフィール文書の利用手法の比較を行った．プロフィール文書のみ (PROF) とツイート集合のみ (TWEET) を用いた精度も比較のために掲載した．性別，年齢，職業，興味について，5-Fold Cross Validation で評価した.

表 6 に，手法，属性ごとの Accuracy と，手法ごとの平均順位を示す．JOIN と AIC の平均順位が最も良く，TWEET に対しすべての属性において McNemar 検定で有意水準 5 %の有意差が認められた．また，プロフィール文書とツイート集合における単語の扱い方のみが異なる JOIN と MIX では，JOIN の方がすべての属性で高精度であった.

この実験により，プロフィール文書を加えると精度は向上すること，その際 JOIN または AIC が最も効果的であることを示した．また，ツイート集合とプロフィール文書に出現する単語は，同じ単語でも別々に扱った方が良いことも示した.

4.4　会話ユーザの利用手法の比較

会話ユーザの最適な利用方法を明らかにするため，3.3 節に示した手法の比較を行った．表 3 のユーザすべてを利用すると，その会話ユーザの数は膨大になるため，表 3 から各クラス最大 200 人ランダムサンプリングしたユーザと，その会話ユーザのデータを用いて実験を行った．結果は表 7 の通りである．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．3.2 節の JOIN に相当する TWEPRO と，会話ユーザのツイート集合のみを用いた NBR を比較のため掲載した．TWEPRO を上回る精度を太字，属性ごとで最高精度のものに*を付与した．性別，年齢，職業，興味について 10-Fold Cross Validation によって評価した.

表 7　会話ユーザの利用手法の精度比較.

	性別	年齢	職業	興味
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

Zamal ら [10] の手法（*-TWTW-*）は年齢のみ TWEPRO より精度が向上したが，提案手法（JOIN-TPPR-*）はすべての属性で向上した．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，すべての属性で TWEPRO よりも安定して高精度であった．年齢は NBR-ALL が TWEPRO より高精度であり，同類性が強い．そのため，情報量を増やすほど精度が向上する傾向にあるが，同類性が弱い性別や興味では逆に精度が低下した.

この実験により，会話ユーザのプロフィール文書のみを加えるのが有効であることを示した．また，同類性の強い属性ほど情報量を増やすことで高精度になることも示した.

5. まとめ

本研究では，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，Twitterユーザの属性（性別，年齢，職業，興味）を推定する手法を提案した．教師ラベルの抽出が困難であり，推定に使用できる言語情報量が少ないという2つの問題点を解決するため，3つの貢献を行った．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．Blogプロフィールの嘘や，BlogとTwitterにおける投稿内容の書き分けの影響は少なく，ラベルをそのまま伝播させることで様々な属性に対して人手ラベリングよりも高精度な推定が行えることを示した．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手法を網羅的に検証し，3.2節で示したJOINとAICが有効であ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に出現する単語は，同じ単語でも別々に扱った方が良いことも示した．3つめは，推定対象ユーザに加えて会話ユーザの情報も利用する手法を網羅的に検証し，会話ユーザのプロフィール文書のみを加えるのが有効であることを示した点である．同類性の強い属性ほど情報量を増やすことで高精度になることも示した．

本稿では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．

[文献]

[1] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 11, pp. 2169-2188, 2009.

[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] 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.

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

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

[6] 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.

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

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

[9] 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.

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

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

伊藤 淳 Jun ITO

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 社員．2011年早稲田大学大学院基幹理工学研究科情報理工学専攻修士課程修了．同年日本電信電話株式会社入社．Webマイニングの研究開発に従事．情報処理学会，日本データベース学会，各会員．

西田 京介 Kyosuke NISHIDA

NTTレゾナント株式会社．2004年北海道大学工学部情報工学科卒業．2006年同大学大学院情報科学研究科修士課程修了．2008年同博士課程修了．同年日本電信電話株式会社入社．2013年より現職．Webマイニングの研究開発に従事．博士（情報科学）．情報処理学会，電子情報通信学会，日本データベース学会，各会員．

星出 高秀 Takahide HOSHIDE

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 主任研究員．1993年九州大学大学院総合理工研究科修士課程修了．同年日本電信電話株式会社入社．遠隔教育システム，Webマイニングの研究開発に従事．情報処理学会正会員．

戸田 浩之 Hiroyuki TODA

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 主任研究員．1999年名古屋大学大学院工学研究科博士課程前期課程修了．同年日本電信電話株式会社入社．以来，情報検索，情報抽出，Webマイニングの研究開発に従事．2007年筑波大学大学院システム情報工学研究科博士後期課程修了．博士(工学)．ACM，情報処理学会，日本データベース学会，各会員．

内山 匡 Tadasu UCHIYAMA

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 主幹研究員．1987年名古屋大学大学院理学研究科修士課程修了．同年日本電信電話株式会社入社．1998-2001年NTTコミュニケーションズ，2004-2006年NTTレゾナントにてポータルサービスの開発等に従事．2007年より現職．行動モデリングの研究開発に従事．情報処理学会，電子情報通信学会，日本応用数理学会，各会員．